Capabilities of Data Mining As a Cognitive Tool: Methodological Aspects

Genady Shevchenko [1], Oleksander Shumeiko [2] and Volodymyr Bilozubenko [3]

[1] *Scientific Center, Noosphere Company, Gagarin avenue, 103-A, Dnipro, 49055, Ukraine*
[2] *Dniprovsk State Technical University, Dniprobudivska Street, 2, Kamyanske, 51900, Ukraine*
[3] *Scientific Center, Noosphere Company, Gagarin avenue 103-A, Dnipro, 49055, Ukraine*

Abstract

Gaining a competitive advantage in many industries is possible only if the available digitized data contains genuine knowledge. In this respect, it is necessary to take a step to preliminary identify their hidden and non-obvious regularities using Data Mining (DM) methods. It is critical to know the capabilities and limits of the use of DM methods as a cognitive tool in order to build the effective strategy for addressing the real-life business problems.

The aim of this paper: within the methodology of scientific cognition to specify the capabilities and limits of the applicability of DM methods. This will enhance the efficiency of using these DM methods by experts in this field as well as by a wide range of professionals in other fields who need an analysis of empirical data.

The paper specifies and supplements the basic stages of scientific cognition in terms of using DM methods. The issue regarding the contribution of DM methods to the methodology of scientific cognition was raised, and the level of cognitive value of the results of their use was determined.

The scheme illustrating the relationship between the methodology of the levels of scientific cognition, which supplements the well-known schemes of their classification and demonstrates the maximum capabilities of DM methods, was developed. In terms of the methodology of scientific cognition, a crucial fact was established - the limit of applicability of any DM method is the lowest, the first level of the methodology of scientific cognition – the level of techniques. The result of the processing in the form of ER can serve as a basis for these techniques.

Keywords

Data Mining, data, scientific cognition, methodology, empirical regularity, hypothesis.

1. Introduction

The enhanced opportunities of the existing cognitive tools and a search for new tools have always aroused a great interest, owing to their crucial importance for the development of human civilization, because knowledge gained as a result of the use of these tools is the primary means of transforming the reality.

In recent decades, Data Mining (DM) methods and tools have become widely used (Data Mining — it is not a single method, but a variety of a large number of different methods for identification of regularities. In the English-speaking world, they commonly use the term “Machine Learning”, denoting all Data Mining technologies.). This happened in response to the practical needs in different sectors of the national economy, as well as in the context of evolving capacities of computers, which enabled to accumulate and process large amounts of heterogeneous data.

ISIT 2021: II International Scientific and Practical Conference «Intellectual Systems and Information Technologies», September 13–19, 2021, Odesa, Ukraine

EMAIL: nikk.gena@gmail.com (A. 1); shumeiko_a@ukr.net (A. 2); bvs910@gmail.com (A. 3)
ORCID: 0000-0003-3984-9266 (A. 1); 0000-0002-8170-9606 (A. 2); 0000-0003-1269-7207 (A. 3)

© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3126/paper2.pdf

2. Main result

DM algorithms, implemented as computer programs, have actually developed new research tools. At the same time, a widespread use of DM methods raises methodological questions whether we have a correct understanding of their capabilities and limits as well as data processing results in terms of scientific cognition. At first glance, it seems an abstract question, but its clarification will enable the concerned parties to achieve better results and organize more effective business processes.

It should be noted that, to varying degrees, the attention has already been paid to the image recognition methodology, as DM methods were formerly called, by such internationally acclaimed scientists as [1-7]. However, these scientists have not conducted an analysis in terms of the theory of cognition.

In fact, almost all the time, most studies on DM methods raise the question which is rather related to the methodology of cognition[2]: "What knowledge can be derived from the accomulated data and what is its level?" This question demonstrates the immaturity of our concept of DM in terms of the theory of cognition, and it also summarizes multiple practical problems of DM application, which are not addressed by enchancing the computing capabilities or parallel computing in the field of Big Data processing [6]. Besides the difficulties of the right choice and application of DM methods to the addressed problems, there is no full understanding of its capabilities and limits for the application as well as of the process (phasing) itself and the obtained results in terms of the theory of cognition. At the same time, an understanding of the capabilities and limits of DM can lead to a significant modification of the methodology for the study and for addressing the practical problems as well as improving the efficiency of applying the methods under consideration.

The practice of analytics shows that DM methods are indeed a powerful tool of scientific cognition, which is of multidisciplinary nature. Moreover, it is DM methods that can serve as a basis for the convergence of the approaches to scientific cognition in the humanities as well as in natural sciences. Based on DM, a huge number of the applied problems is addressed, and the data mining algorithms are improved. However, in terms of the methodology, very little effort is made and almost no researches are carried out in this field, which substantially hinders further development of DM that, generally speaking, could become a basis for disciplinary revolution in the theory of cognition, and could even enable to generate major innovations in the field of intelligent technologies.

The aim of the study: to specify the capabilities and limits of applying DM methods in terms of the methodology of scientific cognition.

The process of cognition is a process of gaining and using knowledge, which is of staged nature [8]. The first stage of cognition – singling out and statement of the problem, then – experience, observation, experiment, studying the phenomenon: the second stage - 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 abstractions found, i.e., hypotheses or conclusions that were made before, are being tested. This is a universal scheme of cognition (Fig.1).

These issues became particularly pronounced when computers started to be used for data mining. The key issue, being critical in terms of cognition, is what the use of DM introduced into the methodology of scientific cognition and what the application of its outcomes can result in?

The application of DM tools starts only when the data has already been prepared in the form of datasets, where the objects are represented by the sets of multidimensional data – for example, in the form of training dataset (TD). It is generally acknowledged that all DM methods are based on the inductive method of cognition, i. e., in case of DM (inductive learning), the program learns based on the presented empirical data. In other words, the program builds some kind of a general rule based on the presented empirical data, which is obtained, in particular, through observation or experiment[3]. When using any DM methods, the final outcome is represented in the form of one or another model that reflects certain regularities intrinsic to the data under study, which might logically be called empirical regularities (ER) and which, probably, are hypotheses in nature (that was very cautiously assumed by Zakrevsky [4].

[2] Although, most often, it is raised in purely practical terms– how far we can trust the knowledge we gain.

[3] The matters of choosing the feature vector and data pre-processing are beyond the competence of DM.

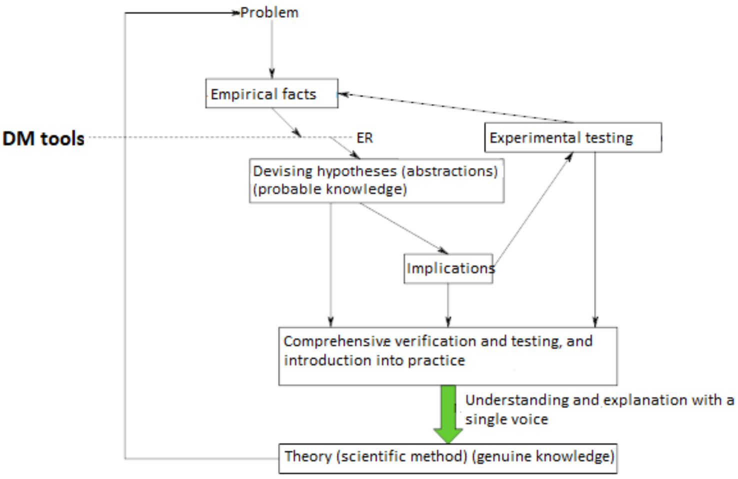
Problem
Empirical facts
DM tools
ER
Experimental testing
Devising hypotheses (abstractions)
(probable knowledge)
Implications
Comprehensive verification and testing, and
introduction into practice
Understanding and explanation with a
single voice
Theory (scientific method) (genuine knowledge)

Figure 1: General Scheme of scientific cognition (using DM methods)

Therefore, the major outcome of applying DM methods is ER in the subject area under study, obtained with the use of these methods, which can be represented in different forms and types. These ER are, in fact, “drafts”, a critical auxiliary material for preparation and development of dialectical “leap” or complicated transition from the empirical level of cognition to the theoretical one through devising hypotheses are the driver of science (Fig.1). In order to clarify the issue of the level of knowledge derived in terms of the theory of scientific cognition when analyzing the data accumulated in a certain subject area, we cannot do it without the methodology of scientific cognition that “studies the methods for building the scientific knowledge and methods which are used to gain new knowledge, i.e., methods and forms of scientific study, dealing with the technical aspect to a minimum extent” [9]. It is customary to distinguish the following levels of the methodology of scientific cognition [9]:

1. Technique – the lowest level, the examples – directions, techniques, etc.;
2. Scientific method, relying on knowledge of the respective regularities, i.e. the theory of the given subject area;
3. General scientific method – 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 methods of scientific cognition, and their study is the subject of philosophical methodology (philosophy of science).

In view of the foregoing, it is proposed to supplement the above classification of the levels of the **methodology of scientific cognition** in the form of the list of items 1-4, suggested by V. Shtoff, with the scheme presented in Fig.2 – some kind of graphical supplement to these items, illustrating the outcomes of the work in a specific subject area of the inductive approach under study, which is a basis of all DM methods, related to the levels of scientific cognition.

The main purpose of this scheme is to show the relationship between the levels of cognition, and, the most important thing, to demonstrate the limit of the capabilities of DM methods. It follows from the above statement and the illustration that the limit of the level of the scientific cognition methodology, achieved through DM methods or tools, is the lowest of these levels – the level of techniques.

As a result, ER is quite understood by the expert in the subject area and is applicable for further processing as a basis for possible transition to the hypothesis, which is not the automated result of induction and not an inductive inference, but one of the possible answers to the problem encountered, including in the form of assumptions, suggestions and their implications with further testing in practice. However, the emergence of hypothesis is mandatory[4].

[4] The need for hypothesis stems from the fact that the laws are not directly seen in individual facts, no matter how many of them are accumulated, as the essence does not coincide with phenomena. Hypothesis is the statement, the truth or falsity of which has not yet been established. The process of establishing the truth or falsity of the hypothesis is the process of cognition as a dialectic unity of practical (experimental, object-tool) and theoretical activity. However, eventually it is only confirmation by practice that converts a hypothesis into the true theory, converts probable knowledge into the credible one, and vice versa, the refutation in practice and experiment discards the hypothesis as false assumption [9].

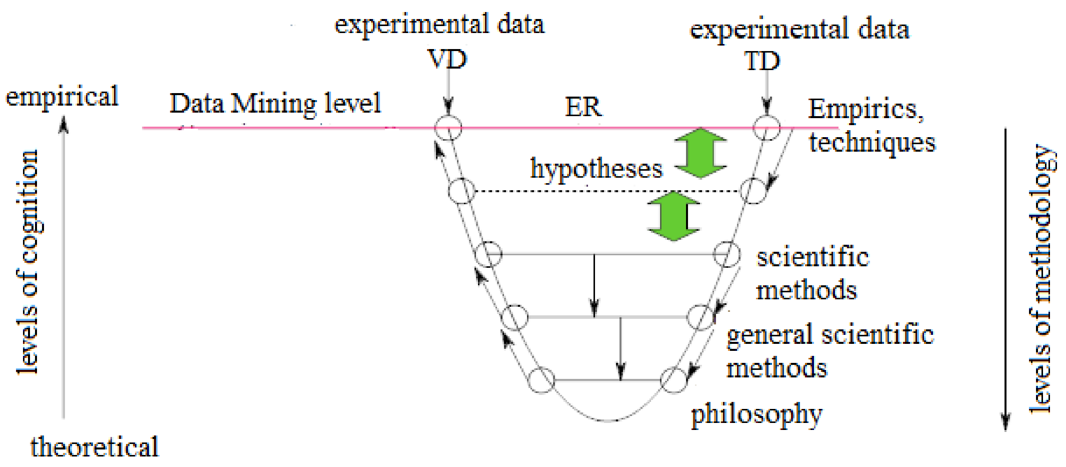

Figure 2: Relationship between the levels of cognition

Abbreviations: ER – empirical regularities. TD – training dataset. VD – validation dataset

Using DM, it becomes possible to automatically generate ER, being the "bricks" for advancing and building hypotheses as a part of addressing a specific problem. That is, the emergence of hypothesis is preceded by a very important stage of generation (search) of ER - 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 and downloads the data).

At the same time, possible transition from ER to hypothesis as a probable knowledge – is not so easy and straightforward way. There is an intersection or convergence of dialectical logic, methodology of scientific cognition and psychology of scientific creativity (Fig.3). The analysis of the structure of such a complex dialectic intersection is one of the challenges in the way of transition from the empirical basis to the theoretical building [9].

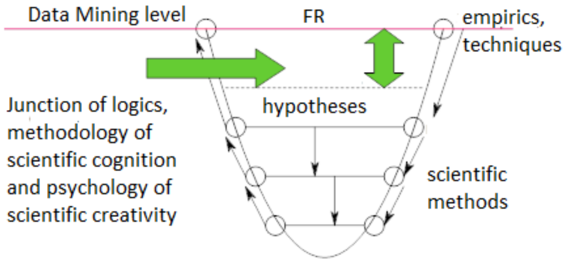

Figure 3: Transition from ER to hypothesis

This also requires performing considerable and nontrivial intellectual work, taking certain efforts by the researcher and, most probably, carrying out additional researches, which, to a large degree, can be considered an extension of DM. This is the case with almost all known DM methods. Therefore, the ultimate outcome that might be obtained directly in the application of any DM tools is ER level, and, methodologically speaking, the level of techniques. Such class of DM models as neural networks needs to be separately mentioned. The use of neural networks, in some cases, yields rather good results; however, unfortunately, they produce no effect in terms of the methodology of scientific cognition – we cannot build ER in this case and, even more, we are unable to proceed to formulate and devise hypotheses! Their level is limited by the level of "primitive" (like animals do it) recognition (classification) and nothing more, and it is not itself a new knowledge. From the cognitive and methodological points of view, it is a dead-end type of DM or a completely different paradigm of the scientific cognition. Actually, this is also discussed in the work [10] where the authors try to "feel out" the ways of understanding the work of neural networks.

It should be noted that it is advancement of ER that the cytogramm processing web service (URL: https://www.data4logic.net/ru/Services/CellsAttributes) is focused on, enabling cytologists-researchers to generate ER and, with a high probability of success, to devise on their basis the hypotheses to address the problems that they face. The pictures stipulated by the paper related to leukemia diagnostics [11, 12] can be used as an example of this approach.

In many cases, solving specific practical problems is actually limited, in terms of cognition, to the level of ER, which is used as a basis for further formulation, in a best-case scenario, of a decision-making direction or rule, and it remains at the first empirical level of cognition, being the lowest of all possible levels [13, 14, 15]. In the short run, it suits business as a sphere of practical activities; however, in the long run, the main think is lost – finding really new knowledge which can be implemented in innovations, or developing a new method, modus operandi, business model, etc., that will provide higher-order competitive advantage.

In a similar way, the level of "primitive" classification inherent to neural networks often suits 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 Fig.1 and Fig.2.

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 is a new man-machine methodology of empirical cognition.
2. These methods have their limit in the form of ER represented in different forms.
3. ER can serve as "drafts" for preparation, generation and formulation of hypotheses aimed at further more in-depth cognition of the subject area.
4. In order to select the best strategy for the use of DM tools, a clear understanding of the goals of problem-solving is needed.
5. The use of DM tools requires a close cooperation with the experts in a specific subject area that, in its turn, raises a number of questions related to: 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 cognition; there is still a limit represented by the 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. The full-fledge interaction between the experts in subject areas and Data Scientist is significantly more painstaking in terms of organizational and communicative cost, but, in our opinion, this approach is able to ensure major 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 was inefficient. From a methodological standpoint, the most critical fact has been established – the limits of the applicability of any DM methods are the level of ER, i.e. the level of techniques and directions in a specific subject area, where data mining methods are used, or provisional (working) hypothesis. As of today, it is the only visible and obvious achievement of all DM algorithms. It should be noted that one of the available web services, suitable for researchers who have no special training on mathematics and informatics, which is designed to find ER, is implemented on ScienceHunter portal (https://www.sciencehunter.net).

3. Conclusions

Knowing the applicability limits of DM tools, it is possible to more fully understand how to set goals when selecting appropriate DM methods; for example, to choose ones that produce a relatively large set of ER, or to use those ones that produce a limited set of such patterns characterized by greater accuracy. From the methodological point of view, the most important fact has been established – the limits of applicability of DM methods is the level of ER. A huge number of methods, techniques, a variety of developed computer programs, cloud services and other software – all this ends up with one thing that is the level of ER. Currently, this is the only observable and obvious achievement of all DM algorithms. Should the result be considered important in terms of cognition? It is quite possible to answer positively. Although it should be emphasized that all this refers to a particular subject area, which applies methods of data mining. It should be noted that DM can be understood as an evidentiary or constructive method of cognition, with all the advantages and disadvantages. Finding

ER today is implemented in the form of web services (for example, ScienceHunter portal: https://www.sciencehunter.net), so future research will focus on the development of an automated system concept for DM, suitable for researchers with no special training in mathematics and computer science.

4. References

[1] M.M. Bongard, Recognition problem, Nauka, Moscow, 1967.

[2] N.G. Zagoruiko, Recognition methods and their application, Soviet radio, Moscow, 1972.

[3] N.G. Zagoruiko, Applied methods of data and knowledge analysis, IM SO RAN, Novosibirsk, 1999.

[4] A.D. Zakrevsky Recognition logic. Minsk: Nauka i tekhnika, 1988, 118 p.

[5] L.G. Malinovsky, Classification processes - the basis for constructing the sciences of reality, Algorithms for processing experimental data (1986) 155-182.

[6] A. Carbon, M. Jensen, A.-H. Sato, Challenges in data science: a complex systems perspective, Chaos, Solitons & Fractals 90 (2016), 1-7. doi:10.1016/j.chaos.2016.04.020

[7] L. Cao, Data Science: Challenges and Directions, Communications of the ACM, 60(8) (2017) 59-68. doi:10.1145/3015456

[8] N.N. Moiseev, Man, environment, society. Problems of formalized description, Nauka, Moscow, 1982.

[9] V.A. Shtoff, Problems of the methodology of scientific knowledge, Vysshaia shkola, Moscow, 1978.

[10] Z. Chen, Y. Bei, C. Rudin, Concept Whitening for Interpretable Image Recognition, Nature Machine Intelligence, 2 (2020) 772-782. doi:10.1038/s42256-020-00265-z

[11] D.F. Gluzman (Ed.), Diagnosis of leukemia. Atlas and Practical Guide, MORION, 2000.

[12] V.A. Lekakh, Sick issues of modern oncology and new approaches to the treatment of oncological diseases, Librokom, Moscow, 2011.

[13] W. Chen, H. R. Pourghasemi, S. Zhang, J. Wang, 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, Elsevier, 2019, pp. 467-484). doi:10.1016/B978-0-12-815226-3.00021-1

[14] K. Gibert, J. Izquierdo, M. SànchezMarrè, S.H. Hamilton, I. Rodríguez-Roda, G. Holmes, Which method to use? An assessment of data mining methods in Environmental Data Science, Environmental Modelling & Software 110 (2018) 3-27. doi:10.1016/j.envsoft.2018.09.021

[15] G. Agapito, P. Guzzi, M. Cannataro, Parallel and Distributed Association Rule Mining in Life Science: a Novel Parallel Algorithm to Mine Genomics Data, Information Sciences 26.07 (2018). doi:10.1016/j.ins.2018.07.055