MATHEMATICAL METHODS OF MULTIDIMENSIONAL DATA INTELLIGENT ANALYSIS

Ismailov V.,
PhD student, Department of Information systems software
Dniprovsky State Technical University, Kamyanske, Ukraine
Yalova K.
PhD, Associate Professor, Department of Information Systems Applications,
University of information technology and management, Rzeszów, Polska
https://orcid.org/0000-0002-2687-5863
https://doi.org/10.5281/zenodo.10975075

Abstract

The presented article is of an overview nature, the purpose of which is to describe the main methods that can be used to analyze and visualize the multidimensional and big data. Multidimensional data analysis provides a deeper and more complex analysis of the data, exploring the relationships and dependencies between different aspects of the data. There are a lot of mathematical methods, approaches and algorithms that can be used to provide the analysis of the multidimensional data. Mathematical methods and information technologies using allow to process big data and present them in a convenient form for making effective management decisions in different data domain. The results of the multidimensional data analysis also can be used to construct generalized strategies for effective decisions making process.

Keywords: intelligent analysis, mathematical methods, multidimensional data, data cube

Introduction. An integral part of the mankind existence is analysis. Nevertheless, the human brain is limited in the perception of big and multidimensional data, and the modern world dynamics of the processes is the cause of numerous problems. The impossibility of their trivial solution leads to use of mechanisms for processing big data and extracting meaningful information from data domain. For these reasons, some of the tasks can be executed with intelligent analysis methods and due to automation of the information processing processes. Particularly difficult to solve are the problems associated with the data domain, where the inputs have a large number of variable parameters and depend on the environmental conditions changing in an indeterminate way. This type of task includes making a decision on planning organization purchases and the budget, forecasting the change in payables and receivables, forecasting the exchange rate or even adapting the automatic transmission operation of the car. Therefore, the search for effective ways to apply intelligent methods of multidimensional data analysis is a modern scientific and practical task with a wide space for use.

The origins of intelligent data analysis can be traced back to the late 80s, when the term began to be used in the research community, although then there was no understanding covering the term. The sources for processing at that time were tabular data, and the main problem was the limitation in the calculation productivity and the amount of the databases queries. In addition, until the early 90s, data analysis was generally recognized as a sub-process in a knowledge mining process in databases [1]. Now with the growth of computing productivity, this is not a cause for concern. More important is a development and improvement of data processing methods and algorithms. It is possible to distinguish several algorithms among them, that are widespread when trying to intelligent analysis, namely: the k-mean clustering method developed by H. Steinhaus in the 50s, the decision tree construction algorithm created by R. Quinlan, the algorithm for associative rules searching in apriority data, which was proposed by R. Agrawal and R. Srikant a method of reference vectors used for classification and regression analysis, authored by V. Vapnik, a method for visualizing multidimensional data created by T. Kohonen, and other. Based on even this, a small list of approaches to data processing, it is possible to say that there is a well-developed set of available methods applied to an increasing variety of data. Another driver for research in the field, however, is the ever-growing size of the data, requiring new practices for data analysis. Therefore, despite the large number of works and scientists, this issue remains an actual scientific and practical task and needs further development.

Results. Multidimensional data is data consisting of one or more numbers or one or more variables. Analysis of such data combines basic methods of data mining with multidimensional analysis based on real-time analytical processing [2]. Multidimensional model construction can be provided for the different data domain. There are a lot of approaches, mathematical methods and algorithms that are used to analyze and visualize multidimensional data, namely [3]:

- multiple regression, that is used to predict events and involves creating a contingency table;
- multivariate analysis of variance (MANOVA), that allows to analyze the relationship between dependent and independent variables. The analysis performs regression and variance analysis for multiple dependent variables on one or more factor variables or covariates.
- discriminant analysis (DA), which is used for classification in homogeneous groups and for data analysis when the dependent variable is categorical and the predictors (independent variables) are interval. It helps to decide which variables discriminate between two or more emergent aggregates (groups).
- factor analysis that is used to identify dependencies between variables and determine which of them are

most significant and which do not affect the overall picture. There are some terms used in factor analysis: the factor is a kind of hidden variable that explains the relationship between the set of variables; factor load is a factor that shows how strongly each indicator affects a given factor; a factor space is a multidimensional space in which each variable is represented by a factor.

Multidimensional data analysis is a difficult task, since it becomes necessary to carry out multiple correlations or pairwise comparisons, also unnecessary predicts are appeared. Therefore, the problem of reducing the dimensionality of data is an integral part of solving the problem of multidimensional data analysis. Reducing the data dimension allows you to extract only the most significant of them, however, reducing the dimension leads to the partially data loss. The most popular mathematical methods of reducing data dimensionality are:

- principal component analysis (PCA) is one of the most commonly used methods of dimensionality reduction, which allows you to rotate the coordinate system of a multidimensional space so that the first axes describe the greatest data distribution, and the last ones – the smallest. In this case, the last axes can be discarded;
- exploratory factor analysis (EFA) is an unrestricted analysis, used to find factors that provide the best correlation between variables;
- confirmatory factor analysis (CFA) is used to separate errors from the true values when generating the measurement model;
- cluster analysis that is used to simplify objects and divide them into different sets and groups.

To improve the visibility of multidimensional data and the results of their analysis, specialized mathematical methods are also used, for example, the most famous of them:

1. Multidimensional scaling (MDS). The result of the multidimensional scaling method is to construct the data distribution so that the distance between objects corresponds to the values given in the distance matrix [4]. The coordinate axes that occur in this case can be interpreted as some factors which values determine the differences between objects. The main dimensions can be determined by the method of the main components. The multidimensional scaling method can be implemented as a linear and non-linear. Linear multidimensional scaling method uses the method of the main components not for the output distances matrix, but for the double-centered matrix, that has 0 value as an average in each row and column. Double-centered matrix is calculated using the output data. After that, it becomes possible to determine the dimension of the space, which provides accurate reproduction of the distances matrix of. In non-linear methods, the dimension of the space is first given and the quality functional is optimized using gradient methods, describing the measure of the error of the distances matrix. Mathematically the main idea of MDS can be described as follows [5]. Let's assume there are points $x_1, x_2, \dots, x_n$ in the dimensions space k. The distance between points x_i and x_j is defined as $\delta_{i,j}$. Let's find points $y_1, y_2, \dots, y_n$ in a space of smaller dimension so that the distance between them, specified as $d_{i,j}$, is found according to the $\delta_{i,j}$. The ideal situation is when there is a direct correspondence $d_{i,j} = \delta_{i,j}$, but in the non-trivial case, when moving to a space of smaller dimension, it is impossible to get such equality. The problem can be solved with method that find the dependence di,j on i,j as a minimization of some target function. Also it is possible to use minimization of the standard deviation using the formula:

$$J(y) = \frac{\sum_{i<j}(d_{i,j} - \delta_{i,j})^2}{\sum_{i<j}\delta_{i,j}^2}$$

In this case the gradient descent can be used. Multidimensional data scaling mapped to the plane does not solve all the issues of input data visualization. The actual task is to map not only the data points themselves to a two-dimensional or three-dimensional map, but also represent various information accompanying the initial data, for example:

- the positions of points in the original space;
- the number of different subsets;
- other continuously distributed values specified in the original feature space.

The main disadvantage of MDS is the complexity in data interpretation.

2. Method of t-Distributed Stochastic Neighbor Embedding (t-SNE), that is mostly used for visualization multidimensional data with the complex nonlinear relations between variables in the space with the lower dimension [6]. The t-SNE algorithm works by giving each data point a suitable location on a two-dimensional or three-dimensional maps. It does is good in fixing the local structure of multidimensional data and detecting global structures such as clusters at different scales. The t-SNE algorithm begins by transforming the distances between data points in multidimensional Euclidean space into conditional probabilities reflecting the degree of similarity [7]. The degree of similarity between the x_j data point and the x_i data point is expressed by the conditional probability $p_{j|i}$, which means that the x_i point selects the x_j point as its neighbor, proportional to their Gaussian density, centered at x_j. For near data points, the probability $p_{j|i}$ is high, whereas for separated data points, this probability is almost infinitesimal. Conditional probability, is defined as:

$$p_{j|i} = \frac{\exp(-||x_i - x_j||^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-||x_i - x_k||^2/2\sigma_i^2)}$$

where σ_i – is the variance of the Gaussian coefficient centered at data point x_i.

If the points on the map, marked as y_i and y_j, correctly reproduce the degree of similarity between high-dimensional data points x_i and x_j, then the conditional probabilities $p_{i|j}$ and $q_{i|j}$ will be the same. By this t-SNE attempts to find a low-dimensional data representation that minimizes the discrepancy between the $p_{i|j}$ and $q_{i|j}$.

In addition to mathematical methods that allow formalizing multidimensional analysis, information technologies and tools with built-in constructors are widely used for the purposes. For example, one modern way to conduct multidimensional analysis is the OLAP (online analytical processing) technology [8], that consists in preparing summary information based on large amounts of data structured according to a multidimensional principle. In this context, an OLAP cube is a multidimensional data aggregate, and an OLAP system is a pivot table interface for examining a cube using a query language or graphical user interface. A multidimensional data model is usually organized from a given data domain that represented by a table of facts and dimension tables. Facts are numerical measures by which it is possible to analyze the relationships between dimensions, also it can be presented as the numerical data of a business. In general, dimensions can be characterized as points of view or entities within a given data domain. Tables of facts are used to store data domain numeric data [9].

Each dimension may have an associated table called a dimension table, which describes the dimension more detailed, with context and background information. In general, these data can be considered in the form of a cube, which allows to model, search, analyze, and present data in several dimensions. Although cubes usually presented as three-dimensional geometric structures, in a data storage the data cube is multidimensional [10].

In the paper there is the result of the multidimensional model development. The created multidimensional model contains data with the average results of an admission company in educational institutions of Ukraine with parameters: years, regions, exam subjects and specialties for applicants. Selected parameters allow to track changes in subjects scores, analyze data by regions and specialties.

For example, a dimension table for specialties can contain attributes: name, form of training, type of bid. Dimension tables can be defined by users or experts, or automatically generated and adjusted based on data distribution.

Tabular data representation has two-dimensional data, for example as it is shown in Table1.

Table 1.

Two-dimension tabular data representation

Dnipropetrovsk region				
Year (time dimension)	*Subject dimension*			
	Maths, average exam score	*Ukrainian language, average exam score*	*Foreign language, average exam score*	*Geography, average exam score*
2020	*167*	*158*	*132*	*144*
2021	*189*	*170*	*167*	*165*
2022	*167*	*140*	*155*	*163*
2023	*179*	*177*	*143*	*179*

Sourse: own work

Table 2 presents two-dimensional data with the average results of an admission company in educational institutions for Dnipropetrovsk region with using time and subject dimensions. Data can be presented two-dimensional tabular form, including the third dimension, for example regions, as it is shown in Table 2.

Table 2.

Three-dimensional tabular data representation

YYear	*Regions dimension*											
	Dnipropetrovsk region				*Ternopil region*				*Sumy region*			
	Subject dimension											
	Maths	*Ukrainian language*	*Foreign language*	*Geography*	*Maths*	*Ukrainian language*	*Foreign language*	*Geography*	*Maths*	*Ukrainian language*	*Foreign language*	*Geography*
2020	*167*	*158*	*132*	*144*	*143*	*179*	*133*	*145*	*167*	*169*	*143*	*187*
2021	*189*	*170*	*167*	*165*	*154*	*145*	*164*	*162*	*169*	*171*	*145*	*134*
2022	*167*	*140*	*155*	*163*	*144*	*166*	*159*	*153*	*147*	*155*	*167*	*176*
2023	*179*	*177*	*143*	*179*	*174*	*180*	*145*	*136*	*169*	*173*	*176*	*154*

Sourse: own work

The three-dimensional data in Table 2 is presented as a series of two-dimensional tables. It is also possible to represent the same data in the form of a three-dimensional data cube, as it is shown in Figure 1.

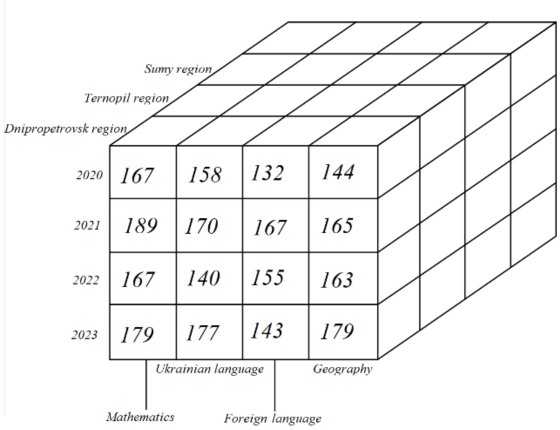

Figure 1. Three-dimensional graphical data representation using data cube

As data cube in the multidimensional data analysis may have more than three dimensions, it can be presented in the form of 4D data cube. For this purpose, it is possible to add one more dimension, for example «specialties». Graphical representation in such a case starts to be complex, but the data cube can be displayed as the set of 3D data cubes, as it is shown in Figure 2.

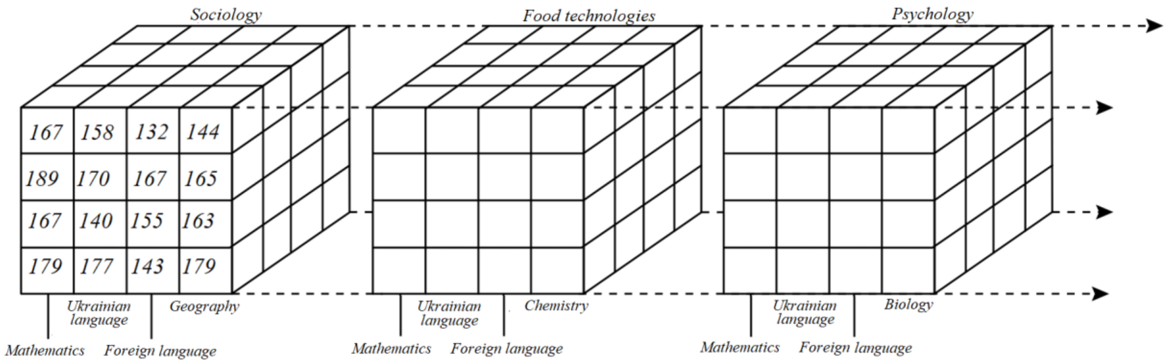

Figure 2. Four-dimensional graphical data representation using data cube

If to continue add dimensions, any n-dimensional data as a matrix of $(n - 1)$ -dimensional cubes can be constructed. At the same time the actual physical storage of such data may differ from their logical graphical representation. Using a matrix of dimensions, a set of cubes or cuboids for each of the possible subsets of given dimensions can be built for the further analysis. The result forms an array of cuboids, each showing data at a different level of summation, aggregation or grouping.

Software implementation of the described OLAP cubes can be provided, for example, with Power BI Desktop using SQL Server Analysis Services or using Python and the framework cubes.

Conclusion. In the paper the task of the multidimensional analysis is discusses. Nowadays all data domains have complex, big, multidimensional data, that are used to ensure the effectiveness of the control operations and the decision making process. Mathematical methods are used to analyze data in different aspects. To simplify the perception of data multidimensionality and to increase the clarity of multidimensional analysis, mathematical methods can also be used, for example, such as: Multidimensional scaling or Method of t-Distributed Stochastic Neighbor Embedding. Online analytical processing technology is used to analyze and visualize the multidimensional data as well as the mathematical methods.

References:

1. Stefan R.-M., Serban M., Rudareanu C. Multidimensional data analysis – representation, security and management. Procedia Economics and Finance (2014), 16, 281-287.
2. Adnan K., Akbar R. An analytical study of information extraction from unstructured and multidimensional big data. Journal of Big Data (2019), 6, 1-38.
3. Liu B., Han T., Qi L., Ji F. Incorporating Multidimensional Data Analysis Methods for Probability Theory and Mathematical Statistics Teaching Reform and Practical Exploration. Applied Mathematics and Nonlinear Sciences, (2024), 9(1) 1-19
4. Delicado P., Gracia P. Multidimensional Scaling for Big Data. ARXiv:2007.11919v4 (2024), 1-24.
5. Jahan S. Discriminant analysis of regularized multidimensional scaling. Numerical Algebram Control and Optimization (2021), 11, 255-267.
6. Tripathy B.K., Anveshrithaa S., Ghela S. t-Distributed Stochastic Neighbor Embedding (t-SNE). In book: Unsupervised Learning Approaches for Dimensionality Reduction and Data Visualization (2021), pp.127-135.

7. Chen W., Wang H., Zhang Y., Deng P., Luo Z. T-Distributed Stochastic Neighbor Embedding for Co-Representation Learning. ACM Transactions on Intelligent Systems and Technology, (2023), 15, 1-18

8. Sarsimbaeva S., Dimitrov V. Multidimensional data analysis with OLAP. CEUR (2019), 2464, 94-98.

9. Kovacic I., Schetz C., Neumayr B., Schrefl M. OLAP patterns: a pattern-based approach to multidimensional data analysis. Data and Knowledge Engineering (2021), 1167-1174

10. Francia M., Gallinucci E., Golfarelli M., Marcel P., Peralta V., Rizzi S. Describing Multidimensional Data Through Highlights. CEUR Proceedings (2022), 3194, 1-8.