ANALELE UNIVERSITĂȚII
“EFTIMIE MURGU” REȘIȚA
ANUL XVIII, NR. 3, 2011, ISSN 1453 - 7397

Mirela Danubianu, Tiberiu Socaciu

Efficient Selection of Data Mining Method

Data mining tools can access large amounts of data and find patterns that can solve various problems, often with surprising solutions. We have analyzed the data mining methods, techniques and algorithms with their characteristics, with their advantages and weakness. Taking into account the tasks to be resolved in order to discover the different types of knowledge, the kind of databases to work on and the type of data, as well as the area for which on desire the implementation of the data mining system we have try to find a way to efficiently choose the proper methods in a given situation. ExpertDM system has the aim to find the best data mining methods for solving a task and specifying the transformation which need to be made for bringing the data at a proper form for applying these methods.

Keywords: *data mining, expert system,*

1. Introduction

Last years we were witness of the phenomenon of explosive development of information technology, materialized, among other things, in systems with large capacities for collecting and storing data. Due to the availability of powerful database systems millions of databases have been used in business management, scientific and engineering data management or in many other applications. In these conditions, we have a huge volume of data, but, unfortunately we have not, also, a huge volume of information. To achieve this, we need new techniques and tools that can automatically transform the data into useful information and knowledge. Data mining can be viewed as a result of the natural evolution of information technology and it has become common in the 1990’s due to data warehousing, powerful desktop computers and access to large databases via the Internet.

2. What is Data Mining

Many researchers [Gregory Piatesky-Shapiro] treat data mining as a synonym for the term Knowledge Discovery in Databases. In this case a formal definition of Data Mining is: *"Data Mining means a process of nontrivial extraction of implicit, previously unknown and potentially useful information (such as knowledge rules, constraints, regularities) from data in databases."* In other cases data mining is viewed as an essential step in the process of Knowledge Discovery in Databases. This process consist in the following steps: data cleaning to remove noise and irrelevant data, data integration to combine data from multiple sources, data selection and transformation in order to retrieve from the database only the relevant data for the analyze and to consolidate data into forms appropriate for mining, data mining is the phase where the algorithms are applied in order to extract data patterns, pattern evaluation to find the interesting patterns who representing new knowledge and, finally knowledge presentation.

There are two main models for KDD: SEMMA and CRISP-DM. According to CRISP-DM [1], the reference model for this process, KDD consists of a sequence of steps. These steps are presented in Figure 1.

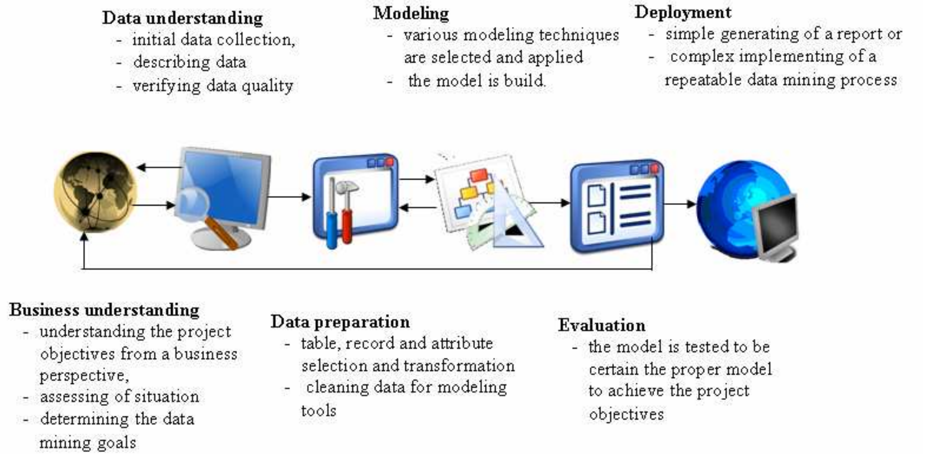
Data understanding
- initial data collection,
- describing data
- verifying data quality
Modeling
- various modeling techniques are selected and applied
- the model is build.
Deployment
- simple generating of a report or
- complex implementing of a repeatable data mining process
Business understanding
- understanding the project objectives from a business perspective,
- assessing of situation
- determining the data mining goals
Data preparation
- table, record and attribute selection and transformation
- cleaning data for modeling tools
Evaluation
- the model is tested to be certain the proper model to achieve the project objectives

Figure 1. CRISP-DM Model for KDD [1]

Data-mining step consist in exploration and analysis, by automate or semiautomatic means of large quantities of data in order to discover meaningful patterns and rules.[2] Consequently, data mining is the search of valuable information in large volumes of data, which combine the efforts of human, who design the databases, describe the problems and set the goals, and computers who process the data looking for patterns that match this goals.

Actually, even the economists consider data mining will become much more important, because, if it provide actionable results that improve business processes, data mining is a competitive weapon.

Data mining techniques

There are different criteria used to categorize data mining methods and systems. Some of them are: the kind of databases to be studied, the kind of knowledge to be discovered or the kind of techniques to be utilized. A data mining system can be classified according to the kinds of databases on which the data mining is performed. For example, a system is a relational data miner if it discovers knowledge from relational data, or an object-oriented one if it mines knowledge from object-oriented databases. In general, a data mining system can be classified according to its mining of knowledge from the following different kinds of databases: relational databases, transaction databases, object oriented databases, deductive databases, spatial databases, temporal databases, multimedia databases, heterogeneous databases, active databases, legacy databases, and the Internet information-base.[2] Several typical kinds of knowledge can be discovered by data mining systems, including association rules, characteristic rules, classification rules, discriminant rules, clustering, evolution, and deviation analysis, which will be discussed in detail in the next subsection. Moreover, data mining can also be categorized according to the abstraction level of its discovered knowledge, which may be classified into generalized knowledge, primitive-level knowledge, and multiple-level knowledge. A flexible data mining system may discover knowledge at multiple abstraction levels.[3]

Among many different classification schemes, this paper uses mainly the kind of knowledge to be mined because such a classification presents a clear image on different data mining requirements and techniques.

Data Mining Methods and Algorithms

Data mining problems can be divided into two general categories: prediction and knowledge discovery. While prediction is the strongest goal, knowledge discovery is the weaker approach. It is usually performed prior to prediction. For example, in a medical application for a disease recognition, which belongs to predictive data mining, we must mine the database for a set of rules that describes the diagnosis. Then this knowledge is used for the prediction of the disease when a new patient comes in. Each of these two problems has some associated methods. For prediction we can use classification or regression while for knowledge discovery we can use deviation detection, clustering, association rules, database segmentation or visualization.

Data classification is a two-step process. In the first step, a model is built describing a predetermined set of data classes or concepts. The model is constructed by analyzing database tuples described by attributes. Each tuple is assumed to belong to a predefined class, as determined by one of the attributes, called the class label attribute. In the context of classification, data tuples are also referred to as samples, examples, or objects. The data tuples analyzed to build the model collectively form the training data set. The individual tuples making up the training set are referred to as training samples and are randomly selected from the sample population. Since the class label of each training sample is provided, this step is also known as supervised learning (i.e., the learning of the model is 'supervised' in that it is told to which class each training sample belongs). It contrasts with unsupervised learning (or clustering), in which the class labels of the training samples are not known, and the number or set of classes to be learned may not be known in advance. Whereas classification determines the set membership of the samples, the answer of *regression* [4][5] is numerical

Cluster analysis is a set of methodologies for automatic classification of samples into a number of groups using a measure of association, so that the samples in one group are similar and samples belonging to different groups are not similar. The input for a system of cluster analysis is a set of samples and a measure of similarity between two samples. The output from cluster analysis is a number of groups (clusters) that form a partition, or a structure of partitions, of the data set. One additional result of cluster analysis is a generalized description of every cluster, and this is especially important for a deeper analysis of the data set's characteristics.

Association rule mining finds interesting association or correlation relationships among a large set of data items. A typical example of association rule mining is market basket analysis. This process analyzes customer buying habits by finding associations between the different items that customers place in their shopping baskets.

The algorithms used in data mining are often well-known mathematical algorithms, but in this case they are applied to large volumes of data and to general business problems. The most used are: neural networks, decision trees, nearest neighbor methods, rule induction and data visualization.

Neural networks simulate the pattern finding capacity of the human brain and hence some researchers have suggested applying this method to pattern mapping. One of the most widespread architectures for neural network, multilayered perceptron with back propagation of errors, emulates the work of neurons incorporated in a hierarchical network. In this case the input of each neuron of the current layer is connected with the outputs of all neurons of the previous layer. The data to be analyzed are treated as neuron excitation parameters and are fed to inputs of the first layer. These excitations of a layer neurons are propagated to the next layer neurons, being weakened or amplified with the weights assigned to corresponding intraneural connection. At the end of this process a single neuron, situate at the topmost neuron layer, acquires some values considered to be a prediction. In order to make accurate predictions a neural network must to be trained on data describing previous situations for which input parameter and correct reactions to them are known. Training consists in selecting weights assigned to intraneural connections that provide the maximal closeness of reactions obtained from the network to the known correct reactions.

Decision trees can be applied for solution of classification or clustering tasks. As result of applying this method to a training set is created a hierarchical structure of classification rules of the type *if…then*. In order to decide to which class an object should be assigned one has to answer questions located at the tree nodes, starting from the root. These questions are of the form: "The value of variable A respect the condition…?" If the answer is "Yes" one follows the right branch of the tree to a node of the next level. If the answer is "No" one follows the left branch. Then a question in this node should be answered, and so on, until on arrive to the final nodes where on finds a class whom the object should be assigned. When decision trees are used for clustering, they divide a population into segments with similar characteristics. The main advantage of this method is that this form of representation of rules is intuitive and easily understood by the human.

K-nearest neighbor technique is most often used for classification, but it can also be used for estimation and prediction. It is an example of instance-based learning, in which the training data set is stored, so that a classification for a new unclassified record may be possible simply by comparing it to the most similar records in the training set.

Association rules algorithms [6] [7] are suited for sequences and associations or affinity tasks. The tools build on these algorithms calculate the support level, which is the percentage of all records when multiple events occurred, and the confidence level, which is the percentage of all transactions when the events from antecedent and the events from consequent occurred together.

For choosing a proper solution of mining a certain knowledge type is needed the comparison of the different approaches with focus of the aspects based on efficiency and scalability, all these seen in the domain context for which hat certain solution is being built.

From this point of view it is needed those techniques that minimize the number and the degree of difficulty of the transformation operations which need to be made in order to obtain results from the date. Thus, in the database that will be mined we can meet the following situations: preponderance of categorial data – data that has value from a certain domain that can not be compared or ordered; preponderance of numeric data- it can be compared and allow arithmetic operations; a large number of fields per record. Most of data mining applications use a single targeted field or dependent variables, and all the other targeted fields are treated as independent variables. Data mining methods have different possibilities of suitable approach of a large number of independent variables, thus, this can be an appropriate way of selecting the right method; multiple target fields, when the prediction of different ways out is needed, based on the same data entrance; records of length variable often meet in the transactional data and free text data that contain valuable information.

3. Expert System

Expert systems are defined as computer application that use expert knowledge to attain high levels of performance in a narrow problem area. Figure 2. present the most important parts of a typical rule-based expert system.

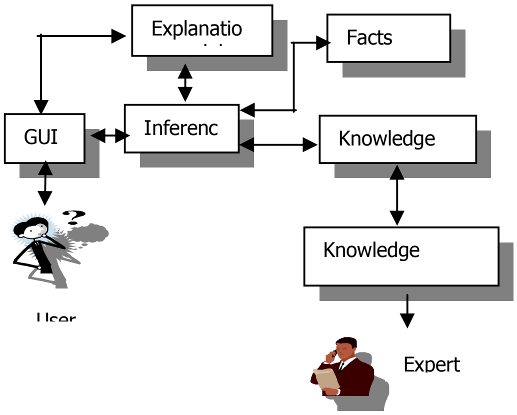

Figure 2 System Architecture

The main elements of this architecture are: the knowledge base which contains the knowledge of the human experts of a specific area: the facts base, which contains the data from a problem to be resolved as well as the facts, resulted from the reasoning made by inference engine over the knowledge base; the inference engine is the module that performs the transformation. It starts from facts, activates the correspondent knowledge from the knowledge base and builds the reasoning which leads to new facts; the explanation module has the role to present in accessible forms the justification of the reasoning made by the inference engine; the knowledge acquisition module transform the human experts knowledge in the appropriate form for the system use and the user graphic interface (GUI) allow to the users to access the expert system

The expert systems are characterized by a declarative approach of the problems. It allows the specification of the knowledge, which is going to be exploited in a dynamic mode and the reasoning mechanism. The most used technique of knowledge representation is the rule-based. A rule has the form "IF condition THEN action". In a rule-based expert system, the domain knowledge is represented as sets of rules that are checked over a collection of facts or knowledge about the current situation. When the IF section of a rule is satisfied by the facts, the action specified by the THEN section is performed. The IF section of a rule is compared with the facts and the rules whose IF section matched the facts are executed. This action may modify the set of facts in the knowledge base.

4. ExpertDM

We have realized a close analyze of the data mining methods, techniques and algorithms with their characteristics, with their advantages and weakness. As well we have taken into account the tasks to be resolved in order to discover the different types of knowledge, the kind of databases to work on and the type of data.

All these aspects as well as the area for which on desire the implementation of the data mining system was been taken into account for the realization of the ExpertDM system which has the aim to finding the best data mining methods for solving a task and specifying the transformation which need to be made for bringing the data at a proper form for applying these methods. Moreover the system can be used to suggest the best algorithm, which could be implemented in a new instrument or in a new data mining system for a specific area. Figure 3 presents the way for obtaining the final information starting from a data entry set.

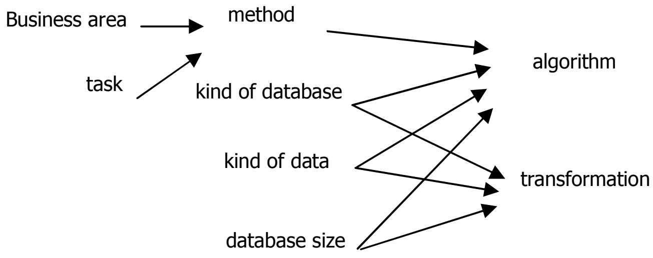

Figure 3.

We have started to implement this system using Clips 6. The knowledge base of the system contains rules as:

- if the business area is retail and the task is space optimization the association rules are used
- if the dates contains taxonomies the generalized association rules algorithms are used
- if the method used is association rules, the database is transactional, the data are categorical and the size of database is medium then the Apriori algorithm is used
- if the method used is association and the data are quantitative, it must transform this data by discrimination into categorial data before applying the appropriate algorithm

5. Conclusion

The aim of this paper is to present a way to choose the best method or algorithm in order to solve a given task. Also we proposed an expert system (ExpertDM) which can select the appropriate method, algorithm and transformation for implementing in a new data mining system. The start point is the business area and the task that must be performed. This system may be extended for suggesting the appropriate commercial product of data mining for a special purpose.

References

[1] Danubianu M., Pentiuc S.G., Tobolcea I, Schipor O.A., Advanced *Information Technology- Support of Improved Personalized Therapy of Speech Disorders*, International Journal of Computers, Communications and Control, . Vol 5 Issue 5, ISSN 1841-9836 (to be published)

[2] Chen M., Han J. , Yu P. *Data Mining: An Overview from a Database Perspective*, IEEE, 1996

[3] Matheus C., Chan P., Piatetsky-Shapiro *System for Knowledge Descovery in Databases*, IEEE, 1993

[4] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (eds.). *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press, 1996

[5] Weiss S. M., Kulikowski C. A.. *Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems.* Morgan Kaufman, 1991.

[6] R.Agrawal, T.Imielinski and A.Swami. *Mining association rules between sets of items in large databases.* In Proc. Of the 1993 ACM SIGMOD Intl. Conference on Management of Data,1993, p.207-216

[7] Lui C.L. *Mining generalized association rules in fuzzy taxonomic structures*. PHD thesis,Hong Kong Polytechnic University, 2001

Addresses:

- Lect. dr. eng. Mirela Danubianu, "Stefan cel Mare" University of Suceava, Str. Universitatii nr. 1, Suceava, mdanub@eed.usv.ro
- Lect. PhD. mat. Tiberiu Socaciu, "Stefan cel Mare" University of Suceava, Str. Universitatii nr. 1, Suceava, socaciu@seap.usv.ro