S. D. Shtovba, Dc. Sc. (Eng.), Prof.; V. V. Mazurenko; D. A. Savchuk

GENETIC ALGORITHM FOR SELECTING RULES OF THE FUZZY KNOWLEDGE BASE BALANCED ACCORDING TO THE ACCURACY AND COMPACTNESS CRITERIA

A genetic algorithm of searching for the set of rules is proposed in order to form a fuzzy knowledge base balanced according to the accuracy and compactness criteria. The algorithm is distinguished by introduction of the linear constraint into the statement of optimization problem. The constraint sets the level of the model accuracy compensation by its compactness. This approximates the region of feasible solutions to the Pareto front.

Key words: *fuzzy knowledge base, accuracy, compactness, selection of rules, Pareto front, genetic optimization.*

Introduction

Fuzzy knowledge base is a set of “If – then” fuzzy rules that describe the relationship between the inputs and the outputs of a certain object using linguistic terms [1]. One of a fuzzy knowledge base design problems is selection of rules from a certain pre-defined set of candidates. Candidate rules could be formed by an expert or obtained by processing the existing experimental data.

Ideally, a fuzzy knowledge base should be both compact and adequate. This cannot be achieved in real problems and so, in practice, attempts are made to choose a knowledge base with a correct balance between these conflicting criteria. The necessary condition of such balance is the knowledge base location on the Pareto front in the “model complexity – model accuracy” coordinates.

Selection of the fuzzy knowledge base rules can be reduced to the binary knapsack problem. An object that can be placed into the knapsack corresponds to a rule of the knowledge base, utility of the knapsack – to the knowledge base accuracy and the total amount of selected objects – to the number of rules. Difference between the problems is in the different types of utility function that is linear in the knapsack problem and is nonlinear – in the problem of selecting the knowledge base rules. By analogy to the classical statements of the knapsack problem [2] the problem of selecting the fuzzy knowledge base rules is generated. The main works in this area are papers [3, 4] on forming a set of knowledge bases of the fuzzy classifier which belong to the Pareto front of non-dominant alternatives in the “number of rules – accuracy" coordinates. For this optimization problems are used with the purpose of 1) maximizing the accuracy with the limited number of rules; 2) minimization of the number of rules for a given level of accuracy and 3) minimization of the integral quality criterion of the knowledge base in the form of a linear convolution of accuracy and of the number of rules [4] or of the accuracy, the number of rules and the total length of the antecedents of the rules [5]. To obtain the Pareto front, optimization is carried out repeatedly for different threshold values in the constraints of problems 1 and 2 and of the weighting coefficients of the objective function in problem 3. Similar approaches are used while selecting the rules of the fuzzy knowledge bases for objects with continuous output [6].

The problem of selecting the fuzzy knowledge base rules as well as the knapsack problem is NP-complete. Accordingly, the algorithm of the exact solution of this problem has exponential computational complexity and, therefore, will be acceptable only for a small number of candidate rules. In practice, this task is usually solved with the application of genetic algorithms. Encoding of the variants is performed according to the Pittsburg approach [7], representing the version of solution by a chromosome, each gene of which specifies the corresponding rule belonging to the knowledge base [6].

The threshold constraint on the knowledge base complexity or on the fuzzy model accuracy [3, 4] generates a fairly large region of feasible solutions, most of which is located far from the Pareto front. This slows down finding optimum solutions that are located on the Pareto front. **The goal of this paper** is to reduce computational complexity of the fuzzy knowledge base rule selection by the development of a new method for finding optimal solutions in the neighborhood of Pareto-front. This neighbourhood is set by a linear constraint that describes the model accuracy compensation by its compactness. We estimate the constraint factors by the Pareto front end points, which correspond to the almost empty and almost filled knowledge bases, as well as by its upper limit that is found by the greedy algorithm on the basis of the ideas of the approximate Sahni method for the knapsack problem [2]. Computational complexity of this problem is quadratic and, therefore, it will not significantly increase the optimization time. Search for optimal solutions will be performed by a genetic algorithm.

1. Mathematical statements of the problems

Let us assume that

- the sampling consisting from M pairs of experimental data about the influence of factors $X = (x_1, x_2, ..., x_n)$ on the continuous output in the investigated dependence:

$$(X_r, y_r),\ r = \overline{1, M}\,, \tag{1}$$

where X_r is the input vector in the r-th line of the sampling; y_r – the corresponding output value;

- set R consisting from N candidate rules to the fuzzy knowledge base, $N = |R|$

are known.

The model, based on fuzzy rules $R' \subseteq R$ that ties inputs X with the output y of the dependence being investigated we designate as $y = F(R', X)$. The root mean squared error in sampling [1] is chosen as the fuzzy model accuracy criterion:

$$RMSE(R') = \sqrt{\frac{1}{M}\sum_{r=1,\,M}\left(y_r - F(R', X_r)\right)^2}\,. \tag{2}$$

In a general case the task is to determine such set of rules R' that provides:

$$\begin{cases} RMSE(R') \to \min \\ C(R') \to \min \end{cases}, \tag{3}$$

where $C(R')$ is the fuzzy model complexity determined by the number of rules $C(R') = |R'|$ or by the completeness level of the knowledge base $C(R') = \frac{|R'|}{N}$.

Multicriteria optimization problem (3) is transformed into the following scalar problems [2, 3]:

$$\begin{cases} RMSE(R') \to \min \\ C(R') \le C^* \end{cases}, \tag{4}$$

$$\begin{cases} C(R') \to \min \\ RMSE(R') \le RMSE^* \end{cases}, \tag{5}$$

where C^* and $RMSE^*$ are maximally permissible values of complexity and error.

Statements (4) and (5) form a large region of feasible solutions, its considerable part being located far from the Pareto front (Fig.1a and 1b). We suggest writing the optimization problem constraint as

$$RMSE(R') \le a \cdot C(R') + b\,, \tag{6}$$

where $a < 0$ and $b > 0$ are the parameters that could be chosen so that the region of feasible solutions will be formed in the neighborhood of Pareto front (Fig. 1c).

Using constraint (6), the following problems of selecting the fuzzy knowledge base rules are formulated:

$$\begin{cases} RMSE(R') \to \min \\ RMSE(R') \le a \cdot C(R') + b \end{cases}, \qquad (7)$$

$$\begin{cases} C(R') \to \min \\ RMSE(R') \le a \cdot C(R') + b \end{cases}. \qquad (8)$$

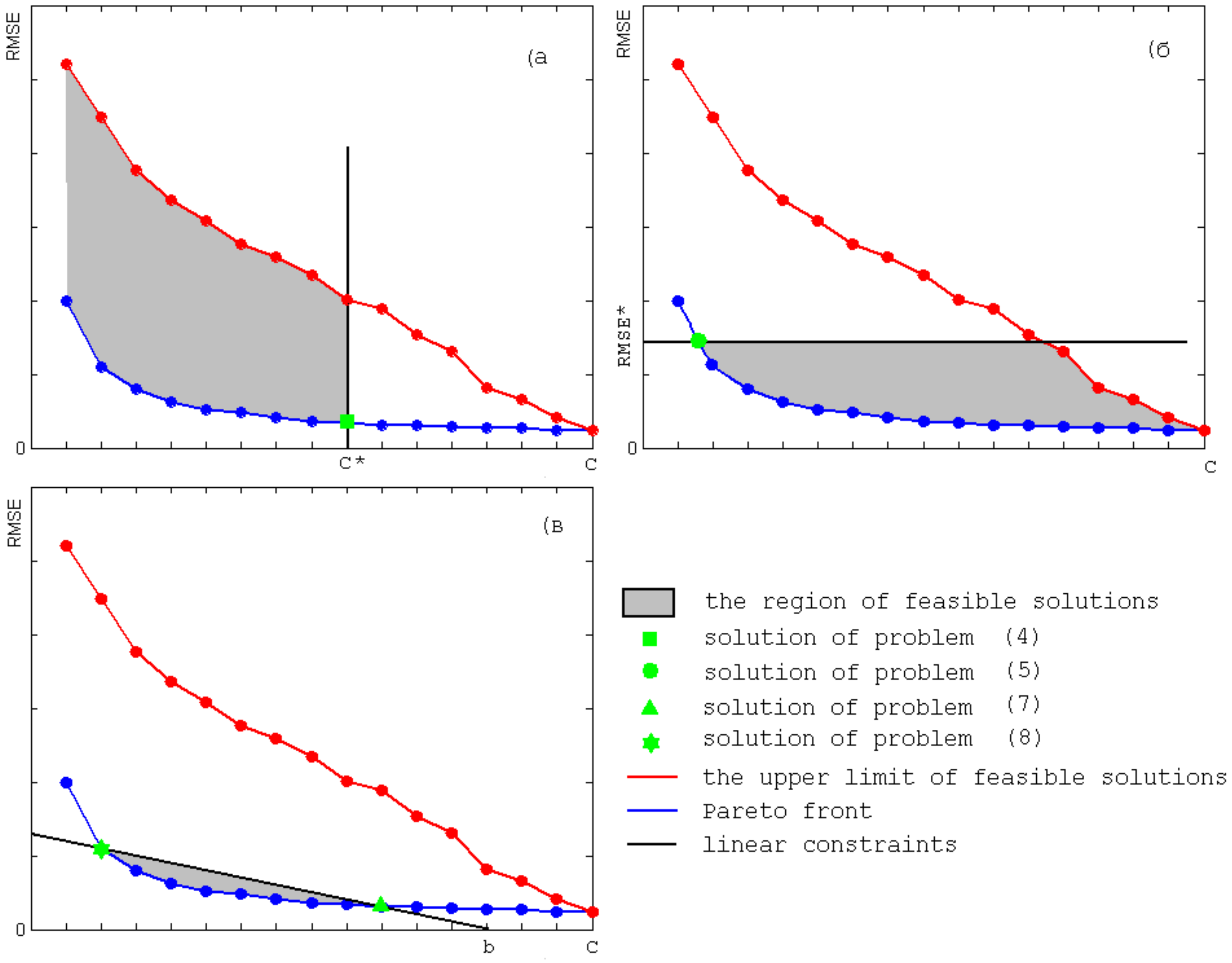

Fig. 1. The region of feasible solutions
a) problem (4); b) problem (5); c) problems (7) and (8)

2. Estimation of the linear constraint parameters

To determine parameters a and b of constraint (6), it is sufficient to know corresponding characteristics of two knowledge bases that satisfy the user. Let us denote them as $(C_1, RMSE_1)$ and $(C_2, RMSE_2)$. Then:

$$\begin{cases} a = \dfrac{RMSE_2 - RMSE_1}{C_2 - C_1} \\ b = RMSE_1 - a \cdot C_1 \end{cases}. \qquad (9)$$

Parameter a could be interpreted as coefficient of accuracy compensation through compactness. It can be calculated from the user's answer to the question "To what extent can the model accuracy be reduced if the number of rules is decreased by 1?" Then, in order to determine the second parameter b it is sufficient to know characteristics of one acceptable knowledge base.

Parameters a and b may be determined by drawing a linear constraint through any two end points of the Pareto front. One of the end points should be on the left and the second one – on the right (Fig. 2). Computational complexity of the full search in order to identify 5 end points of the Pareto front for knowledge bases consisting from 1, 2, N–2, N–1 and N rules is quadratic $O(N^2)$ and, therefore, such approach can be also applied to high-dimensionality problems.

A linear constraint could be also drawn through two points of the learning curve in the form of dependence of the residual on the knowledge base complexity. The learning curve is proposed to be built on the results of the greedy algorithm execution for selecting the rules. This algorithm consists in adding one rule to the knowledge base at each step, which results in maximum reduction of the residual. The obtained learning curve will never be lower than the Pareto front (see Fig. 2). As the initial base for the greedy algorithm a knowledge base containing 2 or N – 2 rules could be chosen from the Pareto front. The greedy algorithm has quadratic computational complexity.

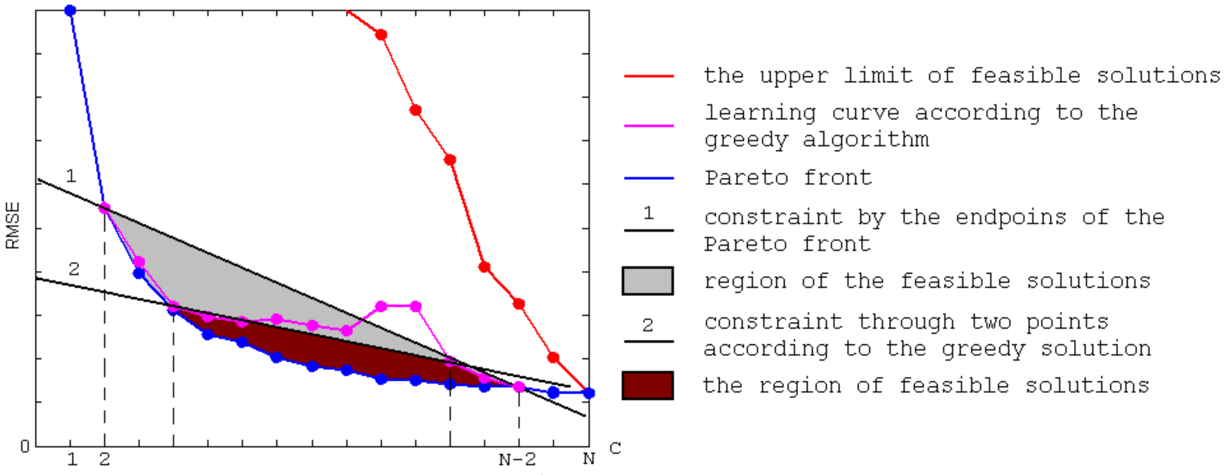

Fig. 2. For linear constraint parameters calculation

3. A genetic algorithm of solving the problem

In order to solve the optimization problem, we use a genetic algorithm based on Pittsburg approach. Each chromosome of the population describes a fuzzy knowledge base with its own set of rules R'. Each of N genes of this chromosome can take the following values: 1 (if the corresponding rule gets into the knowledge base) and 0 (if it does not get there).

The initial population is generated randomly but with the inclusion of suboptimal solutions found by the greedy algorithm.

The probability of selecting a chromosome for crossover is determined as follows:

$$p = \frac{n-j}{2n}, \tag{10}$$

where n is the size of population; j – rank of the chromosome that is determined by the fitness function.

β-fraction of the chromosomes obtained as a result of crossover is subjected to mutation.

Selection is carried out by a deterministic choice of n best chromosomes.

4. Computer experiments

The experiments are carried out for singleton fuzzy knowledge bases where the antecedents of the rules are defined by fuzzy sets and the consequents – by numerical values [1]. As in our previous papers on the formalized multicriteria design of fuzzy knowledge bases [8 - 11], experiments are conducted using three target dependences (Fig. 3):

$$\text{Growing} - y = a\sqrt{b} \quad a \in [2;22],\ b \in [2;14]; \tag{11}$$

$$\text{unimodal} - \; y = -a^2 - b^2, \; a \in [-7;3], \; b \in [-5;5]; \qquad (12)$$

$$\text{multiextremal} - \; y = (1 + \sin(a)^2)^b, \; a \in [0;5], \; b \in [0.5;2]. \qquad (13)$$

For each dependence (11) – (13) a full singleton fuzzy knowledge base from $N = 16$ rules is created (Table 1). Fuzzification of the input variables has been carried out by Gaussian membership functions [1] (Fig. 4). Consequents of the rules are calculated by functions (11) - (13) with the arguments equal to the cores of fuzzy sets consisting from the antecedents of the rules.

Table 1

Full sets of rules (R) for each dependence

№	a	b	y, for dependence (11)	y, for dependence (12)	y, for dependence (13)
1	Very low	Very low	5,04	-71,91	0,95
2	Low	Very low	14,04	-48,94	0,81
3	Medium	Very low	24,84	-45,27	0,94
4	High	Very low	33,84	-62,14	0,79
5	Very low	Low	7,59	-46,08	1,04
6	Low	Low	21,14	-23,11	1,23
7	Medium	Low	37,4	-19,44	1,06
8	High	Low	50,95	-36,3	1,26
9	Very low	Medium	9,82	-31,08	1,17
10	Low	Medium	23,37	-8,1	2,02
11	Medium	Medium	48,42	-4,44	1,25
12	High	Medium	65,96	-21,3	2,17
13	Very low	High	11,36	-31,91	2,29
14	Low	High	31,64	-8,94	3,04
15	Medium	High	55,97	-5,27	1,45
16	High	High	76,25	-22,14	3,40

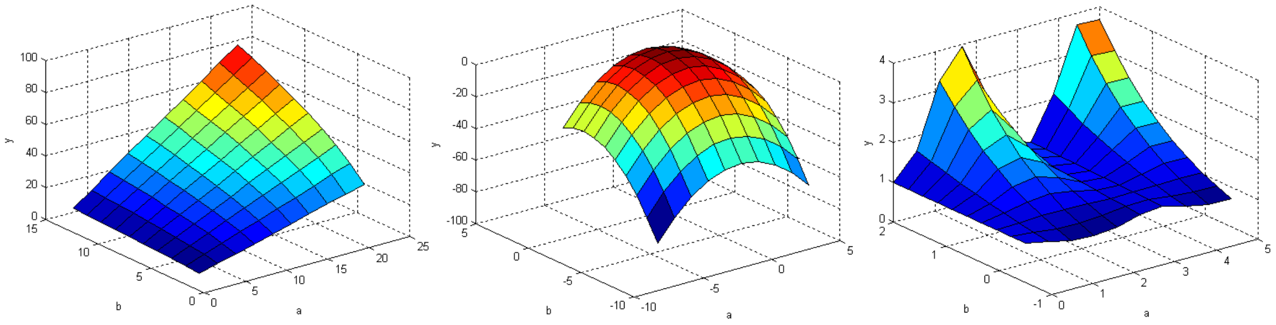

Fig. 3. Surfaces of dependences (11) – (13)

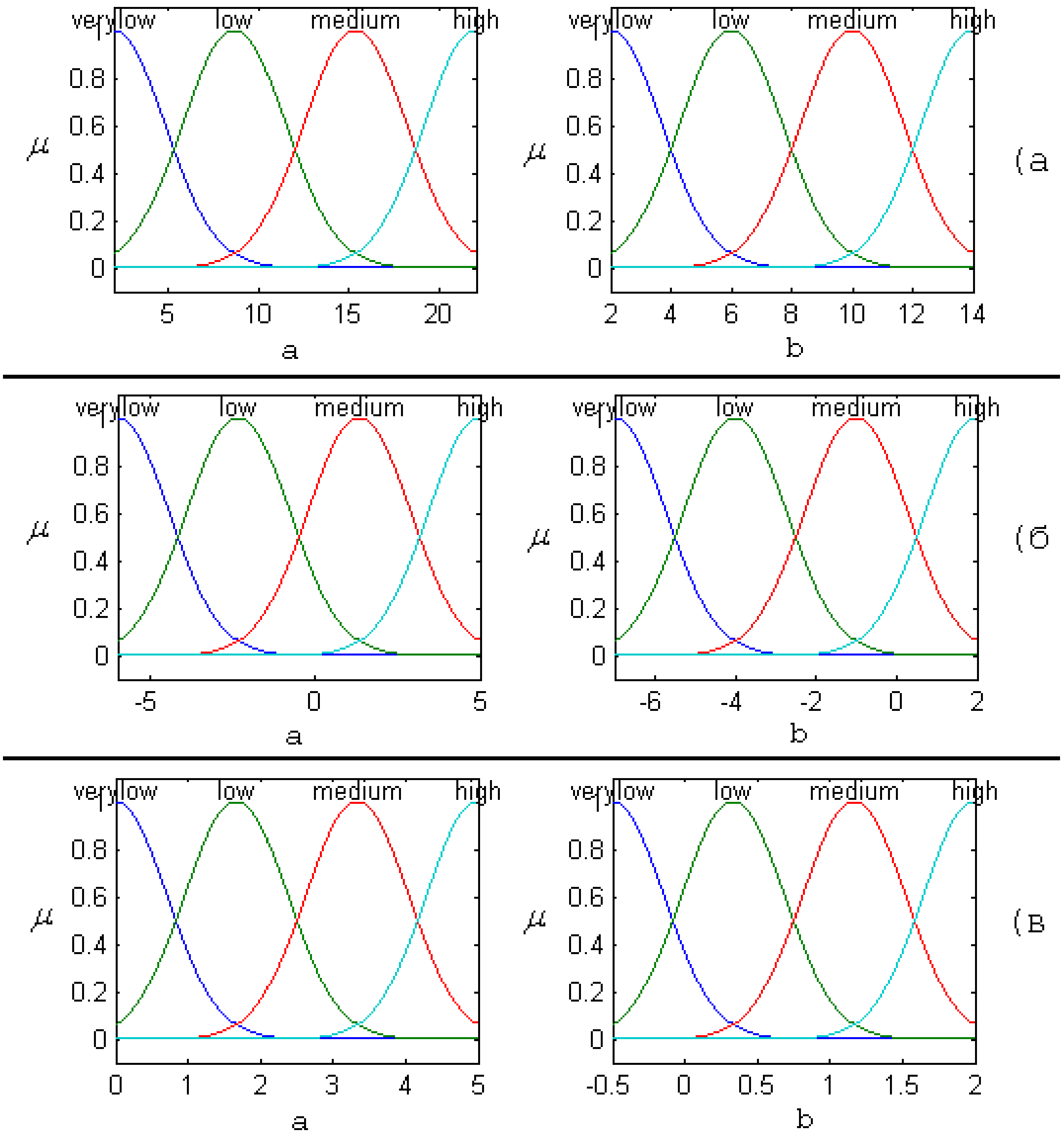

Fig. 4. Membership functions of the terms of input variables for:
a) dependence (11); b) dependence (12); c) dependence (13)

Parameters of the linear constraint in (7) – (8) are found for each experiment separately. First, using the greedy algorithm, we find the best knowledge bases with different numbers of rules for the problem with target dependence (11). Considering the obtained learning curve to be the reference, we set the desired value of the residual for the knowledge base with 4 rules so that it is slightly less than that in Fig. 5, for example, at the level of $RMSE \le 0.55$. As the second linear constraint point, a knowledge base with 10 rules is chosen with the residual value being not higher than that for the full knowledge base, i.e. $RMSE \le 0.22$. Substituting this into (9), we obtain $a = -0.0367$ and $b = 0.6968$. For dependence (12) a knowledge base having 6 rules with the residual $RMSE \le 0.75$ and knowledge base having 10 rules with the residual $RMSE \le 0.58$ are considered to be acceptable. Substituting these values into (9), we obtain $a = -0.0425$ and $b = 1.005$. For dependence (13) a knowledge base with 9 rules with the residual $RMSE \le 0.0365$ and the residual compensation level $\Delta RMSE \le -0.00125$ we consider to be an acceptable one. Hence, $a = -0.00125$ and $b = 0.0365 + 9 \cdot 0.00125 = 0.04775$.

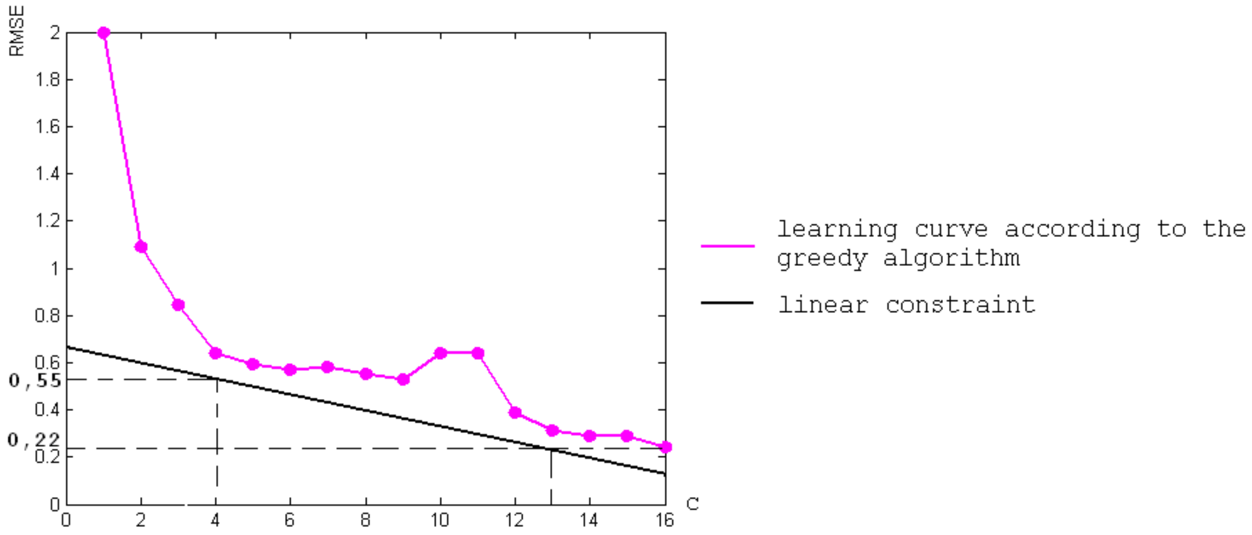

Fig. 5. Estimation of the constraint parameters for experiments with target dependence (11)

The experiments were carried out for the following parameters of the genetic algorithm: size of the population $n = 160$, the number of genes $N = 16$, mutation pressure $\beta = 0.3$, the number of epochs $k = 10$. The obtained solutions of problems (7) and (8) are summarized in table 2. In each of the 6 cases the obtained knowledge bases are located on the Pareto front, i.e. they have the least residual with a fixed number of rules (fig. 6). The Pareto front as well as the upper limit of the region of feasible solutions were found in our previous papers [10, 11] using the exhaustive search among all possible combinations of the fuzzy knowledge base rules (Fig. 6). The computational complexity of the exhaustive search is exponential $O(2^N)$ and, therefore, in all test problems 65536 variants of the fuzzy knowledge base are to be checked. The proposed genetic algorithm has found the global solution after going through 1600 variants for each problem.

Table 2

Results of the experiments

Target dependence	(11)	(11)	(12)	(12)	(13)	(13)
Problem statement	(7)	(8)	(7)	(8)	(7)	(8)
Constraint parameters	a=-0,0367 b=0,6968	a=-0,0367 b=0,6968	a=-0,0425 b=1,005	a=-0,0425 b=1,005	a=-0,00125 b=0,04775	a=-0,00125 b=0,04775
Solution (R')	(1; 2; 5; 6; 7; 9; 10; 12; 16)	(2; 3; 6; 7; 11; 16)	(1; 2; 4; 5; 6; 7; 8; 10; 11; 13; 16)	(1; 3; 6; 11; 12)	(2; 5; 7; 9; 10; 11; 12; 13; 14; 15; 16)	(1; 3; 11; 14; 16)
C(R')	10	5	11	5	11	5
RMSE(R')	0,3098	0,5128	0,5134	0,7235	0,0334	0,0387

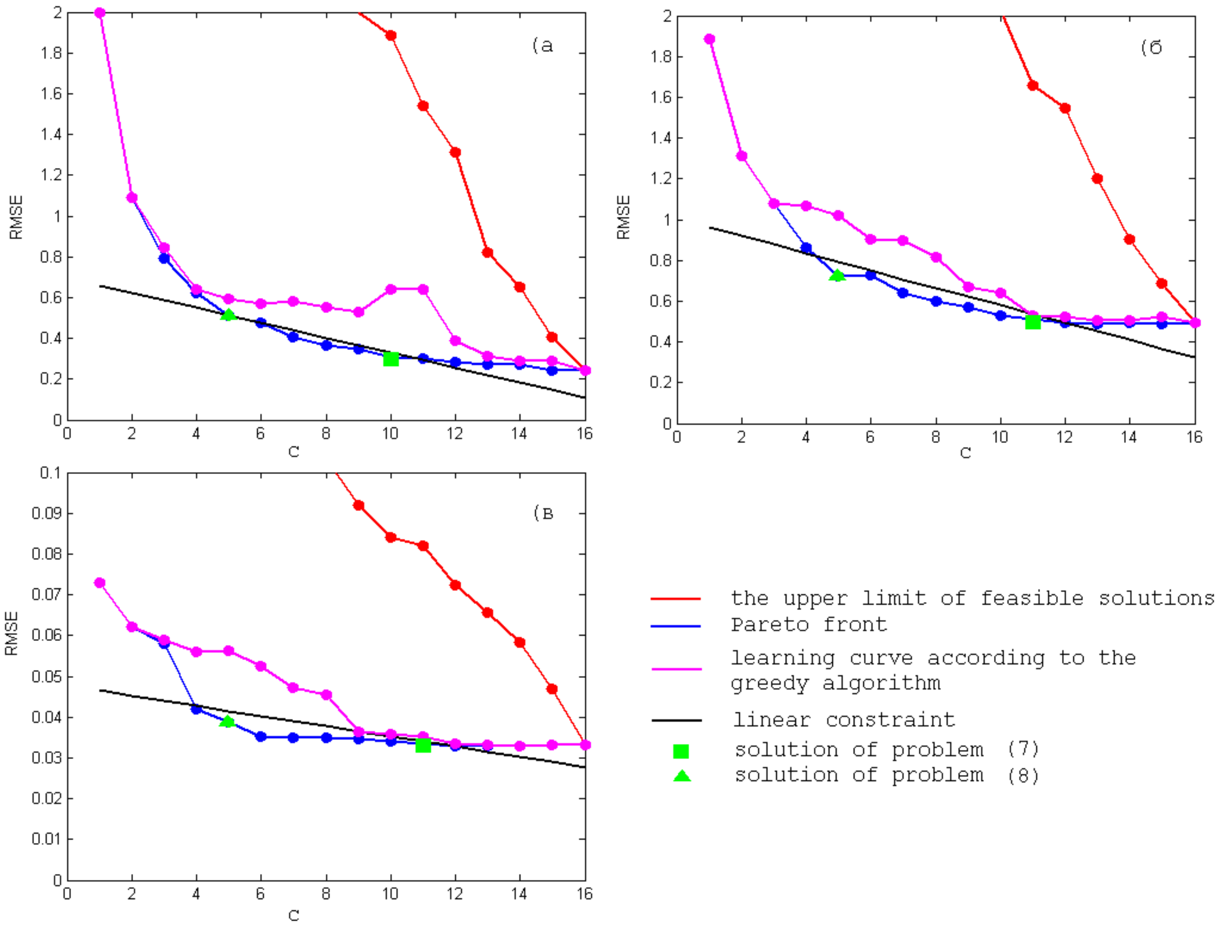

Fig. 6. Learning curves of the fuzzy knowledge base for:
a) dependence (11); b) dependence (12); c) dependence (13)

Conclusions

A new method for solving one of the fuzzy identification problems that aims at selecting a fuzzy knowledge base rules taking into account accuracy and compactness is proposed. The novelty of the method consists in the use of the linear constraint instead of the limit restrictions on accuracy and complexity, which sets the level of compensation between these two conflicting criteria. Application of new constraints enables significant narrowing of the region of feasible solutions confining it to the neighborhood of Pareto front. The computer experiments have proved that with the new statement of the problem the genetic algorithm finds the global optimum after generating tens of times less variants of fuzzy knowledge bases than in the case of exhaustive search.

REFERENCES

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

2. Martello S. Knapsack problems: algorithms and computer implementations / S. Martello , P. Toth. – New York: John Wiley & Sons, Inc, 1990. – 296 p.

3. Ishibuchi H. Selecting fuzzy if-then rules for classification problems using genetic algorithms / H. Ishibuchi, K. Nozaki, N. Yamamoto, H. Tanaka // IEEE Transactions on Fuzzy Systems. – 1995. – Vol. 3, No. 3. – P. 260 – 270.

4. 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 – 150.

5. Ishibuchi H. Three-objective genetics-based machine learning for linguistic rule extraction / H. Ishibuchi, T. Nakashima, T. Murata // Inform. Sci. – 2001. – Vol. 136, No. 1. – P. 109 – 133.

6. Cordon O. A historical review of evolutionary learning methods for Mamdani-type fuzzy rule-based systems: Designing interpretable genetic fuzzy systems / O. Cordon // International Journal of Approximate Reasoning. – 2011. – Vol. 52. – P. 894 – 913.

7. Cordon O. Ten years of genetic fuzzy systems: current framework and new trends / O. Cordon, F. Gomideb, F. Herreraa, F. Homannc, L. Magdalenad // Fuzzy Sets and Systems. – 2004. – Vol. 141. – P. 5 – 31.

8. 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.

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

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

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

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

Mazurenko Viktor – Postgraduate student of the Department of Computer Control Systems.

Savchuk Dmytro – Student of the Institute of Automatics, Electronics and Computer Control Systems.
Vinnytsia National Technical University.