Expert Systems with Applications 42 (2015) 864–872

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Data mining for feature selection in gene expression autism data

Tomasz Latkowski [a,1], Stanislaw Osowski [a,b,*]

[a] Military University of Technology, Faculty of Electronics, Kaliskiego 2, 00-908 Warsaw, Poland
[b] Warsaw University of Technology, Faculty of Electrical Engineering, Koszykowa 75, Warsaw, Poland

ARTICLE INFO

Article history:
Available online 6 September 2014

Keywords:
Gene expression microarrays
Feature selection
Clustering
Classification
Autism

ABSTRACT

The paper presents application of data mining methods for recognizing the most significant genes and gene sequences (treated as features) stored in a dataset of gene expression microarray. The investigations are performed for autism data. Few chosen methods of feature selection have been applied and their results integrated in the final outcome. In this way we find the contents of small set of the most important genes associated with autism. They have been applied in the classification procedure aimed on recognition of autism from reference group members. The results of numerical experiments concerning selection of the most important genes and classification of the cases on the basis of the selected genes will be discussed. The main contribution of the paper is in developing the fusion system of the results of many selection approaches into the final set, most closely associated with autism. We have also proposed special procedure of estimating the number of highest rank genes used in classification procedure.

1. Introduction

Gene microarray technology is a sophisticated technique used in molecular biology for detecting alterations in the expression of thousands of genes simultaneously between different biological conditions (De Rinaldis, 2007). The analysis of the expression levels allows to detect altered gene expression of particular genes in a given disease when compared to healthy controls. From the practical point of view biologists need to identify only a small number of the most significant genes that can be used as biomarkers in the disease tracing. The most relevant genes increase our understanding of the mechanism of disease formation and allow to predict the potential danger of being affected by such disease.

The main problem in this analysis is a limited number of observations related to very large number of gene expressions. Number of observations is usually in the range of hundreds and number of genes tens of thousands. Because of the large imbalance of the number of genes and observations (patients) the selection is an ill conditioned problem. Moreover, data stored in medical databases are typically noisy and some gene sequences have large variance (Alter et al., 2011). It makes the gene selection in DNA microarrays even more difficult task.

Progress in bioengineering and data mining, which has been observed in recent years, has created the solid foundations for discovering the genes which are the best associated with the particular disease. Data analysis of microarrays is widely examined and introduced in literature starting with pioneering Golub investigation in 1999 (Golub et al., 1999).

Actual approaches performing this task include different clustering methods (Eisen, Spellman, & Brown, 1998), application of neural networks and Support Vector Machines (Alonso-González & Moro-Sancho, 2012; Guyon, Weston, Barnhill, & Vapnik, 2002; Wilinski & Osowski, 2012), statistical tests (Baldi & Long, 2001), linear regression methods applying forward and backward selection (Huang & Pan, 2003), fuzzy expert system based algorithms (Kumar, Victoire, Renukadevi, & Devaraj, 2012; Woolf & Wang, 2000), rough set theory (Wang & Gotoh, 2010), use of chaotic binary particle swarm optimization (Chuang, Yang, Wu, & Yang 2011), application of ReliefF method combined with different classifiers (Alonso-González & Moro-Sancho, 2012), various statistical methods (Golub et al., 1999; Mitsubayashi, Aso, Nagashima, & Okada, 2008), as well as fusion of many selection methods (Wilinski & Osowski, 2012; Yang, 2011). Most of the papers studied particular methods and then selected the best one as the most appropriate for the gene selection task, neglecting the others.

Although the progress in this field is high, there is still a need for better understanding and improvement of the research, especially in the medical area not well covered in recent research. To such examples belong autism data (Alter et al., 2011; Esteban & Wall, 2011; Hu & Yinglei, 2013). These data belong to the most

[*] Corresponding author at: Warsaw University of Technology, Faculty of Electrical Engineering, Koszykowa 75, Warsaw, Poland. Tel.: +48 22 234 7235; fax: +48 22 234 5642.

E-mail addresses: tlatkowski@wat.edu.pl (T. Latkowski), sto@iem.pw.edu.pl (S. Osowski).

[1] Tel.: +48 22 234 7235.

http://dx.doi.org/10.1016/j.eswa.2014.08.043
0957-4174/© 2014 Elsevier Ltd. All rights reserved.

demanding, because of very large variability of gene expression values representing the same class of data (Alter et al., 2011). The large variance in the distribution of gene expression levels is associated with many types of symptomatic profiles of autism represented in the base. Therefore, the application of standard methods, which serve very well in recognition of other cases, for example different types of cancer, does not lead to the acceptable results for autism.

Autism is a neurodevelopmental disorder that impairs the normal development of emotional interactions and other forms of social communication (Yang & Gill, 2007). Genetic approaches to autism study aim to identify risk variance at specific genes and in this way to find association of their expression level with the disease. There is a general idea that alterations at the level of gene expression might be important sign in mediating the risk for autism.

This paper is devoted to the task of selection of the genes and gene sequences which are the most closely associated with the disease. The selected genes of the particular expression levels form the most characteristic pattern for the autism. Applying a classifier to such chosen data, should lead to the improved accuracy of the recognition between autism and reference (healthy) cases. These two tasks (gene selection and classification problem) will be considered in the paper.

In the numerical experiments we will analyze different gene ranking methods. It is known that different selection algorithms may provide differing results for the same datasets (Wilinski & Osowski, 2012). The results of individual selection methods will be fused and lead to the final set of genes. The application of several methods gives opportunity to look on the selection problem from different points of view. After fusing their results the probability of proper selection of the most important genes is increased. The results of numerical experiments concerning selection of the most important genes in autism as well as classification of cases on the basis of the selected genes will be discussed.

The other contribution of the paper is developing the fusion system of the results of many selection approaches into the final set, most closely associated with the disease. This is in contrast to the majority of papers, where different methods have been tried, but only one (the best) was treated as the final solution. We have also proposed special procedure of estimating the number of higher rank genes using the self-organization procedure. In the task of classification we have implemented the ensemble of classifiers integrated into the final system, which is responsible for recognition of autism from the reference cases. The trained classifier system may then be used to predict the autism or non-autism class of the newly acquired data.

2. Applied feature selection methods

Feature selection is the most important operation in processing the data stored in gene microarrays. The application of feature selection methods allows to identify a small number of important genes that can be used as biomarkers of the appropriate disease. In this paper some chosen feature selection methods will be examined and integrated into the final system. Using the set of methods instead of single one will increase the probability of finding the globally optimal set of genes which are the best associated with the particular disease.

The paper will apply the following methods: Fisher discriminant analysis, ReliefF algorithm, two sample *t*-test, Kolmogorov–Smirnov test, Kruskal–Wallis test, stepwise regression method, feature correlation with a class and SVM recursive feature elimination. These methods rely their operation principle on different foundations and thank to this allow to look on the selection problem from different points of view.

2.1. Fisher discriminant analysis

In Fisher discriminant analysis the greatest weight is assigned to feature which is characterized by a large difference of the mean values in two studied classes and a small value of standard deviations within each class. The two class discrimination measure of the feature f is defined in the form (Duda, Hart, & Stork, 2003; Guyon & Elisseeff, 2003):

$$S_{12}(f) = \frac{|c_1 - c_2|}{\sigma_1 + \sigma_2} \tag{1}$$

where c_1 and c_2 represent the mean values for classes 1 and 2, respectively, while σ_1 and σ_2 are the appropriate standard deviations. A large value of $S_{12}(f)$ indicates good class discriminative ability of the feature.

2.2. ReliefF algorithm

The ReliefF algorithm ranks the features according to its the highest correlation with the observed class while taking into account the distances between opposite classes (Robnik-Sikonja & Kononenko, 2003). The main idea of the ReliefF algorithm is to estimate the quality of the features according to how well their values distinguish between observations that are near to each other. ReliefF selects randomly an instance R_i of observation and then searches for k of its nearest neighbors from the same class, called nearest hits H_j and also k nearest neighbors from each of the different classes, called nearest misses $M_j(C)$. It updates the quality estimation $W(A)$ for all attributes A depending on their values for R_i, hits H_j and misses $M_j(C)$. If instances R_i and H_j have different values of the attribute A then the attribute A separates two instances with the same class which is not desirable. So the quality estimation $W(A)$ is decreased. If instances R_i and M_j have different values of the attribute A then this attribute separates two instances of different class values which is desirable. So the quality estimation $W(A)$ is increased. The algorithm averages the contribution of all hits and misses. The detailed description of the procedure can be found in Robnik-Sikonja and Kononenko (2003).

2.3. Two-sample t-test

The next used selection method is a two-sample Student *t*-test. One explicit assumption of *t*-test is that each of two compared populations of genes (autism and controls) should follow a normal distribution. Checking the condition of normality distribution of genes in our data base we found that in about 80% cases it was fulfilled. The null hypothesis of *t*-test is that data in the class 1 and 2 are independent random samples of normal distributions with equal means and equal but unknown variances against the alternative hypothesis that the means are not equal. The test statistic is formulated in the form

$$t = \frac{c_1 - c_2}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \tag{2}$$

where and n and m represent the sample sizes of both classes (Sprent & Smeeton, 2007).

Two sample *t*-test is implemented in MATLAB as *ttest2* function (Matlab user manual – statistics toolbox, 2013). The test result returns h, which is equal 1 or 0. The value of 1 indicates a rejection of the null hypothesis at the 5% significance level, while $h = 0$ indicates a failure to reject the null hypothesis at the same significance level. The function returns also the p-value of the test. Low value of p indicates that the compared populations are significantly different.

2.4. Kolmogorov–Smirnov test

The other statistical feature selection method applied in the research was the Kolmogorov–Smirnov (KS) test. It compares the medians of the groups of data to determine if the samples come from the same population (Sprent & Smeeton, 2007). The null hypothesis is that both classes are drawn from the same continuous distribution. The alternative hypothesis is that they are drawn from different distributions. The KS test statistic is based on the relation

$$KS = \max(|F_1(x) - F_2(x)|) \tag{3}$$

where $F_1(x)$ and $F_2(x)$ are the cumulative distribution of samples of feature f belonging to class 1 and 2. High value of this coefficient indicates that the feature has good class discriminative ability. On the other hand, a small value of this factor indicates that feature should be rejected at the selection stage. Fig. 1 illustrates the results of KS test for two types of genes: the significant one and the randomly chosen, assessed as a non-significant.

In the case of significant gene (Fig. 1(a)) the p value was equal 0.0039 and $KS = 0.2881$. In the case of not significant gene we got $p = 0.3374$ and $KS = 0.1536$.

2.5. Kruskal–Wallis test

In this method medians of the samples are compared, but test uses ranks of the data rather than the numeric values (Sprent & Smeeton, 2007). It finds ranks by ordering the data from the smallest to the largest across all groups and taking the numeric index of this ordering. The Kruskal–Wallis test does not make any assumptions about normality. It returns the p value for the null hypothesis that all samples are drawn from the same population.

2.6. Stepwise regression method

Stepwise regression is a systematic method for adding and removing features to the set of input attributes based on their statistical significance in a regression. The method begins with an initial model and then compares the explanatory power of incrementally larger and smaller models. At each step, the p value of F-statistics (Sprent & Smeeton, 2007) is computed to test models with and without selected feature. Based on the statistic result algorithm makes a decision whether feature should be included in a model or not. If a feature is not currently in the model, the null hypothesis is that the term would have a zero coefficient if added to the model. If there is sufficient evidence to reject the null hypothesis, the feature is added to the model. Conversely, if a feature is currently in the model, the null hypothesis is that the term has a zero coefficient. If there is insufficient evidence to reject the null hypothesis the term is removed from the model.

Presented method may build different sets of features, depending on the initial model and order of adding and removing features from the set of attributes. Considering that outcomes may not be reproducible the stepwise regression method provides locally optimal result.

2.7. Feature correlation with class

In this method, the direct correlation of the feature values with a class is examined. The discriminative value $S(f)$ of the feature f for recognizing one class from the other K classes is defined as follows (Guyon & Elisseeff, 2003; Wilinski & Osowski, 2012):

$$S(f) = \frac{\sum_{k=1}^{K} P_k (c_k - c)^2}{\sigma^2(f) \sum_{k=1}^{K} P_k (1 - P_k)} \tag{4}$$

where c is a mean value of feature for all data, c_k is a mean value of the feature for the kth class data, $\sigma^2(f)$ is a variance of feature, P_k is a probability of kth class occurrence in dataset (the uniform distribution is usually assumed). The large value of $S(f)$ indicates good discriminative ability of feature f for recognition of the particular class from the other K classes. In this paper the number of classes is $K = 2$.

2.8. SVM recursive feature elimination

SVM network can be configured for solving selection problem in the form of recursive feature elimination (SVM-RFE) (Guyon et al., 2002). In this approach SVM network with linear kernel is used. Network is learned applying all available features used simultaneously as an input attributes. The sign function is added for matching the input values to the appropriate class label. The output y at presentation of the features organized in the form of vector $\mathbf{f}$ is defined by the following equation

$$y(f) = \mathrm{sgn}(u) = \mathrm{sgn}(\mathbf{w}^T \mathbf{f} + b) \tag{5}$$

where $\mathbf{w} = [w_1, w_2, \ldots, w_n]^T$ is the weight vector, $\mathbf{f} = [f_1, f_2, \ldots, f_n]^T$ is a vector of features and b is a bias. Large absolute value of weight connecting feature f with the network denotes strong ability of this feature to distinguish two classes.

In SVM-RFE approach to feature selection, the features are eliminated step by step according to the assumed criterion related to their support in the discrimination of the classes. The SVM is retrained using smaller and smaller population of features. In each step the features associated with the smallest absolute weights are eliminated. In this approach we eliminate 20% of the actual number of genes. The process is repeated until the appropriate number of the most important features is obtained.

2.9. Fusion of selection methods

The results of the separate selection processes are combined together to perform the second step of selection, which lead to

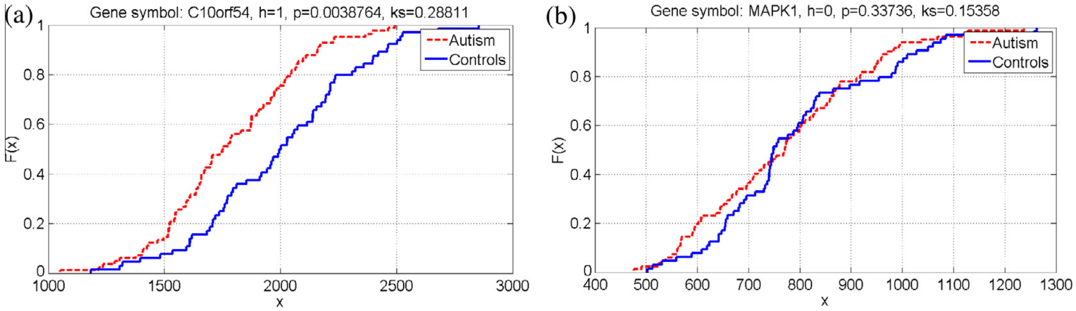


Fig. 1. The cumulative distribution functions for two classes (autism and controls) for the significant gene (a) and for the not significant gene (b) in Kolmogorov–Smirnov test.

the final quasi optimal ranking of genes. We expect the genes of two classes be located in two clusters (the autism class and reference class), each gathering the cases similar to each other. The cluster purity with respect to the class membership, analyzed at different size of the best gene subsets, should indicate the optimal, representative number of genes for recognition of autism.

The results of a single analysis are not representative for the problem, because of small population of available data. To get reliable results of selection the individual methods will be repeated many times on a randomly chosen samples of the original data set. In our experiments we have performed each selection method 10 times by using 60% of the randomly selected rows of the data set. In each run we observe the position of particular gene among the others. The final position of the gene depends on the sum of its positions in all runs. We assign the global weight $w(f)$ to each gene. The following formula has been used

$$w(f) = \sum_{i=1}^{K}\sum_{j=1}^{N_r} w_{ij}(f) \qquad (6)$$

The index K is the number of applied selection methods ($K = 8$ in this research), N_r – the number of repeated runs of selection, w_{ij} is the position of genes in ith method of selection and jth run. The genes are arranged according to the decreasing values of the global weight. The best gene is the one of the smallest value.

3. Numerical results of gene selection

3.1. Materials

The numerical experiments of gene selection have been done using the dataset related to the autism. The database is publicly available and was downloaded from GEO (NCBI) repository (NCBI data base, 2011). Number of observations in this dataset equals 146 and number of genes 54,613. The database consists of two classes: the first one is related to children with autism (number of such observations $n = 82$) and the second to the control group of healthy children ($n = 64$). Blood draws for all subjects were done between the spring and summer of 2004. Total RNA was extracted for microarray experiments with Affymetrix Human U133 Plus 2.0 39 Expression Arrays. Our main task is to find a small subset of genes with good class discriminative abilities. This problem is resolved by using several gene selection methods combined into one final system.

3.2. The main stages of experiments

This section describes the numerical experiments of gene selection for the autism data. We will present the results concerning the selection, clusterization and PCA transformation. In the introductory phase of experiments we have excluded the genes of very similar values of expression means for both classes. The features, for which the ratio of the means in both classes was above 0.98, were removed from the base. In this way the number of genes was reduced from 54,613 to 16,230.

In the next stage eight feature selection methods were applied to discover the importance of the genes and their order. The selection procedure has been repeated 10 times on randomly selected 60% of the available observations. The positions of each gene was noted in each run and then summed up. On this basis the genes have been arranged in a sequence starting from the best to the least significant. In the following stage the most significant genes have been fused into one common set of the reduced dimension. To find the optimal population size of the most significant genes we have applied the cluster purity. The final outcome of the clusterization will be illustrated and examined using PCA transformation (Haykin, 2000).

3.3. Comparative results of selection methods

Feature selection methods described in the previous sections have been applied to get the order of genes, sorted in a decreasing fashion. The following abbreviations are applied: ***FDA*** – the Fisher discriminant analysis, ***RFA*** – the ReliefF algorithm, ***TT*** – the two-sample t-test, ***KST*** – the Kolmogorov–Smirnov test, ***KWT*** – the Kruskal–Wallis test, ***SWR*** – the stepwise regression method, ***COR*** – the feature correlation with a class, ***SVM*** – the SVM-RFE method.

We compare the selection results on the basis of 100 the most relevant genes chosen in each method. As was expected the methods have selected different contents of the best genes. Table 1 shows how many identical genes among the first 100 of the most important have been selected by different methods.

The contents of the best selected sets differ from method to method. Analyzing them we found that few methods identified a large percentage of the same genes. For example the correlation feature with a class and t-test produced exactly the same sets of genes. Kruskal–Wallis test has found 63% of genes which were identical with the COR and TT tests. On the other hand some of the methods have resulted in very different sets, i.e., stepwise regression and hypothesis test outcomes are overlapping only in 1% or 2%. Very low is the agreement of SVM-RFE results and all other methods (only one common gene among the first 100).

The quality of the selection processes has been checked by analyzing the expression profiles of the best identified genes for the opposite classes. Fig. 2 illustrates the exemplary expression levels of the patients for the most important gene selected by the Fisher and SVM-RFE methods. As we can see in both cases the mean value of the observations belonging to the autism class differs significantly from the reference class. At the same time we observe large variability of the gene expressions for the subsequent observations.

In the case of the Fisher method (Fig. 2(a)) we got the mean equal 232.41 ± 140.38 (autism) and 144.59 ± 97.81 (control group). The difference of the mean values is 87.82. For the best gene in SVM-RFE method (Fig. 2(b)) we got 390.06 ± 128.21 (autism) and 342.23 ± 67.36 (control group). The difference of the mean values in this case is 47.83. For comparison, the differences measured for the least significant gene were equal 1.20 in Fisher and 0.89 in SVM-RFE method.

The next step is fusing the results of the individual methods into one common outcome. To find the best genes which are the most representative for all analyzed methods we have assigned the global weight $w(f)$ to each gene according to the formula (6). Analyzing the contents of all sets selected by different methods we have found 501 different genes among the first 100 selected in each method.

Table 1
The redundancy rate achieved by different algorithms among the top 100 genes selected by different methods.

	FDA	RFA	TT	KST	KWT	SWR	COR	SVM
FDA	100	9	17	10	12	3	17	1
RFA	9	100	44	34	45	4	44	1
TT	17	44	100	31	63	2	100	1
KST	10	34	31	100	45	1	31	1
KWT	12	45	63	45	100	1	63	1
SWR	3	4	2	1	1	100	2	1
COR	17	44	100	31	63	2	100	1
SVM	1	1	1	1	1	1	1	100

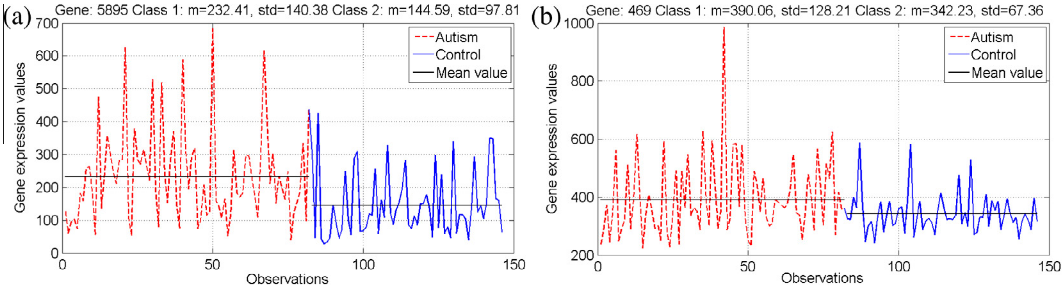


Fig. 2. Expression levels for the best genes selected by: (a) Fisher method, (b) the SVM-RFE method.

3.4. Clusterization of gene space

Good way for assessing the quality of the selected genes is to cluster the data in a multidimensional space. The optimal set of genes should provide the clusters of highest purity with respect to the class membership. Different approaches to clustering are possible: K-means, fuzzy c-means or expectation maximization algorithm (Tan, Steinbach, & Kumar, 2006). In this work we applied the simplest K-means. It is a method of vector quantization that aims to partition n observations into K clusters ($K < n$). Each N-dimensional observation belongs to the cluster with the nearest mean (centroid) which serves as a prototype of the cluster.

This aim is achieved by minimizing the squared sum distances between centroids and the vectors within each cluster (Tan et al., 2006). K-means can be developed in two approaches: off-line and on-line. We use the off-line Matlab version with batch updates. Each iteration consists of reassigning all points to their nearest clusters, followed by recalculation of the cluster centroids.

In our problem the number of clusters is equal two ($K = 2$), the same as the number of investigated classes. The K-means will be used by us to find the number of the most significant genes, which provide the highest class purity of the clustered space. The procedure consists of repeating the K-means algorithm at varying number of genes. In each step we increase this number by one. The clusters are assessed by comparing their purity index (Tan et al., 2006), defined as follows

$$p_i = \max \frac{n_{ij}}{n_i} \tag{7}$$

In this definition n_{ij} is a number of observations of jth class inside of ith cluster and n_i is a number of observations forming the ith cluster ($i, j = 1, 2$). In the next step the total purity index p of the clustered space is determined

$$p = \sum_{i=1}^{K} \frac{n_i}{n} p_i \tag{8}$$

where K is the number of clusters. In this way we can calculate the total purity of the clustered space at varying dimension (number of the most significant genes) of the representative vectors. According to the results presented in Table 1 the TT and COR methods have produced the same results. Therefore, to avoid their domination, only one of them has been taken into account in a fusion process.

Fig. 3 presents the change of the total purity index versus the number of the most relevant genes found in the final fusion procedure. We can observe that the best purity is obtained for the top 32 features. Its value in our experiments was equal 0.83.

The best result obtained at application of the fusion approach has been compared to the outcomes of 8 individual selection methods. Table 2 presents the highest values of the total purity index for all investigated selection methods and the population of genes at which these maxima happened.

We can notice that total purity of the clustered space at application of the investigated methods differs significantly. Moreover, the highest purity is obtained at different number of genes. For instance, in the ReliefF method the best purity occurs for 13 the most significant genes, whereas in the Fisher algorithm for 68 (the highest number). The best purity of clusterization corresponds to the fusion approach at presence of 30 the best genes. However, the same purity value has been obtained for TT/COR methods. Comparing the contents of TT/COR and fusion results we have found 22 identical genes. It seems that this selection method dominated in fusion.

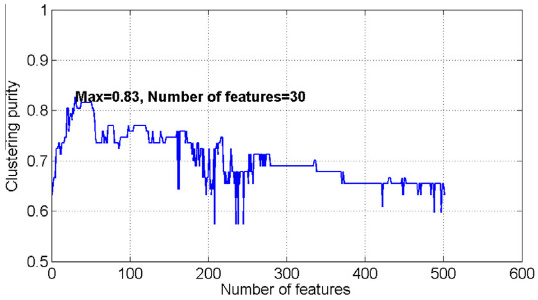


Fig. 3. Total purity index of clustered space versus number of the most significant genes.

To illustrate the results in a graphical form we have presented the expression levels of the selected genes in the form of image. Fig. 4(a) shows the image of the expression profiles for the top thirty-two genes selected by fusion in the form of the colormap of hot. The vertical axis represents observations and the horizontal – the genes arranged according to their importance. There is a visible border between 82 observations of the autism group and the remaining 64 representing the reference one. For comparing purposes the image of the expression profiles for 30 genes chosen randomly from the base is presented in Fig. 4(b). There is a significant difference between both images, which confirms good performance of the proposed selection procedure.

3.5. Illustration of selection results using PCA

The next study is concerned with the graphical representation of the multidimensional observation vectors using the Principal Component Analysis (PCA). PCA is a statistical method mapping the original vectors $\mathbf{x}$ from the high dimensional space to vectors $\mathbf{y}$ in the space of the reduced dimension (Haykin, 2000). The transformation is done through the linear relation

$$\mathbf{y} = \mathbf{W}\mathbf{x} \tag{9}$$

in which the transformation matrix $\mathbf{W}$ is formed from the chosen number of eigenvectors corresponding to the largest eigenvalues of the correlation matrix of the original data $\mathbf{x}$.

We have mapped the multidimensional observations into 2-dimensional space formed by two the most important principal

Table 2
The highest values of the total purity index corresponding to the set of genes selected individually by different methods and after their fusion.

Method	FDA	RFA	KST	KWT	SWR	COR/TT	SVM	Fusion
Purity	0.67	0.76	0.77	0.80	0.63	0.83	0.60	0.83
Number of genes	68	13	44	23	15	24	42	30

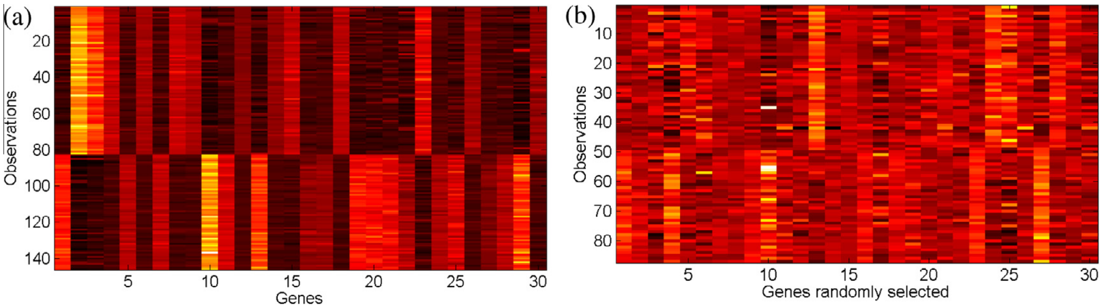
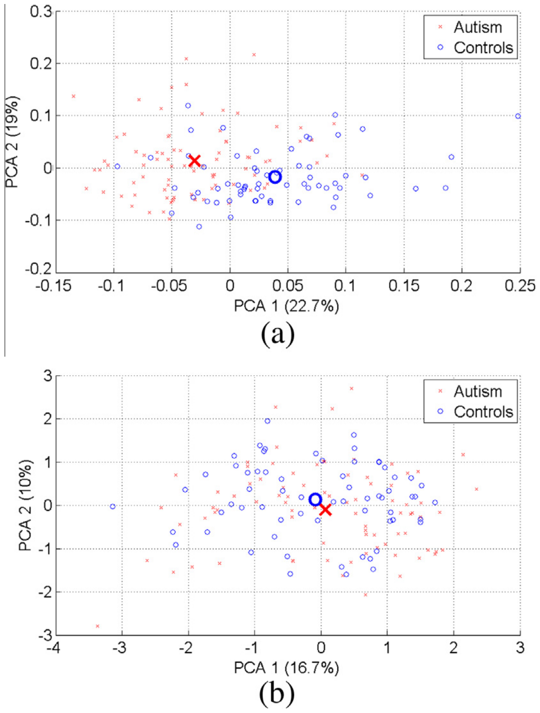


Fig. 4. The colormap of the expression profiles for 30 most significant genes selected by the fusion procedure (a) and for 32 randomly chosen genes (b).

components. Two cases have been investigated. In the first approach the original vectors contained only 30 genes selected by the fusion procedure. Fig. 5(a) depicts the case in which we use only the best representative genes in the vectors **x**. For comparison we have repeated PCA on the full size original 16,230 element vectors containing all genes. The graphical results of the sample distribution are presented in Fig. 5(b). Large bold symbols of the circle and **x** represent the centroids of the data belonging to two classes.

We can observe a significant difference in distribution of samples belonging to both classes. In the first case (Fig. 5(a)) we see two compact regions, in which there is a domination of one class. In the second case (Fig. 5(b)) the situation is different. The observations belonging to different classes interlace each other. It is practically impossible to separate them into two regions. In both cases we see large dispersion of information. For the best 30 selected genes the first and second principal components contain 22.7% and 19.8% of total information, respectively. In the case of all genes these values are much smaller and equal 16.7% (the first component) and 10% (the second component).



(a)

(b)

Fig. 5. The distribution of the two-class samples mapped on two the most important principal components at representation of vectors **x** by 30 most significant genes (a) and at application of all genes (b).

To characterize the dispersion of data in the numerical way the average Euclidean distance between the observations and their respective centroids have been computed. Table 3 shows these values, which represent the average distances and their standard deviations in both investigated cases, calculated for their 2-dimensional mapping provided by PCA.

The results show that the total dispersion in the first case (30 top genes forming the vectors **x**) is much smaller than in the second one (application of all genes). At the same time we can observe in Fig. 5 the significant difference of distances between the centroids representing both classes of data in the 2-dimensional space. At representation of data by 30 genes this distance related to the maximum range of data was equal 0.226. In the second case it was only 0.034 (both centroids occupy similar positions).

4. Classification system for autism prediction

4.1. Applied classifiers

The genes selected in the previous phase can be used for classification of the microarray data. The task in this part of research is recognition of two classes: class 1 – autism and class 2 – control (healthy) subjects. To get the most reliable results we apply the ensemble of classifiers composed of the Support Vector Machine (SVM) of Gaussian kernel (Schölkopf & Smola, 2002) supplied by different sets of features, which were selected by various methods. The classification results of the ensemble members will be integrated by using random forest (RF) of decision trees.

The SVM belongs to the best binary classifiers. It was developed by Vapnik is a linear machine, working in the high dimensional

Table 3
The average relative distances of samples from their centroids and their standard deviations for the top 30 genes and all genes.

Number of genes	Autism	Reference group
Top 30 genes	0.08 ± 0.04	0.09 ± 0.06
All genes (16,230)	1.38 ± 0.79	1.13 ± 0.86

feature space formed by the non-linear mapping of the N-dimensional input vector $\mathbf{x}$ into a L-dimensional feature space ($L > N$) through the use of a kernel function $K(\mathbf{x},\mathbf{x}_i)$. The learning problem of SVM is formulated as the task of separating the learning vectors into two classes of the destination values either $d_i = 1$ (one class) or $d_i = -1$ (the opposite class), with the maximal separation margin. The SVM of the Gaussian kernel has been used in our application. The hyperparameters (the regularization constant C and Gaussian kernel width) have been adjusted by repeating the learning experiments for the set of their predefined values and choosing the best one on the validation data sets.

The Breiman random forest (Breiman, 2001) is an ensemble learning method for classification that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes output by individual trees. To improve the generalization property the randomness in selecting the learning data is implemented. The trees are grown using randomly selected input variable sets for each node.

Random forest will be used by us to integrate the results of 8 different classifiers into the final outcome. The output signals of these 8 units form the input attributes for the random forest, which is responsible for generating the final outcome.

4.2. The results of classification

The experiments of selection and classification have been performed on the separate parts of data in order to get the most objective results. Sixty percent of the data base (chosen randomly) has been used in gene selection and the rest (40%) in the classification only. This split was repeated 10 times and then followed by the gene selection and classification stages. Such approach to the selection and classification allows separating these two phases of data analysis and making them data independent. Repeating these experiments 10 times at different compositions of observations allows drawing more objective assessment of the efficiency of the proposed method. Moreover, thanks to 10-fold cross validation procedure the whole data set takes parts in both phases of experiment.

The applied selection methods have resulted in a particular order of the genes from the most to the least significant. The important step is to determine the optimal number of them used as the input attributes for the classifiers. It is known fact that classification using the limited number of genes, following from the purity of clusters, results in not satisfactory accuracy (Tan et al., 2006; Wilinski & Osowski, 2012). The less important genes following directly the best group, have also some positive impact on recognition accuracy. Therefore, the additional step for determining the optimal population size of the best genes should be made. We have done it by checking the classification accuracy achieved by increasing step by step the number of genes used as the input attributes.

Fig. 6 shows the exemplary results representing the error rate of classification at increasing number of the best selected genes for the Fisher method. The continuous curve presents the actual error rate of class recognition, which was obtained by SVM classifier at differing number of genes. The discrete peaks at the bottom of the figure represent the values of standard deviation for the following equal size ranges of the number of genes. The depicted results refer to 10 equal size ranges extending from the first to the last 100th gene. The numbers located above these peaks show the mean error rate in the appropriate range. The smallest mean error rate of the value 0.21 was observed in the range corresponding to the number of genes extending from 35 to 55. The mean of these two numbers of genes (45) has been used in the 10-fold cross validation of the classification experiments. Similar procedure of determining the optimal number of genes for each selection method has been repeated. As a result of such procedure we got 8 different sets of input attributes for the applied SVM classifiers. In further experiments we have used them in 10-fold cross validation procedure.

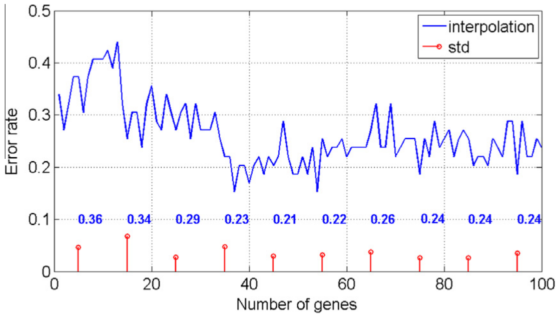


Fig. 6. The error rate of class recognition at changing number of genes used as the input attributes for SVM (the results of testing on the data set not taking part in learning). The upper curve depicts the averaged error in 10 run cross-validation experiments at changing number of genes. The numbers below the curve represent the moving average values of error calculated for the window of 10. The vertical stripes at the bottom represent values of standard deviation of error for different ranges of the number of genes.

Each run of the cross validation procedure applied to 8 classifiers was followed by the integration of their results into the final outcome. This task was performed by the random forest network (Breiman, 2001). Fig. 7 shows the statistical importance of the applied methods in forming the final results by the random forest integrator. Fisher and ReliefF methods have been found the most important in taking the final classification decision by the RF integrator. On the other hand the least important were Kruskal–Wallis results.

To check the importance of genes selected in all methods we have replaced the first 20 best genes in the set by the next 20, while preserving the same number of genes in the input set. Additional investigations have been also done for the same number of genes chosen in a random way from the base.

Table 4 presents the averaged class recognition accuracy of the individual classification systems in the 10-fold cross validation procedure. The columns represent the solutions for different selection methods. The last column presents the final accuracy achieved after fusion, which was done by the random forest. The numbers within the table represent the mean values and standard deviations obtained in 10 independent runs of the selection and classification procedures.

The final class recognition accuracy obtained by using the random forest as an integrator is the best. The classification accuracy has been increased from the best 70.16 ± 5.38% (FDA) to 78.43 ± 2.66% after integration. Not only the accuracy of recognition was increased but also standard deviation was reduced almost two times.

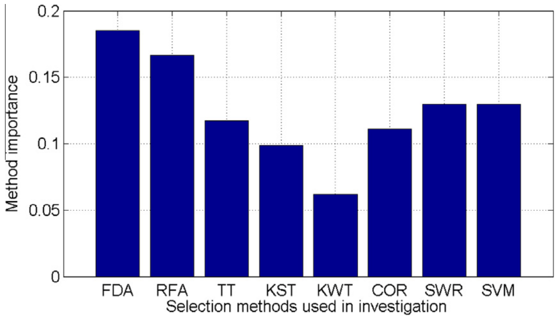


Fig. 7. The relative importance of the selection methods in the final integration stage of the ensemble.

Table 4
The averaged class recognition accuracy and the standard deviations of the SVM classifier supplied by the set of genes selected in different methods (all values in percent).

	FDA	RFA	TT	KST	KWT	COR	SWR	SVM	Fusion
Optimal number of the best genes	70.16 ± 5.38	58.35 ± 7.02	59.92 ± 6.11	61.68 ± 5.61	60.22 ± 5.61	59.92 ± 6.11	60.60 ± 6.89	59.73 ± 6.87	78.43 ± 2.66
20 The best genes substituted by the less important	68.10 ± 7.39	57.48 ± 7.13	58.74 ± 6.63	60.56 ± 5.62	58.11 ± 6.76	58.74 ± 6.63	58.84 ± 7.31	57.48 ± 5.82	76.90 ± 3.56
Random choice of genes	55.12 ± 4.65	51.36 ± 10	54.72 ± 7.15	54.19 ± 4.77	51.01 ± 3.45	55.28 ± 5.61	53.67 ± 3.59	55.22 ± 5.16	67.16 ± 2.09

The results of the final recognition will be also depicted in the form of a confusion matrix representing the average results of all 10 runs of the system (the results correspond to the testing data). The rows represent the percentage of the real class membership of the data and the columns the results of classification. The diagonal entries ($i = j$) depict the percentage of properly recognized classes. Each entry outside the diagonal represents the percentage of the misclassified cases. The entry in the (i,j)th position of the matrix means false assignment of ith class to the jth one. Table 5 presents the average results of class recognition.

The weighted average value of the relative error of class recognition was equal 20.19%. In this weighting the population sizes of both classes have been taken into account.

To compare the importance of the gene ranking we have repeated 10 times the classification procedure at randomly chosen composition of 30 genes. We have used 30 genes because this number was most often taken by our selection methods in previous experiments. The obtained results in the form of confusion matrix are depicted in Table 6. This time the weighted averaged misclassification rate is very high (33.90%).

It is evident that application of the best selected genes in the representation of the samples provides the highest accuracy (the least relative error) in all experiments.

The accuracy measure treats every class as equally important. In medical practice it is not sufficient to assess the method in an objective way. The important aspect is associated with the importance of recognition of autism cases (called here true positive – *TP*) from the normal cases (true negative – *TN*). By the symbol *FN* we understand the number of autism falsely recognized as healthy and by *FP* the healthy cases recognized as the autism. On the basis of these notations four quality measures are defined (Tan et al., 2006).

The first one is the true positive rate (*TPR*), called also sensitivity, defined as the fraction of all positive examples predicted correctly by the classifier $TPR = TP/(TP + FN)$. 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 other measures include false alarm rate (*FA*) defined as the ratio of the negative cases recognized by the classifier as the positive $FA = FP/(FP + TN)$ and the false negative rate, defined as $FNR = FN/(TP + FN)$. Table 7 presents the values of these quality measures when applied to the recognition of autism cases from the healthy ones for the genes selected by us, and for the randomly chosen 30 genes.

The sensitivity and specificity values of the proposed selection and classification system have assumed much higher levels in comparison to the randomly selected genes. These results confirm the superiority of our method over the random choice of genes.

Table 5
The confusion matrix of the class recognition results at application of the best genes after integration.

	Class 1	Class 2
Class 1	0.82	0.18
Class 2	0.23	0.77

Table 6
The confusion matrix of classification results by applying 10 randomly selected genes (after integration).

	Class 1	Class 2
Class 1	0.69	0.31
Class 2	0.39	0.61

Table 7
The values of the quality measures in recognition of autism cases from the healthy ones for the best genes selected by our method and for the randomly chosen 30 genes.

	TPR	TNR	FNR	FA
Best selected genes	0.82	0.77	0.18	0.23
30 Random genes	0.70	0.62	0.30	0.38

5. Conclusions

The paper has examined several data mining methods cooperating in an ensemble for selecting the most important genes in the expression microarray of autism. The most relevant genes have been selected using two stage approach. In the first step we apply eight different feature selection methods working independently. The final set is identified by fusing all obtained subsets in the second step of the procedure.

The expression levels of the selected genes have been analyzed. We applied different tools and methods, including the clusterization of the data combined with the measures of its quality, principal component analysis and statistical characterization of the clustered space.

The selected genes have been used as the input attributes for the classifiers, which were responsible for recognition of autism cases from the reference ones. To get the most reliable results we have applied the ensemble of classifiers composed of SVM classification units and random forest fulfilling the role of integrating unit. These classifiers have been supplied by different sets of features selected by various methods. The obtained results confirmed good performance of such selection and classification system.

The main contribution of the paper in the research of autism is simultaneous application of many selection methods, based on different principle of operation and fusing them into final system providing better recognition of autism cases from the reference class. We have proposed special method of assessing the quality of gene, which takes into account its ranking position in many runs of selection.

In the second phase of study we have proposed the novel solution of two-stage classification system. In the first stage we use the SVM classifiers supplied by the sets of genes selected by different methods. Their results are combined together into final decision in the second stage by random forest of decision trees. In this way we are able to use many selection results to improve the accuracy of autism recognition.

The analysis of the performance of different steps of our approach leads to the interesting observations. The individual methods analyze the problem from different point of view and find the optimum from this point. Usually the solutions are different and depend on the particular choice of observations used in the processing. Our two-step procedure of gene selection tries to find the genes, which have been chosen as the most significant in all runs of different methods. In this way we increase the probability of selecting the genes, which are the best associated with all observations.

In the classification stage we use once again the results of individual gene selections. This time we form many sets of the best genes created by different methods and use them as the input attributes in the ensemble of SVM classifiers. The results presented in Table 4 depict the difference between the performance of the classifiers. The individual results change a lot (the accuracy of class recognition was changing from 58% to 70%). However, combining them into one final result by using random forest classifier has increase this accuracy to above 78%. The random forest classifier has made proper use of the information delivered by different methods of gene selection. The experiments of reducing the number of SVM classifiers by neglecting some selection methods led to.

Our study presented in this paper needs to be continued in few directions. First, more data bases related to autism should be examined. The important point is also to increase the number of selection methods. Our introductory experiments have shown that more methods result in more accurate recognition. Additional study is needed to find the alternative methods for determining the optimal number of input attributes (the best selected genes). The natural candidate is application of genetic algorithm, which was successful in selecting the optimal set of feature in some other recognition problems (Siroic, Osowski, Markiewicz, & Siwek, 2009). The interesting is also developing the alternative ways of assessing the discriminative ability of the selection methods, which might be competitive to ranging the positions, applied in this solution. The natural candidate is using ranking by frequency, in which the order of genes is arranged according to their frequency of appearance within the defined set of the best genes in each run.

References

Alonso-González, C. J., & Moro-Sancho, Q. I. (2012). Microarray gene expression classification with few genes: Criteria to combine attribute selection and classification methods. *Expert Systems with Applications, 39*, 7270–7280.

Alter, M., Kharkar, R., Ramsey, K., Craig, D., Melmed, R., Grebe, T., et al. (2011). Autism and increased paternal age related changes in global levels of gene expression regulation. *Plos One, 6*, 1–10.

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

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32.

Chuang, L., Yang, C., Wu, K., & Yang, C. (2011). Gene selection and classification using Taguchi chaotic binary particle swarm optimization. *Expert Systems with Applications, 38*, 13367–13377.

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

Duda, R. O., Hart, P. E., & Stork, P. (2003). *Pattern classification and scene analysis*. New York: Wiley.

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

Esteban, F., & Wall, D. (2011). Using game theory to detect genes involved in autism spectrum disorder. *Top, 19*(1), 121–129.

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., et al. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science, 286*, 531–537.

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

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

Haykin, S. (2000). *Neural networks, a comprehensive foundation*. New York: Macmillan College Publishing Company.

Hu, V., & Yinglei, L. (2013). Developing a predictive gene classifier for autism spectrum disorders based upon differential gene expression profiles of phenotypic subgroups. *North American Journal of Medicine and Science, 6*(3), 107–116.

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

Kumar, P. G., Victoire, T. A., Renukadevi, P., & Devaraj, D. (2012). Design of fuzzy expert system for microarray data classification using a novel genetic swarm algorithm. *Expert Systems with Applications, 38*(2), 1811–1821.

Matlab user manual – statistics toolbox. (2013), Natick, USA: MathWorks.

Mitsubayashi, H., Aso, S., Nagashima, T., & Okada, Y. (2008). Accurate and robust gene selection for disease classification using a simple statistics. *Biomedical Informatics, 391*, 68–71.

NCBI data base. (2011). <http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS4431>.

Robnik-Sikonja, R., & Kononenko, I. (2003). Theoretical and empirical analysis of ReliefF and RReliefF. *Machine Learning, 53*, 23–69.

Schölkopf, B., & Smola, A. (2002). *Learning with kernels*. Cambridge MA: MIT Press.

Siroic, R., Osowski, S., Markiewicz, T., & Siwek, K. (2009). Application of support vector machine and genetic algorithm for improved blood cell recognition. *IEEE Transactions on Instrumentation and Measurement, 58*(2), 2159–2168.

Sprent, P., & Smeeton, N. C. (2007). *Applied nonparametric statistical method*. Boca Raton: Chapman & Hall/CRC.

Tan, P. N., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining*. Boston: Pearson Education Inc.

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

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

Woolf, P. J., & Wang, Y. (2000). A fuzzy logic approach to analyzing gene expression data. *Physiological Genomics, 3*, 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.

Yang, M. S., & Gill, M. (2007). A review of gene linkage, association and expression studies in autism and in assessment of convergent evidence. *International Journal of Developmental Neuroscience, 25*, 69–85.