UDC 004.048

S. Babichev

The Modern State and Perspectives of Clastering Methods Development for High Dimension Data Analysis

Проведен анализ современного состояния методов кластеризации для обработки высокоразмерных данных. Для каждого этапа эксперимента проведен анализ существующих методов и средств осуществления текущей операции, сформулированы преимущества и недостатки данного метода. На основе анализа выделены задачи, решение которых способствует повышению эффективности процесса кластеризации.

The modern state analysis of the clustering methods for high dimension data processing is conducted. The analysis of the existing methods and means of a current operation realization for each experiment phase is carried out, the basic advantages and disadvantages of these methods are formulated. Based on the analysis some tasks were allocated, solution of which promotes the process effectiveness of clustering.

Проведено аналіз сучасного стану методів кластеризації для обробки високорозмірних даних. Для кожного етапу експерименту проведено аналіз існуючих методів і засобів здійснення поточної операції, сформульовано переваги та недоліки даного методу. На підставі аналізу виокремлено задачі, розв'язання яких сприяє підвищенню ефективності процесу кластеризації.

Statement of a problem. Methods of level gene expression estimate with analysis of DNA micromatrix is used in many fields of biomedical searches at the present time. The gene expression analysis allows to identify the studied biological object by set of features that are typical for given objects. This modern technology allows to carry out the quantity analysis of gene expression of ten thousand gens concurrently. Cluster analysis is a method used for decomposition of the set of the studied objects on the subset by identifying the degree their similarity. Object is showing as vector of data or as coordinate of points in multidimension space. There are a lot of decomposition methods of the objects division into clusters, metrics estimate of the proximity level of objects and clusters, criterions of clustering quality. However, depending on the application area, the choice of metric estimation of proximity level of objects and clusters, the criterions of clustering, the established number of clusters, the completeness of knowledge about objects, clustering algorithm, the character of objects decomposition into the clusters may be different.

Analysis of the main achievement and publications

The publication analysis of the present problem shows that majority of clustering methods and algorithms that are used in different fields of the scientific research at the present time, are oriented to a small feature vector dimension of the tested object (no more then 1000) [1–5]. In [6, 7] clustering process is presented as a model that allows to transform all main theory methods of models self-organization at the basis of group accounting of arguments method to cluster analysis theory, namely: multistage search of the best clustering; estimate of clustering quality at basis of criterions set; use of feature space methods and clusters formation; choice of similarity measure between objects, clusters and between objects and clusters.

High dimension (≈ 80000) and high level of noise are peculiarity of data vector which determines the level of gene expression and received by analyze of DNA microarray, what is caused by biological and technological factors that appear at the process of preparation and experiment realization on creation of DNA microarray and reading the information from it [8, 9]. As a result, for increasing clustering accuracy, methods and algorithms of clustering development, oriented to high dimension data, which include the methods of effective filtration and transformation to necessary range, gain the high actuality at the present time.

The unsolved parts of general problem are: absence of universal methods of information fea-

Key words: clustering, high dimension data, clustering algorithm.

ture isolation of high dimension data; imperfection of proximity level estimation metrics of objects and clusters at cluster structure formation; absence of effective tests of quality clustering estimation of high dimension data; absence of effective algorithms or clustering algorithms set for high dimension data processing.

Purpose of article is to analyze the modern works in field of high dimension data clustering and perspectives definition of creation and development of cluster analysis methods and algorithms for high dimension data.

The basic material

The solution of data clustering problem supposes the presence of the following steps:

– the formation of data matrix representing of experimental objects;

– the formation of informative features matrix to define the location of the research object at given metric space;

– selection or definition of the metric that determines the objects proximity level;

– selection of objects clustering algorithm;

– the formation of clusters, interpretation and analysis of the received results.

Block diagram of clustering process have presented at the Fig. 1.

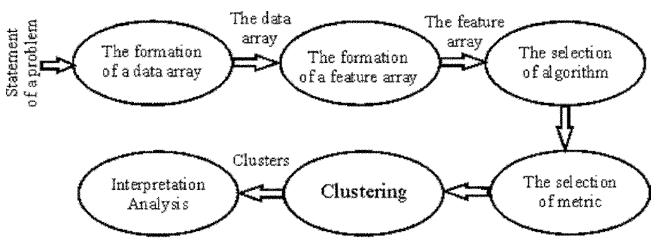

Fig. 1. Block diagram of objects clustering process

The complication of feature space forming problem is related to rapidly increasing of dimension of researching signals during last years, that creates additional problems in the systems of machine learning and data mining [10, 11]. The input data are presented as matrix where rows are objects and columns are features that characterize of consistent object:

$$X = \{x_{ij}\}, \quad i = 1,...,n, \quad j = 1,...,m \qquad (1)$$

n – quantity of objects, m – quantity of feature i-th object.

The choice of feature set is accomplished as result of tapered reflection:

$$\{X^m\} \xrightarrow{F} \{X^k\}, \quad k < m, \qquad (2)$$

where the extremum of quality criterion achieved $J_X(F)$. F at (2) is the functional transforming from set $\{X^m\}$ to set $\{X^k\}$, k – dimension of new feature set. Each reflection (2) conforms to some value of criterion $J_X(F)$. The result of received reflection is function $g(J_X)$. The task is to find such reflection when the extremum this function is achieved. Particularity of this problem is multiextremeness of function $g(J_X)$ and absence of analytical solution in the most case. Therefore the different searched methods for finding of global extremum are used at the present time, the own set of quality criterion is introduced in each of them. The estimate of method work effectiveness is accomplished on the basis of complex criterion.

The method of principal component analysis that allows substantially decrease the dimension of input data with the maximal maintenance of input information has obtained in the biggest extension for this problem solution nowadays [10, 12, 13]. The initial information is presented as matrix of size $(n \times m)$, where n – number of researched objects, m – number of objects features. The output matrix X transforms to matrix $(n \times k)$ by use of principal component method, where k – number of columns, that contains the projections of input data or their remains to corresponding principal components. The model of principal components analysis looks like:

$$z_{ij} = \sum_{\nu=1}^{k} a_{j\nu} f_{i\nu}, \qquad (3)$$

where $i = 1,...,n$ – number of researched objects, k – dimension of new feature space, z_{ij} – j-th feature i-th object, $a_{j\nu}$ – weight coefficient of ν-th principal component of j-th variable, $f_{i\nu}$ – ν-th principal component of i-th object. Each variable is linearly dependent on ν uncorrelated to each

other components f_{iv}, where it is necessary to note that every next component gives the maximal investments to summary dispersion of parameters. The number of the remained components depends on requirement to accuracy by the solution of the given problem, but anyway, the dimension of initial matrix is decreased to size, that allows to process it in real time without additional computing resources. The main disadvantage of principal component method is high sensitivity to the method of data prenormalization. Data distribution in transformed matrix substantially depends on normalization method choice that leads to the different results at the next data processing steps. Moreover, the amount choice of usable principal components is generated separately at each case and accordingly to the experimenter experience, that brings subjectivism to given problem solution.

An alternative of principal component analysis is the factor analysis [14], variations of which are: the method of maximum similarity, method of maximum remainders, method of principal factors, centroid method etc. In contrast to principal component analysis, where the approach of maximum disperse is realized, the factor analysis approximate correlations between the variables, even so, received general factors take into account the dispersion that is peculiar to array of correlation variables, that determines this factor. The main disadvantage of factor analysis is indeterminacy of factor model by high percent of subjectivism at different steps of factor analysis. The finding of factor loading matrix, which would restore latent correlation dependences on the input date with high accuracy, is unsolved problem nowadays. The second disadvantage is the problem of community, i.e. the problem of the variable estimation which is the sum of squares of principal factors loading. Diagonal coefficients of correlation matrix are replaced by the values of generality, after normalization and centering operations before the factors extraction, the condition of maximum maintenance of input useful information which is contained at researches data. The next problem is rotation that lays in the finding coordinate system optimum position, where the projection of received vector data at one axis will be maximal and minimal at others.

The technology of decrease of feature space dimension by self-organizing maps basis is shown at [15], which later got the name: «Kochonen self-organizing map» (SOM). The technology of neuron use with Hebb's adaptation rule of synaptic weight as filter for separation of principal components that has settled as basis of work of Kochonen self-organizing maps is described at [16, 17]. One of the neural network variant that allows to implement this technology is shown at the Fig. 2. The network consists of one layer of neurons, the amount of which equals the amount of principal components and another layer of receptors, where each of them is joined with all neurons of the network. The quantity of receptors equals the input amount or dimension of feature space of researched objects. All neurons of the network output layer are linear. The principle of concurrency learning is based on work of the network. It begins with the initial of synaptic weight of network:

$$w_j(k) = \left(w_{j1}(k), w_{j2}(k), ..., w_{jm}(k)\right), \quad (4)$$

$j = 1, ..., k$ – number of network output.

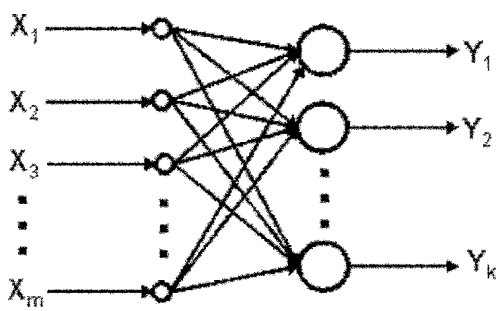

Fig. 2. Forward propagation neural network

The output signal of neuron j determines at entering the input of feature vector n-th object by the next is shown in the following formula:

$$y_j(n) = \sum_{i=1}^{m} w_{ji}(n) x_i(n), \quad (5)$$

In the process of network working an adjustment of synaptic weights is taking place, which connect receptors with calculating nodes of output

layer according to formula of learning Hebba's rule [18]:

$$\Delta w_{ji}(n) = \eta\left[y_j(n) x_i(n) - y_j(n) \sum_{h=1}^{j} w_{hi}(n) y_h(n) \right], \quad (6)$$

$i = 1, \ldots, m$; $j = 1, \ldots, k$, η – the parameter of learning speed of network or iteration step.

Only one neuron that has a name «winner neuron» becomes active at the output of network at the stage of competition when one vector enters the input of network. At [17] it is shown that the first active neuron produces the basis component, and the second – the first component and so on. Thus, each network outputs is represented as a response to concrete own vector of correlation matrix of input vector, the outputs are oriented by decrease of their own values.

An advantage of this technology is automation of principal components isolation that allows to use this method at automatic system of data processing at a stage of their preprocessing. Moreover, the wide spectrum of used neurons transmitting function creates conditions for minimization of useful input information loss by optimal choice of required transmitting function. The disadvantages of this method are the limitation of feature vector dimension that characterized the researched object and necessity of numbers of principal components a priori definition.

The questions of creating and using the distance and proximity levels between objects and clusters are described at works [19–21]. Database that contains the information about gene expression level is the set of numerical values. If $d(a,b)$ – distance between objects a and b at specific metric space, the proximity level of objects is: $s(a,b) \sim \frac{1}{d(a,b)}$. The coefficient of proportionality is calculated by empirical way in each case. The main metrics of objects similarity, characteristics of which are showed by quantitative values, is described at [19]:

– euclidean distance:

$$d_e(a,b) = \left(\sum_{i=1}^{m} (x_{ai} - x_{bi})^2 \right)^{\frac{1}{2}} \quad (7)$$

x_{ai} and x_{bi} – i-th feature, that characterizes objects a and b accordingly. The use this metric is reasonable in case when the researched components are independent to each other, they are isotropic at their physical property and have the same level of significance and same disperse.

– Weighted euclidean distance:

$$d_{ve}(a,b) = \left(\sum_{i=1}^{m} \omega_i (x_{ai} - x_{bi})^2 \right)^{\frac{1}{2}} \quad (8)$$

ω_i – weighting coefficient, that means the level of i-th feature importance. This metric is applied at case if each of the components has certain positive weight depending on defined importance level of this component that has determined a priori.

– Manhattan distance:

$$d_{mh}(a,b) = \sum_{i=1}^{m} |x_{ai} - x_{bi}| \quad (9)$$

It is defined as sum of difference values of comparison objects appropriate features. The weight coefficients are appropriated according to the feature difference at importance level and formula (9) assume the view:

$$d_{mh}(a,b) = \sum_{i=1}^{m} \omega_i |x_{ai} - x_{bi}| \quad (10)$$

– Hamming distance:

$$d_h(a,b) = \sum_{i=1}^{m} |x_{ai} - x_{bi}| \quad (11)$$

x_{ai} and x_{bi} – dichotomous features that take the value 0 or 1 depending on presence or absence of this feature at researched object. It is applied as measure of objects' difference with dichotomous features and it shows the quantity of value distinction of proper features of compared objects.

– Mahalanobis distance:

$$d_{ml}(a,b) = \left((a-b)^T S^{-1} (a-b) \right)^{\frac{1}{2}} \quad (12)$$

S – covariance matrix of sample. The Mahalanobis distance opposed to Euclid distance is recog-

nized by correlations between variables and it is invariant to the scale. The use this metric is reasonable if it is known that the researched objects include the feature with normal distribution and that they belong to one universal set with the same covariance matrix.

The task of the intermediate clusters unification for the formation of general cluster, objects of which have general for intermediate clusters features, is arisen at the final stage cluster analysis. The main metrics of proximity level estimating between clusters that are used in cluster analysis at the present time are shown at [19]. Let the S_i – *i*-th cluster that contains *n* objects, $\bar{X}(S_i)$ – mass center of *i*-th cluster that is calculated as arithmetic mean of objects' position which are in the cluster at selected metric system; $\rho(S_k, S_m)$ – distance between S_k and S_m clusters. The next proximity levels are used to estimate the distance between clusters:

– distance that is measured by «the nearest neighbouring» principles:

$$\rho_{\min}(S_k, S_m) = \min_{x_i \in S_k, x_j \in S_m} d(x_i, x_j); \qquad (13)$$

– distance that is measured by «the longest neighbouring» principles:

$$\rho_{\max}(S_k, S_m) = \max_{x_i \in S_k, x_j \in S_m} d(x_i, x_j); \qquad (14)$$

– distance that is measured by «mass center» principles:

$$\rho(S_k, S_m) = d\left(\bar{X}(S_k), \bar{X}(S_m)\right); \qquad (15)$$

– distance that is measured by «mean connection». It's determined as arithmetic mean of everything pairwise distances between objects, which are in researched clusters:

$$\rho_{cp}(S_k, S_m) = \frac{1}{n_k n_m} \sum_{x_i \in S_k} \sum_{x_j \in S_m} d(x_i, x_j), \qquad (16)$$

n_k і n_m – number of objects, that are in the clusters S_k і S_m accordingly.

It's obvious, that the choice of proper measure depends on the character of the solving task. Moreover the quality of proximity level estimation of clusters is determined by chosen metric of objects closeness definition. Thus, the task of optimization of proximity measures of objects and clusters, with the purpose of maximization the functional quality of clusters division should be solved by increasing of clustering efficiency.

The process of division of researched objects set into clusters may be carried out by set of discriminant functions. The quality functional of division of current objects into clusters is represented as data set of functions, and the best division to extremum corresponds quality functional. At [19] the most prevailing functional of quality division are shown at the present time. Let *n* objects have distributed to *k* clusters at chosen metrics of distance between objects and clusters definition. Then the quality of this division may be estimated by next functional:

– weighted sum of intracluster dispersion:

$$Q_1(S) = \sum_{a=1}^{k} \sum_{x_i \in S_a} d^2\left(x_i, \bar{X}(S_a)\right) \qquad (17)$$

$\bar{X}(S_a)$ – mass center of objects in S_a-th cluster;

– the sum of pairwise intracluster distance between objects:

$$Q_2(S) = \sum_{a=1}^{k} \left(\sum_{x_i, x_j \in S_a} d(x_i, x_j) \right); \qquad (18)$$

– the average of sum of pairwise intracluster distance between objects:

$$Q_3(S) = \frac{1}{k} \sum_{a=1}^{k} \left(\sum_{x_i, x_j \in S_a} d(x_i, x_j) \right); \qquad (19)$$

– generalized intracluster dispersion:

$$Q_4(S) = \det\left(\sum_{a=1}^{k} n_{S_a} W_{S_a} \right), \qquad (20)$$

W_{S_a} – covariance matrix of S_a cluster, elements of which is calculated by formula:

$$\omega_{ch}(S_a) = \frac{1}{n_{S_a}} \sum_{x_i \in S_a} \left(x_i^{(c)} - \bar{X}^{(c)}(S_a)\right)\left(x_i^{(h)} - \bar{X}^{(h)}(S_a)\right), \qquad (21)$$

where $c, h = 1, 2, ..., p$ – dimension of covariance matrix, $x^{(g)}$ – *g*-th component of *i*-th object of S_a cluster, $\bar{X}^{(g)}(S_a)$ – average of *g*-th component of S_a-th cluster.

The choice of appropriate quality criterion of division of objects set into clusters is determined by experimenter depending on *a* type of researched data and character of their distribution in each concrete case. Situation of different criterion contradictoriness is possible during the analysis of high dimension data of compound biological nature that can influence on the quality of clustering process. The task of complex criterion formation that is adapted to tested data appears at this case. The solution of this task will allow to decrease the subjectivism in the process of optimal cluster model choice.

The review of main types of existing cluster structures nowadays is described at [22]. Their block-diagram is shown at the Fig. 3.

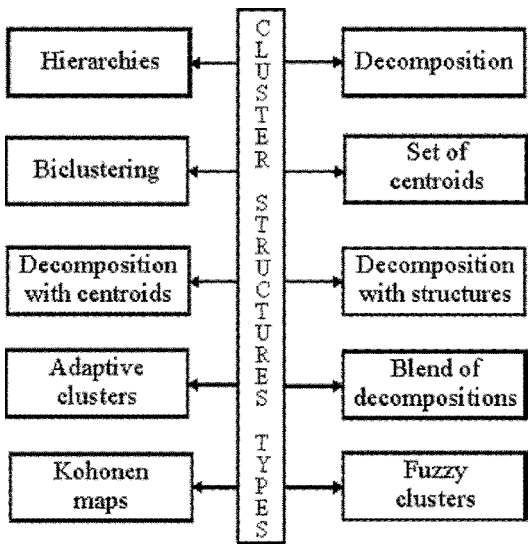

Fig. 3. Blok-diagram of the main types of cluster structures

The set of centroids supposes a prior assignment of finite set of clusters centers, at the same time, each point of assignment metric space is ascribed to one of clusters accordingly to minimum-distance principle. Decomposition is a set of non-empty and non-overlapping clusters. The decomposition with centroids is a structure that contains clusters with their centroids. The objects are grouped into clusters by minimum distance principle. The cluster hierarchy of researched objects set is a set of such nested subset or clusters and their intersection is empty or it is coincided with one of clusters. This type clustering is realized by division of big clusters into smaller or by unification of small clusters into bigger. The adaptive clustering is a set of individual intersection clusters each of them is associated with some positive characteristic – intensity of cluster. Similarity between researched objects is established as a sum of clusters intensity which contains these objects. Decomposition with structures is a variety of hierarchical type of clustering, the result of which is hierarchical graph, each node of which is represented as a subset of initial objects.

Biclustering structure has been developed in the end of the last century and its actual is increasing nowadays [22–25]. The main idea is next: let $Y = \left(y_{ij}\right)$ – data matrix, where $i \in I$ – the set of rows, $j \in J$ – the set of columns. The decomposition into clusters is carried by two decompositions: at set of rows: $S = \left\{S_1, ..., S_K\right\}$; and at set of columns: $T = \left\{T_1, ..., T_V\right\}$ so that each of block $\left\{S_k, T_v\right\}$ is bicluster. Methodic of bicluster structure creating for arbitrary data is shown in [26]. In compliance with this methodic the rows are divided V times for each class T independently on to each other. As a result, the block-structure is received, for creation of which not the components of matrix Y are used, but the first principal component of each feature subset T_v. This approach gives bigger interpretation of principal components by the reason of their unnecessary orthogonality. The advantage of biclustering is a high precision of data processing by reason of maximum accounting of useful information that input data contains. Such accuracy is achieved by consecutive comparison of all rows of initial matrix to appropriate columns that contain the information about features of researched objects. The disadvantage of this method is high complexity of high dimension data processing by reason of a large number of iteration, however, this disadvantages is compensated by better accuracy of data processing by increasing of power and performance of modern computers.

Fuzzy clustering is actual at case of cluster intersection, the grade of membership of object to one or other cluster on the ground of a prior speci-

fied proximity level of objects is estimated in this case. The mix of distribution supposes that each of clusters is presented by single-mode function of objects compactness distribution in cluster. The position of cluster is determined by mean vector.

The self organizing Kohonen maps are oriented to objects visualization in integer lattice, the size of which is determined by user. The main disadvantage of this method is absence of interpretation methods of decomposition to clusters result.

The methods and algorithms that allow to perform one or another clustering are underlaid of appropriate decomposition. Block-diagram of the main methods and algorithms of cluster analysis which are used nowadays is shown at Fig. 4.

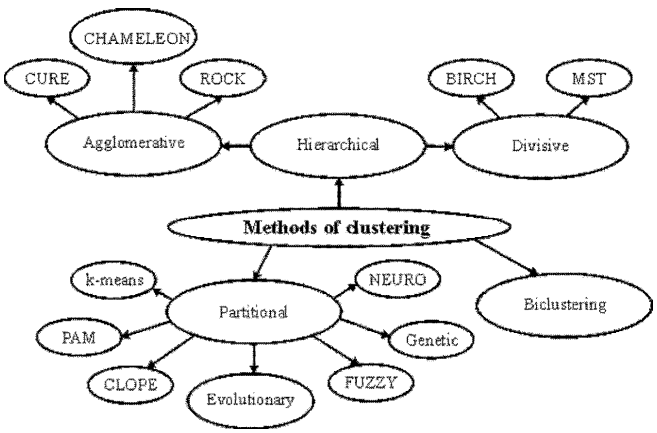

Fig. 4. Block-diagram of the main clustering methods

The methods of hierarchical clustering suppose the successive unification of smaller clusters into bigger (agglomerative) or the successive dividing of big clusters into smaller (divisive). The result of hierarchical methods working is dendrogram, which allows to allocate the quantity of required clusters at defined level of objects decomposition or grouping. A lot of algorithms are created nowadays for agglomerative hierarchical methods realization [27–29].

Algorithm CURE (Clustering Using Representatives) supposes the next steps:

– construction of initial cluster tree that includes each row of input data set;

– compute of distance between clusters using preselected metric of objects proximity level;

– formation of «heap» at mean storage, the clusters are grouped in this case in conformity with increasing of distance from cluster to nearest neighbour;

– confluence of the nearest clusters into one using selected metric of distance between clusters computing;

– if number of clusters is less than it should be, go to step 3.

Algorithm ROCK (Robust Clustering using Links) had been developed for clustering of large volumes of numeric and nominal data. The distinctive feature of this algorithm is presence of new feature for each pair of objects. It is number of new neighbours (references). The algorithm becomes insensitive to overshoots because of references accounting and it does not need an accurate decomposition of objects into clusters. However, computational complexity of process increases sharply at large volumes of data. The proximity between objects is performed by quantity of references to one or other object.

Algorithm Chameleon has been created as modification of algorithms Cure and Rock. It uses the dynamic modeling for hierarchical clustering. The work of algorithm supposes the calculation and comparison interconnection and proximity level of two clusters with concentration of objects distribution in the middle of cluster. Confluence of clusters is performed if calculated proximity levels are less some limited value that it is determined a prior.

An advantage of these methods is the high level of clustering quality even in case of overshoots presence. The agglomerative methods allow selecting clusters of irregular shape and different size. The disadvantage is high percent of subjectivism in choice of limited value and quantity of clusters and in choice of estimation metric of objects proximity level.

Divisive hierarchical methods are described at [30, 31]. Successive decomposition of initial cluster into smaller using predefined proximity metric and limited criterion is performed at the work of this group of methods.

The work of BIRCH algorithm supposes the next steps:

– loading data in a store and creating an initial cluster tree, cluster elements of which have pre-

sented by a triple of numbers (N, LS, SS), where: N – quantity of elements of input data that are in the cluster; LS – sum of elements of input data; SS – square sum of input data elements;

– contraction of data by using of limited coefficient correction to receiving tree of an acceptable size;

– global clustering at leaves components of cluster tree with use of chosen clustering algorithm;

– optimization of obtained distribution, the data between the nearest clusters are distributed in this case, that the nearest data are in the same cluster.

An advantage of BIRCH algorithm is clustering of large volume data by use of limited volumes of the store. At the same time, the fact that data is distributed in the space of inhomogeneously and areas with high density are treated as one cluster. The disadvantages of this method are the necessity of numerical threshold a prior assignment and high error the processing of a wrong form clusters.

The work of MST algorithm supposes the creating of minimum spanning tree that is shown as connected non oriented graph with weights at the lines and decomposition it into clusters, the clusters are divided by arcs with big weights. Advantages of this algorithm are the fact that efficiently assign clusters of any form and it choices an optimal solution from many others. Disadvantage of it is high sensitivity to overshoots.

K-means method is the most popular among partitional clustering methods [32, 33]. The use of this algorithm supposes the next steps:

– choice of k points in given metric space by random way, that are centers of clusters at the initial step;

– distribution of objects to clusters with the nearest center of mass;

– computing of new mass center of obtained clusters;

– if stopping condition of algorithm work is not implemented, go to step 2.

Limited quantity of iteration or clusters stabilization, i.e. absence of objects transition from one cluster into another, are used as criterions of algorithm stopping. The main advantages of this algorithm are the simplicity and quickness of use, intelligibility and transparence. Its disadvantages are high percent of subjectivism at choice of initial number of clusters, high sensitivity to overshoots and low speed of high volumes data processing.

Modification of k-means algorithm is PAM algorithm, the objects redistributed during the work of algorithm relatively to the median, but not relatively to the center mass. It makes it less sensitive to overshoots because of the median is less sensitive to influence of overshoots. The main disadvantages of this algorithm are low speed of large volumes data processing and necessity of a prior determination of cluster quantity.

Algorithm CLOPE is used for clustering of categorical data set large volume [34]. Its main advantages are: high scaling and high speed of work; high quality of clustering by use of global optimization criterion on basis of maximization of gradient high histogram of cluster; absence of necessity of clusters quantity of a prior definition because the automatic selection of clusters quantity is performed during the work of algorithm. This process is regulated by one parameter – coefficient of repulsion. However, it should be noted, that this algorithm is not oriented to numeric sequence, and as a result, the use of it for data set of DNA microarray analysis is not rational.

Kohonen maps are used for searching the regularity in a large data array and to their visualization [35–37]. A typical architecture is one-layer neural network, iteration adjustment of synaptic weight is carried out during its learning for optimal division of set of initial objects into clusters. Advantages of self organizing Kohonen networks are simplicity of their realization and formation of obvious two-dimension reflection of objects set. Disadvantages of this method are the necessity of a prior quantity of clusters definition and absence of interpretation methods of creating model.

The grade of membership of elements to one or another cluster is determined by using fuzzy algorithms of cluster analysis. This type of algorithm is effective in case of clusters intersection, as each of cluster is fuzzy set at this case. Algorithm C-

means is the most popular algorithm of this type [38]. In comparison of *k*-means algorithm, the grade of membership to one or another cluster at the work of *C*-means algorithm are given to the objects, that allows to process the objects which find at the interface between neighbouring clusters. The main disadvantages of this algorithm are computational complexity, necessity of a prior definition of clusters quantity and presence of indefiniteness for objects processing which are distant to centers of all clusters.

The approach about general use of membership and plausibility functions in fuzzy clustering systems of data at ordinal scale is suggested at [39]. Its advantage is more high efficiency of information data processing through the character of processing data distribution accounting. The complex approach about the use of fuzzy logic and Kohonen maps with the use of probabilistic and possibilistic clustering methods for processing of text information is considered at [40, 41], which allows to increase the accuracy of clustering at presence of overshoots and anomalies. However, it should be noted that the use of this technology is problematic for DNA microarray analysis because of high dimension of the vector features space which presents object. Evolutional and genetic algorithms are widely used in the field of data mining and artificial intelligence nowadays [42–44]. Their main advantage is accounting of nature character of data distribution. The concept of populations or chromosomes as a set of different variants of grouping and evolution operators, which are the procedures that allow to receive the chromosome-descendants from chromosome-parents, is used during work of these algorithms. The main disadvantage of these algorithms is their high computer complexity that limits their use for high dimension data.

The ensemble of models which works concurrently is proposed to use for analysis of complex processes at [45]. Each model processes the concrete subset of input information, for which it is adapted, and then, intermediate results are integrated in more high level. An advantage of this approach is decrease of labor intensiveness of information processing due to the absence of necessity of data set integration into one vector of large dimension. Moreover, accuracy of information processing is increased due to the rational use of models, which are in the ensemble. However, it should be noted, that the preliminary researches for adjustment of models at appropriate data, rational division of input set to subsets in accordance with character of experimental information are necessary for rational use of this technology.

Summary

The considerable success has been achieved in the field of development of cluster analysis methods at the last 50 years. The huge amount of methods and algorithms of different type data grouping have been developed, which allows to receive necessary object's grouping depending on character of researched feature space and criterion of decomposition. Methods that are based on fuzzy set theory and genetic algorithms, taking into account a nature character of features distribution that characterize the researched object, with the increase of modern computers power and speed are becoming more popular. The access to databases that characterizes the biologic objects is opening with development of bioinformatics. Each of these objects is DNA sequence of the genes, which are the base of DNA microarray. Levels of genes expression are used as a feature at this case. Distinctive feature of this data is high dimension and high level of noise, that complicate the use of the traditional methods of numeric data clustering. The analysis of modern state of work in this domain object allows to formulate the next unsolved or partially solved problems:

1. Universal methods of information feature space separation are absent in the field of high dimension data preprocessing. The principal components method or factor analysis are used as basic methods nowadays, which have the high sensitivity to quality of data normalization and filtration, that doesn't allows to get the high quality of objects clustering.

2. Estimate criterions of proximity level of objects and clusters for biological data sequence require adjustments, as in high dimension feature space their efficiency is decreased. In this case, the actual task is a development of relative criteri-

ons or complex criterion, taking into account different metrics that are existed nowadays.

3. Substantiation criterions of quality of analysis results for objects with high dimension of feature space require an essential revision because of the estimate of the grouping process has a subjective character nowadays. The same data set may be classified differently depending on application field, completeness of knowledge about objects, etc. Therefore, the necessity of development of appropriate criterions of grouping quality such objects appears.

4. Necessity of development of methods ensembles and algorithms, each of which is oriented on the processing of specified subset of input data set appears for increasing of clustering quality of complex object's nature. Complex parallel use of different methods increases the work content on the one hand, but this disadvantage is compensated by better accuracy of objects clustering, considering the high speed of computer engineering development.

5. Biclustering methods didn't get the wide use in systems of analysis of numeric data high dimension nowadays, because of high work content of information processing. Therefore, development of clustering methods of objects on the base complex use of effective methods of data preprocessing, considering methods of decreasing of feature space dimension and biclustering algorithms is an actual task nowadays.

1. *Data* Analysis of Bio-Medical Data Mining using Enhanced Hierarchical Agglomerative Clustering / J.V. Krishnaiah, D.V. Chandra Sekar, K. Ramchand et al. // J. of Engin. and Innov. Technol. – 2012. – **2**, Issue 3. – P. 43–49.
2. *Liang J.*, *Kachalo S.* Computational analysis of microarray gene expression profiles: clustering, classification, and beyond // Chemometrics and Intelligent Laboratory Syst. – 2002. – N 62. – P. 199–216.
3. *Rezankova H.* Cluster analysis of economic data // Statistica. – 2014. – N 94(1). – P. 73–86.
4. *Li Y.*, *Chung S.M.*, *Holt J.D.* Text document clustering based on frequent word meaning sequences // Data & Knowl. Engin. – 2008. – N 64(1). – P. 381–404.
5. *Jain A.K.*, *Murty M.N.*, *Flynn P.J.* Data clustering: A review // ACM Computing Surveys. – 1999. – **31**, N 3. – P. 264–323.
6. *Ивахненко А.Г.* Объективная кластеризация на основе теории самоорганизации моделей // Автоматика. – 1987. – № 5. – С. 6–15.
7. *Ивахненко А.Г.* Алгоритмы метода группового учета аргументов (МГУА) при непрерывных и бинарных признаках // Ин-т кибернетики им. Глушкова. – Киев, 1992. – 49 с.
8. *Ивахно С.С.*, *Корнелюк А.И.*, *Минцер О.П.* Методы кластеризации в программе *Microarraytool* для анализа данных ДНК-микроарреев // Медична інформатика та інженерія. – 2008. – № 3. – С. 33–40.
9. *Івахно С.С.*, *Корнелюк О.І.* Мікроареї: Огляд технологій та аналіз даних // Укр. біохім. журн. – 2004.– № 2(76). – С. 5–19.
10. *Huan Liu*, *Hiroshi Motoda*. Computational Methods of Feature Selection // Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. – New York, 2007. – 419 p.
11. *Feature* selection for clustering – a filter solution / D. Manoranjan, C. Kiseok, P. Scheuermann et al. // Proc. of the Second Int. Conf. on Data Mining. – Washington, USA, 2002. – P. 115–122.
12. *Principal* Manifolds for Data Visualization and Dimension Reduction / A. Gorban, B. Kegl, D. Wunsch et al. // Series: Lecture Notes in Computational Science and Engineering. – Berlin – Heidelberg – New York. – 2008. – **580**. – 340 p.
13. *Roth V.*, *Lange T.* Feature selection in clustering problems. Advances in neural information processing systems. – Massachusetts Institute of Technology, 2004. – P. 473–481.
14. *Хохлов В.В.* Многомерный факторный анализ временных рядов банковских депозитов. – Севастополь: Изд-во СевНТУ, 2009. – 204 с.
15. *Kochonen T.* Self-Organizing Maps: Third, extended edition. – Berlin: Springer-Verlag, 2001. – 501p.
16. *Oja E.* A simplified neuron model as a principal component analyzer // Mathematical Biology. – 1982. – **15**. – P. 267–273.
17. *Хайкин С.* Нейронные сети. – Москва – Санк-Петербург – Киев, 2006. – С. 509–621.
18. *Sanger T.D.* Optimal unsupervised learning in a single-layer linear feedforwards neural network // Neural Networks. – 1989. – N 12. – P. 459–473.
19. *Прикладная* статистика. Классификация и снижение размерности / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков и др. – М.: Финансы и статистика, 1989. – 608 с.
20. *Раушенбах Г.В.* Об измерении близости между множествами в задачах анализа данных // Программно-алгоритмическое обеспечение анализа данных в медико-биологических исследованиях. – М.: Наука, 1987. – С. 41–54.
21. *Стрижов В.В.*, *Кузнецов М.П.*, *Рудаков К.В.* Метрическая кластеризация последовательностей аминокислотных остатков в ранговых шкалах // Мате-

матическая биология и биоинформатика. – 2012. – Т. 7, № 1. – С. 345–359.
22. *Миркин Б.Г.* Методы кластер-анализа для поддержки принятия решений: Обзор. – М.: Высш. шк. экономики, 2011. – 88 с.
23. *Enhanced* biclustering on expression data / J. Yang, W. Wang, H. Wang et al. // The 3th IEEE Conf. on Bioinform. and Bioeng. – 2003. – P. 321–327.
24. *Tanay A*., *Sharan R*., *Shamir R*. Discovering statistically significant biclusters in gene expression data // Bioinformatics. – 2002. – **18**. – P. 136–144.
25. *Spectral* biclustering of microarray data: coclustering genes and conditions / Y. Klugar, R. Basri, J.T. Chang et al. // Genome Research. – 2003. – **13**. – P. 703–716.
26. *Лингвистический* подход к задаче обработки больших массивов информации / Э.М. Браверманн, А.А. Дорофеюк, И.Б. Мучник и др. // Автоматика и телемеханика. – 1974. – № 11. – С. 73–88.
27. *Sudipto G*., *Rajeev R*., *Kyuseok S*. Cure: An efficient clustering algorithm for large databases // Inform. Syst. – 2001.– **26**, N 1. – P. 35–58.
28. *Sudipto G*., *Rajeev R*., *Kyuseok S*. Rock: A robust clustering algorithm for categorical attributes // Ibid. – 2000. – **25**, N 5. – P. 345–366.
29. *Ляховец А.В*. Исследование результатов применения модифицированного алгоритма Хамелеон в области лечения поясничного стеноза // Вост.-Европ. ж. передовых технол. – 2012. – № 3(11). – С. 13–16.
30. *Zhang T*., *Ramakrishnan R*., *Livny M*. Birch: An efficient data clustering method for very large databases // SIGMOD '96 Proc. of the int. conf. on Management of data. – Montreal, Canada. – 1996. – P. 103–114.
31. *Damodar R.E*., *Prasanta K.J.* Minimum Spanning Tree Based Clustering Using Partitional Approach // Proc. of the Int. Conf. on Frontiers of Intel. Comp.: Theory and Applications (FICTA). – 2013. – **199**. – P. 237–244.
32. *An Efficient* k-Means Clustering Algorithm: Analysis and Implementation / T. Kanungo, M. David, D. Mount et al. // IEEE translation on pattern analysis and machine intelligence. – 2002. – **24**, N 7. – P. 881–892.
33. *Ding C*., *He X*. K-means Clustering via Principal Component Analysis // Proc. of the 26 int. conf. on Machine learning. – New York, USA. – 2004. – P. 29–37.
34. *Yang Y*., *Guan H*., *You J*. CLOPE: A fast and Effective Clustering Algorithm for Transactional Data // 8th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining. – Edmonton, Canada. – 2002. – P. 682–687.
35. *Бодянский Е.В*., *Волкова В.В*., *Колчигин Б.В*. Самообучающаяся нейро-фаззи система для адаптивной кластеризации текстовых документов // Бионика интеллекта. – 2009. – № 1(70). – С. 34–38.
36. *Манжула В.Г*., *Федяшов Д.С*. Нейронные сети Кохонена и нечеткие нейронные сети в интеллектуальном анализе данных // Фундаментальные исследования. – 2011. – № 4. – С. 108–115.
37. *Бодянский Е.В*., *Колчигин Б.В*. Адаптивная нейрофаззи сеть Кохонена типа 2 // Материалы междунар. научн. конф. «Автоматизация: проблемы, идеи, решения», Севастополь: СевНТУ, 2010. –Т. 1. – С. 135–137.
38. *Yambal M*., *Gupta H*. Image Segmentation using Fuzzy C Means Clustering: A survey. // Int. J. of Advanced Research in Comp. and Commun. Eng. – 2013. – **2**, Issue 7. – P. 2927–2929.
39. *Бодянский Е.В*., *Самитова В.А*. Нечеткая кластеризация данных в порядковой шкале на основе совместного использования функций принадлежности и правдоподобия: Зб. наук. праць Харків. ун-ту повітряних сил. – 2010. – № 3(25). – С. 91–95.
40. *Bodyanskiy Ye.V*., *Kolchygin B.V*., *Pliss I. P*. Adaptive neuro-fuzzy Kohonen network with variable fuzzifier // Inform. Theories and Appl. – 2011. – **18**, N 3. – P. 215–223.
41. *Бодянский Е.В*., *Колчигин Б.В*., *Волкова В.В*. Нечеткая возможностная кластеризация текстовых документов на основе самоорганизующихся карт Кохонена // Тез. докл. II междунар. науч.-практ. конф. «Математическое и программное обеспечение интеллектуальных систем (*MPZIS*–2008)». – Днепропетровск: ДНУ, 2008. – С. 47–48.
42. *Демидова Л.А*., *Коняева Е.И*. Кластеризация объектов с использованием *FCM*-алгоритма на основе нечетких множеств второго типа и генетического алгоритма // Вестн. РГРТУ. – 2008. – № 4 (26). – С. 46–54.
43. *Литвиненко В.И*. Кластерный анализ данных на основе модифицированной иммунной сети. // УСиМ. – 2009. – № 1. – С. 54–61.
44. *Литвиненко В.И*. Анализ применимости адаптированной иммунной сети для решения задач кластеризации спиральных структур // Зб. наук. праць «Моделювання та керування станом еколого-економічних систем регіону». – 2008. – **4**. – С. 145–155.
45. *Dietterich T.G*. Ensemble Methods in Machine Learning // First Int. Workshop «Multiple Classifier Systems (MCS 2000)», Cagliari, Italy. – 2000. – **1857**. – P. 1–15.

Поступила 08.12.2014
E-mail: *bsa*63*@mail.ru*