Evaluation of the Gene Expression Profiles Complex Proximity Metric Effectiveness Based on a Hybrid Technique of Gene Expression Data Extraction

Lyudmyla Yasinska-Damri[a], Igor Liakh[b], Sergii Babichev[c] and Bohdan Durnyak[a]

[a] *Ukrainian Academy of Printing, Pid Goloskom street, 19, Lviv, 79000, Ukraine*
[b] *Uzhhorod National University, University street, 14, Uzhhorod, 88000, Ukraine*
[c] *Kherson State University, University street, 27, Kherson, 73000, Ukraine*

Abstract

Gene expression data processing in order to develop the systems of complex diseases diagnostic or/and gene regulatory networks (GRN) reconstruction is one of the actual direction of modern bioinformatics. One of the important stages of this problem solving is an extraction of mutually correlated gene expression profiles (GEP) considering the used proximity metric. Within the framework of our research, we evaluate the complex metric of GEP proximity calculated as the combination of modified mutual information criterion and Pearson's chi-squared test using OPTICS clustering algorithm implemented using principles of the objective clustering inductive technique (OCIT). The examined objects classification accuracy was used as the main criterion to access the applied method effectiveness. The simulation results have shown that the proposed technique allows us to form an optimal GEP cluster structure in terms of maximum values of the patterns classification accuracy quality criterion.

Keywords

Gene expression profiles, proximity metrics, OPTICS clustering algorithm, gene expression profiles classification, inductive methods of objective clustering, clustering quality criteria, classification accuracy

1. Introduction and literature review

The development of models of diseases diagnostics or/and gene regulatory networks (GRN) reconstruction using gene expression data (GED) is one of the actual directions of modern bioinformatics. As a rule, the initial GED is formed as a high dimensional array with components represented the studied patterns and genes. The value of gene expression is depended on the amount of this type of gene that determines the appropriate properties of the examined biological organism. Gene expression profile (GEP) means the vector of gene expressions the values of which are evaluated for the examined patterns.

Reconstruction of gene regulatory network (GRN) which adequate reflect the nature of genes interaction under the different states of a biological organism in order to develop both effective medicine and disease diagnostic and treating methods is possible provided the extraction of groups of highly and mutually expressed genes. For this reason, the stage of gene expression data pre-processing is very important at the early stage of GRN forming or under the development of a disease diagnosing model. Figure 1 illustrates a stepwise procedure for implementing this process. The filtration procedure, in this case, involves removing genes with zero expression at the first step and genes with low expression in terms of the empirically established threshold at the second step.

IDDM-2021: 4th International Conference on Informatics & Data-Driven Medicine, November 19–21, 2021 Valencia, Spain
EMAIL: Lm.yasinska@gmail.com (L. Yasinska-Damri); ihor.lyah@uzhnu.edu.ua (I. Liakh); sbabichev@ksu.ks.ua (S. Babichev); durnyak@uad.lviv.ua (B. Durnyak)
ORCID: 0000-0002-8629-8658 (L. Yasinska-Damri); 0000-0001-5417-9403 (I. Liakh); 0000-0001-6797-1467 (S. Babichev); 0000-0003-1526-9005 (B. Durnyak)

Moreover, data can contain gene expression profiles that are statistically significantly different from the GEP of the main group. It is obvious that such genes do not correlate with the profiles of other genes and they can also be removed from the data. Qualitative implementation of this stage allows significantly reducing the number of genes for further research. This fact also contributes to enhancing the quality of further steps of GED processing for the solving hereinafter described problem.

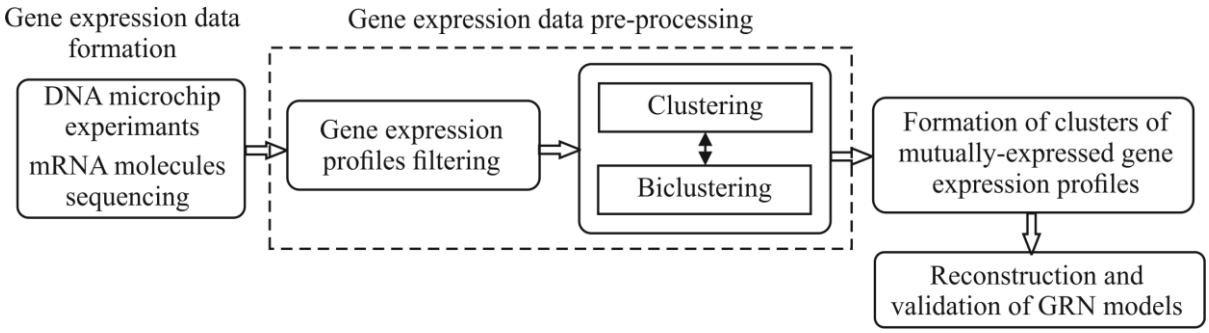

Figure 1: Block-chart of a step-by-step procedure of GED processing to form clusters of highly and mutually expressed GEP

In [1], the authors presented the “limma” module (Linear Models for Microarray and RNA-Seq Data), which contains various functions for generating, filtering and interpreting gene expression data obtained using both DNA microchips experiments and mRNA molecules sequencing method. This module is to some extent an alternative to the “Bioconductor” package, implemented in the data mining and machine learning R software [2] and it is based on the use of linear models to allocate differently expressed genes in a multifactor experiment. This module also contains functions for the genes ontology analysis, which is very important for adequate GRN reconstruction, because the interpretation of genes and their interactions based on the analysis of conceptual interconnections allows identifying target genes, to establish the nature of interconnections between target and other genes taking into account appropriate disease.

The papers [3-5] considered a various tools and techniques of GED filtering that are available in the "Bioconductor" package using quantitative quality criteria for GED received by DNA microarray method [3,4] and mRNA molecules sequencing [5]. As a simulation result, the authors proposed a stepwise algorithm for extracting highly and mutually expressed gene expression profiles for their further grouping into clusters. In a review [6], the authors conducted a comparative analysis of current software to process the GED for purpose of extracting the most informative genes. The analysis of the authors' research allows concluding on the feasibility of using the R software for GEP processing in order to form clusters of highly and mutually expressed genes because this software contains all necessary modules and functions to process gene expression data according to the solved task.

The review [7] presents the research results focused on the study of various hybrid techniques to extract the clusters of mutually correlated GEP to solve the problem of creation of the system of cancer disease diagnostic. In the reviewed works, various combinations of filtering, clustering and classification techniques using various types of statistical criteria and gene expression profiles proximity metrics were applied. The examined objects classification accuracy was applied as the principal quality metric to assess the appropriate hybrid model effectiveness. The following filtration techniques and methods to estimate the gene expression profiles proximity were analyzed in this review: mutual information maximization method [8], χ^2 Pearson's test [9], correlation-based feature selection technique [10], Laplacian and Fisher score [11], information gain method [12], Fisher criterion [13], independent component analysis [14], maximum relevance minimum redundancy [15], probabilistic random function [16], random forest ranking [17], Fisher-Markov selector [18], symmetrical uncertainty [19] and logarithmic transformation [20] method. However, we would like to note that in the analyzed research high classification accuracy in most cases is achieved when using a low number of the extracted GEP. Moreover, the parameters of the respective technique used in the appropriate hybrid models are set upped empirically when the simulation process is performed. Undoubtedly, this fact is one of the main disadvantages of the analyzed models.

The works [21,22] presents the partial decision of this task. A stepwise procedure of GEP extraction on the basis of the joint application of Shannon entropy, statistical criteria, clustering technique based on the SOTA clustering algorithm and random forest binary classifier was developed in these papers. The suitable algorithm parameters considering the classification accuracy were set a priory according to the OCIT principles**.** However, only correlation proximity metric was used within the framework of the authors' research. Thus, the presented hereinbefore brief review allows concluding that an effective model of GEP extraction based on joint application of various proximity metrics, clustering and classification techniques is absent now. This problem can be solved on the basis of joint application of various techniques used successfully in current data science directions of scientific research nowadays [23-26].

In this work, we consider the GEP hybrid proximity metric calculated as a combination of modified mutual information maximization method and Pearson's χ^2 test. The modified mutual information maximization method, in this instance, takes into account various methods of Shannon entropy evaluation.

The **objective of the research** is the development and evaluation of a hybrid model of GEP extraction on the basis of joint application of hybrid proximity metric, OPTICS clustering algorithm implemented using principles of OCIT and random forest binary classifier.

2. Materials and methods

In the general instance, the clustering internal quality criterion should consider both the gene expression profiles allocation inside clusters and clusters' medians allocation relative to each other. Thus, this criterion should be complex and contains two components. If we assume that *K* is the number of clusters, then the formula for assessing the first component of this criterion can be calculated in the following way:

$$QCW = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{N_k}\sum_{i=1}^{N_k} d(e_i, C_k) \tag{1}$$

where: N_k and C_k are the number of GEP in *k*-th cluster and the median of *k*-th cluster respectively; $d(e_i, C_k)$ is the distance between *i*-th profiles and median of this cluster calculated using complex proximity metric which contained both the modified mutual information maximization method (considered various methods of Shannon entropy calculation) and Pearson's χ^2 test the effectiveness of which is proved in [27].

The second component of the internal criterion can be assessed as the average distance between the allocated clusters' medians:

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

In [21], the authors performed modelling to assess the performance of different types of internal criteria, containing (1) and (2) as the components. As a result, a hybrid internal criterion formed as a ratio of Calinski-Harabasz criterion and WB index has been proposed:

$$QC^{\text{int}} = \frac{K(K-1)QCW^2}{(N-K)QCB^2} \tag{3}$$

where *N* is the number of objects that should be grouped. This criterion was used as the internal one during the modelling procedure performing.

Assessment of the efficiency of both the GEP hybrid proximity metric and quality criteria when the profiles grouping into clusters was performed based on the application of density clustering algorithm Optics [28], which is a logical development of DBSCAN density algorithm and allows us to form a multicluster structure based on the application of respective proximity metric. The feasibility of using the OPTICS clustering algorithm is determined by the fact that its application allows us not only to form a multicluster structure containing clusters of close gene expression profiles by density in their allocation in feature space but also to allocate profiles identified as noise because of density of

their allocation relative to other GEP is much lower compared to the density of the main groups of GEP distribution.

We would like to note that the criterion calculated by formulas (1) – (3) does not always allow us to objectively form an adequate clustering due to the reproducibility error, which is inherent to most prevailing clustering algorithms. In other words, satisfactory results of data grouping gotten using one dataset are not always repeated when applying another similar dataset. In [29], the authors proposed the idea of reducing the reproducibility error by using "fresh data" (not used when creating the model) during the process of verifying the obtained model of object distribution into clusters and making the final decision regarding the cluster structure formation by joint using the internal, external and balance criteria, which considered possible discrepancies between internal and external criteria. This idea was further developed in [30,31] where the objective clustering inductive technology was described and implemented. The authors proposed an external quality criterion assessed in the form of normalized distinction of the internal criteria assessed on two equivalent subsets (contained the same number of pairwise similar objects) at the appropriate hierarchical level of cluster structure formation:

$$QC^{ext} = \frac{\left| QC_1^{\text{int}} - QC_2^{\text{int}} \right|}{QC_1^{\text{int}} + QC_2^{\text{int}}} \tag{4}$$

The main idea was as follows. The minimal reproducibility error matches the maximum degree of the similarity of objects allocation in clusters obtained on two equivalent subsets. Since the internal criteria consider the nature of both the patterns distribution in clusters and the clusters' medians allocation relative to each other, objective clustering (minimum value of reproducibility error) in this case corresponds to the minimal difference between the corresponding values of the internal criteria. The normalizing correction in formula (4) transforms the range of the external criteria values variation from 0 (zero reproducibility error) to 1 (maximum error). The balance criterion was calculated using the Harrington desirability function according to the algorithm described in detail in [30,31].

The random forest classifier was used to implement this step. This choice is determined by the previous authors' research, presented in [21], where various types of binary classifiers were studied to classify the samples of patients examined on lung cancer. These samples contained gene expression data as attributes too. The effectiveness of the respective model was assessed using the examined samples classification accuracy.

Figure 2 shows a block chart of the stepwise procedure performed within the framework of the modelling procedure executing. The practical implementation of this algorithm assumes the following stages:

Stage I. Formation of GEP data and functions to calculate respective criteria.

1.1. Forming a array of GED, the components of which represent the assessed patterns and genes whose expression determines the relative amount of a given type of gene for the examined patterns respectively.

1.2. Formation of the function to estimate the proximity metrics between GEP on the basis of the joint application of the modified mutual information maximization proximity metric and Pearson's χ^2 test [28].

1.3. Formation of the functions to calculate the internal, external and hybrid balance quality criteria.

1.4. Formation of the function to calculate the examined samples classification accuracy.

1.5. Formation of two equivalent subsets of GEP by the iterative distribution of the two nearest GEP according to a hybrid proximity metric into two equivalent subsets.

Stage II. Setup of density-based OPTICS clustering algorithm.

2.1. Setup of range for changing the minimum number of points within the *ε-neighborhood*: $MinPts_{min}$, $MinPts_{max}$.

2.2. Creating a reachability chart. Setup of both the range and step of variation of the *ε-neighborhood* values: Eps_{min}, Eps_{max}, *dEps.*

2.3. Calculation of distances between all pairs of gene expression profiles in equal-power subsets and formation of matrixes of distances between the corresponding profiles. The obtained distance

matrixes will be used as input data when the clustering procedure is implemented by applying the OPTICS algorithm.

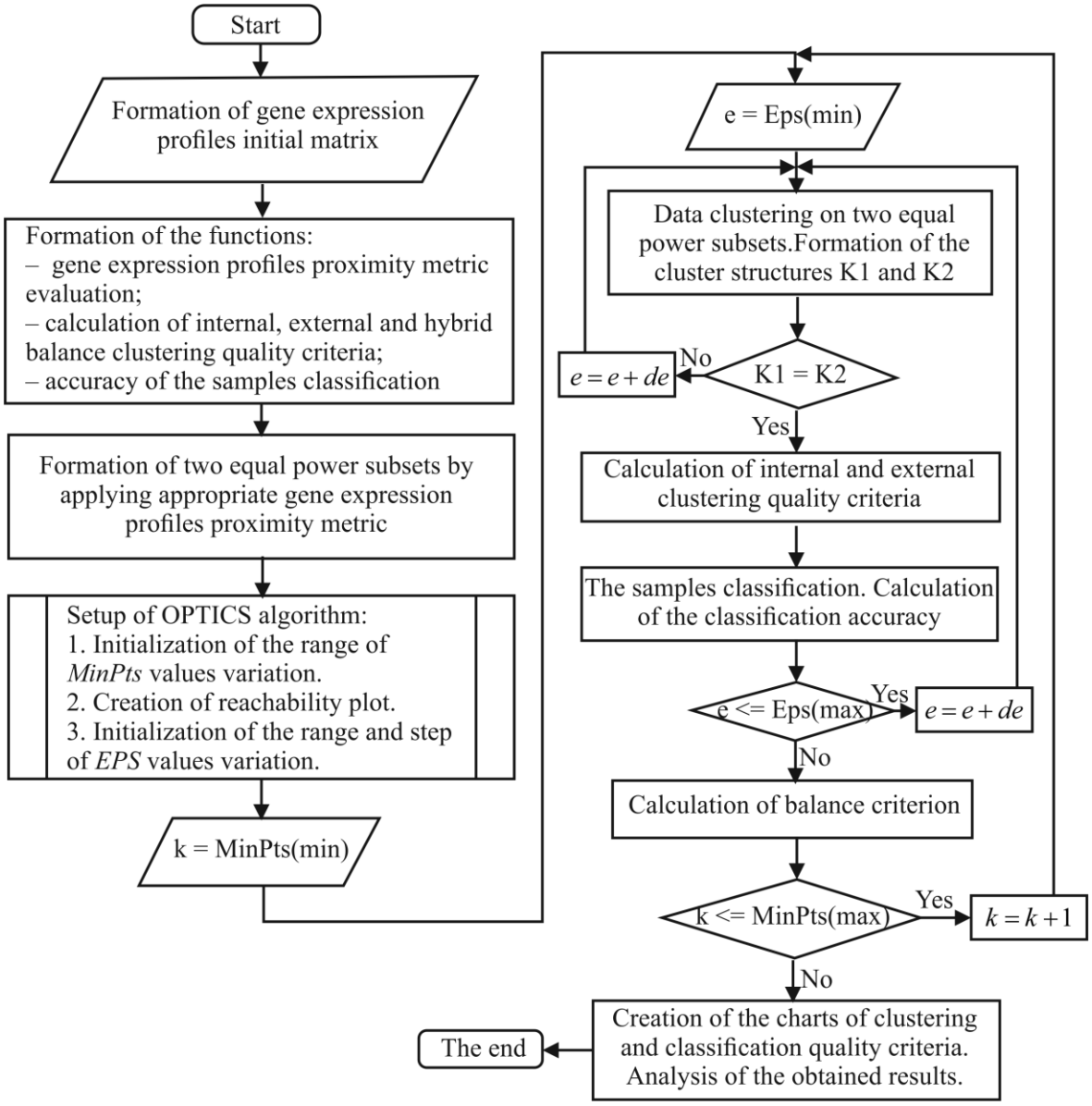

Figure 2: Structural block-chart of the algorithm for forming a multicluster structure based on the OPTICS algorithm implemented using the principles of OCIT

Stage III. Stepwise clustering of GEP within the specified ranges of the algorithm appropriate parameters variation.

3.1. *MinPts* value initialization: $k = MinPts_{min}$.

3.2. *Eps* value initialization: $e = Eps_{min}$.

3.3. Clustering of gene expression profiles contained in equivalent subsets, forming the partitions with the number of clusters *K1* and *K2*.

3.4. If $K1 = K2 > 2$, calculation of internal and external quality criteria by formulas (1) - (4). Otherwise, increase the value of *Eps* parameter (e = e + de) and go to step 3.3 of this procedure.

3.5. Classification of objects that contain gene expression profiles in each of the allocated clusters. Calculation of the classification quality criterion (Accuracy).

3.6. If $e \le Eps_{\max}$, go to step 3.3 of this procedure. Otherwise, calculate the hybrid balance criterion and increase the *MinPts* value by one: $k = k + 1$.

3.7. If $k \le MinPts_{\max}$, go to step 3.2 of this procedure. Otherwise, the creation of charts of the clustering and classification quality criteria depending on the *Eps* value for each of the *MinPts* values.

Stage IV. An analysis of the obtained results.

4.1. An analysis of the obtained charts. Forming conclusions regarding the effectiveness of hybrid metrics of GEP proximity in the process of forming subsets of informative genes for their further use when the creation of disease diagnosing systems or/and GRN reconstruction.

3. Experiment, results and discussion

The practical implementation of the proposed algorithm was carried out using the GSE19188 gene expressions dataset of patients studied for the early stage of lung cancer [32]. The data were obtained using a DNA microchips experiment and contained 156 microchips, 65 of them contained GED of healthy patients and 91 ones included the GED of patients with lung cancer tumor (mild form). 400 the most informative GEP in terms of classification accuracy (approximately 93%) [20,21] were used during the simulation procedure implementation.

The *MinPts* value was changed within the limits of 3 to 5. This interval was established empirically. The results of the modelling showed that a larger quantity of points within the Eps neighborhood degrades the simulation results both in terms of the number of clusters in the equal over subsets and in terms of gene expression profiles clustering quality criteria and the samples classification accuracy. The Eps values were varied from the minimum, which was calculated as the minimum distance between gene expression profiles in equal-power subsets to a 1.5 minimum distance. This range was also set empirically. When the Eps values was larger, the GEP were allocated into 2 clusters, and the clustering results were repeated. The resulting range of the Eps values variation was divided into 20 equal sections. The width of the section was equal to the step of the Eps value changing. According to the hereinbefore presented algorithm, the clustering and classification quality criteria were calculated only for cases where the number of clusters allocated on equal-power subsets was equal. This condition minimizes the reproducibility error. Tables 1 and 2 and Figures 3 and 4 present the modelling results.

Table 1

The result of the division of GEP into clusters when *MinPts* = 3

EPS,$*10^{-3}$	Clusters					
	1	2	3	4	5	6
0.66435	24	311	6	14	6	7
0.69525	24	322	6	14	6	–
0.71070	24	323	6	14	6	–
0.72615	24	329	6	14	6	–
0.74160	24	332	6	14	6	–
0.91155	24	359	6	–	–	–
0.92700	24	359	6	–	–	–

Table 2

The result of the division of GEP into clusters when *MinPts* = 4 and 5

EPS,$*10^{-3}$	*MinPts* = 4				*EPS*,$*10^{-3}$	*MinPts* = 5			
	Clusters					Clusters			
	1	2	3	4		1	2	3	4
0.64890	24	158	115	11	0.64890	23	155	115	11
0.69525	24	322	14	–	0.66435	24	308	14	–
0.71070	24	323	14	–	0.67980	24	314	14	–
0.72615	24	326	14	–	0.69525	24	321	14	–
0.74160	24	331	14	–	0.71070	24	322	14	–
–	–	–	–	–	0.72615	24	326	14	–
–	–	–	–	–	0.74160	24	331	14	–

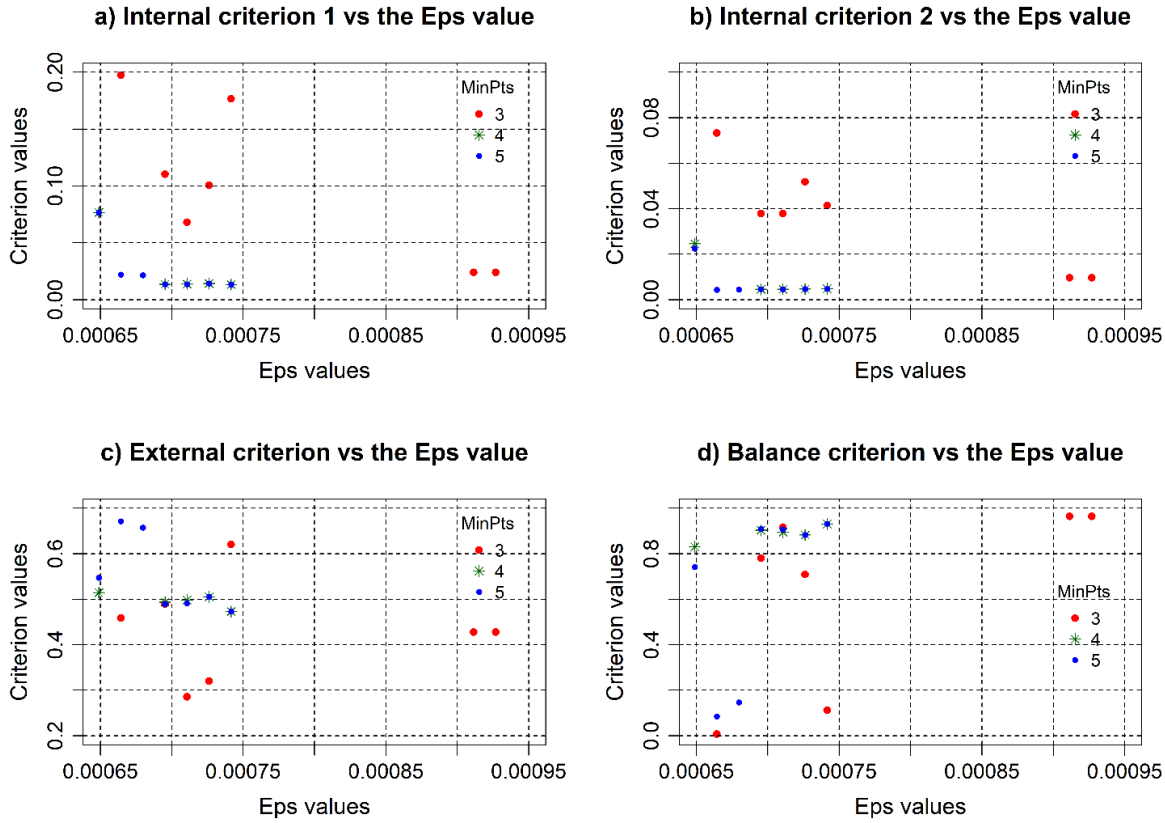

Figure 3: The simulation results regarding the criterial analysis of cluster structure using OPTICS algorithm implemented on the basis of OCIT: distribution of the internal criteria assessed on the first (a) and second (b) equivalent subsets of GEP; external (c) and hybrid balance criterion (d) when the *Eps* and *MinPts* values are varied from minimum to maximum values

The analysis of the obtained results allows concluding on the feasibility of using the proposed GEP proximity metric for the selection of mutually correlated profiles in the case of using a multicluster structure which is formed by applying the OPTICS clustering algorithm. The proposed method crate the condition to assess the algorithm suitable parameters in terms of the optimal nature of the GEP grouping into clusters on the one hand, and the minimum value of the reproducibility error on the other hand. As can be seen from Tables 1 and 2, when the *MinPts* parameter value is 3, there are seven clusters' structures. The first clustering contains six clusters, the four clustering contain five clusters, and the last two clustering contain three clusters. In the cases when *MinPts* values are 4 or 5, the first clustering contained four clusters, in other cases, three clusters were obtained in each clustering. It should be noted that the initial data contained approximately 400 gene expression profiles that were carefully selected by stepwise application of the SOTA clustering algorithm [20,21]. The accuracy of the samples classifying when the full set of gene expression profiles was used as attributes was approximately 93%.

Analysis of the results shows also that in all cases, some of the gene expression profiles are identified as noise. These genes are not contained in any cluster. The presence of "noise" genes can be explained by the fact that the density of these GEP in terms of the used proximity metric is less than the conditional boundary value assessed by the OPTICS clustering algorithm. Analysis of the charts presented in Figure 3 has also shown that the internal and external criteria do not optimal to assess the OPTICS algorithm suitable parameters because the minimum values of these metrics do not matched to the maximum values of the object classification accuracy in the corresponding clusters. The maximum value of the hybrid balance criterion, which contains as components both the internal and external criteria is achieved in the case in a three-cluster structure with the parameters of the OPTICS algorithm: *MinPts* = 3, *Eps* = 0.00091155 or *Eps* = 0.00092700 (the same results are achieved in these instances). The results of the classification of objects contained in the corresponding clusters and

presented in Figure 4, confirm the hereinbefore conclusions. As it can be seen from the charts, with these parameters of the algorithm, the classification accuracy is maximal for the first two clusters, while the second cluster contains the largest number of genes, i.e. it is the main in terms of the number of gene expression profiles. The third cluster contains only six genes. The classification results in the fourth, fifth and sixth clusters are not adequate because they are the same in all cases and slightly worse than the classification results in the first three clusters. It should be noted that the maximum values of the hybrid balance criterion that determines the quality of gene expression profiles clustering correspond to the maximum values of the samples classification accuracy that contain as the attributes the extracted gene expression profiles. This fact indicates the high efficiency of the proposed hybrid proximity metric and technique to asses the quality of GEP clustering.

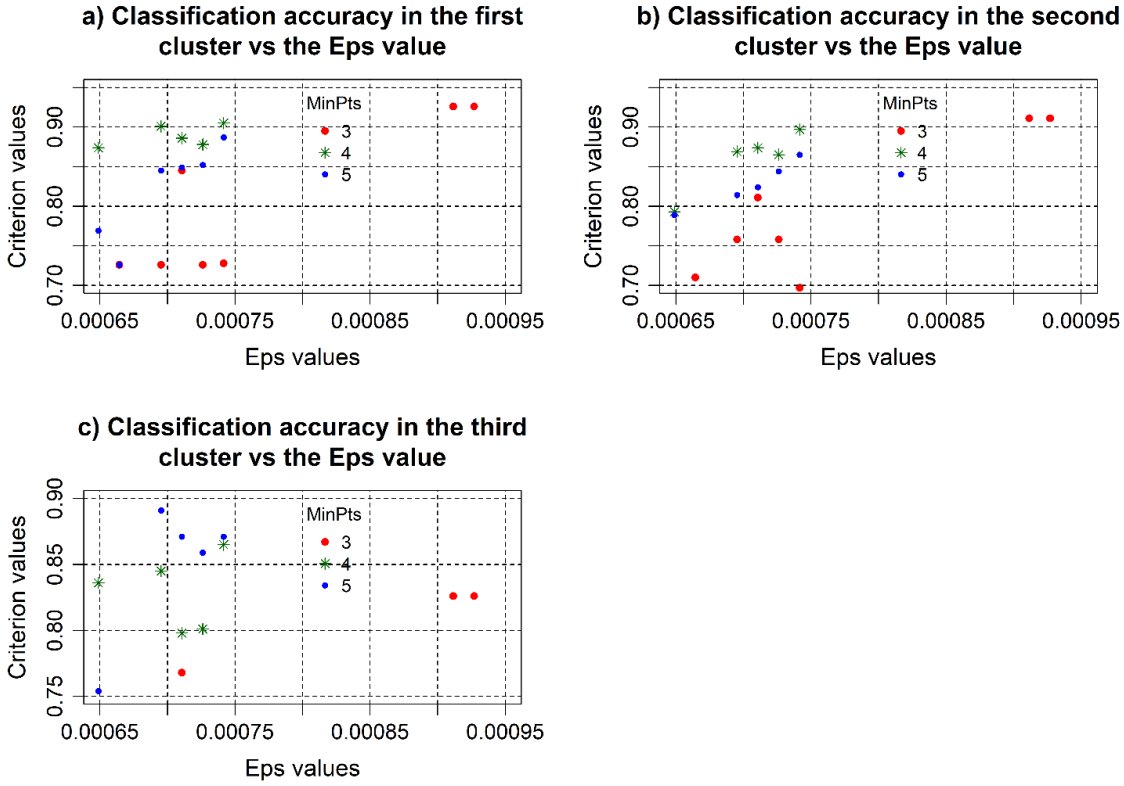

Figure 4: The results of the simulation regarding assessing the objects classification accuracy whose attributes are the gene expression profiles allocated to clusters using the OPTICS algorithm: a) the first cluster; b) the second cluster; c) the third cluster

4. Conclusions

A hybrid model of GEP clusters formation in order to extract the groups of mutual similar GEP in terms of applied proximity metrics based on the application of OPTICS clustering algorithm implemented on the basis of OCIT principles has been described in this paper. The hybrid proximity metric to access the distance between GEP has been applied during the simulation. This metric has been calculated on the basis of the joint applying the modified mutual information maximization metric (considered various methods of Shannon entropy evaluation) and Pearson's χ^2 test. The effectiveness of this hybrid proximity metric has been proved in [27]. The structural block chart of the stepwise algorithm for set the OPTICS algorithm suitable parameters in terms of a hybrid balance clustering quality criterion, which contains as components the internal and external clustering quality criteria has been presented. The high efficiency of the proposed model has been confirmed by the convergence of quality criteria for clustering gene expression profiles and the classification of objects that contain these GEP as attributes.

An analysis of the simulation results has indicated that the internal and external clustering quality criteria do not allow determining the OPTICS algorithm optimal parameters. The minimal values of these criteria do not matched to the maximum values of the object classification accuracy in the corresponding clusters. The maximal value of the hybrid balance criterion, which is formed considering both the internal and external criteria has been achieved for a three-cluster structure with the parameters of the OPTICS algorithm: *MinPts* = 3, Eps = 0.00091155 or Eps = 0.00092700 (the same results are achieved in these instances).

The analysis of the results of objects classification has confirmed the high effectiveness of the proposed technique since the classification accuracy is maximal for the first two clusters, while the second cluster contains the largest number of genes, i.e. it is the main in terms of the number of gene expression profiles. The third cluster contains only six genes. The fourth, fifth and sixth clusters contained the same number of gene expression profiles. Additionally, classification accuracy in these cases is slightly worse than the classification results in the first three clusters. It should be noted that the maximum values of the hybrid balance criterion that determines the quality of GEP clustering matched to the maximum values of the samples classification accuracy that contain as the attributes the extracted gene expression profiles. This fact indicates the high efficiency of the proposed hybrid proximity metric and model to assess the quality of GEP clustering. However, we would like to note that the proposed proximity metric is appropriate for high dimensional gene expression profiles. In the case of the other data use, it is necessary to investigate other more suitable for this type of data metrics. This is the limitation of the proposed model.

The further perspectives of the authors' research are an application of the proposed hybrid proximity metric within the framework of gene expression profiles hybrid clustering and classification techniques implemented based on other clustering and classification algorithms.

5. References

[1] M.E. Ritchie, B. Phipson, D. Wu, et al. limma powers diff. express. analysis for RNA-sequencing and microarray studies. Nucl. Acids Res., 2015, vol. 43(7), art. no. e47. doi: 10.1093/nar/gkv007

[2] R. Ihaka, R. Gentleman. R: a lang. for data analysis and graphics. J. of Comp. and Graph. Statistics, 1996, vol. 5(3), pp. 299-314. doi:10.2307/1390807

[3] S. Babichev, A. Kornelyuk, et al. Computat. analysis of microarray GEP of lung cancer. Biopolymers and Cell, 2016, vol. 32(1), pp.70–79. doi: 10.7124/bc.00090F

[4] S. Babichev, B. Durnyak, et al. Techniques of DNA microarray data pre-processing based on the complex use of Bioconductor tools and Shannon entropy. CEUR Workshop Proceedings, 2019, vol. 2353, pp. 365-377.

[5] S. Babichev, B. Durnyak, V. Senkivskyy, et al. Exploratory analysis of neuroblast. data genes expr. based on Bioconductor package tools. CEUR Workshop Proceedings, 2019, vol. 2488, pp. 268-279.

[6] C.S.Tan, W.S. Ting, M.S. Mohamad, et al. A Review of Feature Extraction Soft. for Microarray Gene Expr. Data. BioMed Res. Int., 2014, vol. 2014, art. no. 213656. doi: 10.1155/2014/213656

[7] N. Almugren, H. Alshamlan. A survey on hybrid feature selection meth. in microarray gene express. data for cancer classific.. IEEE Access, 2019., vol. 7, art. no. 8736725, pp. 78533-78548. doi: 10.1109/ACCESS.2019.2922987

[8] H. Lu, J. Chen, K. Yan, et al. A hybrid feature selection algorithm for GED classification. Neurocomp., 2017, vol. 256, pp. 56-62. doi: 10.1016/j.neucom.2016.07.080

[9] C.P. Lee, Y. Leu. A novel hybrid feature selection method for microarray data analysis. Appl. Soft Comput., 2011, vol. 11(1), pp. 208-213. doi: 10.1016/j.asoc.2009.11.010

[10] L.Y. Chuang, C.H. Yang, K.C. Wu, C.H. Yang. A hybrid feature selection method for DNA microarray data. Comput. Biol. Med., 2011, vol. 41(4), pp. 228-237. doi: 10.1016/j.compbiomed.2011.02.004

[11] M.A. Valizade Hasanloei, R. Sheikhpour, et al. A combined Fisher and Laplacian score for feature selection in QSAR based drug design using compounds with known and unknown activities. J. Comput Aided. Mol. Des., 2018, vol. 32(2), pp. 375-384. doi: 10.1007/s10822-017-0094-6

[12] J.R. Quinlan, Induction of decision trees. Mach Learn, 1986, vol. 1, pp. 81–106. doi: 10.1007/BF00116251
[13] L. Xiaowei, J. Chenglin, et al. A Fisher's Criterion-Based Linear Discriminant Analysis for Predicting the Critical Values of Coal and Gas Outbursts Using the Initial Gas Flow in a Borehole. Mathematical Problems in Engineering, 2017, vol. 2017, art. no. 7189803. Doi: 10.1155/2017/7189803
[14] A. Hyvärinen. Independent component analysis: recent advances. Phil. Trans. R. Soc., 2013, vol.371, art. no. 20110534. doi: 10.1098/rsta.2011.0534
[15] H. Alshamlan, G. Badr, et al. mRMR-ABC: A hybrid gene selection algorithm for cancer classification using microarray GEP. Biomed. Res. Intern., 2018, vol. 2015, art. no. 604910. doi: 10.1155/2015/604910
[16] P. Moradi, M. Gholampour. A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy. Applied Soft Computing, 2016, vol. 43, pp. 117-130. doi: 10.1016/j.asoc.2016.01.044
[17] E. Pashaei, M. Ozen, N. Aydin. Gene selection and classification approach for microarray data based on random forest ranking and BBHA. In Proc. IEEE-EMBS Int. Conf. Biomed. Health Inform. (BHI), 2016, pp. 308-311. doi: 10.1109/BHI.2016.7455896
[18] X. Li, M. Yin. Multiobjective binary biogeography based optimization for feature selection using gene expression data. IEEE Trans. Nanobiosci., 2013, vol. 12(4), pp. 343-353. doi: 10.1109/TNB.2013.2294716
[19] S.S. Shreem, S. Abdullah, M.Z.A. Nazri. Hybrid feature selection algorithm using symmetrical uncertainty and a harmony search algorithm. Int. J. Syst. Sci., 2016, vol. 47(6), pp. 1312-1329. doi: 10.1080/00207721.2014.924600
[20] P. Tumuluru, B. Ravi. GOA-based DBN: Grasshopper optimization algorithm-based deep belief neural networks for cancer classification. Int. J. Appl. Eng. Res., 2017, vol. 12(24), pp. 14218-14231.
[21] S. Babichev, J. Škvor. Technique of Gene Expression Profiles Extraction Based on the Complex Use of Clustering and Classification Methods. Diagnostics, 2020, vol. 10 (8), art. no. 584. doi: 10.3390/diagnostics10080584
[22] S. Babichev, V. Lytvynenko, et al. Information Technology of Gene Expression Profiles Processing for Purpose of Gene Regulatory Networks Reconstruction. (2018) Proceedings of the 2018 IEEE 2nd International Conference on Data Stream Mining and Processing, DSMP 2018, art.no. 8478452, pp. 336-341. doi: 10.1109/DSMP.2018.8478452
[23] I. Izonin, R. Tkachenko, V. Verhun, et al. An approach towards missing data management using improved grnn-sgtm ensemble method. International Journal Engineering Science and Technology, 2020, p. in press. doi: 10.1016/j.jestch.2020.10.005
[24] V. Lytvyn, T. Salo, V. Vysotska, et al. Identifying textual content based on thematic analysis of similar texts in big data. In: IEEE 2019 14th International Scientificc and Technical Conference on Computer Sciences and Information Technologies, CSIT 2019 – Proceedings, 2019, vol. 2, pp. 84-91. doi: 10.1109/STC-CSIT.2019.8929808
[25] A. Rzheuskyi, O. Kutyuk, V. Vysotska, et al. The architecture of distant competencies analyzing system for it recruitment. In: IEEE 2019 14th Int. Sc. and Techn.l Conf. on Comp. Sc. and Inf. Techn., 2019, vol. 3, pp. 254-261. doi: 10.1109/STC-CSIT.2019.8929762
[26] R. Tkachenko, I. Izonin, N. Kryvinska, et. al. An approach towards increasing prediction accuracy for the recovery of missing iot data based on the grnn-sgtm ensemble. Sensors (Switzerland), 2020, vol. 20(9), art. no. 2625. doi: 10.3390/s20092625
[27] S. Babichev, L. Yasinska-Damri, I. Liakh, B. Durnyak. Comparison Analysis of Gene Expression Profiles Proximity Metrics. Symmetry, 2021, vol. 13(10), art no 1812. doi: 10.3390/sym13101812
[28] M. Ankerst, M.M. Breunig, H.P. Kriegel, J. Sander. OPTICS: Ordering Points to Identify the Clustering Structure. SIGMOD Record (ACM Special Interest Group on Management of Data), 1999, vol. 28(2), pp. 49-60, doi: 10.1145/304181.304187
[29] H.R. Madala, A.G. Ivakhnenko. Inductive Learning Algorithms for Complex Systems Modeling. CRC Press, 1994, 365 p.

[30] S. Babichev, B. Durnyak, et al. Application of Optics Density-Based Clustering Algorithm Using Inductive Methods of Complex System Analysis. International Scientific and Technical Conference on Computer Sciences and Information Technologies, 2019, vol. 1, art. no. 8929869, pp. 169-172. doi: 10.1109/STC-CSIT.2019.8929869
[31] S. Babichev, V. Lytvynenko, M.A. Taif. Estimation of the inductive model of objective clustering stability based on k-means algorithm for different level of data noise. Radio Electronics, Comp. Science, Control, 2016, vol. 4, pp. 54-60. doi: 10.15588/1607-3274-2016-4-7
[32] J. Hou, J. Aerts, B. den Hamer, et al. Gene expression-based classification of non-small cell lung carcinomas and survival prediction. PLoS ONE 2010, vol. 5, art no e10312. doi:10.1371/journal.pone.0010312.