БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ . BULGARIAN ACADEMY OF SCIENCES

ПРОБЛЕМИ НА ТЕХНИЧЕСКАТА КИБЕРНЕТИКА И РОБОТИКАТА, 53
PROBLEMS OF ENGINEERING CYBERNETICS AND ROBOTICS, 53

София . 2002. Sofia

On a Data-Driven Method for Research of Models

Vladimir Jotsov, Vassil Sgurev

Institute of Information Technologies, 1113 Sofia
Emails: jotsov@ieee.org, sgurev@bas.bg

1. Introduction

All modern data-driven methods for information acquisition can be divided in two groups according to the results obtained: algorithmic and non-algorithmic. Some of the machine-learning methods are a typical representative of the first group, allowing such data grouping, in which new knowledge is formed as cause-sequence relations (rules). The use of a set of similar rules allows the design of algorithms solving different types of problems. At that, from all the possible algorithms, only those are selected, that solve most efficiently the apriori defined problem (objective) *G*. In other words, *G* is a more specific or more general direction for knowledge discovery in a given subject area.

For some of the known data-driven realizations the algorithms are constructed using a wide spectrum of heuristic inferences, applications of non-classic logics and/or technologies of man-machine interaction [5, 10, 11, 20, 21] and they can represent more complex combinations of knowledge together with the connecting causal relations.

The second group of methods involves heuristic methods mainly. Some of the well-known researchers [3, 13] juxtapose the model and the algorithm as notions, taking into consideration that it is more natural to form knowledge of non-algorithmic character from the models. On the other hand, there exist entire scientific and applied directions, in which methods from the first group only are applied – in data mining, for example [16]. Some of the authors developing algorithmic solutions stick to the quite old thesis that the solutions of non-algorithmic character are so imperfect, that they are practically inapplicable. The truth in this classical dispute transferred to modern intelligent systems must be searched for in the defining of appropriate areas of application for each of the groups and their combination, if possible.

The present paper represents the results from the usage of one method of the first group, which is a suggestion of the authors. The connections between the processes of modeling, data processing with following development and adaptation of the formed (current) solutions of algorithmic character, are discussed. The topic is a logical continuation of the problems considered in [8]. Different variants for development of the newly formed algorithms in environments of several different models consisting of numerical data are discussed

In modern intelligent systems the more complex the system itself is, the higher the requirements towards modeling of the subject area are. In systems of CASE type – with experience gaining – in discovering systems, data mining, in the direction of artificial life [4, 6, 9, 15, 16, 18], as well as in the prevailing part of systems with inference by analogy, the results are in direct relationship with the various models, in connection with the subject domain selected. Machine-learning procedures depend also on the way and completeness in modeling the necessary knowledge [12].

The paper offers the application of different models for data description, including such models that describe one and the same data in different ways. The alteration of the models enables the evaluation of the stability and correctness of the solutions of algorithmic character.

2. Functions for data representation

In some of the cases the way of data representation defines their further usage and the acquisition of new knowledge from them. In order to illustrate the example above given the following example is considered.

Example 1

Let us consider an arithmetic progression $\{1+1k\}_{k-1}^{\infty}$, presenting all the natural numbers. Let the following objective G be investigated: laws of prime numbers p, i.e., numbers that are divided without a remainder by two whole integer numbers only: 1 and p. For the purpose of the study each integer positive number is accepted as a type of data. The data can be ranked and grouped in many ways, for example as shown in Figs. 1–5, where the data are located in linear, spiral, complex pendulum or sinusoid form, the prime numbers being marked by circles.

The prime numbers form different images in Figs. 1–5. In order to investigate the goal G, i.e., to study the set of prime numbers $p \in P$ and its distribution among natural numbers, every researcher would make an attempt to expand the fragment from a given figure and to find repeated parts from p or similarities between different parts of the set P, or analogies between fragments from different figures, etc.

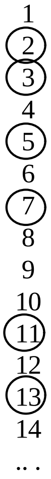

Fig. 1

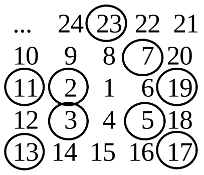

Fig. 2

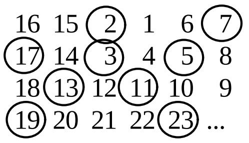

Fig. 3

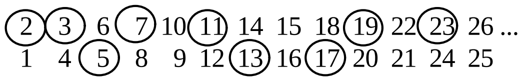

Fig. 4

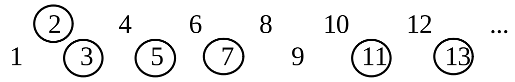

Fig. 5

The survey of modern intelligent systems, in which the analogous data patterns are segmented, clustered or classified according to the type of technologies, which study the nearest neighbours or by other methods, is beyond the frames of the present paper. Most of the methods pointed out are classical [1, 19] and the innovation of modern investigations is rather in finding such a minimal construction in them, that does not lead to exponential complexity of the corresponding applied software than in the application of similar methods. Due to this reason we did not find possibilities to apply or derive new information from digital data with the exception of the particularly efficient inference by analogy in man. The complexity of the solutions obtained was near to exponential, and its reducing by heuristics usage lead to restricting the method applications as a whole. Hence in [7] we suggested the simplest group of procedures for knowledge discovery after data analysis: observation (O), juxtaposition (J) and mathematic induction (MI). The results in the working paper cited can be used by other models of data, for example in models of work [14].

In data-driven methods the formulation of problem G is often set in a specific, characteristic way and it can accomplish another role compared to its purpose in other methods, for example in the wide spread deductive inference in artificial intelligence systems. In data-driven cases G can "hide from view" among the control heuristic meta knowledge, which is often replaced with regard to "system orientation" as in Lenat [10]. But even if G be given in explicit form as a direct task (not only as a wish or general direction for knowledge discovery), it has a more suggestive than instruction character for us. Analogy can be done between the setting of G in this case and the fitness functions of evolutionary programming [2], though this analogy is quite distant. As in the case of evolutionary programming, we assume that current solutions – in analogy with the population members – may not only approach, but also go away from the objective. Referring to example 1, where G is quite a general goal "to study the distribution in p"; we "divert from the objective". For example, we may investigate the properties not of the prime but of the compound numbers. According to our opinion, it is appropriate to control the data-driven systems not by one, but by a hierarchic system of objectives G_j, with the main objective $G = G_0$. Under it, objective G_1 is set which controls the form of the desired results. In the example considered “the results should be in the form of a functional relation between the numbers”. G_1 defines the orientation towards construction of theorems in automatic or man-machine operation. In the indicated system of objectives and “free search” of new knowledge, some situations are possible, when neither G no G_1 are completely satisfied, but the current results obtained are so interesting, that they do not require alteration or modification of the objectives. According to our data a similar type of systems is more appropriate for man-machine work. The procedure above described is called observation (O) and it is discussed in details in [7]. A second example is below given, where the results from operation with digital models and with a system of objectives G_0 and G_1 are shown.

Example 2

Let us assume that in the process of man-machine operation two hypotheses have been formed – H1 and H2. They are considered as satisfied if they do not contradict to the data from the digital model M^* selected – one of the models in the figures above given – and they are executed for each appropriate combination of data. Let us assume that the two auxiliary procedures in connection with “validating of the results” from observation (O) are the following:

- test of data inaccordance (I),
- test of data accordance (A).

They are applied to the hypotheses in the example without any success and hence it is accepted that the two hypotheses do not satisfy the objective G_1. The following H3 is established in the research of the hypotheses in M^*: if H2 is true, then H1 is also true. Then if H3 passes successfully tests I and A, new information about the connection between H1 and H2 is added to the data basis – in spite of the fact that they are already confirmed.

Procedure (O) lies in the basis of the processes for new knowledge discovery. Applying (O) to the digital sets in the figures above given, we shall obtain different results depending on the figure choice, because the data distribution gives additional information about them.

The next procedure from the set pointed, juxtaposition (J) has an assisting character with respect to (O). The juxtaposition between different data sets leads to the discovery of repeating segments or to any other new information. In our method the juxtapositions are introduced with the purpose to replace some algorithmically more complex procedures for search of similarities and analogies.

The mathematical induction (MI) procedure is introduced with the purpose to distribute new knowledge from the data fragment investigated on the whole model M^* – in example 1 it is an infinite set of data (numbers). Procedure MI has no new element – in a theoretical aspect – and has auxiliary character in the processing of the results from the other two procedures (O) and (J).

3. Adaptation of the results obtained

The choice of an appropriate (optimal) model M^* is in the basis of new knowledge formation. As above shown the results depend on the selection of model M among all the possible ones. The choice of the most appropriate model M^* for the objectives G_i is a very responsible process, and its automatic realization in modern intelligent systems is not recommended.

After the solution S_i – which represents the new knowledge in an algorithmic type – is formed and checked in M^*, it has to be checked in the other appropriate models $\{ M_j\}$ which contain data from M^* in one or another form – see example 1. Depending on the results from the transfer of the newly formed algorithm to other models, the following situations appear:

I. No additional alteration of S_i is necessary in order to adapt the algorithmic insurance for operation in M_p, because the algorithm functions with equal success in M^* and M_p also. The above said is formulated as:

$$T(S_i, M^*).T(S_i, M_p),$$

and after that M_p is replaced by the next model in $\{M_j\}$. The applicability of S_i for the next model is checked. It is assumed that the more $T(S_i, M_j)$ are valid for it, the more successful S_i is.

Let us assume that S_i is applied with equal success to the whole set $\{M_j\}$, i.e., $T(S_i, \{ M_j\})$. Since no alteration (adaptation) of S_i is necessary with respect to the elements in $\{ M_j\}$, it is accepted that the whole S_i is invariant to the alteration of the models from the set selected: $\text{inv}(S_i) = S_i$.

II. Let S_i be inapplicable to M_p. At the same time it follows by definition that $T(S_i, M_p)$. In this case it is necessary to investigate the reasons for the inapplicability of S_i to M_p.

If M^* and M_p comprise one and the same data as described in example 1, the inapplicability of S_i in M_p indicates its possible incorrectness (refer to situation V}. The intermediate results are incorrect when formed as a result of the operations (O) or (J) based on rough geometric observation on M^*, for example. After these ideas are transferred to M_p they will fail and S_i is to be corrected or rejected in model M^* initiating the solution (see situations IV and V).

Another reason for the appearance of situation II is the inaccordance between M_p and M^*. In practice the whole set of models from $\{M_j \}$ may not correspond to M^* to such an extent, that the algorithmic solution to be applicable to the rest of the models. On the other hand, if after the exhaustion of $\{M_j\}$, only $T(S_i, M^*)$ is valid, S_i is subject to serious reconsidering after additional investigations in M^*.

The appearance of situation II leads to formation and study of questions of the type: what are the differences between the application of S_i in M^* and in M_p; in what way S_i can be realized outside M^*; why S_i is not compatible with any M_p, etc. These questions cause more investigations concerning S_i and as a result lead to its improvement.

III. The apriori set condition T (S_i, M^*). is not valid for M_p, but there exists a solution S_i, which is a modification or which resembles S_i (i.e. the approximate idea from the initial algorithm is used). The following condition is valid at that:

$$T(S'_i, M^*).$$

In this case the checking procedure described before situation I, is applied towards S'_i. It is also investigated which part in S'_i could be applied for all or for most of the models in $\{ M_i\}$. In other words, inv(S'_i) is studied. If

$$\text{inv}\ (S'_i) > \text{inv}\ (S_i),$$

it is assumed that the new solution is stronger than the initial solution S_i.

Regardless of the result from the comparisons between S'_i and S_i the mere existence of S'_i increases the role of S_i, since it represents the distribution of the ideas from S_i in other models M_p

$$S'_i = A(S_i, M_p).$$

If the power of each S_i is measured by any quantitative system, then

$$0< A\ (S_i, M_p)< T\ (S_i, M_p).$$

The existence of $A(S_i, M_p)$ attracts the attention to this part of S_i, which is contained in S'_i, and places questions of the type: why is there a common part in inv(S_i), and inv(S'_i), why $S_i \neq$ inv(S_i) and so on. The study of similar questions often leads to the discovery of stronger and universal solutions of algorithmic character. The most important point in this case is that in the systems with built in procedures for analysis of situation II, there is a chance to implement and apply automatic strategies for self-directing study and data-driven adaptation of the current results obtained, which is an element of self-learning.

IV. In case S_i is an incomplete solution from M^*, its applying to other M_p will cause the appearance of inaccordances or incomplete applicability of S_i. In this case such investigation of S_i in M^* and $\{M_j\}$, is necessary that leads to the modification of S_i and the coming out of S_i^*, which is stronger than S_i – because it is a fuller renewed variant of the former solution. The process of development S_i – S_i^*, leads to the formation of more universal solutions. The practice shows that the initial solutions connected with complex problems are not complete and final; they pass an “evolutionary way” from the rough idea to the final results, for example – mathematical proofs. The replacement of M^* – elements from $\{M_j\}$, plays an important role for the development and completing of the intermediate results of algorithmic character.

V. In case S_i is an incorrect hypothesis, the way of alteration of the models leads to clarifying its inconsistency. The procedure pointed – see situation I or III enables the reservation of the “rational grain” in S_i which is inv(S_i). In other words, if a part of hypothesis S_i is correct, then it is contained in the non-empty set inv(S_i) and can be modified, stored or included in other algorithms.

Depending on the behavior of the investigated situations I–V, every algorithm can be checked, improved or altered by different means and tools, described in the next chapters of the present paper.

4. Improvement, correction and development of the intermediate solutions

After the suggested sequence of steps in the corresponding algorithm is applied to different models, it becomes more universal (*powerful*), because properties are found in it, that are invariant with respect to some of the parameters. In this way the significance of the juxtapositions (J) in the approach as a whole increases, because (J) enables the transition and change in A_i, and (M) confirms and distributes the newly found properties [10].

A specific difference of the iterative procedure of “algorithm–models” type suggested is that the algorithmic solutions are evaluated not only according to some known parameters such as speed, volume of the necessary memory or other quantitative estimates. The qualitative estimates have greater significance in the method discussed – flexibility, adaptiveness of the algorithms, etc., that are in the base of modern discovering and/or creative systems. Parallel to them some quantitative notions are considered – such as power, used with respect to the factor “invariability” of the different parts in one algorithm.

At the beginning of the study the processes *can* operate in automatic mode and new *knowledge* (hypotheses) can be formed transforming different structures of the data, by finding repeating fragments and *data* grouping, parts of (H). The induction rule (M) is applied to the hypotheses. The operation mode pointed out is proposed and the model investigated is selected by the specialist, learning the problem. He should be able to **direct** the process setting the problem formulation and the general priorities of work. In connection with the matter discussed in the previous chapter it can be pointed out that two absolutely different formulations may lead to equal results under certain conditions. For example, one formulation is for the prime numbers-twins, the other one is a heuristic formulation of the type “search anything nice, i.e. it is an unambiguous or complex logical task, clear geometric interpretation and simple explanations”. In both cases it is necessary to formulate and prove one and the same basic theorem (see Theorem 1 in [9]).

The leading role of man is typical for the technology of knowledge acquisition suggested. The system can provide interesting observations and hypotheses for him and in this way avoid his efforts to compare large data bases. Knowledge is derived in the system and its evaluation and the confirmation of the necessary theorems is on behalf of the user.

The correct evaluation of the results obtained is still an intuitive process. But there exist situations when the specialist cannot make an evaluation mistake in practice. For example, the results cannot be neglected, when the research development causes rapid increase of the role of the algorithm formed. This phenomenon can be compared to an avalanche. Several steps are sufficient in the zone and everything around rotates and quickly gains inertia. As above noted, the approach suggested *occupies* a given area of knowledge, closing it along the perimeter. The area closed may contain a number of unsolved or superficially solved problems of the type “HOW” and “WHY” an inference is obtained. There are ways to improve the solutions obtained. In order to evaluate the situation under these conditions, the question “what is going to happen, if the hypothetic inference is true”, is being set. In case the applications are sufficient or *interesting*, the area bordered is studied in detail with the purpose to transfer the logical inference into a mathematically correct proof.

The shortcomings of the solutions offered are connected with the fact that they cannot be applied to non-formalized knowledge. When operating with data or knowledge outside numbers theory, they have not any proved efficiency. At the same time the approach has the following advantages. A larger part of the theoretic approaches, schemes and the accompanying algorithmic and software insurance can be used in other domains as well. Nowadays some investigations are realized to use these approaches in signal processing and in quality control.

The final result from the application of the approach proposed has been checked in numbers theory. Algorithms solving the problem defined are obtained, which have the power of a mathematical proof. They are formed in the process of knowledge acquisition from the digital data bases, constructing a logical inference, with generation of hypotheses concerning the problem set. The mathematical proof is formed after development or *evolution* of the logical inference. The proof does not require citation of the initially obtained logical tools for inference that lead to its generation, since they have not any proving effect. On the other hand, this knowledge must not be neglected because it can lead to improvement of the existing solution or to other, sometimes unexpected solutions.

The system supports the decision process in the following manner. Algorithms are built on the basis of three procedures only: mappings, observations and induction. Starting from one of the models, each algorithm should be tested (and strengthened) in the other *possible* models. If the algorithm used is good enough, and the models are appropriate, the user finds out an algorithmic part, which is invariant to model shifting. The analysis of this part shows the way for new perspective investigations. On the other hand, if the iterative procedure pointed out fails, the user will obtain information that the algorithm could be applied to a very restricted domain or that the current algorithms or the model should be changed because of inconsistent knowledge.

5. Application and perspectives

The approach above described possesses many possibilities for applications in artificial intelligence systems or in decision support systems. One of the more characteristic cases involves the application of specific heuristics in knowledge acquisition. For example, let us suppose that during the dialogue with the specialist, a solution of the apriori defined problem is obtained. In the “classical” case this ends the job, but not in the variant proposed. On the contrary, the next iteration of the solution test starts. The system issues a message for an algorithm test in *close* or in *analogous* models. In many of the cases this will not alter the results formed, but it may improve the algorithm or find a contra example or other obstacles for its realization.

It is necessary to pay attention to the fact that different formulations of the problem have been used in the second chapter, including multiobjective formulations. Depending on the study process of the apriori set model, the operation may pass from one problem to another problem, close by model or close by algorithm. A similar transition is not possible or it is not well developed in traditional algorithmically directed approaches. New results are expected with the use of information technologies in hard or in weakly formalized subject areas.

Examples from other scientific areas can be pointed, when one and the same algorithms scheme is repeated with the use of different terminology. This example is included in the description of situation III, where the unaltered part is the basis of the algorithms, and the altering part is a kind of a superstructure above the basis and it can comprise different terminologies.

Some other specific features and advantages of the approach in applied investigations can be described by analogy with the matter from the second chapter. The investigations show that there are no obstacles for the application of the approach as a whole or of its separate parts in heuristic and in formal systems and applied information technologies.

6. Conclusion

The paper proposes a method, in which the choice and alteration of different models and algorithms connected with the solution of the problems set, is a process, realized successively in man-machine operation. The selection of model M^*, from which the initial algorithmic solution is obtained, is of particular importance. It is recommended the selection of models to be done after solution of the respective optimization problems, which is a future perspective.

The algorithm formed at the first iteration, is successively improved by its use in other models. The components that do not alter with models change are found in it and additional check is simultaneously done for correctness of the algorithmic solutions with respect to different quantitative and qualitative parameters.

The approach is subject-independent. On the basis of the matter discussed different applications in intelligent systems as well as in other areas of mathematics and informatics are expected.

References

1. Banerji, R. The logic of learning: a basic for pattern recognition and for improvement of performance. - Advance in Computers, **24**, 1985, 187-215.
2. Evolutionary Algorithms in Engineering Applications. D. Dasgupta, Z. Michalewicz (Eds.). Berlin etc., Springer, 1997.
3. Evgenev, G. Models instead of algorithms: shifting the paradigm for design of application systems. - Information Technologies, **3**, 1999, 38-44.
4. Fayyad, U., G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (Ads.). Advances in Knowledge Discovery and Data Mining. MIT Press, 1996.
5. Fayyad, U., G. Grinstein, A. Wierse. Information Visualization in Data Mining and Knowledge Discovery. Morgan Kaufmann, 2001.

6. Fogel, D. B. Blondie24: Playing at the Edge of AI. Morgan Kaufmann, 2001.

7. Jotsov, V. On the usage of discovering systems approaches in Number Theory. - IIT Working Paper WP/83B, 1999, p. 14.

8. Jotsov, V. An iterative approach for building algorithms. - In: Proc. XXXV Int. Conf. “Communication, Electronic and Computer Systems”. Tecnical University of Sofia, 2000, 131-136

9. Langley, P. et al. Scientific Discovery: Computational Explorations of the Creative Processes. Cambridge, MA, MIT Press, 1987.

10. Lenat, D. The nature of heuristics. Artificial Intelligence, **19**, 1982, 189-249.

11. Lenat, D. Why AM and EURISKO appear to work. - Artificial Intelligence, **23**, 1984, 269-294.

12. MichalskI, R., R. Stepp. Learning from observation: conceptual clustering. - In: R. Michalski et al. (Eds.) - In: Machine Learning: Artificial Intelligence Approach. Vol. I. Morgan Kaufmann, 1983, 331-363.

13. Narin’yani, A. Model or algorithm: a new paradigm for an information technology. – Information Technologies, **4**, 1997, 11-16.

14. Quinlan, J. C4.5: Programs for Machine Learning. Morgan Kaufmann, 1996.

15. Ram, A. Creative conceptual change. - In: Proc. Fifteenth Ann. Conf. of the Cogn. Sci. Society, Boulder, CO, 1993, 17-26.

16. Ramakrishnan, N., A. Grama. Data mining: from serendipity to science. – In: IEEE Computer, August 1999, 34-37.

17. Verykios, V. et al. Automating the analysis of option pricing algorithms. – In: IEEE Trans. SMC - Part A, **31**, 2001, No. 6, 573-586.

18. Weber, R., D. Perkins (Ads.). Inventive Minds: Creativity in Technology. New York, Oxford University Press, 1992.

19. Winston, P. Learning new principles from precedents and examples. - Artificial Intelligence, **19**, 1982, 321-350.

20. Zenkin, A. Cognitive Computer Graphics. Moscow, Nauka, 1991, http://www.alex2zen.ru

Метод, управляемый данными, и его интерпретации в различных моделях

Владимир Йоцов, Васил Сгурев

Институт информационных технологий, 1113 София

E-mails: jotsov@ieee.org, sgurev@bas.bg

(Резюме)

В статъе представлен метод, позволяющий формирование алгоритмических решений в одной априорно избранной модели M* и их последующую оценку, применение и развитие в других моделях. Представлены примеры работы с некоторыми моделями представляющими множества целых чисел. Даны ссылки на пред-дущие работы авторов, раскрывающие более подробные примеры работы с предложенным новым методом и специфические особенности работы в данном направлении. Подход предметно-независим и может быть использован с почти одинаковым успехом как при решении абстрактных задач математического характера, так и в прикладных областях науки и техники. Метод представляет оригинальную авторскую разработку и сравнивается с известными представителями из близких и смежных научных областей.