Inductive Sorting-Out GMDH Algorithms with Polynomial Complexity for Active Neurons of Neural Network

A.G. Ivakhnenko[1], D. Wunsch[2], G.A. Ivakhnenko[3]

[1] Cybernetic Center of NASU, Kyiv, Ukraine, Gai@gmdh.kiev.ua
[2] Texas Tech University, Lubbock, USA, DWunsch@aol.com
[3] National Institute for Strategic Studies, Kyiv, Ukraine, http://come.to/GMDH

Abstract: *Neural networks with active neurons which self-organize their structure can use inductive sorting-out GMDH algorithms for their neurons. New threshold type GMDH algorithm with polynomial complexity is developed to decrease computing time in case of large input data sample.*

Introduction

Idea of twice-multilayered neural networks self-organization originally was published in [1]. There was shown that any learning unsupervised forecasting or pattern recognition algorithm, having self-organizing abilities can be used as an active neuron in twice-multilayered neural network (TMNN), which has self-organizing abilities too. Theoretically, external supplement given by a man is displaced by supplement taken from an external error criterion, used one time only in active neurons [6]. Further self-organization process can be realized by any internal or external criterion.

GMDH algorithms can be used as active neurons defining type of problem solved [1]. Special Group Method of Data Handling (GMDH) threshold algorithm is developed, which has polynomial complexity. It means that computation time is proportional to number of input variables. Polynomial complexity is necessary to meet challenge of nowadays increase of number of input variables in data samples. To reach polynomial complexity of GMDH algorithms [3] input variables should be preliminary evaluated on their efficiency on information level. Inductive sorting of models-candidates begins from the most efficient input variables. Complexity of models is steadily increased until minimum of external error criterion will be reached.

Three rules of confluence, consequence control and excluding of non-efficient variables are used for threshold analysis of variables efficiency. They help to take into account D.Gabor's "freedom of choice" principle [4].

Dimension of experimental data samples becomes to be larger and larger. There are two ways to meet this challenge:

1. First way is preprocessing of data samples by dividing them to clusters. Coordinates of clusters centers gives new short data sample. To preprocessing operations should be included also calculation of correlation coefficients between input variables. According to rules used in meteorology, strongly correlated input variables can be summarized into one common input variable.
2. Second way is to develop new algorithms with polynomial complexity where computation time is proportional to number of input variables[5]:

$$t = t_0 + M \cdot t_1,$$

where t_0 - small constant time; M - number of primary and secondary input variables (features), t_1 - time for computing model with one input variable.

Below we shall consider second way of essential computation time decrease. Nowadays GMDH algorithms have exponential complexity. For example, for Combinatorial algorithm computing time increases in two times when one new variable is added to data sample.

The GMDH Modelling Approach

The GMDH is self-organizing approach based on sorting-out of gradually complicated models and evaluation of them by external criterion on separate part of data sample. As input variables can be used any parameters, which can influence on the process. Computer is found structure of model and measures of influence of parameters on the output variable itself. That model is better that leads to the minimal value of external criterion. This inductive approach is different from commonly used deductive techniques or neural networks.

The GMDH was developed for complex systems modelling, prediction, identification and approximation of multivariate processes, decision support after "what-if" scenario, diagnostics, pattern recognition and clusterization of data sample. It was proved, that for inaccurate, noisy or small data can be found best optimal simplified model,

accuracy of which is higher and structure is simpler than structure of usual full physical model.

Method is based on the sorting-out procedure, i.e. consequent testing of models, chosen from set of models-candidates in accordance with the given criterion. Most of GMDH algorithms use the polynomial reference functions.

Components of the input vector X can be independent variables, functional forms or finite difference terms. Other non-linear reference functions, such as difference, logistic, harmonic can also be used for model construction. The method allows to find simultaneously the structure of model and the dependence of modelled system output on the values of most significant inputs of the system.

The GMDH theory using several algorithms solve the problems of:

- long-term forecasting;
- short-term forecasting of processes and events;
- identification of physical regularities;
- approximation of multivariate processes;
- physical fields extrapolation;
- data samplings clusterization;
- pattern recognition in the case of continuous-valued or discrete variables;
- diagnostics and recognition by probabilistic sorting-out algorithms;
- vector process normative forecasting;
- modeless processes forecasting using analogues complexing;
- self-organization of twice-multilayered neuronet with active neurons.

In [6] were obtained the theoretical grounds of GMDH effectiveness as adequate method of robust forecasting models construction. Essence of it consists of automatically generation of models in given class by sequential selection of the best of them by criteria, which implicitly by sample dividing take into account the level of indeterminacy.

Since 1967 a big number of GMDH technique implementations for modelling of economic, ecological, environmental, medical, physical and military objects were done in several countries. The special peculiarities of GMDH are following:

1. External supplement: Following S.Beer work, only the external criteria, calculated on new independent information, can produce the minimum of sorting-out characteristic. Because of this data sampling is divided into parts for model construction and evaluation.
2. Freedom of choice: Following D.Gabor work, in multilayered GMDH algorithms are to be conveyed from one layer to the next layer not one, but F best results of choice to provide "freedom of choice";
3. The rule of layers complication: Partial descriptions (forms of a mathematical description for iteration) should be simple, without quadratic members in them;
4. Additional model definition: In cases, when the choice of optimal physical model is difficult, because of noise level or oscillations of criterion minima characteristic, auxiliary discriminating criterion is used;
5. All algorithms have multilayered structure and parallel computing can be implemented for their realization;
6. All questions that arise about type of algorithm, criterion, variables set etc. should be solved by minimal value of external criterion.

The main criteria used are: cross-validation PRR(s), regularity AR(s) and balance of variables criterion BL(s).

Difference of the GMDH algorithms from another algorithms of structural identification, genetic and best regression selection algorithms consists of three main peculiarities:

1. Usage of external criteria, which are based on data sample dividing and are adequate to problem of forecasting models construction, by decreasing of requirements to volume of initial information;
2. Much more diversity of structure generators: usage like in regression algorithms of the ways of full or reduced sorting of structure variants and of original multilayered (iteration) procedures;
3. Better level of automatization: to receive optimal model there is needed to enter initial data sample and type of external criterion only;
4. Automatic adaptation of optimal model complexity and external criteria to level of noises or statistical violations – effect of noiseimmunity cause robustness of the approach;
5. Implementation of principle of inconclusive decisions in process of gradual models complication.

The Combinatorial GMDH algorithm

As the input data sample is considered a matrix containing N levels (points) of observations over a set of M variables. The sample is divided into two parts. Before dividing, points are ranged by variation value. Approximately two-thirds of points make up the learning subsample N_A, and the remaining one-third of points (e.g. every third point) with same variance form the check subsample N_B. The learning sample is used to derive estimates for the coefficients of the polynomial, and the check subsample is used to choose the structure of the optimal model, that is, one for which the external regularity criterion AR(s) takes on a minimal value:

$$AR(s) = \frac{1}{N_B} \sum_{i=1}^{N_B} (y_i - y_i(B))^2 \rightarrow \min(2)$$

or better to use the cross-validation criterion PRR(s):

$$PRR(s) = \frac{1}{N}\sum_{1}^{N}\left[y_i - y_i(B)\right]^2 \to \min$$

$$N_A = N-1;\ N_B = 1.$$

To obtain a smooth exhaustive-search curve, which would permit one to formulate the exhaustive-search termination rule, the exhaustive search is performed on models classed into groups of an equal complexity. For example, the first layer can use the information contained in every column of the sample; that is full search is applied to all possible models of the form:

$$y = a_0 + a_1 x_i, \quad i = 1,2,..M. \qquad (3)$$

Non-linear members can be taken as new input variables in data sampling. The output variable is specified there in advance by the experimenter. At next layer are sorted all models of the form:

$$y = a_0 + a_1 x_i + a_2 x_j, \quad j = 1,2,..M. \quad (4)$$

The models are evaluated for compliance with the criterion, and so on until the criterion value decrease. For limitation of calculation time recently it was proposed during full sorting of models to range variables according to criterion value after some time of calculation or after some layers of iteration. Then full sorting procedure continues for selected set of best variables till the minimal value of criterion will be found. This gives possibility to set much more input variables at input and to save effective variables between layers to found optimal model.

A salient feature of the GMDH algorithms is that, when they are presented continuous or noisy input data, they will yield as optimal some simplified non-physical model. If is only in the case of discrete or exact data that the exhaustive search for compliance with the precision criterion will yield what is called a physical model, the simplest of all unbiased models. With noisy and continuous input data, simplified (Shannon) models prove more precise [6] in approximation and for forecasting tasks.

The well-known problems of an optimal (subjective) choice of the neural network architecture are solved in the GMDH algorithms by means of an adaptive synthesis (objective choice) of the architecture. There are to estimate networks of the right size with a structure evolved during the estimation process to provide a parsimonious model for the particular desired function. Such algorithms combining the best features of neural nets and statistical techniques in a powerful way discover the entire model structure - in the form of a network of polynomial functions, difference equations and other. Models are selected automatically based on their ability to solve the task (approximation, identification, prediction, and classification).

Threshold GMDH algorithm for each single active neuron

To receive polynomial complexity second auxiliary criterion in selection of optimal model is introduced. In ordinary GMDH algorithms all the problems are solved by search of minimum of external error criterion taking into account D.Gabor principle of freedom of choice [4]. External criterion is used for choice of the most effective input variables and their sets, and also for choice of optimal model structure. We can receive polynomial complexity of sorting-out selection if we apply for choice of variables separate criterion, for example module of correlation coefficient between output and input variables under evaluation. Ranked on efficiency, series of input variables, received by two different criteria can be different too, but taking into account principle of freedom of choice eliminates this difference. Such way we can found correct evaluation of input variables efficiency and their sets.

Example: suppose we have received following seria of efficiency indexes for seven variables:

$$MCC(v_1) = 0.9; \qquad MCC(v_2) = 0.75;$$
$$MCC(v_3) = 0.6; \qquad MCC(v_4) = 0.45;$$
$$MCC(v_5) = 0.4; \qquad MCC(v_6) = 0.25;$$
$$MCC(v_7) = 0.1$$

Using linear polynomial reference function we can organize following set of models-candidates:

$$S = 1 \quad y = a_0;$$
$$S = 2 \quad y = a_0 + a_1 v_1;$$
$$S = 3 \quad y = a_0 + a_1 v_1 + a_2 v_2;$$
$$S = 4 \quad y = a_0 + a_1 v_1 + a_2 v_2 + a_3 v_3;$$
$$S = 5 \quad y = a_0 + a_1 v_1 + a_2 v_2 + a_3 v_3 + a_4 v_4;$$
$$S = 6 \quad y = a_0 + a_1 v_1 + a_2 v_2 + a_3 v_3 + a_4 v_4 + a_5 v_5;$$
$$S = 7 \quad y = a_0 + a_1 v_1 + a_2 v_2 + a_3 v_3 + a_4 v_4 + a_5 v_5 + a_6 v_6;$$
$$S = 8 \quad y = a_0 + a_1 v_1 + a_2 v_2 + a_3 v_3 + a_4 v_4 + a_5 v_5 + a_6 v_6 + a_7 v_7.$$

Only this set of models-candidates should be checked by external error criterion if the principle of freedom of choice is not taken into account to select model of optimal complexity. For this principle we should change the order of variables in the ranked on efficiency seria of variables according to efficiency criterion. Especially should be tried to change each pair of variables, which are almost equal on efficiency. In example, we have only one such pair of variables: v_4 and v_5. Difference in indexes of efficiency

is equal to 0.05. This is reason to change variables in equations where variables v_4 and v_5 take part. We receive additional following model-candidate:

$$S = 5 \qquad y = a_0 + a_1 v_1 + a_2 v_2 + a_3 v_3 + a_4 v_5 .$$

Instead of eight we receive nine models-candidates. Each pair of input variables, almost equal on efficiency, leads to one additional model-candidate. Similar operation should be organized for each pair of variables, almost equal on efficiency.

In the case when the number of input variables are comparatively small instead of consequence control rule described above, can be used the rule of excluding of non-effective variables. For example we can exclude from our consideration all the variables for which the module of correlation coefficient is smaller than definite threshold value. Then ordinary GMDH Combinatorial algorithm can be used.

Self-organization of twice-multilayered neural network by GMDH threshold type algorithm

Twice-multilayered neuronets with active neurons realize twice-multilayered structure: neurons are multilayered and they are connected into multilayered structure. This gives possibility to optimize the set of input variables at each layer, while the accuracy increases.

Not only GMDH algorithms, but many modelling or pattern recognition algorithms can be used as active neurons. Its accuracy can be increased in two ways:

- each output of algorithm (active neuron) generate new variable which can be used as a new factor in next layers of neuronet;
- the set of input factors can be optimized at each layer. In usual once-multilayered NN the set of input variables can be chosen once only. The output variables of previous layers in such networks are very effective secondary inputs for the neurons of next layers.

This corresponds to the actions of human nervous system, where the connections between several neurons are not fixed but change depending on the neurons themselves. GMDH algorithms are examples of complex active neurons, because they choose the effective inputs and corresponding coefficients of them by themselves, in process of algorithm self-organization. The problem of neuronet links structure self-organization is solved in a simple way.

Neuronet can be described as matrix, which unites active neurons in several layers. The neurons of each layer differs one from another by their output and input sets of variables (Fig.1). The output variables of each layer of active neurons are used as the input variables for next layer. Extension of regression area always can only perfect the result of regression. In the neuronet, considered below, extension is realized by very special way. For example, if the first layer of active neurons obtains set of input variables $x_1,x_2,...,x_M$ and generates the set of output variables $y_1,y_2,...,y_L$, then the neurons of second layer obtains the both set of variables on its input. Extension of variables set always is accompanied by reasonable narrowing of variables number to prevent the exceed of computer ability (for allowed computer calculation time).

Programmed by GMDH algorithm module acts similar to a widely known Kalman filter. Output variables repeat values of input variables but with less dispersion of noise. This is reason to use GMDH algorithms as neurons of neural network. This can be called as active neurons because they can find the most effective connections themselves. Number of active neurons in the first layer of neural network is equal to number of input variables, given in experimental data sample. Action of each active neuron is evaluated by usual internal error criterion. Second layer of active neurons should be organized for that variables which are more accurately measured then corresponding variables in input data sample.

Some variables do not need filtration of noises, but some variables measured with noise need organization of next neuronet layers. The number of layers of neuronet is steadily increased, until the error of each variable approximation decreases. So, for different variables are to be used different number of neuronet layers.

The task for self-organization of such networks of active neurons by selection is to estimate the number of layers of active neurons and the set of possible potential inputs and outputs of every neuron. The sorting characteristic - ”number of neuronet layers - variables, given in data sample“ - defines the optimum number of layers for each variable separately. During learning active neurons self-organize the structure of the entire neural network.

At present, inductive GMDH algorithms, used as active neurons, give us the way to get accurate identification and forecasts of different complex processes in the case of noised and short input sampling. In distinction to neural networks, the results are explicit mathematical models, obtained in a relative short time. Neural nets with active neurons should be applied to rise up accuracy of complex objects modelling algorithms.

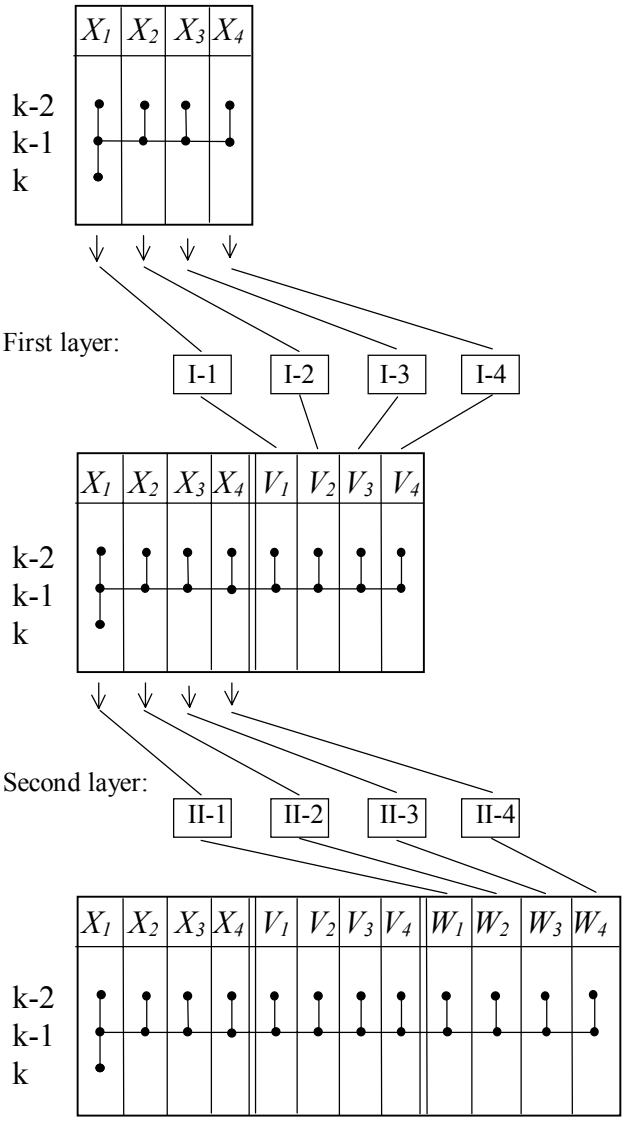

Fig 1. Schematic arrangement of the first two rows of a neural network.

References

1 Ivakhnenko,A.G., Ivakhnenko,G.A. and Müller,J.A., Self-Organization of Neural Networks with Active Neurons. Pattern Recognition and Image Analysis, 1994, vol.4, no.2, pp.185-196.

2 Ivakhnenko A.G., Bogachenko N.N., Li Tian Min Stages of Optimization of Random Processes Forecasting by Analogues Complexing Algorithm J. of Automation and Information Sciences, no.4, 1997, pp.111-118.

3 Madala H.R., Ivakhnenko A.G. Inductive Learning Algorithms for Complex Systems Modeling, CRC Press Inc., Boca Raton, 1994, p.384.

4 Gabor D. Perspectives of Planing. 1) Organization of Economic Cooperation and Development. Imperial College of Science and Technology. London, 1971. 2) Sov.Autom.Control, vol.5, 2, 1972.

5 Lawler E.L. Sequencing jobs to minimize total weighted completion time subject to precedence constrains, Ann. Discrete Math., 1978.

6 Aksenova,T.I. and Yurachkovsky,Yu.P. A Characterisation at Unbiased Structure and Conditions of Their J-Optimality, Sov. J. of Automation and Information Sciences, 21, no.4, 1988, pp.36-42.