UDC 004.9

O.S. Bulgakova

DATA MININGMETHODSBASED ONSELF-ORGANIZINGMODELS

Abstract. The paper proposes analysis ofdata miningmethodsbased onself-organizingmodels. Reviewing the basic idea of the modified Group method of data handling (GMDH) known as the Group of Adaptive Models Evolution (GAME) network. Presents an original idea of active neurons (different activation function are sorted), this idea might be used to improve the efficiency of models based on the generalization of structures of iterative and combinatorial type algorithms.

Keywords: data mining, domain ontology, ontological information, generalized iterative algorithm , inductive modeling, structures of data, handling and storing of information.

Introduction

Problems of data mining complex systems can be solved using both the logical deductive methods and sorting-out inductive ones. Deductive methods have advantages in the case of simple modeling tasks, if there is known the theory of an object being modeled and therefore it is possible to build a model based on physically based principles using knowledge of processes in an object. But these methods can not give satisfactory results for complex systems. In this case, an approach of knowledge extraction directly from the data based on experimental measurements has an advantage. A priori information about the properties of such objects may be only minimal or even absent.

One of the most well-known modeling techniques devoid of problems described above is the group method of data handling that discovers knowledge about the object directly from the data sample.

Group method of data handling is currently widely used in solving various problems and actively applied in tasks where usage of conventional algorithmic solutions is ineffective or even impossible. To increase accuracy and expand horizons for GMDH application, many researchers have been studied some aspects of GMDH and proposed as well as developed hybrid-type algorithms. At present many domestic and foreign researchers like E.Bodyanskiy (Ukraine), P. Kordik (Czech Republic), T.Kondo (Japan) and others are actively developing GMDH-like systems based on multilayered algorithm. For instance, in [1] the use of genetic algorithm is described for optimization of multilayered GMDH structure, and factors are finding at that by the method of singular decomposition (SVD, Singular Value Decomposition).

© BulgakovaO.S., 2014

1. GMDH as a polynomial neural network

Both the GMDH author Ivakhnenko and many users of his method, especially in the last 20 years when artificial neural networks have gained wide popularity, began to call its typical structure also as a neural network. Moreover, in recent years GMDH algorithm among the professionals abroad is called often as Polynomial Neural Network (PNN).

Here one of the main elements of iterative GMDH algorithms, namely any partial description, can also be considered as an elementary neuron of the neural network PNN GMDH. The structure of such neuron for the quadratic partial description is shown in Fig. 1. The originality of the neural network with such neurons consists in speed of the process of local adjustment of weights of neurons and automatic global optimization of the network structure (number of units and number iterations or hidden layers).

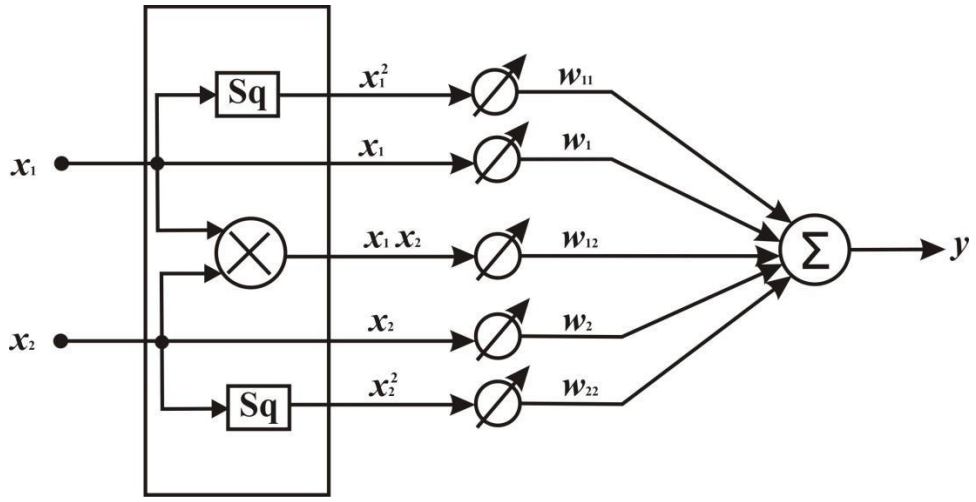

Fig. 1. Structure of GMDH neuron with the quadratic particular description [2]

Elements of polynomial GMDH neural networks and its use is discussed in [2-7] in details.

2. The idea of neural networks with active neurons

At the early stages of the GMDH theory development, the similarity between neural networks and multilayered GMDH algorithm was observed. O.Ivakhnenko in one of the articles stated that the differences between the perceptron and GMDH are not fundamental, so it is acceptable to call GMDH-systems as "perceptron-like systems".

Before explaining what the active neuron is like, there is a need to characterize firstly the passive neuron.

Mechanisms of optimization of the input variables set are not founded in all kinds of standard (passive) neurons. They are set in the complex process of self-organization of the whole system of many neurons in general. Structural optimization by an external criterion and receiving optimal non-physical models in the perceptron is not provided, only parametric optimization using e.g. inverse calculation of errors ("back propagation") is performed. Perceptron is almost not inferior to polynomial GMDH algorithms with respect to accuracy if the learning set is rather long, noise variance is small, and the set contains variables quantized on a small number of levels, i.e. when the physical model appears to be optimal. The advantage of GMDH models in accuracy is reached at short data samples of continuous noisy variables, i.e. if a non-physical model becomes optimal. Advantages of perceptron and polynomial GMDH algorithms are combined in Neural Network with active neurons.

The similarity between the structure of neural networks and GMDH algorithms inspired researchers to explore how they can be combined. O. Ivakhnenko et al. [3] propose the combined method that extends the theory of self-organization from fixed structures to active neural networks. Proposed algorithm known as "neural network with active neurons" is used instead of a passive neuron in the GMDH algorithm [4-6].

Both multilayered and combinatorial GMDH algorithm can be used as active neuron. O. Ivakhnenko [2] proposes to modify the combinatorial algorithm into the algorithm with active neurons that leads to increasing of accuracy and reducing of calculation time.

The advantage of GMDH neural networks with active neurons as compared with conventional neural network with binary neurons consists in that self-organizing of the network is simplified: each neuron finds necessary connections and its structure by itself in the process of self-organization.

Neural networks with active neurons were successfully used for prediction of the processes in economical and ecological systems [4, 5-9].

Idea of active neurons described above served also as the basis for generalization of the previous modifications in order to significantly improve the efficiency of iterative GMDH algorithms.

3. GMDH-type neural network with feedback

T. Kondo proposed a new multilayered GMDH-type neural network with feedback [7] which may by itself automatically choose the optimum neural network architecture using the idea of self-organization. Architecture of the GMDH-like neural network has a cycle of feedback. In this algorithm outputs of neurons are connected

with the system inputs (feedback) for further calculation. Thus, the complexity of the neural network increases gradually. This GMDH-type neural network algorithm has an ability of self-selecting optimum neural network architecture from three variants such as sigmoid function, radial basis function (RBF) and polynomial neural network. In addition, structural parameters such as number of neurons in hidden layers and input variables are selected automatically by minimizing the error criterion defined as Akaike criterion:

$$AIC(s_f) = -2\ln \varphi(X, \hat{\theta}_f) + 2s_f, \quad (1)$$

where φ is the likelihood function.

For each combination, optimum neuron architectures are automatically selected from the two type neurons: neuron architecture with two and with r inputs.

Proposed by T.Kondo the GMDH-type neuron network with feedback has been successfully used for the medical problem of recognizing 3-dimensional images of lungs [9].

It may be noted that this idea with submitting outputs to the input was partially implemented in the multilayered algorithm with selection of initial arguments. In this research the idea is also reflected in the developed generalized algorithm.

4. Group of Adaptive Models Evolution

An evolutionary algorithm based on GMDH and called GAME (group of adaptive models evolution) has developed in the Czech Technical University in Prague by P. Kordik [43]. The coefficients of unit transfer functions are estimated in GAME on the basis of the data set describing the system being modelled.

Major modifications of the GMDH-type system are [10]:

the transfer function of the unit is of several types (linear, polynomial, logistic, etc.) and it can be provided by a perceptron network too. Each type of unit has its own learning algorithm for coefficients estimation. Choice of the type of units that form a network is determined by the given criterion. This is so-called heterogeneous network structure;

the number of unit inputs increases together with the depth of the unit in the network. Transfer functions of units reflect growing number of inputs.

there exist interlayer connections in the network.

the network construction process does not search all possible layouts of units. It searches just the random subset of these layouts. The original GMDH produces one

optimal model. This method produces the group of models that are locally optimal, each for its specific subset of unit layouts.

The Modified GMDH generates a group of models on a single training data set. The random processes influence the construction procedure. Weights and coefficients of units are randomly initialized. Transfer functions of many units types are defined pseudo-randomly when the unit is initialized. Inputs for units are selected pseudo-randomly as well. It results in the fact that the topology of models developed on the same training data set differs (fig.2, fig.3).

After reviewing the basic idea of the modified GMDH known as the GAME network, we can conclude that it like as in the algorithm Kondo also presents an original idea of active neurons (different activation function are sorted), so this idea might be used to improve the efficiency of models based on the generalization of structures of iterative and combinatorial type algorithms.

5. Hybrid Differential Evolution and Group Method of Data Handling for Inductive Modeling

The group method of data handling and differential evolution (DE) population-based algorithm are two well-known nonlinear methods of mathematical modeling. In paper [11] of Godfrey C. Onwubolu, a new design methodology which is a hybrid of GMDH and DE was proposed. The proposed method constructs a GMDH network model of a population of promising DE solutions. The new hybrid implementation was applied to modeling and prediction of practical datasets and its results were compared with the results obtained by GMDH-related algorithms.

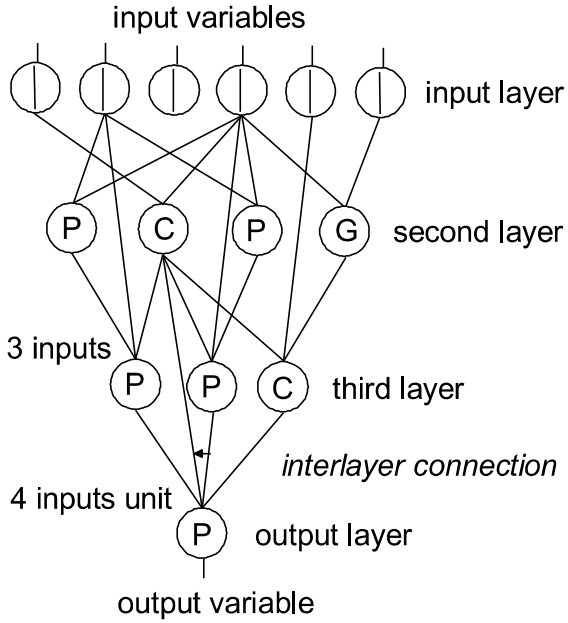

Fig. 2. GAME-network structure shema

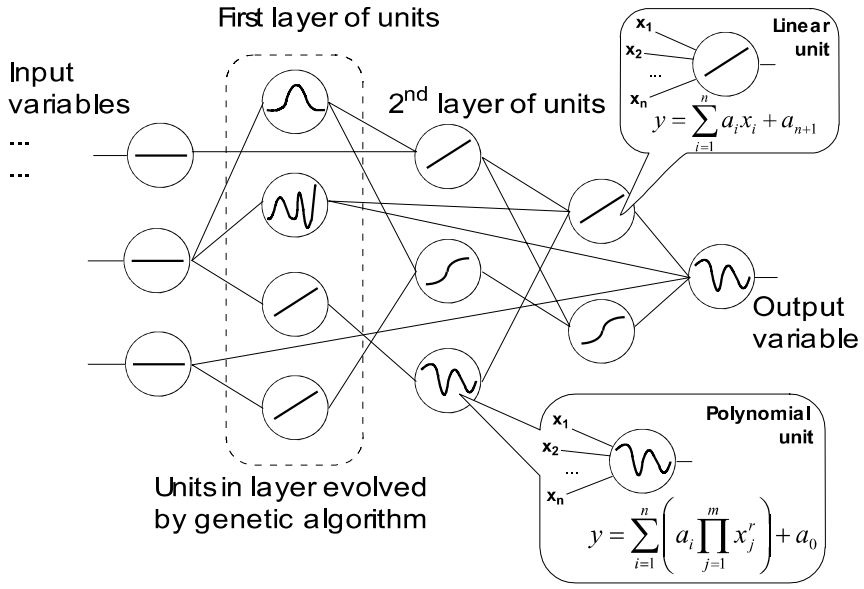

Fig. 3. GAME-network structure

The modeling methods have many common features, but, unlike the GMDH, DE does not follow a pre-determined path for input data generation. The same input data elements can be included or excluded at any stage in the evolutionary process by virtue of the stochastic nature of the selection process. A DE algorithm can thus be seen as implicitly having the capacity to learn and adapt in the search space and thus allow previously bad elements to be included if they become beneficial in the later stages of the search process. The standard GMDH algorithm is more deterministic and would thus discard any underperforming elements as soon as they are realized.

Using DE in the selection process of the GMDH algorithm, the model building process is free to explore a more complex universe of data permutations. This selection procedure has three main advantages over the standard selection method. Firstly, it allows unfit individuals from early layers to be incorporated at an advanced layer where they generate fitter solutions. Secondly, it also allows those unfit individuals to survive the selection process if their combinations with one or more of the other individuals produce new fit individuals. Thirdly, it allows more implicit non-linearity by allowing multi-layer variable interaction.

The new DE-GMDH algorithm is constructed in exactly the same manner as the standard GMDH algorithm except for the selection process. The selected fit individuals were entered in the GMDH algorithm as inputs at the next layer. The whole procedure is repeated until the criterion for terminating the GMDH run has been reached. Presented in [44] results show that the proposed algorithm appears to perform

reasonably well and hence can be applied to real-life prediction and modeling problems.

Conclusion

This paper considers in detail the basic structural elements of data mining methods based on self-organizing models based on typical GMDH algorithm, analyzes its advantages and shortcomings, describes the historical ways to overcome these shortcomings. Presented an original idea of active neurons (different activation function are sorted), this idea might be used to improve the efficiency of models based on the generalization of structures of iterative and combinatorial type algorithms.

REFERENCES

1. Ivakhnenko A.G. Inductive learning algorithms for complex system modeling / Ivakhnenko A.G., Madala H.R. // – London, Tokyo: CRC Press, 1994.
2. IvakhnenkoG.A. Self-Organization of Neuronet with Active Neurons for Effects of Nuclear Test Explosions Forecastings/ IvakhnenkoG.A //System Analysis Modeling Simulation (SAMS), 1995, vol.12, no.1, 1-10 pp.
3. Bodyanskiy Yevgeniy. Hybrid radial-basis neuro-fuzzy wavelon in the non-stationary sequences forecasting problems / Bodyanskiy Yevgeniy, Vynokurova Olena// Proceedings of the II International Conference on Inductive Modelling ICIM-2008, 15-19 September 2008, Kyiv, Ukraine. – Kyiv: IRTC ITS NANU, 2008. – P. 144-147.
4. Ivakhnenko A.G., Inductive sorting-out GMDH algorithms with polynomial complexity for active neurons of neural networks/ Ivakhnenko A.G., D. Wunch and Ivakhnenko G.A. // Proceedings of the International Joint Conference on Neural Networks, Piscataway, New Jersey, USA, IEEE, 1999.
5. Valenca M. Self-organizing modelling in forecasting daily river flows / Valenca M. and Ludermir T. // Proceedings of the 5th Brazilian Symposium on Neural Networks, 1998. –210-214 pp.
6. Ivakhnenko A.G., Self-organization of neuronet with active neurons for effects of nuclear testsexplosions forecasting// Systems Analysis Modelling Simulation, vol.20, no.1-2, 1995. –107-116 pp.
7. Kondo T. GMDH neural network algorithm using the heuristic self-organization method and its application to the pattern identification // Proceedings of the 37th SICE Annual Conference -SICE'98, pp.1143-1148, 1998.
8. Akaike H. A new look at the statistical model identification // IEEE Trans. Automat. Control.- 1974.- v.19.- N 3.- P 716-723.
9. Tadashi KondoFeedback GMDH-Type Neural Network SelfSelecting Optimum Neural Network Architecture and Its Application to 3-Dimensional Medical Image Recognition of the Lungs / Tadashi Kondo, Junji Ueno // Proceedings of

the II International Workshop on Inductive Modelling IWIM-2007, 19-23 September 2007, Prague, Czech Republic. – Prague: Czech Technical University, 2007. – P. 63-70. – ISBN 978-80-01-03881-9.

10. Kordík The Modified GMDH Method Applied to Model Complex Systems / Kordík, Náplava, Šnorek, Genyk-Berezovskij // proceedings of ICIM'2002 Conference, Ukraine, Lviv.
11. Godfrey C. Onwubolu. Design of Hybrid Differential Evolution and Group Method of Data Handling for InductiveModeling / Godfrey C. Onwubolu // – Proceedings of the II International Workshop on Inductive Modelling IWIM-2007, 19-23 September 2007, Prague, Czech Republic. – Prague: Czech Technical University, 2007. – P. 87-95. – ISBN 978-80-01-03881-9.