Problems of Further Development of the Group Method of Data Handling Algorithms. Part I

A. G. Ivakhnenko* and G. A. Ivakhnenko**

**Glushkov Cybernetics Institute, National Academy of Sciences of Ukraine, pr. Akademika Glushkova 40, Kiev, 252207 Ukraine*

***National Institute for Strategic Studies, ul. Pirogova 7a, Kiev, 252030 Ukraine*
e-mail: gai@gmdh.kiev.ua; gai@niss.gov.ua

Abstract—The GMDH algorithms for solving interpolation problems of artificial intelligence differ from each other in the form of the reference function and the iteration rules of the multilayer model structure. In some multilayer algorithms, the number of terms in the iteration rule is constant, which leads to the skipping of some models. In the algorithm called combinatorial, the iteration rule increases by one term when passing to each next row, which ensures an exhaustive search through all of the equations. For exact and complete data, the minimum of the external criterion is nonsharp, and to determine an optimal method, extrapolation of the locus of points of the minimum of the external criterion should be performed. A comparison of linear, polynomial, and ratio-polynomial (with respect to the coefficients) functions may give a method for improving the accuracy of problem solutions. To reduce computational time, a threshold GMDH algorithm is developed which preliminarily estimates the effectiveness of the input variables at the information level and searches for model-candidates based on the most effective input variables (arguments or features).

1. PREPROCESSING OF THE INITIAL DATA SAMPLE

It is expedient to accept the following procedure for preprocessing large samples of initial data.

(1) The "wild points" (the values of variables are obviously impossible) are removed and replaced by the tripled mean deviation.

(2) The mean value of the variables is calculated after the wild points are removed.

(3) The missing values in the sample are replaced by the mean value.

(4) The quantitative variables are normalized to fit into the range between 0 and 1.

(5) Each qualitative variable is assigned the value 0 or 1, depending on its class.

If it is required to recognize several patterns, the sample is divided into several subsamples, and the patterns or classes are recognized pairwise.

2. ACHIEVEMENT OF THE MODEL UNIQUENESS AND ESTIMATION OF ITS REPRESENTATIVENESS

In recursive search modeling methods, the initial data are represented in the sample of experimental data as a table with N rows, which are called observations, or realizations, or images (in pattern recognition). The model sought for is an equation that expresses the value of the output variable through the current and past (i.e., retarded) values of the input variables. In the GMDH algorithms, the model is sought in the form of a linear polynomial. The complexity of the model structure is determined by the number of terms in the polynomial.

The optimal model (most accurate under the preset noise level) corresponds to a minimum of the external accuracy criterion.

We use the term *specific complexity* for the ratio of the number of terms in the polynomial S describing the model to the number $M + 1$ of the input variables (primary and secondary features) increased by one (to take into account the presence of the free term).

The graphically represented dependence of the external criterion on the specific complexity of the model answers all questions about the number of minima in the model and the sufficiency or representativeness of the sample. The number of models that can be obtained based on a given sample equals the number of sufficiently sharp minima of the dependence. A sample is considered sufficient or representative if it gives only one sharp minimum. The uniqueness of the minimum can be achieved by dividing the sample into parts according to the clusters of the optimal physical clustering. The resulting sample must contain only the rows included in the first cluster (in decreasing order of the number of points).

For the second model, the points of the second cluster are used, etc. The model is considered most representative if it gives the absolute minimum of the external criterion.

Received December 27, 1999

Pattern Recognition and Image Analysis, Vol. 10, No. 2, 2000, pp. 187–194.

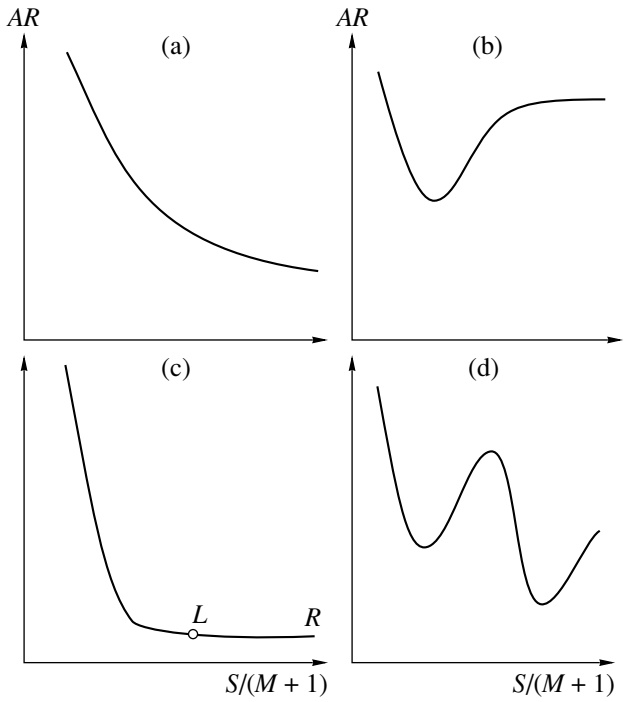


Fig. 1. Various forms of dependence of an external criterion *AR* on the model complexity: (a) the dependence has no minimum; (b) the data are incomplete or inaccurate; (c) the dependence has an uncertainty zone; (d) the sample contains a mixture of data corresponding to two models.

Figure 1 shows the typical forms of a characteristic that expresses the dependence of the external accuracy criterion on the specific complexity of the model. The form of the characteristic allows us to judge the initial data sample. In the case shown in Fig. 1a, we can conclude that the sample contains no sufficiently effective variables. Figure 1b corresponds to a large noise variance or incomplete data. Figure 1c shows the formation of an uncertainty zone in the region of exact data, which impedes the choice of a unique nonphysical optimal model.

Finally, Fig. 1d shows a sample containing data corresponding to two models. In this case, we can obtain more accurate models by dividing the sample into two parts with the use of clustering.

3. PROBLEMS RELATED TO THE UNCERTAINTY ZONE IN THE DOMAIN OF EXACT AND COMPLETE DATA

Interpolation problems of artificial intelligence, such as the problem of predicting random processes, dependence detection, pattern recognition, etc., can be solved by either constructing mathematical models or searching for analogs in prehistory. The deductive logical modeling methods only apply to comparatively simple cases, where the mechanism of operation of the object is clear and the set of arguments is known. These methods give a complete physical model of the object, which is only optimal if the set of the input exactly measured variables is complete. For an unknown set of input variables measured with errors, more accurate results are given by the simplified nonphysical model obtained with the use of the recursive search Group Method of Data Handling (GMDH) [1–3].

The GMDH is based on sorting models of gradually increasing complexity and estimating them according to an external criterion on an independent data subsample. As the input variables, any parameters that may affect the process can be used. A computer automatically determines the structure of the model and the degree of influence of the parameters on the output value. The model corresponding to the minimum value of the external criterion is considered the best.

The GMDH has been developed for complex system modeling, prediction, identification and approximation of multifactor systems, diagnostics, pattern recognition, and data sample clustering. It is proved analytically that only this recursive method of self-organization gives an optimal nonphysical model whose accuracy is higher and whose structure is simpler than the structure of the usual complete physical model for inaccurate, noisy, or short data samples.

The recent GMDH developments led to the creation of decision support systems based on normative prediction (according to an "if–then" scenario) and control optimization with the use of simplified linear programing algorithms and neural networks with active neurons. These neural networks implement the doubly multilayer structure: neurons with multilayer structures are gathered into a multilayer network. This makes it possible to optimize the set of input parameters at each level while the accuracy increases. The accuracy of prediction, approximation, or pattern recognition can be enhanced beyond the values reached by the usual neural networks with simple neurons or by the usual statistical methods. Very accurate predictions of the New York stock exchange and other complex objects were obtained with the use of a doubly multilayer neural network in which every neuron was operated by the GMDH algorithm.

In the combinatorial GMDH algorithm, the number of terms in the polynomial model gradually increases, while an external accuracy criterion decreases; the external criterion is calculated based on an independent material, i.e., on a separate test data subsample. The dependences of only external criteria on the model complexity have minima. The internal criteria, which are calculated from the same data from which the coefficients of models are estimated, result in increasing the accuracy of the model with its complexity; this leads to the choice of an overcomplicated nonoptimal model. As an external accuracy criterion, either the regularity criterion or the cross-validation criterion [4] is recommended.

An application of the combinatorial GMDH algorithm may involve two difficulties. First, the criterion minimum becomes nonunique for exact experimental data, i.e., under low noise. In Fig. 2, an entire uncertainty zone (the *LOR* interval) appears instead of a point of minimum. The optimal physical model then corresponds to the midpoint of the uncertainty zone. However, for very accurate and complete data, when the coordinate of the point *R* is one, an optimal simplified nonphysical model can only be found by searching for a model according to some external or internal accuracy criterion in the uncertainty zone. It is sufficient to apply the search according to the external criterion only once to find a criterion minimum, which determines an optimal nonphysical model. The physical complete model (more precisely, its polynomial approximation) can be found by extrapolation of the line that is the locus of the points of minimum (LPM) of characteristics constructed for various noise variances. Under certain conditions, the LPM line passes through the point *P* with coordinates $1/(M + 1)$, 1 and through the point of minimum obtained based on the available data. Extrapolation of the LPM up to its intersection point with the abscissa axis specifies the structure of the sought physical model, which should be applied under exact input data. For noisy and incomplete data, the nonphysical model corresponding to the minimum of the external accuracy criterion remains optimal.

Solving pattern recognition problems does not require extrapolation of the LPM, because the optimal model or the discriminant function is usually supposed to be used in an environment with the same noise level under which the data sample is obtained. For this reason, in pattern recognition, nonphysical models corresponding to the minima of external accuracy criteria are applied. The higher the noise, the simpler the nonphysical recognizing model; under exact input data, the physical and nonphysical models coincide.

Predicting random processes also requires constructing a nonphysical model. The physical model is largely needed to logically interpret the nonphysical model and reveal the mechanism of operation of the object used to obtain the data sample.

4. ACHIEVEMENT OF THE POLYNOMIAL COMPLEXITY OF THE ALGORITHM

Another difficulty involved in applying the combinatorial algorithm is that the algorithm has exponential computing-time complexity. Adding one new variable doubles the computing time *T*:

$$T = t_0 + t_1 2^M,$$

where t_0 and t_1 are time constants, and *M* is the number of features.

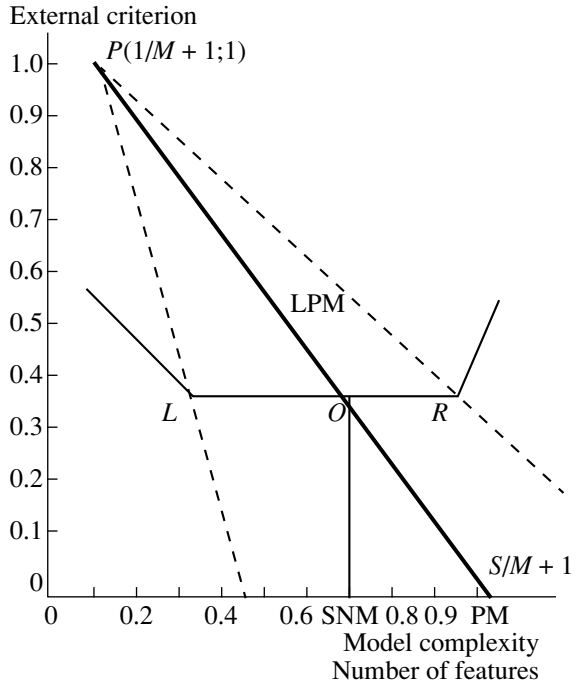


Fig. 2. The dependence of the external accuracy criterion on the complexity of the model structure for the physical model in the absence of noise (PM) and a simplified nonphysical model for the same noise variance as in the initial data sample (SNM); LPM is the locus of the points of minimum of the external accuracy criterion.

For polynomial algorithms, the computing time *T* is proportional to the number of features to be taken into account [6]:

$$T = t_0 + t_1 M.$$

To achieve polynomial complexity in the new threshold GMDH algorithm, we introduce a second auxiliary criterion for the effectiveness of the input variables, which are called predictors or features. For binary features, the criterion of the number of resolved disagreements [6] is recommended.

The effectiveness of continuous features is estimated based on the absolute value of the correlation coefficient of the output variable and the feature to be estimated [7]. If the variables have different dimensions, then they are normalized to fit into the range between zero and one.

The primary features are those included in the data sample. As is shown in Fig. 3, in different interpolation problems, different primary features are used. The least number of features is required to detect dependences and recognize patterns. The values of the primary features are used to generate values of additional secondary features. The secondary features to start with are the coordinates of the two first analogs. The analog of a row in a sample is the nearest row in the same sample.

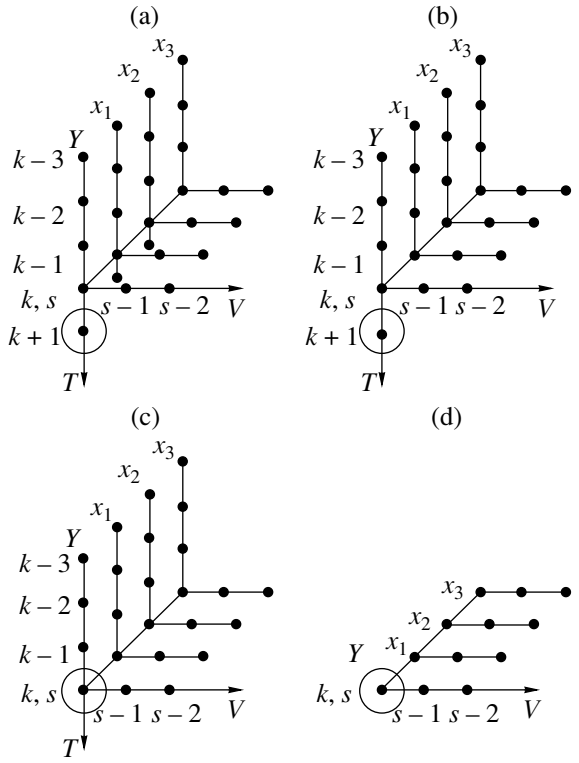


Fig. 3. Discrete templates of GMDH filters for (a) the implicit predicting filter, (b) the explicit predicting filter, (c) the smoothing filter, and (d) the filter of dependence detection and pattern recognition; T is the time axis and V is the axis corresponding to the virtual process of image formation.

Analogs can be considered retarded arguments of some virtual process of row formation in the data sample [8]. As secondary features, the covariances (pairwise products of normalized values) of primary features can also be used [7]. All primary and secondary features together are ranked according to effectiveness. The ordered set of features is subjected to a threshold analysis according to the following rules:

(1) *The rule of merging two features.* Two neighboring features are replaces by one average feature if they have close effectivenesses that differ by no more than a certain threshold value *E1*. This rule follows from the metrology law about averaging close values of two measurements.

(2) *The rule of changing the order of features.* If the difference of the effectivenesses of two neighboring features is larger than *E1* but smaller than another threshold *E2*, then, in addition to the ordered set under analysis, another sequence of features is formed, in which the close features are exchanged. Testing the two orderings of the features implements Gabor's principle of freedom of choice. In practice, this rule reduces to estimating one more additional model [3].

As it is known, in a single GMDH algorithm without the use of a network, the number of variables over which the external criterion search is performed is either constant (if the search is exhaustive) or reduces when passing from one row to another at the expense of eliminating noneffective input variables. The distinguishing feature of the threshold GMDH algorithm is that the number of search variables increases by one when passing to each next row.

In the first row, only the output variable is used, and its mean value is found. In the second search row, one most effective variable is used. In the second row, the search involves two variables, etc. If the effectiveness of the added search variable is close to the effectiveness of the preceding variable, then one more equation of a model-candidate is formed. The search continues until the minimum of the external criterion is attained.

5. VERIFICATION OF THE NECESSITY OF AN ORTHOGONAL DATA SAMPLE

If the sequence of features ranked according to effectiveness contains regions of accumulation, i.e., the features are distributed nonuniformly, then it is recommended to preliminarily orthogonalize the sample by the Karhunen–Loéve algorithm [9].

Example. Suppose that a sample of experimental data contains values of seven input variables, V_1, V_2, V_3, V_4, V_5, V_6, and V_7 (the number of features is $M = 7$). Suppose also that computation of the correlation coefficient between the features and the output variable yields the following sequence of features ranked according to effectiveness:

$$E(V_1) = 0.9 \quad E(V_2) = 0.75 \quad E(V_3) = 0.6$$
$$E(V_4) = 0.45 \quad E(V_5) = 0.4$$
$$E(V_6) = 0.23 \quad E(V_7) = 0.1$$

Let us set the following threshold values: the threshold value for the merging of features is $E = 0.01$, and the threshold value for the change of order is $e = 0.1$. The results of an analysis of the series of features ranked by effectiveness are as follows.

(1) No two features must be merged.

(2) We must try to change the order of only one pair of features, V_4 and V_5.

(3) The features are uniformly distributed over the effectiveness interval, and there is no need to apply preliminary orthogonalization.

As a reference basis function, we choose a linear polynomial. Gradually decreasing the effectiveness threshold, we obtain the following set of models-candidates:

$$S = 1 \quad y = a_{01};$$

$$S = 2 \quad y = a_{02} + a_{12}V_1;$$

$$S = 3 \quad y = a_{03} + a_{13}V_1 + a_{23}V_2;$$

$$S = 4 \quad y = a_{04} + a_{14}V_1 + a_{24}V_2 + a_{34}V_3;$$

$$S = 5 \quad y = a_{05} + a_{15}V_1 + a_{25}V_2 + a_{35}V_3 + a_{45}V_4;$$

$$S = 5 \quad y = b_{05} + b_{15}V_1 + b_{25}V_2 + b_{35}V_3 + b_{45}V_4;$$

$$S = 6 \quad y = a_{06} + a_{16}V_1 + a_{26}V_2 + a_{36}V_3 + a_{46}V_4 + a_{65}V_5;$$

$$S = 7 \quad y = a_{07} + a_{17}V_1 + a_{27}V_2 + a_{37}V_3 + a_{47}V_4 + a_{57}V_5 + a_{67}V_6;$$

$$S = 8 \quad y = a_{08} + a_{18}V_1 + a_{28}V_2 + a_{38}V_3 + a_{48}V_4 + a_{58}V_5 + a_{68}V_6 + a_{78}V_7.$$

For the complexity $S = 5$, we obtain two models-candidates; the external criterion determines which should be included in the extremal characteristic shown in Fig. 2. We can stop increasing the complexity of model equations as soon as the coordinates of the point O of the minimum of the external criterion are found.

Calculation of the external criterion makes it possible to construct its dependence on the complexity of the model structure (see Fig. 2). If the problem is to recognize two patterns, then the optimal model obtained is the best discriminant function [7].

If the number of input variables is not too large, then, instead of the rule of order verification, the simpler and less rigid rule of noneffective variable elimination is used. For example, all input variables with a correlation coefficient less than 0.2 in absolute value are disregarded. This drastically reduces the amount of computations involved in the usual combinatorial GDMH algorithm, especially if the search through model-candidates terminates as soon as the minimum of the external accuracy criterion is attained.

A calculation of a sequence of initial variables ranked according to the effectiveness criterion, threshold analysis of this sequence, and organization of a search for the structure of a model-candidate optimizing the external accuracy criterion solve the problem of constructing a recursive GMDH algorithm with polynomial complexity.

The algorithm for estimating feature effectiveness has polynomial complexity, because its computational time is proportional to the number of features. The threshold analysis of the sequence of features ranked by effectiveness does not require much time. The construction of models-candidates and the estimation of their coefficients also have polynomial complexity due to the application of the bordering procedure [1, 2]. The coefficients of each model are calculated based on the estimates of the coefficients of models of lower complexity.

The computer programs implementing the bordering procedure are designed for both the inclusion and the removal of a term of the polynomial model [1, 2]. The bordering procedure substantially reduces computational time in searching GMDH algorithms. But estimating the feature effectiveness according to the absolute value of the correlation coefficient reduces the computational time even further. The ordering of features determined by this absolute value is subsequently verified by the method of threshold analysis of the ranked sequence of features based on the external accuracy criterion. Thus, the threshold algorithm of model self-organization as a whole has polynomial complexity, which is confirmed by experimental computations.

6. THE THRESHOLD GMDH ALGORITHM WITH SELECTION OF THE SET OF EFFECTIVE VARIABLES

The optimal threshold value separating effective and noneffective variables can be obtained by repeatedly applying the combinatorial GMDH algorithm while the number of the effective variables subject to search successively increases. In the example considered above, the combinatorial algorithm should be applied to data samples containing the following input variables:

$$S = 1 \quad Y;$$

$$S = 2 \quad YV_1;$$

$$S = 3 \quad YV_1V_2;$$

$$\ldots\ldots\ldots\ldots\ldots$$

$$S = 8 \quad YV_1V_2\ldots V_7.$$

Increasing the number of the effective variables that form the input of the algorithm terminates as soon as the minimum of the external accuracy criterion is attained.

The described algorithm is often applied instead of the verification of the order of variables, especially when the choice of the threshold value for the order verification is impeded.

Complication of the ratio-polynomial reference function

Complexity	Number of variables	Reference function	Number of modifications
$S = 1$	$M = 0$	$y = a_0$	1
$S = 2$	$M = 1$	$y = (a_0 + a_1x_1)/1$	2
		$y = a_0/(1 + b_1x_1)$	
$S = 3$	$M = 2$	$y = (a_0 + a_1x_1 + a_2x_2)/1$	
		$y = (a_0 + a_1x_1)/(1 + b_2x_2)$	4
		$y = (a_0 + a_1x_2)/(1 + b_2x_1)$	
		$y = a_0/(1 + b_1x_1 + b_2x_2)$	
$S = 4$	$M = 3$	$y = (a_0 + a_1x_1 + a_2x_2 + a_3x_3)/1$	
		$y = (a_0 + a_1x_1 + a_2x_2)/(1 + b_1x_3)$	
		$y = (a_0 + a_1x_1 + a_2x_3)/(1 + b_1x_2)$	
		$y = (a_0 + a_1x_2 + a_2x_3)/(1 + b_1x_1)$	
		$y = (a_0 + a_1x_1)/(1 + b_2x_1 + b_2x_3)$	
		$y = (a_0 + a_1x_2)/(1 + b_2x + b_2x_2)$	
		$y = (a_0 + a_1x_3)/(1 + b_1x_1 + b_2x_2)$	
		$y = a_0/(1 + b_1x_1 + b_2x_2 + b_3x_3)$	

7. THE APPLICATION OF THE RATIO-POLYNOMIAL REFERENCE FUNCTION TO SOLVE INTERPOLATION PROBLEMS OF ARTIFICIAL INTELLIGENCE

Problems of process prediction, dependence detection, and pattern recognition are usually solved with the use of a polynomial reference (or support) function, linear with respect to the coefficients. Both primary and secondary features are used as arguments; the secondary features are expressed by simple nonlinear dependencies via the primary features.

For essentially nonlinear objects, the solution accuracy can be improved by extending the search by forming ratio-polynomial functions. The table shows an example of such an extension performed under the following two constraints: (i) each argument is included in the ratio-polynomial function only once and (ii) all coefficients are positive.

These constraints are not vital and can be removed, but this would increase the amount of the model-candidates to be searched through to optimize the external accuracy criterion, which is already fairly large. For this reason, we only consider the constrained extension of search domains, as in the table.

8. A METHOD FOR SUBSTANTIALLY REDUCING THE SEARCH

In the table, only ratio-polynomial functions with positive coefficients are to be searched through. For such functions, the output variable is related to the variables in the numerator directly and to the variables in the denominator reciprocally. Using this observation, we can apply the following method to determine an optimal ratio-polynomial model. First, an optimal polynomial model for the data sample under consideration is determined. The variables whose coefficients in this model are positive are included in the numerator of the ratio-polynomial model, and the variables with negative coefficients are included in its denominator.

Numerical example. The problem of recognizing a pairwise dependence of neurons is thoroughly described in [10]. Here, we give the data of the numerical example from [10]. For a polynomial reference function, the recognition accuracy is characterized by

$$R^2 = I - RR^2 = 0.8,$$

where R^2 is the squared multiple determination coefficient [10].

For a ratio-polynomial reference function, the recognition accuracy is much higher:

$$RR^2 = 0.9.$$

With the use of the method described above, the computational time almost does not change:

$$T = 10 \text{ min}.$$

9. A GENERAL DESCRIPTION OF A NONLINEAR GMDH ALGORITHM WITH A RATIO-POLYNOMIAL REFERENCE FUNCTION

The recursive search approach to modeling, which is also called the self-organization of models, performs all tasks except those related to the choice and computation of a criterion by searching, i.e., testing a large number of possible solutions by an external criterion, which is calculated on a separate part of the data sample. The less the value of the external accuracy criterion, the better the corresponding model or algorithm.

An important role is also played by Gabor's choice principle [3]. The GMDH algorithms search through a large number of candidates to solve the following two problems.

(1) The effectiveness of the primary and secondary features and the sets of two, three, and more features are estimated. The freedom-of-choice principle requires that the sets to be estimated be formed not only by the most effective features selected in the preceding row, but also by the next most effective features and sets.

(2) The models optimizing the external criterion in every row are compared to each other, and an optimal model is selected.

10. A METHOD FOR ESTIMATING THE COEFFICIENTS OF RATIO-POLYNOMIAL MODELS

To estimate the coefficients of the models, we transform all ratio-polynomial functions into auxiliary polynomials and include additional columns of digits into the data sample for all of the terms of these polynomials. Then, we apply algorithms for estimating coefficients of polynomial models.

Example. It is required to estimate the coefficients of the ratio-polynomial model

$$y = \frac{a_0 + a_1 V_1}{1 + b_1 V_2}.$$

Let us form the auxiliary polynomial

$$y = a_0 + a_1 V_1 - b_1 y V_2.$$

The variables y, V_1, and yV_2 are treated as arguments whose values are easy to calculate with the use of the data sample. The coefficients are estimated by the usual least-square method [11].

11. PREPROCESSING OF LARGE DATA SAMPLES

The first descent is performed by applying the correlation analysis to the elements in the data sample, reducing the sample to the one-moment form, and dividing it into uniform subsamples used to obtain models by GMDH algorithms.

Reduction to the one-moment form involves an increase in the number of arguments. To each sample row, which contains the current variable values indexed by k, we add several future values indexed by $k + 1$, $k + 2$, etc., and retarded arguments indexed by $k - 1$, $k - 2$, etc. This allows us to arbitrarily change the order of rows, similar to the problem of pattern recognition.

The correlation analysis of sample elements is based on the idea of the genetic approach to modeling. The combinatorial GMDH algorithm, where the iteration rule increases by one term when passing to each next row, is very similar to the genetic optimization of the choice of the number of genders of living organisms during the evolution of species in Nature, which took millions of years.

Another, multilayer GMDH algorithm [1, 2] with an invariable rule of iteration resembles the actions of a selectionist who obtains the required features in several generations. Thus, all GMDH algorithms have genetic analogs. However, only those of them should be called genetic in which the number of models-candidates subject to search is reduced by taking into account the generic (correlation) relations. Close arguments and models merge, and noneffective ones are disregarded. The threshold GMDH algorithm implements such a selection.

The concept of optimal physical clustering, which is implemented by the objective computer clustering (OCC) algorithm, makes it possible to divide the data sample into subsamples of uniform vectors contained in one cluster. In solving problems of random process prediction, it is recommended to form predictions for all uniform samples and select the most accurate prediction. In recognizing patterns and situations, it is also recommended to obtain discriminant functions for all uniform subsamples to select the most effective decision rule. Thus, the reduction of computational time is due to the processing of small subsamples instead of one large sample.

In processing large data samples, each element is characterized by two values, namely, of the output variable or argument and of the contribution of the element to the absolute value of the correlation coefficient between the variable under consideration and the output variable. We refer to this contribution as the effectiveness of the sample element. A rational reduction of sample size consists in the elimination of the least effective elements. The effectiveness of a sample element equals the difference of two correlation coefficients, one calculated for the argument value specified in the sample, and the other, for the element equal to the mean value of the variable. A simplified calculation of the effectiveness of an element can be performed by the formula

$$e = \frac{|(W - \overline{W})*(V - \overline{V})|}{\max((W - \overline{W})*(V - \overline{V}))},$$

where W is the value of the output variable, $\overline{W}$ is its mean value, V is the value of the argument at the sample element under consideration, and $\overline{V}$ is the mean value of the variable to which this element belongs.

The denominator equals the maximum effectiveness over all of the sample elements. Thus, the effectiveness of each element is normalized by the largest value, and we can set a certain threshold effectiveness value, say, 0.3. The rows and columns in the sample that contain only elements with effectiveness less than the threshold value are removed from the sample. The effectiveness threshold is raised until the size of the sample becomes acceptable for calculations on an available computer. The limiting parameters of models that can be computed on a PC can be expressed by the empirical equation $m \le 27 - 0.046M - 1$, where M is the number of sample arguments-candidates and m is the number of arguments in the model.

We see that, if the sample contains 350 rows, then a model with ten terms can be obtained in a reasonable time, which is sufficient for the majority of practical problems.

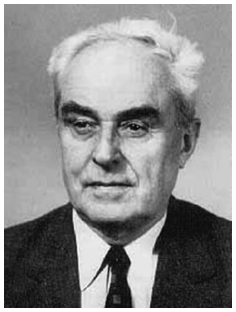
The further processing of large samples reduces to splitting them into subsamples of uniform vectors by the OCC algorithm, and the subsample that gives a most accurate prediction (in process prediction) or

most effective decision rule (in pattern recognition) is selected. If the problem is to recognize an image with a known vector, then the corresponding uniform subsample can be found by pattern recognition methods.

The accuracy of the obtained models may be improved by applying secondary arguments and neural networks with active neurons. As secondary arguments, the coordinates of the first two analogs or pairwise covariances (products of normalized values of the primary sample arguments) may be used. Samples of secondary arguments are also subject to a correlation analysis, which may enlarge the set of effective arguments and thereby improve the accuracy of the model. The same goal can be achieved by constructing a twice-multilayer neural network with active neurons (i.e., by applying implicit templates), which is considered during further descents. Effectiveness of elements in the samples of secondary arguments can be estimated based on the initial material or on the uniform subsamples. The use of uniform subsamples is preferable, because it requires a less computational time.

REFERENCES

1. Ivakhnenko, A.G. and Stepashko, V.S., *Pomekhoustoichivost' modelirovaniya* (Noise Immunity of Modeling), Kiev: Naukova Dumka, 1985.
2. Ivakhnenko, A.G. and Yurachkovskii, Yu.P., *Modelirovanie slozhnykh sistem po eksperimental'nym dannym* (Modeling of Complex Systems from Experimental Data), Moscow: Radio i Svyaz', 1986.
3. Ivakhnenko, A.G., Zaichenko, Yu.P., and Dimitrov, V.D., *Prinyatie reshenii na osnove samoorganizatsii* (Decision Making Based on Self-Organization), Moscow: Sovetskoe Radio, 1976.
4. Ivakhnenko, A.G. and Krotov, G.I., Self-Organization of Models with Variable Coefficients in Optimization of Designing Water Objects, *Sov. J. of Automat. Inform. Sci.*, 1980, vol. 13, no. 6, pp. 11–30.
5. Pavlov, O.A. and Pavlova, L.O., *PDS-algorytmy dlya vazhkovyrishuvannykh kombinatornykh zadach: Teoriya ta metodologiya rozrobky* (PDS Algorithms for Hard-to-Solve Combinatorial Problems: Theory and Methodology of Development), Uzhgorod: Polychka "Karpatc'kogo Krayu," 1998.
6. Ivakhnenko, A.G., *Samoobuchayushchiesya sistemy raspoznavaniya i avtomaticheskogo upravleniya* (Self-Learning Systems of Recognition and Automatic Control), Kiev: Naukova Dumka, 1969.
7. Krug, G.K. and Krug, O.Yu., Mathematical Method for Classification of Ancient Pottery, in *Trudy Instituta Arkheologii Akademii Nauk SSSR* (Proc. of Inst. of Archeology, Academy of Sciences of USSR), Moscow: Nauka, pp. 318–325.
8. Ivakhnenko, A.G., Ivakhnenko, G.A., Tetko, I.V., and Sarychev, A.P., Application of Analog Coordinates as Retarded Arguments of Virtual Processes of the Formation of Rows of a Data Sample, *Pattern Recognit. Image Anal.*, 1999, vol. 9, no. 3, pp. 401–407.
9. Yurachkovskii, Yu.P., Application of the Kahrunen–Loéve Expansion to Construct Scalar Convolutions of Vector Criteria (on the Example of Estimation of the Quality of Surface Waters on Land), *Sov. J. of Automat. Inform. Sci.*, 1987, vol. 20, no. 1, pp. 17–23.
10. Ivakhnenko, A.G., Ivakhnenko, G.A., Tetko, I.V., and Sarychev, A.P., Recognition of the Type of Neurons' Interaction from the Histograms of Pulse Delay of Their Activity, *Pattern Recognit. Image Anal.*, 2000, vol. 10, no. 1, pp. 163–167.
11. Ivakhnenko, A.G., *Induktivnyi metod samoorganizatsii modelei slozhnykh sistem* (A Recursive Method of Self-Organization of Models of Complex Systems), Kiev: Naukova Dumka, 1987.



Aleksei G. Ivakhnenko. Born 1913. Graduated from the Leningrad Institute of Electrical Engineering in 1938. Received doctoral degree in 1953. Author of the Group Method of Data Handling (GMDH) which is widely used in modeling. Scientific interests: inductive sorting modelling methods for forecasting random processes in fuzzy systems of ecology, biology, medicine, and economics.

Grigorii A. Ivakhnenko. Born 1966. Graduated from the Kiev Polytechnic Institute in 1989. Leading Specialist in the National Institute for Strategic Studies. Scientific interests: data mining and complex systems analysis by inductive methods, pattern recognition and clusterization. Author of 22 papers.