на основе выборки данных и технологии интеграции несбалансированных данных.

Методика. Во-первых, традиционный алгоритм SMOTE был улучшен до K-SMOTE (метод увеличения числа примеров миноритарного класса, объединяющий стратегию семплинга SMOTE и метод K-средних). В K-SMOTE, набор данных подлежал кластеризации, а интерполяция проводилась между центром кластера и точкой исходных данных. Во-вторых, был предложен алгоритм ECA-IBD (улучшенная SMOTE-стратегия классификации несбалансированных данных на основе ансамблевого алгоритма). В ECA-IBD, увеличение числа примеров миноритарного класса проводилось с помощью K-SMOTE, а уменьшение числа примеров мажоритарного класса проводилось методом случайного отбора, с целью уменьшения масштаба проблемы и формирования нового набора данных. Целый ряд слабых классификаторов и методов интеграции был использован для формирования конечного сильного классификатора.

Результаты. Эксперимент проводился на UCI наборе несбалансированных данных. Результаты показали, что предложенный алгоритм эффективен при использовании F-значения и G-среднего значения в качестве оценочных индексов.

Научная новизна. Улучшен алгоритм SMOTE и скомбинированы стратегии увеличения числа примеров миноритарного класса и уменьшения числа примеров мажоритарного класса, а также технология бустинга для решения задач классификации несбалансированных данных.

Практическая значимость. Предложенный алгоритм имеет важное значение для классификации несбалансированных данных. Он может быть применен во многих областях, таких как обнаружение неисправностей, обнаружение вторжения и т. п.

Ключевые слова: *несбалансированные данные, композиционное обучение, увеличение числа примеров миноритарного класса, уменьшение числа примеров мажоритарного класса, классификация данных*

Рекомендовано до публікації докт. техн. наук В. В. Гнатушенком. Дата надходження рукопису 22.04.15.

Guangbin Sun[1],
Hongqi Li[1],
Haiying Huang[2]

1 – China University of Petroleum, Beijing, China
2 – Daqing Oilfield Engineering Co,Ltd, Daqing, Heilongjiang, China

IMPROVED K-MEANS ALGORITHM AUTOMATIC ACQUISITION OF INITIAL CLUSTERING CENTER

Гуанбінь Сунь[1],
ХунціЛі[1],
Хайїн Хуан[2]

1 – Китайський університет нафти, м. Пекін, КНР
2 – Дацин Ойлфілд Інжиніринг Ко, Лтд, м. Дацин, КНР

УДОСКОНАЛЕНИЙ АЛГОРИТМ К-СЕРЕДНІХ АВТОМАТИЧНОГО ВИЗНАЧЕННЯ ПОЧАТКОВИХ ЗНАЧЕНЬ ЦЕНТРІВ КЛАСТЕРІВ

Purpose. The traditional K-means algorithm requires the K value, and it is sensitive to the initial clustering center. Different initial clustering centers often correspond to the different clustering results, and the K value is always required. Aiming at these shortcomings, the article proposes a method for getting the clustering center based on the density and max-min distance means. The selection of the clustering center and classification can be carried out simultaneously.

Methodology. According to the densities of objects, the noise was eliminated and the densest object was selected as the first clustering center. The max-min distance method was used to search the other best cluster centers, at the same time, the cluster, which the object belongs to, was decided.

Findings. Clustering results are related to the selection of parameters θ. If the sample distribution is unknown, only test method can be used through multiple test optimization. With prior knowledge for the selection of θ, it can be converged quickly. Therefore, θ should be optimized.

Originality. This article proposes the new method based on the density to get the first initial clustering center, and then the new method based on the maximum and minimum value. The improved algorithm obtained through experimental analysis insures higher and stable accuracy.

Practical value. The experiments showed that the algorithm allows for automatic obtaining of the k clustering centers and have a higher clustering accuracy in unknown datasets processing.

Keywords: *clustering, K-means clustering, max-min distance method,density*

Introduction. Clustering means that a given object is divided into several clusters based on the given definition of similarity so that the objects within a cluster can be as similar as possible and the objects of different clusters can be as different as possible. According to the clustering rules, the clustering algorithm can be divided into: based

© Guangbin Sun, Hongqi Li, Haiying Huang, 2016

on partition, hierarchical,density or grid, and other clustering algorithms.

K-means algorithm is based on partition. James MacQueen proposed this algorithm in 1976. It can effectively classify large data sets, but the algorithm also has some problems [1]. Firstly, it needs to specify the K value. However, prior knowledge of data is needed to set K value. Secondly, the problem also concerns the initial clustering center selection. The improper selection of the initial clustering center affects the clustering results greatly. In addition, the Euclidean distance is defined as the distance measure in the algorithm, but the Euclidean distance is more sensitive to noise and outliers. In view of these problems, some methods are proposed to improve the k-means. DBS can is a typical clustering algorithm based on density, which aims finding high-density area cut by low-density area, can deal with the cluster of arbitrary shape and size, and find the cluster, for which an average K value was not found. There are many studies about the DBS can [2−4]. However, DBS can will meet more difficult problem than defining density when it deals with the data with great change and high dimension. Onoda T., Sakai M. and Yamada S. [5] proposed a method to select the initial center based on the independent component. Reddy D., Jana P. K. and Member I. S. [6] used Voronoi diagram to select the initial cluster centers. Zhang Y. J. and Cheng E.[7] made a conclusion on the improved method of choosing the initial clustering centers in the K-means algorithm.

Based on the density and the max-min distance means, a k-means algorithm is to find k initial clustering centers automatically, which makes the initial clustering center as far as possible to reflect the actual distribution of the data. The feasibility and effectiveness of this method was verified experimentally.

Introduction to the algorithm. ***The traditional K-means clustering algorithm*****.** The basic idea of the traditional K-means clustering is based on such a condition that K is regarded as the parameter, and N objects are divided into K clusters so that the similarity within the classes is high, and the similarity between the classes is low.

The process flow of K-means algorithm is as follows:

Input: the number of clusters K and data sets containing N objects.

Output: a collection of K clusters, which minimize the square error.

Method:

1. Select K objects as the cluster centers of the initial class at random. It can be operated as follows.

2. Assign each object (again) to the most similar cluster according to the mean of objects among classes.

3. Update the cluster mean. That is, to calculate the average value of the objects in each cluster.

4. Return to second step for loop executing until there is no more change and the algorithm ends.

Max-min distance means. The max-min distance means are based on the Euclidean distance. The most distant clustering center was identified, and then the others were determined, until no new clustering centers were generated. Finally, the samples were classified to the nearest class according to the principle of minimum distance. The main idea of such operation is to select the object as far as possible as the clustering center. In this way,the k-means algorithm avoids the situation when the initial clustering centers appear too dense. This allows for automatic determination of the number of initial clustering centers, and for improvement of the efficiency of the data set partitioning.

Steps of the algorithm:

1. Determine the sample point, provide a parameter of θ ($0.5 < \theta < 1$), then take a sample point as the first clustering center z_1.

2. Find a new clustering center,calculate the distances from all other samples to z_1, take the sample points with the maximum distance to z_1 as the second clustering center z_2. Calculate the distances between all samples to z_1 and z_2. Take the shortest distance between the clustering center z_1 and z_2 as the distance between each sample point to the clustering center. Then, determine whether the distance of sample point to the clustering center is greater than θ multiples of the distance between z_1 to z_2. If it is greater, take the sample point as the third clustering center z_3. Continue doing the same, until both the maximum and minimum distances stop appearing greater. The calculation for the clustering center is completed.

3. Classification: take the remaining samples into the nearest class according to the principle of minimum distance. Through clustering by the max-min distance means, we can see that the relationship between clustering result and the selection of parameter $\theta$and the initial object z_1 is significant. Without a prior knowledge of the sample distribution, we can only adopt the tentative and optimizing methods to get θ and z_1. In addition, if the size of the sample data is very large, the execution efficiency of the algorithm will be very low.

The improved k-means algorithm. The traditional K-means algorithm requires the K value, and it is sensitive to the initial clustering center. Different initial clustering centers often correspond to different clustering results, and the K value is always required. Aiming at these shortcomings, the article proposes a method for getting the clustering center based on the density and max-min distance means. The selection of the clustering center and classification can be carried out simultaneously.

Basic definitions. Data sets, which will be clustered are $X = \{x_i \mid x_i \in R^P, i = 1, 2, \ldots, n\}$; the clustering centers are $z_1, z_2, \ldots, z_k$; clustering results are represented with W_j ($j = 1, 2, \ldots, k$).

Definition 1

$N\mathrm{Eps}(x_i)$:Represent the Eps neighborhood of the object x_i. That is the data set in the super sphere region with the Eps as the radius and the object x_i as the center.

Definition 2

$\mathrm{Density}(x_i)$: Represent the density of object x_i. That is the number of samples in a certain area.

Definition 3

$\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $\mathbf{x}_j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ are two p-dimensional data objects. The Euclidean distance between them is

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{ip} - x_{jp})^2}.$$

Definition 4

The distance between the data object x_i and the set U is

$$\mathrm{d}(\mathbf{x}_i, U) = \min(d(\mathbf{u}_m, \mathbf{x}_i)), \mathbf{u}_m \in U.$$

Definition 5

The longest distance between the set X and set U is

$$d(X, U) = \max(d(x_i, U)), x_i \in X.$$

The basic ideas. In the improved k-means algorithm, the Euclidean distance in definition 3 is defined as the similarity measure. The k data objects, which have the furthest distance from each other, are selected as the initial centers. However, noise data often exists in the actual data. If only the furthest distance object is selected, it is very likely that the noise object is selected. Thus, the clustering result will be affected. Therefore, the noise reduction should be undertaken to remove the noise points and improve the accuracy of the algorithm.

The basic idea of this algorithm: first, calculate the density (x_i) of the area where the data object is located according to definition 1 and definition 2. Figure out the results according to the density. The data objects in the low-density area are regarded as the noise while the data object with the highest density is regarded as the first clustering center z_1. Calculate the distances between the remaining objects and z_1 according to definition 3. Take the object with a high density, which is the most distant from z_1, as the second clustering center z_2. Then calculate the distances between

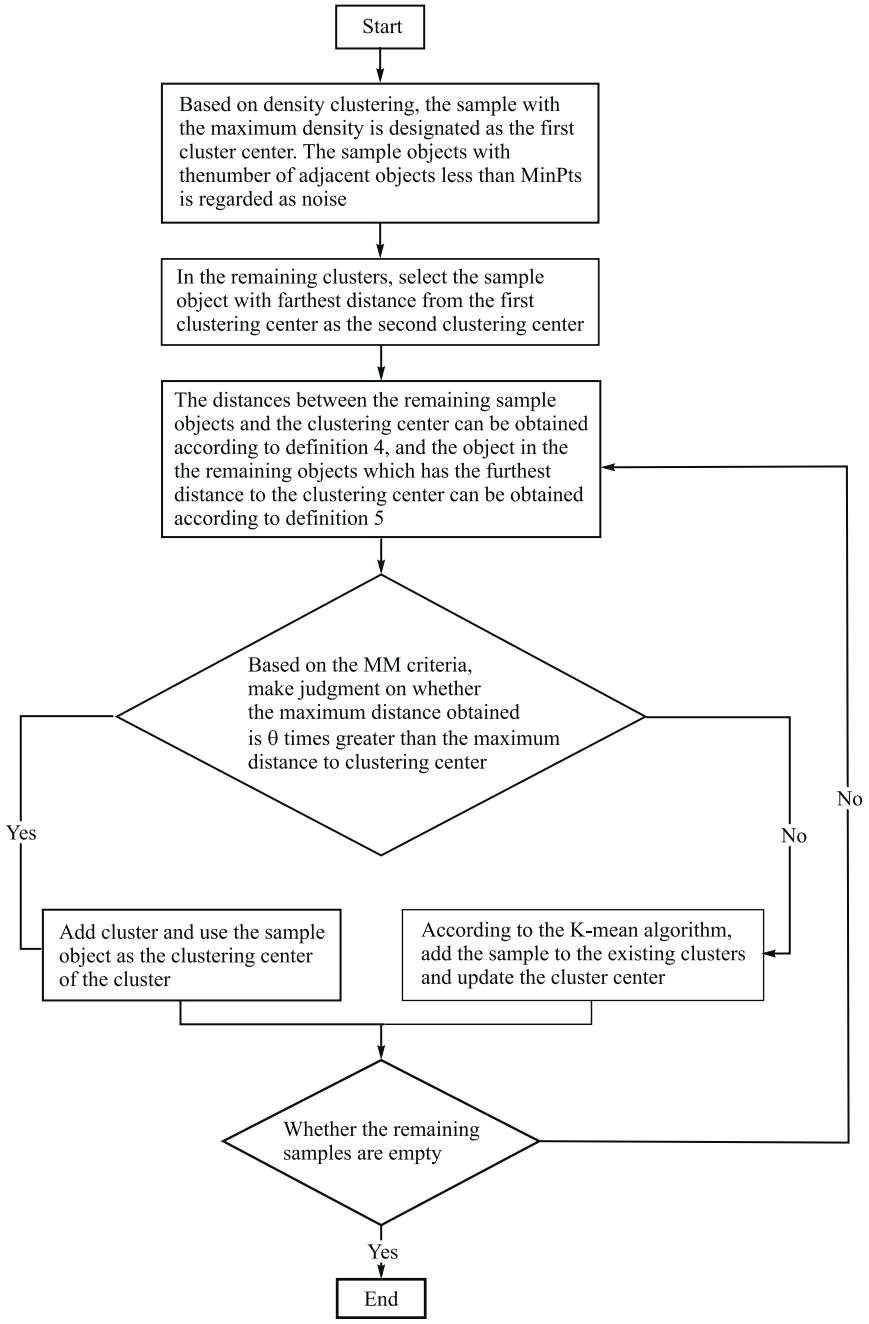

Fig. 1. Flow chart of the improved K-means algorithm

the remaining data objects and the clustering center according to the definition 4. Get the furthest distance between the remaining objects to the clustering center according to definition 5. Determine whether the point is the clustering center based on the conditions of the max-min distance means (here take $\theta = 0.6$). If it is a clustering center, we can take it as a new clustering center z_3. Select the furthest distance between the two objects in the clustering centers, as the distance in MM condition criterion, or add it to the nearest cluster. In this way, similar judgment can be made on the rest of the data objects, until the classification of all objects is completed. The specific process is shown in Fig. 1.

Algorithm analysis and experimental results. In order to test the validity and accuracy of the algorithm, a contrast experiment of the algorithm was carried out with both artificial and standard data sets before and after the improvement.

Artificial data set. Using the Random RBF of data mining the software Weka automatically generates a data set. It consists of 100 data sets of 3D data, which are classified into 2 categories. Add 5 noise points, and then test the data set. The experimental results are shown in Table 1.

Standard data set. For Standard data sets acquisition, Iris in UCI [8] database is often employed as test data. UCI database is a kind of database used to test machine learning and data mining algorithms. All the data in the database are clearly classified, so the accuracy of the algorithm can be directly used to indicate the quality of clustering.

In the experiment, the traditional K-means algorithm was tested by Weka, while the improved algorithm was implemented on Eclipse platform with Java. The experiments were conducted under both noise-free and noisy conditions to compare the accuracy of both the improved and original algorithms. The Iris data set in experiments had four properties: sepal length, sepal width, petal length and petal width, with 150 data objects. Predefination was divided into 3 categories: class Iris-setosa, Iris-versicolor and Iris-virginica. There were 50 data objects in each category. In order to test the impact of noise on the clustering results, we added five noises in the end of the Iris followed by {8.4, 4.6, 3.9, 0.8}, {3.2, 1.2, 0.4, 0.2}, {7.4, 4.4, 6.9, 0.7}, {3.8, 2.4, 0.8, 2.8}, {9.0, 5.0, 8.0, 3.0}. The test results for the Iris data set are shown in Table 2.

Table 1

Test results of artificial data

Algorithm	Sequence	Number of Cluster	Accuracy without Noise, %	Accuracy with Noise, %
Traditional	1	2	75.00	81.90
	2	2	98.00	71.43
	3	2	75.00	71.43
	4	2	88.00	70.48
	5	2	97.00	76.19
	6	2	97.00	71.48
	7	2	74.00	71.43
	8	2	77.00	71.48
	9	2	89.00	71.43
	10	2	75.00	88.57
	average	2	84.10	74.58
Improved	1	2	98.00	98.95

Table 2

Test results of Iris data

Algorithm	Sequence	Number of cluster	Accuracy without noise, %	Accuracy with noise, %
Traditional	1	3	88.00	49.03
	2	3	90.67	92.25
	3	3	56.67	87.10
	4	3	88.67	64.52
	5	3	68.67	89.03
	6	3	52.67	55.48
	7	3	86.67	70.97
	8	3	84.67	61.29
	9	3	52.00	61.94
	10	3	85.33	49.68
	average	3	75.40	68.13
Improved	1	2	98.00	98.95

Table 1 and Table 2 show that the traditional K-means algorithm is easily affected by the initial center. For example, in the noise-free test in Table 2, the accuracy of the results in the second and sixth experiments varies greatly. Besides, the traditional K-means is susceptible to noise. The accuracy of the algorithm always decreases in the case of noise. The traditional K-means algorithm cannot eliminate the noise points. The noise points often interfere the choice of the initial center and update of clustering center, which affects the clustering results. While by the improved K-means algorithm, the choice of the initial center is fixed when all the parameters are set. So only one clustering result can be produced. Therefore, the clustering algorithm is relatively stable. Fig. 2 and Fig. 3 show the differences made by using the improved algorithm for the clustering results of Iris data set (with noise). Compared with the clustering results in the paper, improved K-means algorithm proposed in this paper can often produce better clustering results. Besides, noise data can be effectively separated by using this algorithm and the impact on the clustering results can be reduced.

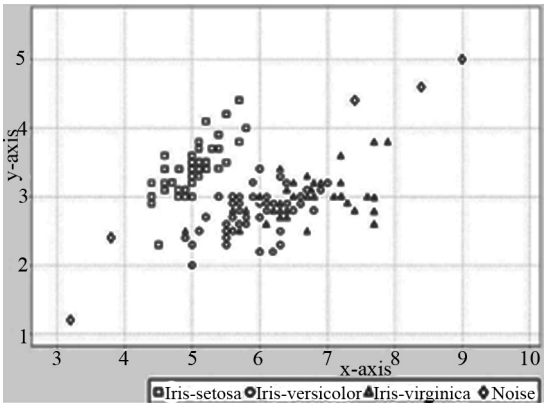

Fig. 2. Iris data set (with noise) sepal attribute clustering results: X-axis– length; Y-axis– width; (□ – *iris-setosa*; ○ – *iris-versicolor*; Δ – *iris-virginica;* ◊ – *noise*)

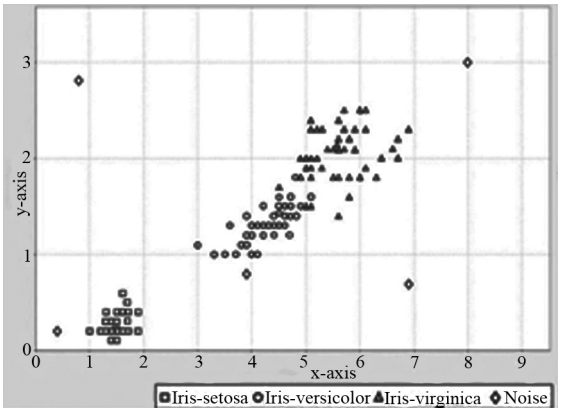

Fig. 3. Iris data set (with noise) petal attribute clustering results: X-axis – length; Y-axis – width; (□ – *iris-setosa;* ○ – *iris-versicolor;* Δ – *iris-virginica,* ◊ – *noise*)

Conclusion. The general process of the traditional K-means algorithm was considered and the influence of the selection of initial clustering center on clustering results was analyzed. Then, the authors proposed a new method based on the density to get the first initial clustering center, and a new method based on the maximum and minimum value. The improved algorithm obtained through experimental analysis can produce a higher and stable accuracy, which is more suitable for the clustering of the actual data. In the future, it can be improved in the following aspects: clustering results are related to the selection of parameters θ. Without knowledge of the sample distribution, only test method can be used through multiple test optimization. With prior knowledge for the selection of θ, it can be converged quickly. Therefore, θ should be optimized.

References / Список літератури

1. Celebi, M.E., Kingravi, H.A. and Vela, P.A., 2013. A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert* S*ystems with Applications*, vol. 40, no. 1, pp. 200–210.

2. Tran T.N. and Drab K., Daszykowski M., 2013. Revised DBSCAN algorithm to cluster data with dense adjacent clusters. *Chemometrics and Intelligent Laboratory Systems*, vol. 120, pp. 92–96.

3. Chakraborty, S. and Nagwani, N.K. 2014. Analysis and study of Incremental DBSCAN clustering algorithm. *Eprint Ar Xiv*, vol. 1406, no. 4754, pp. 401–410.

4. Smiti, A. and Eloudi, Z. 2013., Soft DBSCAN: Improving DBSCAN Clustering method using fuzzy set theory. In: *Proc. of the 6th International Conf. onHuman System Interaction* (*HSI*), pp. 380–385.

5. Onoda, T., Sakai, M. and Yamada, S. 2012. Careful seeding method based on independent components analysis for k-means clustering. *Journal of Emerging Technologies in Web Intelligence*, vol. 4 no. 1, pp. 51–59.

6. Reddy, D., Jana, P.K. and Member, I.S., 2012. Initialization for K-means clustering using Voronoi diagram, *Procedia Technology*, vol. 4, pp. 395–400.

7. Zhang, Y.J. and Cheng, E. 2013. An optimized method for selection of the initial centers of k-means clustering. *Integrated Uncertainty in Knowledge Modelling and Decision Making*. Springer Berlin Heidelberg, pp. 149–156.

8. Frank, A. and Asuncion A. 2012, UCI machine learning repository. Available at: <http:// archive.ics.uci.edu/ml> (2012-05-20)

Мета. Традиційний метод K-середніх вимагає наявності значення K і чутливий до початкового значення центрів кластерів. Різні початкові значення центрів кластерів часто призводять до різних результатів кластеризації, а наявність значення K завжди обов'язкове. З метою усунення цих недоліків, у роботі запропоновано спосіб отримання значення центру кластера на підставі щільності й мінімаксної відстані. Вибір центру кластера та класифікація можуть проводитись одночасно.

Методика. Відповідно до щільності об'єктів був зменшений шум, а в якості початкового значення центру кластера обраний об'єкт з найбільшою щільністю. Метод мінімаксної відстані використаний для пошуку інших кращих центрів. Обирається кластер, до якого належить об'єкт.

Результати. Результати кластеризації пов›язані з вибором параметрів θ. В умовах відсутності знань про розподіл вибірки може використовуватися тільки тестовий метод за допомогою багаторазової оптимізації тестування. У разі, коли θ заздалегідь відоме, можливо швидке сходження. Отже, θ має бути оптимізовано.

Наукова новизна. У роботі запропоновані нові методи отримання початкового центру кластера на основі щільності й мінімаксної відстані. Вдосконалений алгоритм, отриманий за допомогою експериментального аналізу, стабільно показує більш високу точність.

Практична значимість. Експерименти показали, що алгоритм може автоматично отримувати K значень центрів кластерів і показує більш високу точність кластеризації за обробки невідомих наборів даних.

Ключові слова: *кластеризація, кластеризація за методом K-середніх, метод мінімаксної відстані, щільність*

Цель. Традиционный метод K-средних требует наличия значения K и чувствителен к начальному значению центров кластеров. Различные начальные значения центров кластеров часто приводят к разным результатам кластеризации, а наличие значения K всегда обязательно. С целью устранения этих недостатков, в работе предложен способ получения значения центра кластера на основании плотности и минимаксного расстояния. Выбор центра кластера и классификация могут проводиться одновременно.

Методика. В соответствии с плотностями объектов был уменьшен шум, а в качестве начального значения центра кластера выбран объект с наибольшей плотностью. Метод минимаксного расстояния использован для поиска других лучших центров. Выбирается кластер, к которому принадлежит объект.

Результаты. Результаты кластеризации связаны с выбором параметров θ. В условиях отсутствия знаний о распределении выборки может использоваться только тестовый метод посредством многократной оптимизации тестирования. В случае, когда θ заранее из-

вестно, возможно быстрое схождение. Следовательно, θ должно быть оптимизировано.

Научная новизна. В работе предложены новые методы получения начального центра кластера на основе плотности и минимаксного расстояния. Усовершенствованный алгоритм, полученный с помощью экспериментального анализа, стабильно показывает более высокую точность.

Практическая значимость. Эксперименты показали, что алгоритм может автоматически получать K значений центров кластеров и показывает более высокую точность кластеризации при обработке неизвестных наборов данных.

Ключевые слова: *кластеризация, кластеризация по методу K-средних, метод минимаксного расстояния, плотность*

Рекомендовано до публікації докт. техн. наук В.В.Гнатушенком Дата надходження рукопису 24.04.15.