Tomasz LATKOWSKI[1], Stanisław OSOWSKI[1,2]

Military University of Technology (1), Warsaw University of Technology (2)

Feature selection methods in application to gene expression: autism data

Abstract. *The paper presents the application of several different feature selection methods for recognizing the most significant genes and gene sequences (treated as features) stored in dataset of gene expression microarray related to autism. The outcomes of each method have been examined by analyzing gene expression profiles of selected genes. In the next step fusion of the most relevant features selected by different methods, has been implemented. The optimal number of features has been defined as the set providing the best clustering purity.*

Streszczenie. *Praca prezentuje badanie wybranych metod selekcji cech diagnostycznych w celu wyodrębnienia najbardziej znaczących sekwencji genowych z mikromacierzy ekspresji genów dotyczącej autyzmu. Dla wyselekcjonowanych cech przeanalizowano wartości poziomów ekspresji genów. W kolejnym etapie dokonano fuzji wyselekcjonowanych cech. Optymalny zbiór cech wyznaczono na podstawie czystości przestrzeni klasteryzacji. (**Metody selekcji cech diagnostycznych w zastosowaniu do ekspresji genów: baza danych autyzmu**).*

Keywords: feature selection, gene expression microarrays, clustering, autism.
Słowa kluczowe: selekcja cech, mikromacierze ekspresji genów, klasteryzacja, autyzm.

doi:10.12915/pe.2014.08.47

Introduction

Gene expression microarray is a sophisticated technique used in molecular biology. DNA microarrays are mainly applied for analyzing the expression of genes in specific cells at given time and under certain conditions [4]. This technique generates a huge number of features (genes and gene sequences) which create a large information dataset. The main problem from the data mining point of view 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. Figure 1 shows the organization scheme of the DNA microarray.

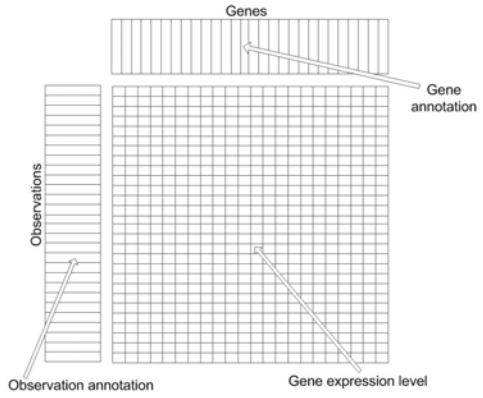

Fig.1 . DNA microarray organization in the form of matrix.

Each row in the above figure corresponds to one observation (patient) and each column to one gene or gene sequence (feature). Because of the large imbalance of the number of genes and patients the selection is an ill conditioned problem. Moreover, data stored in medical databases are typically noisy and some gene sequences have large variance [19]. It makes the gene selection in DNA microarrays very 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 [2,9-11,13,18-22] starting with pioneering Golub investigation in 1999 [7]. This area of science is being studied due to the fact that is strictly connected with early disease prediction (especially in tumors). It is thought that some diseases, i.e., autism, breast cancer, leukemia, etc., have their reflections in genetic changes [1]. 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. In the next step these selected markers can be observed in order to understand or find association with the disease.

Present data mining methods allow to look at the gene selection problem from many points of view. It is known that the selection algorithms may provide various results for different datasets (diseases) [19]. It makes selection issue more complicated due to the fact that the model developed for one problem may not be suitable for another dataset.

Many publications have examined gene selection problem using only one method of ranking. If we take into account that DNA microarrays create ill conditioned problem such approach may not provide the globally optimal result. Moreover, examining different methods on one dataset we can observe various outcomes [20].

In this paper we follow the direction pointed in [20] and present more complex approach based on using several methods simultaneously. The results of individual selection methods are fused, leading to the final set of genes. The application of several methods gives opportunity to look at the selection problem from different points of view. In this way we increase the probability of proper selection of the most important genes.

Developed model is verified on the publicity available database related to autism. Expression levels of the selected genes have been examined and compared to the randomly chosen ones. The cluster purity idea has been applied for obtaining the optimal number of the most relevant genes. The results of selection have been illustrated in a graphical way using the PCA mapping.

Dataset description

This paper presents the problem of gene selection applied to the dataset related to the autism. The database is publicity available and was downloaded from GEO (NCBI) repository [23].

Autism is a severe neurodevelopmental disorder with characteristic social and communication deficits and ritualistic or repetitive behaviors [1]. Many etiologies have been suggested and numerous risk factors identified. Autism is associated with a high degree of heritability. Few specific genetic mutations have been identified accounting for a minority of cases [1], while the majority of cases are considered sporadic.

Number of observations in this dataset equals 146 and number of genes 54613. The database consists of two classes: the first one is related to children with autism (n=82) and the second to control (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.

The important challenge 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 system.

Feature selection methods

Feature selection is an important operation in processing the data stored in gene microarrays. The most relevant genes (treated as the features) increase our understanding of the mechanism of disease formation and allow to predict the potential danger of being affected by such disease. 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 different feature selection methods will be examined and integrated in the final system. Using the set of methods well increase the probability of finding the globally optimal set of genes which are the best associated with the particular disease.

The paper will study 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 multi-input linear Support Vector Machine network. These methods rely their operation principles on different foundations and thank to this allow to look on the selection problem from different points of view. The following sections present short description of each method used in investigation.

Fisher discriminant analysis

In Fisher approach the greatest wage 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 [14] :

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

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. On the other hand small value is an indication of the insignificance of the feature in the recognition of these two classes.

ReliefF algorithm

The reliefF algorithm ranks the features according to the highest correlation with the observed class. It can be implemented for incomplete and noisy data. According to the reported results [15] it represents an important approach to gene selection.

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 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 contribution for each class of misses is weighted with the prior probability of that class $P(C)$ which is estimated from the training set. The process is repeated *m* times, where *m* is a user-defined parameter. The detailed description of the procedure can be found in [15].

Two-sample t-test

The next used selection method is a two-sample *t*-test. One explicit assumption of *t*-test is that each of two compared populations should follow a normal distribution. 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 *n* and *m* represent the sample sizes of both classes [12].

Two sample *t*-test is implemented in MATLAB as *ttest2* function [12]. 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.

Figure 2 illustrates the histogram of the distribution of values of the randomly selected gene for the autism and references classes.

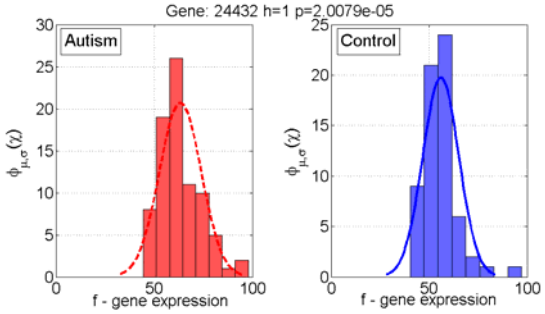

Fig. 2. The histograms and their Gaussian approximations for randomly selected gene representing two classes (autism and controls)

In the above case, the null hypothesis at the 5% significance level was rejected (*h*=1). It means that the distributions of these classes differ. This randomly selected gene can be accepted as a candidate for the relevant feature. Checking the condition of normality distribution of genes we found that in about 80% cases it was fulfilled.

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 [12].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

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

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 selection stage. Figure 3 illustrates the result of KS test for randomly selected gene.

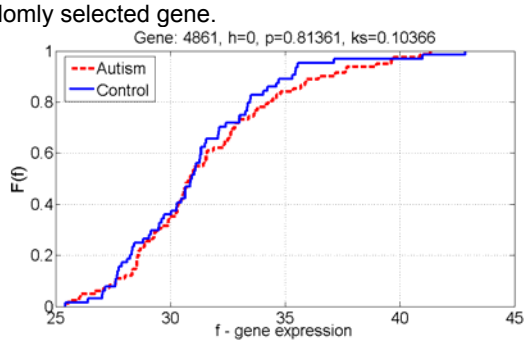

Fig. 3. The cumulative distribution function for two classes (autism and controls) for randomly selected gene in Kolmogorov-Smirnov test.

From the discriminative point of view this gene is not significant (*p*=0.8136) and should not be taken into account in further selection procedure.

Kruskal-Wallis test

In this method medians of the samples are compared, but test uses ranks of the data rather than the numeric values. 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 assumes that the observations in each group come from populations with the same shape of distribution. It returns the *p* value for the null hypothesis that all samples are drawn from the same population.

The Kruskal-Wallis test is implemented in MATLAB as *kruskalwallis* function. The returned value of *p* indicates *kruskalwallis* test rejects the null hypothesis that all data from two classes come from the same distribution at 1% significance level.

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 an *F*-statistic [14] 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. The method proceeds as follows [12]:

1. Fit the initial model.
2. If any terms not in the model have *p*-values less than an entrance tolerance add the one with the smallest *p* value and repeat this step; otherwise go to step 3.
3. If any terms in the model have *p*-values greater than an exit tolerance remove the one with the largest *p* value and go to step 2; otherwise end.

The algorithm is interrupted if none of steps leads to the increase of the model accuracy. Presented method may build different sets of features depending on 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.

Feature correlation with class

In this method, the 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 [14]:

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

where *m* is a mean value of feature for all data, m_k is a mean value of the feature for the *k*th class data, $\sigma^2(f)$ is a variance of feature, P_k is a probability of *k*th class occurrence in dataset (the uniform distribution is assumed).

In this paper number of classes is 2 (*K*=2). At the uniform distribution of both classes the above equation can be simplify to the following formula:

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

The large value of $S_{12}(f)$ indicates good discriminative ability of feature for recognition of two classes.

Multi-input linear SVM network

SVM network is mainly used in regression and classification tasks [8]. It can be also configured for solving selection problem. In this approach SVM network with linear kernel is used. Network is learned applying all available features used simultaneously as an input. The sign (*sgn*) function is added for matching the input values to the appropriate class label [14].

The output *y* at presentation of the features organized in the form of vector **f** is defined by the following equation

$$y(\mathbf{f}) = \mathrm{sgn}(u) = \mathrm{sgn}(\mathbf{w}^T\mathbf{f} + b) \tag{6}$$

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 practice the recursive feature elimination (RFE) approach is used [20]. In RFE, the features are eliminated step by step according to an assumed criterion related to their support in the discrimination of the classes. The SVM is re-trained at each step using smaller and smaller population of features. In first step the linear SVM network is learned using all features. The weights are adapted and then sorted in descending order. In the next steps the features associated with the smallest absolute weights are eliminated. The process is repeated until we obtain an appropriate number of the most important features.

Numerical results of experiments

This section describes the numerical experiments using the autism data. We will present the results concerning the selection, clusterization and PCA transformation. In the first

stage eight feature selection methods are used to discover the sequence of the most significant genes. In the next stage, we consider only 100 top selected genes from each method. These subsets are fused into one reduced common set. To find the optimal number of genes we use the notion of cluster purity. The final outcome of the clusterization is illustrated and examined by using PCA transformation.

Gene selection results

Feature selection methods described in the previous sections have been applied to get the order of genes, sorted in a decreasing fashion. As a result of these experiments eight subsets of 100 the most relevant genes are selected. The following abbreviations were applied: ***FD*** - the Fisher discriminant analysis, ***RF*** - the ReliefF algorithm, ***TT*** - the two-sample *t*-test, ***KS*** - the Kolmogorov-Smirnov test, ***KW*** - the Kruskal-Wallis test, ***SWR*** - the stepwise regression method, ***COR*** - the feature correlation with a class, ***SVM*** - the multi-input linear SVM network.

As was expected the methods have selected different sets of genes. Table 1 shows how many identical genes among the first 100 of the most important have been selected by different methods.

Table 1. The percentage of the same genes among top 100 selected by different methods.

	FD	RF	TT	KS	KW	SWR	COR	SVM
FD	100	46	90	30	66	3	90	1
RF	46	100	46	33	46	4	46	1
TT	90	46	100	30	68	3	100	0
KS	30	33	30	100	46	3	30	1
KW	66	46	68	46	100	3	68	1
SWR	3	4	3	3	3	100	3	1
COR	90	46	100	30	68	3	100	0
SVM	1	1	0	1	1	1	0	100

The contents of the selected sets differ from method to method. Analyzing them we may find that few methods identified a large number of the same genes. For example the correlation feature with a class and *t*-test produced the same sets of genes. The overlapping results between the Fisher and correlation with a class methods cover 90% of genes. On the other hand some of them have resulted in very different sets, i.e., stepwise regression and *t*-test outcomes are overlapping only in 3%.

The quality of the selection processes has been confirmed by analyzing the expression profiles for the identified genes. Fig. 4 illustrates the expression levels of all patients for the most important gene selected by the Fisher method. As we can see the mean value of the observations belonging to the autism class differs significantly from the reference class.

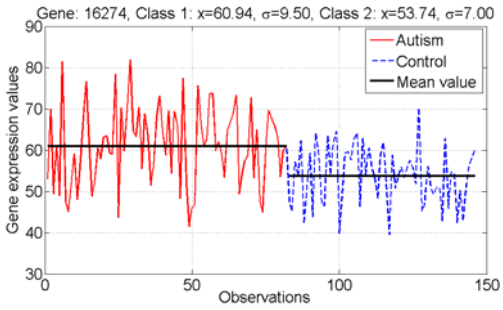

Fig. 4. Expression levels for gene belonging to the most significant group.

For comparison the appropriate expression levels representing gene selected randomly from the least significant subset are shown in Fig. 5. This time the difference between the means of both classes is unnoticeable.

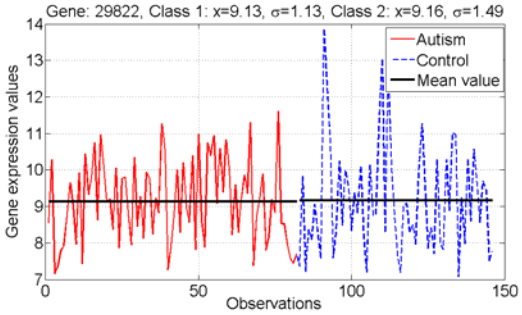

Fig. 5. Expression levels for gene representing the least significant group.

In the case of the carefully selected genes (Fig. 4) we got the mean equal 60.94±9.50 (autism) and 53.74±7.00 (control group). The difference of the mean values is 7.20 (11.81% in relative terms). For the gene representing the least significant subset (Fig. 5) we got 9.13±1.13 (autism) and 9.16±1.49 (control group). The spread of mean values in this case is equal only 0.03 (0.33%). These differences confirm the advantage of the proposed selection procedure.

The next important step is fusing the results of the individual methods into one common outcome. To find the most important genes valid for all analyzed methods we have to assign the global weight to each selected gene. The following formula has been proposed for assigning the global weight *w* of the gene *f*

$$w(f) = \sum_{i=1}^{k} w_k(f) \tag{7}$$

The index *k* is the number of applied selection methods (*k*=8 in this research), w_k is the position of genes in the appropriate method of selection. The best gene is the one of the smallest value of the global weight *w*.

Analyzing the contents of all selected sets we have found 427 different genes. They have been sorted according to their global weights in an increasing order (the best gene is the one of the smallest weight). To the best 10 genes selected by the fusion algorithm belong: 16274 (206827_s_at), 46183 (236933_at), 38133 (228878_s_at), 335 (1552729_at), 50099 (240849_at), 18235 (208819_at), 45248 (235998_at), 2923 (1556314_a_at), 26879 (217593_at), 4077 (1558154_at).

Clusterization of gene space

Good way for assessing the quality of the selected genes is to apply the clustering of data in the multidimensional space. Good set of genes should provide the clusters of high purity with respect to the class membership. Different approaches to clustering are possible: K-means, fuzzy c-means or expectation maximization algorithm [14]. In this work we apply 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) serving as a prototype of the cluster.

This aim is achieved by minimizing the squared sum distances between centroids and the vectors within each cluster [14]. 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 cluster followed by recalculation of the cluster centroids.

In our paper, we assume the number of clusters 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 providing the highest class purity of the clustered space. The procedure consists of repeating the K-means algorithm at varying number of the genes. In each step we increase this number by one. The clusters are assessed using their purity index, defined as follows

$$p_i = \max \frac{n_{ij}}{n_i} \tag{8}$$

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 ith cluster (i,j =1,..,K).

In the next step the total purity of the clustered space is calculated using the following equation:

$$p = \sum_{i=1}^{K} \frac{n_i}{n} p_i \tag{9}$$

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.

Figure 6 presents the change of the total purity index versus the number of the most relevant genes after final fusion procedure. We can observe that the best purity is obtained for the top 24 features. Its value in our experiments was equal 0.84.

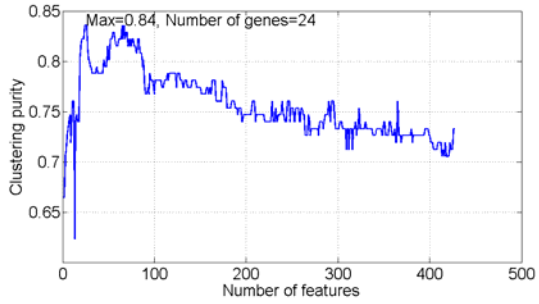

Fig. 6. Total purity index of clustered space versus number of the most significant genes.

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 number of genes at which these maxima happened.

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

	FD	RF	TT	KS	KW	SWR	COR	SVM	Fusion
purity	0.84	0.77	0.83	0.77	0.82	0.70	0.83	0.62	**0.84**
number of genes	52	75	34	53	24	17	34	9	**24**

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 stepwise regression method the best purity occurs at only nine the most significant genes, whereas in the ReliefF algorithm at seventy-five. The best clusterization of the space corresponds to the fusion approach at presence of only 24 best genes. These 24 genes have been taken into account in further experiments.

To illustrate these results in a graphical form we have presented the expression levels of the selected genes in the form of image. Figure 7 shows the image of the expression profiles for the top twenty four genes using the colormap of hot. There is a visible border between the 82 observations of the autism group and the remaining 64 representing the reference one.

For comparison the image of the expression profiles for 24 genes chosen randomly from the base is presented in Figure 8. There is a significant difference confirming good performance of the proposed selection procedure.

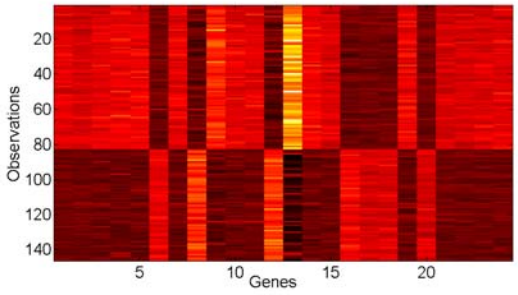

Fig. 7. The color image of the expression profiles for 24 the most significant genes selected by the fusion procedure.

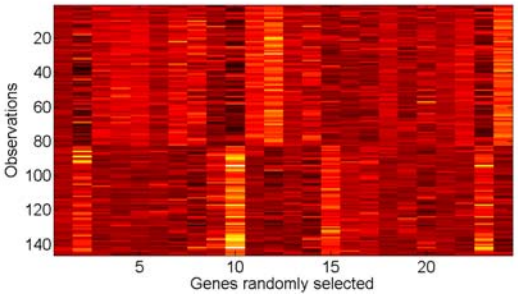

Fig. 8. The color image of the expression profiles for 24 randomly chosen genes.

Illustration of selection results using PCA

The next considerations are 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 **x** from the large space to vectors **y** in the space of the reduced dimension. The transformation is done through the linear relation

$$\mathbf{y} = \mathbf{W}\mathbf{x} \tag{10}$$

in which the transformation matrix **W** is formed from chosen number of eigenvectors corresponding to the largest eigenvalues of the correlation matrix of the original data **x**.

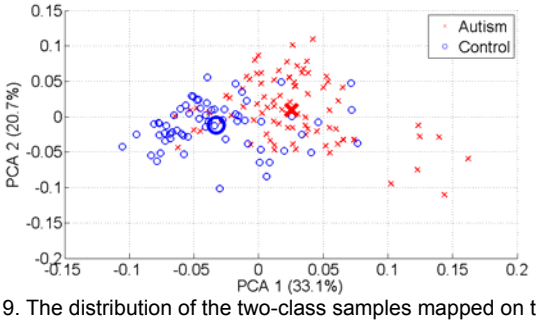

Fig. 9. The distribution of the two-class samples mapped on two the most important principal components at representation of vectors **x** by twenty four most significant genes.

We have mapped the multidimensional observations into 2-dimensional space formed by two most important principal components. Two cases have been investigated. In the first approach the original vectors contained only the selected 24 genes. Figure 9 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 vectors containing all genes. The graphical results of sample distribution are presented in Fig. 10.

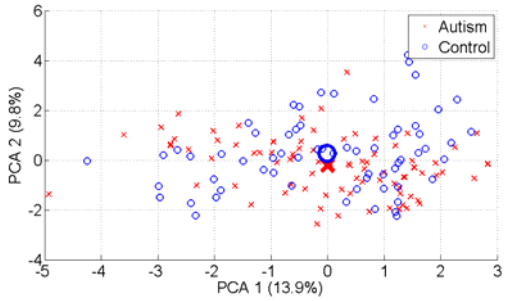

Fig. 10. The distribution of the two-class samples mapped on two most important principal components at application of all genes.

We can observe a significant difference in distribution of the samples belonging to both classes. In the first case (Fig. 9) we can see two compact regions relatively pure with respect to the classes. In the second case (Fig. 10) the situation is different. The observations belonging to different classes interlace each other. It is practically impossible to separate them into two classes.

To confirm this graphical impression the Euclidean distance between the observations and their centroids have been computed. Table 3 shows the values of their average distance to the appropriate centroids and standard deviation of the observations for both investigated cases. They correspond to the data mapped to the 2-dimensional space.

Table 3. The average distance and standard deviation of the observations from their centroids.

Number of genes	Autism	Control group
Top 24 genes	0.05±0.03	0.04±0.03
All genes	1.70±0.90	1.64±1.34

The results show that the total dispersion in the first case (24 genes forming the vectors **x**) is much smaller than in the second (all genes). At the same time we see significant difference of distances between the centroids representing both classes of data in 2-dimensional space. At representation of data by 24 genes this distance related to the maximum range of data was equal 0.226. In the second case this distance was only 0.034.

Conclusions

The paper has examined several data mining methods for selection of 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 applied eight different feature selection methods working independently. The final set has been identified by fusing all obtained subsets.

Next, the expression levels of the selected genes have been investigated. We applied different tools and methods, including the clusterization of the data and the measures of its quality, principal component analysis and statistical characterization of the clustered space. The notion of the cluster purity has allowed us to identify the optimal number of the genes.

Good quality of the presented selection approach has been confirmed by projecting the selected (limited) number of genes on the 2-dimensional space formed by two most important principal components using PCA algorithm.

The results presented in the paper form the first step of exploring the microarray data related to autism. They allow to identify the genes which are the best associated with the disease. In the next step we can use them in early recognition of this disease by applying the predictive classifier system.

REFERENCES

[1] Alter M., Kharkar R. Ramsey K., Craig D., Melmed R., Grebe T., Curtis-Bay R., Ober-reynolds S., Kirwan J., Jones J., Blake-Turner J., Hen R., Stephan D., Autism and increased patternal age related changes in global levels of gene expression regulation, *Plos One*, 2011, vol. 6, pp. 1-10.

[2] Baldi P., Long A.D., A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inference of gene changes, *Bioinformatics*, 2001, vol. 17, pp. 509-519.

[3] Fan R.E., Chen P.H., Lin C.J., Working set selection using second order information for training SVM, Journal of Machine Learning Research, ., 2005, vol. 6, pp.1889-1918.

[4] De Rinaldis E., *DNA microarrays: current applications*, Horizon Scientific Press, Norfolk, 2007.

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

[6] Eisen M., Spellman P., Brown P., Cluster analysis and display of genome wide expression patterns, *Proc. Natl. Acad. USA*, 1998, vol. 95, pp. 14863-14868.

[7] Golub T. et al., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*, 1999, vol. 286, pp. 531-537.

[8] Guyon I., Weston A.J., Barnhill S., Vapnik V., Gene selection for cancer classification using SVM, *Machine Learning*, 2002, , vol. 46, pp. 389-422.

[9] Guyon I., Elisseeff A., An introduction to variable and feature selection, *Journal of Machine Learning Research*, 2003, vol. 3, pp. 1158 – 1182.

[10] Hewett R., Kijsanayothin P., Tumor classification ranking from microarray data, *BMC Genomics*, 2008, vol. 9(2), pp. 1-11.

[11] Huang X., Pan W., Linear regression and two-class classification with gene expression data, *Bioinformatics,* 2003, vol. 19, pp. 2072-2078.

[12] Matlab user manual – *Statistics toolbox*, MathWorks, Natick, USA, 2012.

[13] Mitsubayashi H., Aso S., Nagashima T., Okada Y., Accurate and robust gene selection for disease classification using a simple statistics, *Biomedical Informatics*, 2008, v. 391, 68-71.

[14] Osowski S, *Methods and tools in data mining* (in Polish), BTC, Warsaw, 2013.

[15] Robnik-Sikonja R, Kononenko I., Theoretical and empirical analysis of ReliefF and RReliefF, *Machine Learning*, 2003, vol. 53, pp. 23-69.

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

[17] Tan P. N., Steinbach M., Kumar V., *Introduction to data mining*, Pearson Education Inc., Boston, 2006.

[18] Wang X., Gotoh O., Cancer classification using single genes, *Genom Informa*tics, 2009, vol. 23 (1): pp. 179-188..

[19] Wang X., Gotoh O., A Robust Gene Selection Method for Microarray-based Cancer Classification, *Cancer Informatics*, 2010, vol. 9, pp. 15-30.

[20] Wiliński A., Osowski S., Ensemble of data mining methods for gene ranking, *Bulletin of the Polish Academy of Sciences*, 2012. vol. 60, pp. 461-471.

[21] Woolf P. J., Wang Y., A fuzzy logic approach to analyzing gene expression data, *Physiological Genomics*, 2000, v. 3, pp. 9-15.

[22] Yang F., Robust feature selection for microarray data based on multicriterion fusion, *IEEE Trans. Computational Biology and Bioinformatics*, 2011, vol. 8(4), pp. 1080-1092.

[23] http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS4431

***Authors**:*
mgr inż. Tomasz Latkowski, Military University of Technology, Email: tlatkowski@wat.edu.pl;
prof. dr hab. inż. Stanisław Osowski, Warsaw University of Technology, Military University of Technology, Email: sto@iem.pw.edu.pl.