S. D. Shtovba, Dc. Sc. (Eng.), Prof.; O. V. Shtovba, Cand. Sc. (Economics);

O. D. Pankevich, Cand. Sc. (Eng.), Assist. Prof.

ACCURACY AND COMPACTNESS CRITERIA FOR EVALUATING THE QUALITY OF FUZZY KNOWLEDGE BASES IN IDENTIFICATION PROBLEMS

The knowledge base quality is viewed as the property of the model of dependence under study to satisfy customer's requirements according to many criteria, accuracy and compactness being the most popular among them. Fuzzy knowledge bases that simulate three types of dependences are investigated with the output in the form of a crisp number, fuzzy number or a class of solutions. For each type of the dependences accuracy criteria of corresponding knowledge bases are systemized. For evaluating the compactness of a fuzzy knowledge base 9 known partial criteria are described and 5 new criteria are proposed.

Keywords: *fuzzy knowledge base, quality criteria, accuracy, compactness, fuzzy identification.*

Introduction

Definition of the term “quality of fuzzy knowledge bases” is a prerequisite for successful development of the theory of formalized fuzzy knowledge base design. It is most simple to identify quality with accuracy, i.e. with deviation of the fuzzy inference results from the experimental data. This approach, which dominates in the modern theory of fuzzy identification, has led to a number of negative results. Since 1990s a race for “accuracy” has started in the fuzzy scientific community, which resulted in elaboration of a number of methods for designing highly accurate fuzzy knowledge bases. However, fuzzy knowledge bases developed in accordance with these methods have lost a significant competitive advantage – the ability to describe the dependence under study with literally few natural-language statements understandable to the customers – experts in the applied areas without specialized mathematical qualification. Customers perceive such highly accurate fuzzy knowledge bases as an incomprehensible set of numbers that they are not used to trust when making important decisions. Thus, in addition to accuracy, other quality criteria of fuzzy knowledge bases should be taken into account while solving applied fuzzy identification problems.

The issue of the necessity to provide balance between accuracy and the number of fuzzy knowledge base rules was first raised in [1]. In this paper fuzzy knowledge bases are used for classification problems. Later, works have appeared about the balance between accuracy and compactness of fuzzy knowledge bases of other types as well as about the balance between another pair of criteria – accuracy and interpretability. Over the past two years several important papers [2 – 9] on the multicriterial fuzzy identification theory were published.

In [2, 3] with the help of computer experiments it is shown that for the rule selection problem the dependence of fuzzy knowledge base accuracy on its compactness is quadratic. For parametric identification problems with the tuning of membership functions of fuzzy knowledge base terms exponential dependence of accuracy on compactness was observed [4, 5]. In [6] a method is elaborated for solving one of the structural fuzzy identification problems, namely, that of selecting fuzzy knowledge base rules taking into account accuracy and compactness. The novelty of the method consists in the following: instead of the standard threshold levels of accuracy and compactness [1], the linear constraint that determines the compensation mechanism between these conflicting criteria is used. With the new constraint it is possible to reduce significantly the region of feasible solutions narrowing it to the neighborhood of Pareto front.

Paper [7] proposes an approach to identification based on singleton fuzzy knowledge bases taking into account, in addition to accuracy requirements, also transparency, i.e. consistency of the internal model structure and its external behavior. In this paper consistency is understood as the possibility to predict the model behavior at the key points of factor space without performing fuzzy inference. These key points are the cores of fuzzy antecedents of the rules. To provide transparency, the following constraint is introduced: at each key point a corresponding rule of the knowledge base is to be implemented with a single membership degree and other rules are not to be activated.

Paper [8] proposes an approach to identification taking into account the following three components of the fuzzy knowledge base interpretability: 1) fuzzy partitioning of the variables must meet the semantic integrity conditions; 2) knowledge base must be compact and sufficient; 3) each rule must contain only informative features, i.e. rules could be incomplete (short).

Paper [9] proposes an approach to identification of dependences in classification problems, which takes into account accuracy and interpretability of the fuzzy knowledge base. Interpretability is proposed to be evaluated by consistency of implicit and explicit semantics of the fuzzy knowledge base.

Paper [10] presents an overview of the approaches to defining the term “fuzzy knowledge base interpretability” and proposes their taxonomy. Taxonomy is performed in accordance with complexity and semantics of the models at two levels: at the level of knowledge base rules and at the level of fuzzy partition of linguistic variables. It is shown that the term “fuzzy knowledge base interpretability” and its metrics are not generally accepted yet and, therefore, require further formalization and generalization. Overview [10] is written on 75 works.

Thus, there is a necessity in a formalized definition of the fuzzy knowledge base quality as the model property to satisfy customer’s requirements on many criteria, such as accuracy, compactness, transparency, interpretability, etc. The next step will be development of the algorithms of calculating the above partial criteria for any fuzzy knowledge base. And, finally, methods should be developed for ensuring the desired quality taking into account many criteria on the basis of systematic approach to fuzzy modeling. The **aim** of the paper is to develop criteria of accuracy and compactness of fuzzy knowledge bases that are used for modeling dependences with continuous, discrete and fuzzy output.

1. Classification of fuzzy identification problems

Let us consider the identification object of MISO type with n inputs $\mathbf{X}=(x_1,x_2,...,x_n)$ and a single output y. We propose to classify identification problems according to the type of output variable y (fig. 1).

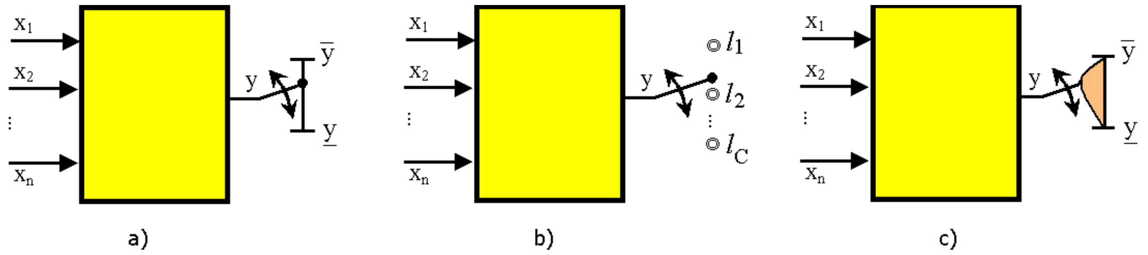

Fig. 1. Classification of fuzzy identification problems
a) continuous output; b) discrete output; c) fuzzy output

For a continuous output (fig. 1a) the value of the output variable y will be an ordinary real number from the range $[\underline{y}, \overline{y}]$. Identification of such dependences is most often carried out using fuzzy knowledge bases of Sugeno and Mamdani [11 – 13].

For a discrete output (fig. 1b) the value of the output variable y will be an element from the countable set $\{l_1, l_2,..., l_C\}$ that is meaningfully interpreted as a decision. Identification of such dependences is expedient to be carried out using fuzzy classified knowledge bases [1, 11, 14].

For a fuzzy output (fig. 1c) the value of the output variable will be a fuzzy number. Accordingly, map $\mathbf{X}=(x_1,x_2,...,x_n)\to\tilde{y}$ can be viewed as a fuzzy function. Three approaches are used for identification of such dependences.

In the first approach [15, 16] a fuzzy knowledge base is used, where consequents of the rules are given by fuzzy sets. A common key feature of the inference for such knowledge bases is production of a fuzzy set at the output

$$\tilde{y}=\int\limits_{y\in[\underline{y},\,\overline{y}]}\mu_{\tilde{y}}(y)\,/\,y\,, \tag{1}$$

where $\mu_{\tilde{y}}(y)$ – degree of number $y\in[\underline{y},\overline{y}]$ membership to the output fuzzy set $\tilde{y}$.

The obtained result will be in the form of fuzzy set (1) after the inference on Mamdani knowledge base [12], on relational knowledge base of Pedrycz [17] and on a fuzzy knowledge base with fuzzy regression equations [18]. Inference on these fuzzy knowledge bases is performed without deffuzification. After that additional procedures for normalization and correction of non-convexity of the fuzzy output set are sometimes necessary (fig. 2). Conversion of non-convex fuzzy sets into fuzzy numbers is carried out by means of approximation by parametric membership functions. In this case for calculating the residual between two fuzzy sets $\tilde{A}=\int\limits_{y\in[\underline{y},\,\overline{y}]}\mu_{\tilde{A}}(y)\,/\,y$ and $\tilde{B}=\int\limits_{y\in[\underline{y},\,\overline{y}]}\mu_{\tilde{B}}(y)\,/\,y$ the following formula is used:

$$RMSE(\tilde{A},\ \tilde{B})=\sqrt{\frac{\int_{\underline{y}}^{\overline{y}}\left(\mu_{\tilde{A}}(y)-\mu_{\tilde{B}}(\,y)\right)^2dy}{\overline{y}-\underline{y}}}\ . \tag{2}$$

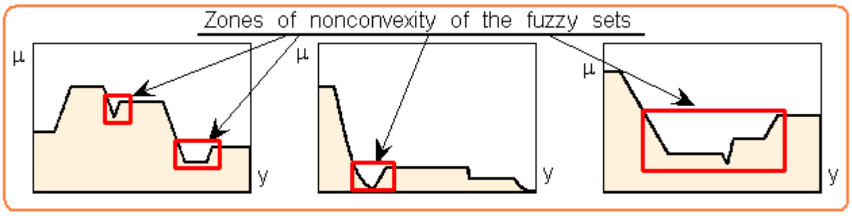

Fig. 2. Nonconvex fuzzy sets obtained by Mamdani inference

In the second approach a fuzzy knowledge base with crisp consequents is used, e.g. a singleton fuzzy knowledge base or Sugeno knowledge base. The final inference procedure – defuzzification – is not performed. Accordingly, at the output we obtain a fuzzy set on a discrete support that is converted to fuzzy number $\tilde{y}$, by means of approximation using a continuous parametric membership function. Parameters of this membership function we select so as to minimize the mean square residual between membership degrees of the two fuzzy sets. In this case the results of defuzzification of the discrete and continuous fuzzy sets must coincide. Fig. 3 presents the examples of such approximation using typical and special membership functions from [16].

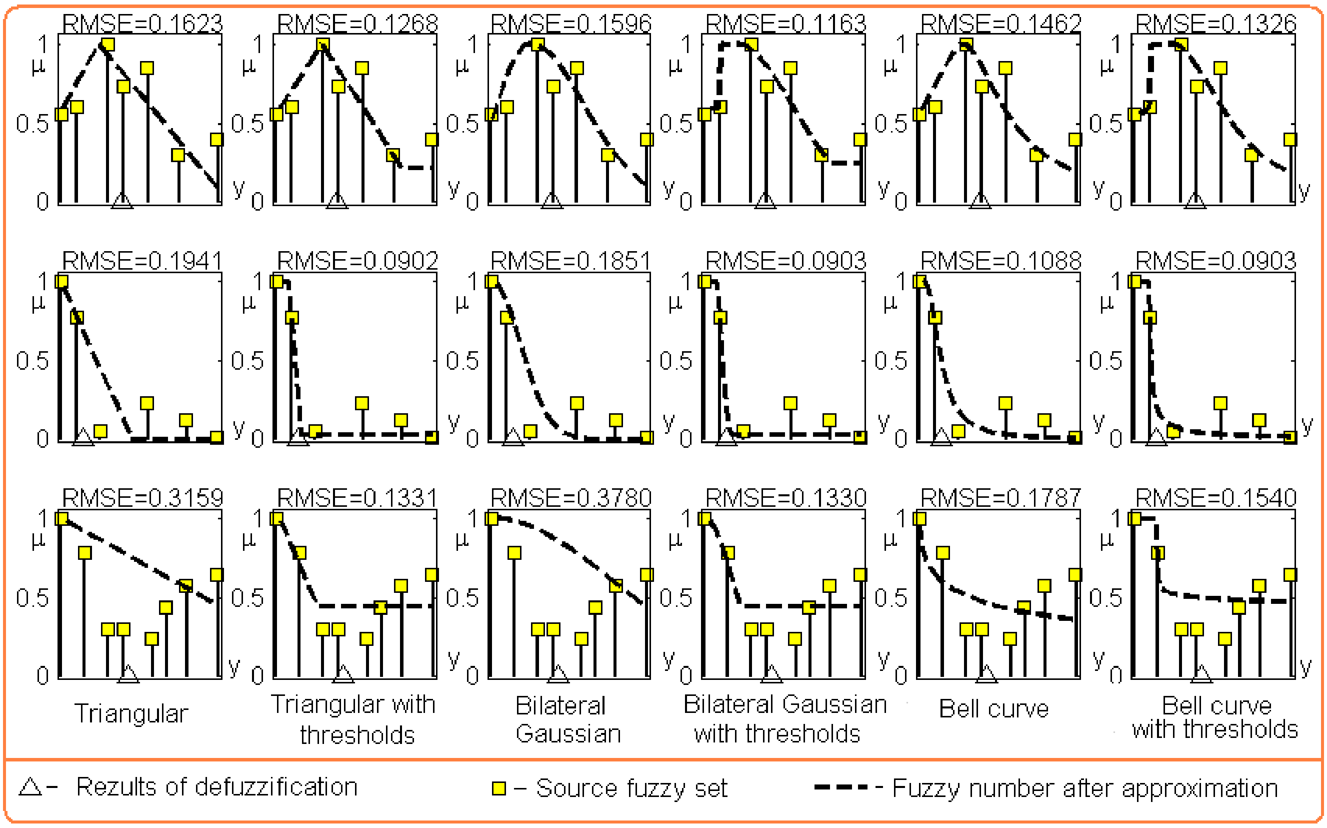

Fig. 3. Aproximation of nonconvex discrete fuzzy sets by parametric membership functions

The third approach [15, 16] assumes that for each points of the factor space corresponding fuzzy number $\tilde{y}$ has parametric membership function of the same type. To describe the dependence of the parameters of this membership function on $(x_1, x_2, ..., x_n)$, a MIMO knowledge base, which has several input and several output variables, is used. Input variables are identical to inputs $(x_1, x_2, ..., x_n)$ of the investigated dependence. Each output variable of the knowledge base sets one parameter of the membership function of the fuzzy number $\tilde{y}$. For describing such dependences different fuzzy knowledge bases are suitable – those, the inference on which results in obtaining crisp numbers. E. g., fuzzy knowledge base of Mamdani will consist from the rules of the type:

If x_1=”Low” and x_2=”High”, then y_1=”Average” и y_2=”Small”,

where y_1 and y_2 – parameters of the membership function of output fuzzy number $\tilde{y}$, e. g. concentration factor and coordinate of the maximum of the bell-type membership function.

2. Accuracy criteria for problems with continuous output

For accuracy evaluation we consider the following data sample to be known:

$$(\mathbf{X}_r, y_r),\quad y_r \in [\underline{y}, \overline{y}],\ r = \overline{1, M}\,, \tag{3}$$

where $\mathbf{X}_r$ and y_r – a pair of «inputs – output» data in the r-th row of the sample; M – the sample length

Let us denote the model based on the fuzzy knowledge base linking inputs $\mathbf{X}$ with output y of the investigated dependence as $y = F(\mathbf{X})$. For the problem with continuous output identification accuracy for sample (3) is determined by the method of least squares:

$$RMSE = \sqrt{\frac{1}{M} \sum_{r=\overline{1,M}} (y_r - F(\mathbf{X}_r))^2}\ .$$

Popularity of *RMSE* accuracy criterion is determined by the smoothness of the objective function at the stage of parametric identification, which allows application of fast gradient optimization methods. A disadvantage is the residual minimization on the average, which does not exclude considerable errors in certain areas of the factor space. For protection against such runs maximum absolute residual could be chosen as the accuracy criterion [19]:

$$MaxErr = \max_{r=1,M} \left| y_r - F(\mathbf{X}_r) \right| .$$

RMSE and *MaxErr* criteria are correlated, although, as a rule, a fuzzy knowledge base that minimizes the residual for one of them is not the best according to the other criterion as well.

3. Accuracy criteria for the problems with a discrete output

Let us assume that the following data sample is known:

$$(\mathbf{X}_r, y_r), \quad y_r \in \{l_1, l_2, ..., l_C\}, \quad r = \overline{1, M}, \tag{4}$$

where $\{l_1, l_2, ..., l_C\}$ – classes of decision

For the problem with discrete output classification error rate for sample (4) is usually chosen as accuracy criterion:

$$MCR = \frac{\sum\limits_{r=1,M} \Delta_r}{M}, \tag{5}$$

where $\Delta_r = \begin{cases} 1, & if \ \ y_r \neq F(\mathbf{X}_r) \\ 0, & if \ \ y_r = F(\mathbf{X}_r) \end{cases}.$

The advantages of criterion (5) are its simplicity and visual interpretation. The disadvantage is related to the fact that membership degrees of the chosen and the competitive alternatives are not taken into account, i.e. “confidence” of the fuzzy classifier remains without attention while selecting one solution from the set $\{l_1, l_2, ..., l_C\}$. This happens because, as a result of inference, for current input vector $\mathbf{X}^*$ from the obtained fuzzy class

$$\tilde{y}(\mathbf{X}^*) = \left(\frac{\mu_{l_1}(\mathbf{X}^*)}{l_1}, \frac{\mu_{l_2}(\mathbf{X}^*)}{l_2}, ..., \frac{\mu_{l_C}(\mathbf{X}^*)}{l_C} \right) \tag{6}$$

a single alternative with maximal membership degree is selected. Doubts as to the correctness of the chosen solution arise when membership degrees of the alternatives are approximately equal.

To take into account the level of confidence when making decisions, in [14] an accuracy criterion is proposed in the form of distance between experimental data and the inferred results. For this, the value of output variable in sample (4) is fuzzified in the following way:

$$\left. \begin{array}{ll} \tilde{y} = (1/l_1, 0/l_2, ..., 0/l_C), & if \ \ y = l_1 \\ \tilde{y} = (0/l_1, 1/l_2, ..., 0/l_C), & if \ \ y = l_2 \\ \vdots & \\ \tilde{y} = (0/l_1, 0/l_2, ..., 1/l_C), & if \ \ y = l_C \end{array} \right\}. \tag{7}$$

After that, the distance between fuzzy classification results (6) and the desired fuzzy values of the output variable (7) is calculated for each object. Accordingly, the accuracy criterion for the entire sample (4) is written as follows [14]:

$$FD = \sqrt{\frac{1}{\mathrm{M}} \sum_{\mathrm{r}=1,M} D_r} , \tag{8}$$

where $D_r = \sum_{j=\overline{1,C}} \left(\mu_{l_j}(y_r) - \mu_{l_j}(\mathbf{X}_r)\right)^2$ – distance between the desired and the inferred fuzzy sets in the r-th object classification;

$\mu_{l_j}(y_r)$ – membership degree of the variable value from the r-th row of data sample (4) of class l_j in accordance with (7);

$\mu_{l_j}(\mathbf{X}_r)$ – degree of the input vector $\mathbf{X}_r$ membership to class l_j, calculated as a result of fuzzy inference

Computer experiments conducted in [11, 20, 21] show that in some cases minimizing fuzzy knowledge base (8) does not provide the rate of classification errors close to the minimal one (5). It is explained by the fact that objects close to the interfaces between classes make almost equal contribution to the accuracy criterion (8) for both correct and erroneous classification. Therefore, in [20] a new criterion that inherits the advantages of the previous ones is proposed. The idea consists in increasing the distance D_r for incorrectly classified objects. As a result, the accuracy criterion becomes as follows:

$$PFD = \sqrt{\frac{1}{\mathrm{M}} \sum_{\mathrm{r}=\overline{1,M}} (\Delta_r \cdot penalty + 1) \cdot D_r}\ , \tag{9}$$

where $penalty > 0$ – penalty factor.

With the known prices for classification errors of different types, values of Δ_r in (5) and (9) are determined according to the payoff matrix.

4. Accuracy criterion for the problems with fuzzy output

Let us assume that the following data sample is known:

$$(\mathbf{X}_r, \tilde{y}_r),\ \mathrm{supp}(\tilde{y}_r) \in [\underline{y}, \overline{y}],,\ r = \overline{1,M}, \tag{10}$$

where $\tilde{y}_r$ – output value in the r-th row of the sample, which is given by a fuzzy number at the $\mathrm{supp}(\tilde{y}_r)$ support.

For the problem with fuzzy output the accuracy criterion for sample (10) is determined as follows [15, 16]:

$$RMSE = \sqrt{\frac{1}{M} \sum_{r=\overline{1,M}} RMSE\left(\tilde{y}_r, \tilde{F}(\mathbf{X}_r)\right)^2}\ ,$$

where $\tilde{F}(\mathbf{X}_r)$ – fuzzy number obtained as a result of inference on the fuzzy knowledge base for input vector $\mathbf{X}_r$; $RMSE\left(\tilde{y}_r, \tilde{F}(\mathbf{X}_r)\right)$ – residual (2) between two fuzzy numbers corresponding to the desired and the inferred results.

5. Compactness criteria of a fuzzy knowledge base

To evaluate the compactness of fuzzy knowledge bases the following partial criteria are used [1 – 5, 8, 10]:

n – the number of input variables of the model;

N_{rules} – the number of rules in the knowledge base;

N_{r1} – the number of the fuzzy knowledge base rules, the antecedents of which contain only one variable, i.e. the number of rules with antecedents of the single-element length;

N_{r2} – the number of the fuzzy knowledge base rules with the antecedents of two-element length;

N_{r3} – the number of the fuzzy knowledge base rules with the antecedents of three-element length;

N_{vr} – total length of the antecedents of all the fuzzy rules;

N_{x_i} – power of the term set of the input variable x_i, $i=\overline{1,n}$;

$N_x^{total}=\sum\limits_{i=\overline{1,n}} N_{x_i}$ – total number of the terms of input variables;

$RF=\dfrac{N_{rules}}{N_{\max}}$ – the level of the knowledge base fullness with rules where $N_{\max}=\prod\limits_{i=\overline{1,n}} N_{x_i}$ – maximally possible number of rules.

In addition to these criteria we propose the following ones:

RF_a – average level of a hierarchal knowledge base fullness with rules;

$AF=\dfrac{N_{vr}}{n\cdot N_{\max}}$ – fullness level of the antecedents of the knowledge base rules;

AF_a – average fullness level of the antecedents of a hierarchal knowledge base;

N_H – total number of the adjustable parameters in a fuzzy knowledge base;

N_{infer} – the number of inferences on a hierarchical knowledge base.

Integral compactness criterion can be determined by the convolution of the set of the above partial criteria. Feasibility of including certain particular criteria into the integral criterion is determined by the specifics of a corresponding identification problem.

Conclusions and prospects for future research

Quality of a fuzzy knowledge base should be viewed as the property of the model of dependence under study to satisfy the customer’s requirements on many criteria such as accuracy, compactness, transparency, interpretability, etc. We have considered fuzzy knowledge bases of three types, as a result of inference on which we obtain a crisp number, a fuzzy number or a class of decision at the output. For each knowledge base type accuracy criteria are systemized. For compactness evaluation 9 known criteria are described and 5 new criteria are proposed. These criteria take into account the number of the model input variables, the number of rules in the knowledge base, the number of short rules in the knowledge base, total length of the antecedents of all rules in the knowledge base, powers of term sets of the input variables, the level of the knowledge base fullness with rules, the number of adjustable parameters of the knowledge base, etc. The new criteria are mainly related to the hierarchical fuzzy knowledge bases. Future research will be directed towards formalization of transparency and interpretability of fuzzy knowledge bases with further development of multicriterial methods for ensuring the desired quality of fuzzy identification of multifactor dependences.

REFERENCES

1. Ishibuchi H. Single-objective and two-objective genetic algorithms for selecting linguistic rules for pattern classification problems / H. Ishibuchi, T. Murata, I. B. Turksen // Fuzzy Sets and Systems. – 1997. – Vol. 89, No. 2 – P. 135 – 50.

2. Штовба С. Д. Вплив кількості нечітких правил на точність бази знань Мамдані / С. Д. Штовба, В. В. Мазуренко, О. Д. Панкевич // Вісник Хмельницького національного університету. Технічні науки. – 2011. – № 2. – С. 185 – 188.

3. Штовба С. Д. Залежність точності ідентифікації від обсягу нечіткої синглтоної бази знань / С. Д. Штовба, О. Д. Панкевич, В. В. Мазуренко // Інформаційні технології та комп’ютерна інженерія. – 2011. – № 1. – С. 73 – 78.

4. Штовба С. Д. Дослідження навчання компактних нечітких баз знань типу Мамдані / С. Д. Штовба , В. В. Мазуренко // Штучний інтелект. – 2011. – № 4. – С. 521 – 529.

5. Штовба С. Д. Дослідження навчання компактних нечітких синглтонних баз знань / С. Д. Штовба, В. В. Мазуренко // Вимірювальна та обчислювальна техніка в технологічних процесах. – 2011. – № 1. – С. 133 – 139.

6. Genetic algorithm for selecting rules of the fuzzy knowledge base balanced according to the accuracy and

compactness criteria [Електронний ресурс] / S. D. Shtovba, V.V. Mazurenko, D. A. Savchuk // Наукові праці Вінницького національного технічного університету. – 2012. – №3. Режим доступу до журн.: http://www.nbuv.gov.ua/e-journals/VNTU/2012_3/2012-3_en.files/en/12sdsacc_en.pdf.

7. Riid A. Identification of Transparent, Compact, Accurate and Reliable Linguistic Fuzzy Models / A. Riid, E. Rüstern // Information Sciences. – 2011. – Vol. 181, № 20. – P. 4378 – 4393.

8. Guillaume S. Learning interpretable fuzzy inference systems with FisPro // Information Sciences / S. Guillaume, B. Charnomordic // Information Sciences. – 2011. – Vol. 181, № 20. – P. 4409 – 4427.

9. Mencar C. Design of fuzzy rule-based classifiers with semantic cointension / C. Mencar, C. Castiello, R. Cannone, A. M. Fanelli // Information Sciences. – 2011. – Vol. 181, № 20. – P. 4361 – 4377.

10. Gacto M. J. Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures / M. J. Gacto, R. Alcala, F. Herrera // Information Sciences. – 2011. – Vol. 181, № 20. – P. 4340 – 4360.

11. Штовба С. Д. Проектирование нечетких систем средствами MATLAB / С. Д. Штовба. – М.: Горячая линия. – Телеком, 2007. – 288 с.

12. Mamdani E. H. An Experiment in Linguistic Synthesis with Fuzzy Logic Controller / E. H. Mamdani, S. Assilian // Int. J. Man-Machine Studies. – 1975. – Vol. 7, № 1. – P. 1 - 13.

13. Takagi T. Fuzzy Identification of Systems and Its Applications to Modeling and Control / T. Takagi, M. Sugeno // IEEE Trans. on Systems, Man, and Cybernetics. – 1985. Vol. 15, № 1. – P. 116 - 132.

14. Ротштейн А. П. Интеллектуальные технологии идентификации: нечеткая логика, генетические алгоритмы, нейронные сети / А. П. Ротштейн. – Винница: УНІВЕРСУМ–Вінниця, 1999. – 320 с.

15. Shtovba S. Fuzzy Model Tuning Based on a Training Set with Fuzzy Model Output Values / Shtovba S. // Cybernetics and System Analysis. – 2007. –Vol.43, №3. – P. 334 – 340.

16. Штовба С. Д. Навчання нечіткої бази знань за вибіркою нечітких даних / С. Д. Штовба // Штучний інтелект. – 2006. – № 4. – С. 560 – 570.

17. Pedrycz W. An Identification Algorithm in Fuzzy Relational Systems / W. Pedrycz // Fuzzy Sets and Systems. – 1984. – № 13. – P. 153 – 167.

18. Штовба С. Д. Моделювання залежностей за допомогою нечіткої бази знань з нечіткими регресійними рівняннями / С. Д. Штовба // Вісник Вінницького політехнічного інституту. – 2011. – № 3. – С. 195 – 199.

19. Rotshtein A. Modeling of the Human Operator Reliability with the Aid of the Sugeno Fuzzy Knowledge Base / Rotshtein A., Shtovba S. // Automation and Remote Control. – 2009. – Vol. 70, № 1. – P. 163 – 169.

20. Shtovba S. Tuning the Fuzzy Classification Models with Various Learning Criteria: the Case of Credit Data Classification / S. Shtovba, O. Pankevich, G. Dounias // Fuzzy Sets and Soft Computing in Economics and Finance : intern. conference, 17 – 20 June 2004 : proc., Vol. 1. – St. Petersburg (Russia), 2004. – P. 103 – 110.

21. Штовба С. Д. Порівняння критеріїв навчання нечіткого класифікатора / С. Д. Штовба // Вісник Вінницького політехнічного інституту. – 2007. – № 6. – С. 84 – 91.

Shtovba Serhiy – Prof., Dc. Sc. (Eng.) of the Department of Computer Control Systems.

Shtovba Olena – Cand. Sc. (Economics), Ass. Prof. of the Department of Management and Modeling in Economics.

Pankevich Olga – Cand. Sc. (Eng.), Ass. Prof. of the Department of Heat and Gas Supply.
Vinnytsia National Technical University.