

9. Киселев, А. С. Прибор для риносинусопневмометрии [Текст] / А. С. Киселев, К. В. Герасимов // Журн. ушных, носовых и горловых болезней. – 1990. – № 5. – С. 76–77.
10. Гарюк, О. Г. Поведение давления воздуха в верхнечелюстной пазухе в норме [Текст] / О. Г. Гарюк, А. Ю. Меркулов, А. С. Нечипоренко, А. В. Новак // Международный научно-практический журнал «Отоларингология. Восточная Европа». – 2013. – № 3(12). – С. 23–27.
11. Захаров, И. П. Оценивание неопределённости измерений для дифференциальной функции [Текст] / И. П. Захаров, Е. А. Климова, О. О. Волков, Ю. Г. Жарко // Метрология та прилади. – 2014. – № 1(45). – С. 78–80.
12. Guide to the Expression of Uncertainty in Measurement [Text]. – Geneva: ISO, First Edition, 1995. – 101 p.
13. Yerokhin, A. Hardware-software complex for biomedical measurement of differential pressure in the maxillary sinus [Text] / A. Yerokhin, I. Zakharov, A. Nechiporenko, O. Garyuk // Proceedings of the symposium 24th national scientific symposium with international participation, 2014. – P. 290–294.

Запропоновано нове правило прийняття рішення, яке є модифікованою альтернативною стандартною в алгоритмі попередньої кластеризації. Дане правило було перевірене на експериментальних даних, а результати були порівняні із результатами, отриманими із використанням критерію сферичної роздільності. Представлені переваги та недоліки модифікованого правила прийняття рішення

Ключові слова: модифіковане правило прийняття рішення, алгоритм попередньої кластеризації

Предложено новое правило принятия решения, которое является модифицированной альтернативной стандартной в алгоритме предварительной кластеризации. Данное правило было проверено на экспериментальных данных, а результаты были сравнены с результатами, полученными с использованием критерия сферической раздельности. Представлены преимущества и недостатки модифицированного правила принятия решения

Ключевые слова: модифицированное правило принятия решения, алгоритм предварительной кластеризации

UDK 004.9

DOI: 10.15587/1729-4061.2015.51214

ANALYSIS OF THE MODIFIED ALTERNATIVE DECISION RULE IN THE PRECLUSTERING ALGORITHM

V. Mosorov

Doctor of Technical Science
Department of Computer Science in Economics
University of Lodz
Narutowicha str., 65, Lodz, Poland, 90-131
E-mail: wmosorow@uni.lodz.pl

T. Panskyi

Postgraduate student*
E-mail: panskyi@gmail.com

S. Biedron

Postgraduate 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

Clustering analysis or simply clustering is a process of dividing a set of data objects into two or more subsets in such a way that objects in one subset (cluster) are characterized by a high degree of similarity, but differ from objects in other clusters. The concept and application of clustering is quite wide, they have been described repeatedly in various literature sources. So, it seems reasonable to omit well-known features of cluster analysis, its application in different fields of science and technology [1] and the description of popular clustering algorithms [2], and focus on a preclustering algorithm.

2. Analysis of published data and problem statement

The most known preclustering algorithms require a user setting of certain input parameters, one of the examples is

a canopy clustering algorithm, presented by [3]. 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 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) T1 and T2. Another example is the usage of a BIRCH pre-clustering algorithm [4]. This algorithm is an efficient data reduction method in the case of large data sets. However, BIRCH requires the set of the optimization key parameters (like branching factor, quality threshold and selection of the separator line). Some clustering algorithms are part of already created algorithms and make up its preprocessing step [5]. For example, an algorithm for preprocess k-means clustering. The preprocessed k-means requires a lower number of iterations and produces very accurate

clusters in large number of data sets, but still it requires the initial parameter to be set.

Another example of analysis the adequate choice of input parameters is the usage of the validation criteria.

Clustering belongs to unsupervised learning methods, so it does not require a priori information about investigated data set. However, to obtain valid results, it is necessary to set an input parameter. For example, k-means and CURE algorithms require setting the number of clusters as an input parameter. In this context the question what number of clusters is optimal comes into existence. Today the use of validation criteria is the means of answering the posed question. In general, two types of clustering techniques are marked out: one of them uses external criteria, the other uses internal ones.

External validation criteria are based on some partial information about the input data set. Some of principal external criteria are listed below [6–8]:

- Rand index is a measure of similarity between two data clusterings. This index is a ratio of the sum of a true positive value and a true negative one to the sum of a true positive, false positive, true negative and false negative value.

- Jaccard's coefficient (or Jaccard similarity coefficient) is a statistic measure for comparing the similarity and diversity of input data sets. Jaccard's coefficient measures similarity between finite sample sets and is defined as the size of the intersection divided by the size of the union of the sample sets.

- Fowlkes-Mallows index is an external evaluation criterion that is used to determine the similarity between two clusterings (for clusters obtained after a clustering algorithm). A higher value of the Fowlkes–Mallows index indicates a greater similarity between the clusters.

- Dunn index (DI) measures the ratio between the minimal intracluster distance (distance between two sets, or the minimum distance between two points taken from different sets) to maximal intercluster distance (maximum distance between two points in the set).

- SD validity index is determined on the basis of the average scattering of clusters and total separation of clusters.

Real situations and practical tasks not always open up possibilities of getting a priori information about input data. Therefore, external criteria are seldom applied at primary analysis of data clustering. They are used together with internal criteria at last stages of clustering analysis as auxiliary criteria for correct decision making.

Internal validation criteria are based on the information inside a cluster and on the manner of arrangement of objects according to this information. The correct choice of internal criteria causes clustering at which all objects inside the cluster are located close to one another and, in addition, clusters are well divided. Some of the internal criteria are listed below [9–11]:

- Davies-Bouldin's (DB's) Index measures the average similarity between all groups of objects and finds the most similar clusters. This index considers each cluster individually. For each cluster it determines which other cluster has maximum ratio of the mean inter cluster distance for objects located in two clusters to the distance between clusters. The smaller this value is, the more dense clusters are and the further they are located from one another.

- Cluster density is a measure which considers each cluster individually and determines a mean distance between all pairs of objects in the cluster multiplying it by the number

of objects in the cluster. When density value of the cluster tends to zero, the number of clusters is large, but low density values mean that clusters become denser.

- Sum of squares of objects scatter considers each cluster individually and divides the number of objects in the cluster by the total number of objects in all clusters. This value is squared and then these values for all clusters are added up. If input data represent one large cluster, this value is close to 1. If sizes of all clusters are equal, this value equals $1/n$, where n is the number of clusters.

- Average centroid distance considers each cluster individually and calculates average distance between every object in the cluster and the centroid. If clusters are dense, this value decreases.

In spite of the versatility of proposed external and internal validation criteria at the clustering analysis and their advantages, it is necessary to mention that they also have some disadvantages. As it was mentioned above, internal validation criteria do not require a priori information about input data. However, the use of one criterion will not cause absolute reliability of clustering results. So, it is advisable to use as more criteria as possible for increasing the reliability of the clustering results. The majority of validation criteria are based on a multiple choice and the change of input clustering parameters (for example, the number of the clusters) and on the choice of the most optimal input parameter. One of the disadvantages is the dependence of the clustering results on a user, since even if the criteria for result validation are used, the input parameters are likely to be chosen erroneously.

The concept of preclustering concerns the elimination of user influence on the clustering results. The preclustering algorithm proposes the possibilities of “artificial intelligence”, that is the determination of the number of clusters in the input data set without a priori information about input data and without additional means of checking (for example, repeated multiple testing or the choice of optimal plausibility criterion).

The main task of the analysis of input data is an answer to the question whether it is necessary to perform data clustering or input data have no inner structure and the clustering process will result not in its revealing but in occurrence of artifacts (artificial structures). The preclustering algorithm allows us to analyze and evaluate input data and decide whether input data represent a single cluster which does not need further clustering, or two different clusters which can be identify by a further clustering process.

Preclustering is a procedure of detecting the possibility of input data clustering. The preclustering algorithm forces the division of input data set into two preclusters. The precluster is a group of objects which is not a single cluster, but can become one after checking. To decide whether a given precluster is a single cluster or a part of a bigger cluster, the preclustering algorithm has been used. After the forced division of the input data set, the empirical decision rule of the preclustering algorithm makes use of average distances between objects in found preclusters $d(K_1)$ and $d(K_2)$, between preclusters K_1 and K_2 accordingly and average distances between objects $d(K)$ of the whole input data set K using the Euclidean distance in the 2D space [12]. The decision rule evaluates the possibility of the precluster to be a cluster.

The preclustering algorithm as opposed to other existed algorithms does not require a priori information about cluster location and about additional means of control (as,

for example, threshold meanings or measures of object similarity) for correct detecting the number of clusters. This preclustering algorithm is multipurpose and promising for a primary analysis of investigated input data.

The universality of the preclustering algorithm can be explained by the ability of using all kinds of numerical attributes, that is, measured numerical quantities produced as integral or real values. On the other hand, the universality is achieved by the possibility for applying this algorithm to the majority of continuous distribution laws (normal distribution, truncated normal distribution, Student's t-distribution, uniform distribution, Weibull distribution and others).

In spite of advantages of the preclustering algorithm and the simplicity of its decision rule, it also has some disadvantages which cause introduction of some limitations for the correct detecting the number of clusters. The main disadvantage of this rule is the dependence of the results on the calculated average distances. If clusters include isolated objects or anomalies (single objects located at a large distance from other cluster objects), the results of calculating average distances become strongly dependent on these objects which results in wrong decision making. This disadvantage strongly influences on the decision making particularly in cases when input data set is not infinitely large and includes the limited number of objects (for example, less than 100).

For eliminating the strong influence of isolated objects on decision making the modification of the existed rule is proposed.

3. Purpose and objectives of the study

The main objective of this publication is to present the modified decision rule for the preclustering algorithm.

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

1. Analysis of the modified decision rule of preclustering algorithm.
2. Testing of the decision rule and its comparison with the criterion of spherical resolution.

4. Modified decision rule

The idea of preclustering algorithm represented in the form block scheme in Fig. 1.

The disadvantage of the decision rule is the strong dependence of calculated distances on the nature of input data set. If input data are spherical, the density of the objects corresponds to the normal distribution law and the standard object deviation is small. In this case the decision rule will detect the correct number of clusters. But if the shape of the input data is arbitrary and they contain anomalies, the decision rule may work erratically.

For decreasing the influence of the factors mentioned above on the decision rule the modification of the rule is proposed. It can be performed by replacement of the calculation of mean distances in a precluster by the mean distances from the center of the precluster to all objects in the chosen precluster.

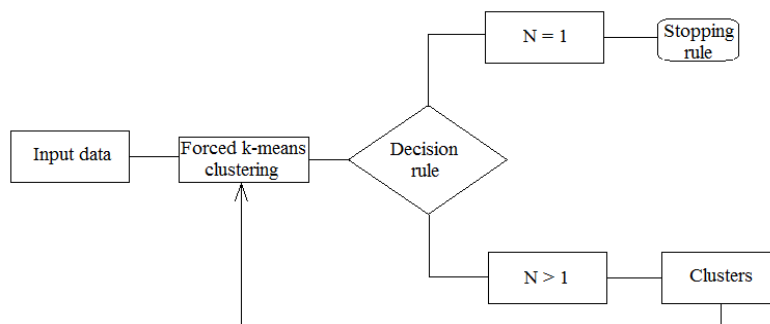


Fig. 1. Block scheme of a preclustering algorithm

The beginning of preclustering algorithm performance remains invariable, that is, the forced k-means clustering (or c-means clustering, or k-medians clustering) of input data set is performed. After the forced division, each cluster is checked by the decision rule, which later determines if this precluster is a single cluster or the part of some bigger cluster.

Modified preclustering algorithm and the transformed decision rule require performing following steps:

1. Forced k-means clustering (or other clustering requiring that the input parameter considering the number of clusters should be set) is performed. In given algorithm the forced clustering always divides input data into only two preclusters.
2. In each divided precluster mean distances from the centre of this precluster to all objects inside it are determined, where $R(K_1)$ and $R(K_2)$ are mean distances from the precluster centre accordingly.
3. Before the forced division the mean distance between all objects is calculated being the distance in the general cluster $d(K)$. The general cluster is considered to be the input data set.
4. By the decision rule the possibility of cluster existence (that is, the possibility that the divided preclusters can be clusters) is determined.

The modified decision rule in the preclustering algorithm can be written as follows: (Table 1).

Table 1

Modified decision rule and its explanation

Decision rule	The number of found clusters	Conclusion
if $R(K_1) + R(K_2) > d(K)$ and $C_1((x_1, y_1), R(K_1)) \cap C_2((x_2, y_2), R(K_2))$	1	Input data set is a single cluster
Otherwise	More than 1	Input data set contains two or more separate clusters

After the forced clustering for each found precluster K_1 and K_2 the mean distance from the precluster centre to all objects inside it $R(K_1)$ and $R(K_2)$ is calculated. Then the circle whose centre coincides with the precluster centre is built, (x_1, y_1) being the centre of the first precluster and (x_2, y_2) being the centre of the second precluster. If the first inequality of the decision rule is satisfied and built circles intersect $C_1((x_1, y_1), R(K_1)) \cap C_2((x_2, y_2), R(K_2))$ (circle C_1 intersects circle C_2), it shows that in this data set one cluster exists (preclusters divided by forced clustering are not clusters). In all other cases found preclusters are independent ones.

5. Choosing the centre of the cluster

In the modified decision rule the mean distances from all objects of the precluster to its centre are calculated. In many popular algorithms the centre of the group of objects is denoted in different ways. In the k-means algorithm the centre of the group is considered to be a centroid. The centroid is a mean value of all analyzed objects in one group. In the k-medians algorithm the median of the group of objects is calculated instead of calculating the mean value of all objects in the group for determining the centroid.

The proposed decision rule determines the centre of the group as a local density maximum of the group of objects (before clustering) or of the precluster (after clustering). The most significant disadvantage of choosing group centers (centroids, or mean values) is its strong dependence on anomalies.

At high density of objects in the investigated group, the great number of objects and in the case of a spherical shape of the group the difference between the centroid, median and maximum density is insignificant, that is shown in Fig. 2, *a*. When the shape of the group of objects is arbitrary and when the density is variable, the difference between the centroid, median and maximum density becomes bigger (Fig. 2, *b*). Choosing the maximum density of the group of objects is explained by the fact that the input data set (without a priori information about it) can contain any number of anomalies. The centroid and median are influenced by this factor and can react inadequately, which can cause erratic results, but the density of the group of objects is resistant to anomalies and their influence.

Visualization of the modified decision rule parameters are shown in Fig. 3. This rule can be introduced as the criterion of spherical resolution, when the sum of radii of two groups of objects is less than the distance between their centers (in such a data set only one cluster exists). In the criterion of spherical resolution (Fig. 4) the centre of the cluster is a centroid and its radius is determined as maximum distance from the centre of the cluster, or the radius of the least circle surrounding all objects in the cluster.

The disadvantage of the criterion of spherical resolution is the fact that maximum radius from the centre of the group of objects heavily depends on the anomalies. At the significant standard deviation and in the presence of anomalies the criterion of spherical resolution causes the distortion of results.

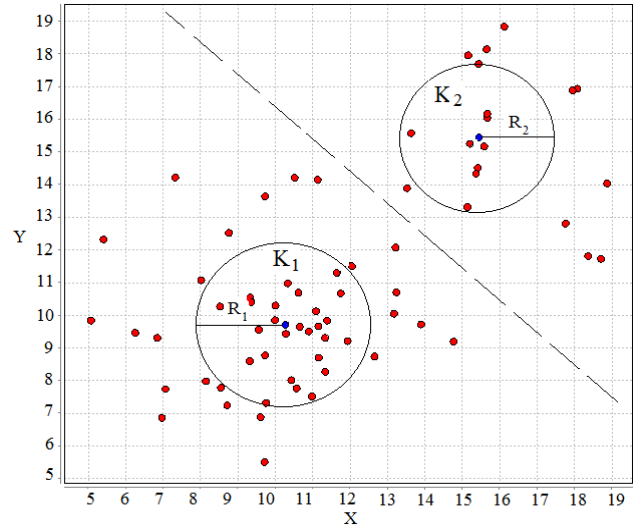


Fig. 3. Visualization of the parameters of the modified decision rule

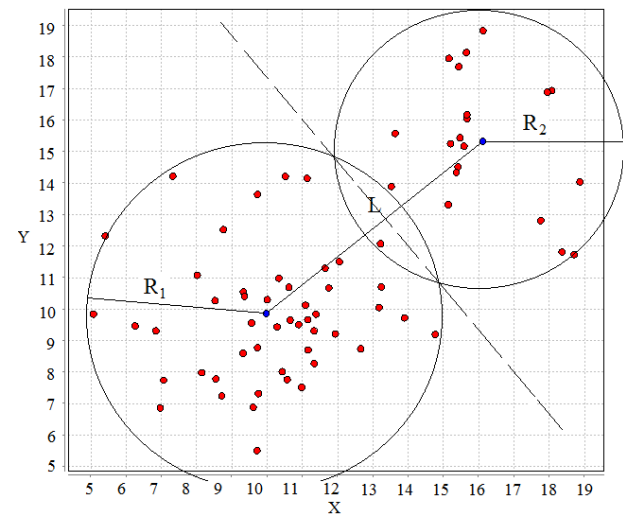


Fig. 4. Visualization of the parameters of the criterion of spherical resolution

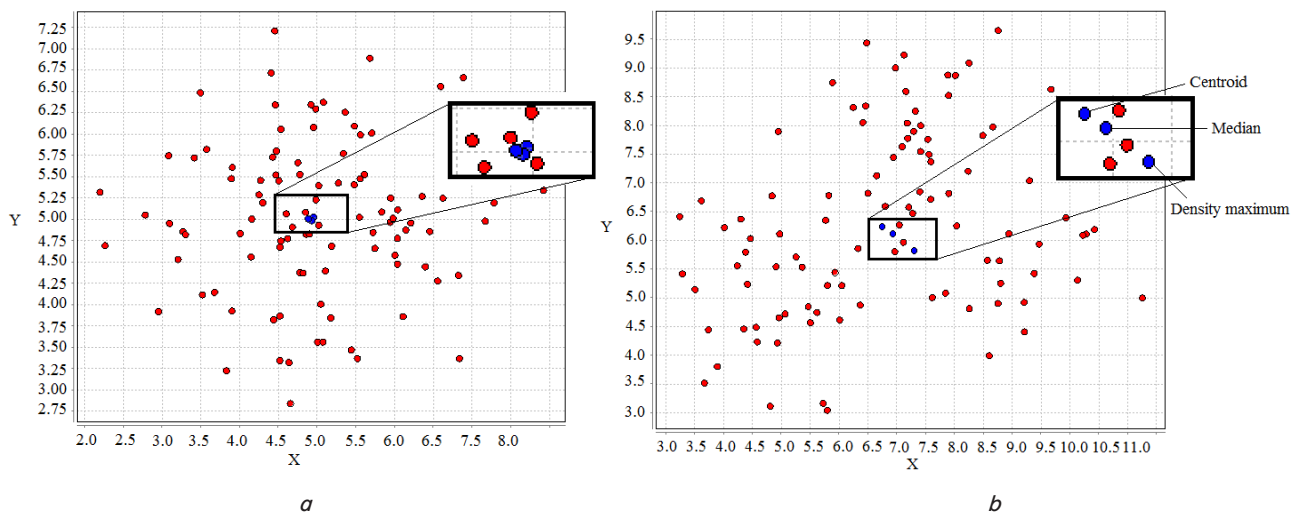


Fig. 2. Centers of the group of objects: *a* – case with the high density of objects, *b* – case with variable density and the arbitrary shape of the group


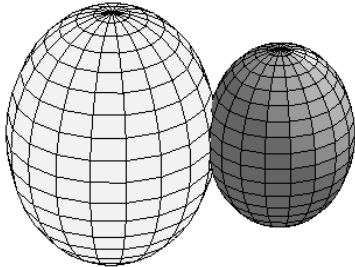
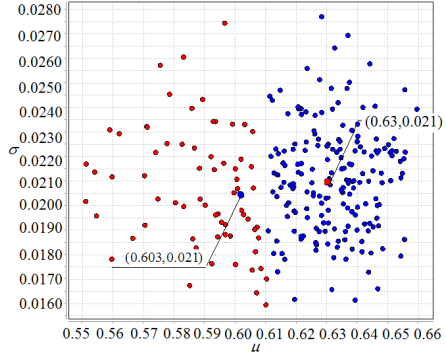

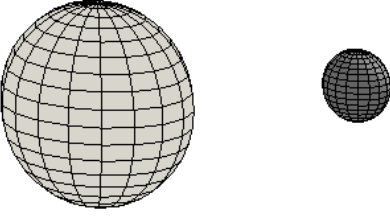
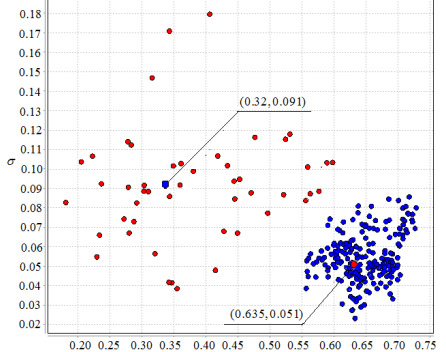

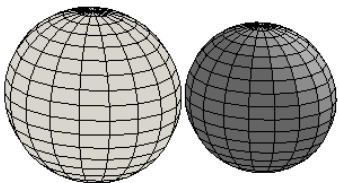
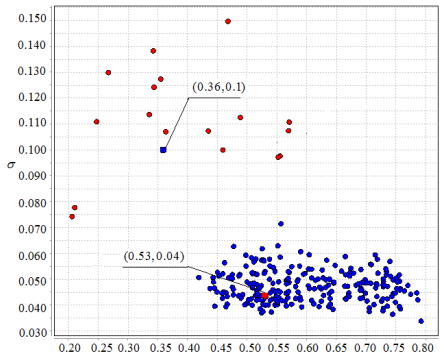

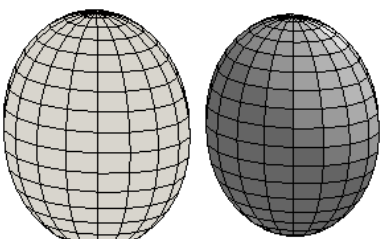
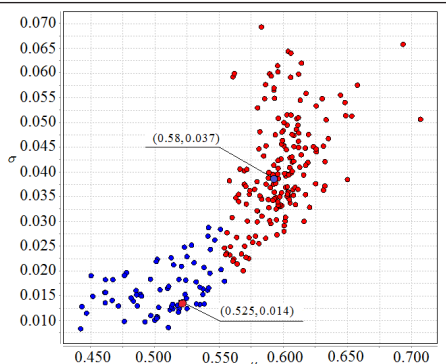
6. Experimental results obtained by the modified decision rule

The decision rule was tested using the series of experimental data being the images of the defects of human skin (disease, pigmentation). The input image of the size 256x256 was divided into parts of the size 16x16. For each part of the image the standard deviation and the mean value was determined. There-

by, the image was transformed into the data set. After that, the forced k-means clustering was performed, in which the input parameter of the number of clusters always equals 2. Using the decision rule we determine whether in the input data set more than one cluster exists. The image of the skin without defects is considered to be a single cluster. The results of the experimental investigation of the decision rule are shown in Table 2.

Table 2

Experimental results of the action of the decision rule

№	Image 256x256	Intersection of built circles (sphere is for better visualization)	Visualization of k-means clustering (with density centers)
1		 <p data-bbox="544 868 791 925"> $C_1((0.603,0.021),0.017) \cap$ $C_2((0.63,0.021),0.01)$ </p>	
2		 <p data-bbox="539 1231 794 1288"> $C_1((0.635,0.051),0.038) \cap$ $C_2((0.32,0.091),0.108)$ </p>	
3		 <p data-bbox="555 1605 778 1662"> $C_1((0.36,0.1),0.103) \cap$ $C_2((0.53,0.04),0.094)$ </p>	
4		 <p data-bbox="544 2002 791 2059"> $C_1((0.58,0.037),0.026) \cap$ $C_2((0.525,0.014),0.028)$ </p>	

The first case shows that the conditions of the decision rules are satisfied, and as a result in the input data set one cluster is detected.

The second case demonstrates that preclusters are located in the significant distant from each other. The result of the analysis of the third case is detecting two independent clusters in the input data set.

At third case both condition of the decision rule are satisfied. In given data set two separate clusters exist, but it is still possible that it makes up one general cluster. In this case it is necessary to use additional means of checking and control (tests, criteria).

The analysis of the fourth case demonstrated the existence of two separate clusters. On this image the color and structure of skin look like normal and using the decision rule can cause inadequate results. Such a set of input data can be analyzed as one stretched out cluster, although actually there are two of them. In such a case one more parameter should be added to the standard deviation and mean value and 3D/4D analysis should be performed.

Table 3 shows the comparison of the results obtained by the decision rule and the criterion of spherical resolution at the condition that normal skin image is considered to be one single cluster.

Table 3

Comparison of the rules and the number of found clusters

№	The number of clusters (visual analysis)	The number of clusters (modified decision rule)	The number of clusters (criterion of spherical resolution)
1	1	1	1
2	2	2	1
3	2	2 (additional means of checking)	1
4	2	2 (additional parameter)	1

The decision rule not always detects the precise number of clusters, but for primary analysis of clustering possibilities the use of this rule together with the preclustering algorithm provides the stronger probability of correct cluster detecting than in the case of criterion of spherical resolution.

7. Conclusions

The majority of popular algorithms of image analysis are able to easily detect the presence and the number of defects. However, given experimental images allow only visual comparison of the correctness of decision rule application. In practical tasks a priori information about the number of the clusters is absent, that is, there is no input image. Input data are only the set of objects without any additional information.

In this article the modified decision rule in the preclustering algorithm has been presented. Also this decision rule was tested on a series of experimental data, and the results were compared with the criteria of spherical resolution. The modified decision rule allows to obtain a better results than classical, one and much better than criterion of spherical resolution.

In this article experimental data were divided into one or two clusters, but if the input data contain more than two clusters the stopping rule for the preclustering algorithm should be applied.

This decision rule has its disadvantages. One of them is still the dependence of the parameters on calculated distances. When objects are significantly scattered and their number is small, there are possibilities for existing anomalies and, accordingly, the difficulties in obtaining adequate results.

In further investigations it is proposed not to calculate distances and pay attention only to the density of the objects.

References

1. North, M. A. Data mining for the masses [Text] / M. A. North – A global text project book, 2012. – 264 p.
2. Aggarwal, C. C. Data Clustering: Algorithms and Applications. 1st Edition [Text] / C. C. Aggarwal. – Chapman & Hall, 2013. – 652 p.
3. 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, 2000. – P. 169–178. doi: 10.1145/347090.347123
4. Kovács, L. Parameter Optimization for BIRCH Pre-Clustering Algorithm [Text] / L. Kovács, L. Bednarik // 12th IEEE International Symposium on Computational Intelligence and Informatics, 2011. – P. 475–480. doi: 10.1109/cinti.2011.6108553
5. Khan, M. A. H. Pre-processing for K-means Clustering Algorithm [Text] / M. A. H. Khan. – Senior Projects Spring, 2015. – 260 p.
6. Everitt, B. S. Data Clustering. 5th Edition [Text] / B. S. Everitt, S. Landau, M. Leese, D. Stahl. – Willey Series In Probability And Statistics, 2011. – 348 p.
7. Gan, G. Data clustering Theory, Algorithms, and Applications [Text] / G. Gan, C. Ma, J. Wu. – ASA-SIAM Series on Statistics And Applied Probability, 2007. – 488 p.
8. Hofmann, M. RapidMiner: Data Mining Use Cases and Business Analytics Applications [Text] / M. Hofman, R. Klinkenberg. – Chapman & Hall/CRC, 2013. – 431 p.
9. Kovács, F. Cluster Validity Measurement Techniques [Text] / F. Kovács, C. Legány, A. Babos // Proceeding AIKED'06 Proceedings of the 5th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases, 2006.
10. Rendón, E. Internal versus External cluster validation indexes [Text] / E. Rendón, I. Abundez, A. Arizmendi, E. M. Quiroz // International journal of computers and communications. – 2011. – Vol. 5, Issue 1. – P. 25–34.
11. Liu, Y. Understanding of Internal Clustering Validation Measures [Text] / Y. Liu, Z. Li, H. Xiong, X. Gao, J. Wu // IEEE International Conference on Data Mining, 2010. – P. 911–916. doi: 10.1109/icdm.2010.35
12. 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, Issue 4. – P. 3013–3022. doi: 10.1007/s13369-013-0920-7