ISSN 1993-6842 (on-line); ISSN 0233-7657 (print)
Biopolymers and Cell. 2017. Vol. 33. N 5. P 379–392
doi: http://dx.doi.org/10.7124/bc.000961

UDC 004.048

Objective clustering inductive technology of gene expression profiles based on SOTA clustering algorithm

S. A. Babichev[1], A. Gozhyj[2], A. I. Kornelyuk[3], V. I. Lytvynenko[4]

[1] University of J. E. Purkyně in Ústí nad Labem
1, Pasteur Str, Ústí nad Labem, Czech Republic, 400 96

[2] Petro Mohyla Black Sea State University
10,68-Desantnykiv Str. Mykolayiv, 54003

[3] Institute of Molecular Biology and Genetics, NAS of Ukraine
150, Akademika Zabolotnoho Str., Kyiv, Ukraine, 03680

[4] Kherson National Technical University
24, Beryslavske sh, Kherson, Ukraine, 73008
sergii.babichev@ujep.cz; alex.gozhyj@gmail.com; kornelyuk@imbg.org.ua; immun56@gmail.com

Aim. Development of an inductive technology of objective clustering of gene expression profiles based on a self-organizing SOTA clustering algorithm. **Methods.** Inductive methods of complex system analysis were used to implement the inductive technology of objective clustering of gene expression profiles. The optimal parameters of clustering algorithm were estimated using internal clustering quality criteria, external criteria and complex balance criteria. **Results.** Here we present the architecture of the inductive technology of objective clustering based on SOTA clustering algorithm and step-by-step procedure of its implementation. Charts of the internal, external and complex balance criteria versus the algorithm parameters were obtained during simulation. This allowed us to determine the optimal parameters of the algorithm. **Conclusion.** We have shown a high efficiency of the proposed technology. In case of analysis of gene expression profiles, this approach allows to implement a step-by-step cluster-bicluster technology of data grouping at an early stage of gene regulatory network reconstruction.

Keywords: objective clustering, inductive modeling, SOTA algorithm, clustering quality criteria, gene expression profiles.

Introduction

Gene regulatory network reconstruction based on the gene expression profiles is one of the current directions of modern bioinformatics. Gene regulatory network is a set of genes, which interact with each other to control the specific cell functions. Qualitatively reconstructed gene regulatory network allows us to study the influence of the corresponding group of genes or individual genes on abilities of the biology objects. Gene expression profiles, which are obtained by DNA microarray experiments or by RNA sequences technology are the basis to reconstruct gene regulatory networks. High dimension of the features space is one of the gene expression profiles peculiarities. About tens of thousands genes are contained in gene expression profiles. It is obvious that reconstruction of the gene regulatory network based on full dataset is very difficult task because this process requests large capacity of computer resources and complicity of the obtained network complicates the results of its work interpretation. Therefore, it is necessary at the early stage of network reconstruction to group studied gene profiles according to the level of their similarity. Biclustering technology is current one for solving this problem. Implementation of this technology allows grouping objects and genes according to their mutual correlation. So, in the paper [1] authors provide a review of a large quantity of biclustering approaches existing in literature with analysis of their advantages and disadvantages. In [2] authors have proposed and implemented convex biclustering method using gene expression profiles of the lung cancer patient. The authors have shown the efficiency of the proposed method during simulation process. However, it should be noted that one of the significant problems of this technology qualitative implementation is selection of the biclustering level during objects and genes grouping. Qualitative validation of the obtained model is another task, which has no solution currently. High dimension of features space promotes to large quantity of the obtained biclusters. Limitation of their quantity by removing of small biclusters leads to the loss of some useful information. To solve this problem we propose cluster-bicluster technology, the implementation of which involves two stages: clustering of gene expression profiles at the first step and biclustering of the obtained clusters at the second step. To decrease the reproducibility error of clustering process the data clustering is performed within the framework of the objective clustering inductive technology the implementation of which involves the use of external information to correct verification of the obtained model and the use of internal clustering quality criteria, external criterion and complex balance clustering quality criterion. High objectivity is achieved by using two equal power subsets during clustering process. The term equal power means that these subsets contain the same quantity of pairwise similar objects.

The idea and conceptual basis of the objective clustering methods have been proposed by Madala and Ivakhnenko [3] and further developed in [4, 5]. The authors' research is based on the inductive method of complex systems self-organization models on the basis of Group Method of Data Handling (GMDH), the idea and main principles of which are presented in [6, 7]. Implementation of the proposed method involves enumeration of the models from simple to complex ones and selection of the best model based on qualitative criteria of the studied process estimation. However, it should be noted that the authors' research is focused mainly on low dimensional data processing. The [8] presents objective clustering inductive technology of high dimensional data. The authors have developed an architecture of this technology and step-by-step procedure of its implementation. Practical implementation of objective clustering inductive technology based on agglomerative hierarchical clustering algorithm is presented in [9]. However, in spite of the progress achieved there are some unsolved issues in this field. They are connected with practical implementation of the objective clustering inductive technology based on self-organizing hierarchical clustering algorithms and verification of the obtained models using different high dimensional data.

The unsolved parts of the general problem are:

- Absence of complex criterial analysis of clustering results which are obtained concurrently on two equal power subsets based on complex balance clustering quality criterion that takes into account: character of objects distribution relative to mass center of clusters where these objects are and character of cluster's mass centers distribution in features space; difference between clustering results which are obtained using two equal power subsets.
- Practical implementation of objective clustering inductive technology based on existing clustering algorithm using gene expression profiles in order to select the best clustering algorithms for studied data and to determine the optimal parameters of this algorithm operation and its practical implementation within the framework of hybrid models of gene expression profiles grouping.

The Aim of the paper is the development of objective clustering inductive technology of gene expression profiles based on self-organizing SOTA clustering algorithm.

Materials and methods

Three principles of inductive methods of complex systems analysis are the basis of objective clustering inductive technology:

- the principle of heuristic self-organization or enumeration of clustering models in order to select from them the best ones based on extremum values of internal and external clustering quality criteria;
- the principle of external edition or necessity of the use of several equal power subsets which contain the same quantity of pairwise similar objects to perform objective verification of the obtained model;
- the principle of inconclusiveness of solutions or generation of the set of intermediate results in order to select from them the best variants based on extremum value of complex balance criterion.

The architecture of objective clustering inductive technology [8] is shown in Fig. 1. The initial dataset is presented as a matrix:

$$A = \left\{x_{ij}\right\},\ i = 1...n,\ j = 1...m,$$

where n – is the quantity of the studied objects, m – is the quantity of the objects features. The aim of the clustering is partition of the objects into non-empty subsets of pairwise non-intersecting clusters in accordance with the clustering quality criteria taking into account the properties of the studied objects:

$$\begin{gathered} K = \left\{K_1, K_2, ..., K_n\right\},\ 1 \le k \le n \\ K_1 \cup K_2 \cup ... \cup K_k = A,\ K_i \cap K_j = \varnothing,\ i \ne j,\ i, j = 1, 2, ..., k \end{gathered} \tag{1}$$

where k – is the clusters quantity. Objective clustering inductive technology is based on the inductive methods of complex systems analysis, which involves sequential enumeration of clustering in order to select from them the best variants. Let W – is the set of available clustering for equal power datasets A and B. Clustering is optimal if the following condition is performed:

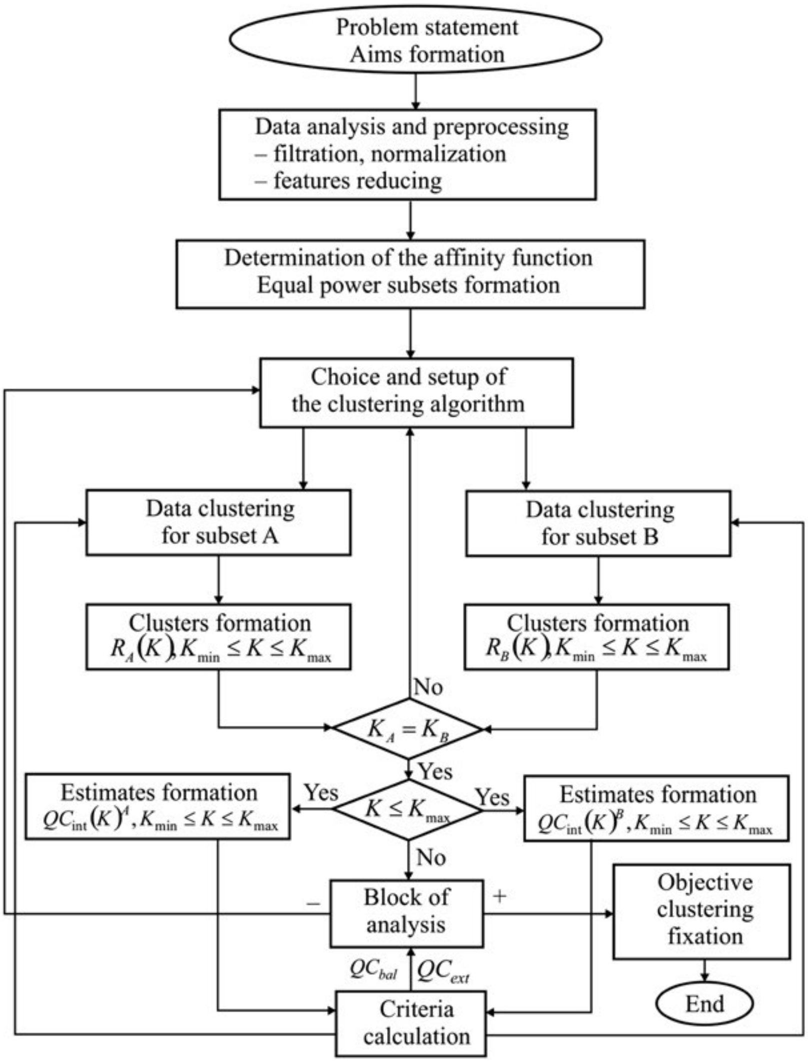

Fig. 1. Architecture of objective clustering inductive technology

$$K_{opt} = \underset{K \subseteq W}{\operatorname{argmin}}\, QC(K) \quad \text{or} \quad K_{opt} = \underset{K \subseteq W}{\operatorname{argmax}}\, QC(K) \tag{2}$$

where $QC(K)$ – is the clustering quality criterion for K clustering.

Clustering $K_{opt} \subseteq W$ is the objective if difference between distribution of objects and clusters in different clustering for equal power subsets A and B is minimal:

$$QC(K_{obj}) = \underset{K_{opt} \subseteq W}{\operatorname{argmin}}\, (QC(K_{opt})^A \ ?\ (QC(K_{opt})^B) \tag{3}$$

Implementation of objective clustering inductive technology involves the following steps:

1. Studied data analysis and preprocessing. Formation of clustering aims.

2. Determination of affinity function (level of similarity) between objects, clusters and between objects and clusters. Division of the initial dataset into two equal power subsets using chosen affinity function.

3. Selection of clustering algorithm. Setup of its initial parameters, intervals and steps of these parameters changing during the algorithm operation.

4. Data clustering on the equal power subsets A and B concurrently within the given range of the algorithm's parameters variation. Clusters formation at each stage of the clustering process.

5. Internal, external and complex balance clustering quality criteria calculation at each stage of the clustering algorithm operation.

6. Analysis of the obtained results. If clusters quantity differs or if the extremum values of clustering quality criteria are more than admissible values, choice and setup another clustering algorithm for the studied data. Otherwise, fixation of objective clustering corresponds to the extremum value of the complex balance criterion.

Comparison analysis of different clustering quality criteria within the framework of objective clustering inductive technology is carried out in [10]. Analysis of the obtained results allows us to determine the complex multiplicative criterion based on Calinski-Harabasz [11] and WB-index [12]. This criterion was used as an internal clustering quality criterion:

$$QC_{int} = \frac{K\,(K-1)\;QCW^2}{(N-K)\;QCB^2} \tag{4}$$

where K and N – are the quantity of the clusters and studied objects respectively; QCW and QCB – are the components which allow us to estimate quantitative of the objects character distribution within the clusters and character of the clusters distribution in features space. The first component is calculated as an average distance from objects to mass centers in clusters, where these objects are:

$$QCW = \frac{1}{N} \sum_{S=1}^{K} \sum_{i=1}^{N_S} d(x_i^S, C_S) \tag{5}$$

The second component is calculated as an average distance between clusters mass centers:

$$QCB = \frac{2}{K\,(K-1)} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} d(C_i, C_j) \tag{6}$$

where N_S – is the quantity of objects in cluster S; x_i^S – is the i-th object in S cluster; C_i, C_j and C_S – are mass centers of the clusters i, j and S respectively; $d()$ – is the similarity metric used to estimate proximity level of the studied vectors. Correlation distance was used as a similarity metric in case of high dimensional gene expression profiles analysis:

$$d(X_s, X_p) = (1 - r) = 1 - \frac{\sum_{i=1}^{m} ((x_{si} - \overline{x}_s) \cdot (x_{pi} - \overline{x}_p))}{\sqrt{\sum_{i=1}^{m} (x_{si} - \overline{x}_s)^2 \cdot \sum_{i=1}^{m} (x_{pi} - \overline{x}_p)^2}} \quad (7)$$

where m – is the features quantity of the studied vector; $\overline{x}_s$ and $\overline{x}_p$ – are the average values of the vectors s and p respectively. In case of low dimensional data, correlation distance is not effective and Euclidean distance was used as a similarity metric:

$$d(X_s, X_p) = \sqrt{\sum_{i=1}^{m} (x_{si} - x_{pi})^2} \quad (8)$$

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

$$QC_{ext}(A, B) = \frac{QC_{\text{int}}(A) - QC_{\text{int}}(B)}{QC_{\text{int}}(A) + QC_{\text{int}}(B)} \quad (9)$$

It is obvious that objective clustering corresponds to the minimum values of internal and external clustering quality criteria. However, it is possible that the extremums of these criteria correspond to different clustering. Thus, it is necessary to determine complex balance clustering quality criterion which takes into account both the character of the objects and the clusters distribution in various clustering and the difference between clustering results, which are obtained on the equal power subsets A and B. To calculate complex balance clustering quality criterion Harrington desirability function [13] was used. Implementation of this function involves transformation of scales of internal and external criteria into reaction scale the values of which are changed linearly within the range from –2 to 5:

$$Y = a - b \cdot Q \quad (10)$$

The coefficients a and b are determined empirically. Then the private desirabilities of the appropriate criteria are calculated by the formula:

$$d = \exp(-\exp(-Y)) \quad (11)$$

General desirability value is calculated as geometric average of private desirabilities:

$$D = \sqrt[n]{\prod_{i=1}^{n} d_i} \quad (12)$$

The largest value of the general Harrington desirability function corresponds to the best parameters of clustering algorithm operation.

SOTA clustering algorithm (Self-Organizing Tree Algorithm) [14] which is a type of self-organizing neural networks based on Kohonen maps and Fritzke algorithm of spatial cell structure growing [15] was used within the framework of objective clustering inductive technology. 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. Effectiveness of SOTA clustering algorithm operation is determined by the two parameters: weight coefficient of the sister's cell (scell) and maximum divergence coefficient (E). Weight coefficients of the parent's and winner's cells are calculated automatically. To calculate the optimal parameters of algorithm operation we propose to use the objective clustering inductive technology.

Block-scheme of the inductive algorithm of objective clustering based on SOTA clustering algorithm is shown in Fig. 2. Implementation of this model involves the following steps:

Step 1. Formation of the initial set Ω of the objects. Data preprocessing (filtration and normalization). Presentation of data as a matrix $n \times m$, where n – is the quantity of the studied objects or the quantity of the rows and m – is the quantity of the features characterizing objects or the quantity of the columns.

Step 2. Determination of the similar metric depending on the type of the studied vectors by formulas (7) or (8). Division of the initial dataset into two equal power subsets.

Step 3. Setup of SOTA clustering algorithm. Setting of E and b parameters and initial value of scell weight parameter, interval and step of its change. The pcell and wcell parameters are changed automatically by formulas: $pcell = scell \cdot 5$; $wcell = pcell \cdot 2$.

Step 4. Data clustering on the equal power subsets A and B concurrently. Clusters formation and internal clustering quality criteria calculation by formulas (4)–(6) within a range of the algorithm's parameter interval change.

Step 5. External and balance criteria calculation by formulas (9)–(12).

Step 6. Fixation of the optimal scell parameter corresponding to the maximum value of the balance criterion.

Step 7. Setting of the initial value of the maximum divergence parameter (E), interval and step of its change. Repetition of the steps 4–5 of this algorithm. Fixation of the optimal E parameter.

Step 8. Data clustering by SOTA clustering algorithm using the optimal parameters of the algorithm operation.

Results and Discussion

Implementation of the proposed technology was performed using three well known databases: gene expression profiles of the lung cancer patients, which were obtained by DNA microchip experiments [16], Seeds data [17] which contained the examined group comprised kernels belonging to three different varieties of wheat: Kama, Rosa and Canadian, each of these groups contains 70 elements randomly selected for the experiment, and Fisher's Iris [18] which was used as the third dataset. This dataset consists of three species of Iris: setosa, virginica and versicolor. Each of the groups

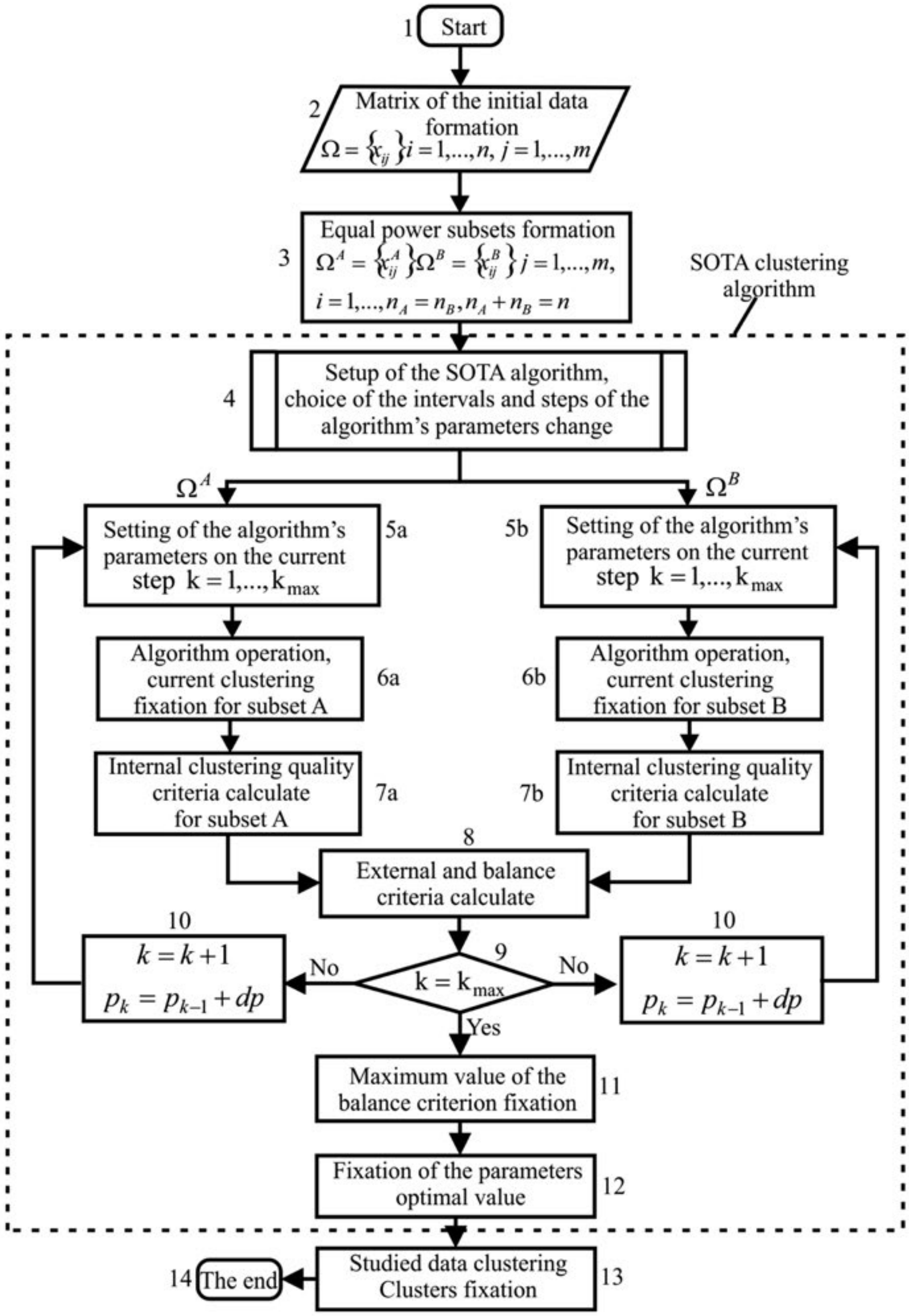
1 Start
2 Matrix of the initial data formation
$\Omega = \{x_{ij}\} i = 1,...,n, j = 1,...,m$
3 Equal power subsets formation
$\Omega^A = \{x_{ij}^A\} \Omega^B = \{x_{ij}^B\} j = 1,...,m,$
$i = 1,...,n_A = n_B, n_A + n_B = n$
SOTA clustering algorithm
4 Setup of the SOTA algorithm, choice of the intervals and steps of the algorithm's parameters change
Ω^A
Ω^B
5a Setting of the algorithm's parameters on the current step $k = 1,..., k_{max}$
5b Setting of the algorithm's parameters on the current step $k = 1,..., k_{max}$
6a Algorithm operation, current clustering fixation for subset A
6b Algorithm operation, current clustering fixation for subset B
7a Internal clustering quality criteria calculate for subset A
7b Internal clustering quality criteria calculate for subset B
8 External and balance criteria calculate
10 $k = k + 1$ $p_k = p_{k-1} + dp$
9 $k = k_{max}$
No
No
Yes
10 $k = k + 1$ $p_k = p_{k-1} + dp$
11 Maximum value of the balance criterion fixation
12 Fixation of the parameters optimal value
13 Studied data clustering Clusters fixation
14 The end

Fig. 2. Block-scheme of the inductive algorithm of the objective clustering based on SOTA clustering algorithm

contains 50 vectors. Correlation metric was used to estimate the proximity level of the gene expression profiles. To determine the distance between the studied objects in case of Seed and Iris data we used Euclidean metric since the studied vectors in these cases have low dimension of features space. The length of the vectors in case of gene expression profiles was 96 (it equals the studied objects quantity). The steps of these data preprocessing in order to increase the informativity of gene expression profiles are described in [19]. The aim of the clustering in this case is grouping of gene expression profiles to decrease the dimension of feature space. Vectors of Seeds and Iris data consist of 7 and 4 features respectively. The interval of the scell parameter in case of gene expression profiles dataset was changed within a range from 0,001 to 0,2 with the step 0,001. The results of internal criteria for the equal power subsets A and B, external criterion and complex balance clustering quality criterion versus weight parameter of the sister's cell value are presented in Fig. 3. Maximum divergence val-

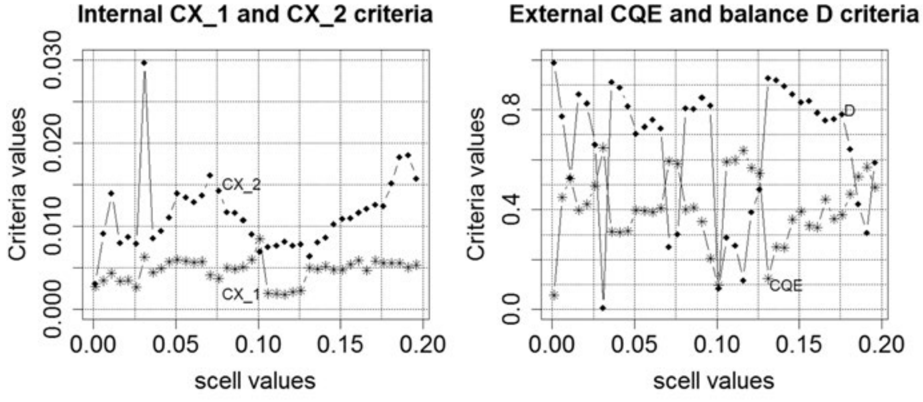

Fig. 3. Charts of internal, external and balance criteria versus weight coefficient of the sister's cell values for lung cancer data

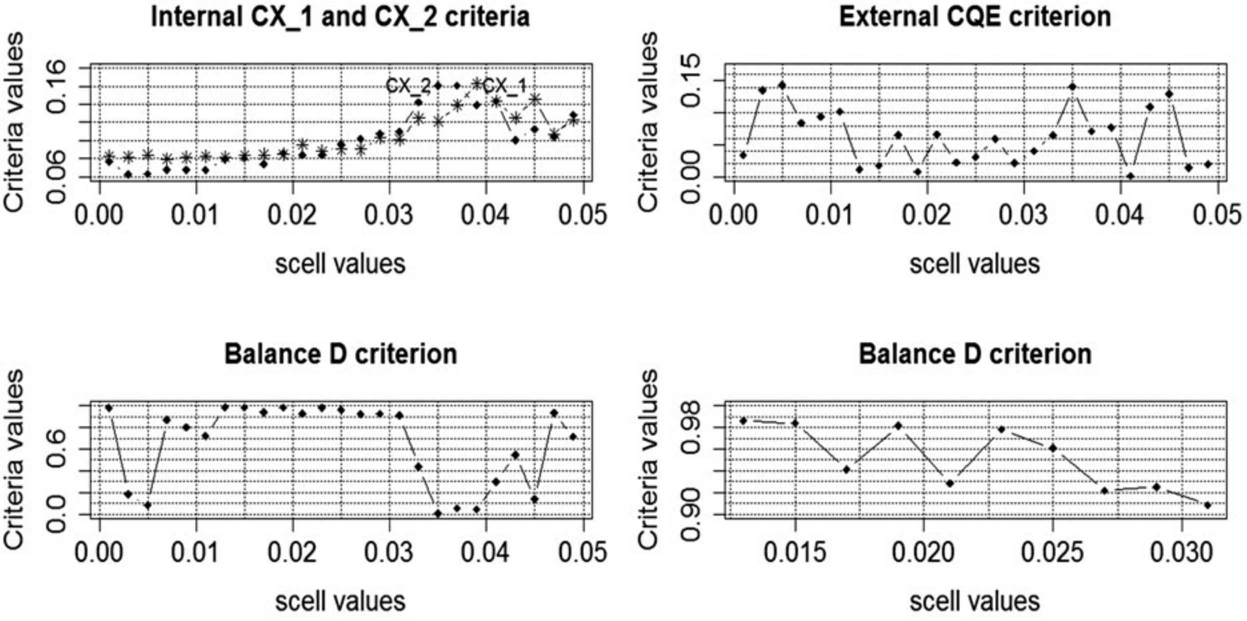

Fig. 4. Charts of internal, external and balance criteria versus weight coefficient of the sister's cell values for Seeds data

ue in this case E = 0,001 was taken. As it can be seen from Fig. 3, the internal clustering quality criteria CX_1 and CX_2, which have been calculated on equal power subsets A and B do not allow us to determine the optimal scell value corresponding the objective clustering of the studied data. External clustering quality criterion CQE has several local minimums corresponding to the successful grouping of the studied vectors. However, the analysis of general Harrington desirability values, which takes into account both internal and external criteria, allows us to conclude that the best clustering corresponds to the scell = 0,001. In this case 6659 profiles were divided into two clusters. The first cluster contained 4276 profiles and the second – 2383 ones. Variation of maximum divergence value in the range from 0,001 to 1 has not changed the obtained results. Fig. 4 presents the same charts for Seeds data.

The scell value in this case was changed within the range from 0,001 to 0,05 with the step 0,002. The analysis of the charts shows that the largest value of balance criterion is achieved for scell = 0,013. This value corresponds also to the least value of external clustering quality criterion and the least difference of clustering results for the equal power subsets A and B (minimum difference between internal clustering quality criteria values). Fig. 5 presents the charts of internal criteria, external criterion and complex balance criterion versus maximum divergence value, which was changed within the range from 0,05 to 1 with step 0,05.

Analysis of the charts shows that the most optimal and the most objective clustering corresponds to 0,7 maximum divergence value. During clustering with the use of full dataset the obtained results have shown that in case of scell = 0,013 and E = 0,7 values using the

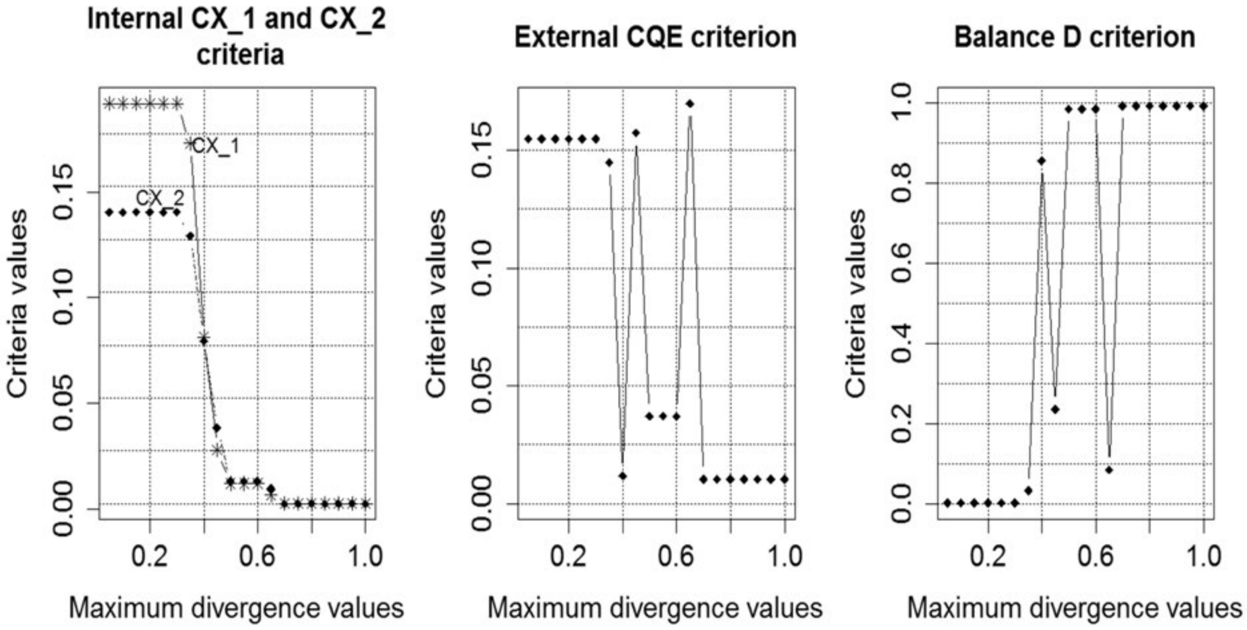

Fig. 5. Charts of internal, external and balance criteria versus maximum divergence values for Seeds data

percent of correctly distributed objects equals 85,5 %. It should be noted that in case of a small change of these parameters the percent of correctly distributed objects is decreased. The same results for Iris data are presented in Fig. 6 and 7. Analysis of the charts allows us to conclude that the best clustering result is achieved in case of scell = 0,029 and E = 0,5 values use. The studied Iris data were divided into 5 clusters. The first cluster contained 50 setosa vectors. In the second clusters there were 27 virginica vectors. The third and the

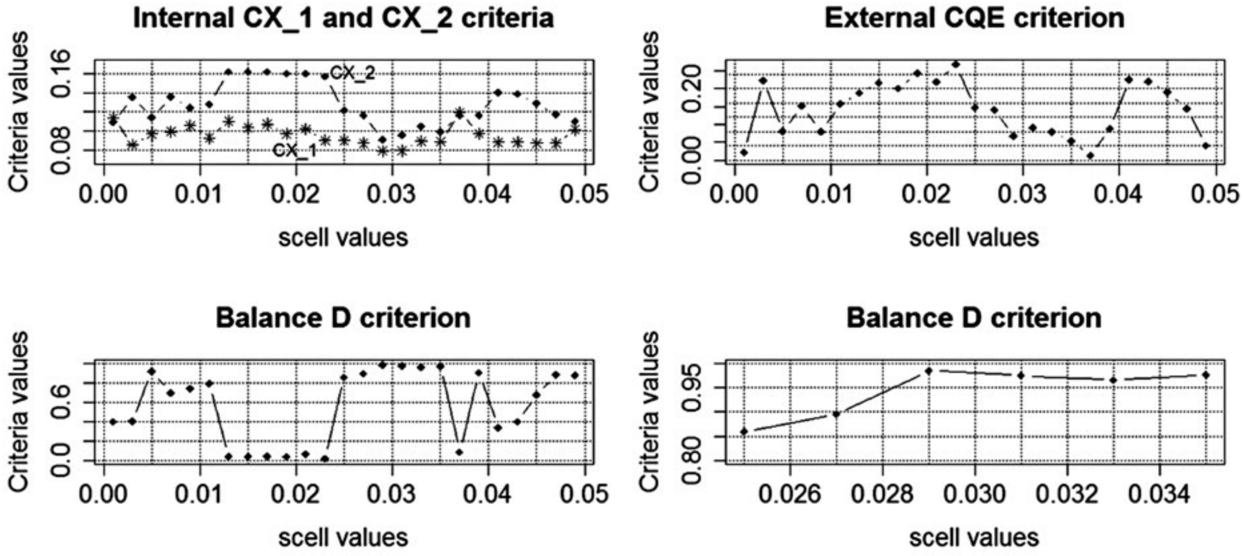

Fig. 6. Charts of internal, external and balance criteria versus weight coefficient of the sister's cell value for Iris data

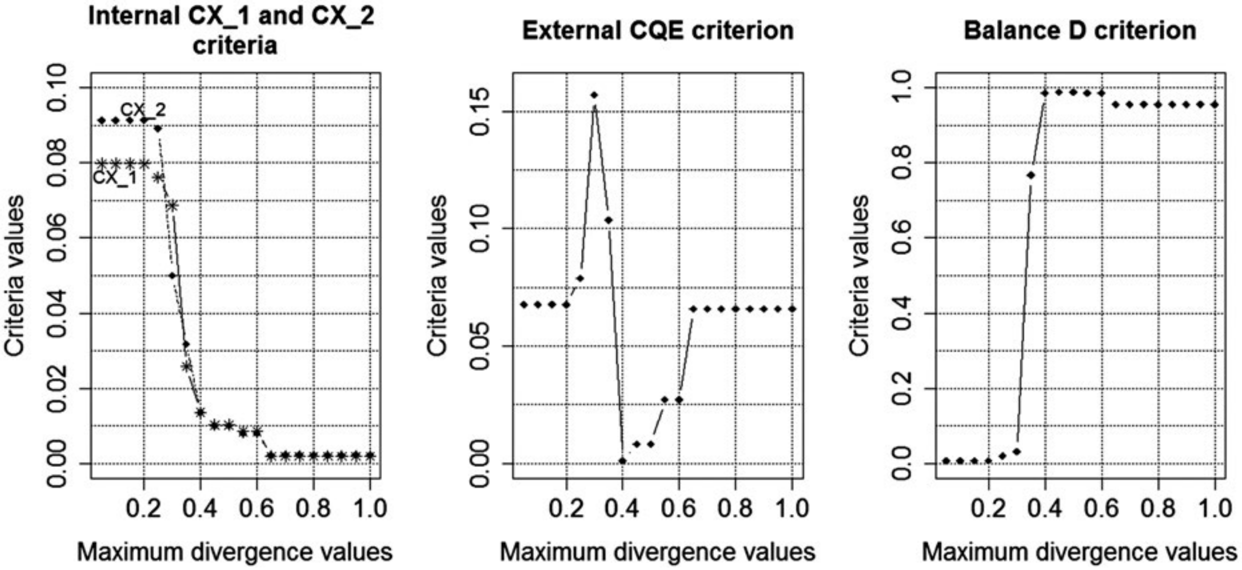

Fig. 7. Charts of internal, external and balance criteria versus maximum divergence value for Iris data

fourth clusters contained 20 and 21 versicolor vectors. In the fifth cluster, there were both virginica and versicolor vectors. It should be noted that virginica and versicolor data have some intersection a priory.

As a conclusion, we would like to say that the obtained results for Seeds and Iris data are not perfect. Self-organizing SOTA clustering algorithm is focused mainly on high dimensional gene expression profiles. Better results for Seeds and Iris data can be obtained using other clustering algorithms. However, the effectiveness of the objective clustering inductive technology based on SOTA clustering algorithm was shown during the simulation process. Implementation of this technology allows us to select objectively the optimal parameters of SOTA algorithm operation, which corresponds to maximum value of general Harrington desirability index.

Conclusions

The problem of gene expression profiles grouping at the early stage of gene regulatory network reconstruction is one of the current problems of the modern bioinformatics. Qualitatively performed profiles grouping determines high quality of gene regulatory network implementation. The paper presents the inductive technology of complex high dimensional data grouping, high objectivity of which is determined by the use of equal power subsets during clustering algorithm operation. Implementation of the proposed technology involves estimation of clustering results for different clustering within a given range of clustering algorithm parameters variation using internal and external clustering quality criteria. The final decision about the character of the studied vectors grouping is taken basing on complex balance criterion, which takes into account both character of objects and clusters distribution in various clustering and difference of clustering results on two equal power subsets. Harrington desirability function was used to calculate the complex balance criterion. Simulation of clustering process was carried out based on self-organizing SOTA clustering algorithm using three well know databases: gene expression profiles of lung cancer patient, Seeds dataset and Fisher's Iris dataset. Results of the simulation have shown high effectiveness of the proposed technology. The use of objective clustering inductive technology has allowed us to determine the optimal parameters of SOTA clustering algorithm operation, which correspond to high objectivity of the studied data grouping. During simulation process in case of lung cancer gene expression profiles maximum value of general Harrington desirability index corresponded to weight coefficient of the sister's cell 0,001. Weight coefficients of the parents (mother) cell and winner's cell were 0,005 and 0,01 respectively. Maximum divergence value was taken 0,001. 6659 gene expression profiles were divided into two clusters. 4276 profiles were in the first cluster and 2383 profiles were in the second one. It should be noted that the variation of maximum divergence value within the range from 0,001 to 1 has not changed the character of objects partition. Three clusters were obtained in case of Seeds data processing. Weight coefficients of the sister's cell, parent's cell and winner's cell were determined as 0,013, 0,052 and 0,104 respectively. Maximum divergence value was changed within the range from 0,05 to 1 with step 0,05. Maximum of Harrington desirability function corresponds to maximum divergence value E = 0,7. The percent of correctly distributed objects in this case was 85,5 %. Small change of the determined parameters decreased the percentage of correctly distributed objects. Thus, it can be concluded that the obtained combination of the parameters is optimal in terms of clustering objectivity. Interesting results were obtained in case of Fisher's Iris data use. The studied data were divided into five clusters. Fifty setosa objects were in one cluster. Virginica and versicolor objects were divided into four clusters. In the second cluster there were 27 virginica data of 50. The third and the fourth clusters contained only 20 and 21 versicolor vectors. The fifth cluster contained both virginica and versicolor vectors. It is enough logically because the virginica and the versicolor data have some intersection a priory. The optimal parameters in case of Iris data using were the following: weight coefficients of the sister's cell, parent's cell and winner's cell were 0,029, 0,116 and 0,232 respectively. Maximum divergence value was taken as 0,5. Similarly to Seeds data the small change of the determined parameters made the obtained clustering results worse. As the next step of our research we plan to create hybrid technology of gene expression profiles grouping based on complex use of objective clustering inductive technology at the first step of the data processing and biclustering technology on the obtained clusters at the final stage of data grouping.

REFERENCES

1. *Pontes B, Giráldez R, Aguilar-Ruiz JS.* Biclustering on expression data: A review. *Journal of Biomedical Informatics.* 2015; **57**: 163–18.
2. *Chi EC, Allen GI, Baraniuk RG.* Convex Biclustering. *Biometrics.* 2017; **73**: 10–10.
3. *Madala HR, Ivakhnenko AG.* Inductive Learning Algorithms for Complex Systems Modeling. *CRC Press, 1994.* 365 p.
4. *Osypenko VV.* Two approaches to solving the problem of clustering in the broad sense from the standpoint of inductive modeling. *Power and Automation.* 2014; **1**: 83–15. [In Ukraine].
5. *Sarycheva LV.* Objective cluster analysis of the data on the basis of the Group Method of Data Handling. *Problem of Management and Informatics.* 2008; **2**: 86–19. [In Russian].
6. *Ivakhnenko AG.* Inductive method for self-organizing of complex systems models. *Kiev: Scientific Thought.* 1982. 296 p. [In Russian].
7. *Ivakhnenko AG.* Objective clustering based on the theory of self-organizing models. *Automatics.* 1987; **5**: 6–10. [In Russian].
8. *Babichev S, Lytvynenko V, Korobchynskyi M, Osypenko V.* Objective clustering inductive technology of gene expression profiles features. Communications in Computer and Information Science. *Proceeding of the 13th International Conference Beyond Databases, Architectures and Structures (BDAS 2017), Ustron, Poland.* 2017; 359–14.
9. *Babichev S, Taif MA, Lytvynenko V.* Inductive model of data clustering based on the agglomerative hierarchical algorithm. *Proceeding of the 2016 IEEE First International Conference on Data Stream Mining and Processing (DSMP 2016), Lviv.* 2016; 19–4.
10. *Babichev S, Taif MA, Lytvynenko V, Osypenko V.* Criterial analysis of the gene expression sequences to create the objective clustering inductive technology. *Proceeding of the 2017 IEEE 37th International Conference on Electronics and Nanotechnology (ELNANO 2017), Kiev, Ukraine.* 2017; 244–5.
11. *Calinski T, Harabasz J.* A dendrite method for cluster analysis. *Communication in statistics.* 1974; **3**: 1–27.
12. *Zhao Q, Xu M, Fränti P.* Sum-of-Squares Based Cluster Validity Index and Significance Analysis. *Proceeding of International Conference on Adaptive and Natural Computing Algorithms.* 2009; 313–10.
13. *Harrington J.* The desirability function. *Industrial Quality Control.* 1965; **21**(10): 494–5.
14. *Dorazo J, Carazo JM.* Phylogenetic reconstruction using an unsupervised growing neural network that adopts the topology of a phylogenetic tree. *Journal of Molecular Evolution.* 1997; **44**(2): 226–34.
15. *Fritzke B.* Growing Cell Structures A Self-Organizing Network for Unsupervised and Supervised Learning. *Neural Networks.* 1994; **7**(9): 1441–20.

16. *Beer DG, Kardia SL, and all.* Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine.* 2002; **8**(8): 816–9.
17. *Charytanowicz M, Niewczas J, Kulczycki P, Kowalski PA, Lukasik S, Zak S.* A Complete Gradient Clustering Algorithm for Features Analysis of X-ray Images. *Information Technologies in Biomedicine. Springer-Verlag, Berlin-Heidelberg*. 2002; 15–10.
18. *Fisher RA.* The use of multiple measurements in taxonomic problems. *Annals of Eugenics.* 1936; **7**(2): 179–10.
19. *Babichev SA, Kornelyuk AI, Lytvynenko VI, Osypenko VV.* Computational analysis of microarray gene expression profiles of lung cancer. *Biopolymers and Cell.* 2016; **32**(1): 70–10.

Індуктивна технологія об'єктивної кластеризації профілів експресій генів на основі алгоритму кластеризації SOTA

С. А. Бабічев, О. Гожий, О. І. Корнелюк,
В. І. Литвиненко

Мета. Розробка індуктивної технології об›єктивної кластеризації профілів експресій генів на основі самоорганізуючого алгоритму кластеризації SOTA. **Методи.** Індуктивні методи аналізу складних систем було використано у якості базової основи при створенні індуктивної технології об›єктивної кластеризації профілів експресій генів. Оптимальні параметри роботи алгоритму кластеризації визначалися на основі комплексного використання внутрішніх та зовнішніх критеріїв якості кластеризації та комплексного критерію балансу. **Результати.** У статті представлено архітектуру індуктивної технології об›єктивної кластеризації на основі алгоритму кластеризації СОТА та покрокова процедура її реалізації. У процесі моделювання було отримано графікі залежності внутрішніх, зовнішніх та комплексного критерію балансу від параметрів роботи алгоритму кластеризації, аналіз яких дозволяє визначити оптимальні параметри роботи алгоритму кластеризації. **Висновки.** Отримані результати моделювання показали високу ефективність запропонованої технології. У випадку обробки профілів експресій генів дана технологія створює умови для реалізації покрокової кластер-бікластер технології групування даних на ранньому етапі реконструкції генної регуляторної мережі.

Ключові слова: об›єктивна кластеризація, індуктивне моделювання, алгоритм кластеризації SOTA, критерії якості кластеризації, профілі експресій генів.

Индуктивная технология объективной кластеризации профилей экспрессий генов на основе алгоритма кластеризации SOTA

С. А. Бабичев, А. Гожий, А. И. Корнелюк,
В. И. Литвиненко

Цель. Разработка индуктивной технологии объективной кластеризации профилей экспрессий генов на основе самоорганизующегося алгоритма кластеризации SOTA. **Методы.** Индуктивные методы анализа сложных систем были использованы в качестве базовой основы при создании индуктивной технологии объективной кластеризации профилей экспрессии генов. Оптимальные параметры работы алгоритма кластеризации определялись на основе комплексного использования внутренних и внешних критериев качества кластеризации и комплексного критерия баланса. **Результаты.** В статье представлена архитектура индуктивной технологии объективной кластеризации на основе алгоритма кластеризации СОТА и пошаговая процедура ее реализации. В процессе моделирования были получены графики зависимости внутренних, внешних и комплексного критерия баланса от параметров работы алгоритма кластеризации, анализ которых позволяет определить оптимальные параметры работы алгоритма кластеризации. **Выводы.** Полученные результаты моделирования показали высокую эффективность предложенной технологии. В случае обработки профилей экспрессии генов данная технология создает условия для реализации пошаговой кластер-бикластер технологии группировки данных на раннем этапе реконструкции генной регуляторной сети.

Ключевые слова: объективная кластеризация, индуктивное моделирование, алгоритм кластеризации SOTA, критерии качества кластеризации, профили экспрессий генов.

Received 12.06.2017