ПРИРОДНИЧО-НАУКОВА ОСВІТА: РОЗРОБКА ТА ВПРОВАДЖЕННЯ ІННОВАЦІЙНИХ ПРОЄКТІВ, ПРОГРАМ, МЕТОДИК ТА ТЕХНОЛОГІЙ

УДК 004.02;004.6 DOI: 10.32626/2307-4507.2023-29.7-10

Tetiana PYLYPIUK[1], Viktor SHCHYRBA[2]
Kamianets-Podilskyi National Ivan Ohiienko University
e-mail: [1]pylypyuk.tetiana@kpnu.edu.ua; [2]shchyrba.viktor@kpnu.edu.ua;
ORCID: 10000-0002-4676-9830; 20000-0002-2520-5825

DATA MINING METHODS

Annotation: research is devoted to Data Mining methods. A comparison of classical and mathematical and statistical methods of data analysis was made. One of the variants of correlation analysis method for intelligent data analysis is proposed and described in an argumentative manner. The question of applying different methodologies for Data Mining is actual.

Classically, the following methods of knowledge discovery and analysis are offered in Data Mining: classification; regression; forecasting time sequences (series); clustering; association.

As mathematical and statistical methods of analysis in applied research, the most of authors offer such methods as: statistical hypothesis testing, regression models construction and research. Since most real models are not amenable to analysis using classical methods, including regression analysis, the authors propose to use correlational analysis method in Data Mining.

Key words: Data Mining, analysis methods, statistical hypothesis testing, regression models, correlation analysis.

I. Introduction

The concept of intelligent data analysis (Data Mining) first appeared in 1978 and gained high popularity in the modern interpretation from about the first half of the 1990s. Data processing and analysis were carried out using applied statistics methods until now, and at the same time, the tasks of processing small databases were mainly solved. The basis of modern Data Mining technology is the concept of patterns (templates) that reflect fragments of multifaceted relationships in data. These templates are regularities that are characteristic of data subsamples and can be compactly expressed in a human-understandable form. The search for templates is carried out by methods that are not limited by a priori assumptions about the sample structure and the type of values distribution of the indicators which are analyze [2].

Large volumes of data are generated at modern enterprises, in research projects and on the Internet. Therefore, there is a need for in-depth data analysis and, accordingly, the application of certain methods (technologies), as well as the selection of certain software for data analysis and interpretation of the obtained results.

II. Mathematical and statistical methods application

According to the educational edition, the essence and purpose of the Data Mining technology can be characterized as follows: it is a technology designed to search for non-obvious, objective and practically useful patterns in large volumes of data [2]. Non-obvious – this means that the patterns found cannot be detected by standard information processing methods or by expert means. Objective – this means that the revealed regularities fully correspond to reality, unlike expert opinion, which is subjective most of the time. Practically useful – this means that the conclusions have a concrete meaning that can be used in practice.

Therefore, the question of applying different methodologies for Data Mining is actual.

Classically, the following methods of knowledge discovery and analysis are offered in Data Mining: classification; regression; forecasting time sequences (series); clustering; association [3].

The authors of [2] propose mathematical methods for solving such problems:

- mathematical statistics: regression, cluster analysis, method of main components, statistical hypotheses testing;
- artificial neural networks, including deep learning methods; genetic and evolutionary optimization algorithms;
- methods of dependencies identification and “machine learning” (machine learning); methods of artificial intelligence and data mining technologies;
- various modeling methods (including non-classical and empirical) [2].

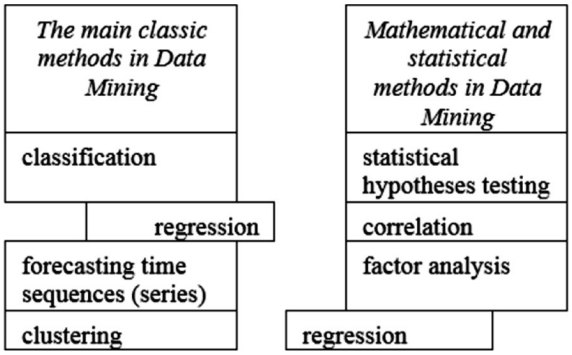

Fig. 1. *Comparison of classical and mathematical and statistical data analysis methods*

As mathematical and statistical methods of analysis in applied research, the authors of the educational edition [1] offer such methods as: statistical hypothesis testing, correlation analysis, regression models construction and research.

As we can see from *Fig. 1,* regression analysis (regression, construction of regression models and their research) is a method that is most often offered both in educational publications and by a group of authors who are determined with mathematical and statistical data analysis methods [2; 6]. Considerable attention was paid to regression analysis by the authors in [6], because «Regression is one of the oldest areas of data mining that this paper covers, but it is still one of the most active fields of research, given its important and wide uses across fields of science. In this section, we will cover some of these recent developments».

Regression analysis is used when the relationship between variables can be expressed quantitatively in the form of some combination of these variables. The resulting combination is used to predict the value that the target (dependent) variable can take, which is calculated on a given set of input (independent) variable values. In the simple case, standard statistical methods such as linear regression are used for this, but most real models do not fit within its scope. For example, sales volumes or stock prices are difficult to predict because they may depend on a complex interrelationship of variables. That is why we suggest using the correlation analysis method as one of the classic data analysis methods. The study of statistical relationship is considered a very complex and time-consuming process, in which it is necessary to analyze multidimensional data tables. Therefore, as a rule, not a statistical, but a correlational relationship between X and Y features is studied [4].

It will be appropriate to give one more (another) definition of Data Mining.

Data Mining – it is a process with the goal of discovering new meaningful correlations, patterns, and trends as a result of sifting through large volumes of stored data using pattern recognition techniques plus the application of statistical and mathematical methods (Gartner Group definition).

The main tasks of correlation analysis are:

- study of the strength of connection (influence) between two or more features of the object under investigation;
- establishment of factors that have the most significant influence on the resulting feature;
- detection of unknown cause-and-effect relationships between object features.

Let's give a classic example. The company has been investing in advertising for many years. There is a task: to analyze whether and how investment in advertising affects the profit of the enterprise? It is also possible to analyze the "power of influence" of investing in equipment modernization, for example, on the profit of the enterprise, etc. Making the right investment decision depends on this.

It is known that the sample data for correlation studying between X and Y features usually have the form of their values pairs: $(x_1; y_1)$, $(x_2; y_2)$, …, $(x_n; y_n)$, x_i – values of the variable X, y_i – values of the variable Y, n – number of value pairs, $i = \overline{1,n}$. If their number is large enough, then for the convenience of calculations the data are grouped and a statistical series is constructed. This statistical series contains the values of X, the corresponding average values of Y, and frequencies n. That is, the functional dependence between the values of X and the average values of Y is correlation one: $Y = f(X)$.

It is possible to display the data graphically: plot points with coordinates $(x_i; y_j)$ ($i = \overline{1,n}$, $j = \overline{1,m}$) on the plane [1].

We will get a plane as a result, which is divided into rectangles and each of them can have a set of points. This graphical representation of the sample data is called a correlation field, which may look like on *Fig. 2,* for example:

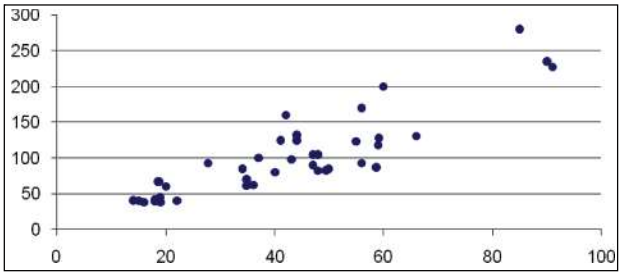

Fig. 2. *Example of correlation field*

To determine the closeness (or strength) of the connection (influence) between X and Y, the correlation coefficient is calculated. The Pearson correlation coefficient is used in the case when there is a linear relationship between X and Y and the sample data are distributed according to the normal law. The Pearson correlation coefficient is also called the parametric correlation coefficient. Is calculated according to the formula (1) (for example, a slightly simplified formula for ungrouped data):

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}. \quad (1)$$

Pearson's correlation coefficient takes values on the interval [-1;1]. In order to avoid cumbersome calculations, we can calculate the correlation coefficient value using a special function in the spreadsheet editor or perform a corresponding correlation analysis using one of the software tools that has this capability, for example, in the SPSS software package [5].

The calculated value of the correlation coefficient is analyzed according to the well-known Chaddock scale (*Table 1*) and the corresponding conclusions are drawn.

Table 1

The Chaddock scale

The connection closeness values (r)	The connection strength characteristics
0,1-0,3	weak
0,3-0,5	moderate
0,5-0,7	noticeable
0,7-0,9	high
0,9-0,99	very high

If $r > 0$, then the relationship is called positive, i.e. as the X values increase, the Y values also increase. If $r < 0$, then the relationship is negative, i.e. as X values increase, Y values decrease.

Let's note that the Pearson correlation coefficient shows the linear relationship strength. If there is a strong nonlinear relationship between X and Y, the Pearson correlation coefficient can be zero ($r = 0$).

Since the relationship strength between X and Y is estimated based on sample data, it is necessary to check its statistical significance [6], that is, to assess the possibility of spreading the obtained results to the entire general population.

As an example of the application of the proposed method, let`s consider an elementary problem. Determine the strength of the connection (influence) between height X (cm) and body weight Y (kg) of students at the age of 20 according to experimental measurement data:

X_i = (157; 158; 160; 165; 167; 162; 171; 174; 168; 176; 170; 180) and

Y_i = (56; 55; 57; 57; 58; 60; 63; 65; 67; 72; 79; 80).

To estimate the strength of the relationship between X and Y, let`s calculate the correlation coefficient using the formula (1). To find the required values, let`s fill in the additional table (*Table 2*) and make the appropriate calculations using the spreadsheet editor:

Table 2

Additional calculations

X_i	Y_i	$X_i \cdot Y_i$	X_i^2	Y_i^2
157	56	8792	24649	3136
158	55	8690	24964	3025
160	57	9120	25600	3249
165	57	9405	27225	3249
167	58	9686	27889	3364
162	60	9720	26244	3600
171	63	10773	29241	3969
174	65	11310	30276	4225
168	67	11256	28224	4489
176	72	12672	30976	5184
170	79	13430	28900	6241
180	80	14400	32400	6400
∑	∑	∑	∑	∑
2008	769	129254	336588	50131

After substituting the found values into the formula, we get the value of the correlation coefficient $r = 0{,}816$. According to the value of the correlation coefficient, we can conclude that there is a strong relationship between the height and mass of students at the age of 20 according to the measurement data.

We obtained this conclusion based on the sample data of the dimension $n = 12$. Since the strength of the connection between X and Y is estimated on the basis of sample data, it is necessary to check its statistical significance [6], that is, to assess the possibility of spreading the obtained results to any general population.

After applying the appropriate mathematical and statistical algorithm, we can get the conclusion that the correlation coefficient can be considered significant at the chosen levels ($\alpha_1 = 0{,}05$; $\alpha_2 = 0{,}01$; $\alpha_3 = 0{,}001$).

Let's consider another simple problem. Let`s analyze the dependence (influence) of the master's students age on the study results. We will not change the number of data in the experimental sample and leave it for convenience $n = 12$.

Let's use the calculation additional table created in the spreadsheet editor (*Table 2*) and substitute the values for

X_i = (22; 22; 30; 27; 23; 23; 35; 22; 22; 30; 27; 22) – age and Y_i = (4,9; 5; 3,2; 4; 5; 4,5; 3,1; 4,3; 4,5; 3,8; 4,3; 3) – average score of success into it.

We will get the correlation coefficient value $r = -0{,}64$ in this case. By known properties of the correlation coefficient $0{,}5 < |r| \le 0{,}7$ the relationship is average, that is, we will not conclude that age has an effect on the results of education. After applying the appropriate mathematical and statistical algorithm, we get the conclusion that the correlation coefficient can be considered significant at the at the chosen levels ($\alpha_1 = 0{,}05$; $\alpha_2 = 0{,}01$; $\alpha_3 = 0{,}001$). That is, the possibility of spreading the results to any general population is also available

We demonstrated the application of correlation analysis on the simple problems example, calculating the Pearson coefficient and taking into account its properties for the linear data dependence.

The Pearson correlation coefficient shows the strength of the linear relationship. If there is a strong nonlinear relationship between X and Y, the Pearson correlation coefficient may be zero.

To evaluate the strength of the relationship between X and Y in the case where there is a non-linear relationship between X and Y or the sample data are not distributed according to the normal law, other studies are performed. The Spearman correlation coefficient should be used in this case [1]. And in the case when the researched object or phenomenon is characterized by more than two features $X_1, X_2, \ldots, X_k$ it is necessary to study multiple dependencies. To assess the strength of the connection between a certain feature X_i and all other features, a multiple correlation coefficient is used [1].

III. Conclusions

Therefore, the proposed correlation analysis method in Data Mining is effective, especially given that for its implementation we have a sufficient arsenal of software today that replace cumbersome calculations. The use of mathematical and statistical methods in Data Mining is justified due to the fact that a sample can be generated for research from a large data warehouse and the results of the analysis can be transferred to the entire general population. It is also possible to group data for research.

The results of data analysis are used to create analytical models that can help the manager(s) of a company (enterprise, institution) in making a decision.

Data processing and analysis professionals present results to managers and users through data visualization, using various publication methods.

References:

1. Vasylenko O.A., Sencha I.A. Mathematical and statistical methods of analysis in applied research: teaching manual. Odesa, 2011. 166 p.
2. Nesvit M.I., Nesvit K.V. Mathematical methods of intelligent data analysis in real time. *Materials of the 71st scientific and methodological conference of the Kharkiv National University of Construction and Architecture "Trends of the development of higher technical education in Ukraine: European choice"*, April 12-13, 2016, P. 9-12.
3. Chernyak O.I., Zakharchenko P.V. Intelligent data analysis: Textbook. Kyiv, 2014. 599 p.
4. Tatiana PYLYPIUK. Classical technologies of intelligent data analysis. *Scientific works of Kamianets-Podilskyi Ivan Ohiienko National University:* collection based on the results of the reported scientific conference of teachers, doctoral students and postgraduates. [Electronic resource]. Kamianets-Podilskyi: Kamianets-Podilskyi Ivan Ohiienko National University, 2023. Issue 22. P. 676-677.
5. Pylypiuk T.M. Computer statistical packages. Laboratory practice. Kamianets-Podilskyi, 2021. 232 p.
6. Pinto da Costa, J.; Cabral, M. Statistical Methods with Applications in Data Mining: A Review of the Most RecentWorks. *Mathematics 2022*, 10, 993. DOI: https://doi.org /10.3390/math10060993

Тетяна ПИЛИПЮК, Віктор ЩИРБА

Кам'янець-Подільський національний університет імені Івана Огієнка

МЕТОДИ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ

Анотація: дослідження присвячене методам інтелектуального аналізу даних. Проведено порівняння класичних і математично-статистичних методів аналізу даних. Запропоновано та аргументовано описано один із варіантів методу кореляційного аналізу для інтелектуального аналізу даних.

Питання застосування різних методів для інтелектуального аналізу даних є актуальним. Класично в інтелектуальному аналізі даних пропонуються наступні методи виявлення й аналізу знань: класифікація; регресія; прогнозування часових послідовностей (рядів); кластеризація; об'єднання.

В якості математично-статистичних методів аналізу в прикладних дослідженнях більшість авторів пропонують такі методи як: статистична перевірка гіпотез, побудова та дослідження регресійних моделей. Оскільки більшість реальних моделей не піддаються аналізу за допомогою класичних методів, включаючи регресійний аналіз, автори пропонують використовувати метод кореляційного аналізу.

Ключові слова: інтелектуальний аналіз даних, методи аналізу, перевірка статистичних гіпотез, регресійні моделі, кореляційний аналіз.

Отримано: 10.09.2023

УДК 37.01(477)"202" DOI: 10.32626/2307-4507.2023-29.10-14

Roksolyana SHVAY

Pomorska Szkoła Wyższa w Starogardzie Gdańskim

e-mail: Roksolyanash@yahoo.com; ORCID: 0000-0003-3859-5196

WYBRANE PROBLEMY WSPÓŁCZESNEJ EDUKACJI

Adnotacja. W artykule poddano analizie problemy współczesnej edukacji. System edukacji ciągle potrzebuje zmian ze względu na współczesne technologie informatyczne, mające wpływ na życie, uczenie się i sposób komunikowania się, myślenie. Rozwój technologii, wysokie tempo życia powodują zmiany w neuronalnej budowie mózgu, zmienia się aktywność mózgu na poziomie biochemicznym. Młodzi ludzie nie są zdolni do głębszej refleksji, nie potrafią wyciągać wniosków, interpretować informacji, są mniej kreatywni, mniej empatyczne, tolerancyjne, całkowicie obojętni na to, co ich nie dotyczy osobiście, są kłopoty z wyrażamiem swoich uczuć, rozumieniem cudzego punktu widzenia i utrzymywaniem prawidłowych relacji społecznych. Strategię pracy mózgu zmienia zjawisko multitasking (wielozadaniowość) jak wykonywanie różnych chynności jednocześnie. To zjawisko prowadzi do gorszych wyników w nauczaniu, wzrostu poziomu lęku, zmniejszania satysfakcji z życia. Zachodzące zmiany w architekturze mózgu wymagają zaistnienia nowych koncepcji nauczania. W procesie dydaktycznym wdrażają się technologie mobilne, personalne środowiska dydaktyczne, używa się zasobu otwartego oraz otwartych platform dydaktycznych. Poddano analizie kluczowe kompetencji nowoczesnej epoki cyfrowej, wady i zalety e-learningu. Tworzenie nowoczesnego modelu kształcenia osobowości innowacyjnej jest odpowiedzią na wyzwania epoki cyfrowej.

Słowa kluczowe: edukacja, mózg, metoda, formy, koncepcji, technologie informatyczne, e-learning.

Technologie XXI wieku szybko się rozwijają oraz społeczeństwo cyfrowe ewoluuje bardzo szybko. System edukacji ciągle potrzebuje zmian – tak w całości, jak też w jego poszczególnych częściach. Jest to system otwarty, który z jednej strony podlega oddziaływaniom środowiska, z drugiej zaś – stymuluje zmiany. System edukacji nie tylko dostosowuje się do potrzeb społeczeństwa, ale może również tworzyć przyszłość. Tylko jedno jest prawie bez zmian – edukacja zawsze opiera się na interakcji „uczeń-nauczyciel", w której ważna jest komunikacja ustna, wspomagana różnymi środkami uczenia się. W procesie uczenia się uczymy się nie tylko myśleć i zdobywać wiedzę, ale także zdobywać doświadczenie, umiejetności życia w społeczeństwie.

Szybki rozwój nauki stale pogłębia dysproporcję między wzrostem wiedzy naukowej a możliwością przyswajania dużej ilości informacji uczniem. Narasta dysproporcja między poziomem rozwoju mózgu ucznia, poziomem jego gotowości umysłowej do percepowania nowych informacji a osiąganym poziomem wiedzy naukowej. Stale zwiększający się rozłam między rozbujałą wiedzą naukową i mozliwością przyswojenia dużego zasobu wiadomości określa zagadnienia szkolnej wiedzy uczniów. Programy nauczania przeobciążone są wiado-

