Information Technology and Management Science
ISSN 2255-9094 (online)
2022, vol. 25, pp. 16–23
https://doi.org/10.7250/itms-2022-0003
https://itms-journals.rtu.lv

Hybrid Classification Model for Biomedical Data Analysis

Natalia Novoselova[1*], Igor Tom[2]
[1,2]*United Institute of Informatics Problems, Minsk, Belarus*

***Abstract* – The paper describes a method for constructing a hybrid classification model that allows combining several sources of biological information in order to build a classifier to identify subtypes of complex diseases. The distinctive feature of the method is its adaptive nature, i.e. the ability to build efficient classifiers regardless of data types, as well as a multi-criteria approach to evaluate the effectiveness of a classification. The testing results on real biomedical data showed the advantages of the proposed hybrid model in comparison with individual classifiers.**

***Keywords* – Classification, efficiency criteria, gene expression, hybrid classifier.**

I. Introduction

Different sources of biological information characterise various changes that occur in the body at the cellular level during the development of a complex disease. In order to take into account the variety of data sources and their limited sample size we have developed a method for constructing a hybrid classification model in order to improve the accuracy of diagnosing subtypes of complex diseases.

The hybrid model is a classification ensemble where classification models built on the same or different sources of biomedical data are considered as base classifiers. The method using several a priori specified classification models [1]–[4] allows determining both the individual classifiers of the ensemble and the structure of the entire hybrid model.

In machine learning ensembles of classifiers have a rather rich history and are mainly constructed on a single data set in order to improve the accuracy of classification [5]–[10]. The theoretical justification for improving the accuracy of classification using an ensemble is the Condorcets theorem [7]. According to this theorem, for a binary classification problem and L base classifiers whose error is less than 0.5, the majority of the ensemble has an error lower than an individual classifier if the errors of individual members of the ensemble are not correlated.

For example, if we have 21 classifiers, and the probability of error for each base classifier is $p = 0.3$ and the errors are independent, the majority vote ensemble error probability is calculated using the binomial distribution, as in Eq. (1). For the case where more than $L/2$ classifiers gave an erroneous result the probability of error P_{error} is:

$$P_{\text{error}} = \sum_{(i=\lfloor L/2 \rfloor)}^{L} \binom{L}{i} p^i (1-p)^{L-i} \Rightarrow P_{\text{error}} = 0.026 \ll p = 0.3 \quad (1)$$

An important component of building ensembles of classifiers is the search for a compromise between the accuracy and independence of base classifiers, since more accurate classifiers tend to be more dependent [5]. From this point of view, ensembles built on heterogeneous data sources can be quite guaranteed to improve the accuracy of classification and prediction of subtypes of complex diseases or to differentiate a case from control.

The proposed method for constructing a hybrid classification model is based on the fact that, as a rule, different classifiers are most effective for different data sets, and therefore, it is a difficult task to initially select the best classifier for data from a particular source of biomedical information. The method combines the bagging procedure and aggregation of ranked lists to build base classifiers, which allows adaptively adjusting the ensemble considering the type of data being classified. To ensure independence in the construction of the ensemble, various types of base classifiers are considered [1]. In addition, the assessment of the base classifiers of the ensemble is based on the calculation of several classification efficiency criteria [1] and allows selecting the most optimal combination of base classifiers corresponding to the maximum value of the efficiency of the entire ensemble.

II. Efficiency Criteria of Classification

To assess the efficiency of classifiers, in addition to the standard accuracy criterion, criteria are used that allow considering the imbalance between the number of elements in individual classes. More informative criteria can be obtained by evaluating the correspondence between the actual and predicted class labels on the test set.

Let $B = \{B_1, B_2, \ldots, B_c\}$ define the partition of the test set into groups according to their real class label l_i, where

$$B_j = \left\{ x_i \in B \mid l_i = k_j \right\}. \quad (2)$$

* Corresponding author. E-mail: novos65@gmail.com

Let $|B_i| = n_i$ define the power of class k_i. The set $R = \{R_1, R_2, \ldots, R_c\}$ defines the partitioning of test objects based on the predicted class label $\hat{l}_i$, where

$$R_j = \left\{ x_i \in B \mid \hat{l}_i = k_j \right\}. \quad (3)$$

Let $|R_i| = m_i$ define the number of objects predicted to belong to class k_i. Thus, using the sets R and B the contingency matrix N is defined as

$$N(i,j) = n_{ij} = \left| R_i \cap B_j \right| = \left| \left\{ x_a \in B \mid \hat{l}_a = k_i \; u \; l_a = k_j \right\} \right|, \quad (4)$$

where $1 \le i, j \le c$.

The value n_{ij} corresponds to the number of objects of the class k_i for which the class k_i is predicted, and n_{ij} is the number of objects for which the prediction matches the actual class label, otherwise there is a discrepancy between the predictions and the real class of the objects.

Using the contingency table, a number of information criteria for evaluating the efficiency of classifiers are calculated. The precision $prec_i$ of the classifier D for the class k_i, $i = 1,\ldots, c$ defined as the ratio of correct predictions to all objects for which the class k_i is predicted and is defined as

$$prec_i = \frac{n_{ii}}{m_i}, \quad (5)$$

where m_i is the number of objects for which the class label k_i is predicted.

The overall classifier accuracy is the weighted average of the precisions of the classes as follows

$$Accuracy = \sum_{i=1}^{c} \left(\frac{m_i}{n} \right) prec_i = \frac{1}{n} \sum_{i=1}^{c} n_{ii}. \quad (6)$$

The recall for an individual class k_i, $i = 1,\ldots, c$ is the ratio of correct predictions to all objects of the class k_i and is defined as

$$recall_i = \frac{n_{ii}}{n_i}, \quad (7)$$

where n_i is the number of objects of the class k_i.

In the case of a binary classification, the positive class recall corresponds to the classification sensitivity, and the negative class recall corresponds to the classification specificity.

If the number of classes is relatively small (no more than 100–150 classes), the considered criteria allow evaluating the results of the classification. The higher the accuracy and recall, the better it is. To simultaneously assess the accuracy and recall, a complex metric F-score is used. F-score is the harmonic mean between precision and recall and for class k_i is defined as

$$F_i = \frac{2}{1/prec_i + 1/recall_i} = \frac{2 \cdot prec_i \cdot recall_i}{prec_i + recall_i} = \frac{2n_{ii}}{n_i + m_i}. \quad (8)$$

The overall F-score for the classifier is the average of the F-scores for the individual classes.

III. Method for Constructing a Hybrid Classification Model

As part of the research, a general scheme of a hybrid classification model has been developed, which allows combining several sources of biological information about patients in order to build a classification model that allows diagnosing subtypes of complex diseases or differentiating the case from control. The proposed hybrid model is a classification ensemble with the following distinctive features:

- unified presentation of information from various data sources, including harmonization of the list of cases and genes/proteins;
- implementation of the procedure for selecting classification features for each individual data source [11], [12];
- construction of the base or individual classifiers of a hybrid model, which can be either a single classifier or an ensemble of classifiers built on the same data source;
- implementation of several integrating schemes of the individual classifiers.

The general scheme of the hybrid model is shown in Fig. 1.

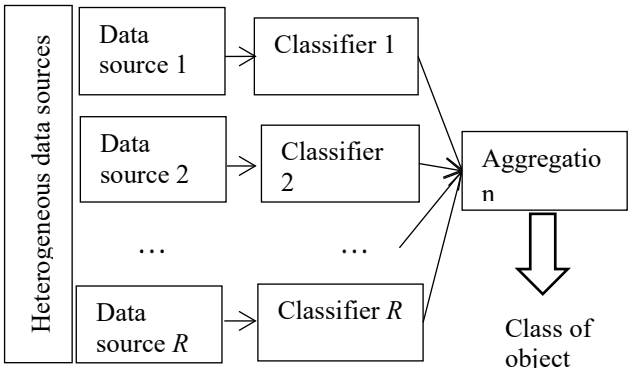

Fig. 1. Diagram of a hybrid classification model.

The proposed adaptive method for constructing the structure of a hybrid classification model combines the procedure of bagging and aggregation of ranked lists.

The method simultaneously uses several efficiency criteria, such as accuracy, sensitivity and specificity, to select the optimal base classifier for a particular data source. This allows both increasing the stability of the model selection under conditions of a limited training sample and increasing the generalizing ability of the hybrid classification model.

Each classification efficiency criterion allows ranking the classifiers according to its values. In the case of simultaneous evaluation of classification results by several criteria, the most optimal classifier is determined using weighted aggregation of ranked lists. The ordered lists $L_1,\ldots, L_K$ of classifiers, where K is the number of efficiency criteria, are aggregated to obtain a single combined list of classifiers ordered simultaneously by all K criteria. The optimization objective function in this case is defined as follows:

$$\Phi(\delta) = \sum_{i=1}^{K} w_i d(\delta, L_i), \quad (9)$$

where δ is a ranked list of simple classification models of size M; d is the proximity function of any pair of ordered lists and w_i is the weight coefficient corresponding to the value of the efficiency criterion.

In the proposed method, when M is small enough ($M < 8$), the aggregation of the ranked lists in order to minimize the function Φ(δ) is carried out using the exhaustive search procedure. For more complex optimization problems, combinatorial optimization algorithms can be used. The weight coefficients w_i make it possible to increase or decrease the influence of individual performance criteria and contribute to the construction of an adaptive structure of the hybrid model.

General scheme of the proposed method for constructing the base classifiers of the hybrid model is presented below.

Step 1. Initialization.

A set of training data $\{X_{(n\times p)}, y_{(n\times 1)}\}$ is given, where n is the size of the training sample, p is the dimension of the feature space. Determine N as the number of subsamples for the bagging procedure.

Step 2. Sampling.

Generate the j-th subsample $\{X_j^*, y_j^*\}$ of size n using random selection with replacement. The selection continues until the objects of each class are presented in the subsample. As a result, some objects can be selected more than once, while a number of objects will not be included in the subsample. Unpresented objects form an OOB (out-of-bag) set.

Step 3. Classification.

Train M classifiers using the j-th subsample.

Step 4. Performance evaluation.

Use M trained classifiers to predict class labels for objects in the OOB set $\{X_j^{oob^*}, y_j^{oob^*}\}$ not included in the j-th subsample. Using known class labels, calculate the values of K efficiency criteria. Rank the classifiers according to the values of each efficiency criterion and make K ordered lists $L_1, \ldots, L_K$ of size M.

Step 5. List aggregation.

The ordered lists $L_1, \ldots, L_K$ are aggregated using weighted rank aggregation, which will allow determining the best classifier $A_{(1)}^j$.

Repeat steps 2–5 N times.

The above procedure allows you to build the base classifiers of the hybrid model, which are both ensembles of classifiers and components of the hybrid classification model.

To predict the class for a new object it is necessary to calculate the class label for all N classifiers of the ensemble. The base classifiers of the ensemble may represent different classification models, and not the same classifier, as, for example, in the tree forest method [6]. As a result, the final class label of the ensemble is determined by the combinational scheme of the ensemble elements. Below are the steps of the class prediction procedure.

Step 1. Individual predictions.

Use N "best" individual models $A_{(1)}^1, \ldots, A_{(1)}^N$ built for each subsample of the training set to compute N class label predictions for each new object. Given a sample $x_{(p\times 1)}$ let $(\hat{y}_1, \ldots, \hat{y}_N)$ determine N predictions from N individual classifiers.

Step 2. Combination by majority vote.

The final classification is based on the choice of the most frequently occurring class among the N predicted labels, and corresponds to the majority vote classification

$$argmax_c \sum_{i=1}^{N} I(\hat{y}_i = c), \quad (10)$$

where N is the number of subsamples of the bagging procedure and c is one of the class labels.

Step 3. Determination of class probabilities.

The probability of the class c is determined using the proportion of votes for this class

$$P(C = c \mid X = x) = \frac{1}{N} \sum_{i=1}^{N} I(\hat{y}_i = c) \quad (11)$$

IV. Testing the Method and Results of Experiments on Data

To test the method and to compare the effectiveness of individual classification models of ensembles of classifiers, a number of experiments were carried out both on artificially generated and real gene expression data from the TCGA database [13].

The artificial dataset consists of two classes. The objects of the first class are generated according to one of two normal distributions $N(\{a, a, \ldots, a\}, I)$ and $N(\{-a, -a, \ldots, -a\}, I)$ where I is the identity matrix. Objects of the second class are generated according to the multivariate distribution $N(\{a, -a, a, -a, \ldots, a, -a\}, I)$, $a = 2/\sqrt{d}$, where d is the number of features. Fig. 2 shows the data set with $d = 10$.

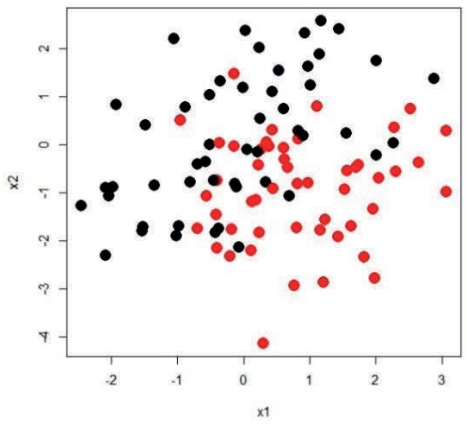

Fig. 2. Projection of an artificially generated data set onto a two-dimensional space x1-x2.

The general scheme of the experiments consists of the following steps:

- to perform the data preprocessing, including substitution of missing values, data standardization;
- to use of training/test samples or cross-validation procedures to evaluate the efficiency of classification models, selection of the number of groups for cross-validation. Choice of efficiency evaluation criteria (accuracy, sensitivity, specificity of the model, AUC (area under curve));
- to select the types of classification algorithms for building a hybrid model;
- to select subsample from data set or several data sets. One subsample for each data set includes the same cases. To construct the number of classifiers for each subsample and select the best one (together with data source) as the base classifier;
- to build a hybrid model by determining the best base classifier in terms of efficiency criteria for each subsample of the bagging procedure using the aggregation function of ranked lists;
- to predict the class for objects of the test sample (or for all groups of the cross-validation procedure). To calculate the values of efficiency criteria, both for individual classifiers and for the hybrid model;
- in the case of cross-validation procedure, to calculate the average values of the efficiency criteria for all cross-validation groups. When using training and test samples, repeat the above steps *n* times ($n = 100$) and calculate the average values of the efficiency criteria.

A. Artificially Generated Data

To test the method a training sample of 100 objects with 1000 features was generated. The following classifiers were selected for building a hybrid classification model: support vector machine (SVM), logistic regression with regularization (PLR), forest of trees (RF), dimensionality reduction (PLS) and forest of trees (PLS+RF), dimensionality reduction (PLS) and linear discriminant analysis (PLS+LDA), dimensionality reduction (PLS) and quadratic discriminant analysis (PLS+QDA), dimensionality reduction (PCA) and linear discriminant analysis (PCA+LDA) and multivariate regression using partial least squares (PLS). Using the PCA and PLS dimensionality reduction methods, five components were extracted. The number of samples of the bagging procedure was 50.

The effectiveness of both individual classifiers and the proposed hybrid classification model was evaluated using the same test sample with 100 objects.

The results of testing the proposed method on artificially generated data are shown in Table I. For the data, the linear discriminant analysis with preliminary dimensionality reduction using PCA and the SVM method is significantly inferior in efficiency to other individual classifiers. Classifiers using PLS dimensionality reduction, such as PLS + RF, PLS + LDA, PLS + QDA, have sufficiently high efficiency criteria and outperform decision tree without dimensionality reduction. PLS + RF has the highest accuracy and specificity criteria, while PLS +LDA has the highest sensitivity value among all individual classifiers. The values of the three criteria for the hybrid model are comparable to the maximum values for the individual classifiers, with the AUC criterion having the highest value.

B. Real biomedical Data

The experiments were carried out on two datasets (Table II), which are the gene expression measurements for cancer patients.

TABLE I

EFFICIENCY CRITERIA FOR THE ARTIFICIAL DATASET

	Accuracy	Sinsitivity	Specificity	AUC
SVM	0.4500* (0.0147)**	0.4620 (0.0222)	0.4380 (0.0203)	0.4384 (0.0194)
RF	0.5550 (0.0047)	0.5668 (0.0333)	0.5432 (0.0300)	0.5836 (0.0057)
PLS+LDA	0.6050 (0.0074)	0.6004 (0.0112)	0.6096 (0.0115)	0.6049 (0.0085)
PCA+LDA	0.4940 (0.0065)	0.5000 (0.0157)	0.4880 (0.0172)	0.4862 (0.0079)
PLS+RF	0.6098 (0.0075)	0.6140 (0.0109)	0.5996 (0.0106)	0.6456 (0.0075)
PLS+QDA	0.5980 (0.0066)	0.5900 (0.0156)	0.6060 (0.0149)	0.5981 (0.0066)
PLR	0.5214 (0.0089)	0.4572 (0.0100)	0.5856 (0.0100)	0.5283 (0.0107)
PLS	0.6004 (0.0070)	0.5668 (0.0163)	0.6020 (0.0158)	0.6448 (0.0081)
Hybrid model	0.6068 (0.0072)	0.6092 (0.0119)	0.6060 (0.0114)	0.6504 (0.0076)

*Mean values of precision, sensitivity, specificity, and AUC for 100 datasets with $N = 100$ data items and $d = 1000$ features.

**Standard deviations.

TABLE II

DESCRIPTION OF REAL DATASETS

Dataset	Number of objects	Number of features	Class labels
DLBCL lymphomas	77	6286	1. DLBCL -58 objects 2. FL – 19 objects
Breast cancer	130	11217	1. Tumor – 67 objects 2. Normal – 63 objects
Lung cancer	203	3312	1 AD – 139 objects 2 SQ – 21 objects 3 COID – 20 objects 4 SMCL – 6 objects 5 NL – 17 objects

The DLBCL dataset contains samples of two subtypes of lymphoma: B-cell and follicular lymphoma. These samples are characterized by the expression of 7129 genes. After sliding window preprocessing, the measurements were bounded from above by 1600, the lower threshold was set at 20. The genes for

which the ratio of the maximum to minimum expression value was less than three or the absolute value of variation was less than 100 were excluded from consideration. Samples were normalized with mean and standard deviation.

Breast dataset contains samples of normal and cancerous tissue. Each sample is characterised by the expression of 11 217 genes.

Lung dataset contains 203 samples with five subtypes of lung cancer characterised by 12 600 genes. Of these, 3312 genes were selected with the standard deviation more than 50.

To build a hybrid classification model for each dataset, we used a double cross-validation procedure. Internal cross-validation was used to select the most efficient classifier based on an aggregate efficiency criterion (accuracy, sensitivity, specificity), external cross-validation was used to evaluate the performance of an ensemble of classifiers.

Hybrid classification model was built for each dataset and its performance was compared with individual classifiers. For the DLBCL dataset, a list of features ranked by information content was loaded from a file, and the number of features 100, 250, 500 was sequentially selected to build a classifier. The models SVM, PLS+LDA, PLS+QDA, PLS+RF, PCA+LDA, PCA+QDA, PCA+RF, PLR were selected as the base classifiers.

TABLE III

EFFICIENCY CRITERIA VALUES FOR THE DLBCL DATASET

Number of features	Classifier	Accuracy*	Sensitivity*	Specificity*	AUC*	Number**
100	svm	0.8482	0.9333	0.6	0.95	156
	pls_lda	0.9	0.9167	0.85	0.9417	147
	pls_qda	0.8857	0.9167	0.8	0.9667	103
	pls_rf	0.8857	0.9333	0.75	0.9667	78
	pca_lda	0.6375	0.7967	0.15	0.5458	4
	pca_qda	0.65	0.83	0.1	0.5233	0
	pca_rf	0.7232	0.8933	0.2	0.5983	3
	plr	0.7429	0.6767	0.95	0.9417	19
	Hybrid model	0.8875	0.95	0.8	0.9667	–
250	svm	0.9071	0.95	0.75	0.9617	135
	pls_lda	0.9375	0.95	0.9	0.95	168
	pls_qda	0.9375	0.95	0.9	0.9458	99
	pls_rf	0.8982	0.95	0.7	0.925	81
	pca_lda	0.6232	0.7933	0.1	0.4425	1
	pca_qda	0.6321	0.8267	0.05	0.4092	0
	pca_rf	0.6875	0.86	0.15	0.4525	0
	plr	0.8054	0.7433	1	0.95	26
	Hybrid model	0.9357	0.9667	0.85	0.9667	–
500	svm	0.9239	0.9833	0.75	0.9267	144
	pls_lda	0.9339	0.95	0.9	0.9667	181
	pls_qda	0.9382	0.9667	0.9	0.9417	66
	pls_rf	0.9214	0.9667	0.8	0.9692	96
	pca_lda	0.5304	0.67	0.1	0.2842	1
	pca_qda	0.6232	0.77	0.15	0.4325	1
	pca_rf	0.6482	0.82	0.1	0.2675	0
	plr	0.8179	0.7767	0.95	0.9717	21
	Hybrid model	0.9357	0.9833	0.9	0.9833	–
All features	svm	0.944	0.9833	0.8	0.9517	80
	pls_lda	0.975	0.9833	0.95	0.9708	220
	pls_qda	0.8875	0.95	0.7	0.8917	30
	pls_rf	0.975	0.9833	0.95	0.9917	120
	pca_lda	0.7607	0.8267	0.55	0.7433	5
	pca_qda	0.6881	0.9133	0	0.6033	0
	pca_rf	0.7089	0.8067	0.4	0.7142	0
	plr	0.95	0.9333	1	0.9917	95
	Hybrid model	0.9835	0.9833	0.95	1	–

* Mean value of 10-fold cross-validation. The number of samples of the bagging procedure is $N = 51$.
** The distribution of individual classifiers included in the ensembles.

The number of groups for cross-validation was chosen to be 10, the number of subsamples to select the base classifiers was chosen to be 50. Classification efficiency criteria depending on the number of features are presented in Table III.

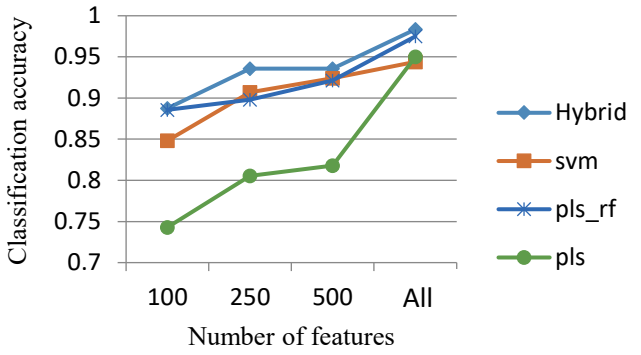

Fig. 3. Dependence of classification accuracy on the number of features for DLBCL dataset.

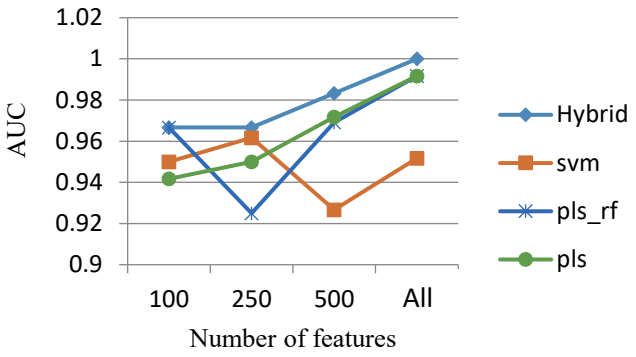

Fig. 4. Dependence of AUC on the number of features for DLBCL dataset.

TABLE IV
EFFICIENCY CRITERIA VALUES FOR THE BREAST DATASET

Number of features	Classifier	Accuracy*	Sensitivity*	Specificity*	AUC*	Number**
50	svm	0.5097	0.3333	0.6714	0.5365	95
	pls_lda	0.5644	0.5476	0.5833	0.6284	140
	pls_qda	0.5868	0.531	0.6452	0.6463	50
	pls_rf	0.5874	0.5643	0.6095	0.6183	100
	pls	0.5951	0.5762	0.6095	0.628	165
	Hybrid model	**0.5726**	**0.5619**	**0.5833**	**0.592**	–
100	svm	0.5053	0.2643	0.7238	0.4848	85
	pls_lda	0.5378	0.4952	0.5762	0.6246	120
	pls_qda	0.5296	0.4929	0.5595	0.6086	90
	pls_rf	0.5581	0.5381	0.5738	0.6103	110
	pls	0.593	0.5381	0.6429	0.6356	145
	Hybrid model	**0.6162**	**0.5381**	**0.6495**	**0.6483**	–
500	svm	0.4797	0.3643	0.5905	0.5029	120
	pls_lda	0.5695	0.5881	0.55	0.6082	115
	pls_qda	0.5855	0.5738	0.5952	0.619	85
	pls_rf	0.6154	0.5881	0.6405	0.5969	100
	pls	0.5951	0.6881	0.5095	0.6083	130
	Hybrid model	**0.626**	**0.6571**	**0.5943**	**0.6529**	–
All features	svm	0.539	0.4071	0.681	0.567	110
	pls_lda	0.5627	0.5881	0.5405	0.6311	115
	pls_qda	0.5627	0.5548	0.569	0.5954	65
	pls_rf	0.5378	0.5048	0.569	0.5882	100
	pls	0.5918	0.681	0.5119	0.6325	160
	Hybrid model	**0.6156**	**0.6414**	**0.6167**	**0.6443**	–

* Mean value of 10-fold cross-validation. The number of samples of the bagging procedure is $N = 51$.
** The distribution of individual classifiers included in the ensembles.

According to Table III, all classifiers for the DLBCL dataset, with the exception of classifiers with a preliminary reduction of the feature space using PCA, have efficiency criteria close to one, which is due to the good separability between the two classes. Figs. 3–4 show classification accuracy and AUC versus the number of features for several classification models.

According to Figs. 3–4, in most cases the hybrid classification model has higher classification accuracy and AUC regardless of the number of features. For all models (with the exception of SVM), the values of classification accuracy and AUC increase with the number of features. Moreover, according to experiments, increasing the number of features does not lead to significant changes in the classification efficiency.

Using the cross-validation procedure and the hybrid classification model, only three objects were classified incorrectly, two FL objects were classified as DLBCL and one DLBCL object was classified as FL.

For the Breast dataset, a list of features ranked by information content was loaded from a file, and the number of features 50, 100, 500 was sequentially selected to build a hybrid classifier. Models SVM, PLS+LDA, PLS+QDA, PLS+RF, PLS were selected as base classifiers. The number of groups for cross-validation was chosen to be 10, the number of subsamples of data objects to select the base classifiers in the hybrid model was chosen to be 50. Classification efficiency criteria depending on the number of features are presented in Table IV.

According to Table IV, no individual classifier outperforms the others in all three efficiency criteria. The PLS model has higher accuracy and sensitivity values, the SVM model has better specificity values for most feature sets. The hybrid model outperforms individual classifiers in terms of accuracy and sensitivity, and also has the maximum value of the AUC criterion.

From Table IV, it can be seen that for all models (with the exception of SVM), the classification accuracy and AUC increase with the number of features. Unlike other classifiers, the hybrid model has equally high values for all efficiency criteria.

TABLE V
EFFICIENCY CRITERIA VALUES FOR THE LUNG DATASET

Classifier	Accuracy*	Number**
svm	0.9367	117
pls_lda	0.9256	346
pls_rf	0.9213	47
pca_lda	0.6363	0
Hybrid model	0.9411	

* Mean value of 10-fold cross-validation. The number of samples of the bagging procedure is $N = 51$.
** The distribution of individual classifiers included in the ensembles.

Classification efficiency criteria for the Lung dataset with all the features are presented in Table V. As the number of subtypes for the Lung dataset equals five, only overall classification accuracy was considered as a performance criterion. According to Table V, SVM classification model has the highest accuracy among the individual classifiers. PCA+LDA is the least efficient classifier. The hybrid model has the best classification accuracy.

V. CONCLUSION

The proposed method for constructing a hybrid classification model allows considering the variety of data sources of biological information and their limited sample size in order to build a classifier to identify subtypes of complex diseases. The hybrid model is a classification ensemble where base classifiers built on varied sources of biomedical data are considered as its basic elements.

The advantage of the proposed method is the simultaneous use of several performance criteria, such as accuracy, sensitivity, and specificity for selecting the optimal base classifiers for the considered data sources. It will increase both the stability of model selection under conditions of a limited training sample and the generalizing ability of the hybrid classification model. The proposed method using several a priori specified classification models and criteria for evaluating the effectiveness of classification allows consistently determining the structure of both the base classifiers of the ensemble and the entire hybrid model.

The distinctive features of the method are its adaptive nature, i.e. the ability to build efficient classifiers regardless of data types, as well as a multi-criteria approach to evaluate the classification efficiency using weighted aggregation of ranked lists.

The testing results on the artificial and real biomedical data show the advantages of the proposed hybrid classification model over other types of classifiers.

REFERENCES

[1] M. J. Zaki and W. Meira, *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press, 2014. https://doi.org/10.1017/CBO9780511810114
[2] A. Statnikov *et al*. "A comprehensive evaluation of multicategory classification methods for microbiomic data," *Microbiome*, vol. 1, no. 1, Art. no. 11, Apr. 2013. https://doi.org/10.1186/2049-2618-1-11
[3] S. Dudoit, J. Fridlyand, and T. P. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," *Journal of the American Statistical Association*, vol. 97, no. 457, pp. 77–87, Dec. 2002. https://doi.org/10.1198/016214502753479248
[4] J. R. Quinlan, *C4.5: Programs for Machine Learning*. Elsevier, 2014.
[5] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*, Wiley, 2014. https://doi.org/10.1002/9781118914564
[6] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001. https://doi.org/10.1023/A:1010933404324
[7] G. Valentini and F. Masulli, "Ensembles of learning machines," in *Lecture Notes in Computer Science*, vol. 2486, Neural Nets WIRN Vietri-2002, R. Tagliaferri, M. Marinaro, Eds. Springer, Berlin, Jan. 2002, pp. 3–19. https://doi.org/10.1007/3-540-45808-5_1

[8] R. E. Schapire and Y. Freund, *Boosting: Foundations and Algorithms*. MIT Press, 2012. https://doi.org/10.7551/mitpress/8291.001.0001
[9] O. Okun and H. Priisalu, "Dataset complexity in gene expression based cancer classification using ensembles of k-nearest neighbors", *Artificial intelligence in medicine*, vol. 45, no. 2–3, pp. 151–162, Feb.–Mar. 2009. https://doi.org/10.1016/j.artmed.2008.08.004
[10] T. Hastie, "Multi-class adaboost", *Statistics and its Interface*, vol. 2, no. 3, pp. 349–360, Jan. 2009. https://doi.org/10.4310/SII.2009.v2.n3.a8
[11] Y. Wang *et al*. "Gene selection from microarray data for cancer classification – a machine learning approach", *Computational biology and chemistry*, vol. 29, no. 1, pp. 37–46, Feb. 2005. https://doi.org/10.1016/j.compbiolchem.2004.11.001
[12] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering", *IEEE Trans. on Knowledge and Data Engineering*, vol. 17, no. 4, pp. 491–502, Apr. 2005. https://doi.org/10.1109/TKDE.2005.66
[13] The Cancer Genome Atlas. [Online]. Available: http://cancergenome.nih.gov/abouttcga. Accessed on: Oct. 6, 2022.

Natalia Novoselova received the PhD degree in Computer Science from the United Institute of Informatics Problems (UIIP), National Academy of Sciences of Belarus (NASB) in 2008. Since 2000, she has been a Senior Scientific Researcher at the Laboratory of Bioinformatics, United Institute of Informatics Problems, National Academy of Sciences of Belarus (NASB) in Minsk, Belarus. Her research interests include data mining methods, notably the neural network, genetic algorithms and fuzzy systems and their application to analysis of medical and biological data. She is the author of more than 20 research publications in these areas.
Contact information: Department of Bioinformatics, United Institute of Informatics Problems, Surganova str. 6, Minsk, 220012, Belarus. Phone: +375-17-2842092.
E-mail: novos65@gmail.com

Igor Tom received the PhD degree in Computer Science from the Institute of Engineering Cybernetics, National Academy of Sciences of Belarus (NASB) in 1986. Since 1999, he has been the Chief of the Department Bioinformatics at the United Institute of Informatics Problems of NASB. His current research interests are in the fields of the development of intelligent methods of data analysis and information technologies for medical and industrial applications. He is the author and co-author of more than 170 scientific publications, including 50 full papers in journals.
E-mail: ietom143@gmail.com