gene selection, partial least squares
stability selection, bootstrap .632+
multiclass classification, BBFR

Sebastian STUDENT[1], Krzysztof FUJAREWICZ[1]

STABILITY OF GENE SELECTION METHODS FOR MULTICLASS CLSSIFICATION

A big problem in applying DNA microarrays for classification is dimension of the dataset. Recently we proposed a gene selection method based on Partial Least Squares (PLS) for searching best genes for classification. The new idea is to use PLS not only as multiclass approach, but to construct more binary selections that use one versus rest and one versus one approaches. Ranked gene lists are highly instable in the sense, that a small change of the data set often leads to big change of the obtained ordered list. In this article, we take a look at the assessment of stability of our approaches. We compare the variability of the obtained ordered lists from proposed methods with well known Recursive Feature Elimination (RFE) method and classical t-test method. This paper focuses on effective identification of informative genes. As a result, a new strategy to find small subset of significant genes is designed. Our results on real cancer data show that our approach has very high accuracy rate for different combinations of classification methods giving in the same time very stable feature rankings.

1. BACKGROUND

A big problem in applying microarrays in classification problem is dimension of this data [12]. Traditional statistical methodology for classification does not work well when there are more variables than samples. Thus, methods able to cope with the high dimensionality of the data are needed. In this paper we describe multiclass classification and dimension reduction which are intrinsically more difficult than binary ones [15]. The gene selection for the classifier is a very important problem. Over the past few years many algorithms were proposed to resolve this problem. However, most of the studies are designed to binary dimension reduction problems and only a few involve multiclass cases. The optimal selection of informative genes for multiclass analysis is still an open problem. Recently we proposed a gene selection method [14] based on Partial Least Squares (PLS) [11]. Then we compare the results with Recursive Feature Elimination (RFE) method [10] and the classical t-statistic.

The standard way to use PLS algorithm is only for dimension reduction, not for selecting significant features. Here, we use this method for searching best genes for classification. The new idea is to use PLS not only as multiclass approach, but to construct more binary selections that use one versus rest (OvR) and one versus one (OvO) methods.

An important aspect of features selection methods is the stability of obtained ordered lists [3], [12]. In this paper by stability, we mean the invariability of the ranked lists obtained with the same method, but with a little bit modified dataset. Hence, the term stability is used in different sense that in classical system theory. The problem of gene lists stability is very important for the confirmation of the obtained lists with biological methods and for the clinical applicability of molecular markers. For example for long genes lists, the investigators will test only the most important genes, in this case the top-ranked genes.

2. FEATURE SELECTION

In this section we describe our approach to select the most significant genes for our dataset. Partial least squares (PLS) is known method for dimension reduction [13], but the standard way to use this algorithm is only for dimension reduction and not for selecting significant genes.

In contrast to dimensionality reduction techniques like those based on projection (e.g. principal component analysis) or compression (e.g. using information theory), feature selection techniques do not

[1] Silesian University of Technology, Automatic Control, Electronics and Computer Science; Automatic Institute, Akademicka 16, 44-100 Gliwice, Poland

alter the original representation of the variables, but merely select a subset of them. Since we need method for selecting genes for classification, we modify the PLS method. In this article we use the weight vector obtained from PLS method to find genes that differentiate cancer types.

PLS aims at finding uncorrelated linear transformations of the original input features which have high covariance with the response features. Based on these latent components, PLS predicts response features, the task of regression, and reconstruct original dataset matrix, the task of data modeling, at the same time. For dataset matrix $\boldsymbol{X}$ of size $l \times p$ with l probes and p genes we denote the $l \times 1$ vector of response value $\boldsymbol{y}$. In the PLS the components $\boldsymbol{t}_i \quad i = 1,...,q$ are constructed to maximize the objective criterion based on the sample covariance between $\boldsymbol{y}$ and linear combination of genes (PLS components) $\boldsymbol{t} = \boldsymbol{X}\boldsymbol{w}$. We search sequentially the weight vector $\boldsymbol{w}$ to satisfy the following criterion

$$\boldsymbol{w}_i = \arg\max_{w^T w=1} \operatorname{cov}^2(\boldsymbol{X}\boldsymbol{w}, \boldsymbol{y}) \tag{1}$$

subject to the orthogonality constraint

$$\begin{aligned} &\boldsymbol{w}_i^T \boldsymbol{S}_X \boldsymbol{w}_j = 0 \\ &1 \le i < j \\ &\boldsymbol{S} = \boldsymbol{X}'\boldsymbol{X} \end{aligned} \tag{2}$$

To derive the components, $\boldsymbol{t}_i \quad (i = 1,...,q)$, the PLS decomposes $\boldsymbol{X}$ and $\boldsymbol{y}$ to produce a bilinear representation of the data

$$\begin{aligned} \boldsymbol{X} &= \boldsymbol{t}_1 \boldsymbol{w}_1^T + \boldsymbol{t}_2 \boldsymbol{w}_2^T + ... + \boldsymbol{t}_q \boldsymbol{w}_q^T + e \\ \boldsymbol{y} &= \boldsymbol{t}_1 \boldsymbol{v}_1^T + \boldsymbol{t}_2 \boldsymbol{v}_2^T + ... + \boldsymbol{t}_q \boldsymbol{v}_q^T + f \end{aligned} \tag{3}$$

where $\boldsymbol{v}$ is weight vector for matrix $\boldsymbol{y}$ and e, f are residuals. The idea of PLS is to estimate $\boldsymbol{w}$ and $\boldsymbol{v}$ by a regression. The PLS fits a sequence of bilinear models by least squares. At every step i $(i = 1,...,q)$ the vector $\boldsymbol{w}_i$ is estimated to obtain the PLS component that has maximal sample covariance with the response variable $\boldsymbol{y}$. Each component $\boldsymbol{t}_i$ is uncorrelated with all previously constructed components. The first PLS component $\boldsymbol{t}_1$ is obtained on a basis of the covariance between $\boldsymbol{X}$ and $\boldsymbol{y}$. Component $\boldsymbol{t}_i \quad (i = 2,...,q)$, is computed using the residuals of $\boldsymbol{X}$ and $\boldsymbol{y}$ from the previous step, which account for the variations left by the previous components. Maximal number of components q is equal to the rank of $\boldsymbol{X}$.

There are two main PLS algorithms described in literature: NIPALS algorithm [9] and SIMPLS algorithm [4]. In contrast with weight vector $\boldsymbol{w}$ obtained from very popular NIPALS algorithm, the weight vector $\boldsymbol{r}$ in SIMPLS algorithm is calculated directly form $\boldsymbol{X}$ without deflation. De Jong [4] showed that we can calculate the weights $\boldsymbol{r}$ directly from the NIPALS algorithm

$$\boldsymbol{r}_i = \boldsymbol{w}_i (\boldsymbol{p}_i' \boldsymbol{w}_i)^{(-1)} \tag{4}$$

where:

$\boldsymbol{p}_i$ - are the loadings,

$\boldsymbol{w}_i$ - are the weights vector for i-th component.

In this paper, we use the weights vector $\boldsymbol{r}$ from SIMPLS algorithms to determine our ranked list. To test the optimal number of components we use only the first weights vector and the multiplication of weights vectors from first 3, 5 and 10 components.

In our approach the sorted $\boldsymbol{r}$ presents the genes ranking and the "best genes" have the biggest absolute value in this vector. First g genes with the highest value in the weights vector are selected for classifier. The standard way to use PLS is multiclass approach, is to search the best direction for maximize the covariance between responses with all classes and linear combination of genes. For multiclass problem the known methods are based on the idea of selecting genes to distinguish all classes. The new idea is to use PLS not only as multiclass approach, but to construct a set of two-class selections that use one versus rest (OvR) and one versus one (OvO) methods. For each two-class selections "best genes" are selected and one ranked genes list is constructed as follows: genes with the highest weights in all binary selections have place at the top of the list, then genes with the second highest weights, and so on. We must underline, that $\mathbf{y}$ for two-class selections is coded as a vector with value 1 for first class and -1 for second class. For multiclass approach $\boldsymbol{y}$ is matrix with N rows. In each row class label equals to row number has value 1 and -1 otherwise. For our needs we introduce the notation PLS+MCLASS for multiclass feature selection approach and similarly PLS+OvO, PLS+OvR for binary approaches.

3. ANALYSIS METHODS

It is shown that bootstrap methodology [5], [6] gives better performance than cross-validation and resubstitution for relatively small sample microarray classification [2]. In this paper we use the balanced bootstrap to reduce error variance and bias over the bootstrap method. The 0.632+ bootstrap estimator is used for an accuracy of the classifier calculation. 500 resampling iterations of all stages of the classifier construction (i.e. gene preselection, gene selection and classifier learning) were performed. We generate the 500 bootstrap sample only ones for all tested methods to reduce the variability of the randomization in the tests results of different methods and parameters. The distribution of the misclassification rate obtained during all bootstrap runs was used to estimate the 95% confidence interval. The accuracy of the classifier and the confidence interval were calculated up to 30 genes.

3.1. STABILITY OF ORDERED GENE LISTS

The stability of obtained gene list in the meaning of similarity between lists from the same experiment, but slightly changed data set. To show the distance between different gene selection methods we use method based on bootstrap resampling. This approach is based on the comparison of sets consisting of a fixed number of top g genes. In our framework we consider the list $\boldsymbol{L}$ with first g top-genes obtained from whole dataset and lists $\boldsymbol{L}_b; b = 1, 2, \ldots, B$ obtained from every b of B bootstrap iterations. The relative s score is introduced to estimate the similarity between all lists:

$$s = 1 - \sum_{b=1}^{B} \sum_{j=1}^{g} \frac{| r_j - r'_{bj} |}{Bg(g+1)/2} \tag{5}$$

where:

r_j - is the ranking (placement in the list) of the j-th gene in the list $\boldsymbol{L}$

r'_{bj} - is the ranking of the j-th gene in the list $\boldsymbol{L}_b$.

The ranking for the gene that is out of $\boldsymbol{L}_b$ is set to $g+1$. The value of function s is scaled to the interval $<0,1>$ and the higher value indicates better stability of obtained genes list. In this approach we don't ignore the rank of the selected genes within the considered subset.

To visualize the stability of ordered gene lists we plot the boxplots of rank for each gene in the list $\boldsymbol{L}$ against ranks in all bootstrap iteration lists $\boldsymbol{L}_b; b = 1, 2, \ldots, B$. The limit to determine which points are extreme we set to the rank out of the g genes list.

Another indicator used to estimate the stability of obtained gene list we use the number of genes that were selected at least one time in all bootstrap samples. The best value is g and the worst is

$\mathrm{Max}(G, Bg)$ where G is the number of all genes. This approach is equal to the number of genes with non-zero score in the Bootstrap Based Feature Ranking (BBFR) (described in the next section).

3.2. BOOTSTRAP BASED FEATURE RANKING

Bootstrap can be used not only for estimating an accuracy our classifier with appropriate confidence intervals. We also apply Bootstrap Based Feature Ranking (BBFR) [8] method that use the information collected during bootstrapping. In this approach we use the information collected during bootstrap-based validation step of the classifier.

Let the data contains m instances (observations). One instance is a vector of G features (gene expression values) with a corresponding class label. To find the optimal size for the feature set we select g feature sets $\Omega_1, \Omega_2, ..., \Omega_g$, of sizes 1, 2, ..., g respectively. In general, selected sets may not overlap, but in most commonly used feature selection methods, based on feature ranking or backward/forward searching, feature subsets satisfy the relation

$$\Omega_1 \subset \Omega_2 \subset ... \subset \Omega_g . \tag{6}$$

Let r_{bj} be a number of subsets $\Omega_i, i = 1, 2, ..., g$ obtained in *b*-th bootstrap iteration where the gene j belongs to. For gene selection methods satisfying relation (6) we have

$$q_{bj} = g - r_{bj} + 1 . \tag{7}$$

The BBFR score Q_j of the feature j is defined as a sum of q_{bj} over all bootstrap runs

$$Q_j = \sum_{b=1}^{B} q_{bj} . \tag{8}$$

The maximum possible value of the Q_j score is Bg. It means that one gene were top-ranked in all B bootstrap iterations. The score Q_j takes into account the ranking r_{bj} of *j*-th gene in all B bootstrap iterations. The modified BBFR score Q'_j takes into account only the presence of the gene j in the lists $\boldsymbol{L}_b; b = 1, 2, \ldots, B$:

$$Q'_j = \sum_{b=1}^{B} q'_{bj} , \tag{9}$$

where $q'_{bj} = 1$ if $j \in L_b$ and $q'_{bj} = 0$ otherwise. The maximum possible value of the Q'_j score is B.

4. RESULTS AND DISCUSSION

We chose public available multiclass microarray LUNG dataset for our experiments. This dataset is the published by [1]. It consists of 254 samples of 4 subtypes of lung carcinomas and normal samples. Samples was normalized by RMA and GA annotation [7]. Each sample has 8359 genes expression levels after re-annotation. The data is available at: http://www.broadinstitute.org/mpr/lung/.

Numerical experiment include classification methods (SVM OvO, SVM OvR, MSVM, and LDA). We demonstrate the usefulness of the proposed methodology to select significant genes with PLS. All approaches: PLS+OvO, PLS+OvR and PLS+MSVM was tested and compared with RFE method and classical t-test (both implementation as OvO and OvR) . As we said before for each approaches we executed 500 bootstrap iteration.

In every case we perform gene selection using the PLS procedure choosing 30 best genes according to approaches PLS+OvO, PLS+OvR and PLS+MCLASS. As we can see in the Tab.1 very good accuracy rate and stability index we obtain for the PLS+OvR gene selection method. Our methods, especially PLS+OvR perform comparably well accuracy rate to RFE method, but looking at stability index and number of reselected genes outperform significantly to other methods(Fig.1). We must underline, that RFE method need to select at least 471 genes in all bootstrap iteration.

Table 1 The bootstrap based classification accuracies, stability index and number of reselected genes in all bootstrap samples based on all tested gene selection methods, on the LUNG dataset. The number of selected genes is set to 30, together with their bootstrap based standard deviations. The number of reselected genes is the sum of non zero bootstrap-based feature ranked genes (BBFR).

Selection method			stability index	reselected genes	Classification method: SVM OvO acc	SVM OvR acc	MSVM acc	LDA acc
SIMPLS	OvO	1 comp.	0.69	81	0.956±0.028	0.945±0.043	0.951±0.036	0.956±0.025
	OvR	**1 comp.**	**0.74**	**77**	0.955±0.031	0.945±0.04	0.947±0.044	**0.965±0.029**
	MCLASS	1 comp.	0.79	69	0.927±0.046	0.882±0.069	0.891±0.079	0.894±0.046
	OvO	5 comp.	0.51	182	0.955±0.034	0.928±0.067	0.944±0.045	0.965±0.034
	OvR	5 comp.	0.61	171	0.959±0.033	0.946±0.04	0.952±0.034	0.966±0.029
	MCLASS	5 comp.	0.58	100	0.955±0.036	0.944±0.04	0.948±0.036	0.916±0.045
	OvO	10 comp.	0.43	211	0.953±0.036	0.926±0.077	0.941±0.044	0.945±0.036
	OvR	10 comp.	0.49	225	0.957±0.034	0.944±0.055	0.95±0.042	0.961±0.03
	MCLASS	10 comp.	0.47	132	0.957±0.033	0.948±0.037	0.949±0.037	0.941±0.047
RFE	OvO		0.36	471	0.961±0.031	0.95±0.036	0.96±0.036	**0.976±0.028**
	OvR		0.23	763	**0.966±0.03**	0.962±0.031	0.965±0.028	0.965±0.029
T-TEST	OvO		0.02	331	0.956±0.037	0.935±0.054	0.95±0.039	0.951±0.036
	OvR		0.44	629	0.942±0.041	0.925±0.052	0.934±0.045	0.851±0.061

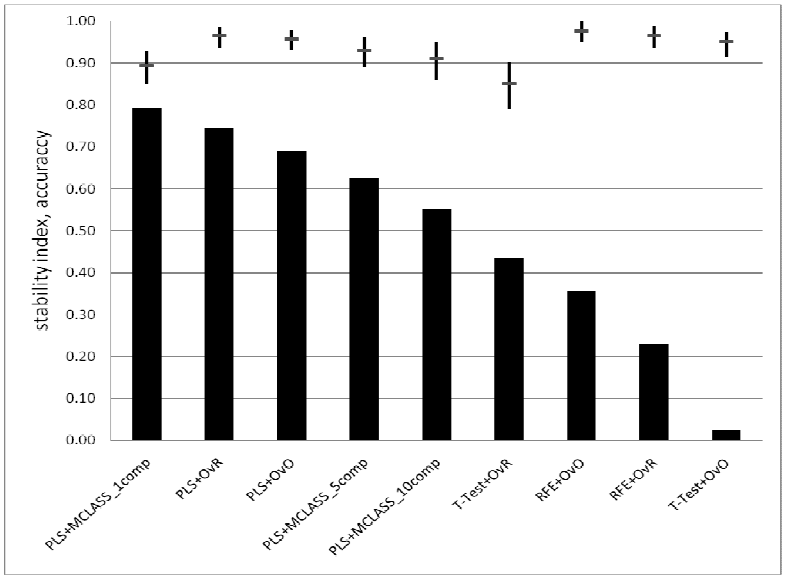

Fig.1. Stability index (bar chart) and accuracy of classification (dot chart) with the 95% confidence interval of the LDA classifier on the tested features selection methods.

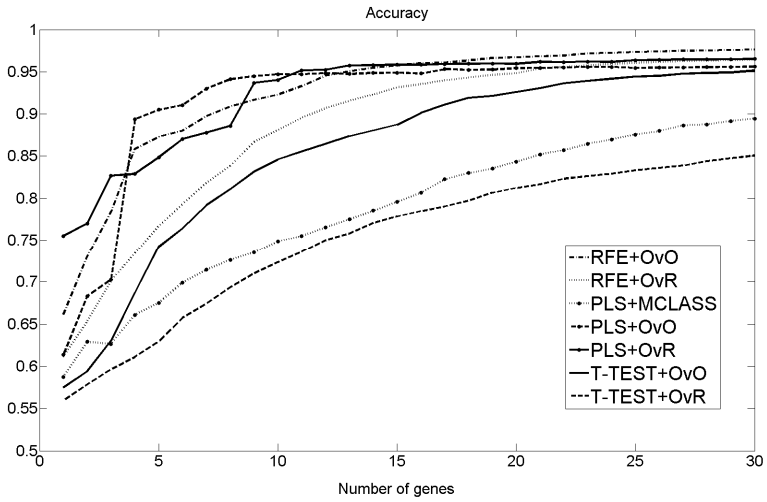

Fig.2. Accuracy of classification obtained by successive gene set reduction, selected with all features selection methods of the LDA classifier.

Fig.2 shows comparison of different gene selection methods for the LDA classifier. In most cases 30 genes is sufficient enough for obtain maximum accuracy rate. Fig.3 shows results of bootstrap-based future ranking where every selected gene revives one point in every bootstrap iteration. Our gene selection methods based on PLS outperform the other methods. There are more genes selected in each bootstrap samples and there are smaller number genes selected at least one time in comparison to RFE and t-test.

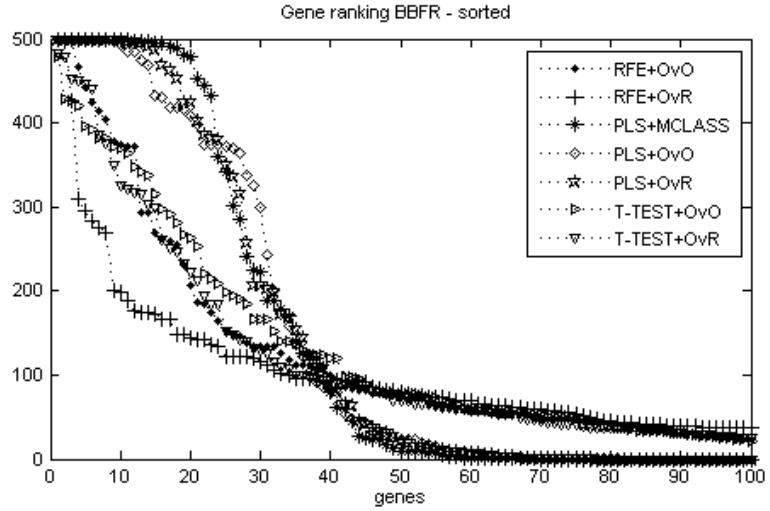

Fig.3. Results of bootstrap-based feature ranking (BBFR). Every dot represents one gene.

The comparison of rank boxplots in the bootstrap samples against rank in the original dataset on two gene selection methods PLS+OvR and RFE +OvO proofs the instability of genes ranking obtained with RFE method. As we can see there are a lot of selected genes (obtained from whole dataset) not included in the bootstrap based gene lists.

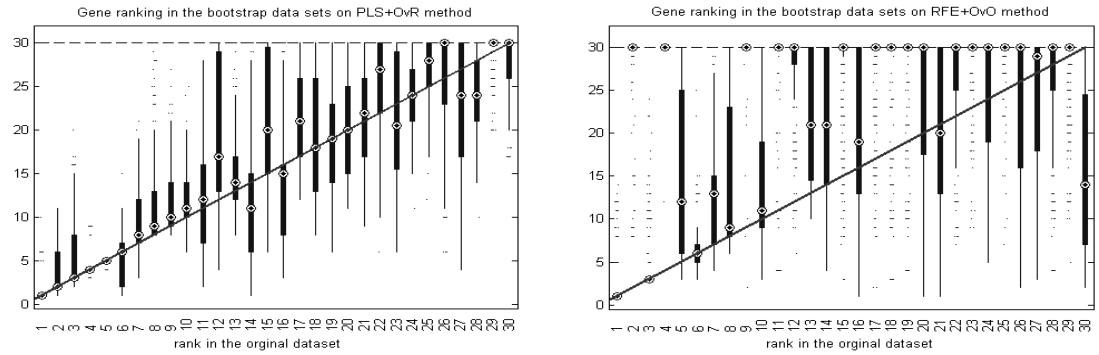

Fig.4. Comparison of rank boxplots in the bootstrap samples against rank in the original data set on two genes selection methods PLS+OvR (left) and RFE+OvO (right).

In this paper we described a new PLS-based method to select significant genes. PLS method with OvR approach for gene selection brings the best results for all tested methods, when we take into consideration classification accuracy and stability of obtained gene list. For this dataset it is more effective to solve a multiclass gene selection by splitting it into a set of two-class problems and merge the results in one gene list. The stability of gene selection methods should be investigated as an important part of genomic analyses, because some gene selection methods show high gene lists variability.

BIBLIOGRAPHY

[1] BHATTACHARJEE A., Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses, 98(24): PNAS 2001, pp. 13790–13795.
[2] BRAGA–NETO U, DOUGHERTY ER., Is cross–validation valid for small–sample microarray classification? 20(3), Bioinformatics 2004, pp. 374–380.
[3] BOULESTEIX AL, SLAWSKI M., Stability and aggregation of ranked gene lists, Brief Bioinform (2009) 10: pp. 556–568.
[4] DE JONG S., SIMPLS: An alternative approach to partial least squares regression, Vol. 18, Chemometrics Intell. Lab. Syst. 1993, pp. 25–263.
[5] EFRON B., Bootstrap methods: another look look at the jackknife, Vol. 7, Annals of Statistics 1979, pp. 1–26.
[6] EFRON B., TIBSHIRANI R., Improvements on cross–validation: the 632+ bootstrap method. Vol. 92, J. Amer. Statist. Assoc. 1997, pp. 548–560.
[7] FERRARI F., BORTOLUZZI S., COPPE A., SIROTA A., SAFRAN M., Novel definition files for human GeneChips based on GeneAnnot. 8(446), BMC Bioinformatics 2007.
[8] FUJAREWICZ K., A multigene approach to differentiate papillary thyroid carcinoma from benign lesions: gene selection using bootstrap–based Support Vector Machines. Vol. 14, Endocrine – Related Cancer 2007, pp. 809–826.
[9] GELADI P., KOWALSKI BR., Partial Least–Squares Regresion: a Tutorial, Vol. 185, Analytica Chimica Acta 1986, pp. 1–17.
[10] GUYON I., WESTON J., BARNHILL S., VAPNIK V., Gene selection for cancer classification using support vector machines, Vol. 46 , Machine Learning 2002, pp. 389–422.
[11] HŐSKULDSSON A., PLS regression methods. Vol. 2(3), J. Chemometrics 1988, pp. 211–228.
[12] MEINSHAUSEN N., BÜHLMANN P., Stability selection (with discussion). Journal of the Royal Statistical Society 2010: Series B, pp. 417−473.
[13] NGUYEN DV., ROCKE DM., Tumor classification by partial least squares using microarray gene expression data. Vol. 18(1), Bioinformatics 2002, pp. 39–50.
[14] STUDENT S., FUJAREWICZ K., Multiclass cancer classification and biomarker Discovery on microarray data. XV Krajowa Konferencja Zastosowań Matematyki w Biologii i Medycynie, Szczyrk 2009, pp. 130−136.
[15] ZHANG T., LI C., OGIHARA M., A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression, Vol. 20(15), Bioinformatics 2004, pp. 2429–2437.

This work was supported by Silesian University of Technology under grant BW/Rau1/2010.