УДК 004.8; 681.3.06-519.852.6

І.Є. Андрущак[1], Ю.Я. Матвіїв, І.В. Андрощук, А.А. Ящук, В.П. Марценюк[2]
[1]Луцький національний технічний університет, Україна
[2]Академія технічно-гуманістична в Бяльско-Бялі, Польща

ОСОБЛИВОСТІ АЛГОРИТМІВ РЕАЛІЗАЦІЇ DATA MINING В БАЗАХ ДАНИХ

У статті розглядаються і аналізуються особливості технології Data Maning. Детально описуються завдання, які вирішує дана технологія, перераховуються методи вирішення цих завдань. Приділено окрему увагу нечіткій логіці, генетичним алгоритмам і нейронним мережам, продемонстровано процес вирішення завдання методами Data Mining.

Ключові слова: *Data Mining, аналіз даних, нейронні мережі, генетичні алгоритми, статистичні методи.*

И.Е. Андрущак[1], Ю.Я. Матвиив, И.В. Андрощук, А.А. Ящук, В.П. Марценюк[2]
[1]Луцкий национальный технический университет, Украина
[2]Академия техническо-гуманистическая в Бяльско-Бялей, Польша

ОСОБЕННОСТИ АЛГОРИТМОВ РЕАЛИЗАЦИИ DATA MINING В БАЗАХ ДАННЫХ

В статье рассматриваются и анализируются особенности технологии Data Maning. Подробно описываются задачи, которие решает данная технология, перечисливаются методы решения этих задач. Уделено особое внимание нечеткой логики, генетическим алгоритмам и нейронным сетям, продемонстрировано процесс решения задачи методами Data Mining.

Ключевые слова: *Data Mining, анализ данных, нейронные сети, генетические алгоритмы, статистические методы.*

I.Ye. Andrushchak[1], Yu.Ya. Matviiv, I.V. Androshchuk, A.A. Yashchuk, V.P. Marcenuyk[2]
[1]Lutsk National Technical University, Ukraine
[2]Akademia Techniczno-Humanistyczna w Bielsku-Bialej, Poland

PECULIARITIES OF DATA MINING REALIZATIONS IN DATA BASES

In the article the features of the Data Maning technology are considered and analyzed. Detailed description of the problem, which solves this technology, are listed methods for solving these problems. Special attention is paid to fuzzy logic, genetic algorithms and neural networks, and demonstrated the process of solving the problem by Data Mining methods.

Key words: *Data Mining, Data Analysis, Neural Networks, Genetic Algorithms, Statistical Methods.*

Formulation of the problem. In today's business, when companies accumulated huge amounts of data, which are often rather chaotic in their work years, standard reporting tools are no longer sufficient. There is a well-known paradox: the more information about the subject area of business accumulates, the more difficult it is to analyze them more effectively and get meaningful conclusions and results. Nevertheless, this information contains a lot of useful information that can and should be used to optimize business processes and improve the quality of the company's work. To do this, it is necessary to generalize past experience, find regularities, extract rules and apply this knowledge in the management process. Therefore, here we need the mechanisms of constructing analytical models capable of finding non-trivial and, at first glance, non-obvious regularities in large volumes of data. In short, Data Mining (DM) systems are required.

Data Mining is a process of searching for correlations, trends, interconnections and regularities through various mathematical and statistical algorithms: clustering, sub-selection, regression, correlation analysis, time series analysis. The purpose of this search is to present data in a form that clearly reflects business processes, as well as to construct a model by which it is possible to predict the processes critical for business planning (for example, the dynamics of demand for certain goods or services, the dependence of their acquisition on the characteristics of the consumer etc.). The use of DM makes sense in the presence of a fairly large amount of data in the corporate storage (QCD). Data in QCD is a constantly replenished set, unified and uniform for the entire enterprise and allows you to reproduce the picture of its activities at any time.

The DM system "sift" the data ("sifts" through the data), revealing the previously hidden information. However, the market offers Data Mining tools that are able to search for regularities, correlations and trends not only in traditional QCDs, but also in other sets of pre-processed statistics. The dedication of using the Data Mining tool for solving business tasks becomes noticeable fairly soon, with the cost of implementing it can pay off fairly quickly. The main areas in which DM is applied are finance, insurance, manufacturing, telecommunications, e-commerce. Note that it is advisable to use Data Mining

wherever there is a large amount of data. In this article, we do not strive to cover all possible spheres, let's dwell only on several life examples.

Setting up tasks. As you know, the main idea of Data Mining reflects the concept of patterns that reflect the fragments of the multidimensional relationships inherent in the data and the employee in order to identify the dependencies and patterns inherent in the sub-samples of data that are suitable for presentation in a convenient, understandable form. Various methods are used to find these patterns, and their peculiarity is that they are not limited to representations about the structure of the sample and the form of distribution of the values of its indicators.

The cornerstones of Data Mining are classification, modeling and forecasting, based on the use of such methods as decision trees, neural networks, genetic algorithms, evolutionary programming, fuzzy logic, etc. Data Mining methods also often include some statistical methods, such as descriptive analysis, correlation and regression analysis, factor analysis, dispersion analysis, component analysis, discriminant analysis, time series analysis, survival analysis, link analysis, and others.

The following standard types of regularities have been adopted that make it possible to identify Data Mining methods:

- association;
- sequence;
- classification;
- clustering;
- forecasting.

The association determines the degree of interconnection of several events among themselves. The sequence defines the relationship between the events by the principle of determination, i.e. what event precedes to what. Classification is the identification of a certain set of characteristics defining the group to which the object belongs, by learning from the example of the previous stage of classification of objects and determining the rules for such groups. Clustering, unlike the classification, involves not attributing to groups but forming new ones. Forecasting involves the discovery or development of templates that adequately reflect the dynamics of the behavior of the targets over time in order to predict their values in the future on their basis [1].

Analysis of recent researches and publications. Numerous firms accumulate a large amount of data over a long period of time, counting that all of them will undoubtedly help them in making the right decisions. Suppose to find out that at that or another definite period, the consumer acquired some product at the mall 123 - it's not so difficult. But here you need knowledge - knowledge about the fact that, for example, retail outlets 123 and 130 implement product X several times faster than other trading outlets. In this case, we can use different algorithms, analyze data and receive results that will have a beneficial effect on the company's profits. Thus, in Data Mining (DM) there is a necessary set of procedures for detecting such clusters of necessary information about the commercial side.

Let's consider the application of Data Mining. This analysis helps to improve the work of the enterprise, since when applying this analysis, researchers can give a more accurate assessment of the results of events occurring in the firm. Data Mining is widely used in various areas of human activity, such as wholesale and retail trade, healthcare, education, industrial production. [1,2,7].

Recall that Data Mining technology is based on the concept of patterns, which are patterns. As a result of the discovery of these laws hidden from the naked eye, Data Mining problems are solved. Certain types of Data Mining correspond to different types of patterns that can be expressed in a form that is understandable to man.

There is no consensus on what tasks should be assigned to Data Mining. Most authoritative sources list the following: classification, clustering, forecasting, association, visualization, analysis and detection, deviations, estimation, link analysis, summing up.

The purpose of the description, which follows, is to give a general idea of the tasks of Data Mining, to compare some of them, and also to present some of the methods by which these problems are solved. The most common tasks of Data Mining are classification, clustering, association, forecasting and visualization. Thus, tasks are divided into types of information produced; this is the most general classification of Data Mining tasks.

Basic material presentation. The approach to research embedded in Data mining is a very popular data processing method and is used mainly in the processing of large data arrays, for example, in the web environment. For students of specialties and areas of study related to information technology and information processing, the concept of Data mining should be familiar, and it is better to form a basic level of knowledge on this subject, especially when it comes to undergraduates or graduate students. For

this reason, it makes sense to consider the possibilities for using the concept of Data mining as a methodology for educational research.

The development of methods for recording and storing data has led to a rapid growth in the volume of information collected and analyzed. The data volumes are so impressive that it is simply impossible for a person to analyze them independently, although the need for such an analysis is quite obvious, because these raw data contain knowledge that can be used when making decisions. In order to conduct automatic data analysis, Data Mining is used [3].

Data mining is the process of discovering previously unknown nontrivial, practically useful and accessible interpretations of knowledge necessary for making decisions in various spheres of human activity in "raw" data. Data mining is one of the steps of Knowledge Discovery in Databases.

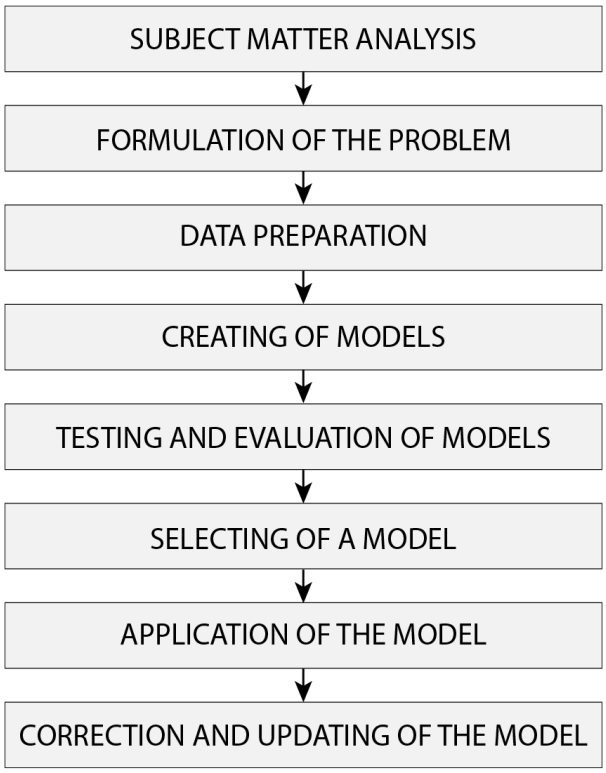
Information found in the process of applying Data Mining methods should be non-trivial and previously unknown, for example, average sales are not. Knowledge should describe new relationships between properties, predict the values of some attributes based on others, etc. Found knowledge should be applicable to new data with some degree of reliability. The usefulness lies in the fact that this knowledge can bring certain benefits in their application. Knowledge should be clear to the user in the form of math. For example, it is easiest for a person to perceive logical constructions "if ... then ...". Moreover, such rules can be used in various SQL queries. In the case when the extracted knowledge is opaque to the user, post-processing methods should exist that allow them to be brought to an interpreted form. Algorithms used in Data Mining require a large amount of computation. Previously, this was a deterrent to the wide practical application of Data Mining, but today's growth in the performance of modern processors has alleviated this problem. Now, in a reasonable time, you can conduct a qualitative analysis of hundreds of thousands and millions of records. Consider approaches to the organization of research activities of students based on the concept of Data Mining.

The task of dividing a set of objects or observations into a priori defined groups, called classes, within each of which they are assumed to be similar to each other, having approximately the same properties and characteristics. In this case, the solution is obtained based on the analysis of attribute values (attributes). Classification is one of the most important tasks of Data Mining. It is used in marketing in assessing the creditworthiness of borrowers, determining customer loyalty, pattern recognition, medical diagnostics, and many other applications. If the analyst knows the properties of objects of each class, then when a new observation belongs to a certain class, these properties are automatically distributed to it.

If the number of classes is limited to two, then a binary classification takes place, to which many more complex problems can be reduced. For example, instead of defining such credit risks as “High”, “Medium” or “Low”, you can use only two - “Issue” or “Refuse”.

Genetic algorithms can be attributed to the universal methods of solving problems of various types. In the field of Data Mining, this is the search for the most optimal model and the determination of significant parameters of the operational basis. The integration of genetic algorithms and neural networks is very effective. This approach allows us to solve the problem of finding the optimal values of the weights of the inputs of neurons. The integration of genetic algorithms and fuzzy logic provides a more optimized system of production rules that are used to control the operators of genetic algorithms.

Consider the process of solving a problem using Data Mining methods. It includes certain steps, which are illustrated in Pic. 1. Data mining process can be both successful and unsuccessful. If the problem solving process was completed unsuccessfully, it may be worth trying to solve the problem using other methods or to change the model parameters.



Pic. 1. **Data Mining Process**

For classification, Data Mining uses many different models: neural networks, decision trees, support vector machines, the k-nearest neighbors method, coverage algorithms, etc., which are built using training with a teacher, when the output variable (class label) is set for each observation . Formally, the classification is based on the division of feature space into regions, within each of which multidimensional vectors are treated as identical. In other words, if an object falls into the area of space associated with a particular class, it is related to it.

The development in the Data Mining sector of the global software market employs both world-renowned leaders and emerging companies. Data Mining tools can be presented either as a standalone application or as additions to the main product. The latter option is implemented by many leaders of the software market. So, it has already become a tradition that the developers of universal statistical packages, in addition to the traditional methods of statistical analysis, include in the package a certain set of Data Mining methods [4-6].

The problems of business analysis are formulated differently, but the solution of most of them comes down to one or another Data Mining task or a combination of them. For example, risk assessment is a solution to a regression or classification problem, market segmentation is clustering, and demand stimulation is an association rule. In fact, Data Mining tasks are elements from which you can assemble a solution to the vast majority of real business problems.

To solve the above problems, various methods and algorithms of Data Mining are used. Due to the fact that Data Mining developed and develops at the intersection of such disciplines as statistics, information theory, machine learning, database theory, it is quite natural that most of the algorithms and methods of Data Mining were developed on the basis of various methods from these disciplines. For example, the k-means clustering procedure was simply borrowed from statistics. The following Data Mining methods have gained much popularity: neural networks, decision trees, clustering algorithms, including scalable ones, algorithms for detecting associative connections between events.

Deductor is an analytical platform that includes a complete set of tools for solving Data Mining problems: linear regression, neural networks with a teacher, neural networks without a teacher, decision trees, search for associative rules and many others. For many mechanisms, specialized visualizers are provided that greatly facilitate the use of the resulting model and the interpretation of results. The strength of the platform is not only the implementation of modern analysis algorithms, but also the provision of the ability to arbitrarily combine various analysis mechanisms [5].

The first one is C4.5 - one of the most popular decision tree generation algorithms. This method processes the input data to determine their class membership. More specifically, in the input data, each object must have a set of attributes, on the basis of which the algorithm determines to which class it can be attributed.

The algorithm, based on the training sample of examples, C4.5 builds a tree, gradually dividing the set into subsets with different values of attributes. Then the resulting subsets are divided further, but the difference of a different attribute is already checked. The procedure continues until the generated set contains examples from the same class, or it is empty. Such a system can be used for decision-making, if each class is assigned a decision, the action that will be applied to each object in it. The disadvantages of the algorithm are that it is inapplicable to fuzzy logic (when examples belong to a class with a certain probability), and also that it needs an initial sample of examples. But nevertheless, decision trees are simply interpreted and have great speed.

Creates k-groups from a data set in such a way that the group objects are the most homogeneous. This is a widely used cluster analysis technique for examining the data set provided. First, let's look at what cluster analysis is. Cluster analysis is a set of algorithms designed to form groups in such a way that the objects of the group are most similar to each other and differ from the elements that are not included in the group. Cluster (the union of several similar elements, which can be considered as a separate unit, possessing certain properties and a group - are synonymous with cluster analysis issues.

Consider a step-by-step (Pic. 2) algorithm for performing the k-means method [8-9]:

1. The k-means method selects positions from the multidimensional space that will represent k-clusters. These elements are called centers of gravity.

2. We will locate each element at the closest distance to one of the points. Through this iteration, several groups are created.

3. Now we have k-clusters, and each object is a member of one of them.

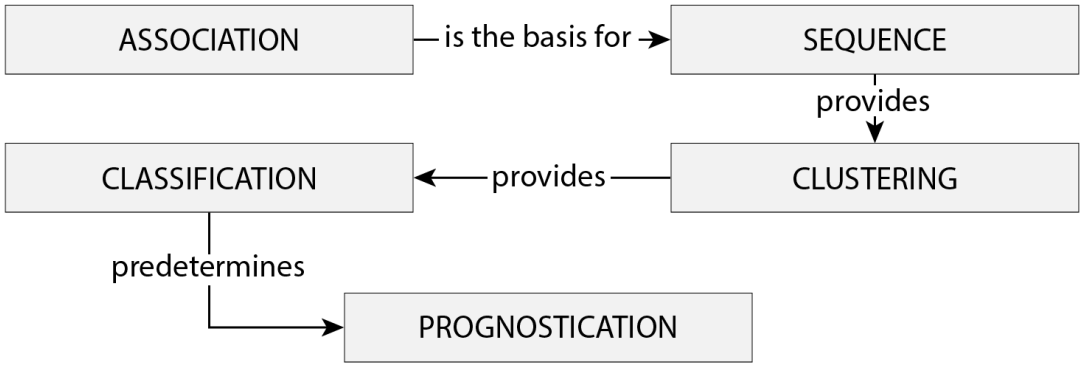
4. The k-means method, taking into account the position of the cluster elements, finds the center of each of the k-clusters.

5. The calculated center becomes the new center of gravity of the cluster.

6. As the center of gravity has moved, the elements can probably shift to be closer to other centers of gravity. Thus, a cluster change may occur.

7. Steps 2-6 are repeated until the center of gravity ceases to change and the groups stabilize. This is called convergence.

The Apriori algorithm finds associative links and applies to each element of a database that contains a large number of different transactions.



Pic. 2. **Data Mining Steps**

Associative rules is a technique used in data mining to study the relationships and relationships between database elements. Let us give an example of the use of associative rules. Suppose we have a database of transactions made daily in the supermarket. Alternatively, such a base is a large table, in it each row is the number of a specific transaction, and each bar is a separate purchase.

Through the use of the Apriori method, we can determine the products purchased together - that is, establish association rules. In this way, we can identify products that are often bought together. The main goal of marketing is to make buyers choose and buy more units of goods. Related units are called sets.

Apriori is usually considered as a self-learning algorithm, so it is often used to detect important elements and required relationships. Currently, a modification of the Apriori method, which is capable of classifying tagged elements, is often used. Apriori is good because it is simple to implement, understandable in explanation and has a large number of modifications.

A significant disadvantage of the algorithm is that in the process of implementation, the algorithm spends a large amount of resources and, as a result, the iterations performed can be performed for a long time. This method is widely applicable. There are a large number of implementations of Apriori. The most commonly used are: ARtool, Weka and Orange.

The support vector machine (SVM - Support vector machine) method is also an algorithm used for classification tasks, but unlike C4.5, it uses hyperplanes instead of trees [10]. Thus, if the initial set of examples can be divided into 2 classes by a certain line, the subsequent objects will be divided into classes, respectively, on one and the other side of this line. In this case, the dividing line must be such that the distance from it to each object is maximum, then the line will be optimal. But it is not always possible to build a reference line, in this case they do this: the elements of the set are placed in a space of higher dimension so that they are separable there, then they search for the optimal hyperplane in the new space. By transferring this method to multidimensional sets, it is possible to distribute data to more classes.

SVM also requires an initial set of examples, moreover, it is poorly interpreted, but the advantage may be that it is a fairly quick method and quite accurate [11].

Algorithm for creating an application that performs Data Mining. Naturally, in our age of modern technology, I would like to automate this algorithm. Next, we describe how to create an application that implements Data Mining. In order to create this product, you should step by step execute the following algorithm:

- find out the scope of the project, describing what information should be obtained as a result. It is important that the plan be focused on the implementation of the necessary entrepreneurial tasks;

- develop a database for Data Mining. The necessary information can be located in several bases, sometimes some of the information is not stored in electronic form. Data from different databases must be consolidated and corrected. In fact, the development of database technology no longer requires the application of DM algorithms to a separate data mart. In fact, effective analysis requires a corporate data warehouse, which, in terms of investment, is cheaper than using separate storefronts. It should be noted that as the introduction of DM projects across the enterprise increases, the number of users grows, and more and more often there is a need for access to large data infrastructures. Modern Storage provides not only an efficient way to store all corporate data and eliminates the need to use other windows and sources, but also becomes an ideal basis for data mining projects. Enterprise repository provides consistent and up-to-date customer data. By implementing the Data Mining functions in the Warehouse, companies reduce costs in two ways. In this case, firstly, it is no longer necessary to acquire and maintain additional equipment for Data Mining. Secondly, the company does not need to transfer data from the Warehouse to special sources for DM projects, while saving time and material resources. Another important point is data cleansing. This means checking integrity and handling missing values. The accuracy of Data Mining methods depends on the quality of the underlying information. Note that the first two stages may take half (or even more) of the time allotted for the entire project;

- give quantitative estimates to data elements. Collaboration with domain experts will help resolve such issues and highlight the data elements that carry the maximum sense from a business perspective.

- apply data mining algorithms to determine the relationship between data. And it is possible that to identify the necessary dependencies will have to use several different algorithms. Some of them will be suitable at the first stages of the process, others at later stages. In certain cases, it makes sense to run several algorithms in parallel in order to analyze data from different points of view.

- explore the relationships identified in the previous stages for applicability across the project. At this stage, you may need the help of an expert in the subject area. He will determine if these or other relationships are too specific or too general, and indicate in which areas the analysis should be continued.

- present the results in the form of a report in which all interpreted relations will be listed. Such a report will bring only one-time benefit, while an application that allows an expert to creatively identify relationships is much more useful. Therefore, the supplier company should not only teach the client how to find dependencies in the data, but also pay special attention to training in working with the program itself [12].

Conclusion. So, Data Mining is a decision support process that is based on the search for hidden data in large volumes of data. This technology has its advantages and disadvantages, but it obviously has prospects for development. At the moment, the most serious drawback is the price. Data Mining facilities

are very expensive software tools and the main consumers are large trading enterprises, banks, insurance and financial companies. However, the gradual popularization of technology should lead to the emergence of more budget software tools that everyone can use. Despite the abundance of Data Mining methods, the priority is gradually shifting more and more towards logical search algorithms in the data if-then rules. With their help, the tasks of forecasting, classification, pattern recognition, database segmentation, extraction of "hidden" knowledge, data interpretation, establishment of associations in the database, etc. are solved. The results of such algorithms are effective and easily interpreted.

At the same time, the main problem of logical methods for detecting patterns is the problem of sorting options in a reasonable time. Known methods either artificially limit such sorting (KORA, WizWhy algorithms), or build decision trees (algorithms CART, CHAID, ID3, See5, Sipina, etc.) that have fundamental limitations on the effectiveness of the if-then rules search. Other problems are related to the fact that the known methods of searching for logical rules do not support the function of generalizing the found rules and the function of finding the optimal composition of such rules. A successful solution to these problems can be the subject of new competitive developments.

References

1. Barseghyan F. Methods and models of data analysis of OLAP and DataMining. / Barseghyan F., Kupriyanov M., Stepaneenko V., Kholod I. - SPb BHV-Petersburg, 2010. – 384 c.

2. Chubukova I.A. Data Mining: a Manual. - M .: Internet University of Information Technologies: BINOM: Knowledge Lab, 2006. - 382 p.

3. Data Mining and Image Processing Toolkit. - http: //datamining.itsc.uah. edu / adam /.

4. Dyuk V. Data Mining: the course (+ CD) /. Dyuk V., Samoilenko A. - St. Petersburg: Izd. Peter, 2001. - 368 pp.

5. IanH. Witten, Eibe Frank, Mark A. Hall, Morgan Kaufmann, Data Mining: Practical Machine Learning Tools and Techniques (Third Edition), ISBN 978-0-12-374856-0

6. Knowledge Discovery Through Data Mining: What Is Knowledge Discovery? - Tandem Computers Inc., 1996.

7. Krechetov N. Products for the analysis of data. - // Market of Software, N14-15_97, c. 32-39

8. Kiselev M. Means of obtaining knowledge in business and finance / Kiselev M., Solomatin E. - / / Open Systems, No. 4, 1997, p. 41-44.

9. Pertrenko A.I. Grid and intelligent data mining. / A.I. Pertrenko // System Research & Information Technologies, 2008, No. 4 97-110.

10. Petrushin V.A. , Khan L. , Multimedia Data Mining and Knowledge Discovery

11. Methods and models for data analysis OLAP and Data Mining / F. Barseghyan, M. Kupriyanov, V. Stepanenko, I. Kholod. - SPb .: BHV. - 2008 - 267 pp.

12. Six of the Best Open Source Data Mining Tools // The New Stack. URL: http://thenewstack.io/six-of-the-best-open-source-data-mining-tools/ (дата обращения: 2.05.2016).

Стаття надійшла до редакції 11.03.2019