CLASSIFICATION OF TASKS OF DATA MINING AND DATA PROCESSING IN THE ECONOMY

Aleksey MINTS[1]
State Higher Educational Establishment "Priazovsky State Technical University", Ukraine

Abstract. *The subject* of the paper is methods of classification of data mining and data processing tasks in the economy, as well as their classification characteristics. *Methodology*. The research used the general scientific methodology of analysis, synthesis, generalization, comparison. Taxonomy methods are used in the compilation of the classification. The selection of actual economic problems is carried out by researching scientific publications on analysis and data processing. *The purpose* of the article is to compile an actual classification of data mining and data processing tasks in the economy and to refine the terminology of this branch of science. *Results*. The classification of data mining tasks is proposed, consisting of four levels. All economic tasks of data mining are divided into two large groups: predictive and descriptive. Each group is subdivided into several classes, which combine tasks with similar taxonomic features. These are the classes of tasks classification, regression, clustering, link analysis, and outlier analysis. Classes of tasks are divided into types. An important criterion for this is the dimensionality of the input data representation. It means the number of neighbours for each individual data element. The data can be presented in the form of series, matrices, and graphs. The methods of analysing each form of data presentation vary considerably. The definition of data processing is clarified and a classification of relevant tasks is proposed. At the top level of classification, the data processing tasks are divided into two groups, depending on whether the order of the elements in the input data changes or not. The main classes of data processing tasks are identified, such as ranking, sorting, filtering, cleansing, recovery, and quantization. *Practical use*. The results of the research can be used to model intellectual decision-support systems. They allow improving the processes of problem formulation in data mining and data processing tasks. Also, it helps in the formalization of the procedures for selecting methods of their solutions. This leads to an increase in economic efficiency.

Key words: data mining, data processing, classification, economical tasks, regression, clustering, link analysis, outlier analysis.

JEL Classification: B41, C10, C45, C80

1. Introduction

Systematization and classification are the most important components of scientific research and in a number of sources are considered as one of the main functions of science. Work on the systematization of knowledge contributes to the development of science and its transition from the empirical to the system level. If the classification is based on objectively existing (natural) signs, then it itself can be an instrument of scientific knowledge and serve to obtain new knowledge and patterns.

At the present time, the theory and practice of intelligent computing are rapidly developing. There are constantly emerging new applications of the data mining and data processing. However, for the classification of these tasks, the methods that are proposed in the last century are still used. They do not take into account many areas of modern research, which complicates the systematization of knowledge in the field of intelligent computing. The current paper aims to compile an actual classification of data mining and data processing tasks in the economy and to refine the terminology of this branch of science. For this, it is necessary to systematize the existing approaches to the classification of data mining tasks. It is needed to select tasks that go beyond the existing classification and to propose natural classification features of data mining tasks in the economy. Then it is needed to create a new classification, taking into account the actual economic tasks. Similar problems must be solved for the data processing tasks classification.

2. Analysis of existing approaches to data mining tasks classification

Depending on the purpose of the actions performed, it is necessary to select the tasks of *data mining* and *data processing*. We will dwell on this classification in detail.

Corresponding author:
[1] Department of Finance and Banking, State Higher Educational Establishment "Priazovsky State Technical University".
E-mail: mints_a_y@pstu.edu

Data mining can be defined as the process of translating data into information (Larose, 2004). In other words, it is extracting the information from the "raw" data useful to the researcher. These data include the relationships hidden in the data array among its various subsets.

By the term *data processing* we mean the processes of data transformation, that is, actions resulting in a different one from the same data array, with the given properties. The aim of these actions is to "improve" the data within the specified criteria.

It should be noted that in the analysis, the dimensionality of the data at the output does not usually depend directly on the dimension of the input sample. At the same time, when processing data, the dimensions of the input and output samples are closely related.

The tasks of data mining are divided into two large groups: *predictive* and *descriptive*.

The *predictive* tasks are related to the construction of a model that can be used to predict the behaviour of the analysed system in a situation that has not previously been observed. They include the tasks of predicting bankruptcy, forecasting financial time series, and many others.

The purpose of solving descriptive problems is to search for hidden regularities in the data, their description, and derivation of rules that can be used in the future to improve work efficiency. Therefore, this group of tasks is also called the tasks of structured data mining. These tasks include a large number of marketing tasks related to the analysis of various target groups and the identification of preferences of their participants, the behaviour analysis problems, and others.

Descriptive and predictive problems should not be considered in isolation. Thus, having studied the patterns of behaviour of a certain system within the framework of solving a descriptive problem, the results obtained can also be used to predict its behaviour.

Currently, data mining tasks usually use different classification options, shown in Fig. 1. In this form, it is used in many studies devoted to the problems of data analysis, for example, by (Gartner Group, 1995), or (Neelamadhab, Pragnyaban, & Panigrahi, 2012), or (Larose, 2004).

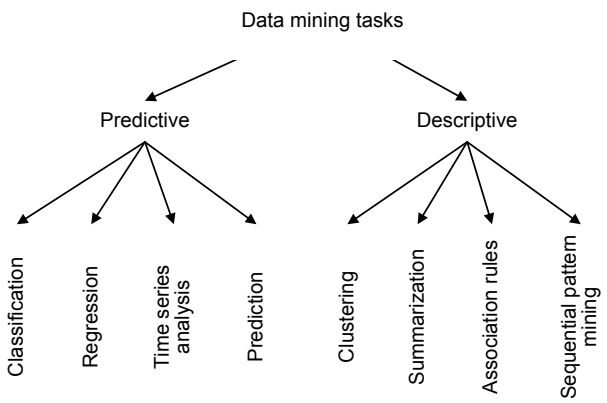

Fig. 1. The generally accepted classification of data analysis problems

Let us consider the set of tasks shown in Fig. 1.

Within the *classification tasks*, the problem of dividing a certain set of data into a predetermined set of classes is solved. The solution of the classification problem allows not only to study the existing data but also makes it possible to predict the future behaviour of the system.

According to the number of classes, into which the input sample is divided, it is necessary to distinguish the problems of *binary* and *polynary* classification. With a *binary classification*, only two classes are allocated in the input sample (more precisely, one class and everything else). The binary classification reduces a large number of economic problems associated with the choice of two alternatives, one of which is interpreted as positive, and the other as negative. For example – to issue or not to issue a loan; to accept or reject the project; to be or not to be.

With a *polynary classification*, the input sample is divided into three or more classes. Examples of such tasks are: pattern recognition, the definition of the borrower's class, the segmentation of clients into well-known classes and the like.

Despite the similarities in the formulations of binary and polynomial classification problems, the methods and tools for their solution vary considerably.

Solving the problem of *regression* involves identifying the relationship between independent (input) variables and dependent (output) ones. The essence of the solution is reduced to the derivation of a mathematical formula by a heuristic or analytical way, expressing the relationship between input and output data.

The example of a regression problem is to determine the amount of credit that can be given to a client.

The aim of solving the *prediction* problem is to approximate the determination of the values of some indicators in the future on the basis of the given values in the past and present.

The examples of tasks of this class are forecasting the growth of the organization, forecasting the implementation of the budget, forecasting the need to recruit new employees, and the like.

The purpose of solving the problem of *time series analysis* is to predict the future values of a certain set of data, where the value of the output variable depends not only on the past values of the variable but also on time. A characteristic feature of time series is the uniform distribution of input data over time. Time series analysis is a kind of regression problem but as it uses specific input data and decision methods, it is allocated to a separate class.

The problems of this class can be, for example, the study of trends in the stock market for forecasting exchange rate fluctuations.

Let us move on to the tasks of the *descriptive group*.

When solving the *clustering* problem, it is required to find regularities in the array of input data, to

allocate a number of zones (clusters) in it and to distribute data over these clusters. The task of clustering resembles the classification problem, with the essential difference that the classes themselves are not defined in advance. To solve the clustering problem, learning algorithms without a teacher are used.

An example of the task of clustering is the grouping of potential consumers of goods in marketing. Another example is the clustering of banks in the country's banking system to analyse their sustainability.

Summarization task has become relevant in connection with the development of the Internet and a sharp increase in the volume of information resources. The essence of it is to reduce a text document to a short summary that retains the most important points of the original document. In particular, such a task is the automatic compilation of articles annotations for issuing them in the results of search systems.

Initially, the *summarization* task was solved solely for test documents but now the scope of its application has expanded and covers all main types of information (images, sound, video). Among its applications is a thematic content search, contextual search, identification of materials that violate the interests of rightsholders and legislative restrictions, etc.

Searching for *association rules* allows you to establish ties and relationships between variables in large databases. Associative rules allow us to find patterns among related events, that is, they give an opportunity to answer the question: "With what probability are events A and B connected?" The sequence of occurrence of events does not matter.

A classic example of the application of search algorithms for associative rules is the analysis of the shopping basket. So, bread and milk, tequila and lemon; diapers and baby food are often bought together. This makes it possible to rationally arrange the goods in the store, which increases its throughput and turnover. Methods of searching for associative rules find application in medical diagnostics, geological prospecting, and other similar tasks, and at the stage of preliminary analysis of data, they find application in economic problems of all varieties.

Sequential pattern mining. Unlike the search for associative rules, sequential pattern mining implies the identification of Cause-effect rules, that is, takes into account the time factor and allows one to answer the question: "With what probability does occurrence of event A entail event B?" Events are usually assumed to be described by discrete values, which distinguishes this task from the task of analysing time series.

Sequential pattern mining is used, for example, in information and financial security systems to detect fraudulent transactions.

3. Clarifying the classification of data mining tasks

The classification shown in Fig. 1 covers a wide range of urgent economic problems. However, the complexity of socio-economic processes and methods of their analysis constantly causes the emergence of new tasks. Thus, in the fundamental research of Han, Kamber & Pei (2012) on the concepts and technologies of data analysis, the problems of identifying fragments in data sequences (subsequences search), searching for data anomalies, the analysis of social networks are considered.

The *subsequence search* function can be considered as a kind of clustering task, as applied to the analysis of dynamic, or time series. Within the scope of the task, it is required to select and to find a sequence of data (fragment) corresponding to another sequence (query). In this case, the dimension of the query and the fragment (the number of elements of the series included in them) can vary significantly.

A variation of the same problem is the search for fractals, that is, sequences of data that have the property of self-similarity (exact or approximate coincidence of data with its part). Fractals can occur in sequences that describe the collective behaviour of people (crowd effect), which is relevant, for example, in the analysis of stock data. Also, these tasks are solved in medical diagnostics, in the modelling of climate changes etc.

Detecting anomalies and outliers. It includes a wide range of tasks related to the finding and identification of such data elements that do not correspond to previously identified patterns. And such anomalies can be both new patterns (for example, new trends in stock data), and signals about the abnormal behaviour of the observation object.

The object of anomalies search can be the results of measurements, time series, text information, graphs.

The problems of finding anomalies are solved within the framework of information security systems, systems for detecting fraudulent transactions with bank cards, medical diagnostics and in many other areas. In addition, they can be used to clear data from noise.

As a matter of fact, the problem of detecting anomalies and outliers is close to the task of identifying significant features, which gives grounds for uniting them under one taxon.

Analysis of social networks is a process of researching social structures, by presenting them in the form of vertices and connections, using the tools of graph theory and network theory.

Classical methods of analysis allow us to analyse the topology of social networks effectively, in particular, to identify key nodes and links, to identify weak links, and to find the shortest paths between nodes. However, in modern conditions, this is not enough since the role of social networks is continuously growing, and,

consequently, the complexity of requests to them is growing.

A significant number of practical problems in the analysis of social networks are reduced to the allocation of participants groups in the network for some generalizing feature. The more abstract this feature is, the more difficult the task becomes.

The tasks of analysing social networks are relevant in economics, marketing, sociology, politics, historical research and many other areas (Hunt, Gentzkow, 2017).

The study and comparison of the analysed data analysis problems made it possible to update the classification shown in Fig. 1. The specified classification is shown in Fig. 2.

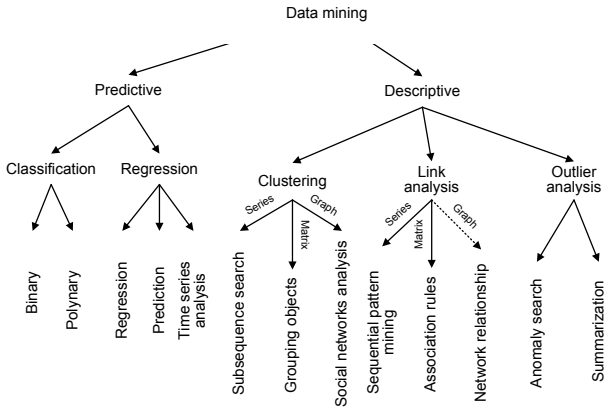

Fig. 2. Clarified classification of data analysis problems

As can be seen from the comparison of Fig. 1 and Fig. 2, the number of grading grades is increased due to the allocation and order of taxonomic features. This allows you to better understand the relationship between different classes of problems and methods for their solution better, which ultimately increases the efficiency of data analysis.

Let us consider the features of the proposed classification.

The dotted link line of the problems of prediction and analysis of time series with a regression taxon emphasizes the existence of regression problems in its pure form as problems of revealing the relationship between input and output variables.

Let us analyse the taxonomic category of clustering tasks. The analysis showed that the tasks of sequence search in data and analysis of social networks are also actually clustering tasks, differing only in the representation of input data and in the dimension of the grouping space.

Thus, subsequence search in dynamic series actually represents a clustering task in a *one-dimensional space*, where each element has only two neighbours – the previous row value and the subsequent one.

Clustering in its pure form is usually called the task of grouping an array of input data, where each element (except for extreme and angular) has a *fixed number of neighbours*. In two-dimensional space, there are four neighbours, in three-dimensional – six, and so on. Input data, in this case, are presented in matrix form.

Finally, the analysis of social networks is a task where each element of the input data (the top of the graph) can have an arbitrary *number of neighbours*, which in real social networks can reach several thousand (and for individual people even more).

Thus, we can conclude that it is the dimension sign of the grouping space (the number of neighbouring elements in the input data) that is a natural classification feature for clustering tasks, which is reflected in Fig. 2.

The tasks of searching for associative rules and identifying cause-effect relationships are placed under the taxonomy of the "Link Analysis" taxonomy. The basis for this is the similarity of tasks. As for the Clustering taxon, the main difference between them is the representation of the input data and the dimensionality of the input data space.

When the cause-effect relationships are identified, the events (analysed elements) are separated in time and, therefore, can be represented as a dynamic series.

When searching for associative rules, the input data set is represented in a matrix form describing events that occur simultaneously or are close in time.

Since the structure of the resulting taxon "Link Analysis" resembles the structure of the Clustering taxon, this gives grounds to assume the existence of another previously unknown problem of link analysis, the input data for which are presented as a graph.

If we formulate its essence by analogy with other tasks of this taxon, it can be assumed that the search should be carried out for epicentres of network activity, that is, nodes, exerting a great influence on the processes occurring in the network or initiating such processes. In a broad sense, this task is reduced to identifying node control and managed nodes for each network. Thus, it can be formulated as an analysis of interactions.

At present, a similar task – the search for a "zero patient" exists in medicine but the methods of its solution provide only a retrospective analysis and cannot be used to predict who can become such a "zero patient" in the future.

In the context of economics, the analysis of interdependence can be used to predict the spread of crisis phenomena, to increase the effectiveness of advertising campaigns, to identify target customers.

Thus, the refinement of the classification of data analysis problems (Fig. 2) made it possible not only to show the relationship between different classes of problems and methods for their solution but also to identify the criterion for the division of tasks into groups within the framework of individual taxa (presentation of input data in the form of a series, matrix, or graph), and also to formulate a problem that was not previously allocated to a separate class.

4. Classification of data processing tasks

A clear classification of data processing tasks is still missing but, from the perspective of the definition proposed above, sorting, filtering, ranking, restoring, cleaning and quantizing data can be classified as such. Let us consider these problems.

Sorting is a task associated with the arrangement of data elements in a given sequence or by grouping them. In the presence of a single sorting criterion or criteria in strict hierarchical submission, the task of sorting data is trivial. Its complexity rises sharply with the complexity of the criteria field structure and its certainty. So, the task of sorting goods in ascending order, or decreasing client preferences, does not have a trivial solution.

The term *filtering* has many meanings in various fields of knowledge (for example, in physics, chemistry, mathematics, radio engineering). With respect to economic data, filtration is considered to be the sampling of information that meets the specified criteria. The initial data for filtering can be represented as a series, an array or a graph. The complexity of filtering increases with the complexity of the criteria field structure and the decrease in their certainty.

The task of *ranking* differs from sorting by the fact that for its solution it is necessary to define a method for determining the rank of each element of the input data sequence, that is, a method that allows the entire vector of element values to be rolled up into a single parameter-rank. This allows you to compare any arbitrary number of elements from the input set, without its full sorting, which is effective for large volumes of processed data. The complexity of ranking increases with the complexity of the criteria field structure. For example, the task of ranking overdue loans in terms of the probability of their return cannot have a unique solution since the methods for determining this probability are themselves based on assumptions.

The need for *data recovery* arises if the input sample contains omissions or some data is missing in it but there are hypotheses about the nature of their occurrence, allowing estimating the most probable values. The solution of this problem allows increasing the efficiency of machine learning by expanding the input data sample. The necessity of data recovery can arise when they are presented in any of the considered forms – in the form of series, matrices or graphs. In the latter case, the object of reconstruction is the connection between the vertices of the graph that are absent in the input sample.

The *data cleansing* task, in fact, is inverse, in relation to the filtering task and involves the exclusion of "extra" data from the input sample.

Purification of the *series* involves the elimination of "emissions", that is, data clearly beyond the main trend. The cause of such emissions can be both measurement errors and conscious distortions of information. In any case, the distorted data strongly influence the quality of the subsequent analysis, especially when using only formal methods, including machine learning.

Cleaning of the data presented in the *matrix* form is performed to eliminate factors that have little effect on the output indicators from the input sample of data or are completely unrelated to them. This procedure must necessarily precede the analysis of data.

With respect to *graphs*, purification provides for the decimation of links and is applied both to input data and to certain types of economic-mathematical models, for example, to artificial neural networks. This eliminates weak links that have little effect on the overall result but significantly complicate the analysis.

Quantization is used in relation to dynamic series of data, if necessary, to reduce the number of their elements, that is, with the simplification of the series. There are such types as quantization by level and time quantization.

In the first case, the range of values of a continuous or discrete quantity is divided into a finite number of intervals. If for a certain period of time the value of the dynamic series values did not exceed the limits of one interval, then as a result of quantization all these quantities will be replaced by a single value (D'yakonov, Kruglov, 2001).

In the second case, the partitioning of the dynamic series into intervals occurs over time. The values accepted by the elements of the series within the boundaries of each interval are replaced by one averaged value or several ones reflecting the limit values of the indicator for the period. The latter method is used to represent stock data.

There are also varieties of quantization according to the level and time with a variable quantization step (Mints, 2016).

Thus, the concept of "data processing" unites a wide range of tasks, which vary in complexity from trivial to those, for which obtaining an exact solution is almost impossible. Taking into account that the latter can occur in each of the given classes, we can talk about the tasks of

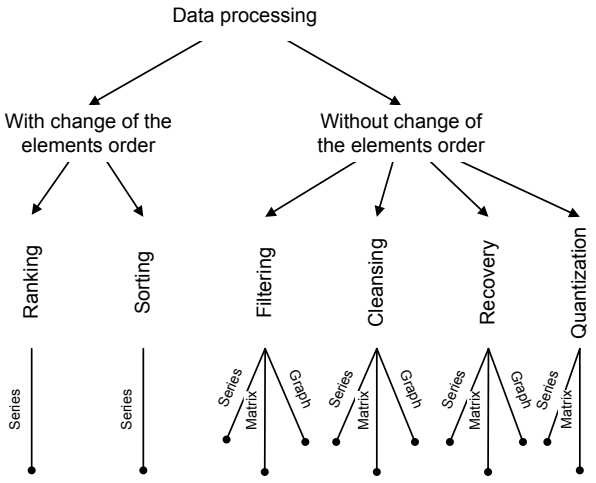

Fig. 3. Classification of tasks of intellectual data processing

intellectual data processing, the classification of which is represented as follows (Fig. 3).

As can be seen from the analysis of Fig. 3, the main classification feature suggests a sign of changing the order of the input data sample elements.

Changing of the elements order occurs while solving the tasks of ranging and sorting. Input data are usually presented in the form of series or are reduced to them.

In other tasks being concerned, there is no change in the order of the input data elements. In this case, it should be noted that methods for solving the problem with respect to data represented in the form of series, matrices, and graphs differ significantly.

It should be noted that the quantization problem, as discussed above, is currently formulated exclusively with respect to dynamic data series. However, its location in one taxon with problems that can be solved with respect to matrices and graphs suggests that it is possible to consider quantization problems with respect to other forms of data representation.

5. Conclusion

Thus, the analysis of data mining and data processing tasks allowed us to classify them according to the main taxa. It should be noted that, in view of the dynamic development of this area, it is not possible to form a complete and consistent classification of analysis and data processing tasks. Therefore, the results obtained should be considered as plausible hypotheses about the representation of these problems.

References:

D'yakonov, V., & Kruglov, V. (2001). Matematicheskiye pakety rasshireniya Matlab. Spetsial'nyy spravochnik. St. Petersburg, Russia: Piter, 480 p. [in Russian].

Gartner Group Advanced Technologies and Applications Research Note (1995). Evolution of data mining. Retrieved from: http://www.thearling.com/text/dmwhite/dmwhite.htm

Han, J., Kamber, M. & Pei, J. (2012). Data Mining: Concepts and Techniques. Waltham, MA: Morgan Kaufmann, 740 p.

Hunt, A., Gentzkow M. (2017). Social Media and Fake News in the 2016 Election. Journal of Economic Perspectives, 2(31), p. 211–236.

Larose, D. T. (2004). Discovering Knowledge in Data: An Introduction to Data Mining. Hoboken, New Jersey: Wiley & Sons, 240 p.

Mints, A. Yu. (2016). Metod uproshcheniya dinamicheskikh ryadov s ispolzovaniyem geneticheskikh algoritmov. Ekonomichnyi visnyk zaporizkoi derzhavnoi inzhenernoi akademii (Economic Bulletin of Zaporozhye State Engineering Academy), 4, p. 120–124 [in Russian].

Neelamadhab, P., Pragnyaban, M., & Panigrahi, R. (2012). The Survey of Data Mining Applications and Feature Scope. International Journal of Computer Science, Engineering and Information Technology, 3, p. 43–58.

Алексей МИНЦ

КЛАССИФИКАЦИЯ ЗАДАЧ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА И ОБРАБОТКИ ДАННЫХ В ЭКОНОМИКЕ

Аннотация. *Предметом исследования* являются методы классификации задач анализа и обработки данных, а также их классификационные признаки. *Методология*. В исследовании использована общенаучная методология анализа, синтеза, обобщения, сравнения. При составлении классификации использованы методы таксономии. Выделение актуальных экономических задач производилось путем исследования научных публикаций по анализу и обработке данных. *Целью* статьи является составление актуальной классификации задач анализа и обработки данных в экономике, а также уточнение терминологии этой отрасли науки. *Результаты*. Предложена классификация задач анализа данных, состоящая из четырех уровней. Задачи анализа данных задачи делятся на две большие группы: предсказательные и описательные. В рамках каждой группы выделены классы, объединяющие задачи со схожими таксономическими признаками. Это классы задач классификации, регрессии, кластеризации, анализа связей, анализа отклонений. Классы задач делятся на виды. Важным критерием для этого является размерность представления входных данных, то есть количество соседей у каждого отдельного элемента данных. Данные могут быть представлены в виде рядов, матриц и графов. Методы анализа таких данных существенно различаются. Уточнено определение обработки данных и предложена классификация соответствующих задач. На верхнем уровне классификации задачи обработки данных разделены на две группы, в зависимости от того, происходит изменение порядка элементов во входных данных, или нет. Выделены основные классы задач обработки данных, такие как ранжирование, сортировка, фильтрация, очистка данных, квантование. *Практическое применение*. Результаты исследования могут быть использованы для моделирования интеллектуальных систем принятия решений. Они позволяют усовершенствовать процессы постановки задач анализа и обработки данных, а также формализовать процедуры выбора методов их решения. Это ведет к повышению экономической эффективности.