UDK 004.048

S. [1]Babichev[1], V. Lytvynenko[2], M. A. Taif[2], A. Sharko[2]
[1]Jan Evangelista Purkinje University in Usti nad Labem, Czech Republic
[2]Kherson National Technical University, Kherson, Ukraine

HYBRID MODEL OF INDUCTIVE CLUSTERING SYSTEM OF HIGH-DIMENSIONAL DATA BASED ON THE SOTA ALGORITHM

Подано модель системи кластеризації високорозмірних даних на основі комплексного використання самоорганізуючого алгоритму COTA і методів індуктивного моделювання складних систем. Якість кластеризації оцінюється на двох рівнопотужних підмножинах з використанням комплексного критерію балансу, у якому враховані як зміщення центрів мас відповідних кластерів різних підмножин, так і розподіл об'єктів у відповідних кластерах відносно центра мас. Для кластеризації об'єктів на кожній з підмножин запропоновано використовувати алгоритм COTA, що являє собою тип самоорганізуючих нейронних мереж на основі карт Кохонена і алгоритму вирощування просторової клітинної структури Fritzke.

Ключові слова: кластеризація, індуктивне моделювання, алгоритм COTA, критерій балансу, високорозмірні дані.

Model of high-dimensional data clustering system based on the complex use of Self-organizing SOTA algorithm and inductive modeling methods of complex systems is presented in the article. The quality of clustering is evaluated at two equal power subsets with the use of complex balance criterion, which takes into account both the displacement the mass centers of the appropriate clusters of different subsets and distribution of objects in the appropriate clusters relative to the mass center. The SOTA algorithm, which is a type of Self-organizing neural networks based on Kohonen maps and algorithm of spatial cell structure of Fritzke growing, is proposed to use for the clustering of objects in each of the subsets.

Key words: clustering, inductive modeling, SOTA algorithm, criterion of balance, high-dimensional data.

Problem statement

Actuality of the problem is determined by the modern state of works in the field of high-dimensional data of complex nature processing for purpose of create of clustering and classification systems of researched objects, and prediction systems of their further functioning. As high-dimensional will understand data, the dimension of the feature space which is equal or more of the number of researched objects. Such data are the gene expression profiles of DNA nucleotides, encephalograms of a biological organism, the chromatogram of narcotic substances, end so on. Peculiarities of investigated data are a high level and specificity of the noise component due to the biological processes running in the studied object, and the imperfection of the system of data formation for further processing, and high dimensional of feature space. Traditional algorithms of data clustering in the case of complex nature data processing are inefficient due to the high error of a final result obtain. Clustering method based on Self-organizing models that use new topologies and different learning strategies are widely used recently. SOTA algorithm (Self-Organizing Tree Algorithm) [1] represents a type of Self-organizing neural networks based on the Kohonen maps and the algorithm of growing of Fritzke spatial cell structure [2]. As opposed to the Kohonen maps that reflect a set of input data of high dimensionality on the elements of a two-dimensional array of small dimension, the SOTA algorithm generates a binary topological tree. The algorithm Fritzke assumes the self-organization of output nodes of the network so that the number of nodes increases in the area with a higher density of objects concentration and decreases in the area with a lower density. Thus, the heterogeneity of objects distribution in space is taken into account. However, it should be noted that the

high quality of clustering on one data sample does not guarantee a similar result for another sample of similar data using the same algorithm of clustering. Increasing of clustering objectivity is possible by the use of inductive methods of complex systems modeling in which the data processing is carried out in parallel on two equal power subsets and the final decision is taken on the basis of external criterion balance of clustering results on two subsets.

Analysis of recent research and publications

The basic foundations of inductive method of complex systems self-organization models on the basis of Group Method of Data Handling (GMDH) are set forth in the works [3-5]. Further development of this theory is reflected in [6-8]. The concept of the objective cluster analysis is presented in [9] and was further developed in [10]. However, it should be noted that investigations of the authors are focused mainly at 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 inefficiently. The objective clustering algorithm which guesses the results of clustering comparing at two equal power subsets at different hierarchical levels of objects clustering is proposed in [9]. Modification of this algorithm is implemented to solve various practical problems nowadays. However, it should be noted that an objective clustering model based on analysis systems of clustering has not practical realization at present.

Unsolved parts of the general problem are the absence of complex use of effective models of high dimensional noisy data clustering algorithms and methods of inductive modeling complex systems.

The Aim of the article is the development of hybrid model of high-dimensional data clustering based on complex use of inductive modeling methods of complex systems and Self-organizing clustering SOTA algorithm.

The presentation the basis material

Let $A=\{x_{ij}\}, i=1...n,\ j=1...m$ – is the matrix of features of researches objects, where n – is the number of observed objects, m – is the number of features that characterize an object. The goal of clustering is a partition of objects into non-empty subsets of pairwise intersecting clusters in accordance with the criteria of remoteness object and cluster, taking into account the properties of the objects [10]:

$$K=\{K_1, K_2,..., K_k\}, \quad 1\le k\le n$$
$$K_1\cup K_2\cup ...\cup K_k = A, \quad K_i\cap K_j=\varnothing, \quad i\ne j, \quad i,j=1,2,...,k \tag{1}$$

An objective clustering model based of method of complex systems inductive modeling is selected using a minimum of external criterion of balance characterizing the quality of the corresponding model clustering on two the same power subsets.

Formally the objective clustering model can be represented as follows:

$$M:\left\{R(K)\,|\,e\le e_0 \xrightarrow{\{CQ\}} opt\right\} \tag{2}$$

where $R(K)$ – is the result of clustering, e – is the error of clustering on training and test simples, e_0 – is the threshold error of clustering, CQ – is the balance criterion of clustering quality assessment. At this work is proposed to use the complex balance criterion of clustering quality that takes into account the displacement of the mass center of corresponding clusters at the clustering of two equal power subsets and character of the objects distribution at the corresponding cluster relative to mass. The mass center of objects in k cluster can be found as follows: $c_k=\frac{1}{n\cdot m}\sum_{i=1}^{n}\sum_{j=1}^{m}x_{ij}$, where n – is the number of objects in cluster k, m – is the number of features that characterize an object. As a measure of the objects distribution relative to the mass center you can use the sum of squared distances from the mass center of the object to the mass center of the corresponding cluster:

$$D_j=\sum_{i=1}^{n} d^2\left(\bar{x}_i, c_j\right) \tag{3}$$

where $d(\cdot)$ – is a measure of similarity between mass centers of corresponding object and cluster. Then the balance criterion based on the formula (3) can be represented as follows:

$$CQ_1(Q,R)=\sum_{j=1}^{q}\left(\sum_{i=1}^{n_1}d^2\left(\bar{x}_i,c_j(Q)\right)-\sum_{i=1}^{n_2}d^2\left(\bar{x}_i,c_j(R)\right)\right)\to\min \tag{4}$$

where q – is the number of clusters, n_1 and n_2 – are the numbers of objects in clusters $j=1,2,...,q$ of clustering Q and R respectively, $c_j(Q)$ and $c_j(R)$ – are the mass centers of cluster j of clustering Q and R.

Another component of the balance criterion assumes that in the case of the objective clustering the value of the squared difference of distances between mass centers of corresponding clusters in the various clustering should be minimal [8]:

$$CQ_2(Q,R)=\sum_{j=1}^{q}\left(c_j(Q)-c_j(R)\right)^2\to\min \tag{5}$$

Normalization of these criteria limits the range of their changes from 0 to 1, and the formulas (4) and (5) take the form:

$$CQN_1(Q,R)=\frac{\sum_{j=1}^{q}\left(\sum_{i=1}^{n_1}d^2\left(\bar{x}_i,c_j(Q)\right)-\sum_{i=1}^{n_2}d^2\left(\bar{x}_i,c_j(R)\right)\right)}{\sum_{j=1}^{q}\left(\sum_{i=1}^{n_1}d^2\left(\bar{x}_i,c_j(Q)\right)+\sum_{i=1}^{n_2}d^2\left(\bar{x}_i,c_j(R)\right)\right)}\to\min \tag{6}$$

$$CQN_2(Q,R)=\frac{\sum_{j=1}^{q}\left(c_j(Q)-c_j(R)\right)^2}{\sum_{j=1}^{q}\left(c_j(Q)+c_j(R)\right)^2}\to\min \tag{7}$$

Each of these criteria has its advantages and disadvantages. The first criterion takes into account the nature of the objects distribution in the cluster relative to the corresponding mass center, but not taken into account displacement of the mass center of corresponding clusters for different clustering. The second criterion compensates this imperfection, but it does not take into account the nature of the objects distribution in clusters. Complex criterion of balance combines the advantages of the two criteria and therefore takes into account their disadvantages. It can be calculated as the average of the criteria (6) and (7):

$$CCQ(Q,R)=\frac{1}{2}\left(CQN_1(Q,R)+CQN_2(Q,R)\right)\to\min \tag{8}$$

Algorithm of the original set of objects Ω division to 2 equal power non-intersecting subsets Ω^A and Ω^B consists the following steps [9,10]:

1. Calculation of $n\cdot(n-1)/2$ pairwise distances between objects in the original sample of data;
2. allocation of pairs of objects X_s, X_p, the distance between which is minimal:

$$d\left(X_s,X_p\right)=\min_{i,j}d\left(X_i,X_j\right) \tag{9}$$

3. distribution of the object X_s to subset Ω^A, and the object X_p to subset Ω^B;
4. repeat steps 2 and 3 for the remaining objects. If the number of objects is odd, the last object is distributed to the two subsets.

During the SOTA-algorithm running, the sequence of binary tree nodes is adapted to the characteristics of the feature space of the input data set. Herewith, the number of output nodes in the model fitting process is

determined by varying of the input data of feature space. As a measure of the vectors similarity was used the correlation or Euclidean metric, determined in accordance with formulas (10) и (11):

$$d(X_s, X_p) = \sqrt{\sum_{i=1}^{m} (x_{si} - x_{pi})^2} \tag{10}$$

$$d(X_s, X_p) = (1 - r) = 1 - \frac{\sum_{i=1}^{m} \left((x_{si} - \bar{x}_s) \cdot (x_{pi} - \bar{x}_p)\right)}{\sqrt{\sum_{i=1}^{n} (x_{si} - \bar{x}_s)^2 \cdot \sum_{i=1}^{n} (x_{pi} - \bar{x}_p)^2}} \tag{11}$$

where r – is the Pearson's correlation coefficient, $\bar{x}$ – is the average value of the corresponding vector characteristics.

In the initial state, the system is composed of two cell units pooled by the external root node, i.e. it has the structure of a binary tree (Fig. 1a). Each node is characterized by a vector of features, the number of which is equal to the dimension of the studied genes feature space. The value of each feature in a column of the vector defines the conditions under which the measurement of expression of the corresponding gene.

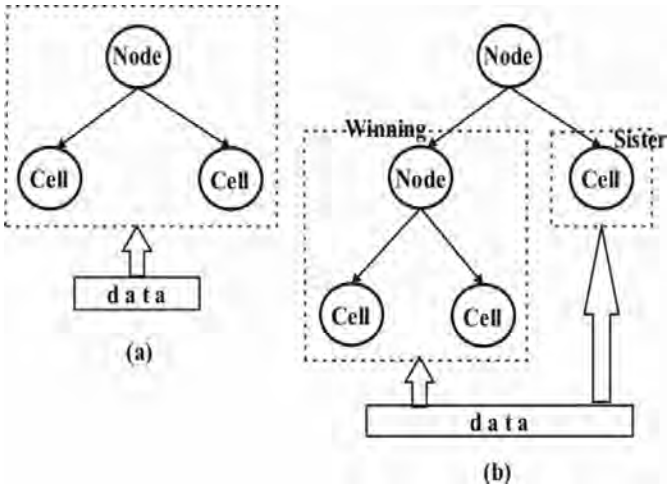

Fig. 1. The cell formation structure by the algorithm SOTA:
a – the initial state of the system; b – the status of the system after one cycle

The algorithm SOTA consists of the following stages:

1. *initialization.* For features of the root node vectors and cells are assigned the weights which values are equal to the average features value of the all studied data columns. The length of weights vector is equal to the dimension of studied data feature space;

2. *adaptation.* During algorithm operation the input of the all external cells is sequentially supplied by feature vector of studied objects. Then the degree of closeness of this vector with weights the cells is calculated by formulas (10) or (11). In accordance with the principle of "winner takes all" is allocated the cell-winner, which vector of weights has the smallest distance from the investigated gene profiles vector. The weights of the cell-winner and its vicinity are adjusted in accordance with the formula:

$$C_i(\tau + 1) = C_i(\tau) + \eta \cdot (P_j - C_i(\tau)) \tag{12}$$

where $C_i(\tau)$ and $C_i(\tau + 1)$ – are the weight vectors of cells i at step τ and τ+*1* respectively, P_j – is the profiles vector j-th gene on the input system, η – is the parameter that determines the adjusting step of the cell-winner weights which at iteration t is defined as [1]:

$$\eta_t = \alpha \cdot \frac{1 - t}{n} \cdot (1 - b\tau) \tag{13}$$

where t – is the total number of objects, n – is the maximum number of studied objects, τ – is the number of operations per cycle, b – coefficient that determines the rate of change of the parameter η, α – is the parameter determined by empirically proceeding from a condition: $\alpha_w > \alpha_m > \alpha_s$, where α_w, α_m and α_s – are the coefficients for adjusting weights of the cell-winner that binds the node with its neighbor cell respectively. The values of α and b for neighboring cells and the node are selected empirically;

3. the convergence of the algorithm and the network formation. To determine the clusterization tree structure, the variation coefficient of each cell as the arithmetic average value between the values of the cell weights and gene expression profiles values in this cell is calculated as:

$$R_i = \frac{1}{q}\sum_{j=1}^{q} d\left(P_j, C_i\right) \tag{14}$$

where q – is the number of objects in cell i, C_i – is the weight vector of cell i. The total value of the variation coefficient is defined as the sum of the variation coefficients for all external cells:

$$\varepsilon_r = \sum_{i=1}^{s} R_i \tag{15}$$

where s – is the number of external cells for r-th clustering. The assessment criterion of the algorithm convergence is the relative change in the total coefficient of variability:

$$\left|\frac{\varepsilon_r - \varepsilon_{r-1}}{\varepsilon_{r-1}}\right| < E \tag{16}$$

where E is the threshold value of variability criterion. The cycle ends if the condition is (16).

The further growth of the tree begins with the cell having the largest value of the coefficient of variation. This cell is divided into two parts and becomes a node (Fig. 1b). The weighting values of daughter cells and a node are identical to each other. The growth of network is finishes when the total value of the variation coefficient reaches a certain threshold value. At the zero value the threshold coefficient the number of clusters equal to the number of studied objects.

The scheme of the inductive cluster analysis model based on the SOTA algorithm is shown in Fig. 2. The implementation of this model guesses the next steps:

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

step 2. Division of set Ω into two equal power subsets in accordance with hereinbefore algorithm. These subsets Ω^A and Ω^B can be formally represented as follows:

$$\begin{aligned} &\Omega^A = \left\{x_{ij}^A\right\},\ \Omega^B = \left\{x_{ij}^B\right\},\ j = 1,\ldots,m \\ &i = 1,\ldots,n_A = n_B,\ n_A + n_B = n \end{aligned} \tag{17}$$

step 3. Setup of clustering procedure using of SOTA algorithm. Setting the initial parameters of the algorithm: $\alpha_w > \alpha_m > \alpha_s$, b, E. Assigning of the weights to root node and cells in accordance with step 1 of the SOTA algorithm. The calculation of the initial value of the variation coefficient for subsets Ω^A and Ω^B. Assigning of a maximum number n of clusterization;

step 4. Clustering of objects Ω at subsets Ω^A and Ω^B by SOTA algorithm. Inductive multiple-row clustering procedure is shown in Fig. 2 and can be represented as follows:

1-th row of selection:

4.1.1. successive giving of objects of subsets Ω^A and Ω^B to the input of external cells, weights of the cells and the root node adjustment according to the formulas (12) и (13);

4.1.2. calculation of the external cells variation coefficient by the formulas (14) and (15);

4.1.3. determining of the variation coefficient relative change for the last two clusterization. When the condition (16) is true, forming a cluster structure for subsets Ω^A and Ω^B. Otherwise, the cycle is repeated until the condition (16) will be true;

4.1.4. calculation of balance complex criterion by formulas (6), (7) and (8) for obtained clustering on fulfillment of condition $k_r^A = k_r^B$ (r – is the number of clustering, k_r^A and k_r^B – are the number of clusters in r-th clustering for subsets Ω^A and Ω^B respectively) ;

4.1.5. comparison of the obtained criteria for clustering r and r-1.

2-th row of selection:

4.2.1. division of the external cell having the largest value of the variation coefficient in two parts. Assignment to new cells of weight coefficient values of initial cell;

4.2.2. implementation of the steps 4.1.1-4.1.5 of this algorithm for new cells;

3-th and the next rows of selection: implementation of steps 4.2.1 and 4.2.2 and fixation of clustering k_r^A and k_r^B for each stage.

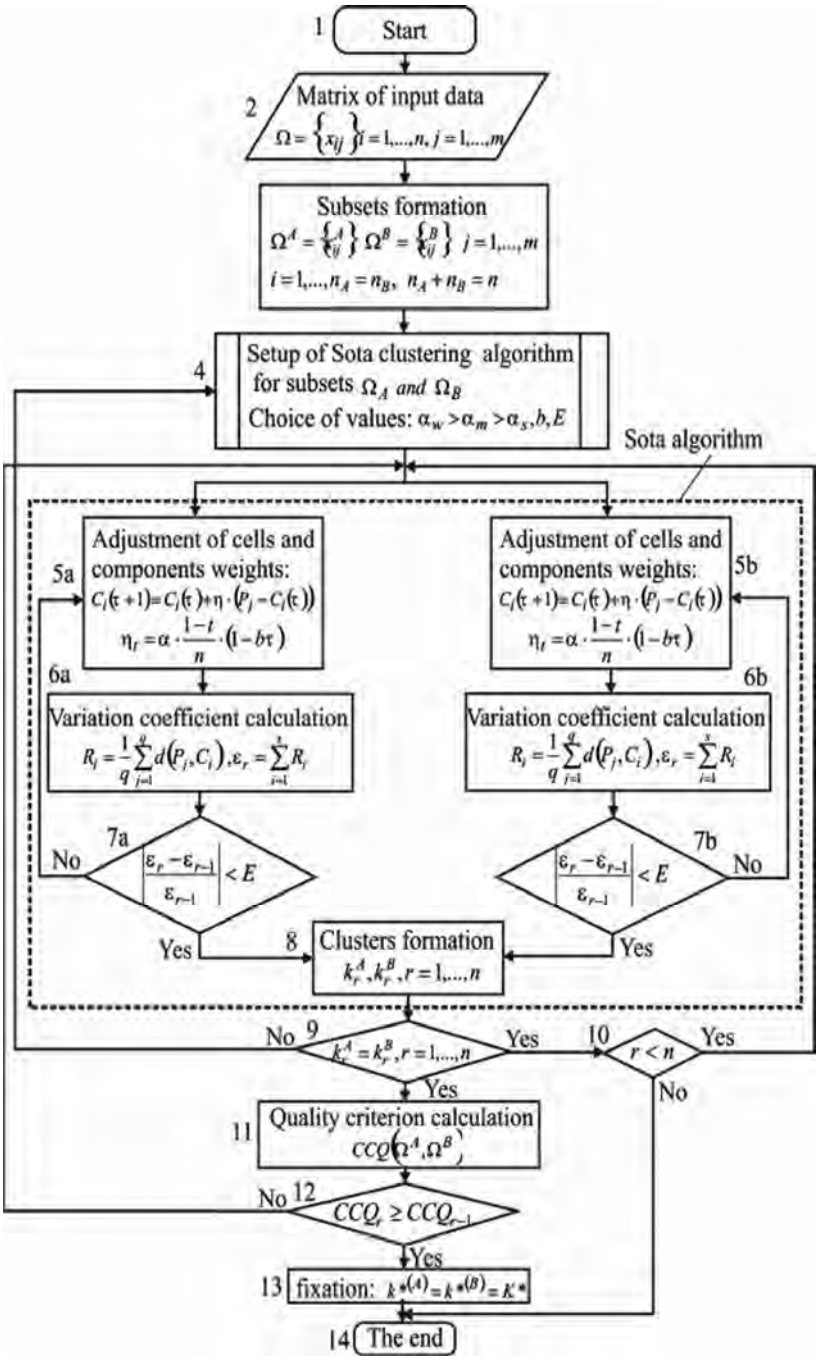

Fig. 2. Scheme of the inductive cluster analysis model based on the SOTA algorithm

The minimum of a complex criterion balance achieving is a condition of stopping algorithm, whereupon it is fixed the clustering that corresponds to extreme value of this criterion.

Conclusions

The hybrid model of induction clustering system of high-dimensional data based on of complex use of Self-organizing SOTA algorithm and complex systems inductive modeling methods is presented in the article. To evaluate the quality of clustering at the initial stage the initial set of the objects is divided into two equal power subsets by the dipole algorithm of objects division. The complex criterion that takes into account both the displacement of the mass center of corresponding of clusters for different clustering and objects distribution in clusters relative to the mass center is proposed to use as a external balance criterion. The minimum of balance complex criterion corresponds to the optimal clustering at this case. Multi rows clustering of equal power subsets was performed using the SOTA algorithm that is a type of Self-organizing neural networks based on Kohonen maps and algorithm of spatial cellular structure of Fritzke growing. As a result of work, it is developed a structural block diagram of this model, which shows the stages of information processing for complex nature data optimal clustering choice. The practical implementation of the proposed model and evaluation of its work effectiveness for complex data at different noise-to-signal ratio are further perspectives of the authors work.

1. Dorazo J., Carazo J.M. Phylogenetic reconstruction using an unsupervised growing neural network that adopts the topology of a phylogenetic tree // Journal of Molecular Evolution, 1997. – No. 44(2).– P. 226–259. 2. Fritzke B. Growing Cell Structures A Self-Organizing Network for Unsupervised and Supervised Learning // Neural Networks, 1994. – Vol. 7, No. 9. – P. 1441–1460. 3. Ivakhnenko A. G. Group method of data handling – a competitor of stochastic approximation method // Automatics, 1968. – No. 3. – P. 58–72. [In Ukraine]. 4. Ivakhnenko A. G. Inductive method for self-organization of complex systems models. – Kiev: Scientific Thought, 1982. – 296 p. [In Russian]. 5. Ivakhnenko A. G. Objective self-organization based on the theory of self-organization models // Automatics, 1987. – No. 5. – P. 6–15. [In Russian]. 6. Stepashko V. S. Theoretical aspects of GMDH as a method of inductive modeling // Managing Systems and Machines, 2003. – No. 2. – P. 31–38. [In Russian]. 7. Stepashko V. S. Elements of the inductive modeling theory / State and prospects of informatics development in Ukraine: Monograph / Team of authors. – Kiev: Scientific Thought, 2010. – 1008 p. – P. 471–486. [In Ukraine]. 8. 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. – No. 1. – P. 83–97. [In Ukraine]. 9. Madala H. R., Ivakhnenko A. G. Inductive Learning Algorithms for Complex Systems Modeling. – CRC Press, 1994. – 365 p. 10. Sarycheva L. V. Objective cluster analysis of the data on the basis of the Group Method of Data Handling // Problem of Management and Informatics, 2008. – № 2. – P. 86–104. [In Russian].