e-ISSN 1607-3274 Радіоелектроніка, інформатика, управління. 2019. № 1
p-ISSN 2313-688X Radio Electronics, Computer Science, Control. 2019. № 1

НЕЙРОІНФОРМАТИКА ТА ІНТЕЛЕКТУАЛЬНІ СИСТЕМИ

НЕЙРОИНФОРМАТИКА И ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ

NEUROINFORMATICS AND INTELLIGENT SYSTEMS

UDC 004.048

IMPLEMENTATION OF DBSCAN CLUSTERING ALGORITHM WITHIN THE FRAMEWORK OF THE OBJECTIVE CLUSTERING INDUCTIVE TECHNOLOGY BASED ON R AND KNIME TOOLS

Babichev S. – PhD, Associate Professor, Associate Professor of the Department of Informatics, Jan Evangelista Purkyne University in Usti nad Labem, Czech Republic. Docent of the Department of Information Technologies, IT Step University, Lviv, Ukraine.

Vyshemyrska S. – PhD, Associate Professor, Associate Professor of the Department of Informatics and Computer Science, Kherson National Technical University, Kherson, Ukraine.

Lytvynenko V. – Dr.Sc., Professor, Head of the Department of Informatics and Computer Science, Kherson National Technical University, Kherson, Ukraine.

ABSTRACT

Context. The problem of the data clustering within the framework of the objective clustering inductive technology is considered. Practical implementation of the obtained hybrid model based on the complex use of R and KNIME tools is performed. The object of the study is the hybrid model of the data clustering based on the complex use of both DBSCAN clustering algorithm and the objective clustering inductive technology.

Objective. The aim of the work is the creation of the hybrid model of the objective clustering based on DBSCAN clustering algorithm and its practical implementation on the basis of the complex use of both R and KNIME tools.

Method. The inductive methods of complex systems modelling have been used as the basis to determine the optimal parameters of DBSCAN clustering algorithm within the framework of the objective clustering inductive technology. The practical implementation of this technology involves: the use of two equal power subsets, which contain the same quantity of pairwise similar objects; calculation of the internal and the external clustering quality criteria; calculation of the complex balance criterion, maximum value of which corresponds to the best clustering in terms of the used criteria. Implementation of this process involves two main stages. Firstly, the optimal values of the *EPS* parameter were determined at each step within the range of the min*Pts* value changes. The charts of the complex balance criterion versus the *EPS* value were obtained for each min*Pts* value as the results of this stage implementation. Then, the analysis of the obtained intermediate results was performed in order to determine the optimal solution, which corresponds to both the maximum value of the complex balance criterion on the one side and the aims of the current clustering on the other side.

Results. The developed hybrid model has been implemented based on software KNIME with the use of plugins, which have been written in software R. The efficiency of the model was tasted with the use of the different data: low dimensional data of the computing school of East Finland University; Fisher's iris; gene expression profiles of the patients, which were investigated on lung cancer.

Conclusions. The results of the simulation have shown high efficiency of the proposed method. The studied objects were distributed into clusters correctly in all cases. The proposed method allows us to decrease the reproducibility error, since the solution concerning determination of the clustering algorithm optimal parameters was taken based on both the clustering results obtained on equal power subsets separately and the difference of the clustering results obtained on the two equal power subsets.

KEYWORDS: Objective clustering, clustering quality criteria, inductive modelling, DBSCAN clustering algorithm.

ABBREVIATIONS

R is a free software environment for statistical computing and graphics;

KNIME is a Konstanz Information Miner;

DBSCAN is a Density-Based Spatial Clustering of Applications with Noise;

CH is Calinski-Harabasz criterion;

WB is Within-Between index;

QCI is an internal clustering quality criterion;

© Babichev S., Vyshemyrska S., Lytvynenko V., 2019
DOI 10.15588/1607-3274-2019-1-8

QCE is an external clustering quality criterion;
QCB is a complex balance clustering quality criterion;
OCIT is an objective clustering inductive technology.

NOMENCLATURE

n is a number of the investigated objects;
m is a number of features or attributes of the objects;
k is a number of clusters;
K is a set of the clusters;
R(*K*) is a clustering result;
e is a clustering error in the case of the two equal power subsets use;
EPS is an epsilon-neighborhood of point;
MinPts is a minimal quantity of points inside EPS;
e_0 is a boarding admissible clustering error;
$\{QC\}$ is the set of the internal and the external clustering quality criteria;
N is a number of the studied objects;
N_s is a number of the objects in *s* cluster;
X_i^s is *i-th* object in *s* cluster;
C_s is a mass center of the *s* cluster;
$d(\cdot)$ is a metric used to estimate the proximity level of the studied objects;
r is the number of the internal and external clustering quality criteria;
D is a set of points, each of them is determined the allocation of the studied object in m-dimensional feature space;
q is a point inside EPS;
$N_{EPS}(p)$ is a number of points inside *EPS* of the point *p*.

INTRODUCTION

Relevance of the problem is determined by the current works in the field of complex data clustering in different fields of scientific research. There are a lot of clustering algorithms nowadays. Each from them has its advantages and disadvantages and is focused to a specific type of data. The results of the data clustering depend on: affinity metrics between objects, clusters and objects and clusters; type of the clustering algorithm and the parameters of its operation; type of the clustering quality criteria, which are used to estimate the character of the object distribution within clusters. However, it should be noted, that in spite of grate quantity of the clustering algorithms and different types of internal and external clustering quality criteria this problem has not final solution nowadays. One of the unsolved tasks in this subject area is the reproducibility error. In other words, successful clustering results, which are obtained on one dataset do not repeat in the case of the use of another similar dataset. The solution of this problem can be achieved by careful determination of the parameters of the used clustering algorithm operation in all cases of the data clustering. However, this fact complicates the data processing, since the researcher in all cases should determine the optimal parameters of the current clustering algorithm operation in terms of the extremums of the used clustering quality criteria. To solve this problem, we propose to carry out the data clustering on two equal power subsets concurrently with following calculation of both the internal and external clustering quality criteria at each step of the algorithm operation. The final decision concerning the studied objects grouping is performed on the basis of maximum value of the complex balance criterion, which contains the internal and external clustering quality criteria as the components.

The aim of the work is practical implementation of DBSCAN clustering algorithm within the framework of the objective clustering inductive technology based on the complex use of R and KNIME tools.

1 PROBLEM STATEMENT

The initial dataset of the studied data is presented as a matrix: $A = \{x_{i,j}\}, i = 1,...,n; j = 1,...,m$. The clustering process involves a partition of the investigated objects into non-empty subsets of the pairwise non-intersection clusters:

$$K = \{K_s\}, s = 1,...,k; \quad K_1 \cup K_2 \cup ... \cup K_k = A;$$
$$K_p \cap K_q = \varnothing, \ p \neq q \ ; \ p,q = 1,...,k \ .$$

Model of the objective clustering based on the inductive methods of complex systems modelling involves sequential enumeration of admissible clustering in order to select from them the best variants [1]. The strategy *S* of the objects grouping within the framework of the objective clustering inductive technology can be presented as the following:

$$S : \left\{R(K) \mid (e < e_0) \xrightarrow{\{QC\}} opt\right\}.$$

Under the strategy in this case we understand a purposeful process of sequential actions, which are performed for the objects grouping according to the current task within the framework of the admissible error. The clustering error is determined based on analysis of the complex balance criterion values, which contains both the internal and external clustering quality criteria as the components.

2 REVIEW OF THE LITERATURE

Classification of several clustering algorithms by their categories are presented in [2]. Each of them has its advantages and disadvantages. The choice of the appropriate clustering algorithm is determined by type of the investigated data and goal of the current task. One of the essential disadvantages of the existing clustering algorithms is the reproducibility error. The main idea to solve this problem was proposed in [1]. The authors have shown that decreasing of the reproducibility error can be achieved based on the use of the inductive methods of complex systems modelling, which are a logical continuation of the group methods of data handling [3,4]. The questions concerning creation of the methodology of inductive systems analysis as a tool of engineering research analytical planning are considered in [5]. The authors proposed the strategy of analytical project design

© Babichev S., Vyshemyrska S., Lytvynenko V., 2019
DOI 10.15588/1607-3274-2019-1-8

based on the inductive principles. The final decision within the framework of the proposed methodology was done with the complex use of both the internal and external quality criteria. However, it should be noted, that authors' research is not focused to complex high-dimensional data clustering.

The results of the research concerning development of the objective clustering inductive technology of high-dimensional complex data are presented in [6]. The authors have shown that implementation of this technology based on some clustering algorithm involves determination of the affinity function between objects, clusters and objects and clusters at the first step. Then, division of the studied data into two equal power subsets, which contain the same quantity of pairwise similar objects should be performed. Formation of the internal, external and complex balance clustering quality criteria should be carried out at the next step. The optimal clustering is determined based on the extremum value of the used criteria during sequential enumeration of the admissible clustering. In [7] authors present the results of the research concerning criterial analysis of the gene expression profiles within the framework of the objective clustering inductive technology. The implementation of this technology based on k-means and agglomerative hierarchical clustering algorithms were presented in [8, 9]. The authors conducted the comparison analysis of the different internal and external clustering quality criteria with evaluation of their effectivity in the case of gene expression profiles use. The results of the simulation have shown higher effectiveness of the appropriate algorithm in the case of its implementation within the framework of the objective clustering inductive technology in comparison with standard method of this algorithm use. However, it should be noted that the used algorithms do not allow us to divide the complex data correctly. The solution of this problem can be achieved by the use of the modern methods of complex data processing [10, 11] within the framework of the objective clustering inductive technology and implementation of this technology based on other clustering algorithms.

In this work we propose the hybrid model of the objective clustering inductive technology based on DBSCAN clustering algorithm. The practical implementation of the proposed model has been performed on the basis of the complex use of both R and KNIME tools.

3 MATERIALS AND METHODS

Three fundamental principles, which are borrowed from various scientific fields, are the basis of the methodology of complex systems inductive modeling. In the case of the OCIT these principles can be presented as the following [1, 12]:

1. The principle of sequential enumeration, i.e., sequential enumeration of admissible clustering within a given range in order to select from them the best variants by the used clustering quality criteria;

2. The principle of external edition, i.e., a necessity of the use of two equal power subsets, which contain the same quantity of pairwise similar objects;

3. The principle of inconclusive of solution, i.e., generation of several sets of intermediate results in order to select from them the best variant in terms of the goal of the current task.

Fig. 1 presents the structural block chart of the OCIT.

The practical implementation of this technology involves the following stages:

Stage I. Problem statement. Data analysis and preprocessing.

1. Problem statement and aim formation.

2. Data analysis and its formation as a matrix, where number of rows is a number of the studied objects and number of columns is a number of the features, which characterized the objects.

3. The data preprocessing. This step involves: missing value processing (in the case of necessity); filtering; normalization.

Stage II. Choice of the affinity metrics and equal power subsets formation.

4. Choice the affinity metrics between objects, clusters, objects and clusters.

5. Formation of the two equal power subsets A and B in accordance with the following algorithm:

Step 1. Calculation of $\frac{n\times(n-1)}{2}$ pairwise distances between all pairs of the studied objects. The triangular

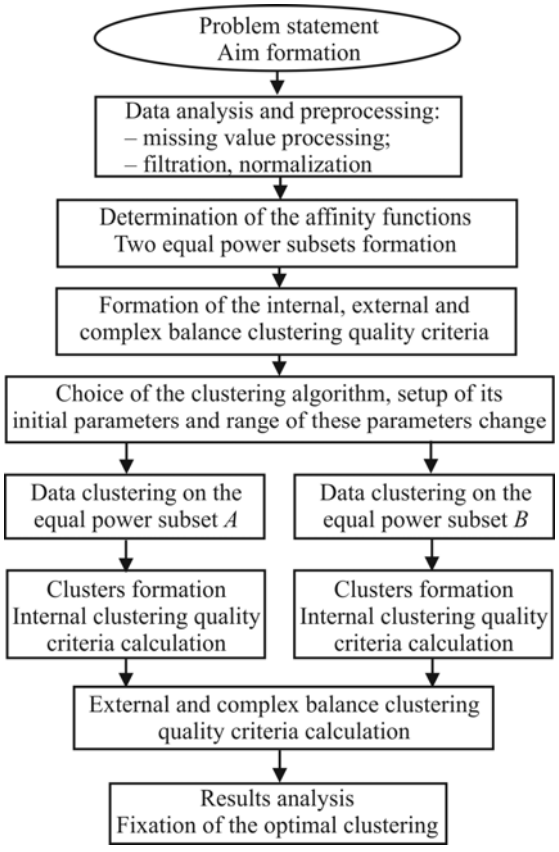

Figure 1 – Structural block chart of the OCIT

© Babichev S., Vyshemyrska S., Lytvynenko V., 2019
DOI 10.15588/1607-3274-2019-1-8

matrix of distances is the result of this step implementation:

$$Dist = \begin{Vmatrix} 0 & d_{12} & d_{13} & \dots & d_{1n} \\ & 0 & d_{23} & \dots & d_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ & & & 0 & d_{n-1,n} \\ & & & & 0 \end{Vmatrix}.$$

Step 2. Allocation of the pair of the objects X_s and X_p, the distance between with is minimal:

$$d(X_s, X_p) = \min_{i,j} d(X_i, X_j).$$

Step 3. Distribution of the object X_s to subset A, and the object X_p to subset B.

Step 4. Repetition of the steps 2 and 3 for remaining objects. If the number of the objects is odd, the last object is distributed into the both subsets.

Stage III. Calculation the internal, external and complex balance clustering quality criteria.

6. Calculation of the internal clustering quality criterion. This criterion allows us to evaluate the quality of the objects grouping in single clustering. It is obvious that quality clustering corresponds to both the high density of the objects distribution inside cluster and the less density of the clusters distribution in the features space. So, the internal clustering quality criterion should be complex and takes into account both the character of the objects distribution within clusters and the character of the mass centers of the obtained clusters distribution. The first component of this criterion within the framework of the proposed technology was calculated as an average distance from objects to the mass centers of the cluster, where these objects are by the formula (1):

$$QCW = \frac{1}{N} \sum_{s=1}^{k} \sum_{i=1}^{N_s} d(X_i^s, C_s). \quad (1)$$

The second component of this criterion is calculated as an average distance between mass centers of the clusters in current clustering by the formula (2):

$$QCB = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} d(C_i, C_j). \quad (2)$$

Various combinations of this components in different internal clustering quality criteria in the case of the use of numeric data were considered in [7]. The authors have shown that Calinski-Harabasz criterion (CH) [13] and Within-Between index (WB) [14] show better results in the case of the use if high dimensional gene expression profiles. As the results of the simulation the complex internal clustering quality criterion was proposed in [15]. This criterion is calculated as the multiplicative combination of *CH* criterion and *WB* index by the formula (3):

$$QCI = \frac{k(k-1)QCW^2}{(N-k)QCB^2} \to \min. \quad (3)$$

7. Calculation of the external clustering quality criterion. This criterion takes into account the difference of the clustering results obtained on the two equal power subsets. The minimal value of this criterion corresponds to higher level of the clustering objectivity. The value of this criterion is calculated as normalised difference of the internal clustering quality criteria calculated on the two equal power subsets for the current clustering level by the formula (4):

$$QCE = \frac{|QCI(A) - QCI(B)|}{QCI(A) + QCI(B)} \to \min. \quad (4)$$

8. Calculation of the complex balance criterion. The necessity of this criterion is determined by possible disagree between the extremums of both the internal and external clustering quality criteria. The Harrington desirability function was proposed to calculate the complex balance criterion [16]. The plot of this function is presented in Fig. 2.

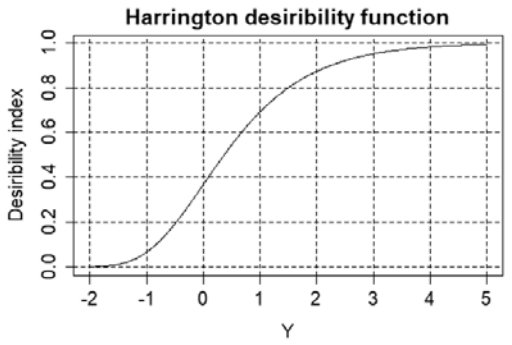

Figure 2 – Chart of the Harrington desirability function

Determination of the general Harrington desirability index involves the following steps:

Step 1. Transformation of scales of the internal and external clustering quality criteria into reaction scale Y, values of which are changed within the range from –2 to 5 by the formula (5):

$$Y = a - b \cdot QC. \quad (5)$$

The parameters a and b are determined empirically for each of the used criteria taking into account its boundary values according to the equations (6):

$$\begin{cases} Y_{\max} = a - b \cdot QC_{\min} \\ Y_{\min} = a - b \cdot QC_{\max} \end{cases}. \quad (6)$$

Step 2. Calculation of the nondimensional parameter Y_i for each of the used clustering quality criteria QC_i by the formula (7):

$$Y_i = a - b \cdot QC_i. \quad (7)$$

Step 3. Calculation of the private desirabilities for each of the criteria by the formula (8):

$$d = \exp(-\exp(-Y)). \quad (8)$$

Step 4. Calculation of the general Harrington desirability index as geometric average of all private desirabilities for appropriate clustering level by the formula (9):

$$QCB = \sqrt[r]{\prod_{i=1}^{r} d_i}. \quad (9)$$

© Babichev S., Vyshemyrska S., Lytvynenko V., 2019
DOI 10.15588/1607-3274-2019-1-8

The largest value of the criteria (9) corresponds to the best clustering in terms of the used criteria.

Stage IV. Data clustering on the equal power subsets *A* and *B* concurrently.

9. Choice of the clustering algorithm depend on type of the used data and goal of the research. Setup of its initial parameters, ranges and steps of these parameters change.

10. Data clustering on the equal power subsets A and B within the range of the algorithm parameters change. Calculation of the internal and the external clustering quality criteria at each step of this procedure implementation.

11. Calculation of the complex balance clustering quality criterion within the range of the algorithm parameters change.

Stage V. Analysis of the obtained results. Fixation of the optimal clustering.

12. Analysis of the obtained results. Fixation of the best clustering, which correspond to the maximum value of the complex balance criterion.

13. Comparison analysis of the intermediate solutions. Fixation of the optimal clustering, which corresponds to both the maximum value of the complex balance clustering quality criterion and the goal of the current task.

DBSCAN clustering algorithm was proposed in 1996 as a solution of the problem to divide the data into clusters of arbitrary shapes [17–19]. The following definitions are the basis of this algorithm operation [18]:

Definition 1. The *Eps-neighborhood* of a point p is defined by the following:

$$EPS(p) = \{q \in D | d(p,q) \le EPS\}.$$

Definition 2. A point q is directly density-reachable from a point p if the following conditions are performed:

$$\begin{cases} q \in EPS(p) \\ N_{EPS}(p) \ge MinPts \end{cases}.$$

Definition 3. A point q is density-reachable from a point p if there is a chain of points $q_1,...,q_n$, $q_1 = p$, $q_n = q$ such that q_{i+1} is directly density-reachable from q_i.

Definition 4. A point q is density-connected with a point p if there is a point k such that both the points q and p are density-reachable from the point k.

Definition 5. A cluster C is a non-empty subset of a set of points D if the following conditions are performed:

1. $\forall p,q$: if $p \in C$ and q is density-reachable from p, then $q \in C$;

2. $\forall p,q$: if q is density-connected with p, then $p,q \in C$.

Definition 6. Let $C_i, i = 1,...,k$ is a set of the allocated clusters. The noise is the set of points of the database D, which not belonging to any cluster C_i:

$$noise = \{p \in D | \forall i : p \notin C_i, i = 1,...,k\}.$$

Result of DBSCAN clustering algorithm operation depends on the two parameters: *EPS* and Min*Pts*. To determine the optimal *EPS* value for appropriate Min*Pts* the technology based on sorted *k*-dist graph was proposed in [18]. However, it should be noted, that implementation of this technology does not allow us to determine the *EPS* value exactly. This fact influences the quality of the algorithm operation. The implementation of the proposed technology allows us to determine only the range of the *EPS* values change for appropriate Min*Pts* value. To solve this problem, we propose the use of DBSCAN clustering algorithm within the framework of the OCIT. The structural block chart of the algorithm to implement this process is presented in Fig. 3. The implementation of this algorithm involves the following steps:

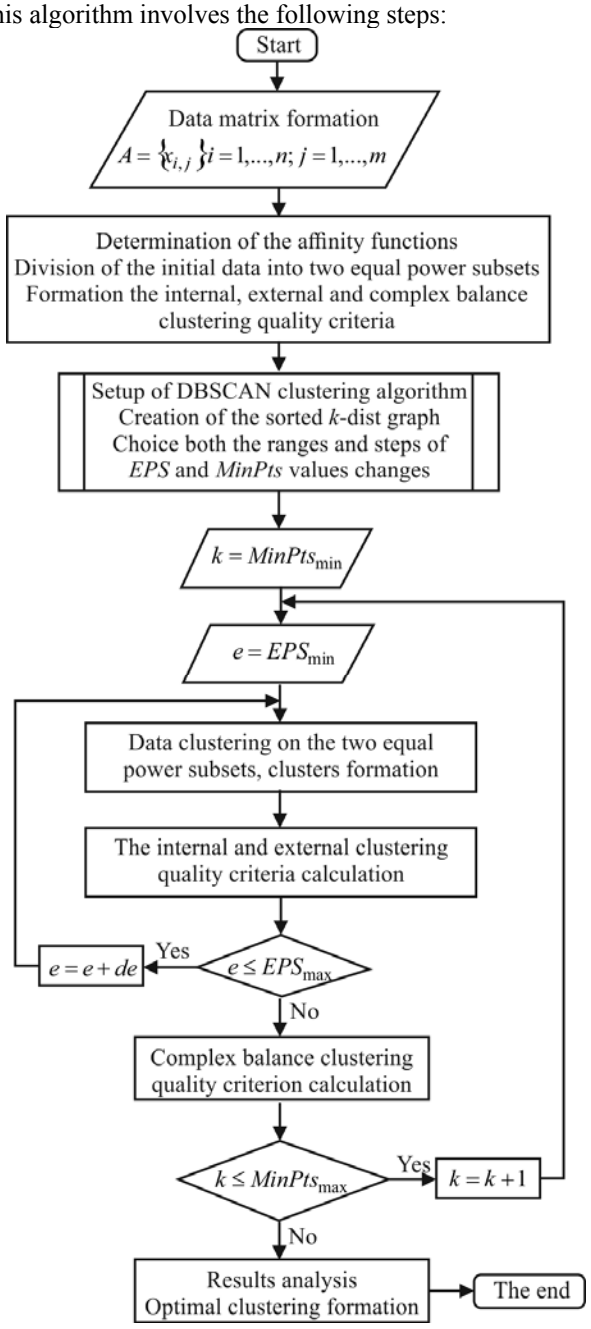

Figure 3 – Structural block chart of the algorithm of the OCIT hybrid model based on DBSCAN clustering algorithm

© Babichev S., Vyshemyrska S., Lytvynenko V., 2019
DOI 10.15588/1607-3274-2019-1-8

Step 1. Formation of the initial data as a matrix, where number of rows is the number of the studied objects and number of columns is a number of the features, which characterized the objects.

Step 2. Determination of the affinity functions in dependence on type of the studied data. Division of the initial data into two equal power subsets.

Step 3. Formation of the internal, external and complex balance clustering quality criteria.

Step 4. Setup of DBSCAN clustering algorithm. Determination of the range of the Min*Pts* value change. Creation of the sorted *k*-dist graph within this range. Determination of both the range and step of the *EPS* value change.

Step 5. Setup of the initial value of the Min*Pts* algorithms parameter (k = min(Min*Pts*)).

Step 6. Setup of the initial value of the *EPS* algorithms parameter (e = min(*EPS*)).

Step 7. Data clustering on the two equal power subsets concurrently. Clusters formation.

Step 8. Calculation of both the internal and external clustering quality criteria by formulas (3) and (4).

Step 9. If the condition $e \leq$ max(*EPS*) is true increasing the EPS value (e=e+de) and repetition of the steps 7 and 8 of this procedure. Otherwise, calculation of the complex balance criterion by the formulas (5)–(9).

Step 10. If the *MinPts* value is less than maximum ($k \leq$ max(Min*Pts*)) increasing the Min*Pts* value (k=k+1) and transition to the step 6 of this algorithm. Otherwise, creation of the charts of the complex balance criterion versus the *EPS* for each Min*Pts* value.

Step 11. Analysis of the obtained results. Fixation of the optimal clustering.

4 EXPERIMENTS

The simulation of the proposed technology was performed based on KNIME analytics platform [20] using R software [21]. The structure of the used model is presented in Fig. 4. To estimate the effectiveness of the proposed technology the data "Aggregation" [22], "Compound" [23], "Multishapes" [24] and "Jain" [25] of the school of computing of the Eastern Finland University were used. These data are presented in the two-dimensional space and they include the clusters of different shapes. Fig. 5 shows the character of the studied data distribution.

Other datasets were the Fisher's iris [26] and gene expression profiles of the patients, which were investigated on lung cancer [27]. The data of the gene expression profiles was presented as a matrix, where the number of rows is the number of the studied genes (2000) and the number of columns is the number of the studied objects or the conditions of the experiment performing (96). The gene expression profile in this case is a vector of gene expressions, which were determined for the different conditions of the experiment performing. To estimate the proximity level between the studied vectors we used Euclidean distance in the case of low-dimensional data and the correlation distance in the case of the gene expression profiles use. In accordance with algorithm presented in Fig. 3, the studied data were normalized and divided into two equal power subsets with the use of the hereinbefore presented algorithm. Then, the sorted k-dist graphs were created within the boundary range of the MinPts value change from 3 to 8. These k-dist graphs were used to determine the range of the *EPS* value change. The date clustering on the two equal power subsets within the range of the EPS value change for each MinPts value were performed at the next step of data processing.

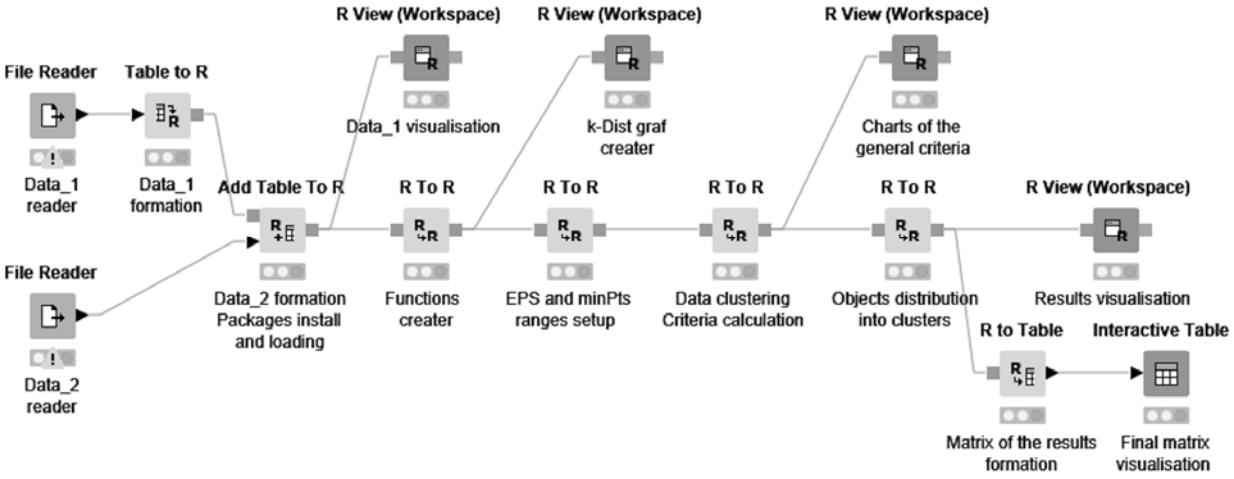

Figure 4 – The model of the objective clustering inductive technology based on DBSCAN clustering algorithm

© Babichev S., Vyshemyrska S., Lytvynenko V., 2019
DOI 10.15588/1607-3274-2019-1-8

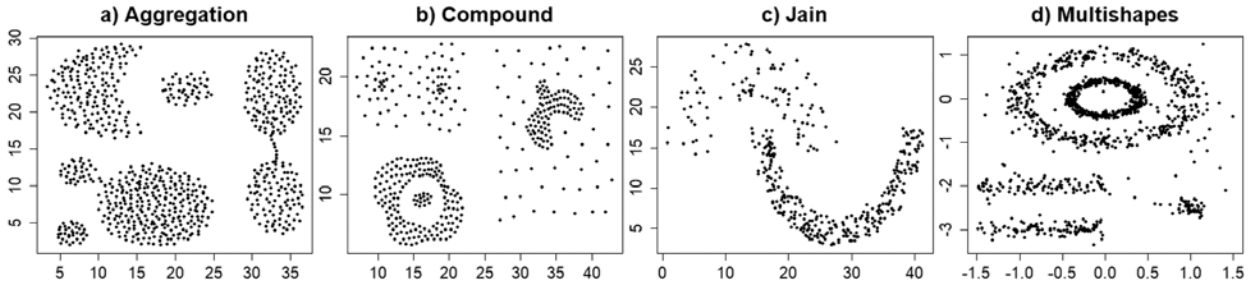

Figure 5 – The two-dimensional data of the school of computing of the Eastern Finland University

As the results, we have obtained the charts of the complex balance clustering quality criterion versus the *EPS* for each Min*Pts* value. The analysis of these charts allows us to determine the best clustering in terms of both the used criteria and goal of the current task.

5 RESULTS

Fig. 6 shows the sorted k-dist graphs for Aggregation data. The similar graphs were obtained in the case of the other data use.

The analysis of the obtained results allows us to determine both the ranges and steps of the EPS value changes for each type of the investigated data. These parameters are presented in the Table 1. Charts of the complex balance criterion versus the EPS for different MinPts value in the case of the "Aggregation" data use are presented in Fig. 7. The similar charts were obtained for the other investigated data. The analysis of the obtained charts allows us to select the subset of the intermediate solutions (new less ranges and steps of the *EPS* value change), which correspond to the maximum values of the complex balance criterion.

Then, the detail analysis of the selected solutions is performed in order to determine the optimal clustering in terms of the goal of the current task. The optimal parameters of DBSCAN clustering algorithm operation, which were determined within the framework of the proposed technology for the investigated data are presented in Table 2. Fig. 8 presents the results of the two-dimensional data clustering. Fig. 9 and Fig. 10 presents the same results in the cases of the use of both "iris" data and the gene expression profiles.

Table 1 – The range and step of the EPS value change

Data	Aggregation	Compound	Multishapes
EPS_{min}	0.1	0.1	0.1
EPS_{max}	0.25	0.6	0.4
Step	0.005	0.01	0.005
Data	Jain	Iris	Gene expression
EPS_{min}	0.15	0.5	0.01
EPS_{max}	0.5	1.5	0.5
Step	0.005	0.01	0.01

Table 2 – The optimal parameters of DBSCAN clustering algorithm operation

Data	Aggregation	Compound	Multishapes
EPS	0.136	0.157	0.237
Min*Pts*	5	5	4
Data	Jain	Iris	Gene expression
EPS	0.305	0.646	0.421
Min*Pts*	3	3	4

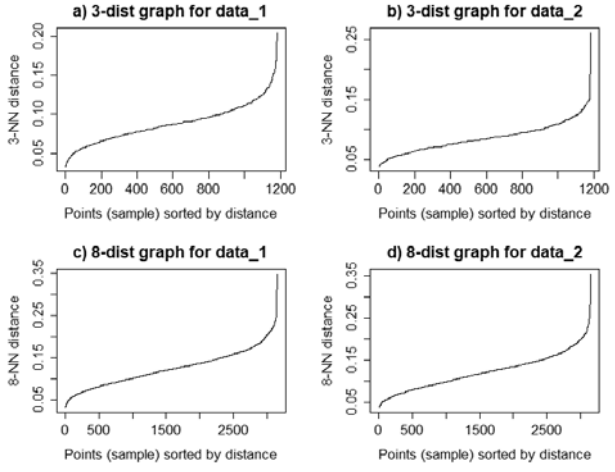

Figure 6 – Sorted k-dist graph for "Aggregation" data

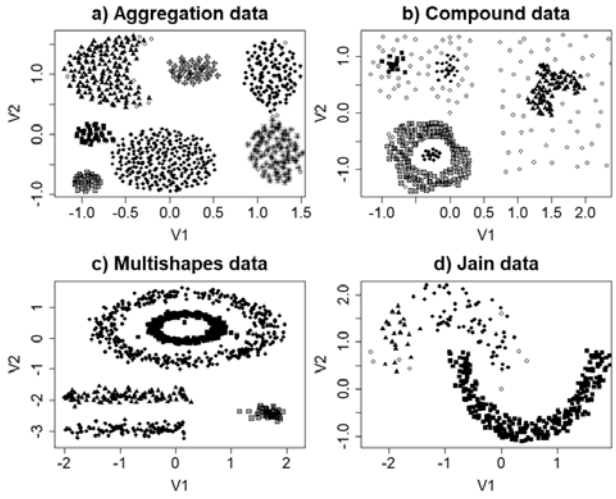

Figure 8 – Results of the two-dimensional data clustering

© Babichev S., Vyshemyrska S., Lytvynenko V., 2019
DOI 10.15588/1607-3274-2019-1-8

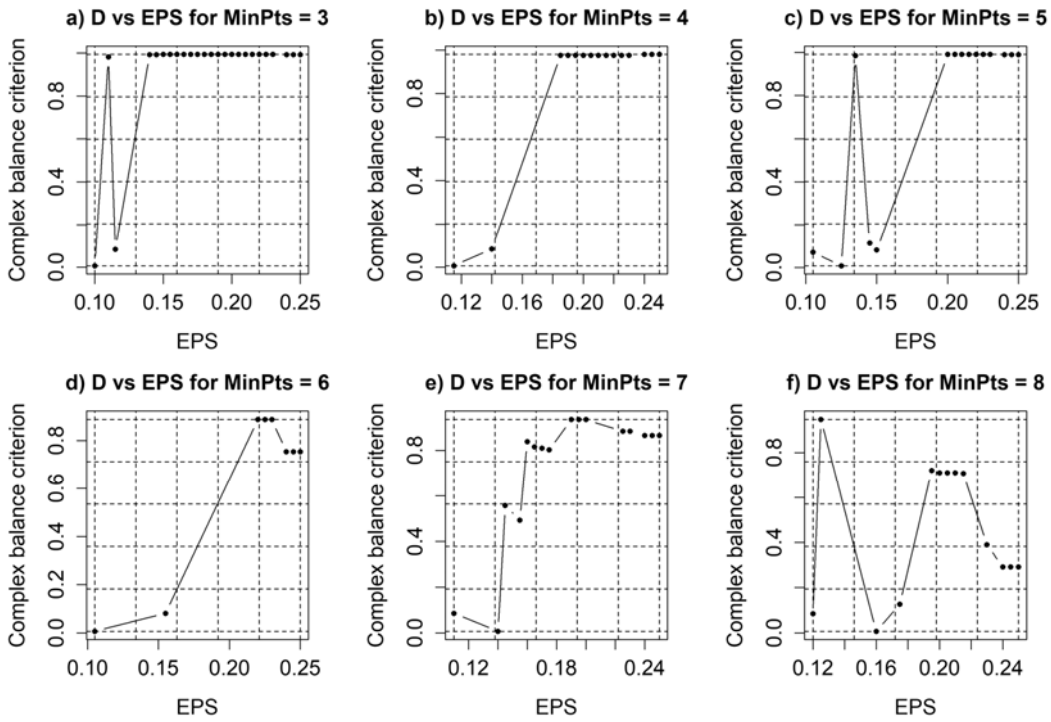

Figure 7 – The charts of the complex balance criterion versus the EPS for different MinPts values in the case of "Aggregation" data use

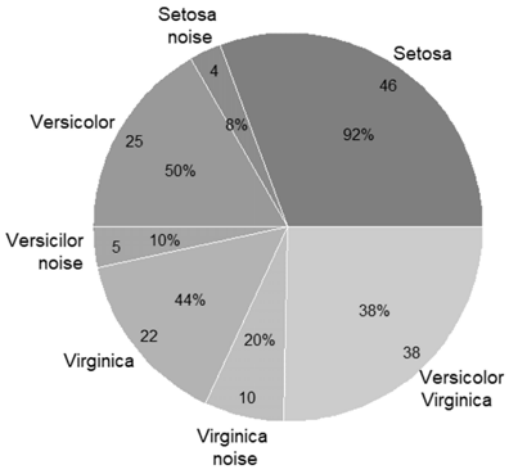

Figure 9 – Results of "Iris" data clustering

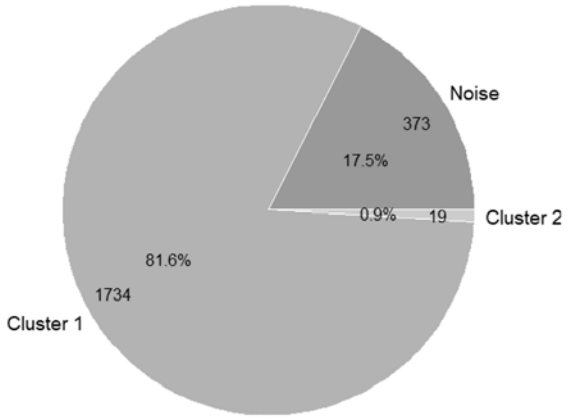

Figure 10 – Results of the gene expression profiles clustering

6 DISCUSSION

The analysis of the obtained results allows us to conclude that the objects were distributed into clusters correctly in all cases.

So, in the case of "Aggregation data" (Fig. 8a) we have as the result seven clusters. Several objects were identified as noise since the density of their distribution in the feature space is less than the density of the other objects distribution within the obtained clusters. It should be noted that in this case the connected clusters were divided correctly. The result of "Compound" data clustering is presented in Fig. 8b. As it can be seen, in this case the objects are distributed into clusters correctly too. We have as the result five clusters that corresponds to the character of the objects distribution in the feature space. A lot of the objects are identified as noise. It is naturally, since the density of these objects distribution in the feature space is significantly less in comparison with density of the other objects distribution. The same results are observed in the case of "Multishapes" (Fig. 8c) and "Jain" (Fig. 8d) data use. "Multishapes" data contains clusters different shapes and sizes. As it can be seen, the studied objects are distributed into clusters correctly. Five clusters were allocated in this case. The objects of "Jain" data were distributed into three clusters. It should be noted that the little change of the DBSCAN algorithm parameters decreases of the clustering quality in all cases. The

© Babichev S., Vyshemyrska S., Lytvynenko V., 2019
DOI 10.15588/1607-3274-2019-1-8

intersected or non-divided clusters are appeared in this case.

The analysis of the result of "Iris" data clustering (Fig. 9) allows us to conclude that the objects of "Setosa" class were allocated in the first cluster. This cluster has no any intersection with the other clusters. However, the four objects of "Setosa" class were identified as the noise. The detail analysis of the parallel coordinates plot for objects of "Setosa" class has shown that this class contains several objects, the profiles of which are distinguished from the profiles of other objects of this class. Thus, the obtained result is adequate. The analysis of the parallel coordinates plot for the objects of "Virginica" and "Versicolor" classes have shown that these classes have some intersection a priory. Fifty percent of the objects of "Versicolor" class and Forty-four percent of the objects of "Virginica" class were distributed into the second and the third clusters accordingly. Ten percent of the objects of "Versicolor" class and twenty percent of the objects of "Virginica" class were identified as the noise. The analysis of the parallel coordinates plot has shown that these classes contained the objects, profiles of which have distinguishes from the profiles of the other objects of these classes. Moreover, the results of the analysis have shown also that the second and the third clusters have thirty-eight percent of intersection in this case. However, this is correctly in view of the type of the studied data.

In the case of the use of the gene expression profiles of the patients, which were investigated on lung cancer, the data were distributed into three clusters. The first cluster contained 81.6% of the gene expression profiles, which determine the main of the processes in the investigated object. The second cluster contained only 0.9% of the investigated gene expression profiles. 17.5% of the gene expression profiles were identified as the noise. The first cluster in this case presents the best interest for the following processing since this cluster contains the genes, which define the main functions in the studied object.

As the results it should be noted that implementation of the proposed technology allows us to determine the optimal parameters of DBSCAN clustering algorithm in terms of the clustering quality. The analysis of the obtained results has shown that the investigated data were distributed into clusters correctly in the case of the use of different types of the data.

ACKNOWLEDGEMENTS

The work is supported by the state budget scientific research project of Kherson National Technical University "Development of hybrid neuro-phase-immune algorithms for information systems and technologies for solving problems in bioinformatics and computational biology" (state registration number 0116U002841).

CONCLUSIONS

The relevant problem concerning increase of the complex data clustering quality is solved based on the use of DBSCAN clustering algorithm within the framework of the objective clustering inductive technology.

The scientific novelty of the proposed hybrid model is the following:

– the data clustering is performed on the two equal power subsets concurrently within the range of the algorithm parameters change;

– the optimal parameters of the clustering algorithm are determined based on the maximum values of the complex balance criterion, which contain as the components both the internal and external clustering quality criteria;

– the final solution concerning selection of the optimal clustering is performed based on the comparison analysis of the best intermediate solutions takes into account the goal of the current task.

The implementation of the proposed information technology allows us to increase the quality of the data clustering due to the paralleling of the data processing and the use of both the internal and the external clustering quality criteria.

The practical significance of the obtained results is the practical implementation of the proposed hybrid model based on the complex use of the R and KNIME tools. The hybrid model was tested on the different types of the investigated data. The analysis of the obtained results has shown the high effectiveness of the proposed technology since the investigated data were distributed into clusters correctly in all cases.

The prospects for further research are the implementation of the objective clustering inductive technology based on other clustering algorithms.

REFERENCES

1. Madala H. R., Ivakhnenko A. G. Inductive learning algorithms for complex systems modeling. CRC Press, 1994, 365 p.
2. Soni N., Ganatra A. Categorization of several clustering algorithms from different perspective: a review, *International Journal of Advanced Research in Computer Science and Software Engineering,* 2012, Vol. 2, Issue 8, pp. 63–68.
3. Stepashko V., Bulgakova O. , Zosimov V. Construction and research of the generalized iterative GMDH algorithm with active neurons, *Advances in Intelligent Systems and Computing II,* 2018, pp. 492–510. DOI: 10.1007/978-3-319-70581-1_35
4. Bulgakova O., Stepashko V., Zosimov V. Numerical study of the generalized iterative algorithm GIA GMDH with active neurons, *Proceedings of the 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies*, 2017, 1, art. no. 8098836, pp. 496–500. DOI: 10.1109/STC-CSIT.2017.8098836
5. Osypenko V. V., Reshetjuk V. M. The methodology of inductive system analysis as a tool of engineering researches analytical planning, *Ann. Warsaw Univ. Life Sci,* 2011, SGGW. No. 58, pp. 67–71. [Electronic resource]. Access mode: http://annals-wuls.sggw.pl/?q=node/234
6. [Babichev S. , Taif M. A. , Lytvynenko V. , Korobchinskyi M.] Objective clustering inductive technology of gene expression sequences features, *Communications in Computer and Information Science: In the book "Beyond*

© Babichev S., Vyshemyrska S., Lytvynenko V., 2019
DOI 10.15588/1607-3274-2019-1-8

Databases, Architectures and Structures", edited by S. Kozelski and D. Mrozek, 2017, pp. 359–372.
7. Babichev S. , Taif M. A. , Lytvynenko V. , Osypenko V. Criterial analysis of gene expression sequences to create the objective clustering inductive technology, *Proceeding of the 2017 IEEE 37th International Conference on Electronics and Nanitechnology (ELNANO),* 2017, pp. 244–249.
8. Babichev S., Taif M. A. , Lytvynenko V. Inductive model of data clustering based on the agglomerative hierarchical algorithm, *Proceeding of the 2016 IEEE First International Conference on Data Stream Mining and Processing (DSMP,* 2016, pp. 19–22. [Electronic resource]. Access mode: http://ieeexplore.ieee.org/document/7583499/
9. Babichev S., Taif M. A. , Lytvynenko V. Estimation of the inductive model of objects clustering stability based on the k-means algorithm for different levels of data noise, *Radio Electronics, Computer Science, Control,* 2016, No. 4, pp. 54–60.
10. Puchala D., Yatsymirskyy M. M. Joint compression and encryption of visual data using orthogonal parametric transforms, *Bulletin of the Polish Academy of Sciences-Technical Sciences,* 2016, Vol. 64, Issue 2, pp. 373–382.
11. Rashkevych Y., Peleshko D., Vynokurova O., Izonin I., Lotoshynska N. Single-frame image super-resolution based on singular square matrix operator, *1st IEEE Ukraine Conference on Electrical and Computer Engineering (UKRCON), MAY 29-JUN 02*, 2017, pp. 944–948.
12. Ivakhnenko A. Group method of data handling as competitor to the method of stochastic approximation, *Soviet Automatic Control,* 1968, Vol. 3, pp. 64–78.
13. Calinski T., Harabasz J. A dendrite method for cluster analysis, *Communication in Statistics,* 1974, Vol. 3, pp. 1-27.
14. Zhao Q. Xu M., Fränti P. Sum-of-squares based cluster validity index and significance analysis, *Proceeding of International Conference on Adaptive and Natural Computing Algorithms,* 2009, pp. 313–322.
15. Babichev S. , Krejci J. , Bicanek J. , Lytvynenko V.] Gene expression sequences clustering based on the internal and external clustering quality criteria, *Proceedings of the 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies*, 2017, 1, art. no. 8098744, pp. 91–94. DOI: 10.1109/STC-CSIT.2017.8098744.
16. Harrington J. The desirability function, *Industrial Quality Control,* 1965, Vol. 21(10), pp. 494–498. [Electronic resource]. Access mode: http://asq.org/qic/display-tem/?item=4860.
17. Ester M., Kriegel H. , Sander J. A density-based algorithm for discovering clusters in large spatial databases, *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996, pp. 226–231.
18. Ester M. , Kriegel H. , Sander J. , Xu X.] A density-based algorithm for discovering clusters in large spatial databases with noise, *KDD-1996 : proceedings*, 1996, pp. 226–231.
19. Kriegel H.-P., Kröger P., Sander J., Zimek A. Density-based clustering, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* 2011, No. 1(3), pp. 231–240.
20. [Electronic resource]. Access mode: https://www.knime.com/ knime-software/knime-analytics-platform
21. Ihaka R., Gentleman R. R: a linguage for data analysis and graphics, *Journal of Computational and Graphical Statistics,* 1996, Vol. 5(3), pp. 299–314.
22. Gionis A., Mannila H. , Tsaparas P. Clustering aggregation, *ACM Transactions on Knowledge Discovery from Data (TKDD),* 2007, Vol. 1(1), pp. 1–30.
23. Zahn C. T. Graph-theoretical methods for detecting and describing gestalt clusters, *IEEE Transactions on Computers,* 1971, Vol. 100(1), pp. 68–86.
24. [Electronic resource]. Access mode: http://www.sthda.com/english/rpkgs/factoextra
25. Jain A., Law M. Data clustering: A user's dilemma, *Lecture Notes in Computer Science*, 2005, Vol. 3776, pp. 1-10.
26. Fisher R. A. The Use of Multiple Measurements in Taxonomic Problems, *Annals of Eugenics,* 1936, Vol. 7, pp. 179-188.
27. Beer D. G. , Kardia S. L. and al Gene-expression profiles predict survival of patients with lung adenocarcinoma, *Nature Medicine*, 2002, Vol. 8(8), pp. 816–824.

Received 25.11.2018.
Accepted 16.12.2018.

УДК 004.048

ЗАСТОСУВАННЯ АЛГОРИТМУ КЛАСТЕРІЗАЦІЇ DBSCAN У РАМКАХ ІНДУКТИВНОЇ ТЕХНОЛОГІЇ ОБ'ЄКТНОЇ КЛАСТЕРІЗАЦІЇ НА ОСНОВІ ПРОГРАМНИХ ЗАСОБІВ R ТА KNIME

Бабічев С. А. – канд. техн. наук, доцент, доцент кафедри інформатики університету імені Яна Євангеліста Пуркінє в Усті на Лабі, Усті над Лабем, Чехія. Доцент кафедри інформаційних технологій ІТ Степ Університету, Львів, Україна.

Вишемирська С. В. – канд. техн. наук, доцент, доцент кафедри інформатики і комп'ютерних наук Херсонського національного технічного університету, Херсон, Україна.

Литвиненко В. І. – д-р техн. наук, професор, завідувач кафедри інформатики і комп'ютерних наук Херсонського національного технічного університету, Херсон, Україна.

АНОТАЦІЯ

Актуальність. Досліджено проблему кластеризації складних даних у рамках індуктивної технології об'єктивної кластеризації. Зроблено практичну реалізацію запропонованої гібридної моделі кластеризації даних на основі комплексного застосування програмних засобів R і KNIME. Об'єктом дослідження є гібридна модель кластеризації даних на основі комплексного застосування алгоритму кластеризації DBSCAN і індуктивної технології об'єктивної кластеризації. Мета роботи полягає у створенні гібридної моделі об'єктивної кластеризації на основі алгоритму кластеризації DBSCAN та практична реалізація моделі на основі комплексного застосування програмних засобів R і KNIME.

Метод. Індуктивні методи моделювання складних систем були використані як основа для визначення оптимальних параметрів алгоритму кластеризації DBSCAN в рамках індуктивної технології об'єктивної кластеризації. Практична реалізація даної технології передбачає: застосування рівнопотужніх підмножин даних, які містять однакову кількість попарно близьких об'єктів; розрахунок внутрішнього та зовнішнього критеріїв якості кластеризації; розрахунок комплексного критерія балансу, максимальне значення якого відповідає найкращій кластеризації з точки зору критеріїв, що

© Babichev S., Vyshemyrska S., Lytvynenko V., 2019
DOI 10.15588/1607-3274-2019-1-8

використовуються. Реалізація процесу визначення оптимальних параметрів алгоритму DBSCAN передбачає два етапи. Першим етапом є визначення оптимального значення параметра *EPS* в межах діапазону зміни значень параметру *minPts*. Результатом реалізації даного етапу є отримання діаграм залежності комплексного критерію балансу від відповідних значень *EPS* для кожного значення min*Pts*. Потім проводився аналіз отриманих проміжних результатів для визначення оптимального рішення, що відповідає максимальному значенню комплексного критерію балансу в залежності від мети поставленої задачі.

Результати. Розроблена гібридна модель індуктивної технології об'єктивної кластеризації на основі алгоритму DBSCAN, яка практично реалізована на основі програмних засобів KNIME R. Виконано оцінку ефективність моделі з використанням різних типів даних: низько-розмірних даних школи обчислень університету східної Фінляндії; ірисів Фішера; профілів експресії генів пацієнтів, які досліджувалися на рак легенів.

Висновки. Результати моделювання показали високу ефективність запропонованої технології. Досліджені об'єкти були розподілені у кластери коректно в усіх випадках. Запропонований метод дозволяє зменшити значення похибки відтворюваності, оскільки остаточне рішення щодо визначення оптимальних параметрів алгоритму кластеризації приймається на основі паралельного аналізу результатів кластеризації, отриманих на рівнопотужніх підмножинах даних, так і на основі аналізу різниці результатів кластеризації, отриманих на даних підмножинах.

КЛЮЧОВІ СЛОВА: об'єктивна кластеризація, критерії якості кластеризації, індуктивне моделювання, алгоритм кластеризації DBSCAN.

УДК 004.048

РЕАЛИЗАЦИЯ АЛГОРИТМА КЛАСТЕРИЗАЦИИ DBSCAN В РАМКАХ ИНДУКТИВНОЙ ТЕХНОЛОГИИ ОБЪЕКТИВНОЙ КЛАСТЕРИЗАЦИИ НА ОСНОВЕ ПРОГРАММНЫХ СРЕДСТВ KNIME И R

Бабичев С. А. – канд. техн. наук, доцент, доцент кафедры информатики университета имени Яна Евангелиста Пуркине в Усти на Лабе, Усти над Лабем, Чехия. Доцент кафедры информационных технологий IT Университета Шаг, Львов, Украина.

Вышемирская С. В. – канд. техн. наук, доцент, доцент кафедры информатики и компьютерных наук Херсонского национального технического университета, Херсон, Украина.

Литвиненко В. И. – д-р техн. наук, профессор, заведующий кафедрой информатики и компьютерных наук Херсонского национального технического университета, Херсон, Украина.

АННОТАЦИЯ

Актуальность. Исследована проблема кластеризации сложных данных в рамках индуктивной технологии объективной кластеризации. Выполнена практическая реализация гибридной модели кластеризации данных на основе комплексного использования программных средств R и KNIME. Объектом исследования является гибридная модель кластеризации данных на основе комплексного использования алгоритма кластеризации DBSCAN и индуктивной технологии объективной кластеризации. Цель работы – разработка гибридной модели объективной кластеризации на основе алгоритма кластеризации DBSCAN и практическая реализация данной модели на основе комплексного использования программных средств R и KNIME.

Метод. Индуктивные методы моделирования сложных систем использовались в качестве основы для определения оптимальных параметров алгоритма кластеризации DBSCAN в рамках индуктивной технологии объективной кластеризации. Практическая реализация данной технологии предполагает: использование равномощных подмножеств данных, содержащих одинаковое количество попарно близких объектов; определение значений внутреннего и внешнего критериев качества кластеризации; расчет комплексного критерия баланса, максимальное значение которого соответствует наилучшей кластеризации с точки зрения используемых критериев. Реализация процесса определения оптимальных параметров алгоритма кластеризации DBSCAN предполагает два этапа. Первым этапом является определение оптимального значения параметра *EPS* в пределах диапазона изменения значений параметра min*Pts*. Результатом реализации данного этапа является получение диаграмм зависимости комплексного критерия баланса от соответствующих значений *EPS* для каждого значения min*Pts*. Реализация следующего этапа предполагает анализ полученных промежуточных результатов для определения оптимального решения, соответствующего максимальному значению комплексного критерия баланса при условии выполнения цели текущей задачи.

Результаты. Разработана гибридная модель индуктивной технологии объективной кластеризации на основе алгоритма DBSCAN, которая практически реализована на основе комплексного использования программных средств KNIME и R. Выполнена оценка эффективности модели с использованием различных типов данных: низкоразмерных данных школы вычислений университета восточной Финляндии; ирисов Фишера; профилей экспрессии генов пациентов, исследуемых на рак легких.

Выводы. Результаты моделирования показали высокую эффективность предложенной технологии. Исследованные объекты были распределены в кластеры корректно во всех случаях. Предложенный метод позволяет уменьшить значение ошибки воспроизводимости, поскольку конечное решение по определению оптимальных параметров алгоритма кластеризации принимается на основе параллельного анализа результатов кластеризации, полученных на равномощных подмножествах данных с учетом разницы в результатах кластеризации, полученных на данных подмножествах.

КЛЮЧЕВЫЕ СЛОВА: объективная кластеризация, критерии качества кластеризации, индуктивное моделирование, алгоритм кластеризации DBSCAN.

© Babichev S., Vyshemyrska S., Lytvynenko V., 2019
DOI 10.15588/1607-3274-2019-1-8

ЛІТЕРАТУРА / ЛИТЕРАТУРА

1. Madala H. R. Inductive learning algorithms for complex systems modeling / H. R. Madala, A. G. Ivakhnenko. – CRC Press, 1994. – 365 p.
2. Soni N. Categorization of several clustering algorithms from different perspective: a review / N. Soni, A. Ganatra // International Journal of Advanced Research in Computer Science and Software Engineering. – 2012. – Vol. 2, Issue 8. – P. 63–68.
3. Stepashko V. Construction and research of the generalized iterative GMDH algorithm with active neurons / V. Stepashko, O. Bulgakova, V. Zosimov // Advances in Intelligent Systems and Computing II. – 2018. – P. 492–510. DOI: 10.1007/978-3-319-70581-1_35
4. Bulgakova O. Numerical study of the generalized iterative algorithm GIA GMDH with active neurons / O. Bulgakova, V. Stepashko, V. Zosimov // Proceedings of the 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies, 2017. – 1, art. no. 8098836. – P. 496–500. DOI: 10.1109/STC-CSIT.2017.8098836
5. Osypenko V. V. The methodology of inductive system analysis as a tool of engineering researches analytical planning / V. V. Osypenko, V. M. Reshetjuk // Ann. Warsaw Univ. Life Sci. – 2011. – SGGW. No. 58. – P. 67–71. [Electronic resource]. – Access mode: http://annals-wuls.sggw.pl/?q=node/234
6. Objective clustering inductive technology of gene expression sequences features / [S. Babichev, M. A. Taif, V. Lytvynenko, M. Korobchinskyi] // Communications in Computer and Information Science: In the book "Beyond Databases, Architectures and Structures", edited by S. Kozelski and D. Mrozek, – 2017. – P. 359–372.
7. Criterial analysis of gene expression sequences to create the objective clustering inductive technology // [S. Babichev, M. A. Taif, V. Lytvynenko, V. Osypenko] // Proceeding of the 2017 IEEE 37th International Conference on Electronics and Nanitechnology (ELNANO). – 2017. – P. 244–249.
8. Babichev S. Inductive model of data clustering based on the agglomerative hierarchical algorithm / S. Babichev, M. A. Taif, V. Lytvynenko // Proceeding of the 2016 IEEE First International Conference on Data Stream Mining and Processing (DSMP). – 2016. – P. 19–22. [Electronic resource]. – Access mode: http://ieeexplore.ieee.org/document/7583499/
9. Babichev S. Estimation of the inductive model of objects clustering stability based on the k-means algorithm for different levels of data noise / S. Babichev, M. A. Taif, V. Lytvynenko // Radio Electronics, Computer Science, Control. Zaporizhzhya Ukraine. – 2016. – № 4. – P. 54–60.
10. Puchala D. Joint compression and encryption of visual data using orthogonal parametric transforms / D. Puchala, M. M. Yatsymirskyy // Bulletin of the Polish Academy of Sciences-Technical Sciences. – 2016. – Vol. 64, Issue 2. – P. 373–382.
11. Single-frame image super-resolution based on singular square matrix operator / [Y. Rashkevych, D. Peleshko, O. Vynokurova et al.] // 1st IEEE Ukraine Conference on Electrical and Computer Engineering (UKRCON), MAY 29-JUN 02. – 2017. – P. 944–948.
12. Ivakhnenko A. Group method of data handling as competitor to the method of stochastic approximation / A. Ivakhnenko // Soviet Automatic Control. – 1968. – Vol. 3. – P. 64–78.
13. Calinski T. A dendrite method for cluster analysis / T. Calinski, J. Harabasz // Communication in Statistics. – 1974. – Vol. 3. –P. 1–27.
14. Zhao Q. Xu M. Sum-of-squares based cluster validity index and significance analysis / Q. Xu M. Zhao, P. Fränti // Proceeding of International Conference on Adaptive and Natural Computing Algorithms. – 2009. – P. 313–322.
15. Gene expression sequences clustering based on the internal and external clustering quality criteria / [S. Babichev, J. Krejci, J. Bicanek, V. Lytvynenko] // Proceedings of the 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies, 2017. – 1, art. no. 8098744, P. 91–94. DOI: 10.1109/STC-CSIT.2017.8098744.
16. Harrington J. The desirability function / J. Harrington // Industrial Quality Control, 1965. – Vol. 21(10). – P 494–498. [Electronic resource]. – Access mode: http://asq.org/qic/display-tem/?item=4860.
17. Ester M. A density-based algorithm for discovering clusters in large spatial databases / M. Ester, H. Kriegel, J. Sander // Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. – 1996. – P. 226–231.
18. A density-based algorithm for discovering clusters in large spatial databases with noise / [M. Ester, H. Kriegel, J. Sander, X. Xu] // KDD–1996 : proceedings. – 1996. – P. 226–231.
19. Density-based clustering / [H.-P. Kriegel, P. Kröger, J. Sander, A. Zimek] // Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. – 2011. – № 1(3). – P. 231–240.
20. [Electronic resource]. – Access mode: https://www.knime.com/ knime-software/knime-analytics-platform
21. Ihaka R. R: a linguage for data analysis and graphics / R. Ihaka, R. Gentleman. // Journal of Computational and Graphical Statistics. – 1996. – Vol. 5(3). – P. 299–314.
22. Gionis A. Clustering aggregation / A. Gionis, H. Mannila, P. Tsaparas // ACM Transactions on Knowledge Discovery from Data (TKDD). – 2007. – Vol. 1(1). – P. 1–30.
23. Zahn C. T. Graph-theoretical methods for detecting and describing gestalt clusters / C. T. Zahn // IEEE Transactions on Computers. – 1971. – Vol. 100(1). – P. 68–86.
24. [Electronic resource]. – Access mode: http://www.sthda.com/english/rpkgs/factoextra
25. Jain A. Data clustering: A user's dilemma / A. Jain, M. Law // Lecture Notes in Computer Science. – 2005. – Vol. 3776. – P. 1–10.
26. Fisher R. A. The Use of Multiple Measurements in Taxonomic Problems / R. A. Fisher // Annals of Eugenics. – 1936. – Vol. 7. – P. 179–188.
27. Gene-expression profiles predict survival of patients with lung adenocarcinoma / [D. G. Beer, S. L. Kardia and al.] // Nature Medicine. – 2002. – Vol. 8(8). – P. 816–824.

© Babichev S., Vyshemyrska S., Lytvynenko V., 2019
DOI 10.15588/1607-3274-2019-1-8