

IntechOpen

# Data Mining and Knowledge Discovery in Real Life Applications

*Edited by Julio Ponce and Adem Karahoca*





**DATA MINING AND KNOWLEDGE DISCOVERY  
IN REAL LIFE APPLICATIONS**

EDITED BY  
JULIO PONCE  
AND  
ADEM KARAHOGA

## **Data Mining and Knowledge Discovery in Real Life Applications**

<http://dx.doi.org/10.5772/97>

Edited by Julio Ponce and Adem Karahoca

### **© The Editor(s) and the Author(s) 2009**

The moral rights of the and the author(s) have been asserted.

All rights to the book as a whole are reserved by INTECH. The book as a whole (compilation) cannot be reproduced, distributed or used for commercial or non-commercial purposes without INTECH's written permission.

Enquiries concerning the use of the book should be directed to INTECH rights and permissions department ([permissions@intechopen.com](mailto:permissions@intechopen.com)).

Violations are liable to prosecution under the governing Copyright Law.



Individual chapters of this publication are distributed under the terms of the Creative Commons Attribution 3.0 Unported License which permits commercial use, distribution and reproduction of the individual chapters, provided the original author(s) and source publication are appropriately acknowledged. If so indicated, certain images may not be included under the Creative Commons license. In such cases users will need to obtain permission from the license holder to reproduce the material. More details and guidelines concerning content reuse and adaptation can be found at <http://www.intechopen.com/copyright-policy.html>.

### **Notice**

Statements and opinions expressed in the chapters are those of the individual contributors and not necessarily those of the editors or publisher. No responsibility is accepted for the accuracy of information contained in the published chapters. The publisher assumes no responsibility for any damage or injury to persons or property arising out of the use of any materials, instructions, methods or ideas contained in the book.

First published in Croatia, 2009 by INTECH d.o.o.

eBook (PDF) Published by IN TECH d.o.o.

Place and year of publication of eBook (PDF): Rijeka, 2019.

IntechOpen is the global imprint of IN TECH d.o.o.

Printed in Croatia

Legal deposit, Croatia: National and University Library in Zagreb

Additional hard and PDF copies can be obtained from [orders@intechopen.com](mailto:orders@intechopen.com)

Data Mining and Knowledge Discovery in Real Life Applications

Edited by Julio Ponce and Adem Karahoca

p. cm.

ISBN 978-3-902613-53-0

eBook (PDF) ISBN 978-953-51-5835-6

# We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,400+

Open access books available

118,000+

International authors and editors

130M+

Downloads

151

Countries delivered to

Our authors are among the  
Top 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index  
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?  
Contact [book.department@intechopen.com](mailto:book.department@intechopen.com)

Numbers displayed above are based on latest data collected.  
For more information visit [www.intechopen.com](http://www.intechopen.com)





## Preface

Knowledge Discovery and Data Mining are powerful data analysis tools. Data mining (DM) is an important tool for the mission critical applications to minimize, filter, extract or transform large databases or datasets into summarized information and exploring hidden patterns in knowledge discovery (KD). Sort of the data mining algorithms can be used in different real life applications. Classification, prediction, clustering, segmentation, summarization can be done on the large amount of data by using DM methods. The rapid dissemination of these technologies calls for an urgent examination of their social impact. The terms "Knowledge Discovery" and "Data Mining" are used to describe the 'non-trivial extraction of implicit, previously unknown and potentially useful information from data. Knowledge discovery is a concept that describes the process of searching on large volumes of data for patterns that can be considered knowledge about the data. The most well-known branch of knowledge discovery is data mining.

Data mining is a multidisciplinary field with many techniques. With these techniques you can create a mining model that describe the data that you will use, some of these techniques are clustering algorithms, tools for data pre-processing, classification, regression, association rules, etc..

Data Mining is a fast-growing research area, in which connections among and interactions between individuals are analyzed to understand innovation, collective decision making, problem solving, and how the structure of organizations and social networks impacts these processes. Data Mining finds several applications; for instance, in different areas like biological, medicine, industrialist, business, economic, e-commerce, security, others.

This book presents four different ways of theoretical and practical advances and applications of data mining in different promising areas like Industrialist, Biological, and Social. Twenty six chapters cover different special topics with proposed novel ideas. Each chapter gives an overview of the subjects and sort of the chapters have cases with offered data mining solutions. We hope that this book will be served as a Data Mining bible to show a right way for the students, researchers and practitioners in their studies. The twenty six chapters have been classified in four corresponding parts.

- Knowledge Discovery
- Clustering and Classification
- Challenges and Benchmarks in Data Mining
- Data Mining Applications

The first part contains four chapters related to Knowledge Discovery. The focus of the contributions in this part, are the fundamental concepts and tools for extraction, representation, and retrieving of Knowledge in large Data Bases.

The second part contains seven chapters related to Clustering and Classification. The focus of the contributions in this part, are the techniques and tools used to made clusters and classify the data on the basis of specific characteristics.

The third part contains three chapters related to Challenges and Benchmarks in Data Mining. The focus of the contributions is the applications and instances of problems that can be challenges or benchmarks for the developed tools to realise the Data Mining.

The fourth part contains twelve chapters related to Data Mining Applications. The focus of the contributions in this part, are the real applications, e applications were classified in three topics that are Social, Biological and Industrialist, on the basis of the applications that are approached in them.

### **Part I: Knowledge Discovery**

Chapter 1 analysis SE (software engineering) process models and proposes a joint model based on two SE standards and claims that this comparison revealed that CRISP-DM does not cover many project management, organization and quality related tasks at all or at least thoroughly enough (Marbán, et.al.).

Chapter 2 offers an explanatory review on mining large-scale datasets and the extraction, representation, and retrieving of knowledge on Grid systems. Different trends and a domain independent solving environment ADMIRE also presented (Aouad, et.al.).

Chapter 3 focuses on the main advantages of rough set theory (RST) in data analysis. Also, mentioned that RST does not need any preliminary or additional information concerning data, such as basic probability assignment in Dempster-Shafer theory, grade of membership or the value of possibility in fuzzy set theory (Rissino and Lambert-Torres).

Chapter 4 introduces a novel robust data mining method by integrating a DM method for pre-processing unclear data and finding significant factors into a multidisciplinary RD method for providing the best factor settings (Shin, et.al.).

### **Part II: Clustering and Classification**

Chapter 5 presents association rule mining on the selection of meaningful association rules. As an application of semantic analysis and pattern analysis, real practice case study of traffic accidents are demonstrated (Marukatat).

Chapter 6 explores to compare hybrid cluster techniques for cognitive mapping with traditional intellectual subject-classifications schemes based on the external validation of clustering results by expert knowledge present in ISI subject categories (Janssens, et.al.).

Chapter 7 describes a system to inspect electronic components and devices, specifically, LCDs Panels that are the core parts of an LCD monitor store inspection result data in the RFID TAG and the Reader/Writer for efficient production control. C4.5 algorithm and Neural Nets are applied in the manufacturing process for TFT LCDs and suggest methods by which to locate defective parts (Kim, et.al.).

Chapter 8 addresses steps for processing the hyperspectral remote sensing images that atmospherically corrected before processing, and a endmember was extracted by PPI algorithm from the intersection area of multi-segmentation and geology map. Dioritic porphyrite area was extracted from hyperspectral remote sensing images by Spectral Angle Mapper (SAM), Multi Range Spectral Feature Fitting (Multi Range SFF) and Mixture Tuned Matched Filtering (MTMF) using the extracted end member respectively. Finally, classification results were outputted by combining three classification results using Classification and Regression Trees (CART) (Wen, et.al.).

Chapter 9 reports content based image classification via visual learning that is a learning framework which can autonomously acquire the knowledge needed to recognize images using machine learning frameworks (Nomiya and Uehara).

---

Chapter 10 presents a similarity based data mining on the trends of data streams, and shows that correlation coefficient method is better than Euclidean distance used data stream clustering methods. A novel partitioning algorithm offers for clustering parallel data streams based on correlation coefficients, which supports online clustering analysis over both fixed and flexible time horizons (Chen).

Chapter 11 describes an incremental mining algorithm for multiple-level association rules. It may lead to discovery of more specific and important knowledge from data. The proposed algorithm is based on the pre-large concept and can efficiently and effectively mine knowledge with taxonomy in an incremental way (Hong, et.al.).

### **Part III: Challenges and Benchmaks in Data Mining**

Chapter 12 reviews some known promises and challenges of the data mining applications and presents foot steps to achieve successful implementation strategies (Athauda, et.al.).

Chapter 13 discusses available techniques and current research trends in the field of spatiotemporal data mining. An overview of the proposed approaches to deal with the spatial and the temporal aspects of data has been presented. Approaches that aim at taking into account both aspects were also surveyed (Kechadi,et.al.).

Chapter 14 proposes a benchmarking for different data mining methods with adaptive neuro fuzzy inference networks (ANFIS) to investigate the best data mining model(s) for churn management in a telecom company (Karahoca, et.al.).

### **Part IV: Data Mining Applications**

#### *Social Applications*

Chapter 15 suggests Apriori association rule algorithm to investigate the behaviour of the video rental customers (Hsia,et.al.).

Chapter 16 offers a novel GCP model to practice global customer retention. It aims to find the useful pattern with the combination of IFS,  $\alpha$ -cuts, expert knowledge, and data mining techniques (Lin and Xu).

Chapter 17 describes the data mining in web applications and outlines Social Networks, Web radio, Security and Internet Frauds. Also, this chapter presents a conceptual model to develop systematically web radio applications taking into account social acceptability factor using the social data mining and cultural algorithms (Ponce, et.al.).

#### *Biological Applications*

Chapter 18 provides an overview of applied data-mining techniques to the analyses of data to investigate the status about the safety of medicine usage, such as the data in the database of medical near-miss cases which have been collected by Japanese government, the investigation data to understand how patients handle the injection device for anti-diabetic drug (Kimura).

Chapter 19 describes the processes of data mining in the molecular biology. Also, explained that the analysis of biological systems requires extensive use of bioinformatics resources for data management, mining, and modelling. Case study of carbohydrates biosynthesis and accumulation in plants is demonstrated (Vicentini and Menossi).

Chapter 20 provides a brief overview of microarray gene expression profiles data and some biological resources available to be used for pathway analysis and examines three kinds of data mining approaches; clustering-based methods, gene-based methods, and gene

set-based methods, which can be used in different contexts for understanding biological pathways and describes some case studies related with the leukemia disease data (Miyoung and Kim).

Chapter 21 addresses one of the important bioinformatics issue DNA sequences and presents development of microsatellite markers by using Data Mining and authors claims that development of SSRs by data mining from sequence data is a relatively easy and cost-saving strategy for any organisms with enough DNA data (Tong, et.al.).

#### *Industrialist Applications*

Chapter 22 proposes a knowledge based six-sigma model where DMAIC for six-sigma was used along with data mining techniques for identifying potential quality problems and assisting quality diagnosis in manufacturing processes (He, et.al.).

Chapter 23 explores a data mining process model and appropriate data mining based decision support system to support decision processes in the direct marketing in publishing sector and claims that direct mailing process gives positive results (Rupnik and Jaklic).

Chapter 24 presents a data mining based methodology to group and organize data from a dam instrumentation system aiming to assist dam safety engineers in order to select, cluster and rank 72 rods of 30 extensometers located at the F stretch of Itaipu's dam (Villwock, et.al.).

Chapter 25 considers a data mining algorithm for monitoring PCB assembly quality in order to minimize visual defects (Zhang).

Chapter 26 reviews data mining applications which are used in power systems. This chapter also presented results comparing different figures of merit for evaluating fault classification systems (Morais, et.al.).

January 2009

Editors

**Julio Ponce**

*Aguascalientes University  
Artificial Intelligence Laboratory  
Department of Computer Science  
Aguascalientes, Ags., Mexico  
E-mail: jcponce@correo.uaa.mx*

**Adem Karahoca**

*Bahcesehir University  
Faculty of Engineering  
Department of Software Engineering  
Ciragan Cad. Besiktas, Istanbul, Turkey  
E-mail : akarahoca@bahcesehir.edu.tr*

# Contents

Preface	VII
Part I: Knowledge Discovery	
1. A Data Mining & Knowledge Discovery Process Model <i>Óscar Marbán, Gonzalo Mariscal and Javier Segovia</i>	001
2. Knowledge Discovery on the Grid <i>Lamine Aouad, An Le-Khac and Tahar Kechadi</i>	017
3. Rough Set Theory – Fundamental Concepts, Principals, Data Extraction, and Applications <i>Silvia Rissino and Germano Lambert-Torres</i>	035
4. Robust Data Mining: An Integrated Approach <i>Sangmun Shin, Le Yang, Kyungjin Park and Yongsun Choi</i>	059
Part II: Clustering and Classification	
5. On the Selection of Meaningful Association Rules <i>Rangsipan Marukatat</i>	075
6. Hybrid Clustering for Validation and Improvement of Subject-Classification Schemes <i>Frizo Janssens, Lin Zhang and Wolfgang Glänzel</i>	089
7. Automatic Product Classification Control System Using RFID Tag Information and Data Mining <i>Cheonshik Kim, Eun-Jun Yoon, Injung Park and Taek-Young</i>	119
8. Hyperspectral Remote Sensing Data Mining Using Multiple Classifiers Combination <i>Xing-Ping Wen, Xiao-Feng Yang and Guang-Dao Hu</i>	129
9. Content-based Image Classification via Visual Learning <i>Hiroki Nomiya and Kuniaki Uehara</i>	141

---

10. Clustering Parallel Data Streams	167
<i>Yixin Chen</i>	
11. Mining Multiple-level Association Rules Based on Pre-large Concepts	187
<i>Tzung-Pei Hong, Tzu-Jung Huang and Chao-Sheng Chang</i>	
Part III: Challenges and Benchmarks in Data Mining	
12. Data Mining Applications: Promise and Challenges	201
<i>Rukshan Athauda, Menik Tissera and Chandrika Fernando</i>	
13. Mining Spatio-Temporal Datasets: Relevance, Challenges and Current Research Directions	215
<i>M-Tahar Kechadi, Michela Bertolotto, Filomena Ferrucci and Sergio Di Martino</i>	
14. Benchmarking the Data Mining Algorithms with Adaptive Neuro-Fuzzy Inference System in GSM Churn Management	229
<i>Adem Karahoca, Dilek Karahoca and Nizamettin Aydın</i>	
Part IV: Data Mining Applications	
<i>Social Applications</i>	
15. Using Data Mining to Investigate the Behavior of Video Rental Customers	243
<i>Tai-Chang Hsia and An-Jin Shie</i>	
16. A Novel Model for Global Customer Retention Using Data Mining Technology	251
<i>Jie Lin and Xu Xu</i>	
17. Data Mining in Web Applications	265
<i>Julio Ponce, Alberto Hernández, Alberto Ochoa, Felipe Padilla, Alejandro Padilla, Francisco Álvarez and Eunice Ponce de León</i>	
<i>Biological Applications</i>	
18. Application of Data Mining Techniques to the Data Analyses to Ensure Safety of Medicine Usage	291
<i>Masaomi Kimura</i>	

- 
- |  |     |
|--|-----|
| 19. Data Mining in the Molecular Biology Era – A Study Directed to Carbohydrates Biosynthesis and Accumulation in Plants<br><i>Renato Vicentini and Marcelo Menossi</i>              | 307 |
| 20. Microarray Data Mining for Biological Pathway Analysis<br><i>Miyoung Shin and Jaeyoung Kim</i>   | 319 |
| 21. Development of Microsatellite Markers by Data Mining from DNA Sequences<br><i>Jingou Tong, Dan Wang and Lei Cheng</i>  | 337 |
| <i>Industrialist Applications</i>  |     |
| 22. Quality Improvement using Data Mining in Manufacturing Processes<br><i>Shu-guang He, Zhen He, G. Alan Wang and Li Li</i>   | 357 |
| 23. The Deployment of Data Mining into Operational Business Processes<br><i>Rok Rupnik and Jurij Jaklič</i>  | 373 |
| 24. Data Mining Applied to the Instrumentation Data Analysis of a Large Dam<br><i>Rosangela Villwock, Maria Teresinha Arns Steiner, Andrea Sell Dyminski and Anselmo Chaves Neto</i> | 389 |
| 25. A Data Mining Algorithm for Monitoring PCB Assembly Quality<br><i>Feng Zhang</i>   | 407 |
| 26. An Overview of Data Mining Techniques Applied to Power Systems<br><i>Jefferson Morais, Yomara Pires, Claudomir Cardoso and Aldebaro Klautau</i>                                  | 419 |



## **PART I: KNOWLEDGE DISCOVERY**



# A Data Mining & Knowledge Discovery Process Model

Óscar Marbán<sup>1</sup>, Gonzalo Mariscal<sup>2</sup> and Javier Segovia<sup>1</sup>

<sup>1</sup> *Facultad de Informática (Universidad Politécnica de Madrid)*

<sup>2</sup> *Universidad Europea de Madrid*  
Spain

## 1. Introduction

The number of applied in the data mining and knowledge discovery (DM & KD) projects has increased enormously over the past few years (Jaffarian et al., 2008) (Kdnuggets.com, 2007c). As DM & KD development projects became more complex, a number of problems emerged: continuous project planning delays, low productivity and failure to meet user expectations. Neither all the project results are useful (Kdnuggets.com, 2008) (Eisenfeld et al., 2003a) (Eisenfeld et al., 2003b) (Zornes, 2003), nor do all projects end successfully (McMurchy, 2008) (Kdnuggets.com, 2008) (Strand, 2000) (Edelstein & Edelstein, 1997). Today's failure rate is over 50% (Kdnuggets.com, 2008) (Gartner, 2005) (Gondar, 2005).

This situation is in a sense comparable to the circumstances surrounding the software industry in the late 1960s. This was what led to the 'software crisis' (Naur & Randell, 1969). Software development improved considerably as a result of the new methodologies. This solved some of its earlier problems, and little by little software development grew to be a branch of engineering. This shift has meant that project management and quality assurance problems are being solved. Additionally, it is helping to increase productivity and improve software maintenance.

The history of DM & KD is not much different. In the early 1990s, when the KDD (Knowledge Discovery in Databases) processing term was first coined (Piatetsky-Shapiro & Frawley, 1991), there was a rush to develop DM algorithms that were capable of solving all the problems of searching for knowledge in data. Apart from developing algorithms, tools were also developed to simplify the application of DM algorithms. From the viewpoint of DM & KD process models, the year 2000 marked the most important milestone: CRISP-DM (CRoss-Industry Standard Process for DM) was published (Chapman et al., 2003). CRISP-DM is the most used methodology for developing DM & KD projects. It is actually a "de facto" standard.

Looking at the KDD process and how it has progressed, we find that there is some parallelism with the advancement of software. From this viewpoint, DM project development entails defining development methodologies to be able to cope with the new project types, domains and applications that organizations have to come to terms with. Nowadays, SE (software engineering) pay special attention to organizational, management or other parallel activities not directly related to development, such as project completeness

and quality assurance. The most used DM & KD process models at the moment, i.e. CRISP-DM, SEMMA, has not yet been sized for these tasks, as it is very much focused on pure development activities and tasks (Marbán et al., 2008). In (Yang & Wu, 2006) one of the 10 challenging problems to be solved in DM research is considered to be the need to build a new methodology to help users avoid many data mining mistakes.

This chapter is moved by the idea that DM & KD problems are taking on the dimensions of engineering problems. Hence, the processes to be applied should include all the activities and tasks required in an engineering process, tasks that CRISP-DM might not cover. The proposal is inspired by the work done in SE derived from other branches of engineering. It borrows ideas to establish a comprehensive process model for DM that improves and adds to CRISP-DM. Further research will be needed to define methodologies and life cycles, but the basis of a well-defined process model will be there.

In section 2 we describe existing DM & KD process models and methodologies, focusing on CRISP-DM. Then, section 3 shows the most used SE process models. In section 4, we propose a new DM & KD process model. And, finally, we discuss the conclusions about the new approach and future work in section 5.

## 2. DM & KD process models

Authors tend to use the terms process model, life cycle and methodology to refer to the same thing. This has led to some confusion in the field.

A process model is the set of tasks to be performed to develop a particular element, as well as the elements that are produced in each task (outputs) and the elements that are necessary to do a task (inputs) (Pressman, 2005). The goal of a process model is to make the process repeatable, manageable and measurable (to be able to get metrics).

Methodology can be defined as the instance of a process model that lists tasks, inputs and outputs and specifies how to do the tasks (Pressman, 2005). Tasks are performed using techniques that stipulate how they should be done. After selecting a technique to do the specified tasks, tools can be used to improve task performance.

Finally, the life cycle determines the order in which each activity is to be done (Moore, 1998). A life cycle model is the description of the different ways of developing a project.

From the viewpoint of the above definitions, what do we have in the DM & KD area? Does DM & KD have process models and/or methodologies?

### 2.1 Review of DM & KD process models and methodologies

In the early 1990s, when the KDD process term was first coined (Piatetsky-Shapiro & Frawley, 1991), there was a rush to develop DM algorithms that were capable of solving all problems of searching for knowledge in data. The KDD process (Piatetsky-Shapiro, 1994) (Fayyad et al., 1996) has a process model component because it establishes all the steps to be taken to develop a DM project, but it is not a methodology because its definition does not set out how to do each of the proposed tasks. It is also a life cycle.

The 5 A's (Martínez de Pisón, 2003) is a process model that proposes the tasks that should be performed to develop a DM project and was one of CRISP-DM's forerunners. Therefore, they share the same philosophy: 5 A's proposes the tasks but does not suggest how they should be performed. Its life cycle is similar to the one proposed in CRISP-DM.

A people-focused DM proposal is presented in (Brachman & Anand, 1996): Human-Centered Approach to Data Mining. This proposal describes the processes to be enacted to

carry out a DM project, considering people's involvement in each process and taking into account that the target user is the data engineer.

SEMMA (SAS, 2008) is the methodology that SAS proposed for developing DM products. Although it is a methodology, it is based on the technical part of the project only. Like the above approaches, SEMMA also sets out a waterfall life cycle, as the project is developed right through to the end.

The two models by (Cabena et al., 1998) and (Anand & Buchner, 1998) are based on KDD with few changes and have similar features.

Like the KDD process, Two Crows (Two Crows, 1999) is a process model and waterfall life cycle. At no point does it set out how to do the established DM project development tasks.

CRISP-DM (Chapman et al., 2003) states which tasks have to be carried out to successfully complete a DM project. It is therefore a process model. It is also a waterfall life cycle. CRISP-DM also has a methodological component, as it gives recommendations on how to do some tasks. Even so these recommendations are confined to proposing other tasks and give no guidance about how to do them. Therefore, we class CRISP-DM as a process model.

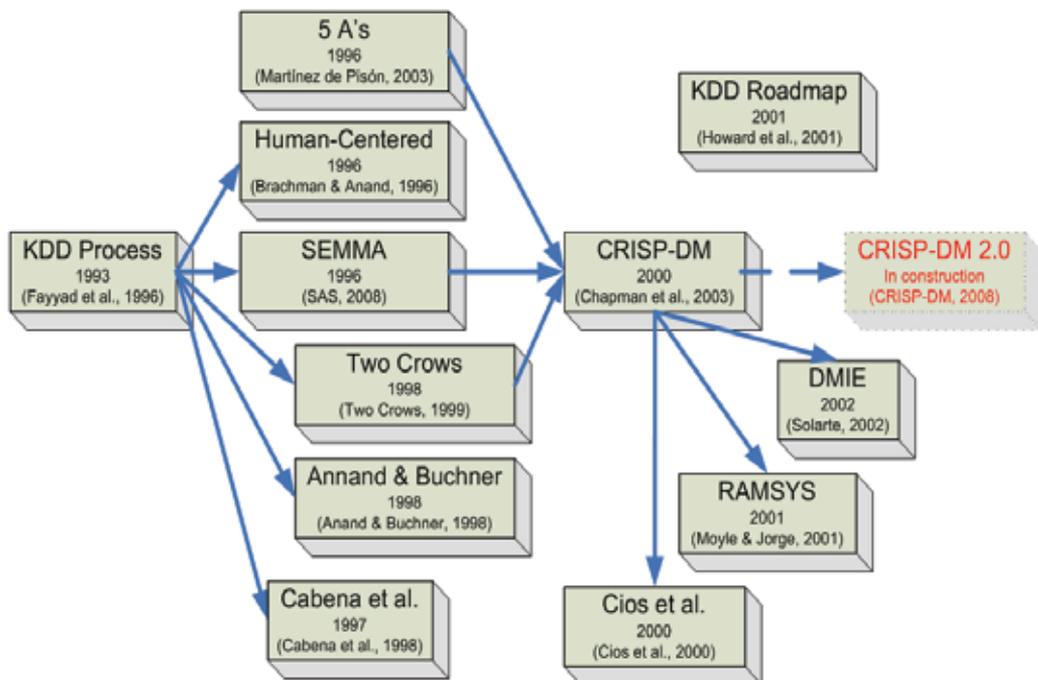


Fig. 1. Evolution of DM & KD process models and methodologies

Figure 1 shows a diagram of how the different DM & KD process models and methodologies have evolved. It is clear from Figure 1 that CRISP-DM is the standard model. It borrowed ideas from the most important pre-2000 models and is the groundwork for many later proposals.

The CRISP-DM 2.0 Special Interest Group (SIG) was set up with the aim of upgrading the CRISP-DM model to a new version better suited to the changes that have taken place in the business arena since the current version was formulated. This group is working on the new methodology (CRISP-DM, 2008). The firms that developed CRISP-DM 1.0 have been joined

by other institutions that intend to input their expertise in the field to develop CRISP-DM 2.0. Changes such as adding new phases, renaming existing phases and/or eliminating the odd phase are being considered for the new version of the methodology.

Cios et al.'s model was first proposed in 2000 (Cios et al., 2000). This model adapted the CRISP-DM model to the needs of the academic research community, providing a more general, research-oriented description of the steps.

The KDD Roadmap (Howard et al., 2001) is a DM methodology used in the DM Witness Miner tool (Lanner Group, 2008). This methodology describes the available processes and algorithms and incorporates experience derived from successfully completed commercial projects. The focus is on the decisions to be made and the options available at each stage to achieve the best results for a given task.

The RAMSYS (RAPid collaborative data Mining SYStem) methodology is described in (Moyle & Jorge, 2001) as a methodology for developing DM & KD projects where several geographically diverse groups work remotely and collaboratively to solve the same problem. This methodology is based on CRISP-DM and maintains the same phases and generic tasks.

Model	Fayyad et al.	Cabena et al.	Anand & Buchner	CRISP-DM	Cios et al.
No of steps	9	5	8	6	6
Steps	Developing and Understanding of the Application Domain	Business Objectives Determination	Human Resource Identification Problem Specification	Business Understanding	Understanding the Data
	Creating a Target Data Set	Data Preparation	Data Prospecting	Data Understanding	Understanding the Data
	Data Cleaning and Pre-processing		Domain Knowledge Elicitation		
	Data Reduction and Projection		Methodology Identification	Data Preparation	Preparation of the data
	Choosing the DM Task		Data Pre-processing		
	Choosing the DM Algorithm				
	DM	DM	Pattern Discovery	Modeling	DM
	Interpreting Mined Patterns	Domain Knowledge Elicitation	Knowledge Post-processing	Evaluation	Evaluation of the Discovered Knowledge
	Consolidating Discovered Knowledge	Assimilation of Knowledge		Deployment	Using the Discovered Knowledge

Table 1. Comparison of DM & KD process models and methodologies (Kurgan & Musilek, 2006)

DMIE or Data Mining for Industrial Engineering (Solarte, 2002) is a methodology because it specifies how to do the tasks to develop a DM project in the field of industrial engineering. It is an instance of CRISP-DM, which makes it a methodology, and it shares CRISP-DM's associated life cycle.

Table 1 compares the phases into which the DM & KD process is decomposed according some of the above proposals. As Table 1 shows, most of the proposals cover all the tasks in CRISP-DM, although they do not all decompose the KD process into the same phases or attach the same importance to the same tasks.

However, some proposals described in this section and omitted the study by (Kurgan & Musilek, 2006), like 5 A's and DMIE, propose additional phases not covered by CRISP-DM that are potentially very useful in KD & DM projects. 5 A's proposes the "Automate" phase. This phase entails more than just using the model. It focuses on generating a tool to help non-experts in the area to perform DM & KD tasks. On the other hand, DMIE proposes the "On-going support" phase. It is very important to take this phase into account, as DM & KD projects require a support and maintenance phase. This maintenance ranges from creating and maintaining backups of the data used in the project to the regular reconstruction of DM models. The reason is that the behavior of the DM models may change as new data emerge, and they may not work properly. Similarly, if other tools have been used to implement the DM models, the created programs may need maintenance, e.g. to upgrade the behavior of the user application models.

## 2.2 CRISP-DM

We focus on CRISP-DM as a process model because it is the "de facto standard" for developing DM & KD projects. In addition, CRISP-DM is the most used methodology for developing DM projects (KdNuggets.com, 2002; KdNuggets.com, 2004; KdNuggets.com, 2007a).

Analyzing the problems of DM & KD projects, a group of prominent enterprises (Teradata, SPSS - ISL, Daimler-Chrysler and OHRA) developing DM projects, proposed a reference guide to develop DM & KD projects. This guide is called CRISP-DM (CRoss Industry Standard Process for Data Mining) (Chapman et al., 2000). CRISP-DM is vendor-independent so it can be used with any DM tool and it can be applied to solve any DM problem.

CRISP-DM defines the phases to be carried out in a DM project. CRISP-DM also defines for each phase the tasks and the deliverables for each task. CRISP-DM is divided into six phases (see Figure 2). The phases are described in the following.

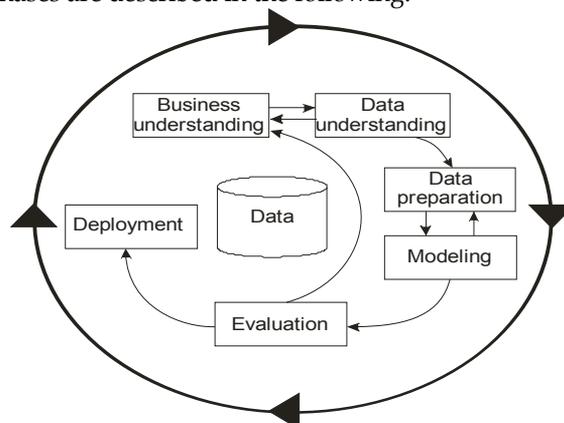


Fig. 2. CRISP-DM process model (Chapman et al., 2000)

Business understanding	Data understanding	Data preparation	Modeling	Evaluation	Deployment
Determine business objectives	Collect initial data	Select data	Select modeling techniques	Evaluate results	Plan deployment
Assess situation	Describe data	Clean data	Generate test design	Review process	Plan monitoring & maintenance
Determine DM objectives	Explore data	Construct data	Build model	Determine next steps	Produce final report
Produce project plan	Verify data quality	Integrate data	Assess model		Review project
		Format data			

Table 2. CRISP-DM phases and tasks.

- **Business understanding:** This phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a DM problem definition and a preliminary plan designed to achieve the objectives.
- **Data understanding:** The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.
- **Data preparation:** The data preparation phase covers all the activities required to construct the final dataset from the initial raw data. Data preparation tasks are likely to be performed repeatedly and not in any prescribed order.
- **Modeling:** In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values. Typically, there are several techniques for the same DM problem type. Some techniques have specific requirements on the form of data. Therefore, it is often necessary to step back to the data preparation phase
- **Evaluation:** What are, from a data analysis perspective, seemingly high quality models will have been built by this stage of the project. Before proceeding to final model deployment, it is important to evaluate the model more thoroughly and review the steps taken to build it to be certain that it properly achieves the business objectives. At the end of this phase, a decision should be reached on how to use of the DM results.
- **Deployment:** Model construction is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it.

Table 2 outlines the phases and generic tasks that CRISP-DM proposes to develop a DM project.

### 3. Software engineering process models

The two most used process models in SE are the IEEE 1074 (IEEE, 1997) and ISO 12207 (ISO, 1995) standards. These models have been successfully deployed to develop software. For this reason, our work is founded on these standards. We intend to exploit the benefits of this experience for application to the field of DM & KD.

Figure 3 compares the two models. As Figure 3 shows, most of the processes proposed in IEEE 1074 are equivalent to ISO 12207 processes and vice versa. To select processes as optimally as possible, the IEEE 1074 and ISO 12207 processes should be mixed. The selection criterion was to choose the processes that either IEEE 1074 or ISO 12207 defined in more detail. We tried to not to mix processes from different groups in either process model. In compliance with the above criteria, we selected IEEE 1074 processes as the groundwork, because they are more detailed. As IEEE 1074 states that it is necessary to acquire or supply software but not how to do this, we added the ISO 12207 (ISO, 1995) acquisition and supply processes.

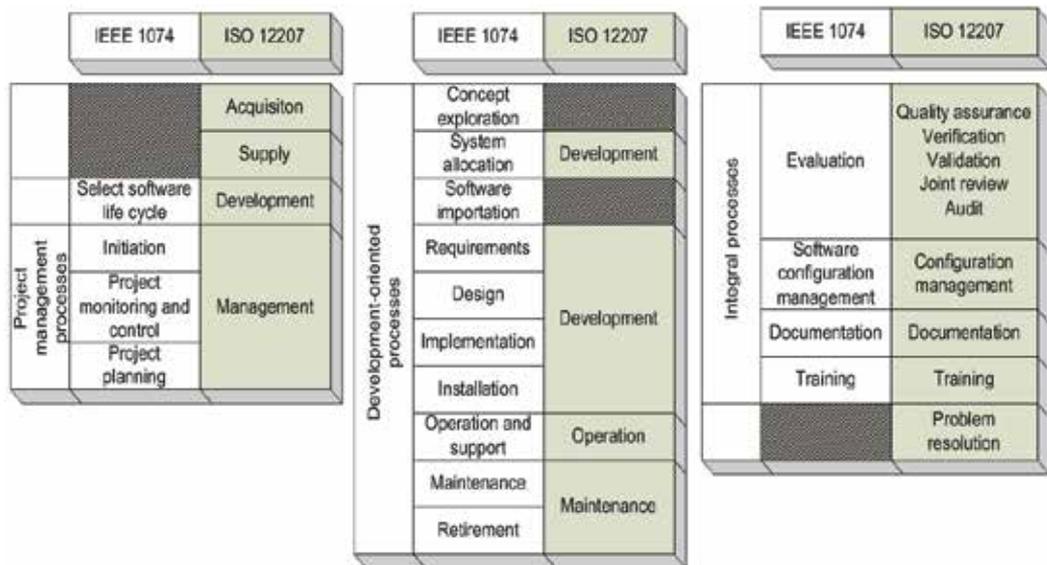


Fig. 3. Mapping ISO 12207 to IEEE 1074 (Marbán et al., 2008)

Figure 4 shows the joint process model developed after examining IEEE 1074 and ISO 12207 according to the above criteria. Figure 4 also shows the activities in each major process group according to the selected standard for the group in question.

We describe the key processes selected for this joint process model below:

- The acquisition and supply processes are taken from the Primary Life Cycle Processes set out in (ISO, 1995). These processes are part of the software development process initiation and determine the procedures and resources required to develop the project.
- The software life cycle selection process (IEEE, 1997) identifies and selects a life cycle for the software under construction.
- The project management processes (IEEE, 1997) are the set of processes that establish the project structure, and coordinate and manage project resources throughout the software life cycle.
- Development-oriented processes (IEEE, 1997) start with the identification of a need for automation. It may take a new application or a change of all or part of an existing application to satisfy this need. With the support of the integral process activities and under the project management plan, the development processes produce software (code and documentation) from the statement of the need. Finally, the activities for installing,

operating, supporting, maintaining and retiring the software product should be performed.

- Integral processes (IEEE, 1997) are necessary to successfully complete the software project activities. They are enacted at the same time as the software development-oriented activities and include activities that are not related to development. They are used to assure the completeness and quality of the project functions.

In the following section we are going to analyze which of the above activities are in CRISP-DM and which are not. The aim is to build a process model for DM projects that is as comprehensive as possible and organizes the activities systematically.

PROCESS	ACTIVITY	PROCESS	ACTIVITY
Acquisition		Design	Perform architectural design
Supply			Design data base
Software life cycle selection	Identify available software life cycles		Design interface
	Select software life cycle		Perform detailed design
Project management activities		Implementation	Create executable code
Initiation	Create software life cycle process		Create operating documentation
	Allocate project resources		Perform integration
	Perform estimations	Post-Development	
	Define metrics	Installation	Distribute software
Project monitoring and control	Manage risks		Install software
	Manage the project		Accept software in operational environment
	Retain records	Operation and support	Operate the system
	Identify software life cycle process improvement needs		Provide technical assistance and consulting
	Collect and analyze metric data		Maintain support request log
Project planning	Plan evaluations	Maintenance	Identify software improvement needs
	Plan configuration management		Implement problem reporting method
	Plan system transition		Maintenance support request log
	Plan installation	Retirement	Notify user
	Plan documentation		Conduct parallel operations
	Plan training		Retire system
	Plan project management	Integral activities	
	Plan integration	Evaluation	Conduct reviews
Deployment oriented activities			Create traceability matrix
Pre-development			Conduct audits
Concept exploration	Identify ideas or needs		Develop test procedures
	Formulate potential approaches		Create test data
	Conduct feasibility studies		Execute test
	Refine and finalize the idea or need		Report evaluation results
System allocation	Analyze functions	Software configuration management	Develop configuration identification
	Decompose system requirements		Perform configuration control
	Develop system architecture		Perform status accounting
Software importation	Identify imported software requirements	Documentation development	Implement documentation
	Evaluate software import sources		Produce and distribute documentation
	Define software import method	Training	Develop training materials
	Import software		Validate the training program
Development			Implement the training program
Requirements	Define and develop software requirements		
	Define interface requirements		
	Prioritize and integrate software requirements		

Fig. 4. Joint process model

#### 4. A data mining engineering process model

A detailed comparison of CRISP-DM with the SE process model described in section 3 is presented in (Marbán et al, 2008). From this comparison, we found that many of the processes defined in SE that are very important for developing any type of DM engineering project are missing from CRISP-DM. This could be the reason why CRISP-DM is not as effective as it should be. What we proposed there was to take CRISP-DM tasks and processes and organize them by processes as SE researchers did. The activities missing from CRISP-DM are primarily project management processes, integral processes (that assure project function completeness and quality) and organizational processes (that help to achieve a more effective organization).

Note that the correspondence between CRISP-DM and SE process model elements is not exact. In some cases, the elements are equivalent, but the techniques are different. In other cases, the elements have the same goal but are implemented completely differently. This obviously depends on the project type. In SE the project aim is to develop software and in DM & KD it is to gather knowledge from data.

Figure 5 shows an overview of the proposed process model, including the key processes. The KDD process is the project development core. In the following we describe the processes shown in Figure 5. We also explain why we think they are necessary in a DM project.

#### 4.1 Organizational processes

This set of processes helps to achieve a more effective organization. They also set the organization's business goals and improve the organization's process, product and resources. Neither the IEEE 1074 nor the ISO 12207 SE process models include these processes. They were introduced in ISO 15504 or SPICE ISO (ISO, 2004). These processes affect the entire organization, not just one project.

This group includes the following processes (see Figure 5):

- **Improvement.** This activity broadcasts the best practices, methods and tools that are available in one part of the organization to the rest of the organization.
- **Infrastructure.** This task builds the best environment in the organization for developing DM projects.
- **Training.** This activity is related to training the staff participating in current or ongoing DM projects.

No DM methodology considers any of these activities. We think that they could be adapted from the SPICE standard because they are all general-purpose tasks common to any kind of project.

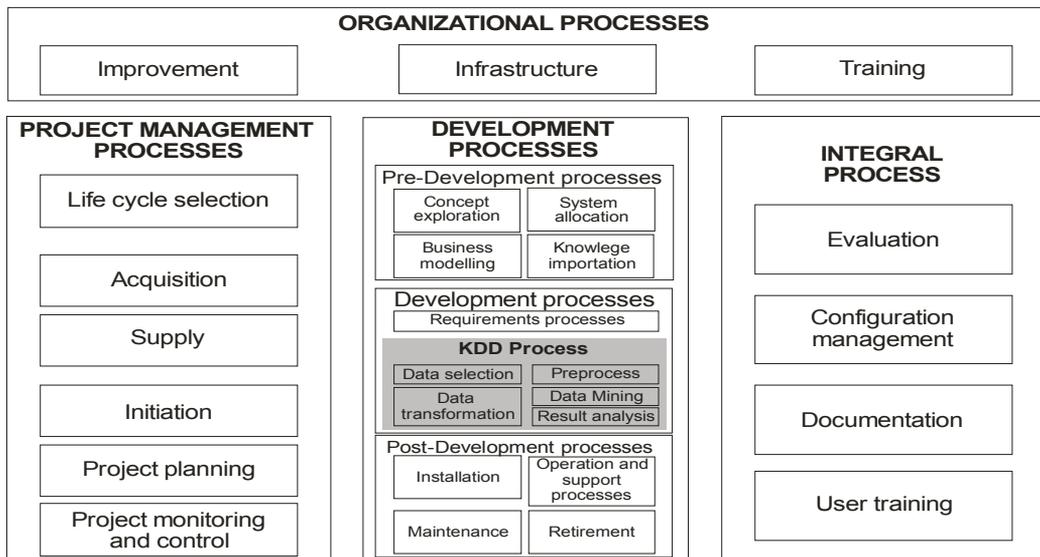


Fig. 5. DM engineering process model (Marban et al., 2008)

#### 4.2 Project management processes

This set of processes establishes the project structure and also how to coordinate and manage project resources throughout the project life cycle. We define six main processes in the project management area. Existing DM methodologies or process models (such as CRISP-DM) take into account only a small part of project management, i.e. the project plan.

The project plan is confined to defining project deadlines and milestones. All projects need other management activities to control time, budget and resources. The project management processes are concerned with controlling these matters.

- **Life cycle selection.** This process defines the life cycle to be used in the DM project (Pressman, 2005). Until now, all DM methodologies had a similar life cycle to CRISP-DM: a waterfall life cycle with backtracking. However, there is a fair chance of new life cycles being developed to meet the needs of the different projects. This is what happens in SE from where they could be adapted.
- **Acquisition.** The acquisition process is related to the activities and tasks of the acquirer (who outsources work). Model building is one possible example of outsourcing. If there is outsourcing, the acquirer must define the acquisition management, starting from the tender and ending with the acceptance of the outsourced product. This process must be included in DM processes because DM projects now developed at non-specialized companies are often outsourced (KdNuggets.Com, 2007d). The acquisition process could be an adaptation of the process proposed in the ISO 12207 standard. It defines software development outsourcing management from requirements to software (this depends on which part is outsourced).
- **Supply.** The supply process concerns the activities and tasks that the supplier has to carry out if the company acts as the developer of an outsourcing project. This process defines the tasks the supplier has to perform to interact with the outsourcing company. It also defines the interaction management tasks. As above, this process can be adapted for DM & KD projects from ISO 12207.
- **Initiation.** The initiation process establishes project structure and how to coordinate and manage project resources throughout the project life cycle. This process could be divided into the following activities: create DM life cycle, allocate project resources, perform estimations, and define metrics. CRISP-DM's Business Understanding phase partly includes these activities, but fails to suggest how to estimate costs and benefits (Fernandez-Baizan et al., 2007). Although some DM metrics have been defined (ROI, accuracy, space/time, usefulness...) (Pai, 2004) (Shearer, 2000) (Biebler et al., 2005) (Smith & Khotanzad, 2007), they are not being used in ongoing DM developments (Marbán et al., 2008).
- **Project planning.** The project planning process covers all the tasks related to planning project management, including the contingencies plan. DM has not shown much interest in this process. DM methodologies focus primarily on technical tasks, and they overlook most of the project management activities. The set of activities considered in this process are: Plan evaluation (Biffel & Winkler, 2007), Plan configuration management (NASA, 2005), Plan system transition (JFMIP, 2001), Plan installation, Plan documentation (Hackos, 1994), Plan training (McNamara, 2007), Plan project management (Westland, 2006), and Plan integration. CRISP-DM does consider the Plan installation and Plan integration activities through its Deployment phase. Although documentation is developed in DM projects, no documentation plan is developed. CRISP-DM states that the user should be trained, but this training is not planned. The other plans are not considered in DM & KD methodologies.
- **Project monitoring and control.** The project monitoring and control process covers all tasks related to project risk and project metric management. The activities it considers within this process are: Manage risks, Manage the Project, Retain records, Identify life

cycle process improvement needs, and Collect and analyze metrics (Fenton & Pfleger, 1998) (Shepperd, 1996). CRISP-DM considers Manage risks in the Business Understanding phase and Identify life cycle process improvement needs in the Development phase. The Manage the project activity is new to CRISP-DM, which considers project planning but not plan control. The other two activities (Retain records and Collect and analyze metrics) are new to DM & KD methodologies.

#### 4.3 Development processes

These are the most highly developed processes in DM. All DM methodologies focus on these processes. This is due to the fact that development processes are more related to technical matters. Consequently, they were developed at the same time as the techniques were created and started to be applied. These processes are divided into three groups: pre-development, development, and post-development processes.

The development process is the original KDD process defined in (Piatetsky-Shapiro, 1994). The pre- and post-development processes are the ones that require a greater effort.

##### 4.3.1 Pre-development processes

These processes are related to everything to be done before the project kicks off. From the process model in Figure 5, we can identify the following Pre-development processes:

- **Concept exploration.** In CRISP-DM these activities are considered in the Business Understanding phase. However, these tasks do not completely cover the activities because they focus primarily on the terminology and background of the problem to be solved. In (Marbán et al., 2008) we conclude that some tasks adapted from SE standards need to be added to optimize project development.
- **System allocation.** CRISP-DM considers these tasks through its Business Understanding phase
- **Business modeling.** This is a completely new activity. The CRISP-DM business model is described in the Business Understanding phase, but there are no business modeling procedures or formal tools and methods as there are in SE (Gordijn et al., 2000).
- **Knowledge importation.** This process is related to the reuse of existing knowledge or DM models from other or previous projects, something which is very common in DM. CRISP-DM does not consider this process at all, and its SE counterpart is related to software and cannot be easily adapted. Consequently, the process must be created from scratch.

##### 4.3.2 Development processes

This is the most developed phase in DM methodologies, because it has been researched since late 1980s. CRISP-DM phases include all these processes in one way or another. In SE the process is divided into requirements, analysis, design, and implementation phases. We can easily map the requirements, design and implementation phases to DM projects. The design and implementation phases match the KDD process, and we will stick with this process and its name.

- **Requirements processes.** CRISP-DM covers this set of processes, but they are incomplete. The requirements are developed in CRISP-DM's Business Understanding phase. In CRISP-DM this process produces a list of requirements, but the CRISP-DM user guide does not specify or describe any procedure or any formal notation, tool or

technique to obtain the requirements from the business models. Neither does it specify or describe how to translate requirements into DM goals and models for proper use in the subsequent design and implementation phases: the KDD process. We believe that requirements can be described formally like they are in SE (Kotonya, G. & Sommerville, 1998). For example, something like use-case models could be adapted to specify the project requirements. Further work and research, possibly inspired by SE best practices, should be put into developing a core of formal methods and tools adapted to this process in the DM area.

- **KDD process.** The KDD process matches the design and implementation phases of a software development project. This set of processes is responsible for acquiring the knowledge for the DM project. KDD includes the following activities: data selection, pre-processing, data transformation, DM, and result analysis. CRISP-DM covers the KDD process (Marbán et al., 2008).

#### 4.3.3 Post-development processes

Post-development processes are the processes that are carried out after the knowledge is gathered. They are applicable during the later life cycle stages.

- **Installation.** This process is commended with transferring the knowledge extracted from the DM results to the users. The knowledge can be used as it is, i.e. to help managers to make decisions about a future marketing campaign, or could involve some software development, i.e. to improve an existing web-based recommender system. CRISP-DM considers planning for deploying the knowledge at the client site, but it does not regard the development of software installed and accepted in an operational environment as part of this deployment (Reifer, 2006).
- **Operation and support process.** This process is necessary to validate the results and how they are interpreted by the client, and, if software is developed, to provide the client with technical assistance. CRISP-DM only includes results monitoring. In addition, we propose tasks to validate the results (this is a new task) and to provide technical assistance if necessary (this task could be directly incorporated from IEEE 1074).
- **Maintenance.** The maintenance process has two different paths. On the one hand, if knowledge is embedded in software, this process will provide feedback information to the software life cycle and lead to changes in the software. For this path, the task can be adapted from the IEEE 1074 maintenance process. On the other hand, CRISP-DM does not include a task for knowledge used as it is, and this needs to be developed from scratch.
- **Retirement.** The knowledge gathered from data is not valid forever, and this task is in charge of retiring obsolete knowledge from the system. CRISP-DM does not cover this process, but it can be adapted from IEEE 1074.

#### 4.4 Integral processes

Integral processes are necessary to successfully complete the project activities. These processes assure project function completeness and quality. They are carried out together with development processes to assure the quality of development deliverables. The integral processes group the four processes described below.

- **Evaluation.** This process is used to discover defects in the product or in the process used to develop the DM project. CRISP-DM covers the evaluation process through

evaluation activities spread across different phases: Evaluation, Deployment, Business Understanding and Modeling. But we think the organization of the SE process is more appropriate and covers more aspects.

- **Configuration management.** This process is designed to control system changes and maintain system coherence and traceability. The ultimate aim is to be able to audit the evolution of configurations (Buckley, 1992). We consider this to be a key process in a DM project because of the amount of information and models generated throughout the project. Surprisingly, DM methodologies do not account for this process at all.
- **Documentation.** This process is related to designing, implementing, editing, producing, distributing and maintaining the project documentation. CRISP-DM considers this process across different phases: Deployment, Deployment, Modeling, and Evaluation.
- **User training.** Current DM methodologies do not consider user training at all. This process is related to training inexperienced users to use and interpret the results of the DM project.

## 5. Conclusions and future development

After analyzing the SE process models, we have developed a joint model based on two standards to compare, process by process and activity by activity, the modus operandi in SE and DM & KD. This comparison revealed that CRISP-DM does not cover many project management-, organization- and quality-related tasks at all or at least thoroughly enough. This is now a must due to the complexity of the projects being developed in DM & KD these days. These projects not only involve examining huge volumes of data but also managing and organizing big interdisciplinary human teams.

Consequently, we proposed a DM engineering process model that covers the above points. To do this, we made a distinction between process model, and methodology and life cycle. The proposed process model includes all the activities covered by CRISP-DM, but distributed across process groups that conform to engineering standards established by a field with over 40 years' experience, i.e. software engineering.

The model is not complete, as the need for the processes, tasks and/or activities set out in IEEE 1074 or ISO 12207 and not covered by CRISP-DM has been stated but they have yet to be adapted and specified in detail.

Additionally, this general outline needs to be further researched. First, the elements that CRISP-DM has been found not to cover at all or only in part would have to be specified and adapted from their SE counterpart. Second, the possible life cycle for DM would have to be examined and specified. Third, the process model specifies that what to do but not how to do it. A methodology is what specifies the "how to" part. Therefore, the different methodologies that are being used for each process would need to be examined and adapted to the model. Finally, a methodology is associated with a series of tools and techniques. DM has already developed many such tools (like Clementine or the neural network techniques), but tools that are well-established in SE (e.g. configuration management techniques) are missing. It remains to be seen how they can be adapted to DM and KD processes.

## 6. References

Anand, S.; Patrick, A.; Hughes, J. & Bell, D. (1998). A data mining methodology for cross-sales. *Knowledge Based Systems Journal*. 10, 449-461.

- Biebler, K.; Wodny, M., & Jager, B. (2005). Data mining and metrics on data sets. In CIMCA '05: *Proceedings of the International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce*. Vol-1 (CIMCA-IAWTIC'06), pages 638–641, Washington, DC, USA. IEEE Computer Society.
- Biffl, S. & Winkler, D. (2007). Value-based empirical research plan evaluation. In *First International Symposium on Empirical Software Engineering and Measurement*, 2007, pages 494–494. IEEE.
- Brachman, R. J. & Anand, T. (1996) *The process of knowledge discovery in databases*. pp 37–57.
- Buckley, F. (1992). Configuration Management: Hardware, Software and Firmware. *IEEE Computer Society Press*, USA.
- Cabena, P.; Hadjinian, P.; Stadler, R.; Verhees, J. & Zanasi, A. (1998) *Discovering Data Mining: From Concepts to Implementation*. Prentice Hall.
- Chapman, P.; Clinton, J.; Kerber, R.; Khabaza, T.; Reinartz, T.; Shearer, C. & Wirth, R. (2000). CRISPDm 1.0 step-by-step data mining guide. *Technical report*, CRISP-DM
- CRISP-DM (2008). Crisp-2.0: Updating the methodology, [www.crisp-dm.org/new.htm](http://www.crisp-dm.org/new.htm)
- Edelstein, H.A. & Edelstein, H.C. (1997). Building, Using, and Managing the Data Warehouse. In: *Data Warehousing Institute*, 1st edn., Prentice Hall PTR, Englewood Cliffs
- Eisenfeld, B.; Kolsky, E. & Topolinski, T. (2003a). *42 percent of CRM software goes unused*, <http://www.gartner.com>
- Eisenfeld, B.; Kolsky, E.; Topolinski, T.; Hagemeyer, D. & Grigg, J. (2003b). *Unused CRM software increases TCO and decreases ROI*, <http://www.gartner.com>
- Fayyad, U.; Piatetsky-Shapiro, G.; Smith, P. & Uthurusamy, R. (1996) *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press, MA
- Fenton, N. E. & Pfleeger, S. L. (1998). *Software Metrics: A Rigorous and Practical Approach*. *Course Technology*, 2nd edition.
- Fernandez-Baizan, C.; Marban, O. & Menasalvas, E. (2007). A cost model to estimate the effort of data mining projects (DMCoMo). *Information Systems*.
- Gartner (2005). Gartner Says More Than 50 Percent of Data Warehouse Projects Will Have Limited Acceptance or Will Be Failures Through 2007. *Analysts to Show How To Implement a Successful Business Intelligence Program During the Gartner Business Intelligence Summit, March 7-9 in Chicago, IL*. 2005 Gartner Press Releases.
- Gondar, J.E. (2005). *Metodología Del Data Mining*. Data Mining Institute, S.L
- Gordijn, J.; Akkermans, H. & van Vliet, H. (2000). Business modelling is not process modelling. In *ER Workshops*, pages 40–51.
- Hackos, J. T. (1994). *Managing your documentation projects*. John Wiley & Sons, Inc., New York, NY, USA.
- Howard, C.M.; Debusse, J. C.W.; de la Iglesia, B. & Rayward-Smith, V.J. (2001). Building the KDD Roadmap: A Methodology for Knowledge Discovery, chapter of *Industrial Knowledge Management*. pages 179–196. Springer-Verlag.
- IEEE (1997) Std. for Developing Software Life Cycle Processes. IEEE Std. 1074-1997. *IEEE*, Nueva York (EE.UU.)
- ISO (1995). ISO/IEC Std. 12207:1995. Software Life Cycle Processes. *International Organization for Standardization*, Geneva (Switzerland).

- ISO (2004). ISO/IEC Standard 15504:2004. Software Process Improvement and Capability determination (SPICE). *International Organization for Standardization*, Geneva (Switzerland).
- Jaffarian, T.; Mok, L.; McDonald, M. P.; Bloesch & M. and Stevens, S. (2006). *Growing it's contribution: The 2006 CIO agenda*. [www.gartner.com](http://www.gartner.com)
- KdNuggets.Com (2002). *Data Mining Methodology*. <http://www.kdnuggets.com/polls/2002/methodology.htm>
- KdNuggets.Com (2004). *Data Mining Methodology*. [http://www.kdnuggets.com/polls/2004/data\\_mining\\_methodology.htm](http://www.kdnuggets.com/polls/2004/data_mining_methodology.htm)
- KdNuggets.Com (2007a). *Data Mining Methodology*. [http://www.kdnuggets.com/polls/2007/data\\_mining\\_methodology.htm](http://www.kdnuggets.com/polls/2007/data_mining_methodology.htm), 2007.
- KdNuggets.Com (2007b). *Is data mining a mature field?* [http://www.kdnuggets.com/polls/2007/data\\_mining\\_mature\\_field.htm](http://www.kdnuggets.com/polls/2007/data_mining_mature_field.htm)
- KdNuggets.Com (2007c). *Data Mining Activity in 2007 vs 2006*. [http://www.kdnuggets.com/polls/2007/data\\_mining\\_2007\\_vs\\_2006.htm](http://www.kdnuggets.com/polls/2007/data_mining_2007_vs_2006.htm).
- KdNuggets.Com (2007d). *Outsourcing data mining*. [http://www.kdnuggets.com/polls/2007/outsourcing\\_data\\_mining.htm](http://www.kdnuggets.com/polls/2007/outsourcing_data_mining.htm)
- KdNuggets.Com (2008). *Data Mining ROI* <http://www.kdnuggets.com/polls/2008/roi-data-mining.htm>
- JFMIP (2001). White paper: *Parallel operation of software - is it a desirable software transition technique?* Joint Financial Management Improvement Program.
- Kotonya, G. & Sommerville, I. (1998). *Requirements Engineering. Processes and Techniques*. Wiley, USA.
- Lanner Group (2008). *Witness Miner On-line Help System*. <http://www.witnessminer.com>
- Marbán, O.; Segovia, J.; Menasalvas, E. & Fernandez-Baizan, C. (2008). Towards Data Mining Engineering: a software engineering approach. *Information Systems Journal*. doi: 10.1016/j.is.2008.04.003
- Martínez de Pisón Ascacibar, F.J. (2003). *Optimización Mediante Técnicas de Minería de Datos Del Ciclo de Recocido de Una Línea de Galvanizado*. PhD thesis, Universidad de La Rioja, 2003.
- McMurchy, N. (2008). *Toolkit Tactical Guideline: Five Success Factors for Effective BI Initiatives*. Gartner.com
- McNamara, C. (2007). *Complete guidelines to design your training plan*. <http://www.managementhelp.org/trng dev/gen plan.htm>.
- Moore, J. (1998). *Software Engineering Standards: A User's Road Map*. IEEE Computer Science Press, Los Alamitos, California
- NASA (2005). *Software Engineering Process Guidebook, Software Configuration Management Planning*. Software Engineering NASA LaRC and Analysis Lab.
- Naur, P. & Randell, B. (1969). *Software Engineering: Report on a conference sponsored by the NATO science committee*.
- Pai; W. C. (2004). Hierarchical analysis for discovering knowledge in large databases. *Information Systems Management*, 21:81-88.
- Piatetsky-Shapiro, G. & Frawley, W. (1991) *Knowledge Discovery in Databases*. AAAI/ MIT Press, MA.
- Piatetsky-Shapiro, G. (1994). *An overview of knowledge discovery in databases: Recent progress and challenges*. *Rough Sets, Fuzzy Sets and Knowledge Discovery*, pages 1-11.

- Pressman, R. (2005). *Software Engineering: A Practitioner's Approach*. McGraw-Hill, New York.
- Reifer, D. J. (2006). *Software Management*. Wiley-IEEE Computer Society Press, 7th. edition.
- SAS Institute (2008). *SEMMA data mining methodology*, <http://www.sas.com>
- Shepperd, M. (1996). *Foundations of Software Measurement*. Prentice Hall.
- Smith, M. & Khotanzad, A. (2007). Quality metrics for object-based data mining applications. In ITNG '07: *Proceedings of the International Conference on Information Technology*, pages 388–392, Washington, DC, USA, 2007. IEEE Computer Society.
- Solarte, J. (2002). *A proposed data mining methodology and its application to industrial engineering*. Master's thesis, University of Tennessee, Knoxville
- Strand, M. (2000). *The Business Value of Data Warehouses - Opportunities, Pitfalls and Future Directions*. PhD thesis, University of Skövde
- Two Crows Corp (1999). *Introduction to Data Mining and Knowledge Discovery*. 3rd edn.
- Westland, J. (2006). *The Project Management Life Cycle: A Complete Step-by-Step Methodology for Initiating, Planning, Executing and Closing the Project Successfully*. Kogan.
- Yang, Q. & Wu, X (2006). 10 Challenging Problems in Data Mining Research. *International Journal of Information Technology & Decision Making*. Vol. 5, No. 4 (2006) 597–604. World Scientific Publishing Company.
- Zornes, A. (2003) *The top 5 global 3000 data mining trends for 2003/04*. META Group Research-Delta Summary.

# Knowledge Discovery on the Grid

Lamine Aouad, An Le-Khac and Tahar Kechadi

*University College Dublin  
School of Computer Science and Informatics  
Belfield, Dublin 4  
Ireland*

## 1. Introduction

In the last few decades, Grid technologies have emerged as an important area in parallel and distributed computing. The Grid can be seen as a computational and large-scale support, and even in some cases as a high-performance support. In recent years, the data mining community have been increasingly using Grid facilities to store, share, manage and mine large-scale data-driven applications. Indeed, data mining and knowledge discovery applications are by nature distributed, and are using the Grid as their execution environment. This particularly led to a great interest of the community in distributed data mining and knowledge discovery on large Grid platforms. Many Grid-based Data Mining (DM) and Knowledge Discovery (KD) frameworks were initiated, and proposed different techniques and solutions for large-scale datasets mining. These include the ADMIRE project initiated by the PCRG (Parallel Computational Research Group) at the University College Dublin, the Knowledge Grid project at the University of Calabria, The GridMiner project at the University of Vienna, among others.

These knowledge discovery<sup>1</sup> frameworks on the Grid aim to offer high-level abstractions and techniques for distributed management, mining, and knowledge extraction from data repositories and warehouses. Most of them use existing Grid technologies and systems to build specific knowledge discovery services, data management, analysis, and mining techniques. Basically, this consists of either porting existing algorithms and applications on the Grid, or developing new mining and knowledge extraction techniques, by exploiting the Grid features and services. Grid infrastructures usually provide basic services of communication, authentication, storage and computing resources, data placement and management, etc. For example, the Knowledge Grid system uses services provided by the Globus Toolkit, and the ADMIRE framework uses a Grid system called DGET, developed by our team at the University College Dublin. We will give some details about the best-known DM/KD frameworks in section 2. Note that this chapter is not intended to Grid systems or the way they are interfaced with knowledge discovery frameworks. Indeed, beyond the architecture design of Grid systems, the resources and data management policies, the data integration or placement techniques, and so on, these DM and KD frameworks need

---

<sup>1</sup> *Knowledge Discovery* is a more general term that includes the *Data Mining* process. It refers to the overall knowledge extraction process.

efficient algorithmic approaches and implementations to optimise their performance and to address the large-scale and heterogeneity issues. This chapter focuses on this aspect of data mining on the Grid. In other words, we will not discuss the actual mapping of the knowledge discovery processes onto Grid jobs or services.

Unlike centralised knowledge extraction and data mining techniques, which are quite well established, there are many challenges and issues that need to be addressed in Grid-based DM, in order to fully benefit from the massive computing power and storage capacity of the Grid. These include issues related to the global model generation using local models from different sites, the heterogeneity of local datasets which might record different feature vectors, the global validity assessment, the scaling behaviour of the distributed techniques and the knowledge representation, etc. In addition, there are many Grid related issues such as security or overheads. There exists a limited amount of literature in distributed DM on Grids. This will be briefly reviewed in the following.

For the sake of clarity, we give a brief definition of what are knowledge discovery, and the data mining process. Knowledge discovery applies a set of appropriate algorithms and mechanisms to extract and present the knowledge from a given dataset, i.e. identifying valid, novel, potentially useful, and ultimately understandable patterns in the dataset. This process generally involves three main steps:

- Data cleaning, pre-processing and transformation. This basically prepares the dataset for the data mining process. Many algorithms have input-related constraints, and some conversions are required. These include noise removal, missing data treatment, data sampling, etc.
- Data Mining. This step describes the application and parameters of a specific algorithm to search for useful patterns within the dataset. There are three basic datasets analysis tasks in data mining, namely classification, association, and cluster analysis.
- Knowledge extraction and interpretation. This presents the results, i.e. a set of patterns, in a human-readable manner on the basis of the end-user interest. Various presentation forms exist in both centralised and distributed systems. In ADMIRE, we have developed an innovative knowledge map representation well adapted for large Grids. This will be discussed in section 3.

The rest of this chapter is organised as follow. The next section presents related work in Grid-based DM. In section 3, we describe the ADMIRE project, some of its well-adapted algorithmic techniques for the Grid, and the innovative knowledge map layer for high-level knowledge representation. The next section discusses some of the fundamental issues of both the knowledge discovery field and Grid computing. Finally, concluding remarks are made in section 5.

## **2. Related work**

This section gives a brief review of the best-known existing projects in distributed data mining and knowledge discovery on the Grid. These emerging frameworks can be roughly classified either as domain-specific or domain-independent. Most of the KD frameworks on the Grid attempt to build domain-independent systems allowing the user to express specific problems.

### **2.1 TeraGrid**

TeraGrid (Berman F., 2001) is a discovery infrastructure combining resources at eleven partner sites to create an integrated and persistent computational and data management

resource. Understanding and making scientific contributions of terabytes and petabytes of distributed data collections via simulation, modeling, and analysis are some of the challenges addressed by the TeraGrid project. TeraGrid tackles a range of domains and data collections including biomedical images, repositories of proteins, stream gauge measurements of waterways, digital maps of the world and the universe, etc. Then, the synthesis of knowledge, through mining for instance, from these massive amounts of data is among the most challenging applications on the TeraGrid. This allows the TeraGrid to achieve its potential and enable the full use of the TeraGrid infrastructure as a knowledge Grid system.

## 2.2 Knowledge grid

Knowledge Grid is a distributed framework that integrates data mining techniques. In the Knowledge Grid architecture, data mining tools are integrated within Grid mechanisms and services provided by the Globus Toolkit. It aims to deal with large datasets available on the Grid for scientific or industrial applications. The Knowledge Grid project was initiated by Cannataro et al. at the University of Calabria in Italy. The architecture of Knowledge Grid is designed in such a way that specialised data mining and knowledge discovery techniques fit with lower-level mechanisms and services provided by GTK. It is composed of two layers: the *Core K-Grid* and the *High-level K-Grid*. Basically, the *Core K-Grid* layer implements basic KD services on top of generic Grid services, while the *High-level K-Grid* layer is intended to design, compose, and execute distributed KD over it.

The lower layer of knowledge Grid comprises two main services: The *Knowledge Directory service* (KDS), and the *Resource allocation and execution management service* (RAEMS). KDS manages metadata including data sources and repositories, data mining tools and algorithms, a distributed execution plans which are basically a graph describing the interactions between processes and the dataflow, and finally the results of the computation, i.e. models or patterns. There are different metadata repositories associated with these services. RAEMS is used to map the application, i.e. the graph, into available resources. This component is directly based on the GRAM services of Globus, and execution plans generate requests which are translated into the Globus RSL language.

The higher layer of KG includes services for composing, validating, and executing distributed KD computations, as well as storing, analysing, and presenting services. This is offered by four main services: the *Data Access Service* (DAS), the *Tools and Algorithms Access Service* (TASS), the *Execution Plan Management Service* (EPMS), and the *Results Presentation Service* (RPS). The first is basically used for the search, selection, extraction, transformation, and access of the datasets. The second is responsible of searching, selecting, and downloading data mining tools and algorithms. The third service generates an abstract execution plan describing the computation and the mapping onto Grid resources. The last service generates, presents and visualises the discovered models and patterns. For more details about the architecture of KG, and other aspects such as the design of applications within KG, we refer the reader to (Cannataro M. et al. 2004).

This framework is more focused in providing a distributed DM architecture that can benefit from 'standard' Grid services provided by Globus. The algorithmic aspect, i.e. well-adapted approaches for the Grid, is not taken into account. In addition, Knowledge Grid does not provide global management and coordination of the overall knowledge on the Grid. These aspects are taken into account in the ADMIRE framework which makes the knowledge discovery on distributed Grids more flexible and efficient.

### 2.3 GridMiner

GridMiner is a Grid and Web based architecture for distributed KD, based on the OGSA architecture. Each service in GridMiner is implemented as a Grid service specified by Open Grid Services Architecture. The data mining process within GridMiner is supported by several Grid services that are able to perform data mining tasks and OLAP (Brezany P. et al., 2005). It also provides a workflow engine (*Dynamic Service Control Engine*) which controls the Grid services composition provided as a *DSCL (Dynamic Service Control Language)* document by the *DSCE client*. GridMiner uses OGSA-DAI (*Data Access and Integration*) as a standard middleware implementation of its GDS (*Grid Data Service*) for supporting access and integration of data within the Grid. GridMiner has a GUI that offers a friendly front-end for the end-user and the system administrator.

GridMiner is quite similar to Knowledge Grid; the main difference is that the KG framework is based on a non-OGSA version of the Globus Toolkit (Version 2). This project is also an architecture-oriented effort and does not address the algorithmic aspect on the Grid as well as the high-level knowledge representation.

### 2.4 Discovery net

Discovery Net is a service-oriented computing model for knowledge discovery which allows the end-user to connect to and use data mining and analysis tools as well as data sources that are available on-line. The overall architecture of Discovery Net is composed of three main servers: the *Knowledge Servers* allow the user to store/retrieve or publish knowledge, the *Resources Discovery Servers* publish service definitions/locations, and the *Discovery Meta-information Server* stores information about each type of knowledge. Discovery Net also provides a composition language called *DPML (Discovery Process Markup Language)* representing the graph of services. Details about architectural aspects of Discovery Net can be found in (Curcin V. et al., 2001). Note that this framework focuses on remote services composition and has a centralised knowledge representation for each of the composed graph services. This approach is not feasible for large-scale and complex heterogeneous scenarios.

## 3. ADMIRE: a grid-based data mining and knowledge discovery framework

In this section, we present the architecture of the ADMIRE framework. The overall organisation of ADMIRE is presented in Fig. 1. We will be focusing on two fundamental parts of ADMIRE, namely the Grid-based algorithms and the knowledge map layer. We will present two lightweight algorithms: for clustering analysis and mining association rules on the Grid, as well as the concept of the knowledge map and its structure.

ADMIRE is organised on a layered architecture built on top of the DGET Grid middleware developed at the University College Dublin (Hudzia B. et al., 2005a), (Hudzia B. et al., 2005b). Details about the use of the specific Grid services provided by DGET within ADMIRE are not discussed in this chapter. Indeed, following the Grid architecture approach, the knowledge discovery services can be developed and implemented in different ways using the available Grid services. This is of little use for the understanding of the algorithmic approaches and the knowledge management within ADMIRE presented below. There are two main hierarchical levels in ADMIRE: the data mining level, and the knowledge map level. Modules and services within these levels will be described in the following.

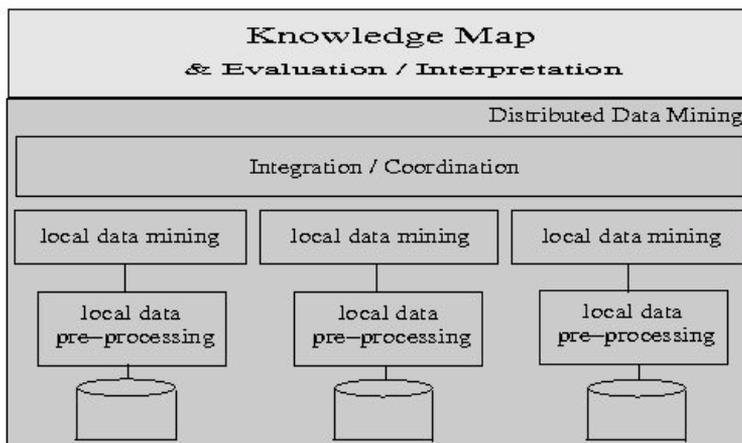


Fig. 1. The ADMIRE layered organisation.

### 3.1 Grid-based algorithms for large-scale datasets mining

Many parallel, distributed, and Grid-based algorithms have been already proposed in the literature for a large range of data mining applications. Most of them are based on the aggregation of local models according to some collected local information or statistics. In Grid environments, the mining algorithms have to deal with distributed datasets, different administration domains, and probably plural ownership and users. Thus, moving these datasets to a single location for performing a global mining is not always possible due to different reasons related to policies or technical choices. The Grid also faces a scalability issue of its DM applications and their implementations. We believe that the communication efficiency of an algorithm is often more important than the accuracy of its results. Indeed, communication issues are the key factors in the implementation of any distributed and grid-based algorithm. A suitable algorithm for high-speed networks is more likely to be of little use in WAN-based and Grid platforms. Efficient Grid-based algorithms need to exchange a few data and avoid synchronisations as much as possible. Following this reasoning, we proposed some well-adapted algorithms for the Grid. The rest of this section describes two of them, for clustering and frequent itemsets generation.

#### 3.1.1 Variance-based distributed clustering

Clustering is one of the basic tasks in data mining applications. Basically, clustering groups data objects based on information found in the data that describes the objects and their relationships. The goal is to optimise similarity within a cluster and the dissimilarities between clusters in order to identify interesting patterns. This is not a straightforward task in unsupervised knowledge discovery. There exists a large amount of literature on this task ranging from models, algorithms, validity and performances studies, etc.

Clustering algorithms can be divided into two main categories, namely partitioning and hierarchical. Different elaborated taxonomies of existing clustering algorithms are given in the literature. Details about these algorithms is out of the purpose of this chapter, we refer the reader to (Jain A. K. et al., 1999) and (Xu R. & Wunsch D., 2005). Many parallel clustering versions based on sequential and centralised algorithms, such as the widely used k-means

algorithm or its variants have been proposed. These include (Dhillon I. S. and Modha D., 1999), (Ester M. et al., 1996), (Garg A. et al., 2006), (Geng H. et al., 2005), (Joshi M. N., 2003), (Xu X. et al., 1999), among others. Most of these algorithms are message-passing versions of their corresponding centralised algorithms and need either multiple synchronisation constraints between processes or a global view of the dataset or both.

Many of the proposed distributed approaches are based on algorithms that were developed for parallel systems. Indeed, most of them typically produce local models followed by the generation of a global model by aggregating the local results. The distributed processes, participating to the computation, have to be quite independent. After local phases, the global model is then obtained based on only local models, without a global view of the whole dataset. These algorithms usually perform global reduction of so-called *sufficient statistics*, probably followed by a broadcast of the results. Some research works are presented in (Januzaj E. et al., 2003), (Zhang B. and Forman G., 2000), (Januzaj E. et al., 2004a), (Januzaj E. et al., 2004b), or (Jin R. et al. 2006). These are mostly related to the k-means algorithm or its variants and the DBSCAN density-based algorithm.

There are still several open questions in the clustering process. These include:

- What is the optimal number of clusters?
- How to assess the validity of a given clustering?
- How to allow different shapes and sizes rather than forcing them into balls and shapes related to the distance functions?
- How to prevent the algorithms initialization and the order in which the features vectors are read in from affecting the clustering output?
- How to find which clustering structure for a given dataset, i.e. why would a user choose an algorithm instead of another?

These questions, among others, come from the fact that there is no general definition of what is a cluster. Indeed, algorithms have been developed to find several kinds of clusters; spherical, linear, dense, or drawnout.

In the process of addressing some of these fundamental issues, we proposed a lightweight Grid-based clustering technique, based on a merging of independent local sub-clusters according to an increasing variance constraint. This was shown to improve the overall clustering quality and finds the number of clusters and the global inherent clustering structure of the global dataset with a very low communication overhead. The algorithm finds a proper variance criterion for each dataset based on a statistical global assessment that does not violate the locality principle of this algorithm and for each dataset. This parameter can also be available from the problem domain for a given data. In the rest of this section, we will give the algorithm foundations and the complexity and performance analysis.

### **The algorithm foundations**

The most used criterion to quantify the homogeneity inside a cluster is the variance criterion, or sum-of-squared-error. The traditional constraint used to minimize this criterion is to fix the number of clusters to an a priori known number, as in the widely used k-means and its variants (Xu R. & Wunsc D., 2005), (Ng R. T. & Han J., 1994), (Zhang B. et al., 1999), etc. This constraint is very restrictive since this number is most likely not known in most cases. Many approximation techniques exist including the gap statistic which compares the change within cluster dispersion to that expected under an appropriate reference null distribution (R. Tibshirani et al., 2000) and (Mingjin Y. & Keying Y., 2007), or the index due

to Calinski & Harabasz (Calinski R. B. & Harabasz J., 1974), among other techniques. The imposed constraint in our method states that the increasing variance of the merging, or union of two sub-clusters is below a given dynamic threshold. This parameter is highly dependent on the dataset and is computed using a global assessment method.

The key idea of the algorithm is to start with a relatively high number of clusters in local sites which are referred to as sub-clusters. An optimal local number using an approximation technique or a method that finds the number of clusters automatically, such as those described earlier, can be considered. Then, the global merging is done according to an increasing variance criterion requiring a very low communication overhead. Recall that this algorithm finds a proper variance criterion for each dataset based on a statistical global assessment. This allows us to comply with the locality criterion for different datasets.

In local sites, the clustering can be done using different algorithms depending on the characteristics of the dataset. This may include k-means, k-harmonic-means, k-medoids, or their variants, or the statistical interpretation using the expectation-maximization algorithm, etc. The merging of local sub-clusters exploits the locality in the feature space, i.e. the most promising candidates to form a global cluster are sub-clusters that are the closest in the features space, including sub-clusters from the same site. Each processing node can perform the merging and deduce the global clusters formation, i.e. which sub-clusters are subject to form together a global cluster. Fig. 2. shows how sub-clusters from different sites (Gaussian distributions) are merged together to form a global cluster.

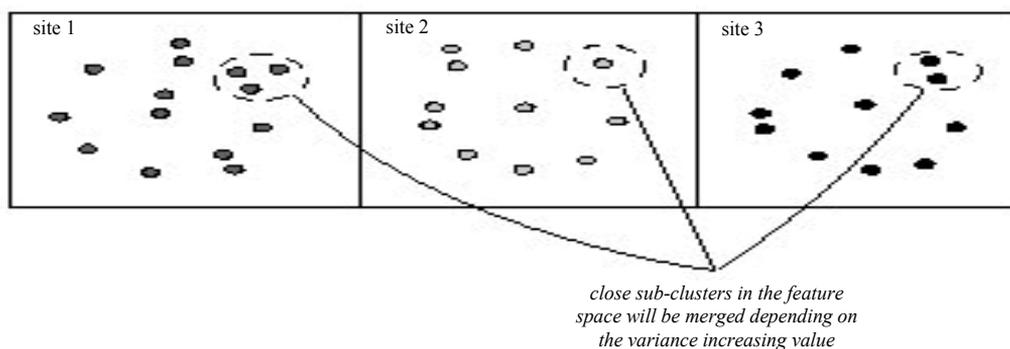


Fig. 2. Close sub-clusters in the features space are merged on a global cluster.

One important notion used in this algorithm is the global cluster border which represents local sub-clusters at its border. These sub-clusters are candidate to be moved to neighbouring global clusters in order to contribute to an improvement of the clustering output, with respect to the variance criterion, i.e. that minimises the sum-of-squared-error. These sub-clusters are referred to as perturbation candidates. The initial merging order may affect the clustering output, as well as the presence of non well-separated global clusters. This is intended to reduce the input order impact. The global clusters are then updated. The border is collected by computing the common Euclidean distance measure. The  $b$  farthest sub-clusters are then selected to be the perturbation candidates, where  $b$  depends on the local number of sub-clusters at each site and their global composition. This process naturally affects sub-clusters initially assigned to multiple global clusters.

Another important aspect of this algorithm is that the merging is a labelling operation, i.e. each local site can generate the global model which is the correspondences between local sub-clusters, without necessarily reconstructing the overall clustering output. That is

because the only bookkeeping needed from the other sites is *centres*, *sizes* and *variances*. The aggregation is defined as a labelling process between local sub-clusters in each site. No datasets move is needed. On the other hand, the perturbation process is activated if the merging action is no longer applied. The perturbation candidates are collected for each global cluster from its border, which is proportional to the overall size composition. Then, this process moves these candidates by trying the closest ones and with respect to the gain in the variance criterion when moving them from the neighbouring global clusters. Formal definitions and details of this algorithm are given in (Aouad L. M. et al., 2007) and (Aouad L. M. et al. 2008b).

### Complexity and performance

The computational complexity of this distributed algorithm depends on the algorithm used locally, the global assessment algorithm, the communication time which is a gather operation, and finally the merging computing time. If the local clustering algorithm is K-means for example, the clustering complexity is  $O(N_{max} k_{max} d)$ , where  $d$  is the number of attributes. The complexity of the global assessment depends on the size of local statistics. If the gap statistic is used on local centers, this will be  $O(B(\sum k_i)^2)$ , where  $B$  is the number of the reference distributions. The communication cost is  $3d \sum t_{comm}^i k_i$ . Since  $k_i$  is much smaller than  $N_i$ , the generated communication overhead is very low.

The merging process is executed  $u$  times. This is the number of iterations until the merging condition is no longer applied. This requires  $ut_{newStatistics} = O(d)$ . This is followed by a perturbation process which is of order of  $O(bk_g k_{max})$ . Indeed, since this process computes for each of the  $b$  chosen sub-clusters at the border of a given cluster  $C_i$ ,  $k_i$  distances for each of the  $k_g$  global clusters. The total complexity is then  $O(dN_i (\sum k_i)^2)$  ( $T_{comm} \ll O(N_i k_i d)$ ).

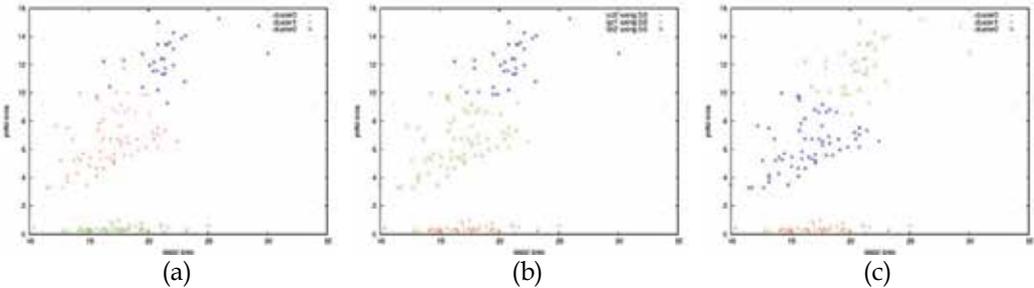


Fig. 3. Generated clustering .using 5 (a) and 7 (b) sub-clusters, and a centralised clustering using K-HarmonicMeans in (c).

The algorithm is tested on a range of artificial and real world datasets including large Gaussian distributions, the well-known Iris dataset, the animal dataset, and the PUMS census dataset available from the UC Irvine KDD Archive. The algorithm finds the inherent number of clusters by varying the maximum variance constraint, independently of the local clustering algorithm and the number of sub-clusters. An example using the Iris dataset, where the maximum variance constraint was twice the highest individual variance, is shown in the Fig. 3. In this case, since the k-harmonic-means does not impose a variance constraint it can find a lower sum-of-squared-error locally. However, the variance-based

clustering finds the 3 initial classes based on 5 and 7 sub-clusters locally using the same increasing variance value. More experiments and evaluation are shown in (Aouad L. M. et al., 2007) and (Aouad L. M. et al. 2008b).

### 3.1.2 Grid-based frequent itemsets mining

The frequent itemsets mining task is at the core of various data mining applications. Since its inception, many frequent itemsets mining algorithms have been proposed (Agrawal R. & Srikant R., 1994), (Brin S. et al., 1997), (Han J. et al., 2000), (Park J. S. et al., 1995), (Savasere A. et al., 1995), among others. Many of these approaches are based on the Apriori or the FP-Growth principles. Basically, frequent itemsets generation algorithms analyse the dataset to determine which combination of items occurs together frequently. For instance, considering the commonly known market basket analysis; each customer buys a set of items representing his/her basket. The input of the algorithm is a list of transactions giving the sets of items among all existing items in each basket. For a fixed support threshold  $s$ , the algorithm determines which sets of items of a given size  $k$  are contained in at least  $s$  transactions.

The focus then is on mining frequent itemsets on distributed datasets over the Grid. Such a Grid-based approach is motivated by the challenge of developing scalable solutions for a highly computationally expensive and data intensive application. Indeed, effective distributed approaches for large-scale data mining should take into account both the challenges raised by the underlying Grid system and the complexity of the task itself. For the purpose of developing well-adapted Grid implementations, we introduce a performance study of frequent itemsets mining of large distributed datasets on the Grid, based on the widely used Apriori principle.

We studied the distributed aspect and the performance of Apriori-based approaches both theoretically and experimentally (Aouad L. M. et al., 2008). The theoretical study presents a performance model of distributed algorithms based on the Apriori principal. Note that the main factor of an Apriori-based distributed algorithm is the number of candidates generated at each step or level. This factor, which governs the algorithm complexity, can be exponential of the size of the input. Considering the case where every transaction, in every site, contains every item, the algorithm must output and may communicate each subset of the whole set of items. we show that local pruning strategies are sufficient and that global phases in classical distributions affect the performance of the system when using the Apriori principal.

The proposed approach has two main phases. The first phase consists of generating frequent itemsets on each node based only on their local datasets. This phase is the *local mining phase* and it uses the sequential Apriori algorithm. After this phase, the result will be the set of all locally frequent itemsets in each node. This information is sufficient for determining all globally frequent itemsets, using a top-down search. The second phase is the *global collection phase*. Each node broadcasts its frequent itemsets, of size  $k$  (*requested size*) and the maximal ones, to the others nodes of the system and asks for their respective support counts. The globally frequent itemsets are then identified by merging local support counts from each node. Then, the algorithm iterates on the subsets of itemsets that fail the global frequency test. More precisely the globally frequent itemsets are generated as follows:

1. Initially collect support counts of frequent itemsets of the requested size  $k$  and all smaller frequent itemsets that are not subsets of any larger frequent itemset (maximal itemsets),

2. Generate globally frequent itemsets and put all the itemsets that are not globally frequent in a set  $F$ .
3. If  $F$  is not empty, collect support counts of subsets of itemsets in  $F$  and go to (2).

This top-down search has been shown to be efficient in large Grids, and the overheads due to synchronisations and communications are significantly reduced. Indeed, this leads to much fewer communication passes. The global pruning steps in classical distributed approaches are computationally inefficient in local nodes and affects the global system performance.

### Discussion and evaluation

Comparisons with a classical Apriori-based distributed approach, namely the Fast Distributed Mining of association rules (FDM), show that in terms of computation, both algorithms perform approximately the same amount of work as they have the same amount of candidates in the local Apriori generation. However, in terms of communication, the proposed top-down approach performs better and has only two communication passes for a range of synthetic and real datasets, namely the PUMS census dataset, and datasets generated using the IBM Quest code. The IBM Quest code is a simulation model for supermarket basket data. It has been used in several frequent itemsets generation studies such as (Han J. et al., 2000), (Purdum P. W. et al., 2004), and (Schuster A. et al., 2003), etc.

As an example, Fig. 4 shows plots of different candidates sets on different nodes using various support thresholds, on the two mentioned datasets. The lower bound, which is the ratio between the number of candidate sets of the two techniques, is  $0.78$ . This value is close to  $1$  in most cases, with an average of  $0.93$ . If we look at the ratio of the number of 1-itemsets for the two techniques we can see the same behaviour with an average of  $0.94$ . One can conclude that the difference in terms of candidate set generation between the two

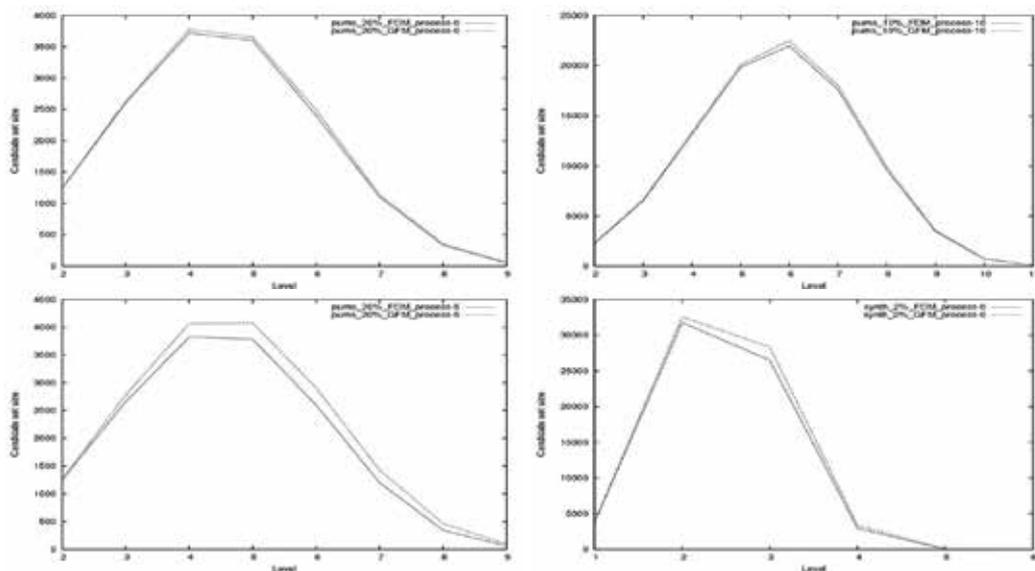


Fig. 4. Generated candidate sets using both approaches, on different processes.

techniques is not significant. In terms of processing time, this is in the order of few seconds in all cases. For the overall computation costs, the proposed technique has a gain factor of up to 82%. However, this highly depends on the size of frequent itemsets and the number of

communication passes. Also, the inputs and outputs requirements were not considered in this model for simplicity. This is likely to be more costly for classical distributed approaches since the proposed approach generates less important overall sets for remote support count collection.

The results show that distributed implementations of the Apriori algorithm do not need global pruning strategies. Therefore, classical distributions are less efficient than the adopted global strategy in our approach, starting from the requested size and using a top-down search. Note that remote support counts computations can be very expensive in classical distribution, especially in lower levels where the numbers of locally frequent itemsets are very high. This was avoided and reduced to a minimum in the proposed approach since only a few passes of remote computations are required and with smaller sizes. Formal definitions and more detailed results are presented in (Aouad L. M. et al., 2008a). This method is intended not only to reduce synchronisation and communication overheads but also the grid tools overheads related to jobs preparation or scheduling for instance.

### 3.2 Knowledge map

The knowledge map concept represents what is called *the knowledge about knowledge*. This basically means the sources, structures, and representation of the acquired knowledge. Different knowledge map structures can be found in the literature including hierarchical or radial knowledge structure maps, networked knowledge maps, knowledge source maps and knowledge flow maps (Jetter A., 2006). These knowledge maps do not codify the knowledge itself, but rather guide the way and help to find the knowledge. Often, geographical maps are used as a metaphor for knowledge maps in the sense that it simplifies complex reality, downsizes it to the important aspects, and add relevant information that help to detect the way, i.e. the knowledge and its source. In the ADMIRE project, we have developed a well-adapted knowledge mapping approach for an efficient knowledge retrieval on large Grid systems. The following of this section will briefly discuss the knowledge representation in general, and then focus on the structure of our knowledge map and its evaluation.

There are different ways of representing the acquired knowledge from mined data. This includes tables, decision trees, rules-based, instance-based, clusters, etc. One of the most popular approaches to knowledge representation is production rules, also called the *if-then* rules. In cluster approaches, the output takes the form of a diagram showing how instances fall into clusters. There are many kinds of cluster representations such as space partitioning, Venn diagram, table, tree, etc. Other knowledge representation approaches, such as Petri net, Fuzzy Petri nets, or G-net were also developed and used.

#### 3.2.1 The knowledge map structure

This section briefly describes the Knowledge Map (KM) structure. More details can be found in (Le-Khac N-A. et al., 2007) and (Le-Khac N-A. et al., 2008). In our context, the KM facilitates the deployment of distributed DM by supporting users' coordination and interpretation of the results. The objectives of our KM architecture are: 1) to provide an efficient way to handle a large amount of data collections in large-scale distributed systems; 2) retrieving easily, quickly, and accurately the knowledge; and 3) supporting the integration process of the results. In order to achieve these goals, KM system consists of the

following components: knowledge navigator, knowledge map core, knowledge retrieval, local knowledge map and knowledge map manager (Fig. 5).

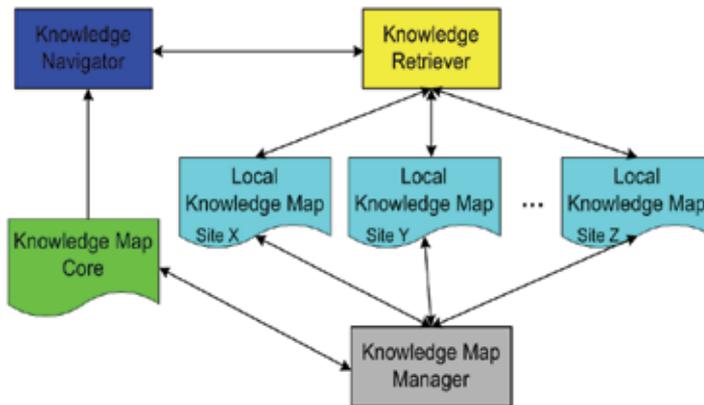


Fig. 5. Knowledge Map structure.

### Knowledge navigator

The knowledge navigator component is responsible for guiding users to explore the *KM* and for determining the knowledge of interest. The result of this task is not the knowledge but its meta-data, called *meta-knowledge*, which includes related information such as data mining tasks used, data type, and a brief description of this knowledge and its location. For example, a user might want to retrieve some knowledge about tropical cyclone. The application domain "meteorology" is used by this component to navigate the user through tropical cyclone area and then a list of information related to it will be extracted. Next, based on this meta-knowledge and its application domain, the users will decide which knowledge and its location are to be retrieved.

### Knowledge Map Core

This component (Fig. 6) is composed of two main parts: *concept tree repository* and its *meta-knowledge repository*. The former is a repository storing a set of application domains. Each application domain is represented by a *concept tree* that has a hierarchical structure such as a concept map (Novak J. D., 1984). A node of this tree, so called *concept node* represents a sub-application domain and it includes a unique identity in the whole *concept tree* repository and the name of its sub-application domain. The content of each *concept tree* is defined by the administrator before using *KM* system. In our approach, a mined knowledge is assigned to only one sub-application domain and this assignment is given by the user. By using *concept tree*, we can deal with the problem of knowledge context. For instance, given the distributed nature of the knowledge, some of them may have variations depending on the context in which it is presented locally.

### Knowledge Retrieval

The role of this component is to seek the knowledge that is potentially relevant. This task depends on the information provided by the users after navigating through application domains and getting the meta-knowledge needed. This component is similar to a search engine which interacts with each site and collects the local knowledge.

### Local Knowledge Map

This component is local to each site of the system. *Local knowledge map* is a repository of knowledge entries. Each entry, which is a knowledge object, represents a mined knowledge

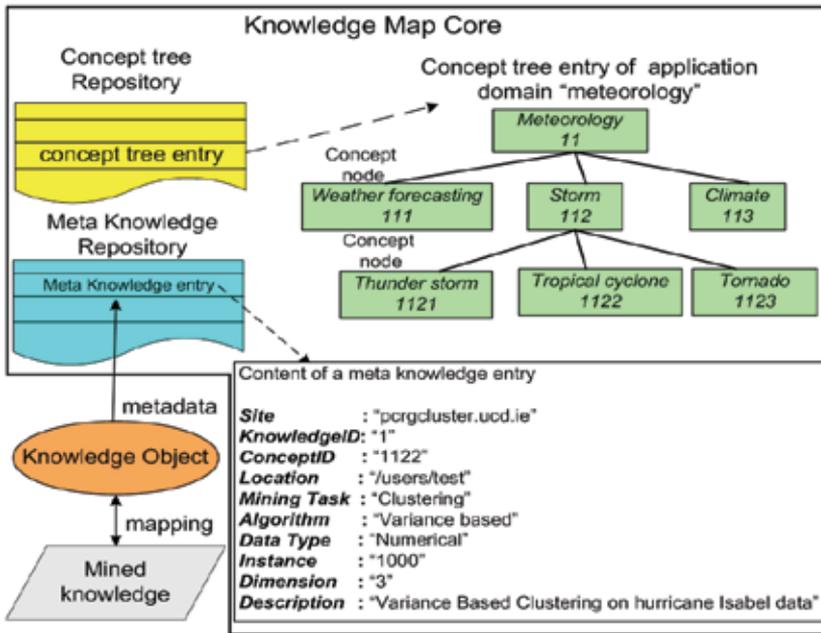


Fig. 6. Knowledge Map Core structure

and contains two parts: *meta-knowledge* and a *representative*. *Meta-knowledge* includes information such as the identity of its mined knowledge that is unique in this site, its properties, and its description. This *meta-knowledge* is also submitted to the *Knowledge map core* and will be used in *meta-knowledge entry* of its repository to be used at the global level. The *representative* of a knowledge entry depends on a given mining task. Currently, *KM* supports two kinds of representatives: one for clustering task and another for rule-based knowledge. Moreover, our system has the capacity of adding more representative types for other mining tasks.

For rule-based knowledge, the mined knowledge is represented as a set of production rules (Buchanan B. G. and Shortlife E. H., 1984). As mentioned above, a rule is of the form "if {*cause expression*} then {*conclusion expression*}" and an expression (cause or conclusion) contains a set of items. A rule also includes its attributes such as *support* and *confidence* in the association rules task or *coverage* and *accuracy* in the classification task (Han J. and Kamber M., 2006).

In the clustering case, a representative of the mined knowledge stands in one or many clusters. A cluster has one or more representative elements and each element consists of fields filled by the user. The number of fields as well as data type of each field depends on the clustering algorithm used. The meta-data of these fields is also included in each representative. A cluster also contains information about its creation. This information shows how this cluster was created: by clustering or integration process. In the former case, the information is a tuple of {*hostname*, *cluster filename*, *cluster identity*} and in the latter, it is a tuple of {*hostname*, *knowledge identity*, *cluster identity*}, where *hostname* is the location of the clustering results, which are stored in files called cluster files with their *cluster filenames*. Each cluster has a *cluster identity* and it is unique in its knowledge entry. In (Le-Khac N-A. et al., 2008), the authors describe both kinds of representatives in details.

### Knowledge Map Manager

The knowledge map manager is responsible for managing and coordinating the local knowledge maps and the knowledge map core. For *local knowledge map*, this component provides primitives to create, add, delete, and update knowledge entries and their related components in knowledge repository. It also allows to submit local meta-knowledge to its repository in *knowledge map core*. This component provides also primitives to handle the meta-knowledge in the repository as well as the concept node in the concept tree repository.

#### 3.2.3 Evaluation

The KM layer is presently being tested in the ADMIRE system. It is difficult to evaluate our approach by comparing it to other systems because it is unique so far. Therefore, this new approach is validated by evaluating different aspects of the system architecture for supporting the management, mapping, representing and retrieving the knowledge. First, we evaluate the complexity of search/retrieve the knowledge object of the system. This operation includes two parts: searching relative concept and search/retrieve the knowledge. Let  $N$  be the number of *concept tree* entries and  $n$  be the number of *concept nodes* for each *concept tree*. The complexity of the first part is  $O(\log N + \log n)$  because the concept tree entries are indexed according to a tree model. However, the number of concept entries as well as of concept nodes of a concept tree is negligible compared to the number of knowledge entries. So this complexity depends strongly on the cost of search/retrieve operations. Let  $M$  be the number of meta-knowledge entries in the *KM core*, so the complexity of searching a meta-knowledge entry at this level is  $O(\log M + C_s)$ , where  $C_s$  is the communication cost between a node  $s$  and the host node where the meta-knowledge repository is stored. This depends on the bandwidth between two nodes and the size of the data size. The complexity of retrieving a knowledge object is the same as for the search operation. However, the retrieve operation depends on the number of knowledge entries  $m$  in the *local KM*.

Tests have been done to evaluate the search/retrieve performance. More details about these tests can be found in (Le-Khac N-A. et al., 2007) and (Le-Khac N-A. et al., 2008). Next, we estimate the performance of the *KM* architecture. Firstly, the structure of *concept tree* is based on the concept map. We can avoid the problem of semantic ambiguity as well as reduce the domain search to improve the speed and accuracy of the results. In the 1-n model (one server-n client nodes), the concept tree is implemented either only at the server node or at each client node. The client-server communication is needed when we interact with concept tree via the operations add, search, delete concept nodes or get the concept identity when adding new knowledge. In a large distributed system, this concept tree can be cached at each local node to reduce the communication cost because the numbers of operations of add/delete a concept node is very small compared to the number of search operations.

Secondly, the division of knowledge map into two main components (local and core) has some advantages: (i) the core component acts as a summary map of knowledge and it is a representation of knowledge about knowledge when combined with local *KM*; (ii) avoiding the problem of having the whole knowledge on one master node (or server), which is not feasible on very large distributed systems such as the Grid. By representing knowledge meta-data by their relationship links, the goal is to provide an integration view of these knowledge.

Finally, this approach offers a knowledge map with flexible and dynamic architecture where users can easily update the *concept tree* repository as well as meta-knowledge entries. The current index technique used in a rule representative is an inverted list. However, we can improve it without affecting to whole system structure by using other index algorithms (Martynov M. and Novikov B, 1996) or by applying compressed technique as discussed in (Zobel J. and Moffat A., 2006). Moreover, flexible and dynamic features are also reflected by mapping a knowledge to a *knowledge object*. The goal here is to provide a portable approach where knowledge object can be represented by different techniques such as an entity, an XML-based record, or a record of database, etc.

#### 4. Discussion

Grid-based knowledge discovery has to address key issues related to the three main aspects of this area, namely the distributed algorithms for generating efficient global models, the knowledge representation and retrieval, and some specific Grid issues which raise questions on how to use a given Grid architecture and mechanisms to build algorithms and knowledge-based operations, i.e. whether high-level approaches for mining and representing knowledge are suitable for the Grid. We discussed different projects which focus, for the large part, on the architectural aspect, i.e. how to interface classical data mining and knowledge discovery operations with existing Grid technologies. In contrast, the ADMIRE framework offers more focus on scalable algorithms, and means to codify the knowledge about knowledge in large distributed Grids.

Grid algorithms are unlikely to scale if they require an extensive communication load. Indeed, straightforward distributions of many DM tasks have little choice but to exchange information between every possible pair of sites or nodes in the Grid. This is not scalable on large distributed systems. However, we might be able to decompose and/or approximate the problem and eliminate this communication needs. The notion of locality is then very important in DM and KD on the Grid. The typical way introduced in this chapter involves *local* data analysis, followed by the generation of a *global* data and knowledge model through the aggregation of the local results in different manners. For instance, many algorithms in peer-to-peer Grids use different network topologies and organisations in order to use properly the neighbourhood notion. Several algorithms have been developed to address basic problems, however multiple challenges still exist in terms of performance, accuracy, communication and scaling behaviour, convergence properties and stability for approximation techniques, privacy-preserving, and trust management, among others.

As for the knowledge-related operations, the chapter has briefly discussed the knowledge map notion as means to codify the knowledge about knowledge. Then, some of the ADMIRE's mapping operations and knowledge representation were presented. The main objective here is to codify both the knowledge and its navigation in order to improve the detection and retrieval of knowledge in large Grids. However, additional means of codification or communication can be taken into account in order to capture user-specific knowledge domains. This is part of the knowledge assessment which takes place prior to the the knowledge mapping itself. In addition, different application domains might lead to different results and could demonstrate the need of other operations and/or different underneath structure. This has to be taken into account in the future.

On the other hand, the Grid offers a large range of technologies, architectures and implementations, although standardisation works have been undertaken in recent years.

This makes it difficult to propose an *open* and *flexible* distributed and Grid-based knowledge discovery architecture that can be configured on top of various Grid middleware in a simple way. Furthermore, typical computational issues in the Grid, such as the important computing overhead, make the straightforward adaptation of classical DM infeasible. Some other inherent characteristics might not have been taken into account at the middleware level, and have to be addressed at higher levels, such as properties of the data management policies, replication, authentication, data protection and privacy, among others.

## 5. Conclusion

Mining large-scale datasets and the extraction, representation, and retrieving of knowledge on Grid systems still an active and challenging research area, either domain-specific or not. Several research work have been done so far including the most known projects shortly reviewed in this chapter. We also discussed different trends and focuses of these projects. Then, the motivations, design, and original aspects of ADMIRE have been presented. ADMIRE is a domain-independent solving environment that allows the user to express a problem using his/her own domain specific knowledge to build its application using basic data analysis and data mining operations. It offers lightweight distributed approaches able to perform large-scale computation to leverage the Grid in an efficient way.

ADMIRE also tackles the issue of clear and easy representation and manipulation of the knowledge by proposing its knowledge map layer. While the concept of the knowledge map itself is not new, its structure and implementation offer a novel and robust knowledge management and retrieval in large distributed Grids. In future works, we will take into account some domain-specific and real-world applications constraints and properties, in order to achieve its potential and enable the full use of the Grid for each of them.

## 6. References

- Agrawal R. and Srikant R. (1994). Fast Algorithms for Mining Association Rules. VLDB'94: Proceeding of the 20th Int. Conf. Very Large Data Bases.
- Aouad L. M., Le-Khac N-A. and Kechadi M-T. (2008). Performance Study of Distributed Apriori-like Frequent Itemset Mining. Technical report, University College Dublin, 2008.
- Aouad L. M., Le-Khac N-A. and Kechadi M-T. (2008). A Multi-Stage Clustering Algorithm for Distributed Data Mining Environments. COSI 2008, Colloque sur l'Optimisation et les Systemes d'Information. Tizi-Ouzou, Algeria.
- Aouad L. M., Le-Khac N-A. and Kechadi M-T. (2007). Lightweight Clustering Technique for Distributed Data Mining Applications. The 7th Industrial Conference on Data Mining ICDM 2007, Springer LNAI 4597.
- Bellec J., Kechadi M-T., and Carthy J. (2005). A New Efficient Clustering Algorithm for Network Alarm Analysis. The 17th IASTED Intl. Conference on Software Engineering and Applications PDCS 2005.
- Berman F. 2001. Viewpoint: From TeraGrid to knowledge grid. Commun. ACM, Vol. 44, No. 11, 2001.
- Brin S. and Motwani R. and Ullman J. D. and Tsur S. (1997). Dynamic Itemset Counting and Implication Rules for Market Basket Data. SIGMOD'97: Proceedings ACM SIGMOD Int. Conf. on Management of Data.

- Buchanan B. G. and Shortliffe E. H. (1984) Rule-Based Expert Systems: The MYCIN Experiments of The Stanford Heuristic Programming Projects Reading, MA: Addison-Wesley.
- Calinski R. B. and Harabasz J. (1974). A dendrite method for cluster analysis. *Communication in statistics*, Vol. 3, 1974.
- Campieta P., Di Martino S., Bertolotto M., Ferrucci F., and Kechadi M-T. (2007). Exploratory Spatio-Temporal Data Mining and Visualization. *Journal of Visual Languages and Computing*, Vol. 18, No. 3, 2007.
- Dhillon I. S. and Modha D. (1999). A Data-Clustering Algorithm on Distributed Memory Multiprocessors. *Workshop on Large-Scale Parallel KDD Systems, SIGKDD, 1999.*
- Ellahi T. N. and Kechadi M-T. (2004). Distributed Resource Discovery In Wide Area Grid Environments. *LNCSE on Computational Science*, 3038, May 2004.
- Ester M., Kriegel H-P and Sander J. and Xu X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *2nd Int. Conf. on Knowledge Discovery and Data Mining (KDD)*, 1996.
- Garg A., Mangla A., Bhatnagar V., and Gupta N. (2006). PBIRCH : A Scalable Parallel Clustering algorithm for Incremental Data. *10th International Database Engineering and Applications Symposium, IDEAS 2006.*
- Geng H. and Deng X. and Ali H. (2005). A New Clustering Algorithm Using Message Passing and its Applications in Analyzing Microarray Data. *Proceedings of the Fourth International Conference on Machine Learning and Applications, ICMLA 2005.*
- Han J. and Jian Pei and Yiwen Yin (2000). Mining frequent patterns without candidate generation. *Proceedings of the 2000 ACM SIGMOD int. conference on Management of Data.*
- Han J. and Kamber M. (2006). *Data Mining: Concepts and Techniques*. 2nd ed Morgan Kaufmann Publishers.
- Hegarty D. F. and Kechadi M-T. (1996). Topology Preserving Dynamic Load Balancing for Parallel Molecular Simulations. *ACM/IEEE Int. Conf. on Supercomputing*, 1997.
- Hudzia B., Kechadi M-T., and Ottewill A. (2005). TreeP: A Tree-Based P2P Network Architecture. *IEEE, International Workshop on Algorithms, Models and tools for parallel computing on heterogeneous networks, HeteroPar 2005.*
- Hudzia B., McDermott L., Illahi T. N., and Kechadi M-T. (2005). Entity Based Peer-to-Peer in a Data Grid Environment. *The 17th IMACS World Congress Scientific Computation, Applied Mathematics and Simulation*, 2005.
- Jain A. K., Murty M. N. and Flynn P. J. (1999). *Data Clustering: A Review*. *ACM Computing Surveys*, Sep. 1999.
- Januzaj E., Kriegel H-P., and Pfeifle M. (2003). Towards Effective and Efficient Distributed Clustering. *Int. Workshop on Clustering Large Data Sets*, 3rd Int. Conf. on Data Mining, *ICDM 2003.*
- Januzaj E. and Kriegel H-P. and Pfeifle M. (2004). Scalable Density-Based Distributed Clustering. *8th European Conference on Principles and Practice Discovery in Databases*, *PKDD 2004.*
- Januzaj E. and Kriegel H-P. and Pfeifle M. (2004). DBDC: Density-Based Distributed Clustering. *9th Int. Conf. on Extending Database Technology*, *EDBT 2004.*
- Jin R. and Goswami A. and Agrawal G. (2006). Fast and Exact Out-of-Core and Distributed K-Means Clustering. *Knowledge and Information Systems*, Vol. 10, 2006.
- Joshi M. N. (2003). *Parallel K-Means Algorithm on Distributed Memory Multiprocessors*. Technical report, University of Minnesota, 2003.

- Le-Khac N.A. and Kechadi M-T. (2005). ADMIRE Framework: Distributed Data Mining on Data-Grid Platforms. Intl. Conference on Software and Data Technologies, ICSoft 2006.
- Le-Khac N-A., Aouad L. M. and Kechadi M-T. (2007). Knowledge Map: Toward a new approach supporting the knowledge management in Distributed Data Mining, KUI Workshop, IEEE International Conference on Autonomic and Autonomous Systems ICAS'07, Athens, Greece, 2007.
- Le-Khac N-A., Aouad L. M. and Kechadi M-T. (2008). An Efficient Knowledge Management Tool for Distributed Data Mining Environments", International Journal of Computational Intelligence Research, ISSN 0973-1873, 2008.
- Martynov M. and Novikov B. (1996). An Indexing Algorithm for Text Retrieval, Proceedings of the International Workshop on Advances in Databases and Information system (ADBIS'96), Moscow: 171-175.
- Mingjin Y. and Keying Y. (2007). Determining the Number of Clusters Using the Weighted Gap Statistic. Biometrics, Vol. 63, No. 4, 2007.
- Ng R. T. and Han J (1994). Efficient and Effective Clustering Methods for Spatial Data Mining. VLDB, Proceedings of 20th International Conference on Very Large Data Bases, September 12-15, 1994.
- Novak J.D. and Gowin D.B. (1984). Learning how to learn, Cambridge University Press.
- Purdom P. W. and Van Gucht D. and Groth D. P. (2004). Average-Case Performance of the Apriori Algorithm. SIAM Journal on Computing, Vol. 33, No. 5. 2004.
- Savasere A. and Omiecinski E. and Navathe S. B. (1995). An Efficient Algorithm for Mining Association Rules in Large Databases. VLDB'95: Proceedings of the 21th International Conference on Very Large Databases.
- Schuster A. and Wolff R. and Trock D. (2003). A High-Performance Distributed Algorithm for Mining Association Rules. ICDM'03: Proceedings of the Third IEEE International Conference on Data Mining.
- Soo Park J. and Ming-Syan Chen and Philip S. Yu (1995). An effective hash-based algorithm for mining association rules. SIGMOD'95: Proceedings of the 1995 ACM SIGMOD international conference on Management of Data.
- Tibshirani R. and Walther G. and Hastie T. (2000). Estimating the number of clusters in a dataset via the Gap statistic. Technical report, Stanford University, March 2000.
- Xu R. and Wunsch D. (2005). Survey of Clustering Algorithms. IEEE Transactions on Neural Networks, Vol. 16, May 2005.
- Xu X. and Jager J. and Kriegel H.-P. (1999). A Fast Parallel Clustering Algorithm for Large Spatial Databases. Journal of Data Mining and Knowledge Discovery, Vol. 3, 1999.
- Zhang B. and Hsu M. and Dayal U. (1999). K-Harmonic Means - A Data Clustering Algorithm. Technical report, HP Labs, 1999.
- Zhang B. and Forman G. (2000). Distributed Data Clustering Can be Efficient and Exact. Technical report, HP Labs, 2000.
- Zobel J., Moffat A. (2006). Inverted Files for Text Search Engines, Journal of ACM Computing Surveys, Vol. 38, No.2, Article 6.

# Rough Set Theory – Fundamental Concepts, Principals, Data Extraction, and Applications

Silvia Rissino<sup>1</sup> and Germano Lambert-Torres<sup>2</sup>

<sup>1</sup>*Federal University of Rondonia,*

<sup>2</sup>*Itajuba Federal University*

*Brazil*

## 1. Introduction

Rough Set Theory, proposed in 1982 by Zdzislaw Pawlak, is in a state of constant development. Its methodology is concerned with the classification and analysis of imprecise, uncertain or incomplete information and knowledge, and of is considered one of the first non-statistical approaches in data analysis (Pawlak, 1982).

The fundamental concept behind Rough Set Theory is the approximation of lower and upper spaces of a set, the approximation of spaces being the formal classification of knowledge regarding the interest domain.

The subset generated by lower approximations is characterized by objects that will definitely form part of an interest subset, whereas the upper approximation is characterized by objects that will possibly form part of an interest subset. Every subset defined through upper and lower approximation is known as Rough Set.

Over the years Rough Set Theory has become a valuable tool in the resolution of various problems, such as: representation of uncertain or imprecise knowledge; knowledge analysis; evaluation of quality and availability of information with respect to consistency and presence a not of date patterns; identification and evaluation of date dependency; reasoning based an uncertain and reduct of information data.

The extent of rough set applications used today is much wider than in the past, principally in the areas of medicine, analysis of database attributes and process control. The subject of this chapter is to present the Rough Set Theory, important concepts, and Rough Set Theory used with tools for data mining, special applications in analysis of data in dengue diagnosis.

The chapter is divided into the four following topics:

- Fundamental concepts
- Rough set with tools for data mining
- Applications of rough set theory;
- Case - Rough set with tools in dengue diagnosis.

## 2. Fundamental concepts

Rough Sets Theory has been under continuous development for over years, and a growing number of researchers have become its interested in methodology. It is a formal theory derived from fundamental research on logical properties of information systems. From the

outset, rough set theory has been a methodology of database mining or knowledge discovery in relational databases. This section presents the concepts of Rough Set Theory; which coincide partly with the concepts of other theories that treat uncertain and vagueness information. Among the existent, most traditional approaches for the modeling and treatment of uncertainties, they are the Theories of the Uncertainty of Dempster-Shafer and Fuzzy Set (Pawlak et al., 1995). The main concepts related to Rough Set Theory are presented as the following:

### 2.1 Set

A set of objects that possesses similar characteristics it is a fundamental part of mathematics. All the mathematical objects, such as relations, functions and numbers can be considered as a set. However, the concept of the classical set within mathematics is contradictory; since a set is considered to be "grouping" without all elements are absent and is know as an empty set (Stoll, 1979). The various components of a set are known as elements, and relationship between an element and a set is called of a pertinence relation. Cardinality is the way of measuring the number of elements of a set. Examples of specific sets that treat vague and imprecise date are described below:

#### a. Fuzzy Set

Proposed by mathematician Loft Zadeh in the second half of the sixties, it has as its objective the treatment of the mathematical concept of vague and approximate, for subsequent programming and storage on computers.

In order for Zadeh to obtain the mathematical formalism for fuzzy set, it was necessary to use the classic set theory, where any set can be characterized by a function. In the case of the fuzzy set, the characteristic function can be generalized so that the values are designated as elements of the Universe Set  $U$  belong to the interval of real numbers  $[0,1]$ .

The characteristic Function Fuzzy is  $\mu_A: U \rightarrow [0,1]$ , where the values indicate the degree of pertinence of the elements of set  $U$  in relation to the set  $A$ , which indicated as it is possible for an element of  $x$  of  $U$  to belong to  $A$ , this function is known as Function of Pertinence and the set  $A$  is the Fuzzy Set (Zadeh, 1965).

#### b. Rough Set

An approach first forwarded by mathematician Zdzislaw Pawlak at the beginning of the eighties; it is used as a mathematical tool to treat the vague and the imprecise. Rough Set Theory is similar to Fuzzy Set Theory, however the uncertain and imprecision in this approach is expressed by a boundary region of a set, and not by a partial membership as in Fuzzy Set Theory. Rough Set concept can be defined quite generally by means of interior and closure topological operations know approximations (Pawlak, 1982).

Observation:

It is interesting to compare definitions of classical sets, fuzzy sets and rough sets. Classical set is a primitive notion and is defined intuitively or axiomatically. Fuzzy set is defined by employing the fuzzy membership function, which involves advanced mathematical structures, numbers and functions. Rough set is defined by topological operations called approximations, thus this definition also requires advanced mathematical concepts.

### 2.2 Information system or information table

An information system or information table can be viewed as a table, consisting of objects (rows) and attributes (columns). It is used in the representation of data that will be utilized by Rough Set, where each object has a given amount of attributes (Lin, 1997).

These objects are described in accordance with the format of the data table, in which rows are considered objects for analysis and columns as attributes (Wu et al., 2004). Below is shown an example of an information Table 1.

Patient	Attributes			
	Headache	Vomiting	Temperature	Viral illness
#1	No	Yes	High	Yes
#2	Yes	No	High	Yes
#3	Yes	Yes	Very high	Yes
#4	No	Yes	Normal	No
#5	Yes	No	High	No
#6	No	Yes	Very high	Yes

Table 1. Example of information table

### 2.3 Indiscernibility relation

Indiscernibility Relation is a central concept in Rough Set Theory, and is considered as a relation between two objects or more, where all the values are identical in relation to a subset of considered attributes. Indiscernibility relation is an equivalence relation, where all identical objects of set are considered as elementary (Pawlak, 1998)

In Table 1, presented in section 2.2, it can be observed that the set is composed of attributes that are directly related to the patients' symptoms whether they be headache, vomiting and temperature. When Table 1 is broken down it can be seen that the set regarding {patient2, patient3, patient5} is indiscernible in terms of headache attribute. The set concerning {patient1, patient3, patient4} is indiscernible in terms of vomiting attribute. Patient2 has a viral illness, whereas patient5 does not, however they are indiscernible with respect to the attributes headache, vomiting and temperature. Therefore, patient2 and patient5 are the elements of patients' set with unconcluded symptoms.

### 2.4 Approximations

The starting point of rough set theory is the indiscernibility relation, generated by information concerning objects of interest. The indiscernibility relation is intended to express the fact that due to the lack of knowledge it is unable to discern some objects employing the available information. Approximations is also other an important concept in Rough Sets Theory, being associated with the meaning of the approximations topological operations (Wu et al., 2004). The lower and the upper approximations of a set are interior and closure operations in a topology generated by the indiscernibility relation. Below is presented and described the types of approximations that are used in Rough Sets Theory.

#### a. Lower Approximation ( $B''$ )

Lower Approximation is a description of the domain objects that are known with certainty to belong to the subset of interest.

The Lower Approximation Set of a set  $X$ , with regard to  $R$  is the set of all of objects, which certainly can be classified with  $X$  regarding  $R$ , that is, set  $B''$ .

#### b. Upper Approximation ( $B^*$ )

Upper Approximation is a description of the objects that possibly belong to the subset of interest. The Upper Approximation Set of a set  $X$  regarding  $R$  is the set of all of objects which can be possibly classified with  $X$  regarding  $R$ , that is, set  $B^*$ .

c. Boundary Region (BR)

Boundary Region is description of the objects that of a set  $X$  regarding  $R$  is the set of all the objects, which cannot be classified neither as  $X$  nor  $-X$  regarding  $R$ . If the boundary region is a set  $X = \emptyset$  (Empty), then the set is considered "Crisp", that is, exact in relation to  $R$ ; otherwise, if the boundary region is a set  $X \neq \emptyset$  (empty) the set  $X$  "Rough" is considered. In that the boundary region is  $BR = B^* - B''$ .

Mathematically speaking, let a set  $X \subseteq U$ ,  $B$  be an equivalence relation and a knowledge base  $K = (U, B)$ . Two subsets can be associated:

1. B-lower:  $B'' = \cup \{Y \in U/B : Y \subseteq X\}$
2. B-upper:  $B^* = \cup \{Y \in U/B : Y \cap X \neq \emptyset\}$

In the same way,  $POS(B)$ ,  $BN(B)$  and  $NEG(B)$  are defined below (Pawlak, 1991).

3.  $POS(B) = B'' \Rightarrow$  certainly member of  $X$
4.  $NEG(B) = U - B^* \Rightarrow$  certainly non-member of  $X$
5.  $BR(B) = B^* - B'' \Rightarrow$  possibly member of  $X$

Figure 1 presents a graphic representation of these regions (Lambert-Torres et al., 1999).

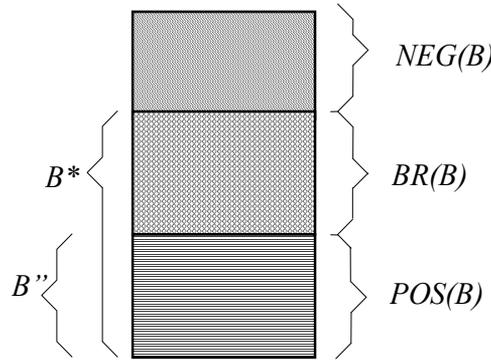


Fig. 1. Definition of B-approximation sets and B-regions

d. Quality Approximation

It is obtained numerically using its own elements, specifically those of lower and upper approximation. The coefficient used in measuring the quality is represented by  $\alpha_B(X)$ , where  $X$  is a set of objects or registrations regarding  $B$ . The quality of approximation uses two coefficients that are presented below:

- Imprecision coefficient  $\alpha_B(X)$

Where  $\alpha_B$  is the quality of approximation of  $X$ , being denoted by:

$$\alpha_B(X) = |B''(X)| / |B^*(X)| \tag{1}$$

Where  $|B''(X)|$  and  $|B^*(X)|$  it represents the cardinality of approximation lower and upper, and the approximation are set  $\neq \emptyset$ . Therefore,  $0 \leq \alpha_B \leq 1$ , if  $\alpha_B(X)=1$ ,  $X$  it is a definable set regarding the attributes  $B$ , that is,  $X$  is crisp set. If  $\alpha_B(X) < 1$ ,  $X$  is rough set regarding the attributes  $B$ .

Applying it formulates for Table 1, it has  $\alpha_B(X) = 3/5$  for the patients with possibility of they are with viral illness.

- Quality Coefficient of upper and lower approximation

- Quality Coefficient of upper approximation  $\alpha_B(B^*(X))$

It is the percent of all the elements that are classified as belonging to X, being denoted for:

$$\alpha_B(B^*(X)) = |B^*(X)| / |A| \tag{2}$$

In the Table 1,  $\alpha_B(B^*(X)) = 5/6$ , for the patients that have the possibility of they be with viral illness.

- Quality Coefficient of lower approximation  $\alpha_B(B''(X))$

It is the percentage of all the elements that possibility are classified as belonging to X, and is denoted as:

$$\alpha_B(B''(X)) = |B''(X)| / |A| \tag{3}$$

In the Table 1,  $\alpha_B(B^*(X)) = 3/6 = 1/2$ , for the patients that have viral illness.

Observation: In Quality coefficient upper and lower, presented in number 2, of this section,  $|A|$  represents the cardinality of any given set of objects.

### 2.5 Decision tables and decision algorithms

A decision table contains two types of attributes designated as the condition attribute and decision attribute. In Table 1, shown in section 2.2, the attributes of headache, vomiting and temperature can all be considered as condition attributes, whereas the viral illness attribute is considered a decision attribute.

Each row of a decision table determines a decision rule, which specifies the decisions (actions) that must be taken when conditions are indicated by condition attributes are satisfied, e.g. in Table 1 the condition (Headache, no), (vomiting, yes), (Temperature, high) determines the decision (Viral illness, yes).

Table 1 shows that both patient2 and patient5 suffer from the same symptoms since the condition attributes of headache, vomiting and temperature possess identical values; however, the values of decision attribute differ. These set of rules are known as either inconsistency, non-determinant or conflicting. These rules are known as consistency, determinant or non conflicting or simply, a rule.

The number of consistency rules, contained in the decision table are known as a factor of consistence, which can be denoted by  $\gamma(C, D)$ , where C is the condition and D the decision. If  $\gamma(C,D) = 1$ , the decision table is consistent, but if  $\gamma(C,D) \neq 1$  the table of decision is inconsistent.

Given that Table 1,  $\gamma(C,D) = 4/6$ , that is, the Table 1 possesses two inconsistent rules (patient2, patient5) and four consistent rules (patient1, patient3, patient4, patient6), inside of universe of six rules for all the Table 1 (Ziarko & Shan, 1995). The decision rules are frequently shown as implications in the form of "if... then... ". To proceed is shown one rule for the implication viral illness:

If	Headache	= no	and
	Vomiting	= yes	and
	Temperature = high		
Then	Viral Illness = yes		

A set of decision rules is designated as decision algorithms, because for each decision table it can be associated with the decision algorithm, consisting of all the decision rules that it occur in the respective decision table. A may be made distinction between decision algorithm and decision table. A decision table is a data set, whereas a decision algorithm is a collection of implications, that is, a logical expressions (Pawlak, 1991).

## 2.6 Dependency of attributes

In the analysis of data, it is important discover the dependence between attributes. Intuitively, a set of attributes  $D$  depends totally on a set of attributes  $C$ , denoted as  $C \Rightarrow D$ , if all values of attributes from  $D$  are uniquely determined by values of attributes from  $C$ , then  $D$  depends totally on  $C$ , if there exists a functional dependency between values of  $D$  and  $C$ . For example, in Table 1 there are no total dependencies whatsoever, if in Table 1, the value of the attribute Temperature for patient p5 was "no" instead of "high", there would be a total dependency  $\{\text{Temperature}\} \Rightarrow \{\text{viral illness}\}$ , because to each value of the attribute Temperature there would correspond a unique value of the attribute viral illness.

It would also necessitate a more global concept of dependency of attributes, designated as partial dependency of attributes, in Table 1, the temperature attribute determines some uniquely values of the attribute viral illness. That is, (temperature, very high) implies (viral illness, yes), similarly (temperature, normal) implies (viral illness, no), but (temperature, high) does not imply always (viral illness, yes). Thus the partial dependency means that only some values of  $D$  are determined by values of  $C$ . Formally the dependence among the attributes can be defined in the following way: If  $D$  and  $C$  are subsets of  $A$ , can be affirmed that  $D$  depends on  $C$  in degree  $K$  ( $0 \leq k \leq 1$ ), denoted  $\Rightarrow_k D$  and if  $k = \gamma(C, D)$ . If  $K=1$ ,  $D$  depends totally on  $C$ , if  $K < 1$ , it is said that  $D$  depends partially (in a degree  $K$ ) on  $C$ .

The concept of dependent attributes is strongly coupled to the concept of consistency of decision table (Pawlak, 1991). For example, for dependency attributes  $\{\text{Headache, Vomiting, Temperature}\} \Rightarrow \{\text{viral illness}\}$ , it get  $k=4/6= 2/3$ , because four out of six patients can be uniquely classified as having viral illness or not, employing attributes Headache, Vomiting and Temperature.

It can be interesting to note exactly how patients can be diagnosed using only the attribute Temperature, that is, the degree of the dependence  $\{\text{Temperature}\} \Rightarrow \{\text{viral illness}\}$ , diagnose is obtained  $k = 3/6 = 1/2$ , in this case there are only three patients {patient1, patient3, patient6} out of six, only these tree, can be classified exclusively as having viral illness. In contrast, with the case of patient4, who cannot be classified as having viral illness, since the value of the attribute temperature, in this case, is normal. Hence the single attribute Temperature offers worse classification than the whole set of attributes Headache, Vomiting and Temperature. It is interesting to observe that neither Headache nor Vomiting can be used to recognize viral illness, because for both dependencies  $\{\text{Headache}\} \Rightarrow \{\text{viral illness}\}$  and  $\{\text{Vomiting}\} \Rightarrow \{\text{Viral illness}\}$  it has  $k=0$ .

It can be easily seen that if  $D$  depends totally on  $C$  then  $I(C) \subseteq I(D)$ . That means that the partition generated by  $C$  is finer than the partition generated by  $D$ , and that the concept of dependency presented in the section corresponds to that considered in relational databases.

## 2.7 Reduction attributes in information system

For many application problems, it is often necessary to maintain a concise form of the information system, but there exist data that can be removed, without altering the basic

properties and more importantly the consistency of the system (Cerchiari et al, 2006). If is subtract relative data from the headache and vomiting, the resultant data set is equivalent to original data in relation to approximation and dependency, as it has the same the approximation precision and the same dependency degree using the original set of attributes, however with one fundamental difference, the set of attributes to be considered will be fewer.

The process of reducing an information system such that the set of attributes of the reduced information system is independent and no attribute can be eliminated further without losing some information from the system, the result is known as reduct. If an attribute from the subset  $B \subseteq A$  preserves the indiscernibility relation  $RA$ , then the attributes  $A - B$  are dispensable. Reducts are such subsets minimal, i.e., that do not contain any dispensable attributes. Therefore, the reduction should have the capacity to classify objects, without altering the form of representing the knowledge (Geng & Zhu, 2006). When the definition above is applied, the information system presented in the Table 1, B is a subset of A and a belongs to B:

- a is dispensable in B if  $I(B) = I(B - \{a\})$ ; otherwise a is indispensable in B;
- Set B is independent if all its attributes are indispensable;
- Subset B' of B is a reduct of B if B' is independent and  $I(B') = I(B)$ ; and

A reduct is a set of attributes that preserve the basic characteristics of the original data set; therefore, the attributes that do not belong to a reduct are superfluous with regard to classification of elements of the Universe.

### 3. Rough set with tools for data mining

The great advances in information technology have made it possible to store a great quantity of data. In the late nineties, the capabilities of both generating and collecting data were increased rapidly. Millions of databases have been used in business management, government administration, scientific and engineering data management, as well as many other applications. It can be noted that the number of such databases keeps growing rapidly because of the availability of powerful database systems. This explosive growth in data and databases has generated an urgent need for new techniques and tools that can intelligently and automatically transform the processed data into useful information and knowledge. One of the processes used to transform data into knowledge is Knowledge Discovery Database (KDD), which is divided in three stages (preprocessing, data mining and post processing) that are shown in the section 3.1.

#### 3.1 Knowledge Discovery in Database - KDD

Knowledge Discovery in Database - KDD is a process, with several stages, no trivial, interactive and iterative, for the identification of comprehensible patterns, valid, new and potentially useful starting from great groups of data (Fayyad et al., 1996a). KDD is characterized as a process composed of several operational stages: preprocessing, data mining and the post processing. Figure 2 presents the sequence of the stages executed during the process of KDD.

##### a. Preprocessing Stage

The preprocessing stage understands the functions related to the reception, the organization and to the treatment of data, this stage has as its objective the preparation of the data for the following stage of the data mining.

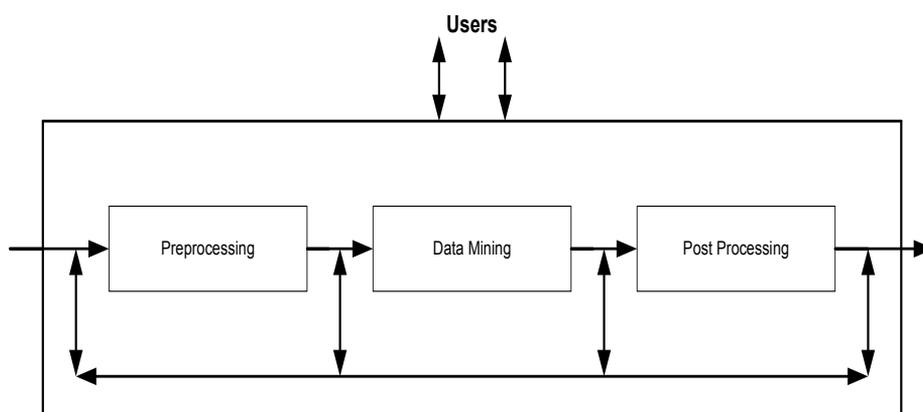


Fig. 2. The KDD process

#### b. Data Mining Stage

The data mining stage defines the techniques and the algorithms to be used by the problem in question, as are examples of techniques that can be used in this stage such as neural network, rough set, genetic algorithms, statistical models and probabilistic. The choice of technique depends, in many cases, on task type to be developed. In Table 2 is shown a summary of tasks to be accomplished and some alternative methods that can be useful. It is important to observe that Table 2 does not drain the universe of methods of data mining that can be used in each task of KDD and is purely a summary (Fayyad et al., 1996b).

Tasks of KDD	Methods of Data Mining
Discovery Associations	Basic, Apriori, DHP, Partition, DIC, ASCX-2P
Discovery Generalize of Associations	Basic, Apriori, DHP, Partition, DIC, ASCX-2P
Discovery of Sequences	GSP, MSDD, SPADE
Discovery Generalize of Sequences	GSP, MSD, SPADE
Classification	Neural Network, C4.5, Rough Sets, Genetic Algorithms, CART, K-NN, Bayes Classifier
Regression	Neural Network, Fuzzy Set
Summarization	C4.5, Genetics Algorithms
Clustering	K-Means, K-Modes, K-prototypes, Fuzzy K-Means, genetics Algorithms, Neural Network
Forecast of Temporal Series	Neural Network, Fuzzy Set

Table 2. Methods of Data Mining that can be applied in the tasks of KDD

During the data mining stage much useful knowledge is gained in respect of the application. Many authors consider data mining synonymous with KDD, in the context this stage, the

KDD process is often known as Data Mining; in the research it will Data Mining be indented as KDD (Piatetsky-Shapiro & Matheus, 1995; Mitchell, 1999; Wei, 2003).

Data mining has become an area of research increasing importance, and is also referred to as knowledge discovery in databases (KDD), consequently this has resulted in a process of non trivial extraction of implicit, previously unknown and potentially useful information, such as knowledge rules, constraints, regularities from data in databases

#### c. Post Processing Stage

In the post processing stage the treatment of knowledge obtained during the data mining stage. This stage is not always necessary; however, it allows the possibility of validation of the usefulness of the discovered knowledge.

### 3.2 Rough set in data mining

Rough set theory constitutes a consistency base for data mining; it offers useful tools for discovering patterns hidden in data in many aspects. Although in theory rough set deals with discreet data, rough set is commonly used in conjunction with other techniques connected to discretization on the dataset. The main feature of rough set data analysis is both non-invasive and notable ability to handle qualitative data. This fits into most real life applications nicely.

Rough Set can be used in different phases of the knowledge discovery process, as attribute selection, attribute extraction, data reduction, decision rule generation and pattern extraction (Komorowski et al., 1999). Furthermore, recent extensions of rough set theory with rough mereology have brought new methods of decomposition of large data sets, data mining in distributed and multi-agent based environments and granular computing (Polkowski, 2002). It includes mechanisms for defining partial memberships of sets, but does not introduce additional measures of probabilities or degrees of membership.

The basic idea is that there is some information or data associated with each object in the universe of discourse. Based on this information, it is possible to tell some of the objects apart, while others are impossible to distinguish. The latter objects are indiscernible from each other, and form a set. Each set of indiscernible objects is a knowledge granule, and they form the building blocks for knowledge about the universe. The rough set community has been a very active research community since its inception in the eighties, and a large number of rough set methods for knowledge discovery and data mining have been developed. The entire knowledge discovery process has been subject to research, and a wide range of contributions has been made.

Data mining technology provides a new thought for organizing and managing tremendous data. Rough set theory is one of the important methods for knowledge discovery. This method can analyze intact data, obtain uncertain knowledge and offer an effective tool by reasoning.

Rough set has shed light on many research areas, but seldom found its way into real world application. Data mining with rough set is a multi-phase process consisted of mainly: discretization; reducts and rules generation on training set; classification on test set.

Rough Set Theory, since it was put forward, has been widely used in Data Mining, and has important functions in the expression, study and conclusion of uncertain knowledge, it is a powerful tool, which sets up the intelligent decision system. The main focus is to show how

rough set techniques can be employed as an approach to the problem of data mining and knowledge extraction.

#### **4. Applications of rough set theory**

Rough set theory offers effective methods that are applicable in many branches of Artificial Intelligence, one of the advantages of rough set theory is that programs implementing its methods may easily run on parallel computers, but several problems remain to be solved.

Recently, much research has been carried out in Rough Set together with other artificial Intelligence methods such as Fuzzy Logic, Neural Networks, and Expert Systems and some significant results have been found. Rough set theory allows characterization of a set of objects in terms of attribute values; finding dependencies total or partial between attributes; reduction of superfluous attributes; finding significance attributes and decision rule generation.

The applications of Rough Set have resolved complex problems; and therefore have been attractive to researchers in recent years, already it has been applied successfully in a number of challenging fields such the a soft computing method.

This section provides a brief overview of some of the many applications of rough set. There are several properties of rough sets that make the theory an obvious choice for use in dealing with real problems:

##### **a. Pattern Recognition**

Pattern Recognition using Rough Set is one such successful application field, but in 2001 A. Mrozek and K. Cyran (2001) proposed a hybrid method of automatic diffraction pattern recognition based on Rough Set Theory and Neural Network. In this new method, the rough set is used to define the objective function and stochastic evolutionary algorithm for space search of a feature extractor, and neural networks are employed to model the uncertain systems. The features obtain end by optimized sampling of diffraction patterns are input to a semantic classifier and the pattern recognition algorithm is performed with optimized and standard computer-generated holograms.

##### **b. Emergency room diagnostic medical**

A common and diagnostically challenging problem facing emergency department personnel in hospitals is that of acute abdominal pain in children. There are many potential causes for this pain, most usually non-serious. However, the pain may be an indicator that a patient has a serious is illness, requiring immediate treatment and possibly surgery. Experienced doctors will use a variety of relevant historical information and physical observations to assess children. Such attributes occur frequently in recognizable patterns, allowing a quick and efficient diagnosis. Inexperienced doctors, on the other hand, may lack the knowledge and information to be able to recognize these patterns. The rough set based clinical decision model is used to assist such inexperienced doctors. In this research, rough sets are used to support diagnosis by distinguishing between three disposition categories: discharge, observation/further investigation, and consultation. Preliminary results show that the system gives accuracy comparable to doctors, though it is dependent on a suitably high data quality (Rubin et al., 1996)

##### **c. Acoustical analysis**

Rough Set was applied for the assessment of concert hall acoustics. Rough set algorithms are applied to the decision table containing subjectively quantified parameters and the results of

overall subjective preference of acoustical objects described by the parameters. Fuzzy membership functions map the test results to approximate the tested parameter distribution, which is determined on the basis of the separate subjective test of individual parameter underlying overall preference. A prototype system based on the rough set theory is used to induce generalized rules that describe the relationship between acoustical parameters of concert halls and sound processing algorithms (Kotek, 1999)

d. Power system security analysis

Rough Set is a systematic approach used to help knowledge engineers during the extraction process of facts and rules of a set of examples for power system operation problems. This approach describes the reduction the number of examples, offering a more compact set of examples to the user (Lambert-Torres et al., 1999).

e. Spatial and meteorological pattern classification

Some categories of sunspot groups are associated with solar flares. Observatories around the world track all visible sunspots in an effort to detect flares early, the sunspot recognition and classification are currently manual and labor intensive processes which could be automated if successfully learned by a machine. The approach employs a hierarchical rough set based learning method for sunspot classification. It attempts to learn the modified Zurich classification scheme through rough set-based decision tree induction. The resulting system has been evaluated on sunspots extracted from satellite images, with promising results (Nguyen et al., 2005).

A new application of rough set theory for classifying meteorological radar data has been introduced. Volumetric radar data is used to detect storm events responsible for severe weather. Classifying storm cells is a difficult problem as they exhibit a complex evolution throughout their lifespan. Also, the high dimensionality and imprecision of the data can be prohibitive. Rough set approach is employed to classify a number of meteorological storm events (Shen & Jensen, 2007).

f. Intelligent control systems

An important application field of rough set theory is that of intelligent control systems especially when incorporated with fuzzy theory (Xie et al., 2004).

g. Measure the quality of a single subset

Ant Colony System algorithm and Rough Set Theory proposed a hybrid approach to feature selection, in Rough Set Theory offers a heuristic function in order to measure the quality of a single subset. It has studied the influence of the setting of the parameters for this problem, in particular for finding a reduct. Experimental results show this hybrid approach is a potential method for features selection (He et al., 2007).

There are infinite possibilities in the development of methods based on Rough Set Theory such as nonstandard analysis, nonparametric statistics and qualitative.

## 5. Case – rough set with tools in dengue diagnosis

In this section, several patients data set is shown with possible dengue symptoms. Through data are analysis is accomplished, using a Rough Set approach for the elimination of redundant data and the development of a set of rules that it can aid the doctor in the elaboration of the diagnosis. Below the Table 3 is shown with the patients data set and respective symptoms, and the data are of the discreet type.

### 5.1 Information system or information table

Patient	Conditional Attributes			Decision Attribute
	blotched_red_skin	muscular_pain_articulations	Temperature	Dengue
P1	No	No	Normal	No
P2	No	No	High	No
P3	No	No	Very High	Yes
P4	No	Yes	High	Yes
P5	No	Yes	Very High	Yes
P6	Yes	Yes	High	Yes
P7	Yes	Yes	Very High	Yes
P8	No	No	High	No
P9	Yes	No	Very High	Yes
P10	Yes	No	High	No
P11	Yes	No	Very High	No
P12	No	Yes	Normal	No
P13	No	Yes	High	Yes
P14	No	Yes	Normal	No
P15	Yes	Yes	Normal	No
P16	Yes	No	Normal	No
P17	Yes	No	High	No
P18	Yes	Yes	Very High	Yes
P19	Yes	No	Normal	No
P20	No	Yes	Normal	No

Table 3. Patients with respective symptoms

Where, B are all of the objects or registrations of the system, given set  $B=\{P1, P2, P3, P4, P5, P6, P7, P8, P9, P10, P11, P12, P13, P14, P15, P16, P17, P18, P19, P20\}$  the set conditional attributes is represented by  $C=\{\text{blotched\_red\_skin, muscular\_pain\_articulations, Temperature}\}$  and the set D represented the decision attribute, where  $D=\{\text{dengue}\}$ . The set A or Table 3, can be shown in relation to the function of nominal values of considered attributes, in the Table 4:

	Attributes	Nominal Values
Conditional Attributes	blotched_red_skin,	Yes, No
	muscular_pain_articulations	Yes, No
	Temperature	Normal, High, Very High
Decision Attributes	Dengue	Yes, No

Table 4. Nominal Values of Attributes

### 5.2 Indiscernibility relation

Indiscernibility Relation is the relation between two objects or more, where all the values are identical in relation to a subset of considered attributes. In Table 3, presented in section 5.1, it can be observed that the set is composed of attributes that are directly related to patients'

symptoms, where  $C = \{\text{blotched\_red\_skin}, \text{muscular\_pain\_articulations}, \text{temperature}\}$ , the indiscernibility relation is given to  $\text{INDA}(C)$ . When Table 3 is broken down it can be seen that indiscernibility relation is given in relationship to conditional attributes:

- The blotched\_red\_skin attribute generates two indiscernibility elementary sets:  $\text{INDA}(\{\text{blotched\_red\_skin}\}) = \{\{P1, P3, P4, P5, P8, P12, P13, P14, P20\}, \{P6, P7, P9, P10, P11, P15, P16, P17, P18, P19\}\}$ .

Patient	Conditional Attributes			Decision Attribute
	blotched_red_skin	muscular_pain_articulations	Temperature	Dengue
P1	No	No	Normal	No
P12	No	Yes	Normal	No
P13	No	Yes	High	Yes
P14	No	Yes	Normal	No
P2	No	Yes	High	No
P20	No	Yes	Normal	No
P3	No	No	Very High	Yes
P4	No	Yes	High	Yes
P5	No	Yes	Very High	Yes
P8	No	No	High	No
P10	Yes	No	High	No
P11	Yes	No	Very High	No
P15	Yes	Yes	Normal	No
P16	Yes	No	Normal	No
P17	Yes	No	High	No
P18	Yes	Yes	Very High	Yes
P19	Yes	No	Normal	No
P6	Yes	Yes	High	Yes
P7	Yes	Yes	Very High	Yes
P9	Yes	No	Very High	Yes

Table 5. Table 3 organize in relations blotched\_red\_skin attribute

- The muscular\_pain\_articulations attribute generates two indiscernibility elementary sets:  $\text{INDA}(\{\text{muscular\_pain\_articulations}\}) = \{\{P1, P2, P3, P8, P9, P10, P11, P16, P17, P19\}, \{P4, P5, P6, P7, P12, P13, P14, P15, P18, P20\}\}$ .

Patient	Conditional Attributes			Decision Attribute
	blotched_red_skin	muscular_pain_articulations	Temperature	Dengue
P1	No	No	Normal	No
P2	No	No	High	No
P3	No	No	Very High	Yes
P8	No	No	High	No
P9	Yes	No	Very High	Yes
P10	Yes	No	High	No

P11	Yes	No	Very High	No
P16	Yes	No	Normal	No
P17	Yes	No	High	no
P19	Yes	No	Normal	No
P4	No	Yes	High	Yes
P5	No	Yes	Very High	Yes
P6	Yes	Yes	High	Yes
P7	Yes	Yes	Very High	Yes
P12	No	Yes	Normal	No
P13	No	Yes	High	Yes
P14	No	Yes	Normal	No
P15	Yes	Yes	Normal	No
P18	Yes	Yes	Very High	Yes
P20	No	Yes	Normal	No

Table 6. Table 3 organize in relation muscular\_pain\_articulations attribute

- The temperature attribute generates three indiscernibility elementary sets:  $INDA(\{temperature\}) = \{P2, P4, P6, P8, P10, P13, P17\}, \{P3, P5, P7, P9, P11, P18\}, \{P1, P12, P14, P15, P16, P19, P20\}$ .

Patient	Conditional Attributes			Decision Attribute
	blotched_red_skin	muscular_pain_articulations	Temperature	Dengue
P13	No	Yes	High	Yes
P2	No	Yes	High	No
P4	No	Yes	High	Yes
P8	No	No	High	No
P10	Yes	No	High	No
P17	Yes	No	High	No
P6	Yes	Yes	High	Yes
P1	No	No	Normal	No
P12	No	Yes	Normal	No
P14	No	Yes	Normal	No
P20	No	Yes	Normal	No
P15	Yes	Yes	Normal	No
P16	Yes	No	Normal	No
P19	Yes	No	Normal	No
P3	No	No	Very High	Yes
P5	No	Yes	Very High	Yes
P11	Yes	No	Very High	No
P18	Yes	Yes	Very High	Yes
P7	Yes	Yes	Very High	Yes
P9	Yes	No	Very High	Yes

Table 7. Table 3 organized in relation temperature attribute

### 5.3 Approximation

The lower and the upper approximations of a set are interior and closure operations in a topology generated by an indiscernibility relation. Below is presented and described the types of approximations are followed using in Rough Set Theory; the approximations concepts are applied in the Table 3, shown to proceed:

Patient	Conditional Attributes			Decision Attribute
	blotched_red_skin	muscular_pain_articulations	Temperature	Dengue
P1	No	No	Normal	No
P2	No	No	High	No
P8	No	No	High	No
P10	Yes	No	High	No
P11	Yes	No	Very High	No
P12	No	Yes	Normal	No
P14	No	Yes	Normal	No
P15	Yes	Yes	Normal	No
P16	Yes	No	Normal	No
P17	Yes	No	High	No
P19	Yes	No	Normal	No
P20	No	Yes	Normal	No
P3	No	No	Very High	Yes
P4	No	Yes	High	Yes
P5	No	Yes	Very High	Yes
P6	Yes	Yes	High	Yes
P7	Yes	Yes	Very High	Yes
P9	Yes	No	Very High	Yes
P13	No	Yes	High	Yes
P18	Yes	Yes	Very High	Yes

Table 8. Table 3 organized in relation decision attribute

- a. Lower Approximation set  $B''$ 
  - Lower Approximation set ( $B''$ ) of the patients that are definitely have dengue are identified as  $B'' = \{P3,P4,P5,P6,P7,P13,P18\}$
  - Lower Approximation set ( $B''$ ) of patients that certain have not dengue are identified as  $B'' = \{P1 ,P2 ,P8 ,P10 ,P12, P14, P15, P16, P17, P19,P20\}$
- b. Upper Approximation set  $B^*$ 
  - Upper Approximation set ( $B^*$ ) of the patients that possibly have dengue are identified as  $B^* = \{P3,P4,P5,P6,P7, P9, P13,P18\}$
  - Upper Approximation set ( $B^*$ ) of the patients that possibly have not dengue are identified as  $B^* = \{P1, P2, P8, P10, P11, P12, P14, P15, P16, P17, P19, P20\}$
- c. Boundary Region (BR)
  - Boundary Region ( $B^*$ ) of the patients that not have dengue are identified as:  $BR = \{P1,P2,P8,P10,P11,P12,P14,P15,P16,P17,P19,P20\} - \{P1,P2,P8,P10,P12,P14,P15,P16,P17, P19,P20\} = \{P11\};$

- Boundary Region ( $B^*$ ), the set of the patients that have dengue are identified as:  $BR = \{P3, P4, P5, P6, P7, P9, P13, P18\} - \{P3, P4, P5, P6, P7, P13, P18\} = \{P9\}$

Observation: Boundary Region (BR), the set constituted by elements P9 and P11, which cannot be classified, since they possess the same characteristics, but with differing conclusions differ in the decision attribute.

#### 5.4 Quality of approximations

The two coefficients of quality of approximation are:

- Imprecision coefficient, using Eq. (1):
  - for the patients with possibility of they are with dengue  $\alpha_B(X) = 7/8$ ;
  - for the patients with possibility of they are not with dengue  $\alpha_B(X) = 8/12$ .
- Quality Coefficient of upper and lower approximation, using Eq. 2 and 3:
  - $\alpha_B(B^*(X)) = 8/20$ , for the patients that have the possibility of they be with dengue;
  - $\alpha_B(B^*(X)) = 11/20$ , for the patients that not have the possibility of they be with dengue;
  - $\alpha_B(B''(X)) = 7/20$ , for the patients that have dengue;
  - $\alpha_B(B''(X)) = 8/20$ , for the patients that not have dengue.

Observations:

1. Patient with dengue:  $\alpha_B(B''(X)) = 7/20$ , that is, 35% of patients certainly with dengue.
2. Patient that don't have dengue:  $\alpha_B(B''(X)) = 11/20$ , that is, approximately 55% of patients certainly don't have dengue.
3. 10% of patients (P9 and P11) cannot be classified neither with dengue nor without dengue, since the characteristics of all attributes are the same, with only the decision attribute (dengue) not being identical and generates an inconclusive diagnosis for dengue.

## 6. Data reduction in information system

The form in which data is presented within an information system must guarantee that the redundancy is avoided as it implicates the minimization of the complexly computational in relation to the creation of rules to aid the extraction knowledge. However, when the information system possesses redundancy situations, it is necessary to treat it One of the ways of accomplishing this is to use the concept of reduct, without altering the indiscernibility relations.

A reduct is a set of necessary minimum data, since the original proprieties of the system or information table are maintained. Therefore, the reduct must have the capacity to classify objects, without altering the form of representing the knowledge.

The process of reduction of information is presented below in Table 3, it can be observed that the data is of a discreet type.

a. Verification inconclusive data

Step 1 – Analysis of data contained in Table 3 shows that possess information inconclusive, being that the values of conditional attributes same and the value of decision attribute is different.

Conclusion of Step 1: The symptoms of patient P9 and patient P11 are both inconclusive, since they possess equal values of conditions attributes together with a value of decision

attribute that is different. Therefore, the data of patient P9 and patient P11 will be excluded from Table 3.

b. Verification of equivalent information

Step 2 - Analysis of data contained in Table 3 shows that it possesses equivalent information.

P2	No	No	High	No
P8	No	No	High	No
P4	No	Yes	High	Yes
P13	No	Yes	High	Yes
P7	Yes	Yes	Very High	Yes
P18	Yes	Yes	Very High	Yes
P10	Yes	No	High	No
P17	Yes	No	High	No
P12	No	Yes	Normal	No
P14	No	Yes	Normal	No
P20	No	Yes	Normal	No
P16	Yes	No	Normal	No
P19	Yes	No	Normal	No

Conclusion of Step 2 - The Table 3 has it reduced data presented in a revised version in Table 9 shown below:

Patient	Conditional Attributes			Decision Attribute
	blotched_red_skin	muscular_pain_articulations	Temperature	Dengue
P1	No	No	Normal	No
P2	No	No	High	No
P3	No	No	Very High	Yes
P4	No	Yes	High	Yes
P5	No	Yes	Very High	Yes
P6	Yes	Yes	High	Yes
P7	Yes	Yes	Very High	Yes
P8	No	No	High	No
P10	Yes	No	High	No
P12	No	Yes	Normal	No
P15	Yes	Yes	Normal	No
P16	Yes	No	Normal	No
P19	Yes	No	Normal	No

Table 9. Reduct of Information of Table 3

Step 3 - Analysis of each condition attributes with the attributes set.

Patient	Conditional Attributes	Decision Attribute
	blotched_red_skin	Dengue
P1	No	No
P2	No	No
P3	No	Yes
P4	No	Yes
P5	No	Yes
P6	Yes	Yes
P7	Yes	Yes
P8	No	No
P10	Yes	No
P12	No	No
P15	Yes	No
P16	Yes	No
P19	Yes	No

Table 10. Analysis of Attribute blotched\_red\_skin in Table 9

Patient	Conditional Attributes	Decision Attribute
	muscular_pain_articulations	Dengue
P1	No	No
P2	No	No
P3	No	Yes
P4	Yes	Yes
P5	Yes	Yes
P6	Yes	Yes
P7	Yes	Yes
P8	No	No
P10	No	No
P12	Yes	No
P15	Yes	No
P16	No	No
P19	No	No

Table 11. Analysis of Attribute muscular\_pain\_articulations in Table 9

Patient	Conditional Attribute	Decision Attribute
	Temperature	Dengue
P1	Normal	No
P2	High	No
P3	Very High	Yes
P4	High	Yes
P5	Very High	Yes
P6	High	Yes

P7	Very High	Yes
P8	High	No
P10	High	No
P12	Normal	No
P15	Normal	No
P16	Normal	No
P19	Normal	No

Table 12. Analysis of Attribute Temperature in Table 9

Conclusion of analysis: In this analysis, no data was excluded.

- c. Given analysis of condition attributes in Table 9, it can be observed that the same data exists in proceeding tables.
- Analysis of attributes blotched\_red\_skin and muscular\_pain\_articulations in Table 9.

Patient	Conditional Attributes		Decision Attribute
	blotched_red_skin	muscular_pain_articulations	Dengue
P1	No	No	No
P2	No	No	No
P3	No	No	Yes
P4	No	Yes	Yes
P5	No	Yes	Yes
P12	No	Yes	No
P6	Yes	Yes	Yes
P7	Yes	Yes	Yes
P8	No	No	No
P16	Yes	No	No
P19	Yes	No	No
P10	Yes	No	No
P15	Yes	Yes	No

Table 13. Analysis of Attributes blotched\_red\_skin and muscular\_pain\_articulations in Table 9

Result of analysis

Patient	Conditional Attributes		Decision Attribute
	blotched_red_skin	muscular_pain_articulations	Dengue
P1	No	No	No
P3	No	No	Yes
P4	No	Yes	Yes
P6	Yes	Yes	Yes
P10	Yes	No	No
P15	Yes	Yes	No

Table 14. Result of Analysis of Attributes blotched\_red\_skin and muscular\_pain\_articulations in Table 9

- Analysis of Attributes attributes blotched\_red\_skin and temperature.

Patient	Conditional Attributes		Decision Attribute
	blotched_red_skin	Temperature	Dengue
P1	No	Normal	No
P12	No	Normal	No
P2	No	High	No
P8	No	High	No
P3	No	Very High	Yes
P5	No	Very High	Yes
P7	Yes	Very High	Yes
P4	No	High	Yes
P6	Yes	High	Yes
P10	Yes	High	No
P15	Yes	Normal	No
P16	Yes	Normal	No
P19	Yes	Normal	No

Table 15. Analysis of Attributes blotched\_red\_skin and temperature in Table 9

Result of analysis

Patient	Conditional Attributes		Decision Attribute
	blotched_red_skin	Temperature	Dengue
P1	No	Normal	No
P2	No	High	No
P3	No	Very High	Yes
P4	No	High	Yes
P6	Yes	High	Yes
P10	Yes	High	No
P15	Yes	Normal	No

Table 16. Result of it Analysis of Attributes blotched\_red\_skin and temperature in Table 9

- Analysis of attributes muscular\_pain\_articulations and temperature in Table 9.

Patient	Conditional Attributes		Decision Attribute
	muscular_pain_articulations	Temperature	Dengue
P1	No	Normal	No
P16	No	Normal	No
P19	No	Normal	No
P2	No	High	No
P8	No	High	No
P10	No	High	No

P3	No	Very High	Yes
P4	Yes	High	Yes
P6	Yes	High	Yes
P5	Yes	Very High	Yes
P7	Yes	Very High	Yes
P12	Yes	Normal	No
P15	Yes	Normal	No

Table 17. Analysis of Attributes muscular\_pain\_articulations and temperature in Table 9  
Result of analysis

Patient	Conditional Attributes		Decision Attribute
	muscular_pain_articulations	Temperature	Dengue
P1	No	Normal	No
P2	No	High	No
P3	No	Very High	Yes
P4	Yes	High	Yes
P5	Yes	Very High	Yes
P12	Yes	Normal	No

Table 18. Result of it analysis of Attributes muscular\_pain\_articulations and temperature in Table 9

Step 4 – Verification of equivalent (intersection) data in the Tables 14, 16 and 18 correspond where data is the element of reduct information in relation to Table 9.

Patient	Conditional Attributes			Decision Attribute
	blotched_red_skin	muscular_pain_articulations	Temperature	Dengue
P1	No	No	Normal	No
P3	No	No	Very High	Yes
P4	No	Yes	High	Yes

Table 19. Table with result of information reduct of Table 9

### 6. Decision rules

With the information reduct shown above, it can be generated the necessary decision rules for aid to the dengue diagnosis. The rules are presented to proceed:

Rule-1

R1: If patient  
 blotched\_red\_skin = No and  
 muscular\_pain\_articulations = No and  
 temperature = Normal  
 Then dengue = No.

**Rule-2**

R2: If patient  
    blotched\_red\_skin = No and  
    muscular\_pain\_articulations = No and  
    temperature = Very High  
Then dengue = Yes.

**Rule-3**

R3: If patient  
    blotched\_red\_skin = No and  
    muscular\_pain\_articulations = Yes and  
    temperature = High  
Then dengue = Yes.

## 7. Conclusion

This study, it has discussed the Rough set theory, was proposed in 1982 by Z. Pawlak, as an approach to knowledge discovery from incomplete, vagueness and uncertain data. The rough set approach to processing of incomplete data is based on the lower and the upper approximation, and the theory is defined as a pair of two crisp sets corresponding to approximations.

The main advantage of rough set theory in data analysis is that it does not need any preliminary or additional information concerning data, such as basic probability assignment in Dempster-Shafer theory, grade of membership or the value of possibility in fuzzy set theory. The Rough Set approach to analysis has many important advantages such as (Pawlak, 1997): Finding hidden patterns in data; Finds minimal sets of data (data reduction); Evaluates significance of data; Generates sets of decision rules from data; Facilitates the interpretation of obtained result

Different problems can be addressed though Rough Set Theory, however during the last few years this formalism has been approached as a tool used with different areas of research. There has been research concerning be relationship between Rough Set Theory and the Dempster-Shafer Theory and between rough sets and fuzzy sets. Rough set theory has also provided the necessary formalism and ideas for the development of some propositional machine learning systems.

Rough set has also been used for knowledge representation; data mining; dealing with imperfect data; reducing knowledge representation and for analyzing attribute dependencies.

Rough set Theory has found many applications such as power system security analysis, medical data, finance, voice recognition and image processing; and one of the research areas that has successfully used Rough Set is the knowledge discovery or Data Mining in database.

## 8. References

- Cerchiarì, S.C.; Teurya, A.; Pinto, J.O.P.; Lambert-Torres, G.; Sauer, L. & Zorzate, E.H. (2006). Data Mining in Distribution Consumer Database using Rough Sets and Self-Organizing Maps, *Proceedings of the 2006 IEEE Power Systems Conference and*

- Exposition*, pp. 38-43, ISBN 1-4244-0177-1, Atlanta-USA, Oct. 29–Nov. 1, 2006, IEEE Press, New Jersey-USA.
- Fayyad, U.; Piatetsky-Shapiro, G. & Smyth, P. (1996a). From Data Mining to Knowledge Discovery: An Overview, In: *Advances in Knowledge Discovery & Data Mining*, Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. & Uthurusamy, R. (Ed.), pp. 1-34. AAAI Press, ISBN 978-0-262-56097-9, Menlo Park-USA.
- Fayyad, U., G. Piatetsky-Shapiro, & P. Smyth. (1996b). Knowledge Discovery and Data Mining: Towards a Unifying Framework, *The Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pp. 82–88, ISBN 978-1-57735-004-0, Portland-USA, Aug. 2–4, 1996, AAAI Press, Menlo Park-USA.
- Geng, Z. & Qunxiong, Z. (2006). A New Rough Set-Based Heuristic Algorithm for Attribute Reduct, *Proceedings of the 6th World Congress on Intelligent Control and Automation*, pp. 3085-3089, ISBN 1-4244-0332-4, Dalian-China, Jun. 21-23, 2006, Wuhan University of Technology Press, Wuhan-China.
- He, Y.; Chen, D.; Zhao, W. (2007). Integration Method of Ant Colony Algorithm and Rough Set Theory for Simultaneous Real Value Attribute Discretization and Attribute Reduction, In: *Swarm Intelligence: Focus on Ant and Particle Swarm Optimization*, Chan, F.T. S. & Tiwari, M.K. (Ed.), pp. 15–36, I-TECH Education and Publishing. ISBN 978-3-902613-09-7, Budapest-Hungary.
- Komorowski, J.; Pawlak, Z.; Polkowski, L. & Skowron, A. (1999). Rough Sets Perspective on Data and Knowledge, In: *The Handbook of Data Mining and Knowledge Discovery*, Klossgrn, W. & Zylkon, J. (Ed.), pp. 134–149, Oxford University Press, ISBN 0-19-511831-6, New York-USA.
- Kostek, B. (1999). Assessment of Concert Hall Acoustics using Rough Set and Fuzzy Set Approach, In: *Rough Fuzzy Hybridization: A New Trend in Decision-Making*, Pal, S. & Skowron, A. (Ed.), pp. 381-396, Springer-Verlag Co., ISBN 981-4021-00-8, Secaucus-USA.
- Lambert-Torres, G.; Rossi, R.; Jardini, J.A.; Alves da Silva, A.P. & Quintana, V.H. (1999). Power System Security Analysis based on Rough Classification, In: *Rough Fuzzy Hybridization: A New Trend in Decision-Making*, Pal, S. & Skowron, A. (Ed.), pp. 263-300, Springer-Verlag Co., ISBN 981-4021-00-8, Secaucus-USA.
- Lin, T. Y. (1997). An Overview of Rough Set Theory from the Point of View of Relational Databases, *Bulletin of International Rough Set Society*, Vol. 1, No. 1, Mar. 1997, pp. 30-34, ISSN 1346-0013.
- Mitchell, T. M. (1999). Machine learning and data mining, *Communications of the ACM*, Vol. 42, No. 11, Nov. 1999, pp. 30-36, ISSN 0001-0782.
- Mrozek, A. & Cyran, K. (2001). Rough Set in Hybrid Methods for Pattern Recognition, *International Journal of Intelligence Systems*, Vol. 16. No. 2, Feb. 2001, pp.149-168, ISSN 0884-8173.
- Nguyen, S.H.; Nguyen, T.T. & Nguyen, H.S. (2005). Rough Set Approach to Sunspot Classification Problem, *Proceedings of the 2005 International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing - Lecture Notes in Artificial Intelligence 3642*, pp. 263–272, ISBN 978-3-540-28653-0, Regina-Canada, Aug. 31-Sept. 3, 2005, Springer, Secaucus-USA.
- Piatetsky-Shapiro, G. & Matheus, C. J. (1995). The Interestingness of Deviations, *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, pp. 23–36,

- ISBN 978-0-929280-82-0, Montreal-Canada, Aug. 1995, AAAI Press, Menlo Park-USA.
- Pawlak, Z. (1982). Rough Sets, *International Journal of Information and Computer Sciences*, Vol. 11, pp. 341- 356, 1982.
- Pawlak, Z. (1991). *Rough Sets: Theoretical Aspects of Reasoning about Data.*, Kluwer Academic Publishers, ISBN 0-79231472, Norwell-USA.
- Pawlak, Z.; Grzymala-Busse, J.; Slowinski, R. & Ziarko, W. (1995) Rough Set, *Communications of the ACM*, Vol. 38, No. 11, Nov. 1995, pp. 88-95, ISSN 0001-0782.
- Pawlak, Z. (1997). Vagueness - a Rough Set View, In: *Lecture Notes in Computer Science- 1261*, Mycielski, J; Rozenberg, G. & Salomaa, A. (Ed.), pp. 106-117, Springer, ISBN 3-540-63246-8, Secaucus-USA.
- Pawlak, Z. (1998). Granularity of Knowledge, Indiscernibility and Rough Sets, *The 1998 IEEE International Conference on Fuzzy Systems Proceedings - IEEE World Congress on Computational Intelligence*, pp. 106-110, ISBN 0-7803-4863-X, May 4-9, 1998, Anchorage-USA, IEEE Press, New Jersey-USA.
- Rubin, S.; Michalowski, W. & Slowinski, R. (1996). Developing an Emergency Room Diagnostic Check List using Rough Sets: A Case Study of Appendicitis, In: *Simulation in the Medical Sciences*, Anderson, J. & Katzper, M. (Ed.), pp. 19-24, The Society for Computer Simulation Press, San Diego-USA.
- Shen, Q. & Jensen, R. (2007). Rough Sets, Their Extensions and Applications, *International Journal of Automation and Computing*, Vol. 4, No. 3, Jul. 2007, pp. 217-228, ISSN 1476-8186.
- Stoll, R.R. (1979). *Set Theory and Logic*, Dover Publications, ISBN 0-486-63829-4, Mineola-USA.
- Xie, G.; Wang, F. & Xie, K. (2004). RST-Based System Design of Hybrid Intelligent Control, *Proceedings of the 2004 IEEE International Conference on Systems, Man and Cybernetics*, pp. 5800-5805, ISBN 0-7803-8566-7, The Hague-The Netherlands, Oct. 10-13, 2004, IEEE Press, New Jersey-USA.
- Zadeh, L. A. (1965). Fuzzy Sets, *Information and Control*, No. 8, pp. 338-353.
- Ziarko, W. & Shan, N. (1995). Discovering Attribute Relationships, Dependencies and Rules by using Rough Sets, *Proceedings of the 28th Annual Hawaii International Conference on System Sciences*, pp. 293-299, ISBN 0-8186-6930-6, Wailea-USA, Jan. 3-6, 1995, IEEE Press, New Jersey-USA.
- Wei, J. M. (2003). Rough Set based Approach to Selection of Node, *International Journal of Computational Cognition*, Vol. 1, No. 2, pp. 25-40, ISSN 1542-8060.
- Wu, C.; Yue, Y.; Li, M. & Adjei, O. (2004). The Rough Set Theory and Applications, *Engineering Computations*, Vol. 21, No. 5, pp.488-511, ISSN 0264-4401.

# Robust Data Mining: An Integrated Approach

Sangmun Shin, Le Yang, Kyungjin Park and Yongsun Choi  
*Department of Systems Management & Engineering, Inje University*  
Korea

## 1. Introduction

The continuous improvement and application of information system technologies have become widely recognized by the industry as critical for maintaining a competitive advantage in the marketplace (Shin et al., 2006). It is also recognized that improvement and application activities are the most efficient and cost-effective when implemented during an early process/product design stage. Data mining (DM) has emerged as one of the key features of many applications in computer science. Often used as a means for predicting the future directions and extracting hidden limitations and specifications of a product/process, DM involves the use of data analysis (DA) tools to discover previously unknown and valid patterns and relationships from a large database. Most DM methods for factor selection reported in literature may yield a number of factors associated with interesting response factors without providing detailed information, such as relationships between the input factor and response, statistical inferences, and analyses (Yang et al., 2007; Witten & Frank, 2005). Based on this, Gardner and Bieker (Gardner & Bieker, 2000) suggested an alternative DA approach toward resolving semiconductor manufacturing problems in order to determine the significant factors. Furthermore, Su et al. (Su et al., 2005) developed an integrated procedure combining a DM method and Taguchi methods.

DA is a term coined to describe the process of sifting through large databases for discovering interesting patterns and relationships. This field spans several disciplines such as databases, machine learning, intelligent information systems, statistics, and expert systems. Two approaches that enable the application of standard machine learning algorithms to large databases are factor selection and sampling. Factor selection is known to be an effective method for reducing dimensionality, removing irrelevant and redundant data, increasing mining accuracy, and improving result comprehensibility (Yu & Liu, 2003). Consequently, factor selection has been a fertile field for research and development since the 1970s and proven to be efficient in removing irrelevant and redundant features, increasing efficiency in mining tasks, improving mining performance like predictive accuracy, and enhancing comprehensibility of the learned results. The factor selection algorithm performs a search through the space of feature subsets (Allen, 1974). In general, two categories of the algorithm have been proposed to resolve the factor selection problem. The first category is based on a filter approach that is independent of the learning algorithms and serves as a filter to sieve out the irrelevant factors. The second category is based on a wrapper approach, which uses an induction algorithm itself as part of the function evaluating the factor subset (Langley, 1994). Since most of the filter methods are based on a heuristic

algorithm for general characteristics of the data rather than a learning algorithm that evaluates the merits of the factor subsets as done by wrapper methods, filter methods are generally much faster and have more practical capabilities to utilize high dimensionality than wrapper methods.

While a large number of factors are considered, there are three important issues when handling data analysis problems, namely, missing values, outliers, and noise factors. The results from DA that uses a number of data sets including many outliers may often be misleading. An outlier is an observation that lies outside of the overall pattern of a distribution (Bakar et al., 2006). Missing values can often seriously affect the data analysis results if a large number of factors and their associated missing values are ignored. Next, in many scientific and engineering fields, there are a number of data sets that are uncontrollable and difficult to handle, since the nature of the measurement of a performance variable may often be a destructive or very expensive characteristic, which is known as the noise factor (Yang et al., 2007).

Existing studies in DM mostly focus on finding patterns in large data sets and further using them for organizational decision making (Yang et al., 2007). DM methods also may not discuss the robustness of solutions, either by considering data pre-processes for outliers and missing values or by considering uncontrollable noise factors.

In order to address this limitation, we have developed an enhanced DA method incorporating the robust design (RD) principle. Among the process/product design methods currently studied in the science and engineering community, researchers often identify RD as one of the most effective methodologies for process/product improvement. Because of their practicability in reducing the inherent uncertainty associated with input factors and process performance, the widespread applications of RD techniques have resulted in significant improvements in process quality, manufacturability, and reliability at low cost. However, most RD methods reported in the literature may obtain the most favorable solution for a small number of given input control factors without considering the reduction in dimensionality for large databases. Although traditional RD methods consider the selection of potential significant factors when they confront a data set including many factors with an interesting response factor, the process is frequently far from the objective as individual egos because the selection process is based on drawing insight from a number of readily available sources relying on the practitioners' opinion and their experience.

For this reason, we propose an integrated approach called robust data mining (RDM), which can reduce the dimensionality of large data sets, may provide detailed statistical relationships among the factors, and robust factor settings, as shown in Fig. 1. This RDM approach has neither been adequately addressed in the literature nor properly applied in industrial processes. As a result, the primary objective of this paper is three-fold. First, the proposed RDM applies outlier test and expectation maximum (EM) algorithm to carry out the data pre-process. Then, the proposed RDM reduces the dimensionality to find the significant factors among a large number of input factors using correlation-based feature selection (CBFS) method and best first search (BFS) algorithm. These methods can evaluate the worth of a subset including the input factors by considering the individual predictive ability of each factor along with the degree of redundancy between the pairs of input factors. This method is far more effective than any other method when a large number of input factors are considered in a process design procedure. Finally, the proposed model utilizes the theory of robust design to handle noise factors using the concept of surrogate variables and response surface methodology (RSM). Our numerical example clearly shows

that the proposed RDM method can efficiently find significant factors and optimal settings by reducing the dimensionality.

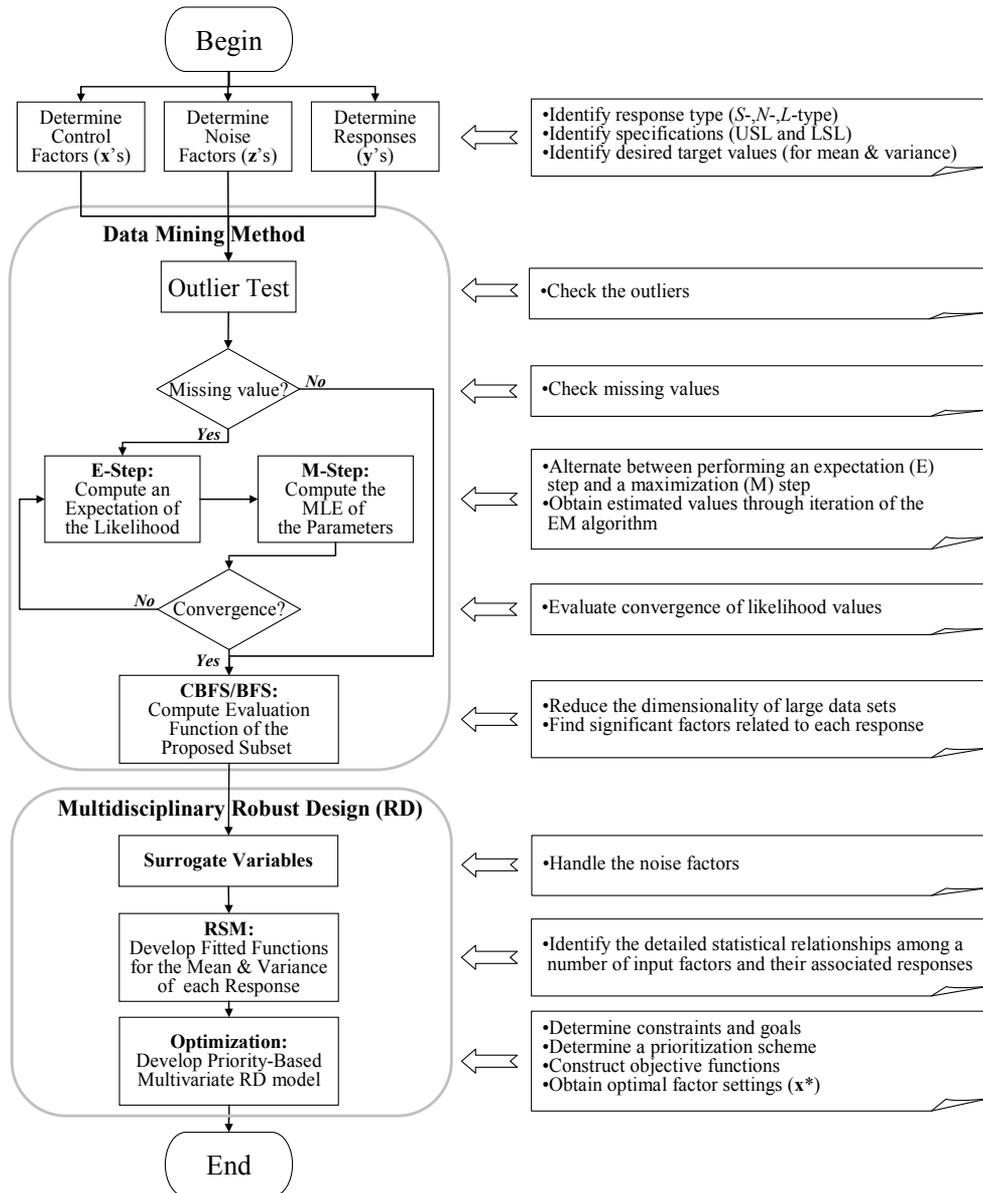


Fig. 1. Overview of the RDM model

## 2. Stage I: data mining method

### 2.1 Data pre-process

The issues of outliers and missing values are the two most important problems in the data pre-process procedure. As shown in Fig. 2, the proposed procedure conducts outlier tests to

detect the outliers in a large number of data sets. If the results of the outlier tests include a number of unusual observations, these outliers are deleted and regarded as missing values. To address the missing values, the EM algorithm is utilized.

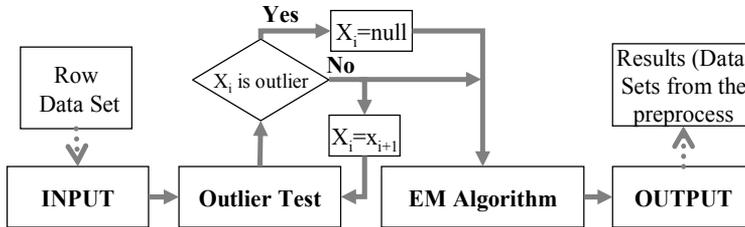


Fig. 2. Proposed data pre-process procedure

### 2.1.1 Outliers test in data mining

Recently, a number of studies have been conducted on an outlier test for large datasets, which can be categorized into (1) the statistical approach, (2) distance-based approach, and (3) deviation-based approach (Bakar et al., 2006).

The statistical approach to outlier detection assumes a distribution or probability model for the given data set and then identifies the outliers with respect to the model using a discordancy test (Witten & Frank, 2005). One of the drawbacks of the statistical approach is the requirement of knowledge about the parameters of the data set, such as data distribution (Bakar et al., 2006). However, the distance-based approach is based on two parameters that are given in advance using the knowledge about the data or may be changed during the iterations to select the most representative outliers. Deviation-based methods identify the outliers by examining the main characteristics of the objects in a group. Objects that “deviate” from this description are considered outliers. Hence, in this approach, the term deviation is typically used to refer to outliers (Witten & Frank, 2005).

### 2.1.2 Expectation Maximization (EM) algorithm

The EM algorithm is used in statistics for finding the maximum likelihood estimates of parameters in probabilistic models, where the model depends on the unobserved latent variables (Pernkopf, 2005). The EM alternates between performing an expectation (E) step, which computes the expectation of the likelihood by including the latent variables as if they were being observed, and a maximization (M) step, which computes the maximum likelihood estimates of the parameters by maximizing the expected likelihood found in the E step. The parameters found in the M step are then used to begin another E step, and the process is repeated.

Let  $Y$  denote the random vector corresponding to the observed data  $y$ , having a probability density function of  $g(y; \psi)$ , where  $\psi = (\psi_1, \dots, \psi_d)^T$  is a vector of unknown parameters within the parameter space  $\Omega$ . The observed data vector  $y$  is viewed as being incomplete and is regarded as an observable function of the complete data. The notion of incomplete data includes the conventional sense of missing data. Let  $x$  denote the vector containing the augmented or complete data. Let  $g_c(x; \psi)$  denote the probability density function of the random vector  $X$  corresponding to the complete-data vector  $x$ . Then, the complete-data log-likelihood function that could be formed for  $\psi$  if  $x$  were fully observable is given by

$$\log L_c(\boldsymbol{\psi}) = \log g_c(\mathbf{x}; \boldsymbol{\psi}) \quad (1)$$

Formally, we have two sample spaces  $\alpha$  and  $\beta$  and many-to-one mapping from  $\alpha$  to  $\beta$ . Instead of observing the complete-data vector  $\mathbf{x}$  in  $\alpha$ , we observe the incomplete-data vector  $\mathbf{y} = \mathbf{y}(\mathbf{x})$  in  $\beta$ . It follows that

$$g(\mathbf{y}; \boldsymbol{\psi}) = \int_{\alpha(\mathbf{y})} g_c(\mathbf{x}; \boldsymbol{\psi}) d\mathbf{x}, \quad (2)$$

where  $\alpha(\mathbf{y})$  is the subset of  $\alpha$  determined from the equation  $\mathbf{y} = \mathbf{y}(\mathbf{x})$ .

The EM algorithm approaches the problem of solving the incomplete-data likelihood function indirectly by proceeding iteratively in terms of the complete-data log likelihood function  $\log L_c(\boldsymbol{\psi})$ . As this function is unobservable, it is replaced by its conditional expectation given  $\mathbf{y}$  by using the current fit for  $\boldsymbol{\psi}$ . On the  $(k+1)$ -th iteration, the E and M steps are defined as follows (McLachlan, 1996):

E-step. Calculate  $Q(\boldsymbol{\psi}; \boldsymbol{\psi}^{(k)})$ ,

where  $Q(\boldsymbol{\psi}; \boldsymbol{\psi}^{(k)}) = E_{\boldsymbol{\psi}^{(k)}} \{ \log L_c(\boldsymbol{\psi}) | \mathbf{y} \}$ .

M-step. Choose  $\boldsymbol{\psi}^{(k+1)}$  to be any value of  $\boldsymbol{\psi} \in \Omega$  that maximizes  $Q(\boldsymbol{\psi}; \boldsymbol{\psi}^{(k)})$ ; that is,

$Q(\boldsymbol{\psi}^{(k+1)}; \boldsymbol{\psi}^{(k)}) \geq Q(\boldsymbol{\psi}; \boldsymbol{\psi}^{(k)})$  for all  $\boldsymbol{\psi} \in \Omega$ .

The E and M steps are alternated repeatedly until the difference  $L(\boldsymbol{\psi}^{(k+1)}) - L(\boldsymbol{\psi}^{(k)})$  changes by an arbitrarily small amount in the case of convergence of the sequence of likelihood values  $\{L(\boldsymbol{\psi}^{(k)})\}$ . An overview of the EM algorithm is shown in Fig. 3.

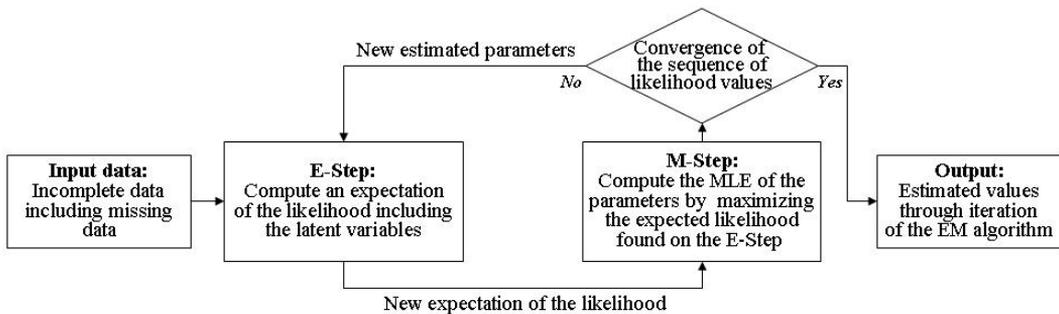


Fig. 3. Overview of the EM algorithm used in data pre-processing

## 2.2 Data mining procedure

### 2.2.1 Correlation-Based Feature Selection (CBFS) method

CBFS is a filter algorithm that ranks the subsets of the input features according to a correlation-based heuristic evaluation function. The bias of the evaluation function is toward the subsets that contain a number of input factors, which are not only highly correlated with a specified response but also uncorrelated with each other (Xu et al., 2004). Among the input factors, irrelevant factors should be ignored because they may have low correlation with the given response. Although some selected factors are highly correlated with the specified

response, redundant factors need to be screened out because they are also highly correlated with one or more of these selected factors. The acceptance of a factor depends on the extent to which it predicts the response in areas of the instance space not already predicted by other factors. The evaluation function of the proposed subset is

$$EV_s = \frac{n\bar{\rho}_{FR}}{\sqrt{n+n(n-1)\bar{\rho}_{FF}}} \quad (3)$$

where  $EV$ ,  $\bar{\rho}_{FR}$ , and  $\bar{\rho}_{FF}$  represent the heuristic evaluation value of a factor subset  $S$  containing  $n$  factors, mean of the factor-response correlation ( $F \in S$ ), and mean of the factor-factor inter-correlation, respectively. Further,  $\sqrt{n+n(n-1)\bar{\rho}_{FF}}$  and  $n\bar{\rho}_{FR}$  indicate the prediction of the response based on a set of factors and redundancy among the factors, respectively. In order to measure the correlation between two factors or a factor and response, an evaluation of a criterion called symmetrical uncertainty is conducted (Hall, 1998).

The symmetrical measure represents that the amount of information gained about  $Y$  after observing  $X$  is equal to the amount of information gained about  $X$  after observing  $Y$ . Symmetry is a desirable property for a measure of the factor-factor inter-correlation or factor-response correlation. Unfortunately, information gain is not apt for factors with more values. In addition,  $\bar{\rho}_{FR}$  and  $\bar{\rho}_{FF}$  should be normalized to ensure they are comparable and have the same effect. Symmetrical uncertainty can minimize the bias in information gain toward features with more values and normalize its value within the range  $[0, 1]$ . The coefficient of symmetrical uncertainty can be calculated as

$$C_{SU} = 2.0 * \left[ \frac{gain}{H(Y) + H(X)} \right] \quad (4)$$

where

$$H(Y) = -\sum_{y \in Y} P(y) \log_2(P(y))$$

$$H(Y | X) = -\sum_{x \in X} p(x) \sum_{y \in Y} p(y | x) \log_2(p(y | x))$$

$$gain = H(Y) - H(Y | X) = H(X) - H(X | Y) = H(Y) + H(X) - H(X, Y)$$

and where  $H(Y)$ ,  $p(y)$ ,  $H(Y | X)$ , and  $gain$  represent the entropy of the specified response  $Y$ , probability of  $y$  value, conditional entropy of  $Y$  given  $X$ , and information gain—a symmetrical measure that reflects additional information about  $Y$  given  $X$ , respectively.

### 2.2.2 Best First Search (BFS) algorithm

In many literatures, finding the best subset is seldom achieved in many industrial situations when using an exhaustive enumeration method. In order to reduce the search spaces for evaluating the number of subsets, one of the most effective methods is the BFS method—a heuristic search method that implements the CBFS algorithm (Langley, 1994). This method

is based on an advanced search strategy that allows backtracking along a search space path. If the path being explored begins to look less promising, the BFS algorithm can backtrack to a more promising previous subset and continue searching from there. The procedure for using the proposed BFS algorithm is given below:

- Step 1. Begin with the OPEN list containing the start state, CLOSE list empty, and BEST  $\leftarrow$  start state (put the start state to BEST).
- Step 2. Let a subset  $\theta = \arg \max \text{EVS}(\text{subset})$ , (get the state from OPEN with the highest evaluation EVS).
- Step 3. Remove  $s$  from OPEN and add to CLOSE.
- Step 4. If  $\text{EVS}(\theta) \geq \text{EVS}(\text{BEST})$ , then  $\text{BEST} \leftarrow \theta$  (put  $\theta$  to BEST).
- Step 5. For each next subset  $\xi$  of  $\theta$  that is not in the OPEN or CLOSE list, evaluate and add to OPEN.
- Step 6. If BEST changed in the last set of expansions, go to step 2.
- Step 7. Return BEST.

The evaluation function given in equation (3) is a fundamental element of CBFS that imposes a specific ranking on the factor subsets in the search spaces. In most cases, enumerating all the possible factor subsets is extremely time-consuming. In order to reduce the computational complexity, the BFS method is utilized to find the best subset. The BFS method can start with either no factor or all the factors. The former search process moves forward through the search space adding a single factor into the result, and the latter search process moves backward through the search space deleting a single factor from the result. To prevent the BFS method from exploring the entire search space, a stopping criterion is imposed. The search process may terminate if five consecutive fully expanded subsets show no improvement over the current best subset. The overview of the CBFS and BFS methods is shown in Fig. 4.

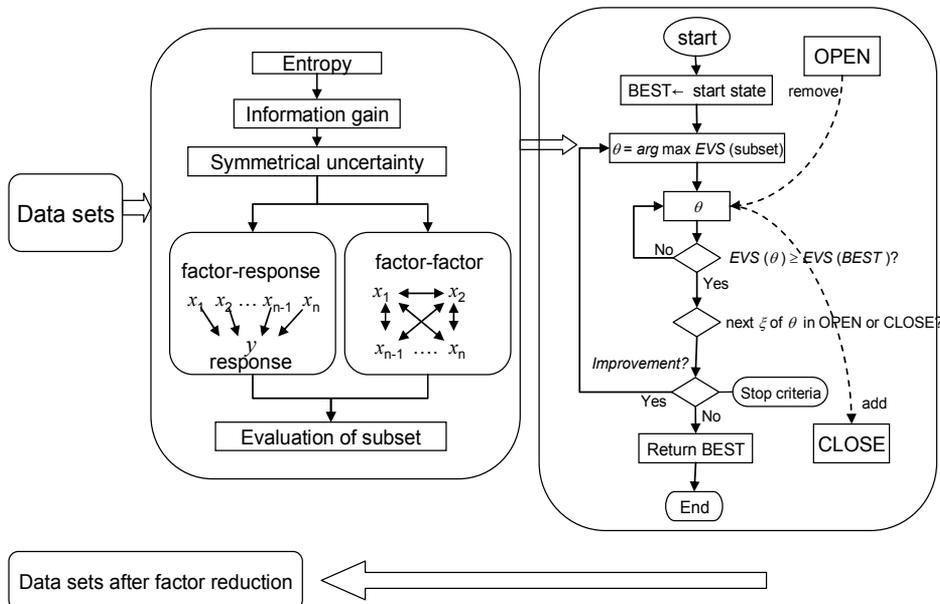


Fig. 4. Overview of the DM method

Data sets after factor reduction ←

### 3. Stage II: robust design

#### 3.1 Surrogate variables

The surrogate variable technique is a subbranch of the screening inspection method. Generally, the nature of measurements or observations on a response (i.e., a dependent variable) may be exceptionally expensive, destructive, or difficult to obtain, forcing a reduction in the overall sample size used to fit the model. To avoid these without dramatically increasing the cost of the experiment, one may use cheaper or more easily collected “surrogate” variables to supplement the expensive input factors.

In our approach, the noise factors of significant factors—both in response and in input—are referred to the destructive or very expensive performance variables to be measured. By using CBFS, we can easily find the candidate surrogate variables from the redundancy factors for every noise factor. Fig. 5 shows two cases of surrogate ( $k, i, n, m \in \text{int}$ ). One is when one of the interesting responses  $y_n$  exhibits the characteristics of noise, while another interesting response  $y_m$  is not only highly correlated to the noise one but also controllable; the surrogate between  $y_n$  and  $y_m$  can be considered. Another is when we focus on a specific interesting response  $y_k$  corresponding to some input factors ( $x_k, x_i, \dots, x_n$ ), where factor  $x_i$  is noise; however, factor  $x_1$  is neither noise and irrelevant to  $x_i$  nor corresponding to the interesting response  $y_k$ . Then,  $x_1$  will be the available surrogate variable candidate for  $x_i$ .

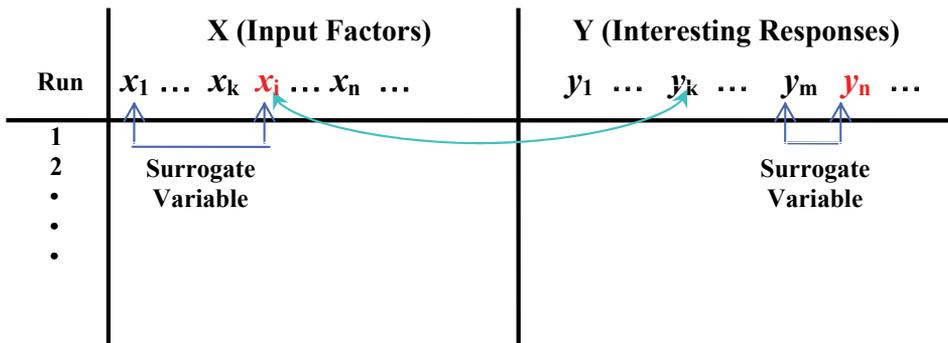


Fig. 5. Concepts of surrogate variables for input factors and responses

#### 3.2 Response Surface Methodology (RSM)

RSM is a statistical tool that is useful for modeling and analyses in situations where the response of interest is affected by several factors. RSM is typically used to optimize the response by estimating an input-response functional form when the exact functional relationship is unknown or is very complicated. For a comprehensive presentation of RSM, Box et al. (Box et al., 1998) and Shin and Cho (Shin & Cho, 2005) provided insightful comments on the current status and future direction of RSM.

In many industrial situations, a manufacturing or service process often contains both control and noise factors that cannot be handled (Montgomery, 2001). Supposing that there are  $k$  controllable variables  $\mathbf{x} = [x_1, x_2, \dots, x_k]$  and  $r$  noise variables  $\mathbf{z} = [z_1, z_2, \dots, z_r]$ , the response model incorporating both control and noise factors can be given by

$$y(\mathbf{x}, \mathbf{z}) = f(\mathbf{x}) + h(\mathbf{x}, \mathbf{z}) + \varepsilon \quad (5)$$

where  $f(\mathbf{x})$ ,  $h(\mathbf{x}, \mathbf{z})$ , and  $\varepsilon$  denote the portion of the model that involves only the control factors, term involving the main effects of the noise factors and the interactions between the control and noise factors, and random error assumed to be normally distributed with zero mean and certain variance, respectively. The detailed calculation of  $h(\mathbf{x}, \mathbf{z})$  is

$$h(\mathbf{x}, \mathbf{z}) = \sum_{i=1}^r \gamma_i z_i + \sum_{i=1}^k \sum_{j=1}^r \delta_{ij} x_i z_j \quad (6)$$

where  $\gamma_i$  and  $\delta_{ij}$  are the coefficients of noise factors and interactions between the control and noise factors, respectively. Denoting the variance of the noise variables as  $\sigma_z^2$  and assuming that the noise variables and random errors  $\varepsilon$  have zero covariance, the mean response model by taking the expectation of the response model in equation (5) can be derived as follows:

$$E_z[y(\mathbf{x}, \mathbf{z})] = \hat{\mu}(\mathbf{x}) = f(\mathbf{x}) \quad (7)$$

By using Taylor series expansion, the variance model for the response can be simplified as follows:

$$Var_z[y(\mathbf{x}, \mathbf{z})] = \hat{\sigma}(\mathbf{x}) = \sigma_z^2 \sum_{i=1}^r \left( \frac{\partial y(\mathbf{x}, \mathbf{z})}{\partial z_i} \right)^2 + \sigma^2 \quad (8)$$

where  $\sigma^2$  is the mean-square error on the analysis of variance (ANOVA).

### 3.3 Robust Desirability Function (RDF) model

The quality of pharmaceutical products is often judged on multiple responses that are not of the same type. Pharmaceutical quality characteristics typically have one of the three possible goals and are therefore categorized as follows:

1. Smaller-the-better (S type): Minimize the quality characteristic of interest.
2. Nominal-the-better (N type): The quality characteristic of interest has a specific target value.
3. Larger-the-better (L type): Maximize the quality characteristic of interest.

Hence, a special multi-objective optimization model is required. It must be able to handle all the three types of quality characteristics simultaneously and consider robustness to reduce both process bias and variability. To address these issues, we propose a RDF model that can resolve the design problems involving multiple responses of several different types by considering the effect of noise factors. Our proposed model integrates the desire function (DF) that involves a popular approach to formulate and resolve the problem as a multi-objective optimization problem into the mean-squared error (MSE) approach, yielding robust solutions by considering a tradeoff between the process mean and variability. Detailed descriptions on desirability function and MSE model can be found in (Myers, 2002) and (Cho, 1994).

Let S, N, and L represent the indexes of the S-, N-, and L-type quality characteristics, respectively. For  $MSE_{kS}$ -related S-type characteristics, the maximum allowable value ( $MSE_{kS}^{\max}$ ) is specified, while for  $MSE_{kL}$ -related L-type quality characteristics, the maximum

allowable value ( $MSE_{kL}^{\max}$ ) is specified. It is noted that the maximum value ( $MSE_{kN}^{\max}$ ) needs to be specified for  $MSE_{kN}$ -related N-type quality characteristics. Suppose we denote the lower and upper bounds for the control factors as  $\tilde{x}_i, \hat{x}_i$ , respectively, and represent the maximum and minimum allowable values for the S-, N-, and L-type quality characteristics as  $\tilde{y}_{k(S,N,L)}, \hat{y}_{k(S,N,L)}$ , respectively. Denoting the target values and weights for desirability of the k-th S-, N-, and L-type characteristics by  $\tau_{k(S, N, L)}$  and  $w_{k(S, N, L)}$ , respectively, we propose the following RDFs:

$$\text{Maximize } D = \left[ \prod_{k=1}^n d_{kt} \right]^{1/k} \quad \text{for } t = S-, N- \text{ and } L- \text{ type} \quad (9)$$

$$\text{where } d_{kt} = \begin{cases} 1 & \text{if } MSE_{kS}(\mathbf{x}) \leq \tau_{kS}, MSE_{kN}(\mathbf{x}) \leq \tau_{kN}, \\ & MSE_{kL}(\mathbf{x}) \leq \tau_{kL} \\ \left( \frac{MSE_{kS}^{\max} - MSE_{kS}(\mathbf{x})}{MSE_{kS}^{\max} - \tau_{kS}} \right)^{w_{kS}} & \text{if } \tau_{kS} \leq MSE_{kS}(\mathbf{x}) \leq MSE_{kS}^{\max} \\ & \text{for } k = 1, 2, \dots, l \\ \left( \frac{MSE_{kN}^{\max} - MSE_{kN}(\mathbf{x})}{MSE_{kN}^{\max} - \tau_{kN}} \right)^{w_{kN}} & \text{if } \tau_{kN} \leq MSE_{kN}(\mathbf{x}) \leq MSE_{kN}^{\max} \\ & \text{for } k = l + 1, \dots, m \\ \left( \frac{MSE_{kL}^{\max} - MSE_{kL}(\mathbf{x})}{MSE_{kL}^{\max} - \tau_{kL}} \right)^{w_{kL}} & \text{if } \tau_{kL} \leq MSE_{kL}(\mathbf{x}) \leq MSE_{kL}^{\max} \\ & \text{for } k = m + 1, \dots, n \\ 0 & \text{if } MSE_{kS}(\mathbf{x}) \geq MSE_{kS}^{\max}, \\ & MSE_{kN}(\mathbf{x}) \geq MSE_{kN}^{\max}, \\ & MSE_{kL}(\mathbf{x}) \geq MSE_{kL}^{\max} \end{cases}$$

Constraints  $\tilde{x}_i \leq x_i \leq \hat{x}_i$  and  $\tilde{y}_{kt} \leq y_{kt} \leq \hat{y}_{kt}$  for  $i = 1, 2, \dots, h$

Note that the objective function D, called the RDF, uses the geometric mean of the individual desirability. It is possible to design a function where the values exceeding the threshold, but still rather less than the target, are only slightly penalized by choosing  $0 < w < 1$ ; higher w values are assigned when you need penalize even further. This allows optimization to take into account the relative importance of each quality characteristic or response, while selecting the most appropriate form of the partial desirability function.

#### 4. Numerical example

To effectively demonstrate the implementation of our proposed methodology, actual case studies of processes that produce a placebo tablet have been conducted in which a number of design variables were considered. The data used in this numerical example is obtained from a continuous real-time tablet manufacturing process. The tablet manufacturing process is classified into three stages, namely, flow, compression, and ejection. In the first step,

granules are fed to be compressed into tablets; at the compression stage, granules are compressed into tablets. At the ejection stage, the tablets are ejected. The objective of this study is to commonly optimize each desired bias and variability value of three tablet quality characteristics including friability ( $y_1$ ), hardness ( $y_2$ ), and disintegration ( $y_3$ ). Then, based on prior information about the system under investigation, it logically follows that the first pressure ( $x_1$ ) to remove air in the granules, second pressure ( $x_2$ ) to produce tablets, first dwell time ( $x_3$ ) to remove air in the granules, second dwell time ( $x_4$ ) to produce tablets, speed to remove the first punch ( $x_5$ ), speed to remove the second punch ( $x_6$ ), speed to eject tablets ( $x_7$ ), amount of overfill ( $x_8$ ), amount of dust ( $x_9$ ), and particle size ( $x_{10}$ ) are the control factors and humidity ( $z_1$ ) and temperature ( $z_2$ ) are the noise factors considered in this study. Friability refers to the brittleness of a tablet and it is measured as the percentage of material lost as it passes through a motorized rotary drum. The effect of the motorized rotary drum allows researchers to predict how the tablets will withstand packaging and transportation. Hardness is an important quality characteristic because it is a major concern in tablet manufacturing. A soft tablet will cause problems during compression and a hard tablet can damage teeth. Lastly, disintegration refers to the time (in minutes) that is required for a tablet to dissolve in a suitable liquid at 37°C and is an estimator of how effectively the tablet will release its ingredients within the body. Hardness is measured by applying a uniform force (measured in Newtons) on the tablet until it breaks. In this particular case, the quality characteristics of interest have conflicting objectives, as shown in Table 1. In order to satisfy the goals of all the three quality characteristics, the goal programming approach is used to establish that the hardness objective is the most important and the friability objective is the least important. Table 2 shows the data from the tablet manufacturing process. The data set is an incomplete-data set and each of the five factors —  $x_1, x_3, x_5, x_7,$  and  $x_9$  — have a missing value.

Quality characteristic	Units	Imp. Rating	Goal	Type	Lower limit	Upper limit
Friability ( $F$ )	%	3	Minimize	S-type	0.4	10
Hardness ( $H$ )	N	1	Target Value (50)	N-type	20	80
Disintegration ( $D$ )	Min	2	Maximize	L-type	0.5	10

Table 1. Quality characteristics of friability ( $F$ ), hardness ( $H$ ), and disintegration ( $D$ )

No.	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$	$x_7$	$x_8$	$x_9$	$x_{10}$	$z_1$	$z_2$	$y_1$	$y_2$	$y_3$
1	4.57	29.69	0.77	2.74	0.27	0.29	0.34	13.23	0.13	134.46	52.08	20.82	4.14	38.79	6.54
2	4.48	38.51	0.75	3.82	0.26	0.47	0.48	21.33	0.12	139.32	41.28	27.42	7.92	67.24	13.19
3		30.58	0.65	4.51	0.23	0.31	0.56	14.04	0.13	138.78	57.60	21.12	4.68	40.71	8.23
4	4.66	32.34	0.78	4.47	0.27	0.53	0.56	23.76		140.67	49.44	22.14	6.66	63.55	11.28
5	3.77	34.99	0.63	3.90	0.22	0.28	0.49	12.42	0.15	138.24	52.32	24.06	8.55	82.50	14.17
6	3.95	29.84		4.11	0.23	0.46	0.51	20.79	0.13	135.54	61.68	16.74	3.72	33.62	6.74
7	4.48	42.19	0.75	3.63	0.26	0.51	0.45	22.95	0.14	132.30	63.36	23.52	8.58	82.48	14.17
8	4.40	27.64	0.74	4.60	0.26	0.31		14.04	0.12	138.78	53.28	21.48	6.54	57.39	11.07
9	5.02	28.37	0.84	3.59		0.41	0.45	18.36	0.13	131.76	51.60	21.42	4.95	46.52	8.59
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
100	5.02	32.78	0.84	3.02	0.30	0.32	0.38	14.31	0.15	148.23	57.36	22.98	8.12	47.03	4.56

Table 2. Data set for the tablet manufacturing process

#### 4.1 Stage I: EM algorithm for data pre-process

DA using the EM algorithm provides the results of the data pre-treatment for the missing values in order to conduct the DM procedure. Table 3 lists the estimated mean, standard deviation, and value after performing 25 iterations of the EM algorithm using the SPSS software package. Consequently, reasonable values based on the estimated mean, standard deviation, and covariance parameters for the five missing values among the twelve factors are found to be  $(x_1, x_3, x_5, x_7, x_9) = (3.86, 0.66, 0.30, 0.58, 0.14)$ .

Significant factors	Estimated mean	Estimated standard deviation	Estimated values
$x_1$	4.44	0.35	3.86
$x_3$	0.74	0.06	0.66
$x_5$	0.26	0.02	0.30
$x_7$	0.49	0.10	0.58
$x_9$	0.13	0.01	0.14

Table 3. Estimated mean, standard deviation, and value after performing 25 iterations of the EM algorithm

#### 4.2 Stage II: DM for dimensionality reduction

The CBFS method, a DM technique, was used to seek the highly correlated factors associated with interesting responses (i.e., friability, hardness, and disintegration) by reducing the dimensionality related to a large number of factors and removing irrelevant and redundant data. As shown in Table 4, data mining results obtained using the Weka software package indicate that two uncorrelated factors (i.e.,  $x_2$  and  $z_2$ ), four uncorrelated factors (i.e.,  $x_1, x_2, x_9$ , and  $z_2$ ), and three uncorrelated factors (i.e.,  $x_2, x_9$ , and  $z_2$ ) are significant for  $y_1, y_2$ , and  $y_3$ , respectively. Among these solutions, the temperature ( $z_2$ ) often cannot be controlled in the tablet manufacturing process. Consequently, we consider  $z_2$  as the noise factor, and consider the others input factors (i.e.,  $x_1, x_2$ , and  $x_9$ ) among the DM results as the control factors.

DM	Responses		
	$y_1$	$y_2$	$y_3$
Search method	Best first	Best first	Best first
Search direction	forward	forward	forward
Total number of subsets evaluated	64	79	67
Selected factors	$x_2, z_2$	$x_1, x_2, x_9, z_2$	$x_2, x_9, z_2$

Table 4. DM results for responses  $y_1, y_2$ , and  $y_3$

#### 4.3 Stage III: results of multidisciplinary RD using RSM

Based on the results of the significant factor selection, RSM was performed by using the MINITAB software package to identify comprehensive relationships among a large number of factors and their associated responses. DA using the RSM provides the following fitted polynomial models for each quality characteristic:

$$y_{1S} = -18.63 + 14.84x_1 + 0.07x_2 - 175.15x_9 - 0.30z_2 - 0.77x_1^2 + 0.02x_2^2 + 2840.08x_9^2 + 0.34x_1x_2 - 139.15x_1x_9 - 0.01x_1z_2 - 11.37x_2x_9 - 0.05x_2z_2 + 18.24x_9z_2 \quad (10)$$

$$y_{2N} = -149.00 + 132.40x_1 - 1.50x_2 - 1665.40x_9 - 3.20z_2 - 6.70x_1^2 + 0.20x_2^2 + 26047.70x_9^2 + 3.20x_1x_2 - 1281.40x_1x_9 - 0.00x_1z_2 - 90.90x_2x_9 - 0.50x_2z_2 + 167.30x_9z_2 \quad (11)$$

$$y_{3L} = -22.35 + 21.98x_1 + 0.07x_2 - 297.09x_9 - 0.56z_2 - 1.16x_1^2 + 0.03x_2^2 + 4477.10x_9^2 + 0.55x_1x_2 - 216.36x_1x_9 + 0.00x_1z_2 - 17.50x_2x_9 - 0.08x_2z_2 + 28.53x_9z_2 \quad (12)$$

The response models for  $y_{1S}$ ,  $y_{2N}$ , and  $y_{3L}$  are adequate for use as a response function since the results yield 76.2, 77.4, and 76.1% R-sq, respectively. Let  $\sigma_{z_2}^2$ ,  $\sigma_{1S}^2$ ,  $\sigma_{2N}^2$ , and  $\sigma_{3L}^2$  denote the variance of the noise variables and mean-square error of the ANOVA values for friability, hardness, and disintegration, respectively. Using equations (10)-(12), the fitted polynomial models of the mean and variance can be written as

$$\hat{\mu}_{1S}(\mathbf{x}) = -18.63 + 14.84x_1 + 0.07x_2 + 175.15x_9 - 0.77x_1^2 + 0.02x_2^2 + 2840.08x_9^2 + 0.34x_1x_2 - 139.15x_1x_9 - 11.37x_2x_9 \quad (13)$$

$$\hat{\sigma}_{1S}(\mathbf{x}) = 1.23 + 0.04x_1 + 0.19x_2 - 69.39x_9 + 0.63e - 3x_1^2 + 0.63e - 2x_1x_2 - 2.31x_1x_9 + 0.02x_2^2 - 11.56x_2x_9 + 2109.30x_9^2 \quad (14)$$

$$\hat{\mu}_{2N}(\mathbf{x}) = -149.00 + 132.40x_1 - 1.50x_2 - 1665.40x_9 - 6.70x_1^2 + 0.20x_2^2 + 26047.70x_9^2 + 3.20x_1x_2 - 1281.40x_1x_9 - 90.90x_2x_9 \quad (15)$$

$$\hat{\sigma}_{2N}(\mathbf{x}) = 122.10 + 20.29x_2 - 6788.37x_9 + 1.59x_2^2 - 1060.68x_2x_9 + 177452.10x_9^2 \quad (16)$$

$$\hat{\mu}_{3L}(\mathbf{x}) = -22.35 + 21.98x_1 + 0.07x_2 - 297.09x_9 - 1.16x_1^2 + 0.03x_2^2 + 4477.10x_9^2 + 0.55x_1x_2 - 216.36x_1x_9 - 17.50x_2x_9 \quad (17)$$

$$\hat{\sigma}_{3L}(\mathbf{x}) = 3.56 + 0.57x_2 - 202.59x_9 + 0.04x_2^2 - 28.94x_2x_9 + 5160.51x_9^2 \quad (18)$$

where  $\hat{\mu}_{1S}(\mathbf{x})$ ,  $\hat{\sigma}_{1S}(\mathbf{x})$ ,  $\hat{\mu}_{2N}(\mathbf{x})$ ,  $\hat{\sigma}_{2N}(\mathbf{x})$ ,  $\hat{\mu}_{3L}(\mathbf{x})$ , and  $\hat{\sigma}_{3L}(\mathbf{x})$  represent the fitted polynomial models for the mean and variance of friability, hardness, and disintegration, respectively.  $\sigma_{z_2}^2$ ,  $\sigma_{1S}^2$ ,  $\sigma_{2N}^2$ , and  $\sigma_{3L}^2$  are 6.34, 0.66, 57.18, and 1.57 from the ANOVA result for each quality characteristic, respectively. Equations (13)-(18) are used in the proposed RDF model. The target value for the process mean of friability, hardness, and disintegration are 0.4, 50, and 10, respectively (i.e.,  $\tau_{1S} = 0.4$ ,  $\tau_{2N} = 50$ , and  $\tau_{3L} = 10$ ). Additionally, the constraints on  $x_1$ ,  $x_2$ , and  $x_9$  can be expressed as

$$3 \leq x_1 \leq 6, \quad 24 \leq x_2 \leq 45, \quad 0.1 \leq x_9 \leq 0.2 \quad (19)$$

We then convert the three quality characteristics into three MSEs of the same type. The target values and upper limits for the three MSE models are 0, 0, 0, 403.34, 3101.21, and

143.01, respectively. Suppose that the weights for the desirability of MSEs based on the friability, hardness, and disintegration are 1 (i.e.,  $w_{(1S, 2N, \text{ and } 3L)} = 1$ ).

By utilizing the proposed RDF model, which makes the multi-objective optimization problem inherently easier to solve due to the fact that single-objective optimization approaches can be applied, the multi-objective optimization problem can be transformed into a single response problem. In order to resolve this single response problem to maximize the geometric mean of the individual desirability of MSEs, MINITAB can be used. Using the MINITAB package, the optimal solutions are found to be  $(x_1^*, x_2^*, x_9^*) = (4.847, 41.315, 0.160)$ . The optimal solution and predicted value of the mean and variability of each process for this case are listed in Table 10.

Ingredients	Optimal Solution	Quality Characteristics	Predicted Value	
			$\hat{\mu}(\mathbf{x})$	$\hat{\sigma}^2(\mathbf{x})$
First pressure ( $x_1$ )	4.847	Friability ( $y_1$ )	1.934	2.272
Second pressure ( $x_2$ )	41.315	Hardness ( $y_2$ )	61.280	110.890
Amount of dust ( $x_9$ )	0.160	Disintegration ( $y_3$ )	4.780	4.673

Table 10. Optimal solutions for the tablet manufacturing process

## 7. Conclusion

In this paper, we developed a RDM method by integrating a DM method for pre-processing unclear data and finding significant factors into a multidisciplinary RD method for providing the best factor settings. Based on the results of the DM method, we found important factors for placebo tablet manufacturing among a large data set. By using the BFS method, the CFBS method in its pure form is exhaustive, but the use of a stopping criterion expedites the probability of searching the entire data set. We then conducted RD optimization using the RSM and RDF methods, while incorporating an uncontrollable noise factor. We finally showed that the proposed RDM method could efficiently find significant factors and optimal settings by reducing the dimensionality through the numerical example. In order to examine the proposed RDM method, the consideration of different case studies can be a possible future research issue.

## 8. Acknowledgement

This work was supported by the Korea Science and Engineering Foundation (KOSEF) grant funded by the Korea government (MOST) (No. R01-2007-000-21070-0).

## 9. References

- Allen, D. (1974). The Relationship between Variable Selection and Data Augmentation and a Method for Prediction, *Technometrics*, Vol. 16, No. 1, (Feb. 1974) 125–127
- Bakar, Z.A.; Mohamad, R.; Ahmad, A. & Deris, M.M. (2006). A Comparative Study for Outlier Detection Techniques in Data Mining, *proceedings of IEEE Conference on Cybernetics and Intelligent Systems*, pp. 1-6, ISBN: 1-4244-0023-6, Bangkok, Thailand, June 2006, IEEE Press New York

- Box, G.E.P.; Bisgaard, S. & Fung, C. (1998). An Explanation and Critique of Taguchi's Contributions to Quality Engineering. *International Journal of Reliability Management*, Vol. 4, No. 2, (Jan. 1998) 123-131, ISSN: 1099-1638
- Cho, B.R. (1994). Optimization issues in quality engineering, Ph.D. dissertation, School of Industrial Engineering, University of Oklahoma, OK, U.S.
- Gardner, M. & Bieker, J. (2000). Data Mining Solves Tough Semiconductor Manufacturing Problems. *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 376-383, ISBN: 1-58113-233-6, Boston, U.S., Aug. 2000, ACM, New York
- Hall, M. A. (1998). Correlation-based Feature Selection for Machine Learning. Ph.D Dissertation, Waikato University, Department of Computer Science. Hamilton, New Zealand
- Langley, P. (1994). Selection of Relevant Features in Machine Learning, *Proceedings of the AAAI Fall Symposium on Relevance*, pp. 140-144, ISBN 978-0-929280-76-9, New Orleans, U.S., Nov. 1994, AAAI Press, U.S.
- McLachlan, G. J., & Krishnan, T. (1996) The EM algorithm and extensions, John Wiley & Sons, New York, ISBN: 978-0-471-12358-3, New York
- Montgomery D.C. (2001). *Introduction to Statistical Quality Control*, John Wiley & Sons, ISBN: 0-471-39412-2, New York
- Myers, R. H. & Montgomery, D. C. (2002). *Response surface methodology: process and product optimization using designed experiments*, John Wiley & Sons, ISBN: 978-0-470-17446-3, New York
- Pernkopf, F.; Bouchaffra, D. (2005). Genetic-based EM algorithm for learning Gaussian mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 27, No. 8, (Aug. 2005) 1344 - 1348, ISSN:0162-8828
- Su, C.T., Chen, M.C., & Chan, H.L. (2005). Applying Neural Network and Scatter Search to Optimize Parameter Design with Dynamic Characteristics. *Journal of the Operational Research Society*, Vol. 56, No. 10, 1132-1140, ISSN 0160-5682
- Shin, S. & Cho, B.R. (2005). Bias-specified robust design optimization and its analytical solutions. *Computer & Industrial Engineering*, Vol. 48, No. 1, (Jan. 2005) 129-140, ISSN: 0360-8352
- Shin, S.; Guo Y.; Choi Y. & Choi M. (2006). Development of a Robust Data Mining Method Using CBFS and RSM, *LNCS 4378*, pp. 337-388, ISBN: 978-3-540-70880-3, Novosibirsk, Russia, June 2006, Springer-Verlag, Berlin Heidelberg
- Witten, I.W.H. & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, ISBN: 0-12-088407-0, San Francisco
- Xu, Q.; Kamel, M. & Salama, M.M.A. (2004). Significance Test for Feature Subset Selection on Image Recognition, *LNCS 3211*, pp.244-252, ISBN: 978-3-540-23223-0, Porto, Portugal, Sept. 2004, Springer-Verlag, Berlin Heidelberg
- Yang, L.; Shin, S.; Choi, Y.; Choi, M. & Lee, Y. (2007). A Surrogate Variable-Based Data Mining Method using CFS and RSM, *Proceedings of the 6th WSEAS International Conference on Applied Computer Science*, pp.651-657, ISBN: 978-960-8457-61-4 , Hangzhou, China, April 2007, World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, U.S.
- Yang, L.; Shin, S. ; Choi, Y. ; Park, K.; Kaewkuekool, S.; Chantrasa, R. & Lila, B. (2007). Development of an Extended Robust Data Mining (ERDM) Model, *Proceedings of*

*International Conference on Control, Automation and Systems*, pp. 1523-1528, ISBN: 978-89-950038-6-2, Seoul, Korea, Oct. 2007, IEEE Press, New York

Yu, L. & Liu, H. (2003). Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution, *Proceedings of the 20th International Conference on Machine Learning*, pp. 856-863, ISBN 978-1-57735-189-4, Washington D.C., U.S., Aug. 2003, AAAI Press, U.S.

## **PART II: CLUSTERING AND CLASSIFICATION**



# On the Selection of Meaningful Association Rules

Rangsipan Marukatat  
Mahidol University  
Thailand

## 1. Introduction

In recent years, data mining has been recognized as a powerful technique to extract hidden patterns from an enormous volume of data. These patterns can be expressed in many forms. One of them is as a set of association rules. From a transactional data set, an association rule " $x \rightarrow y$ " (support =  $s\%$ , confidence =  $c\%$ ) indicates the co-occurrence of items  $x$  and  $y$  in the same transaction, with certain levels of support and confidence. Association rule mining is useful in many applications. For example, in retailing, highly associated products can be identified and sold together as a special offer package (Svetina & Zupancic, 2005). Ma et al. (2003) extracted association rules from microbiology transactions. They detected outbreak of nosocomial infection, or infection acquired by patients during their hospital stay, from low-support and low-confidence rules.

A primeval but elegant association rule mining method, *Apriori* (Agrawal & Srikant, 1994), first discovers itemsets (or sets of data items) that satisfy a minimum support criterion. It then uses these itemsets to generate rules that satisfy a minimum confidence criterion. After *Apriori*, a number of advanced algorithms have been developed. The list includes Brin et al. (1997), Zaki et al. (1997), Liu et al. (1999b), Han et al. (2000), Yun et al. (2003), and Ko and Rountree (2005). Nevertheless, it is the original *Apriori* that is still the most popular one and becomes a standard function in many data mining software.

In real practices, data is usually rudimentary and the probability that each item occurs in a data set may be very low. Association rule mining might be performed several times, each with different sets of parameters, so that plenty of rules are generated and some useful, non-trivial ones can be spotted among them. This is where *Apriori*'s simplicity is traded off. It straightforwardly delivers rules that pass the thresholds given by users, but lacks effective pruning mechanisms. It is up to the users to handle the usually overwhelming amount of rules afterwards.

There have been techniques for post-pruning or post-selecting the association rules. For example, Ableson and Glasgow (2003) proposed statistic-based pruning. Others attempted to identify general and specific rules. Once identified, specific rules could be pruned or kept separately for further analysis (Berrado & Runger, 2007; Liu et al., 1999a; Toivonen et al., 1995). Techniques that exploit semantics conveyed in the rules include Klemettinen et al. (1994), Ma et al. (2003), and Silberschatz and Tuzhilin (1996). According to Li and Sweeney (2005), rules were selected and combined to form a new rule that expressed the knowledge more thoroughly. But the selection was performed as part of the rule generation.

### 1.1 Chapter contribution

This chapter presents an alternative approach to the selection of association rules. It suggests using an already available method, Apriori, to generate association rules. Then, the rules are selected or pruned based on their degrees of semantic redundancy and patterns. The rest of the chapter is organized as follows.

- Section 2 introduces basics of association rule mining, the Apriori algorithm, and the post-mining of association rules. It reviews how rules are summarized, interpreted, and selected or pruned in other works.
- Section 3 describes semantic analysis and pattern analysis. The former classifies rules into four groups: strongly meaningless, weakly meaningless, partially meaningful, and meaningful. The latter prunes repetitive patterns and retains ones that convey the most information.
- Section 4 demonstrates how the semantic analysis and pattern analysis were applied to a real-world application, an analysis of traffic accidents in Nakorn Pathom, Thailand.
- Section 5 discusses the proposed techniques and identifies their drawbacks.
- Section 6 concludes the chapter.

## 2. Association rule mining

Let  $I = \{i_1, i_2, \dots, i_m\}$  be a set of distinct items, and  $D = \{T_1, T_2, \dots, T_n\}$  be a transactional database. Each transaction  $T$  contains a subset of  $I$ . Table 1 shows an example data set and its binary representation.

Transaction ID	Items	Binary Representation				
		A	B	C	D	E
1	A, B, C	1	1	1	0	0
2	B, D	0	1	0	1	0
3	A, C, E	1	0	1	0	1
4	A, B, D, E	1	1	0	1	1
5	C, E	0	0	1	0	1

Table 1. An example data set

" $A \rightarrow B$ " is an association rule, given the following conditions:  $A \subset I$ ,  $B \subset I$ , and  $A \cap B = \emptyset$ . There are three common measures for an association rule. They are *support*, *confidence*, and *lift* (or *interest*). Support is the probability that both  $A$  and  $B$  occur in a transaction, i.e.  $P(A \cap B)$ . Confidence is the probability that  $B$  occurs in a transaction that  $A$  has occurred, i.e.  $P(B|A)$  or  $P(A \cap B)/P(A)$ . Lift normalizes the confidence with the probability of  $B$ , i.e.  $P(A \cap B)/(P(A) \times P(B))$ . The lift equalling one implies that  $A$  and  $B$  are independent of one another.

As mentioned in the introduction, Apriori is a classic association rule mining algorithm that is incorporated in many data mining software. It performs two tasks: (1) generating itemsets that pass a minimum support threshold; and (2) generating rules that pass a minimum confidence threshold. For many users, finding the right thresholds is not easy because low support leads to abundant rules but high support may cause important rules to be missed. Moreover, rules with high support usually have low confidence, and vice versa. Algorithm 1 is one possible implementation that wraps the main tasks in a loop (University of Waikato, n.d.). The support threshold is gradually adjusted at the end of each loop iteration. Either confidence or lift can be used as a criterion for the rule generation.

## Algorithm 1. Apriori

---

```

1. // Parameters given by users are UpperMinSupport, LowerMinSupport, Delta,
2. // Criterion, MinScore, and NumRules
3. Set of association rules =  $\emptyset$ 
4. N = 0
5. do {
6. // Task 1: generate frequent itemsets that satisfy minimum support criterion
7.   for k = 1 to NumItems {
8.     Find frequent k-itemset,  $S_k$ , that satisfies the condition:
9.     LowerMinSupport  $\leq$  support( $S_k$ )  $\leq$  UpperMinSupport
10.  }
11. // Task 2: generate rules that satisfy minimum confidence (or lift) criterion
12. for each frequent itemset S {
13.   for each subset SS of S {
14.     Rule R = " $SS \rightarrow (S - SS)$ "
15.     Compute confidence(R) and lift(R)
16.     if (Criterion == "lift") then score = lift(R)
17.     else score = confidence(R)
18.     if (score  $\geq$  MinScore) then {
19.       Add R to set of association rules
20.       N = N+1
21.     }
22.   }
23. }
24. UpperMinSupport = UpperMinSupport - Delta
25. } until (UpperMinSupport  $\leq$  LowerMinSupport) or (N == NumRules)
26. Sort set of association rules by Criterion

```

---

During the itemset generation, if an item is treated as an asymmetric attribute, it will be counted only when its value is not zero. This approach disallows negative association rules, or rules consisting of absent items such as  $pasta = 0 \rightarrow noodle = 1$ . By ignoring these rules, one could miss valuable information about conflict or competition between items (Antonie & Zaiane, 2004; Yuan et al., 2002). In contrast, if an item is treated as a categorical attribute, each of its distinct values will be counted as a category. This allows negative association rules to be generated, but plenty of them could be redundant ones.

## 2.1 Post-mining of discovered rules

Learning what is conveyed in association rules usually begins with getting an overview of the findings. Toivonen et al. (1995) searched for subsets of rules that cover or summarize the whole set. Among those having the same consequences, the most general rules, or the rules sharing common antecedents with the majority, are selected. For example, given four rules  $\{a\} \rightarrow \{z\}$ ,  $\{b\} \rightarrow \{z\}$ ,  $\{a, b\} \rightarrow \{z\}$ , and  $\{a, b, c\} \rightarrow \{z\}$ :

- $\{a\} \rightarrow \{z\}$  covers itself, the third and the fourth rules.
- $\{b\} \rightarrow \{z\}$  covers itself, the third and the fourth rules.
- $\{a, b\} \rightarrow \{z\}$  covers itself and the fourth rule.
- $\{a, b, c\} \rightarrow \{z\}$  covers itself only.

Hence, the first and the second rules are selected as summary of this group. They are also called direction setting (DS) rules because, guided by them, one could focus on further detail by going through their related non-DS rules (Liu et al., 1999a). To ensure that a DS rule offers useful information about items' relationship, it is picked only if the chi-square correlation between its antecedents and consequences is positive.

Another rule summarization method constructs meta rules which express relationship between the discovered association rules (Berrado & Runger, 2007).

Seeing an overall picture about the domain, one can proceed to in-depth analysis by looking at specific or non-DS rules. However, going through a huge amount of them would be too onerous. Statistical pruning employs rule's common measures and their derivations. A basic idea is that if adding items to a general rule, to make it more specific, does not improve the rule's measures, then the supposedly specific rule is too trivial and is consequently pruned out (Webb & Zhang, 2002). Recall that association rules were generated on the grounds that their support and confidence had passed the criteria set by users. One may argue that they should not be pruned only because their measures are too low. Moreover, if the goal is to conduct detailed analysis, any extra information might be worth retaining.

Taking semantics or information carried by the rules into account, Klemettinen et al. (1994) and Ma et al. (2003) allowed users to construct a set of templates for rule selection. They used two types of templates, inclusive and exclusive ones. A rule is selected if it fits at least one inclusive template and does not fit any of the exclusive templates.

Silberschatz and Tuzhilin (1996) suggested that a rule is interesting if (1) it is unexpected or surprises the users; or (2) it is actionable or allows the users to use it to their advantage. The former criterion requires semantic perception. A user's belief system is first defined. Rules are compared against this system. The ones that affect or do not follow the user's belief will be considered unexpected and thus interesting. General impression can be used instead of the complex belief (Liu et al., 1997). The general impression is defined loosely as the users may have only vague feelings about the domain. Three types of rules are identified against the general impression: conforming, unexpected conclusion, and unexpected condition.

Li and Sweeney (2005) constructed robust rules from sets of rules that carry the same pieces of information. In each set, the most general expression is assigned as the robust rule's antecedent, and the most specific expression as the robust rule's consequence. A rationale is that the general expression states a broad hypothesis whereas the specific expression gives exact description about the knowledge. This technique requires concept hierarchy and the rule selection is performed during, not after, the rule generation.

### 3. Semantic classification and pattern analysis

Given positive and negative association rules, one approach to pruning or selecting them is based on how their meanings can be formulated. A general idea is that rules are useful if their meanings offer insight about the domain. Naturally, a domain is composed of several perspectives. If a rule reveals only one of them and contains all negative terms (nonexistent items), then it is not useful from a decision making point of view. An example is a rule describing vehicles involved in a traffic accident  $\{bicycle = 0, sedan = 0\} \rightarrow \{truck = 0\}$ . On the other hand, a rule that reveals more than one perspectives and contains only a few negative terms offers useful, albeit incomplete, insight about those perspectives.

Suppose that a data set corresponds to a domain. Variables are grouped into *subjects* which correspond to the domain's perspectives. For example, in a traffic accident data set, binary

variables are grouped into three subjects: vehicles involved, causes of accidents, and human losses. The details are as follows.

1. Vehicles involved. Binary variables and their item representation are
 

V0	represents	bicycle or tricycle
V1		motorcycle
V2		sedan
V3		van or bus
V4		pick-up
V5		truck or trailer
V6		pedestrian
2. Causes of accidents. Binary variables and their item representation are
 

C0	represents	vehicle overloaded or malfunctioned
C1		speeding
C2		violating traffic signs
C3		illegal blocking
C4		illegal overtaking
C5		swerving in close distance
C6		driving in the wrong lane or direction
C7		failing to signal
C8		careless driving
C9		following in close distance
3. Human losses. Binary variables and their item representation are
 

H1	represents	dead
H2		seriously injured
H3		slightly injured

### 3.1 Formal definitions

Let  $\{S_1, S_2, \dots, S_n\}$  be subjects;  $\{V_{a1}, V_{a2}, \dots, V_{ap}\}$  be antecedent variables in a rule; and  $\{V_{c1}, V_{c2}, \dots, V_{cq}\}$  be consequence variables in a rule. Let *nil* refer to “absent” or “unknown” category, or categories outside the scope of interest. This definition enables the semantic analysis to cover not only binary variables but also categorical ones. As a refinement on previous work (Marukatat, 2007), criteria to determine whether a rule is semantically useful are as follows.

1. A rule is classified as strongly meaningless if it has the following form:

$$\{V_{ai} = nil \mid V_{ai} \in S, i = 1 \text{ to } p\} \rightarrow \{V_{ck} = nil \mid V_{ck} \in S, k = 1 \text{ to } q\}. \quad (1)$$

That is, all the variables have absent or unknown values, and they are members of the same subject. For instance,  $\{V0 = nil, V2 = nil\} \rightarrow \{V5 = nil\}$  implies that an accident *not* involving bicycle and sedan tends to *not* involve truck. In other words, these vehicles are all absent from the accident. Since there are many types of vehicles in the domain, it is impossible to infer which and how the remaining ones would fit into this accident. This type of rules does not make an individual aspect of the domain (vehicles involved, in this case) any clearer and, therefore, can be removed from the analysis.

2. A rule is classified as weakly meaningless if it has the following form:

$$\{V_{ai} = nil \mid V_{ai} \in S_t, i = 1 \text{ to } p\} \rightarrow \{V_{ck} = nil \mid V_{ck} \in S_t, k = 1 \text{ to } q\} \text{ where } t = 1 \text{ to } n. \quad (2)$$

That is, all the variables have absent or unknown values, and they are members of more than one subjects. For instance,  $\{V6 = nil, H1 = nil\} \rightarrow \{C1 = nil\}$  implies that an accident *not* involving pedestrian and *not* resulting in human death tends to *not* being caused by speeding. Although this rule does not offer insight about individual subjects (vehicles involved, human losses, and causes of accidents, in this case), it reveals some vague interaction between them. Nevertheless, it may not be worth squeezing information out of these rules if there are plenty other rules available.

3. A rule is classified as partially meaningful if it has the following form:

$$\{V_{ah} \neq nil, V_{ai} = nil \mid h, i = 1 \text{ to } p; h \neq i\} \rightarrow \{V_{cj} \neq nil, V_{ck} = nil \mid j, k = 1 \text{ to } q; j \neq k\}. \quad (3)$$

Some variables have absent or unknown values. For instance,  $\{V3 = 1, C1 = nil\} \rightarrow \{V1 = 1\}$  implies that an accident involving bus and *not* being caused by speeding tends to involve motorcycles. These rules are complementary to meaningful ones as they help understand negative association between variables.

4. A rule is classified as meaningful if it does not fall into any of the above categories.

### 3.2 Further rule selection

Among rules classified as meaningful and partially meaningful, there are still redundant or repetitive patterns. This is because Apriori generated rules by permuting items in frequent itemsets and choosing ones that passed the user's criteria without pruning. In addition, the algorithm may have been executed multiple times, with multiple sets of parameters, leading to even more redundancy.

Let  $S_1$  and  $S_2$  be sets of items in rules  $R_1$  and  $R_2$ , respectively. When the rules are compared, their relationship is defined as follows.

1. If  $S_1$  equals  $S_2$ , then  $R_1$  is equivalent to  $R_2$ .
2. If  $S_1$  includes all items in  $S_2$  and at least one item in  $S_1$  does not exist in  $S_2$  ( $S_2 \subset S_1$  and  $|S_1| > |S_2|$ ), then  $R_1$  covers  $R_2$ . In other words,  $R_2$  is covered by  $R_1$ .

The comparison takes every item into account irrespective of whether it is antecedent or consequence. The effect of it being one or the other is captured by the rule's measures, as illustrated by the following example. Given rules  $R_1$  and  $R_2$ :

1.  $R_1: \{V2 = 1, C1 = nil\} \rightarrow \{V1 = 1\};$   
 $R_2: \{V2 = 1\} \rightarrow \{C1 = nil, V1 = 1\};$   
 $R_1$  and  $R_2$  are equivalent since  $S_1 = S_2 = \{V2=1, C1=nil, V1=1\}$ .
2. Confidence( $R_1$ ) =  $P(V2=1 \cap C1=nil \cap V1=1) / P(V2=1 \cap C1=nil)$ .
3. Confidence( $R_2$ ) =  $P(V2=1 \cap C1=nil \cap V1=1) / P(V2=1)$ .
4. Lift( $R_1$ ) =  $P(V2=1 \cap C1=nil \cap V1=1) / (P(V2=1 \cap C1=nil) \times P(V1=1))$ .
5. Lift( $R_2$ ) =  $P(V2=1 \cap C1=nil \cap V1=1) / (P(V2=1) \times P(C1=nil \cap V1=1))$ .

From a set of equivalent rules, the most significant one is selected. The rules' significance are determined according to user's criteria such as lift and confidence. Rules covering the others are selected while the ones being covered are pruned out. The definitions of "cover" and "being covered" are in the opposite direction of those mentioned in Section 2.1. There, the aim was to summarize the entire domain by using general rules. Here, the aim is to dig out as much information as possible from specific rules.

#### 4. Case study: An analysis of traffic accident

This section demonstrates an application of semantic analysis and pattern analysis to real practice. Nakorn Pathom is a province located near Bangkok, the capital of Thailand. Over the past years, economic and human losses due to traffic accidents in Nakorn Pathom have been ranked among the highest of the country. Traffic accident data, dated from January 1<sup>st</sup>, 2003 to March 31<sup>st</sup>, 2006, were collected from its local police stations. The data set contains more records and more variables than the one used in previous publication (Marukatat, 2007). There are 1103 records, 20 binary variables, and 8 categorical variables in total. Subjects of binary variables were described in Section 3. Categorical variables were also grouped into the following subjects.

1. District. This subject includes only one categorical variable.
  - D1 represents Nakorn Pathom's district area
2. Time. This subject includes three categorical variables.
  - T1 represents quarter of day
  - T2 represents day of week
  - T3 represents quarter of year
3. Scenes of accidents. This subject includes four categorical variables.
  - S1 represents type of road (highway, local road, etc.)
  - S2 represents road feature (straight, intersection, etc.)
  - S3 represents road material (concrete, laterite, etc.)
  - S4 represents traffic direction (one-way, two-way)

Weka's Apriori (University of Waikato, n.d.) was employed to extract association rules from this data set. The algorithm was described in Section 2. The data set can be transformed to enable or disable the generation of negative association rules. In case that only positive rules are allowed, the *nil* value must be replaced by "?" which represents Weka's missing value. If negative rules are also allowed, *nil* is a user-defined value that represents a *nil* category. This case study used the latter setting. The other parameters were set as follows:

- NumRules = 500
- LowerMinSupport = 0.1
- UpperMinSupport = 0.4, 0.5, ..., 0.9
- Delta = 0.05
- Criterion = lift
- MinScore = 1.5, 2, 3, 4

The algorithm was executed 24 times by using different combinations of UpperMinSupport and MinScore. Only 12 combinations produced the results. There are 4368 rules in total, as summarized in Table 2.

The rules' contents varied from 2 to 9 items. There was slim chance that every item in the rule was *nil*. Consequently, only 2.2% of the discovered rules were classified as meaningless (see Table 3). About 32% were classified as meaningful, and 65.8% as partially meaningful. Next, pattern analysis was performed to remove redundant patterns and find the most specific ones. As a result, 5.3% and 5.4% of the meaningful and partially meaningful rules were retained, respectively. Their confidence and lift measures are displayed in Fig. 1 and Fig. 2.

Parameters		Results		
MinScore (Lift)	UpperMinSupport	Max Lift	Max Confidence	Number of Rules
4.0	0.8	4.43	0.77	24
3.0	0.8	4.43	0.95	210
2.0	0.8	2.99	0.91	500
2.0	0.7	4.37	0.91	500
2.0	0.6	4.37	0.95	500
2.0	0.5	4.37	0.95	500
2.0	0.4	3.42	0.95	38
1.5	0.8	1.65	0.90	500
1.5	0.7	2.15	0.97	500
1.5	0.6	2.32	0.91	500
1.5	0.5	3.73	0.94	500
1.5	0.4	3.42	0.95	96
Total				4368

Table 2. Summary of rules discovered by Apriori

Semantic Classification	Before		After	
	No. of Rules (1)	% of Total	No. of Rules	% of (1)
Meaningful	1398	32	74	5.3
Partially meaningful	2874	65.8	155	5.4
Meaningless	96	2.2	-	-
Total	4368	100	229	5.2

Table 3. Semantic classification, before and after pattern analysis

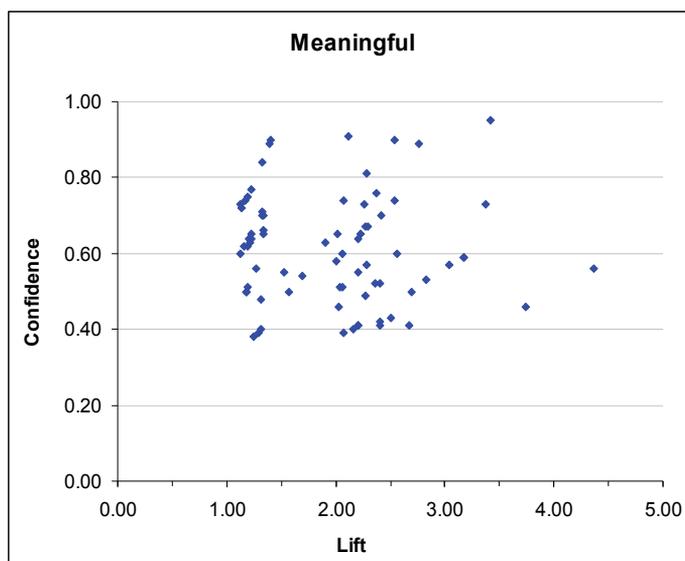


Fig. 1. Distribution of meaningful rules (after pattern analysis)

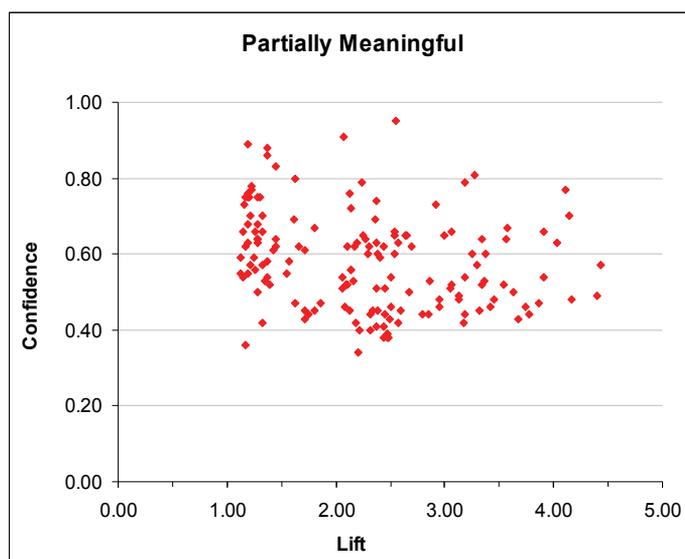


Fig. 2. Distribution of partially meaningful rules (after pattern analysis)

The following are some of the retained rules. To make them more understandable, items were substituted by categories' description, e.g.  $S1 = 1$  was substituted by *highway*.

---

#### Meaningful

---

1. 12.01-18.00, local road, intersection → illegal blocking
  2. 12.01-18.00, straight, dead → truck
  3. 00.01-06.00, truck → dead
  4. 00.01-06.00, highway, straight → speeding
  5. 18.01-24.00, pedestrian → speeding, dead
  6. Swerving in close distance → bicycle
  7. Bus, swerving in close distance → intersection
  8. Local road, illegal overtaking → curve
  9. Violating traffic signs → pedestrian
  10. Violating traffic signs → 06.01-12.00, community area
- 

#### Partially meaningful

---

11. Truck → highway, *no sedan, not speeding, not swerving in close distance*
  12. Local road, slightly injured → *two ways, asphalt surface, no sedan, not speeding*
  13. Bicycle, truck → *straight, no pick-up, no slightly injured*
  14. *No motorcycle, dead* → *highway, speeding*
  15. 18.01-24.00, pedestrian → *no motorcycle, dead*
- 

Nakorn Pathom is known as a gateway to the western and the southern regions of Thailand. Heavy vehicles usually travel through the province at night and in the early morning, rather than in the afternoon. Thus, rules 3 and 4 look like typical accident patterns in Nakorn Pathom while rule 2 is slightly unexpected.

Rule 1 implies that an accident occurring at the intersection of local roads, in the afternoon, is likely to be caused by illegal blocking. Rule 10 implies that an accident caused by violating traffic signs is likely to occur in community area, in the morning. An investigation into the

amount of traffic during rush hours, the adequacy of traffic lights around the intersections, and the law/regulation enforcement should be made to complete the picture.

Problems with law/regulation enforcement is confirmed by rule 9, but this rule describes a rather typical accident pattern in many parts of Thailand. That is, pedestrians often cross the roads wherever they like and vehicles seldom stop for them.

The analysis of partially meaningful rules offers more insight about traffic accidents, or at least spawns a few questions for further investigation. For example, rule 11 implies that an accident which involves truck and highway does not involve sedan, and is not caused by speeding or swerving in close distance. One might suspect if there is any type of vehicles other than truck potentially fitting the pattern described by rule 4.

Rule 15 can serve as a complement to rule 5. From rule 5, one learns an accident pattern that happens in the evening until midnight and involves pedestrian. It is likely to be caused by speeding and result in human death. From rule 15, one also learns that this accident is likely to not involve motorcycle.

## 5. Discussion

### 5.1 Semantic analysis

Association rules can be classified by their meanings or semantics. A rule is meaningful if its whole content describes something which exists in the domain. In contrast, it is meaningless if its whole content describes something which does not exist. A partially meaningful rule is somewhere in the middle. It is comparable to a negative association rule in " $A \rightarrow \sim B$ " or " $\sim A \rightarrow B$ " forms. As mentioned in Section 2, this type of association is useful in marketing research. It helps identify conflicting items that should not be promoted together, or a replacement item in case that the other is in short supply. Wu et al. (2004) gave another example. Suppose that A normally triggers an alert of event B. Rule " $A \rightarrow \sim B$ " suggests that the alert can be postponed because B has not yet happened. This chapter has also demonstrated how such rules were used to gain a better understanding about the domain. However, not all of them are useful for the analysis. Consider the following:

1.  $\{V3 = 1, C5 = 1\} \rightarrow \{S2 = 2, V1 = 1, V2 = nil, V4 = nil, V5 = nil, V6 = 1\}$
2.  $\{V3 = 1, C5 = 1\} \rightarrow \{S2 = 2, C1 = nil, C3 = nil, C4 = nil, C8 = nil\}$

Both rules say that accidents involving bus and being caused by swerving in close distance is likely to occur at the intersection. They give further detail about vehicles involved and causes of accidents, respectively. The extra detail regarding vehicles involved is interesting because it is normal that an accident involves more than one vehicles. On the other hand, it is unlikely (but sometimes possible) that an accident is caused by so many reasons. Hence, adding that there is no other cause of accident is unnecessary.

The above example shows different natures of the subjects. In some subjects, multiple or all the items can exist at the same time. But in the others, one or only a couple of items can co-exist. Further refinement should be made to the semantic analysis to handle this.

A few works have mentioned negative association rules in " $\sim A \rightarrow \sim B$ " form (Antonie & Zaiane, 2004; Wu et al., 2004; Yuan et al., 2002). None explained how to exploit such rules. Only Yuan et al. (2002) suggested that " $\sim A \rightarrow \sim B$ " is equivalent to " $B \rightarrow A$ ", but did not elaborate any further. To make sense of this, " $\sim A \rightarrow \sim B$ " is interpreted as that the absence of A causes the absence of B. Therefore, the presence of B would imply the presence of A. It is probable if the association ( $\rightarrow$ ) is perceived as cause-and-effect relationship.

The cause-and-effect perception is weak when every item belongs to the same subject, as in a strongly meaningless rule  $\{V0 = nil, V2 = nil\} \rightarrow \{V5 = nil\}$ . It is more natural when items belong to different subjects, as in a weakly meaningless rule  $\{V6 = nil, H1 = nil\} \rightarrow \{C1 = nil\}$ . However, it is imprudent to infer that any form of the inverse, e.g.  $\{V6 = 1, H1 = 1\} \rightarrow \{C1 = 1\}$  or  $\{C1 = 1\} \rightarrow \{V6 = 1, H1 = 1\}$ , is true.

Although weakly meaningless rules give a little insight about the domain, it is still unclear how to exploit them effectively. At the moment, they serve only as a confirmation of what has been learned from meaningful rules.

## 5.2 Pattern analysis

The pattern analysis helps remove repetitive or redundant patterns. It aims to retain rules which describe the most information. These rules normally have much lower support and confidence than general ones. In spite of this, the measures are supposed to be accepted by the users, according to their own (Apriori's) criteria. Otherwise, the support and confidence thresholds could have been raised to avoid generating these rules.

Based on the analysis, rules  $\{V2 = 1\} \rightarrow \{C1 = nil, V1 = 1\}$  and  $\{V2 = 1, C1 = nil\} \rightarrow \{V1 = 1\}$  are equivalent. The one with the higher confidence or lift will be selected. In some practices, both of them are considered important. The likes of  $\{V2 = 1\} \rightarrow \{C1 = nil, V1 = 1\}$  are characteristic rules while the likes of  $\{V2 = 1, C1 = nil\} \rightarrow \{V1 = 1\}$  are discriminant rules (Brijs et al., 2000; Cheung et al., 2000). The former are useful for description purpose, since they characterize single antecedent items (concepts) with multiple consequence items. The latter are useful for prediction purpose, since they discriminate consequence items (classes) by using multiple antecedent items. However, this chapter sees both of them as conveying the same piece of information, only in slightly different forms. Therefore, keeping any one of them would be sufficient.

This strategy has some drawbacks. First, consider the following cases:

1.  $R_0: \{V2 = 1\} \rightarrow \{S1 = 1, H2 = 1\}$  and  $R_1: \{V2 = 1, C1 = nil\} \rightarrow \{V1 = 1\}$  are selected.
2.  $R_0: \{V2 = 1\} \rightarrow \{S1 = 1, H2 = 1\}$  and  $R_2: \{V2 = 1\} \rightarrow \{C1 = nil, V1 = 1\}$  are selected.

The second case would make analysis job easier because the rules can be grouped easily and insight about the accidents involving sedan can be obtained quickly. But the first case may happen if  $R_1$  has higher confidence or lift than  $R_2$ . The fact that specific rules usually have many more items than general ones makes the analysis even harder.

## 6. Conclusion

Association rule mining produces a large amount of rules. Many of them are redundant ones. This chapter has presented techniques to select rules that are semantically useful and carry the most information. They aim for a complete understanding, rather than an overall picture, about the domain. Prior to the analysis, variables are grouped into subjects which correspond to the domain's perspectives. These subjects are key factors to classify the rules into strongly meaningless, weakly meaningless, partially meaningful, and meaningful ones. Rules that have equivalent patterns are identified and the most significant one is selected. Furthermore, between a general and a more specific rule, the latter is selected since it offers more insight about the domain.

These rule selection strategies still have some drawbacks, as discussed in Section 5. Further refinement on the semantic analysis would help filter out even more semantically redundant

rules. Throughout this chapter, it was assumed that rules which offer as much information as possible are the ones users would like to see. But as shown in Section 4, some rules only describe what the users already knew. There are techniques, as mentioned in Section 2, that take the users' existing knowledge and the unexpectedness of the rules into account (Liu et al., 1997; Silberschatz & Tuzhilin, 1996). Their ideas could be incorporated to improve the rule selection capability. Finally, visualization systems (Blanchard et al., 2003; Bruzzese & Davino, 2005; Techapichetvanich & Datta, 2005) would make the analysis job easier.

## 8. References

- Ableson, A. & Glasgow, J. (2003). Efficient statistical pruning of association rules, In: PKDD 2003, *Lecture Notes in Computer Science*, Vol. 2838, Lavrac, N. et al. (Eds.), pp. 23-34, Springer.
- Agrawal, R. & Srikant, R. (1994). Fast algorithms for mining association rules, *Proceedings of the 20<sup>th</sup> International Conference on Very Large Databases (VLDB)*, pp. 487-499, Santiago, Chile, September 1994, Morgan Kaufmann.
- Antonie, M.-L. & Zaiane, O. R. (2004). Mining positive and negative association rules: An approach for confined rules, In: PKDD 2004, *Lecture Notes in Computer Science*, Vol. 3202, Boulicaut, J. F. et al. (Eds.), pp. 27-38, Springer.
- Berrado, A. & Runger, G. C. (2007). Using meta rules to organize and group discovered association rules, *Data Mining and Knowledge Discovery*, Vol. 14, No. 3, pp. 409-431, Springer.
- Blanchard, J., Guillet, F. & Briand, H. (2003). Exploratory visualization for association rule rummaging, *The 4<sup>th</sup> International Workshop on Multimedia Data Mining (MDM/KDD)*, Washington, DC, August, 2003.
- Brijs, T., Vanhoof, K. & Wets, G. (2000). Reducing redundancy in characteristic rule discovery by using integer programming techniques, *Intelligent Data Analysis*, Vol. 4, No. 3-4, pp. 229-240, IOS Press.
- Brin, S., Motwani, R., Ullman, J. D. & Tsur, S. (1997). Dynamic itemset counting and implication rules for market basket data, *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 255-264, Tucson, AZ, May 1997, ACM Press.
- Bruzzese, D. & Davino, C. (2003). Visual post-analysis of association rules, *Journal of Visual Languages and Computing*, Vol. 14, No. 6, pp. 621-635, Elsevier.
- Cheung, D. W., Hwang, H. Y., Fu, A. W. & Han, J. (2000). Efficient rule-based attribute-oriented induction for data mining, *Journal of Intelligent Information Systems*, Vol. 15, No. 2, pp. 175-200, Kluwer Academic Publishers.
- Han, J., Pei, J. & Yin, Y. (2000). Mining frequent patterns without candidate generation, *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 1-12, Dallas, TX, May 2000.
- Klemettinen, M., Mannila, H., Ronkainen, P., Toivonen, H. & Verkamo A. I. (1994). Finding interesting rules from large sets of discovered association rules, *Proceedings of the 3<sup>rd</sup> International Conference on Information and Knowledge Management (ICKM)*, pp. 401-407, Gaithersburg, MD, November 1994.
- Ko, Y. S. & Rountree, N. (2005). Finding sporadic rules using Apriori-Inverse, In: PAKDD 2005, *Lecture Notes in Computer Science*, Vol. 3518, Ho, T. B. et al. (Eds.), pp. 97-106, Springer.

- Li, Y. & Sweeney, L. (2005). *Adding semantics and rigor to association rule learning: the GenTree approach*, Technical Report CMU ISRI 05-101, School of Computer Science, Carnegie Mellon University.
- Liu, B., Hsu, W. & Chen, S. (1997). Using general impressions to analyze discovered classification rules, *Proceedings of the 3<sup>rd</sup> International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 31-36, Newport Beach, CA, August 1997, AAAI Press.
- Liu, B., Hsu, W. & Ma, Y. (1999a). Pruning and summarizing the discovered association, *Proceedings of the 5<sup>th</sup> ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 125-134, San Diego, CA, August 1999.
- Liu, B., Hsu, W. & Ma, Y. (1999b). Mining association rules with multiple minimum supports, *Proceedings of the 5<sup>th</sup> ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 337-341, San Diego, CA, August 1999.
- Ma, L., Tsui, F.-C., Hogan, W. R., Wagner, M. M. & Ma, H. (2003). A framework for infection control surveillance using association rules, *AMIA Annual Symposium Proceedings*, pp. 410-414, American Medical Informatics Association.
- Marukatat, R. (2007). Structure-based rule selection framework for association rule mining of traffic accident data, In: *CIS 2006, Lecture Notes in Computer Science*, Vol. 4456, Wang, Y. et al. (Eds.), pp. 231-239, Springer.
- Silberschatz, A. & Tuzhilin, A. (1996). What makes patterns interesting in knowledge discovery systems, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 8, No. 6, pp. 970-974.
- Svetina, M. & Zupancic, Joze. (2005). How to increase sales in retail with market basket analysis. *Systems Integration*, pp. 418-428.
- Techapichetvanich, K. & Datta, A. (2005). VisAr: A new technique for visualizing mined association rules, In: *ADMA 2005, Lecture Notes in Computer Science*, Vol. 3584, Li, X. et al. (Eds.), pp. 88-95, Springer.
- Toivonen, H., Klemettinen, M., Ronkainen, P., Hatonen, K. & Mannila, H. (1995). Pruning and grouping of discovered association rules, *Workshop Notes of the ECML 95 Workshop in Statistics, Machine Learning, and Knowledge Discovery in Databases*, pp. 47-52, Greece, 1995.
- University of Waikato (n.d.). Weka 3 - data mining software in Java (version 3.4.12) [software]. Available from <http://www.cs.waikato.ac.nz/ml/weka/>.
- Webb, G. I. & Zhang, S. (2002). Removing trivial association in association rule discovery, *Proceedings of the 1<sup>st</sup> International NAISO Congress on Autonomous Intelligent Systems (ICAIS)*, Geelong, Australia, 2002, NAISO Academic Press.
- Wu, X., Zhang, C. & Zhang, S. (2004). Efficient mining of both positive and negative association rules, *ACM Transactions on Information Systems*, Vol. 22, No. 3, pp. 381-405.
- Yuan, X., Buckles, B. P., Yuan, Z. & Zhang, J. (2002). Mining negative association rules, *Proceedings of the 7<sup>th</sup> IEEE International Symposium on Computers and Communications (ISCC)*, pp. 623-628, Taormina, Italy, July 2002, IEEE Computer Society.
- Yun, H., Ha, D., Hwang, B. & Ryu, K. H. (2003). Mining association rules on significant rare data using relative support, *Journal of Systems and Software*, Vol. 67, No. 3, pp. 181-191, Elsevier.

Zaki, M. J., Parthasarathy, S., Ogihara, M. & Li, W. (1997). New algorithms for fast discovery of association rules, *Proceedings of the 3<sup>rd</sup> International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 283-286, Newport Beach, CA, August 1997, AAAI Press.

# Hybrid Clustering for Validation and Improvement of Subject-Classification Schemes

Frizo Janssens<sup>1,2,3</sup>, Lin Zhang<sup>1,4</sup> and Wolfgang Glänzel<sup>1,5</sup>

<sup>1</sup>*K.U. Leuven, Steunpunt O&O Indicatoren, Dept. MSI, Leuven*

<sup>2</sup>*Attentio SA/NV, StudioTROPE building, Bloemenstraat 32, B-1000 Brussels,*

<sup>3</sup>*K.U. Leuven, ESAT-SCD, Leuven*

<sup>4</sup>*WISE Lab, Dalian University of Technology, Dalian*

<sup>5</sup>*Hungarian Academy of Sciences, IRPS, Budapest*

<sup>1,2,3</sup>*Belgium*

<sup>4</sup>*China*

<sup>5</sup>*Hungary*

## 1. Introduction

The history of cognitive mapping of science is as long as the history of computerised scientometrics itself. While the first visualisations of the structure of science were considered part of information services, i.e., an extension of scientific review literature (Garfield, 1975, 1988), bibliometricians soon recognised the potential value of structural science studies for science policy and research evaluation as well. At present, the identification of emerging and converging fields and the improvement of subject delineation are in the foreground.

The main bibliometric techniques are characterised by three major approaches, particularly the analysis of citation links (cross-citations, bibliographic coupling, co-citations), the lexical approach (text mining), and their combination. The widely used method of co-citation clustering was introduced independently by Small (1973, 1978) and Marshakova (1973). Although the principle of bibliographic coupling had already been discovered earlier by Fano (1956) and Kessler (1963), coupling-based techniques have been used for mapping the structure of science only decades after co-citation analysis had become a standard tool in visualising the structure of science (e.g., Glänzel & Czerwon, 1996; Small, 1998). Cross-citation based cluster analysis for science mapping has to be distinguished from the previous two methods; while the former two types can be – and usually are – based on links connecting individual documents, the latter approach requires aggregation of documents to units like journals, subject categories, etc., among which cross-citation links are established. The obvious advantages of this method (e.g., the possibility to analyse directed information flows among these units or the assignment/aggregation of units to larger structures) are contrasted by some limitations and shortcomings such as possible biases caused by the use of predefined units. Thus, for instance, Leydesdorff (2006), Leydesdorff and Rafols (2008), and Boyack et al. (2008) used journal cross-citation matrices, while Moya-Anegón (2007) used subject co-citation analysis to visualise the structure of science and its dynamics.

Earlier, a completely different approach was introduced by Callon et al., (1983) and Callon, Law and Rip (1986). Their mapping and visualisation tool Leximappe was based on a lexical approach, particularly, co-word analysis. The notion of lexical approach, which was originally based on extracting keywords from records in indexing databases, was later on deepened and extended by using advanced text-mining techniques in full texts (cf. Kostoff et al., 2001, 2005; Glenisson et al., 2005a,b).

Whatever method is used to study the structure of science, cluster algorithms have beyond doubt become the most popular technique in science mapping. The sudden, large interest the application of these techniques has found in the community is contrasted by objections and criticism from the viewpoint of information use in the framework of research evaluation (e.g., Noyons, 2001; Jarneving, 2005). For instance, clustering based on co-citation and bibliographic coupling has to cope with several severe methodological problems. This has been reported, among others by Hicks (1987) in the context of co-citation analysis and by Janssens et al. (2008) with regard to bibliographic coupling. One promising solution is to combine these techniques with other methods such as text mining (e.g., combined co-citation and word analysis: Braam et al., 1991; combination of coupling and co-word analysis: Small (1998); hybrid coupling-lexical approach: Janssens et al., 2007b, 2008). Most applications were designed to map and visualise the cognitive structure of science and its change in time, and, from a policy-relevant perspective, to detect new, emerging disciplines. Improvement of subject-classification schemes was in most cases not intended. Jarneving (2005) proposed a combination of bibliometric structure-analytical techniques with statistical methods to generate and visualise subject coherent and meaningful clusters. His conclusions drawn from the comparison with 'intellectual' classification were rather sceptical. Despite several limitations, which will be discussed further in the course of the present study, cognitive maps proved useful tools in visualising the structure of science and can be used to adjust existing subject classification schemes even on the large scale as we will demonstrate in the following.

The main objective of this study is to compare (hybrid) cluster techniques for cognitive mapping with traditional 'intellectual' subject-classifications schemes. The most popular subject classification schemes created by Thomson Scientific (Philadelphia, PA, USA) are based on journal assignment. Therefore journal cross-citation analysis puts itself forward as underlying method and we will cluster the document space using journals as predefined units of aggregation. In contrast to the method applied by Leydesdorff (2006), who uses the Journal Citation Reports (JCR), we calculate citations on a paper-by-paper basis and then assign individual papers indexed in the *Web of Science* (WoS) database to the journals in which they have been published. The use of the JCR would confine us to data as available in the JCR and prevent us from combining cross-citation analysis with a textual approach. What is more, proceeding from the document level allows us to control for document types and citation windows, and to combine bibliometrics-based techniques with other methods like text mining. This results in a higher precision since irrelevant document types and 'low-weight journals' can be excluded. This way we can present the results of a hybrid (i.e., combined/integrated) citation-textual cluster analysis to compare those with the structure of an existing 'intellectual' subject classification scheme created and used by Thomson Scientific. The aim of this comparison is exploring the possibility of using the results of the cluster analysis to improve the subject classification scheme in question.

### 1.1 Cognitive mapping vs. subject classification

The objective of the present study is two-fold. The first task is not merely visualising the field structure of science by presenting yet another map based on an alternative approach, but to validate and improve existing subject classifications used for research evaluation. In particular, the question arises of in how far observed ‘migration’ of journals among science fields can be adopted to improve classification. The second issue is, however, a methodological one, namely to evaluate improved methods of hybrid clustering techniques. The 22-field subject classification scheme of the Essential Science Indicators (ESI) of Thomson Scientific, which actually forms a partition of the Web of Science universe with practically unique subject assignment, is used as the “control structure”. In particular, we propose the following approach in seven steps to solve the integration of cluster analysis and cognitive mapping into subject classification.

1. Evaluation of existing subject-classification schemes and visualisation of their cross-citation graph
2. Labelling subject fields using cognitive characteristics
3. Studying the cognitive structure based on hybrid cluster analysis and visualisation of the cross-citation graph
4. Evaluation of science areas resulting from cluster analysis
5. Labelling clusters using cognitive characteristics and representative journals suggested by the PageRank algorithm
6. Comparison of subject fields and cluster structure
7. Migration of journals among subject fields

## 2. Data sources and data processing

In order to accomplish the above objectives, more than six million papers of the type article, letter, note and review indexed in the Web of Science (WoS) in the period 2002–2006 have been taken into consideration. Citations to individual papers have been aggregated from the publication year till 2006. The complete database has been indexed and all terms extracted from titles, abstracts and keywords have been used for “labelling” the obtained clusters.

Citations received by these papers have been determined for a variable citation window beginning with the publication year, up to 2006, on the basis of an item-by-item procedure using special identification-keys made up of bibliographic data elements extracted from first-author names, journal title, publication year, volume and first page.

In a first step, journals had to be checked for name changes, merging or splitting and identified accordingly. Journals which were not covered in the entire period have been omitted. Furthermore, only journals that have published at least 50 papers in the period under study were considered. A second threshold was used afterwards to remove all journals for which the sum of references and citations was lower than 30. The resulting number of retained journals was 8,305. Most of the subsequent analyses were performed in Java and MATLAB. We also made use of the MATLAB Tensor Toolbox (Bader, 2006).

## 3. Methods

In this section we briefly describe the methodological background and the algorithms and procedures that have been applied. The first subsection refers to the outlines of the textual approach; this is followed by the description of the cross-citation analysis. The journal

clustering techniques described in the subsequent paragraphs are applied to the textual and citation data separately and used for combined (hybrid) clustering as well. This procedure is described in the following step by step.

### 3.1 Text analysis

All textual content was indexed with the Jakarta Lucene platform (Hatcher, 2004) and encoded in the Vector Space Model using the TF-IDF weighting scheme (Baeza-Yates, 1999). Stop words were neglected during indexing and the Porter stemmer was applied to all remaining terms from titles, abstracts, and keyword fields. The resulting term-by-document matrix contained nine and a half million term dimensions (9,473,061), but by ignoring all tokens that occurred in one sole document, only 669,860 term dimensions were retained. Those ignored terms with a document frequency equal to one are useless for clustering purposes. The dimensionality was further reduced from 669,860 term dimensions to 200 factors by Latent Semantic Indexing (LSI) (Deerwester, 1990; Berry, 1995), which is based on the Singular Value Decomposition (SVD). The reduction of the number of features in a vector space by application of LSI improves the performance of retrieval, clustering, and classification algorithms. Text-based similarities were calculated as the cosine of the angle between the vector representations of two papers (Salton, 1986).

### 3.2 Citation analysis

Since the present study analyses the structure of science on the level of journals, all local citations between papers are aggregated to form a journal cross-citation graph. For cluster analysis we ignored the direction of citations by symmetrising the journal cross-citation matrix. At the level of journal clusters, the journal cross-citations can be further aggregated into inter-cluster citations.

From the raw number of cross-citations between two journals (or clusters, respectively), a normalised similarity can be calculated by dividing it by the square root of the product of the total number of citations to or from the first journal (cluster), and the total number of citations to or from the second. Intra-cluster 'self-citations' are counted only once.

For visualisation of the networks we use the similarities just described as edge weights between two clusters or fields (see Figure 2 for an example). For clustering, however, we calculated the similarity of two journals somewhat differently because we didn't want to ignore, for instance, that both journals could be highly cited by a third one. That's why we opted to use "second order" journal cross-citation similarities for clustering. The journal cross-citation numbers are usually stored in a square, symmetric matrix. With "second-order similarities" we mean that the cross-citation values between a journal and all other journals (i.e., row or column of the matrix with cross-citation numbers) are used as input for another step of pairwise similarity calculation. The second-order similarities are found by calculating the cosine of the angle between pairs of vectors containing all symmetric journal cross-citation values between the two respective journals and all other journals. Hence, the ultimate similarity of two journals is based on their respective similarities with all other journals.

The journal cross-citation graph is also analysed to identify important high-impact journals. We use the PageRank algorithm (Brin, 1998) to determine representative journals in each cluster. Besides, the graph can also be used to evaluate the quality of a clustering outcome.

### 3.3 Clustering

In order to subdivide the journal set into clusters we used the agglomerative hierarchical cluster algorithm with Ward's method (Jain, 1988). It is a hard clustering algorithm, which means that each individual journal is assigned to exactly one cluster.

#### 3.3.1 Number of clusters

Determination of the optimal number of clusters in a data set is a difficult issue and depends on the adopted validation and chosen similarity measures, as well as on data representation. In general, the number of clusters is determined by comparing the quality of different clustering solutions based on various numbers of clusters. Cluster quality can be assessed by internal or external validation measures. Internal validation solely considers the statistical properties of the data and clusters, whereas external validation compares the clustering result to a known gold standard partition. Halkidi, Batistakis and Vazirgiannis (2001) gave an overview of quality assessment of clustering results and cluster validation measures. The strategy that we adopted to determine the number of clusters is a combination of distance-based and graph-based methods. This compound strategy encompasses observation of a dendrogram, text- and citation-based mean Silhouette curves, and modularity curves. Besides, the Jaccard similarity coefficient and the Rand index are used to compare the obtained results with an intellectual classification scheme.

#### 3.3.2 Dendrogram

A preliminary judgment is offered by a dendrogram, which provides a visualisation of the distances between (sub-) clusters (see Figure 4 for an example). It shows the iterative grouping or splitting of clusters in a hierarchical tree. A candidate number of clusters can be determined visually by looking for a cut-off point where an imaginary vertical line would cut the tree such that resulting clusters are well separated. Because of the difficulty to define the optimal cut-off point on a dendrogram (Jain, 1988), we complement this method with other techniques.

#### 3.3.3 Silhouette curves

A second appraisal for the number of clusters is given by the curve with mean *Silhouette values*. The Silhouette value for a document ranges from -1 to +1 and measures how similar it is to documents in its own cluster vs. documents in other clusters (Rousseeuw, 1987). The average Silhouette value for all clustered objects (e.g., journals) is an intrinsic measurement of the overall quality of a clustering solution with a specific number of clusters. Since Silhouette values are based on distances, depending on the chosen distance measure and reference data different Silhouette values can be calculated. For instance, we use the complement of cosine similarity applied to text and citation data.

The quality of a specific partition can be visualised in a *Silhouette plot*. In a Silhouette plot (see Figures 1 & 5), the sorted Silhouette values of all members of each cluster (or field) are indicated with horizontal lines. The more the Silhouette profile of a cluster (field) is to the right of the vertical line at the value 0, the more coherent the cluster (field) is, whereas negative values indicate that the corresponding objects should rather belong to another cluster (field).

### 3.3.4 Modularity curves

The quality of a clustering can also be evaluated by calculating the modularity of the corresponding partition of the cross-journal citation graph (Newman & Girvan, 2004; Newman, 2006). Up to a multiplicative constant, modularity measures the number of intra-cluster citations minus the expected number in an equivalent network with the same clusters but with citations given at random. Intuitively, in a good clustering there are more citations within (and fewer citations between) clusters than could be expected from random citing. The expected number of citations between two journals is based on their respective degrees and on the total number of citations in the network.

For an additional 'external validation' of clustering results, we also use modularity curves computed from a network containing all journals as nodes, but with edge weights equal to the number of ISI Subject Categories commonly assigned to both journals by Thomson Scientific (out of the total of 254).

### 3.3.5 Jaccard similarity coefficient and Rand index

The Jaccard index is the ratio of the cardinality of the intersection of two sets and the cardinality of their union. The Jaccard similarity coefficient is an extension of the Jaccard index and can be used as a measure for external cluster validation. The Rand index is another external validation measure to quantify the correspondence between a clustering outcome and a ground-truth categorisation (Jain, 1988). In contrast to the Jaccard coefficient, the Rand index does take into account negative matches as well. Both measures result in a value between 0 and 1, with 1 indicating identical partitions. In Figure 8, we use the Jaccard index to compare each cluster with every field from the intellectual ESI classification, in order to detect the best matching fields for each cluster.

### 3.3.6 Hybrid clustering

As mentioned at the outset, in general four major approaches are used for clustering sets of scientific papers, particularly, the lexical approach and three citation-based methods, namely cross-citation, bibliographic coupling, and co-citation analysis. Each of the methods alone suffers from severe shortcomings. For example, typical problems with bibliographic coupling and co-citations are sparse matrices, the lack of consensual referencing in some areas (Braam et al., 1991b; Jarneving, 2007), document types with insufficient number of references (e.g., letters) that have to be excluded (bibliographic coupling), the incompleteness due to missing citations to recent years (co-citation analysis), the missing 'critical mass' for emerging field detection (co-citation analysis, cf. Hicks, 1987), and the bias towards high-impact journals (co-citation analysis). If strict citation-based criteria are applied, then the resulting citations-by-document matrix is extremely sparse. In this case, rejection of relationship between two entities (e.g., journals or documents) tends to be unreliable. On the other hand, any lexical (text-based) approach is usually based on rather rich vocabularies and peculiarities of natural language. The result is, according to our observations, a rather 'smooth' or gradual transition between what is related and what is not. Therefore, the relationship is somewhat fuzzy and not always reliable. Hence, both the textual and citation-based approaches provide different perceptions of similarities among the same data. Textual information might indicate similarities that are not visible to bibliometric techniques, but true document similarity can also be obscured by differences in

vocabulary use, or spurious similarities might be introduced as a result of textual pre-processing, or because of polysemous words or words with little semantic value. The combination of the two worlds helps to improve the reliability of relationship and therefore of the clustering algorithm as well.

Therefore, the present study combines cross-citation analysis with text mining. The former can be applied to directed links as well as to the symmetrised transaction matrix. Symmetrisation also compensates for the incompleteness caused by the lack of citations to recent years and allows links between journals to be considered strong and subject-relevant even if these are asymmetric or even unidirectional. In order to reduce noise caused by 'small' journals and extremely weak citation links, thresholds have been applied to both citation links and number of papers (see previous section).

The text mining analysis supplements the citation analysis. In particular, the textual information is integrated with the bibliometric information before the clustering algorithm is applied. In the present study, the actual integration is achieved by weighted linear combination of the corresponding distance matrices. The methodology and advantages of hybrid clustering have been substantiated in more detail in earlier studies devoted to the analysis of different research fields (see Glenisson et al., 2005; Janssens et al., 2007a, 2007b, 2008). In addition, the lexical approach allows to 'label' clusters using automatically detected salient terms.

In Section 4.3, Silhouette and modularity curves will be used to compare results of text-based, citation-based and hybrid clustering, and we will substantiate that the hybrid method in general outperforms the other two.

### 3.4 Multidimensional scaling

Multidimensional scaling (MDS) can be used to represent high-dimensional vectors (for example, the centroids of journal clusters) in a lower dimensional space by explicitly requiring that the pairwise distances between the points approximate the original high-dimensional distances as precisely as possible (Mardia, 1979). If the dimensionality is reduced to two or three dimensions, these mutual distances can directly be visualised. It should, however, be stressed that interpretations concerning such a low-dimensional approximation of very high-dimensional distances must be handled with care.

## 4. Results

### 4.1 Evaluation of existing 'intellectual' subject-classification schemes

The multidisciplinary databases *Science Citation Index Expanded* (SCIE) and *Social Sciences Citation Index* (SSCI) of Thomson-Reuters (formerly Institute for Scientific Information, ISI, Philadelphia, PA, USA) traditionally did not provide a direct subject assignment for indexed papers. The annual Science Citations Index Guides, the Journal Citation Reports (JCR) and more recently the Website of Thomson Scientific, however, contain regularly updated lists of (S)SCI journals assigned to one or more subject matters (ISI Subject Categories) each. For lack of an appropriate subject-heading system, more or less modified versions of this Subject Category scheme were often used in bibliometric studies too, namely as an indirect subject assignment to individual papers based on the journals in which they had been published. Such assignment systems based on journal classification have been developed among others

by Narin and Pinski (see, for instance, Narin, 1976; Pinski & Narin, 1976). This was followed by classification schemes developed by other institutes as well. Nowadays two ISI systems are widely used, in particular, the ISI Subject Categories, which are available in the JCR and through journal assignment in the Web of Science as well, and the Essential Science Indicators (ESI).

Field #	ESI Field	Field #	ESI Field
1	Agricultural Sciences	12	Mathematics
2	Biology & Biochemistry	13	Microbiology
3	Chemistry	14	Molecular Biology & Genetics
4	Clinical Medicine	15	Multidisciplinary
5	Computer Science	16	Neuroscience & Behavior
6	Economics & Business	17	Pharmacology & Toxicology
7	Engineering	18	Physics
8	Environment/Ecology	19	Plant & Animal Science
9	Geosciences	20	Psychology/Psychiatry
10	Immunology	21	Social Sciences
11	Materials Sciences	22	Space Science

Table 1. The 22 broad science fields according to the *Essential Science Indicators* (ESI)

While the first system assigns multiple categories to each journal and is too fine grained (254 categories) for comparison with cluster analysis, the ESI scheme is forming a partition (with practically unique journal assignment) and the 22 fields are large enough. Therefore the ESI classification seems to be a good choice for our analysis.

Subject fields will be considered like automatically generated clusters. One precondition for easy comparison with results from hard clustering is that the reference classification system must form a partition of the WoS universe, while most schemes allow multiple assignments (e.g., the above-mentioned ISI Subject Categories). The only commonly known subject scheme for ISI products that meets the criterion is the ESI classification system. This subject classification scheme is in principle based on unique assignment; only about 0.6% of all journals were assigned to more than one field over a five-year period. For the present exercise, assignment has to be de-duplicated in the case of journals which merged or split up during the period of 5 years, declaredly a somewhat arbitrary procedure. Nonetheless, the assignment remains correct and results in no more than a slightly narrower scope for several journals. The field structure of the ESI scheme is presented in Table 1.

The question arises whether field classification according to the ESI scheme could still be improved. In particular, we will analyse whether journal assignments to fields can be considered optimum. Figure 1 presents the evaluation of the 22 ESI fields based on the cross-citation- (left) and text-based (right) Silhouette values (see Section 3.3.3). Several fields seem not to be consistent enough from both perspectives. Above all, the Silhouette values of field #2 (Biology & Biochemistry), #4 (Clinical Medicine), #7 (Engineering), #19 (Plant & Animal Science) and #21 (Social Sciences) substantiate that at least five of the 22 fields are not sufficiently consistent.

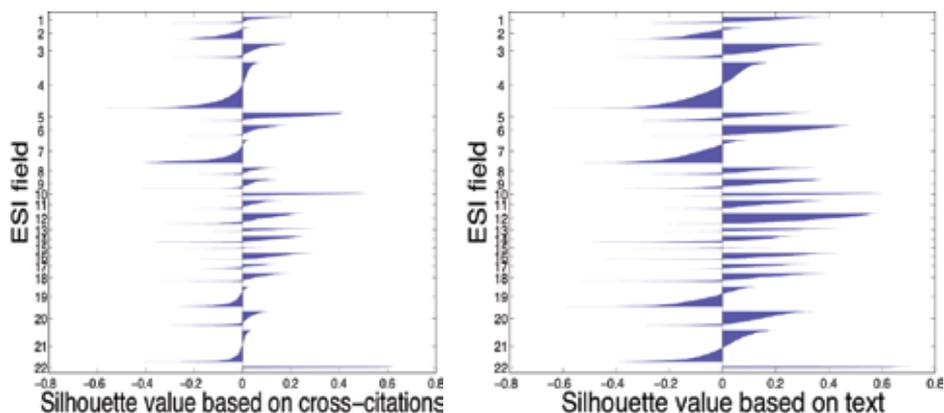


Fig. 1. Silhouette plot for 22 ESI fields based on journal cross-citations (left) and based on text (right)

#### 4.2 Labelling subject fields using cognitive characteristics and visualization of the cross-citation network

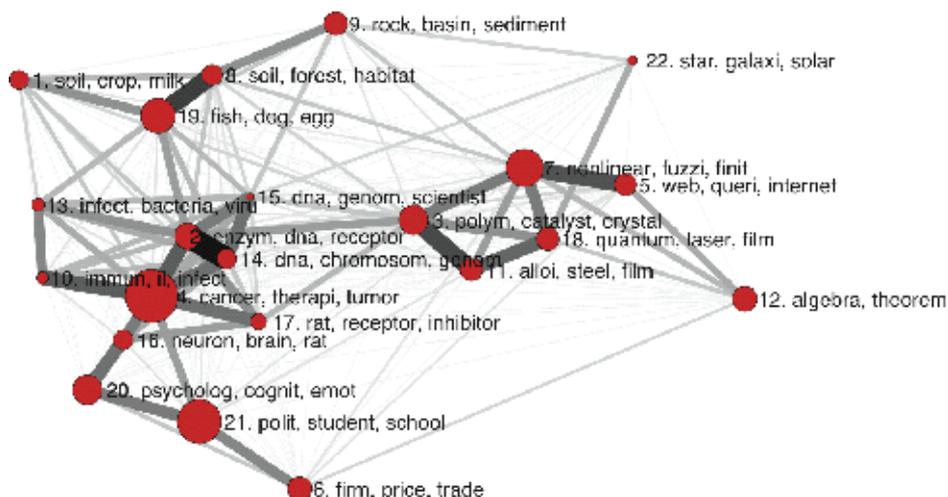


Fig. 2. Network of the 22 ESI fields based on cross-citation links

Simultaneously to the above validation, the textual approach also provides the best TF-IDF terms – out of a vocabulary of 669,860 terms – describing the individual fields. These terms are presented in Table 2. Although these terms already provide an acceptable characterisation of the topics covered by the 22 fields, considerable overlaps are apparent between pairs of fields, respectively: Engineering (#7) and Computer Science (#5), Chemistry (#3) and Materials Science (#11), Plant & Animal Science (#19) and Environment/Ecology (#8), as well as Biology & Biochemistry (#2), Molecular Biology & Genetics (#14) and Clinical Medicine (#4). In addition, the terms characterising the social sciences (#21) reflect a pronounced heterogeneity of the field. The structural map of the 22 ESI fields based on cross-citation links is presented in Figure 2. For the visualisation we used Pajek (Batagelj & Mrvar, 2002). The network map confirms the strong links we have found based on the best terms between fields #2 & #14, #3 & #11, #5 & #7, and #8 & #19, respectively.

Field	Best 50 terms
1	soil; crop; milk; fruit; seed; cultivar; wheat; dry; rice; ha; chees; diet; fat; ferment; nutrit; meat; farm; grain; starch; fertil; irrig; agricultur; dietari; intak; wine; flour; antioxid; sensori; fatti; sugar; juic; nutrient; moistur; harvest; maiz; veget; cook; leaf; soybean; nitrogen; farmer; season; vitamin; potato; weed; textur; dairi; bacteria; fresh; corn;
2	enzym; dna; receptor; rat; peptid; metabol; lipid; genom; insulin; muscl; transcript; ca2; amino; glucos; mutat; rna; molecul; diabet; kinas; inhibitor; hormon; mice; mrna; neuron; fluoresc; mutant; cancer; assai; serum; vitro; secret; bone; recombin; mitochondri; coli; brain; tumor; ligand; liver; antibodi; subunit; ion; apoptosi; yeast; intracellular; vivo; cholesterol; biologi; affin; calcium;
3	polym; catalyst; crystal; ion; bond; molecul; solvent; atom; ligand; hydrogen; film; polymer; adsorpt; aqueou; poli; nmr; methyl; spectroscopi; thermal; chemistri; bi; electro; spectra; cu; catalyt; cation; mol; copolym; anion; angstrom; amino; chiral; nm; ir; electrochem; salt; reactor; copper; chlorid; ionic; surfact; arom; ni; h2o; fluoresc; column; chromatographi; alkyl; cl; alcohol;
4	cancer; therapi; tumor; infect; surgeri; pain; hospit; arteri; syndrom; diabet; injuri; bone; lesion; chronic; symptom; surgic; renal; breast; carcinoma; serum; transplant; lung; mortal; muscl; liver; coronari; cardiac; physician; rat; hypertens; recurr; malign; pulmonari; receptor; oral; men; therapeut; postop; ci; hiv; vascular; mutat; ct; hepat; infant; diagnos; tumour; pregnanc; antibodi; il;
5	web; queri; internet; graph; schedul; wireless; semant; logic; node; busi; video; processor; traffic; execut; fuzzi; server; machin; packet; finit; fault; ltd; grid; hardwar; messag; cach; mesh; xml; multimedia; qo; bandwidth; custom; scalabl; bit; multicast; 3d; iter; java; ip; onlin; metric; platform; polynomi; retriev; neural; circuit; heurist; algebra; robot; topolog; broadcast;
6	firm; price; trade; economi; busi; capit; invest; wage; tax; financi; organiz; incom; bank; compani; sector; corpor; employ; stock; monetari; custom; labor; privat; strateg; welfar; incent; asset; profit; employe; polit; household; game; worker; inflat; job; union; foreign; brand; earn; forecast; labour; reform; export; unemploy; insur; retail; volatil; team; credit; pai; financ;
7	nonlinear; fuzzi; finit; machin; robot; sensor; motion; veloc; nois; crack; thermal; ltd; circuit; vehicl; neural; fuel; voltag; vibrat; elast; beam; shear; turbul; schedul; fault; deform; film; plane; stochast; iter; steel; compress; custom; wind; friction; actuat; concret; logic; soil; geometr; laser; graph; antenna; cylind; traffic; oscil; calibr; autom; geometri; grid; reactor;
8	soil; forest; habitat; river; sediment; ecolog; lake; pollut; land; ecosystem; climat; season; veget; fish; seed; landscap; biomass; nutrient; predat; agricultur; sludg; toxic; groundwat; bird; stream; wast; sea; island; wastewat; wetland; nitrogen; fire; ha; emiss; urban; coastal; flood; biodivers; reproduct; basin; nest; pesticid; seedl; crop; dry; microbi; watersh; graze; winter; rainfal;
9	rock; basin; sediment; sea; fault; ocean; miner; seismic; climat; isotop; earthquak; ic; tecton; ma; soil; southern; volcan; atmospher; mantl; geolog; wind; northern; reservoir; metamorph; precipit; river; cretac; lake; faci; eastern; assemblag; veloc; sedimentari; crust; melt; marin; continent; magma; or; deform; east; flux; granit; belt; fractur; shallow; earth; slope; cloud; clai;

Field	Best 50 terms
10	immun; il; infect; antigen; antibodi; mice; vaccin; receptor; cytokin; hiv; cd4; lymphocyt; ifn; autoimmun; dc; cd8; macrophag; viru; inflammatori; peptid; hla; mhc; tnf; nk; ig; molecul; tumor; lp; serum; tcr; pathogen; innat; assai; chemokin; dendrit; allergen; viral; igg; interleukin; monocyt; apoptosi; neutrophil; epitop; allerg; immunolog; secret; inflamm; dna; vitro; th2;
11	alloy; steel; film; coat; corros; glass; crack; microstructur; ceram; powder; fiber; grain; thermal; sinter; polym; crystal; deform; fabric; weld; fibr; fatigu; concret; fractur; si; specimen; cast; tensil; melt; cement; ni; silicon; shear; bond; microscopi; fe; ion; wear; adhes; cu; copper; nanoparticl; lamin; nanotub; aluminum; compress; roll; elast; creep; atom; al2o3;
12	algebra; theorem; finit; asymptot; infin; manifold; let; polynomi; graph; nonlinear; invari; omega; inequ; singular; lambda; convex; proof; compact; ellipt; conjectur; bar; epsilon; infinit; sigma; phi; symmetr; stochast; hyperbol; banach; topolog; metric; integ; matric; lie; exponenti; markov; curvatur; norm; eigenvalu; kernel; hilbert; cohomolog; geometr; quadrat; covari; dirichlet; semigroup; iter; parabol; theta;
13	infect; bacteria; viru; bacteri; pathogen; dna; genom; pcr; parasit; coli; enzym; mutant; yeast; microbi; viral; hiv; rna; vaccin; immun; encod; virul; antibiot; transcript; sp; assai; escherichia; virus; plasmid; clone; candida; 16; soil; biofilm; antibodi; microorgan; fungal; amino; antigen; bacillu; recombin; fungi; albican; gram; mutat; phyloenet; mice; pseudomona; ferment; rrna; genotyp;
14	dna; chromosom; genom; transcript; mutat; receptor; kinas; mous; mice; rna; allel; mutant; apoptosi; cancer; mrna; rat; phenotyp; muscl; polymorph; embryo; tumor; drosophila; phosphoryl; ca2; neuron; actin; clone; encod; prolifer; mitochondri; enzym; genotyp; vitro; assai; vivo; il; embryon; epitheli; recombin; pcr; chromatin; mammalian; regulatori; linkag; transgen; loci; delet; haplotyp; homolog; yeast;
15	dna; genom; scientist; receptor; brain; soil; climat; earth; molecul; neuron; rna; chromosom; mice; mutat; africa; transcript; biologi; ocean; infect; fossil; india; sea; evolutionari; rock; fuel; logic; southern; island; enzym; marin; insect; fluoresc; cancer; quantum; sediment; scienc; bone; viru; australia; immun; ecolog; fish; china; atmospher; your; mind; rat; bird; ic; colour;
16	neuron; brain; rat; receptor; cortex; motor; cognit; cortic; cerebr; mice; neural; stroke; sleep; nerv; lesion; synapt; seizur; epilepsi; axon; schizopfhrenia; hippocamp; spinal; symptom; pain; alzheimer; hippocampu; dopamin; injuri; parkinson; neurolog; deficit; syndrom; eeg; nervou; sensori; stimuli; dementia; ms; stimulu; glutam; muscl; nucleu; astrocyt; chronic; gaba; frontal; sclerosi; auditori; cord; alcohol;
17	rat; receptor; inhibitor; toxic; therapeut; cancer; metabol; vitro; mice; liver; pharmacokinet; oral; therapi; pharmaceut; enzym; antagonist; assai; vivo; pharmacolog; dna; tablet; inflammatori; tumor; metabolit; lipid; brain; agonist; diabet; cytotox; antioxid; kinas; lung; peptid; apoptosi; ca2; serum; administ; molecul; potent; chronic; insulin; mug; mum; liposom; p450; renal; hepat; inhibitori; immune; ligand;

Field	Best 50 terms
18	quantum; laser; film; beam; spin; atom; scatter; crystal; ion; nonlinear; excit; photon; latic; noise; thermal; oscil; dope; symmetri; veloc; emiss; finit; decai; spectra; wavelength; si; diffract; neutron; nm; plane; acoust; fiber; hole; superconduct; motion; spectral; dielectr; collis; coher; glass; semiconductor; neutrino; perturb; detector; algebra; elast; soliton; waveguid; relativist; amplitud; alloy;
19	fish; dog; egg; forest; genu; breed; habitat; seed; infect; diet; sp; season; larva; reproduct; leaf; bird; nest; hors; cow; soil; predat; sea; cat; taxa; flower; fruit; veget; parasit; pig; milk; seedl; prei; mate; shoot; cattl; southern; trait; genera; fed; island; nov; ecolog; lake; insect; pollen; viru; river; juvenil; farm; pathogen;
20	psycholog; cognit; emot; student; mental; adolesc; anxieti; symptom; school; item; child; psychiatr; gender; sexual; attitud; cope; mother; interview; schizophrenia; suicid; skill; questionnaire; belief; abus; therapi; men; word; psychotherapi; aggress; mood; verbal; teacher; cue; stimuli; satisfact; judgment; job; infant; development; violenc; trait; ptsd; stimulu; style; interperson; peer; prime; esteem; distress; recal;
21	polit; student; school; teacher; gender; urban; nurs; court; reform; war; legal; discours; profession; parti; disabl; interview; capit; rural; attitud; child; ethnic; privat; welfar; democraci; democrat; ethic; employ; justic; feder; violenc; worker; agenc; teach; sexual; economi; incom; academ; immigr; sociolog; moral; african; skill; mental; librari; men; sector; land; crime; china; civil;
22	star; galaxi; solar; orbit; radio; telescop; emiss; stellar; veloc; disk; galact; earth; planet; flux; atmospher; satellit; wind; mar; cosmic; binari; cloud; flare; dust; spectral; luminos; redshift; jet; accret; dwarf; planetari; cosmolog; mission; motion; observatori; burst; spectra; photometr; gravit; comet; sun; bright; infrar; grb; shock; ngc; dark; supernova; spacecraft; radial; halo;

Table 2. The best 50 TF-IDF terms describing the 22 ESI fields

#### 4.3 Cluster analysis: text-based, citation-based and hybrid

Figure 3 compares the performance of text-based, cross-citation and hybrid clustering by several evaluation methods, for various numbers of clusters. For each of the three clustering types, Figure 3(1) presents for various cluster numbers (2 to 30) the modularity calculated from the journal cross-citation graph. Since this evaluation is based on cross-citation data, it is not a surprise that the text-only clustering provides worse results than cross-citation clustering, which performs best here. However, very interesting to note is that the hybrid clustering (integrated text and cross-citation information) provides results highly comparable to those from cross-citation clustering, especially for 7 or more than 12 clusters. The modularity scores for cross-citation clustering indicate that any number of clusters larger than 9 is acceptable. On the other hand, the modularity curve for text-only clustering contains a maximum for eight clusters.

In Figure 3(2), Silhouette curves based on (the complement of) cross-citation values show the somewhat counter-intuitive but beneficial result that hybrid clustering always performs better than cross-citation clustering, although the evaluation only considers citations here. This again demonstrates the power of hybrid clustering: the combined heterogeneous

citation–textual approach is superior to both methods applied separately. Nevertheless, this figure does not provide a clear clue with respect to the number of clusters to choose.

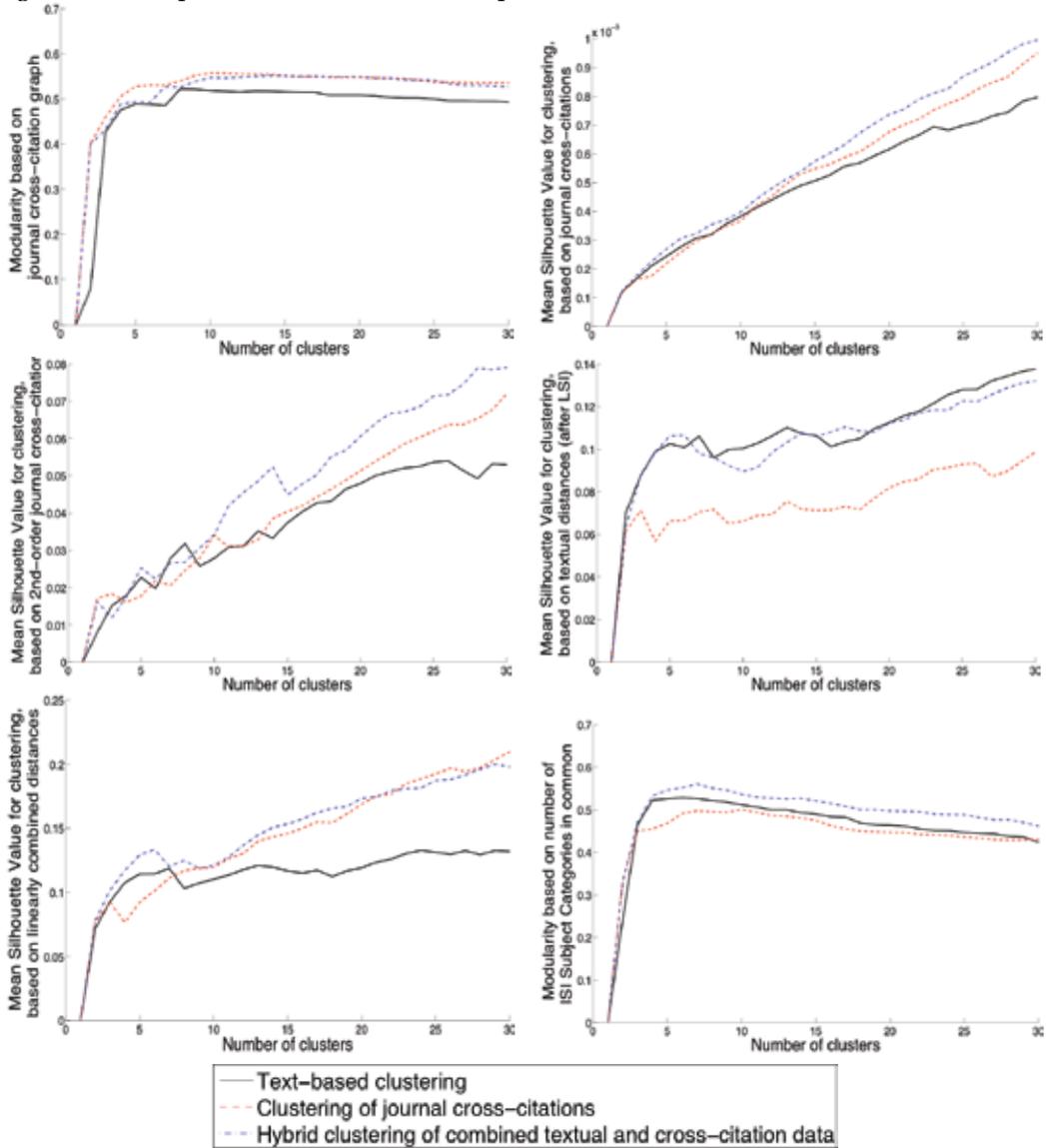


Fig. 3. Performance evaluation of text-based, citation-based and hybrid clustering based on (1) modularity calculated from the journal cross-citation graph, and based on Silhouette curves calculated from (2) journal cross-citations, (3) second-order journal cross-citations, (4) text-based distances, and (5) linearly combined distances. For an additional ‘external validation’ of clustering results compared to ISI Subject Categories, the lower-right figure (6) uses modularity computed from a network containing all journals as nodes, but with edge weights equal to the number of ISI Subject Categories commonly assigned to the corresponding journals by ISI/Thomson Scientific (out of the total of 254 categories).

Silhouette curves based on the complement of second-order cross-citations are shown in Figure 3(3). Again, the hybrid clustering almost always performs best.

In Figure 3(4), the Silhouette values are computed only from textual distances. Naturally, the citation-based clustering performs worst here, while the integrated clustering scores almost as good as the text-only clustering and for some cluster numbers even better.

Figure 3(5) shows Silhouette curves based on linearly combined text-based and citation-based distances (with equal weight). Here, combined data and mere citations give comparable results, which might be an indication that there is a preponderance of citation over text data in the combined Silhouette values.

Finally, Figure 3(6) provides an *external validation* of clustering results by expert knowledge available in the ISI Subject Categories assigned to journals by ISI/Thomson Scientific. The modularity curves are computed from a network containing all journals as nodes, but with edge weights equal to the number of ISI Subject Categories in common (out of the total of 254 categories). Again very interesting to see is that hybrid clustering outperforms text-only and citation-based clustering. The optimal number of clusters according to this type of evaluation is 7.

	Modularity based on journal cross-citation graph	Modularity based on common ISI Subject Categories	MSV based on textual distances	MSV based on 2 <sup>nd</sup> order journal cross-citations	MSV based on linearly combined distances	Rand index with 22 ESI fields as reference classification
22 ESI fields	0.47533	<b>(0.52604)</b>	0.057237	0.016017	0.062807	(1)
22 citation-based clusters	<b>0.54676</b>	0.44244	0.09319	<b>0.057337</b>	<b>0.18938</b>	0.90463
22 text-based clusters	0.50451	0.45091	0.11829	0.035447	0.12987	0.90582
22 Hybrid clusters	<b>0.54677</b>	<b>0.48839</b>	<b>0.1206</b>	0.05453	<b>0.18951</b>	<b>0.90867</b>

Table 3. Evaluation of 22 ESI fields and 22 citation-based, text-based and hybrid clusters by modularities and mean Silhouette values (MSV). Highest values in each column are shown in bold.

In Table 3 we compare the quality of the partition of 22 ESI fields with the quality of the 22 clusters resulting from citation-based, text-based and hybrid clustering. The only evaluation measure for which the 22 human-made ESI fields score best is modularity based on ISI Subject Categories. As already explained before, this evaluation type computes modularity from a network containing all journals as nodes and with edge weights equal to the number of ISI Subject Categories commonly assigned to the corresponding journals by ISI/Thomson Scientific (out of the total of 254 categories). Since there is a direct correspondence between the 22 ESI fields and these 254 Subject Categories (a field is an aggregation of multiple subject categories), it is not at all surprising (not to mention unfair) that the ESI fields outperform the clusters for this type of evaluation. For all other data-driven evaluation types it is clear that automatic clustering does better than human expert classification.

Hybrid clustering always performs at least as good as text-based or citation-based clustering, except for evaluation by second order cross-citations. However, small the difference, the last column shows that the 22 hybrid clusters correspond best to the 22 ESI fields. It should be noted that the values in Table 3 can differ somewhat from the values in Figure 3 because, for the sake of a fair comparison with ESI fields, in the table only 7729 journals were considered for which a field assignment was available.

#### 4.4 Evaluation of hybrid clusters

The cluster dendrogram shows the structure in a hierarchical order (see Figure 4). We visually find a first clear cut-off point at three clusters, a second one around seven, and 22 clusters also seemed to be an acceptable/ appropriate number. This value coincides with the number of fields according to the ESI classification scheme. The Silhouette plots in Figure 5 and the mean Silhouette values in Table 3 substantiate that the 22 hybrid clusters are furthermore acceptable for both the citation and the text-mining approach. The same conclusion can be drawn from computed modularity scores.

The number of three clusters results in an almost trivial classification. Intuitively, these three high-level clusters should comprise natural and applied sciences, medical sciences, and social sciences and humanities. The solutions with 3 and 22 clusters will be analysed in more detail in Section 4.5. The solution comprising of seven clusters results in a non-trivial classification. The best TF-IDF terms (see Table 5) show that three of these clusters represent the natural/applied sciences, whereas two classes each stand for the life sciences and the social sciences and humanities. This situation is also reflected by the cluster dendrogram in Figure 4. A closer look at the best TF-IDF terms reveals that social sciences cluster (#1 of the 3-cluster solution) is split into the cluster #1 (economics, business and political science) and #6 (psychology, sociology, education), the life-science cluster (#3 in the 3-cluster scheme) is split into clusters #3 (biosciences and biomedical research) and #7 (clinical, experimental medicine and neurosciences) and, finally, the sciences cluster #2 of the 3-cluster scheme is distributed over three clusters in the 7-cluster solution, particularly, the cluster comprising biology, agriculture and environmental sciences (#2), physics, chemistry and engineering (#4) as well as mathematics and computer science (#5).

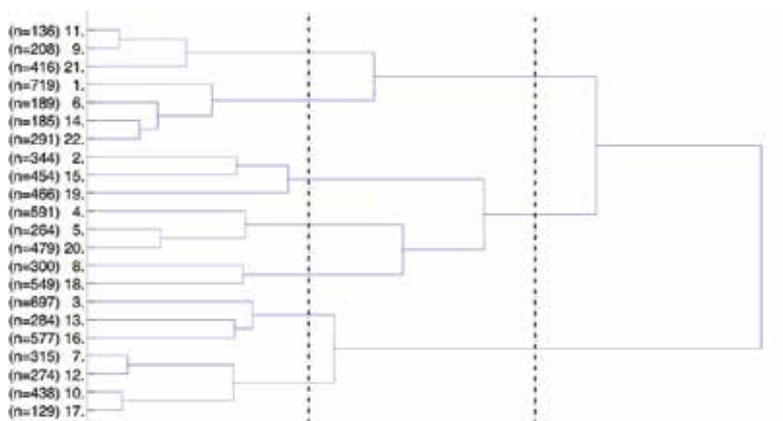


Fig. 4. Cluster dendrogram for hybrid hierarchical clustering of 8305 journals, cut off at 22 clusters on the left-hand side. Two other vertical lines indicate the cut-off points for 7 and 3 clusters.

The hybrid, i.e. the combined citation-textual based clustering yields acceptable results (see Figure 5), and is distinctly superior to both methods applied separately. Nonetheless, we must not conceal that we can also find clusters of lesser quality, notably cluster #1, in the hybrid classification.

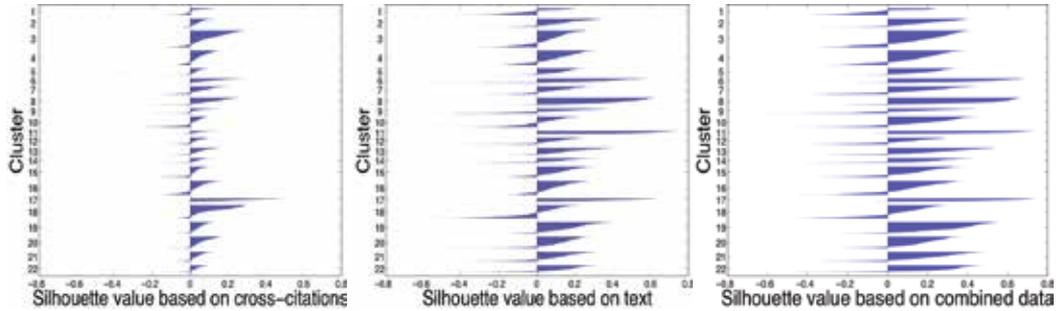


Fig. 5. Evaluation of the hybrid clustering solution with 22 clusters by citation based Silhouette plot (left), text based Silhouette plot (centre) and the plot with Silhouette values based on combined data (right).

#### 4.5 Cognitive characteristics of clusters

As already mentioned in the previous section, another nice point to cut off the dendrogram is at three clusters (cf. the right-most vertical line in Figure 4). Although this refers to a rather trivial case, it might be worthwhile to have a look at term representation of this structure before we deal with ‘labelling’ the 22 clusters that we have obtained from the hybrid algorithm. This will also help us to understand the hierarchical architecture of the subject structure of science. Table 4 lists the best 50 terms for each of the three top-level clusters which definitely confirm the presence of the expected clusters. Indeed, cluster #1 comprises the social sciences, cluster #2 the natural and applied sciences and cluster #3 the medical sciences. The distribution of journals over clusters is surprisingly well-balanced.

Cluster (# journals)	Best 50 terms
1 (n=2144)	polit; student; school; firm; cognit; psycholog; war; gender; price; emot; mental; capit; teacher; trade; economi; reform; adolesc; child; busi; discours; attitud; urban; skill; court; organiz; moral; text; employ; privat; interview; narr; profession; sexual; parti; legal; incom; english; job; music; anxieti; invest; german; welfar; academ; belief; write; sector; violenc; religi; teach
2 (n=3447)	soil; finit; film; nonlinear; thermal; ion; crystal; algebra; polym; ltd; forest; atom; veloc; sediment; laser; quantum; motion; graph; theorem; seed; alloi; asymptot; deform; sea; fish; bond; coat; grain; sensor; beam; polynomi; hydrogen; fiber; fault; machin; season; emiss; crack; fuzzi; shear; habitat; nois; steel; dry; plane; fe; catalyst; elast; sp; glass
3 (n=2714)	cancer; infect; therapi; tumor; receptor; rat; dna; pain; diabet; mice; bone; brain; muscl; hospit; syndrom; chronic; injuri; mutat; surgeri; serum; lesion; arteri; neuron; immun; liver; hiv; il; symptom; antibodi; metabol; inhibitor; renal; enzym; breast; surgic; lung; therapeut; mortal; vaccin; genom; transcript; nurs; assai; transplant; inflammatori; peptid; insulin; cardiac; carcinoma; oral

Table 4. Best 50 TF-IDF terms describing the 3 top-level clusters

According to the terms, economics, business and psychology are the dominant issues in the first cluster which represents the social sciences. The most characteristic terms of the second cluster represent the full spectrum of the sciences including mathematics, geosciences and engineering. Also some subfields of agriculture & environment are covered. Cluster #3, finally, covers biosciences, biomedical research, clinical & experimental medicine and neurosciences.

Cluster (# journals)	Best 50 terms
1 (n=1384)	polit; firm; war; price; trade; economi; capit; busi; reform; urban; court; parti; gender; privat; invest; organiz; sector; corpor; employ; moral; labor; legal; incom; financi; discours; tax; music; compani; contemporari; welfar; essai; union; foreign; democraci; job; land; wage; civil; china; labour; book; narr; worker; democrat; german; school; liber; internet; text; religi
2 (n=1264)	soil; forest; sediment; fish; seed; habitat; sea; season; river; lake; sp; basin; rock; genu; veget; crop; leaf; climat; southern; ecolog; egg; land; ocean; fruit; dry; island; biomass; northern; miner; nutrient; predat; marin; reproduct; nest; larva; bacteria; taxa; winter; cultivar; ha; nitrogen; ecosystem; seedl; eastern; ic; atmosphere; flower; breed; wheat; bird
3 (n=1558)	cancer; infect; tumor; receptor; dna; rat; therapi; mice; mutat; immun; il; antibodi; liver; serum; genom; enzym; transcript; hiv; diabet; assai; inhibitor; viru; antigen; vaccin; peptid; apoptosi; metabol; carcinoma; lung; renal; chromosom; bone; kinas; breast; vitro; chronic; muscl; mrna; therapeut; transplant; syndrom; insulin; dog; inflammatori; hepat; lesion; rna; pcr; diet; molecul
4 (n=1334)	film; ion; crystal; polym; thermal; atom; alloy; laser; bond; coat; quantum; beam; steel; hydrogen; catalyst; crack; glass; fiber; molecul; nm; spectroscopi; spectra; veloc; ltd; finit; cu; vibrat; solvent; deform; electro; shear; powder; spin; elast; fabric; adsorpt; si; nonlinear; excit; sensor; fuel; fe; poli; polymer; diffract; emiss; aqueou; ni; nmr; corros
5 (n=849)	algebra; finit; nonlinear; graph; theorem; asymptot; polynomi; fuzzi; infin; manifold; let; invari; stochast; schedul; inequ; convex; robot; singular; proof; logic; omega; machin; iter; topolog; noise; traffic; infinit; metric; motion; lambda; web; compact; epsilon; neural; integ; circuit; symmetr; ellipt; bar; fault; node; matric; geometr; markov; sigma; exponenti; queri; custom; wireless; video
6 (n=760)	student; school; cognit; psycholog; teacher; mental; adolesc; emot; child; symptom; anxieti; gender; psychiatr; skill; attitud; abus; teach; item; word; interview; disabl; mother; schizophrenia; sexual; alcohol; speech; instruct; belief; cope; english; profession; questionnaire; suicid; violenc; classroom; verbal; youth; academ; peer; therapi; men; development; semant; stimuli; discours; linguist; phonolog; deficit; infant; offend
7 (n=1156)	pain; therapi; hospit; injuri; arteri; nurs; brain; surgeri; neuron; symptom; physician; syndrom; muscl; bone; diabet; rat; lesion; coronari; chronic; stroke; cancer; mortal; cardiac; surgic; receptor; infect; nerv; hypertens; men; infant; implant; cognit; ct; ey; cerebr; smoke; pregnanc; fractur; tumor; mri; cardiovascular; elderli; ci; motor; spinal; sleep; oral; questionnaire; myocardi; vascular

Table 5. Best 50 TF-IDF terms describing the 7 top-level clusters

The 50 best TF-IDF terms describing the 22 hybrid citation–lexical clusters are listed in Table 6. Cluster #1 of the 3-cluster scheme is split up in seven clusters, particularly, in #1, #6, #9, #11, #14, #21 and #22. However, this sub-classification of the social sciences is less straightforward. Cluster #6 represents economics and business and political science, cluster #9 stands for psychology and linguistics, cluster #21 covers psychology and psychiatry, #11 comprises sociology and education, and cluster #1 is rather focussed on the humanities. Cluster #14 and #22 seem to have more heterogeneous profiles among these ‘social and humanity clusters’. Although cluster #14 largely covers information and library science, the terms reflect a large overlap with other clusters. The same applies to cluster #22, which has obviously an even fuzzier structure. On the other hand, #9 and #21 are both covering psychology but focussing on different aspects, namely cognitive (#9) and medical (#21) issues.

Similarly to the structural analysis in Section 4.2, we use now a network with citation links among clusters to study the relationship of clusters based on hybrid cross-citation/textual information (see Figure 6). The observations made on the basis of most characteristic terms are confirmed by the link structure. The social-sciences and humanities clusters form two groups that are each strongly interlinked; one consists of clusters #1, #6, #14 and #22 with focus on humanities, economics, business, political and library science, the other one comprises #9, #11 and #21 with sociology, education and psychology. This is in line with the hierarchical structure shown in Figure 4. These two groups correspond to the two social-sciences clusters in the 7-cluster solution (cf. Section 4.4).

In the natural and applied sciences, we have found eight clusters, particularly, #2, #4, #5, #8, #15, #18, #19 and #20. On the basis of the most important TF-IDF terms (see Table 6) we can assign clusters #2, #15 and #19 to geosciences, environmental science, biology and agriculture, which, in turn, form a larger group corresponding to the first of the three “mega-clusters” in the 7-cluster solution. The graphic network presentation in Figure 6 confirms this interpretation. These three clusters form a group at the bottom. The other sciences clusters are more clearly recognisable, and distinctly separate fields. Thus clusters #4 represents chemistry, #20 physics, #5 engineering, #8 mathematics, and cluster #18 computer science. These science clusters form two groups, #4, #20 and #5 form one group of chemistry, physics and engineering, while #8 and #18 form the third group comprising mathematics and computer science. The network presentation and the hierarchical architecture in the dendrogram confirm the term characterisation.

The interpretation of the most characteristic terms of the life-science clusters is somewhat more complicated. Here we have a biomedical and a clinical group. These two groups are in line with the hierarchical structure of the dendrogram in Figure 4 but less clearly distinguished in the graphical network presentation (Figure 6). Nonetheless, the terms provide an excellent description for at least some of the medical clusters: cluster #7 stands for the neuro- and behavioural sciences, #3 for bioscience, #10 for the clinical and social medicine, #13 microbiology and veterinary science, #12 non-internal medicine, #16 hematology and oncology and #17 cardiovascular and respiratory medicine. According to the dendrogram clusters 3, 13, 16 and clusters 7, 10, 12, 17 form one larger cluster each. On the basis of the best terms, we can characterise these groups as the bioscience–biomedical and the clinical and neuroscience group, respectively.

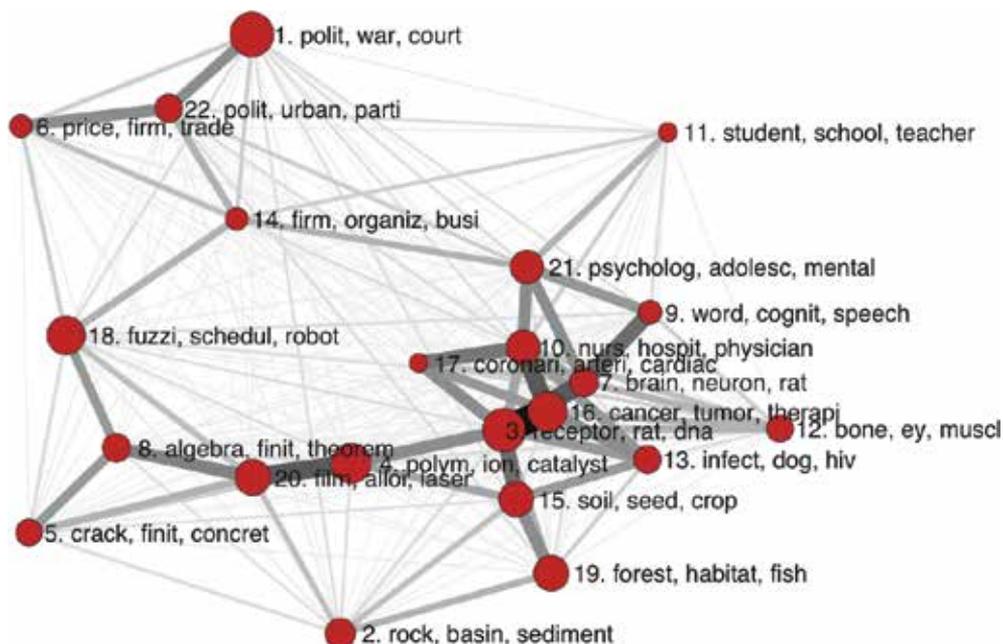


Fig. 6. Network structure of hybrid clusters represented by the three most important TF-IDF terms

Cluster	50 best terms
1	polit; war; court; music; moral; essai; legal; philosophi; narr; text; literari; book; contemporari; french; religi; write; german; discours; ethic; civil; reform; christian; philosoph; justic; fiction; coloni; nineteenth; archaeolog; religion; aesthet; british; english; poetri; stori; feder; truth; church; russian; artist; liber; revolut; historian; gender; roman; america; militari; god; democraci; ideolog; china;
2	rock; basin; sediment; fault; sea; climat; soil; ic; miner; ocean; seismic; atmospher; river; wind; isotop; veloc; earth; star; tecton; earthquak; solar; precipit; ma; volcan; mantl; southern; lake; satellit; geolog; cloud; land; northern; groundwat; metamorph; flux; rainfal; shear; deform; forecast; weather; melt; crust; slope; faci; flood; sedimentari; clai; galaxi; season; magma;
3	receptor; rat; dna; genom; enzym; transcript; mutat; mice; metabol; peptid; diabet; cancer; insulin; chromosom; kinas; inhibitor; lipid; ca2; muscl; mrna; rna; neuron; molecul; vitro; apoptosi; mous; liver; tumor; glucos; assai; brain; hormon; mutant; amino; vivo; serum; mitochondri; embryo; fluoresc; diet; secret; phosphoryl; therapeut; bone; phenotyp; polymorph; prolifer; toxic; therapi; antibodi;
4	polym; ion; catalyst; crystal; bond; molecul; film; solvent; atom; hydrogen; ligand; nmr; polymer; poli; aqueou; adsorpt; thermal; methyl; spectroscopi; spectra; copolym; cation; fiber; cu; nm; bi; coat; mol; blend; nanoparticl; anion; chemistri; catalyt; ms; electro; resin; chromatographi; ir; surfact; silica; copper; gel; amino; salt; column; uv; spectrometri; chiral; angstrom; fluoresc;

Cluster	50 best terms
5	crack; finit; concret; veloc; elast; turbul; vibrat; shear; thermal; nonlinear; beam; deform; fuel; motion; ltd; steel; cylind; combust; convect; flame; fatigu; compress; fractur; vehicl; jet; reynold; plane; wind; stiff; pipe; buckl; shell; friction; damp; vortex; cool; turbin; coal; fire; blade; bend; porou; lamin; axial; reservoir; rotor; specimen; cement; actuat; mesh;
6	price; firm; trade; tax; economi; capit; incom; wage; invest; bank; financi; monetari; stock; welfar; labor; inflat; sector; privat; incent; household; earn; game; asset; employ; insur; forecast; reform; foreign; unemploy; volatil; profit; worker; polit; labour; fiscal; corpor; debt; monei; credit; investor; busi; export; financ; shock; macroeconom; fund; currenc; equiti; school; agricultur;
7	brain; neuron; rat; stroke; lesion; receptor; pain; cerebr; ct; mri; motor; cognit; injuri; spinal; nerv; mr; seizur; epilepsi; cortex; neurolog; tumor; arteri; dementia; sleep; cortic; syndrom; muscl; therapi; symptom; parkinson; neural; cord; alzheim; mice; synapt; chronic; axon; aneurysm; patholog; pet; eeg; nervou; surgeri; elderli; surgic; sclerosi; ms; deficit; rehabilit; sensori;
8	algebra; finit; theorem; manifold; infin; nonlinear; polynomi; let; graph; asymptot; singular; omega; invari; inequ; lambda; ellipt; convex; compact; conjectur; epsilon; hyperbol; infinit; proof; bar; symmetr; phi; sigma; topolog; banach; lie; eigenvalu; metric; matric; curvatur; perturb; integ; norm; parabol; cohomolog; hilbert; geometr; plane; lattic; semigroup; explicit; stochast; iter; dirichlet; exponenti; holomorph;
9	word; cognit; speech; semant; english; linguist; phonolog; stimulu; stimuli; lexic; cue; sentenc; speaker; verb; prime; perceptu; student; acoust; text; item; discours; verbal; auditori; emot; recal; syntact; brain; hear; skill; deficit; write; motor; judgment; listen; nois; letter; noun; mental; learner; psycholog; neuropsycholog; voic; instruct; vowel; motion; german; execut; ey; grammar; grammat;
10	nurs; hospit; physician; therapi; cancer; pain; mortal; pregnanc; infant; symptom; ethic; smoke; diabet; infect; birth; ci; men; interview; hiv; injuri; profession; chronic; syndrom; questionnair; worker; school; adolesc; student; surgeri; mental; child; neonat; sexual; breast; healthcar; caregiv; visit; matern; elderli; mother; satisfact; hypertens; vaccin; attitud; rural; cohort; doctor; cardiac; staff; gender;
11	student; school; teacher; teach; classroom; instruct; skill; academ; curriculum; literaci; disabl; learner; profession; colleg; cognit; peer; child; faculti; gender; reform; write; psycholog; pupil; graduat; attitud; undergradu; text; emot; interview; belief; discours; pedagog; think; gift; adolesc; preschool; english; inquiri; elementari; girl; boi; development; leadership; pedagogi; polit; web; tutor; team; item; intellig;
12	bone; ey; muscl; sport; athlet; pain; implant; surgeri; fractur; injuri; knee; dental; hip; surgic; nerv; anterior; retin; postop; oral; tendon; corneal; teeth; flap; periodont; graft; lesion; ocular; posterior; therapi; radiograph; ligament; neck; nasal; fixat; cartilag; dentin; glaucoma; laser; femor; shoulder; cari; player; ankl; cement; acuiti; tooth; syndrom; symptom; arthroplasti; rehabilit;

Cluster	50 best terms
13	infect; dog; hiv; vaccin; viru; hors; cow; milk; parasit; cat; pig; cattl; antibodi; diet; immun; pcr; calv; breed; viral; herd; sheep; pathogen; serum; antigen; therapi; farm; malaria; antibiot; veterinari; hospit; assai; egg; dairi; dna; bird; genotyp; chicken; pneumonia; hepat; fed; tuberculosi; bovin; goat; mortal; lesion; epidemiolog; outbreak; canin; virus; respiratori;
14	firm; organiz; busi; librari; web; internet; custom; compani; employe; job; onlin; brand; team; strateg; journal; career; corpor; satisfact; student; price; trust; advertis; academ; profession; librarian; attitud; organis; leadership; cognit; enterpris; commerc; invest; sector; manageri; financi; psycholog; polit; retail; skill; commit; interview; emot; ventur; capit; purchas; intent; citat; book; retriev; text;
15	soil; seed; crop; cultivar; leaf; fruit; bacteria; wheat; dry; rice; enzym; pathogen; shoot; nitrogen; microbi; ferment; bacteri; ha; pollut; sediment; milk; nutrient; dna; fertil; germin; biomass; seedl; season; agricultur; coli; sludg; irrig; veget; infect; wast; grain; flower; co2; yeast; toxic; fungi; starch; genom; maiz; sp; grown; sugar; inocul; mutant; forest;
16	cancer; tumor; therapi; il; carcinoma; transplant; breast; immun; infect; lung; liver; renal; antibodi; receptor; antigen; mice; malign; serum; chronic; prostat; lesion; surgeri; tumour; chemotherapi; mutat; inflammatori; dna; bone; recurr; surgic; cytokin; syndrom; hepat; apoptosi; biopsi; lymphoma; lymphocyt; pancreat; gastric; resect; therapeut; rat; histolog; symptom; assai; prolifer; invas; inhibitor; asthma; median;
17	coronari; arteri; cardiac; ventricular; myocardi; hypertens; cardiovascular; aortic; atrial; infarct; diabet; therapi; valv; vascular; stent; endotheli; surgeri; pulmonari; mortal; cholesterol; systol; bypass; syndrom; lv; graft; diastol; renal; vein; rat; echocardiographi; ischemia; hospit; chronic; dysfunct; receptor; angiotensin; aneurysm; af; fibril; atherosclerosi; mitral; venou; perfus; ischem; serum; reperfus; implant; inhibitor; ldl; vessel;
18	fuzzi; schedul; robot; logic; machin; graph; nonlinear; traffic; web; asymptot; circuit; neural; nois; finit; stochast; fault; queri; custom; wireless; video; node; semant; heurist; antenna; motion; markov; polynomi; bayesian; iter; processor; sensor; covari; busi; execut; bandwidth; server; vehicl; internet; packet; wavelet; voltag; queue; alloc; algebra; intellig; bit; hardwar; theorem; ltd; paramet;
19	forest; habitat; fish; genu; egg; predat; season; sp; nest; sea; ecolog; larva; reproduct; lake; bird; prei; island; taxa; seed; river; veget; soil; breed; southern; ecosystem; nov; mate; genera; diet; biomass; insect; phylogenet; parasit; northern; marin; juvenil; forag; sediment; landscap; larval; winter; coastal; ocean; eastern; nutrient; summer; leaf; land; fisheri; assemblag;
20	film; alloy; laser; quantum; crystal; ion; steel; thermal; atom; beam; coat; glass; si; grain; microstructur; corros; silicon; dope; spin; ceram; powder; nm; scatter; fabric; neutron; diffract; dielectr; photon; cu; electro; excit; ni; fe; emiss; sinter; deform; microscopi; fiber; voltag; hydrogen; sensor; anneal; spectra; lattic; spectroscopi; weld; semiconductor; nonlinear; machin; discharg;

Cluster	50 best terms
21	psycholog; adolesc; mental; emot; cognit; symptom; child; anxiety; psychiatr; abus; student; school; alcohol; schizophrenia; sexual; mother; gender; attitud; suicid; interview; cope; violenc; therapi; questionnair; youth; disabl; offend; men; belief; item; psychotherapi; aggress; mood; ptsd; client; satisfact; victim; peer; profession; distress; development; infant; interperson; crime; adhd; style; therapist; esteem; skill; childhood;
22	polit; urban; parti; gender; reform; economi; capit; democrat; democraci; employ; sector; war; land; sociolog; geographi; union; labour; rural; elect; welfar; ethnic; labor; discours; immigr; privat; actor; trade; civil; poverti; firm; citizen; busi; china; worker; incom; feminist; vote; liber; eu; household; elector; contemporari; agenc; job; inequ; domest; foreign; ideolog; agricultur; organiz;

Table 6. The 50 best TF-IDF terms describing the 22 hybrid citation-lexical clusters

In order to gain a better understanding of the cluster structure, we have ranked the journals of each of the 22 clusters according to a modified version of Google's PageRank algorithm (Brin & Page, 1998) in which the number of citations is taken into account, normalised by the number of published papers. The following equation was used,

$$PR_i = \frac{(1-\alpha)}{n} + \alpha \sum_j PR_j \frac{a_{ji}/P_i}{\sum_k \frac{a_{jk}}{P_k}} \quad (1),$$

where  $PR_i$  is the PageRank of journal  $i$ ,  $\alpha$  is a scalar between 0 and 1 ( $\alpha=0.9$  in our implementation),  $n$  is the number of journals in the cluster,  $a_{ji}$  the number of citations from journal  $j$  to journal  $i$ , and  $P_i$  is the number of papers published by journal  $i$ , all in the period under study. Both sums iterate over the journals in the same cluster that contains journal  $i$ . Journal self-citations were removed prior to application of the algorithm. The five journals with highest PageRank are presented in Table 7. The PageRank of a journal can be understood here as the probability that a random reader will be reading that journal, when he randomly, continuously, and with equal probability looks up cited references to other journals (different from the current one), but once in a while randomly picks another journal from the library (cluster). Journals from arts & humanities (according to the ISI Subject Categories) were removed prior to application of the PageRank algorithm because of the low reliability of citation indicators in these disciplines. Zhang and Glänzel (2008) have shown that high entropy of journal cross-citations, relatively low impact and high share of journal self-citation makes it difficult to build reliable citation indicators for the humanities. This has to do with the subject-specific peculiarities in scholarly communication.

In general, the journals ranking highest represent their cluster in an adequate manner (cf. Table 7). Results of the PageRanking thus provide a realistic and representative picture of the hybrid clustering.

#### 4.6 Comparison of subject and cluster structure

In this subsection we compare the structure resulting from the hybrid clustering with the ESI subject classification. This comparison is based on the *centroids* of the clusters and fields.

The centroid of a cluster or field is defined as the linear combination of all documents in it and is thus a vector in the same vector space. For each cluster and for each field, the centroid was calculated and the MDS of pairwise distances between all centroids is shown in Figure 7.

Cluster 1	Cluster 2	Cluster 3	Cluster 4
1. YALE LAW J 2. HARVARD LAW REV 3. STANFORD LAW REV 4. AM HIST REV 5. COLUMBIA LAW REV	1. ANNU REV ASTRON ASTR 2. ASTROPHYS J SUPPL S 3. EARTH-SCI REV 4. REV MINERAL GEOCHEM 5. ANNU REV EARTH PL SC	1. ANNU REV BIOCHEM 2. CELL 3. NAT REV MOL CELL BIO 4. ANNU REV CELL DEV BI 5. ANNU REV GENET	1. CHEM REV 2. PROG POLYM SCI 3. ACCOUNTS CHEM RES 4. ANNU REV PHYS CHEM 5. ADV DRUG DELIVER REV
Cluster 5	Cluster 6	Cluster 7	Cluster 8
1. PROG ENERG COMBUST 2. ANNU REV FLUID MECH 3. PROG AEROSP SCI 4. P COMBUST INST 5. COMBUST FLAME	1. Q J ECON 2. J FINANC 3. J ECON LIT 4. J POLIT ECON 5. J FINANC ECON	1. ANNU REV NEUROSCI 2. NAT REV NEUROSCI 3. NEURON 4. NAT NEUROSCI 5. PROG NEUROBIOL	1. J AM MATH SOC 2. ANN MATH 3. ACTA MATH- DJURSHOLM 4. INVENT MATH 5. COMMUN PUR APPL MATH
Cluster 9	Cluster 10	Cluster 11	Cluster 12
1. PSYCHOL REV 2. BEHAV BRAIN SCI 3. COGNITIVE PSYCHOL 4. J EXP PSYCHOL GEN 5. COGNITION	1. MILBANK Q 2. ANNU REV PUBL HEALTH 3. JAMA-J AM MED ASSOC 4. HEALTH SERV RES 5. J HEALTH SOC BEHAV	1. REV EDUC RES 2. AM EDUC RES J 3. EDUC EVAL POLICY AN 4. EDUC PSYCHOL-US 5. J LEARN SCI	1. CRIT REV ORAL BIOL M 2. PROG RETIN EYE RES 3. AM J SPORT MED 4. PERIODONTOL 2000 5. SPORTS MED
Cluster 13	Cluster 14	Cluster 15	Cluster 16
1. J ACQ IMMUN DEF SYND 2. AIDS 3. CLIN MICROBIOL REV 4. J INFECT DIS 5. CLIN DIAGN VIROL	1. ADMIN SCI QUART 2. ACAD MANAGE J 3. ORGAN SCI 4. ACAD MANAGE REV 5. MIS QUART	1. ANNU REV PLANT BIOL 2. PLANT CELL 3. CURR OPIN PLANT BIOL 4. ANNU REV PHYTOPATHOL 5. MICROBIOL MOL BIOL R	1. ANNU REV IMMUNOL 2. NAT REV IMMUNOL 3. NAT IMMUNOL 4. CA-CANCER J CLIN 5. IMMUNITY
Cluster 17	Cluster 18	Cluster 19	Cluster 20
1. CIRCULATION 2. CIRC RES 3. J AM COLL CARDIOL 4. ARTERIOSCL THROM VAS 5. CARDIOVASC RES	1. ACM COMPUT SURV 2. J ACM 3. J R STAT SOC B 4. VLDB J 5. IEEE T ROBOTIC AUTOM	1. ANNU REV ECOL EVOL S 2. SYSTEMATIC BIOL 3. ANNU REV ENTOMOL 4. OCEANOGR MAR BIOL 5. TRENDS ECOL EVOL	1. REV MOD PHYS 2. MAT SCI ENG R 3. ANNU REV NUCL PART S 4. PHYS REP 5. PROG MATER SCI
Cluster 21	Cluster 22		
1. PSYCHOL BULL 2. ANNU REV PSYCHOL 3. J PERS SOC PSYCHOL 4. ARCH GEN PSYCHIAT 5. PERS SOC PSYCHOL REV	1. AM POLIT SCI REV 2. WORLD POLIT 3. AM J POLIT SCI 4. AM SOCIOL REV 5. ANNU REV SOCIOL		

Table 7. The five most important journals of each cluster according to a modified version of Google's PageRank algorithm (see Equation 1).

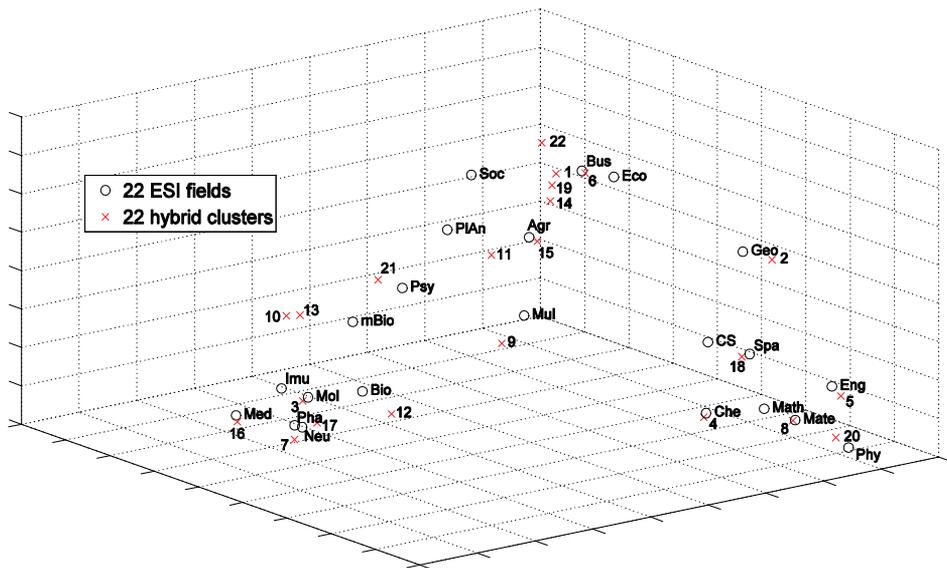


Fig. 7. Three-dimensional MDS map visualising distances between the centres (centroids) of the 22 ESI fields and the 22 clusters containing 8305 WoS journals.

In Figure 8, we use the Jaccard index to determine the concordance between our clustering solution and the ESI Scheme by comparing each cluster with every field, in order to detect the best matching fields for each cluster. The darker a cell in the matrix, the higher the Jaccard index, and hence the more pronounced the overlap between the corresponding cluster and ESI field. For example, cluster #4 (Chemistry) definitely corresponds to ESI field

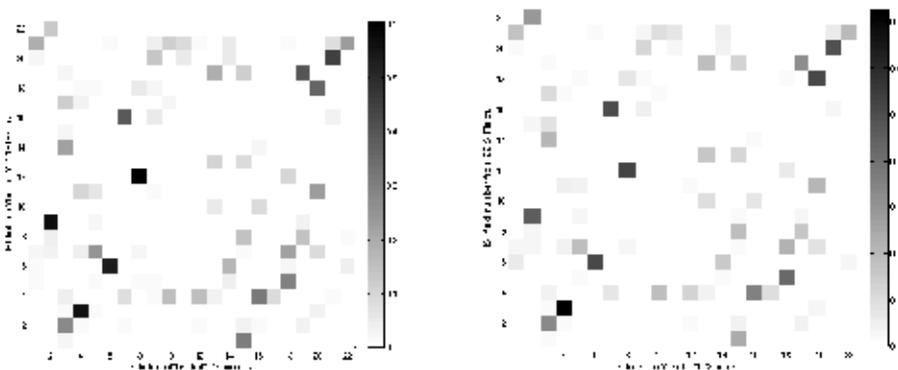


Fig. 8. Concordance between our clustering solution and the ESI Scheme visualised by coloured cells representing the Jaccard index for each cluster and field pair. The darkest cells represent the best matching pairs of fields and clusters. In the upper figure, the Jaccard index is computed from the number of journals a cluster and a field have in common, while the lower figure takes the size of each journal into account by counting the numbers of overlapping papers.

#3 (Chemistry). The same applies to field and cluster #6 (Economics and business). Clearly, ESI field #21 has the least concordance as this field is spread over seven clusters. It is defined as one single field in social sciences. It is not a surprise that the strongest match is found with our somewhat ‘fuzzy’ multidisciplinary social cluster. On the other hand, clusters #13 and #14 are quite similarly spread over four ESI fields each.

#### 4.7 Migration of journals among subject fields and clusters

If clustering algorithms are adjusted or changed, one can observe the following phenomenon. Some units of analysis are leaving clusters they formerly belonged to and end up in different clusters. This phenomenon is called ‘migration’. We can distinguish between ‘good migration’ and ‘bad migration’. ‘Good migration’ is observed if the goodness of the unit’s classification improves, otherwise we speak about ‘bad migration’. We can also apply this notion of migration to the comparison of clustering results with any reference classification. In the following we will use the ESI scheme as reference classification.

In the previous section we visualised the concordance between the clustering and the ESI classification. To determine for each ESI field the cluster that best matches the field, we used the Jaccard index on basis of the number of overlapping journals (cf. upper part of Figure 8). Out of 8305 journals under study, there were more than one third, namely, 3204 journals that were not assigned to the cluster which best matches their ESI field. As already mentioned above, we call these journals ‘migrated journals’. The largest ‘exodus’ comprising 226 migrating journals occurred from the ESI “Engineering” field to cluster #18 (Computer science), whereas the best matching cluster for the Engineering field is actually Cluster #5 (Engineering). The top 10 strongest patterns of migration are listed in Table 8, which indicate possible improvements of journal assignments.

Migration pattern	Number of migrated journals
From ESI field 7 (Engineering) to Cluster 18	226
From ESI field 14 (Molecular Biology & Genetics) to Cluster 3	159
From ESI field 21 (Social Sciences, general) to Cluster 10	145
From ESI field 11 (Materials Science) to Cluster 20	139
From ESI field 4 (Clinical Medicine) to Cluster 7	132
From ESI field 19 (Plant & Animal Science) to Cluster 15	108
From ESI field 21 (Social Sciences, general) to Cluster 21	98
From ESI field 7 (Engineering) to Cluster 20	95
From ESI field 4 (Clinical Medicine) to Cluster 3	86
From ESI field 8 (Environment/Ecology) to Cluster 15	86

Table 8. Top 10 strongest migration patterns

To measure the quality of migrations, we calculated the differences in Silhouette values before and after migration (based on textual and citation distances), for each migrated journal. Most migrated journals improved their Silhouette values. In the following, we will give some examples of good migrations and bad migrations.

'Good migrations' are observed if journals improved their Silhouette values after migration. Based on their titles and scopes (not shown), apparently they should indeed be assigned to the cluster to which they have moved. We observed numerous good migrations and the following cases will serve just as examples.

The *Journal of Analytical Chemistry* and *Chemia Analityczna* migrated from ESI field #7 (Engineering) to cluster 4 (chemistry). The best matching ESI cluster were field #3 (Chemistry) in this case (cf. Figure 8). Similarly, *Land Economics*, *Developing Economies* and *Economic Development and Cultural Change* migrated from field #21 (Social Sciences, general) to the more specific cluster 6 (economics and business). Here, the corresponding ESI field were #6 (Economics & business). In the life sciences, we found the following good migration. The journals *Neuropathology*, *Revista de Neurologia*, *Current Opinion in Neurology*, *Revue Neurologique*, *Lancet Neurology*, *European Journal of Neurology*, *Neurologist*, *Nervenheilkunde*, *Visual Neuroscience*, *Seminars in Neurology*, *Epilepsy & Behavior* and *Journal of Neuroimaging* migrated from field #4 (Clinical Medicine) to cluster #7 (neuroscience and behaviour) which rather corresponds to ESI field #16 (Neuroscience and behavior). Finally, we mention a migration between engineering and mathematics. The journals *Quarterly of Applied Mathematics*, *Bit Numerical Mathematics*, *Siam Journal on Discrete Mathematics* and *Discrete Applied Mathematics*, which were assigned to the ESI field Engineering (field #7), were found in our 'Mathematics' cluster (#8) which in turn corresponds to WSI field #12 (Mathematics). In the case of bad migration, the Silhouette values decreased after migration, that is, their Silhouette values in the ESI scheme were better than in the hybrid clustering. The reasons for this phenomenon are not always clear. According to their titles and scopes this migration is not always convincing. For instance, *Journal of Astrophysics and Astronomy*, *New Astronomy*, *Astrophysical Journal* and *Astronomy & Astrophysics* migrated from the ESI field 22 (Space Science) to Cluster 2 (geosciences) corresponding to ESI field #9, where we have to admit that journals in astronomy and astrophysics are in general spread over the geosciences and physics clusters. *Viral Immunology* migrated from field #10 (Immunology) to cluster #13 (microbiology and veterinary science) and *Canadian Journal of Microbiology* migrated from field #13 (Microbiology) to cluster #15 (agricultural and environmental sciences). Both clusters are rather spread over several ESI fields each (see Figure 8).

The distinction between good and bad makes a target-oriented adjustment of the existing classification scheme possible. Good migration can be used to reassign journals within the old scheme on the basis of the concordance with the results of clustering.

## 5. Conclusions

The hybrid clustering using textual information and cross-citations provided good results and proved superior to its two components when applied separately. The goodness of the resulting classification was even better than that of the "intellectual" reference scheme, the ESI subject scheme. Both classification systems form partitions of the Web of Science so that the direct comparison of clusters and fields was possible. According to our expectations, not all clusters have a unique counterpart in the ESI scheme and *vice versa* although the number of clusters coincided with the number of ESI fields. Although the Silhouette and modularity values substantiate a more coherent structure of the hybrid clustering as compared with the

ESI subject scheme, not all clusters are of high quality. Problems have been found, for instance, in clusters #1 and #12 where interdisciplinarity and strong links with other clusters distort the intra-cluster coherence. However, intellectual classification schemes usually do have a category “multidisciplinary sciences” as well. Although the result of a hard clustering algorithm often does contain a cluster with objects (journals) not strongly related to any other cluster, forming a “multidisciplinary sciences” cluster is not an inherent goal of the algorithm, and actually is not really meaningful either in the light of our outset goal to improve the classification of the sciences. Consequently, real multidisciplinary journals are scattered around different clusters.

Based on the external validation of clustering results by expert knowledge present in ISI subject categories, seven clusters seem to yield best results. Although there is no adequate subject classification scheme with 7 categories to be used as reference system, a more detailed analysis of this solution will be part of future research. Additional ideas for future research are a further improvement of the hybrid clustering algorithm by iterative cleaning of clusters as a post-processing step; allowing multiple assignments by fuzzy clustering; evaluating other algorithms like spectral clustering; and, finally, dynamic analysis by dynamic hybrid clustering.

The continuous rise of computing power might one day allow a large-scale mapping of the scientific universe explorable at various levels of detail. What’s more, application of advanced natural language processing and machine summarization at the scale of large bibliographic corpora might offer some insight into semantics beyond mere statistical processing.

## 6. References

- Bader, B.W. & Kolda, T.G. (2006). Algorithm 862: MATLAB tensor classes for fast algorithm prototyping. *ACM Transactions on Mathematical Software*, 32, 4, 635-653, ISSN: 0098-3500
- Baeza-Yates, R. & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*, Addison-Wesley, ISBN-10: 020139829X, Cambridge
- Batagelj, V. & Mrvar, A. (2002). Pajek – analysis and visualization of large networks. *Graph Drawing*, 2265, 477-478, ISSN: 0302-9743
- Berry, M.; Dumais, S.T. & O’Brien, G.W. (1995). Using linear algebra for intelligent information retrieval. *SIAM Review*, 37, 4, 573-595, ISSN: 0036-1445
- Brin, S. & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30, 1-7, 107-117, ISSN: 0169-7552
- Boyack, K.W.; Börner, K. & Klavans, R. (2008). Mapping the structure and evolution of chemistry research. *Scientometrics*, forthcoming
- Braam, R.R.; Moed, H.F. & Van Raan, A.F.J. (1991a). Mapping of science by combined co-citation and word analysis, Part 1: Structural aspects. *JASIS*, 42, 4, 233-251, ISSN: 0002-8231
- Braam, R.R.; Moed, H.F. & Van Raan, A.F.J. (1991b). Mapping of science by combined co-citation and word analysis, Part II: Dynamical aspects. *JASIS*, 42, 4, 252-266, ISSN: 0002-8231

- Callon, M.; Courtial, J.P.; Turner, W.A. & Bauin, S. (1983). From translations to problematic networks – An introduction to co-word analysis. *Social Science Information*, 22, 2, 191–235, ISSN: 0539-0184
- Callon, M.; Law, J. & Rip, A. (Eds.). (1986). *Mapping the Dynamics of Science and Technology: Sociology of Science in the Real World*, Macmillan Press, ISBN-10: 0333372239, London
- Deerwester, S.; Dumais, S.T.; Furnas, G.W.; Landauer, T.K. & Harshman, R. (1990). Indexing by latent semantic analysis. *JASIS*, 41, 6, 391–407, ISSN: 0002-8231
- ESI, *Essential Science Indicators* (accessible via: <http://www.esi-topics.com/fields/index.html>)
- Fano, R.M. (1956). Information Theory and the Retrieval of Recorded Information, In: *Documentation in Action*, J. H. Shera, A. Kent, J.W. Perry (Ed.), 238–244, Reinhold Publ. Co., New York
- Garfield, E. (1975). ISIS Atlas of Science may help students in choice of career in science. *Current Contents*, 29, 5–8
- Garfield, E. (1988). The encyclopedic ISI Atlas of Science launches three new sections – biochemistry, immunology, and animal & plant sciences. *Current Contents*, 7, 3–8
- Glänzel, W. & Czerwon, H.J. (1996). A new methodological approach to bibliographic coupling and its application to the national, Regional and institutional level. *Scientometrics*, 37, 2, 195–221, ISSN: 0138-9130
- Glenisson, P.; Glänzel, W. & Persson, O. (2005a). Combining full-text analysis and bibliometric indicators. A pilot study. *Scientometrics*, 63, 1, 163–180, ISSN: 0138-9130
- Glenisson, P.; Glänzel, W.; Janssens, F & De Moor B. (2005b). Combining full text and bibliometric information in mapping scientific disciplines. *Information Processing & Management*, 41, 6, 1548–1572, ISSN: 0306-4573
- Hatcher, E. & Gospodnetic, O. (2004). *Lucene in Action*, Manning Publications Co, ISBN-10: 1932394281, New York
- Hicks, D. (1987). Limitations of co-citation analysis as a tool for science policy. *Social Studies of Science*, 17, 2, 295–316, ISSN: 0306-3127
- Jain, A. & Dubes, R. (1988). *Algorithms for Clustering Data*. Prentice Hall, ISBN-10: 013022278X, New Jersey
- Janssens, F. (2007a). *Clustering of Scientific Fields by Integrating Text Mining and Bibliometrics*, Ph.D. Thesis, Faculty of Engineering, Katholieke Universiteit Leuven, Belgium, <http://hdl.handle.net/1979/847>
- Janssens, F.; Glänzel, W. & De Moor, B. (2007b). *Dynamic hybrid clustering of bioinformatics by incorporating text mining and citation analysis*. Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 360–369, ISBN 978-1-59593-609-7, San Jose, CA, USA, August 2007, ACM, New York
- Janssens, F.; Glänzel, W. & De Moor B. (2008). A hybrid mapping of information science. *Scientometrics*, 75, 3, 607–631
- Jarneving, B. (2005). *The Combined Application of Bibliographic Coupling and the Complete Link Cluster Method in Bibliometric Science Mapping*. PhD Thesis, University College of Borås/Göteborg University, Sweden

- Jarneving, B. (2007). Bibliographic coupling and its application to research-front and other core documents. *Journal of Informetrics*, 1, 4, 287–307, 1751-1577, ISSN: 1751-1577
- Kessler, M.M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, 14, 1, 10–25, ISSN: 0096-946X
- Kostoff, R.N.; Toothman, D.R.; Eberhart, H.J. & Humenik, J.A. (2001). Text mining using database tomography and bibliometrics: A review. *Technological Forecasting and Social Change*, 68, 3, 223–253, ISSN: 0040-1625
- Kostoff, R.N.; Buchtel, H.A.; Andrews, J. & Pfeil, K.M. (2005). The hidden structure of neuropsychology: Text mining of the journal *Cortex*, 1991–2001. *Cortex*, 41, 2, 103–115, ISSN: 0010-9452
- Leydesdorff, L. (2006). Can scientific journals be classified in terms of aggregated journal-journal citation relations using the Journal Citation Reports? *JASIST*, 57, 5, 601–613, ISSN: 1532-2882
- Leydesdorff, L. & Rafols, I. (2008). A Global Map of Science Based on the ISI Subject Categories. *JASIST*, to be published, ISSN: 1532-2882
- Mardia, K.V.; Kent, J.T. & Bibby, J.M. (1979). *Multivariate Analysis*, Harcourt Brace & Co, Academic Press, ISBN-10: 0124712525, London, UK.
- Marshakova, I.V. (1973). System of connections between documents based on references (as the Science Citation Index), *Nauchno-Tekhnicheskaya Informatsiya, Seriya 2*, 6, 3–8
- Moya-Anegón, F. de; Vargas-Quesada, B.; Chinchilla Rodríguez, Z. ; Corera-Alvarez, E.; Muñoz Fernández, F.J. & Herrero-Solana, V. (2007). Visualizing the Marrow Science. *JASIST*, 58, 14, 2167-2179.
- Narin, F. (1976). *Evaluative Bibliometrics: The Use of Publication and Citation Analysis in the Evaluation of Scientific Activity*. Computer Horizons, Inc., Washington, D.C
- Newman, M.E.J. & GIRVAN, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69, 2, ISSN: 1063-651X
- Newman, M.E.J. (2006). Modularity and community structure in networks. *PNAS US*, 103, 23, ISSN: 0027-8424
- Noyons, E.C.M. (1999). *Bibliometric Mapping as a Science and Research Management Tool*, DSWO Press, Leiden University, ISBN 9090132503, Leiden, The Netherlands
- Pinski, G. & Narin, F. (1976). Citation influence for journal aggregates of scientific publications. *Information Processing and Management*, 12, 5, 297–312, ISSN: 0306-4573
- Rousseeuw, P. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 1, 53–65.
- Salton, G. & Mcgill, M.J. (1986). *Introduction to Modern Information Retrieval*, McGraw-Hill, Inc, ISBN: 0070544840, New York.
- Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *JASIS*, 24, 4, 265–269, ISSN: 0002-8231
- Small, H. (1978). Cited documents as concept symbols. *Social Studies of Science*, 8, 327–240, ISSN: 0306-3127
- Small, H. (1998). A general framework for general large-scale maps of science in two or three dimensions: The SciViz system. *Scientometrics*, 41, 1–2, 125–133, ISSN: 0138-9130

Zhang, L. & Glänzel, W. (2008). Journal cross-citation matrices reconsidered. Tracing the role of individual journals in the communication network. In: *Proceedings of WIS 2008*, H. Kretschmer and F. Havemann (Ed.), Berlin. (accessible via: <http://www.collnet.de/Berlin-2008/ZhangLinWIS2008jcm.pdf>)

# Automatic Product Classification Control System Using RFID Tag Information and Data Mining

Cheonshik Kim<sup>1</sup>, Eun-Jun Yoon<sup>2</sup>, Injung Park<sup>3</sup> and Taek-Young<sup>4</sup>

<sup>1</sup>*Anyang University,*

<sup>2</sup>*Daegu Polytechnic College,*

<sup>3</sup>*Dankuk University,*

<sup>4</sup>*UM technology*

*South Korea*

## 1. Introduction

Information systems in the modern process and manufacturing sector bridge several layers, from the boardroom to the shop floor. At one end of this spectrum the ubiquitous ERP systems exist that have become essential to today's IT-enabled enterprises. At the opposite end of the spectrum, sensors, actuators and other field devices are found that are equally vital for ensuring perfect process control. Between these extremes a diverse range of systems with varying degrees of interconnection, fineness and cohesion exist (TATA, 2006). Information systems span a wide range of data, processing power and time scales. The objectives and abilities of each of these systems are different, yet there is a clear need for them to operate in sync with each other. Any disconnect among them leads to inefficient operations, higher costs and lower quality, which ultimately translates into lower profits. Therefore, it is vital that there should be tight integration and perfect communication between all systems.

A clear need exists for a set of systems that seamlessly bridge this gap. The manufacturing execution system (MES) fills this need. The MES controls the operations that enable realization of the plans, close the execution gap by providing links among shop floor instrumentation, control hardware, planning and control systems, process engineering, production execution, the sales force and customers. The MES has multiple advantages. Nevertheless, there is no function regarding analysis techniques concerning the manufacturing process in the MES. Therefore, we designed and implemented the MES for the manufacturing process of TFT LCDs. Also, as reported in this manuscript, we investigated and analyzed the defects of LCDs that are produced in the manufacturing process using a data mining technology.

## 2. Manufacturing execution system

A Manufacturing Execution System (MES) is a system that companies can use to measure and control production activities with the aim of increasing productivity and improving

quality. The ISA has defined standards regarding the structuring of MES and its integration in a larger company-wide IT architecture. MES fits in between ERP and process automation level. MES gets production order and those are scheduled by ERP and that is also not in detail. Material requirement planning does the scheduling at the ERP level. MES collects the production order and does a detail scheduling for a small period. The industry wants to know how to reduce the manufacturing cycle time, improve the quality, lower the cost and get more profit. Since MES can provide real time monitor and integrate with ERP and other information system, the potential utility will appear after the enterprise used MES. MES collected the manufacturing data, and managers can make the strategic decision by it and carry out the decision. It is information presented on-line to the production operator and to the desk of manufacturing management. Under MES implementation, integration with your accounting system, order entry system, inventory system, scheduling system and others become easy. These all become plug and play pieces because the data is sharable. While the ERP focus is in areas such as finance, HR, etc., new investments focused on improving and optimizing production and logistic resources. An Advanced Planning and Scheduling (APS) system uses advanced programming techniques to improve/optimize production planning and scheduling, to allow the company to achieve pre-defined objectives, such as improvements on delivery performance without raising inventory levels, or maximize plant throughput (Watford, 2004). (Tao et al., 2004) proposed an implementation process model for integrating the extensible Markup Language (XML) into an enterprise application, which also meets the inter-organizational data exchange standard of RosettaNet. Farahvash and Boucher (Farahvash et al. 2004) [4] introduced an architecture that integrated shop floor agents for scheduling, cell control, transportation, and material management.

### **3. Design of MES for TFT LCD's process control**

#### **3.1 System architecture design**

Product inspection is handled in real-time, because the RFID (Juels et al., 2003) concept is applied to the MES. If the LCD Panel or AD Board arrives in the examination line, the RFID tag will be printed. Information for parts is also provided because the information is linked with the ERP or SCM. However, because the product is selected from the database in this study, the RFID TAG LABEL is printed. The system architecture (Fig. 1) was based on the MES component principle. Quality management process management, labor management, data collection and acquisition, dispatching production units and document control of the MES function were modeled.

#### **3.2 Data flow diagram of the system**

The data flow diagram of the part inspection system using RFID in an electronic device manufacturing process, in which parts are produced or arrived at the subcontractors, the part information is inputted, the tag is attached by using an RFID printer, and it is moved to the inspection conveyor is depicted in Fig. 2. An inspector reads the tag information using a RFID reader, and outputs the information for the product concerned on the monitor. An LCD electronic component is conveyed to an appropriate box or a pallet after judgment is made regarding whether to send the part to a shipment conveyor line, re-inspect it, or send it to disposition on the conveyor line due to inferior quality, when tag information is read through the final RFID reader. The system stores inspection data on various parts in the database, and

permits them to shown on the monitor of a subcontractor, or an office manager, through the EDI/WEB. The system is configured to link with the existing SCM, and the ERP DB.

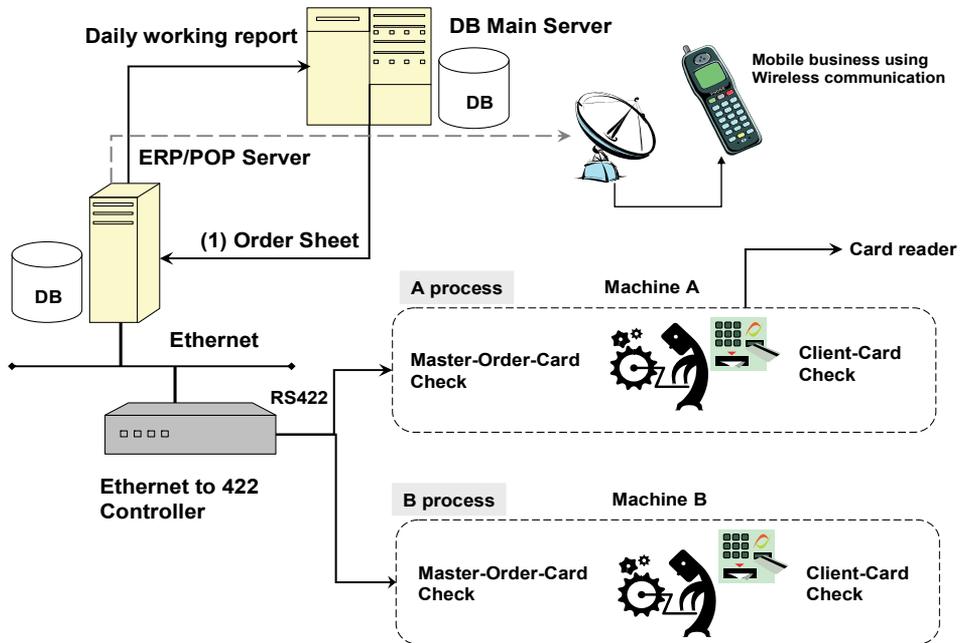


Fig. 1. Architecture of system for the parts

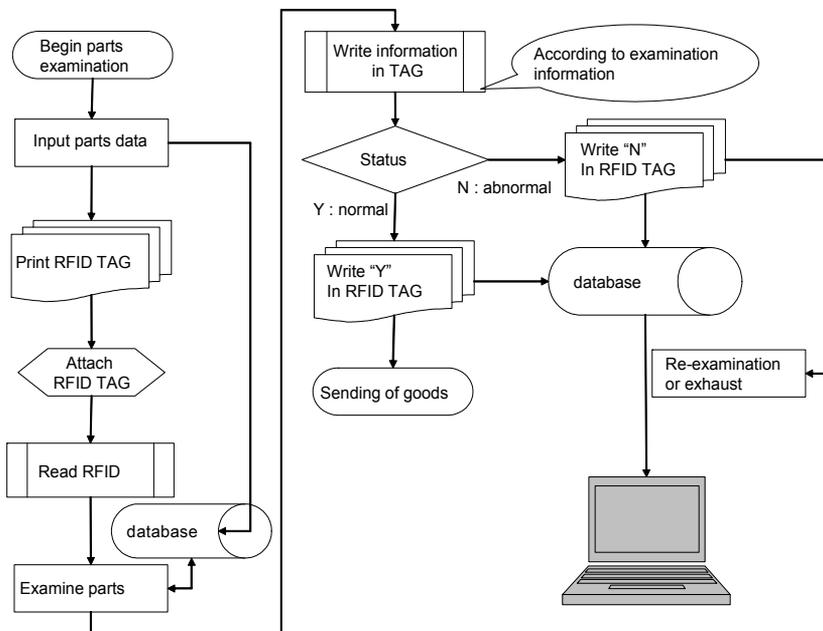


Fig. 2. Title of figure, left justified

The diagram in Fig. 3 illustrates a detailed explanation of processes. When an electronic component arrives at the inspection line via the conveyor belt, the RFID reader automatically reads the appropriate information from the tag and judges whether or not it is a panel inspection or a board inspection.

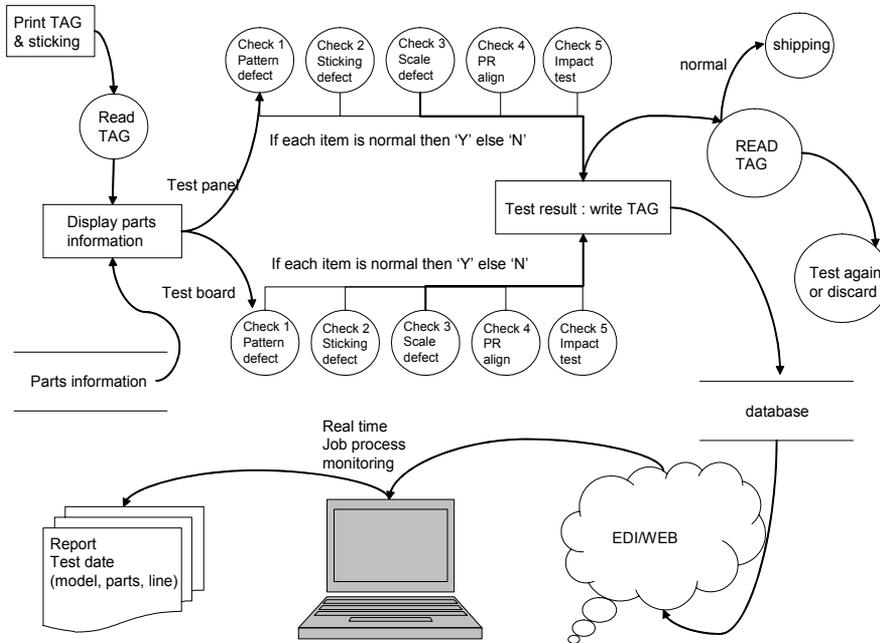


Fig. 3. Data flow diagram for the detail processing

Fig. 4 shows the main screen of the parts inspection system. The screen is on standby in the state depicted in Fig. 4. The left hand side of the screen configuration automatically indicates the details of the electronic component concerned. On the right hand side, the current date, time, inspector's name and ID, and inspection line are displayed. In the centre of the display, which is the core of the program, the inspection items are automatically displayed in line with the electronic component concerned. Therefore, the appropriate inspection items can be easily understood by non-skilled personnel. The main screen of the parts checking system (Fig. 4) is implemented by the MES.

#### 4. Data mining for efficient production control

In this chapter, we will explain neural networks and C 4.5 algorithms. Also, we apply these algorithms in the manufacturing process for TFT LCDs and suggest methods by which to locate defective parts. The method proposed may contribute to improve the efficiency of the manufacturing process for TFT LCDs. A neural network algorithm is a technology used in estimate modelling. Neural network algorithms are very efficient; however, this type of algorithm has shortcomings that can not explain the estimate sequence. Therefore, we used C4.5 algorithms and supplemented the shortcomings of the neural network algorithm. The C4.5 algorithm appropriately explains the sequence.

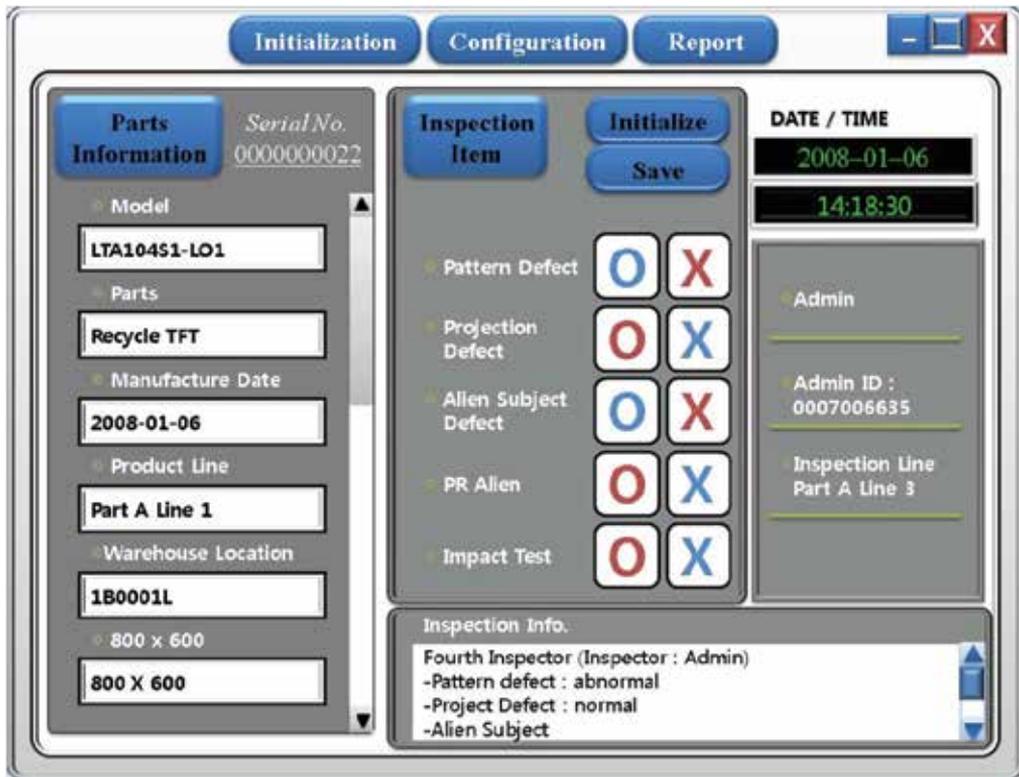


Fig. 4. Main screen for parts inspection

#### 4.1 Neural network algorithm

Neural network (Quinlan, 1993) technology uses a multilayered approach that approximates complex mathematical functions to process data. Today neural networks can be trained to solve problems that are difficult for conventional computers or human beings. As shown a situation in Fig 5, neural networks are trained that a particular input leads to a specific target output. Based on a comparison of the output and the target, the network is trained until the network output matches the target. Typically many such input/target pairs are used to train a network.

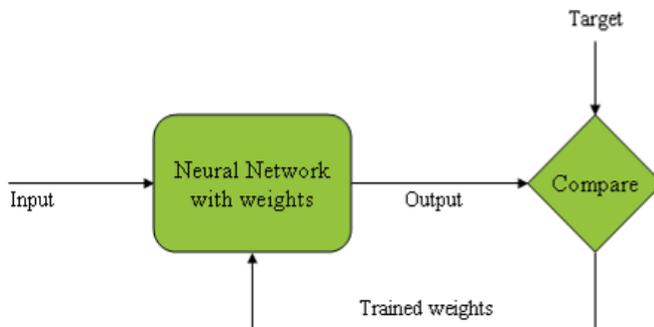


Fig. 5. Neural network algorithm

Neural network consists of many processing elements or nodes that work in parallel. Nodes are connected to each other in layers, and layers are interconnected. These nodes are simple mathematical functions; the connections between these nodes, which weight the data transformation from each node and send the information to the next node or output layer, are how neural networks "think." As the complexity of the task increases, the network size increases, and the number of nodes increases rapidly. To properly train a neural network, the developer feeds the model a variety of real-life examples, called training sets. The data sets normally contain input data and output data. The neural network creates connections and learns patterns based on these input and output data sets. Each pattern creates a unique configuration of network structure with a unique set of connection strengths or weights. A neural network adapts to changing inputs and learns trends from data. A set of examples of the data or images is presented to the neural network, which then weights the connections between nodes based on each training example. Each connection weight builds on previous decision nodes, propagating down to a final decision (Equ. 1).

- Signals passed from each input node are gathered and become a linear combination. That is, the hidden node,  $L$ , is expressed as follows if  $(x_1, \dots, x_p)$  is the explanatory variable.
- Connection weights for each input are summed, resulting in a unique complex function each time the neural network is trained with a set of inputs and outputs. Successively summed weights define the algorithm that the neural network uses to make a pattern-matching decision.

$$L = w_1 X_1 + \dots + w_p X_p \quad (1)$$

After the neural network reaches a final decision, it compares its answer against an answer provided in the training set. If there is a match, within a predefined tolerance, the neural network stores these connection weights as successful. If the decision outcome is outside the tolerance, then the neural network cycles through the training set again. A neural network may cycle thousands of times to reach an acceptable tolerance. Table 1 is used to determine the variable used to forecast the defect in TFT LCDs.

Input variable	Meaning of variable	Category
Variable 1	Pattern defect	Y/N
Variable 2	Output voltage	Y/N
Variable 3	Decide projection	Y/N
Variable 4	Terminal R	Y/N
Variable 5	Terminal Y	Y/N
Variable 6	PR Align	Y/N
Variable 7	Terminal W	Y/N
Variable 8	Impact test	YES/NO

Table 1. Factor for determining the quality of TFT LCDs

#### 4.2 C4.5 algorithm

Statisticians developed a tree-structured classification of many members known as machine learning. Characteristics of the tree model are described as follows.

If A, then B, Else C

The decision tree C4.5 algorithm is a practical method for inductive inference (Abdi, 1994, Mitchell, 1997, Quinlan, 1986). C4.5 algorithm is a method for approximating discrete-valued functions, in which the learned function is represented by a decision tree (Polat, 2008). The aim of C4.5 algorithm is recursively partition data into sub-groups. The decision tree C4.5 algorithm can be used an entropy standard. Entropy is a concept used to measure randomness in thermodynamics. Suppose that data set, T, depends on Y and is divided into k. Then, the ratio ( $p_1, \dots, p_k$ ) of the category can be classified. Therefore, the entropy of T is defined in equation 2.

$$Entrop(T) = \sum_{i=1}^k p_i \log p_i \quad (2)$$

The C4.5 model must find a separation variance that generates the lowest entropy in the entropy test. In this paper, we will determine factors concerning the defects of LCDs using C4.5 algorithms. The factors in Table 1 were used in an experimental design. The sequence of results is illustrated in Fig. 6.

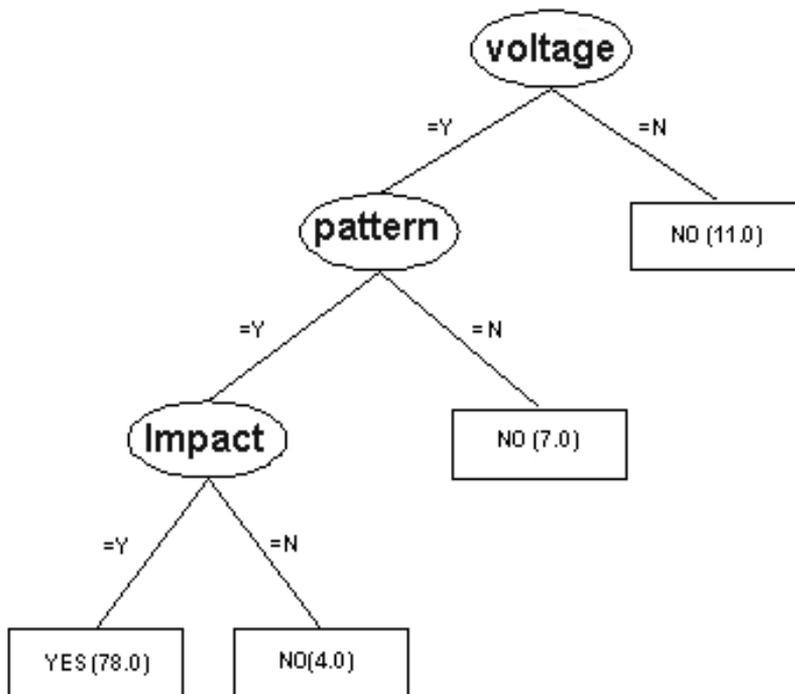


Fig. 6. Application of C4.5 algorithm to LCD line

### 4.3 Analysis of the proposed algorithm

Table 2 and Table 3 are results from the manufacturing process that apply to the neural network algorithm and the C4.5 algorithm, respectively. Table 2 and Table 3 present data on precision and recall.

Precision	Recall	F-Measure	ROC Area	Class
0.98	0.987	0.985	0.946	YES
0.943	0.943	0.971	0.946	NO

Table 2. Experiment result of neural network algorithm

The results of analysis are as follows.

- 11 items had a voltage defect in the whole parts production number.
- 7 items had both a pattern defect and a voltage defect in the whole parts production number.
- 4 items simultaneously had a pattern defect, a voltage defect and an impact defect in the whole parts production number.

Therefore, the incidence of voltage defects should be reduced. Also, we determined that a voltage defect causes a pattern defect. Finally, voltage defects in the manufacturing process should be managed specifically.

We used RFID technology and demonstrated that the process of producing TFT LCDs can generate monitoring in real time. Also, we produced real-time data information by chart to more easily confirm manufacturing information. Fig. 7 illustrates productivity according to each line of TFT LCD production.

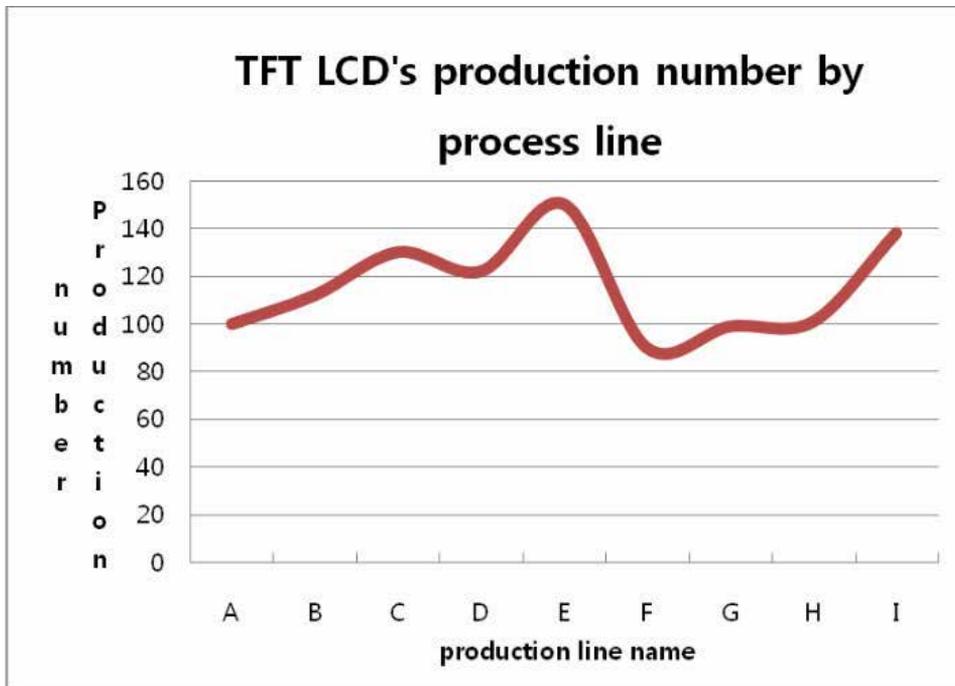


Fig. 7. Statistical analysis of the parts using RFID tag

Precision	Recall	F-Measure	ROC Area	Class
0.97	0.985	0.977	0.983	YES
0.971	0.943	0.957	0.983	NO

Table 3. Experiment result of C4.5 algorithm

## 5. Conclusion

In this paper, the system proposed was established to actualize an overall inspection system for electronic components or devices using RFID. This study developed a system to inspect electronic components and devices, specifically, LCDs Panels that are the core parts of an LCD monitor store inspection result data in the RFID TAG and the Reader/Writer. The existing system managed the inspection result data by manually attaching stickers containing inspection values. As a result of implementing the system developed in this research, the inspection time in the real parts inspection line was greatly reduced. The system developed consists in a way that the inspection data of multiple types of parts or devices can be displayed in real time to raise the efficiency of the concerned inspector or manager. This system is comprised of a MS-SQL SERVER or MySQL, which is a general purpose database, and can be linked with various ERPs and SCM. This system is forecast to provide many benefits to LCD panel and parts inspection companies. The MES is an excellent system for process control. However, the MES lacks an analysis function concerning information that occurs in process control. As a result, this approach was effective for closely examining the cause of necessary product defects in process control. We expect that the system proposed in this paper will be useful in various fields.

## 8. References

- TATA consultancy services (2002). *MANUFACTURING EXECUTION SYSTEMS: A Concept*. February.
- Watford (2004), *Integration of MES with Planning and Scheduling Solutions*. Broner Metals Solutions Ltd , UK.
- Yu-Hui Tao, Tzung-Pei Hong, Sheng-sun (2004). *An XML implementation process model for enterprise applications*, *Computer in Industry* 55,
- Pooya Farahvash, Thomas O. Boucher(2004). *A multi-agent architecture for control of AGV system*. *Robotics and computer-Integrated manufacturing* 20.
- Juels, R. Rivest, M. Szydlo(2003). *The Blocker Tag: Selective Blocking of RFID TAG for Consumer Privacy*". 10th ACM CCS.
- Quinlan, J. R.(1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers.
- Abdi, H.(1994). *A neural network primer*. *Journal of Biological Systems*, 2, 247-281.
- Mitchell, M. T.(1997). *Machine learning*. Singapore: McGraw-Hill.

Kemal Polat, Salih Gunes (2008), *A novel hybrid intelligent method based on C4.5 decision tree classifier and one-against-all approach for multi-class classification problems*, In press of Expert Systems with Applications.

# Hyperspectral Remote Sensing Data Mining Using Multiple Classifiers Combination

Xing-Ping Wen<sup>1</sup>, Xiao-Feng Yang<sup>1</sup> and Guang-Dao Hu<sup>2</sup>

<sup>1</sup>*Faculty of Land Resource Engineering, Kunming University of science and technology,*

<sup>2</sup>*Institute of Mathematic Geology and Remote Sensing Geology,*

*China University of Geosciences (Wuhan)*

*China*

## 1. Introduction

Since the advent of remote sensing in the second half of 20th century, nowadays there have been great changes in theory and technology. The advent of hyperspectral was one of the most significant breakthroughs in remote sensing. Hyperspectral remote sensing has higher spectral resolution as the same time retain higher spatial resolution, so its capability of distinguishing the different and describing the same ground objects in details enhanced greatly. It acquires image in a large number (typically over 40), narrow (typically 10 to 20 nm in width) and contiguous spectral bands to enable the extraction of reflectance spectra at a pixel scale, so it can produce data with sufficient resolution for the direct identification of those materials with diagnostic spectral features (Goetz et al., 1985). The objective of hyperspectral remote sensing is to measure quantitatively the components of the Earth System from calibrated spectra acquired as images for scientific research and applications (Vane & Goetz, 1988). The rationale behind this technology for geological applications is that mineral species have diagnostic absorption features from 20 to 40 nm wide in electromagnetic wavelength ranges which is larger than hyperspectral spectral resolution (van der Meer & Bakker, 1997). Goetz demonstrated firstly that direct identification of carbonates and hydroxyl-bearing minerals is possible by remote measurement from Earth orbit (Goetz et al., 1982).

There are two main categories of extracting information method from hyperspectral remote sensing image: based on feature space and based on spectral space. Many statistics-based classification methods based on feature space have been successfully applied to multi-spectral remote sensing data in the past years (Pal & Mather, 2003, Wen et al., 2008b). However, they are not effective for hyperspectral remote sensing data. The problem is caused by curse of dimensionality and Hughes phenomenon (Hsu, 2007), which refer to the fact that the sample size required for training a specific classifier grows exponentially with the number of spectral bands. Usually simple but sometimes effective ways to overcome this problem is to increase sample numbers or to reduce the dimensionality of hyperspectral remote sensing data. The former needs a lot of sample numbers, so it will cost many human and material resources; the latter will lead to some useful information lost. The Matched Filtering methods based on spectral space are successfully used in hyperspectral data. These

methods based on the hypothesis that all spectra have been calibrated to apparent reflectance and the dark current of sensor and path radiation is removed. It is only the ideal state for these effects are hard to remove successfully, especially in low reflectance object, so this will lead to error using Matched Filtering.

In order to obtain the best classification performance in pattern recognition, the data set should be classified using different methods, and then choose the best classification result as the final conclusion. With the complexity of pattern recognition increased and novel algorithm developed, researchers find that although different classifiers have different classification performance, their misclassification set are not consistent with each other. That is, some sample misclassified by one classifier may be recognized by another classifier. Different classifiers are complementary to each other. If only the best performance classifier is chose, some valuable information from other classifiers may be ignored. In order to solve this problem, multiple classifiers combination was put forward. This chapter proposed the methods to improve the hyperspectral remote sensing classification accuracy using multiple classifiers combination. The method included two parts: one is to improve the performance of the single classifier, and the other is to combine multiple classifiers. In the former, the chapter investigated the methods of atmospheric correction and extracting the purer endmember, and a novel endmember extracting method combining multi-segmentation and geology map was proposed. In the latter, combining multiple classifiers based on decision tree was proposed.

The study area is located in southwest of China, and the composition of the deposits is mainly dioritic porphyrite. Firstly, the image was atmospherically corrected before processing, and an endmember was extracted by PPI algorithm from the intersection area of multi-segmentation and geology map; Secondly, dioritic porphyrite area was extracted from hyperspectral remote sensing image by Spectral Angle Mapper (SAM), Multi Range Spectral Feature Fitting (Multi Range SFF) and Mixture Tuned Matched Filtering (MTMF) using the extracted endmember respectively. At last, final classification results was outputted by combining three classification results using Classification and Regression Trees (CART). Comparing all classification results and geology map, it is concluded that combining multiple classifiers has the best classification performance and Multi Range SFF has the better capable of pixel un-mixing than SAM and MTMF.

## **2. Study area and hyperspectral remote sensing data**

The study area is located at Pulang porphyry copper and gold deposit in Shangri-La of Yunnan province in southwest of China (fig. 1). The longitude and latitude scope are between 26°54'N - 28°43'N and 99°37'E - 100°10'E. Its ore-bearing lithology mainly contains dioritic porphyrite. The hyperspectral remote sensing employed in this paper was acquired by Hyperion sensor on board EO-1 satellite in December 2, 2003. The EO-1 satellite was launched on November 21, 2000 as part of a one-year technology validation/demonstration mission by NASA. The original EO-1 Mission was successfully completed in November 2001. As the end of the Mission approached, an agreement was reached between NASA and the USGS to allow continuation of the EO-1 Program as an Extended Mission based on the remote sensing research and scientific communities interest and willingness to assist in funding continued operations. The three primary instruments on the EO-1 spacecraft are the Advanced Land Imager (ALI), the Hyperion, and the Linear Etalon Imaging Spectrometer Array (LEISA) Atmospheric Corrector (LAC). The Hyperion covers 400-2500 nm with 242

spectral bands over a 7.5 km wide swath at approximately 10nm (sampling interval) spectral resolution and 30m spatial resolution on a 705 km orbit.



Fig. 1. China provincial boundaries and study area

### 3. Atmospheric correction of hyperspectral remote sensing data

Atmospheric effects play a dominant role in the optical part of the electromagnetic spectrum. It generates a variety of effects upon the satellite image that must subsequently be accounted for through atmospheric corrections. To have lasting quantitative value, remotely sensed data must be calibrated to physical units of reflectance (Smith & Milton, 1999). In order to remove accurately atmospheric absorption and scattering effects, atmospheric correction algorithms have evolved from the earlier empirical line method and flat field method to more recent methods based on rigorous radiative transfer modeling. MODTRAN, the Air Force Research Laboratory/Geophysics Directorate moderate spectral resolution (2  $\text{cm}^{-1}$ ) background radiance and transmittance model, has been widely used to analyze hyperspectral data to its computational speed and its ability to model molecular and aerosol/cloud emissive and scattered radiance contributions as well as the atmospheric attenuation (Berk et al., 1998). It accurately and efficiently calculates the scattering and absorption signatures of realistic molecular, aerosol and cloudy environments in the lower and middle atmosphere. There are several MODTRAN-based atmospheric correction software packages, such as FLAASH (Fast Line-of-sight Atmospheric Analysis of Spectral Hyper-cubes) (Anderson et al., 1999), ATCOR (Atmospheric and Topographic Correction) (Richter, 1996) and others. ATCOR was originally developed by Richter at DLR, the German Aerospace Centre. The algorithm is a fast atmospheric correction for imagery of medium and high spatial resolution satellite sensors. There are two ATCOR models available, one for satellite imagery, the other one for airborne imagery. For historic reasons, the satellite codes are called ATCOR-2 (Richter, 1996) and ATCOR-3 (Richter, 1998). ATCOR-2 is used for flat terrain and ATCOR-3 is for Mountainous Terrain. ATCOR3 includes all of the capabilities of ATCOR2 and can be integrated with a DEM for atmospheric correction of images depicting

rugged terrain. FLAASH (Fast Line-of-sight Atmospheric Analysis of Spectral Hypercubes) is being developed by the Air Force Research Laboratory, Hanscom AFB, and Spectral Sciences, Inc., to support current and planned SWIR/visible/UV hyperspectral and multispectral sensors, typically in image format.

In order to compare correction effect of different algorithms, in this chapter, ACTOR-2, ACTOR-3 and FLAASH were used to remove atmospheric effects from the hyperspectral data. Fig. 3 is the reflectance of the same pixel from Hyperion remote sensing data corrected atmospherically by ACTOR-2, ACTOR-3 and FLAASH. Due to complex terrain of study area, where maximum elevation difference is greater than 1200 KM (fig.2), the correction effect of ACTOR-3 is not efficient. Some reasons may be originated from the rough DEM data and geometric correction error. As a result, the Hyperion remote sensing data calibrated to apparent reflectance using the FLAASH was selected to extract information.

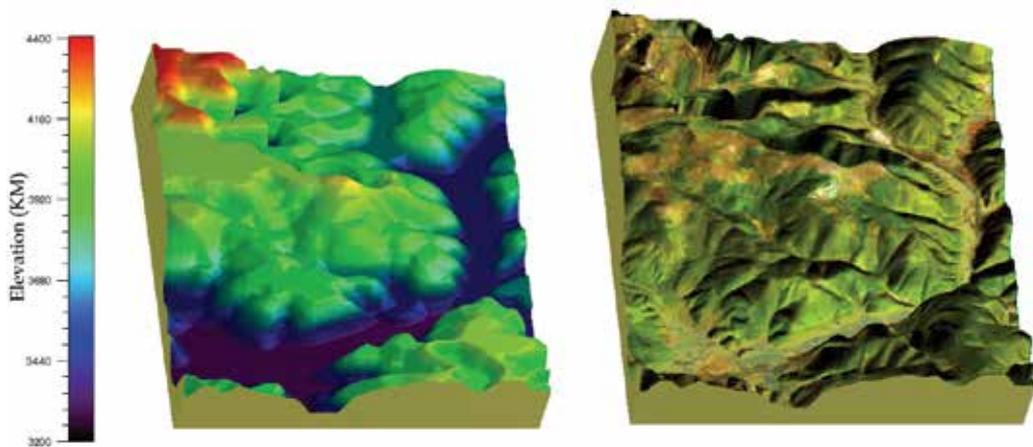


Fig. 2. Digital elevation model ( DEM) of study area (left) and 3D surface view image using hyperion remote sensing data

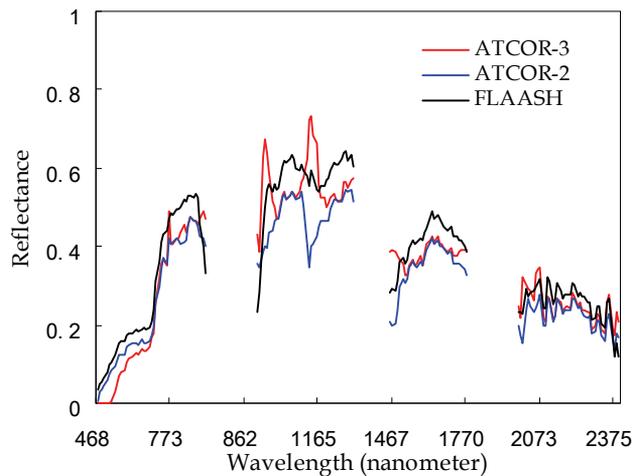


Fig. 3. The reflectance of the same pixel using different atmospheric correction model

## 4. Methodology

### 4.1 Endmember extraction

Mineral mixtures and mixtures with vegetation in an individual pixel can be separated if the components have unique spectral features (endmember). Many researchers have used an n-dimensional analysis approach to determine key endmember and map their distribution and abundance (Boardman, 1993). A variety of methods have been proposed to find endmembers in multispectral and hyperspectral images. Iterated Constrained Endmembers (ICE) is an automated statistical approach to identifying endmembers from hyperspectral images (Berman et al., 2004). Winter proposed to find a unique set of purest pixels based upon the geometry of convex sets (Winter, 1999). Probably the most widely used algorithm is Pixel Purity Index (PPI) (Boardman et al., 1995). PPI usually used to find the most spectrally pure (extreme) pixels in multi-spectral and hyperspectral images. (Kruse, 2005) used multi-resolution segmentation (Baatz & Schäpe, 2000) to separate adjacent regions in an image, then using PPI to extract endmember. Multi-resolution segmentation is a bottom up region-merging technique starting with one-pixel objects. In numerous subsequent steps, smaller image objects are merged into bigger ones. It can slice the image into a network of homogeneous image regions at any chosen resolution, even when the regions themselves are characterized by a certain texture or noise. In this chapter, multi-resolution segmentation regions in hyperspectral remote sensing image were intersected with the dioritic porphyrite area in geologic map (fig. 4), so the overlap areas contain homogeneous dioritic porphyrite area. Then, one overlap area was selected to extract dioritic porphyrite endmember using PPI.

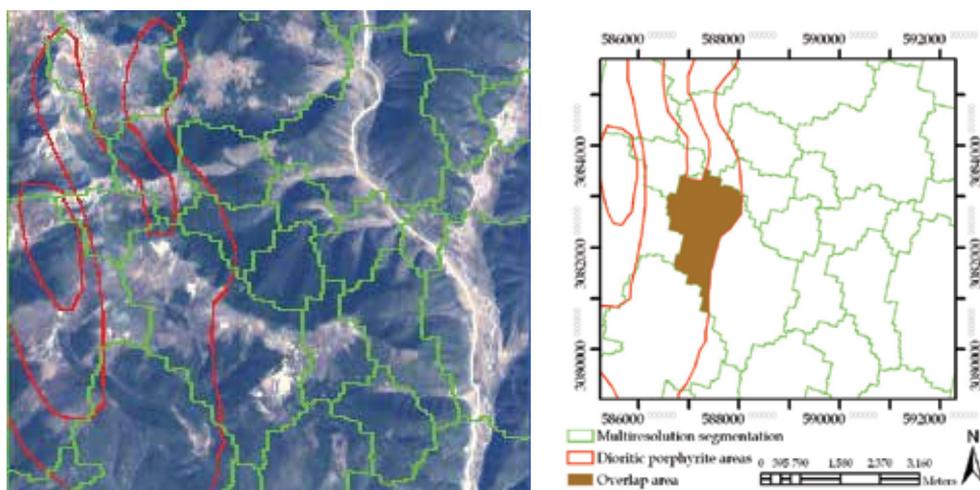


Fig. 4. The hyperspectral remote sensing image (left) and the overlap area where multi-resolution segmentation intersects dioritic porphyrite area in geologic map (right)

### 4.2 Matched filtering methods

Each pixel in the image is a mixture of responses from multiple materials and nearly no pure pixels are present in the image due to large sampling distance (30 m). Matched Filtering (MF) method can find the abundances of user-defined endmembers using a partial unmixing. There are many MF methods. SAM, MTMF and Multi Range SFF are widely used to

extract specific materials based on matches to library or image endmember spectra and does not require knowledge of all the endmembers within an image scene.

#### 4.2.1 SAM algorithm

SAM is a physically-based spectral classification that uses an n-dimension angle to match pixels to endmember spectra. The algorithm determines the spectral similarity between two spectra by calculating the angle between the spectra, treating them as vectors in a space with dimensionality equal to the number of bands (Kruse et al., 1993). The endmember spectra can be either laboratory or field spectra or extracted from the image. This method assumes that the data have been reduce to apparent reflectance, with all dark current and path radiance biases removed. This technique, when used on calibrated reflectance data, is relatively insensitive to illumination and albedo effects. SAM algorithm has successfully used in geological mapping based on remote sensing data (Baugh et al., 1998, Wen et al., 2007b). It computes the "spectral angle" between pixel spectra and the endmember spectra. Smaller angles represent closer matches to the endmember spectra. The result is a rule image that indicates the radian of the spectral angle (fig. 6(a)). The radian of the spectral angle calculated by applying the following equation:

$$\alpha = \cos^{-1} \left( \frac{\sum_{i=1}^{nb} t_i r_i}{\left( \sum_{i=1}^{nb} t_i^2 \right)^{1/2} \left( \sum_{i=1}^{nb} r_i^2 \right)^{1/2}} \right) \quad (1)$$

Where nb=the number of bands,  $t_i$  =pixel spectrum,  $r_i$ =endmember spectrum.

#### 4.2.2 Multi range SFF

SFF is an absorption-feature-based methodology. The endmember spectra are scaled to match the image spectra after the continuum is removed from both data sets. Multi Range SFF uses Continuum Removal (CR) to normalize reflectance spectra so individual absorption features from a common baseline can be compared. An apparent continuum in a reflectance spectrum is modeled as a mathematical function that is used to isolate a particular absorption feature for analysis, and this continuum should be removed by dividing it into the reflectance spectrum for each pixel in the image (Clark & Roush, 1984):

$$S_{CR} = S/C \quad (2)$$

Where:  $S_{CR}$  = Continuum removed spectra,  $S$  = Original spectra,  $C$  = Continuum curve.

The continuum curve is a convex hull fit over the top of a spectrum using straight-line segments that connect local spectra maxima, so the first and last spectral data values are on the hull, therefore, the first and last bands in the output continuum removed data are equal to 1.0. Absorption feature analysis using CR has been shown to enhance the differences in shape between the absorption features of interest (Kokaly & Clark, 1999). Multi Range SFF compares the CR of image spectra to the CR of endmember spectra at each wavelength using a least-squares technique. Scale image and RMS image are output for each

endmember spectrum. The scale image is a measure of absorption feature depth, which is related to material abundance. The brighter pixels in the scale image indicate a better match to the endmember material in those pixels. The RMS error is calculated for each endmember spectrum, and dark pixels in the RMS image indicate a low error. Finally, the fit image with the higher abundance and a low error value which was calculated by dividing RMS image into the Scale image is output for the endmember spectrum (fig. 6(b)).

#### 4.2.3 MTMF

MTMF combines the best parts of the Linear Spectral Mixing model and the statistical Matched Filter model while avoiding the drawbacks of each parent method (Boardman, 1998). It is a useful Matched Filter method without knowing all the possible endmembers in a landscape especially in case of subtle, sub-pixel occurrences. Firstly, pixel spectra and endmember spectra require a minimum noise fraction (MNF) (Green et al., 1988, Boardman, 1993) transformation. MNF reduces and separates an image into its most dimensional and non-noisy components. Once the data is in a less noisy form, it can then be compared to endmember through MTMF processes to determine composition. Its results appear as two gray-scale images for a selected endmember spectrum, one scale image estimate the relative degree of match to the endmember spectrum and the approximate sub-pixel abundance, the other is an infeasibility image used to reduce the number of false positives. Better matched spectra from combination image which was calculated by dividing infeasibility image into the scale image will have a higher abundance and a low infeasibility value (fig. 6(c)).

#### 4.3 Multiple classifiers combination

Due to the complexity of land cover, statistical distribution characteristics of the remote sensing data vary in time and space, so the classification accuracy of different classifiers are obviously different. Experiments reported by the authors and other researchers have clearly shown that the superiority of one algorithm over another cannot be claimed for remote-sensing image classification (Giacinto et al., 2000). Multiple classifiers combination will improve the classification accuracy by using different classifiers complementarities.

Multiple classifiers systems are special cases of approaches that integrate several data-driven models for the same problem. Its key goal is to obtain a better composite global model, with more accurate and reliable estimates or decisions (Ghosh, 2002). The theory of multi-classifier systems can be traced back at least as far as 1965 (Nilsson, 1965). Previous multiple classifiers combination algorithm include the voting (Lam & Suen, 1994), Bayes rule (Xu et al., 1992), Dempster-Shafer theory (Mandler & Schuermann, 1988), decision tree and other methods. Based on classifier outputs, the multiple classifiers combination methods can be classified into three different levels: abstract level, ranked list of classes, and measurements (Suen & Lam, 2000, Xu et al., 1992). Based on manipulating training samples, they can be classified into two approaches: boosting (Freund & Schapire, 1996) and bagging (Breiman, 1996). Ideally, the combination should take advantage of the merit of the individual classifiers, avoid their weaknesses and improve classification accuracy.

Multiple classifiers combination was successfully used in remotely sensed data classification. However, the standard multiple classifiers combination methods were seldom used in remotely sensed data classification. (Wen et al., 2007a) combined three MF results

using adjusting threshold method, and demonstrated the result was better than only one MF result. However, it is hard to get the effective split value only using adjusting the threshold by hand. In this chapter, CART (Breiman et al., 1984, Wen et al., 2008a) was used to extract the rock body information from the hyperspectral remote sensing image by combine three MF results. CART method was suggested by Breiman et al. It partitions the data into two subsets of records with similar values for the target attribute so that the records within each subset are more homogeneous than in the previous subset. It is a recursive process: each of those two subsets is then split again, and the process repeats until the homogeneity criterion is reached or until some other stopping criterion is satisfied. It allows unequal misclassification costs to be considered in the tree growing process. The decision trees produced by CART are strictly binary, containing exactly two branches for each decision node. The CART algorithm grows the tree by conducting for each decision node, an exhaustive search of all available variables and all possible splitting values, selecting the optimal split according to the certain criteria.

## 5. Conclusion

The decision tree constructed by CART algorithm is shown in fig. 5. "Y" in the rectangle means the pixel contains dioritic porphyrite. On the contrary, "N" in the rectangle means the pixel contains no dioritic porphyrite. The value near the rectangle is confidence. From fig. 5, when the value of Multi Range SFF and MTMF rule image are greater than the certain threshold, it concludes the pixel contains dioritic porphyrite. They are coincided with their physical meaning for that the higher value of Multi Range SFF and MTMF rule image is

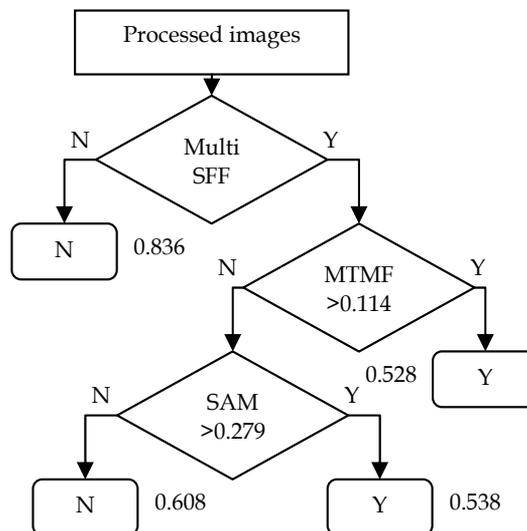


Fig. 5. Decision tree constructed by CART algorithm

related to higher material abundance. However, when the value of SAM rule image is greater than the certain threshold, it concludes the pixel contains dioritic porphyrite. It is inconsistent with its physical meaning for the smaller value of SAM rule image represents closer matches to the endmember spectra. The reason maybe is that SAM only compares the

shape of spectra, so its ability of pixel un-mixing is limited. Comparing three methods, the decision tree also show that Multi Range SFF has great pixel un-mixing ability; followed by MTMF; SAM is the least. Most alteration minerals have diagnostic spectral absorption features in the short wave and mid-infrared. Multi Range SFF compares enhanced spectra absorption features, so it is effective. Three MF results and combining three MF results using CART were used to extract dioritic porphyrite respectively. The rule images generated by different methods are shown in fig. 6. Comparing four rule images, the combination rule image is better than others significantly. Using dioritic porphyrite areas in geologic map as ground truth, the classification accuracy are calculated and results are shown in table 1. As is shown from table 1, multiple classifier combination has the best classification performance, and it generates the highest overall accuracy and hit/false alarm. The final Dioritic porphyrite map of Pulang using combining multiple classifiers is shown in fig. 7. Comparing with the field data and geology map, it concludes that this method is effective.

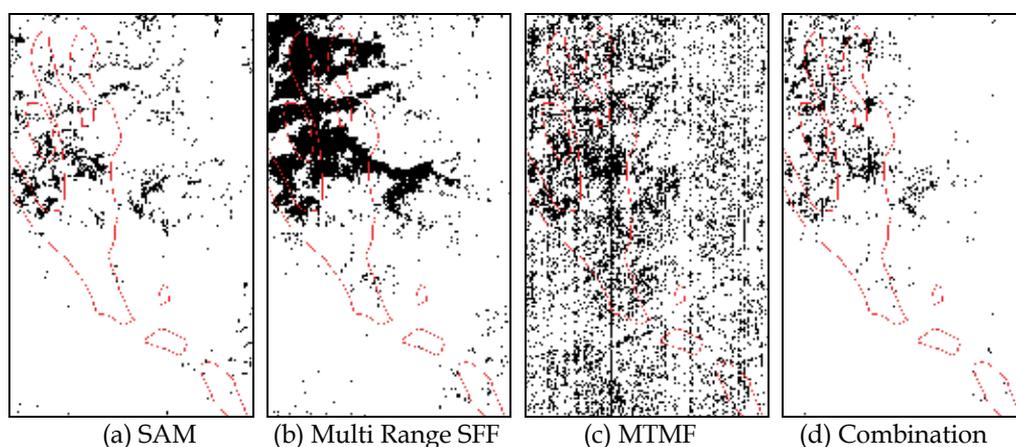
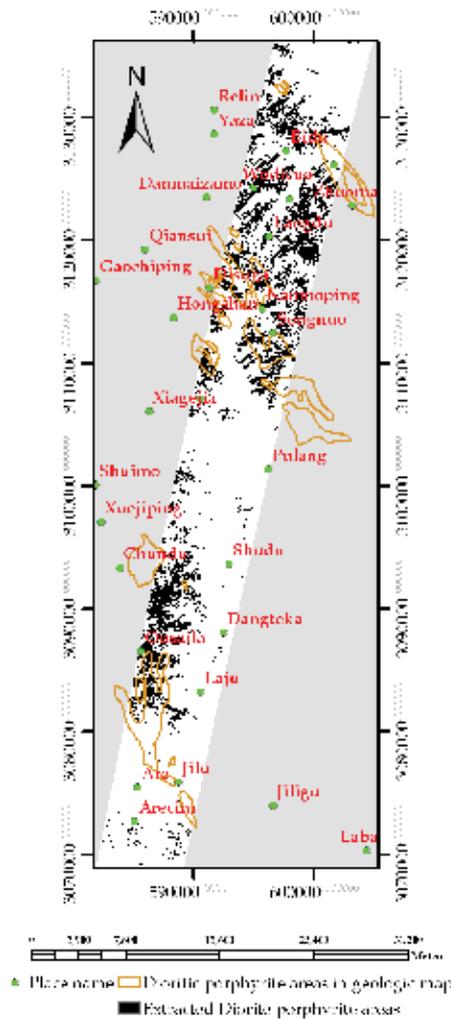


Fig. 6. Rule images generated by different methods

Method	Hit (pixels)	False alarm (pixels)	Overall accuracy	Hit/false alarm
SAM	1858	2972	0.769	0.625
Multi SFF	6971	9137	0.756	0.763
MTMF	6809	13702	0.698	0.497
Combination	2524	2239	0.786	1.127

Table 1. The classification result using different method

SAM is a physically-based spectral classification that determines the spectral similarity. Multi Range SFF is an absorption feature based methodology. MTMF, a special type of spectral mixture analysis, is based on well-known signal processing methodologies. Combining of three matched filtering method can take full advantage of three MF methods, so the classification accuracy is significant improvements than only one approach.



Map Projection: Gauss Kruger, False Easting: 500 kilometer, Central Meridian: 99°.

Fig. 4. Dioritic porphyrite map of Pulang using combining multiple classifiers

## 6. Reference

- Anderson, G. P., et al., (1999). FLAASH and MODTRAN4: state-of-the-art atmospheric correction for hyperspectral data. *IEEE Proceedings of Aerospace Conference*, pp. 177-181, Snowmass at Aspen, CO, USA.
- Baatz, M. & Schäpe, A., (2000). Multiresolution segmentation-an optimized approach for high quality multiscale image segmentation. *AGIT XIII*, pp. 12-23, Wichmann, Heidelberg, Germany.
- Baugh, W. M., et al., (1998). Quantitative geochemical mapping of ammonium minerals in the southern Cedar Mountains, Nevada, using the airborne visible/infrared

- imaging spectrometer (AVIRIS). *Remote Sensing of Environment*, Vol. 65, No. 3, 292-308.
- Berk, A., et al., (1998). MODTRAN Cloud and Multiple Scattering Upgrades with Application to AVIRIS. *Remote Sensing of Environment*, Vol. 65, No. 3, 367-375.
- Berman, M., et al., (2004). ICE: a statistical approach to identifying endmembers in hyperspectral images. *Geoscience and Remote Sensing, IEEE Transactions on*, Vol. 42, No. 10, 2085-2095.
- Boardman, J. W., (1993). Automated spectral unmixing of AVIRIS data using convex geometry concepts: in Summaries. *Fourth JPL Airborne Geoscience Workshop*, pp. 11-14, Arlington, Virginia, JPL Publication.
- Boardman, J. W., (1998). Leveraging the high dimensionality of AVIRIS data for improved sub-pixel target unmixing and rejection of false positives: mixture tuned matched filtering. *Summaries of the Seventh Annual JPL Airborne Geoscience Workshop*, pp. 55-56, Pasadena, CA: NASA Jet Propulsion Laboratory, JPL Publication 97-1.
- Boardman, J. W., et al., (1995). Mapping Target Signatures Via Partial Unmixing of Aviris Data. *Summaries of the Fifth Annual JPL Airborne Earth Science Workshop*, pp. 23-26, Washington, D. C, JPL Publication 95-1.
- Breiman, L., (1996). Bagging Predictors. *Machine Learning*, Vol. 24, No. 2, 123-140.
- Brieman, L., et al., (1984). *Classification and Regression Trees*, Chapman & Hall/CRC Press. Boca Raton, FL.
- Clark, R. N. & Roush, T. L., (1984). Reflectance spectroscopy-Quantitative analysis techniques for remote sensing applications. *Journal of Geophysical Research*, Vol. 89, No. B7, 6329-6340.
- Freund, Y. & Schapire, R. E., (1996). Experiments with a new boosting algorithm. *Machine Learning: Proceedings of the Thirteenth International Conference*, pp. 156, Morgan Kaufman, San Francisco.
- Ghosh, J., (2002). Multiclassifier Systems: Back to the Future. *Proceedings of the Third International Workshop on Multiple Classifier Systems*, pp. 1-15, Cagliari, Italy, Springer.
- Giacinto, G., et al., (2000). Combination of neural and statistical algorithms for supervised classification of remote-sensing images. *Pattern Recognition Letters*, Vol. 21, No. 5, 385-397.
- Goetz, A. F. H., et al., (1982). Mineral Identification from Orbit: Initial Results from the Shuttle Multispectral Infrared Radiometer. *Science*, Vol. 218, No. 4576, 1020-1024.
- Goetz, A. F. H., et al., (1985). Imaging Spectrometry for Earth Remote Sensing. *Science*, Vol. 228, No. 4704, 1147-1153.
- Green, A. A., et al., (1988). A transformation for ordering multispectral data in terms of imagequality with implications for noise removal. *Geoscience and Remote Sensing, IEEE Transactions on*, Vol. 26, No. 1, 65-74.
- Hsu, P.-H., (2007). Feature extraction of hyperspectral images using wavelet and matching pursuit. *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 62, No. 2, 78-92.
- Kokaly, R. F. & Clark, R. N., (1999). Spectroscopic Determination of Leaf Biochemistry Using Band-Depth Analysis of Absorption Features and Stepwise Multiple Linear Regression. *Remote Sensing of Environment*, Vol. 67, No. 3, 267-287.
- Kruse, F. A., (2005). Multi-resolution segmentation for improved hyperspectral mapping. *Proceedings, SPIE Symposium on Defense & Security*, pp. 161, Orlando, FL.

- Kruse, F. A., et al., (1993). The Spectral Image-Processing System (SIPS) - Interactive Visualization and Analysis of Imaging Spectrometer Data. *Remote Sensing of Environment*, Vol. 44, No. 2-3, 145-163.
- Lam, L. & Suen, C. Y., (1994). A theoretical analysis of the application of majority voting topattern recognition. *Proceedings of the 12th IAPR International Conference on Pattern Recognition and Computer Vision & Image Processing*, pp. 418 - 420, Jerusalem.
- Mandler, E. & Schuermann, J., (1988). Combining the classification results of independent classifiers based on the Dempster/Shafer theory of evidence. *Pattern Recognition and Artificial Intelligence*, pp. 381-393, Amsterdam, Netherlands, Elsevier Science.
- Nilsson, N. J., (1965). *Learning machines*, McGraw-Hill. New York.
- Pal, M. & Mather, P. M., (2003). An assessment of the effectiveness of decision tree methods for land cover classification. *Remote Sensing of Environment*, Vol. 86, No. 4, 554-565.
- Richter, R., (1996). A spatially adaptive fast atmospheric correction algorithm. *International Journal of Remote Sensing*, Vol. 17, No. 6, 1201-1214.
- Richter, R., (1998). Correction of satellite imagery over mountainous terrain. *Applied Optics*, Vol. 37, No. 18, 4004 - 4015.
- Smith, G. M. & Milton, E. J., (1999). The use of the empirical line method to calibrate remotely sensed data to reflectance. *International Journal of Remote Sensing*, Vol. 20, No. 13, 2653-2662.
- Suen, C. Y. & Lam, L., (2000). Multiple classifier combination methodologies for different output levels. *Proceedings of First International Workshop on Multiple Classifier Systems (MCS 2000)*, pp. 52-66, Sardinia, Italy, Springer.
- van der Meer, F. & Bakker, W., (1997). Cross correlogram spectral matching: Application to surface mineralogical mapping by using AVIRIS data from Cuprite, Nevada. *Remote Sensing of Environment*, Vol. 61, No. 3, 371-382.
- Vane, G. & Goetz, A. F. H., (1988). Terrestrial imaging spectroscopy. *Remote Sensing of Environment*, Vol. 24, No. 1, 1-29.
- Wen, X., et al., (2007a). Combining the Three Matched Filtering Methods in Mineral Information Extraction from Hyperspectral Data. *Journal of China University of Geosciences*, Vol. 18, No. Special Issue, 294-296.
- Wen, X., et al., (2007b). A Simplified Method for Extracting Mineral Information From Hyperspectral Remote Sensing Image Using SAM Algorithm. *12th Conference of International Association for Mathematical Geology, Geomathematics and GIS Analysis of Resources, Environment and Hazards*, pp. 526-529, Beijing, China.
- Wen, X., et al., (2008a). CBERS-02 remote sensing data mining using decision tree algorithm. *First International Workshop on Knowledge Discovery and Data Mining*, pp. 86-89, Adelaide, Australia.
- Wen, X., et al., (2008b). An investigation of the relationship between land cover ratio and urban heat island. *2008 International Congress on Image and Signal Processing*, pp. 682-686, Sanya, Hainan, China.
- Winter, M. E., (1999). Fast autonomous spectral endmember determination in hyperspectral data. *Proceedings of the Thirteenth International Conference on Applied Geologic Remote Sensing*, pp. 337-344, Vancouver, British Columbia, Canada.
- Xu, L., et al., (1992). Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions on Systems Man and Cybernetics*, Vol. 22, No. 3, 418-435.

# Content-based Image Classification via Visual Learning

Hiroki Nomiya<sup>1</sup> and Kuniaki Uehara<sup>2</sup>

<sup>1</sup>*Graduate School of Science and Technology, Kobe University*

<sup>2</sup>*Graduate School of Engineering, Kobe University  
Japan*

## 1. Introduction

As the information processing ability of computers improves, real-world images are increasingly used in various data mining applications. Thus, flexible and accurate image recognition methods are strongly needed. However, real-world images generally contain a wide variety of objects which have complex features (e.g., shapes and textures). Therefore, the accurate recognition of real-world objects is difficult because of three main problems. Firstly, although an image is generally given as a set of pixels, pixels alone are insufficient for the description and recognition of complex objects. Thus, we must construct more discriminative features from pixels. Secondly, finding useful features to describe complex objects is problematic because appropriate features are dependent on the objects to be recognized. Thirdly, real-world images often contain considerable amounts of noise, which can make accurate recognition quite difficult. Because of these problems, the recognition performance of current recognition systems is far from adequate compared with human visual ability.

In order to solve these problems and facilitate the acquisition of a level of recognition ability comparable to that of human visual systems, one effective method consists of introducing learning schemes into image understanding frameworks. Based on this idea, *visual learning* has been proposed (Krawiec & Bhanu, 2003). Visual learning is a learning framework which can autonomously acquire the knowledge needed to recognize images using machine learning frameworks. In visual learning, given images are statistically or logically analyzed and recognition models are constructed in order to recognize unknown images correctly for given recognition tasks. Visual learning attempts to emulate the ability of human beings to acquire excellent visual recognition ability through observing various objects and identifying several features by which to discriminate them.

The key to the development of an efficient visual learning model resides in features and learning models. Image data contain various types of informative features such as color, texture, contour, edge, spatial frequency, and so on. However, these features are not explicitly specified in input image data. Therefore, *feature construction* is needed. Feature construction is the process of constructing higher-level features by integrating multiple lower-level (primitive) features. Appropriate feature construction will greatly contribute to recognition performance. In addition, since useful features depend on the given image data,

*feature selection* is also needed to determine appropriate features according to the given image data. Thus, we propose both feature construction and feature selection methods to develop a better visual learning framework.

In order to utilize features efficiently, it is necessary to develop an appropriate visual learning model. In most existing visual learning models, a single learner based on a single learning algorithm is trained using input image data. However, this learning model is inefficient compared with human visual models. Human visual systems can be divided into multiple modules, accomplishing excellent visual ability through the cooperation of these modules (Marr 1982). In other words, introducing modularity into visual learning framework leads to improved recognition performance. To introduce modularity, we adopt an *ensemble approach*, a kind of learning approach in which multiple learners called base learners (they correspond to modules) are simultaneously trained and their learning results integrated into a single hypothesis. Based on this ensemble approach, we develop a novel visual learning model in which multiple base learners can be trained through cooperation with each other. Through the introduction of cooperation among multiple learners, recognition accuracy can be considerably improved. The learning strategy of the proposed visual learning model enables more flexible and accurate recognition applicable to a wide variety of data mining tasks using various types of visual data. We verify the flexibility and recognition performance of our method through an object recognition experiment using real-world image data, and a facial expression recognition experiment using video data.

## 2. Visual learning

At the beginning of computer vision and image understanding research, the recognition process was human-intensive. That is, a large part of domain knowledge was defined and given by hand. As the amount and complexity of image data increase, however, conventional image recognition methods have difficulty in recognizing real-world objects for several reasons. First, it is quite difficult to provide sufficient domain-specific knowledge manually due to the complexity of large-scale real-world recognition problems. Next, although the recognition performance of conventional methods is sufficient for limited domains, such as facial recognition and hand-written character recognition, these methods have great difficulty in effecting a flexible recognition that can discriminate a wide variety of real-world objects. Finally, the recognition performance of these methods tends to be affected by noisy images which contain cluttered backgrounds, bad lighting conditions, occluded or deformed objects, and so on. These problems must be solved in order to develop a recognition framework comparable to human visual systems.

The first problem can be solved by the framework of visual learning. Visual learning autonomously acquires domain-specific knowledge in order to recognize images. This knowledge is derived by machine learning frameworks which statistically or logically analyze input image data and construct recognition models to recognize unknown images correctly for the given recognition task. Most machine learning frameworks can be easily applied to a wide variety of recognition problems and can provide domain-specific knowledge.

Although machine learning frameworks automatically derive domain-specific knowledge, however, they often have difficulty in acquiring the knowledge because of the second problem caused by the data structure of image data. That is, an image has various features which are useful for the discrimination of objects in the image: for example, color

histograms and spatial frequency are widely used to describe images. However, these features are not explicitly specified in input data because an input image is usually given only as a set of pixels. Generally, since an image consists of a large number of pixels, the input images contain a large amount of irrelevant or redundant data. To solve this issue, feature construction is required to construct more informative features from the given image data. For instance, color histograms can be constructed by analyzing the distribution of the intensity values of the given pixels. Since the learning efficiency and performance depend on the features input into the machine learning algorithm, the feature construction method has a great influence on recognition performance. In other words, constructing appropriate features leads to an accurate recognition which is able to solve the second problem. For real-world image recognition, a crucial problem stems from the large variety of objects to be recognized. The problem is that appropriate features are generally dependent on the given object. Therefore, it is essential for flexible recognition to develop an efficient feature selection method to select appropriate features according to the given object.

As for the third problem, noisy images, which contain occlusion, deformation, or bad lighting conditions, often worsen the learning performance. To reduce the influence of these noisy images, some kind of image preprocessing such as image filtering is frequently used (Krawiec & Bhanu, 2003). However, the elimination of any type of noise using image preprocessing is extremely difficult. Thus, we attempt to deal with noisy images by developing a noise-robust learning model based on the ensemble approach. In the learning model, interaction among multiple base learners provides robustness to noise by detecting and eliminating noisy images. We show an example of the learning model in Fig. 1.

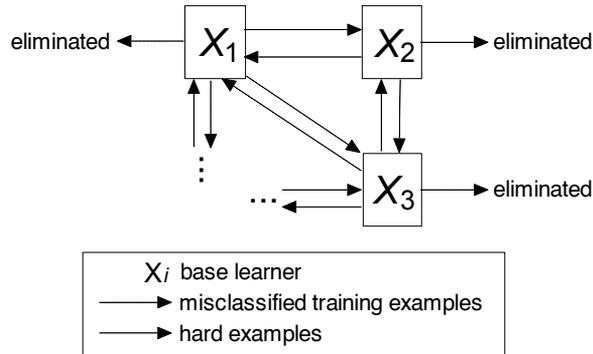


Fig. 1. The learning model with the collaboration of multiple base learners.

In this learning model, an arbitrary number of base learners are collaboratively trained. Specifically, each base learner  $X_i$  ( $i = 1, 2, \dots, n$ , where  $n$  is the number of base learners) sends its misclassified training examples to the other base learners  $\{X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n\}$ . The examples misclassified by  $X_i$  are eliminated from its training set. Similarly, the other base learners send their misclassified training examples to  $X_i$  and eliminate the examples from their own training sets.  $X_i$  is then trained using the examples sent from the other base learners.

The examples that are misclassified by all of the base learners are regarded as *hard examples* and are eliminated from the training sets of all base learners. Hard examples mean the examples which are very difficult to classify correctly; thus, noisy images correspond to

hard examples. Hard examples should be eliminated because they cause overfitting. Overfitting is the phenomenon in which base learners become too specialized through the recognition of hard examples, so that they often fail to recognize non-hard examples, which are the majority of given examples. Since resolving the problem of overfitting is sometimes crucial for ensemble learning, this learning model considerably reduces the influence of noisy images and improves recognition accuracy.

### 3. Collaborative visual learning

Since the key point of the learning model mentioned in the previous section is the collaboration of multiple base learners, we call this model *collaborative ensemble learning*. Its learning framework is based on a boosting algorithm (Freund & Schapire, 1997). Each example (i.e., image) is assigned a weight that measures the difficulty of correctly classifying the example, and each base learner is iteratively trained using these weights. The iteration step is called a round. The weights of all examples are updated at the end of each round to assess their classification difficulty. A higher weight means that the example is more difficult to classify correctly. In the machine learning domain, it has been proven that training a base learner using examples with high weights improves the classification performance of the base learner (Dietterich, 2000). At the end of each round, a base learner generates a hypothesis to classify unseen examples. Through the learning process, multiple hypotheses are generated and are ultimately integrated into a final hypothesis. The integration is performed by, for example, voting by multiple hypotheses. The final hypothesis corresponds to the prediction of an ensemble classifier and generally has much better classification performance than a hypothesis by a single base learner.

#### 3.1 A weighting algorithm to detect hard examples

To describe the collaborative ensemble learning model, we first formulate an object recognition task as a classification problem. The training set  $S$  is represented as  $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$ , where  $m$  is the number of training examples.  $x_i$  and  $y_i \in \{1, \dots, C\}$  correspond to an image and a class label respectively.  $C$  is the number of classes—that is, the recognition problem is to distinguish  $C$  kinds of objects.

Our learning algorithm is based on AdaBoost, which has the crucial problem of susceptibility to hard examples. This problem stems from the fact that AdaBoost gives excessively high weights to hard examples, so that overfitting tends to occur. In order to prevent overfitting, we improve the weighting algorithm that determines weights so that the weights of hard examples are appropriately controlled. We propose a weighting algorithm based on the following two points: (1) To prevent overfitting, when an example is regarded as a hard example by a base learner  $X$ , that example should not be used as a training example for  $X$ ; and (2) To train all base learners collaboratively, when  $X$  misclassifies a training example, its weight for other base learners should be increased so that the example is used to train the other base learners. We thus define the weight distribution for our weighting algorithm. The weight distribution represents the importance of each example. The examples which have high weight distribution values are useful to improve the classification performance of base learners. The weight distribution  $D_{i,t}^l$  of the  $i$ -th example  $x_i$  for the  $l$ -th classifier  $X^l$  at the  $t$ -th round is calculated as follows:

$$D_{i,t}^l = \frac{\delta_{i,t}^l d_{i,t}^l}{\sum_{i=1}^m \delta_{i,t}^l d_{i,t}^l}, \quad (1)$$

where  $\delta_{i,t}^l$  is 0 if  $x_i$  is regarded as a hard example by  $X^l$  (i.e., the weight of  $x_i$  becomes higher than the threshold<sup>1</sup>); otherwise  $\delta_{i,t}^l$  is 1, and

$$d_{i,t}^l = \frac{1}{n-1} \sum_{j \neq l}^n \left( \frac{w_{i,t}^j}{\sum_{i=1}^m w_{i,t}^j} \right), \quad (2)$$

where  $n$  is the number of base learners and  $w_{i,t}^j$  is the weight of  $x_i$  (calculated in the same way as a weight used in AdaBoost) for  $X^l$  at the  $t$ -th round. When an example  $x_i$  is regarded by  $X^l$  as a hard example,  $\delta_{i,t}^l = 0$ . Thus, from equation (1),  $D_{i,t}^l$  is 0 and  $x_i$  is not used in any subsequent rounds. In this way, hard examples are removed from the training set. Equation (2) represents the collaboration of base learners.  $d_{i,t}^l$  is determined based on the weights of other base learners. Since the weight  $w_{i,t}^j$  of a misclassified example increases, the values of both  $d_{i,t}^l$  and  $D_{i,t}^l$  increase when most of the other base learners misclassify the example. Consequently, the example is learned by  $X^l$ .

Here, we show an example of the weighting process. We consider a case in which the training set consists of six examples  $\{x_1, x_2, x_3, x_4, x_5, x_6\}$  whose class labels are  $\{c_1, c_1, c_2, c_2, c_3, c_3\}$  respectively, and three base learners  $X^1, X^2$  and  $X^3$  are trained simultaneously. Assuming that each base learner classifies each example at the first round as shown in Table 1 (a) (the class labels shown in boldface represent the correct classification), the weight distribution of each example for the second round is calculated according to equation (1) as shown in Table 1(b).

For example,  $x_4$  was correctly classified by  $X^1$  and  $X^2$  while misclassified by  $X^3$ . On the other hand,  $x_2$  was correctly classified only by  $X^3$ . Thus,  $x_4$  should be learned by  $X^1$  and  $X^2$  while  $x_2$  should be learned by  $X^3$ . From Table 1 (b), the weight distribution of  $x_4$  for  $X^1$  and  $X^2$  is much higher than for  $X^3$ . The weight distribution of  $x_2$  for  $X^3$  is higher than for  $X^1$  and  $X^2$ . Since each base learner learns the examples which have higher weight distributions, both  $x_2$  and  $x_4$  (as well as the other examples) are learned by appropriate learners in the next round.

	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$
$X^1$	<b><math>c_1</math></b>	$c_2$	$c_3$	<b><math>c_2</math></b>	<b><math>c_3</math></b>	<b><math>c_3</math></b>
$X^2$	$c_2$	$c_3$	<b><math>c_2</math></b>	<b><math>c_2</math></b>	$c_1$	<b><math>c_3</math></b>
$X^3$	<b><math>c_1</math></b>	<b><math>c_1</math></b>	<b><math>c_2</math></b>	$c_3$	<b><math>c_3</math></b>	$c_2$

(a) Predicted class labels

	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$
$X^1$	.20	.13	.08	.20	.20	.20
$X^2$	.09	.15	.22	.22	.09	.22
$X^3$	.21	.22	.21	.08	.21	.08

(b) Weight distributions

Table 1. Predicted class labels and weight distributions.

<sup>1</sup> The threshold is determined by searching for the optimal value using the beam search, based on the NadaBoost algorithm (Nakamura et al., 2004), which is more robust to hard examples than AdaBoost.

### 3.2 Construction of an ensemble classifier using collaboration of base learners

Assuming that the number of rounds is  $T$ , the  $l$ -th base learner trained in the  $t$ -th round is represented as  $X_t^l$  ( $t = 1, \dots, T$ ). At the end of the  $T$ -th round, we integrate the base learners  $\{X_t^l\}_{t=1}^T$  into an ensemble classifier  $X^l$  and determine the prediction of  $X^l$  by integrating the predictions of all base learners  $\{X_t^l\}$  using a weighted voting method. Specifically, the prediction of the ensemble classifier  $X^l$  is determined as the class label which is predicted by the majority of base learners. The prediction  $X^l(x)$  of the ensemble classifier  $X^l$  for an example  $x$  is defined as follows:

$$X^l(x) = \arg \max_{c \in \{1, \dots, C\}} \sum_{t=1}^T \alpha_t^l [X_t^l(x) = c], \quad (3)$$

where  $[X_t^l(x) = c]$  is 1 if  $X_t^l(x) = c$  and otherwise is 0.  $\alpha_t^l = \log\{(1 - \varepsilon_t^l)/\varepsilon_t^l\}$ , where  $\varepsilon_t^l$  is the classification error of  $X_t^l$ . Thus, a higher value for  $\alpha_t^l$  means a lower classification error. In the learning process, each base learner is specialized to classify non-hard examples precisely. Thus, we must determine whether a given example is a hard example or a non-hard example. To distinguish hard examples from non-hard examples, we define a criterion and call it *class separability*. Class separability is defined so that it is proportional to the classification performance of a base learner for each class. When  $X_t^l$  can correctly classify the examples whose class labels are  $c$ , the class separability for class  $c$  is high because  $X_t^l$  can distinguish these examples from the other examples. On the other hand, if  $X_t^l$  misclassifies these examples, the class separability for class  $c$  is low. Since hard examples are frequently misclassified, the class separability of a hard example will be much lower than that of a non-hard example. Here, we consider the case in which  $X_t^l$  predicts the class label of an unknown example  $x$  as class  $c$ . If the class separability of  $X_t^l$  for class  $c$  is high,  $x$  will be a non-hard example. That is, the possibility that the prediction is correct will be high. We define the class separability  $s_t^l(c)$  for class  $c$  as follows:

$$s_t^l(c) = \begin{cases} s_{t+}^l(c) & \text{if } X_t^l(x) = c, \\ s_{t-}^l(c) & \text{otherwise,} \end{cases} \quad (4)$$

where

$$X_t^l(x) = \arg \max_{c \in \{1, \dots, C\}} \sum_{\tau=1}^t \alpha_\tau^l [X_\tau^l(x) = c],$$

$$s_{t+}^l(c) = \frac{n_{t,c,c}^l}{\sum_{i=1}^C n_{t,c,i}^l}, \quad s_{t-}^l(c) = \frac{\sum_{i \neq c}^C \sum_{j \neq c}^C n_{t,i,j}^l}{\sum_{i \neq c}^C \sum_{j=1}^C n_{t,i,j}^l}.$$

$n_{t,i,j}^l$  denotes the number of examples whose class label is  $i$  and which were classified into class  $j$ .  $s_{t+}^l(c)$  is high when the examples whose class labels are  $c$  are correctly classified

into class  $c$ .  $s_{i-}^l(c)$  is high when the examples whose class labels are  $c'$  ( $c' \neq c$ ) are classified into class  $c'$ . If the prediction  $X_i^{l'}(x)$  of the  $l$ -th classifier is  $c$ , then  $s_{i+}^l(c)$  is used as the class separability. Otherwise,  $s_{i-}^l(c)$  is used as the class separability. For example, in the case shown in Table 1,  $s_{i+}^1(c_1) = 1/2$  because  $X^1$  correctly classifies  $x_1$  and misclassifies  $x_2$ .  $s_{i-}^1(c_1) = 4/4 = 1$  because  $X^1$  correctly classifies  $x_3, x_4, x_5$  and  $x_6$ . Similarly, the class separabilities of the other base learners for the class  $c_1$  are calculated as follows:  $s_{i+}^2(c_1) = 0$ ,  $s_{i-}^2(c_1) = 3/4$ ,  $s_{i+}^3(c_1) = 1$ , and  $s_{i-}^3(c_1) = 1$ . These values of class separability for  $c_1$  indicate that  $X^3$  will give the most accurate prediction in the classification of the examples whose class labels are  $c_1$ .

When classifying an unseen example, we first obtain the predictions of all classifiers  $\{X^l\}_{l=1}^n$ , where  $n$  is the number of features. We next calculate the class separability of each classifier and select the classifier with the highest class separability as the most reliable classifier. The prediction  $F(x)$  of the ensemble classifier for an example  $x$  is then determined by the following equation:

$$F(x) = X^{l^*} \text{ such that } l^* = \arg \max_l \sum_{\tau=1}^l s_{i-}^{\tau'}(X_{i-}^{\tau'}(x)). \quad (5)$$

Finally, we show the algorithm list of the collaborative ensemble method as follows:

1. Initialize:  $t = 1$ ,  $D_{i,1}^l = 1/m$  and  $\delta_{i,1}^l = 1$  for all  $i$  and  $l$ ,  $T\_list = \{\}$ .  $T\_list$  is the list to retain up to  $b$  weight thresholds, where  $b$  is beam width.
2. For each base learner, construct the training set  $S_i^l$  by sampling from the original training set  $S$  according to the weight distribution  $D_{i,t}^l$ .
3. Train each base learner using  $S_i^l$  and obtain  $X_i^l$ .
4. If  $t = T$  and  $T\_list$  is empty, then make the final prediction  $F$  and finish the learning process; otherwise, go to step 5.
5. For all  $l$ , classify all training examples using  $X_i^l$ , then decrease the weights of correctly classified examples and increase the weights of misclassified examples.
6. Obtain the possible thresholds  $W_i^l$ .
7. Calculate the accuracy of each threshold by estimating the classification accuracy using the threshold.
8. Add the threshold to  $T\_list$ .
9. If the number of thresholds in  $T\_list$  is more than  $b$ , remove the thresholds which have lower accuracy.
10. Select the most accurate threshold from  $T\_list$  to detect hard examples, then remove the threshold from  $T\_list$ .
11. Set  $\delta_{i,t+1}^l$  to 0 if  $x_i$  is regarded as a hard example by  $X_i^l$ , otherwise to 1.
12. Calculate the weight distribution  $D_{i,t+1}^l$  for each  $l$  and  $i$ .

13.  $t \leftarrow t + 1$  and go to step 2.

In the above algorithm, we efficiently search for the optimal (or suboptimal) threshold based on beam search. The process corresponds to steps 5 to 10 above. After all base learners are trained, each weight is updated in step 5. According to the weights, the possible thresholds  $W_t^l$  are determined in step 6. We next evaluate the classification accuracy for each threshold in step 7. In step 8, the thresholds are added to a list  $T\_list$ , which is used by beam search to restrict the search space.  $T\_list$  retains at most  $b$  thresholds, where  $b$  corresponds to the beam width. If the number of possible thresholds is higher than  $b$ , the thresholds which have a lower accuracy are removed from  $T\_list$  in step 9 and are not used for the search. In step 10, a threshold  $W_t^{l*}$  which has the highest accuracy is selected and removed from  $T\_list$ . Hard examples are detected using  $W_t^{l*}$ . This process is repeated while  $T\_list$  is not empty. As a result, the optimal (or suboptimal) threshold can be found.

### 3.3 Experiment

We carried out several object recognition experiments to verify the performance of our method using the images in the ETH-80 Image Set database (Leibe & Schiele, 2003). This data set contains 8 different objects: apples, cars, cows, cups, dogs, horses, pears, and tomatoes. We used 20% of the examples as training examples and the remainder as test examples. The number of rounds was experimentally set to 100. We constructed five types of base learners using five types of features as given in Fig. 2.

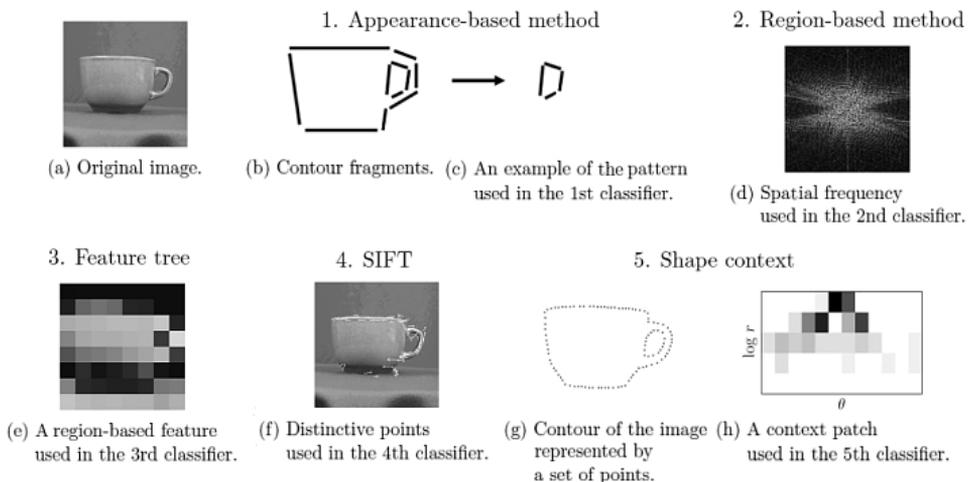


Fig. 2. The features used in this experiment.

The first base learner is an appearance-based recognition method which utilizes contour fragments (Nomiya & Uehara, 2007), as shown in (b). The set of contour fragments is grouped into meaningful structures called patterns as depicted in (c). The second base learner is based on the distributions of pixel intensity values (Nomiya & Uehara, 2007). The distributions are represented by a Generic Fourier Descriptor (GFD). A GFD is obtained by calculating the spatial frequency of an image as described in (d). The third base learner is a feature tree (Nomiya & Uehara, 2005). This method generates region-based features by image filtering as shown in (e). It then combines several features into several decision trees

called feature trees. The fourth base learner is based on Scale Invariant Feature Transform (SIFT) (Lowe, 2004). SIFT generates deformation-invariant descriptors by finding distinctive points in objects (indicated by white arrows in (f)). These points represent the characteristics of the object based on image gradients. The fifth base learner uses the shape context method (Belongie et al., 2001). In this method, the contour of an object is described by a set of points as illustrated in (g). Using this set of points, log-polar histograms of the distance and angle between two arbitrary points, called shape contexts, are calculated for all points. An example of shape context is given in (h). This method discriminates the object by matching its shape contexts with the shape contexts in the training set.

In this experiment, we evaluate the proposed method from the following three viewpoints: firstly, in order to verify the effectiveness of the learning model of the proposed method, we evaluate the recognition performance of the collaborative ensemble learning model. Secondly, we assess the usefulness of the proposed method as an object recognition method by comparing it with several existent object recognition methods. Finally, we use noisy image data to verify the robustness of the proposed method to noise.

### 3.3.1 Evaluation of collaborative ensemble learning model

To verify the effectiveness of our collaborative ensemble learning model, we construct four types of ensemble classifiers,  $L_2$ ,  $L_3$ ,  $L_4$  and  $L_5$ , by integrating two, three, four and five base learners respectively.  $L_i$  consists of the first, second, ..., and  $i$ -th base learners. We then compare their performance. The result of the experiment is provided in Fig. 3.

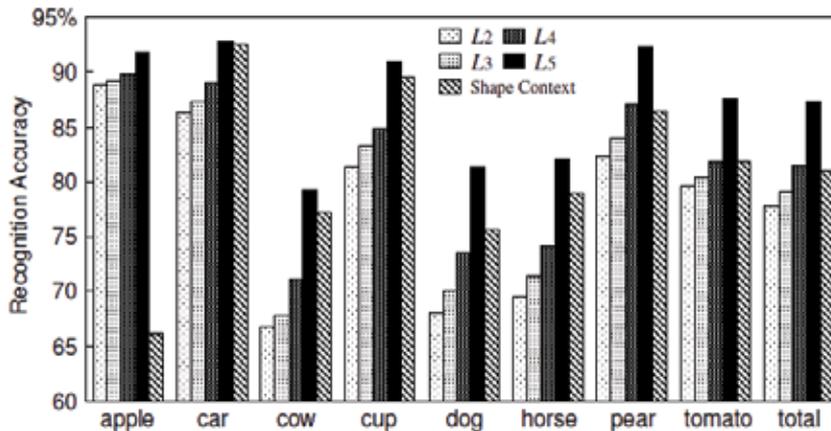


Fig. 3. The recognition accuracy for each ensemble classifier and shape context method.

The recognition accuracy is proportional to the number of integrated base learners. In particular, the accuracy improves significantly for animals with complex shapes and textures. This result implies that diverse features are required to discriminate correctly between complex objects and that our method can effectively utilize various features.  $L_5$  achieved much higher accuracy than the other ensemble classifiers. Although this improvement is due to the high classification performance of the shape context learner (i.e., the fifth base learner), our method fully outperforms the shape context method. Since the shape context method depends on the shapes of objects, it sometimes fails to distinguish between objects which have similar shapes, such as apples and tomatoes. Thus, our method

selects an appropriate base learner other than the shape context learner for the classification of apples and tomatoes, leading to a higher recognition accuracy for our method.

### 3.3.2 Comparison with other object recognition methods

We compare our recognition performance with those of the following six object recognition methods. The first is the shape context (Belongie et al., 2001). The second is the multidimensional receptive histogram (Schiele & Crowley, 2000), which describes the shapes of objects using statistical representations. The third is color indexing (Swain & Ballard, 1991), which discriminates an object using RGB histograms calculated from all the pixels in the object. The fourth is based on local invariant features (Grauman & Darrell, 2005) that are generated by a gradient-based descriptor and are robust to the deformation of images. The fifth is the learning-based recognition method (Marée et al., 2005), in which an object is described by randomly extracted multi-scale subwindows in the image and classified by an ensemble of decision trees. The sixth is the boosting-based recognition method (Tu, 2005), in which a probabilistic boosting-tree framework is introduced to construct discriminative models. The recognition accuracy is shown in Table 2.

Swain & Ballard (1991)	64.85	Grauman & Darrell (2005)	81
Marée et al. (2005)	74.51	Belongie (2001)	86.40 <sup>2</sup>
Tu (2005)	76	<b>Proposed method</b>	<b>87.27</b>
Schiele & Crowley (2000)	79.79		

Table 2. The recognition accuracy for each object recognition methods (in %).

Our recognition accuracy is higher than those of all other recognition methods. We utilize multiple features for recognition and thus can discriminate a wider variety of objects than single-feature recognition methods. In addition, this result indicates that our learning strategy for selecting optimal base learners is effective. Since the criterion for the determination of optimal base learners is determined by observing the collaborative learning process, this result indicates the effectiveness of our collaborative learning framework.

### 3.3.3 Robustness over hard examples

In order to verify the robustness to hard examples of our method, we carry out an experiment using the Caltech image data set, which contains many hard examples. We use the images of six kinds of objects from the data set: airplanes, cars, Dalmatians, faces, leopards and motorbikes. We use 20% of the examples as training examples and the remainder as test examples. We construct two types of ensemble classifiers and compare these ensemble classifiers with our method. The first ensemble classifier,  $X_1$ , does not eliminate any hard examples and never increases the weights of hard examples even if they are misclassified. The second ensemble classifier,  $X_2$ , also does not eliminate hard examples,

---

<sup>2</sup> This recognition accuracy is reported by (Leibe & Schiele, 2003). This accuracy has been achieved, however, using over 98% of the examples in the training set while we use only 20%. In addition, its recognition accuracy is proportional to the number of the training examples as shown in (Belongie et al., 2001). The recognition accuracy using 20% of examples in the training set is 81.06%.

but it does increase the weights of all misclassified examples even if they are hard examples. First, we show the recognition accuracy for each class and total recognition accuracy in Fig. 4.

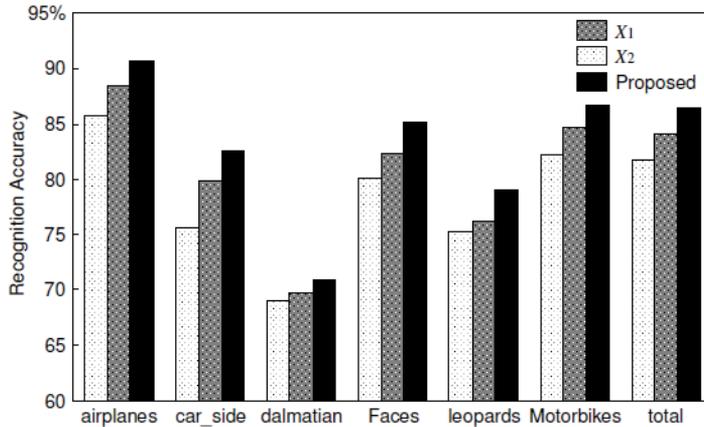


Fig. 4. The recognition accuracy for each class.

The recognition accuracy of  $X_2$  is the worst because it does not take hard examples into consideration.  $X_1$  outperforms  $X_2$  by giving smaller weights to hard examples and thus reducing their influence to some extent. Given that our recognition accuracy is the best for all objects, it seems that detecting and eliminating hard examples from training examples leads to more efficient learning.

Through these experiments, we confirm the effectiveness of our collaborative ensemble learning framework. However, it is difficult to determine the appropriate combination of base learners because finding the optimal combination is an open problem. Although our method enables an ensemble classifier to be less sensitive to the combination of classifiers due to the determination of the optimal base learner for a given example, the accuracy of the proposed method can be improved if we can determine the optimal combination of base learners. Thus, we should find an efficient method to determine the optimal combination.

## 4. Application to facial expression classification problem using multistream time-series data

### 4.1 Visual learning in facial expression recognition

In this section, we mention the application of our visual learning framework to problems in facial expression classification which is a challenging domain in computer vision understanding. In facial expression recognition, distinguishing slight differences in face images is required. However, it is quite difficult to accomplish this because the number of characteristic points on a face is small. Moreover, the movement of each point is subtle, while multiple points are mutually correlated in most facial expressions. For accurate recognition, a variety of recognition methods have been proposed using various features which can be extracted or constructed from input image data. We divide the features typically used in facial expression recognition into three levels: low-level features, medium-level features and high-level features.

Low-level features are based on the intensity of each pixel in an image. Since an image is generally given as a set of pixels and a pixel is described by its intensity value, low-level

features can be obtained by directly processing the given image data using some kind of statistical analysis and dimensionality reduction techniques, such as principal component analysis and linear discriminant analysis (Donato et al., 1999). Since low-level features can be directly and easily generated from the given image data, they can be utilized for a variety of recognition problems (e.g., object recognition and handwritten character recognition). However, low-level features seem to be rather insufficient for accurately distinguishing slight difference among various types of facial expressions.

In facial expression recognition, it is effective to observe the movement of particular parts in a face because facial expressions can be described as the combination of the movement of facial muscles. According to this idea, several features which are able to describe partial movement in a face have been proposed (Bourel et al., 2002). These features, which we call medium-level features, can more precisely describe characteristic parts of a face. By selecting appropriate parts which are relevant to the facial expression and analyzing their movement, more accurate recognition can be performed. However, since useful parts are highly dependent on the expression, finding appropriate parts is a difficult task.

Features more complex than medium-level features are defined by modeling whole face (Essa & Pentland, 1997). We call these features high-level features. The models are able to analyze multiple parts in a face simultaneously and can be more effective than medium-level features. High-level features are typically constructed by taking multiple images from various directions and generating 3D models from input 2D images. This can be a crucial problem, however, because it is troublesome and time-consuming to construct facial models through such processes.

Taking the tradeoff between the better quality of higher-level features and the lower cost of constructing lower-level features into account, we propose a facial expression recognition method based on medium-level features. Although the representational ability of a single medium-level feature is lower than that of a single high-level feature, more discriminative features can be constructed by combining multiple medium-level features. As was mentioned in Section 3, ensemble learning enables the utilization of multiple features. Thus, we can deal with multiple medium-level features by introducing our visual learning framework.

Generally, only a few medium-level features around salient parts on a face (e.g., eyes, eyebrows and mouth) are used for recognition when analyzing various facial parts, and finding the appropriate combination of facial parts is essential to distinguishing a wide variety of facial expressions. In order to solve this issue and find an appropriate combination depending on facial expressions, we utilize a motion capture system which can precisely observe the movement of diverse facial parts (Osaki et al., 2000).

The facial expression data obtained from the motion capture system are represented as multistream time-series data. Multistream time-series data consist of multiple time-series sequences which are mutually correlated. Since multistream time-series data generally contain a large amount of information which includes redundant data, it is necessary to select useful streams from the given streams in order to achieve accurate recognition. Thus, we propose an efficient method of assessing the usefulness of each stream and finding appropriate streams based on an effective criterion to measure the similarity among multiple streams. To verify the effectiveness of the proposed method, we perform several facial expression recognition experiments.

#### 4.2 Feature construction method

In order to generate facial expression data, we utilize a motion capture system to capture the movement of several points on a face. Specifically, the facial expression data are captured by 35 markers on the subject's face as depicted in Fig.5 and described by 35 streams which represent the movement of each marker.



Fig. 5. Markers used by motion capture system.

The location of each marker is determined according to the definition of Facial Action Coding System (FACS) (Ekman & Friesen, 1978), which is designed for measuring and describing facial behavior. FACS was developed by analyzing and determining the relationship between the contraction of each facial muscle and the appearance of the face. In FACS, specific measurement units called Action Units (AUs) are defined to describe facial expressions. AUs represent the muscular activity that leads to the changes in facial expression. Although numerous AUs are specified, the following 17 AUs are considered to be sufficient to describe basic facial expressions.

No.	Name	No.	Name	No.	Name
1	Inner Brow Raiser	9	Nose Wrinkler	17	Chin Raiser
2	Outer Brow Raiser	10	Upper Lip Raiser	20	Lip Stretcher
4	Brow Lowerer	12	Lip Corner Puller	23	Lip Tightener
5	Upper Lid Raiser	14	Dimpler	25	Lips Part
6	Cheek Raiser	15	Lip Corner Depressor	26	Jaw Drop
7	Lid Tightener	16	Lower Lip Depressor		

Table 3. Main Action Units (AUs).

In Table 3, for example, AU 1 corresponds to the raising of the inner corner of the eyebrow, while AU 4 corresponds to the puckering up of the outer corner of the eyebrow. Combining these AUs allows for various types of facial expressions to be described. We show the combinations of AUs for several basic facial expressions (Surprise, Anger, Happiness and Sadness) in Table 4.

Expression	AU numbers (intensity)
Surprise	1(100), 2(40), 5(100), 10(70), 12(40), 16(100), 26(100)
Anger	2(70), 4(100), 7(60), 9(100), 10(100), 12(40), 15(50), 26(60)
Happiness	1(60), 6(60), 10(100), 12(50), 14(60), 20(40)
Sadness	1(100), 4(100), 15(50), 23(100)

Table 4. Combination of AUs for each expression.

In Table 4, the numerical values in parentheses are intensity values of AUs. A higher intensity value means stronger activity of an AU, and the maximum value is 100. For

example, in Sadness, AUs 1, 4 and 23 are strongly activated and AU 15 is weakly activated because brows and lips show characteristic changes in the expression of sadness. Based on the combination of AUs, corresponding combinations of streams (i.e., markers shown in Fig. 5) are determined for each expression as illustrated in Fig. 6. In Fig. 6, the markers encircled by squares denote the corresponding streams.

The input data given by the motion capture system simply represents the movement of each marker. It is a kind of primitive feature and is thus inadequate for recognition. To construct higher-level features, we estimate the stress from each marker. This is because each facial expression is described as the movement of particular facial muscles which can be represented by the stress for each marker. Due to the difficulty of directly measuring the stress, we instead estimate the stress using finite element method (FEM). FEM is widely used for estimating the deformation of an object caused by the given stress. Since our goal is to obtain the stress given to each marker, we consider the inverse problem of FEM. That is, we estimate the stress from the deformation of facial muscles. In order to describe the stress estimation process, we first show the settings of this problem in Fig. 7.

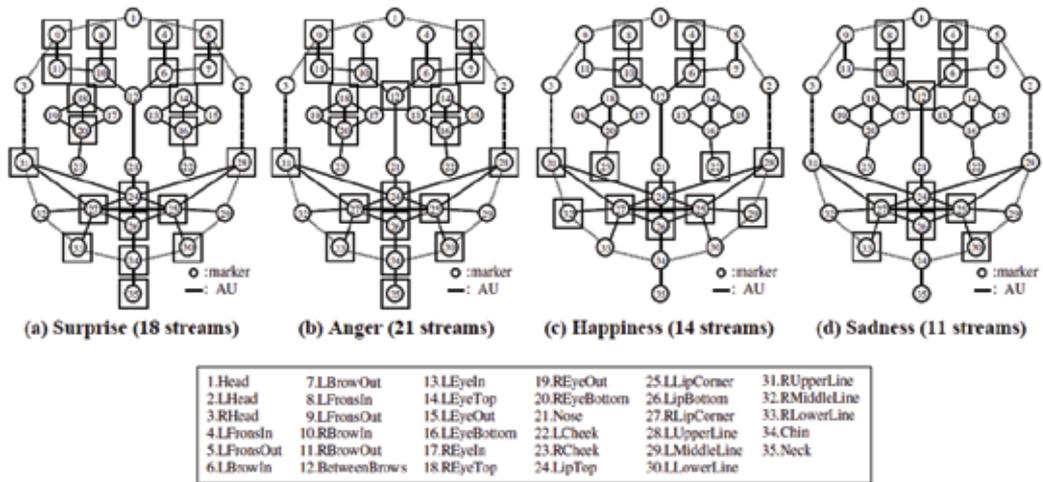


Fig. 6. Corresponding markers (streams) for each expression.

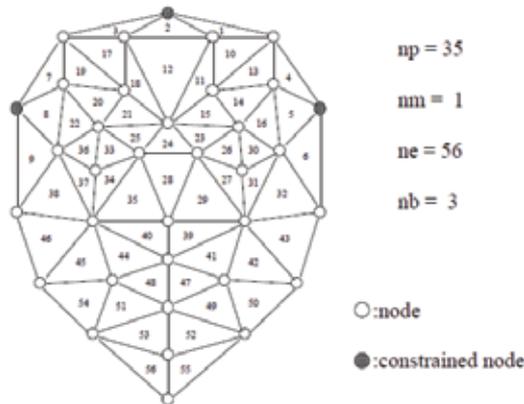


Fig. 7. The problem settings of inverse FEM for facial expression recognition.

Fig. 7 represents the settings of the inverse problem of FEM for facial expression recognition. A face is divided into several regions by the markers on the face. The regions and markers are called elements and nodes, respectively. Since we use 35 markers, the number of nodes  $np$  is 35. A node is represented by a circle in Fig. 7. Each element is defined as a triangular region whose vertices consist of three nodes. Thus, the number of elements  $ne$  is 56. We fix the number of materials  $nm$  at 1 because a face is uniformly covered with skin and set the number of constraints (i.e., the number of fixed points)  $nb$  at 3 because the three markers (Head, LHead and RHead) represented by  $\bullet$  are fixed.

Under this setting, we observe the deformation of each element by measuring the movement of each marker and then estimate the stress given to each node. The process of solving the inverse problem of FEM proceeds as follows:

1. Calculating the element rigidity matrix

Using the above settings, calculate the element rigidity matrix  $[EK]$  as follows:

$$[EK] = tS[B]^T[D][B]$$

where  $[B]^T$  denotes the transpose of  $[B]$  and

$$S = \frac{1}{2} \begin{bmatrix} 1 & x_i & y_i \\ 1 & x_j & y_j \\ 1 & x_k & y_k \end{bmatrix}, [D] = \frac{E}{(1+\nu)(1-\nu)} \begin{bmatrix} 1-\nu & \nu & 0 \\ \nu & 1-\nu & 0 \\ 0 & 0 & \frac{1-2\nu}{2} \end{bmatrix}$$

$$[B] = \frac{1}{2S} \begin{bmatrix} y_j - y_k & 0 & y_k - y_i & 0 & y_i - y_j & 0 \\ 0 & x_k - x_j & 0 & x_i - x_k & 0 & x_j - x_i \\ x_k - x_j & y_j - y_k & x_i - x_k & y_k - y_i & x_j - x_i & y_i - y_j \end{bmatrix},$$

$x_n$  and  $y_n$  ( $n=i, j, k$ ) are the  $x$ -coordinate and  $y$ -coordinate of the three nodes  $i, j$  and  $k$  which form an element.  $E, \nu$  and  $t$  denote Young's modulus, Poisson's ratio, and board thickness, respectively. We experimentally set the value of  $E$  and  $\nu$  to 0.14[MPa] and 0.45, respectively.

2. Constructing the whole rigidity matrix

Construct the whole rigidity matrix  $[TK]$  using the element rigidity matrix  $[EK]$  so that each element of  $[TK]$  corresponds to  $[EK]$  calculated for each element in step 1.

3. Estimating the stress for each node

Calculate the node force vectors  $\{F\}$  according to the following equation:

$$\{F\} = [TK]\{d\}$$

where  $\{d\}$  denotes the displacement of the nodes output by the motion capture system. The node force vectors  $\{F\}$  represent the stress given to each node. Thus, we utilize them as the higher-level features.

### 4.3 Feature selection method

Through the above feature construction process, the higher-level features are constructed for each node (i.e., marker). However, using all of the features is inefficient because the large amount of information which they contain often includes redundant data. Since such redundancy makes the computational complexity excessively large and does not contribute to the improvement of the recognition accuracy, it is necessary to select useful features. To perform efficient feature selection, because the constructed features are represented as multistream time-series data, we propose an effective method to evaluate the usefulness of streams in multistream time-series data.

We propose an effective criterion to assess the usefulness of each stream based on a novel similarity measure called Angular Metrics for Shape Similarity (AMSS) (Nakamura et al., 2007). To measure the similarity between time-series data, AMSS first divides a time-series sequence into several subsequences which are represented by a set of vectors. It then calculates the similarity based on the angles between two subsequences. Using angles for calculating similarity, AMSS can be robust to the difference in spatial locations of two time-series sequences compared with conventional similarity measures such as Dynamic Time Warping (Berndt & Clifford, 1996).

In order to evaluate the usefulness of a stream based on AMSS, we consider a  $C$ -class multistream time-series data classification problem. We assume that the input data consist of several examples. An example  $x$  contains a set of streams, which are described by vector sequences, and is represented as  $x = \{x^1, \dots, x^p\} = \{(\vec{x}_{11}, \dots, \vec{x}_{1q_1}), \dots, (\vec{x}_{p1}, \dots, \vec{x}_{pq_p})\}$ , where  $p$  is the number of streams,  $\vec{x}_{rs}$  is the  $s$ -th vector in the  $r$ -th stream, and  $q_k$  is the length of the  $k$ -th stream. Each example has the class label  $y$  ( $y \in \{1, \dots, C\}$ ). We represent a multistream time-series data set as  $\{(x_1, y_1), \dots, (x_m, y_m)\}$  where  $m$  is the number of examples.

To define a criterion to evaluate the usefulness of a stream, we assume that examples with the same class label have similar streams while examples with different class labels have dissimilar streams. The streams which satisfy these assumptions are useful to classifying examples accurately because if these assumptions are satisfied, the examples can be appropriately separated into each class. Thus, we measure the similarity between arbitrary combinations of two streams and, by determining whether the streams satisfy these assumptions, find the optimal streams.

According to the first assumption, we first define the similarity  $SS(n)$  among the examples which have the same class label for each stream ( $n = 1, \dots, p$ ) as follows:

$$SS(n) = \sum_{k=1}^{m-1} \sum_{l=k+1}^m A_s(x_k^n, x_l^n) w(x_k^n) w(x_l^n), \quad (12)$$

where

$$A_s(x_k^n, x_l^n) = \begin{cases} AMSS(x_k^n, x_l^n) & \text{if } y_k = y_l \\ 0 & \text{otherwise,} \end{cases} \quad (13)$$

$$w(x_j^n) = \sum_{i=1}^{q_n} \|\vec{x}_{jmi}\|, \quad (14)$$

$\bar{x}_{jni}$  corresponds to the  $i$ -th vector in the  $n$ -th stream  $\bar{x}_{ni}$  of the  $j$ -th example  $x_j$ .

$AMSS(x_k^n, x_l^n)$  denotes the similarity between  $x_k^n$  and  $x_l^n$  as calculated by AMSS. Due to space limitations, we do not show the details of AMSS here. The detailed algorithm of AMSS is described in (Nakamura et al., 2007). In equation (12), the coefficients  $w(x_k^n)$  and  $w(x_l^n)$  are used as weights. In order to utilize informative streams, we introduce these weights. We show an example in Fig. 8.

In Fig. 8,  $x_i^1$  is similar but not identical to  $x_j^1$ . On the other hand,  $x_i^2$  and  $x_j^2$  are identical. Thus, the similarity between  $x_i^2$  and  $x_j^2$  is higher than that between  $x_i^1$  and  $x_j^1$ . From the viewpoint of information theory, however, the entropy of  $x_i^1$  and  $x_j^1$  is much higher than those of  $x_i^2$  and  $x_j^2$ . Thus,  $x_i^1$  and  $x_j^1$  are more informative and represent the characteristics of streams. As a result, comparing  $x_i^1$  with  $x_j^1$  is more effective than comparing  $x_i^2$  and  $x_j^2$ . Since the weights in equation (12) reflect the information contained in each stream, introducing the weights enables the utilization of informative streams.

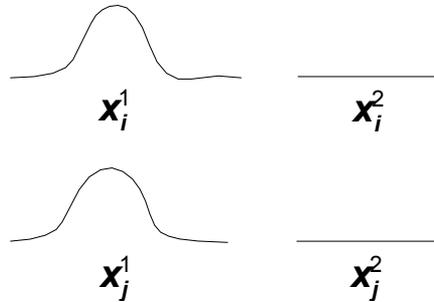


Fig. 8. Informative streams (left) and uninformative streams (right).

Next, with respect to the second assumption, we define the similarity  $DS(n)$  for each stream among the examples which have different class labels as follows:

$$DS(n) = \sum_{k=1}^{m-1} \sum_{l=k+1}^m A_d(x_k^n, x_l^n), \quad (15)$$

where

$$A_d(x_k^n, x_l^n) = \begin{cases} AMSS(x_k^n, x_l^n) & \text{if } y_k \neq y_l \\ 0 & \text{otherwise.} \end{cases} \quad (16)$$

In the calculation of  $DS(n)$ , we do not use weights. Without weighting, uninformative streams as shown in Fig. 8 make the value of  $DS(n)$  higher, and a higher  $DS(n)$  means that the stream is less useful. Consequently, the streams are regarded as useless streams. Based on  $SS(n)$  and  $DS(n)$ , we define the usefulness of the  $n$ -th stream  $U(n)$  by the following equation:

$$U(n) = \frac{SS(n)}{DS(n)}. \quad (17)$$

Using equation (17), we can estimate the usefulness of each stream. However, multistream time-series data generally contain a number of streams, with the number of possible combinations of the streams exponentially increasing as the number of streams increases. Thus, selecting useful streams is a difficult task. In order to determine the optimal number of streams, we propose an effective method based on the idea of class separability mentioned in Section 3.2. For the stream selection, the class separability is defined as follows:

$$s_+^k(c) = \frac{e_{c,c}^k}{\sum_{i=1}^C e_{c,i}^k}, \quad s_-^k(c) = \frac{\sum_{i \neq c} \sum_{j \neq c} e_{i,j}^k}{\sum_{i \neq c} \sum_{j=1}^C e_{i,j}^k}, \quad (18)$$

where  $e_{i,j}^k$  denotes the number of examples whose class labels are  $i$  and which are classified into the class  $j$  using the  $k$  most useful streams (i.e. streams that have the  $k$  highest values of usefulness). Based on the class separability, we determine the optimal number of streams  $k^*$  so that the following evaluation function is maximized:

$$k^* = \arg \max_k \left\{ \sum_{c=1}^C s_+^k(c) s_-^k(c) \right\}. \quad (19)$$

That is,  $k^*$  streams should be selected where  $k^*$  is the number of streams which maximizes the products of  $s_+^k(c)$  and  $s_-^k(c)$  for each class. When classifying an unseen example  $x$ , the predicted class label  $y$  is given by

$$y = \arg \max_c \left\{ \frac{\sum_{i=1}^m \sum_{n=1}^{k^*} f(x^n, x_i^n, c)}{m_c} \right\}, \quad (20)$$

where

$$f(x^n, x_i^n, c) = \begin{cases} AMSS(x^n, x_i^n) & \text{if } y_i = c \\ 0 & \text{otherwise,} \end{cases} \quad (21)$$

and  $m_c$  is the number of examples whose class labels are  $c$ .

#### 4.4 Experiment

For the evaluation of the proposed method, we perform two types of experiments. In the first experiment, we verify the usefulness of our feature construction and feature selection methods by comparing the recognition performance using the streams selected by the proposed method with the performance using the streams defined to be useful by FACS. We then apply our collaborative ensemble learning framework to a facial expression recognition problem and assess the effectiveness through the second experiment.

The motion data used in this experiment are obtained using the optical motion capture system HiRES (4 cameras) by Motion Analysis company. The motion data has a sampling frequency of 60Hz and a length of 5 seconds. Thus, a total of 300 frames are used per markers. We divide these frames into 30 groups, so that each group contains 10 frames, then generate time-series data for each stream whose length is 30 by averaging the frames in each group. Since we use the time-series data of horizontal and vertical movement of the markers, each stream consists of 2-dimensional time-series data. The motion data contains four types of expressions: *Surprise*, *Anger*, *Happiness*, and *Sadness*. Thus, the classification task is to distinguish these four expressions (i.e., a 4-class classification problem).

#### 4.4.1 Comparison with FACS

We carry out several facial expression recognition experiments using facial expression data from five subjects. For each subject, we obtain 24 examples (6 examples for each expression), for a total of 120 examples. We perform 5-fold cross-validation using 96 examples as training examples and 24 examples as test examples. In addition, we perform person-independent and person-dependent experiments. A person-independent experiment is an experiment in which the training set consists of examples from four subjects and the test set consists of examples from the remaining one subject. In a person-dependent experiment, the training set and test set include examples from the same subject (but not identical examples).

For the collaborative ensemble learning, we construct four base learners which are specialized to recognize *Surprise*, *Anger*, *Happiness* and *Sadness*. These base learners were generated using the following equation instead of equation (16).

$$A_s(x_k^n, x_l^n) = \begin{cases} AMSS(x_k^n, x_l^n) & \text{if } y_k = y_l = c \\ 0 & \text{otherwise,} \end{cases} \quad (22)$$

where  $c$  is the class label (that is, *Surprise*, *Anger*, *Happiness*, or *Sadness*). To classify an unseen example, we perform weighted voting using these four base learners and using their class separability as weights. For the comparison with FACS, we construct four base learners based on the stream defined by FACS and integrate them using weighted voting. We show the result of experiment in Table 5.

Method	Avg. # of streams	Surprise	Anger	Happiness	Sadness	Total
Person-independent						
Proposed	24.3	100	63.3	86.7	73.3	80.8
FACS	17.0	100	43.3	66.7	60.0	67.5
Person-dependent						
Proposed	27.4	100	83.3	90.0	76.7	87.5
FACS	17.0	100	60.0	73.3	80.0	78.3

Table 5. The recognition accuracy of the proposed method and FACS (in %).

Our method outperforms FACS in all expressions except for *Surprise* and total recognition accuracy. Since *Surprise* is most discriminative expression because of the intensive movement of facial muscles, both methods perfectly classified the example of *Surprise*. On the other hand, because distinguishing *Anger* from the other expressions is relatively difficult, the recognition accuracy for *Anger* is generally lowest. From this result, we verify the effectiveness of our feature construction and selection method.

The number of selected streams of our method is quite larger than that of FACS. For example, in the person-independent experiment, the average numbers of selected streams for *Surprise*, *Anger*, *Happiness* and *Sadness* are 26.4, 19.0, 27.4, and 24.2, respectively. Thus, the overall average number of selected streams is 24.3. We show an example of selected streams for the person-independent experiment in Fig. 9.

As for *Anger* and *Sadness*, most streams are regarded as useful streams while the number of streams defined by FACS is relatively small. This is because *Anger* and *Sadness* are more difficult to classify correctly than the other two expressions. In fact, the recognition accuracy for *Anger* and *Sadness* is relatively low. The number of selected streams for *Surprise* and *Happiness* is smaller than for the other expressions, but larger than those of FACS. This result implies that most AUs can contribute to the discrimination of facial expressions.

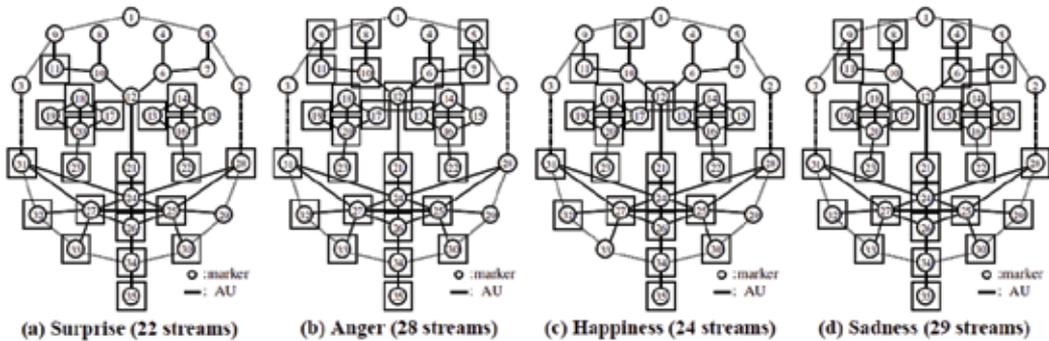


Fig. 9. Example of selected streams for each expression.

Although the recognition performance of our method is higher than that of FACS, the number of streams used is also higher than that of FACS. Since this may be unfair, we perform a recognition experiment using the same number of streams as FACS. We show the result in Table 6.

	Surprise	Anger	Happiness	Sadness	Total
Person-dependent	100	56.7	90.0	80.0	81.7
Person-independent	100	83.3	83.3	66.7	83.3

Table 6. The recognition accuracy of the proposed method using the same number of streams as FACS (in %).

Compared with the recognition accuracy of FACS in Table 5, our method fully achieves high recognition performance. Thus, we confirm the advantage of employing the streams selected by our method over the streams defined by FACS. In the person-dependent experiment, the recognition accuracy is higher than the accuracy shown in Table 5. Generally, person-independent problems are more difficult than person-dependent problems because the distributions of training set and test set can be considerably different. Therefore, we should introduce a method of more accurately estimating the distribution of the test set. However, there is no significant difference between the recognition accuracy of the person-independent cases in Table 5 and Table 6 under the *t*-test with a 5% significance level. This result indicates that our method is able to find the sub-optimal combination of streams which is comparable to the optimal combination with respect to recognition performance. Therefore, our feature selection method seems to be fully effective.

#### 4.4.2 Verification of the effectiveness of collaborative ensemble learning

We introduce our collaborative ensemble learning framework, as shown in Section 3. We use the aforementioned four base learners and train them according to the collaborative ensemble learning algorithm shown in Section 3.2. The prediction of the resulting ensemble classifier  $f(x)$  for a given example  $x$  is determined based on equation (5) as follows:

$$f(x) = Y^{l^*} \text{ such that } l^* = \arg \max_{l \in \{1,2,3,4\}} \sum_{\tau=1}^l s_{\tau}^l(Y_{\tau}^l(x)) , \quad (23)$$

where  $Y_{\tau}^1, Y_{\tau}^2, Y_{\tau}^3$  and  $Y_{\tau}^4$  correspond to the hypotheses generated at the  $\tau$ -th round by the base learners for *Surprise*, *Anger*, *Happiness* and *Sadness*, respectively.  $s_{\tau}^l$  ( $l = 1,2,3,4$ ) represents the class separability for each base learner and is calculated by equation (4). We perform a person-independent experiment with 100 rounds of learning. The result of our experiment is shown in Table 7.

Collaborative ensemble learning						Weighted voting
1 round	5 rounds	10 rounds	20 rounds	50 rounds	100 rounds	
81.5%	81.5%	82.5%	83.0%	85.0%	85.2%	80.8%

Table 7. Recognition accuracy by collaborative ensemble learning and weighted voting.

The recognition accuracy is proportional to the number of rounds. Although the recognition accuracy of the collaborative ensemble learning method is slightly higher than that of the weighted voting method in early rounds, the difference is significant after approximately the 50th round. This result shows that the interaction of multiple base learners is effective in dealing not only with image data but also with multistream time-series data. From this result, we confirm the flexibility of our learning model.

These experimental results show that the streams selected by the proposed method lead to better recognition results than the streams defined by FACS. The results reflect the effectiveness of our method as a data-mining framework in facial expression recognition. However, there is room for improvement in the determination of the optimal combinations of streams in the person-independent case. Thus, we should try to estimate the distribution of each stream more accurately and improve the performance of our stream selection method. In addition, we verify the applicability of our collaborative ensemble learning framework to facial expression recognition problems. In the experiment, the difference between the recognition accuracy at the 50th round and that at the 100th round is insignificant although the computational complexity differs considerably. However, the number of rounds is determined experimentally. In order to improve the learning efficiency, a method of automatically determining the optimal number of rounds is needed.

## 5. Visual learning revisited

The performance of a visual learning model is closely related to (1) the learning model and (2) features. We refer to several visual learning methods shown in Table 8 from these two viewpoints.

At an early stage of the study of visual learning, a successful object detection method was proposed based on AdaBoost (Viola & Jones, 2001). In this method, a cascade classifier,

which is a kind of ensemble classifier, is constructed. An example of a cascade classifier for facial recognition is shown in Fig. 10.

Reference	Learning model	Features
Viola and Jones, 2001	Cascade classifier + AdaBoost	Primitive feature + feature selection
Marée et al., 2005	Decision tree + ensemble learning	Primitive feature
Krawiec and Bhanu, 2003	Evolutionary computation + closed-loop learning	Feature construction + feature selection
Proposed method	Collaborative learning + modular approach	Feature construction + feature selection

Table 8. Visual learning models.

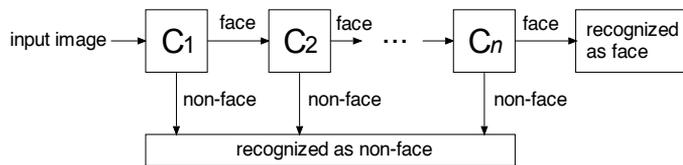


Fig. 10. An example of cascade classifier.

In Fig. 10,  $C_i$  ( $i = 1, 2, \dots, n$ , where  $n$  is the number of base classifiers) represents the base classifier. For a facial recognition task, if all the base classifiers classify the given image as a face image, then the image is recognized as a face image. Otherwise, the given image is recognized as a non-face image. This recognition process is efficient because, when a given image is classified by a certain classifier as a non-face image, subsequent classifiers are not used. In addition, the cascade classifier is able to select a small number of useful features from a large number of extracted features and can thus quickly and accurately detect objects. However, the features used in this method are primitive features because they are obtained only from pixel intensity values. Moreover, all the base learners use the same features, called Harr-like features. Since recognition performance is greatly dependent on the representational ability and diversity of the features (i.e., various types of high-level features are required), this method seems to be insufficient to describe and recognize complex objects. In addition, the learning model of this method is rather simple because it only optimizes the parameters used in AdaBoost rather than constructing or selecting features. Therefore, this method can be regarded as the most primitive visual learning model.

To utilize multiple features effectively, the ensemble approach has already been introduced into visual learning. For example, in (Marée et al., 2005), an ensemble classifier is constructed using decision tree classifiers as base learners. The learning model of this method is shown in Fig. 11. In Fig. 11,  $T_i$  ( $i = 1, 2, \dots, n$ , where  $n$  is the number of base learners) represents the base learner (i.e., decision tree classifier). The learning result of each base learner is integrated into an ensemble classifier using weighted voting by all base learners. However, in this ensemble approach, the base learners are separately constructed

with only their learning results integrated. Thus, there is no interaction among the base learners. In addition, all visual learners are based on the same primitive features directly obtained from the color information of each pixel. Since the performance of the ensemble approach tends to be proportional to the diversity of features, higher-level features are needed from the viewpoints of the flexibility and accuracy of recognition.

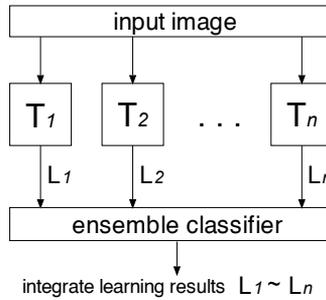


Fig. 11. Ensemble visual learning model.

As a principal learning structure of human visual systems, Krawiec et al. introduced a closed-loop learning scheme based on evolutionary computation (Krawiec & Bhanu, 2003). In this method, high-level features are constructed from the given primitive features (the intensity values of pixels) by combining several image processing operations, such as image filtering and the application of mathematical or logical computation for some pixels. In order to construct appropriate high-level features, the optimal combination of image processing operations is sought through the learning loop. In the learning loop, the combination of image processing operations is determined and its effectiveness is evaluated using evolutionary computation. The learning framework is shown in Fig. 12.

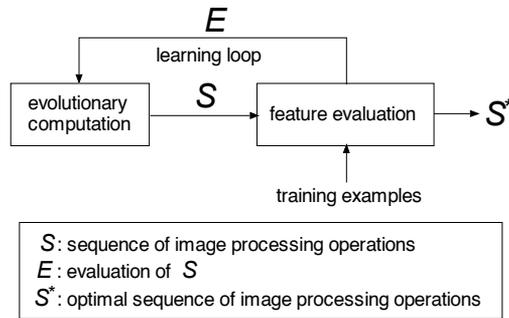


Fig. 12. Evolutionary (closed-loop) visual learning model.

The process of constructing and evaluating features is iteratively performed during the learning process. The evaluation of the constructed features is fed back to the evolutionary computation algorithm and better features then searched. Finally, the best feature  $S^*$  is output at the end of learning process. This feature construction strategy represents a sophisticated learning framework that is consistent with human visual learning process. However, the effectiveness of the feature construction method is dependent on the

predefined image processing operations, and the determination of appropriate image processing operations is an open problem. The proposed visual learning approach has the properties of modularity and closed-loop learning, both essential properties in human visual systems; thus, they make the proposed method more efficient than conventional visual learning methods. However, our method still has the following two main problems.

The first problem is related to features. In the facial expression recognition, we propose a feature construction method based on stress estimation. Additionally, we propose a feature selection method based on the evaluation of the usefulness of each stream. We verify the effectiveness of our method through the comparison of its recognition performance with that of FACS. However, the experimental result shows that our feature construction and selection methods cannot always find the optimal combination of streams. This implies that our method is rather simple because we construct higher-level features for each stream separately. A facial expression is represented by the complex movements of several points on a face. This means that multiple streams are mutually correlated. Therefore, we should improve the feature construction process so that the higher-level features are constructed by integrating multiple streams which are mutually correlated. More generally, we should further analyze human visual systems and attempt to model them in order to develop satisfactory feature construction frameworks for various visual recognition problems. The second problem with our method resides in the representation of knowledge obtained through the learning process. Our method can provide the knowledge for the recognition of visual data as the useful features. This knowledge can be used for data mining, but in order to utilize the learning results of our method fully in some data mining domains, the knowledge should be systematized by analyzing and organizing it in the learning process.

## 5. References

- Belongie, S.; Malik, J. & Puzicha, J. (2001). Matching Shapes, *Proceedings of the 8th IEEE International Conference on Computer Vision*, pp. 454-463
- Berndt, D. J & Clifford, J. (1996). Finding Patterns in Time Series: A Dynamic Programming Approach, *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press, Menlo Park, CA, pp. 229-248
- Bourel, B.; Chibelushi, C. C. & Low, A. A. (2002). Robust Facial Expression Recognition Using a State-Based Model of Spatially-Localised Facial Dynamics, *Proceedings of the 5th International Conference on Automatic Face and Gesture Recognition*, pp. 106-111
- Das, G.; Gunopulos, D & Mannila, H. (1997). Finding Similar Time Series, *Proceedings of the 1st European Symposium on Principles of Data Mining and Knowledge Discovery*, pp. 88-100
- Dietterich, T. G. (2000). Ensemble Methods in Machine Learning, *Proceedings of the 1st International Workshop on Multiple Classifier Systems*, pp. 1-15
- Donato, G.; Bartlett, M. S.; Hager, J.C.; Ekman, P. & Sejnowski, T. J. (1999). Classifying Facial Actions, *Pattern Analysis and Machine Intelligence*, Vol. 21, No. 10, pp. 974-989
- Ekman, P & Friesen, W. (1978). Facial Action Coding System: A Technique for the Measurement of Facial Movement, *Consulting Psychologists Press*, Palo Alto, CA

- Essa I. A. & Pentland, A. P. (1997). Coding, Analysis, Interpretation, and Recognition of Facial Expressions, *Pattern Analysis and Machine Intelligence*, Vol. 19, No. 7, pp. 757-763
- Freund, Y. & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting, *Journal of Computer and System Sciences*, Vol. 55, No.1, pp. 119-139
- Grauman, K. & Darrell, T. (2005). Efficient Image Matching with Distributions of Local Invariant Features, *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol.2, pp. 627-634
- Leibe, B. & Schiele, B. (2003). Analyzing Appearance and Contour Based Methods for Object Categorization, *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pp. 409-415
- Lowe, D. (2004). Distinctive Image Features from Scale-Invariant Keypoints, *International Journal of Computer Vision*, Vol. 60, No. 2, pp. 91-110
- Krawiec, K. & Bhanu, B. (2003). Visual Learning by Evolutionary Feature Synthesis, *Proceedings of the 20th International Conference on Machine Learning*, pp. 376-383
- Marée, R.; Geurts, P.; Piater, J. & Wehenkel L. (2005). Random Subwindows for Robust Image Classification, *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol.1, pp. 34-40
- Marr, D. (1982). *Vision*, W. H. Freeman and Company
- Nakamura, M.; Nomiya, H & Uehara, K. (2004). Improvement of Boosting Algorithm by Modifying the Weighting Rule, *Annals of Mathematics and Artificial Intelligence*, Vol.41, pp. 95-109
- Nakamura, T.; Taki, K. & Uehara, K. (2007). Time Series Classification by Angular Metrics for Shape Similarity, *Proceedings of Workshop and Challenge on Time Series Classification (CTSC '07/KDD 2007)*, pp. 42-49
- Nomiya, H. & Uehara, K. (2005). Feature Construction and Feature Integration in Visual Learning, *Proceedings of ECML2005 Workshop on Sub-symbolic Paradigm for Learning in Structured Domains*, pp. 86-95
- Nomiya, H. & Uehara, K. (2007). Multistrategical Image Classification for Image Data Mining, *Proceedings of International Workshop on Multimedia Data Mining*, pp.22-30
- Osaki, R.; Shimada, M. & Uehara, K. (2000). A Motion Recognition Method by Using Primitive Motions, *Proceedings of the 5th IFIP 2.6 Working Conference on Visual Database Systems*, pp. 117-128
- Schiele, B. & Crowley, J. L. (2000). Recognition without Correspondence using Multidimensional Receptive Field Histograms, *International Journal of Computer Vision*, Vol. 36, No. 1, pp. 31-50
- Swain, M. J. & Ballard, D. H. (1991). Color indexing, *International Journal of Computer Vision*, Vol.7, No.1, pp. 11-32
- Tu, Z. (2005). Probabilistic Boosting-Tree: Learning Discriminative Models for Classification, Recognition, and Clustering, *Proceedings of the 10th IEEE International Conference on Computer Vision*, Vol.2, pp. 1589-1596
- Turk, M. A. & Pentland, A. P. (1991). Face Recognition Using Eigenfaces, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp.586-591

Viola, P. & Jones M. (2001). Rapid Object Detection using a Boosted Cascade of Simple Features, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp.511-518

# Clustering Parallel Data Streams

Yixin Chen

*Department of Computer Science,  
Washington University  
St. Louis, Missouri*

## 1. Introduction

Massive volumes of data streams can be found in numerous applications such as network intrusion detection, financial transaction flows, telephone call records, sensor streams, and meteorological data. In recent years, there are increasing demands for mining data streams. Unlike the finite, statically stored data sets, stream data are massive, continuous, temporally ordered, dynamically changing, and potentially infinite [5]. For example, Cortes et al. report that AT&T long distance call records consist of 300 million records per day for 100 million customers. For the stream data applications, the volume of data is usually too huge to be stored or to be scanned for more than once. Further, in data streams, the data points can only be sequentially accessed. Random access to data is not allowed.

Extensive research has been done for mining data streams, including those on the stream data classification [3, 20], mining frequent patterns [9, 17, 18], and clustering stream data [1, 2, 8, 9, 10, 11, 12, 13, 14, 16, 19].

In this paper, we study the clustering of multiple and parallel data streams. Our study should be differentiated from some previous studies on clustering stream data [19, 1]. Our goal is to group multiple streams with similar behavior and trend together, instead of to cluster the data records within one data stream.

There are various applications where it is desirable to cluster the streams themselves rather than the individual data records within them. For example, the price of a stock may rise and fall from time to time. To reduce the financial risk, an investor may prefer to spread his investment over a number of stocks which may exhibit different behaviors. As another application, in meteorological study and disaster prediction, it is useful to cluster meteorological data streams from different geographical regions of similar curvature trends in order to identify regions with similar meteorological behaviors. Yet another example is that a super market may record sales on different merchandizes. There may be some relationship among the sales of different merchandizes and thus the merchant can make use of the correlation to manipulate the prices to maximize the profit.

Clustering refers to partition a data set into clusters such that members within the same cluster are similar in a certain sense and members of different clusters are dissimilar. Current clustering techniques can be broadly classified into several categories: partitioning methods (e.g.,  $k$ -means and  $k$ -medoids), hierarchical methods (e.g. BIRCH [22]), density-based methods (e.g. DBSCAN [15]), and grid-based methods (e.g. CLIQUE [4]). However, these methods are designed only for static data sets and can not be directly applied to data streams.

An abundant body of researches on clustering data in one data stream has emerged. O'Callaghan et al. [19] provided an algorithm called STREAM based on  $k$ -means, which adopts a divide-and-conquer technique to deal with buckets to get clusters. CluStream [1] is an algorithm for the clustering of evolving data streams based on user-specified, on-line clustering queries. It divides the clustering process into on-line and off-line components. The online component computes and stores summary statistics about the data stream using micro-clustering, while the offline component does macro-clustering and answers various user questions using the stored summary statistics. In these works, the problem refers to cluster the elements of one individual data stream. These works motivate the online compression and offline computation framework used in the paper but are obviously not applicable to our problem.

**Euclidean Distance vs. Correlation Distance** The problem of clustering multiple data streams views each data stream as an element to be clustered, which pays attention to the similarity between streams. There are several previous works on this problem. Yang [21] uses the weighted aggregation of snapshot deviations as the distance measure between two streams, which can observe the similarity of data values but ignore the trends of streams.

Beringer et al. [6] proposed a preprocessing step which uses a discrete Fourier transforms (DFT) approximation of the original data, uses the few low-frequency (instead of all) coefficients to compute the distance between two streams, and applies an online version of the  $k$ -means algorithm. Such a method acts like a low-pass filter to smooth the data stream. The DFT transformation preserves the Euclidean distances. Thus, the DFT distance is equivalent to the Euclidean distance of the smoothed data streams.

A serious limitation of the above previous works is that they are based on the Euclidean distance of data records, but important trend information contained in data streams is typically discarded by clustering methods based on Euclidean distance. This is true because data streams with similar trends may not be close in their Euclidean distance. For example, in stock markets, the absolute prices of stocks from similar area or similar industry can be very different but their trends are close. To capture such similarity, it is more appropriate to use **correlation analysis**, a statistical methods on time series, which measures the resemblance of the trends of multiple data streams.

As an illustration, Figure 1 shows the trends of three stocks on Nylon, chemical fiber, and CPU chip, respectively. Although the two stocks on CPU chip and chemical fiber are closer in their data values and thus their Euclidean distance, the trends of the two stocks on Nylon and Chemical fiber are clearly more close, which can be suggested by their higher correlation coefficient. If using Euclidean distance, we may conclude that the two stocks on CPU chips and chemical fiber are more similar, which does not properly reflect the more interesting trend similarity between nylon and chemical fiber.



Fig. 1. Price of stocks on nylon, chemical fiber, and CPU chips.

In this paper, we propose an algorithm for clustering multiple data streams based on correlation analysis. Performing correlation analysis on data streams under the one-scan requirement poses significant technical challenges since we cannot not store the raw data. In this paper, we develop a novel scheme that compresses the data online and stores only the compressed measures, called the synopsis, in the offline system. We propose a new theory that facilitate efficient computation of correlation coefficients based on the compressed measures. The correlation coefficients are used to define distances between streams which are in turn used by a  $k$ -means algorithm to generate the clustering results. An attenuate coefficient is also introduced to allow the algorithm to discover the evolving behaviors of data streams and adjust the clusters dynamically.

**Clustering on Demand (COD)** More recently, a clustering on demand (COD) framework [7] is proposed to give approximative answers to user queries for clustering sub-streams within certain time window. The framework consists of two components, including the online maintenance phase and the offline clustering phase. The online maintenance phase provides an efficient mechanism to maintain summary hierarchies of data streams with multiple resolutions. The offline phase uses an adaptive clustering algorithm to retrieve approximations of desired sub-streams from summary hierarchies according to clustering queries.

In this paper, we also extend our clustering algorithm to support real-time COD. We propose an innovative scheme to partition the time horizon into segments and store statistical information for each time segment. We prove that the scheme enables us to accurately approximate the correlation coefficients for an arbitrary time duration, and thus allows the users to obtain clusters for data streams within the requested time window.

The paper is organized as follows. After introducing the basic concepts and problem definitions in Section 2, we propose our algorithm in Section 3. We then discuss the extension to COD in Section 4. In Section 5, we show experiment results on synthetic data sets and real data sets which demonstrate the high accuracy, efficiency, and scalability of our algorithm compared with others. We conclude the paper in Section 6.

## 2. Background

In this section, we introduce the background of the work and several basic concepts.

### 2.1 Clustering data streams

A data stream  $X$  is a sequence of data items  $x_1, \dots, x_k$  arriving at discrete time steps  $t_1, \dots, t_k$ . We assume that there are  $n$  data streams  $\{X_1, \dots, X_n\}$  at each time step  $m$ , where  $X_i = \{x_{i1}, \dots, x_{ik}\}$ ,  $1 \leq i \leq n$ , and  $x_{ij}$  ( $j = 1, \dots, m$ ) is the value of stream  $X_i$  at time  $j$ .

The problem of clustering multiple data streams is defined as follows. Given the time horizon for clustering  $L$  and the number of clusters  $k$ , the clustering algorithm partitions  $n$  data streams into  $k$  clusters  $C(L) = \{C_1(L), \dots, C_k(L)\}$  that minimizes some objective function measuring the quality of clustering in the period  $[t - L + 1, t]$ , where  $t$  is the time when the analysis is performed. The given clusters  $C_j(L)$ ,  $j = 1, \dots, k$ , should satisfy:

$$\bigcap_{j=1}^k C_j(L) = \emptyset \text{ and } \bigcup_{j=1}^k C_j(L) = \{X_1(L), \dots, X_n(L)\},$$

where  $X_i(L) = \{x_{i(t-L+1)}, \dots, x_{it}\}$ ,  $i = 1, \dots, n$ .

## 2.2 Attenuation coefficient

In stream data analysis, in order to recognize the evolving characteristics of data streams, newer data records are often given more weights than older ones. Therefore, we use an *attenuation coefficient*  $\lambda \in [0, 1]$  to gradually lessen the significance of each data record over time. Suppose  $t$  is the current time and a data point  $x_i$  is received at time  $i$ , then, in our analysis, we replace the original value of  $x_i$  with

$$x_i(t) = \lambda^{t-i} x_i. \quad (1)$$

Applying this adjustment to every element in the data streams in the time horizon  $[t - L + 1, t]$ , we replace the original values by

$$\begin{aligned} & \{x_{t-L+1}(t), \dots, x_t(t)\} \\ = & \{\lambda^{L-1} x_{t-L+1}, \lambda^{L-2} x_{t-L+2}, \dots, \lambda x_{t-1}, x_t\}. \end{aligned} \quad (2)$$

## 2.3 Time segment

We first consider fixed-length clustering and then extend the algorithm for arbitrary-length clustering in COD in Section 4. Given a fixed length, at any time  $t$ , we report in real time the clustering results for the data streams in the time horizon  $[t - L + 1, t]$ . To support efficient processing, we partition the data streams of length  $L$  into  $m$  *time segments* of equal length  $l = L/m$ . Whenever a new segment of length  $l$  accumulates (Figure 2), we re-compute the clustering results.

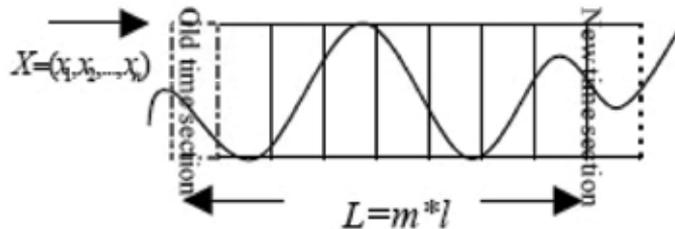


Fig. 2. A fixed length  $L$  is divided into  $m$  segments of size  $l$ .

## 3. The CORREL-cluster algorithm

In the following, we propose a general framework called **CORREL-cluster**, a correlation-based clustering algorithm for data streams. Unlike the widely-used Euclidean distance which only measures the discrepancy of the data values, our correlation-based distance considers two data streams with similar trends to be close to each other.

We first overview the overall framework of the proposed CORREL-cluster algorithm. The framework of the algorithm is in Figure 3.

CORREL-cluster continuously receives new data records from all the data streams at each time step (Line 5 in Figure 3). It keeps outputting the clusters for the most recent data streams of fixed length  $L$ . For every  $l$  time steps, it first computes a compressed representation of the data streams for the time segment  $[t - l + 1, t]$  (Line 7), discards the raw data, and updates the compressed representation of the data streams for the target

clustering time  $[t-L+1, t]$  (Lines 9-10). It then calls a correlation-based  $k$ -means algorithm (Line 11) to compute the clustering results. Since the number of clusters  $k$  may be changing, CORRELcluster also employs a new algorithm to dynamically adjust  $k$  in order to recognize the evolving behaviors of the data streams (Line 12). The algorithm is schematically shown in Figure 4.

```

1. procedure CORREL-cluster
2.  $t = 0$ ;
3. while data stream is not terminated
4.   set  $t = t + 1$ ;
5.   read new data record  $x_{kt}(t)$ ,  $k = 1 \dots n$ , one from each of
     the  $n$  data streams;
6.   if ( $t \bmod l == 0$ ) then /* form a new segment. */
7.     calculate the CCR of the time segment  $[t-l+1, t]$ ;
8.     update other  $m-1$  CCRs, where  $m = L/l$ ;
9.     if  $t == L$  then compute initial  $CCR_L$ , the compressed
     correlation representation for  $[t-L+1, t]$ ;
     else incrementally update  $CCR_L$ ;
10.    call correlation_ks_means(); /* compute clusters */
11.    call adjust_k();
12.    output the clustering result;
13.  end if
14. end while
15. end_procedure

```

Fig. 3. The overall process of CORREL-cluster.

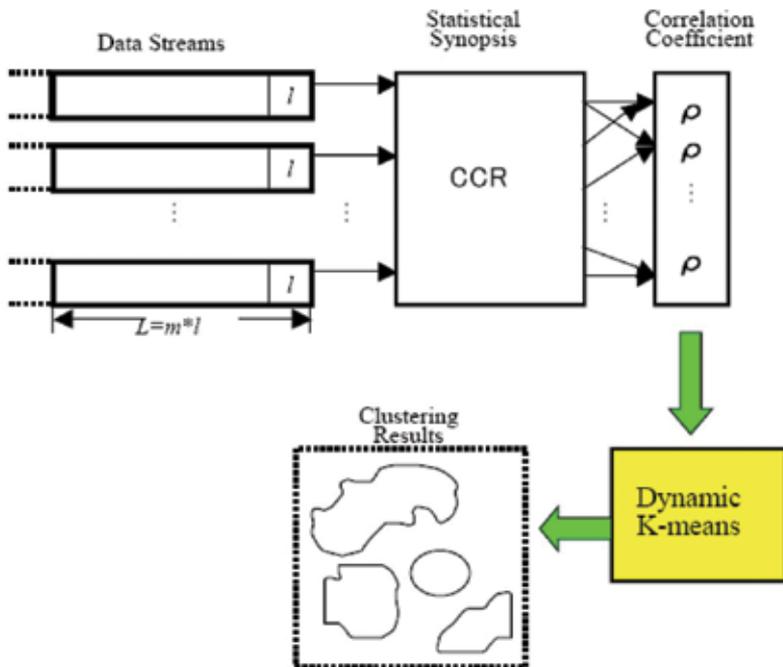


Fig. 4. Illustration of CORREL-cluster.

### 3.1 Correlation analysis of data streams

Before describing the details of our algorithm, we first overview the concepts of correlation analysis.

We first define the correlation coefficients for two data streams. For two data streams  $X = (x_1, \dots, x_n)$  and  $Y = (y_1, \dots, y_n)$ , the correlation coefficient between them is defined as

$$\rho_{XY} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{j=1}^n (y_j - \bar{y})^2}}, \quad (3)$$

where  $\bar{x} = (\sum_{i=1}^n x_i)/n$  and  $\bar{y} = (\sum_{i=1}^n y_i)/n$ .

From the definition, we can see that  $|\rho_{XY}| \leq 1$ . A large value of  $|\rho_{XY}|$  indicates strong correlation between streams  $X$  and  $Y$ , and  $\rho_{XY} = 0$  means  $X$  and  $Y$  are uncorrelated.

Since it is often impossible to store all the past raw data in the stream, we need to compress the raw data and only retain a synopsis for each time segment of each data stream.

The following theorem shows that, to compute the correlation coefficient we only need to save  $\sum_i x_i$ ,  $\sum_i x_i^2$  for each stream  $X$  and  $\sum_i x_i y_i$  between any two streams to compute the correlation coefficients between any two streams.

**Theorem 3.1** *Correlation coefficient  $\rho_{XY}$  between two sequences  $X$  and  $Y$  can be calculated using the information  $\sum_i x_i$ ,  $\sum_i x_i^2$ ,  $\sum_i y_i$ ,  $\sum_i y_i^2$ , and  $\sum_i x_i y_i$ .*

**Proof.** In (3), the numerator can be rewritten as:

$$\begin{aligned} & \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \\ = & \sum_{i=1}^n (x_i y_i - \bar{x} y_i - x_i \bar{y} + \bar{x} \bar{y}) \\ = & \sum_{i=1}^n x_i y_i - \sum_{i=1}^n \bar{x} y_i - \sum_{i=1}^n x_i \bar{y} + \sum_{i=1}^n \bar{x} \bar{y} \\ = & \sum_{i=1}^n x_i y_i - n \bar{x} \bar{y} - \bar{x} \bar{y} + \bar{x} \bar{y} \\ = & \sum_{i=1}^n x_i y_i - \frac{1}{n} \sum_{i=1}^n x_i \sum_{i=1}^n y_i. \end{aligned} \quad (4)$$

And in the denominator, we have

$$\begin{aligned} \sum_{i=1}^n (x_i - \bar{x})^2 &= \sum_{i=1}^n (x_i^2 - 2\bar{x}x_i + \bar{x}^2) \\ &= \sum_{i=1}^n x_i^2 - n\bar{x}^2 \\ &= \sum_{i=1}^n x_i^2 - \frac{1}{n} \left( \sum_{i=1}^n x_i \right)^2. \end{aligned} \quad (5)$$

Similarly, we have

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - \frac{1}{n} \left( \sum_{i=1}^n y_i \right)^2. \quad (6)$$

Therefore, we can compute both the numerator and denominator in (3) using  $\sum_i x_i$ ,  $\sum_i x_i^2$ ,  $\sum_i y_i$ ,  $\sum_i y_i^2$ , and  $\sum_i x_i y_i$ .

Based on the above result, in CORREL-cluster, for any time segment, we store a compressed synopsis for the parallel data streams instead of the raw data.

**Definition 3.1 (Compressed Correlation Representation (CCR))** Given  $n$  data streams  $X_1, \dots, X_n$ , suppose the current time is  $t$ , the time segment length is  $l$ , then we store the following quantities in  $CCR = (\vec{S}, \vec{Q}, C, t)$ , where the components of the vectors  $\vec{S}, \vec{Q}$  and matrix  $C$  are defined as:

$$\begin{aligned} S_i &= \sum_{k=t-l+1}^t x_{ik}(t), \quad i = 1, \dots, n \\ Q_i &= \sum_{k=t-l+1}^t x_{ik}^2(t), \quad i = 1, \dots, n \\ C_{ij} &= \sum_{k=t-l+1}^t x_{ik}(t)x_{jk}(t), \quad i, j = 1, \dots, n, i < j \end{aligned}$$

Based on Theorem 3.1, CCR provides enough information to compute the correlation coefficients between any two streams in  $n$  data streams  $X_1, \dots, X_n$ .

**Theorem 3.2** For two streams  $X_i$  and  $X_j$ ,  $i, j = 1..n$ , their correlation coefficients can be computed as

$$\rho_{X_i X_j} = \frac{C_{ij} - \frac{1}{n} S_i S_j}{\sqrt{(Q_i - \frac{1}{n} S_i^2)(Q_j - \frac{1}{n} S_j^2)}}$$

The theorem can be seen from equations (3), (4), (5), and (6).

### 3.2 Update of the CCR synopsis

For a given time segment, at the current time  $t_c$ , the CCR for the  $n$  streams is

$$\left\{ \sum x_{ik}(t_c), \sum x_{ik}^2(t_c), \sum x_{ik}(t_c)x_{jk}(t_c) \right\}.$$

Then at any time  $t > t_c$ , the data values are updated to  $\{x_{ik}(t)\}$  instead of  $\{x_{ik}(t_c)\}$ . To perform online clustering analysis, we also need to update the CCR.

Let  $t - t_c = \Delta_t$  and  $\lambda$  be the attenuation coefficient, then

$$x_{ik}(t) = \lambda^{\Delta_t} x_{ik}(t_c).$$

Therefore, we can update the saved information as follows.

$$\begin{aligned}
\vec{S}'_i &= \sum_{k=t_c-l_c+1}^{t_c} x_{ik}(t) = \sum_{k=t_c-l_c+1}^{t_c} \lambda^{\Delta t} x_{ik}(t_c) \\
&= \lambda^{\Delta t} \sum_{k=t_c-l_c+1}^{t_c} x_{ik}(t_c) = \lambda^{\Delta t} \vec{S}_i
\end{aligned} \tag{7}$$

$$\begin{aligned}
\vec{Q}'_i &= \sum_{k=t_c-l_c+1}^{t_c} x_{ik}^2(t) = \sum_{k=t_c-l_c+1}^{t_c} [\lambda^{\Delta t} x_{ik}(t_c)]^2 \\
&= \lambda^{2\Delta t} \sum_{k=t_c-l_c+1}^{t_c} x_{ik}^2(t_c) = \lambda^{2\Delta t} \vec{Q}_i
\end{aligned} \tag{8}$$

$$\begin{aligned}
\mathbf{C}'_{ij} &= \sum_{k=t_c-l_c+1}^{t_c} x_{ik}(t)x_{jk}(t) = \sum_{k=t_c-l_c+1}^{t_c} \lambda^{2\Delta t} x_{ik}(t_c)x_{jk}(t_c) \\
&= \lambda^{2\Delta t} \sum_{k=t_c-l_c+1}^{t_c} x_{ik}(t_c)x_{jk}(t_c) = \lambda^{2\Delta t} \mathbf{C}_{ij}
\end{aligned} \tag{9}$$

We note that such an update happens not at every time step, but every  $l$  steps (Line 8 in Figure 3). Whenever a new time segment comes in, we compute the new CCR. Also, there are  $m = L/l$  existing segments and thus  $m$  CCRs. We discard the oldest CCR, and update the other  $m-1$  segments according to the above formulae.

### 3.3 Aggregating CCR to CCR<sub>L</sub>

In CORREL-cluster, for a user specified clustering length  $L$ , we need to cluster streams within the time  $[t - L + 1, t]$ . Therefore, for each pair of streams  $(X, Y)$ , we need to compute the correlation coefficients for  $X[t - L + 1, t]$  and  $Y[t - L + 1, t]$ . Since we compute for each time segment with length  $l$  a CCR, we need to combine them into  $CCR_L$ , the CCR for the time window  $[t - L + 1, t]$ .

To formalize the problem, we have  $m$  time segments and  $m$  CCRs. Let  $CCR(v)$  be the CCR for time segment  $[t - vl + 1, t - (v - 1)l]$ , for  $v = 1..m$ , we denote the components of  $CCR(v)$  as

$$CCR(v) = (\vec{S}(v), \vec{Q}(v), \mathbf{C}(v)), \tag{10}$$

and  $CCR_L$  as

$$CCR_L = (\vec{S}_L, \vec{Q}_L, \mathbf{C}_L), \tag{11}$$

We have, at time  $t$ :

$$\begin{aligned}
\vec{S}_{L_i} &= \sum_{k=t-L+1}^t x_{ik}(t) = \sum_{v=1}^m \left( \sum_{k=t-vl+1}^{t-(v-1)l} x_{ik}(t) \right) \\
&= \sum_{v=1}^m \vec{S}_i(v), \quad \forall i = 1, \dots, n,
\end{aligned} \tag{12}$$

which can be compactly written as

$$\vec{S}_L = \sum_{v=1}^m \vec{S}(v). \quad (13)$$

Similarly, we have:

$$\vec{Q}_L = \sum_{v=1}^m \vec{Q}(v), \quad \text{and} \quad \mathbf{C}_L = \sum_{v=1}^m \mathbf{C}(v). \quad (14)$$

We use the above equations to compute  $CCR_L$  when we receive the first  $m$  time segments (Line 9 of Figure 3).

For later updates (Line 10), we do not need to redo the summation. In fact, we can incrementally update  $CCR_L$ . Given the existing  $m$  CCRs:  $CCR(i)$ ,  $i = 1, \dots, m$ , and the newly generated  $CCR(new) = (\vec{S}(new), \vec{Q}(new), \mathbf{C}(new))$ . We update the  $CCR_L$  as:

$$\vec{S}'_L = \vec{S}_L + \vec{S}(new) - \vec{S}(1) \quad (15)$$

$$\vec{Q}'_L = \vec{Q}_L + \vec{Q}(new) - \vec{Q}(1) \quad (16)$$

$$\mathbf{C}'_L = \mathbf{C}_L + \mathbf{C}(new) - \mathbf{C}(1). \quad (17)$$

We also update the saved CCRs by

$$CCR(i) = CCR(i+1), \quad i = 1, \dots, m-1 \quad (18)$$

$$CCR(m) = CCR(new). \quad (19)$$

### 3.4 Dynamic $k$ -means algorithm

We use a  $k$ -means clustering algorithm to generate the cluster for data streams in the user-specified window  $[t-L+1, t]$ . In the  $k$ -means algorithm, the distance  $d(X, Y)$  between two data streams  $X$  and  $Y$  is measure using the reciprocal of the correlation coefficient:

$$d(X, Y) = 1/\rho_{XY}. \quad (20)$$

The clustering quality is measured by an objective function

$$G = \sum_{i=1}^k \sum_{j=1}^n (1/\rho_{X_j C_i}), \quad (21)$$

where  $\rho_{X_j C_i}$  is the correlation coefficient between data stream  $X_j$  and cluster center  $C_i$ , which is a stream from  $X_1$  to  $X_n$ .

Let  $n$  be the number of streams to be clustered. The correlation-based  $k$ -means algorithm is shown in Figure 5.

In practice, a major advantage of the algorithm is that it typically takes very few steps to converge. This is due to the fact that the clusters are not changing very fast over a time gap  $l$ .

Consider two consecutive clustering calls at time  $t$  and  $t+l$ , then the time windows for clustering,  $[t-L+1, t]$  and  $[t+l-L+1, t+l]$ , significantly overlap with each other. As a result, each clustering often converges in a few steps if we start from the previous clustering result. As the data streams change over time, new clusters may emerge and existing clusters may disappear. A drawback of the conventional  $k$ -means algorithm is that the user needs to specify the number of clusters  $k$ .

To capture this dynamic evolution of data streams, we continuously update the number of clusters  $k$  every time a new time segment with length  $l$  is received. We assume that when  $k$  is updated regularly and frequently, it will not change abruptly. Therefore, if the number of clusters given by the previous clustering is  $k$ , we will only consider  $k-1$ ,  $k$  or  $k+1$  to be the current number of clusters  $k'$ . Then we choose  $k'$  as the one that produces the smallest objective function  $G$ . That is,

$$G_{k'} = \min\{G_{k-1}, G_k, G_{k+1}\}.$$

**Algorithm correlation- $k$ -means( $k, Center_k, R_k$ )**

**Input:** number of the clusters  $k$ , sets of the center points  $Center_k$ , current clustering result  $R_k$ ;

**Output:** updated clustering results  $R_k$  and its objective function value  $G_k$ ;

**begin**

1. **repeat**
2.   **for**  $i = 1$  **to**  $n$ 
  - calculate the correlation distances between stream  $X_i$  and centers of  $k$  clusters;
  - assign  $X_i$  to the cluster with the shortest distance;
- end for**
3.   compute the new center of each cluster, update the set of centers  $Center_k$ ;
4. **until** no change of clustering result
5. calculate the objective function  $G_k$

**end**

Fig. 5. The algorithm for clustering.

When considering  $k' = k-1$ , we initialize the clusters by merging the two clusters that are closest in their centers; when considering  $k' = k+1$ , we initialize the clusters by forming a new cluster as the stream in existing clusters that is the farthest from the its cluster center. After the initial clustering is given, we run  $k$ -means until it converges in order to measure the quality of the adjusted clusters. The adjust  $k$  algorithm is shown in Figure 6.

#### 4. Clustering on demand

Our above discussion only considers clustering data streams over a time period of fixed length  $L$ . In some applications, the length of the time period depends on users' demands. Here, we extend our clustering algorithm to support clustering on demand (COD), i.e. clustering over any time horizon at user's request [7]. We call the extended algorithm CORREL-COD. The CORREL-COD algorithm has an online component and an offline component. The online component calculates the summary information in CCRs, whereas the offline part performs clustering.

**Algorithm**  $\text{adjust}_k(k, \text{Center}_k, R_k)$ **Input:** number of the clusters  $k$ , sets of the center points  $\text{Center}_k$ , current clustering result  $R_k$ **Output:** updated number of clusters  $k'$ , updated clustering result  $R_{k'}$  and its objective function value  $G_{k'}$ **begin**

1. Calculation of  $R_{k+1}$ 
  - (a) Among all the clusters, choose the data stream  $X$  which is the farthest from its cluster center, set a new cluster with  $X$  as its center;
  - (b)  $\text{Center}_{k+1} = \text{Center}_k \cup \{X\}$
  - (c)  $\text{correlation-}k\text{-means}(k+1, \text{Center}_{k+1}, R_{k+1})$
2. Calculation of  $R_{k-1}$ 
  - (a) Choose two closest clusters, suppose their centers are  $C_1$  and  $C_2$ , respectively
  - (b) combine these two clusters into a new cluster, compute the center  $C_3$  of the new cluster
  - (c)  $\text{Center}_{k-1} = \text{Center}_k \cup \{C_3\} - \{C_1, C_2\}$
  - (d)  $\text{correlation-}k\text{-means}(k-1, \text{Center}_{k-1}, R_{k-1})$
3. choose the one with the best  $G$  from  $R_{k-1}, R_k, R_{k+1}$ , set  $k'$  and  $R_{k'}$  accordingly;

**end**Fig. 6. The algorithm for adjusting  $k$ .

We assume that the maximum time horizon over which the user will demand is  $L$ . Namely, we only need to preserve information for time period  $[t - L + 1, t]$ .

We perform  $k$ -means clustering in the offline part to meet the user's demands. The offline part first receives summary information from the online processor and calculates correlation coefficients and distances between streams, then performs the  $k$ -means algorithm to cluster the data streams. Since we cannot afford to store the raw data, we divide the time horizon into multiple segments and store the CCR synopsis for each segment. Therefore, for an arbitrary time horizon, it is often impossible to extract information over the exact horizon. We assemble a combination of partitioned time segments in such a way that minimizes the difference between the user-specified time horizon  $w$  and the best approximative time horizon  $w'$  over which we can get CCR synopsis.

Let  $L = 2^l$  and  $m$  be the maximum number of segments the memory can store. We need to design appropriate **partitioning scheme** on the lengths of segments. One way is to divide  $L$  into  $m$  parts of equal length  $L/m$ . The maximum difference between  $w$  and  $w'$  will be  $L/m$ . Since all segments have the same length, there may be excessive loss for the recent segments that contain newer, and hence more interesting and valuable, information. Another way is to assign segment lengths as  $1, 2, 2^2, 2^3, \dots, 2^{l-1}$ , which means that we store CCRs for time segments  $[t - 1, t], [t - 4, t - 2], [t - 8, t - 5], \dots, [t - L + 1, t - L/2]$ . The segments containing newer data will have shorter lengths, leading to higher clustering accuracy for newer data. However, in this way, the maximum difference between  $w$  and  $w'$  is will be  $L/2$ , which is excessively large.

We propose a new scheme that places more weights on recent data while making the difference  $|w - w'|$  as small as possible. We assume  $m > \log L$ . In our scheme, we arrange

segments with lengths of  $1, 2, 2^2, 2^3, \dots, 2^{l-1}$ , respectively. Let  $m' = m - \log L = m - l$ , and  $S_i$  denote the segment with length  $2^i$ . For the  $m'$  unassigned segments, we assign them according to the following rule.

We first remove  $S_{l-1}$ , replace the region covered by  $S_{l-1}$  by two more  $S_{l-2}$ , and then reduce  $m'$  by one. If  $m' > 0$ , we will keep splitting a larger segment into two smaller segments. When there are still  $S_{l-2}$  segments, we split the most recent  $S_{l-2}$  into two  $S_{l-3}$ , and then reduce  $m'$  by one, until all  $S_{l-2}$  segments are removed or  $m' = 0$ . If all  $S_{l-2}$  segments are removed and  $m' > 0$ , we will split the most recent  $S_{l-3}$  into two  $S_{l-4}$  and reduce  $m'$  by one. We repeat this process until  $m' = 0$ .

Given  $L, m$ , and  $m' = m - \log L$ , it is easy to show that using our scheme, the maximum length of any segment is  $2^k$ , where

$$k = \operatorname{argmax}_i (m' \leq C_i) - 1 \quad (22)$$

where

$$C_i = 2^{l-i+2} - (l - i + 4). \quad (23)$$

Let  $T_i$  the number of segments of type  $S_i$ , we can prove that (the proof is omitted):

$$T_k = C_{k+1} - m' + 1 \quad (24)$$

$$T_{k-1} = 2(2^{l-k} - T_k) - 1. \quad (25)$$

$$T_i = 1, \quad i = 0, 1, \dots, k-2 \quad (26)$$

$$T_i = 0 \text{ for } i > k. \quad (27)$$

**Example:** Suppose  $L = 1024, m = 20, l = 10$ . Then  $m' = m - \log L = 10$ . By (23), we get  $C_9 = 3, C_8 = 10, C_7 = 25$ . This gives  $k = 7$  since  $\operatorname{argmax}_i (m' \leq C_i) = 8$ . Then by (24), (25), and (26), we get

$$T_7 = 10 - 10 + 1 = 1, \quad T_6 = 2(2^{(10-7)} - 1) - 1 = 13$$

and

$$T_5 = T_4 = T_3 = T_2 = T_1 = T_0 = 1.$$

**Theorem 4.1** *Using the above partitioning scheme, we have  $m$  segments in total.*

**Proof.** The number of segments is

$$\begin{aligned} \sum_{i=1}^k T_i &= T_k + 2(2^{l-k+1} - T_k) - 1 + (k-2) \\ &= 2^{l-k+2} - T_k + k - 3 \\ &= 2^{l-k+2} - (C_k - m' + 1) + k - 3 \\ &= 2^{l-k+2} - 2^{l-k+2} + (l - k + 4) + m' + k - 4 \\ &= m' + l = m. \end{aligned}$$

**Theorem 4.2** *Using the above partitioning scheme, the total length of the  $m$  segments is  $L - 1$ .*

**Proof.** The total length of all segments is

$$\begin{aligned}
\sum_{i=0}^k 2^i T_i &= T_k 2^k + T_{k-1} 2^{k-1} + 2^{k-2} + 2^{k-3} + \dots + 2 + 1 \\
&= T_k 2^k + [2(2^{l-k} - T_k) - 1] 2^{k-1} + 2^{k-1} - 1 \\
&= 2^l - 1 = L - 1.
\end{aligned}$$

**Theorem 4.3** Suppose that the user demands a query for segments of length  $r \leq L$ . Using the above partitioning scheme, suppose  $r'$  is the total length of a set of selected segments that is closest to  $r$ . Then  $r' - r \leq 2^k$ , where  $k = \text{argmax}_i (m_i \leq C_i) - 1$ .

**Proof.** To form a set of the most recent segments with a total length closest to  $r$ , we can incrementally add  $S_1, S_2, \dots$  to the set until the total length of the selected segments  $r'$  is larger than  $r$ . Suppose the longest segment in the resulting set is  $S_g$ . Since  $r < L$  we have  $g \leq k$ , and  $r' - r \leq 2^g \leq 2^k$ .

The above results show that our partitioning schemes achieves two goals. First, we assign shorter segments to newer data and thus give newer data higher precision. Second, the largest possible difference between the length of the approximative combination of segments and the user requested length is lowered to  $2^k$ , which is much smaller than  $L/2$ .

## 5. Experimental results

To evaluate the performance of our algorithms, we test it using both synthetic data and real data on a PC with 1.7GHz CPU and 512 MB memory running Window XP. The systems are implemented using Visual C++ 6.0.

### 5.1 Testing data

We generate the synthetic data in the same way as in [6]. For each cluster, we first define a prototype  $p(\cdot)$  which is a stochastic process defined by a second-order difference equation:

$$\begin{aligned}
p(t + \Delta t) &= p(t) + p'(t + \Delta t) \\
p'(t + \Delta t) &= p'(t) + u(t), \quad t = 0, \Delta t, 2\Delta t, \dots
\end{aligned}$$

where  $u(t)$  are independent random variables uniformly distributed in an interval  $[-a, a]$ . The data streams in the cluster are then generated by “distorting” the prototype, both horizontally (by stretching the time axis) and vertically (by adding noise). The formulation for a data stream  $x(\cdot)$  is defined as:

$$x(t) = p(t + h(t)) + g(t),$$

where  $h(\cdot)$  and  $g(\cdot)$  are stochastic processes generated in the same way as the prototype  $p(\cdot)$ . The constant  $a$  that determines the smoothness of a process can be different for  $p(\cdot)$ ,  $h(\cdot)$ , and  $g(\cdot)$ , such as 0.04, 0.04, 0.5, respectively.

We can then generate different clusters by generating different prototype function  $p(\cdot)$ . For each prototype function, we randomly distort the prototype to general multiple data streams in that cluster.

The real data set (Figure 7) that we use contains average daily temperatures of 169 cities around the world, recorded since January 1, 1995 to present. Each city is regarded as a data stream and each stream has 3,416 points.

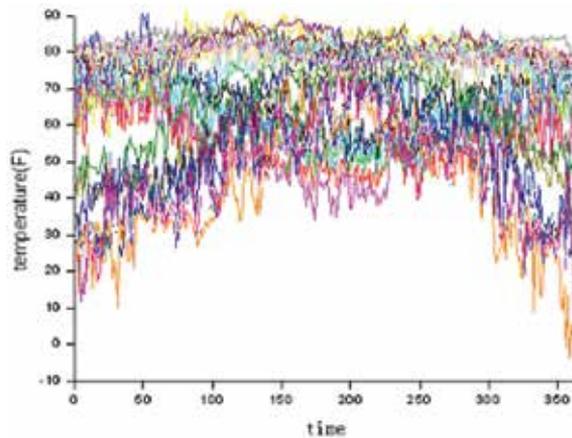


Fig. 7. Daily temperatures for 169 cities around the world.

## 5.2 Performance analysis on CORREL-cluster

### 5.2.1 Clustering results

We ran CORREL-cluster on the real data set to cluster cities based on the recorded daily temperatures. We set  $L = 360$  and  $l = 30$ . The input of the algorithm is the daily temperatures of cities around the world. CORREL-cluster gave five clusters each of which contains cities mostly in the same continent and belonging to the same temperature zone. The correct rate is around 85% to 89%. The results are shown in Figures 8-12, where each graph shows one cluster.

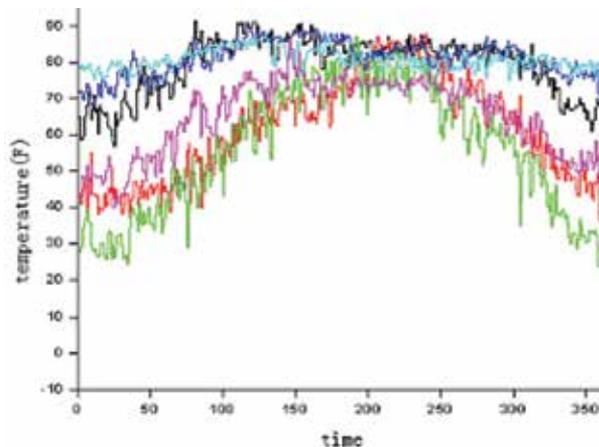


Fig. 8. Cluster 1: cities in Asia

### 5.2.2 Quality

We evaluate the quality of the clustering from CORREL-cluster by comparing with that from DFT-cluster [6] (30 DFT coefficients). Figure 13 shows a comparison of the quality of clustering on the real city-temperature data set by CORREL-cluster and DFT-cluster for various number of segments. The quality is measured by correct rate, the ratio of the number of cities that are correctly labelled to the total number of cities.

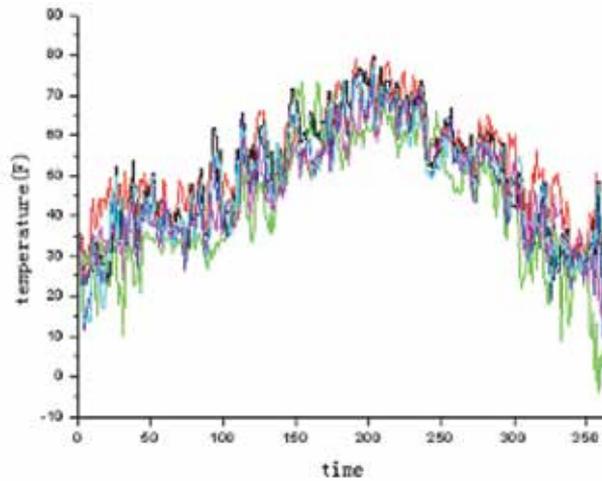


Fig. 9. Cluster 2: cities in Europe.

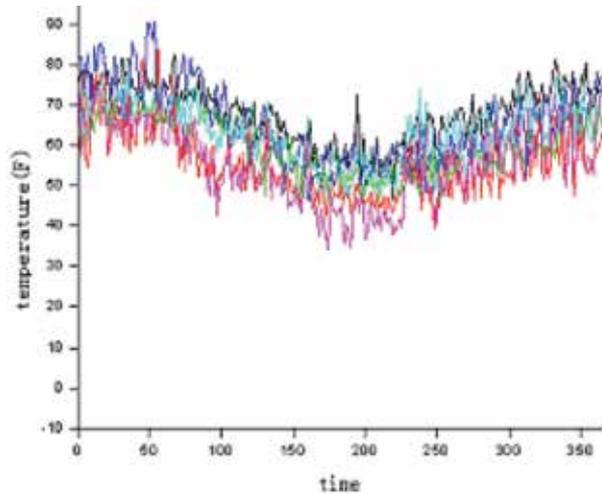


Fig. 10. Cluster 3: cities in Oceania.

Since we use a fixed time horizon  $L = 360$ , the larger the number of segments is, the more frequently clustering is executed. Thus, for both algorithms, the quality improves when the number of segments increases. However, as we can see from Figure 13, CORREL-cluster always has a better quality than DFTcluster.

### 5.2.3 Speed

Since the clustering on the real data set is too fast, we use synthetic data sets to test the processing speed of CORREL-cluster. We generate 6 synthetic data sets each containing 100 data streams. Each data stream has 65,536 data elements. Again, we compare with DFT-cluster (250 DFT coefficients). The experimental results show that the executing time for CORREL-cluster is shorter than that of DFT-cluster for every synthetic data set. Figure 14 shows that, the average processing time per segment for CORREL-cluster isn 0.928 seconds whereas 1.2 seconds for DFT-cluster using 250 DFT coefficients. DFT-cluster needs even

longer processing time when more coefficients are used. When using 1500 DFT coefficients, DFT-cluster takes in average over 7 seconds. Reducing the number of DFT coefficients can save time but will lead to worse quality. As we see in Figure 15, DFT-cluster with 250 DFT coefficients has much worse quality than CORREL-cluster on these synthetic data sets.

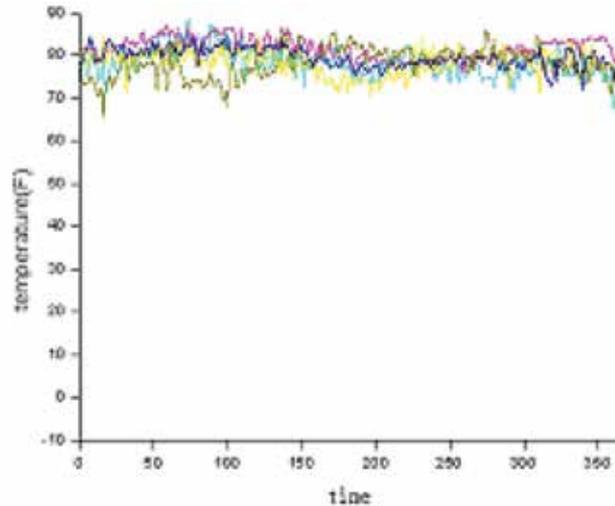


Fig. 11. Cluster 4: cities in Africa.

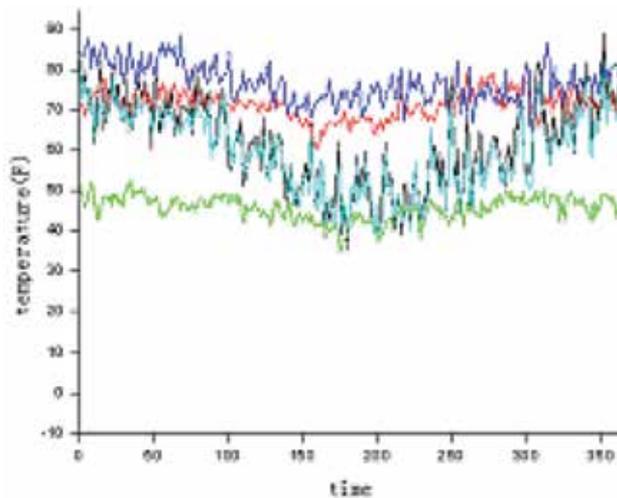


Fig. 12. Cluster 5: cities in South America.

#### 5.2.4 Dynamic number of clusters

CORREL-cluster requires an initial value  $k$  for the number of clusters. We study its sensitivity to  $k$  by trying different values of  $k$ . Figure 16 shows the clustering results of CORREL-cluster on a synthetic data set with three clusters for different initial values of  $k$ , including 2, 3, 4, 6, 8, 10 and 20. We see that the number of clusters given by CORREL-

cluster soon becomes the same regardless the initial value, which indicates that the initial value of  $k$  has little influence on the clustering performance. This stability is due to our adjust  $k()$  algorithm that adaptively changes the number of clusters in the process of clustering. This adaptiveness can be seen in Figure 16. For example, five clusters are found at time 32, four at time 38, and three at 52.

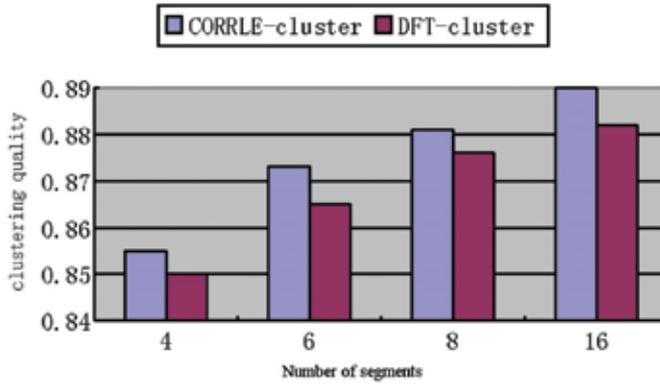


Fig. 13. Clustering quality of CORREL-cluster and DFT-cluster on real data.

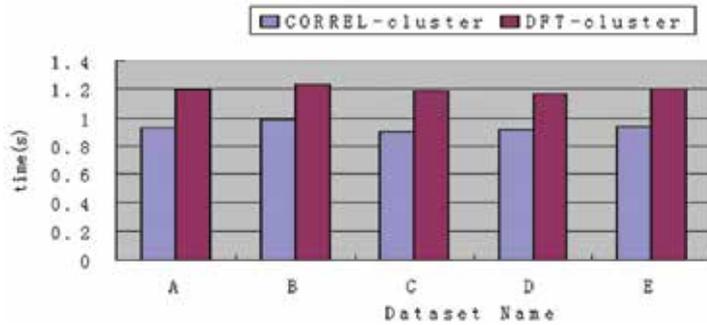


Fig. 14. Computation time of CORREL-cluster and DFT-cluster on synthetic data.

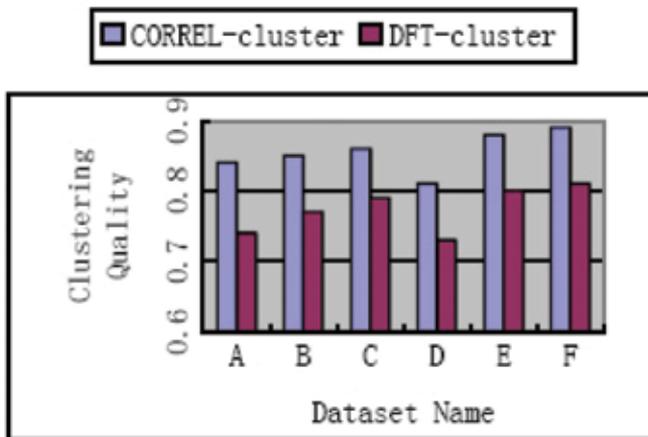


Fig. 15. Quality of CORREL-cluster and DFT-cluster on synthetic data.

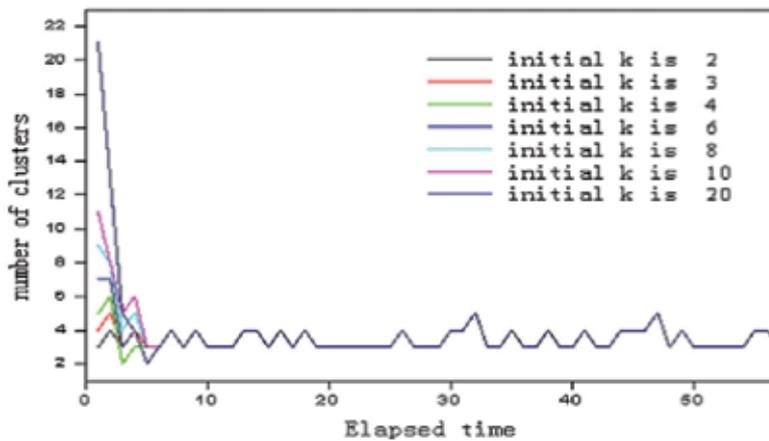


Fig. 16. Number of clusters found by CORREL-cluster for different initial values of  $k$ .

### 5.3 Performance analysis on CORREL-COD

#### 5.3.1 Scalability

To evaluate the scalability of the online processing, we test our CORREL-COD algorithm and ADAPTIVEcluster [7] using several randomly generated synthetic data sets of sizes varying from 1,000 to 10,000. As we see in Figure 17, the execution time for both algorithms increases linearly with the number of data points, but CORREL-COD is always more efficient than ADAPTIVE-cluster.

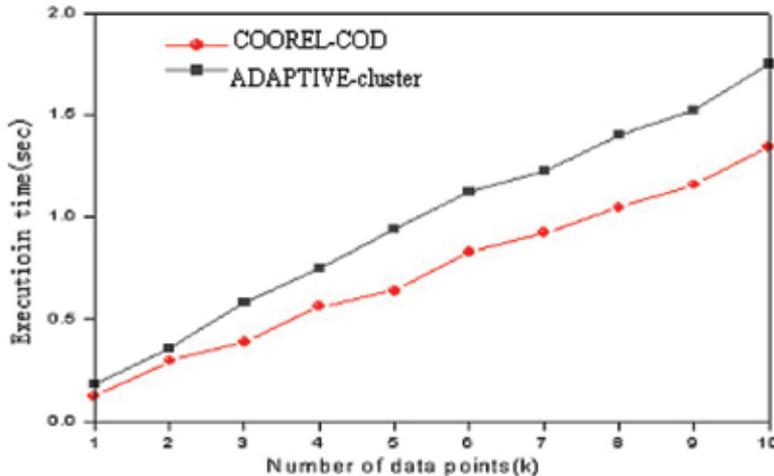


Fig. 17. Comparison of the scalability of CORREL-COD and ADAPTIVE-cluster.

#### 5.3.2 Quality

We measure the quality of clustering using  $G_{rawdata}/G_{COD}$ , where  $G_{rawdata}$  denotes the objective function obtained by clustering the raw data directly without segmentation and  $G_{COD}$  denotes the objective function by clustering based on the summary information retrieved by the online processor.

Figure 18 shows the quality of clustering on two simulated data sets for different number of segments. We see that the larger the number of segment is, the more precise our algorithm achieves. This is because the difference between the user specified length and the length of the approximated statistical information becomes smaller when the number of segments increases. From Figure 8, we can see the clustering quality is always above 95%, which means that results by our COD algorithm are close to the optimal results that can be obtained from raw data.

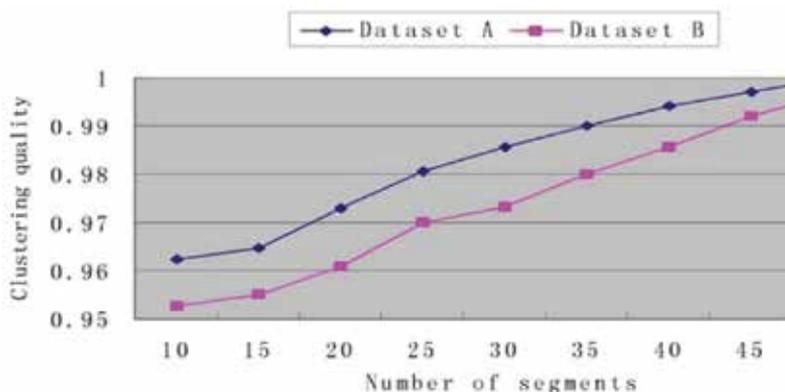


Fig. 18. Clustering quality for different numbers of segments.

## 6. Conclusions

When our aim is to mine the similarity on the trends of data streams, correlation coefficient is a more appropriate measure for similarity between data streams than Euclidean distance used by previous data stream clustering methods. In this paper, we have developed algorithms CORREL-cluster and CORREL COD for clustering multiple data streams based on correlation coefficients, which supports online clustering analysis over both fixed and flexible time horizons. Since data streams have high speed and massive volume, we can not retain the raw data to perform correlation analysis. We have proposed a compression scheme that supports an one-scan algorithm for computing the correlation coefficients for. We have developed an adaptive algorithm to dynamically determine the number of clusters so that CORREL-cluster can adjust to the evolving behaviors of data streams. Moreover, we have developed a novel partitioning algorithm to support clustering of arbitrary length per user's request. Experimental results on real and synthetic data sets show that our algorithms have high clustering quality, efficiency, and scalability.

## 7. References

- [1] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu. A framework for clustering evolving data streams. In *Proc. of conference of very large databases*, pages 81-92, 2003.
- [2] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu. A Framework for projected clustering of high dimensional data streams. In *Proc. of conference of very large databases*, pages 852-863, 2004.
- [3] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu. On-Demand Classification of Evolving Data streams. In *Proc. Of International Conference on Knowledge Discovery and Data Mining*, 2004.

- [4] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In *Proc. of the ACM SIGMOD Conference*, pages 94–105, Seattle, WA, 1998.
- [5] B. Babcock, M. Datar, R. Motwani, and L. O’Callaghan. Maintaining variance and k-medians over data stream windows. In *Proceedings of the twenty-second ACM symposium on Principles of database systems*, pages 234–243, 2003.
- [6] J. Beringer and E. Hüllermeier. Online-clustering of parallel data streams. *Data and Knowledge Engineering*, 58(2):180–204, 2006.
- [7] B.R. Dai, J.W. Huang, M.Y. Yeh, and M.S. Chen. Adaptive clustering for multiple evolving streams. *IEEE Transaction On Knowledge and data engineering*, 18(9), 2006.
- [8] P. Domingos and G. Hulten. A general method for scaling up machine learning algorithms and its application to clustering. In *Proc. of the Eighteenth International Conference on Machine Learning*, pages 106–113, 2001.
- [9] C. Giannella, J. Han, J. Pei, X. Yan, and P. S. Yu. Mining frequent patterns in data streams at multiple time granularities. In H. Kargupta, A. Joshi, K. Sivakumar, and Y. Yesha, editors, *Next Generation Data Mining*. AAAI/MIT, 2003.
- [10] S. Guha, N. Mishra, R. Motwani, and L. O’Callaghan. Clustering data streams. In *Annual IEEE Symposium on Foundations of Computer Science*, pages 359–366, 2000.
- [11] S. Guha, R. Rastogi, and K. Shim. CURE: An efficient Clustering algorithm for large databases. In *ACM SIGMOD Conference*, pages 73–84, 1998.
- [12] M.R. Henzinger, P. Raghavan, and S. Rajagopalan. Computing on data streams. Technical report, Digital Systems Research Center, 1998.
- [13] G. Hulten, L. Spencer, and P. Domingos. Mining time changing data streams. In *Proc. of ACM SIGKDD*, pages 97–106, 2001.
- [14] E. Keogh and S. Kasetty. On the need for time series data mining benchmarks: A survey and empirical demonstration. In *8th ACM SIGKDD Int’l Conference on Knowledge Discovery and Data Mining*, pages 102–111, 2002.
- [15] J. Sander X. Xu M. Ester, H.-P. Kriegel. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Int. Conf. on Knowledge Discovery and Data Mining (KDD’96)*, Portland, Oregon, 1996. AAAI Press.
- [16] S. Madden and M. Franklin. Fjording the stream: an architecture for queries over streaming sensor data. In *Proc. of ICDE*, pages 555–566, 2002.
- [17] G. Manku and R. Motwani. Approximate Frequency counts over data streams. In *Proceedings of conference of very large databases*, 2002.
- [18] A. Metwally, D. Agrawal, and A. El Abbadi. Efficient Computation of frequent and top-k elements in data streams. In *Proc. Of International Conference on Database Theory*, 2005.
- [19] L. O’Callaghan, N. Mishra, A. Meyerson, S. Guha, and R. Motwani. Streaming-data algorithms for high-quality clustering. In *Proc. of 18th International Conference on Data Engineering*, pages 685– 694, 2002.
- [20] H. Wang, W. Fan, P. S. Yu, and J. Han. Mining Concept-Drifting data streams using ensemble classifiers. In *Proc. Of International Conference on Knowledge Discovery and Data Mining*, 2003.
- [21] J. Yang. On the need for time series data mining benchmarks: A survey and empirical demonstration. In *Proc. of IEEE Int’l Conf. Data Mining*, pages 695–697, 2003.
- [22] T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH: an efficient data clustering method for very large databases. In *Proc. of ACM SIGMOD international conference on Management of data*, pages 103– 114, 1996.

# Mining Multiple-level Association Rules Based on Pre-large Concepts

Tzung-Pei Hong<sup>1,2</sup>, Tzu-Jung Huang<sup>1</sup> and Chao-Sheng Chang<sup>3</sup>

<sup>1</sup>*National University of Kaohsiung*

<sup>2</sup>*National Sun Yat-Sen University*

<sup>3</sup>*I-Shou University*

*Kaohsiung, Taiwan, R.O.C.*

## 1. Introduction

The goal of data mining is to discover important associations among items such that the presence of some items in a transaction will imply the presence of some other items. To achieve this purpose, Agrawal and his co-workers proposed several mining algorithms based on the concept of large itemsets to find association rules in transaction data (Agrawal et al., 1993a) (Agrawal et al., 1993b) (Agrawal & Srikant, 1994) (Agrawal & Srikant, 1995). They divided the mining process into two phases. In the first phase, frequent (large) itemsets are found based on the counts by scanning the transaction data. In the second phase, association rules were induced from the large itemsets found in the first phase. After that, several other approaches were also proposed (Fukuda et al., 1996) (Han & Fu, 1995) (Mannila et al., 1994) (Park et al., 1997) (Srikant & Agrawal, 1996) (Han et al., 2000) (Li et al., 2003).

Many of the above algorithms for mining association rules from transactions were executed in level-wise processes. That is, itemsets containing single items were processed first, then itemsets with two items were processed, then the process was repeated, continuously adding one more item each time, until some criteria were met. These algorithms usually considered the database size static and focused on batch mining. In real-world applications, however, new records are usually inserted into databases, and designing a mining algorithm that can maintain association rules as a database grows is thus critically important.

When new records are added to databases, the original association rules may become invalid, or new implicitly valid rules may appear in the resulting updated databases (Cheung et al., 1996) (Cheung et al., 1997) (Lin & Lee, 1998) (Sarda & Srinivas, 1998) (Zhang, 1999). In these situations, conventional batch-mining algorithms must re-process the entire updated databases to find final association rules. Cheung and his co-workers thus proposed an incremental mining algorithm, called FUP (Fast UPdate algorithm) (Cheung et al., 1996), for incrementally maintaining mined association rules and avoiding the shortcomings mentioned above. The FUP algorithm modified the Apriori mining algorithm (Agrawal & Srikant, 1994) and adopted the pruning techniques used in the DHP (Direct Hashing and Pruning) algorithm (Park et al., 1997). It first calculated large itemsets mainly from newly

inserted transactions, and compared them with the previous large itemsets from the original database. According to the comparison results, FUP determined whether re-scanning the original database was needed, thus saving some time in maintaining the association rules. Although the FUP algorithm can indeed improve mining performance for incrementally growing databases, original databases still need to be scanned when necessary. Hong et al. thus proposed a new mining algorithm based on two support thresholds to further reduce the need for rescanning original databases (Hong et al., 2001). They also used a data structure to further improve the performance (Hong et al., 2008).

Most algorithms for association rule mining focused on finding association rules on the single-concept level. However, mining multiple-concept-level rules may lead to discovery of more specific and important knowledge from data. Relevant data item taxonomies are usually predefined in real-world applications and can be represented using hierarchy trees. This chapter thus proposes an incremental mining algorithm to efficiently and effectively maintain the knowledge with a taxonomy based on the pre-large concept and to further reduce the need for rescanning original databases. Since rescanning the database spends much computation time, the mining cost can thus be reduced in the proposed algorithm.

The remainder of this chapter is organized as follows. Some related researches for incremental mining are described in Section 2. Data mining at multiple-level taxonomy is introduced in Section 3. The proposed incremental mining algorithm for multiple-level association rules is described in Section 4. An example to illustrate the proposed algorithm is given in Section 5. Conclusions are summarized in Section 6.

## 2. Some related researches for incremental mining

In real-world applications, transaction databases grow over time and the association rules mined from them must be re-evaluated because new association rules may be generated and old association rules may become invalid when the new entire databases are considered. Designing efficient maintenance algorithms is thus important.

In 1996, Cheung proposed a new incremental mining algorithm, called FUP (Fast Update algorithm) (Cheung et al., 1996) (Cheung et al., 1997) for solving the above problem. Using FUP, large itemsets with their counts in preceding runs are recorded for later use in maintenance. Assume there exist an original database and newly inserted transactions. FUP divides the mining process into the following four cases (Figure 1):

Case 1: An itemset is large in the original database and in the newly inserted transactions.

Case 2: An itemset is large in the original database, but is not large (small) in the newly inserted transactions.

Case 3: An itemset is not large in the original database, but is large in the newly inserted transactions.

Case 4: An itemset is not large in the original database and in the newly inserted transactions.

Since itemsets in Case 1 are large in both the original database and the new transactions, they will still be large after the weighted average of the counts. Similarly, itemsets in Case 4 will still be small after the new transactions are inserted. Thus Cases 1 and 4 will not affect the final association rules. Case 2 may remove existing association rules, and case 3 may add new association rules. FUP thus processes these four cases in the manner shown in Table 1.

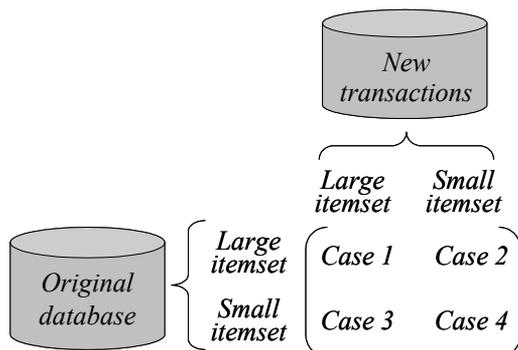


Fig. 1. Four cases in the FUP algorithm

Cases: Original - New	Results
Case 1: Large - Large	Always large
Case 2: Large - Small	Determined from existing information
Case 3: Small - Large	Determined by rescanning original database
Case 4: Small - Small	Always small

Table 1. Four cases and their FUP results

FUP thus focuses on the newly inserted transactions and can save some processing time in rule maintenance. But FUP still has to scan an original database for managing Case 3 in which a candidate itemset is large in newly inserted transactions but is small in the original database. This situation may often occur when the number of newly inserted transactions is small. For example, suppose only one transaction is inserted into a database. In this situation, each itemset in the transaction is large. Case 3 thus needs to be processed in a more efficient way.

Hong et al. thus propose a new mining algorithm based on pre-large itemsets to further reduce the need for rescanning original databases (Hong et al., 2001). A pre-large itemset is not truly large, but promises to be large in the future. A lower support threshold and an upper support threshold are used to realize this concept. The upper support threshold is the same as that used in the conventional mining algorithms. The support ratio of an itemset must be larger than the upper support threshold in order to be considered large. On the other hand, the lower support threshold defines the lowest support ratio for an itemset to be treated as pre-large. An itemset with its support ratio below the lower threshold is thought of as a small itemset. Pre-large itemsets act like buffers in the incremental mining process and are used to reduce the movements of itemsets directly from large to small and vice-versa.

Considering an original database and transactions newly inserted using the two support thresholds, itemsets may thus fall into one of the following nine cases illustrated in Figure 2. Cases 1, 5, 6, 8 and 9 above will not affect the final association rules according to the weighted average of the counts. Cases 2 and 3 may remove existing association rules, and cases 4 and 7 may add new association rules. If we retain all large and pre-large itemsets with their counts after each pass, then cases 2, 3 and case 4 can be handled easily. Also, in the maintenance phase, the ratio of new transactions to old transactions is usually very small. This is more apparent when the database is growing larger. An itemset in case 7 cannot possibly be large for the entire updated database as long as the number of

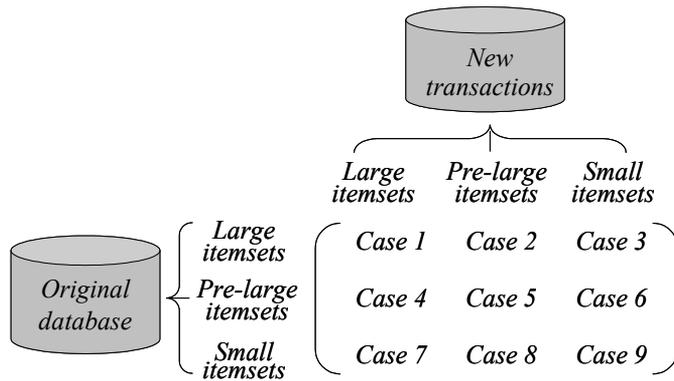


Fig. 2. Nine cases arising from adding new transactions to existing database

transactions is small when compared to the number of transactions in the original database. Let  $S_l$  and  $S_u$  be respectively the lower and the upper support thresholds, and let  $d$  and  $t$  be respectively the numbers of the original and new transactions. They showed that an itemset that is small (neither large nor pre-large) in the original database but is large in newly inserted transactions is not large for the entire updated database if the following condition is satisfied:

$$t \leq \frac{(S_u - S_l)d}{1 - S_u} \tag{1}$$

In this chapter, we will generalize Hong et al’s approach to maintain the association rules with taxonomy.

#### 4. Mining multiple-level association rules

Most algorithms for association rule mining focused on finding association rules on the single-concept level. However, mining multiple-concept-level rules may lead to discovery of more specific and important knowledge from data. Relevant data item taxonomies are usually predefined in real-world applications and can be represented using hierarchy trees. Terminal nodes on the trees represent actual items appearing in transactions; internal nodes represent classes or concepts formed by lower-level nodes. A simple example is given in Figure 3.

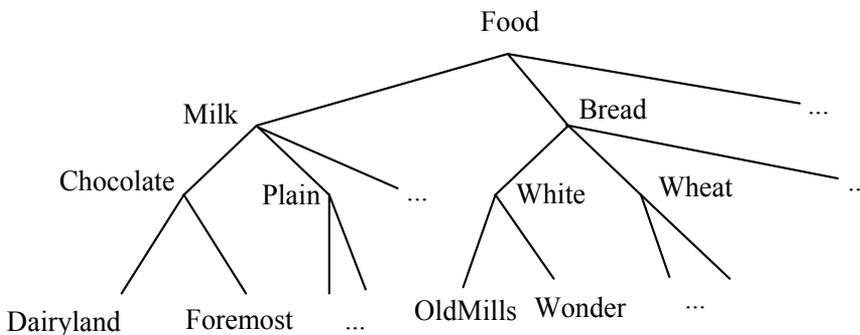


Fig. 3. An example of taxonomy

In Figure 3, the root node is at level 0, the internal nodes representing categories (such as “milk”) are at level 1, the internal nodes representing flavors (such as “chocolate”) are at level 2, and the terminal nodes representing brands (such as “Foremost”) are at level 3. Only terminal nodes appear in transactions.

Han and Fu proposed a method for finding level-crossing association rules at multiple levels (Han & Fu, 1995). Nodes in predefined taxonomies are first encoded using sequences of numbers and the symbol "\*" according to their positions in the hierarchy tree. For example, the internal node "Milk" in Figure 3 would be represented by 1\*\*, the internal node "Chocolate" by 11\*, and the terminal node "Dairyland" by 111. A top-down progressively deepening search approach is used and exploration of “level-crossing” association relationships is allowed. Candidate itemsets on certain levels may thus contain other-level items. For example, candidate 2-itemsets on level 2 are not limited to containing only pairs of large items on level 2. Instead, large items on level 2 may be paired with large items on level 1 to form candidate 2-itemsets on level 2 (such as {11\*, 2\*\*}).

## 5. The proposed incremental mining algorithm for multiple-level association rules

The proposed incremental mining algorithm integrates Hong et al’s pre-large concepts and Han and Fu’s multi-level mining method. Assume  $d$  is the number of transactions in the original database. A variable,  $c$ , is used to record the number of new transactions since the last re-scan of the original database. Details of the proposed mining algorithm are given below.

### The incremental multi-level mining algorithm:

INPUT: A set of large itemsets and pre-large itemsets in the original database consisting of  $(d+c)$  transactions, a set of  $t$  new transactions, a predefined taxonomy, a lower support threshold  $S_l$ , an upper support threshold  $S_u$ , and a predefined confidence value  $\lambda$ .

OUTPUT: A set of multi-level association rules for the updated database.

STEP 1: Calculate the safety number  $f$  of new transactions as follows:

$$f = \left\lfloor \frac{(S_u - S_l)d}{1 - S_u} \right\rfloor. \quad (2)$$

STEP 2: Set  $l = 1$ , where  $l$  records the level of items in taxonomy.

STEP 3: Set  $k = 1$ , where  $k$  records the number of items in itemsets.

STEP 4: Find all the candidate  $k$ -itemsets  $C_k$  and their counts in the new transactions.

STEP 5: Divide the candidate  $k$ -itemsets into three parts according to whether they are large, pre-large or small in the original database.

STEP 6: For each itemset  $I$  in the originally large  $k$ -itemsets  $L_k^D$ , do the following substeps:

Substep 6-1: Set the new count  $S^U(I) = S^T(I) + S^D(I)$ .

Substep 6-2: If  $S^U(I)/(d+t+c) \geq S_u$ , then assign  $I$  as a large itemset, set  $S^D(I) = S^U(I)$  and keep  $I$  with  $S^D(I)$ ;

otherwise, if  $S^U(I)/(d+t+c) \geq S_l$ , then assign  $I$  as a pre-large itemset, set  $S^D(I) = S^U(I)$  and keep  $I$  with  $S^D(I)$ ;

otherwise, neglect  $I$ .

- STEP 7: For each itemset  $I$  in the originally pre-large itemset  $P_k^D$ , do the following substeps:
- Substep 7-1: Set the new count  $S^U(I) = S^T(I) + S^D(I)$ .
  - Substep 7-2: If  $S^U(I)/(d+t+c) \geq S_{ll}$ , then assign  $I$  as a large itemset, set  $S^D(I) = S^U(I)$  and keep  $I$  with  $S^D(I)$ ;
  - otherwise, if  $S^U(I)/(d+t+c) \geq S_{li}$ , then assign  $I$  as a pre-large itemset, set  $S^D(I) = S^U(I)$  and keep  $I$  with  $S^D(I)$ ;
  - otherwise, neglect  $I$ .
- STEP 8: For each itemset  $I$  in the candidate itemsets that is not in the originally large itemsets  $L_k^D$  or pre-large itemsets  $P_k^D$ , do the following substeps:
- Substep 8-1: If  $I$  is in the large itemsets  $L_k^T$  or pre-large itemsets  $P_k^T$  from the new transactions, then put it in the rescan-set  $R$ , which is used when rescanning in Step 9 is necessary.
  - Substep 8-2: If  $I$  is small for the new transactions, then do nothing.
- STEP 9: If  $t + c \leq f$  or  $R$  is null, then do nothing; otherwise, rescan the original database to determine whether the itemsets in the rescan-set  $R$  are large or pre-large.
- STEP 10: Form candidate  $(k+1)$ -itemsets  $C_{k+1}$  from finally large and pre-large  $k$ -itemsets  $(L_k^U \cup P_k^U)$  that appear in the new transactions.
- STEP 11: Set  $k = k+1$ .
- STEP 12: Repeat STEPs 5 to 11 until no new large or pre-large itemsets are found.
- STEP 13: Prune the kept large or pre-large itemsets in the next level which are not the descendents of those found after STEP 12.
- STEP 14: Set  $l = l + 1$ .
- STEP 15: Repeat STEPs 3 to 14 until all levels are processed or there are no large and pre-large itemsets on level  $l - 1$ .
- STEP 16: Modify the association rules according to the modified large itemsets.
- STEP 17: If  $t + c > f$ , then set  $d = d + t + c$  and set  $c = 0$ ; otherwise, set  $c = t + c$ .

After Step 17, the final multi-level association rules for the updated database have been determined.

## 6. An example

An example is given to illustrate the proposed mining algorithm. Assume the original database includes 8 transactions as shown in Table 2.

Each transaction includes a transaction ID and some purchased items. For example, the eighth transaction consists of three items: Foremost plain milk, Old Mills white bread, and Wonder wheat bread. Assume the predefined taxonomy is as shown in Figure 4.

The food in Figure 4 falls into four main classes: milk, bread, cookie and beverage. Milk can further be classified into chocolate milk and plain milk. There are two brands of chocolate milk, Dairyland and Foremost. The other nodes can be similarly explained. Each item name in the taxonomy can then be encoded by Han and Fu's approach. Results are shown in Table 3.

TID	ITEMS
100	Dairyland chocolate milk, Foremost chocolate milk, Old Mills white bread, Wonder white bread, Linton green tea beverage
200	Dairyland chocolate milk, Foremost chocolate milk, Dairyland plain milk, Old Mills wheat bread, Wonder wheat bread, present lemon cookies, 77 lemon cookies
300	Foremost chocolate milk, Old Mills white bread, Old Mills wheat bread, 77 chocolate cookies, 77 lemon cookies
400	Dairyland chocolate milk, Old Mills white bread, 77 chocolate cookies
500	Old Mills white bread, Wonder wheat bread
600	Dairyland chocolate milk, Foremost plain milk, Wonder white bread, Nestle black tea beverage
700	Dairyland chocolate milk, Foremost chocolate milk, Dairyland plain milk, Old Mills white bread
800	Foremost plain milk, Old Mills white bread, Wonder wheat bread

Table 2. The original database in this example

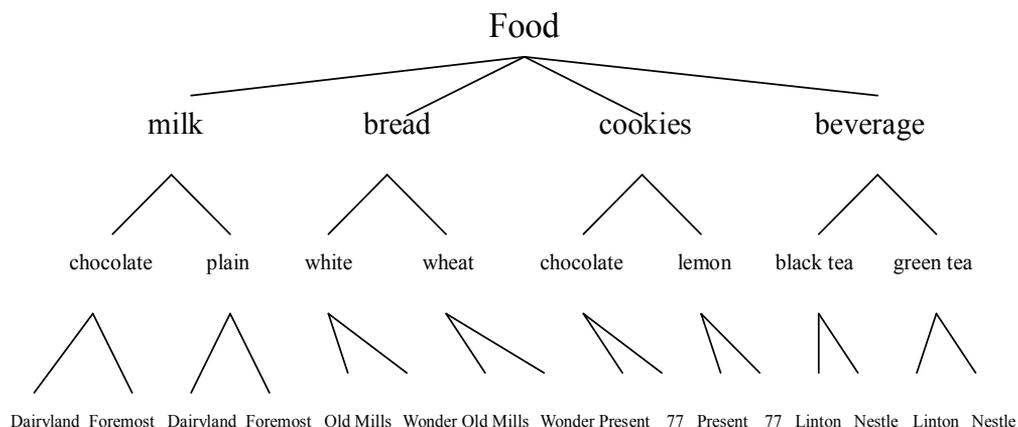


Fig. 4. The predefined taxonomy in this example

For example, the item "Foremost chocolate milk" is encoded as '112', in which the first digit '1' represents the code 'milk' on level 1, the second digit '1' represents the flavor 'chocolate' on level 2, and the third digit '2' represents the brand 'Foremost' on level 3. All the transactions shown in Table 2 are then encoded using the above coding table. Results are shown in Table 4

For  $S_l=30\%$  and  $S_{il}=50\%$ , the sets of large itemsets and pre-large itemsets on any level for the given original transaction database are shown in Table 5 and 6, respectively. They are then kept for later incremental mining.

Item name (terminal node)	Code	Item name (internal node)	Code
Dairyland chocolate milk	111	milk	1**
Foremost chocolate milk	112	bread	2**
Dairyland plain milk	121	cookies	3**
Foremost plain milk	122	beverage	4**
Old Mills white bread	211	chocolate milk	11*
Wonder white bread	212	plain milk	12*
Old Mills wheat bread	221	white bread	21*
Wonder wheat bread	222	wheat bread	22*
Present chocolate cookies	311	chocolate cookies	31*
77 chocolate cookies	312	lemon cookies	32*
Present lemon cookies	321	black tea beverage	41*
77 lemon cookies	322	green tea beverage	42*
Linton black tea beverage	411		
Nestle black tea beverage	412		
Linton green tea beverage	421		
Nestle green tea beverage	422		

Table 3. Codes of item names

TID	ITEMS
100	111, 112, 211, 212, 421
200	111, 112, 121, 221, 222, 322, 321
300	112, 211, 221, 312, 322
400	111, 211, 312
500	211, 222
600	111, 122, 212, 412
700	111, 112, 121, 211
800	122, 211, 222

Table 4. Encoded transaction data in the example

Level-1	Count	Level-2	Count	Level-3	Count
{1**}	7	{11*}	6	{111}	5
{2**}	8	{12*}	4	{112}	4
{1**, 2**}	7	{21*}	7	{211}	6
		{22*}	4		
		{11*, 21*}	5		

Table 5. The large itemsets on all levels for the original database

Level-1	Count	Level-2	Count	Level-3	Count
{3**}	3	{11*, 12*}	3	{222}	3
{1**, 3**}	3	{12*, 21*}	3	{111, 112}	3
{2**, 3**}	3	{21*, 22*}	3	{111, 211}	3
{1**, 2**, 3**}	3			{112, 211}	3

Table 6. The pre-large itemsets on all levels for the original database

Assume now the two new transactions shown in Table 7 are inserted to the original database.

New transactions	
TID	Items
900	112, 221, 311, 412
1000	111, 122, 412, 422

Table 7. Two new transactions

The proposed mining algorithm proceeds as follows. The variable  $c$  is initially set at 0.

STEP 1: The safety number  $f$  for new transactions is calculated as:

$$f = \left\lfloor \frac{(S_u - S_l)d}{1 - S_u} \right\rfloor = \left\lfloor \frac{(0.5 - 0.3)8}{1 - 0.5} \right\rfloor = 3. \tag{3}$$

STEP 2:  $l$  is set to 1, where  $l$  records the level of items in taxonomy.

STEP 3:  $k$  is set to 1, where  $k$  records the number of items in itemsets currently processed.

STEP 4: All candidate 1-itemsets  $C_1$  on Level 1 and their counts from the two new transactions are found, as shown in Table 8.

Candidate 1-itemsets	
Items	Count
{1**}	2
{2**}	1
{3**}	1
{4**}	2

Table 8. All candidate 1-itemsets on level

STEP 5: From Table 8, all candidate 1-itemsets {1\*\*}{2\*\*}{3\*\*}{4\*\*} are divided into three parts: {1\*\*}{2\*\*}, {3\*\*}, and {4\*\*} according to whether they are large, pre-large or small in the original database. Results are shown in Table 9.

Originally large 1-itemsets		Originally pre-large 1-itemsets		Originally small 1-itemsets	
Items	Count	Items	Count	Items	Count
{1**}	2	{3**}	1	{4**}	2
{2**}	1				

Table 9. Three partitions of all candidate 1-itemsets from the two new transactions

STEP 6: The following substeps are done for each of the originally large 1-itemsets  $\{1^{**}\}$  $\{2^{**}\}$ :

Substep 6-1: The total counts of the candidate 1-itemsets  $\{1^{**}\}$  $\{2^{**}\}$  are calculated using  $S^T(I) + S^D(I)$ . Table 10 shows the results.

Items	Count
$\{1^{**}\}$	9
$\{2^{**}\}$	9

Table 10. The total counts of  $\{1^{**}\}$  $\{2^{**}\}$

Substep 6-2: The new support ratios of  $\{1^{**}\}$  $\{2^{**}\}$  are calculated. For example, the new support ratio of  $\{1^{**}\}$  is  $9/(8+2+0) \geq 0.5$ .  $\{1^{**}\}$  is thus still a large itemset. In this example, both  $\{1^{**}\}$  and  $\{2^{**}\}$  are large.  $\{1^{**}\}$  $\{2^{**}\}$  with their new counts are then retained in the large 1-itemsets for the entire updated database.

STEP 7: The following substeps are done for itemset  $\{3^{**}\}$ , which is originally pre-large:

Substep 7-1: The total count of the candidate 1-itemset  $\{3^{**}\}$  is calculated using  $S^T(I) + S^D(I)$  ( $= 4$ ).

Substep 7-2: The new support ratio of  $\{3^{**}\}$  is  $4/(8+2+0) \leq 0.5$ .  $\{3^{**}\}$  isn't a large 1-itemset for the whole updated database.  $\{3^{**}\}$  with its new count is, however, a pre-large 1-itemset for the entire updated database.

STEP 8: Since the itemset  $\{4^{**}\}$ , which was originally neither large nor pre-large, is large for the new transactions, it is put in the rescan-set  $R$ , which is used when rescanning in Step 9 is necessary.

STEP 9: Since  $t + c = 2 + 0 \leq f$  ( $=3$ ), rescanning the database is unnecessary, so nothing is done.

STEP 10: After STEPs 8 and 9, the final large 1-itemsets for the entire updated database are  $\{1^{**}\}$  $\{2^{**}\}$  and the final pre-large 1-itemset is  $\{3^{**}\}$ . Since all of them are in the new transactions, the candidate 2-itemsets are shown in Table 11.

Candidate 2-itemsets
$\{1^{**}, 2^{**}\}$
$\{1^{**}, 3^{**}\}$
$\{2^{**}, 3^{**}\}$

Table 11. All candidate 2-itemsets for the new transactions

STEP 11:  $k = k + 1 = 2$ .

STEP 12: STEPs 5 to 11 are repeated to find large and pre-large 2-itemsets on level 1. Results are shown in Table 12.

Large 2-Itemsets		Pre-large 2-Itemsets	
Items	Count	Items	Count
$\{1^{**}, 2^{**}\}$	8	$\{2^{**}, 3^{**}\}$	4
		$\{1^{**}, 3^{**}\}$	4

Table 12. All large and pre-large 2-itemsets on level 1 for the updated database

Large or pre-large 3-itemsets are then found in the same way. The results are shown in Table 13.

Large 3-Itemsets		Pre-large 3-Itemsets	
Items	Count	Items	Count
		{1**, 2**, 3**}	3

Table 13. All large and pre-large 3-itemsets on level 1 for the updated database

STEP 13: The large and pre-large itemsets on level 1 are then used to prune the originally kept itemsets on level 2. If an itemset originally kept on level 2 is not a descendent of any one on level 1, it is pruned. In this example, since all itemsets originally kept on level 2 are descendants of those on level 1, no pruning is made.

STEP 14:  $l = l + 1 = 2$ . Large and pre-large itemsets on the second level are to be found.

STEP 15: Steps 3 to 14 are repeated to find all large and pre-large itemsets on level 2. The results are shown in Table 14.

Large itemsets			Pre-large itemsets		
1 item	2 items	3 items	1 item	2 items	3 items
{11*}	{11*, 21*}		{31*}	{11*, 12*}	
{12*}				{11*, 22*}	
{21*}				{12*, 21*}	
{22*}				{21*, 22*}	
				{11*, 31*}	

Table 14. All the large and pre-large itemsets on level 2 for the updated database

Similarly, all large and pre-large itemsets on level 3 are found and shown in Table 15.

Large itemsets			Pre-large itemsets		
1 item	2 items	3 items	1 item	2 items	3 items
{111}			{122}	{111, 112}	
{112}			{221}	{111, 211}	
{211}			{222}	{112, 211}	
				{112, 221}	

Table 15. All the large and pre-large itemsets on level 3 for the updated database

The large itemsets on all levels are listed in Table 16.

Level-1	Level-2	Level-3
{1**}	{11*}	{111}
{2**}	{12*}	{112}
{1**, 2**}	{21*}	{211}
	{22*}	
	{11*, 21*}	

Table 16. The large itemsets on all levels for the updated database

STEP 16: Assume the minimum confidence value is set at 0.7. The association rules are then modified according to the modified large itemsets as follows:

$$\begin{aligned}
 &1^{**} \Rightarrow 2^{**} \text{ (Confidence=8/9),} \\
 &2^{**} \Rightarrow 1^{**} \text{ (Confidence=8/9),} \\
 &21^* \Rightarrow 11^* \text{ (Confidence=5/7),} \\
 &11^* \Rightarrow 2^{**} \text{ (Confidence=7/8),} \\
 &2^{**} \Rightarrow 11^* \text{ (Confidence=7/9), and} \\
 &21^* \Rightarrow 1^{**} \text{ (Confidence=6/7).}
 \end{aligned}$$

STEP 17: Since  $t (= 2) + c (= 0) < f (= 3)$ ,  $c = t + c = 2 + 0 = 2$ .

After Step 17, the final association rules for the updated database are found.

## 7. Conclusion

In this chapter, we have proposed an incremental mining algorithm for multiple-level association rules. It may lead to discovery of more specific and important knowledge from data. The proposed algorithm is based on the pre-large concept and can efficiently and effectively mine knowledge with a taxonomy in an incremental way. The large itemsets play a very critical role to reduce the need for rescanning original databases. Since rescanning the database spends much computation time, the mining cost can thus be greatly reduced in the proposed algorithm. An example is given to illustrate the proposed approach. It can also be easily observed from the example that much rescanning can be avoided. Using the concept of pre-large itemsets is thus a good way to incremental mining, no matter for single levels or multiple levels.

## 8. References

- R. Agrawal, T. Imielinski and A. Swami, "Mining association rules between sets of items in large database", The ACM SIGMOD Conference, pp. 207-216, Washington DC, USA, 1993. (a)
- R. Agrawal, T. Imielinski and A. Swami, "Database mining: a performance perspective", IEEE Transactions on Knowledge and Data Engineering, Vol. 5, No. 6, pp. 914-925, 1993. (b)
- R. Agrawal and R. Srikant, "Fast algorithm for mining association rules," The International Conference on Very Large Data Bases, pp. 487-499, 1994.
- R. Agrawal and R. Srikant, "Mining sequential patterns," The Eleventh IEEE International Conference on Data Engineering, pp. 3-14, 1995.
- R. Agrawal, R. Srikant and Q. Vu, "Mining association rules with item constraints," The Third International Conference on Knowledge Discovery in Databases and Data Mining, pp. 67-73, Newport Beach, California, 1997.
- D.W. Cheung, J. Han, V.T. Ng, and C.Y. Wong, "Maintenance of discovered association rules in large databases: An incremental updating approach," The Twelfth IEEE International Conference on Data Engineering, pp. 106-114, 1996.

- D.W. Cheung, V.T. Ng, and B.W. Tam, "Maintenance of discovered knowledge: a case in multi-level association rules," The Second International Conference on Knowledge Discovery and Data Mining (KDD-96), pp. 307-310, 1996.
- D.W. Cheung, S.D. Lee, and B. Kao, "A general incremental technique for maintaining discovered association rules," In Proceedings of Database Systems for Advanced Applications, pp. 185-194, Melbourne, Australia, 1997.
- T. Fukuda, Y. Morimoto, S. Morishita and T. Tokuyama, "Mining optimized association rules for numeric attributes," The ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, pp. 182-191, 1996.
- J. Han and Y. Fu, "Discovery of multiple-level association rules from large database," The Twenty-first International Conference on Very Large Data Bases, pp. 420-431, Zurich, Switzerland, 1995.
- J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," The 2000 ACM SIGMOD International Conference on Management of Data, pp. 1-12, 2000.
- T. P. Hong, C. Y. Wang and Y. H. Tao, "A new incremental data mining algorithm using pre-large itemsets," Intelligent Data Analysis, Vol. 5, No. 2, 2001, pp. 111-129.
- T. P. Hong, J. W. Lin, and Y. L. Wu, "Incrementally fast updated frequent pattern trees", Expert Systems with Applications, Vol. 34, No. 4, pp. 2424 - 2435, 2008. (SCI)
- Y. Li, P. Ning, X. S. Wang and S. Jajodia. "Discovering calendar-based temporal association rules," Data & Knowledge Engineering, Vol. 44, No. 2, pp. 193-218, 2003.
- M. Y. Lin and S. Y. Lee, "Incremental update on sequential patterns in large databases," The Tenth IEEE International Conference on Tools with Artificial Intelligence, pp. 24-31, 1998.
- H. Mannila, H. Toivonen, and A. I. Verkamo, "Efficient algorithm for discovering association rules," The AAAI Workshop on Knowledge Discovery in Databases, pp. 181-192, 1994.
- J. S. Park, M. S. Chen, P. S. Yu, "Using a hash-based method with transaction trimming for mining association rules," IEEE Transactions on Knowledge and Data Engineering, Vol. 9, No. 5, pp. 812-825, 1997.
- N. L. Sarda and N. V. Srinivas, "An adaptive algorithm for incremental mining of association rules," The Ninth International Workshop on Database and Expert Systems, pp. 240-245, 1998.
- R. Srikant and R. Agrawal, "Mining generalized association rules," The Twenty-first International Conference on Very Large Data Bases, pp. 407-419, Zurich, Switzerland, 1995.
- R. Srikant and R. Agrawal, "Mining quantitative association rules in large relational tables," The 1996 ACM SIGMOD International Conference on Management of Data, pp. 1-12, Montreal, Canada, 1996.

S. Zhang, "Aggregation and maintenance for database mining," *Intelligent Data Analysis*, Vol. 3, No. 6, pp. 475-490, 1999.

## **PART III: CHALLENGES AND BENCHMARKS IN DATA MINING**



# Data Mining Applications: Promise and Challenges

Rukshan Athauda<sup>1</sup>, Menik Tissera<sup>2</sup> and Chandrika Fernando<sup>2</sup>

<sup>1</sup>*The University of Newcastle*

<sup>2</sup>*Sri Lanka Institute of Information Technology*

<sup>1</sup>*Australia*

<sup>2</sup>*Sri Lanka*

## 1. Introduction

Data mining is an emerging field gaining acceptance in research and industry. This is evidenced by an increasing number of research publications, conferences, journals and industry initiatives focused in this field in the recent past.

Data mining aims to solve an intricate problem faced by a number of application domains today with the deluge of data that exists and is continually collected, typically, in large electronic databases. That is, to extract useful, meaningful knowledge from these vast data sets. Human analytical capabilities are limited, especially in its ability to analyse large and complex data sets. Data mining provides a number of tools and techniques that enables analysis of such data sets. Data mining incorporates techniques from a number of fields including statistics, machine learning, database management, artificial intelligence, pattern recognition, and data visualisation.

A number of definitions for data mining are presented in literature. Some of them are listed below:

- “Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques” (Gartner Group, 1995).
- “Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner” (Hand et al., 2001).
- “Data mining is an interdisciplinary field bringing together techniques from machine learning, pattern recognition, statistics, databases, and visualization to address the issue of information extraction from large data bases” (Cabena et al., 1998).
- “The extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data” (Han & Kamber, 2001).

We present application of data mining (also known as “Data Mining Applications”) as an “experiment” carried out using data mining techniques that result in gaining useful knowledge and insights pertaining to the application domain. Figure 1 below depicts this process.

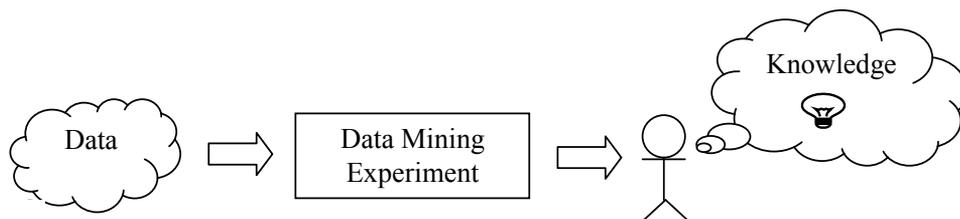


Fig. 1. A Data Mining Experiment

The analogy between conducting an experiment and applying data mining will become evident as we discuss further the issues and challenges faced in data mining applications.

Data mining (DM) application shows the promise to unlocking previously unknown, interesting knowledge from vast collections of data in many different application domains. DM has been successfully applied in a number of different domains (for e.g. astronomy (Weir et al., 1995), decision support (Wang & Weigend, 2004), business (Ester et al., 2002), IT security (Lee et al., 2001; Julisch & Dacier, 2002), medical and health research (Antonie et al., 2001; Au et al., 2005; Yu et al., 2004; Yan et al., 2004; Daxin et al., 2004; Liang & Kelemen, 2002; Chun et al., 2003), text mining (Hearst, 1999), marketing (Kitts et al., 2000), financial forecasting (Li & Kuo, 2008), education (Druzdzal & Glymour, 1994; Tissera et al., 2006; Ma et al., 2000), fraud detection (Senator et al., 1995; Rosset et al., 1999), and others (Fernando et al., 2008; Last et al., 2003; Fayyad et al., 1996c). Although a number of successful data mining applications exist, applying data mining is neither trivial nor straight-forward. Due to the inherent exploratory nature in data mining, there is a significant risk with little or no guarantee of successful results or Return on Investment (RoI) at the onset of a data mining initiative. Conducting a DM experiment is resource intensive and requires careful planning, critical analysis and extensive human judgement.

Today, a number of process models exists which aims to provide structure, control and standard methodology in applying data mining (Fayyad et al., 1996a; Cabena et al., 1998; Anand & Buchner, 1998; Chapman et al. 2000; Cios et al., 2000; Han & Kamber, 2001; Adriaans & Zantinge, 1996; Berry & Linoff, 1997; SAS Institute, 2003; Edelstein, 1998; Klösgen & Zytkow, 2002; Haglin et al. 2005). These process models are also known as Knowledge Discovery and Data Mining (KDDM) process models (Kurgan & Musilek, 2006). The KDDM process models outline the fundamental set of steps (typically executed iteratively with loopbacks) required in data mining applications covering the entire lifecycle in applying data mining from the initial goal determination to the final deployment and maintenance of the discovered knowledge.

Although KDDM process models provide a high-level structure to conduct a data mining experiment, following a KDDM process model by itself does not guarantee success. Applying data mining is an iterative process with extensive human intervention, whereby the DM team gains insights to the patterns and trends in data through the application of DM techniques with each iteration. This results in more-or-less a trial-and-error approach. The authors believe that a number of fundamental questions needs to be resolved in order to provide a more predictive, less risky approach with higher certainty of success in applying data mining. This, in turn, will enable fulfilling the promise of data mining resulting in a proliferation of data mining applications. The chapter aims to bring to light some of these fundamental questions.

Presently, research in data mining has mainly focused on different approaches for data analysis (techniques such as clustering, classification, associations, neural networks, genetic algorithms and others). A number of scalable algorithms enabling analysis of heavy volumes of data are proposed (Apriori (Agrawal et al., 1993); Auto Regression Trees (Meek et al., 2002); CURE (Guha et al.,1998); Two-Dimensional Optimized Association Rules (Fukuda et al.,1996); ML-T2LA (Han & Fu, 1995) and many others (Wu et al. 2008) ). These contributions have provided a rich set of techniques for data analysis. However, in comparison, we observe only a few research work in data mining that discusses practical aspects pertaining to data mining application (Feelders et al., 2000; Karim, 2001). There is a dire need for research focusing these challenges and issues in data mining application.

In comparison to many other fields of study, data mining is evolving and still in its infancy. The analogy between data mining and the field of software development is presented to illustrate this evolution in perspective. Initially, research in software development focused on the development of programming languages and techniques. Later research focused on issues pertaining to large scale software development and methodologies with an entire discipline of software engineering in existence today. Similarly, the field of data mining is still in its infancy where presently more focus is on different techniques and data mining tasks in data analysis. A predictive, controlled approach to data mining application is yet to be realised.

This chapter aims to bring attention to some of the fundamental challenging questions faced in applying data mining with the hope that future research aims to resolve these issues. This chapter is organised as follows: Section 2 briefly discusses the KDDM process models and basic steps proposed for applying data mining. Section 3 discusses the fundamental questions faced during data mining application process. Section 4 provides a discussion and recommendations for conducting a data mining experiment. Section 5 concludes the paper.

## **2. Knowledge discovery and data mining process models**

KDDM process models provide the contemporary guidelines in applying data mining. A number of KDDM process models have been discussed in literature (Fayyad et al.,1996a; Cabena et al., 1998; Chapman et al. 2000; Anand & Buchner, 1998; Cios et al., 2000; Han & Kamber, 2001; Adriaans & Zantinge, 1996; Berry & Linoff, 1997; SAS Institute, 2003; Edelstein , 1998; Klösgen & Zytkow, 2002; Haglin et al. 2005). The basic structure for KDDM process models was initially proposed by Fayyad et al. (Fayyad et al.,1996a, Fayyad et al.,1996b) (popularly known as the “KDD Process”) with other models proposed later. A survey and comparison of prominent KDDM process models are presented in (Kurgan & Musilek, 2006).

The KDDM process models outline the fundamental set of steps (typically executed iteratively with loopbacks) required in data mining applications. KDDM process models span the entire lifecycle in applying data mining, from the initial goal determination to the final deployment of the discovered knowledge. It is noteworthy to point out that data mining is considered as a single step in the overall process of applying data mining in KDDM process. The different KDDM process models are similar except mainly for terminology, orientation towards research vs. industry contexts and emphasis on different steps. In (Kurgan & Musilek, 2006), the prominent KDDM process models (Fayyad et al.,1996a; Cabena et al., 1998; Chapman et al. 2000; Anand & Buchner, 1998; Cios et al., 2000)

are compared. Figure 2 illustrates a comparison table presented in (Kurgan & Musilek, 2006) mapping the different steps in the prominent process models. It is evident that all process models follow more-or-less a similar set of steps. Some of the similarities outlined in (Kurgan & Musilek, 2006) are given below:

- *All process models follow a set of steps:* "All process models consist of multiple steps executed in a sequence, which often includes loops and iterations. Each subsequent step is initiated upon the successful completion of a previous step, and requires a result generated by the previous step as its inputs".
- *All process models cover entire lifecycle of a data mining experiment:* "Another common feature of the proposed models is the span of covered activities. It ranges from the task of understanding the project domain and data, through data preparation and analysis, to evaluation, understanding, and application of the generated results."
- *All process models are iterative in nature:* "All proposed models also emphasize the iterative nature of the model, in terms of many feedback loops and repetitions, which are triggered by a revision process."

Further discussion on KDDM process models are presented in (Kurgan & Musilek, 2006).

Model	Fayyad <i>et al.</i>	Cabeza <i>et al.</i>	Azand & Buchner	CRISP-DM	Cios <i>et al.</i>	Generic model
Area	Academic	Industrial	Academic	Industrial	Academic	N/A
No of steps	9	5	8	6	6	6
Refs	(Fayyad <i>et al.</i> , 1996d)	(Cabeza <i>et al.</i> , 1998)	(Azand & Buchner, 1998)	(Shearer, 2000)	(Cios <i>et al.</i> , 2000)	N/A
Steps	1 Developing and Understanding of the Application Domain	1 Business Objectives Determination	1 Human Resource Identification	1 Business Understanding	1 Understanding the Problem Domain	1 Application Domain Understanding
	2 Creating a Target Data Set	2 Data Preparation	2 Problem Specification	2 Data Understanding	2 Understanding the Data	2 Data Understanding
	3 Data Cleaning and Preprocessing		3 Data Prospecting			
	4 Data Reduction and Projection		4 Domain Knowledge Elicitation	3 Data Preparation	3 Preparation of the Data	3 Data Preparation and Identification of DM Technology
	5 Choosing the DM Task		5 Methodology Identification			
	6 Choosing the DM Algorithm		6 Data Preprocessing			
	7 DM	3 DM		4 Modeling	4 DM	4 DM
	8 Interpreting Mined Patterns	4 Domain Knowledge Elicitation	7 Pattern Discovery	5 Evaluation	5 Evaluation of the Discovered Knowledge	5 Evaluation
	9 Consolidating Discovered Knowledge	5 Assimilation of Knowledge	8 Knowledge Post-processing	6 Deployment	6 Using the Discovered Knowledge	6 Knowledge Consolidation and Deployment

Fig. 2. Comparison of steps in prominent KDDM process models (Source: Kurgan & Musilek, 2006)

Due to the differences in terminology and the number of steps between different KDDM process models, we adopted the Generic Model presented in (Kurgan & Musilek, 2006) to describe the KDDM process:

1. *Application Domain Understanding:* In this step, the high-level goals of the data mining project (i.e. business/customer objectives) are determined and explicitly stated. The context in which the experiment is conducted is understood. The business/customer goals are mapped to data mining goals. The preliminary project plan is developed.
2. *Data Understanding:* This step pertains to identifying the relevant data sources and parameters required by the relevant data mining tasks. It includes elicitation of relevant domain knowledge about the data sets, selection and merging of data, identifying quality issues (completeness, redundancy, missing, erroneous data etc.). Exploring data

at this stage may also lead to hypothesis creation and determination of data mining tasks (Chapman et al., 2000).

3. *Data Preparation and Identification of DM Technology:* At this stage, all necessary tasks needed to perform data mining are finalised. Data mining techniques and algorithms are decided. The data sets are pre-processed for specific data mining tasks. Pre-processing includes selecting, cleaning, deriving, integrating and formatting data in order to apply specific data mining tasks.
4. *Data Mining:* In this step, the data mining tasks are applied to the prepared data set. At this stage, a DM model is developed to represent various characteristics of the underlying data set. A number of iterations in fine-tuning the model may take place with the involvement of data mining experts.
5. *Evaluation:* In this phase, the results of the generated models are visualised and evaluated. Both domain and data mining expertise is incorporated for visualisation and interpretation of results to find useful and interesting patterns. Based on the evaluation of results, the decisions to iterate any of the previous or current steps, or to conclude the project is made.
6. *Knowledge Consolidation and Deployment:* At this stage, the final report documenting the experiment and results are published. In addition, incorporation and formulation of how to exploit the discovered knowledge is taken into consideration. Deployment of discovered knowledge includes plan for implementation, monitoring and maintenance.

Further discussion and comparison of individual steps of the different KDDM process models are presented in (Kurgan & Musilek, 2006) (see Table 2. pp. 9-12 of Kurgan & Musilek, 2006).

Although the KDDM process models provide a structure and a set of high-level steps in applying data mining, there are a number of issues, challenges and decisions faced by the DM team during the project. Decisions taken by the DM team at each stage has a significant impact on the outcome of the experiment. Typically these issues, challenges and decisions are not covered in generic KDDM process models. Research focusing on “guidelines” and “best-practices” to assist in resolving some of these challenging questions would be of significant value – especially to reduce risks and uncertainty associated in conducting a data mining experiment. The authors believe that successful strides in these areas will enable proliferation of data mining applications in many different domains. The next section highlights some of these issues and challenges faced during data mining applications.

### 3. Challenges in data mining application

Data mining by definition is exploratory in nature – that is, we are in search for previously unknown, hidden and interesting patterns in data. The fact that we are in search for unknown, hidden knowledge makes the outcome of data mining application difficult to predict at the onset of a DM project making it a risky and an uncertain endeavour. “Does interesting, relevant knowledge exist?”, “What types of knowledge are we looking for?”, “What method should we consider in order to find what we are looking for?”, “How do we know whether we haven’t missed any interesting ‘knowledge’ in the data set?” are some of the fundamental questions that pertain to data mining application. At present, these questions are answered based on the judgement of the DM team. To assist in these judgements, the iterative nature of KDDM process allows the DM team to try out certain data mining tasks, if failed, to re-tract and repeat until satisfactory results are achieved (or in

the worst case, resources are exhausted). This approach makes contemporary data mining application, a risky endeavour and typically follows a trial-and-error process.

In this section, we present some of the challenging questions faced by the DM team during DM application:

**Application Domain Understanding:** In this step, the overall goal for a data mining experiment is determined. The significance of the impact of determining goals on the outcome of the data mining experiment is highlighted in (Chapman et al., 2000) as follows: "A possible consequence of neglecting this step is to expend a great deal of effort producing the right answers to the wrong questions".

Although the objective of this step is clear, determining a goal for a data mining application may not necessarily be simple or straight-forward. A challenging question that is faced by the DM team is "How do you determine the 'correct' / 'valid' / 'best' objective or goal for a data mining project?"

The fact that data mining is exploratory in nature makes it difficult. In a typical engineering project, the objectives/goals of the project are straight-forward and easily determined. In a data mining experiment, due to its inherent exploratory nature, the outcomes are typically unknown. How to determine the suitable goals for a data mining experiment can be a challenging task.

We observe three main approaches taken in practice to determine goals for a DM initiative:

- *Using domain knowledge to determine goals for the data mining experiment:* The use of domain knowledge to determine suitable goals for data mining is popular. For instance in (Chapman et al., 2000), asking questions such as "What is the motivation of the project?" (see p35 in Chapman et al., 2000) refers directly to finding out what interests domain experts and therefore considered useful and interesting. This will assist in aligning the project goals with what the domain experts consider to be useful, interesting and relevant. Hence it reduces the risk and increases the possibility to mine knowledge that is useful, interesting and relevant. This is the approach typically considered first in a data mining experiment in an industry context.

This was the approach taken by the authors in (Fernando et al., 2008). A key goal of the DM initiative was to determine the correlation between production and export levels of tea in Sri Lanka.

- *Exploring the data to determine goals:* In this approach, it is more data driven, whereby existing data is examined at first (types of data, available data etc.) and understood. Initially, goals may be broad (i.e. "What kind of analysis is possible?"). Typically a preliminary analysis takes into consideration, domain knowledge, available data and data mining techniques, to determine a suitable goal(s). Careful analysis of preliminary results enables to find clues on the promising direction to be taken before further mining is performed. The goals are selected with the aim of increasing the possibility of discovering interesting knowledge.

At the preliminary stage, if a particular path of analysis is not promising, different techniques may be considered. This may result in a more-or-less trial-and-error approach, however, without involving much effort as it is a preliminary analysis. The DM team gets a "look-and-feel" of the possibilities with data mining before focusing/selecting a particular direction/goal.

This approach was taken by authors in (Tissera et al., 2006). In this instance, the authors had access to a data set from an educational institute containing student information, including student performances of various courses in the undergraduate degree programmes. A

particular goal/aim did not exist initially. However, after examining the data, data mining techniques and tools, and considering the domain context, the authors determined that finding related courses in the undergraduate programme as a suitable goal/direction for the project based on student performance.

- *Blind search*: In this approach, certain data mining algorithms are executed on data sets with the hope of finding interesting “nuggets” of knowledge, without having an initial pre-conceived goal/direction. At the evaluation phase, the patterns returned by the different algorithms are considered for relevancy and usefulness. However, this approach is more suitable in a research context than an industry setting – where typically a project’s resources/goals need strict justification prior to initiating. In (Chapman et al., 2000), this aspect is described as follows: “Blind search is not necessarily useless but a more directed search towards business objectives is preferable”.

A hybrid of the above-mentioned approaches may be considered in determining a suitable goal for data mining.

Other challenges faced in data mining goal determination stage include:

- *How do we know the selected goals are “valid”?*: For instance, we may like to find “patterns” or “interesting” clusters of customers. For us to gain such knowledge, such patterns must exist. It is difficult to validate the existence of such patterns prior to embarking on a DM experiment. Otherwise, we may even try to find patterns that do not exist (i.e. patterns which are simply a product of random fluctuations, rather than representing any underlying structure (Hand, 1998)). This process is also known as “data dredging” or “fishing” in statistics (Hand, 1998). However, to complicate matters further, not discovering a pattern using a certain DM technique doesn’t necessarily mean that such a pattern/correlation does not exist!
- *How do we ensure that “good” goals are not missed*: During the selection of goals, the data mining team focuses on selected goals. These goals were selected based on the DM team’s judgement. It is possible that there exist hidden patterns and knowledge undiscovered in the vast data sets. How can we ensure that all relevant “knowledge” is discovered via “valid” goal selection?

Today’s state-of-art mining methodologies, selects goals for a data mining application based on judgement of the data mining team considering domain knowledge, data available and data mining techniques. All KDDM process models emphasise the iterative nature of the process which a data mining application is conducted. Typically, goals are selected, an experiment is conducted, based on results at each stage, a step is revisited or moves to the next step. The iterative nature of KDDM process models allows retracting and considering different approaches/paths (goals, techniques and methods) in conducting a data mining experiment as a way to address this uncertainty. This approach, however is not optimal and results in a trial-and-error process which is resource-intensive and risky with no guarantee of favourable results. Approaches to minimise unsuccessful attempts and provide certain guarantees would be highly beneficial.

Answers to questions such as:

- How to select valid goals in a data mining project? What is a suitable goal?
- How to ensure that valid goals are not missed when conducting a data mining application? When do we stop looking? How do we know nothing interesting exists in the deluge of this data?

are still challenging open research problems.

**Data Understanding:** The major goal of this step is to understand the data sources, data parameters and quality of data. Data sets are selected for analysis in later steps of the KDDM process.

The major challenge at this stage includes:

- How to determine whether the relevant data sets and parameters are selected for the data mining tasks?

Since data mining considers patterns which are hidden or unknown, how do we ensure that the data parameters omitted in the data understanding and selection process does not result in ignoring parameters which have the possibility to affect the outcome of the DM experiment.

We observe that both domain expertise and technical expertise are required in order to determine the relevant data sets and also determine whether certain types of data are useful for mining. As in the previous stage, answers to these questions depend on the experience and judgement of the data mining team conducting the experiment. The iterative nature of conducting a data mining experiment enables to re-consider and re-do a step. However, as previously stated this leads to a trial-and-error approach taking considerable time and effort.

Another challenge is completeness: Whether successful or not in the initial data selection and data mining steps, how do we know we haven't missed any data parameters or considered some data source (maybe even external to the organisation) that, if considered, may have resulted in discovery of valuable knowledge?

Challenging questions such as

- How to determine the 'ideal' data set and parameters to satisfy a data mining goal? How do we ensure that all relevant data parameters and data set are considered?
- How do we ensure that data sets with the potential to discover 'useful' patterns are considered and not missed/omitted? What is the best way to identify which data parameters and data sets are relevant and will enable discovery of 'novel' nuggets of knowledge?

are still open research issues.

**Data Preparation and Identification of DM Technology:** This pertains to preparation of data sets to perform data mining tasks and finalising of the data mining techniques and tools.

Selection of data mining techniques, algorithms and tools is one of the crucial and significant questions that need to be decided by the DM team. There are a number of possible data mining techniques such as classification, clustering, association rule mining, machine learning, neural networks, regression and others. Also, a plethora of data mining algorithms exists. The DM team based on their judgement selects the data mining techniques and algorithms to apply. The question of

- What DM techniques and algorithms to apply that will result in useful and relevant knowledge?

is one of the fundamental open research questions in Data Mining Applications.

Preprocessing is a popularly used term to mean the preparation of data for data mining tasks. Preprocessing takes a significant effort, typically most of the effort in a data mining experiment (see Kurgan & Musilek, 2006 for a discussion on relative efforts spent of specific steps in the KDDM process). There are several reasons; the main reason being the fact that data being collected by enterprises and in different domains is not originally planned for analysis and therefore often contains missing, redundant, inconsistent, outdated and sometimes erroneous data.

As stated in (Edelstein, 1998), data quality is highly important in a data mining application. "GIGO (garbage in, garbage out) is very applicable to data mining. If you want good

models, you need good data.” (Edelstein, 1998). The DM team makes a number of decisions at preprocessing stage which has a significant impact on the results of the DM tasks (for instance, sampling techniques, quantitative vs. qualitative data, omission/reduction of dimensions, data imbalance and others). We observe a number of research efforts addressing these issues in literature (imbalance data sets - e.g. (Nguyen et al., 2008); missing values - e.g. (Farhangfar et al., 2007), and others).

**Data Mining:** In this step, data mining techniques are applied. The primary goals of data mining in practice tend to be prediction and description (Fayyad et al., 1996b). The DM techniques fit models to the data set. These models either are used to describe the characteristics of the data set (descriptive – for e.g., clustering), with the hope of discovering novel and useful patterns and trends in data; or used to predict future or unknown values of other variables of interest (predictive – for e.g., regression).

The DM team typically iterates through process to find a best-fit model for the data by adjusting various parameters of the model (e.g. threshold values). It is possible that the DM team may even modify the DM technique itself in order to determine a best-fit model. For instance in (Julisch & Dacier, 2002), the Attribute Oriented Induction technique is modified to gain favourable results in prediction. In order for the DM team to modify an existing or develop a new DM technique (such as SBA scoring function in (Ma et al., 2000)), requires an in-depth understanding of the DM technique.

A number of challenging issues may be faced by the DM team at this stage as outlined below:

- What are the optimal values in the model that would provide the ‘best-fit’ for the data set?
- Are there any other models that provide a better fit (in terms of accuracy, performance, etc.)?

At present, the DM team makes a judgement call when faced with these challenging questions, typically after applying DM techniques iteratively to the data set.

**Evaluation:** At the evaluation step of the data mining experiment, the results of the particular data mining tasks are visualised and interpreted. Although DM tasks reveal patterns and relationships, this by itself is not sufficient. Domain knowledge and data mining expertise is required to interpret, validate and identify interesting and significant patterns. The DM team incorporates domain expertise and data mining expertise in evaluating and visualising models in order to identify interesting patterns and trends.

In addition to evaluation of results, a number of significant decisions are considered at this stage with respect to the progress of the project:

- *Iterate?:* As discussed earlier, DM tasks may be iterated adjusting the model. Also, during evaluation, it is possible to discover a novel pattern or promising trend. Further investigation may be required for validation and verification purposes. This may be considered as a part of the existing project or as an entirely different DM initiative.

For instance, when conducting the DM experiment discussed in (Fernando et al., 2008), a spike in tea prices is observed in 1996. To determine the reasons causing this spike requires further investigation and can be considered as an entirely different DM project. Similarly, in (Tissera et al., 2006), when mining for correlated courses, the results revealed that the capstone project course did not demonstrate a strong relationship to any other course. The reasons for this fact can be investigated as a part of the existing DM project or fresh DM initiative.

- *Conclude project?:* The decision to conclude the project may be considered at this stage of KDDM process due to a number of reasons:
  - If satisfactory results are achieved in applying data mining, the decision to conclude the project can be considered at this stage. Note that lack of new patterns or correlations can itself be an insight to the non-existence of dependencies in parameters in the data set.
  - If unsatisfactory results are achieved, or due to the lack further resource commitment (personnel, funds etc.) for future analysis and minimal possibility for novel discoveries, the DM project may be concluded.

**Knowledge Consolidation and Deployment:** At this stage of the KDDM process, the results are published and the main stakeholders of the DM project are informed. Also, strategies to incorporate and exploit the discovered knowledge is considered. The implementation of the discovered knowledge and monitoring is also taken into consideration.

It is important to ensure that the support and “buy-in” of the project’s stakeholders are maintained throughout the project. This aspect will assist in speedy acceptance and action on results of the project. A DM experiment conducted in isolation will require more convincing to build trust and acceptance of results prior to deployment.

#### 4. Discussion

It is evident that a number of challenges and issues are faced during data mining applications. Some of them include:

- How do we determine goals for a DM application?
- How do we select the data that will result in achieving the goal?
- What type of DM technique(s) should be considered?
- How complete is the exploration? Did we miss any useful “nuggets“?

The responses to these challenging questions have a major impact on the direction taken and results obtained in a DM initiative. At present, there aren’t definitive answers or processes to determine answers to these challenging questions. It is more-or-less a judgement made by the DM team.

Today, to address this uncertainty, data mining applications are conducted in an iterative manner. This process allows the DM team to develop a better understanding of the data, domain and data mining techniques and gain insights with each iteration. The insights gained assists in decision making. The iterative nature of data mining application is reflected in the KDDM process models as well. However, a disadvantage of this approach is that the iterative nature results in a trial-and-error process to data mining applications, which is resource-intensive. Finding approaches that enable a more predictive, controlled, and risk-averse methodology to data mining applications is a challenge that remains to the DM research community. The authors believe that such methodologies require addressing the challenging questions presented in section 3.

At the beginning of the chapter, the authors described Data Mining Application as an “experiment“ where the goal is discovery of knowledge from large data sets. The analogy between a research experiment and application of DM is evident. Similar to a scientific experiment, where researchers explore the unknown, data mining applications require the DM team to explore for unknown knowledge hidden in the vast data sets. This analogy assists us in determining conducive environments for conducting a DM experiment. Based

on the discussion thus far, we present some recommendations for conducting a data mining experiment.

- *Expertise required in a DM project team:* We believe that selecting the appropriate team with relevant expertise is crucial as human judgement plays a major role in a data mining application. The team must include domain expertise and a high-level of technical expertise. Domain knowledge will play a crucial role throughout the DM application especially in goal determination and evaluation of results. Also, technical expertise, especially relevant to data mining techniques and algorithms, is required. As demonstrated in many DM applications (for instance, Julisch & Dacier, 2002; Ma et al., 2000; Tissera et al., 2006), the DM models are tweaked, modified and even new techniques are developed to enable best-fit models that describe the data. This requires an extensive and in-depth understanding of the data mining models.
- *Management stakeholders perspective of a DM project:* The authors believe that DM project's management stakeholders' understanding of the exploratory nature and uncertainties associated with a DM initiative is beneficial. This is especially true in an industrial context, where typical projects are deterministic. This fact will enable the project's managing stakeholders to effectively support/facilitate a DM project and manage expectations/outcomes of a DM initiative.

## 5. Conclusion

Data mining shows promise in enabling the discovery of hidden, previously unknown knowledge from vast data sets. Today, proliferation of databases technology has made it possible to collect and access such data sets in many different domains. Application of data mining to these large data sets has the possibility to unravel knowledge that can impact many different fields of study in society.

To achieve such proliferation in data mining applications, we believe a consistent, risk-averse and predictable methodology is required. Today, a number of Knowledge Discovery and Data Mining (KDDM) process models are proposed in literature with the aim of providing structure, control and standard methodology in applying data mining. KDDM process models outline the fundamental steps (executed iteratively) in applying data mining covering the entire lifecycle from goal determination to deployment of the discovered knowledge. Although KDDM process models provide a high-level structure to conduct a data mining application, following a KDDM process model by itself does not guarantee success. The exploratory nature of data mining makes DM application a risky, resource-intensive and an uncertain endeavour with no guarantee of success.

To address this issue, we believe some challenging questions need to be answered by the data mining research community. This chapter brings to attention some of the fundamental questions that need to be addressed for a more predictive, less-risky and controlled approach to data mining. We believe that significant strides in resolving these fundamental questions will enable a proliferation of data mining applications in many different application domains.

## 6. References

Adriaans, P. & Zantinge, D. (1996) *Data Mining*, Addison-Wesley.

- Agrawal, R., Imieliski, T. and Swami, A. (1993), Mining association rules between sets of items in large databases, *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 207-216, Washington, D.C., USA.
- Anand, S. S. & Buchner, A. G. (1998) *Decision Support Using Data Mining*, Trans-Atlantic Publications.
- Antonie, M.-L.; Za'iane, O. R. & Coman, A. (2001) Application of data mining techniques for medical image classification, *Second Intl. Workshop on Multimedia Data Mining (MDM/KDD'2001)*, pp. 94-101, San Francisco, USA.
- Au, W.-H.; Chan, K. C. C.; Wong, A. K. C. & Wang, Y. (2005) Attribute Clustering for Grouping, Selection, and Classification of Gene Expression Data, *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 2, 2, 83 - 101.
- Berry, M. J. A. & Linoff, G. (1997) *Data Mining Techniques: For Marketing, Sales, and Customer Support*: Wiley, 0471179809.
- Cabena, P.; Hadjinian P.; Stadler R.; Verhees, J. & Zanasi, A. (1998) *Discovering data mining: From Concept to Implementation*, Prentice-Hall, Inc.
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. (2000) CRISP-DM 1.0: Step by step data mining guide, CRISP-DM Consortium, pp. 79.
- Chun, T., Aidong, Z. & Jian, P. (2003) Mining phenotypes and informative genes from gene expression data, *Proceedings of the ninth ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, pp. 655-660, Washington, D.C.
- Cios, K. J.; Teresinska, A., Konieczna, S.; Potocka, J. & Sharma, S. (2000), A knowledge discovery approach to diagnosing myocardial perfusion, *IEEE Engineering in Medicine and Biology Magazine*, 19, 4, 17-25.
- Daxin, J.; Jian, P.; Murali, R.; Chun, T. & Aidong, Z. (2004) Mining coherent gene clusters from gene-sample-time microarray data, *Proceedings of the tenth ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, pp.430-439, Seattle, WA, USA.
- Druzdzel, M. & Glymour, C. (1994) Application of the TETRAD II program to the study of student retention in U.S. colleges, *Proceedings of the AAAI-94 Workshop on Knowledge Discovery in Databases (KDD-94)*, pp. 419-430, Seattle, WA.
- Edelstein, H. (1998) Data Mining – Let's Get Practical, *DB2 Magazine*, 3, 2.
- Ester, M.; Kriegl, H.-P. & Schubert, M. (2002), Web site mining: a new way to spot competitors, customers and suppliers in the world wide web, *Proceedings of the eighth ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, pp. 249-258, Edmonton, Alberta, Canada.
- Farhangfar, A.; Kurgan, L. A. & Pedrycz, W. (2007) A Novel Framework for Imputation of Missing Values in Databases, *IEEE Transactions on Systems, Man and Cybernetics, Part A*, 37, 5, 692-709.
- Fayyad, U.; Piatetsky-Shapiro, G. & Smyth, P. (1996a), The KDD Process for Extracting Useful Knowledge from Volumes of Data, *Communications of the ACM*, 39,11, 27-34.
- Fayyad, U. M.; Piatetsky-Shapiro G. & Smyth P. (1996b), From data mining to knowledge discovery in databases, *AI Magazine*, 17, 3, 37-54.
- Fayyad, U.; Haussler D. & Stolorz P. (1996c), Mining scientific data, *Communications of the ACM*, 39, 11, 51-57.
- Feelders, A.; Daniels, H. & Holsheimer, M. (2000), Methodological and practical aspects of data mining, *Information & Management*, 37, 271-281.

- Fernando, H. C.; Tissera, W. M. R. & Athauda, R. I. (2008), Gaining Insights to the Tea Industry of Sri Lanka using Data Mining, *Proceedings of the International MultiConference of Engineers and Computer Scientists (IMECS 2008)*, pp. 462-467, Hong Kong.
- Fukuda, T.; Morimoto, Y.; Morishita S. & Tokuyam, T. (1996), Data Mining Using Two-Dimensional Optimized Association Rules: Scheme, Algorithms, and Visualization, *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*, pp. 13-23, Montreal, Canada.
- Gartner Group (1995), Gartner Group Advanced Technologies and Applications Research Note <http://www.gartner.com>
- Guha, S.; Rastogi, R. & Shim, K. (1998), CURE: An Efficient Clustering Algorithm for Large Databases", *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*, pp. 73-84. Seattle, WA, USA.
- Haglin, D.; Roiger, R.; Hakkila, J. & Giblin, T. (2005), A tool for public analysis of scientific data, *Data Science Journal*, 4, 30, 39-53.
- Han, J. & Fu, Y. (2005), Discovery of Multiple-Level Association Rules from Large Databases, *Proceedings of the 21st International Conference on Very Large Data Bases (VLDB 95)*, pp. 420-431, Zurich, Switzerland.
- Han, J. & Kamber, M. (2001), *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers.
- Hand, J. D. (1998), Data Mining: Statistics and More?, *The American Statistician*, 52, 2, 112-118.
- Hand, D. J.; Mannila, H. & Smyth, P. (2001), *Principles of Data Mining*: MIT Press.
- Hearst, M. A. (1999), Untangling text data mining, *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics Association for Computational Linguistics*, pp. 3-10., College Park, Maryland, USA.
- Julisch, K. & Dacier, M. (2002), Mining Intrusion Detection Alarms for Actionable Knowledge, *The 8th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pp. 366-375, Edmonton, Alberta, Canada.
- Karim, K. H. (2001), Exploring data mining implementation, *Communications of the ACM*, 44, 7, 87-93.
- Klösgen, W. & Zytkow, J.M. (2002). The knowledge discovery process, In : *Handbook of data mining and knowledge discovery*, Klösgen, W. & Zytkow, J.M. (Ed.), 10-21, Oxford University Press, 0-19-511831-6, New York.
- Kurgan, L. A. & Musilek P. (2006), A Survey of Knowledge Discovery and Data Mining Process Models, *The Knowledge Engineering Review*, 21, 1, 1-24.
- Last, M.; Friedman, M. & Kandel, A. (2003), The data mining approach to automated software testing, *Proceedings of the ninth ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, pp. 388-396, Washington, D.C., USA.
- Lee, W.; Stolfo, S. J.; Chan, P. K.; Eskin, E.; Fan, W.; Miller, M.; Hershkop, S. & Zhang, J. (2001), Real time data mining-based intrusion detection, *DARPA Information Survivability Conference & Exposition II, 2001 (DISCEX '01)*, pp. 89-100. Anaheim, CA, USA.
- Li, S.-T & Kuo, S.-C. (2008), Knowledge discovery in financial investment for forecasting and trading strategy through wavelet-based SOM networks, *Expert Systems with Applications*, 34, 935-951.
- Liang, Y. & Kelemen, A. (2002), Mining heterogeneous gene expression data with time lagged recurrent neural networks, *Proceedings of the eighth ACM SIGKDD*

- international conference on Knowledge Discovery and Data Mining*, pp. 415-421, Edmonton, Alberta, Canada.
- Kitts, B.; Freed, D. & Vrieze, M. (2000), Cross-sell: a fast promotion-tunable customer-item recommendation method based on conditionally independent probabilities, *Proceedings of the International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pp. 437-446 Boston, Massachusetts, USA.
- Nguyen, T. H.; Foitong, S.; Udomthanapong, S. & Pinngern, O. (2008) Effects of Distance between Classes and Training Datasets Size to the Performance of XCS: Case of Imbalance Datasets, *Proceedings of the International MultiConference of Engineers and Computer Scientists 2008 (IMECS 2008)* pp. 468-473, Hong Kong.
- Ma, Y.; Liu, B.; Wong, C. K.; Yu, P. S. & Lee, S. M. (2000), Targeting the right students using data mining, *Proceedings of the sixth ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, pp. 457-464, Boston, Massachusetts, USA.
- Meek, C.; Chickering, D.M. & Heckerman, D. (2002), Autoregressive Tree Models for Time-Series Analysis, *Proceedings of the Second SIAM International Conference on Data Mining*, pp. 229-244, Arlington, VA, USA.
- Rosset, S.; Murad, U.; Neumann, E.; Idan, Y. & Pinkas, G. (1999), Discovery of fraud rules for telecommunications - challenges and solutions, *Proceedings of the International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pp. 409-413, San Diego, California, USA.
- SAS Institute (2003), *Data mining Using SAS Enterprise Miner*, SAS Publishing, 1590471903.
- Senator, T. E.; Goldberg, H. G.; Wooton, J.; Cottini, M. A.; Khan, A. F. U.; Klinger; C. D., Llamas; W. M.; Marrone M. P. & Wong, R. W. H. (1995), The Financial Crimes Enforcement Network AI System (FAIS): Identifying Potential Money Laundering from Reports of Large Cash Transactions, *AI Magazine*, 16, 21-39.
- Tissera, W. M. R.; Athauda, R. I. & Fernando, H. C. (2006), Discovery of Strongly Related Subjects in the Undergraduate Syllabi using Data Mining, *International Conference on Information and Automation (ICIA 2006)*, pp. 57-62, Colombo, Sri Lanka.
- Wang H. & Weigend, A. S. (2004), Data mining for financial decision making, *Decision Support Systems*, 37, 4, 457-460.
- Weir N.; Fayyad, U. M. & Djorgovski, S. (1995), Automated star/galaxy classification for digitized POSS-II, *Astronomical Journal*, 109, 2401-2412.
- Wu X.; Kumar V., Quinlan J. R., Yang J. G. Q., Motoda H., McLachlan G. J., Ng A., Liu B., Yu P. S., Zhou Z.-H., Steinbach M., Hand D. J. & Steinberg D. (2008), Top 10 algorithms in data mining, *Knowledge and Information Systems*, 14, 1, 1-37.
- Yan, L.; Verbel D. & Olivier, S. (2004), Predicting prostate cancer recurrence via maximizing the concordance index, *Proceedings of the tenth ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, pp. 479-485, Seattle, WA, USA.
- Yu, L. T. H.; Chung, F-I; Chan, S. C. F. & Yuen, S.M.C. (2004), Using emerging pattern based projected clustering and gene expression data for cancer detection, *Proceedings of the Second Conference on Asia-Pacific Bioinformatics*, Vol. 29, pp. 75-84, Dunedin, New Zealand.

# Mining Spatio-Temporal Datasets: Relevance, Challenges and Current Research Directions

M-Tahar Kechadi<sup>1</sup>, Michela Bertolotto<sup>1</sup>,  
Filomena Ferrucci<sup>2</sup> and Sergio Di Martino<sup>2</sup>

<sup>1</sup>*School of Computer Science and Informatics, Univeristy College Dublin*

<sup>2</sup>*Dipartimento di Matematica e Informatica Università di Salerno,*

<sup>1</sup>*Ireland*

<sup>2</sup>*Italy*

## 1. Introduction

Spatio-temporal data usually records the states over time of an object, an event or a position in space. Spatio-temporal data can be found in several application fields, such as traffic management, environment monitoring, weather forecast, etc. In the past, huge effort was devoted to spatial data representation and manipulation with particular focus on its visualisation. More recently, the interest of many users has shifted from static views of geospatial phenomena, which capture its “spatiality” only, to more advanced means of discovering dynamic relationships among the patterns and events contained in the data as well as understanding the changes occurring in spatial data over time.

Spatio-temporal datasets present several characteristics that distinguish them from other datasets. Usually, they carry distance and/or topological information, organised as multidimensional spatial and temporal indexing structures. The access to these structures is done through special methods, which generally require spatial and temporal knowledge representation, geometric and temporal computation, as well as spatial and temporal reasoning. Until recently, the research in spatial and temporal data handling has been mostly done separately. The research in the spatial domain has focussed on supporting the modelling and querying along spatial dimensions of objects/patterns in the datasets. On the other hand, the research in the temporal domain has focussed on extending the knowledge about the current state of the system governed by the temporal data. However, spatial and temporal aspects of the same data should be studied in conjunction as they are often closely related and models that integrate the two can be beneficial to many important applications.

Indeed the amount of available spatio-temporal datasets is growing at exponential speed and it is becoming impossible for humans to effectively analyse and process. Suitable techniques that incorporate human expertise are required. Data mining techniques have been identified as effective in several application domains. In this chapter we discuss the application of data mining techniques to effectively analyse very large spatio-temporal datasets.

Spatio-temporal data mining is an emerging field that encompasses techniques for discovering useful spatial and temporal relationships or patterns that are not explicitly stored in spatio-temporal datasets. Usually these techniques have to deal with complex objects with spatial, temporal and other attributes. Both spatial and temporal dimensions add substantial complexity to the data mining process. Following the above mentioned

separation traditionally applied to the analysis of spatial and temporal dimensions, spatial data mining (SDM) and temporal data mining (TDM) have received much attention independently within the KDD (Knowledge Discovery in Databases) research community. Motivations that have kept these fields separate include:

- **Complexity:** the investigation of spatio-temporal relations complicates the data mining process. The existing spatial or temporal techniques are not suitable for such additional constraints in terms of data types, data representation and structures. Therefore, the exploration of efficient techniques for spatio-temporal data becomes a necessity.
- **Data Models:** the absence of efficient space-time data models makes it difficult for the development of data mining techniques that deal with space and time simultaneously. Most existing data mining techniques applied to spatio-temporal datasets use very simple representations of the objects and their relationships, which usually based on spatial or temporal, but not both. The first model that integrates the time and space was proposed in (Peuquet et al., 1995). Since 1995 other models have been proposed in (Yuan, 1997). However, these models are too generic to handle special cases and implicit complex relationships, as it is often the case in spatial and temporal real-world data mining applications.
- **Application Domain:** An application is usually classified as spatial or temporal depending on the target problem at hand and how the datasets are collected. For instance, consider data describing three regions in Ireland, collected in three successive years. Each region is divided into grids of cells of the same dimensions. The goal is to predict the spatial distribution of agriculturally beneficial herbs e.g. Birdsfoot trefoil. The collected data, which consists of a certain number of attributes (soil pH, N, P, K status, herb abundance, distance from field boundary structure, e.g., hedgerow, etc.), chosen according to the domain knowledge, records the values of these attributes in each cell of the three regions. In this example, even though the overall goal is prediction, the time dimension is not taken into account in the model, as there are only three timestamps. Therefore, more effort is put into the modelling of the space dimensions and the data is recorded accordingly.

The development of efficient techniques for combined spatial and temporal data mining is an open and challenging issue within the research community. In this chapter we investigate this topic and discuss current trends in the area. In particular, given the visual aspect of spatial data, we discuss how visualisation techniques have been applied to support the spatio-temporal data analysis and mining process.

The remained of the chapter is organised as follows. In Section 2 and 3 we explain spatial and temporal data mining, respectively, the challenges of the corresponding mining tasks, review existing approaches, and discuss the main research directions. Section 4 is dedicated to on-going efforts to develop data mining techniques that take into account both spatial and temporal aspects of data. In Section 5 visual techniques applied in support of the mining process are discussed, while in Section 6 we present some innovative techniques for effectively mining very large spatio-temporal datasets as part of our work in this area. Finally, Section 7 concludes summarising new research directions and challenges.

## 2. Spatial data mining

Spatial data is characterised by spatial location attributes or dimensions. These spatial attributes are usually stored as coordinates and topology. Moreover spatial data contains

also non-spatial attributes. The spatial dimensions and their features add another level of complexity compared to relational data in terms of both the mining efficiency and the complexity of possible patterns that can be extracted from spatial datasets (Hinnenburg and Keim, 1998), (Roddick and Lees, 2001), and (Shekhar et al., 2001). The main reasons include:

1. The attributes of neighbouring patterns may have significant influence on a pattern and should also be considered.
2. Classical analytical methods for spatial data were introduced when the technology for collecting such data was very expensive and not powerful and, consequently, the available datasets were small and few. Nowadays, massive amounts of spatial data are collected daily at relatively low cost and classical methods cannot cope with them.
3. While non-spatial datasets are mainly composed of discrete objects stored in databases with well-defined relationships, spatial objects are embedded in a continuous space and characterised by implicit topological, distance and directional relationships (Bedard et al., 2001).

Much research has been dedicated to addressing these issues. Techniques that have been proposed include spatial classification, spatial association, spatial clustering, spatial outlier analysis and spatial prediction. Approaches based on spatial classification group spatial objects into categories based on distance, direction and connectivity relationships among them. For example, (Koperski et al., 1998) introduced spatial buffers to classify objects based on attribute similarity measures and proximity. (Ester et al., 2001) generalised this approach to other different spatial relationships. Spatial association rules are rules that rely on spatial predicates. The work by (Koperski et al., 1996) and (Yoo et al., 2005) provide examples of methods based on the spatial association approach. Methods for spatial clustering borrow heuristics proposed for general clustering algorithms to define meaningful groupings within the input data. Traditional techniques such as k-means and expectation maximisation can take distance relationships into account and can therefore be applied to spatial data in a straightforward way (Han et al., 2001). Alternative approaches include: hierarchical, grid-based, constraint-based, and density-based methods. Spatial outliers are spatial objects whose non-spatial attribute values are inconsistent with other objects, which are in the same local (spatial) neighbourhood. (Shekar et al., 2003) proposed a method for detecting spatial outliers efficiently. (Ng, 2001) adopted a similar approach to identify unusual trajectories based on endpoints, speed and geometry of the trajectories.

Prediction methods combine classification with inductive learning and/or artificial neural networks approaches to extract information from different types of spatial datasets. Examples include techniques applied to topographic maps (Malerba et al., 2001), soil-landscape maps (Qi & Zhou, 2003) and remotely sensed imagery (Gopal et al., 2001).

As previously mentioned, many spatial datasets also have an associated temporal dimension and, therefore, are referred to as spatio-temporal. Integrating time introduces additional complexity to the mining and knowledge discovery process. Simply adding another dimension does not represent an effective solution. Indeed, time and space have different semantics. We will develop spatio-temporal data mining in section 4. The next section is devoted to temporal data mining.

### 3. Temporal data mining

In general, temporal data mining techniques are designed for mining large sequential datasets. A set of data is said to be sequential if its data is ordered with respect to some

index, such as time. For instance, time series datasets are a common class of sequential data, which is recorded according to the index time. Moreover, there are many other sequential data that are not depending on the time. These include protein and gene sequences, Web click streams, sequences of moves in games such as Go and chess, alarms generation in telecommunication networks, etc. The overall goal is to discover sequential relationships or patterns that are implicitly present in the data. These sequences (sequential patterns) can be very useful for many purposes, for example, prediction of future event sequences. Consider a telecommunication company, the sequential patterns can be used for customer churn, marketing new products, pricing, and many others tasks.

The time series problems can behave in four different way: linearly, stationary, periodically, or randomly. A time series application can be represented as a linear problem when the future observation can be a linear function of the past observations. A time series can be stationary when it has a constant mean and variance. For non-stationary time series, the future observations cannot be foreseen, as they are very difficult to model. Time series can be periodic; displaying dominant periodic components with regular periodic variations. Finally, time series can be a random noise problem, which means that there can be a random noise included in some parts or the entire frequency spectrum of the time series.

There are several techniques implemented to handle time series applications. For instance, the simple moving average and exponential moving average techniques are used to deal with linear and stationary time series problems, the simple regression methods, auto regressive and auto regressive average deal with non-stationary time series problems, and decomposition methods deal with seasonal time series problems. One difficulty with regression techniques is that the correlation between the component variables, which affects the observation demand, is not stationary but depends on spatial-temporal attributes. Therefore, they are not capable of tackling this chronological disparity. Moreover, most of these techniques are problem-dependent, which means that while they return reasonably good solutions to the application at hand, they are not suitable for other types of temporal datasets.

### 3.1 Output patterns

Temporal data mining have been heavily investigated mainly for **prediction**, **classification**, and **clustering**. The **prediction** task in the time-series applications deals with forecasting future values of the time series based on its current and past values. Usually, prediction requires an efficient predictive model for the data. By model we mean an abstract representation of the data. For example neural network or Markov models have been widely used for prediction in sequence classification and forecasting applications (Rashid et al., 2006).

**Classification** assumes that some classes or categories have already been predefined. The main objective is to automatically identify for each input sequence its corresponding class or category. There are several sequence classification applications. These include handwriting recognition (Fitzgerald et al. 2004), gene sequences, speech recognition, currency exchange rates, stock market prices, etc. The temporal data mining technique for classification task are divided into two categories (Laxman et al., 2006): *model-based* methods and *pattern-based* methods. Pattern-based methods use a database of prototype feature sequences. Each class is represented by a set of prototype feature sequences (Ewens et al., 2001). For any given input sequence, the classifier searches over all prototypes looking for the closest (or most

similar) to the features of the new input. The model-based methods are techniques that use some powerful existing models such as Hidden Markov models, neural networks, support vector machines, etc. Usually these models consist of two phases: learning phase and testing phase. During the learning phase the model is trained on examples of each pattern class. The model assigns to the new pattern a class that contains the most likely pattern to generate it.

Unlike classification, **clustering** does not assume the class labels. Clustering groups the sequences together based on their similarity. Basically, sequences that are similar are grouped together and those that are dissimilar are assigned to different groups or clusters. Clustering is particularly interesting as it provides a dynamic mechanism for finding some structures (or clusters) in large datasets without any assumption about the number of the groups (clusters) in advance.

### 3.2 An example

In this section we consider the task of electric load forecasting for large and complex buildings. A correct estimation of the energy in this case is crucial as it can result in substantial savings for a power system. Once modelled correctly, it allows planning and/or designing of new future plants, providing security and reliability, and savings in the operational cost of a power system. It is apparent that there are relationships between the energy load and factors affecting it, yet these relations have not been clearly defined and understood. This problem is complex and, in order to learn something from this historical dataset, a very good data model and attributes' selection are required. For many years this application has been classified as one of the important class of time-series. We will see later in this chapter that this can also be modelled as a spatio-temporal data mining application when it involves various regions/locations of the world.

In this section, we consider the datasets that reflects the behaviour of the electricity supply in the Republic of Ireland. The main recorded attributes are the load, temperature, cloud rate, wind speed and humidity at 15mn intervals. This is a typical time-series application; each sequence of recorded data represents a value of a particular feature; these features are observed at different time intervals.

The choice of a data model is a very crucial and complex task in data mining. Not only should the model represent the data precisely but it should also be appropriate for the mining technique used. For instance the data inputs of a neural network technique are different from the inputs of a support vector machine or a hidden Markov model. We usually divide the dimensions into two categories: primary and secondary. Primary dimensions are the main dimensions that characterise the data itself. The secondary dimensions are informative but they can play a huge role when associated with the inputs of a given mining technique. The difficulty here is that there is no general rule for how to select appropriate secondary dimensions (inputs) for a given mining process. The selection depends largely on the experience and the expertise of the user within that specific domain or application.

In the case of energy load forecasting, we can define four types of attributes:

1. **The time and seasonal attributes:** As for any time series application, the time is one of the most important dimensions in this task. However, the time on its own is not enough. The variation in daily load profile is mostly affected by the localised weather effects and seasonal changes, which introduces weather patterns. Therefore, it is

- imperative to include time information such as the time of the day, the day of the week, and the season of the year in order to model appropriately the forecasting behaviour.
2. **Direct and indirect weather attributes:** Direct weather attributes play a key role in the energy load model. These attributes include temperature, cloud rate, wind speed, wind direction, humidity, rainfall, etc. Along with time dimensions, all these attributes constitute principal dimensions of the model. Indirect or secondary attributes provide extra information about the application. These include relative change in temperature and relative change in load linked with the day, month or season of the year.
  3. **The status of the day:** The load consumption depends on other factors such as special days; weekends, holidays, and other special events. Therefore these external factors should be included in the model. Some different behaviour was also noticed on the days before and after weekends and holidays. They are also treated as another different status of the day.
  4. **Historical absolute/relative change in the load:** With this attribute one wants to model relationships between the attributes of a sequence pattern. In other words, one wants to identify the relationship between two consecutive days or the change in the load between two consecutive days. This notion should be extended to other attributes such as cloud rate, humidity, wind speed, temperature, etc.

Generally, some of the attributes involved have to be normalised in order to avoid errors in prediction due to the increase in electricity load consumption that is necessary for the economy growth. Moreover, the economy growth rate is needed for the final calculations of the predicted values; this should be part of the data pre-processing phase. As mentioned above, there are some attributes, which are not primary but necessary for a time-series application, such as energy load forecasting. These are called secondary or endogenous attributes. Example of daily temperature collected for January 1997 and 1998, shown in Figure (1) illustrates the difference between primary (exogenous) attribute and its corresponding secondary \*endogenous) attribute. For instance in Figure 1a, it is difficult to see the patterns between the temperature in January 1997 and 1998, while in Figure 1b we can easily see that the change in temperature through the two months presents some patterns. The goal is to exploit these changes in temperature in order to extract useful sequences (patterns) and use them for future predictions.

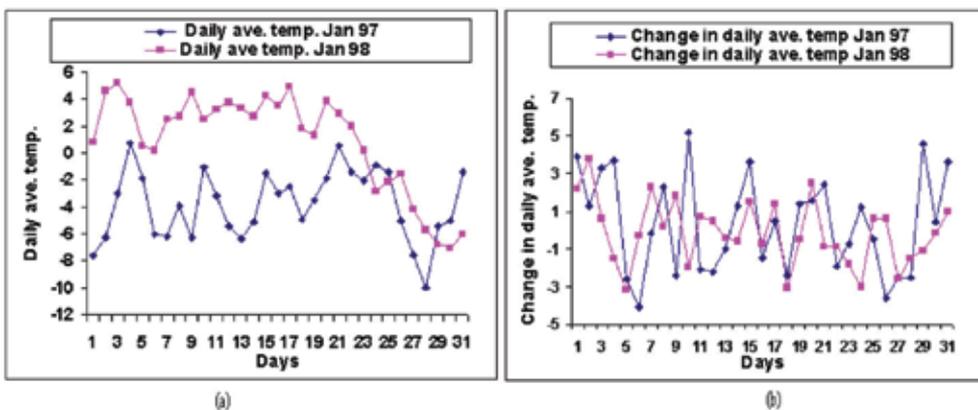


Fig. 1. (a) Daily Average Temperature (DAT). (b) the difference of DAT between two consecutive days.

Different data mining techniques have been used for this application. These techniques are mainly based on recurrent neural networks (RNN) (Elman, 1990) and Kohonen networks (Kohonen, 1995), Hidden Markov Models (HMM) (Picone, 1990), and Support Vector Machines (SVM) (Cortes et al., 1995). Essentially, we have developed several different variants of these techniques. In the case of SRNs, we proposed innovative architecture that can cope with time-series characteristics (Huang et al., 2006, Huang et al. 2005, Huang et al. 2004). It was shown that this new network architecture models more accurately the energy load forecasting and is about 20% more efficient than SVM or HMM based techniques (Tarik et al., 2006). We have also developed another hybrid technique that combines Kohonen networks with traditional data mining technique (k-means), the results were very promising, approaching 98,6% prediction accuracy (Gleeson et al., 2006). This motivated us to develop another hybrid technique based on the ensemble networks principal. The first version of this approach consisted of the combination of the previous developed techniques. While it was quite tricky to find a good system for combining the results of each technique, the overall results were better than the average of the results of the individual solutions. We are currently exploring a different approach of getting the benefit of each approach by building some interactions between them. These interactions are represented as some intermediate rules, allowing some of the techniques to change the its behaviour based on this new input.

#### 4. Spatio-temporal mining

Spatio-temporal Data Mining is an emerging research area (Roddick and Lees, 2001), encompassing a set of exploratory, statistical and computational machine learning approaches for analysing very large spatial and spatio-temporal datasets. Presently, several open issues can be identified in this research field ranging from the definition of suitable mining techniques able to deal with spatio-temporal information to the development of effective methods to analyse the produced results.

In spatio-temporal datasets an event is described as a spatial and temporal phenomenon, as we consider that it happens at a certain time  $t$  and at a location  $x$ . Examples of event types include hurricanes, tornados, road traffic jam, road accidents, etc. In real world many of these events interact with each other and exhibit spatial and temporal patterns, which may help us understand the physical phenomenon behind them. Therefore, it is very important to identify efficiently the spatial and temporal features of these events and their relationships from large spatio-temporal datasets of a given application domain.

Spatio-temporal data mining presents many challenges to both researchers and users communities (Compieta et al., 2007). For researchers, the key challenges are to develop efficient and general methods that can support complex spatio-temporal data types structures as well as the scalability issue as the amount of currently collected data increases at exponential rates. The relationships between the spatial and temporal aspects of the data are often not clear or well defined. Providing means of extracting or defining such relationships is difficult and it is one the main objectives of data mining approaches in this area. Finally, the granularity levels have direct impact on these relationships between spatial and temporal features, so deciding or determining which level(s) is(are) more appropriate to investigate is a very challenging issue. For the users, even though they might have a deep understanding of the application at hand, the key issue is have to find a proper model that supports the given data mining approach. This step is not easy as there are no general rules

of how to build such a model. This type of experimentations with the data and expertise are still part of the current research efforts.

While these challenges remain hot topics in the area, there has been a good progress, in data preparation, models, dependency analysis, etc., in recent years. Based on the models that have been developed for spatial and temporal data of real-world applications, we can classify these applications into four categories (Yao, 2003): 1) Applications where the time is not part of the recorded data (may be not important). In this case there is no way of extracting patterns from the data that include the time dimension. These include all purely spatial data mining applications. 2) Applications where the data is recorded as ordered sequences of events according to a specific relation such as before and after, time-stamp, etc. 3) Applications where the data is recorded at regular intervals, and finally 4) applications where the dimension time is fully integrated in the recorded data. The application example given in section 3.2 can be classified into the third category.

We can define spatio-temporal data as a set of spatio-temporal sequences,  $S$ . Each element of the sequence is represented by its spatial and temporal attributes  $(x_1, x_2, \dots, x_n, t)$ , where  $x_i, 1 \leq i \leq n$ , is a spatial attribute and  $t$  a temporal attribute. For sake of clarity, we consider only one principal temporal attribute and we are not mentioning here non-spatial attributes. The goal is to study the behaviour of some objects (events) in the space or through the time. An example would be the study of the movement of a Hurricane. In this example, one should define what is an object in a Hurricane and then track its movement in the space. Moreover, we need to take into account the surrounding of an object of a Hurricane; therefore more than one object should be tracked at the same time. Note that some objects are dynamic, they can appear and disappear at any time or change in shape, become bigger or smaller or experience a major change in its original shape. This complicates the task of the data mining technique employed. There are different models that can be explored depending on the how the data is collected. For instance the object can be well defined in the space and its locations are recorded in regular timestamps. Things can become more complicated when the object is not defined clearly and its location is not recorded at the same time interval. For the application of energy load forecasting, we have developed approaches based only on temporal analysis. In the case where the data collections exist for different regions, then it is worth looking at models taking into account the spatial dimensions. Currently, we have collections of data from different European countries. Our aim is to develop spatio-temporal models and evaluate both the accuracy and the complexity of such models.

To summarise, depending on the application at hand the knowledge discovery approach can be very complicated involving different inter-dependent steps to be dealt with. Until now, scientists and researchers have proposed concise solutions to specific problems. However, while these solutions work very well on the original targeted problems, they may, it is often the case, deliver poor performance on another problem. In section 6, we will describe a general framework for spatio-temporal data mining by trying to address their main challenges. We will leave very specific details of a given application to the user.

## 5. Visualisation for spatial data mining

Visualisation involves the use of visual/graphics techniques to represent information, data or knowledge. These techniques can be employed where complex datasets need to be explained or analyzed for developing and communicating conceptual information. The essential idea is that visual representations can help the user to get a better understanding of

content of the datasets, since the human visual system is more inclined to process visual rather than textual information. Thus, visualisation techniques may act as intelligence amplification tools for aiding and enhancing human intelligence, improving the perceptive, cognitive, and analytical abilities of people to allow them to solve complex tasks.

Nowadays, there exist very powerful computers able to quickly perform very complex and tedious tasks. Nevertheless, it is recognized that human performs better than computers in some areas, such as pattern recognition, evaluation, the overall sense of context that allows previously unrelated information to become related and useful, flexibility, and imagination. Based on these considerations, visualisation is widely recognized as essential during data analysis and knowledge discovery for gaining an insight of data and underlying phenomena they represent, since it takes advantage of human abilities to perceive visual patterns and to interpret them (Andrienko et al., 2003), (Andrienko et al., 2005), (Johnston, 2001), (Kopanakis & Theodoulidis, 2003), (Costabile & Malerba, 2003). Moreover, it takes advantage of human ability to deal with non-completely defined problems (as often is the case for decision problems) thus overcoming an evident weakness of computers (Adrienko et al., 2007). As observed in (Walker, 1995):

*“Programming a computer to “look for something interesting” in a database is a major undertaking, but given appropriate tools, it is a task for which humans are well equipped.”*

On the other hand, the powerful processing capabilities of computers are essential to deal with the huge amounts of data currently available. Thus, the idea of **Visual data mining** is to synergistically combine the computer processing capabilities with the unique and great human capabilities, to gain knowledge on the considered phenomena (Adrienko et al., 2007). In this context, a crucial role is played by the interaction techniques provided to the user for directly exploring the visual representation of data. Indeed, Visual Data Mining usually refers to methods, approaches and tools for the exploration of large datasets by allowing users to directly interact with visual representations of data and dynamically modify parameters to see how they affect the visualised data. Indeed, the usual approach follows the Information Seeking Mantra (Shneiderman, 1996) consisting of three steps: *Overview* first, *zoom and filter*, and then *details-on-demand*. Starting from an overview of the data the user identifies interesting patterns or subsets and focuses on one or more of them. Then, to analyse the identified patterns he/she requires to access details of the data exploiting a drill-down capability (Keim et al., 2003). For all the three steps of the process, effective visualisation techniques and interaction facilities have to be provided.

Visual Data Mining techniques, shifting the load from the user's cognitive system to the perceptual system, are able to enhance the effectiveness of the overall mining process, by supporting analytical reasoning, and have proven to be very valuable in many application domains.

In the context of **mining large spatial-temporal datasets** where geographical or physical space is involved, the visual exploratory approach is especially useful. Indeed:

- The heterogeneity of the space and the variety of properties and relationships in it cannot be adequately represented for fully automatic processing, thus there is the need to complement the computers capabilities with more sophisticated human capabilities.
- At the same time, an isomorphic visual representation, such as a map or an orthophoto, allows a human analyst or decision maker to perceive spatial relationships and patterns directly.

- Furthermore, a map or photo portraying coasts, rivers, relief, state boundaries, cities, roads, etc. exhibits not only the heterogeneity of the space but establishes the geographic context within which decisions can be made. The analyst or decision maker can grasp this information and relate it to his/her background knowledge about the properties of different parts of the space and take the variation of the properties into account.

However it is widely recognized that spatial visualisation features provided by existing geographical applications are not adequate for decision-support systems when used alone, but alternative solutions have to be defined. Existing tools (in particular, GIS, which are most commonly used as spatial decision aids) are often incapable to cope with the size and complexity of real-life problems, which forces the users to reduce the problems in order to adapt them to the capabilities of the tools (Adrienko et al. 2007). Moreover, as we have pointed out above, visualisation techniques for data exploration should not only include a static graphical view of the results produced by the mining algorithms, but also the possibility to dynamically obtain different spatial and temporal views as well as to interact in several ways with them. For example, the functionality of dynamically changing some of the involved parameters, and for that of quickly switching between different views for fast comparisons should be provided. This could allow the discovery of details and patterns that might remain hidden otherwise.

Some challenges can be identified for visual data mining of large spatial-temporal datasets:

- First of all, it is crucial to identify the most effective way to visualise the spatio-temporal multidimensional dataset taking into account the specific characteristics of the dataset, in order to communicate the useful and relevant information and to amplify the human capabilities.
- These visualisation methods and tools must be scalable with respect to the amount of data, dimensionality, number of data sources and heterogeneity of information, data quality and resolution, and characteristics of various displays and environments such as size, resolution, interaction possibilities, etc.
- It is important to provide effective visual interfaces for viewing and manipulating the geometrical and temporal attributes of the spatial-temporal data.
- Both the visualisation techniques and the interaction techniques should take into account that several persons are usually involved in the decision-making processes that data mining should support. These people have different role (administrators, politicians, data mining experts, domain experts, people affected by the decisions made) and very different background, thus requiring different needs. So, a user and task centered approach should be adopted to define appropriate visualisation and interaction techniques to effectively support each one of the identified actors and at the same time to allow them to fruitfully collaborate.

Visual data mining for spatio-temporal dataset is an interdisciplinary research area where techniques and expertise from information visualisation, visual perception, visual metaphors, diagrammatic reasoning, 3D computer graphics need to be suitably combined with the ones from data mining and geographic information systems.

## 6. Our approach

In this section we describe our work on the combination of visualisation and data mining techniques for spatio-temporal analysis and exploration.

As a case study and test bed we considered the Hurricane Isabel dataset (<http://www.tpc.ncep.noaa.gov/2003isabel.shtml>), which struck the US east coast in 2003

(National Hurricane Center 2003). This dataset is represented by a space of (500x500x100) with 25x106 real valued points in each of the 48 time-steps (approximately 62.5 GB). The main problem for analysing very large datasets such as this is that the hardware resources are not able to deal with the storage (memory) and processing (CPU) within the response time expected by the user to perform interactive queries.

To tackle this problem we have developed a 2-pass strategy (Di Martino et al 2006, Bertolotto et al., 2007). The goal of this type of strategy is to reduce the amount of memory used during the mining process as well as the processing time of a user query. Ideally, the data reduction or compression should not affect the knowledge contained in the data. The first task, then, in these strategies is to find the data points that are most similar according to their static (non spatial and temporal) parameters. This first phase is the key to the whole success of the compression, so that we do not lose any important information from the data that could have an adverse effect on the result obtained from mining the data at a later stage. The second task is to cluster these groups of closely related data points in a meaningful way to produce new "meta-data" sets that are more suitable and acceptable for data mining techniques to analyse and produce results (i.e. models, patterns, rules, etc.).

We have implemented two traditional clustering algorithms; DBSCAN and CURE (Kechadi et al. 2007). The first algorithm is more suitable for similarity measures that can be represented by a distance measure. Each cluster is represented by one data object. CURE can accept any similarity measure and the clusters can be represented by more than one representative. This is very important to represent clusters of different shapes. There are locations in the space that are highly similar; these are represented within each of the small location groups. It is important that no location group overlaps with another location group so that the integrity of the data is not affected.

In spatial and spatio-temporal data mining the effective analysis of results is crucial. Visualisation techniques are fundamental in the support of this task. We employed geo-visualisation in support of spatio-temporal data mining by developing two alternative interfaces, one based on the Google Earth application and one based on a Java 3D implementation. These tools have been developed to fit the complementary requirements posed by domain and mining experts, to allow the definition of a distributed, collaborative environment, and to deal with a dataset that could take advantage of a three-dimensional visualization in a geo-referenced space. The Google Earth-based tool renders in 3D the mining outcomes over a geo-referenced satellite image, enhanced by additional informative layers. The Java 3D-based tool provides more advanced user interaction with the mining results, by providing a set of features oriented to data-mining experts. These tools can be used in conjunction, as well as linked to any number of other instances, over an IP-based network, to create a collaborative and distributed environment.

In particular within our system, we developed an interacting visualisation functionality that allows not only to visualise the shape of clusters but also the shape of rules extracted by the mining algorithm. Therefore, a cluster or a rule produced at the mining layer is accessed directly and represented by its shape at the visualisation layer. This shape represents the region of space where the extracted rule holds (i.e., the set of locations in which the rule and hence all objects involved in it are well supported). While simply removing confusion and overload of visual information from the screen, it also help to highlight the structure of any pattern embedded in the data and to focus the user's attention only on the subset of the dataset involved in the rule being studied. This allows a more efficient and light visualization process, even when displaying millions of points.

The developed system has been successfully tested against the meteorological dataset, gathering positive results, since the system allowed us to detect both expected and

unexpected behaviours, as well as to find interesting relationships and specific patterns/characteristics about hurricane data (see Bertolotto et al., 2007 and Kechadi et al. 2007 for more details).

## 7. Conclusion

This chapter discusses available techniques and current research trends in the field of spatio-temporal data mining. An overview of the proposed approaches to deal with the spatial and the temporal aspects of data has been presented. Approaches that aim at taking into account both aspects were also surveyed.

Many challenges are still to be addressed. In particular, in the cartography and GIS community background/domain knowledge plays a significant role in the analysis of data. Therefore one of the biggest challenges is to integrate background geographic knowledge within the mining process. This is still an unexplored issue.

Currently, huge volumes of data are collected daily are often heterogeneous, geographically distributed and owned by different organisations. For example, an application by its nature is distributed such as an environmental application for which the data is collected in different locations and times using different instruments, and therefore these separate datasets may have different formats and features. So, traditional centralised data management and mining techniques are not adequate anymore. Distributed and high performance computing knowledge discovery techniques constitute a better alternative as they are scalable and can deal efficiently with data heterogeneity. So distributed data mining has become necessary for large and multi-scenario datasets requiring resources, which are heterogeneous and distributed. This constitutes an additional complexity to spatio-temporal data mining. We will look at this problem in our ADMIRE framework (Le-Khac 2006).

Visual techniques are essential for effective interpretation of mining results and as support to the mining process itself. We have discussed such techniques and presented our work in the area.

## 8. Acknowledgements

Research presented in this paper was funded by a Strategic Research Cluster grant (07/SRC/I1168) by Science Foundation Ireland under the National Development Plan. The authors gratefully acknowledge this support.

## 9. References

- Andrienko N., Andrienko G., and Gatalsky P., Exploratory Spatio-Temporal Visualization: an Analytical Review. *Journal of Visual Languages and Computing*, special issue on Visual Data Mining, December 2003, v.14 (6), pp. 503-541.
- Andrienko N., Andrienko G., *Exploratory Analysis of Spatial and Temporal Data - A Systematic Approach*, Springer, 2005.
- Andrienko, G.L., Andrienko, N.V., Jankowski, P., Keim, D.A. , Kraak, M-J., MacEachren, A.M. Wrobel, S., Geovisual analytics for spatial decision support: Setting the research agenda. *International Journal of Geographical Information Science* 21(8): 839-857, 2007.
- Bertolotto, M., Di Martino, S., Ferrucci, F., Kechadi, M-T., Visualization System for Collaborative Spatio-Temporal Data Mining, *Journal of Geographical Information Science*, Vol. 21, No. 7, July, 2007.

- Camossi, E., Bertolotto, M., Kechadi, M-T., Mining Spatio-Temporal Data at Different Levels of Detail, Association Geographic Information Laboratories Europe, Girona, Spain, May 5-8, 2008.
- Compieta, P., Di Martino, S., Bertolotto, M., Ferrucci, F., Kechadi, M-T., Exploratory Spatio-Temporal Data Mining and Visualization, *Journal of Visual Languages and Computing*, Vol. 18, No. 3, June 2007.
- Cortes, C., Vapnik, K., Support Vector Networks, *Machine Learning*, 20(3): 273-297, 1995.
- Costabile, M.F., Malerba, D. (Editors), Special Issue on Visual Data Mining, *Journal of Visual Languages and Computing*, Vol. 14, December 2003, 499-501.
- Di Martino, S., Ferrucci, F., Bertolotto, M., and Kechadi, M-T., Towards a Flexible System for Exploratory Spatio-Temporal Data Mining and Visualization, *Workshop on Visualization, Analytics & Spatial Decision Support (in GIScience'06)*, Münster, Germany, September 20-23, 2006.
- Elman, J.L., Finding Structure in Time, *Cognitive Science*, Vol 14, No. 2, 1990, 179-211.
- Ester, M., Kriegel, H.-P., Sander, J., Algorithms and applications for spatial data mining, in H. J. Miller and J. Han (eds.) *Geographic Data Mining and Knowledge Discovery*, London: Taylor and Francis, 160-187, 2001.
- Ewens, W.J., Grant, G.R., *Statistical methods in bioinformatics: An introduction* (New York: Springer-Verlag), 2001.
- Fitzgerald, T., Kechadi, M-T., Geiselbrechtinger, F., Application of fuzzy logic to online recognition of handwritten symbols, *IEEE, 9th Int'l. Workshop on Frontiers in Handwriting Recognition*, Tokyo, Japan, October 26-29, 2004.
- Gleeson, B. and Kechadi, M-T., Electric Load Forecasting Using Weather Data with a Kohonen Network and Data Mining Approach, *The 26th International Symposium on Forecasting*, Santander, Spain, June 11-14, 2006.
- Gopal, S., Liu, W., Woodcock, X. Visualization based on fuzzy ARTMAP neural network for mining remotely sensed data, in H. J. Miller and J. Han (eds.) *Geographic Data Mining and Knowledge Discovery*, London: Taylor and Francis, 315-336, 2001.
- Han, J., Kamber, M., Tung, A. K. H. (2001) "Spatial clustering methods in data mining: A survey," in H. J. Miller and J. Han (eds.) *Geographic Data Mining and Knowledge Discovery*, London: Taylor and Francis, 188-217.
- Hinneburg, A., Keim, D. A., (1998) An efficient approach to clustering in large multimedia databases with noise, in *Proceedings KDD*, New York, 58-65.
- Huang, B.Q., Rashid, T., Kechadi, M-T., A New Modified Network Based on the Elman Network, *Int'l. Conference on Artificial Intelligence and Applications*, Innsbruck, Austria, Feb. 16-18, 2004.
- Huang, B.Q., Kechadi, M-T., A Recurrent Neural Network Recogniser for Online Recognition of Handwritten Symbols, *The 7th Int'l. Conference on Enterprise Information Systems (ICEIS'05)*, Miami, FL, USA, May 24-28, 2005.
- Huang, B.Q., Rashid, T., Kechadi, M-T., Multi-Context Recurrent Neural Network Time Series Applications", *International Journal of Computational Intelligence*, Vol. 3, No. 3, February 2006.
- Johnston W.L., Model visualization, in: *Information Visualization in Data Mining and Knowledge Discovery*, Morgan Kaufmann, Los Altos, CA, 2001, pp. 223-227.
- D.A. Keim, C.Panse, and M. Sips "Visual Data Mining of Large Spatial Data Sets", N. Bianchi-Berthouze (Ed.): *DNIS 2003, LNCS 2822*, pp. 201-215, 2003.
- Kechadi, M-T., Bertolotto, M., Di Martino, S., Ferrucci, F., Scalable 2-Pass Data Mining Technique for Large Scale Spatio-Temporal Datasets, *LNCS on Knowledge-Based Intelligent Information & Engineering Systems*, 4693, 785-792, 2007.

- Kohonen T., *Self-Organising Maps*, Springer Series in Information Sciences, Vol. 30, 1995.
- Kopanakis I., Theodoulidis B., Visual data mining modeling techniques for the visualization of mining outcomes. *Journal of Visual Languages and Computing*. 14(6): 543-589, 2003.
- Koperski K., Adhikary, J., Han, J., *Spatial Data Mining: Progress and challenges*, Proceedings of the ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, Montreal, Canada, 55-70, 1996.
- Koperski, K., Han, J., and Stefanovic N., An efficient two-step method for classification of spatial data, Proceedings of the Spatial Data Handling Conference, Vancouver, Canada, 1998.
- Laxman, S., Sastry, P.S., Unnikrishnan K.P., *Discovering Frequent Episodes and Learning Hidden Markov Models: A formal Connection*, IEEE Trans. Knowledge data Eng., 17 1595-1517, 2005.
- Le-Khac, N.A., Kechadi, M-T., *ADMIRE Framework: Distributed Data Mining on Data-Grid Platforms*, Int'l. Conference on Software and Data Technologies (ICSOF'T'06), Setubal, Portugal, September 11-14, 2006.
- Malerba, D., Esposito, F., Lanza, A., Lisi, F.A., *Machine learning for information extraction from topographic maps*, in H. J. Miller and J. Han (editors) *Geographic Data Mining and Knowledge Discovery*, Taylor and Francis, pp. 291-314, 2001.
- Ng, R., *Detecting outliers from large datasets*, in H. J. Miller and J. Han (editors) *Geographic Data Mining and Knowledge Discovery*, Taylor and Francis, 218-235, 2001.
- Peuquet, D.J. and Duan, N., *An event-based spatiotemporal data model (ESTDM) for temporal analysis of geographical data*. *International journal of Geographic Information systems*. 9:7-24, 1995.
- Shneiderman, P., *The eye have it: A task by data type taxonomy for information visualizations*. In *Visual Languages*, 1996.
- Picone, J., *Continuous speech recognition using hidden Markov models*, *IEEE Signal processing magazine*, 7:26-41, July 1990.
- Qi, F. and Zhu, A.-X. *Knowledge discovery from soil maps using inductive learning*, *International Journal of Geographical Information Science*, 17, 771-795, 2003.
- Rashid, T., Huang, B.Q., Kechadi, T-M., *Auto-Regressive Recurrent Neural Network Approach for Electricity Load Forecasting*, *International Journal of Computational Intelligence*, Vol. 3, No. 3, February 2006.
- Roddick, J. F. and Lees, B., *Paradigms for spatial and spatio-temporal data minig*, in H. J. Miller and J. Han (eds.) *Geographic Data Mining and Knowledge Discovery*, London: Taylor and Francis, 33-49, 2001.
- Shekhar, S., Huang, Y., Wu, W., Lu, C.T., Chawla, S., *What's spatial about spatial data mining: three case studies*, in R. Grossman, C. Kamath, V. Kumar, R. Namburu (editors), *Data Mining for Scientific and Engineering Applications*, Kluwer Academic Publishers, Dordrecht, 487-514, 2001.
- Shekhar, S., Lu, C. T., Zhang, P., *A unified approach to detecting spatial outliers*, *GeoInformatica*, 7, 139-166, 2003.
- Walker, G., *Challenges in Information Visualisation*, *British Telecommunications Engineering Journal*, Vol. 14, pp17-25, April 1995.
- Yao, X., *Research Issues in Spatio-Temporal Data Mining*, workshop on Geospatial Visualization and Knowledge Discovery, Virginia, USA, Nov. 18-20, 2003.
- Yuan, M., *Use of knowledge acquisition to build wildfire representation in geographical information systems*. *International Journal of Geographic Information Science*. 11:723-745, 1997.

# Benchmarking the Data Mining Algorithms with Adaptive Neuro-Fuzzy Inference System in GSM Churn Management

Adem Karahoca, Dilek Karahoca and Nizamettin Aydın  
*Software Engineering Department, Bahcesehir University  
Turkey*

## 1. Introduction

Turkey has started to distribute Global Services of Mobile (GSM) 900 licences in 1998. Turkcell and Telsim have been the first players in the GSM market and they bought licenses respectively. In 2000, GSM 1800 licenses were bought by ARIA and AYCELL respectively. After then, GSM market has saturated and customers started to switch to other operators to obtain cheap services, number mobility between GSM operators, and availability of 3G services.

One of the major problems of GSM operators has been churning customers. Churning means that subscribers may move from one operator to another operator for some reasons such as the cost of services, corporate capability, credibility, customer communication, customer services, roaming and coverage, call quality, billing and cost of roaming (Mozer et al., 2000). Therefore churn management becomes an important issue for the GSM operators to deal with. Churn management includes monitoring the aim of the subscribers, behaviours of subscribers, and offering new alternative campaigns to improve expectations and satisfactions of subscribers. Quality metrics can be used to determine indicators to identify inefficiency problems. Metrics of churn management are related to network services, operations, and customer services.

When subscribers are clustered or predicted for the arrangement of the campaigns, telecom operators should have focused on demographic data, billing data, contract situations, number of calls, locations, tariffs, and credit ratings of the subscribers (Yu et al., 2005).

Predictions of customer behaviour, customer value, customer satisfaction, and customer loyalty are examples of some of the information that can be extracted from the data stored in a company's data warehouses (Hadden et al., 2005).

It is well known that the cost of retaining a subscriber is much cheaper than gaining a new subscriber from another GSM operator (Mozer et al., 2000). When the unhappy subscribers are predicted before the churn, operators may retain subscribers by new offerings. In this situation in order to implement efficient campaigns, subscribers have to be segmented into classes such as loyal, hopeless, and lost. This segmentation has advantages to define the customer intentions. Many segmentation methods have been applied in the literature. Thus,

churn management can be supported with data mining modelling tools to predict hopeless subscribers before the churn. We can follow the hopeless groups by clustering the profiles of the customer behaviours. Also, we can benefit from the prediction advantages of the data mining algorithms.

Data mining tools have been used to analyze many profile-related variables, including those related to demographics, seasonality, service periods, competing offers, and usage patterns. Leading indicators of churn potentially include late payments, numerous customer service calls, and declining use of services.

In data mining, one can choose a high or low degree of granularity in defining variables. By grouping variables to characterize different types of customers, the analyst can define a customer segment. A particular variable may show up in more than one segment. It is essential that data mining results extend beyond obvious information. Off-the-shelf data mining solutions may provide little new information and thus serve merely to predict the obvious. Tailored data mining solutions can provide far more useful information to the carrier (Gerpott et al., 2001).

Hung et al., 2006 proposed a data mining solution with decision trees and neural networks to predict churners to assess the model performances by LIFT (a measure of the performance of a model at segmenting the population) and hit ratio. Data mining approaches are considered to predict customer behaviours by CDRs (call detail records) and demographics data in (Wei et al., 2002; Yan et al., 2005; Karahoca, 2004).

In this research, main motivation is investigating the best data mining model(s) for churning, according to the measure of the performances of the model(s). We have utilized data sets which are obtained from a Turkish mobile telecom operator's data warehouse system to analyze the churn activities.

## 2. Material and methods

Loyal 24900 GSM subscribers were randomly selected from data warehouses of GSM operators located in Turkey. Hopeless 7600 candidate churners were filtered from databases during a period of a year. In real life situations, usually 3 to 6 percent annual churn rates are observed. Because of computational complexity, if all parameters within subscriber records were used in predicting churn, data mining methods could not estimate the churners. Therefore we selected 31% hopeless churners for our dataset and discarded the most of the loyal subscribers from the dataset.

In pattern recognition applications, the usual way to create input data for the model is through feature extraction. In feature extraction, descriptors or statistics of the domain are calculated from raw data. Usually, this process involves in some form of data aggregation.

The unit of aggregation in time is one day. The feature mapping transforms the transaction data ordered in time to static variables residing in feature space. The features used reflect the daily usage of an account. Number of calls and summed length of calls to describe the daily usage of a mobile phone were used. National and international calls were regarded as different categories. Calls made during business hours, evening hours and night hours were also aggregated to create different features. The parameters listed in Table 1, were taken into account to detect the churn intentions.

Abbreviation	Meaning
City	Subscribers' city
Age	Subscribers' age
Occupation	Occupation code of Subscriber
Home	Home city of Subscriber
Month-income	Monthly income of Subscriber
Credit-limit	Credit Limit of Subscriber
Avglen-call-month	Average length of calls in last month
Avg-len-call4-6month	Average length of calls in last 4-6 months
Avg-len-sms-month	Number of SMS in last month
Avglensms4-6month	Number of SMS in last 4-6 months
Tariff	Subscriber tariff
Marriage	Marital status of the subscriber
Child	Number of children of the subscriber
Gender	Gender of the subscriber
Month-Expense	Monthly expense of subscriber
Sublen	Length of service duration
CustSegment	Customer segment for subscriber
AvgLenCall3month	Average length of calls in last 3 months
TotalSpent	Total expenditure of the subscriber
AvgLenSms3month	Number of sent SMS in last 3 months
GsmLineStatus	Status of subscriber line
Output	Churn status of subscriber

Table 1. Extracted features

Summarized attributes that are augmented by different data sources are given in Table 1. These attributes are used as input parameters for churn detection process. These attributes are expected to have higher impact on the outcome (whether churning or not). In order to reduce computational complexity of the analysis, some of the fields in the set of variables of corporate data warehouse are ignored. The attributes with highest Spearman's Rho values are listed in Table 2. The factors are assumed to have the highest contribution to the ultimate decision about the subscriber.

Attribute Name	Spearman's Rho	Input Code
Month-Expense	0.4804	in1
Age	0.4154	in2
Marriage	0.3533	in3
TotalSpent	0.2847	in4
Month-income	0.2732	in5
CustSegment	0.2477	in6
Sublen	0.2304	in7

Table 2. Ranked Attributes

A number of different methods were considered to predict churners from subscribers. In this section the brief definitions of data mining methods are given. The methods that are described in this section are general methods that are used for modelling the data. These methods are *JRip*, *PART*, *Ridor*, *OneR*, *Nnge*, *Decision Table*, *Conjunction Rules*, *AD Trees*, *IB1*, *Bayesian networks* and *ANFIS*. Except *ANFIS*, all the methods are executed in WEKA (Waikato Environment for Knowledge Analysis) data mining software [Frank & Witten, 2005].

### 2.1 JRip method

JRip implements a propositional rule learner, “Repeated Incremental Pruning to Produce Error Reduction” (RIPPER), as proposed by [Cohen, 1995]. JRip is a rule learner. In principle it is similar to the commercial rule learner RIPPER. The RIPPER rule learning algorithm is an extended version of learning algorithm IREP (Incremental Reduced Error Pruning). It constructs a rule set in which all positive examples are covered, and its algorithm performs efficiently on large, noisy datasets. Before building a rule, the current set of training examples are partitioned into two subsets, a growing set (usually 2/3) and a pruning set (usually 1/3). The rule is constructed from examples in the growing set. The rule set begins with an empty rule set and rules are added incrementally to the rule set until no negative examples are covered. After growing a rule from the growing set, condition is deleted from the rule in order to improve the performance of the rule set on the pruning examples. To prune a rule, RIPPER considers only a final sequence of conditions from the rule, and selects the deletion that maximizes the function [Frank & Witten, 2005].

### 2.2 PART method

The PART algorithm combines two common data mining strategies; the divide-and-conquer strategy for decision tree learning and the separate-and-conquer strategy for rule learning. The divide-and-conquer approach selects an attribute to place at the root node and “divides” the tree by making branches for each possible value of the attribute. The process then continues recursively for each branch, using only those instances that reach the branch. The separate-and-conquer strategy is employed to build rules. A rule is derived from the branch of the decision tree explaining the most cases in the dataset, instances covered by the rule are removed, and the algorithm continues creating rules recursively for the remaining instances until none are left. The PART implementation differs from standard approaches in that a pruned decision tree is built for the current set of instances, the leaf with the largest coverage is made into a rule and the tree is discarded. By building and discarding decision trees to create a rule rather than building a tree incrementally by adding conjunctions one at a time avoids a tendency to over prune. This is a characteristic problem of the basic separate and conquer rule learner.

The key idea is to build a partial decision tree instead of a fully explored one. A partial decision tree is an ordinary decision tree that contains branches to undefined sub trees. To generate such a tree, we integrate the construction and pruning operations in order to find a stable subtree that can be simplified no further. Once this subtree has been found tree building ceases and a single rule is read. The tree building algorithm splits a set of examples recursively into a partial tree. The first step chooses a test and divides the examples into subsets. PART makes this choice in exactly the same way as C4.5. [Frank & Witten, 2005].

### 2.3 Ridor method

Ridor generates the default rule first and then the exceptions for the default rule with the least (weighted) error rate. Later, it generates the best exception rules for each exception and iterates until no exceptions are left. Thus it performs a tree-like expansion of exceptions and the leaf has only default rules but no exceptions. The exceptions are a set of rules that predict the improper instances in default rules [Gaines & Compton, 1995].

### 2.4 OneR method

OneR, generates a one-level decision tree, that is expressed in the form of a set of rules that all test one particular attribute. 1R is a simple, cheap method that often comes up with quite good rules for characterizing the structure in data [Frank & Witten, 2005]. It turns out that simple rules frequently achieve surprisingly high accuracy [Holte, 1993].

### 2.5 Nnge

Nearest-neighbor-like algorithm is using for non-nested generalized exemplars Nnge which are hyper-rectangles that can be viewed as if-then rules [Martin, 1995]. In this method, we can set the number of attempts for generalization and the number of folder for mutual information.

### 2.6 Decision tables

As stated by Kohavi, decision tables are one of the possible simplest hypothesis spaces, and usually they are easy to understand. A decision table is an organizational or programming tool for the representation of discrete functions. It can be viewed as a matrix where the upper rows specify sets of *conditions* and the lower ones sets of *actions* to be taken when the corresponding conditions are satisfied; thus each column, called a *rule*, describes a procedure of the type "if conditions, then actions".

The performance of this method is quite good on some datasets with continuous features, indicating that many datasets used in machine learning may not require these features, or these features may have few values [Kohavi, 1995].

### 2.7 Conjunctive rules

This method implements a single conjunctive rule learner that can predict for numeric and nominal class labels. A rule consists of antecedents "AND"ed together and the consequent (class value) for the classification/regression. In this case, the consequent is the distribution of the available classes (or mean for a numeric value) in the dataset.

### 2.8 AD trees

AD Trees can be used for generating an alternating decision (AD) trees. The number of boosting iterations needs to be manually tuned to suit the dataset and the desired complexity/accuracy tradeoffs. Induction of the trees has been optimized, and heuristic search methods have been introduced to speed learning [Freund & Mason, 1999].

### 2.9 Nearest neighbour Instance Based learner (IB1)

IBk is an implementation of the k-nearest-neighbors classifier that employs the distance metric. By default, it uses just one nearest neighbor ( $k=1$ ), but the number can be specified [Frank & Witten, 2005].

## 2.10 Bayesian networks

Graphical models such as Bayesian networks supply a general framework for dealing with uncertainty in a probabilistic setting and thus are well suited to tackle the problem of churn management. Bayesian Networks was coined by Pearl (1985).

Graphical models such as Bayesian networks supply a general framework for dealing with uncertainty in a probabilistic setting and thus are well suited to tackle the problem of churn management. Every graph of a Bayesian network codes a class of probability distributions. The nodes of that graph comply with the variables of the problem domain. Arrows between nodes denote allowed (causal) relations between the variables. These dependencies are quantified by conditional distributions for every node given its parents.

## 2.11 ANFIS

A Fuzzy Logic System (FLS) can be seen as a non-linear mapping from the input space to the output space. The mapping mechanism is based on the conversion of inputs from numerical domain to fuzzy domain with the use of fuzzy sets and fuzzifiers, and then applying fuzzy rules and fuzzy inference engine to perform the necessary operations in the fuzzy domain [Jang,1992 ; Jang,1993]. The result is transformed back to the arithmetical domain using defuzzifiers. The ANFIS approach uses Gaussian functions for fuzzy sets and linear functions for the rule outputs. The parameters of the network are the mean and standard deviation of the membership functions (antecedent parameters) and the coefficients of the output linear functions (consequent parameters). The ANFIS learning algorithm is used to obtain these parameters. This learning algorithm is a hybrid algorithm consisting of the gradient descent and the least-squares estimate. Using this hybrid algorithm, the rule parameters are recursively updated until an acceptable error is reached. Iterations have two steps, one forward and one backward. In the forward pass, the antecedent parameters are fixed, and the consequent parameters are obtained using the linear least-squares estimate. In the backward pass, the consequent parameters are fixed, and the output error is back-propagated through this network, and the antecedent parameters are accordingly updated using the gradient descent method.. Takagi and Sugeno's fuzzy if-then rules are used in the model. The output of each rule is a linear combination of input variables and a constant term. The final output is the weighted average of each rule's output. The basic learning rule of the proposed network is based on the gradient descent and the chain rule [Werbos, 1974]. In the designing of ANFIS model, the number of membership functions, the number of fuzzy rules, and the number of training epochs are important factors to be considered. If they were not selected appropriately, the system will over-fit the data or will not be able to fit the data. Adjusting mechanism works using a hybrid algorithm combining the least squares method and the gradient descent method with a mean square error method. The aim of the training process is to minimize the training error between the ANFIS output and the actual objective. This allows a fuzzy system to train its features from the data it observes, and implements these features in the system rules. As a Type III Fuzzy Control System, ANFIS has the following layers as represented in Figure 1.

*Layer 0:* It consists of plain input variable set.

*Layer 1:* Every node in this layer is a square node with a node function as given in Eq. (1);

$$\mu_{A_i}(x) = \frac{1}{1 + \left[ \left( \frac{x - c_i}{a_i} \right)^2 \right]^{b_i}} \tag{1}$$

where  $A$  is a generalized bell fuzzy set defined by the parameters  $\{a,b,c\}$ , where  $c$  is the middle point,  $b$  is the slope and  $a$  is the deviation.

*Layer 2:* The function is a T-norm operator that performs the firing strength of the rule, e.g., fuzzy AND and OR. The simplest implementation just calculates the product of all incoming signals.

*Layer 3:* Every node in this layer is fixed and determines a normalized firing strength. It calculates the ratio of the  $j$ th rule's firing strength to the sum of all rules firing strength.

*Layer 4:* The nodes in this layer are adaptive and are connected with the input nodes (of layer 0) and the preceding node of layer 3. The result is the weighted output of the rule  $j$ .

*Layer 5:* This layer consists of one single node which computes the overall output as the summation of all incoming signals.

In this research, the ANFIS model was used for churn data identification. As mentioned before, according to the feature extraction process, 7 inputs are fed into ANFIS model and one variable output is obtained at the end. The last node (rightmost one) calculates the summation of all outputs [Riverol & Sanctis, 2005].

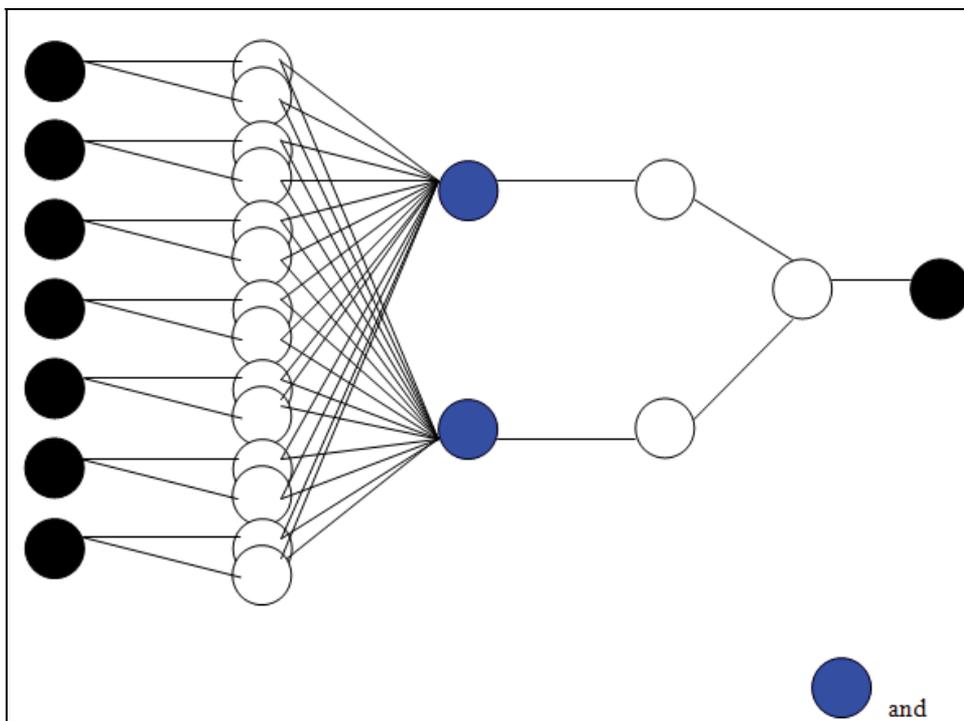


Fig. 1. ANFIS model of fuzzy inference

### 3. Findings

Benchmarking the performance of the data mining methods' efficiency can be calculated by confusion matrix with the following terms [Han & Kamber, 2000]:

1. True positive (TP) corresponding to the number of positive examples correctly predicted by the classification model.
2. False negative (FN) corresponding to the number of positive examples wrongly predicted as negative by the classification model.
3. False positive (FP) corresponding to the number of negative examples wrongly predicted as positive by the classification model.
4. True negative (TN) corresponding to the number of negative examples correctly predicted by the classification model.

The true positive rate (TPR) or sensitivity is defined as the fraction of positive examples predicted correctly by the model as seen in Eq.(2).

$$TPR = TP / (TP + FN) \quad (2)$$

Similarly, the true negative rate (TNR) or specificity is defined as the fraction of negative examples predicted correctly by the model as seen in Eq.(3).

$$TNR = TN / (TN + FP) \quad (3)$$

Sensitivity is the probability that the test results indicate churn behaviour given that no churn behaviour is present. This is also known as the true positive rate.

Specificity is the probability that the test results do not indicate churn behaviour even though churn behaviour is present. This is also known as the true negative rate.

Method	Training data	
	Sensitivity	Specificity
Jrip	0.85	0.88
Ridor	0.90	0.91
PART	0.60	0.93
OneR	0.79	0.85
Nnge	0.68	0.89
Decision Table	0.85	0.84
Conj.R.	0.66	0.75
AD tree	0.75	0.85
IB1	0.77	0.89
BayesNet	0.98	0.95
ANFIS	<b>0.86</b>	<b>0.85</b>

Table 3. Training Results for the Methods Used

Correctness is the percentage of correctly classified instances. RMS denotes the root mean square error for the given dataset and method of classification. Precision is the reliability of the test (F-score).

RMS, prediction and correctness values indicates important variations. Roughly JRIP and Decision Table methods have the minimal errors and high precisions as shown in Table 3 for training. But in testing phase ANFIS has highest values as listed in Tables 4 and 5.

RMSE (root mean squared error) values of the methods vary between 0.38 and 0.72, where precision is between 0.64 and 0.81. RMS of errors is often a good indicator of reliability of methods. Decision table and rule based methods tend to have higher sensitivity and specificity. While a number of methods show perfect specificity, the ANFIS has the highest sensitivity.

Testing data		
Method	Sensitivity	Specificity
JRIP	0.49	0.70
Ridor	0.78	0.78
PART	0.58	0.64
OneR	0.65	0.51
Nnge	0.58	0.57
Decision Table	0.75	0.73
Conj.R.	0.65	0.71
AD tree	0.74	0.77
IB1	0.72	0.65
Bayes Net	0.75	0.76
ANFIS	<b>0.85</b>	<b>0.88</b>

Table 4. Testing Results for the Methods Used

Testing data		
Method	Precision	Correctness
JRIP	0.64	0.43
Ridor	0.72	0.66
PART	0.69	0.51
OneR	0.70	0.55
Nnge	0.66	0.48
Decision Table	0.72	0.71
Conj.R.	0.70	0.55
AD tree	0.71	0.60
IB1	0.74	0.61
Bayes Net	0.72	0.71
ANFIS	<b>0.81</b>	<b>0.80</b>

Table 5. Testing Results for the Methods Used

Three types of Fuzzy models are most common; the Mamdani fuzzy model, the Sugeno fuzzy model, and the Tsukamoto fuzzy model. We preferred to use Sugeno-type fuzzy model for computational efficiency.

Sub-clustering method is used in this model. Sub clustering is especially useful in real time applications with unexpectedly high performance computation. The range of influence is 0.5, squash factor is 1.25, accept ratio is 0.5; rejection ratio is 0.15 for this training model. Within this range, the system has shown a considerable performance.

As seen on Figure 2, test results indicate that, ANFIS is a pretty good means to determine churning users in a GSM network. Vertical axis denotes the test output, whereas horizontal axis shows the index of the testing data instances.

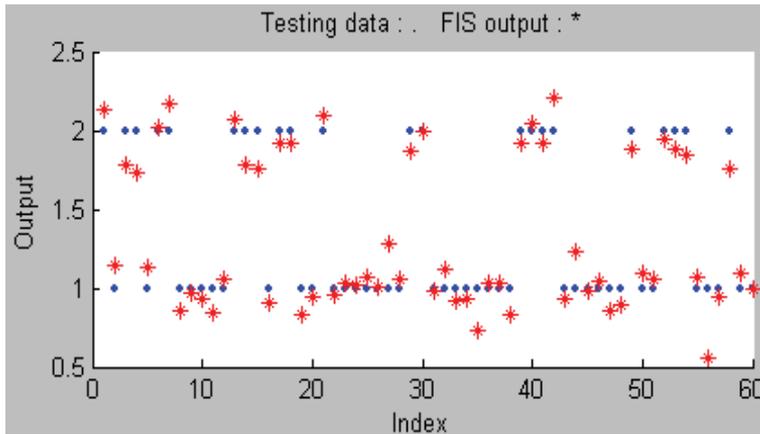


Fig. 2. ANFIS classification of testing data

Figure 2 and Figure 3, show plot of input factors for fuzzy inference and the output results in the conditions. The horizontal axis has extracted attributes from Table 2. The fuzzy inference diagram is the composite of all the factor diagrams. It simultaneously displays all parts of the fuzzy inference process. Information flows through the fuzzy inference diagram that is sequential.

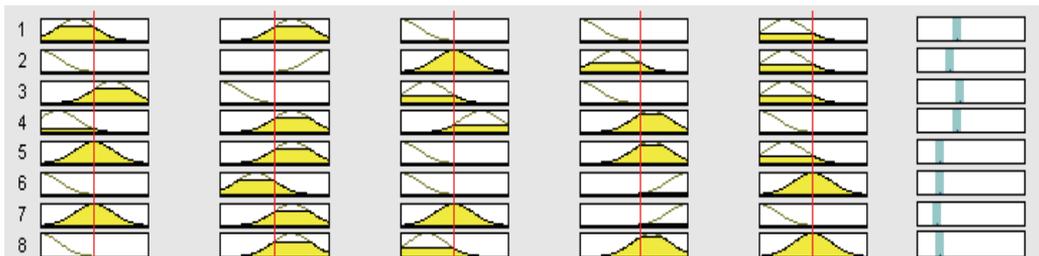


Fig. 3. Fuzzy Inference Diagram

ANFIS creates membership functions for each input variables. The graph shows Marital Status, Age, Monthly expense and Customer Segment variables membership functions. In these properties, changes of the ultimate (after training) generalized membership functions with respect to the initial (before training) generalized membership functions of the input parameters were examined.

Whenever an input factor has an effect over average, it shows considerable deviation from the original curve. We can infer from the membership functions that, these properties has considerable effect on the final decision of churn analysis since they have significant change in their shapes.

In Figures 4 to 7, vertical axis is the value of the membership function; horizontal axis denotes the value of input factor.

Marital status is an important indicator for churn management; it shows considerable deviation from the original Gaussian curve as seen in Figure 4, during the iterative process.

Figure 5 shows the initial and final membership functions. As expected, age group found to be an important indicator to identify churn. In network, monthly expense is another factor affecting the final model most. Resultant membership function is shown in Figure 6.

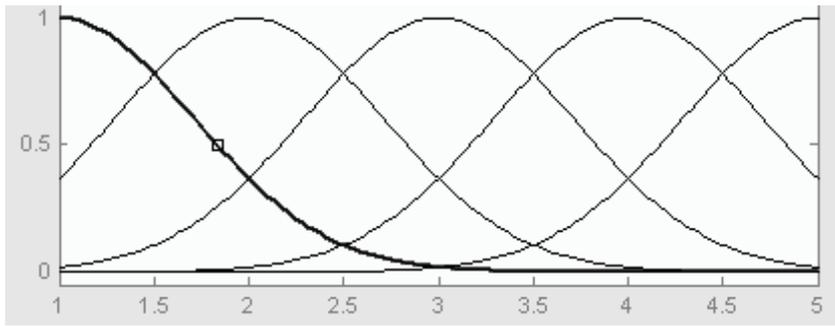


Fig. 4. Membership function for Marital Status

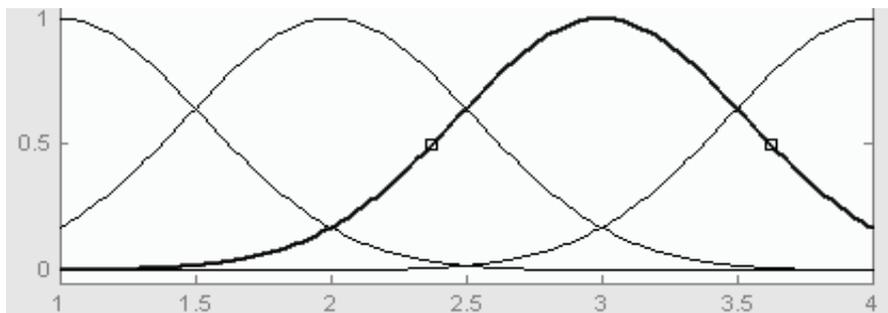


Fig. 5. Membership function for Age Group

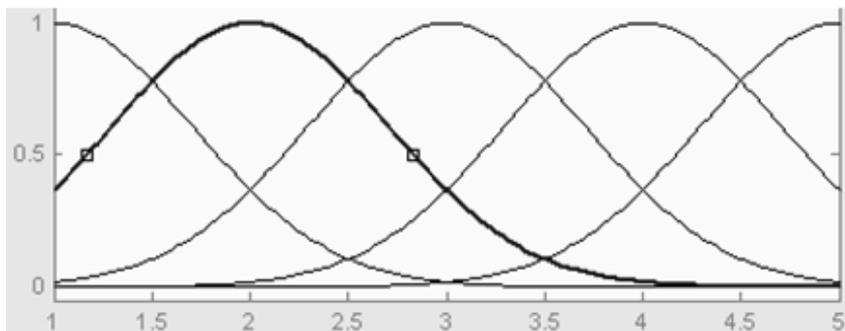


Fig. 6. Membership function for Monthly Expense

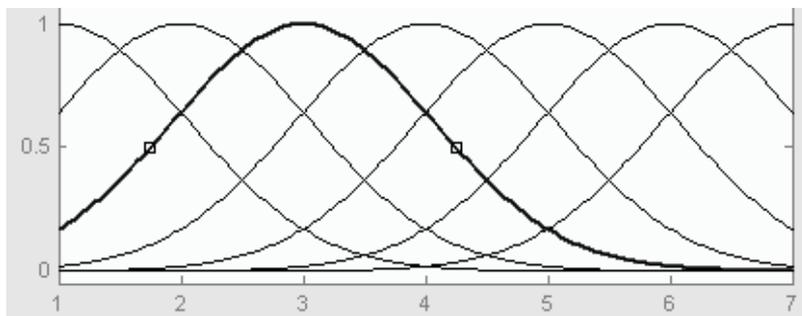


Fig. 7. Membership function for Customer Segment

Subscriber's customer segment also critically affects the model. As seen on Figure 7, deviation from original curve is significant.

These attributes represented in Figures 4 to 7 has the highest effect on final classification, the process has changed the membership functions significantly giving the values more emphasis for the final decision.

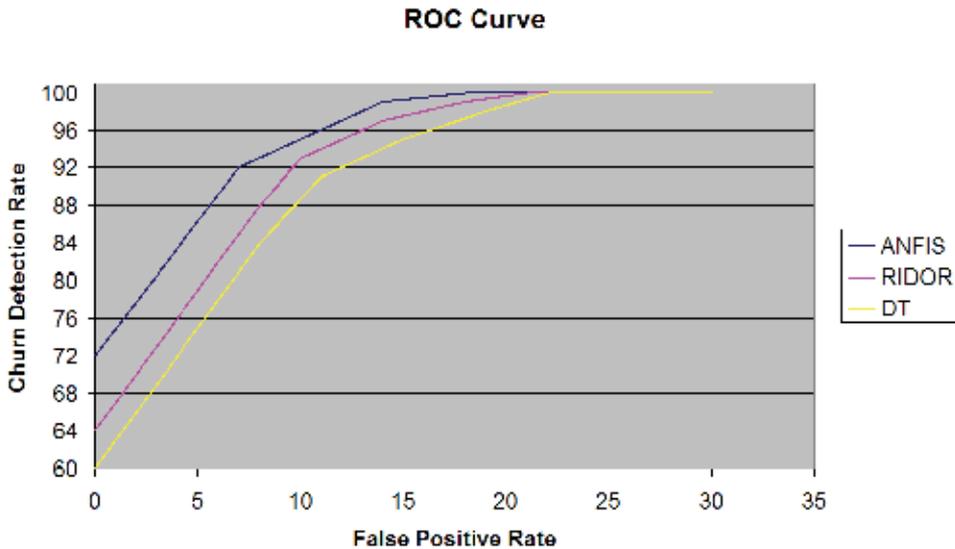


Fig. 8. Receiver Operating Characteristics Curve for Anfis, Ridor and decision trees

Figure 8 illustrates the ROC curve for the best three methods, namely ANFIS, RIDOR and Decision Trees. The ANFIS method is far more accurate where the smaller false positive rate is critical. In this situation where preventing churn is costly, we would like to have a low false positive ratio to avoid unnecessary customer relationship management (CRM) costs.

#### 4. Conclusions

The proposed integrated diagnostic system for the churn management application presented is based on a multiple adaptive neuro-fuzzy inference system. Use of a series of ANFIS units greatly reduces the scale and complexity of the system and speeds up the training of the network. The system is applicable to a range of telecom applications where continuous monitoring and management is required. Unlike other techniques discussed in this study, the addition of extra units (or rules) will neither affect the rest of the network nor increase the complexity of the network.

As mentioned in Section 2, rule based models and decision tree derivatives have high level of precision, however they demonstrate poor robustness when the dataset is changed. In order to provide adaptability of the classification technique, neural network based alteration of fuzzy inference system parameters is necessary. The results prove that, ANFIS method combines both precision of fuzzy based classification system and adaptability (back propagation) feature of neural networks in classification of data.

One disadvantage of the ANFIS method is that the complexity of the algorithm is high when there are more than a number of inputs fed into the system. However, when the system

reaches an optimal configuration of membership functions, it can be used efficiently against large datasets.

Based on the accuracy of the results of the study, it can be stated that the ANFIS models can be used as an alternative to current CRM churn management mechanism (detection techniques currently in use). This approach can be applied to many telecom networks or other industries, since it is once trained, it can then be used during operation to provide instant detection results to the task.

## 5. Acknowledgement

Authors thanks to Mert Şanver for his helps and works.

## 6. References

- Cohen, W. (1995). Fast effective rule induction. *Proceedings of the 12th International Conference on Machine Learning*, pp.115–123 Lake Tahoe, CA, 1995.
- Frank, E. & Witten, I. H. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 0-12-088407-0, San Francisco.
- Freund, Y. & Mason, L. (1999). The alternating decision tree learning algorithm. In *Proceedings of the Sixteenth International Conference on Machine Learning*. Pp. 124-133, Bled, Slovenia, 1999.
- Gaines, B.R. & Compton, P. (1995). Induction of ripple-down rules applied to modelling large databases. *Journal of Intelligent Information Systems* Vol.5, No.3, pp.211-228.
- Gerpott, T.J.; Rams, W. & Schindler, A. (2001). Customer retention loyalty and satisfaction in the German mobile cellular telecommunications market, *Telecommunications Policy*, Vol.25, No .10-11, pp.885-906.
- Hadden, J.; Tiwari, A.; Roy, R. & Ruta, D. (2005). Computer assisted customer churn management: state of the art and future trends. *Journal of Computers and Operations Research*, Vol.34, No .10, pp.2902-2917.
- Han, J. & Kamber, M. (2006). *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 978-1-55860-901-3, San Francisco.
- Hung, S-Y.; Yen, D. C. & Wang, H. Y. (2006). Applying data mining to telecom churn management. *Journal of Expert Systems with Applications*, Vol.31, No.3, pp.515-524.
- Holte, R.C. (1993). Very simple classification rules perform well on most commonly used datasets, *Machine Learning* Vol.11, pp.63-91.
- Jang, J-SR. (1992). Self-learning fuzzy controllers based on temporal back propagation, *IEEE Trans Neural Networks*, Vol.3, No.5, pp.714–723.
- Jang, J-SR. (1993). ANFIS: adaptive-network-based fuzzy inference system, *IEEE Trans. Syst. Man Cybernet*, Vol.23, No.3, pp.665–685.
- Karahoca, A. (2004). Data Mining via Cellular Neural Networks in the GSM sector, *Proceedings of 8th IASTED International Conference: Software Engineering Applications*, pp.19-24, Cambridge, MA. 2004.
- Kohavi, R. (1995). The power of decision tables, In *Proceedings of European Conference on Machine Learning*, pp.174-189 1995.
- Martin, B. (1995). Instance-based learning: Nearest Neighbor with generalization. Unpublished Master Thesis, University of Waikato, Hamilton, New Zealand.

- Mozer, M.C.; Wolniewicz, R.; Grimes, D.B.; Johnson, E. & Kaushansky, H.(2000). Predicting Subscriber Dissatisfaction and Improving Retention in the Wireless Telecommunications Industry, *IEEE Transactions on Neural Networks*, Vol.11, No.3, pp.690-696.
- Pearl, J. (1985). Bayesian Networks: A Model of Self-Activated Memory for Evidential Reasoning. In *Proceedings of the 7th Conference of the Cognitive Science Society*, pp.329-334, University of California, Irvine, CA, 1985, August 15-17.
- Riverol, C. & Sanctis, C. D. (2005). Improving Adaptive-network-based Fuzzy Inference Systems (ANFIS): A Practical Approach, *Asian Journal of Information Technology*, Vol.4, No.12, pp.1208-1212.
- Wei, C. P. & Chiu I-T. (2002). Turning telecommunications call details to churn prediction: a data mining approach, *Journal of Expert Systems with Applications*, Vol.23, pp.103-112.
- Werbos, P. (1974). Beyond regression, new tools for prediction and analysis in the behavioural sciences. Unpublished PhD Thesis, Harvard University.
- Yan, L.; Fasson, M. & Baldasare, P.(2005). Predicting customer behavior via calling links, *Proceedings of Int. Joint Conference on Neural Networks*, pp.2555-2560, Montreal, Canda, 2005.
- Yu, W.; Jutla, D. N. & Sivakumar, S. C. (2005). A churn management alignment model for managers in mobile telecom, *Proceedings of the 3rd annual communication networks and services research conferences*, pp.48-53.

## **PART IV: DATA MINING APPLICATIONS**



## SOCIAL APPLICATIONS



# Using Data Mining to Investigate the Behavior of Video Rental Customers

Tai-Chang Hsia<sup>1</sup> and An-Jin Shie<sup>2</sup>

<sup>1</sup>*Department of Industrial Engineering and Management, Chienkuo Technology University*

<sup>2</sup>*Department of Industrial Engineering and Management, Yuan Ze University*

<sup>1</sup>*Changhua, Taiwan, R.O.C.*

<sup>2</sup>*Taoyuan, Taiwan, R.O.C.*

## 1. Introduction

Living standards have been steadily rising in Taiwan, and indoor recreational activities are now receiving much attention. As a consequence, the market for video rentals has been flourishing in recent years. Renting videotapes, VCDs, and DVDs, to watch at home has become a common consumer entertainment activity. With the arrival of the US video store chain Blockbuster in Taiwan, stagnant personal incomes, and inflation, competition among video rental stores has grown, and their management is becoming more difficult. Like many retail businesses, video rental stores must be guided by the pattern of customer demand. Therefore, the purpose of this study is to investigate how to help rental stores using information in their databases to understand each customer's preferences and demand, so as to increase the rental ratio.

Using data mining theory, this study investigated the customer database of a single store of a local chain of video rental stores in a medium-sized city in central Taiwan, for the period January to March, 2007. First, the records were explored and analyzed in detail by decision tree algorithm. We determined the relationships among customer gender, occupational, favorites leisure activities, and video categories. Second, using the Apriori association rule algorithm for a rougher analysis, we explored and analyzed customers' personal preferences and video categories. We determined favorite videos and personal preferences, and developed rules for predicted which video types will be rented next time. Using these results, video rental stores can recommend personal favorites to each customer and invite customers to rent videos so that rental stores can increase their operating achievements.

## 2. Literature review

### 2.1 Consumer behavior

A survey by UDNJOB.Com CO. LTD (2006) showed that 27.8% of people in Taiwan watch TV videos to relax, the most popular activity, followed by travel, physical activities, dining out, shopping, playing on-line games, and enjoying culture. Obviously video entertainment is the most common and most basic form of consumer entertainment.

Scholars have formulated different definitions of consumer behavior. Schiffman and Kanuk (1991) categorized consumer behavior into searching, purchasing, using, estimation, and

handling products or services that consumers desire in order to meet their own needs. Engel et al. (1982) said that consumer behavior means the decision-making and actual realization processes the consumer engages in to meet his demand when he seeks, purchases, uses, appraises, and handles the product or the service. Wilkie (1994) observed that consumer behavior refers to the activities produced in the mind, the emotion and the body of consumers to meet their demand and desire while they choose, purchase, use, and handle products and services.

Based on the foregoing, we know that consumer behavior is the consumption processes and transaction activities that consumers carry out to satisfying their desires. Understanding the factors that shape consumer behavior can help video rental stores to design appropriate marketing strategies. Bazerman (2001) argued that the focus of research into consumer behavior should be changed from sales and marketing personnel to the customer himself. Above all, Bazerman says that if you can follow the consumer, you can follow the market. Given these understandings, this study uses the records of customer gender, occupational categories, personal preferences, and video rental categories in a video rental database to discover customer preferences and behaviors, for use in designing marketing strategies for video rentals.

## 2.2 Data mining

Scholars use a variety of different definitions for data mining. Frawley et al. (1991) declared that data mining is actually a process of discovering of nonobvious, unprecedented, and potentially useful information. Curt (1995) defined data mining as a database transformation process, in which the information is transformed from unorganized vocabulary and numbers to organized data, and later turned into knowledge from which a decision can be made. Fayyad et al. (1996) stated that data mining is an uncomplicated process of discovering valid, brand new, potentially useful, and comprehensive patterns from data. Hui and Jha (2000) defined data mining as an analysis of automation and semi-automation for the discovery of meaningful relationships and rules from a large amount of data in a database. They further categorized the data mining process into seven steps: 1. Establishing the mining goals; 2. Selection of data; 3. Data pre-processing; 4. Data transformation; 5. Data warehousing; 6. Data mining; 7. Evaluating the data mining results. Peacock (1998) declared that data mining can be categorized as Narrow and Broad. The narrow definition is limited by the methodology of mechanical learning, which emphasizes the discovery process and uses artificial intelligence such as Neural Networks, Correlation Rule, Decision Tree Algorithms and Genetic Algorithms. By contrast, the broad definition emphasizes the knowledge discovery in database (KDD), the process of obtaining, transforming, clarifying, analyzing, confirming and enduring the meaning of the data within existing customers or outside of the cooperation, which then results in a backup system of decision making that is continuously modified and maintained. Hand et al. (2000) stated that data mining is a process that discovers interesting and valuable information from a database. Berson et al. (2001) argued that the appeal of data mining lies in its forecasting competence instead of merely in its ability to trace back.

To summarize the foregoing definitions, data mining is a process of obtaining knowledge. The key to the process is comprehension of the research application, and then construction of a data set by collecting data relevant to the research field, purifying the data in the targeted database to eliminate erroneous data, supplementing missing data, simplifying and transforming the data, and then discovering the patterns among the data and presenting them as useful knowledge.

The range of data mining is extensive as artificial intelligence systems have made remarkable progress. A diversity of data mining algorithms have been developed for application to different types of data. When data mining needs to be performed, data features, research goals and predicted results should be considered first so that the most useful algorithms may be applied to the data.

Decision tree algorithm uses the technique of Information Gain and the theory of classification standards to automatically discover the correlation of targeted and forecasted variables. Based on a preset significance standard, the data is classified and clustered automatically. The Apriori association rule algorithm can find correlations among the attributes of different kinds of variables.

Because of their usefulness to this research, the above-mentioned two algorithms were used to classify the video rental database in order to find the correlation between consumer favorite activities and video categories. When the video rental stores have new videos, they can recommend videos to customers based on consumer preference and the correlation.

### 3. Methodology

This study used information contained in a database of 778 customers and their video rental data from a single store of a local chain of video rental stores in Changhua City, Taiwan, collected from January to March, 2007. Using the Chi-squared Automatic Interaction Detection (CHAID) decision tree data mining algorithm and the Apriori association rule, the behavior of the consumers was explored and analyzed. Prior to data mining, based on the purpose of the study, four variables were defined, i.e., customer gender, professional category, personal preference category, and video category. The four variables and their attributes are as follows:

#### 1. Customer gender category

In the study the attributes of customer gender variable are male and the female. Each customer is defined as one or the other.

#### 2. Occupational category

There are 6 items in the attributes of this category: students, manufacturing, engineering and commerce, government officials and teachers, freelancers, and finance and insurance. Each customer is assigned to one category.

#### 3. Favorite leisure activities category

Based on the survey of UDNJOB.Com CO. LTD (2006), excluding watching videos, the favorite recreational activities of people in Taiwan are: travel, physical activities, dining out, shopping, on-line games, and enjoying culture. Each customer may be assigned to more than one choice.

#### 4. Video category

Based on the classifications of the video rental store, there are nine attributes: comedy, action, horror, science fiction, crime, soap operas, romance, animation and adventure. Each customer may be assigned to more than one choice.

#### 3.1 Decision tree algorithm

Prior to the application of CHAID, targeted and forecasted variables should be defined. CHAID can automatically find the correlation between forecasted and targeted variable. Therefore, the study defined the three categories of customer occupation, favorite leisure

activities and video rentals, and used their characteristic data as the forecasted variables. Customer gender and its characteristic data were used as the targeted variables.

The calculation process of CHAID in decision-making tree was developed using the classification of the attributes of male and female. The, in light of the gender attributes, the attributes of customer occupational category were classified. Similarly, the attributes of male and female, favorite leisure activity, and video were next clustered. The classification of forecasted variables continued until the patterns in the data were discovered. This data mining approach creates a detailed analysis of the data.

One of the advantages of Decision Tree Algorithm is that the data can be automatically split and clustered based on the preset significance standard, and then be built up into a tree structure from the clustering event. Based on the tree structure, certain rules can be obtained, and the correlation between events found for further forecasting.

Each internal node in the tree structure is tested by a preset significance level. Branches are processed by the values of groups or multiple values of groups. This means that the branch of each internal node may have another branch that may at the same time be the internal node of another branch. They are tested in the order of the significance level until the branch cannot be split and comes to an end. The terminal node of the branch is called the leaf node. The path of each leaf node explains the reasons for and the results of each event. The study used the CHAID of the decision tree algorithm to carry out data mining and to build the tree correlation figure of the targeted and forecasted variables, as depicted in Figure 1.

Based on the tree correlation structure of Figure 1, we find there are differences in the video choices among the customers from different occupations and favorite leisure activities. Customer occupation may be divided into three categories: students (62.2%)(Node 1); engineering and commerce, finance and insurance, freelancers, government officials and teachers, (23.7%)(Node 8); and, manufacturing (14.1% )(Node 15).

First we carried out an analysis of students (Node 1). We found two kinds of student customers: those who like playing on-line games (44.5%) and those who dislike playing on-line games (17.7%), on two nodes. The former were called Node 3, the latter, Node 2. Because the p-value preset significance level is less than 0.05, Node 2 and Node 3 may each be divided into two nodes. Node 3 can be divided into Node 6 and Node 7, representing the proportion of student customers who like playing on-line games and choose action movies (31.6%) of which 95.1% are male and 4.9% female. Similarly, 12.9% of those students who like playing on-line games and choose non-action movies. In that group, 78% are male, while 22% are female. Nodes 6 and Node 7 have a p-value greater than the preset significance standard of 0.05 and thus cannot be further subdivided. On Nodes 4 and 5, subdivided from Node 2, no choice of any movie appeared. Hence, these nodes will not be considered in the analysis.

Next, we analyzed customers of the second sort of professional category (Node 8) working in engineering and commerce, finance and insurance, as freelancers, and as government officials and teachers. We found they choose crime movies (6.4%, Node 10) and non-crime movies (17.9%, Node 9), two nodes in total. With the p-value preset significance level of less than 0.05, Nodes 9 and 10 may each be further subdivided into two nodes. Node 10 can be subdivided into Nodes 13 and 14, meaning that 2.1% of the group of professional occupations choose both crime movies and romance movies. In this group 50% are male and 50% female. Non-romance movies are chosen by 4.4% of this occupational grouping, with males constituting 82.4%. At the preset significance level of less than 0.05, Nodes 13 and 14 cannot be further subdivided. Nodes 11 and 12, subdivided from Node 9, contain no individuals choosing movies and will not be considered in this study.

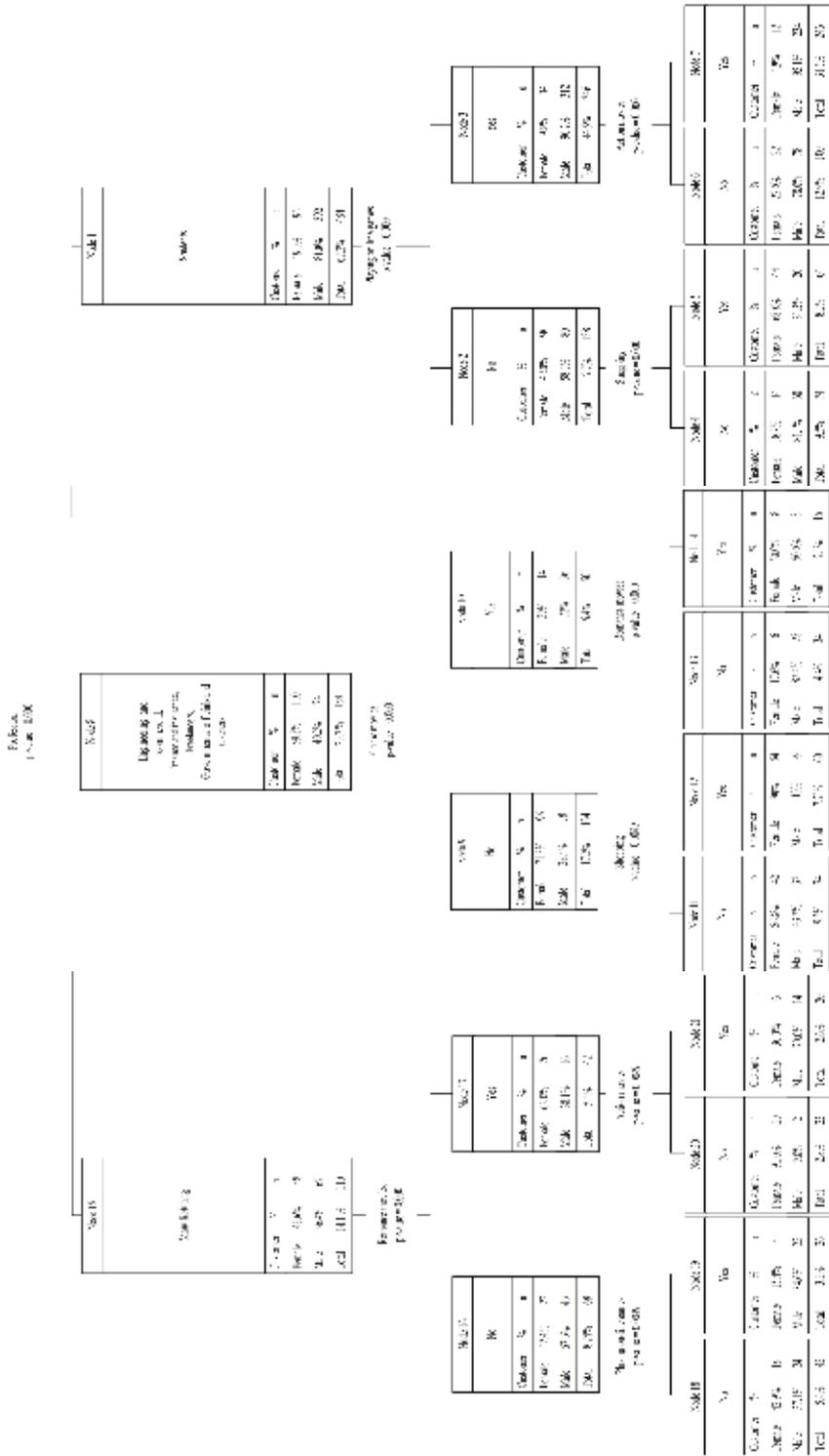


Fig. 1. Structure tree relation between video rentals and customer gender, profession, and favorite activities.

Finally we analyzed customers (Node 15) of the third occupational category working in the manufacturing industry. We found the proportion of customers working in the manufacturing industry and choosing romance movies is 5.4% (Node 17), while Node 16, manufacturing industry and non-romance movies, comprises 8.7%. At the preset significance level of less than 0.05, Nodes 16 and 17 may each be further subdivided. Node 17 can be subdivided into Nodes 20 and 21, representing manufacturing industry customers who choose romance and action movies (2.6%, 70% male). Among such customers choosing non-action movies (2.8%) 9% are male. Nodes 20 and 21 have a p-value greater than the preset significance standard of 0.05 and thus cannot be further divided. Nodes 18 and 19, subdivided from Node 16, do not contain anyone choosing movies and will not be considered.

### 3.2 Apriori algorithm

The Apriori algorithm was first published by Agrawal and Srikant in 1994. During the application of the Apriori association rule algorithm, no defined targeted variables or forecasted variables need be defined, but the attributes of two or more than two variables which meet the following two conditions are necessary. The first condition is that the attributes of variables must correspond with Boolean type ("1"/"0"). The second condition is that the attributes of the variables must be multiple-choice. With those two conditions fulfilled, the Apriori algorithm can be used to find the common correlation among the attributes.

Among the four defined variables of the study, both favorite leisure activities and video category met the above-mentioned conditions. The Apriori algorithm could thus be applied to find correlations among the attributes. These might include discovering what leisure activities are correlated with what video preferences, or what videos will be rented by individuals with a given personal favorite when they rent the next time. Apriori algorithm can find a portion of the correlations among the variables that need to be explored and does not need to discover the entire set of correlations among all defined variables. This data mining method provides a rougher analysis of the data than the Decision Tree approach.

The Apriori association rule algorithm explores the attributes of various variables in which 0 and 1 coexist. This coexistence relationship causes some level of correlation. Rules which rely on one another can be discovered and the proper conclusion stated at last. For instance, imagine that customers bought hamburgers, fried chicken, and potato chips at McDonald's. It could then be deduced that such consumers also tend to buy Coca Cola. Similarly, when individuals have high cholesterol, high triglycerides, and high uric acid, Apriori would deduce that they would also have high blood pressure. Of course, the strength of such correlations may vary. The strength of a given correction is called confidence and is shown by a percentage (%). The higher the percentage is, the higher the degree of correlation between Antecedent and Consequent.

In this study Apriori algorithm was used to analyze favorite leisure activities and video category. After application of the Apriori algorithm, 13 rules were produced with confidence ranging from 85.4% to 90.2%. Five examples are given below.

Rule 1: With a preference for crime movies, customers whose favorite activity is physical activities and like renting science fiction videos tend to rent action movies next. The confidence is 90.2%.

Rule 2: With a preference for crime movies, those customers whose favorites are physical activities and playing on-line games tend to rent action movies next (confidence = 87.8%).

Rule 3: With a preference for crime movies, those customers whose favorite activity is enjoying culture and like renting science fiction movies tend to rent action movies next (confidence = 87.5%).

Rule 4: Customers whose favorite activities are shopping and enjoying culture and who have a preference for renting action movies tend to rent comedy next (confidence = 87.2%).

Rule 5: Customers whose favorite activities are physical activities, travel, and enjoying culture tend to rent comedy next (confidence = 85.4%).

In the above-mentioned rules, the calculation of confidence is performed via the following formula (1).

$$\frac{\text{Support \% of Antecedent and Consequent}}{\text{Support \% of the Antecedent}} = \text{Confidence\%} \quad (1)$$

Using Rule 1 as an example, in formula (1), the Support % of the Antecedent is 10.5%. There are 82 instances in which customers whose favorite activity is sports and who have rented crime movies and science fiction videos ( $82/778 = 10.5\%$ ). Support % of Antecedent and Consequent means there are 74 instances of those customers whose favorite activity is sports and who rented crime films, science fiction movies and action videos (9.47% of 778). Thus, Support % of Antecedent divided by Support % of Antecedent and Consequent gives 90.2% ( $74/82 = 90.2\%$ ).

Rule	Antecedent	Consequent	Instances of Antecedent	Support % of Antecedent	Instances of Antecedent and Consequent	Support % of Antecedent and Consequent	Confidence %
1	crime movies and physical activities and science fiction movies	action movies	82	10.5%	74	9.47%	90.2%
2	crime movies and physical activities and playing on-line games	action movies	82	10.5%	72	9.22%	87.8%
3	crime movies and science fiction movies and enjoying culture	action movies	96	12.3%	84	10.76%	87.5%
4	shopping and enjoying culture and action movies	comedy movies	78	10.0%	68	8.72%	87.2%
5	physical activities and travel and enjoying culture	comedy movies	82	10.5%	70	8.97%	85.4%

Table 1. Rules Produced by Apriori Association Rule Algorithm

#### 4. Result analysis

This study applied the decision tree algorithm to construct a detailed exploration and analysis of the video rental database of a video rental store in central Taiwan. Three interesting results were found. First, the profession of customers who most often rented videos was students. Those customers whose favorite activity was playing on-line games and who liked action movies the best were generally male. Second, the number two customer group renting videos consisted of four professions: engineering and commerce, finance and insurance, freelancers, and government officials and teachers. Most of these customers preferred to rent crime movies and most of them are male. Romance movies are their second choice. The rate of male customers renting romance videos is the same as that of females. Third, the least common occupation for video customers is the manufacturing

industry. Most of these customers prefer to rent romance movies and most of them are female. In this group, action movies are the second most popular preference. Most of the customers who rent action videos are male.

Further, after performing a rough exploration and analysis of the video rental database using the Apriori association rule algorithm, two possible marketing strategies are suggested. First, no matter what the customer's favorite activity is, clerks should recommend action movies to the customers who come to rent videos. Second, clerks should recommend comedy videos to customers who rented action movies before or whose favorite leisure activities are not dining or playing on-line games. The study suggests that the customers will rent more videos with only a simple reminder.

## 5. Conclusion

Performing data mining using decision tree algorithm and Apriori association rule algorithm, we suggested a marketing approach for a single store of a local chain of video rental stores. The store should recommend videos to customers based on their professions and leisure activities, increasing store performance. Using information techniques and data mining of customer data and rental records, video rental marketing strategies may be designed. By expanding the size and time period of the data on customer rentals, the precision of the research may be improved. This method also shows promise for application to similar retail situations.

## 6. References

- Agrawal, R. & Srikant, R. (1994). Fast algorithm for mining association rules, In Proceedings of the 20th International Conference on Very Large Data Bases, pp.487-499, Santiago, Chile
- Berry, M. J. A. & Linoff G. (1997). Data mining technique for marketing, sale, and customer support, Wiley Computer and Sons, Inc.
- Bazerman, M. H. (2001). Consumer research for consumers, Journal of Consumer Research, Vol. 27, No.4, pp.499-504.
- Berson, A.; Smith, S. & Thearling, K. (2001). Building data mining application for CRM, McGraw-Hill Inc, New York.
- Curt, H. (1995). The Devil's in the detail: techniques, tool, and applications for data mining and knowledge discovery-part 1, Intelligent Software Strategies, Vol. 6, No. 9, pp.1-15.
- Engel, J. F.; Roger, D. B. and David, T. K. (1982). Consumer behavior, 4th ed., Hwa-Tai Co, Taipei.
- Frawley, W. J.; Piatetsky-Shapiro, G. & Matheus, C. J. (1991). Knowledge discovery in database: an overview, Knowledge Discovery in Database, pp.1-27, AAAI/MIT Press, California.
- Fayyad, U. M.; Piatetsky-Shapiro, G. & Smyth, P. (1996). From data mining to knowledge discovery: an overview, Advances in knowledge Discovery and Data Mining, pp.1-34, MIT Press, California.
- Hand, D. J.; Blunt, G.; Kelly, M. G. & Adams, N. M. (2000). Data mining for fun and profit, Statistical Science, Vol.15, No. 2, pp.111-131.
- Peacock, P. R. (1998). Data mining in marketing: part 1, Marketing Management, Vol. 6, No.4, pp.8-18.
- Schiffman, L. G., & Kanuk, L. Lazar. (2004) Consumer behavior, Prentice-Hall Inc Englewood Cliffs, New Jersey.
- Wilkie William L. (1994). Consumer behavior, 3th ed., John Wiley & Sons, New York.
- UDNJOB.Com CO. LTD. (2006). <http://udnjob.com/fe/jobseeker/index.shtml>.

# A Novel Model for Global Customer Retention Using Data Mining Technology

Jie Lin and Xu Xu

*School of Economics and Management, Tongji University  
China*

## 1. Introduction

This chapter deals with how to use data mining technology to find interesting pattern, which can be organized for global customer retention. Customer relationship management (CRM) comprises a set of processes and enabling systems supporting a business strategy to build long term, profitable relationships with specific customers. Customer data and information technology (IT) tools shape into the foundation upon which any successful CRM strategy is built. Although CRM has become widely recognized as an important business strategy, there is no widely accepted definition of CRM. Parvatiyar (2001) defines CRM as the strategic use of information, processes, technology, and people to manage the customer relationship with the company across the whole customer life cycle. Kincaid (2003) defines CRM as a company approach to understanding and influencing customer behavior through meaningful communications in order to improve customer acquisition, customer retention, customer loyalty, and customer profitability. These definitions emphasize the importance of viewing CRM as a comprehensive process of retaining customers, with the help of business intelligence, to maximize the customer value to the organization.

According to (Swift 2001; Kim et al., 2003) CRM consists of four dimensions: Customer Identification, Customer Attraction, Customer Retention, and Customer Development. They share the common goal of creating a deeper understanding of customers to maximize customer value to the organization in the long term. Customer retention has a significant impact on enterprise profitability. Analyzing and understanding customer behaviors and characteristics are the foundation of the development of a competitive customer retention strategy, so as to acquire and retain potential customers and maximize customer value. Gupta et al. (2004) find that a 1% improvement in retention can increase enterprise value by 5%. As such, elements of customer retention include one-to-one marketing, loyalty programs and complaints management. One-to-one marketing refers to personalized marketing campaigns which are supported by analyzing, detecting and predicting changes in customer behaviors. Loyalty programs involve campaigns or supporting activities which aim at maintaining a long term relationship with customers.

Customer satisfaction is the central concern for customer retention. Customer satisfaction, which refers to the comparison of customer expectations with his or her perception of being satisfied, is the essential condition for retaining customers (Chen et al., 2005). Bolton and Ruth N. (1998) have established the positive effect of customer satisfaction on loyalty and

usage behavior. With comprehensive customer data, data mining technology can provide business intelligence to generate new opportunities. Brijs et al. (2004) and Wang et al. (2005) concern with the discovery of interesting association relationships, which are above an interesting threshold, hidden in local databases. These researches mainly aimed at regional customer retention.

Globalization is no longer choice for most marketers; any company with a Web site has instant global exposure. Players at all levels are being pulled into global marketing by customer interaction over the Web, but not all are prepared for it. The message to company is clear: If your business is incapable of handling global trade, you are missing the point of conducting business. Company needs an efficient approach to find the pattern of global customer retention. One of the most important reasons companies are failing to take advantage of new global marketing opportunities is that the effective pattern of customer retention needed to do is woefully lacking. That is especially problematic for companies wishing to implement the modern analytical and targeting techniques of global customer retention (Matthew et al., 2006). For example, success with global customer retention requires familiarity with cultural practices. A database of Korean customers needs to include not only date-of-birth but also another data element not usually included in western databases-the wedding anniversary date of customers so the company can send the expected card. These main problems of global customer retention can be summarized as follows:

- Customer retention on a global scale demands the ability to apply theory with systems that reflects local cultures, local attitudes, and the real needs of individual customers in each market.
- Customer retention on a global scale demands to identify the sources of customer data and ascertain that all data was obtained in a manner that complies with privacy laws.
- Customer retention on a global scale demands to judge the values of min-support and min-confidence when utilizing data mining technique to discovery the interesting pattern.
- Customer retention on a global scale demands to avoid the signals that sent by customers are noise and confusing.

This chapter presents a comprehensive and effective model to solve the facing problems of customer retention shown above. This chapter is organized as follows. Customer retention overview is described in section 2. Sections 3 to 4 describe our novel model for global customer retention. This proposed model combines data mining technology with intuitionistic fuzzy set theory (IFS),  $\alpha$ -cuts, and expert knowledge to discovery the interesting pattern of global customer retention (sections 3); some definitions for global customer retention are defined in section 4. An example according to the proposed model is expatiated in section 5. Following sections show our experimental results of the proposed approach and model (section 6), discuss its properties and conclude (section 7) the chapter.

## **2. Customer retention overview**

### **2.1 Customer retention**

At the heart of any contractual or subscription-oriented business model is the notion of the retention rate. An important managerial task is to take a series of past retention numbers for a given group of customers and project them into the future in order to make more accurate predictions about customer tenure, lifetime value, and so on. Churn refers to the tendency

for customers to defect or cease business with a company. Marketers interested in maximizing lifetime value realize that customer retention is a key to increasing long-run enterprise probability. A focus on customer retention implies that enterprises need to understand the determinants of customer churn and are able to predict those customers who are at risk of defection at a particular point in time. Customer churn is the loss of existing customers to a competitor. The phenomenon has the potential to result in considerable profit loss for a company. As such the prevention of customer churn is a core.

Given the importance of customer retention, companies use a variety of mechanisms for reducing churn. These efforts can be grouped into three main areas: improving service quality, targeting interventions to prevent churn, and loyalty programs.

Firms' investment in improving service quality and customer satisfaction is based on the assumption that they improve customer retention. While some studies have found a link between satisfaction and retention (Rust & Zahorik, 1993), others have questioned this link. For example, Mittal & Kamakura (2001) find the link between customer satisfaction and retention to be moderated by customer characteristics.

Recent research finds that retention rates are affected by the channel utilized by the customer. Ansari et al. (2004) find that e-mails tend to drive persons to the Internet, and that purchases on the Internet lessen inertia in buying and loyalty. They conjecture that this arises from lower service levels and lower switching costs. Zhang & Wedel (2004) find the opposite effect in the context of grocery purchases, perhaps due to the use of e-shopping lists, which might actually raise switching costs. In light of these conflicting findings it would be desirable to better ascertain the role of optimal channel mix in retention.

Since the introduction of frequent flier program by American Airlines in the 1980s, loyalty programs have become ubiquitous in almost every industry. The interest in loyalty programs has increased over time as more and more companies use them for developing relationships, stimulating product or service usage, and retaining customers (Kamakura et al., 2000).

In spite of the pervasiveness of loyalty programs, their effectiveness is far from clear. Some studies find that loyalty programs increase customer retention and customer development (Bolton et al., 2000; Leenheer et al., 2004; Verhoef, 2003; Lixia Du et al., 2008); others find no impact on retention but improvement in share of wallet (Sharp & Sharp, 1997); and yet others find almost no difference in the behavior of loyalty program members and non-members (Dowling & Uncle, 1997). Kopalle and Neslin (2003) investigate the economic viability of frequency reward programs in a competitive environment, and find brands benefit from reward programs when customer's value future benefits, reward programs expand the market and if the brand has a higher preference.

Optimal targeting of loyalty programs is also an open issue. Conventional wisdom suggests that loyalty programs should be designed to reward a firm's best customers. However, Lal & Bell (2003) found that, in the context of grocery stores, loyalty programs do not affect the behavior of best customers. Instead, these programs have the biggest impact on a store's worst customers.

Several questions pertain to loyalty program design, including whether rewards should use cash or merchandise, offer luxury or necessity goods, be probabilistic or deterministic, or whether to use the firm's own products. Recent behavioral research provides some guidelines on these important issues (Kivetz, 2003; Kivetz & Simonson, 2002), and these findings have implications for modeling loyalty program design.

## 2.2 Data mining and customer retention

Data mining combines the statistic and artificial intelligence to find out the rules that are contained in the data, letters, and figures. The central idea of data mining for customer retention is that data from the past that contains information that will be useful in the future. Appropriate data mining tools, which are good at extracting and identifying useful information and knowledge from enormous customer databases, are one of the best supporting tools for making different customer retention decisions. There are many methods of data mining including classification, estimation, prediction, clustering, and association rules. Among these, association rules can discover the high frequency pattern and discover which things appear frequently and simultaneously.

Within the context of customer retention, data mining can be seen as a business driven process aimed at the discovery and consistent use of profitable knowledge from organizational data. Each of the customer retention elements can be supported by different data mining models, which generally include association, classification, clustering, forecasting, regression, sequence discovery.

- **Association:** Association aims to establishing relationships between items which exist together in a given record. Market basket analysis and cross selling programs are typical examples for which association modeling is usually adopted. Common tools for association modeling are statistics and apriori algorithms.
- **Classification:** Classification is one of the most common learning models in data mining. It aims at building a model to predict future customer behaviors through classifying database records into a number of predefined classes based on certain criteria. Common tools used for classification are neural networks, decision trees and if then-else rules.
- **Clustering:** Clustering is the task of segmenting a heterogeneous population into a number of more homogenous clusters. It is different to classification in that clusters are unknown at the time the algorithm starts. In other words, there are no predefined clusters. Common tools for clustering include neural networks and discrimination analysis.
- **Forecasting:** Forecasting estimates the future value based on a record's patterns. It deals with continuously valued outcomes. It relates to modeling and the logical relationships of the model at some time in the future. Demand forecast is a typical example of a forecasting model. Common tools for forecasting include neural networks and survival analysis.
- **Regression:** Regression is a kind of statistical estimation technique used to map each data object to a real value provide prediction value. Uses of regression include curve fitting, prediction (including forecasting), modeling of causal relationships, and testing scientific hypotheses about relationships between variables. Common tools for regression include linear regression and logistic regression.
- **Sequence discovery:** Sequence discovery is the identification of associations or patterns over time. Its goal is to model the states of the process generating the sequence or to extract and report deviation and trends over time. Common tools for sequence discovery are statistics and fuzzy set theory.

## 2.3 Customer retention models

Many customer retention models are concerned churn models. Many aspects of churn have been modeled in the literature. First, whether churn is hidden or observable influence the overall approach to modeling. In some industries, customer defection is not directly

observed, as customers do not explicitly terminate a relationship, but can become inactive. In other industries, however, the defection decision is observable as customers cease their relationship via actively terminating their contract with the firm (Schmittlein et al., 1987; Fader et al., 2004).

Models that are better at explanation may not necessarily be better at prediction. The empirical literature in marketing has traditionally favored parametric models (such as logistic or regression or parametric hazard specifications) that are easy to interpret. Churn is a rare event that may require new approaches from data mining, and non-parametric statistics that emphasize predictive ability (Hastie et al., 2001). These include projection-pursuit models, jump diffusion models, neural network models, tree structured models, spline-based models such as Generalized Additive Models (GAM), and Multivariate Adaptive Regression Splines (MARS), and more recently approaches such as support vector machines and boosting (Lemmens & Croux, 2003).

The above models are mainly concerned on technology. To the real markets, more and more customer retention models integrate human nature and culture factors with technology to be built. The following models consider human nature and culture factors.

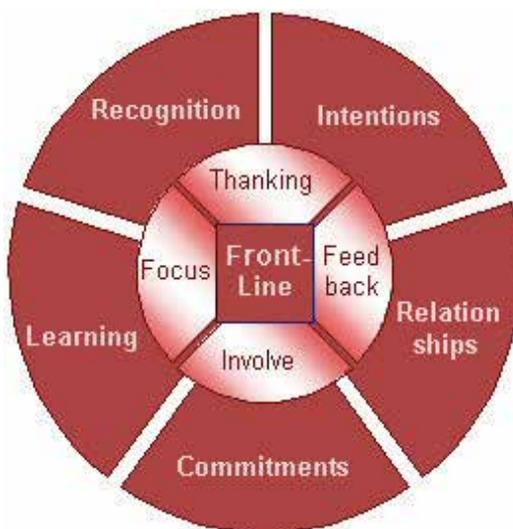


Fig. 1. Customer retention model architecture

This model (see Fig.1) is a bottom-up or user-centered approach based on leading workplace relationships and experiences to compliment your existing management of them. The model is based on a best practice combining two critical processes, or sets of actions, of cultures and human nature. Those processes being asking and thanking. Handling customer concerns, both complements as well as complaints, by asking the questions to get everyone on the service team involved. Retention customers by demonstrating intentions to continue to better serve them and increase value--beginning with those providing products and services. Working backwards, with the support of middle managers, to compliment and provide continuation for existing managerial efforts--especially for the "people" part of the enterprise. Customers may not remember what the employees said but they will always remember what how you made them feel. In other words, best practice is important to customers (see URL: <http://thankingcustomers.com/model.html>).

We proposed a model on customer retention in 2007 (shown in Fig.2). The model contains two functions, i.e. classification and making policies parts. In the classification part, firstly, it constructs a database which concludes a number of satisfaction factors, then it forms a new database based on the original database by combining intuitionistic fuzzy set theory and  $\alpha$ -cuts (see in section 3). Secondly, it gets the churn probability and classifies the customers into different groups. In the making policies part, it employs data mining technique to find the interesting pattern and association rules to each customer group. This is then used to create appropriate policies for different customer group. The most significant feature of this model is that it not only predicts churning but also makes proactive attempts to reduce it.

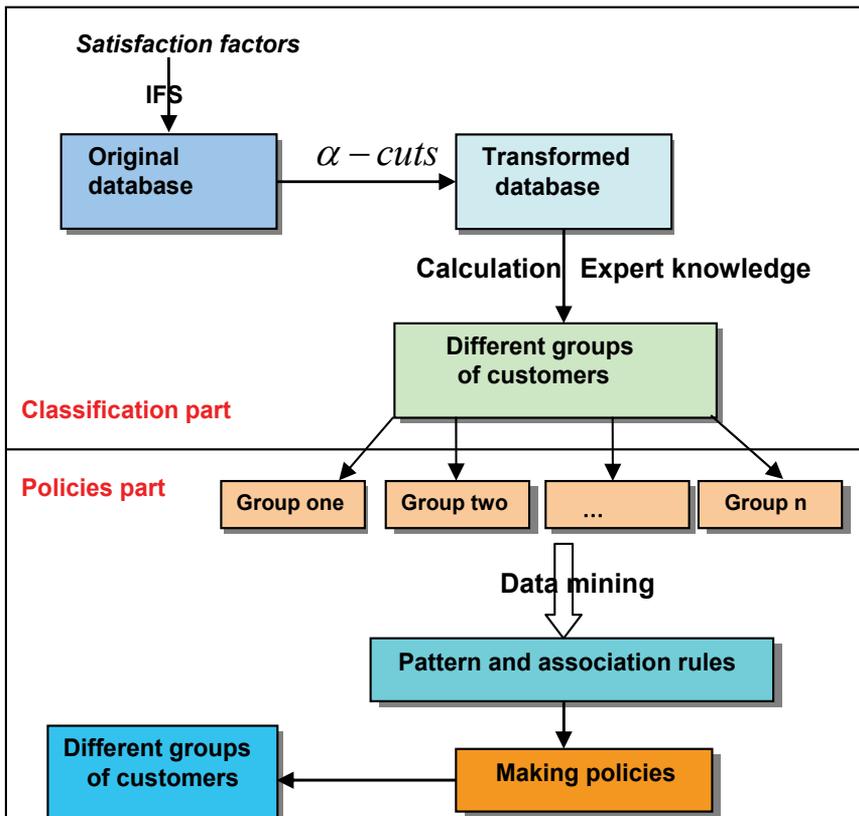


Fig. 2. Customer retention model-CP Model

Although exist a lot of customer retention models, there is a lack of comprehensive and effective approach and model to realize global customer retention by now. CP model has potential merits (such as churn probability, customer behavior character), but it cannot satisfy global customer retention problem. We combine potential merits of CP model to propose a novel global customer model (see section 3).

### 3. Global customer retention model

As the nature of research in global customer retention, data mining technology is difficult to confine to specific disciplines. In this section, intuitionistic fuzzy set theory,  $\alpha$ -cuts, expert knowledge, and CP model are combined for the proposed global customer retention model.

### 3.1 Proposed global customer retention model

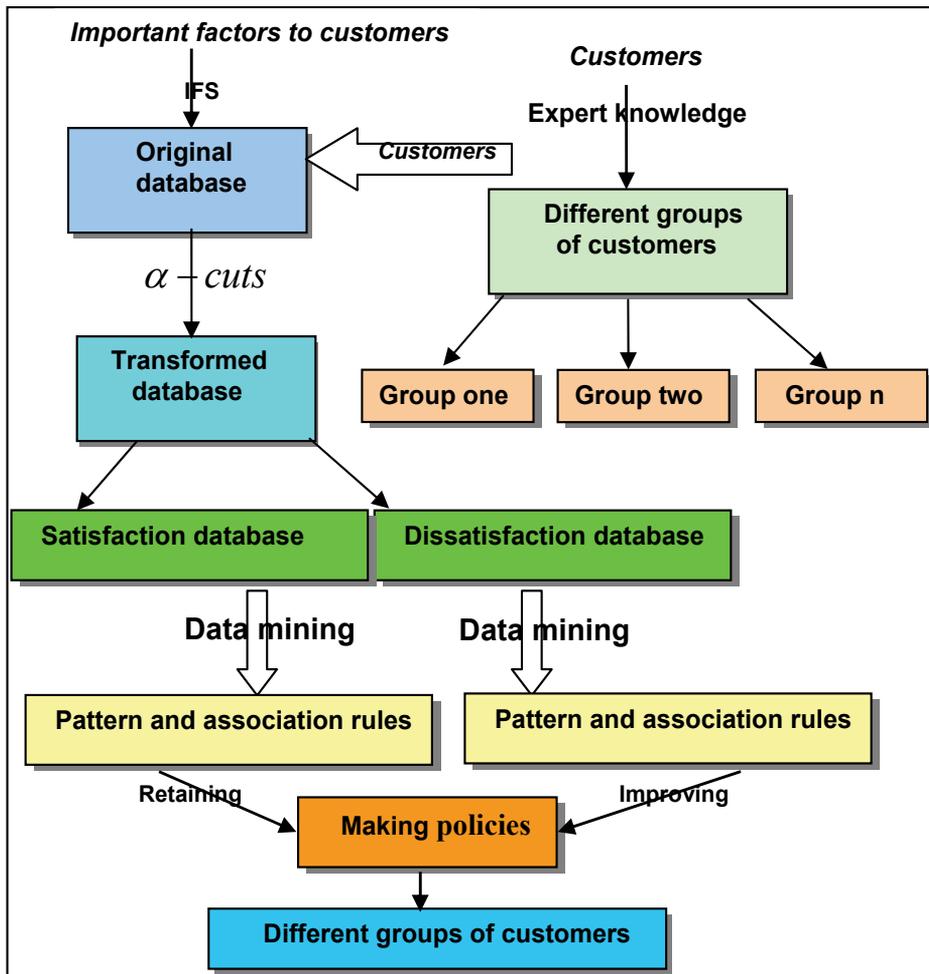


Fig. 3. Proposed global customer retention model-GCP Model

We proposed GCP Model (see Fig.3). Intuitionistic fuzzy set theory,  $\alpha$ -cuts, and data mining technology are all employed in CP Model and GCP Model. The main differences between CP Model and GCP Model are that the function of expert knowledge and original customer database. GCP Model includes dissatisfaction database.

### 3.2 Intuitionistic fuzzy set theory

Fuzzy set theory was firstly presented by Professor L.A.Zadeh in California University in 1965. It transforms the meaning and spoken description into fuzzy set instead of general set, then studies and deals with subjective and undefined data with membership functions, qualifies the data then transforms it into useful information through systemic fuzzy operations.

Intuitionistic fuzzy set (IFS) is intuitively straightforward extensions of L.A.Zadeh's fuzzy sets. IFS theory basically defies the claim that from the fact that an element  $x$  "belongs" to a

given degree to a fuzzy set  $A$ , naturally follows that  $x$  should “not belong” to  $A$  to the extent  $1-\mu$ , an assertion implicit in the concept of a fuzzy set. On the contrary, IFS assign to each element of the universe both a degree of membership  $\mu$  and one of non-membership  $\nu$  such that  $\mu + \nu \leq 1$ , thus relaxing the enforced duality  $\nu=1-\mu$  from fuzzy set theory. Obviously, when  $\mu + \nu = 1$  for all elements of the universe, the traditional fuzzy set concept is recovered (Atanassov, 1986). IFS owe its name to the fact that this latter identity is weakened into an inequality, in other words: a denial of the law of the excluded middle occurs, one of the main ideas of intuitionism.

In CP and GCP Model, intuitionistic fuzzy set theory is an extension of fuzzy set theory that defies the claim that from the fact that an element  $x$  belongs to a given degree  $\mu_{A(x)}$  to a fuzzy set  $A$ , naturally follows that  $x$  should not belong to  $A$  to the extent  $1-\mu_{A(x)}$ , an assertion implicit in the concept of a fuzzy set. On the contrary, IFS assigns to each element  $x$  of the universe both a degree of membership  $\mu_{A(x)}$  and one of non-membership  $\nu_{A(x)}$  such that

$$\mu_{A(x)} + \nu_{A(x)} \leq 1 \quad (1)$$

Thus relaxing the enforced duality  $\mu_{A(x)} + \nu_{A(x)} = 1$  from fuzzy set theory. Obviously, when  $\mu_{A(x)} + \nu_{A(x)} = 1$  for all elements of universe, the traditional fuzzy set concept is recovered. Here, IFS builds the bridge between customer satisfaction and retention. The most important virtue is that IFS can express the customers’ psychology and behavior exactly.

### 3.3 $\alpha$ -cuts

An element  $x \in X$  that typically belongs to a fuzzy set  $A$ , when its membership value to be greater than some threshold  $\alpha \in [0, 1]$ . The ordinary set of each element is the  $\alpha$ -cut  $A_\alpha$  of  $A$ :

$$A_\alpha = \{x \in X, \mu_{A(x)} \geq \alpha\} \quad (2)$$

Didier Dubois and Henri Prade (2001) also define the strong  $\alpha$ -cut:

$$A_\alpha = \{x \in X, \mu_{A(x)} > \alpha\} \quad (3)$$

CP and GCP Model all employ formula (2) to transform the original database.

The membership function of a fuzzy set can be expressed in terms of the characteristic function of its  $\alpha$ -cuts according to the following expressions:

- $\mu_{A(\alpha)} = 1$  iff  $x \in A_\alpha$
- Otherwise  $\mu_{A(\alpha)} = 0$

With the help of  $\alpha$ -cuts, the attribute values of database can be transformed to 0 or 1.

### 3.4 Database architecture

At the core of any sizable global trading effort is the database of information needed to guide and drive customer interactions through locally appropriate customer retention strategies. Customer satisfaction is the essential condition for retaining customers. Well-developed customer information base will smooth a path to the customer; a poorly assembled core of data might well compound global trading difficulties beyond repair. Customer retention is to stay abreast of data privacy rule making. The notion of privacy-preserving data mining is to identify and disallow such revelations as evident visible to the third parties in the kinds of patterns learned using traditional data mining techniques. So the essential factors of data privacy preserving are considered during the research. From the

architecture of the customer retention database, some expressions can be organized as the attributes of database (shown in Table 1).

Description of factors	Satisfaction and dissatisfaction reply
Enterprise offers competitive price (D <sub>1</sub> )	[0,1.0]
Enterprise is abreast of developing new products (D <sub>2</sub> )	[0,1.0]
Complains are taken by enterprise's employees (D <sub>3</sub> )	[0,1.0]
It is easy to get enterprise's contact with the right person at call center (D <sub>4</sub> )	[0,1.0]
The employees at enterprise's center are competent and professional (D <sub>5</sub> )	[0,1.0]
The enterprise's sales representative understands the enterprise (D <sub>6</sub> )	[0,1.0]
The enterprise's sales representative is competent and has profound knowledge (D <sub>7</sub> )	[0,1.0]
The enterprise offers gifts to customers in special days (D <sub>8</sub> )	[0,1.0]
The enterprise's society responsibility (D <sub>9</sub> )	[0,1.0]

Table 1. The attributes of database

### 3.5 Expert knowledge

In CP Model, expert knowledge and percentage of customer satisfaction are combined to classify the customers. Experts of this field are employed to confirm the boundary of customer satisfaction. For example:

- 0 1 0 1 0 1 1 1 0,  $P=5/9=55.6\%$
- 1 1 1 1 1 1 0 0 0,  $P=6/9=66.7\%$
- 1 1 1 1 1 1 1 1 0,  $P=8/9=88.9\%$
- 1 1 1 1 1 1 1 1 1,  $P=9/9=100\%$

The customers can be divided into different groups according to the percentage of customer satisfaction and expert knowledge.

- Group one:  $P < 60\%$
- Group two:  $60\% \leq P < 80\%$
- Group three:  $80\% \leq P < 90\%$
- Group four:  $P = 100\%$

In GCP Model, with regard to the levels of different customers, expert knowledge used to set the weights of the customers, such as customer categories combined with data privacy preserving principle, age, gender characters and regions. The value of  $\alpha$  is related to the weights of customers, in other words, the weights of customers satisfy different criterion  $K_j$ , with  $j=1, 2, \dots, n$ ,  $\alpha = \max(K_j)$ .

## 4. Proposed definitions

This section introduces the following definitions to discover the pattern of global customer retention.

- Definition 1 An association rule has the form  $X \rightarrow Y$ , where  $X \subset I$ ,  $Y \subset I$ ,  $I$  is a transaction set, and  $X \cap Y = \Phi$ .
- Definition 2 The support of the association rule  $X \rightarrow Y$  is the probability that  $X \cup Y$  exists in a transaction in the database  $D$ .
- Definition 3 The min-support is defined as follows,  $\text{min-support} = C \cdot (N/P)$ , where  $C$  denotes the weight of the region;  $N$  denotes the number of a special group of customers in the research;  $P$  denotes the number of all customers is involved in the research.
- Definition 4 The confidence of the association rule  $X \rightarrow Y$  is the probability that  $Y$  exists that a transaction contains  $X$ , i.e.,  $\Pr(Y/X) = \Pr(X \cup Y) / \Pr(X)$ .

## 5. Example of GCP model

Practicing global customer retention demands the ability to apply strategies in a way that reflect local cultures, local attitudes, and the real needs of individual customers in each market. GCP Model is expatiated by an example in this section.

### 5.1 Steps of GCP model

The steps of GCP Model are described below:

- Construct the database of customer retention with support of IFS.
- Ascertain the value of  $\alpha$  according to different group of customers and  $\alpha = \max(K_j)$ .
- Select the satisfaction reply whose value is one to form transaction tables with the help of  $\alpha$ -cuts; select the dissatisfaction reply whose value is zero to form transaction tables with the help of  $\alpha$ -cuts.
- Discover the pattern of customer retention combined with data mining technique and expert knowledge.

### 5.2 An example

In order to explain the proposed GCP Model and find the pattern of customers under globalization, the section takes the data (age:  $A_1$  (20-29); gender:  $M_2$ ) from the cooperative enterprise to set an example. Here we just show the example of satisfaction pattern based on GCP Model.

- Step one: Table 2 shows the original database according to IFS.
- Step two:  $M_2=0.6$ ,  $A_1=0.3$ , according to  $\alpha = \max(K_j)$ ,  $\alpha=0.6$ . Table 3 shows the transformed table with the help of  $\alpha$ -cuts.

$D_1$	$D_2$	$D_3$	$D_4$	$D_5$	$D_6$	$D_7$	$D_8$	$D_9$
0.3	0.6	0.8	0.5	0.4	0.3	0.5	0.6	0.8
0.3	0.4	0.6	0.4	0.3	0.2	0.4	0.5	0.9
0.5	0.3	0.5	0.4	0.2	0.4	0.4	0.4	0.7
0.4	0.3	0.4	0.5	0.2	0.3	0.3	0.4	0.6
0.5	0.6	0.3	0.5	0.1	0.4	0.3	0.6	0.5
0.6	0.7	0.3	0.4	0.1	0.3	0.3	0.3	0.6
0.6	0.3	0.2	0.4	0.1	0.2	0.2	0.5	0.3
0.6	0.3	0.2	0.3	0.2	0.3	0.4	0.6	0.6
0.6	0.3	0.1	0.3	0.1	0.3	0.3	0.6	0.5
0.7	0.2	0.1	0.3	0.2	0.3	0.4	0.7	0.6

Table 2. Original database

D <sub>1</sub>	D <sub>2</sub>	D <sub>3</sub>	D <sub>4</sub>	D <sub>5</sub>	D <sub>6</sub>	D <sub>7</sub>	D <sub>8</sub>	D <sub>9</sub>
0	1	1	0	0	0	0	1	1
0	0	1	0	0	0	0	0	1
0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	1
0	1	0	0	0	0	0	1	0
1	1	0	0	0	0	0	0	1
1	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	1	1
1	0	0	0	0	0	0	1	0
1	0	0	0	0	0	0	1	1

Table 3. Transformed database

- Step three: Select the satisfaction reply whose value is one to form transaction database. The transaction database is shown in table 4. Each transaction is a subset of I, and is assigned a transaction identifier (TID).

TID	Items	TID	Items	TID	Items	TID	Items
1	D <sub>2</sub> D <sub>3</sub> D <sub>8</sub> D <sub>9</sub>	2	D <sub>3</sub> D <sub>9</sub>	3	D <sub>9</sub>	4	D <sub>9</sub>
5	D <sub>2</sub> D <sub>8</sub>	6	D <sub>1</sub> D <sub>2</sub> D <sub>9</sub>	7	D <sub>1</sub>	8	D <sub>1</sub> D <sub>8</sub> D <sub>9</sub>
9	D <sub>1</sub> D <sub>8</sub>	10	D <sub>1</sub> D <sub>8</sub> D <sub>9</sub>				

Table 4. Transaction database

- Step four: Calculate the appearing times of every transaction item and show in Table 5.
- Step five: Acquire the support of every appearing item, D<sub>1</sub>=0.5, D<sub>2</sub>=0.3, D<sub>3</sub>=0.2, D<sub>8</sub>=0.5, D<sub>9</sub>=0.7.

Item	Appearing times	Item	Appearing times	Item	Appearing times
D <sub>1</sub>	5	D <sub>2</sub>	3	D <sub>3</sub>	2
D <sub>8</sub>	5	D <sub>9</sub>	7		

Table 5. The appearing times of every item

- Step six: Compare the support of every appearing item with min-support (0.3). It is obviously to find frequent item set I<sub>1</sub>, I<sub>2</sub>. I<sub>1</sub>= {D<sub>1</sub>, D<sub>2</sub>, D<sub>8</sub>, D<sub>9</sub>}, I<sub>2</sub>= {D<sub>1</sub>D<sub>8</sub>, D<sub>8</sub>D<sub>9</sub>}.
- Step seven: The min-confidence is say, 60%, and then the association rules are shown in table 6.

Association rules	Confidence	Association rules	Confidence
D <sub>1</sub> →D <sub>8</sub>	60%	D <sub>8</sub> →D <sub>1</sub>	60%
D <sub>8</sub> →D <sub>9</sub>	60%		

Table 6. Association rules

However, to women customers whose age is between 20 and 29, the satisfaction patterns are presented below:

- If customers value the competitive price offered by the enterprise, they will like the gifts provided by the enterprise in special days.
- If customers like the gifts provided by the enterprise in special days, they will value the competitive price offered by the enterprise.
- If customers like the gifts provided by the enterprise in special days, they will concern social responsibility of the enterprise.

Dissatisfaction patterns of customers are also can be achieved by the same approach. Finally, the achieved patterns and rules are established for the decision makers.

## 6. Experiments analysis

The study of global customer retention under GCP Model with the cooperative enterprises started in 2007. Although the experiments of proposed approach are still in the early stage, some interesting pattern and rules are achieved in the research. For instance, the proposed approach can increase response rate of the customer loyalty by segmenting customers into groups with different characteristics; the proposed approach can predict how likely an existing customer is to take his/her business to a competitor. It is interesting to note that, dissatisfaction patterns of customers are mainly related to the enterprise call center. Owing to the different culture and location, the satisfaction patterns are almost different. For example, some (especially for western customers) care about the enterprise responsibility; some (especially for eastern customers) care about the gifts provided by the enterprise. All of these useful pattern and rules help the decision makers to make effective policies.

## 7. Conclusions

Study of global customer retention is an emerging trend in the industry. This chapter has proposed a novel model (GCP Model) to practice global customer retention. It aims to find the useful pattern with the combination of IFS,  $\alpha$ -cuts, expert knowledge, and data mining technique. This proposed model might have some limitations, for instance, some sensitive data of customers are not involved in it, and this will be the future research direction. With respect to the research findings, in order to maximize an organization's profits, policy makers have to retain valuable customers and increase the life-time value of the customer. As such, global customer retention is so important to maintaining a long term and pleasant relationship with customers according to different cultures and locations.

## 8. Acknowledgements

The research is supported by the National High-Tech. R & D Program for CIMS, China (No.2007AA04Z151), Program for New Century Excellent Talents in University, China (No.NCET-06-0377), and the National Natural Science Foundation China (No.70531020).

## 9. References

- Anderson, Eric. & Duncan S. (2004). Long-run effects of promotion depth on new versus established customers: Three field studies, *Marketing Science*, Vol.23, No.1, pp.4–20.

- Atanassov, K.T. (1986). Intuitionistic fuzzy sets, *Fuzzy Sets and Systems*, Vol.20, pp.87-96.
- Bolton & Ruth N. (1998). A dynamic model of the duration of the customer's relationship with a continuous service provider: the role of satisfaction, *Marketing Science*, Vol.17, pp.45-65.
- Bolton, Ruth, P. K. Kannan & Matthew B. (2000). Implications of loyalty program membership and service experiences for customer retention and value, *Journal of the Academy of Marketing Science*, Vol.28, No.1, pp.95-108.
- Brijs, T.; Swinnen, G.; Vanhoof, K. & Wets, G. (2004). Building an association rules framework to improve product assortment decisions, *Data Mining and Knowledge Discovery*, Vol.8, pp.7-23.
- Chen, M. C.; Chiu, A. L. & Chang, H. (2005). Mining changes in customer behavior in retail marketing, *Expert Systems with Applications*, Vol.28, pp.773-781.
- Didier Dubois & Henri P. (2001). *Fuzzy Set Theory and its Applications*, Kluwer-Nijhoff Publishing, London.
- Dowling, Grahame R. & Mark U. (1997). Do customer loyalty programs really work? *Sloan Management Review*, Vol.38, No.4, pp.71-82.
- Fader, P. S.; B. G. S. Hardie & K. L. Lee. (2004). Counting your customers' the easy way: An alternative to/NBD model, Working Paper, Wharton Marketing Department.
- Gupta, Sunil, Donald R. Lehmann & Jennifer Ames Stuart. (2004). Valuing Customers, *Journal of Marketing Research*, Vol.41, No.1, pp.7-18.
- Hastie, T.; R. Tibshirani, & J. Friedman. (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, New York: Springer-Verlag.
- Kincaid, J. W. (2003). *Customer Relationship Management: Getting it Right*, Prentice Hall PTR, New Jersey.
- Lal, Rajiv & David E. Bell. (2003). The impact of frequent shopper programs in grocery retailing, *Quantitative Marketing and Economics*, Vol.1, No.2, pp.179-202.
- Leenheer, Jorna, Tammo H. A. Bijmolt, Harald J. & Ale Smidts. (2004). Do loyalty programs enhance behavioral loyalty? A Market-wide analysis accounting for endogeneity, Working Paper, Tilburg University.
- Lixia Du, Xu Xu, Yan Cao & Jiyang L. (2008). A novel approach to find the satisfaction pattern of customers in hotel management, *Proceedings of IEEE International Conference on Fuzzy Systems*, Hongkong, pp.11-14.
- Kamakura, Wagner A. & Michel W. (2000). Factor analysis and missing data, *Journal of Marketing Research*, Vol.37, pp. 490-498.
- Kim, E. ; Kim, W. & Lee, Y. (2003). Combination of multiple classifiers for the customer's purchase behavior prediction, *Decision Support Systems*, Vol.34, pp.167-175.
- Kivetz, Ran. (2003). The effects of effort and intrinsic motivation on risky choice, *Marketing Science*, Vol.22, pp.477-502.
- Kivetz, R. & I. Simonson. (2002). Earning the right to indulge: Effort as a determinant of customer preferences toward frequency program rewards, *Journal of Marketing Research*, Vol.39, No.2, pp.155-170.
- Kopalle & Neslin. (2003). The economic viability of frequency reward programs in a strategic competitive environment, *Review of Marketing Science*, Vol.1, pp.163-172.
- Lemmens, Aurelie & Christophe C. (2003). Bagging and boosting classification trees to predict churn, Working Paper, Teradata center.

- Matthew B. Gersper & Randell C. (2006). Creating a competitive advantage in global trade, *Global Data Mining*.
- Mittal, Vikas & Wagner A. K. (2001). Satisfaction, repurchase intent and repurchase behavior: Investigating the moderating effect of customer characteristics, *Journal of Marketing Research*, Vol.38, pp.131-142.
- Parvatiyar, A. J. N. (2001). Customer relationship management: Emerging practice, process, and discipline, *Journal of Economic and Social Research*, Vol.3, pp.1-34.
- Rust, Roland & Anthony Z. (1993). Customer satisfaction, customer retention, and market share, *Journal of Retailing*, Vol.69, No.2, pp.33-48.
- Schmittlein, David, Donald Morrison, & Richard C. (1987). Counting your customers: Who are they and what will they do next? *Management Science*, Vol.33, No.1, pp.32-46.
- Sharp Byron & Anne Sharp. (1997). Loyalty programs and their impact on repeat-purchase loyalty patterns, *International Journal of Research in Marketing*, Vol.14 , pp.473-86.
- Swift, R. S. (2001). *Accelerating Customer Relationships: Using CRM and Relationship Technologies*, Prentice Hall PTR, New Jersey.
- Verhoef, Peter C. (2003). Understanding the effect of CRM efforts on customer retention and customer share development, *Journal of Marketing*, Vol.67, No.4, pp.30-45.
- Wang, K. ; Zhou, S. ; Yang, Q. & Yeung, J. M. S. (2005). Mining customer value: From association rules to direct marketing, *Data Mining and Knowledge Discovery*, Vol.11, pp.58-79.
- Zhang, Jie & Michel W. (2004). The effectiveness of customized promotions in online and offline stores, Working Paper, University of Michigan Business School.

# Data Mining in Web Applications

Julio Ponce<sup>1</sup>, Alberto Hernández<sup>2</sup>, Alberto Ochoa<sup>4,5</sup>, Felipe Padilla<sup>3</sup>,  
Alejandro Padilla<sup>1</sup>, Francisco Álvarez<sup>1</sup> and Eunice Ponce de León<sup>1</sup>

<sup>1</sup>*Aguascalientes University,*

<sup>2</sup>*CIICAp-UAEM,*

<sup>3</sup>*UQAM,*

<sup>4</sup>*Juarez City University*

<sup>5</sup>*CIATEC*

<sup>1,2,4,5</sup>*México*

<sup>3</sup>*Canada*

## 1. Introduction

The World Wide Web is rapidly emerging as an important medium for commerce as well as for the dissemination of information related to a wide range of topics (e.g., business and government). According to most predictions, the majority of human information will be available on the Web. These huge amounts of data raise a grand challenge, namely, how to turn the Web into a more useful information utility (Garofalakis et al., 1999).

At the moment with the popularity of Internet, people are exhibited to a lot of information that is available for study. Nowadays there is also a great amount of applications and services that are available through Internet as they are seeking, chats, sales, etc., nevertheless much of that information is not useful for many people, but in the area of Data Mining, all the information available in the Internet represents a work opportunity and it is possible to do a lot of analysis on the basis of these with specific purposes.

Knowledge Discovery and Data Mining are powerful data analysis tools. The rapid dissemination of these technologies calls for an urgent examination of their social impact. We show an overview of these technologies. The terms “Knowledge Discovery” and “Data Mining” are used to describe the ‘non-trivial extraction of implicit, previously unknown and potentially useful information from data (Wahlstrom & Roddick, 2000). Knowledge discovery is a concept that describes the process of searching on large volumes of data for patterns that can be considered knowledge about the data. The most well-known branch of knowledge discovery is data mining.

### 1.1 Data mining

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential. Data mining is a knowledge discovery process in large and complex data sets, refers to extracting or “mining” knowledge from large amounts of data. Moreover, data mining can be used to predict an outcome for a given entity (Hernández et al., 2006).

Thus clustering algorithms in data mining are equivalent to the task of identifying groups of records that are similar between themselves but different from the rest. (Varan, 2006).

Data mining is a multidisciplinary field with many techniques. With this techniques you can create a mining model that described the data that you will use. (see Fig. 1).

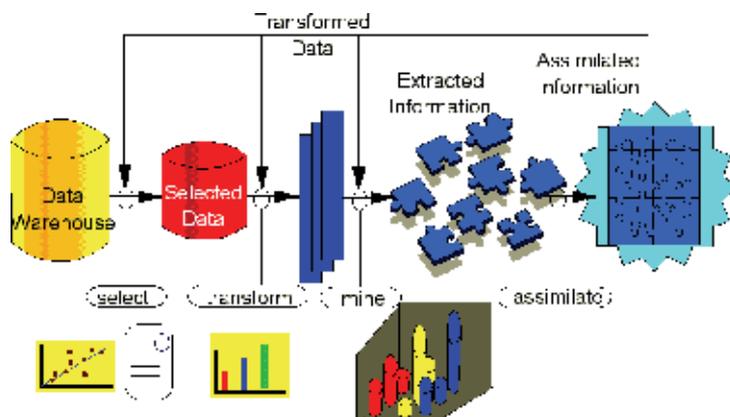


Fig. 1. Data Mining Process

Some elements in Data Mining Process are:

- A. **Data Set.** It is a collection of data, usually presented in tabular form. Each column represents a particular variable. Each row corresponds to a given member of the data set in question. It lists values for each of the variables, such as height and weight of an object or values of random numbers. Each value is known as a datum. The data set may comprise data for one or more members, corresponding to the number of rows.
- B. **Pre-processing.** Data mining requires substantial pre-processing of data. This is especially the case of the behavioural data. To make the data comparable, all data needs to be normalized.
- C. **General Results.** This activity is related to overall assessment of the effort in order to find out whether some important issues might have been overlooked. This is the step where a decision upon further steps has to be made. If all previous steps were satisfactory and results fulfil problem objectives, the project can move to its conclusive phase.
- D. **Decision Trees.** Decision trees are powerful and popular tools for classification and prediction. The attractiveness of decision trees is due to the fact that, these decision trees represent rules. Decision tree induction is a typical inductive approach to learn knowledge on classification. The key requirements to do mining with decision trees are: Attribute-value description, Predefined classes, Discrete classes and Sufficient data.
- E. **Association Rules.** Association rules describe events that tend to occur together. They are formal statements in the form of  $X \Rightarrow Y$ , where if  $X$  happens,  $Y$  is likely to happen (Márquez et al., 2008).

## 1.2 Weka

Weka (Waikato Environment for Knowledge Analysis) is a collection of algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java program. This contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization (Witten & Frank 2005). Weka was developed at the University of Waikato in New Zealand, is an open source software issued under the GNU General Public License. The Data Mining process with Weka includes: reading the

.arrf file in the Weka Explorer, proceeding to classify, visualize clusters and discover associations in the data. Start the hidden patterns finding, remember to keep mind open (No prejudices).

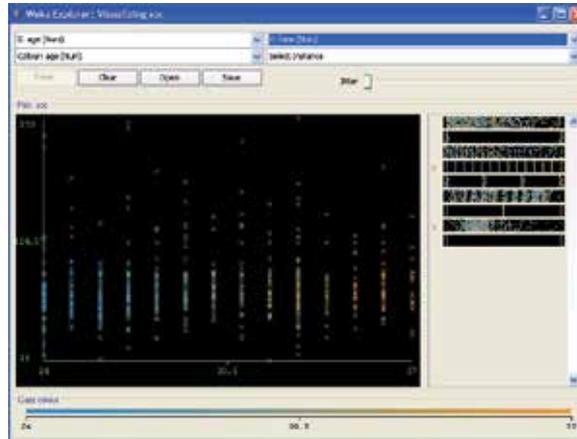


Fig. 2. Visualizing some data in Weka Explores

### 1.3 Web mining

Usually, web mining is categorized as web content mining and web usage mining. The first studies the search and retrieval of information on the web, while the second discovers and analyzes user's access pattern (Xu et al., 2003). A knowledge discovery tool, WebLogMiner, is discussed in (Zaiane et al., 1998), which uses OLAP and data mining techniques for mining web server log files. In (Mobasher et al, 2000), a web mining frame-work which integrated both usage and content attributes of a site is described. Some techniques based on clustering and association rules are proposed. In (Yao & Yao, 2003; Yao, 2003) presents a framework and information retrieval techniques to support individual scientists doing research.

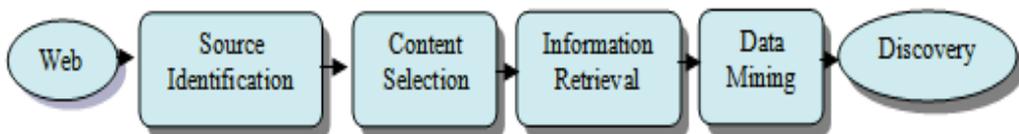


Fig. 3. Framework for Web Data Minig

### 1.4 Social data mining

Social data mining is a new and challenging aspect of data mining. It is a fast-growing research area, in which connections among and interactions between individuals are analyzed to understand innovation, collective decision making, problem solving, and how the structure of organizations and social networks impacts these processes. Social data mining includes various tasks such as the discovery of communities, searching for multimedia data (images, video, etc) personalization, search methods for social activities (find friends), text mining for blogs or other forums. Social data mining finds several applications; for instance, in e-commerce (recommender systems), in multimedia searching (high volumes of digital photos, videos, audio recordings), in bibliometrics (publication patterns) and in homeland security (terrorist networks).

Social data mining systems enable people to share opinions and obtain a benefit from each other's experience. These systems do this by mining and redistributing information from computational records of social activity such as Usenet messages, system usage history, citations, and hyperlinks among others. Two general questions for evaluating such systems are: (1) is the extracted information valuable? , and (2) do interfaces based on extracted information improve user tasks performance?.

Social data mining approaches seek analogous situations in the computational world. Researchers look for situations where groups of people are producing computational records (such as documents, Usenet messages, or web sites and links) as part of their normal activity. Potentially useful information implicit in these records is identified, computational techniques to harvest and aggregate the information are invented, and visualization techniques to present the results are designed. Figure 4. Shows a traditional Data mining process. Thus, computation discovers and makes explicit the "paths through the woods" created by particular user communities. And, unlike ratings-based collaborative filtering systems (Hill & Terveen, 1996)., social data mining systems do not require users to engage in any new activity; rather, they seek to exploit user preference information implicit in records of existing activity. The "history-enriched digital objects" line of work (Resnick et al., 1994) was a seminal effort in this approach. It began from the observation that objects in the real world accumulate wear over the history of their use, and that this wear – such as the path through the woods or the dog-eared pages in a paperback book or the smudges on certain recipes in a cookbook – informs future usage. Edit Wear and Read Wear were terms used to describe computational analogies of these phenomena. Statistics such as time spent reading various parts of a document, counts of spreadsheet cell recalculations, and menu selections were captured. These statistics were then used to modify the appearance of documents and other interface objects in accordance with prior use. For example, scrollbars were annotated with horizontal lines of differing length and color to represent amount of editing (or reading) by various users.

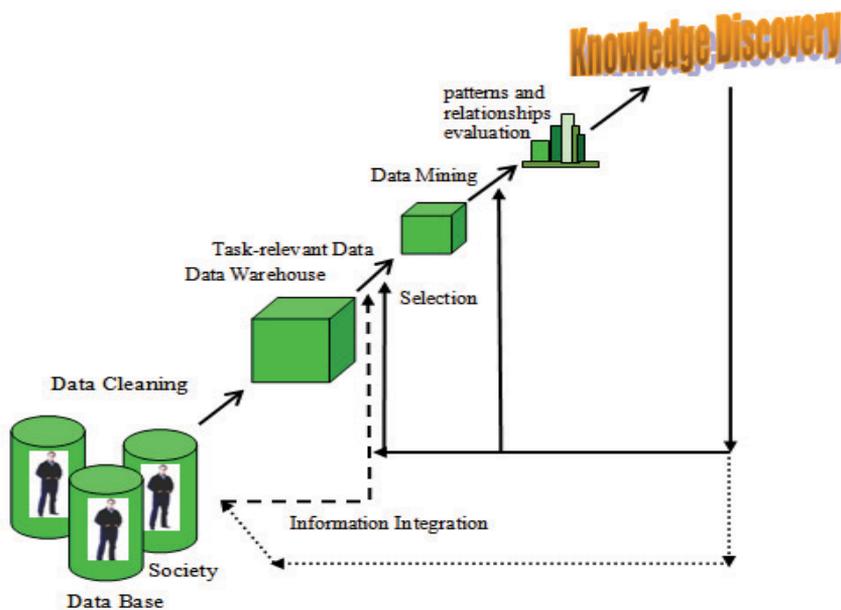


Fig. 4. A traditional Social Data Mining process (Ochoa, 2006).

The examples above mentioned are activities to which we are exposed and that without knowing we make use of the Data Mining, Due to this reason in the last years, Data Mining has had great advances in artificial intelligence in order to offer a better support to user task (Ochoa, 2006).

## 2. Social networks

Web communities have risen rapidly in recent years with benefits for different types of users. For individuals, the Web community helps the users in finding friends of similar interests, providing timely help and allowing them to share interests with each other. For commercial advertisers, they can exploit the Web community to find out what the users are interested on, in order to focus their targets. It would be straightforward to discover the Web community if we had the detailed and up-to-date profiles of the relations among Web users. However, it is not easy to obtain and maintain the profiles manually. Therefore, the automatic approaches in mining users' relationship are badly needed.

Social network describes a group of social entities and the pattern of inter-relationships among them. What the relationship means varies, from those of social nature, such as values, visions, ideas, financial exchange, friendship, dislike, conflict, trade, kinship or friendship among people, to that of transactional nature, such as trading relationship between countries. Despite the variability in semantics, social networks share a common structure in which social entities, generically termed actors, are inter-linked through units of relationships between a pair of actors known as: tie, link, or pair. By considering as nodes and ties as edges, social network can be represented as a graph.

A social network is a social structure made of nodes (which are generally individuals or organizations) that are tied by one or more specific types of interdependency (See Fig. 5).

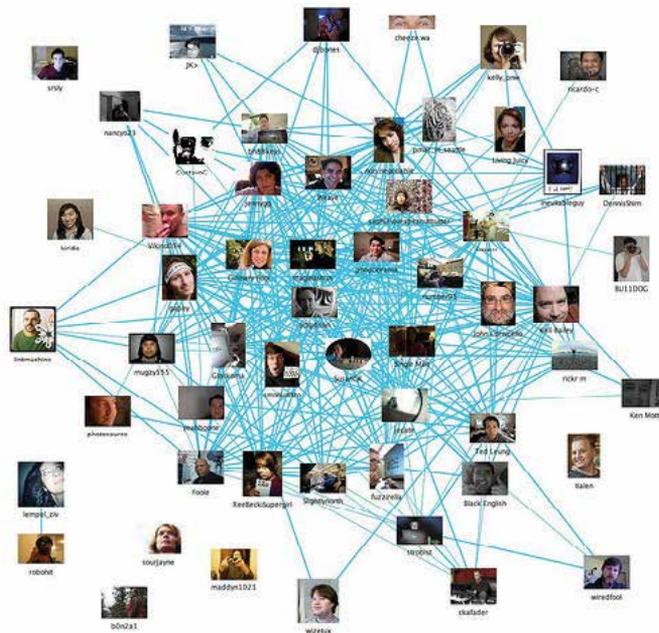


Fig. 5. Social Network Diagram.

### 2.1 Social networks analysis

Social Network Analysis (SNA) became a hot research topic after the seminal work by (Milgram, 1967). SNA is the study of mathematical models for relationships among entities such as people, organizations and groups in a social network. The relationships can be various. For example, they can be friendship, business relationship, and common interest relationship. A social network is often modelled by a graph, where the nodes represent the entities, and an edge between two nodes indicates that a direct relationship exists between the two entities. Some typical problems in SNA include discovering groups of individuals sharing the same properties (Schwartz & Wood, 1993) and evaluating the importance of individuals (Domingos & Richardson, 2001). Previously, the research in the field of SNA has emphasized binary interaction data, with direct and/or weighted edges (Lorrain & White, 1971) and focused almost exclusively on very small networks, typically, in the low tens of entities (Wasserman & Faust, 1994).

Moreover, only considering the connectivity properties of networks without leveraging the information of the entities limits the application of SNA.

Social network analysis has attracted much attention in recent years. Community mining is one of the major directions in social network analysis. Most of the existing methods on community mining assume that there is only one kind of relation in the network, and moreover, the mining results are independent of the users' needs or preferences. However, in reality, there exist multiple, heterogeneous social networks, each representing a particular kind of relationship, and each kind of relationship may play a distinct role in a particular task. Thus mining networks by assuming only one kind of relation may miss a lot of valuable hidden community's information and may not be adaptable to the diverse information needs from different users (Cai et al., 2005).

A social network can be analyzed for many useful insights. For instance, the important actors in the network, those with more connections, or the greatest influence, can be found. Alternatively, it may be the connection paths with actors that are of interest. Analysts may look for the shortest paths, or the most novel types of connections. Sometimes, the focus may even be on finding subgroups that are especially cohesive or interesting.

Knowledge of social networks is useful in various application areas. In law enforcement concerning organized crimes such as drugs and money laundering or terrorism, knowing how the perpetrators are connected, would assist the effort to disrupt a criminal act or to identify suspects. In commerce, viral marketing exploits the relationship between existing and potential customers to increase sales of products and services. Members of a social network may also take advantage of their connections to meet other members, for instance through web sites facilitating networking or dating among their users (Lauw et al., 2005).

### 3. Web radio

A research is detailed to acquire knowledge about how to develop a Web Radio using Social Data Mining and Cultural Algorithms (Reynolds, 1998), to a best functionality of it. Main thematic of the web radio is music to dance, that includes all the rates that consider equipment to dance, its directed to an ample segment of the society, whose only restriction is focused towards the different musical likes, any person who has desires to listen music and why not, to take advantage of it to dance, doesn't matter sex, age, civil state, nationality or many other factors, can access to our web counting on access to internet. Data Mining is very useful to any kind of projects, for that reason, we decided to use it inside, with this

web, users and developers can interact among them easily. We can help users using Data Mining in the creation of their lists of songstakers into account previous experiences with the same characteristics for new users, according to the classification that belongs to it according to the information that it provides in its user profile. In order to obtain that the user stay on line into our Web, we have many rewards for them, agreement with their localization and the number of hours that they stay on our site.

### **3.1 Web radio introduction**

Throughout history, the advance of the technology is in constant growth, one of the greatest discoveries has been the Internet that has facilitated the growth of other technologies as well as, thanks to this, we can practically be in contact with any people with the entire world, and ensure communications between the societies.

The radio has been another mass media between the people, also it has had changes through the time. The radio was one of first mass media with which the society had contact, by means of them emitted news, music and soap opera radio. The network has supposed a significant change in the way of transmission of this media, and has caused the birth of stations that they exclusively emit through them.

### **3.2 Problem outline**

Why do you think that the radio has turned upside down so much with Internet? Because thanks to different services (World wide web, electronic mail, the news, internet, chat, among others) of internet, it is possible to undergo with other forms of information and expression that go beyond the wireless sound and to incorporate, therefore, new contents. In addition, also is feasible to generate new forms of consumption and relation that a listener can have with means (Hill & Terveen, 1996), (Ochoa et al., 2006).

In order to be able to implement the Data Mining it was decided to do a web radio using diverse tools that the new technology provides, as well as different software types to facilitate the work in its creation, in order to practice and to know more about the behavior of the human beings before this type of technologies. The design and development of the Web Radio called "Wave Radio" allowed increasing knowledge in different areas from science and technology. The creation of the Web Radio is a tool That was designed to investigate the user's behaviour of the same. Also to be able to integrate user groups (clusters) according to a stable classification that explains the tastes, characteristics that they share to each other.

In a traditional interactive application there is not factor during the design and development process (Brooks, 1994). If a web radio is considered as an interactive application on line, then it requires of human factors coming from Human Computer-Interaction (HCI) area (Nielsen & Loranger, 2006).

## **4. Security in web applications**

With the rapid growth of interest in the Internet, network security has become a major concern to companies throughout the world. The fact that the information and tools needed to penetrate the security of corporate networks are widely available has increased that concern. Data mining has been loosely defined as the process of extracting information from large amounts of data. In the context of security, the information we are seeking is the

knowledge of whether a security breach has been experienced, and if the answer is yes, who is the perpetrator (Barbará & Jajodia 2002). Among well known criminals are:

- A. **Hackers.** Hackers are criminals who try to break into your computer system and steal your personal information or cause other troubles.
- B. **Online advertising impostors.** Online marketing techniques may be used to trick you or your family into doing something that may have a negative outcome.
- C. **Online predators.** These are usually adults who are interested in grooming children online for their own sexual pleasure.
- D. **Identity theft.** Criminals can steal your personal information and pretend they are you for financial benefit.

Other types of online crime also exist where people can obtain a financial advantage illegally. Because of this increased focus on network security, network administrators often spend more effort protecting their networks than on actual network setup and administration. Tools that probe for system vulnerabilities, such as the Security Administrator Tool for Analyzing Networks (SATAN), and some of the newly available scanning and intrusion detection packages and appliances, assist in these efforts, but these tools only point out areas of weakness and may not provide a means to protect networks from all possible attacks. Thus, as a network administrator, you must constantly try to keep abreast of the large number of security issues confronting you in today's world. Data Mining in Web Security concentrates heavily in the area of intrusion detection.

Private information can reside in two states on a network. It can reside on physical storage media, such as a hard drive or memory, or it can reside in transit across the physical wired or wireless network in the form of packets. These two information states present multiple opportunities for attacks from users on your information, as well as those users on the Internet. We are primarily concerned with the second state, which involves network security issues. The following are five common methods of attack that present opportunities to compromise the information on your network:

- Network packet sniffers
- IP spoofing
- Password attacks
- Distribution of sensitive internal information to external sources
- Man-in-the-middle attacks

When protecting your information from these attacks, your concern is to prevent the theft, destruction, corruption, and introduction of information that can cause irreparable damage to sensitive and confidential data. According (Barbará & Jajodia 2002) the use of data mining is based on two important issues. First, the volume of data dealing with both network and host activity is so large that it makes it an ideal candidate for using data mining techniques. Second, intrusion detection is an extremely critical activity.

#### **What is Network Intrusion Detection?**

Intrusion detection starts with instrumentation of a computer network for data collection. Pattern-based software 'sensors' monitor the network traffic and raise 'alarms' when the traffic matches a saved pattern. Security analysts decide whether these alarms indicate an event serious enough to warrant a response. A response might be to shut down a part of the network, to phone the internet service provider associated with suspicious traffic, or to simply make note of unusual traffic for future reference. Intrusion detection systems are software and/or hardware components that monitor computer systems and analyze events occurring in them for signs of intrusions (Kumar et al., 2005).

Data mining based intrusion detection techniques generally fall into one of two categories; misuse detection and anomaly detection. In misuse detection, each instance in a data set is labelled as 'normal' or 'intrusion' and a learning algorithm is trained over the labelled data. A key advantage of misuse detection techniques is their high degree of accuracy in detecting known attacks and their variations. Their obvious drawback is the inability to detect attacks whose instances have not yet been observed (Dokas et al., 2002). Anomaly detection, on the other hand, builds models of normal behaviour, and automatically detects any deviation from it, flagging the latter as suspect. Anomaly detection techniques thus identify new types of intrusions as deviations from normal usage (Javitz & Valdes 1993)

#### 4.1 Derived data for intrusion detection

A single connection between an outside machine and a single port on a machine inside your network is not malicious – unless it is part of a series of connections that attempted to map all the active ports on that machine. For this reason you will want to add additional fields containing values from the base. Example, you could distinguish traffic originating from outside your network from traffic originating inside your network. Another type of derived data, called an aggregation, is a summary count of traffic matching some particular pattern. Example, we might want to know, for a particular source IP X, and a particular IP Y, how many unique destinations IP were contacted in a specific time window Z. A high value of this measure could give an indication of IP mapping, which is a pre-attack reconnaissance of the network. Aggregations are generally more expensive to compute than other kinds of derived data that are based upon only a single record. A third type of derived data is a flag indicating whether a particular alarm satisfies a heuristic rule. Because data mining methods handle many attributes well, and because we don't know for sure which one will be useful, our approach is to compute a large number of attributes (over one hundred) and store them in the database with the base alarm fields (Bloedorn et al., 2001).

Due to widespread diversity and complexity of computer infrastructures, it is difficult to provide a completely secure computer system. There are numerous security and intrusion detection systems that address different aspects of computer security. Below we present a common architecture of intrusion detection systems and its basic characteristics.

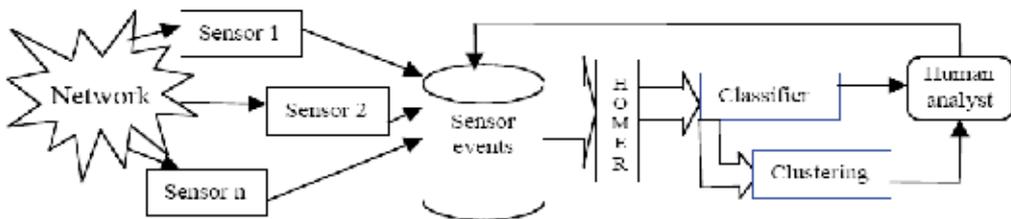


Fig. 6. How sensors feed into overall intrusion detection system (Bloedorn et al., 2001)

The figure above shows a proposed configuration to perform intrusion detection. First, network traffic is analyzed by a variety of available sensors. This sensor data is pulled periodically to a central server for conditioning and input to a relational database.

Before security specialists can start providing input to the data mining effort, this traffic must be filtered. It is a straightforward task to create a filter that can find these patterns within a data table of traffic. At figure 6, this preliminary filter is called HOMER (Heuristic

for Obvious Mapping Episode Recognition). The heuristic operates on aggregations by source IP, destination port, and protocol and then check to see if a certain threshold of destination IPs were hit within a time window. If the threshold is crossed, an incident is generated and logged to the database.

Even though the bulk traffic due to the mapping activity is not shown to the analyst, the source host itself is placed on the radar screen of our system. Please note that some normal activity (e.g., name servers, proxies) within an organization's intranet can match the profile of an IP mapping. HOMER handles this situation by means of an exclusion list of source IPs. HOMER filters events from the sensor data before they are passed on to the classifier and clustering analyses. Data mining tools filter false alarms and identify anomalous behaviour in the large amounts of remaining data. A web server is available as a front end to the database if needed, and analysts can launch a number of predefined queries as well as free form SQL queries from this interface. The goal of this operational model is to have all alarms reviewed by human analysts.

Catching new attacks can not depend on the current set of classification rules. Since classification assumes that incoming data will match that seen in the past, classification may be an inappropriate approach to finding new attacks. K-means clustering is used to find natural groupings of similar alarm records. Records that are far from any of these clusters indicate unusual activity that may be part of a new attack. Finally, we discussed a hot subject on web security: preserving privacy data mining.

#### **4.2 Preserving Privacy Data Mining (PPDM)**

What is privacy?

In (Shoeman, 1984) was defined privacy as "the right to determine what (personal) information is communicated to others" or "the control an individual has over information about himself or herself." More recently, Garfinkel (Garfinkel, 2001) stated that "privacy is about self-possession, autonomy, and integrity."

Another view is corporate privacy - the release of information about a collection of data rather than an individual data item. For example: "I may not be concerned about someone knowing my birth date, mother's maiden name, or social security number; but knowing all of them enables identity theft" (Clifton et al., 2004).

#### **4.3 Knowledge discovery and privacy**

When people talk of privacy, they say "keep information about me from being available to others". However, their real concern is that their information not be misused. The fear is that once information is released, it will be impossible to prevent misuse. Utilizing this distinction - ensuring that a data mining project won't enable misuse of personal information - opens opportunities that "complete privacy" would prevent (Clifton et al. 2004). The key finding is that knowledge discovery can open new threats to informational privacy and information security if not done or used properly (Oliveria & Zaiane, 2004).

Privacy is viewed as a social and cultural concept. However, with the ubiquity of computers and the emergence of the Web, privacy has also become a digital problem (Rezgur et al. 2003). With the Web and the emergence of data mining, privacy concerns have posed technical challenges different from those that occurred before the information era.

In data mining the definition of privacy preservation is still an unclear topic. A notable exception is the work presented in (Clifton et al. 2002), in which PPDM is defined as

"getting valid data mining results without learning the underlying data values." According to (Oliveria & Zaiane, 2004) PPDm encompasses the dual goal of meeting privacy requirements and providing valid data mining results. This definition emphasizes the dilemma of balancing privacy preservation and knowledge disclosure.

## 5. Internet frauds

Fraud is the crime or offense of deliberately deceiving another in order to damage them usually, to obtain property or services unjustly. Fraud can be accomplished through the aid of forged objects. In the criminal law of common law jurisdictions it may be called "theft by deception", "larceny by trick," "larceny by fraud and deception" or something similar.

Sentinel Top Complaint Categories January 1 - December 31, 2007			
Rank	Top Categories	Complaints	Percentage
1	Identity Theft	258,427	32%
2	Shop-at-Home/Catalog Sales	62,811	8%
3	Internet Services	42,266	5%
4	Foreign Money Offers	32,868	4%
5	Prizes/Sweepstakes and Lotteries	32,162	4%
6	Computer Equipment and Software	27,036	3%
7	Internet Auctions	24,376	3%
8	Health Care	16,097	2%
9	Travel, Vacations and Timeshare	14,903	2%
10	Advance-Fee Loans and Credit Protection/Repair	14,342	2%
11	Investments	13,705	2%
12	Magazines and Buyers Clubs	12,970	2%
13	Business Opps and Work-at-Home Plans	11,362	1%
14	Real Estate (Not Timeshare)	9,475	1%
15	Office Supplies and Services	9,211	1%
16	Telephone Services	8,155	1%
17	Employ Agencies/Job Counsel/Overseas Work	5,932	1%
18	Debt Management/Credit Counseling	3,442	<1%
19	Multi-Level Mktg/Pyramids/Chain Letters	3,092	<1%
20	Charitable Solicitations	1,843	<1%

Table 1. Internet Frauds made in January-December 2007

Internet fraud generally refers to any type of fraud scheme that uses one or more online services - such as chat rooms, e-mail, message boards, or Web sites - to present fraudulent solicitations to prospective victims, to conduct fraudulent transactions, or to transmit the proceeds of fraud to financial institutions or to others connected with the scheme. Unfortunately, people who engage in fraud often operate in "Internet time" as well. They seek to take advantage of the Internet's unique capabilities -- for example, by sending e-mail messages worldwide in seconds, or posting Web site information that is readily accessible

from anywhere in the world - to carry out various types of fraudulent schemes more quickly than was possible with many fraud schemes in the past.

The types of Internet Fraud, in general, the same types of fraud schemes that have victimized consumers and investors for many years before the creation of the Internet are now appearing online (sometimes with particular refinements that are unique to Internet technology). With the explosive growth of the Internet, and e-commerce in particular, online criminals try to present fraudulent schemes in ways that look, as much as possible, like the goods and services that the vast majority of legitimate e-commerce merchants offer. In the process, they not only cause harm to consumers and investors, but also undermine consumer confidence in legitimate e-commerce and the Internet. There are some of the major types of Internet fraud that law enforcement and regulatory authorities and consumer organizations are seeing in the USA:

- Auction and Retail Schemes Online. This type of fraudulent schemes appearing on online auction sites are the most frequently reported form of Internet fraud. These schemes, and similar schemes for online retail goods, typically offer high-value items - ranging from items that are likely to attract many consumers. These schemes induce their victims to send money for the promised items, but then deliver nothing or only an item far less valuable than what was promised (e.g., counterfeit or altered goods).
- Business Opportunity/"Work-at-Home" Schemes Online. Fraudulent schemes often use the Internet to advertise purported business opportunities that will allow individuals to earn thousands of dollars a month in "work-at-home" ventures. These schemes typically require the individuals to pay a guaranteed amount of money, but fail to deliver the materials or information that would be needed to make the work-at-home opportunity a potentially viable business.
- Identity Theft and Fraud. Some Internet fraud schemes also involve identity theft - the wrongful obtaining and using of someone else's personal data in some way that involves fraud or deception, typically for economic gain.
- Investment Schemes Online
  - Market Manipulation Schemes. Enforcement actions by the Securities and Exchange Commission and criminal prosecutions indicate that criminals are using two basic methods for trying to manipulate securities markets for their personal profit. First, in so-called "pump-and-dump" schemes, they typically disseminate false and fraudulent information in an effort to cause dramatic price increases in thinly traded stocks or stocks of shell companies (the "pump"), then immediately sell off their holdings of those stocks (the "dump") to realize substantial profits before the stock price falls back to its usual low level. Second, in short-selling or "scalping" schemes, the scheme takes a similar approach, by disseminating false or fraudulent information in an effort to cause price decreases in a particular company's stock.

There are other Schemes of Internet Fraud besides before mentioned (<http://www.usdoj.gov/criminal/fraud/internet/>). The fraud detection is becoming increasingly important in revealing and limiting revenue loss due to fraud. Fraudsters aim to use services without paying or illicitly benefit from the service in other ways, causing service providers financial damage. To reduce losses due to fraud, one can deploy a fraud detection system. However, without tuning and through testing, the detection system may cost more in terms of human investigation of all the false alarms than the gain from

reduction of fraud. Test data suitable for evaluating detection schemes, mechanisms and systems are essential to meet these requirements. The data must be representative of normal and attack behaviour in the target system since detection systems can, and should, be very sensitive to variations in input data.

## 6. Diverse applications and domains where analysis through data mining

In this section we show some application related to the topics before mentioned.

### 6.1 Social networks application

Orkut is a system of social networks used in Brazil by 13 million users, many of them, create more of a profile, and generate different relationships from their different profiles, this takes more to think that they develop Bipolar Syndrome, to be able to establish communications with people of different life styles, and when they doing to believe other users that they are different people (Zolezzi-Hatsukimi, 2007).

Some problems in the social network of Orkut exist that dislike much to their users. The loss of privacy, the lack of materialization of the relations established through the network. The false profiles are created for: to make a joke, to harass other users, or to see who visualizes its profile. As the profile is false, the friends of this profile are also generally false, making difficult the tracking of the original author. The users can make denunciations against those false profiles, but without clear the profile of Orkut. Anyway, the original author can create a new profile. Often a user does not wish to exhibit his photo in Orkut and put a photo of a celebrity. That is more common and is accepted by the community for those who wishes to remain anonymous. In this case, it uses his real name and places photos for the album, but with a photo of any profile.

The most serious event in Orkut is the creation of communities with a concept of racism, xenophobia and mistreat against the animals and making vindication to the consumption and sale of drugs and pedofilia. Unfortunately, the users who denounce these facts to be eliminated do not reach their objective: the criminals create these data again, deceiving to the server of Orkut. Nevertheless, due to pressures on the part of the Brazilian government and of the American press, new actions on the matter of this were announced on the part of the server, in a definitively effective action.

Many critics are made against Orkut. The main one is the libertinism of that place, can be spoken on racism, murders, without nothing happens, if the person knows to hide it. The moderation in Orkut is of ear, nevertheless is not sufficient. The police tries to find guilty of certain crimes, since some of them agree all through Orkut. To some users its own profile is kidnapped, rob password of email, MSN or banking accounts.

Using the tool of Data Mining denominated WEKA, it was come to develop a de-nominated Model "Ahankara" of prediction of profiles in users of Orkut, which allows to understand the motivations of this type of profile and to determine if it has generated Syndrome Bipolar, to see figure 3 (Ponce et al., 2007).

The model obtained Ahankara once used WEKA to look for the relations that us could be of utility to process the data. see Figure 7.

Some of the relations that obtained when observing the data with aid of the WEKA are the following ones.

- The region has to do, with the number of fans.

- The number of communities is based on the number of people with interests in common, some factors can take part like sex and civil state.
- Regions exist in which there is people in all the communities, regions exist in which the people of that region this in a single community.
- The people who participate in many communities, the majority have less fans, so that in theory they spend long time in entering all the communities to which they have themselves you incorporate.
- The Region if it influences in the number of communities which the people enter, due to the activities who are used to doing has certain interests in common in each region.

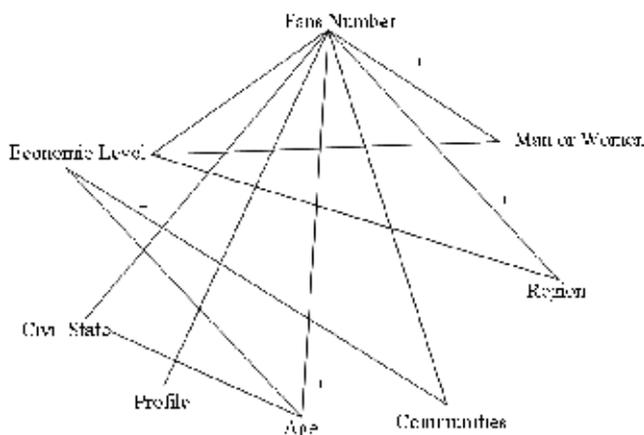


Fig. 7. Ahankara Model, This is the propose model (Ponce et al., 2007).

Through Mining Data it is possible to find relations between the data, often these can be hidden and others are evident, another type of analyses that can be realised are especially the groups of people by some type of characteristics, this can be realised through diverse techniques of clusters or heuristic technical using as is the used in (Ponce et al., 2006) where the clique maximum problem is solved, which can used to find the clusters with major number to relationship in bases of a certain characteristic.

## 6.2 Web radio application

### 6.2.1 Data mining and cultural algorithms for develop the intelligent web radio

We purpose to develop web radio applications taken into account in particular social acceptability factor using Social data mining and cultural algorithms. By so doing, we purpose a conceptual model which include the operation of the web radio, everything begins at the moment in which the user creates his profile, after for creating his data are incorporate in a data base in which all the profiles are stored, to be able to make the analysis with them, so that if at some time a user with different preferences arrives and that he does not share with the other users already registered it stores a new case in the case library so that in the future it can be reused and help the users to create their lists of songs on the basis of its profile. And this process is repeated whenever a new user enters his/her information; this is shown in figure 3.

When people listen to music they do not like, their initial reaction is to fast forward, followed by changing moods if they do not hear acceptable music within a reasonable

number of fast forwards. We believe that people appreciate having these two options. This makes our selection mechanism different from radio. Using in the web radio with the cultural algorithms, it could improve the performance because it is possible to motivate a society with different preferences and analyze the user preferences. With the cultural algorithms is possible analyze what kind of music is someone listening, our Web Radio can deduce the songs, the singers and the genders the person prefers, and by using this information recommend additional songs (see Figure 8).

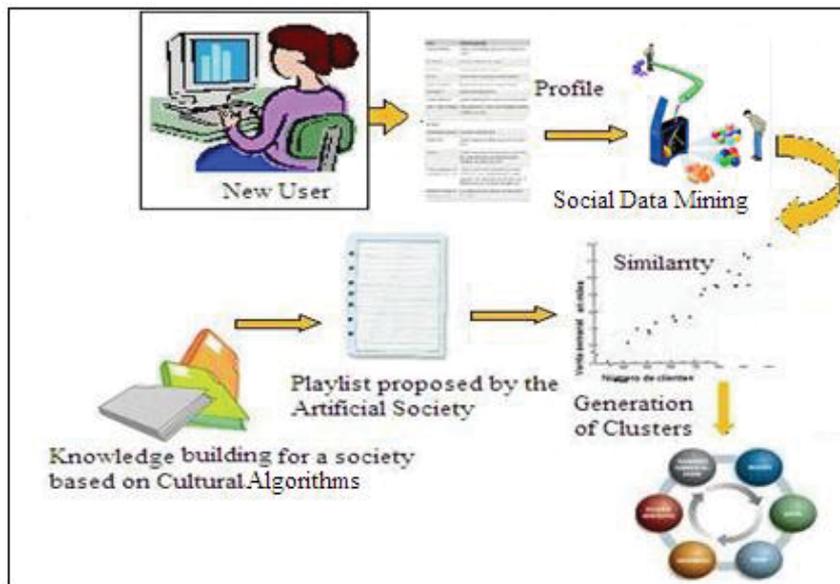


Fig. 8. Design and development of a Web radio based on Social Data Mining and cultural algorithms.

We made a system that allows users to view individual and group historical listening lists and define his new lists. These systems learn of the user preferences, after songs are selected to be played on a shared physical environment, based on the preferences of the whole people present, similar to Tibetan Avenue (<http://tibetanavenue.com/>).

According to users profile, They are classified of a way in which those that occupies that group feel identified with other users who perhaps are not of the same country but who share other characteristics such as; age, musical sort, zodiacal sign. When grouping of this form to the users groups with an almost equal personality. One says almost because they do not share all the characteristics, but in his majority they have the same pleasures. In the social network there is one or two people whom a greater number of features with other users and there shares it is where it is the base of the social network.

An extra on which it tells to the Web Radio is that it has prizes of reward for the users who spend major time in tuning, this according to the region where is being in tune and also vary with the participation that has within the Web Radio. For it every month related to the musical sorts will be realized one trivia on which it tells to the Web Radio. If the user has a major number of options to answer correctly, he/her will be made creditor have access to unload a video of steps to learn to dance, the users only can answer one trivia per week.

### 6.2.2 A thematic web radio application

In general a music delivery system is classified in two broad categories, content purchasing and audio broadcasting. In the case of content purchasing, the consumer pays for specific music (e.g. CDs, cassette tapes, LPs) to build up a collection of personal favorites over which he/she has complete control. Broadcast audio (e.g. radio, TV, Internet radio) provides more content, but at the cost of limited consumer control (consumers choose radio stations, not music). Personalized audio music attempts to bridge the gap between the two. Our hypothesis is that by matching information about the users (listener profiles) with the knowledge building for a society based on cultural algorithms (content metadata), it should be possible to automatically generate more pertinent playlists for individual listeners. In this paper we test this hypothesis.

Thematic web radio is a web application developed using the programming language PHP by means of the Macromedia family was developed the graphical interface so that it is more efficient, pleasant and easy the handling of the same for the user (Field et al., 2001) .

Within the diverse functions that realize the system is had as primary target the handling, administration, search and analysis of users, profiles and tastes, among others features, similar at previous research, which was developed a prototype (Ochoa et al., 2007).

One of the functions that support in the good operation of the web radio is, by means of the use of a profile, the user can provide data of personal interests to the creators of the Web such as: sex, age, date of birth, e-mail, areas of interest, among others, and so it is possible to be grouped to the users according to the characteristics that share to generate clusters.

The registered users can have access to the creation of a play list, at the time of which the user creates his/her list, of the songs on which he/she tells to the data base of the Web Radio.

The list is stored in a data base and this can be used by other users, since, to give pursuit to the social network, we must have like minimum ten cases, so that we pruned to follow with the implementation of the Social Data Mining (see Figure 9).

Archivo Ver Control Ayuda

Play List

Cancion	Interprete	Duracion
Si tu amor no vuelve	Chetes	3.04
Nookie	Lim Bizkit	4.05
Amaral	Sabe	5.23
Its psycho	Infected	5.55
La duranguería	Celeste	3.43
Eso que onda?	Chale	2.43
Pa la banda	Barrabas	4.12

Tiempo de Escucha: 45:36 minutos restantes

Cancion	Interprete	Duracion
La duranguería	Celeste	3.43
Nookie	Lim Bizkit	4.05
Pa la banda	Barrabas	4.12
Si tu amor no vuelve	Chetes	3.04

Cancion:

Intérprete:

Duración:  Mins

Fig. 9. Play list user interface.

Another function of the Web Radio is a finder that allows to locate songs, the searches are based on hierarchies such as title, interprets, among others. By means of the finder one facilitates the use and the permanence of the user.

Within the operation of the web radio is a function that allows that the users has power by deciding the ranking of the songs according to its preference, the scale that is used for the measurement is the "Lickert Scale" that is defined like a series of items or phrases that carefully have been selected, so that they constitute a valid criterion, trustworthy and precise to measure of some form the social phenomena.

The functions before mentioned allow having a better control and thus we can give a pursuit to the social network to which our Web Radio is associated. The profiles are stored and with them we can compute the range between the people using their profiles (to look for the similarities) and in this way of specifying the limits of clusters (Users with similar preferences). We use the similarity function used in Social Data Mining to organize the people in different clusters:

$$\frac{\sum_{i=1}^n w_i \times \text{sim}(f_i^I, f_i^R)}{\sum_{i=1}^n w_i} \quad (1)$$

In where:

$w_i$  is the weight of importance about an attribute.

$\text{sim}$  is the function of similarity.

$f_i^I, f_i^R$  are the values of attribute  $i$  in the entrance profile (I) and the recover profile (R).

### 6.2.3 Experiments related with the artificial social net that support this web radio

We have worked with two scenarios, where web users use our thematic web radio. In the first scenario, we compared the performance of 27 communities of 50 agents, and on the other hand 27 communities of 500 agents each one. The optimal number of songs in the playlist was 17. One of the most interesting characteristics observed in this experiment is the diversity of cultural patterns established for each community. For the solution with the same number of songs the provided result for the "beliefspace" is entirely different. The structured scenarios associated to the agents cannot be reproduced in general due they belong to a given instant in the time and space.

They represent a unique, precise and innovative form of adaptive behavior which solves a computational problem followed by a complex change of relationships. The generated configurations can be metaphorically related to the knowledge of the community behavior regarding to an optimization problem (to decide the music related with the playlist), or a tradition with which to emerge from the experience and with which to begin a dynamics of the process [1]. Comparing the 50 agents of the first community regarding the 500 agents community, this last obtained a better performance in terms of the average songs (17.05 versus 18.30), as well as a smaller standard deviation (1.96 versus 2.64). They also had a greater average number of changes in the paradigm (5.85 versus 4.25), which indicates that even the "less negotiating" generations. In the second experiment, we consider the same scenario for the experiment one, except that after having obtained a solution from a community of 50 agents. In this experiment, it was surprising to see initially how the community of 500 agents uses the solution offered by the 50 agents, whenever these

solutions were close the optimal grade, instead of finding entirely complete new solutions. This can be compared metaphorically with the concept of culture.

### **6.3 Using biometry and data mining in online assessments to detect who is there?**

This application is a clear example of the use of data mining on security, by means of the system described is possible to avoid impersonation as well as academic frauds in online assessments. Unless photo IDs are checked and all course work occurs inside of a monitored classroom, faculty really does not know for sure whether the student is who they say they are in the classroom or online (MSU, 2006). On online assessments in which we are not sure who is taking the test; students will be under pressure, some students perform unfairly poorly under pressure and this is a good incentive to cheat (Rove, 2004). We have a wide spectrum of documented techniques to commit cheat on online assessments: modify a grade in the database (DB), to steal answers for questions, to copy from another student or cheat sheets, impostor or substitute remote students, to search for answers on the Internet or in blogs or purchase the list of answer for an specific exam, on the messenger or cellular phone, in single words to “commit cheat” to obtain a “better grade” in an online assessment. Biometrics is becoming a powerful tool to improve security on transactions and reduce frauds (ITEDU, 2006).

An advanced security measure can be implemented by means of biometric technologies; much of the hot discussion about biometrics has come about due to the level of research and interest shown in large scale implementations of the technology by the US and UK Governments and the European Union (Clarke & Furnell 2005). They may provide added robustness in access control to high security facilities within higher education. As the unit price for biometric devices continues to fall is possible to employ these to replace the current systems used for workstation and network access (Wasniowski, 2005). These devices are likely to become a standard computer peripheral, built into future workstations.

#### **6.3.1 The problem**

The main problem on online assessments is to know who's there (Wisher et al. 2005). In this section, we propose the use of biometrics and data mining, particularly the use fingerprint recognition and web cam monitoring on real time to verify student's identity during online assessment; we propose also the analysis of the student's behavioural patterns by means of data mining to deal with the well-known problem of: who is taking the exam? The contribution of this paper is the use of hybrid technologies in online assessments as a new approach for remote identification of students on real time.

#### **6.3.2 Performance schema (3-tier client-server system)**

We separated the application in three main modules: the first one is on charge of the conduction the online assessment, the second one on charge of the fingerprint recognition and web cam monitoring on real time, the third on charge of data mining analysis on students behavioural patterns. Server must be in listening mode waiting for Clients that requires a service. In order to use fingerprint recognition, the first step is to enrol students – top, right side in Figure 10-, the students fingerprint is saved and indexed in the Features Database, we highly recommend to separate this from the Assessment System Database, using even separated servers, to improve system overall performance. In the features

database is assigned the Student Personnel ID that is used to link the students' personnel information with the fingerprint image.

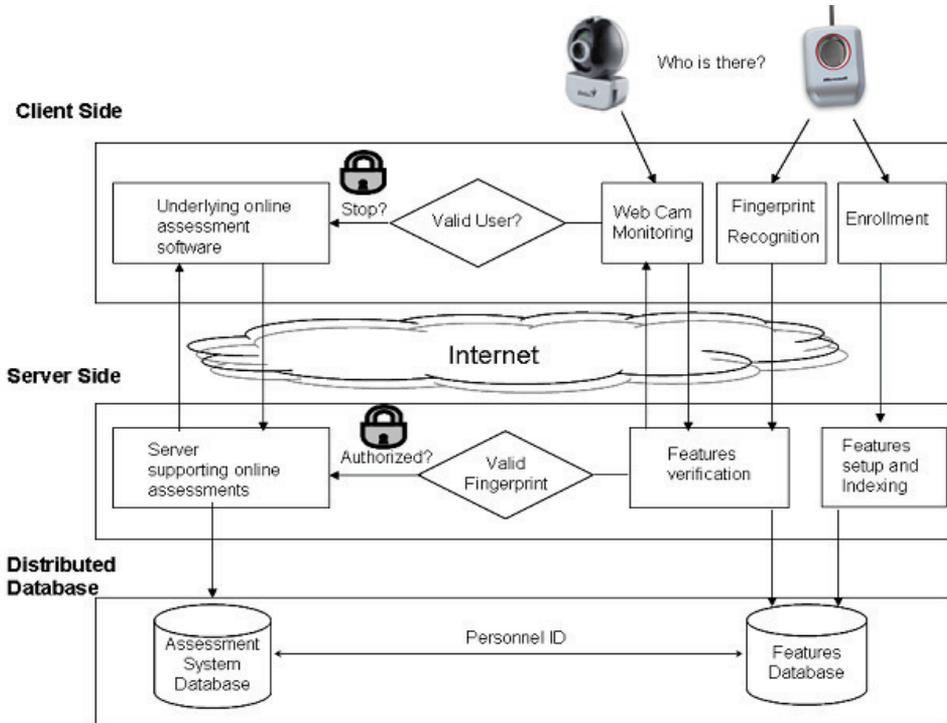


Fig. 10.- Student's biometric recognition on real time & data mining analysis

### 6.3.3 Implementation

- Hardware
  - Client System Requirements (minimal). Pentium class (i386) processor (200 MHz or above) with 128Mb or higher, 100Mb disk space.
  - Fingerprint mouse. 250 DPI (Digits per Inch) or higher; 500 DPI is recommended.
  - Web cam. Genius VideoCaM GE111 or VideoCam GF112.
  - Broad-band Internet. Minimum 128 Kbps, recommended 256 Kbps.
- Software
  - Biometrics SDK. Griaule GrFinger SDK 4.2 allows you to integrate biometrics in a wide variety of applications. Provides Support for dozens of programming languages -including java- and integration with several Database Management Systems. Besides, provides multiple fingerprint reader support, and even after application development or deployment, makes you able to change the fingerprint reader you're using, without modifying your code.
  - Fingerprint template size: 900 bytes average.
  - RapidMiner (Version 4.1). To carry out data mining process.
  - Programming language. Java due the online assessment software tool was developed using this technology, and JMF (Java Media Framework) to allow transmission of video and/or photographs over the Internet.

- Web Server. Apache 2.2.
- Database Management System. MySQL.
- Operating System: Windows 98, Windows ME, Windows NT. Windows 2000, Windows XP, Windows 2003 or Windows Vista.

The fingerprint is verified in the Features Database, and if it is recognized as a valid, then the Server authorizes access to the online assessment application, else an error message is sent to the Client to try again. In other hand, if the student's fingerprint is valid, the user is authenticated into system (see Figure 11), the evaluation process starts and web cam transmission is initialized at Client Side to conduct real time monitoring by means of multitasking, students' activities (keystroke dynamics, navigation and performance patterns) are stored on features database and log files.

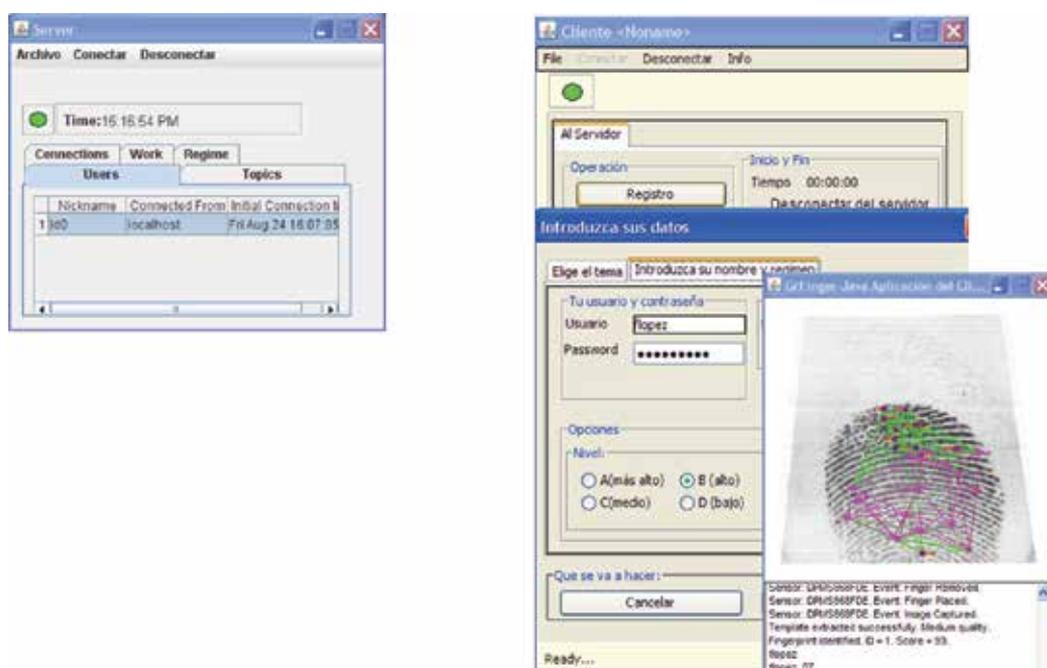


Fig. 11. The Client-Server Application supporting fingerprint recognition to authenticate students in online assessments

If someone else tries to get the control of the computer during the online assessment, the evaluation process is finished prematurely, and the results are sent to server side to be processed as they are. To the contrary, the evaluation process is finished successfully, the assessment is processed at Server Side, and the final results of evaluation and security status are shown at Client Side (Hernández et al. 2008). The stage of data mining is executed asynchronously to verify students' behaviour patterns: keystroke dynamics (Gutiérrez et al. 2002), navigation patterns (Xing & Shen 2004) and performance patterns (Hernández et al. 2006). After repeatedly usage of the system, these combined patterns can be used to verify students' identity and even to substitute the usage of webcams.

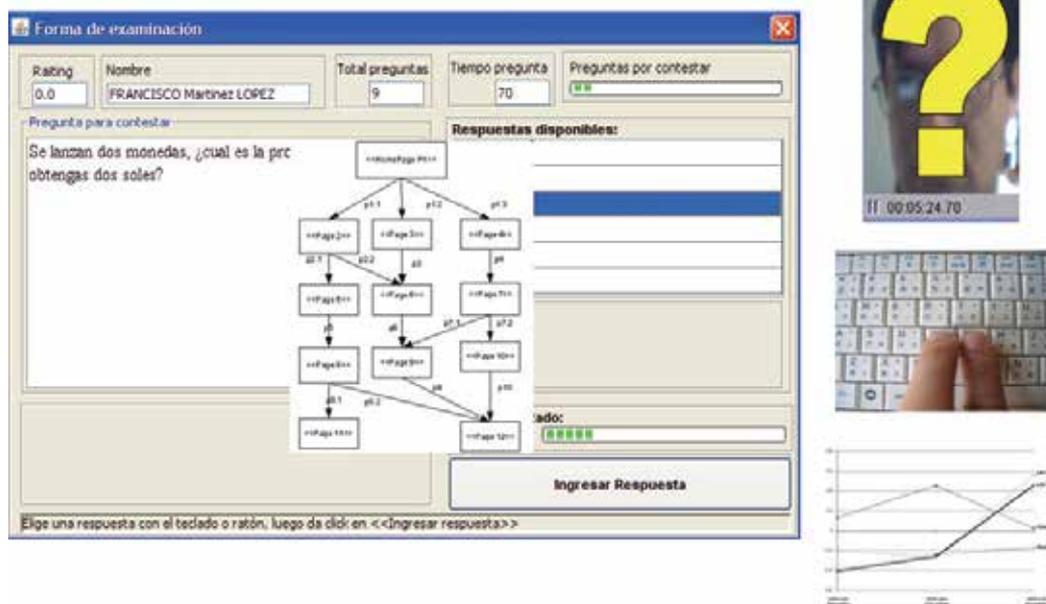


Fig. 12.- Assessment process: recording of keystroke, navigation and performance patterns.

### 6.3.4 Preliminary results

We develop an experiment to try the above mentioned technology. We selected a random sample of students ( $n=54$ ) from the José María Morelos y Pavón High School, located in Temixco, Morelos, México. On this test we obtain a False Acceptance Rate (FAR) of 99.99% and a False Reject Rate (FRR) of 97.09%, only one student could not be recognized despite several trials, although we try enrolling her trying different fingers of her left hand, simply we could not, she has tiny long fingers and the enrolment results were always the same. Her fingerprint template can not be understood by the system due to confusion, her fingerprints templates seem like stains. Something related is registered in literature, Asiatic persons have similar problems to be identified by fingerprint readers (Michigan Org, 2007). We faced this problem by providing this student an user and a strong password to allow permission to the system.

In general, students perceived our system as faster, easy to use and secure; fingerprint recognition plays an important role in this last point. However 13% dislike web cam monitoring. When we asked them directly if they dislike being monitored, 33% answered bothers this fact. They felt under pressure, get nervous and dislike being monitored or watched.

A 20% noticed a way to commit cheat using a system like ours, the ways are: turn the camera to some else, use a photo, use a cheating list, and just one person thinks to dirt the fingerprint reader. We made in-depth analysis and discover that students with poor performance (low grades) are willing to commit cheat. Finally, 78 % of the students would like the system being implemented at their high school.

Data mining process shows promising results to identify remote students by using keystroke dynamics; meanwhile navigation and performance patterns are useful to identify suspicious behaviour (i.e. unexplained grades, unidentified remote IPs, and completely different navigation patterns –regarding previous- when solving an exam).

We consider that the online assessment system with biometric recognition was very well accepted, but must be adapted to be more user friendly and the process to enrol users must be improved too.

## 7. Conclusion and the future research

The quickly grown of the Web has done that it is a great information source in many areas, which can be used to obtain important data in different areas like social, psychological, marketing, among others. As one saw from point psychological are possible to be studied some behaviors and found certain landlords with the help of Data Mining, also it is possible to determine future behaviors on the basis of certain antecedents as one is in the application on the basis of social networks information like Orkut. On the other hand it can predict certain preferences by some product or service; this could be observed through the Web Radio application. The Web Data Mining can help us to understand more some things and provide a base for the decision make. In the Web one can find a great amount of information sources, the unique thing that there is to think is those that are wanted to obtain. In this case our applications analyzed were Social Networks, Web radio, Security and Internet frauds. In this work to show a conceptual model to develop systematically web radio applications taking into account social acceptability factor using the social data mining and cultural algorithms. With the matching information about the users (listener profiles) with the knowledge building for a society based on cultural algorithms (content metadata), it could be possible to automatically generate more pertinent playlists for individual listeners. Then, it is feasible that a web radio purpose could interpret the human behaviour under certain situations to which it is exposed, this by means of the behaviour that will have the users when interacting with the Web Radio. This is only one part of everything what it is possible to be done with the aid of the Social Data Mining. With the creation of the web radio one hopes that one undertakes new projects that are developed with the aid of the Social Data Mining and cultural algorithms.

Nevertheless there are a lot of research work that will be doing with Data Mining for Web applications and some future works that can be: In the area of social networks it is the search of criminal networks in the diverse social networks like Orkut, MySpace, Badoo, Hi5, etc. these criminal networks are possible to be dedicated to the kidnapping, traffic of bodies, drug traffic, among others activities. Another application is to make an analysis of these networks to obtain information that can be used to offer products and services in specific sectors, an example of this is the Web radio, but it is possible to be analyzed another types of services with the security that these are going well to be received by a certain Web user group.

In the Internet area fraud we want to improve human-computer interface and assessment methodology by including student's comments and users feedback. We want to test the tool with different groups at different high schools and Universities. Regarding biometric recognition, we want to improve facial recognition, due at this point of our research we can

detect student's presence or absence only, and our intention is comparing face patterns automatically by means of photo IDs stored at our features databases. We want to test the newest fingerprint scanners included in mouses, keyboards and in some laptops and try to incorporate them to work within our system.

## 8. References

- Bloedorn, E.; Christiansen, D.; Hill, W.; Skorupka, C.; Talbot, L. & Tivel J. (2001). Mining for Network Intrusion Detection: How to Get Started, *MITRE Technical Report*, August 2001
- Brooks, P. (1994). Adding value to usability testing, In: *Usability Inspection Methods*, Nielsen, Jakob y Mack, Robert, 1, pp. 255-271, Published by John Wiley & Sons, New York, NY
- Cai, D.; Shao, Z.; He, X.; Yan, X. & Han J. (2005). Mining Hidden Community in Heterogeneous Social Networks, *Proceedings of LinkKDD'05*, Chicago, USA, August 2005, ACM
- Clarke, N. & Furnell, S. (2005). Biometrics: "The promise versus the practice", In *Computer Fraud & Security*, Volume 2005, pp. 12-16, September 2005
- Clifton, C.; Kantarcioglu, M. & Vaidya J. (2002). Defining Privacy For Data Mining, *Proceedings of the National Science Foundation Workshop on Next Generation Data Mining*, pp. 126-133, Baltimore, USA, November 2002
- Dokas, P.; Ertoz, L.; Kumar, V.; Lazarevic, A.; Srivastava, J. & Tan, P.-N. (2002). Data mining for network intrusion detection, *Proceedings of NSF Workshop on Next Generation Data Mining*, pp. 21-30, November 2002
- Domingos, P. & Richardson M. (2001). Mining the network value of customers, *Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining*, pp. 57-66, San Francisco, USA, August 2001, ACM, California City
- Field, A.; Hartel, P. & Mooij, W.(2001). Personal DJ, an Architecture for Personalised Content Delivery, *Proceedings of WWW10*, Hong Kong, China, May 2001, ACM 1-58113-348-0/01/0005
- Garfinkel, S. (2001). Database Nation: The Death of the Privacy in the 21st Century, O'Reilly & Associates, Sebastopol, CA, USA, 2001.
- Garofalakis, M.; Rastogi, R.; Seshadri, S. & Shim, K. (1999). Data Mining and the Web: Past, Present and Future, *Proceedings of 2nd ACM International Workshop on Web Information and Data Management (WIDM)*, pp. 43-47, Missouri, USA, November 1999, ACM, Kansas City
- Gutiérrez, F.; Lerma, M.; Salgado, L. & Cantú, F. (2002). Biometrics and Data Mining: Comparison of Data Mining-Based Keystroke Dynamics Methods for Identity Verification, *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, pp. 221-245
- Hernández, J.; Ochoa, A.; Andaverde, J. & Burlak. G. (2008). Biometrics in online assessments: A Study Case in High School Students, *Proceedings of 18th International Conference on Electronics, Communications and Computers Conielectcomp 2008*, pp. 111-116

- Hernández, J.; Ochoa, A.; Muñoz, J. & Burlak, G. (2006). Detecting cheats in online student assessments using Data Mining, *Proceedings of The 2006 International Conference on Data Mining (DMIN'2006)*, pp. 204-210, Las Vegas, USA, June 2006, Nevada City
- Hill, W. & Terveen, L. (1996). Using Frequency-of-Mention in Public Conversations for Social Filtering, *Proceedings CSCW'96*, pp. 106-112, Boston, USA, November 1996, ACM, MA. City
- ITEDU (2006). Biometrics. Consulted in <http://et.wcu.edu/aidc/> August, 2006
- Javitz, H. & Valdes, A. (1993). The NIDES Statistical Component: Description and Justification, Technical Report, Computer Science Laboratory, SRI International
- Kumar, V.; Srivastava, J. & Lazarevic, A. (2005). Intrusion Detection: A survey, Managing Cyber Threats Issues, Approaches, and Challenges Chapter 2, Springer Verlag
- Lauw, H.; Lim, E.; Tan, T. & Pang, H. (2005). Mining Social Network from Spatio-Temporal Events, *Proceedings of SIAM Data Mining Conference*, April 2005, Newport Beach
- Lorrain F. & White H. (1971). structural equivalence of individuals in social networks, *In Journal of Mathematical Sociology*, pp. 49-80
- Lundin, E.; Kvarnstrom, H. & Jonsson, E. (2002). A synthetic fraud data generation methodology, *In Lecture Notes in Computer Science ICICS 2002, Laboratories for Information Technology*, Singapore, December 2002, Springer Velag
- Márquez, M.; Ojeda, S. & Hidalgo, H.(2008). Identification of behavior patterns in household solid waste generation in Mexicali's city: Study case, *Journal of Resources, Conservation and Recycling*, volumen 52, Issue 11, September 2008, pp. 1299-1306
- Michigan Org (2007). Consulted on line at [http://www.reachoutmichigan.org/funexperiments/agesubject/lessons/prints\\_ext.html](http://www.reachoutmichigan.org/funexperiments/agesubject/lessons/prints_ext.html), August 2007
- Milgram S. (1967). *The small world problem*, *Psychology Today*, Vol., 2, pp. 60-67
- Mobasher, B.; Dai H., Luo T.; Sun, Y. & Zhu J. (2000). Integrating web usage and content mining for more effective personalization, *Proceedings of Electronic Commerce and Web Technologies, First International Conference, EC-Web 2000*, pp. 165-176, ISBN 3-540-67981-2, London, UK , September 2000, Lecture Notes in Computer Science 1875 Springer 2000
- MSU Michigan State University (2006). Quizzes and Exams: Cheating on exams and Quizzes Consulted in [http://teachvu.vu.msu.edu/public/pedagogy/assessment/index.php?page\\_num=3](http://teachvu.vu.msu.edu/public/pedagogy/assessment/index.php?page_num=3), August 2006
- Nielsen, J. & Loranger, H. (2006). *Prioritizing Web Usability*, *In New Riders Press*, , pp. 165-176, ISBN-13: 978-0-321-35031-2, Berkeley, CA
- Ochoa, A. (2006). Más allá del Razonamiento Basado en Casos y una Aproximación al Modelado de Sociedades utilizando Minería de Datos, *Univ. Autónoma de Aguascalientes*, México, 1th Edition, Aguascalientes
- Ochoa, A.; Agüero, M.; Mügerza, F.; Alvarado C.; Espino, M.; Jiménez, C.; Limón, J. & Valádez, M. (2007). Design and implementation of a Thematic Web Radio based on Social Data Mining and Cultural Algorithms, *Proceedings of CONTECSI'07*, Guanajuato, México, November 2007, León City
- Ochoa, A.; Tcherassi, A.; Shingareva, I.; . Padméterakiris A.; Gyllenhaale J. & Hernández A. (2006). Italianità: Discovering a Pygmalion effect on Italian communities using data

- mining, *Proceedings of 7o Congreso de Computación CORE'2006*, ISSN: 1665-9899, México, México, May 2006, In Journal in Computing Science
- Oliveira, S. & Zaiane, O. (2004). Toward Standardization in Privacy-Preserving Data Mining, *Proceedings of 3rd Workshop on Data Mining Standards, ACM SIGKDD*
- Ponce, J.; Ochoa, A.; Pietsch, W. & Zolezzi-Hatsukimi, Z. (2007). Ahankara: Identify Bipolar Síndrome in User of Orkut with Data Mining, *Proceedings of ENC 2007*, Michoacan, Mexico, September 2007, IEEE, Morelia City
- Ponce, J.; Ponce de León, E.; Padilla, F. Padilla, A. & Ochoa, A. (2006). Ant Colony Algorithm to solve Clique problem with Local Optimizer K-Opt (In Spanish), *Journal Hifen, Urugaiana, Brazil*, November 2006, pp. 191, ISSN 0103-1155
- Rove, N. (2004). Cheating in Online Student Assessment: Beyond Plagiarism, *Online Journal of Distance Learning Administration*, Volume VII, Number II, Summer 2004 State University of West Georgia, Distance Education Center
- Resnick, P.; Iacovou, N.; Suchak, M.; Bergstrom, P. & Riedl, J. (1994). GroupLens: An Open Architecture for Collaborative Filtering of Netnews, *Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work CSCW'94*, pp. 175-186, Chapel Hill, NC, October 1994
- Reynolds, R. (1998). An Introduction to Cultural Algorithms, *In Cultural Algorithms Repository*, <http://www.cs.wayne.edu/~jcc/car.html>
- Rezgur, A.; Bouguettaya, A. & Eltoweissy, M. (2003). Privacy on the Web: Facts, Challenges, and Solutions, *In IEEE Security & Privacy*, pp. 40-49, November-December 2003
- Schoeman, F. (1984). *Philosophical Dimensions of Privacy*, Cambridge University. Press
- Schwartz, M. & Wood, D. (1993). Discovering shared interests using graph analysis., *Communications, ACM*, pp. 78-89
- Varan, S. (2006). *Crime Pattern Detection Using Data Mining*, Oracle Corporation
- Wahlstrom K., & Roddick J. (2000). On the Impact of Knowledge Discovery and Data Mining, *Proceedings of Australian Institute of Computer Ethics Conference (AiCE2000)*, Canberra, Australia, April 2000, Sydney City
- Wasniowski, R. A. (2005). Using Data Fusion for biometric verification, *Transactions on Engineering Computing and Technology*, April 2005. pp. 72-74
- Wasserman, S. & Faust, K. (1994). *Social Network analysis: methods and applications*, Cambridge University Press, ISBN-13: 9780521387071, Cambridge, UK
- Wisher, R.; Curnov, C.; and Belanich, J. (2005). Verifying the Learner in distance learning, *Proceedings of 18 Annual Conference on Distance Teaching and Learning 2005*
- Witten, I. & Frank E. (2005). *Data Mining: Practical machine learning tools and techniques*, Morgan Kaufmann, 2nd Edition, pp. 1-525, ISBN 0-12-088407-0, San Francisco
- Xing, Dongshan; and Shen, Junyi (2004). Efficient data mining for web navigation patterns, *Journal of Information and Software Technology*. Volume 46, Issue 1, 1 January 2004, Pages 55-63
- Xu, J.; Huang, Y. & Madey, G. (2003). A Research Support System Framework for Web Data Mining, *Proceedings of WSS'03: WI/IAT 2003 Workshop on Applications, Products of Web-based Support Systems*, ISBN 0-9734039-1-8, Halifax, Canada, October 2003

- Yao, J. & Yao Y. (2003). Web-based information retrieval support systems: building research tools for scientists in the new in-formation age, *Proceedings of the IEEE/WIC International Conference on Web Intelligence*, Halifax, Canada
- Yao, Y. (2003). A framework for web-based research support systems, *Proceedings of Computer Software and Application Conference, COMPOSAC 2003*, Dallas, Texas
- Zaiane, O.; Xin, M. & Han J. (1998). Discovering web access patterns and trends by applying OLAP and data mining technology on weblogs, In *Advances in Digital Libraries*, pp. 19-29
- Zolezzi-Hatsukimi, Z. (2007). Implement social nets using Orkut, *Proceedings of CHI'07*, Nagoya, Japan

## BIOLOGICAL APPLICATIONS



# Application of Data Mining Techniques to the Data Analyses to Ensure Safety of Medicine Usage

Masaomi Kimura  
*Shibaura Institute of Technology*  
Japan

## 1. Introduction

Not only the safety of medicines themselves, from the perspective of medicinal properties, but also the safety of medicine usage is important, in order to ensure the right use of the right medicines and prevent medical accidents. The author has applied data-mining techniques to the analyses of data to investigate the status about the safety of medicine usage, such as the data in the database of medical near-miss cases which have been collected by Japanese government, the investigation data to understand how patients handle the injection device for anti-diabetic drug by themselves after the guidance by a doctor, the questionnaire data asking opinions about the mark indicating a medical property on some cardiac transdermal patch, which were provided by medical experts and/or pharmaceutical companies.

The analyses on such data have been traditionally based on the statistical approaches, which bring us the information about the single attribute of data, for instance, the tables showing the frequency of events under some condition and the analysis whether the difference of the frequency is statistically significant, and so on. It is, however, important to extract the information not only of the each attribute independently but also about the relations among them, since each data in the results of the investigations is supposed to be generated under complex conditions, some set of which is an essential cause. In order to find the relationships of data or the items of the data, it is useful to utilize the clustering algorithms to classify the data into some clusters, in each of which data have the same characteristics, and the decision tree algorithms to find the condition to distinguish the data based on the classes to which they belong, from in rough to in detail. By means of both algorithms, we can find a structure among the data, which kinds of data we have obtained and which conditions generate such kinds of data.

In this paper, we first make a brief review of some clustering algorithms including agglomerated hierarchical clustering algorithm, TwoStep clustering algorithm and K-means algorithm with the number of seeds optimized by Bayesian information criterion (BIC) and also show how to combine them to the idea of specialization coefficients. Next, we introduce the review of the application to the data related to the safety of medicine usage shown above. Last, we summarize the principle of application of data mining to such analyses.

## 2. Clustering algorithms and their variations for applications to questionnaire data / survey data

As we stated in the introduction, clustering algorithms help us find the structures that lie within data (Berry & Linoff, 1997). There are two types of algorithms, hierarchical one and non-hierarchical one. The representative algorithm of the former type is the agglomerated hierarchical clustering algorithm, which incorporates data or clusters sitting neighbourhood in the order of short distance. This provides us with graphical output, dendrogram, which shows the global structure of data in the form of a tree diagram. When we apply this method, it is important to adopt appropriate distance measure between data and/or clusters. The representative algorithm of the latter type is K-means algorithm, which classifies data into the predetermined number of clusters and shows masses within data. Though this offers easy-to-understand results, unfortunately, it is not clear which number of cluster is appropriate to adopt in the typical way of application. Another typical algorithm, which is the intermediate algorithm of hierarchical and non-hierarchical one, is TwoStep clustering algorithm (SPSS inc., 2003; Zhang et al., 1996), which can be effectively applicable to massive data. This consists of two procedures. To deal with data effectively, the data is not handled homogeneously but divided into groups (pre-clusters) that consist of similar data by means of CF Tree (the first step). After that, clustering algorithm is again applied to the pre-clusters (the second step).

In order to apply such algorithms to questionnaire data or survey data, we propose some improvements in following subsections.

### 2.1 Agglomerated hierarchical clustering using selection co-occurrence measure

Let us consider the case to deal with multiple-choice questions. In general, the analysis of the answers for such questions is based on the majority vote for each option. However, the combinations of the selected options also have a significant meaning, since the answerers do not choose each option independently but they express their intent as a group of options. Hence, it is important to find the option patterns in the data that most answerers select, by utilizing clustering algorithms.

Let  $d$  denote the number of options, and  $n$ , the number of respondents. Let us also define 'answer matrix'  $A_{ij}$ , whose value is '1' if the  $i^{\text{th}}$  respondent selects the  $j^{\text{th}}$  option and '0' if not. To see the co-occurrence relationship between the selection of options, we apply an agglomerated hierarchical clustering algorithm to the column vector of the answer matrix, which we call selection vector  $A_{i*} = (A_{i1}, A_{i2} \dots A_{in})$ .

When we perform agglomerated hierarchical clustering algorithm, we usually adopt the Euclidean distance or Manhattan distance to measure the distance of the vectors. If we apply them to two selection vectors, it is equivalent to counting the number of different elements of the vectors, or counting the number of '1' obtained from the result of 'exclusive or' of each corresponding element of the vectors. This results in not only the most simultaneously selected options, but also simultaneously unselected options being judged to neighbour each other. The options that are unselected by most respondents therefore have a short distance to each other.

The similarity index, however, should measure the co-occurrence of simultaneously selected options, not of those unselected. We therefore adopt other similarity measures than the Euclidean or Manhattan distance.

Let us remember the definition of the inner product of the vectors. If the elements of the vector are either '1' or '0', the inner product counts up the number of elements whose value is both '1', in other words, the number of '1' obtained from the result of the 'and' operation to each corresponding element of vectors. The contributing value comes only from the elements that the respondents choose as a cluster. This suggests us that the inner product is suitable for the similarity index. We take this into account and propose the 'distance' that counts the number of elements whose value is not 1 in either vector. Let  $D^\#$  denote the 'distance', which can be calculated as

$$D^\#(A_{i^*}, A_{j^*}) = \sum_{k=1}^n (1 - A_{ik} A_{jk}) = n - \sum_{k=1}^n A_{ik} A_{jk} \tag{1}$$

Note that  $D^\#$  does not satisfy one of the axioms of distance, namely  $D^\#(A_{i^*}, A_{i^*})$  is not equal to 0. This is because the contribution to  $D^\#$  of the element that has the value 0 in each vector is defined as 1. The requirement of distance for agglomerated hierarchical clustering is to supply measurement that allows us to compare the similarity. Since measure  $D^\#$  satisfies the requirement, we adopt  $D^\#$  as 'distance'. After this, we call  $D^\#$  the selection co-occurrence measure (Kimura et al., 2006a).

As merging methods of clusters, many methods are known, such as the nearest-neighbour method, the furthest-neighbour method and the Ward method. We applied these methods and found that they provide similar dendrograms as a result. Not some but all of the options in the clusters should have a similar pattern of selection if they are merged. We therefore adopt the furthest-neighbour method as the clustering algorithm.

Table 1 shows the sample data that we used to verify our method. Figure 1 shows the dendrogram for which the Euclidean distance is used and that for which we use the selection co-occurrence measure. We can see that Option 2 and Option 5 are judged to be neighbouring for Euclidian distance, although no respondents chose them simultaneously. However, for selection co-occurrence measure, these options are judged to be far from each other as we expected.

	Option 1	Option 2	Option 3	Option 4	Option 5
Answerer1	1	0	1	1	0
Answerer2	1	0	1	1	0
Answerer3	1	0	1	1	0
Answerer4	0	0	1	1	1
Answerer5	0	0	0	0	1
Answerer6	0	1	0	0	0
Answerer7	1	1	1	1	0
Answerer8	1	0	1	0	0
Answerer9	1	0	1	0	0

Table 1. Sample data to verify selection co-occurrence measure.

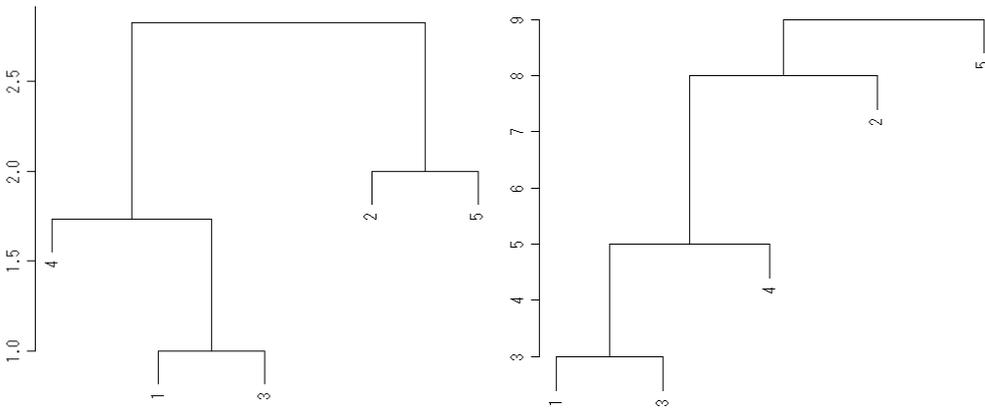


Fig. 1. Dendrogram for Euclidian distance (left) and selection co-occurrence measure (right).

### 3.2 K-means algorithm with the number of clusters determined by BIC

The K-means algorithm classifies data into clusters (the group of data that neighbor each other), the number of which is specified beforehand. (Let  $K$  denote the number of clusters.) The algorithm is realized by the iteration of the following procedures:

- A. Centroid  $X_i$  is obtained as Equation 2, where  $C_i$  denotes the  $i^{\text{th}}$  cluster and  $N_i$  denotes the number of elements of cluster  $C_i$ .

$$X_i = \frac{1}{N_i} \sum_{x \in C_i} x \quad (2)$$

- B. Measuring the distance from the data to each centroid, the data is reassigned to the  $i_x^{\text{th}}$  cluster specified as Equation 3.

$$i_x = \arg \min_i \|x - X_i\| \quad (3)$$

We adopt converged clusters after the iteration of Step A and B. Note that we do not have  $C_i$  initially in Step A, the randomly selected data are used as substitutes.

As we mentioned above, the number of clusters  $K$  has to be specified before performing the algorithm and is usually determined by trial and error. In this study, we determined it by utilizing Bayesian information criteria (BIC). Although it should be simple for the model to suppress overfitting to the data, there is a trade-off relationship between the simplicity and the accuracy of the model. BIC is the index that balances the simplicity and the accuracy of model, whose minimum value gives the best model for describing the data.

To calculate the value of BIC, we translate the K-means method into a probability model, or more specifically, a likelihood function. We assume the likelihood function of the K-means method as follows:

$$p(\{x\}) = \prod_{i=1}^K \prod_{x \in C_i} P_i(x; X_i, \sigma) \quad (4)$$

$$P_i(x; X_i, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\|x - X_i\|^2}{2\sigma^2}\right) \tag{5}$$

where  $\{x\}$  denotes the data group that has  $n$  elements, and  $X_i$  and  $\sigma$  are parameters whose maximum likelihood estimates are the centroid of cluster  $C_i$  and the mean value of the variance in each cluster. With these equations, the K-means algorithm can be reproduced as follows.

A'. Calculate the maximum likelihood estimates of  $X_i$  and  $\sigma$  of  $p(\{x\})$  as

$$\frac{\partial}{\partial X_i} p(\{x\}) = 0 \tag{6}$$

$$\frac{\partial}{\partial \sigma^2} p(\{x\}) = 0 \tag{7}$$

As a result, we obtain the centroid of each cluster.

B'. Each item of data is assigned to the cluster that has the maximum value of probability  $P_i$  as is shown in Equation 8. In fact, this condition is equivalent to Equation 2, since exponential function is monotonic.

$$i_x = \arg \max_i P_i(x; X_i, \sigma) \tag{8}$$

We again obtain the same clusters that is obtained by iterations A) and B), since steps A') and B') are equivalent to A) and B), respectively. We therefore adopt  $p(\{x\})$  as a likelihood function that corresponds to the K-means algorithm.

The definition of BIC is given as Equation 9 (Shimodaira et al., 2004):

$$BIC = -2 \sum_x \ln P(x; \hat{\theta}) + F \ln n \tag{9}$$

where the first term is the logarithm of the likelihood function with the maximum likelihood estimates,  $F$  is the degree of freedom of the parameters and  $n$  is the number of data. The first term decreases its value if we make many small clusters that fit the data. The second term, however, decreases its value if we make a small number of clusters. The smallest value of BIC gives us the optimal value of  $K$ , which provides a small number of clusters approximating the data well. The model that we use in our study has  $K$  centroids and one variance as parameters. If the dimension of the data is  $D$ , the degree of freedom is  $KD+1$ . BIC for our model (Kimura et al., 2006b) is therefore given as:

$$BIC = n \left( 1 + \ln \left( \frac{2\pi}{n} \sum_{i=1}^K \sum_{x \in C_i} \|x - X_i\|^2 \right) \right) + (KD + 1) \ln n \tag{10}$$

For clustering, we calculate BIC for the target data and find  $K$  that minimizes BIC.

### 3. Application

We review our studies where the methods shown in the previous section are applied. First, we show the application of agglomerated hierarchical clustering with selection co-

occurrence measure to the questionnaire data that investigate the evaluation of design of 'therapeutic classification mark' of cardiac transdermal patch. (Kimura et al., 2006a) Next, we introduce the application of BIC to K-means algorithm to evaluate the characteristics of patients who handle the injection device for anti-diabetic drugs in the wrong way even after guidance. (Kimura et al., 2006b)

### 3.1 The application of agglomerated hierarchical clustering with selection co-occurrence measure

Our target data was the answers for the multiple-choice questions listed in Table 2, which is a part of the questionnaire on the 'therapeutic classification mark' and product name label to ensure the safety of drug use, performed in 2004. Each question is answered by 7,078 doctors, 7,018 nurses and 7,361 pharmacists.

It was important to conduct and analyze such nation-wide investigations using questionnaire to obtain feedback from medical experts (doctors, nurses, pharmacists) and patients about the 'therapeutic classification mark' printed on isosorbide dinitrate transdermal patches, which is a cardiac medicine, since there had never been a study that estimates the measures from the position of medical experts or patients. We compare the result obtained by agglomerated hierarchical clustering algorithm with selection co-occurrence measure with the one obtained by TwoStep clustering algorithm in order to see what groups of respondents select the patterns (combinations) of options. Figure 2 shows the result that is obtained by applying the method to the sample data in Table 1. Roughly speaking, this suggests that there are two clusters, one of which contains Options 1, 3 and 4 and the other of which contains Option 2 and 5.

Comparing this with Fig. 1, we can see the relationship among the results as follows:

- Although Option 4 is found to be a neighbor of the cluster of Option 1 and 3 in Fig. 1, it is difficult to determine whether we should regard these three options as one group. Figure 2, however, suggests that the options can be identified as one group.
- Figure 2 suggests that Option 2 and Option 5 are chosen simultaneously, although they have never actually been chosen together. This is because each of the options is independently selected with Options 3 and 4, which mediate Options 2 and 5 to be in the same cluster. Figure 1, however, shows that Option 2, and 5 cannot be included in the same cluster.

Considering these results together, we can identify the group of options that the respondents select together, by comparison with these two results. On this point, these two methods complement each other.

Figure 3 shows the results of Question A. We can see that Option 1 and 3 are selected together. Although the frequency of co-occurrence is smaller, Option 5 can be regarded as being selected simultaneously with those two options. This suggests that the combination of systemic transdermal absorbent preparations that medical experts deal with is mainly a cardiac drug and an asthma drug, and that there are some cases where cancer pain-relief medicine is also used.

The results for Question B are shown in Fig. 4. From these results, we can see that Option 1 and Option 3 are selected together by most respondents, and Option 5 is the one that is simultaneously selected next to these options. This suggests that the main reason to adopt systemic transdermal absorbent preparations as a dosage form is that no burden is imposed on the digestive tract and that the effect continues for many hours. Additional to this, it is also because diet does not have any impact. Since we can also see that the selection of

Options 2, 4 and 6 are associated with these options, the effect of liver, administration termination and good compliance can be regarded as associated reasons to select transdermal patches.

<b>A</b>	<b><i>What systemic transdermal absorbent preparations do you usually deal with?</i></b>
1	Cardiac medicine
2	Hormone replacement
3	Asthma medicine
4	Smoking-cessation medicine
5	Cancer pain-relief medicine
<b>B</b>	<b><i>Why did you select the systemic transdermal absorbent preparation?</i></b>
1	Burden is not imposed on the digestive tract.
2	First pass effect of the liver does not have an effect.
3	Effect lasts for many hours.
4	Administration can be terminated by peeling off.
5	Eating meals does not have an effect.
6	I can ensure good compliance.
7	I do not select.
8	Others
<b>C</b>	<b><i>What do you think about the design of the therapeutic classification mark and product name label of Frandol tape S?</i></b>
1	The concept is valid for medical accident prevention.
2	More innovation of the concept is necessary to prevent medical accidents.
3	The mark is favourable for cardiac transdermal patches.
4	More innovation of the mark is necessary.
5	The print colour, white, is easy to see and favourable.
6	The print colour should be more vivid.
7	The mark, label and layout are valid for medical accident prevention.
8	More innovation of the size of the mark, the number of labels, and layout is necessary.
<b>D</b>	<b><i>What are preventive measures against medical accidents related to the systemic transdermal absorbent preparation?</i></b>
1	Displaying the mark and the label is good enough.
2	The mark for the same efficacy should be integrated.
3	The mark for the same efficacy should not be integrated but unique to each company.
4	The mark should be displayed for other systemic transdermal absorbent preparations.
5	Displaying the mark and the label is unnecessary.
6	The effort is necessary to earn recognition from medical experts, patients and their families.

Table 2. Questions and options (originally in Japanese)

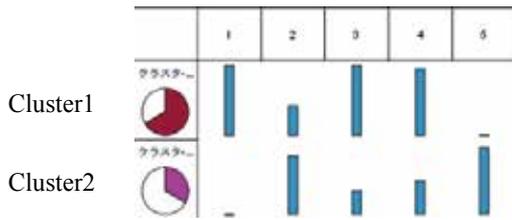


Fig. 2. The result of TwoStep algorithm applied to the sample data in Table 1.

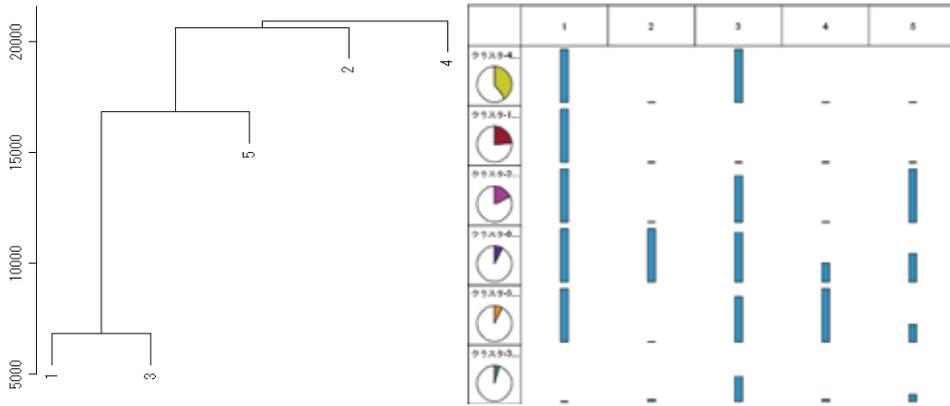


Fig. 3. The dendrogram (left) and clusters obtained by TwoStep algorithm (right) of selected options of Question A.

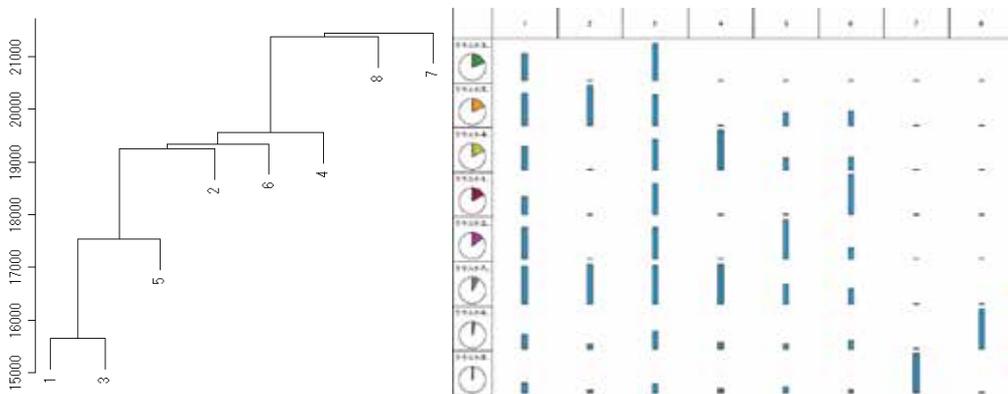


Fig. 4. The dendrogram (left) and clusters obtained by TwoStep algorithm (right) of selected options of Question B.

The results of Question C are indicated in Fig. 5. Comparing these figures, we can see that Options 1, 3, 5 and 7 are selected together. This can be interpreted as the design being valid for the prevention of medical accidents caused by transdermal patches, having a high level of visibility and being preferable. Since the cluster that includes 1, 3, 5, and 7 in the result of TwoStep algorithm contains 81% of the respondents, this suggests that most medical experts

have a favorable opinion on the mark and the product name label on the target transdermal patch. Though some reader might think only 5,000 respondents select the pair of Options 1 and 3 in the dendrogram, they should notice that the result of TwoStep algorithm counts the number of respondents who chose any combination of Options 1, 3, 5 and 7.

Figure 6 shows the results of Question D. These figures indicate that most respondents selected Options 2, 4 and 6. This suggests that most medical professionals think that the therapeutic classification mark and product name label are necessary, should be integrated for the same efficacy and should be widely recognized. The respondents in the cluster in Fig. 6 account for about 58% of the whole. This indicates that more than half of the respondents do not satisfy the current situation and think that use of the mark and product name label should be widely spread.

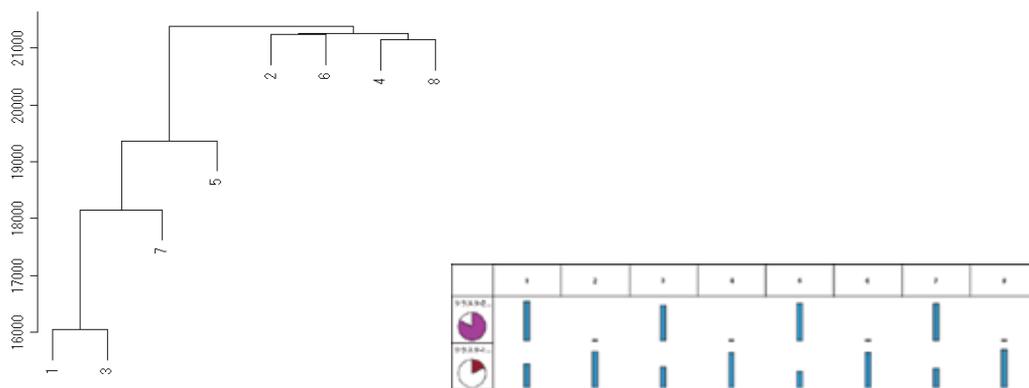


Fig. 5. The dendrogram (left) and clusters obtained by TwoStep algorithm (right) of selected options of Question C.

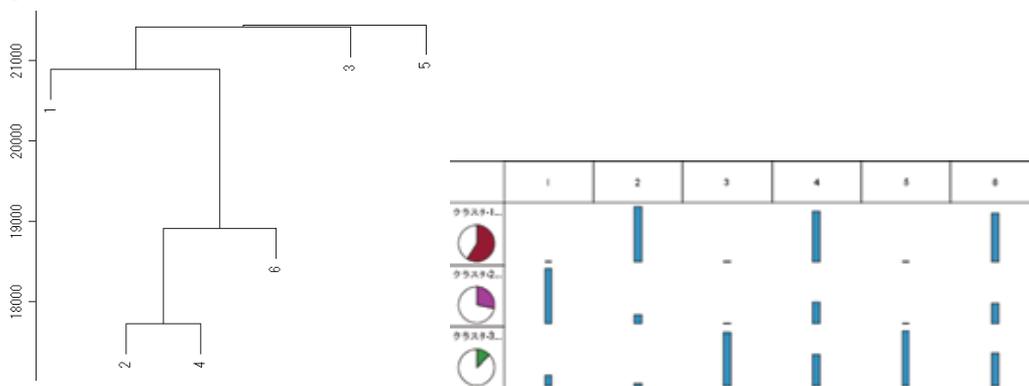


Fig. 6. The dendrogram (left) and clusters obtained by TwoStep algorithm (right) of selected options of Question D.

値▲	割合	%	度数
Doctor		27.05	3420
Nurse		35.21	4451
Pharmacist		37.73	4770

Fig. 7. The distribution of occupation in Cluster 1.

値▲	割合	%	度数
Doctor		42.5	2557
Nurse		31.37	1887
Pharmacist		26.13	1572

Fig. 8. The distribution of occupation in Cluster 2.

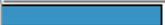
値▲	割合	%	度数
Doctor		39.32	1101
Nurse		24.29	680
Pharmacist		36.39	1019

Fig. 9. The distribution of occupation in Cluster 3.

Figures 7, 8, and 9 show the distribution of respondents in each cluster in 12 by occupation. From Fig. 7, we can see the trend that the opinion that the mark and label should be promoted comes from pharmacists and nurses. Figure 8 suggests that doctors tend to satisfy the current situation. Figure 9 shows that the profession people who least select the negative options are nurses.

### 3.2 The application of K-means algorithm with the number of clusters determined by BIC

In this subsection, we review the study whose target data was a self-injection device of an antidiabetic drug, insulin. Before using the device, patients need to conduct procedures such as setting the medicinal solution cartridge and the injection needle, confirmation by pulling the cartridge to ensure secure attachment and test injection.

In this procedure, it is found that 'confirmation by pulling the cartridge system before setting the units of the test injection', has the lowest success rate. In the procedure, patients confirm that the cartridge is firmly mounted to the device in the trial test injection. A test injection is important to prevent the mixture of air in the injection device. If patients do not perform the confirmation procedure in the right way, it is not ensured that test injection has been correctly performed. In this study, we classify the type of examinees who could not accomplish the confirmation procedure by four parameters: age, length of use of the pre-improved device, length of use of a pen-type insulin injection device, and length of supervision by medical experts.

The investigation consists of two trials that are performed at certain intervals for the same examinee. The total number of examinees is 589, the number of failed examinees in the first trial is 199 and the number of failed examinees in the second trial is 264. We also compare the results of the trials.

In this study, we utilize K-means algorithm with the number of clusters determined by BIC and obtain knowledge on the tendency of patients who cannot accomplish the procedure. After performing K-means clustering, we compare the distribution of the failed examinees in each trial as mentioned above. To do this, we utilize the specialization coefficient, which is obtained as follows:

- Calculate the relative frequency of the number of examinees in each cluster who could not accomplish the procedure.
- Calculate the relative frequency of the number of all examinees in each cluster.
- Obtain the ratio of the value of a) to b).

The reason that we use the ratio of relative frequency rather than the ratio of the number of examinees is that we can normalize the ratio so as to let it be 1 if the distribution of the failed examinees is the same as the distribution of all examinees for each cluster.

To compare the tendency of the failed examinees in detail, we also obtain the specialization coefficient of the groups, which is classified by cluster, gender and age.

We calculate BIC for K whose value is between 1 and 8 (Fig. 10) and found the minimum value of BIC is given by K=3 for our target data. We therefore applied this value to the K-means algorithm for the profile data of examinees that fail to accomplish the operation in the first trial.

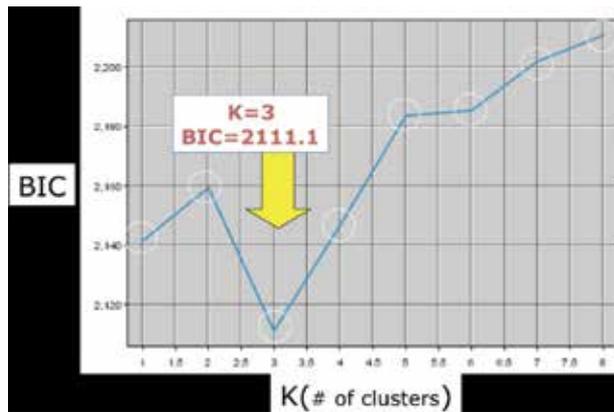


Fig. 10. The relationship between BIC and K (the number of clusters) for the target data.

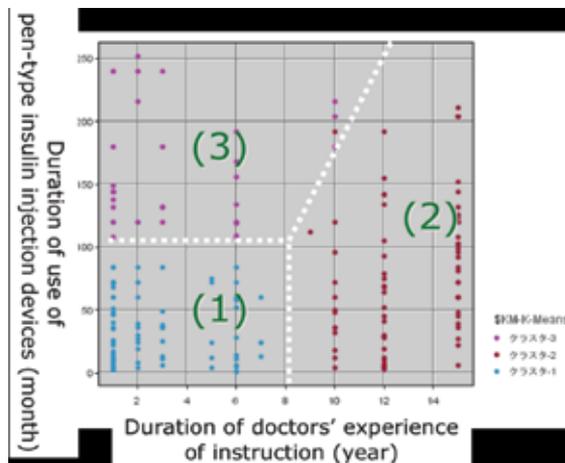


Fig. 11. The resultant clusters.

The result of the K-means algorithm is shown in Figure 11. From this, we can see that two parameters, the length of use of a pen-like insulin injection device and the length of supervision by medical experts can classify the failed examinees in the first trial. Cluster 1 corresponds to the group of examinees in which both the examinees and the medical experts supervising them have had a short career, Cluster 2 to examinees whose supervisors have had a relatively long career and Cluster 3 to examinees who have had a long career to some extent.

Since the classification by the K-means algorithm depends on the source data obtained by investigation, this classification might not be universal for all users of any insulin injection device. However, at least the data obtained in this investigation have the structure shown in Fig. 11.

	Total number of examinees	A: Relative Frequency (%)
Cluster1	215	36.9
Cluster2	265	45.5
Cluster3	103	17.7

Table 3. The total number of examinees in each cluster and its relative frequency.

1 <sup>st</sup> research	The number of failed examinees	B: Relative Frequency (%)	Specialization Coefficient (B/A)
Cluster1	82	41.2	1.12
Cluster2	78	39.2	0.86
Cluster3	39	19.6	1.11

Table 4. The number of failed examinees in each clusters and its specialization coefficient (1<sup>st</sup> research).

2 <sup>nd</sup> research	The number of failed examinees	C: Relative Frequency (%)	Specialization Coefficient (C/A)
Cluster1	117	44.3	1.20
Cluster2	103	39.0	0.86
Cluster3	44	16.7	0.94

Table 5. The number of failed examinees in each clusters and its specialization coefficient (2<sup>nd</sup> research).

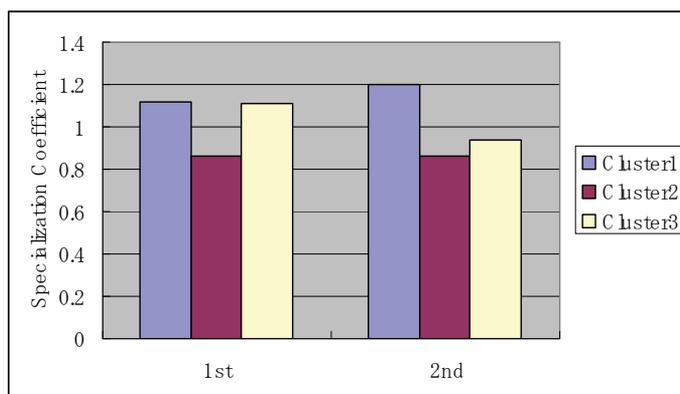


Fig. 12. The distribution of specialization coefficient in each research.

The specialization coefficients of the number of failed examinees are shown in Table 3, 4, 5 and Fig. 12. From these, we can see that Cluster 2 has less value than the other clusters. This can be interpreted as medical experts with a long career tending to succeed in guiding the examinees to accomplish the operation, and shows the importance of proper guidance by experienced experts.

It can also be seen that although in the first trial, Cluster 1 has as great a value as Cluster 3, in the second trial, Cluster 1 has a greater value than Cluster 3. This result suggests that, in the first trial, the career of the medical experts mainly has an effect on the result since it was the first time for examinees to use the device. In addition, in the second trial, the result can be considered as being affected by some examinees with a short use period of insulin injection device forgetting or omitting the operation since they do not recognize its importance.

The fact to be noticed is that the number of failed examinees in the second trial increased compared with the number in the first trial. This indicates that there is a general tendency to omit or forget the operation after a while, even if they first do it in the right way. It is important to notify patients of the correct method of using the device repeatedly.

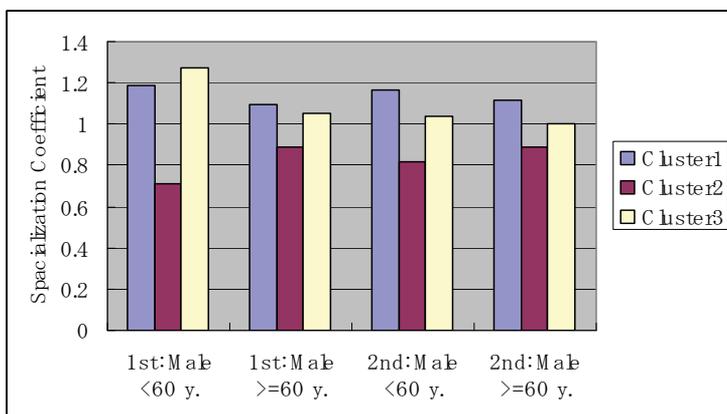


Fig. 13. The distribution of specialization coefficient for male examinees in each research.

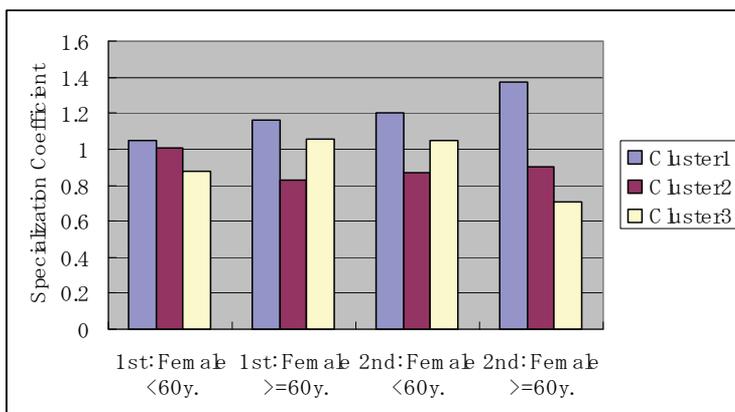


Fig. 14. The distribution of specialization coefficient for female examinees in each research.

Figure 13 and Fig. 14 show the results of analysis by gender and age. We divided the examinees into two groups by age, more or less than 60 years old, since the retirement of the principal or spouse usually changes the style of life and may cause a change in thinking and feeling about the operation.

There is a characteristic for male examinees younger than 60 years old in the first trial that Cluster 3 has the biggest value among the clusters. This shows that the rate of examinees with a long career who operate incorrectly is high. This suggests that the experienced examinees tend to omit the procedure due to the habituation of the pen-like injection device. Female examinees have the feature that the rate of examinees in Cluster 2 whose age is less than sixty in the first trial is higher than the other cases, and the rate in Cluster 1 of those over 60 years old is high in the second trial. This shows that, in the case of females younger than sixty, there is a tendency to fail for the first time even if the medical expert supervising them is fully experienced, and that, in the case of females older than sixty, they fail if they are not very used to the pen-type injection device and their supervisors have had a short career. This suggests that there is a tendency for young females to find usage of the new device not easy, and that older females tend to forget the correct operation after a while.

This indicates that, at the time of switching to a new device, young males need to be paid attention to if they have a long career and tend to operate by habituation, and young females should be paid attention to if they have a problem with a new device even if medical expert supervisor is fully experienced, and that older females need to be trained in the correct operation after a certain period time.

#### 4. Summary and conclusion

In this chapter, we introduced some clustering algorithms, their variation to analyze questionnaires and investigations and their applications.

We first pointed out the weakness of agglomerated hierarchical clustering algorithm in application to multi-choice question, introduced selection co-occurrence measure to estimate the difference of how respondents select options and compared its result with the results of TwoStep algorithm method. The selection co-occurrence measure is applied to the vectors that correspond to the options and whose elements are assigned a value 1 or 0, depending on the response of the corresponding respondent. It is a distance-like index counting up the NAND value of each element of the two vectors. TwoStep algorithm supplied by SPSS is applied to the vectors each of which correspond to the respondents and whose elements correspond to each option and is suitable in the case to classify a number of respondents.

Next, we show the method to fix the suitable number of clusters in K-means algorithm, whose derivation has been done by trial and error in practice. We reformulated K-means algorithm as likelihood functions and applied Bayesian information criterion (BIC).

As the application of agglomerated hierarchical clustering algorithm with selection co-occurrence measure, we reviewed the analysis of the multiple-choice question part of the questionnaire about systemic transdermal absorbent preparations and identified the combinations of the simultaneously selected options that occur frequently as the set of opinions expressed by medical experts and obtained the results as follows:

- The combination of systemic transdermal absorbent preparations that medical experts deal with is mainly cardiac drugs and asthma drugs.
- The reason that medical experts select transdermal patches as a dosage form is that no damage is imposed on the gastrointestinal tract and that it maintains its effect for many

hours. In particular, we can make a good guess that the statement 'no damage is imposed on the gastrointestinal tract' comes from the fact that it is better for patients who have difficulty in eating or who have to take many oral drugs if the gastrointestinal tract is not affected.

- Many medical experts think that the design of the transdermal patch is desirable from the viewpoint of the prevention of medical accidents and that the mark should be integrated for the same efficacy and be widely recognized. The answer regarding the design can be attributed to the fact that, in such a situation, nobody can find what medicine the patch is after it has been put on the patient.
- Although pharmacists and nurses tend to answer that therapeutic classification mark and product name label should be promoted, doctors tend to be satisfied with current situation. This suggests that the people who deal directly with patients and medicines feel the necessity of the mark and the label strongly.

As the application of K-means algorithm whose number of clusters is determined by BIC, we review the analysis of the profile data of the examinees of the investigation to calculate how many patients can handle the injection device for antidiabetic drugs in the right way after guidance, and to identify difficult procedures for users. We classified the type of examinees who could not accomplish the procedure of 'confirmation by pulling the cartridge system before setting the units of the test injection', which has the lowest success rate, and obtained knowledge on the patients to whom information should be provided with special attention.

First, we applied the K-means algorithm to the profile data such as age, length of use of the pre-improved device, length of use of a pen-type insulin injection device and length of supervision by medical experts and calculated BIC to find the number of clusters approximating the data well. Next, in order to analyze the tendency of the failed examinees, we compare the specialization coefficient in each cluster obtained as the ratio of the relative frequency of the number of failed examinees to the relative frequency of all examinees. The investigation consists of two trials that are performed at certain intervals for the same examinee. We compared the results in each trial and found that, in the first trial, the career of medical experts mainly has an effect on the result, but, in the second trial, some examinees with a short use period of the insulin injection device forget or omit the operation since they do not recognize its importance. We also found that medical experts with a long career tend to succeed in guiding the examinees to accomplish the operation. This shows the importance of proper guidance by experienced experts.

We divided each cluster into examinee groups by gender and age (more or less than 60 years old) and analyzed the groups in the same way. It was found that, at the time of switching to the new device, young males with a long career need to be paid attention to if they tend to operate by habituation, and young females should be paid attention to if they have a problem with a new device even if the medical expert supervisor is fully experienced, and that older females need to be trained in the correct operation after a certain period time. Of course, it is insufficient only to analyze the data with which we deal in the above review. Since data mining is an activity to utilize data and to suggest improvement of operation or service, it is important to feed back the results to appropriate people and/or organizations. We fed back our results to medical experts and pharmaceutical companies by reporting the results and presentation at an academic conference. It is not necessarily easy to measure effectiveness of our feedback, since the investigations introduced in this review are

conducted on a too broad scale to be conducted again as verification in the short term. However, we suppose that it is important to measure to what extent the results contribute to the improvement of the operations, e.g. better compliance to safely treat the injection device for antidiabetic drugs in our review.

## 5. Acknowledgement

The author would like to thank Dr. F. Tsuchiya and Prof. Dr. M. Ohkura for suggestion of the studies that are reviewed in this chapter and fruitful discussion.

The author would also like to thank Dr. H. Furukawa, TOA EIYO LTD., and Sanofi-Aventis K.K. for the valuable investigation data.

## 6. Reference

- Berry, M.; Linoff, G. (1997) *Data mining techniques: for marketing, sales and customer support*, John Wiley & Sons Inc.
- Kimura, M.; Ohkura, M.; Tsuchiya, F. (2006) Application of data-mining techniques to questionnaires about safety of drug use, *Proceedings of IEA2006 Congress* (CDROM).
- Kimura, M.; Furukawa, H.; Ohkura, M.; Tsuchiya, F. (2006) Study on the safety of the usage of antidiabetic drug injection devices, *Proceedings of IEA2006 Congress* (CDROM).
- SPSS inc. (2003). *Clementine 8.0 Algorithms Guide*, Integral Solutions Limited, pp.53-59.
- Shimodaira, H.; Itoh, S.; Kubokawa, T.; Takeuchi, K. (2004). *Information theoretic model selection and its confidence evaluation* (in Japanese), a book in the series of Frontier of Statistical Science, Iwanami-Shoten.
- Zhang, T.; Ramakrishnan, R.; Livny, M.(1996). BIRCH: an efficient data clustering method for very large databases, *Proceedings of the 1996 ACM SIGMOD international conference on Management of data*, pp.103-114.

# Data Mining in the Molecular Biology Era – A Study Directed to Carbohydrates Biosynthesis and Accumulation in Plants

Renato Vicentini and Marcelo Menossi

*Universidade Estadual de Campinas, Departamento de Genética e Evolução  
Brazil*

## 1. Introduction

The last revolutionary advance in biological research was driven by the concepts of molecular biology, which links information about genetic traits to DNA and proteins (Katagiri, 2003). The central dogma of molecular biology is that information stored in DNA is processed through RNA to produce proteins that execute various cellular functions. At the basic hierarchical level, information flow is from the genes, to mRNA, and proteins. The current adopted approaches are based in the attribution of the biological phenomena to the actions of one or a few genes. Unfortunately with these approaches it is difficult to reconstitute a model for a whole biological system by simple combining the information they generate. Recently, ideas about linear flow of information have been revised to permit the development of a more integrated view of cellular functions as being distributed among groups of elements that all interact within large networks.

Over the last years, there has been an explosion of information in biology. The sequencing of more than a hundred genomes has detailed thousands of genes. High-throughput technologies, especially DNA arrays, generate information about the expression of these genes under different conditions (Ideker et al, 2001). The next step towards a more comprehensive understanding of the biological system is to integrate these data into a conceptual framework. The ultimate goal is to understand biological systems in sufficient detail to enable accurate and quantitative predictions about the behaviors of biological systems, including predictions of the effects of modifications of the system (Katagiri, 2003).

Systems biology applies the methods of biology, mathematics, computer science, engineering, and physics to understanding living systems. Developing accurate predictive methods capable of scaling from genotype to phenotype can be approached through systems biology coupled with genomics and gene expression data (Fig. 1). The way to build bridges from molecular biology to physiology is to recognize that a network of interacting genes and proteins is a dynamic system evolving in space and time according to laws of reaction, diffusion and transport (Tyson, 2007).

A system can be generally defined as a network of interacting elements receiving certain inputs and producing certain outputs. Quantitative models are generated as tools to aid understanding and prediction of system output in response to the environment and system inputs. Models are simplified representations of system dynamics, usually in mathematical

form (Janes & Yaffe, 2006). In biology, a system can be described equally at the level of gene action, biochemical pathway, an organelle, a cell, an organ, a whole organism, or a community (Hammer et al., 2004).

## 2. Genomics and others 'omics' approaches

The use of high-throughput technologies in recent year has generated extensive information on the various levels of cellular and developmental process in many organisms. The major challenge, however, remains in the integration of this information towards a broad understanding on how the different biological layers interact to form higher functional units (Girke, 2003).

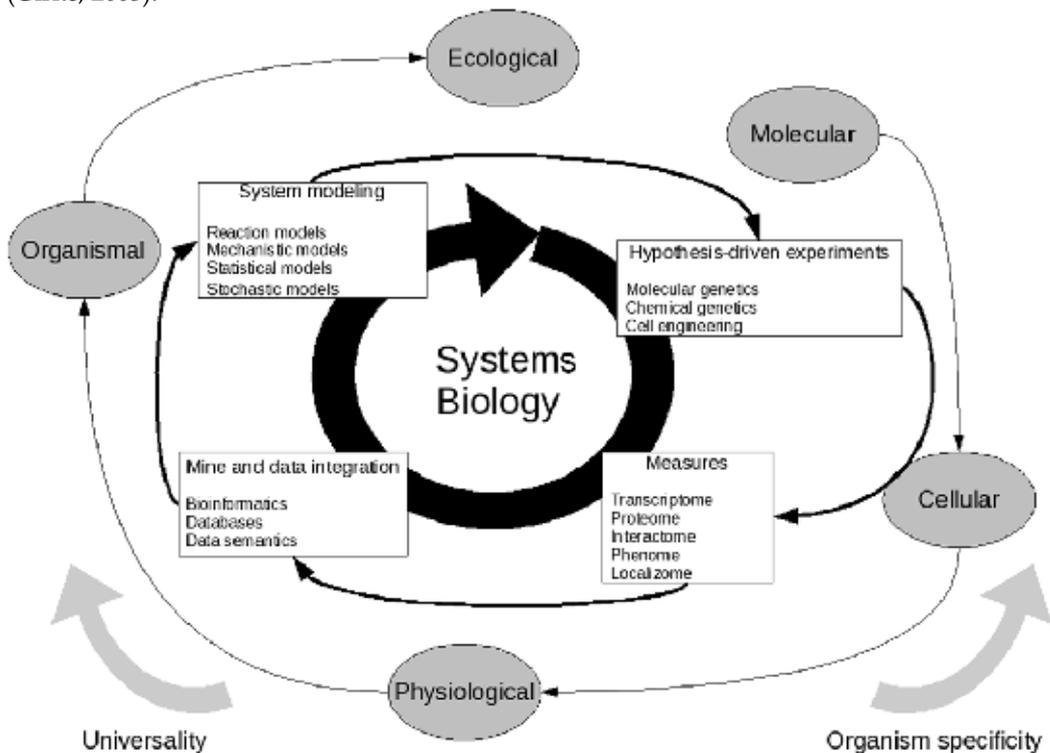


Fig. 1. Systems biology viewed as a combination of 'omic' approaches, mine and data integration, system modeling and hypothesis-driven experiments. The goal of systems biology is to generate a model of the whole organism that describes processes across the layers of biological organization (molecular, cellular, physiological, organismal, and ecological). With the availability of complete genome and transcriptome sequences, functional genomics and proteomic approaches are used to map the transcriptome, proteome, interactome, phenome, and localzome of a given organism. Computational methods can be then used to model biological process based on integrated data. The hypotheses lead to the design of new experiments that start a new round of the cycle. Although the individual components (in molecular level) are unique to a given organism at a particular time, the grouped components (cellular, physiological, and organismal) share similarities with other large-scale processes.

Using genomic techniques, we can now identify all the genes in an organism. Moreover, using microarray and proteomic techniques, we now have the ability to resolve which genes are activated or inactivated during development or in response to an environmental change. The DNA arrays approach has allowed investigators to evaluate simultaneously thousands of genes, measuring which ones are turned on or turned off in a genome in response to an experimental treatment. Proteomic methods reveal the proteins translated from the mRNA molecules that are the direct result of gene expression, and can be used to determine not just whether a protein is present but how much of the protein is present and in some cases how active it is. On a genome-wide scale, combining data from several unrelated measure profiling experiments can result in more detailed and informative module assignments (Ge et al., 2003). Such integration should not only improve functional annotation but also help to formulate biological hypotheses. However, identifying all the genes and proteins in an organism is comparable with listing all the parts of a machine. Although such a list provides a catalog of the individual components, it is not sufficient to understand the complexity underlying the engineered object (Minorsky, 2003). The massive acquisition of data in molecular and cellular biology has led to the simulations of biological systems. Simulations, increasingly paired with experiments, are being successfully and routinely used by computational biologists to understand and predict the quantitative behavior of complex system, and to drive new experiments (Di Ventura, 2006).

### 3. Data mining and modeling

Data mining refers to the analysis of large-scale data sets for the purpose of general inference and the extraction of specific information that relates to some initial data of interest (Kersey, 2006). Data mining for large data sets can be quite different from extracting data about individual sequences. The types of analysis performed are generally similar, but the development of automated procedures is usually essential as the data volume is increased. Large data sets enable 'knowledge discovery' through the identification of patterns within the data.

The analysis of biological systems requires extensive use of bioinformatics resources for data management, mining, and modeling. An additional dimension of complexity will be added by the incoming data from new technologies. The importance of automatic methods for linking gene, protein and literature data has grown as the number of known sequences has increased (Kersey, 2006; and Fig. 2). To manage these multidimensional data sets, it will be necessary to develop a new generation of integrated databases to allow complex queries across diverse types combined with new algorithms and flexible software for mining and simulating network architectures (Girke, 2003).

#### 3.1 'Omics' data mining

DNA array analysis is exploratory and very high dimensional, and the primary purpose is to generate a list of differentially regulated genes that can provide insight into the biological phenomena under investigation. While it is possible to interpret DNA array experiments a single gene at a time, most studies generate long lists of differentially expressed genes whose interpretation requires the integration of prior biological knowledge. This prior knowledge is stored in various public and private databases (Fig. 2) and covers several aspects of gene function and biological information (Coulibaly & Page, 2008). Below are described the main features of the types of bioinformatics tools and analysis that permit mining the microarray data (Coulibaly & Page, 2008).

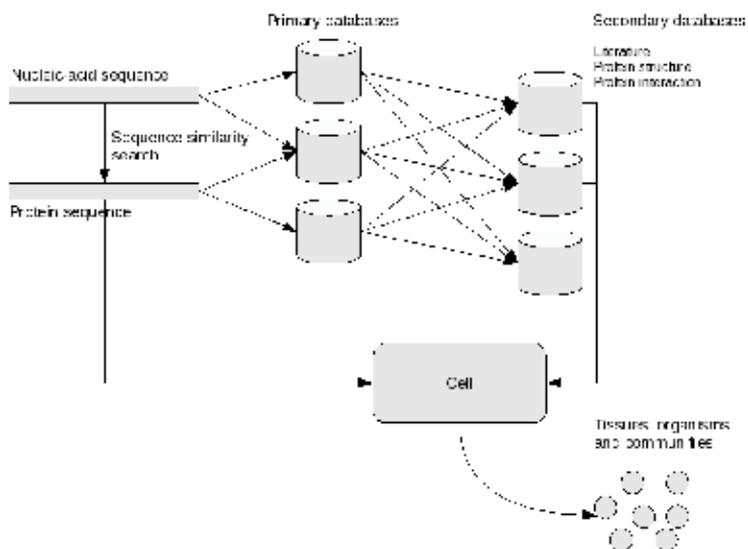


Fig. 2. A schematic representation of a workflow for bioinformatics analysis. The workflow performs the analysis from sequence to functional annotation, protein structure and literature. The complete analysis may be carried out entirely within a bioinformatics warehousing system, or as a sequence of separate operations performed in different environments. Starting with the sequence of a gene or protein, identical and/or similar sequences are identified in the primary databases. The database records describing these sequences also contain general information about the sequence, and accurate links to secondary database. Finally, modeling analyzes are performed to build a cell model than can be expanded to a complex systems biology modeling of tissues, organisms or communities.

*Functional annotation tools.* The goal of these tools is to relate the expression data to other attributes such as cellular localization, biological process, and molecular function. The most common way to functionally analyze a gene list is to gather information from the literature or from databases covering the whole genome.

*Gene coexpression analysis tools.* In most DNA array studies, gene expressions are measured on a small number of arrays or samples; however, large collections of arrays are available in several databases. These tools provide the opportunity to analyze the transcriptome by pooling gene expression information from multiple data sets. It has been demonstrated that genes which protein products cooperate in the same pathway.

*Gene network analysis.* Genes and their protein products are related to each other through a complex network of interactions. By using systems biology approach we can analyze the behavior and relationships of all the elements in a particular biological system to arrive at a more complete description of how the system functions. These analyses permit the development of models for gene regulatory networks, the display of a simplified view of large amount biological components, and the associate network nodes and edges with biological information.

*Biological pathway resources.* One of the downstream applications of the reconstruction of a gene regulatory network or the identification of clusters of functionally related genes is to associate the genes and their interconnections with known metabolic pathways.

### 3.2 Literature data mining

The systematic application of automated high-throughput molecular biology techniques has led to the generation of an immense quantity of data. However, the interpretation of these data is still dependent on inference drawn from hypothesis-driven experimentation, the details of which reside in free-text articles. The value of bioinformatics data is thus utterly dependent on the ability to make the correct links from the sequences to the scientific literature (Kersey & Apweiler, 2006).

Text mining refers to computational methods for the automatic analysis of semi-structured text, and has gained considerable attention in recent years in the molecular biology field (Hakenberg et al., 2004). Literature data mining has progressed from simple recognition of terms to extraction of interaction relationships from complex sentence (Hirschman et al., 2002). The current research focused on tasks needing limited linguistic context and processing at the level of words, like identifying protein names or on tasks relying on word co-occurrence and pattern matching.

### 3.3 System modeling

Biologists commonly use the term 'model' for verbal or graphical description of a mechanism underlying a cellular process. However, the mathematical modeling and description of complex biological process has become more important in the last years. Knowledge about the system is essential and needs to be formalized for the chosen framework (Di Ventura, 2006).

Clearly, useful modeling will depend on having a large amount of high-quality quantitative information about all aspects of biological processes, and many new types of data have to be systematically determined. Recently, Bayesian network, a probabilistic graphic model representation, has been widely used to analyze expression data. Compared with clustering analysis, Bayesian network has the advantage of uncovering conditional independency among genes, which provides a promising way to survey direct interaction of gene regulation (Chen et al., 2006). A complete understanding of regulation requires quantitative information about kinetic laws and the concentrations of metabolites and enzymes. This quantitative knowledge in combination with the known network of metabolic pathways allows the construction of mathematical models that describe the dynamic changes in metabolite concentrations over time. A variety of pathways modeling tools such a CellDesigner (Funahashi et al., 2003) has been developed which simplify model construction and analysis. Most of these tools are able to store and exchange models in the Systems Biology Markup Language (SBML, Hucka et al., 2003) and to fit parameters for a given set of experimental data.

## 4. Systems biology

The emerging field of systems biology is a new branch of biology that attempts to discover and understand biological properties that emerge from the interactions of many elements (Minorsky, 2003). Systems biology examines the structure and dynamics of cellular and organismal function, rather than the characteristics of isolated parts of a cell or organism. These systems may be gene expression networks, signal transduction pathways, metabolic networks, or combinations of them. In contrast to previous approaches, systems biology endeavors to quantitatively model and simulate complex biological process and systems comprising thousands of chemical compounds and reactions.

There is an emphasis on complexity and large data sets, which are typically produced by a variety of high-throughput genomic, proteomic and metabolomic techniques. The major reason behind the increasing interest in systems biology is that progress in molecular biology, particularly in genomics, proteomics, and high-throughput measurements, is enabling scientists to collect comprehensive data sets on the mechanisms underlying responses to perturbations on a biological system (Minorsky, 2003).

A systems approach to understanding biology can be described as an interactive process that includes (1) data collection and integration of all available information, (2) system modeling, (3) experimentation at a global level, and (4) generation of new hypotheses (Gutiérrez et al., 2005; and Fig. 1).

## 5. Case study of carbohydrates biosynthesis and accumulation in plants

Recently, genes encoding the enzymes of carbohydrates biosynthesis in plants have been isolated, cloned, and used in experiments to transform the plant to increase or decrease expression of the enzyme with the goal of altering the carbohydrates accumulation. However, results of this reductionist approach towards understanding sucrose accumulation have fallen short of expectations, mainly because of the complex interactions among the multitude of simultaneous processes. Insights into the complex interactions will require systems-level approaches, from the molecular, biochemical, and physiological levels. Carbohydrates accumulations in plants are the products of a large complex network of interactions that can be analyzed from several perspectives. Increasing evidence suggests that sucrose is involved in signaling to modulate expression of genes controlling transporters and storage proteins, division and differentiation of cells, and accumulation of storage products (Lunn and MacRae 2003). Each of the reactions involved is controlled by activation of specific genes in response to an interaction of the genotype of the plant and the environment.

### 5.1 Approach

Availability of abundant, high-quality data sets from DNA array expression experiments has stimulated rapid progress in gene networks analysis for a variety of plant species. By examining correlated expression patterns between genes, we can infer new functions for previously uncharacterized genes and identify potential causal relationships between regulators and their targets (Srinivasainagendra et al., 2008). The idea that correlated expression implies biologically relevant relationships between gene products were use in the present study.

We use the AraCyc database to select the sucrose biosynthesis and degradation pathways. The AraCyc database (Table 1) is a reference database for visualization of *Arabidopsis thaliana* biochemical pathways. With the selected enzymes from the sucrose AraCyc pathways, we performed a search by coexpressed genes in the *Arabidopsis* mRNA arrays experiments storage in the Nottingham Arabidopsis Stock Center (NASC). For this purpose we utilized the Cress-express mining tool (Table 1) with the RMA processing method in four specific carbohydrates experiments. Cress-express estimates the coexpression between an user-provided list of genes and all genes from Affymetrix Ath1 platform using up to 1779 arrays. Cress-express also performs pathway-level coexpression (PLC). PLC identifies and ranks genes based on their coexpression with a group of genes. The tool has the data processed with a variety of image processing methods: RMA, MAS5, and GCRMA (Srinivasainagendra et al., 2008).

With the identified coexpressed genes, we constructed a Bayesian networks by using the BNArray tool (Table 1). It allows the reconstruction of significant submodules within regulatory networks using an extended subnetwork mining algorithm. Under this framework and the assumption of parameter independence, an initial Bayesian network structure is learned from the training data and a user specified prior network. From this initial network, greedy search algorithm with random restarts was performed to get the highest score posterior network to avoid local maxima. Finally, we obtained an optimized Bayesian network that maximizes the Bayes factor.

We utilized the Kinetikon and SABIO-RK softwares (Table 1) for mining kinetic reactions in the biological literature. Finally, the network representation of carbohydrates biosynthesis pathways of coexpressed genes was created by Cytoscape network tool (Table 1), and the simulation of the kinetic dynamics of the network created was modeling by the CellDesigner (Table 1).

Tools and databases	Description	Reference
AraCyc	<i>Arabidopsis thaliana</i> database of metabolic pathways and enzymes.	(Mueller et al., 2003)
Cress-express	Coexpression analysis tool for Arabidopsis microarray expression data that computes patterns of correlated expression between user-entered query genes and the rest of the genes in the genome.	(Srinivasasainagendra et al., 2008)
BNArray	R package for construct gene regulatory networks from microarray data by using Bayesian network.	(Chen et al., 2006)
CellDesigner	Structured diagram editor for drawing gene-regulatory and biochemical networks based on standardized technologies.	(Funahashi et al., 2003)
Cytoscape	Software environment for the large scale integration of molecular interaction network data.	(Shannon et al., 2003)
KEGG	Kyoto Encyclopedia of Genes and Genomes that link lower-level information (i.e. proteins) with higher-level information (i.e. pathways).	(Kanehisa et al., 2000)
Kinetikon	A collection of detailed knowledge about biochemical reaction kinetics.	( <a href="http://kinetikon.molgen.mpg.de">http://kinetikon.molgen.mpg.de</a> )
SABIO-RK	Web-based application that contains information about biochemical reactions, their kinetic equations with their parameters.	(Wittig et al., 2006)

Table 1. Software and databases adopted in this study.

## 5.2 Results

We used the AraCyc database to look up AGI codes for genes associated with the sucrose biosynthesis and degradation pathway. We used Cress-express to determine the degree to

which these genes are coexpressed with other. Using Cress-express tool default parameters, we performed a coexpression analysis of all genes, comparing them both to each other as well as to all other genes represented on the ATH1 array. The Fig. 3 presents the diagrams showing the network obtained by the coexpression genes related with the genes present in the pathways. Each connection in the network represents a pair of connected genes exhibit expression correlation. This study revealed that many genes are highly coexpressed with each other.

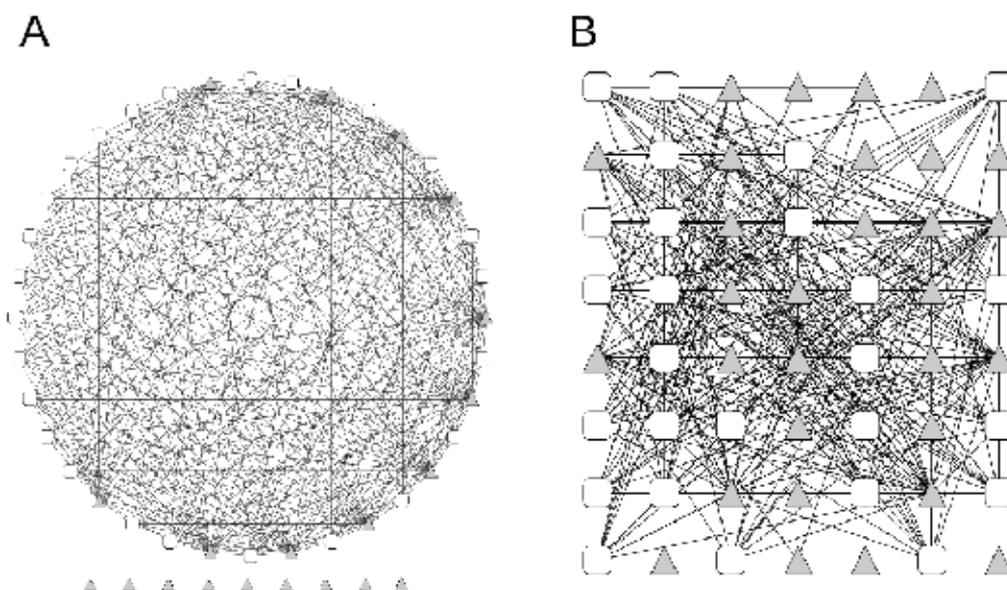


Fig. 3. Network representation of protein coexpressed with proteins in sucrose biosynthesis pathway (A) and sucrose degradation pathway (B). In these networks, the 'bait' genes are represented by triangles.

Bayesian network modeling was used to capture regulatory interactions between genes based on genome-wide expression measurements. To build the network, we integrated expression data from 'bait' genes (from biosynthesis pathway) with others coexpressed genes in the *Arabidopsis* genome. The generated network (Fig. 4) is a set of relationships that defines the probability that genes function together; this regulatory interactions among genes and their directions are derived from expression data. The subnetwork in Fig. 4 is the one that best explains the observed data, and provides the direction of regulatory interactions.

We use text mining systems (Table 1) that supports researchers in their search for experimentally obtained parameters to build our kinetic model. The kinetic modeling of biological reaction networks deals with the question of how the concentrations of substances change over time. The dynamics are determined by (a) the concentrations of substrates and products, (b) the structure of the whole reaction network, and (c) the kinetic parameters of the involved enzymes (Hakenberg et al., 2004).

We based our model (Fig 5) on the knowledge contained in the Kyoto Encyclopedia of Genes and Genomes (KEGG; Kanehisa & Goto, 2000). KEGG is a set of databases that constitute a computer representation of biological knowledge at different levels, i.e.



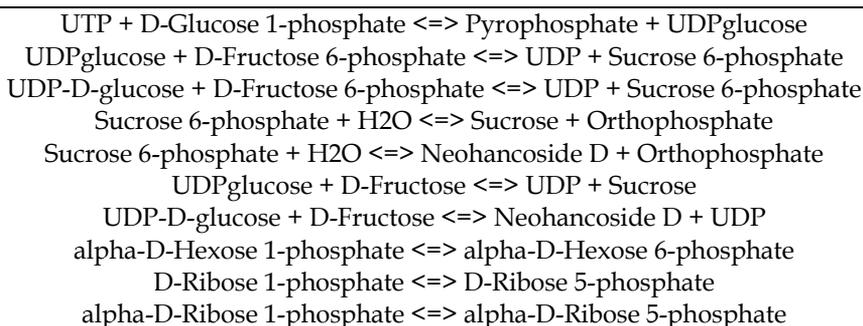


Table 2. Reactions list used to build the kinetic model.

Kinetic modeling of biological systems depends on sets of different kinetic data and values measured in expensive experiments. Such data are published in thousands of scientific articles. It is infeasible for humans to read and analyze this number of papers with reasonable time constraints (Hakenberg et al., 2004). To simulate our model we mined literature and databases for kinetic data using the SABIO-RK and Kinetikon databases. Finally, we applied the kinetic data to the model of carbohydrate biosynthesis pathway (Fig. 5). The result is the concentration curve showed in Fig. 6.

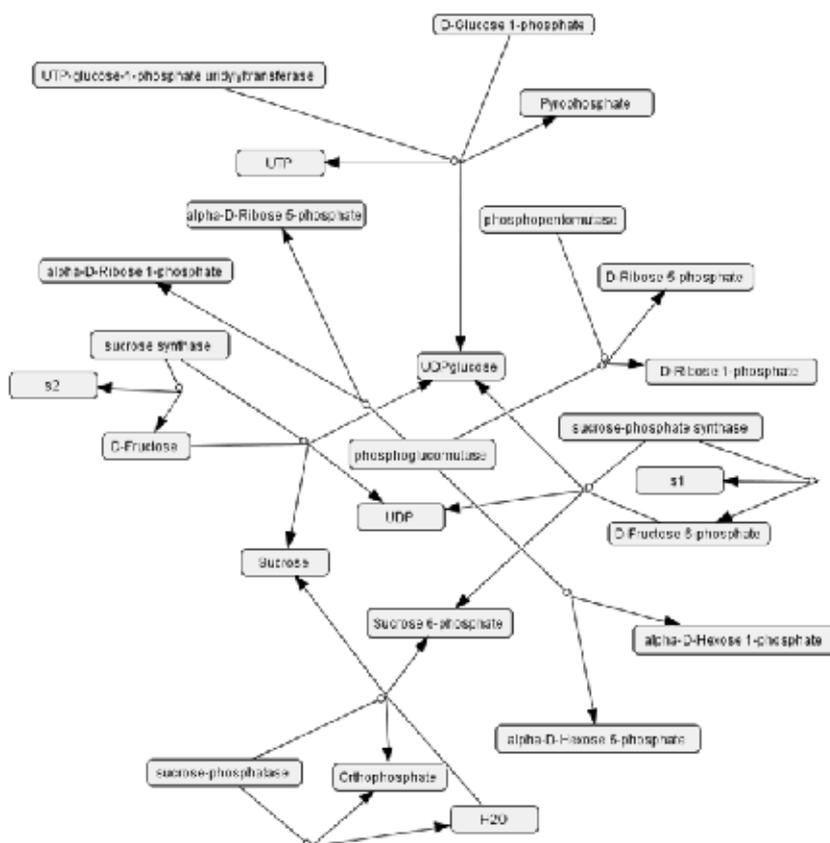


Fig. 5. Kinetic model for reaction present in sucrose biosynthesis.

## 6. Conclusion

The immensity of the information explosion in biology presents the challenge of how these data sets will be organized and mined in standard and easily accessible forms. To successfully iterate through the cycle described in Fig. 1, we need high-quality quantitative data and a flexible software platform that integrates arbitrary data types and that is coupled to data visualization and analysis tools. It will allow scientists to study and understand biological dynamics, to create a detailed model of cell function, and to provide system level knowledge for the network of signaling that are essential for physiological function. To reach this goal, we must adopt mathematical and computational methods for modeling and simulating complex biological systems (Girke, 2003).

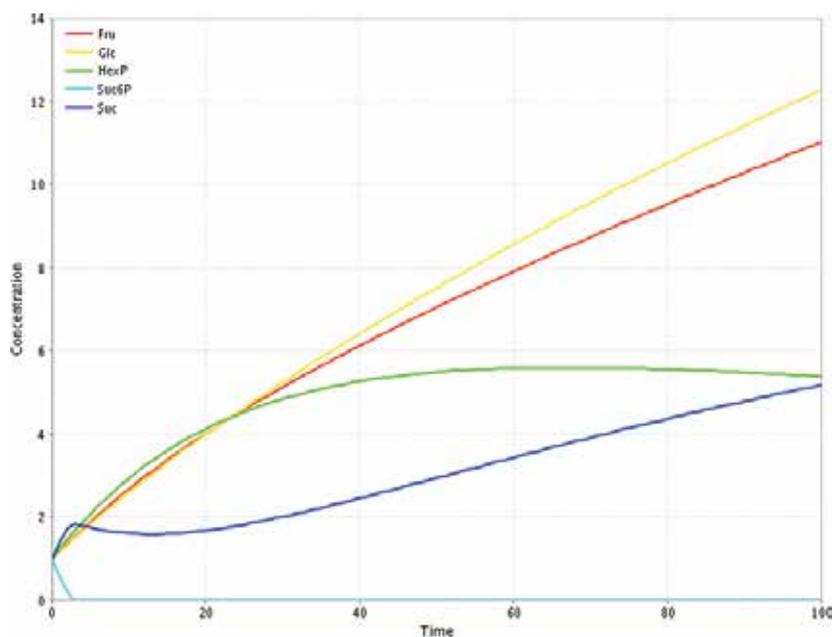


Fig. 6. Simulated concentration curve (in time) for metabolites of sucrose biosynthesis pathway. Fru: fructose; Glc: glucose; HexP: hexose-phosphate; Suc6P: sucrose 6-phosphate; Suc: sucrose.

## 7. References

- Chen, X.; Chen, M. & Ning, K. (2006). BNArray: an R package for constructing gene regulatory networks from microarray data by using Bayesian network. *Bioinformatics*, 22, 2952-2954, ISSN 1367-4803
- Coulibaly, I. & Page, G.P. (2008) Bioinformatic tools for inferring functional information from plant microarray data II: Analysis beyond single gene. *International Journal of Plant Genomics*, 893-941, ISSN 1687-5370
- Di Ventura, B.; Lemerle, C.; Michalodimitrakis, K. & Serrano, L. (2006). From *in vivo* to *in silico* biology and back. *Nature*, 443, 527-533, ISSN 0028-0836
- Funahashi, A.; Tanimura, N.; Morohashi, M. & Kitano, H. (2003). CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. *BIOILICO*, 1, 159-162, ISSN 1741-8364

- Ge, H.; Walhout, A.J.M. & Vidal, M. (2003). Integrating 'omics' information: a bridge between genomics and systems biology. *TRENDS in Genetics*, 19, 551-559, ISSN 0168-9525
- Girke, T.; Ozkan, M.; Carter, D. & Raikhel, N.V. (2003). Towards a modeling infrastructure for studying plant cells. *Plant Physiology*, 132, 410-414, ISSN 0032-0889
- Gutiérrez, R.A.; Shasha, D.E. & Coruzzi, G.M. (2005). Systems biology for the virtual plant. *Plant Physiology*, 138, 550-554, ISSN 0032-0889
- Hakenberg, J.; Schmeier, S.; Kowald, A.; Klipp, E. & Leser, U. (2004). Finding kinetic parameters using text mining. *OMICS*, 8, 131-152, ISSN 1536-2310
- Hammer, G.L.; Sinclair, T.R.; Chapman, S.C. & van Oosterom E. (2004). On systems thinking, systems biology, and the *in silico* plant. *Plant Physiology*, 134, 909-911, ISSN 0032-0889
- Hirschman, L.; Park, J.C.; Tsujii, J.; Wong, L. & Wu, C.H. (2002). Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, 18, 1553-1561, ISSN 1367-4803
- Hucka, M.; Finney, A.; Sauro, H.M.; Bolouri, H.; Doyle, J.C.; Kitano, H. et al. (2003) The systems biology markup language (SBML): A medium for representation and exchange of biochemical network models. *Bioinformatics*, 19, 524-531, ISSN 1367-4803
- Ideker, T.; Thorsson, V.; Ranish, J.A.; Christmas, R.; Buhler, J.; Eng, J.K.; Bumgarner, R.; Goodlett, D.R.; Aebersold, R. & Hood, L. (2001). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, 292, 929-934, ISSN 0036-8075
- Janes, K.A. & Yaffe, M.B. (2006). Data-driven modeling of signal-transduction networks. *Nature Reviews Molecular Cell Biology*, 7, 820-828, ISSN 1471-0072
- Kanehisa, M. & Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28, 27-30, ISSN 1362-4962
- Katagiri, F. (2003). Attacking complex problems with the power of systems biology. *Plant Physiology*, 132, 417-419, ISSN 0032-0889
- Kersey, P. & Apweiler, R. (2006). Linking publication, gene and protein data. *Nature Cell Biology*, 8, 1183-1189, ISSN 1465-7392
- Lunn, J.E. & MacRae, E. (2003). New complexities in the synthesis of sucrose. *Current opinion in plant biology*, 6, 208-214, ISSN 1369-5266
- Minorsky, P.V. (2003). Achieving the *in silico* plant. System biology and the future of plant biological research. *Plant Physiology*, 132, 404-409, ISSN 0032-0889
- Mueller, L.A.; Zhang, P. & Rhee, S.Y. (2003). AraCyc: A biochemical pathway database for Arabidopsis. *Plant Physiology*, 132, 453-460, ISSN 0032-0889
- Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N.S.; Wang, J.T.; Ramage, D.; Amin, N.; Schwikowski, B. & Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*. 13, 2498-2504, ISSN 1088-9051
- Srinivasasainagendra, V.; Page, G.P.; Mehta, T.; Coulibaly, I. & Loraine, A.E. (2008) CressExpress: A tool for large-scale mining of expression data from Arabidopsis. *Plant Physiology*, 147, 1004-1016, ISSN 0032-0889
- Tyson, J.J. (2007). Bringing cartoons to life. *Nature*. 445, 823, ISSN 0028-0836
- Wittig, U.; Golebiewski, M.; Kania, R.; Krebs, O.; Mir, S.; Weidemann, A.; Anstein, S.; Saric, J. & Rojas I. (2006) SABIO-RK: Integration and curation of reaction kinetics data. In: *Data Integration in the Life Sciences*, 94-103, Springer Berlin / Heidelberg, ISBN 978-3-540-36593-8

# Microarray Data Mining for Biological Pathway Analysis

Miyoung Shin and Jaeyoung Kim

*School of Electrical Engineering and Computer Science, Kyungpook National University,  
Korea*

## 1. Introduction

In recent years, microarray gene expression studies have been actively pursued for extracting significant biological knowledge hidden under a large volume of gene expression profiles accumulated by DNA microarray experiments. Particularly great attentions have been paid to a variety of data mining schemes for gene function discovery [Eisen et al., 1998], disease diagnosis [Saiki et al., 2008], pathway analysis [Werner et al., 2008], pharmaceutical target identification [Corn et al., 2007], and etc. Out of these, the pathway analysis is one of the most significant problems in current bioinformatics researches. Pathway analysis concerns about identifying significant pathways, which are the groups of genes actively involved in some biological processes, based on the gene expression profiles. By doing so, our objective is to understand the role of such biological processes in a given experiment condition and their associated gene activities. In this chapter, we investigate several computational techniques that are often used in a variety of contexts for pathway analysis. Specifically, we first give a brief overview of microarray gene expression profile data and some biological resources available to be used for pathway analysis. Then we examine three different approaches, i.e. clustering-based methods, gene-based methods, and Gene set-based methods, all of which can be employed for understanding biological pathways in various environments. Subsequently we perform some case studies with the leukemia disease data and finally conclude this chapter with some remarks and discussions.

### 1.1 Overview of microarray gene expression profile data

DNA microarray is generally a glass or plastic substrate, or silicon chip, onto which tens of thousands of DNA molecules (*probes*) are deposited in a regular grid-like pattern [Zhang, 2006; Draghici, 2003]. Each grid spot corresponds to a DNA sequence of a specific gene. The idea of a microarray is to detect the presence and abundance of specific DNA molecules (*targets*) in biological samples of interests. For this purpose, from two mRNA samples (a test sample and a control sample), cDNAs are obtained and labeled with fluorescent dyes and the solutions including the labeled targets are hybridized on the surface of the chip. Then the chips are scanned to read the expression intensities emitted from the labeled and hybridized targets. Thus, by doing so, the microarray enables us to monitor the expression levels of tens of thousands of genes simultaneously. Once the raw microarray gene expression profiles are obtained in this way, several pre-processing steps are usually

performed which include data transformation, data filtering, missing value imputation, data normalization, and etc. Consequently, for the analysis of microarray data, the gene expression profiles are typically employed in a  $p \times n$  matrix form as follows:

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ & & \vdots & \\ x_{p1} & x_{p2} & \cdots & x_{pn} \end{pmatrix}$$

where  $x_{ij}$  denotes the expression intensity of the  $i$ th gene ( $i=1, \dots, p$ ,  $p$  is the number of genes) in the  $j$ th sample ( $j=1, \dots, n$ ,  $n$  is the number of experiment conditions). Thus, the procedure of obtaining gene expression data matrix from a collection of raw data obtained by microarray experiments can be summarized as in Fig. 1.

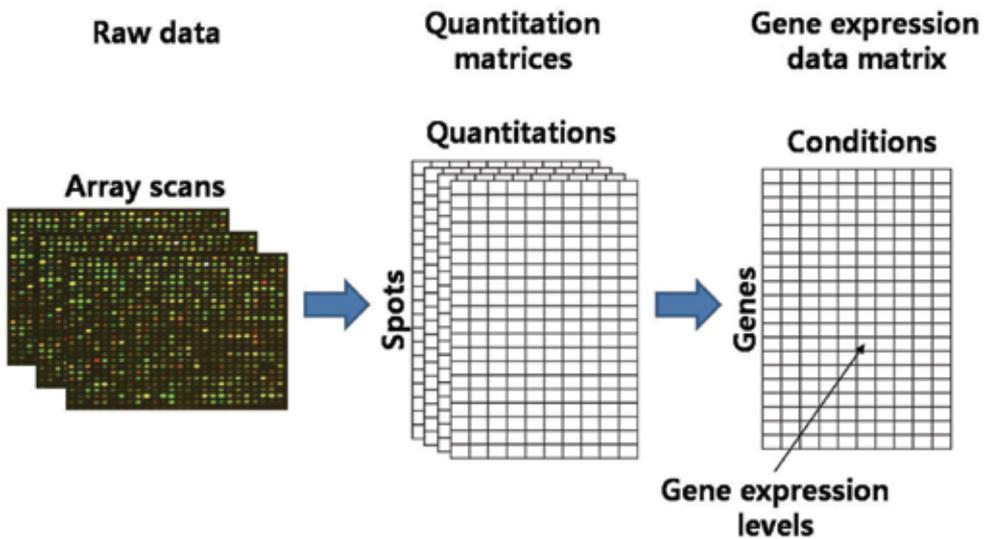


Fig. 1. The procedure of obtaining gene expression data matrix from a collection of raw microarray data (adapted from [Brazma et al., 2001]).

## 1.2 Biological resources for pathway analysis

There exist several types of biological resources which can be used for pathway analysis, such as pathway databases, gene annotation databases, Gene Ontology, and etc. In particular, approximately 179 kinds of pathway databases are currently available and some of them are given in Table 1. These databases are easily accessible through a web and include a collection of pathway maps already known for various organisms. The pathway maps represent the knowledge on molecular interaction and reaction networks in metabolic pathways, genetic networks, signaling pathways, and complexes. Some of the pathway databases provide application program interfaces (APIs) enabling us to access pathway data in various downloadable formats. Among them, the KEGG (Kyoto Encyclopedia of Genes and Genomes) database is one of the most well-known extensive pathway databases

including the information about the genes of 181 organisms, their associated pathways and graphical diagrams, which can be downloadable in KGML and XML formats. Similarly, there are some other useful pathway databases such as BioCyc, GenMapp, BioCarta, and etc.

Pathway DB	Organisms	Pathway Types	Downloadable Formats
KEGG	181(varied)	metabolic, genetic, signaling, complexes	KGML, XML
BioCyc	E.Coli, human(20 others)	metabolic and complexes	BioPax, SBML
GeneMAPP	human, mouse, rat, fly, yeast	metabolic, signaling, complexes	MAPP format
Reactome	human, rat, mouse, chicken, fugu, zebrafish	metabolic, signaling, complexes	SBML, MySQL
BioCarta	human, mouse	metabolic, signaling, complexes	Just Images
TransPATH	human, mouse	Signaling, genetic	XML

Table 1. Some examples of currently available pathway databases (adapted from <http://bioinformatics.ca/>)

These pathway databases are also useful for evaluating and interpreting the analysis results of microarray gene expression profiles from the biological aspects. Fig. 2 shows an example of the acute myeloid leukaemia pathway map in KEGG pathway database.

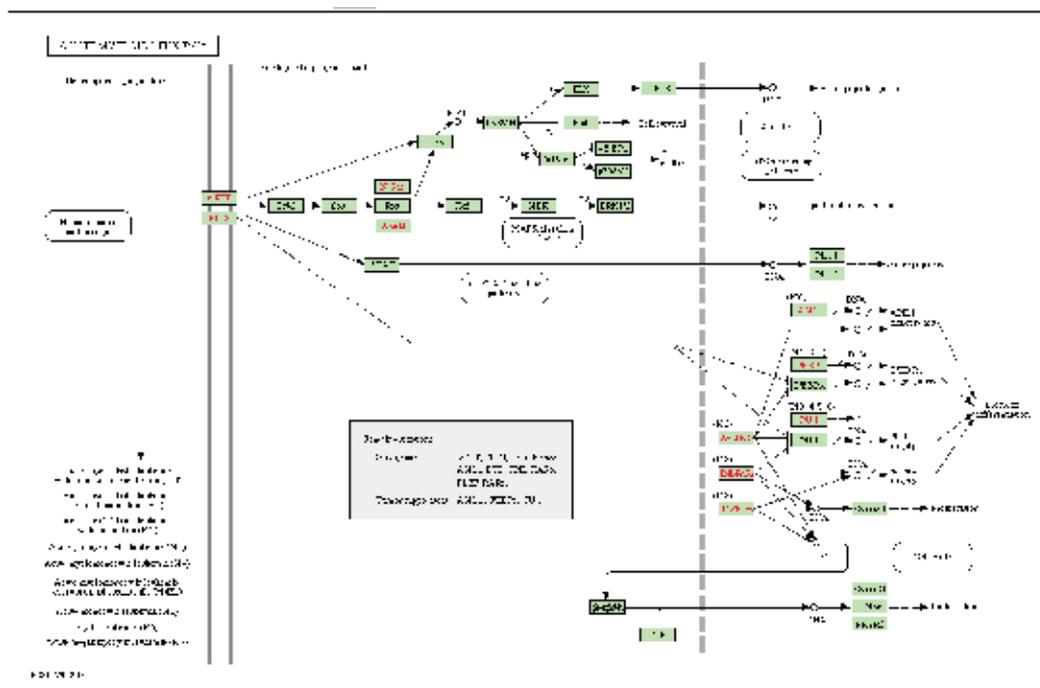


Fig. 2. An example of acute myeloid leukemia Pathway map in KEGG pathway database.

## 2. Computational methods for pathway analysis

There are several computational approaches for pathway analysis which employ microarray gene expression profiles and *a priori* known biological knowledge such as Gene Ontology, KEGG pathway database, and etc. According to the aim of microarray experiments and analyses, an appropriate computational method can be chosen. Here three different approaches are detailed. First, the clustering-based methods are based on the assumption that the genes having similar expression profiles would have similar biological functions. Thus, the genes showing similar expression Patterns under a series of conditions are grouped into a cluster and its corresponding common functional pathways are identified. By doing so, the functional pathways of uncharacterized genes in a cluster may be possibly conjectured from the pathway categories of characterized genes in the same cluster. Second, the gene-based methods are the ones that can be applied for two-grouped samples data (e.g., treatment vs. control, cancer vs. normal). That is, their main purpose is to identify the differentially expressed genes (DEGs) between two groups and find out which functional pathways these DEGs are significantly involved in. Third, the gene set-based methods are intended to identify significant pathways (i.e. gene-sets) showing good differential expression between two groups from the candidate gene-sets each of which is generated by taking all the genes belonging to a certain pathway. The approach of gene set enrichment analysis (GSEA) by Subramanian et al. (2005) belongs here. Unlike earlier two approaches, this enables us to identify the most significant pathways in a unified analytical framework by employing *a priori* known biological knowledge along with gene expression profiles for the analysis.

### 2.1 Clustering-based methods

Conventionally cluster analysis for identifying groups of objects having similar characteristics. In our context, clustering methods are employed to generate groups of genes, i.e. gene clusters, showing similar expression profiles. The clustering-based approach for pathway analysis is based on the assumption that the co-expressed genes are of similar biological functions. Thus, once gene clusters are generated, the corresponding common functional pathways for each cluster are identified. For this purpose, Fisher's exact test [Trajkovski et al., 2008] can be used with some biological resources like pathway databases or gene annotation databases. Also, for cluster generation, any clustering methods are possibly used. Here we describe one of the most popular methods for microarray data analysis, which is hierarchical clustering.

#### 2.1.1 Cluster generation by hierarchical clustering

To generate gene clusters, two types of hierarchical clustering methods can be used, top-down and bottom-up approaches. The top-down approach is to start with one large cluster consisting of all genes and keep dividing them into smaller clusters until they become singleton clusters. On the other hand, the bottom-up approach is to start with as many singleton clusters as the gene size and keep grouping together two closest genes or clusters until they reach a single large cluster consisting of all genes. For gene expression data analysis, the bottom-up approach is generally used and the clustering results are Given in a tree-like graph, called *dendrogram*.

According to the way of defining the distance between two clusters in hierarchical clustering, which is called *linkage method*, some variations [Han et al., 2000] are possible as follows.

- **Single Linkage:** The distance between two clusters  $C_A$  and  $C_B$  is defined as the minimum distance between any two genes each belonging to  $C_A$  and  $C_B$ , respectively. This method has the characteristics that the between-cluster distances are relatively small while the within-cluster distances are relatively large.

$$d(C_A, C_B) = \min_{i \in C_A, j \in C_B} d_{ij} \quad (1)$$

- **Complete Linkage:** The distance between two clusters  $C_A$  and  $C_B$  is defined as the maximum distance between any two genes each belonging to  $C_A$  and  $C_B$ , respectively. This method has the characteristics that the within-cluster distances are relatively smaller than other linkage methods, since the maximum distance between two genes within a cluster is minimized. Because of this reason, it is the most popularly used.

$$d(C_A, C_B) = \max_{i \in C_A, j \in C_B} d_{ij} \quad (2)$$

- **Average Linkage:** The distance between two clusters  $C_A$  and  $C_B$  is defined as an Average of the distances between any of two genes each belonging to  $C_A$  and  $C_B$ , respectively. Here  $n_A$  is the number of genes in a cluster  $C_A$  and  $n_B$  is the number of genes in a cluster  $C_B$ . This method is considered as a compromise between single linkage and complete linkage.

$$d(C_A, C_B) = \frac{1}{(n_A + n_B)} \sum_{i \in C_A} \sum_{j \in C_B} d_{ij} \quad (3)$$

Also, several types of distance or similarity measures can be used for cluster generation, which include Euclidean distance, Manhattan distance, correlation coefficient, and etc. Each of these measures is described as below.

Distance measures	Formula
Euclidean distance	$d_{ij} = \sqrt{\sum_{k=1}^n (x_{ik} - x_{jk})^2}$
Manhattan distance (city block)	$d_{ij} = \sum_{k=1}^n  x_{ik} - x_{jk} $
Correlation Coefficient	$\rho_{ij} = \frac{\sum_{k=1}^n (x_{ik} - \bar{x}_{i\cdot})(x_{jk} - \bar{x}_{j\cdot})}{\sqrt{\sum_{k=1}^n (x_{ik} - \bar{x}_{i\cdot})^2} \sqrt{\sum_{k=1}^n (x_{jk} - \bar{x}_{j\cdot})^2}}, \quad -1 \leq \rho_{ij} \leq 1$ $d_{ij} = 1 - \rho_{ij}$

### 2.1.2 Pathway analysis for gene clusters

Once gene clusters showing similar expression profiles are obtained by clustering method, for each cluster, the most representative common functional pathways need to be found. For this purpose, Fisher's exact test [Trajkovski et al., 2008] can be applied to estimate the probability of having *at least*  $x$  genes in a given cluster annotated by a specific functional pathway. The probability of having exactly  $x$  genes, out of  $k$  genes in a cluster, annotated by a specific functional pathway  $S$  is given as follows.

$$P(X = x | N, M, k) = \frac{\binom{M}{x} \binom{N-M}{k-x}}{\binom{N}{k}} \quad (4)$$

Here  $N$  is the number of total genes on a microarray,  $M$  is the number of genes *a priori* known as belonging to a specific functional pathway  $S$ ,  $N-M$  is the number of genes not included in  $S$  out of total  $N$  genes. The Fisher's score  $p$ -value is thus calculated by using Eq. (5). For each of gene clusters, the corresponding significant pathways are extracted based on these  $p$ -values. In usual, the significance of functional pathways for a cluster are judged with  $p$ -value  $< 0.05$ .

$$p = 1 - \sum_{i=0}^{x-1} \frac{\binom{M}{i} \binom{N-M}{k-i}}{\binom{N}{k}} \quad (5)$$

## 2.2 Gene-based methods

In analyzing the gene expression profiles consisting of two groups of samples (e.g. normal vs. cancer), some genes are expected to have significant difference in expression levels between two groups. We call these genes *differentially expressed genes* (DEGs), which are taken as the significant ones in a given experiment. Thus, the gene-based approach for pathway analysis is intended to identify DEGs first and then discover in which functional pathways the DEGs are actively involved. Prior to finding DEGs, some pre-processing like data filtering, log-transformation and normalization is generally performed on microarray expression profiles.

### 2.2.1 Identifying DEGs

When two-grouped microarray sample data are given, the differentially expressed genes can be identified by using one of the following methods, such as  $k$ -fold change, signal-to-noise ratio and  $t$ -test.

- $k$ -Fold Change [Draghici, 2003; Schena et al., 1996]: This method is to find significant genes by calculating the ratio of the averaged expression intensities over each group of samples as follows.

$$\phi(i) = \frac{\mu_A(i)}{\mu_B(i)} \quad (6)$$

Here  $\mu_A(i)$  and  $\mu_B(i)$  denote the averaged expression intensities of the  $i$ th gene over the samples each belonging to group A and group B, respectively. If such ratio for a gene is larger than the threshold, the gene is considered as being significantly changed in expression levels between two groups. In usual, by taking an arbitrary threshold  $k=2$  or 3, the DEGs are identified.

- Signal-to-Noise Ratio (SNR) [Golub et al., 1999]: This method is to find significant genes by calculating the difference between the averaged expression intensities of two groups divided by the sum of their standard deviations. The SNR is computed as follows.

$$\rho(i) = \frac{\mu_A(i) - \mu_B(i)}{\sigma_A(i) + \sigma_B(i)} \quad (7)$$

In Eq. (7),  $\mu_A(i)$  and  $\mu_B(i)$  denote the averaged expression intensities of the  $i$ th gene over the samples each belonging to group A and group B, respectively. Also,  $\sigma_A(i)$  and  $\sigma_B(i)$  denote the standard deviations of group A and group B, respectively. Once the SNRs for entire genes are computed and arranged in a decreasing order, the corresponding graph is typically shown as in Fig. 3. The DEGs are taken from the both ends of the graph.



Fig. 3. A typical form of SNR graph for the entire gene-list where the genes are arranged in a decreasing order of SNR values

- $t$ -test [Tusher et al., 2001; Dudoit et al., 2002]: This method is to find significant genes by calculating  $t$ -statistic and using  $t$ -distribution to obtain  $p$ -value. The  $t$ -statistic for a specific gene is computed as in Eq. (8).

$$t(i) = \frac{\mu_A(i) - \mu_B(i)}{\sqrt{\frac{\sigma_A(i)^2}{n_A} + \frac{\sigma_B(i)^2}{n_B}}} \quad (8)$$

Here  $\mu_A(i)$  and  $\mu_B(i)$  denote the averaged expression intensities of the  $i$ th gene over the samples each belonging to group A and group B, respectively. Similarly,  $\sigma_A(i)$  and  $\sigma_B(i)$

denote the standard deviations of group A and group B, respectively, while  $n_A$  and  $n_B$  denote the numbers of samples in group A and group B, respectively. As the absolute value of  $t$ -statistic for a specific gene is larger, the corresponding gene would have smaller  $p$ -value which implies higher significance in a statistical sense. Once the  $t$ -statistics and  $p$ -values of all genes are obtained, we can take the genes having  $p$ -values less than a given significance level (generally, 0.01 or 0.05) as the DEGs.

### 2.2.2 Pathway analysis for DEGs

To identify functional pathways significantly involved by the DEGs, the Fisher's exact test can be used in the same way as described earlier, except that the meanings of some parameters are different. Specifically, in Eqs (4) and (5),  $k$  corresponds to the total number of the identified DEGs and  $x$  is the number of genes included in a specific functional pathway  $S$  out of  $k$  DEGs. Once the Fisher's scores are computed for each of all the genes, significant functional pathways can be identified by taking the pathways having less than a specific  $p$ -value. Also, the specific roles of the DEGs in these functional pathways might be understood by further analysis of the inter-relationships among the gene expression profiles of the DEGs.

### 2.3 Gene set-based methods

As a gene set-based method for pathway analysis, the gene set enrichment analysis (GSEA) approach has been attracting lots of attentions lately. In particular, the GSEA employs microarray expression profiles and *a priori* known biological resources in a unified analytical framework to identify significant pathways. That is, it generates the candidate gene-sets of interest, where a gene-set consists of the genes belonging to a specific pathway, by using *a priori* known biological resources such as pathway databases, gene annotation databases, literatures, and etc. Then, the significance of each candidate gene-set (i.e. pathway) is evaluated by using microarray gene expression profiles. Specifically, for each gene-set, the enrichment score is computed and its statistical significance is estimated. The detailed steps of GSEA are summarized in the following [Subramanian et al., 2005; Taskesen, 2006]:

- *Step 1*: Computation of enrichment scores for gene-sets

This step is to compute an enrichment score of a given gene-set  $A$  gene-set consists of the genes *a priori* known as being involved in a specific pathway and many candidate gene-sets can be constructed by using pathway databases, Gene Ontology, and etc. To compute ES, the entire gene-list should be rearranged in the order of Ranking statistic such as SNR or Fisher's criterion [Kim et al., 2008]. Then, with the ordered gene-list, the Kolmogorov-Smirnov(KS) score is computed for each gene-set. For KS score, the empirical cumulative distribution functions For  $P_{hit}$  and  $P_{miss}$  are employed as shown in Eq. (11).

$$P_{hit}(i) = \sum_{j=1}^i \frac{E(j)}{N_H} \tag{11}$$

$$P_{miss}(i) = \sum_{j=1}^i \frac{(1 - E(j))}{N - N_H}$$

Here  $P_{hit}$  is the empirical cumulative distribution function of which cumulative sum becomes 1 when the first  $i$  genes in the ordered gene-list completely match the genes

included in a specific gene-set  $S$ . On the other hand,  $P_{\text{miss}}$  is the one of which cumulative sum becomes 1 when there is no match between them. In Eq. (11),  $E(j)$  is 1 if the  $j$ th gene in the ordered gene-set is included in a given gene-set (i.e. hit), or is 0 if the  $j$ th gene in the ordered gene-set is not included in a given gene-set (i.e. miss). Also  $N$  is the total number of genes in entire gene-list and  $N_H$  is the number of genes in a specific gene-set  $S$ . At times, the ranking statistic can be used for computing  $P_{\text{hit}}$  and  $P_{\text{miss}}$  [Taskesen, 2006]. For instance, assuming that the SNR-based gene ranking is used,  $E(j)$  can be replaced by the SNR value of the  $j$ th gene in the ordered gene-list while  $N$  is the sum of the SNR values for total genes included in entire gene-list and  $N_H$  is the sum of the SNR values for the genes included in a specific gene-set  $S$ . In either way, by taking the maximum deviation between  $P_{\text{hit}}$  and  $P_{\text{miss}}$ , as shown in Eq. (12), the ES for a specific gene-set  $S$  is obtained.

$$ES(S) = \max_{i=1, \dots, N} |P_{\text{hit}}(i) - P_{\text{miss}}(i)| \quad (12)$$

- *Step 2:* Estimation of statistical significance of ES

To estimate statistical significance of the ES obtained in step 1,  $k$  random permutations of a given microarray expression profile data on labels are generated and for each permutation, the corresponding ES is calculated. By using the computed ESs for  $k$  permuted data, we can obtain the null distribution of ES which will be used to calculate a *nominal* p-value of the ES.

- *Step 3:* Adjustment for multiple hypothesis testing

The estimated significance level is now adjusted to account for multiple hypothesis testing. First, the ES for each candidate gene-set is normalized for the gene-set size by dividing it with the mean of the  $k$  ESs for the permuted data obtained in Step 2. Then, for each normalized ES, the proportion of false positives is controlled by calculating the corresponding false discovery rate (FDR) or the family-wise error rate (FWER).

### 3. Case studies

For experiments, the leukemia dataset published by Golub et al. [1999] was used. The leukemia disease [Knudsen, 2006] is known as a cancer of the blood or bone marrow that affects blood-forming cells (usually white blood cells) in the bone marrow. The white blood cells are divided into granulocytes, monocytes, and lymphocytes. Leukemias starting in granulocytes or monocytes are called *myeloid leukemias*, and leukemias starting in lymphocytes are called *lymphocytic leukemias*. Further, leukemias are divided into acute and chronic. In acute leukemias, the cells cannot mature properly, whereas in chronic leukemias, the cells mature partly but do not obtain their full function. The Golub's leukemia dataset originally consists of 7129 human gene expression profiles shown in total 78 leukemia samples of two classes including 47 acute lymphoblastic leukemia (ALL) samples and 25 acute myeloid leukemia (AML) samples. Out of them, Golub et al. used 38 samples for training and 34 samples for testing to diagnose the subtypes of the leukemia. For our analyses, these 38 training samples including 27 ALL samples and 11 AML samples were used.

#### 3.1 Pathway analysis with clustering

For cluster analysis, the data-filtering and data-transformation were applied for microarray expression profiles in the same way as in [Dudoit et al., 2002]. Specifically, the data-

transformation was done to replace the expression intensities less than 100 with 100 and expression intensities larger than 16000 with 16000. Also, the log-transformation with base 2 was applied for all the gene expression profiles. After such data-transformation, data-filtering was performed to remove the genes whose the difference between the maximum and the minimum of expression intensities is less than 500 or the ratio of the maximum to the minimum of expression intensities is less than 5. As a result, 3051 genes were remained and used for cluster generation. Clusters were generated by using complete-linkage hierarchical clustering with Euclidean distance and the results are shown in Fig. 6. By visual inspection of Fig. 6, the four was chosen to be an appropriate number of clusters. Thus, we generated four clusters and obtained the results as shown in Fig. 7, where four clusters include 666 genes, 656 genes, 427 genes, and 1302 genes, respectively.

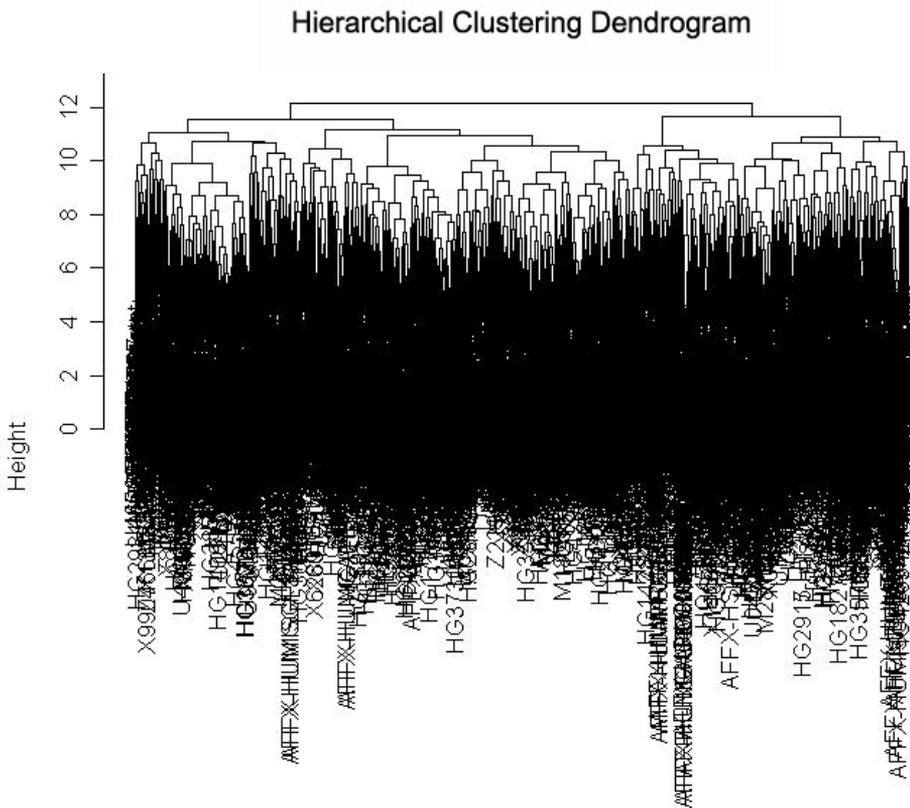


Fig. 6. Clustering result of Golub's leukemia dataset by complete-linkage hierarchical clustering with Euclidean distance.

To identify significant functional pathways involved in each cluster, we used KEGG pathway database to perform Fisher's exact test on the genes in a cluster. By using  $p$ -value $\leq 0.05$ , the significant pathways for each of the four clusters were identified and shown in Tables 2-5.

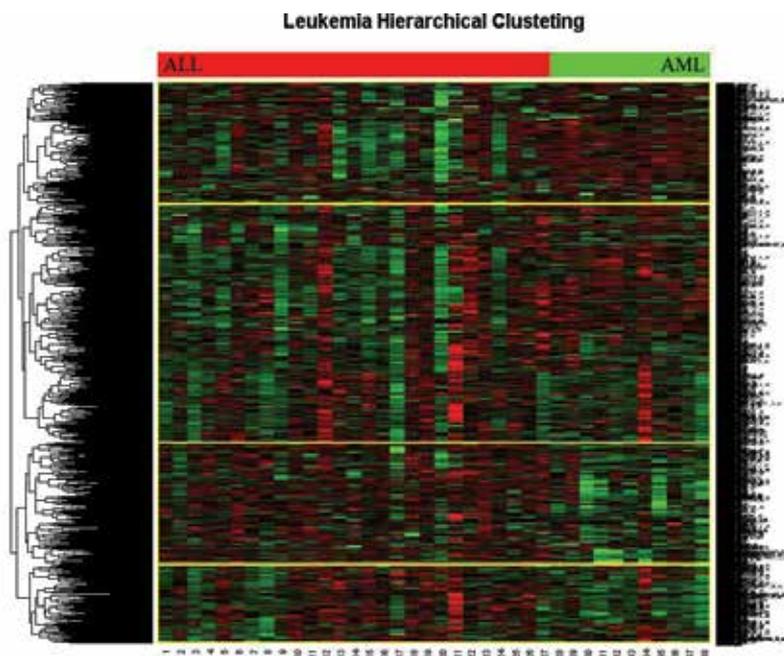


Fig. 7. Heatmap for four clusters of Golub’s leukemia dataset obtained by complete-linkage hierarchical clustering with Euclidean distance measure

functional pathways	<i>p</i> -values
Hematopoietic cell lineage	3.7e-7
Type I diabetes mellitus	2.9e-5
Antigen processing and presentation	5.1e-5
Cytokine-cytokine receptor interaction	0.00013
Epithelial cell signaling in Helicobacter pylori infection	0.00042
Cell adhesion molecules (CAMs)	0.00067
Adipocytokine signaling pathway	0.00086
Toll-like receptor signaling pathway	0.00097
Porphyrin and chlorophyll metabolism	0.0012
Acute myeloid leukemia	0.0029
Aminosugars metabolism	0.035
Glutathione metabolism	0.042

Table 2. 1<sup>st</sup> cluster’s significant functional pathways (*p* -value≤0.05)

functional pathways	<i>p</i> -values
Hematopoietic cell lineage	6.8e-7
B cell receptor signaling pathway	0.0048
Aminoacyl-tRNA biosynthesis	0.011
Insulin signaling pathway	0.016
Long-term potentiation	0.043
Alzheimer's disease	0.049

Table 3. 2<sup>nd</sup> cluster’s significant functional pathways (*p* -value≤0.05)

functional pathways	<i>p</i> -values
Metabolism of xenobiotics by cytochrome P450	0.003
Linoleic acid metabolism	0.013
Colorectal cancer	0.013
VEGF signaling pathway	0.013
Small cell lung cancer	0.015
Pancreatic cancer	0.016
gamma-Hexachlorocyclohexane degradation	0.018
Non-small cell lung cancer	0.028
Neurodegenerative Diseases	0.029
Focal adhesion	0.032
Apoptosis	0.033
Acute myeloid leukemia	0.035
Prostate cancer	0.04
Chronic myeloid leukemia	0.047

Table 4. 3<sup>rd</sup> cluster's significant functional pathways (*p* -value≤0.05)

functional pathways	<i>p</i> -values
Glioma	2.0e-5
Chronic myeloid leukemia	5.7e-5
Prostate cancer	0.0001
Cell cycle	0.00024
Glycolysis / Gluconeogenesis	0.00074
Endometrial cancer	0.00078
T cell receptor signaling pathway	0.00092
Thyroid cancer	0.00098
Non-small cell lung cancer	0.0012
Small cell lung cancer	0.0018
Regulation of actin cytoskeleton	0.0027
Adherens junction	0.0035
Long-term potentiation	0.0051
Focal adhesion	0.0072
Oxidative phosphorylation	0.0073
Natural killer cell mediated cytotoxicity	0.0081
Gap junction	0.012
Proteasome	0.014
Acute myeloid leukemia	0.015
Colorectal cancer	0.016
ErbB signaling pathway	0.016
Renal cell carcinoma	0.016
Carbon fixation	0.018
Melanoma	0.021
Bladder cancer	0.022
Huntington's disease	0.022
Pancreatic cancer	0.033
Calcium signaling pathway	0.038
Insulin signaling pathway	0.041
Neurodegenerative Diseases	0.047

Table 5. 4<sup>th</sup> cluster's significant functional pathways (*p* -value≤0.05)

According to earlier Golub's study on this leukemia dataset, it was observed that their gene expression profiles can be grouped into AML related samples, ALL-T-cell related samples, ALL-B-cell related samples. To understand the biological meaning of our analysis results, thus, we analyzed whether significant functional pathways identified for each cluster are somewhat related to AML or ALL-T-cell or ALL-B-cell types and found some interesting observations. In Tables 2 and 4, it was observed that the genes in the 1<sup>st</sup> cluster and 3<sup>rd</sup> cluster are significantly involved in AML-related functional pathways such as hematopoietic cell lineage, apoptosis, and acute myeloid leukemia. Also, in Tables 3 and 5, it was observed that the genes in the 2<sup>nd</sup> cluster are significantly involved in ALL-B-cell related pathways such as hematopoietic cell lineage and B cell receptor signaling pathway, while the genes in the 4<sup>th</sup> cluster are significantly involved in ALL-T-cell related pathways such as T cell receptor signaling pathway and cell cycle. Consequently, from our experiments, it was found that each of the four clusters is closely related to the subtypes of the leukemia which include AML, ALL-T-cell, and ALL-B-cell types.

### 3.2 Pathway analysis with DEGs

To apply the gene-based method described earlier for pathway analysis, we need to find differentially expressed genes first. For this purpose, the data-filtering and data-transformation were applied for Golub's 7129 gene expression profiles in the same way as done for pathway analysis with clustering, and 3051 genes were obtained. Out of these genes, we extracted the most 50 differentially expressed genes by using SNR and the results are shown in Fig. 8. In this figure, we can see 25 DEGs showing relatively higher expression in ALL than AML, and the other 25 DEGs showing relatively higher expression in AML than ALL. This is why the DEGs were chosen by taking the genes having the 25 largest SNRs in positive region and the 25 smallest SNRs in negative region. Also, for comparisons, we applied *t*-test for 3051 gene expression profiles and selected 50 genes as the DEGs by having  $p\text{-value} \leq 0.000004$ , as shown in Fig. 9. When the *t*-test is used for finding DEGs, the different number of DEGs can be chosen depending on the choice of *p*-value. In our case, we adjusted *p*-value in such a way to choose 50 genes, for comparative purpose with the SNR result. As seen in Figs 8 and 9, even if the same number of DEGs are chosen by SNR method and *t*-test, respectively, it is observed that the identified DEGs are quite different.

To understand significant functional pathways in which the DEGs identified by SNR method and *t*-test are actively involved, the Fisher's exact test was applied for the both cases. Table 6 shows the result of significant functional KEGG pathways identified for the SNR DEGs while Table 7 shows the pathway analysis result for the *t*-test DEGS. As seen in Tables 6 and 7, the same pathways were identified for the two different sets of 50 DEGs chosen by SNR method and *t*-test, respectively. In particular, the pathway of hematopoietic cell lineage is known as being related to both ALL and AML types, and the B cell receptor signaling pathway is related to ALL-B-cell type. Thus, the DEGs showing significant expression difference between two groups of ALL and AML are empirically found to be actively involved in hematopoietic cell lineage and B cell receptor signaling pathway. In addition, it is observed that the DEGs identified by the *t*-test seem to be more meaningful in terms of *p*-values than the DEGs identified by the SNR.



functional pathways	<i>p</i> -values
Hematopoietic cell lineage	0.0081
B cell receptor signaling pathway	0.041

Table 6. Significant functional KEGG pathways identified for 50 DEGs by SNR method (*p*-value≤0.05)

functional pathways	<i>p</i> -values
Hematopoietic cell lineage	0.005
B cell receptor signaling pathway	0.01

Table 7. Significant functional KEGG pathways identified for 50 DEGs by *t*-test method (*p*-value≤0.000004)

### 3.3 Pathway analysis by gene set enrichment analysis

To perform pathway analysis with GSEA, the *z*-score normalization was first applied for 7129 gene expression profiles. Also, 136 candidate gene-sets were generated from KEGG pathway database in such a way to take only the pathways each of which include at least 10 genes out of 7129 genes. Then, for each of the candidate gene-sets, the corresponding ES was computed, and its statistical significance level and the normalized ES were obtained with 1000 random permutations of 7129 gene expression profiles on class labels. In particular, the normalized ESs for the candidate gene-sets has a two-modal distribution as shown in Fig. 10. Thus, the most 40 significant pathways were identified as done in [Subramanian et al., 2005] by taking 20 gene-sets from the rightmost end in a positive region and 20 gene-sets from the leftmost end in a negative region. The results are shown as in Table 8.

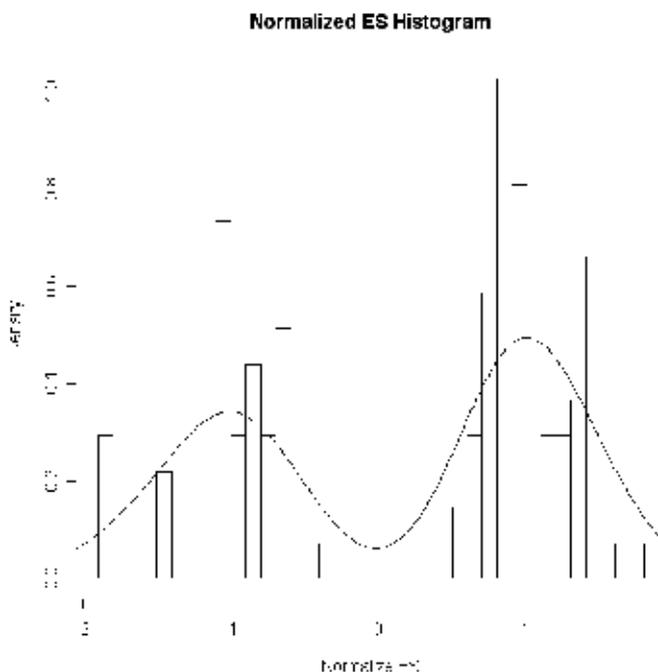


Fig. 10. The distribution of normalized ESs for the candidate gene-sets

functional pathways	Normalized ES
Pyruvate metabolism	1.759
Cell cycle	1.718
Galactose metabolism	1.511
Alanine and aspartate metabolism	1.500
Basal transcription factors	1.497
DNA polymerase	1.496
Aminoacyl-tRNA biosynthesis	1.481
Purine metabolism	1.463
Citrate cycle (TCA cycle)	1.445
Proteasome	1.440
Pentose and glucuronate interconversions	1.377
One carbon pool by folate	1.369
1- and 2-Methylnaphthalene degradation	1.360
Pyrimidine metabolism	1.327
Biosynthesis of steroids	1.311
Lysine degradation	1.296
Wnt signaling pathway	1.266
Folate biosynthesis	1.260
Butanoate metabolism	1.248
RNA polymerase	1.209
Apoptosis	-1.150
Neurodegenerative Disorders	-1.177
ECM-receptor interaction	-1.204
Metabolism of xenobiotics by cytochrome P450	-1.216
Glycosphingolipid biosynthesis - neo-lactoseries	-1.219
Cytokine-cytokine receptor interaction	-1.249
Methane metabolism	-1.280
Sphingolipid metabolism	-1.308
Leukocyte transendothelial migration	-1.323
Hematopoietic cell lineage	-1.378
Glycan structures - degradation	-1.438
Nitrogen metabolism	-1.477
Complement and coagulation cascades	-1.497
Toll-like receptor signaling pathway	-1.656
Glycosaminoglycan degradation	-1.682
Arachidonic acid metabolism	-1.754
Glutathione metabolism	-1.755
Epithelial cell signaling in Helicobacter pylori infection	-1.778
Porphyryn and chlorophyll metabolism	-1.805
Adipocytokine signaling pathway	-1.810

Table 8. The most 40 significant pathways identified by GSEA for Golub's 7129 gene expression profiles

As seen in Table 8, It is interesting that the pathways including an apoptosis pathway known as being related to cell death, or the hematopoietic cell lineage known as being related to generating an antibody in a blood were identified as significant pathways. Also, some other cancer-related pathways were found.

#### 4. Concluding remarks

In this chapter we introduced several mining methods for biological pathway analysis with microarray gene expression profiles. For pathway analysis, our concern is how to identify significant functional pathways in which many genes showing differential expression between treatment and control groups are actively involved. In earlier approaches, the computational techniques only for microarray data had been more focused than biological interpretation of the results. On the other hand, the recent approaches concentrate more on the effective use of a variety of biological resources in analyzing large volume of microarray expression data to obtain more biologically meaningful results. By applying them for a variety of fields such as drug discovery [Bild et al., 2006] or disease diagnosis[Watters et al., 2006], pathway analysis with microarray expression data can play a key role in the path to new scientific discoveries and allows us to understand what biological phenomena causes the observed expression patterns. Furthermore, by making an attempt to use new types of biological resources along with expression profiles for the analysis, our understanding of the expression data could be enhanced in a deeper manner.

#### 5. Acknowledgement

This work was supported by the Korea Science and Engineering Foundation(KOSEF) grant funded by the Korea government(MEST) (No. R01-2008-000-11089-0)

#### 6. References

- Werner, T., "Bioinformatics applications for pathway analysis of microarray data.", *Current opinion in biotechnology*, 19(1):50-4.(2008)
- Corn, P.G. et al., "Microarray analysis of p53-dependent gene expression in response to hypoxia and DNA damage," *Cancer biology & therapy*, 6(12):1858-66.(2007)
- Saiki, T. et al., "Identification of marker genes for differential diagnosis of chronic fatigue syndrome," *Molecular Medicine*, 14(9-10):599-607.(2008)
- Zhang, A., "Advanced analysis of gene expression microarray data," World scientific publishing co. (2006).
- Draghici, S., "Data Analysis Tools for DNA microarrays," Chapman & Hall/CRC(2003)
- Brazma, A. et al., "A quick introduction to elements of biology - cells, molecules, genes, functional genomics, microarrays", European Bioinformatics Institute, Draft. (2001)
- Canadian Bioinformatics Workshops, <http://bioinformatics.ca/>, Protein Pathways and Pathway Databases
- KEGG: Kyoto Encyclopedia of Genes and Genomes, <http://www.genome.ad.jp/kegg/>
- BioCyc, <http://biocyc.org/>
- K. D. Dahlquist et al., "GenMAPP: A new tool for viewing and analyzing microarray data on biological pathways," *Nature genetics* 2002 May;31(1):19-20.
- BioCarta, <http://www.biocarta.com/>

Gene Ontology, <http://www.geneontology.org/>

- Subramanian, A. et al., "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.", *Proc. Natl Acad Sci USA* 102: 15545-50. (2005)
- Eisen, M. B. et al., "Clustering analysis and display of genome-wide expression patterns," *PNAS*, 95, 14863-14868. (1998)
- Han, J. et al., *Data Mining : Concepts and Techniques*, Academic Press. (2000)
- Trajkovski, I. et al., "SEGS: Search for enriched gene sets in microarray data," *Journal of biomedical informatics* 41, 588-601. (2008)
- Schena, M. et al., "Parallel human genome analysis: microarray-based expression monitoring of 1000 genes," *Proc Natl Acad Sci U S A*, 93(20):10614-10619. (1996)
- Golub, T. R. et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring", *Science (Wash. DC)*, 286: 531.537. (1999)
- Tusher, V. G. et al., "Significance analysis of microarrays applied to ionizing radiation response," *Proc Natl Acad Sci U S A*, 98(9), 5116-5121. (2001)
- Dudoit, S. et al., "Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments," *Statistica Sinica*, 12, 111-140. (2002)
- E. Taskesen, "Sub-typing of model organisms based on gene expression data," *Bioinformatics technical University of Delft Research Assignment*. (2006)
- Knudsen, S. "Cancer diagnostics with DNA microarrays," *Wiley-Liss* (2006)
- Bild, A. H. et al., "Linking oncogenic pathways with therapeutic opportunities," *Nature reviews. Cancer*, 6(9):735-41. (2006)
- Watters, J. W. et al., "Developing gene expression signatures of pathway deregulation in tumors", *Molecular cancer therapeutics*, 5(10):2444-9. (2006)
- Kim, J.Y. et al., "Identifying biologically significant pathways by gene set enrichment analysis using Fisher's criterion," *To appear in proc. of bioscience and biotechnology* 2008.

# Development of Microsatellite Markers by Data Mining from DNA Sequences

Jingou Tong, Dan Wang and Lei Cheng  
*State Key Laboratory of Freshwater Ecology and Biotechnology  
Institute of Hydrobiology, Chinese Academy of Sciences  
P.R.China*

## 1. Introduction

### 1.1 What are microsatellites

Microsatellites are tandem repeats of 1-6 nucleotides found at high frequency in the nuclear genomes of most taxa (Beckmann and Weber, 1992). As such, they are also known as simple sequence repeats (SSR), variable number tandem repeats (VNTR) and short tandem repeats (STR). For example, (A)<sub>11</sub>, (GT)<sub>12</sub>, (ATT)<sub>9</sub>, (ATCG)<sub>8</sub>, (TAATC)<sub>6</sub> and (TGTGCA)<sub>5</sub> represent mono-, di-, tri-, tetra-, penta- and hexa-nucleotide repeats, respectively. A microsatellite locus typically varies in length between 5 and 40 repeats, but longer strings of repeats are possible. Dinucleotide, trinucleotide and tetranucleotide repeats are the most common choices for molecular genetic studies. Dinucleotides are the dominant type of microsatellite repeats in most vertebrates characterized so far, although trinucleotide repeats are most abundant in plants (Beckmann & Weber, 1992; Chen et al., 2006; Kantety et al., 2002).

Despite the fact that the mechanism of microsatellite evolution and function remains unclear, SSRs were being widely employed in many fields soon after their first description (Litt & Luty 1989; Tautz 1989; Weber & May 1989) because of the high variability which makes them very powerful genetic markers. Microsatellites have proven to be an extremely valuable tool for genome mapping in many organisms (Schuler et al., 1996; Knapik et al., 1998), but their applications span over different areas ranging from kinship analysis, to population genetics and conservation/management of biological resources ( Jarne & Lagoda 1996).

Microsatellites can be amplified for identification by the polymerase chain reaction (PCR), using two unique sequences which are complementary to the flanking regions as primers. This process results in production of enough DNA to be visible on agarose or polyacrylamide gels; only small amounts of DNA are needed for amplification as thermocycling in this manner creates an exponential increase in the replicated segment. With the abundance of PCR technology, primers that flank microsatellite loci are simple and quick to use, but the development of correctly functioning primers is often a tedious and costly process. However, once they are developed and characterized in an organism, microsatellites are powerful for a variety of applications because of their reproducibility, multiallelic nature, codominant inheritance, relative abundance and good genome coverage (Liu & Cordes, 2004).

Unlike conserved flanking regions, microsatellite repeat sequences mutate frequently by slippage and proofreading errors during DNA replication that primarily change the number

of repeats and thus the length of the repeat string (Eisen 1999). Because alleles differ in length, they can be distinguished by high-resolution gel electrophoresis, which allows rapid genotyping of many individuals at many loci for a fraction of the price of sequencing DNA. Many microsatellites have high-mutation rates (between  $10^{-2}$  and  $10^{-6}$  mutations per locus per generation, and on average  $5 \times 10^{-4}$ ) that generate the high levels of allelic diversity necessary for genetic studies of processes acting on ecological time scales.

### 1.2 Progress in the development of microsatellites

As aforementioned, the major drawback of microsatellites is that they need to be isolated and characterization before to be used for the first time. Generally, microsatellites can be developed by the following approaches:

#### 1. Cross-species amplification

Because the sequences of flanking region are generally conserved across individuals of the same species and sometimes of different species, a particular microsatellite locus can often be identified by its flanking sequences. The presence of highly conserved flanking regions has been reported for some microsatellite loci in cetaceans (Schlötterer et al., 1991), turtles (FitzSimmons et al., 1995) and fish (Rico et al., 1996), allowing cross-amplification from species that diverged as long as 470 million years ago (Ma).

In this way, the first step is to search published literature and public databases for any existing microsatellite primers for the target species or closely-related species. The availability of microsatellite markers for a given species will be a combination of past interest in that species (and related species) and the inherent success rate of microsatellite development for that taxon. There are clear differences in the frequency of microsatellite regions in the genomes of plants, animals, fungi and prokaryotes (Toth et al. 2000), and the success rate of isolating microsatellite markers often scales with their frequency in the genome (Zane et al. 2002).

Currently, many microsatellite markers are reported as primer notes in a specialized journal "Molecular Ecology Notes" (now changed as "Molecular Ecology Resources"). There is a searchable database online for any microsatellite primers published in this journal (<http://tomato.bio.trinity.edu/>). The sequences themselves are archived in GenBank, and are often submitted long before their use appears in published studies. GenBank can be searched with a web-based engine run by the National Center for Biotechnology Information (<http://www.ncbi.nlm.nih.gov/>) by typing in the species, genus or family name, the term microsatellite and selecting the Nucleotide database (Benson et al. 2008; Wheeler 2008).

#### 2. Genomic library- based method

Traditionally, microsatellite loci have been isolated from partial genomic libraries (selected for small insert size) of the species of interest, screening several thousands of clones through colony hybridization with repeat-containing probes (Rassmann et al. 1991). Although relatively simple, especially for microsatellite-rich genomes, this approach can turn out to be extremely tedious and inefficient for species with low microsatellite frequencies. Therefore, several alternative strategies have been devised in order to reduce the time invested in microsatellite isolation and to significantly increase yield.

Conventional library-screening methods, established before 2000, had low efficiency and they could be time-consuming. A repeat-enriched method by using an AFLP procedure, named as FIASCO, was reported and increased the efficiency of microsatellite isolation significantly (Zane et al. 2002).

### 1.3 Mining microsatellites from nucleotide sequences

Methods of SSR-mining have gone through a rapid evolution during the past few years. The first approaches relied on visual inspection of sequence. Although manual comparison of a small number of sequences is feasible, standard accuracy criteria are hard to establish, and this method does not scale well for multiple sequences and many microsatellite location. The efficiency of visual inspection is increased when it is performed aided by computer programs that are capable of displaying sequence traces. Computer-aided manual examination was used in the analysis of overlapping regions of genomic clone sequences to detect microsatellites. Although visual inspection remains an integral part of software testing and tuning, demands for fast and reliable detection in large data sets have necessitated the development of automated, computational methods of microsatellites discovery.

Once batches of nucleotide sequences with the length higher than approximate 200 base pairs have been accumulated in a species, mining microsatellites from them would be a cheapest way. Recently, with the great progress in genomics and bioinformatics, many *in silico* approaches are increasingly being used for the development of microsatellite markers in many species. Structured, classified and easy to use microsatellite data have been compiled in various microsatellite databases that have been developed and made available online by various institutions in recent years (Table 1). Many of these resources are dedicated to mine microsatellites, although they are sometimes by-products of completed or ongoing genome-sequencing projects.

A number of algorithms already existed which either directly or indirectly detect tandem repeats, all suffer from significant limitations. One group of algorithms was based on computing alignment matrices, and their primary limitation was excessive running time. Another group of algorithms found tandem repeats indirectly using methods from the field of data compression, which may require that the approximate pattern size and a range for the number of copies be specified. Benson (1999) overviewed microsatellite-finding softwares and presented a new algorithm for finding tandem repeats which works without the need to specify either the pattern or pattern size. The algorithm presented in this paper is designed to overcome many of the aforementioned limitations: (i) it uses the method of *k-tuple matching* to avoid the need for full scale alignment matrix computations; (ii) it requires no *a priori* knowledge of the pattern, pattern size or number of copies; (iii) there are no restrictions on the size of the repeats that can be detected; (iv) it uses percentage differences between adjacent copies and treats substitutions and indels separately; (v) it determines a consensus pattern for the smallest repetitive unit in the tandem repeat. The program has already been used as a preprocessor in a new alignment algorithm where tandem duplication augments the standard mutation set of insertion, deletion and substitution.

This chapter aims to give readers basic concept and know-how about the development of microsatellite markers by data mining from DNA sequences.

Database	Species	Host	Description	Weblink
Mouse Microsatellite Database of Japan (MMDBJ)	Mouse	National Institute of Genetics, Japan	Collection of 6119 microsatellites. Also includes PCR conditions for all entries of primer sets and keyword searches for the information	<a href="http://www.shigen.nig.ac.jp/mouse/mmdbj/">www.shigen.nig.ac.jp/mouse/mmdbj/</a>

Simple-Sequence Repeat Database (SSRD)	Human	Center for Cellular and Molecular Biology, India	Provides summary and detailed view of SSRs, the flanking genomic regions and their associations with genes and sequence tagged sites (STS) markers	<a href="http://www.ccmb.res.in/ssr">www.ccmb.res.in/ssr</a>
Satelog	Human	Michael Smith Genome Science Center, Canada	Catalogs 1-16 repeat-unit perfect repeats in the human genome	<a href="http://satelog.bcgsc.ca">http://satelog.bcgsc.ca</a>
Microsat2006	Human	King's College, UK	Catalogs human microsatellite repeats	<a href="http://www.microsatellites.org/db_search.php">www.microsatellites.org/db_search.php</a>
Molecular Mycology SSR Database	Nine fungal genomes	Westmead Hospital, UK	Mono- to hexa-nucleotide repeats of fungal genomes with complete or draft sequences available	<a href="http://www.mmrl.med.usyd.edu.au/ssr.html">www.mmrl.med.usyd.edu.au/ssr.html</a>
TRBase	Human	University of Exeter, UK	Perfect and imperfect repeats of 1-2000 bp unit lengths from human-sequence data and annotation files for 11 chromosomes	<a href="http://trbase.ex.ac.uk">http://trbase.ex.ac.uk</a>
InSatdb	Five fully sequenced insect genomes	Center for DNA Fingerprinting and Diagnostics, India	Microsatellite information according to size, genomic location, nature and sequence composition (repeat motif and GC%) as well as microsatellite cluster	<a href="http://210.212.212.8/PHP/INSATDB/home.php">http://210.212.212.8/PHP/INSATDB/home.php</a>
TRDB	Data imported from genome.ucsc.edu	Center for Advanced Genome Technology (CAGT), Boston University, USA	Microsatellite collection along with information on their primers, marker potential, etc., in addition to the facility to screen user's sequence resources, while enabling a user to store and organize their data in allocated 100 Mb of storage space	<a href="http://cagt.bu.edu/page/TRDB_about">http://cagt.bu.edu/page/TRDB_about</a>

Table 1. Some online microsatellite resources (from Prakash 2007)

## 2. Mining microsatellites from nucleotide sequences

### 2.1 Sources of the data

Sequences from both genomic DNA and cDNA can be used for microsatellite mining. Though some researchers produced DNA sequences and kept them in their own laboratories, publicly accessible nucleotide databases are the major source in many studies today including microsatellite mining. The GenBank sequence database is an annotated collection of all publicly available nucleotide sequences and their protein translations. This database is produced at National Center for Biotechnology Information (NCBI) as part of an international collaboration with the European Molecular Biology Laboratory (EMBL) Data Library from the European Bioinformatics Institute (EBI) and the DNA Data Bank of Japan (DDBJ). GenBank and its collaborators receive sequences produced in laboratories throughout the world from more than 100,000 distinct organisms. GenBank continues to grow at an exponential rate, doubling every 10 months. Release 134, produced in February 2003, contained over 29.3 billion nucleotide bases in more than 23.0 million sequences. GenBank is built by direct submissions from individual laboratories, as well as from bulk submissions from large-scale sequencing centers. GenBank nucleotide records are located in separate databases that must be searched independently. These include dbEST and dbGSS, plus multiple databases for the CoreNucleotide division, including nr, htgs, wgs and env\_nt. **ESTs** [<http://www.ncbi.nlm.nih.gov/dbEST/>] are generally short (<1 kb), single-pass cDNA sequences from a particular tissue and/or developmental stage. However, they can also be longer sequences that are obtained by differential display or Rapid Amplification of cDNA Ends (RACE) experiments. ESTs are particularly attractive for marker development since they represent coding regions of the genome and putative function can often be deduced by homology searches although little is known about many of the ESTs. While ESTs provide means for the identification of genes, microsatellites provide high level of polymorphism. Microsatellites identified in ESTs are typically referred to as EST-SSRs or genic SSRs, contrasting to type II SSRs which come from random sequences of the genome. The identification of ESTs has preceded rapidly, with approximately 39 million ESTs sequences now available in public databases (e.g. GenBank 4/2008, all species). As a by-product of EST or BAC sequencing projects in many organisms, microsatellite-mining from SSR-containing ESTs is inexpensive and time-saving, and has proved to be an effective approach to develop microsatellites for genetic map and population genetics studies in animals and plants (e.g. Yue et al., 2004; Wang et al., 2005; Caire et al., 2005).

**STS** [<http://www.ncbi.nlm.nih.gov/dbSTS/>]s are short genomic landmark sequences (1). They are operationally unique in that they are specifically amplified from the genome by PCR amplification. In addition, they define a specific location on the genome and are, therefore, useful for mapping.

**GSS** [<http://www.ncbi.nlm.nih.gov/dbGSS/>]s are also short sequences but are derived from genomic DNA, about which little is known. They include, but are not limited to, single-pass GSSs, BAC ends, exon-trapped genomic sequences, and AluPCR sequences.

EST, STS, and GSS sequences reside in their respective divisions within GenBank, rather than in the taxonomic division of the organism. The sequences are maintained within GenBank in the dbEST, dbSTS, and dbGSS databases.

ESTs are particularly attractive for marker development represent coding regions of the genome and putative function can often be deduced by homology searches. While ESTs provide means for the identification of genes, microsatellites provide high levels of polymorphism.

## 2.2 Finding and characterizing repeat motifs

Traditionally, SSR isolation has relied on the screening of genomic libraries using repetitive probes and sequencing of positive clones in order to develop locus-specific primers. These processes are necessary for many organisms but normally time-consuming and labor-intensive. Mining SSR from public databases has been streamlined with technological advance and protocol optimization to make the process cheaper, more efficient and more successful, and has proved to be an effective approach to develop microsatellites for genetic map and population genetics studies in animals (Serapion et al., 2004; Yue et al., 2004; Wang et al., 2005; Chen et al., 2005; Pérez et al., 2005; Maneeruttanarungroj et al., 2006) and plants (Cordeiro et al., 2001; Kantety et al., 2002; Chen et al., 2006).

Here, we demonstrate how to mine SSRs from common carp EST data step by step.

### 1. Download EST sequences from public databases

The target ESTs from the NCBI dbEST database were downloaded into VectorNTI software (InforMax Inc.). First, "common carp EST" was used as a keyword to search nucleotide sequences at the NCBI databases (<http://www.ncbi.nlm.nih.gov>). EST sequences of common carp were downloaded from GenBank, DDBJ and EMBL databases between January 1, 2002 and October 18, 2005. All matched sequences were downloaded by changing the "display" window to FASTA, and the "send to" window to FILE. A file containing 10,088 sequences was saved as a text file.

### 2. Tools for microsatellite mining

In general, microsatellite-finding tools can be classified broadly into three subcategories based on their architecture: first, such as MISA and TROLL etc; second, Tandem-Repeats Finder (TRF) etc; third, ATR and ETR, etc (Table 2). (Prakash et al., 2007)

Name, acronym and weblink of the tool	Salient features	Limitations
Repeatmasker <a href="http://www.repeatmasker.org">www.repeatmasker.org</a>	Available online and stand-alone; mines perfect, imperfect and compound repeats; accepts data in multiple formats; presents statistical analysis; returns flanking sequences; MaskerAid, a performance enhancement is available	Runs only on Unix/Linux systems; not specific for microsatellites
Sputnik ( <a href="http://espressosoftware.com/pages/sputnik.jsp">http://espressosoftware.com/pages/sputnik.jsp</a> and <a href="http://cbl.labri.fr/outils/Pise/sputnik.html">http://cbl.labri.fr/outils/Pise/sputnik.html</a> )	C-language program available online and stand-alone; mines perfect, imperfect and compound repeats; accepts data in multiple formats; improved versions include Modified Sputnik-I and Modified Sputnik-II	Automated statistical analysis files not generated; runs only on Unix/Linux systems; hexanucleotide repeats are not screened
Tandem Repeats Finder (TRF) ( <a href="http://tandem.bu.edu/trf/trf.html">http://tandem.bu.edu/trf/trf.html</a> )	Both online and stand-alone versions are GUI; mines perfect, imperfect and compound repeats; platform	Accepts input as fasta files only; automated statistical analysis file not generated (TRAP;

	independent	www.coccidia.icb.usp.br/trap/ [54] can be used); process limited-size files only; output files are numerous and difficult to manage
Repeatfinder (www.cbcb.umd.edu/software/RepeatFinder/)	Available online and stand-alone; mines perfect, imperfect and compound repeats; accepts multiple formats as input	Runs on Unix/Linux systems; not specific for microsatellites
eTandem and eQuicktandem (http://bioweb.pasteur.fr/seqanal/interfaces/etandem.html)	Perl script available online and stand-alone; parts of EMBOSS suite; mines perfect, imperfect and compound repeats; accepts input in multiple formats; generates statistics	Runs only on SGI Irix, Linux, Sun solaris and Tru64 Unix
REPuter (http://bibiserv.techfak.uni-bielefeld.de/reputer/)	Available online and stand-alone; stand-alone version can handle large genomic sequences; output cataloged in a format similar to BLAST; statistical and graphical analysis provided; excellent connectivity to BLAST, FASTA.	Limited capacity of online version; accepts data in fasta/plain format only; runs only on Unix; not specific for microsatellites
Simple-Sequence Repeat Identification Tool (SSRIT) and Clemson University Genomics Institute Simple-Sequence Repeat Tool (CUGIssr) (www.gramene.org/db/searches/ssrtool)	Perl scripts available online and stand-alone; platform independent (CUGIssr is a modified version of SSRIT)	Finds only perfect repeats; accepts only fasta-formatted files; automated statistical analysis not generated
Tandem Repeats Occurrence Locator (TROLL) (http://wsmartins.net/cgiocal/webtroll/troll.cgi) and WebTROLL (http://wsmartins.net/webtroll/troll.html)	C++ program available online and stand-alone (TROLL downloadable, WebTROLL web interface); identifies perfect, imperfect and compound repeats; also designs primers	Accepts fasta-formatted files only as input; executes only on Linux systems; statistical analysis not provided
Microsatellite Analysis Server (MICAS)	An exclusively web-based utility	Scans only one file at a time;

( <a href="http://210.212.212.7/MIC/index.html">http://210.212.212.7/MIC/index.html</a> )		compound and imperfect repeats are not identified; statistical analysis is not performed
MISA ( <a href="http://pgrc.ipkgatersleben.de/misa/">http://pgrc.ipkgatersleben.de/misa/</a> )	Perl script executing only offline; large sequences are handled easily; statistical analysis is generated; platform independent; can design primers using Primer3 by running supplementary scripts	Inappropriate clustering of microsatellite motifs in statistical analysis file; only fasta-formatted files are taken as input; identifies only perfect repeats and compound repeats
mreps ( <a href="http://bioinfo.lifl.fr/mreps/mreps.php">http://bioinfo.lifl.fr/mreps/mreps.php</a> and <a href="http://bioweb.pasteur.fr/seqanal/interfaces/mreps.html">http://bioweb.pasteur.fr/seqanal/interfaces/mreps.html</a> )	Available online and stand-alone; identifies compound and imperfect repeats; accepts data in multiple formats; platform independent; can design primers	Statistical analysis is not performed
Search for Tandem Repeats in Genomes (STRING) ( <a href="http://www.caspur.it/_castri/STRING/">http://www.caspur.it/_castri/STRING/</a> )	C-language program available online and stand-alone; finds perfect, imperfect and compound repeats; runs well with large genomic sequences; platform independent	Only fasta files taken as input; no automated statistical analysis
Search for Tandem Approximate Repeats (STAR) ( <a href="http://atgc.lirmm.fr/star">http://atgc.lirmm.fr/star</a> )	Available online and stand-alone; searches for 'approximate' tandem repeats of a given motif; platform independent	Does not generate statistical analysis
MicrosatDesign ( <a href="http://daphnia.cgb.indiana.edu/wfleabase/software">http://daphnia.cgb.indiana.edu/wfleabase/software</a> )	Perl scripts executing as a stand-alone tool; builds database and designs primers from the nascent DNA-sequencer outputs; DNA-sequence trace files are taken as an input; combination of phredPhrap, Primer 3 and GCG software/eTandem software; identifies compound repeats and imperfect repeats as well	Specific in its use; does not generate statistical analysis
Poly ( <a href="http://bioinformatics.org/poly/">http://bioinformatics.org/poly/</a> )	Downloadable Python script; statistical analysis is provided; platform independent	Slow

<p>Exact Tandem Repeats Analyzer (E-TRA) and Tandem Repeats Analyzer (TRA) (<a href="ftp.akdeniz.edu.tr/Araclar/">ftp.akdeniz.edu.tr/Araclar/</a>)</p>	<p>C++ program available online and stand-alone; search microsatellites in ESTs combining with key-word match searches; multiple sequences and multiple files can be handled simultaneously; provide flanking sequences and capable of designing primers; fast; GUI; find perfect, imperfect and compound repeats; accept input in multiple formats; provides statistical analysis</p>	<p>Redundancy in output</p>
<p>msatminer (<a href="http://www.genomics.ceh.ac.uk/msatminer/">www.genomics.ceh.ac.uk/msatminer/</a>)</p>	<p>Perl scripts executing online and stand-alone; finds compound repeats and imperfect repeats also; accepts input in multiple formats; statistical analysis can be obtained on executing additional scripts; separate scripts for designing primers</p>	<p>Runs on Unix and Mac OS environment; stand-alone version complicated owing to requirements to execute as many as four scripts for complete analysis</p>
<p>msatcommander (<a href="http://code.google.com/p/msatcommander/">http://code.google.com/p/msatcommander/</a>)</p>	<p>Python script available for download; GUI; capable of searching perfect, imperfect and compound repeats with flexibility; output in CSV format; platform independent; primer designing utility available</p>	<p>No online interface; only fasta formatted files accepted as input; statistical analysis is not generated automatically</p>
<p>SciRoko (<a href="http://www.kofler.or.at/bioinformatics/SciRoKo/index.html">www.kofler.or.at/bioinformatics/SciRoKo/index.html</a>)</p>	<p>C-language program available for stand-alone execution; identifies perfect, imperfect and compound repeats; highly flexible; extremely fast; GUI; provides statistical analysis; platform independent</p>	<p>Depends on .NET framework</p>
<p>Imperfect Microsatellite Extraction (IMEx) (<a href="http://203.197.254.154/IMEX/">http://203.197.254.154/IMEX/</a>)</p>	<p>C-language program executing stand-alone; finds perfect and imperfect repeats; efficient, fast and user-</p>	<p>Executes on Linux</p>

	friendly; returns the coding/ noncoding information of microsatellites; highly flexible; can design primers as well; statistics are generated	
--	---	--

Table 2. Characteristics of some important microsatellite search tools

In our study in common carp (*Cyprinus carpio*), we use software “Tandem Repeat Finder” (Benson, 1999). All the ESTs were screened for potential microsatellites by using the TRF with the following parameters: match: 2; mismatch 7; indel: 7; PM: mini-score; 30; and max period size 500. Strings of oligo sequences were used to search for microsatellites: 6 repeats for dinucleotides; 4 repeats for trinucleotides, and 3 repeats for tetranucleotides and pentanucleotides as described by Stalling et al (1991).

### 3. Frequency and distribution of microsatellites

A total of 10,088 ESTs of common carp with an average length of 531 bp were downloaded from public databases and subject to bioinformatic analyses. The results showed that 555 (about 5.5%) of these ESTs contained SSRs inside, which is lower than values reported in some aquaculture animals e.g. black tiger shrimp (*Penaeus monodon*) (13.7%, Maneeruttanarungroj et al., 2006), Japanese pufferfish (*Fugu rubripes*) (11.5%, Edwards et al., 1998) and channel catfish (11.2%, Serapion et al., 2004), but higher than those in some other species e.g. Chinese shrimp (*Fenneropenaeus chinensis*) (2.2%, Wang et al., 2005), bay scallop (*Argopecten irradians*) (3.9%, Zhan et al., 2005), and red sea bream (*Chrysophrys major*) (4%, Chen et al., 2005). The abundance of EST-derived microsatellites seems to be highly species-specific in aquacultured animals studied.

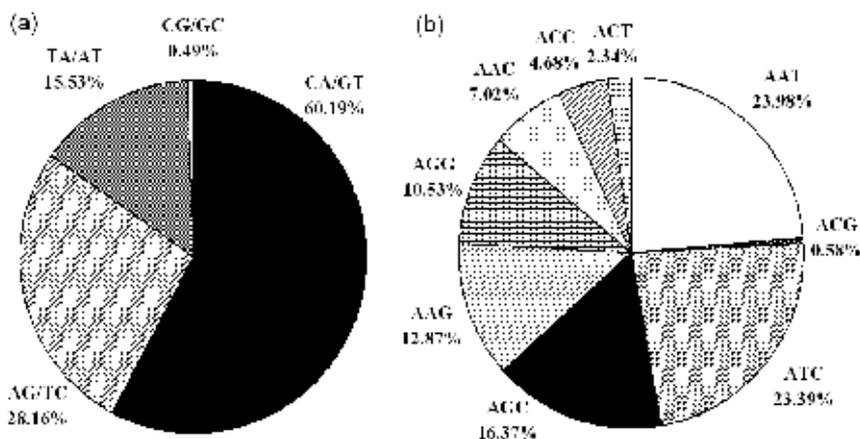


Fig. 1. Distribution of the repeat types of dinucleotides (a) and trinucleotides (b) in common carp EST-SSRs identified by mining public expressed sequence tags databases.

Most of these common carp EST-SSRs were composed of dinucleotide and trinucleotide repeats. Specifically, the abundance of di-, tri-, tetra-, and penta-nucleotide motifs among these ESTs is 37.2%, 30.8%, 20.4%, and 11.7%, respectively. For dinucleotides, AC/TG is the most abundant (Figure 1a), which is consistent with previous findings for both Type I and Type II microsatellites in fish (Edwards et al., 1998; David et al., 2001; Serapion et al., 2004),

various plant species (Gupta & Varshney, 2000), and vertebrates as a whole (Neff & Gross, 2001). The proportion of the trinucleotide repeats was also not evenly distributed, with the two most frequent types (AAT and ATC) accounting for 24.0% and 23.4% of the total motifs, respectively (Figure 1b).

Dinucleotides are the dominant type of microsatellite repeats in most aquaculture species characterized so far, although trinucleotide repeats are most abundant in plants (Cho et al. 2000; Chen et al., 2006; Kantety et al., 2002) (Fig.2).

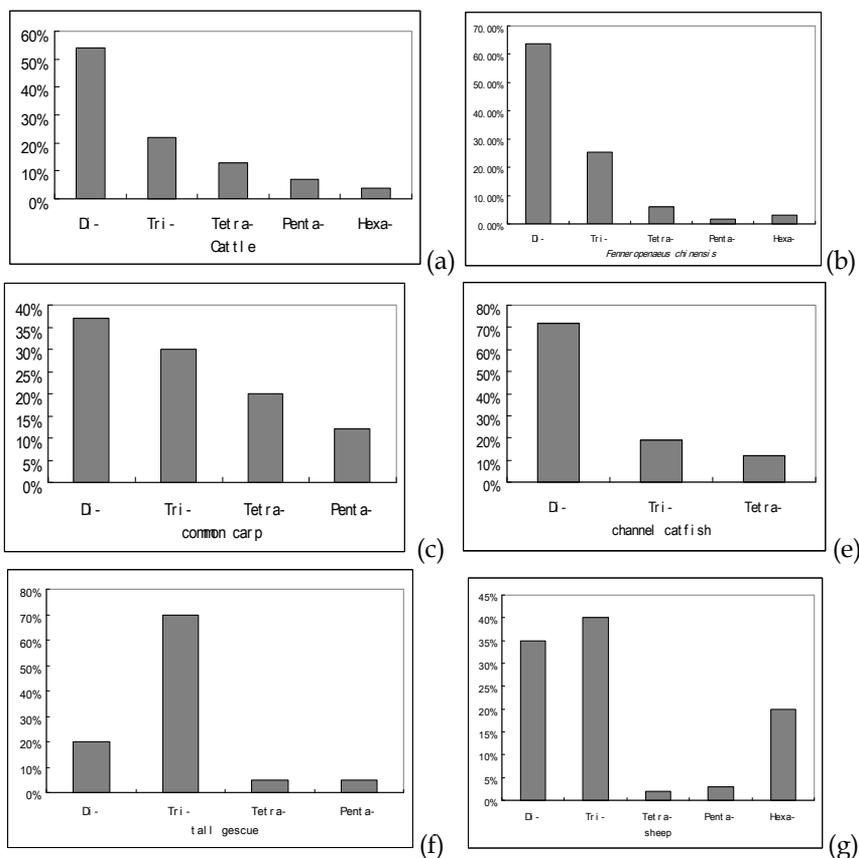


Fig. 2. Distribution of microsatellites in EST sequences from various species.

### 2.3 Other related bioinformatic work

#### 1. Clustering analysis

EST sequences were analyzed by cluster analysis using the ContigExpress module in VectorNTI package (available at <http://download.invitrogen.com>) and linear assembly algorithm was applied. The criteria for clustering were set at a minimum overlap of 30 bases (default is 20 bases). Each cluster was visually inspected to ensure the fidelity of alignment to avoid pseudo-clusters caused by repetitive elements or long strings of microsatellite repeats. In our study, after clustering and assembly, 465 unique microsatellite-containing ESTs were identified, including 400 singletons and 65 contigs (Wang et al., 2007).

## 2. Identification of the known genes

The unique ESTs were then subjected to BLASTx search against the GenBank (protein database) for putative identification of gene function. When accumulated probability of sequence similarity was less than  $1 \times 10^4$ , the tentative identities were established. The BLASTx results revealed that about 165 of these ESTs showed similarity to genes or proteins of known function (Wang et al., 2007).

## 3. Primer design for microsatellites

In our study, 60 of the 465 unique ESTs or genes were randomly chosen for pilot tests for primer design, locus amplification and polymorphism. Software 'Primer 3' (<http://www.genome.wi.mit.edu/cgi-bin/primer/>) was used to design primers for the amplification of repeat regions of interest across the flanking regions. During the primer design, the range of annealing temperature was set up to be between 45 and 55°C, and that of expected size of PCR products 150-250 bp. A single pair of "best" primers was designed and synthesized for each unique EST or gene that contains SSR, and no repeated designs and syntheses of primers were carried out. Here we introduce several tools for primer design (Table 3).

Name of the tools	Features	Limitations
Primer 3	Work on line.	web; C-language
Primer 5	Designing primers for long PCR of sequences up to 50 kb is possible.	Windows
Oligo 6	The graphic features allow screens to be displayed in either a bar or a dot graph.	Windows; Macintosh
DNASTar	Sequence assembly and SNP discovery; gene finding; utility for importing unusual file types. Primer design function included.	Windows
FASTPCR	Automatically SSR loci detection; direct PCR primers design	Windows

Table 3. Characterize important software for microsatellites design

## 3. Laboratory verification of predicted microsatellites

### 3.1 PCR amplification and polymorphism test for microsatellites

In our study, PCR amplifications of microsatellites were carried out on a thermocycler (PTC-100, MJ Research) by using the following program: 94°C for 5 min, followed by 34 cycles of 94°C for 35s, appropriate annealing temperature for 35s, and 72°C for 50s, and a final extension of 72°C for 10 min. The PCR reactions were performed in a 25 µl-reaction mixture, which contained 2.5 µl 10×reaction buffer, 2 µl  $Mg^{2+}$  ( $1.5 \text{ mmol} \cdot \text{L}^{-1}$ ), 1 µl dNTP ( $10 \text{ mmol} \cdot \text{L}^{-1}$ ), 0.5U *Taq* polymerase ( $2 \text{ U} / \mu\text{l}$ ), 2 µl template DNA, 0.25 µl each of the primer ( $5 \mu\text{mol} \cdot \text{L}^{-1}$ ), and 17 µl sterile water. PCR products were separated in 6% denaturing polyacrylamide gel and visualized by silver staining. Allele sizes were determined by comparison with pBR322

DNA/*Msp* I markers (Sino-American, Luoyang, China) combined with image analysis as described previously (Tong et al., 2005).

Out of the 60 common carp EST-SSRs for which primers were designed, 54 primers worked (25 polymorphic, 11 monomorphic, 18 with multiple bands) and 6 failed in the common carp. Some of polymorphic EST-SSRs are shown in Fig 3.

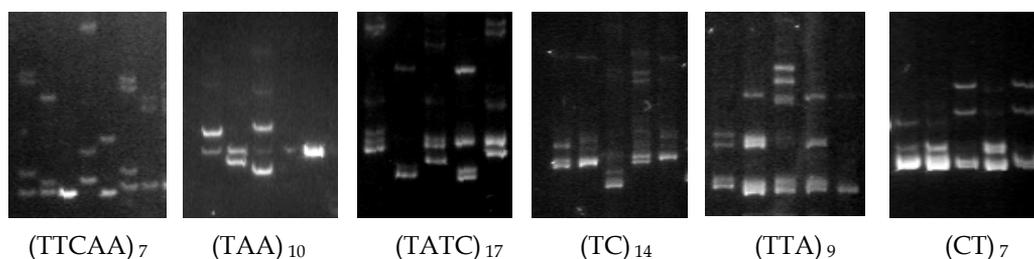


Fig 3. Polymorphism test for parts of EST-SSRs in common carp (Wang et al. 2007).

Twenty-five of the 60 EST-SSRs were found to be polymorphic in a common carp population. The observed heterozygosity of these polymorphic loci ranged from 0.13 to 1.00, and expected heterozygosity ranged from 0.12 to 0.91. The number of alleles of the polymorphic EST-SSRs in common carp ranged from 3 to 17 (mean 7).

Of the 60 common carp EST-SSRs, 10 (17%) of them showed polymorphism in a pilot panel in crucian carp (*Carassius auratus*). In silver carp (*Hypophthalmichthys molitrix*), only 3 (5%) of these loci were found to be polymorphic. In general, these loci are less polymorphic in crucian carp and silver carp than in their source species (common carp).

### 3.2 Hardy-Weinberg Equilibrium (HWE)

In our study, when the frequencies and distributions of the alleles and genotypes were compared under the HWE expectation for an ideal population (random mating, no mutation, no drift, and no migration), 6 of the 25 loci showed significant departure after Bonferroni correction ( $P < 0.002$ ), and the remaining 19 EST-SSRs were in HWE.

A heterozygote excess (also known as homozygote deficit) occurs when the data set contains fewer homozygotes than expected under HWE, and a heterozygote deficit (also known as homozygote excess) occurs when there are more homozygotes than expected under HWE. Currently, tests used to determine statistically significant deviation from HWE have low power when allelic diversity is high and sample sizes are moderate (Guo & Thompson 1992). However, failure to meet HWE is not typically grounds for discarding a locus. Heterozygote deficit, the more common direction of HWE deviation, can be due to biological realities of violating the criteria of an ideal population, such as strong inbreeding or selection for or against a certain allele. Alternatively, when two genetically distinct groups are inadvertently lumped into a single sampling unit, either because they co-occur but rarely interbreed (unbeknownst to the sampler), or because the spatial scale chosen for sampling a site is larger than the true scale of a population, there will be more homozygotes than expected under HWE. This phenomenon is called a Wahlund effect and may be a common cause of heterozygote deficit in population genetic studies. Both of these causes of heterozygote deficit should affect all loci, instead of just one or a few.

### 3.3 Null alleles

Null alleles are those that fail to amplify in a PCR, either because the PCR conditions are not ideal or the primer-binding region contains mutations that inhibit binding. In our study, primers of the six loci failed to amplify in the common carp, and primers of some other loci could not amplify specific products. This could be due to one or both primers being designed across the junction of the spliced ends of exons in the EST sequence, which in genomic DNA is interrupted by an intron (Cordeiro et al., 2001), or due to the inaccuracy of some EST sequences.

As a result of null alleles, some heterozygotes are genotyped as homozygotes and a few individuals may fail to amplify any alleles. Often the mutations that cause null alleles will only occur in one or a few populations, so a heterozygote deficit might not be apparent across all populations. A simple way to identify a null allele problem is to determine if any individuals repeatedly fail to amplify any alleles at just one locus while all other loci amplify normally (suggesting the problem is not simply poor quality DNA). If re-extraction and amplification still fail to produce any alleles at that locus, it is likely that the individual is homozygous for a null allele. In addition, a statistical approach to identifying null alleles can match the pattern of homozygote excess to the expected signatures of several different causes of homozygote excess and estimate the frequency of null alleles for each locus. The software MICROCHECKER (Van Oosterhout et al. 2004) is designed for this aim. A more technical way to detect null alleles is to examine patterns of inheritance in a pedigree (e.g. Paetkau & Strobeck 1995). Redesigning primers to bind to a different region of the flanking sequence, or adjusting PCR conditions can often ameliorate null allele problems. Many researchers are quick to use highly stringent PCR conditions without considering the downside that it inflates the chances for null alleles. A low incidence of null alleles is usually only a minor source of error for most types of analyses, but for certain analyses e.g. parentage analysis, even rare null alleles can confound results and any loci with strong evidence of null alleles should be excluded.

### 3.4 Mendelian inheritance

Mendelian inheritance of alleles is a requirement for almost all population genetic analyses for diploid vertebrate species (Jarne & Lagoda 1996). Because relatively few studies report tests for Mendelian inheritance, it is still unclear how common non-Mendelian inheritance is across taxa. Potential causes of true non-Mendelian behaviour are sex linkage, physical association with genes under strong selection, centres of recombination, transposable elements, or processes during meiosis such as non-disjunction or meiotic drive (segregation distortion). These processes can have severe effects, such as only one parental allele being passed on to all offspring. Performing defined crosses and genotyping a large number of offspring can be quite challenging or impractical in some species, and straightforward in others, such as those that brood their young. Microsatellite loci in any polyploidy species have a high likelihood of occurring multiple times throughout the genome and this will confound analysis, so in particular inheritance should always be examined for polyploidy. Even in diploid or haploid species, duplication of loci can be common and potentially problematic. Any case of a locus displaying more than two alleles per individual (that is not traceable to cross

contamination of samples) should be discarded from most analyses. It is important to note that automated sequencers are set by default to call only two alleles per locus, and will return apparently valid allele calls regardless of the actual number of amplification products produced; for this reason, automated sequencer allele calling should always be double checked by an experienced operator.

### 3.5 Gametic disequilibrium

When two loci are very close together on a chromosome, they may not assort independently and will be transmitted to offspring as a pair. Even if loci are not linked physically on a chromosome, they can be functionally related or under selection to be transmitted as a pair (hence the more accurate term gametic disequilibrium is starting to replace the term linkage disequilibrium). While functional linkage would be unusual for microsatellite loci, microsatellites can be clustered in the genome and gametic disequilibrium should always be tested. Gametic disequilibrium creates pseudo-replication for analyses in which loci are assumed to be independent samples of the genome. Like tests of HWE, gametic disequilibrium testing has low power for highly polymorphic loci, so examining confidence intervals on estimates is recommended. Several user-friendly software programs (most of them are accessible online), such as ARLEQUIN, FSTAT, GENEPOP, GENETIX, and MICROSATELLITE ANALYZER, include tests for gametic disequilibrium by searching for correlations between alleles at different loci. One type of linkage that this test will not catch is sex linkage; however, sex linkage will produce an apparent heterozygote deficit that resembles a null allele problem. Lastly, there are many ecological questions that can benefit from the study of linked loci (Gupta et al. 2005). For instance, inter population variation in linkage can correlate with the history of bottlenecks (Tishkoff et al. 1996).

## 4. Prospects

The option of mining microsatellites from DNA-sequence databases has clearly advanced our understanding of evolutionary processes, leading to the formation of repeats in the genome and their selective advantage for the organism. Information on microsatellite distribution in the genomes is a prerequisite for an in-depth understanding of processes determining the formation of microsatellite regions in genomes. This can be obtained either by *de novo* mining of repeats in genomic sequences or by accessing a database cataloging microsatellite repeats along with their genomic positions.

Despite many advantages, microsatellite markers also have several challenges and pitfalls that at best complicate the data analysis, and at worst greatly limit their utility and confound their analysis. For example, there are some taxa for which new marker isolation is still fraught with considerable failure rate, such as some marine invertebrates (Cruz et al., 2005), lepidopterans (Megleck et al., 2004) and birds (Primmer et al., 1997). If mutations occur in the primer region, some individuals will have only one allele amplified, or will fail to amplify at all (Paetkau & Strobeck 1995). Several taxa seem more often beset by amplification problems than others, notably, bivalves, corals and some other invertebrate taxa (Hedgecock et al., 2004). On the other hand, because the cDNAs from which ESTs are derived lack introns, one possible concern with EST-SSRs is that unrecognized intron splice

sites could disrupt priming sites, resulting in failed amplification. Alternatively, large introns could fall between the primers, resulting in a product that is either too large or, in extreme cases, failed amplification. In some cases, it may be possible to redesign the primers to exclude troublesome introns.

A large amount of organisms on the earth are directly or indirectly important to human life. However, only a small fraction of them are under comprehensive studies using modern science and technology. Due to the limitation of investment and funding, only a very low percentage of organisms have enough DNA or protein sequences, although they may be economically or ecologically important. Sequence data are expected to accumulate in more diverse species.

An optimistic trend in recent years is that with the advance in sequencing technique (e.g. 454 sequencing by Roche) and the increase of invest by government and private companies, full genomic sequences, EST or BAC sequences, have been increasing rapidly, especially in some domestic animals and plants as well as some model organisms.

The recent trend is to cross-amplify molecular markers across a set of closely related genomes. Microsatellites associated with quantitative trait loci (QTLs) and agronomically important genes remain a good candidate for the development of specific markers. The low cost of their generation and ease in documentation are two of the important relative advantages of these sequences over equally promising single nucleotide polymorphisms (SNPs). Microsatellites can thus firmly be expected to have an important role in genomics research in the future and mining microsatellites from DNA databases is likely to take center stage to come.

## 5. Conclusion

With the increasing accumulation of the nucleotide sequence data in both private and public databases, and the invention of more efficient computer-based tools, mining some valuable biological resources, such as microsatellites and SNPs, from the raw DNA data, has become one of the most popular areas of biological studies today, bioinformatics. Development of SSRs by data mining from sequence data is a relatively easy and cost-saving strategy for any organisms with enough DNA data. This is a very good example from data to knowledge, and from knowledge to basic and applied studies for biology, production, conservation and management of many organisms.

## 6. References

- Beckmann J.S. & Weber J.L. (1992) Survey of human and rat microsatellites. *Genomics*, 12, 627-631.
- Benson D.A., Karsch-Mizrachi I., Lipman D.J., Ostell J. & Wheeler D.L. (2008) GenBank. *Nucleic Acids Research*, 36 (Database issue), 25-30.
- Benson G. (1999) Tandem repeats finder: a program to analyze DNA Sequences. *Nucleic Acids Research*, 1999, 27, 573-580.
- Chen, C.X., Zhou, P., Choi, Y.A., Huang, S., Gmitter, F.G., 2006. Mining and characterizing microsatellites from citrus ESTs. *Theor. Appl. Genet.* 112, 1248-1257.

- Chen, S.L., Liu, Y.G., Xu, M.Y., Li, J., 2005. Isolation and characterization of polymorphic microsatellite loci from an EST-library of red sea bream (*Chrysophrys major*) and cross-species amplification. *Mol. Ecol. Notes* 5, 215-217.
- Cho, Y.G., Ishii, T., Temnykh, S., Chen, X., Lipovich, L., McCouch, S.R., Park, W.D., Ayres, N., Cartinhour, S., 2000. Diversity of microsatellites derived from genomic libraries and GenBank sequences in rice (*Oryza sativa* L.). *Theor. Appl. Genet.* 100, 713-722.
- Cordeiro, G.M., Casu, R., McIntyre, C.L., Manners, J.M., Henry, R.J., 2001. Microsatellite markers from sugarcane (*Saccharum* spp.) ESTs cross transferable to erianthus and sorghum. *Plant Sci.* 160, 1115-1123.
- Cruz, F., Perez, M. & Presa, P. (2005). Distribution and abundance of microsatellites in the genome of bivalves. *Gene*, 346, 241-247.
- David, L., Rajasekaran, P., Fang, J., Hillel, J., Lavi, U., 2001. Polymorphism in ornamental and common carp strains (*Cyprinus carpio* L.) as revealed by AFLP analysis and a new set of microsatellite marker. *Mol. Genet. Genomics* 266, 353-362.
- Edwards, Y.J., Elgar, G., Clark, M.S., Bishop, M.J., 1998. The identification and characterization of microsatellites in the compact genome of the Japanese pufferfish, *Fugu rubripes*: perspectives in functional and comparative genomic analyses. *J.Mol.Biol.* 278, 843-854.
- Eisen, J.A. 1999. Mechanistic basis for microsatellite instability. In: *Microsatellites: Evolution and applications* (eds Goldstein, D.B. & Schlotterer, C.). Oxford University Press, Oxford, UK, pp.34-48.
- FitzSimmons N.N., Moritz C. & Moore S.S. (1995) Conservation and dynamics of microsatellite loci over 300 million years of marine turtle evolution. *Molecular Biology and Evolution*, 12, 432-440.
- Guo, S.W., Thompson, E.A., 1992. Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics* 48, 361-372.
- Gupta, P.K., Varshney, R.K., 2000. The development and use of microsatellite markers for genetic analysis and plant breeding with emphasis on bread wheat. *Euphytica* 113, 163-185.
- Hedgecock, D., Li, G., Hubert, S., Bucklin, K. & Ribes, V. (2004). Widespread null alleles and poor cross-species amplification of microsatellite DNA loci cloned from the Pacific oyster, *Crassostrea gigas*. *J. Shellfish Res.*, 23, 379-385.
- Jarne P. & Lagoda P.J.L. (1996) Microsatellites, from molecules to populations and back. *Trends in Ecology and Evolution*, 11, 424-429.
- Kantety, R.V., Rota, M. L., Matthews, D.E., Sorrells, M.E., 2002. Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. *Plant Mol. Biol.* 48, 501-510.
- Knapik E.W., Goodman A., Ekker M., Chevrette M., Delgado J., Neuhauss S., Shimoda N., Driever W., Fishman M.C. & Jacob H.J. (1998) A microsatellite genetic linkage map for zebrafish (*Danio rerio*). *Nature Genetics*, 18, 338-343.
- Litt, M. & Luty, J.A. (1989) A hypervariable microsatellite revealed by *in vitro* amplification of a dinucleotide repeat within the cardiac muscle actin gene. *American Journal of Human Genetics*, 44, 397-401.

- Liu, Z.J., Cordes, J.F., 2004. DNA marker technologies and their applications in aquaculture genetics. *Aquaculture* 238, 1-37.
- Maneeruttanarungroj, C., Pongsomboon, S., Wuthisuthimethavee, S., Klinbunga, S., Wilson, K.J., Swan, J., Li, Y., Whan, V., Chu, K.H., Li, C.P., Tong, J., Glenn, K., Rothschild, M., Jerry, D., Tassanakajon, A., 2006. Development of polymorphic expressed sequence tag-derived microsatellites for the extension of the genetic linkage map of the black tiger shrimp (*Penaeus monodon*). *Anim. Genet.* 37, 363-368.
- Meglecz, E., Petenian, F., Danchin, E., D'Acier, A.C., Rasplus, J.-Y. & Faure, E. (2004). High similarity between flanking regions of different microsatellites detected within each of two species of Lepidoptera: *Parnassius apollo* and *Euphydryas aurinia*. *Mol. Ecol.*, 13, 1693-1700.
- Neff, B.D., Gross, M.R., 2001. Microsatellite evolution in vertebrates: inference from AC dinucleotide repeats. *Evolution* 55, 1717-1733.
- Nonneman, D., Waldbieser, G.C., 2005. Isolation and enrichment of abundant microsatellites from a channel catfish (*Ictalurus punctatus*) brain cDNA library. *Anim. Biotechnol.* 16, 103-116.
- Paetkau, D. & Strobeck, C. (1995). The molecular-basis and evolutionary history of a microsatellite null allele in bears. *Mol. Ecol.*, 4, 519-520.
- Pérez, F., Ortiz, J., Zhinaula, M., Gonzabay, C., Calderón, J., Volckaert, F.A.M.J., 2005. Development of EST-SSR markers by data mining in three species of shrimp: *Litopenaeus vannamei*, *Litopenaeus stylirostris*, and *Trachypenaeus birdy*. *Mar. Biotechnol.* 7, 554-569.
- Prakash C., 2007. Mining microsatellites in eukaryotic genomes. *Trends in biotechnology*.
- Primmer, C.R.; Raudsepp, T; Chowdhary, B.P.; Moller, A.P.; Ellegren, H. 1997. Low frequency of microsatellites in the avian genome. *Genome Research.* 7, 471-482.
- Rassmann K., Schlötterer C. & Tautz D. (1991) Isolation of simple sequence loci for use in polymerase chain reaction-based DNA fingerprinting. *Electrophoresis*, 12, 113-118.
- Rexroad, C.E.3rd., Rodriguez, M.F., Coulibaly, I., Gharbi, K., Danzmann, R.G., Dekoning, J., Phillips, R., Palti, Y., 2005. Comparative mapping of expressed sequence tags containing microsatellites in rainbow trout (*Oncorhynchus mykiss*). *BMC Genomics* 6, 54.
- Schlötterer, C., Amos B. & Tautz D. (1991) Conservation of polymorphic simple sequence loci in cetacean species. *Nature*, 354, 63-65.
- Schuler G.D., Boguski M.S., Stewart E.A., Stein L.D., Gyapay G., Rice K., White R.E., Rodriguez-Tom P., Aggarwal A., Bajorek E., Bentolila S., Birren B.B., Butler A., Castle A.B., Chiannikulchai N., Chu A., Clee C., Cowles S., Day P.J.R., Dibling T., East C., Drouot N., Dunham I., Duprat S., Edwards C., Fan J.B., Fang N., Fizames C., Garrett C., Green L., Hadley D., Harris M., Harrison P., Brady S., Hicks A., Holloway E., Hui L., Hussain S., Louis-Dit-Sully C., Ma J., MacGilvery A., Mader C., Maratukulam A., Matise T.C., McKusick K.B., Morissette J., Mungall A., Muselet D., Nusbaum H.C., Page D.C., Peck A., Perkins S., Piercy M., Qin F.,

- Quackenbush J., Ranby S., Reif T., Rozen S., Sanders C., She X., Silva J., Slonim D.K., Soderlund C., Sun W.L., Tabar P., Thangarajah T., Vega-Czarny N., Vollrath D., Voyticky S., Wilmer T., Wu X., Adams M.D., Auffray C., Walter N.A.R., Brandon R., Dehejia A., Goodfellow P.N., Houlgatte R., Hudson J.R., Jr., Ide S.E., Iorio K.R., Lee W.Y., Seki N., Nagase T., Ishikawa K., Nomura N., Phillips C., Polymeropoulos M.H., Sandusky M., Schmitt K., Berry R., Swanson K., Torres R., Venter J.C., Sikela J.M., Beckmann J.S., Weissenbach J., Myers R.M., Cox D.R., James M.R., Bentley D., Deloukas P., Lander E.S. & Hudson T.J. (1996) A Gene Map of the Human Genome. *Science*, 274, 540-546.
- Serapion, J., Kucuktas, H., Feng, J.N., Liu, Z.J., 2004. Bioinformatic mining of type I microsatellites from expressed sequence tags of channel catfish (*Ictalurus punctatus*). *Mar. Biotechnol.* 6, 364-377.
- Stallings, R.L., Ford, A.F., Nelson, D., Torney, D.C., Hildebrand, C.E., Moyzis, R.K., 1991. Evolution and distribution of (GT)<sub>n</sub> repetitive sequences in mammalian genomes. *Genomics* 10, 807-815.
- Tautz D. (1989) Hypervariability of simple sequences as a general source for polymorphic DNA markers. *Nucleic Acids Research*, 17, 6463-6471.
- Tishkoff, S.A., Dietzsch, E., Speed, W., Pakstis, A.J., Kidd, J.R., Cheung, K. et al. (1996). Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science*, 271, 1380-1387.
- Tong, J., Yu, X., Liao, X., 2005. Characterization of a highly conserved microsatellite marker with utility potentials in cyprinid fishes. *J. Appl. Ichthyol.* 21, 232-235.
- Toth, G., Gaspari, Z. & Jurka, J. (2000). Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res.*, 10, 967-981.
- Van Oosterhout C., Hutchinson W.F., Wills D.P.M., Shipley P. (2004). MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. *Molecular Ecology Notes*, 4, 535-538.
- Wang D, Liao XL, Cheng L, Yu XM, Tong J. (2007). Development of novel EST-SSR markers from common carp by data mining from public EST sequences. *Aquaculture* 271: 558-574.
- Wang, H.X., Li, F.H., Xiang, J.H., 2005. Polymorphic EST-SSR markers and their mode of inheritance in *Fenneropenaeus chinensis*. *Aquaculture* 249, 107-114.
- Weber J.L. & May P.E. (1989) Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *American Journal of Human Genetics*, 44, 388-396.
- Wheeler D.L., Barrett T., Benson D.A., Bryant S.H., Canese K., Chetvernin V., Church D.M., DiCuccio M., Edgar R., Federhen S., Feolo M., Geer L.Y., Helmberg W., Kapustin Y., Khovayko O., Landsman D., Lipman D.J., Madden T.L., Maglott D.R., Miller V., Ostell J., Pruitt K.D., Schuler G.D., Shumway M., Sequeira E., Sherry S.T., Sirotkin K., Souvorov A., Starchenko G., Tatusov R.L., Tatusova T.A., Wagner L. & Yaschenko E. (2008) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 36 (Database issue), 13-21.
- Yue, G.H., Ho, M.Y., Orban, L., Komen, J., 2004. Microsatellites within genes and ESTs of common carp and their applicability in silver crucian carp. *Aquaculture* 234, 85-98.

- Zane, L., Bargelloni, L. & Patarnello, T. (2002). Strategies for microsatellite isolation: a review. *Mol. Ecol.*, 11, 1-16.
- Zhan, A.B., Bao, Z.M., Wang, X.L., Hu, J.J., 2005. Microsatellite markers derived from bay scallop *Argopecten irradians* expressed sequence tags. *Fisheries Sci.* 71, 1341-1346.

## INDUSTRIALIST APPLICATIONS



# Quality Improvement using Data Mining in Manufacturing Processes

Shu-guang He<sup>1</sup>, Zhen He<sup>1</sup>, G. Alan Wang<sup>2</sup> and Li Li<sup>3</sup>

<sup>1</sup>*School of Management, Tianjin University, Tianjin,*

<sup>2</sup>*Pamplin College of Business, Virginia Polytechnic Institute and State University,*

<sup>3</sup>*School of Electronics and Information Engineering, Tianjin Professional College, Tianjin,*

<sup>1,3</sup>*P.R.China*

<sup>2</sup>*U.S.A.*

## 1. Introduction

Nowadays, manufacturing enterprises have to stay competitive in order to survive the competition in the global market. Quality, cost and cycle time are considered as decisive factors when a manufacturing enterprise competes against its peers. Among them, quality is viewed as the more critical for getting long-term competitive advantages. The development of information technology and sensor technology has enabled large-scale data collection when monitoring the manufacturing processes. Those data could be potentially useful when learning patterns and knowledge for the purpose of quality improvement in manufacturing processes. However, due to the large amount of data, it can be difficult to discover the knowledge hidden in the data without proper tools.

Data mining provides a set of techniques to study patterns in data “that can be sought automatically, identified, validated, and used for prediction” (Witten and Frank 2005). Typical data mining techniques include clustering, association rule mining, classification, and regression. In recent years data mining began to be applied to quality diagnosis and quality improvement in complicated manufacturing processes, such as semiconductor manufacturing and steel making. It has become an emerging topic in the field of quality engineering. Andrew Kusiak (2001) used a decision tree algorithm to identify the cause of soldering defects on circuit board. The rules derived from the decision tree greatly simplified the process of quality diagnosis. Shao-Chuang Hsu (2007) and Chen-Fu Chien (2006 and 2007) demonstrated the use of data mining on semiconductor yield improvement. Data mining has also been applied to product development process (Bakesh Menon, 2004) and assembly lines (Sébastien Gebus, 2007). Some researchers combined data mining and traditional statistical methods and applied to quality improvement. Examples are the use of MSPC (multivariate statistical control charts) and neural networks in detergent-making company (Seyed Taghi Akhavan Niaki, 2005; Tai-Yue Wang, 2002), the combination of automated decision system and six sigma in the General Electric financial Assurance businesses (Angie Patterson, 2005), the combined used of decision tree and SPC with data from Holmes and Mergen (Ruey-Shiang Guh, 2008), the use of SVR (support vector regression) and control charts (Ben Khediri ISSam, 2008), the use of ANN (artificial neural

network), SA (simulated annealing) and Taguchi experiment design (Hsu-Hwa Chang, 2008). Giovanni C Porzio (2003) has presented a method for visually mining off-line data with combination of ANN and  $T^2$  control chart and to identify the assignable variation automatically.

Although techniques are available to learn numerous patterns and knowledge hidden in mass data, it can be difficult to identify those patterns that can be directly applicable. Kaidi Zhao (2005) developed a visualization tool, named as Opportunity Map, which can help identify useful and actionable knowledge quickly. Mu-Chen Chen (2007) proposed a method that ranks the rules learned by data mining techniques using DEA (Data Envelopment Analysis).

This work focuses on using data mining for quality improvement in manufacturing processes. The chapter is organized as follows. In section 2, we introduce a knowledge-based continuous quality improvement model in comparison with DMAIC for six-sigma. In section 3, we explain parameter optimization, quality diagnosis and service data analysis for quality improvement. In section 4, we propose a system framework for quality improvement in manufacturing processes. Finally, section 5 concludes the chapter and highlights the areas for further work.

## 2. Knowledge-based continuous quality improvement in manufacturing processes

Continuous quality improvement is an important concept in the field of manufacturing quality management. DMAIC (Define-Measure-Analyze-Improve-Control) for six-sigma is the most commonly used model of continuous quality improvement. Fig.1 illustrates the processes of DMAIC.

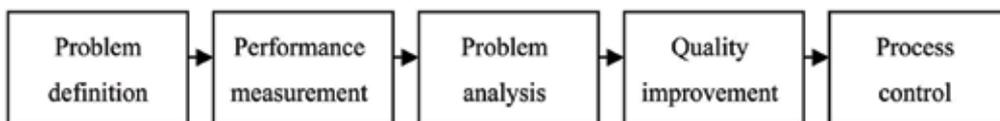


Fig. 1. The processes of DMAIC

We can observe from Fig. 1 that DMAIC is a problem driven approach. The entire process begins with locating a problem. However, in a complicated manufacturing process, such as semiconductor manufacturing and steelmaking, it may not be easy to identify and define a proper problem. Moreover, in high-speed manufacturing processes quality problems must be quickly identified and eliminated. Otherwise it may lead to a large amount of loss in both cost and productivity. Therefore, we propose a knowledge-based quality improvement model (see Fig. 2). Different from DMAIC, this model is a goal-driven process. The central idea of the knowledge-based quality improvement mode is to mine the mass data collected from a manufacturing process using automated data mining techniques. The goal is to improve the quality performance of manufacturing processes by quickly identifying and eliminating quality problems.

In the knowledge-based quality improvement model, the first step is to define the goal. The goal here may be defect elimination, efficiency improvement, or yield improvement. Data mining is used to analyze the quality related data for finding the knowledge between the goal and the factors such as machinery parameters, operators, and material vendors. After

the knowledge has been verified, opportunities of quality improvement can be identified using the knowledge and patterns learned by data mining techniques. The scope of the problem can be broad across different phases of a manufacturing process. In the following sections, we explained how to apply the model to parameter optimization, quality diagnosis and service data analysis.

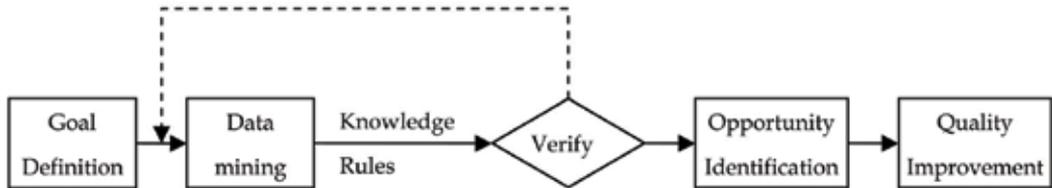


Fig. 2. The knowledge-based quality improvement mode

### 3. Using data mining to improve quality in manufacturing processes

There are three important processes in manufacturing enterprises, including design, manufacturing and service. Data mining techniques can be used in all three processes to improve the quality of manufacturing processes and final products.

#### 3.1 Parameter optimization with data mining

Parameter optimization is critical to quality improvement in manufacturing processes. There are various methods proposed for parameter optimization. DOE (Design Of Experiments) is considered as the most important one (Myers R H, 1985). RSM (Response Surface Methodology) is a technique developed based on DOE and has been most widely used. The main idea of DOE and RSM is to build a function between the inputs (factors) and outputs (responses) of a process. The parameters of the function can be optimized using mathematical methods.

Although DOE is valuable for parameter optimization, there are also drawbacks. DOE is a static method in which the parameters are optimized in a certain situation and the parameters cannot be adjusted during the real manufacturing process that is often dynamic in all kinds of situations. In addition the DOE model only considers a small number of factors with the constriction of cost.

To overcome the problems of DOE, we propose a technique that combines DOE and data mining. DOE is used for parameter optimization for certain static modes. Data mining is employed to analyze actual quality measures with different parameters settings. It aims to learn the patterns between the parameters settings and quality outcome. The pattern can then be used to dynamically adjust the parameters during the manufacturing process so that undesired outcome can be avoided. In the following, we use an injection modeling process in semiconductor assembly as an example to illustrate the process.

In the molding process of LED packaging, there can be many factors affecting the quality of final products. We chose four most important factors with screening experiments. The factors are explained below.

1. Mold temperature (°C). The mold temperature will affect the shape of final products, which is a critical quality characteristic. If the temperature is too high, the number of products with defects will increase.

Standard Order	Run Order	Mold temperature(°C)	Warm-up temperature(°C)	Screw pressure(N)	Screw duration(s)	Defect rate (%)
16	1	0	1	0	-1	2.25
21	2	-1	0	0	-1	2.24
9	3	0	0	-1	-1	2.96
24	4	0	-1	0	1	4.95
8	5	0	1	1	0	1.40
13	6	-1	1	0	0	10.05
4	7	0	0	1	1	4.78
12	8	0	0	-1	1	5.50
20	9	0	0	0	0	1.16
7	10	0	-1	0	-1	2.62
26	11	0	0	0	0	0.69
1	12	0	-1	1	0	7.17
22	13	1	-1	0	0	4.67
2	14	1	0	0	-1	4.17
17	15	-1	0	-1	0	10.62
6	16	-1	-1	0	0	11.20
11	17	0	1	-1	0	7.47
5	18	0	-1	-1	0	1.67
10	19	-1	0	0	1	2.15
19	20	0	0	1	-1	6.85
18	21	1	1	0	0	0.76
3	22	1	0	-1	0	2.35
23	23	1	0	1	0	0.69
15	24	0	1	0	1	2.24
25	25	-1	0	1	0	2.77
14	26	0	0	0	0	3.96

Table 1. The RSM table for parameter optimization

2. Warm-up temperature (°C). In injection molding process, epoxy compound, which is a kind of material, must be softened before being used. The warming up temperature is to ensure the material be softened in a proper temperature. The warm-up temperature is also very important to the process.
3. Screw pressure (Newton). If the screw pressure is too high, the wire attached to the dice and the PCB will be skewed. On the other hand, if the screw pressure is too low, the epoxy compound will be cooled and also lead to quality problems.
4. Screw duration (Second.). Screw duration is the time intervals between the beginning and the end of molding processes. The epoxy compound will be cooled with too long screw duration. Screw duration is also a vital parameter of this process.

Experimental data are collected in a RSM table (Table 1) and analyzed using the Minitab© software. And the optimized manufacturing parameters are found using statistical methods. The mold temperature is 142.3 °C, the warm-up temperature is 65.7 °C, the screw pressure is 22.4 Newton and the screw duration is 40 seconds.

It must be pointed out that these optimized parameters are obtained in a controlled experiment. The parameters may not achieve the optimal results in real manufacturing

conditions. Other uncontrollable factors may also have an effect on the quality of final products. Thus the process should be monitored and optimized continuously using real quality related data. Data mining can be used to serve that purpose by automatically obtaining knowledge and patterns about the manufacturing process.

We collect 1000 records randomly from the molding processes. Each record consists of values of the four factors. The records were classified based on their defect rates. Those records whose defect rates were higher than 3.0% were categorized into a negative class and labeled with *L*. All other records belonged to a positive class and were labeled with *H*. We use the C5.0 decision tree algorithm to analyze the data. Fig. 3 shows the result of the decision tree. We can observe that the right branch of the tree achieves a better performance than the left branch. The path indicated by the circled nodes provides us guidance for parameter optimization.

The decision tree shown in Fig. 3 can also be presented by a set of rules. We have identified five rules. Two of them lead to the classification of the positive class (*H*) where the other three predict the negative class (*L*).

1) Rule 1 for *H* (144; 0.973)

If Mold\_Tem > 135.345 and  
Warm\_Tem > 66.762 and  
Pressure > 21.235  
Then *H*.

2) Rule 2 for *H* (127; 0.961)

If Mold\_Tem ≤ 135.746 and  
Warm\_Tem > 62.106 and  
Warm\_Tem ≤ 66.762 and  
Pressure > 20.733  
Then *H*.

3) Rule 1 for *L* (44; 0.848)

If Mold\_Tem ≤ 135.345 and  
Warm\_Tem > 66.761  
Then *L*.

4) Rule 2 for *L* (390; 0.625)

If Pressure ≤ 21.235  
Then *L*.

5) Rule 3 for *L* (790; 0.598)

If Warm\_Tem ≤ 66.762  
Then *L*.

Although a decision tree does not provide precisely optimized parameters like what the RSM method does, decision tree can analyze a very large amount of quality related data with noise which is still a constriction in DOE. The combination of these two methods can provide more feasible results than using DOE only.

### 3.2 Quality diagnosis with data mining

During a manufacturing process, product quality can be affected by two types of variation: random variations and assignable variations. Random variations are caused by the intrinsic characteristics of a manufacturing process and cannot be eliminated completely. Assignable variations are often produced by a flawed manufacturing setup that involves machinery,

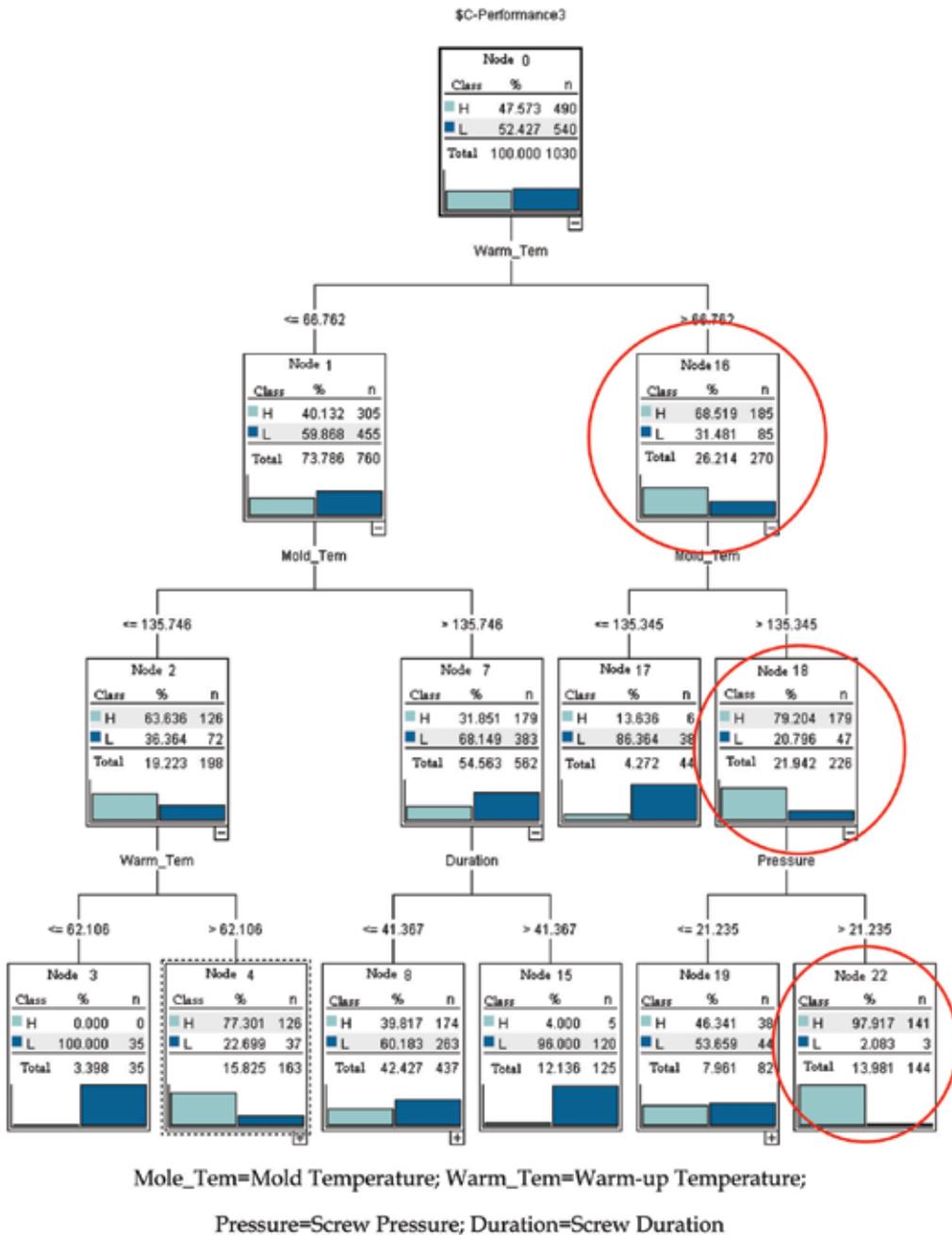


Fig. 3. The decision tree of manufacturing data analysis

operators, materials and environment. Assignable variations are predictable and such can be eliminated as soon as we identify them.

SPC (Statistical Process Control) and MSPC (Multivariate SPC) are the most widely used tools in manufacturing for finding assignable variations. Although they can effectively

detect assignable variations in manufacturing processes, they give no clue to identifying the root causes of the assignable variations. Data mining techniques can again be employed in this case to provide insights for quality diagnosis.

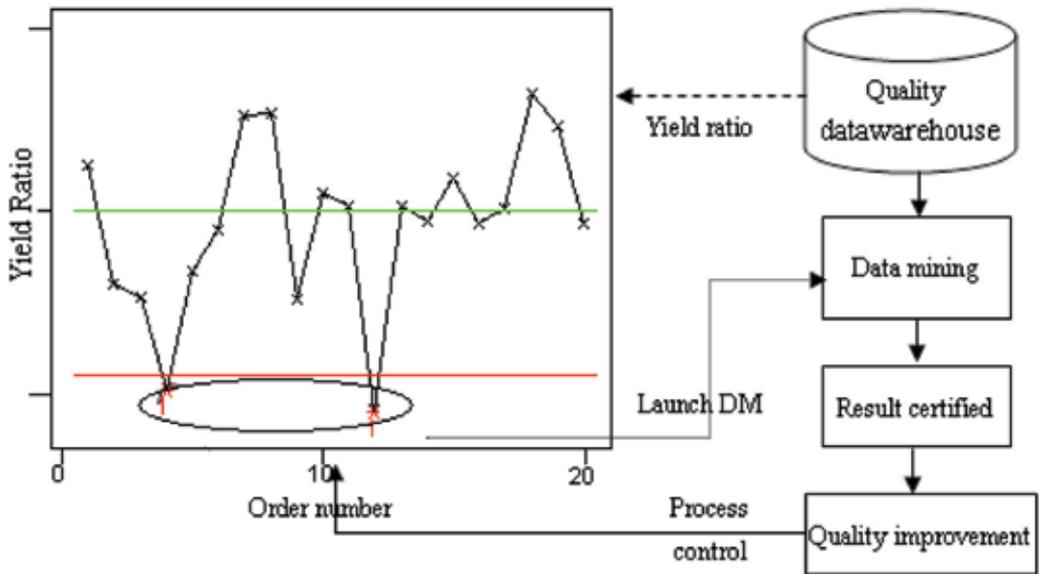


Fig. 4. The combination of SPC and data mining

In the model shown in Fig. 4, the yield ratio of a product is defined as the index of the quality performance of a manufacturing process. The chart on the left is a control chart. When the chart shows alarming signals, i.e., points located beyond the control limit, the data mining process will be engaged. Data related to quality are stored in a data warehouse. Data mining techniques such as a decision tree and association rule mining can be applied to the data to identify the causes of the alarming signals.

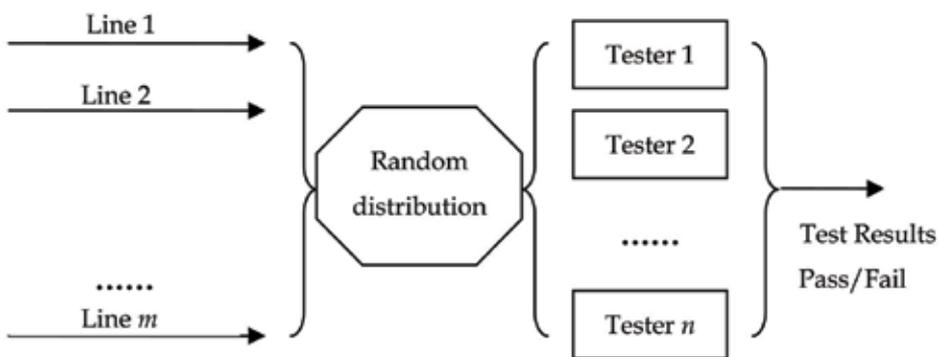


Fig. 5. A mobile phone assembly line model

We use a mobile phone assembly line as an example to explain the process described above. Electronic test is an important process in a mobile phone assembly line. During electronic test, all the products will be tested using an electronic test instrument. Quality related data

can be collected off the test instrument and stored in a database. We can then analyze the data using data mining in order to find ways to improve the quality of the manufacturing process. A mobile phone assembly line model is presented in Fig. 5. There are  $m$  assembly lines while products assembled in these lines are randomly distributed to  $n$  testers. The test results are classified into two classes: **Pass** and **Fail**.

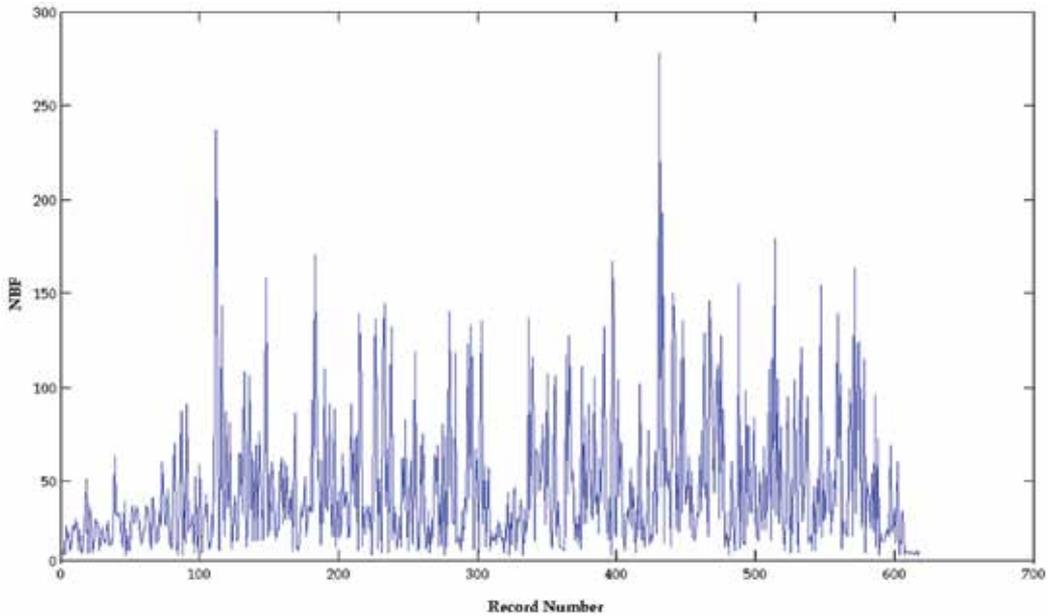


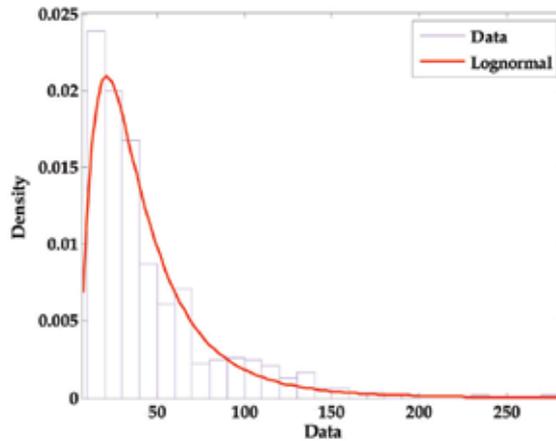
Fig. 6. The distribution of NBF values

In a mobile phone assembly line, there is a very important process named electronic test. In this process, all the products will be tested with an electronic test instrument. And the data are collected and stored in a database. We can analyze the data with data mining to find the opportunity to improve the quality of manufacturing process. The mobile phone assembly line is illustrated in fig. 5. There are  $m$  assembly lines and the products assembled in these lines are randomly distributed to  $n$  testers. The test results are classified into two classes, marked with **Pass** and **Fail**. Furthermore, there are also 27 numerical attributes in the test results, but they are not concerned in this study.

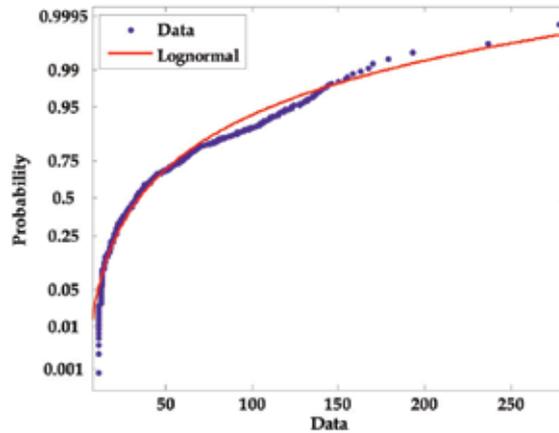
Testers are assumed to work under the same condition with same precision. The number between failures (NBF), i.e., the number of products that passed the test between two successive failure tests, is used to measure the quality performance of the manufacturing process. A larger NBF denotes a more stable manufacturing process.

In the experiment we collected 62,592 records from the electronic test process of the mobile phone assembly line. The NBF values of the records are shown in Fig. 6 that provides a concise description of the data.

We used Matlab software to analyze the distribution of NBF values. Fig. 7 shows that the data can fit to a lognormal distribution. Fig. 7(a) shows the probability density function plot while Fig. 7(b) draws the probability plot of the NBF values and the lognormal distribution curve. The parameter estimation of the lognormal distribution is presented in Table 6.



(a) The PDF plot



(b) The probability plot

Fig. 6. Distribution fitting of the NBF data

Parameter	Estimate	Standard Error
Mean	3.54087	0.0284031
Sigma	0.70609	0.0201084

Table 6. The estimated parameters of the lognormal distribution

The standard error of both the mean and the standard deviation of the NBF data are really small. Therefore we can say that the NBF data can approximately fit a lognormal distribution with mean=3.54087 and sigma=0.70609. With the estimated parameters, we can get the 0.0027 fractile, which is a vital parameter in SPC, of the lognormal distribution is 4.78. This means that if the NBF is smaller than 5, we can make the decision that there are assignable variations in the process and the expected probability of an error decision is less

than 0.27%. Thus we can set the  $CL=5$  as a control line of the SPC chart for the NBF data. If there are points distributed beyond the control limits, the data mining methods will be used for quality diagnosis.

Next an association rule mining tool is used to find the root cause of the assignable variations of the process. Attributes such as tester ID, assembly line number, and test results are supplied into our association rule mining tool with test results being the consequent variable. We used the Apriori algorithm as the association rules analysis method. The minimum antecedent support was set to 0.65% while the minimum rule confidence was 80%. The obtained association rules are presented in Table 7.

Consequent	Antecedent	Instances	Support %	Confidence %	Rule ID
Fail	Line = L15	3,810	6.087	99.948	1
	Line = L20	4,209	6.725	99.81	2
	Line = L20 and Tester_ID = T106	440	0.703	100	3
	Line = L20 and Tester_ID = T21	429	0.685	100	4
	Line = L20 and Tester_ID = T31	431	0.689	100	5
	Line = L20 and Tester_ID = T73	411	0.657	100	6
	Line = L20 and Tester_ID = T97	425	0.679	98.824	7
	Line = L20 and Tester_ID = T35	410	0.655	100	8
	Line = L20 and Tester_ID = T63	423	0.676	100	9
	Line = L33	3,216	5.138	100	10
	Line = L33 and Tester_ID = T5	409	0.653	100	11
	Line = L36	4,297	6.865	100	12
	Line = L36 and Tester_ID = T10	432	0.69	100	13
	Line = L36 and Tester_ID = T14	503	0.804	100	14
	Line = L36 and Tester_ID = T57	455	0.727	100	15
	Line = L36 and Tester_ID = T70	516	0.824	100	16
	Line = L36 and Tester_ID = T75	416	0.665	100	17
	Line = L36 and Tester_ID = T81	438	0.7	100	18
	Line = L36 and Tester_ID = T89	499	0.797	100	19
	Line = L36 and Tester_ID = T104	472	0.754	100	20
	Tester_ID = T104	585	0.935	83.077	21

Table 7. The obtained association rules

Some conclusions drawn from the association rules are listed as below.

1. There were 3,810 mobile phones assembled by line 15, 99.948% of which failed the test (Rule 1).

2. There were 4,209 mobile phones assembled by line 20. 99.81% of which failed the test (Rule 2). Rules 3- extended Rule 2 by considering each different tester.
  3. There were 3,216 mobile phones assembled by line 33, all of which failed the test (Rules 10 and 11).
  4. There were 4,297 mobile phones assembled by line 36, all of which failed the test (Rules 12-20).
  5. There were 585 products tested by tester 104, 83.077% of which failed the test (Rule 21).
- In all the 62,592 products, there were 16,905 products that failed the test. The percentage of the failed tests was 27%. The failed products assembled by lines 15, 20, 33, 36 or tested by tester 104 are summed up to 16,117. That means nearly 95% of the products that failed in the tests were caused by the four assembly lines and/or tester 104. This example shows how data mining techniques can be used to identify the root causes of quality problems in a manufacturing process. That kind of knowledge is valuable for quality diagnosis and quality improvement.

### 3.3 Quality improvement using customer service data

Manufacturing companies have obligations to repair sold products covered by the manufacturer's warranty. Customer service records can be a valuable information source for quality improvement. Those records usually are multi-dimensional with many attributes. It is a challenging issue for companies to extract useful information from the vast amount of customer service records and provide feedback to product design and manufacturing quality improvement.

We propose a research design for customer service data analysis. It includes four major steps as we discuss below.

#### Step 1: Data structure design

Customer service data usually includes customer ID, product ID, service items and duration of the owner ship. Fig. 7 shows an example of the relational database used to store the customer service data. The database design follows the guidelines of developing a relational database (Edgar F Codd, 1970).

This is the fact table of customer service data. We can define different dimensions for data analysis. For example, in a mobile phones customer service domain, dimensions are customer, product, service items and the duration of ownerships'.

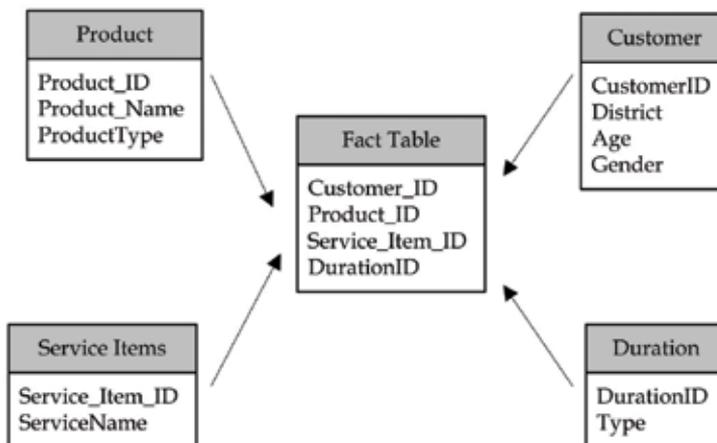


Fig. 7. The database design for the customer service data

**Step 2: MDA (Multi-Dimensional Analysis)**

The stored customer data can be analyzed in different dimensions. And the MDA can give a statistical interpretation of the data. With the drill-down and drill-up operations, we can analyze the data at a detailed level or a summarized level.

**Step 3: Data mining**

With the integrated customer service data, data mining techniques such as decision tree or association rule mining can be used to find the relationship among customers, products, durations and service items. These kinds of relationships will help improve product design.

**Step 4: Rules certification and usage**

The patterns and knowledge obtained by data mining techniques must be verified by domain experts or customer interview before being applied.

We illustrating the methods discussed above using 2000 records extracted from a mobile phones manufacturer's service process. The purpose is to demonstrate how the learned knowledge can improve the quality of products by design.

The data were rearranged to a table with seven columns, including customer ID, Age, District, Gender, Product Type, Service Item and Duration. The customer's age was divided into five intervals of above 55, between 45 and 55, between 35 and 45, between 25 and 35 and below 25. The location of customers was divided into four districts; they are southeast, southwest, northeast and northwest of China. There are four types of products named as series A, series B, series C and series D. The service items were classified into four types including the product appearance, product function, product performance, and others. The duration of ownerships was divided into four intervals: 1-6 months, 7-12 months, 13-18 months and 19-24 months.

To identify the relationship among these factors, we used the CHAID decision tree algorithm (J. A. McCarty, 2007) to analyze the data. Fig. 8 shows the decision tree.

There is an interesting pattern in the decision tree. 48.3% of the products had function defects. If the products were used in the district of southeast of China, the percentage increased to 79.13%. Among them, 69.95% were the product of series C, 21.7% series B, and only 8.3% product of series D. Obviously we can say that if the products of series C are used in southeast of China, they are inclined to have functional defects.

Furthermore, we should analyze the rule deeply. After a technical examination of the repaired products, it was found that the repaired products have almost the same erosion in circuit boards. Compared with the environment conditions, it was found that the reason of the functional defects of the products was the humid atmospheric corrosion. Erosion occurred more easily on the products of series C than other types of products. The defects were eliminated by design improvement. But this kind of problems was not very easy to identify without the deep analysis of the service data.

**4. System framework**

Based on the above discussion, we propose a system infrastructure for quality improvement using data mining techniques (Fig. 9). There are three layers in the model, namely data collection layer, data analysis layer and data view layer. The function of each layer is described below.

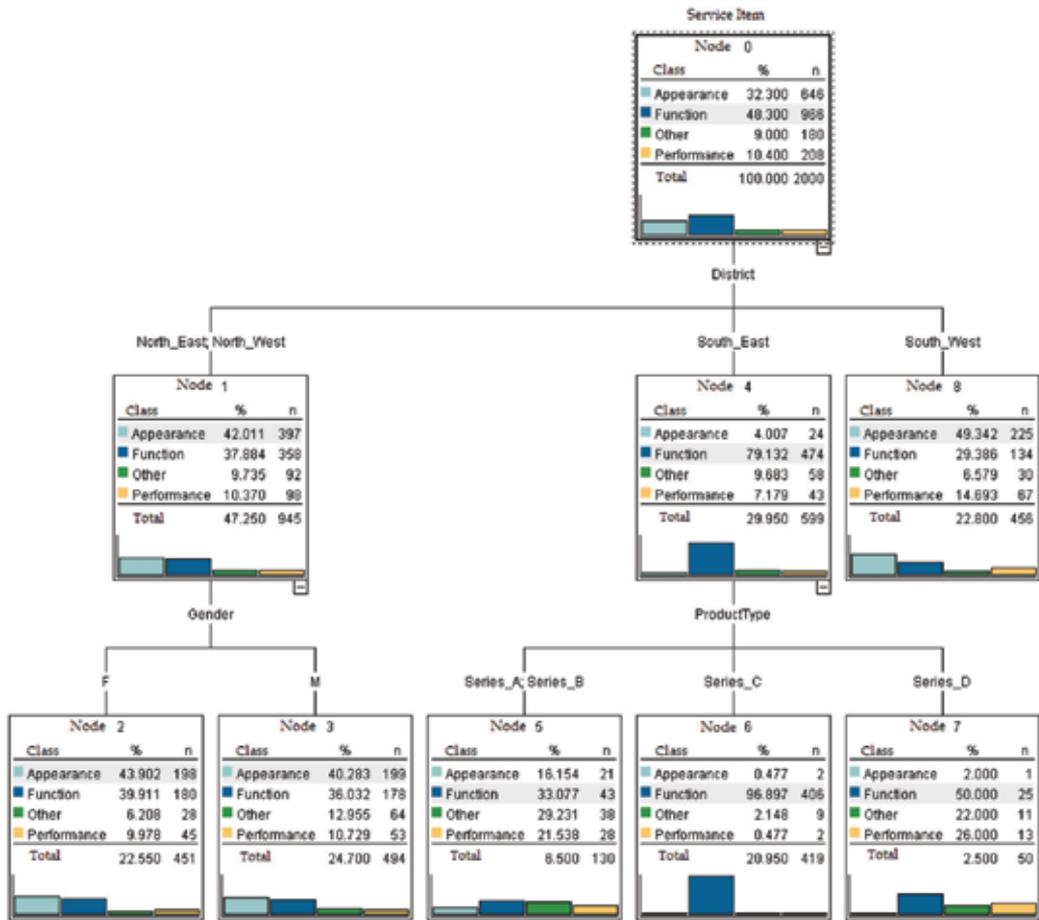


Fig. 8. The decision tree of service data analysis

### 1. Data collection layer

Functions in this layer are mainly used by operators for daily operations. They include quality data maintenance, simple data analysis and query, and generating daily reports.

### 2. Data analysis layer

Functions of this layer are mainly used by mid-level managers and quality engineers. This layer is the center of the system. Most of the quality data analysis functions are provided in this layer including SPC, statistical analysis, data mining, quality diagnosis, etc.

### 3. Data view layer

Functions of this layer are mainly used by top managers for an integrated data view. In this layer, data are often shown in graphs or reports.

Among the four layers, the data collection layer is the basis of the system. All the quality data are collected and stored by this layer. The data analysis layer is the middle layer, where the raw data are processed and analyzed. Results of the data analysis layer can be transferred to the data view layer where can help high-level managers make quality improvement decisions.

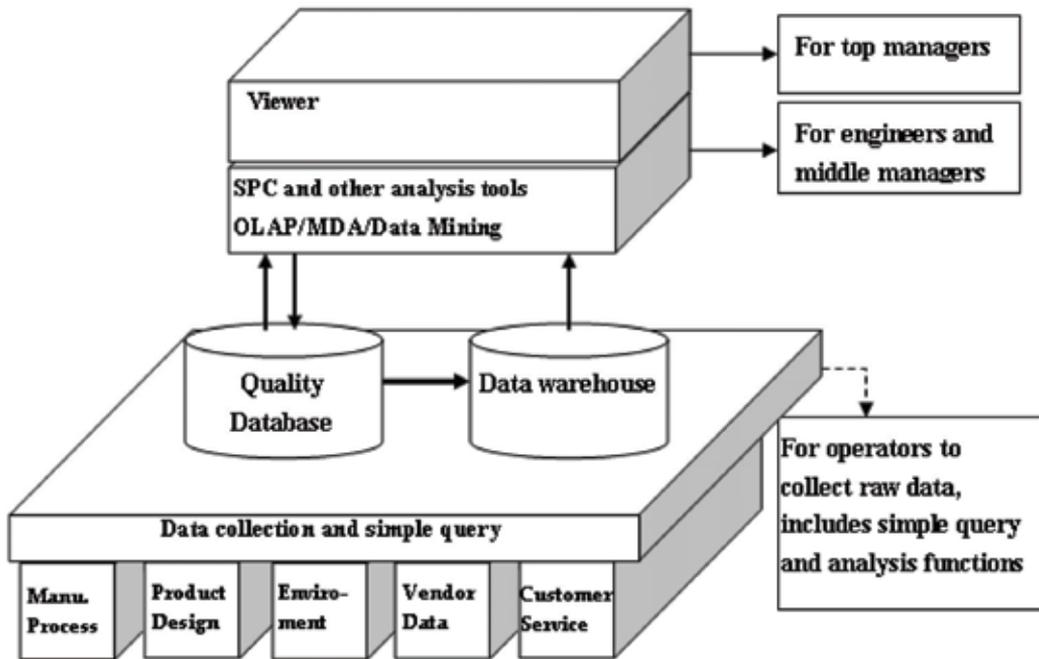


Fig. 9. A system infrastructure for quality improvement using data mining techniques

## 5. Conclusions and further research

In a competitive global market, manufacturing enterprises must stay agile when making quality improvement decisions. The development of IT and other related technologies makes the collection of quality related data easy and cost-effective. However, it is still an open question on how to leverage the large amount of quality data to improve manufacturing quality. This chapter has approached the problem of quality improvement in manufacturing processes using data mining techniques.

Firstly, we proposed a knowledge based six-sigma model where DMAIC for six-sigma was used along with data mining techniques. The knowledge learned by the data mining techniques was helpful in identifying potential quality problems and assisting quality diagnosis. We also examined the problem of parameter optimization by applying data mining techniques to DOE. A decision tree was built in order to dynamically adjust parameter optimization. In addition, we applied data mining to quality diagnosis where an association rule mining technique was used to analyze the electronic test data. The rules obtained by data mining provided a direct guidance in identifying the root causes of the quality problems. And the findings were beneficial for quality diagnosing that is still a difficult problem in six-sigma. Furthermore, a decision tree was also used in the service data analysis. The findings were valuable to improve product design. Finally, we presented a system infrastructure for quality improvement in manufacturing processes.

On the other hand, there are still a lot of open questions to be studied in this field. Firstly, the manufacturing processes are quite different with each other. The quality of the collected data from each manufacturing process varies significantly. There can be a significant portion of missing values and errors in the raw data. It is still a challenging issue as how to

preprocess raw data for data mining algorithms. Secondly, patterns and knowledge learned by data mining techniques are not always usable. How to ascertain the usable knowledge in a large amount rules and patterns is also a problem that deserves attention. Finally, the learned rules and patterns have to be analyzed by domain experts with their domain knowledge. How to present the domain knowledge and build an automated knowledge ascertain system is also a challenging issue in this field.

## 6. Acknowledgement

The authors of this work would like to thank the National Natural Science Foundation of China (NSFC) who sponsored this research (grant no. 70572044). The authors would also like to thank the anonymous reviewers for their constructive comments on this work.

## 7. References

- Andrew Kusiak, Christian Kurasek. Data mining of printed-circuit board defects, *IEEE transactions on robotics and automation*, vol.17, No.2, 2001.
- Angie Patterson, Piero Bonissone, Marc Pavese. Six sigma applied throughout the lifecycle of an automated decision system, *quality and reliability engineering international*, 2005.21: 275-292
- Ben Khediri Issam, Limam Mohamed. Support vector regression based residual MCUSUM control chart for autocorrelated process, *Applied mathematics and computation*, 2008 (In press)
- Chen-Fu Chien, Wen-Chih Wang, Jen-Chieh Cheng. Data mining for yield enhancement in semiconductor manufacturing and an empirical study, *Expert Systems with Applications*, 2007.33: 192-198
- Chen-Fu Chien, Huan-Chung Li and Angus Jeang. Data mining for improving the solder bumping process in the semiconductor packaging industry, *Intelligent systems in accounting, finance and management*, 2006.14: 43-57
- Edgar F. Codd, A Relational Model of Data for Large Shared Data Banks, *Communications of the ACM* 13(6): 377-387.
- Giovanni C Porzio, and Giancarlo Ragozini. Visually mining off-line data for quality improvement, *Quality and reliability engineering international*, 2003.19:273-283
- Hsu-Hwa Chang. A data mining approach to dynamic multiple responses in Taguchi experimental design, *Expert systems with applications*, 2008.35: 1095-1103
- Ian H Witten and Eibe Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann Publishers, 2005.
- J A McCarty, Manoj Hastak. Segmentation approaches in data-mining: a comparison of RFM, CHAID, and logistic regression, *Journal of Business Research*, 2007.60: 656-662.
- Kaidi Zhao, Bing Liu, Tomas M Tirpak and Weimin Xiao. A visual data mining frame work for convenient identification of useful knowledge, *proceedings of the fifth IEEE international conference on data mining*, 2005.
- Myers R H, Montgomery D C. *Response surface methodology*. John Wiley & Sons, New York, 1985.
- Mu-Chen Chen. Ranking discovered rules from data mining with multiple criteria by data envelopment analysis, *Expert systems with application*, 2007.33:1110-1116

- Rakesh Menon, Loh Han Tong, S Sathiyakeerthi, Aarnout Brombacher and Christopher Leong. The needs and benefits of applying textual data mining within the product development process, *Quality and reliability engineering international*, 2004.21:1-15
- Ruey-Shiang Guh, Yeou-Ren Shiue. An effective application of decision tree learning for on-line detection of mean shifts in multivariate control charts, *Computers & Industrial Engineering*, 2008 (In Press)
- Seyed Taghi Akhavan Niaki, Babak Abbasi. Fault diagnosis in multivariate control charts using artificial neural networks, *Quality and reliability engineering international*, 2005.21: 825-840
- Sébastien Gebus, Kauko Leiviskä. Knowledge acquisition for decision support systems on an electronic assembly line, *Expert systems with applications*, 2007 (In press).
- Shao-Chuang Hsu, Chen-Fu Chien. Hybrid data mining approach for pattern extraction from wafer bin map to improve yield in semiconductor manufacturing, *Int J Production Economics*, 2007.107: 88-103
- Tai-Yue Wang, Long-hui Chen. Mean shifts detection and classification in multivariate process: a neural-fuzzy approach, *Journal of intelligent manufacturing*, 2002.12: 211-221

# The Deployment of Data Mining into Operational Business Processes

Rok Rupnik and Jurij Jaklič  
*University of Ljubljana  
Slovenia*

## 1. Introduction

Data mining is progressively used in information systems as a technology to support decision making on the tactical level, as well as to enable decision activities within operational business processes. In general there are three categories of business decision making approaches: (1) decision making based on precisely defined business rules, (2) analytical decision making based on the analysis of information and (3) decision making based on intuition. In many cases there are all three categories used at a time. Business rule based decision making is typical for operational business processes, whereas other two are typical for managerial processes.

Data mining is predominantly used to support analytical decision making, which is typically based on models acquired from a huge quantities of data and therefore makes possible to acquire patterns and knowledge. Based on before introduced discussion one could assume that data mining can be used only in managerial processes. But, there are also operational business processes that require analytical decision activities, e.g. loan approval and classifying the set of customers for promotional mailings. Through data mining methods we can acquire patterns and rules, which can be used as business rules in operational processes. Thus, rules acquired through data mining methods can be used in operational business processes instead of or to support analytical decision activities. Data mining models should in such cases be acquired and used on a daily basis.

Some recent technology achievements, such as JDM API (Java Data Mining Application Interface), enable the possibility to develop application systems which utilize data mining methods and as such do not demand expertise in data mining technology for business users. Through JDM API we can develop transactional application systems or any other application systems which create and use data mining models. It means that application systems which use JDM API can be used to support operational business process with the possibility to utilize business rules acquired through data mining methods.

CRISP-DM 1.0 as the most used data mining methodology introduces four tasks within the deployment phase: plan deployment, plan monitoring and maintenance, produce final report and review project. None of those tasks provides detailed directions for the deployment of data mining models into business processes. The indicator that CRISP-DM 1.0 lacks such detailed directions mentioned is the fact that the CRISP-DM methodology update efforts intend to fulfill the following aim: "Integration and deployment of results with operational systems such as call centers and Web sites".

The deployment of data mining is in fact the deployment of an information technology into business processes. It is therefore recommended to use the same general principles as at deployment of new technologies into business processes. It is true that data mining is in business processes used only in those steps where analytical decisions are needed. In spite of this, when data mining is deployed into business process it should at least to some extent be renovated. It should be renovated because the decision making process is changed and as a consequence it can affect other parts of a business process.

The challenge of deployment of the use of data mining in business processes also encompasses the roles granted to various actors within business process. In the early stages of data mining tools the vendors often emphasized that in the future their tools will be so simple that managers will be able to use them without any assistance. As we know the reality was quite different and often interdisciplinary teams (data mining experts, database experts, experts of statistics) were needed for data mining projects.

As evident from the discussion above, the classical data mining methodologies have to be extended for the cases where data mining is deployed in operational business processes. A proposal of such a methodological framework is presented in this chapter. The next section provides some background material on which the proposed methodology is based. Afterwards the proposal is presented in the form of a process model and described into more details. A case study is used to show an example of how a process can and has to be redesigned when deploying data mining. Besides, the case study serves to evaluate the previously proposed approach. At the end of the chapter, some general conclusions and lessons that can be learned from the case study are given.

## 2. Background

In this section we briefly introduce areas that are important for the mission of the paper.

### 2.1 Data mining

Data mining is the process of analyzing data in order to discover implicit, but potentially useful information and uncover previously unknown patterns and relationships hidden in data (Witten & Frank, 2005). In the last decade, the digital revolution has provided relatively inexpensive and available means to collect and store the data. The increase in data volume causes greater difficulties in extracting useful information for decision support. Traditional manual data analysis has become insufficient, and methods for efficient computer-based analysis indispensable. From this need, a new interdisciplinary field of data mining was born. Data mining encompasses statistical, pattern recognition, and machine learning tools to support the analysis of data and discovery of principles that lie within the data.

The data mining learning problems can be roughly categorized as either *supervised* or *unsupervised*. In supervised learning, the goal is to predict the value of an outcome based on a number of input measures; in unsupervised learning, there is no outcome measure, and the goal is to describe associations and patterns among a set of input measures (Rupnik et al., 2007).

Potential buyers classification naturally fits in supervised learning problems. Finding interesting subgroups of potential buyers and generally interesting associations among attributes requires using unsupervised learning techniques, such as association rules and clustering.

## 2.2 CRISP-DM process model

A data mining process model defines the approach for the use of data mining, i.e. phases, activities and tasks that have to be performed. Data mining represents a rather complex and specialized field. A generic and standardized approach is needed for the use of data mining in order to help organizations use the data mining.

CRISP-DM (CRoss-Industry Standard Process for Data Mining) is a non-proprietary, documented and freely available data mining process model. It was developed by the industry leaders and the collaboration of experienced data mining users, data mining software tool providers and data mining service providers. CRISP-DM is an industry-, tool-, and application-neutral model created in 1996 (Shearer, 2000). Special Interest Group (CRISP-DM SIG) was formed in order to further develop and refine CRISP-DM process model to service the data mining community well. CRISP-DM version 1.0 was presented in 2000 and it is being accepted by business users (Shearer, 2000).

CRISP-DM process model breaks down the life cycle of data mining project into the following six phases which all include a variety of tasks (Shearer, 2000; Rupnik et al., 2007):

- **Business understanding:** focuses on understanding the project objectives from business perspective and transforming it into a data mining problem (domain) definition. At the end of the phase the project plan is produced.
- **Data understanding:** starts with an initial data collection and proceeds with activities in order to get familiar with data, to discover first insights into the data and to identify data quality problems.
- **Data preparation:** covers all activities to construct the final data set from the initial raw data including selection of data, cleaning of data, the construction of data, the integration of data and the formatting of data.
- **Modelling:** covers the creation of various data mining models. The phase starts with the selection of data mining methods, proceeds with the creation of data mining models and finishes with the assessment of models. Some data mining methods have specific requirements on the form of data and to step back to data preparation phase is often necessary.
- **Evaluation:** evaluates the data mining models created in the modeling phase. The aim of model evaluation is to confirm that the models are of high quality to achieve the business objectives.
- **Deployment:** covers the activities to organize knowledge gained through data mining models and present it in a way users can use it within decision making.

Even though CRISP-DM recognizes that creation of the model is generally not the end of the project and introduces four tasks within the deployment phase, it lacks more detailed directions for the deployment of the data mining results into an operational business process, which requires implementation of a repeatable data mining process (Chung & Gray, 1999).

## 2.3 Business processes and business process renovation

Throughout the last twenty years business process orientation gained importance in business community. There are several definitions of business processes. One of the widely used is the one by Davenport and Short (1990) that defines a business process as a set of logically related tasks performed to achieve a defined business outcome. Generally, there are two groups of generic business processes: operational and management processes.

Information is vital to all business processes that make up an organization's operations and management (Chaffey & Wood, 2005), therefore it should not be a surprise, that business process renovation research demonstrated the critical role of information technology in business process renovation (Broadbent et al., 1999). Nowadays information technology (IT) offers very good solutions for implementing business process renovation. The contributions of IT in business process renovation could be categorized in two different ways (Chang, 2000). Firstly, IT contributes heavily as a facilitator to the process of renovation. Secondly, IT contributes in the reengineering process as an enabler to master the new process in the most effective way (Davenport and Short, 1990). However, IT should be the enabler, but not the initiator of business process renovation projects.

In addition to IT, business process renovation requires consideration of organizational and managerial issues, such as cross-functional integration, stakeholder involvement, leadership qualities, and employee motivation. According to (Chaffey & Wood, 2005) the high failure rates for information systems projects are often a consequence of managers' neglect of how users will react to new ways of working.

### **3. The use of data mining in operational business processes**

The use of data mining in business processes is increasing, but has still not reached the level appropriate to the potential benefits of its use. The literature review reveals that it is used mainly for the purposes of decision support (Rupnik et al., 2007). There are only few examples introducing daily use of data mining in business processes (Kohavi & Provost, 2001; Feelders et al., 2000; Chung & Gray, 1999). We can say that data mining is predominantly used to support decision making on a tactical level in decision processes and business processes.

What about the use of data mining to support business processes on operational level, i.e. operational business processes? We believe that the use of data mining in operational business processes represents a potential in cases where decisions can be operationalized based on stable models representing rules, in this case data mining models.

#### **3.1 Related work**

As mentioned before there is not much research done in the area of the use of data mining in operational business processes. There are also some papers discussing the use of data mining in business processes in general and they also represent important basis for our research.

Chung and Gray (Chung & Gray, 1999) argue that there is a lot of research done in the areas of data mining model creation, but there is a lack of research done in the area of the use of data mining models in operational business processes.

Kohavi and Provost (2001) argue that it is important to enable the use of data mining in business processes through automated solutions. They discuss the importance of the ease of integration of data mining in business processes. In their paper they discuss the integration of data mining in business processes as the consequence of the need to incorporate background knowledge in business processes. They state that deploying automated solutions to previously manual processes can be a rife with pitfalls. Authors also argue that one must deal with social issues when deploying automated solutions to previously manual processes.

Feelders et al. (2000) discuss the use of data mining in business processes with the emphasis on the integration of data mining models and solutions into existing application systems within information systems. Authors argue that it is essential that the results of data mining are used to support operational business processes like direct mailing for the selection of potential customers.

Gray (2005) discusses data mining as an option of knowledge sharing within the enterprise. Through the integration of data mining in operational business processes, knowledge sharing is not only present in business processes on a tactical level, but also on operational level.

### 3.2 Prior work

One of the motivations for our research presented in this chapter was also our prior work. We developed a data mining process model and appropriate data mining based decision support system to support decision processes in the telecommunication company (Rupnik et al., 2007). In our research we explored the use of data mining application systems approach of the use of data mining. In data mining application systems approach the data mining is not used in ad-hoc projects, but through data mining based decision support systems. Our aim was to define how business users can use data mining models to facilitate decision making where data mining experts create data mining models and business users use them. One of the responds of the company that we got through the deployment of the system was the estimation that the use of data mining would also bring value added in operational level business processes.

### 3.3 The potentials of the use of data mining in operational business processes

We define operational business process as business process on operational level and they are significantly more structured than managerial processes. Although they are executed on an operational level, there are in many cases also decision-making components present in those processes. The decision-making components and the level of their use in operational business processes can be analyzed on classification of business rules and procedures (Raghu & Vinze, 2007) (Figure 1).

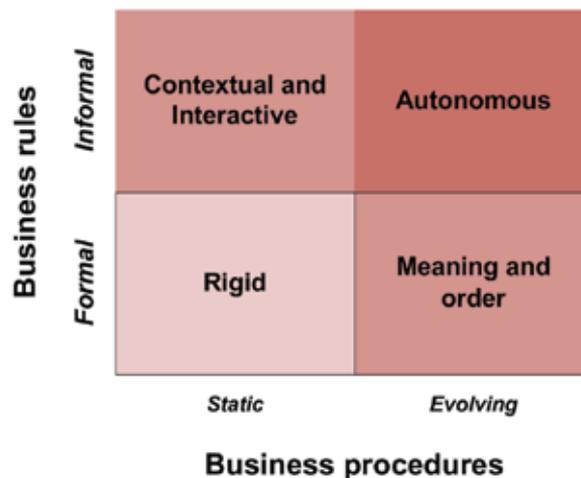


Fig. 1. Classification of decision-making components through business rules and business procedures

Operational business processes are typically based on rather formal business rules and static business procedures (Raghu & Vinze, 2007). We could say that the majority of operational business processes are *rigid* in their decision-making structure, i.e. they have rigid decision-making structure. Formal business rules leave no room for alternative interpretations, which means that those processes practically do not include analytical decision-making components. It is obvious that there is no room for the use of data mining in rigid decision making structures.

When decision-making structure is evolving in business procedures and still has formal business rules, then it is considered as being oriented towards *meaning and order*. Rules remain formal in order to ensure exact equity in the process. The procedures that implement the rules are evolving over time to allow some flexibility in the interpretation of rules. The evolving of decision-making structure is as a result achieved through innovativeness of flexibility in the interpretation of rules in business procedures. As higher level of decision-making structure is achieved through flexibility in the interpretation of rules there is no room for data mining to enhance business rules through lessons learned on past episodes.

*Contextual and interactive* decision-making structures are governed by effective representation of informal business rules. For this kind of knowledge-making structure is important to store and retrieve lessons learned from decision-making episodes from the past. This enables business rules to become more informal and adapted to patterns and acquired by data mining models. Through the use of knowledge acquired in the past highly formal business procedures can be better executed. The role of data mining for contextual and interactive decision-making structures is rather clear. The use of data mining enables the transformation of knowledge-making structure from rigid to contextual and interactive.

*Autonomous* decision-making structure is evolving in business procedures and is informal in business rules. Decision-making structures of this kind support and enable knowledge sharing, storing of knowledge and knowledge retrieval. Interactivity among decision makers both within and outside of process domain has positive effect on autonomous decision-making structures. Further positive effect is the possibility to retrieve knowledge related to solutions and procedures applied to similar decision problems from within and outside the problem domain. Data mining can as technology for acquiring of knowledge and knowledge retrieval in this case also contribute to the transformation to achieve autonomous decision-making structure.

### **3.4 The methodology of the implementation of data mining into operational business processes**

Based on approaches to BI implementations (e.g. Moss & Atre, 2003; Williams & Williams, 2007), on a methodological framework to business process renovation and IS development (Kovačič & Bosilj-Vukšič, 2005), and according to our experience with the implementation of data mining in analytical business process we propose a methodology of the implementation of data mining into operational business processes. The main elements of the proposal are:

- business case assessment,
- DM readiness assessment,
- business process renovation, and
- CRISP-DM as the methodology used for DM itself.

The methodology has three phases where each of them has several activities. We describe the methodology, its phases and activities through subchapters.

### 3.4.1 First phase: exploratory data mining

This purpose of this phase is to evaluate the readiness of operational business process and people involved in it for the implementation of data mining in their operational business process. Beside that the first phase also defines the business value of the use of data mining in the process and evaluates risks and opportunities (Figure 2).

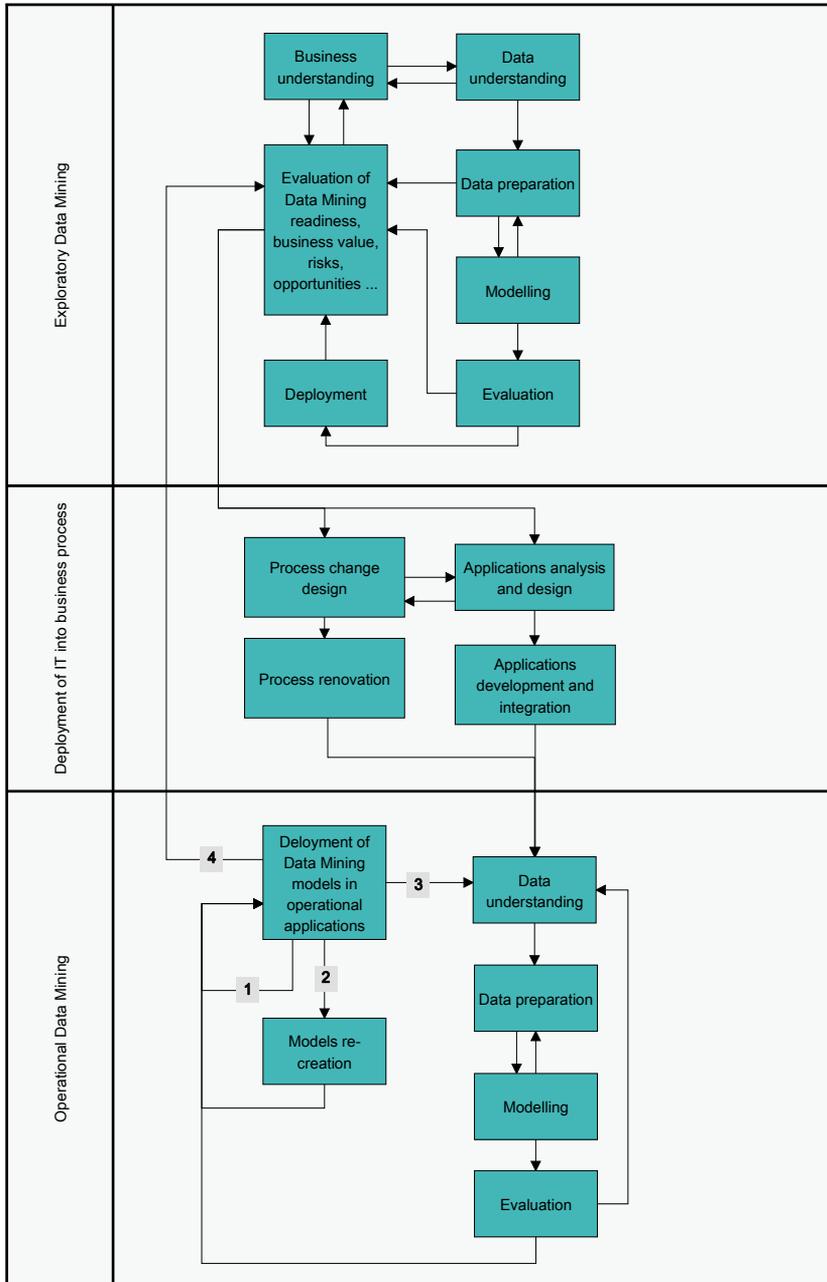


Fig. 2. The methodology of implementation of data mining into operational business processes

The process model of the first phase includes the CRISP-DM process model activities and the evaluation activity, which justifies and confirms (or declines) the implementation of data mining into operational business process. Through CRISP-DM the problem domain is first explored and transformed to a problem definition suitable for data mining implementation through *business understanding* activity. After that the data needed for modeling is defined and prepared through *data understanding* and *data preparation* activities. Data mining models are then created, evaluated and deployed through the following activities: *modeling*, *evaluation* and *deployment*.

It is important to note that the aim of methodology is to implement data mining for daily use, not for ad-hoc use. As a result *data preparation* activity tends to automate data preparation to the highest possible stage and implement it as daily procedure which is executed automatically during night or on demand at any time.

The final activity is *Evaluation of DM readiness, business value, risks and opportunities*. This is core activity of the first phase which carries out the mission of the first phase: it evaluates the readiness of operational business process and people involved in it for the implementation of data mining in their operational business process. Beside that it also defines the business value of the use of data mining in the process, evaluates risks and opportunities. If the evaluation gives positive results, then the second phase is initiated.

### 3.4.2 Second phase: deployment of data mining into operational business process

This phase is in responsible for deployment of data mining in operational business process (Figure 2). For successful implementation of data mining there are two key activities which must be executed more or less simultaneously: *process change design* and *applications analysis and design*. The aim of the former is to define and design the changes that happen in operational business process in order that the use of data mining brings added value. The aim of the latter is to make analysis and design of the application that will use data mining and must be developed. There is not only new application that must be developed, there are also existing applications that must be adapted to the use of data mining models and changes in operational business process. Both activities are very dependent on each other. For example, the application must provide the functionalities suitable for changes designed in activity *process change design*. After the activity *applications analysis and design* is finished the development of application is initiated through activity *applications development and integration*. The aim of this activity is not only to develop application, but also to integrate it into information systems of the company.

When the activity *process change design* is finished the activity *process renovation* is initiated. The aim of this activity is to renovate the operational business process according to the changes defined and designed in previous activity.

### 3.4.3 Third phase: operational data mining

The last phase supports the daily use of data mining in operational business process. The first step of the phase is the re-execution of activity *data understanding*. This activity was already executed in the first phase. But, since between first and the third phase there can be quite a time difference, the activity is re-executed in order to adapt to the possible changes of the databases. After this activity the daily use of data mining through one or more applications is enabled. The activities that are performed daily have symbols in blue color (Figure 2).

Activity *data preparation* is responsible for daily creation of data sets that represents input for *modeling*. Data sets are either created automatically during night or by demand by business users. Modeling and evaluation are executed by the use of application developed in the second phase. The core activity for daily use of data mining in the operational business process is *deployment of data mining models in operational applications*. In this activity data mining models are used in various applications by business users. According to the experience of the use of data mining models the following scenarios are possible (Figure 2):

1. Business users do not notice any disadvantages in data mining models or effects of their use. In this case activity *deployment of data mining models in operational applications* is re-executed daily.
2. Business users notice some disadvantages in data mining models or disadvantages in effects of their use. If they know that there were no such changes done in the databases that represent sources for data preparation, then the activity *models re-creation* is invoked. This way the data mining models are re-created and they reflect the last changes in the contents of source databases.
3. Business users notice some disadvantages in data mining models or disadvantages in effects of their use. If they know that there were recent changes in the databases that represent sources for data preparation, then the activity *data understanding* is invoked. The aim of the re-invoking is to detect the effect of recent changes in the source databases and to find out if there are any changes necessary in the area of data preparation. After the activity *data preparation* is finished activities of *modeling* and *deployment* can be executed, i.e. the daily use of data mining models can go on.
4. Business users notice some disadvantages in data mining models or disadvantages in effects of their use. It can also happen that they know that there are changes happening in the company that affect the domain of operational business process and its mission. In this case the activity *Evaluation of DM readiness, business value, risks and opportunities* is re-executed in order to evaluate (and re-define) the use of data mining in the operational business process.

#### **4. The use of data mining to support direct marketing – case study**

In this section we present an example of data mining deployment into operational business process called *direct marketing*. We also discuss the renovation of the process needed to successfully deploy data mining.

The case study analyses the direct marketing process in a Slovenian publishing company (Publisher), which has published over 20,000 titles in print run approximately 100 millions in the last 60 years. Sales department, which is responsible for direct marketing sales, is a part of the Marketing and Sales unit. Sales through direct sales channels represent approximately 80 % of all sales in the last couple of years. There are several direct sales channels use: telemarketing, direct mail, email marketing, direct selling. Current practice shows that the target groups for different sales channels almost don't overlap. For example, elderly customers prefer telemarketing, which is mostly not preferred by other customer segments. This case study is limited to direct mailing (via snail mail), yet one could expect similar findings for other direct sales channels.

Sales marketing processes are currently supported by the *Libris* application, which is in use for more than 30 years. There are several problems with the *Libris* as not all the direct marketing activities are adequately supported. The main drawback is that each data

processing for selecting prospective buyers from the *Libris* database requires several hours. Therefore queries are run during nights. Consequently, running times for lists of potential customers are very long.

In the current (As-Is) direct marketing process (Figure 3) a product manager first defines a target group for the book that is to be marketed. There are several possible ways or criteria to define the group:

- Through demographic data (gender, age, city) and other characteristics of customer that have bought similar titles in the past are analyzed. *Demographs* and *geographs* are used for this type of analysis (Robertson, 2005).
- Based on the characteristics of "similar books". Product manager creates the list of those books according to his experience, i.e. his business knowledge. The *Libris* application enables several ways for selecting "similar titles", e.g. using the book classification.
- By the use of the RFM (Recency, Frequency, Monetary) method. Each customer that has bought at least one title in the last 5 years has a 3 digit RFM code.

The product manager then fills a paper form with the selected characteristic of the target group. As we can see, the selection criteria are defined using a combination of data that is derived from the entire history of sales and from the knowledge, intuition and experience of product managers. It is obvious that data analysis and business knowledge are very important for direct marketing process, what makes this process rather knowledge intensive.

At this point of the process the IT staff is drawn in the process. The product manager sends them the request for preparing the prospective buyers list that includes the above mentioned paper form with the selection criteria. In most cases the attributes that are used for defining the criteria are standard and predefined. In these cases the IT staff enter the criteria in the *Libris* and the list is created over the night.

There are some cases when the product manager wants to use complex criteria or additional attributes, besides the standard ones. In such situations a preliminary data processing is done in the earlier phase, at the time of criteria definition. This even prolongs the time required for preparing the prospective buyers list. Preferably, the IT staff that is involved in preparing the list should have quite a great deal of business knowledge, too. For example, they have to understand whether the result makes sense if there are X customers on the list. At the moment, there is only one such person (IT staff member – analyst) employed with the Publisher, therefore he is the critical element in the process. This is one of the reasons for which the Publisher has started to consider using modern information technology and data analytical methods in the direct marketing processes.

When the prospective buyers list is prepared, it is reviewed by the product manager and the criteria are modified if necessary. If so, the data processing has to be repeated the next night. When the final list is prepared in the required form, it is forwarded to the printing office by the product manager for printing and personalization of the marketing materials. Responses of the customers are recorded in the database, which enables analysis of the response and the creation of data mining models to improve future campaigns.

An important role of IT staff can be observed from the above description and the model (Figure 3) of the direct marketing process. However, IT staff contributes little to the process value added, as they act mostly as data stewards. Besides, the average process cycle time is rather long due to the night processing. And, as mentioned, each requirement for the mailing list change increases this time for a least one day. Consequently, flexibility and possibility for on-line analyses is diminished.

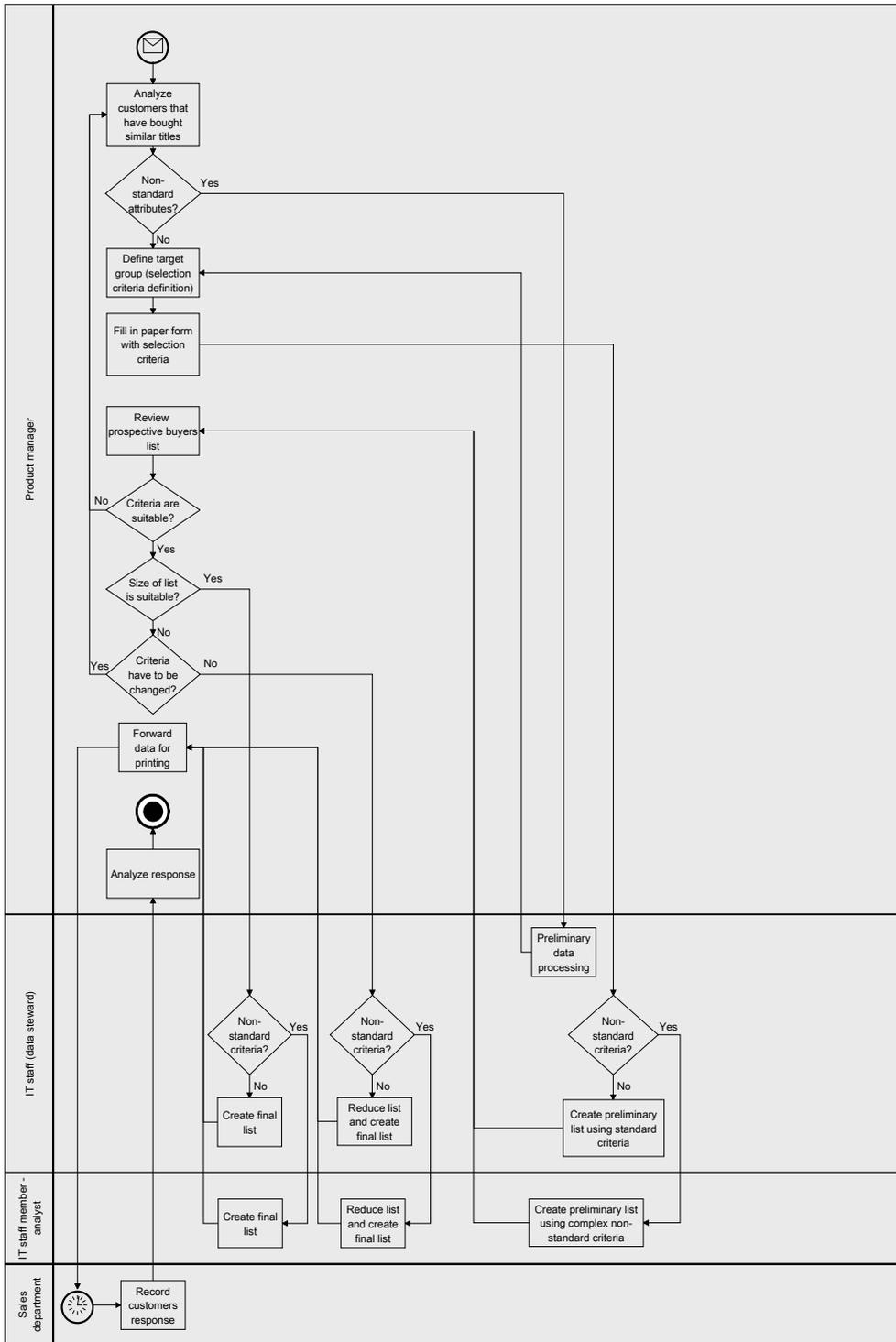


Fig. 3. The AS-IS direct marketing process model

Therefore, an analytical application is under deployment. It is expected to offer a possibility for autonomous analyses to the business users, such as product managers and other sales staff. They will be able to create and analyze prospective buyers lists, to pivot data cubes, to query the data, to perform *what-if* analyses, to define their own filtering criteria, etc. As already noted, the analytical activities in the process are knowledge intensive, thus it is reasonable to expect that there is a vast potential for the use of data mining. Publisher is already considering to implement data mining, in the first phase for determining prospective buyers and for predicting response to marketing campaigns, i.e. to support direct marketing process.

At the moment, all purchases are recorded in the Publisher's database. Besides that, they want to record each contact and communication with customers in the future. This will enable to upgrade from the purchase analysis to the customers analysis.

The TO-BE process model can be easily mapped to the methodological framework, shown on Figure 4. The labels 1, 2, and 3 match to the corresponding labels in the figure that shows the methodological framework.

#### 4.1 TO-BE direct marketing process

The use of Data mining will enable new ways for customer segmentation and discovering customer groups for marketing campaigns. The standard attribute set will not be the only way for segmentation any more, the segmentation on various attributes will be enabled. Publisher is aware that deployment of data mining would require a renovation of the direct marketing process, including changes in the employee roles and responsibilities. After all, the latter is the reason for consideration to deploy new information technologies. As with other information technologies, the data mining is an enabler for business process renovation on one side, however it also requires process changes on the other side.

Step-by step changes of the direct marketing process are planned; first the analytical application will be deployed and later on the data mining application will be launched. Accordingly, the transition is expected to be smoother, particularly the changes related to the changing of roles. In the first phase the product managers will autonomously query and analyze the database. They will learn more about the database content and in this way they will get ready for the data mining, where the knowledge and understanding of the database content is one of the key success factors. As it can be learned from the direct marketing experiences, around 50 % of the success is in good understanding of the data. Only in the second phase, the data mining models for generating the prospective buyers list will be used inside the analytical application. An integration of both systems is planned.

The TO-BE process (Figure 4) shows the direct marketing process flow after the analytical application and data mining will be deployed. Significant differences can be noticed when the AS-IS and TO-BE processes are compared:

- A significant shift in the workload of the IT staff toward the business users can be noticed. The number of activities performed by the IT staff is minimized, and moreover they are not involved in each marketing campaign. They have to acquire the appropriate level of new knowledge in the data mining, as they are mostly involved in building new models and re-creating of models. They are also involved in the designing of problem definitions together with business users, integration of models into the direct marketing application and into the analytical application, etc. Thus, the IT staff will have the role of a Data Mining model constructor.

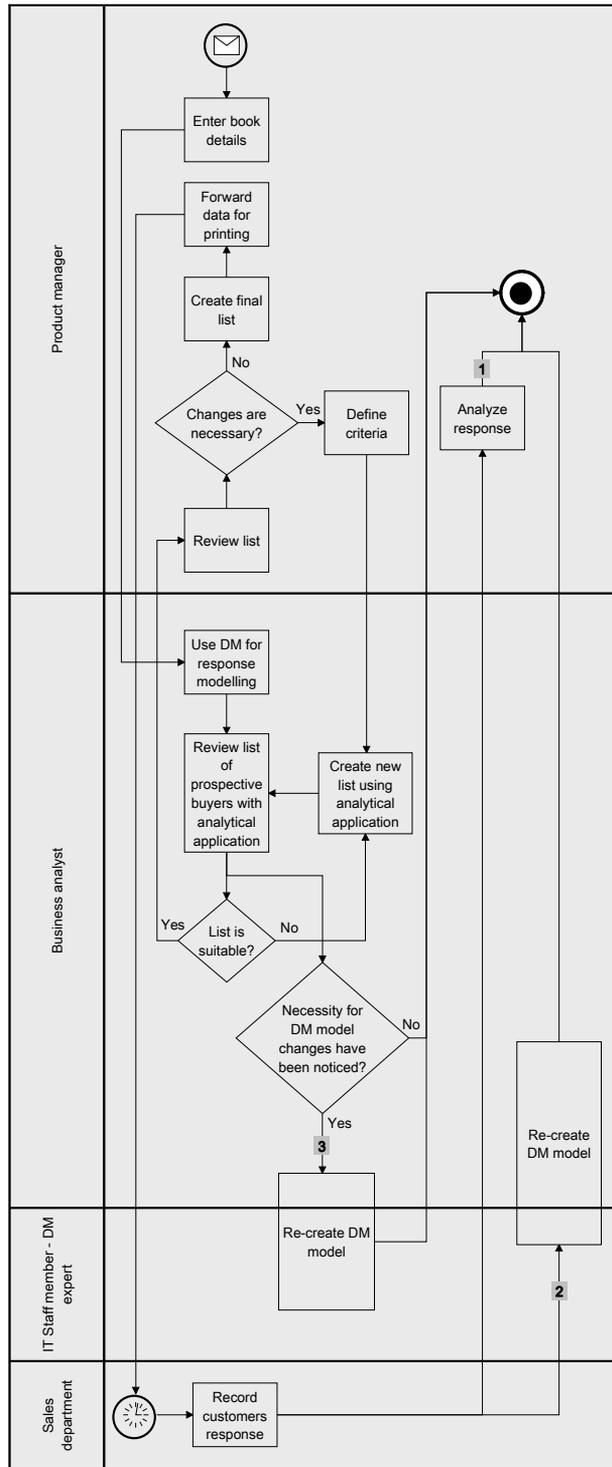


Fig. 4. The TO-BE process model – after the data mining deployment

- It can not be expected that the product managers will be able to do all the additional analyses by themselves, therefore a new role appears on the business users' side: business analyst. Despite this is a business user, an advanced level of information technology knowledge is desired for such a person.
- Analytical activities are much more emphasized and the number of activities with the low added value is decreased. Moreover, the data mining supports the analytical activities.
- An increased number of possible iterations (see the closed loops in the process model) as the consequence of the increased interactivity can be noticed. This enables greater flexibility during the preparation of the prospective buyers list.

## 5. Lessons learned

According to the methodology presented (Figure 2) we have already finished the first phase and the part of the second phase which covers process renovation. The data mining models acquired in the first phase and test instance of direct mailing process gave positive results.

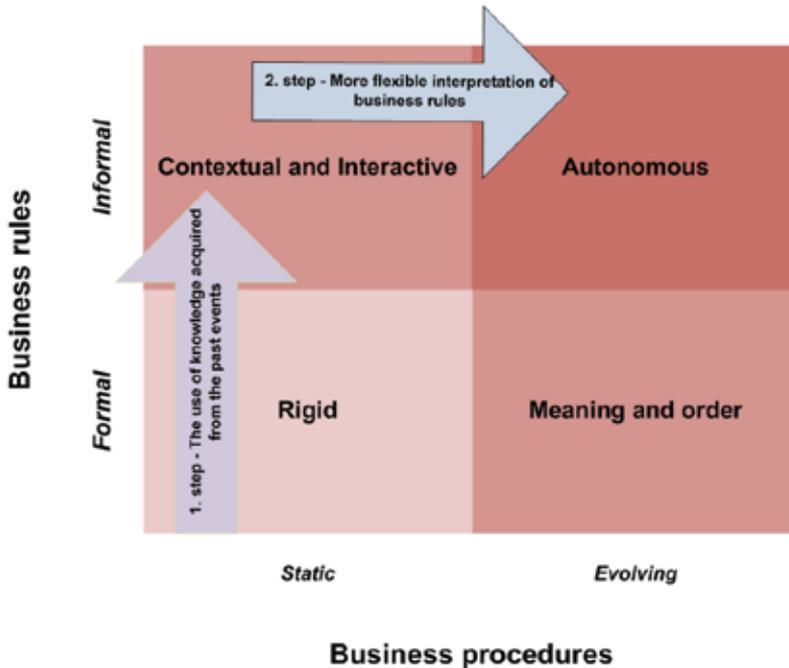


Fig. 5. Horizontal and vertical shift to higher level of decision-making structures

It has become clear that business process renovating and deploying automated solutions is not an easy task. Business users in general agree with the need to deploy data mining into direct marketing process. But, there is a little resistance felt and we know that the third phase will be not an easy task at all.

We believe that through our methodology there will be two shifts to a higher level of decision-making structures (Figure 5). The first step, the vertical shift from formal business rules to informal business rules will be achieved through the use of data mining, i.e. through knowledge acquired through decision-making episodes and patterns from the past.

According to our estimation this will be reached after about one year of use of data mining in the renovated process. The second step, the horizontal shift from static business procedures to evolving business procedures can be achieved through more flexible interpretation of business rules. We believe that for the successful implementation of data mining in operational business processes first the vertical shift must be achieved (Figure 5). After reaching the stage of *contextual and interactive* decision-making structures the advances in interpretation of business rules, also those acquired through data mining models, enable the shift to *autonomous* decision-making structure. We believe that the Publisher, for which we did the research introduced, must reach *autonomous* decision-making structure within their operational business processes. In our opinion successful sales oriented companies must reach autonomous decision-making structure in order to be successful and more agile than their competition (Raghu & Vinze, 2007).

The advances in interpretation of data mining models can be achieved after some time when the company reaches higher level of maturity of the use of data mining. Data mining models with their rules represent the basis to acquire business rules (Figure 6). As the maturity of the use of data mining in the company grows, the business users develop business rules through the evolution of data mining models.

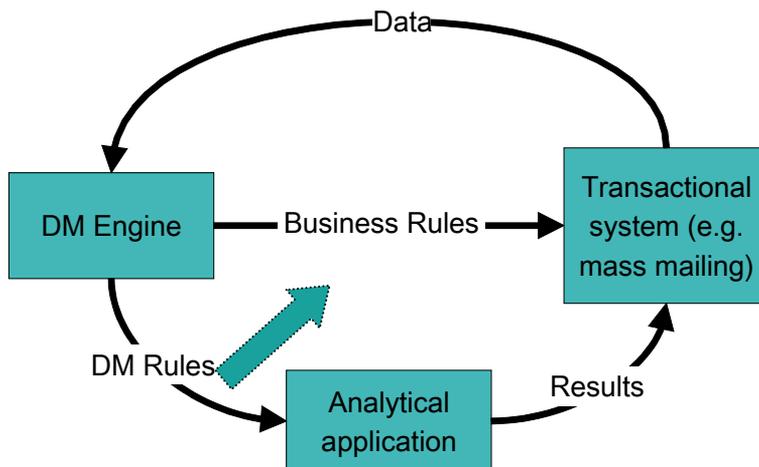


Fig. 6. From data mining rules to business rules

### 5.1 Future work

We already plan activities for our future work. Our contribution delivered so far is methodology of deployment of data mining in operational business processes and direct mailing process renovation. When the third phase starts we will monitor the efficiency of our TO-BE process model. We will do monitoring together with product managers and executives responsible for marketing and sales. Based on monitoring we intend to do modifications in TO-BE process model when it will be needed. We also intend to do modifications in our data mining deployment methodology when there will be any indications requiring it. In our future work we see big potentials in defining criteria for selecting/determining scenario within activity *deployment of data mining models in operational applications* (Figure 2; section 3.4.3).

## 6. References

- Broadbent, M.; Weill, P. & St.Clair, D. (1999). The implications of information technology infrastructure for business process redesign. *MIS Quarterly*, Vol. 23, No 2, 159-182, ISSN 0276-7783
- Chaffey, D. & Wood, S. (2005). *Business Information Management: Improving Performance Using Information Systems*, Prentice Hall, ISBN 0-273-68655-0, Essex
- Chang, S. L. (2000). Information technology in business processes. *Business Process Management Journal*, Vol. 6, No. 3, 224-237, ISSN 1463-7154
- Chung, H.M. & Gray, P. (1999). Special Section: Data mining. *Journal of management information systems*, Vol. 16, No. 1, 11-16, ISSN 0724-1222
- Davenport, T.H. & Short, J.E. (1990). The new industrial engineering: information technology and business process redesign. *Sloan Management Review*, Vol. 31, No. 4, 11-27, ISSN 0019-848X
- Feelders, A.; Daniels, H. & Holsheimer, M. (2000). Methodological and practical aspects of data mining. *Information&Management*, Vol. 37, No. 5, 271-281, ISSN 0378-7206
- Gray, P. (2005). New thinking about the enterprise. *Information systems management*, Vol. 11, No. 1, 91-95, ISSN 1058-0530
- Kohavi, R. & Provost, F. (2001). Applications of data mining to electronic commerce. *Data mining and knowledge discovery*, Vol. 5, No. 5, ISSN 1384-5810
- Kovačič, A. & Bosilj-Vukšič, V. (2005). *Business process management (In Slovene)*. GV založba, ISBN 86-7061-390-5, Ljubljana
- Moss, L. T. & Atre, S. (2003). *Business Intelligence Roadmap: The Complete Project Lifecycle for Decision-Support Applications*. Addison-Wesley Professional, ISBN 0-201-78420-3, Boston
- Raghu, T.S. & Vinze, A. (2007). A business process context for knowledge management. *Decision support systems*, Vol. 43, No. 3, 1061-1079, ISSN 0167-9236
- Robertson, S.P. (2005). Voter-centered design: toward a voter decision support system. *ACM transactions on computer-human interaction*, Vol. 12, No. 2, 263-292, ISSN 1073-0516
- Rupnik, R.; Kukar, M., & Krisper, M. (2007). Integrating data mining and decision support though data mining based decision support system. *Journal of computer information systems*, Vol. 47, No. 3, 89-104, ISSN 0887-4417
- Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. *Journal for data warehousing*, Vol. 5, No. 4, 13-22, ISSN 1548-3924
- Williams, S., & Williams, N. (2007). *The Profit Impact of Business Intelligence*. Morgan Kaufmann, ISBN 978-0-12-372499-1, San Francisco
- Witten, I.H., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan-Kaufmann, New York, USA

# Data Mining Applied to the Instrumentation Data Analysis of a Large Dam

Rosangela Villwock, Maria Teresinha Arns Steiner,  
Andrea Sell Dyminski and Anselmo Chaves Neto  
*Federal University of Paraná  
Brazil*

## 1. Introduction

Dams are conceived with the purpose of bringing great benefits to society. It is expected that their construction, operation and eventual decommissioning should occur safely. If a dam breaks down the destruction scale may be very high; it may put not only the environment in the surrounding areas at risk but also human lives. Therefore, adequate design, construction and operation of dams are a worldwide concern. International guidelines aiming at dams' safety and many productive discussions about this theme have been proposed by the ICOLD - International Commission on Large Dams (ICOLD, 2007).

An adequate auscultation system must be present in dams in order to monitor their structures and foundations during life cycle period. Generally an auscultation system is composed by a set of instruments installed in important points of a dam and of the subsoil where its foundation is based on. These instruments generate a large amount of data, which should be used to understand dam behavior and help engineers in decision making process involving dam safety. Usually the instrumentation readings compose a huge set where important information is mixed with non relevant data. So it would be very useful to have an automatic tool capable to point the significant information or hierarchically organize instrumentation data.

The objective of this work is to present a data mining based methodology to group and organize data from a dam instrumentation system aiming to assist dam safety engineers. The purpose with this work was to select, cluster and rank 72 rods of 30 extensometers located at the F stretch of Itaipu's dam, by means of Multivariate Statistical Analysis techniques. The Principal Components Analysis was used as a method to select the extensometers' rods, Clustering Analysis identified the extensometer rods that were similar and Factor Analysis was used to rank the extensometer rods.

This text is organized as follows: the second section is about Instrumentation system and its relevance to dam safety, the third section describes the KDD Process, the fourth section is a brief description of what Clustering Analysis is, the fifth section describes Itaipu's Dam, the sixth section introduces the Methodology, the seventh section shows the results and the eighth one has the conclusions.

## 2. Dam safety

The concept of “Dam Safety” should involve structural, hydraulic, geotechnical, environmental and operational aspects. All these features must be considered during a dam’s lifespan. A proper instrumentation system capable of monitoring the dam’s geotechnical and structural behavior is essential to assess its behavior and integrity. Good overviews about the relevance of instrumentation to evaluate dam safety can be found in Dibiagio (2000) and Duarte *et al.* (2006).

Some objectives of dam instrumentation and its relationship with structural safety are described in two Engineering Manuals published by U.S.Army Corp of Engineers (1987 and 1995). There, the main objectives of a geotechnical instrumentation plan were grouped into four categories: analytical assessment; prediction of future performance; legal evaluation; and development and verification of future designs. Instrumentation may achieve these objectives by providing quantitative data to assess useful information like groundwater pressure, deformation, total stress, and water levels. Combining visual and periodical inspections with careful data analysis a critical condition can be revealed (FEMA, 2004).

Dam stability must be firstly analyzed during design phase. The geometry of the structures and the properties of the involved materials must be considered as well as the loading conditions. Some basic loading conditions like dam weight ( $W$ ), hydrostatic pressure acting against dam wall (which resultant is  $F_{\text{reservoir}}$ ) and uplift pressure due to seepage in the foundation rock mass (resultant  $F_{\text{uplift}}$ ) are shown in figure 1. The effects of the loads in dam failure process can be many and two of them are shown in figure 1: sliding and overturning. Loading conditions and materials properties can change along dam life cycle and the instrumentation can catch some of these changes.

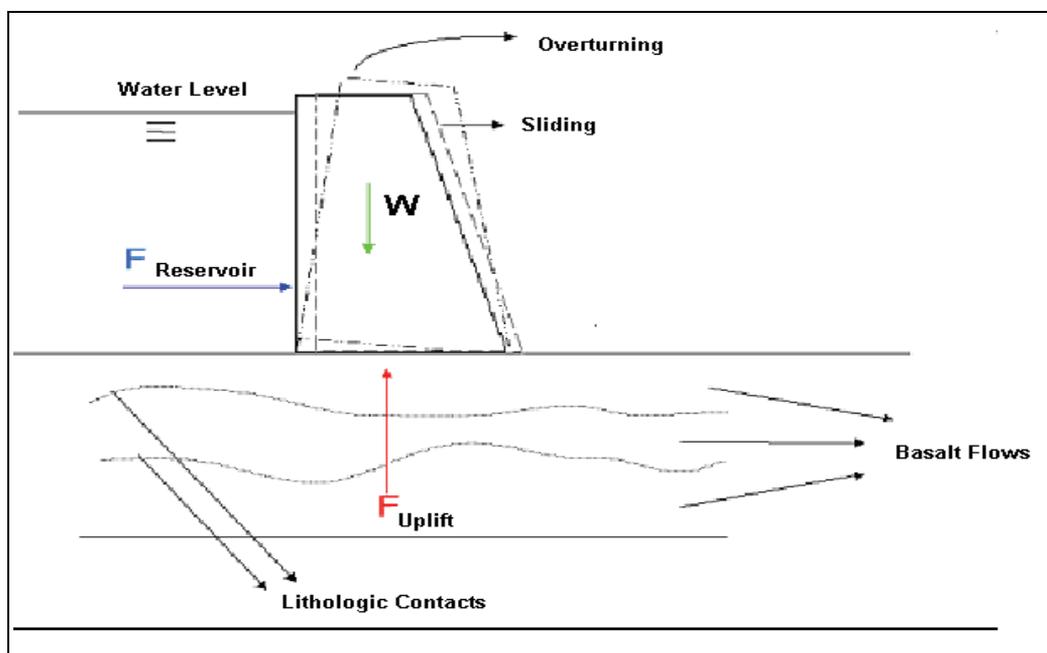


Fig. 1. Basic load condition and some failure processes in concrete gravity dams.

The instrumentation monitoring generates a large data set composed of periodical readings taken during several years. It is important to select the correct array of information to better understand the dam's behavior and solve occasional problems as soon as they occur. Decision making based on instrumentation data usually occurs throughout the lifespan of a dam. An interesting discussion about risk assessment and decision making in dam safety is presented in Bowles *et al.* (2003). Harrald *et al.* (2004) made a good review about some decision making systems and methodologies to help prioritization of tasks and mitigation of failure risk.

Many times, a large amount of data contains useful information, called knowledge. Generally this information is not easily available or identified. Human analysts can spend weeks to discover this knowledge. Because of this fact some huge data sets never receive a detailed analysis (Tan *et al.*, 2005). The more the data volume increases, the more useful are the Data Mining techniques. According to Witten e Frank (2000), ingeniously analyzed data are an invaluable resource for decision-making.

### 3. The KDD process

The acronym KDD means "Knowledge Discovery in Databases" and Fayyad *et al.* (1996) define it as a non-trivial discovery process of valid, new, useful and accessible patterns. The main advantage of the discovery process is that no hypotheses are needed and knowledge is extracted from the data without previous knowledge.

KDD is related to the broad process of discovering information in a database in which there is an emphasis on a high-level application of the particular Data Mining (DM) method. While the DM step is characterized by the extraction of patterns hidden in the data, the whole KDD process is broader and includes all the processing (data selection, pre-processing and transformation) that is needed for this to occur, making it possible to evaluate and interpret the results that were obtained after the DM techniques were used.

The KDD process is a set of continuous activities that include five steps: Data Selection, Pre-processing, Formatting, Data Mining and Interpretation, as shown in Figure 2.

The process starts by understanding the application's domain and the targets that must be reached. Then, a selection can be drawn from these data so that one may work with the data that are of interest. The pre-processing step is the one in which missing or inconsistent data are analyzed and treated. During the formatting step data are prepared so Data Mining can be used as, for instance, to map categorical data among numerical data or using methods to reduce dimensions in the data. According to Silver (1996), pre-processing and formatting may take up to 80% of the time needed for the whole process.

Advancing along the process, there is the Data Mining step, the main one in the KDD process, in which several methods can be used to extract information, which are then presented to the last step, the interpretation, where knowledge is acquired.

If results are not satisfactory, the whole process may be fed back, changing some of the information, which may be reprocessed in the previous steps.

The main purpose with the KDD process is to obtain knowledge hidden in data that may be useful for decision-making, by using methods, algorithms and techniques from different scientific areas. According to Tan *et al.* (2005), these include Statistics, Artificial Intelligence, Machine Learning and Pattern Recognition.

According to Fayyad *et al.* (1996), Data Mining tasks are predictive and descriptive. The predictive ones use some variables to forecast unknown or future values of other variables,

while the descriptive ones find patterns to describe the data. The main tasks of Data Mining are related to pattern Classification, Association and Clustering. In this work, the Data Mining task is to cluster patterns.

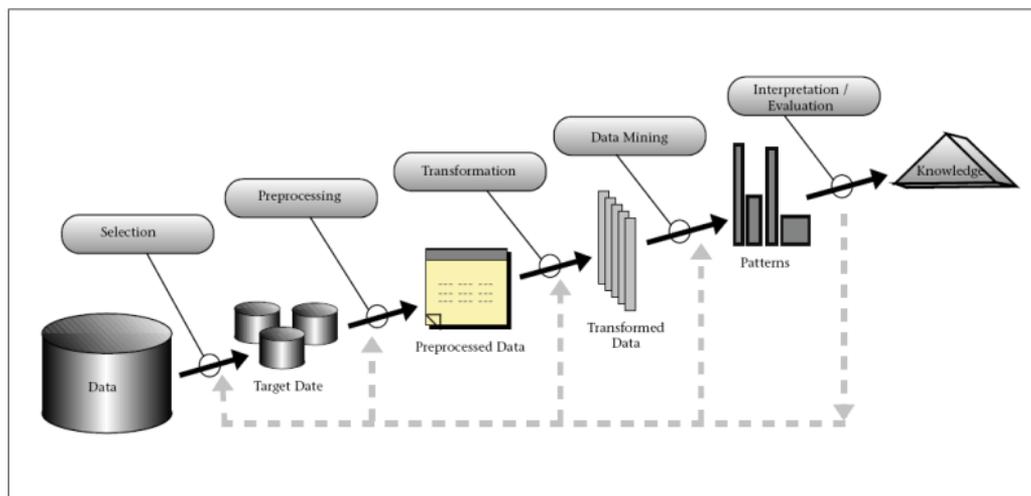


Fig. 2. Steps in the KDD process, Fayyad *et al.* (1996)

#### 4. Clustering analysis

According to Tan *et al.* (2005), Clustering or Segmentation searches for clusters of patterns so that patterns that belong to a same cluster are more similar to one another and dissimilar to patterns in other clusters. According to Hair Jr *et al.* (2005), clustering analysis is an analytical technique to develop objects significant subclusters. Its purpose is to classify the objects into a small number of mutually excluding clusters. Freitas (2002) thinks that in Clustering Analysis it is important to favor a small number of clusters.

Clustering algorithms can be divided into categories, in several ways, according to some characteristics. The two main classes of clustering algorithms are the hierarchic and the partitioning methods. This work used the Ward Method, a hierarchical one.

Hierarchical methods include techniques that search clusters hierarchically and, for this, they admit that several clustering levels can be obtained. According to Diniz e Louzada-Neto (2000), hierarchical methods can be subdivided into dividing and agglomerative ones. The agglomerative hierarchical method first considers each pattern as a cluster and iteratively clusters the pair of clusters with greater similarity into a new cluster, until it forms one single cluster that contains all patterns. On the contrary, the dividing hierarchical method starts with one single cluster and runs a process with successive subdivisions.

The most usual way to represent a hierarchical cluster is through a dendrogram. A dendrogram represents the cluster of patterns and the similarity levels in which clusters are formed. According to Jain *et al.* (1999), a dendrogram can be “split” into different levels to show the different clusters. In the dendrogram showed in Figure 3, admitting a cut at the level shown in the figure, two clusters can be seen: the first one made up by patterns P1, P2 and P5, and the second one made up by patterns P3 and P4.

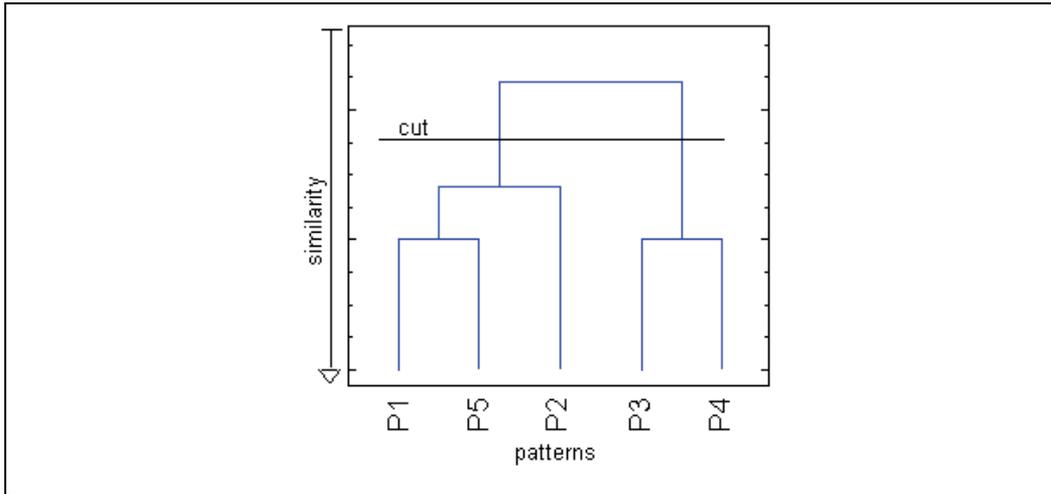


Fig. 3. Dendrogram example

## 5. The Itaipu Dam

According to ITAIPU (2008), ITAIPU Binacional, the largest hydroelectric power plant in the world, started to be built in 1973 at a part of the Paraná River known as ITAIPU, meaning “the singing rock” in tupy, and located at the heart of South America, at the border between Paraguay and Brazil. On 14<sup>th</sup> of November of 1978 were cast 7,207 m<sup>3</sup> of concrete, the equivalent to erecting 24 ten-story buildings in the same day, a Civil Engineering record in South America. In October 1982, the dam works were concluded and on the 5<sup>th</sup> of November, the same year, once the reservoir was filled, the presidents of Brazil, João Figueiredo, and of Paraguay, Alfredo Stroessner, put into operation the mechanism that automatically raises the spillway’s 14 gates, releasing the dammed waters of the Paraná River and officially inaugurating the largest hydroelectric power plant in the world. Presently, ITAIPU has 20 generating units, with 700 MW (*megawatts*) each, adding up to a total installed capacity of 14,000 MW. In 2000, ITAIPU Binacional broke its own record of power production: approximately 93.4 billion kilowatts-hour (KWh) were generated that year. ITAIPU Binacional is responsible for supplying 95% of the power consumed in Paraguay and 24% of all the Brazilian demand.

The ITAIPU dam is 7,919 m long with a maximum height of 196 m, the equivalent to a 65-story building. It took 12.3 million m<sup>3</sup> of concrete and the iron and steel used to build it would be enough to build 380 Tower Eiffel, dimensions that have made the power plant a reference for studies in concrete and dam safety. Figure 4 shows ITAIPU dam’s general structure and Table 1 shows the main features of the different stretches at the dam.

At the ITAIPU dam two segments are earth dams, one is a rockfill dam and another is made of concrete. Along it, and to follow-up the performance of concrete structures and foundations, there are 2,218 instruments (1,362 in concrete and 856 in the foundations and landfills). Of these, 210 are automated and 5,239 are drains (949 in the concrete and 4,290 in the foundations), and their readings occur at different frequencies, which may be, for instance, daily, weekly, biweekly or monthly, according to the type of instrument.

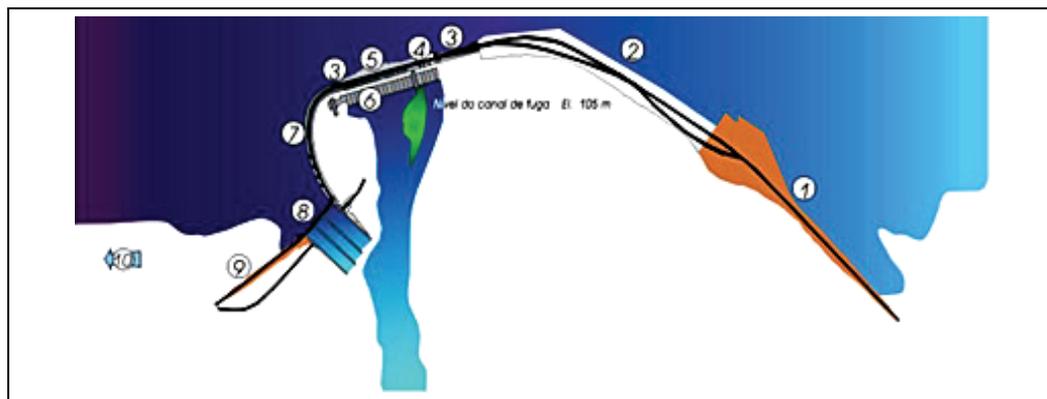


Fig. 4. General Structure of the ITAIPU complex (ITAIPU, 2008)

Stretch		Structure	Length (m)	Maximum Height (m)
1 (L)	Auxiliary Dam	Earth	2,294	30
2 (K)	Auxiliary Dam	Rockfill	1,984	70
3 (E e I) 7 (D)	Side Dams	Buttress	1,438	81
4 (H)	Deviation Structure	Solid Concrete	170	162
5 (F)	Main Dam	Hollow Gravity	612	196
9 (Q)	Auxiliary Dam	Earth	872	25
Others Stretches		Features		
8 (A)	Spillway	350 m wide		
6 (U)	Power House	20 Generating Units		

Table 1. Features of the stretches at the ITAIPU Dam

Although all nine stretches of the ITAIPU dam are instrumented and monitored, the Main Dam (stretch F) is a highlight and deserves a deeper study. Stretch F is where the power generating turbines are located and it is also where the highest water column and greater number of instruments are. Stretch F is constituted by several blocks and each one of them has instruments that supply data about their physical behavior, regarding their concrete structure and foundations.

In order to simplify, but without losing generality, once this same study may be carried out for the other instruments, the extensometer was the instrument that was chosen to apply the methodology in this work, because it is considered one of the most important instruments to monitor a dam and it is one of the instruments that ITAIPU's engineer team has automated. Measurements of settlements at a dam can be made with rod multiple extensometers (Figure 5) installed into probing holes. It is one of the most important observations to supervise the structure's behavior during the dam building, reservoir filling and operation periods. Installing extensometers upstream and downstream at the blocks where there are access galleries that are transversal to the axis allows, according to Silveira (2003), the measurement of angular displacements the dam may show close to the foundations.

By using the extensometers, it is possible to measure vertical displacements of the basaltic rock mass where dam foundation is based on. A typical geological profile of rock mass

foundation can be observed in Figure 6. Settlement monitoring is very important and special attention is given to rock mass discontinuities, such as joints, faults and rock contacts. Each extensometer is installed at a specific location and can be composed by multiple rods of different lengths. Thus, it is possible to separately monitor the vertical displacement of each geological discontinuity.



Fig. 5. Automated extensometer installed in Itaipu's rock foundation.

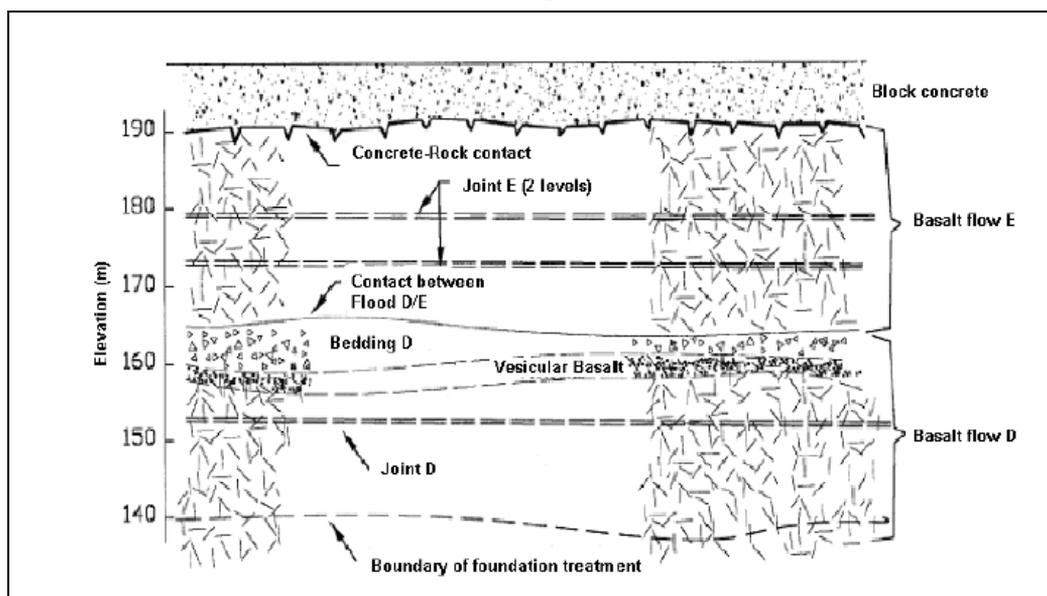


Fig. 6. Basaltic geological profile of Itaipu Dam rock foundation. (Adapted from OSAKO, 2002)

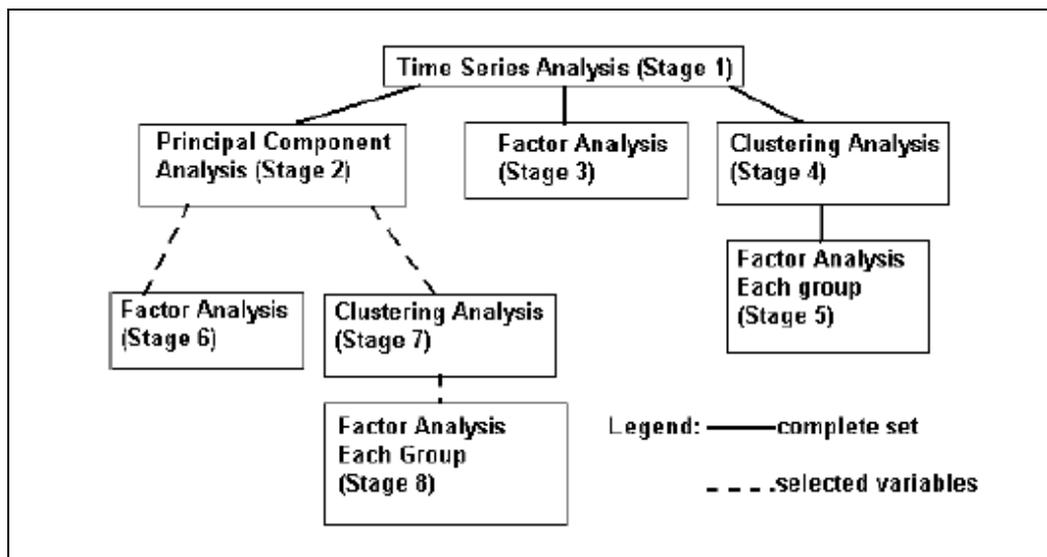
## 6. Methodology

The methodology that is presented in this work was applied to monitoring instruments at the ITAIPU dam, specifically at the dam's F stretch. The chosen instrument was the extensometer. Thirty extensometers are placed in the F stretch, each one bearing one, two or three rods, totaling 72 displacement measurements. These measurements will be identified as follows: equip4\_1, means rod 1 of extensometer 4, for instance.

The data that will be used to develop this work will be monthly data, collected from January 1995 through December 2004, totaling 120 readings. This period was established after a suggestion of ITAIPU's engineering team, because the automatic data acquisition system was implanted after it. During this system's installation phase some instruments had no manual readings. Besides, the automated instruments suffered changes that may have influenced the later readings.

Most instruments render a monthly reading, but some of them render more than a reading per month and in these cases the monthly average was obtained. On the other hand some instruments had missing readings and in these cases interpolations were made by means of time series (stage 1, Flowchart 1, below), thus assuring that all instruments had exactly 120 readings.

Once the time series interpolations were made, were applied simultaneously: Principal Component Analysis (stage 2 - to select the extensometers' rods), Factor Analysis (stage 3 - to rank the extensometers' rods) and Clustering Analysis (stage 4 - to cluster the rods of similar extensometers). Factor Analysis was also applied to each cluster formed by the Clustering Analysis (stage 5). Steps 3, 4 and 5 were once again applied, considering only the extensometer rods selected in step 2, and these were called steps 6, 7 and 8, respectively. The several steps involved in producing this work are shown in Flowchart 1, below.



Flowchart 1. Methodology application steps.

The Principal Component Analysis, as per Hair *et al.* (2005), was used to analyze the relationship between the variables of a set, by transforming the original set into a new one composed by non-correlated variables, called Principal Components, which have special properties in terms of variance. The Principal Components consist in linear combinations of the original variables and they were obtained in a descendant priority order. Most of the data variability can be explained by a small number of principal components.

The main objectives with the Principal Component Analysis were: to reduce the number of variables and indicate which variables, or variable sets, explain most part of the total variability, in order to reveal the type of relationship that exists between them. In the

Principal Components Analysis, it was possible to observe, for instance, that some components represent a non-significant part of the total variability (less than 1%) and that some variables are important (weights greater than 0.5 or smaller than -0.5) for these components. The most important variables that correspond to the least important components should not be selected.

Factor Analysis, as per Hair *et al.* (2005), was the technique that was used to explain the correlations between a large set of variables in terms of a set of few non-observable random variables that are called factors. Within the same cluster, variables can be highly correlated between one another and correlations can be only a few from one cluster to another. Each cluster represents a factor that is responsible for the observed correlations. Communality is the variable's variance part that is distributed throughout the factors.

There are several criteria to establish the number of factors. Kaiser's criterion is the mostly used one and it says that the number of factors should be equal to the number of eigenvalues that are greater than 1.

According to Johnson & Wichern (1998), by means of a rotation one may obtain a structure for the weights, such that each variable has a high weight in one single factor, and low or moderate weights in all the other factors. Kaiser suggested an analytical measure known as the Varimax criterion.

In Factor Analysis there are three types of variance: common, specific and of error. Common is the variance in a variable that is shared with all the other variables in the analysis. Specific variance is associated only with a specific variable. Error variance is the one that is due to the data clustering process' non-reliability, to a measuring error or to a random component in the measured phenomenon. Communalities are estimates of the common variance among the variables.

As communality is the part of the variable's variance that is ascribed to the factors and this represents a percentage of the variable's variation that is not random, the criterion to rank the extensometers' rods consists in sorting the extensometers' rods according to their communalities.

In the present work the method used for the Clustering Analysis was the Ward Method. The Ward Method is an agglomerative hierarchical method. According to Johnson e Wichern (1998), the Ward Method joins two clusters based on the "loss of information" criterion. This criterion can be represented by the Square Quadratic Error (SQE). For each cluster  $i$  are calculated the cluster's mean (or centroid) and square quadratic error ( $SQE_i$ ).  $SQE_i$  is obtained through the sum of the square error of each variable in the cluster in relation to the cluster's mean. For  $k$  clusters there were  $SQE_1, SQE_2, \dots, SQE_k$ , making possible the following definition:

$$SQE = SQE_1 + SQE_2 + \dots + SQE_k \quad (1)$$

For each pair of clusters  $m$  and  $n$ , the mean of the new-formed cluster (cluster  $mn$ ) was calculated. Then, the square quadratic error of cluster  $mn$  ( $SQE_{mn}$ ) was calculated. The square quadratic error (SQE) could be recalculated by using:

$$SQE = SQE_1 + SQE_2 + \dots + SQE_k - SQE_m - SQE_n + SQE_{mn} \quad (2)$$

Clusters  $m$  and  $n$  that resulted in the smallest SQE increase (i.e., the lowest "loss of information") could be merged. According to Hair Jr *et al.* (2005), this method tends to create clusters with the same size because of the minimization of the internal variation.

## 7. Results

As shown in the flowchart, for the 1<sup>st</sup> stage composed by Time Series, the model was automatically chosen according to the Akaike criterion (AIC) and also by observing the root mean squared error (RMSE). We observed the residual integrated periodogram and in some cases, after analyzing the p-values in the parameters “t” testing, the model was substituted by other considered more adequate.

Once the interpolations by Time Series were finished at the 2<sup>nd</sup> and 3<sup>rd</sup> steps, respectively, the Principal Component Analysis and the Factor Analysis were performed in order to select the most important extensometer rods among the 72.

The Principal Component Analysis (Stage 2) showed that 63 components represent a non-significant part of the total variability (less than 1%). Considering the extensometer rods that are important for these components ( $-0.5 \leq \text{weight} \leq 0.5$ ) nine extensometer rods are obtained: equip4\_1, equip4\_2, equip6\_1, equip6\_2, equip13\_2, equip18\_2, equip21\_1, equip22\_1 and equip29\_1.

By means of this analysis the remaining 63 extensometer rods were considered important and should be used in steps 6, 7 and 8. This is not a good criterion to select the most important extensometer rods, once the number of selected rods is very big. However, this reduction of the number of rods may be interesting when other techniques are applied after that (such as in steps 6, 7 and 8).

In the Factor Analysis, stage 3, the variables with low communalities should have been discarded, but no variable had a communality that was smaller than 0.71. Communalities equal to 0.71 indicate that 71% of the variable’s variance is distributed among the factors and only 29% is random. This meant that the corresponding instrument or rod was working well. Table 2 shows the 25 extensometer rods with the highest communalities.

Communality	Rod	Communality	Rod	Communality	Rod
0.988861	<i>equip29_1</i>	0.968213	<i>equip23_2</i>	0.957609	<i>equip14_3</i>
0.981763	<i>equip21_2</i>	0.968083	<i>equip4_1</i>	0.953036	<i>equip25_1</i>
0.976523	<i>equip23_1</i>	0.967029	<i>equip21_1</i>	0.950395	<i>equip33_2</i>
0.975655	<i>equip22_1</i>	0.966632	<i>equip4_2</i>	0.949394	<i>equip24_2</i>
0.972231	<i>equip3_1</i>	0.965999	<i>equip29_2</i>	0.949108	<i>equip24_1</i>
0.971971	<i>equip1_1</i>	0.965522	<i>equip34_3</i>	0.948646	<i>equip5_1</i>
0.971798	<i>equip22_3</i>	0.964925	<i>equip6_1</i>	0.943644	<i>equip28_1</i>
0.970804	<i>equip11_1</i>	0.963139	<i>equip6_2</i>		
0.970397	<i>equip1_2</i>	0.960121	<i>equip22_2</i>		

Table 2. 25 extensometer rods with the highest communalities

The dendrogram in Figure 7 shows how the Clustering Analysis formed the clusters (4<sup>th</sup> step). From the first cut, two clusters result. The first cluster, herein called Cluster A, is a cluster composed by extremely important instruments for dam monitoring. They are instruments installed at the block’s axis, upstream the dam and inclined 60° upstream.

From the second cut we have how the two additional clusters were formed. The first one, called Cluster B, has most of the extensometer rods installed at Spillings B, C and D, and at Contacts B/C and C/D. The second cluster, called Cluster C, has most of the extensometer rods installed at Joints A and B, and at Contact A/B. Table 3 shows the extensometer rods in

each cluster, the inclination, distance from the dam's axis and the attribute where the rod is installed.

Cluster	Rod	Inclination	Distance from the dam's axis	Shape
A	EMF001/1	60° upstream	125.5 Meters upstream	Joint B
A	EMF001/2	60° upstream	105.4 Meters upstream	Contact B/C
A	EMF004/1	60° upstream	65.3 Meters upstream	Contact C/D
A	EMF004/2	60° upstream	60.4 Meters upstream	Fractured Rock
A	EMF006/1	60° upstream	150.8 Meters upstream	Joint A
A	EMF006/2	60° upstream	110.5 Meters upstream	Spilling B
A	EMF021/1	60° upstream	159.8 Meters upstream	Joint A
A	EMF021/2	60° upstream	135.1 Meters upstream	Spilling B
A	EMF026/1	60° upstream	139.2 Meters upstream	Joint B
A	EMF026/2	60° upstream	115.6 Meters upstream	Contact B/C
A	EMF031	60° upstream	64.7 Meters upstream	Contact C/D
B	EMF002/1	0	32.0 Meters upstream	Contact C/D
B	EMF002/2	0	32.0 Meters upstream	Fractured Rock
B	EMF003/1	0	32.0 Meters upstream	Spilling C
B	EMF003/2	0	32.0 Meters upstream	Spilling D
B	EMF005/1	0	13.0 Meters downstream	Contact C/D
B	EMF005/2	0	13.0 Meters downstream	Spilling D
B	EMF007/3	0	13.0 meters downstream	Spilling B
B	EMF008/2	0	84.0 meters upstream	Fractured Rock
B	EMF008/3	0	84.0 meters upstream	Spilling B
B	EMF012/1	60° downstream	47.2 meters downstream	Fractured Rock
B	EMF012/2	60° downstream	42.5 meters downstream	Dense Basalt
B	EMF013/2	0	44.0 meters downstream	Fractured Rock
B	EMF013/3	0	44.0 meters downstream	Spilling B
B	EMF014/2	0	54.0 meters downstream	Fractured Rock
B	EMF014/3	0	54.0 meters downstream	Spilling B
B	EMF015/1	0	80.0 meters upstream	Fractured Rock
B	EMF015/2	0	80.0 meters upstream	Spilling B
B	EMF018/3	0	33.0 meters downstream	Spilling B
B	EMF019/3	0	55.0 meters downstream	Spilling B
B	EMF020/2	0	82.0 meters upstream	Fractured Rock
B	EMF020/3	0	82.0 meters upstream	Spilling B
B	EMF023/3	0	36.0 meters downstream	Fractured Rock
B	EMF024/3	0	62.0 meters downstream	Spilling B
B	EMF025/2	0	75.0 meters upstream	Spilling B
B	EMF025/3	0	75.0 meters upstream	Spilling B
B	EMF027/1	30° upstream	16.6 meters downstream	Joint B
B	EMF027/2	30° upstream	22.6 meters downstream	Contact B/C
B	EMF029/2	30° downstream	55.7 meters downstream	Contact B/C
B	EMF032/1	30° upstream	36.5 meters upstream	Joint B

Cluster	Rod	Inclination	Distance from the dam's axis	Shape
B	EMF032/2	30° upstream	14.6 meters upstream	Spilling C
B	EMF032/3	30° upstream	7.5 meters upstream	Contact C/D
B	EMF033/1	0	0.0	Joint B
B	EMF033/2	0	0.0	Spilling C
B	EMF033/3	0	0.0	Contact C/D
B	EMF034/3	30° downstream	7.5 meters downstream	Contact C/D
B	EMF035/1	90° upstream	0.0	Concrete
B	EMF035/2	90° upstream	0.0	Concrete
C	EMF007/1	0	13.0 meters downstream	Joint A
C	EMF007/2	0	13.0 meters downstream	Contact A/B
C	EMF008/1	0	84.0 meters upstream	Contact A/B
C	EMF011	0	81.0 meters upstream	Joint A
C	EMF013/1	0	44.0 meters downstream	Contact A/B
C	EMF014/1	0	54.0 meters downstream	Contact A/B
C	EMF018/1	0	33.0 meters downstream	Joint A
C	EMF018/2	0	33.0 meters downstream	Fractured Rock
C	EMF019/1	0	55.0 meters downstream	Joint A
C	EMF019/2	0	55.0 meters downstream	Fractured Rock
C	EMF020/1	0	82.0 meters upstream	Fractured Rock
C	EMF022/1	0	68.0 meters upstream	Joint A
C	EMF022/2	0	68.0 meters upstream	Fractured Rock
C	EMF022/3	0	68.0 meters upstream	Spilling B
C	EMF023/1	0	36.0 meters downstream	Joint A
C	EMF023/2	0	36.0 meters downstream	Fractured Rock
C	EMF024/1	0	62.0 meters downstream	Joint A
C	EMF024/2	0	62.0 meters downstream	Fractured Rock
C	EMF025/1	0	75.0 meters upstream	Joint A
C	EMF028/1	0	40.0 meters downstream	Joint B
C	EMF028/2	0	40.0 meters downstream	Contact B/C
C	EMF029/1	30° downstream	63.5 meters downstream	Joint B
C	EMF034/1	30° downstream	36.6 meters downstream	Joint B
C	EMF034/2	30° downstream	21.0 meters downstream	Spilling C

Table 3. Extensometer rods in each cluster

Observing the three clusters A, B and C obtained at the 4<sup>th</sup> step, at the 5<sup>th</sup> step Factor Analysis was applied within each cluster in order to rank the extensometer rods. Tables 4, 5 and 6 show the extensometer rods, and their communalities, for clusters A, B and C, respectively.

Considering the third cut, Cluster B was divided into two clusters called B1 and B2. The highlight is Cluster B2, which is formed mostly by extensometer rods installed at Spilling B. Cluster C was divided into two clusters called C1 and C2. Dyminski *et al.* (2008) introduces a methodology to identify the most important extensometer rods in these five clusters, by using Factor Analysis applied within each cluster. Visual Data Mining were used to examine relationships between extensometers in Silva Neto *et al.* (2008).

Communality	Rod	Communality	Rod
0.961839	<i>equip21_1</i>	0.881852	<i>equip26_1</i>
0.956979	<i>equip21_2</i>	0.854339	<i>equip6_2</i>
0.953791	<i>equip4_1</i>	0.809062	<i>equip1_2</i>
0.94278	<i>equip4_2</i>	0.798566	<i>equip26_2</i>
0.911982	<i>equip6_1</i>	0.677401	<i>equip31_1</i>
0.885099	<i>equip1_1</i>		

Table 4. Extensometer rods and their communalities - Cluster A

Communality	Rod	Communality	Rod	Communality	Rod
0.958451	<i>equip3_1</i>	0.899792	<i>equip35_1</i>	0.833353	<i>equip19_3</i>
0.957903	<i>equip29_2</i>	0.896858	<i>equip2_2</i>	0.832945	<i>equip15_1</i>
0.956985	<i>equip34_3</i>	0.895818	<i>equip3_2</i>	0.826735	<i>equip18_3</i>
0.949573	<i>equip14_3</i>	0.895145	<i>equip8_2</i>	0.818713	<i>equip15_2</i>
0.942628	<i>equip33_3</i>	0.894009	<i>equip14_2</i>	0.80722	<i>equip33_1</i>
0.938697	<i>equip33_2</i>	0.890046	<i>equip23_3</i>	0.77312	<i>equip12_2</i>
0.930857	<i>equip32_3</i>	0.888025	<i>equip25_3</i>	0.693537	<i>equip27_1</i>
0.92853	<i>equip5_1</i>	0.859488	<i>equip5_2</i>	0.692067	<i>equip20_3</i>
0.928364	<i>equip27_2</i>	0.854581	<i>equip8_3</i>	0.688618	<i>equip25_2</i>
0.925976	<i>equip13_2</i>	0.85328	<i>equip32_2</i>	0.635662	<i>equip32_1</i>
0.91628	<i>equip20_2</i>	0.847534	<i>equip35_2</i>	0.624787	<i>equip7_3</i>
0.905482	<i>equip12_1</i>	0.84403	<i>equip2_1</i>		
0.903051	<i>equip13_3</i>	0.843003	<i>equip24_3</i>		

Table 5. Extensometer rods and their communalities - Cluster B

Communality	Rod	Communality	Rod	Communality	Rod
0.975487	<i>equip29_1</i>	0.924851	<i>equip22_3</i>	0.783119	<i>equip7_1</i>
0.966626	<i>equip23_1</i>	0.917463	<i>equip19_2</i>	0.766859	<i>equip34_2</i>
0.952144	<i>equip23_2</i>	0.909212	<i>equip11_1</i>	0.732343	<i>equip7_2</i>
0.946772	<i>equip24_1</i>	0.899795	<i>equip14_1</i>	0.730472	<i>equip18_1</i>
0.945604	<i>equip22_1</i>	0.866861	<i>equip28_2</i>	0.680493	<i>equip18_2</i>
0.943369	<i>equip24_2</i>	0.853862	<i>equip13_1</i>	0.660694	<i>equip28_1</i>
0.942178	<i>equip34_1</i>	0.845538	<i>equip25_1</i>	0.647952	<i>equip19_1</i>
0.928596	<i>equip22_2</i>	0.812306	<i>equip20_1</i>	0.60656	<i>equip8_1</i>

Table 6. Extensometer rods and their communalities - Cluster C

Considering only the 63 rods selected by the Principal Component Analysis (2<sup>nd</sup> step), Factor Analysis and Clustering Analysis were performed during the 6<sup>th</sup> and 7<sup>th</sup> steps, respectively.

Through the Factor Analysis that was applied to the 63 rods (6<sup>th</sup> step), it was observed that no extensometer rod showed a communality that was smaller than 0.7. A communality equal to 0.7 means that 70% of the rod's variance is ascribed to the factors and that only 30% of the variance is random, this is, the corresponding rods worked well. Table 7 shows the 25 extensometer rods with the highest communalities.

During the Clustering Analysis that was applied to the 63 rods (7<sup>th</sup> step) were considered the 3 clusters formed by the second cut, shown in Figure 8, through equivalence with the 4<sup>th</sup>

step. Table 8 shows the extensometer rods in each cluster, the inclination, distance from the dam's axis and the attribute where the rod is installed.

Communality	Rods	Communality	Rods	Communality	Rods
0.974392	equip23_1	0.959987	equip29_2	0.939298	equip34_1
0.970051	equip21_2	0.956602	equip14_3	0.939156	equip33_3
0.96932	equip3_1	0.954664	equip23_2	0.938739	equip14_1
0.968294	equip1_1	0.952711	equip33_2	0.928829	equip27_2
0.968093	equip11_1	0.950157	equip24_1	0.928208	equip13_1
0.966944	equip34_3	0.947958	equip5_1	0.92547	equip26_2
0.96429	equip1_2	0.943773	equip19_2	0.92472	equip20_2
0.960271	equip22_2	0.941997	equip28_1		
0.96008	equip22_3	0.94105	equip24_2		

Table 7. 25 extensometer rods with the highest communalities - 63 rods

Cluster	Rod	Inclination	Distance from the dam's axis	Shape
A	EMF001/1	60° upstream	125.5 meters upstream	Joint B
A	EMF001/2	60° upstream	105.4 meters upstream	Contact B/C
A	EMF021/2	60° upstream	135.1 meters upstream	Spilling B
A	EMF026/1	60° upstream	139.2 meters upstream	Joint B
A	EMF026/2	60° upstream	115.6 meters upstream	Contact B/C
A	EMF031	60° upstream	64.7 meters upstream	Contact C/D
B	EMF002/1	0	32.0 meters upstream	Contact C/D
B	EMF002/2	0	32.0 meters upstream	Fractured Rock
B	EMF003/1	0	32.0 meters upstream	Spilling C
B	EMF003/2	0	32.0 meters upstream	Spilling D
B	EMF005/1	0	13.0 meters downstream	Contact C/D
B	EMF005/2	0	13.0 meters downstream	Spilling D
B	EMF007/3	0	13.0 meters downstream	Spilling B
B	EMF008/2	0	84.0 meters upstream	Fractured Rock
B	EMF008/3	0	84.0 meters upstream	Spilling B
B	EMF012/1	60° downstream	47.2 meters downstream	Fractured Rock
B	EMF012/2	60° downstream	42.5 meters downstream	Dense Basalt
B	EMF013/3	0	44.0 meters downstream	Spilling B
B	EMF014/2	0	54.0 meters downstream	Fractured Rock
B	EMF014/3	0	54.0 meters downstream	Spilling B
B	EMF015/1	0	80.0 meters upstream	Fractured Rock
B	EMF015/2	0	80.0 meters upstream	Spilling B
B	EMF018/3	0	33.0 meters downstream	Spilling B
B	EMF019/3	0	55.0 meters downstream	Spilling B
B	EMF020/2	0	82.0 meters upstream	Fractured Rock
B	EMF020/3	0	82.0 meters upstream	Spilling B
B	EMF023/3	0	36.0 meters downstream	Fractured Rock
B	EMF024/3	0	62.0 meters downstream	Spilling B

B	EMF025/2	0	75.0 meters upstream	Spilling B
B	EMF025/3	0	75.0 meters upstream	Spilling B
B	EMF027/1	30° upstream	16.6 meters downstream	Joint B
B	EMF027/2	30° upstream	22.6 meters downstream	Contact B/C
B	EMF029/2	30° downstream	55.7 meters downstream	Contact B/C
B	EMF032/1	30° upstream	36.5 meters upstream	Joint B
B	EMF032/2	30° upstream	14.6 meters upstream	Spilling C
B	EMF032/3	30° upstream	7.5 meters upstream	Contact C/D
B	EMF033/1	0	0.0	Joint B
B	EMF033/2	0	0.0	Spilling C
B	EMF033/3	0	0.0	Contact C/D
B	EMF034/3	30° downstream	7.5 meters downstream	Contact C/D
B	EMF035/1	90° upstream	0.0	Concrete
B	EMF035/2	90° upstream	0.0	Concrete
C	EMF007/1	0	13.0 meters downstream	Joint A
C	EMF007/2	0	13.0 meters downstream	Contact A/B
C	EMF008/1	0	84.0 meters upstream	Contact A/B
C	EMF011	0	81.0 meters upstream	Joint A
C	EMF013/1	0	44.0 meters downstream	Contact A/B
C	EMF014/1	0	54.0 meters downstream	Contact A/B
C	EMF018/1	0	33.0 meters downstream	Joint A
C	EMF019/1	0	55.0 meters downstream	Joint A
C	EMF019/2	0	55.0 meters downstream	Fractured Rock
C	EMF020/1	0	82.0 meters upstream	Fractured Rock
C	EMF022/2	0	68.0 meters upstream	Fractured Rock
C	EMF022/3	0	68.0 meters upstream	Spilling B
C	EMF023/1	0	36.0 meters downstream	Joint A
C	EMF023/2	0	36.0 meters downstream	Fractured Rock
C	EMF024/1	0	62.0 meters downstream	Joint A
C	EMF024/2	0	62.0 meters downstream	Fractured Rock
C	EMF025/1	0	75.0 meters upstream	Joint A
C	EMF028/1	0	40.0 meters downstream	Joint B
C	EMF028/2	0	40.0 meters downstream	Contact B/C
C	EMF034/1	30° downstream	36.6 meters downstream	Joint B
C	EMF034/2	30° downstream	21.0 meters downstream	Spilling C

Table 8. Extensometer rods in each cluster – 63 rods

Observing the three clusters obtained at the 7<sup>th</sup> step, at the 8<sup>th</sup> step Factor Analysis was applied within each cluster in order to rank the extensometer rods. Tables 9, 10 and 11 show the extensometer rods, and their communalities, for clusters 1, 2 and 3, respectively.

## 8. Conclusions

This work presents a methodology that can be included in the KDD area. The purpose in selecting, clustering and ranking the extensometers' rods is to be able to maximize the

efficacy and efficiency of readings analyses, by identifying similar extensometer rods, as well as the main instruments.

In this work the methodology is applied to only one instrument: the extensometers located at the dam's Stretch F. There are a total of 30 extensometers with one, two or three rods, located at different points of Stretch F, totaling 72 displacement measurements. It is worth pointing out that from the 72 measurements, the company has automated 24.

During the Principal Component Analysis (2<sup>nd</sup> step), from the nine extensometer rods important for insignificant components, this is, they would not be considered in further analyses, seven are common with the ones Itaipu has automated. On the other hand, of the remaining 63 extensometer rods that were considered important, 17 are common with the ones Itaipu has automated. This is not a good criterion to select the important extensometer rods, once the number of selected rods is very big. However, this reduction of the number of rods may be interesting when other techniques are applied after that.

During the Factor Analysis (3<sup>rd</sup> step) it was observed that the extensometer rods worked well. Table 2 presents the 25 extensometer rods with the highest communalities. The 14 rods Itaipu's Engineering team automated are highlighted in italic.

During the Clustering Analysis (4<sup>th</sup> step) it was evidenced that it is possible to discover technical justifications for the cluster formation.

Observing the three clusters A, B and C obtained at the 4<sup>th</sup> step, at the 5<sup>th</sup> step Factor Analysis was applied within each cluster in order to rank the extensometer rods. The extensometer rods marked in italic in Tables 4, 5 and 6 are those Itaipu has automated. One can notice that the automated extensometer rods are, in most times, among the first of each cluster's ranking.

During the Factor Analysis applied to the 63 rods (6<sup>th</sup> step) it was observed that the extensometer rods worked well. Table 7 shows the 25 extensometer rods with the highest communalities. Highlighted in italic are the 10 rods Itaipu's Engineering team has automated, a figure that is smaller than the one found during the 3<sup>rd</sup> step.

Observing the three clusters obtained at the 7<sup>th</sup> step, at the 8<sup>th</sup> step Factor Analysis was applied within each cluster in order to rank the extensometer rods. The extensometer rods marked in italic in Tables 9, 10 and 11 are those Itaipu has automated. One can notice that the automated extensometer rods are, in most times, among the first of each cluster's ranking. The Principal Component Analysis (2<sup>nd</sup> step), however, excluded some of extensometer rods from this analysis. Therefore, the number of rods Itaipu's Engineering team has automated is smaller in this step than in the 5<sup>th</sup> step.

The results of steps 6 and 8 show the reduction of the number of rods that resulted from the Principal Component Analysis was not favorable when other techniques were applied after it. This happened because of the automated extensometer rods that were excluded.

## 9. Acknowledgements

The authors would like to thank FINEP for financially supporting the research project "AIEVC - "Uncertainty Analysis and Estimation of Control Values for the Geotechnical-structural Monitoring System at the Itaipu Dam", to CAPES for the author's first scholarship and to Itaipu's Civil Engineering team for instrumentation data and technical contributions.

Communality	Rods	Communality	Rods
0.949996	<i>equip1_1</i>	0.826814	<i>equip21_2</i>
0.901233	<i>equip1_2</i>	0.486577	<i>equip31_1</i>
0.845959	<i>equip26_2</i>	0.308738	<i>equip26_1</i>

Table 9. Extensometer rods and their communalities – Cluster 1

Communality	Rods	Communality	Rods	Communality	Rods
0.958367	<i>equip3_1</i>	0.89854	<i>equip2_2</i>	0.844444	<i>equip2_1</i>
0.957442	<i>equip29_2</i>	0.897657	<i>equip14_2</i>	0.833835	<i>equip15_1</i>
0.956486	<i>equip34_3</i>	0.895226	<i>equip3_2</i>	0.830048	<i>equip19_3</i>
0.948978	<i>equip14_3</i>	0.894602	<i>equip8_2</i>	0.827687	<i>equip18_3</i>
0.943927	<i>equip33_3</i>	0.89268	<i>equip23_3</i>	0.824894	<i>equip15_2</i>
0.937446	<i>equip33_2</i>	0.887947	<i>equip25_3</i>	0.806434	<i>equip33_1</i>
0.934316	<i>equip32_3</i>	0.877563	<i>equip13_3</i>	0.77671	<i>equip12_2</i>
0.929148	<i>equip27_2</i>	0.859427	<i>equip5_2</i>	0.702823	<i>equip27_1</i>
0.928345	<i>equip5_1</i>	0.855579	<i>equip32_2</i>	0.69919	<i>equip20_3</i>
0.916552	<i>equip20_2</i>	0.852541	<i>equip8_3</i>	0.687451	<i>equip25_2</i>
0.905877	<i>equip12_1</i>	0.849741	<i>equip24_3</i>	0.636324	<i>equip32_1</i>
0.899833	<i>equip35_1</i>	0.84772	<i>equip35_2</i>	0.625665	<i>equip7_3</i>

Table 10. Extensometer rods and their communalities – Cluster 2

Communality	Rods	Communality	Rods	Communality	Rods
0.965763	<i>equip23_1</i>	0.915313	<i>equip22_3</i>	0.791565	<i>equip34_2</i>
0.949473	<i>equip23_2</i>	0.911498	<i>equip11_1</i>	0.791028	<i>equip7_1</i>
0.945071	<i>equip24_1</i>	0.899258	<i>equip14_1</i>	0.755515	<i>equip19_1</i>
0.939808	<i>equip34_1</i>	0.869832	<i>equip28_2</i>	0.733123	<i>equip7_2</i>
0.939418	<i>equip24_2</i>	0.849429	<i>equip13_1</i>	0.70484	<i>equip18_1</i>
0.93064	<i>equip19_2</i>	0.83898	<i>equip20_1</i>	0.676714	<i>equip28_1</i>
0.921589	<i>equip22_2</i>	0.833704	<i>equip25_1</i>	0.594188	<i>equip8_1</i>

Table 11. Extensometer rods and their communalities – Cluster 3

## 10. References

- Bowles, D.S., Anderson, L.R., Glover, T.F. e Chuhan, S.S. (2003) *Dam Safety Decision-Making: Combining Engineering Assessments With Risk Information*, Proc. of 2003 US Society on Dams Annual Lecture
- CBGB (1983)- *Comitê Brasileiro de Grandes Barragens*. Diretrizes para a inspeção e avaliação de segurança de barragens em operação. Rio de Janeiro-Brasil
- Dibiagio, E. (2000) *Question 78 - Monitoring of Dams and Their Foundations – General Report*. Proc. Of Twentieth Congress on Large Dams, ICOLD, 1459-1545, Beijing
- Diniz, C. A. R. e Louzarda Neto, F. (2000) *Data mining: uma introdução*. ABE. São Paulo-Brasil
- Duarte, J. M. G., Calcina, A. M. e Galván, V. R. (2006), *Instrumentação Geotécnica de Obras Hidrelétricas Brasileiras: Alguns Casos Práticos Atuais*. In: COBRAMSEG' 2006, Curitiba-Brasil

- Dyminski, A. S., Steiner, M. T. A. e Villwock, R. (2008) *Hierarchical Ordering of Extensometers Readings from Itaipu Dam*. In: First International Symposium on Life-Cycle Civil Engineering -IALCCE' 8, Varenna - Italia
- Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P. e Uthurusamy, R. (1996) *Advances in knowledge Discovery & Data Mining*. AAAI\_MIT
- FEMA - Federal Emergency Management Agency. (2004) *Federal Guidelines For Dam Safety*, U. S. Department Of Homeland Security, USA
- Freitas, A. A. (2002) *Data Mining and Knowledge Discovery with Evolutionary Algorithms*. Springer. New York
- Hair Jr, J.F., Anderson, R.E., Tatham, R.L. e Black, W.C. (2005) *Análise Multivariada de Dados (tradução)*. Bookman. São Paulo-Brasil
- Harrald, J. R.; Renda-Tanali, I.; Shaw, G.L.; Rubin, C.B.; Yeletaysi, S. (2004) *Review of Risk Based Prioritization/Decision Making Methodologies for Dams*. Technical Report, US Army Corps of Engineers
- ICOLD - International Commission on Large Dams. (2008) <http://www.icold-cigb.org>
- ITAIPIU. ITAIPIU Binacional. (2008) <http://www.itaipu.gov.br>
- Jain, A. K., Murty, M. N. e Flynn, P. J. (1999) *Data clustering: a review*. ACM Computing Surveys
- Johnson, R.A. e Wichern, D.W. (1998) *Applied Multivariate Statistical Analysis*. 4nd. Edition. Ed. Prentice Hall
- Osako, C. I.. (2002) *A Manutenção dos Drenos nas Fundações de Barragens - O Caso da Usina Hidrelétrica de Itaipu*. Dissertação de Mestrado do Programa de Pós-Graduação em Construção Civil - PPGCC, UFPR, Curitiba-Brasil
- Silva Neto, M. A., Villwock, R., Steiner, M. T. A., Dyminski, A. S. e Sccheer, S. (2008) *Mineração Visual de Dados Aplicada à Extração do Conhecimento nos Dados de Instrumentação da Barragem de Itaipu*. In: XL Simpósio Brasileiro de Pesquisa Operacional - SBPO, João Pessoa - Brasil.
- Silveira, J. F. A. (2003) *Instrumentação e Comportamento de Fundações de Barragens de Concreto*. Oficina de Textos. São Paulo-Brasil
- Silver, D.L. (1996) *Knowledge Discovery and Data Mining*. Technical Report MBA6522 CogNova Technologies London Health Science Center
- Tan, P. N., Steinbach, M. e Kumar, V. (2005) *Introduction to Data Mining*. Addison-Wesley Longman Publishing Co. Inc. Boston, MA, USA
- U.S. Army Corps of Engineers. (1987) *Instrumentation for Concrete Structures*. Engineering and Design. Engineer Manual N° No. 1110-2-4300, Washington, DC
- U.S. Army Corps of Engineers. (1995) *Instrumentation of embankment dams and levees*. Engineering and Design. Engineer Manual N° 1110-2-1908, Washington, DC
- Witten, I. H. e Frank, E. (2000) *Data Mining: Pratical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann Publishers. San Francisco, California

# A Data Mining Algorithm for Monitoring PCB Assembly Quality

Feng Zhang

*Fairchild Semiconductor South Portland, Maine 04106,  
USA*

## 1. Introduction

When surface mount technology (SMT) evolves as driven by the continuing miniaturization of electronic components and ever-growing board complexity, in-line defect inspection has become common for ensuring reliable production. For example, as an in-line measurement technique, visual defect metrology is now widely utilized in assessing process capability (Cunninggham & MacKinnon 1998; Rao et al. 1996; Barajas et al. 2003). In discrete printed circuit board (PCB) assembly, the boards within each shift are visually inspected to monitor the variation on operational conditions. Often the visual inspections are performed by automated machines, which utilize sophisticated optical and image processing techniques to detect the defects that lead to the process yield loss.

Literature study on semiconductor industry shows that over 60% of end-of-the-line defects can be traced back to solder paste printing process (Breed 1998; Venkateswaran et al. 1997). Improving the printing process performance is expected to produce reduced rework and lower cost in the downstream stages of PCB assembly by preventing small shifts and twists of components from being defects. Moreover, when components have a large number of pins such as ball grid array (BGA), it is crucial to reduce the variation between the deposits of electronic components after printing so that all joints will be soldered properly (Dempster et al. 1977).

Therefore, inspection systems built in paste printing process should not only detect the defects, but also help the operators identify the underlying root causes of poor yield resulting from inappropriate printing operations, and then develop corrective measures to avoid defective boards (Barajas et al. 2001; Litman 2004). A proper understanding of the patterns of variability among the measured solder paste profile is thus required to facilitate operators adjust the influential stencil printing parameters before a significant damage has occurred. To accommodate such quality control and yield improvement motivations, this paper proposes an effective identification method on root causes of solder paste defects by integrating statistical analysis of solder paste measurements and engineering knowledge of stencil printing process.

Very often, in semiconductor fabrication, the outputs of visual defect inspection constitute a list of binary values. That is, when hundreds of integrated circuits are assembled on a printed circuit board, the inspection machine will indicate each solder joint either good or defective. Classical statistical process control (SPC) techniques have been applied to monitor the process disturbances by charting the percentage of defects per PCB. If the total number

of visual defects exceeds a predefined control limit, the identified offending equipment should be tuned up and returned to in-control operational conditions. In PCB assembly, the optimal operational parameters for running the process in control are usually designated by operator's experience, or based on a small sample of measurements in that the cost of replicating massive quantities of PCB for inspection prevents the application of experimental design approaches (Bartholomew & Knott 1999; Gopladrishnan & Srihari 1999). A new diagnosing scheme for identifying the fault pattern present in binary inspection data is addressed in this paper, which is shown to serve as a tool to extract clustered patterns from inspected pastes on PCB and thereby identify corresponding root causes for each cluster of defects. Note that this method does not assume any prior knowledge about the nature of stencil printing faults, or any particular distribution on the size, shape or location of solder joint patterns. In short, this chapter introduces a method for routinely monitoring binary visual inspection data to detect the presence of clustered defects caused by certain assignable causes in stencil printing. As a key aspect of quality control and diagnosing, this root cause identification involves searching for systematic faults that explain the observed variability behavior by incorporating process knowledge.

## 2. The solder past printing process

A substantial proportion of the defects in PCB assembly occur in solder paste printing. For instance, insufficient solder paste volume may result in solder opens while excess solder paste volume increases the chances of bridging (O'Hara & Lee 1996). Maximizing the uniformity of solder paste profile to reduce subsequent assembly defects is then expected to improve the overall quality of PCB fabrication. On the other hand, the detection and reduction of defects in the earlier stage of SMT manufacturing such as stencil printing also diminishes the cost for other downstream stages (Pan et al. 1999). Therefore, a proper control of stencil printing has become significantly important over the years in yield management. As a major step in SMT manufacturing, stencil printing involves the allocation of adequate amount of solder paste on each component pad. In practice, various potential process factors (e.g., printer alignment, squeegee pressure, printing speed, and separation speed) may impact solder paste printing in achieving high quality. Fig. 1 schematically illustrates the stencil printing operation, where metallic stencil is first placed over a PCB and solder paste is kneaded on one side of the stencil. As shown in Fig. 1(a) and 1(b), the squeegee is pushed over the stencil under predefined pressure and moved to the other side of stencil with specific speed. This procedure makes the solder paste roll to fill the apertures in the stencil and the squeegee blade removes the excess of material, followed by the separation of the stencil from PCB at a slow snap-off speed.

In stencil printing, operation parameters should be adjusted as controllable variables by process engineers. However, such parameter adjustment relies heavily on ad-hoc algorithms or expert knowledge, because the direct printing performance evaluation given visual inspection data is not readily achievable. The lack of an analytical process monitoring mechanism comes from the difficulties in deriving a direct mathematical function between the paste defects and process parameters. Thus, a challenging problem arises on how to utilize the binary inspection information to identify the influential process factors (or systematic causes) that affect solder paste quality. When the sample of inspection data becomes available, as discussed below, a logistic regression model will characterize the

correlation of binary solder paste defects and measured physical profile (e.g., solder paste volume, height, area, etc.), the results of which are then incorporated into a latent variable framework for clustering the systematic causes to explain the variation on solder paste and consequent binary defects.

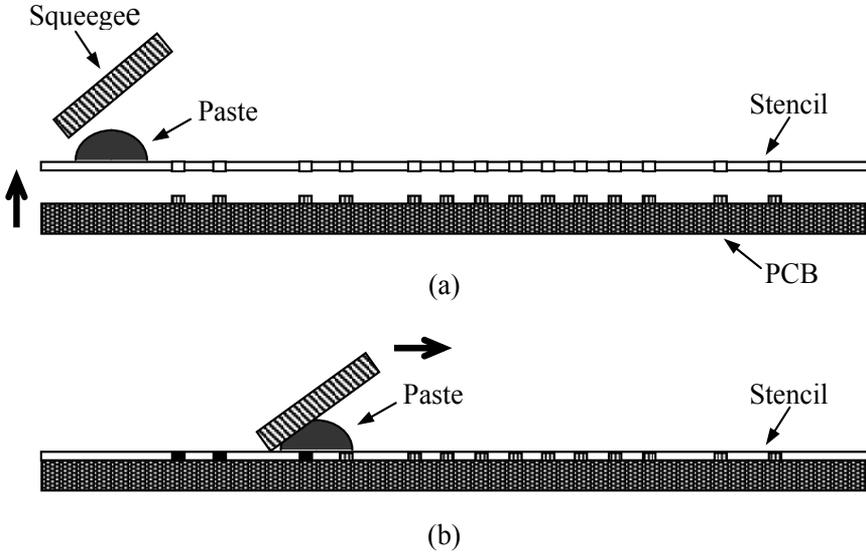


Fig. 1. The schematic illustration of stencil printing process.

### 3. Logistic regression model

As a common statistical approach for analyzing binary data, logistic regression model has been applied to various data mining and machine learning disciplines such as data classification and predicting the certainty of binary outcome (Bartholomew & Knott 1999; Jaakkola & Jordan 1997; McCulloch 1997). Under the present problem setting, for each solder paste in PCB assembly, let  $y$  denote the binary inspection such that 1 for good paste and  $-1$  for failure, and  $x$  be a  $d$ -dimensional vector representing a set of physical characteristics (called solder paste profile). The logistic regression analysis usually assumes the following quantitative relationship between  $y$  and  $x$ :

$$p\{y|x,\beta\} = \sigma(y\beta^T x) \equiv \frac{1}{1 + \exp(-y\beta^T x)}, \tag{1}$$

where  $\beta = [\beta_1 \ \beta_2 \ \dots \ \beta_d]^T$  is regression coefficient.

For a set of  $m$  measurement couples  $\{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$ , the log-likelihood of vector  $\beta$  in Equation (1) is:

$$L(\beta) = -\sum_{j=1}^m \log(1 + \exp(-y_j \beta^T x_j))$$

It is straightforward to obtain the gradient of log-likelihood function  $L(\beta)$

$$g = \nabla L(\boldsymbol{\beta}) = \sum_{j=1}^m (1 - \sigma(y_j \boldsymbol{\beta}^T \mathbf{x}_j)) y_j \mathbf{x}_j'$$

and the second-order Hessian matrix

$$\mathbf{H} = \frac{d^2 L(\boldsymbol{\beta})}{d\boldsymbol{\beta} d\boldsymbol{\beta}^T} = -\sum_{j=1}^m \sigma(\boldsymbol{\beta}^T \mathbf{x}_j) (1 - \sigma(\boldsymbol{\beta}^T \mathbf{x}_j)) \mathbf{x}_j \mathbf{x}_j^T. \quad (2)$$

For notation simplicity, the Hessian matrix  $\mathbf{H}$  in Equation (2) is often written in matrix form, i.e.,  $\mathbf{H} = -\mathbf{XAX}^T$ , where the non-zero element of diagonal matrix  $\mathbf{A}$  is

$$a_{jj} = \sigma(\boldsymbol{\beta}^T \mathbf{x}_j) (1 - \sigma(\boldsymbol{\beta}^T \mathbf{x}_j)), \quad j = 1, 2, \dots, m,$$

and  $x_j$  is the  $j$ th column of  $d \times m$  sample matrix  $\mathbf{X} = [x_1 \ x_2 \ \dots \ x_m]$ .

Newton optimization algorithm works as an efficient way to estimate the  $d \times 1$  regression coefficient vector by maximizing  $L(\boldsymbol{\beta})$  through the second derivatives (2), which provides the following iterative calculation (3) to estimate  $\boldsymbol{\beta}$ :

$$\begin{aligned} \boldsymbol{\beta}_{new} &= \boldsymbol{\beta}_{old} + (\mathbf{XAX}^T)^{-1} \sum_{j=1}^m (1 - \sigma(y_j \boldsymbol{\beta}^T \mathbf{x}_j)) y_j \mathbf{x}_j \\ &= (\mathbf{XAX}^T)^{-1} \mathbf{XA} (\mathbf{X}^T \boldsymbol{\beta}_{old} + \left[ \frac{y_1}{\sigma(y_1 \boldsymbol{\beta}_{old}^T \mathbf{x}_1)} \quad \dots \quad \frac{y_m}{\sigma(y_m \boldsymbol{\beta}_{old}^T \mathbf{x}_m)} \right]^T). \end{aligned} \quad (3)$$

The maximum-likelihood (ML) estimation of  $\boldsymbol{\beta}$  is also called iterative re-weighted least squares (IRLS) algorithm, where the computation complexity within each iteration is  $O(md^2)$ .

As discussed in previous research work, the logistic regression model (1) has been used mostly to understand the role of input variables  $x$  in predicting the binary response variable  $y$ . In manufacturing practice, however, many of the measurement variables in  $x$  are correlated due to some common physical phenomena, which encourage us to seek a parsimonious form of the input variables to summarize their effects on binary outcomes. In other words, the effects on the measured physical profile can be explained by a reduced set of latent variables without loss of statistical information, as described in the next section. Thus, we would refit the regression model (1) with fewer latent variables to provide an interpretation of their influence on binary outputs as observed in defect inspection. This statistical interpretation, equipped with proper pattern clustering and visualization, is shown to enhance the diagnosing of solder paste quality.

## 4. MLPCA based pattern clustering algorithm

### 4.1 Latent variable model and MLPCA

When correlations are present among the measured variables  $x$  for a product, this implies the existence of common systematic causes that govern such interrelated manners. Therefore, multivariate statistical techniques such as PCA have been proposed to investigate the correlations when multiple variables are involved (Crida et al. 1997). A latent variable model is introduced to relate  $d$  characteristics of solder pastes to  $p$  unknown systematic causes  $v$ , by assuming that  $v$  affects the solder paste profile  $x$  through a linear model, i.e.,

$$x = Cv + w, \tag{4}$$

where  $C = [c_1, c_2, \dots, c_p]$  is a  $d \times p$  constant matrix with full rank, and  $v = [v_1, v_2, \dots, v_p]^T$  is a  $p \times 1$  zero-mean random vector with independent components, each scaled without loss of generality to have unit variance.

As assumed in PCA, the latent variables are of smaller dimension (i.e.,  $p < d$ ) so that the dependencies among observed data  $x$  can be described by a reduced set of variables  $v$ . Noise  $w$  denotes the aggregated effects that are not due to any systematic causes, which is assumed to be white noise, i.e.,  $w \sim N(0, \sigma_w^2 \mathbf{I})$ , and independent of  $v$ . It is reasonable to assume that each root cause is associated with distinct physical dynamics so that the latent variables  $v$  can be represented by normalized independent Gaussians, that is,  $v \sim N(0, \mathbf{I})$ . As such, the impacts on measured solder joint profile  $x$  from  $v$  are quantified by the magnitude of corresponding rows in matrix  $C$ . Equipped with prior distributions over  $v$  and  $w$ , model (4) now provides a parsimonious probabilistic description for multivariate measurement data  $x$  (Hamada & Nelder 1997; Tipping & Bishop 1999). Moreover, the probabilistic assumptions enable an ML estimate for  $C$  (denoted by CML) that is shown to span the principal subspaces of  $x$  (Tipping & Bishop 1999).

For isotropic Gaussian noise  $w$ , model (4) yields the conditional probability of  $x$  as:

$$p(x|v) = (2\pi\sigma_w^2)^{-d/2} \exp\left\{-\frac{1}{2\sigma_w^2} \|x - Cv\|^2\right\}. \tag{5}$$

The Gaussian assumption on  $v$  implies that the marginal density of data  $x$  can be readily obtained by integrating out  $v$  so that  $x \sim N(0, \Sigma)$ , and covariance  $\Sigma = \sigma_w^2 \mathbf{I} + C C^T$ .

For a sample of  $\{x_j: j = 1, 2, \dots, m\}$  from model (4) and (5), the log-likelihood is

$$L = \sum_{j=1}^m \log(p(x_j)) = -\frac{m}{2} \{d \log(2\pi) + \ln |\Sigma| + \text{tr}\{\Sigma^{-1} S\}\}, \tag{6}$$

where  $S$  is the sample covariance matrix. The estimate of  $C$  that maximizes the log-likelihood (6) is shown to satisfy (Tipping and Bishop 1999):

$$C_{ML} = U_p (\Lambda_p - \sigma_w^2 \mathbf{I})^{1/2} R. \tag{7}$$

The interpretation of Equation (7) is that the maximum of log-likelihood is achieved when the column vectors of  $d \times p$  matrix  $U_p$  are eigenvectors of  $S$  corresponding to the  $p$  largest eigenvalues. The eigenvalues  $\lambda_i$  are stored in descending order within matrix  $\Lambda_p = \text{Diag}\{\lambda_k\}$  ( $k = 1, 2, \dots, p$ ). The column vectors in  $U_p$  are also called principal eigenvectors due to their relationship with respect to the eigenvectors, and  $R$  is a  $p \times p$  orthogonal matrix. Furthermore, the ML estimate of  $\sigma_w^2$  is given by

$$\sigma_{ML}^2 = \frac{1}{d-p} \sum_{k=p+1}^d \lambda_k,$$

in which noise variance is viewed as the average of the  $d-p$  smallest eigenvalues.

The maximum-likelihood estimate of  $C$  in Equation (7) can be calculated by an iterative expectation-maximization (EM) algorithm between the following equations (Booth & Hobert 1999; Dempster et al. 1977):

$$\mathbf{C}_{new} = \mathbf{S}\mathbf{C}_{old} (\sigma_{w,old}^2 \mathbf{I} + \mathbf{M}^{-1} + \mathbf{C}_{old}^T \mathbf{S}\mathbf{C}_{old})^{-1}, \quad (8)$$

$$\sigma_{w,new}^2 = \frac{1}{d} \text{tr}(\mathbf{S} - \mathbf{S}\mathbf{C}_{new} \mathbf{M}^{-1} \mathbf{C}_{new}^T), \quad (9)$$

where  $\mathbf{M} = (\sigma_w^2 \mathbf{I} + \mathbf{C}^T \mathbf{C})$ . Thus, the optimal  $\mathbf{C}$  and noise variance  $\sigma_w^2 \mathbf{I}$  are obtained when Equations (8) and (9) converge. Note that the rotation matrix  $\mathbf{R}$  brings somewhat ambiguity in the ML estimation for matrix  $\mathbf{C}$ . In the proposed method, this ambiguity can be resolved by determining the rotation matrix from  $\mathbf{C}_{ML}^T \mathbf{C}_{ML} = \mathbf{R}^T (\mathbf{\Lambda}_p - \sigma_w^2 \mathbf{I}) \mathbf{R}$ , i.e.,  $\mathbf{R}$  is the eigenmatrix of  $\mathbf{C}_{ML}^T \mathbf{C}_{ML}$ . As implied in Equation (7), latent variable model (4) effects a mapping from the latent space into the principal subspace of multivariate data  $\mathbf{x}$ . In this sense the ML estimate  $\mathbf{C}_{ML}$  for model (4) is indeed a form of principal component analysis. Therefore, we choose to term the proposed method as maximum-likelihood PCA.

One major advantage of latent variable model and corresponding MLPCA estimate is to offer an effective way to link the variability analysis on solder paste profile and subsequent binary inspections to a candidate set of process faults. Suppose that multivariate measurements  $\mathbf{x}$  on solder pastes are correlated due to common unobservable process factors  $\mathbf{v}$ , this paper tries to provide an analytical tool for diagnosing product quality by relating variation pattern on physical characteristics to these hypothesized systematic causes. As demonstrated in later case study, this method is developed on a process-oriented basis, which applies MLPCA to determine the latent space of systematic root causes and then project logistic regression coefficients onto this reduced space for pattern clustering and interpretation. The visualization of clustered variation pattern, combined with appropriate engineering knowledge, will help identify the underlying process faults. On the other hand, classical PCA is a data-oriented approach that tries to explain the variance of  $\mathbf{x}$  by seeking the principal eigenvectors. PCA works well for situations when a single process fault occurs (i.e.,  $p = 1$ ), but can not produce interpretable results for process diagnosing when  $p > 1$  (Apley & Shi 2001). The limitations of PCA on root cause recognition or fault interpretation thus hamper its diagnostic capabilities in complicated multivariate process control.

The latent variable model (4) also considers the effects from measurement noise on solder paste, which has been a non-neglectable factor when accurate process modeling and diagnosing are required. The probabilistic formulation enables the introduction of likelihood measure for obtaining ML estimate CML. It is worth noting that CML is built on the assumption that  $p$  is known. However, the probabilistic model itself does not provide a mechanism to determine  $p$ . For practical implementation, we need to address how to define the dimension of latent variable  $\mathbf{v}$  prior to parameter estimation. For  $p = d-1$ , the model is equivalent to a full covariant Gaussian distribution, while in case of  $p < d-1$  it implies that the remaining  $d-p$  directions is caused by noise variance  $\sigma_w^2$ . As a possible approach, cross-validation may compare all potential values of  $p$ , however, it becomes expensive in computation when  $d$  increases. Simulation results over numerous examples with varying  $p$  and  $d$ , suggest the following practical rule to determine  $p$  and substitute it into the iterative EM algorithm:

$$p = \inf_l \left\{ \frac{\frac{1}{l+1} \sum_{k=1}^{l+1} \lambda_k - \frac{1}{l} \sum_{k=1}^l \lambda_k}{\frac{1}{l} \sum_{k=1}^l \lambda_k - \frac{1}{l-1} \sum_{k=1}^{l-1} \lambda_k} \gg 1 \right\}. \quad (10)$$

**4.2 The regression coefficient clustering algorithm**

The dramatic advances in in-process sensor and data collection technologies enable vast quantities of physical features to be measured about the manufacturing system. For instance, in PCB assembly, laser-optical measurement machines are commonly installed to record detailed dimensional characteristics of wet solder paste after it is deposited onto the board in stencil printing. When electronic components are positioned and the solder is cured in the re-flow oven, dimensional characteristics are obtained via X-ray laminography (Crida et al. 1997; Litman 2004; Neubauer 1997). As in any quality control applications, one fundamental objective considered in this paper is to explain as precisely as possible the nature of variation on solder paste and identify the root causes of binary defects by utilizing the earlier measured physical information.

Although the aforementioned logistic regression method can estimate coefficients for each measurement variable, the high dimensionality of solder profile makes it not efficient for engineers to explore the nature of how the underlying process factors cause the defective outputs. On the other hand, as shown in Fig. 2, the latent variable model helps recognize the patterns of solder paste variation and thereby identify the corresponding systematic causes during stencil printing operation. By integrating the logistic regression method with latent variable model (4), the proposed methodology will quantify the effects on solder profile  $x$  and defects  $y$  from process faults  $v$ , which is performed entirely on the collected sample data with no a priori knowledge about the patterns of variation. Therefore, a core component of this approach includes the proper clustering over regression coefficients with respect to variables  $v$ , which provides more intuitive insight into the interdependencies among multiple measurement variables.

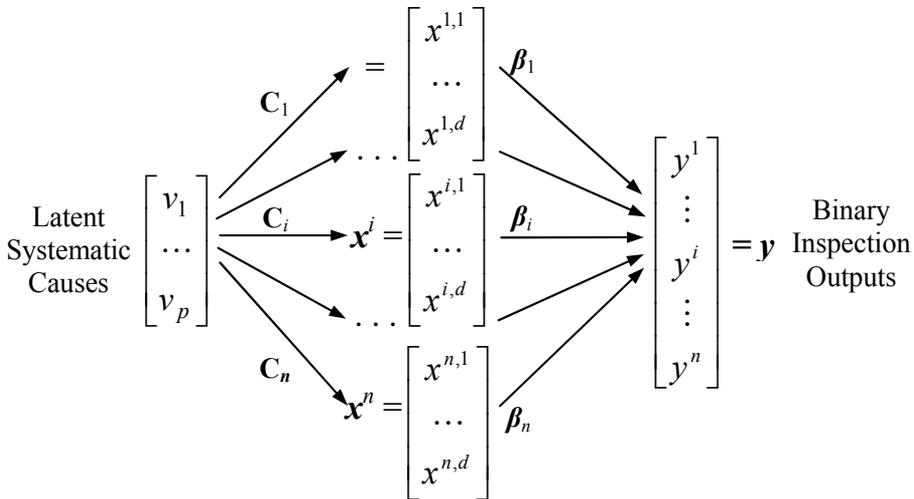


Fig. 2. Illustration of latent variable model that explains the relationship between systematic cause  $v$ , solder paste profile  $x$ , and final defect inspection output  $y$ .

Following the assumptions on model (4), let  $x = [x_1, x_2 \dots x_d]^T$  represent the measurable characteristics of a solder paste, and  $\{x_i; i = 1, 2, \dots, n\}$  be a set of  $n$  solder pastes in the board. Fig. 2 implies that  $p$  independent causes  $v_j$  apply their joint effects on the variation of physical profile  $x_i$  through a constant matrix  $C_i$ , and produce the consequent binary outputs

$y_i$  through logistic regression coefficient  $\beta_i$ . In particular, the effect from cause  $v_j$  is represented by the  $j^{\text{th}}$  column vector  $c_{i,j}$  in  $C_i$ . Since each  $v_j$  is scaled to have unit variance,  $c_{i,j}$  indicates the magnitude or severity of variation caused by  $v_j$ . After clustering a sample of solder pastes based on the distribution of their regression coefficients in terms of  $v_j$ , quality diagnosing of stencil printing becomes possible by assigning a process fault to the solder pastes within the same group.

Prior to the clustering analysis over regression coefficients, Equation (4) is substituted into the logistic regression model, yielding new coefficients  $\beta_{v,i} = C_i^T \beta_i = [\beta_{v,i,1} \ \beta_{v,i,2} \ \dots \ \beta_{v,i,p}]^T$  for latent variable  $v$ . Now binary data  $y_i$  can be explained by systematic causes  $v_j$ , which takes the form of a logit function, that is,

$$\text{logit}\{y_i = 1 \mid v, \beta_{v,i}\} = \sum_{j=1}^p \beta_{v,i,j} v_j + \beta_i^T w \equiv v^T \beta_{v,i} + \varepsilon_w$$

where  $\varepsilon_w$  denotes the transformed noise effect. The new coefficients  $\beta_{v,i,j}$  correspond to the change in the log odds per unit change in  $v_j$  when  $v_j$  does not interact with other sources (this is reasonable given the latent variable model assumptions). Or, the effect of increasing  $v_j$  by 1 is to increase the odds that  $y_i = 1$  by a factor  $\exp(\beta_{v,i,j})$ .

Since  $\beta_{v,i}$  depends on the systematic causes  $v$ , the regression coefficients can be classified so that each cluster describes the similar pattern of solder paste variation. In other words, the proposed clustering method is used to separate the impacts from cause  $v$ . Once all inspected solder pastes on a PCB are clustered in terms of  $\beta_{v,i}$ , process diagnosis for variation reduction can be performed since each cluster is mapped to a specific process fault or assignable cause.

### 4.3 MLPCA based clustering algorithm for quality diagnosing

As a statistical tool for diagnosing the quality of solder pastes, the proposed MLPCA based regression coefficients clustering algorithm is now summarized as follows:

*Step 1.* Apply logistic regression model (1) to binary inspection data collected from  $m$  PCBs, yielding the estimates of coefficients  $\beta_i$  for the  $i^{\text{th}}$  solder paste through sample  $\{y_j^i : i = 1, 2, \dots, n; j = 1, 2, \dots, m\}$ .

*Step 2.* Given the set of measured solder paste profile  $x_j^i$ , determine the dimension  $p$  via rule (10) and estimate the matrix  $C_i$  in model (4) by MLPCA method.

*Step 3.* Calculate new regression coefficients  $\beta_{v,i} = C_i^T \beta_i$ , followed by a  $k$ -means clustering algorithm (Hastie et al. 2001) over  $\beta_{v,i}$  to recognize the coefficient clusters.

*Step 4.* Present the geometrical clustering results on the board to process operators to identify the process faults by utilizing their engineering knowledge. The diagnosing results of solder paste quality will then lead to appropriate stencil printing operation adjustments.

By taking advantage of the diagnostic information from latent variable model and logistic regression coefficients, the MLPCA based clustering algorithm provides a visual way to relate stencil printing process problems to the variation on solder paste profile and consequent binary defects. Case study in the following section shows that the proposed method is favorable in improving process quality by developing a more interpretable relationship between variation pattern and physical faults.

## 5. Application in PCB assembly

In stencil printing process, each solder paste is deposited on the board automatically by printing machines, then registered with the screen and printed. Stencil printing is known to be an established technology, however, there are some uncontrolled factors that influence the quality of solder pastes (Lathrop 1997; Liu et al. 2001), and hence cause component failures in PCB assembly. In order to produce pastes with minimal variation on physical profile, the controllable parameters for printing operation should be monitored and adjusted by appropriate diagnosing of solder paste quality. In the present study, solder paste printability was denoted by a physical profile collected from laser triangulation and X-ray based measurement machine. The purpose of the present experimental research is to identify the systematic factors in solder deposition process by quantifying their impacts on paste quality. The set of process factors include printer steel squeegee angle, printing direction, and squeegee speed, etc.

The variation on measured solder paste profile that leads to binary inspection results stems from improper parameter settings of stencil printing, called systematic factors. Their effects on solder paste (such as solder paste volume, area, and height, etc.) are present in multivariate profile  $x$ . Due to the common factors  $v$ , variables in  $x$  are always highly correlated, as shown in the scatter plots in Fig. 3. For a specific solder paste, the plots were drawn over pairs of distinct physical solder paste features from a sample of  $m = 30$  inspected boards. In semiconductor fabrication, the solder paste profile often includes paste thickness, paste volume, shape of heel fillet, shape of toe or center fillet, alignment between pad and lead, pad average width, pad average height, and pad volume, etc. For purpose of illustration, we chose ten physical features as element variables of vector  $x$ , that is,  $d = 10$ .

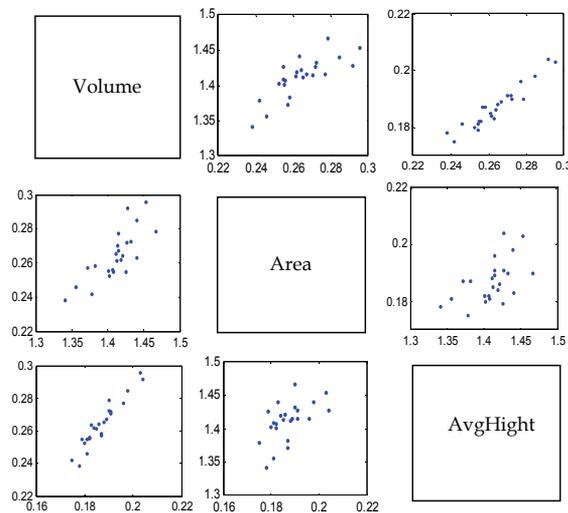


Fig. 3. Scatter plot of selected pairs of measured variables in solder paste profile (e.g., paste volume, area, and average height).

Next, the coefficients clustering algorithm proposed in Section 4 was applied to map the pattern of solder paste defects on PCB to the latent systematic causes, given the assumption that variation of solder paste profile was not completely random due to the measurement noise. To accommodate pattern clustering and visualization, the present experimental study

was undertaken over a region of PCB that consists of more than 3000 solder joints (e.g.,  $n = 3012$ ), as shown in Fig. 4. Given the sample of binary inspection  $y_i$  and corresponding physical profile  $x_i$  ( $i = 1, 2, \dots, n$ ), we first calculated the estimation of coefficients  $\beta_i$  by logistic regression model (1). MLPCA was then applied to estimate variation pattern matrix  $C_i$ , in which the dimension of systematic cause  $v_i$  was always determined as two by the rule (7) (i.e.,  $p = 2$  for all  $i$ ). As indicated in the algorithm summary, after projecting original  $\beta_i$  onto the latent space spanned by  $C_i$ , the new coefficients  $\beta_{v,i}$  became available for  $k$ -clustering algorithm (Hastie et al. 2001), which classify them into two clusters.

The graphical illustration of clustered  $\beta_{v,i}$  in Fig. 4 also validate the presence of two clusters as identified by the standard  $k$ -means algorithm. Since each solder paste is positioned on PCB by the unique X-Y coordinates, we can visualize the clustered coefficients on a printed circuit board such that the pastes in each cluster are denoted by the same symbol (e.g., “+” for cluster 1 and “×” for cluster 2).

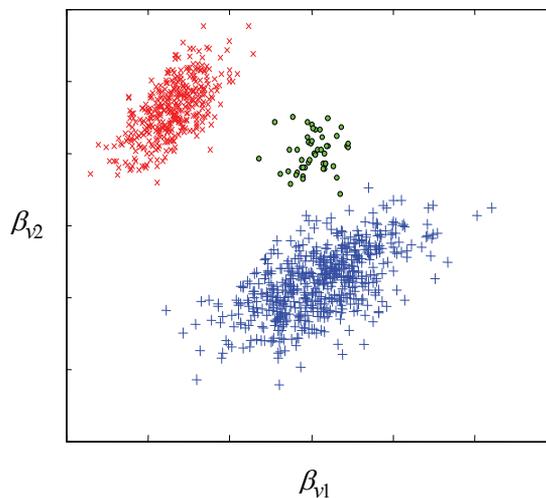


Fig. 4. Scatter plot of their logistic regression coefficients  $\beta_v = [\beta_{v1} \beta_{v2}]^T$ .

MLPCA implied that there were 2 systematic causes that governed the variation over the measured solder paste profile. The pastes denoted by ‘+’ in Fig. 4, for example, were dominantly affected by the first systematic cause  $v_1$ , which almost lie on the horizontal direction of the board, while the pastes denoted by ‘×’ were distributed along the vertical direction and influenced mainly by the second cause  $v_2$ . The graphical demonstration of clustered coefficient results in Fig. 4 thus helped process engineers to adopt their expert knowledge and experiences in diagnosing the solder paste defects. For instance, the solder pastes denoted by “+” in cluster 1 had relatively large coefficients  $\beta_{v,i,1}$  and were mostly distributed along the length of PCB. That is, the systematic cause corresponding to this cluster should influence the solder pastes along the horizontal direction to a greater extent during stencil printing. Intensive discussions with process engineers have provided a reasonable explanation for the causes to be inappropriate parameter settings in controlling the stencil printing speed and printing pressure. These process factors are expected to generate large variation of solder paste profile along the horizontal and vertical direction, respectively, and correspondingly more inspected defects.

Further investigation on other potential process faults implied that the above identified systematic causes are most likely to produce the consistent results. The diagnostic results also agreed with the natural speculation on stencil printing diagram in Fig. 1, where printing speed usually influences the solder pasts along the length of PCB, while printing pressure has greater impacts on solder paste quality than other process factors (e.g., separation distance, printer alignment) along the width of PCB. In addition, detailed inspections revealed that a substantial portion of the quality deficiencies (such as slumping, bridging and bleeding of paste underneath the stencil) along the width of PCB was caused by abnormal high printing pressure during stencil printing.

The case study shows that the systematic pattern on PCB assembly defects is often owing to specific process faults such as inappropriate process operations, rather than completely random due to the environmental or measurement noise. The proposed coefficients clustering algorithm provides an effective process-oriented diagnosis tool for identifying such production irregularities. By assuming the potential systematic causes are mapped to the clusters of solder pastes with similar coefficients, the variation on solder paste profile and corresponding binary defects can be mapped to improper parametric control that deviates from optimal conditions, which will suggest informative corrections to adjust the stencil printing to improve process quality.

## 6. Conclusion

The distillation of massive quantities of solder paste inspection data into relevant quality information allows rapid understanding of the low production yield in PCB assembly. The statistical diagnosis method proposed in this paper provides more meaningful insights into the defect mechanisms than traditional yield analysis methods, which can identify the assignable causes of defects and their effects on yield by integrating MLPCA and logistic regression model. This offers a systematic representation on the impacts of process condition changes to the variation of solder paste profile. The probabilistic latent variable model allows ML estimation to determine the latent space by iteratively maximizing the likelihood function. In contrast to standard PCA, this approach is also efficient for multivariate process analysis when some sample data are missing. The clustering algorithm over the projected regression coefficients onto the latent space is relatively easy to implement with affordable computational effort. Experimental study demonstrates that the statistical interpretation of solder defect distributions can be enhanced by intuitive pattern visualization for process fault identification and variation reduction.

## 7. References

- Apley, D. & Shi, J. (2001) A Factor-Analysis Method for Diagnosing Variability in Multivariate Manufacturing Processes. *Technometrics*, Vol. 43 (1), pp. 84-95.
- Barajas, L. G. ; Kamen, E.W. & Goldstein, A. (2001) On-line Enhancement of the Stencil Printing Process. *Circuits Assembly*, March, pp. 32-36.
- Barajas, L.G. et al. (2003) Process Control in a High-noise Environment with Limited Number of Measurements. *Proceedings of American Control Conference*, Vol. 1, pp. 597-602, Denver, June 2003.
- Bartholomew, D.J. & Knott, M. (1998). *Latent Variable Models and Factor Analysis*. Oxford University Press, London.

- Booth, J. G. & Hobert, J. P. (1999). Maximizing Generalized Linear Mixed Model Likelihoods with An Automated Monte Carlo EM Algorithm. *Journal of the Royal Statistical Society, Series B*, Vol. 61, pp. 265-285.
- Breed, S. (1998). Advances in Intelligent Stencil Printing. *Proceedings of NEPCON West 98*, Anaheim, California, pp. 253-256, 1998.
- Crida, R. C. ; Stoddart, A. J. & Illingworth, J. (1997). Using PCA to Model Shape for Process Control. *International Conference on Recent Advances in 3-D Digital Imaging and Modeling*, pp. 318-325, Ottawa, Canada, 1997.
- Cunningham, S. P. & MacKinnon, S. (1998). Statistical Methods for Visual Defect Metrology. *IEEE Transactions on Semiconductor Manufacturing*, Vol. 11(1), pp. 48 -53.
- Dempster, A. P. ; Laird, N. M. & Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm (with Discussion). *Journal of the Royal Statistical Society, Series B*, Vol. 39, pp. 1-38.
- Gopaladrishnan, L. & Srihari, K. (1999). Process Development for Ball Grid Array Assembly using a Design of Experiments Approach. *Journal of Advanced Manufacturing Technology*, Vol. 15, pp. 587-596.
- Hamada, M. & Nelder, J. A. (1997). Generalized Linear Models for Quality-improvement Experiences. *Journal of Quality Technology*, Vol. 29, pp. 292-304.
- O'Hara, W. & Lee, N.C. (1996). How Voids Develop in BGA Solder Joints. *Surface Mount Technology*, pp. 44-47, January 1996.
- Hastie, T. ; Tibshirani, R. & Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.
- Jaakkola, T. S. & Jordan, M. I. (1997). A Variational Approach to Bayesian Logistic Regression Models and Their Extensions. *Proceedings of the Sixth International Workshop on Artificial Intelligence and Statistics*, Fort Lauderdale, Florida, 1997.
- Lathrop, R. R. (1997). Solder Paste Print Qualification using Laser Triangulation. *IEEE Transactions on Components and Packaging Manufacturing Technologies*, Vol. 20, pp. 174-182.
- Litman, E. (2001). Solder Paste Printing: An Inside Look. *Surface Mount Technology*, pp. 30-34, January 2004.
- Liu, S. et al. (2001). A Novel Approach for Flip Chip Solder Joint Quality Inspection: Laser Ultrasound and Interferometric System. *IEEE Transactions on Components and Packaging Technologies*, Vol. 24(4), pp. 616-624.
- McCulloch, C. E. (1997). Maximum Likelihood Algorithm for Generalized Linear Mixed Models. *Journal of the American Statistical Association*, Vol. 92, pp. 162-170.
- Neubauer, C. (1997). Intelligent X-Ray Inspection for Quality Control of Solder Joints. *IEEE Transactions on Components, Packaging and Manufacturing Technology-Part C*, Vol. 20(2), pp. 111-120.
- Pan, J. et al. (1999). Critical Variables of Solder Paste Stencil Printing for Micro-BGA and Fine Pitch QFP. *IEEE/CPMT International Electronics Manufacturing Technology Symposium*, Austin, October 1999.
- Rao, S. et al. (1996). Monitoring Multistage Integrated Circuit Fabrication Processes. *IEEE Transactions on Semiconductor Manufacturing*. Vol. 9(4), pp.495-505.
- Tipping, M. E. & Bishop, C.M. (1999). Mixtures of Probabilistic Principal Component Analyzers. *Neural Computation*, Vol. 11(2), pp. 443-482.
- Venkateswaran, S. et al. (1997). A Realtime Process Control System for Solder Paste Stencil Printing. *Proceedings of International Electronics Manufacturing Technology Symposium*, pp. 62-67, Austin, October 1997.

# An Overview of Data Mining Techniques Applied to Power Systems

Jefferson Morais, Yomara Pires, Claudomir Cardoso and Aldebaro Klautau  
*Signal Processing Laboratory (LaPS) Federal University of Pará (UFPA)*  
*Belém PA Brazil*

## 1. Introduction

The growth of available data in the electric power industry motivates the adoption of data mining techniques. However, the companies in this area still face several difficulties to benefit from data mining. One of the reasons is that mining power systems data is an interdisciplinary task. Typically, electrical and computer engineers (or scientists) need to work together in order to achieve breakthroughs, interfacing power systems and data mining at a mature level of cooperation. Another reason is the lack of freely available and standardized benchmarks. Because of that, most previous research in this area used proprietary datasets, which makes difficult to compare algorithms and reproduce results.

This chapter has two main goals and, consequently, is divided in two parts. In the first part, the goal is to present a brief overview on how data mining techniques have been used in power systems. There are several works, such as (Mori, 2002), that introduce data mining techniques to people with background in power systems. In contrast, this text assumes previous knowledge of data mining, describes some fundamental concepts of power systems and illustrates the kind of problems that the electric industry tries to solve with data mining.

The second part of the work presents a thorough investigation of a specific problem: classifying time series that represent short-circuit faults in transmission lines. Studies show that these faults are responsible for 70% of the disturbances and cascading blackouts (Kezunovic & Zhang, 2007). Besides, there is a large and growing number of publications about this problem.

Two types of fault classification systems are discussed: *on-line* and *post-fault*. On-line fault classification must be performed on a very short time span, with the analysis segment (or frame) being located approximately at the instant the fault begins. Post-fault classification can be performed off-line and its input consists of a multivariate time series with variable length (duration). Post-fault is a sequence classification problem, while in on-line classification the input is a fixed-length vector. Both fault classification systems (and most data mining applications) require a preprocessing or front end stage that converts the raw data into sensible parameters to feed the back end (in this case, the classifier).

Besides its practical importance, one reason for the popularity of fault classification is that it is relatively easy to artificially generate a dataset through simulators. Here, the well-known Alternative Transients Program (ATP) (ATP, 1987) simulator was used to create a public and

comprehensive labeled dataset. Such datasets are key components to allow reproducing research results in different sites, which is crucial given the large number of parameters to be tuned in a fault classification system. Therefore, in order to promote reproducible research, this work also provides detailed information about the adopted parameters.

The chapter is organized as follows. In Section 2 a brief description of data mining applications in power systems is provided. The section also introduces basic concepts of power systems. For a more detailed treatment, the reader is referred to (Casazza & Delea, 2005). The second part begins in Section 3, which poses the fault classification problem and discusses solutions. Section 4 presents simulation results and is followed by the conclusions.

## 2. Power systems for data miners

### 2.1. Concepts of interest

A typical power system is represented in Figure 1 and can be divided into three parts: generation, transmission and distribution. The distribution system delivers power to the end users (*loads*). Most systems adopt three *phases* (A, B and C), using three conductors to carry sinusoidal voltage waveforms that have an offset in time equivalent to 120 degrees. While the customers need low voltage values (hundreds of Volts), the transmission system typically uses much higher values for efficiency. The transformers are responsible for the up and down conversions of voltage values and are located in different parts of the system.

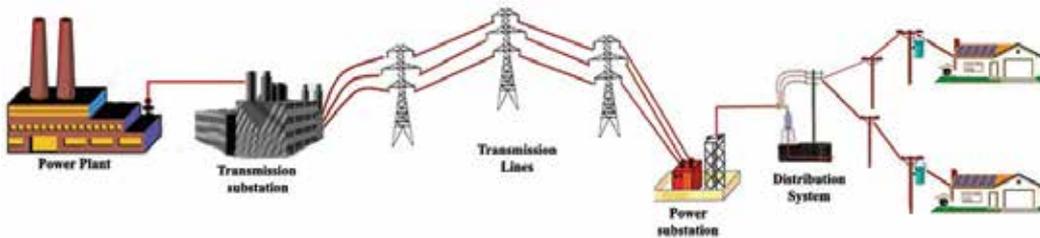


Fig. 1. An example of an electric power system.

Under normal conditions, the voltage waveform  $x(t)$  has a pre-established frequency (e.g., 60 or 50 Hz). Knowing the nominal value of the amplitude (e.g., 500 kV), it is convenient to normalize  $x(t)$  by this value and report the amplitude in p.u. (per unity). In this case, the ideal waveform could be expressed by  $x(t) = \cos(2\pi ft + \theta)$ , where  $f$  is the frequency and  $\theta$  the initial angle. Figure 2(a) illustrates a segment of the ideal voltage waveform for a frequency  $f = 60$  Hz and  $\theta = 0$  radians. Figure 2(b) depicts simultaneously all three voltage waveforms in a segment containing a fault recorded by an *oscillograph* recording equipment: a short circuit between the conductor corresponding to phase B and ground (G). It can be seen that, besides phase B, the other two phases are also disturbed. Such faults belong to a category of events that is called *transients* because they tend to disappear after proper operation of the system to recover normal conditions, as occurs after approximately 0.05 seconds in Figure 2(b).

In order to properly operate it, a power system contains several data acquisition equipments. For example, some of these equipments register the status of logical (boolean) variables at each minute, while others store waveforms digitized at relatively high sampling

rates (e.g., 5 kHz). In many cases it suffices to monitor the root mean-square (RMS) value of each waveform estimated at each second. Figure 3 illustrates the information about the voltage amplitude that is provided by the RMS estimation.

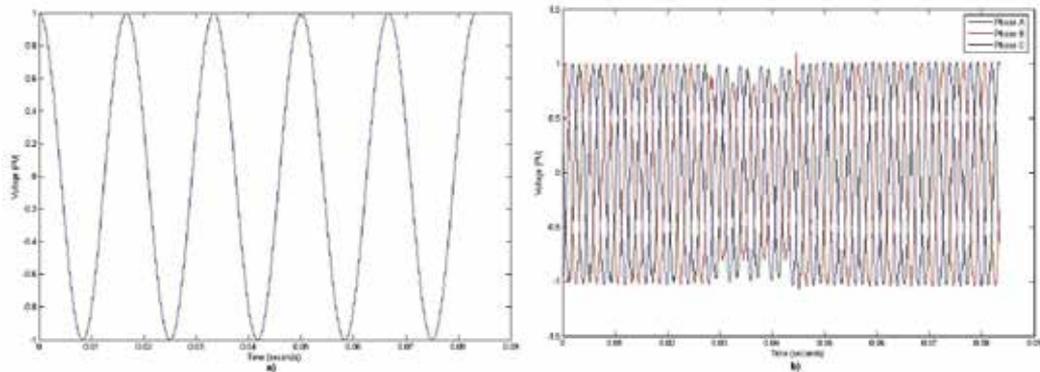


Fig. 2. a) Example of ideal normalized voltage waveform of one phase. b) All three phases with short-circuit between phase B and ground, as registered by an oscillograph equipment.

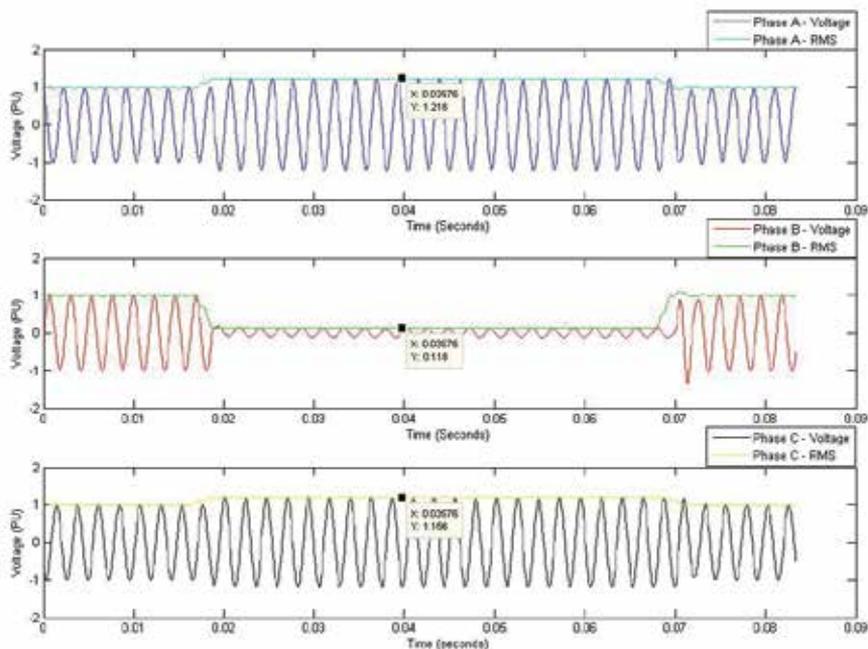


Fig. 3. Example of the RMS waveforms superimposed to the voltage waveforms in a fault between phase B and ground. It can be seen that the voltage in phase B drops to a value over than 1 p.u., while phases A and C achieve values above the nominal.

In summary, most power systems data can be considered as multivariate time series, but the sampling frequencies may differ significantly and the variables are eventually categorical (not numerical). Roughly, one can organize the data originated in power systems into three categories:

- a. Raw waveforms (voltages and currents) sampled at relatively high sampling frequencies;
- b. Pre-processed waveforms (e.g., RMS) typically sampled at low sampling frequencies;
- c. Status variables (e.g., if a relay is opened or closed) typically sampled at low sampling frequencies.

Due to its higher volume of information, the first category data is sometimes organized in specific databases, such as the oscillographic database, which stores all data from oscillographs. The other two categories are sometimes organized in the so-called Supervisory Control And Data Acquisition (SCADA) systems (Boyer, 1999). SCADAs are complex systems that periodically store several thousands of heterogeneous variables and are an important source of information for data mining. For example, automatically mining cause-effect relationships in SCADA data is an incipient but promising activity. Some power systems are affected by events that repeatedly cause troubles but their causes remain undetected. However, most of the times it is necessary to organize a data warehouse in order to be able to mine data from a SCADA, and fewer works use such data when compared to the first category.

Data mining can also alleviate another problem in power systems: when a disturbance is detected, a large amount of messages and alarms are generated. Protection equipments are responsible for detecting a problem, and act appropriately, isolating the defective part of system, for example. Part of this operation is automatic, but some tasks depend on a specialist. The amount of information regarding the problem cannot be excessive but should be enough for making decisions. Data mining techniques can be used to filter alarms and messages and provide the important information to the operator.

Failures in the performance of protection equipments, remote terminal, communications link and acquisition of data online, and variations in the voltage levels after of the occurrence disturbance, are factors that difficult the assessment and diagnosis in real time initial cause of power off.

Another problem of interest to the electric power industry is load forecasting, in which the goal is to predict the demand for power in specific regions. This can be cast as a conventional regression problem. Power quality is another area that can benefit from data mining. Here the goal is to help characterizing how close to the ideal (nominal) parameter values the system is operating. Small and large deviations are categorized by detection and classification modules. The location where a power quality event happened is also of interest.

The next section describes tasks and techniques used in data mining.

## **2.2 Review of data mining applications in power systems**

This section briefly describes typical applications of data mining in electrical power systems via a collection of 18 papers. Table 1 lists the technique, task and application area.

Among the several applications listed in Table 1, the second part of this work concentrates in fault classification, as discussed in the next section.

Reference	Technique	Task	Application Area	Problem
Ramos, 2008	Decision tree	Classification	Distribution system	Characterization and classification of consumers
Hagh, 2007	Neural network	Classification	Transmission lines	Faults classification and locations
Saibal, 2008	WN <sup>1</sup>	Classification	Distribution system	Classification of transients
Chia-Hun & Chia-Hao 2006	Adaptative wavelet networks	Detection and discrimination	14-bus power system	Power-quality detection for power system disturbances
Pang & Ding 2008	wavelet transform and self-organizing learning array	Power quality disturbances classification	Distributed power system	Power-quality detection for power system disturbances
Bhende, 2008	Neural network	Classification	Not defined	Detection and classification of Power quality disturbances
Figueiredo, 2005	Decision tree	Classification	Distribution system	Electric energy consumer
Silva, 2006	Neural network	Detection and classification	Transmission lines	Faults detection and classification
Costa, 2006	Neural network	Classification	Transmission lines	Fault classification
Dola, 2005	Decision tree and neural network	Classification	Distribution system	Faults classification
Tso, 2004	Statistical analysis	Detection	Transmission and distribution systems	Detection the substations most sensitive to the disturbances
Mori, 2002	Regression tree and neural network	Forecasting	Distribution system	Load forecasting
Dash, 2007	Support Vector Machine	Classification identification	Transmission lines	Classification and identification of series-compensated

---

<sup>1</sup>Wavelet Networking (WN) can be considered as an extension of perceptron networks.

Monedero 2007	Neural network	Classification	Not defined	Classification of electrical disturbances in real time.
Vasilic, 2005	Fuzzy/ neural network	Classification	Transmission lines	Faults classification.
Vasilic, 2002	Neural network	Classification	Transmission lines	Faults classification.
Kezunovic, 2002	Neural network	Detection and diagnostic	Transmission lines	Detection and diagnosis of transient and faults.
Huisheng, 1998	Fuzzy/ neural network	Classification	Transmission lines	Faults classification.

Table 1. Summary of tasks, techniques and applications of data mining in power systems.

### 3. Classification of time series representing faults

As mentioned, most transmission systems use three **phases**: A, B and C. Hence, a short-circuit between phases A and B will be identified as "AB". Considering the possibility of a short-circuit to "ground" (G), the task is to classify a time series into one among ten possibilities: AG, BG, CG, AB, AC, BC, ABC, ABG, ACG and BCG. The ABC and ABCG faults are typically not distinguished because in well-balanced circuits (or ATP simulations) there is no current flow through the ground (Anderson, 1995). Algorithms to solve this classification problem are used by digital fault recorders (DFRs), distance relays and other equipments (Luo & Kezunovic, 2005).

The signal capturing equipments are typically located at both endpoints of transmission line. Most of them are capable of digitizing both voltage and current waveforms. It is assumed that a *trigger* circuit detects an anomaly and stores only the interval of interest: the fault and a pre-determined number of samples before and after the fault. The trigger is out of the scope of the present work and the simulations assumed a perfect trigger algorithm, with the fault endpoints being directly obtained from the simulator.

The next subsection describes the front end, the stage that is responsible for providing a suitable parametric representation of the time series. At some points the notation may look abusive, but there are many degrees of freedom when dealing with time series and a precise notation is necessary to avoid obscure points.

#### 3.1 Front end

Each fault is a variable-duration multivariate time series. Then  $n$ -th fault  $\mathbf{X}_n$  in dataset (oscillography records, for example) is represented by a  $Q \times T_n$  matrix. A column of  $x_t$ ,  $t = 1, \dots, T_n$ , is a multidimensional sample represented by a vector of  $Q$  elements. For example, if we consider voltage and current waveforms of phases A, B and C, then  $Q = 6$  in the experiments. In some situations, it is possible to obtain *synchronized samples* from both

endpoints of a given line. In these cases the sample is an augmented vector with twice the dimension of the single endpoint scenario. For the previous example, the sample dimension for double endpoint measures would be  $Q = 12$ .

A *front end* converts samples into *features* for further processing. An example of a modern front end algorithm is the wavelets decomposition (Vertelli & Kovacevic, 1995). Independent of the adopted parametric representation, a single sample typically does not carry enough information to allow performing reasonable decisions. Hence, it is useful to consider that a *front end* converts the matrix  $X$  in to a matrix  $Z$  with dimension  $K \times N$ , as depicted in Figure 4 (the processing is performed on  $Z$ , not  $X$ ), where  $K$  is the number of features and  $N$  the number of feature vectors.

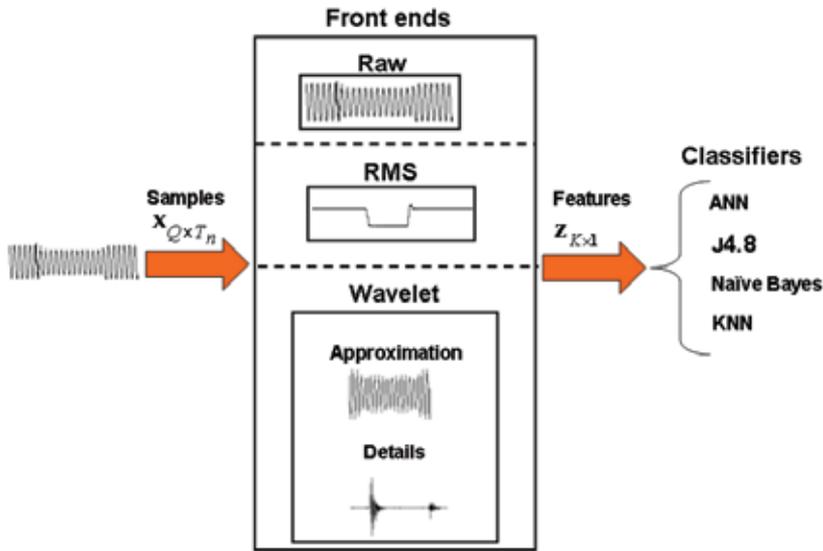


Fig. 4. The input and output matrices of the front end stage.  $Q$  and  $K$  are the dimensions of the sample and feature vectors, respectively, while  $T_n$  is the number of samples.

A *front end* is called *raw* when it outputs features that correspond to values of the original samples, without any processing other than organizing the samples into a matrix  $Z$ . In the framed raw front end, this organization is obtained through an intermediate representation called *frame*. A frame  $F$  has dimension  $Q \times L$ , where  $L$  is the number of samples called *frame length* and their concatenation  $\hat{z} = [F_1, \dots, F_n]$  is a matrix of dimension  $Q \times LN$ , where  $N$  is the number of frames.

The frames can overlap in time such that the *frame shift*  $S$ , i.e., the number of samples between two consecutive frames, is less than the frame length. Hence, the number of frames for a fault  $X_n$  is:

$$N_n = 1 + \lfloor (T_n - L) / S \rfloor \tag{1}$$

where  $\lfloor \cdot \rfloor$  is the flooring function.

The frames  $\mathbf{F}$  (matrices) are conveniently organized as vectors of dimension  $K = QL$ , and  $\hat{\mathbf{Z}}$  resized to create  $\mathbf{Z} = [\mathbf{z}_1 \dots \mathbf{z}_N]$  of dimension  $K \times N$ .

It should be noticed that, if  $S = L$  (no overlap) and a frame is a concatenation of samples

$$\mathbf{F} = [\mathbf{x}_{t-0.5(L-1)}, \dots, \mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{x}_{t+1}, \dots, \mathbf{x}_{t+0.5(L-1)}] \tag{2}$$

the matrices  $\mathbf{X}$  and  $\hat{\mathbf{Z}}$  would coincide, i.e.,  $\hat{\mathbf{Z}} = \mathbf{X}$ .

For example, in (Kezunovic & Zhang, 2007) the frames are composed by the concatenation of raw samples and vectors  $\mathbf{Z}$  have dimension  $K = 198$ . In more details, for example, If  $Q = 6$  (currents and voltages), a concatenated raw front end could obtain frames  $\mathbf{F}$  of dimension  $6 \times 5$  by concatenating to each central sample its four neighbours, two at the left and two at the right. In this case, assuming a fault with  $T = 10$  samples and  $S = L = 5$ , one would have  $K = 30$  and  $N = 2$ , such that  $\hat{\mathbf{Z}} = \mathbf{X}$ . In this case,  $\hat{\mathbf{Z}}$  and  $\mathbf{Z}$  would have dimensions  $6 \times 10$  and  $30 \times 2$ , respectively. Figure 5 illustrates the segmentation in vectors  $\mathbf{z}$  of features for one fault (ABG) with 4 frames. In this example,  $L = 3$ ,  $S = 1$  and this leads to three vectors  $\mathbf{z}$ , each of dimension  $K = 18$ .

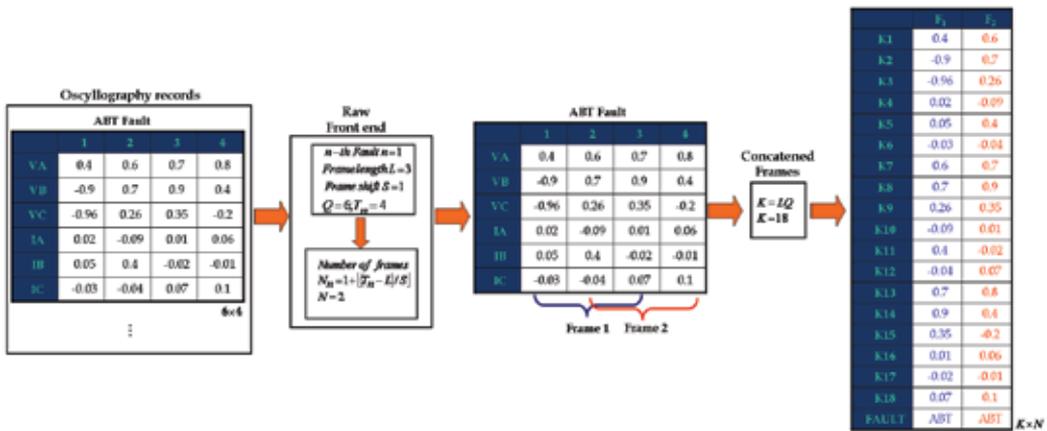


Fig. 5. Organizing feature vectors  $\mathbf{z}$  in a concatenated raw front end. In this case, the ABG fault with a total of four frames,  $L = 3$  and  $S = 1$  lead to two vectors  $\mathbf{z}$  of dimension  $K = 18$ .

As an alternative to the raw front end, the wavelet transform provides information via a multi-resolution analysis (MRA) (Vertelli & Kovacevic, 1995). When adopting this front end special care needs to be exercised to fully describe the processing, given their large number of degrees of freedom.

It is assumed a  $\gamma$ -level dyadic wavelet decomposition, which has  $\gamma$  stages of filtering and decimation (Vertelli & Kovacevic, 1995) and transforms each of the  $Q$  waveforms into  $\gamma+1$  waveforms. More specifically, the  $q$ -th waveform is decomposed into approximation  $\mathbf{a}^q$  and details  $\mathbf{d}_1^q, \mathbf{d}_2^q, \dots, \mathbf{d}_\gamma^q$ , for  $q = 1, \dots, Q$ . For simplicity, the dependence on  $q$  is omitted hereafter.

Some works in the literature use only one of the details or calculate the average power of the coefficients (Morais et al., 2007). In contrast, the framed wavelet front end keeps all the

coefficients by taking in account that for  $\gamma > 1$  they have different sampling frequencies and organizing them as matrix  $\mathbf{Z}$ . For that, instead of using a single  $L$ , the user specifies a value  $L_{min}$  for the waveforms with lowest  $f_s$  ( $\mathbf{a}$  and  $\mathbf{d}_\gamma$ ) and a large value  $L_i = 2^{\gamma-i} L_{min}$ , where  $S_{min}$  is another user-defined parameter.

The values are organized in a Frame  $\mathbf{F}$  of dimension  $Q \times L$ , where  $L = 2^\gamma L_{min}$ . The number of frames for this organization of a wavelet decomposition is

$$N = 1 + \lfloor (T_a - L_{min}) / S_{min} \rfloor \tag{3}$$

where  $T_a$  is the number of elements in  $\mathbf{a}$ .

The notation is flexible enough to easily describe several wavelet front ends, such as the concatenated wavelet (*wavelet-concat*, for which  $L_{min} = S_{min}$ ) and *wavelet-energy* described in (Morais et al, 2007).

Many recent works adopt the wavelet front end ((Saibal, 2008); (Silva, 2006); (Costa, 2006); (Chia-Hun & Chia-Hao, 2006); (Pang & Ding, 2008)). However, some of these works do not compare the wavelet with other (and eventually) simpler front ends. In Section 4, some results are presented with a simple RMS-based front end. It consists in taking the minimum RMS value of each phase during the whole fault duration. As a first step, a normalization is adopted (Morais et al., 2007) to represent the voltage values in p.u. Because the feature vectors have dimension 3, it is relatively simple to visualize them. Figure 6 illustrates 1,000

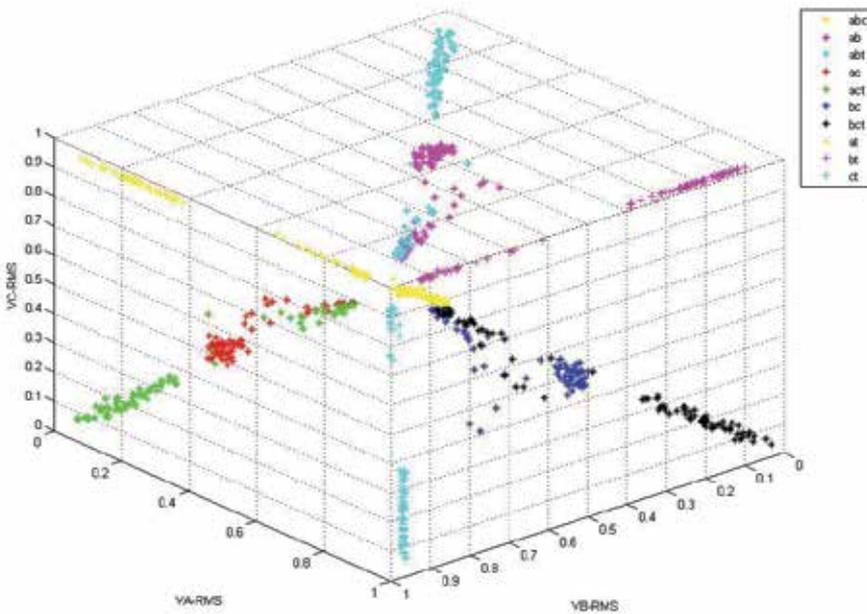


Fig. 6. Vector of features obtained with a simple RMS front end, which represents each fault by a three-elements vector. The color and shape indicate the fault category according to the legend.

of these vectors obtained via different simulations (more details in Section 4) with the color indicating the kind of fault. It can be seen, for example, that the monophasic faults (AG, BG e CG) and the triphasic fault (ABC) can be distinguished from the biphasics faults. Moreover, the biphasic faults that do not involve the ground are relatively similar to those biphasic faults involving ground.

### 3.2 On-line and post-fault classification

Fault classification systems can be divided into two types. The first one aims at performing a decision (classification) for each feature vector  $\mathbf{z}$  or, equivalently, a frame  $\mathbf{F}$  (giving that  $\mathbf{z}$  is just a representation  $\mathbf{F}$  as a vector). This is typically the goal in on-line scenarios, at the level of, e.g., a protection relay (Kezunovic & Zhang, 2007). Alternatively, the decision can be made at a supervisory center in a post-fault stage. The latter case makes a decision having available the whole matrix  $\mathbf{Z}$  of variable dimension  $K \times N_n$ , where  $n$  distinguishes the individual faults, which having distinct durations in the general case. The on-line and post-fault systems try to solve problems that can be cast as *conventional classification* (Witten & Frank, 2005) and *sequence classification* (Ming & Sleep, 2005) problems, respectively.

On-line fault classification must be performed on a very short time span with the frame located in the beginning of the fault. It is often based on a frame corresponding to half or one cycle of the sinusoidal signal (typically of 60 or 50 Hz). For example, assuming 60 Hz and a sampling frequency of  $f_s = 2$  kHz, one cycle corresponds to  $L = 2000/60$  approximate 33 samples.

As mentioned, on-line classification corresponds to the conventional scenario, where one is given a *training set*  $\{(\mathbf{z}_1, y_1), \dots, (\mathbf{z}_M, y_M)\}$  containing  $M$  examples. Each example  $(\mathbf{z}, y)$  consists of a vector  $\mathbf{z} \in \mathfrak{R}^K$  called *instance* and a *label*  $y \in \{1, \dots, Y\}$ . A conventional classifier is a mapping  $\Phi : \mathfrak{R}^K \rightarrow \{1, \dots, Y\}$ . Some classifiers are able to provide *confidence-valued scores*  $f_i(\mathbf{z})$  for each class  $i = 1, \dots, Y$  such as a probability distribution over  $y$ . For convenience, it is assumed that all classifiers return a vector  $\mathbf{y}$  with  $Y$  elements. If the classifier does not naturally return confidence-valued scores, the vector  $\mathbf{y}$  is created with a unitary score for the correct class  $f_i(\mathbf{z}) = 1$  while the others are  $f_i(\mathbf{z}) = 0, i \neq y$ . With this assumption, the final decision is given by the *max-wins* rule.

$$F(\mathbf{z}) = \arg \max_i f_i(\mathbf{z}) \quad (4)$$

Contrasting to the on-line case, a post-fault module has to classify a sequence  $\mathbf{Z}$ . The classifier is then a mapping  $\varphi : \mathfrak{R}^{K \times N} \rightarrow \{1, \dots, Y\}$  and the training set  $\{(\mathbf{z}_1, y_1), \dots, (\mathbf{z}_M, y_M)\}$  contains  $M$  sequences and their labels. The technique adopted in this work is the *frame-based sequence classification* (FBSC) (Morais et al., 2007).

In FBSC systems, the fault module repeatedly invokes a conventional classifier  $F(\mathbf{z})$  (e.g., a neural network or decision tree) to obtain scores  $\mathbf{y} = (f_1(\mathbf{z}), \dots, f_Y(\mathbf{z}))$  for each class. To come up with the final decision, the fault module can then take in account the scores of all

frames. Two possible options consist in calculating an accumulated score  $g_i(\mathbf{Z})$  for each class and then using the max-wins rule

$$G(\mathbf{Z}) = \arg \max_i g_i(\mathbf{Z}) \quad (5)$$

Where:

$$g_i(\mathbf{Z}) = \sum_{n=1}^N f_i(\mathbf{z}_n) \quad (6)$$

or

$$g_i(\mathbf{Z}) = \sum_{n=1}^N \log(f_i(\mathbf{z}_n)) \quad (7)$$

The accuracy of the system  $G(\mathbf{Z})$  can be evaluated according to the misclassification rate and it is clearly dependent on the accuracy of the classifier  $F(\mathbf{z})$ . The misclassification rates are  $E_s$  and  $E_f$ , for the post-fault (sequence) and on-line (frame) modules, respectively. In the case of post-fault systems, in spite of  $E_s$  being the actual figure of merit, it is sometimes useful to also calculate  $E_f$ . However, one should note that estimating  $E_f$  takes in account all frames that compose a fault (frames in the beginning, middle and end of the fault). In on-line applications, such as relaying, taking a decision in the beginning of the fault is the most important. In order to take this situation in account, this work defines  $E_o$  as the misclassification rate obtained when one considers only the first frame of the fault.

The next section presents simulation results for fault classification.

#### 4. Simulation results

The experiments used the UFPAFaults4 dataset, which can be downloaded from [www.laps.ufpa.br/freedatasets/UfpaFaults](http://www.laps.ufpa.br/freedatasets/UfpaFaults). The UFPAFaults4 dataset is composed by 5,500 faults, organized into five sets of 100, 200, . . . , 1000 faults each. The division into these sets is to facilitate obtaining *sample complexity* curves (Vapnik, 1999). The sample complexity indicates how many training examples are required to train the classifier. It can be evaluated by observing how the performance varies with the number of training examples.

Each fault in the dataset corresponds to three voltage and three current waveforms stored as binary files with an associated text (ASCII) files, which stores a description of the fault (its endpoints, label, etc.). The waveform samples are stored as real numbers represented as the primitive type float in Java (big-endian, 32-bits, IEEE-754 numbers).

The faults are generated with the software AMAZONTP (Pires et al., 2005). Some parameters for the simulations are randomly generated. The values of all four resistances were obtained as i.i.d. samples draw from a uniform probability density function (pdf)  $U(0.1; 10)$ , with support from 0.1 to 10 Ohms. The begin and duration (both in seconds) of the fault were draw from  $U(0.1; 0.9)$  and  $U(0.07; 0.5)$ , respectively. The location was draw from  $U(2; 98)$  (percentage of the total line length). Eleven types of faults (AG, BG, CG, AB, AC, BC, ABC, ABG, ACG, BCG, ABCG) were generated using a uniform distribution.

The voltage and current waveforms generated by the ATP simulations had a sampling period equal to 0.25 microseconds, corresponding to a sampling frequency  $f_s = 40$  kHz. It is possible to obtain versions with smaller values for  $f_s$  by decimating the original waveforms. This operation requires low-pass filtering to avoid aliasing. Details about decimation and filtering can be found in digital signal processing textbooks, e.g. (Oppenheim, 1989).

#### 4.1 Normalization

The elements of feature vectors  $\mathbf{z}$  may have very different dynamic ranges (e.g., voltage in kV and currents in Amperes). This can cause the learning algorithms to perform poorly. Therefore, as a pre-processing stage, it is important to apply a normalization process. There are many algorithms for normalization of time series. This work adopted the so-called *allfault* (Morais et al., 2007), which takes in account all duration of the waveforms for getting the maximum and minimum amplitudes of each phase, and the converting the values to the range  $[-1, 1]$ . A distinct normalization factor is calculated for each of the Q waveforms.

#### 4.2 Model selection

Often, the best performance of a learning algorithm on a particular dataset can only be achieved by tedious parameter tuning. This task is called model selection and corresponds, for example, to choosing parameters such as the number of neurons in the hidden layer for a neural network. A popular strategy for model selection is cross-validation (Witten & Frank, 2005). This is a computationally intensive approach, but avoids tuning the parameters by repeatedly evaluating the classifier using the test set. The test set should be used only once, after model selection, such that the error rate on this test set is a good indicator of the generalization capability of the learning algorithm. When dealing with frames extracted from sequences, it should be noted that, in conventional classification, the examples are assumed to be i.i.d. "samples" from an unknown but fixed distribution  $P(\mathbf{z}, y)$ . Because examples are independent, they can be arbitrarily split into training and test sets. Similarly, when organizing the folds for cross-validation, examples can be arbitrarily assigned to the training and validation fold. However, the i.i.d. assumption becomes invalid, for example, when examples  $(\mathbf{z}, y)$  are extracted from contiguous frames of the same sequence given the relatively high similarity among them. Hence, in practice it is important to use cross-validation properly, to avoid overfitting due to a training set with similar vectors extracted from the same waveform.

This work performed model selection via a validation set, disjoint to both training and test sets. A grid (Cartesian product) of model parameters is created and the point (set of parameters) that leads to the smallest error in the validation set is selected. For each coordinate, the user specifies the minimum and maximum values, the number of values and chooses between a linear or logarithmic spacing for the values.

#### 4.3 Results

The simulations in this work relied on Weka (Witten & Frank, 2005), which has many learning algorithms. Specifically, the work used decision trees (J4.8, which is a Java version of C4.5 (Witten & Frank, 2005)), multilayer artificial neural network (ANN) trained with backpropagation, naïve Bayes and K-nearest neighbor (KNN). The choice of these classifiers

was based in the fact that they are popular representatives of different learning paradigms (probabilistic, lazy, etc.).

The parameters obtained by model selection for each classifier are summarized in Table 2. The KNN used the squared-error as distance measure and  $K = 1$  neighbors. The naïve Bayes used Gaussian pdfs and does not have parameters to be tuned. For the ANN,  $H$  is the number of neurons in the hidden layer,  $N$  the maximum the number of epochs,  $L$  the learning rate and  $M$  the momentum (Witten & Frank, 2005). For J4.8,  $C$  is the confidence and  $M$  the minimum number of examples in a leaf.

Front end	L	S	K	ANN	J4.8
Raw	1	1	6	-H 8 -N 1500 -L 0.2 -M 0.3	-C 0.35 -M 10
	5	5	30	-H 20 -N 1500 -L 0.2 -M 0.3	-C 0.5467 -M 10
	7	7	42	-H 26 -N 1500 -L 0.2 -M 0.3	-C 0.7433 -M 10
	9	9	54	-H 32 -N 1500 -L 0.2 -M 0.3	-C 0.35 -M 10
	11	11	66	-H 38 -N 1500 -L 0.2 -M 0.3	-C 0.5467 -M 10
	33	33	198	-H 104 -N 1500 -L 0.2 -M 0.3	-C 0.35 -M 10
RMS	33	33	3	-H 30 -N 2000 -L 0.1 -M 0.2	-C 0.35 -M 5

Table 2. Summary of parameters for the front ends and two classifiers.

The results for frame-based classification using the concatenated raw front end are shown in Figure 7. The best results were obtained by the ANN, followed by the J4.8 classifier. The best frame length was  $L = 9$ . It is interesting to note that for  $L = 1$ , ANN achieved an error rate more than three times the one achieved by J4.8, which could be due to problems in convergence.

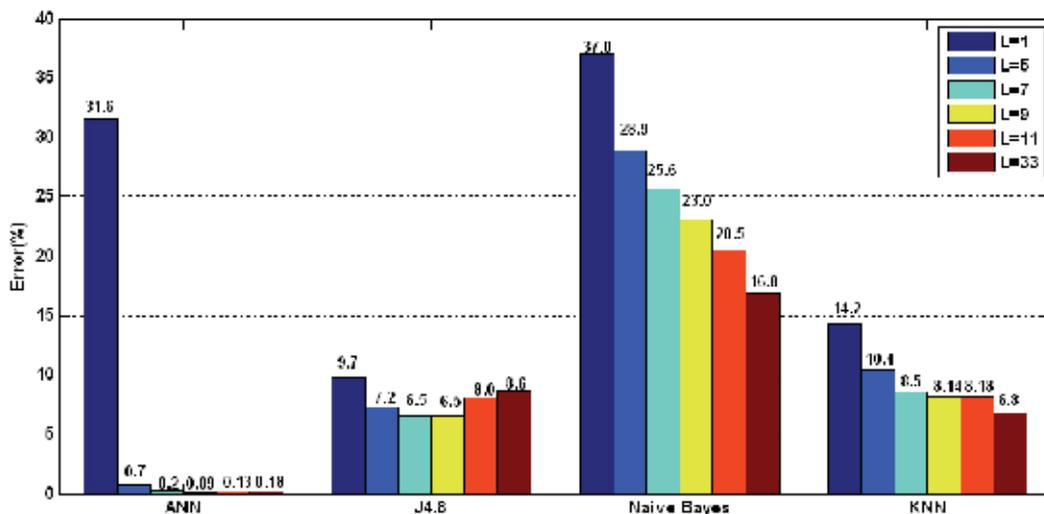


Fig. 7. Error rate  $E_f$  for several classifiers and frame lengths  $L$  using the concatenated raw front end.

The results obtained with RMS front end (Figure 8) were inferior to the ones in Figure 7, for the raw front ends. But it is interesting to note that the described RMS front end, which uses only three numbers obtained reasonable results.

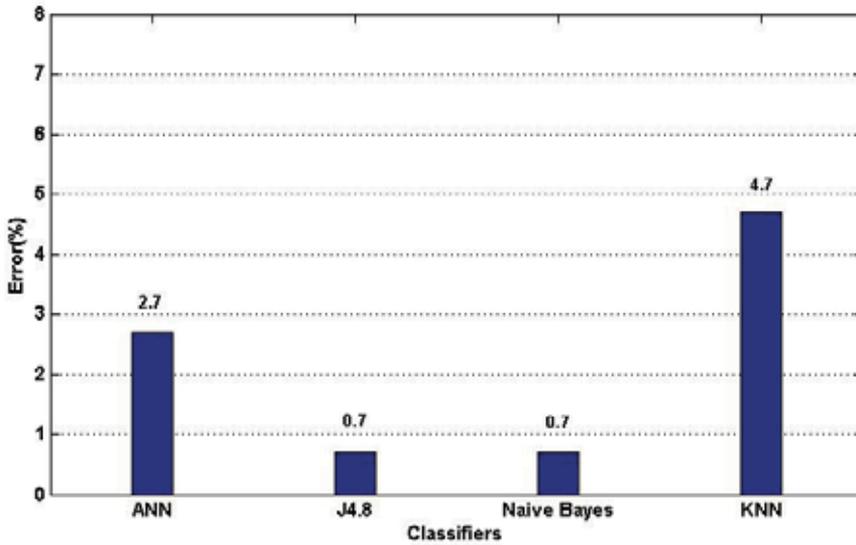


Fig. 8. Error rate  $E_s$  using RMS front end for several classifiers and frame length  $L = 33$ .

Besides, the interpretation of decision trees trained with the RMS parameters provides a good insight about the problem. Figure 9 shows an example. In this case, the participation of a phase in the short circuit can be inferred by a relatively low minimum RMS value. For example, if this minimum value is above the threshold estimated from the data, then the corresponding phase should not be involved in the short-circuit.

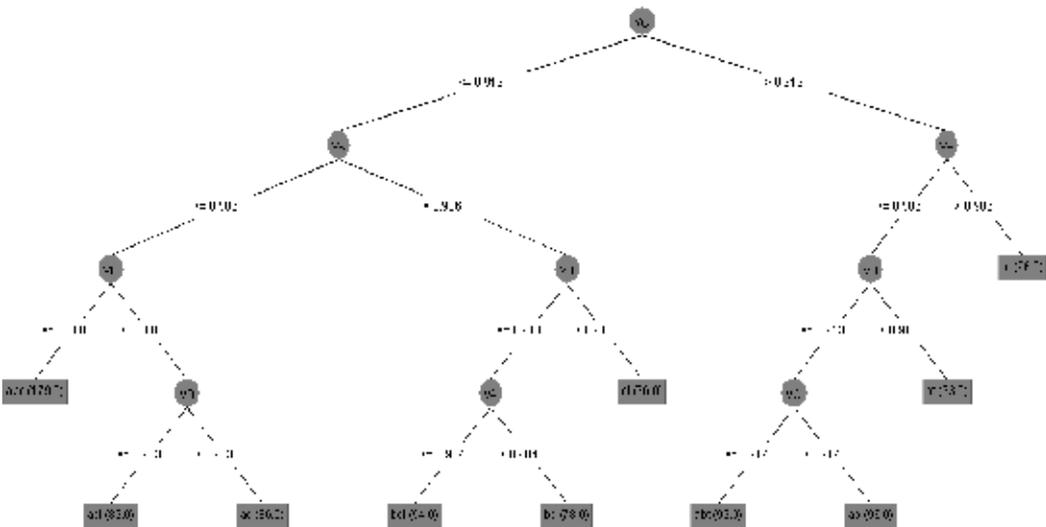


Fig. 9. Decision tree for the RMS front end that represents each fault with only three parameters: VA, VB and VC, which correspond to the minimum RMS value of each phase.

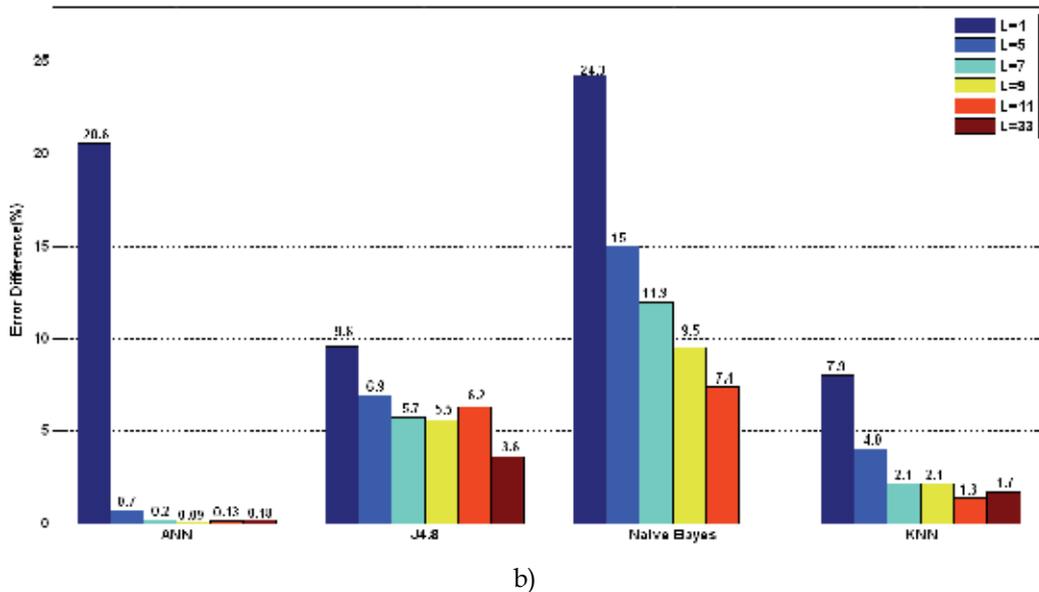
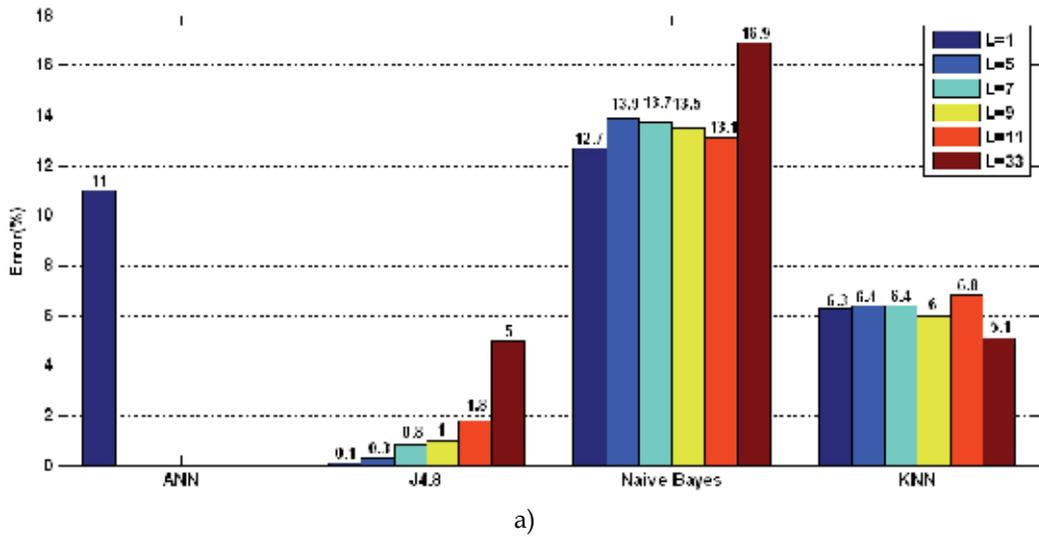


Fig. 10. Results for post-fault classification. a) Error  $E_s$  for post-fault classification. The ANN-based FBSC achieved  $E_s = 0$  for  $L > 1$ . b) Difference  $E_f - E_s$  between the error rates for frame-by-frame and sequence classification.

Figure 10 shows results for post-fault classification. Figure 10 a) shows absolute values while Figure 10b) indicates the difference between  $E_s$  and  $E_f$ . As expected, post-fault classifiers can achieve smaller error rates than the ones that operate at one frame only. One can see that the ANN-based FBSC achieves  $E_s = 0\%$  for all values of  $L$  but  $L = 1$ . The classifier J4.8 achieves  $E_s = 0.1\%$  with a computational cost smaller than the ANN.

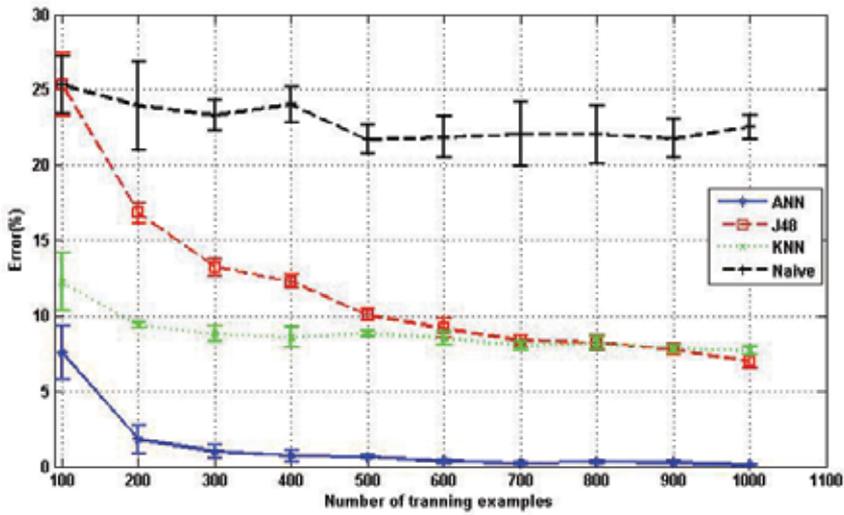


Fig. 11. Sample complexity for frame-based classification (the error is  $E_f$ ) with  $L = S = 9$ . The figure shows the average and standard deviation. It can be seen that approximately  $M = 700$  examples suffices to train the classifiers.

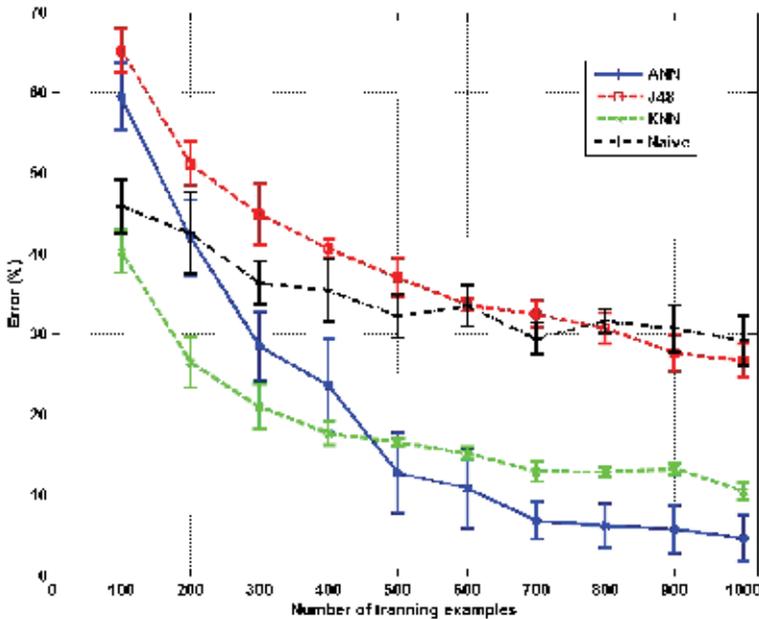


Fig. 12. Sample-complexity for one-frame-based classification with  $L = S = 9$  (error  $E_o$ ). The figure shows the mean and standard deviation. It can be seen that the error has a more erratic behavior than in Figure 11.

Figures 11 and 12 show, respectively, the results for sample complexity of frame-by-frame  $E_f$  and one-frame  $E_0$  classification with  $L = S = 9$ . Model selection was used for each value of  $M$ , given that the best parameters for the classifier typically depend on the number of training examples (Rifkin & Klautau, 2004). Comparing Figures 11 and 12 one can conclude that more examples are needed to train a classifier that observes only the first frame of the fault and its misclassification rate  $E_0$  is typically higher than  $E_f$  under the simulated conditions.

## 5. Conclusions

This work presented an overview of data mining techniques used in power systems. Among several data mining tasks, fault classification is popular especially because it is relatively easy to generate artificial data using simulators such as ATP. Other applications of data mining will potentially impact the electric power industry, but this will require data warehouses to cope with preprocessing and organizing heterogeneous and large datasets.

Even within fault classification, the research methodology needs improvement for an easier conversion of academic results into effective products. One important issue is the robustness of proposed algorithms to distinct power system constituent elements, such as the transmission line lengths. Several algorithms are tested with only one simulated scenario. This work shows that a very simple RMS front end, which represents each fault by only three values, can lead to misclassification rates under 1% in controlled conditions. Hence, it is important to improve benchmarks with publicly available datasets, such as the UFPAFaults4, and use them to properly evaluate new approaches.

This chapter also presented results comparing different figures of merit for evaluating fault classification systems. It was shown that post-fault classifiers, which can take the whole fault segment in account to make a decision, achieve smaller error rates than classifiers based on a fixed-length (and short) frame. In fact, neural networks precisely classified all test examples (zero error) in some configurations. Another aspect that was emphasized is that, as vastly discussed in the machine learning literature, the number of examples to train a classifier depends on the learning algorithm and the domain (dataset).

## 6. References

- Mori, H.; Kosemura, N.; Kondo, T.; Numa, K.; (2002). Data mining for short-term load forecasting. *Power Engineering Society Winter Meeting*. pp. 623 - 624, ISBN 0-7803-7322-7, Jan 2002.
- Kezunovic, M.; Zhang, N. (2007). A real time fault analysis tool for monitoring operation of transmission line protective relay, *Electric Power Systems Research*, Vol.77, No.3-4, (March 2007) 361-70.
- ATP, (1987). Alternative Transients Program, Rule Book, Leuven EMTP Center (LEC).
- Casazza, J.; Delea, F. (2005). *Understanding Electric Power Systems - An Overview of the Technology and the Marketplace*. Electrical Insulation Magazine, IEEE. ISBN: 978-0-471-44652-1.

- Boyer, Stuart A.(1999). SCADA: Supervisory Control and Data Acquisition , 2nd Edition, ISA International Society for Measurement. ISBN 1-55617-660-0.
- Ramos, S.; Vale, Z. (2008). Data mining techniques application in power distribution utilities. *Transmission and Distribution Conference and Exposition*, pp. 1-8. ISBN. 978-1-4244-1903-6, Chicago, April 2008.
- Hagh, M.T.; Razi, K.; Taghizadeh, H. (2007). Fault classification and location of power transmission lines using artificial neural network, *International Power Engineering Conference*, pp. 1109 - 1114, ISBN 978-981-05-9423-7, Singapore, Dec 2007.
- Saibal Chatterjee, Sivaji Chakravorti, Chinmoy Kanti Roy and Debangshu Dey. (2008) Wavelet network-based classification of transients using dominant frequency signature, *Electric Power Systems Research*, Vol. 78, No. 1, (January 2008) 21-29.
- Chia-Hung Lin; Chia-Hao Wang.(2006).Adaptive wavelet networks for power-quality detection and discrimination in a power system, *IEEE Transactions on Power Delivery*, Vol 21, No. 3,(July 2006) 1106 - 1113, ISSN: 0885-897.
- Pang, P.; Ding, G.(2008). Power quality detection and discrimination in distributed power system based on wavelet transform. *27th Chinese Control Conference (CCC 2008)*, pp. 635 - 638, ISBN 978-7-900719-70-6, China, July 2008.
- Bhende C. N.; Mishras S.; Panigrahi B. K. (2008). Detection and classification of power quality disturbances using S-transform and modular neural network . *Electric Power Systems Research*, Vol. 78, (February 2008) 122-128.
- Figueredo, V.; Rodrigues F.; Vale, Z.; Gouveia, J. B. (2005). An electric energy Consumer characterization Framework based on data mining techniques. *IEEE Transactions Power Systems*, Vol. 20, No. 2., May 2005 , 596- 60), ISSN 1558-0679.
- Silva, K. M.; Souza, B. A.; Brito, N. S. D. (2006). Fault detection and classification in transmission lines based wavelet transform and ANN. *IEEE Transaction on Power Delivery*, Vol 21 , No. 4, (October 2006) 2058-2063, ISSN 0885-8977.
- Costa, F. B.; Silva, K. M.; Souza, B. A.; Dantas, K. M. C.; Brito, N. S. D. (2006). A method for fault classification in transmission lines bases on ANN and wavelet coefficients energy. *International Joint Conference Neural Networks*. pp. 3700 - 3705, ISBN 0-7803-9490-9, Vancouver, July-2006.
- Dola, H.M.; Chowdhury, B.H. (2005). Data mining for distribution system fault classification. *Power Symposium, 2005. Proceedings of the 37th Annual North American*, pp. 457 - 462, ISBN 0-7803-9255-8, October 2005.
- Tso, S.K.; Lin, J.K.; Ho, H.K.; Mak, C.M.; Yung, K.M.; Ho, Y.K. (2004). Data mining for detection of sensitive buses and influential buses in a power system subjected to disturbances. *IEEE Transactions on Power Systems*, Vol. 19, No.1, (February 2004) 563 - 568, ISSN 1558-0679.
- Dash, P.K.; Samantaray, S.R.; Panda, G. (2007). Fault Classification and Section Identification of an Advanced Series-Compensated Transmission Line Using Support Vector Machine. *IEEE Transactions on Power Delivery*, Vol 22. No. 22, (January 2007) 67 - 73, ISSN 0885-8977.

- Monedero, I.; Leon, C.; Roperro, J.; Garcia, A.; Elena, J.M.; Montano, J. C. (2007). Classification of Electrical Disturbances in Real Time Using Neural Networks. *IEEE Transactions on Power Delivery*, Vol. 22, No. 3, (July 2007) 1288 - 1296, ISSN 0885-8977.
- Vasilic, S.; Kezunovic, M. (2005) Fuzzy ART Neural Network Algorithm for Classifying the Power System Faults. *IEEE Transactions on Power Delivery*, Vol. 20, No. 2, (April 2005) 1306-1314, ISSN 0885-8977.
- Vasilic, S.; Kezunovic, M. (2002). An Improved Neural Network Algorithm for Classifying the Transmission Line Faults. *IEEE Power Engineering Society Winter Meeting*, pp. 918 - 923. ISBN 0-7803-7322-7, Jan 2002.
- Kezunovic, M. ; Vasilic, S.; Gul-Bagriyanik, F. (2002). Advanced Approaches for Detecting and Diagnosing Transients and Faults. 2002.
- Huisheng Wang; Keerthipala, W.W.L. (1998). Fuzzy-neuro approach to fault classification for transmission line protection. *IEEE Transactions Power Delivery*, Vol.13, No. 4, 1093-1104, ISSN 0885-8977.
- Anderson, P. M. (1995). *Analysis of faulted Power Systems*. IEEE Press Series on Power Engineering, ISBN 978-0-7803-1145-9.
- Luo, X.; Kezunovic, M.(2005). Fault Analysis Based on Integration of Digital Relay and DFR. *Power Engineering Society General Meeting*, pp. 746 - 751, ISBN 0-7803-9157-8, Jun 2005.
- Vertelli, M.; Kovacevic, J. (1995). *Wavelets and Subband Coding*. Prentice Hall, ISBN 978-0130970800.
- Morais J.; Pires, Y.; Cardoso, C.; Klautau, A. (2007). Data mining applied to the electric power industry: Classification of short-circuit faults in transmission lines. In *IEEE International Conference on Intelligent Systems Design and Applications (ISDA)*, pp. 943-948, ISBN: 978-0-7695-2976-9 , October 2007.
- Witten, I.; Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann, 2nd Edition, ISBN 0-12-088407-0.
- Ming Li.; Sleep, R. (2005). A robust approach to sequence classification, *International Conference on Tools with Artificial Intelligence*, pp.5, ISBN 0-7695-2488-5, November 2005.
- Vapnik, V.N (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, Vol 10, No.5, Sept. 1999.988 - 999, ISSN: 1045-9227.
- Pires, Y. P.; Santos, A.; Borges, J.; Carvalho, A.; Nunes, M. V. A.; Santoso, S.; Klautau, A. (2005). A framework for evaluating data mining techniques applied to power quality. *Brasilizn Conference on Neural Network*, October 2005.
- Yang, K.; Shahabi, C. (2004). A PCA-based similarity measure for multivariate time series. 2nd, ACM International Workshop on Multimedia DataBases, pp. 65-74, ISBN 1-58113-975-6, Washington, DC, USA.
- Oppenheim, A.; Schafer, R. (1989). *Discrete-time Signal Processing*. Prentice-Hall, 2nd Edition , ISBN 9780137549207.

Rifkin, R.; Klautau, A. (2004). *In defense of one-vs-all classification*. Journal of Machine Learning Research, 5:101-141, 2004.



*Edited by Julio Ponce and Adem Karahoca*

This book presents four different ways of theoretical and practical advances and applications of data mining in different promising areas like Industrialist, Biological, and Social. Twenty six chapters cover different special topics with proposed novel ideas. Each chapter gives an overview of the subjects and some of the chapters have cases with offered data mining solutions. We hope that this book will be a useful aid in showing a right way for the students, researchers and practitioners in their studies.

Photo by solarseven / iStock

**IntechOpen**

