# DATA MINING AND APPLICATION OF IT TO CAPITAL MARKETS

**Cenk AKKAYA**

Faculty of Economics and Administrative Sciences, Dokuz Eylül University

Dokuz Eylül Üniversitesi, İİBF, İşletme Bölümü, Dokuzçeşmeler Kampüsü, Buca/İZMİR

e-mail: cenk.akkaya@deu.edu.tr

**Ceren UZAR**

Fethiye A.S.M.K Vocational School of Higher Education, Muğla University

Fethiye A.S.M.K M.Y.O. Fethiye/MUĞLA

e-mail: cerenuzar@mu.edu.tr

## —Abstract —

Nowadays with the development of technology importance given to knowledge increases gradually. Data mining enables to form forecasts and models regarding future by making use of past data. Any method which helps to discover data can be used as a data mining method. Enterprises gain important competitive advantage by data mining methods. Data mining is used in different fields. In finance field it is a specially used in financial performance applications, guessing the enterprise bankruptcies and failures, determining transaction manipulation, determining financial risk management, determining customer profile and depth management. It can be costly, risky and time consuming for enterprises to gain knowledge. Thus today enterprises use data mining as an innovative competitive mean. The aim of the study is to determine the importance of data mining applications to capital markets.

**Key Words:** *Data Mining, Capital Markets, Data Mining Methods*

**JEL Classification:** M13

## 1. INTRODUCTION

The application of data mining (DM) and its related techniques and also technologies has been greatly expanded in the last few years. Data mining is the process of handling information from databases which can not be seen directly. Therefore data mining can be used for a variety of purposes in private sector. Industries such as banking, insurance, medicine commonly use data mining to reduce costs, enhance research, and increase sales. Data analysis techniques that have been traditionally used for such tasks include regression analysis, cluster analysis, numerical taxonomy, multidimensional analysis, other multivariate statistical methods, stochastic models, time series analysis, nonlinear estimation techniques, and others (Sumathi and Sivanandam, 2006: 24). Data mining takes advantage of advances in the fields of artificial intelligence (AI) and statistics. Both disciplines have been working on problems of pattern recognition and classification (Two Crows, 2005: 4).

## 2. DEFINITION OF DATA MINING

Today's databases have reached to dimensions described by terabytes. In time, it has been determined that such a large volume of data has secret information with strategical importance. The main question is about how to disclose such secret information. The most update and popular answer to this question is Data Mining (DM) (Koyuncugil, 2007; 7). DM may also be regarded as the natural development processed of the information Technologies. Large scale data may be regarded as a data mine comprising valuable data within large scale databases in different areas. And, DM is defined as the process of producing substantive information, which used to be unknown before these data (Albayrak and Yılmaz, 2009; 32).

The purpose of data mining is to create decision making models for estimation of the behaviours in the future based on the analysis of the past actvities (Koyuncugil and Özgülbaş, 2009; 24). At this point, it is a means which supports the decision process to be given for reaching to the solution rather than being a solution alone and provides the information required for solving the problem. Data mining refers to rendering assistance to the analyst for finding the patterns and relations between the data created in the stage of working (Akgöbek and Çakır, 2009; 802).

In general the data mining process iterates through five basic steps (Sumathi and Sivanandam: 2006; 43):

- Data selection. This step consists of choosing the goal and the tools of the data mining process, identifying the data to be mined, then choosing appropriate input attributes and output information to represent the task.
- Data transformation. Transformation operations include organizing data in desired ways, converting one type of data to another (e.g., from symbolic to numerical), defining new attributes, reducing the dimensionality of the data, removing noise, "outliers," normalizing, if appropriate, deciding strategies for handling missing data.
- Data mining step per se. The transformed data is subsequently mined, using one or more techniques to extract patterns of interest. The user can significantly aid the data mining method by correctly performing the proceeding steps.
- Result interpretation and validation. For understanding the meaning of the synthesized knowledge and its range of validity, the data mining application tests its robustness, using established estimation methods and unseen data from the database. The extracted information is also assessed (more subjectively) by comparing it with prior expertise in the application domain.
- Incorporation of the discovered knowledge. This consists of presenting the results to the decision maker who may check/resolve potential conflicts with previously believed or extracted knowledge and apply the new discovered patterns.

Based on the type of knowledge that is mined, data mining can be mainly classified into the following categories.

## 3. CLASSIFICATION OF DATA MINING METHODS

Data mining methods may be categorized as either supervised or unsupervised. In unsupervised methods, no target variable is identified as such. Instead, the data mining algorithm searches for patterns and structure among all the variables. The most common unsupervised data mining method is clustering (Larose, 2005; 90).

When the data is unlabelled and each instance does not have a given class label the learning task is called unsupervised. If we still want to identify which instances belong together, that is, form natural clusters of instances, a clustering algorithm can be applied (Olafsson et all., 2008:1439). Clustering techniques can be used to identify stable dependencies for risk management and investment management (Zhang and Zhou, 2004: 514).

Another unsupervised learning approach is association rule discovery that aims to discover interesting correlation or other relationships among the attributes. Association rule mining was originally used for market basket analysis, where items are articles in the customer's shopping cart and the supermarket manager is looking for associations among these purchases (Olafsson et all., 2008:1441).

One of the most common learning tasks in data mining is classification. Contrary to clustering, classification is a supervised way of learning. The database contains one or more attributes that denote the class of a tuple and these are known as predicted attributes, whereas the remaining attributes are called predicting attributes (Sumathi and Sivanandam, 2006: 577).

Many methods have been studied for classification. One of the techniques for classification is the top-down induction of decision trees. One of the main reason behind their popularity appears to be their transparency, and hence relative advantage in terms of interpretability (Olafsson et all., 2008:1436). Examples of classification applications are pattern recognition, medical diagnosis like that.

Bayes method is another simple but yet effective classifier. This method learns the conditional probability of each attribute given the class label from the training data. Classification is then done by applying Bayes rule to compute the probability of a class value given the particular instance and predicting the class value with the highest probability (Olafsson et all., 2008:1437). Bayes methos uses graphical models, allow representing dependencies among attribute subsets.

Another approach for classification is neural networks. The inductive learning of neural networks from data is referred to as training this network, and the most popular method of training is back-propagation (Olafsson et all., 2008:1438).

Another definition just lists methods of data mining: Decision Trees, Neural Networks, Rule Induction, Nearest Neighbors, Genetic Algorithms. Less formal, but the most practical definition can be taken from the lists of components of current data mining products (Kovalerchuk and Vityaev, 2000:16).

There are dozens of products, including, Intelligent Miner (IBM), SAS Enterprise Miner (SAS Corporation),Recon (Lockheed Corporation), MineSet (Silicon Graphics), RelationalData Miner (Tandem), KnowledgeSeeker (Angoss Software), Darwin (Thinking Machines Corporation), ASIC (NeoVista Software), Clementine (ISL Decision Systems, Inc), DataMind Data Cruncher (DataMind Corporation), BrainMaker (California Scientific Software), WizWhy(WizSoft Corporation) (Kovalerchuk and Vityaev, 2000:16).

## 4. SUGGESTIONS FOR IMPLEMENTATION OF DATA MINING ON THE CAPITAL MARKETS

Companies active in the financial markets use data mining to determine market and industry characteristics as well as to predict individual company and stock performance (Two Crows, 2005: 5), since risk management has become an important topic for all institutions, especially for SMEs, banks, credit rating firms, and insurance companies.

In this respect, surveillance and early warning systems, financial failure predictions and studies for predicting anormal stock exchange revenues can be exemplified as follows.

### 4.1. Surveillance and Warning Systems

Early warning systems in finance are vital tools for monitoring and detecting events in financial markets. New York Stock Exchange uses a computerized early warning system called ‘Stock Watch’ used for manipulation and pre-determining the insider training. NYSE (2011) defines Stock Watch as “The state of art computerized surveillance unit of NYSE which monitors the anormal price and quantity movements which may not be legal for the securities recorded in NYSE”. The NYSE's state-of-the-art computer surveillance unit, which monitors the market in NYSE-listed stocks for aberrant price and volume activity, which may indicate illegal transactions (www.nyse.com, 20.02.2011).

The Stock Watch unit of Market Surveillance combines the human judgment of analysts with electronic data-mining and pattern-detection systems, links to news and research, as well as public databases of listed company officers, directors, and other insiders to detect possible insider trading and market manipulation. Market Surveillance forwards cases that involve possible rules violations for further investigation to the NYSE Regulation Enforcement division or to the SEC for matters outside NYSE jurisdiction.

When unusual trading is detected by a senior official on the Floor or by the Market Surveillance Stock Watch unit, the Exchange will contact the listed company and request it to issue a news release that addresses the unusual market activity. If there is material corporate news to account for the activity, trading will be interrupted on a "news pending" basis. If the listed company declines to issue a news release, the NYSE will issue its own release stating the company’s position. The re-opening process will begin with the public dissemination of an indication of interest to bring supply and demand more closely in balance (www.nyse.com, 20.02.2011).

The market surveillance functions at the Stock Exchange of Thailand are enhanced by a computerized electronic monitoring system. Comprised of three main systems for the alert, detection and documentation functions, the ATOMS system was introduced in 1995 to monitor, analyze and facilitate the investigation of suspicious activities. Its effectiveness and reliability continue to play a key role in ensuring proper surveillance, exposing trading malpractice, supporting Exchange investigations and effective enforcement in the process (www.set.or.th, 20.02.2011).

Koyuncugil (2006) has developed a similar system to those already used in New York and Thailand Stock Exchanges for Istanbul Securities and Equities Stock Exchange Market (HSP). The system is an early warning system based on data mining for determining manipulation and successful operation of the system has been determined by operation using real data.

The system determines the following automatically,

- Manipulated equities,
- Manipülative transactions,
- Mediators mediating the manipulative transactions,
- Investors performing the manipulation,

It is considered that in case a similar system is to be realized in Capital Markets Board of Turkey or Istanbul Stock Exchange, it will provide a large contribution.

### 4.2. Predicting Corporate Bankrupties Using Decision Trees

Evaluating the financial healthiness of a firm and assessing the default risk have been of great interest to many stakeholders such as creditors, investors, and government (Cho et all., 2009: 403).

The early warning model for detecting the bankruptcy risk is useful for the firms. Analysis is based on data mining technıques in order to identify the firms' categories accordingly to the bankruptcy risk levels. Therefore, the researchers chose decision trees as their analysis method, because of the transparency of the algorithm.

In this subject, the studies conducted by Vasilescu et all., (2011) may be summarized as follows.

The first step was the collection of necessary data regarding the Romanian SMEs from Dolj county (12,496 firms) using the data basis from the Ministry of Public Finance, in the year 2007.

There will be selected financial ratios tahat seperate the bankrupcty firms from the low bankrupcty ones. 15 Ratios can be selected and grouped in 4 categories, such as profitability ratios, risk ratios, liguidity and solvability ratios, rotation ratios. Thus, for the study there were be selected the most discriminant financial ratios, grouped in 5 classes, in function of the risks involved, as follows:

- First class: very high risk,
- Second class: high risk,
- Thirth class: medium risk,
- Fourth class: low risk
- Fifth class: very low risk.

Besides the financial ratios, in their analysis there were introduced the non-financial indicators. The using of the non-financial indicators has as purpose the achievement of the progresses made through the application of specific measures which ensure the success of the firm.

In the third phase, qualitative and quantitative data to be obtained through phases 1 and 2 will be analyzed with data mining (Koyuncugil and Özgülbaş, 2009: 41).

CHAID method was applied because it has the advantage of simplicity in comparision with other possible methods to be used. As a result of CHAID method, firms were grouped in function of several parameters in 5 risk classes regarding the risk bankrupcty. It was used a data basis with 12,496 firms which included the 15 financial ratios, determined on the information from the SMEs' balance sheets of the year 2007 and 10 non-financial indicators, determined in a previous research from a market research on the analyzed firms.

### 4.3. Predicting Abnormal Stock Market Returns

Insider traders usually make abnormal returns. The dominant data mining technique used in stock market prediction so far is neural network modeling. Some researchers are interested in using data mining methodology to increase the ability to predict abnormal stock price returns arising from legal insider trading.

There are several important design issues involved in applying neural network approach to stock prediction (Zhang and Zhou, 2004: 516):

- Determine the optimal length of time in the past from which to analyze data. Many studies take an aggregate of insider activities one month before the current date and then predict the future trend;
- Select time-sensitive indicators as network inputs; and
- Decide what to do with the lagged data. In general, the inputs to neural networks include daily transaction volume, interest rates, stock prices, moving average, and/or rate of change, etc.

In this subject, the studies conducted by Safer, (2002) may be summarized as follows.

The insider trading data used in this study are from January 1993 to mid June 1997.The data was collected from Securities and Exchange Commission.

The stocks used in the analyses included all stocks in the S&P 600 (small cap), S&P 400 (midsize cap) and S&P 500 (large cap) as of June 1997 that had insider records for the entire period of the study. There were 946 stocks in the three market caps which had available data in January1993. From the list of 946 stocks, the sample included every stock that averaged at least 2 buys per year. This resulted 343 stocks being used for the study.

The variables in the original data set include the company, name and rank of the insider, transaction date, stock price, number of shares traded, type of transaction (buy or sell), and number of shares held after the trade. To assess an insider’s prior trading patterns, the study examined the previous 9 and 18 weeks of trading history. The prediction time frames for predicting abnormal returns were established as 3, 6, 9, and 12 months. Then the data can be split into a training set (80% of the data) and validation set (20%). A neural network model is applied.

Safer found that the prediction of abnormal returns could be enhanced in the following ways:

- Extending the time of the future forecast up to 1 year;
- Increasing the period of back aggregated data;
- Narrowing the assessment to certain industries such as electronic equipment and business services and
- Focusing on small and midsize rather than large companies.

## 5. CONCLUSION

Data mining can be expressed as disclosure of hidden values and usable information among a large amount of data. The purpose of data mining is to create decision making models for future predictions based on the analysis of the past activities.

Data mining can find an application usage field in almost every media in which the data is produced intensively and consequently the databases are created. It can be used in the field of finance, particularly for early determination of financial failure, determination of financial information manipulation and internal learners, preventing law appricaiton in the initial public offerings, determining the market anomalities, analysis of investor risk perception depending on education, income level, gender and such factors, analysis of integration level in financial markets, analysis of the level of transmission of the capital provided from the initial public offerings to the reel investments, presupposition of bear and bull periods in the stock exchanges, creating risk (volatility) index for markets, company evaluation and equities price depth analysis of SMEs in entrepreneur capital market and forecasting og anotmal stock exchange revenues.

In this study, information on data mining concept and data mining methods has been rendered. Furthermore, the suggestions for applying data mining on capital markets have been described by means of some related studies.

## BIBLIOGRAPHY

Akgöbek, Ömer, Çakır Fuat (2009), “Expert System Design with Data Mining”, *11. Akademic Information Conference*, 11-13 February, Harran University, Şanlıurfa, pp.801-806.

Albayrak, A.Sait, Yılmaz Şebnem Koltan (2009), “Data Mining: Decision Tree Algorithms and an Application on Ise Data”, Süleyman Demirel University, *The Journal of Faculty of Economics and Administrative Sciences*, Vol. 14, No. 1, pp.31-52.

Cho, Sungbin, Kim Jinhwa and Bae J. Kwon (2009), “An Integrative Model with Subject Weight Based on Neural Network Learning for Bankruptcy Prediction”, *Expert Systems with Applications*, Vol. 36, No. 1, pp.403–410.

Kovalerchuk, Boris, Vityaev Evgenii E., (2000), Data Mining In Finance, Advances in Relational and Hybrid Methods, Newyork: Kluver Academic Publishers.

Koyuncugil, A. Serhan (2006). “Fuzzy Data Mining and its Application to Capital Markets”, Unpublished doctoral dissertation, Ankara University, Ankara.

Koyuncugil, A. Serhan (2007), “Determination of Stock Market Corporation’s Sectoral Risk Profile with Data Mining”, *Capital Market Board Resarch Report*.

Koyuncugil, A. Serhan, Özgülbaş Nilgün (2009), “Risk Modeling by CHAID Decision Tree Algorithm”. *ICCES*, Vol. 11, No. 2, pp.39-46.

Koyuncugil, A.Serhan, Özgülbaş Nermin (2009), “Data Mining: Data Mining: Using and Applications in Medicine and Healthcare”, *Journal of Information Technology*, Vol .2, No. 2, pp.21-32.

Larose, D.T., 2005. Discovering Knowledge In Data, An Introduction to Data Mining, New Jersey: John Wiley & Sons.

Newyork Stock Ehxchange (2011), *Stock Watch*, http://www.nyse.com/glossary/glossarylinks.html?a=1048903219379, [Accessed 20.02.2011]

Newyork Stock Exchange (2011), *About NYSE Regulation*, http://www.nyse.com/regulation/nyse/1045516499685.html, [Accessed 20.02.2011]

Olafsson, Sigurdur, Li Xiaonan and Wu, Shuning (2008), “Operations Research and Data Mining”, *European Journal of Operational Research*, Vol. 187, No. 3, pp. 1429–1448.

Safer, Alan M. (2002), “The Application of Neural Networks to Predict Abnormal Stock Returns Using Insider Trading Data”, *Applied Stochastic Models in Business and Industry*, Vol.18, No. 4, pp. 381–389.

Sumathi, Sai, Sivanandam S.N., (2006), Introduction to Data Mining and its Applications, Vol. 29, ISBN 3-540-34350-4. Verlag Berlin Heidelberg: Springer.

The Stock Exchange of Thailand (2011), *Investor Protection*, http://www.set.or.th/en/regulations/protection/protection_p1.html, [Accessed 20.02.2011]

TWO CROWS (2005), Introduction to Data Mining and Knowledge Discovery, USA: Two Crows Corporation.

Vasilescu, L. Giurca, Siminica Marian, Pirvu Ceraselai, Ionascu Costel, Mehedintu Anca (2011), “Data Mining Used for Analyzing the bankruptcy Risk of the Romanian SMEs”, (in Ali Serhan Koyuncugil and Nermin Özgülbaş- Ed., Surveillance Technologies and Early Warning Systems, pp.144-181. Hershey New York: Information Science Reference.

Zhang, Dongsong, Zhou Lina (2004), "Discowering Golden Nuggets: Data Mining in Financial Application", *Ieee Transactıons on Systems, Man, and Cybernetıcs*—Part C: Applicatıons and Reviews, Vol. 34, No. 4, pp. 513-522.