Intelligent Data Analysis: from Theory to Practice

Tatur M.M.
Belarusian State University
of Informatics and Radioelectronics
Minsk, Republic of Belarus
tatur@bsuir.by

Iskra N.A.
Belarusian State University
of Informatics and Radioelectronics
Minsk, Republic of Belarus
niskra@bsuir.by

***Abstract*—In this paper common issues in practical intelligent data analysis are addressed. Recommendations for productive practical applications and development of intelligent data analysis systems are proposed.**

***Keywords*—data analysis, structured data, knowledge acquisition, statistical analysis, machine learning, data visualization**

I. Introduction

The term "data" in the common sense is a very broad concept, and "Intelligent Data Analysis", if desired, can be used almost in any area of information processing. Therefore we initially want to concretize (restrict) the concept of "data", "data analysis", and further "intelligent data analysis".

In this paper by "data" we refer to the ***structured data***, i.e. vector of measured parameters describing an object, phenomenon or event. For example:

- While paying at the store you "leave" data about the time of your visit, the purchased goods with the corresponding amounts. In case you use a buyer's card, this transaction is personalized (i.e. it can contain additional information such as age, sex, place of residence, birthday, etc.). Obviously, this information can potentially be used to plan and carry out marketing programs to retain clients, promotion-actions, etc.
- Each call to emergency services (911) is recorded and initiates data collection: the time, place, cause of the incident and actions to be taken, as well as the results of these actions or consequences. Such data are accumulating constantly and can be used to generalize and form useful conclusions for practice.

Under the definition above do not fall unstructured and weakly structured data, for example, signals, images, video data, texts, etc., that require an additional stage of preprocessing.

The simplest case of structured data processing (apart from counting, extracting, copying, updating, etc.) is ***statistical analysis*** [1]. For example, it is easy to calculate how many visitors to the supermarket made purchases in a certain period of time, what is the average purchase price, how much did the revenue increase or decrease in the analyzed periods, etc. Such statistical analysis requests are formed by the user and are aimed at covering the current issue (problem). Such analysis, as a rule, operates with one of the parameters, while other parameters are fixed or predefined.

However, very often in order to obtain more complex information about an object, event, or phenomenon, it is necessary to carry out a deeper analysis on data, taking into account many parameters, and in the worst case - all parameters available. Thus, in each industry there are historically gained and constantly accumulated specific knowledge, that are associated with a deep data analysis. This knowledge is represented by complex characteristics that link a number of measured parameters, as well as laws and regularities related to these parameters.

For example, in economics, the efficiency of the enterprise is characterized by the enterprise's profit index for a certain period of time. This index combines information on incomes and expenditures. Incomes and expenditures in their turn are also derivative characteristics that are calculated (extracted) from a whole series of primary data. It is even more difficult to calculate such a complex index as a labor productivity. Economic laws of breakevenness of the enterprise, a stock of circulating assets, etc. are well-known. Similar complex indexes and indicators can be found in medicine, sociology, criminology and other fields [2]. This kind of knowledge is undoubtedly the result of the intelligent data analysis carried out by specialists in particular industry, relying on previous experience, empirical observations, hypotheses formation and testing, heuristics, and the like. In this context, "intelligent data analysis" is a special case of "data analysis", and there is hardly a clear boundary between trivial (simplest) and intelligent analysis, at least this issue can be classified as rhetorical.

Note: By this we don't mean knowledge structuring, formal knowledge representation or formal knowledge management, which are usually referred to as ***artificial intelligence***.

In Fig. 1 a mnemonic diagram of data analysis is given. It can be divided into a trivial and an intelligent analysis. The meaning of the picture is that the user generates a database query (request) to obtain information (and, ultimately, knowledge) about an object or phenomenon. Based on this query, the data is analyzed: trivially and / or intelligently. If to obtain knowledge of an object or a phenomenon (for example, about the production efficiency), it will be sufficient to formally apply already acquired knowledge (for example, the formula for calculating the labor productivity, without any additional interpretation and explanation), this analysis hardly could be attributed to the intelligent type. In Fig. 1

this circumstance is marked by a dotted arrow.

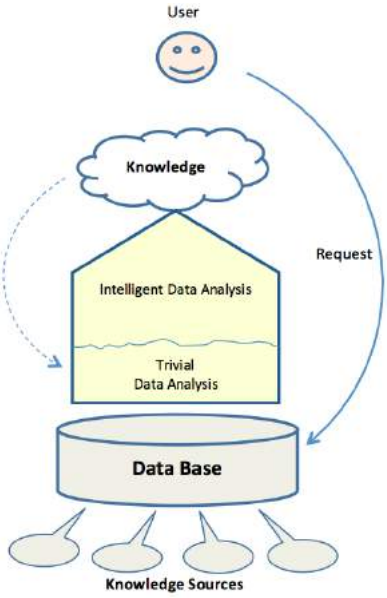

Figure 1. Place and gradation of data analysis of the certain object or phenomenon in user-oriented knowledge acquisition system.

Investigating the processes of obtaining new knowledge (hidden dependencies, regularities, trends, etc.) from structured data, developing and formalizing the mathematical apparatus, the scientific direction of the intelligent analysis of data was gradually formed. This direction has many close (rather, synonymous) names: Decision Making, Machine learning, Pattern Recognition, Data Mining, Knowledge Discovery, Advanced Data Analysis. This area is constantly updated with new terms, for example, Deep Learning, Big Data, Data Science, etc., a number of which, unfortunately, are of a mere conjunctive nature.

Today, processing and data centers, in which huge data flows are accumulated, are being created everywhere. Each major organization, each department collects, stores, protects and provides data, maintains databases up-to-date. For this purpose, software and hardware processing technologies, such as SPARK, Hadoop, Cassandra, EMC, etc., which contain tools for both trivial and intelligent analysis, are used industrially.

However, in view of our previous collaboration experience with data center representatives and direct database users [3], it turns out that in the vast majority of cases intelligent data analysis based on modern scientific achievements and computer technologies is almost never applied. As a result, databases exist and are constantly accumulating new data, but data analysis is still conducted by means of trivial methods. This, in turn, limits the value and efficiency of the useful information obtained.

Further in this paper we analyze the reasons for such state of affairs and propose an informal approaches (recommendations) to activate the practical application of the intelligent data analysis apparatus.

II. Basic Intelligent Data Analysis Terminology

Analysis of the mathematical methods and algorithms underlying the intelligent data analysis, shows that the said directions (Machine Learning, Advanced Data Mining, etc.) closely overlap. In this case, the basic concepts remain:

- feature (or informative feature) – in essence, the measured parameter, a unit of structured data;
- pattern – a vector of informative features.

Then the structured database is nothing more than a repository of patterns. And with the patterns, typical formal processing tasks can be performed: clustering, classification, ranking, forecasting, associative search, regression and some others.

These tasks can be solved by a limited list of formal methods or algorithms, such as neural network of a certain type, k-means, SVM, k-nearest neighbors, decision trees, etc. Most algorithms are based on the principle of measuring the distance between patterns using different measures (or metrics).

Some algorithms rely on the principle of differentiation or separation of patterns. These algorithms, as well as metrics for estimating the distance between patterns and ways of pattern differentiation, have long been described and sufficiently investigated in the textbook literature [4]. Moreover, most algorithms are already implemented in the form of program libraries of a number of systems and language environments, such as R, Weka, Wolfram Mathematics, Caffe, Tensorflow, cuDNN, scikit-learn, etc. It remains only to learn how to use this mathematical apparatus [5].

However, this often does not happen. The reasons for such a "modest" practical application of the mathematical intelligent data analysis apparatus is evidently due to:

- on the one hand, the problem of developing applied intelligent data analysis systems, or rather, the domestic experience of creating such systems is practically absent;
- on the other hand - the problem of a potential user who is not ready to master the achievements in the field of "Data Science" and solves pressing problems at the level of trivial data analysis.

Let's try to understand the reasons and give some advice or recommendations.

III. About Some Data Analysis "secrets" openly …

Based on personal long-term teaching experience in one way or another related to intelligent data analysis [6], as well as on the experience of scientific papers, articles and dissertations examination and research in this field, the following observations and conclusions were made.

A. Variety of Methods and Algorithms

Sometimes developers are simply lost in a variety of methods and algorithms related to the field of data mining, and the terminology used, which abounds with overlays and collisions and only aggravates the problem of entering the field of research.

Recommendation. We recommend using the ontology of the Intelligent Data Analysis presented in Fig. 2. It is limited to two or three most simple for understanding tasks (for example, ranking, classification and clustering) with studying two or three most popular algorithms for solving each of the problems. Moreover, for the solution of the applied problem, in 99 cases out of 100 there is no need to develop an original method. The main thing is to master and learn how to effectively apply the already accumulated arsenal of methods and algorithms. Moreover, to create an original algorithm is half the battle, it is necessary that the created or modified algorithm shows positive winning qualities in comparison with the known ones. Carrying out an objective comparative analysis of algorithms represents an independent scientific problem, the discussion of which is beyond the scope of this thesis.

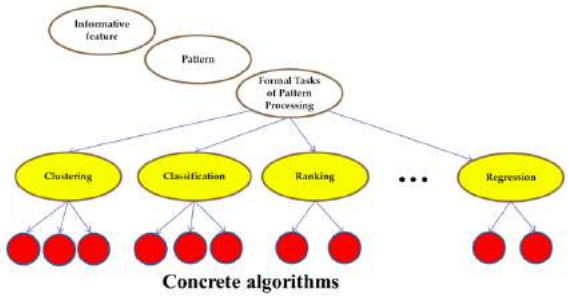

Figure 2. Key terms and typical intelligent data analysis tasks.

B. "Blind" Application

Unfortunately, some of the work of aspiring scientists is presented as a formal ("blind") application of one of the methods or algorithms of intelligent data analysis, followed by sometimes incorrect result interpretation in "their own favor".

Recommendation. The algorithmic complexity of solving an applied data analysis application usually goes beyond the single data analysis algorithm (machine learning). In general case, when solving problems from a particular domain, a chain of algorithms is used, both from the data analysis list and the conventional deterministic ones. At the same time, the task as a whole is informal, and an important role is assigned to specialists in data science who must correctly formalize the initial problem, correctly apply the mathematical (algorithmic) apparatus of the intelligent data analysis and correctly interpret the result.

C. "Small Data" Analysis

To date, a lot of data for analysis has been accumulated. But often important conclusions need to be made using only small portions (we shall call it "small data", as opposed to "Big Data"). For example, in medicine it is very hard to collect a large amounts of examples, it could be risky, expensive and, often, virtually impossible.

Recommendation. For such tasks it would be good to have metrics that would allow us to evaluate so-called "representativeness" of the training sample, and show how relevant the conclusions can be [7].

D. Data Visualization

Traditionally, the raw data and processing results are displayed as marked points on a two-dimensional plane. At the same time, it is often overlooked that in practice the number of informative features, and therefore the dimensionality of space is much higher and can reach tens and hundreds of parameters. On the other hand, data analysis begins with the visualization of data. On 2D or 3D visualization data specialist can allocate patterns, get an "insight" about the composition and internal structure of the data under examination. Very important is the task of correct and understandable visualization of the data under study, preferably without loss or with controlled losses.

There are methods of visualization that allow to compare the projections of data on the 2D or 3D plane. Insight about the data can also be obtained by analyzing the histograms, viewing the visualizations interactively. Data can be cleaned, normalized and transformed (for example, the principal data analysis), or represent data in the form of curves where the informative features are the coefficients of the curves, and it can be judged about the internal data structure by the shape of these curves [8]. In this case data will not be lost. Example of such visualization is shown in Fig. 3

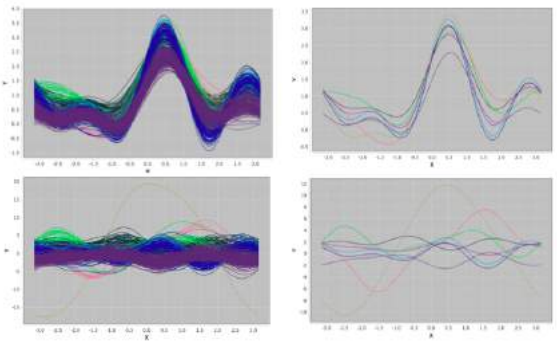

Figure 3. Andrews curves for multidimensional data visualization.

Recommendation. We see the following path in the direction of data visualization:

1) Simple methods using part of the information – the entire picture is not visible, it is difficult to draw conclusions.
2) Methods using all data as coefficients – sometimes difficult to interpret.
3) Methods using preliminary data transformation, representing the most informative aspects of data (the development of such methods can themselves be the object of intelligent data analysis: which features to choose, how to convert them, how to place and / or group images, etc.)

4) Interactive methods that allow data analysts to interact directly with real-time data in 2D or 3D, and in the form of augmented or virtual reality [9]. These methods can and should take advantage of the previous groups of methods - the simplicity and intuition of the first group, the completeness of the second and the "intelligence" of the third.

E. Practical Use

And, most importantly. Only the manager, who is highly interested in obtaining objective results, can formulate the actual task of data analysis in a specific subject area. For example, the director of a trading network is interested in knowing: "What is the reason of the outflow of customers in the current (or, certain) period?" It is assumed that there is no direct connection between this event and supplies, or the dismissal of any of the personnel. Another example, for the adoption of strategic management decisions it is necessary to evaluate and rank alternative versions of legislative initiatives (projects), relying on the statistics of previous periods, etc. In fact it can be revealed that:

- the director is not ready to formulate questions (queries) that go beyond trivial statistical analysis of data;
- the request goes beyond the possibilities of solving methods and means of data analysis, both for objective reasons (for example, an insufficiently representative database, insufficient time resources for research, etc.), and for subjective reasons (for example, the specialist's skill level in the analysis of data does not allow to solve the problem).

Recommendation. It is necessary to establish close cooperation between the user and the data analysis specialist. The user must understand the possibilities and know the objective limitations of data analysis technology. The expert analyst should know the terminology and problems of the subject area.

It is impossible to oppose the computer technology of intelligent data analysis to traditional research, i.e. intelligent analysis of data by an application specialist using trivial statistical analysis.

It must be borne in mind that the technology in question is not a silver bullet; not all tasks can be solved using intelligent data analysis, and for a number of tasks the application of this technology is unnecessary.

IV. Conclusion

We named only some of the most significant, in our opinion, aspects of intelligent data analysis. Of course, there are much more problematic issues worthy of discussion and research. These include: the problem of processing large amounts of data, the problem of providing (and / or evaluating) the representativeness of the data sample, as well as the reliability of the results of intelligent data analysis, the problem of the technical implementation of research and application systems.

A team of engineers and scientists specializing in this field has been working at the computer department of the Belarusian State University of Informatics and Radioelectronics for more than 10 years. At the department there is an innovative enterprise OOO "Intelligent processors". The team's assets include the creation of the first Belarusian neuro-like (neuronegal) computer (2010) [10], more than 30 publications on the research topic. In the period 2016-2020, research on the "Intelligent computing system for processing large data" is being carried out, with the goal of creating a tool for building applied data analysis systems. Our competencies include the development of algorithms and software to meet the customer's requirements, including participation in the formalization of the formulation of the problem, the choice of solutions and the interpretation of the results.

We are currently moving from academic research to practical applications and are starting to work with potential users from the organizations of the Ministry of Emergency Situations, the State Customs Committee, the State Committee for Forensic Expertise, the Center for System Analysis and Strategic Studies of the National Academy of Sciences of the Republic of Belarus.

References

[1] P. Buhlmann, P. Drineas, M. Kane, Mark van der Laan,"Handbook of Big Data," CRC Press, 2016, 464 p.

[2] S. Pyne, B.P. Rao, S.B. Rao, "Big Data Analytics: Methods and Applications," Springer, 2016, 276 p.

[3] A.V. Zhovna, V.M. Prorovskii, M.V. Khodin, N.D. Chistyakov, T.A. Kornacheva, "Analiz obstanovki s chrezvychainymi situatsiyami v Respublike Belarus' v 2016 godu [Analysis of the situation with emergency situations in the Republic of Belarus in 2016]," Chrezvychainye situatsii: preduprezhdenie i likvidatsiya [Emergency situations: prevention and elimination], No1(41), 2017, pp. 24–31.

[4] D. Cielen, A. Meysman, M. Ali, "Introducing data science: Big Data, Machine Learning, and more", Manning Publications, 2016, 325 p.

[5] A. Muller, S.Guido, "Introduction to Machine Learning with Python: A Guide for Data Scientists," O'Reilly Media, Inc., 2016, 394 p.

[6] A.I. Demidchuk, D.Yu. Pertsev, M.M. Tatur, D.V. Krishtal', "Intellektual'naya obrabotka bol'shikh ob"emov dannykh na osnove tekhnologii MPI i CUDA [Intellectual processing of big data volumes based on the MPI and CUDA technology]," Minsk, BGUIR, 2017, 60 p.

[7] V.V. Iskra, N.A. Iskra, M.M. Tatur, "Vliyanie statisticheskikh kharakteristik obuchayushchei vyborki na ee reprezentativnost'[The influence of the statistical characteristics of the training sample on its representativeness]," Nauchno-tekhnicheskii zhurnal Iskusstvennyi intellekt [Scientific-technical journal Artificial Intelligence], No4(62), Donetsk, 2013, pp. 325–332.

[8] C. García-Osorio, C. Fyfe, "Visualization of High-Dimensional Data via Orthogonal Curves," Journal of Universal Computer Science, 11 (11), 2015, pp. 1806-–1819.

[9] M. Huh, K. Kiyeol, "Visualization of multidimensional data using modifications of the Grand Tour," Journal of Applied Statistics, 29.5, 2012, pp. 721–728.

[10] D.N. Odinets, D.A. Lavnikevich, D.I. Samal', V.V. Starovoitov, "Postroenie informatsionnykh intellektual'nykh sistem na osnove neiropodobnogo komp'yutera [Building information intelligence systems based on a neuron-like computer]," Minsk, 2013, 43 p.

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ: ОТ ТЕОРИИ К ПРАКТИКЕ

Татур М.М., Искра Н.А.

В настоящей работе рассматриваются проблемы практического применения интеллектуального анализа данных. Даются рекомендации по созданию прикладных систем интеллектуального анализа данных.