BULLETIN OF THE POLISH ACADEMY OF SCIENCES
TECHNICAL SCIENCES, Vol. 69(3), 2021, Article number: e136748
DOI: 10.24425/bpasts.2021.136748

Fusion of feature selection methods in gene recognition

Fabian GIL[1] and Stanislaw OSOWSKI[1,2]*

[1] Warsaw University of Technology, Pl. Politechniki 1, 00-661 Warsaw, Poland
[2] Military University of Technology, ul. gen. Sylwestra Kaliskiego 2, 00-908 Warsaw, Poland

Abstract. The paper presents the fusion approach of different feature selection methods in pattern recognition problems. The following methods are examined: nearest component analysis, Fisher discriminant criterion, refiefF method, stepwise fit, Kolmogorov-Smirnov criteria, T2-test, Kruskall-Wallis test, feature correlation with class, and SVM recursive feature elimination. The sensitivity to the noisy data as well as the repeatability of the most important features are studied. Based on this study, the best selection methods are chosen and applied in the process of selection of the most important genes and gene sequences in a dataset of gene expression microarray in prostate and ovarian cancers. The results of their fusion are presented and discussed. The small selected set of such genes can be treated as biomarkers of cancer.

Key words: diagnostic features; selection methods; genes; recognition; biomarkers.

1. Introduction

Feature selection is the process of reducing the number of input variables and selecting those that contribute most to the prediction of the output variables. The output variable may represent either the class to which the actual input attributes belong (classification problem) or the real predicted value (regression problem).

The presence of irrelevant features in the data usually leads to a decrease in the accuracy of the machine model and reduces its generalization ability. Therefore, their elimination is the primary task in any machine learning problem, since the reduction of input attributes improves not only the performance of the model but also reduces the computational cost of modeling, allows to avoid the curse of dimensionality, and make them easier to interpret by users.

There are many reported feature selection methods [1–3]. They rely on different principles of operation. Therefore, the selected set may differ a lot concerning its contents and also the order of significance of particular features.

The main task of this study is to identify the features which are most responsible for the decision on class membership in a badly conditioned problem represented by a gene expression array registered for two types of cancers: ovarian and prostate tumor.

The common problem in this task is a very high number of variables (genes) counted in tens of thousands and a very small set of observations (in the range of hundreds). The individual feature selection methods are relied on specific algorithms and usually generate different order of feature importance. To find the most important set of features in an objective sense, we have to apply an additional fusion phase of the individual results of members of the ensemble. We propose weighted voting, in which the importance of a feature is weighted according to its position in the set. The features in the first positions according to the applied methods are regarded as the most significant. They may be treated as biomarkers for the tumors, i.e. substance that is indicative of the presence of cancer in the body.

Many different approaches to the gene selection problem have been presented in the past. They include such as various clustering methods [4], classification approaches using neural networks and support vector machines [5, 6], application of various statistical measures [7], rough set theory [8], or application of deep learning [9], etc. Some papers have proposed also the integration of many selection methods in one system [6, 10–12]. Although the progress in this field is fast, there is still a need for a better understanding and improvement of the research.

This paper proposes the approach, in which many feature selection methods are simultaneously applied in gene selection. Their results are integrated into one definite verdict. The final ranking of genes is based on the sum of positions, taken by each gene in the selection procedures performed. The best features are those with the smallest value of the total sum.

The paper is organized as follows. The next chapter introduces the basic information of the investigated selection methods. In the following chapter, we present the results of feature selection by using the presented methods. In the first step, we test the ability of these methods to identify six predefined known features that characterize the synthetic classification problem. The important point is to find out how resistant they are to the noise that contaminates the input data. Based on these results, the best selection methods were chosen and then applied to the main problem of identifying the genes of the gene expression microarray that are most important for the detection of cancer cases from the reference cases. Two types of tumors will be investigated: ovarian and prostate. Both represent a badly

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

Manuscript submitted 2020-09-28, revised 2020-09-28, initially accepted for publication 2020-11-29, published in June 2021

conditioned problem of a very high number of variables and a small number of observations. The obtained results, presented in a graphical and numerical form, prove that fusion of many selection methods into one common ensemble system leads to very good results of class discrimination. The developed system allows separating the cancer records from reference ones in an efficient way. All numerical experiments have been performed using the Matlab platform [13].

2. Selection Methods

The methods of feature selection are directed at evaluating the relationship between the input attributes and the target variable by using a different form of statistics [14, 15]. As a result, the attributes that have the strongest relationship with the target, and the weakest connection between themselves should be selected. Feature selection is often related to dimensionality reduction. However, feature selection aims at including or excluding particular attributes without changing them, while in dimensionality reduction we usually create the new combinations of attributes, limiting their population.

There are three general classes of feature selection: filter, wrapper, and embedded methods [1, 15]. Filter methods are based on different statistical measures for the relevance of characteristics according to their correlation target. They are fast but lack robustness against interactions among features and feature redundancy. Typical representatives of this method include information gain, correlation coefficient, Fisher score, and statistical tests.

Wrapper methods (feature selections "wrapped" in a learning algorithm) make the selection by measuring the usefulness of a chosen subset of features by training a particular model on it and checking its accuracy. To the typical representatives of these methods belong recursive feature elimination (the most often combined with support vector machine), sequential feature selection algorithms, and genetic algorithm. Generally, wrapper methods are more effective than filter methods, however, their drawback is high computational cost.

Embedded methods check the contribution of particular attributes to the accuracy of the model while the model is being created. Usually, they apply different regularization methods. To this type belong for example LASSO regularization or decision tree.

Recently, hybrid approaches, taking advantage of all these methods have been proposed. Examples of hybrid algorithms include statistics combined with genetic algorithms or correlation method cooperating with recursive feature elimination. The general idea is that the filter method is first applied to reduce the size of the feature set and then the wrapper method is used to find the optimal subset of features from the selected feature pool. This makes the feature selection process faster since the filter method rapidly reduces the effective number of features in the wrapper application.

There are specialized packages devoted to feature selection methods, for example, Weka, Scikit-Learn, or Caret R package [16, 17]. In this paper, we will compare some chosen representatives of these three groups for the importance of genes in the expression microarray of prostate and ovarian cancers.

In the Fisher method, the significance of the feature f for recognizing the samples belonging to two classes is measured by the Fisher score, which is defined by [15]

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

where c_1, c_2, and σ_1, σ_2 represent the mean values and standard deviations of feature in the first and second classes, respectively. The higher this value the more significant is the feature in recognition between classes 1 and 2.

In the chi-square method, we calculate the chi-square metric between the target and the particular attribute. The null hypothesis is that there is no relationship between the attribute and the class, so they are independent. The attributes with the maximum chi-squared values are selected as the most important in class recognition [14].

Correlation of the feature with the class (COR) measures how a particular feature relates to the class to which it belongs [1]. In the case of K classes, the correlation measure is defined by [1]

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

where $m = E\{f\}$ is the mean value of feature f for all data, $m_k = E\{f/k\}$ is this mean for kth class, $\mathrm{var}(f)$ is a variance, and P_k probability of kth class in the data set. The highest values of $S(f)$ represent the best features.

The two-sample t-test (2TT) applies the null hypothesis that data in classes 1 and 2 are independent random samples of normal distributions of equal means c_1, c_2, and equal variances. The alternative hypothesis is that the means are not equal. The test statistic is formulated in the form

$$t(f) = \frac{c_1 - c_2}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}, \tag{3}$$

where n and m represent the sample sizes of both classes. The test returns the value 1 or 0, where 1 indicates a rejection of the null hypothesis (feature well recognizing two classes) and 0 a failure to reject the null hypothesis (lack of class discrimination ability). Additional information delivered by the test is the significance p-value of the feature. A small p (≤ 0.05), rejects the null hypothesis. This is strong evidence that the null hypothesis is invalid (the compared populations are significantly different). A large p (> 0.05) means the alternate hypothesis is weak, so you do not reject the null (the compared populations are not statistically different).

The Kolmogorov-Smirnov (KS) test compares the medians of the groups of data to determine if the samples come from

the same population (the same class) [14]. The null hypothesis is that the attribute values representing both classes have the same continuous distribution (do not recognize the classes). In our implementation, the KS test statistic is based on the expression

$$KS(f) = \max\left(\left|F_1(f) - F_2(f)\right|\right), \tag{4}$$

where $F_1(f)$ and $F_2(f)$ are the cumulative distribution of samples of feature f belonging to classes 1 and 2, respectively. The higher this value the better is the class discriminative ability of the feature.

Kruskal-Wallis (KW) test is similar to KS but uses ranks of the data rather than the numerical values. Moreover, the KW test does not make any normality assumptions. The class recognition ability of the feature is based on the p-value for the null hypothesis that all samples are drawn from the same population. The higher this value the least important is the feature.

Stepwise fit (SWF) applies the strategy of sequential adding and removing features to the set of input attributes based on their statistical significance in a regression. It starts with an initial set of features and then compares the explanatory power of the model with a larger and smaller number of features based on p-value in F-statistics. According to these results, the algorithm decides whether a feature should be included in a model or not. As a result, we get the final set of features, regarded by the method as the most important. Additionally, the p-values corresponding to all features are estimated. The smaller this value the more important is this feature in the class recognition problem [1, 13].

Lasso is another method frequently used for feature selection in regression problems. It is usually introduced in the context of least squares and formulated as follows

$$\min_{\mathbf{w}} \frac{1}{p} \left\| \mathbf{d} - \mathbf{X}\mathbf{w} \right\|^2, \tag{5}$$

subject to $\|\mathbf{w}\|_1 \leq t$. In this formulation, $\mathbf{X}$ is the observation matrix representing known measured samples of dimension $p \times N$, where p represents the number of observations and N – number of attributes, $\mathbf{d}$ is the destination vector, and $\mathbf{w}$ – weighting vector corresponding to the attributes. Constant t is a pre-specified free parameter determining the amount of regularization. The weights w_i of vector $\mathbf{w}$ represent the impact of particular attributes of the linear numerical model on the data representation.

In recursive feature elimination (RFE) the network of the linear kernel (for example linear SVM) is trained applying all available input attributes [2]. As a result of learning the weights associated with the input, attributes are arranged in decreasing order. A large absolute value of weight connecting feature f with the output signal means a high discrimination ability of this feature. In the RFE approach, the smallest value features are eliminated sequentially, step by step, and the network is re-trained using smaller and smaller populations of features. The process is repeated until the appropriate number of the most important features is achieved.

In reliefF algorithm, the features are assessed based on their correlation with the class while taking into account the distances between opposite classes [18]. ReliefF chooses randomly the observation $\mathrm{R_i}$ and searches for k of its nearest neighbors belonging to the same class $\left(\text{nearest hits } \mathrm{H_j}\right)$ and k nearest neighbors of the different classes $\left(\text{nearest misses } \mathrm{M_j(C)}\right)$. Each attribute f is associated with the weight $w(f)$. This value is updated depending on hits $\mathrm{H_j}$ and misses $\mathrm{M_j(C)}$ for each observation $\mathrm{R_i}$. If instance $\mathrm{R_i}$ and its nearest neighbors of the set Hj have different values of the attribute f then this attribute separates the instances of the same class (bad quality of attribute). The value of the weight $w(f)$ is then decreased. On the other side when the instance $\mathrm{R_i}$ and its nearest neighbors of different classes $\mathrm{M_j}$ have different values of attribute f then this attribute separates two instances of a different class (desirable case). In such cases, the value of weight $w(f)$ is increased. The adaptation of weights is according to the formula

$$w(f) := w(f) - \left(w(f) - nearHit_i\right)^2 + \\ + \left(w(f) - nearMiss_i\right)^2, \tag{6}$$

The final result for each attribute is the normalized average of the contribution of all hits and misses calculated for all observations. The higher the value of the weight, the more important is the feature.

The next selection method is based on the nearest neighborhood component analysis (NCA) [19]. Each vector of attributes searches for K nearest neighbors (usually using KNN classifier). The distances between two vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ are scaled by the weight vector $\mathbf{w}$ using $\mathrm{L_1}$ metric

$$D\left(\mathbf{x}_i, \mathbf{x}_j\right) = \sum_{l=1}^{N} w_l^2 \left| x_{il} - x_{jl} \right|, \tag{7}$$

The weight w_l (the element of vector $\mathbf{w}$) is associated with the l-th feature. The higher this value the more important is the feature. In the KNN classification process, the choice of neighbors is done according to the probability p_{ij} that vector x_i is the nearest neighbor of x_j. This probability is defined as follows

$$p_{ij} = \begin{cases} \dfrac{e^{-D(\mathbf{x}_i, \mathbf{x}_j)/\sigma}}{\sum\limits_{k \neq i} e^{-D(\mathbf{x}_i, \mathbf{x}_j)/\sigma}} & i \neq j \\ 0 & i = j. \end{cases} \tag{8}$$

As a result, the features are arranged according to the decreasing values of the weights. The first positions in this series indicate the features of the highest class discrimination ability.

The genetic algorithm used in feature selection represents a special version of biologically inspired genetic processes. The real features represented by vector $\mathbf{x}$ are binary coded in the chromosome population used in the genetic processes of selection, crossover, and mutation. Value 1 represents here the

acceptation of a particular feature as the input attribute and 0 – its elimination from the list of attributes [12, 13]. The fitness function is defined through the cross-validation error E committed by the applied classification system on the chosen validation set of observations while minimizing the number of active features (represented by binary value 1). The typical form of the fitness function is defined as follows

$$f_{fit} = -\left(E + \alpha \frac{N_c}{N}\right), \tag{9}$$

where E represents the cross-validation error, N_c – the actual number of active features, N – population of the original (full) set of features, and α the regularization coefficient. The positions of binary value 1 in an optimal vector indicate the features of the highest discrimination value.

A random forest of decision trees is another efficient method of feature selection [20]. The ensemble of multivariate decision trees is trained on the original samples of the learning database. After finishing the learning process, the particular input attributes (features) in all decision trees which form the forest, are subject to perturbation and the resulting change of accuracy of the ensemble is computed. The degree of worsening of the classification results corresponding to perturbation of the particular feature is its importance measure. The higher the rate of the deterioration of ensemble results the more important is the feature.

3. Numerical experiments

The numerical experiments aim to compare different methods of feature selection and their application in the recognition of gene biomarkers in two types of cancers. Two types of data have been checked. The first one is the synthetic classification problem, for which the set of important features is known in advance. The second is the real-life problem to select the most important genes (the so-called biomarkers) in a dataset of gene expression microarray in prostate and ovarian cancers.

3.1 Synthetic data. The feature selection in synthetic classification problems is aimed at comparing different selection algorithms, especially concerning their sensitivity to the noise corrupting the data. The observations data X belonging to 2 classes (vector d) have been defined based on some deterministic and random variables using the following Matlab code

```
N = 100; %number of observations
% 20 INPUT ATTRIBUTES for observations
i=1:100
X = [sin(2*pi/N*i), cos(2*pi/N*i), sin(3.2*pi/N*i), cos(5.3*pi/N*i)
sin(7.4*pi/N*i), cos(1.5*pi/N*i), rand(1,14)];
%NOISE of alpha subject to change
noise =alpha*rand(size(y));
y =-2*(X(:,1))-(X(:,2)+X(:,3)+X(:,4)+X(:,5)+ 0.3*X(:,6)+noise);
% CLASS d
d=sign(y)
```

The class membership of the observations depends on strictly defined variables: x_1, x_2, x_3, x_4, x_5, and x_6. Only these variables influence the desired output. The other variables (from x_7 to x_{20}) are dummy and have no real impact on the target. However, they will also take part in the selection process. Moreover, the generated target is subject to the noise of different SNR values, disturbing the output signal.

The numerical investigations of the selection algorithms compare the quality of feature selection methods at different amounts of noise. Observe, that the real attributes influencing the target class are known in advance.

Table 1 presents the results of discovering the first six most important features, arranged in an order, suggested by a particular algorithm. The order of features is organized according to their class discrimination ability suggested by the selection method. So, the first feature is regarded as the most important. The distant position in the series means that the feature is less important. Only the first 6 selected features are presented. In the case, when the true feature did not arrive in the set its position in the ranking is presented below.

Table 1
Six the most important features chosen by different selection algorithms at changing the SNR ratio. The true set of the features influencing the output variables involves the variables x_1, x_2, x_3, x_4, x_5, and x_6

Method	Noiseless	SNR = 2 dB	SNR = –2.8 dB	SNR = –7.5 dB
KS	6,1,2,5,4,3	6,1,2,5,4,3	6,2,1,3,4,5	6,2,1,3,4,5
2TT	6,1,2,5,4,3	6,1,2,5,4,3	6,2,1,4,3,5	6,2,1,4,3,5
KW	6,1,2,5,3,17 4 → pos. 8	1,6,2,3,5,17 4 → pos. 8	1,6,2,5,3,7 4 → pos. 7	1,6,2,3,14,5 4 → pos. 10
COR	6,1,2,3,5,19 4 → pos. 8	6,1,2,5,3,8 4 → pos.9	6,2,3,1,5,8 4 → pos.7	6,3,4,1,2,19 4 → pos.17
Fisher	6,1,3,2,5,19 4 → pos. 8	6,1,2,5,3,8 4 → pos. 9	6,2,1,3,5,8 4 → 7 pos.	6,1,3,4,19,2 5 → pos.17
Lasso	1,3,5,4,2,6	1,4,5, 3,6,2	1,6,5,4,14,2 3 → pos.7	6,15,18,1,4,3 5 → pos.7 2 → pos.19
SWF	1,3,5,4,2,6	5,1,3,2,4,6	1,2,4,3,6,5	6,4,15,19,3,12 5 → pos.14 2 → pos.15
NCA	5,2,4,1,3,6	5,2,4,1,3,6	2,1,5,4,6,3	4,1,2,3,15,16 6 → pos.11 5 → pos.12
ReliefF	1,6,4,3,5,2	2,1,3,6,4,5	2,6,5,3,1,13 4 → pos. 9	15,11,2,4,13,19 3 → pos.9 1 → pos.11 5 → pos.14 6 → pos.16
RFE	1,2,5,3,4,6	1,2,3,4,5,6	1,2,3,4,6,5	1,2,3,4,5,7 6 → pos.9
RF	6,5,2,13,1,4 3 → pos. 7	6,5,2,12,3,1 4 → pos.7	6,2,5,16,13,1 3→pos.7, 4→p.10	6,19,5,13,7,2 1 → pos.10 4 → pos.11 3 → pos.16

The results show that selection algorithms that rely their operation on different principles, generate various order of feature importance. Only two methods (KS and 2TT), both applying the statistical hypothesis tests, were able to discover the full set of 6 true features, irrespective of the noise injected into the input data. The correlation between both methods is very high and the Pearson correlation coefficient, in this case, was over 0.94 irrespective of the applied noise level.

The sequence of attribute importance, treated as the diagnostic features, changes a lot in different methods. However, in the noiseless case and small level of noise, the contents of the selected set were practically close to perfect. The most sensitive to the high noise disturbing the data seems to be refiefF (4 wrong features in the set of 6 at the high level of noise) and random forest (3 wrong features in the set of 6). Nonetheless

3.2. Gene expression microarray in ovarian and prostate cancers. The main experiments have been conducted on a real-life problem of discovering the most important genes in the expression microarray of cancers (so-called biomarkers). Tumor formation involves simultaneous changes in hundreds of cells. The variations in gene expression of microarray provide a platform for a simultaneous testing large set of genetic samples. It can help in the identification of cancer biomarkers (the most class discriminative genes). Thanks to such investigations we can compare different patterns of gene expression levels between a group of cancer and a group of reference patients. In this way, we can identify the genes, which are the most associated with particular cancer (so-called biomarkers).

The sources of difficulties in identifying such genes are different. First, there is usually a very small number of observation records concerning the number of genes (hundreds of records compared to tens of thousands of genes). The second problem is the quality of data: many outliers, high variance of data, and bad conditioning of the problem [10, 21–23]. These complexities raise the challenge of how to identify the genes, that are the most informative for such disorder.

The numerical experiments have been performed for two types of tumors: ovarian and prostate cancers. They have been obtained from a publicly available database containing the gene expression arrays [21–23]. In the case of prostate cancer, one class corresponds to the gene expressions of the prostate tumor cases (52 records) and the second to non-tumor cases (50 records) representing the reference class. In the case of ovarian 91 records represent tumor and 162 the reference class. The basic information regarding the distribution of data records is gathered in Table 2.

The second column in this table is composed of three numbers: the first one – the total number of data, the second – the number of patients belonging to the first class, and the third – the number of patients of the second class. The third column provides the number of genes for the particular cancer type.

Table 2
Database of microarrays used in experiments

Prostate Tumor (PRT)	102/50/52	10509
Ovarian (OV)	253/162/91	15155

As a result, the data are organized in the form of matrix **X**, with the rows representing the patient records and columns – the genes in the expression array. The large difference between the number of records (in the range of hundreds) and the number of genes (more than 10 thousand) is evidence of bad conditioning and high difficulty of the recognition task. The second problem is the large diversity of expression values corresponding to different patients within the same group. They change a lot, and there are many outliers that are very different from the average of the population.

Figure 1 shows such an example of one gene expression in prostate data for 102 individuals. The first 52 samples represent cancer cases, the other samples the reference class. Few outliers can be observed in the first group, while the second group is more uniform, although the expression values change a lot within both groups. Moreover, there is no significant difference between statistics within the first and second class samples (after eliminating outliers). A similar situation is observed for other genes. This makes the task of selecting the most class discriminative genes more difficult.

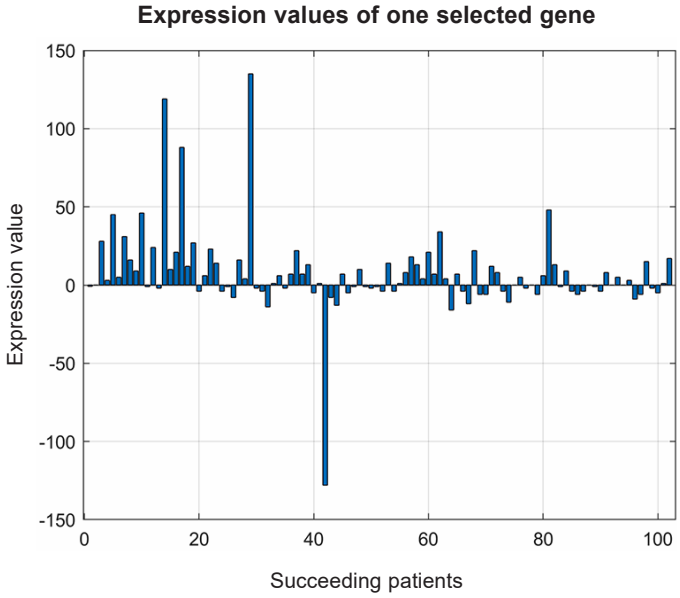

Fig. 1. The exemplary expression values of one gene in prostate cancer

In further experiments the genes with a high number of outliers have been removed from further analysis, treating them as not reliable.

The scale of difficulties in the class recognition problem, while taking into consideration all genes, is well illustrated by the visual distribution of samples belonging to both classes. This is shown in Fig. 2, by mapping multidimensional data sets into the two-dimensional coordinate system at the application of Stochastic Neighbor Embedding (TSNE) [2, 13]. Figure 2a

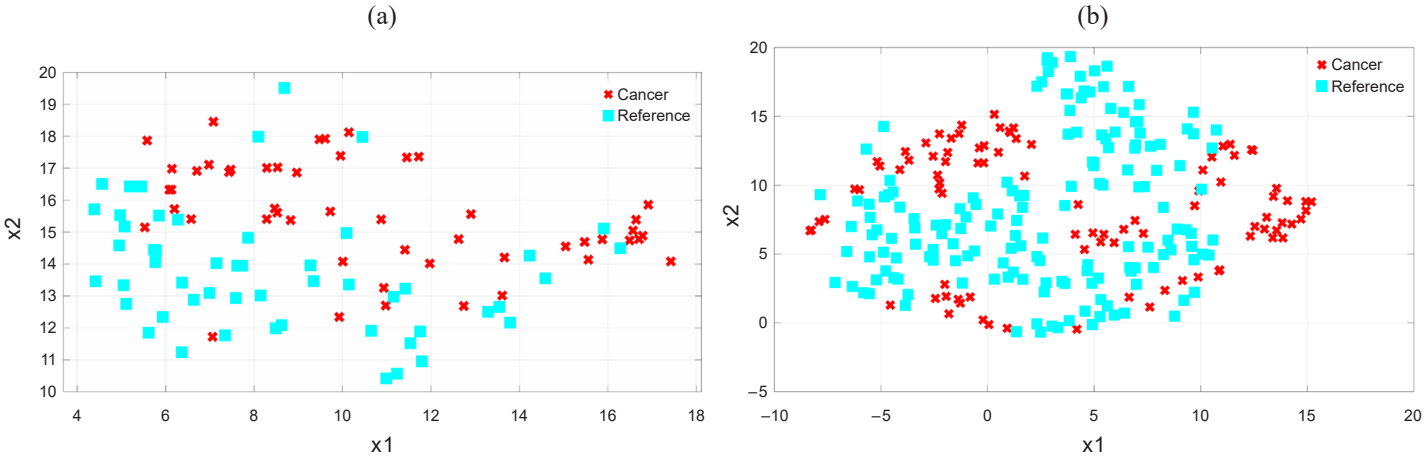

Fig. 2 The distribution of gene expression values represented by all gens mapped into a 2-D system using TSNE: a) prostate, b) ovarian. The regions of samples representing both families are mixed

refers to the prostate and Fig. 2b to ovarian. It is evident, that samples of both classes interlace each other in all regions of 2-D space. The recognition of classes in such cases is dubious and will be loaded by large errors.

The aim is to select these genes, which provide the best visible separation of cancer samples from the reference ones. Such genes can be treated as biomarkers of particular cancer. The choice of such genes is achieved in the work by using different methods of feature selection and the subsequent merging of the individual results. Only the top, most valuable genes in the series are considered in each method.

Based on the introductory experiments we have selected a few methods, which are along with the best in feature selection. To such methods belonged: Fisher, 2TT, ReliefF, KS, KW, COR, SWF, and NCA. Note, that they are not compatible with the results obtained in the synthetic problem. Table 3 shows the indexes of genes, selected by particular methods, presented in the sequence according to their importance [24]. The first position in the list means the highest importance.

High repeatability of some genes can be observed for both types of tumors. For example in ovarian, the gene 1681 was selected by all methods (on the first or second position). It seems to be the most significant. A similar situation is for the genes 1680, 1682, and 1683. These genes might be treated as biomarkers. There are also several genes selected only by a single method. To such genes belong 99, 543, 1675. Additionally, they appear on far positions and therefore, should be excluded from the set of potential biomarkers.

In the case of the prostate, the situation is a bit more complex. The best genes of the numbers: 7516 (selected 6 times), 6069 (selected 6 times), and 2719 (selected 6 times) have been commonly chosen by 6 methods in the first 10 positions (the most important). The results presented in Table 3 suggest, that the stepwise fit (SWF) method is not efficient in the prostate, since none of these mentioned 3 genes were selected among the best ten.

Taking into account the contents of the selected sets of genes and their sequence it is possible to select the most discriminative genes. This was done by summing the positions of genes in the sets generated by the particular methods. The position of the gene is treated as its weight.

Table 3
The set of 10 genes selected individually by different methods

Selection method	Prostate	Ovarian
Fisher	7516, 2719, 6069, 4565, 6822, 7539, 2646, 7530, 8902, 5531	1681, 1680, 1682, 1683, 1679, 1684, 2238, 1685, 2239, 2237
2TT	7516, 2719, 6069, 4565, 6822, 7530, 5531, 2646, 7539, 4353	1681, 1680, 1682, 1683, 2238, 2239, 1679, 2237, 2240, 1684
ReliefF	4565, 7547, 7516, 2719, 8786, 5531, 6822, 3471, 6073, 7638	1681, 1680, 1682, 1683, 1679, 2237, 1684, 2238, 2239, 2240
KS	3125, 6069, 7530, 7516, 2719, 8010, 6822, 4224, 5108, 4274	1680, 1681, 1682, 1679, 1683, 2238, 1684, 2239, 2237, 1685
KW	7516, 6069, 2719, 7530, 3125, 8010, 3618, 6716, 6822, 4565	1680, 1681, 1682, 1679, 1683, 2238, 1684, 2239, 1685, 2237
COR	7516, 2719, 6069, 4565, 6822, 7530, 7539, 5531, 2646, 4353	1681, 1680, 1682, 1683, 1679, 1684, 1685, 1686, 1737, 2193
SWF	60, 212, 647, 949, 1408, 1412, 1674, 1814, 2073, 2296	1681, 2238, 1800, 1488, 1647, 183, 182, 6781, 99, 2310
NCA	8010, 2495, 8902, 7468, 8125, 7312, 4617, 3414, 6883, 6069	1680, 1681, 1682, 182, 183, 1683, 543, 544, 2241, 1675,

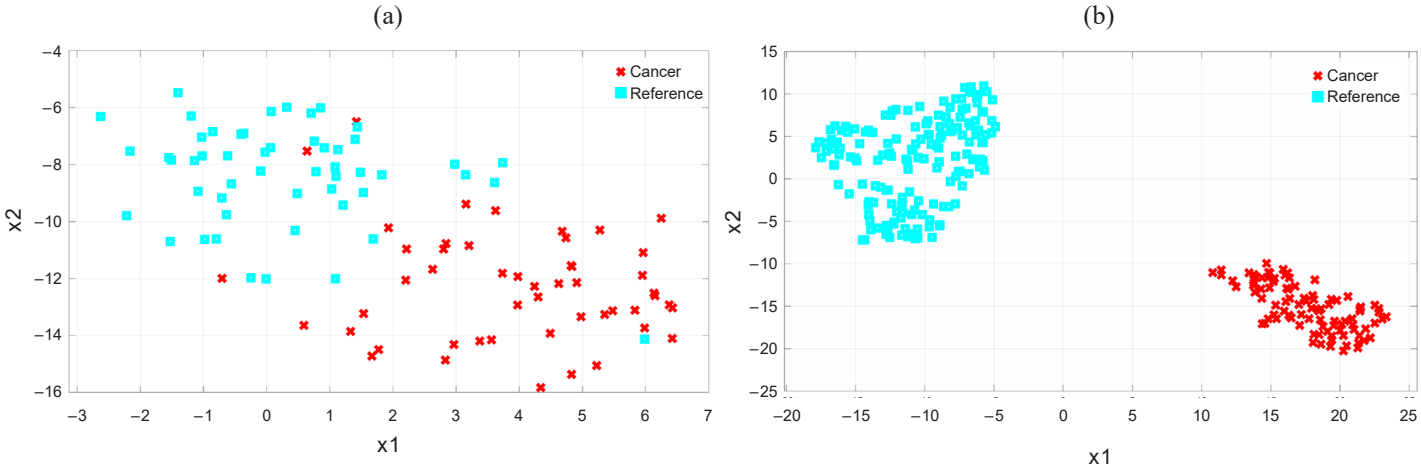

Fig. 3. The visual distribution of gene expression values represented by 10 biomarkers mapped into a 2-D system using TSNE: a) prostate, b) ovarian

The gene with the smallest sum of position numberings is treated as the most valuable biomarker. Only the first 20 genes in the sequence have been taken into account in further considerations. The absence of the particular gene in such a set resulted in assuming its weight equal to 21. Table 4 shows the contents of the most valuable genes (biomarkers) obtained in this way for both considered tumors.

Table 4
The set of 20 selected genes treated as possible biomarkers

Tumor	Genes
Prostate	7516, 2719, 6069, 4791, 8010, 6822, 4565, 5317, 5108, 8398, 3125, 9879, 8427, 3535, 8902, 6716, 7530, 7267, 4353, 4670
Ovarian	1681, 1680, 1682, 2238, 183, 182, 2310, 1739, 1680, 2239, 2240, 1737, 2236, 1738, 1679, 2241, 2237, 1595, 544, 545

To assess the quality of the chosen biomarkers all observations represented by the selected genes have been mapped into a 2-D system using TSNE. A different number of genes was tried out, looking for the smallest possible number capable of separating both classes. The resulting distribution of samples belonging to two classes (cancer versus reference) for both tumors (prostate and ovarian) at the gene representation limited to ten are presented in a visual form in Fig. 3.

Comparing these distributions with the results corresponding to all genes, depicted in Fig. 2, we can see a significant improvement. It is evident now, that 10 selected genes separate well both classes of data. In the case of ovarian, the classes are ideally separated. The distance between their centers is very large and the standard deviation relatively low. In the case of the prostate, the distribution of samples is less ideal, however, still, both classes are relatively well separated. Only a small number of samples interlace with each other (three cancer and one reference). They might be treated as outliers. However, the standard deviation among samples is much higher in comparison to ovarian.

Tables 5 and 6 show the numerical characterization of data distribution in both classes at a representation of data by all genes and 10 randomly selected genes and 10 carefully selected biomarkers.

Table 5
Numerical characterization of the data distribution of both classes in the case of ovarian

	$\lvert c_1 - c_2 \rvert$	σ_1	σ_2	$\frac{\lvert c_1 - c_2 \rvert}{0.5(\sigma_1 + \sigma_2)}$
All genes	0.01	0.18	0.17	0.03
10 random genes	0.06	0.17	0.17	0.38
10 biomarkers	0.24	0.15	0.11	1.86

Table 6
Numerical characterization of the data distribution of both classes in the case of the prostate

		σ_1	σ_2	$\frac{\lvert c_1 - c_2 \rvert}{0.5(\sigma_1 + \sigma_2)}$
All genes	1.25	36.13	25.66	0.04
10 random genes	0.34	15.26	11.96	0.02
10 biomarkers	13.59	14.39	8.39	1.19

The values depicted in the Table represent the distance between the centers of both classes $|c_1 - c_2|$, the standard deviations σ_1 and σ_2 of samples belonging to classes 1 and 2, as well as the relation of the distance between centers to the average of standard deviation.

The advantage of the application of biomarkers is evident. The distance between both classes is much higher than in the

case of a random choice of genes. The same conclusion is true for standard deviation (biomarkers represent smaller deviation). However, the differences are now not as large as in the case of centers. The most significant difference is in the relative distances between classes, taking into account the deviations. In the case of ovarian, the improvement is 5:1, while for the prostate this ratio is 59:1. These numerical relations have been estimated for the same number (10) of chosen genes (random choice and selected set).

4. Conclusions

The paper has studied the class discriminative properties of different feature selection methods. Among the many existing approaches to feature selection, we have selected only 11, the most representative in this area. Two selection problems have been used in testing. The first one was a synthetic problem, in which the true diagnostic features were known in advance. The selection methods have been used to discover them in the presence of different levels of noise.

Based on these experiments, a limited number of the best methods were selected to solve the real problem of selecting gene biomarkers in microarray gene expression data for two types of tumors: ovarian and prostate tumors.

The results generated by particular methods have been combined to define the smallest set of the most important genes (biomarkers). Their application in data representation has shown very good class discrimination ability of these genes. This fact was confirmed by a visual representation of both data classes and by numerical results regarding the statistical quality measures of clusters representing both classes.

The next research will be directed to develop the deep learning approach to feature selection and compare the results with these presented now. More experiments performed on the larger set of cancer cases are also planned.

Acknowledgement: This work was supported by Military University of Technology under research project UGB 22–850.

References

[1] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection", *J. Mach. Learn. Res.* 3, 1158–1182 (2003).

[2] I. Guyon, A.J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using SVM", *Mach. Learn.* 46, 389–422 (2003).

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

[4] H. Chen, Y. Zhang, and I. Gutman, "A kernel-based clustering method for gene selection with gene expression data", *J. Biomed. Inform.* 62, 12–20 (2016).

[5] P. Das, A. Roychowdhury, S. Das, S. Roychoudhury, and S. Tripathy, "sigFeature: novel significant feature selection method for classification of gene expression data using support vector machine and t statistic", *Front. Genet.* 11, 247 (2020), doi: 10.3389/fgene.2020.00247.

[6] A. Wiliński and S. Osowski, "Ensemble of data mining methods for gene ranking", *Bull. Pol. Acad. Sci. Tech. Sci.* 60, 461–471 (2012).

[7] H. Mitsubayashi, S. Aso, T. Nagashima, and Y. Okada, "Accurate and robust gene selection for disease classification using simple statistics, *Biomed. Inform.* 391, 68–71 (2008).

[8] J. Xu, Y. Wang, K. Xu, and T. Zhang, "Feature genes selection using fuzzy rough uncertainty metric for tumour diagnosis", *Comput. Math. Method Med.* 2019, 6705648 (2019), doi: 10.1155/2019/6705648.

[9] B. Lyu and A. Haque, "Deep learning based tumour type classification using gene expression data", *bioRxiv*, p. 364323 (2018), doi: 10.1101/364323.

[10] F. Yang, "Robust feature selection for microarray data based on multi criterion fusion", *IEEE Trans. Comput. Biol. Bioinf*. 8(4), 1080–1092 (2011).

[11] M. Muszyński and S. Osowski, "Data mining methods for gene selection on the basis of gene expression arrays", *Int. J. .Appl. Math. Comput. Sci.* 24(3), 657–668 (2014).

[12] T. Latkowski and S. Osowski, "Data mining for feature selection in gene expression autism data", *Expert Syst. Appl.* 42(2), 864–872 (2015).

[13] Matlab user manual. Natick (USA): MathWorks: (2020).

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

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

[16] Exxact. [Online]. https://blog.exxactcorp.com/scikitlearn-vs-mlr-for-machine-learning/

[17] Tutorialspoint. [Online]. https://www.tutorialspoint.com/weka/weka_feature_selection.htm

[18] R. Robnik-Sikonja, and I. Kononenko, "Theoretical and empirical analysis of Relief", *Mach. Learn.* 53, 23–69 (2003).

[19] W. Yang, K. Wang, and W. Zuo. "Neighborhood Component Feature Selection for High-Dimensional Data", *J. Comput.* 7(1), 161–168 (2012).

[20] L. Breiman, "Random forests", *Mach. Learn.* 45, 5–32 (2001).

[21] NCBI database. [Online]. http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS4431, (2011).

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

[23] http://sdmc.lit.org.sg/GEDatasets/Datasets.html

[24] F. Gil and S. Osowski, "Feature selection methods in gene recognition problem", in *Proc. on-line Conference Computatational Methods in Electrical Engineering*, 2020, pp. 1–4.