APPLICATION OF THE RECURSIVE FEATURE ELIMINATION AND THE RELAXED LINEAR SEPARABILITY FEATURE SELECTION ALGORITHMS TO GENE EXPRESSION DATA ANALYSIS

Joanna Gościk, Tomasz Łukaszuk

Faculty of Computer Science, Bialystok University of Technology, Białystok, Poland

Abstract: Most of the commonly known feature selection methods focus on selecting appropriate predictors for image recognition or generally on data mining issues. In this paper we present a comparison between widely used Recursive Feature Elimination (RFE) with resampling method and the Relaxed Linear Separability (RLS) approach with application to the analysis of the data sets resulting from gene expression experiments. Different types of classification algorithms such as K-Nearest Neighbours (KNN), Support Vector Machines (SVM) and Random Forests (RF) are exploited and compared in terms of classification accuracy with optimal set of genes treated as predictors selected by either the RFE or the RLS approaches. Ten-fold cross-validation was used to determine classification accuracy.

Keywords: gene expression analysis, feature selection, classification

1. Introduction

Gene expression data analysis has become a very important topic nowadays due to the need of understanding mechanisms underlying disease development, phenotypic differences between groups of subjects/patients or the influence of environmental factors on functioning of a particular organism. The existence of variety of vendors (e.g. Affymetrix, Agilent, Illumina), either providing standard or custom-made platforms, enabling measuring gene expression causes a variety of data formats produced with the use of different techniques. As a result, many tools have been created to deal with data preprocessing (e.g. normalization across arrays, background signal correction) – a fist and an inevitable step that must be performed in order to analyse gene expression data. Incorporating a specificity of an experiment's design into the analysis

Advances in Computer Science Research, vol. 10, pp. 39-52, 2013.

is also a fundamental issue when one wants to conduct the analysis. This specificity includes the occurrence of technical and biological replicates since those cannot be treated the same way. An example of a software making possible data preprocessing and discovery of genes significantly differentially expressed in groups under investigation is Limma package [1] freely available as a part of the Bioconductor project [2]. The approach proposed in this package along with many other solutions enables identifying a set of under or over-expressed genes by performing multiple testing (one test for each gene) but gives no information about the discriminative power of this set – whether the set of chosen genes can be treated as a set of features providing good classification accuracy. In this paper we present a comparison between two methods of feature selection: the Recursive Feature Elimination (RFE) with resampling [3] technique and the Relaxed Linear Separability (RLS) [4] approach. Obtained discriminative set of genes is validated using 10-fold cross validation with the use of the following classifiers: K-Nearest Neighbours (KNN), Support Vector Machines (SVM) and Random Forests (RF) and the classification accuracy is reported.

2. Data acquisition and characteristic

All samples composing an example data set used for all calculations were downloaded from Gene Expression Omnibus database [5]. It is a public functional data repository which contains detailed information about platforms used for the experiment, series of experiments conducted with the use of a specific platform and samples gathered within a concrete study (series). GEO database provides users with an advanced search engine allowing finding e.g. experiments carried out with a desired technique as gene expression profiling by array (microarray experiments), experiments carried out with a desired platform - especially useful when one wants to compare own results with another obtained with the same equipment or experiments relating to a specific organism.

Forty eight samples of series GSE27144 available online since June 15th, 2012 were exploited in our study. Technique used in this experiment was Real-Time PCR (qPCR) and samples concerned homo sapiens. Available genetic data, extended by clinical features as BMI was analysed and published [6]. Authors investigated whole saliva as a source of biomarkers to discriminate subjects who have and have not undergone severe and life-threating difficulties. Objective of the research was to evaluate an influence of severe life events and difficulties on: (1) clinical characteristics, (2) salivary analyte metrics and (3) salivary gene expression. Genetic part of the data set related to genes that have previously revealed differential expression in genome-wide analysis of adults experiencing various levels of a chronic stress and only this

part of the data is taken under consideration in our study. Data was gathered within the framework of Oregon Youth Substance Use Project (OYSUP) [7] and divided into two separate groups: low level of stress (L) and high level of stress (H), each consisting of 24 subjects. Every sample contained expression levels for 38 genes represented by normalized C_T value using $\Delta\Delta C_T$ method [8].

2.1 Brief introduction to Real-Time PCR technology

Polymerase chain reaction (PCR) is a method that allows exponential amplification of short DNA sequences within a longer double stranded DNA molecule which consist of subsequent cycles theoretically resulting in doubling the DNA amount when compared to the previous cycle - an exponential reaction. The reaction finally tails off and reaches a plateau. For the most part, this method was developed to enable measuring differences in gene expression. To introduce basic terms related to RT-PCR let us assume that we have two kinds of cells: drug treated (experimental lane) and non-treated (control lane) and two signal intensities for a specific gene - one for each sample. If the amount of signal in the experimental sample was 100 times greater when compared to the control sample we could say that expression of the gene has increased 100-fold in the experimental cells but it could also mean that we have 100 times more RNA in the experimental lane - in other words we have a loading artefact. To prevent this kind of artefacts housekeeping genes were introduced. Housekeeping gene is a kind of gene, that its expression does not change depending on the conditions (e.g. it remains the same in different kinds of tissues). We could call these genes loading control. Let us say that the experiment showed that for our housekeeping gene there is twice as much DNA in the experimental lane comparing to control sample. This means that the real change in gene under investigation expression is equal to 100/2=50 fold. The actual fold change is given by the Equation 1.

$$\text{Fold change} = \frac{\text{Fold change of a target gene}}{\text{Fold change of a reference gene}} \tag{1}$$

The goal of an RT-PCR experiment is to measure level of expression of a certain gene from a specific sample. This measurement is expressed in Cycled to Threshold (C_T) - a relative value representing the cycle number at which the amount of amplified DNA reaches the threshold level - exceeds background intensity. An important property of C_T is that it is reversely proportional to the DNA quantity present in a sample expressed on a logarithmic scale (the amount of DNA doubles with every cycle). In order to make different C_T values comparable across multiple samples normalization is applied. There are many methods of normalization available, but in this article we

will focus on $\Delta\Delta C_T$ method [8]. A detailed description of the process can be found in the cited article so we will just mention the main formula which is given below.

$$\Delta\Delta C_T = \Delta C_T\,(\text{experimental}) - \Delta C_T\,(\text{reference}) \tag{2}$$

Where $\Delta C_T\,(\text{experimental})$ refers to the C_T value for the gene in an experimental lane normalized against housekeeping gene and $\Delta C_T\,(\text{reference})$ refers to the C_T value for that gene in an control lane again normalized against housekeeping gene. The amount of target normalized against housekeeping gene and relative to the control is given by the Equation 3.

$$\text{Amount of target} = 2^{-\Delta\Delta C_T} \tag{3}$$

Quantities obtained with the use of Equation 3 can then be compared across samples.

2.2 Missing values imputation

Source data set, after combining all samples, contained missing values. Missing values characteristics for each class are given in Table 1. where columns represent genes and rows represent appropriate metrics.

Table 1. Missing values occurence in classes with low (L) and high (H) levels of stress.

		ADAR	**ADORA2A**	**BTN3A3**	**C7orf68**	**CD9**	**CSF1R**	**CX3CR1**	**CYLD**	**DPYD**	**FGL2**	**FOLR3**
Class L	N observed	24	24	22	23	24	23	23	23	23	24	22
	Missing	0%	0%	8%	4%	0%	4%	4%	4%	4%	0%	8%
Class H	N observed	23	24	16	24	24	17	18	21	22	20	15
	Missing	4%	0%	33%	0%	0%	29%	25%	12%	8%	17%	37%
		FOSB	**GADD45B**	**GALC**	**GBP1**	**GNA15**	**GORASP2**	**HLA-DQB1**	**HSPA1B**	**IL8**	**MAFF**	**NAIP**
Class L	N observed	24	24	23	23	24	23	12	24	24	24	22
	Missing	0%	0%	4%	4%	0%	4%	50%	0%	0%	0%	8%
Class H	N observed	21	24	16	17	24	22	10	24	24	24	17
	Missing	12%	0%	33%	29%	0%	8%	58%	0%	0%	0%	29%
		NDUFB7	**NGLY1**	**NRG1**	**NSF**	**PUM2**	**RAB27A**	**RPA1**	**SEC24A**	**SERPINB2**	**SLC35A1**	**SLC7A5**
Class L	N observed	24	23	23	22	23	24	23	24	24	22	24
	Missing	0%	4%	4%	8%	4%	0%	4%	0%	0%	8%	0%
Class H	N observed	23	21	16	18	22	19	19	23	19	16	23
	Missing	4%	12%	33%	25%	8%	21%	21%	4%	21%	33%	4%
		STAT1	**STX7**	**THBS1**	**VEGFA**	**WDR7**						
Class L	N observed	24	24	23	24	23						
	Missing	0%	0%	4%	0%	4%						
Class H	N observed	23	23	24	23	18						
	Missing	4%	4%	0%	4%	25%						

Package impute [9] from the Bioconductor project [2] was used to complete the data set. Method implemented in this package uses K-Nearest Neighbours approach to impute missing expression data provided that the data set exploited in the computation is in matrix form with rows corresponding to genes and columns corresponding to samples. For each gene with missing expression levels K nearest neighbours are

found using the Euclidean metric with restriction to samples for which expression level for that gene is defined. Descriptive statistics including mean and its standard error before and after missing values imputation are presented in Table 2 and Table 3 respectively.

Table 2. Descriptive statistics for both Low (L) and High (H) levels of stress for data set before imputing missing values.

		ADAR	**ADORA2A**	**BTN3A3**	**C7orf68**	**CD9**	**CSF1R**	**CX3CR1**	**CYLD**	**DPYD**	**FGL2**	**FOLR3**
Class L	Mean	1.03	1.13	1.33	1.21	1.04	1.49	1.27	1.12	1.22	1.05	1.31
	Std. Err.	0.01	0.01	0.01	0.02	0.01	0.01	0.02	0.01	0.01	0.01	0.02
Class H	Mean	1.02	1.12	1.34	1.18	1.08	1.50	1.27	1.15	1.18	1.07	1.33
	Std. Err.	0.01	0.02	0.01	0.02	0.02	0.02	0.01	0.01	0.01	0.02	0.01
		FOSB	**GADD45B**	**GALC**	**GBP1**	**GNA15**	**GORASP2**	**HLA-DQB1**	**HSPA1B**	**IL8**	**MAFF**	**NAIP**
Class L	Mean	1.17	0.87	1.25	0.98	1.12	1.26	1.29	0.80	0.70	0.98	1.05
	Std. Err.	0.01	0.01	0.01	0.02	0.01	0.01	0.11	0.01	0.02	0.01	0.01
Class H	Mean	1.19	0.86	1.33	1.05	1.09	1.22	1.12	0.79	0.84	0.97	1.10
	Std. Err.	0.01	0.01	0.02	0.03	0.02	0.02	0.07	0.01	0.04	0.01	0.02
		NDUFB7	**NGLY1**	**NRG1**	**NSF**	**PUM2**	**RAB27A**	**RPA1**	**SEC24A**	**SERPINB2**	**SLC35A1**	**SLC7A5**
Class L	Mean	1.18	1.23	1.52	1.21	1.26	1.13	1.26	1.23	1.26	1.27	1.10
	Std. Err.	0.01	0.01	0.02	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.01
Class H	Mean	1.18	1.22	1.52	1.25	1.26	1.17	1.25	1.24	1.30	1.31	1.13
	Std. Err.	0.01	0.01	0.03	0.01	0.01	0.01	0.02	0.01	0.02	0.02	0.02
		STAT1	**STX7**	**THBS1**	**VEGFA**	**WDR7**						
Class L	Mean	1.00	1.08	1.09	1.04	1.31						
	Std. Err.	0.01	0.01	0.02	0.01	0.01						
Class H	Mean	1.00	1.11	1.05	1.03	1.32						
	Std. Err.	0.01	0.01	0.02	0.01	0.02						

Table 3. Descriptive statistics for both Low (L) and High (H) levels of stress for data set after imputing missing values.

		ADAR	**ADORA2A**	**BTN3A3**	**C7orf68**	**CD9**	**CSF1R**	**CX3CR1**	**CYLD**	**DPYD**	**FGL2**	**FOLR3**
Class L	Mean	1.03	1.13	1.33	1.21	1.04	1.49	1.27	1.12	1.21	1.05	1.31
	Std. Err.	0.01	0.01	0.01	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.02
Class H	Mean	1.01	1.11	1.33	1.18	1.08	1.49	1.26	1.15	1.18	1.07	1.33
	Std. Err.	0.01	0.02	0.01	0.02	0.02	0.01	0.01	0.01	0.01	0.02	0.01
		FOSB	**GADD45B**	**GALC**	**GBP1**	**GNA15**	**GORASP2**	**HLA-DQB1**	**HSPA1B**	**IL8**	**MAFF**	**NAIP**
Class L	Mean	1.17	0.87	1.25	0.97	1.12	1.26	1.34	0.80	0.70	0.98	1.05
	Std. Err.	0.01	0.01	0.01	0.02	0.01	0.01	0.07	0.01	0.02	0.01	0.01
Class H	Mean	1.20	0.86	1.34	1.05	1.09	1.22	1.17	0.79	0.84	0.97	1.11
	Std. Err.	0.02	0.01	0.02	0.02	0.02	0.02	0.04	0.01	0.04	0.01	0.01
		NDUFB7	**NGLY1**	**NRG1**	**NSF**	**PUM2**	**RAB27A**	**RPA1**	**SEC24A**	**SERPINB2**	**SLC35A1**	**SLC7A5**
Class L	Mean	1.18	1.23	1.52	1.21	1.26	1.13	1.26	1.23	1.26	1.27	1.10
	Std. Err.	0.01	0.01	0.02	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.01
Class H	Mean	1.18	1.22	1.50	1.25	1.26	1.18	1.24	1.24	1.30	1.30	1.13
	Std. Err.	0.01	0.01	0.02	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.01
		STAT1	**STX7**	**THBS1**	**VEGFA**	**WDR7**						
Class L	Mean	1.00	1.08	1.08	1.04	1.31						
	Std. Err.	0.01	0.01	0.02	0.01	0.01						
Class H	Mean	1.00	1.11	1.05	1.03	1.31						
	Std. Err.	0.01	0.01	0.02	0.01	0.01						

3. Feature selection for classification

Feature selection process can be perceived as a technique of choosing a subset of variables from available feature set with the evaluation of accuracy for each choice. Nowadays, when the amount of genomic data is growing rapidly, feature selection has become very important issue because of the specificity of the data. Data sets containing genetic information are usually in a form of a matrix where genes are considered as features which expression levels are measured for each sample (rows of a matrix). One distinctive feature characterizing these kind of data sets is the disequilibrium between the number of samples and the number of features - the number of features is generally 10^3 greater (or more) than the number of samples as the modern genomic technology (e.g. microarrays, Next Generation Sequencing - NGS) enables quantifying expression of enormous number of different kinds of genetic attributes (e.g. genes, microRNA). This disequilibrium is associated with the experiment's design: it is focused on determining as many genetic attributes as possible (e.g. whole genome sequencing where one can discover individual differences comparing to the reference genome of as small variations as Single Nucleotide Polymorphism - SNP). Moreover, performing these kinds of experiments is quite expensive and time-consuming, so usually very few replications, either technical or biological, are produced. Due to just outlined matters feature selection appears to be very important task in genomic data analysis.

As it was previously mentioned our data set contained 38 features (genes) and 48 samples divided into two separate groups determined by the level of stress. In this work we present a comparison of two feature selection methods and their influence on the classification accuracy. First method we adapted for the purpose of feature subset selection is the Recursive Feature Elimination (RFE) incorporating resampling implemented in the caret package [10]. The second method is based on the Relaxed Linear Separability (RLS) approach [4]. Subsets of features providing the best classification accuracy generated either with the use of the RFE method or the RLS approach were generated for three classifiers: K-Nearest Neighbours (KNN), Support Vector Machines (SVM) and Random Forests (RF). Ten-fold cross validation was used to determine the classification accuracy.

3.1 Recursive Feature Elimination approach

Basic Backward Selection (a.k.a. Recursive Feature Elimination) algorithm firstly fits the model to all of the features. Each feature is then ranked according to its importance to the model. Let S be a sequence of ordered numbers representing the number

of features to be kept ($S_1 > S_2 > S_3$...). At each iteration of feature selection algorithm, the S_i top raked features are kept, the model is refit and the accuracy is assessed. The value of S_i with the best accuracy is assessed and the top S_i features are used to fit the final model. Algorithm 1 gives a description of subsequent steps of this procedure.

Algorithm 1 Basic Recursive Feature Elimination

1: Train the model using all features
2: Determine model's accuracy
3: Determine feature's importance to the model for each feature
4: **for** *Each subset size* S_i, $i = 1 \ldots N$ **do**
5: Keep the S_i most important features
6: Train the model using S_i features
7: Determine model's accuracy
8: **end for**
9: Calculate the accuracy profile over the S_i
10: Determine the appropriate number of features
11: Use the model corresponding to the optimal S_i

Model building process is composed of few successive steps and feature selection is one of these. Due to that, resampling methods (e.g. cross-validation) should contribute to this process when calculating model's accuracy. It has been showed that improper use of resampling when measuring accuracy can result in model's poor performance on new samples [11] [12]. A modification of Algorithm 1 including resampling was suggested and is outlined in Algorithm 2.

3.2 Relaxed Linear Separability approach

Relaxed Linear Separability (RLS) approach to feature selection problem refers to the concept of linear separability of the learning sets [14]. The term "relaxation" means here deterioration of the linear separability due to the gradual neglect of selected features. The considered approach to feature selection is based on repetitive minimization of the convex and piecewise-linear (CPL) criterion functions. These CPL criterion functions, which have origins in the theory of neural networks, include the cost of various features [13]. Increasing the cost of individual features makes these features falling out of the feature subspace. Quality the reduced feature subspaces is assessed by the accuracy of the CPL optimal classifiers built in this subspace.

RLS feature selection method consists of three stages. The first stage is to determine an optimal hyperplane $H(\mathbf{w}, \theta)$ separating objects $\mathbf{x}_j$ ($\mathbf{x}_j = [x_{j1}, ..., x_{jn}]$) from

Algorithm 2 Recursive Feature Elimination with resampling

```
for Each Resampling Iteration do
    Partition data into training and testing data sets via resampling
    Train the model on the training set using all features
    Predict the held-back samples
    Determine feature's importance to the model for each feature
    for Each subset size S_i, i = 1...N do
        Keep the S_i most important features
        Train the model using S_i features
        Predict the held-back samples
    end for
end for
Calculate the accuracy profile over the S_i using held-back samples
Determine the appropriate number of features
Estimate the final list of features to keep in the final model
Fit the final model based on the optimal S_i using the original training set
```

two learning sets $G^+ = \{\mathbf{x}_j; j \in J^+\}$ and $G^- = \{\mathbf{x}_j; j \in J^-\}$ (J^+ and J^- are disjoined sets of indices j ($J^+ \cap J^- = \emptyset$)).

$$H(\mathbf{w},\theta) = \{\mathbf{x} : \mathbf{w}^T\mathbf{x} = \theta\} \tag{4}$$

The hyperplane $H(\mathbf{w},\theta)$ is calculated by the minimization of the criterion function $\Psi_\lambda(\mathbf{w},\theta)$.

$$\Phi_\lambda(\mathbf{w},\theta) = \sum_{\mathbf{x}_j \in G^+} \varphi_j^+(\mathbf{w},\theta) + \sum_{\mathbf{x}_j \in G^-} \varphi_j^-(\mathbf{w},\theta) + \lambda \sum_{i \in \{1,\ldots,n\}} \phi_i(\mathbf{w}) \tag{5}$$

where $\lambda \geq 0$.

The function $\Phi_\lambda(\mathbf{w},\theta)$ is the sum of the penalty functions $\varphi_j^+(\mathbf{w},\theta)$ or $\varphi_j^-(\mathbf{w},\theta)$ and $\phi_i(\mathbf{w},\theta)$. The functions $\varphi_j^+(\mathbf{w},\theta)$ are defined on the feature vectors $\mathbf{x}_j$ from the set G^+. Similarly $\varphi_j^-(\mathbf{w},\theta)$ are based on the elements $\mathbf{x}_j$ of the set G^-.

$$(\forall \mathbf{x}_j \in G^+) \quad \varphi_j^+(\mathbf{w},\theta) = \begin{cases} 1 + \theta - \mathbf{w}^T\mathbf{x}_j & if\ \mathbf{w}^T\mathbf{x}_j < 1 + \theta \\ 0 & if\ \ \mathbf{w}^T\mathbf{x}_j \geq 1 + \theta \end{cases} \tag{6}$$

and

$$(\forall \mathbf{x}_j \in G^-) \quad \varphi_j^-(\mathbf{w},\theta) = \begin{cases} 1 + \theta + \mathbf{w}^T\mathbf{x}_j & if\ \mathbf{w}^T\mathbf{x}_j > -1 + \theta \\ 0 & if\ \ \mathbf{w}^T\mathbf{x}_j \leq -1 + \theta \end{cases} \tag{7}$$

The penalty functions $\phi_i(\mathbf{w},\theta)$ are related to particular features x_i.

$$\phi_i(\mathbf{w}) = |w_i| \tag{8}$$

The criterion function $\Phi_\lambda(\mathbf{w},\theta)$ (5) is the convex and piecewise linear (*CPL*) function as the sum of the *CPL* penalty functions $\varphi_j^+(\mathbf{w},\theta)$ (6), $\varphi_j^-(\mathbf{w},\theta)$ (7) and $\phi_i(\mathbf{w})$ (8). The basis exchange algorithm allows to find the minimum efficiently, even in the case of large multidimensional data sets G^+ and G^- [15].

The vector of parameters $\mathbf{w} = [w_1,...,w_n]^T$ is used in the feature reduction rule [13]:

$$(w_i = 0) \Rightarrow (the\ feature\ i\ is\ omitted) \tag{9}$$

In the result of the first stage of RLS method we obtain optimal hyperplane $H(\mathbf{w},\theta)$ and initial feature set F_k composed of k features not subject to the reduction rule (9).

In the second stage the value of the cost level λ in the criterion function $\Psi_\lambda(\mathbf{w},\theta)$ is successive increased. It causes reduction of some features x_i. It is possible to determine the value of Δ_k ($\Delta_k > 0$) by which to enlarge λ in order to reduce only one feature from feature set F_k. In the result of the second stage we obtain the descended sequence of feature subsets F_k with decreased dimensionality ($F_k \supset F_{k-1}$):

$$F_k \to F_{k-1} \to ... \to F_1 \tag{10}$$

The last step is the calculation of classifier accuracy in each reduced data set corresponding to the subsets of features in sequence (10). The accuracy is usually determined by the use of the CPL [13] classifier and in cross-validation procedure. In this work, different than usual, accuracy is determined by the use of KNN, SVM or RF classifier. As the result of the third stage and the whole RLS method we obtain feature set F^*. It is the feature set characterized by the highest value of classifier accuracy.

4. Results

The classification accuracy determined for the set of genes treated as predictors and obtained with the use of either the Recursive Feature Elimination with resampling (RFE) or the Relaxed Linear Separability (RLS) approach feature selection method is presented for each of the three tested classificators (KNN, SVM, RF). Figures 1 and 2 show the dependence between the number of features selected and the classification accuracy for the KNN classification obtained with the set of features chosen by the two methods of feature selection: RFE and RLS respectively. Figures 3 and 4, 5 and 6 show the same relationship for the remaining two classifiers: SVM and RF also created for the two methods of feature selection. The best classification accuracy and related number of predictors are marked. The correlation coefficient for the

linear dependency between the number of features and classification accuracy is also presented on each figure.

As it can be noticed there are two main facts differentiating the two feature selection methods under investigation and their influence on the classification accuracy. First aspect is the existence of the statistically significant linear dependency between number of features and classification accuracy in case of the RFE method – accuracy rises in conjunction with an increasing number of features and reaches its best value with almost all features included in the predictors set. When considering the RLS method this linear relationship is not observed for the KNN and the SVM classification methods and only in the case of the RF classification statistically significant linear relationship was found, however this relation is somehow disputable – it is probably caused by relatively big decrease in the classification accuracy when considering number of features varying from three to eight. It is also worth to mention that the characteristic of the linear relationship is different for those two methods of feature selection: in the event of RFE we have found directly proportional relationship between the number of features selected and the classification accuracy as opposed to the RLS method where this relationship was rather inversely proportional. Second aspect distinguishing the RFE and the RLS methods is the number of features chosen to acquire the best classification accuracy – approximately a magnitude higher in case of the RFE method.

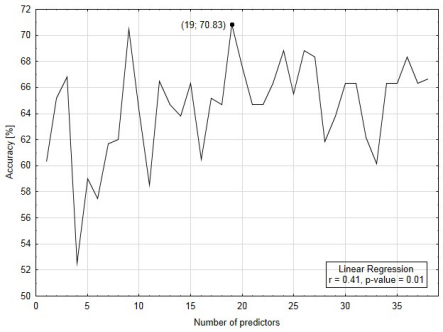

Fig. 1. KNN classification accuracy obtained with an RFE approach.

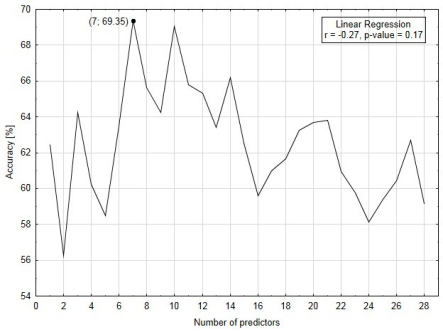

Fig. 2. KNN classification accuracy obtained with an RLS approach.

Table 4 gives a desctiption of genes selected by the RLS method. Given superscripts along with the gene name have the following meaning: 1 - gene was selected for the KNN classifier, 2 - gene was selected for the SVM classifier, 3 - gene was selected

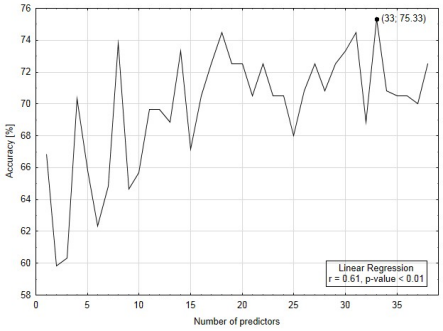

Fig. 3. SVM classification accuracy obtained with an RFE approach.

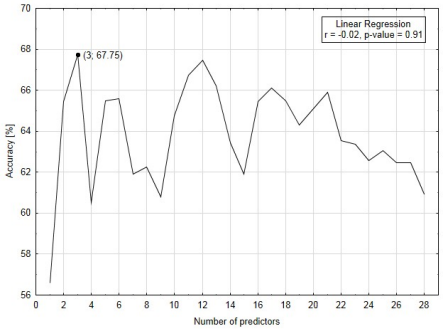

Fig. 4. SVM classification accuracy obtained with an RLS approach.

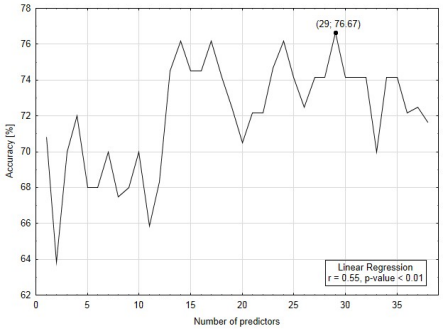

Fig. 5. RF classification accuracy obtained with an RFE approach.

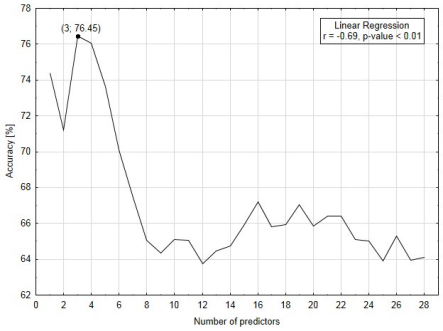

Fig. 6. RF classification accuracy obtained with an RLS approach.

for the RF classifier. As it can be seen: the same set of genes was determined for the SVM and the RF classification methods. The selected set of genes can be characterized by two features: (1) they play an important role in the immune system, (2) they are connected with the nervous system. Those two attributes might be considered as related to the response to stress exposure which is the case in our data set. Lists of genes selected by the RFE method for each classifier are not presented here because they contains almost all of the available features.

5. Conclusions

Classification accuracy obtained with the use of features/genes selected by the RFE method was slightly higher than the one calculated for the set of features/genes se-

Table 4. Genes selected by the RLS method and their description.

Gene name	Description
HLA-DQB1[1, 2, 3]	The HLA-DQB1 gene provides instructions for making a protein that plays a critical role in the immune system. The HLA-DQB1 gene is part of a family of genes called the human leukocyte antigen (HLA) complex. The HLA complex helps the immune system distinguish the body's own proteins from proteins made by foreign invaders such as viruses and bacteria.
IL8[1, 2, 3]	Is one of the major mediators of the inflammatory response. It is released from several cell types in response to an inflammatory stimulus.
NAIP[1, 2, 3]	Acts as a mediator of neuronal survival in pathological conditions. Prevents motor-neuron apoptosis (a type of cell death) induced by a variety of signals.
VEGFA[1]	Is implicated in every type of angiogenic disorder, including those associated with cancer, ischemia, and inflammation. It is also involved in neurodegeneration.
FGL2[1]	May play a role in physiologic lymphocyte functions at mucosal sites.
SLC7A5[1]	Plays a role in neuronal cell proliferation (neurogenesis) in brain.
CSF1R[1]	Plays an important role in innate immunity and in inflammatory processes. Plays an important role in the regulation of osteoclast proliferation and differentiation, the regulation of bone resorption, and is required for normal bone and tooth development.

lected by the RLS method for all three classifiers: KNN, SVM and RF although this measure did not differ substantially. One reason that might have caused this is the specificity of the data set: the RLS method was designed to perform well on sets characterized by considerably great number of features in comparison to the number of samples (few orders of magnitude greater) as presented in [4] and in the data set used in our work the number of samples was almost the same as the number of features. Another issue that is worth to mention is the lack of linear dependency between the number of selected features and the classification accuracy in case of the RLS method – it suggests that the method tends to find the optimal set of predictors providing the best classification accuracy independently from the number of available features which is especially important when one wants to analyse data sets with great amount of variables such as genomic data sets. One disadvantage of the RFE feature selection method that must be intimated is its time-consumingness, which is a common characteristic of all methods exploiting recurrence. In the event of the data set used in our experiment this issue was not so troublesome, although noticeable, but in cases when the number of features is in order of thousands or millions (what is common when analysing genomic data) it might turn out that the trade-off between the classification accuracy and the time spent on computation is not cost-effective.

References

[1] G.K. Smyth, Limma: linear models for microarray data., Bioinformatics and Computational Biology Solutions using R and Bioconductor., R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, W. Huber (eds), Springer, New York, pp. 397-420.

[2] The Bioconductor project. [http://www.bioconductor.org]

[3] F. De Martino, G. Valente, N. Staeren, J. Ashburner, R. Goebel, E. Formisano, Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns., NeuroImage, 43, pp. 44-48, (2008).

[4] L. Bobrowski, T. Łukaszuk, Feature selection based on relaxed linear separability., In: Biocybernetical and Biomedical Engineering, vol.29, nr 2, pp. 43-58, (2009).

[5] Gene Expression Omnibus. [http://www.ncbi.nlm.nih.gov/geo]

[6] A.W. Bergen, A. Mallick, D. Nishita, X. Wei et al., Chronic psychosocial stressors and salivary biomarkers in emerging adults., Psychoneuroendocrinology 2012 Aug; 37(8):1158-70.

[7] J. Andrews, Oregon Youth Substance Use Project (OYSUP), 1998-2010. ICPSR34263-v1., Ann Arbor, MI: Inter-university Consortium for Political and Social Research, 2013-03-29. doi:10.3886/ICPSR34263.v1

[8] K.J. Livak, T.D. Schmittgen, Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the $2^{-\Delta\Delta C_T}$ Method., Methods 25, 402-408 (2001).

[9] T. Hastie, R. Tibshirani, B. Narasimhan, G. Chu, impute: Imputation for microarray data. R package version 1.32.0.

[10] M. Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt and Tony Cooper, caret: Classification and Regression Training (2013). R package version 7.17-7.

[11] C. Ambroise, G.J. McLachlan, Selection bias in gene extraction on the basis of microarray gene-expression data., PNAS vol. 90 (10), pp. 6562-6566, 2002.

[12] V. Svetnik, A. Liaw, C. Tong, T. Wang, Application of Breiman's Random Forest to Modeling Structure-Activity Relationships of Pharmaceutical Molecules., Lecture Notes in Computer Science vol. 3077, pp. 334-343, 2004.

[13] L. Bobrowski, Eksploracja danych oparta na wypukłych i odcinkowo-liniowych funkcjach kryterialnych., Wyd. Politechniki Białostockiej, Białystok, (2005).

[14] L. Bobrowski Feature subsets selection based on linear separbilty, In: Lecture Notes of the VII-th ICB Seminar: Statistics and Clinical Practice, ed. by H. Bacelar-Nicolau, L. Bobrowski, J. Doroszewski, C. Kulikowski, N. Victor, June 2008, Warsaw, 2008.

[15] L. Bobrowski, Design of Piecewise Linear Classifiers from Formal Neurons by Some Basis Exchange Technique, pp. 863–870 in: Pattern Recognition, 24(9), 1991.

REKURENCYJNA ELIMINACJA CECH Z WALIDACJĄ ORAZ RELAKSACJA LINIOWEJ SEPAROWALNOŚCI JAKO METODY SELEKCJI CECH DO ANALIZY ZBIORÓW DANYCH ZAWIERAJĄCYCH WARTOŚCI EKSPRESJI GENÓW

Streszczenie Zdecydowana większość znanych metod selekcji cech skupia się na wyborze odpowiednich predyktorów dla takich zagadnień jak rozpoznawanie obrazów czy też ogólnie eksploracji danych. W publikacji prezentujemy porównanie pomiędzy powszechnie stosowaną metodą Rekurencyjnej Eliminacji Cech z walidacją (ang. Recursive Feature Elimination - RFE) a metodą stosującą podejście Relaksacji Liniowej Separowalności (ang. Relaxed Linear Separability - RLS) z zastosowaniem do analizy zbiorów danych zawierających wartości ekspresji genów. W artykule wykorzystano różne algorytmy klasyfikacji, takie jak K-Najbliższych Sąsiadów (ang. K-Nearest Neighbours - KNN), Maszynę Wektorów Wspierających (ang. Support Vector Machines - SVM) oraz Lasy Losowe (ang. Random Forests - RF). Porównana została jakość klasyfikacji uzyskana przy pomocy tych algorytmów z optymalnym zestawem cech wygenerowanym z wykorzystaniem metody selekcji cech RFE bądź RLS. W celu wyznaczenia jakości klasyfikacji wykorzystano 10-krotną walidację krzyżową.

Słowa kluczowe: analiza ekspresji genów, selekcja cech, klasyfikacja

Artykuł zrealizowano w ramach projektu badawczego N N519 657940 finansowanego przez Ministerstwo Nauki i Szkolnictwa Wyższego oraz pracy badawczej S/WI/2/2013 Wydziału Informatyki PB.