


Article

Technique of Gene Expression Profiles Extraction based on the Complex Use of Clustering and Classification Methods

Sergii Babichev ^{1,2,†,‡}  and Jiří Škvor ^{1,‡}

¹ Jan Evangelista Purkyně University in Ústí nad Labem, Ústí nad Labem, Czech Republic; sergii.babichev@ujep.cz, Jiri.Skvor@ujep.cz

² Kherson State University, Kherson, Ukraine

* Correspondence: sergii.babichev@ujep.cz; Tel.: +420-777-843-785

† Current address: Pasteurova 3632/15, 400 96 Ústí nad Labem, Czech Republic

‡ The authors contributed to this work as follows: the first author – 80%, the second one – 20%.

Abstract: In this paper, we present the results of the research concerning extraction of informative gene expression profiles from high-dimensional array of gene expressions considering the state of patients' health using clustering method, ML-based binary classifiers and fuzzy inference system. Applying of the proposed stepwise procedure can allow us to extract the most informative genes taking into account both the subtypes of disease or state of the patient's health for further reconstruction of gene regulatory networks based on the allocated genes and following simulation of the reconstructed models. We used the publicly available gene expressions data as the experimental ones which were obtained using DNA microarray experiments and contained two types of patients' gene expression profiles: the patients with lung cancer tumor and healthy patients. The stepwise procedure of the data processing assumes the following steps: in beginning, we reduce the number of genes by removing non-informative genes in terms of statistical criteria and Shannon entropy; then, we perform the stepwise hierarchical clustering of gene expression profiles at hierarchical levels from 1 to 10 using SOTA clustering algorithm with correlation distance metric. The quality of the obtained clustering was evaluated using complex clustering quality criterion which is considered both the gene expression profiles distribution relative to center of the clusters were these gene expression profiles are allocated and the centers of the clusters distribution. The result of this stage execution was selection of the optimal cluster at each of the hierarchical levels which corresponded to minimum value of the quality criterion. At the next step, we have implemented classification procedure of the examined objects using four well known binary classifiers: logistic regression, support-vector machine, decision trees and random forest classifier. The effectiveness of the appropriate technique was evaluated based on the use of ROC analysis using criteria included as the components the errors of both the first and the second kinds. The final decision concerning extraction of the most informative subset of gene expression profiles was taken based on the use of fuzzy inference system, the inputs of which are the results of the appropriate single classifiers operation and output is the final solution concerning state of the patient's health. To our mind, the implementation of the proposed stepwise procedure of the informative gene expression profiles extraction create the conditions for increasing effectiveness of the further procedure of gene regulatory networks reconstruction and the following simulation of the reconstructed models considering the subtypes of the disease and/or state of the patient's health.

Keywords: gene expression profiles; lung cancer; clustering; classification; ML-based binary classifiers; SOTA clustering algorithm; clustering quality criteria; ROC analysis; fuzzy inference system

1. Introduction

The use of gene expression datasets for reconstruction of gene regulatory networks (GRN) and simulation of the reconstructed models is one of the topical directions of current bioinformatics [1–4]. GRN in this case is a group of molecular elements interconnections that determines the functional possibilities of biological organism. Qualitatively reconstructed GRN allows us to understand the particularities of genes interconnections, differences of these interconnections for healthy and ill cells in order to create both new effective medicines and methods to treat complex diseases, such as Alzheimer, Parkinson, various types of cancer, etc. The results of both DNA microchip experiments and mRNA molecules sequencing methods are used to form the gene expression data nowadays [5,6]. In the first case, we have as a result the matrix of light intensities, the values of which are proportional to expression of appropriate gene (level of gene activity). Transformation of these light intensity into expression values assumes implementation of four steps: background correction [7–10], normalization [7,11–15], PM correction and summarization [7,13,16,17]. In the second case, the initial data is presented as a matrix of genes count, the values of which are varied in very wide range. In this case, the first step of the data processing involves a transform of this matrix into gene expression matrix using appropriate mathematical functions [18]. However, in any case, we receive as a result the high dimensional matrix of gene expressions, where quantity of genes is varied from 50 to 60 of thousands of genes. Under the gene expression profile in this case, we understand a set of gene expressions, the values of which are evaluated for various samples or under dissimilar conditions of the experiment carrying out. Each of the profile values corresponds to appropriate sample. Informative genes extraction in terms of current problem is the first task which should be solved at the stage of the experimental data pre-processing. The informative genes extraction in this case means that it is necessary to extract mutually correlated gene expression profiles in terms of resolving ability of the studied samples (healthy and not-healthy patients or subtypes of disease). Biclustering technique is applied to solve this problem in the most cases nowadays [19–21]. Each of the biclusters contains a set of mutually correlated genes and samples. However, direct applying this technique to high-dimensional array of gene expressions leads to large number of biclusters and the choice from them the informative sets is very difficult and unsolved task nowadays. Moreover, in the most cases biclusters contains not complete set of samples. This fact also limits the range of the gene expression values variation during further simulation process.

Hereinbefore presented facts indicates the relevance of the research concerning extraction of groups of informative genes considering particularities of the investigated objects for purpose of further reconstruction of GRN based on the extracted genes and simulation of the reconstructed models. Within the framework of this research, we solve this problem based on the complex apply of classification and clustering techniques with the use of fuzzy inference system at the final step of decision making concerning extraction of set of the informative gene expression profiles.

1.1. Problem Statement

The initial dataset is presented as a matrix of gene expressions: $(e_{ij}) \in R^{n \times m}$, where n and m are the number of samples and genes respectively. We suppose that the samples can be divided into previously known classes. The main problem consists of extraction of genes which allow us to divide the samples into classes maximally correctly in terms of the used criteria.

1.2. Literature Survey

There are a lot of works which are devoted to gene expression data processing nowadays. So, in [22] the authors considered reducing the non-informative genes expression profiles using both Shannon entropy and statistical criteria. They supposed that gene expression profile can be removed

from the data if its Shannon entropy value is larger and variance and average of absolute values are less in comparison with appropriate boundary values. To determine the boundary values the authors used fuzzy inference system and clustering quality criteria. The result of the proposed technique applying is removing genes which has zero or low expression values for all samples (lowly expressed genes), low level of gene expression variation for samples various types (do not allow distinguishing samples) and chaotic variation of the expression values for investigated samples (high value of Shannon entropy). In this study, we have applied the results of the authors research.

The papers [23,24] considered the issues concerning bicluster analysis of genes expressions data. Implementation of this technique allows extracting groups of mutually correlated rows and columns. In [23] the reserchers presented an enhanced version of Pearson's correlation coefficient (PCC) to achieve better biclustering-enabled co-expression analysis. the obtained results were established both statistically and biologically using benchmarked gene expression data. In [24] the authors proposed a novel approach for gene expression data biclustering with the use of fusion of differential evolution framework and self-organizing Kohonen's map (SOM). The proposed approach was applied on two real-life microarray gene expression datasets and the obtained results were compared with various current techniques. The papers [25,26] presents the research results concerning implementation of various clustering techniques for single-cell RNA sequencing data processing. Within the framework of the research, the authors carried out four experiments using two big scRNA-seq datasets with the use of twenty models. The obtained results allowed authors to conclude that the proposed feature extraction increased the quality of high-dimensional and sparse scRNA-seq data. The authors have also shown that proposed feature-extraction techniques can promote to the clustering performance.

The issues concerning genes extraction to solve the problem of cancer types classification are considered in [27]. The authors proposed a new hybrid wrapper procedure applying of which allows combining the parameters of teaching learning-based algorithm and gravitational search algorithm. They have shown also that proposed technique is expressively outmatch existing metaheuristic methods relating to convergence rate, classification accuracy and optimal quantity of used features. A new multi-classification technique based on combining the probabilistic support vector machine and elastic net was described in [28]. Applying this technique can allow solving the problem of cancer detection using gene expression profiles data of platelets. The authors applied within the framework of the research the probabilistic support vector machine in order to produce the outputs of the binary classifiers with class-specific features matching. The obtained results have shown that the presented technique is well-suited for traditional multi-classification tasks in the case using datasets with high-dimension of features and small quantity of samples.

In [29], the authors proposed a new approach for semi-supervised classification of time-series. The proposed technique techniques learn both from labeled and unlabeled data. The authors have shown that the proposed approach approach substantially outperforms the state-of-the-art semi-supervised time-series classifier. The results of the research concerning the use of hubness-aware semi-supervised approach for classification of high dimensional gene expression data are presented in [30]. The author proposed a self-training semi-supervised extension of Naive Hubness-Bayesian k-Nearest Neighbor. The author has also shown that the proposed approach can increase the classification accuracy and reduce computational costs. In [31], the authors considered issues focused to classification of gene expression data using extreme learning machines with regularization. The authors compared the proposed technique with different regularization strategies in context of a binary classification task related to gene expression data. The [32] presents the results of the research concerning development of non-invasive method of recognition of finger skin based on K-NN classifier. The authors have shown that the proposed approach can help us to diagnose pathologies of human skin.

However, we would like to note that accuracy of the classifier operation in the case of the use of high dimensional gene expression data depends on the vector of the extracted genes which are used as the classifier inputs. The perspective of our research is reconstruction of gene regulatory network based on the extracted genes and following simulation of the reconstructed models. In this case, the extraction

of optimal subset of gene expression profiles can increase the informativity of the reconstructed gene regulatory network and, as a result, it can create the conditions for better understanding the character of genes interconnections during the following simulation process considering both the state of the patient's health or subtype of disease. In [35], we solve this problem based on stepwise apply of clustering and biclustering techniques. Implementation of this procedure allowed us to remove gene expression profiles which were identified as noise using density based DBSCAN clustering algorithm. Then, we divided the set of remaining genes into two subsets using SOTA clustering algorithm. At the final step, we applied the bicluster analysis to the obtained subset of gene expression profiles. To our mind, the main disadvantage of this technique is the following: the gene expression data were divided without considering the type of the used samples (state of the patients' health or subtype of the disease). We used in this case only appropriate quantitative criteria. This fact can influence the quality of the reconstructed gene regulatory networks. This problem can be solved by using current techniques, models and information technologies which are used successfully in various fields of scientific research nowadays [33,34]. Within the framework of this research, we propose the solution of this problem based on the complex use of clustering techniques, ensemble of binary classifiers and fuzzy inference system using various quantitative quality criteria of both the clustering and classification procedures implementation.

The aim of the paper is the development of a technique of stepwise gene expression data extraction on the basis of complex use of cluster analysis, binary classifiers and fuzzy inference system. To our mind, it can contribute to increase the objectivity of informative genes selection considering the state of the patients' health for purpose of both further gene regulatory networks reconstruction based on the allocated genes and simulation of the reconstructed models.

2. Materials and Methods

2.1. General Procedure of the Problem Solving

Figure 1 shows the structure chart of stepwise procedure of gene expression data processing which was implemented within the framework of current research. As it can be seen from Figure 1, implementation of this procedure assumes solving the following tasks:

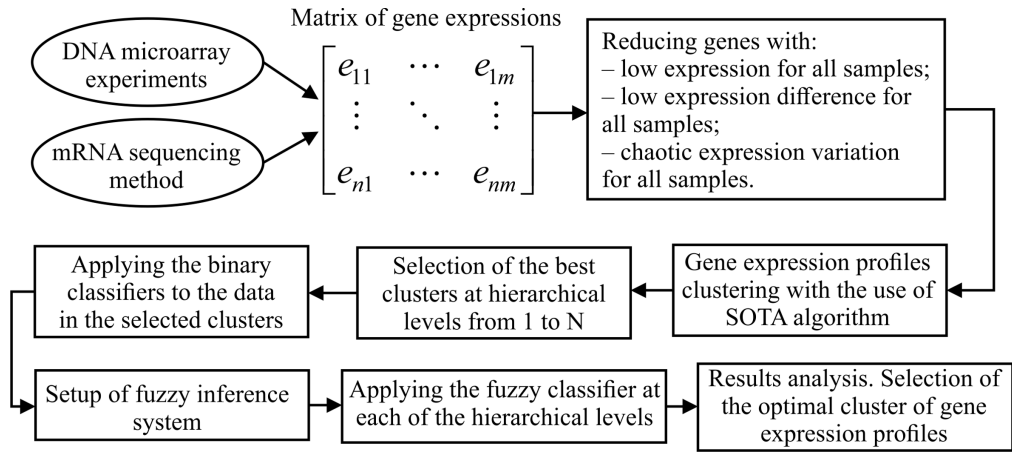


Figure 1. A structural chart of stepwise procedure of gene expression data processing

- formation of the matrix of gene expressions for the investigated samples. In the case of the use of DNA microarray experiments technique, this step involves background correction, normalization, PM correction and summarization. In the case of mRNA molecules sequencing method use, this step assumes allocation of genes count matrix and following transforming the values of this matrix into suitable range;

- extraction of genes which are identified as informative in terms of absolute value of gene expressions, variance and Shannon entropy.
- hierarchical clustering of gene expression profiles at the levels from 1 to N using SOTA clustering algorithm with correlation distance metric;
- division of the samples in allocated clusters of gene expression values into previously known classes and calculation of the quality criteria considering both the samples distribution within the appropriate classes and the distance between the samples in different classes;
- selection of the best clusters in terms of the used criteria at each of the hierarchical levels. These clusters correspond to the extreme values of the used quality criteria;
- applying the binary classifiers to the data in obtained clusters at each of the hierarchical levels. Formation of the intermediate solutions for each of the used classifiers and for each of the selected clusters;
- setup of fuzzy inference system. Definition of the membership functions for both the input and output variables, setup of ranges of the input and output parameters variation, knowledge base formation;
- applying the fuzzy classifier at each of the hierarchical levels;
- results analysis. Selection of the optimal cluster of gene expression profiles in terms of the used criteria.

2.2. Gene Expression Profiles Reducing

Implementation of this stage assumes removing genes, expressions profile of which were identified as non-informative in terms of variance, average of absolute values of gene expressions and Shannon entropy. We assumed that if gene expressions for all samples (averages of gene expressions are small) and differences of gene expression values for different samples (variances) are small, and if expression values for various samples are varied chaotically and this fact does not allow us to identify correctly the classes of the examined samples (Shannon entropy values are large), then, this gene can be removed from dataset as non-informative one. The value of Shannon entropy for each of the gene expression profiles was calculated using James-Stein shrinkage estimator technique [36].

To evaluate the appropriate criteria boundary values, we apply the technique presented in [22]. Applying this technique involves the following:

- calculation of variance, average of absolute values and Shannon entropy for each of the genes expression profiles. Formation of both the ranges of these criteria variation and steps of their values change;
- formation of clusters of the examined samples considering the data annotation. In the case of our dataset use, the samples can be divided into two clusters (with tumor and healthy samples);
- Determination of clustering quality criterion which is calculated at each step of change of the used criteria. Within the framework of our research, we used as the clustering quality criterion the multiplicative combination of WB-index [38] and Calinski Harabasz criterion [37]:

$$QC_{int} = \frac{QC_{WB}}{QC_{CH}} = \frac{K(K-1)QCW^2}{(N-K)QCB^2}; \quad (1)$$

where QCW and QCB are calculated as an average distance from objects to centers of the clusters where these objects are allocated and between centers of the clusters respectively:

$$QCW = \frac{1}{N} \sum_{s=1}^K \sum_{i=1}^{N_s} d(x_i^s, C_s) \quad (2)$$

$$QCB = \frac{2}{K(K-1)} \sum_{i=1}^{K-1} \sum_{j=i+1}^K d(C_i, C_j) \quad (3)$$

Here, K is the clusters quantity; N is the number of samples; N_s is the number of samples in the cluster s ; x_i^s is the i -th sample in the cluster s ; C_i , C_j and C_s are the centers of the clusters i , j and s respectively; $d(\cdot)$ is the distance metric between vectors of gene expressions. Considering high dimension of the gene expressions vectors, we used the correlation distance as the distance metric. Minimum value of the criterion (1) corresponds to the optimal clustering;

- increasing the boundary values of variance and average of absolute values from minimum to maximum ones and Shannon entropy values from maximum to minimum one within the admissible ranges and removing genes for which the variance and average of absolute values are less and Shannon entropy is larger than appropriate boundary values. Calculation of the clustering quality criterion at each step of this procedure execution by the formula (1);
- result analysis. Fixation of the used criteria boundary values which correspond to minimum value of the clustering quality criterion;
- final removing the non-informative genes using determined boundary values of the statistical criteria and Shannon entropy.

Below, we present the algorithm for this step implementation.

Algorithm 1: Gene expression profiles reducing

Initialization:

Formation of the vectors of: variance (var), average of gene expression profiles absolute values (abs) and Shannon entropy ($entr$); set: ranges and steps of these parameters change; fix: iteration counter $t = 1$, $m = 1$, $var_1 = var_{min}$, $abs_1 = abs_{min}$, $entr_1 = entr_{max}$; Create: the empty subsets A and B for allocation of informative and non-informative gene expression profiles; empty vector of clustering quality criterion QC ;

while $t \leq length(var)$ do

fix the bounrary values of the statistical criteria and Shannon entropy: $var_b = var_t$, $abs_b = abs_t$, $entr_b = entr_t$;

while $m \leq ncol(dataset)$ do

calculation of var_m , abs_m , $entr_m$;

if $var_m \leq var_b$ and $abs_m \leq abs_b$ and $entr_m \geq entr_b$ then

| distribution of the gene expression profile into the subset A ;

else

| distribution of the gene expression profile into the subset B ;

end

$m = m + 1$;

end

formation of clusters of the examined samples based on subset A considering the data annotation;

calculation of the clustering quality criterion $QC[t]$ by the formulas (1)-(3);

$t = t + 1$;

end

Results analysis and final decision making:

making the chart: $QC = F(t)$;

fixation of t_{opt} which corresponds to the minimum value of the quality criterion QC ;

final division of the gene expression profiles into informative and non-informative subsets A and B considering the determined boundary values of both the statistical criteria and

Shannon entropy ($var_{t_{opt}}$, $abs_{t_{opt}}$, $entr_{t_{opt}}$);

Return the subsets A of the informative gene expression profiles.

2.3. Stepwise Hierarchical Gene Expression Profiles Clustering

As was noted hereinbefore, the main objective of this research is extraction of the most informative gene expression profiles in terms of their ability to identify the investigated samples considering both

the state of the patient's health or subtype of the disease. For this reason, the next stage of the hereinbefore presented procedure executing is stepwise gene expression profiles clustering at the hierarchical levels from 1 to N . We used Self-Organizing Tree clustering Algorithm (SOTA) [40] with correlation distance metric for this step implementation. This algorithm is a variety of self-organizing neural networks and it is based on the complex apply of Kohonen maps and Fritzke algorithm of spatial cell structure growing [41]. The simulation results have shown that SOTA clustering algorithm with correlation distance metric divides the set of high dimensional gene expression profiles into two clusters at one step of this procedure execution. Thus, the number of clusters is varied from 2 to 2^N at the first and the N -th hierarchical levels respectively. Then, we calculated the quality criterion values for each of the allocated clusters at each of the hierarchical levels using formulas (1)-(3). The vectors of genes expressions which correspond to the studied samples are used in this case as the investigated data. In other words, we evaluate in this case the proximity level of the samples, the attributes of which are the values of genes expressions which are grouped in the cluster. One cluster at each of the hierarchical levels was selected for the further research. These clusters correspond to the minimum value of the used quality criterion. The algorithm for this step implementation is presented below.

Algorithm 2: Stepwise hierarchical gene expression profiles clustering using SOTA clustering algorithm

Initialization:

setup of SOTA clustering algorithm parameters: $scell = 0.001$, $p_{cell} = scell \times 5$,

$w_{cell} = scell \times 2$, $distance = correlation$;

fix: iteration counter $t = 1$, maximal clustering hierarchical level $t_{max} = N$;

create the empty vector of the clustering quality criterion QC ;

while $t \leq N$ **do**

applying the SOTA clustering algorithm. Allocation of the clusters of gene expression profiles for the examined samples;

formation of subsets of the examined samples based on the obtained clusters;

calculation of the clustering quality criteria for the allocated clusters using the formulas (1)-(3);

fixation of the optimal cluster which corresponds to the minimum value of the clustering quality criterion;

$t = t + 1$;

end

Results analysis and final decision making:

making the chart: $QC = F(t)$;

selection of the clusters for the following processing. These clusters correspond to less values of the clustering quality criterion.

Return the list of the optimal clusters.

2.4. Binary Classification of the Investigated Samples

Four binary classifiers were used to evaluate the resolving ability of gene expression profiles in the selected clusters:

- Logistic Regression classifier (GLM) [42];
- Support-Vector Machine classifier (SVM) [43];
- Decision Tree classifier (CART) [44];
- Random Forest classifier (RF) [45].

The quality criteria based on the errors of both the first and the second kinds were used to evaluate the appropriate classifier effectiveness within the framework of the research. We used the gene expression data of patients, which were investigated on lung cancer disease. The data contained two type of samples: for healthy patients and patients with tumor. In this case, the classifier output can take two

states: 0 - healthy; 1 - tumor. The obtained results in this case can be represented using confusion matrix as follows (Table 1):

Table 1. Confusion matrix for lung cancer disease diagnostic

Real state of test-objects	Testing result	
	Tumor predicted	Norm predicted (healthy)
Tumor (1)	True positives (TP)	False negatives (FN)
Healthy (0)	False positives (FP)	True negatives (TN)

The following criteria were used to evaluate the classifiers effectiveness:

- *Accuracy (AC)* determines the total probability that classifier predicts true results:

$$AC = \frac{TP + TN}{TP + FP + TN + FN} \quad (4)$$

- *F-measure (F)* is defined as a harmonic mean of *Precision (PR* - positive predicted values) and *Recall (RC or Sensitivity)* [46]:

$$F = \frac{2 \cdot PR \cdot RC}{PR + RC} \quad (5)$$

where:

$$PR = \frac{TP}{TP + FP}; \quad RC = \frac{TP}{TP + FN}$$

- *Matthews correlation coefficient (MCC)* used in machine learning as a measure of the quality of binary classifiers [47]:

$$MCC = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}} \quad (6)$$

Larger value of each of the criteria corresponds to higher classifier effectiveness.

2.5. Fuzzy Inference System Implementation

Necessity of the use of fuzzy inference system [48] is determined by the possible contradiction of the different classifiers results for individual samples. To solve this problem, we propose to form the final solution using fuzzy inference system. Within the framework of our research the mathematical model can be presented as follows:

$$FS = f(x_{GLM}, x_{SVM}, x_{CART}, x_{RF}), \quad (7)$$

where *FS* is the output parameter of the fuzzy inference system characterized a final state of the investigated object (tumor or health); x_{GLM} , x_{SVM} , x_{CART} , x_{RF} are the input parameters or the results of *GLM*, *SVM*, *CART* and *RF* classifiers respectively. The values of both the input and output variables were varied within the range from 0 to 1. The fuzzy inference process executing assumes the following stages:

1. Setup of the system:

- transforming the values of both the input and output variables into linguistic estimates. Formation of the membership functions for each of the variables;
- formation of a basic term-set with appropriate membership function for each of the terms;
- formation of a set of fuzzy rules which are agreed between input and output variables.

2. Fuzzification procedure.

This step assumes evaluation of the membership functions values for each of the input variables crisp values for each of the terms.

3. *Fuzzy inference process*. This step involves the following:

- aggregation or determination of the conditions truth degree by clipping the levels for the prerequisites of each of the rules using the *min* operation;
- activation or determining the truth degree for each of the fuzzy rules;
- accumulation or forming the resulting membership function for output variable using *max* operation.

4. *Defuzzification* or determining the output variable crisp value.

3. Experiment

The publicly available gene expression data *GSE19188* of patients examined at early stage of lung cancer [49] was used as the experimental data within the framework of the research. This dataset was obtained as a result of DNA microchip experiments. 156 of DNA microchips were obtained during the experiment performing. The data annotation analysis has shown that the examined samples can be divided into two groups: 65 of the patients were healthy and 91 of the patients have lung cancer tumor. The *rma* method of data preprocessing (background correction, normalization, PM correction and summarization) was used to form the array of gene expression profiles. Initially, the data contained 54675 of genes (maximum number of genes at each of the microchips). Thus, the initial dataset was formed as a matrix in size (156×54675) .

At the first step, the non-informative genes in terms of variance, Shannon entropy and average of absolute values were reduced in accordance with technique described hereinbefore in the section 2.1. The simulation process assumed changing the boundary values of Shannon entropy from maximum to minimum value and appropriate statistical criteria values from minimum to maximum ones within the admissible ranges. Then, the gene expression profiles were identified as informative profiles for the following processing, if their average of absolute values and variance were larger and Shannon entropy was less than appropriate boundary values. Two clusters considering the state of the patients' health were formed with following computation of the quality criterion by formulas (1)–(3) at each stage of this procedure execution. Figure 2 presents the diagrams of both the quantity of genes in clusters and the clustering quality criterion values versus the step of the boundary parameters change.

An analysis of the obtained charts allows concluding that the quality criterion achieves its minimum value at 31-st step. 21431 of genes are identified in this case as informative ones. Thus, the initial matrix was transformed into matrix in size (156×21431) as a result of this step implementation. At the next stage, we performed the stepwise gene expression profiles clustering at hierarchical levels from 1 to 10 using SOTA clustering algorithm with following selection of the most informative groups of genes at each of the hierarchical levels in accordance with technique described in the section 3. At the final step, we have performed binary classification of the examined samples and have carried out the fuzzy inference procedure for final solution making.

4. Results and Discussion

Figure 3 displays the dot plot of the clustering quality criterion values of which were computed using the formulas (1)–(3) for the most informative clusters considering the minimum value of the quality criterion at each of the hierarchical levels. This chart also shows the number of genes in the selected clusters. The clusters quantity was changed from 2 to $2^{10} = 1024$ at the first and the tenth hierarchical clustering levels respectively. Six of clusters were selected for the following research as the result of the obtained chart analysis: the clusters which were allocated at hierarchical levels from 5 to 10. The cluster which was allocated at the fourth hierarchical level was not considered due to large quantity of genes.

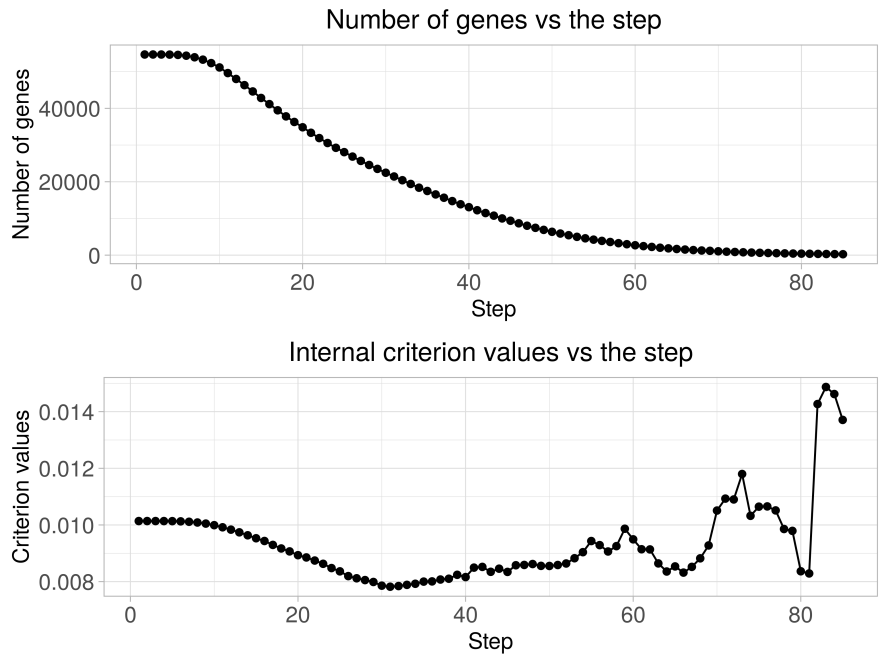


Figure 2. Diagrams of the number of genes in the clusters and the quality criterion values versus the step of the boundary parameters change

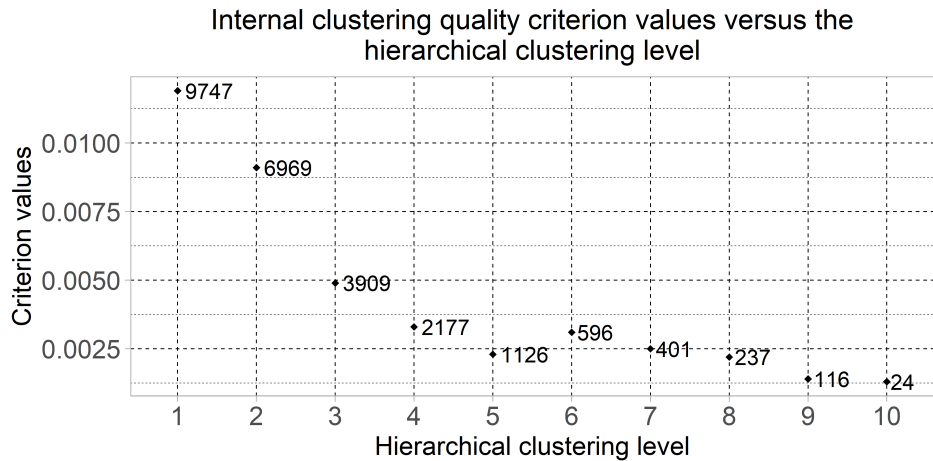


Figure 3. Dot plot of the clustering quality criterion calculated for the most informative clusters and the number of genes in this cluster versus the hierarchical clustering level

301 The simulation process concerning examined samples classification was performed using "Caret"
302 [50], "AER" [51], and "e1071" [52] packages of R software [53]. In the case of SVM classifier use, we
303 used the "linear" kernel (this choice was done empirically. Considering the high dimension of the
304 experimental data, the use of "radial" kernel gave significantly worse classification results). The
305 optimal parameters "gamma" and "cost" were determined in each of the cases empirically using
306 cross validation by the use of tune.svm() function of "e1071" package. The examined samples were
307 divided into two subsets considering the class to which belong the appropriate samples. 60% of
308 samples contained data for the model training and the remaining 40% were used for testing process
309 perform. In the case of logistic regression classifier (GLM) apply, we used glm() function with
310 family = binomial(link = "logit"). Decision tree and random forest classifiers were implemented
311 based on "caret" package using train() function. In both cases we used 10 estimators.

Tables 2 – 5 present the simulation results concerning application of *GLM*, *SVM*, *CART* and *RF* binary classifiers to classify the data in the selected clusters. The tables contain the results of the test datasets classification using previously trained classifiers.

Table 2. Results of logistic regression classifier operation (GLM)

Hierarchical level	Quality criteria				
	AC	PR	RC	F	MCC
5	0.543	0.474	0.450	0.462	0.066
6	0.565	0.526	0.476	0.500	0.118
7	0.565	0.632	0.480	0.545	0.148
8	0.522	0.579	0.440	0.500	0.060
9	0.587	0.579	0.500	0.537	0.169
10	0.804	0.895	0.708	0.791	0.629

Table 3. Results of support-vector machine classifier operation (SVM)

Hierarchical level	Quality criteria				
	AC	PR	RC	F	MCC
5	0.913	0.842	0.941	0.889	0.821
6	0.935	0.895	0.944	0.919	0.865
7	0.935	0.895	0.944	0.919	0.865
8	0.913	0.842	0.941	0.889	0.821
9	0.913	0.895	0.895	0.895	0.821
10	0.848	0.842	0.800	0.821	0.689

Table 4. Results of decision tree classifier operation (CART)

Hierarchical level	Quality criteria				
	AC	PR	RC	F	MCC
5	0.968	1.000	0.929	0.963	0.936
6	0.968	1.000	0.929	0.963	0.936
7	0.968	1.000	0.929	0.963	0.936
8	0.952	0.885	1.000	0.939	0.904
9	0.952	0.885	1.000	0.939	0.904
10	0.790	0.846	0.710	0.772	0.588

Table 5. Results of random forest classifier operation (RF)

Hierarchical level	Quality criteria				
	AC	PR	RC	F	MCC
5	0.952	0.885	1.000	0.939	0.904
6	0.968	0.923	1.000	0.960	0.935
7	0.968	0.923	1.000	0.960	0.935
8	0.952	0.885	1.000	0.939	0.904
9	0.968	0.923	1.000	0.960	0.935
10	0.903	0.923	0.857	0.889	0.805

The obtained results analysis allows concluding that classifier based on logistic regression model (GLM) is not effective to process high-dimensional vectors of gene expressions. The classification results are not satisfactory in all cases. A little better result in terms of the used criteria was obtained in the case of the use of cluster which was allocated at the tenth hierarchical level. This cluster contained only 24 of genes. However, the use of this classifier is not reasonable in the case of gene expression

320 data classification. Significantly better results were obtained in the cases of other binary classifiers
321 applying. It should be noted that all classifiers show worse classification results in the case of the use
322 of data in the smallest cluster (24 of genes). In other cases, the results of the classifications almost
323 agree under the use of *SVM*, *CART* and *RF* classifiers. Some better results were obtained in the case
324 of *CART* and *RF* classifiers use in comparison with the use of *SVM* classifier. Figures 4 - 9 show the
325 ROC curves of classification results for datasets allocated at hierarchical clustering levels from 5 (1126
326 of genes) to 10 (24 of genes) in the case of the use of all ML-based binary classifiers.

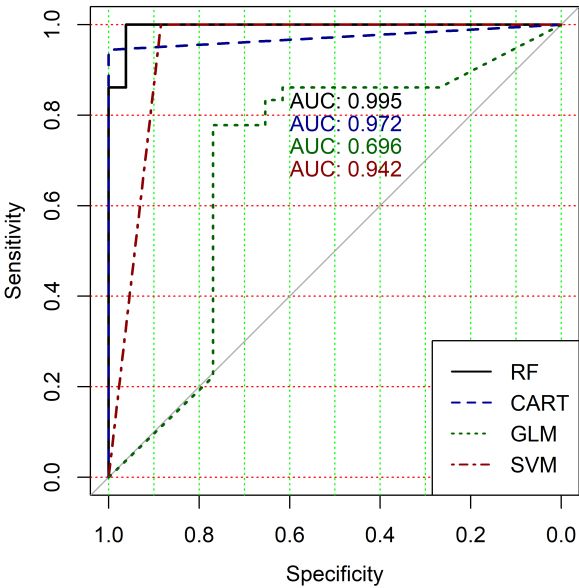


Figure 4. ROC curves for models of ML-based binary classifiers at the hierarchical clustering level 5

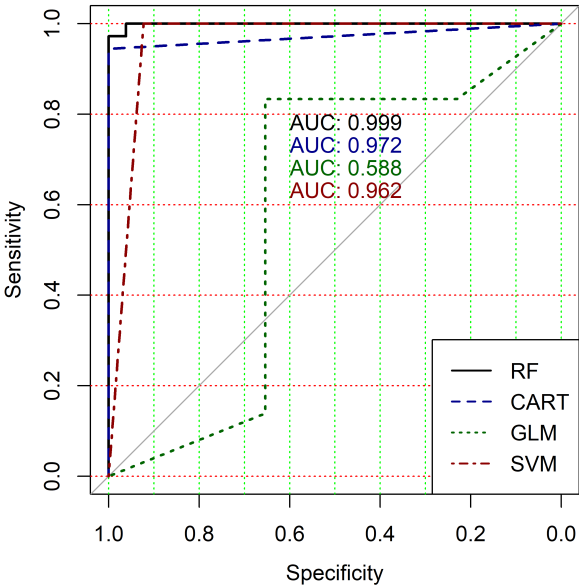


Figure 5. ROC curves for models of ML-based binary classifiers at the hierarchical clustering level 6

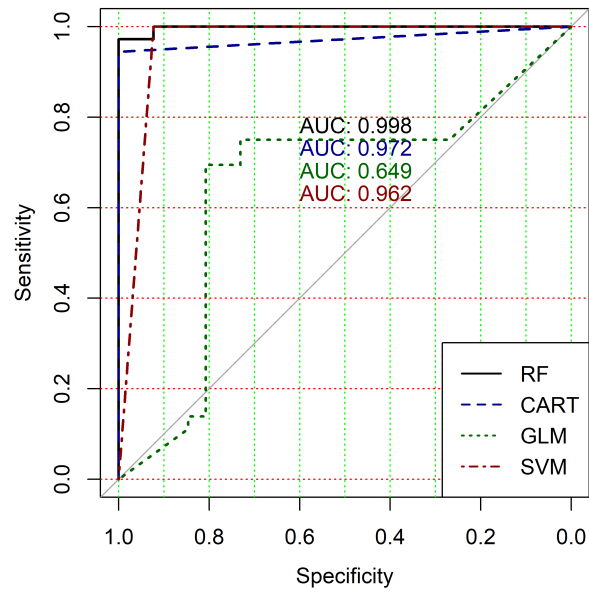


Figure 6. ROC curves for models of ML-based binary classifiers at the hierarchical clustering level 7

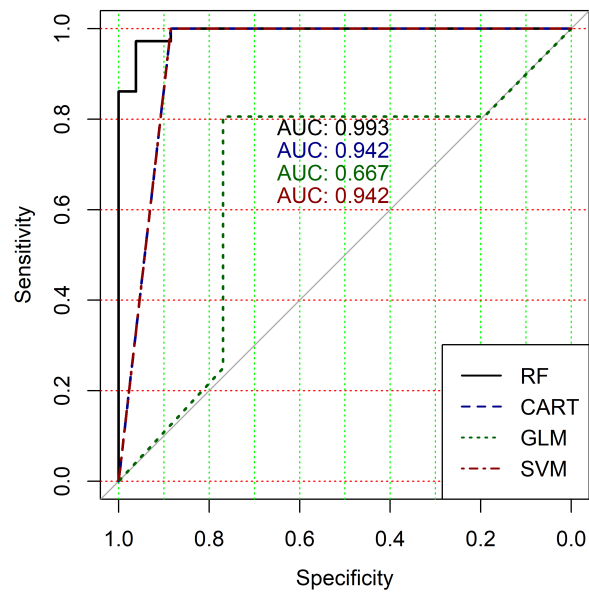


Figure 7. ROC curves for models of ML-based binary classifiers at the hierarchical clustering level 8

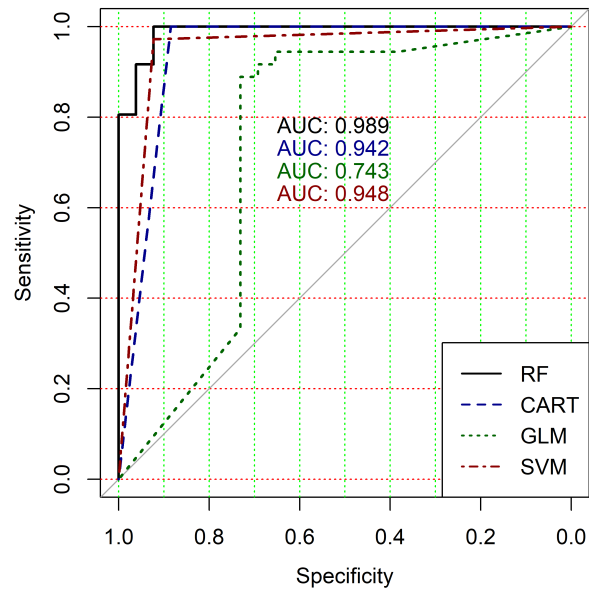


Figure 8. ROC curves for models of ML-based binary classifiers at the hierarchical clustering level 9

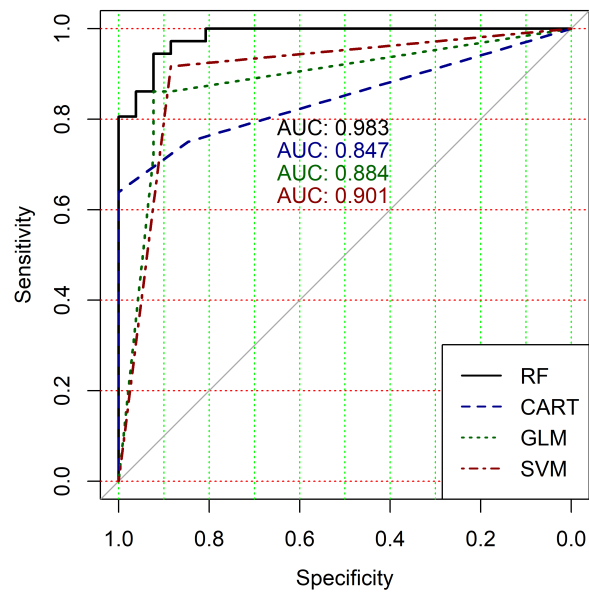


Figure 9. ROC curves for models of ML-based binary classifiers at the hierarchical clustering level 10

The analysis of the ROC curves confirms the conclusion concerning low effectiveness of GLM classifier (areas under the curves are 0.696, 0.588, 0.649, 0.667, 0.743 and 0.884 for clusters obtained at hierarchical levels from 5 to 10 respectively) and high effectiveness of RF, CART and SVM classifiers (areas under the roc-curves are significantly larger in comparison with areas obtained using GLM classifier). For this reason, we will use only the results of RF, CART and SVM classifiers as the input parameters of the fuzzy inference system at the next step of the simulation process.

We defined the terms "Healthy" (*HL*) and "Tumor" (*TM*) for input variables (result of appropriate classifier operation) and we used trapezoidal membership function for each of the terms. For output variable "Final State" (*FS*), we defined the terms: "Healthy" (*HL*); "Probably Healthy" (*PHL*); "Probably

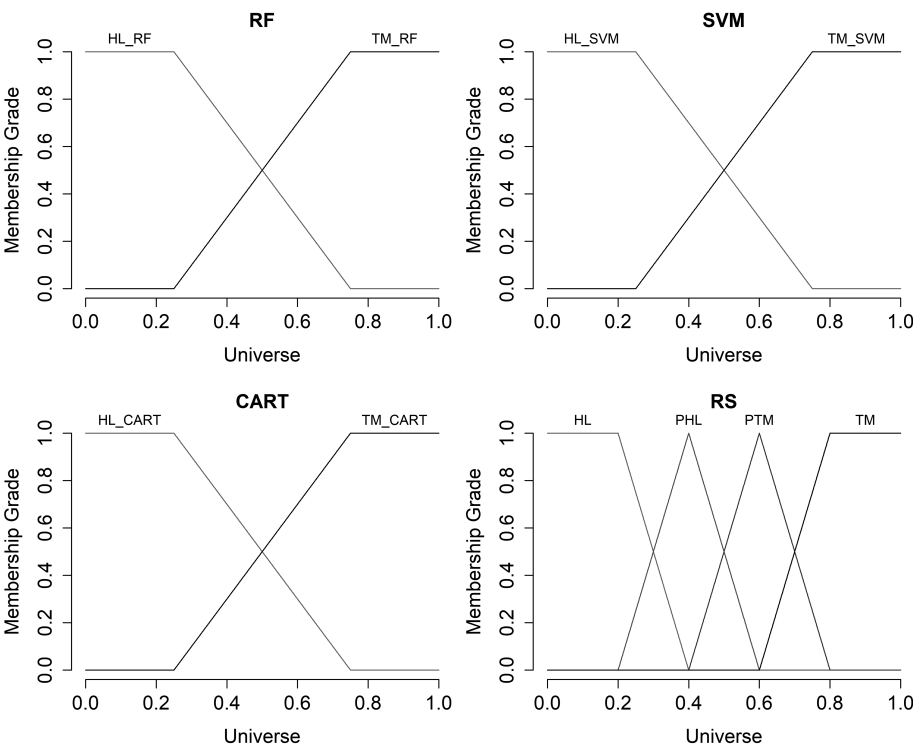


Figure 10. The charts of the membership functions for input and output variables

Tumor" (*PTM*); and "Tumor" (*TM*). We used also trapezoidal membership function for terms *HL* and *TM* and triangular membership function for terms *PHL* and *PTM* respectively.

Figure 10 shows the charts of the hereinbefore defined membership functions for input and output variables.

Table 6 presents the various combinations of the terms values which were used during the fuzzy rules formation. We applied Mamdani inference algorithm for fuzzy inference procedure performing and centroid method (mass center of the resulting membership function) for implementation of the defuzzification process.

Table 6. Terms values of the input and output variables

Number of fuzzy rules	Input and output variables			
	x_{SVM}	x_{CART}	x_{RF}	<i>FS</i>
rule 1	<i>T</i>	<i>T</i>	<i>T</i>	<i>T</i>
rule 2	<i>T</i>	<i>T</i>	<i>H</i>	<i>PT</i>
rule 3	<i>T</i>	<i>H</i>	<i>T</i>	<i>PT</i>
rule 4	<i>H</i>	<i>T</i>	<i>T</i>	<i>PT</i>
rule 5	<i>H</i>	<i>H</i>	<i>T</i>	<i>PH</i>
rule 6	<i>H</i>	<i>T</i>	<i>H</i>	<i>PH</i>
rule 7	<i>T</i>	<i>H</i>	<i>H</i>	<i>PH</i>
rule 8	<i>H</i>	<i>H</i>	<i>H</i>	<i>H</i>

Table 7 presents the results of fuzzy inference system operation.

Table 7. Results of fuzzy inference system operation

Hierarchical level	Quality criteria				
	AC	PR	RC	F	MCC
5	0.913	0.842	0.941	0.889	0.821
6	0.935	0.895	0.944	0.919	0.865
7	0.935	0.895	0.944	0.919	0.865
8	0.935	0.895	0.944	0.919	0.865
9	0.891	0.842	0.889	0.865	0.775
10	0.913	0.895	0.895	0.895	0.821

The obtained results analysis allows us to conclude that in the case of fuzzy inference system use, we get some worse results for clusters obtained at hierarchical level from 5 to 9 and significantly better result for cluster obtained at hierarchical level 10. Moreover, an analysis of the classification result for cluster at ninth hierarchical level shows disagree of various binary classifiers applied at previously step of our research in spite of very good classification results in the case of binary classifiers apply. This fact indicates that the use of this cluster is not reasonable for the following research. Moreover, the complex analysis of both Figure 4 and Tables 3 – 5 and Table 7 indicates the reasonability using for the further research the cluster obtained at hierarchical level 7. The cluster contains 401 of gene expression profiles, the values of the clustering quality criterion are not large too, and classification results in the terms of the used quality criteria are suitable in the case of the use of both separate binary classifiers and hybrid model based on fuzzy inference system.

5. Conclusions

In this paper, we have presented the results of the research concerning extraction of set of informative gene expression profiles in terms of their mutually correlation based on the complex use of both the clustering and classification techniques. The initial data has been presented as a matrix of gene expressions $(e_{ij}) \in R^{n \times m}$, where n and m are the number of samples and genes respectively. The publicly available gene expression data *GSE19188* of patients examined at early stage of lung cancer disease has been used as the experimental data. This data contained 156 of DNA microchips. The annotation of the data has shown that the examined samples can be divided into two groups: 65 of the samples for healthy patients and 91 of the samples belong to patients with lung cancer tumor. 54675 of genes (maximal quantity of genes at DNA microchips) contained the initial dataset.

At the first step, we have extracted the informative gene expression profiles by removing low-informative genes in terms of statistical criteria and Shannon entropy. In this case, we have used the clustering quality criterion as the main measure to evaluate the boundary values of the appropriate criteria. The initial matrix has been transformed into matrix in size (156×21431) as a result of this step implementation. At the next stage, we have performed the step-by-step gene expression profiles clustering at hierarchical levels from 1 to 10 with the use of SOTA clustering algorithm with following selection of the most informative clusters in terms of the used clustering quality criterion at each of the hierarchical levels. The number of clusters was changed from 2 to $2^{10} = 1024$ at the first and at the tenth hierarchical clustering levels respectively. Six of clusters have been selected for the following research as the result of this step implementation: the clusters that were allocated at hierarchical levels from 5 to 10.

Then, we have carried out the classification of the examined samples using four well known binary classifiers: Logistic regression classifier (*GLM*); Support-vector machine classifier (*SVM*); Decision tree classifier (*CART*); Random forest classifier (*RF*). The quality criteria based on errors of both the first and the second kinds have been used to evaluate the appropriate classifier effectiveness. The analysis of the obtained results has shown that classifier based on logistic regression model is not effective to process the high-dimensional vectors of gene expressions. Significantly better results have been obtained in the cases of other binary classifiers applying. However, it should be noted, that all

classifiers have shown worse classification results in the case of the use of data in the smallest cluster (24 of genes). In other cases, the results of the classifications almost agree under the use of *SVM*, *CART* and *RF* classifiers. Some better results have been obtained in the case of *CART* and *RF* classifiers use in comparison with the use of *SVM* classifier. The simulation results have shown also that some of the examined samples were identified differently and applying the fuzzy classifier to increase the objectivity of the gene expression profiles extraction at the final step is reasonable.

The analysis of the results of fuzzy inference system operation allows concluding that we have some worse results for clusters obtained at hierarchical level from 5 to 9 and significantly better result for cluster obtained at 10-th hierarchical level. Moreover, an analysis of the classification result for cluster at ninth hierarchical level has shown disagree of various binary classifiers in spite of very good classification results in the case of the use of individual classifiers. This fact indicates that the use of this cluster is not reasonable for the following research. The analysis of the obtained results has also shown the reasonability using for the further research the cluster obtained at the hierarchical level 7. This cluster contains 401 of genes, the value of the clustering quality criterion is not large, and the classification results in terms of the used quality criteria are suitable in terms of both separate binary classifiers and hybrid model based on fuzzy inference system.

To our mind, the conducted research can allow us to increase the objectivity for extraction of genes which can be used for reconstruction of gene regulatory networks and simulation of the reconstructed models considering the subtype of disease and/or state of the patient's health. In further, we are going to use the obtained results for both the gene regulatory networks reconstruction based on allocated genes and simulation of the reconstructed models in order to better understanding the gene interconnection in the cases of various state of the patient's health. This is the perspective of our research.

Author Contributions: The individual contributions of the authors are the following: Conceptualization, formal analysis, resources, writing–review and editing: Sergii Babichev and Jiří Škvor; methodology, software (R-programming), validation, statistical analysis and investigation, writing–original draft preparation, visualization: Sergii Babichev.

Funding: This research was funded by IGA UJEP, grant number UJEP-IGA-TC-2019-53-02-2.

Acknowledgments: We thank team of the researchers from Cell Biology, Erasmus University Medical Center, Rotterdam, The Netherlands Hou J, Aerts J, den Hamer B, et al. who have performed a genome-wide gene expression analysis on a cohort of 91 patients with tumor and 65 adjacent normal lung tissue samples. We would like also to acknowledge the support from IGA UJEP (grant number UJEP-IGA-TC-2019-53-02-2)

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DNA	Deoxyribonucleic Acid
RNA	Ribonucleic Acid
SOTA	Self-Organizing Tree Algorithm
ROC	Receiver Operating Characteristic
PM	Perfect-Match
GRN	Gene Regulatory Network
PCC	Pearson's Correlation Coefficient
GLM	Generalized Linear Model
SVM	Support-Vector Machine
CART	Classification And Regression Trees
RF	Random Forest

References

1. Lesage, R., Kerkhofs, J., Geris, L. Computational modeling and reverse engineering to reveal dominant regulatory interactions controlling osteochondral differentiation: Potential for regenerative medicine. *Frontiers in Bioengineering and Biotechnology* **2008**, *6*, art. no. 165. <https://doi.org/10.3389/fbioe.2018.00165>
2. Alexiou, A., Chatzichronis, S., Perveen, A., Hafeez, A., Ashraf, G.M. Algorithmic and stochastic representations of gene regulatory networks and protein-protein interactions. *Current Topics in Medicinal Chemistry* **2019**, *19*(6), 413–425. <https://doi.org/10.2174/1568026619666190311125256>
3. Liu, Z.P. Towards precise reconstruction of gene regulatory networks by data integration. *Quantitative Biology* **2018**, *6*(2), 113–128. <https://doi.org/10.1007/s40484-018-0139-4>
4. Byron, K., Wang, J.T.L. A comparative review of recent bioinformatics tools for inferring gene regulatory networks using time-series expression data. *International Journal of Data Mining and Bioinformatics* **2018**, *20*(4), 320–340. <https://doi.org/10.1504/IJDMB.2018.094889>
5. Schena, M., Davis, R.W. In *Microarray biochip technology*; Eaton Publishing, 2008; pp. 1–18.
6. Heather, J.M., Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics*, **2016**, *107*, 1–8.
7. Bolstad, B.M., Irizarry, R.A., Åstrand, M., Speed, T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **2003**, *19*(2), 185–193. <https://doi.org/10.1093/bioinformatics/19.2.185>
8. Affymetrix: Statistical algorithms description document. *Affymetrix*, **2002**. <http://tools.thermofisher.com/content/sfs/brochures>
9. Irizarry, R.A., Hobbs, B., Collin, F., et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Selected Works of Terry Speed* **2012**, pp. 601–616. https://doi.org/10.1007/978-1-4614-1347-9_15
10. Chen, Z., McGee, M., Liu, Q., Kong, M., Deng, Y., Scheuermann, R.H. A distribution-free convolution model for background correction of oligonucleotide microarray data. *BMC Genomics*, **2009**, *10*(1), art. no. 19. <https://doi.org/10.1186/1471-2164-10-S1-S19>
11. Gentleman, R., Carey, V., Huber, W., Irizarry, R., Dudoit, S. Bioinformatics and Computational Biology Solutions Using R and Bioconductor. *Springer* **2005**.
12. Park, T., Yi, S.G., Kang, S.H., Lee, S.Y., Lee, Y.S., Simon, R. Evaluation of normalization methods for microarray data. *BMC Bioinformatics*, **2003**, *4*, art. no. 13. <https://doi.org/10.1186/1471-2105-4-33>
13. Raddatz, B.B., Spitzbarth, I., Matheis, K.A., et al. Microarray-based gene expression analysis for veterinary pathologists: A review. *Veterinary Pathology*, **2017**, *54*(5), 734–755. <https://doi.org/10.1177/0300985817709887>
14. Åstrand, M. Contrast normalization of oligonucleotide arrays. *Journal of Computational Biology*, **2003**, *10*(1), 95–102. <https://doi.org/10.1089/106652703763255697>
15. Chen, Y.J., Kodell, R., Sistare, F., Thompson, K.L., Morris, S., Chen, J.J. Normalization methods for analysis of microarray gene-expression data. *Journal of Biopharmaceutical Statistics*, **2003**, *13*(1), 57–74. <https://doi.org/10.1081/BIP-120017726>
16. Barbara, D., Wu, X. An approximate median polish algorithm for large multidimensional data sets. *Springer-Verlag London Ltd. Knowledge and Information Systems*, **2003**, *5*, 416–438.
17. Lazaridis, E.N., Sinibaldi, D., Bloom, G., Mane, S., Jove, R. A simple method to improve probe set estimates from oligonucleotide arrays. *Mathematical Biosciences*, **2002**, *176*(1), 53–58. [https://doi.org/10.1016/S0025-5564\(01\)00100-6](https://doi.org/10.1016/S0025-5564(01)00100-6)
18. Babichev, S., Durnyak, B., Senkivskyy, V., Sorochynskiy, O., Kliap, M., Khamula, O. Exploratory analysis of neuroblastoma data genes expressions based on bioconductor package tools. *CEUR Workshop Proceedings*, **2019**, *2488*, 268–279.
19. Helgeson, E.S., Liu, Q., Chen, G., Kosorok, M.R., Bair, E. Biclustering via sparse clustering. *Biometrics*, **2020**, *76*(1), 348–358. <https://doi.org/10.1111/biom.13136>
20. Xie, J., Ma, A., Zhang, Y., et al. Qubic2: A novel and robust biclustering algorithm for analyses and interpretation of large-scale rna-seq data. *Bioinformatics*, **2020**, *36*(4), 1143–1149. <https://doi.org/10.1093/bioinformatics/btz692>

21. Karim, M.B., Kanaya, S., Altaf-Ul-Amin, M. Implementation of bicluso and its comparison with other biclustering algorithms. *Applied Network Science*, **2019**, 1(1), art. no. 79. <https://doi.org/10.1007/s41109-019-0180-x>
22. Babichev, S., Barilla, J., Fišer, J., Škvor, J. A hybrid model of gene expression profiles reducing based on the complex use of fuzzy inference system and clustering quality criteria. In: *2019 Conference of the International Fuzzy Systems Association and the European Society for Fuzzy Logic and Technology (EUSFLAT 2019)*. Atlantis Press **2019**. <https://doi.org/10.2991/eusflat-19.2019.20>
23. Patowary, P., Sarmah, R., Bhattacharyya, D.K. Developing an effective biclustering technique using an enhanced proximity measure. *Network Modeling Analysis in Health Informatics and Bioinformatics* **2020**, 9(1), art. no. 6. <https://doi.org/10.1007/s13721-019-0211-7>
24. Saini, N., Saha, S., Soni, C., Bhattacharyya, P. Automatic evolution of bi-clusters from microarray data using self-organized multi-objective evolutionary algorithm. *Applied Intelligence* **2020**, 50(4), 1027–1044. <https://doi.org/10.1007/s10489-019-01554-w>
25. Feng, C., Liu, S., Zhang, H., et al. Dimension reduction and clustering models for single-cell rna sequencing data: A comparative study. *International Journal of Molecular Sciences* **2020**, 21(6), art. no. 2181. <https://doi.org/10.3390/ijms21062181>
26. Babichev, S., Taif, M.A., Lytvynenko, V. Estimation of the inductive model of objects clustering stability based on the k-means algorithm for different levels of data noise. *Radio Electronics Computer Science Control* **2016**, 4, 54–60. <https://doi.org/10.15588/1607-3274-2016-4-7>
27. Shukla, A.K., Shukla, P., Vardhan, M. Gene selection for cancer types classification using novel hybrid metaheuristics approach. *Swarm and Evolutionary Computation* **2020**, 54, art. no. 100661. <https://doi.org/10.1016/j.swevo.2020.100661>
28. Yuan, L.M., Sun, Y., Huang, G. Using class-specific feature selection for cancer detection with gene expression profile data of platelets. *Sensors (Switzerland)* **2020**, 20(5), art. no. 1528. <https://doi.org/10.3390/s20051528>
29. Marussy, K., Buza, K. SUCCESS: A new approach for semi-supervised classification of time-series. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **2013**, 7894 LNAI (PART 1), 437–447. https://doi.org/10.1007/978-3-642-38658-9_39
30. Buza, K. Classification of gene expression data: A hubness-aware semi-supervised approach. *Computer Methods and Programs in Biomedicine* **2016**, 127, 105–113. <https://doi.org/10.1016/j.cmpb.2016.01.016>
31. Varkonyi, D.T., Buza, K. Extreme learning machines with regularization for the classification of gene expression data. *CEUR Workshop Proceedings* **2019**, 2473, 99–103.
32. Glowacz, A., Glowacz, Z. Recognition of images of finger skin with application of histogram, image filtration and K-NN classifier. *Biocybernetics and biomedical engineering* **2016**, 36(1), 95–101. <https://doi.org/10.1016/j.bbe.2015.12.005>
33. Tkachenko, R., Doroshenko, A., Izonin, I., Tsymbal, Y., Havrysh, B. Imbalance data classification via neural-like structures of geometric transformations model: Local and global approaches. *Advances in Intelligent Systems and Computing*, **2019**, 754, 112–122. https://doi.org/10.1007/978-3-319-91008-6_12
34. Izonin, I., Trostianchyn, A., Duriagina, Z., et al. The combined use of the wiener polynomial and SVM for material classification task in medical implants production. *International Journal of Intelligent Systems and Applications*, **2018**, 10(9), 40–47. <https://doi.org/10.5815/ijisa.2018.09.05>
35. Babichev, S., Lytvynenko, V., Skvor, J., Korobchynskiy, M., Voronenko, M. Information Technology of Gene Expression Profiles Processing for Purpose of Gene Regulatory Networks Reconstruction. In *Proceedings of the 2018 IEEE 2nd International Conference on Data Stream Mining and Processing, DSMP 2018*, **2018**, 336–341. <https://doi.org/10.1109/DSMP.2018.8478452>
36. Hausser, J., Strimmer, K. Entropy inference and the james-stein estimator, with application to nonlinear gene association networks. *Journal of Machine Learning Research* **2009**, 10, 1469–1484.
37. Calinski, T., Harabasz, J. A dendrite method for cluster analysis. *Communication in statistics* **1974**, 3, 1–27.
38. Zhao, Q., Xu, M., Fränti, P. Sum-of-squares based cluster validity index and significance analysis. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **2009**, 5495, 313–322. https://doi.org/10.1007/978-3-642-04921-7_32
39. Babichev, S., Lytvynenko, V., Skvor, J., Fiser, J. Model of the objective clustering inductive technology of gene expression profiles based on sota and dbscan clustering algorithms. *Advances in Intelligent Systems and Computing* **2018**, 689, 21–39. https://doi.org/10.1007/978-3-319-70581-1_2

40. Dorazo, J., Carazo, J.M. Phylogenetic reconstruction using an unsupervised growing neural network that adopts the topology of a phylogenetic tree. *Journal of Molecular Evolution* **1997**, *44*(2), 226–260. <https://doi.org/10.1007/PL00006139>
41. Fritzke, B. Growing cell structures a self-organizing network for unsupervised and supervised learning. *Neural Networks*, **1994**, *7*(9), 1441–1461. [https://doi.org/10.1016/0893-6080\(94\)90091-4](https://doi.org/10.1016/0893-6080(94)90091-4)
42. Tolles, J., Meurer, W.J. Logistic regression: Relating patient characteristics to outcomes. *JAMA - Journal of the American Medical Association* **2016**, *316*(5), 533–534. <https://doi.org/10.1001/jama.2016.7653>
43. Pedregosa, F., Varoquaux, G., Gramfort, A., et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.
44. Arunachalam, A.S., Thirumurthi Raja, A., Perumal, S. Enhanced constructive decision tree classification model for engineering students data. *International Journal of Recent Technology and Engineering* **2019**, *8*(1), 2414–2420.
45. Breiman, L. Random forests. *Machine Learning* **2001**, *45*(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
46. Sasaki, Y. The truth of the f-measure. In *Research Fellow* **2007**, pp. 1–5.
47. Matthews, B.W. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *BBA - Protein Structure* **1975**, *405*(2), 442–451. [https://doi.org/10.1016/0005-2795\(75\)90109-9](https://doi.org/10.1016/0005-2795(75)90109-9)
48. Zadeh, L.A., Abbasov, A.M., Shahbazova, S.N.: Fuzzy-based techniques in human-like processing of social network data. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* **2015**, *23*, 17–14.
49. Hou, J., Aerts, J., den Hamer, B., van Ijcken, W., et al. Gene expression-based classification of non-small cell lung carcinomas and survival prediction. *PLoS ONE*, **2010**, *5*(4), art. no. e10312.
50. Kuhn, M., Wing, J., Weston, S., et al. Classification and Regression Training. <https://github.com/topepo/caret/>.
51. Kleiber, C., Zeileis, A. Applied Econometrics with R. <https://cran.r-project.org/web/packages/AER/AER.pdf>.
52. Meyer, D., Dimitriadou, E., Hornik, K., et al. Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. <https://cran.r-project.org/web/packages/e1071/e1071.pdf>.
53. Ihaka, R., Gentleman, R. R: a language for data analysis and graphic. *Journal of Computational and Graphical Statistics* **1996**, *5*(3), 299–314.