THE FUZZY GROUP METHOD OF DATA HANDLING WITH FUZZY INPUTS

Yuriy Zaychenko

Abstract: The problem of forecasting models constructing using experimental data in terms of fuzziness, when input variables are not known exactly and determined as intervals of uncertainty is considered in this paper. The fuzzy group method of data handling is proposed to solve this problem. The mathematical model of the problem mentioned above is built and fuzzy GMDH with fuzzy inputs is elaborated in the paper. The corresponding program which uses the suggested algorithm was developed. The experimental investigations and comparison of FGMDH with neural nets in problems of stock prices forecasting were carried out and presented in this paper.

Keywords Group Method of Data Handling, fuzzy, economic indexes, stock prices, forecasting

ACM Classification Keywords: H.4.2 [Information Systems Applications] Types of Systems - Decision support

Conference: The paper is selected from XVth International Conference "Knowledge-Dialogue-Solution" KDS-2 2009, Kyiv, Ukraine, October, 2009.

Introduction

The problem of forecast in financial sphere is of great importance. Financial processes have some specific features which prevent the application classical statistic methods and stipulate the development of novel approaches and methods based on artificial intelligence. One of such novel methods is *fuzzy group method of data handling (FGMDH)* developed in [Зайченко, 2003, Zaychenko, 2006]. As is well known, fuzzy GMDH allows to construct fuzzy models and has the following advantages:

1. The problem of optimal model finding is transformed to the problem of linear programming, which is always solvable;
2. There is interval regression model built as the result of method work out. Fuzzy GMDH in its original form has one limitation: the input variables must be crisp, non-fuzzy while the model constructed is fuzzy.

At the same time situations in financial sphere often occur when the input data are uncertain, given in the form of uncertainty intervals.

The goal of this paper is the development and investigation of extended version of FGMDH working in fuzzy environment when input variables are fuzzy and its comparison with fuzzy neural networks.

Mathematical Model of Group Method of data Handling with Fuzzy Inputs

The initial fuzzy GMDH considers a linear interval regression model:

$$Y=A_0Z_0+A_1Z_1+\ldots+A_nZ_n, \tag{1.1}$$

- where A_i are fuzzy numbers, which are described by threes of parameters $A_i = (\underline{A_i}, \breve{A}_i, \overline{A_i})$, where $\breve{A}_i$ – interval center, $\overline{A_i}$ – upper border of the interval, $\underline{A_i}$ - lower border of the interval,
- and Z_i are input variables which are also fuzzy numbers determined 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 is output fuzzy number, which parameters are defined as follows (in accordance with L-R numbers multiplying formulas):

Center of interval:

$$\breve{y} = \sum \breve{A}_i * \breve{Z}_i .$$

Deviation in the left part of the membership function:

$$\breve{y} - \underline{y} = \sum (\left|\breve{A}_i\right| * (\breve{Z}_i - \underline{Z}_i) + (\breve{A}_i - \underline{A}_i) * \left|\breve{Z}_i\right|) .$$

Thus the upper border of the interval

$$\overline{y} = \sum (\left|\breve{A}_i\right| * (\overline{Z}_i - \breve{Z}_i) + \left|\breve{Z}_i\right| * (\overline{A}_i - \breve{A}_i) + \breve{A}_i * \breve{Z}_i) .$$

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

So, the general requirements to estimation linear interval model are following:

to find such values of parameters $(\underline{A}_i, \breve{A}_i, \overline{A}_i)$ of fuzzy coefficients, which enable:

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

b) Total width of estimation interval be minimal.

Input data for this task is Z_k=[Z_{ki}]$_i$ - input training sample, and also y_k – known output values, $k = \overline{1, M}$, M is a sample size (number of observation points).

There are two cases of fuzzy values membership function investigated in this work:

- Triangular membership functions
- Gaussian membership functions.

FGMDH with Fuzzy Inputs for Triangular Membership Functions

Let's consider the special case of the linear interval regression model:

$$Y = A_0Z_0+A_1Z_1+\ldots+A_nZ_n, \tag{1.2}$$

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 – 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 is a fuzzy number, whose 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:

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

The lower border of the interval:

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

The upper border of the interval:

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

For the interval model to be correct, the real value of input variable Y should lay in the interval got by the method It can be described in such a way:

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

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 the following task of linear programming:

$$\min_{a_i , c_i} \sum_{k=1}^{M} (\sum (a_i * \overline{Z}_i + c_i \left|\breve{Z}_i\right|) - \sum (a_i * \underline{Z}_i - c_i \left|\breve{Z}_i\right|)) \qquad (1.3)$$

under constraints:

$$\begin{cases} \sum (a_i * \underline{Z}_{ik} - c_i \left|\breve{Z}_{ik}\right|) \le y_k , \\ \sum (a_i * \overline{Z}_{ki} + c_i \left|\breve{Z}_{ik}\right|) \ge y_k , k = \overline{1,M}. \end{cases} \qquad (1.4)$$

Formalized problem formulation in case of triangular membership functions

Let's consider partial description

$$f(x_i , x_j) = A_0 + A_1 x_i + A_2 x_j + A_3 x_i x_j + A_4 x_i^2 + A_5 x_j^2 . \qquad (1.5)$$

Rewriting it in accordance with the model (1.1) needs such substitution $z_0 = 1$, $z_1 = x_i$,

$z_2 = x_j$, $z_3 = x_i x_j$, $z_4 = x_i^2$, $z_5 = x_j^2$.

Then math model (1.3)-(1.4) will take the form

$$
\begin{aligned}
&\min_{a_i, c_i} \; (2Mc_0 + a_1 \sum_{k=1}^{M} (\bar{x}_{ik} - \underline{x}_{ik}) + 2c_1 \sum_{k=1}^{M} |\breve{x}_{ik}| + a_2 \sum_{k=1}^{M} (\bar{x}_{jk} - \underline{x}_{jk}) + 2c_2 \sum_{k=1}^{M} |\breve{x}_{jk}| + \\
&+ a_3 \sum_{k=1}^{M} (|\breve{x}_{ik}| (\bar{x}_{jk} - \underline{x}_{jk}) + |\breve{x}_{jk}| (\bar{x}_{ik} - \underline{x}_{ik})) + 2c_3 \sum_{k=1}^{M} |\breve{x}_{ik} \breve{x}_{jk}| + 2a_4 \sum_{k=1}^{M} |\breve{x}_{ik}| (\bar{x}_{ik} - \underline{x}_{ik}) + \\
&+ 2c_4 \sum_{k=1}^{M} \breve{x}_{ik}^2 + 2a_5 \sum_{k=1}^{M} |\breve{x}_{jk}| (\bar{x}_{jk} - \underline{x}_{jk}) + 2c_5 \sum_{k=1}^{M} \breve{x}_{jk}^2)
\end{aligned}
\tag{1.6}
$$

with the following conditions:

$$
\begin{aligned}
&a_0 + a_1 \underline{x}_{ik} + a_2 \underline{x}_{jk} + a_3 (-|\breve{x}_{ik}| (\breve{x}_{jk} - \underline{x}_{jk}) - |\breve{x}_{jk}| (\breve{x}_{ik} - \underline{x}_{ik}) + \breve{x}_{ik} \breve{x}_{jk}) + \\
&+ a_4 (-2|\breve{x}_{ik}| (\breve{x}_{ik} - \underline{x}_{ik}) + \breve{x}_{ik}{}^2) + a_5 (2|\breve{x}_{jk}| (\breve{x}_{jk} - \underline{x}_{jk}) + \breve{x}_{jk}{}^2) - c_0 - c_1 |\breve{x}_{ik}| - \\
&- c_2 |\breve{x}_{jk}| - c_3 |\breve{x}_{ik} \breve{x}_{jk}| - c_4 \breve{x}_{ik}^2 - c_5 \breve{x}_{jk}^2 \le y_k, \\
&a_0 + a_1 \bar{x}_{ik} + a_2 \bar{x}_{jk} + a_3 (|\breve{x}_{ik}| (\bar{x}_{jk} - \breve{x}_{jk}) + |\breve{x}_{jk}| (\bar{x}_{ik} - \breve{x}_{ik}) - \breve{x}_{ik} \breve{x}_{jk}) + a_4 (2|\breve{x}_{ik}| (\bar{x}_{ik} - \\
&- \breve{x}_{ik}) - \breve{x}_{ik}{}^2) + a_5 (2|\breve{x}_{jk}| (\bar{x}_{jk} - \breve{x}_{jk}) - \breve{x}_{jk}{}^2) + c_0 + c_1 |\breve{x}_{ik}| + c_2 |\breve{x}_{jk}| + c_3 |\breve{x}_{ik} \breve{x}_{jk}| + \\
&c_4 \breve{x}_{ik}^2 + c_5 \breve{x}_{jk}^2 \ge y_k.
\end{aligned}
\tag{1.7}
$$

$$
c_l \ge 0, \; l = \overline{0,5}. \tag{1.8}
$$

As we can see, this is the linear programming problem, but there are still no limitations for non-negativity of variables a_i, so we may pass to a dual problem, introducing dual variables $\{\delta_k\}$ and $\{\delta_{k+M}\}$.

Write down dual problem:

$$
\max(\sum_{k=1}^{M} y_k \cdot \delta_{k+M} - \sum_{k=1}^{M} y_k \cdot \delta_k). \tag{1.9}
$$

Under constraints:

$$
\begin{gathered}
\sum_{k=1}^{M} \delta_{k+M} - \sum_{k=1}^{M} \delta_k = 0. \\
\sum_{k=1}^{M} \bar{x}_{ik} \cdot \delta_{k+M} - \sum_{k=1}^{M} \underline{x}_{ik} \cdot \delta_k = \sum_{k=1}^{M} (\bar{x}_{ik} - \underline{x}_{ik}), \\
\sum_{k=1}^{M} \bar{x}_{jk} \cdot \delta_{k+M} - \sum_{k=1}^{M} \underline{x}_{jk} \cdot \delta_k = \sum_{k=1}^{M} (\bar{x}_{jk} - \underline{x}_{jk}).
\end{gathered}
\tag{2.0}
$$

And several constraints inequalities corresponding to variables C_i are not presented here for brevity.

It was proved that the dual LP problem is always solvable as well as the primary LP problem (1.6)-(1.8).

Experimental Investigations of FGMDH and Comparison with Fuzzy Neural Networks

The numerous experimental investigations of the suggested method-FGMDH with fuzzy inputs were carried out at the problem of forecasting stock prices of shares at Russian stock market RTS and comparison of efficiency of FGMDH and fuzzy neural networks. Some of the obtained results are presented below.

Experiment 1. “LUKOIL” stock price forecasting.

1. Stock price forecasting using fuzzy group method of data handling with triangular MF and Gaussian MF:

a) Forecasting based on previous data about stock prices at the period: 01.12./2005 - 20.12.2005.

The following results were obtained which are presented at the table 1 and Fig. 1

Table 1. Experiment 1 results using FGMDH with fuzzy data and triangular MF

Date	Real Values	Lower Border	Center	Upper Border	Deviation
01.12.2005	58.1	57.5984	58.30609	59.05755	0.2060885
02.12.2005	58.7	58.22393	58.70375	59.29865	0.00375055
05.12.2005	59.4	58.35602	59.10925	59.85003	0.290755
06.12.2005	59	57.94641	58.70476	59.55968	0.2952399
07.12.2005	59.85	58.78311	59.60075	60.4198	0.2492475
08.12.2005	59.6	58.6168	59.42533	60.16382	0.17467435
09.12.2005	59.9	59.43436	60.13189	60.94283	0.2318917
12.12.2005	60.65	59.70792	60.46069	61.22393	0.1893122
13.12.2005	60.65	60.10314	60.77584	61.39809	0.1258443
14.12.2005	61.15	60.63159	61.44362	62.25126	0.2936235
15.12.2005	60.25	59.7932	60.44087	61.0663	0.1908689
16.12.2005	61	60.15238	60.88398	61.60982	0.116021
19.12.2005	61.01	60.58471	61.24406	61.85153	0.234062
20.12.2005	60.7	60.02635	60.71058	61.35923	0.01058345

MSE = 0.215421

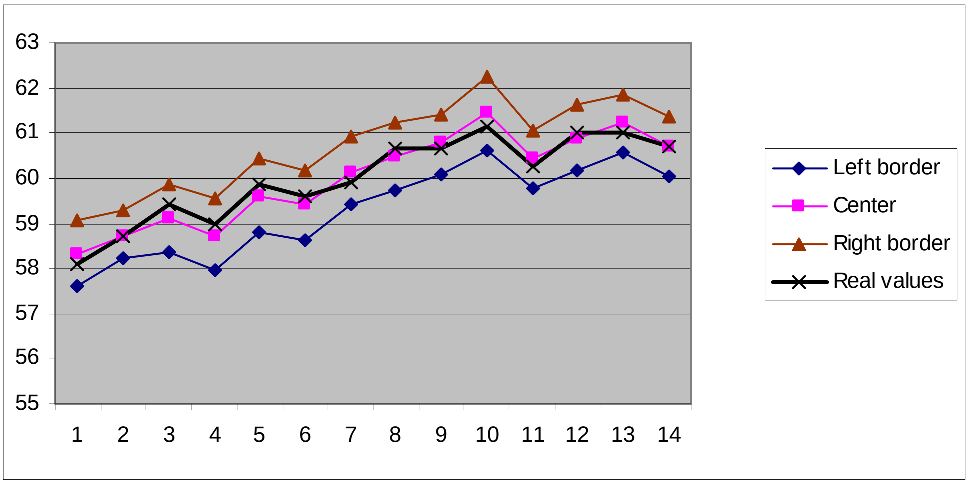

Fig.1. Experiment 3 results using FGMDH with fuzzy input data and triangular MF

Experiment 2. "LUKOIL" stock price forecasting based on previous data about stock prices of leading Russian energetic companies for the same period:

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 following results were obtained:

MSE = 0.115072

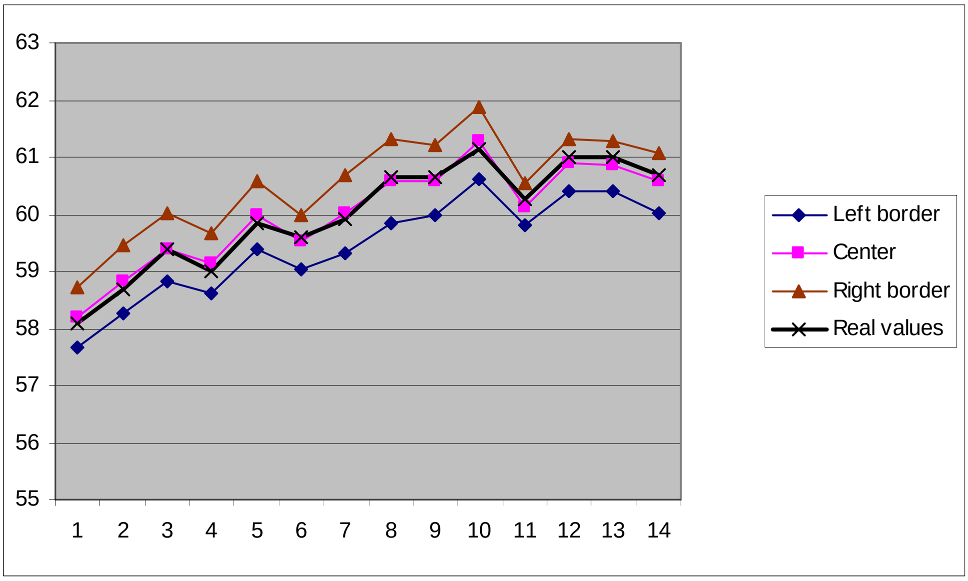

Fig. 2. Experiment 2 results using FGMDH with fuzzy data and Gaussian MF

Experiment 3. Stock price forecasting using fuzzy neural nets with Mamdani and Tsukamoto algorithm.

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 2. Experiment 3 results using FNN

Date	Real Value	Mamdani with Gaussian MF	Mamdani with Triangular MF	Tsukamoto with Gaussian MF	Tsukamoto with Triangular MF
01.12.2005	58.1	58.23	58.41	58.37	58.48
02.12.2005	58.7	58.54	58.46	58.47	58.37
05.12.2005	59.4	59.14	59.11	59.1	59.05
06.12.2005	59	59.11	59.15	59.25	59.27
07.12.2005	59.85	59.97	60.1	60.19	60.23
08.12.2005	59.6	59.416	59.313	59.37	59.23
09.12.2005	59.9	60.12	60.22	60.27	60.32
12.12.2005	60.65	60.5	60.45	60.48	60.4
13.12.2005	60.65	60.54	60.31	60.42	60.28
14.12.2005	61.15	61.32	61.39	61.4	61.42
15.12.2005	60.25	60.1	59.99	60.06	59.97
16.12.2005	61	61.2	61.25	61.22	61.27
19.12.2005	61.01	61.24	61.34	61.28	61.34
20.12.2005	60.7	60.54	60.44	60.48	60.38
MSE		*0.18046*	*0.28112*	*0.268013*	*0. 10093*

As experiment 3 results show, forecasting using FNN with Mamdani controller with Gaussian MF was the best place, Tsukamoto controller with Gaussian MF takes the second place.

Forecasting using FGMDH with fuzzy input data gives best results when using Gaussian MF.

Table 3. Summarizing forecasting results for FGMDH and FNN with Gaussian MF:

	Mamdani controller	Tsukamoto Controller	FGMDH With fuzzy inputs (4 input variables)	FGMDH with fuzzy inputs (previous values on input)
MSD	*0.18046*	*0.268013*	*0.115072*	*0.094002*

For triangular MF:

Table 4. Summarizing forecasting results for FGMDH and FNN with triangular MF

	Mamdani controller	Tsukamoto Controller	FGMDH with fuzzy inputs (4 input variables)	FGMDH with fuzzy inputs (previous values on input)
MSD	*0.28112*	*0.34443*	*0.210865*	*0.215421*

As we can see from tables 2 and 3, the best MSD are given by FGMDH with fuzzy inputs, and this method also allows to build interval estimation of forecasted value. Gaussian MF is more accurate than triangular MF in FGMDH with fuzzy inputs, and also in FNN.

Experiment 4. RTS index forecasting (opening price)

Forecasting of RTS index (opening price) using fuzzy neural nets with Mamdani and Tsukamoto algorithm.

267 everyday indexes of stock prices during period from 1.04.2005 to 30.12.2005 were used for neural net training.

The following final results were obtained: which are presented at the table 5 (for Gaussian MF) and table 6 (triangular MF)

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

Forecasting using FGMDH with fuzzy input data gives best results when using Gaussian MF and using 5 input variables:

Table 5. Forecasting results of experiment 4 for FGMDH and FNN with Gaussian MF

	Mamdani Controller	Tsukamoto Controller	FGMDH with fuzzy inputs (5 input variables)	FGMDH with fuzzy inputs (previous values on the input)
MSD	*3.692981*	*7.002467*	*2.1151183*	*2.886697*
MAPE. %	*0.256091*	*0.318056*	*0.179447*	*0.256547*

Table 6. Forecasting results of experiment 4 for FGMDH and FNN with triangular MF

	Mamdani Controller	Tsukamoto Controller	FGMDH with fuzzy inputs (5 input variables)	FGMDH with fuzzy inputs (previous values of the input variable)
MSD	*3.34179*	*5.119318*	*4.717268*	*4.977901*
MAPE. %	*0.318056*	*0.419659*	*0.40437*	*0.415434*

Experiment 5. RTS 2 index forecasting (opening price)

Table 7. Forecasting results of experiment 5 for FGMDH and FNN with Gaussian MF

	Mamdani controller	Tsukamoto Controller	FGMDH with fuzzy inputs (4 input variables)	FGMDH with fuzzy inputs (previous input values)
MSE	*0.18046*	*0.268013*	*0.115072*	*0.094002*

Table 8. Forecasting results of experiment 5 for FGMDH and FNN with triangular MF

	Mamdani controller	Tsukamoto Controller	FGMDH with fuzzy inputs (4 input variables)	FGMDH with fuzzy inputs (previous values on input)
MSE	*0.28112*	*0.34443*	*0.210865*	*0.215421*

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

Best MSD are given by FGMDH with fuzzy inputs, and this method also allows to build interval estimation of forecasted value. Gaussian MF is more accurate than triangular MF in FGMDH with fuzzy inputs, and also in FNN.

Conclusion

In this paper new method of inductive modeling FGMDH with fuzzy inputs is described. This method represents the development of fuzzy GMDH when information is fuzzy and given in the form of uncertainty intervals. The mathematical model was constructed and corresponding algorithm was elaborated. The experimental results of application of the suggested method in the forecasting of market index and stock prices and their comparison with FNN are presented and discussed. The main advantages of the suggested method are following:

- It operates with fuzzy and uncertain input information and constructs the fuzzy model;
- The constructed model has minimal possible total width and in this sense it is optimal;
- For finding optimal model we solve corresponding linear programming problem which is always solvable.

Experiments have shown that FNN and FGMDH with fuzzy inputs are effective methods for forecasting stock prices and market indexes.

The best results were obtained in cases of Gaussian MF both for fuzzy neural nets and FGMDH.

Besides accurate results, which are in many cases better than the results of FNN, FGMDH also has the following advantages: possibility to work with arbitrary number of fuzzy input variables.

Acknowledgements

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

Bibliography

[Zaychenko Yu, 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 - p.83-96.

[Зайченко, 2001]. Ю.П. Зайченко, І.О. Заєць. Синтез та адаптація нечітких прогнозуючих моделей на основі методу самоорганізації //Наукові вісті НТУУ КПІ, №3, 2001.-с. 34-41.

[Зайченко, 2003] Ю.П. Зайченко. Нечеткий метод индуктивного моделирования в задачах прогнозирования макроэкономических показателей. // Системні дослідження та інформаційні технології, № 3, 2003.-с.25-45.

Authors' Information

Yuriy Zaychenko – Professor NTUU “Kiev Polytechnic Institute”, Institute of Applied System Analysis, Politehnicheskaya 14, Kiev, Ukraine, phone: 8-044-2418693, e-mail: baskervil@voliacable.com