Methods, Models and Information Technology of Complex Data Processing in the Fields of Technical Diagnostics and Bioinformatics

Sergii Babichev and Bohdan Durnyak

Ukrainian Academy of Printing
Lviv
2020

UDC: 004.048;004.94

Babichev S., Durnyak B. Methods, models and information technology of complex data processing in the fields of technical diagnostics and bioinformatics.

This monograph reflects the results of the authors' research concerning development and applying the data science techniques in the fields of technical diagnostic and bioinformatics. The noised signals filtering is very important part of data preprocessing techniques. The authors solve this task based on complex use of Huang transform and wavelet analysis techniques.

Another direction of authors' research is devoted to one of current directions of modern bioinformatics: development of techniques of gene expression profiles processing for purpose of gene regulatory networks reconstruction and validation of the reconstructed models. The monograph presents the authors' solutions concerning: gene expression array formation using Bioconductor package of R software; non-informative gene expression profiles reducing based on the complex use of fuzzy inference system and clustering quality criteria; stepwise cluster-bicluster analysis of gene expression profiles; reconstruction of gene regulatory networks using correlation and ARACNE inference algorithms based on Cytoscape software; validation of the reconstructed models using ROC analysis theory.

The monograph can be interested for scientists specialized in the fields of both development and applying data science techniques in various fields of scientific research.

Reviewers:

1. Prof. *Mykhaylo Yatsymirskyy*, DSc. (Lodz University of Technology, Poland)
2. Doc. *Viktor Mashkov*, DSc. (Jan Evangelista Purkyně University in Ústí nad Labem, Czech Republic)
3. Prof. *Dmytro Peleshko*, DSc. (IT Step University, Ukraine)

ISBN:978-966-322-505-0

Contents

List of Figures

List of Tables

Preface

This monograph is devoted to problem of data processing in the fields of technical diagnostics and bioinformatics. The techniques, models and algorithms presented in this book is a result of many-years authors' research. However, we do not want to say that our solutions are absolute ones. To our mind, each of the tasks can be solved by various ways. In this monograph, we have presented our views concerning solutions of appropriate tasks. Moreover, our opinion concerning one or other solution can be changed during accumulation of knows, skills and experience in the field of data science techniques applying.

The main direction of our research is focused to gene expression profiles processing for purpose of both gene regulatory network reconstruction and validation of the reconstructed models. This problem is one of the main direction of current bioinformatics. The experimental foundation for our research are arrays of gene expressions obtained as a result of both DNA microarray experiments or RNA molecules sequencing technique. Gene expressions in this case is meant a level of gene activity. This value is proportional to number of genes which correspond to appropriate type of protein in the biological organism. Gene expressions profile is a vector of gene expressions determined for differed samples or for different conditions of the experiment performing. Reconstruction of gene regulatory networks and further simulation of the reconstructed models forms the basis for investigation and analysis of both the character of molecular systems elements interconnections and influences of these interconnections to functional possibilities of the investigated objects.

The complexity of gene networks reconstruction is determined by the follows: the experimental data which are used for the reconstruction process usually does not allows defining the network structure and pattern of genes interconnection in the network. Moreover, large quantity of genes complicates the interpretation of the network elements interconnections. In this case, it is necessary to conduct research concerning: experimental data pre-processing in order to determine the optimal ways of gene expression array formation; gene expression profiles reducing for purpose of informative genes allocation in terms of quantitative quality criteria; evaluation of both the network topology and the pattern of genes interconnection in network with

the use of experimental data obtained by the use of both DNA microchip experiment or RNA-molecules sequencing method. Qualitatively reconstructed gene regulatory network allows investigating the pattern of the biological organism development at the genetic level. It creates the conditions for both making new effective medicines and development of methods of early diagnostics and effective treatment of complex diseases. This fact indicates the actuality of the research in this subject area.

Structure of book

Chapter 1 is devoted to development of technique of 1-D signals filtering based on complex use of Huang transform and wavelet analysis. The acoustic emission signals have been used as experimental during the simulation process. The techniques of complex signals processing based on wavelet analysis are widely used in various fields of scientific research. The effectiveness of this technique implementation depends on the choice of the type of the used wavelet, level of the wavelet decomposition and the thresholding coefficient value to process the detail coefficients. It should be noted that effective techniques for these parameters objective determining are absent nowadays. Moreover, the direct implementation of this technique for signals processing increases the requirements to the wavelet filter parameters determination. In this case more effective can be techniques which are based on decomposition of the signal into components with the further allocation and wavelet-processing of the noised components. In this chapter we have solved this problem based on the complex use of Huang transform and wavelet filtering techniques.

The results of the research concerning development of the objective clustering inductive technology have been presented in **chapter 2**. The idea and main conception of this technology were formulated by prof. Ivakhnenko A.G. This chapter presents the research concerning practical implementation of the objective clustering conception. Implementation of this technology involves determination of optimal clustering based on the extremum value of the complex balance criterion which contains as the components both the internal and external clustering quality criteria. The clustering process is carried out on two equal power subsets concurrently. The equal power subsets contain the same quantity of pairwise similar objects. This approach allows us to decrease the reproducibility error, which is one of the main unsolved problems of existing clustering algorithms. In this chapter, we have implemented the technique of objective clustering based on both density-based DBSCAN clustering algorithm and self-organizing SOTA one. The conducted research has allowed us to propose the stepwise procedure of gene expression profiles clustering at the stage of gene expression profiles pre-processing.

In **chapter 3**, we present the results of the research concerning development of gene expression profiles biclustering technique in order to allocate mutually cor-

related genes and samples for the following reconstruction of gene networks and validation of the reconstructed models. In the beginning, we have compared different biclustering techniques using synthetic data contained non-intersectional equal biclusters. At this stage, we have determined the *ensemble* biclustering technique as the most effective in comparison with other biclustering techniques in terms of both the internal and the external biclustering quality criteria. Then, we have proposed the technique of data biclustering based on *ensemble* biclustering algorithm, implementation of which allows determining the optimal algorithm parameters in terms of internal biclustering quality criterion. Finally, we have proposed the technique of step-by-step cluster-bicluster analysis of gene expression profiles. This technique is used as one of the components within the framework of information technology of gene expression profiles processing.

In **chapter 4** we have presented and detail described the structure block chart of the information technology of gene expression profiles processing for purpose of gene regulatory networks reconstruction and validation of the reconstructed models. Then, we have presented the result of the research concerning solution of problem of gene expression array formation based on the use of both DNA microchip experiments or RNA molecules sequencing method. At the next step we have presented the technique of non-informative genes reducing based on complex use of fuzzy inference system and clustering quality criterion.

Chapter 5 is devoted to development of technique of gene regulatory network reconstruction and validation of the reconstructed models. We have proposed the technique of optimizing the gene network topology based on the complex use of the network topological parameters. The optimal network topology corresponded to the maximum value of general topological parameter, which contains the simple topological parameters as the components. The simulation process was performed using both correlation and ARACNE inference algorithms. This chapter contains also the results of the research concerning validation of the reconstructed models based on evaluation of type 1 and type 2 errors.

Acknowledgements

The authors are grateful to professors Sharko A., Lytvynenko V., and Kornelyuk A. for fruitful cooperation during formation of the book content. We would like also to thank the reviewers prof. Yatsemirskyy M, Mashkov V. and Peleshko D. for their remarks and comments, which contributed to the improvement of the book.

Chapter 1

Acoustic Emission Signals Filtering

1.1 Introduction

Acoustic emission (AE) technique is one of the current directions of structural state monitoring methods which are developed as an alternative of non-destructive testing methods. Implementation of this technique allows us to perform both the continuous or on-demand diagnostics and discovering defects using permanently installed sensors [62, 120, 31, 148]. The main advantages of the AE technique are high level of availability and low maintenance costs. Identification of a defect location is performed by evaluation of the time difference of AE signals arrival to the sensors which are allocated at the different places of the object [140, 132]. High level of noise component which appears at the stages of signal generation, propagation and detection is one of the main reasons which complicates the successful application of this technique. Thus, the filtering of initial AE signal in order to remove the noise component is the one of the necessary conditions of the AE signals processing technique successful implementation.

A lot of techniques for different types of signals filtering exist nowadays. So, in [83, 137] the authors presented the signal processing methods based on smoothing the signal by the use of both the extrapolation technique and minimizing the mean square error between the estimated random and the desired processes. The main disadvantage of these techniques is their low effectiveness in the case of processing of non-stationary and non-linear signals with local particularities. Implementation of these techniques in these cases promotes to the loss of the large amount of useful information. The current methods of non-stationary and non-linear signals processing are based on decomposition of the signal with allocation of its components and the following processing of these components in order to remove the noise. The

paper [130] presents the results of the research concerning the use of fast Fourier transform for evaluation of the anisotropic relaxation of composites and nonwovens. Implementation of the fast Fourier transforms technique for time-frequency analysis of pressure pulsation signal is presented in [43]. The frequency spectrum including frequency-domain structure and approximate frequency-scope was obtained during the simulation process. However, it should be noted, that fast Fourier transform technique is effective in the case of stationary signals processing. In the case of non-stationary and non-linear signal processing the effectiveness of this technique decreases.

An alternative and logical continuation of the fast Fourier transforms technique is wavelet analysis [147, 119]. Implementation of this technique involves wavelet-decomposition on levels from 1 to N with calculation of both the approximation coefficients on N-th level and the detail coefficients on levels from 1 to N. In the most cases the detail coefficients contain the noise component, thus these coefficients should be processed during the filtering process. Reconstruction of the signal is performed with the use of both the approximation coefficients and the processed detail coefficients. The effectiveness of this technique implementation depends on the choice of the type of the used wavelet, level of the wavelet decomposition and the thresholding coefficient value to process the detail coefficients. It should be noted that effective techniques for these parameters objective determining are absent nowadays. Moreover, the direct implementation of this technique for signals processing increases the requirements to the wavelet filter parameters determination. In this case more effective can be techniques which are based on decomposition of the signal into components with the further allocation and wavelet-processing of the noised components.

In [69, 71] the authors proposed the use of the empirical mode decomposition (EMD) method based on complex use of both the Huang transform and Hilbert spectrum for non-stationary and non-linear signals analysis and processing. The main concept of this method consists of decomposition of the initial signal into mutually independent intrinsic mode functions (IMFs) based on Huang transform. Then, the Hilbert spectrum is formed by applying the Hilbert transform to the obtained modes (IMFs). The analysis of the Hilbert spectrum for the allocated modes allows us to receive the detail information concerning particularities of the investigated signal. Nowadays, the Hilbert-Huang technique has been implemented in various fields of scientific research. So, the paper [94] presents the technique to decompose the multicomponent micro–Doppler signals based on the complex use of Hilbert-Huang transform and analytical mode decomposition (HHT-AMD). The approach concerning implementation of the Hilbert-Huang transform (HHT) for detection, diagnostic and prediction of the degradation in the ball bearing is proposed in [129]. The papers [133, 134, 135] present the results of the research concerning implementation of the

HHT for analysis of the vibration signals from different objects. In the paper [111] the authors present the results of the research concerning the use of HHT for both the analysing and processing ECG signal in order to diagnose the brain functionality abnormalities. The results of the research concerning implementation of the HHT for analysis of both the non-stationary financial time series and acoustic wave frequency spectrum characteristics of rock mass under blasting damage are presented in the papers [70, 143]. However, it should be noted that in spite of achievements in this subject area the problem of denoising non-stationary and non-linear signals has no effective solution nowadays. This problem can be solved based on the complex use of modern techniques of both the data mining and machine learning which are applied successfully in different areas of the scientific research nowadays. In [19, 16, 18] we propose the technique of synthetic and acoustic emission (AE) signals filtering based on the complex use of both the Huang empirical mode decomposition method and wavelet analysis. The optimal parameters of the wavelet filter for each of the intrinsic modes are determined on the basis of minimum value of the quantitative criterion which is calculated as the ratio of Shannon entropies for the filtered data and for the allocated noise component. The obtained results of the research we present in this chapter.

1.2 Empirical Mode Decomposition and Discrete Wavelet Transform

Huang transform technique involves that initial signal is a complex one and it can be decomposed into intrinsic mode functions (IMFs) [69]:

$$y(x) = \sum_{i=1}^{n} f_i(x) + r_n(x) \tag{1.1}$$

where: n is the number of the IMFs functions; $f_i(x)$ is the IMFs function on i-th level of the signal decomposition; $r_n(x)$ is the residual function, which represent the average trend of the investigated signal. Implementation of the Huang empirical mode decomposition technique (EMD) intendes the following conditions:

- the number of each of the IMFs functions extrema and the number of zero crossing should be equal or not differ by more than one;
- in any point of the IMFs function the mean value of the envelope defined by local maximums and local minimums should be zero.

The signal decomposition process is stopped if one of the following conditions is performed:

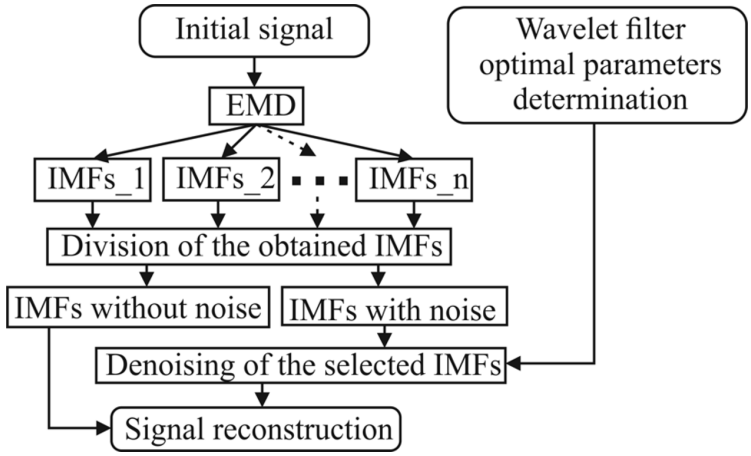

Figure 1.1: A structural block chart of step-by-step process of the signal filtering

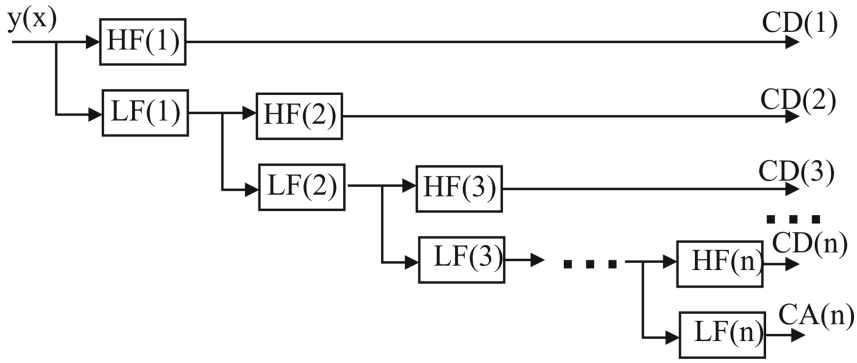

Figure 1.2: A structural block chart of the discrete wavelet decomposition process

- the residual function $r_n(x)$ does not contain more than 2–3 extrema points;
- the residual function $r_n(x)$ in whole interval of x change is insignificant in comparison with appropriate values of the IMFs functions.

A structural block-chart of the step-by-step procedure of the signal filtering based on the complex use of Huang empirical mode decomposition technique and wavelet analysis is presented in Figure 1.1. As it can be seen, the result of the Huang transform is selection of the IMFs functions which contain the noise component for purpose of their further filtering using discrete wavelet transform technique. Figure 1.2 presents the main idea of the discrete wavelet decomposition process. Implementation of this procedure involves calculation of both the approximation coefficients at N-th level and the detail coefficients at levels from 1 to N using the

low frequency (LF) and high frequency (HF) filters:

$$y(x) \rightarrow \{CA(N), CD(N), ..., CD(2), CD(1)\} \tag{1.2}$$

The noise component in the most cases is contained in detail coefficients therefore these coefficients should be processed during the signal processing. To process the detail coefficients we propose to use the soft thresholding in accordance with the following conditions:

$$\begin{cases} d = 0, & if\ d \leq \tau \\ d = d - \tau, & if\ d > \tau \end{cases} \tag{1.3}$$

where d is the detail coefficient and τ is the thresholding coefficient value. It is obvious, that quality of wavelet filtering process depends on type of the used wavelet, level of the wavelet decomposition and thresholding coefficient value to process the detail coefficients. In [19] we proposed the technique to determane the optimal parameters of the wavelet filter based on the use of the Shannon entropy criterion which is calculated on the basis of james-stein shrinkage estimator [67]. This method is based on the complex use of two different models: a high-dimensional model with low bias and high variance, and a lower dimensional model with larger bias but smaller variance. Evaluation of the values distribution probability in a cell in accordance with the James-Stein shrinkage method is calculated in the following way:

$$p_i^{Srink} = \lambda p_i + (1 - \lambda) p_i^{ML} \tag{1.4}$$

where p_i^{ML} is the probability of the values distribution in the i-th cell, which is calculated by the maximum likelihood method; $p_i = \dfrac{1}{n_i}$ is the maximum entropy target in the i-th cell; n_i is the number of the features in this cell. It is obvious, that p_i^{ML} corresponds to the high-dimensional model with low bias and high variance and p_i is the estimation with higher bias and lower variance of the features distribution. Intensity parameter λ in the proposed model is calculated as follows:

$$\lambda = \frac{1 - \sum_{i=1}^{k} (p_i^{ML})^2}{(n-1) \sum_{i=1}^{k} (p_i - p_i^{ML})^2} \tag{1.5}$$

where n is the number of the features in the vector. The value of Shannon entropy is calculated with the use of standard formula taking into account the method of the probability estimation:

$$H^{Shrink} = -\sum_{i=1}^{k} p_i^{Shrink} \log_2 p_i^{Shrink} \tag{1.6}$$

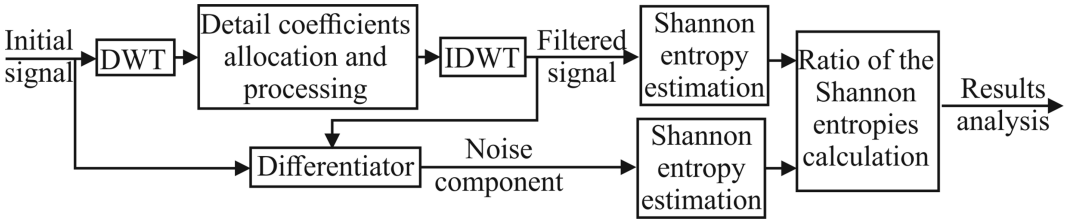

Figure 1.3: A structural block chart of the technique to calculate the Shannon entropies ratio

In [16, 18] the authors proposed the technique of the wavelet filter optimal parameters determination based on the use of the ratio of Shannon entropies which are calculated for both the filtered signal and the allocated noise component:

$$RH = \frac{H(filtered\,signal)}{H(noise\,component)} \tag{1.7}$$

The optimal parameters of the wavelet filter corresponds to the minimum value of the Shannon entropy for filtered signal and the maximum value of this criterion for the allocated noise component. In this case the value of the relative criterion (1.7) should be achieved the minimum one. The structural block chart of the procedure of this criterion calculation within the framework of the proposed technique is presented in Figure 1.3.

1.3 Hybrid model of AE signal filtering

Figure 1.4 shows the structural block chart of the algorithm to determine the wavelet filter optimal parameters. The stages of this algorithm implementation are the following:

Stage I. Signal loading and Huang transform performing.

1. Loading of the investigated signal. Application of Huang transform to the signal. Empirical mode decomposition of the signal.
2. Visualization and analysis of the obtained modes. Allocation of the noised modes for their following processing.

Stage II. Wavelet filtering of the selected modes.

3. Setup the ranges and the steps of the wavelet filter parameters change.

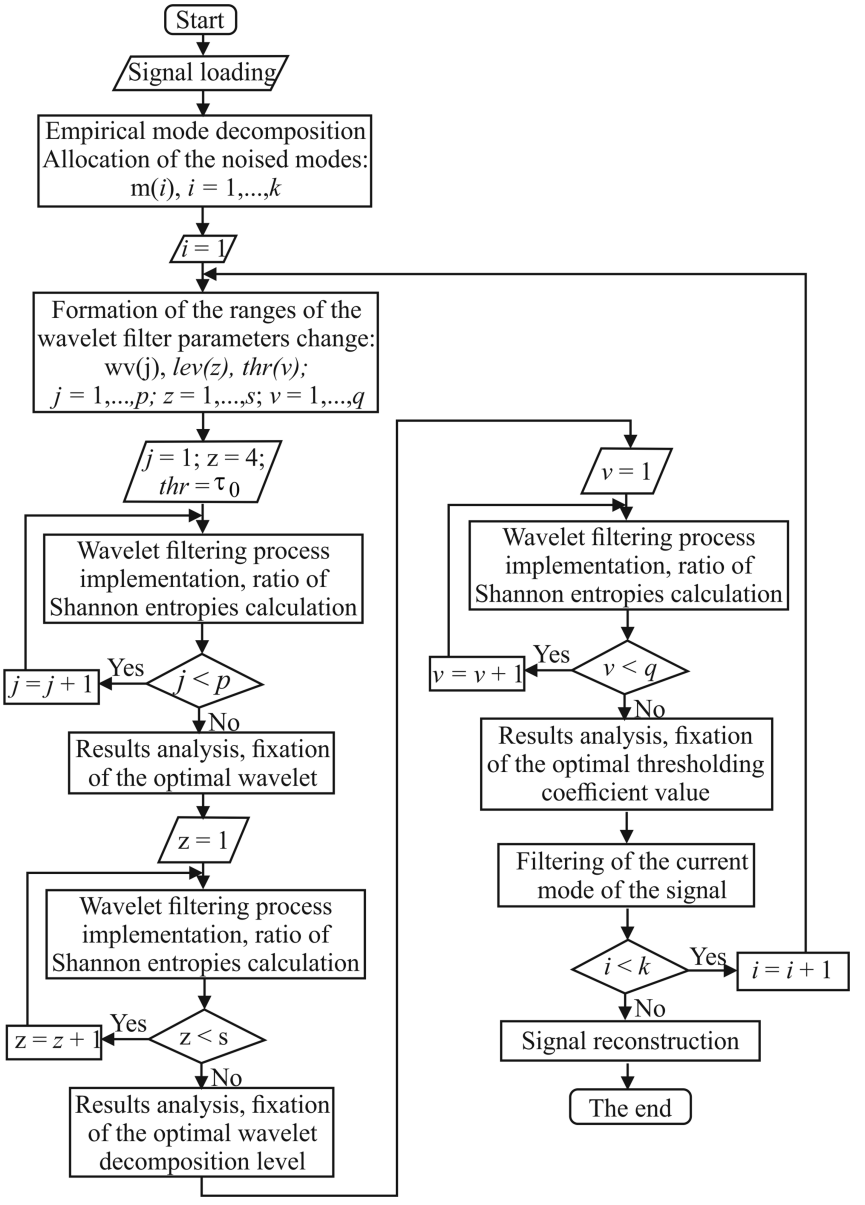

Figure 1.4: A structural block chart of the algorithm to determine the wavelet filter optimal parameters

3.1 Formation of the vector of different types of wavelets for the appropriate mother wavelet.

3.2 Calculation of the thresholding coefficients initial value to process the detail coefficients:

$$\tau_0 = \sigma\sqrt{2\ln k}$$

where k is the length of the investigated signal; σ is the median absolute deviation for the allocated detail coefficients:

$$\sigma = \delta \cdot (|CD(i) - median(CD(i))|)$$

where $i = 1, ..., n$ is the wavelet decomposition level, coefficient δ is determined empirically during the simulation process.

3.3 Formation of the range and the step of the thresholding parameter value change:

$$\tau_{min} = 0.1\tau;\ \tau_{max} = 5\tau;\ d\tau = 0.02 \cdot (\tau_{max} - \tau_{min})$$

3.4 Evaluation of the wavelet decomposition maximum level.

4. Determination of the optimal type of the wavelet.

4.1 Initialization of the counter, which corresponds to the first wavelet in the appropriate sequence ($j = 1$). Setup of the initial values of both the wavelet decomposition level ($N = 3$) and the thresholding coefficient value ($\tau = \tau_0$).

4.2 Discrete wavelet decomposition of the signal with calculation of both the approximation coefficients at the N-th decomposition level and the detail coefficients at the levels from 1 to N.

4.3 Soft thresholding of the detail coefficients using the conditions (1.3).

4.4 Reconstruction of the signal based on both the approximation coefficients and the processed detail coefficients.

5. Calculation of the data processing quality criteria.

5.1 Extraction of the noise component as the difference of both the initial and filtered signals.

5.2 Calculation of the Shannon entropies for the filtered signal and for the allocated noise component by the formula (1.6). Calculation of their ratio by the formula (1.7).

5.3 If the counter value is maximal one, the results analysis and fixation of the optimal type of wavelet which corresponds to the minimum value of the criterion (1.7). Otherwise, increment of the counter value and go to the step 4.2 of this procedure.

6. Determination of the optimal wavelet decomposition level.

   6.1 Initialization of the counter, which corresponds to the first level of wavelet decomposition ($z = 1$).

   6.2 Repetition of the steps from 4.2 to 5.3 of this procedure.

   6.3 If the counter value is maximal one, the results analysis and fixation of the optimal wavelet decomposition level. Otherwise, increment of the counter value and go to the step 6.2 of this procedure.

7. Determination of the thresholding coefficient optimal value.

   7.1 Initialization of the counter (v = 1), which corresponds to the minimum value of the thresholding coefficient: ($\tau = \tau_{min}$).

   7.2 Repetition of the steps from 4.2 to 5.3 of this procedure.

   7.3 If the counter value is maximal one, the results analysis and fixation of the thresholding coefficient optimal value. Otherwise, increment of the counter value and go to the step 7.2 of this procedure.

8. Filtering of the current IMFs function with the use of the wavelet filter optimal parameters.

9. Repetition of the stage 2 for other of the allocated IMFs functions.

Stage III. Reconstruction of the signal.

10. Reconstruction of the signal with the use of both the processed and non-processed components of the signal.

1.4 Experiment

Two type of signals were used during the simulation process. Figure 1.5 presents the synthetic signals which were used to estimate the proposed technique effectiveness. The first test signal contained 8 seconds of data from the 2005 TOMODEC ocean

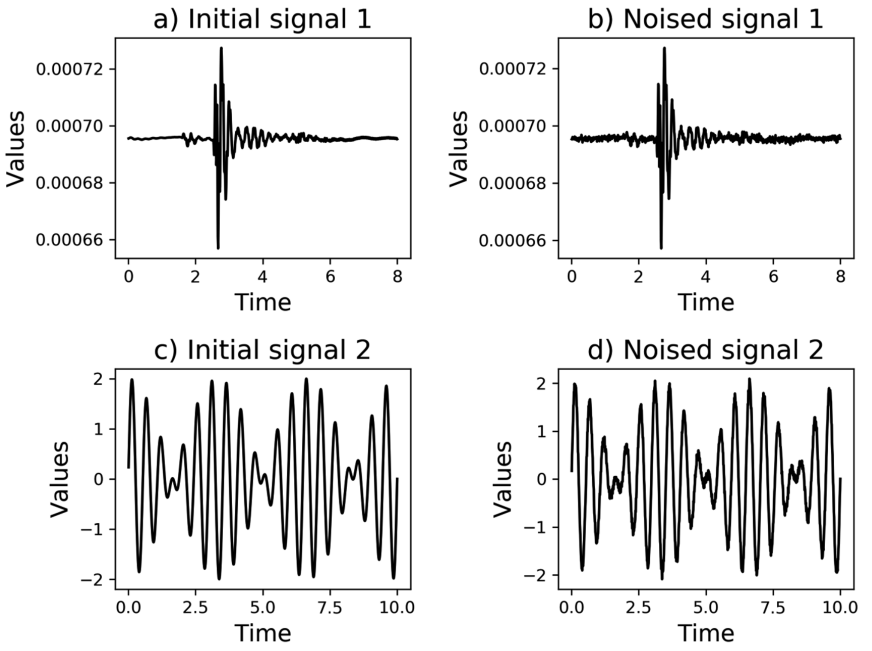

Figure 1.5: Sinthetic signals: a,c) signals without noise; b,d) signals with noise

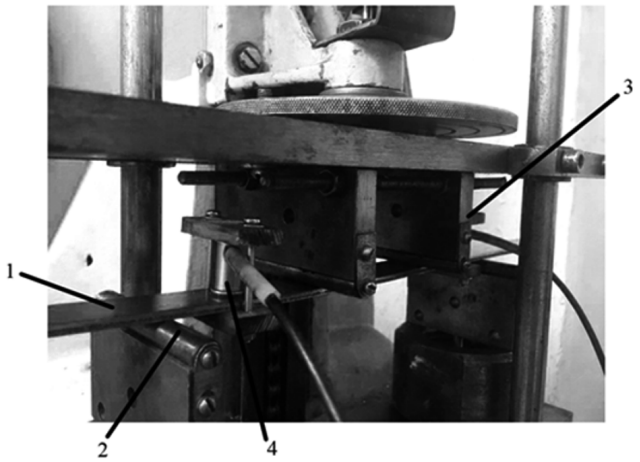

Figure 1.6: Four-point bend test device: 1) test sample; 2) support; 3) deformation indicator; 4)AE signal indicator

bottom seismometer network at Deception Island, South Shetland Islands, Antarctica [35]. The second test signal is the combination of two sinusoids with different frequencies ω_1 and ω_2:

$$y_2(t) = sin(\omega_1 \pi t) + sin(\omega_2 \pi t)$$

The noise component was generated as the vector of random numeric values, the range of their changes corresponded to the condition:

$$range(noise) = 0.02 \cdot (max(signal) - min(signal))$$

The experimental device which were used to generating the acoustic emission signals for the different levels of mechanical loading of the tested material is shown in Figure 1.6. The experimental device contains three main mechanisms: the deformation, the force-fixation and the AE signal fixation mechanisms [3]. The samples for four-point bend test were cut out from steel flat in the size $300 \times 20 \times 4$ mm. The simulation process involved the fixation of both the AE signals and the level of the sample deformation for different levels of the tested sample loading. The broadband sensors for acoustic emission instrument AF15 with a bandwidths of 0.2-2.0 was used as the measurement device. The artificial load was increased step-by-step from 150 N to 400 N with fixation of the AE signals for different values of the sample deformation. The examples of the AE signals which were obtained during the simulation process are shown in Figure 1.7. As it can be seen, the shape of the signals is changed during the load increase. However, the existence the noise component complicates the obtained results interpretation. Thus, at the first step it is necessary to

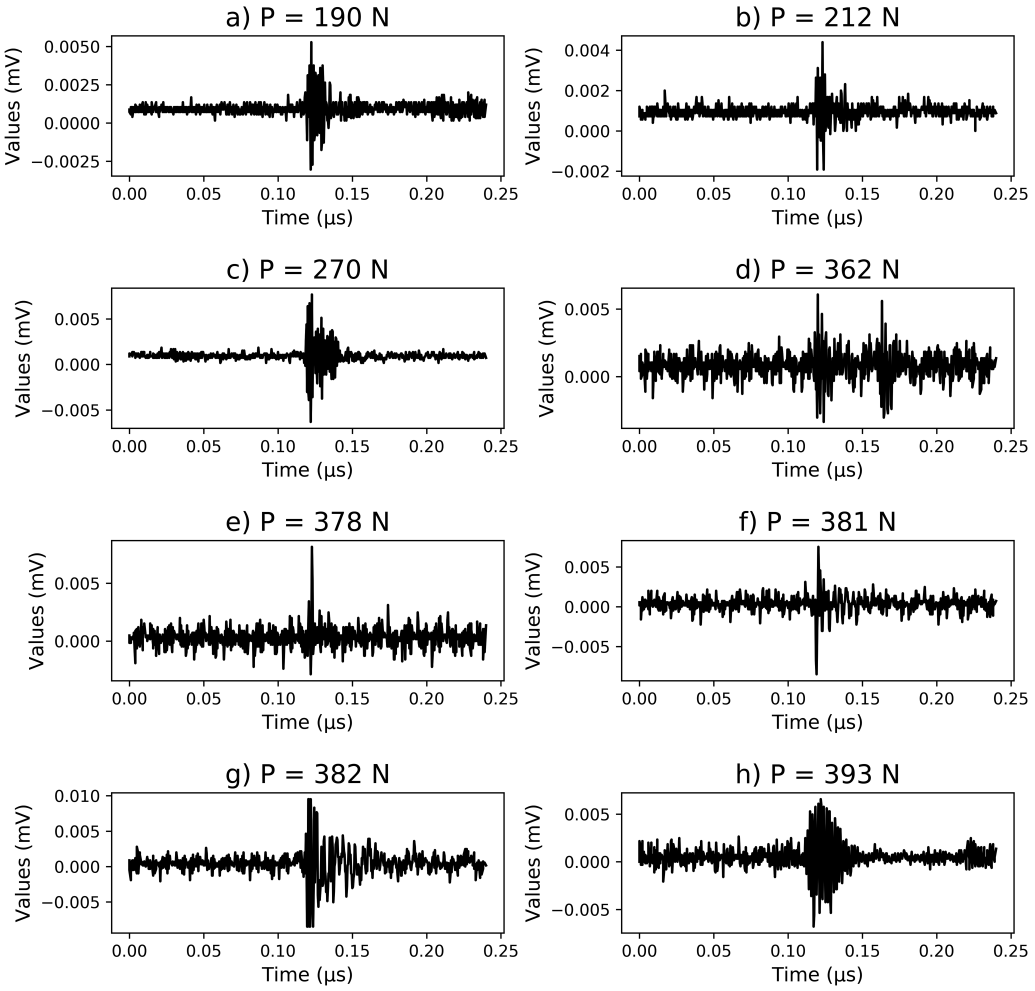

Figure 1.7: AE signals for different levels of the sample deformation

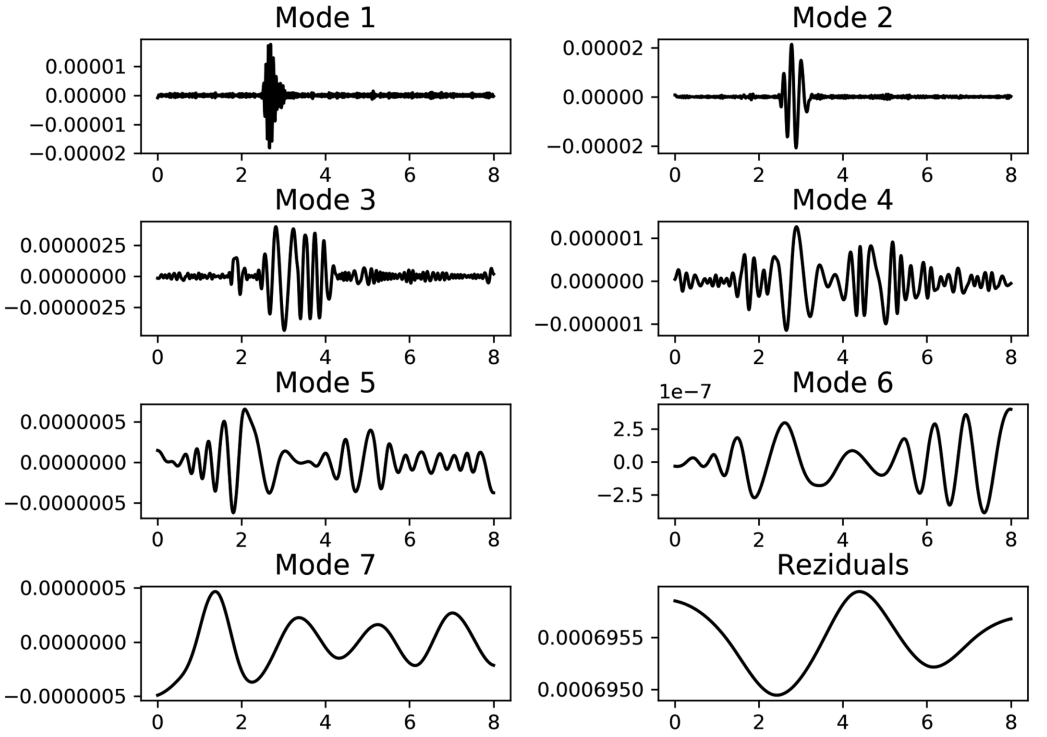

Figure 1.8: Results of the empirical mode decomposition for the synthetic signal 1

decrease the level of the noise component with saving useful information concerning state of the investigated sample.

1.5 Results and Discussion

1.5.1 Results of the Sinthetic Signals Filtering

Figure 1.8 and Figure 1.9 present the results of the Huang transform implementation for the synthetic signals. The first stage of the hereinbefore presented algorithm was used in this case. The analysis of the obtained results allows us to conclude that in the both cases two IMFs functions (Mode 1 and Mode 2) contain the high frequency noise component. Thus, these modes should be processed at the second step of the algorithm implementation.

Figure 1.10 presents the results of the simulation concerning determination of the optimal type of the wavelet for each of the allocated IMFs functions. The biorthogonal wavelet *bior* was used as the mother wavelet in this case. This choice is

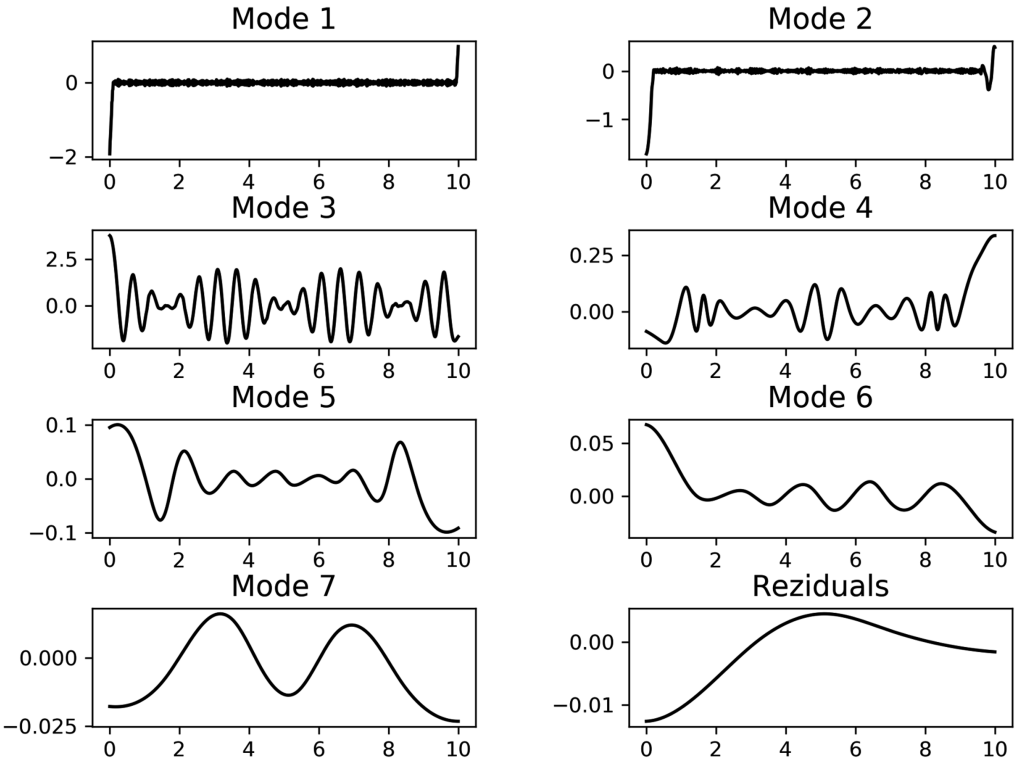

Figure 1.9: Results of the empirical mode decomposition for the synthetic signal 2

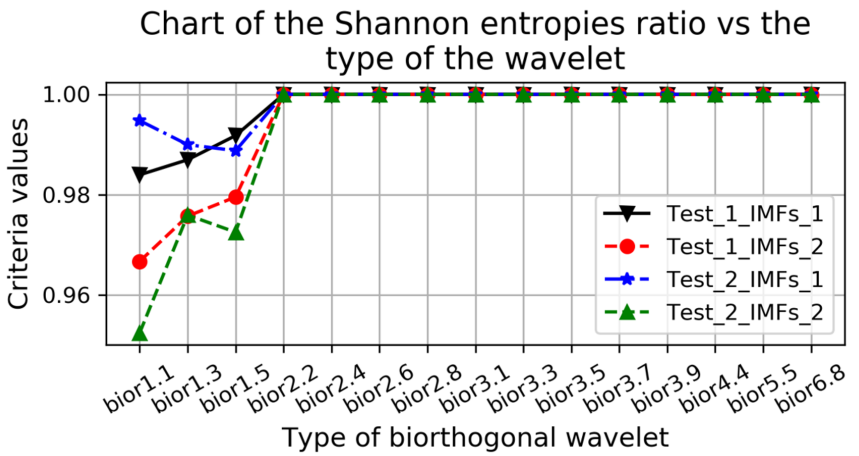

Figure 1.10: Results of the simulation concerning determination of the optimal wavelet

determined by the result of the previous research involving the comparison analysis of the orthogonal and biorthogonal wavelets for complex signals filtering [7]. The authors have shown that the choice of the type of the mother wavelet from orthogonal and biorthogonal wavelets in the case of the gene expression profiles filtering is not determinative. The quality of the signal filtering is determined mainly by the following parameters: type of the wavelet from the family of the mother's wavelets; level of the wavelet decomposition; value of the thresholding parameter to process the detail coefficients. Moreover, in this work was shown too that the use of the biorthogonal wavelet family allows obtaining better results in terms of the used criteria. In this reason we use the family of biorthogonal wavelets for the following signals processing.

The analysis of the obtained results allows us to conclude that biorthogonal wavelet *bior1.1* is optimal to process IMFs_1 and IMFs_2 functions in the case of the first synthetic signal use. In the case of the second synthetic signal processing the wavelet *bior1.5* is optimal for IMFs_1 function and the wavelet *bior1.1* is optimal for IMFs_2 one. The values of the Shannon entropies ratio in these cases are the minimal ones.

The results of the proposed technique implementation for purpose of the wavelet decomposition optimal level determination are presented in Figure 1.11. The level of the wavelet decomposition was changed within the range from 3 to the maximum level depending on the type of the studied signal. The minimal boundary value was determined empirically. The results of the simulation have shown that in the case of the first synthetic signal use, the optimal level of the wavelet decomposition is 8

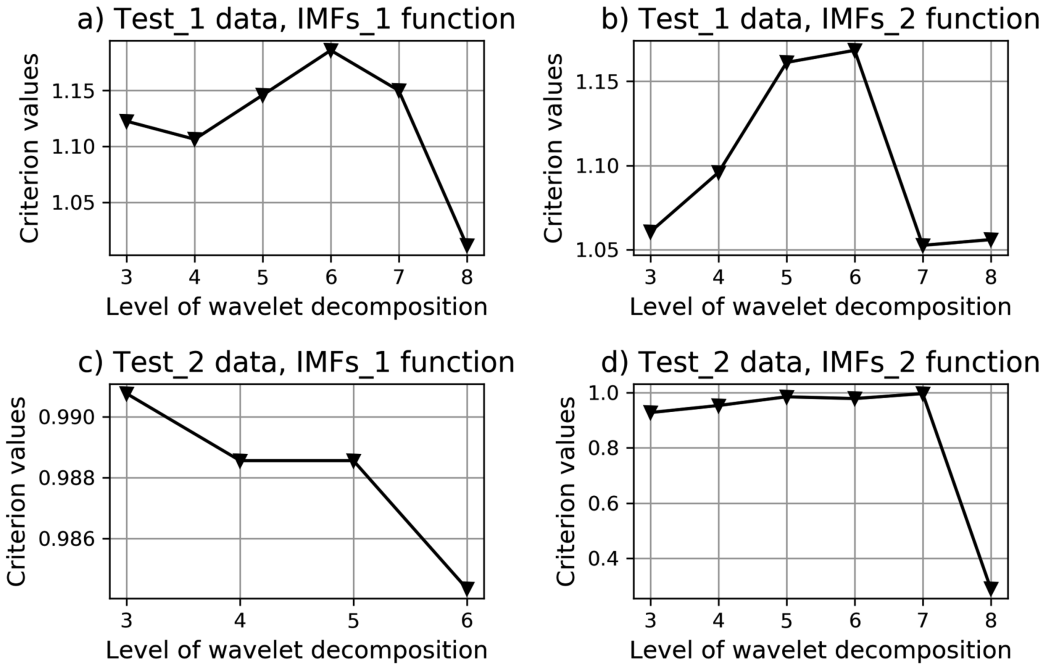

Figure 1.11: Charts of the Shannon entropies ratio vs the wavelet decomposition level

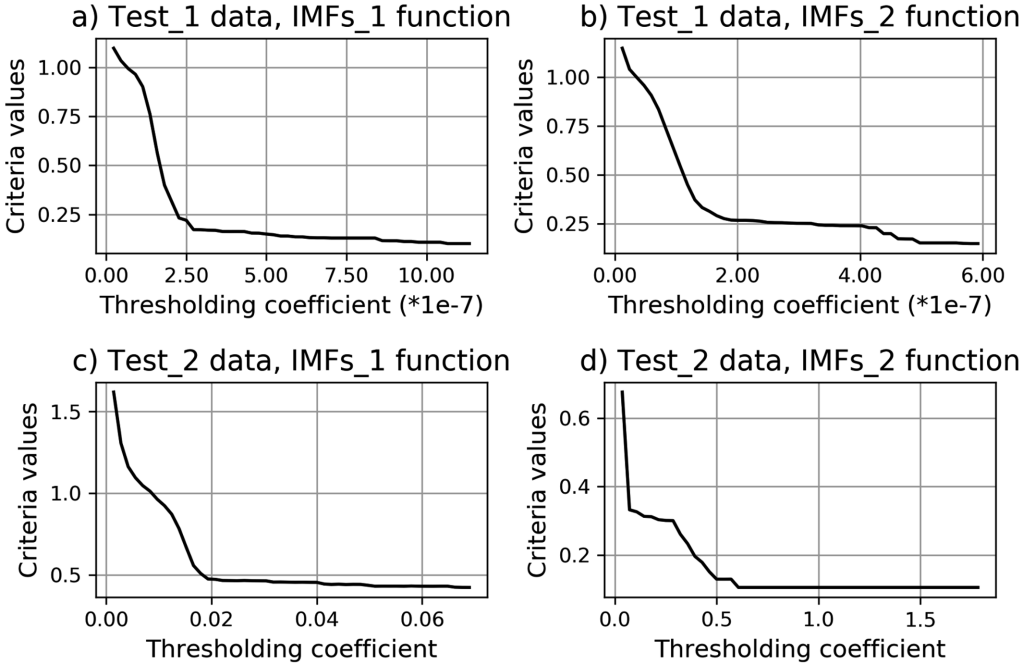

Figure 1.12: Charts of the Shannon entropies ratio vs the thresholding coefficient value

for the mode 1 and 7 for the mode 2. In the case of the second synthetic signal use the optimal levels are 6 and 8 for the mode 1 and the mode 2 respectively. These parameters were used during the following data processing.

Figure 1.12 presents the results of the simulation concerning determination of the thresholding coefficients optimal values to process the detail coefficients at the levels of wavelet decomposition from 1 to n in accordance with the formula (1.3). The range and the step of the thresholding coefficient change were determined in accordance with the step 9 of the hereinbefore described algorithm. As the results, the optimal values of the thresholding coefficients were determined for each of the allocated IMFs functions.

The final stage of the data processing is the signal reconstruction with the use of both the processed and non-processed IMFs functions. The results of the wavelet filter optimal parameters determination and relative change of the Shannon entropy for both the filtered and initial signals in percentages calculated by the formula (1.7) are presented in Table 1.1.

Figure 1.13 presents the results of the synthetic signals filtering. The analysis of the obtained results allows us to conclude that the relative change of the Shannon

Table 1.1: Parameters of the wavelet filter in the cases of the synthetic signals processing

Signal	Function	Parameters			Error
		Wavelet	Level	THR	
1	IMFs_1	*bior1.1*	8	1.07e-6	0.92%
	IMFs_2	*bior1.1*	8	5.8e-7	
2	IMFs_1	*bior1.5*	6	0.069	0.42%
	IMFs_2	*bior1.1*	6	0.607	

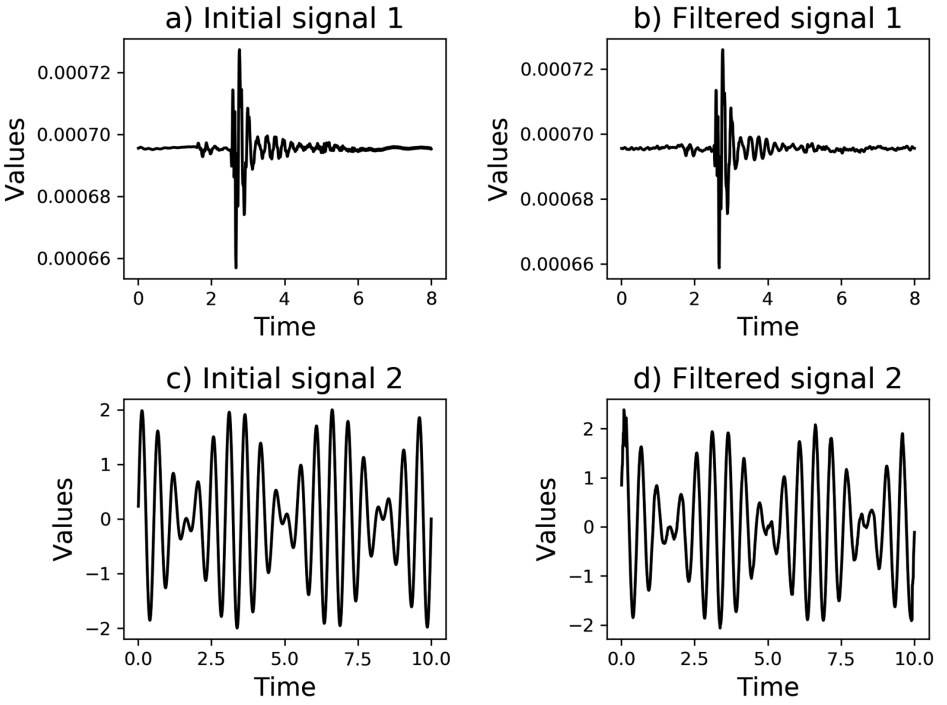

Figure 1.13: Results of the synthetic signals filtering

entropy criterion calculated for the initial and the filtered signals are less than one percent for the both synthetic signals. This fact indicates the high effectiveness of the proposed technique. Moreover, the results of the simulation have also shown that the proposed technique is not so greatly sensitive to the thresholding coefficient value as in the case of the direct use of the wavelet analysis for the signal filtering. This fact can be explained in the following way. The use of the Huang transform allows us to select the noised components of the signal, which are processed at the next stage of the algorithm implementation. The components without noise are not processed. Thus, a little change of the thresholding coefficient value does not significantly influence the results of the signal as in the case of the direct use of the wavelet analysis for the signal denoising.

1.5.2 Results of the AE Signals Filtering

Figure 1.14 presents the result of the Huang empirical modes decomposition implementation for the acoustic emission (AE) signal which is shown in Figure 1.7g. The same results were obtained for other AE signals. The analysis of the IMFs functions allows us to conclude that in this case the first and the second modes contain the noise component too, thus these modes should be processed in order to denoise the signal. Figure 1.15 presents the results of the simulation concerning determination of the optimal wavelet from the family of the biorthogonal ones in terms of the minimum value of the criterion (1.7). The results of the simulation have shown that the biorthogonal wavelet *bior*1.1 is the optimal one for processing both the IMFs 1 and IMFs 2 functions since the criterion values of Shannon entropies ratio which have been calculated by the formula (1.7) achieved the minima values in these cases. Figure 1.16 shows the results of the simulation concerning determination of the optimal wavelet decomposition level in the cases of the use of both the IMFs 1 and IMFs 2 functions. The analysis of the obtained charts allows concluding that the wavelet decomposition levels 7 and 8 are the optimal in terms of the criterion (1.7) minima values for the both IMFs functions. Figure 1.17 presents the same results in the case of the thresholding coefficient optimal values determination. The range and the step of the thresholding coefficient value change were determined in accordance with the steps 3.2 and 3.3 of the hereinbefore described algorithm. The value of multiplier δ was taken as 0.5. The optimal value of the thresholding coefficient was determined as the first achieved of the global minimum within the range of this parameter change. The thresholding coefficient values $\tau_1 = 1.47 \cdot 10^{-3}$ and $\tau_2 = 1.16 \cdot 10^{-3}$ were determined for the IMFs 1 and the IMFs 2 functions respectively as the result of this stage implementation. These values were used to process the detail coefficients within the framework of the proposed technique.

The final stage of the hereinbefore described procedure is the reconstruction of the signal with the use of both the processed and unprocessed IMFs functions.

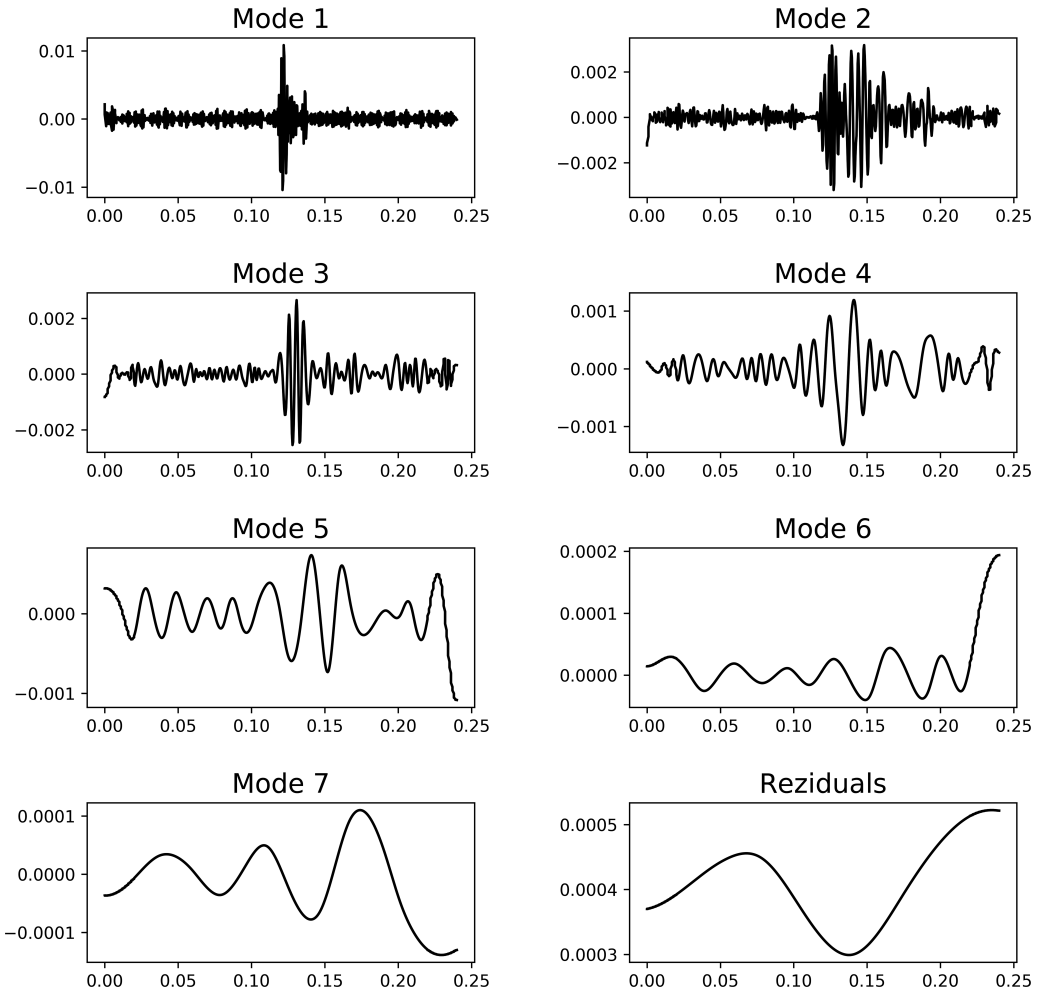

Figure 1.14: The result of the Huang transform

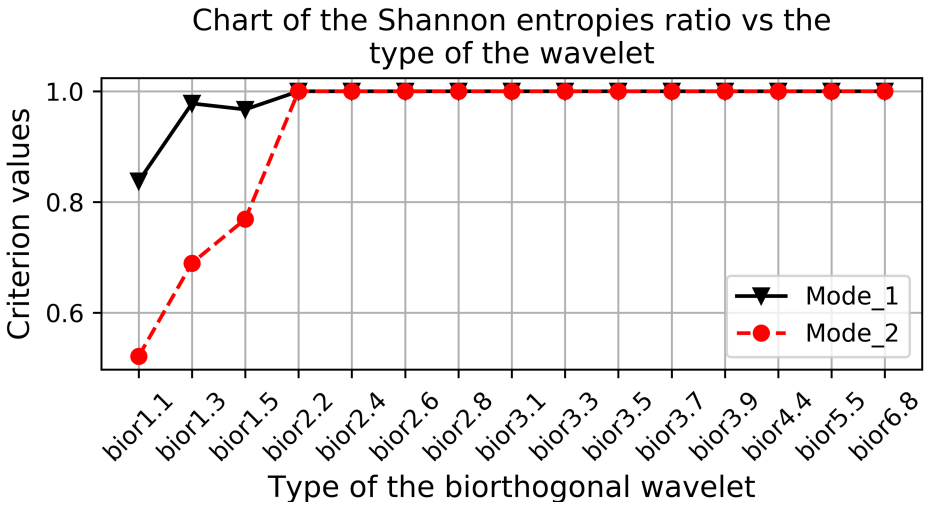

Figure 1.15: Results of the simulation concerning determination of the biorthogonal wavelet optimal type

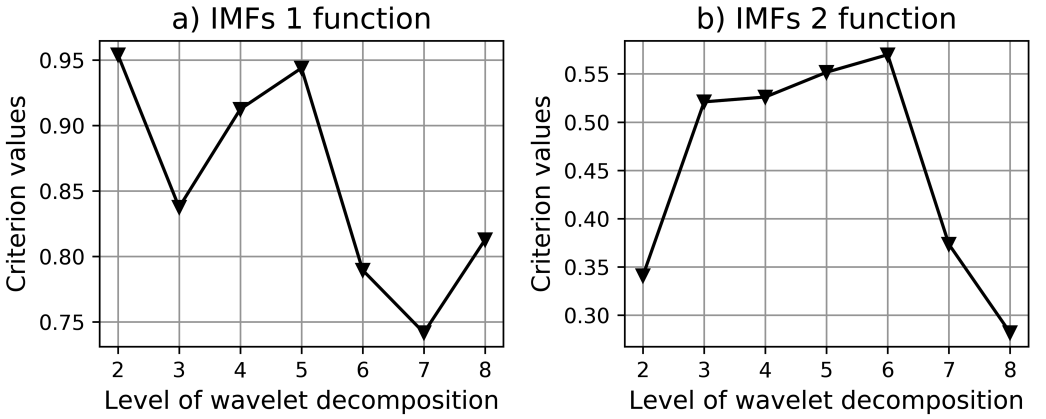

Figure 1.16: Charts of the Shannon entropies ratio vs the wavelet decomposition level

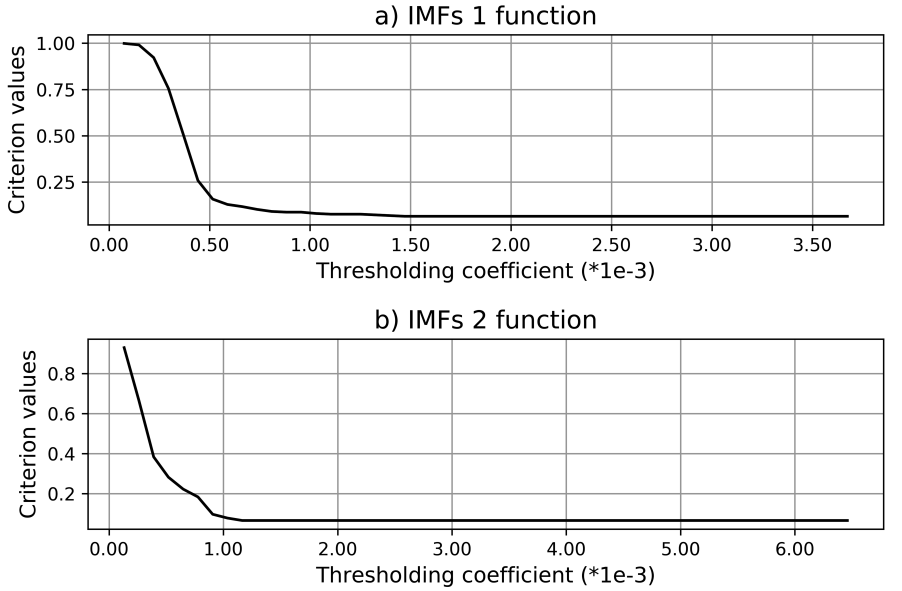

Figure 1.17: Results of the simulation to determine the thresholding coefficient optimal values

Figure 1.18 presents the results of the simulation concerning filtering the AE signal using parameters of the wavelet filter which were determined within the framework of the proposed technique. The analysis of the obtained results indicates the high effectiveness of the proposed technique, since the level of noise in the signal in Figure 1.18b is significantly less in comparison with the level of noise in the initial signal (Figure 1.18a). Table 1.2 presents the values of wavelet filter optimal parameters which were determined within the framework of the proposed technique for the AE signals which have been shown in Figure 1.7. In all cases the biorthogonal wavelet *bior*1.1 was determined as an optimal one in terms of the minimum value of the criterion (1.7). The results of the simulation concerning filtering the AE signals which have been shown in Figure 1.7 are presented in Figure 1.19. The analysis of the obtained results allows concluding that the level of noise component in signals in Figure 1.19 is significantly less in comparison with the noise level in appropriate initial signals which are shown in Figure 1.7. It should be noted, that local particularities of the signals have not been changed during the signals processing. Moreover, the filter parameters are adapted to the filtered component. In all cases the parameters are determined empirically based on the minimum value of the quality filtration criterion.

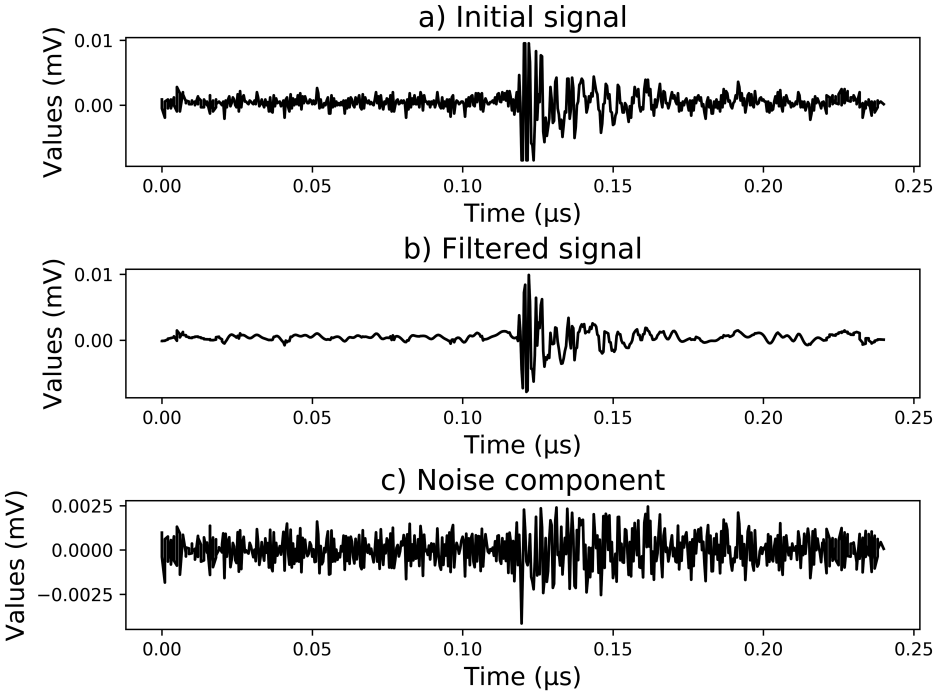

Figure 1.18: Results of the AE signal filtering

Table 1.2: Parameters of the wavelet filter in the cases of the AE signals processing

Weight,N	190		212		270		362	
Modes	M1	M2	M1	M2	M1	M2	M1	M2
Level	8	8	8	2	7	2	8	2
THR,10^{-3}	0.66	0.31	0.37	0.24	0.64	0.25	0.8	0.71
Weight,N	378		381		382		393	
Modes	M1	M2	M1	M2	M1	M2	M1	M2
Level	8	2	3	3	7	8	2	3
THR,10^{-3}	0.95	0.55	0.75	0.82	1.47	1.16	0.89	0.47

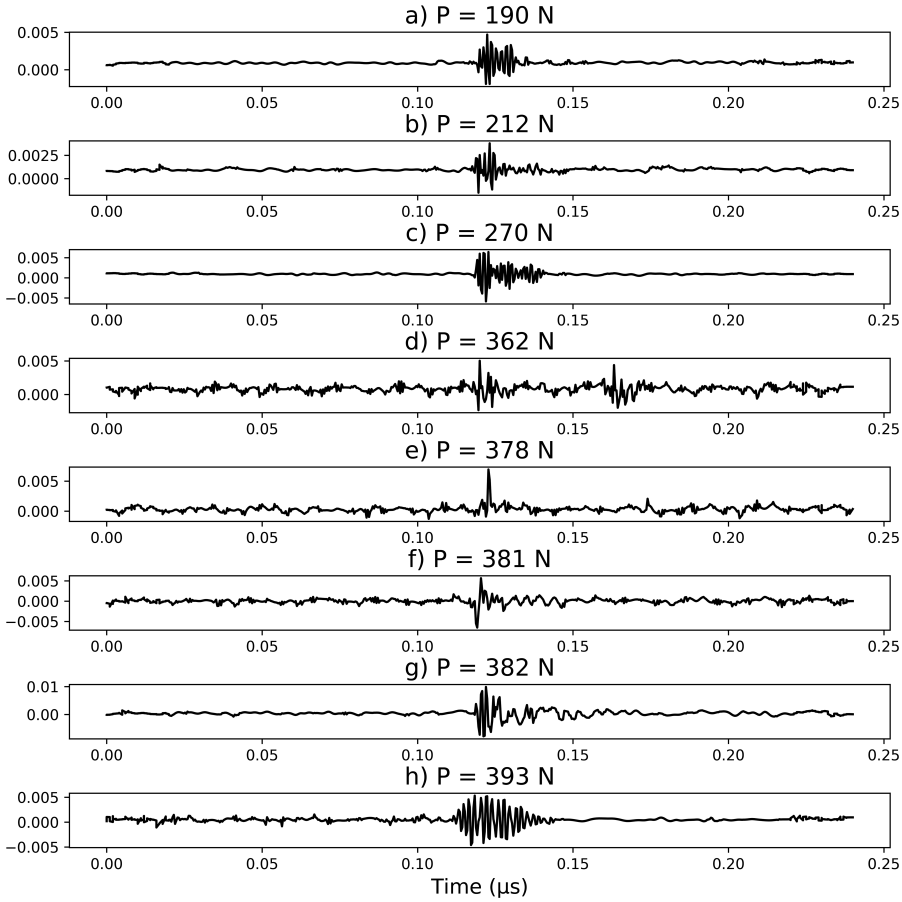

Figure 1.19: The final results of the AE signals filtering

1.6 Conclusions

The technique of 1-D noised signals filtering based on the complex use of empirical mode decomposition (EMD) method (Huang transform) and wavelet analysis has been presented in this chapter. Implementation of this technique involves the following stages: in the beginning, the Huang transform has been performed to decompose the initial signal into the IMFs functions (modes). The modes with noise are allocated at this stage; then, the optimal parameters of the wavelet filter have been determined for each of the selected modes. The wavelet filtering of the allocated modes is performed as the result of this stage implementation; finally, the reconstruction of the signal has been carried out with the use of both the processed and non-processed IMFs functions. The ratio of Shannon entropies which are calculated for both the filtered signal and the allocated noise component has been used as the main criterion to determine the wavelet filter optimal parameters. The effectiveness of the proposed technique was estimated with the use of both the synthetic and acoustic emission (AE) signals obtained as the result of the experiment concerning identification of the structural particularities of the mechanism of the materials deforming based on the AE signals analysis. The family of biorthogonal wavelets has been used during the simulation process. The optimal wavelet decomposition level and the thresholding coefficient value for each of the allocated IMFs functions have been determined during the simulation process. The practical implementation of the proposed technique has shown its high effectiveness since the level of noise component in the filtered signals was significantly less in comparison with the level of the noise in appropriate initial signals. Moreover, the local particularities of the signals have not been changed during the signals processing.

Chapter 2

Objective Clustering Inductive Technology

2.1 Introduction

One of the current directions of modern Data Science is data clustering [92, 33, 38, 136]. Implementation of this technique allows us to divide the objects or features into groups considering the level of their mutual similarity. There are a lot of clustering algorithms nowadays. Choice of the appropriate algorithm is determined by type and particularities of the investigated data. So, in the papers [92, 33, 136] the authors presented the results of the research concerning application of both k-means and fuzzy c-means clustering algorithms to cluster analysis of complex data. The paper [88] is devoted to implementation of DBSCAN clustering algorithm to detect communities in social networks. The tasks concerning implementation of both self-organizing SOTA and hierarchical clustering algorithms are solved in [38, 61, 123]. The results of the research concerning practical implementation of self-organizing neural networks (Kohonen Map) are presented in [55, 128]. However, it should be noted that result of appropriate clustering algorithm operation in the most cases depends on its initial parameters. Setup of these parameters is not easy task and this step is usually implemented empirically during the simulation process taking into account the aim of the solved task.

Evaluation of clustering quality is another task which has not unambiguous solution nowadays. There are a lot of internal clustering quality criteria which allows estimating the character of both the objects distribution within the clusters and the clusters distribution in the features space. These criteria are implemented as the functions in the package clusterCrit [44] of R software [75]. However, the results of the simulation, which were carried out in [21] have shown inconsistence of these criteria in the case of the use of similar datasets. As results of the simulation, the au-

thors in this paper proposed the complex internal clustering quality criterion which is calculated as the multiplicative combination of the Calinski-Harabasz criterion [37] and WB-index [149].

However, as a rule, the internal criteria do not always allow us to divide the objects into clusters objectively. One of the current problems of the existing clustering algorithms is the reproducibility error. In other words, successful clustering results obtained on one dataset do not repeat while using another similar dataset. Reduction of this error can be achieved by careful verification of the obtained model using "fresh information", which was not used during the model making. A higher degree of coincidence between the clustering results on the similar datasets corresponds to a higher degree of the obtained model objectivity. This idea is the basis of the objective clustering inductive technology, the main conception of which was presented in [102, 78] and further developed in [131, 142, 14]. Implementation of this technology involves determination of the optimal clustering based on the extremum value of the complex balance criterion which contains as the components both the internal and external clustering quality criteria. The practical implementation of the objective clustering inductive technology based on the k-means, agglomerative hierarchical, self-organizing SOTA, density-based DBSCAN and OPTICS clustering algorithms were presented in [20, 15, 9, 22].

This chapter presents the results of the authors' research concerning development of the objective clustering inductive technology and its practical implementation based on various clustering algorithms.

2.2 Basic Concepts of the Objective Clustering Inductive Technology

2.2.1 Problem Statement

Let the initial dataset of the objects is a matrix: $A = \{x_{ij}\}, i = 1, \ldots, n;\ j = 1, \ldots, m$, where n is the number of the studied objects; m is the number of features characterizing the objects. The aim of the clustering process is a partition of the objects into non-empty subsets of pairwise non-intersecting clusters, herewith a surface which divides the clusters can take any shape:

$$K = \{K_s\}, s = 1, \ldots, k; K_1 \cup K_2 \cup \cdots \cup K_k = A; K_i \cap K_j = \emptyset, i \neq j,$$

where k is the number of clusters, $i, j = 1, \ldots, k$. Inductive model of objective clustering assumes a sequential enumeration of clustering in order to select from them the best variants. Let W is the set of all admissible clustering for given set A. The best objective on quality criteria $QC(K)$ is the clustering for which is:

$$K_{opt} = \arg \min_{K \subseteq W} CQ(K) \text{ or } K_{opt} = \arg \max_{K \subseteq W} CQ(K)$$

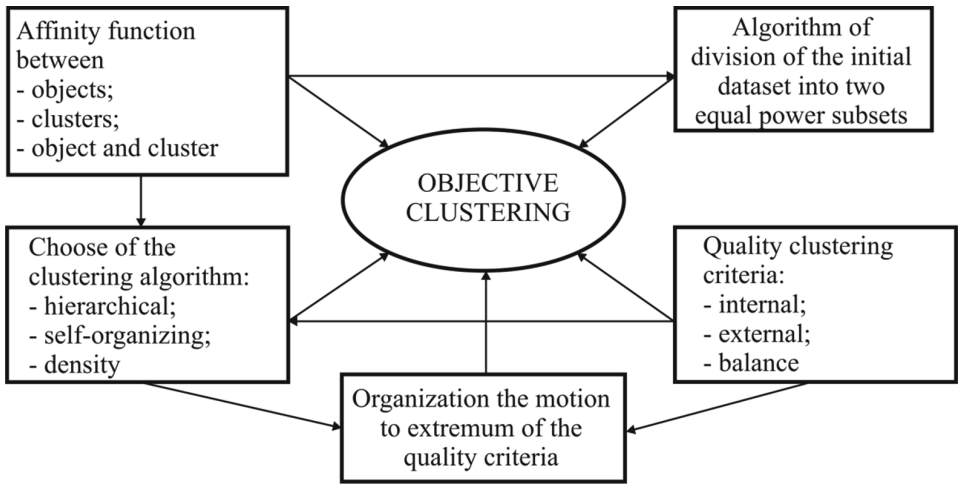

Figure 2.1: Charts of the modules interaction within the objective clustering inductive technology

Clustering $K_{opt} \subseteq W$ is an objective if there is the least difference of this clustering from an expert one in terms of the objects number, the character of the objects distribution in the appropriate clusters and the quantity of discrepancies [102]. Figure 2.1 shows the chart of the modules interaction within the objective clustering inductive technology. As it can be seen, implementation of this technology assumes the following stages:

1. Assignment an affinity function of studied objects, i.e., finding the metric to determine the degree of objects, clusters and objects and clusters similarity.

2. Development of the algorithm to partition the initial set of the objects into two equal power subsets. The equal power subsets are the subsets which contain the same number of pairwise similar objects.

3. Assignment a method of clusters formation (sorting, regrouping, grouping, division, etc.).

4. Assignment the clustering quality criteria: internal, external an complex balance.

5. Organization of motion to the extreme or optimal value of the clustering quality criteria.

6. Assignment an objective clustering fixation method corresponding to the extreme values of the used criteria.

2.2.2 Principles of the Objective Clustering Inductive Technology

Three fundamental principles, borrowed from different scientific fields allowed us to create the complete, organic and interconnected theory. These principles are the basis of the methodology of the complex systems inductive modeling [125, 110]:

- the principle of heuristic self-organization, i.e., sequential enumeration of various complicating models-applicants in order to select from them the best models by a group of criteria for assessing the quality of the model operation;
- the principle of external addition, i.e., the necessity of the use «fresh information» for purpose of objective verification of the models;
- the principle of inconclusive of solution, i.e., generation a set of intermediate results in order to select from them the best variants.

Implementation of these principles in the adapted version provides the conditions to create the methodology of objective clustering inductive technology.

Principle of sequential enumeration

Objective clustering inductive technology assumes a sequential enumeration of the clustering within the admissible range with the use of two equal power subsets which contain the same quantity of pairwise similar objects. The clustering result is estimated at each step of this procedure implementation by calculating the internal and external clustering quality criteria, which consider the character of objects and clusters distribution in various clustering and the difference of the clustering results obtained on two equal power subsets. The model self organizes in such a way that the better clustering correspond to the extreme values of these criteria.

Principle of external addition

Implementation of this principle within the framework of the objective clustering inductive technology assumes the existence of two equal power subsets, which contain the same number of pairwise similar objects. Clustering is carried out on these subsets concurrently during the algorithm operation with sequential comparison of the clustering results by the use of both the internal and external clustering quality criteria. The idea of the algorithm to divide the initial dataset into two equal power subsets A and B is stated in [77] and further developed in [125]. Implementation of this algorithm assumes the following steps:

1. Calculation of $\frac{n \times (n-1)}{2}$ pairwise distances between the objects in the initial dataset. The result of this step is a triangular matrix of the distances.

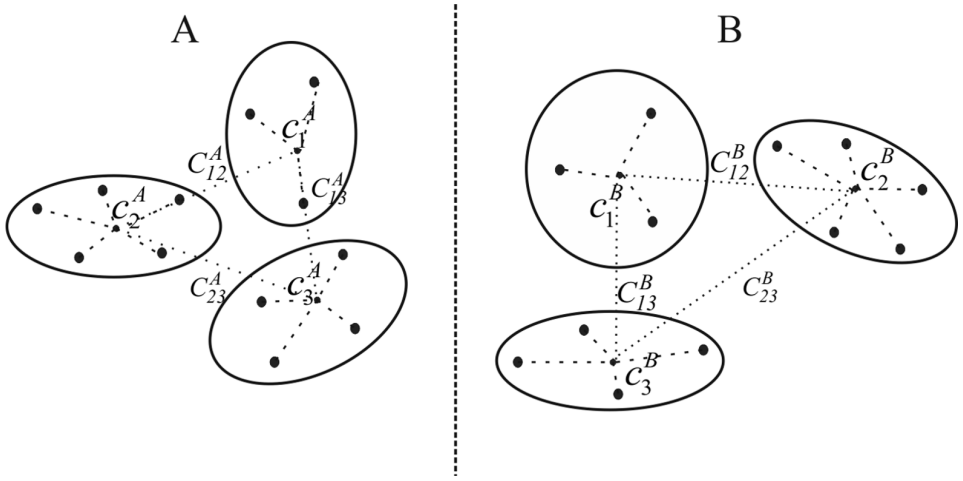

Figure 2.2: An example of objects and clusters distribution in OCIT

2. Allocation of the pairs of objects X_s and X_p, the distance between which is minimal:
$$d(X_s, X_p) = \min_{i,j} d(X_i, X_j);$$

3. Distribution of the object X_s to subset A, and the object X_p to subset B.

4. Repetition of the steps 2 and 3 for the remaining objects. If the number of objects is odd, the last object is distributed to the both subsets.

Principle of inconclusive of solution

Implementation of this principle involves a fixation of clustering which correspond to the extreme values of complex balance clustering quality criterion, which includes as the components both the internal and external criteria. Each local extremum corresponds to an objective clustering with a certain degree of detailing. The final choice and therefore the fixation of the clustering is determined by the goals of the current task.

2.2.3 Clustering Quality Criteria

Figure 2.2 shows an example of both the objects and clusters distribution within the framework of the objective clustering inductive technology (OCIT). It is obvious, that the best clustering corresponds to the higher density of the objects grouping relative to the mass centers of the clusters where these objects are allocated on the one hand, and the less density of the clusters' mass centers distribution in the feature

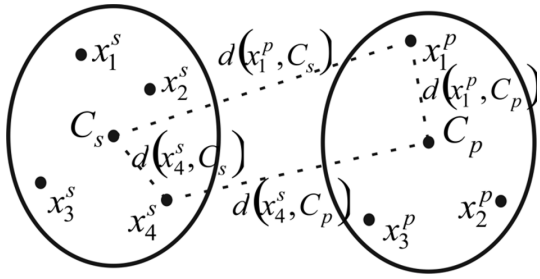

Figure 2.3: Model of objects and clusters distribution in two cluster structure

space on the other hand. Moreover, the difference of the clustering results obtained on the equal power subsets should be minimal. Thus, to implement this technology it is necessary to determine for the investigated data the proximity metric, the internal, the external and the complex balance clustering quality criteria.

Comparison analysis of the proximity metrics

In [21] the authors presents the results of the research concerning comparison of the three well known metrics to estimate the proximity level of high-dimensional numeric vectors: Manhattan, Euclidean and correlation distances. Evaluation of the metrics effectiveness was performed using the synthetic data representing the gene expression profiles of the objects in two different clusters (see Figure 2.3). Gene expression profile in this case means the vector of gene expressions, the values of which were determined for different samples. The gene expression is presented as a numeric value. Centers of the corresponding clusters within the framework of the model were calculated as follows:

$$C_s = \frac{1}{N_s} \sum_{i=1}^{N_s} x_i^s,$$

where N_s is the quantity of gene expression profiles in cluster s, x_i^s is i-th profile in cluster s. The simulation process consisted the following steps:

- calculation of the average distance d_{int} from the profiles to the clusters' centers, where these profiles were allocated:

$$d_{\text{int}}(X^{s,p}, C_{s,p}) = \frac{1}{N} \left(\sum_{i=1}^{N_s} d(x_i^s, C_s) + \sum_{j=1}^{N_p} d(x_j^p, C_p) \right);$$

- calculation the average distance d_{ext} from the profiles to the centers of the

neighbouring clusters:

$$d_{\text{ext}}(X^{s,p}, C_{s,p}) = \frac{1}{N}\left(\sum_{i=1}^{N_s} d(x_i^s, C_p) + \sum_{j=1}^{N_p} d(x_j^p, C_s)\right);$$

- calculation the relative coefficient:

$$d_{\text{rel}}(X^{s,p}, C_{s,p}) = \frac{d_{\text{ext}}(X^{s,p}, C_{s,p})}{d_{\text{int}}(X^{s,p}, C_{s,p})};$$

Higher value of the relative coefficient corresponds to the higher separating ability of the used proximity metric.

Evaluation of the used metrics effectiveness was performed with the use of gene expression profiles of patients which were investigated on lung cancer disease. The data were submitted from database Array Express [30] and they included the gene expression profiles of 96 patients, 10 of which were healthy and 86 patients were divided by the state of the health severity into three groups (Well, Moderate and Poor). Each of the profiles included 7129 of genes. To evaluate the appropriate metric effectiveness the group of the health patient (10 profiles) and the group of patients with poor state of health (21 of profiles) were used. The results of the simulation are shown in Figure 2.4.

The analysis of obtained results allows us to conclude that in the case of high dimensional gene expression profiles use the correlation metric has significantly higher separating ability in comparison with Euclidean and Manhattan metrics since the values of the relative criterion, which is calculated based on the correlation distance, are higher in comparison with the use of Euclidean and Manhattan distances. However, in the case of low dimensional data the Euclidean metric has higher separating ability level. Thus, the choice of the proximity metric depends on type of the investigated data. Hereinafter, we will use the Euclidean and correlation metrics for low and high dimensional data respectively.

The internal and the external clustering quality criteria

As it was noted hereinbefore, it is obvious that the qualitative clustering corresponds to the less density of the clusters distribution and higher density of the objects concentration inside the clusters. Thus, the internal clustering quality criterion should be complex and considering both the objects distribution inside different clusters and the clusters distribution in the features space. The first component of the complex internal criterion is calculated as an average distance from the objects to the mass centers of the clusters, where these objects are allocated:

$$QCW = \frac{1}{N}\sum_{s=1}^{K}\sum_{i=1}^{N_s} d(x_i^s, C_s) \tag{2.1}$$

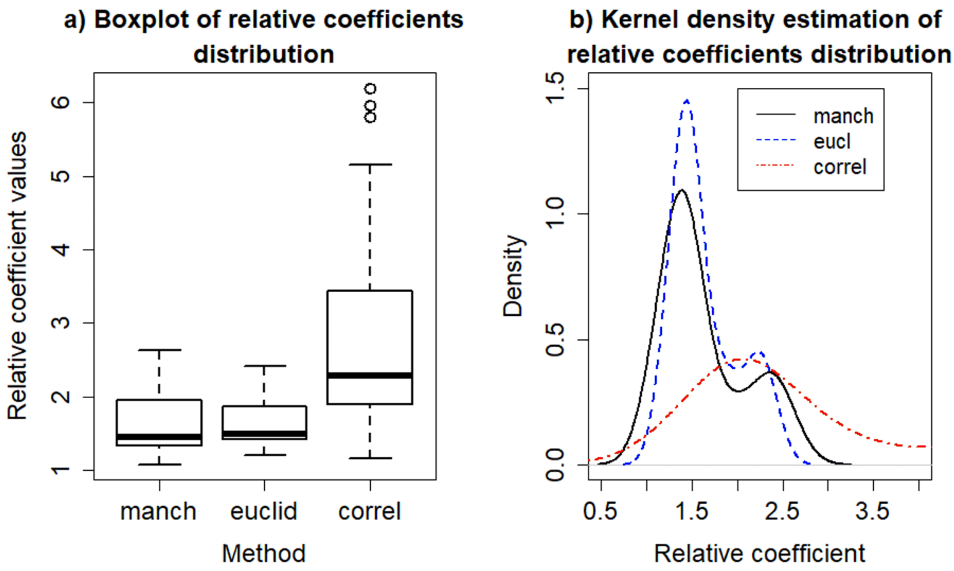

Figure 2.4: Charts of the relative criterion values distribution using different metrics: a) box plot; b) kernel density plot

Table 2.1: Internal clustering quality criteria

N	Index	Short name	Ref.	Rule
1	Banfeld Raftery	BR	[28]	min
2	C index	C_index	[74]	min
3	Calinski Harabasz	CH	[37]	max
4	WB index	WB	[149]	min
5	PBM	PBM	[27]	max
6	Ray Turi	RT	[118]	min
7	Xie Beni	XB	[139]	min
8	Silhouette	SH	[124]	max
9	Gamma	GM	[25]	max

The second component of this criterion, which takes into account the particularities of the clusters distribution in the feature space, is calculated as an average distance between the mass centers of the clusters:

$$QCB = \frac{2}{K(K-1)} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} d(C_i, C_j) \tag{2.2}$$

where K is the quantity of clusters; N is the general quantity of objects; N_s is the quantity of the objects in cluster s; x_i^s is the i-th vector in cluster s; C_i, C_j and C_s are the mass centers of the clusters i, j and s respectively, $d(\cdot)$ is the metric used to estimate the proximity level of the studied vectors. Various combinations of these components allow us to calculate the internal clustering quality criteria.

Package clusterCrit [44] of R software [75] contains various functions to calculate the internal clustering quality criteria. Of course, selection of appropriate criterion is determined by type of the studied data and this choice should be performed in each case empirically using synthetic dataset. Below, we present the technique to determine the internal clustering quality criteria in the case of high dimensional gene expression profiles use. Criteria which were used during the simulation process are presented in Table 2.1.

To estimate the effectiveness of the internal clustering quality criteria the gene expression profiles of lung cancer patients were used [30] as the experimental data. Firstly, the data were divided into two equal power subsets with the use of the algorithm that had been presented hereinbefore. Then, each of the subsets was sequentially divided into clusters from $K_{\min} = 2$ to $K_{\max} = 5$. In the case of two-cluster structure in the first cluster there were the gene expression profiles of healthy patients (NORM) and gene expression of the patients with good state of health (WELL), second cluster included the gene expression of the patients with poor (POOR) and moderate (MODERATE) states of health. In the case of three-

cluster structure the first cluster contained the data of the healthy patients, the second one contained the data of the patients with good state of health, the third cluster included the gene expression of the patients with poor and moderate states of health. In the case of four-cluster structure the first cluster contained the data of the healthy patients, the second cluster contained the data of the patients with good state of health, the third cluster included the gene expression of the patients with poor state of health and the fourth cluster contained the gene expression of the patients with moderate state of health. To obtain a five-cluster structure the gene expression profiles of the patients with moderate state of health were divided into two groups randomly. The optimal clustering in this case corresponded to four-cluster structure. To estimate the proximity level of the appropriate vectors, we used the correlation metric. Figure 2.5 shows the charts of the internal clustering quality criteria for equal power subsets A and B versus the clusters quantity. As it can be seen from Figure 2.5, only the criteria WB,CH,PBM and SH allows derermining optimal clustering since their extrema correspond to four clusters structure.

The external clustering quality criterion was calculated as the normalized difference of the internal clustering quality criteria determined for the equal power subsets A and B:

$$QC_{\text{ext}}(A,B) = \frac{|QC_{\text{int}}(A) - QC_{\text{int}}(B)|}{QC_{\text{int}}(A) + QC_{\text{int}}(B)}. \tag{2.3}$$

This choice is determined by the following reason: the equal power subsets contains the mutually similar objects. In the case of applying the clustering algorithm with the same parameters the character of both the clusters and objects distributions in the clustering obtained on these subsets should be almost the same. As a result, the internal criteria values in this case should have minimal difference between each other and their normalized difference for the objective clustering should has extremum.

Figure 2.6 shows the charts of the external clustering quality criteria, which were calculated based on the selected internal criteria versus the clusters quantity. The analysis of the obtained charts allows us to conclude that PBM criteria is not reasonable in this case since the external criterion which was calculated based on appropriate internal criteria does not allow us to distinguish the objective clustering. The external criterion calculated based on SH index has the brightest minimum. However, the negative values of the SH internal criterion for some clustering (3) can complicate the results interpretation in more complicated cases. Thus, the criteria CH and WB are optimal ones to determine the objective clustering in terms of both the internal and external clustering quality criteria. As a result of the simulation we have proposed the complex internal clustering quality criterion which is calculated as multiplicative combination of CH and WB criteria:

$$QC_{int} = \frac{QC_{WB}}{QC_{CH}} = \frac{K(K-1)QCW^2}{(N-K)QCB^2}; \tag{2.4}$$

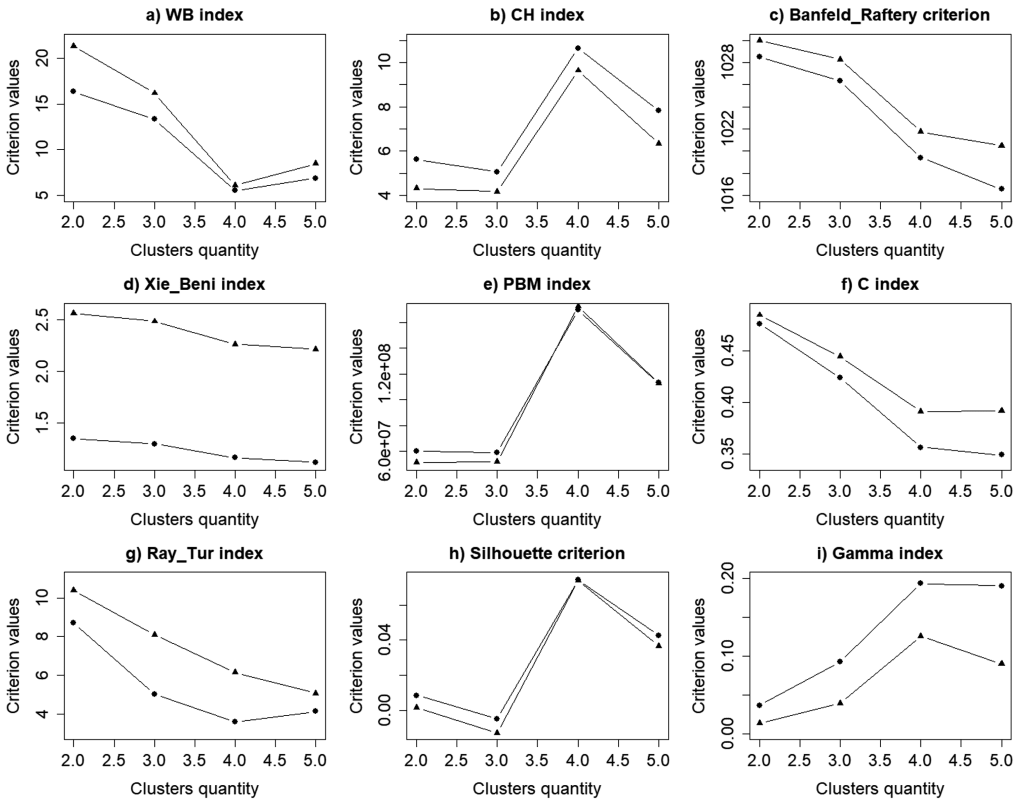

Figure 2.5: Charts of the internal clustering quality criteria versus the clusters quantity

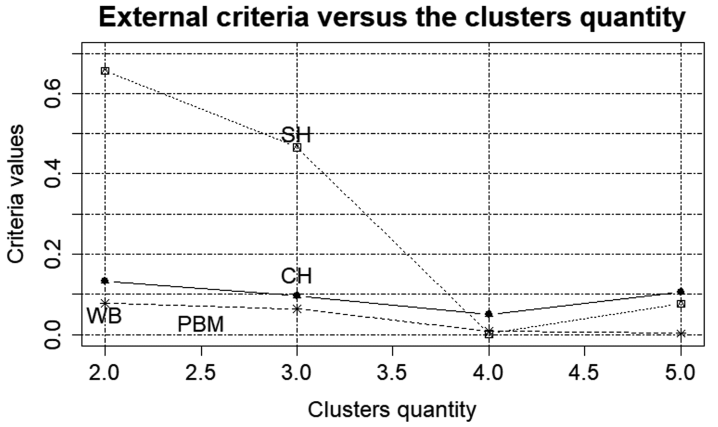

Figure 2.6: Charts of the external clustering quality criteria versus the clusters quantity

where K is the quantity of clusters; N is the quantity of the objects. The minimum value of this criterion corresponds to optimal clustering.

It should be noted, we have estimated the effectiveness of this criterion using various types of data sets and this applying really allowed us to determine the optimal clustering in various cases. So, we will use the criterion (2.4) as the internal one for the following research. However, we do not want to conclude that this criterion is the best in all cases. As we remarked hereinbefore, the choice the internal criterion depends on type of investigated data and this step should be performed in each case empirically.

Balance clustering quality criterion

The necessity of the balance clustering quality criterion is determined by the following reasons. Of course, the objective clustering corresponds to the minimum values of the internal and the external clustering quality criteria. However, it is possible that the extrema of these criteria are disagree between each other. Thus, in this case it is necessary to determine the balance criterion which considers both the character of the objects and the clusters distribution in various clustering and the difference between clustering, which are implemented on the two equal power subsets. To calculate this criterion we used Harrington desirability function [66]. The chart of this function is shown in Figure 2.7. Calculation of this criterion assumes the following steps:

1. Transformation of scales of the internal and the external clustering quality

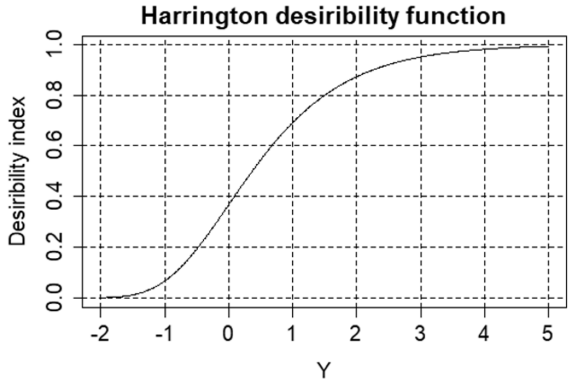

Figure 2.7: Harrington desirability function

criteria into reaction scale Y in the following way:

$$Y = a - b \cdot QC$$

where a and b parameters which are are determined empirically considering the boundary values of the appropriate clustering quality criteria:

$$\begin{cases} Y_{max} = a - b \cdot QC_{min} \\ Y_{min} = a - b \cdot QC_{max} \end{cases}$$

2. Calculation of the Y_i non-dimensional parameter for each of the used criteria:

$$Y_i = a - b \cdot QC_i$$

3. Calculation of the private desirabilities for each of the criteria:

$$d_i = exp(-exp(-Y_i))$$

4. Calculation of the balance clustering quality criterion as the geometric average of all private desirabilities:

$$QC_{bal} = \sqrt[r]{\prod_{i=1}^{r} d_i} \tag{2.5}$$

where r is the number of both the internal and external clustering quality criteria.

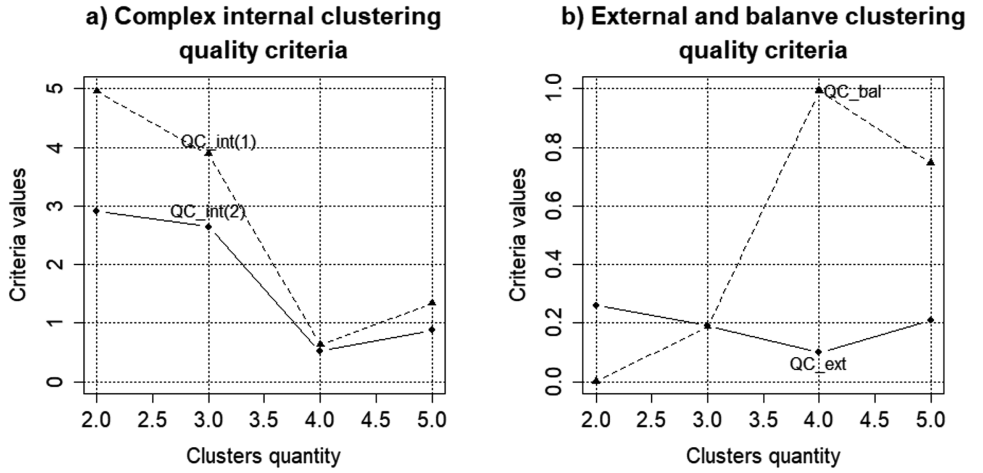

Figure 2.8: Charts of the a) complex internal, b) external and balance clustering quality criteria

The maximum value of this criterion corresponds to the optimal clustering in terms of the used criteria.

Figure 2.8 presents the charts of the complex internal, external and balance clustering quality criteria varsus the number of clusters which were calculated using two equal power subsets. Analysis of the obtained results indicates the high effectiveness of the used criteria for determining the objective clustering. The extreme values of all criteria correspond to four-cluster structure of the investigated objects grouping.

2.2.4 Structural Block-Chart of the OCIT

Figure 2.9 shows the structural block-charts of the objective clustering inductive technology. Its implementation involves the following steps [9]:

1. Preparing, analysing and preprocessing the investigated data. The data is formed as a matrix, where number of rows is a number of the studied objects and number of columns is a number of the features which characterized the objects. The preprocessing stage involves the following: missing values processing, normalization, filtering, at al.

2. Choice of the proximity metric taking into account both the type and particularities of the investigated vectors (Euclidean, Manhattan, correlation at al.).

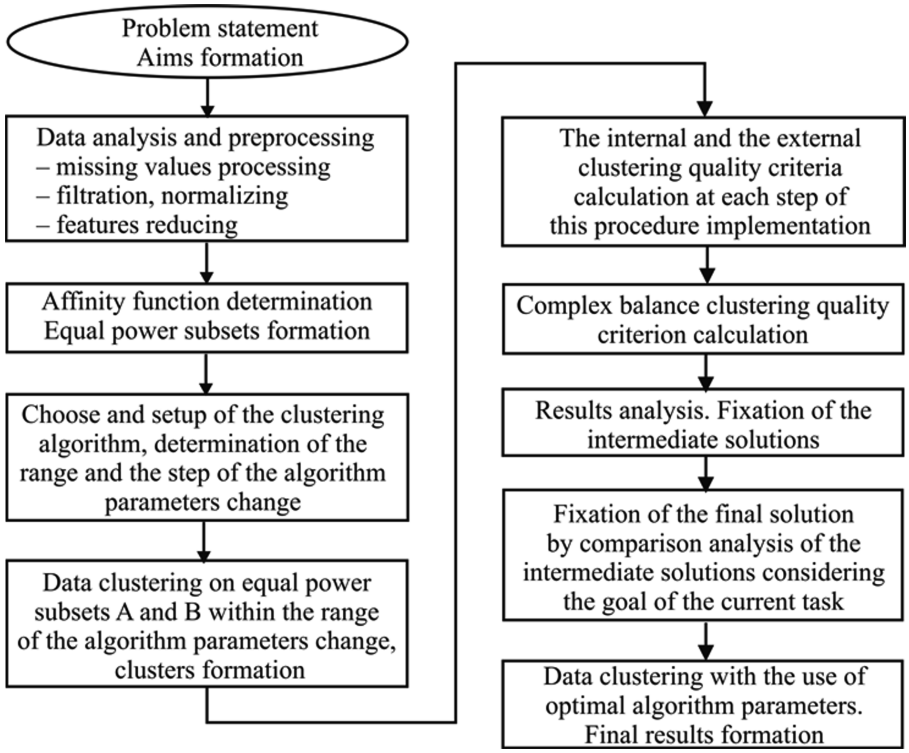

Figure 2.9: Structural block-chart of the OCIT

3. Division of the initial dataset into two equal power subsets (contains the same quantity of the pairwise similar objects).

4. Choice of the clustering algorithm. Setup of its initial parameters and range of these parameters change.

5. Implementation of the clustering algorithm on the equal power subsets within a given range of the algorithm parameters change. Fixation of the clustering at each step of this procedure implementation. Calculation of both the internal and the external clustering quality criteria by the formulas (2.3) and (2.4) in the cases if the quantity of the clusters in the different clustering are the same ones.

6. Calculation of the complex balance clustering quality criterion by the formula (2.5).

7. Results analysis. Fixation of intermediate solutions which correspond to the maxima values of the complex balance criterion.

8. Comparison analysis of the intermediate sollutions and fixation of the final colution consitering the aim of the current task. Determination of the algorithm parameters which correspond to the optimal clustering.

9. Data clustering with the use of the current clustering algorithm using determined before parameters. Final results formation.

2.3 Practical Implementation of the Objective Clustering Inductive Technology

Practical implementation of the OCIT is possible based on various clustering algorithm. Choice of clustering algorithm is determined by type of the studied data and aim of the current task. However, in any case it is necessary to determine the optimal algorithm parameters or to choose the optimal hierarchical level in the case of hierarchical clustering algorithm use. In this section we solve this task within the framework of the OCIT with the use of density-based DBSCAN and self-organizing SOTA clustering algorithms.

2.3.1 Hybrid Model of OCIT Based on DBSCAN Clustering Algorithm

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) clustering algorithm was proposed in 1996 as a solution of the problem to divide the data into clusters of arbitrary shapes [50, 65].

Let D is the database of the points in m-dimensional features space. The following definitions are the basis of DBSCAN clustering algorithm operation [50]:

Definition 1. The *Eps-neighborhood* of a point p is defined as follows:

$$Eps(p) = \{q \in D | dist(p, q) \leq EPS\}$$

where $dist(p, q)$ is the proximity distance between the points p and q.

Definition 2. A point q is directly density-reachable from a point p if the following conditions are performed:

$$\begin{cases} q \in Eps(p) \\ N_{EPS}(p) \geq MinPts \end{cases}$$

where $N_{EPS}(p)$ and $MinPts$ are the number of points and the minimum number of points within Eps-neighborhood of a point p respectively.

Definition 3. A point q is density-reachable from a point p if there is a chain of points $q_1, ..., q_n$, $q_1 = p$, $q_n = q$ such that q_{i+1} is directly density-reachable from q_i.

Definition 4. A point q is density-connected with a point p if there is a point k such that both the points q and p are density-reachable from the point k.

Definition 5. A cluster C is a non-empty subset of a set of points D if the following conditions are performed:

1. $\forall p, q : if\ p \in C$ and q is density-reachable from p, then $q \in C$;
2. $\forall p, q :$ if q is density-connected with p, then $p, q \in C$

Definition 6. Let C_i, $i = \overline{1, k}$ is a set of the allocated clusters. The noise is the set of points of the database D, which not belonging to any cluster C_i:

$$noise = \{p \in D | \forall i : p \notin C_i, i = 1, k\}$$

The key points, which are determined by hereinbefore definitions are shown in Figure 2.10. Here, the point q is directly density-reachable from the point p, but the point p is not directly density-reachable from the point q. The point q is density-reachable from the point k, and the points p and h are density-connected through the point k.

Implementation of DBSCAN clustering algorithm starts from initialisation of EPS and $MinPts$ values. Then, it is necessary to choose arbitrary point p and to retrieve the points, which are density-reachable from p and density-connected to p. If p is a core point, all of the found points are joined into cluster. If p is a border point and no points which are density-reachable from p, escape to the next point of

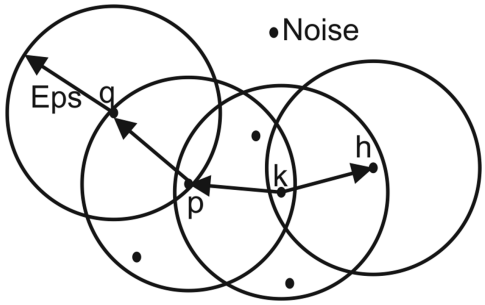

Figure 2.10: Keys points of DBSCAN clustering algorithm (MinPts = 3)

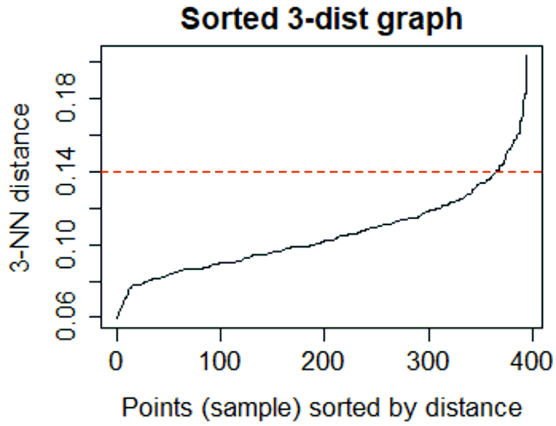

Figure 2.11: An example of sotred k-dist graph

the database. If the object is not a core point and it is not density reachable from other points then, this object is identified as noise.

Thus, result of DBSCAN clustering algorithm operation depends on two parameters: EPS and $MinPts$. To determine the optimal EPS value for appropriate $MinPts$ the authors in [50] proposed the technique based on sorted k-dist graph (see Figure 2.11). To authors' mind, the optimal EPS value for appropriate $MinPts$ value should belong to knee of k-dist graph. However, it should be noted, that implementation of this technique does not allow us to determine the EPS value exactly. This fact influences the quality of the algorithm operation. The implementation of the proposed technique allows us to determine only the range of the EPS values change for appropriate $MinPts$ value. To solve this problem, we propose the technique of the DBSCAN clustering algorithm optimal parameters determination based on the objective clustering inductive technology.

Structure block-chart of algorithm to implement the hybrid model of objective clustering inductive technology based ob DBSCAN clustering algorithm is presented on Figure 2.12. Implementation of the algorithm assumes the following steps:

1. Formation of the matrix of the investifated data. The matrix contains n rows or studied objects and m columns or the objects attributes.

2. Division of the initial dataset into two equal power subsets.

3. Calculation of the distance matrix between the objects for both subsets using correlation distance in the case of high-dimensional data. This distance matrix is the input matrix for the next step of the algorithm operation.

4. Setup of DBSCAN clustering algorithm, choice of both the range and steps of EPS and $MinPts$ values change.

5. Fixation of $MinPts$ value ($MinPts = 3$). Initialization of $EPS = EPSmin$.

6. Data clustering on the two subsets A and B using DBSCAN algorithm in the range from $EPSmin$ to $EPSmax$. Clustering fixation at each step of this procedure implementation.

7. Calculation of both the internal and external clustering quality criteria at each step of the algorithm operation.

8. Calculation of the balance clustering quality criterion.

9. Analysis of the balance clustering quality criterion values. Fixation of the optimal value EPS which corresponds to the maximum value of this criterion.

10. Data clustering on the two equal power subsets A and B within the range from $MinPtsmin$ to $MinPtsmax$. Clustering fixation at each step of this procedure implementation.

11. Repetition of the steps 7–9 of this algorithm for $MinPts$ values. Fixation of EPS and $MinPts$ optimal values which correspond to the maximum of the balance clustering quality criterion.

12. Investigated data clustering using obtained parameters of DBSCAN clustering algorithm operation.

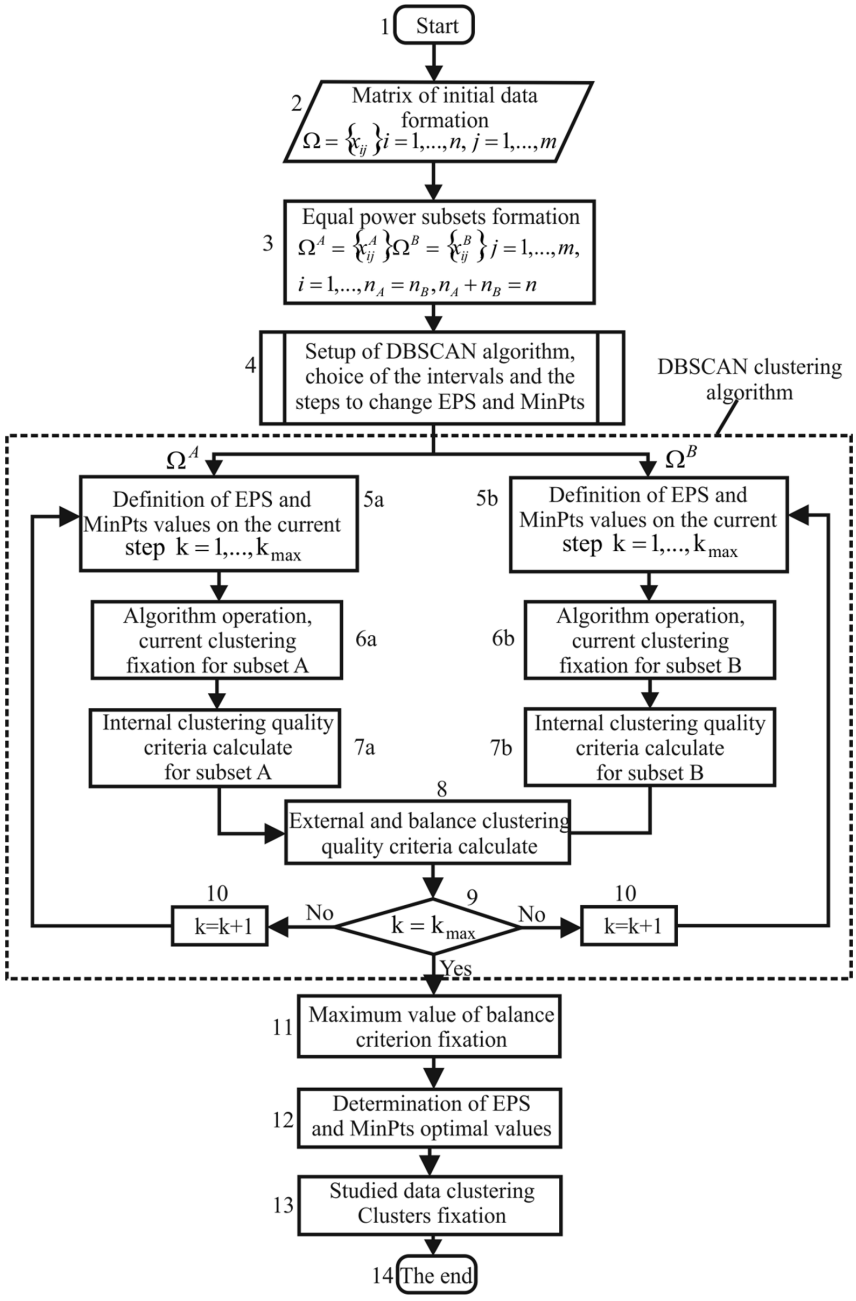

Figure 2.12: Structure block-chart of algorithm to implement the hybrid model of OCIT based on DBSCAN clustering algorithm

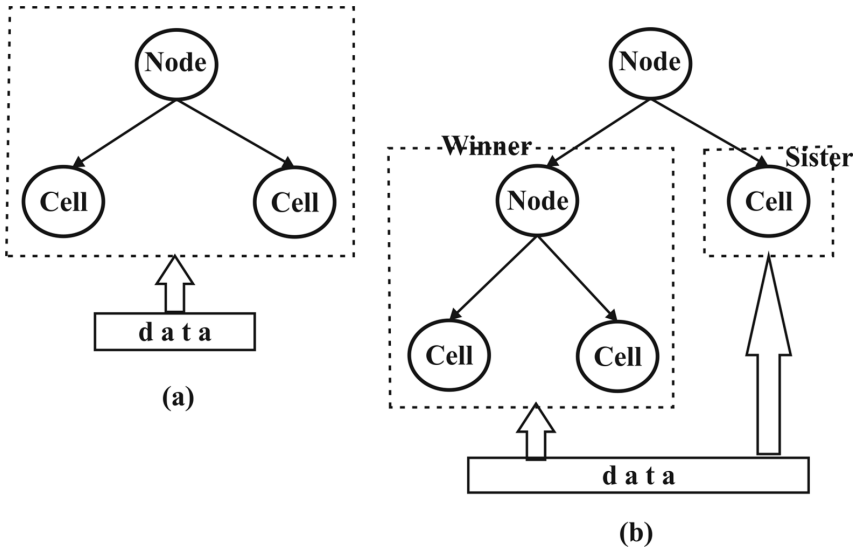

Figure 2.13: Process of cell structure forming: a) initial state of the system; b) state of the system after one cycle

2.3.2 Hybrid Model of OCIT Based on SOTA Clustering Algorithm

SOTA clustering algorithm (Self-Organizing Tree Algorithm) [47] is a type of self-organizing neural networks based on the complex use of Kohonen maps and Fritzke algorithm of spatial cell structure growing [57]. Opposed to Kohonen maps that reflect a set of high dimensional input data on the elements of two-dimensional array of small dimension, SOTA algorithm generates a binary topological tree. Fritzke algorithm performs self-organization of output nodes of network in such a way that quantity of the nodes increases in the field of higher density of objects concentration and decreases in the field of lower density. Figure 2.13 shows the process of cell structure form during SOTA clustering algorithm operation. Initially, the system consists of two cells that are connected through an external root node. In other words, the system has the structure of a binary tree (see Figure 2.13a). Each of the cells or nodes is characterized by a feature vector, the number of elements in which is equal to the dimension of the feature space of the studied data. Implementation of SOTA algorithm assumes the following steps:

1. *Initialization.* Weights which are calculated as an average of the attributes of all of the investigated vectors assign to both the root node and cell vectors. It is obvious that in this case, the length of the vector of weights is equal to the dimension of the data feature space. Parameters for correction of the appropriate cells weight setup in accordance with the following conditions:$\alpha_w > \alpha_m > \alpha_s$,

where α_w, α_m and α_s are parameters for weight correction of the winner cell, root (parent) node and adjacent cell respectively. Also, we should setup a limit value of the variation coefficient E which determines the stopping condition of the algorithm operation.

2. *Adaptation.* The investigated vectors are applied to the input of all external cells during the algorithm operation The degree of proximity of the corresponding vector to the cell weight vector is calculated using the selected affinity function. The winning cell, the vector of weights of which has the smallest distance from corresponding vector is allocated in accordance with the principle of "winner takes everything". The weights of the winner cell, its adjacent and root cells are adjusted in accordance with the formula:

$$C_i(\tau + 1) = C_i(\tau) + \eta \cdot (P_j - C_i(\tau)),$$

where $C_i(\tau)$ and $C_i(\tau + 1)$ are weight vectors for i cell at step τ and $\tau + 1$ respectively; P_j is jth features vector which is entered to the system input; η is the parameter that determines the step of adjusting the weights of the winning cell. The alignment of the weights of the neighbouring relative to winning cell is carried out in accordance with the principle: if the cell adjacent to the winning cell has no offspring, the weights of the winning cell, the neighboring cell and the root node are adjusted. Otherwise, only the winner cells are adjusted. The parameter η at the t iteration is calculated as follows:

$$\eta_t = \alpha \cdot \frac{1 - t}{n} \cdot (1 - b\tau),$$

where t is the total number of vectors presented to the system input; n is the maximum number of the investigated vectors; τ is the number of operations in one cycle; b is the coefficient that determines the rate of the η parameter change; α is the parameter for the weights correction of the appropriate cells.

3. *Convergence of algorithm and network formation.* To determine the structure of a clustering tree, the coefficient of variation for each cell is calculated as an average distance from the cell weights to the feature vectors in a current cell:

$$R_i = \frac{\sum_{k=1}^{K} d(P_k, C_i)}{K},$$

where C_i is the weight vector for ith cell; P_k is the feature vector in this cell; K is the maximum number of the investigated vectors in ith cell. The total value of the variation coefficient is determined as the sum of the variation coefficients of all external cells:

$$\varepsilon_t = \sum_{i=1}^{S} R_i.$$

The criterion for estimating the algorithm convergence is the relative change in the total variation coefficient:

$$|\frac{\varepsilon_t - \varepsilon_{t-1}}{\varepsilon_t}| < E, \tag{2.6}$$

where E is the boundary value of relative change of the variation coefficient. If the condition (2.6) is not fulfilled, the further growth of the tree after each cycle begins from the cell, which has the greatest value of relative change of the variation coefficient. This cell divides into two parts and becomes a root node (see Figure 2.13b). The values of the weight coefficients of the daughter cells and root node are identical. The algorithm is stopped if condition (2.6) is satisfied. Thus, the adjusting the value of the boundary parameter allows us to achieve the desired network structure and, as a result, to get the desired structure of the objects distribution in the clusters.

An analysis of hereinbefore described procedure allows concluding that the clustering results in this case is determined by the parameters for the correction of cell weights and the boundary value of relative change of the variation coefficient. In this section we present the results of the research concerning determination of the SOTA clustering algorithm optimal parameters within the framework of the OCIT [22, 15]. Implementation of this procedure assumes that weight coefficients of the parent's and winner's cells are determined automatically: $pcell = scell \times 5$; $wcell = pcell \times 2$. This ratio is recommended by the authors of the algorithm [47]. The block chart of the algorithm to implement the objective clustering inductive technology based on SOTA clustering algorithm is shown in Figure 2.14. The implementation of this model involves the following steps:

1. Presentation of the studied data as a matrix $n \times m$, where n is the quantity of the studied objects or rows and m is the quantity of attributes or columns.

2. Division of the initial data set into two equal power subsets.

3. Setup of SOTA clustering algorithm. Setting of the *scell* weight parameter initial value, the range and the step of its change.

4. Data clustering on the equal power subsets A and B concurrently. The clusters formation and calculation the internal and the external clustering quality criteria within a range of the algorithm parameters change.

5. Calculation of the balance clustering quality criterion. Fixation of the optimal value *scell* parameter corresponding to the maximum value of the balance criterion.

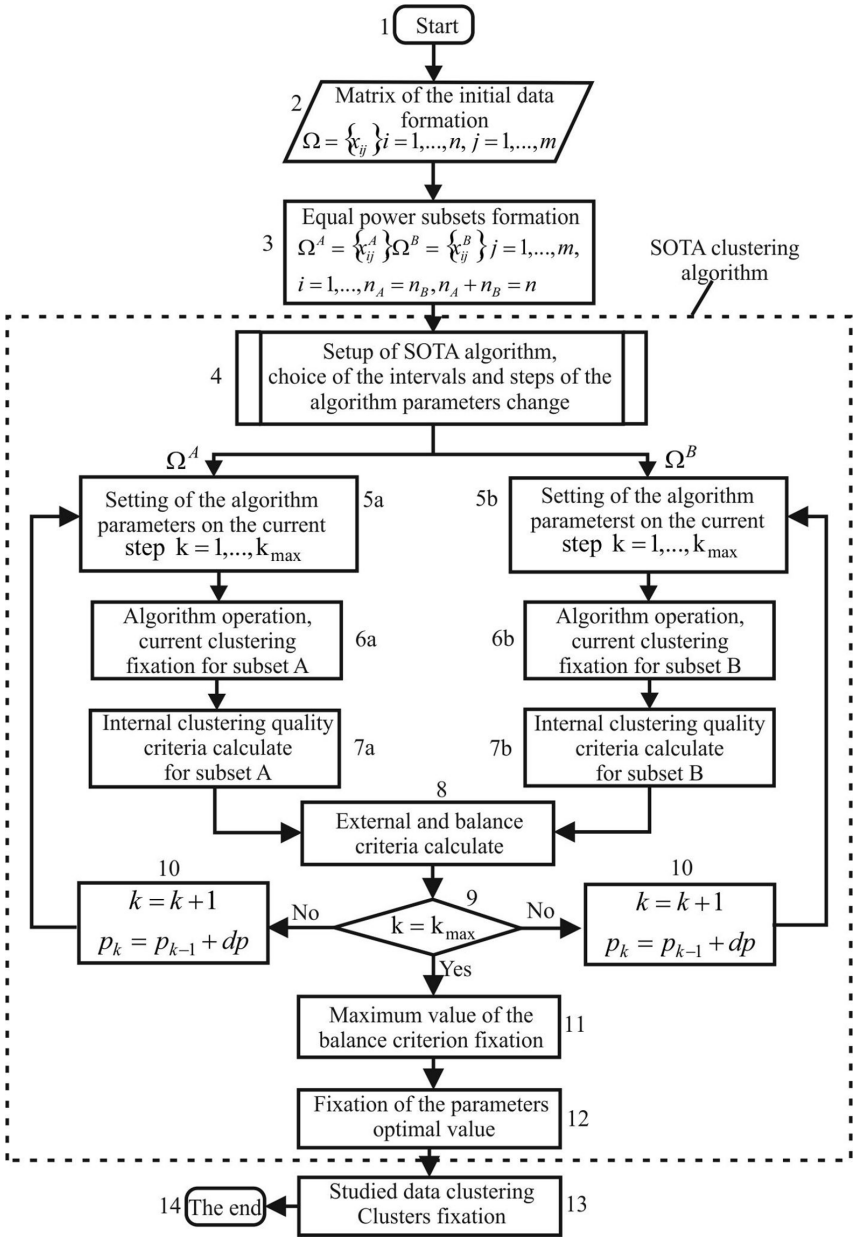

Figure 2.14: Block chart of the algorithm to implement the OCIT based on SOTA clustering algorithm

6. Setting of the initial value of the maximum divergence parameter, range, and step of its change.

7. Repeating step 4 of this procedure. Fixation of the optimal maximum divergence parameter.

8. Full data clustering by SOTA clustering algorithm using the optimal parameters of the algorithm operation.

2.4 Experiments

2.4.1 Experimental Datasets

Evaluation of the hereinbefore presented clustering techniques was performed using the following datasets:

- Datasets of School of Computing of University of Eastern Finland [54]: *Aggregation* [60], *Compound* [145], *Multishapes* [87] and *Jain* [80]. These datasets contain objects that form in two-dimensional space clusters of different shapes. The character of the clusters distribution for the investigated data is shown in Figure 2.15.

- Fisher's Iris [116]. This dataset consists of 3 different types of irises' (Setosa, Versicolour, and Virginica) petal and sepal length, stored in a 150x4 array. The rows being the samples and the columns being: Sepal Length, Sepal Width, Petal Length and Petal Width. Figure 2.16 shows the irises' dataset using parallel coordinates plot. As can be seen from the figure, the profiles of objects belonging to the Setosa class differ from objects of the Virginica and Versicolor classes. This fact allows uniquely identifying the Setosa class. Objects of both Virginica and Versicolor classes have some intersection between each other. Moreover, each of the classes contains objects whose profiles are differed from objects of the general group of the appropriate class. This fact can mean that these profiles can be classified as noise or these objects can be grouped into a separate cluster.

- Database of patients which were investigated on lung cancer disease. The data were submitted from database Array Express [30] and they included the gene expression profiles of 96 patients, 10 of which were healthy and 86 patients were divided by the state of the health severity into three groups: 24, 41 and 21 patients with well, moderate and poor state of health respectively.

- Dataset *moe430a* which contains gene expression profiles obtained by DNA microchip experiments [32]. These profiles contain information concerning

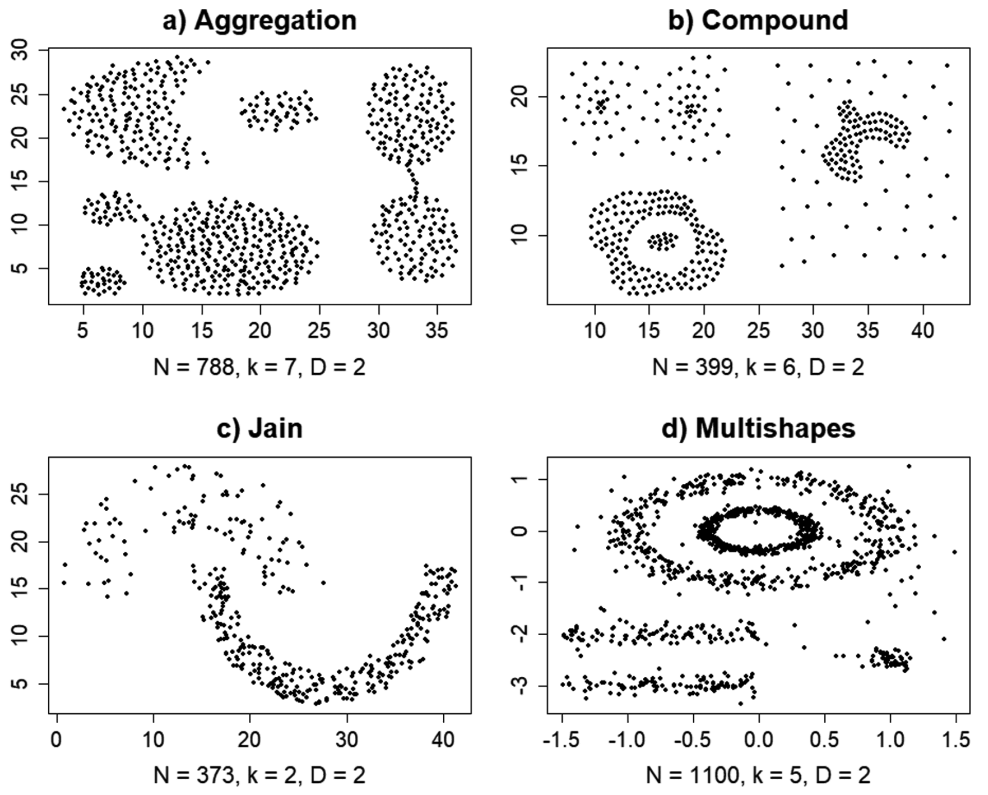

Figure 2.15: Datasets of School of Computing of University of Eastern Finland

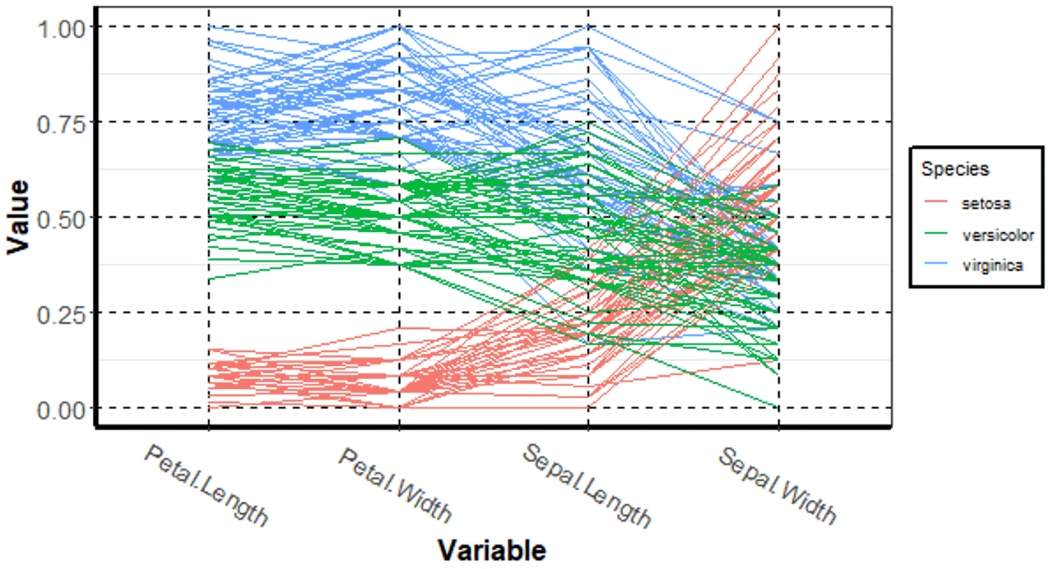

Figure 2.16: Fisher's irises dataset

expression of genes from mesenchymal cells of two types: neural crest and mesoderm. 22 of gene expression profiles were used in the simulation process. In the first case the data contained 147 of genes, in the second case there are 1000 of gene expression profile. This type of profiles was used to evaluate the effectiveness of the model based on the SOTA clustering algorithm.

Initially, the data were divided into two equal power subsets using hereinbefore described algorithm.

2.4.2 Results of DBSCAN Clustering Algorithm Operation

As was described in the section 2.3.1, the result of DBSCAN clustering algorithm operation depends on two parameters: $MinPts$ and EPS. The technique of these parameters determining involves fixing $MinPts = 3$ at the first stage and following step-by-step changing of EPS values in given range with calculation of the balance clustering quality criterion at each step of this procedure implementation. Then, the EPS values corresponding to the maxima of the balance criterion are fixed and the $MinPts$ values are changed from 3 to maximum one with calculation of the balance clustering quality criterion. The range of the EPS value changing is determined based on the sorted k-dist graph analysis. The best decisions concerning choice of the optimal algorithm parameters correspond to the maxima values of the balance

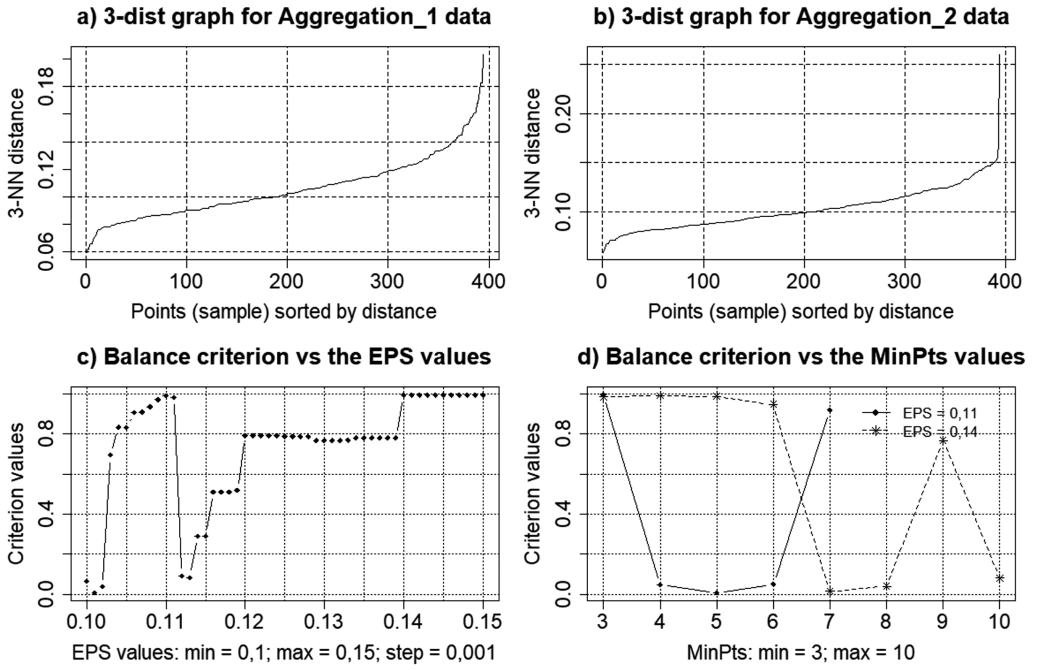

Figure 2.17: Results of the simulation for *Aggregation* dataset: a) sorted k-dist graph for equal power subset A; b) sorted k-dist graph for equal power subset B; c) chart of the balance criterion vs the EPS value for $MinPts = 3$; d) chart of the balance criterion vs the $MinPts$ for optimal EPS values

criterion for each combination of the parameters. Finally, the final solution is taken based on comparison analysis of the intermediate solution considering the aim of the current task.

Figure 2.17 presents the results of the simulation for *Aggregation* dataset. The range of EPS value changing from 0.1 to 0.15 and the step 0.001 were determined as the results of the sorted k-dist graphs analysis (Figure 2.17a,b). Two EPS values (0.11 and 0.14) were determined as the result of the Figure 2.17c analysis. Figure 2.17d shows the charts of the balance criterion versus the $MinPts$ values for the selected EPS values. As a result of this charts analysis the following algorithm parameters combinations were determined: a) $EPS = 0.11$, $MinPts = 3$; b) $EPS = 0.14$, $MinPts = 4$. Results of the data clustering are presented in Figure 2.18. As it can be seen from Figure 2.18, the clustering results is satisfactory in the both cases since there are not any intersection between the obtained clusters. Moreover, in the first case (Figure 2.18a), all clusters are well identified, but the number of points

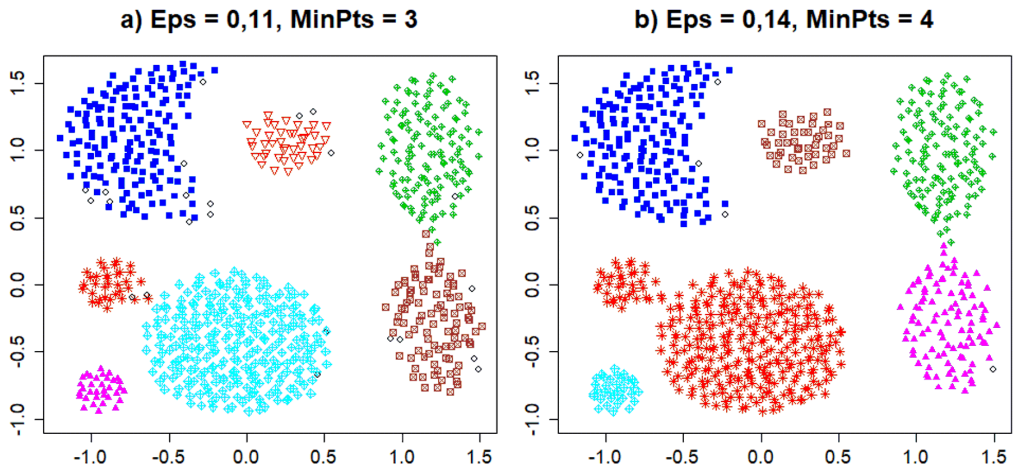

Figure 2.18: Clustering results for *Aggregation* dataset

identified as noise is greater than in the second case (Figure 2.18b). However, in the second case the two clusters are connected with each other and they are not identified as the separate clusters. The choice of parameters in this case is determined by the goals of the current task.

Figure 2.19 presents the results of the simulation in the case of *Compound* dataset use. The range of the EPS value change was setted from 0.1 to 0.2 based on sorted k-dist graphs analysis (Figure 2.19a,b). Step of this parameter changing was taken as 0.002. Three EPS values were selected for the following analysis based on Figure 2.19c analysis: 0.13, 0.166, 0.19. Three combinations of the algorithm parameters were determined as the results of Figure 2.19d analysis: a) $EPS = 0.13$, $MinPts = 3$; b) $EPS = 0.166$, $MinPts = 3$; c) $EPS = 0.19$, $MinPts = 3$. The detail analysis of the obtained results has shown that in the first casethere were the different number of clusters in the obtained clusterings. This fact does not satisfy the condition of the objective clustering. Results of the data clustering for both the second and third cases are presented in Figure 2.20. As it can be seen from Figure 2.20, the algorithm distinguishes well the points with low density of their distribution in the feature space. These points are identified as noise. The smaller EPS value corresponds to the better resolution of the algorithm (Figure 2.20a).

Results of the simulation in the case of *Jain* dataset use are presented in Figure 2.21. The following EPS values were determined as the results of sorted k-dist graphs analysis (Figure 2.21a,b):$EPS_{min} = 0.2$, $EPS_{max} = 0.35$, step of the EPS value change $= 0.005$. Two EPS values were determined as the result of Figure

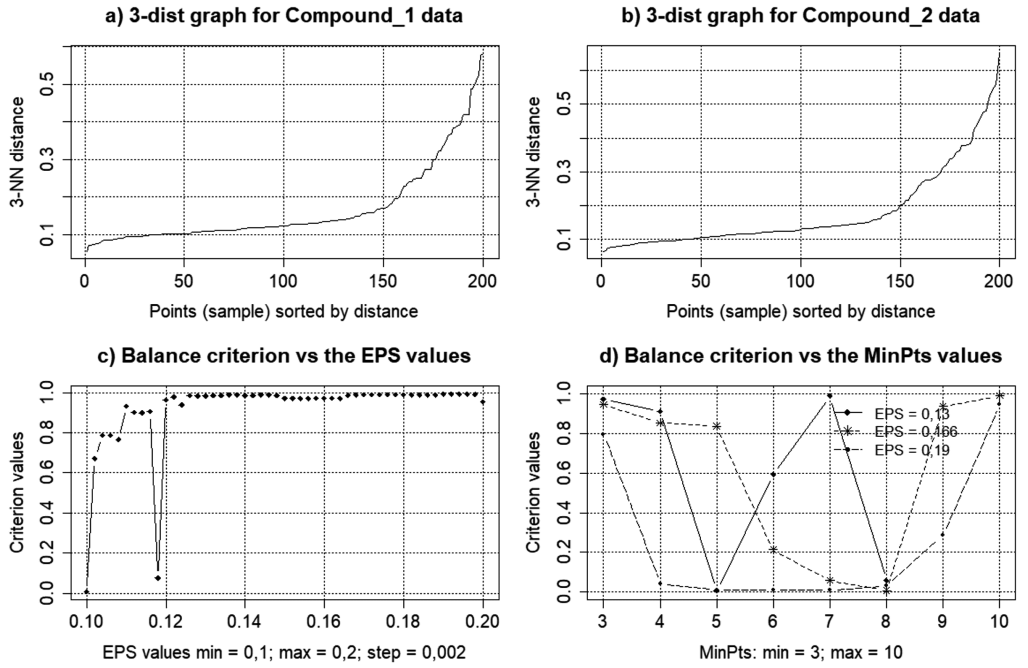

Figure 2.19: Results of the simulation for *Compound* dataset: a) sorted k-dist graph for equal power subset A; b) sorted k-dist graph for equal power subset B; c) chart of the balance criterion vs the EPS value for $MinPts = 3$; d) chart of the balance criterion vs the $MinPts$ for optimal EPS values

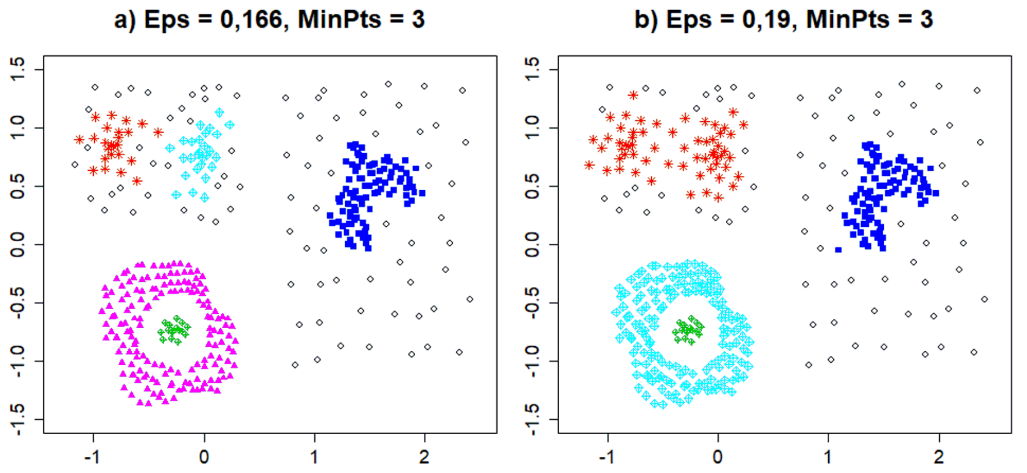

Figure 2.20: Clustering results for *Compound* dataset

2.21c analysis: 0.265 and 0.31. Two combination of the algorithm parameters were determines as the result of Figure 2.21d analysis: a) $EPS = 0.31$, $MinPts = 3$; b) $EPS = 0.31$, $MinPts = 5$. Results of the clustering for *Jain* dataset are presented in Figure 2.22.

Results of both the simulation and clustering for *Multishapes* dataset are presented in Figure 2.23 and Figure 2.24 respectively. As it can be seen, the algorithm parameters combination $EPS = 0.25$, $MinPts = 4$ allows us to obtain correct clustering in the case of *Multishapes* dataset use.

The analysis of the obtained results in the case of the use of two-dimensional synthetic datasets containing clusters of different shapes has shown that the use of DBSCAN clustering algorithm within the framework of the objective clustering inductive technology allows us to group the investigated objects into clusters correctly. The points with smaller density of distribution in the feature space in comparison with density of other objects that make up the clusters are grouped into a separate cluster. These points are identified as noise. In accordance with the principles of the objective clustering inductive technology, the best solutions are formed during the simulation process. These solutions correspond to the maximum values of the balance clustering quality criterion and they are presented as the optimal combinations of algorithm parameters. The choice of the final decision is determined by the goals of the current task.

Figure 2.25 presents the same results in the case of Fisher's irises dataset use. In this case the EPS value was changed within the range from 0.5 to 1 with the

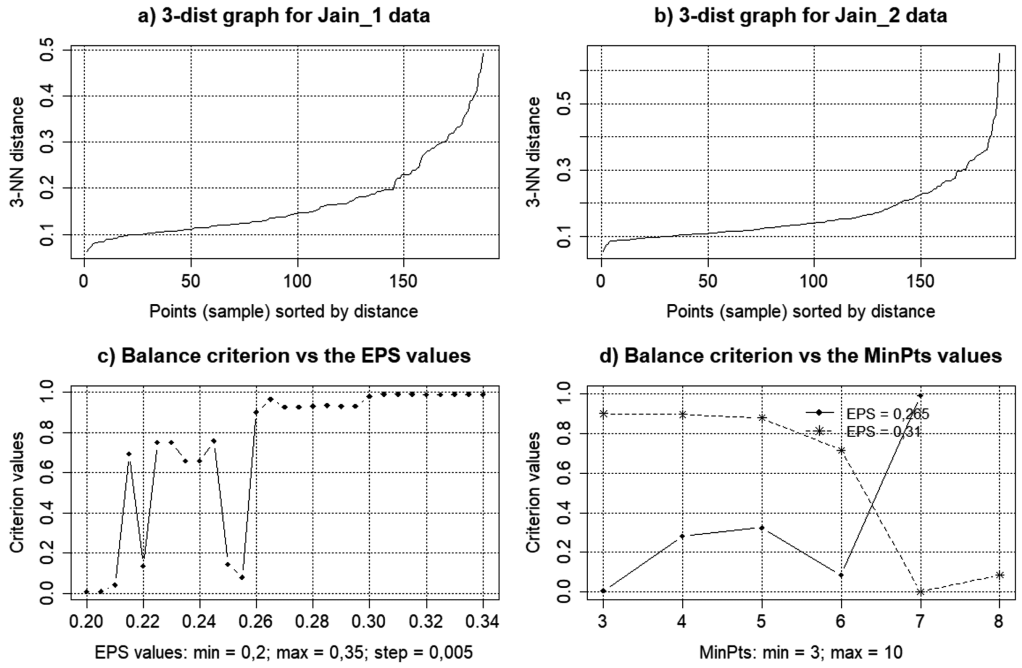

Figure 2.21: Results of the simulation for *Jain* dataset: a) sorted k-dist graph for equal power subset A; b) sorted k-dist graph for equal power subset B; c) chart of the balance criterion vs the EPS value for $MinPts = 3$; d) chart of the balance criterion vs the $MinPts$ for optimal EPS values

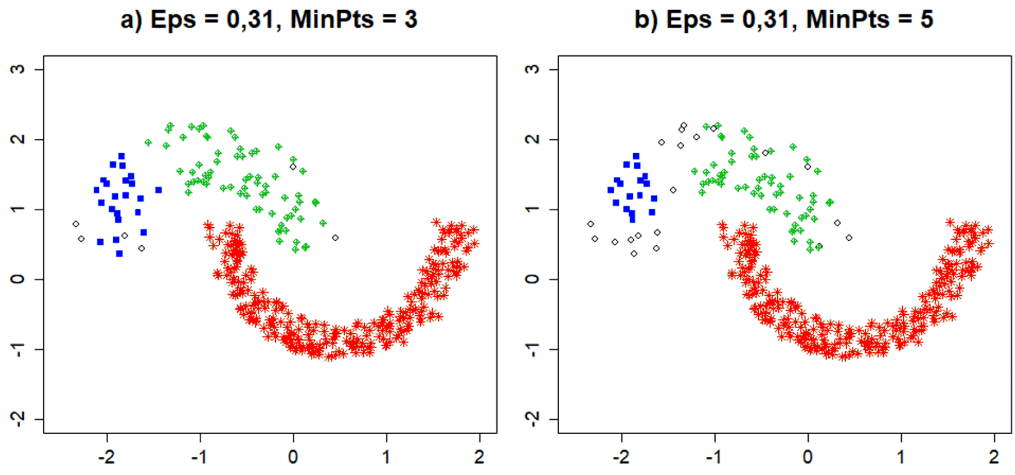

Figure 2.22: Clustering results for *Jain* dataset

step 0.01. The EPS optimal values 0.52 and 0.66 were determined for the following analysis as the result of Figure 2.25c analysis. The simulation results have shown that increasing the EPS value impaired the clustering results since the objects of the *Virginica* and *Versicolor* classes were not distinguish between each other. The following combinations of the algorithm parameters were determined as the result of the Figure 2.25d analysis: a) $EPS = 0.52$, $MinPts = 3$; b) $EPS = 0.66$, $MinPts = 3$. Clustering results for both cases are presented in Figure 2.26. The analysis of the obtained results allows us to conclude that the first combination of parameters leads to unsatisfactory clustering because in this case we get a large percentage of objects that are identified as noise. Moreover, nine of *Setosa* class objects (18%) are identified as objects of *Versicolor* class, what is not correct. In the second case, the clustering results are significantly better. Objects of *Setosa* class are fully identified excepting four objects which were identified as noise. The analysis of the parallel coordinates plot (Figure 2.16) confirms that several objects of *Setosa* class are differ from the other objects of this class. As expected, the objects of *Versicolor* and *Virginica* classes overlap between each other. This fact also confirms the parallel coordinates plot analysis. In this case 54% of objects of *Versicolor* class and 34% of objects of *Virginica* class were identified correctly. 6% of objects of *Versicolor* class and 20% of objects of *Virginica* class were identified as noise. 43% of objects of *Virginica* and *Versicolor* classes were identified no correctly. However, it should be noted that general quantity of objects which were identified as noise in this case was 17 from 150 (11%). This fact justifies the use of the proposed technique to reduce

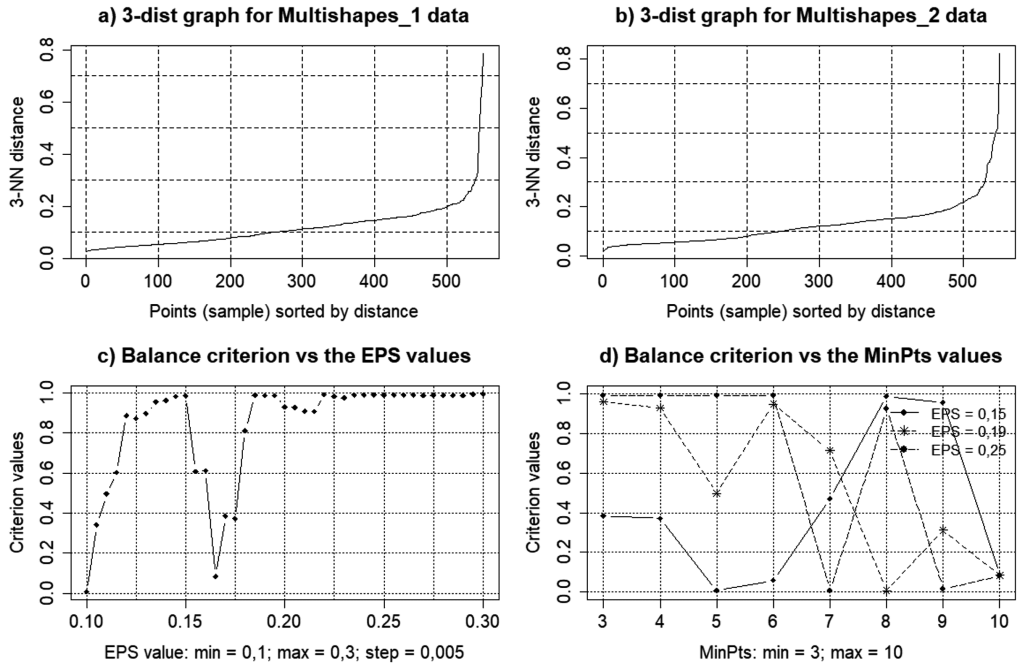

Figure 2.23: Results of the simulation for *Multishapes* dataset: a) sorted k-dist graph for equal power subset A; b) sorted k-dist graph for equal power subset B; c) chart of the balance criterion vs the EPS value for $MinPts = 3$; d) chart of the balance criterion vs the $MinPts$ for optimal EPS values

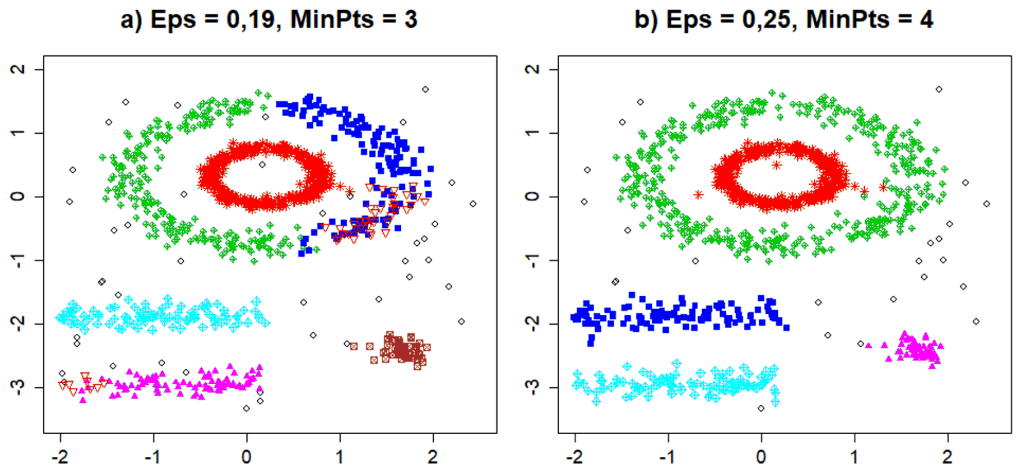

Figure 2.24: Clustering results for *Multishapes* dataset

the feature space of high-dimensional data by removing the noise component.

Two thousand of gene expression profiles of patients which were investigated on lung cancer disease were used as the experimental data in the case of evaluation of effectiveness of the hybrid model of OCIT based on DBSCAN clustering algorithm for high-dimensional gene expression profiles clustering. Initially, the data were divided into two equal power subsets of 1000 profiles in each of them. Then, the triangular distance matrixes were calculated for each of the subsets using correlation metric. These matrixes were used us the input data during the DBSCAN clustering algorithm implementation. Figure 2.27 presents the simulation results. The following EPS values were determines as the results of Figure 2.27a analysis: 0.18, 0.35 and 0.44. At the next step we determined the optimal combinations of the algorithm parameters based on Figure 2.27b analysis. We have obtained three combinations: a) $EPS = 0.18$, $MinPts = 4$; b) $EPS = 0.35$, $MinPts = 5$; c) $EPS = 0.44$, $MinPts = 6$. However, a more detail analysis has shown that in the second case the number of clusters in various clustering is differ. This fact is unsatisfactory in terms of the OCIT. Figure 2.28 shows the results of the gene expression profiles clustering in the case of both the first and third combinations of algorithm parameters use. The analysis of the obtained results allows us to conclude that in the first case (Figure 2.28a) the clustering results are not satisfactory. Large quantity of the objects (1459 of 2000) are identified as noise. The clustering results in the second case (Figure 2.28b) are satisfactory. 1663 of objects from 2000 are allocated in one cluster. 321 of objects are identified as noise. The second cluster contains 16 of

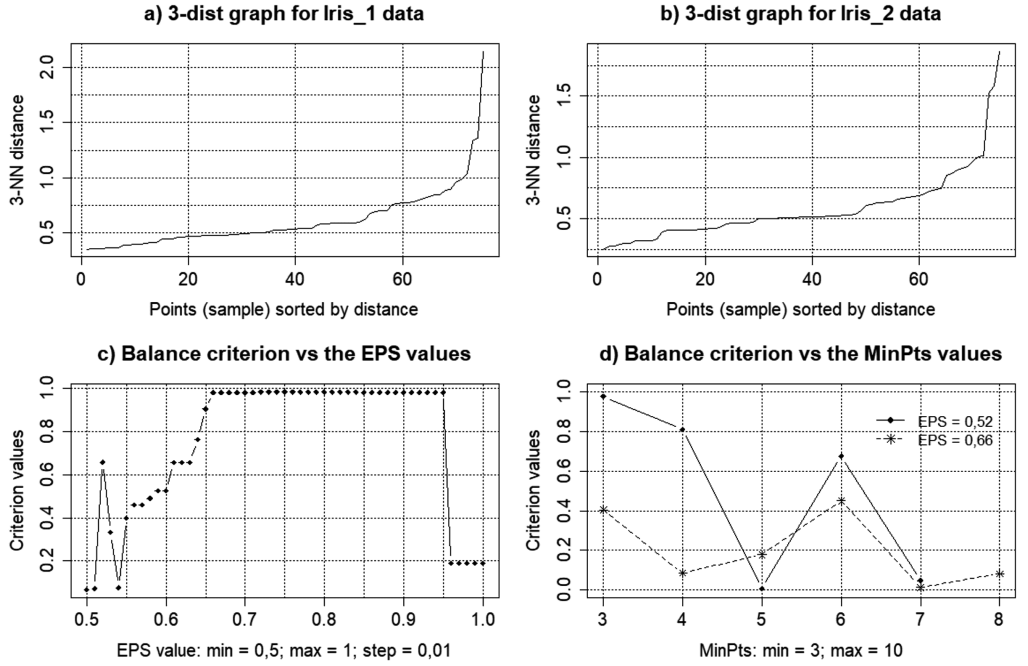

Figure 2.25: Results of the simulation for Fisher's irises dataset: a) sorted k-dist graph for equal power subset A; b) sorted k-dist graph for equal power subset B; c) chart of the balance criterion vs the EPS value for $MinPts = 3$; d) chart of the balance criterion vs the $MinPts$ for optimal EPS values

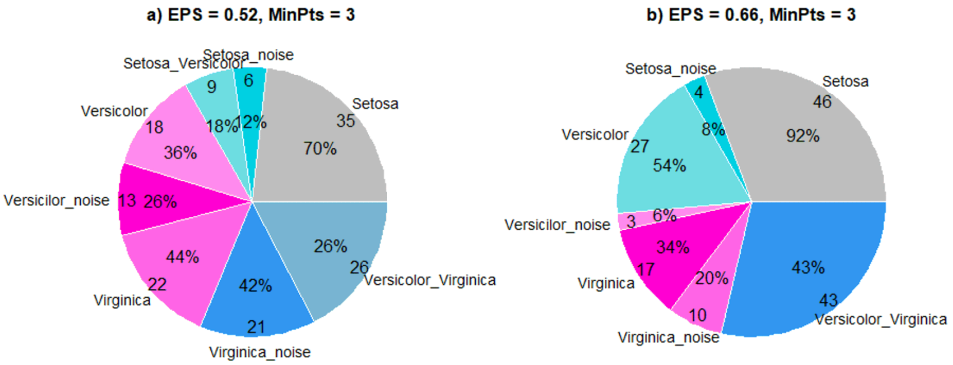

Figure 2.26: Clustering results for Fisher's irises dataset

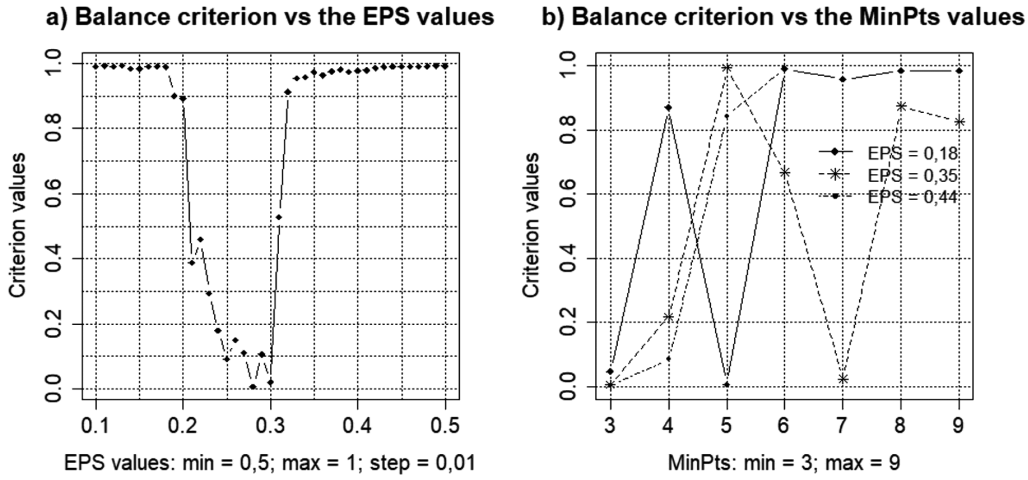

Figure 2.27: Results of the simulation for gene expression profiles of patients which were investigated on lung cancer disease: a) chart of the balance criterion vs the EPS value for $MinPts = 3$; b) chart of the balance criterion vs the $MinPts$ for optimal EPS values

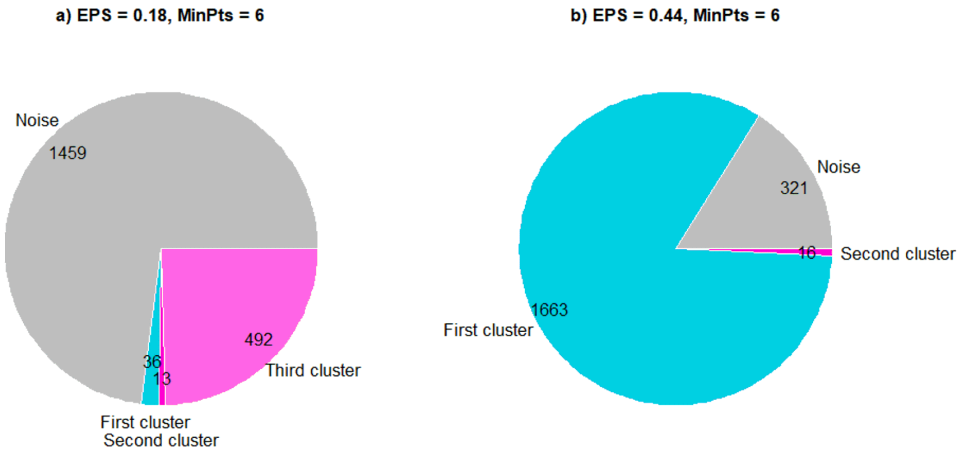

Figure 2.28: Clustering results for gene expression profiles dataset

objects. This distribution is adequate one since the main group of gene expression profiles which are allocated in the first cluster determines the major processes of the biological organism functioning. Noise and a small number of objects of the second cluster can be removed from the data at this stage of the experiment performing.

2.4.3 Results of SOTA Clustering Algorithm Operation

Self-organizing SOTA clustering algorithm in advance is focused to high-dimensional data processing, so, the evaluation of effectiveness of the hybrid model of the objective clustering inductive technology based on SOTA algorithm was performed using Fischer's irises [116], gene expression profiles of *moe430A* [32] and dataset of gene expression profiles of patients, which were investigated on lung cancer disease [30]. As was noted hereinbefore in section 2.3.2, the result of the SOTA clustering algorithm operation is determined by two main parameters: the value of the weight coefficient of the sister's cell (*scell*), and the variation coefficient value (E). Figure 2.29a shows a chart of the balance clustering quality criterion versus the value of the weight coefficient of the sister's cell *scell* for the Fischer's irises data. The value of the variation coefficient E was taken as 0.1 in this case. The *scell* value increased within the range from 0.001 to 0.2 with step 0.002. Two *scell* values were selected for the following processing: 0.001 and 0.013. The simulation results have shown that the use of larger *scell* values impairs the obtained results. Figure 2.29b shows the charts of the balance clustering quality criterion versus the maximum value of the variation coefficient E for selected *scell* values. Two combination of the algorithm

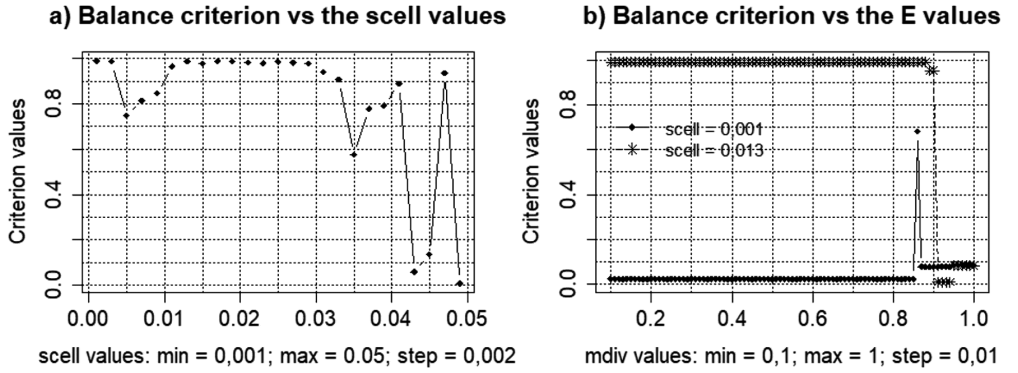

Figure 2.29: Results of the simulation for Fisher's irises dataset: a) chart of the balance criterion vs the *scell* value; b) chart of the balance criterion vs the variation coefficient E for optimal *scell* values

parameters were determined as the simulation results: a) $scell = 0.001$, $E = 0.86$; b) $scell = 0.013$, $E = 0.88$. Figure Figure 2.30 presents the results of the Fisher's irises data clustering when the first algorithm parameters combination was used.

As it can see from Figure 2.30, the algorithm divided the objects into four clusters adequately. In this case, the setosa class objects are contained in the third and fourth clusters, which is quite justified since the profiles of the features of these objects are differed from each other. Objects of both versicolor and virginica classes are contained in the first and in the second clusters. The profiles of these objects are similar, so the first and second clusters have some intersection between each other. In the case of the use of the second algorithm parameters combination the data were divided into six clusters. This results is unsatisfactory for this type of data.

Figure 2.31 presents the simulation results for gene expression profiles from dataset *moe*430*a* in the case of the use of both 147 of genes (Figure 2.31a) and 1000 of genes (Figure 2.31b). The simulation results have shown that changing the variation coefficient value does not affect the obtained results. Thus, the variation coefficient value in this case was equal zero. The *scellvalues* 0.001, 0.016, 0.071, 0.156 were selected as a result of the Figure 2.31a analysis. These values correspond to the maxima of the balance clustering quality criterion. Figure 2.32 shows the clustering results for each of the cases. As it can be seen, in each of the cases the data were divided into two clusters. In addition, increasing the *scell* value is not reasonable, since increasing this parameter value leads to decreasing difference between average values of gene expression profiles in different clusters. This fact indicates a lower algorithm resolution. Figure 2.33 shows the same results in the

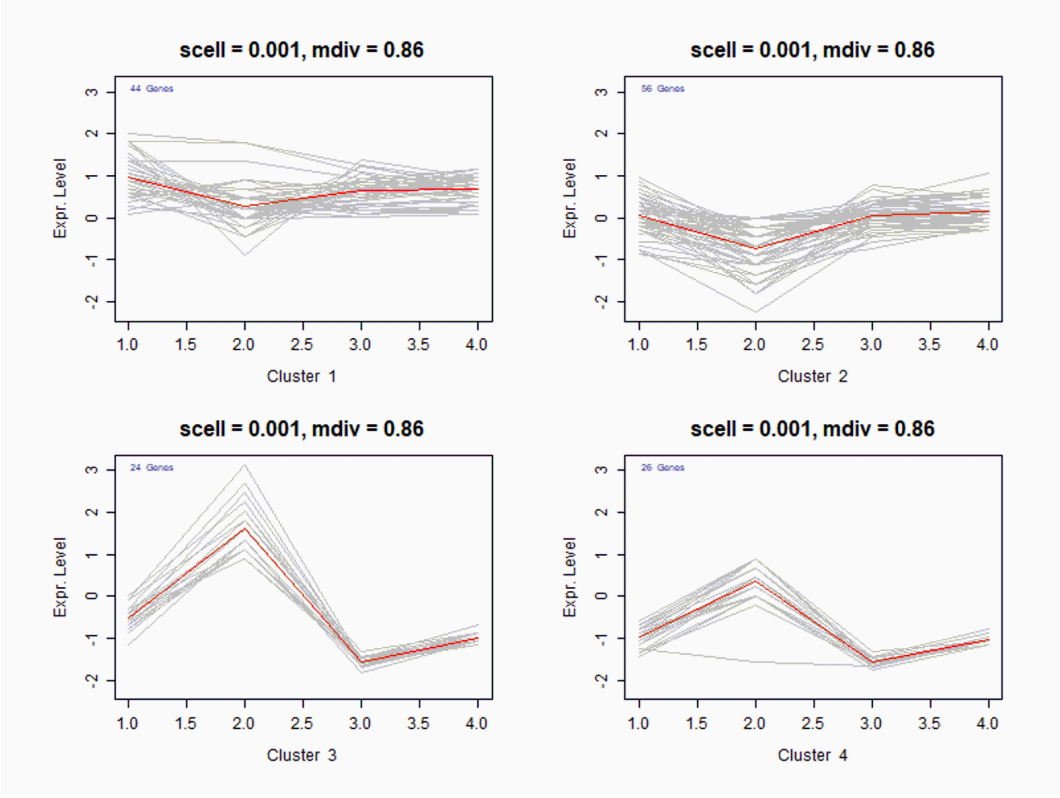

Figure 2.30: Results of Fisher's irises dataset clustering when the first algorithm parameters combination was used: $scell = 0.001$, $E = 0.86$

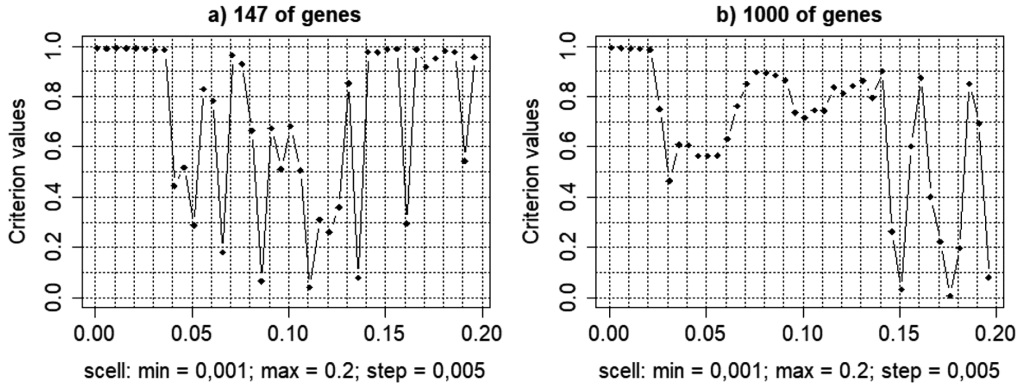

Figure 2.31: Results of the simulation for gene expression profiles from dataset *moe*430*a* in the case of the use: a) 147 of genes; b) 1000 of genes

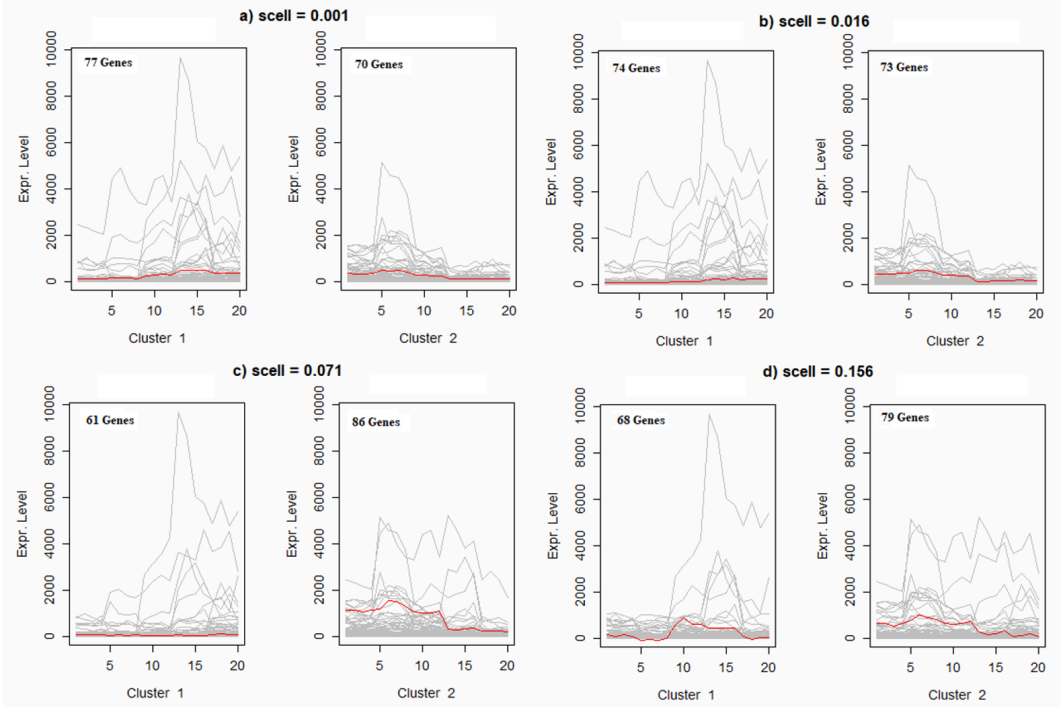

Figure 2.32: Clustering Results of 147 of gene expression profiles from *moe*430*a* dataset

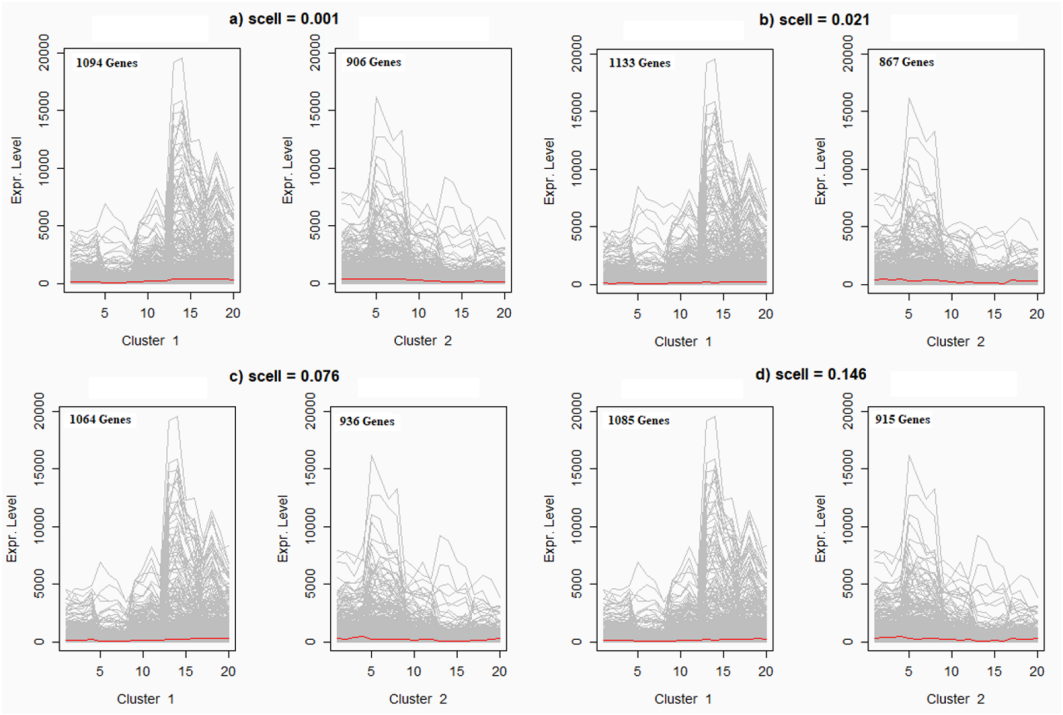

Figure 2.33: Clustering Results of 1000 of gene expression profiles from *moe*430*a* dataset

case of 1000 of gene expression profiles use. The following *scell* values were selected as the result of Figure 2.31b analysis: 0.001,0.021, 0.076, 0.146. As it can be seen, the data in this case also were divided into two clusters and minimal value of the *scell* parameter corresponds to better resolution of the algorithm.

Figure 2.34 presents the chart of the balance clustering quality criterion versus the *scell* value for gene expression profiles of patients which were examined on lung cancer disease. 2000 of gene expression profiles were used during the experiment performing. As it can be seen from Figure 2.34, the *scell* value 0.001 corresponds to maximum value of the balance clustering quality criterion. Considering the results obtained for gene expression profiles of *moe*430*a* dataset, other of the *scell* values did not examine. The clustering results are presented in Figure 2.35. The analysis of the obtained results allows concluding that the use of a hybrid model of objective clustering inductive technology based on SOTA algorithm allows us to divide the gene expression profiles into two almost equal groups. It is obvious that genes in each of the groups perform appropriate functions in the biological organism. Further data processing can be performed on separated clusters, which simplifies the process

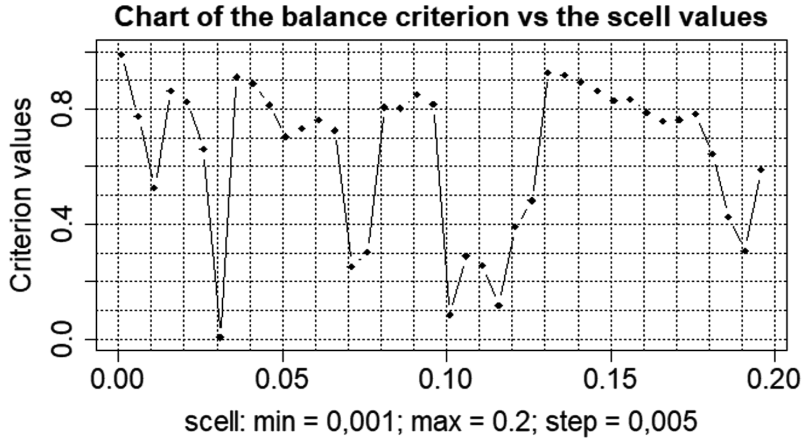

Figure 2.34: Results of the simulation for gene expression profiles of patients which were examined on lung cancer disease

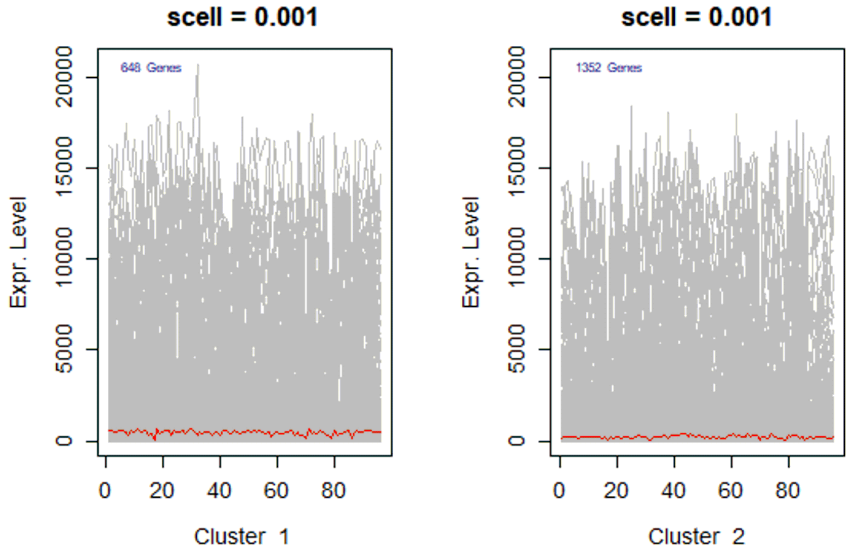

Figure 2.35: Clustering Results of gene expression profiles of patients which were investigated on lung cancer disease

of the gene expression profiles processing in accordance with aim of the current task.

2.5 An Evaluation of the OCIT Robustness to a Level of Noise Component

In this section, we present the results of the research concerning evaluation of the objective clustering inductive technology robustness to a level of noise component [8]. The hybrid model of OCIT based on SOTA clustering algorithm was used as the experimental during the simulation process. Gene expression profiles of 2000 patients who were examined on lung cancer [30] were used in this case. The length of the studied vectors was equal to the number of the studied samples (96). The simulation process involved the following steps:

1. Generation of random values vector. The length of this vector is equal to the length of the studied gene expression profiles and its amplitude corresponds to the minimum value of the studied data genes expression ("white noise").

2. Setup of the vector of coefficients to change the amplitude of the noise component. In the case of the studied gene expression profiles the values of coefficients were changed within the range from 0.2 to 4 with step 0.2. These parameters were determined empirically during the simulation process.

3. Formation of gene expression profiles with the noise by adding of the appropriate noise component to the studied gene expression profiles.

4. Division of the obtained data into two equal power subsets.

5. Gene expression profiles clustering with the use of objective clustering inductive technology using SOTA clustering algorithm. The value of the sister cell weigh coefficient (*scell*) was changed within the small range from 8×10^{-4} to 11×10^{-4} with the step 2×10^{-5}. This range was determined empirically during the previous simulation process. The value of the variation coefficient was taken as zero.

6. Calculation of the complex balance criterion (general Harrington desirability index) for each value of the sister cell weigh coefficient. Creation of the plots of complex balance criterion versus the weigh coefficient value for both the data without noise and the data with different levels of noise component. Determination of the SOTA clustering algorithm optimal parameters, which correspond to the maximum value of the complex balance criterion. Data clustering with the use of SOTA algorithm with its optimal parameters.

7. Calculation of the external clustering quality criteria, which allows us to compare the clustering results for both the data without noise and the data with noise component. The following criteria were used as the external clustering quality criteria in this case:

   - Jaccard index:
   $$J = \frac{a}{a+b+c}. \tag{2.7}$$
   - Kulczynski index:
   $$K = \frac{a}{2 \times (a+b)} + \frac{a}{2 \times (a+c)}, \tag{2.8}$$

   where a is the number of objects distributed in the same clusters in different clustering; b is the number of objects in the clusters of the first clustering, which did not coincide with the appropriate objects in the clusters of the second clustering; c is the number of objects in the clusters of the second clustering, which did not coincide with the appropriate objects in the clusters of the first clustering.

8. Analysis of the obtained results.

Figure 2.36 presents the charts of the complex balance criterion versus the sister's cell weigh coefficient (*scell*) of SOTA clustering algorithm, which was implemented within the framework of the objective clustering inductive technology. The noised gene expression profiles of the patients who were examined on lung cancer disease were used in this case. The optimal value of the *scell*, which corresponds to the maximum value of balance clustering quality criterion was determined during the simulation process. The results of the simulation have shown that the increase of the amplitude coefficient of the noise component from 0.2 to 3.2 does not significantly influence to the character of the balance criterion change. Figure 2.37 shows the charts of the number of objects in the clusters,the values of Jaccard and Kulczynski indexes and the relative changes of these indexes in percentage versus the amplitude coefficient of the noise component. The analysis of the obtained charts allows us to conclude that the character of the objects distribution within the clusters is changed slightly during the increase of the noise amplitude coefficient. It is naturally, since the existence and the increase of the amplitude of the noise component in the studied data changes the gene expression profiles. In this case, the movement of the object between clusters is possible. The values of Jaccard and Kulczynski indexes decrease monotonically in this cases, but the speed of these indexes changes chaotically in the defined range. This character of these parameters change is observed to value of the amplitude coefficient of noise 3.2. The charts of the appropriate parameters are changed significantly in the case of larger value of the noise amplitude. The *scell*

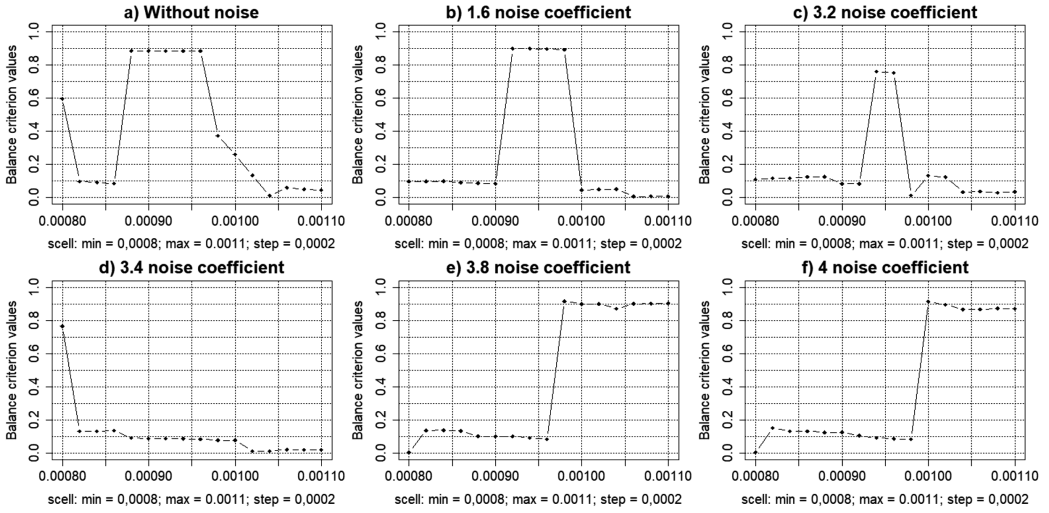

Figure 2.36: Charts of the complex balance criterion versus the sister's cell weigh coefficient (*scell*) for the gene expression profiles with the different levels of noise component

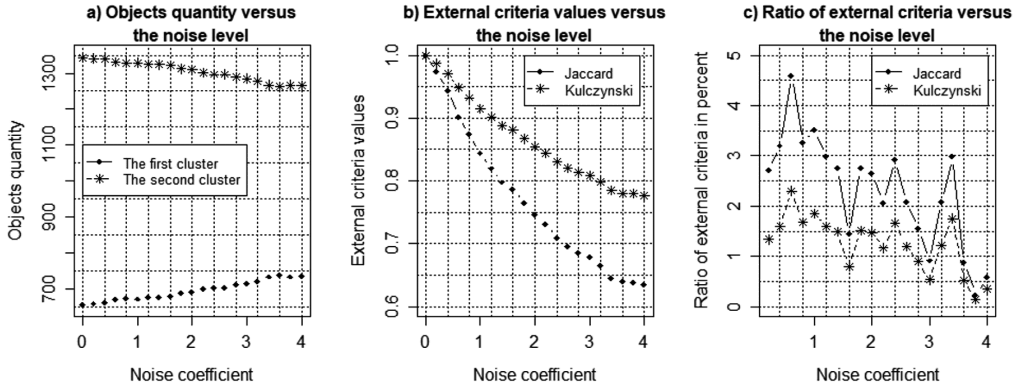

Figure 2.37: Charts of: *a*) the quantity of gene expression profiles in different clusters; *b*) Jaccard and Kulczynski indexes values; *c*) the relative changes of Jaccard and Kulczynski indexes versus the amplitude coefficient of noise component

optimal value of SOTA clustering algorithm, which corresponds to the maximum value of the complex balance criterion is changed chaotically too. This fact indicates the non-stability of the system. The number of the objects in the clusters and the values of Jaccard and Kulczynski indexes in the case of large values of the noise amplitude coefficient are changed very slowly.

As it can be seen in Figure 2.37c, the speed of these parameters changes in this case tends to zero. This fact can be explained in the following way. In the case of high level of the noise component, local particularities of the gene expression profiles become smoother and clustering in this case is carried out by the estimation of the coarse component of the appropriate vector. Therefore, the *scell* value in this case is not determinative. The results of the simulation have shown that the clustering results in the case of the high level of noise component are almost the same and they do not depend on the *scell* value. The conducted research has shown also that the objective clustering inductive technology is effective and efficient in the case of the analysis of the complex data with the local particularities. The use of this technology to group the gene expression profiles is reasonable in the case of low level of the noise component.

2.6 Conclusions

In this chapter, the objective clustering inductive technology is presented as a definite analogue of the inductive technology of complex systems analysis, which has allowed us to formulate this study as a task of complex data objective clustering. Three fundamental principles borrowed from different scientific fields are the basis of the methodology of the complex systems inductive modeling: the principle of heuristic self-organization; the principle of external addition; the principle of inconclusive of solution. Implementation of these principles in the adapted version provides the conditions to create the methodology of objective clustering inductive technology.

The affinity measures of high-dimensional vectors characterizing the investigated data have been considered. Comparative analysis of the proximity metrics of the gene expression profiles has been performed. It has been shown that the use of the correlation metric for high dimensional data analysis allows us to obtain a higher resolution in comparison with Euclidean and Manhattan metrics. A comparative analysis of the internal clustering quality criteria has been performed using synthetic separated equal power subsets. The investigated criteria considered both the character of the objects distribution relative to the mass centers of the clusters were these objects are located, and the character of the cluster centers distribution in the feature space. The results of the simulation using synthetic datasets of gene expression profiles have shown that Calinski-Harabasz criterion and WB index have the greatest effectiveness in selecting the best clustering. A complex multiplicative

internal clustering quality criterion has been proposed as the simulation result. This criterion is calculated as multiplicative combination of Calinski-Harabasz criterion and WB index and it allows us to obtain a higher resolution in comparison with other internal clustering quality criteria. The external clustering quality criterion was calculated within the framework of the objective clustering inductive technology as normalized difference of the internal criteria calculated on equal power subsets. A technique for calculating the complex balance clustering quality criterion based on the Harrington desirability function has been developed. This technique is presented as a stepwise algorithm for determining the vector of the generalized Harrington desirability index for the obtained clusterings. Maximum value of this criterion corresponded the best clustering in terms of the used criteria.

The architecture of the objective clustering inductive technology has been proposed. This model is presented as a step-by-step detailed scheme of the practical implementation of the procedure for objective selecting the optimal clustering based on inductive methods of complex systems analysis.

A comparative analysis of the obtained hybrid models has been done using two-dimensional synthetic data from School of Computing of Eastern Finland University, Fischer's irises, and gene expression profiles. The optimal parameters of the corresponding clustering algorithm have been determined during the simulation process. The applying of the proposed technique has allowed us to obtain the optimal grouping of the investigated objects. It has been shown that in the case of the gene expression profiles analysis, the use of DBSCAN clustering algorithm within the framework of the OCIT allows us to distinguish the gene expression profiles which have the lowest density distribution in feature space. These genes have been identified as noise and they can be removed from the investigated data as non-informative ones. SOTA clustering algorithm allows us to divide the selected gene expression profiles into two approximately equal subsets.

An evaluation of the robustness of the objective clustering inductive technology to the level of the noise component has been done with the use of SOTA clustering algorithm. The gene expression profiles with various level of noise component of patients which were investigated on lung cancer disease were used as the experimental data. Jaccard index and Kulchinsky coefficient have been used as the external independent quality criteria for evaluate the clustering results during grouping gene expression profiles within the framework of the OCIT. The analysis of the obtained results has shown high robustness of the proposed OCIT to the noised data. The proposed technique allows us to determine the optimal clustering algorithm parameters. The further applying of this algorithm with optimal parameters allows dividing the investigated objects into clusters adequately.

Chapter 3

Biclustering Techniques

3.1 Introduction

Bicluster analysis is a technique focused to allocate from high-dimensional data array mutual correlated rows and columns [81]. For example, in the case of gene expression profiles analysis, the rows are the studied genes and the columns are the conditions of the experiment performing. Selection of groups of mutually correlated genes and conditions from microarray allows us to reconstruct the gene network, which will be able to reflect objectively the influence of the appropriate genes to functional possibilities of the studied biological object.

Bicluster analysis of gene expression profiles has been considered in [108, 114]. The authors analysed various biclustering algorithms and specified their advantages and shortcomings. The paper [49] presents a comparative analysis of various biclustering algorithms effectiveness. The gene expression profiles were used as the experimental data in this case. In [89] the authors presented the results of the research concerning spectral biclustering algorithm use to allocate biclusters from gene expression profiles. In [93] the authors considered questions concerning bicluster analysis implementation for data which contain missing values. However, it should be noted, that in spite of the achieved results, technique of biclustering algorithms optimal parameters determination based on qualitative criteria is absent nowadays.

This chapter presents the results of the research concerning comparison analysis of biclustering algorithms effectiveness based on the quantitative biclustering quality criteria with the use of both the synthetic data and gene expression profiles [17].

3.2 Basic Concepts of Bicluster Analysis and Biclustering Quality Criteria

As was noted hereinbefore, bicluster is a set of mutually correlated rows and columns. Let the initial data is presented as a matrix:

$$A = \{a_{ij}\},\ i = \overline{1,n},\ j = \overline{1,m}, \tag{3.1}$$

where n is the number of rows; m is the number of columns; a_{ij} is the value in j-th column for i-th row. By analogy with (3.1), the matrix of data in biclusters takes the form:

$$B = \{a_{ij}\},\ i = \overline{1,k},\ j = \overline{1,s}, \tag{3.2}$$

where k and s are the numbers of rows and columns in biclusters respectively.

In [103] the authors analysed the types of various biclusters considering their structure. Assuming that δ is a some constant for bicluster B, then, depending on the character of the data distribution in the initial matrix, the following types of the biclusters can be allocated:

1. Bicluster with constant values:

$$a_{ij} = \delta$$

2. Bicluster with constant values of rows or columns:

$$a_{ij} = \delta + a_i \ or \ a_{ij} = \delta \times a_i$$
$$a_{ij} = \delta + a_j \ or \ a_{ij} = \delta \times a_j$$

3. Bicluster with coherent values:

$$a_{ij} = \delta + a_i + a_j \ or \ a_{ij} = \delta \times a_i \times a_j$$

4. Bicluster with coherent evolution:

$$a_{ih} \leq a_{ir} \leq a_{id} \ or \ a_{hj} \leq a_{tj} \leq a_{dj}$$

However, it should be noted that in most cases real data contains intersected biclusters, since the character of rows data distribution for different columns can be very complex and various biclusters can include the same rows or columns.

The following biclustering quality criteria were used during the simulation process:

- *Jaccard index* was used as the external biclustering quality criterion in the case of perfect biclustering existence. For two biclusters BC_1 and BC_2 formula to calculate Jaccard index is the following:

$$JI(BC_1, BC_2) = \frac{|BC_1 \cap BC_2|}{|BC_1 \cup BC_2|}$$

In the case of more biclusters, any biclustering algorithm allows us to obtain a result matrix which contains information concerning appropriate biclusters. In this case, the Jaccard index is calculated by the formula:

$$\begin{gathered} JI(BC\_Res_1, BC\_Res_2) = \\ = \frac{1}{n_1} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \frac{|BC_i(BC\_Res_i) \cap BC_j(BC\_Res_j)|}{|BC_i(BC\_Res_i) \cup BC_j(BC\_Res_j)|} \end{gathered} \tag{3.3}$$

where n_1 and n_2 are the number of biclusters in various biclustering; $BC_i(BC\_Res_i)$ and $BC_j(BC\_Res_j)$ are the biclustering results in i-th and j-th biclusterings respectively.

If there are intersection between biclusters, then, the formula (3.3) is transformed in the following way:

$$\begin{gathered} JI_{cor}(BC\_Res_1, BC\_Res_2) = \\ = \frac{JI(BC\_Res_1, BC\_Res_2)}{max(JI(BC\_Res_1, BC\_Res_1), JI(BC\_Res_2, BC\_Res_2))} \end{gathered} \tag{3.4}$$

It should be noted that the formula (3.4) is transformed into the formula (3.3) in the case of absence of any intersection between biclusters (the denominator in the formula (3.4) is equal one).

- *The internal biclustering quality criterion.* This criterion is calculated as follows:

  1. Calculation of average of Euclidean distance between all rows in the appropriate bicluster:

$$QC_1 = \frac{2}{nr \times (nr-1)} \sum_{i=1}^{nr-1} \sum_{j=i+1}^{nc} (\frac{1}{nc} \sum_{k=1}^{nc} (x[i,k] - x[j,k])^2)$$

     where nr and nc are the numbers of rows and columns in bicluster respectively.

  2. Calculation of average of Euclidean distance QC_2 between all columns in the appropriate bicluster.

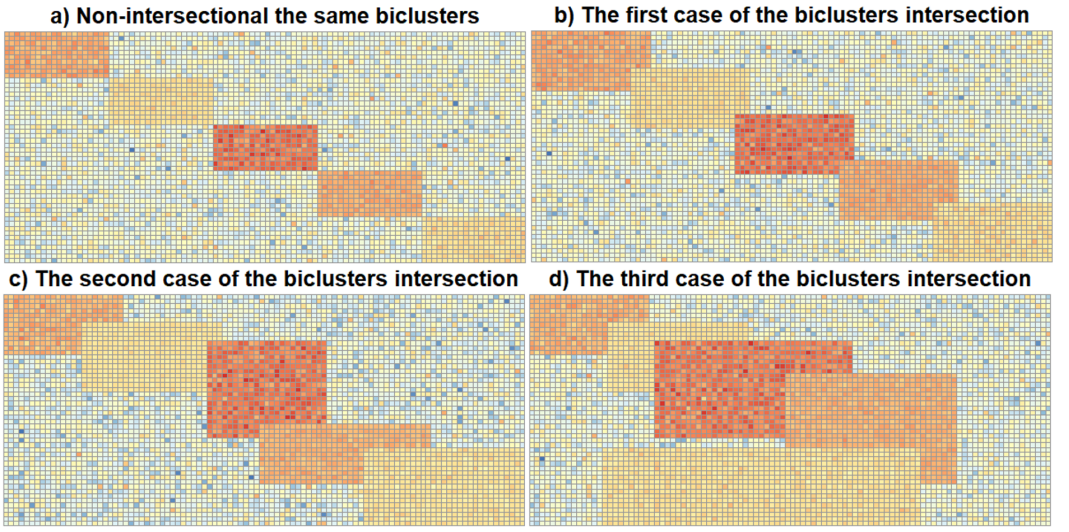

Figure 3.1: Synthetic biclusters

3. Calculation of average value of QC_1 and QC_2 criteria.
4. Calculation of average value of the obtained criterion for all biclusters.

It is obvious, that minimal value of the internal biclustering quality criterion corresponds to the best biclustering.

3.3 Bicluster Analysis With the Use of Synthetic Biclusters

3.3.1 Experimental Synthetic Data

Figure 3.1 shows the synthetic data which were used during the simulation process. Data matrix of random values (50×100) contains five biclusters. In the first case (Figure 3.1a), the same biclusters (10×20) have no intersections, in the second case (Figure 3.1b), the lager biclusters (13×23) have some intersection between each other, in the third and in the fourth cases (Figure 3.1c,d), the biclusters differ and they have various level of mutual intersection.

3.3.2 Bicluster Analysis with the use of Non-intersectional Synthetic Biclusters

Evaluaton of effectiveness of the hereinbefore presented biclustering quality criteria was performed with the use of *biclust* package [82] of R software [75]. Biclustering

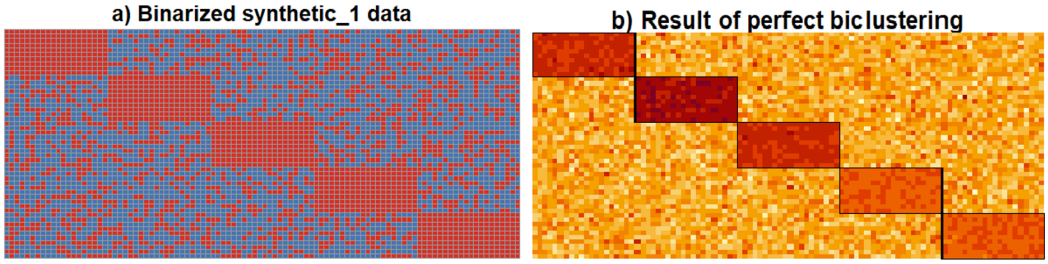

Figure 3.2: Results of *bimax* biclustering algorithm operation: a) binarized data; b) perfect biclustering

algorithms *bimax* [115], *CC* [42], *spectral* [89] and *ensemble* [81] were used during the simulation process. Firstly, the perfect biclustering was obtained using *bimax* algorithm. The implementation of this algorithm assumes binarization of the studied data as follows:

$$x_{ij} = \begin{cases} 0, \ if \ x_{ij} \leq median(X) \\ 1, \ if \ x_{ij} > median(X) \end{cases}$$

Binarization result is shown in Figure 3.2a. To obtain perfect biclustering, a minimum number of rows and columns in biclusters (10 and 20) were determined on a binarized data. Then, a set of biclusters which corresponds to perfect biclustering was obtained (Figure 3.2b). The perfect biclastering was used to calculate Jaccard index. Higher value of Jaccard index corresponds to better biclustering.

Figure 3.3 shows the charts of both Jaccard index and internal biclustering quality criterion versus the minimum number of rows in biclusters in the case of the use of synthetic data with non-intersectional biclusters (Figure 3.1a). Minimum number of columns was twice bigger than minimum number of rows. The analysis of the charts allows us to conclude that value of Jaccard index achieves maximum (JI = 1) in the case of 9 or 10 of rows quantity in biclusters. It means that the obtained and the perfect biclustering are the same. The internal biclustering quality criterion has minimal values in the case. This fact means that both criteria in this case allow us to determine the optimal parameters of the *Bimax* biclustering algorithm. The main disadvantage of this algorithm is that it implementation requires the binarization of the data. In this case, a lot of useful information is lost.

Figure 3.4 shows the results of *CC* biclustering algorithm operation. Delta (δ) is the main parameter, which determines the result of the algorithm operation. Value of the delta parameter was changed within the range from 0.2 to 0.5 by step 0.005. Analysis of the obtained results shows low quality of this algorithm operation, since Jaccard index has maximum value 0.2 if delta parameter value is 0.3. Low value of Jaccard index and high level of these values fluctuation indicate

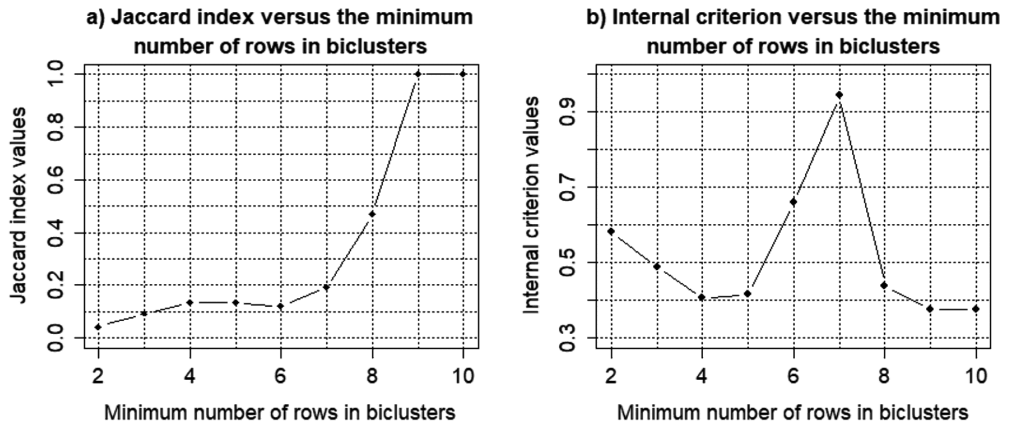

Figure 3.3: Charts of Jaccard index (a) and internal biclustering quality criterion (b) versus the minimum number of rows in biclusters in the case of *bimax* biclustering algorithm applying

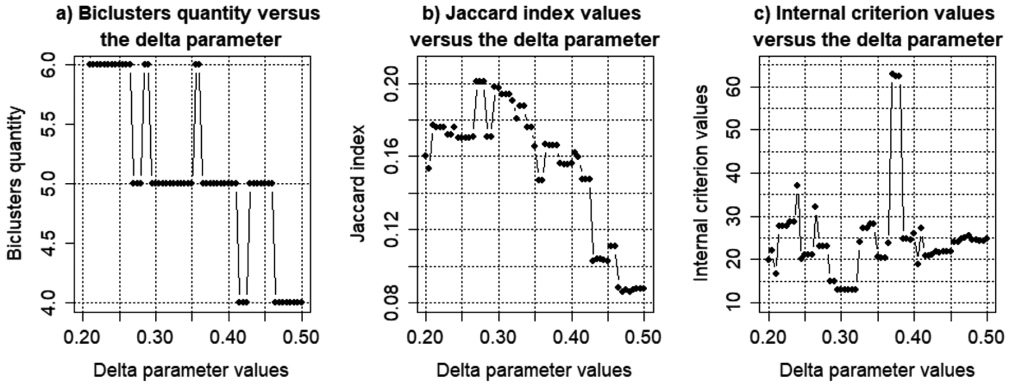

Figure 3.4: Charts of biclusters quantity (a), Jaccard index (b) and internal biclustering quality criterion (c) versus the delta parameter in the case of *CC* biclustering algorithm use

Table 3.1: Results of *CC* biclustering algorithm operation (Distributions of rows and columns in the biclusters in the case of the use of 0.3 delta parameter value)

Bicluster	Perfect	1	2	3	4	5
Rows	10	13	11	12	7	6
Columns	20	22	27	10	18	14

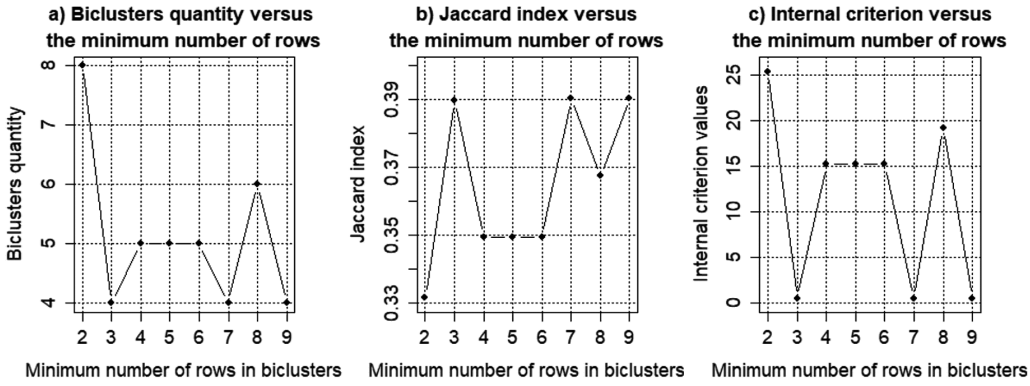

Figure 3.5: Charts of biclusters quantity (a), Jaccard index (b) and internal biclustering quality criterion (c) versus the delta parameter in the case of *spectral* biclustering algorithm use

a big difference between perfect and obtained biclustering. However, it should be noted, that minimum values of the internal biclustering quality criterion correspond to maximum values of Jaccard index. In this case these criteria allow us to select the best biclustering from admissible ones. Distributions of rows and columns in the biclusters in the case of the use of 0.3 delta parameter value are presented in Table 3.1.

The simulation results in the case of *spectral* biclustering algorithm application are shown in Figure 3.5. As it can be seen, in this case the results of biclustering are not perfect. However the values of Jaccard index and the internal biclustering quality criterion change concordantly to each other too. The analysis of the obtained charts has shown also that the optimal biclustering corresponding to both the maximal values of the Jaccard index or the minimal values of the internal biclustering quality criterion are achieved in the cases of the number of rows in the biclusters 3, 7 and 9. However, more detailed analysis has shown that in the case of the use of the minimal number of rows of 3, the Jaccard index value is slightly smaller and the internal criterion value slightly larger than the corresponding values obtained in the case of the use of the minimal number of rows 7 and 9. The optimal clustering in

Table 3.2: Results of *spectral* biclustering algorithm operation (Distributions of rows and columns in the biclusters in the case of minimal number of rows 7(9))

Bicluster	Perfect	1	2	3	4	5
Rows	10	10	11	10	10	-
Columns	20	21	27	20	21	-

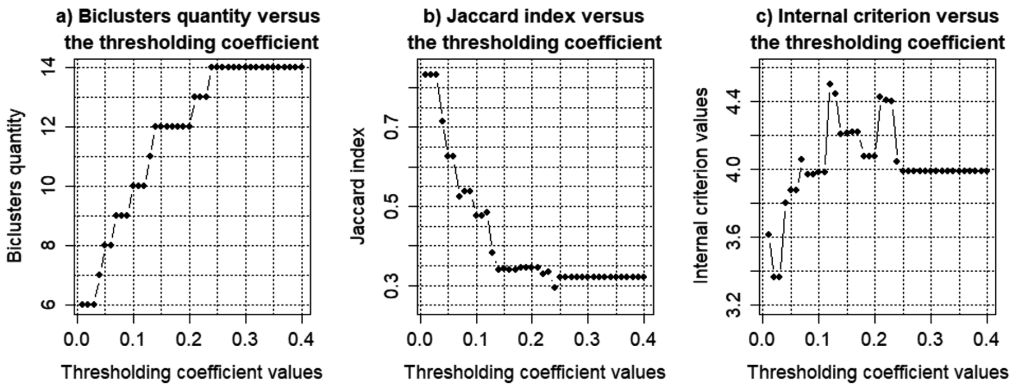

Figure 3.6: Charts of biclusters quantity (a), Jaccard index (b) and internal biclustering quality criterion (c) versus the thresholding coefficient value in the case of *ensemble* biclustering algorithm use

this case corresponds to the four biclusters. The fifth bicluster was not identified. This fact explains the difference of the appropriate criteria values. The results of optimal biclustering obtained using the *spectral* biclustering algorithm are shown in Table 3.2.

Figure 3.6 and Figure 3.7 present the results of the simulation concerning application of *ensemble* biclustering algorithm. The result of the algorithm operation depends on two parameters: thresholding coefficient values and ratio of rows and columns quantity in biclusters. The charts of biclusters quantity, Jaccard index and internal biclustering quality criterion versus the thresholding coefficient value in the case of 0.5 ratio of rows and columns quantity are shown in Figure 3.6. The thresholding coefficient value was changed in this case within the range from 0.01 to 0.4 with step 0.01. Figure 3.7 shows the charts of Jaccard index and the internal biclustering quality criterion versus the ratio of rows and columns quantity in the case of 0.03 value of the thresholding coefficient (this value corresponds to both the maximal value of Jaccard index and minimal value of the internal criterion). At the second step, the ratio of rows and columns quantity was changed within the range from 0.1 to 0.8 with step 0.1.

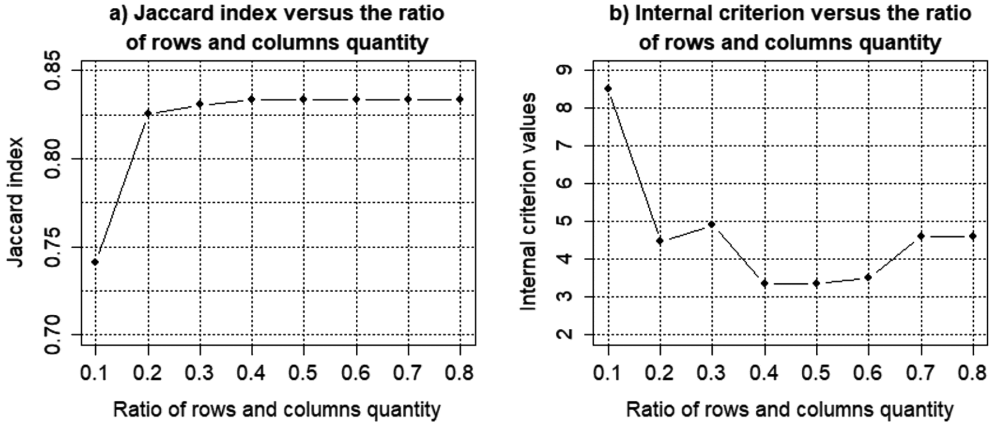

Figure 3.7: Charts of Jaccard index (a) and internal biclustering quality criterion (b) versus the ratio of rows and columns quantity in biclusters in the case of the use of *ensemble* biclustering algorithm

The analysis of the obtained results shows high stability and effectiveness of *ensemble* biclustering algorithm operation. The character of the data distribution in the biclusters is not changed during variation of the thresholding coefficient from 0.01 to 0.32. Six biclusters were obtained in this case. Five biclusters were the same as in perfect biclustering were. The sixth bicluster contained 6 of rows and six of columns. The value of Jaccard index was 0.83, what indicates to high level of proximity between obtained and perfect biclustering. The value of the internal biclustering quality criterion in this case also achieved its minimal one. The analysis of the charts in Figure 3.7 allows us also to conclude that the internal biclustering quality criterion is more sensitive to variation of the ratio coefficient in comparison with Jaccard index. Both criteria have extrema, which correspond to 0.4 value of ratio of the rows and columns quantity in the biclusters. However, the extremum value of the internal biclustering quality criterion is more accurate to compare with the use of Jaccard index.

The results of the simulation concerning bicluster analysis with the use of synthetic data contained the same non-intersectional biclusters have shown that *ensemble* biclustering algorithm is more effective in comparison with other biclustering algorithms. Application of this algorithm allows us to obtain adequate biclustering relative to perfect one. Moreover, the simulation results have also shown high effectiveness of the proposed internal biclustering quality criterion, since the value of this criterion has a local minimum which corresponds to the local maximum of Jaccard index. This fact indicates a high level of similarity between the obtained and perfect

biclustering, but the internal criterion does not require the perfect biclustering.

3.3.3 Technique of Bicluster Analysis Based on *Ensemble* Biclustering Algorithm

The obtained results allow us to propose the technique of bicluster analysis based on *ensemble* biclustering algorithm. Implementation of this technique allows determining the optimal parameters of *ensemble* algorithm which correspond to extrema values of the internal biclustering quality criterion. Structure block-chart of algorithm to implement the proposed technique is shown in Figure 3.8. Practical implementation of the algorithm assumes the following steps:

1. Preparation of data.
   (a) Data preprocessing. Data formation as a matrix $A = \{x_{i,j}\}$, where $i = \overline{1,n}$ and $j = \overline{1,m}$ are the the numbers of rows and columns respectively.
2. Determination of the thresholding coefficient optimal value.
   (a) Fixation of *simthr* parameter value, which determines the ratio of rows and columns quantity in biclusters. Setup of both the range and step of the thresholding coefficient value variation.
   (b) Data biclustering within the range of the thresholding coefficient value change. Biclusters fixation and calculation of the internal biclustering quality criterion at each step of this procedure implementation.
   (c) Analysis of the obtained results, fixation of the thresholding coefficient value, which corresponds to the minimal value of the internal biclustering quality criterion.
3. Determination of optimal value of ratio of rows and columns quantity in biclusters.
   (a) Setup of both the range and step of ratio of rows and columns quantity variation.
   (b) Data biclustering within the range of this parameter change. Fixation of the biclusters, calculation of the internal biclustering quality criterion at each step of this procedure implementation.
   (c) Analysis of the obtained results, fixation of the ratio of rows and columns quantity in biclusters, which corresponds to the minimal value of the internal biclustering quality criterion.
4. Fixation of the optimal biclustering.

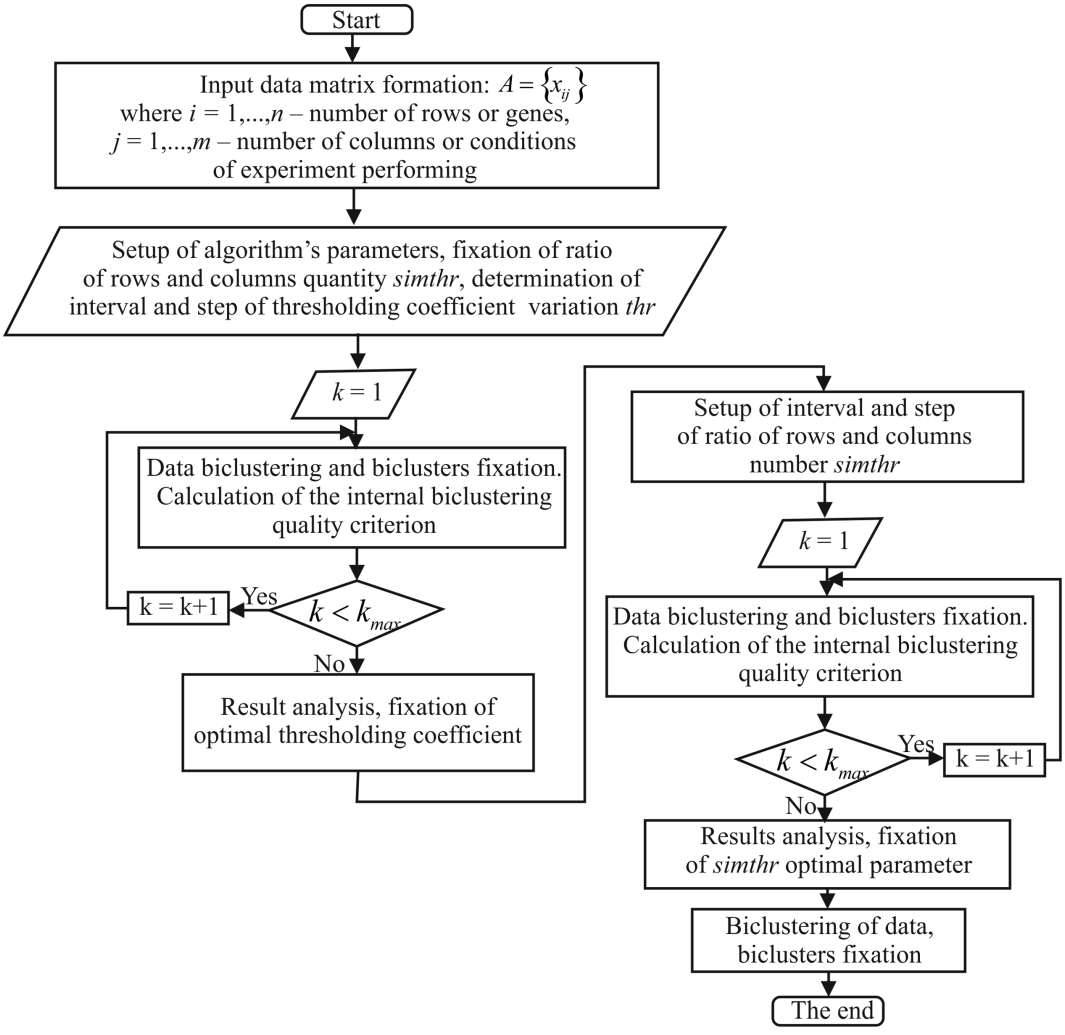

Figure 3.8: Structural block-charts of the algorithm to implement technology of bicluster analysis based on *ensemble* biclustering algorithm

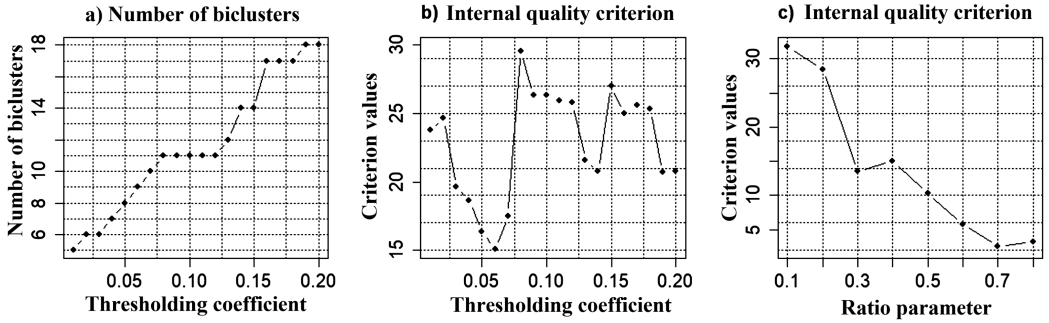

Figure 3.9: Results of the simulation concerning bicluster analysis of the synthetic dataset 2

Table 3.3: Disrtibution of rows and columns in the biclusters obtained from synthetic data 2

Biclusters	1	2	3	4	5	6	7	8	9
Rows	13	12	12	13	4	3	5	13	13
Columns	22	23	23	23	8	6	21	23	27

(a) Data biclustering with the use of *ensemble* biclustering algorithm optimal parameters. Biclusters fixation.

3.3.4 Bicluster Analysis with the use of Intersectional Synthetic Biclusters

Figure 3.9 shows the results of the simulation concerning bicluster analysis of the synthetic data presented in Figure 3.1b. The biclustering process was performed within the framework of the technique presented in Figure 3.8. It is obvious, that in this case the number of biclusters can be more than five due to their mutual intersection. The following optimal algorithm parameters were determined as the result of Figure 3.9 analysis: $thr = 0.06$, $simtfr = 0.7$. Nine biclusters were selected in this case. Distribution of rows and columns in the biclusters are presented in Table 3.3. Analysis of the simulation results allows us to conclude that the obtained biclustering is adequate, since the biclusters 1,2,3,4,8 coincide almost completely with biclusters in the synthetic data 2. However, the ninth bicluster contained 13 rows and 17 columns was allocated too. This bicluster can also be interesting for the following investigation. Other biclusters can be removed due to lower number of rows and columns.

Results of the simulation concerning bicluster analysis of the synthetic data 3 and 4 (Figure 3.1c,d) are presented in Figure 3.10 and Figure 3.11. The investi-

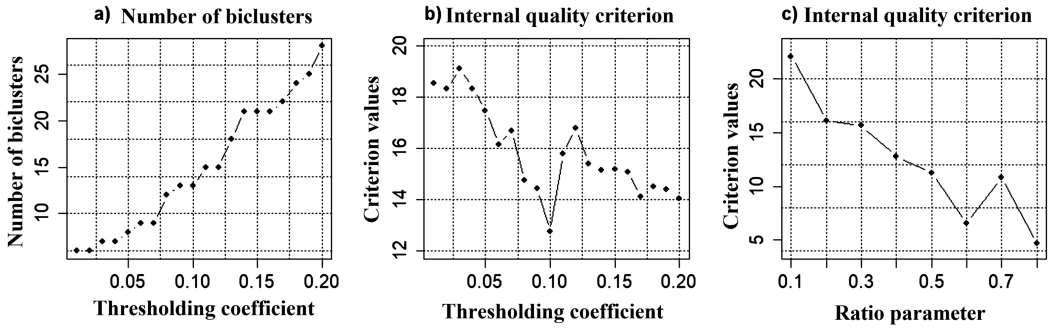

Figure 3.10: Results of the simulation concerning bicluster analysis of the synthetic dataset 3

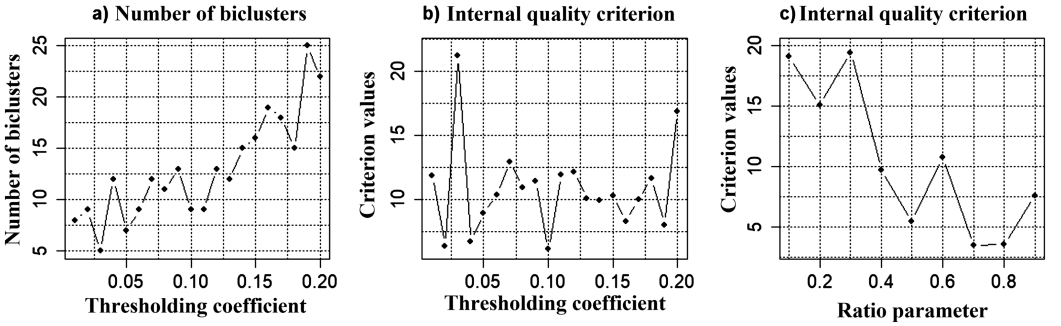

Figure 3.11: Results of the simulation concerning bicluster analysis of the synthetic dataset 4

Table 3.4: Disrtibution of rows and columns in the biclusters obtained from synthetic data 3

Biclusters	1	2	3	4	5	6	7	8
Rows	15	21	14	11	5	13	8	13
Columns	31	33	16	23	25	20	14	16
Biclusters	9	10	11	12	13	14	15	16
Rows	2	6	3	2	2	2	3	4
Columns	28	13	11	15	26	22	31	13

Table 3.5: Disrtibution of rows and columns in the biclusters obtained from synthetic data 4

Biclusters	1	2	3	4	5	6	7	8	9	10	11
Rows	10	21	8	6	6	16	2	2	3	3	2
Columns	23	26	19	15	25	20	12	17	18	13	10

gated datasets contain biclusters of different size with various level of their mutual intersection. Analysis of the obtained results allows concluding that the increasing level of mutual intersection complicates the process of biclusters allocation. This fact can be explained by greater amounts of variants of rows and columns grouping into biclusters. In this case the character of the data grouping into biclusters is changed more intensively during change of the algorithm parameters.

Table 3.4 and Table 3.5 present the results of rows and columns distribution within the obtained biclusters. Parameters of *ensemble* biclustering algorithm in these cases were the following: synthetic_3 data: $thr = 0.1$, $simthr = 0.6$; synthetic_4 data: $thr = 0.1$, $simthr = 0.7$. As it can be seen from the Table 3.4 and Table 3.5, 16 biclusters were selected in the case of synthetic_3 data use, but only nine biclusters (1-8,10) are informative in terms of amount of the data in biclusters. Eleven biclusters were selected in the case of synthetic_4 data use, but only the first six are informative ones.

3.4 Bicluster Analysis With the Use of Gene Expression Profiles

Figure 3.12 shows the results of the simulation concerning bicluster analysis of patients' gene expression profiles, which were investigated on lung cancer disease [30]. The gene expression profile in this case is a vector of numeric values, each of them determines the expression of gene for appropriate sample. The thresholding coefficient value was changed within the range from 0.05 to 0.5 by step 0.01. The

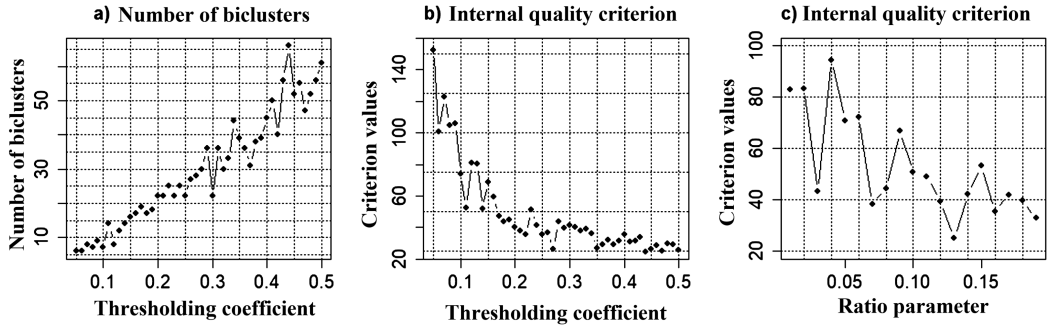

Figure 3.12: Results of the simulation concerning bicluster analysis of gene expression profiles

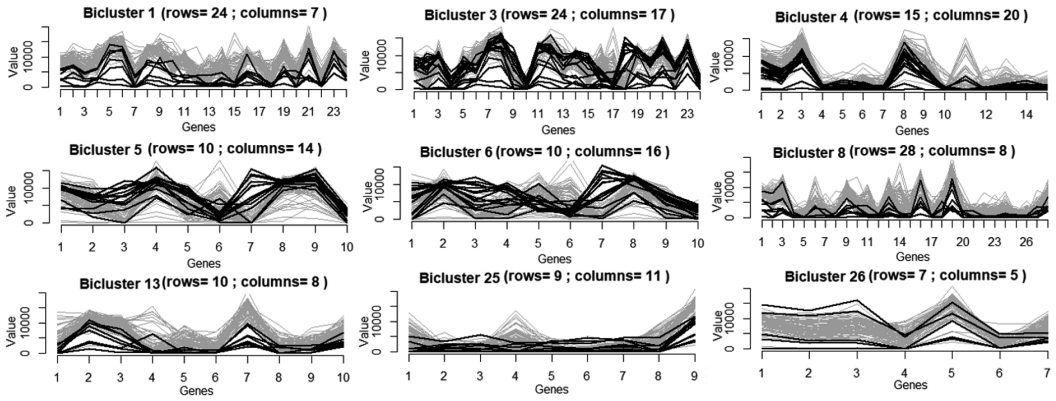

Figure 3.13: Gene expression profiles of nine of the biggest biclusters

value of thresholding coefficient 0.27 was determined as the simulation result (Figure 3.12b). This value corresponds to the global minimum of the internal biclustering quality criterion. The increase of the thresholding coefficient value is not reasonable since it promotes to sharp increase of the number of little biclusters. Figure 3.12c shows the chart of the internal biclustering quality criterion versus the value of ratio coefficient, which determines the ratio of rows and columns in biclusters. The thresholding coefficient value in this case was 0.27. The *simthr* value was changed within the range from 0.01 to 0.2 by step 0.01 during the simulation process. The *simthr* value 0.13 was determined as optimal one because this value corresponds to the minimal value of the internal biclustering quality criterion. 27 of biclusters were allocated as the simulation results. Figure 3.13 shows the gene expression profiles of nine of the biggest biclusters.

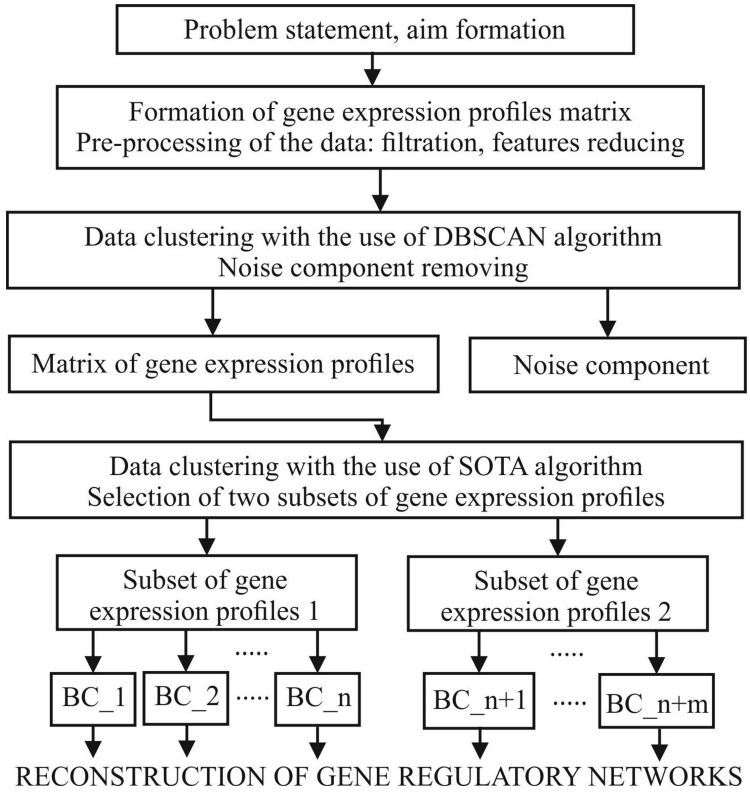

Figure 3.14: Structural block chart of the cluster-bicluster analysis hybrid model

As it can be seen, the gene expression profiles, which are within different biclusters are differed between each other. At the same time, the gene expression profiles in single biclusters are similar to each other. This fact indicates the correctness of the proposed biclustering technique.

3.5 Hybrid Model of Cluster-Bicluster Analysis of Gene Expression Profiles

The conducted research allow us to propose a hybrid model of gene expression profiles grouping based on step-by-step application of clustering and biclustering techniques. Structure block chart of this model is presented in Figure 3.14. Practical implementation of this model involves the following stages:

1. Formation of data and their preprocessing.

   (a) Formation of matrix of the gene expression profiles, data filtration, and

non-informative genes reducing.

2. Gene expression profiles clustering.

   (a) Division of the initial gene expression profiles data set into two equal power subsets, which contain the same quantity of pairwise similar gene expression profiles.

   (b) Determination of the optimal parameters for DBSCAN and SOTA clustering algorithms using the technique presented in the section 2.3.

   (c) Data clustering with the use of DBSCAN algorithm. Selection of a noise component. Formation of a new matrix of gene expression profiles for the following processing.

   (d) Clustering the obtained gene expression profiles with the use of SOTA clustering algorithm. Formation of two subsets of gene expression profiles for the following bicluster analysis.

3. Gene expression profiles biclustering on the obtained clusters.

   (a) Choice of the biclustering algorithm, setup of its parameters.

   (b) Biclustering process. Formation of the biclusters, which contain mutually correlated genes and conditions of experiment performing.

4. Reconstruction of the gene regulatory networks based on the data of the obtained biclusters.

It should be noted that practical implementation of hereinbefore described step-by-step technology allows us to save more useful information at the stage of gene expression profiles preprocessing. Initial data set of the gene expression profiles, which can be obtained by DNA microarray experiments or by RNA sequencing methods, contains tens of thousands of genes. Bicluster analysis with the use of the initial data allows us to receive the biclusters of mutually correlated genes and conditions of experiment performing. However, the parallel gene expression profiles processing with the use of step-by-step data clustering technology promote to higher level of concentration of mutually correlated genes and conditions of the experiment performing. This fact influences the following process of the gene regulatory networks reconstruction and simulation of their operation.

3.6 Conclusions

This chapter presents the comparison analysis of biclustering algorithms with the use of synthetic data contained five biclusters. In the first and in the second cases

the generated biclusters were the same, but in the first case they have no intersection and in the second case biclusters have the same intersection between each other. The third and the fourth data contained different and mutually intersected biclusters. Jaccard index and the internal biclustering quality criterion have been used to estimate the biclustering algorithm effectiveness. *Bimax*, *CC*, *spectral* and *ensemble* biclustering algorithms have been used during the simulation process. Results of the simulation with the use of non-intersectional similar biclusters have shown that Jaccard index and the internal biclustering quality criterion have extrema, which correspond to the same and the best character of the studied data grouping. This fact has allowed us to use the internal biclustering quality criterion as the main criterion to estimate the effectiveness of the studied data grouping into biclusters. Simulation results have shown also that *ensemble* biclustering algorithm is more effective in comparison with other used biclustering algorithms. Technique of biclustering based on *ensemble* algorithm and step-by-step procedure of its implementation has been proposed as the simulation results. Application of this technique allows us to determine the optimal parameters of the biclustering algorithm operation.

Application of the proposed technique with the use of the test synthetic data, which contain intersection biclusters, has shown high efficiency of its operation. The studied data have been divided into biclusters adequately, but the increase of the level of mutual intersection complicates the allocation of biclusters due to greater number of variants of rows and columns grouping into biclusters.

The hybrid model of cluster-bicluster analysis based on the complex use of DBSCAN and SOTA clustering algorithms and bicluster analysis method has been proposed as the simulation result. Practical implementation of this model at the early stage of gene regulatory network reconstruction allow us to increase the quality of the reconstructed gene network by more careful grouping of the mutually correlated genes and conditions of the experiment performing.

Chapter 4

Gene Expression Profiles Pre-Processing

4.1 Introduction

Gene regulatory network is a set of genes which interact between each other to control the specific cells functions [146]. Qualitatively reconstructed gene regulatory network promotes to better understanding of the genes interaction mechanism in order to create new methods of both the early diagnostics and treatment of complex genetic diseases. The gene expression profiles which are obtained by DNA microarray experiments or by RNA molecules sequencing method, are the basis for the gene regulatory networks reconstruction [126, 68]. High dimension of feature space is one of the distinctive peculiarities of the studied data. The reconstruction of gene networks based on the whole dataset of gene expression profiles is very complicated task due to the following aspects: it requests large computer resources; complexity of the reconstructed gene regulatory networks complicates the interpretation of obtained results. Therefore, it is necessary at early stage of gene regulatory network reconstruction to process the gene expression profiles with the use of current computational and information techniques of complex data processing. This process includes data formation as a matrix of genes expression, non-informative genes reducing, data clustering and biclustering in order to select mutually correlated genes and samples.

Figure 4.1 shows the structural block-chart of the information technology of gene expression profiles processing for purpose of both gene regulatory network reconstruction and validation of the reconstructed models.

As it can be seen from Figure 4.1, implementation of the technology assumes five stages:

Stage 1 Formation of gene expression profiles array.

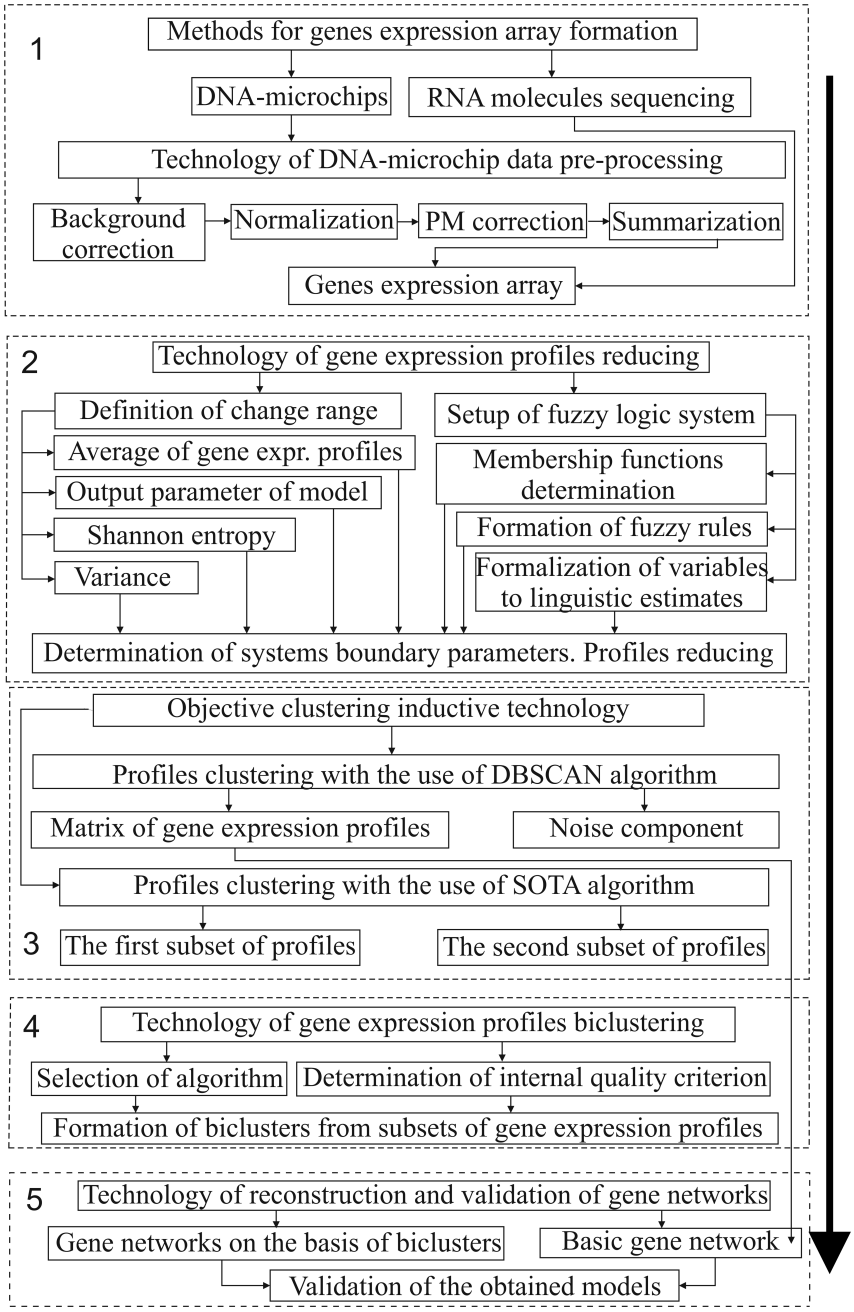

Figure 4.1: Information technology of gene expression profiles processing

In the case of the use of RNA molecules sequencing method, we have the matrix of gene expressions directly. However, in this case, it is necessary to select informative genes in terms of appropriate criteria. In the case of DNA microchip experiment performing, we have as the initial data the matrix of light intensities. Convertation of this matrix into matrix of gene expressions assumes the following steps: background correction, normalization, PM correction and summarization. Determination of optimal combination of the methods to perform this process is one of the current tasks. Further, we will present our versus of this task solve.

Stage 2 Gene expression profiles reducing.

The aim of this stage is division of the studied gene expression profiles into informative and non-informative in terms of complex use of statistical criteria and Shannon entropy. It is assumed that if variance or average of absolute values of gene expression profiles is less and Shannon entropy is greater than appropriate boundary value, then, these profiles are identified as non-informative and they can be removed without significant loss of useful information.

Stage 3 Step-by-step gene expression profiles clustering within the framework of the objective clustering inductive technology.

As was described in the chapter 2, the use of DBSCAN clustering algorithm allows us to allocate the genes, which are identified as noise. These genes are removed from the studied data. At the second step of the clustering process implementation, the gene expression profiles are divided into two clusters with the use of SOTA clustering algorithm. These subsets are used for the following bicluster analysis.

Stage 4 Bicluster analysis of the obtained subsets of gene expression profiles.

Implementation of this stage allows us to allocate subsets of mutually correlated vectors of both gene expression profiles and conditions of the experiment performing. These subsets are used for gene regulatory networks reconstruction at the next step of this procedure implementation.

Stage 5 GRN reconstruction and validation of the reconstructed models.

The optimal topology of the obtained gene networks is determined on the basis of the maximum value of general Harrington desirability index, which contains as the components the network topological parameters. Validation of the reconstructed models is performed based on the comparison analysis of the interconnection between the appropriate genes in the basic network and in the networks reconstructed based on the obtained biclusters. ROC-analysis

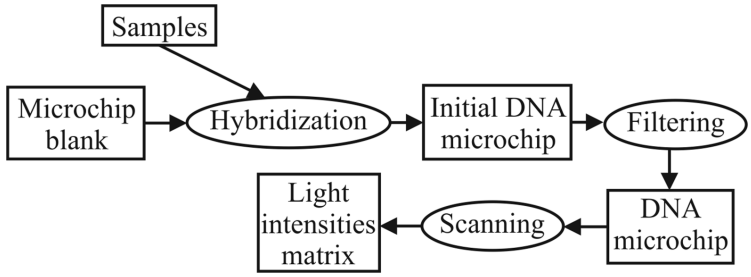

Figure 4.2: A block chart of the procedure of DNA microchip light intensities matrix formation

technique is used to calculate the relative quality criterion, which indicates a quality of the reconstructed gene networks.

4.2 Techniques of Genes Expression Array Formation

As was described hereinbefore, two technique are used nowadays to form array of genes expression: DNA microchip technique and RNA molecules sequencing method. Each of the techniques has its advantages and shortcomings. DNA microchips technique is significantly more cheaper, but exactness of genes expression estimation in this case is significantly lower in comparisson with RNA molecules sequencing method. However, two techniques are used concurrently nowadays. In this reason, below, we describe both techniques.

4.2.1 Technique of DNA-microchips Data Processing

The data, which are obtained during the DNA microchip experiments implementation, are presented as a matrix of light intensities. A block chart of the procedure of DNA microchip light intensities matrix formation during the experiment performing is presented in Figure 4.2. Joining of complementary single-chain nucleotides with fluorescent labels to a single molecule is performed during the hybridization process. It is obvious, that the level of light intensities in appropriate point of the microchip is proportional to quantity of the hybridized RNA molecules, which correspond to appropriate type of the protein. The following stages of the DNA microchip processing are filtering in order to remove unhybridized samples and scanning for purpose of the matrix of light intensities formation. Figure 4.3 presents the step-by-step procedure of transforming the light intensities values to the expression of the appropriate genes. As it can be seen from Figure 4.3, each of the steps assumes the use of various methods and choice of the combination of these methods influences directly to the quality of the obtained genes expression values. Thus, the main prob-

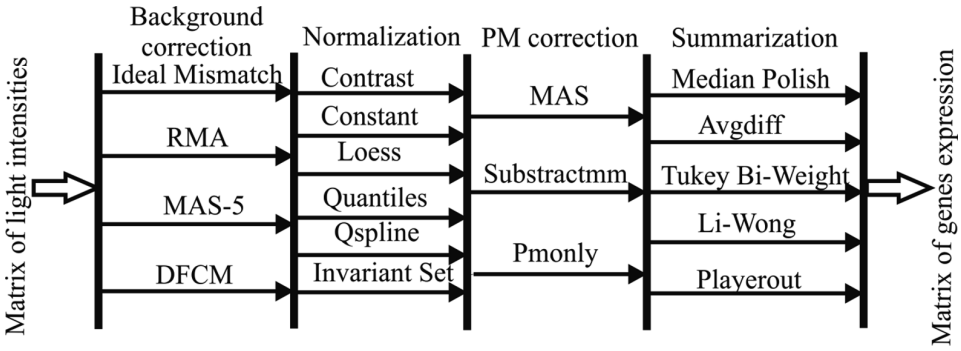

Figure 4.3: A chart of step-by-step procedure of transforming the light intensities matrix to the matrix of genes expression

lem consists in determination of the optimal combination of the methods to process the DNA microchip data in order to increase the informativity of the obtained gene expression data.

The issues concerning DNA microarray data processing are presented in [90, 79, 23]. The authors considered in detail the stages of DNA microarrays creation and the peculiarities of their processing. However, these papers do not contain the investigations concerning determination of optimal combination of the methods based on quantitative criteria.

Classification and detail description of the background correction methods are presented in [34, 2, 76, 41]. Ideal Mismatch method was proposed by Affymetrix company [2]. This method involves the complex use of both the Perfect Match (PM) nucleotide samples which fully correspond to the investigated genes and the Miss Match (MM) samples, in which the mean nucleotide is changed to complementary one. Robust Multichip Average (RMA) background correction method involves the use only PM samples [76]. This fact decreases the costs to the microchip preparing due to absence of the MM samples. The values of light intensities in this case are presented as the sum of the useful signal, which is distributed exponentially, and the normally distributed noise component. Distribution Free Convolution Model (DFCM) background correction method [41] also assumes that values of light intensities are presented as the combination of both the useful signal and the noise component. But in this case do not any assumes about the character of the components distribution. This method involves the use of both the PM and MM samples. The main idea and the detail description of the Affymetrix Micro Array Suite 5.0 (MAS 5.0) technique of background correction are presented in [34, 2].

The techniques of DNA microchip data normalization are presented in [34, 59, 113, 117, 6, 40]. The necessity of this stage is determined by low correlation of the data which were determined when different conditions of the experiment performing.

The aim of the normalization process is the reduction of the microchip empirical data to the same distribution. This step allows minimizing the technological differences between the parameters of different genes and, as a result, to carry out the comparison of the expression values of the corresponding genes obtained under different conditions of the experiment performing. The results of the research concerning comparison analysis of various methods of PM corrections and summarization of the DNA microchip data are presented in [34, 117, 29, 91]. PM correction stage is performed in order to reduce the nonspecific hybridization effect by correction of PM samples light intensities considering the light intensities of the appropriate MM samples. The summarization process assumes the calculation of gene expressions values from light intensities of the samples for investigated genes.

Below, we present the technique of DNA microarray data processing based on the complex use of Bioconductor tools and Shannon entropy for purpose of gene expression array formation [12].

Materials and methods

The Shannon entropy criterion, which is calculated based on James-Stein shrinkage estimator [67], was used as the main criterion to estimate the gene expression informativity during the simulation process. This technique is described in detail in the chapter 1 (formulas 1.4-1.6). Less value of Shannon entropy criterion (1.6) corresponds to the higher level of the investigated vector informativity. A structural block chart of the algorithm which was used to determine the optimal combination of the methods of DNA microarray data processing is shown in Figure 4.4. Implementation of this algorithm involves the following steps:

1. Loading the DNA microarray data.

2. Setup of the stage of data processing (background correction, normalization, PM correction, summarization). Fixation of the methods, which do not correspond to this stage randomly.

3. Choice of the first method for current stage.

4. DNA microarray data processing using selected methods.

5. Calculation of the Shannon entropy for each of the investigated microchips. Calculation of average value of the Shannon entropy for all DNA microarrays.

6. If the number of the method is less than the maximum quantity of the methods at this stage, then choice the next method and go to the step 4 of this procedure. Otherwise fixation of the method which correspond to the minimal value of the Shannon entropy.

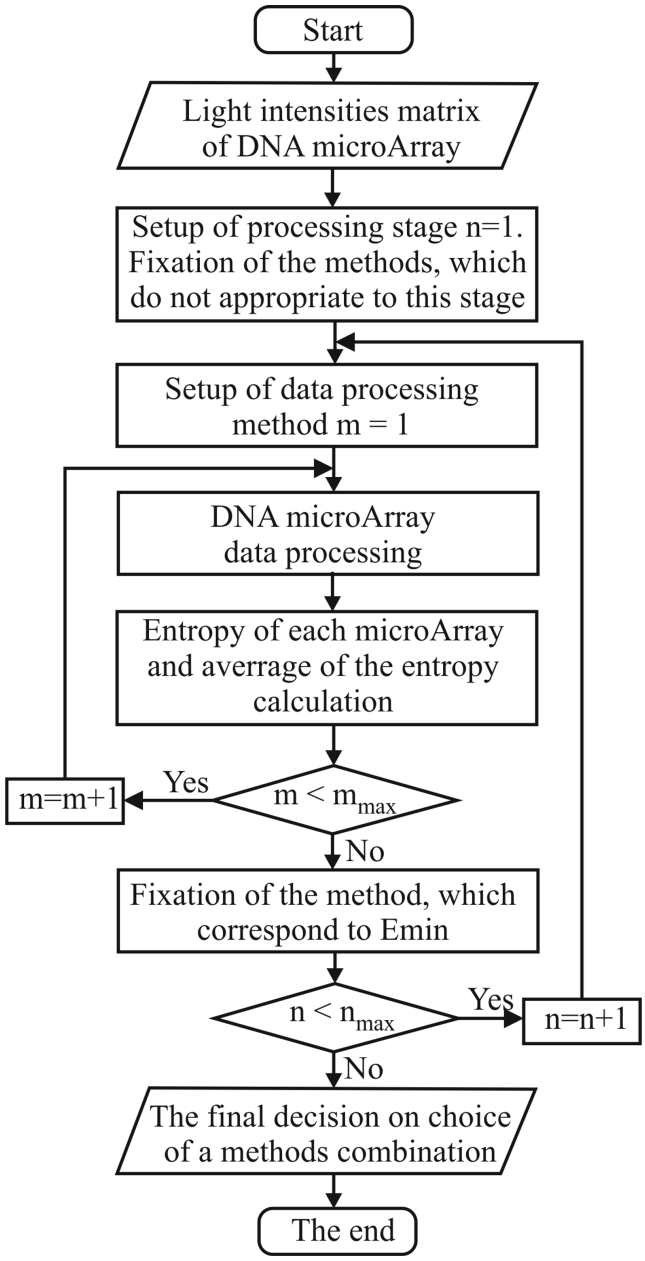

Figure 4.4: A chart of the step-by-step procedure to transform the light intensities matrix to the matrix of genes expression

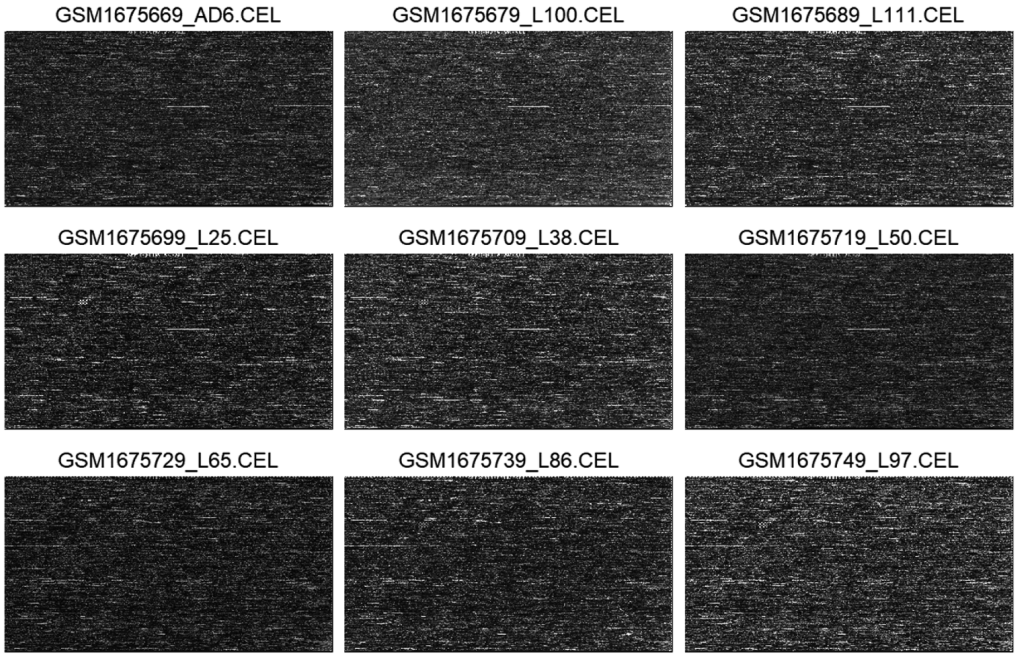

Figure 4.5: Scanning images of nine of the investigated DNA microchips

7. If the number of the stage is less than maximal quantity of the stages, then go to the next stage and go to the step 3 of this procedure. Otherwise, DNA microarray data processing with the use of determined combination of the methods.

Experiments

Simulation process of DNA microchip data pre-processing was performed based on R software [75] using functions of Bioconductor package [73, 59]. The lung cancer patients' gene expression profiles E-GEOD-68571 [30] from database ArrayExpress were used as the experimental ones during the simulation process. These data include 96 of DNA microchips of patients which were investigated on lung cancer. Each of the DNA microchips includes 7129 of genes. 10 patients were identified as healthy and 86 sick patients were divided by the state of their health into three groups. Figure 4.5 shows the images of nine of the investigated DNA microchips which were obtained by scanning of the appropriate objects. The following step of the data processing is the transform of the light intensity matrixes into array of genes expressions using the hereinbefore presented methods.

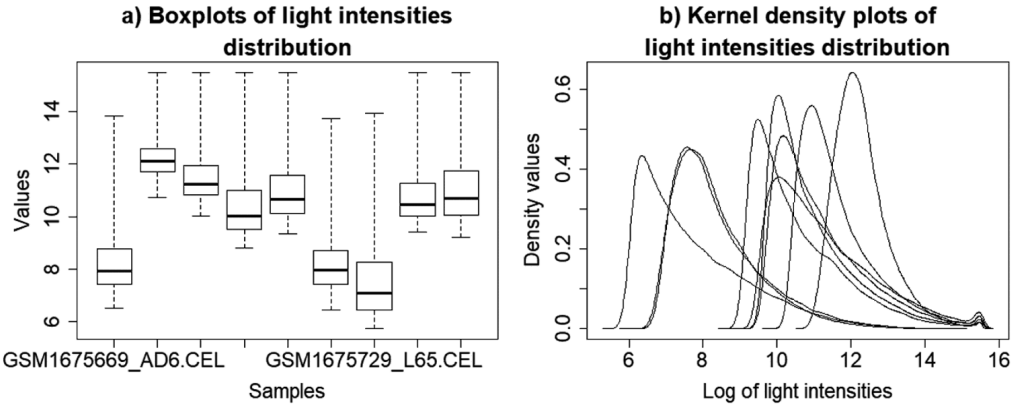

Figure 4.6: Estimation of light intensities distribution in the selected DNA microchips

Results and discussions

The character of light intensities values distribution in the selected DNA microarrays is presented in Figure 4.6. Figure 4.7 shows the MA charts for all pairs of five selected DNA microchips. The MA chart shows the difference of logarithms of the PM (Perfect Math) samples values (M) versus the average of logarithms of the PM samples values (A). The parameters M and A for i-th gene and samples k and n are calculated in the following way:

$$M_k = log_2(\frac{x_{ki}}{x_{ni}}),\ A = \frac{1}{2} log_2(x_{ki} \cdot x_{ni}) \tag{4.1}$$

The chart is created for PM values for all possible pairs of the investigated samples. In the case of highest quality of the data processing, the data should be distributed in a rather narrow range, and the points at MA diagram should be located along the axis of $M = 0$ with the lowest averages.

The analysis of the received diagrams confirms the assumption concerning the necessity of the initial data preprocessing. The character of the data distribution for various microchips is differed significantly (Figure 4.6a). The values at kernel density plots which are shown in Figure 4.6b are distributed along axis of the light intensities logarithm randomly too. Finally, the corresponding points on the MA diagrams (Figure 4.7) have different distributions too. These facts do not allow us to compare the investigated gene expression profiles objectively.

Figure 4.8 and Figure 4.9 present the results of the research concerning background correction of the DNA microarrays by methods: *rma*, *mas* and *DFCM*. *Ideal Mismatch* method has not used due to lower quality of its operation [26]. The

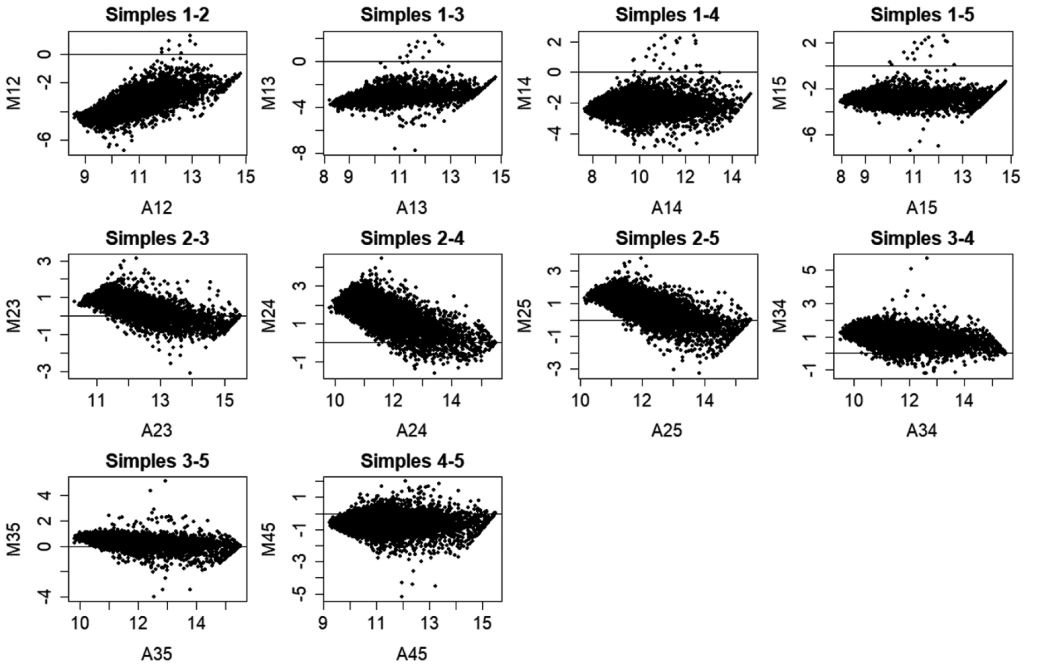

Figure 4.7: *MA* charts of light intensities distribution for PM samples

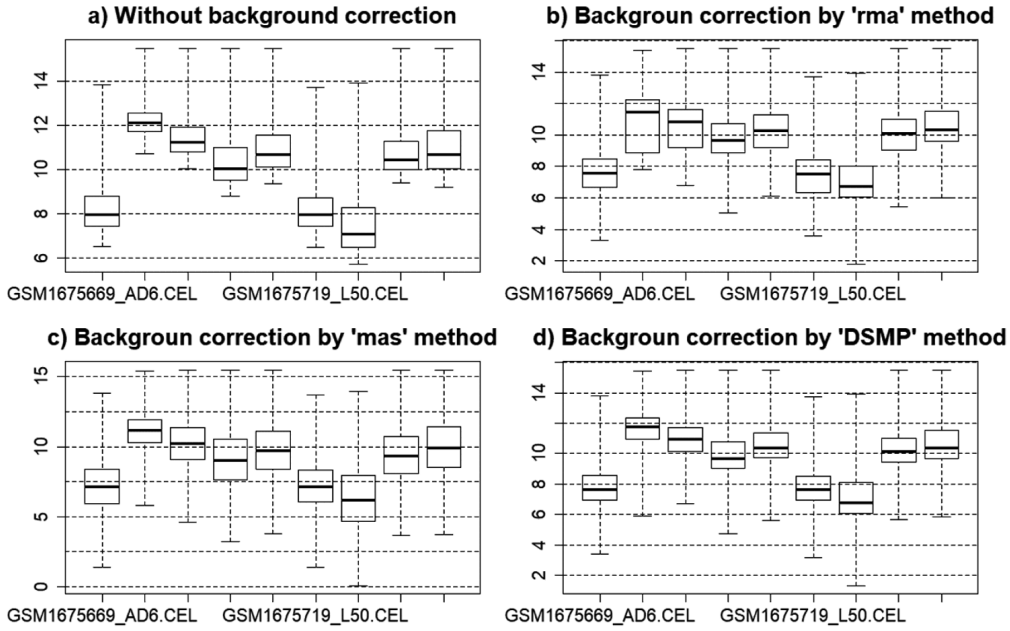

Figure 4.8: Boxplot charts of unprocessed and processed data when the various background correction methods are used

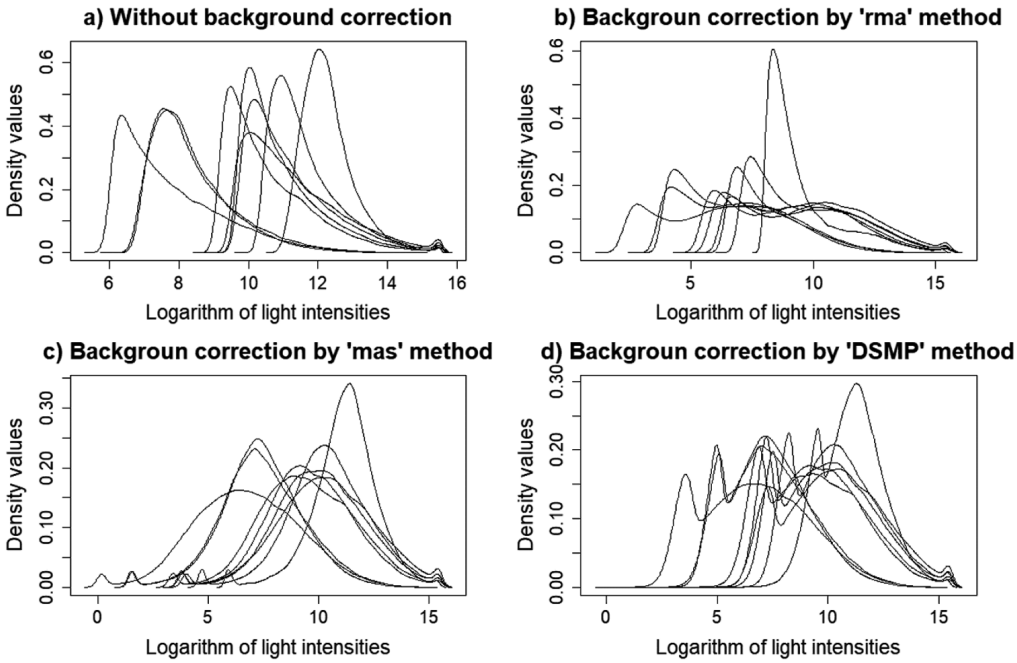

Figure 4.9: Kernel density plots of unprocessed and processed data when the various background correction methods are used

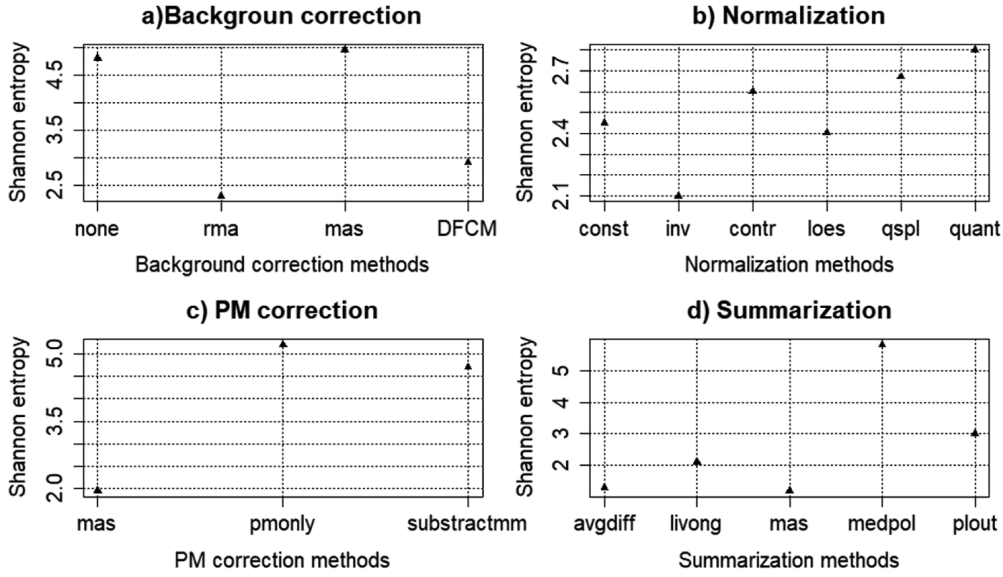

Figure 4.10: Charts of Shannon entropy distribution versus the methods of the data processing at the stages: a) background correction; b) normalization; c) PM correction; d) summarization

analysis of the obtained charts allows us to conclude that the background correction increases the image quality. The processed data are distributed more uniformly in comparison with unprocessed data distribution. However, it should be noted that visual analysis of the diagrams does not allow comparing the quality of the used methods objectively in order to choose the best one. Figure 4.10 presents the results of the research concerning determination of the optimal combination of the methods to process the DNA microarray data based on the minimum value of the Shannon entropy in accordance with hereinbefore presented technique.

The analysis of the obtained charts allows us to conclude that the optimal methods in terms of the minimal value of Shannon entropy criterion are the following ones: *rma* background correction method; *invariant set* normalization method; *mas* methods PM correction and summarization. This combination of the methods was used to process the investigated DNA microarrays. Figure 4.11 presents the boxplots of genes expression profiles for the investigated samples of both the non-processed (Figure 4.11a) and processed (Figure 4.11b) data.

As it can be seen from Figure 4.11b, the values of gene expressions are distributed in the same range. The change of this range can be explained in the following way.

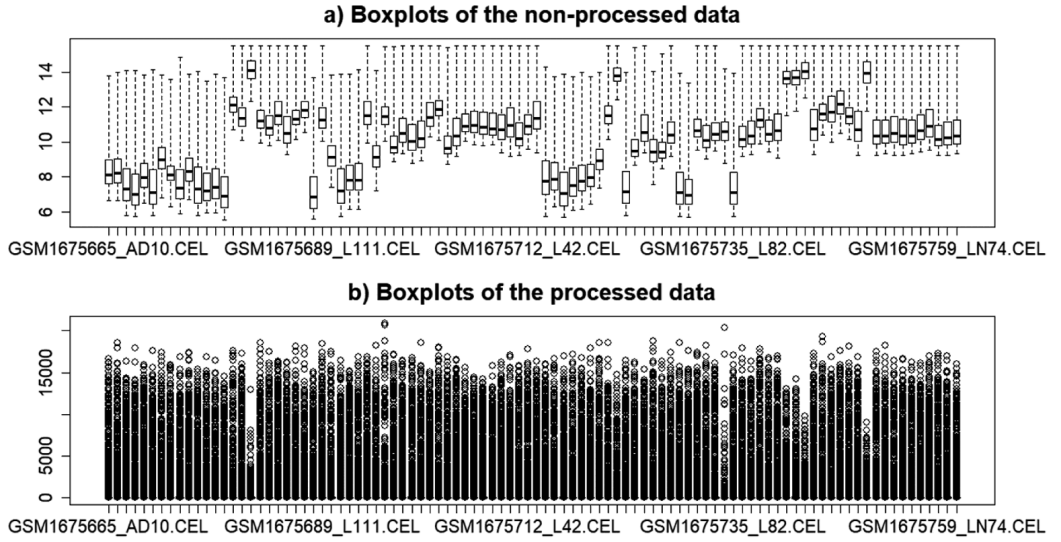

Figure 4.11: Results of the DNA microchips processing

The expression values of the largest quantity of genes are low. But some of the genes have significantly higher values of expression. It means that these genes determine some important processes in the investigated objects. The expression values of these genes determine the variation range of another genes expression. The analysis of the boxplots allows us also to conclude that the values of the largest quantity of gene expressions for various objects lie in a very narrow range. This can mean that these genes are responsible for the functions that are inherent for all investigated objects. However, each of the investigated samples contains genes, the expression of which goes beyond the inter-quartile range. These genes are very important for the following research since they allow us to distinguish the investigated objects by their particularities.

4.2.2 RNA-molecules Sequencing Method

Applying RNA-molecules sequencing method allows obtaining the number of investigated genes for studied samples directly. In this reason, this method is more exact in comparison with DNA-microchip technique. The number of genes determines the level of this gene activity or its expression. At the next step it is necessary to remove non-expressed genes for all samples and gene with low level of expression. At this stage it is appear the problem concerning identification of boundary value

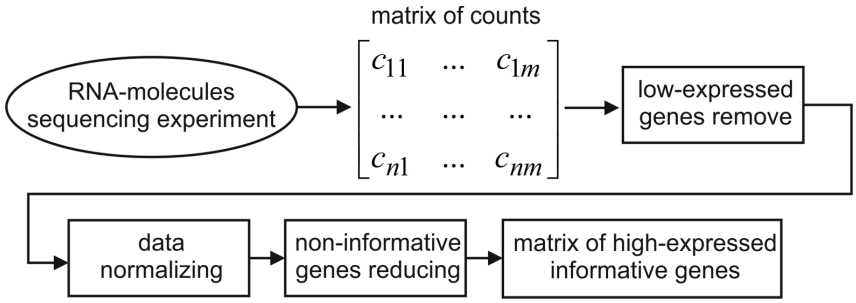

Figure 4.12: A step-by-step procedure to transform a matrix of counts to the matrix of highly-expressed informative genes

which allows dividing genes to lowly-expressed and highly-expressed. Moreover, the matrix of counts of genes is not suitable for the following processing. Thus, initially, the data should be normalized. This step assumes transformation the counts values into the same suitable range. There are various normalized methods to process gene expression values. However, it should be noted, that the task of objective selection of appropriate normalizing method based on the quantitative criteria has not effective solution nowadays.

Formal problem statement

A block chart of procedure to process the experimental data which are obtained by RNA-molecules sequencing technique is presented in Figure 4.12. The studied dataset is presented as a matrix of counts, values of which are the number of genes for appropriate sample. One of the most important steps of this procedure implementation is the data normalizing. The normalized values of gene expression profiles should have the equal ranges and their norms should be distinguish minimally between each other. Moreover, the values of gene expressions should allow us to identified the samples which belong to various clusters. Considering hereinbefore, evaluation of quality of gene expression profiles normalizing will be performed visually by analysis of both box plot and kernel density plot, and based on quantitative criterion.

Various techniques have been proposed over last years to pre-process the results of RNA-molecules sequencing experiments [36, 64, 51, 63, 72, 107, 112, 95, 122]. These tools are differed between each other by types and thresholds which are used to process the counts of genes and by algorithms which are used for filtering and alignment of the investigated values. It should be noted that the choice of the alignment algorithm has very great influence to the evaluation accuracy of RNA molecules abundance in the sequenced samples. In this reason, testing different

tools for processing of data of RNA-molecules sequencing experiments can help us to choose the best technique for current type of data.

After implementation of the alignment procedure, it is necessary to normalize the recovered values of miRNA counts for purpose of removing variations in the data which have not biological origins and, as a result, can influence to the ranges of measured values change. Correct applying the normalizing technique allows minimizing the experimental and the technical bias without noise introduce. In [46, 58] the authors have proposed several normalizing techniques for data of RNA-molecules sequencing experiments. As a result of comparison of different normalization methods effectiveness, the conflicting results were obtained in these works. So, the authors in [58] proposed using the locally weighted linear regression and quantile normalizing techniques. At the same time, they were discouraging against the use of trimmed mean of M values (TMM technique). The obtained results were validated based on the use of polymerase chain reaction (qPCR). In [46] the authors proposed the opposite, to use against quantile normalization the TMM method. The simulation results were used to confirm these findings. An assessment of the relative effectiveness of various pre-processing techniques in terms of statistical criteria, bias, sensitivity and specificity in order to detect the differential expressed genes can be achieved on the basis of complex implementation of both qualitative and quantitative normalizing quality criteria using current techniques of data processing [10].

Data set

We used the dataset *GSE129336* generated from Gene Expression Omnibus (GEO) database [1] as the experimental data during the simulation process. The data contains the results of expression profiling by high throughput sequencing in human SH-SY5Y neuroblastoma cells [53]. The transcriptomic responses to Mn dose (0,1,5,10,50,100 μM MnCl2 for 5 h) in the investigated cells with three biological replicates per Mn treatment were examined during the experiment performing. Thus, the examined samples can be divided into six clusters considering the Mn dose. Each of the clusters contains three samples. This fact can be used to calculate one of the criteria to estimate the quality of gene expression values processing. Each of the samples contained 53186 of genes. So, the initial dataset contained 53186 of genes or rows and 18 of columns or samples. The early analysis has shown, that there were 27838 non-expressed genes (zero for all samples). Of course, these genes can be removed from the data at the first step. Moreover, the lowly-expressed genes for all samples can be removed from the data too. The search of the thresholding value to remove lowly-expressed genes is one of the solved tasks within the framework of this research.

Removing lowly-expressed genes

As was noted before, the studied dataset contains 53186 of genes. However, 27838 of genes are non-expressed for all samples (the count value is zero). Thus, the number of the expressed genes can be changed from 53186 to 25348 of genes.

At the next step, it is necessary to remove lowly expressed genes considering the appropriate thresholding value. The initial values of the counts of genes are not suitable for solve this task since the range of the genes count value change is very large (in the case of our dataset this range is changed from 0 to 47434890). In this case it is necessary to transform the count value scale into other, more suitable scale. To solve this tack, *Bioconductor* package contains *cpm()* function which allows transforming the counts values into count-per-million values as follows:

$$x'_{ij} = \frac{x_{ij}}{\sum_{i=1}^{n} x_{ij}} \cdot 10^6 \tag{4.2}$$

where n is the number of rows, x_{ij} is the value in i-th row and j-th column. Applying this function allows us to obtain the new, more suitable, range of the data values change (from 0 to 380367.8).

The main idea for lowly-expressed genes removing is the following: the use a nominal thresholding of 1 *cpm* value (this value is corresponded to 0 of $log2(cpm)$ value) allows dividing the genes into two groups (expressed and unexpressed). If value of gene expression is more than this threshold, the gene is identified as expressed. Otherwise, the gene is identified as unexpressed. Considering the number of samples in the clusters we can suppose that the genes should be expressed in at least one cluster (three samples) for the further analysis.

Normalizing gene expression profiles

The following normalizing techniques were evaluated during the simulation process: 1) *lcpm*; 2) *TMM*; 3) *TMMwsp*; 4) *RLE*; 5) *upper quartile scaling*. Brief describing each of these techniques is presented below:

1. *lcpm* is the simplest normalizing technique, the $log2(cpm)$ values are calculated during this technique implementation.

2. *TMM* is trimmed mean of M is the normalizing technique by total count of scaling. The counts quantity for an appropriate target for all samples is estimated during TMM technique implementation. If an expression value is identified in the same proportion for all samples, this gene is identified as non-differentially expressed. It should be noted that this technique does not allow considering the potentially different RNA molecules which are presented in the samples. Applying this method allows us to calculate a linear scaling index for

appropriate sample considering weighted average after transforming the data using log fold-changes (M) relative to the absolute intensity in the reference sample (A) [121].

3. *TMMwsp* is TMM with singleton pairing. This technique is a variant of TMM, in which the data with a high proportion of zeros are processed. Implementation of the TMM method assumes that the genes which have zero value in either library are ignored when pairs of libraries are compared between each other. As opposed to TMM method, implementation of the TMMwsp technique assumes that the positive counts from such genes are reused to increase the quantity of features which are used to compare the libraries. The singleton positive counts are paired up between the libraries in decreasing order of size and then a slightly modified TMM method is applied to the reordered libraries.

4. *RLE* is relative log expression technique. Implementation of this method assumes that the median library is calculated from the geometric average of all columns and the median ratio of each sample to the median library is used as the scale factor.

5. *Upper quartile scaling* is the upper-quartile normalizing technique, in which the scale factors are calculated from the 75% quantile of the counts for each of the libraries, after removing genes that are zero in all libraries.

Quantitative criterion to estimate the quality of data normalizing

The main idea to evaluate the quality of data normalizing is the following: as we noted hereinbefore, the samples can be divided into six clusters considering the dose of Mn. Each of the clusters in this case contains three samples. It is naturally that informativity of gene expression profiles is determined by their ability to distinct the samples in different clusters. Thus, the quality of data normalizing can be estimated based on clustering quality criterion which should consider the samples distribution within clusters and the clusters distribution in the feature space. Considering the high dimension of the studied vectors, the correlation metric should be used to estimate the proximity level between the vectors. This quality criterion of the samples and clusters grouping was calculated as multiplicative combination of Calinski-Harabasz criterion and WB-index [37, 149]:

$$QC_{int} = \frac{K(K-1)QCW^2}{(N-K)QCB^2}; \tag{4.3}$$

where: K is the clusters quantity; N is the samples quantity; QCW is an average distance from samples to centers of the clusters where these samples are allocated; QCB is an average distance between cluster centers. It should be noted that minimum value of this criterion corresponds the best normalizing technique.

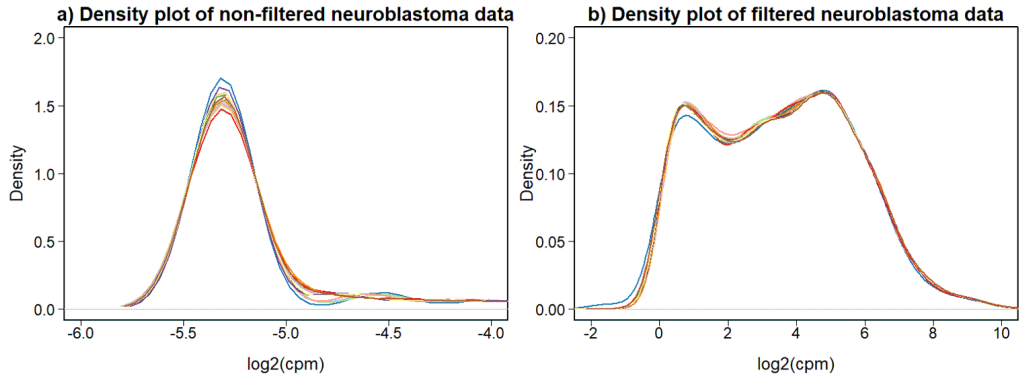

Figure 4.13: Density plots of non-filtered and filtered gene expression values distribution for neuroblastoma data samples

Experiments, results and discussions

Figure 4.13 presents the results of lowly expressed genes reducing in accordance with hereinbefore described technique. To increase the charts informativity the data preliminarily were transformed using $log2(cpm)$ function. The number of genes was reduced at this step from 25348 to 7435. The analysis of the obtained in Figure 4.13 diagrams allows concluding that level of genes expression informativity significantly increased due to remove lowly expressed values. The same conclusion can be done based on the box plots analysis (See Figure 4.14). In the case of filtered data use, the values of gene expressions for all samples are distributed more evenly and they are shifted to the side of larger values.

The next step of the data preprocessing is their normalizing. Figure 4.15 shows the chart of the clustering quality criterion (4.3) versus the normalizing method. To calculate this criterion values the data previously were divided into clusters considering the Mn dose. It should be noted that in the case of non-normalized filtered data the value of this criterion was 100.05.

The analysis of the obtained results allows concluding that normalizing process significantly increases the quality of the data in terms of the quality criterion (4.3). The value of this criterion for non-normalizing data 100.05 and it has been decreased more than 10 time. Comparison analysis of various normalizing methods has shown that the easiest *lcpm* method is showed the worst results in comparison with other methods. The difference between methods *TMM*, *TMMwsp*, *RLE* and *Upper quartile scaling* is very small, however, the value of the criterion (4.3) achieved the minimum one in the case of *Upper quartile scaling* method apply. This fact indicates the reasonable of this method use for normalizing the current type of data.

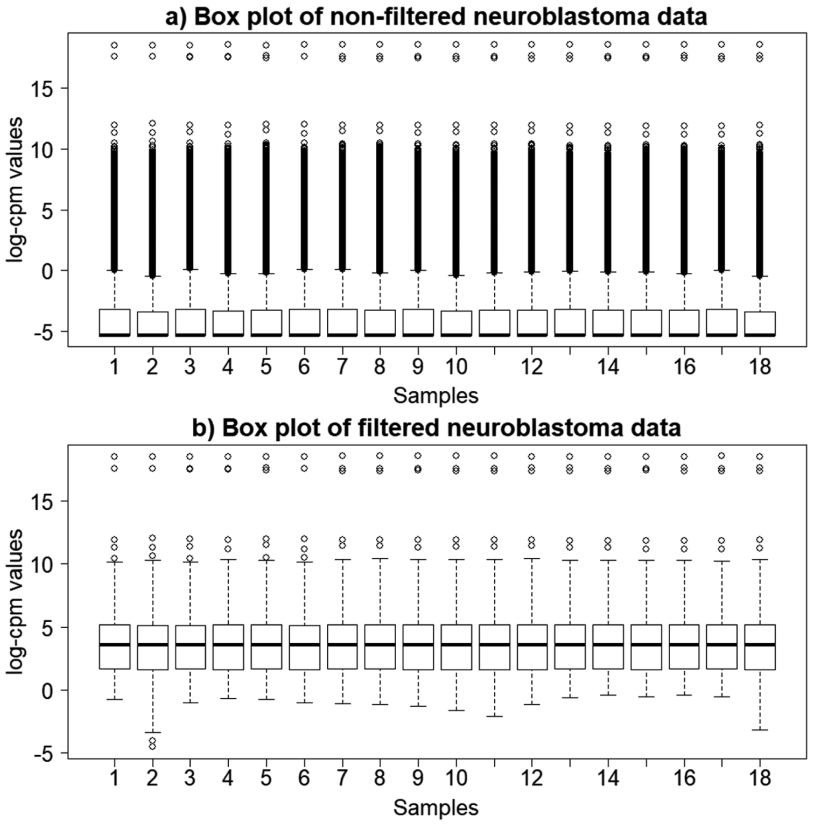

Figure 4.14: Box plots of non-filtered and filtered gene expression values distribution for neuroblastoma data samples

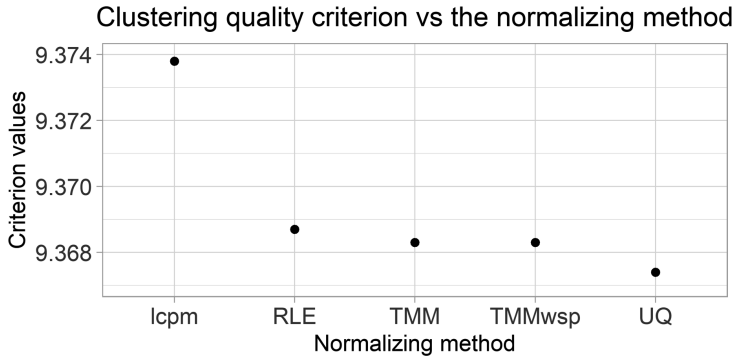

Figure 4.15: Dot plot of the quality criterion VS the normalizing method

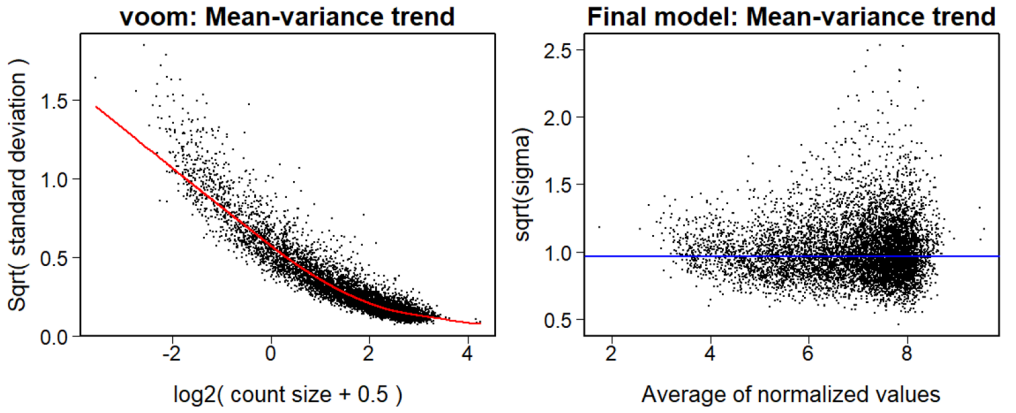

Figure 4.16: Visualization of heteroscedasticity removing from the data

At the next step of this procedure implementation, it is necessary to remove heteroscedasticity from the data. The analysis of the normalized data has shown that in the case of RNA-seq data use, the variance values have not depended on the mean values. Methods which are based on counts with the use of Negative Binomial distribution are based on a quadratic mean-variance relationship. In *limma* package of *R* software, linear modelling is performed using the normalized values. In this case the data should be normally distributed and the mean-variance relationship is evaluated with the use of precision weights calculated by the *voom*() function. Figure 4.16 presents the results of this step implementation. The left chart in Figure 4.16 shows the mean-variance relationship of normalized gene expression values. Usually, the voom-plot shows a decreasing trend between the means and variances which are appeared due to an existence of both the technical incorrectness during the sequencing experiment performing and the biological variation among the replicate samples from various cell samples. Typically, the results of the experiments with high level of biological variation are presented as a flatter trends. The variance values in this case are not significantly changed for high expression values (right chart in Figure 4.16). And otherwise, experiments including data with low biological variation usually have tend to sharp decreasing the variance values. Moreover, the voom-plot allows us to visual evaluate the quality of gene expression filtration process. If filtration process of lowly-expressed genes is insufficient, then, the variance values should be decreased at the low end of the expression scale due to very small gene expression values.

In order to visual summarize the results for all genes in obtained groups, we create a mean-difference plots using the plotMD function of *limma* package. These plots allow us to display log-Fold-change values from the linear model which can be

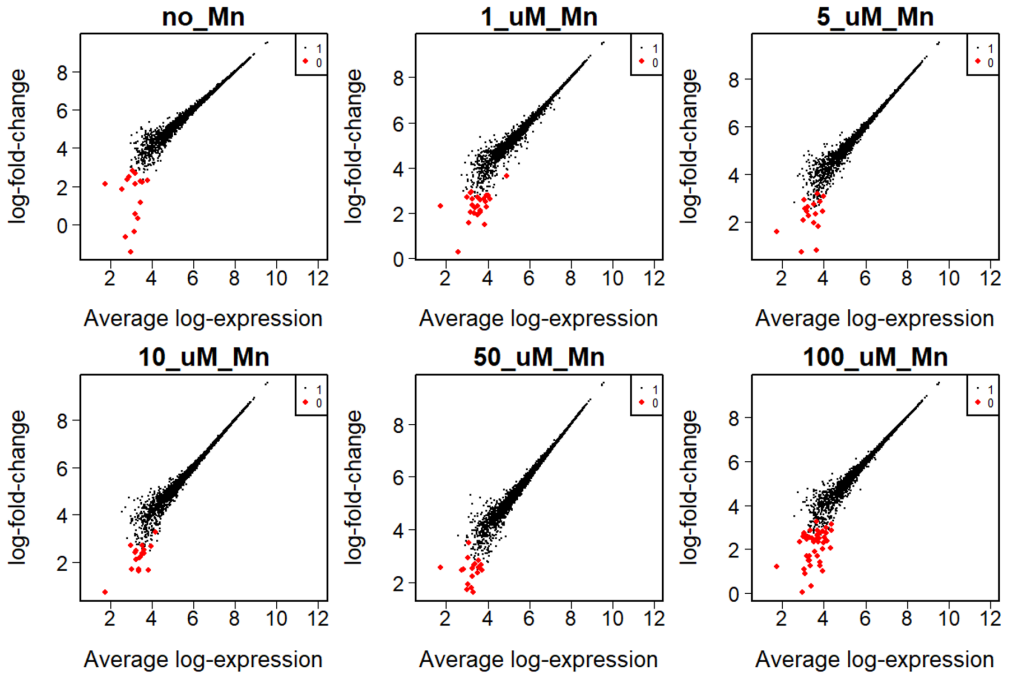

Figure 4.17: Mean-difference plots of gene expression profiles for investigated groups

fitted against the average of log-expression values. These charts allow us to identify differentially expressed genes. The charts are shown in Figure 4.17.

The result of visual analysis of the obtained diagrams allows concluding the greatest number of genes in investigated groups have high level of differentially expression (black colour or number 1). It means that these genes are informative to distinct the samples for the further processing. However, the data contains some quantity of lowly-expressed genes (red colour or zero number). It is means that these data need the following processing for purpose of non-informative genes reducing in terms of various quantitative criteria.

Figure 4.18 shows the box charts of the processed gene expression profiles. The samples previously were reordered considering Mn dose from 0 to 100 μM. Analysis of character of gene expressions distribution allows us to conclude about correctness of data preprocessing step implementation. The values of gene expression have the same and not so much ranges, all genes are enough highly-expressed for all of the samples. However, it should be noted, that there is some quantity of lowly-expressed genes (for example, in Mn_1 sample). This fact indicates about necessity the further data processing on the basis of the use of current techniques of complex

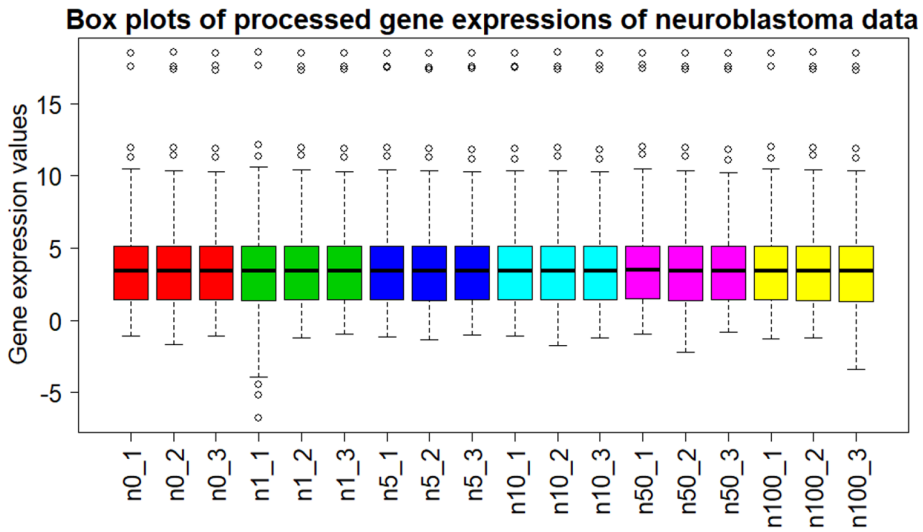

Figure 4.18: Results of gene expression profiles of neuroblastoma data processing

data processing.

4.2.3 Technique of Non-informative Genes Expression Profiles Reducing

The gene expression profiles reducing in terms of the statistical criteria and Shannon entropy is one of the stages of the hereinbefore presented information technology implementation. It is assumed that if variance or average of absolute value of gene expression profiles is less than appropriate boundary values, or if Shannon entropy of the appropriate gene expression profile is greater than the boundary value, then these profiles are not informative and they can be removed from data without significant loss of useful information. However, there is a problem of determining the boundary values of the appropriate criteria, which allow us to divide objectively the gene expression profiles into informative and non-informative ones. In addition, the use of these criteria independently of each other is not objective because the non-informativeness of the corresponding gene according to one of the criteria does not mean that this gene is not informative according to other criteria. The complex use of all criteria for determining the level of gene expression profiles informativity is rational in this case. A gene, identified as non-informative based on the use of all criteria can be removed from the data without significant loss of useful information.

In papers [13, 24] this problem is solved with the use of fuzzy logic technique

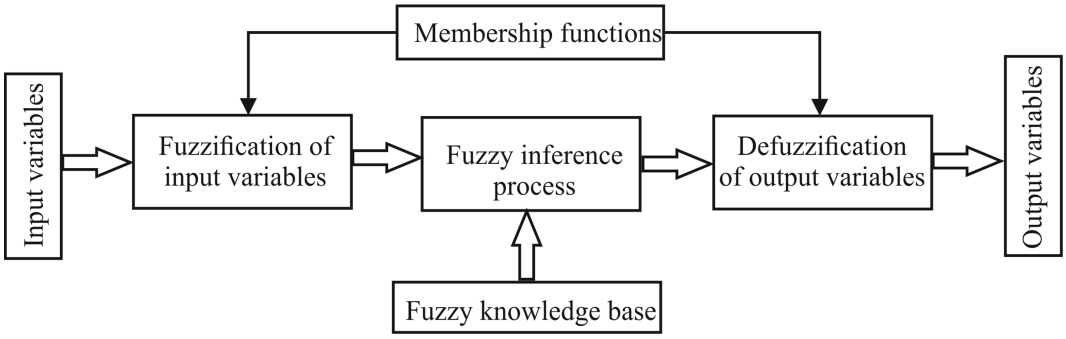

Figure 4.19: Structural block diagram of fuzzy inference process

[144], which involves possibility for complex estimation of various criteria influence on integral parameter, which determines the level of the corresponding gene informativity.

Stages of fuzzy inference process

Implementation of fuzzy inference process assumes transforming the values of the input variables into the values of the output variables using fuzzy rules, which are formed by experts in this subject field. The block diagram of fuzzy inference process is shown in Figure 4.19. The stage of knowledge base formation involves:

- formation of sets of input $X = \{x_1, ..., x_n\}$ and output $D = \{d_1, ..., d_m\}$ variables;
- formation of the basic term-set with the corresponding membership functions for each term: $A = \{a_1, a_2, ..., a_i\}$;
- formation a set of fuzzy rules agreed with the variables used:

$$\bigcup_{k=1}^{m} [\bigcap_{i=1}^{n} (x_i = a_i^k), for\ \omega_k] \longrightarrow D = d_k$$

  where $k = \overline{1, m}$ is the quantity of logical statements, $i = \overline{1, n}$ is the quantity of terms used, ω_k represents the weight of the k-th statement.

The fuzzification stage involves establishing the correspondence between the specific values of the input variables of fuzzy inference system and the values of the membership function of the term corresponding to the given variable. For this, the membership functions defined on the input variables are applied to their actual values. In other words, the values of the membership functions $\mu^{a_i^k}(x_i)$ versus the

variables x_i for the corresponding terms a_i^k are determined. The fuzzification stage is completed when all values of the membership functions $b_i = \mu(a_i)$ for all rules which are included in this knowledge base of the fuzzy inference system are found.

The process of fuzzy inference includes the following steps: aggregation of prerequisites, activation, accumulation of fuzzy rules inferences and defuzzification. Depending on the method of the fuzzy logic process using the classical fuzzy model can be implemented on the basis of the following algorithms: Mamdani, Sugeno, Larsen and Tsukamoto. The choice of model type is determined by the nature of the data used. In the case of creating a fuzzy model for evaluating the quality of gene expression profile on the basis of statistical and entropy criteria, the values of the input functions do not require scaling, all membership functions can be isotropic, resulting membership function can be represented as a simple set and the implementation of the defuzzification process does not require the use of a special functional. In this case, it is reasonable the use of Mamdani fuzzy inference algorithm, implementation of which involves the following:

- determination of the "cutting" levels for prerequisite of each rule using the operation min at the stage of aggregation:

$$\alpha_k = min(\bigwedge_{i=1}^{n}[\mu^{a_i^k}(x_i)])$$

- determination of the degree of truth for each of the statements of the fuzzy inference rules. At this stage the truncated membership functions of the corresponding fuzzy sets are determined:

$$\mu'_k(D) = (\alpha_k \wedge \mu_k(D)),$$

  where $\mu_k(D)$ is the membership function of the output variable corresponding to the statement k, $\mu'(D)$ is the truncated function of the membership of the output variable for the statement k;

- determination of the membership functions for the final fuzzy subset for the output variable, which correspond to the appropriate combination of input variables:

$$\mu_\Sigma(D) = \bigvee_{k=1}^{m}[\mu'_k(D)]$$

- determination of the crisp value of the output variable for the corresponding combination of input variables as the mass center of the obtained final truncated membership function of the output variable.

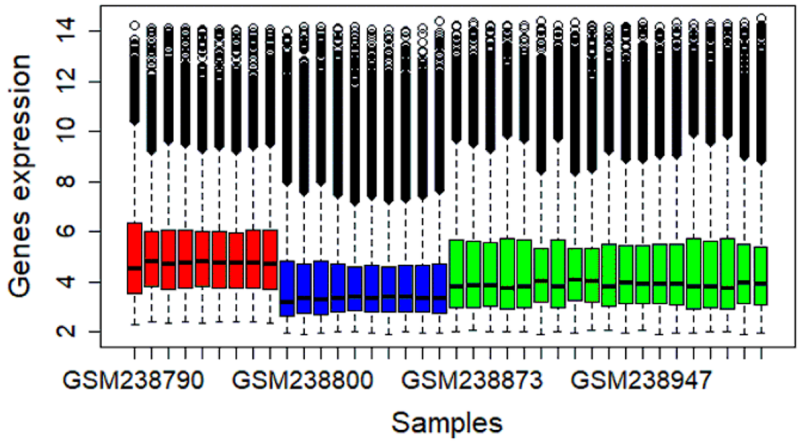

Figure 4.20: Boxplots of the investigated samples

Experiment

DNA microarrays data from database KEGG [84, 86, 85] were used as the experimental ones during the simulation process. This data contains 38 of DNA microarrays of patients which were investigated on Alzheimer disease [97, 98]. The DNA microarrays contain the information concerning genes expression of brain samples from three Alzheimer's Disease Centers. The first data contained 9 samples from entorhinal cortex (EC) of brain. The second data contained 10 samples from hippocampus (HIP) of brain. The third data contained 19 samples from primary visual cortex (VCX) of brain. Each of the samples contained 54675 of genes. So, the initial matrix of genes expression contained 38 of rows (samples) and 54675 of columns (genes). Gene expression profile in this case is presented as a vector of genes expression which are determined for different samples. The character of the genes expression values distribution in the investigated samples is presented in the Figure 4.20. The analysis of the Figure 4.20 allows us to conclude that the gene expression profiles can be divided into three distinguish clusters in dependence on type of the disease.

Three criteria were used for division of the gene expression profiles into informative and non-informative: variance, average of absolute values and Shannon entropy. The main idea for this process implementation is the following: if variance and average of the absolute values are less and Shannon entropy is larger than the appropriate boundary values, then this profile can be removed from the studied data as non-informative. In this case the gene expression values for various samples do not allow us to recognize the investigated samples. The Shannon entropy criterion was

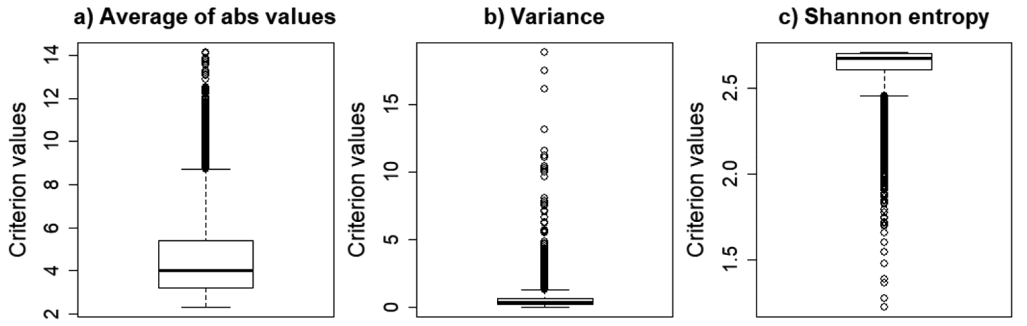

Figure 4.21: Boxplots of the statistical and Shannon entropy criteria distribution

Table 4.1: Statistical analysis of the used criteria distribution

Crit	**Min**	**Quart 1**	**Med**	**Quart 3**	**Max**
Abs	2.3	3.2	3.99	5.41	14.1
Var	0.01	0.2	0.39	0.64	18.8
Entr	1.22	2.61	2.67	2.7	2.71

calculated based on James-Stein shrinkage estimator [67]. This method is based on the complex use of two different models: a high-dimensional model with low bias and high variance, and a low-dimensional model with larger bias but smaller variance. The character of these criteria distribution in the studied gene expression profiles is presented in Figure 4.20 and Table 4.1. Figure 4.22 shows the membership functions for both the input and output (quality of gene expression profiles) variables. As it can be seen, three linguistic terms for input variables (Low, Medium and High) and five for output parameter (Very low, Low, Medium, High, Very high) were used within the framework of the proposed fuzzy inference system for purpose of the fuzzy rules formation. The range of the input variables change was divided into fifty equal sections during the simulation process. Implementation of the simulation process involves step-by-step increasing both the variance and average of the genes expression from minimum to maximum values and decreasing the Shannon entropy from maximum to minimum value. The value of the output parameter (quality) was estimated for each of the investigated gene expression profiles with the use of the fuzzy inference system. Conditions for division of the gene expression profiles into informative and non-informative are the following:

$$var \leq var_{lim}; \quad abs \leq abs_{lim}; \quad entr \geq entr_{lim}; \tag{4.4}$$

If the conditions (4.4) are true, the gene expression profile is removed from the

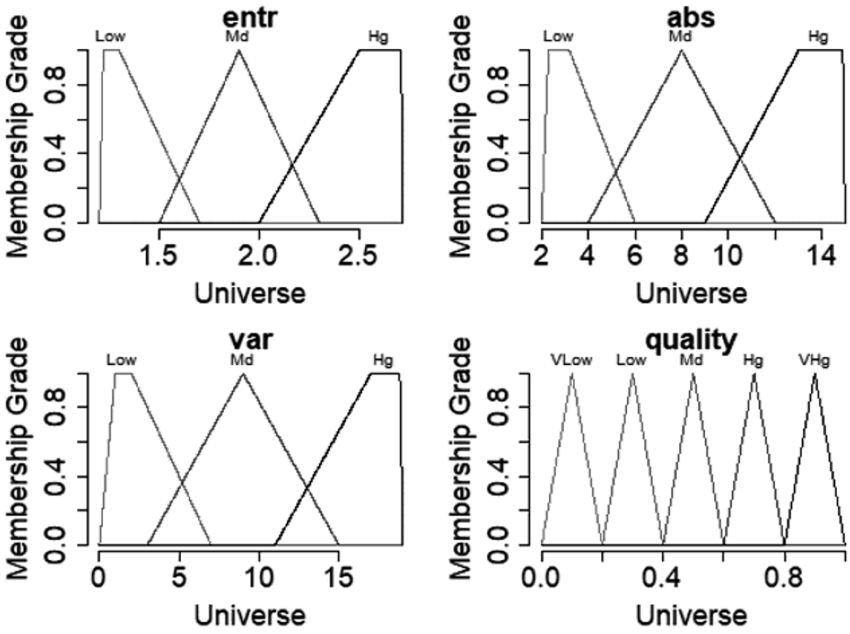

Figure 4.22: Membership functions of the fuzzy inference system

database as non-informative. Otherwise, this profile is identified as informative and it is used for the following processing.

The next stage of the simulation process implementation involved division of the investigated objects which include only informative gene expression profiles into three clusters in accordance with the type of the data. The calculation of the clustering quality criterion was performed at this stage. The correlation distance was used as the metric to estimate the proximity level of the investigated vectors:

$$d(A,B) = 1 - \frac{\sum_{i=1}^{m}(x_{ai} - \bar{x}_a)(x_{bi} - \bar{x}_b)}{\sqrt{\sum_{i=1}^{m}(x_{ai} - \bar{x}_a)^2} \times \sqrt{\sum_{i=1}^{m}(x_{bi} - \bar{x}_b)^2}} \tag{4.5}$$

where m is the number of the informative genes; A and B are the studied objects; x_{ai} and x_{bi} are the expressions of the i-th gene for A and B objects respectively, $\bar{x}$ is the average value of genes expression of the appropriate vector. The clustering quality criterion (4.3) was used to evaluate the character of the data dictribution in the clusters during reducing process implementation.

Figure 4.23 presents the structure block-chart of the algorithm for the investigated data processing within the framework of the proposed technique. Its implementation involves the following steps:

1. Formation of the vectors of the fuzzy inference system input parameters: variance (var), average of gene expression profiles absolute values (abs) and Shannon entropy ($entr$). Setup of both the ranges and steps ($dvar$, $dabs$, $dentr$) of these parameters change.

2. Setup of the fuzzy inference system. Formation of the basic term set of the membership function for both the input and output variables and the set of fuzzy rules agreed between input and output parameters.

3. Initialization of the fuzzy inference system initial parameters: $var_1 = var_{min}$; $abs_1 = abs_{min}$; $entr_1 = entr_{max}$. Setup of the counter initial value of the fuzzy inference process implementation: $m = 1$.

4. Implementation of the fuzzy inference process for current values of the input parameters. Determination of the output parameter values (quality of gene expression profiles).

5. Division of the gene expression profiles into informative and non-informative taking into account the input parameters boundary values at appropriate stage of this process implementation according to the condition (4.4).

6. Formation of the clusters. Clustering quality criterion calculation with the use of the formuls (4.3).

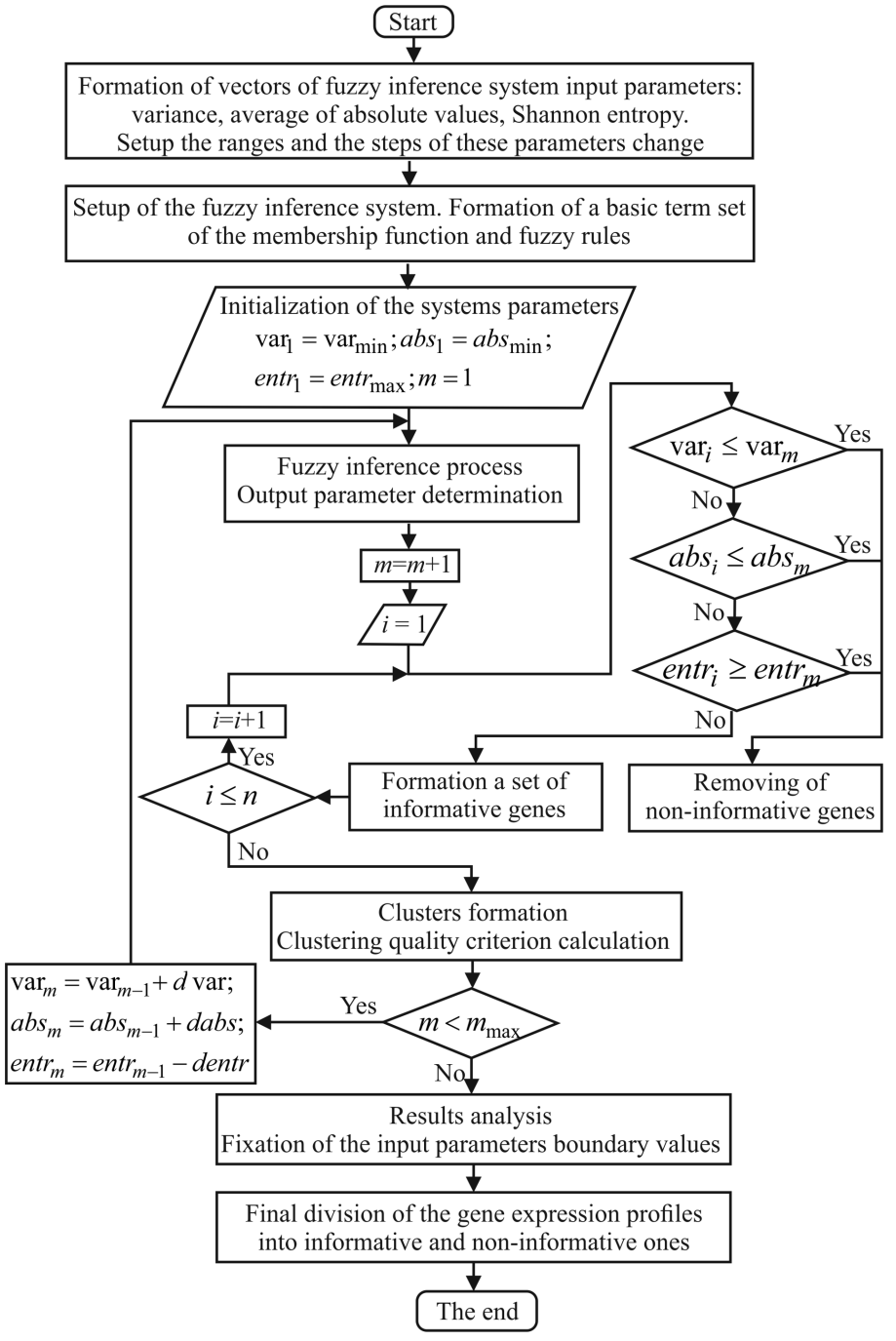

Figure 4.23: Structural block-chart of the algorithm for gene expression profiles reducing

7. If the counter value of fuzzy inference process implementation is less than maximum value, increasing of the boundary values of the input parameters and go to the step 4 of this procedure. Otherwise, results analysis and determination of the optimal boundary values of the input parameters which correspond to the minimum value of the clustering quality criterion.

8. Reducing of the gene expression profiles with the use of the optimal boundary values of the variance, Shannon entropy and average of the gene expression profiles absolute values.

Results and discussion

Figure 4.24 presents the results of the fuzzy inference system simulation within the framework of the proposed model. As was described hereinbefore, the ranges of the input parameters were divided into fifty equal sections. The variance and the average of absolute values of the gene expression profiles were changed from minimum to maximum values (Figure 4.24a,b) and the value of Shannon entropy was changed from maximum to minimum ones (Figure 4.24c) during the simulation process. The value of the output parameter (quality of the gene expression profiles) was calculated at the each step of this process implementation (Figure 4.24d). Table 4.2 presents the linguistic estimates which vere used within the framework of the fuzzy inference system. The logistic operator *And* was used during fuzzy rules creation. The charts of both the clustering quality criterion and the number of the informative gene expression profiles versus the step of the fuzzy inference process implementation are presented in Figure 4.25.

The analysis of the charts allows us to conclude that the number of the informative gene expression profiles is decreased monotonically during the change of the fuzzy inference system input parameters boundary values. At the same time, the clustering quality criterion value is changed chaotically. As it can be seen from Figure 4.25a, the value of this criterion has three local minima during the simulation process. It means that the used gene expression profiles in these cases allow us to distinguish the investigated objects better in comparison with other cases. The number of the informative gene expression profiles in these cases are the following: 615 at 31-*st* step; 222 at 35-*th* step; 35 at 42-*nd* step. The quality of the gene expression profiles which were determined using fuzzy inference system belong to the range from medium to high values in the first and in the second cases. In the third case (step = 42) the quality of the gene expression profiles was identified as very high one. This fact indicates the highest level of the genes expression profiles informativity in terms of separating ability of the investigated objects. The boundary values of the fuzzy inference system input parameters are the following: $var = 11.52$, $abs = 9.55$, $entr = 1.8$ in the first case ($step = 31$); $var = 13.05$, $abs =$

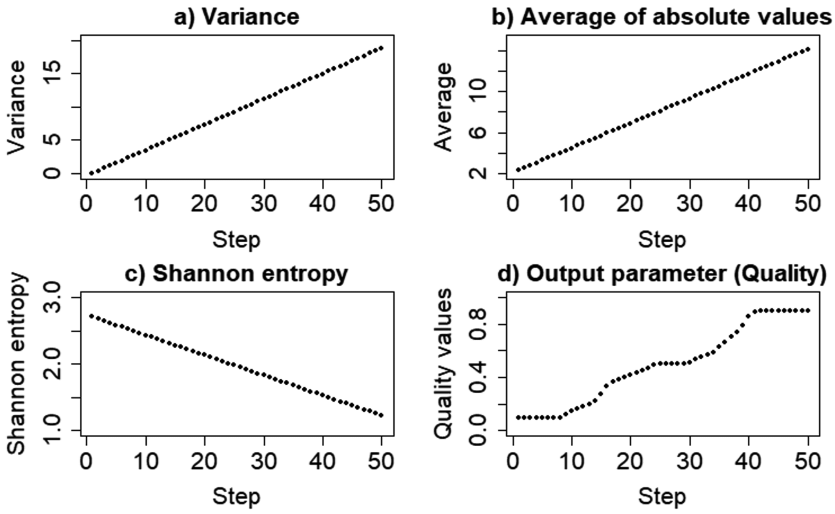

Figure 4.24: Results of the fuzzy inference system simulation

Table 4.2: Linguistic estimates of the input and output parameters

Number of rule	Variance	Average of absolute values	Shannon entropy	Quality
1	Low	Low	Hg	VLow
2	Md	Low	Hg	Low
3	Low	Md	Hg	Low
4	Md	Md	Hg	Low
5	Low	Low	Md	Low
6	Low	Md	Md	Md
7	Md	Low	Md	Md
8	Md	Md	Md	Md
9	Hg	Low	Md	Md
10	Low	Hg	Md	Md
11	Hg	Hg	Low	VHg
12	Md	Hg	Low	Hg
13	Hg	Md	Low	Hg
14	Md	Md	Low	Hg
15	Hg	Hg	Md	Hg

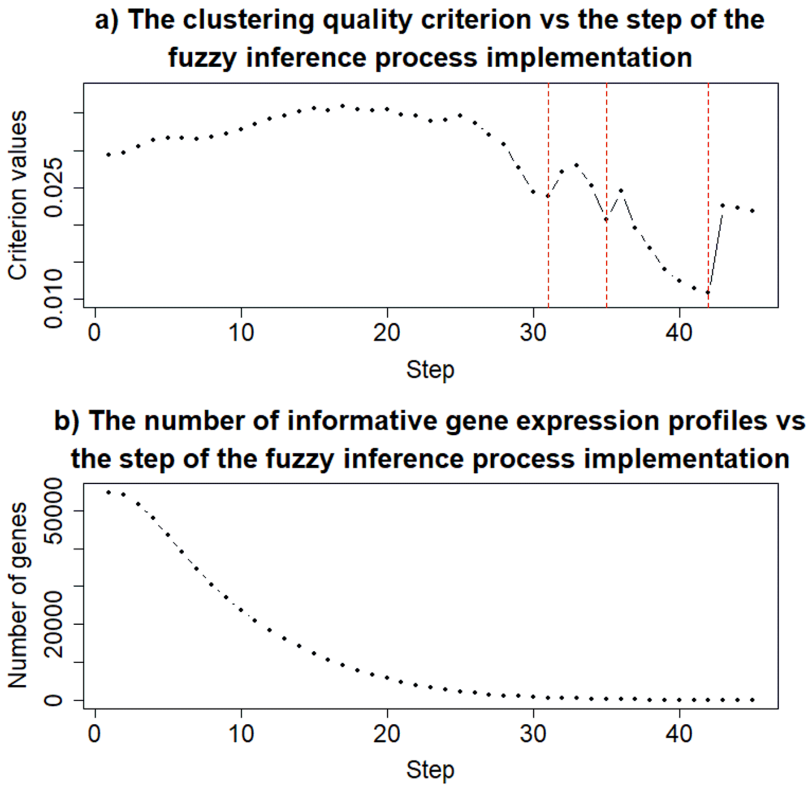

Figure 4.25: Charts of both the clustering quality criterion (a) and the number of the informative gene expression profiles (b) vs the step of the fuzzy inference process implementation

10.31, $entr = 1.679$ in the second case ($step = 35$); $var = 15.74$, $abs = 12.2$, $entr = 1.467$ in the third case ($step = 42$). The choice of the boundary values of the used parameters is determined by the aims of the current task. In the case of the further reconstruction of genes regulatory network the third case is the optimal one since the number of the allocated genes allows us to reconstruct the genes network qualitative for purpose of the following investigation of the character of genes interconnection taking into account the status of the object. In the case of the step-by-step gene expression profiles clustering and biclustering for purpose of genes regulatory networks reconstruction the first or the second cases are the optimal ones since the number of the allocated gene expression profiles allows us to implement the gene expression profiles clustering and biclustering for both the reconstruction of genes regulatory networks and simulation of the reconstructed models.

4.2.4 Conclusions

In this chapter, we have presented the results of the research concerning gene expression array formation and following gene expression profiles reducing in order to select the most informative genes in terms of statistical criteria and Shannon entropy. In the first part of this chapter, we have proposed the technique of gene expression array formation which were obtained based on DNA microarray experiments. The initial data is presented as a set of DNA microchips, each of which contains the matrix of light intensities, the values of which are proportional the expression values of the appropriate genes. Four stages have been performed during the simulation process: background correction, normalization, PM correction and summarization. Each of the stage assumed the use of different methods. The Shannon entropy criterion which is calculated based on James-Stein shrinkage estimator has been used as the main criterion to estimate the genes expression informativity. The simulation process has been performed based on R software with the use of Bioconductor package functions. The lung cancer patients' gene expression profiles E-GEOD-68571 from database ArrayExpress have been used as the experimental data during the simulation process. The results of the simulation have shown that the optimal combination of the methods in terms of the minimum value of the Shannon entropy is the following one: *rma* background correction method, *invariant set* normalization method and *mas* methods PM correction and summarization. This combination of the methods has been used to process the investigated DNA microchips. The boxplots of both the non-processed and processed data have been created as the simulation results. The analysis of the obtained results has shown that the values of the largest quantity of gene expressions for various objects lie in a very narrow range. It means that these genes are responsible for the functions that are inherent for all investigated objects. However, each of the investigated samples contains genes, the expression of which goes beyond the inter-quartile range. This

fact can mean that these genes are very important for the following research since they allow us to distinguish the investigated objects by their particularities.

Then, we have presented the results of the research to process the results of RNA-molecules sequencing experiments. The dataset GSE129336 generated from Gene Expression Omnibus database was used as the experimental one. This data contains the results of expression profiling by high throughput sequencing in human SH-SY5Y neuroblastoma cells. The initial data matrix contained counts of expressed genes for studied samples. At the first step, we have removed lowly-expressed genes. The number of genes was changed from 53186 to 7435. Then, we have compared various normalizing technique using clustering quality criterion as the main criterion of appropriate normalizing method effectiveness estimation. At the next steps we have analyzed the obtained results using various visualization techniques. The analysis of the processed genes expression values distributions allows concluding about high effectiveness of the proposed technique, since its implementation allows allocating a set of similarly distributed highly-expressed genes for the following processing.

Finally, we have proposed the hybrid model of gene expression profiles reducing based on the complex use of fuzzy inference system and clustering quality criterion. This model is presented as the algorithm of step-by-step data processing. The variance, the average of absolute values and Shannon entropy of gene expression profiles have been used as the boundary criteria for division of the gene expression profiles into informative and non-informative ones. Three groups of gene expression profiles of the patients which were investigated on Alzheimer disease have been used during the simulation process. The first data contained 9 samples from entorhinal cortex of brain. The second data contained 10 samples from hippocampus of brain. The third data contained 19 samples from primary visual cortex of brain. Each of the samples contained 54675 of genes. The simulation process was performed in the following way. Firstly, the vectors of both the statistical criteria and Shannon entropy for the investigated gene expression profiles have been formed. The ranges of these criteria changes were divided into fifty equal sections and these values have been used for setup of the fuzzy inference system parameters. Then, the values of variance and average of absolute values of the gene expression profiles were changed monotonically from minimum to maximum and Shannon entropy from maximum to minimum values within the range of these parameters variation. The values of both the quality of gene expression profiles and the clustering quality criterion have been calculated at the each step of this process implementation. The optimal boundary values of the used criteria (var, abs, entr) were determined based on local minima of the clustering quality criterion. The results of the simulation have been shown that the best values of the used parameters in terms of the minimum value of the clustering quality criterion are the following: $var = 15.74$, $abs = 12.2$, $entr = 1.467$. In this case 35 of gene expression profiles were allocated. These genes can be used

for both reconstruction of genes regulatory networks and simulation of the obtained models.

However, it should be noted that the solution concerning determination of the parameters boundary values which correspond to the clustering quality criterion minimum values should be done taking into account both the aims of the current task and the type of the used data. If previously division the samples into clusters is problematic, we can select the genes for the following processing considering number of the genes expression profiles which should be used for the following processing with the use of cluster-bicluster analysis. In any case, the propsed technique allows selecting informative genes in terms of various quantitative criteria objectively.

Chapter 5

Gene Regulatory Networks Reconstruction

5.1 Introduction

In most cases, current systems of information processing are based on the use of analogies of biological processes functioning which are occurred in living organisms. Such systems are the follows: immune system of organism, neural network, gene network, et.al. Their particularities are high level of complexity, ability of self-learning, ability to information recognizing and decisions making, decentralized parallel information processing. In this reason, development of current artificial intelligent models should be carried out based on the complex approach considering complex use of techniques of molecular biology, mathematics, informatics, physics, etc. Implementation of this approach creates the conditions for better understanding particularities of operation of the biological system in order to influence to this process.

Reconstruction of gene regulatory networks and further simulation of the reconstructed models form the basis for investigation and analysis of both the character of molecular systems elements interconnections and influences of these interconnections to functional possibilities of the investigated objects. The complexity of gene networks reconstruction is determined by the follows: the experimental data which are used for the reconstruction process usually does not allows defining the network structure and pattern of genes interconnection in the network. Moreover, large quantity of genes complicates the interpretation of the network elements interconnections. In this case, it is necessary to conduct research concerning evaluation of both the network topology and the pattern of genes interconnection in network with the use of experimental data obtained by the use of both the DNA microchip experiment or RNA-molecules sequencing method. Qualitatively reconstructed gene regulatory network allows investigating the pattern of the biological organism development at

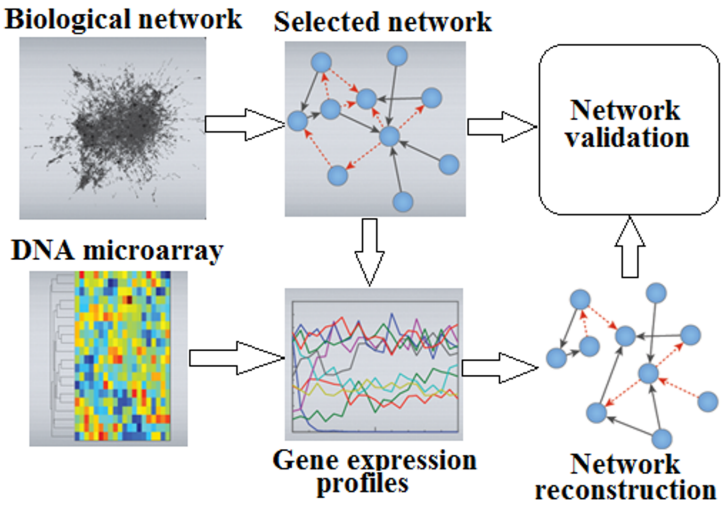

Figure 5.1: Structural block chart of a general process of reconstruction and validation of the model of a gene regulatory network

the genetic level. It creates the conditions for both making new effective medicines and development of methods of early diagnostics and effective treatment of complex diseases. This fact indicates the actuality of the research in this subject area.

There are various databases of biological gene regulatory networks (GRN) of different organisms [99] that allows visualizing and exploring a network topology of the appropriate biological object. One of currents methods employed for the reconstruction of GRN is the identification of a network by comparing it with a known network of the relevant biological organism. In this case, it is necessary to determine the quality criteria for evaluation of the network topology considering appropriate topological parameters and existence of both the genes and links between corresponding genes in reconstructed and biological selected networks. Another important stage when reconstructing a network is the validation process of the reconstructed models based on quantitative criteria for assessing the network quality. A structural block chart of a general process of these procedures implementation aimed to the reconstruction and validation of GRN is shown in Figure 5.1. As it can be seen from Figure 5.1, the process of GRN reconstruction and validation assumes a comparative analysis of parameters of the reconstructed and biological network with the purpose of determining an optimal network topology and the character of interrelations between relevant nodes.

5.2 Literature Review and Problem Statement

Current biological system of living organisms is a complex dynamic network of interacting elements with different purposes whose state can change under the influence of external conditions [106]. Reconstruction and simulation of gene networks are rather complicated tasks that do not have an unambiguous solution nowadays. The first studies concerning reconstruction of biological networks based on experimental data were published at the end of the 90s of the last century [45, 96, 56, 39]. These papers proposed several approaches focused to a given type of simulation. Studies [138, 105, 48] reviewed several methods related to the reconstruction and simulation of gene regulatory networks based on gene expressions data. The authors considered in detail stages of GRN reconstruction and they also performed a comparative analysis of different techniques with outlined advantages and shortcomings of the respective method.

Nowadays, there is sufficient information about the properties of gene regulatory networks in natural biological systems. So, in papers [141, 4] the authors formulated the rules for GRN reconstruction and simulation, which allows significantly limiting the dimensionality of search space for the optimal network. The sparsity property is the most common and important feature of GRN. This property means that the topology of GRN is sparse, meaning that each gene has a small number of regulatory inputs [109]. However, it should be noted that there are a small number of genes (master-genes) which are able to control hundreds of other genes. A given property is used to limit the search space of the optimal solution by limiting the number of regulatory links. Papers [101, 100] show that the frequency distribution of the number of regulatory inputs of nodes at a gene network of biological systems is often governed by the law of Pareto distribution. This means that in the case of a non-scalable network most genes are loosely linked, but there are several nodes with a high number of links. These nodes are named nodes-concentrators. They correspond to genes that perform most of the overall regulation of other nodes in the network. The existence of a concentrator leads to the localization of the network, since all the nodes in the network are connected to concentrators via short links, the quantity of which is limited. In addition, the concentrators improve stability of thenetwork to external influences and various kinds of fluctuations, since they link a network and do not alow splitting the network into separate fragments.

The next important property of GRN is modularity. This property means that genes in the network cannot be regarded as independent elements. In a general case, genes can be divided into functional, perform the function of control over other genes (concentrators), and genes which function concordantly performing a joint function. It is obvious that in this case genes can be grouped in modules or clusters depending on their functional similarity.

Thus, based on hereinbefore presented analysis, we can conclude that solution the problem of gene regulatory network reconstruction based on gene expression profiles with optimal topology, which corresponds to biological gene network, is one of the current direction of modern bioinformatics.

5.3 Topological Parameters of Gene Regulatory Network

The process of gene regulatory network reconstruction based on gene expression profiles assumes a possibility of creation of different network topologies which differ from each other by the number of nodes, the number of arcs that connect respective nodes, and the character of bonds between the nodes. As a result, it is necessary to develop the technique to evaluate a network topology using quantitative quality criteria. This allows us to select reasonably the optimal network topology for respective biological object. Figure 5.2 shows an example of yeast gene regulatory network topology [127].

Analysis of the structure and topology of gene regulatory network allows us to conclude that it can be presented as a directed or a non-directed graph whose arcs can have weight coefficient (if the case of presence of weights which determine the strength of the connections). Therefore, the graph theory can be used to define the parameters of the network topology. Classification of basic topological parameters of GRN is shown in Figure 5.3 [5]. The follows simple topological parameters can be used to evaluate the network topology:

- *Number of network nodes.* This parameter determines the general number of genes (nodes) interconnected between each other.

- *Degree of a network node or its connectivity* is the total weight of connections (arcs) that connect a given node with neighbour nodes:

$$k_i = \sum_{j=1, j \neq i}^{n_i} w_{ij} \tag{5.1}$$

  where n_i is a number of neighbour nodes of i-th node; w_{ij} is the weight of arc which connect the nodes i and j.

- *Average of degree or average of connectivity* is determined as an average of degrees of all network nodes:

$$\overline{k} = \frac{1}{n} \sum_{i=1}^{n} k_i \tag{5.2}$$

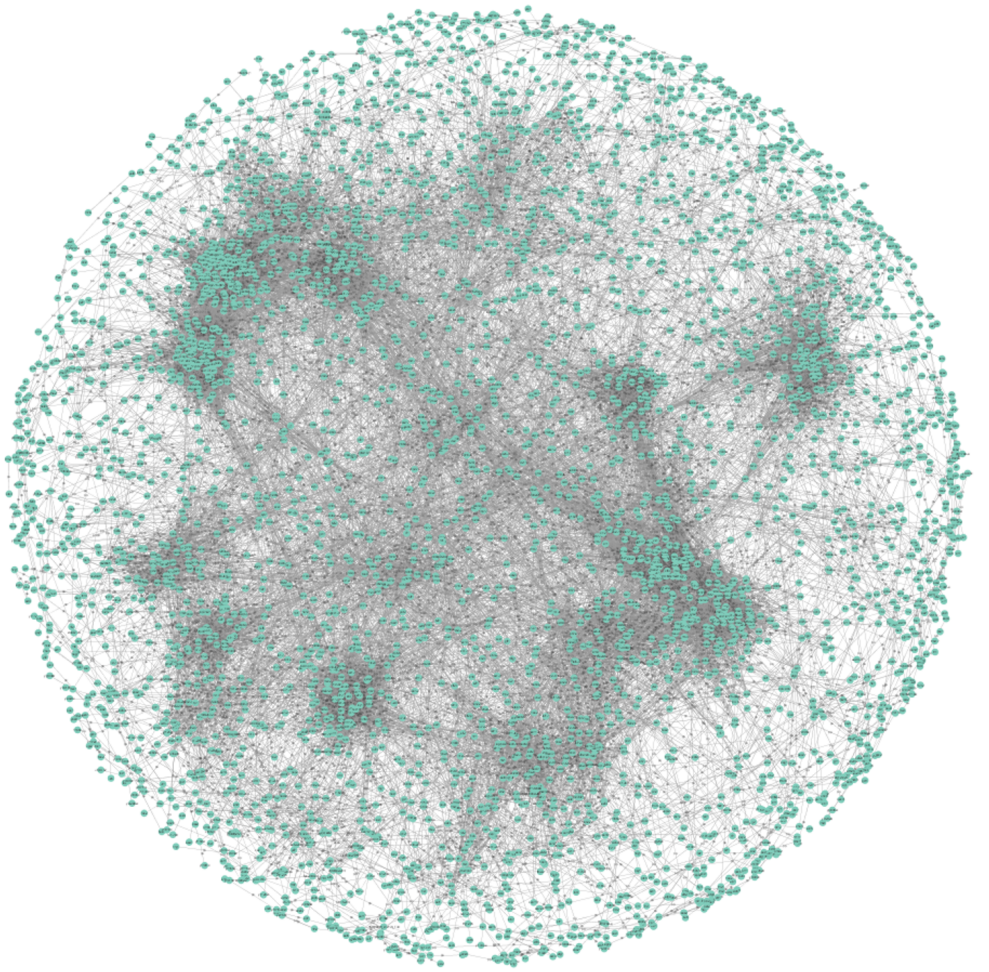

Figure 5.2: An example of yeast gene regulatory network

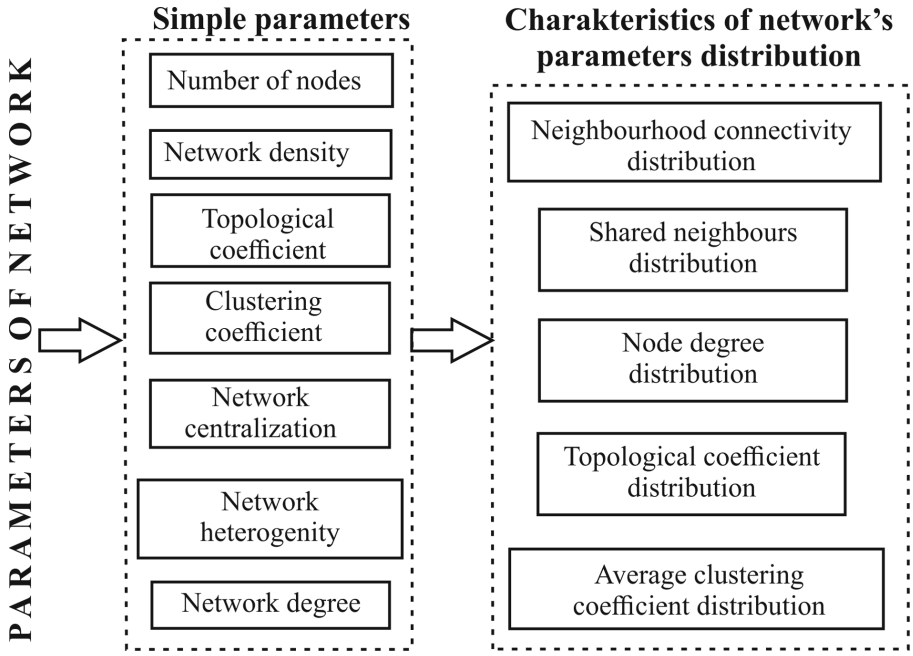

Figure 5.3: Classification of gene regulatory networks topological parameters

- *Maximal degree* determines the maximal value of elements of the connectivity vector for all nodes in the network:

$$k_{max} = max(\overline{k_1, k_n}) \tag{5.3}$$

  The high value of this parameter indicates a high level of complexity, since all network nodes have a large number of connections with their neighbours. This fact complicates the interpretation of the reconstructed network. It is obvious that the minimal value of this parameter is optimal if the number of the network nodes is maximal one.

- *Network density* is defined as the ratio of the number of weighted connections between the nodes to the maximal number of connections between the nodes in the network:

$$DS = \frac{2 \cdot \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} w_{ij}}{n \cdot (n-1)} \tag{5.4}$$

  The density value is varied from 0 to 1. If $DS = 0$, then there is no connection between appropriate nodes, in the case of $DS = 1$ we have a fully connected network.

- The *clustering coefficient* of a node determines the probability that the nearest neighbours of a current node are directly connected between each other. The network clustering coefficient is defined as the average of the clustering coefficients of all network nodes:

$$CL = \frac{1}{n}\sum_{i=1}^{n}\frac{c_i}{0.5 \cdot k_i(k_i - 1)} \tag{5.5}$$

  where n is a number of the network nodes, c_i is a number of connections of i-th node with neighbour nodes, k_i determines the numbers of neighbours of i-th node including this node which can make up a complete cluster. This parameter is a quantitative measure of the network connectivity. If its value is unity, the network is fully connected. In the case of zero values, there are no connections between the neighbours of the network nodes.

- The network *centralization* parameter determines the degree of proximity to the star topology:

$$Centr = \frac{n}{n-2}(\frac{k_{max}}{ni1} - DS) \tag{5.6}$$

  If the network topology looks like a lattice, where all nodes are connected equally, the value of this parameter is zero. A higher value of the centralization parameter indicates a higher degree of network similarity to a star topology.

- Network *heterogeneity* determines the degree of non-uniformity of the network topology and it is expressed using the variance and mean values of the average degree of nodes as follows:

$$GT = \frac{\sqrt{\overline{k}}}{mean(\overline{k})} \tag{5.7}$$

  Homogeneous network has zero heterogeneity value and otherwise, increasing the value of this parameter indicates a greater level of the network non-uniformity.

The indicated parameters make it possible to make a preliminary assessment of the GRN model topology. At a constant number of nodes, lower values of density and clustering of the network and a larger heterogeneity value testifies to the higher quality of network topology. A higher value of centralization coefficient indicates the degree of proximity of network topology to a star-shaped structure. Analysis of hereinbefore defined topological parameters allows us to define the steps of the network topology formation. On the one hand, the network should contain the maximum number of examined genes. On the other hand, network density and clustering coefficient should be minimal, and the coefficients of heterogeneity and

centralization should have maximal values. To make a final decision concerning network topology, it is necessary to calculate the complex topology index which should include the private topological parameters as the components. To solve this task, we have proposed a Harrington's desirability function [66], plot of which is shown in Figure 2.7. Implementation of this technique assumes transforming topological parameters scales into non-dimensional parameter Y, values of which are varied within the range from -2 to 5. Private values of desirabilities are calculated as follows:

$$dpr = exp(-exp(-Y)) \tag{5.8}$$

The algorithm to calculate the general topological index involves the following steps:

1. Transformation of topological parameters scale into non-dimensional parameter Y scale in accordance with following linear equations:

$$\begin{cases} Y_{DS} = a_{DS} - b_{DS} \cdot DS; \\ Y_{CL} = a_{CL} - b_{CL} \cdot CL; \\ Y_{Centr} = a_{Centr} + b_{Centr} \cdot Centr; \\ Y_{GT} = a_{GT} + b_{GT} \cdot GT. \end{cases} \tag{5.9}$$

   The a and b coefficients for appropriate topological parameter are determined empirically considering the appropriate topological parameter boundary values.

2. Calculation of the non-dimensional parameter Y for each of the topological parameters by the equations (5.9).

3. Determination of private desirabilities for the used topological parameters by the equation (5.8).

4. Calculation of the general topological index as geometric average of all private desirabilities:

$$GTP = \sqrt[n]{\prod_{i=1}^{n} dpr_i} \tag{5.10}$$

The maximal value of the index (5.10) indicates the optimal network topology in terms of the used topological parameters.

5.4 Reconstruction of GRN Based on Correlation Inference Algorithm

The process of GRN reconstruction based on correlation analysis assumes the calculation of correlation coefficients between all pairs of the gene expression profiles.

Since in the case of analysis of the matrix of gene expressions, the vectors of profiles are the sequences of rational numbers, it is reasonable to use the Pearson method for calculating a correlation between respective profiles:

$$r(X_a, X_b) = \frac{\sum_{i=1}^{m}(x_{ai} - \overline{x}_a)(x_{bi} - \overline{x}_b)}{\sqrt{\sum_{i=1}^{m}(x_{ai} - \overline{x}_a)^2} \cdot \sqrt{\sum_{i=1}^{m}(x_{bi} - \overline{x}_b)^2}} \tag{5.11}$$

where X_a and X_b are the vectors of the investigated gene expressions profiles; m is the number of attributes in the respective vectors; $\overline{x}_a$ and $\overline{x}_b$ represent the average values of the profiles X_a and X_b respectively. The pair correlation coefficient in the case of its significance represents the strength of the relationship between the corresponding nodes of the network. When using a full matrix of pair correlation coefficients, a gene network is fully connected, since there is connection between all the nodes of a given network. The weight of the arc is equal to the correlation coefficient between a pair of gene expression profiles whose relations are assessed. Network topology in this case is determined by the value of thresholding coefficient τ which defines the thresholding value of the existence of a relationship between a pair of genes in the network. A weight factor of the arc which connects the corresponding genes is defined as follows:

$$w(X_a, X_b) = \begin{cases} 0, \ if \ r(X_a, X_b) < \tau; \\ r(X_a, X_b), \ if \ r(X_a, X_b) \geq \tau. \end{cases} \tag{5.12}$$

Simulation of the process of a gene regulatorz network reconstruction based on gene expression profiles was performed based on Cytoscape software [127] using the gene expression profiles of data *moe430a* from the database ArrayExpress [32]. The data were obtained using DNA-microchip experiments and they contained information concerning the genes expressions of mesenchymal cells of two types: nerve crest and mesoderm. The matrix of initial data consisted of 147 of rows or genes and 20 of columns or the examined samples. A block chart of the algorithm to implements the process of GRN reconstruction based on the correlation inference algorithm is shown in Figure 5.4. Implementation of this process assumes the following steps:

1. Formation of the input data as a matrix where rows are the genes which represent nodes of a gene network, and columns are the investigated samples.

2. Setup the range and the step the thresholding coefficient change. Initializing the initial value of the threshold coefficient $\tau = \tau_{min}$.

3. Reconstruction of gene regulatory network within the range of the thresholding parameter change from minimal to maximal value. Calculation of the topology parameters at each step of this procedure implementation.

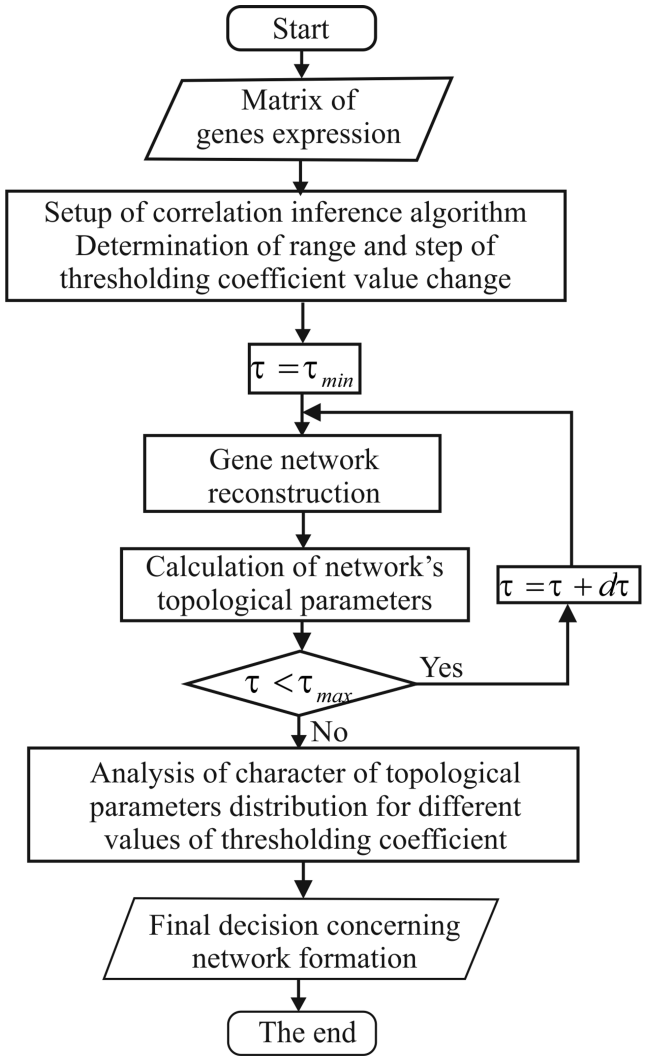

Figure 5.4: Block diagram of algorithm to determine the optimal value of threshold coefficient when using a correlation inference algorithm

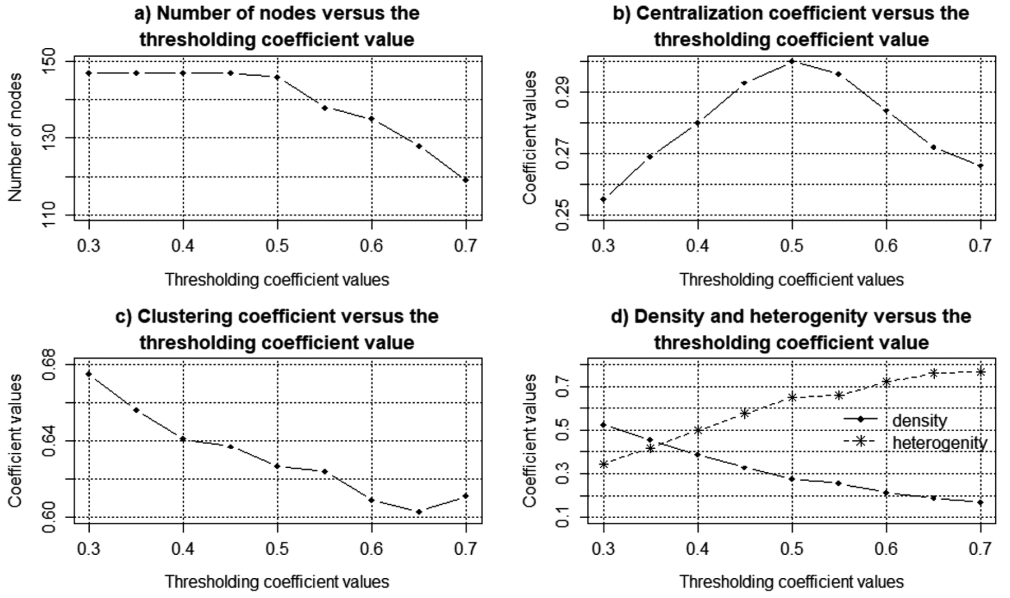

Figure 5.5: Charts of the simple parameters versus the thresholding coefficient: a) number of genes; b) centralization coefficient; c) clustering coefficient; d) density and heterogeneity of the network

4. Calculation of general topology index by formulas (5.8)-(5.10) at each step of this procedure implementation.

5. Analysis of the obtained results, determining the value of threshold coefficient that matches the optimal GRN topology (maximum value of the general topology index).

Figure 5.5 shows the results of the algorithm operation in the form of charts of the simple topological parameters versus the thresholding coefficient. The value of the threshold coefficient was changed within the range from 0.3 to 0.7 step 0.05. An analysis of the charts allows concluding that within the range of the thresholding coefficient from 0.3 to 0.45 the number of genes in the network does not change. In this case, the values of both the centralization and heterogeneity coefficients are increased while the values of the of clustering and density coefficients are decreased. This fact indicates the improvement in the network topology by reducing the number of connections between appropriate nodes at a constant number of genes. When the value of threshold coefficient is achieved 0.5, the number of genes is reduced from 147 to 146 while the centralization coefficient achieved its maximum. Upon

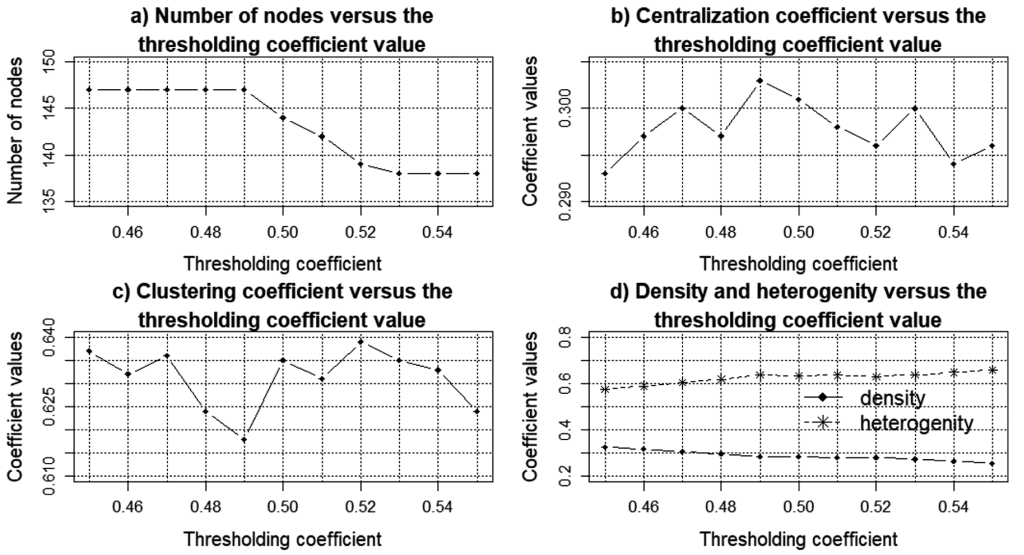

Figure 5.6: Charts of the simple parameters versus the thresholding coefficient, while the value of the thresholding coefficient is changed within the range from 0.45 to 0.55 with a step 0.01: a) number of genes; b) centralization coefficient; c) clustering coefficient; d) density and heterogeneity of the network

further increase of the threshold coefficient value, the number of genes and the value of the centralization coefficient are decreased sharply. This fact indicates the deterioration of the network topology structure. The obtained results allows us to define a narrower range of the threshold coefficient value variation for purpose of determining the optimal topology of a gene network.

Figure 5.6 shows the charts of the simple topological parameters versus the value of thresholding coefficient while it was changed within the range from 0.45 to 0.55 with a step 0.01. In this case, the several genes linked between each other but separated from the main network were removed from the data. Figure 5.6 shows that the values of thresholding coefficient which form the structure of a gene network is determined by four topological parameters: clustering coefficients, centralization, the network homogeneity, and nodes density. It should be noted that the optimal network structure corresponds to the minimal values of the density and clustering coefficients, and to the maximal values of the centralization and heterogeneity coefficients. Figure 5.7 shows the chart of general topological index versus the thresholding coefficient whose value was changed within the range from 0.45 to 0.55 with a step 0.01. Analysis of the Figure 5.7 has shown that the optimal value based on simple parameters for the estimation of the gene network topology is the

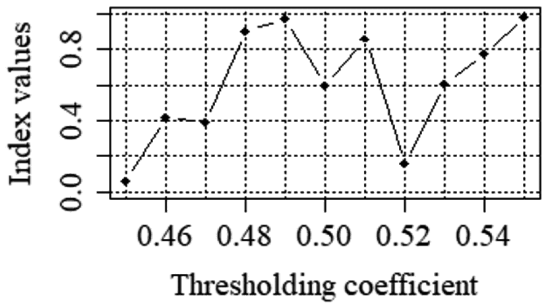

Figure 5.7: Charts of the general topological index versus the thresholding coefficient value while applying correlation inference algorithm

value of thresholding coefficient 0.49. In this case, the network contains 147 of genes, centralization coefficient reaches a maximum, while clustering coefficient reaches a minimum. The values of both the density and heterogeneity coefficients within the range of the thresholding coefficient change from 0.45 to 0.49 are changed monotonically towards to lower and larger values respectively. In the range from 0.49 to 0.51, the rate of these parameters change is zero. The value of the general topological index also reaches a maximum when a value of thresholding coefficient is 0.49. Figure 5.8 shows the result of the gene regulatory network reconstruction when applying the correlation inference algorithm with a thresholding coefficient value of 0.49.

The conducted research allows us to propose a technique of gene regulatory network reconstruction based on the correlation inference algorithm. Structural block chart of the algorithm of this technique implementation is presented in Figure 5.9. Practical implementation of a this algorithm assumes the following stages:

Stage 1. Problem statement. Data formation.

1. Formation of initial data as a matrix where rows are the genes and columns are the stadied samples.

Stage 2. Approximate estimation of both the range and step of the thresholding coefficient value change.

1. Setup the range and step of the thresholding coefficient value change. Initialization of the initial value of thresholding coefficient $\tau = \tau_{min}$.
2. Reconstruction of GRN whose topology matches the assigned value of thresholding coefficient.

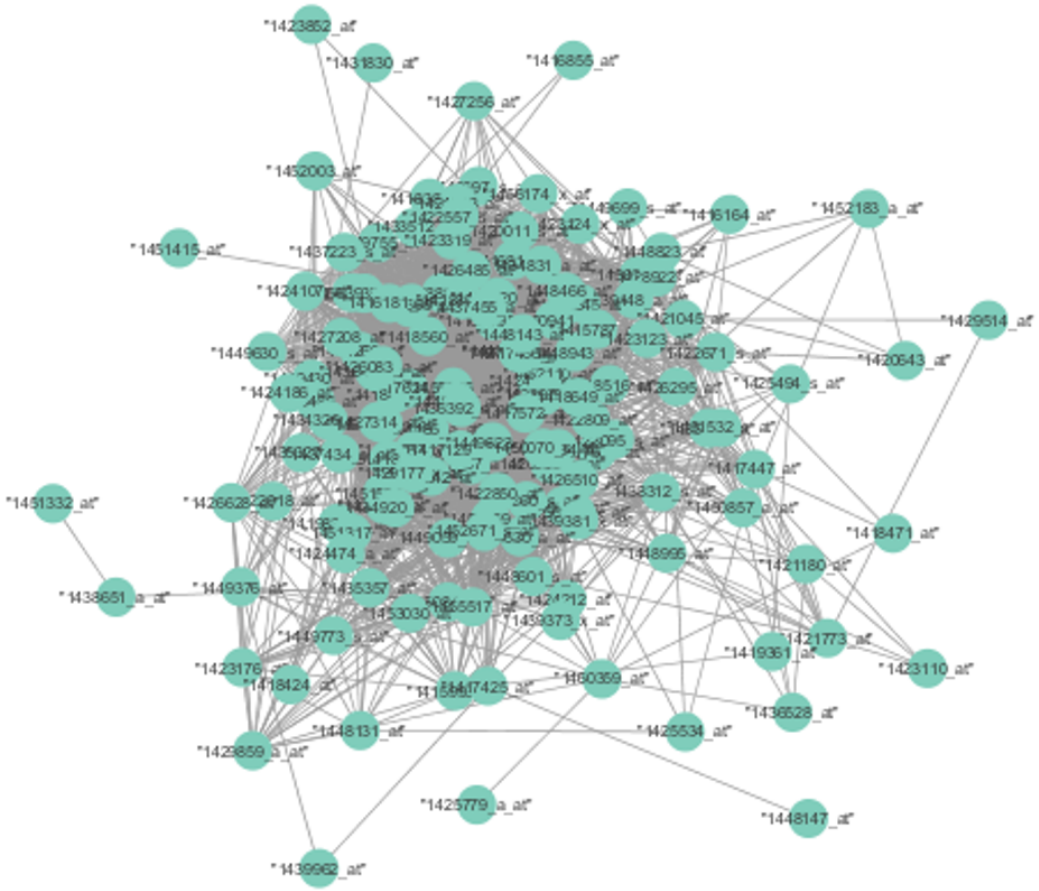

Figure 5.8: Result of gene network reconstruction while applying the correlation inference algorithm

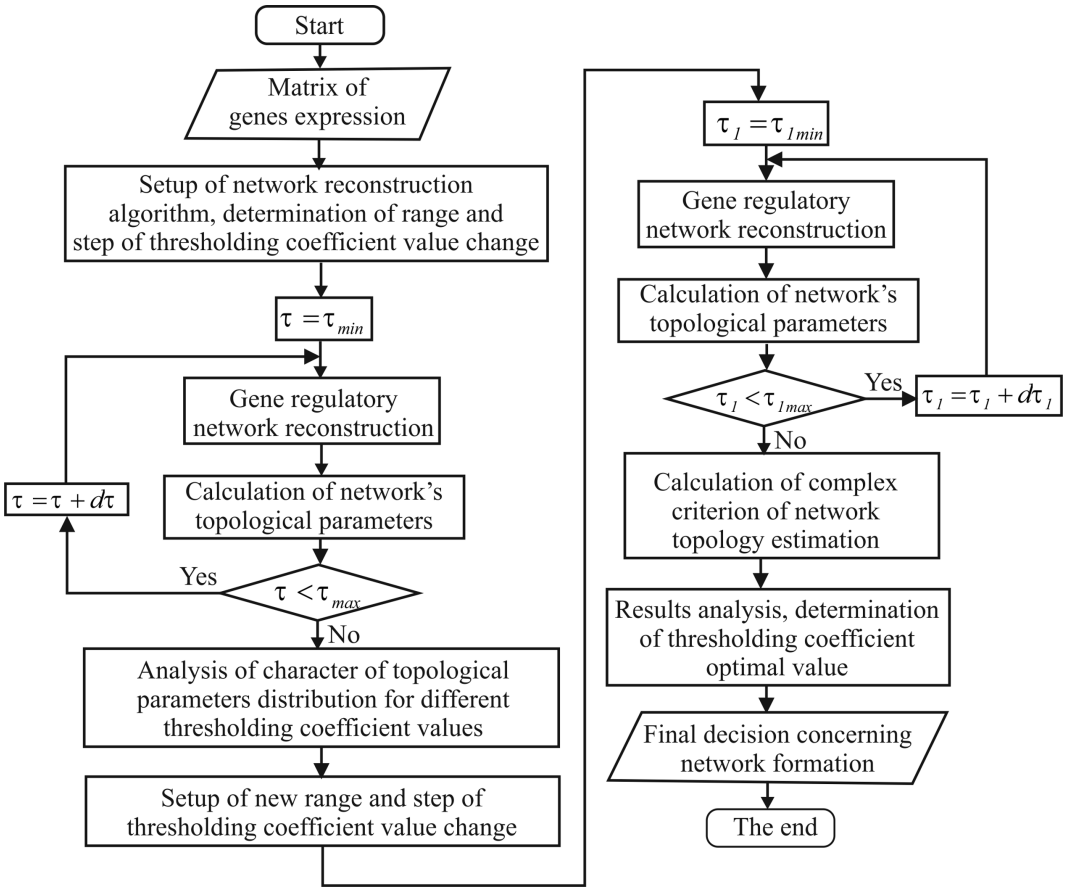

Figure 5.9: Technique of gene network reconstruction based on correlation inference algorithm

3. Calculation of topological parameters for the reconstructed gene regulatory network.
4. If a value of the thresholding coefficient is less than the maximal value, then increasing this value by $\mathrm{d}\tau$ (step for a change in threshold coefficient) and escape to the step 3 of this procedure. Otherwise, creating charts of the topological parameters within the range of the thresholding coefficient change.
5. Analysis of the obtained results. Determining the new, narrower, range and a smaller step of the thresholding coefficient variation.

Stage 3. Determining the optimal value of the thresholding coefficient.

1. Reconstruction of a gene regulatory network within the framework of new interval of the threshold coefficient change. Calculation of the topological parameters at each step of this procedure implementation. Calculation of the general topological index.
2. Create the charts of both the simple topological parameters and the general topological index versus the thresholding coefficient.
3. Analysis of the obtained results. Determination of the thresholding coefficient optimal value.

Stage 4. Reconstruction of gene regulatory network.

1. Reconstruction of GRN applying the optimal value of the thresholding coefficient.

5.5 Reconstruction of GRN Based on ARACNE Inference Algorithm

ARACNE inference algorithm (Algorithm for the Reconstruction of Accurate Cellular Networks) of gene regulatory networks reconstruction forms the links between genes (arcs) based on the analysis of statistical hypotheses concerning presence or absence of appropriate link [104]. As a result, the vector of probabilities is formed for each gene at the stage of gene network reconstruction. Each of the elements of this vector determines both the existence and the force of appropriate link. The estimation of the degree of interconnection between a pair of genes in the case of N connection variants is performed using the Gaussian kernel estimation based on Shannon entropy:

$$I(g_i, g_j) = \frac{1}{N} \sum_{k=1}^{N} log \frac{H_k(g_i, g_j)}{H_k(g_i) H_k(g_j)} \tag{5.13}$$

where $H(g)$ is a Shannon entropy which is calculated for profile of g gene. The principal idea of ARACNE algorithm is the following: if there are different paths of appropriate genes linking in the network, each of them is characterized by degree of the appropriated connection $I(g_i, g_j)$, then, it is selected the connection which satisfied the following condition:

$$I(g_i, g_j) = min[I(g_i, g_s), I(g_s, g_p), ..., I(g_h, g_j)] \tag{5.14}$$

where $g_s, g_p, ..., g_h$ are intermediate genes through which are linked the genes g_i and g_j. As a result, we obtain a network of interacting genes, the weight of the linkage between the corresponding genes is determined by degree of linkage between the genes. The number of relationships in network is also limited by the appropriate thresholding coefficient. It is assumed that if the weight of the corresponding link is less than the value of the thresholding coefficient, the relationship between these genes is broken. In this way a topology of gene network is formed.

Figure 5.10 presents the structure block-chart of the algorithm to estimate the thresholding coefficient optimal value which corresponds to optimal gene network topology in terms of used topological parameters [11]. Implementation of the algorithm assumes the next steps:

1. Problem statement and the experimental dataset formation. The data is formed as a matrix, where rows and columns are the investigated genes and samples respectively.

2. Data preprocessing. Implementation of this step involves data normalizing and non-informative genes reducing. The number if the investigated genes significantly decreases at this step.

3. Setup the range and the step of the thresholding parameter value variation.

4. Reconstruction of gene regulatory network within the range of the thresholding coefficient change from minimum to maximum value. Calculation of the topological parameters for each thresholding coefficient value.

5. Analysis of the obtained results. Setup a new significantly less range and step of the thresholding parameter value variation.

6. Reconstruction of gene regulatory network within the new range of the thresholding parameter variation. Calculation of the general topological index at each step of this procedure implementation.

7. Result analysis. Determination of the optimal value of the thresholding parameter. This value corresponds to the maximum of the general topological index in the case of maximal quantity of genes in network.

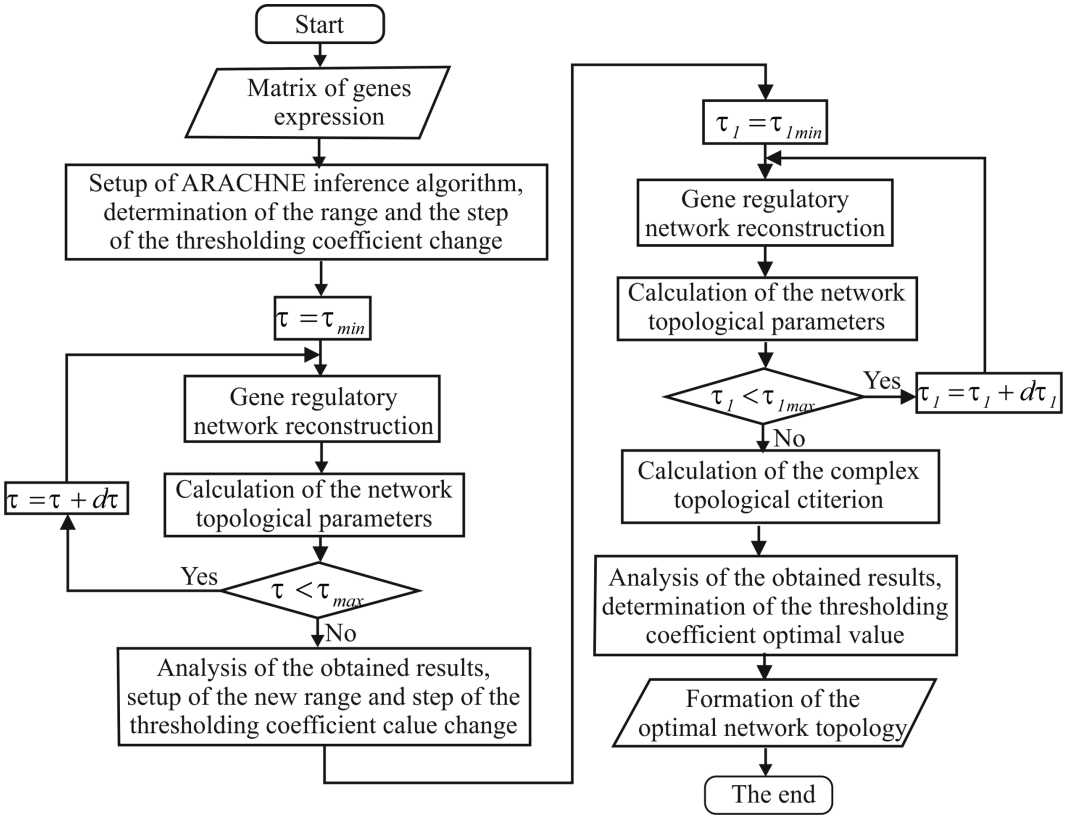

Figure 5.10: A structure block-chart of algorithm to form the gene regulatory network optimal topology while applying ARACNE inference algorithm

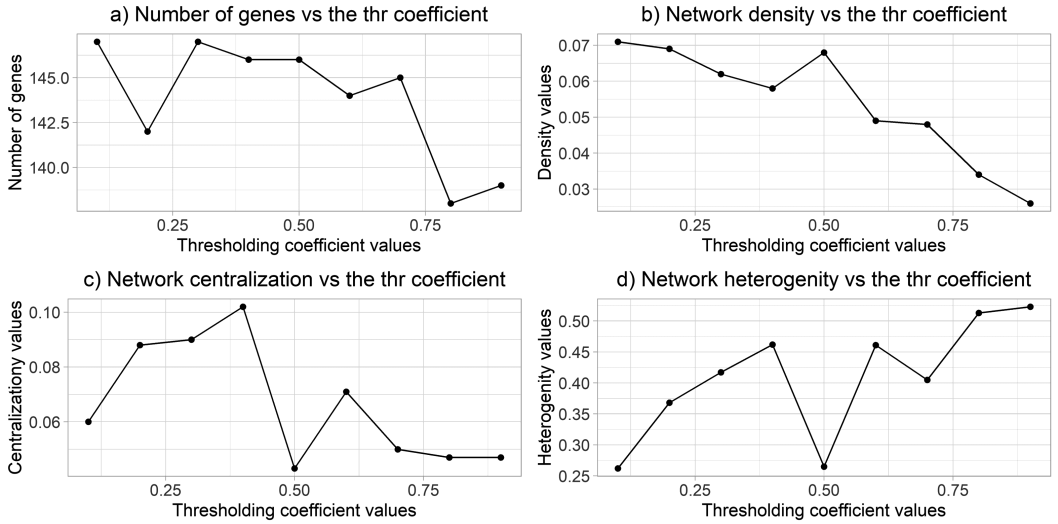

Figure 5.11: Charts of network topological parameters versus the thresholding coefficient values in the case of range of the thresholding coefficient variation from 0.1 to 0.9

8. Gene regulatory network reconstruction using optimal thresholding coefficient value.

5.5.1 Implementation of the Technique of GRN Reconstruction Based on ARACNE Inference Algorithm

Similarly to technique of gene regulatory network reconstruction based on correlation inference algorithm, simulation of the gene network reconstruction based on the ARACNE inference algorithm was also performed based on the Cytoscape software using gene expression profiles of *moe430a* dataset from ArrayExpress database. Initially, the value of the thresholding coefficient was changed within the range from 0.1 to 0.9 with a step 0.1. The charts of the topological parameters versus the thresholding coefficient value are presented in Figure 5.11. The network clustering coefficient was equal zero, that indicates an absence of the relationships between neighbours of genes in network. The analysis of the obtained charts allows us to conclude that the optimal value of the thresholding coefficient is allocated within the range from 0.3 to 0.5, since the values of the centralization and heterogeneity coefficients in this range achieve the local maxima and the value of the density of the network nodes achieves the local minimum. The number of genes in this case varies from 147 to 146, what is quite acceptably. Figure 5.12 shows the same charts for the case of range of the thresholding coefficient from 0.3 to 0.5. Step of this coefficient

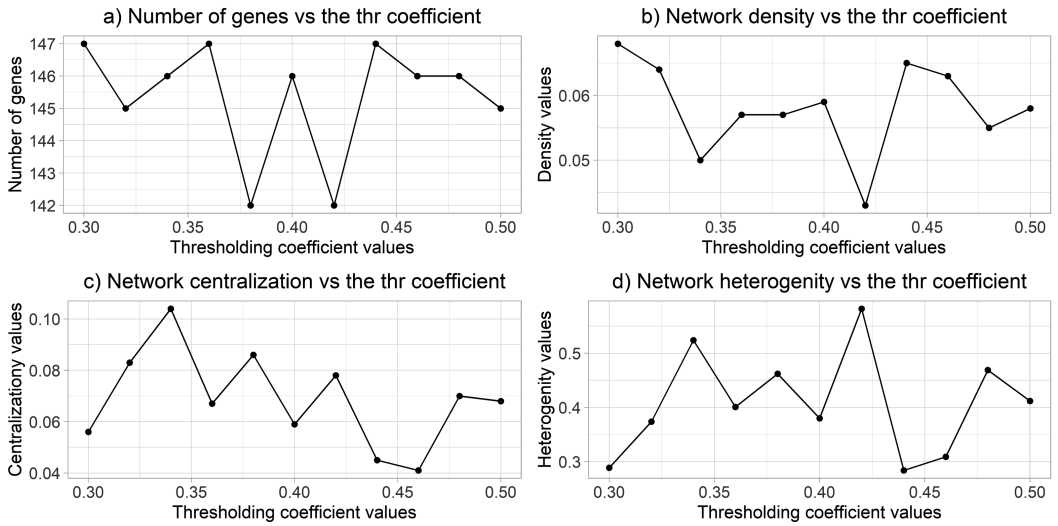

Figure 5.12: Charts of network topological parameters versus the thresholding coefficient values in the case of range of the thresholding coefficient variation from 0.3 to 0.5

change was 0.02. As it can be seen from Figure 5.12, the analysis of the obtained charts does not allow us to determine the optimal value of the thresholding coefficient, since the coefficients of centralization, heterogeneity and density have three local extrema, which disagree between each other. In accordance with hereinbefore presented algorithm it is necessary to calculate the general topology index for each combination of the thresholding parameters. The plot of the general topology index versus the thresholding coefficient is showed in Figure 5.13.

The analysis of the obtained results allows concluding that the maximal value of the general topological index is achieved in the cases of 0.33 and 0.42 values of the thresholding coefficients. However, it should be noted that in the second case, the gene network contains five genes less. Thus, the value of the thresholding coefficient of 0.34 is more acceptable in terms of the number of genes in the network. The result of gene regulatory network reconstruction while applying ARACNE inference algorithm is shown in Figure 5.14. Figure 5.15 shows the charts of the topological parameters distribution for the reconstructed gene regulatory network.

The analysis of the charts indicates the higher level of structuredness of gene networks, reconstructed based on ARACNE inference algorithm, since the number of genes with high degree is not so much, the values of the topological coefficient, number of shared neighbours and average of neighbours connectivity are decreased monotonically during the number of neighbours increase. Moreover, the comparison

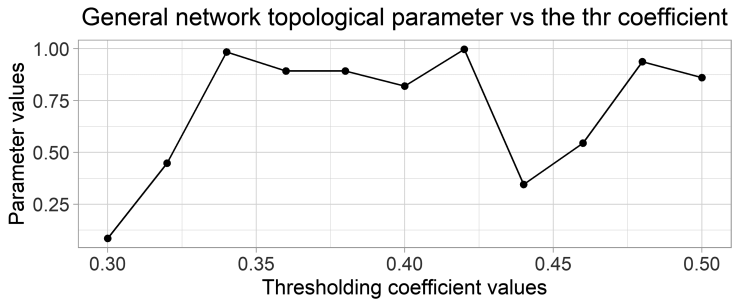

Figure 5.13: Charts of the general topological index versus the thresholding coefficient value while applying ARACNE inference algorithm

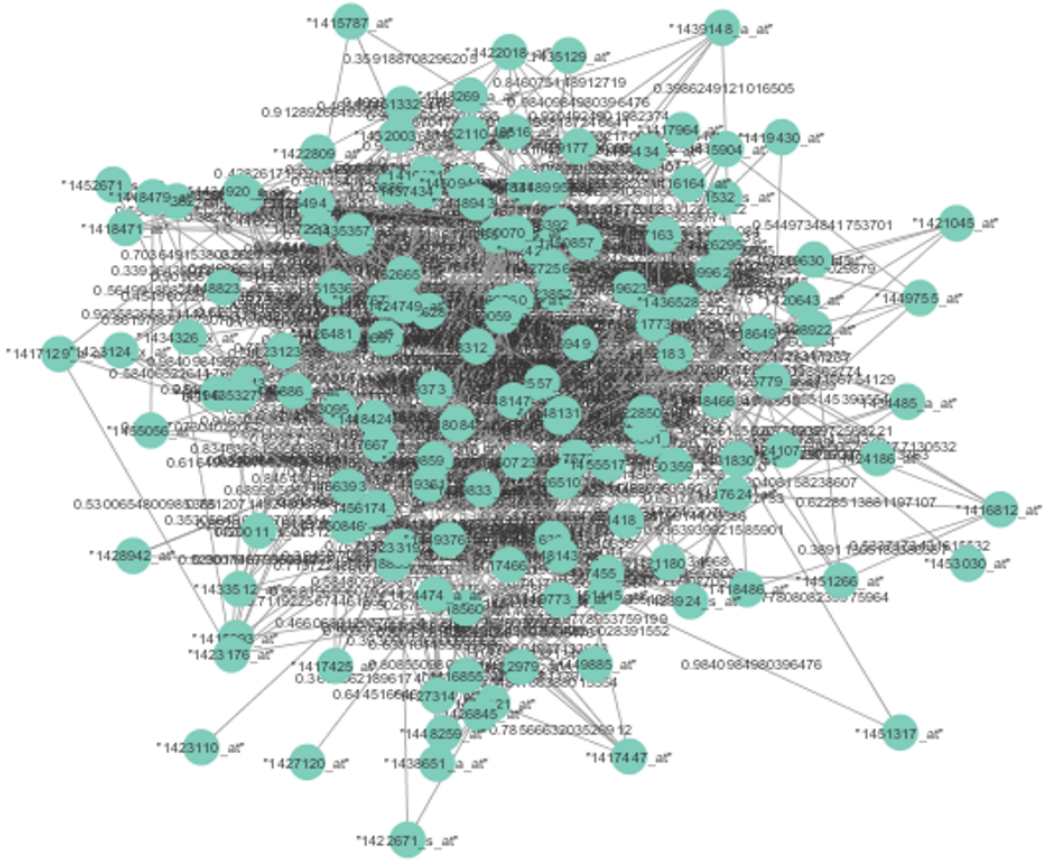

Figure 5.14: Result of gene network reconstruction based on ARACNE inference algorithm

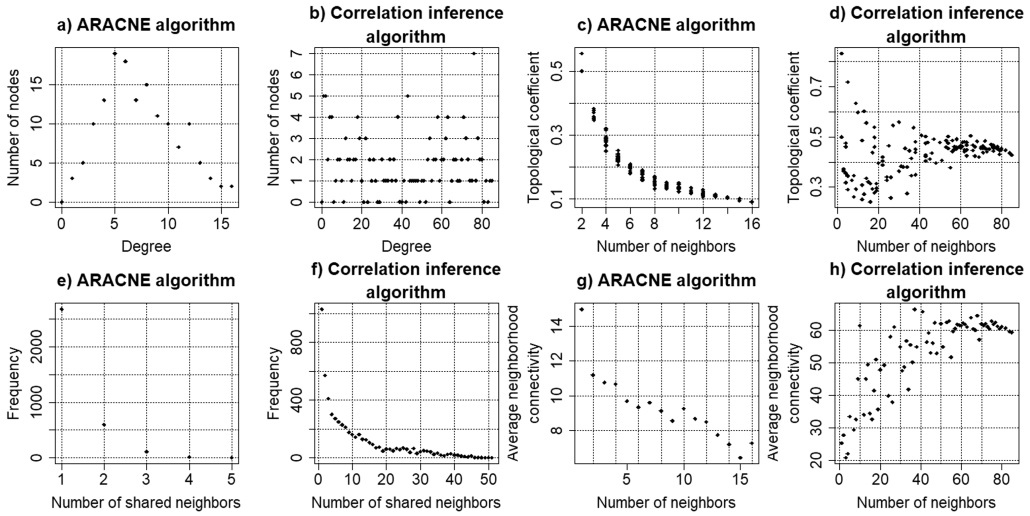

Figure 5.15: Charts of topological parameters distribution for GRN reconstructed using ARACNE and correlation inference algorithms: a,b) distribution of degree of the network nodes; c,d) distribution of the topological coefficient; e,f) distribution of number of shared neighbours; g,h) distribution of average of the neighbours connectivity

of the obtained charts with appropriate charts for other reconstructed valid gene networks [127] shows high level of their similarity. This fact also allows concluding about the correctness of the proposed technique. Of course, the effectiveness of the reconstructed gene regulatory network can be estimated by the further simulation process implementation.

5.6 Technique of Validation of the Reconstructed GRN

Block chart of the process of gene regulatory network reconstruction and validation of the reconstructed models (Figure 5.1) involves a comparative analysis of the selected biological gene network and the reconstructed network based on gene expression profiles. This approach assumes the development of a technique of quality evaluation of the gene network reconstruction procedure by comparison of the topologies of both the basic network and networks reconstructed based on gene expression profiles. Within the framework of our research, gene networks are considered to be completely adequate if the character of relations between relevant genes in the basic and reconstructed networks fully coincides. In this case, we estimate the presence of a relationship between appropriate genes in different networks. If there is a connection, it is considered to be equal 1. In the case of absence of such a connection, this value is 0. The ROC analysis (Receiver Operator Characteristic), which is applied to visualize the results of binary classification using errors of the first and second types [52], was used as the basic technique for comparative analysis of the investigated gene regulatory networks. According to ROC analysis theory, at the first stage we calculate the quality parameters for the classification of relationships between genes in relevant networks. Such parameters are:

- *TP* (True Positives) is the number of relationships between pairs of matching genes that coincide in the two networks (true positive cases).
- *TN* (True Negatives) is the number of matching negative relationships between pairs of corresponding genes in different networks (true negative cases).
- *FN* (False Negatives) is the number of relationships between pairs of genes, reconstructed based on full data which do not identified in the network, reconstructed on the basis of a limited number of genes and conditions (error of the first kind). In this case, the relationship that exists between a pair of genes in the full network is missing between the pair of genes in the examined network.
- *FP* (False Positives) is the number of missing links between the relevant genes in the network, reconstructed based on full data that are identified as existing in the network, reconstructed on the basis of a limited number of genes and conditions (error of the second kind). In this case, a connection between a pair

of genes that is missing in the full network is identified as existing between a given pair of genes in the examined network.

Based on these parameters, we calculate relative parameters for the reconstructed network quality assessment:

- the percentage of true positive cases or the *sensitivity* of the model is the ratio of the number of true positive connections to the full number of connections between the examined genes which is estimated based on the results of analysis of the gene network reconstructed based on complete data:

$$Sc = TPR = \frac{TP}{TP + FN} \cdot 100\% \tag{5.15}$$

- percentage of false positive cases:

$$FPR = \frac{FP}{TN + FP} \cdot 100\% \tag{5.16}$$

- specificity is the percentage of missing links that were correctly identified in the network reconstructed based on a limited number of genes and samples:

$$Sp = \frac{TN}{TN + FP} \cdot 100\% \tag{5.17}$$

It should be noted that the percentage of false positive cases and the specificity are related via ratio: $FPR = 100 - Sp$. A larger specificity value corresponds to a smaller percentage of incorrectly identified cases of the presence of links in the complete network. ROC-curve is presented as a chart of sensitivity Sc against the percentage of incorrect positive cases $FPR = 100 - Sp$. A larger value of sensitivity and a lower FPR value corresponds to a higher degree of the adequacy of the model. In this case, the area under a ROC-curve (AUC) reaches the highest value. Another criterion that determines the adequacy of a model is calculated as the ratio of sensitivity to the percentage of false positive cases:

$$RC = \frac{Sc}{FPR} \tag{5.18}$$

A higher value of this criterion corresponds to a larger level of adequacy of gene network reconstructed on the basis of data in biclusters to the basic gene network.

Figure 5.16 shows a stricture block chart of the gene regulatory networks validation technique. In this case, we used as the basic the gene regulatory network reconstructed based on full set of gene expression profiles. Practical implementation of this technique assumes the following steps:

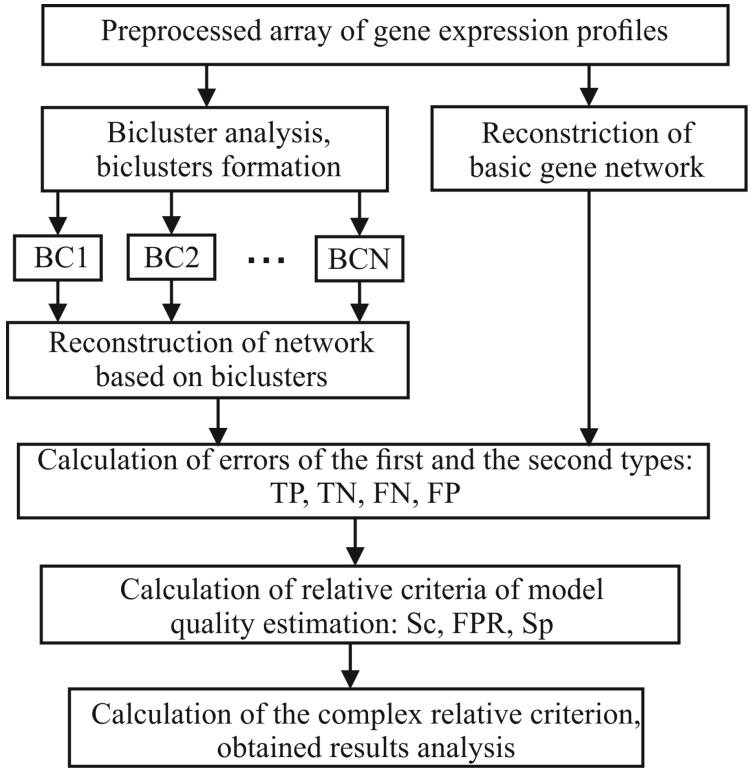

Figure 5.16: Structure block chart of gene regulatory networks validation technique

1. Problem statement. Forming an array of gene expression profiles. Data preprocessing: filtering, reducing and clustering of gene expression profiles.

2. Reconstruction of basic gene networks based on full set of the obtained gene expression profiles.

3. Bicluster analysis of the gene expressions data. Allocation of the biclusters.

4. Reconstruction of gene regulatory networks based on gene expressions data of the relevant biclusters.

5. Determination of the absolute quality parameters for the classification of relationships between appropriate genes in the reconstructed networks.

6. Calculation of relative quality parameters Sc, Sp, FPR in accordance with the formulas (5.15)–(5.17).

7. Creation of ROC curve and evaluation of area under it. Calculation relative quality criterion of gene regulatory networks reconstruction by formula (5.18).

8. Selection of the model whose area under the ROC curve is the largest or which corresponds to the higher value of the relative quality criterion (5.18).

5.6.1 Validation of the GRN Reconstructed Using Correlation Inference Algorithm

Figure 5.17 shows the results of bicluster analysis of gene expression profiles of data *moe*430*a* [32], performed using the biclustering algorithm *ensemble* [81] which is implemented within the framework of the technique presented in the section 3.3.3. At the first stage, the value of a parameter that determines the ratio of the number of rows and columns in biclusters was fixed at the level of 0.15, the value of thresholding coefficient was changed within the range from 0.01 to 0.5 with a step 0.01. The 0.12 value of the thresholding coefficient was fixed as the result of Figure 5.17b analysis (First global minimum). Increasing the value of this criterion is not reasonable, since it leads to sharp increasing the number of small biclusters (Figure 5.17a).

At the second step, the parameter which determines the ratio of rows and columns in biclusters was changed within the range from 0.05 to 0.12 with a step 0.01. As in can be seen from Figure 5.17, the minimal value of the internal biclustering quality criterion corresponds to the 0.1 value of ratio of rows and columns in biclusters. The results of gene expression profiles biclustering while the *ensemble* biclustering algorithm applying with thresholding coefficient 0.12 and ratio of rows and columns in biclusters 0.1 are presented in Table 5.1. Reconstruction of gene regulatory networks was carried out using biclusters contained more than ten of

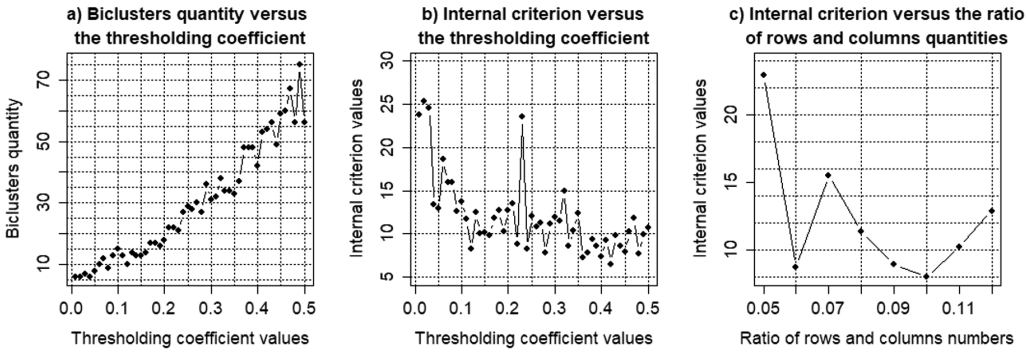

Figure 5.17: Results of the bicluster analysis of gene expression profiles of data *moe430a*: a) chart of the number of biclusters vs the thresholding coefficient; b) chart of the biclustering quality criterion vs the thresholding coefficient; c) chart of the biclustering quality criterion vs the ratio of rows and columns in biclusters

Table 5.1: Distribution of rows and columns in the biclusters obtained as a result of the gene expression profiles of *moe*430*a* data biclustering

BC	1	2	3	4	5	6	7	8	9	10	11	12	13
Genes	23	16	9	13	5	39	11	44	32	28	24	24	6
Samples	8	8	4	9	8	8	7	12	12	6	11	9	6

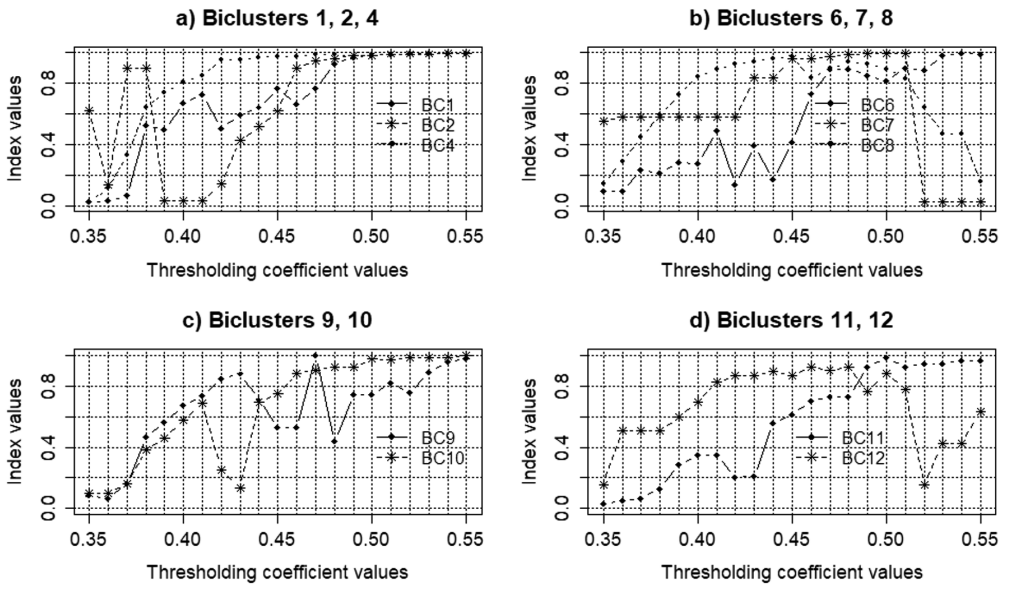

Figure 5.18: Charts of the general topological index versus the value of the thresholding coefficient for gene regulatory networks reconstructed using the correlation inference algorithm based on data from biclusters: a) 1, 2, 4; b) 6, 7, 8; c) 9, 10; d) 11, 12

genes (small biclusters were not dealt with). Thus, 10 biclusters were chosen in this case: $BC1$, $BC2$, $BC4$, $BC6 - BC12$. Figure 5.18 shows the charts of the general topological index versus the thresholding coefficient for the reconstructed gene networks based on the obtained biclusters. The value of the thresholding coefficient was changed within the range from 0.35 to 0.55 with a step 0.01. This range was determined empirically during the simulation process. The values of the thresholding coefficient within this range corresponded to the complete number of genes in the reconstructed networks. The following values of the thresholding coefficients were determined as the result of Figure 5.18 analysis: $BC1$, $BC2$ and $BC7$ - 0.51; $BC4$ and $BC9$ - 0.47; $BC6$ - 0.53; $BC8$ - 0.45; $BC10$ and $BC11$ - 0.5; $BC12$ - 0.48.

Table 5.2 presents the relative criteria values for reconstructed GRN which were calculated by formulas (5.15) - (5.18). The chart of relative quality criterion calculated for GRN reconstructed based on the data in allocated biclusters is presented in Figure 5.19. Horizontal line in Figure 5.19 is drawn at the level of the average value of the relative validation criterion for the reconstructed gene networks. An analysis of the obtained results allows concluding that the value for the specificity

Table 5.2: Relative criteria for the reconstructed GRN based on correlation inference algorithm

BC	1	2	4	6	7
Sp	99.57	99.92	99.95	98.29	99.77
Se	51.47	89.66	90.24	57.42	45.45
RC	120.71	1156.41	1746.99	33.66	195.43
BC	8	9	10	11	12
Sp	97.08	99.35	98.78	98.80	98.9
Se	61.76	85.34	67.36	54.86	52.51
RC	21.15	132.09	55.46	45.65	47.81

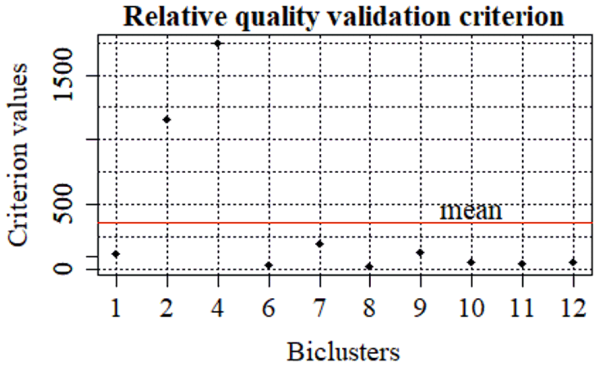

Figure 5.19: Chart of the relative quality validation criterion for GRN reconstructed using correlation inference algorithm

parameter is varied within the range from 97 to 100 percentages, indicating a low percentage of incorrectly identified positive cases. The value of sensitivity in these cases is varied from 45.5% for the gene network based on data of the seventh bicluster to 90.2% for the network, reconstructed on the basis of data from the fourth bicluster. The minimal value of the relative criterion corresponds to the gene network reconstructed based on the data of eighth bicluster and it equals 21. In this case, the maximal value of this criterion is equal 1746 for the GRN reconstructed based on the data of fourth bicluster. The weighted average of relative validation criterion for all reconstructed models is equal 355.5.

The obtained results indicate a high level of adequacy of the proposed technique for the reconstruction of gene regulatory networks, since the values of relative validation criteria for all reconstructed gene networks are substantially larger than unity. The number of incorrectly identified positive cases belongs to the interval from 0 to 3 percentages, and sensitivity values are less than 50% (45.5) only for the network reconstructed based on the seventh bicluster. For the gene networks reconstructed based on other biclusters, the value of this parameter is greater than 50%, and for the fourth bicluster it reaches 90.2%.

5.6.2 Validation of the GRN Reconstructed based on ARACNE Inference Algorithm

Figure 5.20 shows charts of the general topological index versus the thresholding coefficient for the models of gene regulatory networks reconstructed applying ARACNE inference algorithm using the data in obtained biclusters. Thresholding coefficient value in this case was changed within the range from 0.03 to 0.2 with a step 0.01. This range was also determined empirically. The value of the thresholding coefficient in this range corresponded to the complete number of genes in the reconstructed networks. The following values of the thresholding coefficient were determined as a result of the obtained charts analysis: $BC1$, $BC4$, $BC9$ and $BC12$ - 0.13; $BC2$ - 0.06; $BC6$ - 0.19; $BC7$ - 0.07; $BC8$ - 0.17; $BC10$ - 0.14; $BC11$ - 0.09. Table 5.3 and Figure 5.21 present the relative criteria values for GRN reconstructed using ARACNE inference algorithm and chart of the relative quality criterion calculated for the reconstructed GRN respectively.

An analysis of the obtained results allows us to conclude that the level of adequacy of the models of gene regulatory networks reconstructed using ARACNE inference algorithm is significantly less in comparison with the gene networks, reconstructed using correlation inference algorithm. So, in the case of applying ARACNE inference algorithm, sensitivity and specificity values are varied within the ranges from 34.3% to 72.5%, and from 95.1% to 99.2% respectively. The weighted average of the relative validation criterion for the models of gene networks reconstructed based on the ARACNE algorithm is equal to 25.24, that significantly less than the

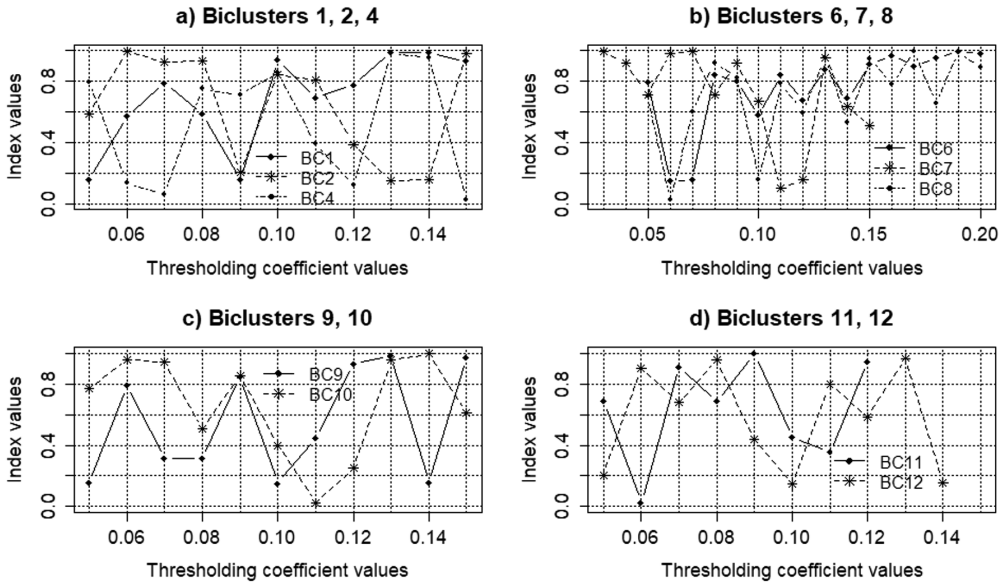

Figure 5.20: Charts of the general topological index versus the value of the thresholding coefficient for gene regulatory networks reconstructed using ARACNE inference algorithm based on data from biclusters: a) 1, 2, 4; b) 6, 7, 8; c) 9, 10; d) 11, 12

Table 5.3: Relative criteria for GRN reconstructed based on ARACNE inference algorithm

BC	1	2	4	6	7
Sp	97.58	95.72	98.86	98.02	96.82
Se	44.55	38.45	72.5	53.00	34.29
RC	18.39	8.98	63.54	26.81	10.78
BC	8	9	10	11	12
Sp	95.11	98.37	99.19	96.59	97.65
Se	48.32	60.00	39.09	46.82	50.26
RC	9.87	36.71	48.54	13.72	21.36

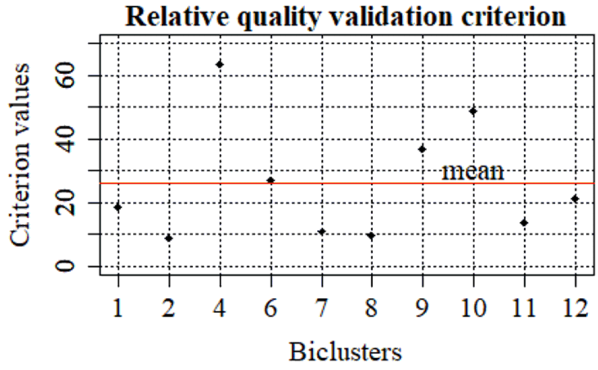

Figure 5.21: Chart of the relative quality validation criterion for GRN reconstructed using ARACNE inference algorithm

corresponding value when applying the correlation inference algorithm. This fact indicates a higher effectiveness of the correlation inference algorithm in comparison with ARACNE algorithm in terms of the used quantitative quality criteria.

5.7 Conclusions

In this chapter, we have presented a techniques of both gene regulatory network reconstruction and validation of the reconstructed models. The technique of GRN reconstruction is presented as a structural block chart of stepwise procedure of determination of the used inference algorithm optimal parameters considering the network topological parameters. The optimal topolology of the network in this case corresponded the maximum value of the general topological index which contained the partial network topological parameters as the components. As the simulation results, we have obtained the charts of both the simple topological parameters and the general topological index versus the value of a thresholding coefficient, which determines the network topology. The simulation process was carried out using both correlation and ARACNE inference algorithms based on Cytoscape software. The values of the thresholding coefficient were 0.49 and 0.33 in the cases of the use of correlation and ARACNE inference algorithms respectively. In these cases the value of the general topological index achieves maximal ones and networks contained 147 of genes from 147.

A technique of the reconstructed gene regulatory network validation is based on ROC analysis, implementation of which assumes a comparative analysis of the character of relations between relevant genes in the basic network and in the networks

reconstructed based on data in the obtained biclusters with calculation of errors of both the first and second types. The relative validation quality criterion has been proposed as the main criterion to evaluate the adequacy of the reconstructed network. This criterion was calculated as the ratio of sensitivity of the model in percentage to the percentage of false positive cases. A larger value of this criterion corresponds to a higher level of adequacy of the networks reconstructed based on biclusters to basic network. It has been shown that when applying the correlation inference algorithm, the value of specificity parameter is varied within the range from 97 to 100 percentages, indicating a low percentage of incorrectly identified positive cases. The value of sensitivity in this case was varied from 45.5% for the gene network based on data from the seventh bicluster, to 90.2% for the network, reconstructed on the basis of data from the fourth bicluster. Minimal value of a relative validation criterion corresponded to the gene network based on the eighth bicluster and this value was equal 21. In this case, the maximal value of this criterion of 1746 matched the fourth bicluster. The weighted average of relative validation criterion for the reconstructed models was equal 355.5. When employing ARACNE inference algorithm, the sensitivity was varied within the range from 34.3% to 72.5%, and the specificity was varied from 95.1% to 99.2%. The weighted average of relative validation criteria for the models of gene networks reconstructed based on the ARACNE inference algorithm was equal 25.24, what is significantly less in comparison with the use of correlation inference algorithm.

Thus, it can be concluded, that in the terms of existence or absence of the links between relevant genes in the basic network and networks reconstructed based on the biclusters, the gene regulatory network reconstructed using correlation inference algorithm has higher level of adequacy in comparison with gene networks reconstructed using ARACNE inference algorithm. However, the comparison analysis of the charts of distributed network topological parameters have shown the higher level of structuredness of gene networks, reconstructed based on ARACNE inference algorithm in comparison with the use on correlation inference algorithm, since the number of genes with high degree is not so much, the values of the topological coefficient, number of shared neighbours and average of neighbours connectivity are decreased monotonically during the number of neighbours increase. Moreover, the comparison of the obtained charts with appropriate charts for other reconstructed valid gene networks shows high level of their similarity. This fact also allows concluding about the corectnesss of the proposed technique. The obtained results allow us to conclude too that the final decision concerning level of the gene regulatory network adequacy can be done for the further simulation process.

Bibliography

[1] Ncbi homepage, https://www.ncbi.nlm.nih.gov/geo/query

[2] Affymetrix: Statistical algorithms description document. affymetrix, inc., santa clara, ca (2002), http://tools.thermofisher.com/content/sfs/brochures

[3] Alekseenko, V., Sharko, A., Sharko, A., Stepanchikov, D., Yurenin, K.: Identification by the method of structural features of deformation mechanisms at bending. Technical Diagnostics and Non-Destructive Testing **1**, 32–39 (2019). https://doi.org/10.15407/tdnk2019.01.04

[4] Alkallas, R., Fish, L., Goodarzi, H., Najafabadi, H.: Inference of rna decay rate from transcriptional profiling highlights the regulatory programs of alzheimer's disease. Nature Communications **8**(1), 909 (2017). https://doi.org/10.1038/s41467-017-00867-z

[5] Assenov, Y., Ramírez, F., Schelhorn, S.E., Lengauer, T., Albrecht, M.: Computing topological parameters of biological networks. Bioinformatics **24**(2), 282–284 (2007). https://doi.org/10.1093/bioinformatics/btm554

[6] Astrand, M.: Contrast normalization of oligonucleotide arrays. Journal of Computational Biology **10**(1), 95–102 (2003). https://doi.org/10.1089/106652703763255697

[7] Babichev, S.: Technology of wavelet-filtration of the gene expression profiles in order to remove the background noise. Control Systems and Computers **5**, 25–42 (2017)

[8] Babichev, S.: An evaluation of the information technology of gene expression profiles processing stability for different levels of noise components. Data **3**(4), 40 (2018). https://doi.org/10.3390/data3040048

[9] Babichev, S., Durnyak, B., Pikh, I., Senkivskyy, V.: An evaluation of the objective clustering inductive technology effectiveness implemented

using density-based and agglomerative hierarchical clustering algorithms. Advances in Intelligent Systems and Computing **1020**, 532–553 (2020). https://doi.org/10.1007/978-3-030-26474-1_37

[10] Babichev, S., Durnyak, B., Senkivskyy, V., Sorochynskyi, O., Kliap, M., Khamula, O.: Exploratory analysis of neuroblastoma data genes expressions based on bioconductor package tools. In: CEUR Workshop Proceedings. vol. 2488, pp. 268–279 (2019)

[11] Babichev, S., Durnyak, B., Senkivskyy, V., Sorochynskyi, O., Kliap, M., Khamula, O.: Technique of gene regulatory networks reconstruction based on *ARACNE* inference algorithm. In: CEUR Workshop Proceedings. vol. 2488, pp. 195–207. CEUR (2019)

[12] Babichev, S., Durnyak, B., Zhydetskyy, V., Pikh, I., Senkivskyy, V.: Techniques of dna microarray data pre-processing based on the complex use of bioconductor tools and shannon entropy. In: CEUR Workshop Proceedings. vol. 2353, pp. 365–377 (2019)

[13] Babichev, S., Lytvynenko, V., Gozhyj, A., Korobchynskyi, M., Voronenko, M.: A fuzzy model for gene expression profiles reducing based on the complex use of statistical criteria and shannon entropy. Advances in Intelligent Systems and Computing **754**, 545–554 (2019). https://doi.org/10.1007/978-3-319-91008-6_55

[14] Babichev, S., Lytvynenko, V., Korobchynskyi, M., Taif, M.: Objective clustering inductive technology of gene expression sequences features. Communications in Computer and Information Science **716**, 359–372 (2017). https://doi.org/10.1007/978-3-319-58274-0_29

[15] 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 **689**, 21–39 (2016). https://doi.org/10.1007/978-3-319-70581-1_2

[16] Babichev, S., Mikhalyov, O.: A hybrid model of 1-d signal adaptive filter based on the complex use of huang transform and wavelet analysis. International Journal of Intelligent Systems and Applications **11**(2), 1–8 (2019). https://doi.org/10.5815/ijisa.2019.02.01

[17] Babichev, S., Osypenko, V., Lytvynenko, V., Voronenko, M., Korobchynskyi, M.: Comparison analysis of biclustering algorithms with the use of artificial data and gene expression profiles. In: (2018) 2018 IEEE 38th International

Conference on Electronics and Nanotechnology, ELNANO 2018 - Proceedings. pp. 298–304 (2018). https://doi.org/10.1109/ELNANO.2018.8477439

[18] Babichev, S., Sharko, O., Sharko, A., Mikhalyov, O.: Soft filtering of acoustic emission signals based on the complex use of huang transform and wavelet analysis. In: Advances in Intelligent Systems and Computing. vol. 1020, pp. 3–19. Springer (2020). https://doi.org/10.1007/978-3-030-26474-1_1

[19] Babichev, S., Skvor, J., Fiser, J., Lytvynenko, V.: Technology of gene expression profiles filtering based on wavelet analysis. International Journal of Intelligent Systems and Applications **10**(4), 1–7 (2018). https://doi.org/10.5815/ijisa.2018.04.01

[20] Babichev, S., Taif, M., 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 **4**, 54–60 (2016). https://doi.org/10.15588/1607-3274-2016-4-7

[21] Babichev, S., Taif, M., Lytvynenko, V., Osypenko, V.: Criterial analysis of gene expression sequences to create the objective clustering inductive technology. In: 2017 IEEE 37th International Conference on Electronics and Nanotechnology. pp. 244–248 (2017). https://doi.org/10.1109/ELNANO.2017.7939756

[22] Babichev, S., Gozhyj, A., Kornelyuk, A., Lytvynenko, V.: Objective clustering inductive technology of gene expression profiles based on sota clustering algorithm. Biopolymers and Cell **35**(5), 379–392 (2017). https://doi.org/10.7124/bc.000961

[23] Babichev, S., Kornelyuk, A., Lytvynenko, V., Osypenko, V.: Computational analysis of microarray gene expression profiles of lung cancer. Biopolymers and Cell **32**(1), 70–79 (2016). https://doi.org/10.7124/bc.00090F

[24] 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 (2019/08). https://doi.org/https://doi.org/10.2991/eusflat-19.2019.20, `https://doi.org/10.2991/eusflat-19.2019.20`

[25] Baker, F., Hubert, L.: Measuring the power of hierarchical cluster analysis. Journal of the American Statistical Association **70**, 31–38 (1975)

[26] Baldi, P., Hatfield, G.: DNA Microarrays and gene expression: From experiments to data analysis modeling. Cambridge: Cambridge University Press (2002). https://doi.org/10.1017/CBO9780511541773

[27] Bandyopadhyay, S., Pakhira, M.K., U., M.: Validity index for crisp and fuzzy clusters. Pattern Recognition **37**, 487–501 (2004)

[28] Banfield, J., Raftery, A.: Model-based gaussian and non-gaussian clustering. Biometrics **49**, 803–821 (1993)

[29] Barbara, D., Wu, X.: An approximate median polish algorithm for large multidimensional data sets. Springer-Verlag London Ltd. Knowledge and Information Systems **5**, 416–438 (2003)

[30] Beer, D.G., Kardia, S.L., at.al.: Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nature Medicine **8**(8), 816–824 (jul 2002). https://doi.org/10.1038/nm733, `http://www.nature.com/nm/journal/v8/n8/full/nm733.html`

[31] Berte, R., Della Picca, F., Poblet, M., Li, Y., Cortés, E., Craster, R., Maier, S., Bragas, A.: Acoustic far-field hypersonic surface wave detection with single plasmonic nanoantennas. Physical Review Letters **121**(25), 253902 (2018). https://doi.org/10.1103/PhysRevLett.121.253902

[32] Bhattacherjee, V., Mukhopadhyay, P., Singh, S., Johnson, C., at al.: Neural crest and mesoderm lineagedependent gene expression in orofacial development. Differentiation **75**(5), 463–477 (2007)

[33] Bi, Y., Wang, P., Guo, X., Wang, Z., Cheng, S.: K-means clustering optimizing deep stacked sparse autoencoder. Sensing and Imaging **20**(1), 6 (2019). https://doi.org/10.1007/s11220-019-0227-1

[34] Bolstad, B., Irizarry, R., Åstrand, M., Speed, T.: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics **19**(2), 185–193 (2003). https://doi.org/10.1093/bioinformatics/19.2.185

[35] Bowman, D., Wilcock, W.: Unusual signals recorded by ocean bottom seismometers in the flooded caldera of deception island volcano: Volcanic gases or biological activity. Antarctic Science **26**(3), 267–275 (2014)

[36] Buermans, H., Ariyurek, Y., van Ommen, G., den Dunnen, J., Hoen, P.: New methods for next generation sequencing based microrna expression profiling. BMC Genomics **11**(1) (2010). https://doi.org/10.1186/1471-2164-11-716

[37] Calinski, T., Harabasz, J.: A dendrite method for cluster analysis. Communication in statistics **3**, 1–27 (1974)

[38] Chang, Y.S., Yoon, S., Kim, J., Baek, S.Y., Cho, Y., Hong, S., Kim, S., Moon, I.: Standard audiograms for koreans derived through hierarchical clustering using data from the korean national health and nutrition examination survey 2009–2012. Scientific Reports **9**(1), 3675 (2019). https://doi.org/10.1038/s41598-019-40300-7

[39] Chen, T., He, H., Church, G.: Modeling gene expression with differential equations. In: Pacific Symposium on Biocomputing. pp. 29–40. Pacific Symposium on Biocomputing (1999)

[40] Chen, Y.J., Kodell, R., Sistare, F., Thompson, K., Morris, S., Chen, J.: Normalization methods for analysis of microarray gene-expression data. Journal of Biopharmaceutical Statistics **13**(1), 57–74 (2003). https://doi.org/10.1081/BIP-120017726

[41] Chen, Z., McGee, M., Liu, Q., Kong, M., Deng, Y., Scheuermann, R.: A distribution-free convolution model for background correction of oligonucleotide microarray data. BMC Genomics **10**(1), 19 (2009). https://doi.org/10.1186/1471-2164-10-S1-S19

[42] Cheng, Y., Church, J.: Biclustering of expression data. In: Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB'00). pp. 93–103 (2000)

[43] Cui, L., Ma, F., Gu, Q., Cai, T.: Time-frequency analysis of pressure pulsation signal in the chamber of self-resonating jet nozzle. International Journal of Pattern Recognition and Artificial Intelligence **32**(11), 1858006 (2018). https://doi.org/10.1142/S0218001418580065

[44] Desgraupes, B.: Compute clustering validation indices (2018), `https://cran.r-project.org/web/packages/clusterCrit`

[45] D'haeseleer, P., Wen, X., Fuhrman, S., Somogyi, R.: Linear modeling of mrna expression levels during cns development and injury. In: Pacific Symposium on Biocomputing. pp. 41–52. Pacific Symposium on Biocomputing (1999)

[46] Dillies, M.A., Rau, A., Aubert, J., at al.: A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. Briefings in Bioinformatics **14**(6), 671–683 (2013). https://doi.org/10.1093/bib/bbs046

[47] Dorazo, J., Carazo, J.: Phylogenetic reconstruction using an unsupervised growing neural network that adopts the topology of a phylogenetic tree. Journal of Molecular Evolution **44**(2), 226–260 (1997)

[48] Emmert-Streib, F., Dehmer, M., Haibe-Kains, B.: Gene regulatory networks and their applications: Understanding biological and medical problems in terms of networks. Frontiers in Cell and Developmental Biology **2**(AUG), 38 (2014). https://doi.org/10.3389/fcell.2014.00038

[49] Eren, K., Deveci, M., Kucuktunc, O., Catalyurek, U.: A comparative analysis of biclustering algorithms for gene expression data. Briefings in Bioinformatics **14**(3), 279–292 (2012)

[50] Ester, M., Kriegel, H., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial datasets with noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. pp. 226–231 (1996)

[51] Farazi, T., Brown, M., Morozov, P., at al.: Bioinformatic analysis of barcoded cdna libraries for small rna profiling by next-generation sequencing. Methods **58**(2), 171–187 (2012). https://doi.org/10.1016/j.ymeth.2012.07.020

[52] Fawcett, T.: Roc graphs: Notes and practical considerations for researches. Machine learning **31**(1), 1–38 (2004)

[53] Fernandes, J., Chandler, J., Lili, L., Uppal, K., Hu, X., Hao, L., Go, Y.M., Jones, D.: Transcriptome analysis reveals distinct responses to physiologic versus toxic manganese exposure in human neuroblastoma cells. Frontiers in Genetics (676) (2019). https://doi.org/10.3389/fgene.2019.00676

[54] Fränti, P., Sieranoja, S.: K-means properties on six clustering benchmark datasets (2018), `http://cs.uef.fi/sipu/datasets/`

[55] Frid, A., Manevitz, L., Mosafi, O.: Kohonen-based topological clustering as an amplifier for multi-class classification for parkinson's disease. In: 2018 IEEE International Conference on the Science of Electrical Engineering in Israel. p. 8646026 (2018). https://doi.org/10.1109/ICSEE.2018.8646026

[56] Friedman, N., Linial, M., Nachman, I., Pe'er, D.: Using bayesian networks to analyze expression data. Journal of Computational Biology **7**(3-4), 601–620 (2000). https://doi.org/10.1089/106652700750050961

[57] Fritzke, B.: Growing cell structures a self-organizing network for unsupervised and supervised learning. Neural Networks **7**(9), 1441–1461 (1994)

[58] Garmire, L., Subramaniam, S.: Evaluation of normalization methods in mammalian microrna-seq data. RNA **18**(6), 1279–1288 (2012). https://doi.org/10.1261/rna.030916.111

[59] Gentleman, R., Carey, V., Huber, W., Irizarry, R., Dudoit, S.: Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Springer (2005)

[60] Gionis, A., Mannila, H., Tsaparas, P.: Clustering aggregation. ACM Transactions on Knowledge Discovery from Data (TKDD) **1**(1), 1–30 (2007)

[61] Gómez, S., Rodríguez, J., Rodríguez, F., Juez, F.: Analysis of the temporal structure evolution of physical systems with the self-organising tree algorithm (sota): Application for validating neural network systems on adaptive optics data before on-sky implementation. Entropy **19**(3), 103 (2017). https://doi.org/10.3390/e19030103

[62] Gong, K., Hu, J.: Online detection and evaluation of tank bottom corrosion based on acoustic emission. In: Springer Series in Geomechanics and Geoengineering. vol. 216039, pp. 1284–1291. Springer-Verlag (2019). https://doi.org/10.1007/978-981-10-7560-5_118

[63] Hackenberg, M., Rodríguez-Ezpeleta, N., Aransay, A.: Miranalyzer: An update on the detection and analysis of micrornas in high-throughput sequencing experiments. Nucleic Acids Research **39**(2), W132–W138 (2011). https://doi.org/10.1093/nar/gkr247

[64] Hackenberg, M., Sturm, M., Langenberger, D., Falcón-Pérez, J., Aransay, A.: miranalyzer: A microrna detection and analysis tool for next-generation sequencing experiments. Nucleic Acids Research **37**(2), W68–W76 (2009). https://doi.org/10.1093/nar/gkp347

[65] Hahsler, M., Piekenbrock, M., S., A., Mount, D.: Density based clustering of applications with noise (dbscan) and related algorithms (2019), `https://github.com/mhahsler/dbscan`

[66] Harrington, J.: The desirability function. Industrial Quality Control, **21**(10), 494–498 (1965), `http://asq.org/qic/display-item/?item=4860`

[67] Hausser, J., Strimmer, K.: Entropy inference and the james-stein estimator, with application to nonlinear gene association networks. Journal of Machine Learning Research **10**, 1469–1484 (2009)

[68] Heather, J., Chain, B.: The sequence of sequencers: The history of sequencing dna. Genomics **107**, 1–8 (2016)

[69] Huang, N., Shen, Z., Long, S., Wu, M., Shih, H., Zheng, N., Yen, N., Tung, C., Liu, H.: The empirical mode decomposition and hilbert spectrum for nonlinear and nonstationary time series analysis. In: Proceedings of the Royal Society a Mathematical, Physical and Engineering Sciences. vol. 454, pp. 903—-995. The Royal Society (1998). https://doi.org/10.1098/rspa.1998.0193

[70] Huang, N., Wu, M.L., Qu, W., Long, S., Shen, S.: Applications of hilbert-huang transform to non-stationary financial time series analysis. Applied Stochastic Models in Business and Industry **19**(3), 245–268 (2003). https://doi.org/10.1002/asmb.501

[71] Huang, N., Wu, Z.: A review on hilbert-huang transform: Method and its applications to geophysical studies. Reviews of Geophysics **46**(2), RG2006 (2008). https://doi.org/10.1029/2007RG000228

[72] Huang, P.J., Liu, Y.C., Lee, C.C., Lin, W.C., Gan, R.C., Lyu, P.C., Tang, P.: Dsap: Deep-sequencing small rna analysis pipeline. Nucleic Acids Research **38**(2), W385–W391 (2010). https://doi.org/10.1093/nar/gkq392

[73] Huber, W., Carey, V., Gentleman, R., et al.: Orchestrating high-throughput genomic analysis with bioconductor. Nat Methods **12**, 115–121 (2015). https://doi.org/doi:10.1038/nmeth.3252

[74] Hubert, L., Schultz, J.: Quadratic assignment as a general data-analysis strategy. British Journal of Mathematical and Statistical Psychologie **29**, 190–241 (1976)

[75] Ihaka, R., Gentleman, R.: R: a linguage for data analysis and graphics. Journal of Computational and Graphical Statistics **5**(3), 299–314 (1996)

[76] Irizarry, R., Hobbs, B., Collin, F., Beazer-Barclay, Y., Antonellis, K., Scherf, U., Speed, T.: Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Selected Works of Terry Speed pp. 601–616 (2012). https://doi.org/10.1007/978-1-4614-1347-9_15

[77] Ivakhnenko, A.: Group method of data handling as competitor to the method of stochastic approximation. Soviet Automatic Control **3**, 64–78 (1968)

[78] Ivakhnenko, A.: Objective clustering based on the theory of self-organization models. Automatics **5**, 6–15 (1987)

[79] Ivakhno, S., Kornelyuk, A.: Microarrays: Technologies overview and data analysis. Ukrainian Biochemical Journal **76**(2), 5–19 (2004)

[80] Jain, A., Law, M.: Data clustering: A user's dilemma. Lecture Notes in Computer Science **3776**, 1–10 (2005)

[81] Kaiser, S.: Biclustering: Methods, software and application (2011)

[82] Kaiser, S., Santamaria, R., Khamiakova, T., Sill, M., at al.: Package 'biclust' (2018), `https://cran.r-project.org/web/packages/biclust`

[83] Kalman, R.: A new approach to linear filtering and prediction problems. Journal of Basic Engineering **82**(1), 35–45 (1960)

[84] Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., Morishima, K.: Kegg: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Research **45**, 353–361 (2017)

[85] Kanehisa, M., Goto, S.: Kegg: Kyoto encyclopedia of genes and genomes. Nucleic Acids Research **28**, 27–30 (2000)

[86] Kanehisa, M., Sato, S., Kawashima, M., Furumichi, M., Tanabe, M.: Kegg as a reference resource for gene and protein annotation. Nucleic Acids Research **44**, 457–462 (2016)

[87] Kassambara, A., Mundt, F.: Extract and visualize the results of multivariate data analyses (2017), `http://www.sthda.com/english/rpkgs/factoextra`

[88] Khatoon, M., Banu, W.: An efficient method to detect communities in social networks using dbscan algorithm. Social Network Analysis and Mining **9**(1), 9 (2019). https://doi.org/10.1007/s13278-019-0554-1

[89] Kluger, Y., Basry, R., Chang, J., Gerstein, M.: Spectral biclustering of microarray data: coclustering genes and conditions. Genome Resources **12**(4), 703–716 (2003)

[90] Kohane, I., Kho, A., Butte, A.: Microarrays for an integrative genomics (2003)

[91] Lazaridis, E., Sinibaldi, D., Bloom, G., Mane, S., Jove, R.: A simple method to improve probe set estimates from oligonucleotide arrays. Mathematical Biosciences **176**(1), 53–58 (2002). https://doi.org/10.1016/S0025-5564(01)00100-6

[92] Li, C., Liu, L., Sun, X., Zhao, J., Yin, J.: Image segmentation based on fuzzy clustering with cellular automata and features weighting. Eurasip Journal on Image and Video Processing **1**, 37 (2019). https://doi.org/10.1186/s13640-019-0436-5

[93] Li, J., Reisner, J., Pham, H., Olafsson, S., Vardeman, S.: Biclustering with missing data. Information Sciences **510**, 304–316 (2020). https://doi.org/10.1016/j.ins.2019.09.047

[94] Li, W., Kuang, G., Xiong, B.: Decomposition of multicomponent micro-doppler signals based on hht-amd. Applied Sciences **8**(10), 1801 (2018). https://doi.org/10.3390/app8101801

[95] Li, Y., Zhang, Z., Liu, F., Vongsangnak, W., Jing, Q., Shen, B.: Performance comparison and evaluation of software tools for microrna deep-sequencing data analysis. Nucleic Acids Research **40**(10), 4298–4305 (2012). https://doi.org/10.1093/nar/gks043

[96] Liang, S., Fuhrman, S., Somogyi, R.: Reveal, a general reverse engineering algorithm for inference of genetic network architectures. In: Pacific Symposium on Biocomputing. pp. 18–29. Pacific Symposium on Biocomputing (1998)

[97] Liang, W.S., Dunckley, T., Beach, T., Grover, A., at al.: Gene expression profiles in anatomically and functionally distinct regions of the normal aged human brain. Physiological Genomics **28**(3), 311–322 (2007)

[98] Liang, W.S., Reiman, E.M., Valla, J.a.a.: Alzheimer's disease is associated with reduced expression of energy metabolism genes in posterior cingulate neurons. Proceedings of the National Academy of Sciences of the United States of America **105**(11), 4441–4446 (2008)

[99] Liu, X.P., Wu, C., Miao, H., Wu, H.: Regnetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. Database pp. 1–12 (2015). https://doi.org/10.1093/database/bav095

[100] Ma, C.Y., Chen, Y.P., Berger, B., Liao, C.S.: Identification of protein complexes by integrating multiple alignment of protein interaction networks. Bioinformatics **33**(11), 1681–1688 (2017). https://doi.org/10.1093/bioinformatics/btx043

[101] Ma, X., Zhou, G., Shang, J., Wang, J., Peng, J., Han, J.: Detection of complexes in biological networks through diversified dense subgraph mining. Journal of Computational Biology **24**(6), 923–941 (2017). https://doi.org/10.1089/cmb.2017.0037

[102] Madala, H., Ivakhnenko, A.: Inductive Learning Algorithms for Complex Systems Modeling, chap. 5:Clusterization and Recognition, p. 380. CRC Press (1994)

[103] Madeira, S., Oliveira, A.: Biclustering algorithms for biological data analysis: A survey. In: IEEE/ACM Transactions on Computational Biology and Bioinformatics. vol. 1, pp. 24–45 (2004)

[104] Margolin, A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R., Califano, A.: Aracne: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics **7**(SUPPL.1), S7 (2006). https://doi.org/10.1186/1471-2105-7-S1-S7

[105] Mayer, G., Marcus, K., Eisenacher, M., Kohl, M.: Boolean modeling techniques for protein co-expression networks in systems medicine. Expert Review of Proteomics **13**(6), 555–569 (2016). https://doi.org/10.1080/14789450.2016.1181546

[106] Meysman, P., Titeca, K., Eyckerman, S., Tavernier, J., Goethals, B., Martens, L., Valkenborg, D., Laukens, K.: Protein complex analysis: From raw protein lists to protein interaction networks. Mass Spectrometry Reviews **36**(5), 600–614 (2015). https://doi.org/10.1002/mas.21485

[107] Morin, R., O'Connor, M., Griffith, M., Kuchenbauer, F., at al.: Application of massively parallel sequencing to microrna profiling and discovery in human embryonic stem cells. Genome Research **18**(4), 610–621 (2008). https://doi.org/10.1101/gr.7179508

[108] Mukhopadhyay, A., Maulik, U., Bandyopadhyay, S.: On biclustering of gene expression data. Current Bioinformatics **5**, 204–216 (2010)

[109] Nair, A., Chetty, M., Wangikar, P.: Improving gene regulatory network inference using network topology information. Molecular BioSystems **11**(9), 2449–2463 (2015). https://doi.org/10.1039/c5mb00122f

[110] Osypenko, V.V., Reshetjuk, V.M.: The methodology of inductive system analysis as a tool of engineering researches analytical planning. Agricultural and Forest Engineering **58**, 67–71 (2011), `http://annals-wuls.sggw.pl/?q=node/234`

[111] Oweis, R., Abdulhay, E.: Seizure classification in eeg signals utilizing hilbert-huang transform. BioMedical Engineering Online **10**, 38 (2011). https://doi.org/10.1186/1475-925X-10-38

[112] Pantano, L., Estivill, X., Martí, E.: Seqbuster, a bioinformatic tool for the processing and analysis of small rnas datasets, reveals ubiquitous mirna modifications in human embryonic cells. Nucleic Acids Research **38**(5), e34.1–e34.13 (2009). https://doi.org/10.1093/nar/gkp1127

[113] Park, T., Yi, S.G., Kang, S.H., Lee, S., Lee, Y.S., Simon, R.: Evaluation of normalization methods for microarray data. BMC Bioinformatics **4**, 13 (2003). https://doi.org/10.1186/1471-2105-4-33

[114] Pontes, B., Giráldez, R., Aguilar-Ruiz, J.: Biclustering on expression data: A review. Journal of Biomedical Informatics **57**, 163–180 (2015)

[115] Prelic, A., Bleuler, S., Zimmermann, P., Wille, A., at al.: A systematic comparison and evaluation of biclustering methods for gene expression data. Bioinformatics **22**(9), 1122–1129 (2006)

[116] R.A., F.: The use of multiple measurements in taxonomic problems. Annals of Eugenetics **7**, 179–188 (1936)

[117] Raddatz, B., Spitzbarth, I., Matheis, K., Kalkuhl, A., Deschl, U., Baumgärtner, W., Ulrich, R.: Microarray-based gene expression analysis for veterinary pathologists: A review. Veterinary Pathology **54**(5), 734–755 (2017). https://doi.org/10.1177/0300985817709887

[118] Ray, S., Turi, R.: Determination of number of clusters in k-means clustering and application in colour image segmentation. In: Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques. pp. 137–143 (1999)

[119] Riabova, S.: Application of wavelet analysis to the analysis of geomagnetic field variations. Journal of Physics: Conference Series **1141**(1), 012146 (2018). https://doi.org/10.1088/1742-6596/1141/1/012146

[120] Ridgley, K., Abouhussien, A., Hassan, A., Colbourne, B.: Characterisation of damage due to abrasion in scc by acoustic emission analysis. Magazine of Concrete Research **71**(2), 85–94 (2019). https://doi.org/10.1680/jmacr.17.00445

[121] Robinson, M., Oshlack, A.: A scaling normalization method for differential expression analysis of rna-seq data. Genome Biology **11**(3) (2010). https://doi.org/10.1186/gb-2010-11-3-r25

[122] Ronen, R., Gan, I., Modai, S., Sukacheov, A., Dror, G., Halperin, E., Shomron, N.: A software for microrna deep sequencing analysis. Bioinformatics **26**(20), 2615–2616 (2010). https://doi.org/10.1093/bioinformatics/btq493

[123] Ros, F., Guillaume, S.: A hierarchical clustering algorithm and an improvement of the single linkage criterion to deal with noise. Expert Systems with Applications **128**, 96–108 (2019). https://doi.org/10.1016/j.eswa.2019.03.031

[124] Rousseeuw, P.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics **20**, 53–65 (1987)

[125] Sarycheva, L.: Objective cluster analysis of data based on the group method of data handling. Problems of Control and Automatics **2**, 86–104 (2008)

[126] Schena, M., Davis, R.W.: Microarray biochip technology. Eaton Publishing pp. 1–18 (2000)

[127] Shannon, P.: Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Research **3**(11), 2498–2504 (2003). https://doi.org/10.1101/gr.1239303

[128] Silva, E., da Silva, E., Silva, D., Novaes, C., Amorim, F., dos Santos, M., Bezerra, M.: Evaluation of macro and micronutrient elements content from soft drinks using principal component analysis and kohonen self-organizing maps. Food Chemistry **273**, 9–14 (2019). https://doi.org/10.1007/978-3-319-91008-6_58

[129] Soualhi, A., Medjaher, K., Zerhouni, N.: Bearing health monitoring based on hilbert-huang transform, support vector machine, and regression. In: IEEE Transactions on Instrumentation and Measurement. vol. 64, pp. 52–62 (2015). https://doi.org/10.1109/TIM.2014.2330494

[130] Staub, S., Andrä, H., Kabel, M.: Fast fft based solver for rate-dependent deformations of composites and nonwovens. International Journal of Solids and Structures **154**, 33–42 (2018). https://doi.org/10.1016/j.ijsolstr.2016.12.014

[131] Stepashko, V.: Inductive modeling from historical perspective. In: Proceedings of the 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT 2017. vol. 1, pp. 537–542 (2017). https://doi.org/10.1109/STC-CSIT.2017.8098845

[132] Su, F., Li, T., Pan, X., Miao, M.: Acoustic emission responses of three typical metals during plastic and creep deformations. Experimental Techniques **42**(6), 685–691 (2018). https://doi.org/10.1007/s40799-018-0274-x

[133] Susanto, A., Liu, C.H., Yamada, K., Hwang, Y.R., Tanaka, R., Sekiya, K.: Application of hilbert–huang transform for vibration signal analysis in end-milling. Precision Engineering **53**, 263–277 (2018). https://doi.org/10.1016/j.precisioneng.2018.04.008

[134] Susanto, A., Liu, C.H., Yamada, K., Hwang, Y.R., Tanaka, R., Sekiya, K.: Milling process monitoring based on vibration analysis using hilbert-huang transform. International Journal of Automation Technology **12**(5), 688–698 (2018). https://doi.org/10.20965/ijat.2018.p0688

[135] Trusiak, M., Styk, A., Patorski, K.: Hilbert–huang transform based advanced bessel fringe generation and demodulation for full-field vibration studies of specular reflection micro-objects. Optics and Lasers in Engineering **102**, 100–112 (2018). https://doi.org/10.1016/j.optlaseng.2018.05.021

[136] Wan, R., Xiong, N., Hu, Q., Wang, H., Shang, J.: Similarity-aware data aggregation using fuzzy c-means approach for wireless sensor networks. Eurasip Journal on Wireless Communications and Networking **1**, 59 (2019). https://doi.org/10.1186/s13638-019-1374-8

[137] Wiener, N.: Extrapolation, Interpolation, and Smoothing of Stationary Time Series. Wiley, New York (1964). https://doi.org/10.7551/mitpress/2946.001.0001

[138] Wong, K.C., Li, Y., Zhang, Z.: Unsupervised learning in genome informatics, pp. 405–448 (2016). https://doi.org/10.1007/978-3-319-24211-8_15

[139] Xie, X., Beni, G.: A validity measure for fuzzy clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence **13**(4), 841–846 (1991)

[140] Xu, J., Shu, S., Han, Q., Liu, C.: Experimental research on bond behavior of reinforced recycled aggregate concrete based on the acoustic emission technique. Construction and Building Materials **191**, 1230–1241 (2018). https://doi.org/10.1016/j.conbuildmat.2018.10.054

[141] Yan, B., Guan, D., Wang, C., Wang, J., He, B., Qin, J., Boheler, K., Lu, A., Zhang, G., Zhu, H.: An integrative method to decode regulatory logics in gene transcription. Nature Communications **8**(1), 1044 (2017). https://doi.org/10.1038/s41467-017-01193-0

[142] Yefimenko, S., Stepashko, V.: Intelligent recurrent-and-parallel computing for solving inductive modeling problems. In: Proceedings - 2015 16th International Conference on Computational Problems of Electrical Engineering, CPEE 2015. pp. 236–238 (2015). https://doi.org/10.1109/CPEE.2015.7333385

[143] Yuan, H., Liu, X., Liu, Y., Bian, H., Chen, W., Wang, Y.: Analysis of acoustic wave frequency spectrum characters of rock mass under blasting damage based on the hht method. Advances in Civil Engineering p. 9207476 (2018). https://doi.org/10.1155/2018/9207476

[144] Zadeh, L.: Fuzzy logic = computing with words. IEEE Transactions on Fuzzy Systems **4**(2), 103–111 (1996). https://doi.org/10.1109/91.493904, `http://ieeexplore.ieee.org/document/493904/`

[145] Zahn, C.: Graph-theoretical methods for detecting and describing gestalt clusters. IEEE Transactions on Computers **100**(1), 68–86 (1971)

[146] Zak, D., Vadigepalli, R., Gonye, E., Doyle, F., at al.: Unconventional systems analysis problem in molecular biology: a case study in gene regulatory network modeling. Computational and Chemical Engineering **29**(3), 547–563 (2005)

[147] Zalik, R.: On orthonormal wavelet bases. Journal of Computational Analysis and Applications **27**(5), 790–797 (2019)

[148] Zhang, X., Zou, Z., Wang, K., Hao, Q., Wang, Y., Shen, Y., Hu, H.: A new rail crack detection method using lstm network for actual application based on ae technology. Applied Acoustics **142**, 78–86 (2018). https://doi.org/10.1016/j.apacoust.2018.08.020

[149] Zhao, Q., Xu, M., Fränti, P.: Sum-of-squares based cluster validity index and significance analysis. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **5495**, 313–322 (2009). https://doi.org/doi:10.1007/978-3-642-04921-7_32

For notes

Scientific edition

Sergii Babichev and Bohdan Durnyak

Methods, Models and Information Technology of Complex Data Processing in the Fields of Technical Diagnostics and Bioinformatics

Monograph

Ukrainian Academy of Printing
19, Pid Holoskom Str., Lviv, 79020
Certificate of State Registration
DK No. 3050 of 11.12.2007

Signed for printing 18.12.2019
Format 70 × 100/16. Offset paper. Offset printing technique.
Print run 16.01.2020. Order No. 500.

Printed in TPLPC
of Ukrainian Academy of Printing
3, Lychakivska Str., Lviv, 79008