Supervised Clustering of Genes for Multi-Class Phenotype Classification

Novoselova N., Tom I.
United Institute of Informatics Problems NAS Belarus
{novosel, tom}@newman.bas-net.by, http://uiip.bas-net.by

Abstract: *The paper presents the new approach to the supervised gene selection by means of gene clustering for the microarray data, which belong to more than two phenotypes (classes). The main distinction from the previous approaches, that are based on the splitting the multi-class task into several binary ones, is the application of the HUM (hypervolume under the manifold) score, that guides the search for the most discriminative gene clusters that simultaneously differentiate all the classes. The results of comparative analysis with other methods shows the advantages of our approach both in classification rate of the new samples and in the lower number of gene clusters. The application of our approach to a randomly permuted data shows that the identified structure is more than just a noise artifact.*

Keywords**:** Microarray technology; gene expression data; clustering.

1. INTRODUCTION

Microarray technology is widely used to identify the molecular characteristics and distinctions of several tissue samples in order to reveal new subtypes of disease, e.g. cancer or to predict the phenotype of sample tissue on the basis of several gene expression values. It's well known that the microarray datasets contain a small number of samples and a large number of genes, and only small portion of genes are related to the discrimination of the sample classes. Therefore the important task of the analysis of microarray datasets is to reveal the most important genes, the so-called biomarkers, which are closely connected to the distinction of phenotypes of disease. There are a plenty of proposed methods for the informative features selection for classification [1], including the gene selection methods [2-5]. These methods are mainly concentrated on the selection of individual genes [3, 4] or the genes subsets [2, 5], that are considered individually. In microarray data analysis, there can be a large number of highly discriminative subsets containing only a couple of genes, and the composition of such subsets can greatly vary as a result of a different choice of the subset of samples or the noise in the microarray data. Therefore it is more reliable and robust to direct the search of discriminative gene clusters and to base the prediction of a new sample on the basis of the collective behavior and coordinated expression of a group of genes, rather than that of the individual genes. Therefore we propose the approach, that is based on supervised clustering scheme, developed in [6] and allows constructing gene clusters, taking into account the sample class labels of the training set. The major differences between our approach and the one, proposed in [6], is the possibility to provide the search for gene clusters for multi-class problem in one run, avoiding its separation into several binary classification tasks. Such an approach allows getting more compact gene clusters, which are able to simultaneously discriminate more then two classes. To estimate the importance of the individual gene or gene cluster for classification we apply the HUM (hypervolume under the manifold) score or the extension of ROC analysis on multiple classes, described in [7]. Using the proposed clustering scheme allows to identify most discriminative gene clusters, where the average expression of genes in the clusters constitute the predictors and are further used to define the phenotype of the unknown tissue sample. The output of our approach is the number and the composition of the gene clusters, which consists of only small number of genes. As in [6] the greedy strategy is used for the searching the best composition of each following cluster, according to the objective function, which measures the cluster's ability for phenotype discrimination and relies on the HUM score calculation. It achieved similar or better prediction accuracy than the methods in [10] and the approach in [6] in our validation process.

We have also investigated the ability of the HUM score to reveal the really discriminative gene clusters instead just some structure by chance.

The output of our algorithm is thus valuable for cancer-type diagnosis. At the same time it is very accessible for interpretation, as the output consists of a very limited number of clusters, each summarizing the information about a few genes. Thus, it may also reveal insights into biological processes and give hints on explaining how the genome works.

We first describe the proposed approach for supervised clustering of gene-expression data, then apply the procedure to publicly available microarray dataset and test the results of its predictive potential.

2. DESCRIPTION OF THE APPROACH

The input data is the data matrix $X = (x_{ij})$ with dimension $n \times m$, where n – the number of experiments, m – the number of genes. Each i th individual experiment or the sample tissue belongs to class y_i and the data for the experiment constitutes the sample expression profile $x_i = (x_{i1}, \ldots, x_{im})$. For K cancer phenotypes the class label y_i is the integer value in interval $[1, K]$ and n_k is the number of samples of k th class. The aim of microarray data analysis is to reveal the discriminative genes and to construct the classifier for K disease subtypes, which splits the set X of gene expression profiles into K disjoint subset (in the case of crisp classification) $A_1, \ldots, A_k$ such that for each sample with gene expression profile $x_i = \left(x_{i1},\dots,x_{im}\right) \in A_k$ the predicted class is k. The classificatory is constructed on the basis of "past" experience, i.e. on the basis of a training set with "a priori" known class labels $L = \left\{\left(x_1, y_1\right),\dots,\left(x_{n_L}, y_{n_L}\right)\right\}$. After construction of the classifier on L the class labels y_i, $i = \overline{1, n_T}$ of test set $T = \left\{x_1,\dots,x_{n_T}\right\}$ can be predicted. When the class labels of the test set are known beforehand, they can be compared with the predicted ones in order to estimate the classification error.

According to the clustering algorithm the classifier, constructed on the discriminative subset of genes $q << m$ is modeled by the conditional probability:

$$P(Y = k \mid X) = f(X_{C_1}, X_{C_2},\dots, X_{C_q}), \qquad (1)$$

where f – nonlinear function from R^q to $[0,1]$, $\{C_1, C_2,\dots, C_q\}$ – functional groups of clusters of genes with $\{\bigcup_{i=1}^{q} C_i\} \subset \{1,\dots,m\}$ and $C_i \bigcap C_j = \varnothing, i \neq j$, $X_{C_i} \in R$ – an average expression value of gene cluster C_i.

As the construction of the model (1) is non-trivial task therefore the authors in [6] have adopted the greedy procedure, which we also follow. The procedure relies on growing the cluster incrementally by adding one gene after the other. During clustering the forward search is followed by the backward search in order to remove non-important genes, that where wrongly added to the cluster at the previous steps. The decision of the inclusion of a new gene in the cluster is made under supervision of class labels and is based on the cluster ability to discriminate classes. As opposed to estimation score in [6] we proposed to use a HUM score [7], that allows estimating the ability of gene cluster to discriminate between more than two classes. HUM is a direct extension of the area under the ROC curve (AUC) and was employed in many foregoing works [8-9]. If the expression values of a particular gene or cluster yield exact separation of the classes, then the expression values for all tissue samples are strongly ordered according to the class label, i.e. if $X_{n_1}, X_{n_2},\dots, X_{n_K}$ – the sets of i th gene values, corresponding to classes $\overline{1, K}$, then there exist such a permutation of class labels $\pi(k), k = \overline{1, K}$, that $X_{\pi(1)} < X_{\pi(2)} < \dots < X_{\pi(K)}$. In such a case the HUM score is maximal and is equal to $HUM = 1$. If the gene expression values are independent of class labels, then the value of HUM score equals $(K!)^{-1}$.

The supervised clustering procedure thereafter consists in following steps:

1. Start with the entire matrix $X = \left(x_{ij}\right)$ with dimension $n \times m$. The values of each gene are normalized to zero mean and unit variance.

2. Determine the HUM score of every gene i, that is, every n-dimensional vector of observed expression values $x_i = \left(x_{i1},\dots,x_{in}\right)$.

3. Define the first gene in cluster C as gene with maximal HUM score.

4. Perform the forward search of gene cluster C.

Average the gene cluster expression profile $x_C = (x_C^1, x_C^2,\dots, x_C^n)$ with each gene profile $x_i = \left(x_{i1},\dots,x_{in}\right)$ $x_{C,i} = \dfrac{1}{|C|+1}(x_i + \sum_{j \in C} x_j), i = 1,\dots,m$.

As candidate to the inclusion in cluster select the gene i^* for which $HUM(x_{C,i^*}) = \max_i(HUM(x_{C,i}))$.

5. Repeat step 4 until the $HUM(x_{C,i^*})$ score for the selected gene i^* is not lower than the previous $HUM(x_C)$ score.

6. Perform the backward search, consequently selecting the possible genes for exclusion. For this calculate the set of HUM scores of a cluster without gene i, define $HUM(x_{C,without\ i^*}) = \max_i(HUM(x_{C,without\ i}))$. After that if

$$HUM(x_{C,without\ i^*}) >= HUM(x_{Ci}), \qquad (2)$$

exclude the uninformative gene i^* from the cluster C.

7. Repeat step 6 until the inequality (2) is no longer valid.

8. Repeat steps 4-7 until the stabilization of the cluster composition and optimization of the HUM score.

9. If more than one cluster C is desired, discard the genes in the former clusters from X and restart the algorithm at step 3 with the reduced matrix.

During the algorithm execution the gene cluster construction is performed together with gene selection for classification. The new samples can be further classified on the basis of the constructed predictors – the average gene cluster values.

3. HUM SCORE RELEVANCE FOR GENE CLUSTER FORMATION

We have tested the significance of distinction between the HUM score of the analyzed data and the ones, received for the unstructured random-noise gene-expression data. If the distinction is significant, than the hypothesis that the clusters found on the original data by the supervised algorithm are just a noise artifact can be rejected.

To perform the test we have simulated L random-noise gene-expression datasets by permuting the set of original class labels $(y_1^{*(l)},\dots, y_n^{*(l)}), l = \overline{1, L}$. After that each of the original gene expression profiles was allocated to the permuted class label, resulting in independent pairs $(x_1, y_1^{*(l)}), (x_2, y_2^{*(l)}),\dots,(x_n, y_n^{*(l)})$ for each $l = 1,\dots, L$. The supervised clustering procedure is then applied $L = 100$ times on such data with randomly permuted responses. For every permuted set of responses, a single cluster (q = 1) was formed on the entire dataset and both its final HUM score were recorded. Empirical distribution of the HUM scores from permuted data together with the HUM score for the original data set are depicted in Fig. 1. According to Fig. 1 we can reject the hypothesis with the p-value of zero.

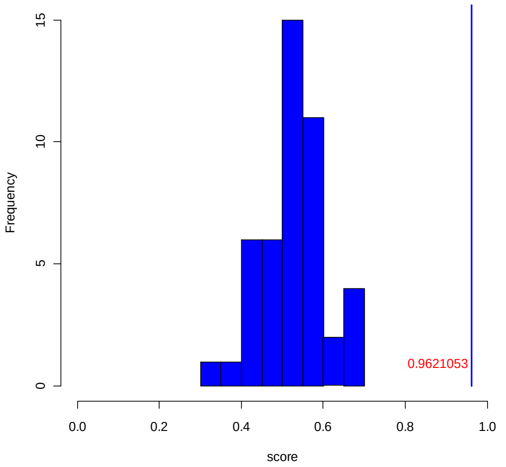

Fig. 1 – Histograms showing the empirical distribution of HUM scores for the leukemia dataset (AML/ALL distinction), based on 100 bootstrap replicates with permuted response variables. The vertical line marks the values of score with the original response variables

4. DATASET AND EXPERIMENTAL RESULTS

We have tested the proposed approach on the real Leukemia dataset. As a rule the dataset is considered consisting of two classes – 47 samples of acute lymphoblastic leukemia (ALL) and 25 samples of acute myeloid leukemia (AML) [10]. In order to validate our approach for multi-class task we have taken into account the two subtypes of ALL: 38 samples of B-cell ALL and 9 samples of T-cell T-ALL, analyzing the classification into 3 classes. All the samples are characterized by the expression of 7129 genes. After data preprocessing with thresholding and filtering the 3571 genes are selected for further analysis.

To estimate the prediction potential of the approach we have applied the following scheme. The data was randomly split into a learning set comprising two thirds, and a test set containing the remaining third of all n data samples. The learning set was used to perform the gene clustering according to the proposed approach and to construct classifier using two methods: nearest-neighbor classifier (NNC) and diagonal linear discriminant analysis (DLDA). After that the prediction of class labels for the test sets was performed using two classifiers and the number of clusters $q = \overline{1,10}$. The misclassification rate is then calculated as the averaged fraction of predicted class labels which differ from the true one. The whole procedure was repeated 100 times and the results are depicted in Table 1. The box plots and median-quartile plots of the misclassification errors for each individual value of q are presented in Fig. 2.

Table 1 – Classification error according to the random splitting study for two classifiers (our approach)

	Value	q= 1	q= 2	q= 3	q= 4	q= 5	q= 6	q= 7	q= 8	q= 9	q= 10
DLDA	Median	0,125	0,083	**0,042**	0,042	0,042	0,042	0,042	0,042	0,042	0,042
	1 quartile	0,125	0,052	0,042	0,042	0,042	0,042	0,042	0,000	0,000	0,000
	3 quartile	0,167	0,125	0,115	0,115	0,073	0,073	0,083	0,042	0,042	0,042
NNC	Median	0,1667	0,0833	**0,0625**	0,0625	0,0833	0,0417	0,0417	0,0417	0,0417	0,0417
	1 quartile	0,1250	0,0833	0,0417	0,0417	0,0417	0,0417	0,0417	0,0417	0,0417	0,0313
	3 quartile	0,2500	0,1250	0,1250	0,0833	0,0833	0,0833	0,0833	0,0833	0,0833	0,0833

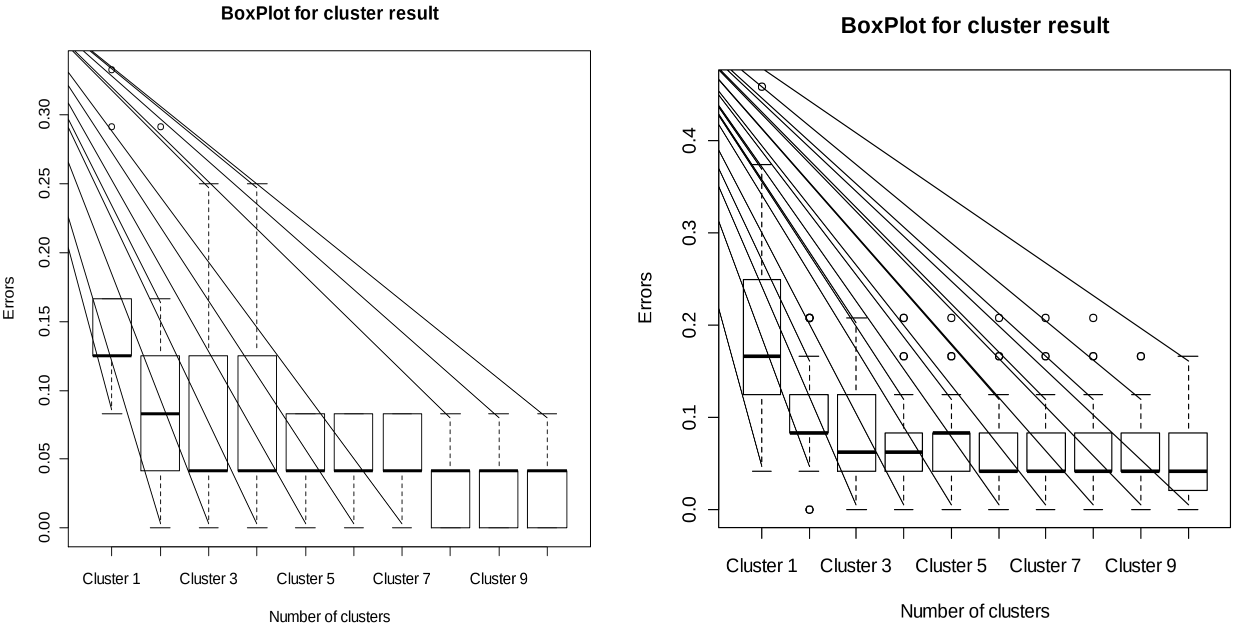

Fig. 2 – Box plots of the misclassification errors for the leukemia data set (3 classes), based on the 100 random divisions into the learning and test sets: a) DLDA; b) NNC

According to Fig. 2 to construct the best classifier (in accuracy) it’s sufficient to select 3 clusters of genes as an input for DLDA and for NNC.

We have performed the supervised clustering for all the set of 72 samples. The Fig.3 depicted the projection of the samples on two first revealed cluster coordinated. It can be seen that the first two coordinate perfectly separate two classes of leukemia.

We have compared the experimental results with the approach in [6], where the multi-class task is considered as the set of binary tasks, the corresponding classification accuracy of the test sets are depicted in Table 2 and in Fig. 4.

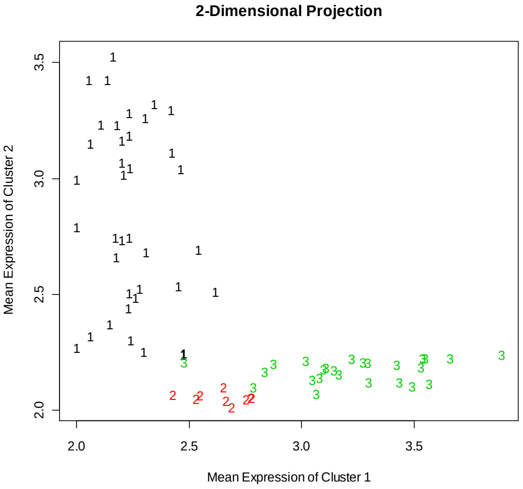

Fig. 3 – The leukaemia dataset in the space of average expression values of two gene clusters

Table 2 – Classification error according to the random splitting study for two classifiers (approach [6])

	Value	q= 1	q= 2	q= 3	q= 4	q= 5	q= 6	q= 7	q= 8	q= 9	q= 10
DLDA	Median	0,167	0,083	0,083	**0,042**	0,042	0,042	0,042	0,042	0,042	0,042
	1 quartile	0,125	0,042	0,042	0,042	0,042	0,042	0,042	0,042	0,042	0,042
	3 quartile	0,250	0,125	0,125	0,083	0,083	0,083	0,083	0,083	0,083	0,083
NNC	Median	0,167	0,083	0,083	0,063	**0,042**	0,042	0,042	0,042	0,042	0,042
	1 quartile	0,083	0,042	0,042	0,042	0,042	0,042	0,042	0,042	0,042	0,042
	3 quartile	0,250	0,125	0,083	0,083	0,083	0,083	0,083	0,083	0,083	0,083

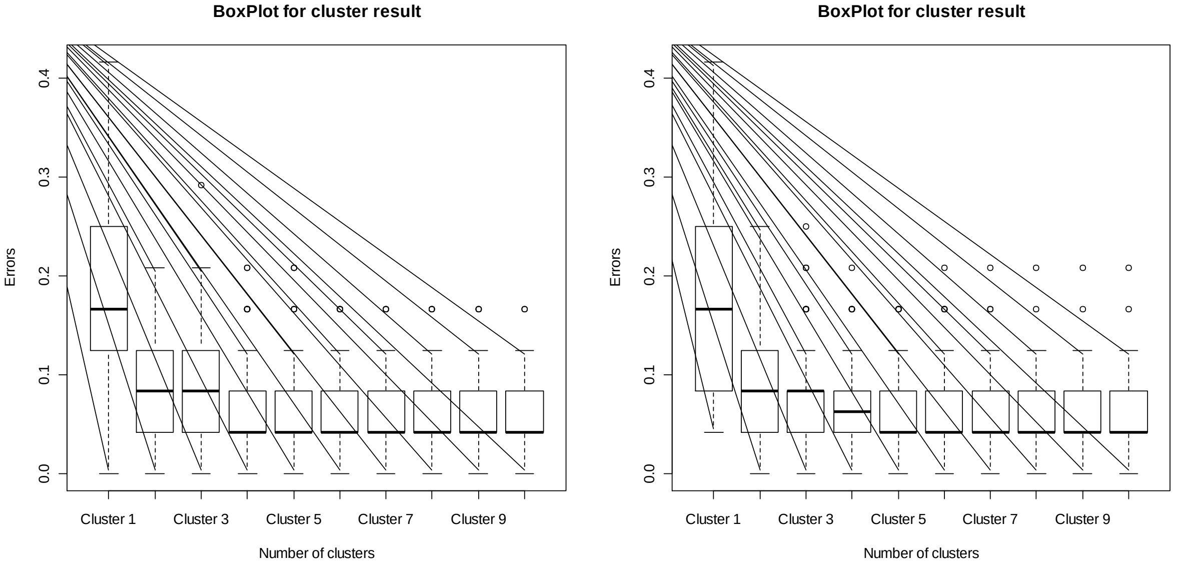

Fig. 4 – Box plots of the misclassification errors for the leukemia data set (3 classes), based on the 100 random divisions into the learning and test sets: a) DLDA; b) NNC

According to the Fig. 4 to construct the best classifier (in accuracy) it’s sufficient to select 4 clusters of genes as an input for DLDA and 5 clusters for NNC. It is noticeably more than in our approach.

In order to compare the results with the approach [6] we have made all the process of supervised clustering and classifier construction on the standard training set of 38 samples and have tested the result on 34 left independent samples. According to the results of the random splitting study we have used the best predictor set and classifier for the independent subset (approach [6] – DLDA with $q = 4$; our approach – DLDA with $q = 3$).

For approach [6] we have received 1 error in the test set (T-ALL sample was wrongly attributed to the AML subtype), classification results according to our approach are error-free. Moreover the number of the gene clusters, received according to [6] – 9 clusters (Table 3) – is higher than in our approach – 3 clusters (Table 4).

Table 3 – The composition of the gene clusters according to the approach [6]

Class 1 (B-cell) versus 2&3	Cluster 1	X82240_rna1_at (TCL1 gene (T cell leukemia))
	Cluster 2	L33930_s_at, M89957_at, K01911_at
	Cluster 3	U05259_rna1_at, U50743_at, D88422_at, M96326_rna1_at, M27891_at (cystatin), M74719_at
	Cluster 4	L08895_at, J03077_s_at, M11722_at, M21005_at, L20971_at, Z22548_at, M57731_s_at
Class 2 (T-cell) versus 1&3	Cluster 1	X59871_at, X03934_at, L08895_at
	Cluster 2	M38690_at, M74719_at, U23852_s_at
	Cluster 3	M28826_at, X00437_s_at, HG987-HT987_at, J04164_at, L19686_rna1_at
	Cluster 4	X82240_rna1_at (TCL1 gene (T cell leukemia))
Class 3 (AML) versus 1&2	Cluster 1	X95735_at (zyxin), M27891_at (cystatin), J04615_at
	Cluster 2	M16336_s_at (CD2 CD2 antigen), M89957_at, U23852_s_at, M27783_s_at
	Cluster 3	L33930_s_at
	Cluster 4	X82240_rna1_at, D87433_at, X77737_at, M57731_s_at, M57466_s_at, M84526_at, U57341_at, U10485_at, X59871_at, U05259_rna1_at

Table 4 – The composition of the gene clusters according to our approach

Class 1 (B-cell) versus Class 2 (T-cell) versus Class 3 (AML)	Cluster 1	M27891_at, X59871_at, M28826_at, J03077_s_at
	Cluster 2	M74719_at, M92934_at, X82240_rna1_at, K01911_at
	Cluster 3	X95735_at, X76223_s_at, X00437_s_at, Z84721_cds2_at

We have received better misclassification rate than in [10], where the authors compared the efficiency of different discrimination methods for several datasets including 3-class leukemia dataset, but unlike our approach they have considered 40 preliminary selected genes.

5. CONCLUSION

We have proposed an approach to supervised construction of the gene clusters from the microarray experiments. The main difference of the proposed approach from the previous similar one [6] consists in the utilizing the HUM measure to estimate the discriminative power of each individual gene or gene cluster for the case, when dataset belongs to more than two classes. Our procedure is potentially useful in the context of medical diagnostics, as it identifies groups of interacting genes that have high explanatory power for given tissue types, and which in turn can be used to accurately predict the class labels for the new samples. The result of comparison with previously proposed methods [6, 10] shows the advantages of our approach in number of clusters and accuracy of the classifier, constructed on the basis of gene cluster values. Moreover each gene cluster is able to perfectly discriminate between more than two classes simultaneously, that is not possible in [6], where each multi-class task splits into several binary tasks and the constructed thereafter gene clusters are responsible to differentiate each individual class from all the other. Such an approach complicates the clustering process and increases the number of clusters that are sufficient for accurate classification of the new samples. We have also shown that an application of our algorithm to a randomly permuted data shows that the identified structure is more than just a noise artifact.

In our further study we aim to moderate the time-consuming process of HUM score calculation.

6. REFERENCES

[1] Kohavi, R. Wrapper for feature subset selection / R. Kohavi, G. John // Artificial Intelligence. – 1997. – Vol. 97, №1-2. – P. 273–324.

[2] Ding, C. Minimum Redundancy feature selection from microarray gene expression data / C. Ding, H. Peng // Journal of Bioinformatics and Computational Biology. – 2005. – Vol. 3, №2. – P. 185–205.

[3] An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles / J.G. Thomas et al. // Genome Res. – 2001. – Vol. 11. – P. 1227–1236.

[4] Filter versus wrapper gene selection approaches in DNA microarray domains / I. Inza et al. // Artif. Intell. Med. – 2004. – Vol. 31, № 2. – P. 91–103.

[5] Liu, X. An entropy-based gene selection method for cancer classification using microarray data / X. Liu, A. Krishnan, A. Mondry // BMC Bioinformatics. – 2005. – Vol. 6, № 76.

[6] Dettling M., Buhlmann P. (2002) Supervised clustering of genes. Genome Biol 3: research0069.1–0069.15.

[7] Li J., Fine J.P. ROC analysis with multiple classes and multiple tests: methodology and its application in microarray studies. Biostatistics 2008; 9; 566–576.

[8] Y.-J. Wu, C.-T. Chiang. ROC Representation for the Discriminability of Multi-Classifcation Markers . Department of Mathematics, National Taiwan University, PREPRINT, 2011.

[9] Ogdie A., Li J., Dai L., Paessler M.E., Yu X., Diaz-Torne C., Akmatov M., Schumacher H.R., Pessler F. Identification of broadly discriminatory tissue biomarkers of synovitis with binary and multicategory receiver operating characteristic analysis. Biomarkers. 2010 Mar;15(2):183–90.

[10] Dudoit S., Fridlyand J., Speed T. Comparison of discrimination methods for the classification of tumors using gene expression data. J. Am. Stat. Assoc. 2002, 97:77–87.