DATA MINING TECHNOLOGY

Georgiana Marin[1]

Abstract

In terms of data processing, classical statistical models are restrictive; it requires hypotheses, the knowledge and experience of specialists, equations, effective knowledge of probabilities distribution and the data must have a high quality, being subject to prior processing and transformations. Because of these disadvantages the concept of data mining has emerged, implementing knowledge extraction algorithms from the large data collections.

Data mining is a process consisting in collecting knowledge from databases or data warehouses and the information collected that had never been known before, it is valid and operational. Nowadays data mining is a modern and powerful IT&C tool, automatizing the process of discovering relationships and combinations in raw data and using the results in an automatic decision support.

Keywords: data mining, data warehouse, knowledge discovery, OLAP, OLAM.

1. What is data mining?

Knowledge discovery and data mining have emerged as an interdisciplinary domain with a fast evolution and merging with *databases*, statistics, data warehouses and willing to extract a big amount of valuable knowledge and information.

The purpose of data mining is to discover unknown new information and due to this fact the results are truly useful. The knowledge discovered through data mining must be valid. Appling the data mining techniques on large amounts of varied data could lead also to false information therefore is essential to check the data validity. We could also refer to the data mining process as a step in discovering the information through a set of algorithms and patterns meaningful in the data structures and showing market trends.

Data mining discovers patterns within data, using predictive techniques. These patterns play a very important role in the decision making because they emphasize areas where business processes require improvement. Using the data mining solutions, organizations can increase their profitability, can detect fraud, or may enhance the risk management activities. The models discovered by using data mining solutions are helping organizations to make better decisions in a shorter amount of time.

[1] PhD Student, Academy of Economic Studies; Teaching assistant, Romanian American University, georgiana.maryn@gmail.com

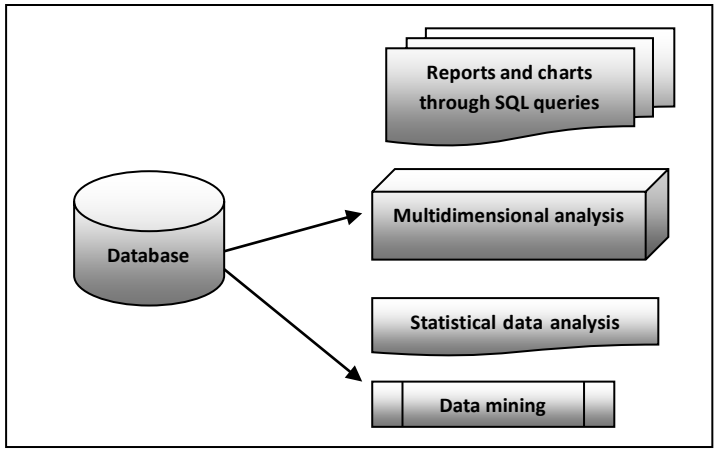


Figure 1 - Differences between traditional data analysis and data mining[2]

Data mining methods derived from statistical calculation, *database* administration and artificial intelligence, SQL queries, analysis in multidimensional databases using OLAP systems. They don't replace the traditional methods of statistics, but are considered to be extensions of graphic and statistical techniques.

Typical data structure suitable for data mining contains the observations placed on lines and the variables placed on columns. Domains or range values for each variable must be precisely defined, avoiding as much as possible vague expressions. Line and column format, similar to the spreadsheet file is required for data mining.

Data mining software is separated into two groups:

1. Data mining tools – are providing techniques that can be applied to any business problems.
2. Data mining applications – incorporate techniques inside an application specially built to address business problems. Our life is influenced by data mining applications. For example, almost any financial transaction is processed by a data mining application to detect fraud. Increasingly more organizations are using both data mining tools and applications to develop predictive analysis.

In order to apply data mining tools, data must go through the following processes:

a. Preprocessing – removing unnecessary data
   - Consistency checking (measurement units);
   - Detecting and removing erroneous information;
   - Removing extreme values (outliners).
b. Data integration – combining variables
c. Transforming variables – through standardization, or by passing to logarithmic scale
d. Separating the database into three categories of data:
   - Training category
   - Validating category

[2]Lungu, Ion and Bâra, Adela. *Sisteme informatice executive.* Bucharest : Editura ASE, 2007

- Testing category

e. Using simple classical descriptive statistic: mean, median, range, standard deviation
f. Using simple diagrams: frequency histograms, bar charts, pie chart.

2. Data mining and knowledge discovery components

The main function of data mining is to extract knowledge patterns from data. Therefore, data mining uses a variety of statistic algorithms, forms recognitions, classifications, fuzzy logic, machine learning, genetic algorithms, and neural networks. The variety of algorithms can be grouped into the main components of data mining.

The main components of data mining are:

1. The model which, like any other computerized model, is represented by a function in single-dimensional or multidimensional space, depending on the parameters. It may be represented either as a linear function of parameters either as a probability or fuzzy function. Different algorithms, such as classification and clustering are leading to the achievement of the model.
2. Preference criteria may have a different nature, some of them being based on ranking and others on interpolation or on the best approximation.
3. Selection algorithms are leading to the selection of three important elements that occur in data mining: the model selected from a model base, data selected from the *database* and setting up the parameters and the preference criteria, selected from a criteria base.
4. Setting residuals generally consist in algorithms of determination of deviation and stability; a particular category of such algorithms are the statistical ones, which are setting the deviations from the ideal model.

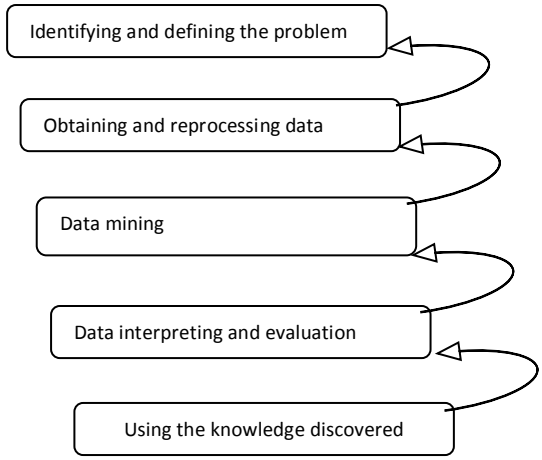
Each commercial product uses several algorithms and in each of them we can find some or all of the above components, in different proportions.

Researches who make the difference between data mining and knowledge discovery consider knowledge discovery as a complex interactive and iterative process, which includes data mining. Thus, it is considered that the knowledge discovery retrieval is accomplished in the following steps:

a. Understanding the applicability domain and the formulation of the problem. This step is an essential condition in extracting relevant knowledge for choosing the most suitable method of data mining for the third stage, according to the destination of the application and the nature of data.
b. Collecting and reprocessing data, including the selection of data sources, removing the outer layers, processing and data reduction.
c. Step three is represented by data mining, the process of extracting models or patterns hidden in data. A model is a global representation of a structure that summarizes the systematic component, underlying the data, or which describes how data can result. A pattern is a local structure, associated with some variables and conditions. The most important data mining methods are classification and predictive regression modeling, clustering, dependences modeling with graphical models and estimation of density.

d. The fourth step is the interpretation or post processing of knowledge found, especially the interpretation in terms of description and prediction, the two main purposes in the discovery system practice. Experience shows that patterns or data patterns are not directly used and the knowledge discovery process is repeated through the knowledge discovered. A standard manner of assessment is to divide data into two sets, working on a data set and testing on the second. We repeat the process a number of times, each time dividing the data differently. The results will be used to estimate the rules of performance.
e. The final step is to put into practice the knowledge discovered. In some cases the new discovered data can be used without the need of an integrated system and, in other cases, it can be used to exploit it through specialized software.

Figure 2 - Knowledge extraction process



3. On-line analytical data mining systems (OLAM)

Data mining and OLAP system are tools for business intelligence. OLAP queries retrieve the database information, at certain levels. OLAP analysis is a deductive process. Based on this hypothesis, data mining is different from OLAP system because it is using its data to discover new patterns. This tool examines the data and interactions between them.
Data mining technology is focused on assessing the predictive power of patterns, this being possible by testing conclusions on a different set of data and by calculating the predictive accuracy. So, we can appreciate that data mining and OLAP are two technologies that complement each other. Data mining could help analyze and design the data warehouse by focusing attention on important variables, identifying exceptions and finding interactions between variables. Due to the interconnection between the two technologies, the OLAM systems have emerged. OLAM systems are also called OLAP systems for data mining. This type of system integrates OLAP multidimensional processing with extracting knowledge form data, in data mining.

Lately, different architectures had been defined but OLAP systems are imposed increasingly due to their advantages:

1. Ensuring a high quality of data in the data warehouse. Most data mining tools require consistent and complete data and therefore expensive processing.

2. Exploiting the data processing infrastructure, available in data warehouses. Processing facilities provided by the data warehouse includes accessing, integration, consolidation and transformation of the heterogeneous databases, multiple Web access, report generating facilities and online analysis.
3. Analyzing data based on the facilities offered by OLAP processing. Data mining techniques require a data analysis, navigating through the database, selecting relevant data, analyzing data in various levels and presenting the results in different forms.
4. Selecting online the processing functions for data mining. Integrating into an OLAP system different data mining functions, the user has greater flexibility in selecting these functions and a faster switch between different types of processing is possible.

An OLAM system must retrieve knowledge in multidimensional data in the same way OLAP systems carries out the data processing. Since the data mining processing in OLAM systems is specific, it requires additional modules that are not found in a basic OLAP system.

4. Conclusion

Data mining tools and applications are helpful in business management, business intelligence, selective marketing, and decision analysis.

Data mining is a technology that uses complex and elaborate algorithms in order to analyze and reveal interesting information useful in the analysis made by decision makers. OLAP organizes data into a pattern suitable for the analysts to operate while data mining carries out data analyses and provides the results to the decision makers. Thus, OLAP enables a model oriented analysis while data mining makes the oriented data analysis easier.

Bibliography

1. Lungu, Ion and Bâra, Adela. *Sisteme informatice executive.* Bucharest : Editura ASE, 2007.
2. Velicanu, Manole, et al., et al. *Sisteme de baze de date evoluate.* Bucharest : Editura ASE, 2009.
3. Chmelar, Petr and Stryka, Lukas. Iterative, Interactive and Intuitive Analytical Data Mining. [Online] 2009.
http://www.fit.vutbr.cz/~chmelarp/public/07znalosti%20i3mining.pdf.
4. Gorunescu, Florin. *Data Mining Concepts, Models and Techniques.* : Springer, 2011.
5. Ullman, Jeffrey D. *Data Mining Lecture Notes.* 2009.
6. Han, Jiawei and Kamber, Micheline. *Data mining: concepts and techniques.* : Morgan Kaufmann Publishers, 2006.