Int. J. Appl. Math. Comput. Sci., 2014, Vol. 24, No. 3, 657–668
DOI: 10.2478/amcs-2014-0048

DATA MINING METHODS FOR GENE SELECTION ON THE BASIS OF GENE EXPRESSION ARRAYS

MICHAŁ MUSZYŃSKI [*], STANISŁAW OSOWSKI [*,**]

[*] Faculty of Electrical Engineering
Warsaw University of Technology, pl. Politechniki 1, 00-661 Warsaw, Poland
e-mail: muszyna22@wp.pl

[**]Faculty of Electronic Engineering
Military University of Technology, ul. Kaliskiego 2, 00-908 Warsaw, Poland
e-mail: sto@iem.pw.edu.pl

The paper presents data mining methods applied to gene selection for recognition of a particular type of prostate cancer on the basis of gene expression arrays. Several chosen methods of gene selection, including the Fisher method, correlation of gene with a class, application of the support vector machine and statistical hypotheses, are compared on the basis of clustering measures. The results of applying these individual selection methods are combined together to identify the most often selected genes forming the required pattern, best associated with the cancerous cases. This resulting pattern of selected gene lists is treated as the input data to the classifier, performing the task of the final recognition of the patterns. The numerical results of the recognition of prostate cancer from normal (reference) cases using the selected genes and the support vector machine confirm the good performance of the proposed gene selection approach.

Keywords: gene expression array, gene ranking, feature selection, clusterization measures, fusion, SVM classification.

1. Introduction

A DNA microarray is a collection of microscopic DNA spots attached to a solid surface used to measure simultaneously the expression levels of a large number of genes. Each DNA spot contains a specific DNA sequence, being the short section of a gene or other DNA element used to hybridize a cDNA or cRNA sample (called the target) under high-stringency conditions. Monitoring gene expression using microarrays is an important problem in the study of cell functions, candidate gene identification, cellular response to different drugs as well as classification of disease states (De Rinaldis, 2007; Ramaswamy *et al.*, 2001).

DNA microarrays typically store data of thousands of expressions of individual genes. A common way to represent the data set produced through DNA microarray experiments is to form a matrix in which the row corresponds to the particular individual and the column represents the expression levels of different genes. Typically, we have the number of rows in the range of hundreds and the number of columns (genes or gene sequence) of a several or tens of thousands.

Comparing the gene expression profiles and selecting those which are best associated with the analyzed types of data in the high dimensional space of a small number of observations represent a formidable challenge in pattern recognition, which can be solved using specialized methods of data mining. Present approaches to this task include various clustering methods (Eisen *et al.*, 1998; Herrero *et al.*, 2001), application of neural networks and support vector machines (Guyon *et al.*, 2002; Huang and Kecman, 2005; Wiliński and Osowski, 2012), statistical tests (Baldi and Long, 2001), linear regression methods applying forward and backward selection (Huang and Pan, 2003), fuzzy logic based algorithms (Woolf and Wang, 2000), rough set theory (Wang and Gotoh, 2009; 2010; Świniarski, 2001), various statistical methods (Mitsubayashi *et al.*, 2008; Golub *et al.*, 1999), as well as a fusion of many selection methods (Wiliński and Osowski, 2012; Yang, 2011). Although the progress in this field is fast, there is still a need for better understanding and improvement of the research.

This paper presents an analysis and a comparison of methods of gene selection which are strongly associated with the prostate cancer. This particular problem was considered, for example, by Wang and Gotoh (2010) by using rough set theory, or by Wiliński and Osowski (2012) by using several methods of selection. The issue is a typical feature selection task of data mining (Duda *et al*., 2003; Guyon and Elisseeff, 2003; Tan *et al*., 2006). The most difficult problem is that the number of genes is extremely large (more than ten thousand) and the number of patterns—very limited (around one hundred). The selected genes of the particular expression levels form the most characteristic pattern for the given type of the cancer. Applying a classifier to given data should lead to an improved accuracy of the recognition of cancer cases from non-cancerous ones (Furey *et al*., 2000; Makinaci, 2007; Wiliński and Osowski, 2012).

In the numerical experiments we will analyse many different gene selection methods in the prostate cancer problem, containing two classes of data, either tumor or non-tumor cases. The experimental data set of patterns $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m \in \mathbb{R}^n$ representing the values of gene expressions with known class labels $d_1, d_2, \dots, d_m \in \{-1, 1\}$ standing for these two particular classes, where $m \ll n$. The results of such analysis will be presented in numerical and visual forms.

The small subset of the most representative features (gene expression coefficients) will be used to train the Support Vector Machine (SVM) classifier (Haykin, 1999; Scholkopf and Smola, 2002), the network generating the decision function $D(\mathbf{x}_i)$ for the particular input pattern vector $\mathbf{x}_i \in \mathbb{R}^l$, where $i \in \{1, \dots, m\}$ and $l \ll n$. The trained classifier may then be used to recognize and assign the newly acquired data to the appropriate class.

The results of numerical experiments concerning selection of the most important genes in prostate cancer as well as classification of cases on the basis of the selected genes will be discussed. The main contribution of the paper is developing a fusion system of various selection approaches into the final set, most closely associated with the cancer. Also, we propose a special procedure which estimates the number of higher rank genes using the self-organization procedure. This is in contrast to the majority of papers, where many methods have been tried, but only the best one was treated as the final solution.

2. Selection of gene ranking methods

2.1. Problem formulation. The main task of selection is to discover the expression activity of the genes that are associated with the type of cancer considered. The difficulty is that an expression array is a huge matrix with a large number of columns (n) representing genes or gene sequences (further called features) and a small number of rows (m) corresponding to the succeeding individuals. It is known that the level of expression of some genes is characteristic for the specific type of illness and is similar for many patients suffering from this illness (De Rinaldis, 2007; Guyon *et al*., 2002; Hewett and Kijsanayothin, 2008).

When visualising the gene expression matrix of all genes for two groups of patients (the group suffering from a particular type of cancer against the healthy class), we cannot observe any visible division of the image into two groups that are associated with these two classes of patients. Thus, an important problem is to select a limited number of the top rank genes that are most representative and discriminative for both the classes. After a selection is made, we should see a clear division of data into two separated graphical regions corresponding to both classes.

In this paper we will limit our consideration of gene ranking to prostate tumor (PRT) data, taken from the benchmark gene data base of Vanderbilt University (Vanderbilt, 2002), represented by two classes.

2.2. Gene ranking methods. Gene ranking is a specific form of the general process of feature selection, in which each gene or gene sequence expression is treated as a feature. Out of the tremendous number of different methods of feature selection, we will limit our consideration to just a few of them: Kolmogorov–Smirnov and Wilcoxon–Mann–Whitney statistical tests, the correlation analysis, a Fisher measure based on the analysis of distribution of centres and variances of the clusters, as well as gene ranking using a linear support vector machine. This choice of methods belongs to different approaches to the feature selection and provides a different point of view on the problem.

The results of the separated selection processes are combined together to perform the second step of selection that leads to the final optimal ranking of genes. The main stress of the analysis will be directed toward the measure of the cluster quality and then to the application of the gene selection results in the classification of the data.

To get reliable results of selection, each of the individual methods presented above will be repeated many times on randomly chosen samples of the original data set. In our experiments we performed each selection method 10 times by using 90% of the randomly selected rows of the data set. The results of these trials will be integrated into a final ranking by applying two further processes:

- Ranking by frequency: the features are arranged according to their frequency of the appearance within the first 100 best features. This is a sufficiently large number to represent the most relevant genes that characterize the classes. The feature which appeared the highest number of times within the 100 best in all performed experiments is treated as most important. In this way, a natural ranking of features was created.

- Ranking by positions: this ranking is based on the sum of the positions that the particular feature took in all experiments of selection. The best feature is the one of the smallest value of this total sum. To suppress the problem of a very large size of vectors, we took into account only the first 100 features. The feature which was absent in the list of the best 100 was punished by assigning the position of 101.

These two methods of fusion will be compared on the example of benchmark data concerning the prostate tumor (Vanderbilt gene data base).

2.2.1. Fisher discriminant measure. In the Fisher method we form the discriminant measure $S_{12}(f)$ of the feature f to recognize Class 1 from Class 2 in the following form (Golub *et al.*, 1999):

$$S_{12}(f) = \frac{|m_1(f) - m_2(f)|}{\sigma_1(f) - \sigma_2(f)}, \tag{1}$$

where $m_k(f) = E\{f|k\}$ is the mean value of the features for data records that form the k-th class $(k = 1, 2)$, while σ_1 and σ_2 denote the standard deviations of the feature in both classes, respectively. Large values of $S_{12}(f)$ indicate a good separation ability of the feature f for recognition of these two classes. A small value means that clusters of both the classes are close to each other and the data samples are widely distributed. Such a feature does not represent a good discriminative property. The results generated by this method will be denoted shortly by FISH.

2.2.2. Correlation of gene expression with classes. The method frequently used in assessing the discriminative power of the candidate feature f for the recognition of the particular class among K classes ($K = 2$ in our case) is the correlation of this feature with the class (Duda *et al.*, 2003).

Let us denote by $\mathbf{d} \in \{1, \ldots, k\}^m$ the vector of class membership, by $m(f) = E\{f\}$ the mean value of the feature f in the whole set of data, and by $\mathrm{var}(f) = E\{(f - m(f))^2\}$ the variance of the feature f for the whole data. The correlation measure of the feature f with the vector $\mathbf{d}$ representing classes is defined through the covariance $\mathrm{cov}(f, \mathbf{d})$ and can be expressed in the form (Wiliński and Osowski, 2012; Duda *et al.*, 2003)

$$S(f) = \frac{\sum_{k=1}^{K} P_k^2 (m_k(f) - m(f))^2}{\mathrm{var}(f) \sum_{k=1}^{K} P_k (1 - P_k)}, \tag{2}$$

with P_k denoting the probability of the k-th class. After calculating this measure for all features, we can arrange them in a decreasing order, from the highest to the smallest discriminative value. With this, we get an automatic ranking of the features. This method of feature ranking will be denoted by COR.

2.2.3. Statistical hypothesis tests. In these tests the feature f is treated as a statistical variable of some distribution related to the type of data (Sprent and Smeeton, 2007). The feature of a good discriminating ability should have a similar distribution for the group of observations belonging to the same class, and different in the case of different classes. We applied here two tests: the Wilcoxon–Mann–Whitney (WMW) and the Kolmogorov–Smirnov (KS) (Sprent and Smeeton, 2007; Matlab, 2012)

The WMW test applied here is based on the ranks of the particular patterns. In this method we arrange the population of X random variables (Class 1) and the population of Y random variables (Class 2) together in the increasing order of magnitude (ordinal fashion). The arrangement where most of the Y's are greater than most of the X's or vice versa would be evidence against random mixing. This would tend to discredit the null hypothesis of identical distributions. In the WMW test we estimate the probability P denoting the degree of similarity of both the sequences. The higher this probability, the more similar the two populations. This method of feature ranking will be denoted shortly by WMW.

The KS test checks the null hypothesis that the samples of both the classes are drawn from the same distribution at the desired significance level (default = 0.05). The Matlab function $kstest2$ (Matlab, 2012) implements this test and determines the distance between the cumulative distribution functions of the data belonging to two compared classes. This distance is regarded as a basis for defining the statistical measures of difference between both the populations. Let us denote by $F_{\mathbf{x}_1}(x)$ and $F_{\mathbf{x}_2}(x)$ the empirical cumulative distributions of two populations of feature f which respectively represent Class 1 (vector $\mathbf{x}_1$) and Class 2 (vector $\mathbf{x}_2$), $([\mathbf{x}_1, \mathbf{x}_2]^T \in \mathbb{R}^m)$. Using the Kolmogorov–Smirnov test, we defined two different discriminative measures (Wiliński and Osowski, 2012) (where $\mathbf{D} = \{x_1, \ldots, x_s\}$, for a given set of $s \in \mathbb{N}$):

- Additive Kolmogorov–Smirnov measure (AKS),

$$S_{12}(f) = \sum_{x \in \mathbf{D}} |F_{\mathbf{x}_1}(x) - F_{\mathbf{x}_2}(x)|. \tag{3}$$

- Scaled maximum Kolmogorov–Smirnov measure (SKS),

$$S_{12}(f) = a(f) \cdot \sup_{x \in \mathbf{D}} |F_{\mathbf{x}_1}(x) - F_{\mathbf{x}_2}(x)|, \tag{4}$$

where the scaling coefficient $a(f)$ is defined as follows:

$$a(f) = \frac{|\mathbb{E}(\mathbf{x}_1) - \mathbb{E}(\mathbf{x}_2)|}{\sigma(\mathbf{x}_1) + \sigma(\mathbf{x}_2)}. \tag{5}$$

High values of these measures indicate that the distributions of points belonging to two classes are different (do not belong to the same population of samples). Such a feature is beneficial.

2.2.4. Application of linear SVM. A support vector machine of a linear kernel is another good tool for feature selection (Guyon *et al.*, 2002). In this work it was applied in two different modes. The first one is its application in one-input arrangement by using only one feature at a time (Vert, 2007). We train equally as many networks as the number of features. Each feature is the only single input signal to the SVM. The discriminative power of a single feature is characterized by the value of the class recognition error, provided by a one-dimensional linear SVM, trained to classify all learning samples using only one feature at a time as the input signal. The results of classification are used for assessing the feature. The smaller this error, the better the class discriminative ability of the feature. The discriminative value of the particular feature f is then defined as

$$S_{12}(f) = \frac{N_r(f)}{N_a}, \tag{6}$$

where $N_r(f)$ represents the number of correctly recognized samples at application of feature f, while N_a is the total number of samples under recognition. On the basis of this value, the ranking of features is made. This method of the feature ranking will be referred to as 1SVM.

The application of the multi-input linear SVM, called also SVM Recursive Feedback Elimination (SVM-RFE), was proposed by Guyon *et al.* (2002). This method found some modifications (Huang and Kecman, 2005; Yang, 2011) used in various research areas. The discrimination power of each feature is tested in the presence of the whole set of features, used as the excitation to the linear kernel SVM. The decision on the membership of the n-dimensional input vector **x** to the particular class relies on the sign of the value of the linear function $y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$ with the weight vector $\mathbf{w} \in \mathbb{R}^n$ and the bias $b \in \mathbb{R}$ dependent on the linear combination of the training patterns forming the support vectors. The absolute values of the weights of the vector **w** produce a feature ranking

$$S_{12}(f) = |w_f|, \tag{7}$$

where w_f is the value of weight joining the input of feature *f* with SVM network. The procedure of the feature elimination is repeated many times by training the SVM classifiers. The application of the shorter and shorter feature vectors forms the input signals. The procedure is ended when we get the state in which there are no weights of significantly smaller magnitudes, or when we achieve the vector of appropriate (desired) size. This method will be shortly denoted by MSVM.

3. Assessment of cluster quality

The samples representing each class form natural clusters of the data. The quality of these clusters depends on the size and composition of vectors representing the samples. In this section we introduce different measures of cluster quality.

3.1. Class oriented measures of cluster quality. In this approach the records of data belong to different classes that are treated as samples coming from two clusters representing the cancer and non-cancer cases, respectively. The clusters are represented by their cluster centres $\mathbf{c}_A$ and $\mathbf{c}_B$. The quality of the clusters may be characterized by different measures of quality (Sabo, 2014). The first one is the minimum distance d_{min} (in Euclidean metric) between closest two points belonging to different clusters. Denote by A and B the clusters and by n_A and n_B their sizes, respectively. Then we get

$$d_{min}(A,B) = \min_{i,j} d(\mathbf{x_i}, \mathbf{x_j}), \tag{8}$$

for $i = 1, 2, \ldots, n_A$ and $j = 1, 2, \ldots, n_B$. The other measure is the centroidal distance d_{cent}, defined between the centers of both clusters,

$$d_{\text{cent}}(A,B) = d(\mathbf{c_A}, \mathbf{c_B}). \tag{9}$$

A higher distance guarantees better separation of the clusters. An important measure is also the dispersion of the clusters. It is defined as the average of the distances among all points belonging to the same cluster A and B,

$$\sigma(i) = \frac{1}{m} \sum_{\mathbf{x}_k \in i, \mathbf{x}_j \in i} d^2(\mathbf{x}_k, \mathbf{x}_j), \tag{10}$$

for $i = A, B$, where $m = n_i(n_i - 1)/2$. The total dispersion of the clustered data may be treated as the sum of cluster dispersions.

3.2. Quality measures at unsupervised clustering. In this approach the original set of high dimensional data is clustered according to the distances into two groups without taking into account their class membership. In this way the clusters will contain the representatives of both classes. The simplest way to such clustering is the application of the K-means algorithm (Tan *et al.*, 2006; Haykin, 1999). After performing such clustering we can assess the quality of clusters by using some specific measures. Let us denote by n_{ij} the number of representatives of the j-th class in i-th cluster (i, j = 1, 2), by n_i the number of elements in the i-th cluster with $n = n_1 + n_2$, and by $p_{ij} = n_{ij}/n_i$ the percentage of the elements of j-th class in the i-th cluster. The purity of the i-th cluster is defined as (Tan *et al.*, 2006)

$$p_i = \max_j(p_{ij}). \tag{11}$$

After dividing the data set into K clusters ($K = 2$ in our case), the total purity of the clustered space is defined as

$$p = \sum_{i=1}^{K} p_i \frac{n_i}{n}. \tag{12}$$

The higher this measure, the more uniform the class composition of the clusters and the better their predictive properties. Its high value means the highest purity of the clustered space (each cluster represents a separate class).

4. Results of gene ranking and fusion

The methods of feature selection presented above were applied simultaneously to the fusion of the most important genes corresponding to the division of a prostate cancer data (Vanderbilt gene data base (Vanderbilt, 2002)) into two classes. The first class of data (52 records) represents the prostate tumour and the second (50 records) reflects the reference class which represents the patients with no prostate cancer. Each vector of gene expression contains 10509 elements. The value ranges of these elements change from gene to gene. The highest and the lowest range were found for the gene numbers 9737 and 3281. Their values were [12, 15753] and [−7, 10.1], respectively. The set of experimental patterns $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{102} \in \mathbb{R}^{10509}$ arranged in the form of row vectors presents the values of gene expressions with known class labels $d_1, d_2, \dots, d_{102} \in \{-1, 1\}$ representing either first or second class cases.

4.1. General procedure of gene fusion. To get the most reliable results, we repeated 10 times the experiments of ranking according to the procedure described in Section 2.1. The statistics of the selection of each gene in these 10 trials were made. In this way we are able to note the positions of the genes in the ranking and the frequency of their appearance among the first 100 best. All genes beyond the first 100 are ignored in these statistics. Our aim is to select only a few genes that best correlate with the class. In each selection run we take into the consideration only the 100 best genes and form a limited, but sufficiently large population of them for the future fusion procedure.

To find the best set of genes, we applied two fusion procedures: by frequency and by positions. This allows us to fuse the results of the individual trials into the final outcome. The two systems were applied within the particular methods of selection and then for integration of all methods into the final result of the gene position. The results of such selection were compared with a random selection of 100 genes (without ranking). In this way, we provided the same size of representative vectors for the comparison purpose. The random choice of the genes was also repeated 10 times to average the results.

4.2. Results of clustering class oriented data. In the first point of our analysis we characterize the distribution of the available data that belong to the particular classes. In this analysis we assume that the data of each class form one cluster. The number of clusters is identical to that of classes (two in our experiments). The representation of the original data by the selected genes should result in different distributions of the clusters. The clusterization results which are generated by the different selection methods are compared on the basis of the distances ($d_{\min}$ and d_{cent}) between both clusters of data as well as the dispersion measure of the clusters representing the cancer cases (A) and the reference (healthy) class (B). In this part of experiments the data belong to the clusters that are associated with the exactly known membership to the classes (the clusters are class uniform). The numerical results of the experiments are presented in Table 1. These include also the statistical results in the form of mean value ± standard deviation that was obtained at 10 trials and a different random composition of 100 genes.

In the table we see that the selection results are better than the random choice of genes. The presented methodological approaches to feature selection increases the distances between clusters in comparison with the random choice of data, but the differences are not very significant. The dispersion measures of the clusters were also reduced.

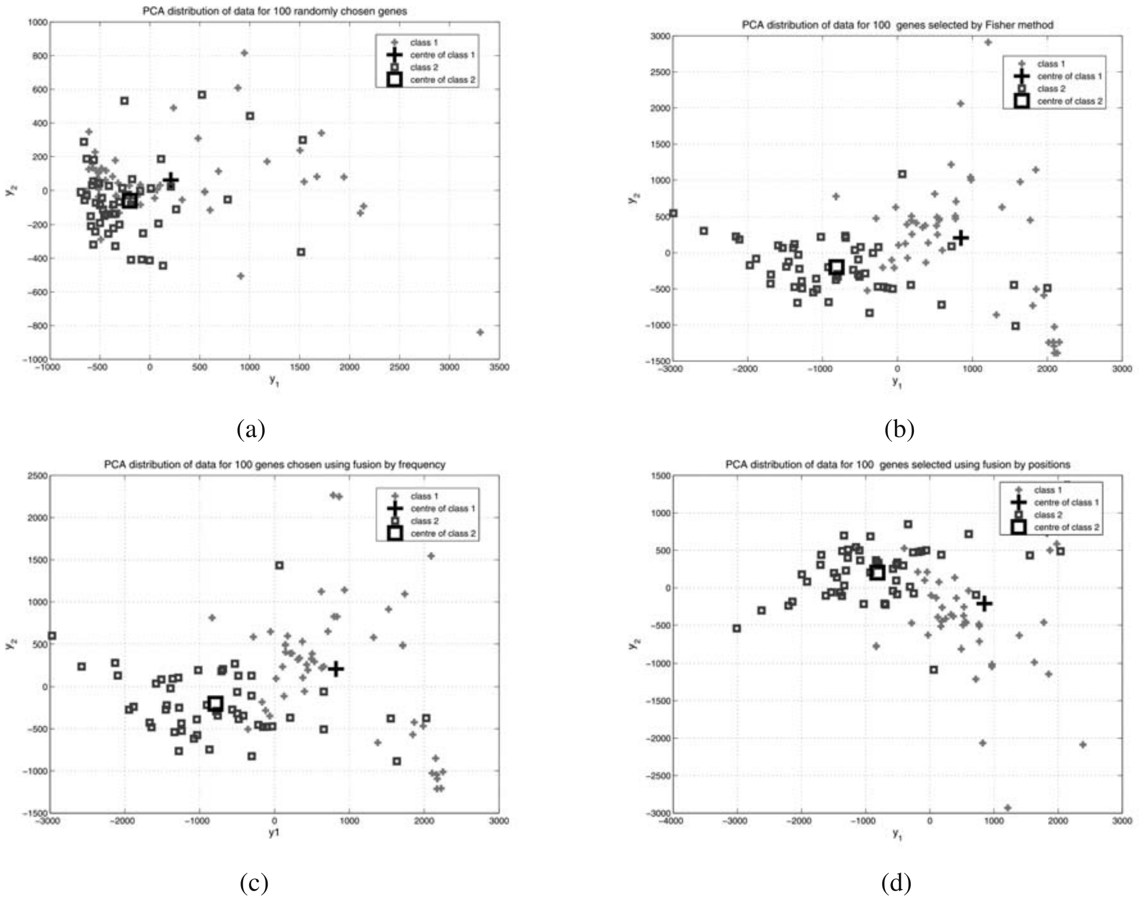
We provided the visualization of the 100 dimensional data by mapping them into a two-dimensional coordinate system to get better insight into the results of selection. We employed two most important components of linear Principal Component Analysis (PCA) (Haykin, 1999). The PCA transformation was performed on the covariance matrix formed on the basis of the available data. The PCA mapping results related to the random choice of 100 genes, use of the Fisher method, the final fusion system by frequency and positions applied to the 100 best genes are presented in Fig. 1. These results correspond to the sample run of procedures.

The visualization of results definitely confirmed the good performance of our selection methods. We obtained a significant improvement of data location applying only the Fisher method. When the genes are randomly chosen, the representatives of both classes are strongly interlaced with each other. There is no visible border between both the classes. The application of gene ranking and selection methods caused that the distribution of data changed completely. Both classes are well separated now and only a few representatives are located in a wrong area.

The discrimination ability of the selected genes is depicted in Fig. 2. It presents the mapping of the gene expression values into colors for both the classes considered. The jet color map was used in this mapping. Figure 2(a) represents the randomly selected set of 10 genes and Fig. 2(b) the results of representing data by

Table 1. Values of quality measures of class uniform clusters for different methods applied to selection of the 100 best genes.

Method	$d_{\min}$	d_{centr}	$\sigma(A)$	$\sigma(B)$
Random	3.84± 0.38	9.63±3.26	12.52±0.47	12.76±0.39
FISH	4.05±0.28	11.55±3.24	11.02±0.42	11.71±0.36
COR	4.30±0.29	11.86±3.19	10.69±0.37	10.08±0.40
WMW	3.95±0.39	9.91±3.27	10.58±0.39	11.08±0.42
AKS	4.37±0.40	11.68±3.25	12.29±0.52	10.51±0.44
SKS	4.03±0.46	9.76±3.24	11.07±0.49	11.98±0.41
1SVM	4.73±0.27	11.23±3.19	11.30±0.39	11.47±0.38
MSVM	6.40±0.24	10.59±3.17	12.13±0.37	12.24±0.33
Fusion by frequency	4.06	11.55	10.51	10.34
Fusion by position	4.45	11.58	10.59	10.35



(a)

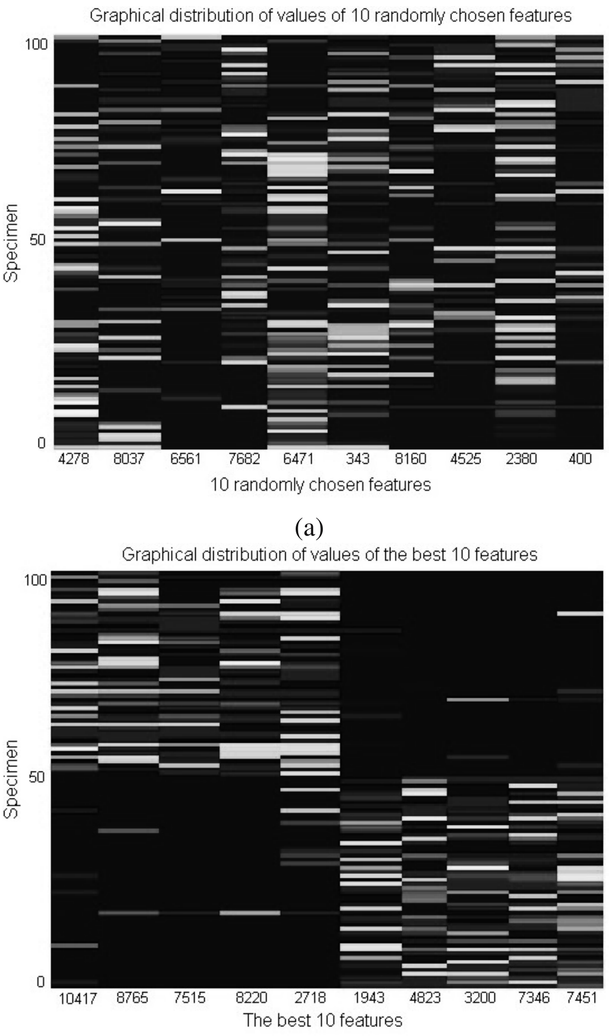
(b)

(c)

(d)

Fig. 1. PCA distribution of data at random choice of 100 genes (a), 100 best genes selected by the Fisher method (b), 100 best genes selected by using the fusion system based on the frequency (c), 100 best genes selected by using the fusion system based on the positions (d).

the 10 best genes selected using the fusion system based on positions. The horizontal axis represents the selected genes identified by their numbers in the original notation. They are presented in sequence according to their position in ranking. In the case of randomly selected genes, there is no visible border between the patterns representing the first 52 specimen of the prostate cancer and the other 50 healthy specimen. The situation radically changed after the application of the developed selection procedure. Now this border is easily recognized, because the pattern of levels of gene expression belonging to opposite classes is significantly different.



(a)

(b)

Fig. 2. Graphical representation of data for the application of 10 genes (features): genes chosen randomly (a), genes selected by our method (b).

We identified the composition of the genes with the highest correlation in relation to prostate cancer. Both the fusion systems (although integration was done on the outcomes of different principles of basic rankings) produced consistent results. Among the 10 most important genes, 9 were the same (10417, 1943, 8765, 7346, 7515, 2718, 7811, 3200, 8220). These genes, together with the next in ranking, will form the basic input information for the classification task.

4.3. Results of unsupervised clustering of data. A very interesting approach to gene selection is unsupervised clusterization of the data (without taking into account their class membership) in the process of self-organization. We performed the simplest clustering in the form of the K-means method for the data set represented by all genes (the vector size 10509) and by shorter vectors formed by the selected genes. In the experiments we changed the population size of the most important genes selected by using different methods. The aim was to find an optimal size of the vector (the number of the most important genes) resulting in the best purity of the clusters. Figure 3 presents the total purity curve as the function of the population of the selected genes after the fusion of the results of all selection methods. The highest total purity value corresponds to the optimal number of features (genes).

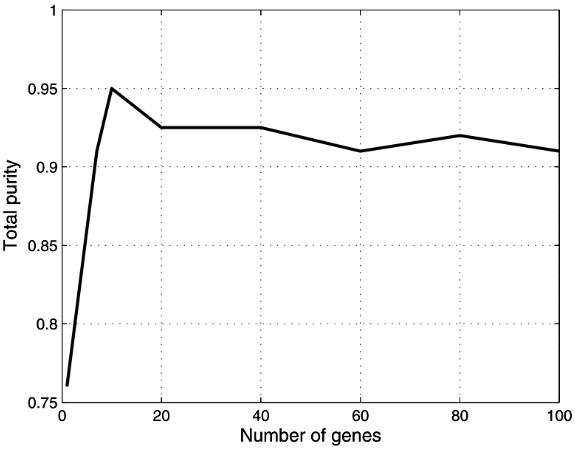


Fig. 3. Dependence of the total purity of clusters on the number of selected genes.

On the basis of this curve, we selected the optimal number of the most important genes, which equals 10. Tables 2 and 3 present the numerical results of the chosen quality measures of the clusters by applying different ranking methods. The tables show the class distribution of data belonging to both classes and clusters, the purity of both clusters as well as the total purity of the clustered space.

As we can see, the selection and reduction of features significantly improved the total purity of the clusterization. The purity of the clusters increased from the total mean value of 0.67 for all genes (no reduction), to 0.95 in the best case of reduction. We observed that different methods had greater or smaller or lower impact on the results. Also, the integration system of selection results carried out within the individual method and between different methods plays an important role. The mean value of purity over all the methods within the first system of selection was equal to 0.843, while for the second one 0.865. The integration of all selection results improved this value to 0.94 (the first integration system) and 0.95 (the second integration system).

It is interesting to compare the composition of the

Table 4. Composition of the set of the 10 best genes selected by the Fisher method and by the proposed fusion by positions, arranged in the original order of their importance.

Fisher	4823	7451	3200	8220	7652	8765	8468	5461	6640	7515
Fusion by position	10417	8765	7515	7346	2718	1943	4823	3200	8220	7451

Table 2. Integrated results of clusterization of gene expression microarray data for all genes and for 10 genes selected using the fusion system based on frequency.

Method	Cluster	Class 1	Class 2	Purity	Total purity
No selection (all genes)	A	5	16	0.76	0.67
	B	47	34	0.58	
FISH	A	8	46	0.85	0.89
	B	44	4	0.92	
COR	A	6	42	0.88	0.87
	B	46	8	0.85	
WMW	A	4	48	0.92	0.92
	B	48	2	0.96	
AKS	A	6	25	0.81	0.70
	B	44	25	0.64	
MKS	A	6	26	0.81	0.70
	B	46	24	0.66	
1SVM	A	6	46	0.88	0.90
	B	46	4	0.92	
MSVM	A	10	50	0.83	0.92
	B	42	0	1.00	
Fusion by frequency	A	5	49	0.91	**0.94**
	B	47	1	0.98	

Table 3. Integrated results of clusterization of gene expression microarray data for all genes and for the 10 best genes selected using the fusion system based on positions.

Method	Cluster	Class 1	Class 2	Purity	Total purity
No selection (all genes)	A	5	16	0.76	0.67
	B	47	34	0.58	
FISH	A	5	48	0.92	0.93
	B	47	2	0.96	
COR	A	6	49	0.89	0.92
	B	46	1	0.98	
WMW	A	11	49	0.82	0.90
	B	41	1	0.98	
AKS	A	6	19	0.76	0.70
	B	46	31	0.60	
MKS	A	8	25	0.76	0.68
	B	44	25	0.64	
1SVM	A	6	49	0.89	0.92
	B	46	1	0.98	
MSVM	A	5	49	0.91	0.92
	B	47	1	0.98	
Fusion by positions	A	4	49	0.92	**0.95**
	B	48	1	0.98	

selected genes found by the best final fusion system corresponding with the last row of Table 3 and the composition of genes selected by the best individual method of only slightly lower purity (for example, the Fisher method—the second row of the results of Table 3). Table 4 depicts the identification of the 10 best genes selected by both the methods (Fisher and fusion system by positions) arranged in the original order of their importance. They represent the fused results of the application of all selection methods at all trials. Among the 10 best genes we found 6 which appeared simultaneously in both the methods (7451, 3200, 4823, 8220, 8765, 7515). The other four were different in both the cases. This means that the integration process significantly influenced the selection process.

If we transpose the results directly from purity to class recognition (the sample falling into a particular cluster is automatically associated with the majority class), we can observe a significant improvement of recognition accuracy after the selection process. In the case of all genes taking part in clusterization, there were 39 misclassified cases (38.2% of the relative error). In the best case of integration, the total number of the misclassified samples was equal to 5, which corresponds to the relative error equal to 4.9%. These results indicate the upper limit of classification accuracy in the application of clusterization as classification tool. Application of more advanced classifiers should lead to an improvement of the classification results. In the next section we will present the results of applying the SVM as the classifier, performing the task of gene pattern recognition.

5. Classification of prostate tumor data

5.1. SVM classifier. The last step of the developed gene ranking approach is performing the recognition of data into two classes in the application of the selected genes. This classification step used for the final recognition of the tumor from the healthy cases was performed by applying once again support vector machine classifiers. However, this time the the polynomial kernel was used (Scholkopf and Smola, 2002). This particular kernel was chosen after introductory experiments involving other types of kernel (linear, Gaussian, sigmoidal). The polynomial kernel was found to have the highest accuracy of classification. In these experiments we used a solver developed and implemented in Matlab by Fan *et al*. (2005) as well as Chang and Lin

(2011). This algorithm is based on modified sequential programming.

The support vector machine is a solution of a feedforward structure with one hidden layer. The application of a special learning method leads to a quadratic program with linear constraints and one well-defined global minimum. Basically, the SVM is a one output linear machine, working in a high dimensional feature space formed by nonlinear mapping of the original n-dimensional input vector $\mathbf{x}$ into the k-dimensional feature space. The nonlinear vector function $\boldsymbol{\varphi}(\mathbf{x})$ is arranged in the form of the kernel $K(\mathbf{x}, \mathbf{x}_i) = \boldsymbol{\varphi}^T(\mathbf{x})\boldsymbol{\varphi}(\mathbf{x}_i)$. More details about the learning phase of this network can be found, e.g., in the book of Scholkopf and Smola (2002).

The regularization constant C plays an important role in the learning phase. It balances the complexity of the network, characterized by the values of weights and the error of the classification, over the learning data. The low value of C means smaller significance of the learning errors in the adaptation stage and leads to a smaller network size of a higher separation margin. Increasing the value of C leads to more complex structures with a smaller separation margin but better performance over the learning data. For normalized input signals, the value of C is usually much bigger than 1 (a typical value is in the range from 100 to 1000). In practice, we adjusted it by a trial-and-error procedure using a small percentage of validation data extracted from the available learning data. In the same way we adjusted the proper value of the degree q of the polynomial kernel function. The parameters C and q were adjusted simultaneously by trying different combinations of their values in the introductory stage of the experiments.

5.2. Results of classification. In this part of research we investigate the accuracy of cancer recognition from the normal cases on the basis of gene expression microarray data at different selections of genes. Well selected genes should result in an improvement of cancer recognition accuracy. The application of all genes together in the recognition is not suggested. The number of records corresponding to the individuals is too small in comparison with the number of genes in order to get the reliable results.

In this research we compare the results of cancer recognition applying a limited number of genes selected by the ensemble of all methods. The quantity of the genes used comes from the results of data clusterization. In these experiments we applied the population of genes which corresponded to the highest purity of clusters. According to the results of Fig. 3, it was equal to 10. For the comparison purposes, we performed additional classification trials with the same number of genes, but selected randomly from the whole population of the available genes.

To get the objective results, we applied a 10-fold cross validation approach, repeating the learning/testing experiments 10 times, changing each time the testing part of the data not used in learning (Haykin, 1999). As the classifier, we used the SVM network of a third-order polynomial kernel. On the basis of these experiments, the percentage error of testing was calculated.

The particular results of the experiments will be depicted in the form of a confusion matrix representing the average results of the cross validation procedure for the testing data. The rows represent the real class membership of the data and the columns—the results of the classification. The diagonal entries ($i = j$) represent the number of properly recognized classes. Each entry outside the diagonal represents the number of misclassified cases. The entry in the (i, j)-th position of the matrix means a false assignment of the i-th class to the j-th one. Table 5 presents the best results of classification corresponding to the fusion system based on positions in the application of the 10 best genes.

As can be seen, the recognition results are nearly perfect. Only few representatives of both classes were misclassified. The relative error of the final classification results was reduced to 1.96%.

To compare the importance of gene ranking, we repeated these 10 cross validation experiments by randomly choosing the composition of 10 genes. The obtained results in the form of a confusion matrix are depicted in Table 6. This time the misclassification rate is very high (32.3%).

From these results it is evident that the application of the highest ranked genes in this representation of samples provides the highest accuracy (the least relative error) in all experiments.

In the medical experiments the accuracy is usually only one measure of the quality. This measure treats every class as equally important, hence it is not sufficient to assess the method in an objective way. Additional aspects of the results associated with the importance of the cancer

Table 5. Confusion matrix of the classification results for the application of the 10 best genes selected by the fusion system based on positions.

	Class 1	Class 2
Class 1	51	1
Class 2	1	49

Table 6. Confusion matrix of the classification results by applying 10 randomly selected genes.

	Class 1	Class 2
Class 1	39	13
Class 2	20	30

class recognition (called here True Positive, TP) from the healthy class (True Negative, TN) should be taken into consideration. By the symbol FN we understand the number of cancer specimen falsely recognized as healthy and by FP—healthy cases recognized as cancerous. On the basis of this notation we define here four quality measures of classification.

The most important is the True Positive Rate (TPR), called also sensitivity, defined as the fraction of all positive examples predicted correctly by the classifier $TPR = TP/(TP + FN)$. Similarly, the True Negative Rate (TNR), called specificity, is defined as the fraction of negative examples predicted correctly by the classifier $TNR = TN/(TN + FP)$. The next used measure is the False Alarm (FA) rate defined as the ratio of negative class cases recognized by the classifier as positive $FA = FP/(FP + TN)$. The last one of very high importance is the False Negative Rate (FNR), defined as $FNR = FN/(TP + FN)$. Table 7 presents numerical results which concern these quality measures when applied to the recognition of a prostate cancer specimen from healthy ones. They correspond to the use of the 10 highest ranked genes selected by the fusion method based on positions. The sensitivity and specificity of the proposed classification method achieve very high values (close to 0.98). On the other hand, the false alarm rate is very low. These results confirm high efficiency of the developed approach.

6. Conclusions

We analyzed different methods of gene selection for the recognition of the prostate tumor from healthy cases using the gene expression array. We applied seven chosen measures of the gene class discrimination ability for this pattern recognition problem. The methods considered represent different approaches to feature selection and apply the correlation of the gene with the classes, the Fisher measure, statistical hypotheses as well as the application of the classification ability of the linear SVM.

We proposed a two-step procedure of ranking the most important genes associated with the classes. In the first step, each method acts in an independent way, estimating its specific value of the discriminative measure for each gene. On the basis of these values the genes are ranked from the most to the least significant. To avoid the problem of scarcity of data in comparison with the number of genes, the ranking procedure was repeated many times using different (randomly chosen) sets of data records. Next, we determined the positions of the genes and the frequency of their selection from the best set in these cross-validation runs. On the basis of these results, we finally create the first step of the gene ranking.

Table 7. Values of quality measures for recognition of prostate cancer cases (class $+$) and healthy ones (class $-$) for application of the 10 highest rank genes after fusion by positions of all methods of selection.

TPR	TNR	FNR	FA
0.981	0.980	0.019	0.020

In the second stage of the ranking procedure, we compared the contents of the high rank genes created by the different methods. Through these results we are able to identify the highest rank genes chosen by most of the selection methods. They form the final set of the best genes (the genes correlated with cancer in the highest way). Two approaches to fusion were investigated. The first one considered the frequency of occurence of the particular feature in the best set in all rankings, and the second one—the position of each feature in the rankings.

The quality of the ranking procedure has been checked in different ways. First, we checked the visual form of the expression level of the selected genes for both classes. The other visualization form applied the mapping of data by using principal component analysis and presenting the transformed data in a 2-D space formed by the two most important components. The numerical analysis of the clusters formed by 2-class data was also important. We defined and investigated various measures of cluster quality by considering the selected genes. On the basis of this analysis, we were able to determine the optimal number of genes used in the final classification step.

In order to verify the results, we applied the SVM classifier with the polynomial kernel, which is responsible for recognition of tumor data from non-tumor cases. In this step, we use the top ranked genes as the input information in the classifier to obtain the best possible recognition of the classes which the succeeding records belonged to.

Observe that the testing data did not take part in the learning phase. The results of classification on this set of data were of very high quality, proving the efficiency of the proposed gene ranking methods. The average accuracy of class recognition in the prostate tumor problem, calculated on the basis of 10 cross-validation runs, was equal to 98.04%. This result is compared favorably with the most recent results for similar data of prostate tumor (Wang and Gotoh, 2010), processed using rough set theory. The declared average accuracy on the PRT data in this paper was equal to 90.98%, while the highest single run achieved the peak value of accuracy below 98%. Our approach guaranteed high quality and stable results.

References

Baldi, P. and Long, A. (2001). A Bayesian framework for the analysis of microarray expression data: Regularized t-test and statistical inference of gene changes, *Bioinformatics* **17**(4): 509–519.

Chang, C.-C. and Lin, C.-J. (2011). LibSVM: A library for support vector machines, *ACM Transactions on Intelligent Systems and Technology* **1**(27): 1–27.

De Rinaldis, E. (2007). *DNA Microarrays: Current Applications*, Horizon Scientific Press, Norfolk.

Duda, R., Hart, P. and Stork, P. (2003). *Pattern Classification and Scene Analysis*, John Wiley, New York, NY.

Eisen, M., Spellman, P. and Brown, P. (1998). Cluster analysis and display of genome wide expression patterns, *Proceedings of the National Academy of Sciences* **95**(25): 14863–14868.

Fan, R.-E., Chen, P.-H. and Lin, C.-J. (2005). Working set selection using second order information for training SVM, *Journal of Machine Learning Research* **6**(12): 1889–1918.

Furey, T., Cristianini, N., Duffy, N., Bednarski, D., Schummer, M. and Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data, *Bioinformatics* **16**(10): 906–914.

Golub, T., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A. and Bloomfield, C.D. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, *Science* **286**(5439): 531–537.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection, *Journal of Machine Learning Research* **3**(3): 1158–1182.

Guyon, I., Weston, A., Barnhill, S. and Vapnik, V. (2002). Gene selection for cancer classification using SVM, *Machine Learning* **46**(1–3): 389–422.

Haykin, S. (1999). *Neural Networks. A Comprehensive Foundation, 2nd Edition*, Prentice-Hall, Englewood Cliffs, NJ.

Herrero, J., Valencia, A. and Dopazon, A. (2001). A hierarchical unsupervised growing neural network for clustering gene expression patterns, *Bioinformatics* **17**(2): 126–136.

Hewett, R. and Kijsanayothin, P. (2008). Tumor classification ranking from microarray data, *BMC Genomics* **9**(2): 1–11.

Huang, T.M. and Kecman, V. (2005). Gene extraction for cancer diagnosis by support vector machines—an improvement, *Artificial Intelligence in Medicine* **9**(35): 185–194.

Huang, X. and Pan, W. (2003). Linear regression and two-class classification with gene expression data, *Bioinformatics* **19**(16): 2072–2078.

Makinaci, M. (2007). Support vector machine approach for classification of cancerous prostate regions, *World Academy of Science, Engineering and Technology* **1**(7): 166–169.

Matlab (2012). *Matlab User Manual—Statistics Toolbox*, MathWorks, Natic.

Mitsubayashi, H., Aso, S., Nagashima, T. and Okada, Y. (2008). Accurate and robust gene selection for desease classification using a simple statistics, *Biomedical Informatics* **3**(2): 68–71.

Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J., Poggio, T., Gerald, W., Loda, M., Lander, E. and Golub, T. (2001). Multiclass cancer diagnosis using tumor gene expression signatures, *Proceedings of the National Academy of Sciences* **98**(26): 15149–15154.

Sabo, K. (2014). Center-based l_1-clustering method, *International Journal of Applied Mathematics and Computer Science* **24**(1): 151–163, DOI: 10.2478/amcs-2014-0012.

Scholkopf, B. and Smola, A. (2002). *Learning with Kernels*, MIT Press, Cambridge, MA.

Sprent, P. and Smeeton, N. (2007). *Applied Nonparametric Statistical Methods*, Chapman and Hall-CRC, Boca Raton, FL.

Świniarski, R.W. (2001). Rough sets methods in feature reduction and classification, *International Journal of Applied Mathematics and Computer Science* **11**(3): 565–582.

Tan, P.N., Steinbach, M. and Kumar, V. (2006). *Introduction to Data Mining*, Pearson Education, Boston, MA.

Vanderbilt (2002). *Data base of prostate cancer*, Vanderbilt University, `http://discover1.mc.vanderbilt.edu/discover/public/mcsvm`.

Vert, J. (2007). Kernel methods in genomics and computational biology, *in* G. Camps-Valls, J.L. Rojo-Alvarez and M. Martinez-Ramon (Eds.), *Kernel Methods in Bioengineering, Signal and Image Processing*, Idea Group, London, pp. 42–64.

Wang, X. and Gotoh, O. (2009). Cancer classification using single genes, *Genom Informatics* **23**(1): 179–188.

Wang, X. and Gotoh, O. (2010). A robust gene selection method for microarray-based cancer classification, *Cancer Informatics* **9**(2): 15–30.

Wiliński, A. and Osowski, S. (2012). Ensemble of data mining methods for gene ranking, *Bulletin of the Polish Academy of Sciences* **60**(3): 461–471.

Woolf, P.J. and Wang, Y. (2000). A fuzzy logic approach to analyzing gene expression data, *Physiological Genomics* **3**(1): 9–15.

Yang, F. (2011). Robust feature selection for microarray data based on multicriterion fusion, *IEEE Transactions on Computational Biology and Bioinformatics* **8**(4): 1080–1092.

Michał Muszyński was born in Poland in 1988. He received the M.Sc. degree in electrical engineering from the Warsaw University of Technology, Warsaw, Poland, in 2012. Currently, he is a Ph.D. student at the Faculty of Electrical Engineering of the same university. His research and teaching interests are in the areas of neural networks and their applications in biomedical engineering.

Stanisław Osowski was born in Poland in 1948. He received the M.Sc., Ph.D., and D.Sc. degrees from the Warsaw University of Technology, Warsaw, Poland, in 1972, 1975, and 1981, respectively, all in electrical engineering. Currently, he is a professor of electrical engineering at the Institute of the Theory of Electrical Engineering and Electrical Measurements of the same university. His research and teaching interests are in the areas of neural networks, optimization techniques and their application in various areas of biomedical engineering. He is an author or co-author of more than 200 scientific papers and ten books.

Received: 17 September 2013
Revised: 15 January 2014
Re-revised: 9 March 2014