Objective Clustering Inductive Technology of Gene Expression Sequences Features

Sergii Babichev[1(✉)], Volodymyr Lytvynenko[2], Maxim Korobchynskyi[3], and Mochamed Ali Taiff[2]

[1] Jan Evangelista Purkyne University, Usti nad Labem, Czech Republic 8, Ceske Mladeze Street, 400 96 Usti nad Labem, Czech Republic
sergii.babichev@ujep.cz
[2] Kherson National Technical University, Kherson, Ukraine
immun56@gmail.com
[3] Military-Diplomatic Academy Named Eugene Bereznyak, Kiev, Ukraine
http://www.sci.ujep.cz

Abstract. Technology of high dimensional data features objective clustering based on the methods of complex systems inductive modeling is presented in the paper. Architecture of the objective clustering inductive technology as a block diagram of step-by-step implementation of the objects clustering procedure was developed. Method of criterial evaluation of complex data clustering results using two equal power data subsets is proposed. Degree of clustering objectivity evaluates on the basis of complex use of internal and external criteria. Researches on the simulation results of the proposed technology based on the SOTA self-organizing clustering algorithm using the gene expression data obtained by DNA microarray analysis of patients with lung cancer GEOD-68571 Array Express database, the datasets "Compound" and "Aggregation" of the Computing School of the Eastern Finland University and the data "seeds" are presented.

Keywords: Clustering · Inductive modeling · Gene expression · High dimensional data

1 Introduction

The process of the gene regulatory networks creation based on the gene expression sequences suggests the steps to group genes using different proximity metrics at the stage of data preprocessing. Currently this problem is solved by various methods. Using the component or factor analysis partially solves the problem of the feature space dimension reducing, however, the partial loss and distortion of the initial information occurs during data transformation that has a direct influence to the problem solution accuracy. Technology of the bicluster analysis preserves the object-feature structure of data, but the feature space dimension of the objects subsets derived much smaller than the dimension of the original data that allows us to construct the gene regulatory networks in real-time with

S. Kozielski et al. (Eds.): BDAS 2017, CCIS 716, pp. 359–372, 2017.
DOI: 10.1007/978-3-319-58274-0_29

the preservation of the information about the influence specifics of the individual genes to the target node. The bicluster analysis questions for gene expression sequences processing are considered in [10,17]. The authors analyzed various biclustering algorithms and extracted their advantages and disadvantages. In [6] was conducted a comparative analysis of different biclustering algorithms for the analysis of gene expression profiles. The [11] presents a study on the use of the spectral biclustering technique for the analysis of the gene expression data on the example of the simulated data. The distribution diagram of objects and the specifics of their grouping in different biclusters are showed. It should be noted that this technology has high actuality in the context of feature extraction for the construction of the gene regulatory networks nowadays. However it should be noted, that in spite of archived progress in this subject area, there are some problems associated with: the choice of the biclusters quantity and the degree of detailing of the objects and features in corresponding biclusters; the choice of the metrics to estimate the proximity of the objects and features vectors concurrently. The use of traditional clustering algorithms to group the feature vectors according to their degree of similarity is an alternative to the bicluster analysis. A lot of clustering algorithms exist nowadays. Each of them has its advantages and disadvantages and is focused on a specific type of data. One of the essential disadvantages of the existing clustering algorithms is the nonreproductivity error, i.e., high clustering quality on a single dataset does not guarantee the same results on another similar dataset. To raise the clustering objectivity is possible by developing the hybrid models based on the inductive methods of complex systems modeling, which is a logical continuation of the group method of data handling (GMDH) [9]. The questions of inductive methods of complex systems objective self-organizing models creation are presented in [14] and further developed in [18]. The authors have presented the researches concerning the implementation of the inductive modeling principles for creating the systems of objects complex nature self-organizing based on the group method of data handling. Researches concerning the use of inductive modeling methods to create the inductive technologies of informational and analytical researches for different nature information analysis are presented in [16]. However, it should be noted that the authors' studies are focused primarily on the low dimensional data, at the same time insufficient attention is paid to the inductive models based on the clustering enumeration for the purpose of their self-organizing with the use of the external balance criteria of clustering quality assessing by equal power subsets.

The aim of the paper is working out the technology of creation the objective clustering inductive model of complex nature high dimensional data and its practical implementation by DNA microarray experiments use.

2 Problem Statement

Let the initial dataset of the objects is a matrix: $A = \{x_{ij}\}, i = 1, \ldots, n;\ j = 1, \ldots, m$, where n – is the number of objects observed, m – is the number of

features characterizing the objects. The aim of the clustering process is a partition of the objects into nonempty subsets of pairwise nonintersecting clusters, herewith a surface which divides the clusters can take any shape [9,18]:

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

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

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

Clustering $K_{opt} \subseteq W$ is an objective if it has the least difference from an expert by the number of objects, the character of the objects distribution in the appropriate clusters and the number of discrepancies [9,18].

The technology of the objective clustering inductive model creation assumes the following stages:

1. Assignment an affinity function of studied objects, i.e., finding the metric to determine the degree of objects similarity in m-dimensional feature space.
2. Development of the algorithm to partition the initial set of the objects into two equal power subsets. The equal power subsets are the subsets which contain the same number of pairwise similar objects.
3. Assignment a method of clusters formation (sorting, regrouping, grouping, division, etc.).
4. Assignment the criterion QC of quality clustering estimation as a measure of the clusters similarity in various clustering.
5. Organization of motion to max, min or optimal value of the criteria QC of quality clustering estimation.
6. Assignment an objective clustering fixation method corresponding to the extremum of the criteria value of quality clustering estimation.

Figure 1 shows the chart of the modules interaction in the objective clustering inductive model. The choice of affinity functions to assess the degree of proximity from objects to clusters is determined by the nature of the studied objects features. The method of clusters formation in inductive model is determined by clustering algorithm used for parallel grouping the objects in equal power subsets. The character of equal power subsets formation is determined by the choice of the objects similarity measure which depends on the objects feature space properties. To choose the objective clustering it is necessary at the early stage to define the internal and the external criteria, extremum value of which will allows to fix an objective clustering for the studied data subsets during clustering enumeration.

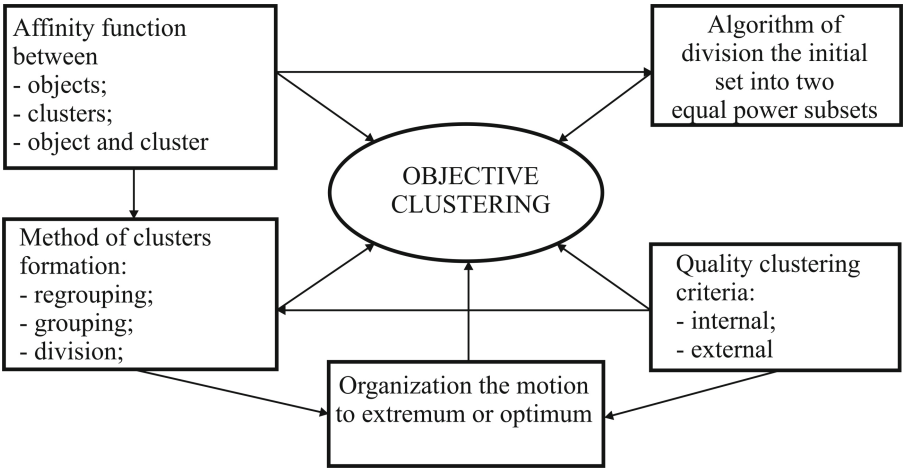

Fig. 1. Charts of the modules interaction in the objective clustering inductive model

3 Principles of the Objective Clustering Inductive Technology

Three fundamental principles borrowed from different scientific fields allowed to create the complete, organic and interconnected theory, are the basis of the methodology of the complex systems inductive modeling [9,16,18]:

- the principle of heuristic self-organizing, i.e., enumeration of the models set aiming to select from them the best on the basis of the external balance criteria;
- the principle of external addition, i.e., the necessity to use several equal power data subsets with the purpose of objective verification of models;
- the principle of solution inconclusive, i.e., generation of certain sets of intermediate results in order to select from them the best variants.

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

3.1 Principle of Heuristic Self-Organizing

Inductive model of objective clustering assumes a sequential enumeration of the clustering by using the two equal power subsets; herewith the result of clustering is estimated at each step by calculating the external balance criterion, which determines the difference of the objects clustering results on the two subsets. The model self organizes so that the best clustering correspond to an extremum value of this criterion depending on the type of the algorithm and the measures of the objects and clusters similarity. During the process of clustering enumeration it is

possible that the value of the external criterion has several local extremums corresponding to different objects clustering. This phenomenon is occurred in case of a hierarchical clustering, when clustering on the two subsets are sufficiently similar during the sequential process of objects grouping or separation that leads to the appearing of the local minimum at a given level of the hierarchy. In this case the choice of an objective clustering is determined by the goals of the task whereas each of the clustering that corresponds to the extremum value of the external balance criteria may be considered as the objective and the choice of the final clustering is determined by the required objects partition or grouping detailed elaboration level.

3.2 Principle of the External Edition

The principle of the external addition in the model of the group method of data handling (GMDH) assumes the use of "fresh information" for an objective verification of the model and selection of the best model during the process of multiserial inductive procedures of optimal model synthesis. The implementation of this principle in the framework of the objective clustering inductive model supposes the existence of the two equal power subsets, which contain the same number of pairwise similar objects in terms of their attributes values of objects. Clustering is carried out on the two equal power subsets concurrently during the algorithm operation with the sequential comparison of clustering results by chosen external balance criteria. The idea of the algorithm to divide the initial dataset of the objects Ω into two equal power subsets Ω^A and Ω^B is stated in [9] and further developed in [18]. Implementation of this algorithm assumes the following steps:

1. Calculation of $\frac{n \times (n-1)}{2}$ pairwise distances between the objects in the original sample of data. The result of this step is a triangular matrix of the distances.
2. Allocation of the pairs of objects X_s and X_p, the distance between which is minimal:
$$d(X_s, X_p) = \min_{i,j} d(X_i, X_j);$$
3. Distribution of the object X_s to subset 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.

3.3 Principle of the Solution Inconclusive

The implementation of this principle, which is relative to the inductive model of objective clustering, assumes a fixation of clustering which correspond to the local minimum or maximum of external balance criterion for different levels of the hierarchical tree. Each local extremum corresponds to an objective clustering with a certain degree of detailing. The final choice and therefore the fixation of the obtained clustering is determined by the goals of the task at this stage of its solving.

4 Criteria in the Objective Clustering Inductive Technology

The necessity of the clustering quality estimation on the several equal power subsets occurs during the process of implementation of the objective clustering inductive technology, herewith separate estimations may not coincide with each other while using different algorithms and different evaluation functions for the same data. Thus, there is a necessity of the estimation of the correspondence of the modeling results to the purposes of the task in view.

4.1 Internal Criteria in the Objective Clustering Inductive Technology

Usually in most cases the number of clusters is unknown, therefor the best solutions which correspond to the extremums of the internal criteria are allocated during the process of clustering algorithm operation. High level of the clustering, obviously, corresponds to a high separating capability of various clusters and high density of objects distribution within clusters. Therefore, an internal criterion of clustering quality evaluation should include two components: the sum of squared deviations of objects relative to the corresponding centroid within clusters QC_W and the sum of the squared deviations of clusters centroids relative to a general mass center of all clusters QC_B. The formulas to calculate these internal criterion components can be presented as follows:

$$QC_W = \sum_{j=1}^{K} \sum_{i=1}^{N_j} d(x_i^j, C_j)^2$$

$$QC_B = \sum_{j=1}^{K} N_j d(C_j, \bar{C})^2$$

where K – is the quantity of clusters, N_j – is the quantity of the objects in j cluster, C_j – is the centroid of cluster j: $C_j = \frac{1}{N_j} \sum_{i=1}^{N_j} x_i^j$, x_i^j – is the object i in cluster j, $\bar{C}$ – is the general centroid of studied objects, $d(X_a, X_b)$ – is the similarity distance between vectors X_a and X_b. The correlation distance was used as similarity distance in the case of gene expression sequences analysis:

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

where m – is the number of sequences features, x_{ai} and x_{bi} – are the i-th features of the X_a and X_b sequences respectively, $\bar{x}$ – is the mean value of the correspond sequence features.

Comparative analysis of the quality clustering internal criteria estimation by the use of various types and combinations of the presented measures are

showed in [7,8,12,15,19,20]. Structural block diagram of the process of the clusters quantity determination on the basis of the internal criteria is shown in Fig. 2. Implementation of this process assumes the following steps:

1. Application of the selected clustering algorithm for clustering K within the limits of allowable range $K = [K_{min}, K_{max}]$.
2. Fixation of the obtained clustering, calculation of the clusters centroids.
3. Calculation of the internal criteria for obtained clustering.
4. Repetition of the steps 1–3 to obtain the required number of clusters within the given range.
5. Construction of the graphs of internal criteria versus the number of clusters. Analysis of the graphs, selection of the optimal clustering.

As can be seen from Fig. 2, an objective clustering corresponds to a local minimum values of the internal criterion, herewith several extremums can be observed within a clustering process. Each of the local minimum corresponds to an adequate grouping of objects with various degree of the process detailing. However, it should be noted that it is not possible to evaluate the clustering objectivity based on the internal criteria because this evaluation is possible if there is a "fresh" information based on an external criteria of evaluation of the corresponding clustering difference by using the two equal power subsets.

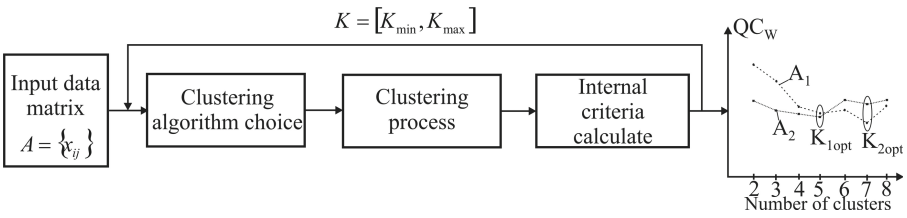

Fig. 2. Flowchart to choose the optimal clustering based on the internal criterion for the data A1 and A2

4.2 External Criteria to Estimate the Quality of the Objects Grouping

As noted hereinbefore, an adequate selection of the criteria to estimate the clustering quality on the different stages of the model operation is one of the major factors which promotes to the high efficiency of the objective clustering inductive model. These criteria should take into account both the location of the objects in the respective clusters relative to the corresponding mass center and the centroid position of the respective clusters according to the relation of clusters to each other in different clustering. An example of a possible location of the objects in a three-cluster objective clustering inductive model is shown in Fig. 3. The

position of the k-th cluster's centroid is defined as the average of the objects features in this cluster:

$$C_k = \frac{1}{n_k} \sum_{i=1}^{n_k} x_{ij},$$

where n_k – is the quantity of objects in k cluster, $j = 1, \ldots, m$ – is the quantity of features what characterize the objects.

The first component of the external criterion based on the assumption that the average value of the total displacement of corresponding clusters mass centers at the different clustering in the case of the objective clustering should be minimal. In case of normalization of criterion value formula takes the form:

$$QC_1(A, B) = \sqrt{\sum_{j=1}^{m} \left(\frac{\sum_{i=1}^{k} (C_i(A) - C_i(B)^2}{\sum_{i=1}^{k} (C_i(A) + C_i(B)^2}\right)^2} \longrightarrow \min$$

The second component of the external criterion takes into account the difference of the character clusters and the objects distribution in the respective clusters in different clustering. The average distance from the objects to the corresponding clusters centroids can be calculated as follows:

$$D_W = \frac{1}{k} \sum_{s=1}^{k} \left(\frac{1}{n_s} \sum_{i=1}^{n_s} d(X_i, C_s)\right)$$

where $s = 1, \ldots, k$ – is the number of clusters, n_s – is the quantity of objects in cluster s, C_s – is the centroid of s cluster, $d(\cdot)$ – is the correlation distance or Euclid distance in case of low dimensional data. The distance between the centroids of the clusters is defined as the average distance from the centroids to the mass center of the studied objects:

$$D_B = \frac{1}{k} \sum_{s=1}^{k} d(C_s, \bar{C}).$$

It is obviously that the clustering will be more qualitative when the density of the objects distribution within clusters is higher and the distance from the centroids of the clusters to the total mass center of objects are more: $D_W \longrightarrow \min$, $D_B \longrightarrow \max$. The complex internal criterion was calculated using Calinski-Harabasz criterion [20]:

$$D = \frac{D_W(K-1)}{D_B(N-K)}.$$

The second component of the external criterion can be represented as an absolute value of the internal criterion difference for various clustering. A normalized form of this formula takes the form:

$$QC_2(A, B) = \frac{|D(A) - D(B)|}{|D(A) + D(B)|}.$$

The objective clustering is selected based on the local minimum analysis of the both external criteria during the enumeration of all accessible clustering.

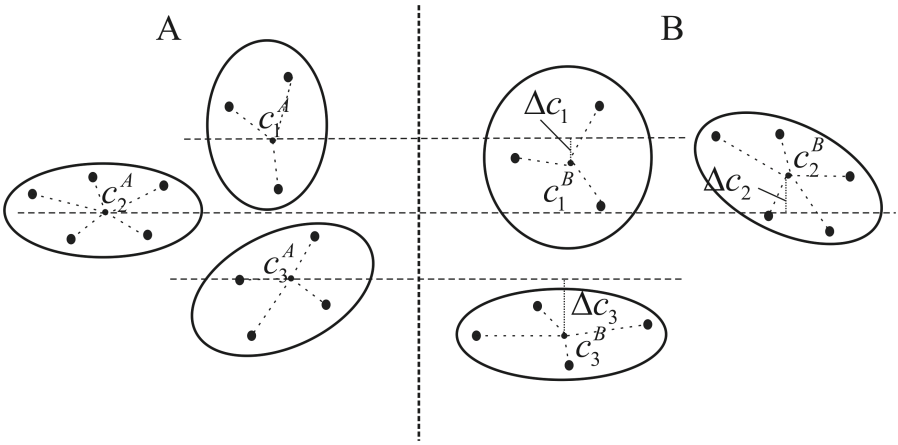

Fig. 3. An example of the objects and clusters location in objective clustering inductive technology

5 Architecture of the Step by Step Procedure of the Objective Clustering Inductive Technology

Figure 4 shows the general architecture of the objective clustering inductive technology implementation. There is a data matrix where the studied objects are given in rows and the features, defining the properties of the given objects are presented in the columns, are applied to the input of the system. The set of clusters, each of which includes a group of objects features which have a high affinity for these objects is the output of the system. Implementation of this technology supposes the following steps:

Phase I

1. Problem statement. Formation of the clustering aims.
2. Analysis of the studied data, definition of the studied objects feature character, brining of the data to a matrix view: $A = \{x_{ij}\},\ i = 1, \ldots, n;\ y = 1, \ldots, m$, where n – is the quantity of the studied objects, m – is the quantity of the features characterizing the objects.
3. Data preprocessing that includes the data filtration, data normalization, missing value restoration, dimension of the feature space reducing.
4. Determination of the affinity function for the further degree of objects' affinity estimation.
5. Formation of the equal power subsets A and B in accordance with hereinbefore algorithm.
6. Choosing and setup of the clustering algorithm. Initialization of the input parameters of this algorithm.

Phase II

1. Data clustering for subsets A and B, clusters formation inside the selected range $K_{min} \leq K \leq K_{max}$. If the number of clusters in a variety of clustering is differed the process is stopped due to the poor algorithm selection or incorrect setup of this algorithm. In this case it is necessary to apply other algorithm from the admissible set or to change the initial parameters of current algorithm.
2. Formation of the current clustering estimation, mass centers $C(K)^A$, $C(K)^B$ and internal criteria $QC_W(K)^A$ and $QC_W(K)^B$ for subsets A and B of current clustering calculation.
3. Calculation of the external criteria QC_1 and QC_2 for this clustering.

Phase III

1. Plotting of the charts of the calculated external criteria versus the number of the obtained clusters within a given range $K_{min} \leq K \leq K_{max}$.
2. Analysis of the obtained results. In case of external criteria local minimum absence or if the values of these criteria are more than permissible norms (Fig. 4 sign "–") selection of another clustering algorithm or reinitialization of the current algorithm initial parameters. Repetition of the steps 2–5 of the Phase 1 of this procedure.
3. In case of the local minimum presence under a condition of enumerating all clustering within given range fixation of the objective clustering corresponding to the minimum of the external criteria.

6 Experiment, Results and Discussion

Approbation of the proposed model was carried out by using the patients' data with lung cancer E-GEOD-68571 of the database Array Express [5], which includes the gene-expression profiles of 96 patients, 10 of which were healthy and 86 patients were divided by the degree of the disease into three groups. The size of the initial data matrix was (96×7129). The researches on the optimization of the gene expression data preprocessing for the purpose of features space informativity increasing and quantity of the genes reducing are presented in the [2,3]. Sample of 400 genes was used at the present stage of the simulation to simplify a computing, herewith the initial dataset was divided into two equal power subsets using hereinbefore algorithm. To compare the simulation results during inductive model operation also the datasets "Compound" and "Aggregation" of the Computing School of the Eastern Finland University [1] and the dataset "seeds", representing the researches of kernels of different kinds of wheat [13] were used. Each kernel was characterized by seven attributes. In work [4] authors used the agglomerative hierarchical clustering algorithm to data clustering within the framework of presented model. In this work the SOTA self-organizing clustering algorithm was used as a base within the framework of the proposed model. The simulation of the clustering was carried out by software R. The charts of

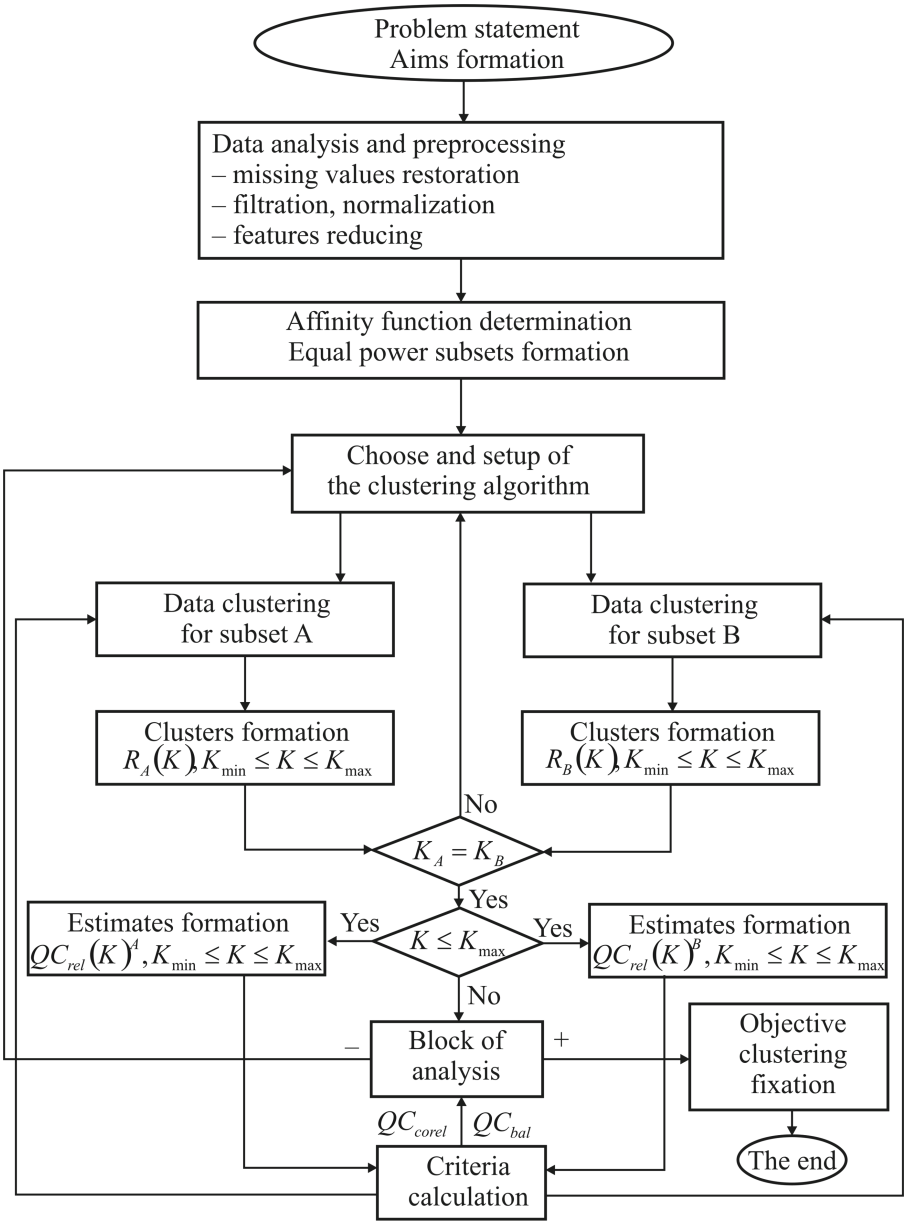

Fig. 4. Architecture of the objective clustering inductive technology

the used criteria versus the obtained clusters quantity for studied datasets are shown in Fig. 5. The number of the obtained clusters was changed from 2 to 11. The analysis of the charts allows to conclude that in terms of all external criteria

using the clustering for division of the objects into 7 clusters is an objective in case of the high dimensional gene expression data. In this case the position of the external criteria local minimums are the same and the value of the internal criterion is insignificantly different from the local minimum value corresponding the objects division into six clusters. In case of other low dimensional data the simulation results shows the low efficiency of the internal criterion because its value decreases while the level of the objects division increases, herewith the local minimums don't allow to make the conclusion about the quality of the model operation. Only in case of the "Compound" data, this criterion works in agreement with the external criteria that allows to fix the clustering with the objects division into 7 clusters. Comparative analysis of the external criteria versus the level of clustering allows to conclude that the external criterion QC_2 shows more adequate results during simulation process. In case of the "seeds" data this criterion has two local minimums corresponding to the objects division into 3 and 5 clusters, while extraction of the 3 clusters is not fixed by other criteria. In case of the "Compound" and "Aggregation" datasets this criterion fixes 7 clusters. These results completely agree with the results, which were presented in these data annotations.

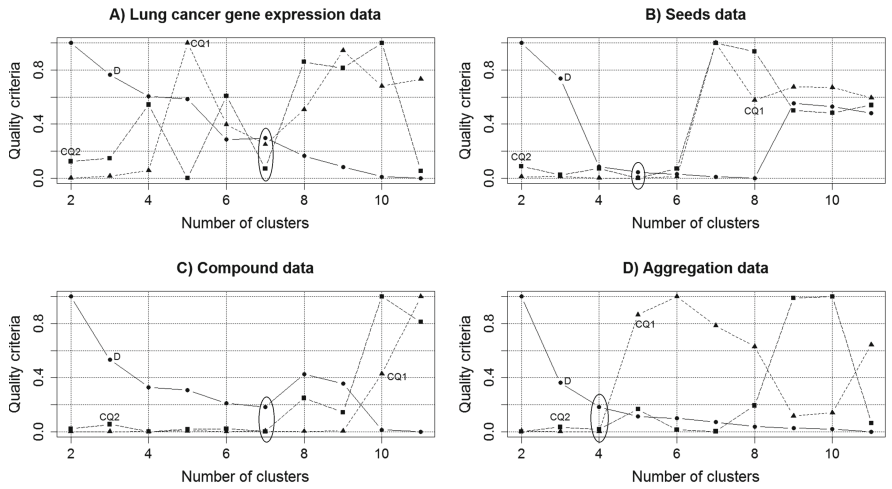

Fig. 5. Internal and external criteria to estimate the clustering quality: (A) lung cancer gene expression data; (B) seeds data; (C) compound data; (D) aggregation data

7 Conclusions

The researches aiming to create the technology of the objective clustering inductive model of the complex nature objects are presented in the paper. To improve the objectivity of the objects grouping the original data set is divided into two

equal power subsets, which contained the same number of the pairwise similar objects in terms of the correlation distance of their attributes profiles. The architecture of the objective clustering inductive model has been developed and practically implemented on the basis of the self-organizing SOTA clustering algorithm, while the evaluation of the partition objects into clusters quality at each step was estimated using an external criteria, which take into account the difference of the objects and the clusters distribution in different clustering. The sample of the lung cancer patients' gene expression profiles which contains 400 profile genes of 96 patients, "Compound", "Aggregation" and "Seeds" data were used to approbate the proposed model. The simulation results proved high efficiency of the proposed model operation. The local minimums values of the internal and external criteria allow to take more adequate solution about the choice of the studied data objective clustering. The further authors' researches will be focused on a more detailed study of the proposal model operation based on various clustering algorithms with the use of different nature data.

References

1. Machine learning school of computing university of eastern finland. Clustering datasets. https://cs.joensuu.fi/sipu/datasets/
2. Babichev, S.A., Kornelyuk, A.I., Lytvynenko, V.I., Osypenko, V.: Computational analysis of microarray gene expression profiles of lung cancer. Biopolymers Cell **32**(1), 70–79 (2016). http://biopolymers.org.ua/content/32/1/070/
3. Babichev, S., Taif, M.A., Lytvynenko, V.: Filtration of dna nucleotide gene expression profiles in the systems of biological objects clustering. Int. Front. Sci. Lett. **8**, 1–8 (2016). https://www.scipress.com/IFSL.8.1
4. Babichev, S., Taif, M.A., Lytvynenko, V.: Inductive model of data clustering based on the agglomerative hierarchical algorithm. In: Proceeding of the 2016 IEEE First International Conference on Data Stream Mining and Processing (DSMP), pp. 19–22 (2016). http://ieeexplore.ieee.org/document/7583499/
5. Beer, D.G., Kardia, S.L., et al.: Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat. Med. **8**(8), 816–824 (2002). http://www.nature.com/nm/journal/v8/n8/full/nm733.html
6. Eren, K., Deveci, M., Kucuktunc, O., Catalyurek, U.V.: A comparative analysis of biclustering algorithms for gene expression data. Briefings Bioinform. **14**(3), 279–292 (2012). https://doi.org/10.1093/bib/bbs032
7. Halkidi, M., Batistakis, Y., Vazirgiannis, M.: Clustering validity checking methods: Part 2. ACM SIGMOD Rec. **31**(3), 19–27 (2002). https://www.researchgate.net/publication/2533655_Clustering_Validity_Checking_Methods_Part_II
8. Halkidi, M., Vazirgiannis, M.: Clustering validity assessment: finding the optimal partitioning of a data set, pp. 187–194 (2001). http://ieeexplore.ieee.org/document/989517/?reload=true&arnumber=989517
9. Ivakhnenko, A.: Group method of data handling as competitor to the method of stochastic approximation. Sov. Autom. Control **3**, 64–78 (1968)
10. Kaiser, S.: Biclustering: methods, software and application (2011). https://edoc.ub.uni-muenchen.de/13073/
11. Kluger, Y., Basry, R., Chang, J., Gerstein, M.: Spectral biclustering of microarray data: coclustering genes and conditions. Genome Res. **13**(4), 703–716 (1985). http://genome.cshlp.org/content/13/4/703.abstract

12. Krzanowski, W., Lai, Y.: A criterion for determining the number of groups in a data set using sum of squares clustering. Biometrics **44**(1), 23–34 (1985). https://www.jstor.org/stable/2531893?seq=1#page_scan_tab_contents
13. Kulczycki, P., Kowalski, P.A., Lukasik, S., Zak, S.: Seeds data set. http://archive.ics.uci.edu/ml/datasets/seeds
14. Madala, H., Ivakhnenko, A.: Inductive Learning Algorithms for Complex Systems Modeling, pp. 26–51. CRC Press (1994). http://www.gmdh.net/articles/theory/ch2.pdf
15. Milligan, G., Cooper, M.: An examination of procedures for determining the number of clusters in a data set. Psychometrika **50**(2), 159–179 (1985). http://link.springer.com/article/10.1007/BF02294245
16. Osypenko, V.V., Reshetjuk, V.M.: The methodology of inductive system analysis as a tool of engineering researches analytical planning. Agric. Forest Eng. **58**, 67–71 (2011). http://annals-wuls.sggw.pl/?q=node/234
17. Pontes, B., Giraldez, R., Aguilar-Ruiz, J.S.: Biclustering on expression data: a review. J. Biomed. Inf. **57**, 163–180 (2015). https://www.ncbi.nlm.nih.gov/pubmed/26160444
18. Sarycheva, L.: Objective cluster analysis of data based on the group method of data handling. Probl. Control Automatics **2**, 86–104 (2008)
19. Still, S., Bialek, W.: How many clusters? An information theoretic perspective. Neural Comput. **16**(12), 2483–2506 (2004). http://www.mitpressjournals.org/doi/abs/10.1162/0899766042321751#.WJst02_hCUl
20. Xie, X., Beni, G.: A validity measure for fuzzy clustering. IEEE Trans. Pattern Anal. Mach. Intell. **13**(8), 841–847 (1991). http://dl.acm.org/citation.cfm?id=117682