VERSITA
Information Technology and Management Science
doi: 10.2478/itms-2013-0015
2013 / 16

Feature Ranking by Classification Accuracy Estimation of Multiple Data Samples

Natalia Novoselova[1], Igor Tom[2], [1-2] *United Institute of Informatics Problems of the National Academy of Sciences of Belarus,* Arkady Borisov[3], Inese Polaka[4], [3-4] *Riga Technical University*

***Abstract* – This article considers the gene ranking algorithm for the microarray data. The rank vector is estimated by classifications of the random data samples. At each iteration, the ranks of genes participating in the successful classification become higher. Unlike other methods of feature selection, the proposed algorithm allows increasing the generality of the classification models by construction of the balanced training samples and taking into account the descriptiveness of the gene combinations by the subset estimation.**

***Keywords* – Biomarker, classification, feature ranking, gene expression**

I. Introduction

The improvement of the cancer diagnostics is one of the topmost aims in medicine. The defined diagnosis is the basis for the selection of the treatment protocol in order to simultaneously maximize the treatment efficiency and minimize the toxicity. Recently, the great attention is paid to the investigation of the potential of the cancer cell genetic information for adjustment of existing cancer subtypes and discovery of the new cancer subtypes at the molecular level. The results will aid in more accurate discrimination of cancer and will assist in more effective therapy.

Microarray technology presents the new era of biotechnologies and allows measuring the expression of thousands of genes simultaneously. Microarray data give the rich material for the deep understanding of differences between tumour subtypes at the molecular level and thereafter their more reliable classification. Microarray data are used both to determine up-regulated and down-regulated genes relative to the cancer subtypes under investigation and to identify gene signatures or biomarkers for the clinical diagnostics and prognosis [1-2].

Unfortunately, the information extraction from such kind of data is hampered by the strong imbalance between the number of samples (usually less than hundred) and the number of genes (usually tens of thousands). This discrepancy as a rule leads to overfitting, namely, the construction of the classification model with poor generality (low accuracy on independent set of data) [3]. Moreover, the microarray data enclose as a rule the measurement and systematic errors, which can considerably influence the classification accuracy. Therefore, the classification process as well as clustering must be preceded by the feature space preprocessing, which in our context means the process of feature (gene) selection by their ranking or feature reduction [4].

The methods of feature selection for the tasks of cancer diagnostics as well as for all classification tasks are divided into two major groups: filter and wrapper methods [5]. Filter methods allow for the deletion of the uninformative features according to general data characteristics. They are computationally simpler and are not connected to a particular classification algorithm. The wrapper methods imply the use of the classification algorithm in the process of feature subset selection, which simultaneously improves the classification accuracy. The limitation of the wrapper methods is the high computational complexity, caused by the repeated application of classification algorithm during the selection process. For the informative gene subset selection, both the filter methods, e.g., standard parametric tests (t-test [6]) or nonparametric tests (Wilcoxon signed-rank test [7]) and the wrapper methods [8-9] are used in the literature.

However, very often the subsets of the selected genes are greatly differentiated by different researchers that can be the result of both the discrepancy of analysed data dimensions and the ignoring of gene dependencies. To date, there are not unique recommendations on the methods of informative gene selection that are more adequate for a particular dataset [10]. The algorithm of gene selection proposed in our research allows overcoming the shortages of both the filter and wrapper approaches and is based on gene ranking by classification of repeated random samples from the initial dataset. The algorithm possesses the following advantages: it avoids the overfitting through forming the low-dimensional training matrices with predominance of sample number to the number of genes; it takes into account the informativity of gene combinations through feature subset estimation by a classification algorithm.

The paper presents the results of testing the algorithm on the leukaemia dataset, verification of the biological significance of the twenty top-ranked genes and a comparative analysis with similar studies of other authors.

II. Algorithm Description

Let $X_{M\times N}$ be the initial matrix of gene expression, where M is the number of genes, N is the number of samples or data objects. $x_i = (x_i^1, x_i^2, \ldots, x_i^N), i = 1, M$ is a vector of gene values for N samples or gene profile. Each data object is marked by the class label and in our research belongs to one of two classes. The algorithm can be automatically extended on datasets with more than two classes. Let us denote $N_i,\ \ i = 1, 2$ as the number of data objects of i th class, then $N = N_1 + N_2$.

The algorithm performs the feature ranking by classification of repeated random samples from initial gene expression matrix $X_{M\times N}$. To form the sample, the process of two-way bootstrapping is applied, namely $m < M$ genes and $n < N$ data objects are selected simultaneously from the initial matrix. The number of selected objects is defined by the parameter β, the number of genes is such that $m \le n/2$. Two-way bootstrapping scheme gives the possibility at each iteration k to form balanced matrix $Y^k_{m\times n}$, which according to research in [3] is less probable to overtraining. The number of iterations K is defined such that each gene in the dataset is selected the number of times sufficient to assess its rank. In addition to the number of iteration, the value of correlation between the rank vectors at the current and preceding iteration steps is used as another stopping criterion.

At each following step of iteration process, i.e., after the predefined number of iterations k_{iter} the output matrix $E^k_{M\times 4}$ is formed, which has the number of rows equal to the initial number of genes ($i = \overline{1, M}$) and four columns. The value of the first column T_i^k ($i = \overline{1, M}$) corresponds to the number of times the gene i is included into the random samples after subsequent k_{iter} iterations. The second column S_i^k ($i = \overline{1, M}$) values correspond to the number of times the gene is selected into the successful classification model. The successful classification model is the model with the classification error less than the predefined value of parameter α. The i th value of the third column $P_i^k = S_i^k / T_i^k$ ($i = \overline{1, M}$) presents the predictive power of the i th gene and the forth column R_i^k ($i = \overline{1, M}$) contains the gene rank values calculated on the basis of P_i^k after each k_{iter} iterations. Genes with higher values of P_i^k ($i = \overline{1, M}$) are more informative and, therefore, have a higher rank. Each iteration $k = \overline{1, K}$ of the algorithm consists in the following steps:

1. To construct the random sample $Y^k_{m\times n}$ at the k th iteration, the two-way bootstrapping procedure for the selection of rows and columns from the initial matrix $X_{M\times N}$ is performed. The matrix $Y^k_{m\times n}$ consists of m randomly selected genes from the whole set of cardinality M, n_1 samples from N_1 initial objects of the first class and n_2 samples from N_2 initial objects of the second class, such as $n_1 / N_1 = n_2 / N_2 = \beta$, $n_1 + n_2 = n$ и $m < n$. As a result, the training set of n data objects and the testing set of $N - n$ data objects are formed.

2. Training matrix $Y^k_{m\times n}$ is classified using the nearest shrunken centroid method [11]. The cross-validation procedure is applied to search for the optimal classifier with minimal classification error on the training sample $Y^k_{m\times n}$. Cross-validation allows defining the threshold value Δ^k, which corresponds to the lowest classification error, achieved with the fewest number of genes in the subset $x_1, x_2, \ldots, x_l$, where $l \le m$. The resultant classification model $C^k = f(x_1, x_2, \ldots, x_l)$ is subsequently verified on the test set with $N - n$ data objects. The classification error e is calculated as in (1).

$$e^k = \frac{FP + FN}{N - n}, \tag{1}$$

where FP – type I error, FN – type II error.

When the minimal classifier error on the training set for all possible threshold values Δ^k and gene subsets $x_1, x_2, \ldots, x_l$, $l \le m$ exceeds the value α, then the verification on the testing set is bypassed and the execution proceeds to step 4.

3. If the classification error on the test set $e^k \le \alpha$, i.e., less than a predefined threshold the classification model is assumed as successful and the output matrix $E_{M\times 4}$ is updated as follows. At first, for each gene x_i in matrix $Y^k_{m\times n}$, analysed at the k th iteration the corresponding gene in $E_{M\times 4}$ is defined and the values of columns T_i^k, S_i^k и P_i^k are sequentially modified as follows:

$$\begin{gathered} T_i^k = \begin{cases} T_i^{k-1} + 1, & x_i \in (x_1, x_2, \ldots, x_m) \\ T_i^{k-1}, & x_i \notin (x_1, x_2, \ldots, x_m) \end{cases} \\ S_i^k = \begin{cases} S_i^{k-1} + 1, & x_i \in (x_1, x_2, \ldots, x_l) \\ S_i^{k-1}, & x_i \notin (x_1, x_2, \ldots, x_l) \end{cases} \\ P_i^k = S_i^k / T_i^k . \end{gathered} \tag{2}$$

4. If the classification error on the test set $e^k > \alpha$, i.e., more than a predefined threshold the classification model is assumed as non-predictive and the selected genes as uninformative. The model is defined as overfitted on the training set and is discarded. The values of the columns T_i^k, S_i^k и P_i^k of the output matrix $E_{M\times 4}$ for each gene from the selected subset $(x_1, x_2, \ldots, x_m)$ are modified as follows:

$$\begin{gathered} T_i^k = \begin{cases} T_i^{k-1} + 1, & x_i \in (x_1, x_2, \ldots, x_m) \\ T_i^{k-1}, & x_i \notin (x_1, x_2, \ldots, x_m) \end{cases} \\ S_i^k = S_i^{k-1} \\ P_i^k = S_i^k / T_i^k . \end{gathered} \tag{3}$$

The rank values are defined by sorting the gene prognostic values P_i^k, $i = \overline{1, M}$ in descending order and are stored in the column R_i^k, $i = \overline{1, M}$ of the output matrix $E_{M\times 4}$.

After each k_{iter} iteration, the rank vector is repeatedly updated and the difference between the rank vector R^k at k th iteration and rank vector R^{k*} at $k*$ th iteration, where $k* = k - k_{iter}$, is estimated using the Spearman rank correlation coefficient like in (4).

$$r = 1 - 6\sum\nolimits_{i=1}^{M} \frac{\left(R_i^k - R_i^{k*}\right)^2}{M\left(M^2 - 1\right)}. \tag{4}$$

The algorithm stops if the correlation coefficient $r > \gamma$ (e.g., $\gamma = 0.99$), i.e., the rank vector is stabilized.

III. Datasets and Experiments

In our research we used the real Leukaemia dataset [12] to test the proposed gene ranking algorithm and to conduct the comparative analysis of the results. The Leukaemia dataset includes the bone marrow samples obtained from acute leukaemia patients at the time of diagnosis: 25 acute myeloid leukaemia (AML) samples and 47 acute lymphoblastic leukaemia (ALL) (24 acute 9 T-lineage acute lymphoblastic leukaemia samples; and 38 B-lineage ALL samples), characterized by expression of 7129 genes. According to the protocol in [13] the preprocessing steps include:

1) thresholding with minimal expression value of 100 and maximal expression value of 1600;

2) filtering the genes according to one of the conditions: $\max/\min < 5$ or $(\max - \min) \le 500$, where max – a maximum expression value and min – a minimum expression value;

3) logarithm transformation of gene values. After preprocessing, 3571 genes were selected. The dataset can be downloaded from the website [14].

For the experiment the following parameters were selected: the number of iterations $K = 300000$, the threshold value of the classification error $\alpha = 0.2$, the fraction of samples randomly selected from the initial dataset at each iteration $\beta = 0.7$, the number of iterations k_{iter} =20000 between the estimations of the stopping condition using the correlation coefficient with a threshold value $r = 0.09$. After each k_{iter} iteration, the output matrix $E_{M\times 4}$ was updated. Twenty top-ranked genes, received as a result of the experiment, are presented in Table I.

TABLE I

Description of the Top-Ranked Genes

No.	Gene definition	Description	No.	Gene definition	Description
1	Cd33	Myeloid cell surface antigen	11	DNTT	DNA nucleotidylexotransferase
2	ZYX	Zyxin	12	MARCKSL1	MARCKS-related protein
3	APLP2	amyloid beta (A4) precursor-like protein 2	13	CD79A	CD79a molecule, immunoglobulin-associated alpha
4	CST3	CST3	14	CTSA	cathepsin A
5	MGST1	microsomal glutathione S-transferase 1	15	CSTA	cystatin A (stefin A)
6	CTSD	Cathepsin D	16	SERPINB1	serpin peptidase inhibitor, clade B (ovalbumin), member 1
7	CFD	Adipsin	17	FAH	Fumarylacetoacetate hydrolase
8	CCND3	cyclin D3	18	CFP	complement factor properdin
9	CD63	Lysosomal-associated membrane protein 3	19	VPREB1	pre-B lymphocyte 1
10	TCF3	Transcription factor E2-alpha	20	SPTAN1	spectrin, alpha, non-erythrocytic 1
			21	MPO	myeloperoxidase

The classification models were constructed for each of the twenty gene subsets for the whole dataset, starting from the one-element subset, which consists of one top-ranked gene (see Table I), and then by adding the genes successively from the ranking list to form subsequent subsets. The classification errors were estimated and the gene subset with the minimal classification error was selected as the most informative one, which could be considered the potential disease biomarkers. Figure 1 shows the point graph of dependence of the classification error on the number of the selected genes, taken in the range from one to twenty.

According to error values (see Fig. 1) received using cross-validation procedure on the whole dataset, the four genes can be selected as the biomarkers (Cd33, Zyxin, APLP2, CST3). The accuracy of the classifier using only four genes equals 98.6%. As shown in Fig. 1, the set of ten genes gives a lower classification error, but requires the expression assessment of greater number of genes.

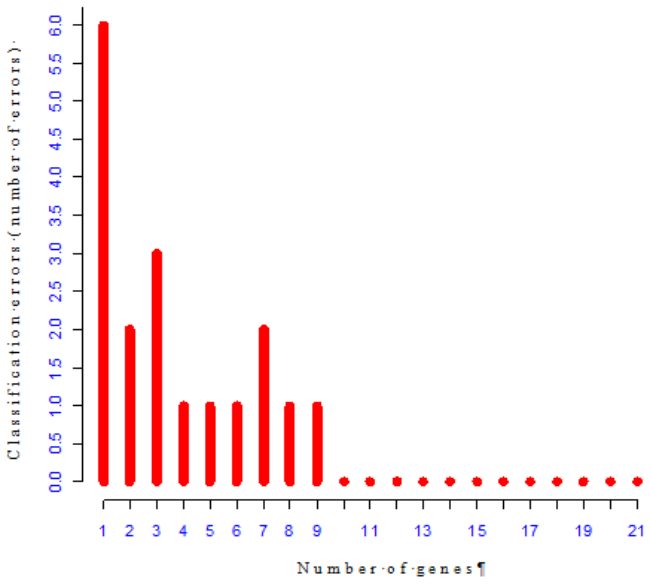

Fig. 1. Classification errors for number of genes from two to twenty

The comparative analysis of the results of the proposed algorithm has shown its efficiency in recognizing the most informative genes, which are also present in the gene lists in [15-18]. In addition, the proposed algorithm reaches the comparative classification accuracy (using the cross-validation procedure) using a smaller number of genes (Table II).

TABLE II
COMPARATIVE ANALYSIS OF THE DIFFERENT FEATURE SELECTION METHODS

	Classification accuracy	Number of genes
HBE method [15]	97.2%	4 genes
Method [16]	98.6%	132 genes
Method [17]	98.6%	3 gene clusters (the cluster size is from 1 up to 23 genes)
Proposed algorithm	98.6%	4 genes

IV. BIOLOGICAL SIGNIFICANCE OF THE SELECTED GENES

In order to assess the biological significance of the results, the ranked gene subset consisting of twenty most informative genes was analysed using a web knowledge resource for innate immunity interactions and pathways InnateDB (http://www.innatedb.ca). The potential functional interconnection of the gene subset with the disease phenotype and the biological pathways were revealed. The pathway over-representation analysis allowed defining the pathways that were significantly presented in the selected gene subset. The significance of the gene participation in the particular biological process was estimated by the p-value. According to the analysis, two (adipsin CFD and complement factor properdin CFP) out of the twenty genes take part in the alternative complement pathway, infectious agent suppression (p-level = 0.00006). In [18], adipsin was also selected as a biomarker of the leukaemia subtype, its role in the myeloid cell differentiation was described.

Two out of 20 genes (CD33 and DNTT) are of the hematopoietic cell lineage (p-level = 0.00429). CD33 is a transmembrane receptor expressed on cells of myeloid lineage and is usually considered myeloid-specific, but it can also be found on some lymphoid cells [19]. DNTT (deoxynucleotidyltransferase) is expressed in a restricted population of normal and malignant pre-B and pre-T lymphocytes during early differentiation.

Two genes of the subset (cyclin D3 и CD79A) are the members of the B Cell receptor signalling and cell cycle pathways (p-level = 0.01489). The cyclin D3 gene together with zyxin takes part in the process of Focal adhesion (p-level = 0.0225). In [20], zyxin was selected as a biomarker of the leukaemia phenotypes. The expression level of zyxin is important for the leukaemia subtype differentiation, but its role in hematopoiesis has not been reported. Zyxin proteins may regulate gene transcription by interaction with transcription factors.

According to the results of the biological significance analysis, the twenty top-ranked genes selected using the proposed algorithm are functionally important both for the process of hemopoietic cell differentiation and for the cancer pathogenesis. Most genes from the resultant gene subset are selected as biomarkers to recognize the leukaemia subtypes in several other studies [18-20].

V. ESTIMATION OF THE ALGORITHM CONVERGENCE

To define the optimal prognostic value of each gene, it is necessary to estimate all the possible samples of m genes out of initial data matrix $X_{M\times N}$, namely $C_M^m = (M-1)!/(m-1)!(M-m)!$ different combinations, which require about $O(M!)$ computations and take the enormous computational resources. The heuristic approach to estimate the prognostic quality of genes used in the proposed algorithm allows assessing the gene prognostic values with a smaller number of iterations, using the repeated analysis of the randomly selected data samples with $m << M$ genes from the initial matrix $X_{M\times N}$. Therefore, it is necessary to analyse the convergence of the proposed algorithm and its ability to receive the reliable estimation of the ranked prognostic values. For this purpose, we have analysed the variation of the rank vector for different number of iterations. As shown in Fig. 2, for twenty genes at the top of the ranked list their rank order differs a lot after 20000 and 40000 iterations.

However, after approximately 200000 iterations the rank order of genes is stabilized without significant changes. Therefore, the proposed algorithm converges to near optimal solution and is able to perform the reliable gene ranking, using a relatively small number of iterations.

VI. CONCLUSION

The proposed algorithm allows selecting the most informative genes by ranking, where the stability of the ranks of individual genes is ensured by estimation of multiple samples from an initial data matrix. Such an approach helps to avoid overfitting and enables the unbiased estimate of the vector of ranks. At each iteration, the classification model constructed on the randomly generated training sample is verified on the test sample. The classification accuracy is the indicator of the prognostic ability of the individual genes. After the successful classification, the ranks of the participating genes become higher, meanwhile the search for the optimal ranking is performed not for each individual gene, but for the whole combination. The output matrix is modified after the pre-determined number of iterations, registering both the prognostic ability and the ranks of the individual genes. The stability of the rank vector serves as an optimality criterion and is estimated by computing Spearman correlation coefficient between the current and previous rank order.

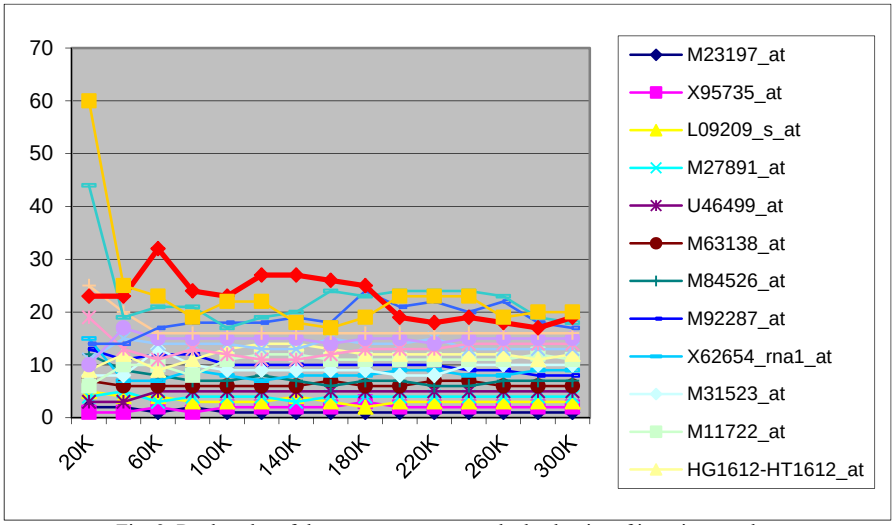

Fig. 2. Rank order of the twenty genes at each checkpoint of iteration number

The proposed algorithm has been tested on the leukaemia dataset and its convergence is analysed considering the twenty top-ranked genes. The analysis of the biological significance of the investigated gene subset allows confirming its obvious functional relevance to the phenotype it predicts and the processes taking place in the leukemic cells. It assures that the top-ranked genes are highly unlikely to be selected by chance. The comparative analysis of the developed algorithm on the leukaemia dataset shows its advantage over analogues, notably the selected set of biomarkers is smaller, consisting of four genes, which provide similar or higher classification accuracy.

REFERENCES

[1] X. Liu, A. Krishnan, and A. Mondry, "An entropy-based gene selection method for cancer classification using microarray data", in *BMC Bioinformatics*, vol. 6, no. 76, 2005.

[2] N. Novoselova and I. Tom, *Methods for gene expression analysis. Survey and perspective directions.* LAMBERT Academic Publishing GmbH&Co, 2012, 68 p.

[3] E.R. Dougherty, J. Hua, and C. Sima, "Performance of feature selection methods", in *Curr. Genomics*, vol.10, 2009, pp. 365–374.

[4] Y. Wang, I.V. Tetko, and M.A. Hall, "Gene selection from microarray data for cancer classification a machine learning approach", in *Comp Biol Chem.*, vol. 29, 2005, pp. 37–46.

[5] R. Kohavi and G. John, "Wrapper for feature subset selection", in *Artificial Intelligence*, vol. 97, no. 1, 1997, pp. 273-324.

[6] J.G. Thomas, J.M. Olson, S.J. Tapscott, and L.P. Zhao, "An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles", in *Genome Res.*, vol. 11, 2001, pp. 1227–1236.

[7] A. Antoniadis, S. Lambert-Lacroix, and F. Leblanc, "Effective dimension reduction methods for tumor classification using gene expression data", in *Bioinformatics*, vol. 19, 2003, pp. 563–570.

[8] I. Inza, P. Larranaga, R. Blanco, and A. Cerrolaza, "Filter versus wrapper gene selection approaches in DNA microarray domains", in *Artif. Intell. Med.*, vol. 31, no. 2, 2004, pp. 91–103.

[9] M. Xiong, Z. Fang, and J. Zhao, "Biomarker identification by feature wrappers", in *Genome Research*, vol. 11, 2001, pp. 1878-1887.

[10] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics", in *Bioinformatics*, vol. 23, 2007, pp. 2507-2517.

[11] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, "Diagnosis of multiple cancer types by shrunken centroids of gene expression", in *Proc Natl. Acad. Sci U S A*, vol. 99, 2002, pp. 6567–6572.

[12] T.R. Golub, D.K. Slonim, P. Tamayo, et al., "Molecular classification of Cancer: class discovery and class prediction by gene expression monitoring", in *Nature*, vol. 286, 1999, pp. 531-537.

[13] S. Dudoit, J. Fridlyand, and T. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data", in *J. Am. Stat. Assoc.*, vol. 97, 2002, pp. 77-87.

[14] Cancer Program Data Sets/ Broad Institute of Harvard and MIT. [Online]. Available: http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi . [Accessed: September 30, 2013].

[15] O. Dagliyan, F. Uney-Yuksektepe, I.H. Kavakli, and M. Turkay, "Optimization Based Tumor Classification from Microarray Gene Expression Data", *PLoS ONE*, 6(2): e14579. doi:10.1371/journal.pone.0014579, 2011.

[16] A. Antonov, I.V. Tetko, M.T. Mader, J. Budczies, and H.W. Mewes, "Optimization models for cancer classification extracting gene interaction information from microarray expression data", in *Bioinformatics*, vol. 20, 2004, pp. 644–652.

[17] M. Dettling and P. Buhlmann, "Supervised clustering of genes", in *Genome Biol.*, vol. 3, 2002: research0069.1–0069.15.

[18] W. Chu, Z. Ghahramani, F. Falciani, and D.L. Wild, "Biomarker discovery in microarray gene expression data with gaussian processes", in *Bioinformatics*, vol. 21, 2005, pp. 3385–3393.

[19] A.J. Yang and X.Y. Song, "Bayesian variable selection for disease classification using gene expression data", in *Bioinformatics*, vol. 26, 2010, pp. 215–222.

[20] Y. Wang et al., "Gene selection from microarray data for cancer classification – a machine learning approach", in *Comput. Biol. Chem.*, vol. 29, no. 1, 2005, pp. 37-46.

Natalia Novoselova has graduated from Belarus State University, Faculty of Mathematics and Mechanics (MSc. in mathematics, 1987). In 2008 she received her Doctoral Degree in Computer Sciences.

Currently she is a Senior Researcher at the Department of Bioinformatics, United Institute of Informatics Problems (UIIP), National Academy of Sciences of Belarus (NASB), Minsk, Belarus. At the beginning of her career she was engaged in mathematical modelling of different engineering processes. Since 2003 she is strongly interested in statistical and intelligent data analysis in relation to medical data. Her main 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 has published more than 30 papers in peer-reviewed journals and conference proceedings.

In 2008 and 2010 she received the two-month scientific grant from the German Academic Exchange Service to conduct research in the field of bioinformatics at the Ostfalia University of Applied Sciences, Wolfenbuettel, Gemany.

Contact information: Department of Bioinformatics, United Institute of Informatics Problems, 6 Surganova Str., Minsk, 220012, Belarus. Phone: +375-17-2842092, e-mail: novosel@newman.bas-net.by

Igor E. Tom received the diploma from the Belarusian State University of Informatics and Radioelectronics in 1977. In 1984–1986 he was the postgraduate student at the Institute of Engineering Cybernetics of National Academy of Sciences of Belarus (IEC NASB). In 1986 he received the Doctoral Degree in Computer Science from IEC NASB.

Since 1999 he has been the Head of the Department of Bioinformatics at the United Institute of Informatics Problems (former IEC) of NASB. His current research interests include the development of intelligent data analysis methods and information technologies for medical and industrial applications. He is the author and co-author of more than 170 scientific publications. His last publications are devoted to intelligent data analysis methods for solving classification tasks, data clustering and revealing of decision rules.

Contact information: Department of Bioinformatics, United Institute of Informatics Problems, 6 Surganova Str., Minsk, 220012, Belarus. Phone: +375-17-2842153, e-mail: tom@newman.bas-net.by

Arkady Borisov received his Doctoral Degree in Technical Cybernetics from Riga Polytechnic Institute in 1970 and Dr.habil.sc.comp. degree in Technical Cybernetics from Taganrog State Radio Engineering University in 1986. He is a Professor of Computer Science at the Faculty of Computer Science and Information Technology, Riga Technical University (Latvia). His research interests include fuzzy sets, fuzzy logic and computational intelligence. He has 235 publications in the fields of computer science and information technology.

He has supervised a number of national research grants and participated in the European research project ECLIPS.

He is a member of IFSA European Fuzzy System Working Group, Russian Fuzzy System and Soft Computing Association, honorary member of the Scientific Board, member of the Scientific Advisory Board of the Fuzzy Initiative Nordrhein-Westfalen (Dortmund, Germany).

Contact information: Riga Technical University (RTU), 1 Kalku Street, Riga LV- 1658. Phone: +371 67089530, e-mail: arkadijs.borisovs@cs.rtu.lv.

Inese Polaka is the fourth-year postgraduate student at Riga Technical University. She finished her Master studies at Riga Technical University majoring in Information Technology in 2010 obtaining the Mg.sc.ing. degree. Her research interests include machine learning methods and classification tasks in bioinformatics, decision tree classifiers, classifier efficiency improvement methods, use of ontology in machine learning, ontology-based classifier design, descriptive statistics, and exploratory data analysis.

Contact information: Riga Technical University, 1 Kalku Street, Riga LV-1658. Phone: +371 67089530, e-mail: inese.polaka@rtu.lv.

Natalija Novoselova, Igors Toms, Arkādijs Borisovs, Inese Poļaka. Īpašību ranžēšana, izmantojot daudzējādu datu kopu klasifikācijas precizitātes vērtējumu

Rakstā izskatīts gēnu ranžēšanas algoritms, kuru var izmantot ar datiem, kas iegūti no mikročipu tehnoloģijas pielietošanas. Piedāvātais algoritms ļauj noteikt informatīvākos gēnus gēnu ekspresijas datos, turklāt atsevišķu gēnu rangu nozīmes stabilitāti garantē vairākkārtējs novērtējums datu kopās, kas iegūtas no sākotnējās datu kopas. Tas ļauj izvairīties no pārapmācības un nerada nobīdes rangu vektora atzīmēs. Atbilstoši algoritmam, katrā iterācijā klasifikācija tiek veikta apmācības datu kopā, kura gadījuma veidā izveidota no sākotnējās datu kopas, kā arī tiek veikta klasifikācijas modeļa verifikācija, izmantojot testa datu kopu. Klasifikācijas precizitāte ir mērs, pēc kura novērtē atlasīto gēnu prognostiskās īpašības. Katrā nākamajā algoritma iterācijā rangs gēniem, kas piedalījās veiksmīgā klasifikācijā, tiek paaugstināts, turklāt optimālā aranžējuma meklēšana notiek nevis katram gēnam atsevišķi, bet gan visai atlasītajai kombinācijai. Izejas matrica tiek modificēta pēc noteikta iterāciju skaita izpildīšanas, fiksējot gan tās prognostiskās spējas, gan atsevišķu gēnu rangu. Rangu vektora optimalitātes kritērijs ir tā vērtību stabilizācija, kuru novērtē, izmantojot Spīrmena korelācijas koeficientu. Piedāvātais algoritms ir notestēts, izmantojot leikēmijas datu kopu, un ir veikta algoritma saderības novērtēšana, izmantojot 20 gēnus, kas atrodas ranžētā saraksta augšpusē. Tika novērtēta apskatītās atlasīto gēnu apakškopas bioloģiskā nozīme, kas ļauj izdarīt secinājumus par to ciešo saistību ar procesiem, kas notiek leikēmijas šūnās. Šādā veidā var secināt, ka atlasītie gēni ar augstāko ranga vietu nav izvēlēti nejauši. Salīdzinošā analīze ar citu autoru darbiem, kas īstenoti, izmantojot šo pašu datu kopu, uzrādīja piedāvātā algoritma priekšrocību, jo, izmantojot to, tika atlasīta vismazākā biomarķieru kopa, kas saturēja četrus gēnus un nodrošināja līdzvērtīgu vai labāku klasifikācijas precizitāti.

Наталья Новоселова, Игорь Том, Аркадий Борисов, Инесе Поляка. Ранжирование признаков для обнаружения биомаркеров в данных генной экспрессии

В статье рассматривается алгоритм ранжирования генов, полученных с использованием технологии микрочипов. Предложенный алгоритм позволяет выделять наиболее информативные гены в данных генной экспрессии, причем стабильность значений рангов отдельных генов обусловлена многократной оценкой выборок из исходной матрицы данных, что помогает избежать переобучения и способствует получению несмещенных оценок вектора рангов. Согласно алгоритму, на каждой итерации выполняется классификация случайным образом сформированной обучающей выборки с верификацией классификационной модели на тестовой выборке. Точность классификации является показателем прогностической способности отобранных генов. На каждой последующей итерации алгоритма ранг генов, участвующих в успешной классификации повышается, при этом поиск оптимального ранжирования осуществляется не для каждого гена в отдельности, а для всей отобранной комбинации. Выходная матрица модифицируется после выполнения заданного количества итераций, фиксируя как прогностическую способность, так и ранг отдельных генов. Критерием оптимальности вектора рангов является стабилизация его значений, что оценивается с использованием критерия корреляции Спирмена. Предложенный алгоритм протестирован на наборе данных по лейкемии, и проведена оценка сходимости алгоритма на примере 20 генов, расположенных вверху ранжированного списка. Оценена биологическая значимость исследуемого подмножества отобранных генов, которая позволяет сделать вывод об их тесной связи с процессами, происходящими в лейкемических клетках. Таким образом, отобранные гены с наивысшим рангом не являются результатом случайного отбора. Сравнительный анализ с работами других авторов по исследуемому набору данных показал преимущество предложенного нами алгоритма, так как после его применения было отобрано наименьшее подмножество из четырех биомаркеров, обеспечивающих схожую или лучшую точность классификации.