BULLETIN OF THE POLISH ACADEMY OF SCIENCES
TECHNICAL SCIENCES, Vol. 60, No. 3, 2012
DOI: 10.2478/v10175-012-0058-x

Ensemble of data mining methods for gene ranking

A.WILIŃSKI[1] and S. OSOWSKI[2,3]*

[1] University of Life Sciences, Faculty of Applied Informatics and Mathematics, 159 Nowoursynowska St., 02-776 Warszawa, Poland
[2] Institute of the Theory of Electrical Engineering, Measurement and Information Systems, Warsaw University of Technology, 1 Politechniki Sq., 00-661 Warsaw, Poland
[3] Military University of Technology, Institute of Electronic Systems,2 S. Kaliskiego St., 00-908 Warsaw, Poland

Abstract. The paper presents the ensemble of data mining methods for discovering the most important genes and gene sequences generated by the gene expression arrays, responsible for the recognition of a particular type of cancer. The analyzed methods include the correlation of the feature with a class, application of the statistical hypotheses, the Fisher measure of discrimination and application of the linear Support Vector Machine for characterization of the discrimination ability of the features. In the first step of ranking we apply each method individually, choosing the genes most often selected in the cross validation of the available data set. In the next step we combine the results of different selection methods together and once again choose the genes most frequently appearing in the selected sets. On the basis of this we form the final ranking of the genes. The most important genes form the input information delivered to the Support Vector Machine (SVM) classifier, responsible for the final recognition of tumor from non-tumor data.

Different forms of checking the correctness of the proposed ranking procedure have been applied. The first one is relied on mapping the distribution of selected genes on the two-coordinate system formed by two most important principal components of the PCA transformation and applying the cluster quality measures. The other one depicts the results in the graphical form by presenting the gene expressions in the form of pixel intensity for the available data. The final confirmation of the quality of the proposed ranking method are the classification results of recognition of the cancer cases from the non-cancer (normal) ones, performed using the Gaussian kernel SVM. The results of selection of the most significant genes used by the SVM for recognition of the prostate cancer cases from normal cases have confirmed a good accuracy of results. The presented methodology is of potential use for practical application in bioinformatics.

Key words: gene expression array, feature selection, gene ranking methods, classification, SVM.

1. Introduction

DNA microarray represents now a powerful tool in biomedical discoveries [1–16]. They store the expressions of thousands of individual genes on a single surface of the size of the microscope slide. Such an array allows to see genes that are induced or represent the medical experiment. The signature of a disease (for example the cancer) may be encrypted in DNA microarrays and then used for diagnosis of the disease of the other patients [2]. The problem in a gene expression analysis is that the number of measurement variables (genes) are very large and extend up to tens of thousands, while the number of observation (number of patients) is usually very limited and is within the range of hundreds. The adjustment of a large number of free parameters from the scarce observation is a difficult and error susceptible task, since the problem is ill conditioned. This means that the problem of selecting the genes strictly associated with the particular type of illness needs special methods of solution.

Nowadays we observe great progression of data mining methods for feature selection. They rely on different principles and possibly generate different results for the same data sets. A good practice in ill defined problems is application of few methods simultaneously and draw the final conclusion by considering the results of all of them [3, 4, 11]. Conflicting results in repeated experiments are resolved through attention to the statistical details.

This paper addresses the problem of the prostate cancer recognition on the basis of the data of gene expression array. The input data is the n-dimensional vector formed of elements of this array, which may be called the features. It means that in such statement of the problem the features are understood as the coefficients of different genes generated by the expression microarray. Each row of data corresponds to one patient. The most important problem is that the number of genes is extremely large (more than ten thousands) and the number of patterns – very limited (around one hundred). The task is to select the genes which are most strongly associated with the cancer. We may regard them as the genes characteristic for this particular type of cancer. These genes input to the classifier, will provide good factors taken into account in recognition of cancer cases from the non-cancer ones. However, to get the satisfactory generalization of the trained classifier [1] we have to apply only limited number of genes, by considering only the most important ones. This is the typical selection problem of data mining.

There are many different methods used at feature selection for gene recognition on the basis of DNA microarray. The measures combined with the correlation analysis, clusterization of data, different distance approaches, statistical hypothe-

*e-mail: sto@iem.pw.edu.pl

ses, Bayesian formulation, application of linear kernel Support Vector Machines, chi-square, information entropy-based, and many others are the most often used [2, 3, 5, 6, 13]. In spite of many existing approaches to the feature selection the problem of efficient and reliable choice of the most important genes is still open in research.

In the numerical experiments we consider the gene ranking in the prostate cancer problem containing two classes of data [12]. An experimental data is the set of patterns $\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_p$ representing the available values of gene expressions with known class labels d_1, d_2, ..., d_p representing either tumor or non-tumor cases. The small subset of the most representative features (the gene expression coefficients) is used to train the Support Vector Machine (SVM) network generating the decision function $D(\mathbf{x})$ for the particular input pattern vector $\mathbf{x}$. The trained classifier is next used to recognize and assign the newly acquired data to the appropriate class.

2. The problem description

The gene expression array is a huge matrix of a very high number of columns representing genes or gene sequences and a small number of rows corresponding to the succeeding patients. The value of expression corresponding to each position in the row describes the intensity of transcription of the particular gene. The expressions of some genes are characteristic for the specific type of illness and are similar for many patients suffering from this illness. The aim of application of data mining methods is to discover these most important genes.

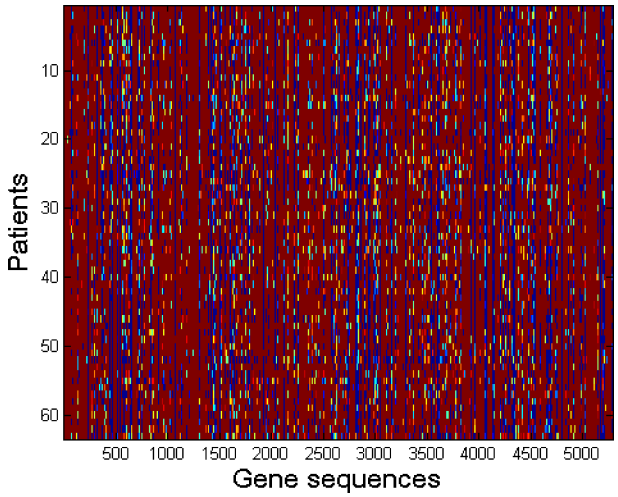

Fig. 1. The visual form of gene expression array corresponding to two classes of leukemia at application of all available gene sequences

When looking at the graphical form of all gene expressions for many patients we can't observe any particular pattern characteristic for different classes. This is well illustrated on the example of typical microarray data representing two classes of leukemia [12, 13]. The first class data (38 records) represent the acute lymphoblastic leukemia of Burkitt type (ALL B-cell) and the second (25 records) the acute myelogenous leukemia (AML). Each vector of the gene expression contains 5327 elements.

Figure 1 is the visual illustration in the form of pixel intensity for the gene expression of these data corresponding to two classes of leukemia. The columns correspond to the genes and the rows to the patients. We cannot observe any visible form of graphical division of the image into two groups associated with two classes of patients.

The important problem is to select the fixed number of top rank genes that are most representative and discriminative for both classes. After a proper selection we should see the clear division of data into two separate graphical regions corresponding to classes.

In this paper we limit our considerations of gene ranking to prostate cancer data, represented by two-classes. One class corresponds to the gene expressions of the prostate tumor cases (52 records) and the second to non-tumor cases (50 records) representing the reference class. The total number of genes in both classes was equal 10509. The data was acquired from the benchmark available in internet [15].

An experimental data form the set of patterns $\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_p$ ($p = 102$), each arranged in the form of row vector and present the available values of gene expressions with known class labels d_1, d_2, ..., d_p representing either first or second class patients. The aim of the work is to select small population of the most important genes well differentiating both classes. This small subset of selected features, representing the gene expression coefficients will be next used to train the Support Vector Machine (SVM) network generating the class membership for the particular input pattern vector $\mathbf{x}$. Each new pattern described by the vector $\mathbf{x}$, delivered to the input of the classifier is classified according to the sign of the decision function $D(\mathbf{x})$: class 1 when $D(\mathbf{x}) > 0$, class 2 when $D(\mathbf{x}) < 0$.

In this paper we discuss various methods of gene selection, including Kolmogorov-Smirnov test of different variants, Wilcoxon method, correlation analysis, the statistical Fisher measure based on the analysis of distribution of centres and variances of the clusters, as well as the gene ranking using linear Support Vector Machine [3, 7]. The results of each separate selection process are combined together in the form of ensemble, to perform the second step of selection, leading to the optimal ranking of genes.

3. Theoretical basis of gene ranking methods

Gene ranking is a specific form of the general process of feature selection, in which each gen or gene sequence expression is associated with a feature. There are many different methods of feature selection [4]. In this work we limit our considerations to only few of them: the correlation of the data of gene expression with a class, the Kolmogorov-Smirnov and Wilcoxon-Mann-Whitney statistical tests, the clusterization measures and ranking of genes by using linear Support Vector Machine.

3.1. The correlation of gene expression with a class. One of the methods assessing the discriminative power of the candidate feature f for the recognition of the particular class

among K classes is the correlation of this feature with the class [8]. Let us denote by **d** =[k_1, k_2, ..., k_K] the vector of class membership, by m(f) = E{f} the mean value of the feature f in the whole set of data by m_k(f) = E{f|k} the mean value of features for data records forming kth class and by var(f) = E{(f – m(f))^2} the variance of the feature f for the whole data. The correlation of feature f with the class vector d is defined through the vector of covariances in the form [8]

$$S(f) = \frac{|\text{cov}(f, \mathbf{d})|^2}{\text{var}(f)\text{var}(\mathbf{d})}, \tag{1}$$

where the vector of covariances is expressed by

$$\text{cov}(f, \mathbf{d}) = E\{f\mathbf{d}\} - E\{f\}E\{\mathbf{d}\} \tag{2}$$

and $E\{f\mathbf{d}\} = \sum_{k=1}^{K} P_k E\{f\mathbf{d}\,|k\}$ with P_k denoting the probability of kth class. On the basis of this we express the covariance vector by

$$\text{cov}(f, \mathbf{d}) = \begin{bmatrix} P_1\,(m_1(f) - m(f)) \\ P_2\,(m_2(f) - m(f)) \\ \cdots \\ P_K\,(m_K(f) - m(f)) \end{bmatrix} \tag{3}$$

and the final expression estimating the discriminative power of feature f for recognition of K classes in the form

$$S(f) = \frac{\sum_{k=1}^{K} P_k^2 (m_k(f) - m(f))^2}{\text{var}(f) \sum_{k=1}^{K} P_k(1 - P_k)}. \tag{4}$$

Limiting the number of classes to only two and assuming equal probability of both classes we can simplify the discriminative measure of feature f to recognize one class from the second to the form

$$S_{12}(f) = \frac{(m_1(f) - m(f))^2 + (m_2(f) - m(f))^2}{2\text{var}(f)} \tag{5}$$

After calculating this measure for all features we can arrange them in a decreasing order, from the highest to the smallest discriminative value. Such arrangement automatically determines the ranking of the features. This method of feature ranking will be denoted shortly by **COR**.

3.2. The Fisher discriminant measure. The other, often used criterion of feature selection is the analysis of variances and means of the feature samples belonging to the clusters, representing each class [1, 2]. The first step is to calculate the means and standard deviations of the feature for both classes. The standard deviation of the feature describing the data belonging to one class should be as small as possible, and at the same time the positions of means of the feature values for the data belonging to different classes should be as far as possible. Denoting these means by m_1 and m_2 and standard deviations by σ_1 and σ_1, respectively we can define the 2-class discrimination measure of the feature f in the form

$$S_{12}(f) = \frac{|c_1(f) - c_2(f)|}{\sigma_1(f) + \sigma_2(f)}. \tag{6}$$

The large value of $S_{12}(f)$ indicates good separation ability of the feature f for recognition of these two classes. Small value means that clusters of both classes are close to each other and the data samples are widely distributed. Such feature does not represent any discriminative value. In the case of many classes the discriminative measure should be calculated for all 2-class combinations of them. The total discrimination power of feature f is the sum over all combinations. The results generated by this method is denoted further by **FISH**.

3.3. The Kolmogorov-Smirnov tests. The next investigated methods of feature selection will be based on statistical hypotheses [9]. We apply here two groups of methods. The first is the Kolmogorov-Smirnov (KS) and the second Wilcoxon-Mann-Whitney (WMW) tests. In these tests we treat the features f as a statistical variables of some special distribution related to the type of data [9]. Well discriminating feature has similar distribution for the group of patients belonging to the same class, and different distribution for patients of different classes.

We compare the statistical distribution of values of the particular feature f corresponding to one class ($\mathbf{x}_1$) and the second class ($\mathbf{x}_2$) of patients. The KS test is performed for these two vectors to determine if they are drawn from the same underlying continuous population. The KS test is relied on checking the null hypothesis that the samples of both classes are drawn from the same distribution at the desired significance level (default = 0.05). If the estimated significance level is below the desired one the null hypothesis is accepted, otherwise is rejected. The KS-test is a robust test that cares only about the relative distribution of the data belonging to two populations and does not take into account the order of features.

The important advantage of it is that no specific assumption about the distribution of data is taken a'priori. It is a non-parametric and distribution free test. The Matlab function *kstest2* [13] implementing this test delivers the distance between the cumulative distribution functions of the data belonging to two compared classes. This distance may be regarded as the statistical measure of difference between the distribution of both populations. Let us denote the cumulative distribution of both populations by F($\mathbf{x}_1$) and F($\mathbf{x}_2$). Using the Kolmogorov-Smirnov test we can define three different discriminative measures.

- Maximum Kolmogorov-Smirnov measure **(MKS)**

$$S_{12}(f) = \text{sup}|F(\mathbf{x}_1) - F(\mathbf{x}_2)|. \tag{7}$$

- Additive Kolmogorov-Smirnov measure **(AKS)**

$$S_{12}(f) = sum|F(\mathbf{x}_1) - F(\mathbf{x}_2)|. \tag{8}$$

- Scaled Kolmogorov-Smirnov measure **(SKS)**

$$S_{12}(f) = a(f) \cdot \text{sup}|F(\mathbf{x}_1) - F(\mathbf{x}_2)|. \tag{9}$$

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

$$a(f) = \frac{|mean(\mathbf{x}_1) - mean(\mathbf{x}_2|}{std(\mathbf{x}_1) + std(\mathbf{x}_2)}. \tag{10}$$

High values of these measures indicate that the distributions of points belonging to two classes are different (don't belong to the same population of samples). Such feature is of high quality. On the other hand low values indicate poor discriminative ability of the feature f.

3.4. The Wilcoxon-Mann-Whitney statistics. The Wilcoxon-Mann-Whitney (WMW) test [9] applied here is another statistic measure used for assessing the similarity of statistical distribution of two vectors. It applies so called U statistics. Using the WMW test, we can decide whether the investigated population distributions are identical without assuming them to follow the normal distribution or assuming that the variances of the two populations are equal. This test is based on the idea that the particular pattern, exhibited when some numbers of X random variables and some numbers of Y random variables are arranged together in increasing order of magnitude, provides information about the relationship between their parent populations.

The WMW test criterion is based on the magnitude of the feature f in class 1 in relation to the class 2, i.e. the position of the first sequence in the combined ordered sequence. A sample pattern of 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 distribution.

In order to calculate the U statistics, the combined set of data is first arranged in an ascending order with tied scores receiving a rank equal to the average position of those scores in the ordered sequence.

Let T denote the sum of ranks for the first set of samples. The Wilcoxon-Mann-Whitney test statistic is then calculated using $U = n_1 n_2 + n_1(n_1+1)/2 - T$, where n_1 and n_2 denote the sizes of the first and second samples respectively. We next compare the value of calculated U with the value given in the Tables of Critical Values for the Mann-Whitney U-test (the critical values are provided for given n_1 and n_2) and get their difference, on the basis of which we accept or reject the null hypothesis.

Finally the probability p denoting the degree of similarity of both sequences is estimated. The higher this probability, the more similar are the two populations. Small value of p means large differences between two populations. The final ranking of features is created on the basis of the value of expression

$$S_{12}(f) = 1 - p. \tag{11}$$

This method of feature ranking will be denoted shortly by **WMW**.

3.5. The one-input linear SVM method. The next method considered for feature selection is the application of the one-input linear kernel Support Vector Machine [10, 12]. The first step of application of this method is training the SVM network on the data set by using only one feature at a time. We train as many networks as is the number of features. The predictive power of the single feature is characterized by the value of the error function of the class recognition provided by a one-dimensional linear SVM trained to classify all learning samples on the basis of only one feature of interest. The smaller this error the better is the quality of the feature. Training many SVM networks by applying one feature at a time, selected in turn from the feature set, will inform of the quality of the features. The final discriminative value of the feature is then defined as

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

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. The features arranged in decreasing order of this discriminant value from highest to lowest create the ranking. This method of feature ranking will be referred later as **1SVM**.

3.6. The multi-input linear SVM method. The application of multi-input linear SVM, called SVM recursive feedback elimination, belongs to the well known and widely used selection method proposed originally by Guyon and Vapnik [3] which found some modifications [5]. The most important distinction to the previous methods is that the discrimination power of each feature is tested in the presence of the whole set of features.

The ranking of the features is done here as a result of simultaneous application of all features in the role of input information to the linear kernel SVM working as a classifier. The linear kernel SVM is used, because this kernel does not deform the original impact of each feature on the result of the classification. The decision function of the N-dimensional input vector $\mathbf{x}$ is a linear function defined as $D(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$ with the weight vector $\mathbf{w}$ and bias b dependent on the linear combination of the training patterns ($\mathbf{x}_k$, d_k) belonging to the support vectors. The positive value of $D(\mathbf{x})$ means membership of vector $\mathbf{x}$ do the class 1 and negative value to the opposite one.

The method is based on the idea, that the absolute values of the weights of a linear classifier produce a feature ranking. The feature associated with the larger weight is more important than that associated with the small one. It means that this time the discriminative power of feature f at recognition of two classes is determined by the relation

$$S_{12}(f) = w_f, \tag{13}$$

where w_f means the value of weight joining the input of the feature f with SVM network. All values of weights are arranged in decreasing order and the least important are rejected, reducing the size of the remaining features. The procedure of the feature elimination is repeated many times by training the SVM classifiers at application of the shorter and shorter feature vectors forming 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 achieved the vector of the appropriate size.

4. The numerical experiments of gene ranking

The presented above methods of feature selection have been applied simultaneously in the process of selection of the most important genes associated with the particular division of data into two classes. All experiments have been performed using the publicly available data [15] by processing the gene expression array corresponding to prostate tumor (PRT). The number of genes for this particular cancer type was equal 10509. The total number of patients was equal only 102, from which 52 correspond to tumor cases and the rest (50) to the non-tumor ones.

4.1. The general procedure of gene ranking. The most important problem with gene ranking is very small number of available records (only 102) in comparison to the number of genes (10509). To get the reliable results we have to perform many experiments of ranking using part of data containing each time the randomly chosen sets, analyzing the results and looking for the genes selected most frequently. Each time we select randomly 90% of available records, changing the contents of the set for each experiment. For each methods 100 trials have been performed using these 90% of randomly selected data from the whole set. The statistics of selection of each gene in these 100 trials are made. In this way we are able to arrange the selected genes according to the frequency of their appearance among the first 100 best. All features beyond the first 100 are ignored in this statistics. As a result we determined the frequency of appearing of each gene among the selected 100 best. We have considered the following ranges of frequencies: 100% (the gene appeared in all 100 trials among the first 100 best) , from 90% to 100%, from 80% to 90% and from 60% to 80%. Since for each selection method we have got different contents of the most frequently selected genes, in the second step we search for the genes most often appearing in all selection methods. The same ranges of frequencies have been considered. In this way we were able to select the sets of genes appearing in the selection methods with different frequency of repeatability. This created the natural ranking of genes of the highest discriminative ability.

4.2. The numerical results of gene ranking. The numerical results concerning the problem of selection of the most discriminative genes have been performed on the available data of prostate tumor. The procedures of gene selection have been repeated 100 times by applying all presented above feature selection methods. For each method the statistics of gene appearance among 100 best were made. Such statistics was prepared for all selection methods. If the particular gene appeared among 100 best in all runs of selection it was added to the list of genes with 100% repeatability. The genes appearing among the best with lower frequency of runs was added to the list of genes of proper range of repeatability, for example [90% – 100%) range, etc.

Table 1 presents the results of the experiments concerning this step of selection. The first column of the table contains the short name of the applied selection methods: COR – correlation analysis, FISH – the Fisher measure based on centres and standard deviation values, MKS, AKS and SKS – the measures based on Kolmogorov-Smirnov test, WMW – the measure corresponding to the Wilcoxon-Mann-Whitney test, 1SVM – the one-input SVM ranking method, MSVM – the multiple input SVM method of selection.

Table 1
The results of the first step of ranking. The values in the table denote the number of genes selected among 100 best by different methods in the cross validation trials at the defined ranges of repeatability

Selection method	[60% – 80%)	[80% – 90%)	[90% – 100%)	100%
COR	73	60	50	31
FISH	76	62	51	32
MKS	78	55	44	25
AKS	71	53	45	32
SKS	83	59	48	31
WMW	72	55	47	32
1SVM	55	58	47	30
MSVM	76	54	45	31

Each column of the table represents the number of genes which has been chosen by each selection method at the defined ranges of repeatability in all 100 experiments. The last column shows the number of genes which have been selected among 100 best in all 100 cross validation experiments. The interesting fact is that this number was quite stable and changed from 30 to 32 in all except one selection method (MKS).

In the next step of ranking we compare the contents of the list of genes within the appropriate ranges of repeatability for all applied methods of selection. The aim is to find the genes commonly selected by all methods. Close checking of the contents of these genes has revealed that only 7 particular genes have been selected in all runs simultaneously by all selection methods. These genes have been treated by us as the most representative for this particular type of cancer. This may be quite important information for the researchers in biology and medicine.

In the same way we have discover the number of genes commonly selected by all methods within the considered ranges of repeatability (between 90% and 100%, between 80% and 90% and between 60% and 80%) in all cross validation experiments. The number of such genes within the considered ranges of repeatability are depicted in Table 2.

Table 2
The results of the second step of ranking. The values in the table denote the number of genes commonly selected by all selection methods

Repeatability range	Number of genes
100%	7
[90% – 100%)	23
[80% – 90%)	42
[60% – 80%)	57

Among all 10509 genes only 129 have been selected among the best with the repeatability higher than 60% in all experiments after application of all selection methods. To check the discriminative quality of the selected genes we have

compared the visual form of gene expression value distribution corresponding to both classes for all genes and for the selected 7 genes, regarded as the best.

Figure 2a presents this distribution for all genes and Fig. 2b for the selected 7 most important genes. The horizontal axis represents genes and vertical – the cases. The first 52 rows represent cancer and the other 50 non-cancer cases. In the case of all genes it is hard to see any border between the colour distribution of gene expression for cancer and non-cancer cases. Reducing the number of genes to only 7 selected by us has shown clear difference between both groups of people. There is visible similarity of the colour patterns for the records representing the same class. At the same time it is evident that the pattern of gene expression intensities of the first class is significantly different from the pattern of records representing the second class. These graphical results confirm close association of the expression activity of the selected genes with their membership in both classes of data.

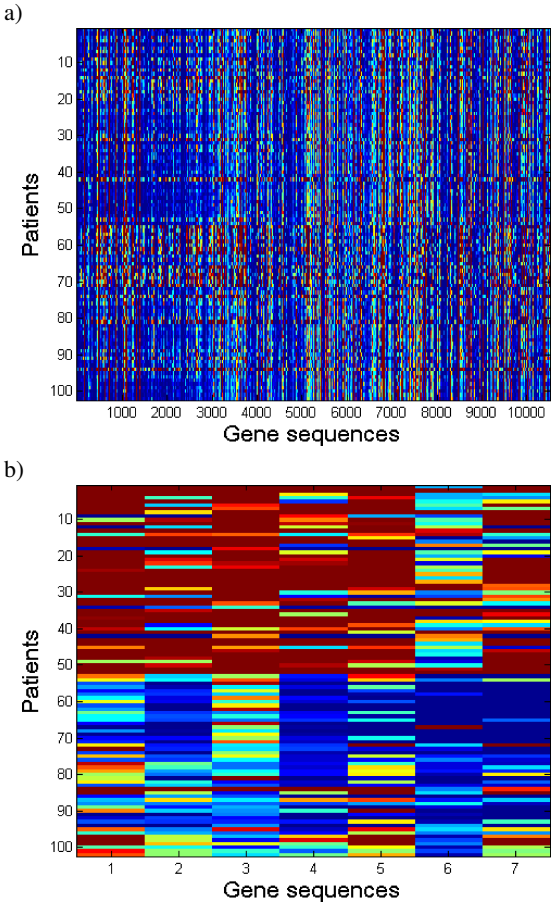

Fig. 2. The visual distribution of gene expression values corresponding to 2 classes of data; a) all genes, b) the selected 7 genes of 100% repeatability. The horizontal axis represents genes and vertical – the individuals. The first 52 individuals represent cancer and the other 50 normal cases

4.3. Principal component analysis of the selection results. To illustrate graphically how the selected genes represent the distribution of data we have applied the principal component analysis mapping the data $\mathbf{x}$ from the higher order space, formed by the specific number of top ranked genes, to the 2-D space defined by two most important principal components: PCA_1 and PCA_2. The PCA is a linear transformation defined as follows [7]

$$\mathbf{y} = \mathbf{W}\mathbf{x} \,, \tag{14}$$

where the transformation matrix $\mathbf{W}$ is formed from two eigenvectors corresponding to two largest eigenvalues of the correlation matrix of the original data $\mathbf{x}$. The PCA_1 and $PCA_2$are the first two elements of vector $\mathbf{y}$.

In the analysis we have limited ourselves to the representation of only 100 best genes selected commonly by all selection methods. To show the significance of the selected genes we will consider and compare two cases of PCA transformation. In the first case we map the data corresponding to 100 top rank genes. In the second one we replace the 30 highest rank genes by the genes occupying the positions from 101 to 130, leaving the total number of genes identical.

In Fig. 3a we present the distribution of points belonging to two classes, representing them by 100 top ranked genes. One class is denoted by symbol x and the second by the circle

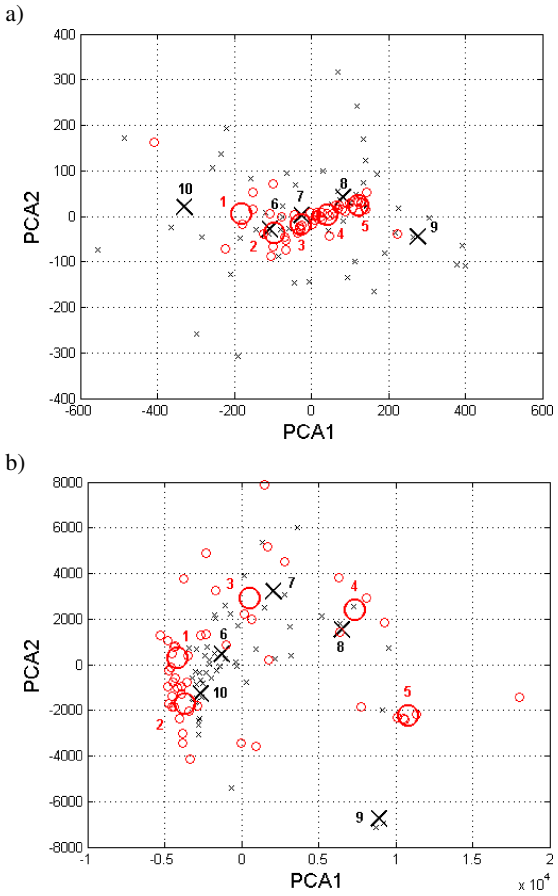

Fig. 3. The distribution of the two-class data of PT (small size symbols) and the representatives of the clusters (large symbols) mapped on two most important principal components for a) 100 top rank genes, b) after replacing 30 top rank genes by the genes of numbers 101 to 130

(small size black symbols). Figure 3b presents the situation after removing 30 highest top rank genes. At consideration of only 100 highest rank genes the dispersion of samples belonging to the same class is the smallest one. After replacing 30 top rank genes by the less representative ones we can observe much higher dispersion (observe different scale of both figures). Moreover there are more samples of one class interlaced with the samples of the second one in the whole region.

In the next phase we have made clusterization of the data, splitting all records into 10 clusters (5 for one class and 5 for the second one). The results are depicted in Fig. 3 in the form of large red symbols of x (first class) and large red symbol of circle (second class). Next, the average distances of the original points to their winning centres have been calculated. In the case of data represented by 100 top rank genes (Fig. 3a) the average distance was equal 0.15e3. After removing 30 top rank genes (Fig. 3b) this distance has rose to 0.23e4. We observe more than 15 fold increase of the relative distance. The appropriate values related to the groups of genes belonging to 2 separate classes are given in Table 3.

It is evident that after removing the best rank genes the dispersion of points measured by their distance to the winning centres has been drastically increased. It confirms the significance of the ranking results in a numerical way.

Table 3
The average distances of the data belonging to 2 classes to their appropriate winning centres

	Class 1	Class 2
100 top rank genes	0.33e3	0.26e3
30 top rank replaced by the less representative genes	9.26e4	0.82e4

5. Classification of the prostate tumor data using Gaussian kernel SVM

5.1. SVM classifier of Gaussian kernel. The last step of checking the importance of the developed gene ranking is performing the classification of data into two classes using different arrangement of genes. The final recognition of tumor and healthy cases has been done by applying Support Vector Machine classifiers of Gaussian kernel [7]. SVM is known of high efficiency at relatively low number of learning data and high dimension of input vectors. The Support Vector Machine (SVM) is the solution of a feedforward structure with one hidden layer, applying special method of learning [7]. In distinction to the classical neural network formulation of the learning problem, where the minimized error function is nonlinear with respect to the optimized variables of many potential minima, SVM leads to the quadratic programming with linear constraints of one, well defined global minimum. Basically, the SVM is the one output linear machine, working in the high dimensional feature space formed by the nonlinear mapping of the original N-dimensional input vector $\mathbf{x}$ into a K-dimensional feature space ($K > N$) through the use of a nonlinear function $\varphi(\mathbf{x})$ arranged in the form of kernel. The details of learning this network can be found in excellent book of Scholkopf and Smola [7].

The important role in practice of learning fulfils the regularization constant C. It determines the balance between the complexity of the network, characterized by the values of weights and the error of classification of learning data. Low values of C mean smaller significance of the learning errors in the adaptation stage and leads to the smaller size networks of higher separation margin. The higher values of C lead to the more complex network structures with a smaller separation margin. For the normalized input signals the value of C is usually much bigger than 1 (typical value is in the range 100–1000). In practice it is adjusted by trials and errors using small percentage of the validation data extracted from the available learning data. In the same way we adjust the proper value of the parameter σ of the Gaussian kernel function. Both parameters C and σ are adjusted simultaneously by trying different combinations of their values in introductory experiments. They have been kept constant in all performed experiments of classification.

5.2. The statistical results of classification. The efficiency of an automatic recognition of cancer on the basis of gene expression microarray data is largely dependent on the proper selection of genes. Application of all genes in recognition is senseless, since the number of records (corresponding to patients) is too small in comparison to the number of genes [3]. Therefore the comparison of the classification results at application of the selected genes should be compared and assessed on a different basis.

In this research we compare the classification results at different compositions of the genes selected by the ensemble of methods. Since the number of genes of 100% repeatability is very small their dimension is not sufficient as the input information to the classifier in the class recognition process. On the basis of the introductory experiments we decided to use 100 high ranked genes. To get the objective results we have applied 100 repetitions of learning/testing experiments at random choice of learning and testing data. In all experiments half of the data records has been used in learning and the remaining half in testing. On the basis of these experiments the average errors of testing have been calculated for all selection methods.

To compare the importance of gene ranking we have made these 100 cross validation experiments at different composition of genes. In the first set we have used 100 best ranked genes (case 1). Next we have eliminated the highest ranked genes of the 90–100% repeatability (including 7 genes pointed by all selection methods) and replaced them with the genes occupying the highest ranking positions starting from 101 (case 2). In the same way we have repeated the classification experiments excluding the genes of repeatability from 80% to 90% (case 3), from 70% to 80% (case 4) and from 60% to 70% (case 5). Figure 4 depicts the mean relative error of class recognition on the testing data not taking part in learning in all 100 experiments at different composition of genes. It is evident that inclusion of all 100 highest ranked genes provides the highest accuracy (the least relative error) in all experiments. The lowest accuracy (the highest relative error)

was observed after elimination of genes of the repeatability in the range [90% – 100%]. The closest to the best selected gene results is the case 5, when only the genes of the repeatability in the range 60–70% were eliminated. These genes represent relatively lower portion of information than the genes of the highest repeatability, so their elimination results in only small decrease of the class recognition accuracy.

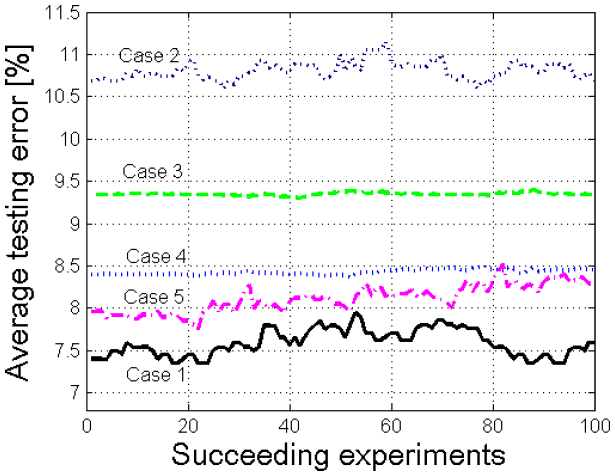

Fig. 4. The distribution of mean relative error of class recognition in the succeeding 100 cross validation experiments at different compositions of genes

Table 4 presents the comparison of the mean values and standard deviations of class recognition errors calculated over all cases of experiments. The best results of recognition have been marked in bold. There is significant difference between the accuracy of recognition at application of only highest rank genes and after elimination of some of them. After elimination of 30 top rank genes (7+23) and substituting them by the lower rank genes the increase of error was the highest one. It rose from 7.6% to 10.8%.

Table 4
The mean values and standard deviations of class recognition errors for all cases of 100 cross validation experiments

Case	Input gene composition	Mean error ± std
1	**100 top rank genes**	**7.6±0.16**
2	Elimination of genes of [90% – 100%) repeatability	10.8±0.11%
3	Elimination of genes of [80% – 90%) repeatability	9.3±0.02%
4	Elimination of genes of [70% – 80%) repeatability	8.5%±0.03%
5	Elimination of genes of [60% – 70%) repeatability	8.1%±0.16%

We have tried to apply only 7 highest rank genes in the classification. However, the results have shown that this number is not sufficient to get the highest accuracy of class recognition. The mean testing error of 100 experiments was equal 8.3%. The additional experiments have been performed at application of only 100 worst discriminating genes (the genes occupying the last positions in ranking). This time the mean error of class recognition has increased to the value of 48%. This result confirms the significance of the presented approach to gene ranking.

5.3. The additional quality measures of classification. In all medical experiments the accuracy is only one measure of quality. Since this measure treats every class as equally important it is not sufficient to assess the method in a satisfactory way. The additional aspects of results associated with the importance of recognition of cancer class should be also analyzed. If we denote the cancer cases by + and non-cancer ones by – we can present the results in the form of confusion matrix, as shown in Table 5.

Table 5
A confusion matrix for a 2-class classification problem in which the classes are not equally important

	Predicted Class +	Predicted Class −
Real class +	TP	FN
Real class −	FP	TN

The following terminology is used when referring to the counts tabulated in a confusion matrix:

- True positive (TP) – corresponds to the number of positive examples correctly predicted as positive by the classifier.
- False negative (FN) – corresponds to the number of positive examples wrongly predicted as negative by the classifier.
- False positive (FP) – corresponds to the number of negative examples wrongly predicted as positive by the classifier.
- True negative (TN) – corresponds to the number of negative examples correctly predicted as negative by the classifier.

The counts in a confusion matrix may be also expressed in terms of percentage. On the basis of the numbers in confusion matrix we may define additional measures of quality. One of the most important is true positive rate (TPR), called also sensitivity, defined as the fraction of positive examples predicted correctly by the classifier

$$TPR = \frac{TP}{TP + FN}. \tag{15}$$

Similarly the true negative rate, (TNR), called specificity is defined as the fraction of negative examples predicted correctly by the classifier

$$TNP = \frac{TN}{TN + FP}. \tag{16}$$

The next used measure is the false alarm rate (FA) defined as the ratio of the negative class cases recognized by the classifier as the positive. It is defined as

$$FA = \frac{FP}{FP + TN} = 1 - TNR. \tag{17}$$

The next one is the positive predictivity value (PPV) defining what part of cases recognized as positive is really positive

$$PPV = \frac{TP}{TP + FP}. \tag{18}$$

In the same way we define the negative predictivity value (NPV) defining what part of cases recognized as negative is really negative

$$NPV = \frac{TN}{TN + FN}. \tag{19}$$

Table 6 presents the numerical results concerning these quality measures at recognition of prostate cancer cases from the healthy ones. They correspond to the use of 100 highest ranking genes selected by the ensemble of methods.

Table 6
The values of the quality measures at recognition of prostate cancer cases (class +) from the healthy ones (class –) at application of 100 highest rank genes

TPR	TNR	FA	PPV	NPV
93.7%	91.1%	8.9%	91.2%	93.6%

To compare how application of ensemble improved the results of recognition we have made similar 100 cross validation experiments at application of the individual selection methods. Each time we applied 100 genes occupying the highest positions in ranking. Table 7 presents the obtained results of classification in the form of mean relative errors at application of individual methods and at application of ensemble (the last row denoted in bold). The first column numerical results correspond to 100 best genes and the second one depict the influence of replacement of the highest rank genes of the repeatability from 90% to 100% by the next rank genes starting from the position 101.

Table 7
The mean values of the class recognition errors at application of individual selection methods (the average results of 100 cross validation experiments)

Selection method	100 top rank genes	Elimination of the best rank genes
COR	10.38%	14.04%
FISH	13.62%	17.95%
MKS	12.09%	16.56%
AKS	12.28%	17.06%
SKS	11.71%	15.92%
WMW	12.86%	17.61%
1SVM	10.89%	15.44%
MSVM	8.78%	12.02%
Ensemble	7.60%	10.8%

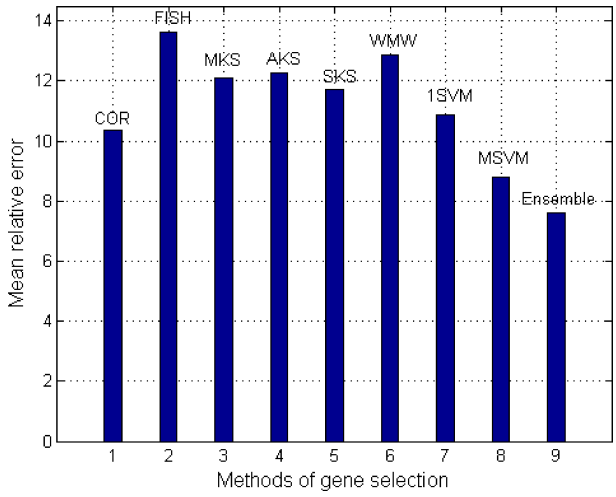

Fig. 5. The comparison of the mean value recognition error at application of different selection methods

Very large differences of the results can be observed. They are following from the application of the particular feature selection method. The best individual selection method was the multiple input SVM denoted as MSVM (8.78% of mean error). However, even this method is significantly worse than ensemble (7.60% of the mean error). The relative difference between the best individual method and the ensemble is equal 15.5%. This ratio is much higher and equal 79.2% when we relate the ensemble to the worst method (Fisher).

Figure 5 presents the graphical form of comparison of the average misclassification rate for the individual methods of selection and the application of ensemble. They reflect the data depicted in Table 7.

6. Conclusions

The paper has compared different methods of gene ranking for the recognition of the prostate tumor on the basis of the gene expression array data. We have applied 8 measures of gene significance for the recognition of two classes of data: the prostate tumor cases and non-tumor cases treated as the reference class. These methods represent different approaches to the selection and make use of the correlation measure, clustering properties, analysis of statistical distribution of data in two classes as well as application of classification ability of linear SVM.

The paper has proposed the two step procedure of ranking the most important genes associated with the class distribution data. In the first step each method acts in an independent way, estimating the value of quality measure for each gene. On the basis of these values the genes are ranked from the best to the least significant. Since the number of records is very scarce in comparison to the number of genes the ranking procedure is repeated many times using different (randomly chosen) set of records. Next we determine the number of times each gen was selected among the best in these cross validation runs and on the basis of this repeatability ratio we finally assume the first step ranking of genes.

In the second step of the ranking procedure we compare the contents of the high rank sets of genes created by different methods. As a results of it we are able to identify the highest rank genes chosen by all selection methods (100% of repeatability) as well as the subsequent genes of lower repeatability. This way of processing arranges the final ranking of genes.

The quality of the ranking procedure has been checked in different ways. First, we check the visual form of recognition of both classes by the selected genes. The other method has applied the clustering of the data and presenting them in 2-D space by using their mapping through the principal component analysis. The final form of checking was the application of SVM classifier with Gaussian kernel, responsible for recognition of the tumor data from the non-tumor cases. In this step we use the top rank genes in different arrangement as the input information to the classifier, trying to get the best possible recognition of classes to which the succeeding records belonged.

On the basis of many experiments, repeated for different sets of learning and testing data formed randomly from the whole data set, we have selected relatively small number of the most important genes, that have been associated the prostate tumor at a highest degree. These genes have formed the input vector **x** applied to SVM classifier performing the classification task.

The results of class recognition on the testing data, not taking part in learning were of a high quality, proving the efficiency of the gene ranking methods. The average accuracy of a class recognition in the prostate tumor problem calculated over 100 cross validation runs was equal 92.4%. This result is in a good relation to the most recent results for similar data of the prostate tumor [11], where the declared average accuracy on the PR data gathered in the base [16] was equal 90.98%.

Observe at the end that the gene ranking methods, considered in the work, do not dictate their optimal number. They only rank them according to their degree of association with a class. An additional analysis is needed to find the optimal number of the genes, providing the highest efficiency of classification.

REFERENCES

[1] R.O. Duda, P.E. Hart, and P. Stork, *Pattern Classification and Scene Analysis*, Wiley, New York, 2003.

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

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

[4] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection", *J. Machine Learning Research* 3, 1158–1182 (2003).

[5] T.M. Huang and V. Kecman, "Gene extraction for cancer diagnosis by support vector machines - An improvement", *Artificial Intelligence in Medicine* 35, 185–194 (2005).

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

[7] B. Schölkopf and A. Smola, *Learning with Kernels*, MIT Press, Cambridge, 2002.

[8] Schurmann, *Pattern Classification, a Unified View of Statistical and Neural Approaches*, Wiley, New York, 1996.

[9] P. Sprent and N.C. Smeeton, *Applied Nonparametric Statistical Methods,* Boca Raton: Chapman & Hall/CRC, London, 2007.

[10] J.P. Vert, "Kernel methods in genomics and computational biology", in *Kernel Methods in Bioengineering, Signal and Image Processing*, eds. G. Camps-Vals, J.L. Rojo-Alvarez, and M. Martinez-Ramon, pp. 42–64, Idea Group, London, 2007.

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

[12] A. Wiliński, "Selected exploration methods of diagnostic features in analysis of gene expression activity", *PhD Dissertation*, Warsaw University of Technology, Warsaw, 2007.

[13] A. Wiliński and S. Osowski, "Gene selection for cancer classification", *COMPEL* 28, 231–241 (2009).

[14] *Matlab User Manual – Statistics Toolbox*, MathWorks, Natick, 1999.

[15] http://discover1.mc.vanderbilt.edu/discover/public/mcsvm.

[16] http://datam.ir2.a-star.edu.sg/datasets/krbd/.