Appl. Math. and Comp. Sci., 1993, vol.3, No.3, 573–593

GROUP METHOD OF DATA HANDLING IN TECHNICAL DIAGNOSTIC TASKS†

JANUSZ KUŚ*, ZBIGNIEW BANASZAK**

The work presents a certain concept of an advisory system, which assists in making decisions in the technical diagnostic tasks. To this end, as a mathematical formalizm, the Group Method of Data Handling (GMDH), from the class of genetic algorithms has been adopted. The rules of the knowledge base creation and the stages of the algorithm GMDH realization have been discussed and commented. Following this theoretical basis, the assumption of the computer–aided diagnostic of technical objects system has been formulated. The work has been completed with an example illustrating advantages and disadvantages of the solution proposed.

1. Computer–Aided Technical Diagnostic

In many fields in which computers begin to take over functions reserved so far only for experts and specialists, diagnostics is already a classic example. This domain remains in close relation with the mathematical apparatus – logic. This apparatus formulated and developed long before the formation of digital computers, is perfectly applicable in a formal representation of knowledge and its implementation in computer memory. The ability to accumulate a large amount of this kind of data, and its simple search, actualization, and processing has created the terms for developing a particular class of computer programs – expert systems.

Apart from the knowledge base (which consists of: a fact set and the connection rules among them), the expert system must be equipped with an inference procedure, and a response quality correctness (logical consistency) procedure. Processing the accumulated knowledge by means of the inference rules, the computer may give answers to questions from a defined subject area.

Thus, the expert systems are perfectly suitable for the solutions of the diagnostic tasks, which are easily adopted by the processing of the conditional logic sentences (*if ..., then ...*). This is why INTERNIST and CADUCEUS (computer diagnosis of internal diseases), CASNET (system of computer–aided glaucoma treatment), or MYCIN (computer diagnosis of blood diseases) were among the first and most successful programs of this type.

† This work was supported by the Polish State Committee for Scientific Research under grant No. 7–1086 91 01

* P.I.W. "EMIO" Ltd Co., Wrocław, ul. Skwierzyńska 21, Poland

** Dept. of Applied Mathematics, Kuwait University, Safat–13060, Kuwait

The first successful results in expert systems raised hopes for extension of applications, but at the same time revealed the development barriers of this class of software. The primary restriction was the formation of a data base. The only source of essential information are experts, i.e. specialists who know deeply both theory and practice of the subject field considered. This knowledge can only partly be put into rigid frames of the applied formalism of a mathematical representation. To omit this problem we may apply the theory of scattered sets.

Still, the knowledge reduced only to definitions, theorems, rules, and arithmetic formulae proves to be unsatisfactory for the solution of problems at the level comparable to that of experts. In diagnostic tasks, an additional problem appears – the limited input information, as diagnostics is based on defining the state of the examined object and the location of the possible defects, only on the basis of external observation. The expert systems intended for these purposes, although still developed and used, have not met the expectations so far.

This is why, the attempts to introduce new mathematical apparatuses and concepts are made, which, by extending a formulae of expert systems, would allow to computerize decision processes. One of such proposals has been presented in this work.

2. Concept of Advisory System. Genetic Algorithms

The concept of the presented computer–aided decision making system in diagnostic tasks differs from the traditional construction of the expert system in a few essential points. Therefore, in the other part of this work, the expression ***advisory system*** will be used.

The main assumptions of the planned advisory system should be specified first:

- the subject of the analysis are real problems with the degree of complexity requiring great human knowledge, which cannot be fully described in a formal way,
- information about the examined object may be incomplete,
- information about the examined object come from the observation of its behavior,
- high quality of advisory system as far as speed and accuracy are required.

Therefore, two tasks should be solved:

i) accumulation and representation of knowledge,
ii) modelling and analysis of complex real objects.

Their common features are simple solutions at a low level of complexity, and lack of effective methods and algorithms for the tasks of medium and high degree of compilation. The properties mentioned indicate, that an adequate mathematical apparatus would be a formalism coming from the class of genetic algorithms.

In this solution, the inference procedure is realized on the following stages:

1. Selecting information from the knowledge base is essential for the solution of the formulated task.
2. Constructing a simple, elementary mathematical model of the object examined.
3. Evolution of a model family to a form with the required degree of modelling accuracy and analytical complexity.
4. Selection of the best model.
5. Solution of a diagnostic task on the basis of a mathematical model analysis.

The structure of the modelling procedure presented, impose the requirements as for the concept of the defined components of the planned advisory system:

- the knowledge about the class of objects considered must be recorded in a form adjusted to the requirements of the inference procedure,
- diagnostic tasks are solved at a level of confidence defined in advance,
- acquiring a response from the advisory system requires formation of a question as well as the definition of the inference procedure parameters (practically the implemented genetic algorithm).

The problems mentioned will be developed in the further part of this work with a detailed description of each of the advisory system components.

3. Concept of Knowledge Representation

The first stage of the advisory system formation is a data base construction. Considering specificity of the class of objects examined, the accepted initial assumptions, and the assignment of the collected knowledge for the diagnostic purposes, it is visible, that the key problem of the stage discussed is the representation of the knowledge. In classic expert systems this problem is solved by means of logic and mathematical apparatus: experts formulate logical sentences and arithmetic formulae describing behavior of the objects considered.

In the complex diagnostic tasks, information of this type is usually incomplete and, as a result, it is not possible to carry out the inference procedure based only on the logical analysis of the collected knowledge. The introduction of the elements of objective approximation (the one the genetic algorithms introduce) is essential. If all the rules essential to make an appropriate diagnosis are not known, the missing knowledge can be completed just through the computer approximation. Thereby, between the data base analysis and the solution of diagnostic tasks, a medium stage – modelling is introduced.

Following such a concept, the analytical apparatus of the advisory system must be based on arithmetics of the real numbers, and in this connection, the knowledge representation must be the same. The only objective information of this type is empirical data recording the selected states of the objects observed.

Therefore, to create the knowledge base one should:

i) make the selection of the parameters describing the state of an object. A part of the selected quantities makes a group of independent variables, the rest – a group of dependent variables,

ii) organize an observation system of the object behavior,

iii) make a registration of parameters value in the selected states.

Thus, the created knowledge base consists of a performed registration (observation) set, i.e. a set of n–element vectors (where n is the number of the parameters observed). The set may be interpreted as a group of points in n–dimensional hyperspace (see Fig. 1).

a)

X	Y
x_1	y_1
x_2	y_2
x_3	y_3
x_4	y_4
x_5	y_5
x_6	y_6
x_7	y_7
x_8	y_8
x_9	y_9

b)

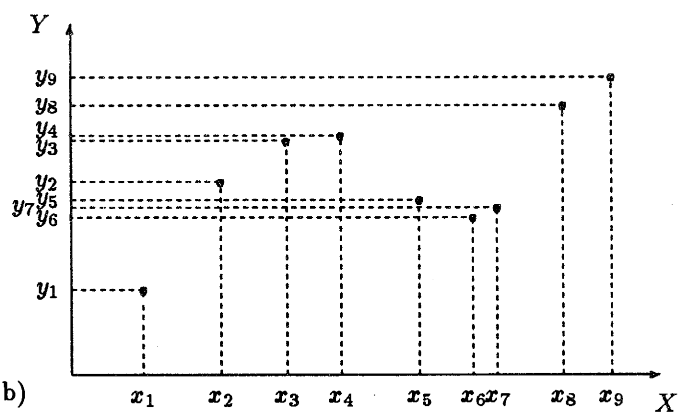


Fig. 1. The example of a knowledge base for an object described by two parameters observed: X and Y, consisting of nine registrations; a)registered observations, b) graphic representation of an observation set.

Thereby, the acceptance of the assumption of the knowledge base creation of the unknown (unidentified) objects allows to create information banks, which in a simple way guarantee the communication between the advisory system and data base systems.

4. Clustering

The concept of the knowledge representation and the inference procedure assume lack of the arbitrary introduced rules governing an object behavior. The solution of a diagnostic task requires a synthesis of these rules. In this connection, a preliminary preparation of the knowledge base is essential.

The solution of a diagnostic task, as a rule, does not require a complete knowledge of the whole object, but only of its certain fragment. Thus, it is advisable to prepare data so as the inference procedure would take advantage only of these observations, which show homogeneity of the selected features, e.i. from the data comprising the searched information.

This sorting of data is handled by a branch of knowledge called clustering or concentration analysis (Bijnen, 1974). Clustering deals with a formation of homogeneous elements into groups, and at the same time, with their separation from other elements. Thereby, the output data set is decomposed into subsets (classes), which may be accepted as the representation of the elements comprised in them.

Introduced by the analysis of concentration, a higher level of abstraction of data representation is essential in expert and advisory systems because of:

- seizure of minuteness of detail,
- reduction of the number of processed data,
- emphasis on the importance of the data connection rules.

4.1. Formal Description of Clustering

Let's consider the Euclidean clustering space VXY, spanned over the field of real numbers R^{m+n}, being a simple sum of the two subspaces

$$VXY = VX \cup VY \tag{1}$$

where:

$$VX(R) = R^m, \tag{2}$$

$$VY(R) = R^n. \tag{3}$$

Let's $(x_1, x_2, ..., x_m, y_1, y_2, ..., y_n)$ be a vector $m+n$ of the observed parameters.

Definition 1. *Clustering* of a point pattern in space VXY is based on:

- the separation of subspace VX_1Y_1, fulfilling the following conditions:

(i) $VX_1 \subset VX,$ (4)

(ii) $VY_1 \subset VY,$ (5)

(iii) $VX_1Y_1 \subset VX_1 \cup VY_1,$ (6)

(iv) $\forall(x \in VX_1)\ \exists(y \in Y_1)\ (y = f(x)).$ (7)

- breaking a point pattern enclosed in a subspace VX_1Y_1 into k subsets:

$$X_1 = \sum_{j=1}^{k} \overline{x}_j \text{ and} Y_1 = \sum_{j=1}^{k} \overline{y}_j, \text{ such that}$$

(v) one–to–one correspondence exists $\overline{x}_j \Leftrightarrow \overline{y}_j,$

(vi) external criterion of clustering is fulfilled.

If conditions (i)–(vi) of Definition 1 are fulfilled, we say, that the external criterion of clustering caused decomposition of space VXY into k clusters. The genetic algorithm of the inference procedure will examine the dependences $y = f(x)$.

4.2. Criteria and Algorithms of Clustering

After analyzing the above mentioned remarks, it becomes obvious, that the better clustering, the better results of the inference procedure operation. To this end, the division of the knowledge base into clusters should make possible to separate such an observation subset, that will comprise all the data essential to solve the formulated task, and at the same time, will not include redundant information.

In literature, it is not difficult to find a great amount of different algorithms and external criteria of clustering (Bijnen, 1974; Ivachnenko, 1987; Kucharczyk, 1982). Therefore, making an appropriate selection must be preceded with an accurate analysis of the problem considered. Below, the example of the solution of a clustering task has been presented, as an illustration (Kucharczyk, 1982). One of the simplest and often applied external criteria of clustering is the critical distance (Cdist). It is connected with graphic interpretation of the knowledge base as a point pattern in Euclidean space (see Section 3).

There are several methods to determine it, e.g.:

$$C\text{dist} = \max_i \min_j d(i,j), \tag{8}$$

$$C\text{dist} = \frac{1}{N}\sum_{i=1}^{N} \min_j d(i,j), \tag{9}$$

where $d(i,j)$ denotes the distance between points i and j, and N is number of points (observations).

One of the algorithms of clustering, taking advantage of the critical distance as an external criterion is Singlm (Single linkage method), called Wroclaw taxonomy as well.

The realization of this algorithm comprises the following steps:

1. Calculate value of the critical distance Cdist.
2. For each point create one–element cluster in a form of hypersphere with a radius Cdist.
3. Find a pair of clusters p and q the least distant from each other, calculating distance $d(p,q) = \min_{i \neq j} \min_j d(i,j)$.
4. If $d(p,q) > C$dist or if there is only one cluster, STOP.
5. Connect p and q into one new cluster, numbering it p (where $p < q$), and removing cluster number q.
   Reduce by 1 cluster numbers greater than q.
6. Return to step 2.

As shown, the Singlm method joins two clusters as long as a minimal intercluster distance (the distance between the two nearest points of a cluster pair) is smaller than Cdist.

The examples provided are to illustrate the problems of the knowledge base preliminary preparation by making full use of the cluster analysis methods. Obviously, in advanced applications, more complicated criteria and algorithms are used. Very often they are individually developed according to the specificity of the tasks solved. In such analyses, the knowledge of the solutions implemented in the functioning systems is very useful. Therefore, the readers interested in these problems are advised to study (Banaszak and Kobyłecki, 1992; Ivachnenko *et al.*, 1986, 1987).

To illustrate the importance of clustering in the initial knowledge processing, Figure 2 presents a graphic interpretation of the knowledge base shown in Figure 1 after carrying out the division of the observations into clusters.

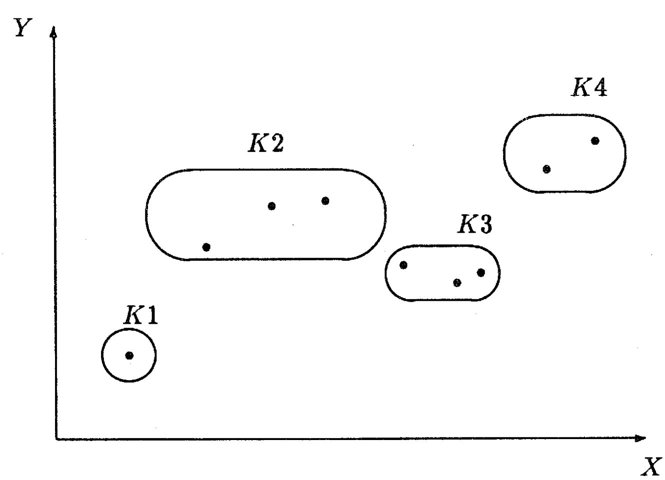


Fig. 2. Graphic representation of the knowledge base (from the example shown in Fig. 1) divided into four clusters $K1$, $K2$, $K3$, $K4$ of homogenous observations.

5. Group Method of Data Handling

When the knowledge base clustering is finished, i.e. the first stage of the object considered behavior rules synthesis, the second stage may start, i.e. modelling. In Section 2, the premises which resulted in the selection of genetic algorithms as an appropriate mathematical apparatus have been presented.

The idea of this class of algorithms is based on the preliminary assumption of a simple, approximate solution (model, estimator, etc.), and then, its evolution to a form, meeting the requirements as for the analytical complexity, accuracy, etc. An interesting property of genetic algorithms is the fact, that accuracy of a final solution depends only in a small degree on accuracy of a initial solution. Thus, such methods prove effective in the application for complex modelling tasks, identification, estimation, forecasting, etc.

In the presented concept of the advisory system, the inference procedure is based on the Group Method of Data Handling (GMDH), as a typical representative of the genetic algorithms class (*Self organizing* ..., 1984; Ivachnenko, 1990). The input data of algorithm GMDH are observations from a selected cluster of the knowledge base (the method of this cluster selection will be described in the further part of the work). They are used to aquire the next population of variables, which better represent their common features. The preliminary solution (model) is accepted to be polynomial of second degree of two variables. Therefore, the creation of the next *generation* of data begins with the solution of equations of the regression:

$$y = A + Bx_i + Cx_j + Dx_i^2 + Ex_j^2 + Fx_ix_j \tag{10}$$

for each pair of input variables x_i and x_j.

After forming the equations of regression, those of them are selected, which model the considered object best are selected. At the next step the same process is repeated for the population of variables (equations of regression) obtained lately. The whole process is continued until the equations of regression at the last step have better approximating properties than at the previous step.

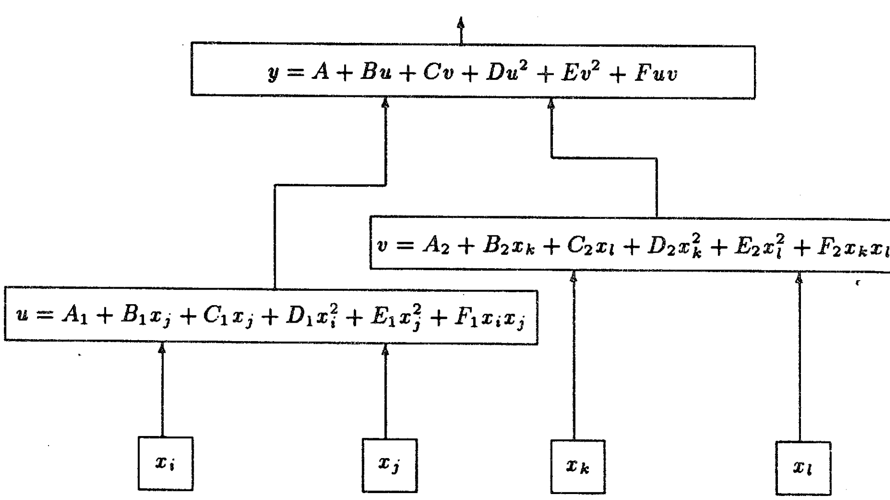


Fig. 3. Diagram of variables propagation in the algorithm GMDH;
x_i, x_j, x_k, x_l – input data (initial population),
u, v – first population of estimators,
y – second population of estimators.

The final solution is a polynomial:

$$y = a + \sum_{i=1}^{n} b_i x_i + \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij} x_i x_j + \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} d_{ijk} x_i x_j x_k + \ldots \tag{11}$$

known as Ivachnenko polynomial, (Ivachnenko *et al.*, 1989). Such the polynomial, in particular, may represent a characteristic function defined in a diagnostic task.

Figure 3 illustrates the described diagram of variables propagation, leading to the construction of Ivachnenko polynomial.

5.1. Description of Algorithm GMDH

The input data for the algorithm GMDH are the observations of the one of the knowledge base clusters. Let's assume the size of this cluster is n observations. Let each observation be the registration of the state described by values $m+1$ parameters, recorded in a form of a state vector: $(x_1, x_2, \ldots, x_m, y)$. The input data can be recorded in a form of matrix of size $n \times m + 1$ (see Fig. 4).

Let's assume, that the task is to find a value of one parameter y in a state, in which the other parameters have values $x_1, x_2, \ldots, x_m$. Let's divide the input data into two observation groups:

- *learning observations* (from 1 to t), for the construction of the regression polynomial,
- *testing observations* (from $t+1$ to m), for the control of modelling accuracy and the selection of the best approximating polynomials.

X			Y	
x_{11} x_{12}	...	x_{1m}	y_1	
...		...	...	learning observations
x_{t1} x_{t2}	...	x_{tm}	y_t	
...		...	...	
				testing observations
x_{n1} x_{n2}	...	x_{nm}	y_n	

Fig. 4. Structure of input data for algorithm GMDH.

Step 1: *Evolution of the new population of variables.*

For each pair of independent variables x_p and x_q, the regression polynomial is constructed:

$$y^* = A_{pq} + B_{pq}x_p + C_{pq}x_q + D_{pq}x_p^2 + E_{pq}x_q^2 + F_{pq}x_p x_q, \tag{12}$$

where $p = 1, 2, \ldots, m-1$, and $q = p+1, p+2, \ldots, m$, which best approximates parameter y in the set of learning observations.

With the mean square deviation (18) as the evaluation criterion of estimation quality, the postulate of the optimal quality of approximation is tantamount to the fulfillment of the condition:

$$S = \sum_{i=1}^{t} (y_i^* - y_i)^2 \rightarrow \min \tag{13}$$

where

$$y_i^* = A_{pq} + B_{pq}x_{ip} + C_{pq}x_{iq} + D_{pq}x_{ip}^2 + E_{pq}x_{iq}^2 + F_{pq}x_{ip}x_{iq}. \tag{14}$$

Thus, the task is to find the values of coefficients A_{pq}, B_{pq}, C_{pq}, D_{pq}, E_{pq}, F_{pq}. which minimize expression (13). This is tantamount to the satisfaction of the system of equation

$$\begin{aligned}
&\frac{\partial S}{\partial A_{pq}} = 0, \qquad \frac{\partial S}{\partial B_{pq}} = 0 \\
&\frac{\partial S}{\partial C_{pq}} = 0, \qquad \frac{\partial S}{\partial D_{pq}} = 0 \\
&\frac{\partial S}{\partial E_{pq}} = 0, \qquad \frac{\partial S}{\partial F_{pq}} = 0
\end{aligned} \tag{15}$$

Equations (13), (14) and (15) lead to the system of equations

$$\left.\begin{aligned}
&\sum_{i=1}^{t} 2(A_{pq} + B_{pq}x_{ip} + C_{pq}x_{iq} + D_{pq}x_{ip}^2 + E_{pq}x_{iq}^2 \\
&\qquad\qquad\qquad + F_{pq}x_{ip}x_{iq} - y_i) = 0 \\
&\sum_{i=1}^{t} 2x_{ip}(A_{pq} + B_{pq}x_{ip} + C_{pq}x_{iq} + D_{pq}x_{ip}^2 + E_{pq}x_{iq}^2 \\
&\qquad\qquad\qquad + F_{pq}x_{ip}x_{iq} - y_i) = 0 \\
&\sum_{i=1}^{t} 2x_{iq}(A_{pq} + B_{pq}x_{ip} + C_{pq}x_{iq} + D_{pq}x_{ip}^2 + E_{pq}x_{iq}^2 \\
&\qquad\qquad\qquad + F_{pq}x_{ip}x_{iq} - y_i) = 0 \\
&\sum_{i=1}^{t} 2x_{ip}^2(A_{pq} + B_{pq}x_{ip} + C_{pq}x_{iq} + D_{pq}x_{ip}^2 + E_{pq}x_{iq}^2 \\
&\qquad\qquad\qquad + F_{pq}x_{ip}x_{iq} - y_i) = 0 \\
&\sum_{i=1}^{t} 2x_{iq}^2(A_{pq} + B_{pq}x_{ip} + C_{pq}x_{iq} + D_{pq}x_{ip}^2 + E_{pq}x_{iq}^2 \\
&\qquad\qquad\qquad + F_{pq}x_{ip}x_{iq} - y_i) = 0 \\
&\sum_{i=1}^{t} 2x_{ip}x_{iq}(A_{pq} + B_{pq}x_{ip} + C_{pq}x_{iq} + D_{pq}x_{ip}^2 + E_{pq}x_{iq}^2 \\
&\qquad\qquad\qquad + F_{pq}x_{ip}x_{iq} - y_i) = 0
\end{aligned}\right\} \tag{16}$$

The system of equations (16) leads, after transformation, to the system of equations

$$\left.\begin{aligned}
&A_{pq}\sum_{i=1}^{t}1+B_{pq}\sum_{i=1}^{t}x_{ip}+C_{pq}\sum_{i=1}^{t}x_{iq}+D_{pq}\sum_{i=1}^{t}x_{ip}^2\\
&\qquad\qquad+E_{pq}\sum_{i=1}^{t}x_{iq}^2+F_{pq}\sum_{i=1}^{t}x_{ip}x_{iq}=\sum_{i=1}^{t}y_i\\
&A_{pq}\sum_{i=1}^{t}x_{ip}+B_{pq}\sum_{i=1}^{t}x_{ip}^2+C_{pq}\sum_{i=1}^{t}x_{ip}x_{iq}+D_{pq}\sum_{i=1}^{t}x_{ip}^3\\
&\qquad\qquad+E_{pq}\sum_{i=1}^{t}x_{ip}x_{iq}^2+F_{pq}\sum_{i=1}^{t}x_{ip}^2x_{iq}=\sum_{i=1}^{t}x_{ip}y_i\\
&A_{pq}\sum_{i=1}^{t}x_{iq}+B_{pq}\sum_{i=1}^{t}x_{ip}x_{iq}+C_{pq}\sum_{i=1}^{t}x_{iq}^2+D_{pq}\sum_{i=1}^{t}x_{ip}^2x_{iq}\\
&\qquad\qquad+E_{pq}\sum_{i=1}^{t}x_{iq}^3+F_{pq}\sum_{i=1}^{t}x_{ip}x_{iq}^2=\sum_{i=1}^{t}x_{iq}y_i\\
&A_{pq}\sum_{i=1}^{t}x_{ip}^2+B_{pq}\sum_{i=1}^{t}x_{ip}^3+C_{pq}\sum_{i=1}^{t}x_{ip}^2x_{iq}+D_{pq}\sum_{i=1}^{t}x_{ip}^4\\
&\qquad\qquad+E_{pq}\sum_{i=1}^{t}x_{ip}^2x_{iq}^2+F_{pq}\sum_{i=1}^{t}x_{ip}^3x_{iq}=\sum_{i=1}^{t}x_{ip}^2y_i\\
&A_{pq}\sum_{i=1}^{t}x_{iq}^2+B_{pq}\sum_{i=1}^{t}x_{ip}x_{iq}^2+C_{pq}\sum_{i=1}^{t}x_{iq}^3+D_{pq}\sum_{i=1}^{t}x_{ip}^2x_{iq}^2\\
&\qquad\qquad+E_{pq}\sum_{i=1}^{t}x_{iq}^4+F_{pq}\sum_{i=1}^{t}x_{ip}x_{iq}^3=\sum_{i=1}^{t}x_{iq}^2y_i\\
&A_{pq}\sum_{i=1}^{t}x_{ip}x_{iq}+B_{pq}\sum_{i=1}^{t}x_{ip}^2x_{iq}+C_{pq}\sum_{i=1}^{t}x_{ip}x_{iq}^2+D_{pq}\sum_{i=1}^{t}x_{ip}^3x_{iq}\\
&\qquad\qquad+E_{pq}\sum_{i=1}^{t}x_{ip}x_{iq}^3+F_{pq}\sum_{i=1}^{t}x_{ip}^2x_{iq}^2=\sum_{i=1}^{t}x_{ip}x_{iq}y_i
\end{aligned}\right\}\qquad(17)$$

Solving the system of equations (17), one can calculate the demanded values A_{pq}, B_{pq}, C_{pq}, D_{pq}, E_{pq}, F_{pq} acquiring for each pair of parameters p and q the optimal regression polynomial (12).

The final operation of this stage of algorithm is to calculate values of estimators y_i^* according to the equation (14). These values are recorded in the separate column of the auxiliary matrix Z. The procedure described in Step 1 is repeated for each pair of variables x. Each column of matrix Z is a variable of a new population.

Step 2: *The selection of the new population variables.*

After the construction of all the estimators of the new population we should select those of them, which approximate the demanded dependency $y = f(x_1, x_2, ..., x_m)$ best. The mean square criterion of quality estimation (18) known as the regularity criterion is used for this purpose. Its domain is the set of the testing observations.

For each column j of matrix $\boldsymbol{Z}$, the coefficient is calculated:

$$r_j^2 = \frac{\sum_{i=t+1}^{n} (z_{ij} - y_i)^2}{\sum_{i=t+1}^{n} y_i^2} \tag{18}$$

In the proposed solution, during the selection of the new population estimators, the criterion of population constancy is also applied, which means that at the next stage of evolution the same number of elements m takes part. Therefore, in further calculation for matrix $\boldsymbol{X}$, one should rewrite m column of matrix $\boldsymbol{Z}$, for which coefficient r_j achieves the least values.

Step 3: *Optimality test.*

After generating the new population and making its selection, one can carry out the evaluation of the efficiency of the evaluated mathematical model. To this end, one should choose the best approximating Ivachnenko polynomial, calculating

$$R_{\min} = \min_j r_j \tag{19}$$

In the consecutive stages of the optimal model evolution value $R_{\min}$ may change according to one of the dependences presented in Figure 5.

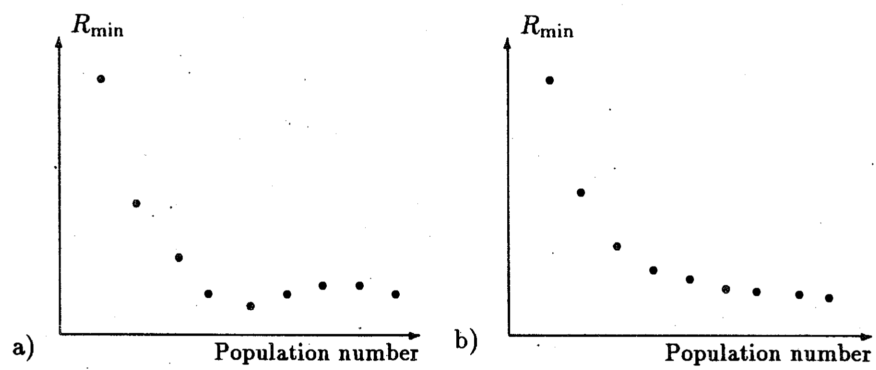


Fig. 5. The diagram of dependences $R_{\min}$ on population number: a) dependence with extreme point (points), b) monotonic–asymptotic dependence.

In case of the dependence described in Figure 5a, indication of the optimal model is tantamount to finding the first minimum. In this case, the execution of the algorithm GMDH is stopped for a moment, the best index of quality estimation R is greater than in the previous population. In such a situation value $R_{\min}$ may tend asymptotically to a certain constant value (most frequently 0). Then, the optimality test is based on either the limitation of the number of evaluated population or testing the boundary (lower) value of index $R_{\min}$.

The optimality test allows to establish the best mathematical model, i.e. Ivachnenko polynomial, which approximates dependence $y = f(x_1, x_2, .., x_m)$ best.

6. Inference Procedure

The Group Method of Data Handling, presented in Section 5, enables computer identification of complex real objects, and the construction of their mathematical models (in a form of characteristic functions recorded as Ivachnenko polynomial) in situations, where theoretical knowledge of the object considered does not give sufficient information to solve diagnostic tasks, to forecast etc. Such information may be acquired through the analysis of the constructed models. Such tasks may carry out the inference procedure based on the algorithm GMDH.

To this end, certain preparatory actions should be executed: the selection of the parameters describing the state of the object, registration of their values during the observation of the object in normal and emergency state, and recording of the registration in the knowledge base, which in turn is submitted to clustering. The method of executing this operation has been described in Section 3 and 4.

In diagnostic tasks it is essential, that at least one of the parameters from the knowledge base should play a role of a state identificator – emergency/normal. It means, that for each parameter, the scope of admissible values in a normal state should be explicitly defined. The other values will signal the state of failure. In technical diagnostics, which is the subject of this work, the identificators of the state are most frequently the output quantities (depended variables) such as: the engine rotational speed, frequency of transmitter circuit, etc.

The task of the diagnostics may include the definition of the object state and/or location of defects. In both cases finding the solution is connected with assigning the values of the unknown parameters, which, while fulfilling the description of the object state, will enable to formulate the answers.

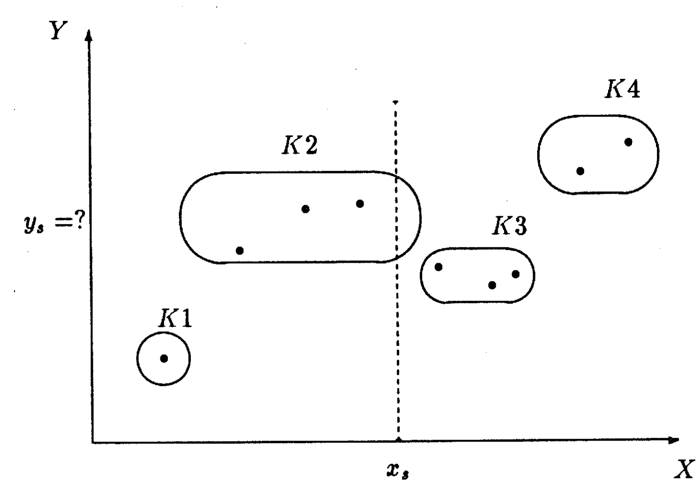


Fig. 6. Graphic interpretation of a preliminary diagnostics task.

Therefore, the first step of the inference procedure is to define the values of the parameters observed in the state analyzed, and indicate the other parameters (unknown) indispensable to generate the answers. Considering the example from Figures 1 and 2, let's analyze the situation, where $x = x_s$. The complete description of the object state requires finding value $y = ?$ (see Fig. 6). Thus, the task is to find a mathematical model, which accurately approximates the dependence $y = f(x)$. The next step is the selection of the data group (cluster), which comprises information enabling to give a precise answer to a given question.

In the discussed advisory system, the selection of a cluster is executed by the method of the nearest neighbour, according to which, after dropping all points on a line (and in general case on hyperplane) $y = 0$, the estimating cluster is indicated by the nearest point to the point defined in question (see Fig. 7). The least distance of the dropping points is $d(x_s, x_4) = x_s - x_4$, and $x_4 \in K2$, $K2$ has been selected as an estimating cluster.

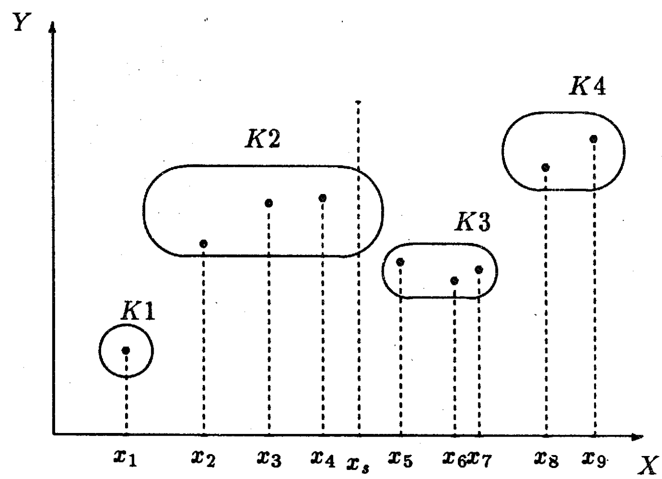


Fig. 7. Illustration of how the estimating cluster is selected.

The estimating cluster observations are at the next stage of the inference procedure divided, into learning and testing observations, and then, the evolution of Ivachnenko polynomial, estimating dependence $y = f(x)$ in the cluster $K2$ neighborhood, is carried out.

After constructing the model, the value $y = y_s$ is calculated (see Fig. 8). The realization of the algorithm GMDH enables to achieve two targets:

- determining values of the parameters describing the object state,
- constructing a mathematical model of the object considered.

The knowledge of the parameter values enables to solve the first of the diagnostic tasks, i.e. state identification. To this end, to the knowledge base of the advisory system one should arbitrarily insert complementary information, which will allow the interpretation of the observed and approximated quantities. The information must comprise the following data:

i) a list of the parameters, which may be used to diagnose an object state,

ii) a list of the boundary values of selected quantities and interpretation of the particular variation intervals.

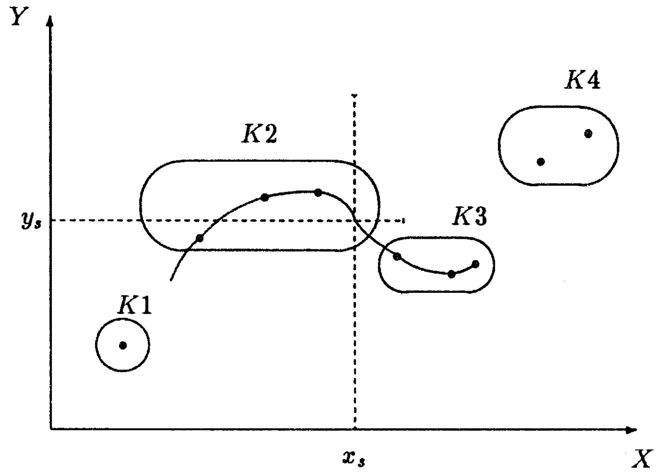


Fig. 8. Illustration of the mathematical model construction and answer calculation.

Figure 9 presents the example of data supplementing knowledge base and the adequate graphical interpretation. In the example presented, the diagnosis, carried out by the inference procedure of the advisory system states on the basis of data collected in the knowledge base, and the results of GMDH approximation, that (x_s, y_s) *is a normal state, but close to the boundary of the emergency state 1.*

In case of the emergency state, the inference apparatus based on the algorithm GMDH may be helpful to solve the second of the diagnostics tasks, i.e. failure location. It is clear, that the advisory system for failure location requires greater number of data than for identification of the object state, if both tasks are to be solved with the same accuracy.

In the technical diagnostic, the object state is characterized by three groups of parameters:

i) input quantities (independent variables) forcing the reaction (action) of the object,

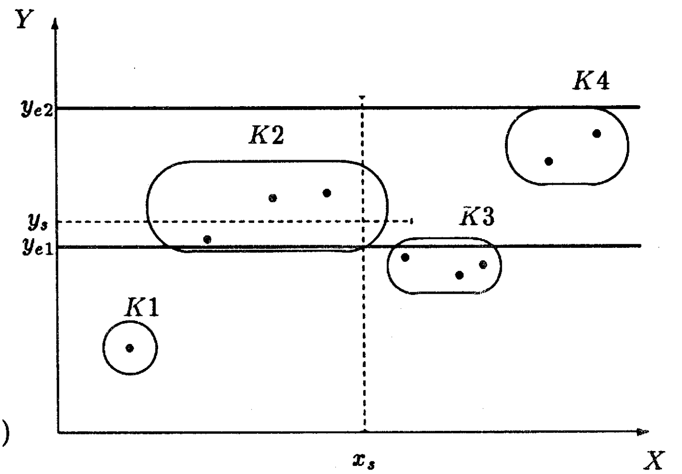
ii) output quantities (dependent variables) describing the action of the object,

iii) internal parameters (dependent variables) describing the internal state of the object or the states of its particular components.

The second group of parameters most frequently comprises information essential to the state identification, whereas the third group – information indispensable for failure location.

a)

Parameter	Variation interval	Interpretation
y	$y < y_{e1}$	emergency state 1
	$y_{e1} < y < y_{e2}$	normal state
	$y_{e2} < y$	emergency state 2



b)

Fig. 9. The example of the data supplementing knowledge base of computeraided technical diagnostic system: a) data supplementing knowledge base, b) graphic interpretation.

Thus, for the inference procedure, to fulfil all the functions connected with the solutions of diagnostic tasks, all the three groups of parameters must be observed and recorded during the knowledge base creation. While the object functioning and diagnosing, usually only input and output quantities are observed.

The principle of the advisory system operation, based on the Group Method of Data Handling is to make use of the potential of the precise approximation. The sequence of actions carried out during failure detection and location is as follows:

1. Diagnostics of the selected object parameters.
2. Realization of approximation GMDH procedure to define the (internal) parameter values not observed during diagnostics.
3. Identification of the object state.
4. If the emergency state has been not indicated, the diagnostic procedure ends. Else go to Step 5.
5. Considering input quantities as data for the algorithm GMDH, calculate the other parameters of a normal state.
6. Considering input and output quantities as data for the algorithm GMDH, calculate the other part of the emergency state parameters.

7. Comparing the calculated vectors of the normal and emergency state, find the values indicating the failure point, which differ.

The steps from 1 to 4 have been discussed and illustrated with the example, thus do not require further explanations. It is worth analyzing Steps 5, 6 and 7, which describe the functioning of the inference procedure from the moment of a failure detection. The knowledge base comprises information about the normal object functioning. Therefore, if only values of input parameters (not including information about failure) are inserted as data for the algorithm GMDH (Step 5 of the inference procedure), then as a result, the other quantities for the state, which should be observed, if the emergency does not occur, will be calculated.

The emergency state has been indicated on the basis of the output quantity analysis. Thus, if the activities mentioned are repeated for input and output quantities as data for the algorithm GMDH (Step 6), then the object parameters in emergency state will be calculated. Comparing both approximated vectors of state (Step 7), it is easy to find these internal parameters, which have other values in normal state and other in emergency state. These parameters indicate the failure point.

The described procedure may not provide satisfactory accuracy, if the number of internal parameters is much greater than the number of input and output quantities. Then, to make a precise diagnosis one should also observe some internal parameters in an emergency state.

7. Technical Diagnostic Advisory System

The operation of the advisory system described in this work will be illustrated by the following example. Let the object considered be full–wave rectifier circuit presented in Figure 10.

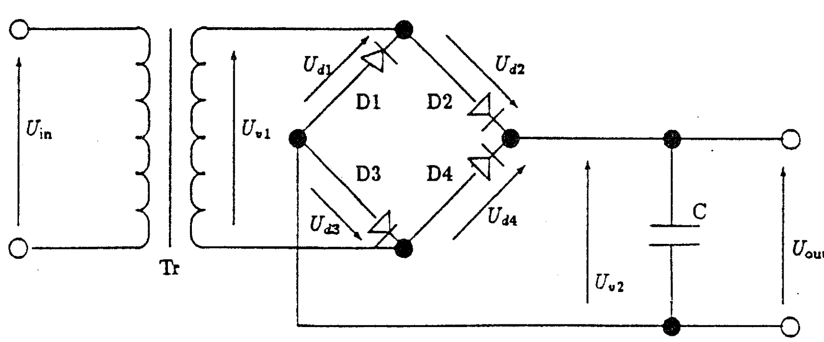


Fig. 10. The diagram of a full–wave rectifier in the Graetz bridge circuit.

To set up a knowledge base, the following parameters are observed:

– input quantities:

U_{in} is variable voltage on the terminals of input rectifier,

– output quantities:
    U_{out} is direct voltage on the terminals of output rectifier,
– internal quantities:
    U_{v1} is variable voltage on the secondary winding of the transformer Tr,
    U_{v2} is variable voltage on the terminals on the capacitor C,
    U_{d1} is direct voltage on the terminals of the diode D1,
    U_{d2} is direct voltage on the terminals of the diode D2,
    U_{d3} is direct voltage on the terminals of the diode D3,
    U_{d4} is direct voltage on the terminals of the diode D4,

During the observation of the rectifier work, the following data have been read and recorded in the knowledge base:

U_{in}	U_{out}	U_{v1}	U_{v2}	U_{d1}	U_{d2}	U_{d3}	U_{d4}
$[V]_{AC}$	$[V]_{DC}$	$[V]_{AC}$	$[V]_{AC}$	$[V]_{AC}$	$[V]_{DC}$	$[V]_{DC}$	$[V]_{DC}$
218	14.17	10.8	0.49	6.44	6.44	6.44	6.44
218.5	14.2	10.83	0.49	6.46	6.44	6.44	6.44
219	14.23	10.85	0.5	6.47	6.47	6.47	6.47
219.5	14.27	10.88	0.5	6.49	6.49	6.49	6.49
220	14.3	10.9	0.5	6.5	6.5	6.5	6.5
220.5	14.33	10.92	0.5	6.51	6.51	6.51	6.51
221	14.36	10.95	0.5	6.53	6.53	6.53	6.53
221.5	14.4	10.97	0.5	6.54	6.54	6.54	6.54
222	14.43	10.99	0.51	6.56	6.56	6.56	6.56
222.5	14.46	11.02	0.51	6.57	6.57	6.57	6.57
223	14.5	11.05	0.51	6.59	6.59	6.59	6.59

The knowledge base has been completed with the information concerning the interpretation of the rectifier state:

Parameter	Variation interval	Interpretation
U_{out}	$U_{out} < 14.0V$	Emergency state: too low output voltage
	$14.0V \leq U_{out} \leq 14.5V$	Normal state
	$14.5V < U_{out}$	Emergency state: too high output voltage

The diagnostics of the rectifier is based on the measurement of the values U_{in} and U_{out}. During the diagnostic test the following states has been indicated: $U_{in} = 219V$, and $U_{out} = 12.1V$.

The value U_{out} fulfills the condition $U_{out} < 13.7V$, therefore the emergency state has been indicated: *too low output voltage.*

As a result the inference procedure has begun, giving the following results:

1. Normal state approximation:
input data: $U_{\text{in}} = 219V$,
approximated normal state:

U_{in}	U_{out}	U_{v1}	U_{v2}	U_{d1}	U_{d2}	U_{d3}	U_{d4}
$[V]_{AC}$	$[V]_{DC}$	$[V]_{AC}$	$[V]_{AC}$	$[V]_{AC}$	$[V]_{DC}$	$[V]_{DC}$	$[V]_{DC}$
219	14.26	10.8	0.49	6.45	6.45	6.45	6.45

2. Emergency state approximation:
input data: $U_{\text{in}} = 219V$, and $U_{\text{out}} = 12.1V$,
approximated emergency state:

U_{in}	U_{out}	U_{v1}	U_{v2}	U_{d1}	U_{d2}	U_{d3}	U_{d4}
$[V]_{AC}$	$[V]_{DC}$	$[V]_{AC}$	$[V]_{AC}$	$[V]_{AC}$	$[V]_{DC}$	$[V]_{DC}$	$[V]_{DC}$
219	12.1	10.83	1.64	6.43	6.43	6.43	6.43

The comparison of both approximating state vectors shows that the greatest (and the only essential) difference of the internal parameters value is for $U_{v2}(U_{v2\text{emerg}} > U_{v2\text{norm}})$. Other internal parameters do not reveal differences indicating an emergency (abnormal) state.

Following the obtained results of the calculations, the advisory system issued a diagnosis: *In the rectifier considered, the emergency state has been indicated: too low output voltage. This state is the result of a failure: too high value of variable voltage on the terminals of the capacitor* C. This enables, in the considered example, to point out the damaged element, if it is connected with the rule (provided by an expert or an expert system): *if the variable voltage on the terminals of the capacitor* C *is too high and the other internal voltages are correct, then the capacitor* C *is damaged.*

The presented example shows the potential of the computer–aided technical diagnostics and, in particular, automatic failure location. There are examples of more complicated objects (here electric circuits), where fast indication of the incorrect values of the internal parameters would require laborious execution of a series of measurements, and then their analysis. Meanwhile, the advisory system applying the approximation methods GMDH, used as an aid by a diagnostic expert (or an expert system), enables fast and simple analysis of freely complicated objects, and an easy failure location. All the calculations have been made by system EVALUATOR (version 2.2), which is the computer implementation of the described algorithm GMDH (Kuś, 1990; Kuś, 1992).

8. Summary

This work has presented a certain concept of the advisory system based on the Group Method of Data Handling, which belongs to the class of genetic algorithms. Apart from the precise description of the knowledge base algorithms and the approximation GMDH, the methods of the organisation of the above mentioned procedures in the computer–aided technical diagnostics have been presented. In the examples, the method of the automatic solution of the detection tasks and failure location has been illustrated.

The most important advantages of system presented are:

- the ability to consider a wide class of complex technical objects,
- easy creation of knowledge base, not requiring the knowledge of rules of the whole object functioning, and its particular components,
- fast and automatic detection and location of a failure,
- the ability to support the work of experts and expert systems.

The application of the GMDH method brings additional advantages, connected with the ability of the mathematical model automatic construction of the freely complex real objects. The analysis of the models may turn out to be useful outside the technical diagnostic area. This problem however, is not the subject of this work. The readers interested in it should study the enclosed list of literature.

The proposed solution of the advisory system has some shortcomings. The basic one is the inability to direct indication of the damaged components, without expert knowledge (or expert systems). Due to this, the presented system has been called advisory, because it can carry out only such functions.

It is also important, that the use of the approximation GMDH is connected with the necessity to define supplementary procedures (as e.g. the algorithm of clustering, the method of the estimating cluster selection, the method of the observations division into learning and testing, the selection of the estimation quality evaluation method, etc.), which have an important influence on the final solution.

It is worth mentioning, that in case of complicated technical objects, which consist of many components, the ability of the fast state analysis or the indication of the failure point often turn out to be a sufficient help to find the damaged element.

The computer experiments carried out confirm the efficiency of the described concepts in the technical diagnostic tasks. It has been confirmed, that there are prospects for the research direction connected with the application of the genetic algorithms associated with the methods of the knowledge base linguistic concept, to diagnose large technical objects.

References

Banaszak Z. and Kobyłecki G. (1992): *Two–step structural identyfication method.* — Zeszyty Naukowe Politechniki Śląskiej, Gliwice, No.109, pp.135–145, (in Polish).

Bijnen E.I. (1974): *Cluster Analysis.* — Berlin: Springer Verlag.

Farlow S.J. (Ed.) (1984): *Self Organizing Methods in Modelling: GMDH–type Algorithms.* — New York: Mersel Dekker, Inc.

Gleichgewicht B. (1983): *Algebra.* — Warszawa: PWN Press, (in Polish).

Ivachnenko A.G. (1990): *Continuity and Discreteness.* — Kiev: Naukova Dumka, (in Russian).

Ivachnenko A.G. (1987): *Objective clustering on the ground of self–organized models theory.* — Avtomatika, No.5, (in Russian).

Ivachnenko A.G., Pietuchova S.A. and Ivachnenko N.A. (1986): *Objective computer clustering. Part 1: Theoretical questions.* — Avtomatika, No.3, (in Russian).

Ivachnenko A.G., Semina L.P. and Ivachnenko N.A. (1987): *Objective computer clustring. Part 2: Record of goal–function information for clusters number decriease.* — Avtomatika, No.1, (in Russian).

Ivachnenko A.G., Tatiejeva E.H. and Ivachnenko N.A. (1989): *Nonparametrical prognostic models GMDH.* — Avtomatika, No.1, (in Russian).

Kucharczyk J. (1982): *Concentration Analysis Algorithm in Algol 60 Language.* — Warsaw: PWN Press, (in Polish).

Kuś J. (1990): *System for computer aided decision making in the tasks of design and utilization of flexible manufacturing systems.* — Report ICT PWr, ser. PRE No.44/90, Technical University of Wroclaw, Poland, (in Polish).

Kuś J. (1992): *Advisory - prognostic system "Evaluator".* — Report No.1/92, Technical University of Zielona Góra, Dept. of Robotics and Software Engngn., Poland, (in Polish).

Received March 25, 1993
Revised June 17, 1993