7N

UDC 004.8

***N. Novoselova*[1], *I. Tom*[1], *S. Ablameyko*[2]**
[1]United Institute of Informatics Problems
National Academy of Sciences of Belarus, Minsk
[2]Belarusian State University, Minsk
{novosel, tom}@newman.bas-net.by, ablameyko@bsu.by

Evolutionary Design of the Classifier Ensemble

This paper[1] presents two novel approaches to evolutionary design of the classifier ensemble. The first one presents the task of one-objective optimization of feature set partitioning together with feature weighting for the construction of the individual classifiers. The second approach deals with multi-objective optimization of classifier ensemble design. The proposed approaches have been tested on two data sets from the machine learning repository and one real data set on transient ischemic attack. The experiments show the advantages of the feature weighting in terms of classification accuracy when dealing with multivariate data sets and the possibility in one run of multi-objective genetic algorithm to get the non-dominated ensembles of different sizes and thereby skip the tedious process of iterative search for the best ensemble of fixed size.

Introduction

According to the literature [1-3] application of the classifier combination to solving the practical tasks allows to improve the classification accuracy. The combined decision is supposed to be better (more accurate, more reliable) than the classification decision of the best individual classifier. Among the existing methods of designing the classifier ensembles the “bagging” and “boosting” [4-6] are the most popular. They are based on the manipulations with initial training set in order to build several classifiers. Theoretical and empirical investigations of the classifier ensembles show that the most prosperous approach is the combination of independent classifiers [7]. One of the effective methods of independent classifiers construction is the training the individuals of ensemble on the different features subsets [8], [9]. Hereby, in most cases the design of the classifier ensemble using the partitioning of the initial features set, which describes the data objects, has the advantages. There are a lot of papers, devoted to the study of the properties of classifier ensembles, constructed with the different feature subsets. For example, in [9] the authors demonstrate the possibility to use the randomized feature subsets for the design of classifier ensemble. However, this approach is inefficient for the high-dimensional feature space. In [2] the heuristic algorithm is applied for the partition of the feature set into several uncorrelated subsets, which because of being locally optimal doesn’t guarantee the best result.

In this paper we present novel approaches to evolutionary design of the classifier ensembles. The approaches utilize the genetic algorithm (GA) for the purpose of simultaneous selection of several feature subsets for the construction of individual classifiers, which constitute the ensemble. The use of GA for solving the optimization task, which consists in the partition of the initial feature set for the construction of efficient classifier ensemble, is adopted by following reasons:

– simplicity of coding the solution of optimization task;

– absence of the restriction to smoothness of the optimizable function that allows to use as such the classification accuracy of classifier ensemble;

[1] This paper was prepared under the financial support of the Belarusian Republican Foundation for Fundamental Research [grant №Ф10ЛАТ-015].

– lack of efficient suboptimal algorithms for the selection of feature subsets to be used by individual classifiers, comprising the ensemble.

The first approach studies the influence of feature weighting on the classification performance of the ensemble. For this purpose we extend the proposed in [10] GA by taking into account both the search for optimal partitioning of feature set and corresponding feature weights. The second approach consists in formulating and solving the multi-objective optimization task of classifier ensemble design by considering it as an objective, where apart from classification accuracy the error independence criteria is optimized. As a result the several non-dominated solutions with different number of the individual classifiers in the ensemble (the size of the ensemble) can be obtained in one run of GA. The single ensemble can be further selected as one, providing the best classification accuracy. The experimental results have shown that the selected ensemble in most cases gives the result, which is comparable with the best solution from several single objective GA runs, each with the fixed number of individual classifiers.

Formal definition of classifier ensemble

Let $C=\{v_1,\dots,v_c\}$ be a set of class labels and $x=\left[x_1,\dots,x_N\right]^T \in R^N$ is the feature set, describing a data object. An individual classifier is the function of the following form:

$$F: R^N \to \left[0,1\right]^c, \quad (1)$$

where $F(x)$ is a c-dimensional vector, the i-th element of which defines the membership degree of the data object x to the class v_i, i=1,…,c. In order to get the final classification decision the outputs of m individual classifiers, which constitute the ensemble, are aggregated as follows:

$$F(x)=A(F_1(x),\dots,F_m(x)), \quad (2)$$

where A is the aggregation operator. The output of each individual classifier for particular data object x is the c-dimensional vector $F_i(x)=\left[f_{i,1}(x),\dots,f_{i,c}(x)\right]^T$, i=1,…,m. The output of the classifier ensemble is the c-dimensional vector: $F(x)=\left[g_1(x),\dots,g_c(x)\right]^T$. The selection of the single class label v_s for the data object x is performed according to the maximal membership degree:

$f_{i,s}(x)\ge f_{i,j}(x)\ \forall j=1,\dots,c$ is for the individual classifiers;

$g_s(x)\ge g_t(x),\ \forall t=1,\dots,c$ is for the whole ensemble.

There exist different aggregation operators, on the basis of which the combination of the outputs of the individual classifiers is executed. Among them are the following operators: maximum, minimum, product, average, majority vote, etc. In our study the individual classifiers are combined using the popular and simple in realization majority vote operator.

Let c-dimensional vector $F_i(x)=\left[f_{i,1}(x),\dots,f_{i,c}(x)\right]^T \in \left[0,1\right]^c$ be the output of individual classifier F_i, i=1,…,m for the input object x. The value $f_{i,j}(x)\in\left[0,1\right]$ is the degree of belonging of x to class v_j, which is defined by classifier F_i. In order to determine the vote of the classifier in support of a single class we use the coarse classification decision and select the class according to the following expression:

$$v_s \Leftrightarrow f_{i,s}(x)=\max_j\left\{f_{i,j}(x)\right\}. \quad (3)$$

Hereby the classification decision for each individual classifier F_i is formulated as binary vector F_i^h with s-th element equals to one and other elements equal to zeros:

$$f_{i,j}^h(x) = \begin{cases} 1, & j = s \\ 0, & j \neq s \end{cases}. \tag{4}$$

A decision by combination of classifiers using majority vote aggregation A_{maj} can be presented as a c-dimensional vector and is calculated as follows:

$$\begin{gathered} A_{maj} \equiv F(x) = \left[f_1(x), \ldots, f_c(x) \right]^T, \quad f_j(x) \in \{0,1\}, \text{ j=1,… c} \\ \text{and} \qquad\qquad\qquad\qquad\qquad\qquad\qquad , \\ f_j(x) = \begin{cases} 1, & \sum_{i=1}^{m} f_{i,j}^h(x) = \max_{s=1,\ldots,c} \sum_{i=1}^{m} f_{i,s}^h(x) \\ 0, & \text{otherwise} \end{cases} \end{gathered} \tag{5}$$

where m is the number of individual classifiers in the ensemble.

In our research for the design of classifier ensemble we used the different subsets of initial features. The k-nearest neighbor classifier is applied as the individual classifier of the ensemble.

Design of classifier ensemble with feature weighting

We propose the novel approach to classifier ensemble design, based on the GA with modified realization scheme [10]. The approach consists in extending the task of optimization of the partitioning of the feature set into subsets to be used by the individual classifiers by simultaneous search for feature weights. The optimization task in this case can be formulated as follows:

Let Φ be the set of all the partitions of the initial feature set, describing the data objects, into m subsets. Each subset corresponds to individual classifier. Each partition is a particular combination of input features from the maximum possible number of combinations $(m+1)^N \times [0,1]^N$, where N is the number of input features. It's required to find such a partition $S \in \Phi$ of a feature set, which is the solution of the optimization task with one optimization criteria

$$\max f_1(S),$$

where $f_1(S)$ is the number of data objects, that was correctly classified using the classifier ensemble.

The Fig. 1 presents the main realization scheme of the evolutionary ensemble design. The GA initial generation is randomly formed by different partitions of the whole feature set B of the training sample into m subsets $B^j, 1 \le j \le m$. The construction of the j-th individual classifier is based on the corresponding j-th feature subset. The classification decisions of the individual classifiers are then combined using the majority voting aggregation operator, thereby defining the decision of the classifier ensemble. After that the genetic operations of recombination and selection of the GA individuals into the new generation are iteratively performed, converging sequentially to the optimal solution. The accuracy of the data set classification by the classifier ensemble stands as the GA fitness function.

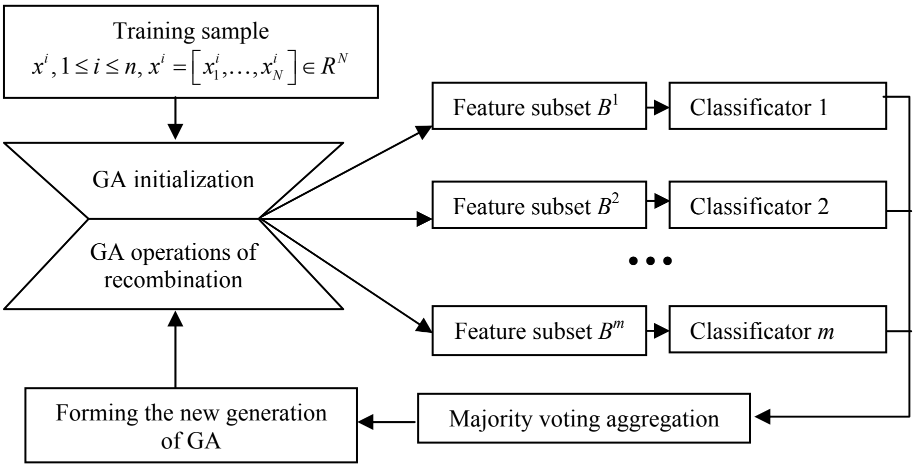

Figure 1 – General realization scheme

The GA individual represents the whole initial feature set, where each feature is related to the particular subset, and i-th gene corresponds to i-th feature. In the previous paper [10] the following two coding schemes are used:

1) according to the first scheme each gene takes the integer value in the interval $[0,m]$, where 0 means that the feature is not used, and an integer k, $1 \le k \le m$ means that the feature is used by k-th classifier. In this case the set of initial features is partitioned into the non-overlapping subsets. The search space of the partitioning task equals $(m+1)^N$, where N is the number of input features. For example, when m=3, and the number of features N=7, the possible graphical view of the GA individual is depicted in Fig. 2.

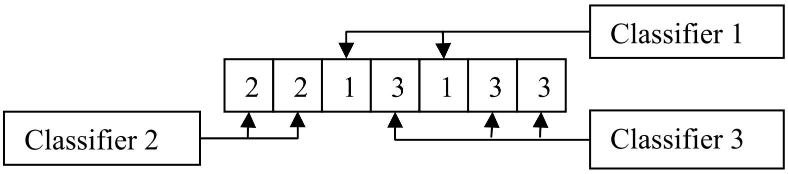

Figure 2 – Example of the first GA encoding scheme

2) according to the second scheme it's possible to encode the overlapping feature subsets. In this case the search space of the partitioning task equals $(2^m)^N$, where N is the number of input features. The example of GA individual, encoding three feature subsets with the number of features N=7 is shown in Fig. 3.

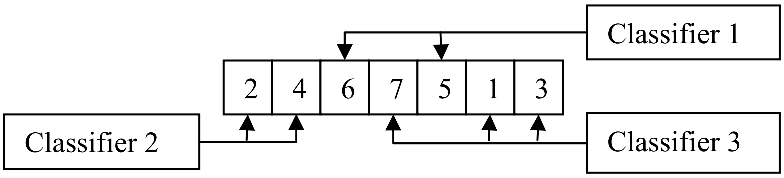

Figure 3 – Example of the second GA encoding scheme

For the encoding of the partition of the feature set into the three overlapped subsets, presented in Fig. 3, the following notations are used:

when the value of gene equals 0, the corresponding feature is not used by any classifier of the ensemble;

when the value of gene equals 1, the corresponding feature is a part of the first subset;

when the value of gene equals 2, the corresponding feature is a part of the second subset;

when the value of gene equals 3, the corresponding feature is a part of the third subset;

when the value of gene equals 4, the corresponding feature is a part of the first and the second subset;

when the value of gene equals 5, the corresponding feature is a part of the first and the third subset;

when the value of gene equals 6, the corresponding feature is a part of the second and the third subset;

when the value of gene equals 7, the corresponding feature is a part of all three subsets.

In our paper the GA individual apart from the feature partitioning encodes the vector of feature weights as a real numbers in the interval [0,1]. The example of GA individual, encoding three feature subsets with feature weights and the number of features N=7 is shown in Fig. 4. By such encoding of GA individual it's possible to simultaneously solve the task of feature scaling and feature partitioning for classifier ensemble design, which allows to define not only the feature subsets for individual classifiers but also their information value.

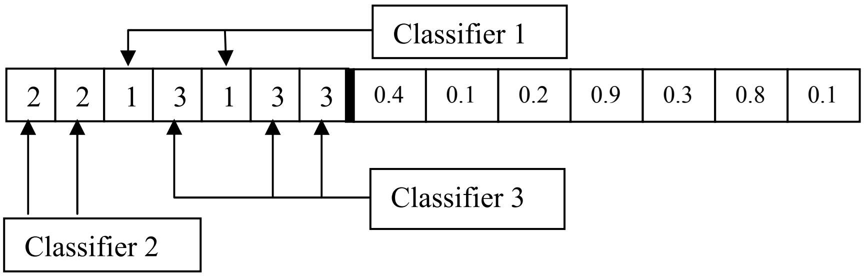

Figure 4 – Example of the GA with feature weights (first encoding scheme)

For the realization of the proposed approach to the evolutionary design of classifier ensemble by means of feature weighting the different genetic operations of crossover and mutation are applied to binary and real part of the GA individual.

Multi-objective optimization task of classifier ensemble design

The second approach consists in formulating and solving the multi-objective task of the classifier ensemble design. The two objective functions are considered: classification accuracy and error independence criteria, which emphasize the independence of individual classifiers. Hence the task of classifier ensemble design can be formulated as follows:

Let Φ be the set of all the partitions of the initial feature set, describing the data objects, into m subsets. Each subset corresponds to individual classifier. Each partition is a particular combination of input features from the maximum possible number of combinations $(m+1)^N$, where N is the number of input features. It's required to find such a partition $S \in \Phi$ of a feature set, which is the solution of the optimization task with two optimization criteria

$$\max f_1(S), \min f_2(S) ,$$

where $f_1(S)$ is the number of data objects, that was correctly classified using classifier ensemble, $f_2(S)$ is the value of the error independence criteria.

As it was stated in the introduction the best classification accuracy of ensemble can be reached by the combination of independent individual classifiers. For providing more effective search of classifier ensembles with varying size we propose the use of error independent criteria *E*. To calculate the value of the error independent criteria the number of the wrong votes for each data object (assignment to the wrong class), that was produced by individual classifiers, is defined. Then the *E* is equal to the maximum of this number for all the data objects and must be minimized. In the case of single criteria $f_2(S)=E$ optimization, the optimal partition will tend to the empty ensemble. We suppose, that simultaneous optimization of the second criteria $f_1(S)$ will compensate this trend.

In our research we used Nondominated Sorting Genetic Algorithm [11] to perform the multi-objective optimization.

Experimental results and discussion

The proposed approaches to the design of the classifier ensemble with GA has been tested on two data sets (Table 1) from the machine learning repository [12] and one real data set on transient ischemic attack (TIA)[1]. The classification accuracies of the proposed approaches are compared with standard *k*-nearest neighbor classifier with all the features and the feature selection using GA in individual classifier.

For the estimation of the accuracy of classifier ensemble we used 10-fold cross-validation. The cross-validation consists in splitting the data set into 10 subsets and iteratively considering each single subset as a test sample, while training the ensemble on the rest nine subsets. For the real dataset TIA we used 5-fold cross-validation algorithm.

Table 1 – Description of the data sets

Data set	Number of objects	Number of features	*Number of classes*
Heart	303	13	*2*
Wine	178	13	*3*
TIA	*101*	*41*	*4*

Five different experiments concerning the design of classifier ensemble have been made:

1. GA-selection of the feature subset for the construction of single classifier: without feature weights and together with feature weighting [13].
2. Design of the ensemble with three classifiers, based on the non-overlapping feature subsets: without feature weights and together with feature weighting.
3. Design of the ensemble with five classifiers, based on the non-overlapping feature subsets: without feature weights and together with feature weighting.
4. Design of the ensemble with three classifiers, based on the overlapping feature subsets: without feature weights and together with feature weighting.
5. Design of several non-dominated ensembles by multi-objective optimization without feature weights.

The GA parameters, selected for the design the classifier ensembles, are as follows:

- Population size: 50-100
- Maximal number of generations: 100
- Crossover probability: $P_{\text{cross}} = 0.8$
- Mutation probability: $P_{\text{mut}} = 0.1$

The experimental results of the proposed approach with feature weighting for each analyzed data set are presented in Tables 2-4. In the column “Classification accuracy” the mean classification accuracy of the train/test samples are indicated.

[1] The authors are much obliged to A.S. Mastikin (Belarus State Medical University) for providing the TIA data set

According to Table 2 the classification accuracy of the Heart data set is gradually improved with the increase of the number of classifier with non-overlapping feature subsets (with exceptions of some classification accuracies for training samples). The accuracy of the classifier with the overlapped feature subsets is the best. The classification accuracy of the classifier ensembles with feature weights is slightly better than for the same in size ensemble without feature weighting.

Table 2 – Results of experiments for data set Heart

	Number of feature subsets	Classification accuracy (%)		The best GA individual (without feature weights)
		without feature weights	with feature weights	
Classifier (*k*-nearest neighbor classifier)	1 classifier	77,7/79,5		All features
Classifier with feature selection	1 classifier	82,5/75,7	85,8/76,1	0,0,1,0,0,0,0,0,0,0,0,1,1
Classifier ensemble (scheme 1)	3 classifiers	83,5/76,7	84,7/77,4	0,3,2,0,1,1,3,0,1,0,1,3,1 or 2,1,2,0,1,0,1,0,1,0,2,3,3
	5 classifiers	80,9/78,1	85,9/80,1	0,1,5,1,1,3,1,0,3,1,2,4,3
Classifier ensemble (scheme 2)	3 classifiers	87,2/78,5	87,2/81,3	3,7,2,2,4,3,4,3,7,6,0,5,3

Table 3 – Results of experiments for data set Wine

	Number of feature subsets	Classification accuracy (%)		The best GA individual (without feature weights)
		without feature weights	with feature weights	
Classifier (*k*-nearest neighbor classifier)	1 classifier	94,9/94,4		All features
Classifier with feature selection	1 classifier	99,4/92,4	99,6/92,6	1,1,0,0,1,0,1,1,0,1,1,0,1
Classifier ensemble (scheme 1)	3 classifiers	99,5/92,4	99,7/94,9	3,1,1,1,2,3,3,0,2,2,3,0,3
Classifier ensemble (scheme 2)	3 classifiers	99,8/96,7	99,8/93,1	6,6,1,0,5,5,7,0,2,6,6,1,7

Table 4 – Results of experiments for data set TIA

	Number of feature subsets	Classification accuracy (%)	
		without feature weights	with feature weights
Classifier (*k*-nearest neighbor classifier)	1 classifier	52,8/57,5	
Classifier with feature selection	1 classifier	80,2/59,1	85,6/60,5
Classifier ensemble (scheme 1)	3 classifiers	75,7/58,4	78,2/59,3
	5 classifiers	76,7/53,1	80,2/59,3
Classifier ensemble (scheme 2)	3 classifiers	77,3/59,4	81,6/59,5

According to Table 3 the accuracy of the classification of the Wine data set by the classifier using the selected informative features (99,4 % for training sample) is better than using the whole feature set (94,4 %). The classification accuracy of the classifier ensembles with three individual classifier and non-overlapping feature subsets and with three classifiers and overlapping feature subset are only slightly better than the classification accuracy of single classifier with selected subset of informative features. It can be explained by the fact, that almost all features of data set Wine are informative, that can be confirmed by the high classification accuracy of the single classifier with the whole feature set.

According to Table 4 the best classification of the TIA data set is provided by the single classifier with selected informative features (80,2% for training sample). Only the classifier ensembles with feature weighting, which consist from 5 individual classifiers without overlapped feature subsets and 3 individual classifiers with overlapped feature subsets, have reached nearly the same classification as the single classifier with feature selection. As a whole the classifier ensembles with feature weighting are definitely better than for the same in size ensemble without feature weighting.

The most accurate non-dominated solution for two data sets according to experimental results with multi-objective evolutionary design of classifier ensemble are presented in Tables 5 and the intermediate and final GA generations are depicted in Fig 5.

Table 5 – Selected solutions from non-dominated sets

Data sets	Number of feature subsets	Classification accuracy (%)	Error independence criteria	The GA individual
Heart	5 classifiers	86,2	5	0,0,1,3,3,0,3,0,6,3,2,7,6
Wine	5 classifiers	99,5	3	2,1,3,1,5,6,3,0,3,6,2,5,3

The non-dominated solutions represent the classifier ensembles of different sizes, the ensembles with the best classification accuracy for the both analyzed data sets have the biggest number of individual classifiers and the higher value of the error independence criterion. Non-dominated ensembles, presented in Table 5 are comparable in terms of accuracy with the ensembles in Tables 2-3.

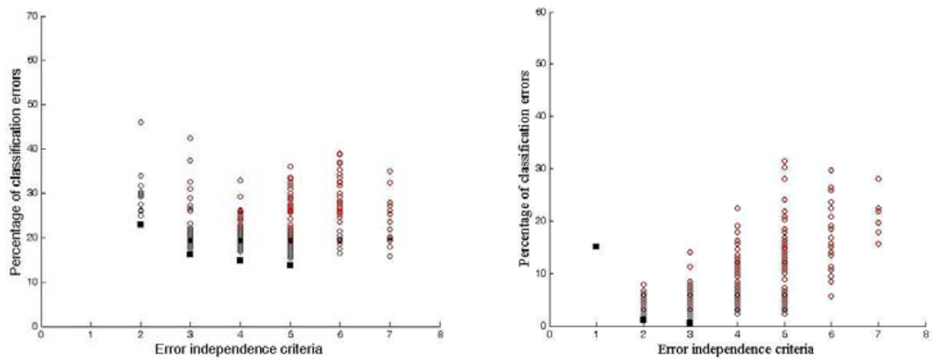

Figure 5 – GA generations and final non-dominated solutions (indicated as filled black squares) in two-dimensional optimization criteria space: data set Heart (left), data set Wine (right)

Conclusions

In the paper two novel approaches to evolutionary design of the classifier ensemble using GA are presented. According to the results of the experiments with three data sets the proposed approach using feature weighting in most cases allows to improve the classification accuracy of the classifier ensembles. The multi-objective optimization for the ensemble design helps to get in one GA run the set of non-dominated solutions with tradeoff between the classification accuracy and error independence criteria. The classification accuracy of the selected ensembles isn't inferior to ones, designed by one-objective optimization.

As the dimensionality of the analyzed data sets is not very high there is a lack of the independent feature subsets and therefore the increase of the number of the individual classifiers doesn't always lead to the increase of the classification accuracy. The further experiments with the multi-dimensional data sets are planned in order to investigate the dependency of the optimal number of feature subsets or the ensemble members and the dimensionality of the data.

Literatura

1. Kotsiantis S. A Combinational Incremental Ensemble of Classifiers as a Technique for Predicting Students' Performance in Distance Education / S. Kotsiantis, K. Patriarcheas, M. Xenos // Knowledge-Based Systems. – 2010. – Vol. 23, № 59. – P. 529-535
2. Vishwath P. Fusion of Multiple Approximate Nearest Neighbor Classifier for Fast and Efficient Classification / P. Vishwath, M.N.Murty, C. Bhatnagar // Information Fusion. – 2004. – Vol. 5. – P. 239-250.
3. Ensemble Classifiers Based on Kernel PCA for Cancer Data Classification / Zhou J., Pan Y., Chen Y., Liu Y. // ICIC 2009. LNCS / Huang D.-S., Jo K.-H., Lee H.-H., Kang H.-J., Bevilacqua V. (eds.). – Vol. 5755. – Springer, Heidelberg (2009). – P. 955-964.
4. Zeng X.D. Optimization of Bagging Classifiers Based on SBCB Algorithm / Zeng X.D., Chao S., Wong F. // Proceedings of the Ninth International Conference on Machine Learning and Cybernetics, IEEE Computer Society. – 2010. – P. 262-267.
5. Breiman L. Bagging predictors / L. Breiman // Machine learning. – 1996. – Vol. 24. – P. 123-140.
6. Hamamoto Y. A bootstrap technique for nearest neighbor classifier design / Y. Hamamoto, S. Uchimura, S. Tomita // IEEE trans. on pattern analysis and machine intelligence. – 1997. – Vol. 19, № 1. – P. 73-79.
7. Tumer K. Decimated Input Ensembles for Improved Generalization / K. Tumer, N.C. Oza // Proceedings of the International Joint Conference on Neural Networks. – Vol. 5. – P.3069-3074.
8. Stefanowski J. Combining Answers of Sub-classifiers in the Bagging-Feature Ensembles / J. Stefanowski // RSEISP 2007. LNCS / Kryszkiewicz M., Peters J.F., Rybinski H., Skowron A. (eds.). – Vol. 4585, Springer, Heidelberg (2007). – P. 574-583.

9. Bryll R. Attribute Bagging: Improving Accuracy of Classifier Ensembles by using Random Feature Subsets / R. Bryll, R. Gutierrez-Osuna, F. Quek // Pattern Recognition. – 2003. – Vol. 36. – P. 1291-1302.
10. Kuncheva L.I. Designing Classifier Fusion Systems by Genetic Algorithms / L.I. Kuncheva, L.C. Jain // IEEE Trans. Evol. Comput. – 2000. – Vol. 4, № 4. – P. 327-336.
11. Deb K. Multi-Objective Optimization using Evolutionary Algorithms / Deb K. – JohnWiley and Sons Ltd, 2002. – [2nd ed.]
12. UC Irvine Machine Learning Repository, http://archive.ics.uci.edu/ml/.
13. Novoselova N.A. Evolutionary Approach to Informative Feature Extraction in the Tasks of Medical Data Analysis (Эволюционный подход к выделению информативных признаков в задачах анализа медицинских данных) / N.A. Novoselova, I.E. Tom, A.S. Mastikin // Artificial Intelligence (Искусственный интеллект ISSN 1561-5367) – 2008. –№ 3. – P. 105-112.

Н.А. Новоселова, И.Э. Том, С.В. Абламейко
Эволюционное построение ансамбля классификаторов
В статье предложены два новых подхода к эволюционному построению ансамбля классификаторов. Первый подход представляет собой задачу однокритериальной оптимизации разбиения множества признаков на отдельные подмножества, которые используются для построения классификаторов ансамбля. Второй подход осуществляет многокритериальную оптимизацию структуры ансамбля классификаторов.

Н.А. Новосьолова, І.Е. Том, С.В. Абламейко
Еволюційна побудова ансамблю класифікаторів
У статті запропоновано два нові підходи до еволюційної побудови ансамблю класифікаторів. Перший підхід є завданням одинкритерійної оптимізації розбиття безлічі ознак на окремі підмножини, які використовуються для побудови класифікаторів ансамблю. Другий підхід здійснює багатокритеріальну оптимізацію структури ансамблю класифікаторів.

Статья поступила в редакцию 22.06.2011.