Algebraic Approach and Optimal Physical Clusterization in Interpolation Problems of Artificial Intelligence

A. G. Ivakhnenko*, G. A. Ivakhnenko, E. A. Savchenko*, and D. Wunsch*****

**International Educational Center of Informational Systems and Technologies, National Academy of Sciences of Ukraine,
pr. Akademika Glushkova 40, Kiev, 252207 Ukraine
e-mail: gai@gmdh.kiev.ua*

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

****Emerson Electric Company, Missouri, USA*

Abstract—Perceptron-type interpolation systems of artificial intelligence are considered. A concept of optimal physical clusterization allows us to divide a second layer of hidden units into the compact sets of units (clusters). Then, an algebraic approach developed for pattern recognition systems may be extended to other systems. To solve the problems of process forecasting, a data sample should be transformed into a single-moment sample.

INTRODUCTION

The problems, which are solved by handling experimental data sample of an object in a passive or active mode of observation, are usually considered as interpolation problems of artificial intelligence. The input data sample is processed, when we solve the problems of pattern recognition, law detection, identification of object equations, forecasting of short-term and long-term random processes in the object. Each problem has such a specific character that a separate theory of solution has been developed for each of them and the biggest success was achieved in the theory of pattern recognition. Here, the fundamental explanation of the results obtained via different algorithms is given by an algebraic theory of pattern recognition [1].

The well-known algebraic rules explain possible results that can be obtained by one or another algorithm of data sample processing. For example, an algebraic theory easily explains the results of inserting the intermediate or hidden layer of units in the one-layered classifier scheme since it can be easily interpreted by considering the corresponding equations.

In the middle of 1950s, F. Rosenblatt published an original work devoted to perceptron [2, 3] where he implemented a layer of hidden units and showed the efficiency of such complication of a classifier. The estimates of the coefficients were simply found by the least squares procedure; an optimal number of the hidden units could also be easily found. However, the algorithms designed for pattern recognition were not applied to solution of other interpolation problems of artificial intelligence, particularly, to the problem of step-by-step forecasting of random processes. This problem was considered to be fundamentally different from pattern recognition problem both in its purpose and in the structure of the algorithm [4].

A general approach to different problems implies that an input data set is divided into compact sets of observations (called clusters) and then, a general theory is applied to each of them. An adequate notion *compact set of processes* should be found to correspond such notions as *compact sets* and *cluster of images*. This is achieved by using a concept of optimal physical clusterization [5]; in all interpolation problems, it allows us to divide an input data sample into elementary clusters for subsequent unification of the processing algorithm. Here, the subset of observations of one of the clusters should be transformed into single-moment form. Below, we consider it in detail.

Further, some differences in solving particular problems will be pointed out, but clusterization allows us to use common algorithms for the solution of general problems. In the engineering perceptron-type recognition systems (see Figs. 1, 2) in the case, when all possible images are presented for recognition, a subset of hidden units, where the biggest signal is received, defines the pattern to which the input image should be referred. In the systems of stepwise forecasting of random processes, similar work can be fulfilled by the layer of single-moment subset of clusters obtained through optimal physical clusterization. A subsample, which corresponds to the cluster containing the output vector, gives the most accurate stepwise forecast. Algorithms for different interpolation problems have a common perceptron-type structure, in which the input set of observations is divided (subjectively or objectively) into optimal number of clusters.

Algebraic rules for systems of equations or inequalities used in algebraic pattern recognition theory can be applied to each separate cluster for the solution of any interpolation problems. This is the base for the general approach for solving different problems. Algebraic

Received March 18, 2000

transformations can be supplemented by estimating the efficiency of primary and secondary arguments-candidates [7, 9] and by minimizing their number by the combinatorial GMDH algorithm [10] in all interpolation problems. In addition, the GMDH algorithms are applied in order to increase the generalization property of the decision function, which is necessary in the case of incomplete input data samples.

1. THE CONCEPT OF PHYSICAL CLUSTERIZATION

According to this concept, for any stationary random process, which is characterized by a subset of observations, we can find such a clusterization of observations, which remains the same in all sufficiently representative samples of observations of this process. Therefore, by comparing all possible clusterizations of two samples, we can find one general clusterization which is called *physical clusterization* because it reflects physical properties of the process.

2. ALGORITHM OF UNSUPERVISED SEARCHING FOR OPTIMAL PHYSICAL CLUSTERIZATION BY CONSTRUCTING TWO CLUSTERIZATION TREES

For rational use of information, two clusterization trees are constructed to provide a search for optimal physical clusterization [5]. A comparison between trees allows us to find optimal physical clusterization. Several ways of efficient data sample division into two subsamples are developed, which are used for the construction of the clusterization trees. Let us describe one of them.

2.1. Discretization of Variables

Discretization of variables to a large number of levels can serve as an example of data sample transformation, which does not change clusterization presented in this sample. Due to this, we do not distinguish a photo of an object and its image in a newspaper, which consists of a large number of points. According to this approach, a data sample, which has N points, should be discretized to N levels [6]. Subsamples obtained in this way are used for the construction of two trees and their subsequent comparison in accordance with a balance criterion. It is recommended to construct one tree using an input data sample and another by the same sample discretized to N levels. Both subsamples should be standardized according to the maximum value. Any other variable transformation, which preserves physical clusterization, can be used in a similar way. For example, we can range a sample according to variance. Even points will form one subsample and odd points—the second subsample. Another example: one tree is constructed from the input data sample, another—from the data sample, which consists of the first analogs of each line of the same data sample, and so on. Analogs are the nearest neighbors of the observation pointed out in the data sample.

2.2. Algorithm for Construction of Hierarchical Trees

To define optimal physical clusterization, the dependence of the balance criterion upon the serial number of step in construction of two trees should be found. The minimum of this dependence should be determined. This minimum points out the optimal clusterization according to this criterion. The balance criterion is calculated by the formula

$$BL = \sum \frac{k_i - k_{eq}}{k_i},$$

where k is a common number of clusters in each compared clusterization and k_{eq} is a number of equal clusters in each compared clusterization.

Clusters, which contain the same points, are considered equal. One of the two possible approaches to the tree construction can be applied:

(1) the closest points should be united in one cluster at each step;

(2) in each cluster, two points with the greatest distance between them should be found and considered as centers; other points should be divided between them [7].

A second approach takes less computational time because optimal clusterization is usually closer to the root of a tree. Comparison of clusterizations obtained via these approaches is calculated by a balance criterion. It is known that it equals zero at the root of a tree, where all points are united into one cluster and at the top of a tree, where the number of clusters is equal to the number of data sample observations. Therefore, constructed dependence is M-shaped. Thus, the purpose of the tree construction is to find an optimal clusterization, which corresponds to the minimum of the balance criterion. There can be several minimums. This indicates that the data sample contains several objects for investigations.

3. ANOTHER DEFINITION OF SUBSAMPLE OF OBSERVATIONS WHICH FORMS THE FIRST CLUSTER AND ITS APPLICATION

A cluster, which contains an output vector followed by a forecasting vector, is considered as the first cluster. This cluster can be found not only by the comparison of two clusterization trees according to the balance criterion, but by simple sorting of variants according to the accuracy of short-term forecasting. For nonstationary processes with some trends of developing, the sorting procedure is simple. In such processes, we *a priori* know that the first cluster consists of a set of observations which are followed by output observation. We steadily increase the number of observations, which are included in a first cluster, as long as the accuracy of the

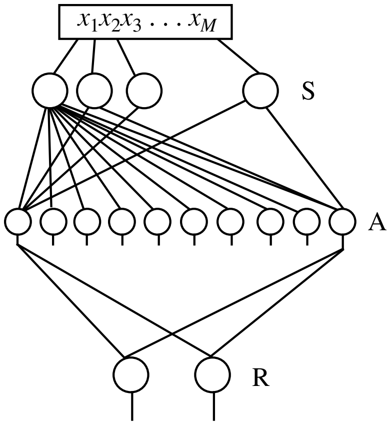


Fig. 1. Perceptron as a model of brain structure. The set of hidden units A is not divided into clusters.

forecast is increasing. Thus, we can find the whole first cluster and, simultaneously, make a most accurate forecast.

In the stationary processes, where the trend of development is absent or horizontal, the first cluster should be found according to the criterion of the clusterization balance. However, if the first cluster includes only the last observations, this indicates a tendency of evolution in the given process hidden for a simple observer. Therefore, the analysis of the structure and time of appearance of the first cluster observations opens a possibility of detecting the implicit tendencies of development. This, for example, is important in forecasting the results of activity of commercial companies. If the business goes well, the observations of the first cluster are gathered mainly in the beginning of the time interval of the representative input data sample. The uniform distribution of observations in the first cluster along the time interval is an indicator of stable situation. In the case, when business is going bankrupt, the observations of the first cluster are gathered near the output observation, as if they want to save it.

4. PERCEPTRON AND ENGINEERING PERCEPTRON-TYPE RECOGNITION SYSTEM

From the standpoint of today, perceptron should be considered as an elementary-multilayered neural network designed for modeling a brain structure, taking into account some specific properties of this object. These specific peculiarities imply the following basic constraints:

Supervised and unsupervised learning of a perceptron represents the black box situation. Only one or two vectors of the input signal are given for processing. A set of observations contained in input data is not used for the estimation of the coefficients by the least squares method. Instead of this, various adaptive procedures are recommended.

A priori information about the cluster where the maximum signal is received is not used. Moreover, a set of hidden units is not divided into clusters. Instead of this, the linear on coefficients polynomial decision function can be received and analyzed.

The vast majority of neural networks may be considered as the committees of perceptrons, in which the problem of interconnection between the perceptrons should be solved. The perceptron-type engineering recognition systems may be defined as a network of elements, where the above constraints are removed. Particularly, a black box situation is not observed and a set of hidden units is divided into standard clusters. A decision is made, taking into account information on the subset of hidden units where the maximum signal is received. Neural network can be considered as a committee of such perceptron-type pattern recognition systems. Self-organization of neural network architecture is performed automatically, i.e., without the participation of an expert.

The use of the set of so-called internal (hidden) units in the second layer of multilayered system is essential for perceptrons. In perceptron-type pattern recognition systems, the set of hidden units is divided into standard clusters.

In Fig. 1, the structure of a perceptron is shown. Its work is based on the analysis of the decision function. The following notation is used in the figure:

S—sensitive units;

A—associative or hidden units; and

R—decision units.

This notation corresponds to the notation of Rosenblatt [2]. More than once, the author of perceptron pointed out that his perceptron is a model for brain structure, and not the effective engineering system for pattern recognition. Actually, from the engineering point of view, a perceptron is not rational. The links between units S and A can be easily calculated so that on a definite (hidden) unit the maximum signal is received for definite image given in input data sample [8]. Thus, if the number of hidden units is no less than the number of observations given in the learning sample, then, to solve pattern recognition problem means to find a place of maximum and to define the cluster of hidden units where a maximum signal is received. We do not need to receive and to analyze the decision function [8]. However, the algorithm without the classification of a special discriminant function, shown at Fig. 2, can be recommended only in the case when data sample contains all possible images of an object (for example, for printed characters recognition). In the case of handwritten character recognition, only several typical images can be presented in input data sample. In this case, the GMDH algorithm with a definition of the discriminant function should be recommended (see Fig. 3).

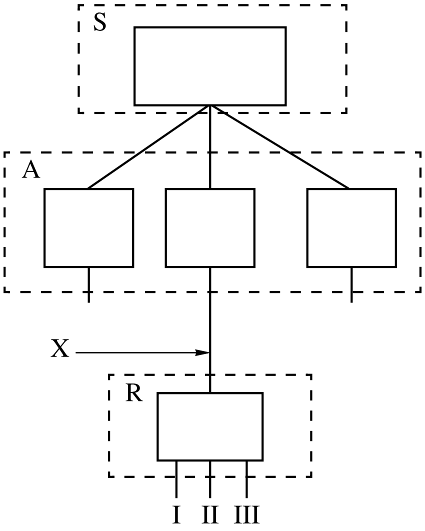


Fig. 2. A recognition system with a low degree of generalization. Decisions are made in accordance with the distance between the input image and some of the clusters. There is no need to apply the GMDH algorithms.

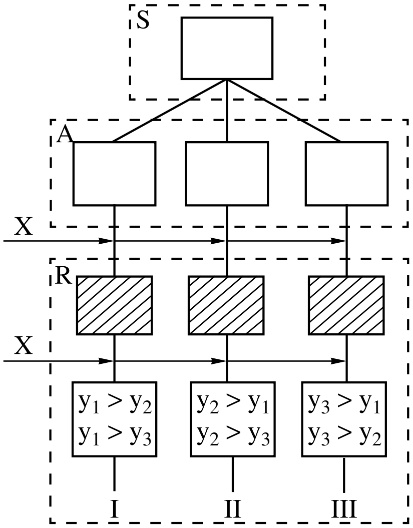


Fig. 3. A perceptron-type recognition system with a high degree of generalization. Decisions are made via discriminant analysis of functions obtained by the GMDH algorithms in all sufficiently representative clusters.

5. APPLICATION OF THE GMDH ALGORITHMS FOR PATTERN RECOGNITION AND DETECTION OF POLYNOMIAL DEPENDENCIES BY PERCEPTRON-TYPE ALGORITHMS

Figure 2 shows the perceptron-type system of pattern recognition for the case when the input data sample consists of all images presented for recognition. By using the notation of perceptron theory, we define

S—input data sample;

A—layer of hidden units, which should be divided into clusters;

R—decision unit, which indicates the cluster on which the biggest signal is received. This signal is equal to the module of the correlation coefficient of the input vector subject to recognition and vector of a given hidden unit;

X—input vector subject to recognition.

In Fig. 3, a perceptron-type pattern recognition system is presented, which is based on the analysis of a decision function. This function is obtained by using the GMDH algorithm. The units corresponding to the combinatorial GMDH algorithm are shaded. In Fig. 3, we denote:

S—input data sample;

A—layer of hidden units, which should be divided into clusters;

R—decision unit, which acts according to discriminant analysis;

X—input vector subject to recognition.

According to the rules of discriminant analysis, a supervisor specifies the values of output variables: one—for all correctly recognized images and zero—for all errors. A pattern recognition system should necessarily have generalization properties. To achieve this, the indicator of the biggest signal is not sufficient: it is necessary to receive polynomial discriminant functions using the combinatorial GMDH algorithm.

The input data sample can contain not only the primary features obtained from the object but the secondary ones, which are obtained by means of a certain technique from the values of primary features. The coordinates of the first analogs [7] and covariations of the first features [9] are very often used as secondary

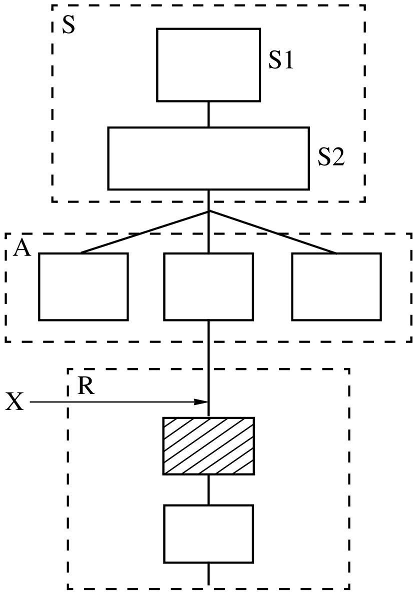


Fig. 4. A perceptron-type system of stepwise forecasting of the random processes. Forecasting is performed with the help of the GMDH algorithm for the standard cluster, in which an output single-moment observation is included.

Table 1

N	X_1	X_2	X_3
1	X_{11}	X_{21}	X_{31}
2	X_{12}	X_{22}	X_{32}
3	X_{13}	X_{23}	X_{33}
4	X_{14}	X_{24}	X_{34}
5	X_{15}	X_{25}	X_{35}
6	X_{16}	X_{26}	X_{36}

features. In the case of supervised learning, the input data sample should be divided into several compact subsamples or clusters. The division of a data sample is performed by a supervisor—an author of modeling. In the case of unsupervised learning, the above-described methods are used for the division of input data sample into compact clusters. In the simple case, the recognition of the input signal consists in assigning a new signal to the closest standard cluster in the feature space. In the problems where wide generalization is not necessary, it is not reasonable to use the GMDH algorithm (see Fig. 2). A simple indicator of the maximum signal value will suffice. It will put the input signal into correspondence with the cluster. Necessity in the interpolation decision rule only appears in the case when the most typical images are presented in the input data sample, and it is necessary to recognize many other images with the close feature values.

Discriminant functions for pattern recognition can be considered as polynomial approximations of dependencies, which connect the value of output variables of models with the feature values. This solves the problem of approximation of dependencies in experimental data by the polynomial linear on coefficients.

6. APPLICATION OF GMDH ALGORITHMS FOR STEPWISE FORECASTING OF RANDOM PROCESSES BY PERCEPTRON-TYPE ALGORITHMS

The algorithm of pattern recognition described above is very similar to the algorithm of stepwise forecasting of random processes, since both are based on the division of a second layer of units into compact subsamples (or clusters) and both implement a perceptron-type scheme. Figure 4 shows a perceptron-type algorithm for forecasting random processes, which performs the analysis of the forecasting function received by GMDH algorithms. The following notation is used here:

S1—input data sample normalized by biggest value;

S2—transformed data sample in single-moment form;

A—compact subsets of data (or clusters);

R—decision unit, which looks for the cluster, where the given output vector takes place.

Simplicity and clarity of pattern recognition algorithm [1] can be explained, first of all, by the following property of input data sample: cross-correlation between input data observations is small and inessential. In contrast to this, in stepwise forecasting problems, correlation of observations is very important. Therefore, it is reasonable to use two data samples during forecasting. In one sample, the rows are correlated with the previous rows, and the second data sample is obtained by extending the set of arguments. Several independent future and delayed variables are introduced into each observation, which characterizes the state of the object. The data sample is extended and takes the so-called single-moment form. In this form, the correlations between observations become immaterial and observations can be used in any sequence, which is important for the solution of certain problems.

Example. The transformation of the data sample from a usual to a single-moment form.

Suppose that the following form of the sample of time series is given (see Table 1).

In order to take into account two delayed arguments, we introduce them into the sample as independent variables. We get the following data sample in a single-moment form (the first two observations and the last one are not taken into account) (see Table 2).

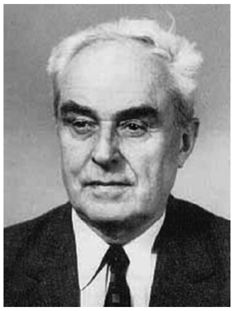
The output variable is usually specified by a supervisor, but we may consider all the variables as output variables, in turn or simultaneously. The last causes the use of implicit patterns in the combinatorial algorithm. The forecasting models express dependence of future variables from current and delayed values of arguments.

Table 2

N	X_{1k+1}	X_{1k}	X_{2k}	X_{3k}	X_{1k-1}	X_{2k-1}	X_{3k-1}	X_{1k-2}	X_{2k-2}	X_{3k-3}
3	X_{14}	X_{13}	X_{23}	X_{33}	X_{12}	X_{22}	X_{32}	X_{11}	X_{21}	X_{31}
4	X_{15}	X_{14}	X_{24}	X_{34}	X_{13}	X_{23}	X_{33}	X_{12}	X_{22}	X_{32}
5	X_{16}	X_{15}	X_{25}	X_{35}	X_{14}	X_{24}	X_{34}	X_{13}	X_{23}	X_{33}

REFERENCES

1. Zhuravlev, Yu.I. and Gurevich, I.B., Pattern Recognition and Image Recognition, *Pattern Recogn. Image Anal.*, 1991, vol. 1, no. 2, pp. 149–181.
2. Rosenblatt, F., *Principles of Neurodynamics Perceptrons and the Theory of Brain Mechanisms*, Washington: Spartan Books, 1962.
3. *Pertseptron—Sistema Raspoznavaniya Obrazov* (Perceptron—A System of Pattern Recognition), Ivakhnenko, A.G., Ed., Kiev: Naukova Dumka, 1975.
4. Vassilyev, V.I., *Recognition Systems*, A Handbook, Kiev: Naukova Dumka, 1983.
5. Jambu, M., *Classification Automatique Pour L'analyse Des Donnees Methodes et Algorithmes*, Dunod Bordas, Paris, 1978.
6. Ivakhnenko, A.G., Ivakhnenko, G.A., and Mueller, J.A., Self-Organization of Optimization of Optimum Physical Clustering of a Data Sample for a Weakened Description and Forecasting of Fuzzy Objects, *Pattern Recogn. Image Anal.*, 1993, vol. 3, no. 4, pp. 415–422.
7. Ivakhnenko, A.G., Kovalishin, V.V., *et al.*, Self-Organization of the Neural Networks with the Active Neurons for Forecasting the Activity of Chemical Compounds with the Help of Algorithm of Analog Searching, *Control and Computer Science Problems*, 1999, no. 1, pp. 69–77.
8. Ivakhnenko, A.G. and Ivakhnenko, G.A., Perceptron Synthesis according to Clusterization-Balance Criterion, *Pattern Recogn. Image Anal.*, 1995, vol. 5, no. 3, pp. 337–341.
9. Krug, G.K. and Krug, O.Yu., Matematicheskii Metod Klassifikatsii Drevnei Keramiki (Mathematical Method for Classification of Ancient Ceramics) in *Trudy Instituta Arkheologii Akademii Nauk SSSR* (Proc. of Inst. of Archeology, Academy of Sciences of USSR), Moscow: Nauka, 1965, pp. 318–325.
10. Madala, H.R. and Ivakhnenko, A.G., *Inductive Learning Algorithms for Complex Systems Modeling*, Boca Raton: CRC, 1994.



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

Grigorii A. Ivakhnenko. Born 1966. Graduated from the Kiev Polytechnical 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 23 papers.

Evgeniya A. Savchenko. Born 1969. Graduated from the Kiev Polytechnical Institute in 1992. In 1995–1998, worked as an engineer in the Institute of Complex Transport Problems. Now, a postgraduate at the International Center of Information Technologies and Systems of National Academy of Sciences of Ukraine. Scientific interests: complex objects modeling, pattern recognition, and forecasting random processes by the GMDH approach.

Donald C. Wunsch. Received his BS degree from the University of New Mexico in 1984, MS degree from the University of Washington in 1987, and PhD degree (Electrical Engineering) in 1991 from the same university. In 1984–1993, was a Senior principal scientist in the Boeing Company. In 1994–1999 was an Associate Professor in Texas Technical University. Now, a Distinguished Professor in the University of Missouri-Rolla (Computer Engineering). Published more than 100 technical papers on adaptive critic design, neural networks, fuzzy systems, reliability, nonlinear adaptive control, complex systems, financial engineering, and optimization.