INDUCTIVE MODELING METHOD GMDH IN THE PROBLEMS OF DATA MINING

Yuriy Zaychenko, Galib Hamidov

Abstract: *The problem of constructing unknown dependencies (laws) in huge data warehouses is considered. For its solution inductive modeling method- so called Group Method of Data Handling (GMDH) is suggested. This method enables to construct automatically optimal models of variables based on experimental data stored in data warehouses. Unlike other modeling methods GMDH enables to find out the structure of the unknown model and solves the problem of not parametric, but structural identification. Additionally for finding unknown laws in incomplete and unreliable data under uncertainty fuzzy GMDH is suggested enabling to construct fuzzy models. The experimental investigations of the suggested methods for models identification in Data Mining problems are presented and the obtained results discussed.*

Keywords: *Data Mining, GMDH, fuzzy, model identification.*

ITHEA Keywords*: I. Computing methodologies; I 2. Artificial intelligence, I 6.5. Model development*

Introduction

Last year's problems of Data Mining in data bases (DB) have become very crucial in IT-applications. Especially it refers to big DB, so-called data warehouses where mountains of raw data are accumulated and hidden laws in these data are to be detected and corresponding models to be constructed [Barsegyan, 2008; Duke, 2001].

Previously several classes of methods were developed for finding unknown dependencies in data, in particularly statistical methods: ARMA, Logit and Probit models, ARCH and GARCH methods and neural networks. But they have drawbacks: statistical methods solve only problems of parametric identification and don't solve structural identification while neural networks allow to determine model structure but in an implicit form. The model structure is hidden in neural weights and its analytical form is unavailable.

Therefore the development of methods for structural models identification constitute important problem in DM. The main goal of this paper is development and investigation of methods for constructing models in data accumulated in data warehouses. For this goal the method of inductive modeling- Group Method of Data Handling (GMDH) is suggested and investigated [Ivakhnenko, 1985].

For finding unknown laws in data under uncertainty new version of GMDH – Fuzzy GMDH is suggested and explored [Zaychenko, 2003; Zaychenko, 2006]. Fuzzy GMDH enables to operate with incomplete or indefinite initial data and constructs fuzzy models whose coefficients are fuzzy.

The significant property of GMDH is that it may operate with high dimensional data (with many variables) and so-called "short samples" when the number of model coefficients m is greater than sample size N. This is achieved due to specificity of GMDG algorithm as at each step of it a set of so-called partial models are constructed consisting only of two variables instead of n initial input variables like other modeling methods. This enables to cut substantially the dimension of model and decrease the time for its construction. This advantage rises with the increase of model complexity: the greater is model dimension (number of variables), the greater is cut in computational time for its construction as compared with conventional modeling methods.

1. Problem Formulation

Consider the problem of model construction. A set of initial data is given, inclusive input variables $\{X(1), X(2), ..., X(N)\}$ and output variables $\{Y(1), Y(2), ..., Y(N)\}$, where $X = [x_1, x_2, ..., x_n]$ *is* -n-tuple vector, N is a number of observations .

The task is to synthesize an adequate forecasting model $Y = F(x_1, x_2, ..., x_n)$, and besides, the obtained model should have the minimal complexity. In particularly, while solving forecasting problem as an output variable Y a forecasting model is used $X(N+K) = f(X(1), ..., X(N))$, where K is a value of a forecasting interval.

The constructed model should be adequate according to the initial set of data, and should have the least complexity (Figure 1).

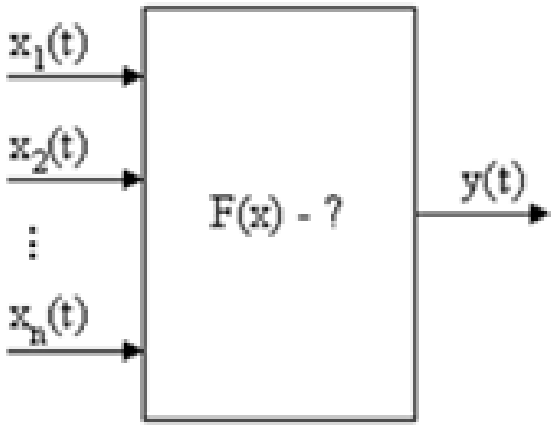

Figure 1. Graphical representation of the problem

The distinguishing features of the problem are the following:

1. Form of functional dependence is unknown and only model class is determined, for example, polynomial of any degree or Fourier time series.
2. Short data samples;
3. Time series $x_i(t)$ in general case is non- stationary.

In this case the application of conventional methods of statistical analysis (e.g. regression analysis) is impossible and it's necessary to utilize methods based on computational intelligence (CI). To this class belongs Group Method of Data Handling (GMDH) developed by acad. A. Ivakhnenko [Ivakhnenko, 1985] and extended by his colleges. GMDH is a method of inductive modeling. The method inherits ideas of biological evolution and its mechanisms:

1. Crossing-over of parents and offspring generation;
2. Selection of the best offsprings.

GMDH method belongs to self-organizing methods and allows to discover internal hidden laws in the appropriate object area.

The advantages of GMDH algorithms are the possibility of constructing optimal models with a small number of observations and unknown dynamics among variables. This method doesn't demand to know the model structure a priori, the model is constructed by algorithm itself in the process of its run.

The basic principles of GMDH

Let's remind the fundamental principles of GMDH [3, 4-6]. The full interconnection between input $X(i)$ and output $Y(i)$ in the class of polynomial models may be presented by so-called generalized polynomial of Kolmogorov- Gabor:

$$Y = a_0 + \sum_{i=1}^{n} a_i x_i + \sum_{j=1}^{n} \sum_{i \le j} a_{ij} x_i x_j \sum_{i=1}^{n} \sum_{j \le i} \sum_{k \le j} a_{ijk} x_i x_j x_k + \ldots \quad (1)$$

where all the coefficients $\bar{a}_0, \bar{a}_i, \bar{a}_{ij}$, are unknown.

While constructing model (search coefficients values) as a criterion of adequacy the so-called regularity criterion (mean squared error- MSE) is used

$$\overline{\varepsilon^2} = \frac{1}{N} \cdot \sum_{i=1}^{N} (y_i - f(X_i))^2 \qquad (2)$$

where *N* is a sample size (number of observations).

It' s demanded to find minimum $\overline{\varepsilon}^2$.

GMDH method is based on the following principles [Ivakhnenko, 1985, Zaychenko, 2003].

The principle of multiplicity of models. There is a great number of models providing zero error on a given sample . It's enough simply to raise the degree of the polynomial model. If N nodes of interpolation are available, then it's possible to construct the family of models each of which gives zero error on experimental points $\overline{\varepsilon}^2 = 0$.

The principle of self-organization. Denote S as model complexity. The value of an error depends on the complexity of a model. As the the level of complexity S grows the error first drops, attains minimum value and then begins to rise (see Fig. 2).

We need to find such level of complexity for which the error would be minimal. In addition if to take into account the action of noise we may make the following conclusions concerning ε:

1. With the increase of noise the optimal complexity $s_0 = \arg\min \overline{\varepsilon}^2$ shifts to the left;
2. With the increase of noise level the value of optimal criterion min $\overline{\varepsilon}^2(s)$ grows.

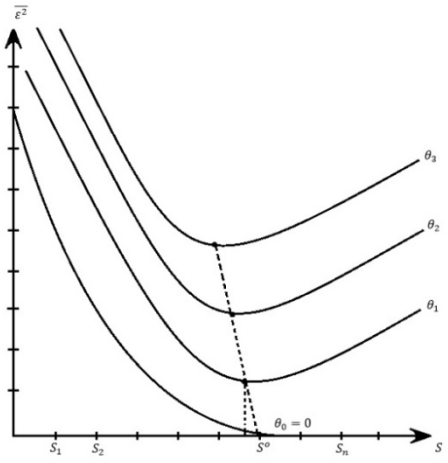

Figure 2. Dependence of criterion $\overline{\varepsilon}^2$ on model complexity S

Theorem of incompleteness by Geodel: In any formal logical system there are some statements which cannot be proved or refuted using the given system of axioms and staying in the margins of this system. That to prove or refute such statement one need go out this system and use some external information (meta information) which is called "external complement". In our case as external information stands additional sample of data which wasn't used for the finding unknown coefficients of the model.

So one way to overcome incompleteness of sample is to use **principle of external complement** which means that the whole sample should be divided into two parts – training subsample and test subsample. The search of optimal model is performed in such a way:

- At the training sample N_{train} the estimates $\bar{a}_0, \bar{a}_i, \bar{a}_{ij}$, are determined;
- At the test sample N_{test} the best model is selected.

The ideas of computational method GMDH

For each pair of inputs x_i and x_j so-called partial descriptions are being built (all in all C_n^2) of the form:

$$\bar{Y}_s = \phi(x_i, x_j) = a_0 + a_i x_i + a_j x_j \ , \ s = 1..C_n^2 \ \text{(linear)};$$
$$\text{or } \bar{Y}_s = \phi(x_i, x_j) = a_0 + x_i + a_j x_j + a_{ii} x_i^2 + a_{ij} x_i x_j + a_{jj} x_j^2 \ , \ s = 1..C_n^2 \ \text{(quadratic)}. \quad (3)$$

1. Determine the coefficients of these model using LSM (least square method) at the training sample (i.e. find estimates $\bar{a}_0, \bar{a}_1, ..., \bar{a}_j, ..., \bar{a}_N, \bar{a}_{11}, ..., \bar{a}_{ij}, ..., \bar{a}_{NN}$.
2. Further at the test sample for each of these models calculate the value of regularity criterion :

$$\bar{\delta}_s^2 = \frac{1}{N_{test}} \cdot \sum_{i=1}^{N_{testx}} [Y(k) - \bar{Y}_s(k)]^2 \quad (4)$$

(where $Y(k)$ is real output value of the k-th point of test; $\bar{Y}_s(k)$ is a value of this criterion on k-th point obtained by model, N_{test} is a number of points at the test sample);

as alternate criterion "unbiasedness" criterion may be used:

$$N_{ub} = \frac{1}{N_1 + N_2} \sum_{k=1}^{N} \left(y_k^* - y_k^{**} \right)^2 \quad (5)$$

where the sample is also divided in two parts N_1 and N_2, y_k^{**} are outputs of the model built on the subsample N_1, y_k^{**} are outputs of model built on subsample N_2, $N = N_1 + N_2$.

3. Determine F (this number is called a freedom of choice) best models using one of these criteria. The selected models y_i are then transferred to the second row of model construction. We search coefficients of new partial descriptions:

$$z_l = \phi^{(2)}(x_i, x_j) = a_0^{(2)} + a_1^{(2)} y_i + a_2^{(2)} y_j + a_3^{(2)} y_i^2 + a_4^{(2)} y_i y_j + a_5^{(2)} y_j^2$$

The process at the second row runs in the same way . The selection of the best models is carried out similarly, but $F_2 < F_1$. The process of rows construction repeats more and more till MSE (regularity criterion) falls. If at the m-th layer occurs the increase of the error $\bar{\varepsilon}^2$ the algorithm stops. In this case find the best model at the preceding layer and then moving backward by its connections find models of preceding layer and successfully passing all the used connections at the end we'll reach the first layer and find the analytical form of the optimal model (with minimal complexity).

2. Fuzzy GMDH. Principal ideas. Interval model of regression

Classical GMDH has some drawbacks:

1. GMDH utilizes least squared method (LSM) for finding the model coefficients but matrix of linear equations may be close to degenerate and the corresponding solution may appear non-stable and very volatile. Therefore, the special methods for regularization should be used;

2. after application of GMDH point-wise estimations are obtained but in many cases it's needed find interval value for coefficient estimates;

3. GMDH doesn't work in case of incomplete or fuzzy input data.

Therefore, in last 10 years the new variant of GMDH – fuzzy GMDH was developed and refined which may work with fuzzy input data and is free of classical GMDH drawbacks [Zaychenko, 2003; Zaychenko, 2006].

In works [Zaychenko, 2003; Zaychenko, 2006] the linear interval model regression was considered :

$$Y = A_0 Z_0 + A_1 Z_1 + \ldots + A_n Z_n \tag{6}$$

where A_i is a fuzzy number of triangular form described by pair of parameters $A_i = (\alpha_i, c_i)$, where α_i is interval center, c_i is its width, $c_i \geq 0$

Then *Y* is a fuzzy number, parameters of which are determined as follows:

the interval center

$$\alpha_y = \sum \alpha_i z_i = \alpha^T \cdot z, \tag{7}$$

the interval width

$$c_y = \sum c_i \cdot |z_i| = c^T |z|. \tag{8}$$

In order the interval is correct it's necessary that real value of output should belong to the interval of uncertainty described by the following constraints:

$$\begin{cases} \alpha^T z - c^T \cdot |z| \leq y \\ \alpha^T z + c^T \cdot |z| \geq y \end{cases}. \tag{9}$$

For example, for the partial description of the kind

$$f(x_i, x_j) = A_0 + A_1 x_i + A_2 x_j + A_3 x_i x_j + A_4 x_i^2 + A_5 x_j^2 \tag{10}$$

it's necessary to assign in the general model (6)

$$z_0 = 1, \; z_1 = x_i \; z_2 = x_j \; z_3 = x_i x_j \; z_4 = x_i^2 \; z_5 = x_j^2$$

Let the training sample be $\{z_1, z_2, ..., z_M\}$, $\{y_1, y_2, ..., y_M\}$. Then for the model (10) to be adequate it's necessary to find such parameters (α_i, c_i) $i = \overline{1,n}$, which satisfy the following inequalities:

$$\begin{cases} \alpha^T z_k - c^T \cdot |z_k| \le y_k \\ \alpha^T z_k + c^T \cdot |z_k| \le y_k \end{cases}, \quad k = \overline{1,M} . \tag{11}$$

Let's formulate the basic requirements for the linear interval model of partial description of a kind (10). It's necessary to find such values of the parameters (α_i, c_i) of fuzzy coefficients for which:

1. Real values of the observed outputs y_k would drop in the estimated interval for Y_k;
2. The total width of the estimated interval for all sample points would be minimal.

These requirements lead to the following linear programming problem:

$$\min(C_0 \cdot M + C_1 \sum_{k=1}^{M} |x_{ki}| + C_2 \sum_{k=1}^{M} |x_{kj}| + C_3 \sum_{k=1}^{M} |x_{ki} x_{kj}| + C_4 \sum_{k=1}^{M} |x_{ki}^2| + C_5 \sum_{k=1}^{M} |x_{kj}^2|, \tag{12}$$

under constraints:

$$\begin{aligned} &a_0 + a_1 x_{ki} + a_2 x_{kj} + a_3 x_{ki} x_{kj} + a_4 x_{ki}^2 + a_5 x_{kj}^2 - (C_0 + C_1 |x_{ki}| + C_2 |x_{kj}| + \\ &+ C_3 |x_{ki} x_{kj}| + C_4 |x_{ki}^2| + C_5 |x_{kj}^2|) \le \mathrm{y_k} \end{aligned} \tag{13}$$

$$\begin{aligned} &a_0 + a_1 x_{ki} + a_2 x_{kj} + a_3 x_{ki} x_{kj} + a_4 x_{ki}^2 + a_5 x_{kj}^2 + (C_0 + C_1 |x_{ki}| + C_2 |x_{kj}| + \\ &+ C_3 |x_{ki} x_{kj}| + C_4 |x_{ki}^2| + C_5 |x_{kj}^2|) \ge \mathrm{y_k} \\ &k = \overline{1,M} , \\ &C_p \ge 0, \quad p = 0,5 , \end{aligned} \tag{14}$$

where *k* is an index of a point.

As we can easily see the task (12) – (14) is linear programing (LP) problem. However, the inconvenience of the model (12) – (14) for the application of standard LP methods is that there are no constraints of non- negativity for variables α_i. Therefore for its solution it's reasonable to pass to the dual LP problem by introducing dual variables $\{\delta_k\}$ and $\{\delta_{k+M}\}$, $k = \overline{1,M}$. Using simplex- method for the dual problem and after finding the optimal values for the dual variables $\{\delta_k\}$ the optimal solutions (α_i, c_i) of the initial direct problem will be also found [Zaychenko, 2003; Zaychenko, 2006].

3. FGMDH with fuzzy input data for triangular membership functions

The generalization and further development of the considered FMGH is Fuzzy GMDH where fuzzy are not only model coefficients but input data as well. Below the correspondent mathematical model is presented. [Zaychenko, 2008]

3.1. The form of math model for triangular MF

Let's consider the linear interval regression model with fuzzy inputs which generalies the model (6) :

$$Y = A_0 Z_0 + A_1 Z_1 + \ldots + A_n Z_n, \tag{15}$$

where A_i – fuzzy number of triangular shape, which is described by threes of parameters $A_i = (\underline{A_i}, a_i, \overline{A_i})$, where a_i – center of the interval, $\overline{A_i}$ – its upper border, $\underline{A_i}$ - its lower border.

Current task contains the case of symmetrical membership function for parameters A_i, so they can be described via pair of parameters (a_i, c_i).

$$\underline{A_i} = a_i - c_i,\ \overline{A_i} = a_i + c_i,\ c_i \text{ – interval width, } c_i \geq 0,$$

Z_i – also fuzzy numbers of triangular shape, which are defined by parameters $(\underline{Z_i}, \breve{Z}_i, \overline{Z_i})$, $\underline{Z_i}$ - lower border, $\breve{Z}_i$ - center, $\overline{Z_i}$ - upper border of fuzzy number.

Then Y – fuzzy number, which parameters are defined as follows:

Center of the interval:

$$\breve{y} = \sum a_i * \breve{Z}_i,$$

Deviation in the left part of the membership function:

$$\breve{y} - \underline{y} = \sum (a_i * (\breve{Z}_i - \underline{Z_i}) + c_i \left|\breve{Z}_i\right|), \text{ thus}$$

Lower border of the interval:

$$\underline{y} = \sum(a_i * \underline{Z}_i - c_i \left|\breve{Z}_i\right|) \tag{16}$$

Deviation in the right part of the membership function:

$$\overline{y} - \breve{y} = \sum(a_i * (\overline{Z}_i - \breve{Z}_i) + c_i \left|\breve{Z}_i\right|) = \sum a_i \overline{Z}_i - a_i \breve{Z}_i + c_i \left|\breve{Z}_i\right| \text{, so}$$

Upper border of the interval:

$$\overline{y} = \sum(a_i * \overline{Z}_i + c_i \left|\breve{Z}_i\right|) \tag{17}$$

For the interval model to be correct, the real value of input variable Y should lay in the interval got by the method workflow.

It can be described in such a way:

$$\begin{cases} \sum(a_i * \underline{Z}_{ik} - c_i \left|\breve{Z}_{ik}\right|) \le y_k \\ \sum(a_i * \overline{Z}_{ki} + c_i \left|\breve{Z}_{ik}\right|) \ge y_k, k = \overline{1,M} \end{cases} \tag{18}$$

Where Z_k=[Z_k]i is input training sample, y_k –known output values, $k = \overline{1,M}$, M – number of observation points.

So, the general requirements to estimation linear interval model are to find such values of parameters (a_i, c_i) of fuzzy coefficients, which enable:

a) Observed values y_k lay in estimation interval for Y_k;

b) Total width of estimation interval is minimal.

These requirements can be redefined as a task of linear programming:

$$\min_{a_i, c_i} \sum_{k=1}^{M} (\sum(a_i * \overline{Z}_i + c_i \left|\breve{Z}_i\right|) - \sum(a_i * \underline{Z}_i - c_i \left|\breve{Z}_i\right|)) \tag{19}$$

under constraints:

$$\begin{cases} \sum(a_i * \underline{Z}_{ik} - c_i \left|\breve{Z}_{ik}\right|) \le y_k \\ \sum(a_i * \overline{Z}_{ki} + c_i \left|\breve{Z}_{ik}\right|) \ge y_k, k = \overline{1,M} \end{cases} \tag{20}$$

3.2. Formalized problem formulation in case of triangular membership functions

Let's consider partial description

$$f(x_i, x_j) = A_0 + A_1 x_i + A_2 x_j + A_3 x_i x_j + A_4 x_i^2 + A_5 x_j^2 \tag{21}$$

Then math model (19)-(20) takes the form

$$\begin{aligned}\min_{a_i, c_i} \; & (2Mc_0 + a_1 \sum_{k=1}^{M} (\bar{x}_{ik} - \underline{x}_{ik}) + 2c_1 \sum_{k=1}^{M} |\breve{x}_{ik}| + a_2 \sum_{k=1}^{M} (\bar{x}_{jk} - \underline{x}_{jk}) + 2c_2 \sum_{k=1}^{M} |\breve{x}_{jk}| + \\ & + a_3 \sum_{k=1}^{M} (|\breve{x}_{ik}| (\bar{x}_{jk} - \underline{x}_{jk}) + |\breve{x}_{jk}| (\bar{x}_{ik} - \underline{x}_{ik})) + 2c_3 \sum_{k=1}^{M} |\breve{x}_{ik} \breve{x}_{jk}| + 2a_4 \sum_{k=1}^{M} |\breve{x}_{ik}| (\bar{x}_{ik} - \underline{x}_{ik}) + \\ & + 2c_4 \sum_{k=1}^{M} \breve{x}_{ik}^2 + 2a_5 \sum_{k=1}^{M} |\breve{x}_{jk}| (\bar{x}_{jk} - \underline{x}_{jk}) + 2c_5 \sum_{k=1}^{M} \breve{x}_{jk}^2)\end{aligned} \tag{22}$$

with the following conditions:

$$\begin{aligned} & a_0 + a_1 \underline{x}_{ik} + a_2 \underline{x}_{jk} + a_3 (-|\breve{x}_{ik}| (\breve{x}_{jk} - \underline{x}_{jk}) - |\breve{x}_{jk}| (\breve{x}_{ik} - \underline{x}_{ik}) + \breve{x}_{ik} \breve{x}_{jk}) + \\ & + a_4 (-2 |\breve{x}_{ik}| (\breve{x}_{ik} - \underline{x}_{ik}) + \breve{x}_{ik}{}^2) + a_5 (2 |\breve{x}_{jk}| (\breve{x}_{jk} - \underline{x}_{jk}) + \breve{x}_{jk}{}^2) - c_0 - c_1 |\breve{x}_{ik}| - \\ & - c_2 |\breve{x}_{jk}| - c_3 |\breve{x}_{ik} \breve{x}_{jk}| - c_4 \breve{x}_{ik}^2 - c_5 \breve{x}_{jk}^2 \le y_k \\ & a_0 + a_1 \bar{x}_{ik} + a_2 \bar{x}_{jk} + a_3 (|\breve{x}_{ik}| (\bar{x}_{jk} - \breve{x}_{jk}) + |\breve{x}_{jk}| (\bar{x}_{ik} - \breve{x}_{ik}) - \breve{x}_{ik} \breve{x}_{jk}) + a_4 (2 |\breve{x}_{ik}| (\bar{x}_{ik} - \\ & - \breve{x}_{ik}) - \breve{x}_{ik}{}^2) + a_5 (2 |\breve{x}_{jk}| (\bar{x}_{jk} - \breve{x}_{jk}) - \breve{x}_{jk}{}^2) + c_0 + c_1 |\breve{x}_{ik}| + c_2 |\breve{x}_{jk}| + c_3 |\breve{x}_{ik} \breve{x}_{jk}| + \\ & c_4 \breve{x}_{ik}^2 + c_5 \breve{x}_{jk}^2 \ge y_k \end{aligned} \tag{23}$$

$$c_l \ge 0, \; l = \overline{0,5}.$$

As we can see, this is the linear programming problem, like the problem (12)-(13)for non-fuzzy inputs but there are still no limitations for non-negativity of variables a_i, so we need go to dual problem, introducing dual variables $\{\delta_k\}$ and $\{\delta_{k+M}\}$.

Write down dual problem:

$$\max\left(\sum_{k=1}^{M} y_k \cdot \delta_{k+M} - \sum_{k=1}^{M} y_k \cdot \delta_k\right) \tag{24}$$

Under constraints:

$$\begin{aligned} & \sum_{k=1}^{M} \delta_{k+M} - \sum_{k=1}^{M} \delta_k = 0 \\ & \sum_{k=1}^{M} \bar{x}_{ik} \cdot \delta_{k+M} - \sum_{k=1}^{M} \underline{x}_{ik} \cdot \delta_k = \sum_{k=1}^{M} (\bar{x}_{ik} - \underline{x}_{ik}) \end{aligned} \tag{25}$$

$$\sum_{k=1}^{M} \bar{x}_{jk} \cdot \delta_{k+M} - \sum_{k=1}^{M} \underline{x}_{jk} \cdot \delta_k = \sum_{k=1}^{M} (\bar{x}_{jk} - \underline{x}_{jk})$$

$$\sum_{k=1}^{M} (|\breve{x}_{ik}|(\bar{x}_{jk} - \breve{x}_{jk}) + |\breve{x}_{jk}|(\bar{x}_{ik} - \breve{x}_{ik}) - \breve{x}_{ik}\breve{x}_{jk}) \cdot \delta_{k+M} -$$
$$-\sum_{k=1}^{M} (-|\breve{x}_{ik}|(\breve{x}_{jk} - \underline{x}_{jk}) - |\breve{x}_{jk}|(\breve{x}_{ik} - \underline{x}_{ik}) + \breve{x}_{ik}\breve{x}_{jk}) \cdot \delta_k =$$
$$= \sum_{k=1}^{M} (|\breve{x}_{ik}|(\bar{x}_{jk} - \underline{x}_{jk}) + |\breve{x}_{jk}|(\bar{x}_{ik} - \underline{x}_{ik}))$$

$$\sum_{k=1}^{M} (2|\breve{x}_{ik}|(\bar{x}_{ik} - \breve{x}_{ik}) - \breve{x}_{ik}^{\,2}) \cdot \delta_{k+M} - \sum_{k=1}^{M} (-2|\breve{x}_{ik}|(\breve{x}_{ik} - \underline{x}_{ik}) + \breve{x}_{ik}^{\,2}) \cdot \delta_k = \sum_{k=1}^{M} |\breve{x}_{ik}|(\bar{x}_{ik} - \underline{x}_{ik})$$

$$\sum_{k=1}^{M} (2|\breve{x}_{jk}|(\bar{x}_{jk} - \breve{x}_{jk}) - \breve{x}_{jk}^{\,2}) \cdot \delta_{k+M} - \sum_{k=1}^{M} (-2|\breve{x}_{ik}|(\breve{x}_{jk} - \underline{x}_{jk}) + \breve{x}_{jk}^{\,2}) \cdot \delta_k = \sum_{k=1}^{M} |\breve{x}_{jk}|(\bar{x}_{jk} - \underline{x}_{jk})$$

$$\sum_{k=1}^{M} \delta_{k+M} + \sum_{k=1}^{M} \delta_k \le 2M$$

$$\sum_{k=1}^{M} |\breve{x}_{ik}| \cdot \delta_{k+M} + \sum_{k=1}^{M} |\breve{x}_{ik}| \cdot \delta_k \le 2\sum_{k=1}^{M} |\breve{x}_{ik}|$$

$$\sum_{k=1}^{M} |\breve{x}_{jk}| \cdot \delta_{k+M} + \sum_{k=1}^{M} |\breve{x}_{jk}| \cdot \delta_k \le 2\sum_{k=1}^{M} |\breve{x}_{jk}|$$

$$\sum_{k=1}^{M} |\breve{x}_{ik}\breve{x}_{jk}| \cdot \delta_{k+M} + \sum_{k=1}^{M} |\breve{x}_{ik}\breve{x}_{jk}| \cdot \delta_k \le 2\sum_{k=1}^{M} |\breve{x}_{ik}\breve{x}_{jk}|$$

$$\sum_{k=1}^{M} \breve{x}_{ik}^2 \cdot \delta_{k+M} + \sum_{k=1}^{M} \breve{x}_{ik}^2 \cdot \delta_k \le 2\sum_{k=1}^{M} \breve{x}_{ik}^2 \qquad (26)$$

$$\sum_{k=1}^{M} \breve{x}_{jk}^2 \cdot \delta_{k+M} + \sum_{k=1}^{M} \breve{x}_{jk}^2 \cdot \delta_k \le 2\sum_{k=1}^{M} \breve{x}_{jk}^2$$

$$\delta_k \ge 0 \ , \delta_{k+M} \ge 0 \ , \ k = \overline{1,M} \qquad (27)$$

The task (24)-(27) can be solved using simplex-method. Having optimal values of dual variables $\{\delta_k\}$, $\{\delta_{k+M}\}$, we easily obtain the optimal values of desired variables c_i, a_i, $i = \overline{0,5}$, and also a desired fuzzy model for given partial description.

4. The description of fuzzy algorithm GMDH

Let's present the brief description of the algorithm FGMDH [Zaychenko, 2006].

1. Choose the general model type by which the sought dependence will be described.
2. Choose the external criterion of optimality (criterion of regularity or non --biasedness).
3. Choose the type of partial descriptions (for example, linear or quadratic one).
4. Divide the sample into training N_{train} and test N_{test} subsamples.
5. Put zero values to the counter of model number k and to the counter of rows r (iterations number).
6. Generate a new partial model f_k (10) using the training sample. Solve the LP problem (12) – (14) or (22)-(23) and find the values of parameters α_i , c_i .
7. Calculate using test sample the value of external criterion ($N_{ubk}^{(r)}$ or $\delta_k^{(2)}(r)$).
8. $k = k + 1$. If $k > C_N^2$ for r=1or $k > C_F^2$ for r>1, then $k = 1$, $r = r + 1$ and go to step 9, otherwise go to step 6.
9. Calculate the best value of the criterion for models of r-th iteration. If $r = 1$, then select F best models and assigning $r = r + 1$, $k = 1$, go to step 6 and execute (r+1)-th iteration otherwise, go to step 10.
10. If $\left|N_{ub}(r) - N_{ub}(r-1)\right| \le \varepsilon$ or $\delta_k^{(2)}(r) \ge \delta_{k-1}^{(2)}(r)$, then go 11,
11. Otherwise select F best models and assigning $r = r + 1$, $k = 1$, go to step 6 and execute (r+1) iteration.
12. Select the best model out of models of the previous row (iteration) using external criterion.

Starting from this model and moving backward by its connection to the models of previous row and successively passing the models of all previous rows by corresponding connections at the last step reach the models of the first row. Having made corresponding reverse substitutions of variables we find the final best model in initial variables $Y = F(x_1, x_2, ..., x_n)$.

Thus, fuzzy GMDH allows to construct fuzzy models and has the following advantages:

1. The problem of ill- conditionality of matrix of normal equalities is absent in fuzzy GMDH unlike classic GMDH as the least squared method isn't used for optimal model determination. The problem of optimal model determination is transferred to the problem of linear programming, which is always solvable.
2. There is interval regression model built as the result of method work unlike GMDH which constructs point-wise models. And the interval width enables to estimate the accuracy of the found model.
3. There is a possibility of the obtained model adaptation.

5. The application of GMDH for forecasting at the stock exchange

For estimation of efficiency of the suggested FGMDH method with non-fuzzy and fuzzy inputs the corresponding software kit was elaborated and numerous experiments of financial markets forecasting were carried out. For the experiments the stock prices of different shares at the Stock exchange "RTS" were chosen. Some of them are presented below

Experiment 1. RTS-2 index forecasting (opening price)

There were 5 fuzzy input variables in this experiment; they were price on shares of "second echelon" Russian energetic companies, which are included to RTS-2 index computation list:

BANE – shares of "Башнефть" joint-stock company,

ENCO – shares of "Сибирьтелеком" joint-stock company,

ESMO – shares of "ЦентрТелеком" joint-stock company,

IRGZ – shares of "Иркутскэнерго" joint-stock company,

KUBN – shares of "Южтелеком" joint-stock company.

Output variable is the value of RTS-2 index (opening price) for the same period (03.04.2006 – 18.05.2006).

Sample size – 32 values.

Training sample size – 19 values (optimal size of training sample for current experiment).

The following results were obtained:

1. **For triangular membership function**

   Criterion for this experiment was MSE=0,061787

   The corresponding results are presented at the Figure 3.

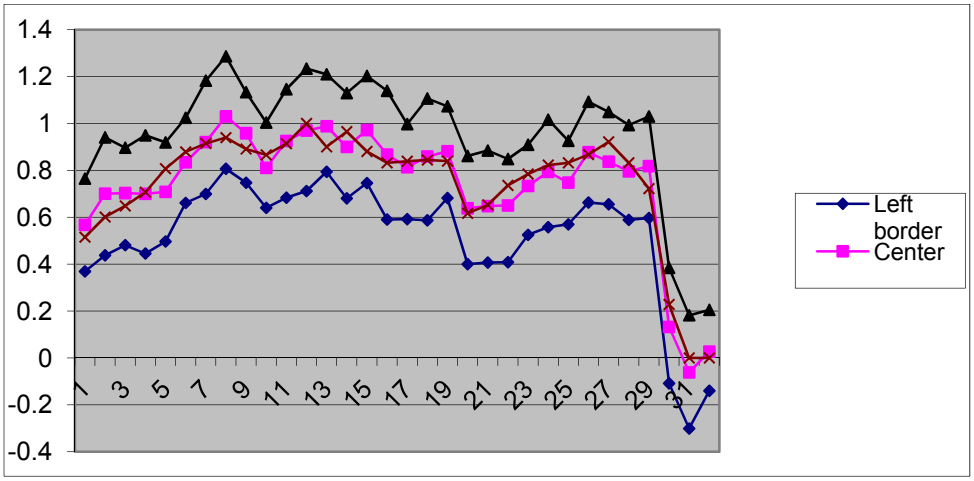

Figure 3. Experiment 1 result for triangular MF and normalized values of input variables

a) For non-normalized input data

Criterion value for this experiment was:

MSE = 6,407928

MAPE =0,24%

2. For Gaussian membership function (optimal level α=0,85)

a) For normalized input data

Criterion value: MSE = 0,033097.

The corresponding results are presented at the Figure 4.

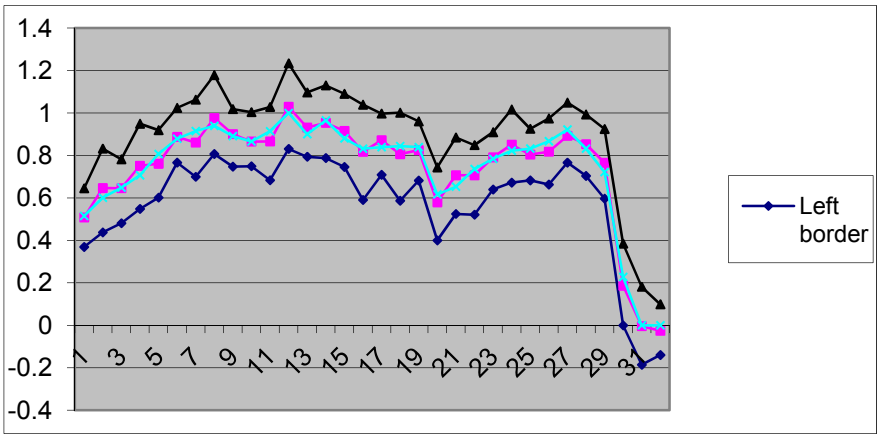

Figure 4. Experiment 1 result for Gaussian MF and normalized values of input variables

b) For non-normalized input data

Criterion value: MSE = 3,432511 MAPE =0,13%

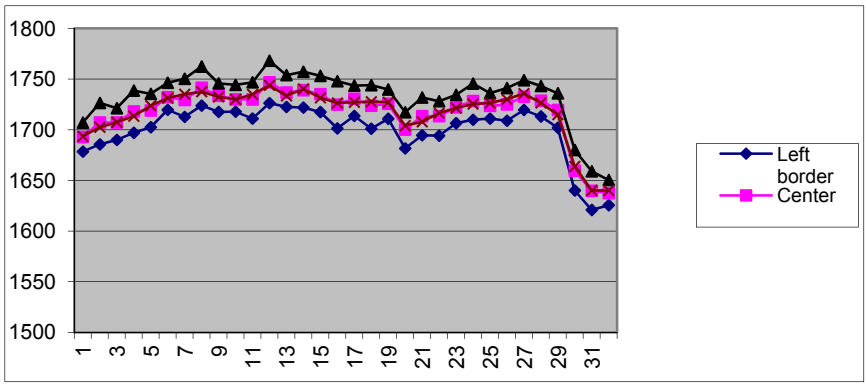

Figure 5. Experiment 1 result for Gaussian MF and non-normalized values of input variables

The total results for triangular and Gaussian MF are presented in the table 1.As we can see from the presented results of experiment 1, forecasting using triangular and Gaussian membership functions gives good results. Results of experiments with Gaussian MF are better than results of experiments with triangular MF.

Table 1. Forecasting results at RTS stock exchange

For non-normalized data	Triangular MF	Gaussian MF
MSE	0,061787	0,033097
For normalized data	Triangular MF	Gaussian MF
MSE	6,407928	3,432511
MAPE	0,24%	0,13%

6. The comparison of GMDH, FGMDH and FGMDH with fuzzy inputs

In the next experiments the comparison of the suggested method FGMDH with fuzzy inputs with known methods: classical GMDH and Fuzzy GMDH was performed

Experiment 2. Forecasting of RTS index (opening price)

Current experiment contains 5 fuzzy input variables, which are the stock prices of leading Russian energetic companies included into the list of RTS index calculation:

Output variable is the value of RTS index (opening price) of the same period (03.04.2006 – 18.05.2006).

Sample size – 32 values.

Training sample size – 18 values (optimal size of the training sample for current experiment).

The following results were obtained presented at the Table 2 and Fig. 6.

Table 2. MSE comparison for different methods of experiment 2

	GMDH	FGMDH	FGMDH with fuzzy inputs, Triangular MF	FGMDH with fuzzy inputs, Gaussian MF
MSE	0,1129737	0,0536556	0,055557	0,028013

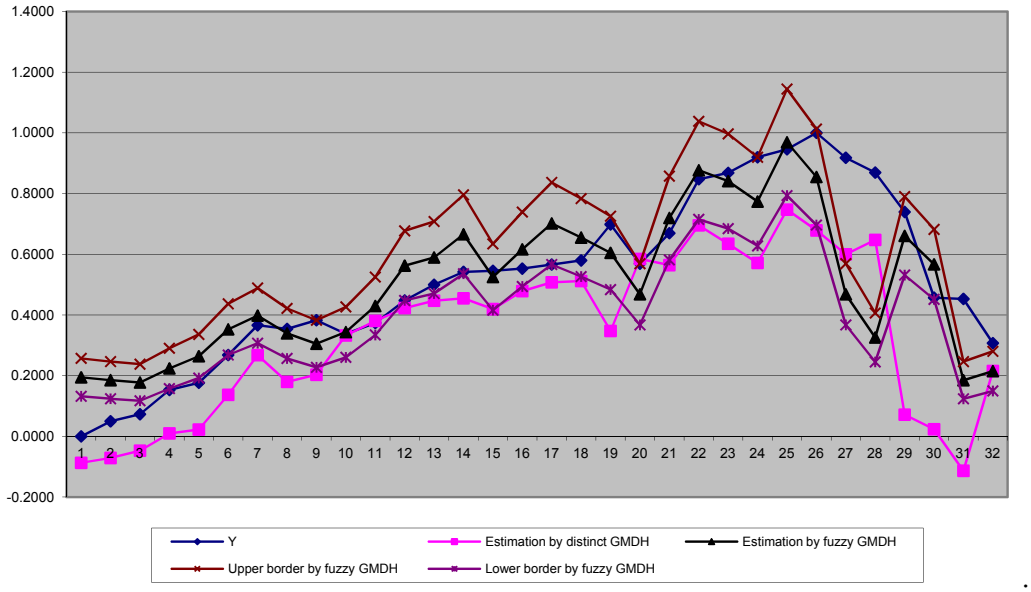

Figure 6. Experiment 2 results using GMDH and FGMDH

As the results of experiment 2 show, fuzzy group method of data handling with fuzzy input data gives more accurate result than FGMDH with triangular membership function or Gaussian membership function. In case of triangular MF FGMDH with fuzzy data gives a little worse than FGMDH with Gaussian MF.

Experiment 3. RTS-2 index forecasting (closing price)

Sample size – 32 values.

Training sample size – 19 values (optimal size of training sample for current experiment).

The following results were obtained, which are presented in Table 3.

Table 3. MSE of different methods of experiment 3 comparison

	GMDH	FGMDH	FGMDH with fuzzy inputs, triangular MF	FGMDH with fuzzy inputs, Gaussian MF
MSE	0,051121	0,063035	0,061787	0,033097

As the results of the experiment 4 show, fuzzy group method of data handling with fuzzy input data gives the better result than GMDH and FGMDH in case of Gaussian membership functions. At the same

time in this experiment GMDH gives the better results, than FGMDH and FGMDH with fuzzy input data in the case of triangular membership functions.

Experiment 4. RTS index forecasting (opening price)

For the efficiency estimation stock indexes forecasting using fuzzy neural nets (FNN) with Mamdani and Tsukamoto algorithms were carried out. Total 267 everyday indexes of stock prices during period from 1.04.2005 to 30.12.2005 were used for neural net training. The following results were obtained

Table 4. Experiment 4 results using FNN

Criterion	Mamdani with Gaussian MF	Mamdani with Triangular MF	Tsukamoto with Gaussian MF	Tsukamoto with Triangular MF
MSE	3,692981	3,341179	7,002467	5,119318
MAPE %	0,256091	0,318056	0,318056	0,419659

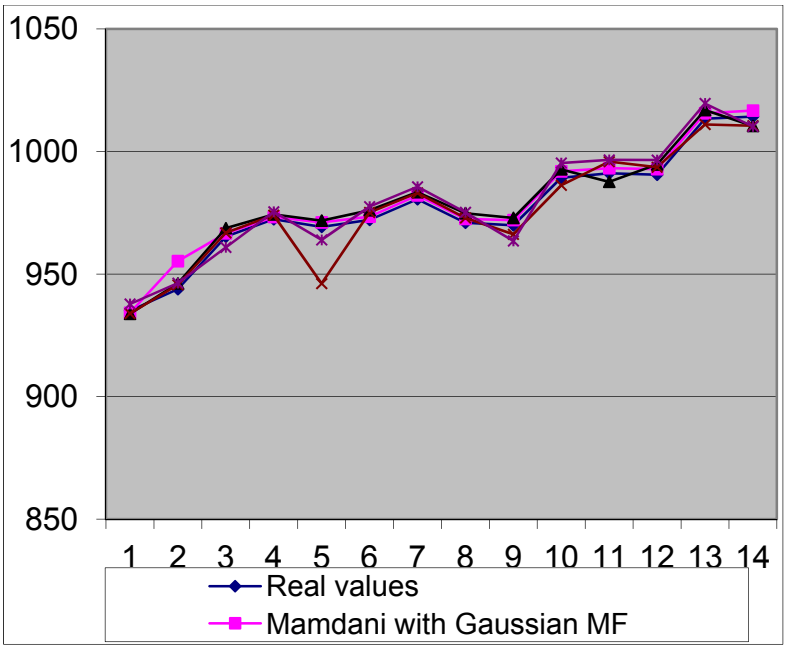

Figure 7. Experiment 4 forecasting results using FNN

As experiment 4 results show, forecasting using Mamdani controller with Gaussian MF was the best, Mamdani controller with triangular MF is on the second place.

The comparative results of forecasting accuracy of FNN and different variants of FGMDH were carried out. The corresponding results are presented at the Table 5 and Figure 8.

Table 5. Forecasting results for FGMDH and FNN

	Mamdani controller	Tsukamoto Controller	FGMDH With fuzzy Inputs (4 input variables)	FGMDH with fuzzy inputs (previous values of forecasted variable used)
MSE for Gaussian MF	0,18046	0,26801	0,115072	0,094002
MSE for triangular MF	0,28112	0,34443	0,210865	0,215421

The best MSE was achieved by FGMDH with fuzzy inputs, and this method also allows to build interval estimation of the forecasted value. FGMDH with fuzzy inputs using Gaussian MF gives more accurate forecast than triangular MF as well as with FNN.

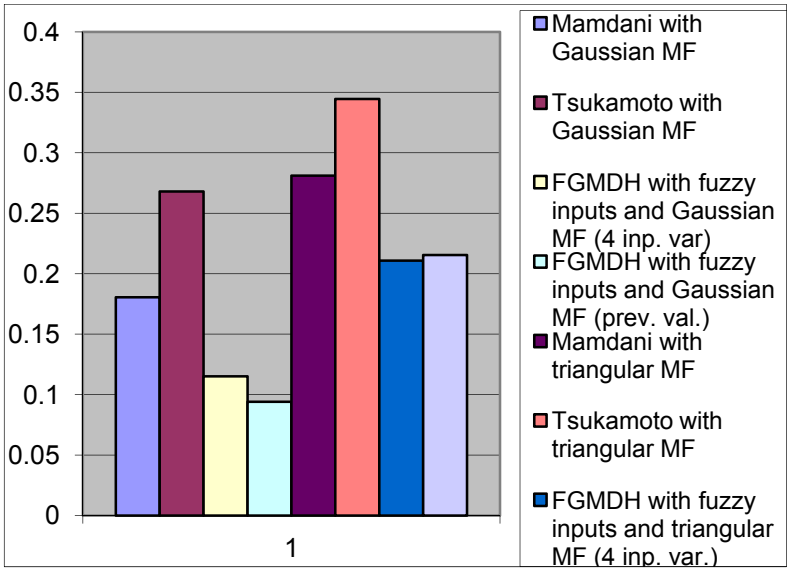

Figure 8. MSE comparison for FMGH and FNN

Experiment 5. "Lukoil" stock prices forecasting based on previous data about stock prices of leading Russian energetic companies for the same period.

Input variables:

EESR – shares of "РАО ЕЭС России" joint-stock company,

YUKO – shares of "ЮКОС" joint-stock company,

SNGSP – privileged shares of "Сургутнефтегаз" joint-stock company,

SNGS – common shares of "Сургутнефтегаз" joint-stock company.

The results are presented at the Table 6.

Table 6. Forecasting results in experiment 4.

		Mamdani Controller	Tsukamoto Controller	FGMDH with fuzzy inputs (4 input variables)	FGMDH with fuzzy inputs (previous values on the input)
Gaussian MF	MSE	3,692981	7,002467	2,1151183	2,886697
	MAPE, %	0,256091	0,318056	0,179447	0,256547
	MSE	3,34179	5,119318	4,717268	4,977901
triangular MF	MAPE, %	0,318056	0,419659	0,40437	0,415434

As current experiment results show, forecasting using FGMDH with fuzzy input data using Gaussian membership function was the best, fuzzy Mamdani controller with Gaussian MF is on the second place.

In a whole the experiments have shown the high accuracy of forecasting using FGMDH in comparison with FNN. The additional advantage od GMDH is its possibility to work with short samples and under uncertainty when input data are fuzzy.

Conclusion

1. The problem of finding unknown dependencies in big data was considered. For its solution inductive modeling method GMDH was suggested which allows constructing models with unknown structure almost automatically. Besides GMDH may work with insufficient data available (Short samples).
2. In case of incomplete or unreliable data fuzzy GMDH with fuzzy inputs was suggested for synthesis of corresponding forecasting models in experimental data.
3. The experimental investigations of the suggested method in the problem of stock prices forecasting with different types of partial descriptions were carried out.
4. The comparison of forecasting accuracy of FGMDH and fuzzy neural networks Mamdani and Tsukamoto was performed confirming the efficiency of FGMDH.

Bibliography

[Barsegyan, 2008] BarsegyanA.A. Technologies of data analysis: Data mining, Visual Mining, TextMining, OLAP. / A.A. Barsegyan, M.C. Kuprianov, V.V. Stepanenko, I.I. Holod.-. -2-nd edition, revised and add.).- SPb.: BHV- Petersburg, 2008.- 384 p. (rus)

[Bodyanskiy, 2009] Bodyanskiy Ye, Zaychenko Yu., Pavlikovskaya E., Samarina M., Viktorov Ye. Neo-fuzzy neural network structure optimization using GMDH for solving forecasting and classification problems // Proc. Int. Workshop on Inductive Modeling 2009. Krynica,Poland, 2009.- 77- 89.

[Duke, 2001] V. Duke, A. Samoilenko. Data Mining: Learning course. Publ. House " Peter". Moscow, Saint- Petersburg, Kharkov, Minsk, 2001.- 366 p. (rus)

[Ivakhnenko, 1985] Ivakhnenko A.G., Mueller I.A. Self-organization of forecasting models.-Kiev: Publ. House "Technika".– 1985. (rus)

[Zaychenko, 2003]Zaychenko Yu. P. Fuzzy Group Method of Data Handling // System research and information technologies.-2003.-№3.-pp.-25-45. (rus)

[Zaychenko, 2006] Zaychenko Yu. The Fuzzy Group Method of Data Handling and Its Application for Economical Processes forecasting // Scientific Inquiry , vol.7, No 1, June, 2006.-pp. 83-98.

[Zaychenko, 2008] Zaychenko Yu. The Fuzzy Group Method of Data Handling with Fuzzy Input Variables // Scientific Inquiry , vol.9, No 1, June, 2008.-pp. 61-76.

Acknowledgement

The paper is published with financial support by the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA (www.ithea.org) and the Association of Developers and Users of Intelligent Systems ADUIS Ukraine (www.aduis.com.ua).

Authors' Information

Yuri Zaychenko *– Professor, doctor of technical sciences, Institute for applied system analysis, NTUU "KPI", 03056, Ukraine, Kyiv, Peremogi pr. 37, Corpus 35; e-mail: baskervil@voliacable.com, zaychenkoyuri@ukr.net*

Major Fields of Scientific Research: Information systems, Fuzzy logic, Decision making theory

Galib Hamidov- *PhD, Azarishig, Head of the Information technologies department, Baku, Azerbaijan , Galib.hamidov@bes.az*

Major Fields of Scientific Research: Information technologies, Data Mining