









“Data mining as a cognitive tool: Capabilities and limits”

AUTHORS	Maxim Polyakov   Igor Khanin  Gennadiy Shevchenko   Vladimir Bilozubenko  
ARTICLE INFO	Maxim Polyakov, Igor Khanin, Gennadiy Shevchenko and Vladimir Bilozubenko (2021). Data mining as a cognitive tool: Capabilities and limits. <i>Knowledge and Performance Management</i> , 5(1), 1-13. doi: 10.21511/kpm.05(1).2021.01
DOI	http://dx.doi.org/10.21511/kpm.05(1).2021.01
RELEASED ON	Thursday, 08 July 2021
RECEIVED ON	Tuesday, 15 December 2020
ACCEPTED ON	Monday, 05 July 2021
LICENSE	 This work is licensed under a Creative Commons Attribution 4.0 International License
JOURNAL	"Knowledge and Performance Management"
ISSN PRINT	2543-5507
ISSN ONLINE	2616-3829
PUBLISHER	LLC “Consulting Publishing Company “Business Perspectives”
FOUNDER	Sp. z o.o. Kozmenko Science Publishing



NUMBER OF REFERENCES

32



NUMBER OF FIGURES

4



NUMBER OF TABLES

0

© The author(s) 2021. This publication is an open access article.



BUSINESS PERSPECTIVES



LLC "CPC "Business Perspectives"
Hryhorii Skovoroda lane, 10,
Sumy, 40022, Ukraine
www.businessperspectives.org

Received on: 15th of December, 2020

Accepted on: 5th of July, 2021

Published on: 8th of July, 2021

© Maxim Polyakov, Igor Khanin,
Gennadiy Shevchenko, Vladimir
Bilozubenko, 2021

Maxim Polyakov, Doctor of Economics,
Associate Professor, Managing Partner
Noosphere Ventures Inc., USA.

Igor Khanin, Doctor of Economics,
Professor, National University of Water
and Environmental Engineering,
Ukraine.

Gennadiy Shevchenko, Candidate of
Technical Sciences, Associate Professor,
Association Noosphere, Ukraine.

Vladimir Bilozubenko, Doctor of
Economics, Professor, Head of the
Department of International Economic
Relations, Regional Studies and
Tourism, University of Customs and
Finance, Ukraine. (Corresponding
author)



This is an Open Access article,
distributed under the terms of the
[Creative Commons Attribution 4.0
International license](https://creativecommons.org/licenses/by/4.0/), which permits
unrestricted re-use, distribution, and
reproduction in any medium, provided
the original work is properly cited.

Conflict of interest statement:

Author(s) reported no conflict of interest

Maxim Polyakov (USA), **Igor Khanin** (Ukraine), **Gennadiy Shevchenko** (Ukraine),
Vladimir Bilozubenko (Ukraine)

DATA MINING AS A COGNITIVE TOOL: CAPABILITIES AND LIMITS

Abstract

Due to the large volumes of empirical digitized data, a critical challenge is to identify their hidden and unobvious patterns, enabling to gain new knowledge. To make efficient use of data mining (DM) methods, it is required to know its capabilities and limits of application as a cognitive tool. The paper aims to specify the capabilities and limits of DM methods within the methodology of scientific cognition. This will enhance the efficiency of these DM methods for experts in this field as well as for professionals in other fields who analyze empirical data. It was proposed to supplement the existing classification of cognitive levels by the level of empirical regularity (ER) or provisional hypothesis. If ER is generated using DM software algorithm, it can be called the man-machine hypothesis. Thereby, the place of DM in the classification of the levels of empirical cognition was determined. The paper drawn up the scheme illustrating the relationship between the cognitive levels, which supplements the well-known schemes of their classification, demonstrates maximum capabilities of DM methods, and also shows the possibility of a transition from practice to the scientific method through the generation of ER, and further from ER to hypotheses, and from hypotheses to the scientific method. In terms of the methodology of scientific cognition, the most critical fact was established – the limitation of any DM methods is the level of ER. As a result of applying any software developed based on DM methods, the level of cognition achieved represents the ER level.

Keywords

data mining, data, scientific cognition, methodology,
empirical regularity, provisional (working) hypothesis

JEL Classification

D80, D83

INTRODUCTION

The enhanced opportunities and a search for new tools have always aroused great interest, owing to their crucial importance for the development of human civilization. This primarily pertains to one of the main branches of cognition – scientific cognition aimed at gaining objective knowledge about the surrounding reality, being a fundamental prerequisite for organization and efficient implementation of almost any human activity. Science studies only the reality that is accessible to observation, and it is represented as a systematic summary and presentation of knowledge mined from practice, i.e., an explanation of empirical data or empirical evidence.

In recent decades, data mining (DM) has become widely used, i.e. a search in the large volumes of empirical data for the unobvious and nontrivial regularities which can be used in practice, for example, for decision-making. This happened in response to the practical needs in the field of industry, business, R&D, medicine, military technology, etc., as well as in the context of evolving capacities of computers, which enabled the accumulation and processing of Big Data.

As a rule, analyzed objects have two basic characteristics. Firstly, these objects are considered multidimensional and are described by a large

number of features that assess their properties. Secondly, these objects are quite numerous, which enables, through studying their similarities and differences, to identify the regularities common for multiple objects. At the same time, the use of conventional mathematical and statistical tools for the analysis of data turned out to show low efficiency. On the contrary, DM methods demonstrated their effectiveness as an indispensable tool in the search for knowledge of practical use, deriving value from data. The rapidly expanding use of DM methods shows their efficiency in analytical activities.

DM algorithms, implemented as computer programs, have developed a new research tool. Mass digitization of historically accumulated data in many industries has led to the emergence of the so-called Big Data technologies, which are similar to DM and embedded in the respective management processes, enabling to process Big Data. At the same time, a widespread vigorous application of DM methods raises new questions about whether there is a correct understanding of their capabilities and limits as well as the outcomes in terms of scientific cognition. Today, there is no holistic DM methodology, and, as a result, it reduces the efficiency of the application of such methods and blocks the formation of this innovative interdisciplinary field (along with Big Data and Data Science).

1. LITERATURE REVIEW

One of the recognized methodologists on pattern recognition – PR (as the identification of regularities in data, i.e., actually DM, was formerly called), Bongard (1967) believed that recognition is one of the most important blocks for the simulation of thinking. The problem of recognition was called to be a part of the problem of cognition. Such aspects of PR as identification and classification of objects, imitation, and search, enumeration of possibilities, etc., being present in a wide range of practical problems, are associated with cognition. This makes it necessary to raise the issue of determining the place of DM in the methodology of cognition.

According to Zagoruiko (1972), PR is one of the ways to build the models that explain multiple experimentally obtained facts from different fields of human activities (diagnosis of diseases in the medical field, hardware failure prediction, analysis of photographs, recognition of speech signals, sociological studies and much more). The absence of the holistic theory and methodology of PR was emphasized. Moreover, Zagoruiko (1999) demonstrated significant addition to methodological best practices (understanding of data, knowledge, basic hypotheses; building the system of attributes; methods and algorithms, etc.) Nevertheless, poor development of DM methodology itself as a cognitive tool was noted. At the same time, the research value of DM was demonstrated using the example of rediscov-

ery of Ohm's law, Mendel's law, and Mendeleev's periodic law, which also proves the importance of methodological issues for the efficient application of DM in science (Zagoruiko, 1999).

Using the example of soil classification, Rozhkov (2011) showed the specifics of the application of the information approach in this field of knowledge and the crucial importance of taxonomy (clustering) and meronomy (search for differences, classification), which are two major challenges of DM. Soil classification demonstrates the evolvement of the paradigms of systematization. It is aimed to develop the knowledge base on soils that is evidence of a more in-depth cognition in this critical area of science. Rozhkov (2011) showed the universal nature and generality of these DM methods for the cognition of the world around and thereby complimented other (physical and chemical, nuclear, biological, etc.) approaches.

Zakrevskii (1988) believed that identification of regularities in the data flow is the basic way of the scientific cognition of the surrounding reality to gain knowledge as a basis for rational actions. Regularity was considered as the central notion and it was understood as a certain quite strong link between the attributes of the observed phenomena. The regularities are less stringent than the laws and are similar to hypotheses.

The increasingly complex challenges of the modern world (process management, organization of interactions, prediction, etc.) are assigned to au-

tomatic technical devices. This, on one hand, addresses the ontological issues and, on the other hand, implies carrying out comprehensive work on data processing using computers, which is far beyond human capabilities. Therefore, the study of DM as a cognitive tool becomes more important, finding out the outcomes of using its methods that are introduced in the methodology of scientific cognition (for example, what regularities are identified, the limits of methods applicability, the search for more advanced solutions, etc.).

A study of complex systems and their exploring at the initial stages have become to a considerable degree relying on experimental data and approaches, including their processing, based on DM. Owing to the progress of computer and measuring technology, the amount of data available for the description of complex systems has significantly increased, as well as data processing capabilities. It creates the required empirical framework for modeling, prediction, and management of the behavior of complex systems (Hastings et al., 2017). At the same time, the issues related to the applicability limits in terms of science and cognitive value of DM methods as well as the processing outcomes are still outstanding.

Survey studies confirm a sustainable growth of publications that focus on the use of DM methods (Liu et al., 2019). DM has combined several disciplines and knowledge systems, such as statistics, computer science, data collection, special mathematical methods, algorithmic, computer calculations, machine learning, visualization, management of databases and repositories (Han et al., 2012), and, sometimes, communication, sociology, and management.

There are many examples of the efficient use of DM methods in various fields of scientific studies, where large volumes of multidimensional data have been accumulated, namely the Earth science (Chen et al., 2019); environmental studies (Gibert et al., 2018); life science (Agapito et al., 2018); chemistry (Szymańska, 2018); medicine (Thakkar et al., 2021).

Industry 4.0 concept involves collection, storage, management, and analysis of the data generated by production systems to manage (for automated

identification of failures, assessment of operating conditions and quality of products, identification of the unplanned stops, etc.). However, many companies refuse to apply DM because of the poor quality of the outcomes, which is mainly due to the reasons raised in this paper (Schuh et al., 2019; Kozjeka et al., 2019).

Often, DM is indispensable to ensure the proper functioning of technologies, for example, to detect malicious network intrusions (Salo et al., 2018) or analyze network alarms (Zheng et al., 2020). This, to a great extent, is a prerequisite for the further development of such modern sophisticated technologies as the Internet of things (Li, 2020) that links multiple “smart devices” equipped with sensors and actuating mechanisms. The conversion of data into practical knowledge is considered a part of the technology itself. Its intelligent environment is required for the efficient operation and management of resources and services (Sunhare et al., 2020).

A wide range of the problems of DM application is found in social sciences. Given the specifics of the objects (phenomena and processes) being studied, major difficulties are associated with the development of their feature vectors and selection of methods that enable to obtain the most accurate results (Santhosh & Mohanapriya, 2020). The specifics of “Big social Data” and the high dynamics of its changes necessitate a combination of various methods and correction of feature vectors (Cuomo et al., 2021).

The analysis has highlighted that the increased level of saturation with data synthesized from different sources increases the need for its analytical processing to derive new knowledge and, respectively, the requests for the application of DM. Significant growth in the volume, dimensionality, diversity, and, sometimes, update speed of data limits the use of conventional mathematical and statistical tools for its analysis because Big Data violates basic assumptions, which form the basis of these tools. On the other hand, a reliable analysis of data is a more serious problem. At the same time, all fields of active DM application experience similar problems of methodological nature. New approaches to planning and implementation of research projects are required, combining various as-

pects of analysis (assessment of the quality of data, selection of the strategy for preliminary processing of data, visualization of data, verification of models, etc.), which can significantly enhance the accuracy of the result.

Most studies raise the question that is rather related to the methodology of cognition: “What knowledge can be derived from the accumulated data and what is its level?” This question demonstrates the immaturity of the DM concept. It also summarizes multiple practical problems of DM, which are not addressed by enhancing computing capabilities or parallel computing in the field of Big Data processing (Carbon et al., 2016). Besides the difficulties of the right choice and application of DM methods, there is no full understanding of its capabilities and methods as well as the process (phasing) and the obtained results in terms of cognition. The study of complex systems as multidimensional objects requires not only complex thinking but also an understanding of what will be obtained after the processing of diverse Big Data, characterizing its features. An understanding of the capabilities and limits of DM may result in the significant modification of the methodology for the study of complex systems in general.

Initially, DM methods were actively used in natural-science areas, and further – in social fields and management. The realized opportunities (and the growth of the volume of data) made multiple sciences and disciplines data-driven. Interdisciplinary nature in no way limits the development of DM as an independent scientific field. However, so far there is too much hype around DM; misunderstanding about these methods, myths, and even speculations about their application are quite widespread, as well as excessive use of essentially similar terms Big Data and Data Science, for example, in the project justifications. This is an intelligence trap for many people who are not experts in mathematics and computer science. Most of the discussions and publications, which take place, are primarily related to the existing concepts of DM, which is not accompanied by any considerable improvements in terms of methodology. What is needed is a more active and wide-ranging discussion related to the internal problems of DM and its gaps as applied to different disciplinary fields (Cao, 2017).

Therefore, the practice of analytical work has shown that DM is indeed a powerful cognitive tool, which has an interdisciplinary significance. Moreover, DM methods may serve as a basis for the convergence of approaches to cognition in both humanities and natural sciences. The analysis also leads to the conclusion that DM addresses a huge number of the applied problems and improves the data mining algorithms themselves. However, in terms of the methodology, very little is being done and almost no activities are carried out in this field, which substantially hinders further development of DM that, generally speaking, could become a basis for a disciplinary revolution in the theory of cognition, and even enable to generate major innovations in the field of intelligent technologies.

2. AIMS

The study aims to specify the basis for the understanding of DM capabilities and limits as well as its results within the methodology of scientific cognition. This will enhance the efficiency of using DM methods by experts in this field as well as by a wide range of professionals in various other fields, who need to analyze empirical data, in particular when the adapted (for non-experts in mathematics and computer science) data mining tools are used.

3. GENERALIZATION OF MAIN STATEMENTS

The process of cognition is a process of gaining and using knowledge, which is of staged nature (Moiseev, 1982). Cognition is based on a strive to know, which links different types of cognitive activities, in particular, imagination, enthusiasm, aggregation of facts, reasoning, etc., often resulting in new ideas, observation, concepts, knowledge, and solutions. Specially developed algorithms and systems to analyze empirical data may facilitate thinking processes. At the same time, it should be noted that data aggregation, computational logic, and operations can be built in machines, and imagination, reasoning, and creative work cannot do it; in the course of analyzing the data, apart from the explicit element, many implicit ones are involved. However, the tools, which simulate human thinking, may not yet encompass, embody and synthesize knowledge, experience, and intel-

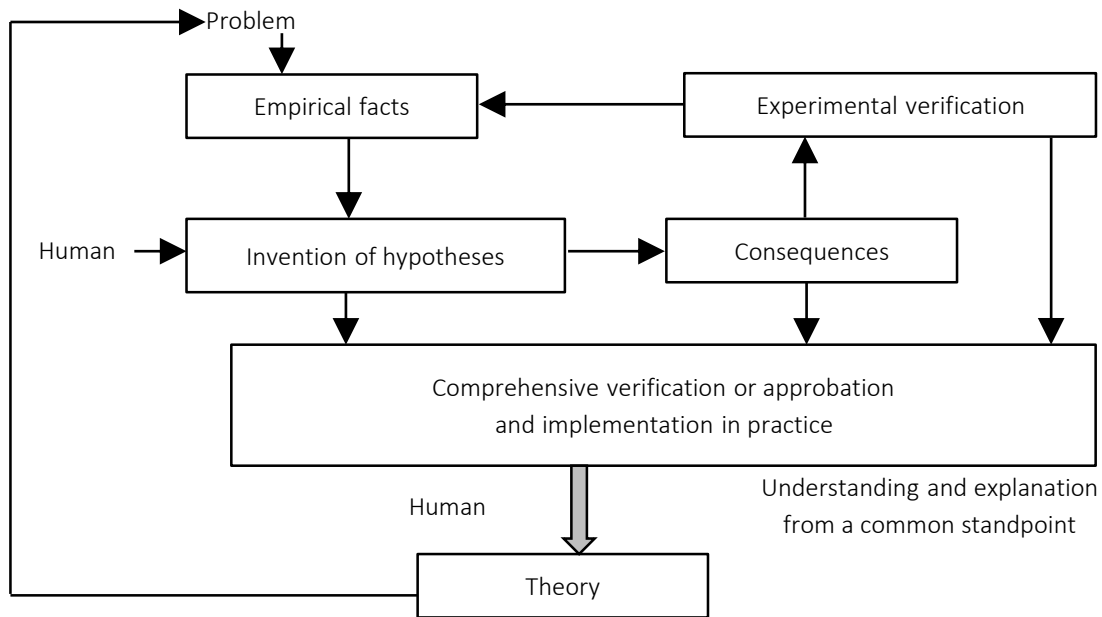


Figure 1. General scheme of cognition

ligence operations of different levels, which are required to analyze the data and verify the results, and that makes human involvement mandatory.

It is necessary to consider the main stages of cognition. The first stage of cognition is experience, observation, experiment, a study of the phenomenon, in other words – accumulation of facts for further analysis. The second stage is summarizing the facts, identifying their essential parts, forming hypotheses and conclusions on their basis, i.e. certain abstraction from the first stage. At the third stage, the hypotheses or conclusions that were made before are verified. This is a universal scheme of cognition. Gnedenko (1983) quite clearly and concisely formulated this scheme: 1 stage – observation or living contemplation, 2 stage – transition to abstraction, 3 stage – testing abstraction in practice. This is the dialectic way of knowing the truth. For illustrative purposes, this scheme can be shown in graphical form (Figure 1).

The knowledge of the world is gained using different methods of cognition. The methodology of science accepts the following methods, enumerated from the most elementary methods to the most complex ones (Shtoff, 1978):

- 1) Technique – the lowest level, the examples are detections, different kinds of empirics (any empirical method, leading to a certain result).

- 2) Scientific method, relying on knowledge of the respective regularities, i.e., the theory of the given subject area.

- 3) General scientific method – a quite general method of scientific study, where the applicability extends the limits of one or another scientific discipline and relies on the existence of regularities, being common for different areas.

- 4) Methods used in all sciences without exception, although, in different forms and modifications. It is the most general method of scientific cognition, and their study is the subject of philosophical methodology.

If a certain study addresses single-type or standard problems, and a certain method is used, then, as a rule, the issues of methodological nature do not arise. Everything is clear – it is an experiment or calculation and it fits the existing paradigm of the study. However, if the study is related to a new field or is at the interface between different sciences, then the methodological issues arise and the answers to them show the level of the study that is carried out, for example: “Is there a technique or scientific method?”; “What should be done or what conditions should be met to move to the next level?”; “Is it possible to automate the intellectual work?”; “Is it possible to convert the knowledge,

which was previously gained in a declarative way, into procedural one, that is to track the way of solving the problem?”, etc. These issues became particularly pronounced when using computers for data mining. The key issue, being critical in terms of cognition, is what DM introduced into the methodology of scientific cognition and what the application of its outcomes can result in?

It is customary in DM to make a difference between two major challenges – classification and clustering. Classification is generally understood as:

- 1) verification – reasonable categorization of the objects under study by classes;
- 2) identification – assignment of new objects to one or another class.

Classification is generally a subject to specific study or management goals, which gives grounds for the formulation of the problem of feature selection to describe the properties of objects and select the method of classification. As a result of the learning procedure, there is a regularity that enables to conclude in respect of the characteristics of

a specific group of objects as well as the features that, in any case, differentiate one group from another, which could enable to establish causality. Clustering is generally understood as the division of the aggregate of objects under study by clusters (classes) based on the special algorithms for assessment of the similarity of object features within clusters, and, accordingly, the distinction between the objects from different clusters. Clustering may be considered as a stage prior to classification (if there is no preliminary division of the dataset by classes) as well as a stand-alone stage of structural and comparative analysis.

Still, at the early stages of PR studies, the issues of their role in cognition arose. In particular, Malinovskii (1986) suggested the scheme of cognition based on the principles of classification and recognition of patterns, which is shown in Figure 2.

The scheme accurately describes a general approach to cognition with the use of classification principles. However, it is impossible to determine the place of DM tools in the above-listed methods of cognition and, therefore, to answer the questions raised above. At the same time, tangible phenomena are one thing, and information artifacts,

Source: Malinovskii (1986).

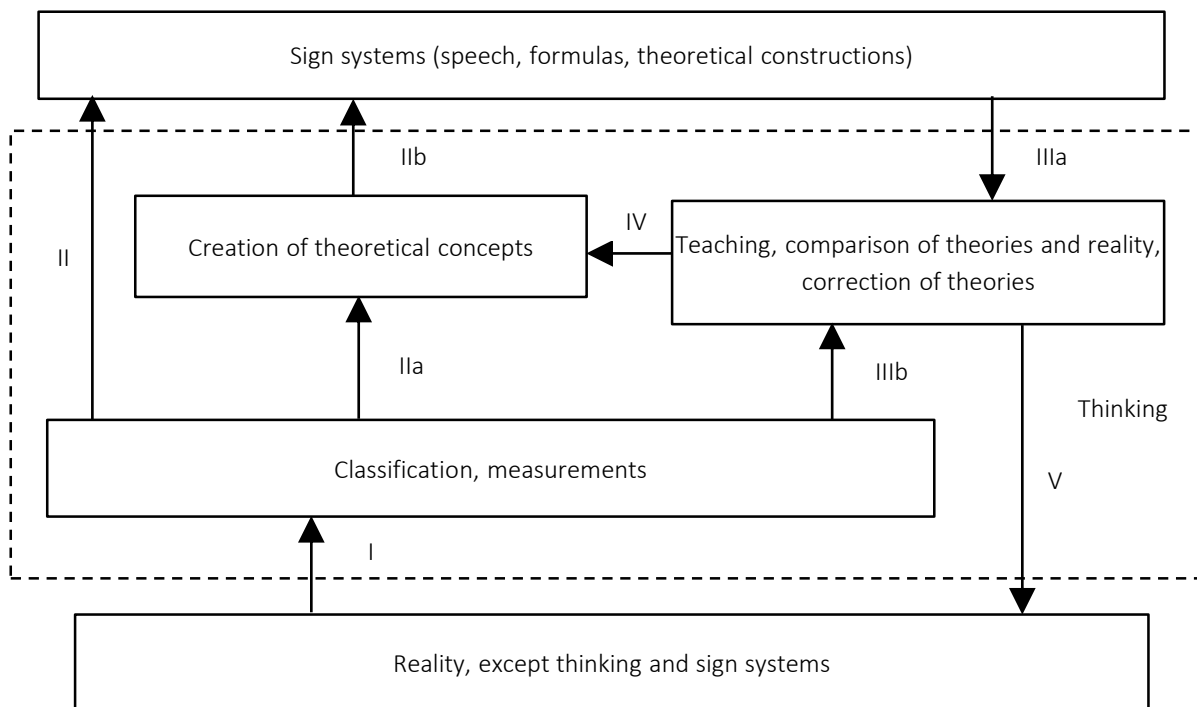


Figure 2. Correlation between thinking, reality, and sign systems

“virtual” phenomena, social phenomena, and processes are entirely different things. There is a wide range of problems, where such kinds of objects must be studied to analyze their features and related facts and to assess the trends.

In general, the application of DM tools starts when the prepared data is available in the form of datasets, where the objects are represented by the sets of multidimensional data (feature vector). These kinds of datasets have different names – object-property table, training dataset (TD), or the table of experimental data. They are considered equivalents to each other. The issues related to the selection of feature vector (feature space) and pre-processing of data are beyond the competence of DM, although these issues arise in solving any of DM problems. At the same time, it is believed that these issues fall within the competence of the expert in the subject area that DM analyst cooperates with or obtains data from.

To answer the question of the contribution of DM to the methodology of scientific cognition, there is a need for a benchmark, which would show the

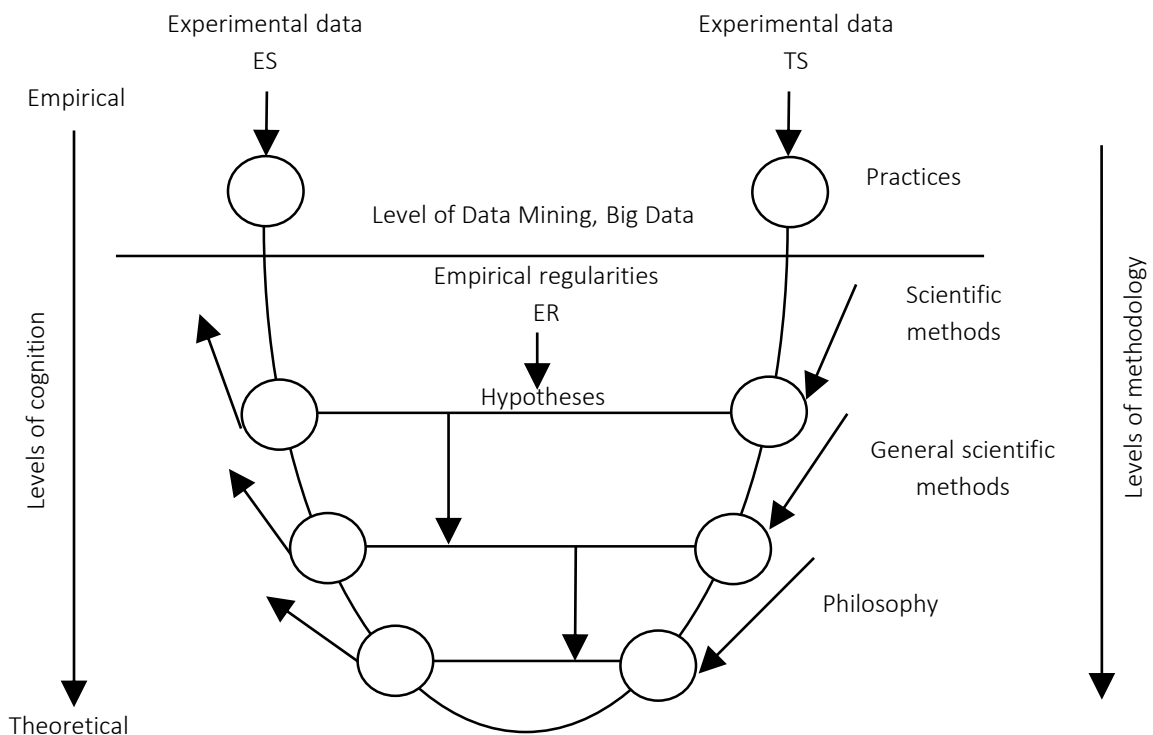
level of the cognitive value of the model that was obtained using DM.

The analysis has shown that these kinds of benchmarks are almost unavailable, and those which are available do not highlight the methodological nuances of the application and are quite general descriptions that do not bear any constructive load.

Therefore, it is proposed to complement the classification of cognition methods in the form of the list of 1-4 paragraphs suggested by Shtoff (1978) in the scheme shown in Figure 3 – a sort of graphical supplement to these paragraphs.

The main purpose of this scheme is to show the relationship between the levels of cognition, and, the most important thing, to demonstrate the limits of DM methods as well as the possibility of the direct and quite understandable transition from practice to the scientific method through the generation of empirical regularities (ER), and further, from the latter to the hypotheses, and from hypotheses to the scientific method.

Source: Authors' elaboration.



Note: ER – empirical regularities; TD – training dataset; VD – validation dataset.

Figure 3. Relationship between the levels of cognition

When using any of DM methods, the outcome is represented in the form of one or another model, reflecting certain regularities intrinsic to the data under study, which might logically be called empirical regularities (ER) and which are hypotheses in nature (Zakrevskii, 1988). In Figure 3, the level of ER is highlighted with the “red line”. It is the basis for the possible transition from practice to provisional hypothesis and further, to scientific methods. ER as an outcome is quite understandable to the expert in the subject area and suitable for further processing. It is required because data is considered in a certain context to be given a meaning, which also affects the use of the outcome that was obtained (Zagoruiko, 1972, 1999)

The answer to the question of the role of DM in the methodology of cognition is clear from Figure 3. DM provides an opportunity for automated generation of ER, being the “building blocks” for making hypotheses as a part of addressing a specific problem. It means that the emergence of hypotheses, which are the driving force of science, is preceded by a very critical stage of ER generation (search) – this is precisely the contribution of DM to the process of cognition. Furthermore, this stage occurs automatically, based on the algorithms invented by human beings and implemented in the form of computer programs (a human just selects the suitable algorithm).

The further process also becomes clear – transition from ER to scientific method – the volume of data “captured” with every new confirmation of such transitions should keep increasing (Figure 3), which constitutes a solid basis for the development of a scientific method to be used in a specific subject area.

The ER found should be considered a prerequisite for formulating a hypothesis, some kind of provisional hypothesis. It also requires carrying out certain intellectual work and, probably, additional studies. The transition from ER to hypothesis does not take place automatically; it requires certain human efforts, although, with the use of DM, a quite clear direction for finding this hypothesis became evident. There is a need for additional studies, which might also be considered to a great extent an extension of DM, and it is well illustrated in Figure 3. It is related to the use of almost all known methods.

Therefore, the outcome that can be achieved directly in the application of any of DM tools is the level of ER, and further, the hypotheses based on them. The rules are defined, but it is not yet possible to explain the observed phenomena. There is one thing related to the use of software algorithms and it consists in the fact that these provisional hypotheses can be called man-machine. Such duality of the name reflects the root of the matter: on one hand, these hypotheses are based on ER that are generated by a computer, to be more precise, by program. On the other hand, the program contains the algorithms to find ER that are developed and selected by a human being, and a human being assesses the obtained results. It is a new phenomenon, which influences the course of cognition itself, and it should be methodologically understood.

It is vital to mention such a class of DM models as neural networks. In many cases, the use of neural networks yields good results; however, unfortunately, it actually gives nothing in terms of cognition. Their level is limited by the level of “primitive” (like it is in animals) recognition (classification) and nothing more, and this in itself is not new knowledge. They represent no regularities like ER and there is no reason to talk about the man-machine generation of hypotheses. In terms of cognition and methodology, it is a dead-end type of DM. A huge number of observations can be organized (similar to neural networks, because they are “fed” with a large amount of data), but no casual relationships can be found and, respectively, understood. Therefore, it is impossible to move forward in terms of methodology – new laws in any subject area will hardly be ever discovered with their help.

A thorough consideration of the scheme in Figure 3 enables to understand the role of DM in the methodology of cognition as well as to clarify if it is possible to transit from hypothesis to scientific method. For this purpose, when testing a man-machine hypothesis, the volume of the encompassed data should increase every time until there is a sufficient number of its confirmations, which will be a guarantee or a sound basis for further development of a scientific method, which might be used in this subject area in future. This is the next step that should be made in terms of the methodology of cognition.

It should be noted that scientific studies need not only the hypotheses that just explain something. According to Pavlov, the explanation is not yet a science; science is distinguished by absolute domination and prediction (cited in Iugov, 1942). There should be hypotheses – predictions and hypotheses on possible ways to solve the problem. Lekakh (2011) called them working hypotheses; this is particularly important for the applied studies that are characterized by the existence of a purpose. It is possible to agree with the opinion that “working hypotheses are the assumptions about possible ways of addressing the problems; they are looking not towards the past, but towards the future, and, if the idea about the solution is determined, the prospects of success will increase significantly”.

It should be emphasized that it is ER obtained using DM methods as provisional hypotheses enable to make working (in the above sense) hypotheses, which in many ways can predetermine problem-solving. It is the promotion of the “working”, according to Lekakh (2011), hypotheses that, for example, the cytochrome processing website Data4logic is designed to, which enables medical researchers to automatically generate ER and, with a high probability of success, to address the problems they faced. The patterns stipulated by the paper related to leukemia diagnostics (Gluzman et al., 2000) can be used as an example of this approach and for attempts to make working hypotheses, being of the form of logical expressions like “if..., then...”, using the service with further processing of the obtained characteristics with an application of the tools of the same website. However, it is related to medical staff or experts in cytology.

In many cases, problem-solving is limited to, in terms of cognition, the level of hypothesis, to be more precise, the level of ER (or provisional hypothesis), used as a basis for the further formulation, in a best-case scenario of a decision-making direction or rule, and it remains at the first, the lowest of all possible levels, empirical level of cognition. In the short run, it suits business as a sphere of practical activities; however, in the long run, the main thing is lost – finding new knowledge that can be implemented in innovations, or development of a new method that will provide

a higher-order competitive advantage. Similarly, the level of “primitive” classification inherent to neural networks often suits the business.

Consequently, it can be ascertained that DM methods are capable of providing only the level of empirical cognition in the specific subject area under study as well as the level of techniques and directions, which completely fits the scheme shown in Figure 3.

Now, it becomes clear why there are no “breakthrough” inventions made using DM – because now such inventions can take place only in a specific subject area, and this requires close cooperation and interaction as well as full-fledged scientific communication with the representatives of the same subject area, which is the biggest obstacle to such kind of achievements.

Hence, the following conclusions can be drawn.

- 1) The methods of DM as well as Big Data are a new man-machine methodology of empirical cognition.
- 2) These methods have their limitations in the form of ER (provisional or working hypotheses, according to Lekakh (2011)) represented in different forms.
- 3) ER can serve as “drafts” for the formulation, validation, and selection of hypotheses aimed at more in-depth cognition of the subject area.
- 4) To select the best strategy for the use of DM tools, a clear understanding of the goals of problem-solving is required.
- 5) The use of DM tools requires close cooperation with the experts in a specific subject area that, in its turn, raises several questions related to the initiation of such cooperation; skillfulness of the experts in the subject area; statement of the problem in the respective context; building the team to solve the problem, etc.
- 6) DM and Big Data experts’ “shifting” to the area of development of the standardized software (cloud services, web-services, desktop applications) does not solve the problem of in-depth

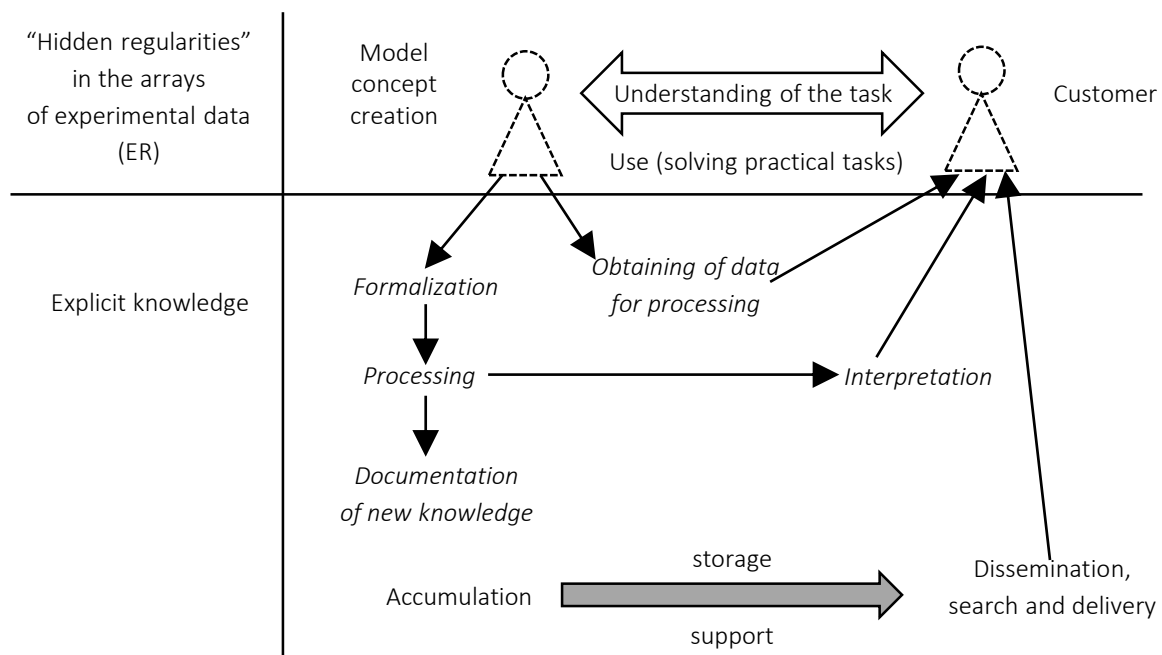


Figure 4. Search scheme for hidden empirical regularities

cognition; there is still a limit represented by empirical cognition – obtaining of ER, i.e., in fact, provisional hypothesis for the given specific subject area. In this case, the burden of solving the specific problem to deepen cognition and clarify the hypotheses is fully transferred to the experts in the subject area.

They face the problems related to learning the respective software, comprehend its major capabilities, and, most important, find out whether the selected software is appropriate to solve practical problems in the given area. At the same time, the following issues, being critical for problem-solving, are “left behind”: selection and development of feature vector, understanding of the subject area, verification of data (quality and usefulness), selection of the proper processing algorithm, assessment of the reliability of the method, interpretation of the results and transition from data to solutions and actions. Most often, such issues are addressed intuitively, based on the experience or analogy, through trial and error.

The full-fledged interaction between the experts in subject areas and data scientists is significantly more painstaking in terms of organizational and communicative cost. However, this approach can ensure breakthroughs in the subject area. An interim option is also possible and now it begins to

be actively used in business. Many companies realized that without efficient “task setters” and analytics well-versed in DM tools, just the use of desktop, web, and cloud services is inefficient.

Therefore, the main conclusion is that, for efficient use of DM and obtaining of maximum possible results in cognition with its help, i.e., hypotheses, the primary focus should be on preliminary analytics, formulation of problems, and selection of features to describe the objects. It becomes the axioms of the use of DM. The problem should be thoroughly studied, reasonably selecting the features, drawing out the feature vector of the objects, collecting and verifying data, and only then it is possible to expect the achievement of acceptable results with the use of various special applications. Hence, the practical importance of the suggested view of DM enables to point out the limit of its applicability. This limit is ER and, as a consequence, provisional or working hypotheses, which can be derived from the dataset, provided that the above-stated steps are made during its development. Moreover, how the process will proceed depends on the understanding of the general problem as well as on the need to achieve a practical result. If business problems require the soonest practical results, one is satisfied with what has already been achieved, if formally all criteria are relevant (Figure 4).

In scientific studies, if there are prerequisites for further work, there is a progression by stages from data to knowledge, and, further, to understanding. Using ER as initial benchmarks, having accepted them as working hypotheses, the field of experiment is expanded and hypotheses using a new material are validated. If the hypotheses are confirmed, then a certain knowledge is generated (in many cases, this generation of knowledge might facilitate the emergence of innovations), and further, it is necessary to find out what its basis is. There is a transition to the understanding of the process, i.e. what it can mean and what conclusions can follow from this knowledge, which

will enable to speak of the possible approaches to problem-solving.

If there is an understanding, it might mean that there is something that the gained knowledge is based on, and, probably, it will be a new method to solve the above-mentioned problems. Knowledge is data, and understanding is the ability to draw conclusions. Currently, this is a limit of applicability of all DM tools. They generate knowledge in the form of ER and working hypotheses, but they do not provide understanding, without which a transition to the next level of cognition – the scientific method – is impossible.

CONCLUSION

Knowing the limits of DM tools, it is possible, with a significantly better understanding, to proceed to the formation of targets when selecting the appropriate methods of DM. For example, to choose the methods, which provide a relatively large set of ER, or to use the methods, which provide a limited set of such regularities characterized by greater precision. The most critical fact in terms of methodology has been established – the limits of the applicability of DM methods are the level of ER or provisional (working) hypothesis. A huge number of methods and techniques developed a variety of computer programs, cloud services, and other support – all this is ultimately reduced to theory or laws, and it is limited only to one thing – the level of ER or provisional (working hypothesis). It is worth continuing in terms of philosophy, as of today, that is the only visible achievement of all DM algorithms. Should the obtained result be considered critical in terms of cognition? The answer is yes. Although, it should be emphasized that everything mentioned above is related to a certain subject area, where data mining methods are applied. It should be noted that DM can be understood as an evidence-based/constructive method of cognition with all advantages and disadvantages. Today, the finding of ER (working hypotheses) is implemented in the form of web services (for example, ScienceHunter portal); therefore, future studies will be focused on the development of the concept of an automated system for DM, which will be appropriate for the experts who do not have any special background in the field of mathematics and computer science.

AUTHOR CONTRIBUTIONS

Conceptualization: Maxim Polyakov, Igor Khanin, Vladimir Bilozubenko, Gennadiy Shevchenko.

Formal analysis: Maxim Polyakov, Vladimir Bilozubenko.

Investigation: Maxim Polyakov.

Methodology: Igor Khanin, Vladimir Bilozubenko, Gennadiy Shevchenko.

Project administration: Gennadiy Shevchenko.

Supervision: Maxim Polyakov, Igor Khanin, Gennadiy Shevchenko.

Writing – original draft: Maxim Polyakov, Igor Khanin, Vladimir Bilozubenko, Gennadiy Shevchenko.

Writing – review & editing: Maxim Polyakov, Igor Khanin, Vladimir Bilozubenko, Gennadiy Shevchenko.

REFERENCES

1. Agapito, G., Guzzi, P., & Cannataro, M. (2018). Parallel and Distributed Association Rule Mining in Life Science: a Novel Parallel Algorithm to Mine Genomics Data. *Information Sciences*. <https://doi.org/10.1016/j.ins.2018.07.055>
2. Bongard, M. M. (1967). *Problema uznvaniya [Recognition Problem]*. Moscow: Nauka. (In Russian).
3. Cao, L. (2017). Data Science: Challenges and Directions. *Communications of the ACM*, 60(8), 59-68. <https://doi.org/10.1145/3015456>
4. Carbone, A., Jensen, M., & Sato, A.-H. (2016). Challenges in data science: a complex systems perspective. *Chaos, Solitons & Fractals*, 90, 1-7. <https://doi.org/10.1016/j.chaos.2016.04.020>
5. Chen, W., Pourghasemi, H. R., Zhang, S., & Wang, J. (2019). 21 – A Comparative Study of Functional Data Analysis and Generalized Linear Model Data-Mining Methods for Landslide Spatial Modeling. In H. R. Pourghasemi & C. Gokceoglu (Eds.), *Spatial Modeling in GIS and R for Earth and Environmental Sciences* (pp. 467-484). Elsevier. <https://doi.org/10.1016/B978-0-12-815226-3.00021-1>
6. Cuomo, M. T., Tortora, D., Foroudi, P., Giordano, A., Festa, G., & Metallo, G. (2021). Digital transformation and tourist experience co-design: Big social data for planning cultural tourism. *Technological Forecasting and Social Change*, 162, 120345. <https://doi.org/10.1016/j.techfore.2020.120345>
7. Data4Logic. (n.d.). *Finding cells attributes*. Retrieved from <https://www.data4logic.net/en/Services/CellsAttributes>
8. Gibert, K., Izquierdo, J., Sánchez-Marrè, M., Hamilton, S. H., Rodríguez-Roda, I., & Holmes, G. (2018). Which method to use? An assessment of data mining methods in Environmental Data Science. *Environmental Modelling & Software*, 110, 3-27. <https://doi.org/10.1016/j.envsoft.2018.09.021>
9. Gluzman, D. F., Abramenko, I. V., Skliarenko, L. M., & Kriachok, I. A. (2000). *Diahnostika leukozvo [Diagnosis of leukemia. Atlas and practice guidelines]*. Kiev: MO-RION. (In Russian).
10. Gnedenko, B. V. (1983). *Matematika i nauchnoye poznaniye [Mathematics and scientific knowledge]*. Moscow: Znanie. (In Russian). Retrieved from <https://www.litmir.me/bpr/?b=578209>
11. Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Methods* (3rd ed.). Elsevier Inc. <https://doi.org/10.1016/C2009-0-61819-5>
12. Hastings, H. M., Davidsen, J., & Leung, H. (2017). Challenges in the analysis of complex systems: introduction and overview. *The European Physical Journal Special Topics*, 226, 3185-3197. <https://doi.org/10.1140/epjst/e2017-70094-x>
13. Iugov, A. K. (1942). *Ivan Petrovich Pavlov*. Moscow: Detgiz. (In Russian).
14. Kozjeka, D., Vrabič, R., Kralj, D., Butala, P., & Lavrač, N. (2019). Data mining for fault diagnostics: A case for plastic injection molding. *Procedia CIRP*, 81, 809-814. <https://doi.org/10.1016/j.procir.2019.03.204>
15. Lekakh, V. A. (2011). *Bolnyye voprosy onkologii i novyye podkhody v lechenii onkologicheskikh zabol-evaniy [Pressing issues of modern oncology and new approaches to the treatment of oncological diseases]*. Moscow: Librokom. (In Russian).
16. Li, L. (2020). Real time auxiliary data mining method for wireless communication mechanism optimization based on Internet of things system. *Computer Communications*, 160, 333-341. <https://doi.org/10.1016/j.comcom.2020.06.021>
17. Liu, J., Kong, X., Zhou, X., Wang, L., Zhang, D., Lee, I., Xu, B., & Xia, F. (2019). Data Mining and Information Retrieval in the 21st century: A bibliographic review. *Computer Science Review*, 34, 100193. <https://doi.org/10.1016/j.cosrev.2019.100193>
18. Malinovskii, L. G. (1986). Protsessy klassifikatsii – osnova postroyeniya nauk o deystvitel'nosti [Classification processes are the basis for the construction of the sciences of reality]. In I. A. Ovseevich (Ed.), *Algoritmy obrabotki eksperimentalnykh daniykh [Experimental data processing algorithms]* (pp. 155-182). Moscow: Nauka. (In Russian).
19. Moiseev, N. N. (1982). *Chelovek, sreda, obshchestvo. Problemy formalizovannogo opisaniya [A person, environment, society. Problems of formalized description]*. Moscow: Nauka. (In Russian). Retrieved from <https://www.libex.ru/detail/book952963.html>
20. Rozhkov, V. A. (2011). On an Information Approach to Soil Classification. *Dokuchaev Soil Bulletin*, 69, 4-24. (In Russian). <https://doi.org/10.19047/0136-1694-2012-69-4-24>
21. Salo, F., Injadat, M., Nassif, A. B., Shami, A., & Essex, A. (2018). Data Mining Methods in Intrusion Detection Systems: A Systematic Literature Review. *IEEE Access*, 6, 56046-56058. <https://doi.org/10.1109/ACCESS.2018.2872784>
22. Santhosh, R., & Mohanapriya, M. (2020). Generalized fuzzy logic based performance prediction in data mining. *Materials Today: Proceedings*, 45(2), 1770-1774. <https://doi.org/10.1016/j.matpr.2020.08.626>
23. Schuh, G., Reinhart, G., Prote, J.-Ph., Sauer mann, F., Horsthofer, J., Oppolzer, F., & Knoll, D. (2019). Data Mining Definitions and Applications for the Management of Production Complexity. *Procedia CIRP*, 81, 874-879. <https://doi.org/10.1016/j.procir.2019.03.217>
24. ScienceHunter. (n.d.). *O nas [About us]*. Retrieved from <https://www.sciencehunter.net>
25. Shtoff, V. A. (1978). *Problemy metodologii nauchnogo poznaniya [Problems of scientific knowledge*

- methodology*]. Moscow: Vysshaya shkola. (In Russian). Retrieved from <https://www.twirpx.com/file/1849590/>
26. Sunhare, P., Chowdhary, R. R., & Chattopadhyay, M. K. (2020). Internet of things and data mining: An application oriented survey. *Journal of King Saud University – Computer and Information Sciences*. <https://doi.org/10.1016/j.jksuci.2020.07.002>
 27. Szymańska, E. (2018). Modern data science for analytical chemical data – A comprehensive review. *Analytica Chimica Acta*, 1028, 1-10. <https://doi.org/10.1016/j.aca.2018.05.038>
 28. Thakkar, H., Shah, V., Yagnik, H., & Shah, M. (2021). Comparative anatomization of data mining and fuzzy logic methods used in diabetes prognosis. *Clinical eHealth*, 4, 12-23. <https://doi.org/10.1016/j.ceh.2020.11.001>
 29. Zagoruiko, N. G. (1972). *Metody raspoznavaniya i ikh primeneniye [Recognition Methods and Their Application]*. Moscow: Sovetskoe radio. (In Russian). Retrieved from <https://www.twirpx.com/file/382297/>
 30. Zagoruiko, N. G. (1999). *Prikladnyye metody analiza dannykh i znaniy [Applied methods of data and knowledge analysis]*. Novosibirsk: Sobolev Institute of Mathematics, SBRAS. (In Russian). Retrieved from <https://www.docme.ru/doc/1762951/zagoruiko-n.g.-prikladnye-metody-analiza-dannyh-i-znani>
 31. Zakrevskii, A. D. (1988). *Logika raspoznavaniya [Recognition logic]*. Minsk: Nauka i tekhnika. (In Russian). Retrieved from <http://www.aiportal.ru/downloads/books/logic-recognition-by-zakrevsky.html>
 32. Zheng, Q., Li, Y., & Cao, J. (2020). Application of data mining technology in alarm analysis of communication network. *Computer Communications*, 163, 84-90. <https://doi.org/10.1016/j.comcom.2020.08.012>