Chapter 1

Survey of Data Mining and Applications (Review from 1996 to Now)

Adem Karahoca, Dilek Karahoca and Mert Şanver

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/48803

1. Introduction

The science of extracting useful information from large data sets or databases is named as data mining. Though data mining concepts have an extensive history, the term "Data Mining", is introduced relatively new, in mid 90's. Data mining covers areas of statistics, machine learning, data management and databases, pattern recognition, artificial intelligence, and other areas. All of these are concerned with certain aspects of data analysis, so they have much in common but each also has its own distinct problems and types of solution. The fundamental motivation behind data mining is autonomously extracting useful information or knowledge from large data stores or sets. The goal of building computer systems that can adapt to special situations and learn from their experience has attracted researchers from many fields, including computer science, engineering, mathematics, physics, neuroscience and cognitive science.

As opposed to most of statistics, data mining typically deals with data that have already been collected for some purpose other than the data mining analysis. Majority of the applications presented in this book chapter uses data formerly collected for any other purposes. Out of data mining research, has come a wide variety of learning techniques that have the potential to renovate many scientific and industrial fields.

This book chapter surveys the development of Data Mining through review and classification of journal articles between years 1996-now. The basis for choosing this period is that, the comparatively new concept of data mining become widely accepted and used during that period. The literature survey is based on keyword search through online journal databases on Science Direct, EBSCO, IEEE, Taylor Francis, Thomson Gale, and Scopus. A total of 1218 articles are reviewed and 174 of them found to be including data mining methodologies as primary method used. Some of the articles include more than one data mining methodologies used in conjunction with each other.

The concept of data mining can be divided into two broad areas as predictive methods and descriptive methods. Predictive methods include Classification, Regression, and Time Series Analysis. Predictive methods aim to project future status before they occur.

Section 2 includes definition of algorithms and the applications using these algorithms. Discussion of trends throughout the last decade is also presented in this section. Section 3 introduces Descriptive methods in four major parts; Clustering, Summarization, Association Rules and Sequence Discovery. The objective of descriptive methods is describing phenomena, evaluating characteristics of the dataset or summarizing a series of data. The application areas of each algorithm are documented in this part with discussion of the trend in descriptive methods. Section 4 describes data warehouses and lists their applications involving data mining techniques. Section 5 gives a summarization of the study and discusses future trends in data mining and contains a brief conclusion.

2. Predictive methods and applications

A predictive model makes a prediction about values of data using known results found from different data sets. Predictive modeling may be made based on the use of other historical data. Predictive model data mining tasks include classification, regression, time series analysis, and prediction (Dunham, 2003).

2.1. Classification methods

Classification maps data into predefined groups or classes. It is often referred to as supervised learning. Classification algorithms require that the classes be defined based on data attribute values. Pattern recognition is a type of classification where an input pattern is classified into one of several classes based on its similarity to these predefined classes (Dunham, 2003). In this section; decision trees, neural networks, Bayesian classifiers and support vector machines related applications are considered.

2.1.1. Decision trees

Decision trees can be construct recursively. Firstly, an attribute is selected to place at root node to make one branch for each possible value. This splits up the example set into subsets, one for every value of the attribute (Witten, Frank; 2000).

The basic principle of tree models is to partition the space spanned by the input variables to maximize a score of class purity that the majority of points in each cell of the partition belong to one class. They are mappings of observations to conclusions (target values). Each inner node corresponds to variable; an arc to a child represents a possible value of that variable. A leaf represents the predicted value of target variable given the values of the variables represented by the path from the root (T. Menzies, Y. Hu, 2003).

Information entropy is used to measure the amount of uncertainty or randomness in a set of data. Gini index also used to determine the best splitting for a decision tree.

Decision trees can be divided into two types as regression trees and classification trees. The trend is towards the regression trees as they provide real valued functions instead of classification tasks. Applications include; Remote Sensing, Database Theory, Chemical engineering, Mobile communications, Image processing, Soil map modeling, Radiology, Web traffic prediction, Speech Recognition, Risk assessment, Geo information, Operations Research, Agriculture, Computer Organization, Marketing, Geographical Information Systems. Decision trees are growing more popular among other methods of classifying data. C5.0 algorithm by R.J. Quinlan is very commonly used in latest applications.

Decision tree applications	Authors
2006 – Geographical Information Systems	Baisen Zhang, Ian Valentine, Peter Kemp and Greg Lambert
2005 – Marketing	Sven F. Crone, Stefan Lessmann and Robert Stahlbock
2005 – Computer Organization	Xiao-Bai Li
2005 - Agriculture	Baisen Zhang, Ian Valentine and Peter D. Kemp
2004 – Operations Research	Nabil Belacel, Hiral Bhasker Raval and Abraham P. Punnen
2004 – Geoinformation	Luis M. T. de Carvalho, Jan G. P. W. Clevers, Andrew K. Skidmore
2004 – Risk assessment	Christophe Mues, Bart Baesens, Craig M. Files and Jan Vanthienen
2003 – Speech Recognition	Oudeyer Pierre-Yves
2003 – Web traffic prediction	Selwyn Piramuthu
2002 – Radiology	Wen-Jia Kuo, Ruey-Feng Chang, Woo Kyung Moon, Cheng Chun Lee
2002 – Soil map modelling	Christopher J. Moran and Elisabeth N. Bui
2002 – Image processing	Petra Perner
2001 – Mobile communications	Patrick Piras, Christian Roussel and Johanna Pierrot-Sanders
2000 – Chemical engineering	Yoshiyuki Yamashita
2000 – Geoscience	Simard, M.; Saatchi, S.S.; De Grandi
2000 – Medical Systems	Zorman, M.; Podgorelec, V.; Kokol, P.; Peterson, M.; Lane, J
1999 – Database Theory	Mauro Sérgio R. de Sousa, Marta Mattoso and Nelson F. F. Ebecken
1999 – Speech Processing	Padmanabhan, M.; Bahl, L.R.; Nahamoo, D
1998 – Remote Sensing	R. S. De Fries M. Hansen J. R. G. Townshend R. Sohlberg

Table 1. Decision Tree Applications

2.1.2. Neural networks

An artificial neural network is an interconnected group of artificial neurons that uses a mathematical or computational model for information processing based on a connectionist approach to computation (Freeman et al., 1991). Formally the field started when neurophysiologist Warren McCulloch and mathematician Walter Pitts wrote a paper on how neurons might work in 1943. They modeled a simple neural network using electrical circuits. In 1949, Donald Hebb pointed out the fact that neural pathways are strengthened each time they are used, a concept fundamentally essential to the ways in which humans learn. If two nerves fire at the same time, he argued, the connection between them is enhanced.

In 1982, interest in the field was renewed. John Hopfield of Caltech presented a paper to the National Academy of Sciences. His approach was to create more useful machines by using bidirectional lines. In 1986, with multiple layered neural networks appeared, the problem was how to extend the Widrow-Hoff rule to multiple layers. Three independent groups of researchers, one of which included David Rumelhart, a former member of Stanford's psychology department, came up with similar ideas which are now called back propagation networks because it distributes pattern recognition errors throughout the network. Hybrid networks used just two layers, these back-propagation networks use many. Neural networks are applied to data mining in Craven and Sahvlik (1997).

Neural Networks Applications	Authors
2006 – Banking	Tian-Shyug Lee, Chih-Chou Chiu, Yu-Chao Chou and Chi-Jie Lu
2005 – Stock market	J.V. Healy, M. Dixon, B.J. Read and F.F. Cai
2005 – Financial Forecast	Kyoung-jae Kim
2005 – Mobile Communications	Shin-Yuan Hung, David C. Yen and Hsiu-Yu Wang
2005 – Oncology	Ta-Cheng Chen and Tung-Chou Hsu
2005 – Credit risk assessment	Yueh-Min Huang, Chun-Min Hung and Hewijin Christine Jiau
2005 – Enviromental Modelling	Uwe Schlink, Olf Herbarth, Matthias Richter, Stephen Dorling
2005 – Cybernetics	Jiang Chang; Yan Peng
2004 – Biometrics	Marie-Noëlle Pons, Sébastien Le Bonté and Olivier Potier
2004 – Heat Transfer Engineering	R. S. De Frıes M. Hansen J. R. G. Townshend R. Sohlberg
2004 – Marketing	YongSeog Kim and W. Nick Street
2004 – Industrial Processes	X. Shi, P. Schillings, D. Boyd
2004 – Economics	Tae Yoon Kim, Kyong Joo Oh, Insuk Sohn and Changha Hwang
2003 – Crime analysis	Giles C. Oatley and Brian W. Ewart
2003 – Medicine	Álvaro Silva, Paulo Cortez, Manuel Filipe Santos, Lopes Gomes and José Neves
2003 – Production economy	Paul F. Schikora and Michael R. Godfrey
2001 – Image Recognation	Kondo, T.; Pandya, A.S

Table 2. Neural Networks Applications

The research in theory has been slowed down; however applications continue to increase popularity. Artificial neural networks are one of a class of highly parameterized statistical models that have attracted considerable attention in recent years. Since the artificial neural networks are highly parameterized, they can easily model small irregularities in functions however this may lead to over fitting in some conditions. Applications of neural networks include; Production economy, Medicine, Crime analysis, Economics, Industrial Processes, Marketing, Heat Transfer Engineering, Biometrics, Environmental Modeling, Credit risk assessment, Oncology, Mobile Communications, Financial Forecast, Stock market, Banking.

2.1.3. Bayesian classifiers

Bayesian classification is based on Bayes Theorem. In particular, naive Bayes is a special case of a Bayesian network, and learning the structure and parameters of an unrestricted Bayesian network would appear to be a logical means of improvement.

However, Friedman (1997) found that naive Bayes easily outperforms such unrestricted Bayesian network classifiers on a large sample of benchmark datasets. Bayesian classifiers are useful in predicting the probability that a sample belongs to a particular class or grouping. This technique tends to be highly accurate and fast, making it useful on large databases. Model is simple and intuitive. Error level is low when independence of attributes and distribution model is robust. Some often perceived disadvantages of Bayesian analysis are really not problems in practice. Any ambiguities in choosing a prior are generally not serious, since the various possible convenient priors usually do not disagree strongly within the regions of interest. Bayesian analysis is not limited to what is traditionally considered statistical data, but can be applied to any space of models (Hanson, 1996).

Application areas include; Geographical Information Systems, Database Management, Web services, Neuroscience. In application areas which large amount of data needed to be processed, technique is useful. The assumption of normal distribution of patterns is the toughest shortcoming of the model.

Bayessian Classifiers	Authors
2005 – Neuroscience	Pablo Valenti, Enrique Cazamajou, Marcelo Scarpettini
2003 – Web services	Dunja Mladeni and Marko Grobelnik
1999 – Database Management	S. Lavington, N. Dewhurst, E. Wilkins and A. Freitas
1998 – Geographical Information Systems	A. Stassopoulou, M. Petrou J. Kıttler

Table 3. Bayesian Classifiers

2.1.4. Support Vector Machines

Support Vector Machines are a method for creating functions from a set of labeled training data. The original optimal hyper plane algorithm proposed by Vladimir Vapnik in 1963 was a linear classifier. However, in 1992, Boser, Guyon and Vapnik suggested a way to create non-linear classifiers by applying the kernel trick to maximum-margin hyper planes. The resulting algorithm is formally similar, except that every dot product is replaced by a non-linear kernel function. This allows the algorithm to fit the maximum-margin hyper plane in the transformed feature space. The transformation may be non-linear and the transformed space high dimensional; thus though the classifier is a hyper plane in the high-dimensional feature space it may be non-linear in the original input space.

In 1995, Cortes and Vapnik suggested a modified maximum margin idea that allows for mislabeled examples. If there exists no hyper plane that can split the binary examples, the Soft Margin method will choose a hyper plane that splits the examples as cleanly as possible, while still maximizing the distance to the nearest cleanly split examples.

A version of a SVM for regression was proposed in 1997 by Vapnik, Golowich, and Smola. This method is called SVM regression. The model produced by classification only depends on a subset of the training data, because the cost function for building the model does not care about training points that lie beyond the margin. Analogously, the model produced by SVR only depends on a subset of the training data, because the cost function for building the model ignores any training data that is close (within a threshold ε) to the model prediction. The function can be a classification function or the function can be a general regression function. A detailed tutorial can be found in Burges (1998).

For classification they operate by finding a hyper surface in the space of possible inputs. Applications of Support Vector Machines include; Industrial Engineering, Medical Informatics, Genetics, Medicine, Marketing.

Support Vector Mechanism	Authors
2005 – Marketing	Sven F. Crone, Stefan Lessmann and Robert Stahlbock
2004 – Medicine	Lihua Li, Hong Tang, Zuobao Wu, Jianli Gong, Michael Gruidl
2004 – Genetics	Fei Pan, Baoying Wang, Xin Hu and William Perrizo
2003 – Medical Informatics	I Kalatzis, D Pappas, N Piliouras, D Cavouras
2002 – Industrial Engineering	Mehmed Kantardzic, Benjamin Djulbegovic and Hazem Hamdan

Table 4. Support Vector Machines

2.2. Time series analysis and prediction applications

Time series clustering has been shown effective in providing useful information in various domains. There seems to be an increased interest in time series clustering as part of the effort in temporal data mining research (Liao, 2003). Unvariate and multivariate time series explained respectively.

2.2.1. Univariate time series

The expression "univariate time series" refers to a time series that consists of particular observations recorded sequentially over equal time increments. Although a univariate time series data set is usually given as a single column of numbers, time is in fact an implicit variable in the time series. If the data are equi-spaced, the time variable, or index, does not need to be explicitly given. The time variable may sometimes be explicitly used for plotting the series. However, it is not used in the time series model itself. Triple exponential smoothing is an example of this approach. Another example, called seasonal loses, is based on locally weighted least squares and is discussed by Cleveland (1993). Another approach, commonly used in scientific and engineering applications, is to analyze the series in the frequency domain. The spectral plot is the primary tool for the frequency analysis of time series. Application areas includes financial forecasting, management, energy, economics, zoology, industrial engineering, emergency services, biomedicine, networks.

Univariate Time Series Applications	Authors
2005 – Financial Forecasting	James W. Taylor and Roberto Buizza
2005 – Enviromental Management	Peter Romilly
2004 – Energy	Jesús Crespo Cuaresma, Jaroslava Hlouskova, Stephan Kossmeier
2003 – Crime Rates Forecasting	Wilpen Gorr, Andreas Olligschlaeger and Yvonne Thompson
2002 – Financial Economics	Per Bjarte Solibakke
2001 – Unemployment Rates	Bradley T. Ewing and Phanindra V. Wunnava
2001 – Forecasting	Juha Junttila
2000 – Zoology	Christian H. Reick and Bernd Page
1998 – Industrial Engineering	Gerhard Thury and Stephen F. Witt
1998 – Emergency Medicine	Kenneth E Bizovi, Jerrold B Leikin, Daniel O Hryhorczuk and Lawrence J Frateschi
1998 – Biomedicine	R. E. Abdel-Aal and A. M. Mangoud
1997 – Economics	Hahn Shik Lee and Pierre L. Siklos
1996 – Sensors	Stefanos Manganaris
1996 – Economics	Apostolos Serletis and David Krause

Table 5. Univariate Time Series Applications

2.2.2. Multivariate time series

Multivariate time series may arise in a number of ways. The time series are measuring the same quantity or time series depending on some fundamental quantity leads to multivariate series. The multivariate form of the Box-Jenkins univariate models is frequently used in applications. The multivariate form of the Box-Jenkins univariate models is sometimes called the ARMAV model, for AutoRegressive Moving Average Vector or simply vector ARMA process. Also, Friedman worked multivariate adaptive regression splines in 1991.

The application areas of the method include neurology, hydrology, finance, medicine, chemistry, environmental science, biology.

Multivariate Time Series Applications	Authors
2006 – Neurology	Björn Schelter, Matthias Winterhalder, Bernhard Hellwig, Brigitte Guschlbauer
2005 – Neurobiology	Ernesto Pereda, Rodrigo Quian Quiroga and Joydeep Bhattacharya
2005 – Hydrology	R. Muñoz-Carpena, A. Ritter and Y.C. Li
2005 – Neuroscience	Andy Müller, Hannes Osterhage, Robert Sowa
2004 – Market analysis	Bernd Vindevogel, Dirk Van den Poel and Geert Wets
2004 – Policy Modelling	Wankeun Oh and Kihoon Lee
2004 – Medicine	Fumikazu Miwakeichi, Andreas Galka, Sunao Uchida, Hiroshi Arakaki
2004 – Economics	Morten Ørregaard Nielsen
2003 – Labor Force Forecasting	Edward W. Frees
2003 – Statistical Planning	Hamparsum Bozdogan and Peter Bearse
2002 – Chemistry	Jan H. Christensen
2002 – Biomedicine	Stephen Swift and Xiaohui Liu
2001 – Marine Sciences	Ransom A. Myers
2001 – Reliability Engineering	S. Lu, H. Lu and W. J. Kolarik
2000 – Environmental Science	Zuotao Li and Menas Kafatos

Table 6. Multivariate Time Series Applications

2.3. Regression methods

Regression is generally used to predict future values base on past values by fitting a set of points to a curve. Linear regression assumes that a linear relationship exists between the input data and the output data. The common formula for a linear relationship;

$$y= c_{0}+c_{1x1}+\ldots\ldots+c_{nxn} \tag{1}$$

Here there are n input variables, that are called predictors or regressors; one output variable, that is called response and n + 1 constants which are chosen during the modeling process to match the input examples. This is sometimes called multiple linear regression because there is more than one predictor (Dunham, 2003).

Following subsections give explanations about non parametric, robust, ridge and nonlinear regressions.

2.3.1. Nonparametric regression

Nonparametric regression analysis is regression without an assumption of linearity. The scope of nonparametric regression is very broad, ranging from smoothing the relationship between two variables in a scatter plot to multiple-regression analysis and generalized regression models. Methods of nonparametric-regression analysis have been rendered practical by advances in statistics and computing, and are now a serious alternative to more traditional parametric-regression modeling. Non-parametric regression is a type of regression analysis in which the functional form of the relationship between the response variable and the associated predictor variables does not to be specified in order to fit a model to a set of data. The applications are mostly in fields of medicine and biology. Also applications in economics and geography exist.

Non parametric Regression Applications	Authors
2005 – Medicine	Hiroyuki Watanabe and Hiroyasu Miyazaki
2005 – Veterinery	A.B. Lawson and H. Zhou
2004 – Economics	Insik Min and Inchul Kim
2004 – Geography	Caroline Rinaldi and Theodore M. Cole, III
2004 – Biosystems	Sunyong Kim, Seiya Imoto and Satoru Miyano
2003 – Surgery	David Wypij, Jane W. Newburger, Leonard A. Rappaport
2002 – Environmental Science	Ronald C. Henry, Yu-Shuo Chang and Clifford H. Spiegelman
2001 – Econometrics	Pedro L. Gozalo and Oliver B. Linton
1999 – Econometrics	Yoon-Jae Whang and Oliver Linton

Table 7. Non parametric Regression Applications

2.3.2. Robust regression

Robust regression in another approach, used to set a fitting criterion which is not vulnerable as other regression methods like linear regression. Robust regression analysis provides an alternative to a least squares regression model when fundamental assumptions are unfulfilled by the nature of the data (Yaffee, 2002) . When the analyst estimates his statistical regression models and tests his assumptions, he frequently finds that the assumptions are substantially violated. Sometimes the analyst can transform his variables to conform to those assumptions. Often, however, a transformation will not eliminate or attenuate the leverage of influential outliers that bias the prediction and distort the significance of parameter estimates. Under these circumstances, robust regression that is resistant to the influence of outliers may be the only reasonable resource. The most common method is M-estimation introduced by Huber in 1964. Application areas varies as epidemiology, remote sensing, bio systems, oceanology, computer vision and chemistry.

Non parametric Regression Applications	Authors
2005 – Epidemiology	Andy H. Lee, Michael Gracey, Kui Wang and Kelvin K.W. Yau
2005 – Remote Sensing	Ian Olthof, Darren Pouliot, Richard Fernandes and Rasim Latifovic
2004 – Biosystems	Federico Hahn
2002 – Oceanology	C. Waelbroeck, L. Labeyrie, E. Michel, J. C. Duplessy, J. F. McManus
2001 – Policy Modeling	Bradley J. Bowland and John C. Beghin
1999 – Biochemistry	V. Diez, P. A. García and F. Fdz-Polanco
1998 – Computer vision	Menashe Soffer and Nahum Kiryati
1997 – Chemistry	Dragan A. Cirovic

Table 8. Non parametric Regression Applications

2.3.3. *Ridge Regression*

Ridge regression, also known as Tikhonoy regularization, is the most commonly used method of regularization of ill-posed problems. A frequent obstacle is that several of the explanatory variables will vary in rather similar ways. As result, their collective power of explanation is considerably less than the sum of their individual powers. The phenomenon is known as near collinearity. Data mining application areas are frequently related with chemistry and chemometrics. Also applications in organizational studies and environmental science are listed in table.

Ridge regression Applications	Authors
2005 – Atmospheric Environment	Steven Roberts and Michael Martin
2005 – Epidemology	L.M. Grosso, E.W. Triche, K. Belanger, N.L. Benowitz
2004 – Chemical Engineering	Jeffrey Dean Kelly
2002 – Chemistry	Marla L. Frank, Matthew D. Fulkerson, Bruce R. Patton and Prabir K. Dutta
2002 – Chemometrics	J. Huang, D. Brennan, L. Sattler, J. Alderman
2001 – Laboratory Chemometrics	Kwang-Su Park, Hyeseon Lee, Chi-Hyuck Jun, Kwang-Hyun Park, Jae-Won Jung
2000 – Food Industry	Rolf Sundberg
1996 – Organizational Behaviour	R. James Holzworth

Table 9. Ridge Regression Applications

2.3.4. *Nonlinear regression*

Almost any function that can be written in closed form can be incorporated in a nonlinear regression model. Unlike linear regression, there are very few limitations on the way parameters can be used in the functional part of a nonlinear regression model. Nonlinear least squares regression extends linear least squares regression for use with a much larger and more general class of functions. Almost any function that can be written in closed form can be incorporated in a nonlinear regression model. Unlike linear regression, there are very few limitations on the way parameters can be used in the functional part of a nonlinear regression model. The way in which the unknown parameters in the function are estimated, however, is conceptually the same as it is in linear least squares regression. Application areas include Chromatography, urology, ecology and chemistry.

Nonlinear Regressin Applications	Authors
2005 – Chromatography	Fabrice Gritti and Georges Guiochon
2005 – Urology	Alexander M. Truskinovsky, Alan W. Partin and Martin H. Kroll
2005 – Ecology	Yonghe Wang, Frédéric Raulier and Chhun-Huor Ung
2005 – Soil Research	M. Mohanty, D.K. Painuli, A.K. Misra, K.K. Bandyopadhyaya and P.K. Ghosh
2005 – Chemical Engineering	Vadim Mamleev and Serge Bourbigot
2004 – Dental Materials	Paul H. DeHoff and Kenneth J. Anusavice
2004 – Metabolism Studies	Lars Erichsen, Olorunsola F. Agbaje, Stephen D. Luzio, David R. Owens
2004 – Biology	David D'Haese, Karine Vandermeiren, Roland Julien Caubergs, Yves Guisez
2003 – Hydrology	Xunhong Chen and Xi Chen
2003 – Chemo metrics	Igor G. Zenkevich and Balázs Kránicz
2003 – Production Economics	Paul F. Schikora and Michael R. Godfrey
2002 – Quality Management	Shueh-Chin Ting and Cheng-Nan Chen
2001 – Agriculture	Eva Falge, Dennis Baldocchi, Richard Olson, Peter Anthoni
2000 – Medicine	Marya G. Zlatnik, John A. Copland
1999 – Pharmacology	Johan L. Gabrielsson and Daniel L. Weiner

Table 10. Nonlinear Regression Applications

3. Descriptive methods and applications

The goal of a descriptive model is describe all of the data (or the process generating the data). Examples of such descriptions include models for the overall probability distribution of the data (density estimation), partitioning of the p-dimensional space into groups (cluster analysis and segmentation), and models describing the relationship between variables (dependency modeling). In segmentation analysis, for example, the aim is to group together similar records, as in market segmentation of commercial databases (Hand, et al., 2001).

3.1. Clustering methods and its applications

Clustering is similar to classification except that the groups are not predefined, but rather defined by the data alone. Clustering is alternatively referred to as unsupervised learning or segmentation. It can be thought of as partitioning or segmenting the data into groups that might or might not be disjointed, clustering is usually accomplished by determining the similarity among the data on predefined attributes (Dunham, 2003).

3.1.1. K-means clustering

The k-means algorithm (MacQueen, 1967) is an algorithm to cluster objects based on attributes into k partitions. It is a variant of the expectation-maximization algorithm in which the goal is to determine the k means of data generated from Gaussian distributions. K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids shoud be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early groupage is done. At this point we need to re-calculate k new centroids as bar centers of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more. It assumes that the object attributes form a vector space. Application areas include sensor networks, web technologies, cybernetics.

K-means Clustering Applications	Authors
2005 – E-commerce	R. J. Kuo, J. L. Liao and C. Tu
2005 – Text clustering	Shi Zhong
2005 – Peer to peer data streams	Sanghamitra Bandyopadhyay, Chris Giannella, Ujjwal Maulik, Hillol Kargupta
2005 – Bioscience	Wei Zhong; Altun, G.; Harrison, R
2004 – Image Processing	Mantao Xu; Franti, P
2003 – Cybernetics	Yu-Fang Zhang; Jia-Li Mao
2000- Adaptive Web	Mike Perkowitz and Oren Etzioni

Table 11. K-means Clustering Applications

3.1.2. Fuzzy c-means clustering

Fuzzy c-means is a method of clustering which allows one piece of data to belong to two or more clusters. This method (developed by Dunn in 1973 and improved by Bezdek in 1981) is frequently used in pattern recognition. It is based on minimization of the following objective function. The method is frequently used in pattern recognition. Application areas include ergonomics, acoustics, and manufacturing. Widely used in image processing.

C-means clustering Applications	Authors
2006 – Ergonomics	Stéphane Armand, Eric Watelain, Moïse Mercier, Ghislaine Lensel
2005 – Neurocomputing	Antonino Staiano, Roberto Tagliaferri and Witold Pedrycz
2004 – Acoustics	Nitanda, N.; Haseyama, M.; Kitajima, H
2001 – Manufacturing	Y. M. Sebzalli and X. Z. Wang
2000 – Image Processing	Rezaee, M.R.; van der Zwet, P.M.J.; Lelieveldt, B.P.E.; van der Geest, R.J.; Reiber, J.H.C
2000 – Signal Processing	Zhe-Ming Lu; Jeng-Shyang Pan; Sheng-He Sun
2000 – Remote Sensing	Chumsamrong, W.; Thitimajshima, P.; Rangsanseri, Y
1999 – Machine Vision	Gil, M.; Sarabia, E.G.; Llata, J.R.; Oria, J.P
1998 – Bioelectronics	Da-Chuan Cheng; Kuo-Sheng Cheng

Table 12. C-means Clustering Applications

3.2. Summarization

Summarization involves methods for finding a compact description for a subset of data. A simple example would be tabulating the mean and standard deviations for all fields (Bao, 2000). More sophisticated methods involve the derivation of summary rules, multivariate visualization techniques, and the discovery of functional relationships between variables. Summarization techniques are often applied to interactive exploratory data analysis and automated report generation.

Summarization applications	Authors
2005 – Genetics	Howard J. Hamilton, Liqiang Geng, Leah Findlater
2005 – Linguistics	Janusz Kacprzyk and Sławomir Zadrożny
2003 – Decision Support Systems	Dmitri Roussinov and J. Leon Zhao

Table 13. Summarization Applications

3.3. Association rules

Association rule mining searches for interesting relationships among items in a given data set. This section provides an introduction to association rule mining introduction to association rule mining.

Let I={i1, i2,…,im} be a set of items. Let D, the task relevant data, be a set of database transactions where each transaction T is a set of items such that $T \subseteq I$. Each transaction is associated with an identifer, called TID. Let A be a set of items. A transaction T is said to contain A if and only if $A \subseteq T$. An association rule is an implication of the form $A \Rightarrow B$, where $A \subset I$, $B \subset I$ and $A \cap B = \varnothing$. The rule $A \Rightarrow B$ holds in the transaction set D with support s, where s is the percentage of transactions in D that contain $A \cup B$. The rule $A \Rightarrow B$ has confidence c in the transaction set D if c is the percentage of transactions in D containing A which also contain B. That is,

$$\begin{aligned} \text{support}(A \Rightarrow B) &= \text{Prob}\{A \cup B\} \\ \text{confidence}(A \Rightarrow B) &= \text{Prob}\{B \mid A\} \end{aligned} \tag{2}$$

Rules that satisfy both a minimum support threshold (min_sup) and a minimum confidence threshold (min_conf) are called strong (Han and Kamber, 2000).

3.3.1. The Apriori algorithm

Apriori employs breadth-first search and uses a hash tree structure to count candidate item sets efficiently. The algorithm generates candidate item sets (patterns) of length k from k − 1 length item sets. Then, the patterns which have an infrequent sub pattern are pruned. If an item set is frequent, then all of its subsets must also be frequent. Apriori principle holds due to the following property of the support measure; support of an item set never exceeds the support of its subsets.

Apriori algorithm Applications	Authors
2004 – MIS	Ya-Han Hu and Yen-Liang Chen
2004 – CRM	Tzung-Pei Hong, Chan-Sheng Kuo and Shyue-Liang Wang
2004 – Banking	Nan-Chen Hsieh
2004 – Methods Engineering	Shichao Zhang, Jingli Lu and Chengqi Zhang
2002 – Thermodynamics	K. T. Andrews, K. L. Kuttler, M. Rochdi

Table 14. Apriori Algorithm Applications

3.3.2. Multidimensional association rules

A top-down strategy is to be used for multi-level association rules considering more than one dimension of the data.

Multi Dimensional Association Rule Applications	Authors
2003 – Behavioral Science	Ronald R. Holden and Daryl G. Kroner
2003 – Health Care	Joseph L. Breault, Colin R. Goodall and Peter J. Fos

Table 15. Multi Dimensional Association Rule Applications

3.3.3. Quantitative association rules

In practice most databases contain quantitative data and are not limited to categorical items only. Unfortunately, the definition of categorical association rules does not translate directly to the quantitative case. It is therefore necessary to provide a definition of association rules for the case of a database containing quantitative attributes. Srikant and Agrawal (1996) extended the categorical definition to include quantitative data. The basis for their definition is to map quantitative values into categorical events by considering intervals of the numeric values. Thus, each basic event is either a categorical item or a range of numerical values.

Quantitive Association Rule Applications	Authors
2001 – Cybernetics	Ng, V.; Lee, J
2001 – Database Management	Shragai, A.; Schneider, M
2002 – Control and Automation	Tian Yongqing; Weng Yingjun

Table 16. Quantitative Association Rule Applications

3.3.4. Distance-based association rules

Distance Based Association Rule Mining can be applied in data mining and knowledge discovery from genetic, financial, retail, time sequence data or any domain which distance information between items is of importance.

Distance Based Association Rule Applications	Authors
2004 – Cardiology	Jeptha P. Curtis, Saif S. Rathore, Yongfei Wang and Harlan M. Krumholz
2003 – Computational Statistics	Thomas Brendan Murphy
1999 – Decision Support Systems	Daniel Boley, Maria Gini, Robert Gross

Table 17. Distance Based Association Rule Applications

3.4. Sequence discovery

Sequence discovery is similar to association analysis, except that the relationships among items are spread over time. In fact, most data mining products treat sequences simply as associations in which the events are linked by time (Edelstein, 1997) In order to find these sequences, not only the details of each transaction should be captured, but also actors are needed to be identified. Sequence discovery can also take advantage of the elapsed time between transactions that make up the event.

Sequence Discovery Applications	Authors
2003 – Cellular Networks	Haghighat, A.; Soleymani, M.R
2003 – System Sciences	Ming-Yen Lin; Suh-Yin Lee
2002 – Biochemistry	Gilles Labesse, Dominique Douguet, Liliane Assairi and Anne-Marie Gilles
1998 – Chemical Biology	Molly B Schmid

Table 18. Sequence Discovery Applications

4. Data mining and data warehouses

A data warehouse is an integrated collection of data derived from operational data and primarily used in strategic decision making by means of online analytical processing techniques (Husemann and et al., 2000) The data mining database may be a logical rather than a physical subset of your data warehouse, provided that the data warehouse DBMS can

Data Mining in Data Warehouses	Authors
2006 – Production Technologies	Pach, F.P.; Feil, B.; Nemeth, S.; Arva, P.; Abonyi, J.
2005 – Supply Chain Management	Mu-Chen Chen and Hsiao-Pin Wu
2005 – Business Management	Nenad Jukić and Svetlozar Nestorov
2005 – CRM	Bart Larivière and Dirk Van den Poel
2005 – Stock Market	Adam Fadlalla
2005 – Computer Integrated Manufactoring	Ruey-Shun Chen; Ruey-Chyi Wu; Chang, C.C
2005 – Customer Analysis	Wencai Liu; Yu Luo
2005 – Power Engineering	Cheng-Lin Niu; Xi-Ning Yu; Jian-Qiang Li; Wei Sun
2004 – Web Management	Sandro Araya, Mariano Silva and Richard Weber
2004 – Biology	Junior Barrera, Roberto M Cesar-, Jr, João E. Ferreira and M.D.Marco D. Gubitoso
2004 – Oil refinery	A. A. Musaev
2004 – Real Estates	Wedyawati, W.; Lu, M.;
2004 – Electrical Insulation	Jian Ou; Cai-xin Sun; Bide Zhang
2004 – Electrical Engineering	Wang, Z
2003 – Management	Qi-Yuan Lin, Yen-Liang Chen, Jiah-Shing Chen and Yu-Chen Chen
2003 – Corporate Databases	Nestorov, S. Jukic, N.
2002 – Oceanography	Nicolas Dittert, Lydie Corrin, Michael Diepenbroek, Hannes Grobe, Christoph Heinze and Olivier Ragueneau
2002 – Financial Services	Zhongxing Ye; Xiaojun Liu; Yi Yao; Jun Wang; Xu Zhou; Peili Lu; Junmin Yao
2002 – Corporate Databases	Hameurlain, A.; Morvan, F.
2002 – Human Resources	Xiao Hairong; Zhang Huiying; Li Minqiang;
2002 – Medical Databases	Miquel, M.; Tchounikine, A
2002 – Machinery Fault Diagnosis	Dong Jiang; Shi-Tao Huang; Wen-Ping Lei; Jin-Yan Shi
2000 – Biomonitoring	A. Viarengo, B. Burlando, A. Giordana, C. Bolognesi and G. P. Gabrielides
1999 – Banking	Gerritsen, R

Table 19. Data Mining in Data Warehouses

support the additional resource demands of data mining. If it cannot, then you will be better off with a separate data mining database. Data warehouses were emerged from the need to analyze large amount of data together. In the 1990's as organizations of scale began to need more timely data about their business, they found that traditional information systems technology was simply too cumbersome to provide relevant data efficiently and quickly. From this idea, the data warehouse was born as a place where relevant data could be held for completing strategic reports for management. As with all technologic development, over the last half of the 20th century, increased numbers and types of databases were seen. Many large businesses found themselves with data scattered across multiple platforms and variations of technology, making it almost impossible for any one individual to use data from multiple sources. A key idea within data warehousing is to take data from multiple platforms and place them in a common location that uses a common querying tool. In this way operational databases could be held on whatever system was most efficient for the operational business, while the reporting / strategic information could be held in a common location using a common language. Data Warehouses take this even a step farther by giving the data itself commonality by defining what each term means and keeping it standard. All of this was designed to make decision support more readily available and without affecting day to day operations. One aspect of a data warehouse that should be stressed is that it is not a location for all of a businesses data, but rather a location for data that is subject to research. In last few years, corporate database producers adopted data mining techniques for use on customer data. It is an important part of CRM services today. Some other application areas include; Production Technologies, Supply Chain Management, Business Management, Computer Integrated Manufacturing, Power Engineering, Web Management, Biology, Oceanography Financial Services Human Resources Machinery Fault Diagnosis, Bio monitoring, Banking.

5. Conclusion

The purpose of data mining techniques is discovering meaningful correlations and formulations from previously collected data. Many different application areas utilize data mining as a means to achieve effective usage of internal information. Data mining is becoming progressively more widespread in both the private and public sectors. Industries such as banking, insurance, medicine, and retailing commonly use data mining to reduce costs, enhance research, and increase sales. In the public sector, data mining applications initially were used as a means to detect fraud and waste, but have grown to also be used for purposes such as measuring and improving program performance.

Sort of the techniques like decision tree models, time series analysis and regression were in use before the term data mining became popular in the computer science society. However, there are also techniques found by data mining practitioners in the last decade; Support Vector Machines, c-means clustering, Apriori algorithm, etc.

Many application areas of predictive methods are related with medicine fields and became increasingly popular with the rise of biotechnology in the last decade. Most of the genetics research depends heavily on data mining technology, therefore neural networks, classifiers and support vector machines will continue to increase their popularity in near future.

Descriptive methods are frequently used in finance, banking and social sciences to describe a certain population such as clients of a bank, respondents of a questionnaire, etc. Most common technique used for description is clustering; in the last decade k-means method has lost popularity against c-means algorithm. Another common method is association rules where Apriori is the most preferred method by far. By increasing importance of corporate databases and information centered production phenomena association rules continue to increase their growth. Sequence discovery is also a growing field nowadays.

Another aspect of subject discussed in this paper was exploiting data warehouses in conjunction with techniques listed. It is expected that data warehousing and usage of data mining techniques will become customary among corporate world in following years. Data warehouses are regularly used by banks, financial institutions and large corporations. It is unsurprising that they will spread through industries and will be adopted by also intermediate sized firms.

Author details

Adem Karahoca
Bahçeşehir University Software Engineering Department, Turkey

Dilek Karahoca
Bahçeşehir University Software Engineering Department, Turkey
Near East University Computer Technology and Instructional Design PhD Program Department, TRNC

Mert Şanver
Google USA, USA

6. References

Abdel-Aal, R. E. and Mangoud, A. M. (1998), Modeling and forecasting monthly patient volume at a primary health care clinic using univariate time-series analysis, Computer Methods and Programs in Biomedicine, 56, 235-247

Agrawal, R. and Imielinski, T. and Swami, A.N. Mining Association Rules between Sets of Items in Large Databases, Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data.

Agrawal, R. and Srikant, R.(1994), Fast Algorithms for Mining Association Rules, Proc. 20th Int. Conf. Very Large Data Bases (VLDB)

Andrews, K.T.; Kuttler, K.L.; Rochdi, M. and Shillor, M. (2002), One-dimensional dynamic thermoviscoelastic contact with damage, Journal of Mathematical Analysis and Applications, Volume 272, 249-275

Araya, S.; Silva, M. and Weber, R. (2004), A methodology for web usage mining and its application to target group identification, Fuzzy Sets and Systems, 148, 139-152

Armand, S. ; Watelain, E. ; Mercier, M. ; Lensel, G. and Lepoutre, F. X. (2006), Identification and classification of toe-walkers based on ankle kinematics, using a data-mining method, Gait & Posture, 23, 240-248

Bandyopadhyay, S. ; Giannella,C. ; Maulik, U. ; Kargupta, H. ; Liu, K. and Datta, S. (2005), Clustering distributed data streams in peer-to-peer environments, Information Sciences, In Press, Corrected Proof

Bao, H. T. (2000), Knowledge Discovery and Data Mining Techniques and Practice, Department of Pattern Recognition and Knowledge Engineering Institute of Information Technology, Hanoi, Vietnam

Belacel, N. ; Raval, H.B. and Punnen, A.P. (2005), Learning multicriteria fuzzy classification method PROAFTN from data, Computers & Operations Research, In Press, Corrected Proof

Bensaid, A.M.; Hall, L.O.; Bezdek, J.C.; Clarke, L.P.; Silbiger, M.L.; Arrington, J.A.; Murtagh, R.F. (1996), Validity-guided (re)clustering with applications to image segmentation, IEEE Transactions on Fuzzy Systems, 4,112 – 123

Bizovi, K. E. ; Leikin, J. B. ; Hryhorczuk, D. O. and Frateschi, L. J. (1998), Night of the Sirens: Analysis of Carbon Monoxide-Detector Experience in Suburban Chicago, Annals of Emergency Medicine, 31, 737-740

Body, M. ; Miquel, M. ; Bedard, Y. ; Tchounikine, A. (2003), Handling evolutions in multidimensional structures, 19th International Conference Proceedings, 581- 591

Boley, D.; Gini, M.; Gross, R.; (Sam) Han, E.H.; Hastings, K.; Karypis, G.; Kumar, V.; Mobasher, B. and Moore, J. (1999), Partitioning-based clustering for Web document categorization, Decision Support Systems, 27, 329-341

Boser, B. E. ; Guyon, I. M. ; Vapnik, V. N. (1992), A training algorithm for optimal margin classifiers, Proceedings of the fifth annual workshop on Computational learning theory, p.144-152, July 27-29,, Pittsburgh, Pennsylvania, United States

Bowland, B. J. and Beghin, J. C. (2001), Robust estimates of value of a statistical life for developing economies, Journal of Policy Modeling, 23, 385-396

Box, G.; Jenkins, G. and Reinsel, G. (1994) Time Series Analysis, 3rd ed., Prentice Hall

Box, G.E.P. and Jenkins, G.M. (1976) Time series analysis forecasting and control. Prentice Hall, Englewood Cliffs, New Jersey.

Bozdogan, H. and Bearse, P. (2003), Information complexity criteria for detecting influential observations in dynamic multivariate linear models using the genetic algorithm, Journal of Statistical Planning and Inference, 114, 31-44

Breault, J.L.; Goodall, C.R. and Fos, P.J. (2003), Data mining a diabetic data warehouse, Artificial Intelligence in Medicine, 27, 227

Bui, E. N. and Moran, C. J. (2001), Disaggregation of polygons of surficial geology and soil maps using spatial modeling and legacy data, Geoderma, 103, 79-94

Burges, C. J. C. (1998) "A Tutorial on Support Vector Machines for Pattern Recognition". Data Mining and Knowledge Discovery 2:121 - 167

Carpena, R. M. ; Ritter, A. and Li, Y.C. (2005), Dynamic factor analysis of groundwater quality trends in an agricultural area adjacent to Everglades National Park, Journal of Contaminant Hydrology, 80, 49-70

Carvalho, L.M. T. ; Clevers, J. G. P. W. ; Skidmore, A.K. and Jong, S.M. (2004), Selection of imagery data and classifiers for mapping Brazilian semideciduous Atlantic forests, International Journal of Applied Earth Observation and Geoinformation, 5,173-186

Chang, J. ; Peng, P. (2005) Decision-Making and Operation of OTDAS, Journal of Automation, 1, 102- 107

Chen, M.C. and Wu, H.P. (2005), An association-based clustering approach to order batching considering customer demand patterns, Omega, 33, 333-343

Chen, O. ; Zhao, P. ; Massaro, D. ; Clerch, L. B. ; Almon, R. R. ; DuBois, D. C: ; Jusko, W. J. and Hoffman, E. P. (2004), The PEPR GeneChip data warehouse, and implementation of a dynamic time series query tool (SGQT) with graphical interface, Nucleic Acids Research, 32, 578-581

Chen, R.S.; Wu, R.C.; Chang, C.C. (2005), Using Data Mining Technology to Design an Intelligent CIM System for IC Manufacturing, SNPD 2005: 70-75

Chen, T. C. and Hsu, T. C. (2006), A Gas based approach for mining breast cancer pattern, Expert Systems with Applications, 30, 674-681

Chen, X. and Chen, X. (2005), Reply to Comment on "Sensitivity analysis and determination of streambed leakance and aquifer hydraulic properties" by S. Christensen, Journal of Hydrology, 303, 322-327

Chen, Y.L. and Hu, Y.H. (2005), Constraint-based sequential pattern mining: The consideration of recency and compactness, Decision Support Systems, In Press, Corrected Proof

Cheng, D.C.; Schmidt-Trucksäss, A.; Cheng, K.S. and Burkhardt, H. (2002), Using snakes to detect the intimal and adventitial layers of the common carotid artery wall in sonographic images, Computer Methods and Programs in Biomedicine, 67, 27-37

Christensen, J. H.; Hansen, A. B.; Karlson, U. ; Mortensen, J. and Andersen, O. (2005), Multivariate statistical methods for evaluating biodegradation of mineral oil, Journal of Chromatography, 1090, 133-145

Chumsamrong, W. ; Thitimajshima, P. ; Rangsanseri, Y. (1999), Wavelet-based texture analysis for SAR image classification, Geoscience and Remote Sensing Symposium (IGARSS '99), 3, 1564-1566

Cirovic, D. A. (1997), Feed-forward artificial neural networks: applications to spectroscopy, Trends in Analytical Chemistry, 16, 148-155

Cortes, C. and Vapnik, V. (1995) Support vector networks. Machine Learning, 20:273–297.

Craven and Shavlik (1997) Using Neural Networks for Data Mining, Future Generation Computer Systems, 13, pp.211-229.

Crone, S. F. ; Lessmann, S. and Stahlbock, R. (2005), The impact of preprocessing on data mining, An evaluation of classifier sensitivity in direct marketing, European Journal of Operational Research, In Press, Corrected Proof

Crone, S.F. ; Lessmann, S. and Stahlbock, R. (2005), The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing, European Journal of Operational Research, In Press, Corrected Proof

Cuaresma, J. C. ; Hlouskova, J. ; Kossmeier, S. and Obersteiner, M. (2004), Forecasting electricity spot-prices using linear univariate time-series models, Applied Energy, 77, 87-106

Curtis, P.; Rathore, S.S.; Wang, Y.; Krumholz, H.M. (2004), The association of 6-minute walk performance and outcomes in stable outpatients with heart failure, J Card Fail

Data Mining and Knowledge Discovery, 2(2), pp.955-974

DeHoff, P. H. and Anusavice, K. J. (2004), Shear stress relaxation of dental ceramics determined from creep behavior, Dental Materials, 20, 717-725

Delgado, M.; Sánchez, D.; Martín-Bautista, M.J. and Vila, M.A. (2001), Mining association rules with improved semantics in medical databases, Artificial Intelligence in Medicine, 21, 241-245

D'Haese, D. ; Vandermeiren, K. ; Caubergs, R. J. ; Guisez, Y. ; Temmerman, L. ; Horemans, N. (2004), Non-photochemical quenching kinetics during the dark to light transition in relation to the formation of antheraxanthin and zeaxanthin. Journal of Theoretical Biology, 227, 175-186.

Diez, V. ; García, P. A. and Fdz-Polanco F. (1999), Evaluation of methanogenic kinetics in an anaerobic fluidized bed reactor, Process Biochemistry, 34, 213-219

Dittert, N.; Corrin, L.; Diepenbroek, M.; Grobe, H.; Heinze, C. and Ragueneau, O. (2002), Management of (pale-)oceanographic data sets using the PANGAEA information system: the SINOPS example, Computers & Geosciences, 28, 789-798

Dunham, M. (2003) Data Mining: Introductory and advanced topics, New Jersey: Prentice Hall

Edelstein, H. (1997), Mining for Gold, Information Week: April 21, 1997

Erichsen, L. ; Agbaje, O. F. ; Luzio, S. D. ; Owens, D. R. and Hovorka, R. (2004), Population and individual minimal modeling of the frequently sampled insulin-modified intravenous glucose tolerance test, Metabolism, 53, 1349-1354

Ewing, B. T. and Wunnava, P. V. (2001), Unit roots and structural breaks in North American unemployment rates, The North American Journal of Economics and Finance, 12, 273-282

Fadlalla, A. (2005), An experimental investigation of the impact of aggregation on the performance of data mining with logistic regression, Information & Management, 42, 695-707

Falge, E. ; Baldocchi, D. ; Olson, R. ; Anthoni, P. ; Aubinet, M. ; Bernhofer, C. ; Burba, G. ; Ceulemans, R. ; Clement, R. ; Dolman, H.(2001), Gap filling strategies for defensible annual sums of net ecosystem exchange, Agricultural and Forest Meteorology, 107, 43-69

Frank, M. L. ; Fulkerson, M. D. ; Patton, B. R. and Dutta, P. K. (2002), TiO2-based sensor arrays modeled with nonlinear regression analysis for simultaneously determining CO and O2 concentrations at high temperatures, Sensors and Actuators B: Chemical, 87, 471-479

Freeman, J. A. and Skapura, D. M. (1991). Neural Networks: Algorithms, Applications, and Programming Techniques, Addison-Wesley, Reading, MA.

Frees, E. W. (2003), Stochastic forecasting of labor force participation rates, Insurance: Mathematics and Economics, 33, 317-336

Friedman, J. H. (1991). Multivariate adaptive regression splines. The Annals of Statistics, 19:1--141.

Fries, R. S. ; Hansen, M. ; Townshend, J. R. G. And Sohlberg, R.(1998), Global land cover classifications at 8 km spatial resolution: the use of training data derived from Landsat imagery in decision tree classifiers, International Journal of Remote Sensing, 19, 3141- 3168

Gabrielsson, J. L. and Weiner, D. L. (1999), Methodology for pharmacokinetic/ pharmacodynamic data analysis, Pharmaceutical Science & Technology Today, 2, 244-252

Gerritsen, R. (1999), Assessing loan risks: a data mining case study, IT Professional, 1, 16-21

Gil, M. ; Sarabia, E.G. ; Llata, J.R. ; Oria, J.P. (1999), Fuzzy c-means clustering for noise reduction, enhancement and reconstruction of 3D ultrasonic images, Emerging Technologies and Factory Automation, 1, 465-472

Gorr, W. ; Olligschlaeger, A. ; Thompson, Y. - International Journal of Forecasting (2003), Short-term forecasting of crime. International Journal of Forecasting 19:44, 579-594

Gozalo, P. L. and Linton, O. B. (2001), Testing additivity in generalized nonparametric regression models with estimated parameters, Journal of Econometrics, Volume 104, 1-48

Gritti, F. and Guiochon, G. (2005), Critical contribution of nonlinear chromatography to the understanding of retention mechanism in reversed-phase liquid chromatography, Journal of Chromatography, 1099, 1-42

Grosso, L.M. ; Triche, E.W. ; Belanger, K. ; Benowitz, N.L. ; Holford, T.R. and Bracken, M.B. (2005), Association of caffeine metabolites in umbilical cord blood with IUGR and preterm delivery: A prospective cohort study of 1609 pregnancies, Annals of Epidemiology, 15, 659-660

Haghighat, A.; Soleymani, M.R. (2003), A Subspace Scheme for Blind User Identification in Multiuser DS-CDMA, IEEE Wireless Communications and Networking Conference

Hahn, F. (2005), Novel Valve for Automatic Calibration of a Chloride Sensor for River Monitoring, Biosystems Engineering, 92, 275-284

Hairong,X. ; Huiying,Z. ; Minqiang, L. (2002), Regional human resource management decision support system based on data warehouse, Proceedings of the 4th World Congress on Intelligent Control and Automation, 3,2118- 212

Hameurlain, A.; Morvan, F. (1995), Scheduling and mapping for parallel execution of extended SQL queries, Proceedings of the fourth international conference on Information and knowledge management, 197 – 204

Hamilton, H.J.; Geng, L.; Findlater, L. and Randall, D.J. (2005), Efficient spatio-temporal data mining with GenSpace graphs, Journal of Applied Logic, In Press, Corrected Proof

Han, J. and Kamber, M.(2000) Data Mining: Concepts and Techniques. Morgan Kaufmann.

Hand, D.; Mannila, H. and Smyth, P. (2001), Principles of Data Mining, The MIT Press.

Hansen, M.C. ; DeFries, R.S. ; Townshend J.R.G. ; Sohlberg, R. ; Dimiceli, C. ; Carroll, M. (2002), Towards an operational MODIS continuous field of percent tree cover algorithm: examples using AVHRR and MODIS data, Remote Sensing of Environment, 83, 303–319

Hanson, R. ; Stutz, J. ; Cheeseman, P. (1990) "Bayesian Classification Theory", Technical Report FIA-90-12-7-01

Hanson, R.D. (1996) Consensus by Identifying Extremists, Theory and Decision, Vol. 44(3), Springer.

Healy, J.V. ; Dixon, M. ; Read, B.J. and Cai, F.F. (2004), Confidence limits for data mining models of options prices, Physica A: Statistical Mechanics and its Applications,344, 162-167

Henry, R. C. ; Chang, Y. S. and Spiegelman, C. H. (2003), Locating nearby sources of air pollution by nonparametric regression of atmospheric concentrations on wind direction, Atmospheric Environment, 36, 2237-2244

Holden, R.R. and Kroner, D.G. (2003), Differentiating Suicidal Motivations and Manifestations in a Forensic Sample, Canadian Journal of Behavioural Science, 35, 35-44

Holzworth, R. J. (1996), Policy Capturing with Ridge Regression, Organizational Behavior and Human Decision Processes, 68, 171-179

Hong, T.P.; Kuo, C.S. and Wang, S.L. (2005), A fuzzy AprioriTid mining algorithm with reduced computational time, Applied Soft Computing, 5, 1-10,

Hsieh, N.C. (2005), Hybrid mining approach in the design of credit scoring models, Expert Systems with Applications, 28, 655-665

Huang, H. C. ; Pan, J. S. ; Lu, Z. M. ; Sun, S. H. ; and Hang, H. M. (2001), Vector quantization based on genetic simulated annealing, Signal Processing, 81, 1513-1523

Huang, J. ; Brennan, D. ; Sattler, L. ; Alderman, J. ; Lane, B. and O'Mathuna, C. (2002), A comparison of calibration methods based on calibration data size and robustness, Chemometrics and Intelligent Laboratory Systems, 62, 25-35

Huang, Y. M. ; Hung, C. M. and Jiau, H. C. (2005), Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem, Nonlinear Analysis: Real World Applications, In Press, Corrected Proof

Hung, S. Y. ; Yen, D.C. and Wang, H. S. (2005), Applying data mining to telecom churn management, Expert Systems with Applications, In Press, Corrected Proof

Husemann, B.; Lechtenborger, J.; Vossen, G. (2000) Conceptual Data Warehouse Design, Institut f¨ur Wirtschaftsinformatik

Jiang, D. ; Huang, S.T. ; Lei, W.P. ; Shi, J.Y. (2002), Study of data mining based machinery fault diagnosis, Conference on Machine Learning and Cybernetics, 1, 536- 539

Jukić, N. and Nestorov, S. (2005), Comprehensive data warehouse exploration with qualified association-rule mining, Decision Support Systems, In Press, Corrected Proof

Junior Barrera, R M Roberto M Cesar, J E João E Ferreira, M D Marco D Gubitoso, Roberto M Cesar, João E Ferreira, Marco D Gubitoso (2004), An environment for knowledge discovery in biology, Comput Biol Med, 34(5), 427-47. ???

Junttila, J. (2001), Structural breaks, ARIMA model and Finnish inflation forecasts, International Journal of Forecasting, 17, 203-230

Kacprzyk, J. and Zadrożny, S. (2005), Linguistic database summaries and their protoforms: towards natural language based knowledge discovery tools, Information Sciences, 173, 281-304

Kantardzic, M. ; Djulbegovic, B. and Hamdan, H.(2002), A data-mining approach to improving Polycythemia Vera diagnosis, Computers & Industrial Engineering, 43, 765-773

Kecman, V. (2001). Learning and Soft Computing - Support Vector Machines, Neural Networks, Fuzzy Logic Systems, The MIT Press, Cambridge, MA.

Kelly, J. D. (2004), Formulating large-scale quantity–quality bilinear data reconciliation problems, Computers & Chemical Engineering (2004), 28, 357-362

Kim, K. J. (2006), Artificial neural networks with evolutionary instance selection for financial forecasting, Expert Systems with Applications, 30, 519-526

Kim, S. ; Imoto, S. and Miyano, S. (2004), Dynamic Bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data, Biosystems, 75, 57-65

Kim, T. Y. ; Joo, K. ; Sohn, O. I. and Hwang, C. (2004), Usefulness of artificial neural networks for early warning system of economic crisis, Expert Systems with Applications, 26, 583-590

Kim, Y. S. and Street, W. N. (2004), An intelligent system for customer targeting: a data mining approach, Decision Support Systems, 37, 215-228

Kondo, T. ; Pandya, A. S. ; Zurada, J. M. (1999), GMDH-type neural networks and their application to the medical image recognition of the lungs, 38th Annual Conference Proceedings of the SICE

Kuo, R. J. ; Liao J. L. and Tu C. (2005), Integration of ART2 neural network and genetic K-means algorithm for analyzing Web browsing paths in electronic commerce, Decision Support Systems, 40, 355-374

Kuo, W. J. ; Chang, R. F. ; Moon, W. M. ; Lee, C. C: and Chen, D. R. (2002), Computer-Aided Diagnosis of Breast Tumors with Different US Systems, Academic Radiology, Volume 9, 793-799

Labesse, G.; Douguet, D.; Assairi, L. and Gilles, A.M. (2002), Diacylglyceride kinases, sphingosine kinases and NAD kinases: distant relatives of 6-phosphofructokinases, Trends in Biochemical Sciences, 27, Pages 273-275

Larivière, B. and Van den Poel, D. (2005), Predicting customer retention and profitability by using random forests and regression forests techniques, Expert Systems with Applications, 29, 472-484

Lavington, S. ; Dewhurst, N. ; Wilkins, E. and Freitas, A. (1999), Interfacing knowledge discovery algorithms to large database management systems, Information and Software Technology,41, 605-617

Lawson, A.B. and Zhou, H. (2001), Spatial statistical modeling of disease outbreaks with particular reference to the UK foot and mouth disease (FMD) epidemic of 2001, Preventive Veterinary Medicine, 71, 141-156

Lee, A. H.; Gracey, M. ; Wang, K. and Yau, K. K.W. (2005), A Robustified Modeling Approach to Analyze Pediatric Length of Stay, Annals of Epidemiology, 15, 673-677

Lee, H. S. and Siklos, P. L. (1997), The role of seasonality in economic time series reinterpreting money-output causality in U.S. data, International Journal of Forecasting, 13, 381-391

Lee, T. S. ; Chiu, C. C. ; Chou, Y. C. and Lu, C. J. (2006), Mining the customer credit using classification and regression tree and multivariate adaptive regression splines, Computational Statistics & Data Analysis, 50, 1113-1130

Li, Z. and Kafatos, M. (2000), Interannual Variability of Vegetation in the United States and Its Relation to El Niño/Southern Oscillation, Remote Sensing of Environment, 71, 239-247

Li, L. ; Tang, H. ; Wu, Z. ; Gong, J., Gruidl, M. ; Zou, J. ; Tockman, M. and Clark, R. A. (2004), Data mining techniques for cancer detection using serum proteomic profiling, Artificial Intelligence in Medicine, 32, 71-83

Li, X.B. (2005), A scalable decision tree system and its application in pattern recognition and intrusion detection, Decision Support Systems, Volume 41, 112-130

Liao, T. W. (2003), Clustering of time series data—a survey, Pattern Recognation, 38, 1857-18

Lin, M.Y.; Lee, S.Y. (1998), Incremental update on sequential patterns in large databases, Proc Int Conf Tools Artif Intell., 24-31

Lin, Q.Y.; Chen, Y.L.; Chen, J.S. and Chen, Y.C. (2003), Mining inter-organizational retailing knowledge for an alliance formed by competitive firms, Information & Management, 40, 431-442

Lu, S. ; Lu, H. and Kolarik, W. J. (2001), Multivariate performance reliability prediction in real-time, Reliability Engineering & System Safety, 72, 39-45

MacQueen J., (1967) Some methods for classification and analysis of multivariate observations, Proceedings of the 5th Berkeley Symposium on Mathematical Statistics

Mamleev, V. and Bourbigot, S.(2005), Modulated thermogravimetry in analysis of decomposition kinetics, Chemical Engineering Science, 60, 747-766

Manganaris, S. (1996), Classifying sensor data with CALCHAS, Engineering Applications of Artificial Intelligence, 9, 639-644

McQueen, J. B. (1967): "Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability", Berkeley, University of California Press, 1:281-297

Min, I. ; Kim, I. (2004), A Monte Carlo comparison of parametric and nonparametric quantile regressions, Applied Economics Letters

Miwakeichi, F. ; Galka, A. ; Uchida, S. ; Arakaki, H. ; Hirai, N. ; Nishida, M. ; Maehara, T. ; Kawai, K. ; Sunaga, S. and Shimizu, H. (2004), Impulse response function based on multivariate AR model can differentiate focal hemisphere in temporal lobe epilepsy, Epilepsy Research, 61, 73-87

Mladenic, D. ; Grobelnik, M. (1999), Feature selection for unbalanced class distribution and Naive Bayes, Machine Learning-International Workshop Then Conference

Mohanty, M. ; Painuli, D.K. ; Misra, A.K. ; Bandyopadhyaya, K.K. and Ghosh, P.K. (2006), Estimating impact of puddling, tillage and residue management on wheat (Triticum aestivum, L.) seedling emergence and growth in a rice–wheat system using nonlinear regression models, Soil and Tillage Research, 87, 119-130

Mues, C. ; Baesens, B. ; Files, C.M. and Vanthienen, J. (2004), Decision diagrams in machine learning: an empirical study on real-life credit-risk data, Expert Systems with Applications, 27, 257-264

Müller, A. ; Osterhage, H; Sowa, R. ; Andrzejak, R. G. ; Mormann, F. and Lehnertz, K (2005), A distributed computing system for multivariate time series analyses of multichannel neurophysiological data, Journal of Neuroscience Methods, In Press, Corrected Proof

Murphy, T.B. and Martin, D. (2003), Mixtures of distance-based models for ranking data, Computational Statistics & Data Analysis, 41, 645-655

Musaev, A.A. (2004), Analytic information technologies in oil refinery, Expert Systems with Applications, 26, 81-85

Myers, R. A. (2001), Stock and recruitment: generalizations about maximum reproductive rate, density dependence, and variability using meta-analytic approaches, ICES Journal of Marine Science58, 937-951

Nestorov, S.; Jukic, N. (2003), Ad-Hoc Association-Rule Mining within the Data Warehouse, HICSS

Nielsen, M. O. (2004), Local empirical spectral measure of multivariate processes with long range dependence, Stochastic Processes and their Applications, 109, 145-166

NIST/SEMATECH e-Handbook of Statistical Methods,
http://www.itl.nist.gov/div898/handbook

Nitanda, N. ; Haseyama, M. ; Kitajima, H. (2004), An Audio Signal Segmentation and Classification Using Fuzzy c-means Clustering, Proceedings of the 2nd International Conference on Information Technology for Application

Niu, C.L.; Yu, X.N.; Li, J.Q.; Sun, W. (2005), The application of operation optimization decision support system based on data mining in power plant, International Conference on Machine Learning and Cybernetics, 3, 1830 - 1834

Oatley, G. C. and Ewart, B.W. (2003), Crimes analysis software: ‘pins in maps’, clustering and Bayes net prediction, Expert Systems with Applications, 25, 569-588

Oh, W. and Lee, K. (2004), Energy consumption and economic growth in Korea: testing the causality relation, Journal of Policy Modeling, 26, 973-981

Olthof, I. ; Pouliot, D. ; Fernandes, R. and Latifovic, R. (2005), Landsat-7 ETM+ radiometric normalization comparison for northern mapping applications, Remote Sensing of Environment, 95, 388-398

Ou, J.; Sun, C.X.; Zhang, B. (2004), Design and building of data warehouse for steam turbine-generator set, Electrical Insulation, Conference Record of the 2004 IEEE International Symposium on, 12-14

Pach, F.P.; Feil, B.; Nemeth, S.; Arva, P.; Abonyi, J. (2006), Process-Data-Warehousing-Based Operator Support System for Complex Production Technologies Systems, Man and Cybernetics,36, 136 – 153

Padmanabhan, M. ; Bahl, L. R. ; Nahamoo, D.(1999), Partitioning the Feature Space of a Classifier with Linear Hyperplanes, IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, 7-3, 282-287

Pan, F. ; Wang, B ; Hu, X. and Perrizo, W. (2004), Comprehensive vertical sample-based KNN/LSVM classification for gene expression analysis, Journal of Biomedical Informatics, 37, 240-248

Pappas, D. ; Piliouras, N. ; Cavouras, D. (2003), Support vector machines based analysis of brain SPECT images for determining cerebral abnormalities in asymptomatic diabetic patientsI Kalatzis, Medical Informatics and the Internet in Medicine, 28, 221 – 230

Park, K. S. ; Lee, H. ; Jun, C. H. ; Park, K. H. ; Jung, J. W. and Kim, S. B. (2000), Rapid determination of FeO content in sinter ores using DRIFT spectra and multivariate calibrations, Chemometrics and Intelligent Laboratory Systems, 51, 163-173

Pereda, E. ; Quiroga, R. O. and Bhattacharya, J. (2005), Nonlinear multivariate analysis of neurophysiological signals, Progress in Neurobiology, 77, 1-37

Perkowitz, M. and Etzioni, O. (2000), Towards adaptive Web sites: Conceptual framework and case study, Artificial Intelligence, 118, 245-275

Perner, P. (2002) Image mining: issues, framework, a generic tool and its application to medical-image diagnosis, Engineering Applications of Artificial Intelligence, 15, 205-216

Piramuthu, S. On learning to predict Web traffic, Decision Support Systems, 35, 213-229

Piras, P. ; Roussel, C. and Pierrot-Sanders, J. (2001), Reviewing mobile phases used on Chiralcel OD through an application of data mining tools to CHIRBASE database, Journal of Chromatography, 906, 443-458

Pons, M. N. ; Bonté, S. B. and Potier, O. (2004), Spectral analysis and fingerprinting for biomedia characterization, Journal of Biotechnology, 113, 211-230

Quinlan, J. R. (1993) C.4.5: Programs for machine learning, San Francisco: Morgan Kaufmann.

Reick, C. H. and Page, B. (1998), Time series prediction by multivariate next neighbor methods with application to zooplankton forecasts, Mathematics and Computers in Simulation, 52, 289-310

Rinaldi, C. and Cole, T. M. (2004), Environmental seasonality and incremental growth rates of beaver (Castor canadensis) incisors: implications for palaeobiology Palaeogeography, Palaeoclimatology, Palaeoecology, 206, 289-301

Roberts, S. and Martin, M. (2005), A critical assessment of shrinkage-based regression approaches for estimating the adverse health effects of multiple air pollutants, Atmospheric Environment, 39, 6223-6230

Romilly, P. (2005), Time series modelling of global mean temperature for managerial decision-making, Journal of Environmental Management, 76, 61-70

Roussinov, D. and Zhao, J.L. (2003), Automatic discovery of similarity relationships through Web mining, Decision Support Systems, 35, 149-166

Schelter, B. ; Winterhalder, M. ; Hellwig, B. ; Guschlbauer, B. ; Lücking, C. M. and Timmer, J. (2006), Direct or indirect? Graphical models for neural oscillators, Journal of Physiology, 99, 37-46

Schikora, P.F. and Godfrey, M. R. (2003), Efficacy of end-user neural network and data mining software for predicting complex system performance, International Journal of Production Economics, 84, 231-253

Schikora, P. F. and Godfrey, M. R. (2003), Efficacy of end-user neural network and data mining software for predicting complex system performance, International Journal of Production Economics, 84, 231-253

Schlink, U. ; Herbarth, O. ; Richter, M. ; Dorling, S. ; Nunnari, G. ; Cawley, G. and Pelikan, E. (2006),Statistical models to assess the health effects and to forecast ground-level ozone, Environmental Modelling & Software, 21, 547-558

Schmid, M.B. (1998), Novel approaches to the discovery of antimicrobial agents, Current Opinion in Chemical Biology, 2, 529-534

Sebzalli, Y. M. and Wang, X. Z. (2001), Knowledge discovery from process operational data using PCA and fuzzy clustering, Engineering Applications of Artificial Intelligence, Volume 14, 607-616

Serletis, A.and Krause, D. (1996), Empirical evidence on the long-run neutrality hypothesis using low-frequency international data, Economics Letters, 50, 323-327

Shi,X. ; Schillings, P. Boyd, D. (2004), Applying artificial neural networks and virtual experimental design to quality improvement of two industrial processes, International Journal of Production Research, 42, 101–118

Shragai, A.; Schneider, M. (2001), Discovering quantitative associations in databases, IFSA World Congress and 20th NAFIPS International Conference, 1, 423-428

Silva, A. ; Cortez, P. ; Santos, M. F. ; Gomes, L. and Neves, J. (2005), Mortality assessment in intensive care units via adverse events using artificial neural networks, Artificial Intelligence in Medicine, In Press, Corrected Proof

Soffer, M. and Kiryati, N. (1998), Guaranteed Convergence of the Hough Transform, Computer Vision and Image Understanding, 69, 119-134

Solibakke, P. B. (2001), A stochastic volatility model specification with diagnostics for thinly traded equity markets, Journal of Multinational Financial Management, 11, 385-406

Sousa, M. S. R. ; Mattoso, M. and Ebecken, N. F. F. (1999), Mining a large database with a parallel database server, Intelligent Data Analysis, 3, 437-451

Srikant, R., & Agrawal, R. (1996). Mining sequential patterns: Generalizations and performance improvements, . Proc. of the Fifth Int'l Conference on Extending Database Technology (EDBT). Avignon, France

Stassopoulou, A. ; Petrou, M. and Kittler, J. (1996), Bayesian and neural networks for geographic information processing, Pattern Recognition Letters, 17, 1325-1330

Sundberg, R. (2000), Aspects of statistical regression in sensometrics, Food Quality and Preference, 11, 17-26

Swift, S. and Liu, X. (2002), Predicting glaucomatous visual field deterioration through short multivariate time series modeling, Artificial Intelligence in Medicine, 24, 5-24

T. Menzies, Y. Hu, (2003) Data Mining For Very Busy People. IEEE Computer, October 2003. 18-25

Tagliaferri, R. and Pedrycz, W. (2005), Improving RBF networks performance in regression tasks by means of a supervised fuzzy clustering, Neurocomputing, In Press, Corrected Proof

Taylor, J. W. and Buizza, R. (2006), Density forecasting for weather derivative pricing, International Journal of Forecasting, 22, 29-42

Thury, G. and Witt, S. F. (1998), Forecasting industrial production using structural time series models, Omega, 26, 751-767

Ting, S. C. and Chen, C. N. (2002), The asymmetrical and non-linear effects of store quality attributes on customer satisfaction,Total Quality Management, 13, 547 - 569

Truskinovsky, A. M. ; Partin, A. W. and Kroll, M. H.(2005), Kinetics of tumor growth of prostate carcinoma estimated using prostate-specific antigen, Urology, 66,577-581

Valenti, P. ; Cazamajou, E. ; Scarpettini, M. ; Aizemberg, A. ; Silva, W. and Kochen, S. (2006), Automatic detection of interictal spikes using data mining models, Journal of Neuroscience Methods, 150, 105-110

Vapnik, V. ; Golowich, S. and Smola, A. (1997) "Support Vector Method for Function Approximation, Regression Estimation and Signal Processing," Advances in Neural Information Processing Systems, vol. 9, Cambridge, Mass.: MIT Press.

Vapnik, V.N. (1995) The nature of statistical learning theory, Springer-Verlag New York, Inc., New York, NY,

Viarengo, A. ; Burlando, B.; Giordana, A. ; Bolognesi, C. and Gabrielides, G. P. (2000), Networking and expert-system analysis: next frontier in biomonitoring, Marine Environmental Research, 49, 483-486

Vindevogel, B. ; Poel, D. V. and Wets, G. (2005), Why promotion strategies based on market basket analysis do not work, Expert Systems with Applications, 28, 583-590

Waelbroeck, C. ; Labeyrie, L. ; Michel, E. ; Duplessy, J. C. ; McManus, J. F.; Lambeck, K. ; Balbon, E. and M. Labracherie (2002), Sea-level and deep water temperature changes derived from benthic foraminifera isotopic records, Quaternary Science Reviews, 21, 295-305

Wang, Y. ; Raulier, F. and Ung, C. H. (2005), Evaluation of spatial predictions of site index obtained by parametric and nonparametric methods—A case study of lodgepole pine productivity, Forest Ecology and Management, 214, 201-211

Watanabe, H. and Miyazaki, H. (2005), A new approach to correct the QT interval for changes in heart rate using a nonparametric regression model in beagle dogs, Journal of Pharmacological and Toxicological Methods, In Press, Corrected Proof

Wedyawati, W. and Lu, M. (2004), Mining Real Estate Listings Using ORACLE Data Warehousing and Predictive Regression, Proc. of IEEE IRI 2004.

Whang, Y. J. and Linton, O. (1999), The asymptotic distribution of nonparametric estimates of the Lyapunov exponent for stochastic time series, Journal of Econometrics, Volume 91, 1-42

Witten, I..E. ; Frank, E. (200) Data Mining: Practical Machine Learning tools and Techniques with Java Implementations, San Francisco: Morgan Kaufmann.

Wypij, D. ; Newburger, J. W. ; Rappaport, L. A. ; duPlessis, A. J. ; Jonas, R. A. ; Wernovsky, G. ; Lin, M. and Bellinger, D. C. (2003), The effect of duration of deep hypothermic circulatory arrest in infant heart surgery on late neurodevelopment: The Boston Circulatory Arrest Trial, Journal of Thoracic and Cardiovascular Surgery, Volume 126, 1397-1403

Xu, M. and Franti, P (2004) Context clustering in lossless compression of gray-scale image, 13th Scandinavian Conf. on Image Analysis

Yaffee, Robert A. (2002), "Robust Regression Analysis:Some Popular Statistical Package Options", Statistics, Social Science, and Mapping Group Academic Computing Services Information Technology Services

Yamashita, Y. (2000), Supervised learning for the analysis of process operational data, Computers & Chemical Engineering, 24, 471-474

Ye, Z.; Liu, X.; Yao, Y.; Wang, J.; Zhou, X.; Lu, P.; Yao, J. (2002), An intelligent system for personal and family financial service, Neural Information Processing, ICONIP'02

Yongqing, T.; Yingjun, W.; Zhongying, Z. (2002), Proceedings of the 4th World Congress on Intelligent Control and Automation, 3, 2203- 2207

Yves, Q. P. (2003), The production and recognition of emotions in speech: features and algorithms, International Journal of Human-Computer Studies, 59,157-183

Zenkevich, I. G. and Kránicz, B. (2003), Choice of nonlinear regression functions for various physicochemical constants within series of homologues, Chemometrics and Intelligent Laboratory Systems, 67, 51-57

Zhang, B. ; Valentine, I. ; Kemp, P. and Lambert, G. (2006), Predictive modelling of hill-pasture productivity: integration of a decision tree and a geographical information system Agricultural Systems, Volume 87, 1-17

Zhang, B. ; Valentine, I. and Kemp, P.D. (2005), A decision tree approach modelling functional group abundance in a pasture ecosystem, Agriculture, Ecosystems & Environment, Volume 110, 279-288

Zhang, S.; Lu, J. and Zhang, C. (2004), A fuzzy logic based method to acquire user threshold of minimum-support for mining association rules, Information Sciences, 164, 1-16

Zhang, Y.F. ; Mao, J. L. ; Xiong, Z. Y. (2003), An efficient clustering algorithm, Machine Learning and Cybernetics, 2003 International Conference

Zhong, S.(2005), Efficient streaming text clustering, Neural Networks, 18, 790-798

Zhong, W. ; Altun, G ; Harrison, R. ; Tai, P. C. ; Pan, Y. (2005), Improved K-means clustering algorithm for exploring local protein sequence motifs, IEEE Transactions On Nanobioscience, Vol. 4, No. 3, September 2005, 255

Zlatnik, M. G. ; Copland, J. A. ; Ives, K. and Soloff, M. S. (2000), Functional oxytocin receptors in a human endometrial cell line, American Journal of Obstetrics and Gynecology, 182, 850-855

Zorman, M. ; Podgorelec, V. ; Kokol, P. ; Peterson, M. ; Lane, J. (2000), Decision tree's induction strategies evaluated on a hard real world problem, 13th IEEE Symposium on Computer-Based Medical Systems, 19-24