Models of Explanation in Intelligent Data Analysis

Michael Zabezhailo [1]

[1] *Federal Research Center "Informatics and Control" of the RAS, 40 Vavilova str., 119333, Moscow, Russia*

Abstract

There are discussed some variants to explain results of Big Data intelligent analysis made by computer AI-systems. So-called partial (open) theories formed basing on consequentially extended collections of empirical data are presented as a platform for knowledge representation and a tool to explain results calculated in the process of intelligent data analysis (IDA). Some characteristic features of the presented approach are discussed, especially – automated analysis of causal dependencies in empirical data. Computational complexity of combinatorial problems related to the generation of open empirical theories is analyzed.

Keywords

Artificial Intelligence (*AI*), intelligent data analysis (*IDA*), Big Data, explanation, causality (*cause-and-effect* relationships), empirical theories

1. Introduction

Last decade *trustability* became one of the critically important features of some classes of modern AI-solutions. From the classical viewpoint the concept of *trustability* (tractable as *acceptability*, *reliability*, …) was based on statistical evaluation [1]. But now explanation of results, conclusions and recommendations generated by means of AI-systems is posed in focus of different *trustability* forming approaches and procedural mechanisms.

Basing on historical experience it's useful and effective inspecting trustability in AI-systems to take into consideration a number of significant effects, including different logical "tricks". Here to form "illusion" of explanation some non-correct arguments and explanation-like schemes may be demonstrated. For example, a well-known demagogic technique is to conclude at the same time from both true and false premises (antecedent), where only true premises are presented to the demand to provide explanations to the declared conclusions (but false ones are hidden). However, it is clear that the presence of false premises (in antecedent) allows you to deduce any (including false) conclusions. In the last couple of years, "explanatory" schemes of the *highly likely* class have become widely known [2, etc.] (but the applicability of them, even in the classical legal practice of the country of their origin, apparently does not have any procedural prospects). Interesting examples of the explanatory construction formation can be found in court proceedings (see the procedure of construction and comparison of versions of what happened that are formed by the defense and prosecution sides), as well as in the so-called legendization of special operations [3,4, etc.].

In turn, the history of the scientific research and development has given a number of interesting results in the field of cognition under discussion. So well-known today are the explanatory constructions of a *deductive* type proposed by K.G.Hempel[2]. No less studied are the so-called *abductive* schemes of explanation by C.S.Pierce[3], significantly developed and detailed over the one and a half centuries that

Russian Advances in Fuzzy Systems and Soft Computing: Selected Contributions to the 10th International Conference «Integrated Models and Soft Computing in Artificial Intelligence» (IMSC-2021), May 17–20, 2021, Kolomna, Russian Federation
EMAIL: m.zabezhailo@yandex.ru
ORCID: 0000-0002-5067-5909

[2] C.G.Hempel, Studies in the Logic of Confirmation, Mind, V.54 (1945), Part I: Pp. 1-26; Part II: pp. 97-121.
C.G.Hempel, Oppenheim P., Studies in the logic of explanation, Philos. of Science, 1948, 15.
C.G.Hempel, Philos. of Natur. Sciences, Englewood Cliffs, N. J.: Prentice-Hall, 1966, 116 p.

[3] C.S.Peirce, On the natural classification of arguments, Proc. of the American Academy of Arts and Sciences, 1867, 7, pp.261-287.
C.S.Peirce. Writings, URL: www.peirce.org/writings.html
Writings of Charles S. Peirce, (Kloesel C.J. et al. - eds.), V.1, 1857-66. V.2, 1867-71. V.3, 1872-78, Bloomington: Indiana Un. Pr., 1982-86.

have passed since their first publication[4]. However, using the Peirce abduction, special attention has to be paid to the problem of the *constructiveness* of the procedures for the formation of *abductive* explanations in AI systems. Thus, the experience of the *Big Mechanism* [5] and *Explainable AI* [6] research projects organized by DARPA has demonstrated how non-trivial and at the same time critically important to determine the reliability of the entire AI system as a whole can be the problem of explanation (especially - in critical applications). The problem of finding such *constructive* procedures for generating *abductive* explanations can become a *bottle neck* preventing purposefully and effectively moving from the initial data to the appropriate conclusions (added by their *abductive* explanations).

2. Explanation in modern data analysis systems

Today, the development and use of procedural explanation schemes in AI-systems has a history spanning more than half a century. One of the first generally recognized circumstances here was the understanding of the fundamental difference between the answers to the questions *HOW (*a particular conclusion has been formed*)*? and *WHY* (this conclusion has been formed*)*?. Back in the 60-70 years of the last century, it was understood that the track of the solution “inference” in the so-called rule-based expert systems[5] does not guarantee the interpretability of the expert conclusion formed in this way. In general case, to explain results, it is necessary to ask experts (both in the analyzed subject area, and in the architectural specifics of the used system of rules and procedures of their combining).

The history of attempts repeatedly made in the last couple of decades to develop a reliable explanation mechanism for the results of so-called *deep* machine *learning* is also informative and instructive. The practice of studying this problem (see, for example, the *Explainable AI* (XAI) project [6] of the DARPA agency, etc.) has demonstrated a set of fundamental barriers that have slowed down to almost zero the effective progress towards the goal outlined here – the development of reliable "tools" for explaining the results of *deep learning*. (It became clear that in general case by “instruments” of the type *HOW*?, it is unlikely that the results generated by one "black box" can be reliably explained using another "black box").

Another fundamentally important step was the identification of well-defined "side" empirical effects - artifacts of computer data analysis, and, first of all, formed in the process of machine learning *artifacts*, which are called *overfitting* (i.e. generation of formally correct conclusions that are formed without any errors in calculations, but characterizing by unacceptably low generalizing ability[6]). As it was shown (see [7-8]), the direct question "*Is it possible to reduce the generation and influence of such artifacts to zero*?", unfortunately, does not have a positive answer. Moreover this circumstance can be given a completely understandable interpretation: in general case, the *semantics* of a phenomenon studied with the help of an AI computer system cannot be accurately (exactly - that is, in a *one- to-one* way) represented by purely *syntactic* means\”instruments” of computer modeling.

In recent years, analyzing the lessons of previous AI research and development experience, new requirements for the "functional capabilities" of the explanation subsystem that reflect the actual problems and challenges of the current moment have become increasingly obvious. The most significant ones here, apparently, should include:

- "*transparency*" of recommendations and conclusions formed in the process of intelligent data analysis (IDA) for Decision Makers. Such "transparency" (*meaningful interpretability* and *explainability*) of the formed conclusions can be achieved, in particular, due to the expert's full understanding of the data analysis method implemented in the IDA used (so-called "seamless" integration of the expert's reasoning method and the AI computer system reasoning scheme),
- *stability* (*heritability*) of the formed conclusions and explanations with variations in the details of the object under analysis knowledge representation. (In particular, it can be stability in the process of finding a balance of *details ~ computational complexity* of the corresponding conclusion generation. An example of such balancing can be found in the so-called QSAR problems – analysis of relationship between the CHEMICAL STRUCTURE and PHYSIOLOGICAL ACTIVITY of

[4] R.W.Burch, A Piercean Reduction Thesis: the Foundations of Topological Logic, Lubbock, TX: Texas Tech Univ. Press, 1991, 152 p.
Abductive Inference: Computation, Philosophy, Technology (Josephson J.R. et al. - Eds.), Cambridge: Cambr. Un. Press, 1994, 320 p.
V.K.Finn, J.S. Mill’s inductive methods in AI systems, Sci. Tech. Inf. Process., P.I:2011,V.38, pp.385–402, P.II:2012, V.39, pp.241–260.

[5] E.V.Popov, Expert systems, M: Nauka, 1987, 288 p.

[6] The possibility to organize correct extrapolation of such empirical dependencies to newly analyzed cases (objects, situations, ...).

chemical COMPOUNDS. Here to describe the structures of physiologically active compounds, you can use both graph representations – for example, in the form of graphs7 themselves or coverings with sets of "typical" subgraphs8, and with descriptions of radiation and/or absorption spectra9),

- taking into account the *managerial "consequences"* of decisions made based on the results of the IDA: the *stability* (*heritability*) of conclusions and explanations when receiving new data (clarifying the description of the object of research). Such stability can be achieved, in particular, by implementing an IDA analysis of *causality* (*cause-and-effect* relationships). Well-known examples of subject areas where this approach works successfully are medical and technical diagnostics, fraud protection in banking and finance, information security (identification and countering computer attacks,...), etc.,
- ensuring the *sufficiency of the grounds* for *accepting the results* of the IDA in specific current conditions, and even in situations of incompleteness of the currently available description of the objects under analysis (see for example small, statistically insignificant – non-representative - training samples, etc.). An interesting example of this type of characteristics is given by the actual requirements for the so-called *third wave* of AI research and development for the "reliability" of empirical theories, including at the early stages of their formation (see, for example, [9], etc.).

3. The problem of explanation and open empirical theories

Historically, the most common methodological basis to assess the acceptability of the computer calculation results (including IDA) has become the reliability paradigm that came from statistical data analysis (as well as its applications in technology), based on the concepts of the *general population*, samples from it and the *representativeness* of such samples. Actually, in the applications of AI technologies and systems, this approach is the basic component of the so-called *second wave* of AI research and development ([9], etc.), associated with *statistical learning* systems (widely understood – from Bayesian inductive inference to deep machine learning by means of artificial neural networks).

However, in general case this approach is no adequate in the situation with Big Data analysis. Here we have to deal not only with the *Big* type effect (i.e. large volumes of analyzed data), but also with the *Open* type effect (i.e. the openness of the arrays of processed data, related to the possibilities to extend collected data by new information, and not necessarily the same as the previously known "nature"). It is easy to see that in situations of the *Open* type, the traditional ideas about ways to assess the reliability of IDA results and conclusions (based on statistical models – see above) are generally not adequate. In consequently expanding (and, in general, in an unpredictable way) collections of big data there may simply be no reason to talk about the concepts of the *general population* and a *representative* sample from it.

One of the resultative alternative ways to work with open collections of empirical data has become[10] the study of *open empirical theories*. Here we propose a problem-oriented mathematical technique to *represent knowledge* (about the object of research) in the form of a *partial theory* (describing the accumulated empirical information – the current state of the **F**act **B**ase). Partial empirical theory is formed by a consistent set of formulas **T**, for which each fact from the current **FB** can be represented as a *logical consequence* of a subset of formulas from **T**. At the same time, each extension of the current **FB** with new facts implies a corresponding modification of the current version of the theory **T**. The problem to allocate *stable\heritable* (with respect to the consequent extensions of the current **FB** formed by new “potions” of empirical data) *fragments* of the corresponding modifications of the theory **T** can be successfully solved by means of *cause-effect* dependencies generation. To extract this type empirical dependencies (hidden in each current **FB**) from sequences of expanding Databases of **F**acts (allocating "inherited" fragments in sequences of corresponding refinements of the theory **T**) allows us to successfully form corresponding “initial” theory **T** and its consequent modifications. For a number of applications (see for example medical and technical diagnostics, fraud protection in banking and finance, countering computer attacks of certain classes, etc – - see, for example, [10-13, etc.]) this type approach demonstrates its power and effectiveness. It is the analysis

[7] A method of description that generates a number of provably hard combinatorial problems.

[8] A simpler (than *graph*-theoretic description) *set*-theoretic version of the representation of knowledge about the analyzed chemical structures.

[9] Also, a *set*-theoretic version of knowledge representation for the objects under study.

[10] See for example [9] and more earlier ones publications of V.K.Finn:
V.K.Finn, J.S. Mill’s inductive methods in AI systems, Sci. Tech. Inf. Process., P.I:2011, V.38, pp.385–402, P.II: 2012, V.39, pp.241–260.
V.K.Finn, About intelligent data analysis, AI News, 2014, №3, p. 3-18.
V.K.Finn, The Synthesis of Cognitive Procedures and the Problem of Induction, Autom. Doc. and Math. Ling., 2009, V.43, №3, pp.149-195.

of causality, which underlies the procedural construction of the corresponding empirical theories formation that became the basis for the effective identification of stable\heritable sets of empirical dependencies (fragments of the generated empirical theories of a *causal* nature) that are inherited when the initial **FB** is expanded. The identification of such *causal factors of influence* can be organized, in particular, by analyzing the *similarities* of the precedent descriptions – "positive" examples - where the presence of the studied effect is identified. In addition it's necessary to check the "non-embeddability" of the effective (identified as similarities of the descriptions of examples) combinations of such factors in the descriptions of counterexamples (precedents of the absence of the studied effect) – see the quoted works of V.K.Finn, etc..

The stability of the causal grounds that are "forcing" the appearance of the investigated effect with respect to the expansion of the studied phenomenon descriptions with more and more complete data describing their properties and characteristics in sensitive applications (see, in particular, the diagnostic tasks mentioned above) allows us to form *inherited control actions*. For example let we see medical therapeutic measures, targeted actions to support the "survivability" of complex technical systems, measures to detect and suppress financial fraud, effective identification and countering certain classes of computer attacks, etc.. In each of these applications, empirical dependencies of a *causal* nature, generated in the process of forming an empirical theory to describe the already accumulated facts, allow us to explain the results of IDA performed by a computer system (answering not only the question *HOW*?, but also the question *WHY*?).

4. Some combinatorial properties of empirical theories

Of course, in terms of procedural basis – mathematical models, methods and algorithms - the implementation of the described scheme (i.e. the formation of partial theories from consequently extended collections of empirical data and the generation of explanations for accumulated facts by causal empirical dependencies forming such theories) has a number of specific "technical" features. Problems with assessing the reliability of empirical dependencies generated during such an IDA by traditional means (see above comments on the concept of the general population in open subject areas) are supplemented by the need to analyze a set of new effects, as well as to offer acceptable solutions, including for the following tasks:

- to evaluate of the (causal) *representativeness* of training samples (current **FB**),
- to estimate of the *complexity* (i.e. *capacity* - number of elements) of the set of empirical dependencies (*EmpD*) that interpolate the training sample **FB** (i.e. estimation of the "size\volume" of the formed empirical theory-**EmpT**),
- to estimate of the *capacity* of the **EmpD -** the set of all *EmpD* inherited during the transition to a given extension of the current **FB**,
- to check the *non-emptiness* of the set **EmpD** of all *EmpD* inherited during the transition to the extension of the current **FB** by specific new data (checking the non-emptiness of the fragments of the **EmpT** that are stable with respect to the expansion of the current **FB** by specific new data) ,
- etc..

The analysis of the presented problems solvability led to the following two types of conclusions.

(1) The inefficiency of the use of "brute force" methods (i.e. an exhaustive complete search of variants) in the considered area is shown. Thus, in particular, demonstrated ([10-13], etc.):

- in general case *exponentially fast* growing complexity (capacity sizes) of the current **EmpD** (including versions of **EmpD**, which are completely changing its internal structure while extending current **FB** by new facts)
- the existence of **EmpD** that are growing *exponentially* in the size and are *extrapolated* to the specified new object, but *not inherited* when you extend the current **FB** by some new facts.

(2) The effective solvability of the basic combinatorial problems describing computational complexity of the discussed type **EmpT** is demonstrated. Thus, in particular, there is shown ([10-13], etc.) the polynomial solvability of following problems:

- the (causal) *representativeness checking* of the current training sample (current **FB**),
- the *non-emptiness checking* of inherited at a given **FB**-extension fragment of **EmpT**,
- the *non-emptiness checking* of inherited at a given **FB**-extension fragment of **EmpT** providing *extrapolation* of (at least some) collected in it *EmpD* for a given newly "diagnosed" *precedent*.

5. Conclusions

The effective solvability of the above problems on the properties of **EmpT** can serve as a basis for (fast - !) formation of *approximate* solutions – "*soft*" *versions* of such theories, which are resultative fragments of the corresponding **EmpT** (allowing you to quickly "diagnose" given new objects when solving specific applied problems). At the same time, it is possible to effectively generate only "useful" fragments of the empirical theory explaining the accumulated facts in this particular case, providing for the Decision Maker full "transparency" of conclusions and recommendations for making control and management decisions formed in the process of IDA.

Assessing the trends and perspectives for further development of this direction of AI research and development, it seems that it is necessary first of all to pay attention to the problem of intellectualization of control "instruments" for so-called *big systems* management. The necessity

- to operate with *large amounts* of information (Big Data),
- to perform appropriate data analysis in the (*process*-) *real*-time mode and
- ensuring the *explainability* (as well as the *predictability* of the consequences) of management decisions made based on the results of such intelligent data analysis, the *responsibility* for which lies (and will lie - !) on the Decision Maker,

is probably the most critical arguments in favor of the relevance of the proposed choice.

6. References

[1] Dependability in technics. Terms and definitions - GOST 27.002-2015.Moscow: Standartinform, 2015, 28 p. URL: https://files.stroyinf.ru/Data2/1/4293754/4293754027.pdf

[2] B. Maksimov, Highly likely: how British English confuses foreigners. BBC News (Russian service). – April 20, 2018. - https://www.bbc.com/russian/features-43804414

[3] Information leaks and internal threats. Legendization. Infowatch (Livejournal). – April 11, 2013. URL: https://infowatch.livejournal.com/394085.html

[4] A.D. Rudchenko, A.V.Yurchenko, Business intelligence analyst. Magister program, Moscow: Inst. of secur. probl., HSE, 2019. URL: https://www.hse.ru/ma/intelligence/courses/296805596.html

[5] P.R.Cohen, DARPA's Big Mechanism Program, Physical Biology, 2015, V.12, №4, pp.1-9. URL: https://iopscience.iop.org/article/10.1088/1478-3975/12/4/045008

[6] IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI), URL: http://home.earthlink.net/~dwaha/research/meetings/ijcai17-xai/

[7] K.V,Vorontsov, Combinatorial theory of learning by precedents, D.Sc. Thesis (Theor. Comp. Sci.), M: CC RAS, 2010, 273 p. URL: https://www.dissercat.com/content/kombinatornaya-teoriya-nadezhnosti-obucheniya-po-pretsedentam

[8] D.V.Vinogradov, Probabilistic-combinatorial formal method of learning based on lattice theory, D.Sc. Thesis (05.13.17 – Theoretical computer science), M.: FRCCSC RAS, 2018, 131 p. URL: http://www.frccsc.ru/diss-council/00207305/diss/list/vinogradov_dv

[9] DARPA Sets Up Fast Track for Third Wave AI, Jul 26, 2018, URL: https://defence.pk/pdf/threads/darpa-sets-up-fast-track-for-third-wave-ai.569563/

[10] M. I. Zabezhailo, A. A. Grusho, N. A. Grusho, E. E. Timonina, Support for solving diagnostic problems, Systems and Means of Informatics, 2021, Vol. 31, Issue 1, pp 69-81

[11] M.I.Zabezhailo, Yu.Yu.Trunin, On the Problem of Medical Diagnostic Evidence: Intelligent Analysis of Empirical Data on Patients in Samples of Limited Size, Automatic Documentation and Mathematical Linguistics, 2019,Vol. 53, No. 6, pp. 322–328

[12] M.I.Zabezhailo, To the Computational Complexity of Diagnostic Predictions Designed by Means of Characteristic Functions, Automatic Documentation and Mathematical Linguistics, 2020, V54, №6, pp. 298-305

[13] M.I.Zabezhailo, On the complexity of characteristic function sets for correct diagnostic problem solving, Artificial Intelligence and Decision Making, 2021, № 2, pp. 44-54