7. Tan, H. L. A Perceptually Relevant MSE-Based Image Quality Metric [Text] / H. L. Tan, Z. Li, Y. H. Tan, S. Rahardja, C. Yeo // IEEE Transactions on Image Processing. – 2013. Vol. 22, № 11. – P. 4447–4459. doi:10.1109/tip.2013.2273671

8. Huynh-Thu, Q. Scope of validity of PSNR in image/video quality assessment [Text] / Q. Huynh-Thu, M. Ghanbari // Electronics Letters. – 2008. – Vol. 44, № 13. – P. 800–801. doi:10.1049/el:20080522

9. Hum, Y. C. Multiobjectives bihistogram equalization for image contrast enhancement [Text] / Y. C. Hum, K. W. Lai, M. I. Mohamad Salim // Complexity. – 2014. – Vol. 20, № 2. – P. 22–36. doi:10.1002/cplx.21499

10. Абламейко, С. В. Обработка изображений: технология, методы, применение [Текст] / С. В. Абламейко, Д. М. Лагуновский. – Минск: Амалфея, 2000. – 304 с.

11. Колмогоров, А. Н. Элементы теории функций и функционального анализа [Текст] / А. Н. Колмогоров, С. В. Фомин. – Москва: Наука, 1976. – 544 с.

12. Wang, Z. A universal image quality index [Text] / Z. Wang, A. C. Bovik // IEEE Signal Processing Letters. – 2002. – Vol. 9, № 3. – P. 81–84. doi:10.1109/97.995823

13. Красильников, Н. Н. Цифровая обработка 2D- и 3D-изображений [Текст] / Н. Н. Красильников. – СПб.: БХВ-Петербург, 2011. – 608 с.

Представлена верифікація та аналіз алгоритму попередньої кластеризації, зокрема його основного елементу – правила прийняття рішення. Цей алгоритм, на відміну від інших, не використовує початкову інформацію про кількість кластерів. Верифікація полягала в тестуванні правила прийняття рішення відповідно до кожного окремого випадку вхідних даних. Представлено переваги та недоліки алгоритму попередньої кластеризації

Ключові слова: кластеризація даних, кластер, верифікація, емпіричне правило, прийняття рішення

Представлена верификация и анализ алгоритма предварительной кластеризации, в частности его основного элемента – правила принятия решения. Этот алгоритм, в отличие от других, не использует исходную информацию о количестве кластеров. Верификация состояла в тестировании правила принятия решения в соответствии каждому отдельному случаю входных данных. Представлены преимущества и недостатки алгоритма предварительной кластеризации

Ключевые слова: кластеризация данных, кластер, верификация, эмпирическое правило, принятие решения

ROBUST VERIFICATION AND ANALYSIS OF THE PRE-CLUSTERING ALGORITHM WITH A-PRIORI NON-SPECIFICATION OF THE NUMBER OF CLUSTERS

V. Mosorov
Doctor of Technical Sciences*
E-mail: volodymyr.mosorov@p.lodz.pl

T. Panskyi
PhD student*
E-mail: panskyy@gmail.com

S. Biedron
PhD student*
E-mail: SBiedron@wpia.uni.lodz.pl
*Institute of Applied Computer Science
Lodz University of Technology
Stefanowskiego str., 18/22, Lodz, Poland, 90-924

1. Introduction

The range of the implementation of cluster analysis is wide, it extends from many technical applications to different branches of science, such as biology, medicine, computer sciences and psychology. The main purpose of the cluster analysis is dividing the investigated objects into homogeneous groups, or clusters, according to certain criteria and investigating the process of natural grouping of these objects. It means solving the task of grouping data and revealing in them a relevant structure. The task of clustering can be defined as follows: taking into account the information about n objects, find K groups based on the measure of similarity, so that the similarity among the object inside one group might be strong, while the similarity among the objects of different groups might be weak.

The presence of noises in the input data makes the reveal of clusters much more difficult. Noise is considered to be the outliers that do not ingress into any cluster and are located at the considerable distance from other objects. In practice, the cluster is a subjective bunch of objects, the analysis of which requires some specific knowledge. Using cluster analysis, the investigator aims to reveal data structure, that is, interconnection of parts of the whole, the inner construc-

tion of the subject of research. At the same time, the cluster analysis implies including the structure into analyzed data. Clusterization can lead to the appearance of the artifacts (finding structures in the data, which do not contain any structures). The main purpose of this article is answering the question whether input data are structured, that is, whether the total clustering procedure is necessary, or there is no interconnection among the input data.

2. Analysis of published clustering techniques and problem statement

According to their nature, humans are excellent cluster searchers in two-dimensional space, but it is necessary to automatize cluster search algorithms for two-dimensional and three-dimensional data. This challenge as well as unknown number of input data clusters has lead to appearance of hundreds of clustering algorithms, which have already been published and still continue to appear. In the case of clusterization with taking into consideration distances, the peculiarity of the research is that the objects in the single cluster are located at a short distance from one another, while objects from different clusters are located at a long distance from one another. We can subdivide the clustering algorithms into two groups concerning two radically different strategies [1, 2].

Hierarchical algorithms serve to hierarchical decomposition of the objects. Such algorithms are subdivided into *agglomerative* (bottom-up) and *divisive* (top-down) ones.

A. The agglomerative algorithms begin their operation from each object, supposing that this object is a separate cluster and sequentially unifying the objects into groups according to the distance function. Clusterization can stop if all objects are in a separate (individual) group or a user decides so, or further clustering results cause the appearance of some undesirable clusters. For example, clustering can stop, when some given number of clusters is achieved, or when using the measure of cluster compactness we abandon building a cluster from two small ones, if objects of the resulting cluster can be spread on the too wide area.

B. Through the divisive algorithms another strategy is realized. Their activity begins from unifying all objects into one group and sequentially subdividing groups into smaller ones till every object will be put into one cluster or the process will be stopped by the user. The divisive algorithms divide the objects into independent groups at every step. The objects are considered according to some order, and every object is assigned to the most appropriate cluster.

The clusterization allows unifying or dividing clusters, or supposing some objects to belong to none of the clusters found (anomalies, isolated points, noise etc.). The drawbacks of the hierarchical algorithms are the necessity of determining threshold values and choosing the measure of cluster proximity.

At the great number of observations the hierarchical methods of cluster analysis are inapplicable. In such cases non-hierarchical methods are applied. They are based on the division and are the iterative methods for the division of output package. At this process new clusters are formed till the rule of stopping will not be applied. There two approaches there. The first one determines the boundaries of the clusters as the densest areas in the multidimensional space of output data, that is, determines the cluster at the place with large

"gathering of objects". The second approach provides the minimization of the level of difference between the objects. This method has the advantage of being more noise-resistant. The disadvantages of the method are the necessity of determining the clusterization parameters in advance, including the number of clusters, as well as the number of iterations or the rule of stopping etc.

Apart two main categories of the hierarchical and non-hierarchical clusterization algorithms, many other methods have been published. They concern certain problems or serve for certain data sets. They include:

1. Density-based clustering. These algorithms unify objects according to the specific density of an objective function. Density is usually determined as the number of the objects in the certain data domain. Cluster grows till the number of elements in certain domain does not exceed the parameter set by the user [3]. This group of methods precisely detects clusters with high dense object aggregation, but these methods are not effective at the analysis of fuzzy groups of objects (for example, at the uniform distribution of objects).

2. Model-based clustering. These algorithms give good approximations of the model parameters, which do for the data best. They can be both hierarchical and non-hierarchical depending on the structure, model, data set, or for better breakdown determining [4]. Main drawbacks are finding the initial distribution parameters, and setting the appropriate model which is user dependence.

3. Categorical data clustering. These algorithms are developed specifically for the data, to which the Euclidean distance or other numeric measurement of a distance cannot be applied. In the literature these approaches are like the hierarchical and non-hierarchical methods [5].

4. Grid-based clustering. These algorithms are in general used for spatial data. Their aim is to quantize the data set and determine the number of cells, and later to work with the objects belonging to these cells [6]. One of the drawbacks of grid algorithms is the strong dependence of the quality of detected clusters on cell dimensions.

The unsupervised learning methods (clusterization) as opposed to the supervised learning methods (classification), marks of output objects, that is, determining each object belonging to the certain cluster, as well as the number of the clusters are not given from the very beginning of the process. The created clustering algorithm without a-priori information about the number of the clusters belongs to the group of pre-clustering algorithms. Pre-clustering is the procedure of checking the possibility of clustering the input data. Checking this possibility answers the question whether data can be divided into more than one cluster. A well-known unsupervised pre-clustering algorithm is a canopy clustering algorithm, presented by [7]. It is often used for the preliminary analysis of input data or for primary clusterization for the k-means algorithm or hierarchical clustering algorithm. The canopy clustering algorithm is intended for speeding up the clusterization of big data arrays, where the use of other algorithms causes incorrect results. The aim of this method is finding the approximate number of the clusters, which make up the input information for other clustering algorithms (for example, k-means algorithm). The disadvantage of this pre-clustering algorithm is the heuristic definition of two threshold values (distances) T_1 and T_2.

Determining the number of clusters in data set is a famous and popular procedure. The correct choice of k is

often ambiguous, with interpretations depending on the shape and scale of the distribution of points in a data set and the desired clustering resolution of the user. Some algorithms use partial information about data set, others based on multiple parameter changes allow to estimate best division of the database, with optimal k (Elbow method, AIC, DIC, BIC, silhouette method, etc.), but there is no clear and simple algorithm that would without the prior information about the data set, and without user help correctly identify the number of clusters [8, 9].

The clustering algorithm created by [10], as opposed to other existing algorithms, does not require the determination of input parameters, threshold values for correct determination of the number of the clusters. So, this pre-clustering algorithm has been chosen as a privileged one from the whole totality of the clustering algorithms for the primary analysis of investigated input data. The principle of operation of the created algorithm is described below.

3. Purpose and objectives of the study

The key purpose of this paper is verification and analysis of pre-clustering algorithm with the calculated criteria of decision making.

In accordance with the set goal the following research objectives are identified:

1. Testing the pre-clustering algorithm for the selected cases of input data.

2. Upgrading of decision rule according to each case of input data.

3. Modification of decision rule taking into account not only Euclidian distances, but also input data parameters (standard deviation and mathematical expectation).

4. Description of analyzed algorithm

The presented pre-clustering algorithm together with calculated criteria of decision making determines the existence of one or two clusters in the input data array.

The following assumption about cluster existence was made:

a) input array can include two clusters K_1 and K_2 accordingly;

b) input array is a single cluster.

For making the decision if one or two clusters exist, forced c-means clustering was performed. Forced clustering is always performed for dividing the analyzed array into two clusters, even if the input array is a single group of objects. Only on the basis of such forced division of the input array the decision is made about the existence of one or two clusters.

After the forced division of the input data array, to found the essential criteria, the calculation of average distance between the objects in the found clusters $d_1(K_1)$ and $d_2(K_2)$, clusters K_1 and K_2 accordingly, and the average distance between the objects $d_{12}(K_1 \cup K_2)$ of unification of two clusters $(K_1 \cup K_2)$ was carried out with the use of the Euclidean distance in the 2D space [11]. This distance is calculated by the known formula $d(a_i, b_j) = \sqrt{(a_x^i - b_x^j)^2 + (a_y^i - b_y^j)^2}$, Euclidean distance between the objects in the 2D space, where

$a = (a_x, a_y)$, $b = (b_x, b_y)$, are two objects in the Euclidean 2D space. After comparing the average distances of the clusters $d_1(K_1)$ and $d_2(K_2)$ with the average distance of the unified cluster $d_{12}(K_1 \cup K_2)$, the strict rule of thumb for decision making was determined:

$$\begin{cases} \text{if} \quad d_{12}(K_1 \cup K_2) > d_1(K_1) + d_2(K_2), \text{ Analyzed array includes two clusters,} \\ \\ \text{otherwise} \qquad\qquad \text{Analyzed data array is a single cluster.} \end{cases} \quad (1)$$

Forced c-means clustering can be replaced by another clustering algorithm (k-means algorithm or algorithms which are based on the density distribution of input data). But the use of the clustering algorithm without the decision making criteria does not always cause the correct determination of the number of the clusters. The invalid primary determination of the number of the clusters for c-means or k-means clustering algorithms can be the example of the case mentioned above. The presented clustering algorithm uses c-means clustering only for rough data estimation, but the decision making criteria are the main components responsible for the correctness and adequacy of the cluster determination.

In spite of all advantages of the created pre-clustering algorithm, it is necessary to mention some of its drawbacks. The first one is a priori decision whether the input data make up one or two clusters. This rule is not ideal when clusters are located near one another, so that it is difficult to suppose if one big cluster or two smaller ones exist. Another disadvantage is the determination of the Euclidean distance between the objects of the clusters taking into account only constant dispersion, and not considering the change of the dispersion and median according to the certain case of input data.

5. Verification of algorithm parameters

The process of verification of the parameters of the pre-clustering algorithm concerns testing the rule of decision making for different types of input data. The selected cases of input data considered here are the input data with the normal distribution law having been grouped into one or two clusters. Because of impossibility to wholly analyze all possible cases, some simplifying condition for the number of objects and distribution parameters were introduced.

Simplification 1: the number of the objects in every case did not exceed $n \le 60$.

Simplification 2: the standard deviation of input data with the normal distribution law did not exceed $\sigma \le 8$.

Simplification 3: all object groups (gatherings) have a globular form.
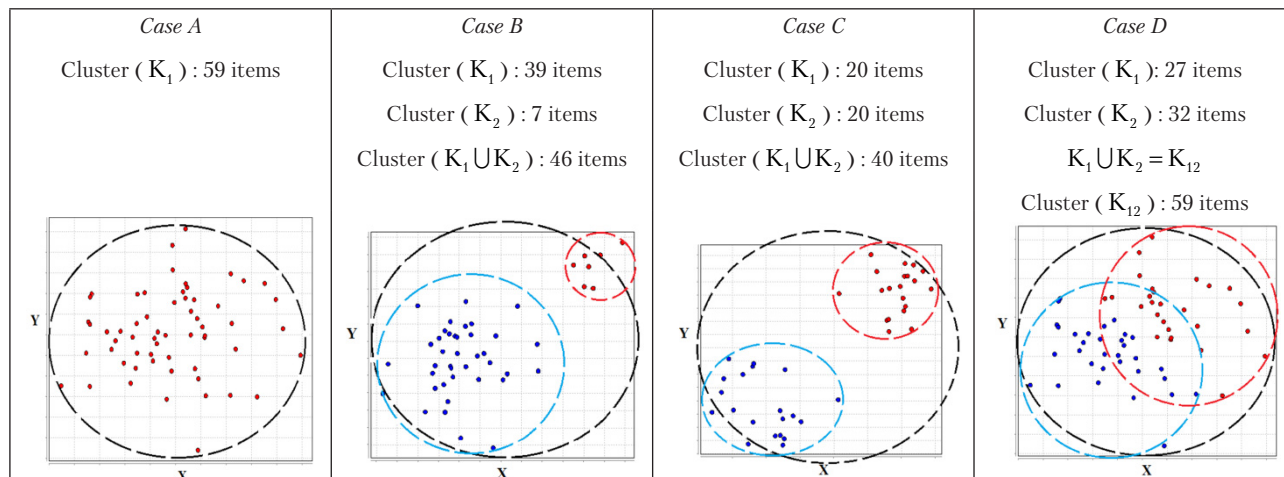
The series of tests were performed for such cases:

A. Test input data make up a single cluster, they are analyzed as one set and are not subdivided into smaller ones;

B. Test input data make up two clusters, one cluster being much bigger than another one;

C. Test input data make up two equal clusters;

D. Test input data make up a single cluster, being divided into two symmetrical subsets each of them being analyzed as a single cluster.

The examples of all mentioned above cases of input data can be shown in Table 1.

Table 1

The examples of input data cases

Case A	Case B	Case C	Case D
Cluster (K_1) : 59 items	Cluster (K_1) : 39 items	Cluster (K_1) : 20 items	Cluster (K_1): 27 items
	Cluster (K_2) : 7 items	Cluster (K_2) : 20 items	Cluster (K_2) : 32 items
	Cluster ($K_1 \bigcup K_2$) : 46 items	Cluster ($K_1 \bigcup K_2$) : 40 items	$K_1 \bigcup K_2 = K_{12}$
			Cluster (K_{12}) : 59 items

The next step after introducing primary simplifications is determining the dependences between the average distances (1) and statistic parameters of input data, such as standard deviation and mathematical expectation.

The size of the groups of objects depended on the case and series of the test and deviated within the range of 10 to 60 objects. The change of standard deviation was $1 \le \sigma \le 8$. The detailed analysis of each case is described below.

Case A. The average distance of the cluster $d_1(K_1)$ depends not only on the distance between all objects of the data array, but also on the change of standard deviation according to the test series. It can be estimated with the use of the formula (2).

$$d_1(K_1) \approx k_1 \sigma_1, \qquad (2)$$

where $k_1 \approx 1.6 \dots 1.9$ is an arbitrage coefficient, σ_1 is the standard deviation of the cluster K_1.

As we can conclude from formula (2), the value of the average distance depends only on the coefficient chosen and does not depend on the size and mathematical expectation of the cluster. The value of the coefficient k_1 is found statistically from ten series of the test data.

Case B. In this case the average distance $d_1(K_1)$ and $d_2(K_2)$ for every cluster can be determined by formula (2). The average distance $d_{12}(K_1 \bigcup K_2)$ for the unified cluster is determined by formula (3).

$$d_{12}(K_1 \bigcup K_2) \approx \frac{\Delta m}{q} \cdot \sigma_1^{-0.2 \ln(\Delta m)} \cdot \sigma_1, \qquad (3)$$

where $q = \begin{cases} 1.6 & \text{if } \Delta m < 5, \\ 2.1 & \text{if } 5 \le \Delta m \le 15, \\ 2.6 & \text{if } \Delta m > 15, \end{cases}$

where $\Delta m = |m_1 - m_2|$ is the difference of mathematical expectations of two clusters.

In this case only a standard deviation σ_1 of a bigger cluster K_1 is used. A standard deviation σ_2 is neglected because of a small size and influence of the cluster K_2 on the average distance of the unified cluster $d_{12}(K_1 \bigcup K_2)$.

Case C. The average distance for clusters $d_1(K_1)$ and $d_2(K_2)$ is separately determined by formula (2). The aver-

age distance for the unified cluster $d_{12}(K_1 \bigcup K_2)$ was determined by formula (4).

$$d_{12}(K_1 \bigcup K_2) \approx 0.8 \Delta m \left(\frac{\sigma_1 + \sigma_2}{2} \right)^p, \qquad (4)$$

where $p = \begin{cases} 0.2 & \text{if } \Delta m < 10, \\ 0.1 & \text{if } 10 \le \Delta m \le 25, \\ 0.05 & \text{if } \Delta m > 25. \end{cases}$

The standard deviations and σ_2 of the clusters K_1 and K_2 accordingly are used in this case. Both clusters are of equal size, and this causes their considerable influence on the unified cluster $d_{12}(K_1 \bigcup K_2)$, and the standard deviation of any cluster cannot be neglected, as it was done in the previous case.

Case D. In this case the average distance of the unified cluster is determined by formula (5).

$$d_{12}(K_1 \bigcup K_2) \approx k_1 \sigma_1. \qquad (5)$$

Formula (5) for the average distance of the unified cluster $d_{12}(K_1 \bigcup K_2)$ is similar to formula (2) of the *Case A* considering the single cluster $d_1(K_1)$. The average distance for two divided symmetric clusters is determined by formula (6).

$$d_1^1(K_1) \approx d_2^2(K_2) \approx 3 \sigma_i^j, \qquad (6)$$

where $\sigma_i^j \approx \frac{\sigma_1}{2}, \quad i = 1, 2, \quad j = 1, 2,$

where $d_1^1(K_1)$ and $d_2^2(K_2)$ are average distances of two symmetric clusters, σ_1^1, and σ_2^2 are standard deviations of two clusters calculated by the formulae for the parameters of truncated normal distribution law. From the properties of symmetry we can conclude that the average distance for both divided clusters is about equal.

Verified parameters of the decision making rule allows us to analyze and estimate input data in accordance to the certain case. Formulae (2)–(6) for average distances take into account

the parameters of input data distribution, which are not considered in rule (1). So, in any single case the calculation of the average distances in the groups of objects can be replaced by obtained formulae (2)–(6), input parameters being the standard deviation and mathematical expectation of the data.

6. Experimental results of modified decision rule

Cases A and *D* are alike, because the input data made up gathering the objects into a single group. In the *case A* the whole data array is considered as a single cluster and its average distance is calculated according to formula (2), but in the *case D* one general group of objects is forced to be divided into two clusters. Hence, at equal initial data it can be assumed that $d_1(K_1)$ (*Case A*)$= d_{12}(K_{12})$ (*Case D*). That is, the average distance of the data array which is considered to be the single cluster (*Case A*) and the identical array divided into two clusters and being then unified (*Case D*) is equal. The average distances of the clusters are linearly ramp values, as it is shown in Fig. 1.
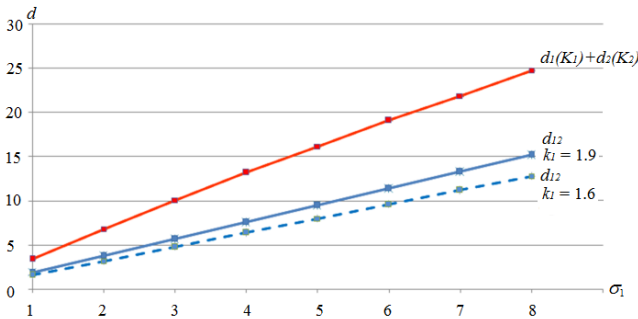

Fig. 1. Decision making rule for the *cases A, D*

As it can be seen in Fig. 1, according to the decision making rule if $d_1(K_1)+d_2(K_2) > d_{12}(K_{12})$, there is always one cluster in the given array.

In *cases B* and *C*, when the input array is forced to be divided into two clusters, one of which is much bigger than the other (*case B*), or when the array is divided into two equal clusters (*case C*), the conclusion about the existence of one or two clusters is not such definite, as in the previous case. This conclusion depends on the measure of the proximity of two clusters, that is, on the difference of mathematical expectations Δm. In the *case B* it depends on the standard deviation of a bigger cluster σ_1, and in the *case C* on the standard deviations of both clusters σ_1 and σ_2. The less Δm is and the bigger is σ_1 (for the *case B*) or σ_1 and σ_2 (for the *case C*), the bigger the probability of the existence of one cluster is. But when Δm increases and σ_1, σ_2 decrease, the existence of two clusters will be evident. The average distance of the unified cluster in the *case B* does not depend on the standard deviation of a smaller cluster σ_2, but it cannot be neglected in the *case C*. The illustration of the decision making rule for the *cases B* and *C* are shown in Fig. 2, 3.

As we can see in Fig. 2 and 3, at small values of the difference of mathematical expectations $\Delta m < 10$ and at small standard deviations of the clusters K_1 and K_2 the input data array can be considered as a single cluster. But at the increase in Δm, that is, at moving away the groups of objects, according to the rule (1) the conclusion can be drawn that two clusters exist.
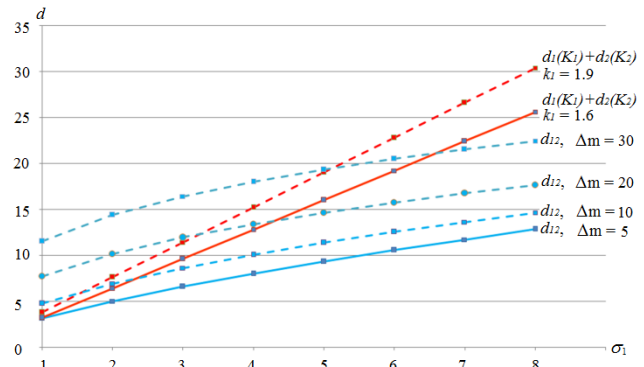

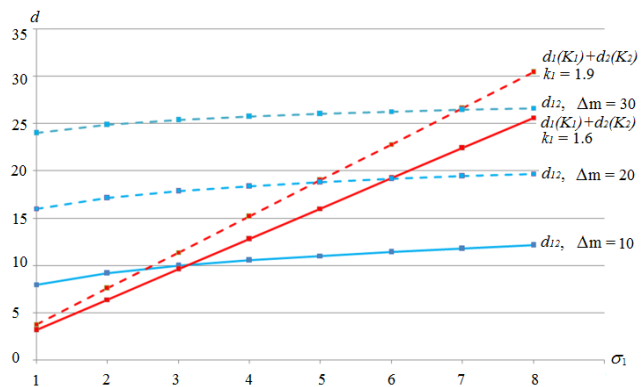Fig. 2. Decision making rule for the *case B*

Fig. 3. Decision making rule for the *case C*

6. Conclusions

In this article the verification of the pre-clustering algorithm, notably, its core element, which is the decision making rule has been made. This algorithm, as opposed to other algorithms, does not use the initial information about the number of clusters. Verification was performed as testing the decision making rule according to certain cases of input data. The parameters of the decision making rule were modified, and it allowed using not only average distances as the main criterion, but also the parameters of input data (standard deviation and mathematical expectation) for efficient algorithm operation. The advantage of the presented decision making rule is the possibility of choosing cases where clustering is to be performed, and where it is not necessary (data do not have evident structure). The drawback of this rule is a firm result dependence on the calculated average distances, which is also a disadvantage of this type of clustering algorithms [12]. The analyzed algorithm is sensitive to noise and to single random isolated objects, which can change the values of calculated distances and cause erroneous decisions.

The next stage of the research will be processing and applying the clusterization based on object density and comparing it with the created pre-clustering algorithm.

References

1. Han, J. Data Mining: Concepts and Techniques [Text] / J. Han, M. Kamber. – Ed. 2. – Morgan Kaufmann Publishers, 2006. – 703 p. – ISSN 1-55860-901-6.

2. Yan, M. Methods of Determining the Number of Clusters in a Data Set and a New Clustering Criterion [Text]: Ph. D thesis / M. Yan. – Blacksburg, Virginia, 2005. – 120 p.

3. Pérez-Suárez, A. An algorithm based on density and compactness for dynamic overlapping clustering [Text] / A. Pérez-Suárez, J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa, J. E. Medina-Pagola // Pattern Recognition. – 2013. – Vol. 46, № 11. – P. 3040–3055. doi:10.1016/j.patcog.2013.03.022

4. HaiJiang, S. S. Model-based clustering [Text] / S. S. HaiJiang. – Ontario, Canada: University of Waterloo, 2005. – 61 p.

5. Dutta, M. QROCK: A quick version of the ROCK algorithm for clustering of categorical data [Text] / M. Dutta, A. K. Mahanta, A. K. Pujari // Pattern Recognition Letters. – 2005. – Vol. 26, № 15. – P. 2364–2373. doi:10.1016/j.patrec.2005.04.008

6. Schikuta, E. Grid-clustering: an efficient hierarchical clustering method for very large data sets [Text] / E. Schikuta // Proceedings of 13th International Conference on Pattern Recognition. – 1996. – Vol. 2. – P. 101–105. doi:10.1109/icpr.1996.546732

7. McCallum, A. Efficient clustering of high-dimensional data sets with application to reference matching [Text] / A. McCallum, K. Nigam, L. H. Ungar // Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining – KDD '00. – Association for Computing Machinery (ACM), 2000. – P. 169–178. doi:10.1145/347090.347123

8. Goutte, C. Feature-space clustering for fMRI meta-analysis [Text] / C. Goutte, L. K. Hansen, M. G. Liptrot, E. Rostrup // Human Brain Mapping. – 2001. – Vol. 13, № 3. – P. 165–183. doi:10.1002/hbm.1031

9. Hofmann, M. RapidMiner: Data Mining Use Cases and Business Analytics Applications [Text] / M. Hofman, R. Klinkenberg. – Chapman & Hall/CRC, 2013. – 431 p. – ISBN:1482205491.

10. Mosorov, V. Image Texture Defect Detection Method Using Fuzzy C-Means Clustering for Visual Inspection Systems [Text] / V. Mosorov, L. Tomczak // Arabian Journal for Science and Engineering. – 2014. – Vol. 39, № 4. – P. 3013–3022. doi:10.1007/s13369-013-0920-7

11. Sisodia, D. Clustering Techniques: A Brief Survey of Different Clustering Algorithms [Text] / D. Sisodia, L. Singh, S. Sisodia, K. Saxena // International Journal of Latest Trends in Engineering and Technology (IJLTET). – 2012. – Vol. 1, № 3. – P. 82–87.

12. Qian, W. Analyzing popular clustering algorithms from different viewpoints [Text] / W. Qian, A. Zhou // Journal of Software. – 2002. – Vol. 13, № 18. – P. 1383–1394.