PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1150 2007
Informatyka Ekonomiczna 10

Leszek Ziora

APPLICATION OF DATA MINING METHODS AND TECHNIQUES IN AN ENTERPRISE. REVIEW OF CHOSEN PRACTICAL EXAMPLES

1. Introduction

Organizations today routinely collect and manage terabytes of data in their databases. In order to use these data for the enterprise's management purposes, organizations must be able to transform the data they have collected into useful information. When dealing with large sets of data, this transformation can be a challenge for organizations. It is difficult to understand information hidden in data without the aid of data analysis techniques. Data mining provides an attractive opportunity for this purpose. It combines work from areas such as statistics, machine learning, pattern recognition, data warehouses, databases and high performance computing. Tools for data mining have the ability to parse enormous amounts of data and discover significant patterns and relationships that might otherwise have taken a person thousands of hours to find it. Data mining is a broad field of data analysis and pattern discovery and there are numerous subfields of data mining. The use of these data may lead to gaining competitive advantage of a given enterprise. Data mining as the fastest growing field has successfully provided solutions for finding information from data in HR, bioinformatics, pharmaceuticals, banking, retail, sports and entertainment, and many more fields. Many important problems in science and industry have been addressed by data mining methods, such as neural networks, fuzzy logic, decision trees, genetic algorithms, and statistical methods.

2. Notion of data mining

Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner (Hand et al. 2001). The another definition is provided by Gartner Group and according to this definition, "Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques”. The ongoing remarkable growth in the field of data mining and knowledge discovery has been fueled by confluence of a variety of factors (Larose 2005):

- the explosive growth in data collection,
- the storing of the data in data warehouses, so that the entire enterprise has access to a reliable current database,
- the availability of increased access to data from Web navigation and intranets,
- the competitive pressure to increase market share in a globalized economy,
- the tremendous growth in computing power and storage capacity.

Although the existence of a data warehouse is not a prerequisite for data mining, in practice, the task of data mining, especially for some large companies, is made a lot easier by having access to a data warehouse. A primary goal of a data warehouse is to increase the “intelligence” of a decision process and the knowledge of the people involved in this process (Kantardzic 2003). Data warehouse is a technology which includes creation of integrated data sets containing unified historical data related to the enterprise and it is strategic investment of the enterprise (*Zarządzanie...* 2004). Contrary to the transactional systems, data in the warehouses are permanent, organized thematically, integrated and properly aggregated and usually have time dimension (Jarke et al. 2003). The data found in the data warehouse are cleansed, integrated and properly organized. To the basic effects of data warehouse appliance belongs quick obtaining of information throughout data drilling which consists of deep analysis called drill-down, aggregation analysis – drill-up and sectional analysis – slicing and dicing. This foundation is precisely what the data miner and the explorer need in order to start the exploration and data mining activity (Nowicki et al. 2006). The data warehouse is often not the only source and external data together with other data can be freely mixed with data warehouse data in the course of doing exploration and mining (Inmon 2002). The practice of contemporary management indicated the need of maintenance of high quality data in the relation to Information Systems and most of gathering information models possess elements of extraction, transformation and load of data connected with the processes of quality and integrity control of data (Simon, Shaffer 2002). The whole data mining process can be presented in four stages. The process starts with a clear definition of the problem – stage 1, followed by stage 2 which is the selection process aimed at identifying all the internal and external sources of information and selecting the subgroup of data necessary for the application of DM in order to deal with the problem. Stage 3 consists of preparing the data, which includes preprocessing. It is divided into visualization tools and data reformatting tools as it was illustrated in Fig. 1. This preparation is crucial for the final quality of the results and because of this, the tools used are very important. The software used at this stage must be capable of performing many different procedures, such as adding values, carrying out conversions, filtering variables, having a format for exporting data, working with relational databases and mapping entry variables.

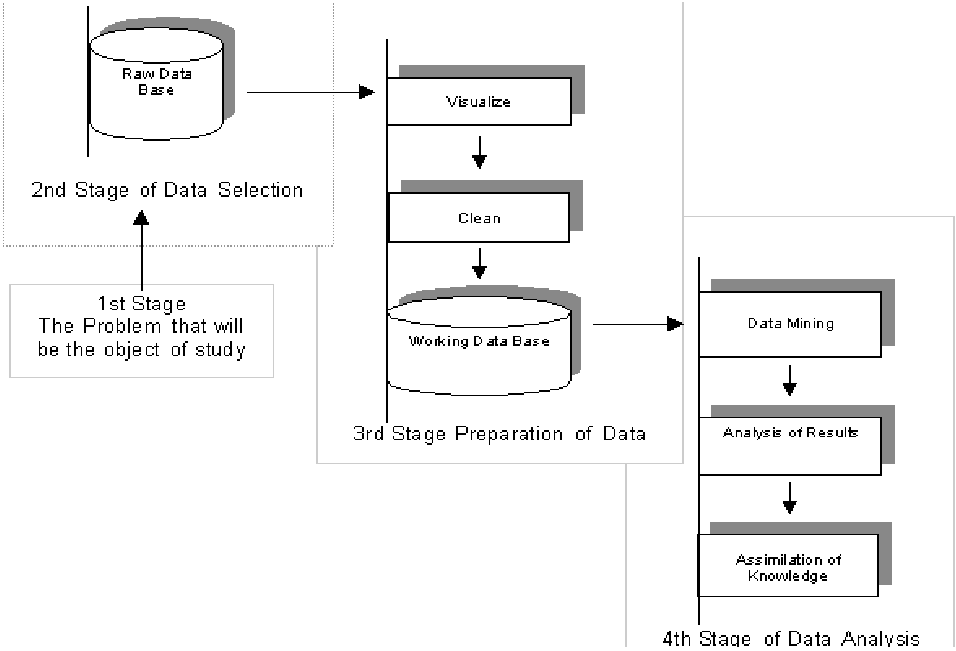
Raw Data Base
2nd Stage of Data Selection
1st Stage
The Problem that will be the object of study
Visualize
Clean
Working Data Base
3rd Stage Preparation of Data
Data Mining
Analysis of Results
Assimilation of Knowledge
4th Stage of Data Analysis

Fig. 1. The process of data mining

Source: (Cabena et al. 1998).

In the stage 4 there is presented the analysis of results obtained through the DM process, two basic aspects of which have to be considered: giving information about new discoveries and presenting them in such a way that they can be potentially exploited (Tarapanoff et al. 2001). Data preprocessing usually includes noise elimination, feature selection, data partition, data transformation, data integration, and missing data processing (Wang, Fu 2005).

3. Classification of data mining tasks

Many problems of intellectual, economic, and business interest can be presented in terms of the following five tasks: **classification, estimation, prediction, clustering, description and profiling**. The first three are all examples of directed data mining, where the goal is to find the value of a particular target variable. One of the most common data mining tasks is **classification**. It consists of examining the features of a newly presented object and assigning it to one of a predefined set of classes. The objects to be classified are generally represented by records in a data-base table or a file, and the act of classification consists of adding a new column with a class code of some kind. The classification task is characterized by a well-defined definition of the classes, and a training set consisting of pre-classified examples. The task is to build a model of some kind that can be applied to unclassified data in order to classify it. Examples of classification tasks may include (Berry, Linoff 2004):

- classifying credit applicants as low, medium, or high risk,
- choosing content to be displayed on a Web page,
- determining which phone numbers correspond to fax machines,
- spotting fraudulent insurance claims,
- assigning industry codes and job designations on the basis of free-text job descriptions.

Decision trees and nearest neighbor techniques are techniques well suited to classification. Neural networks and link analysis are also useful for classification in certain circumstances.

Estimation deals with continuously valued outcomes and in practice it is often used to perform a classification task. The estimation approach has the great advantage that the individual records can be rank ordered according to the estimate. Examples of estimation tasks include e.g. estimating a family's total household income, estimating the lifetime value of a customer and estimating the probability that someone will respond to a balance transfer solicitation. Regression models and neural networks are well suited to estimation tasks. Survival analysis is well suited to estimation tasks where the goal is to estimate the time to an event, such as a customer stopping.

Prediction is the same as classification or estimation, except that the records are classified according to some predicted future behavior or estimated future value. In a prediction task, the only way to check the accuracy of the classification is to wait and see. Any of the techniques used for classification and estimation can be adapted for use in prediction by using training examples where the value of the variable to be predicted is already known, along with historical data for those examples. The historical data are used to build a model that explains the current observed behavior. When this model is applied to current inputs, the result is a prediction of future behavior. Examples of prediction tasks addressed by the data mining techniques may include (Berry, Linoff 2004):

- predicting the size of the balance that will be transferred if a credit card prospect accepts a balance transfer offer,
- predicting which customers will leave within the next 6 months,
- predicting which telephone subscribers will order a value-added service such as three-way calling or voice mail etc.

The choice of technique depends on the nature of the input data, the type of value to be predicted, and the importance attached to explicability of the prediction.

Clustering is the task of segmenting a heterogeneous population into a number of more homogeneous subgroups or *clusters*. What distinguishes clustering from classification is that clustering does not rely on predefined classes. The records are grouped together on the basis of self-similarity. Clustering is often done as introduction to some other form of data mining or modeling, e.g. it might be the first step in a market segmentation effort.

Profiling in data mining is used to describe what is going on in a complicated database in a way that increases our understanding of the people, products, or processes that produced the data in the first place (Berry, Linoff 2004). A good enough *description* of a behavior will often suggest an *explanation* for it as well. It suggests where to start looking for an explanation. Decision trees are a powerful tool for profiling customers with respect to a particular target or outcome.

4. Overview of chosen methods and techniques of data mining

There are many methods and techniques which occur in data mining process. There is worth to mention the most significant ones as: **decision trees, artificial neural neurons, genetic algorithms, nearest neighbor method, rule induction and data visualization**.

The **decision-tree** is tree-shaped structures that represent sets of decisions and is the most widely used logic method. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID) used for classification of a dataset. They provide a set of rules that can be applied to a new dataset to predict which records will have a given outcome. A typical decision-tree learning system adopts a top-down strategy that searches for a solution in a part of the search space. It guarantees that a simple tree will be found (Kantardzic 2003). A decision tree consists of *nodes* where attributes are tested. The outgoing *branches* of a node correspond to all the possible outcomes of the test at the node. A simple decision tree for classification of samples with two input attributes X and Y is presented in Fig. 2.

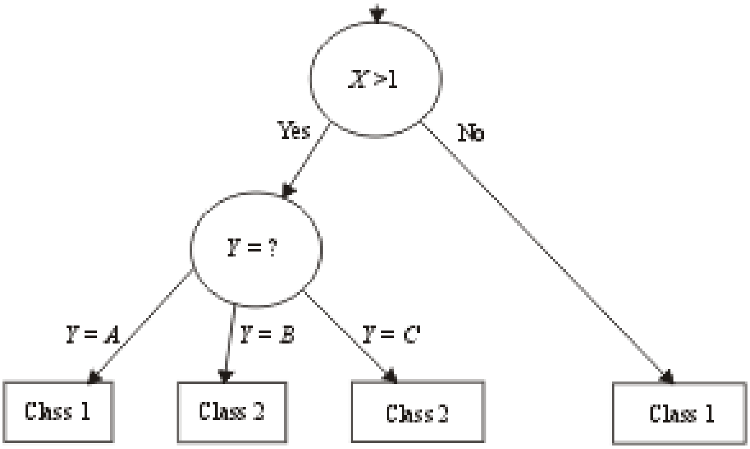

Fig. 2. A simple decision tree with the tests on attributes *X* and *Y*
Source: (Kantardzic 2003).

All samples with feature values *X* >1 and *Y* = *B* belong to Class2, while the samples with values *X* <1 belong to Class1, whatever the value for feature *Y*. Decision trees allow the user to understand the inferred model. A well-known tree-growing algorithm for generating decision trees based on univariate splits is Quinlan's *ID3* with an extended version called *C4.5*. Decision trees provide an effective method of Decision Making because they (*Decision*… 2007): clearly lay out the problem so that all options can be challenged, allow to analyze fully the possible consequences of a decision, provide a framework to quantify the values of outcomes and the probabilities of achieving them and help to make the best decisions on the basis of existing information and best guesses.

Another model of data mining can be **artificial neural network** based onto artificial neuron which is an information-processing unit. It consists of three basic elements (Kantardzic 2003): *a set of connecting links* from different inputs x_i (or synapses), each of which is characterized by a weight or strength w_{kiz}, *an adder* for summing the input signals x_i weighted by the respective synaptic strengths w_{ki} and a*n activation function* for limiting the amplitude of the output y_k of a neuron. In mathematical terms, an artificial neuron is an abstract model of a natural neuron, and its processing capabilities are formalized.

The other models used in data mining are: **genetic algorithms** which are optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution. The next is **nearest neighbor method** which is a technique that classifies each record in a dataset based on a combination of the classes of the *k* record(s) most similar to it in a historical dataset (where *k* = 1). Sometimes called the *k*-nearest neighbor technique. And the another technique is **rule induction** which is the extraction of useful if-then rules from data based on statistical significance and **data visualization** which is the visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

5. The examples of practical applications of data mining

There are many applications of data mining models and techniques in organizations. There are mentioned here: practical application in **Human Resource information, financial data analysis and retail industry.** The data mining models and techniques can be used to support **human resources** decisions especially in employee recruitment and to decrease costs of employee turnover (Wang 2003). In order to perform it there can be applied a regression model or Neural Network mode where a success score would be assigned to employees in the data sample. Then the possible predictor variables would be decided, run on the data to obtain a formula that predicts the success score of an employee with the given predictor variable values. In this modeling approach, two samples – one of high performers and one of low performers – are obtained. Then, a statistical group score is developed. The data for a new potential hire can then be processed to predict in which group a given candidate will occur. The application of data mining models and techniques in Human Resources management is **employee training evaluation** where data mining might offer a solution to evaluate training as a whole. Some of the data stored in HR Information Systems include the content of training programs and records of which courses an employee has taken. By setting up a data-mining program to search for patterns of training activities related to advancement in the organization, a company might uncover data to support further training investments.

The application of data mining solutions in **financial data analysis** refers to most banks and financial institutions. Financial data, collected in the banking and financial industry, undergo systematic data analysis and data mining to improve a company's competitiveness. In the banking industry, data mining is used in the areas of modeling and predicting credit fraud, in evaluating risk, in performing trend analyses, in analyzing profitability, as well as in helping with direct-marketing campaigns. In the financial markets, neural networks have been used in forecasting stock prices, options trading, rating bonds, portfolio management, commodity-price prediction, and mergers and acquisitions analyses. It has also been used in forecasting financial disasters. As the example of organization which uses data mining it is worth to mention US Treasury Department which thanks to the deployment of data mining uncovered more than 400 cases of money-laundering activities, involving more than $1 billion in potentially laundered funds. It is also possible to discover criminal activities that law enforcement in the field would otherwise miss. As the other user of data mining technology can be mentioned American Express where data warehousing and data mining are being used to cut spending. American Express has created a single Microsoft SQL Server database by merging its worldwide purchasing system, corporate purchasing card, and corporate card databases. This allows American Express to find exceptions and patterns to target for cost cutting.

Data mining is also crucial in **retail industry.** The early adoption of data warehouses by retailers has allowed them a better opportunity to take advantage of data mining. The retail industry is a major application area for data mining since it collects huge amounts of data on sales, customer-shopping history, goods transportation, consumption patterns, and service records, and so on. The quantity of data collected continues to expand rapidly, especially due to the increasing availability and popularity of business conducted on the Web, or e-commerce. Retail data mining can help identify customer-buying behaviors, discover customer-shopping patterns and trends, improve the quality of customer services, achieve better customer retention and satisfaction, enhance goods consumption, design more effective goods transportation and distribution policies, and, in general, reduce the cost of business and increase profitability. Almost every type of retailer uses direct marketing, including catalogers, consumer retail chains, grocers, publishers, B2B marketers, and packaged goods manufacturers. The claim could be made that every Fortune 500 company has used some level of data mining in their direct-marketing campaigns. Large retail chains and groceries stores use vast amounts of sale data. Direct marketers are mainly concerned about customer segmentation, which is a clustering or classification problem. Retailers are interested in creating data mining models to answer questions such as: what are the best types of advertisements to reach certain segments of customers, what is the optimal timing at which to send mailers, what is the latest product trend, what types of products can be sold together and what are the significant customer segments that buy products. The examples of data mining systems in retail industry can be Safeway UK company which is a grocery chains and is a big user of data-mining technology applying it to extract business knowledge from its product-transaction data. Safeway is also able to generate customized mailing to its customers *by applying the sequence-discovery function of Intelligent Miner, allowing the company to maintain its competitive edge*. The other example is RS Components UK which is distributor of electronic and electrical components and it also uses the IBM Intelligent Miner to develop a system to do cross-selling and in-warehouse product allocation.

6. Conclusions

All organizations collect different types of data like data connected with employees, financial data, data connected with company's customers etc. The use of the data mining methods and techniques can provide the basis for a competitive advantage of enterprise by allowing it to strategically analyze data needed for the whole process of management. The goal of data mining is to discover interesting and previously unknown information in data sets. It provides valuable, hidden business and scientific "intelligence" from a large amount of historical data. The application of data mining methods and techniques in the enterprise helps in decision making and the entire process is possible thanks to applying computer-based methodology, including new methods and techniques mentioned in this article.

Literature

Berry M., Linoff G., *Data Mining Techniques for Marketing, Sales, Customer Relationship Management*, Wiley Publishing, Indianapolis 2004.

Cabena P. et al., *Discovering Data Mining: From Concept to Implementation*. Prentice Hall, Englewood Cliffs, New York 1998.

Decision Tree Analysis Choosing Between Options by Projecting Likely Outcomes, http://www.mindtools.com/dectree.html, Mind tools, Retreived 12 February 2007.

Hand D., Mannila H., Smyth P., *Principles of Data Mining*, MIT Press, Cambridge, MA, 2001.

Inmon W.H., *Building the Data Warehouse*, John Wileys & Sons, New York 2002.

Jarke M., Lanzerini M., Vassiliou Y., Vassiliadis P., *Hurtownie danych. Podstawy organizacji i funkcjonowania*, WSiP, Warszawa 2003.

Kantardzic M., *Data Mining: Concepts, Models, Methods, and Algorithms*, John Wiley & Sons 2003.

Larose D., *Discovering Knowledge in Data. An Introduction to Data Mining*, John Wiley & Sons, New Jersey 2005.

Nowicki A., Jelonek D., Wydmuch G., Ziora L., *Data Warehouse as the Element of Innovation in the Enterprise. Review of Selected Case Studies. The Challenges of Reconversion*, Wydawnictwo Politechniki Częstochowskiej, Częstochowa 2006.

Simon A.R, Shaffer S.L., *Hurtownie danych i systemy informacji gospodarczej*, Oficyna Ekonomiczna, Kraków 2002.

Tarapanoff K. et al., *Intelligence Obtained by Applying Data Mining to a Database of French Theses on the Subject of Brazil*, ,,Information Research", available at: http://InformationR.net/ir/7--1/paper117.html 2001.

Wang J., *Data Mining. Opportunities and Challenges*, Idea Group Publishing, London 2003.

Wang L., Fu X., *Data Mining with Computational Intelligence*, Springer, Singapore 2005.

Zarządzanie wiedzą w systemach informacyjnych, red. W. Abramowicz, A. Nowicki, M. Owoc, AE, Wrocław 2004.

ZASTOSOWANIE METOD I TECHNIK *DATA MINING* W PRZEDSIĘBIORSTWIE. PRZEGLĄD WYBRANYCH PRZYKŁADÓW PRAKTYCZNYCH

Streszczenie

Celem artykułu jest prezentacja istoty *data mining* (drążenia danych) z ukazaniem jego czteroetapowego procesu oraz praktycznych przykładów zastosowań w przedsiębiorstwie. Ukazane zostały podstawowe zadania *data mining*, takie jak klasyfikacja, estymacja, predykcja, grupowanie, opis wraz z profilowaniem. Artykuł przedstawia również wybrane metody i techniki DM, takie jak drzewa decyzyjne, sztuczne sieci neuronowe, algorytmy genetyczne, wizualizacja danych. Praktyczne zastosowania DM zostały ukazane w takich dziedzinach, jak zarządzanie zasobami ludzkimi, finansowa analiza danych oraz handel detaliczny.