IntechOpen

# Data Mining

*Edited by Ciza Thomas*

# DATA MINING

Edited by **Ciza Thomas**

**Data Mining**
http://dx.doi.org/10.5772/intechopen.71371
Edited by Ciza Thomas

**Contributors**

Mümine Kaya Keleş, Abdullah Emre Keleş, Sien Chen, Klaus Jantke, Oksana Arnold, Taotao Liu, Manaswini Pradhan, Vesa Kalevi Salminen, Päivi Sanerma, Seppo Niittymaki, Patrik Eklund, Mustafa Kemal Pektürk, Muhammet Ünal, Priyank Jain, Derya Birant, Aysegul Pala, Goksu Tuysuzoglu, Milan Vukicevic, Sven Van Poucke, Ana Kovacevic, Jose Ávila, Eugenio Saavedra, Ailed Marenco, Ivon Romero

**Notice**

Statements and opinions expressed in the chapters are these 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 London, United Kingdom, 2018 by IntechOpen
eBook (PDF) Published by IntechOpen, 2019
IntechOpen is the global imprint of INTECHOPEN LIMITED, registered in England and Wales, registration number: 11086078, The Shard, 25th floor, 32 London Bridge Street
London, SE19SG – United Kingdom
Printed in Croatia

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Additional hard and PDF copies can be obtained from orders@intechopen.com

Data Mining
Edited by Ciza Thomas

p. cm.

Print ISBN 978-1-78923-596-8

Online ISBN 978-1-78923-597-5

eBook (PDF) ISBN 978-1-83881-567-7

# Meet the editor

Dr. Ciza Thomas is currently working as professor and head of Electronics and Communication Department of the College of Engineering, Trivandrum, India. Her area of expertise is network security with research interest in the fields of information security, data mining, sensor fusion, pattern recognition, information retrieval, digital signal processing, and image processing. She has published more than 40 international journal articles and international conference proceedings and 50 national conference publications. She has edited five books and published six book chapters. She is a reviewer of more than 10 reputed international journals including IEEE transactions on Signal Processing, IEEE transactions on Neural Networks, International Journal of Network Security, International Journal of Network Management, and IEEE-John Wiley International Journal on Security and Communications Network. She is a recipient of an achievement award in 2010 and the e-learning IT award in 2014 from the Government of Kerala.

# Contents

# Preface

This book on data mining discusses a broad set of ideas and presents some of the advanced research in this field. The book is triggered by pervasive applications that retrieve knowledge from real-world big data. It provides the basics, methods, and tools in processing large data available with various applications. The chapters discuss various applications and research frontiers in data mining with algorithms and implementation details for use in real-world.

The inundation of data from various fields like scientific observations, medical data, social science data, financial data, business data, etc. has been an issue in the recent past and is expected to become worse in future. Manual analysis of this large volume of data is difficult and there is a need to automate the data analysis. Data mining is the intelligent supporting techniques in discovering useful information and knowledge from the big data. This can be through characterization, classification, discrimination, anomaly detection, association, clustering, trend or evolution prediction, etc., and accordingly, data mining can be either descriptive or predictive.

Researchers and practitioners in areas such as statistics, pattern recognition, machine learning, artificial intelligence, data analytics, and visualization are contributing to the field of data mining for better utilization of the data. Data mining finds applications in the entire spectrum of science and technology including basic sciences to life sciences and medicine, to social, economic, and cognitive sciences, to engineering and computers.

This book on data mining consists of ten chapters. The chapters include the real-world problems in various fields and propose methods to address each one of them. Technologies explored in each of these chapters are introduced for the reader in every chapter.

In Chapter I, the ensemble methods in environmental data mining are discussed. Environmental data mining is the nontrivial process of identifying valid, novel, and potentially useful patterns in data from environmental sciences. This chapter proposes ensemble methods like bagging, random forest, boosting, and voting in environmental data mining that combine the outputs from multiple classification models to obtain better results than the outputs that could be obtained with a single model. In the experimental studies, ensemble methods are tested on different real-world environmental datasets in various subjects such as air, ecology, rainfall, and soil.

In order to maintain a good relationship with customers in business or with workers in a production field, there is always the need to collect and analyze the data and interpret the information. Data mining can play an important part in such scenarios as seen in Chapters II and III. Chapter II discusses the estimation of customer lifetime value using machine learning techniques. The chapter includes the passenger network value assessment model for the least resource investment for maximum profit return to help airlines find their significant customers. Chapter III proposes a methodology for determination of crew productivity using data mining methods. The identification of leadership types that will motivate and support employees has great importance in construction businesses where the human element is of significance. For this purpose, the relationship of productivity between the engineers working in construction companies and workers who work with them is examined and analyzed using data mining methods.

Chapter IV discusses the mining human-computer interaction (HCI) data for theory of mind induction. The HCI data has the potential for understanding a human user's intentions, goals, and desires. Knowing what users want and need is key to intelligent system assistance. The theory of mind concept known from studies in animal behavior is adopted and adapted for expressive user modeling. Theories of mind are hypothetical user models representing, to some extent, a human user's thoughts. Theories of mind are induced by mining HCI data.

Chapter V discusses performance-aware high-performance computing for remote sensing big data analytics. The chapter introduces a novel high-performance computing system on the geo-distributed private cloud for remote sensing applications that take advantage of network topology and exploit utilization and workloads of CPU, storage, and memory resources in a distributed fashion and optimize resource allocation for realizing big data analytics efficiently.

Data mining techniques are used these days in the medical field as they have the potential to improve health systems. Data mining uses data and analytics to identify the best practices that improve care at reduced cost. Chapter VI proposes a predictive model for early prediction of patient mortality. Experimental evaluation is conducted on patients admitted to ICUs with renal failure. Chapter VII presents a semantic infrastructure for service environment supporting successful aging. The aging individuals living independently at home need new kinds of services and service environments. Digitalization of services and the data gathered from the individuals creates an opportunity for more optimized and punctual services. The data gathered through digital equipment is used in optimizing service processes. However, service process misses common ontology and semantic infrastructure to use the gathered data for service optimization. This chapter introduces the service environment and semantic infrastructure, which could be used in social and health care. Chapter VIII presents an adaptive neural network classifier-based analysis of big data in health care. An FCM-based MapReduce programming model is used for the parallel computing using the AANN approach. The FCM-based MapReduce clusters the large medical datasets into smaller groups of certain similarity and assigns each data cluster to one Mapper, where the training of neural networks are done by the optimal selection of the interconnection weights by Whale Optimization Algorithm (WOA). Finally, the reducer reduces all the AANN classifiers obtained from the Mappers for identifying the normal and abnormal classes of the newer medical records promptly and accurately.

Chapter IX is on the identification of research thematic approaches based on keywords network analysis in Colombian social sciences. This chapter unveils the structure of knowledge of social sciences in Colombia through the analysis of thematic networks and its association with different disciplines' new knowledge production to define scenarios and trends in each. About 2992 published articles in the period 2006–2015 are revised in this work, all indexed in Web of Science, Scopus and other bibliographic databases, applying the social networks analysis technique to the keywords of all. The analysis includes each discipline's clustering coefficient and group metrics. The results described in this chapter identify how social disciplines in Colombia have mainly focused their research production on topics such as armed conflict, poverty, and human development. Chapter X presents data privacy for big data publishing using newly enhanced PASS data mining mechanism. The growth in data made anonymization using conventional processing methods inefficient. This chapter proposes PASS mechanism in Hadoop framework to reduce the processing time of anonymization. In this work, the whole program is divided into the map and reduced parts. Moreover, the data types used in Hadoop provide better serialization and transport of data.

The intended audience of this book will mainly consist of students, researchers, practitioners, data analysts, and business professionals who seek information on the various data mining techniques and their applications.

I would like to convey my gratitude to everyone who contributed to this book including the authors of the accepted chapters. My special thanks go to the Publishing Process Manager, Mr. Julian Virag, and other staff of InTech publishing for their support and efforts in bringing the book to fruitful completion.

**Prof. Ciza Thomas, Professor and Head**
College of Engineering Trivandrum, India

Chapter 1

# Ensemble Methods in Environmental Data Mining

Goksu Tuysuzoglu, Derya Birant and Aysegul Pala

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.74393

**Abstract**

Environmental data mining is the nontrivial process of identifying valid, novel, and potentially useful patterns in data from environmental sciences. This chapter proposes ensemble methods in environmental data mining that combines the outputs from multiple classification models to obtain better results than the outputs that could be obtained by an individual model. The study presented in this chapter focuses on several ensemble strategies in addition to the standard single classifiers such as decision tree, naive Bayes, support vector machine, and k-nearest neighbor (KNN), popularly used in literature. This is the first study that compares four ensemble strategies for environmental data mining: (i) *bagging*, (ii) bagging combined with random feature subset selection (the *random forest* algorithm), (iii) *boosting* (the AdaBoost algorithm), and (iv) *voting* of different algorithms. In the experimental studies, ensemble methods are tested on different real-world environmental datasets in various subjects such as air, ecology, rainfall, and soil.

**Keywords:** data mining, classification, ensemble learning, environmental data, bagging, random forest, AdaBoost

## 1. Introduction

*Environmental data mining* is defined as extracting knowledge from huge sets of environmental data. It is an interdisciplinary area of both computer and environmental sciences, including but not limited to environmental information management systems, decision support systems, recommender systems, environmental data analytics, and so on.

Environmental data mining based on ensemble learning is a rather young research area where a set of learners are trained sequentially on the dataset to better analyze and understand

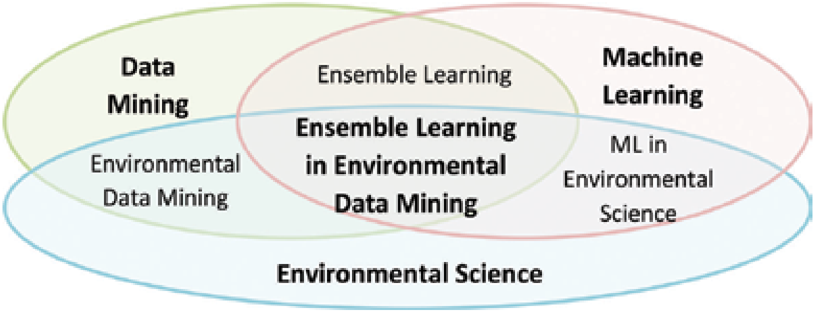

Data Mining
Ensemble Learning
Machine Learning
Ensemble Learning in Environmental Data Mining
Environmental Data Mining
ML in Environmental Science
Environmental Science

**Figure 1.** Interdisciplinary structure of ensemble learning in environmental data mining (ELEDM).

environmental processes and systems. However, it is not well-known yet how ensemble methodology can be utilized in order to improve the performance of a single method. For this purpose, this chapter presents the findings of a systematic survey of what is currently done in the area and aims to investigate the ability of different ensemble strategies for environmental data mining.

Ensemble learning in environmental data mining (ELEDM) can be drawn as a combination of three main areas: data mining (DM), machine learning (ML), and environmental science (**Figure 1**). ML in environmental science is learning-driven, meaning that machines teach themselves to recognize patterns by analyzing environmental data, whereas in contrast, DM is discovery-driven, meaning that patterns are automatically discovered from environmental data. DM uses many ML methods, including ensemble learning methods.

The novelty and main contributions of this chapter are as follows. First, it provides a brief survey of ensemble learning used in environmental data mining. Second, it presents how an ensemble of classifiers can be applied on environmental data in order to improve the performance of a single classifier. Third, it is the first study that compares different ensemble strategies on different environmental datasets in terms of classification accuracy.

## 2. Related work

Data mining techniques have been recently utilized in environmental studies for processing environmental data and converting it to useful patterns to obtain valuable knowledge and make right decisions when dealing with environmental problems. Many of the developed techniques in data mining can often be tailored to fit environmental data.

Recently, ensemble learning has been one of the active research fields in machine learning. Thus, it has been utilized in a very broad range of areas such as marketing, banking, insurance, health, telecommunication, and manufacturing. In contrast to these studies, our work proposes ensemble learning approach that combines several models to produce a result to solve environmental problems.

### 2.1. Ensemble-based environmental data mining

Ensemble classifiers have been applied to different environmental subjects, such as air [1–6], water [7–9], soil [10–12], plant [13], forests [14, 15], climate [16–18], noise [19], rainfall [20], energy [21–23], as well as living organisms [18, 24, 25]. Some of the ensemble-based environmental data mining studies have been compared in **Table 1**. In this table, the scopes of the studies, the year they were performed, the algorithms that were used in the studies, the type of data mining task, the success rate with the validation method, and the ensemble strategy are listed. In addition, if more than one algorithm is presented and compared with each other, the proposed one (the most successful one) is also indicated. As given in the table, ensemble of models for classification or prediction has higher interest than ensemble clustering and anomaly detection [2, 22] in environmental science. Although ensemble clustering has been used in many areas, especially in bioinformatics, only a few studies [4, 25] have been conducted so far in the environmental science.

| Ref. | Year | Type | Description | Data mining task | Ensemble strategy | Algorithms | Validation |
|---|---|---|---|---|---|---|---|
| [22] | 2017 | Energy | Identification of anomalous consumption patterns in building energy consumption | Anomaly detection | 2, 4 | RF, SVR, CCAD-SW using autoencoder and PCA, EAD | TPR = 98.10%<br>FPR = 1.98% (for EAD model) |
| [18] | 2016 | Climate | Determination of the impact of climate change on the habitat suitability for large brown trout | Prediction | 1, 2 | Generalized additive models, MLP with bagging ensembles, RF, SVM, and fuzzy rule-based systems (TSK) | Threefold cross validation<br>Weighted MSE = 0.18 (MLP with bagging ensembles)<br>Overall true skill statistics (TSS) = 0.69 (RF) |
| [11] | 2015 | Soil | Classification of complex land use/land cover categories of desert landscapes using remotely sensed data | Classification | 2, 3 | RF and boosted ANNs | Mean class user's accuracy = 86.7% (for boosted ANN) and 86.6% (for RF ANN) |
| [26] | 2015 | Soil | Solve the problem of rare classes' classification on dust storm forecasting | Classification | 2, 3 | SMOTE with AdaBoost and RF (SARF), SVM, fuzzy ANN | Tenfold cross validation<br>Accuracy = 96.51% (SARF) |
| [4] | 2015 | Air | Forecasting of air pollutant values for the Attica area | Clustering | 2, 4 | SOM for clustering, FFANN and RF ANN for regression, FIS to obtain fuzzy values | Tenfold cross validation<br>RMSE and $R^2$ |

| Ref. | Year | Type | Description | Data mining task | Ensemble strategy | Algorithms | Validation |
|---|---|---|---|---|---|---|---|
| [9] | 2014 | Water | Predictive modeling of groundwater nitrate pollution | Prediction | 2 | RF regression, LR | ROC = 0.923 (for model RF-A)<br>AUC = 0.911 (for model RF-B) |
| [25] | 2013 | Living organisms | Construction of habitat models for living species in the Lake Prespa, Macedonia; in the soils of Denmark; and in the Slovenian rivers | Clustering | 1, 2 | RF and bagged multitarget predictive clustering tree (PCT) and single-target DT | Tenfold cross validation<br>RRMSE |
| [5] | 2012 | Air | Prediction of the Macau's air pollution index | Prediction | 1 | Bootstrap sampling with replacement and random sampling without replacement using ANFIS method as base learner | RMSE = 12.21 (ANFIS with random sampling) |
| [2] | 2011 | Air energy | Detect overconsumption of fuel in aircrafts | Anomaly detection | 1 | Bootstrap sampling on each of the regression tree (tree), elastic network, GP, and stable GP regression methods | ROC = 0.90<br>NRMSE varied consistently between 85 and 90% |

ANN, artificial neural network; SVR, support vector regression; PCA, principal component analysis; MLP, multilayer perceptron; SOM, self-organizing maps; EAD, ensemble anomaly detection; FIS, fuzzy inference system; GP, Gaussian process; MSE, mean squared error; RMSE, root-mean-square error; TPR, true positive rate; ROC, receiver operating characteristic.

**Table 1.** Comparison of ensemble-based environmental data mining studies.

The idea of using an ensemble of classifiers rather than the single best classifier has been proposed in several environmental data mining studies [5, 11, 26]. It is apparent that ensemble learners boost the performance of the single classifiers. Different models pick up different patterns in data. By pooling all these predictions together, as long as they are reasonably independent, informed, and diverse, the outcomes tend to be better.

One of the most popular ensemble learning strategies, *bagging*, is also well adapted to develop models for solving environmental problems. For example, it has been utilized to the forecast air pollution level of a region [5] and to establish habitat models for living species [25].

The second type of ensemble learning strategy, the *random forest* (RF) algorithm, has also been applied for classifying environmental data. It has been applied to predict pollutant occurrences in groundwater [9] and determination of the impact of climate change on the habitat suitability for a fish species [18] and to predict dust storm accurately [26].

Another ensemble learning strategy (*boosting*), the AdaBoost algorithm, has been used in various types of environmental applications such as for the classification of complex land use/land cover categories of desert landscapes using remotely sensed data [11], to solve the problem of rare classes' classification on dust storm forecasting [26] and discovering plant species for automatic weed control [27].

Training with different algorithms in each ensemble (*voting*) is another commonly used ensemble strategy in environmental science. Some of the examples are for the identification of anomalous consumption patterns in building energy consumption [22] and forecasting air pollutant values of a region [4].

Differently from existing studies, the study presented in this chapter focuses on applying four distinct ensemble strategies to environmental datasets using (i) different training sets formed by random sampling with replacement (bagging), (ii) different training sets obtained by random instance and feature subset selection (random forest), (iii) different training sets using random sampling with replacement over weighted data (AdaBoost), and (iv) different algorithms (voting).

### 2.2. Advantages of ensemble-based environmental data mining

Some of the advantages of environmental data mining are given below:

- Prediction of parameters expected based on other parameters or under different cases in environmental studies, for example, prediction of rainfall [20], climate change [16–18] species richness/diversity [24, 25], and atmospheric parameters [28].
- Construction of models to reduce the consumption of energy [21–23] and raw materials [2] such as wood, grass, metal, steel, plastics, glass, paper, fuel, and natural gas.
- Clustering the items in environmental data to describe the current situation more clearly and to plan different activities for different clusters [4, 25].
- Classification of environmental audio and environmental noise [19].
- Processing ecological data for better modeling ecological systems [24, 25].
- Analyzing environmental data toward a better quality control such as air quality [1, 5, 6] and water quality [7–9].
- Identifying unexpected patterns from an environmental data using a data mining algorithm and detection of anomalies in environmental data [2, 22] to identify bad values, changes, errors, noises, frauds, and abnormal activities to realize the purpose of giving an alarm.
- Determination of the most important factor that affects the environment using a data mining technique such as decision tree and random forest [29].
- Development of a model to manage resources effectively [2, 21, 23], including environmental resources such as air, water, and soil; flow resources such as solar power [30] and wind energy; and natural resources such as coal, gas, and forests.
- Discovering patterns that can be used for better waste management and recycling.
- Analyzing the records of financial transactions related to environmental economics for better decision-making, i.e., investigating the financial impacts of environmental policies.

- Using ensemble methods as a preprocessing step before performing the essential environmental study.
- Clustering environmental documents according to their topics and main contents.
- Usage of process mining to improve work management in the environmental science.

## 3. Background information

### 3.1. Ensemble learning

Ensemble learning is a machine learning technique where multiple learners are trained to solve the same problem and their predictions are combined with a single output that probably has better performance on average than any individual ensemble member. The fundamental idea behind ensemble learning is to combine weak learners into one, a strong learner, who has a better generalization error and is less sensitive to overfitting in the presence of noise or small sample size. This is because different classifiers can sometimes misclassify different patterns and accuracy can be improved by combining the decisions of complementary classifiers.

### 3.2. Elements of an ensemble classifier

A typical ensemble framework for classification tasks contains four fundamental components descripted as follows:

- *Training set*: a training set is a special set of labeled examples providing known information that are used for training.
- *Base inducer*(s) or *base classifier*(s): an inducer is a learning algorithm that is used to learn from a training set. A base inducer obtains a training set and constructs a classifier that generalizes relationship between the input features and the target outcome.
- *Diversity generator*: it is clear that nothing is gained from an ensemble model if all ensemble members are identical. The diversity generator is responsible for generating the diverse classifiers and decides the type of every base classifier that differs from each other. Diversity can be realized in different ways depending on the accuracy of individual classifiers for the improved classification performance. Common diversity creation approaches are (i) using different training sets, (ii) combining different inducers, and (iii) using different parameters for a single inducer.
- *Combiner*: the task of the combiner is to produce the final decision by combining all classification results of the various base inducers. There are two main methods of combining: weighting methods and meta-learning methods. *Weighting methods* give each classifier a weight proportional to its strength and combine their votes based on these weights. The weights can be fixed or dynamically determined when classifying an instance. Common weighting methods are majority voting, performance weighting, Bayesian combination, and vogging. *Meta-learning methods* learn from new training data created from the predictions of a set of base classifiers. The most well-known meta-learning methods are stacking

and grading. While weighting methods are useful when combining classifiers built from a single learning algorithm and they have comparable success, meta-learning is a good choice for cases in which base classifiers consistently classify correctly or consistently misclassify.

## 4. Ensemble strategies

In order to construct an ensemble model, any of the following strategies can be performed:

### 4.1. Strategy 1: different training sets using random sampling with replacement

One ensemble strategy is to train different base learners by different subsets of the training set. This can be done by random resampling of a dataset (i.e., *bagging*; **Figure 2a**). When we train multiple base learners with different training sets, it is possible to reduce variance and therefore error.

### 4.2. Strategy 2: different training sets obtained by random instance and feature subset selection

The combination of bagged decision trees is constructed similar to Strategy 1 using one significant adjustment that random feature subsets are used (i.e., *random forest*; **Figure 2b**). When we have enough trees in the forest, random forest classifier is less likely overfit the model. It is also useful to reduce the variance of low-bias models, besides handling missing values easily.

### 4.3. Strategy 3: different training sets using random sampling with replacement over weighted data

This ensemble strategy can be implemented by weighted resampling of the dataset serially by focusing on difficult examples which are not correctly classified in the previous steps (i.e., *boosting*; **Figure 2c**). Boosting helps to decrease the bias of otherwise stable learners such as linear classifiers or univariate decision trees also known as decision stumps.

### 4.4. Strategy 4: different algorithms

The other ensemble strategy (i.e., *voting*; **Figure 2d**) is to use different learning algorithms to train different base learners on the same dataset. So, the ensemble includes diverse algorithms that each takes a completely different approach. The main idea behind this kind of ensemble learning is taking advantage of classification algorithms' diversity to face complex data.

### 4.5. Characteristic of different ensemble classifiers

Although ensemble classifiers have a common goal to construct multiple, diverse and predictive models and finally to combine their outputs, each strategy is carried out in different ways using different training sets, combiner or inducer. **Table 2** summarizes the properties of different ensemble strategies, the popular algorithms under each category and pros and cons of each ensemble classifier.

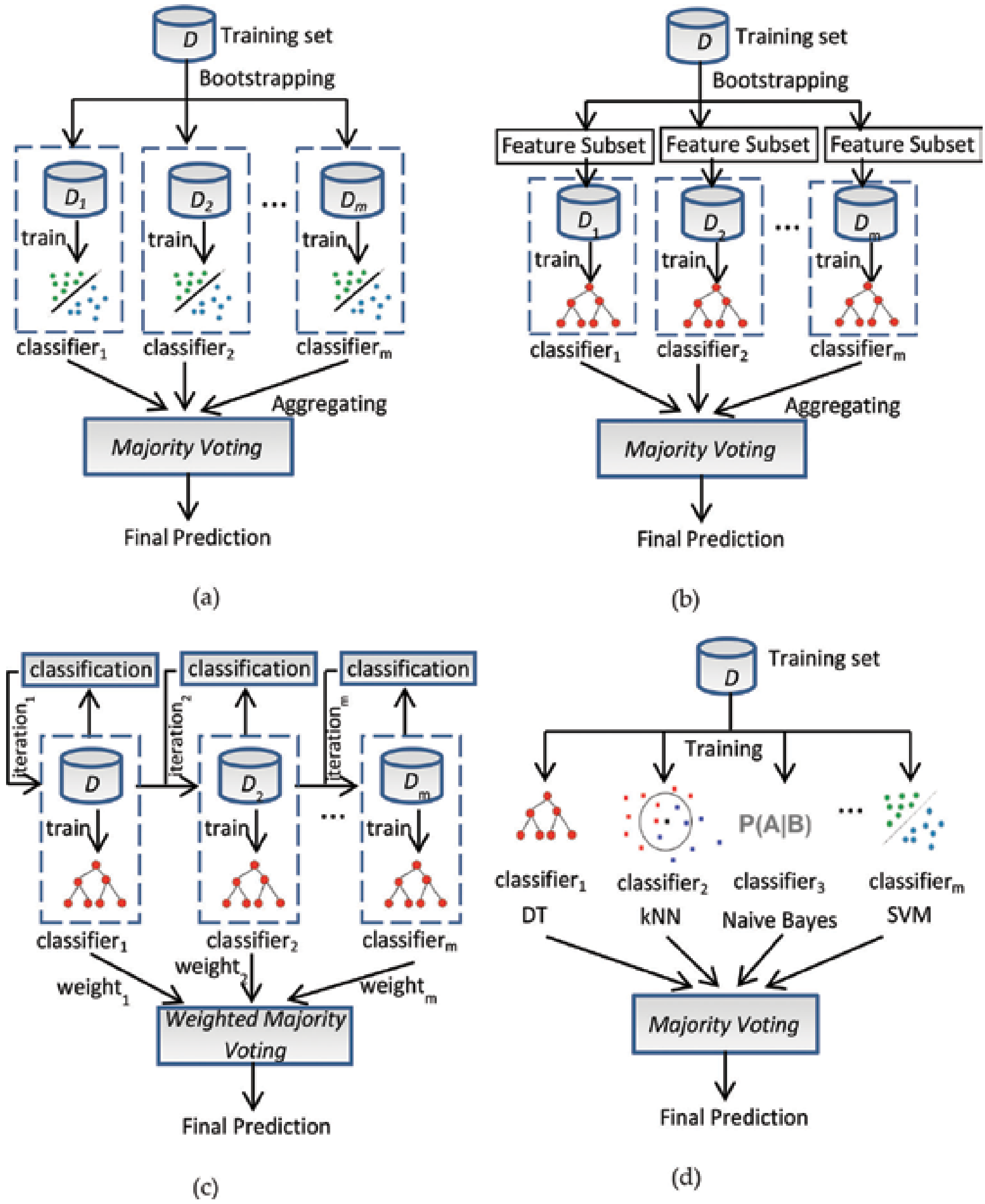


**Figure 2.** Different ensemble strategies: (a) bagging, (b) random forest, (c) AdaBoost, and (d) voting.

### 4.6. Challenges of ensemble learning in environmental data mining

Even ensemble-based environmental data mining is helpful based on the advantages indicated in Section 3; there are also challenges that could be overcome when you are aware. Challenges can be grouped under five main titles: selecting ensemble strategy, determining

| Algorithm | Training set | Classifiers | Combiner | Inducer | Ensemble strategy | Advantage | Weakness |
|---|---|---|---|---|---|---|---|
| Bagging | Random resampling | Inducer independent | Majority voting | Single inducer | 1 | Minimizes variance | A relatively large ensemble size—loss of cooperation with each other |
| Random forest | Random resampling + feature subset | Inducer dependent (decision tree) | Majority voting | Single inducer | 2 | | |
| Boosting | Weighted resampling | Inducer independent | Weighted majority voting | Single inducer | 3 | Boosts the performance of the weak learners | Degrades with noise |
| AdaBoost | Weighted resampling | Inducer independent | Weighted majority voting | Single inducer | 3 | | |
| Stacking | Resampling and k-folding | Inducer independent | Meta-learning | Multiinducer | 1, 4 | Good performance | Storage and time complexity |
| Grading | Resampling and k-folding | Inducer independent | Meta-learning | Multiinducer | 1, 4 | Predictions are graded | Storage and time complexity |
| Voting | Same dataset | Inducer independent | Majority voting | Multiinducer | 4 | Increase predictive accuracy | How classifiers are selected |
| Voting | Same dataset | Inducer independent | Majority voting | Single inducer | 4 | Simple to understand and implement | Limited to a single algorithm performance |

**Table 2.** Characteristic of different ensemble classifiers.

a satisfactory architecture, computational cost, complex nature of environmental data, and finally post processing:

- *Selecting ensemble strategy*: it is a difficult work to determine the best ensemble strategy in terms of accuracy, scalability, computational cost, usability, compactness, and speed of classification. Environmental researchers should know how to construct an ensemble model and be aware of alternative strategies and advantages/disadvantages of them. To overcome this problem, environmental data mining is mostly addressed to computer and environmental scientists working together.

- *Determining a satisfactory architecture*: there are two levels of problems in designing ensemble architecture. First, it is necessary to determine the optimal ensemble size. There are three approaches for determining the ensemble size: (i) preselection of the ensemble size, (ii) selection of the ensemble size while training, and (iii) postselection of the ensemble size (pruning). Second, how are learning algorithms and their respective parameters selected to construct the best ensemble? The best values for the input parameters of the algorithms should be determined through a number of tries. These problems are fundamentally different and should be solved separately to improve classification accuracy. Furthermore, it is necessary to update the model when new environmental data is acquired, allowing the up-to-date model to change over time.

- *Computational cost*: increasing the number of classifiers usually increases computational cost. To overcome this problem, users may predefine a suitable ensemble size limit, or classifiers can be trained in parallel.
- *Complex nature of environmental data*: it is necessary to deal with high dimensionality and complexity of environmental data. To reduce the dimensionality of the feature vector, feature selection techniques can be used such as principal component analysis, information gain, and ReliefF. Another problem is to deal with heterogeneous data by adding problem-specific science algorithms to the solution.
- *Post processing*: another critical issue is determining what the best voting mechanism (majority, weighted, average, etc.) for combining the outputs of base classifiers is. Furthermore, the final results should be presented in an appropriate form to help users understand and interpret easily.

## 5. Experimental study

In this study, different ensemble learning strategies were compared in terms of classification accuracy, precision (PRE), recall (REC), and f-measure (F-MEA). Four ensemble learning strategies were tested on six different real-world environmental datasets. The application was developed by using Weka open source data mining library.

### 5.1. Dataset description

In this experimental study, six different datasets that are available for public use were selected to determine the best ensemble strategy. Basic characteristics of the investigated environmental datasets are given in **Table 3**.

| ID | Dataset name | Year | Attributes | Instances | Type | Link |
|---|---|---|---|---|---|---|
| 1 | Ozone (1 h) | 2008 | 74 | 2536 | Air | http://archive.ics.uci.edu/ml/datasets/Ozone+Level+Detection |
| 2 | Ozone (8 h) | 2008 | 74 | 2534 | | |
| 3 | Leaf | 2014 | 17 | 340 | Ecology | http://archive.ics.uci.edu/ml/datasets/Leaf |
| 4 | Eucalyptus | 1991 | 20 | 736 | Soil | https://weka.wikispaces.com/Datasets |
| 5 | Forest type | 2015 | 28 | 523 | Ecology | https://archive.ics.uci.edu/ml/datasets/Forest+type+mapping |
| 6 | Cloud | 1971 | 8 | 108 | Rainfall | https://github.com/renatopp/arff-datasets/blob/master/statlib/nominal/cloud.arff |

**Table 3.** Environmental datasets and their characteristics.

### 5.2. Comparison of ensemble strategies

Classification accuracies, precision, recall, and f-measure values for the applied algorithms were obtained using tenfold cross validation. Comparison of the classification accuracies of the applied algorithms for each dataset is displayed in **Figure 3**. Four weak learners (support vector machine (SVM), naive Bayes (NB), decision tree (DT) applied with C4.5 algorithm, and K-nearest neighbor (KNN)) and four ensemble learners (bagging, random forest (RF), AdaBoost, and voting) were used to construct classification models from environmental data. The base classifiers for the ensemble learners were selected as the one which gave the best classification accuracy among the applied weak learners for the respective dataset.

The experimental results were obtained with optimum parameters (given in **Table 4**) using grid search. The best parameters of SVM were found for the complexity parameter, *C* for the exponent value, *E* for polykernel parameters in the interval [$10^k$ for $k \in \{-3, \ldots, 3\}$], and [1–10], respectively. To model DT, confidence factor, *C*, for pruning and the minimum number of objects, *M*, for leaf were obtained in the intervals of [0.05–0.95] and [1–10]. The number of neighbors, *N* for KNN classifier, was selected in the range of [1, 25]. For RF classifier, the number of randomly chosen attributes, *K*, and the number of iterations to be performed, *I*, were found in the intervals [0–15] and [10–100], respectively. The number of ensemble classifiers for bagging is 10 for each dataset. Weight threshold for weight pruning, *P*, and the number of iterations to be performed, *I*, were selected in the interval [10–100] for AdaBoost classifier. Voting was performed using the optimum parameters of SVM, NB, DT, KNN, and RF classifiers.

The objective of this experiment is to remark the success of the ensemble strategies in terms of classification accuracy concerning environmental data. According to the experimental results, it is apparent that the number of correctly classified instances is increased if ensemble strategies are applied. Especially, AdaBoost classifier provides significant performance gain compared to other models. SVM has superiority over other single learners; hence, most of the ensemble models selected it as the base learner.

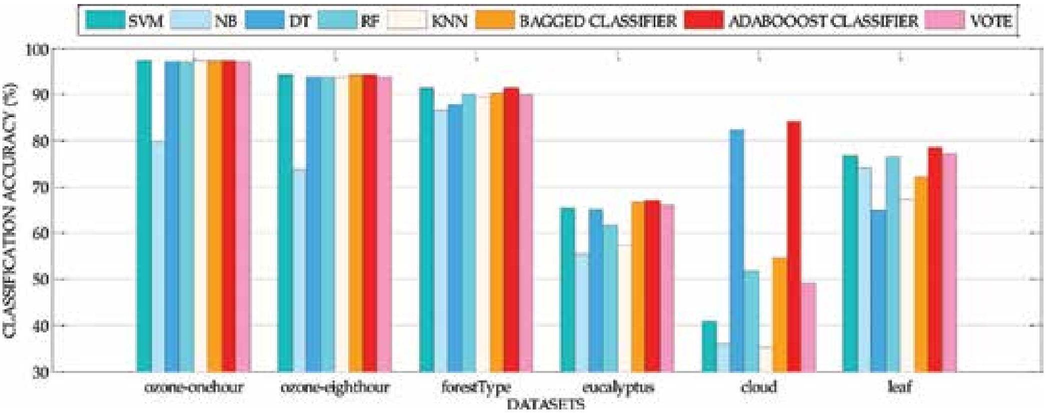


**Figure 3.** Comparison of single and ensemble classifiers in terms of classification accuracies.

| Dataset | SVM | | DT | | KNN | | RF | | AdaBoost | |
|---|---|---|---|---|---|---|---|---|---|---|
| | C | E | C | M | N | Distance metric | I | K | I | P |
| Ozone (1 h) | $10^3$ | 2 | 0.05 | 1 | 17 | Euclidean distance | 10 | 5 | 10 | 10 |
| Ozone (8 h) | $10^3$ | 5 | 0.55 | 1 | 5 | Chebyshev distance | 10 | 5 | 100 | 10 |
| Leaf | $10^0$ | 1 | 0.05 | 2 | 1 | Manhattan distance | 60 | 0 | 100 | 10 |
| Eucalyptus | $10^1$ | 1 | 0.15 | 2 | 9 | Manhattan distance | 50 | 2 | 80 | 40 |
| Forest type | $10^0$ | 1 | 0.15 | 3 | 11 | Manhattan distance | 50 | 11 | 10 | 10 |
| Cloud | $10^0$ | 1 | 0.05 | 1 | 17 | Euclidean distance | 100 | 0 | 100 | 40 |

**Table 4.** Optimum classifier parameters corresponding to each dataset.

There are a number of cases resulting in poor classification performance, such as the following:

- In case of the presence of either noisy or missing data
- If there is an insufficient number of instances available
- If there are too many number of classes
- If a complex relationship is inherent
- If the feature dependencies are ignored
- If the feature selection is not well performed
- If the algorithm parameters are not correctly determined
- If the class labels are imbalanced

For example, because the number of instances in "cloud" dataset is very few (due to the insufficient number of instances), inferior results are obtained for most of the applied algorithms as expected. However, even in such cases while some algorithms fail, some others manage to perform well (e.g., $C_{4.5}$ DT 82%). In this situation, the classifier's performance can also be enhanced by applying ensemble learning methods as in the case of AdaBoost with 84% classification accuracy for the same dataset. AdaBoost is a powerful ensemble learning algorithm because its distribution update step ensures that instances misclassified by the previous classifier are more likely to be included in the training data of the next classifier with the chance of further enhancement.

Due to the fact that classification accuracy as a performance metric is not just enough to decide whether a learner is considerably good or not, the precision, recall, and f-measure values were also calculated for each model (**Table 5**). It is also clear from the table values that applying ensemble strategies compared to single learners makes more sense in terms of classifier performance.

| Dataset | Algorithm | PRE | REC | F-MEA | Dataset | Algorithm | PRE | REC | F-MEA |
|---|---|---|---|---|---|---|---|---|---|
| Ozone (1-h) | SVM | 0.97 | 0.97 | 0.95 | Eeucalyptus | SVM | 0.65 | 0.65 | 0.65 |
| | NB | 0.96 | 0.79 | 0.86 | | NB | 0.62 | 0.55 | 0.55 |
| | $C_{4.5}$ DT | 0.94 | 0.97 | 0.95 | | $C_{4.5}$ DT | 0.66 | 0.65 | 0.64 |
| | RF | 0.94 | 0.97 | 0.95 | | RF | 0.61 | 0.61 | 0.61 |
| | K-NN | 0.97 | 0.97 | 0.95 | | K-NN | 0.57 | 0.57 | 0.56 |
| | K-$NN_{Bagged}$ | 0.97 | 0.97 | 0.95 | | $SVM_{Bagged}$ | 0.66 | 0.66 | 0.66 |
| | K-$NN_{AdaBoost}$ | 0.97 | 0.97 | 0.95 | | $SVM_{AdaBoost}$ | 0.67 | 0.67 | 0.67 |
| | Vote | 0.94 | 0.97 | 0.95 | | Vote | 0.67 | 0.65 | 0.65 |
| Ozone (8-h) | SVM | 0.93 | 0.94 | 0.93 | Cloud | SVM | 0.37 | 0.40 | 0.37 |
| | NB | 0.92 | 0.73 | 0.80 | | NB | 0.49 | 0.36 | 0.32 |
| | $C_{4.5}$ DT | 0.87 | 0.93 | 0.90 | | $C_{4.5}$ DT | 0.82 | 0.82 | 0.82 |
| | RF | 0.91 | 0.93 | 0.91 | | RF | 0.51 | 0.51 | 0.51 |
| | K-NN | 0.87 | 0.93 | 0.90 | | K-NN | 0.33 | 0.35 | 0.32 |
| | $SVM_{Bagged}$ | 0.92 | 0.94 | 0.93 | | $C_{4.5}$ $DT_{Bagged}$ | 0.55 | 0.54 | 0.54 |
| | $SVM_{AdaBoost}$ | 0.93 | 0.94 | 0.93 | | $C_{4.5}$ $DT_{AdaBoost}$ | 0.84 | 0.84 | 0.84 |
| | Vote | 0.93 | 0.93 | 0.91 | | Vote | 0.47 | 0.49 | 0.46 |
| Forest types | SVM | 0.91 | 0.91 | 0.91 | Leaf | SVM | 0.78 | 0.76 | 0.76 |
| | NB | 0.86 | 0.86 | 0.86 | | NB | 0.75 | 0.74 | 0.74 |
| | $C_{4.5}$ DT | 0.88 | 0.88 | 0.87 | | $C_{4.5}$ DT | 0.66 | 0.65 | 0.64 |
| | RF | 0.90 | 0.90 | 0.90 | | RF | 0.77 | 0.76 | 0.76 |
| | K-NN | 0.89 | 0.89 | 0.89 | | K-NN | 0.69 | 0.67 | 0.67 |
| | $SVM_{Bagged}$ | 0.90 | 0.90 | 0.90 | | $SVM_{Bagged}$ | 0.72 | 0.72 | 0.71 |
| | $SVM_{AdaBoost}$ | 0.91 | 0.91 | 0.91 | | $SVM_{AdaBoost}$ | 0.79 | 0.78 | 0.78 |
| | Vote | 0.90 | 0.90 | 0.90 | | Vote | 0.77 | 0.77 | 0.76 |

**Table 5.** Precision (PRE), recall (REC), and f-measure (F-MEA) results using tenfold cross validation for respective algorithms in each dataset.

## 6. Conclusion and future work

This study aims to provide helpful guidelines for future applications by presenting the advantages and challenges of ensemble-based environmental data mining and comparing alternative ensemble strategies through experimental studies. It compares four different ensemble strategies for environmental data mining: (i) bagging, (ii) bagging combined with random feature subset selection, (iii) boosting, and (iv) voting. In the experimental studies, ensemble methods are tested on different real-world environmental datasets.

In the future, the following studies can be carried out:

- Multistrategy ensemble learning that combines several ensemble strategies can be addressed, instead of a single ensemble strategy.
- Text mining, web mining, and process mining have been used in many engineering fields. However, there is very limited usage of them in environmental engineering. Future research can focus on these subjects.
- Some ontologies can be developed for environmental domain. We believe that the future environmental data mining studies will be supported by the ontologies to extract semantic relationships, to improve accuracy, and to develop better decision support systems.

## Author details

Goksu Tuysuzoglu[1], Derya Birant[2]* and Aysegul Pala[3]

*Address all correspondence to: derya@cs.deu.edu.tr

1 Graduate School of Natural and Applied Sciences, Dokuz Eylul University, Izmir, Turkey

2 Department of Computer Engineering, Dokuz Eylul University, Izmir, Turkey

3 Department of Environmental Engineering, Dokuz Eylul University, Izmir, Turkey

## References

[1] Stojić A, Stojić SS, Reljin I, Čabarkapa M, Šoštarić A, Perišić M, Mijić Z. Comprehensive analysis of PM10 in Belgrade urban area on the basis of long-term measurements. Environmental Science and Pollution Research. 2016;**23**:10722-10732. DOI: 10.1007/s11356-016-6266-4

[2] Srivastava AN. Greener aviation with virtual sensors: A case study. Data Mining and Knowledge Discovery. 2012;**24**:443-471. DOI: 10.1007/s10618-011-0240-z

[3] Al Abri ES, Edirisinghe EA, Nawadha A. Modelling ground-level ozone concentration using ensemble learning algorithms. In: Proceedings of the International Conference on Data Mining (DMIN'15); 27-30 July 2015; Las Vegas. USA: The Steering Committee of the World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp); 2015. pp. 148-154

[4] Bougoudis I, Demertzis K, Iliadis L. HISYCOL a hybrid computational intelligence system for combined machine learning: The case of air pollution modeling in Athens. Neural Computing and Applications. 2016;**27**:1191-1206. DOI: 10.1007/s00521-015-1927-7

[5] Lei KS, Wan F. Applying ensemble learning techniques to ANFIS for air pollution index prediction in Macau. In: International Symposium on Neural Networks (ISNN'12); 11-14 July 2012. Berlin, Heidelberg: Springer; 2012. pp. 509-516

[6] Singh KP, Gupta S, Rai P. Identifying pollution sources and predicting urban air quality using ensemble learning methods. Atmospheric Environment. 2013;**80**:426-437. DOI: 10.1016/j.atmosenv.2013.08.023

[7] Granata F, de Marinis G. Machine learning methods for wastewater hydraulics. Flow Measurement and Instrumentation. 2017;**57**:1-9. DOI: 10.1016/j.flowmeasinst.2017.08.004

[8] Budka M, Gabrys B, Ravagnan E. Robust predictive modelling of water pollution using biomarker data. Water Research. 2010;**44**:3294-3308. DOI: 10.1016/j.watres.2010.03.006

[9] Rodriguez-Galiano V, Mendes MP, Garcia-Soldado MJ, Chica-Olmo M, Ribeiro L. Predictive modeling of groundwater nitrate pollution using random Forest and multisource variables related to intrinsic and specific vulnerability: A case study in an agricultural setting (Southern Spain). Science of the Total Environment. 2014;**476**:189-206. DOI: 10.1016/j.scitotenv.2014.01.001

[10] Heung B, Hodúl M, Schmidt MG. Comparing the use of training data derived from legacy soil pits and soil survey polygons for mapping soil classes. Geoderma. 2017;**290**: 51-68. DOI: 10.1016/j.geoderma.2016.12.001

[11] Halmy MWA, Gessler PE. The application of ensemble techniques for land-cover classification in arid lands. International Journal of Remote Sensing. 2015;**36**:5613-5636. DOI: 10.1080/01431161.2015.1103915

[12] Wang Q, Xie Z, Li F. Using ensemble models to identify and apportion heavy metal pollution sources in agricultural soils on a local scale. Environmental Pollution. 2015;**206**: 227-235. DOI: 10.1016/j.envpol.2015.06.040

[13] Crimmins SM, Dobrowski SZ, Mynsberge AR. Evaluating ensemble forecasts of plant species distributions under climate change. Ecological Modelling. 2013;**266**:126-130. DOI: 10.1016/j.ecolmodel.2013.07.006

[14] Engler R, Waser LT, Zimmermann NE, Schaub M, Berdos S, Ginzler C, Psomas A. Combining ensemble modeling and remote sensing for mapping individual tree species at high spatial resolution. Forest Ecology and Management. 2013;**310**:64-73. DOI: 10.1016/j.foreco.2013.07.059

[15] Healey SP, Cohen WB, Yang Z, Brewer CK, Brooks EB, Gorelick N, et al. Mapping forest change using stacked generalization: An ensemble approach. Remote Sensing of Environment. 2018;**204**:717-728. DOI: 10.1016/j.rse.2017.09.029

[16] Gaál M, Moriondo M, Bindi M. Modelling the impact of climate change on the Hungarian wine regions using random forest. Applied Ecology and Environmental Research. 2012;**10**:121-140. DOI: 10.15666/aeer/1002_121140

[17] Nelson TA, Coops NC, Wulder MA, Perez L, Fitterer J, Powers R, Fontana F. Predicting climate change impacts to the Canadian Boreal forest. Diversity. 2014;**6**:133-157. DOI: 10.3390/d6010133

[18] Muñoz-Mas R, Lopez-Nicolas A, Martínez-Capel F, Pulido-Velazquez M. Shifts in the suitable habitat available for brown trout (*Salmo trutta* L.) under short-term climate change scenarios. Science of the Total Environment. 2016;**544**:686-700. DOI: 10.1016/j.scitotenv.2015.11.147

[19] Bravo-Moncayo L, Naranjo JL, García IP, Mosquera R. Neural based contingent valuation of road traffic noise. Transportation Research Part D: Transport and Environment. 2017;**50**:26-39. DOI: 10.1016/j.trd.2016.10.020

[20] Kühnlein M, Appelhans T, Thies B, Nauss T. Improving the accuracy of rainfall rates from optical satellite sensors with machine learning—A random forests-based approach applied to MSG SEVIRI. Remote Sensing of Environment. 2014;**141**:129-143. DOI: 10.1016/j.rse.2013.10.026

[21] Fan C, Xiao F, Wang S. Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques. Applied Energy. 2014;**127**:1-10. DOI: 10.1016/j.apenergy.2014.04.016

[22] Araya DB, Grolinger K, ElYamany HF, Capretz MA, Bitsuamlak G. An ensemble learning framework for anomaly detection in building energy consumption. Energy and Buildings. 2017;**144**:191-206. DOI: 10.1016/j.enbuild.2017.02.058

[23] Jovanović RŽ, Sretenović AA, Živković BD. Ensemble of various neural networks for prediction of heating energy consumption. Energy and Buildings. 2015;**94**:189-199. DOI: 10.1016/j.enbuild.2015.02.052

[24] Knudby A, Brenning A, LeDrew E. New approaches to modelling fish–habitat relationships. Ecological Modelling. 2010;**221**:503-511. DOI: 10.1016/j.ecolmodel.2009.11.008

[25] Kocev D, Džeroski S. Habitat modeling with single-and multi-target trees and ensembles. Ecological Informatics. 2013;**18**:79-92. DOI: 10.1016/j.ecoinf.2013.06.003

[26] Zhang Z, Ma C, Xu J, Huang J, Li L. A novel combinational forecasting model of dust storms based on rare classes classification algorithm. In Geo-Informatics in Resource Management and Sustainable Ecosystem (GRMSE'14); October 2014. Berlin, Heidelberg: Springer; 2015. pp. 520-537

[27] Mathanker SK, Weckler PR, Taylor RK, Fan G. AdaBoost and support vector machine classifiers for automatic weed control: Canola and Wheat. In: 2010 Pittsburgh, Pennsylvania, 20-23 June 2010; American Society of Agricultural and Biological Engineers. 2010. p. 1

[28] Lima AR, Cannon AJ, Hsieh WW. Nonlinear regression in environmental sciences by support vector machines combined with evolutionary strategy. Computers & Geosciences. 2013;**50**:136-144. DOI: 10.1016/j.cageo.2012.06.023

[29] Luo Q, Kathuria A. Modelling the response of wheat grain yield to climate change: A sensitivity analysis. Theoretical and Applied Climatology. 2013;**111**:173-182. DOI: 10.1007/s00704-012-0655-5

[30] Mohammed AA, Yaqub W, Aung Z. Probabilistic forecasting of solar power: An ensemble learning approach. Intelligent Decision Technologies. Smart Innovation, Systems and Technologies. 2015;**39**:449-458. DOI: 10.1007/978-3-319-19857-6_38

Chapter 2

# Estimating Customer Lifetime Value Using Machine Learning Techniques

Sien Chen

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.76990

**Abstract**

With the rapid development of civil aviation industry, high-quality customer resources have become a significant way to measure the competitiveness of the civil aviation industry. It is well known that the competition for high-value customers has become the core of airline profits. The research of airline customer lifetime value can help airlines identify high-value, medium-value and low-value travellers. What is more, the airline company can make resource allocation more rational, with the least resource investment for maximum profit return. However, the models that are used to calculate the value of customer life value remain controversial, and how to design a model that applies to airline company still needs to be explored. In the paper, the author proposed the optimised China Eastern Airlines passenger network value assessment model and examined its fitting degree with the TravelSky value score. Besides, the author combines China Eastern Airlines passenger network value assessment model score with loss model score to help airlines find their significant customers.

**Keywords:** customer lifetime value, estimating, machine learning

## 1. Introduction

In the context of customer relationship management, customer lifetime value (CLV) or customer equity (CE) becomes important because it is a disaggregate metric to evaluate marketing decisions [1], which can be utilised to allocate resources appropriately and identify profitable consumers [2]. Companies are looking forward to better approaches to create value and optimise their market offerings to appeal to customers and make profits [3]. Many firms are utilising CLV regularly to control and supervise the strategies of marketing as well as evaluate the business success. For companies, it is of interest to know how much net benefit it can expect from their customers. It is recognised that clv has become a significant component of companies' central strategy [4, 5]. CLV of customers at present and in the future can be a good proxy of the general corporate value [6]. Meanwhile, at each point in each customer's lifetime with the firm, the firm would like to form some expectation regarding the lifetime value of that customer.

## 2. Definition of CLV

Customer valuation has been discussed by several papers in the customer relationship management literature, for example, Dwyer [7], Berger and Nasr [8], Rust et al. [9] and Blattberg and Malthouse [10].

Dwyer [7] and Berger and Nasr [8] firstly provided a framework using the lifetime value of a customer. Then Gupta and his colleagues [6] found that the earnings of a company, and hence its value, are a function of the total customer lifetime value (CLV), defined as the discounted value of the future profits yielded by customers to the company, in other words, the value of a customer as the expected sum of discounted future earnings, where a customer generates a profit margin for each period. Moreover, a customer lifetime value (CLV) stands for the expected benefits' current value [7] and the equity of customer approaches to marketing [11, 12]. And CLV plays a major role in the marketing of the relationship [13]. The relationship with customers in the relationship marketing can be considered as the capital assets that require proper management [14].

## 3. Related work

In measuring customer lifetime value, a standard approach is to estimate the present value of the net benefit to the firm from the customer over time [1]. Researchers have suggested various methods to use customer-level data to measure the CLV [8, 9, 15–17]. However, the relationship between customer purchase behaviour and customer lifetime is not specific [15–19], if firms observed the customer defections, and longer customer lifetime implies higher customer lifetime value [20–22]. Different models for measuring CLV are different at estimates of the expectations of future customer purchase behaviour.

### 3.1. Methods of CLV prediction

#### *3.1.1. CLV model*

CLV is typically defined and estimated at an individual customer or segment level. This allows us to differentiate between customers who are more profitable than others rather than merely examining average profitability. The issue is to predict the future profits when the timing and the benefit of future transactions are unknown as discussed in Mulhern [23] and Bell et al. [24]. It is proposed by Gupta and other scholars [25] that CLV for a customer is [6, 19]:

$$\mathrm{CLV} = \sum\nolimits_{t=0}^{T} \frac{(P_t - C_t)r_t}{(1+i)^t} - AC \tag{1}$$

It is proposed by Gupta and other scholars [18] that CLV for a customer is [19, 36]:

where:

= price paid by a consumer at time t.

= direct cost of servicing the customer at time t.

= discount rate or cost of capital for the firm.

= probability of customer repeat buying or being 'alive' at time t.

AC = acquisition cost.

T = time horizon for estimating CLV.

Another review of CLV model sees Jain and Singh [26]. Linear regression with the variance that stabilises the transformation forecasted with the ordinary least square is the first approach. Selecting a stable variance transformation can be informed by residual plots [27]. As shown by Neter et al. [28], the linear regression forecasted with iteratively reweighted least square is the second approach of regression. IRLS is another means to solve the heteroscedasticity issue.

### *3.1.2. RFM model*

For the sake of simplicity, the only predictor variables in these models are the recency, frequency and monetary (RFM) type, Buckinx and Van den Poel [29], and the variables of RFM are sound predictors for CLV [15, 16].

The models of RFM have been utilised in direct marketing for three decades developed to target marketing programmes at specific customers with the objective to improve response rates. Studies show that customers' response rates vary the most by their recency, followed by their purchase frequency and monetary value [30]. Before these models, companies typically used demographic profiles of customers for targeting purposes. However, research strongly suggests that past purchases of consumers are better predictors of their future purchase behaviour than demographics.

They have many restrictions though RFM, or other models of scoring try to forecast customers' behaviour in the future and are therefore associated with CLV implicitly [15, 16, 31]. Firstly, in the next periods, the behaviour can be predicted by the models. However, to estimate CLV, we need to estimate customers' purchase behaviour not only in Period 2 but also in Periods 3, 4, 5 and so on. Secondly, RFM variables are real underlying behaviour's imperfect index stemmed from a real distribution. The models of RFM have neglected this part. Thirdly, the previous behaviour of customers can be an outcome of the company's previous marketing promotion, which has been ignored by the models. In spite of the restrictions, due to the implementation in real practice, the models of RFM are the core of the industry.

One fundamental limitation of RFM models is that they are scoring models and do not explicitly provide a number for customer value. However, RFM is essential past purchase variables that should be good predictors of future purchase behaviour of customers. Fader et al. [15, 16] showed how RFM variables could be used to build a CLV model that overcomes many of its limitations.

### 3.1.3. NBD-Pareto model

A popular method is the negative binomial distribution (NBD)-Pareto model introduced by Schmittlein et al. [32], which is referred by several authors [23, 26, 33] as a powerful technique to provide the situation where past customer purchase behaviour is used to predict the future probability of a customer remaining in business with the firm.

To forecast the CLV and integrate the transaction profits, some adoptions are conducted as the model of NBD-Pareto estimates the activity probability and the transaction number of a customer. Made by the NBD-Pareto for the forecast, an essential assumption refers to the independence between the relevant profit for every transaction and the transaction number of a customer. According to the prediction of a majority of papers, a two-step scheme to CLV modelling is being utilised by CLV [16, 17, 34]. Firstly, the transaction number of every person in the future will be forecasted. Subsequently, the mean profit for every transaction can be forecasted. At the level of customers, the values can be predicted. It generates a CLV approximation for every customer if the future transaction number and the mean profit for every transaction can be concluded.

In Fader and Hardie [15, 16], the maximum likelihood estimation (MLE) for an individual with purchase history is shown to describe the NBD-Pareto submodel. Utilising the approach of moments is an alternative to the MLE. However, similar results can be generated [19]. A person can forecast the transaction number that will be made by a customer in the future or predict the possibility for him or her to be alive when the parameters can be forecasted. As discussed by Schmittlein and Peterson [17], in the situation where customer lifetimes are observed, the NBD-Pareto model has limitations and is not suitable.

Another approach that can naturally incorporate past behavioural outcomes into future expectations is a Bayesian approach [35]. Bayesian approaches could integrate the previous data and information into the model's structure via the prior distribution of the CLV drivers.

### 3.1.4. Computer science models

The vast computer science literature in data mining, machine learning and nonparametric statistics has generated many approaches that emphasise predictive ability. These include projection-pursuit models; neural network models [36]; decision tree models; spline-based models such as generalised additive models, multivariate adaptive regression splines and classification and regression trees; and support vector machines. Lots of the methods might be more applicable to the research on the value in customers' lifetime.

In a recent study, Cui and Curry [37] conducted extensive Monte Carlo simulations to compare predictions based on multinomial logit model. Besides, Giuffrida et al. [38] reported that a multivariate decision tree induction algorithm outperformed a logit model in identifying the best customer targets for cross-selling purposes.

Due to the high focus that academics in marketing emphasise on interpretability and a parametric setup, these approaches remain little known in the marketing literature. However, given the importance of prediction in CLV, these methods need a closer look at the future.

#### *3.1.5. RFMc model*

The meaning of individual passenger value is calculating the traveller's particular value for the airline company based on the passenger's consumption data. It also refers to the passengers' profit contribution to the airline company.

Based on the characteristics of civil aviation, the fare discounts corresponding to class C are introduced to represent the level of value which passenger's consumption contributes to airlines. The RFMc model is proposed to calculate the civil aviation passengers' individual value, where R is the closeness coefficient of flight time, F is the total number of flights in a period of time and Mc is the passengers' relative total amount of flights calculated with the class of flight.

(1) Mc: the passengers' relative total amount of flights.

Calculate the total amount of relative consumer consumption Mc based on the fare weight of class c (corresponding fare discount); see formula (2):

$$M_c = \sum_{i=1}^{k} m_i * c_i \tag{2}$$

In the formula (2), $c_i$ represents the fare discount on the traveller's ith flight, $m_i$ is the fare of the traveller's ith plane, and k is the number of tickets purchased.

(2) R: the approximate coefficient of flight time.

The latest flight time t: the interval between the last flight time and the current time (the time when using the model to calculate the passenger's value).

The average turnaround time of flight $t_0$: the average of the two adjacent flights' time interval; see formula (3):

$$t_0 = \begin{cases} \sum_{i=1}^{n-1} t_i/(n-1) & n > 1 \\ t_s & n = 1 \end{cases} \tag{3}$$

In the formula (3), n is the gross number of passenger flights, $t_i$ is the passenger's flight time interval between ith and (i + 1)th, and $t_s$ is the average turnaround time of the precalculated whole passenger set.

The approximate coefficient of flight time R: the possibility that passengers take the plane again; see Eq. (4):

$$R = \begin{cases} 1 & t \leq t_0 \\ \frac{t_0}{t} & t > t_0 \end{cases} \tag{4}$$

The average flight turnaround time $t_0$ reflects the expectation of the interval between passengers' two contiguous flights. As the latest flight time t is less than or equal to the average turnaround time $t_0$, the value of R is 1; when t is greater than $t_0$, the possibility of passengers taking off again is gradually reduced, and R is slowly decreased.

(3) F: the passengers' flight frequency.

The passengers' flight frequency F reflects the activity and loyalty of passengers. It is acknowledged that the activity and loyalty affect the CLV to the airline company. The greater the take-off frequency, the higher the activity and loyalty degree, which can lead to the greater passenger's value to the airline. In general, the passengers' relative total amount of flights, the approximate coefficient of flight time and the passengers' flight frequency weighted sum, to obtain the passengers' value 'v', see Eq. (5):

$$v = \omega_1 M_c + \omega_2 R + \omega_3 F \tag{5}$$

In formula (5), $\omega_1$, $\omega_2$ and $\omega_3$ are each indicator's weight coefficients. Considering the different measurement of different indicators, Mc, R and F should be standardised and then weighted summation.

### *3.1.6. MRE model*

Passenger co-occurrence relationship includes the same order explicit co-occurrence relationship and different orders implicit coordination relationship. MRE multi-relational evaluation model combines order data and departure data, quantifies the explicit and implicit relationship between passengers and integrates time to make the comprehensive multi-relational evaluation.

(1) The same order co-occurrence relationship.

The same order co-occurrence relationship refers to the passenger relationship in the same order. The passenger's the same order relationship includes the number of passengers in the order, the difference between passenger class and order generation date. Based on PNR data to establish the whole passengers' same order relationship, use $P_{ij}$ to show the sequence of the same order relationship between passenger i and passenger j. $P_{ij}$ [k] = [∣ $c_i$ [k] -$c_j$ [k] ∣, s [k], $t_p$ [k]] is the kth record in the sequence, which indicates the data from the passenger i and passenger j's same order, where s [k] is the number of passengers of the order, $t_p$ [k] is the order generation date and $c_i$ [k] is the class of passenger i in the order (corresponding to the fare discount).

According to the sequence of the passenger's same order relation, passenger's same order relationship score is calculated. $P'_{ij}$ shows the total score of the same order relationship between passenger i and passenger j; see formula (6):

$$\mathrm{p}'_{\mathrm{ij}} = \Sigma_k s_P[k] = \Sigma_k \frac{1}{\sqrt{\mathrm{s}[k] \times \left(\left|c_i[k] - c_j[k]\right| + 1\right)}} \tag{6}$$

In the formula (6), $s_p$[k] is the score of the kth same order between passenger i and passenger j.

(2) Passenger company relationship

Company relationship: company relationship is defined by the author as the passenger-company relationships on the same flight which include coincidental company and appointed company. A co-occurrence relationship includes the date of flight departure, passenger seat distance, check-in sequence number distance, class rank difference and other attributes. According to the whole passengers' company relationship based on the departure data, $D_{ij}$ is denoted as the sequence of company relationship between passenger i and passenger j. $D_{ij}$ [k] = [| $d_{ci}$ [k] |, | $d_{seat}$ [k] |, | $d_{class}$ [k], $t_d$ [k]] is the kth record in the sequence, which represents the kth flight data of passenger i and passenger j when they fly together. Among these, $t_d$ [k] is the flight departure date, $d_{ci}$ [k] represents the check-in distance between passenger i and passenger j, $d_{seat}$ [k] represents the Euclidean distance between passenger i and passenger j's flight seats and $d_{class}$ [k] represents the class difference between passenger i and passenger j. According to the processed sequence of passenger-company relationship, the passenger-company relationship score can be calculated, where $D'_{ij}$ is used to show the total company relationship score of passenger i and passenger j, and the formula is given as

$$D'_{ij} = \sum_k S_d[k] \tag{7}$$

$$S_d[k] = \frac{\omega_1}{d_{ci}[k]} + \frac{\omega_2}{d_{seat}[k]} + \frac{\omega_3}{d_{class}[k] + 1} \tag{8}$$

In formulas (7) and (8), $\omega_1$, $\omega_2$ and $\omega_3$ are the impact factors of check-in sequence number distance, seat distance and class difference on passenger-company relationship score. $S_d$ [k] is the kth company relationship score between passenger i and traveller j.

(3) Time involved multi-relational comprehensive evaluation

Passenger value is unevenly distributed according to the edge weight. The scientific and accurate calculation of the edge weight directly affects the result of passenger value for the reason that the closer the passenger relationship is, the higher the value distributed. The RFM model predicts the possibility of customer repurchasing on the basis of customer consumption proximity R. Similarly, we also think that the civil aviation-passenger relationship is also connected with time: The passengers that fly together in the last few days are more likely to travel together again and have a closer relationship. In contrast, even if they have been together for many times, but no record of company in the past 2 years, we also have to consider whether the passenger relationship has disappeared. Due to the above considerations, we set the observation time window to observe the passenger relationship and bring in the time attenuation factor τ to make the passenger's relationship time perceptive. Assuming that the same last order (or same flight) of traveller i and traveller j is t, the time attenuation factor τ of the same order (or company) relationship between passenger i and passenger j can be expressed as

$$\tau = \frac{t - t'}{T - t'} \tag{9}$$

where **T–t** ′is the length of the observation time window, **T** is the end time of the time window and **t′** is the beginning time of the time window. If t ≤ t ′means that the passenger does not have the same order (or company) relationship in the observation time window, then the relationship is considered to disappear, and assume that $\tau = 0$. After introducing the time attenuation factor, the score of the same order passenger relationship can be expressed as formula (10), and the score of passenger-company relationship can be expressed as formula (11):

$$P'_{\mathrm{ij}} = \tau_{pij} \times \sum_{k} S_p[k] \tag{10}$$

$$D'_{\mathrm{ij}} = \tau_{Dij} \times \sum_{k} S_d[k]$$

where $\tau_{\mathrm{Pij}}$ is the time attenuation factor of the passenger i and the passenger j's same order relationship and $\tau_{\mathrm{Dij}}$ is the time attenuation factor of the same order relationship between the passenger i and the passenger j.

Standardise the passengers' company relationship score and the same order relationship score, and then weight and sum to get the total passenger relationship score. The formula is

$$\mathrm{W}_{ij} = \omega_{\mathrm{P}} P'_{\mathrm{ij}} + \omega_{\mathrm{d}} D'_{\mathrm{ij}} \tag{11}$$

where $\mathrm{W}_{\mathrm{ij}}$ represents the total score of the relationship between passenger i and passenger j, $\omega_{\mathrm{p}}$, $\omega_{\mathrm{d}}$, followed by the same order relationship weight and company relationship weight, $\omega_{\mathrm{p}} < \omega_{\mathrm{d}}$.

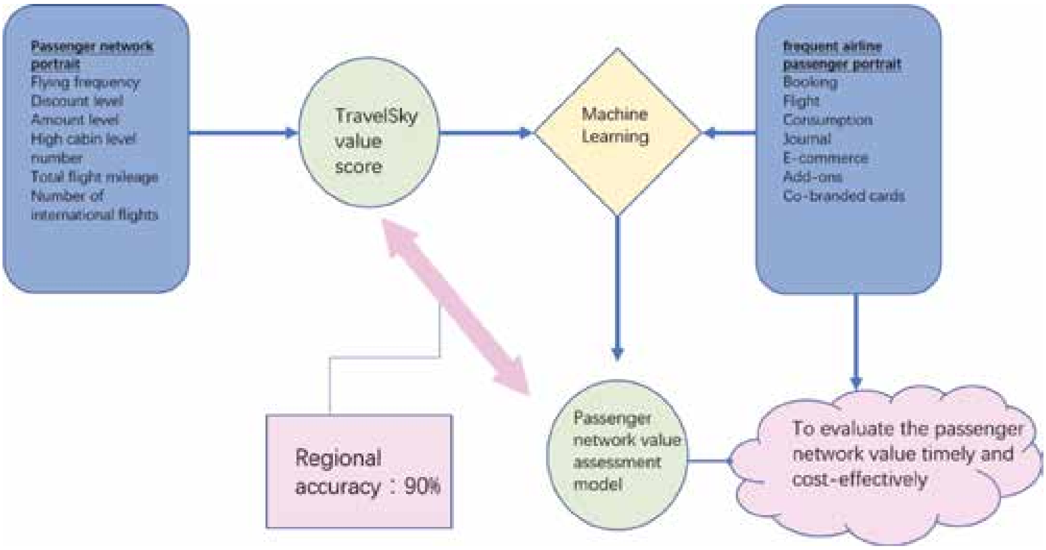
### 3.2. CLV prediction accuracy

Fit is the criterion suggested in the data-mining literature [39–41] for problems where the primary objective is making predictions that are as accurate as possible. As measures of prediction accuracy, Glady et al. [42] used the mean absolute error (MAE) and root mean square error (RMSE) between the actual value and the forecast of value in customers' lifetime. The 1% trimming can be used for the MAE and RMSE to enhance the strength to potential outliers in the set of data.

## 4. Passenger network value assessment model

### 4.1. Model description

Based on the dimensions of flying frequency, discount level, amount level, total flight mileage and number of international flights, etc. in the past year, TravelSky makes a comprehensive assessment on the value of passengers every month and form a scale of 0–100 value score. Which is called TravelSky Value Score. Passenger network value assessment based on the internal data of China Eastern Airlines, using airline frequent personal attributes and the airline's internal flight network's behaviour to estimate the TravelSky Value Score by using the advanced machine learning model. By fitting the TravelSky Value Score to the XGBoost model, a high fitting accuracy rate can be obtained, therefore helping the airline to evaluate the passenger network value timely and cost-effectively and to provide follow-up passenger segmentation and precision marketing services.



## 4.2. Data collection: frequent airline passenger portrait

First, collect data from relational database, log system, file system, document, picture, video, voice and other sources of different formats; analyse and identify the data. Then, focus on the business to identify and comprehend the information from the data. After that, extract valuable data fusion to the data platform. The dimension of frequent airline passenger portrait involves more than 300 variables including booking, flight, consumption, journal, e-commerce, add-ons and co-branded cards.

## 4.3. Passenger network value assessment model

### *4.3.1. The inputs of the model*

The input of the model is regarded as the relevant data or information which is used for computer processing. More specifically, in the process of the model application, input refers to the data of human and human behavioural characteristics. In the case of China Eastern Airlines, the inputs of passenger network value assessment model include 300+ variables, such as member current level, the highest consumption points in the last 3 months, the average delay time, how much changes of the air ticket endorsement, etc. However in general, the 300+ variables can be categorised into booking, flight, consumption, journal, e-commerce, add-ons and co-branded cards.

### *4.3.2. The outputs of the model*

The outputs of CEA passenger network value assessment model is estimated CEA passenger value score.

### *4.3.3. The mechanism of the model*

XGBoost is adopted as the mechanism in the paper.

(1) The introduction of XGBoost

XGBoost is a scalable machine learning system for tree boosting. The system is accessibly regarded as an open source package2.

XGBoost most prominent feature is that it can automatically use the CPU's multithreaded parallel while improving the algorithm to enhance the accuracy. Its debut is the Kaggle Higgs Sub Sign Recognition Contest, because of its superior efficiency and high predictive accuracy and it caught the attention of contestants in the competition forum.

(2) The Objective function of the optimisation model is

$$Ob_j(\theta) = L(\theta) + \Omega(\theta) \tag{12}$$

where $L(\theta)$ is error function which proves how well our model fits the data. $\Omega(\theta)$ is regularisation term, which is used to punish complex models [43].

The error function encourages the optimisation model to fit the training data, while the regularisation term helps the simpler model. Because when the model is simple, the randomness of the fitting degree of the finite data is relatively small and is not accessible to overfitting, making the prediction of the final model more stable.

The optimisation objective function in this case is

$$Ob_j(\theta) = \sum_i^n l\left(y_i, \widehat{y}_i\right) + \sum_{k=1}^K \Omega\left(f_k\right) \tag{13}$$

In this function, $\widehat{y}_i$ is estimated passenger network value score and, $\widehat{y}_i$ is TravelSky value score.

For more objective function derivation process, please refer to 《XGBoost: A Scalable Tree Boosting System》.

## 4.4. The performance of passenger network value assessment model

### *4.4.1. Model main parameters*

Tree depth, 6; step size, 0.1; maximum number of iterations, 66.

### *4.4.2. Model indicators*

rmse: Training Set 11.9455 and Test Set 13.02934.

$R^2$, 0.3939464.

### *4.4.3. The model main feature variables*

| **Feature** | | **Gain** | **Cover** | **Frequency** |
|---|---|---|---|---|
| wd_tk_bef_mean_hur_curr | Average advance booking time (next time window) (hours) | **0.186679089** | 0.055902665 | 0.013308573 |
| wd_tk_bk_next_nbr_curr | The number of future booking (the next time window) | **0.145693046** | 0.049101542 | 0.013618075 |
| wd_tk_bk_mean_intv_curr | Average booking time interval (next time window) | **0.050121893** | 0.05435901 | 0.023522129 |
| travel_max_intv_3m | The maximum flight time interval in the last 3 months | **0.045163104** | 0.019874175 | 0.007118539 |
| wd_tk_bk_mean_intv_prev | Average booking time interval (last time window) | **0.028847336** | 0.008726862 | 0.013308573 |
| ap_income_channel_3m.非航累积 | Accumulate the main channel of the last 3 months: non-flight accumulation | **0.023556519** | 0.015727313 | 0.008047044 |
| ap_income_sum_1y | The total number of points accumulated in the most recent year | **0.018971991** | 0.028146008 | 0.010523058 |
| ticket_bef_max_intv_1y | The largest number of days in advance tickets in the latest year | **0.014451596** | 0.030321255 | 0.009285051 |
| wd_tr_zj_curr | Average early check-in time (the next time window) | **0.01381261** | 0.02388794 | 0.006499536 |
| wd_tk_bk_nbr_curr | Booking times (next time window) | **0.013618875** | 0.01638353 | 0.006499536 |

### *4.4.4. Distribution of TravelSky value and forecast value*

Separately observe their scores, and it can be seen that the scores are all concentrated in the high segment. In particular, 63.13% of the passengers get 100 TravelSky Value scores.

| **Summation items: the number of people** | |
|---|---|
| TravelSky value | **Summary** |
| 0 | 1.04% |
| 91 | 0.36% |
| 92 | 0.44% |
| 93 | 0.61% |
| 94 | 0.82% |
| 95 | 1.17% |
| 96 | 1.82% |
| 97 | 3.14% |
| 98 | 6.35% |
| 99 | 16.56% |
| 100 | 63.13% |
| **Total** | **100.00%** |

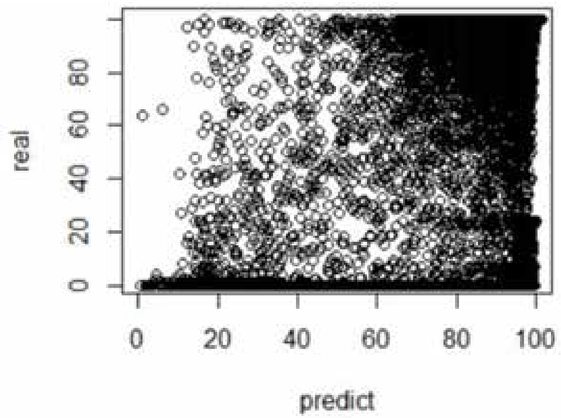
| Summation items: the number of people | |
|---|---|
| forecast value **(rounding)** | **Summary** |
| 0 | 0.00% |
| 91 | 1.08% |
| 92 | 1.40% |
| 93 | 1.94% |
| 94 | 3.03% |
| 95 | 4.58% |
| 96 | 6.66% |
| 97 | 15.69% |
| 98 | 49.32% |
| 99 | 9.08% |
| 100 | 0.07% |
| **Total** | **100.00%** |

## 4.5. Model evaluation report: TravelSky value fit report

Using more than 300 features of CEA loss model and 240,000 passenger data of loss model, the TravelSky value score is fitted to the Xgboost model [44, 45].

### *4.5.1. Cross-contrast the TravelSky value score and the forecast value*

Cross-contrast the TravelSky value score with the forecast value; visualise the data and present it in the form of the charts below.



PivotTable: The horizontal axis represents the 10-point range where the value score fits with CEA data. For example, 1 indicates [0, 10], 2 indicates [11, 20], and similarly, 10 indicates

(90,100). The vertical axis represents the 10-point interval in which the avionics value score is located.

| **Summation items: the number of people** | **Predicted value (divided by 10 and rounded)** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| TravelSky value **(divided by 10 and rounded)** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **Total** |
| 1 | 0.04% | 0.25% | 0.14% | 0.09% | 0.08% | 0.08% | 0.11% | 0.14% | 0.23% | 0.68% | 1.84% |
| 2 | 0.00% | 0.01% | 0.01% | 0.01% | 0.01% | 0.01% | 0.01% | 0.01% | 0.02% | 0.15% | 0.26% |
| 3 | 0.00% | 0.01% | 0.01% | 0.01% | 0.01% | 0.01% | 0.01% | 0.02% | 0.04% | 0.24% | 0.37% |
| 4 | 0.00% | 0.01% | 0.01% | 0.01% | 0.01% | 0.01% | 0.01% | 0.01% | 0.03% | 0.05% | 0.16% |
| 5 | 0.00% | 0.01% | 0.01% | 0.01% | 0.01% | 0.01% | 0.02% | 0.03% | 0.04% | 0.07% | 0.21% |
| 6 | 0.00% | 0.01% | 0.01% | 0.01% | 0.01% | 0.01% | 0.02% | 0.03% | 0.06% | 0.10% | 0.26% |
| 7 | 0.00% | 0.00% | 0.01% | 0.01% | 0.01% | 0.02% | 0.02% | 0.04% | 0.08% | 0.15% | 0.34% |
| 8 | 0.00% | 0.01% | 0.01% | 0.01% | 0.01% | 0.02% | 0.03% | 0.08% | 0.13% | 0.33% | 0.62% |
| 9 | 0.00% | 0.01% | 0.01% | 0.01% | 0.01% | 0.03% | 0.06% | 0.12% | 0.34% | 0.96% | 1.56% |
| 10 | 0.00% | 0.01% | 0.01% | 0.02% | 0.03% | 0.07% | 0.19% | 0.62% | 3.33% | 90.11% | 94.39% |
| **Total** | **0.04%** | **0.32%** | **0.23%** | **0.18%** | **0.20%** | **0.27%** | **0.50%** | **1.11%** | **4.29%** | **92.85%** | **100.00%** |

Evaluation criteria:

1. Calculate the accuracy of 10-point interval: 90.64%.
2. As the forecast scores are mainly concentrated around 98 points, the proportion of people between 0 and 90 points is relatively low which belongs to the low-value area. Therefore, the author will divide '0–90 points' into a category. Using the 400,000 senior frequent passengers of China Eastern Airlines's portrait features to fit the TravelSky value score, the accuracy of the evaluation is up to 90% with ten-point interval. 92.85% of the passenger network value assessment model (CEA model) is located in the 91–100 value range. Ninety-seven percent (90.11%/92.85% = 97%) of the TravelSky value score is also located in the [91,100] value range (as shown in the following table).

| **The proportion of the population** | | **CEA value** | | **Total** |
|---|---|---|---|---|
| | | **[0, 90]** | **[91, 100]** | |
| **Travel Sky** | **[0, 90]** | 2.87% | 2.74% | 5.61% |
| | **[91, 100]** | 4.28% | 90.11% | 94.39% |
| **Total** | | 7.15% | 92.85% | 100.00% |

### 4.6. Module application

Based on the accuracy of the passenger network value assessment model combined with the prediction of passenger loss probability in the next 6 months, it is necessary to give priority to reach the target of 'high network value and high risk of loss in the next 6 months' passenger groups. Thus, it can help marketing accurate positioning.

*4.6.1. Passenger loss model*

**A.** Definition of loss: The number of flight phase in the next 6 months is at least decreasing ten absolute flight phases or reducing 50%.

**B.** The loss model Xgboost main features.
Summary (model).

Importance of features in the XGBoost model.

| Feature | | Gain | Cover | Frequency |
|---|---|---|---|---|
| wd_tr_12h_dep_curr | Delay [1, 2] Number of flight phase (next time window) | 0.085744211 | 0.060886127 | 0.017437145 |
| wd_tr_24h_dep_series_c | Delay [2,4] Number of flight phase (how many changes) | 0.064961789 | 0.032607435 | 0.010948905 |
| wd_tr_24h_dep_series_b | Delay [2,4] Number of flight phase (whether changed) | 0.054268923 | 0.040321268 | 0.01216545 |
| deploy_arr_mean_tm_3m | Average delay of flight arrival in the last 3 months (in minutes) | 0.048203621 | 0.034246522 | 0.016626115 |
| wd_tr_y_nbr_curr | Economy class travel flight phase number (next time window) | 0.044511041 | 0.057238477 | 0.01297648 |
| travel_f_max_intv_3m | The latest 3-month maximum flying time interval | 0.040740831 | 0.047666125 | 0.01865369 |
| deploy_1h_nbr_3m | The number of flight phases which flight delays of 1 hour in the last 3 months | 0.018180959 | 0.014574099 | 0.01540957 |
| wd_pt_aft_lvl_labels_b.银卡 | Member level (end) (from A to B): silver card | 0.016429216 | 0.017211579 | 0.004460665 |
| deploy_arr_mean_tm_1y | Average delay of flight arrival in the most recent year (in minutes) | 0.015352698 | 0.008028182 | 0.00729927 |
| wd_tk_bk_next_nbr_prev | The number of future booking (the last time window) | 0.014210402 | 0.010862506 | 0.01054339 |
| y_hd_cnt_3m | The last 3-month economy class flight phase number | 0.011647629 | 0.01455673 | 0.01865369 |
| wd_tr_dpt_mean_dep_curr | Average delay time (departure, minute) (next time window) | 0.010571347 | 0.01455938 | 0.00486618 |
| wd_pt_aft_lvl_labels_b.小飞人 | Member level (end) (from A to B): supermaster | 0.010329831 | 0.011708833 | 0.0162206 |
| deploy_dpt_mean_tm_3m | Average take-off delay in the last 3 months of flight (in minutes) | 0.01016305 | 0.013088575 | 0.01540957 |
| y_hd_cnt_6m | The last 6-month economy class travel flight phase number | 0.010010692 | 0.004976274 | 0.01459854 |
| wd_tk_sum_amt_curr | Total booking amount (next time window) | 0.009631331 | 0.003969202 | 0.0081103 |
| travel_f_mean_intv_1y | | 0.009385889 | 0.020736252 | 0.011759935 |

| Feature | | Gain | Cover | Frequency |
|---|---|---|---|---|
| | First-class average flying interval in the latest year | | | |
| deploy_dpt_max_tm_3m | The maximum time of take-off delay in the last 3 months (unit: minutes) | 0.009230547 | 0.007903106 | 0.0081103 |
| y_hd_cnt_1y | The number of economy class flight phase in the latest year | 0.008812939 | 0.003531476 | 0.011759935 |
| wd_tr_f_nbr_curr | First-class flight phase number (next time window) | 0.008393582 | 0.019247921 | 0.01135442 |

The resulted model fits the entire dataset, and the relative importance of each variate can be viewed by importance_xgb () or simpler summary () as above.

### *4.6.2. Loss model score combines the forecast value score to select the key population*

| **Summation items: the number of people** | **Estimated loss rate (10% interval)** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Forecast value (divided by ten and rounded)** | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **Total** |
| 1 | 0.00% | 0.00% | 0.01% | 0.01% | 0.01% | 0.01% | 0.00% | 0.00% | 0.00% | 0.00% | 0.04% |
| 2 | 0.00% | 0.04% | 0.07% | 0.09% | 0.06% | 0.04% | 0.02% | 0.00% | 0.00% | 0.00% | 0.32% |
| 3 | 0.00% | 0.03% | 0.04% | 0.05% | 0.05% | 0.03% | 0.01% | 0.01% | 0.00% | 0.00% | 0.23% |
| 4 | 0.00% | 0.02% | 0.04% | 0.04% | 0.03% | 0.03% | 0.01% | 0.00% | 0.00% | 0.00% | 0.18% |
| 5 | 0.00% | 0.02% | 0.03% | 0.04% | 0.04% | 0.03% | 0.02% | 0.01% | 0.00% | 0.00% | 0.20% |
| 6 | 0.00% | 0.02% | 0.05% | 0.06% | 0.05% | 0.05% | 0.03% | 0.01% | 0.00% | 0.00% | 0.27% |
| 7 | 0.00% | 0.04% | 0.10% | 0.11% | 0.10% | 0.08% | 0.04% | 0.02% | 0.01% | 0.00% | 0.50% |
| 8 | 0.00% | 0.08% | 0.21% | 0.25% | 0.23% | 0.17% | 0.11% | 0.04% | 0.01% | 0.00% | 1.11% |
| 9 | 0.01% | 0.31% | 0.73% | 0.98% | 0.93% | 0.71% | 0.41% | 0.17% | 0.04% | 0.01% | 4.29% |
| 10 | 0.24% | 4.26% | 12.40% | 18.59% | 20.64% | 18.18% | 12.02% | 5.32% | 1.13% | 0.07% | 92.85% |
| **Total** | **0.26%** | **4.82%** | **13.68%** | **20.24%** | **22.14%** | **19.31%** | **12.67%** | **5.59%** | **1.21%** | **0.08%** | **100.00%** |

An orange group means that both high loss scores (high likelihood of loss in the next 6 months) and high-value scores (up to 90 points) fit well with TravelSky value score.

## 5. Conclusion

In this paper, the author first described the definition of customer lifetime value (CLV) and demonstrated the approach to estimating customer lifetime value by proposing various customer lifetime value models and illustrating the criterion to predict customer lifetime value accuracy. The aim is to provide the theoretical basis for the airline customer lifetime value estimation research. After that, a numeral case of China Eastern Airlines was given to show the practicability and veracity of China Eastern Airlines passenger network value assessment model with assessing their fitting accuracy rate with the TravelSky value score. The ambition is combining forecast value score calculated by China Eastern Airlines passenger network value assessment model with loss model score to select the critical population.

## Author details

Sien Chen[1,2,3*]

*Address all correspondence to: sien.chen@postgrad.manchester.ac.uk

1 Institute of Internet Industry, Tsinghua University, Beijing, China

2 Alliance Manchester Business School, University of Manchester, Manchester, UK

3 Antai College of Economics and Management, Shanghai Jiao Tong University, Shanghai, China

## References

[1] Blattberg RC, Deighton J. Manage marketing by the customer equity test. Harvard Business Review. 1996;(July-August):136-144

[2] Kumar V, Venkatesan R, Reinartz W. Knowing what to sell, when, and to whom. Harvard Business Review. 2006;**84**(3):131-137

[3] Bendapudi N, Leone R. Psychological implications of customer participation in co-production. Journal of Marketing. 2003;**67**(1):14-28

[4] DeSarbo W, Jedidi K, Sinha I. Customer value analysis in a heterogeneous market. Strategic Management Journal. 2001;**22**(9):845-857

[5] Porter M. Clusters and the new economics of competition. 76th ed. Boston: Harvard Business Review. 1998. pp. 77-90

[6] Gupta S, Lehmannand DR. Valuing customers. Journal of Marketing Research. 2004;**41**(1):7-18

[7] Dwyer FR. Customer lifetime valuation to support marketing decision making. Journal of Direct Marketing. 1997;**11**(4):6-13

[8] Berger PD, Nasr NI. Customer lifetime value: Marketing models and applications. Journal of Interactive Marketing. 1998;**12**(1):17-30

[9] Rust RT, Lemon KN, Zeithaml VN. Return on marketing: Using customer equity to focus marketing strategy. Journal of Marketing. 2004;**68**(1):109-127

[10] Blattberg EC, Malthouse FJ. Can we predict customer lifetime value? Journal of Interactive Marketing. 2005;**19**(1):2-16

[11] Rust RT, Zeithaml VA, Lemon KN. Driving Customer Equity: How Customer Lifetime Value Is Reshaping Corporate Strategy. New York: Free Press; 2000

[12] Blattberg RC, Getz G, Thomas JS. Customer Equity: Building and Managing Relationships As Valuable Assets. Boston: Harvard Business School Press; 2001

[13] Sheth J, Mittal B, Newman B. Consumer Behavior and Beyond. NY: Harcourt Brace; 1999

[14] Hennig-Thurau T, Hansen U. Relationship Marketing-Some Reflections on the State-of-the-Art of the Relational Concept. In: Hennig-Thurau T, Hansen U, editors. Relationship Marketing: Gaining Competitive Advantage Through Customer Satisfaction and Customer Retention. New York: Springer; 2000. pp. 3-27

[15] Fader PS, Hardie BGS, Lee KL. "Counting your customers" the easy way: An alternative to the Pareto/NBD model. Marketing Science. 2005;**24**(2):275-284

[16] Fader PS, Hardie BGS, Lee KL. RFM and CLV: Using CLV curves for customer base analysis. Journal of Marketing Research. 2005;**42**(4):415-430

[17] Schmittlein DC, Peterson RA. Customer base analysis: An industrial purchase process application. Marketing Science. 1994;**13**(1):41-67

[18] Reinartz WJ, Kumar V. The impact of customer relationship characteristics on profitable lifetime duration. Journal of Marketing. 2003;**67**(1):77-99

[19] Reinartz WJ, Kumar V. On the profitability of long-life customers in a noncontractual setting: An empirical investigation and implications for marketing. Journal of Marketing. 2000;**64**:17-35

[20] Bhattacharya CB. When customers are members: Customer retention in paid membership contexts. Journal of Academy of Marketing Science. 1998;**26**(1):31-44

[21] Bolton RN. A dynamic model of the duration of the customer's relationship with a continuous service provider: The role of satisfaction. Marketing Science. 1998;**17**(1):45-65

[22] Thomas JS. A methodology for linking customer acquisition to customer retention. Journal of Marketing Research. 2001;**38**(May):262-268

[23] Mulhern FJ. Customer profltability analysis: Measurement, concentrations, and research directions. Journal of Interactive Marketing. 1999;**13**(1):25-40

[24] Bell D, Deighton J, Reinartz J, Rust R, Swartz G. Seven barriers to customer equity management. Journal of Service Research. 2002;**5**(1):77-85

[25] Gupta S, Lehmann DR. Customers as assets. Journal of Interactive Marketing. 2003;**17**(1):9-24

[26] Jain D, Singh SS. Customer lifetime values research in marketing: A review and future directions. Journal of Interactive Marketing, 2002;**16**(2):34-46

[27] Cook RD, Weisberg S. Residuals and Influence in Regression. New York: Chapman and Hall; 1982

[28] Neter J, Kutner M, Nachtsheim C, Wasserman W. Applied Linear Statistical Models, 4th ed. Chicago: Irwin; 1996

[29] Buckinx W, Van den Poel D. Customer base analysis: Partial defection of behaviorally-loyal clients in a non-contractual fmcg retail setting. European Journal of Operational Research. 2005;**164**(1):252-268

[30] Hughes A. Strategic Database Marketing, 3rd ed. New York: McGraw-Hill; 2005

[31] Kumar V. CLV: A Path to Higher Profitability. Working Paper. Storrs: University of Connecticut; 2006

[32] Schmittlein DC, Morrison DG, Colombo R. Counting your customers: Who are they and what will they do next? Management Science. 1987;**33**(1):1-24

[33] Niraj R, Gupta M, Narasimhan C. Customer profltability in a supply chain. Journal of Marketing. 2001;**65**(3):1-16

[34] Venkatesan R, Kumar V. A customer lifetime value framework for customer selection and resource allocation strategy. Journal of Marketing. 2004;**68**(4):106-125

[35] Rossi PE, Allenby GM. Bayesian statistics and marketing. Marketing Science. 2003;**22**(3):304-328

[36] Venables W, Ripley B. Modern Applied Statistics with S-PLUS. New York: Springer; 1999

[37] Cui D, Curry D. Prediction in marketing using the support vector machine. Marketing Science. 2005;**24**(Fall):595-615

[38] Giuffrida G, Chu W, Hanssens D. Mining Classification Rules from Datasets with Large Number of Many-Valued Attributes. Computer Science. Vol. 1777. Berlin: Heidelberg; 2000. pp. 335-349

[39] Breiman L. Statistical modeling: The two cultures. Statistical Science. 2001;**16**(3):199-231

[40] Breiman L. Heuristics of instability and stabilization in model selection. Annals of Statistics. 1996;**24**(6):2350-2383

[41] Hastie T, Tibshirani RJ, Friedman JF. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer; 2001

[42] Glady N, Baesens B, Croux C. A modified Pareto/NBD approach for predicting customer lifetime value. Expert Systems with Applications. 2009;**36**(2):2062-2071

[43] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining ACM, 2016. pp. 785-794

[44] Gupta S, Lehmann D, Stuart J. Valuing customers. Journal of Marketing Research. 2004;**41**(1):7-18

[45] Malthouse E, Blattberg R. Can we predict customer lifetime value?. Journal of Interactive Marketing. 2005;**19**(1):2-16

Chapter 3

# Determination and Classification of Crew Productivity with Data Mining Methods

Abdullah Emre Keleş and Mümine Kaya Keleş

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.75504

**Abstract**

Turkey is a developing country and the main axis of development is "construction." The construction sector is in a position to create demand for goods and services produced by more than 200 subsectors, and this widespread impact is the most basic indicator of the sector's "locomotive of the economy." In the development of the construction industry, crew productivity plays a very important role. While businesses that do not measure their employees' needs, their locations, and so on are suffering from various losses, rare businesses that take these parameters into account can profit. The identification of leadership types that will motivate employees has great importance in terms of construction businesses where the human element is the foreground. For this purpose, in the province of Adana, the relationship of productivity between the engineers working in construction companies and workers who work at lower departments of these engineers was examined. In this study, bidirectional multiple leadership questionnaire (MLQ) was applied to construction site managers and employees, and according to this survey data, leadership and motivations/productivities were classified using data mining methods. According to the classification analysis results, the most successful data mining algorithm was random forest algorithm with a rate of 81.3725%.

**Keywords:** classification, construction information, construction management, crew productivity, data mining, random forest algorithm, supervised and unsupervised learning, Weka

## 1. Introduction

With the increasing globalization in the construction sector, institutionalization is at the forefront. In addition, under increasing competition conditions, construction companies are forced to differentiate in the methods and technologies they use in business processes in order to be able to share in the sectoral market and to protect and strengthen it [1]. When the construction works are examined in terms of their management functions, the result is that they are still largely human-focused. This situation pushes construction companies to survive in a highly competitive environment and to take positions in construction businesses, especially by increasing their loyalty to operating by transferring employees to business processes with greater motivation, and thus to obtain more efficiency. In short, the human factor, one of the inputs used in construction production, must be managed correctly. Businesses have to deal with systematic approaches through the management of employees.

When the subject of this study and similar studies are examined, it can be seen that the studies related to construction project management give different and highly motivated results in the regions where the study is conducted, but they can give different results when the same study is conducted in another region. These researches have been inspired to make this study. The situation in Adana province, the relationships between the leadership styles and the motivations of employees are determined and only the results based on this province are taken into consideration in terms of the enterprises and employees working on this province.

The determination of the productivity of workers in the construction sector is directly related to the success of the enterprises. In the face of increasing competition, the businesses that do not measure the needs, locations, activities, and so on of their employees are also undergoing various losses, even if they are unaware.

In the light of all these, the main objective of this study is to determine the relationship between the civil engineers who are in managerial leadership position in the construction enterprises operating in Adana province and the subordinates of the master-worker-headman positions they work with in the leadership-motivation axis. In line with this aim, 100 construction companies were selected to conduct construction projects in Adana province and two questionnaires were applied on two sides, one for the leader and one for the employee.

For this purpose, the productivity relationship between the persons who are engineers work in the city of Adana, the ones who produce the building, the ones who work in the construction companies and workers who work in the subhierarchy of these engineers were examined. The identification of leadership types that will motivate and support employees has a great importance in terms of construction businesses where the human element is the fore ground. From the point of view of the construction site managers in charge of the sites, it is thought that it will be useful for the sector representatives, businesses and all employees to determine which leader type will motivate which employees. In this context, association rule mining were made with Apriori Algorithm using data mining methods using Weka [2] and Keel [3] software on the data obtained by the multiple leadership questionnaire (MLQ) applied bidirectionally to site chiefs and employees, and leadership and motivations/efficiencies were classified by using classification algorithms. The impact of leadership styles on employee motivation/productivity has been analyzed. Thus, it is aimed to present the creation of the most suitable rules that can be used in the field of engineer leadership-employee motivation/productivity of the construction companies in Adana province and to present them to the sector.

The results obtained by this method are analyzed with these data mining methods, and they are given and interpreted in "Findings and Discussion" and "Conclusion" sections of this study.

## 2. Literature review

When the literature on the topic is considered, it is determined that there are deficiencies in the "Systematic Leadership Approach in the Construction Sector". It has also been determined that there is not enough consciousness about what employee productivity will be when applying which leadership styles. At the same time, it has been determined that there are deficiencies in how to increase employee productivity on a sectoral basis. Even if the number of works based on the construction sector in the field of employee productivity has increased in recent years, there is little research to measure the productivity by evaluating both sides and the subordinate relationship. It is believed that it is important to provide a systematic productivity analysis to the workers in the construction sector by eliminating this deficiency. Such a systematic development and submission to the use of sector representatives may lead to the deficiencies mentioned earlier and may also provide guidance.

Unfortunately, too many sources were not found when the literature review was conducted in this area. Kaya and his colleagues [4] tried to estimate the productivity values with the help of data obtained from the survey of ceramic workers working in construction companies. These values, which are also used as attributes in the related study, are the number of teams, the work experience and the average age of the people on the crew. In their study, it was focused on how to achieve high productivity in ceramic works by means of mining rules by classifying productivity with the ceramic data obtained by using measurement methods. Andaç and Oral [5] presented the results obtained from their work using the Artificial Bee Colony Algorithm for estimating worker/labor productivity. In the work of Keleş and Kaya [6], they used the association rules and a data mining method, in order to increase employee productivity with the demographic information obtained from the workers working as wall masters/mason in the construction sector. In Keles's PhD thesis [7] "Determining the relationship between leadership of the site managers and motivation of their employees with the data mining in construction projects," a double-sided survey has been applied to site managers and their subordinates who work in the construction sites in Turkey. Following this conducted questionnaire, leadership was identified from two different perspectives. After determining the leadership models, the relationship between the leadership of the site chiefs and the motivations of the workers was determined by using the association rule mining method from the data mining methods.

## 3. Materials and methods

In this chapter, the changes of the productivity of the other employees are discussed according to the leadership models of the construction engineers working as construction engineers. In this respect, data will be obtained from the point of view of the site managers and employees. In other words, it will be determined that, productivity will increase as a result of which modern leadership types—transactional/transformational/passive avoidant leadership behaviors—in the literature studies especially in recent years, are applied to the construction worker/employee group in which characteristics.

In this study, as a method of obtaining data, questionnaires were used to provide "bidirectional evaluation" in order to reach the targeted results realistically. In other words, not only the behaviors, characteristics and the like of the leaders, but also the factors such as expectations, characteristics and style of living are taken into account.

After this data which will be obtained from the construction companies in Adana province through the questionnaire forms, data mining studies have been started. Detailed survey studies for the determination of relations have been applied to civil engineers working in construction companies as construction site supervisors and other employees in construction sites. A bidirectional model designed to be tested in this context is shown in **Figure 1**.

When the relevant model is considered, it has been decided that the implementation of a bidirectional questionnaire will be positive and appropriate, as explained earlier. When all theories and methods in the relevant literature are examined, it was found appropriate to use the multiple leadership questionnaire (MLQ) scale developed by Bass and Avolio [8] in this study to determine the types and characteristics of the engineers, as shown in **Figure 2**.

Information was gathered from the site chiefs about the efficiency value to be calculated for employees who will evaluate themselves. These collected data were added to the end of MLQ surveys applied to engineers. The chiefs assessed the productivity level of their employees by choosing between low, medium and high. In the employee questionnaires, employees indicated their productivity information by choosing one of the low, medium and high options according to the management of the site chiefs and leadership understanding. It is ensured that the data mining methods that form the basis of the study in this way are consistent in the data to be applied.

Since MLQ has a bidirectional survey application system, it reveals how leaders perceive the way of management both from the point of view of themselves and from the point of view of employees. Forty five questions in the short measure are asked to the leaders in active mode and to the employees in passive mode. The relevant scale with this feature, that is, the same type of questions are directed to the people on both sides of the subject, it is possible to obtain more healthy results. With this feature, MLQ is a leadership survey questionnaire that has been used in recent years by many researchers in different disciplines. The survey questionnaire used in this study does not include personal information of the persons, only

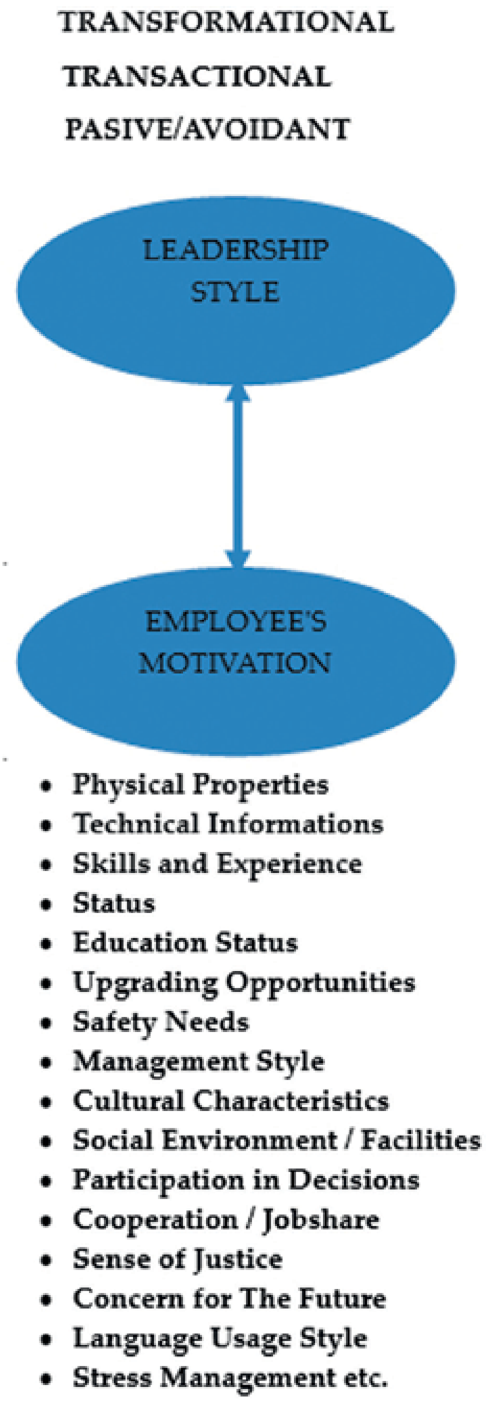


**Figure 1.** Model designed to be tested.

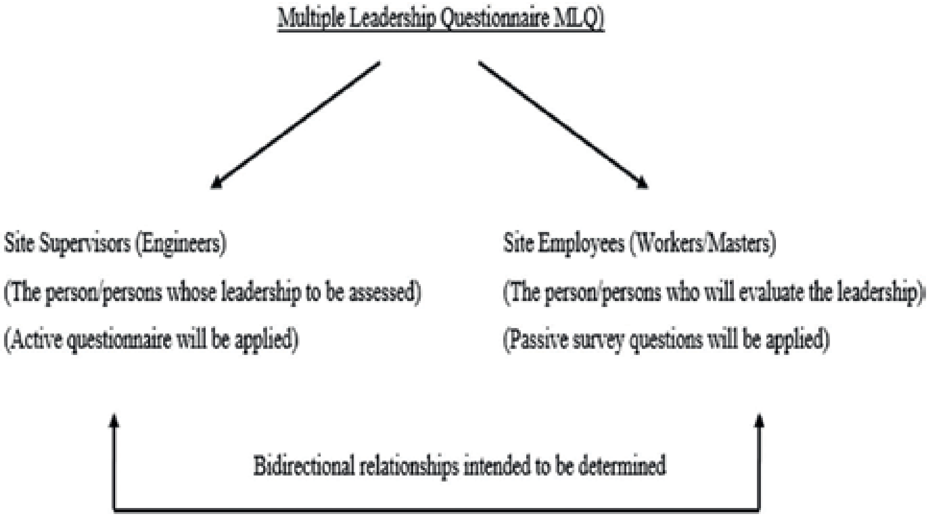

Multiple Leadership Questionnaire MLQ)
Site Supervisors (Engineers)
(The person/persons whose leadership to be assessed)
(Active questionnaire will be applied)
Site Employees (Workers/Masters)
(The person/persons who will evaluate the leadership)
(Passive survey questions will be applied)
Bidirectional relationships intended to be determined

**Figure 2.** Systematization of the related survey study and the goals planned to be achieved.

information like their age range, their gender, how much they are working, and so on were collected. In this study, a scale covering 45 questions, which was revised and abbreviated, was used instead of 72 questions.

According to the main axis of the study, data mining methods are used together with Weka and KEEL programs for classifications. In a sense, this study is supported by a different perspective that is not frequently used in the sector, and the sector has benefited.

Data mining is the process of extracting previously undiscovered information based on a wide variety and quantity of data held in data warehouses and using them to make decisions and action plans. It is the search for relationships and rules that will allow us to make predictions about the future from a large amount of data. Data mining is the semiautomatic discovery of patterns, relations, changes, irregularities, rules and statistically significant structures in data. The computer is responsible for determining the relationships, rules and properties between the data. The goal is to be able to detect previously unrecognized data patterns [9].

It is necessary to address the different types of data for an effective data mining application, to ensure the effectiveness and scalability of the data mining algorithm, to provide usefulness, accuracy and significance of the results, to display the discovered rules in various forms, to process data in different environments and to provide privacy and data security features. Alternatively, data mining is actually regarded as a part of the knowledge discovery process. The stages of the knowledge discovery process are as follows [9]:

1. Data cleaning (remove loud and inconsistent data)
2. Data integration (combining multiple data sources)
3. Data selection (determine the data related to the analysis to be performed)

4. Data transformation (to transform the data to be used by the data mining technique)
5. Data mining (implementing intelligent methods to capture data patterns)
6. Pattern evaluation (to identify interesting patterns representing information obtained according to some measurements)
7. Information presentation (performing the user presentation of the obtained information that has been mined) [10, 11].

The difficulty in obtaining the relevant data from the large data generated by the use of technology in every sector is also valid for leadership. In order to obtain meaningful and useful information from meaningless data heaps, it is planned to use data mining methods in this study. For this reason, the data in the surveys gathered within the scope of the study were primarily preprocessed and then prepared in the relevant file format for operation in data mining programs. Today, both commercial and open source programs have been developed to make data mining studies. There are many algorithms in these programs. By using these algorithms, meaningful information can be extracted from the data available [9]. In this study, KEEL software was used for preprocessing steps and Weka software for classification steps.

Weka [10] is the abbreviation for Waikato Environment for Knowledge Analysis. It is a Java-based data mining and machine learning software developed under the GNU general public license at Waikato University in New Zealand. It includes preprocessing, classification, clustering, association rule mining, feature selection and visualization processes on data sets. Weka works with the Attribute Relationship File Format (.arff) file format. This file format is a specially designed file format that is kept in a text structure. The @relation, @attribute and @data statements are used to specify the file structure. The @relation specifies the general purpose or name of the stack data. While @attribute is used to specify attribute names that correspond to columns in the data set, @data marks the beginning of the raw data set.

KEEL is software written in Java language developed by the University of Granada with the support of the National Science Projects Agency of Spain. KEEL is not rich in terms of classical data mining algorithms such as clustering. Instead, Fuzzy classifiers, artificial intelligence-based classification and rule-based clustering algorithms are included [12]. One of the weakest software in terms of data visualization is the KEEL. Since KEEL software provides highly advanced algorithms in preprocessing parts according to other software, KEEL software was used in the preprocessing phase of the data obtained from the questionnaires in this study; that is, the data preprocessing level, which constitutes the first four steps of the information discovery process, was realized with the help of KEEL software. During this preprocessing phase, normalization is performed, and the data are transformed into the related form.

## 4. Findings and results

In the scope of the study, 102 employee questionnaires and 102 leader questionnaires were applied. The results of the obtained leadership outputs are given in **Table 1**.

| | Transformational leadership | Transactional leadership | Passive avoidant leadership | Total number of surveys |
|---|---|---|---|---|
| According to the employee | 81 | 20 | 1 | 102 |
| According to the leader | 84 | 18 | 0 | 102 |

**Table 1.** Survey results.

The collected surveys in the scope of this study were primarily brought together in an Excel file. The data stored in Excel format are then converted to the .arff file format, which is a file format of Weka, so that one of the data mining program, Weka, can be run. For this, various preprocessing was carried out with the help of KEEL program and transformation of file format was performed.

In the scope of the study, min-max normalization method, which is widely used, is used. In the min-max normalization method, the largest and smallest values in a group are handled. All other data are normalized to these values. The purpose here is to normalize the smallest value to be 0 and the largest value to be 1 and to spread all the other data to this 0–1 range. The prescribed formulation is shown in Eq. (1). According to this formulation, v is the value

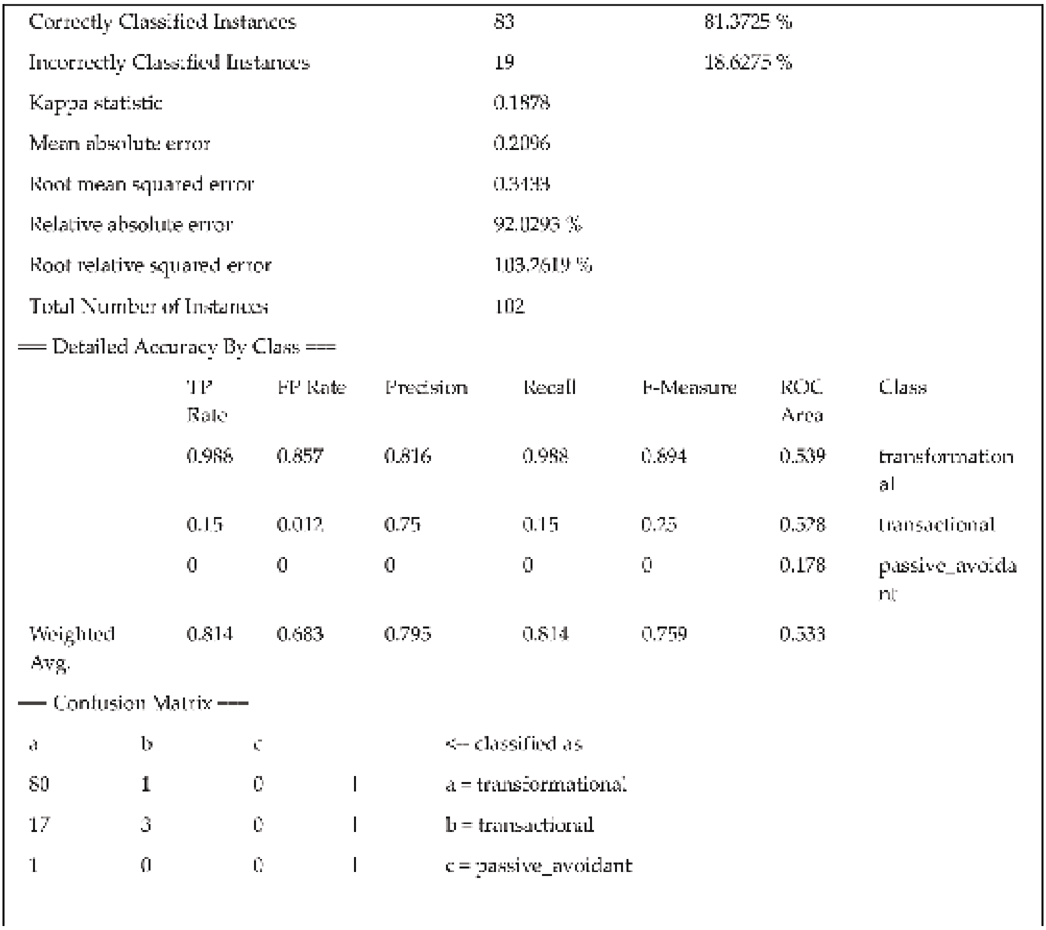
```
@relation bap_calisan_2017
@attribute gorev
{boyaci,demirci,dogalgaz_doseme,formen,isci,kalipci,parke_ustasi,seramikci,sivaci,tesisatci,usta}
@attribute cinsiyet {1,2}
@attribute yas {1,2,3,4,5}
@attribute egitim {1,2,3,4,5,6}
@attribute maas {1,2,3,4}
@attribute tecrube {1,2,3,4}
@attribute benzer_islerdeki_tecrube {1,2,3,4}
@attribute …….
…………
@attribute motivasyon {yuksek,dusuk}
@attribute lideregore_liderliktipi {transformational,transactional,passive-avoidant}
@attribute calisanagore_liderlik_tipi {transformational,transactional,passive_avoidant}
@data
tesisatci,osmaniye,1,2,3,2,3,3,2,3,2,1,3,0,0,1,3,1,2,7,10,1,yuksek,transformational,transformational
sivaci,kahramanmaras,1,3,3,2,4,4,3,3,3,1,3,0,1,0,3,1,2,7,10,2,yuksek,transformational,transformational
sivaci,adana,1,2,3,1,3,4,3,3,1,1,3,0,1,0,3,1,2,7,10,2,dusuk,transformational,transformational1,3,2,4,6,2,1,0,1,0,3,dusuk,D
isci,adana,1,2,3,2,4,4,3,3,3,1,1,0,1,0,3,1,2,7,8,2,dusuk,transformational,passive_avoidant
formen,adana,1,2,3,3,3,3,2,3,2,3,4,1,0,0,3,2,2,6,8,2,yuksek,transformational,transformational
isci,adana,1,3,2,1,4,4,4,3,4,1,3,0,1,0,3,1,2,7,10,2,yuksek,transactional,transactional
kalipci,erzurum,1,3,2,1,4,4,1,3,1,2,1,0,1,0,3,2,3,7,8,1,dusuk,transformational,transactional
parke_ustasi,adana,1,3,2,1,4,4,3,3,2,1,3,0,1,0,3,2,2,6,10,2,dusuk,transactional,transformational
……..
```

**Table 2.** Categorical .arff file.

to be normalized, whereas v′ is the value to be obtained as the result of normalization. In the formulation, new_min is taken as 0 and new_max is taken as 1.

$$v' = \frac{v - min_A}{max_A - min_A}(yeni\_max_A - yeni\_min_A) + yeni\_min_A \tag{1}$$

The format of the .arff file is different for each data mining method. Some of the data mining methods in Weka only work with numerical data, while others work with categorical data. For example, while there is a greater need for numerical data for classification and clustering algorithms, categorical or nominal data are needed for the algorithm of association rules [7]. Numerical data were categorized by the uniform frequency method. In the equal frequency method, the property ranges are divided into N ranges, and an equal number of pieces of data are held in each range. This method is used because it can work with distorted data. **Table 2** shows the summary of the .arff file format prepared in the scope of the study.



```
 Correctly Classified Instances              83              81.3725 %
 Incorrectly Classified Instances            19              18.6275 %
 Kappa statistic                             0.1878
 Mean absolute error                         0.2096
 Root mean squared error                     0.3433
 Relative absolute error                     92.0293 %
 Root relative squared error                 103.2619 %
 Total Number of Instances                   102
=== Detailed Accuracy By Class ===
                TP      FP Rate   Precision    Recall     F-Measure   ROC      Class
                Rate                                                  Area
                0.986   0.857     0.816        0.988      0.894       0.539    transformation
                                                                               al
                0.15    0.012     0.75         0.15       0.25        0.578    transactional
                0       0         0            0          0           0.178    passive_avoida
                                                                               nt
 Weighted       0.814   0.683     0.795        0.814      0.759       0.533
 Avg.
=== Confusion Matrix ===
 a        b        c                    <-- classified as
 80       1        0        |           a = transformational
 17       3        0        |           b = transactional
 1        0        0        |           c = passive_avoidant
```

**Table 3.** The classification results of random forest.

Within the scope of this study, the leadership qualities that employees qualify for their leaders are classified. According to the analysis results obtained, the most successful data mining algorithm in terms of classification has been random forest algorithm with a rate of 81.3725%. This algorithm is a community learning algorithm. This algorithm generates more than one decision tree during the classification process and thus aims to increase the classification value. Individually constructed decision trees come together to form the decision forest. The classification results obtained are given in **Table 3**.

## 5. Conclusion

In this study, some results were obtained by applying "Classification" methods of data mining methods. With the use of classification algorithms in leadership, an analysis of the effect of leaders' leadership styles and behaviors on the motivation of site employees has been made in the construction industry.

According to the MLQ licensed questionnaire score, the leadership style tendencies of construction site supervisors were categorized and evaluated in three different ways as "transformational," "transactional" and "passive/avoidant" within the scope of responses obtained from the survey questionnaires in this study.

The following suggestions can be made to the construction sector, construction companies and construction site managers who are the managers of the construction sites in order to overcome the deficiencies found in the subject in the data presented in this study.

Among the important results of the study, it can be said that sector representatives will benefit from the fact that the motivation of employees can be increased if each employee group has different qualifications and if these leaders are selected and implemented in accordance with these qualities.

At the same time, based on the information obtained as a result of this study, it is considered that important shortcomings can be achieved by sharing the results of the construction sector, which can raise the motivation of the construction site employees with the representatives of the construction sector. In this context, it is thought that it would be beneficial to share the results obtained with the site chiefs and employees in the construction companies with the meetings, seminars and similar sessions.

## Acknowledgements

This study was supported by Adana Science and Technology University Scientific Research Project Commission (Project Number: 17103018 and Project Number: MÜHDBF.BM.2015-2111), Turkey.

## Author details

Abdullah Emre Keleş[1] and Mümine Kaya Keleş[2]*

*Address all correspondence to: mkaya@adanabtu.edu.tr

1 Faculty of Engineering, Department of Civil Engineering, Adana Science and Technology University, Adana, Turkey

2 Faculty of Engineering, Department of Computer Engineering, Adana Science and Technology University, Adana, Turkey

## References

[1] Balaban B. The Effect of the Culture on the Motivation of the Workers in the Turkish Construction Sector [thesis]. İstanbul: İstanbul Technical University; 2006

[2] WEKA. Machine Learning Group at the University of Waikato [Internet]. 1993. Available from: https://www.cs.waikato.ac.nz/ml/index.html [Accessed: 2017-12-05]

[3] KEEL. Knowledge Extraction based on Evolutionary Learning [Internet]. 2005. Available from: http://www.keel.es/ [Accessed: 2017-12-05]

[4] Kaya M, Keleş AE, Laptali Oral E. Construction crew productivity prediction by using data mining methods. Procedia—Social and Behavioral Sciences. 2013;**141**:1249-1253. DOI: 10.1016/j.sbspro.2014.05.215

[5] Andaç MS, Oral EL. Yapım İşlerinde Çalışan Verimliliğinin Yapay Arı Kolonisi Algoritması Kullanılarak Tahmini. In: 3. Proje ve Yapım Yönetimi Kongresi (PYYK 2014); November 06-08, 2014. Antalya. 2014. p. 111

[6] Keleş AE, Kaya M. The Analysis of the Factors Affecting the Productivity in the Wall Construction of the Using Apriori Data Mining Method. In: Proceedings of the XVI. Academic Information Conference (AB2014). Mersin; 05-07 February 2014. p. 831-836

[7] Keleş AE. İnşaat Projelerinde Şantiye Şeflerinin Liderliği ve Çalışan Motivasyonu İlişkisinin Veri Madenciliği ile Belirlenmesi [thesis]. Adana: Çukurova University Natural and Applied Sciences Institute; 2016

[8] Bass BM, Avolio BJ. Mind Garden Tools for Positive Transformation, Multifactor Leadership Questionnaire [Internet]. 1995. Available from: https://www.mindgarden.com/16-multifactor-leadership-questionnaire [Accessed: 2017-12-05]

[9] Dener M, Dörterler M, Orman A. Open Source Data Mining Programs: A Case Study on Weka. In: Proceedings of the XI. Academic Information Conference (AB09). Şanlıurfa; 11-13 February 2009. p. 11-13

[10] Han J, Kamber M. Data Mining: Concept and Techniques. 3rd ed. CA: Academic Press; 2001

[11] Garner SR. Weka: The Waikato environment for knowledge analysis. In: Proceedings of the New Zealand Computer Science Research Students Conference (NZCSRSC). New Zealand; 18-21 April 1995. p. 57-64

[12] Bilgin TT. Data Flow Diagrams Based Data Mining Tools and Software Development Environments. In: Proceedings of the XI. Academic Information Conference (AB09). Şanlıurfa; 11-13 February 2009. p. 807-814

# Mining HCI Data for Theory of Mind Induction

Oksana Arnold and Klaus P. Jantke

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.74400

**Abstract**

Human-computer interaction (HCI) results in enormous amounts of data-bearing potentials for understanding a human user's intentions, goals, and desires. Knowing what users want and need is a key to intelligent system assistance. The theory of mind concept known from studies in animal behavior is adopted and adapted for expressive user modeling. Theories of mind are hypothetical user models representing, to some extent, a human user's thoughts. A theory of mind may even reveal tacit knowledge. In this way, user modeling becomes knowledge discovery going beyond the human's knowledge and covering domain-specific insights. Theories of mind are induced by mining HCI data. Data mining turns out to be inductive modeling. Intelligent assistant systems inductively modeling a human user's intentions, goals, and the like, as well as domain knowledge are, by nature, learning systems. To cope with the risk of getting it wrong, learning systems are equipped with the skill of reflection.

**Keywords:** user modeling, inductive modeling, theory of mind, theory of mind induction, theory induction, inductive learning, identification by enumeration, logical refutation

## 1. Introduction

The present work originates from the authors' earlier work in the field of digital games with emphasis on the impact of game play, in general, and on game-based learning, in particular [1–5]. The original approach has been generalized, and algorithms have been extended toward business applications far beyond the limits of gaming [6]. Essentials have been carried over to the study of scenarios of data analysis, visualization, and exploration [7, 8].

Motivated by questions for the impact of playing digital games, the authors analyzed game play represented as (sets of) sequences of actions. Seen from the application point of view, the task in focus is player modeling or, more generally, user modeling. Seen from the data point of view, it is string mining. Seen from the viewpoint of algorithms deployed, the task is pattern inference.

To achieve a high expressiveness, the authors prefer logical terminology powerful enough to approximately represent human goals, intentions, preferences, desires, fears, and the like. Seen this way, the task is theory induction, and the method is hypotheses refutation [9, 10].

## 2. From software tools to intelligent assistant systems

No doubt, *digitalization* pervades nearly every sphere of life. Humans are facing more and more digital systems at their workplaces, in everyday education, in their spare time, and in health care. With the US Food and Drug Administration's approval of aripiprazole tablets with sensors in November 2017 [11], the digitalization reaches the inside of the human body.

Frequently, the process of digitalization is placing on humans the burden of learning about new digital systems and how to use them appropriately. More digital systems do not necessarily ease the human life. To use them effectively, users need to become acquainted with software tools, have to understand the interfaces, and have to learn how to wield the tools. "A tool is something that does not do anything by itself unless a user is wielding it appropriately. Tools are valuable for numerous simple tasks and in cases in which a human knows precisely how to operate the tool. Those tools have their limitations as soon as dynamics come into play. There are various sources of dynamics, such as a changing world or human users with different wishes, desires, and needs" (see [12], p. xii). As the present authors put it earlier, the digitalization process "bears abundant evidence of *the need for a paradigmatic shift from digital tools to intelligent assistant systems*" (see [7], p. 28).

Thinking about human assistance, the most helpful assistants are those who have own ideas, go their own ways, and—from time to time—surprise us with unexpected outcomes. This does apply to digital assistant systems as well.

Approaches to intelligent system assistance are manifold (e.g., see [13, 14] and the references therein including the authors' contributions [15, 16]).

Digital assistants are programmed to behave differently in different conditions such as varying environmental or infrastructure contexts and varying human users with different prior knowledge, preferences, skills, needs, desires, fears, and the like. To adapt accordingly, *assistant systems need to learn* from the data available. In a sense, a digital assistant system has "to ask itself," so to speak, how to learn what the user needs from sparse information such as mouse clicks or wisps over the screen.

Seen in its right perspective, digital assistant systems are facing problems of learning from incomplete information sometimes called *inductive inference* [17]. Digital assistant systems are necessarily *learning systems*.

The purpose of the system's learning is understanding the context of interaction to adapt to. In this chapter, the authors confine themselves to understanding the human user.

Conventionally, this is called *user modeling* naming a rather comprehensive field of studies and applications (see, e.g., [18–22] or any of the earlier UMAP conference proceedings).

By way of illustration, [23] provides a comprehensive digital game case study of mining large amounts of human-computer interaction data—in fact, data of game playing behavior—for the purpose of classification according to psychologically based personality traits [24].

This exemplifies a particular way of user modeling by means of HCI data mining.

## 3. Perspectives at inductive modeling and data mining

### 3.1. Approaches and opinions

Already for decades, the misconception of *data mining as digging for golden nuggets* is spooking through the topical literature [25, 26]. Some authors believe that data mining means somehow squeezing out insights from the given data and put this opinion in words such as "visualization exploration is the process of extracting insight from data via interaction with visual depictions of that data" (see [27], p. 357).

Instead, *data mining is a creative process of model formation based on incomplete information* (see [7], p. 108). In brevity, *data mining is inductive modeling*.

The details of the inductive modeling process depend substantially on the data, on the goal of modeling, of the algorithmic technologies in use, and on the underlying model concept including the syntax of representing models [28, 29]. For going into the details of the model concept, [30] is of particular interest.

In the thesis [30, 31] is taken as a basis for a systematic framework of data mining design in which the model concept resides in the center. An outer frame, so to speak, consists of the application domain, the methods of model construction, and the methods of model use. Every concrete model depends (a) on the context of modeling, (b) on communities of practice, (c) on the purpose of modeling, and, possibly, (d) on models generated earlier (see [31], p. 37). This covers Fayyad's KDD process [32] and the CRISP data mining model [33].

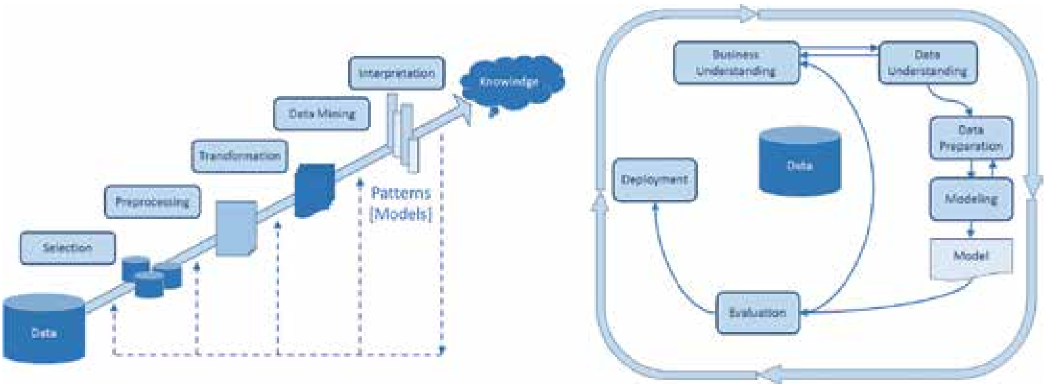


**Figure 1.** Fayyad's KDD process according to [32] vs. the CRISP data mining model as in [33].

With respect to the difficulty of learning from incomplete information, process models of data mining allow for cyclic processing as shown in **Figure 1**. Consequently, data mining has to be seen as a process over time that does not result in an alone model, but in an unforeseeably long *sequence of subsequently generated hypothetical models*. This does perfectly resemble the learning systems perspective of [17].

In the authors' opinion, the thinking about *emerging sequences of hypotheses* is badly underestimated in contemporary data mining investigations. Pondering model concepts is not sufficient. We need to put emphasis on the investigation of suitable *spaces of model hypotheses*.

### 3.2. Theories of mind

Throughout the rest of this chapter, spaces of hypotheses will be *spaces of logical theories*. The concept *theory of mind* is adopted and adapted from behavioral research in animals [34]. There is much evidence that certain animals reflect about intentions and behaviors of other animals [35, 36]. Birds of the species *Aphelocoma californica*—the western scrub jay, esp. the California scrub jay—are food-caching. They do not only cache food but also colorful and shiny objects such as plastic toys. In case such a bird, let us name it A, is caching food or other treasures, and if it is watched by another bird of its species, we name it B, then A returns shortly after to unearth the treasures cached before. The interpretation is, loosely speaking, that the bird A thinks about the possibly malicious thoughts of the bird B. It builds its own theory of mind. More generally speaking, thinking about another one's thoughts means to build a theory of mind.

The authors aim at digital assistant systems able "to understand their human users" by hypothesizing theories of mind. Anthropomorphically speaking, digital assistant systems shall be enabled "to think about their human user's thoughts." The cornerstone has been laid in [1, 2]. Case studies as in [4, 5, 37] demonstrate that this is possible.

For this purpose, user models are seen as theories—just formalizations of theories of mind—such that human user modeling becomes theory induction. The conceptual approach is called *theory of mind modeling and induction* [1, 2] based on human-computer interaction data.

### 3.3. Data mining as theory induction based on HCI data

What the system "knows" about its human user comes from an analysis of interaction data. [3] describes a study based on a commercial digital game. When playing the game, players may learn about pieces of legerdemain. They play successfully when being able to script the necessary steps for doing conjuring tricks. Patterns of game playing behavior reveal the human players' success or failure. Instances of those patterns are shown in recorded game play. It is the system's task to learn patterns from their instances. This approach is generalized toward theory induction.

The concept of a pattern in science dates back to work by Alexander in architecture [38, 39]. Angluin redefines the pattern concept for purposes of formal language investigations and, most importantly, provides algorithmic concepts for learning patterns from instances [40]. Exactly this is what an assistant needs to do when collecting sequences of interaction data.

In game studies such as [3, 5], interaction is abstractly represented by finite sequences over an alphabet A that contains identifiers of all possible activities of all engaged agents such as human players, non-player characters (NPCs), and other computerized components. $A^+$ denotes the set of all those strings, and given any particular game G, $\Pi(G)$ is the subset of all strings that can occur according to the rules and mechanics of G. Angluin's pattern concept describes string properties of a certain type. If instances $\omega_1, \ldots, \omega_n \in \Pi(G)$ occur, the learning task consists in finding a pattern p that holds in all the observed strings. In a sense, p is a theory with $\{\omega_1, \ldots, \omega_n\} \models p$, where $\models$ denotes the logical consequence operator.

The authors generalize the before-mentioned approach toward human-computer interaction in general beyond the limits of game play as undertaken in [6, 7].

The validity of the logical expression $\{\omega_1, \ldots, \omega_n\} \models p$ means *consistency* of the hypothesis p with the set of observations $\Omega_n = \{\omega_1, \ldots, \omega_n\}$ it is built upon. In conditions more general than pattern inference according to [40], the choice of the logic is decisive to consistency. In the conventional case, the question for consistency $\Omega_n \models p$ is recursively decidable but NP-hard. Concerning the background of computability and complexity, the authors rely on [41, 42]. For the moment, let us assume any suitable logic. Details will be discussed as soon as they become interesting.

A closer look at conventional data mining process models as in **Figure 1** reveals that original data appear somehow static. Both models on display show data represented by a drum icon. An emergence over time is beyond the limits of conventional perspectives. In contrast, human-computer interaction data emerge over time [3, 5, 7]. This leads to the learning task of processing sequences $\Omega_1 \subseteq \Omega_2 \subseteq \ldots \subseteq \Omega_n \subseteq \ldots$ of growing finite data sets of observations. When learning patterns according to [39], the learner returns hypotheses $p_1, p_2, \ldots, p_n, \ldots$ such that every hypothesis $p_n$ is consistent with the underlying data set $\Omega_n$.

Consistence is a critical requirement and may be refined by approximations in different ways. In learning theory, it is known that algorithms that are allowed to temporarily return inconsistent hypotheses are of higher effectiveness [17, 43, 44]. The authors refrain from a detailed discussion of these effects, for reasons of space.

Extending the abovementioned approach, one arrives at an understanding of mining HCI data as the induction of theories over emerging sequences $\Omega_1 \subseteq \Omega_2 \subseteq \ldots \subseteq \Omega_n \subseteq \ldots$ of data. The result is a corresponding sequence of logical theories $T_1, T_2, \ldots, T_n, \ldots$ which, if possible, should converge to an ultimate explanation T of the observed human-computer interaction, i.e., $\Omega_n \models T$ for all sets of observations.

By way of illustration, [5] is based on a case study in which the sequence of theories begins with some default $T_0$ and consists of 35 subsequent hypothetical theories. In this sequence, subsequent theories remain unchanged frequently. There are only nine changes of hypotheses. The final theory reasonably explains the overall human user's behavior.

Note that there are varying other approaches to deal with dynamics such as [45, 46]. The authors, however, stick to the logical approach for its declarativity and expressiveness.

Different from other approaches, they dovetail logical reasoning and inductive inference [17, 43]. In this way, logics and recursion theory are underpinning data mining on HCI data.

### 3.4. Formalization and operationalization of theory induction on HCI data

The logical background of the authors' approach includes reasoning about changes in time. This leads directly to temporal logics that are around for already more than half a century [47, 48]. In these good old days, time was tense, but in conditions of digitalization, time became digital as well [49]. It was already known before that this makes a difference [50].

In the simple digital game case study [4, 5], it is sufficient to choose the Hilbert-style logic K (see [49], Section 1.6).

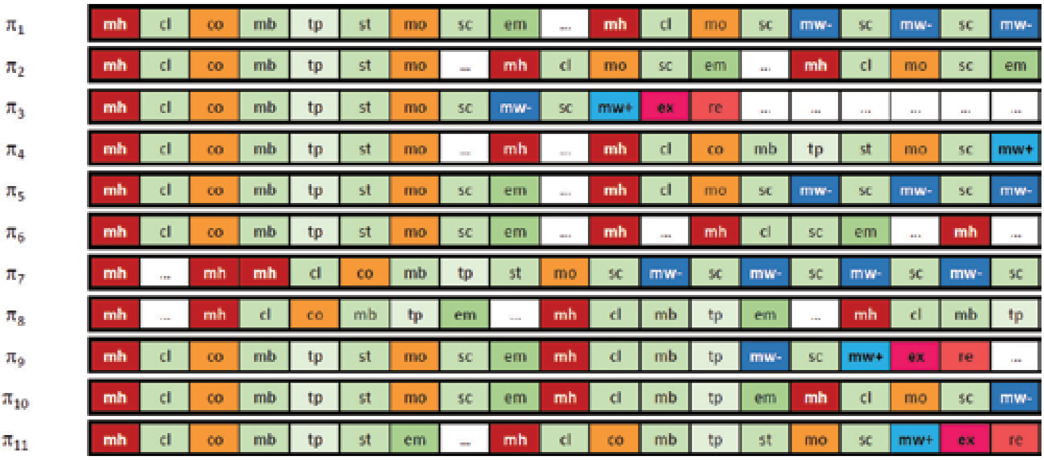
Which logic to choose depends on the particular domain of application. In particular, there is an indispensable need (i) to formalize background knowledge. The logic must allow for the representation of knowledge in such a way that it is easy (ii) to refute hypotheses [9, 10]. Below, we will come back to these two issues. Logics taken into account come from [49–56]. For the generic approach discussed in this section, however, the choice is subordinate.

Speaking about human-computer interaction with the intention of user modeling by theories of mind, the fundamental question is what to take into account. Interaction may be represented on largely varying levels of granularity [57] ranging from keystrokes and wisps over the screen through compound actions to activities on a task level (named quests in the world of digital games, where the approach originated). The authors are engaged in a joint project in which even the exact position of a document on the screen plays a role in mining HCI data and, thus, must be documented in interaction representations [58].

The (finite) set of actions of interest is denoted by A. It is considered an alphabet. As usual in theoretical computer science, A* denotes the set of all finite strings over A including the empty string $\varepsilon$, $A^+ = A^* \setminus \{\varepsilon\}$. If it makes sense, one may restrict $A^+$ to the set $\Pi$ of only those sequences that can occur in practice, $\Pi \subseteq A^+$. In conditions of strictly regulated interaction possibilities, $\Pi$ is a formal language [59].

Every string $\pi \in \Pi$ abstractly represents some process of human-computer interaction such as a game play [3] or a session with a data analysis tool [7]. When trying "to understand the human user," $\pi$ is subject to investigation. Sometimes, there is a finite subset of $\Pi$ available. By way of illustration, see **Figure 2** adopted and adapted from ([3], p. 89, Figure 6.3).

According to the game magazines worldwide, allegedly, the innovation of the commercial game studied in [3] consists in the unprecedented feature of players doing conjuring tricks. To investigate this in more detail, the alphabet A of actions contains player actions such as clicking to a magic head (denoted by mh in the strings on display in **Figure 2**), opening a grimoire, in the game called a magic book (mb), turning pages of the book when searching for an appropriate trick (tp), selecting a trick from the book (st), scripting the steps of the trick in preparation of performance (sc), and presenting the trick by means of a magic wand either successfully (mw) or not (mw-). The digital game system's actions indicated by boxes ranging from yellowish to reddish in **Figure 2** are the presentation of a magic head (mh) indicating that



| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\pi_1$ | mh | cl | co | mb | tp | st | mo | sc | em | ... | mh | cl | mo | sc | mw- | sc | mw- | sc | mw- |
| $\pi_2$ | mh | cl | co | mb | tp | st | mo | ... | mh | cl | mo | sc | em | ... | mh | cl | mo | sc | em |
| $\pi_3$ | mh | cl | co | mb | tp | st | mo | sc | mw- | sc | mw+ | ex | re | ... | ... | ... | ... | ... | ... |
| $\pi_4$ | mh | cl | co | mb | tp | st | mo | ... | mh | ... | mh | cl | co | mb | tp | st | mo | sc | mw+ |
| $\pi_5$ | mh | cl | co | mb | tp | st | mo | sc | em | ... | mh | cl | mo | sc | mw- | sc | mw- | sc | mw- |
| $\pi_6$ | mh | cl | co | mb | tp | st | mo | sc | em | ... | mh | ... | mh | cl | sc | em | ... | mh | ... |
| $\pi_7$ | mh | ... | mh | mh | cl | co | mb | tp | st | mo | sc | mw- | sc | mw- | sc | mw- | sc | mw- | sc |
| $\pi_8$ | mh | ... | mh | cl | co | mb | tp | em | ... | mh | cl | mb | tp | em | ... | mh | cl | mb | tp |
| $\pi_9$ | mh | cl | co | mb | tp | st | mo | sc | em | mh | cl | mb | tp | mw- | sc | mw+ | ex | re | ... |
| $\pi_{10}$ | mh | cl | co | mb | tp | st | mo | sc | em | mh | cl | mb | tp | em | mh | cl | mo | sc | mw- |
| $\pi_{11}$ | mh | cl | co | mb | tp | st | em | ... | mh | cl | co | mb | tp | st | mo | sc | mw+ | ex | re |

**Figure 2.** Excerpts from recorded game play of 11 subjects striving to do a conjuring trick.

there is an opportunity of witchcraft, a comment (co) in response to a human user's click to inform the player what to do, the opening of the magic book (mo) to allow for scripting tricks, and, in case the trick has been scripted correctly and the user has triggered its execution by clicking to the magic wand, a virtual execution (ex) of the trick by means of a cut scene and some response (re) to the player about the success of the performance.

The (cutouts of) strings on display in **Figure 2** have different properties that are indicators of the players' mastery of game play, in general, and of scripting tricks, in particular [3]. For a precise and readable treatment, actions in A are written in brackets such as [mh] and [ex]. […] abbreviates an action not of interest. Using this convention, the cutout of $\pi_1$ is [mh][cl][co][mb][tp][st][mo][sc][em][…][mh][cl][mo][sc][mw-][sc][mw-][sc][mw-]. Readers may easily recognize that the player has a problem. The substring [sc][mw-] indicates a failed effort of scripting and doing a conjuring trick. This will be discussed in some detail.

Suppose that $\preccurlyeq$ denotes the substring relation. $\pi_1 \preccurlyeq \pi_2$ means that there are (possibly empty) strings $\pi'$ and $\pi$ "satisfying $\pi'\pi_1\pi$ "= $\pi_2$. In other words, $\pi_1$ occurs somewhere in $\pi_2$.

By way of illustration, the following two sample formulas $\varphi_2$ = [sc][mw-][sc][mw-] $\preccurlyeq \pi$ and $\varphi_3$ = [sc][mw-][sc][mw-][sc][mw-] $\preccurlyeq \pi$ describe certain string properties. This justifies logical expressions such as $\pi \vDash \varphi_2$ and $\pi \vDash \varphi_3$ meaning that the string $\pi$ satisfies the corresponding property. It is custom to say that $\pi$ is a model of $\varphi_2$ or $\varphi_3$, respectively. The intuitive meaning is quite obvious. When $\varphi_3$ occurs in a string $\pi$ describing human game play, the player appears to stab around in the dark. According to **Figure 2**, it holds $\pi_1 \vDash \varphi_3$, $\pi_5 \vDash \varphi_3$, and $\pi_7 \vDash \varphi_3$. Properties of this type are called *patterns*. Patterns according to Angluin [40] are properties of strings that are decidable. This does obviously apply to both $\varphi_2$ and $\varphi_3$ as well.

Because the information about the other eight strings of game play in **Figure 2** is incomplete, we are not sure whether or not one of the patterns $\varphi_2$ and $\varphi_3$ is satisfied. With respect to the information available, all we know is that we are *not* able to *disprove* one of these patterns.

Needless to state that in computational logics, double negation cannot be removed [60, 61]. In other words, $(\neg\neg p \rightarrow p)$ is no valid axiom of (propositional) computational logics.

Jantke [37, 62] has developed a family of games based on the Post correspondence problem (see [63], Section 2.6, pp. 88ff). Patterns that occur in game play are of higher complexity than those sketched above. Computational learning of these patterns—theory of mind induction—is possible but considerably more involved. As illustrated in [64], a computer program may even learn skills the human player is not aware of. The authors confine themselves to a sketch of the essentials of PCP games within the following four paragraphs.

A Post correspondence system (see [63], p. 89)—in PCP games, this is called a pool—is a finite set of pairs of strings that may be visualized as dominos. Some of these systems have solutions; others have not. Playing a PCP game means to incrementally modify a common pool according to some rules of play. The goal is to make a pool solvable and to prevent others from doing so. Who makes a pool solvable and declares victory accordingly wins the game by showing a solution. If the player's demonstration fails, the game is lost.

Interestingly, the solvability of Post correspondence systems is algorithmically undecidable (for a comprehensive treatment of undecidability, [65] is recommended). As a consequence, a player might be unaware of being able to declare victory and to win the game accordingly. There is the phenomenon of missing a win. This may occur repeatedly.

Using elementary formalizations (see [62, 64] for details), one may write down formulas $\varphi_n$ of first-order predicate calculus saying that a player never misses more than $n$ wins in a game. Whether or not $\psi_n$ holds in recorded game play $\pi$ is effectively undecidable. But the problem is effectively enumerable (some call it semi-decidable).

Therefore, a computer program can watch a human playing PCP games. It can analyze strings describing the human-computer interaction for the occurrence of missing wins. The program's first hypothesis may be $\psi_0$. In case a missing win is detected, the hypothesis is changed to $\psi_1$. If $\psi_n$ is hypothesized, but one more missing win is diagnosed, the hypothesis is changed to $\psi_{n+1}$. The underlying process is identification by enumeration [66].

Let us have a look—quick and dirty—at the principle of identification by enumeration from a logical viewpoint. A space of hypotheses is an effective enumeration $T_0, T_1, T_2, T_3, T_4, T_5, \ldots$ of theories; in the paragraph before, these theories are the singleton sets $\{\psi_n\}$. When sets of observations $\Omega_1 \subseteq \Omega_2 \subseteq \Omega_3 \subseteq \Omega_4 \subseteq \Omega_5 \ldots$ come in subsequently, learning means to search the given enumeration of hypotheses for the first theory that does not contradict the current information. Formally, a learner L getting fed in $\Omega_n$ searches for $k = \mu m\ [\neg(\Omega_n \not\models T_k)]$ and hypothesizes $L(\Omega_n) = T_k$. The symbol $\mu$ represents the minimum operator [41].

As explicated already much earlier [67], the key logical reasoning problem in learning from incomplete information is refutation. This is sound with related philosophical positions [11]. The crux is that $\neg(\Omega_n \not\models T_k)$ is usually undecidable as seen in the PCP game case study. This leads to the authors' original pattern concept. Whereas in [3, 68]—adopted from [40]—the assumption is that the validity of a pattern in a stream of HCI data is decidable, the ultimate approach weakens the requirement (see [37], p. 12): Patterns are logical theories that are co-semi-decidable. In other words, under the assumption of an underlying logic with (i) its consequence operator $\models$, (ii) the operator's implementation $\vdash$, (iii) background knowledge, and (iv) current observations, the implementation $\vdash$ may be used to find out in a uniform way whether any set of observations and any theory are inconsistent. Furthermore, according to scenarios of analyzing human experience of patterns in HCI data [69], patterns should have the property of locality. Informally, once a pattern instance occurred, it does not disappear throughout subsequent interaction. In formal terminology, for any pattern $\varphi$ and for any $\pi_1$, $\pi_2 \in A^*$, the validity of $\pi_1 \models \varphi$ implies the validity of $\pi_1\pi_2 \models \varphi$.

### 3.5. Theory of mind model induction via identification by enumeration

To sum up, theory induction on HCI data is operationalized by construction of theories and sticking to them as long as they are not refuted. The underlying decisive knowledge forms an effectively enumerable space of hypotheses. In formal language learning, the appropriate technical term is called an indexed family of formal languages [70]. For the purpose of theory induction, this concept has been slightly generalized. The authors coined the term of an indexed family of logical formulas [5]. Because logic in general is more expressive than formal languages are, there is a need for requirements that are weaker but still sufficient to allow for inductive learning.

Assume any logic that does not exceed the expressive power of first-order predicate calculus to allow for a completeness theorem [71]. The logic brings with it its well-formed formulas, its consequence operator $\models$ and the operator's implementation $\vdash$ (due to completeness). Practically, refutation completeness is sufficient [67]. By way of illustration, the authors' recent application uses Horn logic and relies on the refutation completeness of Prolog [72].

Given domain-specific background knowledge BK, an indexed family F of logical formulas is defined by the following conditions. $F = \{\varphi_n\}_{n\,=\,0,1,2,\dots}$ such that the sequence of formulas $\varphi_n$ is effectively enumerable. Furthermore, for any two indices m and n with m < n, the formula that occurs later in the enumeration does not imply the earlier one, i.e., $BK \not\models (\varphi_n \rightarrow \varphi_m)$.

Note that the sequence of formulas $\{\psi_n\}_{n\,=\,0,1,2,\dots}$ discussed in the context of PCP games above meets the conditions and, thus, is an example of an indexed family of logical formulas. The corresponding background knowledge comprises the rules of play including Peano arithmetic.

Apparently, the authors' approach is a two-stage process above the granularity of the more conventional processes depicted in **Figure 1**. First, one selects an effectively enumerable space of hypotheses. Second, one performs identification by enumeration as the key learning methodology. Other conventional steps such as data selection, data preparation, and data preprocessing occur as well [8]. However, the latter are not in focus of this chapter.

As the choice of spaces of hypothetic models—an issue ignored in conventional approaches—is decisive, it is worth to take updates and revisions into account. The authors introduced a generalization for which they coined the term dynamic identification by enumeration [73].

## 4. An inductive inference perspective at mining HCI data

In contrast to earlier approaches that are widespread (see **Figure 1**, where in the CRISP-like model on the right, the "model" node is hatched, as it is missing in the original figure [33]), the authors stress the aspects illustrated by the (four groups of) darker boxes in **Figure 3**. First of all, data are not seen as a monolithic object within the process concept but as an emerging sequence. Second, whereas in the Fayyad process (see Figure [1] and the source [32]), the pattern concept appears from nowhere, the terminology of forming hypotheses is seen a central issue—the selection of a logic and the design of suitable spaces of hypotheses, both potentially subject to revision over time. Third, the inductive modeling procedure discussed in some more detail throughout this chapter is identification by enumeration.

Involved logical reasoning may easily become confusing—not so much to a computer or to a logic program [4], but to a human being. Within the digital game case study [4, 5], the generation of a single indexed family of logical formulas has been sufficient. Identification by enumeration works well for identifying even a bit perfidious human player intentions. Business applications as in [6] are more complex and may require unforeseeable revisions of the terminology in use, i.e., the dynamic generation of spaces of hypotheses on demand [73].

The present section is aimed at a clarification of the core ideas and technicalities. For this purpose, the approach is stripped to the essentials. Recursion-theoretic inductive inference as in [17, 43, 44] is the most lucid area in which problems of inductive learning can be explicated

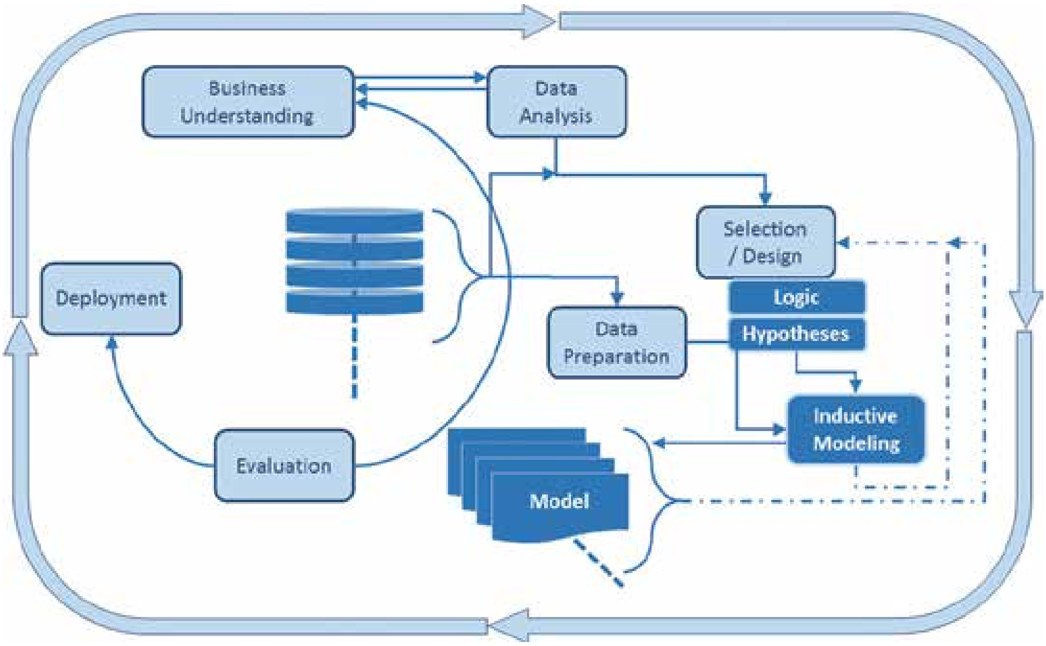


**Figure 3.** HCI data mining approach with emphasis on aspects of inductive modeling.

without any need for dealing with syntactic sugar. The underlying apparatus of mathematics can be found in textbooks such as [41, 63].

In **Figure 4**, the darker boxes with white inscriptions denote conventional concepts of recursion-theoretic inductive inference [44]. The other boxes reflect formalizations of this chapter's core approaches to HCI data mining by means of identification by enumeration. The concepts derived from the present chapter's practical investigations form a previously unknown infinite hierarchy between the previously known concepts NUM and TOTAL.

Throughout the remaining part of this section, the authors confine themselves to only elementary concepts.

Learning logical theories is very much like learning recursive functions. Both have finite descriptions but determine a usually infinite amount of facts—the theorems of a theory and the values of a function, respectively. In both cases, the sets of facts are recursively enumerable but usually undecidable. The deep interplay of logic and recursion theory is well understood for almost a century and provides a firm basis of seminal results [74]. Inductively learning a recursive function means, in some sense, mining the function's graph which is presented in growing chunks over time, a process very similar to mining HCI data.

A few notions and notations are inevitable. IN is the set of natural numbers. $P^n$ denotes the class of n-ary partial recursive functions mapping from $IN^n$ into IN. $R^n \subset P^n$ is the subclass of all total recursive functions. Assume any ordering of IN written in the form $X = \{x_0, x_1, x_2, \ldots\}$. For any function $f \in R^1$, the sequence of observations $(x_0, f(x_0))$, $(x_1, f(x_1))$, $(x_2, f(x_2))$, $(x_3, f(x_3))$, … provides growing but incomplete information about f. With respect to the ordering X, the amount of information up to the timepoint n is encoded in $f_X[n] = ((x_0, f(x_0)), \ldots, (x_n, f(x_n)))$. If $X_0$

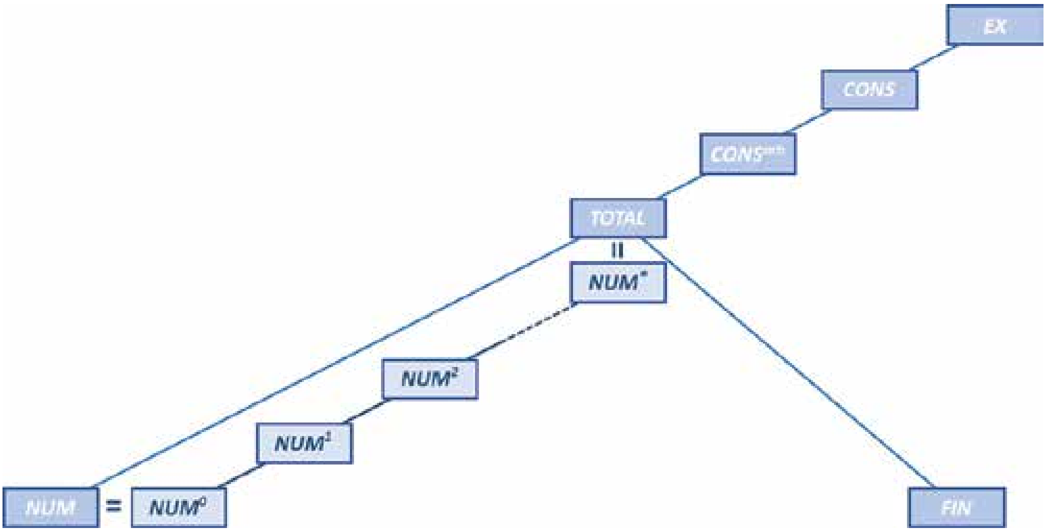


**Figure 4.** Abstractions of fundamental inductive learning concepts compared and related; ascending lines mean the proper set inclusion of the lower learning concept in the upper one.

is the standard ordering 0,1,2,3,4, …, the index is dropped such that the notation is f[n]. Throughout any learning process, hypotheses are natural numbers interpreted as programs according to some Gödel numbering $\varphi$. Because any two Gödel numberings are recursively isomorphic, the choice of a numbering does not matter. Learnability is transcendental.

Assume any class $C \subset R^1$ of total recursive functions. The functions of C are uniformly learnable by an effectively computable learner $L \in P^1$ on the ordering of information $X_0$, if and only if the following conditions are satisfied. For all $f \in C$ and for all $n \in IN$, the learner computes some hypothesis L(f[n]). For every $f \in C$, the sequence of hypotheses converges to some c that correctly describes f, i.e., $\varphi_c = f$. EX denotes the family of all function classes learnable as described. $EX(L) \in EX$ is the class of all functions learnable by L. In the case arbitrary arrangements of information X are taken into account, the definition is changed by substituting $f_X[n]$ for f[n]. The class of functions learnable by L is named $EX^{arb}(L)$, and the family of all function classes learnable on arbitrary X is $EX^{arb}$. The term EX is intended to resemble *explanatory learning*; this is exactly what theory of mind induction is aiming at.

The equality of EX and $EX^{arb}$ is folklore in inductive inference. Therefore, arbitrary orderings are ignored whenever possible without loss of generality.

Intuitively, it seems desirable that a hypothesis reflects the information it is built upon. Formally, $\forall m \leq n\ (\varphi_h(x_m) = f(x_m))$ where h abbreviates $L(f_X[n])$. In the simpler case of $X_0$, every $x_m$ equals m. The property is named *consistency*. The families of function classes uniformly learnable consistently are CONS and $CONS^{arb}$, resp., and $CONS^{arb} \subset CONS \subset EX$ is folklore as well. Apparently, the message is that consistency is a nontrivial property.

Consistency may be easily guaranteed, (T) if all hypotheses are in $R^1$ or (F) if it is decidable whether or not a hypothesis is finally correct. Adding (T) or (F) to the definition of EX and $EX^{arb}$, one gets learning types denoted by TOTAL, $TOTAL^{arb}$, FIN, and $FIN^{arb}$, respectively. In inductive inference, $FIN = FIN^{arb} \subset TOTAL = TOTAL^{arb} \subset CONS^{arb}$ is folklore as well [44].

Under the prior knowledge of $FIN = FIN^{arb}$, $TOTAL = TOTAL^{arb}$, and $EX = EX^{arb}$ (see [44]), all the abovementioned inclusions are on display in **Figure 4**.

NUM is the learning type defined by means of identification by enumeration as discussed in the previous section. A class $C \subset R^1$ belongs to NUM, if and only if there exists a general recursive enumeration h with $C \subseteq \{\varphi_{h(n)}\}_{n \in IN} \subset R^1$. A partial recursive learning device $L \in P^1$ learns via identification by enumeration on h, if and only if $L(f[n]) = h(\mu m[\varphi_{h(m)}[n] = f[n])$. Interestingly, this extremely simple concept reflects exactly the application in [4, 5].

The potential of generalizing the learning principle of identification by enumeration is practically demonstrated in [6]. Accordingly, [73] introduces the novel concept of dynamic identification by enumeration. In terms of recursion theory, this looks as follows.

For simplicity, the authors confine themselves to $X_0$. Note that we adopt a few more notations. If h is an enumeration or, alternatively, if n is an index of the enumeration h, $C_h$ and $C_n$ denote the class of all functions enumerated by h. From Grieser [75], we adopt the notation [C] to denote all initial segments of functions in C, i.e., $[C] = \{f[n] \mid f \in C \wedge n \in IN\}$.

A class of functions $C \subseteq R^1$ belongs to NUM*, if and only if there exists a computable generator function $\gamma \in P^1$ such that for all $f \in C$, it holds (I) for all $n \in IN$ that $\gamma(f[n])$ is defined, $\varphi_{\gamma(f[n])} \in R^1$, $C_{\gamma(f[n])} \subseteq R^1$, and $f[n] \in [C_{\gamma(f[n])}]$ and (II) there is a critical point $m \in IN$ such that for all $n \in IN$ larger than m, it holds $\gamma(f[m]) = \gamma(f[n])$ and (III) $f \in C_{\gamma(f[m])}$.

The criteria (I), (II), and (III) are practically motivated [6]. They are called *operational appropriateness*, *conversational appropriateness*, and *semantic appropriateness*, respectively. Usually, the change of $\gamma(f[n])$ to another $\gamma(f[n + 1])$ means an extension of terminology [6, 73]. The condition (II) of conversational appropriates prevents us from a Babylonian confusion.

According to [73], it holds NUM* = TOTAL. This proves the enormous gain of learning power by means of *dynamic* identification by enumeration. Whereas NUM is incomparable to FIN, NUM* is lying far above FIN; [44] provides much more information about the space between FIN and TOTAL.

In this chapter, the authors are going much further by introducing a family $\{NUM^k\}_{k \in IN}$ of infinitely many refinements of NUM*. A class C in NUM* belongs to $NUM^0$, if and only if there exists some generator function $\gamma$ that is constant and identical for all functions f of C. For a positive number k, a class C in NUM* belongs to $NUM^k$, if and only if there exists a $\gamma$ that, for every function f of C, does generate at most k different spaces of hypotheses $\gamma(f[n])$. Intuitively, $\gamma$ suggest at most k times an extension of terminology for the purpose of more appropriately expressing hypotheses throughout the process of data analysis and learning.

Jantke [76] provides a detailed discussion of benchmarks to prove that $\{NUM^k\}_{k \in IN}$ forms an infinite hierarchy as on display in **Figure 4**. For brevity, just two benchmarks are presented. $C^1{}_{q\text{-like}} = \{f \mid f \in R^1 \wedge \forall x \in IN\ (x > 0 \rightarrow f(x) > 0) \wedge \varphi_{f(0)} = f\}$. Apparently, $C^1{}_{q\text{-like}} \in NUM^1 \setminus NUM^0$. $C^{k+1}{}_{q\text{-like}} = \{f \mid f \in R^1 \wedge \exists g \in C^k{}_{q\text{-like}} \wedge \exists n \in IN\ (f(n) = 0 \wedge \forall x \in IN\ (x > n \rightarrow f(x) > 0) \wedge \forall x \in IN\ (x < n \rightarrow f(x) = g(x)) \wedge \forall x \in IN\ (x > n \rightarrow f(x) = \varphi_{f(n+1)}(x\text{-}n\text{-}1))\}$. This allows to separate $NUM^{k+1}$ from $NUM^k$.

## 5. Process models and heuristics for mining HCI data

By the end of Section 3, the authors have summarized their HCI data mining approach and visualized essentials of inductive modeling in **Figure 3**. We take up the thread once again. The selection or the design of a terminology is essential. The terminology determines the space of hypothetical models that may be found. Throughout the process of data mining, model spaces may be subject to revision repeatedly (see preceding Section 4).

The world of models is overwhelmingly rich. Models may be characterized by properties, by purpose, by function, by model viability, or by model fitness [30]. As Thalheim puts it, "models are developed within a theory" ([30], p. 117).

Every concrete application domain provides such an underlying theory. It is a necessary precondition to data mining to specify all the aspects of the underlying theory that should be taken into account (see [30], p. 115, for mapping, truncation, distortion, and the like). Revisions may turn out to be necessary, when inductive modeling, i.e., learning proceeds. Therefore, the word "data understanding" in the CRISP model (see **Figure 1**) is considered inappropriate and, hence, substituted by "data analysis" in the approach shown in **Figure 3**. This figure is intended to visualize both the dynamics of the data and of the model spaces. *Hypothetical* data understanding is seen as the *preliminary* result of data mining.

When speaking about logics and its algorithmic use, it is strictly advisable to stay within the limits of first-order predicate calculus [71]. The selection or the design of a logic means to decide about the signature of the language and about axiom sets of background knowledge.

Under the assumption of a given logic, business understanding and data analysis underpin an impression of what the current analysis process is about. To say it more practically, what might be typical statements arrived at by the end of the data mining process? In the authors' digital game case study, by way of illustration, typical statements explain a human player's action under conditions of a play state [4, 5]. In their business intelligence application [6–8], formulas relate business data and temporal information of largely varying granularity. As soon as the type of expected formulas becomes clear, the next design task is to specify an indexed family of logical formulas. This forms the first space of hypothetical models.

Within the authors' framework, a crucial step is the modification of a space of hypotheses. There are heuristics discussed in [8] that shall be briefly surveyed. An automation may require, to some extent, natural language processing.

The human user's activities are syntactically analyzed. In case there occur terms that have no corresponding sort, constant, function, or predicate names in the formulas of the current space of hypotheses, a limitation of the terminology is detected. The system is "unable to speak about what the user is doing." A case discussed in [8], p. 234, is "retracement of business volume." Retracement is interpreted as inequality with a (large) factor in it, and some sequence of such formulas of properly increasing strength is automatically generated.

Methodologies, guidelines, and process models aiming at (logical) model space construction are worth much more future research work and practical exploration.

## 6. Summary, conclusion, and outlook

The authors' present approach to mining human-computer interaction data works well in applications that provide larger amounts of data [3–6]. The novel dynamic approach to the generation of model spaces exceeds the power of preceding approaches significantly [6, 73].

However successful in the cited prototypical applications, the approach may fail under conditions of small amounts of data. Consequently, it seems inappropriate to applications such as recommender systems. Perhaps, the authors' approach would work when applied to accumulated data of larger numbers of users. If so, the particular outcome would be something like a theory of mind of a user stereotype. Related questions are still open.

Another question derives from the authors' generalization of identification by enumeration. The authors are convinced that it is possible to generalize their recent approach to dynamic identification by enumeration even further. This requires a careful easing of one or more of the requirements named operational appropriateness, conversational appropriateness, and semantic appropriateness. The related open questions need some more research effort.

Finally, the authors want to attract the audience's attention to a larger and rather involved field of research problems beyond the limits of this chapter: *reflective artificial intelligence*.

There are rarely any bug-free software systems. In the future, there will be rarely any bug-free assistant systems. However, even if a future assistant system were to be totally free of bugs, it would hardly be able to solve every imaginable problem. Digital assistant systems may fail. In response to this severe problem, it is necessary to work toward digital systems able to ponder their own abilities and limitations. Systems that do so are called reflective.

Limitations of learning systems are unavoidable [17]. In response, approaches to reflective inductive learning have been developed and investigated in much detail [75]. The results demonstrate the possibility to design and implement reflective artificial intelligence.

The authors' step from the conventional approach to dynamic identification by enumeration reveals a feature of reflection. A learning digital assistant system that gives up a certain space of hypotheses—in formal terms, $\gamma(f[n]) \neq \gamma(f[n + 1])$ resp. $\gamma(f_X[n]) \neq \gamma(f_X[n + 1])$—with the intention to change or to extend the terminology in use is, in a certain sense, reflective. It "worries" about the limits of its current expressive power and aims at fixing the problem. Vice versa, a system able to change spaces of hypotheses, but not doing so (formally, it holds $\gamma(f[n]) = \gamma(f[n + 1])$ or $\gamma(f_X[n]) = \gamma(f_X[n + 1])$, resp.), shows a certain confidence in its abilities to solve the current problem.

This leads immediately to a variety of possibilities to implement reflective system behavior. First, a system changing its space of hypotheses may inform the human user about its recent doubts as to the limitations of terminology. Second, a bit further, it may inform the human user about details of the new terminology. Third, such a system may also report confidence.

As a side effect, so to speak, the authors' work leads to concepts and algorithmic approaches to reflective AI. This bears strong evidence of the need for further in-depth investigations.

## Acknowledgements

After the second author—inspired by some fascinating results in behavioral sciences—has introduced the concept and coined the term of *theory of mind modeling and induction* in 2012, the two authors' student Bernd Schmidt has undertaken the endeavor to provide the first theory of mind modeling and induction application. The authors are grateful to him for his engaged and excellent work and for his continuous willingness to meet whatsoever requirements.

Working on an internship, Rosalie Schnappauf, then a student of the University of Rostock, took part in a series of experiments demonstrating that Bernd Schmidt's implementation of identification by enumeration does really work and allows for the fully computerized induction of a human game player's goals and intentions—a very first case of, so to speak, *mining HCI data for theory of mind induction*.

Rosalie's and Bernd's success encouraged the authors to attack harder application problems and to develop the generalized approach to *dynamic identification by enumeration*.

Part of the work reported in this chapter has been supported by the German Federal Ministry for Education and Research (BMBF) within the joint research project ODIN aiming at Open Data INovation. The authors' subprojects KVASIR (Erfurt University of Applied Sciences) and BALDUR (ADICOM Software) are currently administrated by the agency Projektträger Jülich (PtJ, see www.ptj.de) under the contracts 03PSWKPA and 03PSWKPB, respectively.

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

WK · POTENZIAL

UNTERNEHMEN REGION

Die BMBF-Innovationsinitiative
Neue Länder

## Author details

Oksana Arnold[1] and Klaus P. Jantke[2]*

*Address all correspondence to: klaus.p.jantke@adicom-group.de

1 Erfurt University of Applied Sciences, Erfurt, Germany

2 ADICOM Software, Weimar, Germany

## References

[1] Jantke KP. User Modeling with Theories of Mind: An Introductory Game Case Study. Report KiMeRe-2012-05, Fraunhofer IDMT, Children's Media Dept., Erfurt, Germany, November 2012

[2] Jantke KP. Theory of Mind Induction in User Modeling: An Introductory Game Case Study. Report KiMeRe-2012-06, Fraunhofer IDMT, Children's Media Dept., Erfurt, Germany, December 2012

[3] Jantke KP. Patterns of game playing behavior as indicators of mastery. In: Ifenthaler D, Eseryel D, Ge X, editors. Assessment in Game-Based Learning: Foundations, Innovations, and Perspectives. New York, Heidelberg, Dordrecht, London: Springer; 2012. pp. 85-103

[4] Schmidt B. Theory of Mind Player Modeling [Bachelor Thesis]. Erfurt, Germany: University of Applied Sciences; 2014

[5] Jantke KP, Schmidt B, Schnappauf R. Next generation learner modeling by theory of mind model induction. In: Proceedings of the 8th International Conference on Computer Supported Education (CSEDU 2016), 21–23 April 2016, Rome, Italy. Sétubal: SCITEPRESS, p. 499-506

[6] Arnold O, Drefahl S, Fujima J, Jantke KP. Co-operative knowledge discovery based on meme media, natural language processing and theory of mind modeling and induction. In: Proceedings of the International Conference on e-Society; 20–22 April 2017; Budapest, Hungary. Sétubal: IADIS Press; 2017. pp. 27-38

[7] Jantke KP, Fujima J. Analysis, visualization and exploration scenarios: Formal methods for systematic meta studies of big data applications. In: Grand E, Kotzinos D, Laurent D, Spyratos N, Tanaka Y, editors. Information Search, Integration, and Personalization, 10th International Workshop (ISIP 2016), 10–12 October 2015, Grand Forks, ND, USA, Revised Selected Papers. Heidelberg, Dordrecht, London, New York: Springer; 2016. pp. 107-127

[8] Fujima J, Arnold O, Jantke KP, Schmidt B. Interaction semantics vs. interaction syntax in data visualization and exploration. Design, implementation and utilization of meme media. In: Proceedings of the 30th International Florida Artificial Intelligence Research Society Conference (FLAIRS-30), 22–24 May. Vol. 2017. Marco Island, FL, USA. Palo Alto, CA: AAAI; 2017. pp. 231-234

[9] Popper KR. Logik der Forschung. Tübingen; 1934

[10] Popper KR. Conjectures and Refutations. London: Routledge & Kegan Paul; 1963

[11] FDA. FDA approves pill with sensor that digitally tracks if patients have ingested their medication [Internet]. 2017. Available from: https://www.fda.gov/NewsEvents/Newsroom/ PressAnnouncements/ucm584933.htm [Accessed: November 22, 2017]

[12] Jantke KP. Foreword. In Kreuzberger G, Lunzer A, Kaschek RH, editors. Interdisciplinary Advances in Adaptive and Intelligent Assistant Systems: Concepts, Techniques, Applications, and Use. Hershey, PA, USA: Idea Group Inc.; 2011. pp. vii-viii

[13] Kaschek RH, editor. Intelligent Assistant Systems: Concepts, Techniques and Technologies. Hershey, PA, USA: Idea Group Inc.; 2007

[14] Kreuzberger G, Lunzer A, Kaschek RH, editors. Interdisciplinary Advances in Adaptive and Intelligent Assistant Systems: Concepts, Techniques, Applications, and Use. Hershey, London, Melbourne, Singapore: Idea Group Inc.; 2011

[15] Jantke KP, Igel C, Sturm R. From e-learning tools to assistants by learner modeling and adaptive behavior. In: Kaschek RH, editor. Intelligent Assistant Systems: Concepts, Techniques and Technologies. Hershey, PA, USA: Idea Group Inc.; 2007. pp. 212-231

[16] Jantke KP, Müller C. Wrapper induction programs as information extraction assistants. In: Kaschek RH, editor. Intelligent Assistant Systems: Concepts, Techniques and Technologies. Hershey, PA, USA: Idea Group Inc.; 2007. pp. 35-63

[17] Jain S, Osherson DN, Royer JS, Sharma A. Systems That Learn: An Introduction to Learning Theory. Cambridge, MA, USA: MIT Press; 1999

[18] Carberry S, Weibelzahl S, Micarelli A, Semeraro G. User Modeling Adaptation, and Personalization, Proceedings of the 21st International Conference (UMAP 2013), 10–14 June 2016, Rome, Italy, Ser. LNCS, Vol. 7899. Heidelberg, Dordrecht, London, New York: Springer; 2013

[19] Dimitrova V, Kuflik T, Chin D, Ricci F, Dolog P, Houben GJ. User Modeling Adaptation, and Personalization, Proceedings of the 22nd International Conference (UMAP 2014), 7–11 July 2016, Aalborg, Denmark, Ser. LNCS, Vol. 8538. Heidelberg, Dordrecht, London, New York: Springer; 2014

[20] Ricci F, Bontcheva K, Conlan O, Lawless S. Editors. User Modeling Adaptation, and Personalization, Proceedings of the 23rd International Conference (UMAP 2015), 29 June - 3 July 2016, Dublin, Ireland, Ser. LNCS, Vol. 9146. Heidelberg, Dordrecht, London, New York: Springer; 2015

[21] Vassileva J, Blustein J, Aroyo L, D'Mello S, editors. User Modeling Adaptation, and Personalization, Proceedings of the 24th International Conference (UMAP 2016), 13–17 July 2016, Halifax, NS, Canada. New York, NY, USA: ACM; 2016

[22] Bielikova M, editor. User Modeling Adaptation, and Personalization, Proceedings of the 25th International Conference (UMAP 2017), 9–12 July 2017, Bratislava, Slovakia. New York, NY, USA: ACM; 2017

[23] Lengyel D. A Bottom-Up Approach to Reveal a Mapping between Personality Traits and Gameplaying Traces [Diploma Thesis]. Ilmenau, Germany: Ilmenau Univ. of Technology; 2012

[24] Leary TF. Interpersonal Diagnosis of Personality: A Functional Theory and Methodology for Personality Evaluation. New York, NY, USA: Ronald Press; 1957

[25] Veluswamy R. Clinical quality data mining in acute care. The Physician Executive. 2008: 48-53

[26] Zhang D, Zhou L. Discovering golden nuggets: Data mining in financial application. IEEE Transactions on Systems, Man, and Cybernetics Part C: Applications and Reviews. 2004; **34**:513-522

[27] Jankun-Kelly T, Ma KL, Gertz M. A model and framework for visualization exploration. IEEE Transactions on Visualization and Computer Graphics. 2007;**13**:357-369

[28] Bridewell W, Langley P, Todorovski L, Džeroski S. Inductive process modeling. Machine Learning. 2008;**71**:1-32

[29] Overmars, KP, de Groot, WT, Huigen, MGA. Comparing inductive and deductive modeling of land use decisions: Principles, a model and an illustration from the Philippines. Human Ecology 2007;**32**:439-452

[30] Thalheim B. The conception of the model. In: Abramowicz W, editor. International Conference on Business Information Systems (BIS 2013). Ser. LNBIP. Vol. 157. Berlin, Heidelberg: Springer; 2013. pp. 113-124

[31] Jannaschk K. Infrastruktur für ein Data Mining Design Framework [Ph.D. thesis]. Kiel, Germany: Christian-Albrechts-Univ.; 2017

[32] Fayyad U, Piatetsky-Shapiro G, Smyth P. The KDD process for extracting useful knowledge from volumes of data. Communication ACM. 1996;**39**:27-34

[33] Wirth R, Hipp J. CRISP-DM. Towards a standard process model for data mining. In: Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining. 2000. pp. 29-39

[34] Carruthers P, Smith PK. Theories of Theories of Mind. Cambridge, UK: Cambridge University Press; 1996

[35] Emery NJ, Clayton NS. Comparative social cognition. Annual Review of Psychology. 2009;**60**:87-113

[36] Emery NJ, Dally JM, Clayton NS. Western scrub-jays (Aphelocoma californica) use cognitive strategies to protect their caches from thieving conspecifics. Animal Cognition. 2007;**7**: 37-43

[37] Jantke, KP. PCP-Spiele. Report KiMeRe-2012-04, Version 2.0, Fraunhofer IDMT, Children's Media Dept., Erfurt, Germany, February 2016

[38] Alexander C. A Pattern Language. New York, NY, USA: Oxford University Press; 1977

[39] Alexander C. The Timeless Way of Building. New York, NY, USA: Oxford University Press; 1979

[40] Angluin D. Finding patterns common to a set of strings. J. Computer and System Sciences. 1980;**21**:46-62

[41] Rogers H. Theory of Recursive Functions and Effective Computability. New York, NY, USA: McGraw-Hill; 1967

[42] Garey MR, Johnson DS. Computers and Intractability: A Guide to the Theory of NP-Completeness. San Francisco, CA, USA: W. H. Freeman & Co.; 1979

[43] Angluin D, Smith CH. Inductive inference: Theory and methods. ACM Computing Surveys. 1983;**15**:237-269

[44] Jantke KP, Beick HR. Combining postulates of naturalness in inductive inference. EIK. 1981;**17**:465-484

[45] Yang Q, Wu X. 10 challenging problems in data mining research. International Journal of Information Technology & Decision Making. 2006;**5**:597-604. DOI: 10.1142/S0219622006002258

[46] Roddick JF, Spiliopoulou M. A bibliography of temporal, spatial and spatio-temporal data mining research. ACM SIGKDD Explorations. 1999;**1**:34-38

[47] Prior AN. Time and Modality. Oxford, UK: Clarendon Press; 1957

[48] Prior AN. Past, Present and Future. Oxford, UK: Clarendon Press; 1967

[49] Blackburne, P; De Rijke, M, Venema, Y. Modal Logic. Ser. Cambridge Texts in Theoretical Computer Science, Vol. 53. Cambridge, UK: Cambridge University Press, 2011

[50] van Benthem, JFAK. Logical Dynamics of Information and Interaction. Cambridge, MA, USA: Cambridge University Press, 2010

[51] Gabbay DM. Hodkinson, I, Reynolds, M. Temporal Logic: Mathematical Foundations and Computational Aspects. Ser. Oxford Logic Guides, Vol. 1. Oxford, UK: Clarendon Press; 1994

[52] van Benthem, JFAK, Smets, S. Dynamic logics of belief change. In: van Ditmarsch, H, Halpern, JY, van der Hoek, W, Looi, B, editors. Handbook of Logics for Knowledge and Belief. London, UK: College Publications, 2015. p. 299-368

[53] Fisher M. Temporal representation and reasoning. In: van Harmelen F, Lifschitz V, Porter B, editors. Handbook of Knowledge Representation. Amsterdam, Oxford: Elsevier; 2008. pp. 513-550

[54] Mueller ET. Event calculus. In: van Harmelen F, Lifschitz V, Porter B, editors. Handbook of Knowledge Representation. Amsterdam, Oxford: Elsevier; 2008. pp. 671-708

[55] Doherty, P, Kvarnström, J. Temporal action logics. In: van Harmelen, F, Lifschitz, V, Porter, B, editors. Handbook of Knowledge Representation. Amsterdam, Oxford: Elsevier, 2008. p. 709-757

[56] Peppas P. Belief revision. In: van Harmelen F, Lifschitz V, Porter B, editors. Handbook of Knowledge Representation. Amsterdam, Oxford: Elsevier; 2008. pp. 317-359

[57] Lenerz C. Layered languages of Ludology – Eine Fallstudie. In: Beyer A, Kreuzberger G, editors. Digitale Spiele – Herausforderung Und Chance. Ser. Game Studies. Boizenburg, Germany: Whv; 2009. pp. 39-52

[58] Schedel T, Atzenbeck C. Spatio-temporal parsing in spatial hypermedia. In: Proceedings of the 27th ACM Conference on Hypertext and Social Media (HT 16), 10–13 July 2016, Halifax, Canada. New York, NY, USA: ACM Press. pp. 149-157

[59] Hopcroft JD, Ullman JE. Introduction to Automata Theory, Languages, and Computation. Menlo Park, London, Amsterdam, Ontario: Reading; 1979

[60] Heyting A. Die formalen Regeln der intuitionistischen Logik. Sitzungsberichte der Preussischen Akademie der Wissenschaften, Physikalisch-mathematische Klasse. 1930: 158-169

[61] Kolmogorov AN. Zur Deutung der intuitionistischen Logik. Mathematische Zeitschrift. 1932;**35**:58-65

[62] Jantke KP. PCP-Spiele. Technical Report KiMeRe-2012-04. Erfurt, Germany, Fraunhofer IDMT, Children's Media Dept.; 2012

[63] Machtey M, Young P. An Introduction to the General Theory of Algorithms. Elsevier North-Holland: New York, NY; 1978

[64] Jantke KP. Theory of mind modeling and induction: Ausdrucksfähigkeit und Reichweite. ADISY Technical Report 03/2016. Weimar, Germany: ADISY Consulting GmbH; 2016

[65] Davis M, editor. The Undecidable: Basic Papers on Undecidable Propositions, Unsolvable Problems and Computable Functions. New York, NY; 1965

[66] Gold EM. Language identification in the limit. Information and Control. 1967;**17**:447-474

[67] Jantke KP. The main proof-theoretic problems in inductive inference. In: Wechsung G, editor. Frege Conference, 10–14 September 1984, Schwerin, Germany. Berlin, Germany: Akademie-Verlag, pp. 321-330

[68] Jantke KP, Arnold O. Patterns – The key to game amusement studies. In: Proceedings of the 3rd Global Conference on Consumer Electronics (GCCE 2014), 7–10 October 2014, Makuhari Messe, Tokyo, Japan. IEEE Consumer Electronics Society 2014, pp. 478-482

[69] Jantke KP. The pattern experience evaluation program. In: Proceedings of the Intl. Workshop on Design, Evaluation and Refinement of Intelligent Systems (DERIS 2009), 28 November. Vol. 2009. Kraków, Poland: AGH Univ. of Science and Technology; 2009. pp. 70-75

[70] Angluin D. Inductive inference of formal languages from positive data. Information and Control. 1980;**45**:117-135

[71] Lindtröm P. On extensions of elementary logic. Theoria. 1969;**35**:1-11

[72] Clocksin WF, Mellish CS. Programming in Prolog. Berlin, Heidelberg, New York: Springer; 1981

[73] Arnold O, Drefahl S, Fujima J, Jantke KP, Vogler C. Dynamic identification by enumeration for co-operative knowledge discovery. IADIS International Journal on Computer Science and Information Systems. 2017;**12**:65-85

[74] Gödel K. Über formal unentscheidbare Sätze der ”Principia Mathematica“ und verwandter Systeme. Monatshefte für Mathematik und Physik. 1931;**38**:173-198

[75] Grieser G. Reflective inductive inference of recursive functions. Theoretical Computer Science. 2008;**397**:57-69

[76] Jantke KP, Drefahl S, Arnold O. Characterizing the power and the limitations of concepts for adaptivity and personalization by benchmark results from inductive inference. In 12th Intl. Workshop on information search, integration, and personalization (ISIP 2018), 14-15 May 2018, Kyushu University, Fukuoka, Japan, 2018

Chapter 5

# Performance-Aware High-Performance Computing for Remote Sensing Big Data Analytics

Mustafa Kemal Pektürk and Muhammet Ünal

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.75934

**Abstract**

The incredible increase in the volume of data emerging along with recent technological developments has made the analysis processes which use traditional approaches more difficult for many organizations. Especially applications involving subjects that require timely processing and big data such as satellite imagery, sensor data, bank operations, web servers, and social networks require efficient mechanisms for collecting, storing, processing, and analyzing these data. At this point, big data analytics, which contains data mining, machine learning, statistics, and similar techniques, comes to the help of organizations for end-to-end managing of the data. In this chapter, we introduce a novel high-performance computing system on the geo-distributed private cloud for remote sensing applications, which takes advantages of network topology, exploits utilization and workloads of CPU, storage, and memory resources in a distributed fashion, and optimizes resource allocation for realizing big data analytics efficiently.

**Keywords:** big data analytics, high-performance computing, real-time analytics, remote sensing, distributed computing, geo-distributed cloud, resource allocation

## 1. Introduction

The extreme increase in the amount of data produced daily by many organizations reveals big challenges in data storage and extracting information from timely data [1–3]. Many sensors designed in today’s technology are used in observation centers and on the Earth to create a continuous stream of data [4]. Real-time, near-real-time geospatial data must be analyzed in a short time in order to be able to provide time-critical decision support in time-critical applications [5]. The development of efficient computing techniques for obtaining information from remote sensing (RS) big data is critical for Earth science [6, 7]. In particular, the recent developments in remote sensing technologies have had a tremendous increase in remote sensor data [8]. The amount of remote sensing (RS) data collected from a single satellite data center has dramatically increased and has reached several terabyte values per day [9]. This is because sensors have high resolution and a large amount of band due to new camera technologies. Thus, RS data reaching high dimensions is defined as "Big data."

Large image files which consist of both voluminous pixel and multiple spectral bands (multispectral/hyperspectral) cause great difficulties to read and store in memory. Besides this, data mining algorithms which extract information from satellite and remote sensing data involve high computational complexity. With these features, remote sensor applications are both data intensive [10, 11] and compute intensive [12]. However, the computational complexity of many data mining algorithms is super linear to the number of samples and the size of the sample. Hence, optimization of algorithms is not enough to obtain better performance when those two variables continue to increase.

When talking about big data analysis, the main difficulties in computer architecture are CPU intensiveness and slow input/output (I/O) operations. According to Moore's Law, CPU and driver performance doubles every 18 months. On the other hand, when the trends in I/O interfaces are examined, the improvement is near-to-network speed improvement but still behind of it [13]. Although I/O interfaces operate at the same speed as the processor bus, I/O lags behind because peripheral hardware cards operate at low speeds. PCI cards are being used to increase I/O performance. Thus I/O performance increases by about 33% annually. In spite of these improvements in traditional approaches, when collected data exponentially increases, usage of relatively slow data processing techniques makes real-time analysis difficult particularly [2]. For this reason, the most important factor determining the performance of analytical processes is the limited structure of hardware resources inherently [14]. Therefore, modern technologies and high-performance computing (HPC) techniques, which have parallel processing capabilities such as multi-core programming, GPU programming, field programmable gate arrays (FPGA), cluster computer, and cloud computing are needed to perform analysis on large volumes of data, which is complex and time-consuming to extract information [14–16].

The HPC system usage attracts more attention in remote sensing applications because of a great deal of data recently [6, 7]. HPC integrates some of the computing environments and programming techniques to solve large-scale problems in the remote sensing era. Many applications of remote sensing such as environmental studies, military applications, tracking and monitoring of hazards, and so on require real-time or near-real-time processing capabilities for urgent intervention that is timely in the necessary situation. HPC systems such as multi- or many-integrated core, GPU/GPGPU, FPGA, cluster, and cloud have become inevitable to meet this requirement. Multi-core, GPU, FPGA, and so on technologies meet parallel computing needs for onboard image processing particularly. In such technologies, developed algorithms run on a single node only but with multiple cores. So they could just scale vertically with hardware upgrades—for example, more CPUs, better graphics cards, more memories [17]. When distributed computing is considered, the most conceivable technology is commercial off-the-shelf (COTS) computer equipment, which is called cluster [6, 7]. In this approach, a cluster is created from a number of computers to work together as a team [18]. These parallel systems, installed with a large number of CPUs, provide good results for both real-time and near-real-time applications that use both remote sensing and data streams, but these systems are both expensive, and scalability cannot exceed a certain capacity. Although much of parallel systems are homogeneous inherently, the recent trend in HPC systems is the use of heterogeneous computing resources, where the heterogeneity is generally the result of technological advancement in progress of time. With increasing heterogeneity, cloud computing has emerged as a technology which aimed at facilitating heterogeneous and distributed computing platforms. Cloud computing is an important choice for efficient distribution and management of big datasets that cannot be stored in a commodity computer's memory solitarily.

Not only the increasing data volume but also the difficulty of indexing, searching, and transferring the data exponentially increases depending on data explosion [19, 20]. Effective data storage, management, distribution, visualization, and especially multi-modal processing in real/near-real-time applications are challenged as open issues for RS [21].

## 2. Big data

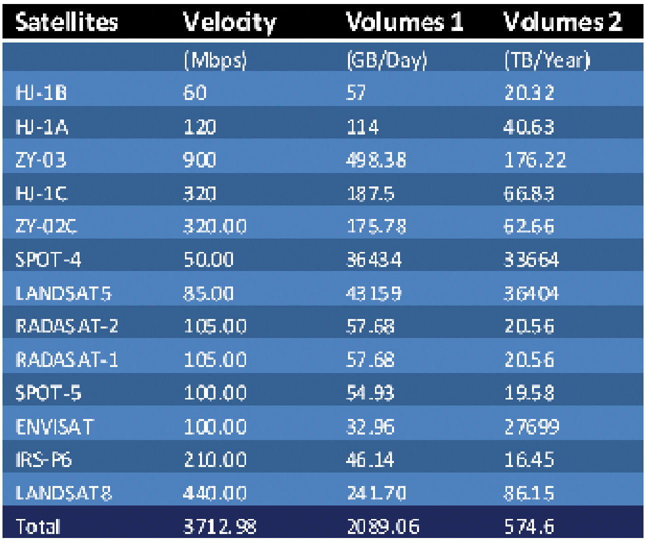
Big data is characterized as 3 V by many studies: volume, velocity, and variety [22, 23]. Volume is the most important big data quality that expresses the size of the dataset. The velocity indicates the rate of production of big data, but the increase in the rate of production also reveals the need for faster processing of the data. Variety refers to the diversity of different sources of data.

Given the variety of the data, the majority of the data obtained is unstructured or semi-structured [21]. Considering the velocity of data, velocity requirements vary according to application areas. In general, velocity is addressed under the heading of processing at a specific time interval, such as batch processing, near-real-time requirement, continuous input–output requirement real time, and stream processing requirements. Application of critical and live analytics-based batches to improve data and analysis processes requires continuous and real-time analysis, and critical applications require immediate intervention, depending on the analysis of incoming data streams.

### 2.1. Remote sensing data

Remote sensing (RS) is defined as the ability to measure the quality of a surface or object from a distance [9]. RS data are obtained from various data acquisition technologies (lidar, hyperspectral camera, etc.) in airplanes and unmanned aerial vehicles. With the recent developments in RS technologies, the amount, production rate, and diversity of remote sensor data have increased exponentially (**Table 1**). Thus, the data received from the remote sensors are being treated as RS "Big Data."

Diversity and multidimensionality are the greatest factors in the complexity of RS big data. RS data is used in a variety of geosciences with environmental monitoring, underground, atmospheric, hydrological, and oceanographic content. Due to such different and wide range of



| Satellites | Velocity | Volumes 1 | Volumes 2 |
|---|---|---|---|
| | (Mbps) | (GB/Day) | (TB/Year) |
| HJ-1B | 60 | 57 | 20.32 |
| HJ-1A | 120 | 114 | 40.63 |
| ZY-03 | 900 | 498.38 | 176.22 |
| HJ-1C | 320 | 187.5 | 66.83 |
| ZY-02C | 320.00 | 175.78 | 62.66 |
| SPOT-4 | 50.00 | 36434 | 33664 |
| LANDSAT5 | 85.00 | 43159 | 36404 |
| RADASAT-2 | 105.00 | 57.68 | 20.56 |
| RADASAT-1 | 105.00 | 57.68 | 20.56 |
| SPOT-5 | 100.00 | 54.93 | 19.58 |
| ENVISAT | 100.00 | 32.96 | 27699 |
| IRS-P6 | 210.00 | 46.14 | 16.45 |
| LANDSAT8 | 440.00 | 241.70 | 86.15 |
| Total | 3712.98 | 2089.06 | 574.6 |

**Table 1.** Satellite data centers, data rates, and volumes.

application areas, the diversity of RS data has increased greatly. There are approximately 7000 different types of RS datasets in NASA archives, as far as is known [9]. Numerous satellites and sensors with different resolutions have emerged due to higher spatial resolution, temporal resolution, and even spectral resolution. As remote sensing data continues to increase and complex, a new architecture has become a necessity for existing algorithms and data management [24].

### 2.2. Remote sensing algorithms

The processing of RS data has an important role in Earth observation systems. The longest RS workflow starts with data acquisition and ends with a thematic application (**Figure 1**). There can be processes which are processed sequentially or concurrently in each step within the workflow. An RS application typically consists of the following stages, respectively [9]:

- Data Acquisition/Collection: Images obtained from satellite or hyperspectral cameras are captured on the downlink channel. With high spatial and temporal resolution, data volume is increasing.
- Data Transmission: The data stream obtained is sent simultaneously to the satellite center. The high bandwidth of the data transfer is critical for applications requiring real-time processing. In reality, it is impossible to transmit real-time data because the data increase due to satellite and camera technologies is faster than the data transmission rate.
- Preprocessing: Preprocessing related to image quality and geographical accuracy such as decomposition, radiometric verification, and geometric verification is performed.

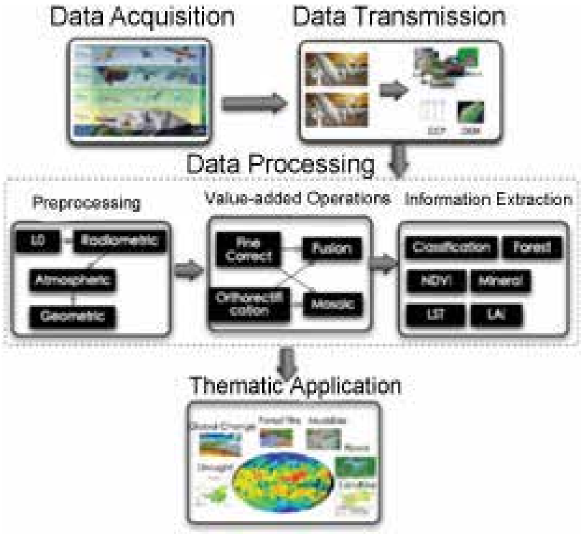


**Figure 1.** Remote sensing data-processing steps [9].

- Value-added Operations: Value-added operations such as fine-tuning, orthorectification, fusion, and mosaic are performed. It is possible to say that the mosaic process has a longer processing time than the others.
- Information Extraction/Thematic Implementation: In this step, classification, attribute extraction, and quantitative deductions are obtained with images subjected to preprocessing and value-added operations, for example, extracting information such as leaf area index, surface temperature, and snowy area on RS images.

### 2.3. Data access pattern in remote sensing algorithm

The data access patterns in the remote sensing algorithms vary according to the characteristics of the algorithms. There are four different access patterns for RS data. Depending on the size of the RS data, the unavoidable I/O load and irregular data access patterns make inapplicable the traditional cluster-based parallel I/O systems [25].

“Sequential row access pattern” is used by algorithms such as pixel-based processing-based radiometric verification, support vector machine (SVM) classifier. When algorithms are implemented in parallel, each processor needs logically multiple consecutive image rows.

“Rectangular block access pattern” is used by algorithms such as convolution filter and resample which require neighbor-based processing. Non-contiguous I/O patterns are visible in these algorithms and are not efficiently supported by normal parallel file systems.

“Cross-file access pattern” is used by algorithms like fusion and normalized difference vegetation index (NDVI) that require inter-band calculations. This data consists of small and non-contiguous fragments in hundreds of image files. Thus, in this type of access, a large number of read/write operations take place, which is a time-consuming process.

"Irregular access pattern" is used by algorithms such as fast fourier transform (FFT), image distortion, and information extraction, which require scattered access or the entire image. These algorithms use diagonal and polygonal access patterns as irregular access. In these patterns, different sized parts can be in different nodes, even though they are small and non-contiguous pieces of data. In addition to I/O difficulties, the problem of identifying irregular data areas also arises.

## 3. Real-time big data architecture

Big data architects often need a distributed system structure for data analysis, which requires data storage. S. Tehranian et al. proposed an architectural model that provides performance, reliability, and scalability, consisting of candidate hardware and software for distributed real-time processing of satellite data at ground stations [26]. The resulting prototype system was implemented using the open source adaptive communication environment (ACE) framework and C ++ and tested on the cluster; real-time systems have achieved significant performance without sacrificing reliability and high availability. Structures and mechanisms for parallel programming are needed so that RS data can be analyzed in a distributed manner. In this context, Y. Ma et al. proposed a generic parallel programming framework for RS applications on high-performance clusters [27]. The proposed mechanism has programming templates that provide both distributed and generic parallel programming frameworks for RS algorithms. The proposed magHD for storage, analysis, and visualization of multidimensional data combines Hadoop and MapReduce technologies with various indexing techniques for use on clusters [28]. Some systems, such as [29], contain all of the different steps of addition, filtering, load balancing, processing, combining, and interpreting. In the related work, a real-time approach to continuous feature extraction and detection aimed at finding rivers, land, and rail from satellite images was proposed using the Hadoop ecosystem and MapReduce.

At a minimum, the components that the big data architecture should have in order to be able to perform real-time analysis are as follows: user interface, distributed file system, distributed database, high-performance computing (**Figure 2**).

A distributed file system is a virtual file system that allows distributed data to be stored on multiple computer clusters. In clusters that may consist of heterogeneous nodes, the application provides a common interface for accessing the data in isolation from the operating and file systems.

A distributed database consists of separate databases on each node/server in a multi-computer cluster connected by a computer network. The distributed database management system allows distribution and management of data on the server.

High-performance computing systems are technologies that provide an infrastructure that provides a sufficiently fast computing environment for parallel analysis of big data. It is crucial for the system to respond to the user within the reasonable time.

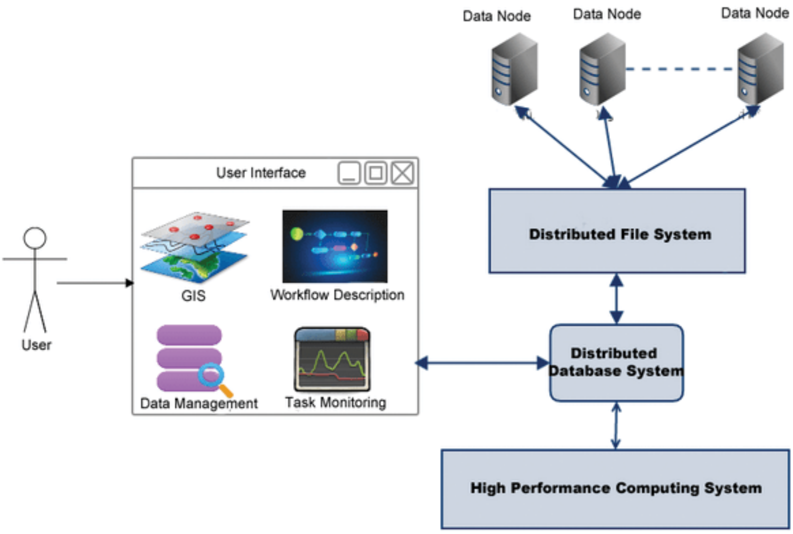


**Figure 2.** Generalized real-time big data architecture.

The user interface is a component that basically allows the user to query the server-based application through a visual interface, query, load, delete, update, request analysis, define workflow. The user interface has not been investigated as a major concern in this chapter.

### 3.1. Storage

#### *3.1.1. Distributed file system*

SSD and PCM devices that are being used instead of HDD as data storage are far from the I/O performance required for big data. Enterprise storage architectures, such as DAS, NAS, and SAN, have drawbacks and limitations when used as distributed systems [2].

Distributed File System (DFS) is a virtual file system that spans multiple nodes on the cloud [15]. It provides the abstraction of heterogeneity of data nodes in different centers. Thus, the distributed file system provides a common interface for applications to access data on heterogeneous nodes that use different operating systems and different file systems on different nodes individually. There are some capabilities that a distributed file system should generally provide. Location transparency is where the application can access data such as being held locally without actually having to hold it. Access transparency is a common interface for access to data independent of the operating system and the file system. Fault tolerance is the ability to keep a replica of a replica on more than one node so that in the event of an error the replica is preserved in the nodes holding the replica. Scalability means that the number of nodes the file system is running on can be increased to the required amount (without the error system hanging down) if needed.

The Hadoop distributed file system (HDFS) is an open-source distributed file system distributed with an Apache license (**Figure 3**). HDFS is designed especially for big datasets and high

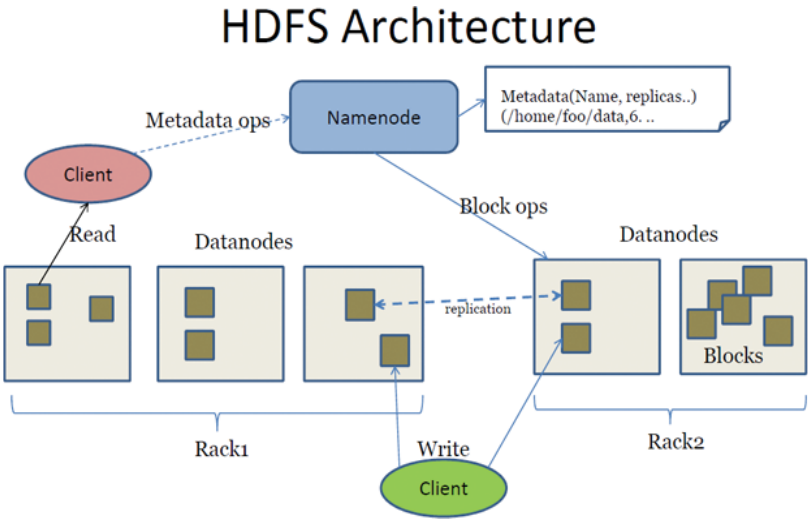


**Figure 3.** HDFS architecture.

availability, apart from the common abilities. It is also platform independent as it is implemented in Java. Applications are accessed via the HDFS API, which is maintained by any filing system. Thus, file access is isolated from local file systems. Compared to other distributed file systems (IRODS, Luster), it is stated that the performance is different in design and HDFS is the only DFS with automatic load balancing [15]. At the same time, because it is platform independent and the availability of MapReduce support makes it easy to use on many systems, it is the preferred choice.

The Google file system (GFS) is a proprietary distributed file system developed by Google for its own use [30]. The reason for the development is the need for a scalable distributed file system that emerges in big data-intensive applications. It is designed to enable reliable, efficient, and fault-tolerant use of data in a multitude of thousands of drives and machines, each with thousands of simultaneous users.

Storage systems such as amazon simple storage (S3), nirvanix cloud storage, openstack swift, and windows azure blob that are used in cloud systems do not fully meet the scalability and replication needs of cloud applications and the concurrency and performance requirements of analysis applications.

General parallel file system (GPFS) is a high-performance clustered file system developed by IBM. GPFS can be built on shared drives or shared-nothing distributed parallel nodes. Since it fully supports the POSIX-based file system, it removes the need to learn the new API set introduced by other storage systems. On the other hand, HDFS and GFS are not completely POSIX compliant and require new API definitions to provide analysis solutions in the cloud. In the study conducted by Schmuck et al., it is stated that GPFS is in terms of file-reading performance of HDFS with a meta-block concept [31]. A meta block is a set of consecutive data blocks located on the same disk. In the proposed approach, a trade-off was attempted between different block sizes.

Parallel virtual file system (PVFS) compared with HDFS [32] shows that PVFS has not shown a significant improvement in terms of completion time and throughput.

Rasdaman database is a database that supports large multidimensional arrays that conventional databases cannot handle and can store large remote sensing data by nature [33]. The architecture of Rasdaman is based on the sequence shredding process called "tiling." The Rasdaman parallel server architecture feature provides a scalable and distributed environment for efficient processing of large numbers of concurrent user requests. Thus it is possible to present distributed datasets over the web. In order to retrieve and process the dataset from Rasdaman, queries of data retrieval in the query language defined by open geospatial consortium's (OGC) web coverage processing service (WCPS) standards should be run. The PetaScope component, developed as a Java Servlet used at this point, provides queries for multidimensional data retrieval, retrieval filtering, and processing by implementing OGC standard interfaces. It also adds support for geographic and temporal coordinate systems.

Depending on the size of the RS data in the remote sensing applications, the unavoidable I/O load and irregular data access patterns are not applicable to traditional cluster-based parallel I/O systems [25]. In the study conducted by L. Wang et al., an RS file-based parallel file system for remote sensing applications was proposed and implemented using the OrangeFS file system. By providing an application-specific data placement policy, efficiency is achieved for different data access patterns. The improvement in the performance of the proposed system is seen as an average of 20%.

### *3.1.2. Distributed database*

The classical approaches used in managing structured data have a schema for data storage and a relational database for retrieving data. Existing database management tools have been inadequate for processing large volumes that grow rapidly and become complex. Data warehouse and data-market approaches have gained popularity in systems with more than one structured data [2]. One of these approaches is the data warehouse, which is used to store, analyze, and report results to the user. The data market (March) approach is an approach that improves data access and analysis based on the data warehouse. The enterprise data warehouse (EDW), which is favored by large organizations, allows the data processing and analysis capability to be used on a very large and unified enterprise database [21]. Some cloud providers can offer a petabyte data and more scaling solution with EDW. For example, Amazon Redshift uses a massively parallel processing (MPP) architecture consisting of a large number of processors for high-performance interrogation, with columnar storage and data compression. In addition, the amount of I/O required by queries is reduced using local attached storage and zone maps.

For storing and managing unstructured or non-relational data, the NoSQL approach is divided into two independent parts: data storage and management [2]. With the key-value storage model in storage, NoSQL's focal point is scalability and high performance of data storage. In the management section, data management tasks can be performed at the application layer through the lower-level access mechanism. The most important features of the NoSQL database are the ability to quickly change the data structure by providing schema freedom and the need to rewrite the data so that the structured data can be stored heterogeneously, providing flexibility. The most popular NoSQL database is the Cassandra database, which was first used by Facebook and published as open source in 2008. There are also NoSQL implementations such as SimpleDB, Google BigTable, MongoDB, and Voldemort. Social networking applications such as Twitter, LinkedIn, and Netflix also benefited from NoSQL capabilities.

According to the method proposed by L. Wang et al. for the management problem of conventional remote sensing data, the image data is divided into blocks based on the GeoSOT global discrete grid system and the data blocks are stored in HBase [34]. In this method, the data is first recorded in the MetaDataInfo table. The satellite-sensor acquisition time is used as the row ID. In the DataGridBlock table, the row ID is kept with the MetaDataInfo row ID as well as the geographic coordinate. HBase tables ensure that blocks that are geographically close to the ascending order of row numbers will be held in adjacent rows in the table. When a spatial query arrives, the GeoSOT codes are first calculated and the DataGridBlock table is filtered by these codes. In addition, a distributed processing method that uses MapReduce model to deal with image data blocks is also designed. When MapReduce starts the job, it splits the table into bounds of regions, each region containing a set of image data blocks. The map function then processes each data block from the region and sends the resulting results to the reduce function.

The analysis of ultra-big databases has attracted many researchers' interest, as traditional databases are inefficient for storing and analyzing large digital data. Apache HBase, the NoSQL distributed database developed on HDFS, is one of the results of these researches. A study by M.N. Vora evaluated a hybrid approach in which HDFS retains data such as non-textual images and HBase retains these data [35]. This hybrid architecture makes it possible to search and retrieve data faster.

### 3.2. HPC systems

When data analysis is considered, the most important difficulty is scalability, depending on the volume of data. In recent years, researchers have focused more on accelerating their analysis algorithms. However, the amount of data is much faster than CPU speed. This has led processors to come to a position to support parallel computing as multi-core. Timeliness for real-time applications comes first. Thus, many difficulties arise not only in hardware development but also in the direction of development of software architects. The most important trend at this point is to make distributed computing improvements using cloud computing technology.

The technologies used in remote sensing applications have difficulties in delivering, processing, and responding in time [36]. Web technologies, grid computing, data mining, and parallel computation on remote sensing data generated by R. Patrick and J. Karpjoo have been scanned. The size of the data volume, the data formats, and the download time are general difficulties.

With the combination of betting technologies, the processing time in some applications can be reduced to as short a time as can be decided by the helpers in a timely manner. Although it is reasonable to process the remote sensing data as soon as possible, it does not seem possible to perform real-time processing automatically.

#### *3.2.1. Onboard architecture*

##### *3.2.1.1. Multi-core processor*

The multi-core processor is an integrated circuit which has two and more processors to process multiple tasks efficiently. After the frequency of processors reached limits due to the heating problem, processing capabilities of new CPUs continue to increase by multiplying the number of cores [37]. Recently new CPUs could handle 8–12 simultaneous threads. To benefit from the multi-core CPUs, the problem should be divided into partitions which can be processed simultaneously. Multi-core programming is needed to achieve this benefit and rewriting of application is needed affordably. Multi-core programming is the implementation of algorithms using a multi-core processor on a single computer to improve performance. Some APIs and standards such as OpenMP and MPI are needed to implement algorithms which could be run simultaneously on multi-core processors.

##### *3.2.1.2. Graphic processing units*

GPU is a specialized circuit dedicated to graphical processing preliminarily. After it has had a brilliant rise in manipulating computer graphics and image processing in recent years, this technology gets used for developing parallel algorithms on RS image data widely [38–40]. RS complex algorithms should be rewritten to benefit from GPU parallelism by using thousands of simultaneous threads.

##### *3.2.1.3. Field programmable gate arrays*

Some onboard remote sensing data processing scenarios require components that can operate with low weight and low power, especially in systems where air vehicles such as unmanned air vehicle and satellite are used. While these components reduce the amount of payload, they can produce real/near-real-time analysis results at the same time as data is being obtained from the sensor. For this purpose, programmable hardware devices such as FPGAs can be used [41, 42]. FPGAs are the digital integrated circuits which consist of an array of programmable logic blocks and reconfigurable interconnects that allow the blocks to be connected simply. But the need for FPGA programming and learning a new set of APIs is emerging.

#### *3.2.2. Distributed architecture*

##### *3.2.2.1. Cluster*

One of the most used approaches when considering hardware-based improvements is commercial off-the-shelf (COTS)-based computer-based solutions. In this approach, a cluster is created from a number of computers to work together as a team [43]. These parallel systems, installed with a large number of CPUs, provide good results for both real-time and near-real-time applications using both remote sensors and data streams, but these systems are both expensive, and scalability does not exceed a certain capacity.

Cavallaro et al. have addressed the classification of land cover types over an image-based dataset as a concrete big data problem in their work [44]. In the scope of the study, PiSVM, an implementation based on LibSVM, was used for classification. The PiSVM code is stale and stable despite the I/O limits. While PiSVM is used in parallel, MPI is used for communication on multiple nodes. For the parallel analysis, the JUDGE cluster in Jülich Supercomputing Center in Germany was used. The training period has been reduced significantly in the PiSVM, which runs parallel to the running of the series MATLAB. In parallel operation, the accuracy of SVM remains the same as in serial operation (97%).

#### *3.2.2.2. Cloud*

Cloud computing is one of the most powerful big data techniques [45]. The ability to provide flexible processing, memory, and drivers by virtualizing computing resources on a physical computer made the supercomputing concept more affordable and easily accessible [46]. The use of the cloud concept, which provides a multi-computer infrastructure for data management and analysis, provides great ease in terms of high scalability and usability, fault tolerance, and performance. Especially considering the critical applications that need to extract information from the data that the next-generation remote sensors can produce near real time, it is very important to use cloud computing technologies for high-performance computing [21]. In addition, the cloud computing infrastructure has the ability to create an efficient platform for the storage of big data as well as for the performance of the analysis process. Thus, together with the use of this technology, expensive computing hardware such as cluster systems, allocated space and software requirements can be eliminated [22].

The Hadoop ecosystem has emerged as one of the most successful infrastructures for cloud and big data analysis [23, 45]. The platform brings together several tools for various purposes, with two major services: HDFS, a distributed file system, and MapReduce, a high-performance parallel-data processing engine. The MapReduce model is an open-source implementation of the Apache Hadoop framework. This model allows big datasets to be distributed concurrently on multiple computers. Remote sensing applications using the MapReduce model have become a research topic in order to improve the performance of the analysis process as a result of exponential growth within the latest developments in sensor technologies [15, 16]. The processing services on the cloud are accessed via the distributed file system. In order to reduce data access, it is reasonable to process the data in the central computer where the data is stored.

## 4. Performance-aware HPC

Processing of big geospatial data is vital for time-critical applications such as natural disasters, climate changes, and military/national security systems. Its challenges are related to not only massive data volume but also intrinsic complexity and high dimensions of the geospatial datasets [47]. Hadoop and similar technologies have attracted increasingly in geosciences communities for handling big geospatial data. Many investigations were carried out for adopting those technologies to processing big geospatial data, but there are very few studies for optimizing the computing resources to handle the dynamic geo-processing workload efficiently.

In existing software systems, computing is seen as the most expensive part. After daily collected data amount is grown exponentially with recent technologies, data movement has a deep impact on performance with a bigger cost than computing which is cheap and massively parallel [48]. At that point, new high-performance systems need to update themselves to adapt to the data-centric paradigm. New systems must use data locality to succeed that adaptation. Current systems ignore the incurred cost of communication and rely on the hardware cache coherency to virtualize data movement. Increasing amount of data reveals more communication between the processing elements and that situation requires supporting data locality and affinity. With the upcoming new model, data locality should be succeeded with recent technologies which contain tiling, data layout, array views, task and thread affinity, and topology-aware communication libraries. Combination of the best of these technologies can help us develop a comprehensive model for managing data locality on a high-performance computing system [49].

### 4.1. Data partition strategy

Domain decomposition is an important subject in high-performance modeling and numerical simulations in the area such as intelligence and military, meteorology, agriculture, urbanism, and search and rescue [50]. It makes possible parallel computing by dividing a large computational task into smaller parts and distributing them to different computing resources. To increase the performance, high-performance computing is extensively used by dividing the entire problem into multiple subdomains and distributing each one to different computing nodes in a parallel fashion. Inconvenient allocation of resources induces imbalanced task loads and redundant communications among computing nodes. Hereby, resource allocation is a vital part which has a deep impact on the efficiency of the parallel process. Resource allocation algorithm should minimize total execution time by taking into consideration the communication and computing cost for each computing node and reduce total communication cost for the entire processing. In this chapter, a new data partitioning strategy is proposed to benefit from the current situation of resources on the cloud. At this new strategy, RS data should be partitioned based on performance metrics which is formulated by a combination of available resources of network, memory, CPU, and storage. After appropriate cloud site and resource nodes are found by stage 1 and 2 as described in Sections 6.2 and 6.3, the receiving data is divided into the selected resource nodes according to the number of available resources on that cloud. For this dividing operation, the system should determine which portion of data will be allocated to found nodes based on some performance metrics such as the below heuristic function:

$$p_i = \frac{t^i_{throughput} + t^i_{processing}}{\sum_{j=1}^{n} t^j_{throughput} + t^j_{processing}} \tag{1}$$

$p_i$ is the portion of node $i$ and $t^i_{throughput}$ are the transfer time needed for data with network throughput of $i$ and processing time needed for data with CPU frequency of $i$. $\sum_{j=1}^{n} t^j_{throughput} + t^j_{processing}$ is the total sum of transfer and processing time for selected nodes. In the formula, transfer time can be computed with $t^i_{throughput} = data\_size/bandwidth_i$ and processing time can be computed with $t^i_{processing} = data\_size/CPU\_frequency_i$.

### 4.2. Geo-distributed cloud

Geo-distributed cloud is the application of cloud computing technologies which consist of multiple cloud sites distributed in different geographic locations to interconnect data and applications [51]. Cloud providers prefer the distributed cloud systems to enable lower latency and provide better performance for cloud consumers. Recently, most of the large online services have been geo-distributed toward the exponential increase in data [52]. The most important reason for realizing the services as geo-distributed is latency. Geo-distributed clouds provide a point of presence nearby clients with reference to reduce latency. In this chapter, we introduce a novel resource allocation technique for managing RS big data in the geo-distributed private cloud. This new approach will select the most appropriate cloud site which has minimum latency. It also finds efficient data layout for data which gives a higher performance for selected data nodes in the related cloud site. Within this context, resource management should match instantaneous application requirements with optimal CPU, storage, memory, and network resources [53]. Putting the data into a more appropriate node with enough resources and providing an efficient layout of the dependent data partitions on the nodes with minimum latency on the network should decrease processing time of algorithms and also minimize transferring time of dependent data which is needed by algorithm processing on different nodes.

In the proposed approach, acquired RS data should be stored on geo-distributed cloud within two stages (**Figure 4**). In the first stage, each cloud site determines a score based on latency, bandwidth capacity, CPU, memory, and storage workloads. At that point, each cloud site has a multi-criteria decision-making (MCDM) process. MCDM is a subdiscipline of operation research that evaluates multiple criteria in decision-making problems [54]. As a criteria value for the resource, workloads of resources could be computed by dividing the used amount of the resource to the total capacity of it. For CPU workload, the equation is given as follows:

$$\mathrm{W^i_{CPU}} = \frac{\left(\sum_{j=1}^{n} \mathrm{c_{ij}}\right)}{\mathrm{t^i_{CPU}}} \tag{2}$$

Let $t^i_{CPU}$ be the total number of physical CPU threads in the cloud site $i$ and $c_{ij}$ be the number of virtual CPUs that are allocated for virtual machines (VMs) in the cloud site. For storage workload, the equation is given as follows:

$$\mathrm{W^i_{STR}} = \frac{\left(\sum_{j=1}^{n} \mathrm{s_{ij}}\right)}{\mathrm{t^i_{STR}}} \tag{3}$$

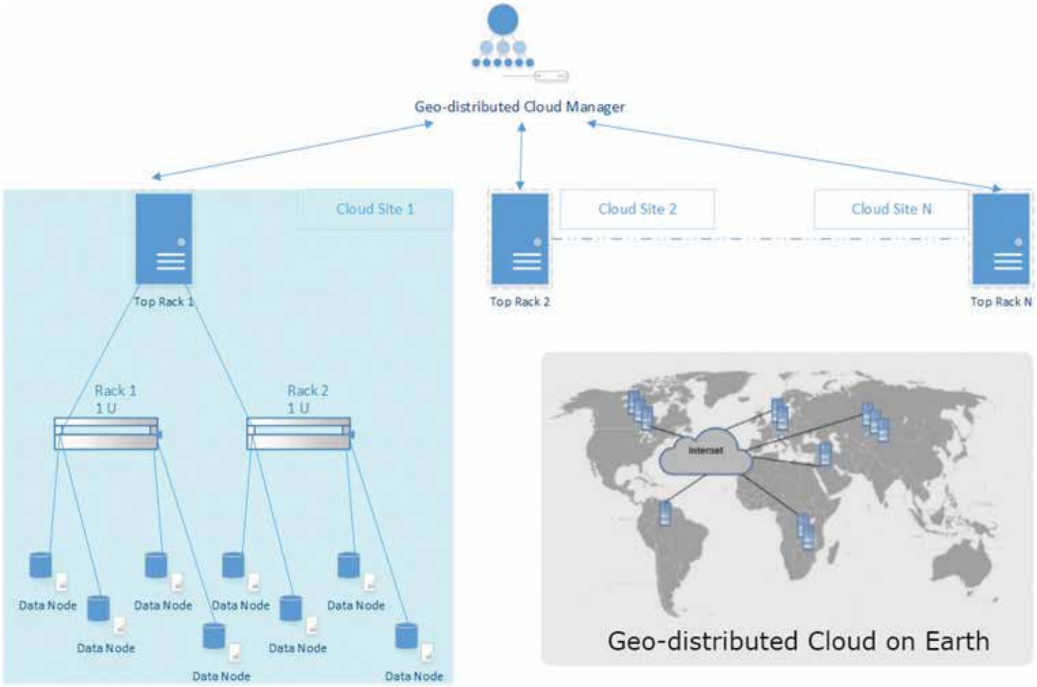


**Figure 4.** Geo-distributed cloud hierarchy and scene on earth.

Let $t^i_{STR}$ be the total size of storage in the cloud site $i$ and $s_{ij}$ be the size of storage that is used for coming data in the cloud site. For memory workload, the equation is given as follows:

$$W^i_{MEM} = \frac{\left(\sum_{j=1}^{n} m_{ij}\right)}{t^i_{MEM}} \tag{4}$$

Let $t^i_{MEM}$ be the total size of memory in the cloud site $i$ and $m_{ij}$ be the size of storage that is used for coming data in the cloud site.

Time of transferring data between consumer and cloud site is described as latency$L_i$, bandwidth between them is $B_i$.

After determining criteria for decision-making, a processing method is needed to compute numerical values using the relative importance of the criteria to determine a ranking of each alternative [54]. Some of the well-known MCDM processing models are weighted sum model (WSP), weighted product model (WPM), and analytic hierarchy process (AHP). If we define the solution with AHP which is based on decomposing a complex problem into a system of hierarchies, the best alternative could be defined as the below relationship:

$$A_{AHP} = \min_i \sum_{j=1}^{N} w_{ij} C_j, \quad \text{for } i = 1, 2, 3, \ldots, M. \tag{5}$$

where $\sum_{j=1}^{N} w_{ij} = 1$, N is criteria amount, and *i* cloud site. If we write the model for earlier five criteria:

$$\mathrm{A_{AHP}} = \mathrm{min_i}\left(\mathrm{w_1 L_i} + \mathrm{w_2}\frac{1}{\mathrm{B_i}} + \mathrm{w_3 W^i_{CPU}} + \mathrm{w_4 W^i_{STR}} + \mathrm{w_5 W^i_{MEM}}\right) \text{ for } \mathrm{i} = 1, 2, 3, \ldots, \mathrm{M} \qquad (6)$$

Although AHP is similar to WSM, it uses related values instead of actual values. This makes it possible to use the AHP in multidimensional decision-making problems by removing the problem of combining different dimensions in various units (similar to adding oranges and apples).

After the minimum valued alternative cloud site is found, the second stage takes place for evaluating which resources should be used optimally in the related cloud site and finding an optimal layout in the network for RS big data.

### 4.3. Resource optimization in cloud network with performance metrics

Large-scale networks and its applications lack centralized access to information [55]. When we interpret RS big data on the cloud as a large-scale network, optimization of resource allocation depends on local observation and information of each node. At this point, control and optimization algorithms should be deployed in a distributed manner for finding optimum resource allocation in such a network. Optimization algorithm should be robust against link or node failures and scalable horizontally.

To succeed distributed optimization of the system, each node should run the optimization algorithm locally which can be called as an agent. At this system which consists of multi-agents connected over a network, each agent has a local objective function and local constraint set which are known by just this agent. The agents try to decide on a global decision vector cooperatively based on their objective function and constraints (**Figure 5**).

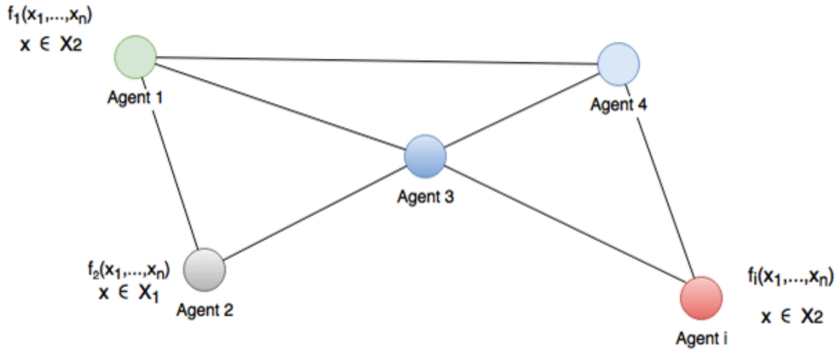


**Figure 5.** Multi-agent optimization problem for resource allocation on a network. $\mathrm{f_i(x)}$: Local objective function where f: $\mathbb{R}^n \to \mathbb{R}$. $\mathrm{X_i}$: Local constraint set where $\mathrm{X_i} \subset \mathbb{R}^n$. x: Global decision vector which agents collectively try to decide on, where $\mathrm{x} \in \mathbb{R}^n$.

The agents cooperatively optimize a global objective function denoted by f(x), which is a combination of the local objective functions, that is:

$$f(x) = T(\ f_1(x), \ldots, f_i(x)\ ) \tag{7}$$

where $T : \mathbb{R}^n \rightarrow \mathbb{R}$ and decision vector x is bounded by a constraint set, where x ∈ C, which consists of local constraints and global constraints that may be imposed by the network structure, that is:

$$C = \left( \bigcap_{i=1}^{n} X_i \right) \cap C_g \tag{8}$$

where $C_g$ represent global constraints. This model leads to following optimization problem:

$$minimize\, f(x)\ subject\ to\ x \in C \tag{9}$$

where $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and the set C is constraint set. The decision vector in Eq. (9) can be considered as resource vector whose component corresponds to resources allocated to each node or global decision vector which is estimated by the nodes on the network using local information.

After defining some basic notation and terminology for optimization of a function, continue with where we left off. The cloud site should decide which resources will be used for coming RS data when it is determined to receive it. Each node in the cloud site evaluates network latency and bandwidth between other nodes for optimum network communications. In addition to that, current amounts of CPU, memory, and storage are also taken into account together for finding the best-fitted solution to store and process RS data. Hence each node solves the defined formula:

$$\min f_i(x) = \sum_{j=1}^{n} x_j \left[ \left( w_1 L_{ij} + w_2 \frac{1}{B_{ij}} + w_3 H_{ij} + w_4 W^j_{CPU} + w_5 W^j_{STR} + w_6 W^j_{MEM} \right) \right]$$

$$\text{for } i = 1, 2, 3, \ldots, N$$

$$\text{subject to} \begin{cases} \sum_{j=1}^{n} x_j A^j_{MEM} \geq I_{size} \\ \sum_{j=1}^{n} x_j A^j_{STR} \geq I_{size} \\ H_{ij} \leq Hop_{max} \\ B_{ij} \leq C_{max} \\ x_{i,j} \in \{0, 1\} \forall i, j \end{cases} \tag{10}$$

where $x_j$ indicates that the jth node would be in the data-receiving group together with node $i$ or not, $L_{ij}$ is latency between $i$ and $j$, $B_{ij}$ is bandwidth between $i$ and $j$, $H_{ij}$ is hop count between $i$ and $j$, and $W^j_{CPU}$, $W^j_{STR}$, and $W^j_{MEM}$ are CPU, storage, and memory resources in node $j$. Each node should solve $f_i(x)$ to minimize decision vector with AHP model in a decentralized manner. The individual optimization problem is a mix integer program for each node. Lagrangian relaxation is a heuristic method that stands for solving mix integer problems with decomposing constraints. The idea of the method is to decompose constraints which complicated the problem by adding them to the objective function with the associated vector μ called the Lagrange multiplier. After applying it to our problem, we derive the dual problem of Eq. (9) for an efficient solution by adding complicated constraints 1 and 2.

$$\min L(x, \mu, \lambda) = \sum_{j=1}^{n} x_j U_{ij} + \mu_j \left( I_{size} - \sum_{j=1}^{n} x_j A^{j}_{MEM} \right) + \lambda_j \left( I_{size} - \sum_{j=1}^{n} x_j A^{j}_{STR} \right)$$

$$\text{subject to} \begin{cases} H_{ij} \leq Hop_{max} \\ B_{ij} \leq C_{max} \\ x_{i,j} \in \{0, 1\} \forall i, j \end{cases} \tag{11}$$

where utilization function for node:

$$U_{ij} = \left[ \left( w_1 L_{ij} + w_2 \frac{1}{B_{ij}} + w_3 H_{ij} + w_4 W^{j}_{CPU} + w_5 W^{j}_{STR} + w_6 W^{j}_{MEM} \right) \right]$$

and $\mu_j \geq 0 \forall j$ and $\lambda_j \geq 0 \forall j$ are the dual variables.

After obtaining above optimization problem, it could be separable in the variables $x_i$ and it also decomposes into sub-problems for each node i. Thus each node needs to solve the one-dimensional optimization problem Eq. (11). This optimization problem consists of its own utility function and Lagrangian multipliers which are available for node i. Generally, the subgradient method is used to solve the obtained dual problem because of simplicity of computations per iteration. First-order methods such as subgradient have slower convergence rate for high accuracy but they are very effective in large-scale multi-agent optimization problems where the aim is to find near-optimal approximate solutions [55].

When one optimal approximate solution is found, RS data should be divided into partitions and proportionally distributed to found nodes in the solution according to performance heuristic (Eq. (1)) which is given in Section 6.1.

## 5. Conclusion

As a result of technological developments, the amount of data produced by many organizations on a daily basis has increased to terabyte levels. Remotely sensed data, which is spatially and spectrally amplified and heterogeneous by means of different sensing techniques, causes great difficulties in storing, transferring, and analyzing with conventional methods. It has become a necessity to implement distributed approaches instead of conventional methods that are inadequate in critical applications when real/near-real-time analysis of relevant big data is needed. Existing distributed file systems, databases, and high-performance computing systems are experiencing difficulties in optimizing workflows for analysis, in the storage, and retrieval of spatial big data of remote sensing and data streams.

According to investigated researches in this chapter, it is observed that existing techniques and systems cannot find a solution that covers the existing problems when analyzing the real and near-real-time big data analysis in remote sensing. Hadoop and similar technologies have attracted increasingly in geosciences communities for handling big geospatial data. Many investigations were carried out for adopting those technologies to processing big geospatial data, but there are very few studies for optimizing the computing resources to handle the dynamic geo-processing workload efficiently.

In this chapter, a two-stage innovative approach has been proposed to store RS big data on a suitable cloud site and to process them with optimizing resource allocation on a geo-distributed cloud. In the first stage, each cloud site determines a score based on latency, bandwidth capacity, CPU, memory, and storage workloads with an MCDM process. After minimum valued alternative cloud site is found, the second stage takes place for evaluating which resources should be used optimally in related cloud site and finding an optimal layout in the network for RS big data with respect to latency, bandwidth capacity, CPU, memory, and storage amount. Lastly, data should be divided into partitions based on a performance metric which could be computed with available network and processing resources of selected nodes in the cloud site.

As future work, optimal replication methods will be searched for preventing failure situations when transferring and processing RS data in a distributed manner. For succeeding that, a performance-based approach is considered to maintain high-performance computing.

## Author details

Mustafa Kemal Pektürk[1]* and Muhammet Ünal[2]

*Address all correspondence to: mkpekturk@havelsan.com.tr

1 HAVELSAN A.Ş., Ankara, Turkey

2 Gazi University, Ankara, Turkey

## References

[1] Fadiya SO, Saydam S, Zira VV. Advancing big data for humanitarian needs. Procedia Engineering. 2014;**78**:88-95

[2] Chen CLP, Zhang C-Y. Data-intensive applications, challenges, techniques and technologies: A survey on big data. Information Sciences. 2014;**275**:314-347

[3] Özköse H, Arı ES, Gencer C. Yesterday, today and tomorrow of big data. Procedia-Social and Behavioral Sciences. 2015;**195**:1042-1050

[4] Rathore MMU et al. Real-Time Big Data Analytical Architecture for Remote Sensing Application. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 2015;**8**(10):4610-4621

[5] Yue P et al. Sensor Web event detection and geoprocessing over big data. In: 2014 IEEE International Geoscience and Remote Sensing Symposium (IGARSS). IEEE; 2014

[6] Lee CA et al. Recent developments in high performance computing for remote sensing: A review. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 2011;**4**(3):508-527

[7] Toma AC et al. Computational challenges in processing large hyperspectral images. In: Tier 2 Federation Grid, Cloud & High Performance Computing Science (RO-LCG), 2012 5th Romania. IEEE. 2012. pp. 111-114

[8] Ma Y et al. Remote sensing big data computing: Challenges and opportunities. Future Generation Computer Systems. 2014;**51**:47-50

[9] Ma Y et al. Towards building a data-intensive index for big data computing–a case study of remote sensing data processing. Information Sciences. 2014;**319**:171-188

[10] Aji A et al. Hadoop GIS: A high performance spatial data warehousing system over mapreduce. Proceedings of the VLDB Endowment. 2013;**6**(11):1009-1020

[11] Zhong Y, Fang J, Zhao X. VegaIndexer: A distributed composite index scheme for big spatio-temporal sensor data on cloud. In: 2013 IEEE International Geoscience and Remote Sensing Symposium (IGARSS). IEEE; 2013

[12] Wickramaarachchi C et al. Real-time Analytics for fast evolving social graphs. In: 2015 15th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid). IEEE; 2015

[13] Oliveira SF, Fürlinger K, Kranzlmüller D. Trends in computation, communication and storage and the consequences for data-intensive science. In: 2012 IEEE 14th International Conference on High Performance Computing and Communication & 2012 IEEE 9th International Conference on Embedded Software and Systems (HPCC-ICESS). IEEE; 2012

[14] Fernández A et al. Pattern recognition in Latin America in the "big data" era. Pattern Recognition. 2015;**48**(4):1185-1196

[15] Lin F-C et al. The framework of cloud computing platform for massive remote sensing images. In: 2013 IEEE 27th International Conference onAdvanced Information Networking and Applications (AINA). IEEE; 2013. pp. 621-628

[16] Krämer M, Senner I. A modular software architecture for processing of big geospatial data in the cloud. Computers & Graphics. 2015;**49**:69-81

[17] Bernabe S et al. Hyperspectral unmixing on GPUs and multi-core processors: A comparison. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 2013;**6**(3):1386-1398

[18] Chen J, Zheng G, Chen H. ELM-MapReduce: MapReduce accelerated extreme learning machine for big spatial data analysis. In: 2013 10th IEEE International Conference on Control and Automation (ICCA). IEEE; 2013

[19] Kambatla K et al. Trends in big data analytics. Journal of Parallel and Distributed Computing. 2014;**74**(7):2561-2573

[20] Hashem IAT et al. The rise of "big data" on cloud computing: Review and open research issues. Information Systems. 2015;**47**:98-115

[21] Song G et al. Constructing gazetteers from volunteered big geo-data based on Hadoop. Computers, Environment and Urban Systems; 2017;**61**:172-186

[22] Yang C et al. A spatiotemporal compression based approach for efficient big data processing on cloud. Journal of Computer and System Sciences. 2014;**80**(8):1563-1583

[23] Douglas CC. An open framework for dynamic big-data-driven application systems (DBDDAS) development. Procedia Computer Science. 2014;**29**:1246-1255

[24] Konstantinos K, Bliziotis D, Karmas A. A scalable geospatial web service for near real-time, high-resolution land cover mapping. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 2015;**8**(10):4665-4674

[25] Wang L et al. A parallel file system with application-aware data layout policies for massive remote sensing image processing in digital earth. IEEE Transactions on Parallel and Distributed Systems. 2015;**26**(6):1497-1508

[26] Tehranian S et al. A robust framework for real-time distributed processing of satellite data. Journal of Parallel and Distributed Computing. 2006;**66**(3):403-418

[27] Ma Y et al. Generic parallel programming for massive remote sensing data processing. In: 2012 IEEE International Conference on Cluster Computing (CLUSTER). IEEE; 2012

[28] Angleraud C. magHD: a new approach to multi-dimensional data storage, analysis, display and exploitation. In: IOP Conference Series: Earth and Environmental Science. Vol. 20. No. 1. IOP Publishing; 2014

[29] Rathore MMU et al. Real-time continuous feature extraction in large size satellite images. Journal of Systems Architecture. 2016;**64**:122-132

[30] SUPPLY, POWER. Zhou et al. (45) Date of Patent: Aug 26, 2014

[31] Schmuck FB, Haskin RL. GPFS: A shared-disk file system for large computing clusters. FAST. 2002;**2**

[32] Tantisiriroj W et al. On the duality of data-intensive file system design: reconciling HDFS and PVFS. In: Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis. ACM; 2011

[33] Assunção MD et al. Big data computing and clouds: Trends and future directions. Journal of Parallel and Distributed Computing. 2015;**79**:3-15

[34] Wang L et al. Massive remote sensing image data management based on HBase and GeoSOT. In: 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS). IEEE; 2015

[35] Vora MN. Hadoop-HBase for large-scale data. In: 2011 International Conference on Computer Science and Network Technology (ICCSNT). Vol. 1. IEEE; 2011

[36] Tumwizere RP, Karpjoo J. A survey on computing technology applications in remote sensing. In: 2012 8th International Conference on Computing and Networking Technology (ICCNT). IEEE; 2012

[37] Setoain J et al. GPU for parallel on-board hyperspectral image processing. The International Journal of High Performance Computing Applications. 2008;**22**(4):424-437

[38] Qu H et al. Parallel acceleration of SAM algorithm and performance analysis. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 2013;**6**(3):1172-1178

[39] Qu H et al. Parallel implementation for SAM algorithm based on GPU and distributed computing. In: 2012 IEEE International Geoscience and Remote Sensing Symposium (IGARSS). IEEE; 2012

[40] Wu Z et al. Parallel implementation of sparse representation classifiers for hyperspectral imagery on GPUs. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 2015;**8**(6):2912-2925

[41] Shan N, Wang X-S, Wang Z-S. Efficient FPGA implementation of cloud detection for real-time remote sensing image processing. In: 2010 Asia Pacific Conference on Postgraduate Research in Microelectronics and Electronics (PrimeAsia). IEEE; 2010

[42] Plaza A et al. Parallel implementation of hyperspectral image processing algorithms. In: IEEE International Conference on Geoscience and Remote Sensing Symposium. IGARSS 2006. IEEE; 2006

[43] Plaza A et al. High performance computing for hyperspectral remote sensing. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 2011;**4**(3):528-544

[44] Cavallaro G et al. Scalable developments for big data analytics in remote sensing. In: 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS). IEEE; 2015

[45] O'Driscoll A, Daugelaite J, Sleator RD. 'Big data', Hadoop and cloud computing in genomics. Journal of Biomedical Informatics. 2013;**46**(5):774-781

[46] Vinay A et al. Cloud based big data analytics framework for face recognition in social networks using machine learning. Procedia Computer Science. 2015;**50**:623-630

[47] Li Z et al. Automatic scaling hadoop in the cloud for efficient process of big geospatial data. ISPRS International Journal of Geo-Information. 2016;**5**(10):173

[48] Tate A et al. Programming abstractions for data locality. In: PADAL Workshop 2014, April 28–29. Lugano: Swiss National Supercomputing Center (CSCS); 2014

[49] Pektürk MK, Ünal M. A review on real-time big data analysis in remote sensing applications. In: 2017 25th Signal Processing and Communications Applications Conference (SIU). IEEE; 2017

[50] Gui Z et al. Developing subdomain allocation algorithms based on spatial and communicational constraints to accelerate dust storm simulation. PLoS One. 2016;**11**(4):e0152250

[51] Wu Y et al. Scaling social media applications into geo-distributed clouds. IEEE/ACM Transactions on Networking (TON). 2015;**23**(3):689-702

[52] Narayanan I et al. Towards a leaner geo-distributed cloud infrastructure. In: HotCloud; 2014

[53] Tso FP, Jouet S, Pezaros DP. Network and server resource management strategies for data Centre infrastructures: A survey. Computer Networks. 2016;**106**:209-225

[54] Triantaphyllou E et al. Multi-criteria decision making: An operations research approach. Encyclopedia of Electrical and Electronics Engineering. 1998;**15**(1998):175-186

[55] Nedic A, Ozdaglar A. 10 cooperative distributed multi-agent. Convex Optimization in Signal Processing and Communications. 2010;**340**

Chapter 6

# Early Prediction of Patient Mortality Based on Routine Laboratory Tests and Predictive Models in Critically Ill Patients

Sven Van Poucke, Ana Kovacevic and
Milan Vukicevic

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.76988

## Abstract

We propose a method for quantitative analysis of predictive power of laboratory tests and early detection of mortality risk by usage of predictive models and feature selection techniques. Our method allows automatic feature selection, model selection, and evaluation of predictive models. Experimental evaluation was conducted on patients with renal failure admitted to ICUs (medical intensive care, surgical intensive care, cardiac, and cardiac surgery recovery units) at Boston's Beth Israel Deaconess Medical Center. Data are extracted from Multi parameter Intelligent Monitoring in Intensive Care III (MIMIC-III) database. We built and evaluated different single (e.g. Logistic regression) and ensemble (e.g. Random Forest) learning methods. Results revealed high predictive accuracy (area under the precision-recall curve (AUPRC) values >86%) from day four, with acceptable results on the second (>81%) and third day (>85%). Random forests seem to provide the best predictive accuracy. Feature selection techniques Gini and ReliefF scored best in most cases. Lactate, white blood cells, sodium, anion gap, chloride, bicarbonate, creatinine, urea nitrogen, potassium, glucose, INR, hemoglobin, phosphate, total bilirubin, and base excess were most predictive for hospital mortality. Ensemble learning methods are able to predict hospital mortality with high accuracy, based on laboratory tests and provide ranking in predictive priority.

**Keywords:** mortality risk prediction, renal failure, metabolic panel, feature selection, ensemble methods

# 1. Introduction

Precision medicine is based on comprehensive models with the potential to elucidate the complexity of health and diseases, including the features of emergence, nonlinearity, self-organization, and adaptation [1].

Laboratory testing is more common among patients admitted to ICU [2, 3]. Blood sample frequencies vary, but routinely tests are ordered by fixed schedule and in clusters as part of the hypothetico-deductive diagnostic exploration. Quantitative predictive analysis of daily sampling might provide new insights into the choice (feature selection) and importance (feature weighting) of each laboratory test [4]. In this chapter, we propose a system for mortality risk prediction of patients with renal failure, based on predictive methods. Renal failure patients were selected based on the Elixhauser Comorbidity Index [5]. For chronic disease, the use of Elixhauser is sensitive for the systemic underrepresentation of chronic conditions [6, 7].

This study quantitatively assessed the predictive power of laboratory tests for hospital mortality in patients admitted to ICU. Based on previous findings, we compared the predictive performance of different single (Decision Tree, Naive Bayes, Logistic and Regression) and ensemble (Random Forest, Boosting, and Bagging) learning methods. Moreover, the predictive power and importance of predictors (laboratory tests) were quantitatively assessed by use of feature weighting and selection techniques: Correlation, Gini Selection, Information Gain and ReliefF [8]. For predictive modeling, feature selection, and visual analytics of the results, we used RapidMiner and R platforms as mentioned in [9–11].

# 2. Materials and methods

## 2.1. Data source and study subjects

The MIMIC-III (version 1.0) clinical database consists of 58,976 ICU admissions for 46,520 distinct patients, admitted to Beth Israel Deaconess Medical Center (Boston, MA) from 2001 to 2012 [12, 13]. The establishment of the database was approved by the Institutional Review Boards of the Massachusetts Institute of Technology (Cambridge, MA) and Beth Israel Deaconess Medical Center (Boston, MA). Accessing the database was approved for authors S.V.P and Z.Z. (certification number: 1712927 and 1132877). Informed consent was waived due to observational nature of the study.

The MIMIC-III clinical database includes data related to patient demographics, hospital admissions and discharge dates, room tracking, death dates (in or out of the hospital), ICD-9 codes, health care providers, and types. All dates were surrogate dates but time intervals were preserved. In addition, physiological data, medications consumption, laboratory investigations, fluid balance calculations and notes, and reports were included in the basic dataset.

### 2.2. Data preparation

RapidMiner was used because it enabled handling unstructured data without the need for coding [9, 14].

The dataset in this study was generated by joining data from the following MIMIC-III tables: admission, patients, ICU stays, diagnoses_icd, and lab events. Patients were assigned to subpopulations including hypertension, paralysis, chronic pulmonary disease, diabetes, renal failure, acquired immunodeficiency syndrome (AIDS), coagulopathy, obesity, and weight loss, and so on, based on the Elixhauser comorbidity score [6, 7]. Renal failure is defined in the Elixhauser comorbidity score, when ICD-9 code is in (70.32, 70.33, 70.54, 456, 456.1, 456.2, 456.21, 571, 571.2, 571.3, ≥ 571.4 ≤ 571.49, 571.5, 571.6, 571.8, 571.9, 572.3, 572.8, and 42.7).

All time stamped measurements in MIMIC-III were zeroed in reference to the moment of hospital admission.

### 2.3. Predictive algorithms

The process compared different learning and ensemble methods (Decision Stump, Decision Tree, Naive Bayes, Logistic Regression (LR), Random Forest, Support Vector Machine, AdaBoost, Bagging, and Stacking) in association with feature weighting and selection, quantitatively assessed in terms of Correlation, Gini Selection, and Information Gain and ReliefF as previously described [8].

#### *2.3.1. Single learning methods*

Decision trees (DT) are predictive algorithms based on "greedy," top-down recursively partitioning of data. DT algorithms perform an exhaustive search over all possible splits in every recursive step. The attribute (predictor) demonstrating the best split by an evaluation measure selected for branching the tree. Regularly used are information theoretic measures (e.g. Information Gain, Gain Ratio, Gini, etc.) or statistical tests quantifying the significance of the association between predictors and class. The procedure is recursively iterated until a stop criterion is met [15, 16]. In this research, we used the J48 algorithm, which is the Java implementation of the C4.5 algorithm [17].

Logistic regression (LR) is a linear classifier modeling the probability of a dependent binary variable y given a vector of independent variables X. For the estimation of the probability, the example belongs to the positive class, a logit model is used:

$$log\left(\frac{p}{1-p}\right) = \Theta_0 + \Theta_1 x_1 + \cdots + \Theta_n x_n \tag{1}$$

where p presents probability that y = 1, θj, j = 1,…,n present the weights of the corresponding dependent variable, while p/(1-p) is called odds ratio, parameters θj, j = 1,…,n of the model can be interpreted as changes in log odds or the results can be interpreted in terms of probabilities [18–20].

#### *2.3.2. Ensemble learning methods*

Ensemble (meta-learning) methods combine multiple models aiming to provide more accurate or more stable predictions. These models can be aggregated from the same model built on different sub-samples of data, from different models built on the same sample or a combination of the previous two techniques. Ensemble methods are often used to improve the individual performance of algorithms that constitute ensembles by exploiting the diversity among the models produced [21]. The ensemble methods implemented in this chapter are: Random Forest [22], Boosting [23], and Bootstrap Aggregating (Bagging) [24]. In our experiments, Boosting and Bagging used J4.8 and Logistic regression as base learners.

Random Forest (RF) is an ensemble classifier that evaluates multiple DT and aggregates their results, by majority voting, in order to classify an example [22]. There is a two-level randomization in building these models. First, each tree is trained on a bootstrap sample of the training data and second, in each recursive iteration of building a DT (splitting data based on information potential of features); a subset of features for evaluation is randomly selected. In this research, we grew and evaluated Random Forest (RF) with 10 trees.

Boosting is an ensemble meta-algorithm developed in order to improve supervised learning performance of weak learners (models whose predictive performance is only slightly better than random guessing). In this study, the adaptive boosting (AdaBoost) algorithm was used [23].

Bagging algorithm builds a series of models (e.g. CHAID Decision Trees) on different data subsamples (with replacement) [24]. For new examples, each model is applied, and predictions are aggregated (e.g. majority voting for classification or average prediction for regression).

### 2.4. Feature weighting and selection

Several filter feature selection schemes were evaluated. Filter selection (FS) methods rely on the evaluation of the information potential of each input feature in relation to the label (hospital mortality). A threshold search and selection of those features, providing most predictive power, was calculated for each predictive model. The first is based on Pearson correlation returning the absolute or squared value of the correlation as attribute weight. Furthermore, we applied Information Gain Ratio and Gini Index, two weighting schemes that are based on information theoretic measures, frequently used with decision trees for evaluation of potential splits [17]. The T-test calculated, for each attribute, a p-value for two-sided, two-sample T-test. Finally, the ReliefF evaluated the impact of an attribute by repeatedly sampling an instance and considering the value of the given attribute for the nearest instance of the same and different class [25].

## 3. Experimental evaluation

### 3.1. Exploratory analyses

The MIMIC-III database consists of 58,976 hospital admissions from 46,520 patients. All patients are characterized by at least one ICU admission.

Guided by the CRoss-Industry Standard Process for Data Mining (CRISP-DM), the ETL process was initiated by retrieving data from the MIMIC-III tables of interest (d_labitems, admissions, patients, and diagnosis_icd) [26]. Next, patients were selected for renal failure by the Elixhauser score, leading to 1477 (3.15%) patients satisfying our inclusion criteria and 20,068 patient days (examples) in total. In a consecutive step, admissions were joined based on hospital admission id (hadm_id) with all laboratory tests (from the 755 item ids in d_labitems), aggregated on a daily level. Mean, standard deviation, and the number of tests per day (*len*) were defined as aggregation functions. As an output feature (*label*), this study focused on hospital mortality (hospital_expire_flag). From all renal failure patients in this study, 399 (27.0%) did not survive during hospital admission. Next, data were split per day in order to examine feature selection and weight changes over time. Therefore, we arbitrarily limited our computations for admission duration of 7 days, where for each day the number of patients was >1000. After that period, the number of patients admitted to ICU declined.

Patients who survived hospital stay were significantly older (69.3 ± 12.4 years vs. 65.9 ± 14.1 years; $p < 0.05$), suffered more frequently from deficiency anemia (15.5 vs. 9.8% $p = 0.01$) and depression (8.3 vs. 3.8% $p = 0.00$). The survivors suffered less frequently from congestive heart failure (40.2 vs. 46.9% $p = 0.02$), valvular disease (9.8 vs. 14.3% $p = 0.01$), lymphoma (1.6 vs. 3.8% $p = 0.01$), and metastatic cancer (1.7 vs. 4.5% $p = 0.00$). **Table 1** displays the basic characteristics of the baseline dataset. Binary variables are reported as prevalence percentages or count, and continuous variables are reported as data mean ± standard deviation.

In **Figure 1**, distributions of numbers of laboratory tests by admission days are described by the box plots for each day demonstrating a decline in the number of different laboratory tests requested by admission days from day 1 to day 4. For the following days, the number of requested laboratory tests was stable.

### 3.2. Automatic model building, feature selection and evaluation

A more detailed technical description of the use of RapidMiner for scalable predictive analytics of medical data, as well as templates of generic processes, can be found in [8] and its supplementary materials.

Initially, all features are weighted by five feature weighting and selection methods (Information Gain ratio, Gini, Correlation, ReliefF, and T-test), for each day. In order to find the adequate number of features that will be used by each predictive model for each day (and to identify optimal feature selection methods for our data), we conducted the following procedure. First, we sorted the features by their weights in descending order (for each feature weighting method). Then we trained each of five predictive models (Decision tree, Logistic regression, Random Forest, Bagging, and Boosting) on subsets of features with highest weights, starting from 10 features up to 100 with the step of 10 (9 different feature sets) [27, 28]. Even though a number of experiments were conducted (315 experiments: 7 algorithms X 5 feature selection schemes X 9 thresholds), this method as previously described [8] allowed ease of implementation of the experimental setup within only one RapidMiner process execution and with complete reproducibility of the results.

| Characteristics | ICU patients with renal failure (n = 1477) | Survival during hospital admission (n = 1078) (73.0%) | Death during hospital admission (n = 399) (27.0%) | p |
|---|---|---|---|---|
| Age (years) | 66.8 ± 13.8 | 69.3 ± 12.4 | 65.9 ± 14.1 | *<0.05* |
| Sex (male, %) | 60.1 | 59.4 | 61.9 | |
| Congestive heart failure | 620 (42.0) | 433 (40.2) | 187 (46.9) | *0.02* |
| Cardiac arrhythmias | 438 (29.7) | 307 (28.5) | 131 (32.8) | 0.10 |
| Valvular disease | 163 (11.0) | 106 (9.8) | 57 (14.3) | *0.01* |
| Pulmonary circulation | 95 (6.4) | 68 (6.3) | 27 (6.8) | 0.75 |
| Peripheral vascular disease | 274 (18.6) | 199 (18.5) | 75 (18.8) | 0.88 |
| Hypertension | 13 (0.9) | 8 (0.7) | 5 (1.3) | 0.35 |
| Paralysis | 19 (1.3) | 13 (1.2) | 6 (1.5) | 0.65 |
| Other neurological | 70 (4.7) | 46 (4.3) | 24 (6.0) | 0.16 |
| Chronic pulmonary disease | 253 (17.1) | 188 (17.4) | 65 (16.3) | 0.60 |
| Diabetes uncomplicated | 322 (21.8) | 225 (20.9) | 97 (24.3) | 0.15 |
| Diabetes complicated | 435 (28.5) | 326 (30.2) | 109 (27.3) | 0.27 |
| Hypothyroidism | 155 (10.5) | 120 (11.1) | 35 (8.8) | 0.19 |
| Renal failure | 1477 (100) | 1078 (100.0) | 399 (100.0) | |
| Liver disease | 76 (5.1) | 51 (4.7) | 25 (6.3) | 0.24 |
| Peptic ulcer | 12 (0.8) | 10 (0.9) | 2 (0.5) | 0.42 |
| Aids | 19 (1.3) | 14 (1.3) | 5 (1.3) | 0.94 |
| Lymphoma | 32 (2.2) | 17 (1.6) | 15 (3.8) | *0.01* |
| Metastatic cancer | 36 (2.4) | 18 (1.7) | 18 (4.5) | *0.00* |
| Solid tumor | 68 (4.6) | 46 (4.3) | 22 (5.5) | 0.31 |
| Rheumatoid arthritis | 44 (3.0) | 31 (2.9) | 13 (3.3) | 0.70 |
| Coagulopathy | 149 (10.1) | 106 (9.8) | 43 (10.8) | 0.59 |
| Obesity | 34 (2.3) | 28 (2.6) | 6 (1.5) | 0.21 |
| Weight loss | 60 (4.1) | 37 (3.4) | 23 (5.8) | *0.04* |
| Fluid electrolyte | 572 (38.7) | 414 (38.4) | 158 (39.6) | 0.68 |
| Blood loss anemia | 0 (0.0) | 0 | 0 | 0 |
| Deficiency anemias | 206 (13.9) | 167 (15.5) | 39 (9.8) | *0.01* |
| Alcohol abuse | 41 (2.8) | 29 (2.7) | 12 (3.0) | 0.74 |
| Drug abuse | 25 (1.7) | 19 (1.8) | 39 (1.5) | 0.73 |
| Psychoses | 42 (2.8) | 32 (3.0) | 10 (2.5) | 0.63 |
| Depression | 105 (7.1) | 90 (8.3) | 15 (3.8) | *0.00* |

**Table 1.** Patient characteristics of the baseline dataset.

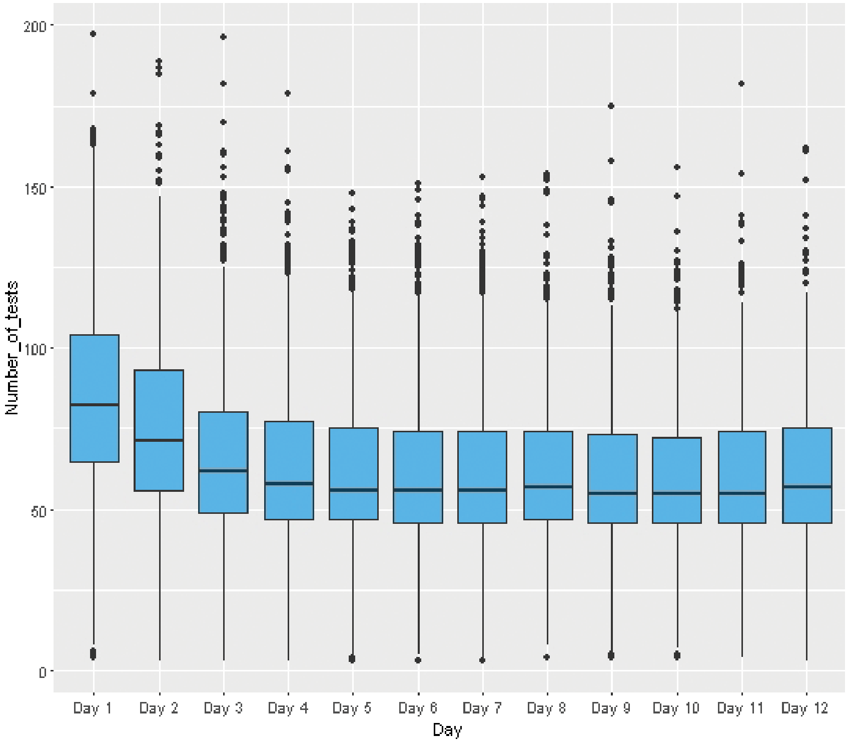


**Figure 1.** Distribution of the number of laboratory tests per patient by days.

Evaluation of all predictive models was performed by AUPRC (area under the precision-recall curve) for model comparison, because of the unbalanced nature of data [27, 28]. Namely, frequently used area under receiver operating curve (AUROC) is calculated based on true positive rate and false positive rate. True positive rate may be high even if recall is low (situation when predictor rarely predicts positive class), and thus it is often misleading in case of imbalanced data.

Because of relatively small number of samples (between 1000 and 1500 for each day), the predictive performance of models on unseen data was estimated by a fivefold cross-validation set created by stratified sampling, preserving the initial distribution of positive and negative classes of the target attribute. This validation on relatively small samples avoids the risk of misleading interpretation of the results caused by biased selection of a test set based on one sample.

### 3.3. Performance and feature selection

First, we present a comparison between feature selection methods, based on maximal predictive performance (in terms of AUPRC) overall algorithms. Next, we restrict further analyses on experiments with the overall best feature selection technique. Values in **Table 2** illustrate the maximal predictive performance for each day and for each feature selection technique. Maximum values by days (rows) are shown in bold. It can be seen that Gini and ReliefF achieved maximum values on all days, except for the first day of ICU admission.

| Day/FS method | Info Gain Ratio | Gini | Correlation | ReliefF | T-test |
|---|---|---|---|---|---|
| 1 | 0.44 | 0.47 | **0.48** | 0.45 | 0.33 |
| 2 | 0.79 | **0.84** | 0.82 | **0.84** | 0.77 |
| 3 | 0.80 | **0.85** | **0.85** | **0.85** | 0.78 |
| 4 | 0.83 | **0.86** | **0.86** | **0.86** | 0.78 |
| 5 | 0.84 | **0.86** | **0.86** | **0.86** | 0.80 |
| 6 | 0.83 | **0.86** | 0.85 | **0.86** | 0.77 |
| 7 | 0.84 | **0.86** | **0.86** | **0.86** | 0.77 |

Bold values represent the best results per rows. Multiple bold Values per row means that there was more than one equally good results.

**Table 2.** Maximum area under the precision-recall curve (AUPRC) performance of algorithms per days (rows) and feature selection measures (columns).

On the first day, the maximal performance is achieved with correlation (0.48). A more detailed inspection of the model performance and the number of features selected for each admission day demonstrated that Random Forest, result in the best predictive performance overall days (except the first day). Logistic regression often achieved a good performance. J4.8. achieved the worst AUPRC performance over all days, but in synergy with the AdaBoost ensemble scheme, it provided a competitive performance with Random Forest and Logistic Regression.

Further, **Table 2** illustrates that predictive performance is increasing over days and stabilizes from day 4 to 7 on AUPRC = 0.86. AUPRC values for days 2 and 3 are also high (0.83 and 0.85, respectively) and illustrate that risk for hospital mortality can be predicted, with high confidence, starting from the second day of admission.

Values of area under precision-recall curve illustrate the general performance of predictive models but do not explain anything related to the selection of the actual thresholds that should be selected for predictions. Therefore, we analyzed possible thresholds by inspecting precision-recall (PR) trade-off. High recall means that most of the positive examples (in this case, hospital death) are predicted correctly. High precision means that there is a low number of false alarming (mortality is predicted, but the patient survived). **Figure 2** shows precision/recall (PR) curves for first 4 days that were generated from the predictions of the best performing models (**Table 3**), built on features from the Gini selection.

On the first day, all models resulted in poor results which are found on the upper left PR curve in **Figure 2**.

Highest recall can be achieved with 0.3 precision (70% of false alarms), so this model is not useful, regardless of the threshold selection. On days 2–4, maximal recall can be achieved with around 20% of false alarms. Considering the cost of false negative predictions (low recall), we argue that optimal threshold values for day 2–4 models should be between 0.8 and 1 of recall.

Further, rankings of features provided by Gini feature selection methods illustrated in **Table 3** demonstrate that different algorithms achieved best predictive accuracies with different numbers of selected features. The number of features selected varied over the days.

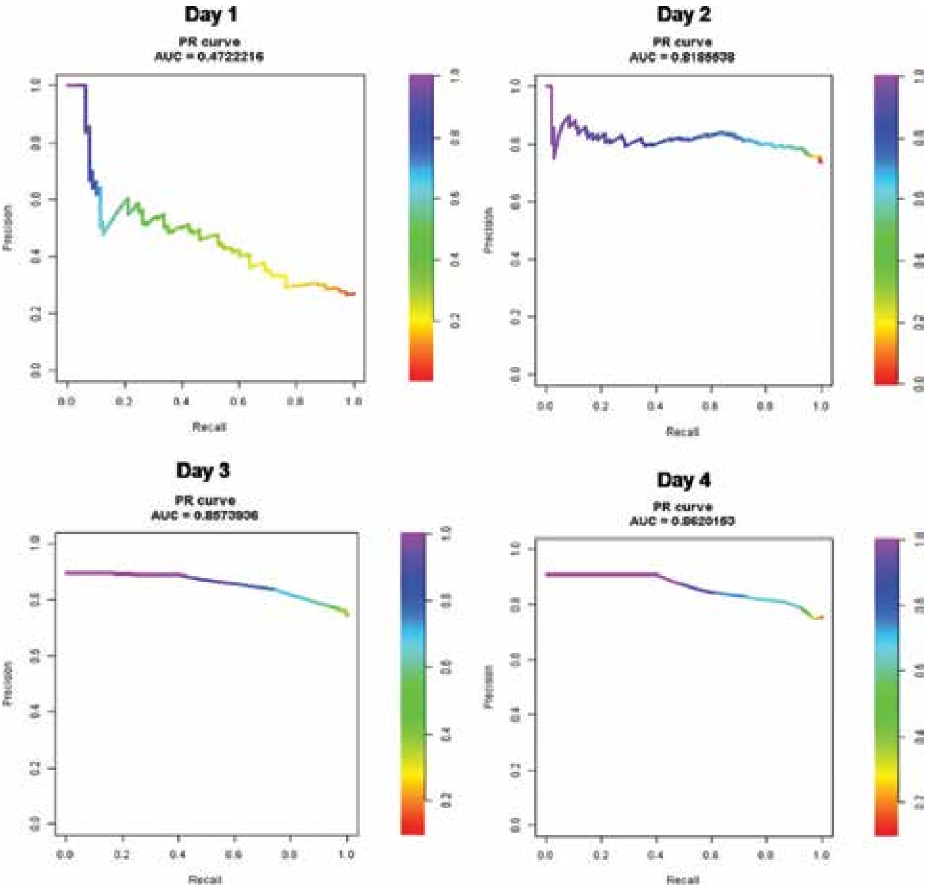


**Figure 2.** PR curves for first 4 days that were generated from the predictions of the best performing models, built on features from the Gini selection.

| Day | J4.8 | Logistic | Random Forest | AdaBoost (J 4.8) | AdaBoost (Logistic) | Bagging (J4.8) | Bagging (Logistic) |
|---|---|---|---|---|---|---|---|
| 1 | 0.35 (10) | **0.47 (60)** | 0.42 (30) | 0.37 (10) | 0.42 (30) | 0.36 (20) | 0.46 (30) |
| 2 | 0.76 (60) | **0.83 (50)** | **0.83 (70)** | 0.81 (30) | 0.82 (60) | 0.78 (50) | 0.82 (60) |
| 3 | 0.78 (30) | 0.84 (20) | **0.85 (40)** | 0.84 (60) | 0.82 (90) | 0.76 (10) | 0.83 (20) |
| 4 | 0.79 (20) | 0.85 (20) | **0.86 (40)** | **0.86 (10)** | 0.83 (20) | 0.78 (10) | 0.84 (10) |
| 5 | 0.81 (20) | 0.85 (30) | **0.86 (80)** | 0.85 (20) | 0.83 (20) | 0.8 (10) | 0.84 (10) |
| 6 | 0.82 (20) | 0.83 (40) | **0.86 (70)** | **0.86 (20)** | 0.82 (90) | 0.8 (20) | 0.83 (10) |
| 7 | 0.80 (10) | 0.84 (30) | **0.86 (80)** | 0.85 (10) | 0.81 (30) | 0.78 (10) | 0.84 (30) |

Bold values represent the best results per rows. Multiple bold Values per row means that there was more than one equally good results.

**Table 3.** Algorithms performance per admission day based on Gini index feature selection measure.

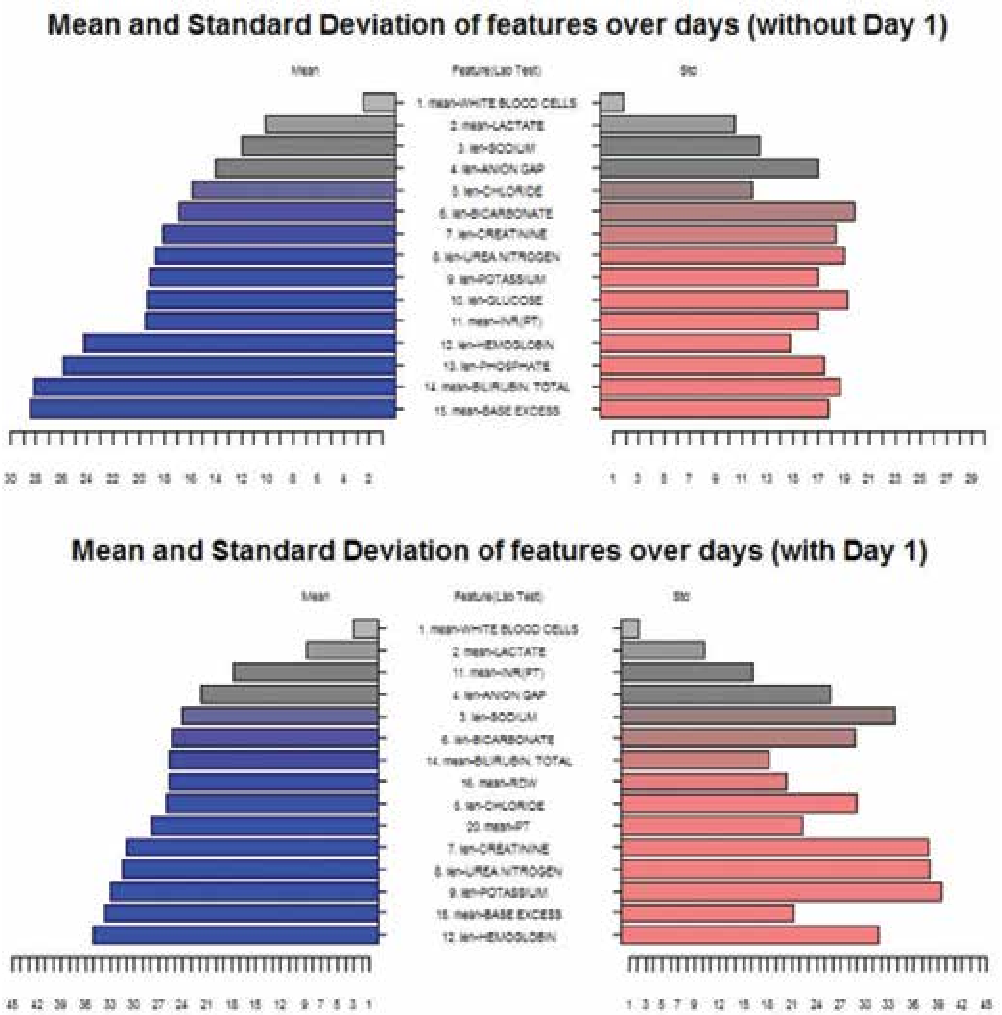


**Figure 3.** Pyramid chart of feature ranking (laboratory tests) for predictive value by mean and standard deviation; in- and -excluding day 1.

Finally, we analyzed if there was the difference, between features selected on the first day, where the predictive performance was consistently poor, and other days, where the predictive performance was acceptable. Rank means and standard deviations were calculated for two groups: all days (with day one) and days 2–7 (without day 1) (**Figure 3**). Standard deviations of ranks are much higher over ranks of features that include day 1 (right parts of **Figure 3**). The average ranks changed (middle part of the figure), but similar laboratory tests were in first 15 ranks in both cases.

## 4. Discussion

This study using ensemble methods demonstrated an improvement in predictive accuracy compared to prediction based on single models. Random Forests seem to provide the best predictive accuracy complying with our previous research [8]. (**Table 3**) Random Forest also resulted in a high predictive accuracy for mortality risk prediction [29]. This study, however, did not analyze different ensemble and feature selection methods and was conducted on a different population. In addition, the feature selection techniques Gini Index and ReliefF scored best in the majority of the cases.

Laboratory tests ranked per day based on Random Forest and Gini Index. (*mean*, *std.*: standard deviation, *len*: number of tests/day) indicated the importance of mean lactate values (ranked first on day 1, 2, and 8), and mean white blood cells count (on day 3, 4 and 6) in the prediction of hospital mortality. In addition, it is fascinating to observe that predictive features for hospital mortality calculated from 755 laboratory related parameters and without any additional patient-related information or medical knowledge, correlated well with the laboratory tests used on a daily basis (sodium, anion gap, chloride, bicarbonate, creatinine, urea nitrogen, potassium, glucose, INR, hemoglobin, phosphate, total bilirubin, and base excess). Laboratory-based clinical decision support may improve physician adherence to guidelines with respect to timely monitoring of chronic kidney disease [30, 31].

As demonstrated in this study, parameters for shock (lactate), sepsis (white blood cells), and multi-organ failure are more important [32].

The ensemble models in this study were able to generate a high predictive accuracy (AUPRC values) from day 4, with acceptable results on the second and third day. On the first day of admission, however, AUPRC values were very low but correlated well with diagnostic uncertainty on the first day of admission.

Surprisingly, patients who survived hospital stay were significantly older but suffered less from lymphoma and metastatic cancer. These findings might indicate some admission bias for certain comorbidities or indicate a constitutional superiority of older people admitted to ICU despite renal failure.

The hospital mortality in this renal failure ICU population was 27.0% (399/1477) [33]. Laboratory testing alone is only a part of the daily assessment of ICU patients. More research could elaborate predictive analysis of laboratory tests and other patient-related data in different patient populations.

## 5. Conclusions

Predictive analytics using ensemble methods are able to predict hospital or ICU outcome of renal patients with high accuracy. Predictive accuracy changes with the length of stay. Feature ranking enables quantitative assessment of patient data (e.g. laboratory tests) for predictive power. Lactate and white blood cell count best predict hospital mortality in this population. From the second day of ICU admission, predictive accuracy based on laboratory tests >80%. This generates opportunities for efficacy and efficiency analysis of other data recorded during ICU stay.

## Acknowledgements

Licenses for RapidMiner/Radoop are provided by RapidMiner (Cambridge, MA, and USA).

Support for data extraction was kindly provided by Wessel Luijben from Anchormen (Amsterdam, The Netherlands). RapidMiner (RapidMinerAcademia) provided a free use of the commercial version of its platform to researchers and other academics at educational institutions.

## Author details

Sven Van Poucke[1], Ana Kovacevic[2] and Milan Vukicevic[3]*

*Address all correspondence to: vukicevicm@fon.bg.ac.rs

1 Department of Anesthesiology, Intensive Care, Emergency Medicine and Pain Therapy, Ziekenhuis Oost-Limburg, Genk, Belgium

2 Saga Ltd., New Frontier Group, Belgrade, Serbia

3 Faculty of Organizational Sciences, University of Belgrade, Belgrade, Serbia

## References

[1] Yan Q. From pharmacogenomics and systems biology to personalized care: A framework of systems and dynamical medicine. Methods in Molecular Biology. 2014;**1175**:3-17

[2] Ullman AJ, Keogh S, Coyer F, et al. "True Blood" The Critical Care Story: An Audit of Blood Sampling Practice Across Three Adult, Paediatric and Neonatal Intensive Care Settings [Internet]. Australian Critical Care. 2015. Available from: http://www.sciencedirect.com/science/article/pii/S1036731415000752

[3] Ezzie ME, Aberegg SK, O'Brien JM. Laboratory testing in the intensive care unit. [Internet]. Critical Care Clinics. 2007;**23**:435-465

[4] Frassica JJ. Frequency of laboratory test utilization in the intensive care unit and its implications for large-scale data collection efforts. Journal of the American Medical Informatics Association. 2005;**12**:229-233

[5] Yurkovich M, Avina-Zubieta JA, Thomas J, et al. A systematic review identifies valid comorbidity indices derived from administrative health data. Journal of Clinical Epidemiology. 2015;**68**:3-14

[6] Garvin JH, Redd A, Bolton D, et al. Exploration of ICD-9-CM coding of chronic disease within the Elixhauser comorbidity measure in patients with chronic heart failure. Perspectives in Health Information Management. 2013;**10**(Fall):1b

[7] Southern DA, Quan H, Ghali WA. Comparison of the Elixhauser and Charlson/Deyo methods of comorbidity measurement in administrative data. Medical Care. 2004;**42**(4): 355-360

[8] Van Poucke S, Zhang Z, Schmitz M, Vukicevic M, Vander Laenen M, Celi LA, De Deyne C. Scalable predictive analysis in critically ill patients using a visual open data analysis platform. PLoS One. 2016;**11**(1)

[9] Ritthoff O, Klinkenberg R, Fisher S, Mierswa I, Felske S. YALE: Yet Another Learning Environment. LLWA'01–Tagungsband der GI-Workshop-Woche Lernen–Lehren–Wissen Adaptivitat. Dortmund, Germany: University of Dortmund. Technical Report 763. 2001. pp. 84-92

[10] Zhang Z. Data management by using R: Big data clinical research series. Annals of Translational Medicine. 2015;**3**(20):303. DOI: 10.3978/j.issn.2305-5839.2015.11.26

[11] Zhang Z. Missing values in big data research: Some basic skills. Annals of Translational Medicine. 2015;**3**(21):323. DOI: 10.3978/j.issn.2305-5839.2015.12.11

[12] Goldberger AL, Amaral LAN, Glass L, et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation. 2000;**101**:215-220

[13] Saeed M, Villarroel M, Reisner AT, et al. Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): A public-access intensive care unit database. Critical Care Medicine. 2011;**39**(5):952-960. DOI: 10.1097/CCM.0b013e31820a92c6

[14] McGregor C, Catley C, James A. A process mining driven framework for clinical guideline improvement in critical care. In: CEUR Workshop Proceedings; 2011

[15] Chao C-M, Yu Y-W, Cheng B-W, et al. Construction the model on the breast cancer survival analysis use support vector machine, logistic regression and decision tree. Journal of Medical Systems. 2014;**38**:106

[16] Ting H, Mai Y-T, Hsu H-C, et al. Decision tree based diagnostic system for moderate to severe obstructive sleep apnea. Journal of Medical Systems. 2014;**38**:94

[17] Quinlan JR. Induction of decision trees. Machine Learning. 1986;**1**:81-106

[18] Hastie T, Tibshirani R, Friedman J, Franklin J. The elements of statistical learning: Data mining, inference and prediction. Mathematical Intelligence. 2005;**27**:83-85

[19] Druss BG, Marcus SC, Rosenheck RA, Olfson M, Tanielian T, Pincus HA. Understanding disability in mental and general medical conditions. The American Journal of Psychiatry. 2000;**157**(9):1485-1491

[20] Post RM, Altshuler L, Leverich GS, Frye MA, Suppes T, McElroy SL, et al. Relationship of clinical course of illness variables to medical comorbidities in 900 adult outpatients with bipolar disorder. Comprehensive Psychiatry. 2015;**56**:21-28

[21] Kuncheva LI, Whitaker CJ. Measures of diversity in classifier ensembles. Machine Learning. 2003;**51**:181-207

[22] Breiman L. Random forests. Machine Learning. 2001;**45**:5-32

[23] Freund Y, Schapire R, Abe N. A short introduction to boosting. Journal of JSAI. 1999;**14**(5):771-780

[24] Breiman L. Bagging predictors. Machine Learning. 1999;**24**(2):123-140

[25] Kononenko I. Estimating attributes: Analysis and extensions of RELIEF. In: European Conference on Machine Learning. Berlin, Heidelberg: Springer; 1994. pp. 171-182

[26] Shearer C. The CRISP-DM model: The new blueprint for data mining. Journal of Data Warehousing. 2000;**5**:13-22

[27] Riley RD, Ahmed I, Debray TPA, Willis BH, Noordzij JP, Higgins JPT, Deeks J. Summarising and validating test accuracy results across multiple studies for use in clinical practice. Statistics in Medicine. 2015;**34**(13):1097-0258

[28] Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine learning; (ICML 2006). New York, NY, USA: ACM; pp. 233-240

[29] Churpek MM, Yuen TC, Winslow C, Meltzer DO, Kattan MW, Edelson DP. Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. Critical Care Medicine. 2016;**44**(2):368-374

[30] Rassa AC, Horne BD, McCubrey RO, Bair TL, Muhlestein JB, Morris DR, Anderson JL. Novel stratification of mortality risk by kidney disease stage. American Journal of Nephrology. 2015;**42**(6):443-450

[31] Ennis J, Gillen D, Rubenstein A, et al. Clinical decision support improves physician guideline adherence for laboratory monitoring of chronic kidney disease: A matched cohort study. BMC Nephrology. 2015;**16**:163. DOI: 10.1186/s12882-015-0159-5

[32] Zhang Z, Ni H. Normalized lactate load is associated with development of acute kidney injury in patients who underwent cardiopulmonary bypass surgery. In: Ricci Z, editor. PLoS One. 2015;**10**(3):e0120466

[33] Timmers TK, Verhofstad MH, Moons KG, et al. Long-term survival after surgical intensive care unit admission: Fifty percent die within 10 years. Annals of Surgery. 2011;**253**: 151-157

# Semantic Infrastructure for Service Environment Supporting Successful Aging

Vesa Salminen, Päivi Sanerma,
Seppo Niittymäki and Patrick Eklund

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.76945

**Abstract**

Demographic changes and the rapid increase of aging people are occurring throughout the world. There is a need for step-by-step developing service environment to support elderly living as old as possible at home. Digital equipment and technology solutions installed at home produce real-time data which can be used for predictive and optimized service creation. New technology solutions have to be tested at home environments to get certainty of usability, flexibility, and accessibility. The implementation of new digitalization has to happen according to ethical rules taking into account the values of elderly people. The data gathered through digital equipment is used in optimizing service processes. However, service process misses common ontology and semantic infrastructure to use the gathered data for service optimization. The service environment and semantic infrastructure, which could be used in social and health care, are introduced in this article.

**Keywords:** aging, home care, data mining, social and healthcare ontology, semantic infrastructure, ethics

## 1. Introduction

World society is facing demographic changes in aging population. Digitalization creates potentiality for service creation. The more frailty elderly are, the more services they need. Senior citizens living independently at their own homes need various kinds of services. All solution data and personal health data gathered from home environment create an opportunity for various types of services. Technology solutions have to be tested before they are accepted at home environments.

The municipal authorities, on the other hand, are closely paying attention to the data gathered and its usability for punctual service creation cost efficiency of the services. The new solutions and gadgets have to be also interesting and easy to use for senior citizens. There is lack of information in behavior research involving especially the oldest group of people.

Digitalization is rapidly increasing, and enterprises must find new ways to innovate for business advantage. Through digital transformation by using technologies, e.g., artificial intelligence, data mining, machine learning, and open data, it is possible to create new smart services or renew health pathway by lean operations. It is essential to manage the available open and gathered life cycle data as data to service process to produce value-added services to home environment for senior citizens.

Caring for elderly people need responsible business leadership and democratic innovation culture and co-innovation. This article introduces also some features of how ethical values are fulfilled when applying digitalization in caring older people at home by the methods of responsible business leadership. In order to sustain competitive advantage, health technology companies making products and solutions for the use of healthcare and well-being are expanding their product offering to service solutions over the product life cycle. Responsible leadership is understood as a social-relational and ethical phenomenon, which occurs in social processes of interaction communication.

Services are created to support the elderly and their families in maintaining a high-quality life at home. It shows that the Internet has had a positive impact on the health information acquired by seniors but not all seniors. There is lack of information behavior research involving especially the oldest group of people. The objectives of digitalization in home care environments are:

- To support and increase the autonomy of older people and independent coping at home
- Efficiency and effectiveness of nursing at health and well-being environments

Healthcare and well-being technology companies have started a project TELI in Finland to promote business opportunities in implementing digitalization into the healthcare and well-being solutions. Häme University of Applied Sciences is focusing on digitalisation of home care in this article. University has together with the municipal authorities established real-life piloting environments to enhance the adoption of the healthcare and well-being sector technology and digitalization in the way of using the data gathered in building semantic infrastructure to support service environment in creating integrated services for elderly customers.

The objective is to determine how the gathered data can be used in optimizing service processes, logistic routing, and focused use of medicine and medical equipment at home visits. However, there is missing generic data to service process, common ontology, and semantic architecture to use the gathered data for service optimization. Common ontology is essential as shared domain knowledge in services [1]. Communication and data transfer infrastructure in home care services require:

- Standardization of the formal semantics to enable data and service management on healthcare pathway

- Approaches to define real-world semantics linking social and healthcare process content with meaning for humans based on care terminology

Novel technology brings new opportunities for responsible business models. The transformation toward responsible business takes a long time, and that is why it is important to fully understand the strategic concept, identify the key issues, and harness the associated opportunities. Service environment and semantic infrastructure, which could be used in social and health care, are introduced in this article.

The concepts in this paper are valid also within European Innovation Partnership of Active and Healthy Aging (EIP AHA), where EIP AHA Task Forces focus on information and process modeling, upscaling, and business models [2].

## 2. Theoretical background

Key elements of the new service approach are the innovative solutions and revised healthcare and social welfare legislative reform proposal. The reform aims also to promote cooperation between municipal officers, local universities, and digital equipment and service system suppliers.

Services are continuously being created in Europe to support the elderly and their families in maintaining a high-quality life at home [3]. The Internet has had a positive impact on the health information acquired by seniors but not all seniors [4]. There is lack of information behavior research involving especially the oldest group of people.

Digitalization of services and information management are no longer a question of if, but of when, and to what extent it will influence a specific well-being sector. It is no longer a negative reactionary tactic to moderate home care service environment but a positive proactive strategy to accelerate long-term service environment. It is not just about data and service management, but a social and healthcare-changing paradigm integrating innovation, differentiation, and transformation.

There is a new logic behind open innovation which embraces external ideas and knowledge in conjunction with internal R&D [5]. This offers a novel way of creating value. New value for the customer is created in the form of a product or service offering, and it results in life cycle innovation [6]. It is essential to know whether there is also a transition into a new business model of the well-being service. System thinking is the art of simplifying complexity [7]. It is about seeing through chaos, managing interdependency, and understanding choice. Concepts are important to explain chaos.

It is substantive to increase awareness of Industry 4.0 outside the group of industrial key stakeholders. There is lack of understanding on the importance of the need of implementing common Internet of technology, IoT, infrastructure on all stakeholders (large and small firms, healthcare authorities, and municipals) [8]. The widespread adoption of information and communication technology (ICT) is increasingly accelerating the blurring of boundaries between the real physical world and the virtual one. The linkage is becoming increasingly smart [9]. Presently Industry 4.0 is more industrial driven, but this will change and broaden out. Company democracy model [10] can be characterized as a multidisciplinary science, as it integrates many management (strategy, leadership, etc.), engineering (process knowledge, innovation), social (human resources, ethos, etc.), financial (marketing, extroversion, etc.), and other disciplines.

## 3. Research questions and research approach

The role of digitalization and growing amount of data collected from home and older person himself as a business driver is growing and has to be carefully taken into account in well-being business transition. The opportunities of digitalization have not been understood in full context and as new service innovation. The main research questions are:

**a.** What does digitalization mean in home care of older persons and well-being service business context?

**b.** How to cope with the dramatically growing amount of data increased through digitalization? How this data can be turned as generic data to be used in optimizing service process in value network operations?

**c.** How the geriatric data and data mining are used in caring elderly people?

**d.** How the semantic infrastructure should be used supporting service environment creation in social and health care for aging people living at home?

**e.** How to cope with ethical values and loneliness when digitalization creates new opportunities for caring of older people at home?

This article introduces a concept model for utilizing digitalization as a business and innovation driver to facilitate the transition toward the new service economy and service environment on well-being business. This paper introduces preliminary results and experiences of the applied research project partially already executed.

## 4. Multidisciplinary and cooperative environment on home care

TELI-project promotes business opportunities in implementing digitalization into the healthcare and well-being solutions and services. Municipal officers have with the help of Häme University of Applied Sciences established real-life piloting environments for enhancing the adoption of the healthcare and well-being sector technology and digitalization. **Figure 1** introduces home care apartment with functions executed by digital assets, equipment, and services.

**Figure 2** illustrates the roles and partners in home care piloting environments. In the center there is a customer, who is utilizing home care services. Red arrows describe equipment and

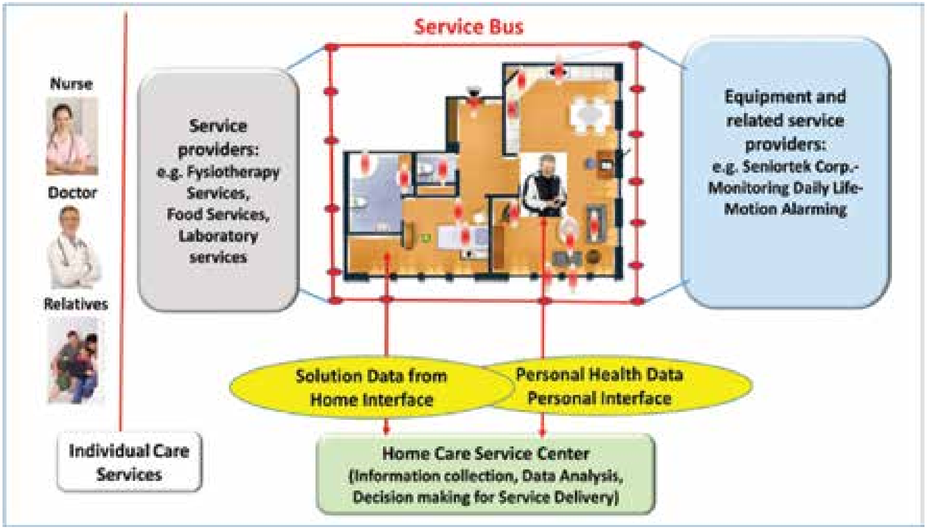

Service Bus
Nurse
Doctor
Relatives
Service providers: e.g. Fysiotherapy Services, Food Services, Laboratory services
Equipment and related service providers: e.g. Seniortek Corp.- Monitoring Daily Life- Motion Alarming
Solution Data from Home Interface
Personal Health Data Personal Interface
Individual Care Services
Home Care Service Center (Information collection, Data Analysis, Decision making for Service Delivery)

**Figure 1.** Elements of home care environment.

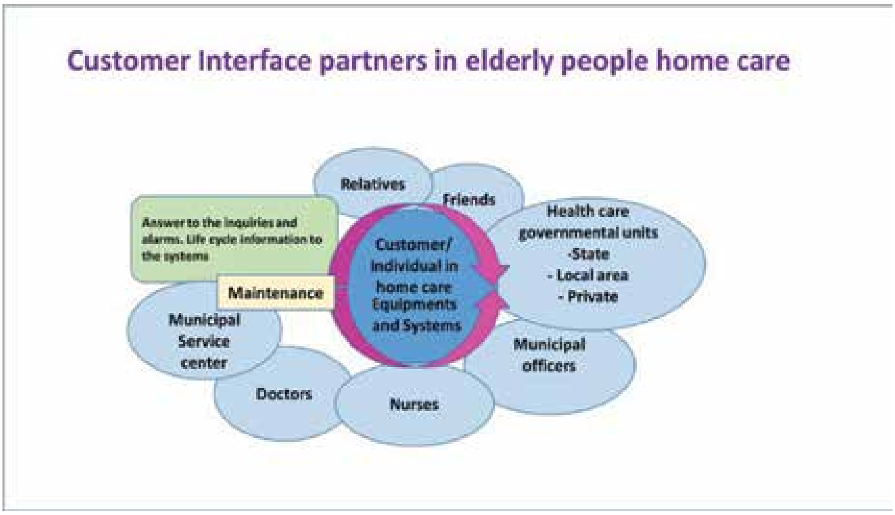

Customer Interface partners in elderly people home care
Relatives
Friends
Answer to the inquiries and alarms. Life cycle information to the systems
Customer/ Individual in home care Equipments and Systems
Health care governmental units -State - Local area - Private
Maintenance
Municipal Service center
Municipal officers
Doctors
Nurses

**Figure 2.** Conceptualized scheme of customer interface partners and systems in home care.

systems, which are involved in home care events. Equipment may be like medication dispenser, wireless smart flower stands monitoring customer daily life, like falling overs. Systems are patient databases; there is all relevant information about customer (basic information like age, gender, medication, etc.).

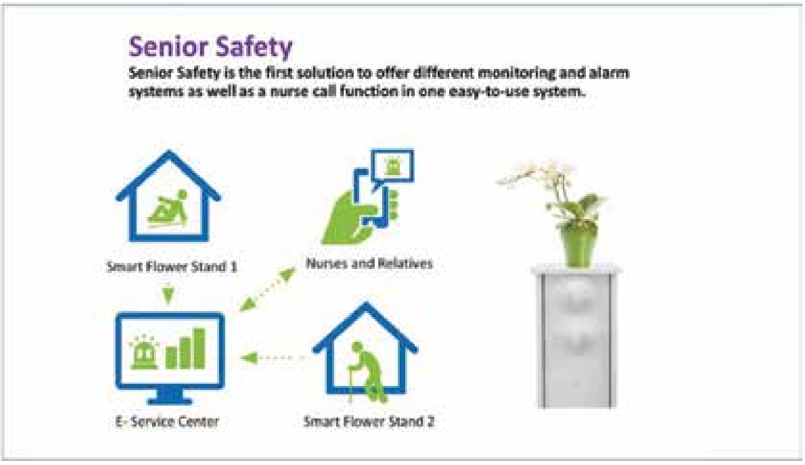
All relevant information is booked into the system, like changes in the medication, etc. Maintenance of systems and equipment should be provided. It will be absolutely important to arrange answer to inquiries and alarms, if the customer has been, e.g., fallen down or has not been taking necessary medicine. Essential events have to be booked into the system, e.g., if medication has been changed. Persons to answer the alarms may be relatives, municipal service centers, nurses, or even friends. Municipal officer or healthcare units will decide how much support will be provided by the municipality or conjoint municipalities. Medical doctors will decide changes, e.g., medications and other actions in their fields.

Fourteen health and well-being equipment, system, and service companies participated in TELI-project and were partially funding it. New health and well-being solutions including wireless motion detecting, medication dispenser, and INR- and EKG-measuring equipment were piloted in real home environment.

User requirements were gathered from elderly people, nurses, and service people of home care center in municipal areas of Forssa and Hämeenlinna cities. The gathering of data and experiences was executed by students of Häme University of Applied Sciences by the help of researchers and teachers. They had opportunity to learn to use the digital equipment and use later the knowledge in education purposes. Elderly customers did learn not to be afraid of digital equipment. Their mindset became more open toward utilization of digital gadgets in their homes to increase safety. Municipal officers at home care service centers recognized the importance of data to be used in optimized service creation and predicting diseases and dangerous situation at home environment. It is possible to streamline service processes, route logistic, and improve the use of medicine and medical equipment at home visits. It is possible to achieve also savings in home care visits.

In **Figure 3** is described what type of possibilities wireless Senior Safety- system will give in home care. In this context smart flower stand is a gateway sensor equipment with individual ID address, which is collecting data from sensor network. Smart senior safety equipment is actually a system itself, which is connected semi-automatically to healthcare system.

When older persons are leaving the apartment at night (random walking), opening the fridge, having epilepsy or other seizures, having problems with circadian rhythm, and forgetting that



Senior Safety

Senior Safety is the first solution to offer different monitoring and alarm systems as well as a nurse call function in one easy-to-use system.

**Figure 3.** Wireless senior safety.

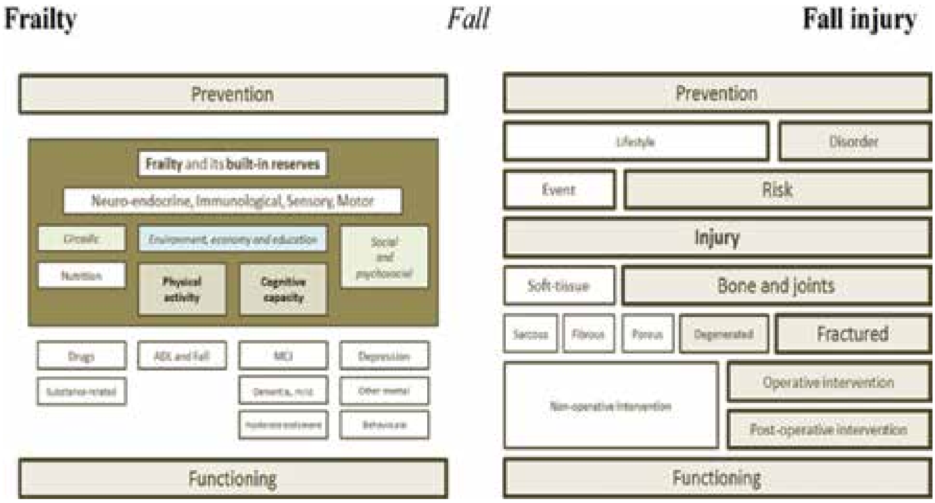


**Figure 4.** Orthopedic view of monitoring and assessment of geriatric data.

the stove is on, alarms can be controlled with mobile phone, optimized to summon the nearest nurse; functional abilities can be adjusted according to customer; and there is an external alarm option. All rooms of an apartment can be provided with motion sensors for different levels.

This then relates to monitoring and assessment of geriatric data [11] and can also be further specified, e.g., for fall risk assessment and assessment related to frailty (**Figure 4**).

The development of business environments is understood to be the responsibility of public sector and government. Public sector is however multilayered (e.g., legislative, national, provincial, regional, municipal, areal). There are many committees and operations, which have the duty to develop business environment.

The digitalization changes functions and operational processes of well-being and home care. Deployment of new functions and operational processes often needs new type of legislation which creates rules for the co-innovation and operations generated and new business opportunities (government). It is seen rather as enabler than restrictor. Industry 4.0 IoT and platform architecture are becoming the standard approach for all domain areas [12]. Planning of social and healthcare areas influences remarkably on settling and placing of enterprises and prerequisites for operation (e.g., nursing and logistics). Health care 4.0 (Industrial Internet, Industry 4.0) enables functional optimization of entire value network. Collected data from the whole value network can be used for purposes of functional development or predicting diseases. New entrepreneurship and new digital services can be created through digitalization activities.

## 5. Continuous development in social and health care

The cooperation between the government, enterprises, and universities is essential to succeed in coevolution when building up cumulative competence in creation of solutions for home care

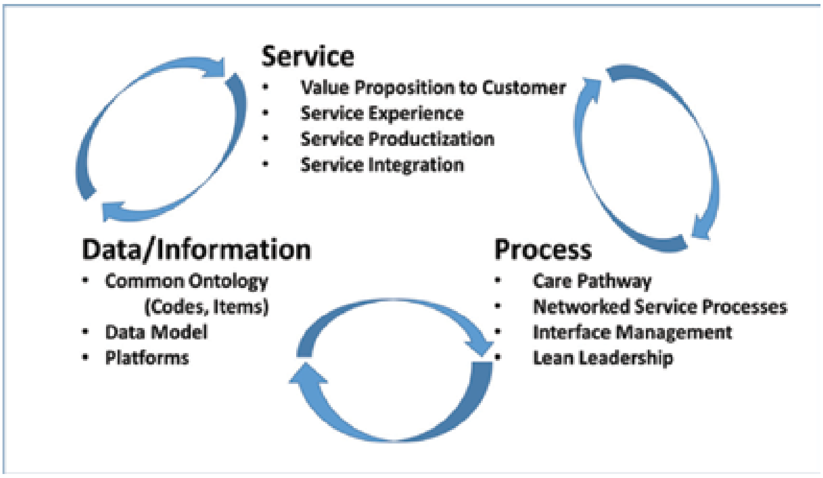


**Figure 5.** System model of continuous development in social and health care.

and older people. It is also essential to have a common vision to direct the local operation and funding. Otherwise the activities can splinter into small pieces and do not form parts of the whole vision.

Continuous development of service environment is highly dependent on linked process (care pathway), service, and data management development (**Figure 5**).

The digitalization changes functions and operational processes of well-being and home care. Deployment of new functions and operational processes often needs new type of legislation which creates rules for the co-innovation and operations generated and new business opportunities (government). It is seen rather as enabler than restrictor. Planning of social and healthcare areas influences remarkably on settling and placing of enterprises and prerequisites for operation (e.g., nursing and logistics).

On legislation, national legislation in the Nordic countries also develops in direction of regulating responsibilities related to injury prevention. In Finland, in the current law of the elderly people's social and health services (L:28.12.2012/980), municipalities have to promote the health and well-being of elderly people, also specifically as related to assuring safety (§14) and assessment of functioning with requirement of care services and levels (§15).

## 6. From data to service-process supporting business coevolution

The amount of scattered and structured data around us is increasing dramatically. It is a great opportunity to exploit that data for business purposes. Well-being and home care consist of huge amount of data, for example, the lifetime data from well-being of older persons and

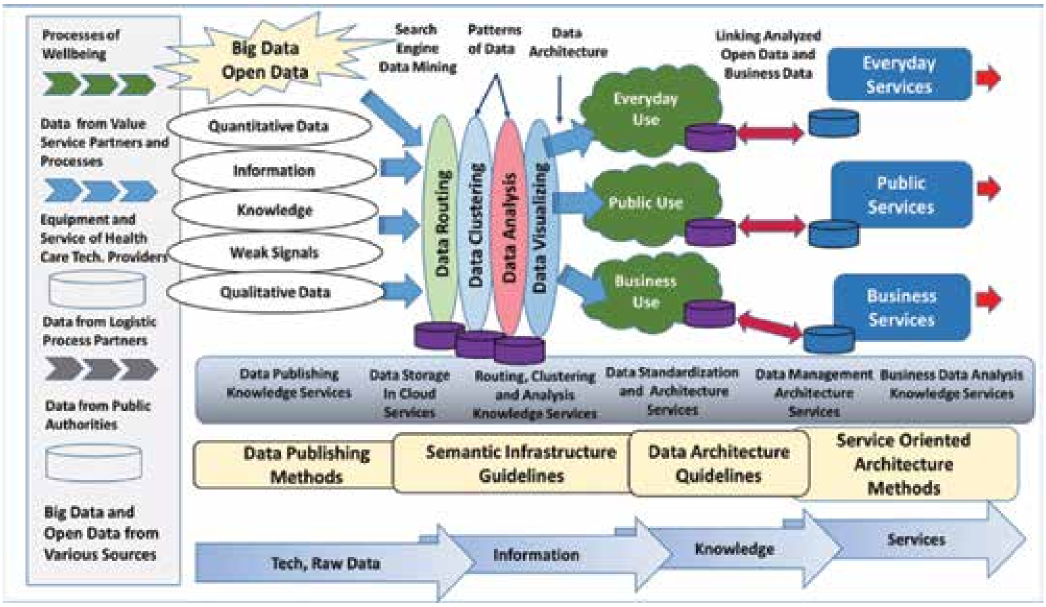

Processes of Wellbeing
Data from Value Service Partners and Processes
Equipment and Service of Health Care Tech. Providers
Data from Logistic Process Partners
Data from Public Authorities
Big Data and Open Data from Various Sources
Big Data Open Data
Quantitative Data
Information
Knowledge
Weak Signals
Qualitative Data
Search Engine Data Mining
Patterns of Data
Data Architecture
Data Routing
Data Clustering
Data Analysis
Data Visualizing
Everyday Use
Public Use
Business Use
Linking Analyzed Open Data and Business Data
Everyday Services
Public Services
Business Services
Data Publishing Knowledge Services
Data Storage In Cloud Services
Routing, Clustering and Analysis Knowledge Services
Data Standardization and Architecture Services
Data Management Architecture Services
Business Data Analysis Knowledge Services
Data Publishing Methods
Semantic Infrastructure Guidelines
Data Architecture Quidelines
Service Oriented Architecture Methods
Tech, Raw Data
Information
Knowledge
Services

**Figure 6.** From data to service process in business coevolution of well-being and home care.

home where he is living. Understanding the value proposition in growing value networks is essential. Management and analysis of data coming from various sources are routed through data-to-service process in business coevolution of well-being and home care (**Figure 6**).

Decision-making of optimizing functionality and creation of new services is cooperative activity by several stakeholders in home service environment. It is essential to gather data from various data sources, different processes, and different systems. Automation system or sensor network (IoT) at home is creating data, which is gathered, clustered, analyzed, and compared with the data gathered earlier. After that decisions are made on what nursing operations, which medication, which type of rehabilitation, and what type of logistics are needed. It is essential to build data mining operations on various parts of the process. It is also important to have all types of experts in virtual network optimizing material, medication, logistic, and nursing processes to support this value network process.

Knowledge and capability of various stakeholders in home care environment are activated in order to manage economical, ethical, and technical risks in service creation. The created service should be evaluated as a value for elderly persons, nursing staff, and network service partners.

## 7. Exploiting digitalization and big data in service coevolution

Integrated services and technologies are connected on the same platform, which is a new innovation in domiciliary home care and at care homes. The definition of interfaces is nowadays missing, and combining of data gathered and life cycle data is not possible, or it is very difficult. The common interface layers are strategy (e.g., contract), process, information, data ownership, and security and communication. It is essential to understand all the stakeholders in home care environment and define functionally interface structure between partners (**Figure 7**).

New Internet of thing, IoT, and other distant service solutions are developed. They support combining data gathered from home digital systems (solution data from home environment

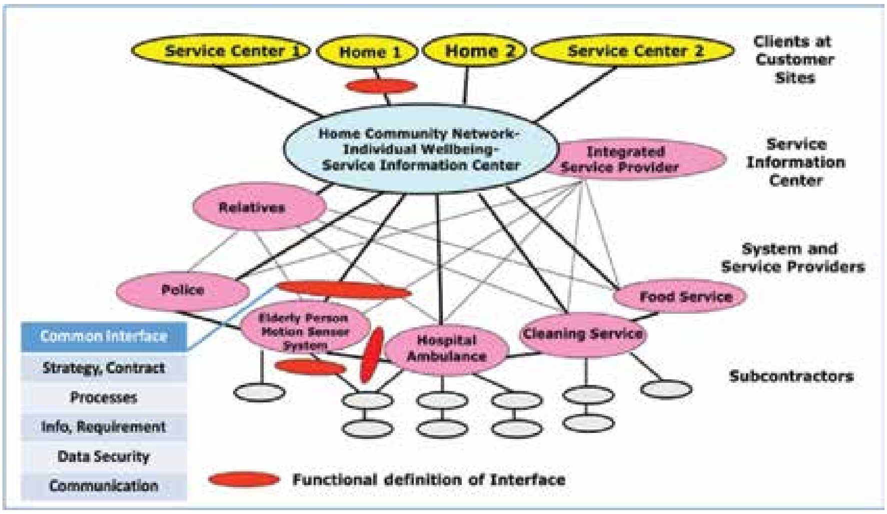

Service Center 1
Home 1
Home 2
Service Center 2
Clients at Customer Sites
Home Community Network-Individual Wellbeing-Service Information Center
Integrated Service Provider
Service Information Center
Relatives
System and Service Providers
Police
Food Service
Common Interface
Elderly Person Motion Sensor System
Hospital Ambulance
Cleaning Service
Subcontractors
Strategy, Contract
Processes
Info, Requirement
Data Security
Communication
Functional definition of Interface

**Figure 7.** Functional definition of interface in networked environment.

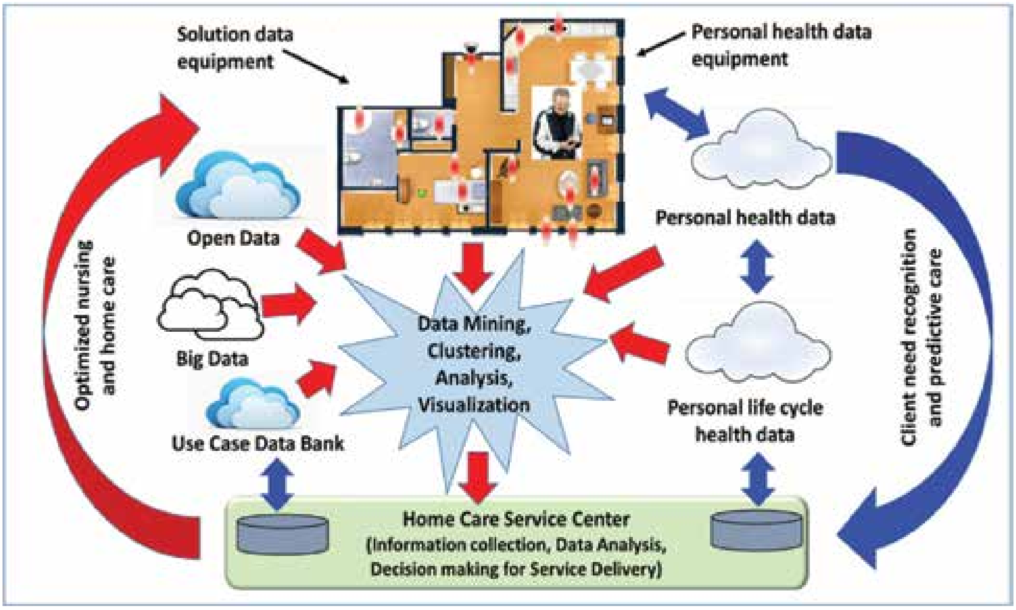

Solution data equipment
Personal health data equipment
Optimized nursing and home care
Open Data
Personal health data
Big Data
Data Mining, Clustering, Analysis, Visualization
Use Case Data Bank
Personal life cycle health data
Client need recognition and predictive care
Home Care Service Center
(Information collection, Data Analysis, Decision making for Service Delivery)

**Figure 8.** Predictive care planning according data mining and analysis.

and individual own well-being data) and information from open data in real time to home care service center, doctors, nurses, dental nurses, and other experts. This information is recorded in patient information systems so that it is available to all persons participating in care process. Digitalization brings new information on health and well-being by following ways:

- Through physiological information recognition, person's activity level can be maintained at a good level, thus making it possible to contribute to living at home and to slow the progression of diseases such as Alzheimer's disease.
- The well-being process allows information management and home care and nursing homes; the collected data (personal data and big data) is processed on the basis of analysis to be used in decision-making. Digitalization can be used to create a new care culture [11], which is based on documented and analyzed data (**Figure 8**).

Digitalization in home care and well-being sector are rather new topics, and there are few experiences in exploiting digitalization in municipal service functions and service providers. It is important to commit stakeholders in care pathway on cooperative co-innovation and demonstrations, when creating new services in home care environment. Most of the innovations will be created at customer interface.

## 8. From ontology formulation to semantic infrastructure

When a request involved in any format should answer the end-to-end performance, the Semantic definition needs to be clear from the request structure [15]. There will be four structural classes, such as health, social, process, and controls. A database links automatically. Social and healthcare

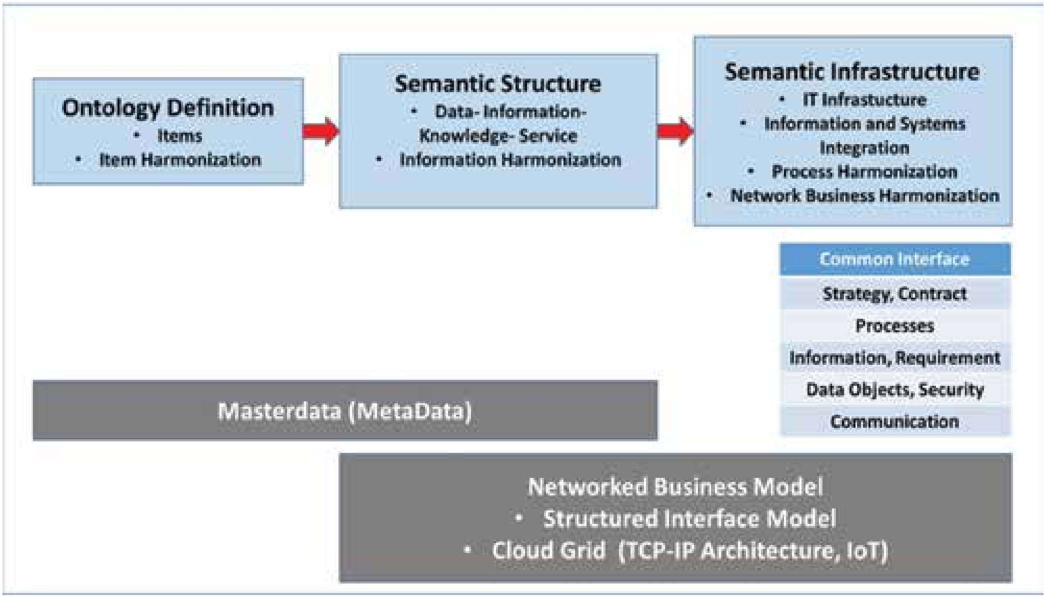


**Figure 9.** Semantic infrastructure creation for home care environment.

ontology definition is starting point in process harmonization. Ontology with item harmonization is the backbone for semantic structure and information harmonization. The functionality of semantic infrastructure has been built upon semantic structure and master data definition (**Figure 9**). Process definition, care pathway, is guiding process harmonization in semantic infrastructure. That is also the key for networked business harmonization. Service environment in elderly care is highly networked.

## 9. Ethical values and digitalization

Older people like to live meaningful life at their own home. Clients of home care have self-determination about their life and care activities. Living at home has to be the person's own choice. Very often elderly are frail and weak persons who need a lot of help. Generally they have many diseases and disabilities, and caring of them is often challenging and needs complex arrangements and careful planning. Elderly want the right to decide whether or not to use the digital services. It is important to develop solutions for supporting the safety and quality of life of elderly. All services provided to the elderly must be based on client's personal needs with respect of their own autonomy. Home care professionals can recognize clients' individual resources. Caring and all services have to support elderly's own performance and individual resources [13]. Unfortunately home care of older people is based on an illness-centered approach that focuses only on their physical resources, and development of home care services is urgent [14]. Digital services and applications can help us to develop elderly care for better direction.

Older people encounter difficulties in many activities of daily life. The market has a lot of different technological solutions for elderly. Older people need versatile services for living at home. Elderly can get benefit from technological applications and solutions that we already have in society. This would be fair for the elderly. The technology for elderly home care and all aging-related applications and solutions are a growing market in the business field. Technology professionals are responsible for developing digital solutions for the elderly. It is important to develop technological solutions from the perspective of the elderly in the first place [15].

In nursing practice we also have effective technology for nurses. Nurses are often in hurry, so applications can help them in decision-making [16] and documentation [13] and in clinical work [17]. We take care of the clients as possible and with the help of updated information. With digital solutions it is possible to develop safer caring pathway in social and health care for elderly.

Demographical change is a large political, social, and economic question, and it is a global phenomenon. In the political point of view, many countries have to make changes and develop organization of services. Elderly need more services, and this is also an economical question. Elderly services have to be flexible and cost-effective [18]. Belgium, France, Italy, Portugal, Spain, and the United Kingdom have an organizational model in which health and social services are separate. In other countries, especially Denmark, Finland, and Sweden, policy-makers recognized the advantages of providing home care as one integrated organization under the responsibility of municipalities [19, 20].

It is most important that person's service package is an integrated and appropriate wholeness. One of the key challenges of home technologies is the need to control and integrate many separate systems, as technologies are often purchased one by one [21, 19, 15]. The services for the elderly must have consistent quality and be fair for all the elderly [22]. Digital services can complement and defragment the customer care and thus improve the quality of care. The longer the elderly are able to live in their own homes, the cheaper it is for the society [15].

Among results of earlier research [14], home care services need to provide a service that meets clients' specific needs including psychological and social resources. Older persons try to find strategies to cope with the changes and difficulties that arise in relation to aging. Technological solutions can be used in activities of daily living to compensate lack of elderly's performance. The technological solutions promote elderly's safety and the feeling of safety also to client's family members and relatives.

Home care professionals need more education for exploiting digital applications in client's care at home. They need more information on what kind of applications is available for care. When the treatment is planned, the possibilities of digital services should be taken into account. Client orientation is most important in developing digital services for the elderly. Elderly as a social and heterogeneous group with diverse interest, variable education, health, and socio-economic level has to be taken in account, when creating and producing services.

It is a well-known fact that loneliness is a significant problem for elderly. Technology is not the answer to everything [15]. It cannot replace human relationships, but suitable technological solutions can support elderly's social life. The ability to live safely at home should be considered carefully respecting older people's own choices. The focus is on technological solutions that can extend the time the elderly continue to live in their own homes independently and safely. There is no universal technological solution that suits everyone; hence they should be applied by configured way. Technological solutions can give elderly opportunities to participate and socialize instead of isolating at home.

## 10. Well-being and home care: conceptual model for adaptive development

Digitalization in home care and well-being sector is a rather new phenomenon. There are still few experiences in exploiting digitalization in municipal service functions. The best and widely accepted innovations and services are created at customer interface. Research and learning environments at municipal customer environment have been used in succeeding business coevolution and continuous innovation. Cooperation requires engagement of municipal authorities and nurses. It requires trust on information and experience delivering. It is ought to be continuous on various organization levels. Cooperation and learning together on research and learning environment supplied in TELI-project case by university are basis for new innovations and continuous development. Development of superior competitive power through principals of well-being and home care is built by lean and digitalized value networks. It is important to succeed in exploiting multidisciplinary competence and open information sharing.

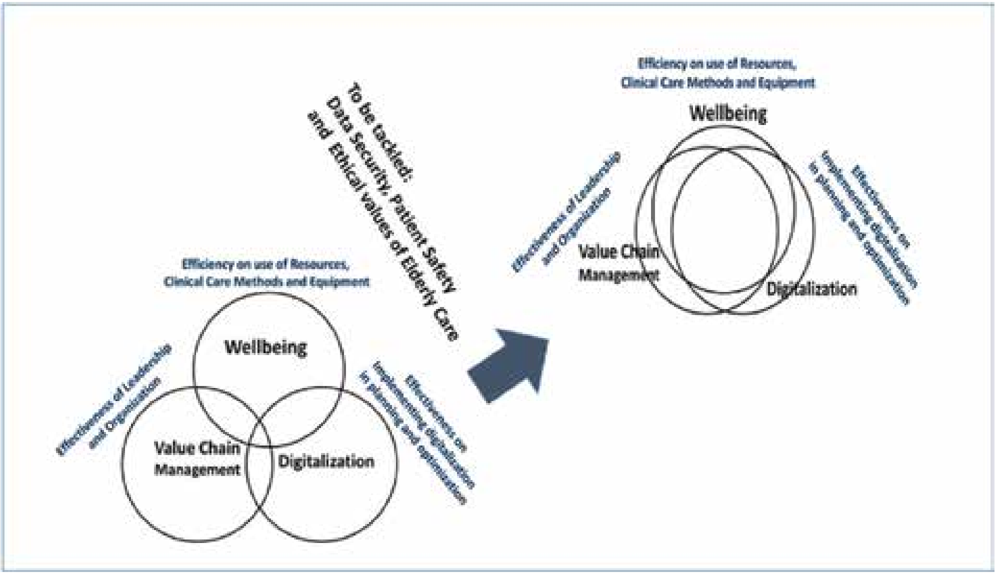

To be tackled:
Data Security, Patient Safety
and Ethical values of Elderly Care
Efficiency on use of Resources,
Clinical Care Methods and Equipment
Wellbeing
Effectiveness of Leadership
and Organization
Value Chain
Management
Digitalization
Effectiveness on
implementing digitalization
in planning and optimization
Efficiency on use of Resources,
Clinical Care Methods and Equipment
Wellbeing
Effectiveness of Leadership
and Organization
Value Chain
Management
Digitalization
Effectiveness on
implementing digitalization
in planning and optimization

**Figure 10.** Proper implementation leading on more intensive overlapping.

Sustainable growth and responsible business management are not possible to achieve by linear way but adaptive way. **Figure 10** describes the conceptual model of adaptive development toward successful well-being operations and home care services. Value chain management, proper digitalization, and data management are the key developing features. The objective of well-being economy is resource efficiency and maintaining ethical values on all operations. Continuous digitalization increases effectiveness on optimized services. Value partners need to streamline effectiveness of leadership and interface processes. The trend on well-being service innovation in home care environment is that ethical care pathway, digitalization, and value network development are increasingly overlapping. The increasing digitalization and management on data-to-service process are key enablers in business coevolution.

Succeeding co-innovation on well-being and home care requires data mining practices, data-to-service management process, and creation of adaptive multidisciplinary cooperation model for common semantic infrastructure and solution development. The experts making applied research with customers have to have content on individual nursing of older persons and home care process knowledge at customer site; they have to be capable to work in teams on distributed way with other experts in value network and have to have certain collaborative skills to work together.

## 11. Ethical considerations

The study followed good scientific practice and guidelines [23, 24, 25]. The clients and all professionals volunteered to participate to the research. Ethical questions related to the research were evaluated. Target organizations gave research permissions and decisions based on ethical evaluations. All students have signed a separate confidentiality agreement.

The clients were informed of the research with a written notice. They were asked for permission to the research by home care staff. The informants gave their consent to interviews both in writing and orally. The anonymity of the participants was secured throughout the research process. Participants were able to withdraw from the study whenever they wanted to do that.

## 12. Discussion and conclusions

Combining the principles of home care service to value network thinking and digitalization with data mining practices gives opportunity for remarkable competitive performance on whole the well-being environment. Recognition of older persons creating actual customer needs combined with life cycle calculation creates opportunities for life cycle services on home care environment.

Experiences on executed TELI-project show that it is essential to engage municipal authorities and public sector on conceptual development work when creating services for home care environment. It is also relevant to develop acceptable legislation, which enables the use of new digital equipment and delivering of created new services.

Universities can support municipal officers and technology application providers to provide and maintain research and learning environments for continuous piloting of new technologies and preparation of new business models on home care environment. Häme University of Applied Sciences supported demonstration of digitalization of versatile home care environment in Finland at Forssa municipal facility and Hämeenlinna Home Care division.

The data gathered through digital equipment can be used in optimizing service processes. It is important to have generic data as common ontology of service process and semantic architecture to route the gathered data. In this article is introduced, what type of ontology-based sematic architecture could be used in social and health care and how geriatric data should be integrated in caring elderly people. This article introduces experiences on co-innovation of home care services cooperatively together with public and private organizations.

This article introduces experiences on responsible business leadership. Older people like to live meaningful life at their own home. Clients of home care have self-determination about their life and care activities. Living at home has to be the person's own choice. Elderly want the right to decide whether or not to use the digital services.

The study followed good scientific practice and guidelines. The clients and all professionals volunteered to participate to the research. Ethical questions related to the research were evaluated. Target organizations gave research permissions and decisions based on ethical evaluations. All students have signed a separate confidentiality agreement.

## Author details

Vesa Salminen[1]*, Päivi Sanerma[1], Seppo Niittymäki[1] and Patrick Eklund[2]

*Address all correspondence to: vesa.salminen@hamk.fi

1 Häme University of Applied Sciences, Hämeenlinna, Finland

2 Umeå University, Umeå, Sweden

## References

[1] Chandrasekharan B, Josephson JR, Benjamins VR. What are ontologies, and why do we need them? IEEE Intelligence Systems. 1999;**14**(1):20-26

[2] Bousquet J, Bewick M, Cano A, Eklund P, et al. Building bridges for innovation in ageing: Synergies between action groups of the EIP on AHA. The Journal of Nutrition, Health & Aging. 2016:20

[3] Gonzales A, Ramirez MP, Viadel V. Attitudes of the elderly towards information and communication technologies. Educational Gerontology. 2012;**38**(9):585-594

[4] Hallows K. Health information literacy and the elderly: Has the internet had an impact? The Serials Librarian. 2013;**65**:39-55

[5] Chesbrough H. Open Innovation: The New Imperative for Creating and Profiting from Technology. Boston: Harvard Business School Publishing Corporation; 2003

[6] Salminen V. Management of Life Cycle Business Transition by Hybrid Innovation. Managing Innovation in Connected World, ISPIM 08, 14–17.12.2008, Singapore; 2008

[7] Jamshid G. System Thinking: Managing Chaos and Complexity. A Platform for Designing Business Architecture. MA, USA: Butterworth- Heinemann; 1999

[8] European Parliament: Briefing: 4.0 Industry digitalization for productivity and growth http://www.europarl.europa.eu/thinktank. September 2015. (Internet)

[9] Deloitte: Industry 4.0 challenges: Challenges and Solutions for the Digital Transformation and use of Exponential Technologies; 2015

[10] Markopoulos E, Vanharanta H. Human Perception, Interpretation, Understanding and Communication of Company Democracy. 14th International and interdisciplinary Conference of the Research Cooperation, Turku, Finland; 2014

[11] Salminen V, Sanerma P, Niittymäki S, Eklund P. Ontology based service environment supports successful aging. In: Kantola JI et al., editors. Advances in Human Factors, Business Management and Leadership, Advances in Intelligent Systems and Computing. Springer International Publishing AG; 2018. p. 594

[12] Salminen V, Pillai B. Interoperability Requirement Challenges- Future Trend. International Symposium on Collaborative Technologies and Systems, CTS 2007. Orland, USA, May 21–25, 2007

[13] Turjamaa R, Hartikainen S, Kangasniemi M, Pietilä A-M. Is it time for comprehensive approach in older home care client's care planning in Finland? Scandinavian Journal of Caring Sciences. 2014;**29**:317-324

[14] Turjamaa R, Hartikainen S, Pietilä A-M. Forgotten resources of older home care clients: Focus group study in Finland. Nursing and Health Sciences. 2013;**15**:333-339

[15] Jännes J, Hämäläinen P, Hanski J, Lanne M. Homelike living for elderly people: A needs-based selection of technological solutions. Home Health Care Management & Practice. 2015;**27**(2):64-72

[16] Johansson P, Petersson G, Nilsson G. Personal digital assistant with a barcode reader—A medical decision support system for nurses in home care. International Journal of Medical Informatics. 2010;**7**(9):232-242

[17] Lagerin A, Carsson A, Nilsson G, Westman J, Törnvist L. District nurses'preventive home visits to 75-year-olds: Opportunity to identify factors related to unsafe medication management. Scandinvian Journal of Public Health. 2014;**42**:786-794

[18] Taylor C, Donoghue J. Innovation and translation. New ways to provide community aged care services. Australian Journal on Aging. 2015;**34**(3):199-200

[19] Tarricone R, Tsouros A, editors. Home care in Europe. World Health Organisation; 2008

[20] Tepponen M. Home Care Integration and Quality. Kuopio University Publication, E. Society Sciences; 2009. p. 171

[21] Geron S, Smith K, Tennstedt K, Jette A, Chassler D, Kasten L. The home care satisfaction measure: A client-centered approach to assessing the satisfaction of frail older adults with home care services. Journal of Gerontology: Social Sciences. 2000;**55B**(5):S259-S270

[22] Pajalic O, Pajalic Z. An evaluation by elderly people living at home of the prepared meals distributed by their municipality – A study with focus on the Swedish context. Global Journal of Health Science. 2015;**7**(3):57-68

[23] Bujnowska-Fedak M. Support for e-health services among elderly primary care patients. Telemedicine and Health. 2014;**2014**:696-704

[24] Research Ethical Councel 2012. Good Scientific Practices. www.tenk.fi

[25] Bashshur R, Shannon G, Smith B. The empirical foundations of telemedicine interventions for chronic disease management. Telemedicine and E-health. 2014;**20**(9):769-800

Chapter 8

# Adaptive Neural Network Classifier-Based Analysis of Big Data in Health Care

Manaswini Pradhan

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.77225

**Abstract**

Because of the massive volume, variety, and continuous updating of medical data, the efficient processing of medical data and the real-time response of the treatment recommendation has become an important issue. Fortunately, parallel computing and cloud computing provide powerful capabilities to cope with large-scale data. Therefore, in this paper, a FCM based Map-Reduce programming model is proposed for the parallel computing using AANN approach. The FCM based Map-Reduce, clusters the large medical datasets into smaller groups of certain similarity and assigns each data cluster to one Mapper, where the training of neural networks are done by the optimal selection of the interconnection weights by Whale Optimization Algorithm (WOA). Finally, the Reducer reduces all the AANN classifiers obtained from the Mappers for identifying the normal and abnormal classes of the newer medical records promptly and accurately. The proposed methodology is implemented in the working platform of JAVA using CloudSim simulator.

**Keywords:** fuzzy C-means clustering (FCM), adaptive artificial neural networks (AANN), whale optimization algorithm (WOA), parallel computing, Map-Reduce model

## 1. Introduction

Big Data has been characterized by it three properties i.e.1.Volume, 2.Velocity and 3.Variety. Volume refers to the huge amount of data being generated by several sources. Velocity refers to the rate at which this data is being generated and the variety refers to the different type of the data being used [1]. Now a days with so much of data all around the world, the trend in healthcare is shifting from cure to prevention. Hospitals and healthcare systems are good repositories of big data (like patient records, test reports, medical images etc.) that can be utilized to cut the cost in healthcare, to improve reliability and efficiency, and to provide better cure to patients.

Healthcare applications require large amounts of computational and communication resources, and involve dynamic access to large amounts of data within and outside the health organization leading to the need for networked healthcare [2]. Data Analysis has always in demand in all the industry as it gives the approximate prediction of how the market is growing [3]. Although the innovations are in the healthcare field, there are some issues that need to be solved, particularly the heterogeneous data fusion and the open platform for data access and analysis [4].

Today, the healthcare industry is turning to big data technology to improve and manage medial systems. For this purpose, healthcare companies and organizations are leveraging big data in health informatics [5]. The analysis of big data carried out through different ways. Machine learning algorithm helps in analysis of big data very efficiently [3].

Feature selection is an important preprocessing technique used before data mining so that it can reduce the computational complexity of the learning algorithm and remove irrelevant/redundant features to remove noise [16]. Decision Tree is a predictive model of classification, which can be viewed as a Tree like structure [6]. It is simple and gives a fast and accurate result.

Neural Network is one of the other machine learning algorithms which showed a lot of modification. Neural Network is an adaptive learning model which adjusts the weight of the connecting links between its neuron [15]. K-Nearest Neighbor model of classification is one of the simplest classification algorithm which work on the classifying the data set based on the nearest neighbor of the existing class label of already trained mode [7]. Naïve Bayesian Classifier has a very good accuracy in classification for large set of data [8].

Clustering algorithm makes the groups or clusters of homogenous data. It is an unsupervised learning technique. In Partitioned Clustering the number of cluster was defined beforehand. In Hierarchical Clustering we do not need to define the number of clusters in advance [9, 10]. In both of the above approaches the stopping criterion is usually the number of clusters to be achieved; once the required number is achieved the algorithm can be stopped. Different methods are used for the analysis of Big Data in Health Care has been discussed below.

## 2. Literature survey

Abdulsalam Yassine et al. [11] have proposed a model that utilizes smart home big data as a means of learning and discovering human activity patterns for health care applications. They proposed the use of frequent pattern mining, cluster analysis and prediction to measure and analyze energy usage changes sparked by occupants' behavior.

Md. Mofijul Islam et al. [12] have proposed a mobility- and resource aware joint virtual-machine migration model for heterogeneous mobile cloud computing systems to improve the performance of mobile Smart health care applications in a Smart City environment.

Mohammad-Parsa Hosseini et al. [13] focused on an autonomic edge computing framework for processing of big data as part of a decision support system for surgical candidacy, an optimized model for estimation of the epileptogenic network, and an unsupervised feature extraction model.

Bernhard Schölkopf et al. [14] have designed a class of support vector algorithms for regression and classification.

Chandra et al. [15] have proposed a approach for using MLP to handle Big data. There was high computational cost and time involved in using MLP for classification of Big data having large number of features. This is a promising technique for handling big data and is the idea extracted for the present research work.

Huan Liu et al. [16] have introduced a concepts and algorithms of feature selection, surveys existing feature selection algorithms for classification and clustering, groups and compares different algorithms with a categorizing framework based on search strategies, evaluation criteria, and data mining tasks.

Malika Bendechache et al. [17] have proposed a distributed clustering approach to deal efficiently with both phases; generation of local results and generation of global models by aggregation.

## 3. Proposed FCM Map-Reduce based adaptive neural network classifier for handling big data in health care

The large amounts of data, driven by record keeping, compliance & regulatory requirements, and patient care will historically render for the healthcare industry. While most data is saved in hard copy form, the current trend is towards quick digitization of these large amounts of data. Driven by mandatory requirements and the potential to develop the quality of healthcare delivery meanwhile minimizing the costs, these massive quantities of data known as 'big data' hold the promise of supporting a wide range of medical and healthcare functions, admitting between others clinical decision support, disease surveillance, and population health management. Some troubles that exist in big data analysis in health care are, i) to succeed, big data analytics in healthcare requires to be packaged so it is menu driven, user-friendly and transparent. ii) The lag among data collection and processing has to be addressed. iii) The crucial managerial issues of ownership, governance and standards have to be conceived. iv) Continuous data acquisition and data cleansing is another issue.

In the increasingly quick generation of large amounts of data, and across several areas of science, technological and conceptual advances are resulting. The collection and organization of data, the volume, variety, and velocity of current 'big data' production inaugurates novel opportunities and challenges in both scale and complexity these are always admitted on research. Also, in health care sector, the dealing of big data has currently get an interesting research topic, as since there are wide amount of medical data's available in cloud storage.

Moreover, the huge number of data records within very large datasets that comprise an extremely high amount of information is conceived to be a very critical issue. Thus processing with sequential algorithm results in greater computational cost in terms of memory space and time complexities. Hence, for discovering the above mentioned issues, a parallel architecture is required to be demonstrated.

In order to minimize the computational complexity and the memory requirement while leading large healthcare data, it is suggested to have a parallel adaptive artificial neural network (AANN) technique applying Map-Reduce programming model for health care analysis from big data in cloud environment. The introduction of abnormality in the medical data records applying the proposed Map-Reduce based Adaptive Artificial Neural Network classification method by the trained data these are determined by suggested approach. The medical data from the cloud has to be first clustered in order to distinguish the similar classes of data associated to any one particular health disorder for better classification of data. Here, the clustering of similar sets of data is done with the help of Fuzzy C means clustering algorithm, so as to develop the classification performance. The dataset separated as sub clusters were afforded to Map-Reducer framework, where AANN is implemented in parallel. Once the clustering is done the normal and abnormal classes of medical data are then learned applying the proposed map decrease programming model based AANN. By training the AANN models, it can be capable to predict for newer data as well.

The map minimized programming model comprises of two phases: (1) Mapper phase and (2) Reducer phase. Data belonging to each cluster are mapped applying separate mappers. Each mapper based AANN receives one training item (i.e. any one data cluster) and then calculates the weights of the network applying the training item in the suggested parallel prediction model. Here, to develop the precision of classification of the data, the proposed AANN method applies the concept of optimization, where the weight factors are maximized by applying Whale Optimization Algorithm. The Reducer separates the test medical record in order to distinguish the health condition established on the mapped data.

Here, the schematic diagram of the proposed healthcare application model for the analysis of large datasets is presented in **Figure 1**.

The proposed FCM based Map-Reduce AANN approach comprises of the following phases, 1) Fuzzy C-means (FCM) based Data Grouping 2) Mapper phase involving assigning each data groups to separate Mappers and training Data using Adaptive ANN 3) Reducer Phase consisting of Testing Phase. Each of the steps is detailed in the following sections.

### 3.1. Phase 1: Fuzzy C-means (FCM) based data grouping

Established on the membership function, Fuzzy C-means (FCM) is a data clustering technique in which each and every data in that group will comes under one cluster. It will group all the data in to particular number of clusters in high dimensional search space. The degrees of the cluster are determined by the membership function in terms of [0, 1] which affords the flexibility that the data point can belong to more than one cluster.

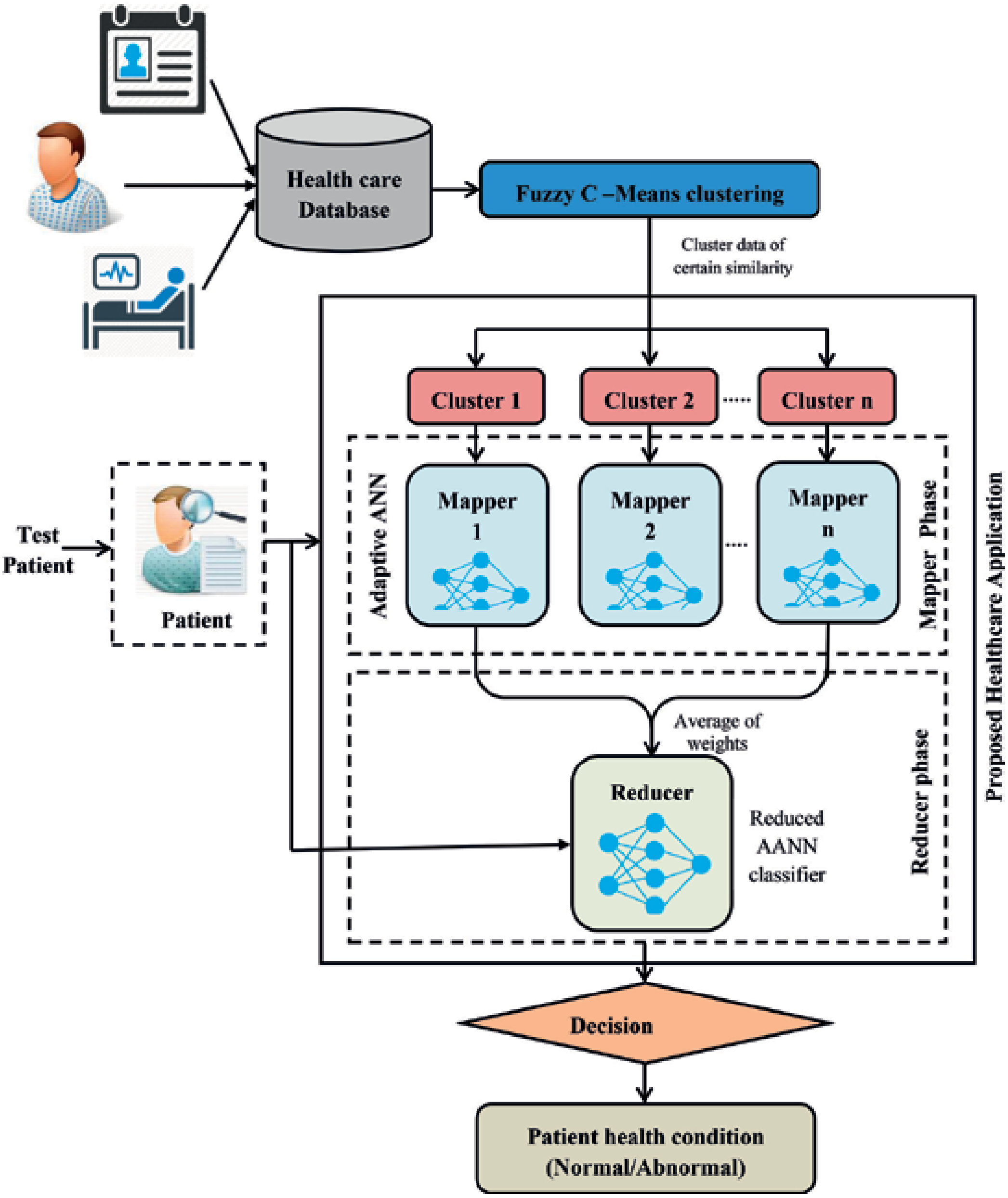


**Figure 1.** Schematic diagram.

The proposed method applies FCM for clustering the input large data into smaller groups of similar data. The input data will be grouped into number of clusters randomly and the centroids will be rendered for the clusters during Fuzzy c-means clustering. The clusters are updated established on the membership grade of the data points and the novel centroid is depicted correspondingly at the each iteration.

Moreover, how the clustering with fuzzy c means algorithm is made for a set of input samples is afforded below.

Let us considering the input sample be,

$$D_u \, (u = 1, 2, \ldots, n) \tag{1}$$

The input sample is to be separated into *'v'* number of clusters. The clustering cannot be exactingly but it will be made by means of the grouping with respective to the grade of membership function.

The objective function of FCM algorithm is effectively explained as follows.

$$Objec\,Fun = \sum_{u=1}^{n} \sum_{v=1}^{C} M_{uv} \|D_u - c_v\|^2 \tag{2}$$

where,

*"$M_{uv}$"* is the membership of $u^{th}$ data ($D_i$) in the $v^{th}$ cluster $c_v$.

*"$c_v$"* is the $v^{th}$ cluster center.

*"$D_u$"* is the $u^{th}$ data record.

*"n"* is the total number of data record.

*"C"* is the required number of clusters.

*"$\|D_u - c_v\|^2$"* is the similarity between $u^{th}$ data record and the center vector of $v^{th}$cluster.

Now the cluster center calculation is done by Eq. (3),

$$c_v = \frac{\sum\limits_{u=1}^{n} M_{uv} D_u}{\sum\limits_{u=1}^{n} M_{uv}} \tag{3}$$

Membership updation is done by Eq. (4),

$$M_{uv} = \frac{1}{\sum\limits_{y=1}^{C} \left( \frac{\|D_u - c_v\|}{\|D_u - c_y\|} \right)^{\frac{2}{b-1}}} \tag{4}$$

where, *'b'* is the fuzziness coefficient. The membership matrix $M^{(x)} = [M_{uv}]$ is calculated for among every iteration. If $\|M^{(x)} - M^{(x-1)}\| < T$ then stop, Where, *"x"* denotes the current iteration and *"T"* is the threshold of termination criterion, which is among 0 and 1.

The input data is clustered into data groups of certain similarity for established on the above procedure of FCM. We found the number of cluster set such as $C_1, C_2, C_3, \ldots, C_n$ at the end of the FCM process. For the parallel implementation, all the clusters are aligned to divide the mappers.

### 3.2. Phase 2: Mapper phase

For large scale mobile data process, the mapper is a programming model and a connected implementation. Programmers only required to specify a Map-Reduce job which is composed of Reducer functions and the mapper. A Mapper function receives a key/value pair and generates a set of intermediate key/value pairs. With the same intermediate key, and a Reducer function merges all intermediate values are connected. Here, in parallel, the Mapper receives the clustered data and trains the AANN. Then established on all the Mappers output network model, the Reducers improve an AANN model to predict for unknown/newer data.

### 3.3. Phase 2(a): assigning each data groups to separate Mappers

In the Mapper phase, the clustered data is now processed. Mapper receives several items of the training sets (i.e. data items from the cluster groups) and accomplishes many mapper tasks. Each Mapper receives one training item (i.e. data items from one cluster group) and then performs AANN learning/training task by maximizing the weight values in the network applying this training item; so as to develop the learning efficiency. Through the AANN algorithm, their outputs are the trained network models resulted. The Mapper process (i.e. the AANN training procedure) is accomplished repeatedly until the expected precision is attained.

### 3.4. Phase 2(b): training data clusters using parallel adaptive ANN

Artificial neural network is otherwise named as Neural Network (NN). For calculation, it contains of an interconnected collecting of artificial neurons and procedures data applying a connectionist approach. Here a feed forward neural network (FFNN) inaugurated by this work. The data moves in just a single direction, forward, from the input layers, through the hidden layers, and to the output layers by this system. There are no cycles or circles in the system. The information handling can stretch out over numerous (layers of) units, yet no criticism associations are available, that is, an association reaching out from outputs of units to contributions of units in a similar layer or past layers is not present. There are associations among the processing elements (PEs) in every layer that have a weight (parameter) connected with them. Amid preparing this weight is balanced. The proposed adaptive ANN renders the optimal training network aligned by optimally selecting the interconnection weights among the hidden and output layers applying Whale Optimization Algorithm.

Input information is displayed to the system and proliferated through the system until it attains the output layer in FFNN. A predicted output is delivered by this forward procedure. The desired output is subtracted from the actual output and error esteem for the systems is ascertained. The error function can be characterized as:

$$e(w) = \sum (desired - actual\ output)^2 \tag{5}$$

For altering weights, a couple of traditional researches has applied Backpropagation learning algorithm. In reverse through the system, the calculation begins with the weights among the output layer PE's and the last hidden layer PE's and works. Once back propagation has fulfilled, the forward procedure begins once more, and this cycle proceeds until the error among is predicted and actual output are reduced. Rather than back propagation algorithm, the Whale Optimization algorithm is displayed because it can acquire valuable output than back propagation calculation.

The proposed Adaptive Artificial Neural Network model is given in below **Figure 2**.

**a.** Whale optimization approach

Recently a novel optimization algorithm named whale optimization algorithm (Mirjalili 2016) has been introduced to metaheuristic algorithm by Mirjalili and Lewis. As highly intelligent animals with motion, the whales are conceived. The WOA is inspired by the unique hunting behavior of humpback whales. The humpback whales prefer to hunt krills or small fishes which are close to the surface of sea at normally. Humpback whales use a special unique hunting method named bubble net feeding method. In this method they swim around the prey and produce distinctive bubbles along a circle or 9-shaped path. The mathematical model of WOA is described in the following sections a) Encircling prey b) Bubble net hunting method

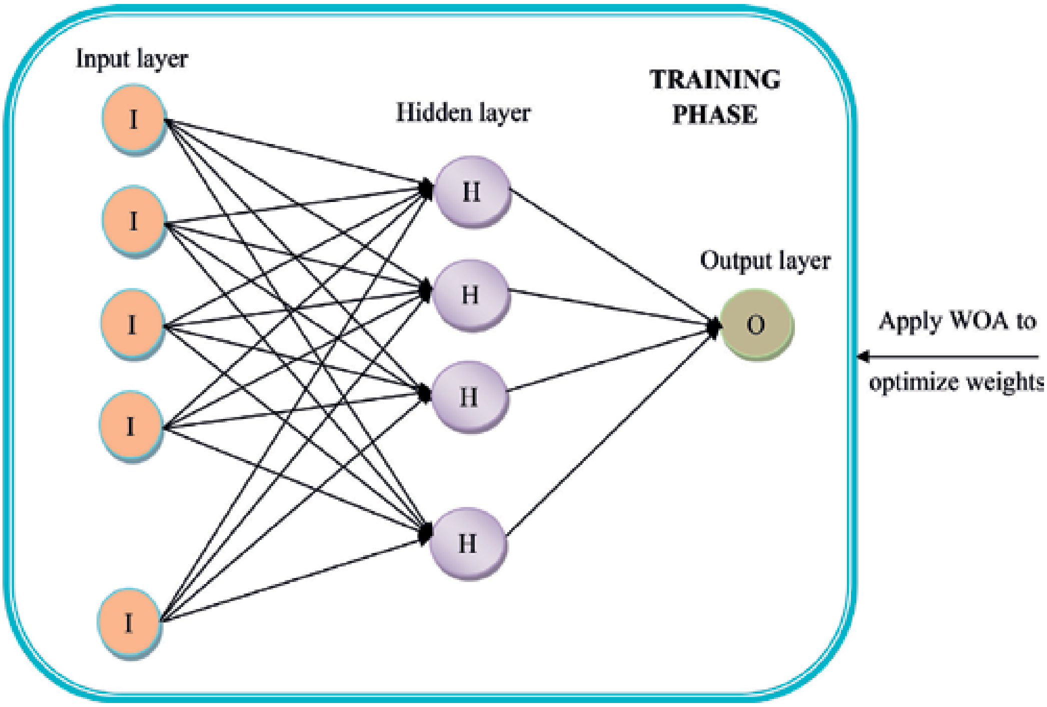


**Figure 2.** Proposed adaptive artificial neural network.

and c) Search the prey. The steps admitted in the proposed Whale optimization algorithm for rendering the optimal network structure by maximizing the interconnection weights of the neurons are afforded as follows,

Step 1: Initialization.

The algorithm is showed by arbitrarily generating the solutions (i.e. the interconnection weight values) that communicates to the result. Here the neural network structure comprising the interconnection weights among the hidden layers and the output layers are referred by the random value in the search space is afforded as:

$$W_m = \{w_{m,1}, w_{m,2}, \ldots, w_{m,h}, \ldots, w_{m,R}\} \tag{6}$$

where, $\{w_{m,1}, w_{w,m}, \ldots, w_{m,h}\}$ represents the set of weights among the input and hidden layer and $\{w_{m,h}, \ldots, w_{m,R}\}$ represents the set of weights among the input and hidden layer. Also, each solution, $W_m = \{w_{m,1}, w_{m,2}, \ldots, w_{m,h}, \ldots, w_{m,R}\}$ is a $R$-dimensional vector where $R$ being the number of optimization parameters. And also start the coefficient vectors of whale such as, $\vec{f}\ ,\ \vec{F}\ , and\ \vec{G}$.

Step 2: Fitness Calculation.

Determine the fitness of the input solutions on the basis of the Eq. (7). To get the best weight values, the fitness value of the solutions are computed. It's revealed in below,

$$Fit_{W_m} = \min(MSE) \tag{7}$$

The minimum of mean square error (MSE) determines that, how correct the network predicted targets are (i.e. high classification accuracy) in above equation. Hence, for further development, the initial solution with minimum error is chosen as best solution and checked.

Step 3: Update position of current solutions towards the best

**A.** Encircling prey

The position of prey (i.e. the current best solution) is distinguished by humpback whale and then it encircles the prey. Towards the best search operator the other search operators will consequently attempt to update their positions when the best search agent is characterized. The updation method is determined by the below equations:

$$\vec{N} = \left| \vec{G} \cdot \vec{W}_m^{best}(x) - \vec{W}_m(x) \right| \tag{8}$$

$$\vec{W}_m(x+1) = \vec{W}_m^{best}(x) - \vec{F} \cdot \vec{N} \tag{9}$$

where '$\vec{W}_m(x+1)$' denotes the newer solutions for next iteration, $\vec{F}$ and $\vec{G}$ denotes a Coefficient vector, $\vec{W}_m^{best}$ denotes a position vector for best solution, $\vec{W}_m(x)$ represents a current position Vector and || represents an absolute value.

The vectors $\vec{F}$ and $\vec{G}$ are calculated as follows:

$$\vec{F}= 2\ \vec{f}\ \cdot \vec{k}\ - \vec{f} \tag{10}$$

$$\vec{G}= 2\cdot \vec{k} \tag{11}$$

where, $\vec{f}$ is linearly reduced from 2 to 0 during the course of iterations (in both exploration and exploitation phases), $\vec{k}\ \in (0,1)$.

**B.** Bubble-net attacking method (exploitation phase)

To model the bubble-net behavior of humpback whales mathematically two approaches developed are a) Shrinking encircling mechanism and b) Spiral updating position

**a.** Shrinking encircling mechanism

The value of $\vec{f}$ in the Eq. (10) is reduced to attain this behavior. Note that $\vec{f}$ is applied to reduce the variation range of, $\vec{F}$. In other words, where $\vec{f}$ is minimized from 2 to 0. The novel position of a search agent can be determined anywhere by setting the random value, $\vec{F}$ from $[-1,1]$.

**b.** Spiral updating position

A spiral equation among the position of whale and prey is produced to mimic the helix-shaped movement of humpback whales is as follows:

$$\vec{W}_m\ (x+1) = N^{'}.\exp^{pq} \cdot \cos\left(2\prod q\right) + \underset{m}{\vec{W}}^{best}(x) \tag{12}$$

where $N^{'} = \left|\vec{W}_B^{best}(x) - \vec{W}_m\ (x)\right|$ and denotes the distance of the $B^{th}$ whale (which is the best solution obtained so far) to the prey, $q$ is the random value from, $[-1,1]$, $p$ denotes the shape of the logarithmic spiral and it is a constant value. During maximization the position of whales is updated by assuming a probability of 50% by choosing either the spiral model or shrinking encircling mechanism to model this simultaneous behavior. The mathematical model is afforded by Eq. (13).

$$\vec{W}_m\ (x+1) = \begin{cases} \vec{W}_m^{best}(x) - \vec{F} \cdot \vec{N}, & if\, L < 0.5 \\ N^{'}.\exp^{pq} \cdot \cos(2\prod q) + \vec{W}_m^{best}(x), & if\, L \geq 0.5 \end{cases} \tag{13}$$

where, $L \in [0,1]$. The humpback whales search for prey randomly to form bubble net.

**c.** Search for prey (exploration phase)

To search for prey in exploration phase, the same search approach applied in the exploitation phase established on the variation of the $\vec{F}$ vector can be applied. In fact, allowing to the position of each other humpback whales search randomly. Therefore, to force search agent to move so far from a reference whale we use $\vec{F}$ with the random values greater than 1 or less than $-1$. In exploitation phase, the position of the search agent is updated. This mechanism and $\left|\vec{F}\right| > 1$ emphasize exploration permit the WOA algorithm to perform a global search. The mathematical model is afforded below:

$$\vec{N} = \left|\vec{G} \cdot \vec{W}_m^{rand}(x) - \vec{W}_m(x)\right| \tag{14}$$

$$\vec{W}_m(x+1) = \vec{W}_m^{rand}(x) - \vec{F} \cdot \vec{N} \tag{15}$$

where, $\vec{W}_m\ rand(x)$ is a current population random position vector. Search agents update their positions at each iteration with respect to either the best solution found so far or a randomly selected search agent. In order to render exploration and exploitation the parameter '$\vec{f}$' is reduced from 2 to 0, respectively. A random search agent is selected when $\left|\vec{F}\right| > 1$, while the best solution is chosen when $\left|\vec{F}\right| < 1$ for the position of the search agents for updating. Depending on the value of '$L$', WOA is able to switch development either a circular or spiral movement.

The solutions are updated established on the best search agent between the current solutions found from the fitness evaluation step during each iteration. Again, at each time of generating newer weight values, it is aligned to the network and the fitness is determined and established on the back propagation error (i.e. the min MSE), which is the fitness function.

Step 4: Termination criteria.

Once the optimal weights are generated for all the networks of the Mappers, the training of the networks is finished. Now the AANN becomes a classifier and it can be generalized to predict for newer data also. The output mapped networks are then forwarded to Reducer phase.

To create the optimal network structure, the WOA algorithm is finished when best weight values are found. Also, the satisfaction of a termination criterion is confirmed when the mean square error is decreased to the needed limit or when the maximum iteration is attained.

Once the optimal weights are rendered for all the networks of the Mappers, the training of the networks is completed. Now the AANN gets a classifier and it can be generalized to predict for newer data also. The output mapped networks are then forwarded to Reducer phase.

### 3.5. Phase 3: Reducer phase

A Reducer accepts the data element of each Mapper for each Reducer task. With the same intermediate key, and a Reducer function merges all intermediate values connected. Established on the requirement, the Reducers can be customized. The proposed healthcare analysis model needs only one Reducer for improving a classifier model that must separate the patient's medical records. Here, the Reducer task is to form a robust classifier model from the parallely trained network models. Since the Reducer results in only one classifier network model, it will average all the maximized weight results for each interconnection links found for each training item and find the final optimal weights of the classifier. Here the analyzing data's (i.e. the unknown/newer data) are separated in the minimized AANN classifier model found from the Reducer phase.

## 4. Result and discussion

This section comprises result and discussion about the proposed parallel AANN (Adaptive Artificial Neural Network) technique for health care analysis from big data in cloud environment. The proposed algorithm is accomplished through JAVA software and the experimentation is carried out applying a system of having 4 GB RAM and 2.10 GHz Intel i-3 processor.

For estimating the performance of the proposed FCM based accuracy, Map-Reduce model, time, memory, precision, and recall are taken into an account and equated with the existing k-means based Map-Reduce and DBSCAN model. The experimental results for the suggested FCM based Map-Reduce model and other being k-means based Map-Reduce model and DBSCAN are tested in this section. The prediction efficiency is evaluated established on differentiating the number of records and number of mappers.

### 4.1. Performance analysis

The performance judgment of the proposed FCM based Map-Reduce model to predict the inauguration of abnormality in the medical data records is established in this section and equated with accomplishing k-means based Map-Reduce and DBSCAN method. The efficiency of our proposed method is evaluated in terms of time, memory, precision, recall and accuracy established on number of records and number of mappers.

### 4.2. Case (1): Performance analysis based on varying data size

#### *4.2.1. Time*

In the medical data records, an effective method should minimize the time needed to predict the inauguration of abnormality. The proposed FCM based Map-Reduce model decreases the time while equating with other accomplishing k-means based Map-Reduce and DBSCAN method.

**Figure 3** establishes the time needed for prediction of abnormality applying our proposed FCM based Map-Reduce model and k-means based Map-Reduce and DBSCAN model while the number of records rises. This clearly establishes that our proposed FCM based Map-Reduce model decreases the time needed for prediction of abnormality while equating with other accomplishing k-means based Map-Reduce and DBSCAN method.

#### *4.2.2. Memory*

An effective method should decrease the requirement of memory. The proposed FCM based Map-Reduce model decreases the requirement of memory while equating with other accomplishing k-means based Map-Reduce and DBSCAN method.

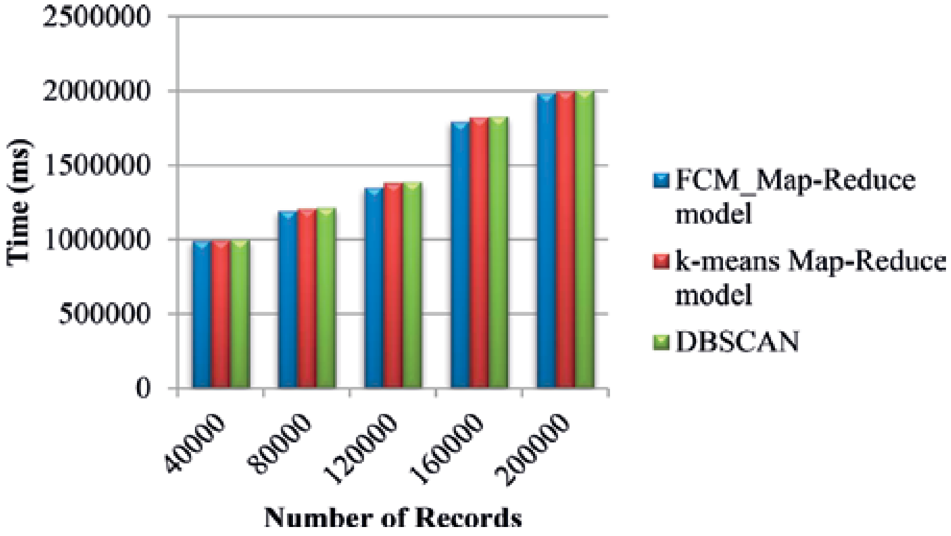


**Figure 3.** Time taken by FCM based Map-Reduce; k-means based Map-Reduce model for prediction.

**Figure 4** demonstrates the memory needed for our proposed FCM based Map-Reduce model and k-means based Map-Reduce and DBSCAN model while the number of records increases. This clearly establishes that our proposed FCM based Map-Reduce model decreases the memory requirement while equating with other accomplishing k-means based Map-Reduce and DBSCAN method.

#### *4.2.3. Precision*

The method with high precision will be more efficient. The proposed FCM based Map-Reduce model maximizes the precision while equating with other being k-means based Map-Reduce and DBSCAN method.

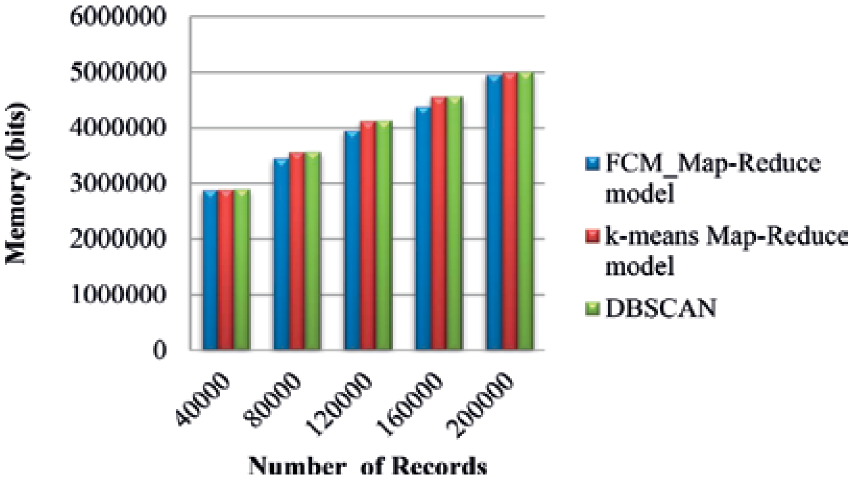


**Figure 4.** Memory requirement for FCM based Map-Reduce and k-means based Map-Reduce model.

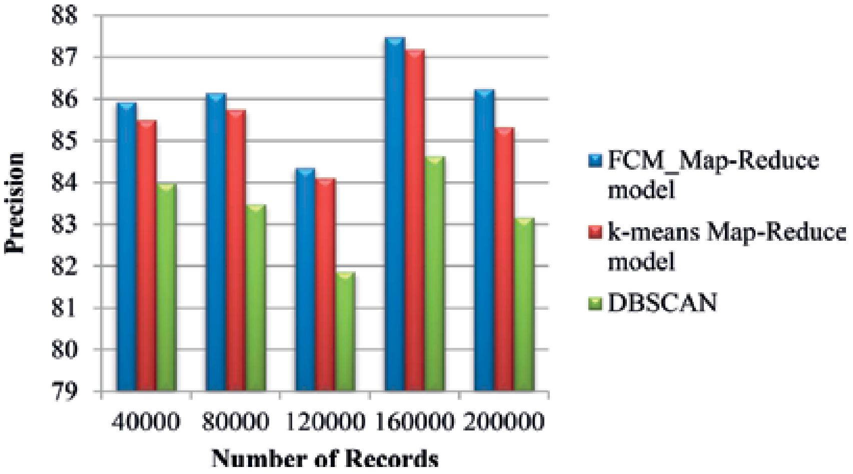


**Figure 5.** Precision for FCM based Map-Reduce; k-means based Map-Reduce and DBSCAN model.

**Figure 5** establishes the precision level for our proposed FCM based Map-Reduce model and k-means based Map-Reduce and DBSCAN model while the number of records rises. This clearly demonstrates that our proposed FCM based Map-Reduce model rises the precision level while equating with other being k-means based Map-Reduce and DBSCAN method.

### *4.2.4. Recall*

The method with high recall is said to be more effective. The proposed FCM based Map-Reduce model rises the recall while equating with other being k-means based Map-Reduce and DBSCAN method.

While the number of records rises, the **Figure 6** establishes the recall for the proposed FCM based Map-Reduce model and k-means based Map-Reduce and DBSCAN model. This clearly

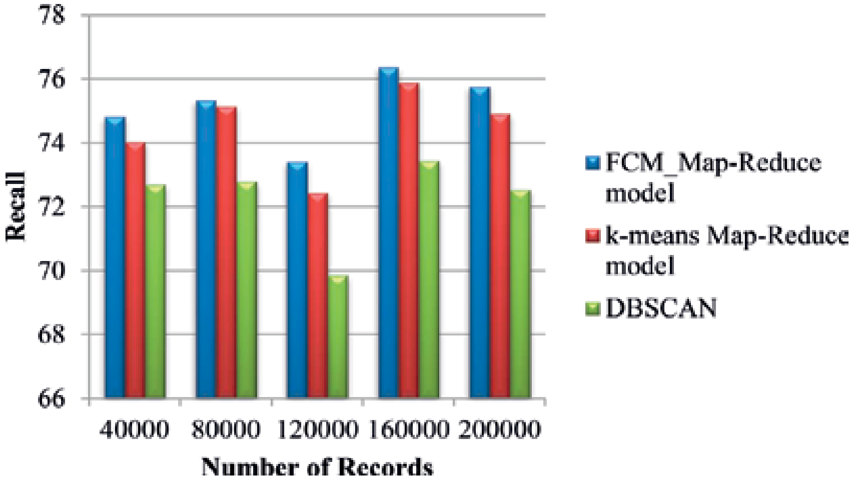


**Figure 6.** Recall for FCM based Map-Reduce; k-means based Map-Reduce and DBSCAN model.

establishes that the proposed FCM based Map-Reduce model rises the recall while equating with other being k-means based Map-Reduce and DBSCAN method.

#### *4.2.5. Accuracy*

The method with high accuracy is said to be more effective. With other being k-means based Map-Reduce and DBSCAN method, for the proposed FCM based Map-Reduce model raises the accuracy while comparing.

**Figure 7** establishes the accuracy for the proposed FCM based Map-Reduce model and k-means based Map-Reduce and DBSCAN model while the number of records rises. This clearly demonstrates that the proposed FCM based Map-Reduce model raises the accuracy while equating with other accomplishing k-means based Map-Reduce and DBSCAN method.

### 4.3. Case (2): performance analysis by varying number of Mapper

#### *4.3.1. Time*

In the medical data records, an effective method should decrease the time needed to predict the presence of abnormality. The proposed FCM based Map-Reduce model decreases the time while equating with other accomplishing k-means based Map-Reduce and DBSCAN method.

**Figure 8** establishes the time needed for prediction of abnormality applying the proposed FCM based Map-Reduce model and k-means based Map-Reduce model while the number of mapper rises. This clearly demonstrates that the proposed FCM based Map-Reduce model decreases the time needed for prediction of abnormality while equating with other accomplishing k-means based Map-Reduce method.

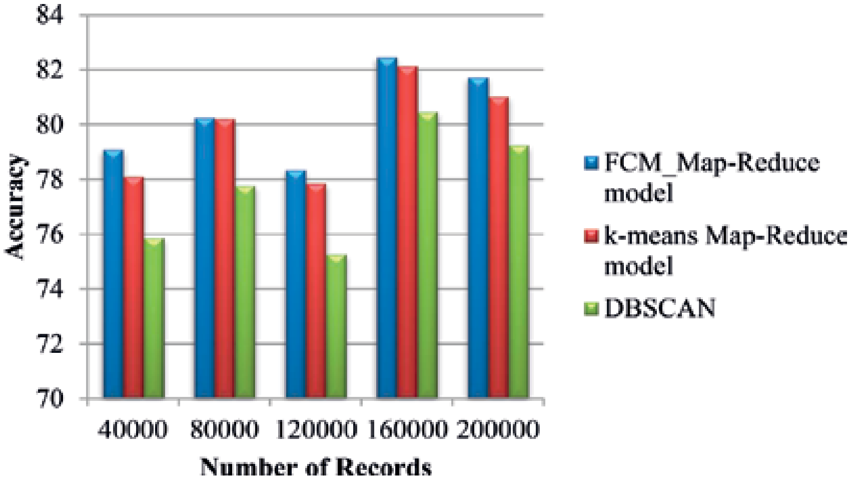


**Figure 7.** Accuracy for FCM based Map-Reduce; k-means based Map-Reduce and DBSCAN model.

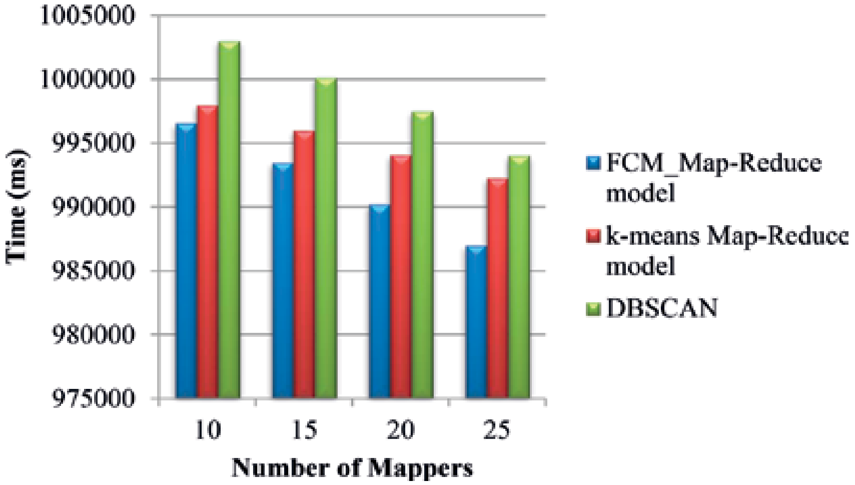


**Figure 8.** Time taken by FCM based Map-Reduce; k-means based Map-Reduce model for prediction.

### 4.3.2. *Memory*

An effective method should decrease the requirement of memory. While equating with other accomplishing k-means based Map-Reduce and DBSCAN method, for the proposed FCM based Map-Reduce model decreases the requirement of memory.

While the number of mapper rises, **Figure 9** establishes the memory needed for the proposed FCM based Map-Reduce model and k-means based Map-Reduce model and k-means base model. This clearly demonstrates that the proposed FCM based Map-Reduce model decreases the memory requirement while equating with other accomplishing k-means based Map-Reduce and DBSCAN method.

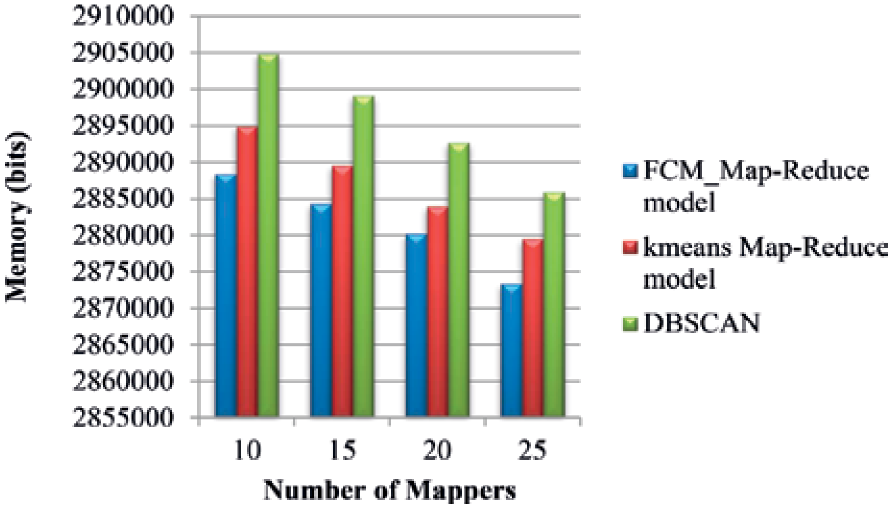


**Figure 9.** Memory requirement for FCM based Map-Reduce; k-means based Map-Reduce and DBSCAN model.

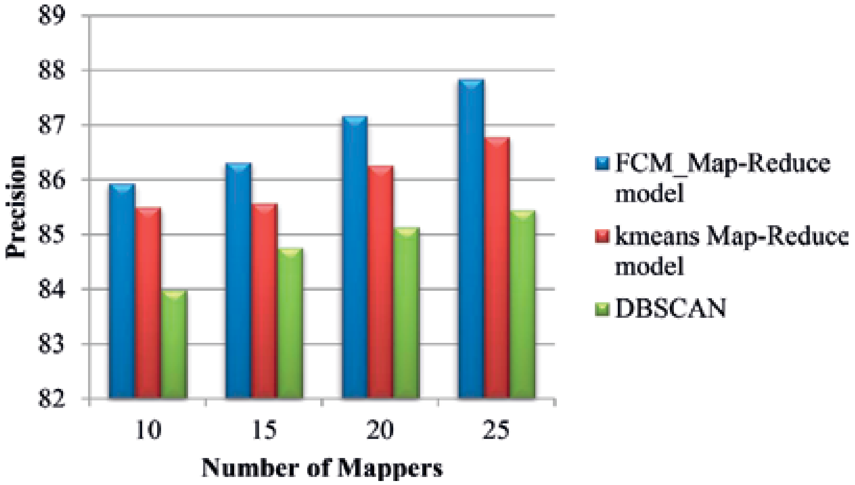


**Figure 10.** Precision for FCM based Map-Reduce; k-means based Map-Reduce and DBSCAN model.

### *4.3.3. Precision*

The method with high precision will be more effective. The proposed FCM based Map-Reduce model raises the precision while equating with other accomplishing k-means based Map-Reduce method.

While the number of mapper rises, **Figure 10** establishes the precision level for the proposed FCM based Map-Reduce model and k-means based Map-Reduce model. While equating with other accomplishing k-means based, Map-Reduce method this clearly demonstrates that the proposed FCM based Map-Reduce model rises the precision level.

### *4.3.4. Recall*

The method with high recall is said to be more effective. Our proposed FCM based Map-Reduce model raises the recall while comparing with other accomplishing k-means based Map-Reduce method.

**Figure 11** establishes the recall for the proposed FCM based Map-Reduce model and k-means based Map-Reduce and DBSCAN model while the number of mapper rises. This clearly demonstrates that the proposed FCM based Map-Reduce model raises the recall while equating with other accomplishing k-means based Map-Reduce and DBSCAN method.

### *4.3.5. Accuracy*

The method with high accuracy is said to be more effective. With other being k-means based Map-Reduce and DBSCAN method, for the proposed FCM based Map-Reduce model raises the accuracy while comparing.

While the number of mapper increases, the **Figure 12** establishes the accuracy for the proposed FCM based Map-Reduce model and k-means based Map-Reduce and DBSCAN model. This

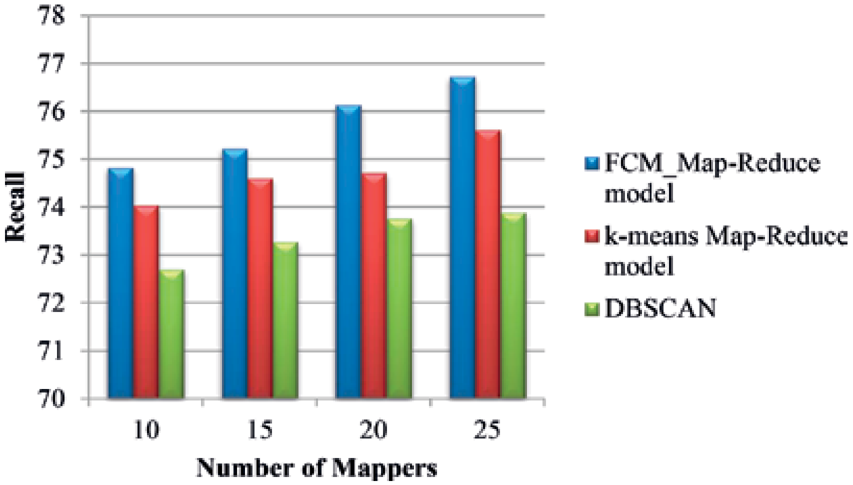


**Figure 11.** Recall for FCM based Map-Reduce; k-means based Map-Reduce and DBSCAN model.

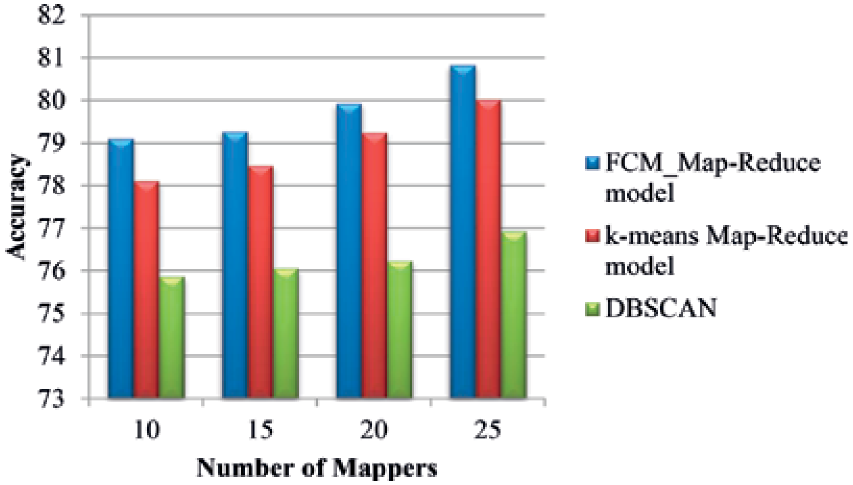


**Figure 12.** Accuracy for FCM based Map-Reduce; k-means based Map-Reduce and DBSCAN model.

clearly demonstrates that the proposed FCM based Map-Reduce model raises the accuracy while equating with other accomplishing k-means based Map-Reduce and DBSCAN method.

## 5. Conclusion

The presented research method have improved a FCM based Mapreduce programming model for the implementation parallel calculating applying Adaptive Artificial Neural Network approach for the prediction of abnormality of medical records. The proposed FCM based Mapreduce model is equated with the accomplishing k-means based Mapreduce and DBSCAN model and tested in terms of different evaluates like time, memory, precision, recall and accuracy by differentiating the data size as well as the number of mappers. It can be seen from the results that, all the values found for the proposed method is better when equated to the being method. Moreover, the time and memory requirements are very much minimized when the number of mappers is raised. This establishes the efficiency of proposed model and so the proposed application can be applicable for handling large healthcare databases in cloud environment.

## Author details

Manaswini Pradhan[1,2]*

*Address all correspondence to: mrs.manaswini.pradhan@gmail.com

1 Visiting Researcher School of Computer Science and Software Engineering, East China Normal University, Shanghai, China

2 P.G. Department of ICT, Fakir Mohan University, India

## References

[1] Suthaharan S. Big data classification: Problems and challenges in network intrusion prediction with machine learning. ACM Sigmetrics Performance Evaluation Review. 2014; **41**(4):70-73

[2] Tawalbeh L'a A, Mehmood R, Benkhlifa E, Song H. Mobile cloud computing model and big data analysis for healthcare applications. IEEE Access. 2016;**4**(99):1-12

[3] Ramesh D, Suraj P, Saini L. Big data analytics in healthcare: A survey approach. In: Microelectronics, Computing and Communications (MicroCom). 2016

[4] Chandola V, Sukumar S, Schryver J. Knowledge discovery from massive healthcare claims data. In: Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining. 2013. pp. 1312-1320

[5] Fang R, Pouyanfar S, Yang Y, Chen S-C, Iyengar SS. Computational health informatics in the big data age: A survey. ACM Computing Surveys (CSUR). July 2016;**49**(1):1-36

[6] Wang Dingxian, Xiao Liu, and Mengdi Wang. "A DT-SVM strategy for stock futures prediction with big data." Computational Science and Engineering (CSE), 2013 IEEE 16th International Conference on. IEEE, 2013

[7] Bijalwan V et al. KNN based machine learning approach for text and document mining. International Journal of Database Theory and Application. 2014;**7**(1):61-70

[8] Tomar D, Agarwal S. A survey on data mining approaches for healthcare. International Journal of Bio-Science and Bio-Technology. 2013;**5**(5):241-266

[9] Fahad A et al. A survey of clustering algorithms for big data: Taxonomy and empirical analysis. Emerging Topics in Computing, IEEE Transactions on. 2014;**2**(3):267-279

[10] Belciug Smaranda et al. "Clustering-based approach for detecting breast cancer recurrence." Intelligent Systems Design and Applications (ISDA), 2010 10th International Conference on. IEEE, 2010

[11] Yassine A, Singh S, Alamri A. Mining human activity patterns from smart home big data for healthcare applications. IEEE Access. 2017;**5**:13131-13141

[12] Islam MM, Razzaque R, Hassan MM, Nagy W, Song B. Mobile cloud-based big healthcare data processing in smart cities. IEEE Access. 2017;**5**:11887-11899

[13] Hosseini M-P, Tran TX, Pompili D, Elisevich K, Soltanian-Zadeh H. Deep learning with edge computing for localization of epileptogenicity using multimodal rs-fMRI and EEG big data. In: Autonomic Computing (ICAC). 2017

[14] Schölkopf B et al. New support vector algorithms. Neural Computation. 2000;**12**(5):1207-1245

[15] Chandra B, Ratnesh K Sharma. "Fast learning for big data applications using parameterized multilayer perceptron." Big Data (Big Data), 2014 IEEE International Conference on. IEEE, 2014

[16] Liu H, Yu L. Toward integrating feature selection algorithms for classification and clustering. Knowledge and Data Engineering, IEEE Transactions on. 2005;**17**(4):491-502

[17] Belciug S, Gorunescu F, Salem A-B, Gorunescu M. Clustering-based approach for detecting breast cancer recurrence. In: Intelligent Systems Design and Applications (ISDA). 2010

# Identification of Research Thematic Approaches Based on Keywords Network Analysis in Colombian Social Sciences

José Hernando Ávila-Toscano,
Ivón Catherine Romero-Pérez,
Ailed Marenco-Escuderos and
Eugenio Saavedra Guajardo

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.76834

**Abstract**

The purpose of this research was to unveil the structure of knowledge of Social Sciences in Colombia through the analysis of thematic networks and its association with different disciplines' new knowledge production to define scenarios and trends in each. 2992 published articles in the period 2006–2015 were revised in this research, all indexed in Web of Science, Scopus and other bibliographic databases, applying the social networks analysis technique to the keywords of all. The analysis included each discipline's clustering coefficient and group metrics. The results described in this chapter identify how social disciplines in Colombia have mainly focused its research production in topics such as armed conflict, poverty and human development.

**Keywords:** social sciences, field of study, network analysis, keywords, Colombia

## 1. Introduction

Social Sciences agglomerate disciplines with a diverse object of study. Such diversity defines a thematic amplitude according to the multiple approaches and perspectives that surround social studies. Among the different disciplines, the appearance of distant methodological offers leads to significantly dissimilar analysis focuses regarding a same phenomenon. For example, psychology applies quantitative research methods that belong to “hard sciences” through which it associates variables, predicts or explains human behavior. But also, psychology develops studies with qualitative approaches focused on the comprehension of the meaning around social and individual phenomena. Both of the approaches imply clearly differentiated theoretical corpora. Educational field’s research is predominantly qualitative [1], which also happens in Laws and Political sciences.

In summary, the study of Social Sciences is related to the variance of its object of study, which provides a complex, diverse and variable scenario of topics. For each scientific discipline, recognizing such topics means a significant contribution in the delimitation of the field of study, approaches posed or connections between investigative and theoretical proposals; it also helps researchers to determine which areas of work to explore or deepen and whom to cooperate with depending on the common interests.

## 2. The thematic-approach based analysis of scientific fields of study

Research in Social Sciences has lately changed its presentation, that is, with the publication formats used or association strategies between authors and institutions. Bibliometric research has mostly focused its interest on the productions indexed in Web of Science (WoS) and Scopus [2, 3], the internationalization of Social Sciences [4], the collaboration between authors and institutions [5–9] i.a. Bibliometric studies have made possible the characterization of the Social Sciences researchers’ scientific production, nevertheless, works surrounding thematic approaches analysis in this field of knowledge are still a proposal in development that has been evolving from the perspective of analysis and textual data visualization.

Studies with the goal to identify the structure of science are more common every day, seeking to understand areas of knowledge’s organization systems, relevant topics and research agenda. In other words, one of the most important tasks in scientometrics is the decomposition of scientific literature into disciplinary and sub-disciplinary structures [10]. This task is highly complex because it requires large volumes of information from which it is only possible to identify patterns through the use of computational tools [11].

The development of methods based on the usage of research production network metrics has been remarkably useful in the mapping construction. These studies have addressed diverse topics such as science interdisciplinary level [12], the multi-centric organization of basic sciences’ structure with bidirectional information flux in disciplines such as Physics, Chemistry or Medicine [13], the identification of edge and central concepts within a scientific discipline [14], the comparison of the different disciplines’ topological organization of quotation networks [15], the structural precision of the scientific panorama according to the size, similarities and interconnection among the different areas of science [16] i.a.

### 2.1. Social networks analysis and co-words method

Science maps or structures analyses also include the revision of scientific production’s semantic components as an identification mechanism of a particular area’s relevant topic of study.

For these purposes, the use of methods that take advantage of the textual resources and techniques of documentary summary has grown [15]. A useful tool that facilitates the identification of approaches and thematic trends in the different knowledge fields is the usage of metrics derived from text networks that define the interconnection of words or semantic fragments.

Social Network Analysis (SNA) is a common-used method in which the mapping procedures of terms' network structure [17] are made considering the relative importance of the concepts, the relational density level between them, that is, real connections given within all the possible connections, and the proximity among the different semantic units considered inside the networks.

Another possibility lies in the identification of clusters, from which the aggrupation of words according to its attraction is conformed, which gives rise to textual regions of high or low frequency where words are common or rare respectively [18].

Based on these analyses, knowledge maps are generated, they are built from the identification of co-occurrence in terms or words. This is a content analysis technique that eases the analysis of relations emerging within the ideas from a specific text [19, 20]. In the analysis, main topics from a scientific area are extracted and co-related to determine bonds between them [21] allowing the construction of hierarchies to define central research problems and (smaller) auxiliary areas [22].

### 2.2. Thematic approaches and contextualization of social problems

An important premise given by the SNA is the dynamic nature of the networks, which means that they can be structurally transformed when influenced by diverse variables [23]. For this reason, knowledge maps or networks are understood as dynamic structures, which contents are adjusted to the evaluation period and the immediate reality that affects the scientific work.

This is an especially important consideration in Social Sciences, given that social's relative nature is a feature inherent to the object of study of the disciplines that deal with these phenomena, therefore, the problems that social research addresses have a high contextual value [24, 25]. In general, the properties of scientific production are different between authors in the central region (North America & EU) and authors from peripheral scenarios as in the case of Latin-America [26, 27], this is noticeable in type of productions generated [28, 29] (central region has a higher number of works indexed in WoS and Scopus while peripheral area has preferences in local sources), as in the contents, themes of predilection and the significance value that the regional context gives to the studied problems.

This leads the social scientist to adjust the researched themes to the main problems of his immediate context [30, 27], which means that the networks of the thematic field can be substantially different depending on the social reality and the geographic position of the researchers. The study of meaning networks helps to understand those local-weighted differences defining the connections that facilitate the instauration of specific thematic axes in a field of knowledge.

This particular study applied the SNA fundamentals and the co-words method to identify the keyword networks in seven disciplines of the Social Sciences in Colombia. The results shared are part of a wider scientometric project that includes collaboration networks between authors and institutions and the assessment of research groups' scientific quality.

## 3. Methodological aspects of the study

2992 scientific articles published between 2006 and 2015 were analyzed for this study, all indexed in WoS, Scopus and other regional and international bibliographic databases (PsycINFO, Scielo, REDALyC, Political Science Complete, ProQuest, and others). The works belonged to the areas of Psychology, Education, Law, Sociology, Political Sciences, Journalism and Other Social Sciences (Economics, Anthropology, i.a). The production was developed by 7774 researchers from 168 research groups classified in the Colombian National System of Science, Technology and Innovation (SCIENTI-Col).

The keyword networks analysis was accomplished through the construction of vertical edge matrices using the NodeXL Excel Template (2016 version) software. For the generation of sub-groups inside networks the cluster coefficient of each area of knowledge was used. The cluster coefficient is the measure in which the nodes of a graph tend to cluster together with a relatively high density of the links. In this study, the clusters were calculated using the Clauset-Newman-Moore algorithm [31], which is highly effective for inferring community structure from network topology, being much faster than other algorithms that precede it, as well as allowing the calculation of community structure analysis in very large networks. Subsequently the group metrics were calculated, being: word counting by semantic group, number of established connections, maximum geodetic distance (DGM), its respective statistic measure (GDμ), and the relational density of each group. As a general criterion, it was defined that the main groups chosen would have a minimum integration of 10 keywords.

The visualizations of the networks (graphs) were organized using *grid algorithm,* which allows to clearly identify the sub-groups and their interaction.

The graph distribution was made with Harel-Koren Fast Multiscale [32], which eases the esthetic drawing of non-directed graphs with edges ordered in straight line, accomplishing the drawing procedure quickly for big networks. The node sizes were assigned according to the gross nodal degree obtained, that is the number of mentions in each term, the visualizations show all the nodes sized above 5, the lesser-graded nodes were overshadowed from 25 to 40% for the purpose of esthetic, simplicity and better readability of the graphs.

## 4. Fields of study in Colombian social sciences. Relevant results

**Table 1** describes the sociometric properties of each group identified in **Other Social Sciences.**

Being an area that integrates several disciplines, it is expectable for thematic groups to have diverse analysis lines. In group 1, for example, at least three sub-groups are differentiated; The first one is focused in science studies and scientific production, the second one emphasizes in the problems associated to violence and the third agglomerates social and anthropologic study proposals and methodological approaches.

Other lines are focused on experiences linked to the armed conflict and the necessity of the social capital recovery, as well as conflicts regarding the environment and the territories. Equally,

| Group | Group metrics | | | | | Grouped thematic lines (group name) |
|---|---|---|---|---|---|---|
| | Words | Connections | MGD[a] | GDμ[b] | Density | |
| G1 | 55 | 179 | 4 | 2.245 | 0.080 | Science studies—social violence—diverse social approaches |
| G2 | 24 | 75 | 6 | 3.007 | 0.170 | Post-conflict and social recovery |
| G3 | 22 | 75 | 4 | 2.112 | 0.212 | Socio-environmental conflicts |
| G4 | 21 | 67 | 4 | 2.222 | 0.210 | Social reintegration and Life quality |
| G5 | 21 | 63 | 4 | 2.150 | 0.190 | Child labor and children development |
| G6 | 20 | 66 | 5 | 2.300 | 0.211 | Urban development, basic needs and Health |
| G7 | 11 | 37 | 2 | 1.455 | 0.400 | Bibliometric studies |
| G8 | 10 | 33 | 2 | 1.380 | 0.467 | Administrative and financial processes |
| G9 | 10 | 23 | 4 | 1.860 | 0.333 | Logistics |

a = Maximum Geodetic Distance.

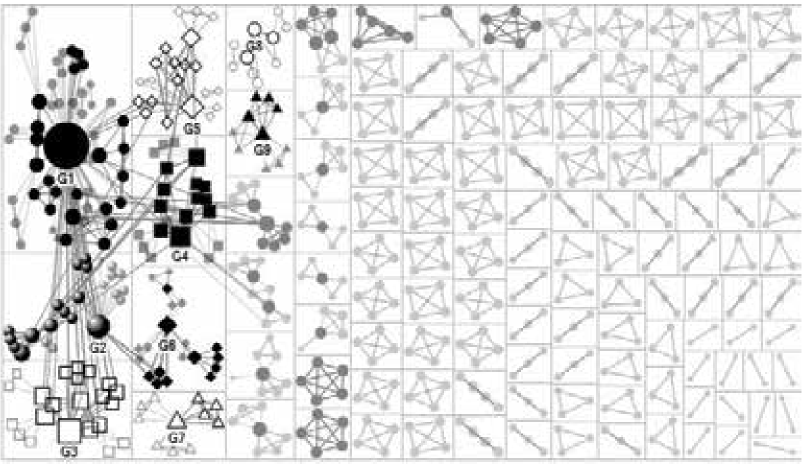
b = Mean geodetic distance.

**Table 1.** Keywords networks analysis in Other Social Sciences: description of metrics by thematic lines.

studies surrounding health, life quality and well-being of individuals in the civil reintegration process are highlighted. The scientific interest is also attracted by child labor, particularly referring to exploitation, violation of rights and exposure to risk conditions associated with work.

Groups 8 and 9 cross the eminent social barrier of the other groups, focusing on financial or administrative nature issues. Clustering terms related to the organizational activities and their financial affairs.

**Figure 1** shows that thematic groups in Other Social Sciences have low density. The main inter-group relations are given between groups 1, 2 and 3, and between 2 and 4. Few thematic



Created with NodeXL (http://nodexl.codeplex.com)

**Figure 1.** Graph of thematic groups with inter-group relations in Other Social Sciences.

interactions are found, perhaps because it is an area that includes research groups from different disciplines.

Information regarding thematic lines identified in **Psychology** appear in **Table 2**. In this discipline, the highest number of interconnected terms is related to neuroscientific issues and mental health, syndromes, conduct disorders and neuropsychological problems.

Several therapeutic approaches also excel inside relevant study topics, with cognitive system as precedence. Gender approach has an important role, specially facing phenomena as violence and sexuality. The qualitative notion of health and illness is seen from a perspective focused on the meanings constructed around the disease experience. The dynamic is also supported with other thematic lines directed to the meaning construction in front of social phenomena.

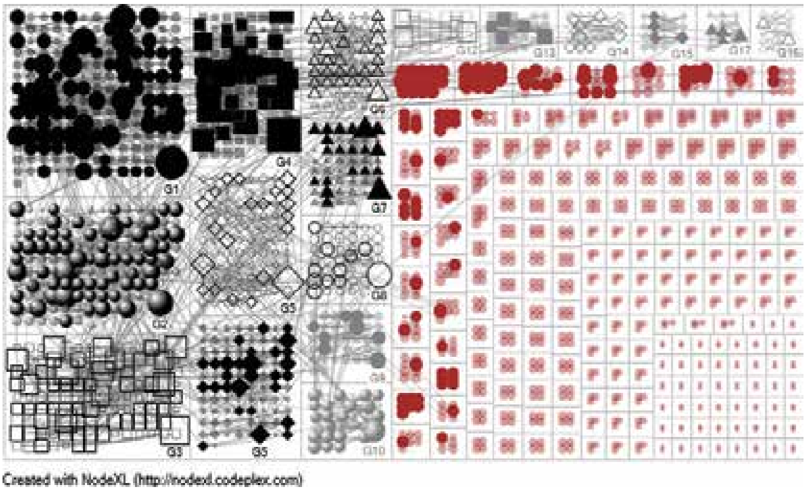
We observed interest in the social nature of psychological conditions of teens and children, by approaching violence, aggression and their related elements as factors that influence mental health and functional and non-functional behavior.

| Group | Group metrics | | | | | Grouped thematic lines (group name) |
|---|---|---|---|---|---|---|
| | Words | Connections | MGD[a] | GDμ[b] | Density | |
| G1 | 183 | 944 | 9 | 3.888 | 0.032 | Health and neuropsychological rehabilitation |
| G2 | 132 | 642 | 12 | 4.720 | 0.039 | Clinical psychology, psychotherapy and public health |
| G3 | 119 | 581 | 10 | 3.908 | 0.044 | Violence, aggression and mental health |
| G4 | 92 | 448 | 11 | 4.164 | 0.058 | Qualitative social research and consumer psychology |
| G5 | 86 | 336 | 7 | 3.096 | 0.051 | Quantitative methods; ethics and social/family conflicts |
| G6 | 82 | 330 | 9 | 4.166 | 0.055 | Family, attachment relationships and gender |
| G7 | 53 | 205 | 5 | 3.104 | 0.082 | Basic and applied psychology epistemology |
| G8 | 46 | 231 | 4 | 2.408 | 0.118 | Basic research and neurosciences |
| G9 | 41 | 113 | 8 | 3.367 | 0.090 | Instrumental studies in clinical psychology |
| G10 | 37 | 128 | 6 | 3.217 | 0.114 | Psychology and psychopharmacology |
| G11 | 36 | 148 | 5 | 2.924 | 0.140 | Health significances, illness and social development |
| G12 | 25 | 82 | 4 | 2.438 | 0.160 | Bibliometric studies |
| G13 | 22 | 74 | 6 | 2.711 | 0.173 | Subjectivities, violence and rights |
| G14 | 21 | 78 | 5 | 2.531 | 0.214 | Basic research in consumer psychology |
| G15 | 18 | 61 | 6 | 2.759 | 0.222 | Psychology of work and social inclusion |
| G16 | 15 | 38 | 5 | 2.347 | 0.238 | Positive approach in health and coping |
| G17 | 15 | 47 | 4 | 1.982 | 0.295 | Psychology of language |

a = Maximum Geodetic Distance.

b = Mean geodetic distance.

**Table 2.** Keywords networks analysis in Psychology: thematic lines metric description.



Created with NodeXL (http://nodexl.codeplex.com)

**Figure 2.** Graph of thematic groups with inter-group relations in Psychology.

Among other thematic groups, psychology research in Colombia approaches: (a) consumer behavior, publicity and alike, (b) analysis of family's cycle of life, affective relationships and family dynamics, highlighting a particular role in the analysis of the differences between men and women in issues related to family development, (c) axiological study of social problems through the integration of values, moral, reflexivity and relational, social, political conflict, (d) methodological studies (qualitative, quantitative), epistemological revisions and reflexive approaches about the discipline's reach. **Figure 2** shows keywords network for Psychology.

Colombian research produced in the area of **Laws** shares with the other described disciplines an interest towards armed conflict, human rights and diverse social and juridical issues associated with violence. Among the production of Law researchers also excel the topics related to international processes that involve human groups migration, States participation and the protection guarantees of the migrants. Another local context issues are clearly relevant such as armed conflict and transitional justice, illegal actors, disarmament processes and the relations with the victims. The approach of mediation stands out as an instrument associated with the construction of peace. In addition, terms inherent to law and juridical exercise ethics are also common (**Table 3**).

Inside the field of study around Laws in Colombia, the following topics have been objects of academic analysis: (a) protection relating to personal interests over real rights and properties, (b) historical sources revision, reflexive analysis of the juridical exercise, (c) ownership and tenure of land, crops and peasant labor, agro-industry and rural development from a perspective of agrarian law and social rights, (d) analysis of criminal law issues, from an accusatory perspective as well as a focus on the defendants and their protection guarantees, (e) the reality of women as a infringed individuals, integrating terms related to intimate, moral and sexual rights of women.

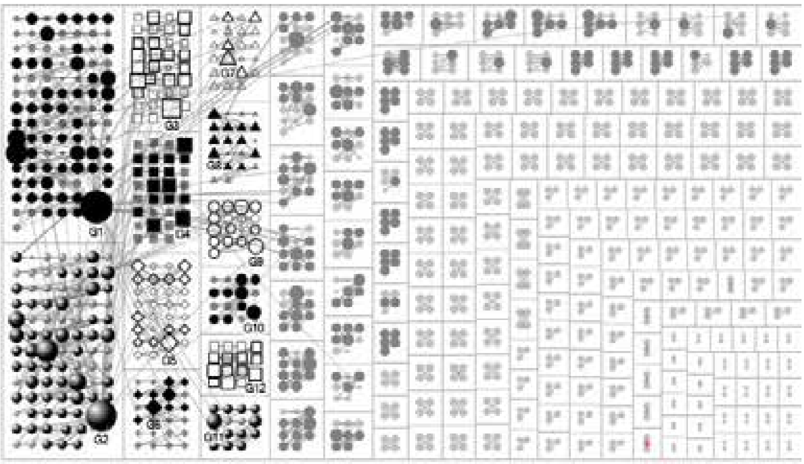
The graph (**Figure 3**) generated for this field of knowledge allows to identify a high connection level between keywords of each thematic group, moreover, there is a notorious relation between groups 2, 3, 5 and 6. For its part, group 1, which thematic is focused on conflict, migrations and

| Group | Group metrics | | | | | Grouped thematic lines (group name) |
|---|---|---|---|---|---|---|
| | Words | Connections | MGD[a] | GDμ[b] | Density | |
| G1 | 179 | 933 | 7 | 3.341 | 0.033 | Armed conflict, international right, migration and minorities with gender approach |
| G2 | 132 | 731 | 7 | 3.580 | 0.041 | Constitutional and administrative law |
| G3 | 95 | 476 | 12 | 4.449 | 0.054 | Conflicts, mediation, peace and law ethics |
| G4 | 88 | 389 | 8 | 4.153 | 0.055 | Special protection groups and social rights |
| G5 | 82 | 567 | 9 | 4.093 | 0.081 | Civil and patrimonial right |
| G6 | 69 | 365 | 10 | 3.751 | 0.081 | Violence, refugee protection and environment |
| G7 | 69 | 483 | 7 | 3.302 | 0.097 | Public law and supranational issues |
| G8 | 57 | 288 | 11 | 4.596 | 0.095 | Epistemology, history and Right |
| G9 | 40 | 207 | 5 | 2.565 | 0.138 | Agricultural, environmental issues and countryside dev. |
| G10 | 29 | 195 | 4 | 2.257 | 0.239 | Law and criminal processes |
| G11 | 24 | 95 | 5 | 2.448 | 0.185 | Violence against women and rights approach |
| G12 | 17 | 88 | 4 | 2.021 | 0.324 | International law studies |
| G13 | 17 | 100 | 3 | 1.647 | 0.404 | Public administration and crime studies |
| G14 | 15 | 81 | 3 | 1.662 | 0.419 | Individual freedoms and role of the state |

a = Maximum Geodetic Distance.

b = Mean geodetic distance.

**Table 3.** Keywords networks analysis in Laws: thematic lines metric description.



Created with NodeXL (http://nodexl.codeplex.com)

**Figure 3.** Graph of thematic groups with inter-group relations in Laws.

gender, constitutes a sub-group that agglomerates a vast number of keywords with high-leveled nodal degree, what exposes this group as the area in the field of Laws in Colombia with the highest exploration level in the last decade.

In the case of **Education** (**Table 4**), 12 thematic lines were identified as relevant, with an important variety of contents due to the reach of the groups registered in the Colombian scientific system, focusing not only in pedagogical processes at a school level, but also in university teaching and the study of sciences.

Research in the education field in Colombia over the period 2006–2015 has shown a strong interest in technology and its application to the education of basic sciences, and also has emphasized in elements related to teaching/learning methods and the pedagogical reflection in sciences and at general level.

A relevant area consists in the analysis of the role of education, school and educational actors in front of a diversity of social issues that affect the individual and its community cores. The notion of human development is certainly appreciated as a transversal element in the educational field, also denoting an interdisciplinary perspective of education.

Investigation, teacher's research formation, knowledge diffusion and its visibility are also themes approached by this discipline, accentuating the usage of technological tools in relation to the abilities and competences on its handle. In the second instance, there is interest regarding the disciplinary performance and the role of education as a mechanism of socio-political transformation (**Table 4**).

| Group | Group metrics | | | | | Grouped thematic lines (group name) |
|---|---|---|---|---|---|---|
| | Words | Connections | MGD[a] | GDμ[b] | Density | |
| G1 | 99 | 335 | 9 | 3.750 | 0.045 | Tech. basic sciences and didactical fundamentals |
| G2 | 90 | 326 | 11 | 4.018 | 0.050 | Education in front of social issues |
| G3 | 34 | 106 | 7 | 3.126 | 0.112 | Educational research, scientific visibility and problem intervention |
| G4 | 32 | 104 | 5 | 2.938 | 0.125 | Disciplinary and informational skills |
| G5 | 31 | 95 | 6 | 2.968 | 0.131 | Education and politics |
| G6 | 24 | 75 | 4 | 2.219 | 0.167 | Ethics and education |
| G7 | 23 | 56 | 7 | 3.009 | 0.142 | Violence and education sociohistorical approach |
| G8 | 22 | 72 | 6 | 2.888 | 0.190 | Cognitive functioning and performance |
| G9 | 17 | 66 | 4 | 2.035 | 0.272 | Teaching and teaching practice |
| G10 | 16 | 51 | 4 | 2.000 | 0.267 | Science teaching and mathematics |
| G11 | 15 | 54 | 3 | 1.831 | 0.305 | Specific teaching performance |
| G12 | 15 | 65 | 3 | 1.938 | 0.305 | Childhood, care and protection |

a = Maximum Geodetic Distance.

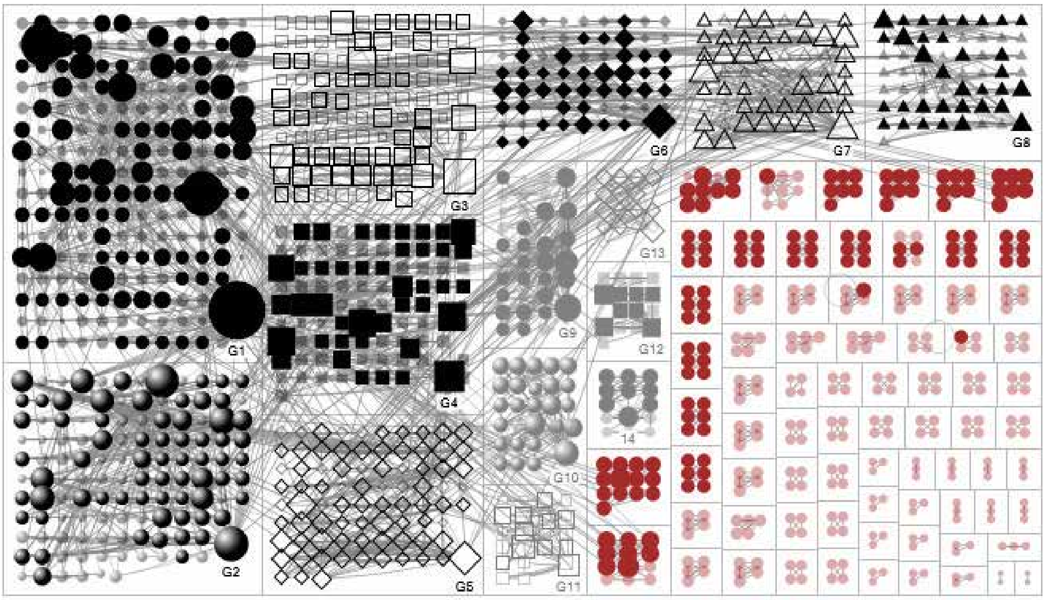
b = Mean geodetic distance.

**Table 4.** Keywords networks analysis in Education: thematic lines metric description.

Other relevant contents in Education are: (a) Political vision assessment as a linked factor of educational teaching. The analysis of politics excels from the citizen participation and the state, and also the concept of subject as political subject, (b) ethical analysis focused on teaching, (c) analysis in Colombian history perspectives about violence, (d) the analysis of cognitive skills and individuals' performance in the task execution, (e) especially in decisive strategies for problems or accomplishment of tasks in basic sciences specific areas of formation, (f) the teaching exercise and the role of the teacher. This includes theoretical points and pedagogical approaches of the educational activity, the teacher's performance and skills from a perspective of academic specialization, directing the interest to professional performance areas, (g) welfare of the children, integrating the education with complementarian areas which focus was centered in the attention, care and prosperity of the minors.

**Figure 4** graphically shows the suggested groups of keywords analysis for Education. It can be observed that groups 1,2,3 and 7 establish tight transitive connections. Highlighting group 1 (*Technology, basic sciences and didactical fundamentals*) as the most developed component in the educational area in terms of interconnected keywords and the number of connections with other thematic cores.

For the case of **Sociology,** the exceled thematic lines (**Table 5**) include topics in common with other disciplines, as violence, armed and social conflict, inequity, discrimination and social/economical gaps in specially protected populations such as women, natives and cultural minorities. Likewise, Colombian scientific production in sociology has focused on the investigative analysis of armed confrontation, sociopolitical issues and the nation's reality in terms of violence. It is stressed a positive approach focused on peace, reconciliation and social change, highlighting



Created with NodeXL (http://nodexl.codeplex.com)

**Figure 4.** Graph of thematic groups with inter-group relations in Education.

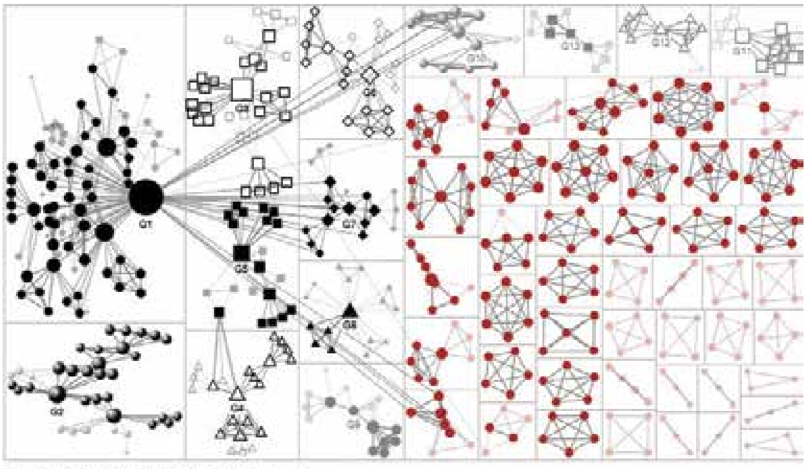
| Group | Group metrics | | | | | Grouped thematic lines (group name) |
|---|---|---|---|---|---|---|
| | Words | Connections | MGD[a] | GDμ[b] | Density | |
| G1 | 86 | 318 | 7 | 3.055 | 0.056 | Sociopolitical conflict and transition to peace |
| G2 | 37 | 157 | 7 | 3.002 | 0.143 | Public policies, migration and colonization |
| G3 | 33 | 126 | 6 | 2.621 | 0.140 | Cultural, gender and ethnic inequity |
| G4 | 22 | 86 | 5 | 2.566 | 0.212 | Educational sociology |
| G5 | 22 | 76 | 5 | 2.339 | 0.203 | Ecological/environmental sociology |
| G6 | 21 | 81 | 4 | 2.150 | 0.233 | Health and environment |
| G7 | 19 | 61 | 4 | 2.155 | 0.234 | Bio-geography, sociology and archaeologic-botany |
| G8 | 16 | 50 | 2 | 1.633 | 0.258 | Cultural socio-anthropology |
| G9 | 15 | 55 | 4 | 2.142 | 0.295 | Nativism, culture and health |
| G10 | 13 | 58 | 3 | 1.716 | 0.397 | Ethno-botany and sociology |

a = Maximum Geodetic Distance.
b = Mean geodetic distance.

**Table 5.** Keywords network analysis in Sociology: thematic groups metric description.

regional experiences and recovery surrounding violence (studies of case). Public policies have also been objects of study, as well as the development of social mobilization and settlement, population and territorial occupation activities, with its respective administrative, social and political implications. There is a notorious emphasis in forced displacement of population.

Besides the aforementioned, there is a diversity of themes identified in the sociological study including (a) sociological aspects in education, formation and educative quality, pedagogical exercise, superior education and teaching of sciences, (b) environmental reality, role of the society in the care or degradation of the ecological systems, (c) experiments related to human health, as well as social discourses and practices in terms of environment, (d) revisions of natural and geographic elements of past archeological eras, including the study of plants or investigations about archeology or anthropology, (e) analyses of human activities from a contextual perspective with qualitative approach, (f) studies focused on the analysis of native communities' health in general from an ethnographic perspective and native's activities towards healthcare and protection, (g) analysis of relation between human groups and vegetal environment. The network analysis also allowed to identify that these groups themselves are constituted as sub-units (mostly independent from each other) in the universe of Sociology keywords (**Figure 5**). Among the analyzed groups, only group 1 (*Sociopolitical conflict and transition to peace*) has inter-group relations defined with groups 7 and 10, nevertheless, this connection is given by the usage of the word *Colombia.* Despite the thematic similarity between groups 7 and 10, there is no direct connection detected among their topics. Apparently in Sociology there are defined thematic fields which production or study area is usually sectorized and with low tendency to interact between approaches within the same discipline.



Created with NodeXL (http://nodexl.codeplex.com)

**Figure 5.** Graph of thematic groups with inter-group relations in Sociology.

| Group | Group metrics | | | | | Grouped thematic lines (group name) |
|---|---|---|---|---|---|---|
| | Words | Connections | MGD[a] | GDμ[b] | Density | |
| G1 | 147 | 540 | 6 | 2.679 | 0.033 | Electoral studies, social violence and organized crime |
| G2 | 81 | 254 | 7 | 3.452 | 0.048 | Socio-economic development policies of nations |
| G3 | 76 | 263 | 8 | 3.704 | 0.058 | Foreign affairs, national security and politics |
| G4 | 72 | 304 | 7 | 3.403 | 0.070 | Political philosophy, transitional conflict and security |
| G5 | 63 | 208 | 9 | 4.204 | 0.068 | Administrative behavior and economic policies |
| G6 | 42 | 205 | 5 | 2.514 | 0.134 | Conflict-peace transition policies and national systems |
| G7 | 29 | 103 | 6 | 2.966 | 0.143 | Post-conflict and implied actors |
| G8 | 24 | 84 | 5 | 2.413 | 0.174 | Economic systems and military/security financing |
| G9 | 24 | 133 | 4 | 2.288 | 0.261 | Crime and role of public force |
| G10 | 22 | 95 | 4 | 2.107 | 0.234 | Productive sector responsibility in post-conflict |
| G11 | 14 | 43 | 4 | 1.939 | 0.275 | Collective participation in peacebuilding |
| G12 | 14 | 51 | 3 | 1.847 | 0.319 | Political behavior and participation |
| G13 | 11 | 45 | 3 | 1.504 | 0.455 | Urbane and environmental public policy |

a = Maximum Geodetic Distance.

b = Mean geodetic distance.

**Table 6.** Keywords network analysis in Political Sciences: thematic groups metric description.

The revision of **Political Sciences** is the last-but-one analysis, its thematic lines are varied (**Table 6**). the most important groups are related to Colombian sociopolitical conflict as the main emerging category, also integrating other types of organized violence and its social consequences. This field of study also focus the scientific interest in phenomena associated to electoral activities and the exercise of democracy, including their own diverse issues such as partisan conflict, clientelism, electoral volatility, a.o.

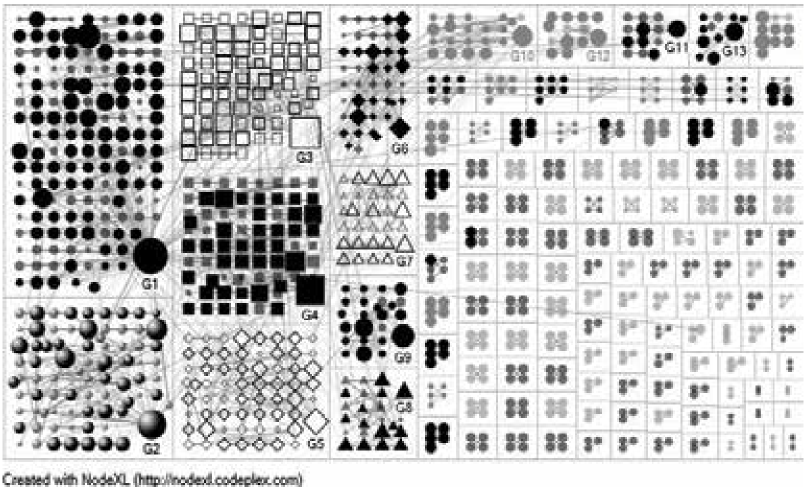
In addition, it is noticed an ethical, axiological and psychosocial perspective of the armed conflict and the restoration of the rights of groups and individuals. This thematic field does not ignore the role of the victims or demobilized people from illegal armed groups, addressing affairs related to the recovery status of political subjects-of-rights in the post-conflict.

The politic role in foreign relations is also a relevant topic, such as political, geographical and economical nature issues that constitute risks for the security of nations, assuming also the national risks based on the internal affairs in Colombia.

Other analytic field is the study of the economic development policies from the public administration including a regional perspective and proposals/reflections of economic development alternatives to the traditional model. Other highlighted topics are military expenses, security policies professionalization and the analysis of the economic development models and approaches. The study of crime is also noted, especially illegal organized activities related to the history of armed confrontation and war financing in Colombia. This group also weights the role of the police forces.

Armed conflict is without doubt the main topic in Social Sciences. Beside the described perspectives, other approaches direct the interest towards the private sector and institutions that contribute to the economy, in the development of productive processes that tend to contribute with social reintegration, as well as in the negotiations regarding the termination of armed conflict in Colombia and the participation of citizen groups in the peace-building.

**Figure 6** shows the distribution of each thematic group with its interrelations. The graph allows to identify an important relation between groups 3 and 8, which emerging topics are indeed related (See **Table 6**), Likewise, group 1 shares bonds with almost all of the other groups, especially because the bonding concept is *Colombia*, being contextualized studies. This group also interacts



Created with NodeXL (http://nodexl.codeplex.com)

**Figure 6.** Graph of thematic groups with inter-group relations in Political Sciences.

| Group | Group metrics | | | | | Grouped thematic lines (group name) |
|---|---|---|---|---|---|---|
| | Words | Connections | MGD[a] | GDμ[b] | Density | |
| G1 | 54 | 184 | 7 | 3.018 | 0.080 | Journalism, diffusion and social visibility |
| G2 | 51 | 193 | 7 | 3.155 | 0.095 | RSI, communication and social conflicts |
| G3 | 36 | 122 | 6 | 2.617 | 0.121 | Media and political news |
| G4 | 18 | 66 | 6 | 2.673 | 0.248 | Semiotics, cinema and literature |
| G5 | 13 | 43 | 4 | 2.178 | 0.308 | Journalism and audiovisual productions |
| G6 | 12 | 48 | 3 | 1.639 | 0.394 | Journalism and childhood protection |
| G7 | 12 | 48 | 3 | 1.639 | 0.394 | Journalism and history |
| G8 | 10 | 25 | 4 | 1.780 | 0.333 | Research in cinematography |

a = Maximum Geodetic Distance.

b = Mean geodetic distance.

**Table 7.** Keywords network analysis in Journalism: thematic groups metric description.

closely with groups 3, 5 and 10, by referring a wide spectrum of themes related to conflict, State's economic financing facing the ending of war, victims' recovery and social reintegration.

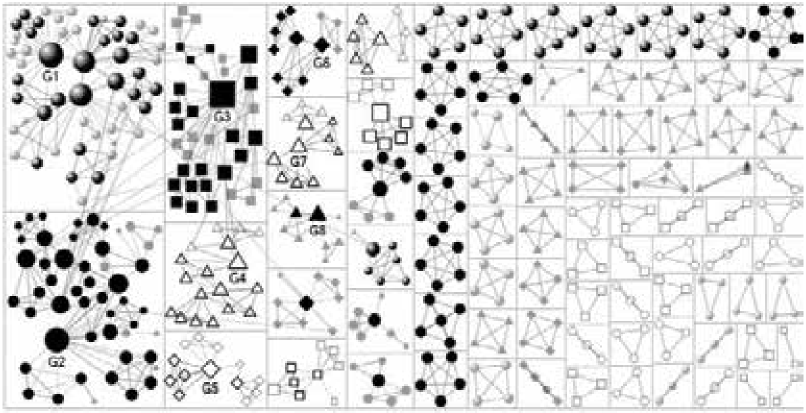
Finally, **Table 7** gathers data related to emerging thematic lines in the area of **Journalism.** The scientific interest in this area is for topics linked to the use of journalism and communications as a social impulse to highlight social minorities; political and sociocultural approaches are articulated to this line. Likewise, associated terms point an approach in the role of media and journalism in communication, information and entertainment, such approach gives relevance to the younger audiences and also emphasizes in the quality of the transmitted information. It is also observed a group of words directed towards formation in journalism.

Other important thematic field is the internet's social networks and their usage with informative ends in front of relevant social/political events, as well as the analysis of the participative ends in the use of web networks and the definition of relations mediated by the technology.

Colombian research in journalism does not ignore the visible reality of this nation, thereof, issues as forced displacement, poverty and security (integrated to the use of communication media) excel as a study topic.

Investigations related to the following topics are also relevant: (a) the role of media (conventional & digital) in the coverage and diffusion of national/regional politic news, (b) television and radio productions and the development of journalism activity. Also headlining the interest in research processes focused in the content and discourses of journalism, and the gaps of information access in the knowledge society, (c) studies related to infancy and childhood, the protection of their rights and the diverse processes that surround their development, (d) the Internet and its properties, this time in relation to the history and how journalism communicates such information.

In journalism, other two thematic lines are accented apart from the interest in social issues, focusing their interest on artistic and esthetic affairs, in one hand, there is research in literary issues, based on the revision of representative authors and works in literature and



Created with NodeXL (http://nodexl.codeplex.com)

**Figure 7.** Graph of thematic groups with inter-group relations in Journalism.

cinematography. On the other hand, there are social research and representations, as well as the use of images and cinema as object of study from journalism and communication.

**Figure 7** represents each one of the main groups descripted and their interaction. The graph shows interaction between most of the thematic groups, although clearly groups 5, 6 and 8 establish subgroups apart from the other terms, which indicates that they are thematic areas less integrated to the traditional lines of journalism in the sample, nevertheless, they are large enough to constitute significant groups of topics. Groups 1, 2 and 3 are connected instead, which shows the closeness of the topics and interests that they share.

## 5. Conclusions

The analysis of thematic approaches in Social Sciences using the keywords of productions (article type) contribute with the detection of important conceptual lines, from which is constructed a semantic identification map capable of exposing thematic scenarios and the most common and relevant topics of analysis for different disciplines [19–22]. This is a scientometric exercise that leads to the disintegration of a scientific discipline into main areas and sub-areas of knowledge, which allows to understand the main interests in scientific research at a disciplinary level. Such results are useful because they orientate researchers around methodological paradigms, theoretical movements and focuses of interest for the scientific community, providing tools to decide what guidance to give to contributions, or even which areas to reinforce with their own productions.

The social disciplines studied present networks of thematic approaches in which fragmentation is common, having multiple subgroups of terms with moderate levels of relational density and few connections between emerged groups of terms.

This clearly is not a phenomenon exclusive for the social investigation in Colombia, but constitutes a common denominator in the Social Sciences for the breadth of its object of study and even the methodological and theoretical oppositions that occur, sometimes within the same discipline.

| Disciplines | Main topics |
|---|---|
| Other Social Sciences | Politic violence, Displacement, Human rights, Human development, Memory, Dispossession, Armed conflict |
| Psychology | Risk factors, Child development, Violence against women, Domestic violence, Mental health |
| Laws | Discrimination, Human rights, Armed conflict, Political constitution, Conflict resolution, Civil society, Reintegration, Dignity, Homosexuality, Matrimonial property regimes |
| Education | Pedagogical practice, Applied didactics, Risk behavior, Maternity, Poverty, Educational policy, Education for peace, Informational skills, Democracy, State |
| Sociology | Conflict and region, Criminality, Social change and region, Victims, Public services, Residual gaps, Education quality |
| Political Sciences | Organized crime, Paramilitarism, Forced displacement, Armed conflict, Social policy, International economic policy, State performance, Humanitarian exchange, Transitional justice |
| Journalism | Minority language, Power, Displaced people, Social mobilization, New media, Online news, Messages, Poverty, Public space, Journalistic coverage, News sources, Infancy, Child abuse |

**Table 8.** Main keywords identified in Colombian Social Sciences.

It has been described that in recent years, research in Social Sciences has become more international [4]. The increase of quantitative methods and the use of information technology have eased the communication and the comparison of research results with geographically distant colleagues [33]. Nevertheless, collaboration or international source quotations are not sufficient criteria to define the contents from an internationalist trend perspective. Term networks with wide thematic variety emerge in the different social disciplines in Colombia, although they clearly share common interests related to the immediate reality of the nation. They are thematic structures highly focused on the local nature issues and impact over the region, thereof topics as armed conflict between illegal forces and the Colombian State, along with the sequels associated to itself, the violation of women and children's rights, the peacebuilding, and the challenges for human and social development constitute the main themes of all the studied disciplines. **Table 8** gathers the most frequent keywords of the field in general, that is, shared by different disciplines.

As described previously [30, 27], the context has a dominant role over the topics and approaches postulated in the study of Social Sciences. The social scientist places among his objectives the knowledge development capable of translating or changing the experienced reality, thereof that the thematic cores identified in Colombian Social Sciences are closely related to the main social, political and economic issues experienced by the nation.

Considering the dynamic character of the knowledge networks, it can be inferred that future research lines focus on the associated processes to the construction of a post-conflict society, this in accordance with the construction of peace processes between the Colombian State and the illegal armed actors.

## Acknowledgements

We express our gratitude to the Colombian Administrative Department of Science and Technology (COLCIENCIAS) for the financing of the Project entitled "Scientology of the field of studies in

Social Sciences in Colombia (2006–2015): Analysis of scientific production, co-authorship networks, institutional cooperation and positioning of research groups" code SIGP1800074054754. We also thank our research assistant Laura Rambal Rivaldo for the design of databases and graph drawing.

## Conflicts of interest

The authors do not report conflicts of interest.

## Nomenclature

| | |
|---|---|
| DGM | maximum geodetic distance |
| GDμ | geodetic distance mean |
| SCIENTI-Col | Colombian National System of Science, Technology and Innovation |
| SNA | social networks analysis |

## Author details

José Hernando Ávila-Toscano[1]*, Ivón Catherine Romero-Pérez[2], Ailed Marenco-Escuderos[3] and Eugenio Saavedra Guajardo[4]

*Address all correspondence to: javila@unireformada.edu.co

1 Research and Innovation Management, Corporación Universitaria Reformada, Barranquilla, Colombia

2 Scientology, Mining and Data Analysis, Universidad Simón Bolívar, Barranquilla, Colombia

3 Faculty of Social Sciences, Arts and Humanities, Corporación Universitaria Reformada, Colombia

4 Universidad Católica del Maule, Talca, Chile

## References

[1] Herrera-González J. La formación de docentes investigadores: el estatuto científico de la investigación pedagógica. Magis, Revista Internacional de Investigación en Educación. 2010;**3**:53-62

[2] Ho Y. Classic articles on social work field in social science citation index: A bibliometric analysis. Scientometrics. 2014;**98**:137-155. DOI: 10.1007/s11192-013-1014-8

[3] Ossenblok T, Engels T, Sivertsen G. The representation of the social sciences and humanities in the web of science. A comparison of publication patterns and incentive structures in Flanders and Norway (2005-9). Research Evaluation. 2012;**21**:280-290. DOI: 10.1093/reseval/rvs019

[4] Verleysen F, Engels TCE. Barycenter representation of book publishing internationalization in the social sciences and humanities. Journal of Informetrics. 2014;**8**:234-240. DOI: 10.1016/j.joi.2013.11.008

[5] Henriksen D. What factors are associated with increasing co-authorship in the social sciences? A case study of Danish economics and political science. Scientometrics. 2018;**114**:1395-1421. DOI: 10.1007/s11192-017-2635-0

[6] Shin J, Cummings W. Multilevel analysis of academic publishing across disciplines: Research preference, collaboration, and time on research. Scientometrics. 2010;**85**(2): 581-594. DOI: 10.1007/s11192-010-0236-2

[7] Lancho-Barrantes B, Guerro-Bote V, Moya-Anegon F. Citation increments between collaborating countries. Scientometrics. 2013;**94**:817-831. DOI: 10.1007/s11192-012-0797-3

[8] Low Y, Ng K, Kabir H, Koh M, Sinnasamy J. Trend and impact of international collaboration in clinical medicine papers published in Malaysia. Scientometrics. 2014;**98**:1521-1533. DOI: 10.1007/s11192-013-1121-6

[9] Prathap G. Second order indicators for evaluating international scientific collaboration. Scientometrics. 2013;**95**:563-570. DOI: 10.1007/s11192-012-0804-8

[10] Leydesdorff L, Rafols I. A global map of science based on the isi subject categories. Journal of the American Society for Information Science and Technology. 2009;**60**:348-362. DOI: 10.1002/asi.20967

[11] Silva F, Amancio D, Bardosova M, Costa L, Oliveira O. Using network science and text analytics to produce surveys in a scientific topic. Journal of Informetrics. 2016;**10**:487-502. DOI: 10.1016/j.joi.2016.03.008

[12] Porter A, Rafols I. Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. Scientometrics. 2009;**81**:719-745. DOI: 10.1007/s11192-008-2197-2

[13] Rosvall M, Bergstrom C. Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences of the United States of America. 2008;**105**:1118-1123. DOI: 10.1073

[14] Silva F, Travencolo B, Viana M, Costa L. Identifying the borders of mathematical knowledge. Journal of Physics A: Mathematical and Theoretical. 2010;**43**:425-448. DOI: 10.1088/1751-8113/43/32/325202

[15] Silva F, Viana M, Travencolo B, Costa L. Investigating relationships within and between category networks in wikipedia. Journal of Informetrics. 2011;**5**:431-5.438. DOI: 10.1016/j.joi.2011.03.003

[16] Boyack K, Klavans R, Borner K. Mapping the backbone of science. Scientometrics. 2005;**64**:351-374

[17] Tonta Y, Darvish H. Diffusion of latent semantic analysis as a research tool: A social network analysis approach. Journal of Informetrics. 2010;**4**:166-174. DOI: 10.1016/j.joi.2009.11.003

[18] Carretero-Campos C, Bernaola-Galván P, Coronado A, Carpena P. Improving statistical keyword detection in short texts: Entropic andclustering approaches. Physica A: Statistical Mechanics and Its Applications. 2013;**392**:1481-1492. DOI: 10.1016/j.physa.2012.11.052

[19] Miguel S, Caprile L, Jorquera-Vidal I. Análisis de co-términos y de redes sociales para la generación de mapas temáticos. El profesional de la información. 2008;**17**:637-646. DOI: 10.3145/epi.2008.nov.06

[20] Verd J. El uso de la teoría de las redes sociales en la representación y análisis de textos. De las redes semánticas al análisis de redes textuales. Empiria, Revista de Metodología de las. Ciencias Sociales. 2005;**10**:129-150

[21] Callon M, Courtid J, Ladle F. Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of poly- mer chemistry. Scientometrics. 1991;**22**:155-205

[22] He Q. Knowledge discovery through co-word analysis. Library Trends. 1999;**48**:133-159

[23] Trujillo H, Mañas F, González-Cabrera J. Evaluación de la potencia explicativa de los grafos de redes sociales clandestinas con UciNet y Net-Draw. Universitas Psychologica. 2010;**9**:67-78

[24] Larivière V, Sugimoto C, Cronin B. A bibliometric chronicling of library and information Science's first hundred years. Journal of the American Society for Information Science and Technology. 2012;**63**:997-1016. DOI: 10.1002/asi.22645

[25] Nederhof A. Bibliometric monitoring of research performance in the social sciences and the humanities: A review. Scientometrics. 2006;**66**:81-100. DOI: 10.1007/s11192-006-0007-2

[26] Gingras Y, Mosbah-Natanson S. Where are social sciences produced? In: World Social Science Report. Knowledge Divides. Paris: International Social Science Council and UNESCO; 2010. pp. 149-153

[27] Mosbah-Natanson S, Gingras Y. The globalization of social sciences? Evidence from a quantitative analysis of 30 years of production, collaboration and citations in the social sciences (1980-2009). Current Sociology. 2014;**62**:626-646. DOI: 10.1177/0011392113498866

[28] Huang M, Lin C. A citation analysis of western journals cited in Taiwan's library and information science and history research journals: From a research evaluation perspective. Journal of Academic Librarianship. 2011;**37**:34-45. DOI: 10.1016/j.acalib.2010.10.005

[29] Hu C, Hu J, Gao Y, Zhang Y. A journal co-citation analysis of library and information science in China. Scientometrics. 2011;**86**:657-670. DOI: 10.1007/s11192-010-0313-6

[30] Gantman E. La productividad científica argentina en Ciencias Sociales: Economía, Psicología, Sociología y Ciencia Política en el CONICET (2004-2008). Revista Española de Documentación Científica. 2011;**34**(3):408-425. DOI: 10.3989/redc.2011.3.829

[31] Clauset A, Newman MA, Moore C. Finding community structure in very large networks. Physical Review E. 2004;**70**:1-6

[32] Harel D, Koren YA. Fast multi-scale method for drawing large graphs. Journal of Graph Algorithms and Applications. 2002;**6**:179-202

[33] Borgman C. The digital future is now: A call to action for the humanities. DHQ: Digital Humanities Quarterly. 2009;**3**:1-21. DOI: 10.3989/redc.2011.3.829

Chapter 10

# Data Privacy for Big Data Publishing Using Newly Enhanced PASS Data Mining Mechanism

Priyank Jain, Manasi Gyanchandani and Nilay Khare

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.77033

### Abstract

Anonymization is one of the main techniques that is being used in recent times to prevent privacy breaches on the published data; one such anonymization technique is k-anonymization technique. The anonymization is a parametric anonymization technique used for data anonymization. The aim of the k-anonymization is to generalize the tuples in a way that it cannot be identified using quasi-identifiers. In the past few years, we saw a tremendous growth in data that ultimately led to the concept of the big data. The growth in data made anonymization using conventional processing methods inefficient. To make the anonymization more efficient, we used the proposed PASS mechanism in Hadoop framework to reduce the processing time of anonymization. In this work, we have divided the whole program into the map and reduce part. Moreover, the data types used in Hadoop provide better serialization and transport of data. We performed our experiments on the large dataset. The results proved the best efficiency of our implementation.

**Keywords:** big data, big data privacy and security, data mining, attribute disclosure, PASS, information loss, membership disclosure

## 1. Introduction

### 1.1. What is big data?

“Big data” is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it.

### 1.2. Big data significance in industry and challenges

While understanding the estimation of big information keeps on residual a test, other down to practical challenges including financing and rate of return and aptitudes keep on remaining at the front line for various distinctive ventures that are embracing huge information. All things considered, a Gartner survey for 2015 demonstrates that over 75% of organizations are putting or are intending to put resources into enormous information in the following 2 years. These discoveries speak to a critical increment from a comparable study done in 2012, which showed that 58% of organizations put or were wanting to put resources into enormous information inside the following 2 years [1].

By and large, most associations have a few objectives for receiving enormous information ventures. While the essential objective of most associations is to upgrade client encounter, different objectives incorporate cost diminishment, better focused on promoting and making existing procedures more effective. Lately, information ruptures have additionally made upgraded security a critical objective that has huge information.

### 1.3. Data stream

Big data associated with the time stamp is called big data stream [2].

Examples of data streams:

1. Sensor data
2. Call center records
3. Clickstreams
4. Healthcare data
5. Constraints associated with data streams

**Privacy protection**: i.e., the data streams are extracted from various sources which consist of many individuals' information; hence, the sensitive data of any individuals must not be leaked.

**Computer security**: Access control and verification guarantee that opportune individual has a right expert to the correct protest at the perfect time and the ideal place. That is not what we need here. A general precept of information security is to discharge all the data as much as the personalities of the subjects (individuals) are ensured.

**Real-time processing**: Since the data is not static in nature, real-time processing is required, and at present, not many algorithms are there to process the dynamic data.

### 1.4. What is MapReduce?

MapReduce, as shown in **Figure 1**, is a preparing method and a program that demonstrates for circulated figuring in light of java. The structure deals with every one of the points of interest

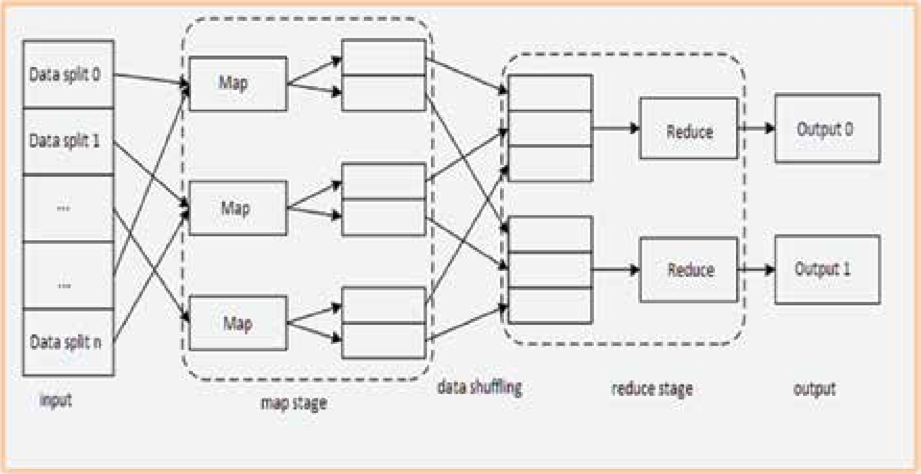


**Figure 1.** Internal working of Map Reduce.

of information passing, for example, issuing errands, confirming assignment culmination, and duplicating information around the group between the hubs [1]:

- Most of the registering happens on hubs with information on neighborhood circles that lessens the system activity.
- After consummation of the given errands, the bunch gathers and diminishes the information to shape a suitable outcome and sends it back to the Hadoop server.

## 2. Anonymization

### 2.1. Anonymization

Generally, the fundamental theme data anonymization [2–5] is the use of one or more techniques designed to make it impossible or at least more difficult to identify a particular individual from stored data related to them. Purposes of data anonymization are:

1. To prevent the privacy of individuals who shared data for various surveys
2. To implement effective techniques to prevent a security breach

It is privacy preservation techniques used for static data. Techniques implemented in anonymization are:

1. Encryption
2. Hashing
3. Generalization

4. Suppression of data
5. Destroy data quality
6. Adding mathematical noise

## 2.2. K-anonymity

A release of data is said to have the *k*-anonymity property if the information for each person contained in the release cannot be distinguished from a least k-1 individuals whose information also appear in the release. For example, if you try to identify a person from a release dataset but you only have information of his/her birth date and gender. There are k people that meet the requirement. This is k-anonymity [6, 7].

### *2.2.1. Classification of attributes*

**Key attribute** is name, address, and cell phone, which can uniquely identify an individual directly. It is always removed before release.

**Quasi-identifier** is a zip code, birth date, and gender, a set of attributes that can be potentially linked with external information to re-identify entities. Eighty-seven percent of the population in the USA can be uniquely identified based on these attributes, according to the census summary data in 1991. There are two tables shown below: **Table 1** is hospital dataset and **Table 2** is voter data.

| DOB | Sex | Zip code | Disease |
|---|---|---|---|
| 1/21/1976 | M | 65715 | Heart disease |
| 4/13/1986 | F | 65715 | Hepatitis |
| 2/28/1976 | M | 65703 | Bronchitis |
| 1/21/1976 | M | 65703 | Broken arm |
| 4/13/1986 | F | 65706 | Flu |
| 2/28/1976 | F | 65706 | Hang nail |

**Table 1.** Medical dataset.

| Name | DOB | Sex | Zip code |
|---|---|---|---|
| Andre | 1/21/1976 | Male | 53715 |
| Beth | 1/10/1981 | Female | 55410 |
| carol | 10/1/1944 | Female | 90210 |
| Dan | 2/21/1984 | Male | 02174 |
| Ellen | 4/19/1972 | Female | 02237 |

**Table 2.** Voter dataset.

From above tables, we can conclude that Andre has heart disease; here the heart disease is the sensitive attribute. It is known as linking attack by combining two different tables. The solution is to consider all of the released tables before releasing the new one and trying to avoid linking. And k-anonymity does not provide privacy if sensitive values in an equivalence class lack diversity [8, 9].

## 3. Related work

### 3.1. FANNST algorithm

#### *3.1.1. Algorithm*

When the numbers of tuples in the processing window reach μ, one round of the clustering algorithm is started to slide again in order to accumulate more tuples in each round [10].

Parameters used in the algorithm are k, u, d:

k defines the parameter for cluster anonymization.

d defines the number of clusters which can be used later.

u defines the processing window size.

#### *3.1.2. Drawback*

The main drawback of FANNST is that some tuples may remain in the system for more than allowable time constraint. In addition, the time and space complexity of the algorithm is O(S*S) and not efficient for a data streaming algorithm. Another weakness of FANNST is that it does not support categorical data.

### 3.2. FADS algorithm

The algorithm considers a set as a buffer and saves at most δ tuples in it [11, 12]. Also, another set (setkc) is considered to hold the newly created cluster for later reuse. Each k-anonymized cluster will be remained in setkc up to the reuse constraint Tkc, and after that, the cluster is removed.

#### *3.2.1. Drawbacks*

The main drawback of the FADS is that the algorithm does not check the remaining time of tuples that hold in the buffer in each round and give their result when they might be considered to have expired. The other important weakness of FADS is that it is not parallel and cannot handle a large number of data streams in tolerable time.

# 4. Terminology of proposed algorithm

## 4.1. Data stream

A sequence of tuples is defined as <sn>n∈N where N is the natural number set. The kth term of <sn> is order pair (t, tk) where k is a number and tk is a tuple.

A data stream S is a potentially infinite sequence of tuples, depicted by $<t_i>$, where all tuples $t_i$ follow the schema $t_i$ = $<ID,a_1,a_m,q_1,q_n,TS>$. ID is an identifier attribute; $q_1$ to $q_n$ are quasi-identifiers, and TS is the time stamp.

## 4.2. Cluster

The cluster is a set of tuples in a stream [12]. Suppose that PS is a set of tuples in stream cluster C which can be defined as follow:

C = {t | t belongsPs}

### *4.2.1. K-anonymized cluster*

If a cluster C is built from the data stream and the number of the unique tuple in the cluster is greater than k, the cluster is called a k-anonymized cluster.

### *4.2.2. Generalization*

Generalization is a function that maps a cluster into a tuple. More formally, generalization function G is defined as G: PowerSet(TUPLE) → TUPLE where TUPLE is the set of all possible tuples.

### *4.2.3. Numerical value generalization*

Numerical values are generalized in between maximum and minimum value, i.e., they are generalized in their domain.

### *4.2.4. Categorical value generalization*

Categorical values are generalized to their lowest common ancestors as shown in **Figure 2**.

### *4.2.5. Example of above two types of generalization*

Considering a cluster of three tuples which contains both numerical and categorical values, the tuples contain the name, profession, and age of employees.

C = <"prof.young", Academic, 43>,

<"Mr.Zhou", non-Academic, 39>,

<"Prof.Chung", Academic, 46>.

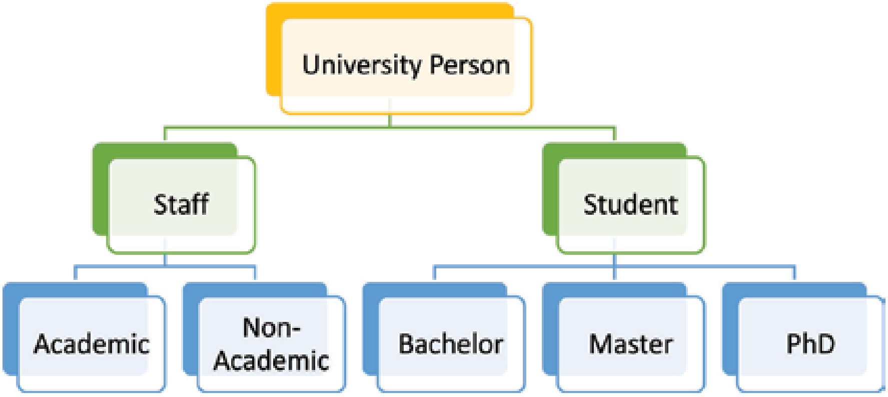


**Figure 2.** University taxonomy tree.

The above tuple can be generalized as follows: gc = <∗,staff,[39–46]>. Since we do not want to disclose the name, we kept * in the first column; here profession is categorical value, and age is numerical value; age is generalized as [max, min], and profession is generalized to lowest common ancestor of academic and nonacademic.

### *4.2.6. Distance*

Distance is used to calculate the similarity or dissimilarity between two tuples. This function is the heart of the clustering. Generally, clustering is done based on distance calculation; the tuples with the closest distance are placed the same cluster.

#### *4.2.6.1. Types of distances*

##### *4.2.6.1.1. The distance between the numerical values*

Let $v_1$,$v_2$ be 2 numerical values.

**The distance between $v_1$,$v_2$ = d($v_1$,$v_2$) = $|v_1 - v_2|/|D|$**

where D is the domain of the values.

##### *4.2.6.1.2. The distance between two categorical values*

If all the categorical values are arranged in the form of a tree where the root is the most generalized value of all the values and lowest most level containing more specialized values of the categorical values, e.g., of a categorical tree as shown in **Figure 3** Country taxonomy tree and **Figure 4** Occupation taxonomy tree.

**Distance between two categorical values $v_1$,$v_2$ = d($v_1$,$v_2$) = (height of the subtree roots at lowest common ancestor of ($v_1$,$v_2$))/(height of tree):**

For example, distance between India and Egypt (considering the tree from the above picture).

=Height of subrooted tree of a lowest common ancestor of India and Egypt/height of the tree.

=Height of the tree with east as root/height of tree = 1/3 = 0.33.

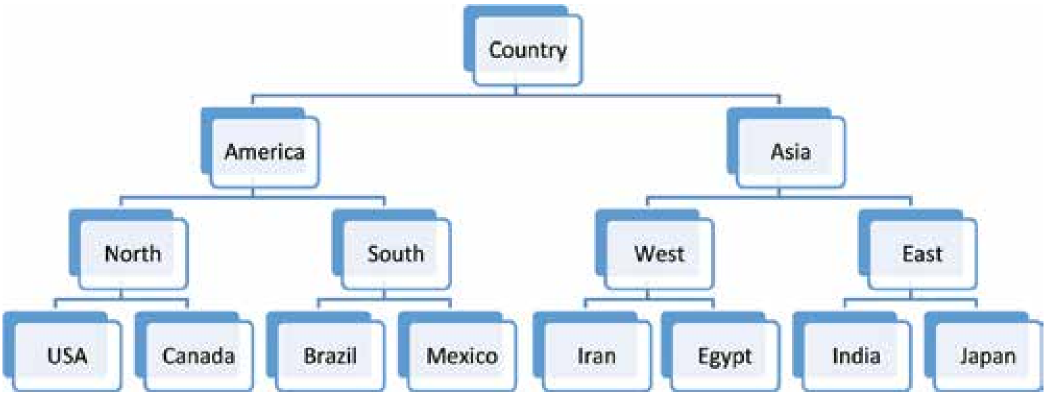


**Figure 3.** Country taxonomy tree.

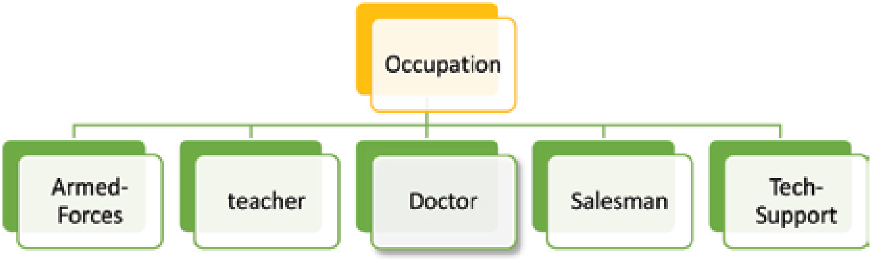


**Figure 4.** Occupation taxonomy tree.

*4.2.6.1.3. The distance between two tuples*

Distance between two tuples t = {N1,…,Nm, C1,…,Cn} is the quasi-identifier of table T, where $N_i$ (i = 1,…,m) is an attribute with a numeric domain and Cj(j = 1,…,n) is an attribute with a categorical domain.

The distance d(r1,r2) (i.e., the distance between two tuples r1, r2) is defined as:

d(r1,r2) = **sum of distances between numerical attributes of two tuples + sum of distances between categorical attributes of two tuples**.

**Information loss**: generalization leads to information loss, but we have to group clusters in such a way that the information loss is minimum.

Information loss of a single cluster is calculated as:

**Total information loss** = sum of information loss of all the clusters.

**Information loss of the cluster** = info loss of all the tuples in the cluster.

**Information loss of the tuple** = information loss of all the attributes (categorical attributes and numerical attributes).

**Information loss of numerical attribute** = (value of attribute)/(domain of the attribute).

**Information loss of categorical attribute** = (height of the tree rooted with categorical attribute)/(height of categorical attribute tree) where h is the height of the tree and k is the height of the tree rooted at the required categorical attribute.

## 5. Proposed PASS algorithm

### 5.1. Details of the PASS algorithm

S = total number of tuples in the dataset.

K = anonymization parameter.

$ = number of tuples to be read before processing.

SetTp = set of $ tuples.

SetKc = set of all unique generalized sets.

Snew = set of K tuples.

Gs = generalized set of Snew.

The algorithm reads $ tuples continuously and inserts them into the SetTp. At First, for each tuple in SetTp procedure finds t's K-1 nearest tuples in SetTp, with the help of tuple t and its K-1 nearest tuples, generate a new set called as Snew and generalize it into Gs. Then a set with minimum information loss (Sk-best) that covers tuple t is chosen from SetKc if Sk-best exists and has smaller information loss than Gs; then tuple t is published Sk-best generalization.

If tuple t does not match with any set of SetKc which has less information loss compared to Gs, then tuple t is published with Snew generalization, i.e., Gs. Then Gs is inserted in SetKc.

In the following, a simple example is illustrated for better understanding. **Table 3** is a portion of a university person data stream, in which quasi-identifiers are age and job. Also $ and K are assumed as $ = 3 and K = 2. Suppose that in thread n, the value of variables is as follows:

| Pid | Age | University person |
|---|---|---|
| Id1 | 22 | Bachelor |
| Id2 | 24 | Master |
| Id3 | 37 | Nonacademic |
| . | . | . |
| . | . | . |
| . | . | . |
| Idn | 45 | Academic |
| Idn + 1 | 26 | Nonacademic |
| Idn + 2 | 39 | PhD |

**Table 3.** University person.

| Pid | Age | University person |
|---|---|---|
| Id1 | [22–24] | Student |
| Id2 | [22–24] | Student |
| Id3 | [15–95] | University person |
| . | . | . |
| . | . | . |
| . | . | . |
| Idn | [44–46] | Staff |
| Idn + 1 | [26–39] | University person |
| Idn + 2 | [26–39] | University person |

**Table 4.** Two anonymized university persons.

In this stage, information loss of Sk-best is compared with Gs information loss. As the information loss of Sk-best is less than Gs, a tuple with idn is published with Sk-best generalization. **Table 4** represents Two anonymized university persons.

- SetTp = {(<idn,45,academic>, <idn + 1,26,Non − academic>,<idn + 2,39,PhD>)}
- SetKc = {(([22–24],university), ([31–39],staff),([44–46],staff))}
- Snew = (<idn,45,academic>,<idn + 2,26,non-academic>)
- Gs = ([26–45],staff)
- Sk-best = ([44–46],staff)

### 5.2. Proposed PASS algorithm

```
Big data Anonymization (S,K,$)
{
while S!=0 do
  Read $ tuples and insert them into
  SetTp.
  For each tuple t do
       1. Select K-1 unique tuples
       which are closest to t among
             the tuples in SetTp and insert them into
       set Snew.
       2. Generalize Snew into Gs.
```

```
            **For** each set which covers t **do**
                Calculate the information loss
            **End for**
            3. Select a set which includes less information loss
            Call the set as Sk-best
            4. **If** (Sk-best exist and Sk-best generate less information loss
             Than Gs) **then**
                Publish t with Sk-best generalization
            **Else**
                Publish t with Gs and insert Gs in set Kc
            **End if**
        **End for**
    **End while**
}
```

## 6. Result and discussion

### 6.1. Experiment environment

This experiment is performed on the system having Intel i5 processor with the processing power of 2.2 GHz and main memory of 4.0 GB using Linux platform. The algorithm is implemented in Java and executed with the help of Hadoop MapReduce framework.

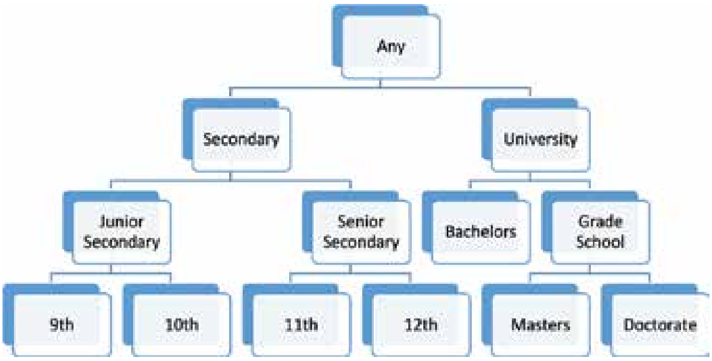


**Figure 5.** Taxonomy tree.

### 6.2. Dataset description

In this experiment, we evaluated the performance of the proposed algorithm on the adult dataset from UCI [13]. The dataset was widely used for the privacy-preserving purpose. The taxonomy tree is defined as per **Figure 5**. The sensitive attribute in the dataset is age (numerical) and profession (categorical).

### 6.3. Results and discussions

The total number of records in the dataset used for the experiment purpose is 32,599 tuples. The efficiency of proposed algorithm is verified by parameter information loss. The average information loss of the proposed PASS algorithm, FADS and FAST, is presented in **Figure 6**. The proposed PASS algorithm publishes data with less information loss, because the SetKc in the proposed approach as shown in **Figure 7** has more entities so that the data tuple has more

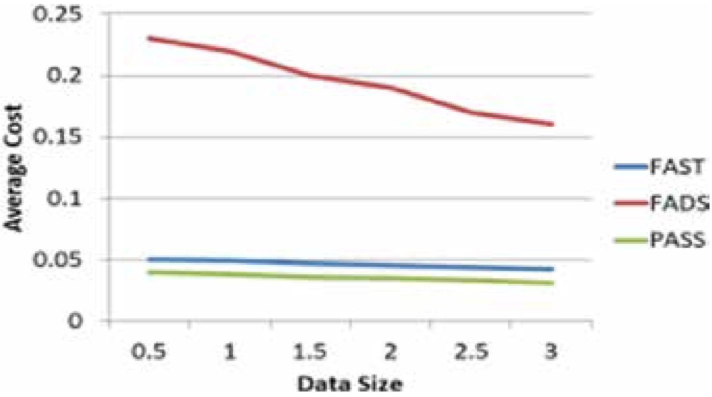


**Figure 6.** Information loss in FAST and FADS algorithms.

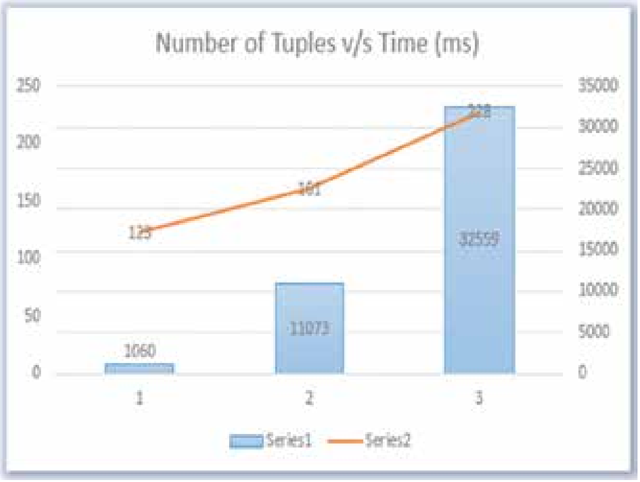


**Figure 7.** Number of tuples vs. running time.

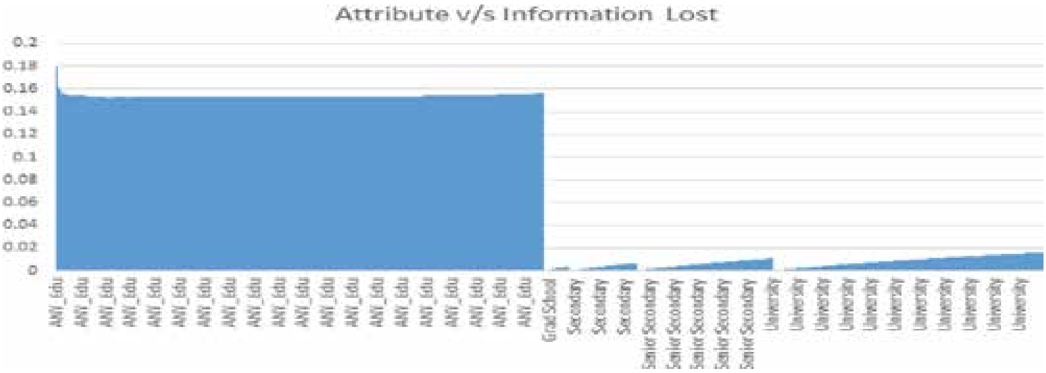


**Figure 8.** Attribute vs. information loss.

options to select, and this decreases the information loss as shown in **Figure 8**, and hence the results of an algorithm show improvement. The average execution time drastically decreases as MapReduce-based newly enhanced PASS mechanism is used.

## 7. Conclusion

All the algorithms which are present for data stream processing are not capable of processing big data, i.e., data with high capacity and volume. The data which is processed using data anonymization (nonparallel) algorithms use old languages (JAVA, SQL) and old techniques, which are not very effective means because they take a lot of time for computation and sometimes provide tuples, which are expired; this lead to loss of accuracy as well as loss of privacy which is very dangerous. Static algorithms need all the computations to be performed on a single node due to which the data and the processing requirements are very high and the computers used are prone to failure which is very expensive to recover.

In this paper, we have proposed PASS algorithm, which uses Hadoop framework to process the data. Using Hadoop, the computer's resources are used to the maximum extent by which time required for computation is reduced which in turn prevents the publishing of expired tuples. Other advantages of this algorithm are that computations can be performed on nodes which have less computation and less storage capacity than that of computers which perform nonparallel data processing. The proposed PASS algorithm publishes data with less information loss. Using Hadoop, the failures in both data and processors can be recovered. These features drastically reduce the maintenance cost and the initial setup cost.

## Author details

Priyank Jain*, Manasi Gyanchandani and Nilay Khare

*Address all correspondence to: priyankjain1984@gmail.com

Department of Computer Science and Engineering, MANIT, Bhopal, India

## References

[1] Aggarwal CC, Han J, Wang J, Yu PS. A framework for clustering evolving data streams. In: Proceedings of the 29th International Conference on Very Large Data Bases-Vol. 29. VLDB Endowment; 2003. pp. 81-92

[2] Cao J, Carminati B, Ferrari E, Tan K-L. Castle: Continuously anonymizing data streams. IEEE Transactions on Dependable and Secure Computing. 2011;**8**(3):337-352

[3] Dwork C. Differential privacy. In: Bugliesi M, Preneel B, Sassone V, Wegener I, editors. ICALP Lecture Notes in Computer Science. Springer, 2006;**4052**(2):1-12

[4] Li F, Sun J, Papadimitriou S, Mihaila GA, Stanoi I. Hiding in the Crowd: Privacy Preservation on Evolving Streams through Correlation Tracking. In: ICDE Istanbul Turkey, 2007. p. 2

[5] Li N, Li T, Venkatasubramanian S. T-Closeness: Privacy beyond K-Anonymity and L-Diversity. In: ICDE Istanbul Turkey, 2007, p. 106-115

[6] Fung B, Wang K, Chen R, Yu PS. Privacy-preserving data publishing: A survey of recent developments. ACM Computing Surveys (CSUR). 2010;**42**(4):14

[7] Kim S, Sung MK, Chung YD. A framework to preserve the privacy of electronic health data streams. Journal of Biomedical Informatics. 2014;**50**:95-110

[8] Fung BC, Wang K, Yu PS. Top-down specialization for information and privacy preservation. In: Proceedings of the 21st International Conference on Data Engineering (ICDE 2005). IEEE; 2005. pp. 205-216

[9] Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. L-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD). March 2007;**1**(1):3. DOI: 10.1145/1217299.1217302

[10] Zakerzadeh H, Osborn SL. FAST: Fast Anonymizing Algorithm for Numerical Streaming Data. Proceedings of the 5th International Workshop on Data Privacy Management, and 3rd International Conference on Autonomous Spontaneous Security. Athens, Greece; September 23, 2010

[11] Guo K, Zhang Q. Fast clustering-based anonymization approaches with time constraints for data streams. Knowledge-Based Systems. 2013;**46**:95-108

[12] Guo K, Zhang Q. Fast clustering-based anonymization approaches with time constraints for data streams. Knowledge-Based Systems. July 2013;**46**:95-108. DOI: 10.1016/j.knosys.2013.03.007

[13] Newman CBD, Merz C. UCI Repository of machine learning databases. Technical report, University of California, Irvine, Department of Information and Computer Sciences. 1998

*Edited by Ciza Thomas*

This book on data mining explores a broad set of ideas and presents some of the state-of-the-art research in this field. The book is triggered by pervasive applications that retrieve knowledge from real-world big data. Data mining finds applications in the entire spectrum of science and technology including basic sciences to life sciences and medicine, to social, economic, and cognitive sciences, to engineering and computers. The chapters discuss various applications and research frontiers in data mining with algorithms and implementation details for use in real-world. This can be through characterization, classification, discrimination, anomaly detection, association, clustering, trend or evolution prediction, etc. The intended audience of this book will mainly consist of researchers, research students, practitioners, data analysts, and business professionals who seek information on the various data mining techniques and their applications.

Published in London, UK
© 2018 IntechOpen
© LukinBrno / iStock

IntechOpen

ISBN 978-1-83881-567-7
9 781838 815677