UDC 330.46:004.89
DOI: 10.20535/SRIT.2308-8893.2022.4.06

DATA MINING TOOLS FOR COMPLEX SOCIO-ECONOMIC PROCESSES AND SYSTEMS

T.V. OBELETS

Abstract. The paper considers discovering new and potentially useful information from large amounts of data that actualizes the role of developing data mining tools for complex socio-economic processes and systems based on the principles of the digital economy and their processing using network applications. The stages of data mining for complex socio-economic processes and systems were outlined. The algorithm of data mining was considered. It is determined that the previously used stages of data mining, which were limited to the model-building process, can be extended through the use of more powerful computer technology and the emergence of free access to large amounts of multidimensional data. The available stages of data mining for complex socio-economic processes and systems include the processes of facilitating data preparation, evaluation, and visualization of models, as well as in-depth learning. The data mining tools for complex socio-economic processes and systems in the context of technological progress and following the big data paradigm were identified. The data processing cycle has been investigated; this process consists of a series of steps starting with the input of raw data and ending with the output of useful information. The knowledge obtained at the data processing stage is the basis for creating models of complex socio-economic processes and systems. Two types of models (descriptive and predictive) that could be created in the data mining process were outlined. Algorithms for estimating and analyzing data for modeling complex socio-economic processes and systems in accordance with the pre-set task were determined. The efficiency of introducing neural networks and deep learning methods used in data mining was analyzed. It was determined that they would allow effective analysis and use of the existing large data sets for operational human resources management and strategic planning of complex socio-economic processes and systems.

Keywords: data mining, complex socio-economic systems, predictive modeling, neural networks, deep learning.

INTRODUCTION

The continuous scaling of data on the Internet is changing the way we interact in economic and social systems. Many users search, publish, and create new data daily, leaving a digital footprint that can help describe their behavior, decisions, and intentions. This highlights the role of developing data mining tools for complex socio-economic processes and systems based on the principles of the digital economy and their processing using network applications.

Analysis of recent research and publications. The most significant results in statistical analysis and applications for data mining were achieved in the works of R. Nisbet, H. Miner, O. Maimon, L. Rokach. Such scientists as H. Jiawei, M. Kamber, P. Jiang, H. Choi, H. Varian, and others devoted their research to the creation of concepts and development of data mining techniques. H. Xiong, H. Pandei, A. Kryzhevsky, I. Sutskever and Jeffrey E. Hinton explored machine-learning capabilities with deep convolution neural networks. Successful results in

the field of artificial intelligence with deep learning were obtained in the works of J. Lekun, Y. Bengio, G. Hinton, J. Schmidhuber, and Ukrainian authors M. Lavrenyuk, N. Kusul, O. Novikov. The analysis of complex socio-economic systems was carried out by foreign authors in their works P. dos Santos, N. Wiener, J. Stefanovsky, and the issues of forecasting socio-economic processes were in the interests of Ukrainian scientists G. Prisenko and E. Ravikovych and others. At the same time, the progress of computing power and the availability of large amounts of multidimensional data make it necessary to develop data mining tools for complex socio-economic processes and systems.

The purpose of the article is to study the process of identifying new and potentially useful information from large amounts of data, outlining the stages of data mining for complex socio-economic processes and systems, and identifying appropriate tools in the context of the progress of computing power and the emergence of a large number of multi-dimensional data in the free-of-use.

Presentation of the main material of the study. As data mining has evolved as a professional activity, it is necessary to distinguish it from previous statistical modeling activities and broader knowledge discovery activities. Data mining is defined as the use of machine learning algorithms to find weak patterns of relationship between data elements in large and disordered datasets, which can lead to actions to increase benefits in one form or another (diagnostics, profit, prediction, management, etc.) [1].

Data mining is also called knowledge discovery in databases (Knowledge Discovery in Databases — KDD), i.e. the process of discovering new and potentially useful information from large amounts of data. The definition of data mining was initially limited to the modeling process, but over time the data analysis tools have included processes to facilitate data preparation, as well as evaluation and visualization of models [2] (Fig. 1).

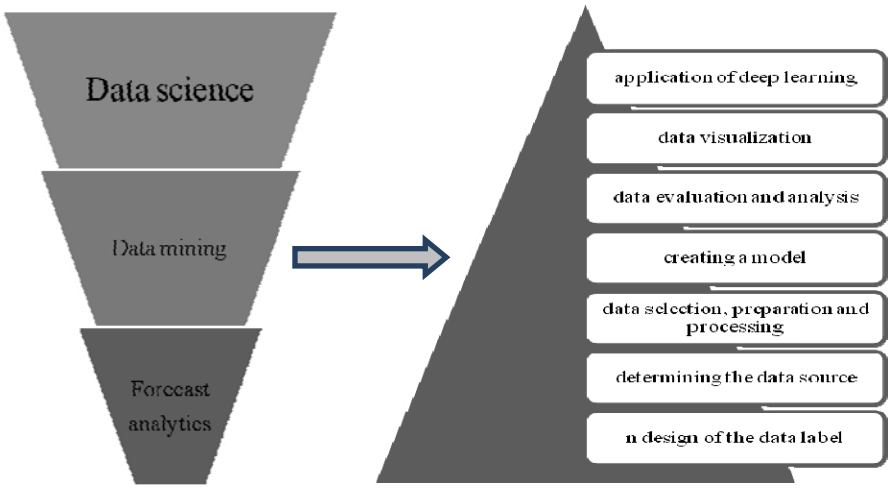

Fig. 1. Algorithm of data analytics and tools of data mining

The process of identifying knowledge in databases combines the mathematics used to identify patterns in the data with the whole process of data selection and the use of models to apply to other datasets and use the information for predetermined purposes. This process combines the development of business systems, statistical methods, and digital technologies to identify the structure of socio-economic processes and systems (relationships, patterns, associations, and basic functions), not just their statistical parameters (averages, weights, etc.) [1].

The data mining algorithm begins with the definition of objectives in the data matrix design process and ends with the introduction of the identified knowledge. At the stage of designing the data matrix, preparatory work is carried out, goals are defined and strategic ideas are formed, for the achievement of which the process of knowledge discovery in databases begins. Understanding the strategic goal, a clear understanding of the end-user of the data, and understanding the environment in which the data will be disseminated, is a prerequisites for an adequate process of datamining.

Data mining uses techniques from different fields of knowledge, such as statistics, machine learning, pattern recognition, database and storage systems, information search, visualization, algorithms, high-performance computing, and many application areas. Statistics examines the collection, analysis, interpretation or explanation, and presentation of data. Statistical models are widely used to model data and data classes. Multivariate graphical methods are used to research. analyze databases and present the results of data analysis [3].

Identification of data sources and searching for documents or information in documents is an important step in data mining. Documents can be text-based or multimedia can be on paper in archives and can be available electronically on the Internet. The main source of information for data mining for complex socio-economic processes and systems today is the Internet. Thus, Google Trends (GT) is an online tool that reports on the volume of search queries for a particular keyword or text. The use of GT data for the current forecast of social and economic variables was introduced in 2009 [4]. Social networking sites and blogs are specifically designed to encourage users to express their feelings and opinions, which can potentially be used to predict social variables. Websites and programs (transactional platforms, opinion platforms, and dissemination of information) created on the Internet by enterprises, public organizations, charitable foundations, or multinational corporations inform about their products, services, organizational structure, and intentions. In addition to providing information, websites are used for transactions, e-commerce, and online services. According to the big data paradigm, this wide variety of sources requires specific tools for processing them.

The selection, preparation and processing of data based on which the intellectual analysis will be carried out is a stage of creating opportunities. The data to be used to identify knowledge must meet the following requirements: the available data must be, firstly, reliable, secondly, up-to-date, thirdly, sufficient to present the information as fully as possible, and fourthly, optimally necessary so as not to overload research database systems and integrate all selected data into one set. The minimum set of available data, if necessary, can be extended with additional necessary data to identify nuances that will be taken into account when creating a model. Sometimes the presence of such minor accents can be fundamental to the success of the knowledge discovery process in databases. A large number of nuances provides more opportunities to create a multidimensional model that will allow the most complete consideration of the studied phenomena and perform intellectual analysis. However, storing, organizing, and managing large and complex databases requires large resources, which are planned in advance and often limited.

Research on database systems and data stores focuses on creating, maintaining, and using databases for organizations and end-users. Database systems are known for their high scalability in processing very large, relatively structured datasets. Choosing the optimal dataset should balance the requirements

of sufficiency and necessity, so this stage of data mining creates the foundation for opportunities. In addition, the choice of data should be guided by their validity and relevance.

Preparing data for further processing increases the validity of the data set. Preparation includes sorting and filtering data that will eventually be used as input data. This involves cleaning up the data by removing missing values and informational noise. Noise removal increases the chances of performing data mining most efficiently.

Removing objects with noise is an important goal of data cleansing, as noise interferes with most types of data analysis. Most existing data cleaning methods focus on noise removal, which is the product of low-level data errors that are the result of an imperfect data collection process, but irrelevant data objects or of little relevance, can also significantly impede data analysis. Thus, if the goal is to improve data analysis as much as possible, these objects should also be considered noise, at least concerning to the main analysis. Therefore, there is a need for data cleansing techniques that eliminate both types of noise. Because datasets can contain a lot of noise, these methods should also be able to discard potentially much of the data [5].

Data processing is the process of converting raw data into useful information through electronic data processing, machining, or automated means. Data processing can take time depending on the complexity of the data and the amount of input data. The preparation step described above helps to make this process faster. Data processing is usually performed step by step: raw data is collected, filtered, sorted, analyzed, stored, and then provided in an accessible format, such as graphs, charts, and documents (Fig. 2).

The data processing cycle consists of a series of steps in which raw data (input data) enters the process to obtain useful information (output). Each step is performed in a certain order, but the whole process can be repeated cyclically. The output of the first data processing cycle can be stored and presented as input for the next cycle. If the data obtained in the processing process is not used as input data for the next processing cycle, this complete process cannot be considered a cycle and will remain a one-time activity for data processing and information obtaining.

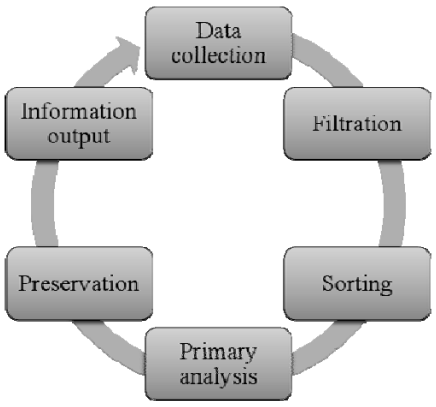

Fig. 2. Data processing cycle

Information is processed and analyzed databases. The information obtained at this stage can be useful and become the basis for the formation of knowledge. This is how the data processing phase discovers knowledge in databases. In the next stages, knowledge of social processes is identified, economic phenomena should be formed and presented in such a way as to create a model of complex socio-economic processes and systems. At the stage of data discovery and recognition of useful information and further knowledge of databases, mathematical, statistical methods, methods of artificial intelligence and machine learning are used.

The evolution of the Internet and social media has led to a huge explosion in the volume and complexity of data, so-called big data. Thus, data mining has also

gone beyond traditional data modeling**,** such as regression and statistical models. Information theory offers tools to make formal conclusions about complex models of economic and social interaction. There are two main theoretical concepts of information that can help guide the observation of the relationship between the economic characteristics of a large number of people: entropy and mutual information. The concepts of entropy and mutual information make it possible to develop non-parametric characteristics of information associations present in the observed data generated by economic and broader social interactions [6].

Creating a model is a stage of data mining that requires special responsibility and diligence. Many algorithms can be used to model complex socio-economic processes and systems in accordance with a predetermined task. Different socio-economic processes and systems have different and complex causal relationships, so it is very important to determine the tactics of finding such connections and choose the optimal algorithm. This step involves choosing a specific method that will be used to find templates and data that most accurately describe the process or system under study. Currently, many algorithms are known to solve data mining problems: the method of reference vectors, the method of k-nearest neighbors, neural networks and decision trees.

When choosing a specific method, it is necessary to take into account the strategic goal of data mining and, accordingly, to determine the priority characteristics of the model that will be created at this stage. On the one hand, we can consider such a characteristic as the accuracy of the model, and on the other – clarity and simplicity of perception. To create a simpler model that should be intuitive, the best choice may be to use the decision tree method, which is one of the most popular methods of solving classification and forecasting problems This is a way of demonstrating rules in a hierarchical, sequential structure, where each object corresponds to a single node that provides the solution. The decision tree method should be used in cases where symbolic representation and good classification are required; the problem does not depend on many attributes; a modest subset of attributes contains relevant information; linear combinations of features are not critical; important learning speed [7].

To create an accurate model, it is appropriate to use neural networks - an extremely powerful method of modeling, which allows you to reproduce extremely complex relationships. For many years, linear modeling has been the main method of modeling in most areas, as it has well-developed optimization procedures. In problems where the linear approximation is unsatisfactory (and there are many of them), linear models work poorly. In addition, neural networks cope with the "curse of dimensionality", which does not allow you to simulate linear relationships in the case of a large number of variables. The advantages of the neural network method are the following [8]:

– nonlinearity, neural networks are nonlinear;

– through controlled learning the network learns according to the examples: after receiving the primary information from the operator, the learning algorithm is started, which automatically perceives the data structure

– adaptability, i.e. the network can adapt its synaptic scales even in real time,

– response capability – in the context of template classification, the network not only provides template selection but also reliability of decision-making,

– fault tolerance due to massive interconnections,

– integrated large scale, i.e. its parallelism makes it potentially faster for certain tasks and thus captures complex patterns of behavior;

– homogeneity in analysis and design, i.e. the same notation is used in all in all areas related to neural networks,

– the analogy of neurobiology [9], in general, neural networks are self- adaptive and nonlinear methods that collect data and do not require specific assumptions about the basic model.

For each strategy of data mining and modeling in this process, there are several possible methods by which you can achieve your goals. The choice of a particular method is explained by the efficiency of the algorithm in a particular problem. Thus, at this stage of data mining, the most acceptable method of modeling is selected in accordance with the conditions. All these algorithms study the data and create models that are closest to the characteristics of the studied data of complex socio-economic systems and processes. Models created during data mining can be of two types: predictive or descriptive (Fig. 3).

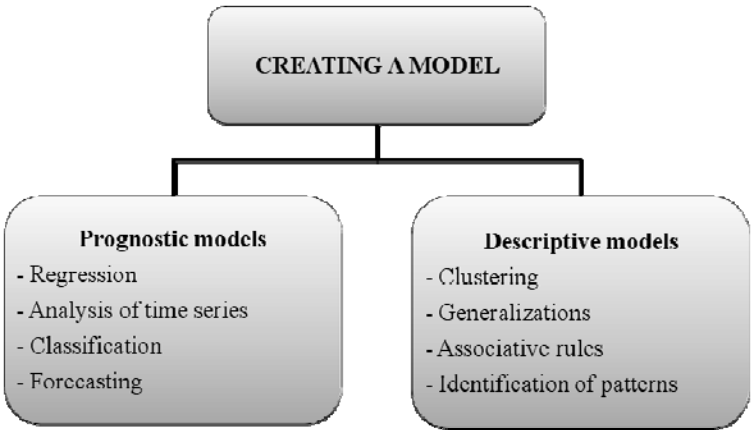

Fig. 3. Types of models, created in the process of data mining

The predictive model is a projection based on the data and information obtained in the earlier stages of data mining. As a rule, the forecast model is created on the basis of the directed analysis of data; that is, a top-down approach, where mappings from a vector input to a scalar output are obtained by applying a specific one. For example, predictive modeling can be performed using a variety of historical and statistical data. When creating a prognostic model in the process of data mining, the following tasks are performed: regression, time series analysis, classification, and forecasting [10].

The prognostic model is also known as a statistical regression. It is a monitoring method that involves explaining the relationship of several attribute values among themselves in similar elements and predicting the development of the model, that is, directional modeling based on these observations. As noted earlier, the two common methods of predictive modeling available in many data mining tools are neural networks and decision trees.

The descriptive model presents in a concise form the main characteristics of the data set. In essence, it is a collection of data points that allows you to study important aspects of a data set. As a rule, the descriptive model is created by indirect data analysis; that is, a bottom-up approach where the data "speaks for itself". Undirected data analysis finds patterns in the data set, but the patterns are inter-

preted by analysts. Data mining specialists determine the usability of the found templates. The most characteristic tasks of descriptive modeling are the following:

- clustering, i.e. decomposing or splitting a data set into groups;
- generalization as the process of providing summary information from data in an easier to understand form;
- association of rules – identification of causal relationships between different features in large data sets;
- sequence detection, which involves the identification of patterns of interest to researchers in the data.

Descriptive models and predictive models can (and often should) be used together in data mining. For example, it seems logical and appropriate to first look for patterns in the data using non-directional methods. These descriptive models can offer segments of data sets and ideas that improve the results of directional modeling when creating predictive models.

The modular design of neural network architecture facilitates the creation of models that simultaneously process data presented in different formats, such as creating text annotations from images, synthesizing language from the text, or through translation. This allows you to solve problems that go beyond traditional classification and regression, and is especially convenient when the data comes from different sources, which is often the case when working with big data. In addition, data obtained from different repositories or databases, presented in the form of object maps, can be reused in other contexts and, if necessary, further configured/ taught.

Evaluation and analysis of data. The purpose of any predictive modeling is to apply the model to new data. Forecasting models are useful only insofar as the quality of their prediction is adequate, therefore, the principle is not the process of creating a model as such, but the creation of a high quality model. Both predictive models and descriptive models have their evaluation criteria. For forecast models, the evaluation criterion is the accuracy of the forecast, measured by the size of the forecast error, i.e. the difference between the forecast and the actual value of the studied indicator. For descriptive models, it is more difficult to define obvious evaluation criteria, but they usually capture a discrepancy between the observed data and the proposed model. Thus, at this stage of data mining, different strategies for assessing the quality of models can be used.

Parametric methods for analyzing the accuracy of forecasts. According to the results of the ex-post-forecast, such indicators of forecast accuracy for m steps as the root mean square error are calculated, the root of standard error, mean absolute error, root of root mean square error in percentage, mean absolute relative error in percentage (MARE). The smaller the value of these values, the higher the quality of the forecast. In practice, these characteristics are used quite often. This approach gives good results, if in the period of the retro forecast there are no fundamentally new patterns. To create a prognostic model of complex socio-economic systems and processes, each time the forecast is built in a new situation, therefore, the comparison of the numerical accuracy of forecasts made at different points in time is not entirely correct. These considerations led to the use of non-parametric methods of analysis of the accuracy of forecasts [11].

Non-parametric methods of forecast accuracy analysis have two types of non-parametric criteria: label criterion and rank criterion. The criterion of labels for comparing the accuracy of two sequences of predictions is based on the percentage of cases when the method of determining the prediction A is better than

the method B. Such a comparison is made for individual predictions of the same events (variables). If the ranks of their criteria are applied, the numerical characteristic of accuracy (absolute error when estimating one forecast, or root mean square error when considering a sequence of predictions) is replaced by ranks, which are then checked for significance. For example, if the sequences of predictions of indicators A and B are obtained using k methods, then first calculate the root mean square error, then the values are ranked from smallest to largest. Although non-parametric methods have their advantages, it is important to realize that they ignore some of the available information. Thus, the criteria of labels and ranks do not take into account the numerical values of errors [11].

Data visualization. Created models used in the process of data mining of complex socio-economic systems and processes including large and complex parameters. To solve the problem of size and complexity, the best methods are used to represent complex systems and data visualization (for example, advanced user interfaces). These technologies increase the level of abstraction, which helps users focus on the most important components and properties of complex models.

In the world of big data, data visualization tools and technologies are needed to analyze large and complex amounts of information and make decisions based on the intellectual analysis of this data. Data visualization is a graphical representation of information and data. Using visual elements such as charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, deviations, and patterns in data. Effective data visualization is a delicate balance between form and function. On the one hand, the simplest schedule can be both a very primitive transfer of information and a vision of the main core of information that is analyzed and must be presented and understood. On the other hand, the most complex visualization can overload information, and say about many details, but not convey the main essence of the message. Data and visual elements must work together to create a better understanding and awareness of information.

There is a choice of visualization methods for efficient and interesting data presentation. Common types of data visualization: charts, tables, graphs, maps, infographics, dashboards, etc.

Construction and illustration of relationships between different objects of the created models of complex socio-economic systems and processes can be done with the help of modern tools. Therefore, Draw.io is a free, intuitive browser-based flowchart builder where users can drag object shapes (including ellipses and parallelograms common to data models) onto a canvas, and then combine them into by means of through connecting lines. Lucidchart Chart Designer is similar to Draw.io, but it reproduces streams that are more complex and has more reliable data protection. SQuirreL is a free and open-source graphical tool supported by most major relational databases.

The most important trend in the field of data in recent years is the proliferation of data catalogs, largely due to privacy rules such as the GDPR and the CCPA (General Data Protection Regulation of 2016; California Consumer Privacy Act of 2018) [12]. This trend has not escaped the field of data mining and modeling. The line between data discovery tools and applications and data modeling tools is increasingly blurred, as exemplified by Amundsen, a metadata-based data discovery platform developed by Lyft [13].

Open source Metabase is a GUI tool with some useful analytics visualizations but does not support modeling tools [14]. Other notable data visualization tools include erwin, ER / Studio, SAP PowerDesigner, IBM InfoSphere Data Architect, and Microsoft SQL Server Management Studio, etc.

Application of deep learning. The knowledge and models of processes created in the socio-economic systems created in the process of data mining will be popular only if they can be included in other complex systems in order to predict their development. Forecasting analytics is interesting and useful in the context of the possibility of making changes to the simulated system and presenting the long-term consequences of these changes. The real structure of complex socio-economic systems is dynamic, data characteristics may change over time, new parameters may appear that were not foreseen in the model, and others may disappear. Therefore, this stage of application of deep learning determines the success and effectiveness of the whole process of data mining.

Progress in the field of deep learning has made it possible to use the powerful capabilities of artificial neural networks in this process. They are a universal tool for data mining and effective for learning based on data presented in various formats. For example, neural networks have demonstrated their effectiveness in performing certain tasks of object or image recognition [15–16]. Recent approaches have shown that deep learning can effectively learn based on data representations in variable length sequences (time series, sound, language, and text), graphs and networks, including social networks, natural language and even source code in computer programs [17–18].

Another aspect of deep neural networks that is closely related to big data is their ability to perform complex functional design. The problem of big data is often related to the difficulty of making reliable predictions when training data needs to be represented, for example, to identify successfully relevant classes of solutions. An important requirement for the use of deep learning methods is the availability of large samples of training, as insufficient training data causes the problem of "overfitting" when the model does not summarize the information obtained during training, but simply remembers it. In this case, the model shows good results on educational data but does not show such accuracy on unfamiliar data [19].

Previously, the search for useful representations had to be conducted by experts using manual design of characteristics or explicit methods of selection and construction of functions [20]. At the present stage of technology development, the most suitable architecture for data processing, which characterizes complex socio-economic processes and systems, and as a consequence to solve the problem of data mining are convolutional neural networks, because they are designed to process data in the form of multidimensional arrays [21].

Some neural models on the ethane of deep learning allow you to synthesize features in the form of hidden variables with certain desired properties. Neural networks help automate the tasks set at the beginning of the data mining process complex socio-economic systems: the construction of the characteristics of these systems becomes an integral part of the process of deep learning, closely related to the search in space for new hypotheses for the development of socio-economic processes.

CONCLUSIONS

Thus, the study of the data mining process showed that the expansion of the data analysis tools in connection with the powerful development of technologies, the formation of big data sets creates the ability to track, evaluate, simulate, and ultimately include key economic and social changes and trends in complex processes and systems. An important step that has increased the efficiency of data mining has been the inclusion of steps to facilitate data production as well as model evaluation and visualization.

The descriptive and predictive models generated by the mining process can and should be used together. The logical sequence of the model application, which will improve the results of the directional modeling, is seen primarily in the search for patterns in the data with the help of descriptive models, and already based on the obtained ideas of directional modeling when creating predictive models of complex socio-economic processes and systems.

At the present stage of technology development, machine learning is widely used in data mining to invent complex models and algorithms that serve to create descriptive and predictive models of complex socio-economic systems and processes. Machine learning gives computers the ability to "learn", recognize complex patterns and make intelligent decisions without explicit programming based on large data samples. These opportunities are the basic application of deep learning methods, designed to process data presented in the form of multidimensional arrays, and allow you to create models of complex socio-economic processes and take into account possible changes to design and manage the development of complex systems. That is, the use of the above tools allows you to perform successfully and efficiently the tasks of data mining of complex socio-economic processes and systems.

Therefore, digital tools are becoming relevant to maintain effective competitiveness, help model complex socio-economic processes and systems, effectively analyze and use existing large data sets for operational human resource management and strategic planning of complex socio-economic processes and systems.

REFERENCES

1. R.R. Nisbet, G. Miner, and K. Yale, "Chapter 2 – Theoretical Considerations for Data Mining," Editor(s): Robert Nisbet, Gary Miner, Ken Yale, *Handbook of Statistical Analysis and Data Mining Applications (Second Edition)*, Academic Press, 2018, pp. 21–37. Available: https://doi.org/10.1016/B978-0-12-416632-5.00002-5
2. O. Maimon and L. Rokach, *Data Mining and Knowledge Discovery Handbook*, 2nd ed. Springer, January 2010, 1285 p. Available: https://doi.org/10.1007/978-0-387-09823-4
3. H. Jiawei, M.Kamber, and J. Pei, *Data mining: concepts and techniques*, 3rd ed. Morgan Kaufmann Publishers, 2012, pp. 23–27.
4. H. Choi and H. Varian, *Predicting the Present with Google Trends*, 2009. Available: http://static.googleusercontent.com/external_content/untrusted_dlcp/www.google.com/en//googleblogs/pdfs/google_predicting_the_present.pdf
5. H. Xiong, G. Pandey, M. Steinbach, and V. Kumar, "Enhancing data analysis with noise removal," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, iss. 3, pp. 304–319, 2006. Available: http://datamining.rutgers.edu/publication/ tkdehcleaner.pdf
6. P.L. dos Santos, and N. Wiener, "Indices of Informational Association and Analysis of Complex Socio-Economic Systems," *Entropy*, 21(4), 367, 2019. Available: https://doi.org/10.3390/e21040367
7. J. Stefanowski, *Discovering Decision Trees*. Institute of Computing Science. Poznań University of Technology, 2010, 45 p. Available: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.176.1423&rep=rep1&type=pdf
8. E. Gómez-Ramos and F. Venegas-Martínez, "A Review of Artificial Neural Networks: How Well Do They Perform in Forecasting Time Series?" *Analítika, Revista de análisis estadístico*, vol. 6, no. 2, pp. 7–15, 2013.
9. R. Tadeusiewicz, *Neural networks: A comprehensive foundation: by Simon HAYKIN.* USA, New York: Macmillan College Publishing, 1995, 696 p.
10. U. Johansson, *Obtaining Accurate and Comprehensible Data Mining Models – An Evolutionary Approach.* Linköping, Sweden: Department of Computer and Information Science, Linköpings universitet, 2007, 272 p. Available: http://www.diva-portal.org/smash/get/diva2:23601/FULLTEXT01.pdf
11. G.V. Prisenko and E.I. Ravikovich, *Forecasting of socio-economic processes*: *Textbook*. K: KNEU, 2005, 378 p.

12. "Comparing privacy laws: GDPR v. CCPA," *Data Guidance and Future of Privacy Forum*, 42 p. Available: https://fpf.org/wp-content/uploads/2018/11/GDPR_CCPA_ Comparison-Guide.pdf
13. *Amundsen. Open source data discovery and metadata engine.* Available: https://www.amundsen.io/
14. *Metabase. Built for data*. Available: https://www.metabase.com/
15. A. Krizhevsky, I. Sutskever, and G.E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, 25, pp. 1097–1105, 2012.
16. Jiquan Ngiam et al., "Multimodal deep learning," *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011.
17. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, 521(7553), pp. 436–444, 2015.
18. J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, 61(C), pp. 85–117, 2015.
19. M. Lavreniuk, N. Kussul, and A. Novikov, "Deep Learning Crop Classification Approach Based on Coding Input Satellite Data Into the Unified Hyperspace," *IEEE 38th International Conference on Electronics and Nanotechnology*, pp. 239–244, 2018.
20. K. Krawiec, "Evolutionary feature selection and construction," in S. Claude and G. Webb (Eds.) *Encyclopedia of Machine Learning and Data Mining*. Boston, MA: Springer, 2016.
21. M. Reichstein et al., "Deep learning and process understanding for data-driven Earth system science," *Nature*, vol. 566, iss. 7743, pp. 195–204, 2019.

Received 16.05.2022

INFORMATIONON THE ARTICLE

Tetyana V. Obelets, ORCID: 0000-0002-1553-5150, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Ukraine, e-mail: obelectv@ukr.net

ІНСТРУМЕНТАРІЙ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ ДЛЯ СКЛАДНИХ СОЦІАЛЬНО-ЕКОНОМІЧНИХ ПРОЦЕСІВ ТА СИСТЕМ / Т.В. Обелець

Анотація. Розглянуто процес виявлення нової та потенційно корисної інформації з великих обсягів даних, що актуалізує роль розроблення інструментарію інтелектуального аналізу даних для складних соціально-економічних процесів та систем на основі принципів цифрової економіки та їх оброблення за допомогою мережевих застосунків. Окреслено етапи інтелектуального аналізу даних для складних соціально-економічних процесів та систем. Розглянуто алгоритм інтелектуального аналізу даних. Визначено, що використовувані раніше етапи інтелектуального аналізу даних, які обмежувалися лише процесом побудови моделі, можуть бути розширені завдяки використанню більш потужної обчислювальної техніки та появи у вільному доступі великої кількості багатовимірних даних. До наявних етапів інтелектуального аналізу даних для складних соціально-економічних процесів та систем включено процеси полегшення підготовки даних, оцінювання та візуалізацію моделей, а також глибинне навчання. Визначено інструментарій інтелектуального аналізу даних для складних соціально-економічних процесів та систем у контексті технологічного прогресу та відповідно до парадигми великих даних. Досліджено циклічність оброблення даних; цей процес складається із серії кроків, починаючи із входу необроблених даних, закінчуючи виведенням корисної інформації. Отримані на етапі оброблення даних знання закладаються в основу створення моделей складних соціально-економічних процесів та систем. Окреслено два типи моделей (описову та прогностичну), що можуть бути створені у процесі інтелектуального аналізу даних. Визначено алгоритми оцінювання та аналізу даних моделювання складних соціально-економічних процесів та систем відповідно до заздалегідь поставленого завдання. Проаналізовано ефективність запровадження нейронних мереж та методів глибинного навчання, що застосовуються у процесі інтелектуального аналізу даних. Визначено, що вони дозволять ефективно аналізувати та використовувати наявні великі масиви даних як для оперативного управління людськими ресурсами, так і стратегічного планування розвитку складних соціально-економічних процесів та систем.

Ключові слова: інтелектуальний аналіз даних, складні соціально-економічні системи, прогностичне моделювання, нейронні мережі, глибинне навчання.