UDK 004.048

S.A. Babichev

OBJECTIVE CLUSTERING INDUCTIVE TECHNOLOGY OF GENE EXPRESSION SEQUENSES

Annotation. *The paper presents the objective clustering technology of gene expression sequences based on the use of inductive methods of complex systems analysis. The clustering objectivity is achieved by concurrent data clustering with the use of two equal power subsets, which contain the same quantity of pairwise similar objects. Clustering quality of data is estimated by using of the internal and the external criteria, which take into account both the character of objects distribution inside individual clusters and the character of clusters' mass centers distribution in the features space.*

Keywords: *objective clustering, gene expression, inductive modeling, internal and external criteria.*

Introduction

Gene regulatory networks creation by DNA microchip data analysis is one of the current problem of modern bioinformatics. Solution of this problem involves early grouping of gene expression sequences according to their functional possibilities, which define a variety of processes occurring in biological objects. Initial data that are obtained during the use of microchip experiments or RNA sequencing technology include expression of tens of thousands genes. Gene regulatory network creation based on full gene expression sequence is very complicated and ineffective because:

- it requires large computer resources;
- it needs high expenses of time to process information;
- complexity of the resulting network complicates the interpretation of the obtained results.

Thereby, it is necessary firstly to divide the gene expression sequences into subsets, each of which includes a group of genes that perform similar functions in studied biological object. Saving information about individual genes is a basic condition in this case. This fact does not allow us to use the factor or the component analysis for this procedure implementation of which involves getting new features that are linear combination of the initial features of the studied data. Nowadays there are two technologies to solve this problem: clustering and biclustering. Grouping of features for the whole set of studied

objects in accordance with a priori chosen proximity metric of corresponding sequences is carried out in case of cluster analysis. Each of the clusters includes a group of features for studied objects. At the same time ideally obtained clusters include different features. Biclusters contain a certain quantity of mutually correlated objects and features. Herewith it is possible that different biclusters contain the same objects and features. One of the main problems of the existing models of the complex objects grouping is the reproducibility error, in other words, satisfactory clustering results obtained on one dataset do not repeat while using another similar dataset. Solution of this problem can be achieved by qualitative selection of gene expression sequences affinity function, quality criteria to estimate data grouping and by developing validation methods of the obtained models with the use of external data, which are not used during creation of an appropriate model of data grouping. This paper presents the objective clustering inductive technology of gene expression sequences using inductive methods of complex systems analysis, which are a logic continuation of the group method of data handling (GMDH) [1-3]. Implementation of this technology allows us to decrease the reproducibility error and, consequently, to increase the gene expression sequences clustering objectivity.

Analysis of recent research and publications. The foundations of inductive self-organizing method of complex systems models on the basis of the Group Method of Data Handling (GMDH) are presented in the works [2-5]. Further development of this theory is reflected in [6-8]. The conception of the objective cluster analysis is presented in [7] and was further developed in [8]. However, it should be noted that investigations of the authors are focused mainly on the low-dimensional data. Herewith, the optimal clustering model is determined during enumeration of different combinations of features of the objects that in case of high-dimensional data is inefficient. The objective clustering algorithm which involves the comparison of clustering results using two equal power subsets for different clustering is proposed in [7,8]. Modification of this algorithm has been implemented to solve various practical problems nowadays. However, it should be noted that in spite of the achieved advantages in this subject area there are several unsolved problems.

Unsolved part of the general problem of high dimensional data clustering is the absence of effective methods and algorithms to group the features sequences based on the use of complex criteria to estimate the objects grouping quality within the framework of the inductive methods of complex systems analysis.

The aim of the paper is the development of the objective clustering inductive technology of gene expression sequences of DNA microarray data.

Presentation of the basis material. Let the initial dataset be presented as a matrix: $A = \{x_{ij}\}, i = 1,\ldots,n; j = 1,\ldots m$, where n – is the quantity of rows or studied objects and m – is the quantity of features that characterize the studied objects. The problem of clustering is the differentiation of objects or features into non-empty subsets of clusters, which ideally have no intersection and the area of which separates the clusters can take any form [8]:

$$K = \{K_s\}, s = 1,\ldots,k;\ \ K_1 \cup K_2 \cup \ldots \cup K_k = A;$$

$$K_i \cap K_j = \varnothing, i \neq j;\ \ i, j = 1,\ldots,k,$$

where k – is the quantity of clusters. Model of objective clustering based on the inductive methods of complex systems modelling supposes sequential enumeration of clustering to choose the best one [7,8]. Let W – is the set of all-admissible clustering for a given dataset A. The best (an optimal) in term of quality criterion is the clustering, for which:

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

Clustering $K_{opt} \subseteq W$ is the objective if by quantity of clusters, by character of objects in appropriate clusters distribution and by quantity of disagreements it the least differs from the expert one [9]. Technology to create the objective clustering model based on the inductive methods of complex systems analysis supposes the following stages [8]:

- choosing of affinity function of studied objects or metrics to determine the degree of profiles sequences similarity in m-dimensional feature space;

• division of initial dataset into two equal power subsets. The term "equal power" in this case means that this subsets contain the same quantity of pairwise similar objects;

• determination of the method of clusters formation (sorting, regrouping, division, integration, and all);

• determination of internal and external criteria to estimate the clustering quality;

• organization of motion to the extremum of clustering quality criteria;

• determination of the objective clustering fixation method, which corresponds to the extreme value of clustering quality criteria.

S strategy to group the objects in m-dimensional feature space within the framework of the objective clustering inductive technology can be represented as follows:

$$S : \left\{ R(K) \,|\, (e \le e_0) \xrightarrow{\{QC\}} opt \right\}, \quad (1)$$

where $R(K)$ – is the clustering result, e – clustering error for two equal power subsets, e_0 – is the maximum permissible error, $\{QC\}$ – is the set of internal and external clustering quality criteria.

Under the strategy in this case is understood the purposeful process of sequential actions implementation to group the objects according to the problem statement within the framework of acceptable error. The error of clustering using two equal power subsets can be determined as the difference between the characters of objects grouping in appropriate clusters in different clustering:

$$e = f(e_1, e_2, \ldots, e_k) \le e_0, \quad (2)$$

where k – is the quantity of clusters in appropriate clustering. Obviously that if the quantity of clusters in different clustering varies the error automatically exceeds the allowable maximum, thus this clustering is unsatisfactory.

Principles of objective clustering inductive technology. Three fundamental principles are the basis of the inductive methodology of the complex systems analysis [1-5]:

 ISSN 1562-9945

1. The principle of heuristic self-organization that is sequential enumeration of different complicate models to choose the best model by a prior determined external clustering quality criterion.

2. The principle of external addition, the basic idea of which is the necessity of "fresh information" use for objective verification of the model.

3. The principle of inconclusiveness of solutions, the idea of which is the generation of a certain set of intermediate results and then selection from them the best variants.

Implementation of these principles in the adapted variant creates the premise to create the objective clustering inductive technology of complex data.

The principle of sequential enumeration. Model of the objective clustering based on the inductive methods of complex systems analysis supposes sequential enumeration of clustering for two equal power subsets, herewith the clustering result is estimated at the each step by computing of external clustering quality criteria, which is determined as the difference of clustering result for two subsets. The model self-organizes in such a way that depending on the type of the used algorithm and the affinity metric of objects and clusters better clustering correspond to the extremum of these criteria that correspond to the objective clustering in terms of these criteria. During the clustering enumeration it is possible that the value of the external criterion has several local extrema, which correspond to different levels of objects clustering. This phenomenon occurs in case of hierarchical clustering process when the clustering on two subsets is very similar during the sequential objects grouping or division. This fact leads to the appearing of local minimum of external criteria for this level of clustering. In this case, the choice of optimal clustering is done by expert based on complex analysis of values of internal clustering quality criteria for two parts of the initial dataset and the values of external clustering quality criteria.

The principle of the external edition. The principle of the external addition assumes the use of "fresh information" for an objective verification of the model and selection the best model during the multi-inductive procedure of the optimal model synthesis. Within the framework of the objective clustering inductive technology the

implementation of this principle presupposes the existence the two equal power subsets, which contain the same quantity of pairwise objects in terms of the used affinity function. The clustering process is carried out on two equal power subsets during the algorithm operation, herewith the choice the best solutions for each subsets is performed by the internal criteria and the final decision about the objects grouping is done on the basis of the external clustering quality criteria. The idea of the algorithm to divide the initial dataset Ω into two equal power subsets Ω^A and Ω^B is described in [7] and further developed in [8]. Implementation of this algorithm involves the following steps:

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

2. Allocation of the pairs of objects X_s, 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 Ω^A, and the object X_p to subset Ω^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.

Implementation of the external edition principle within the framework of objective clustering inductive technology assumes existence of external clustering quality criterion, which can be complex and take into account both clusters distribution in obtained clustering and objects distribution in the appropriate clusters in various clustering.

The principle of inconclusiveness of solutions. Implementation of this principle within the framework of objective clustering inductive technology involves the fixation of clustering, which correspond to the local minimums of the external clustering quality criterion estimating the character of objects distribution in the obtained clustering. Each local minimum corresponds to the objective clustering of the certain level of detailing. The final decision and, consequently, the objective

 ISSN 1562-9945

clustering fixation is determined by expert in accordance with the aim of the task at the current level of its solution.

Thus, it can be concluded that all three principles of the methodology of the complex systems inductive modelling have formed the basis of the objective clustering inductive technology of high dimensional data.

Affinity metrics in objective clustering inductive technology of high dimensional data. Grouping of the objects into clusters and estimation of the clusters and objects proximity is performed based on the similar metrics, the choice of which depends on the properties of the studied data and the nature of their distribution. In case of high dimensional data analysis to estimate the proximity level between the objects, clusters, objects and clusters the proximity level between vectors is estimated, the length of which is determined by feature space dimensional of studied data. In this case the degree of proximity between the vectors is named affinity (proximity, relationship), which can be determined as distance or degree of proximity between the appropriate vectors of features in m-dimensional space. The most prevalent metrics to analyse the high dimensional numeric sequences are:

- Euclid distance:

$$d_e(X_a, X_b) = \left(\sum_{i=1}^{m} (x_{ai} - x_{bi})^2 \right)^{\frac{1}{2}}; \quad (3)$$

- correlation distance:

$$d_{cor}(X_a, X_b) = 1 - \frac{\sum_{i=1}^{m} (x_{ai} - \overline{x}_a)(x_{bi} - \overline{x}_b)}{\sqrt{\sum_{i=1}^{m} (x_{ai} - \overline{x}_a)^2} \cdot \sqrt{\sum_{i=1}^{m} (x_{bi} - \overline{x}_b)^2}}; \quad (4)$$

- Manhattan distance:

$$d_m(X_a, X_b) = \sum_{i=1}^{m} |x_{ai} - x_{bi}|, \quad (5)$$

where X_a and X_b – are the vectors of features, affinity between which should be estimated, x_{ai} and x_{bi} – are the i-th features of the vectors X_a and X_b concurrently.

Hamming distance can be used as proximity metric to analyse vectors with dichotomy features, which in general case allows to determine the quantity of the differences of appropriate features for vectors X_a and X_b:

$$d_h(X_a, X_b) = \sum_{i=1}^{m} |x_{ai} - x_{bi}|, \tag{6}$$

here with, it should be noted that

$$|x_{ai} - x_{bi}| = \delta_i = \begin{cases} 1, if\ x_{ai} \neq x_{bi}, \\ 0, if\ x_{ai} = x_{bi} \end{cases}. \tag{7}$$

Some extensions of the hamming distance for binary series based on the relative quantity of matching or not matching bits are used to assess the degree of similarity of antigens and antibodies for development of artificial immune systems:

$$a = \sum_{i=1}^{m} \varphi_i, \quad \varphi_i = \begin{cases} 1, if\ x_{ai} = x_{bi}, \\ 0, if\ x_{ai} \neq x_{bi} \end{cases}$$

$$b = \sum_{i=1}^{m} \lambda_i, \quad \lambda_i = \begin{cases} 1, if\ x_{ai} = x_{bi}, \\ 0, if\ x_{ai} \neq x_{bi} \end{cases}$$

$$c = \sum_{i=1}^{m} \gamma_i, \quad \gamma_i = \begin{cases} 1, if\ x_{ai} = x_{bi}, \\ 0, if\ x_{ai} \neq x_{bi} \end{cases}$$

$$d = \sum_{i=1}^{m} \psi_i, \quad \psi_i = \begin{cases} 1, if\ x_{ai} = x_{bi}, \\ 0, if\ x_{ai} \neq x_{bi} \end{cases},$$

where x_{ai} and x_{bi} are i-th features of vectors x_a and x_b concurrently; a – is the quantity of single bits that coincide in both vectors; b – is the quantity of single bits of vector x_a that does not match with the

 ISSN 1562-9945

appropriate bits of vector x_b; c – is the quantity of zero bits of vector x_a that does not match with appropriate bits of vector x_b; d – is the quantity of zero bits that coincide in both vectors.

Combinations of the presented estimations in different variants allow receiving different affinity functions for studied vectors. It should be noted that the expediency of metric proximity using in the studied feature space is determined by the type of features and the nature of their distribution, herewith, the choice of the required affinity function for studied data is carried out empirically during the computer simulation.

To determine the optimal affinity metric of high dimensional data vectors the technology, which assumes using of two datasets that previously belong to different clusters have been proposed. An example of the objects and clusters distribution in the proposed technology is presented in fig. 1.

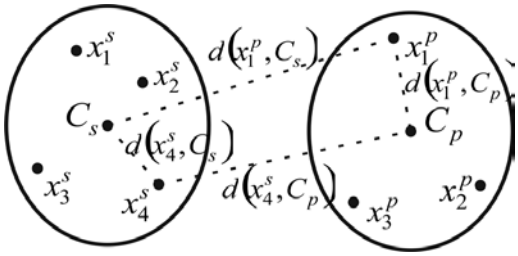

Fig. 1 – Model of the objects and cluster's centers distribution in two-cluster structure

Implementation of this technology includes the following steps:

1. Calculation of mass centres for clusters S and P:

$$C_{s,p} = \frac{1}{N_{s,p}} \sum_{i=1}^{N_{s,p}} x_{ij}^{s,p}, \tag{8}$$

where N_s and N_p are the quantity of objects in clusters S and P concurrently; x_{ij}^s, x_{ij}^p – are the j-th features of i-th vector in clusters S and P; $j = 1,\ldots,m$.

2. Calculation of the average distances between the appropriate feature vectors of studied objects and mass centres of clusters, inside of which these objects are:

$$d_{\text{int}}\left(X^{s,p}, C_{s,p}\right) = \frac{1}{N}\left(\sum_{i=1}^{N_s} d\left(x_i^s, C_s\right) + \sum_{i=1}^{N_p} d\left(x_i^p, C_p\right)\right). \quad (9)$$

3. Calculation of average distances between the appropriate feature vectors of studied objects and mass centres of the neighbour clusters:

$$d_{ext}\left(X^{s,p}, C_{s,p}\right) = \frac{1}{N}\left(\sum_{i=1}^{N_s} d\left(x_i^s, C_p\right) + \sum_{i=1}^{N_p} d\left(x_i^p, C_s\right)\right). \quad (10)$$

4. Calculation of the relative coefficient:

$$d_{rel}\left(X^{s,p}, C_{s,p}\right) = \frac{d_{ext}\left(X^{s,p}, C_{s,p}\right)}{d_{\text{int}}\left(X^{s,p}, C_{s,p}\right)}. \quad (11)$$

Obviously that the higher value of relative coefficient corresponds to the better division ability of the appropriate metric.

Internal criteria in the objective clustering inductive technology. There is the necessity of clustering quality estimation using two equal power data subsets during the objective clustering inductive technology implementation, herewith some estimations for the same data may not coincide with each other using different algorithms and estimation functions. Thus, it is necessary to estimate the accuracy of simulation results to the aims of the solvable task. In reality in the most cases the quantity of clusters is unknown, therefore it is necessary to select the best solutions which correspond to the extrema of the clustering quality criteria during the clustering algorithm operation. Obviously that the qualitative clustering corresponds to the high division ability of different clusters and high density of the objects concentration inside the clusters. Thus, the internal clustering quality criterion should be complex and takes into account both objects distribution inside single clusters and clusters distribution in the features space. The first component of the complex criterion is calculated as average distance from the objects to the mass centre of the cluster, where these objects are:

$$QCW = \frac{1}{N}\sum_{s=1}^{K}\sum_{i=1}^{N_s} d\left(x_i^s, C_s\right). \quad (12)$$

The second component of the complex criterion which takes into account the nature of the clusters distribution in feature space is calculated as an average distance between the centres of the clusters:

$$QCB = \frac{2}{K(K-1)} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} d(C_i, C_j), \tag{13}$$

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 S cluster, C_i, C_j and C_s – are the mass centres of the clusters i, j and s concurrently, $d(\cdot)$ – is the metric used to estimate the proximity level of the studied vectors.

The comparison analysis of the clustering quality criteria using different combinations of metrics (12) and (13) is carried out and described in [11-16]. As the main criteria the following can be selected:

- Calinski-Harabasz [12]:

$$QC_{CH} = \frac{QCB \cdot (N - K)}{QCW \cdot (K - 1)}; \tag{14}$$

- Hartigan [16]:

$$QC_H = \log_2 \frac{QCB}{QCW}; \tag{15}$$

- WB index [11]:

$$QC_{WB} = \frac{K \cdot QCW}{QCB}. \tag{16}$$

Structural block diagram of the process of determining the quantity of the clusters based on the internal clustering quality criteria is shown in Fig. 2.

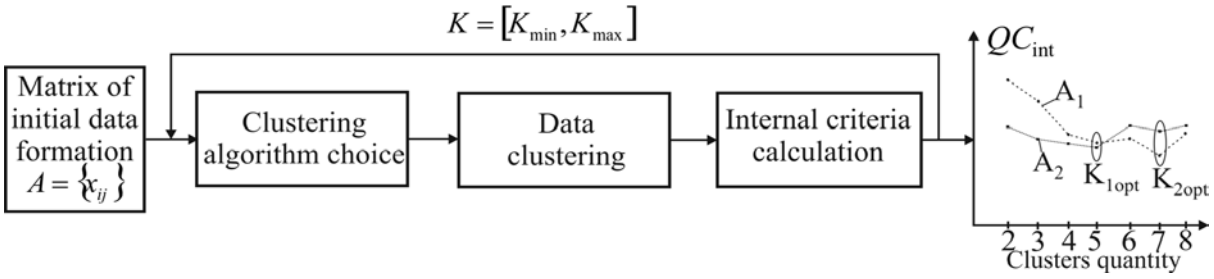

Fig.2 – Block diagram of determination of the optimal clustering based on the internal clustering quality criteria for subsets A_1 and A_2

As it can be seen in Fig. 2, the optimal clustering corresponds to the local minimums (or maximums) of internal clustering quality criterion, herewith, it is possible to obtain several extrema within the

given range. Each of these extrema may correspond to adequate grouping of the objects with different level of process detailing. Block diagram of the algorithm of this process implementation is presented in Fig. 3.

Implementation of this algorithm supposes the following steps:

1) formation of the initial data matrix;

2) setup of the initial clusters quantity: $K = K_{min}$;

3) data clustering and clusters fixation. Centres of clusters calculation;

4) calculation of the internal clustering quality criteria;

5) if quantity of clusters is less than K_{max}, increase of clusters quantity of 1 and return to step 3 of this algorithm. Otherwise, fixation of the internal clustering quality criteria matrix:

$$QC_{\text{int}} = \begin{Bmatrix} QC_{11}^{\text{int}} & \dots & QC_{1k}^{\text{int}} \\ \dots & \dots & \dots \\ QC_{n1}^{\text{int}} & \dots & QC_{nk}^{\text{int}} \end{Bmatrix}, \quad (17)$$

where $k = K_{\min},\dots,K_{\max}$ – is the serial number of clustering; n – is the quantity of the used internal clustering quality criteria.

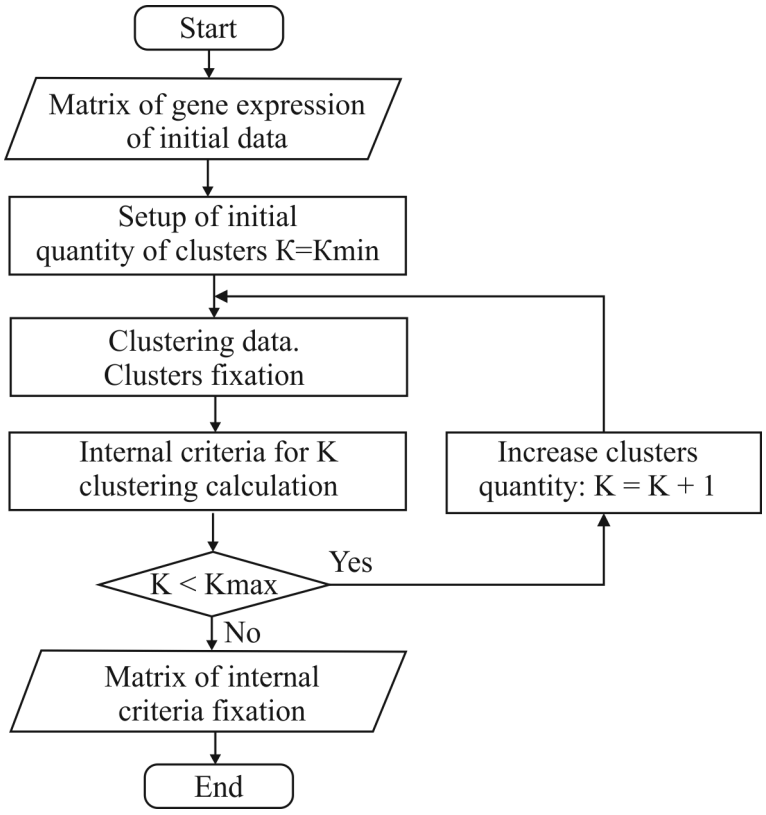

Fig.3 – Algorithm to calculate the internal criteria matrix

 ISSN 1562-9945

However, it should be noted that based on the internal clustering quality criteria it is impossible to assess the objectivity of the objects grouping because this accession is possible in case of the "fresh" information existence, in other words, on the basis of the quantitative estimation of the character of the objects grouping into two equal power subsets using external clustering quality criteria.

External criteria in the objective clustering inductive technology. As noted hereinbefore, one of the significant disadvantages of the existing clustering algorithms is the reproducibility error, in other words, high accuracy of corresponding clustering algorithm operation on a single dataset does not guarantee the similar results on other similar datasets. This problem is solved in the proposed inductive technology with the use of two equal power subsets, while the clustering is performed on the two subsets concurrently with simultaneous comparison of the intermediate results. This approach assumes the necessity of another criterion creation – analogue to the consistency criterion in the theory of the complex system inductive modelling. The technology of this criterion use in inductive modelling of the complex systems involves that the models with the same structure on two parts of the studied dataset give maximally similar results. The result of the simulation on different subsets in objective clustering inductive technology is the matrix of the intermediate results (17). If to use one internal criterion, the matrix (17) is transformed into vector-row:

$$QC_{\text{int}} = \left(QC_1^{\text{int}},\ldots,QC_k^{\text{int}}\right). \qquad (18)$$

An example of possible locations of the objects in three cluster objective clustering model is shown in Fig. 4.

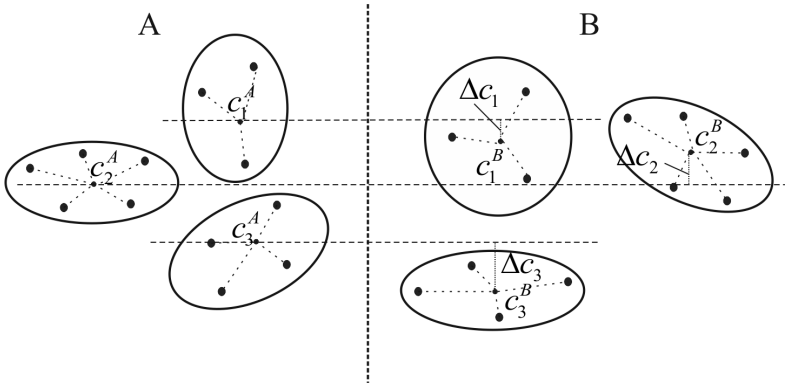

Fig. 4 – An example of the objects location in three cluster objective clustering model

The matrix of external clustering quality criteria within the framework of the objective clustering inductive technology is calculated as a normalized difference of the internal criteria (17), which are calculated for equal power subsets within a given range of possible clustering (from K_{min} to K_{max}):

$$QC_{ext}(A,B)=\begin{Bmatrix} QC_{11}^{ext}(A,B) & \dots & QC_{1k}^{ext}(A,B) \\ \dots & \dots & \dots \\ QC_{n1}^{ext}(A,B) & \dots & QC_{nk}^{ext}(A,B) \end{Bmatrix}. \quad (19)$$

Each component of matrix (19) is calculated by formula:

$$QC_{ij}^{ext}(A,B)=\frac{\left|QC_{ij}^{\text{int}}(A)-QC_{ij}^{\text{int}}(B)\right|}{QC_{ij}^{\text{int}}(A)+QC_{ij}^{\text{int}}(B)}, \quad (20)$$

where $i=1,\dots,n-$ is the quantity of the internal clustering quality criteria, $j=1,\dots,k$ – is the quantity of the clusters within the range from K_{min} to K_{max}. In case of one internal criterion use we have:

$$QC_{ext}(A,B)=\left(QC_{1}^{ext}(A,B),\dots,QC_{k}^{ext}(A,B)\right). \quad (21)$$

The choice of the objective clustering is performed based on the analysis of the columns of matrix (19) or by vector (21) analysis. In the easiest case of one criterion existence the objective clustering corresponds to the extremum (minimum) value of vector (21). In case of larger quantity of the external criteria the task is multicriterial and the choice of an optimal solution is performed based on the complex analysis of the external criteria values of matrix (19) using model data with a priori known value of the target function that corresponds to the objective clustering for the studied data.

Architecture of the objective clustering inductive technology. Fig. 5 shows the general architecture of the objective clustering inductive technology implementation. The data matrix, whose rows are the studied objects, is supplied to the input of the system. The aggregate of clusters, each of which includes the objects with high affinity features is the output of the system. Implementation of this technology supposes the following stages.

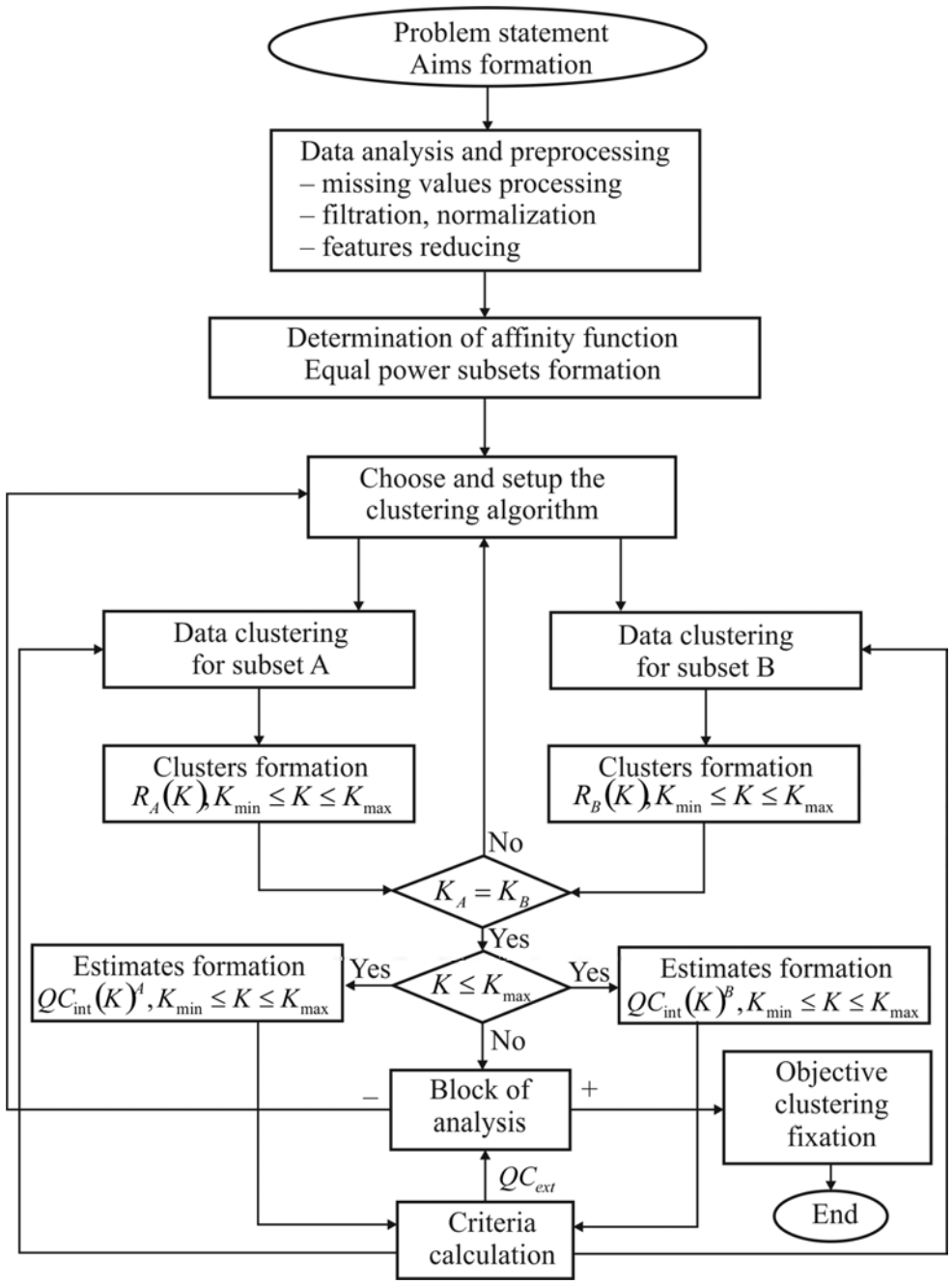

Fig. 5 – Architecture of the objective clustering inductive technology

Stage I.

1. Problem statement. Clustering aims formation according to the solved task.

2. Analysis of the studied data, presentation of the data as a matrix: $A = \{x_{ij}\}, i = 1,\ldots,n; j = 1,\ldots m$, where n – is the quantity of the studied objects, m – is the quantity of features characterizing these objects.

3. Data pre-processing, includes missing values processing (if it is necessary), data filtration and normalization.

4. Reducing the dimension of the feature space of the studied data with the use of the criteria to estimate the informational content of the studied data. In case of gene expression sequences these criteria are: average, variance, Shannon entropy.

5. Affinity function determination in dependence with the type of the studied data.

6. Equal power subsets A and B formation.

7. Choice and setup of the initial parameters of the clustering algorithm. The parameters are changed during further algorithm operation for the purpose of optimal values determination in terms of minimum values of the external clustering quality criterion.

Stage II.

8. Data clustering on subsets A and B. Clusters formation within the range $K_{\min} \le K \le K_{\max}$. If the clusters quantity in different clustering differs, clustering process is stopped due to the unsuccessful choice of the clustering algorithm. In this case it is necessary to use another admissible clustering algorithm and to repeat step 7.

9. Estimation of the private clustering formation, calculation of the internal quality criteria QC_{int} for current clustering on subsets A and B.

10. Calculation of the external quality criteria QC_{ext} for this clustering.

Stage III.

11. Plotting of the external clustering quality criteria depending on the clusters quantity within the given range: $K_{\min} \le K \le K_{\max}$.

12. Analysis of the obtained results. In case of the local minimums absence or if their values are more than admissible standards (sign "–" in fig. 5) choose another clustering algorithm. Repetition of stages II-III of this procedure.

13. Fixation of the objective clustering corresponding to the minimum value of the external clustering quality criteria.

Conclusion

The article presents the objective clustering technology of gene expression sequences based on the methods of complex systems inductive modelling. The architecture of step by step implementation of this technology, which involves a parallel data clustering on the two equal power subsets that include the same quantity of pairwise similar objects,

have been developed. The complex criterion which takes into account both the character of the objects distribution within the clusters and the character of the clusters distribution in the feature space have been proposed to estimate the clustering quality on a different datasets. The paper presents also the algorithm to form the matrix of the clustering quality internal criteria. The final decision about the studied objects grouping is done based on the external clustering quality criterion, which is calculated as normalized difference of the appropriate internal criteria for various clustering. The objective clustering corresponds to the minimum value of this criterion. Implementation of the proposed technology involves the possibility of different clustering algorithms use. Choosing the best clustering algorithm for the studied data within the framework of the presented technology assumes computer simulation of the studied data clustering process with further evaluation of the obtained results. These are the perspectives of further research of the author.

References:

1. Ivakhnenko A.G. Group method of data handling – a competitor of stochastic approximation method // Automatics, 1968. – №3. – P. 58-72. [In Ukrainian].
2. Ivakhnenko A.G. Inductive method for self-organization of complex systems models.– Kiev: Scientific Thought,1982.– 296 p. [In Russian].
3. Ivakhnenko A.G. objective clustering based on the theory of self-organization models // Automatics, 1987. – №5. – P. 6-15. [In Russian].
4. Stepashko V.S. Theoretical aspects of GMDH as a method of inductive modelling // Managing Systems and Machines, 2003. – №2. – P. 31-38. [In Russian].
5. Stepashko V.S. Elements of the inductive modelling theory / State and prospects of informatics development in Ukraine: Monograph / Team of authors. – Kiev: Scientific Thought, 2010. – 1008 p. /– P. 471-486. [In Ukrainian].
6. Osypenko V.V. Two approaches to solving the problem of clustering in the broad sense from the standpoint of inductive modeling // Power and Automation, 2014. - №1. - P.83-97. [In Ukrainian].

7. Madala H.R., Ivakhnenko A.G. Inductive Learning Algorithms for Complex Systems Modelling.– CRC Press, 1994. – 365 p.
8. Sarycheva L.V. Objective cluster analysis of the data based on the Group Method of Data Handling // Problem of Management and Informatics, 2008. – №2. – P. 86-104. [In Russian].
9. Ivakhnenko A.G., Coppa J.V., Petuchova C.A., Ivakhnenko M.A. The use of self-organization to divide the dataset into a priori undetermined number of clusters // Automatics. – 1985. – №5. – P. 9-16. [In Russian].
10. De Castro L.N. Artificial Immune Systems: A New Computational Intelligence Approach / L.N. De Castro, J. Timmis. Springer, Heidelberg, 2002. – 357 p.
11. Q. Zhao, M. Xu, P. Frдnti, Sum-of-Squares Based Cluster Validity Index and Significance Analysis, Proceeding of International Conference on Adaptive and Natural Computing Algorithms, pp. 313-322, 2009.
12. Calinski T., Harabasz J. A dendrite method for cluster analysis // Communication in statistics. – 1974. – №3. – P. 1–27.
13. Halkidi M., Vazirgiannis M. Clustering validity assessment: Finding the optimal partitioning of a data set // Proc. of the 2001 IEEE Int. Conf. on Data Mining (ICDM'01). – 2001. – P. 187-194.
14. Still S., Bialek W. How many clusters? An information theoretic perspective // Neural Computation. – 2004. – №16. – P. 2483-2506.
15. Halkidi M., Batistakis Y., Vazirgiannis M. Clustering validity checking methods: Part II // ACM SIGMOD Record. – 2002. – №31. – P. 19-27.
16. Hartigan J. Clustering algorithms // New York, NY: Wiley. – 1975. – 369 p.

 ISSN 1562-9945