Application of K-means Algorithm in Geological Disaster Monitoring System

Wang Jianguo
College of Computer Science and Engineering
Xi'an Technological University
No.2 Middle Xuefu Road, Weiyang District,
Xi'an, 710021, China
e-mail: wjg_xit@126.com,

Xue Linyao[*]
College of Computer Science and Engineering
Xi'an Technological University
No.2 Middle Xuefu Road, Weiyang District,
Xi'an, 710021, China
e-mail: 1525610807@qq.com

***Abstract*—The K-means algorithm is considered to be the most important unsupervised machine learning method in clustering, which can divide all the data into k subclasses that are very different from each other. As K-means algorithm is simple and efficient, it is applied to data mining, knowledge discovery and other fields. This paper proposes CMU-kmeans algorithm with improved UPGMA algorithm and Canopy algorithm. The experimental results is that the algorithm can not only get the number k of the initial clustering center adaptable, but also avoid the influence of the noise data and the edge data. Also, the improved algorithm can void the initial effect of the random selection on the clustering, which reflects the actual distribution in the dataset.**

Keywords-Clustering Analysis; CMU-kmeans Algorithm; Geological Disaster Monitoring Data

I. INTRODUCTION

The occurrence of geological disasters caused great casualties to humans, the main reasons include landslides and debris flow and rainfall and so on. And these geological disasters always cause many local public facilities to be damaged by large and small, and brought great damage to the people and their property. Also, there are still many such cases in China. Faced with such a severe threat of geological disasters, the state and the government on the prevention and control of geological disasters into a lot of human and material resources, and achieved remarkable results. With the progress of technology and high development of information technology, many new detection equipments have been put into the geological disaster real-time detection, such as GPS, secondary sound wave monitoring, radar and so on.

With the development of geological hazard detection technology, the amount of the monitoring data grew by leaps and bounds, data types are becoming more and more complex as well. K-means algorithm is a clustering algorithm based on the classification of the classic algorithm, the algorithm in the industrial and commercial applications more widely. As we all know, it both has many advantages and many disadvantages. In this paper, we mainly study the optimization of the initial clustering center and the avoidance of the blindness of the k-value selection, and propose the CMU-kmeans algorithm.

The data source of the study is the historical data detected by the geological disaster monitoring system, and 2000 records are randomly selected from the rainfall data of different areas in Shaanxi Province as the research object, which are served as a representative sample of the improved K-means clustering algorithm. The experimental results show that the improved algorithm not only eliminates the sensitivity to the initial input and improve the stability and effectiveness of the algorithm, but also can intelligently determine the initial clustering center number k, which improves the simplicity and operability of the algorithm.

A. Overview of K-means algorithm

The K-means algorithm is a classical unsupervised clustering algorithm. The purpose is to divide a given dataset containing N objects into K clusters so that the objects in the cluster are as similar as possible, and the objects between clusters are as similar as possible. Set the sample set X = {x1, x2, x3, ..., xn}, n is the number of samples. The idea of the K-means algorithm is that the k data objects are randomly selected from the sample set X as the initial clustering center, and then the data is allocated to the most similar cluster according to the similarity degree of each data object and k clustering centers; Recalculate the average of each new cluster and regard it as the next clustering center and repeat the process until the updated cluster center is consistent with the update, that is, the criterion function E converges. The goal is to make the object similarity in the cluster the largest, and the similarity between the objects is the smallest. The degree of similarity between the data can be determined by calculating the Euclidean distance between the data. For the n-dimensional real vector space, the Euclidean distance of two points is defined as form.1:

$$\delta(\xi, \psi)=\sqrt{(x_i - y_i)^2} \tag{1}$$

Here, x_i and y_i are the attribute values of x and y respectively, and the criterion function is defined as form.2:

$$E = \sum_{i=0}^{n} \sum_{x \in c_i} |x - \bar{x}_i|^2 \tag{2}$$

Here, k is the total number of clusters, and $\bar{x}_i$ is the center of cluster c. The flow of K-means algorithm is shown in Fig. 1.

DOI: 10.21307/ijanmc-2019-002

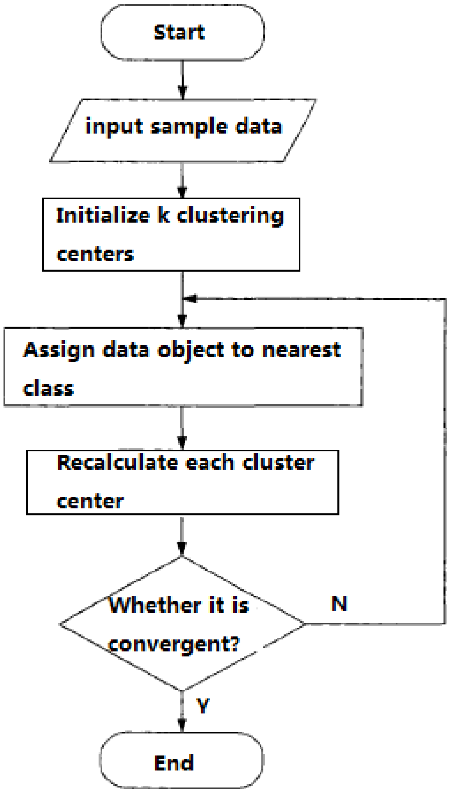
Start
input sample data
Initialize k clustering centers
Assign data object to nearest class
Recalculate each cluster center
Whether it is convergent?
N
Y
End

Figure 1. K-means clustering algorithm flow chart

B. Research status quo of K-means algorithm

For the advantages of K-means algorithm, it has been widely used in practice, but there are many shortcomings as well. In order to get better clustering effect, many researchers have explored the shortcomings of improving K-means. Aiming at the shortcomings of K-means algorithm in selecting the initial point, many scholars have proposed an improved method. Duan Guiqin [1] uses the method of product based on mean and maximum distance to optimize the initial clustering center. The algorithm first selects the set of data objects which are the farthest from the sample set to join the clustering center, and then the set of mean and current poly The largest data object of the class center is added to the clustering center set, which improves the accuracy. Yi Baolin [2] et al. proposed another improved K-means algorithm, which first calculates the density of the region to which the data object belongs, and then selects k points as the initial center in the high density region. The experimental results show that the algorithm reduces the initial center point Impact. Yiu-Ming Cheng[3] and others proposed a new clustering technique called K * -means algorithm. The algorithm consists of two separate steps. A center point is provided for each cluster in the first step; and then adjust the unit through adaptive learning rules in the second step. The algorithm overcomes the shortcomings of K-means algorithm initial center sensitivity and K value blindness, but the calculation is complicated. Xie and others [4] proposed a k-means algorithm to optimize the initial clustering center by using the minimum variance based on the sample space distribution compactness information. The algorithm chooses the samples with the smallest variance and a distance away from each other as the initial clustering center. Liu Jiaxing et al.[5] proposed a radius-based k-means + λ algorithm. When selecting the initial center point of the cluster, the distance ratio between points is calculated from the λ parameter and rounded at a specific distance. In the circle, an initialized center point is selected according to the distance ratio, and the algorithm has higher performance in error rate and operation time. Ren Jiangtao[6] proposed an improved K-means algorithm for text clustering, which is improved by using feature selection and dimension reduction, sparse vector selection, initial center point search based on density and spreading, Class accuracy, stability and other aspects have improved.

C. The performance analysis of K-means algorithm

K-means clustering algorithm uses the Euclidean distance to calculate the distance between each sample point. For the convex and spherical data distribution, the clustering effect is better and has been widely used in many fields. However, the Euclidean distance criterion adopted by the algorithm also has some limitations. For the more complicated or non-convex data, the clustering effect is often not very satisfactory. Clustering algorithm in the iterative process, if you do not meet the termination criteria will recalculate the average clustering center, this operation also improves the convergence rate of the clustering algorithm. In summary, K-means clustering algorithm has the following advantages and disadvantages of the following aspects.

1) The main advantages of K-means algorithm:

a) K-means clustering algorithm has high stability and scalability, clustering effect is very well.

b) The results of the treatment is intuitive and easy to understand. When dealing with the target data in numerical form, its geometric meaning is very clear. When clustering images and texts, the extracted eigenvalues can be regarded as clustering result values for the convenience of people's understanding.

c) K-means clustering algorithm When dealing with numerical data sets, the input data sequence will not affect the clustering result.

d) It can be a good judge of the data set shape is convex cluster.

2) The main shortcomings of K-means algorithm:

a) The K value in the K-means algorithm needs to be given in advance. According to the K value determined in advance, the clustering samples are classified into K class, so that the sum of squares of all the samples in the clustering domain to the clustering center is minimized.

b) Clustering results are highly dependent on the selection of initial clustering centers. The K-means algorithm uses the stochastic method to select the initial clustering center. If the initial clustering center is chosen improperly, it is difficult to obtain the ideal clustering effect. This dependence on the initial value may lead to the instability of the clustering results, and it is easy to fall into the local optimal rather than the global optimal results.

c) Sensitive to noise and isolated points.

d) The time complexity of the algorithm is large.

II. IMPROVEMENT OF K-MEANS ALGORITHM AND ITS APPLICATION

Aiming at the shortcomings of traditional K-means algorithm, this paper mainly improves on the optimization of initial clustering center to enhance the clustering effect.

A. The selection of data object in Cluster analysis

The preliminary data are collected firstly when data selecting, then know about the characteristics of data to identify the quality of the data and to find a basic observation of the data or assume the implied information to monitor the subset of data of interest. The data object segmentation variable determines the formation of clustering, which in turn affects the correct interpretation of the clustering results, and ultimately affects the stability of the clustering clusters after the new data objects are added. Before the K-means clustering related data mining, the sample data set related to the data mining clustering analysis should be extracted from the original data object set, and it is not necessary to use all the historical data. In addition, we should pay attention to the quality of data, only high-quality data to the correct analysis of conclusions everywhere, to provide a scientific basis for clustering.

The source of this research object is the historical monitoring data of the geological disaster monitoring system. From the records of geological monitoring data from 2016 to 2017, a representative sample of K-means clustering algorithm for this improved algorithm is selected as the object of study in 2000, and the two samples of rainfall are randomly selected in different regions.

The sample data attributes show as table1:

TABLE I. THE SAMPLE DATA ATTRIBUTES

Field number	**Field name**	**Field code**	**Type of data**
1	Id	Xx	Number
2	Sno	Yy	Varchar
3	Type	type	Varchar
4	Gettime	time	Datatime
5	Alarm Level	alarm	Integer
6	Value	value	Double
7	Day Value	d_value	Double

For the cluster analysis, there are obviously redundant ones in the data attributes of the above geological hazard monitoring system, and it does not have the objectivity of the cluster analysis data. Therefore, the redundant ones should be eliminated. Finally, only four data object attributes reflecting the characteristics of rainfall data are selected as the research object. The optimized data attributes show as table2:

TABLE II. THE OPTIMIZED DATA ATTRIBUTES

Field number	**Field name**	**Field code**	**Type of data**
1	Id	xx	Number
2	Sno	yy	Varchar
3	Gettime	time	Datatime
4	Day Value	d_value	Double

B. Improvement of K-means algorithm

It is not difficult to see that, through the above study of the status quo, we can see that most of the above algorithm improvements are only a single defect in the traditional k-means algorithm is optimized. Although these improvements have optimized the k-means algorithm to some extent, there are still many shortcomings. For the above geological disaster monitoring system rainfall data characteristics, the K-means algorithm is very sensitive to the initialization center, and the initial clustering center is very easy to make the clustering result into the local optimum and the influence of the isolated point is large. In this paper, the simple random sampling technique is used to reduce the scale of the data set on the original dataset, and then the improved UPGMA algorithm and Canopy algorithm are combined to propose the CMU-kmeans algorithm. The improved algorithm can select the points with the furthest distance k in the high density region as the initial clustering center according to the regional density of each data, so that the improved k-means algorithm can produce high quality poly The results show that the sensitivity of the algorithm is not only eliminated, but also the stability and validity of the algorithm are improved.

1) Improved UPGMA algorithm

a) The basic idea of improved UPGMA algorithm

At the beginning of the UPGMA algorithm, each data object in the sample data set is considered as a separate class； and calculates the distance between each two data objects to obtains the distance matrix, then merges the two data objects that are closest to each other to obtain a new subclass, repeat the process .The UPGMA algorithm stops until no new class is generated or the stop condition is satisfied. It can be found that the first subclasses are usually located in the dense area of the data set, so the subclass center selected by this algorithm can be used as the initial clustering center candidate point for the next step. In this way, the selection of the initial clustering center is optimized and its accuracy is improved. The distance between two data objects is measured using the Euclidean distance formula, as form3:

$$\mathrm{d} = sqrt\left(\sum_{k=1}^{m}\left(X_{ik} - X_{jk}\right)^{2}\right) \tag{3}$$

Here, Xi and Xj represent the data objects in the sample data set.

Xi={Xi1,Xi2,…,Xik,…,Xim}，k=1,2,…,m

Xj={Xi1,Xi2,…,Xik,…,Xjm}，k=1,2,…,m

The formula for calculating subclasses is as form4:

$$Z = \frac{1}{n}\sum_{j=1}^{n} X_j \qquad (4)$$

Here, n refers to the number of data objects contained in a subclass, and Xj refers to a data object in the subclass.

b) The description of improved UPGMA algorithm

Input: All data in the sample data set, parameters m, p, Q;

Output: initial clustering center candidate point.

(1) set each data object as a separate class;

(2) Calculate the distance between two data objects, and then merge the nearest two classes into a new subclass to determine whether the subclass of the data object containing no less than m% of the total amount of data continues to produce , If not, then go to (4);

```
(3) For (i = 1 to maxcluster) {
    { for (j = i + 1 to maxcluster) {
        If the number of data objects in subclasses i and j is less than or equal to m% of the total amount of data, calculate the distance between them to obtain the distance matrix.
      }
    }
```

Find the nearest two subclasses i and j and merge them into a new subclass, then add the new subclass to the end of the sequence Q to go to (2);

(4) Select the former p% subclasses in the sequence Q as the candidate subclasses and calculate the centers of all candidate subclasses as the initial clustering center candidate points.

Using the advantage of the improved UPGMA hierarchical clustering algorithm, we can find the dense region of the data set，which avoid the edge data and the noise data become the initial center candidate point. At the same time, considering the relative intensity of the region, we propose new clustering conditions and filter conditions to change the traditional UPGMA algorithm, so that the generation of subtrees can be stopped at different clustering levels to adapt to the actual density distribution data set. But the improved UPGMA algorithm also has some shortcomings. For example, if the m% and p% values are not set properly, the selection of the initial clustering center candidate points may be too dense and centralized. However, the Canopy algorithm, which introduces the idea of maximum and minimum distance, can select the data points that are far apart from each other. It is necessary to make up the deficiencies of the improved UPGMA algorithm. Therefore, it is necessary to introduce the Canopy algorithm to ensure that the distribution of the initial clustering center is decentralized, which can correctly reflect the data distribution of the original data set.

2) Improved Canopy algorithm

In order to avoid the clustering process is locally optimal, it is necessary to make Canopy get the center point spacing as large as possible. The maximum and minimum distance method [30] is a kind of test-based algorithm in the field of pattern recognition. Its basic idea is to take the object as far as possible as a cluster center, trying to get a better initial division. The algorithm not only intelligently determines the number k in the initial clustering, but also improves the efficiency of dividing the initial data set.

a) The description of improved Canopy algorithm

The Euclidean distance method is used to measure the degree of dissimilarity between data objects. Set the data set, S={X1,X2,…,Xn}, and the initial cluster center set is V = {v1, v2, ..., vn}. The improved Canopy algorithm is described as follows:

Input: Improve the initial clustering center candidate point of the UPGMA algorithm output, the parameter θ;

Output: Optimize the initial clustering center.

(1) Arbitrarily select a data object from the data set S as the first cluster center point v1 and put it into V;

(2) Calculate the distance between v1 and all the data objects remaining in the data set S, and put the farthest data object into V as the second cluster center v2;

(3) Calculate the distance Di between all the data objects Xi and all the data objects remaining in the data set S, select the smaller distance and denote Min (Di);

(4) Selects the maximum value in all the Min(Di) , marked as Max (Min (Di)), and regard the corresponding data Xi as the candidate cluster center, then judgment is made by the discriminant formula Max (Min (Di))> θ ‖ v1-v2 ‖. If the condition is satisfied, Xi is added to the initial clustering center set V, and if it is not satisfied,

(5) To (3);

(6) Output optimization of the initial clustering center.

The most critical step in the improved Canopy algorithm is the step (4), which takes the corresponding point of Max as the candidate of the new clustering center, thus avoiding the fact that the distance from an existing clustering center is closer to the other Clustering centers are far away as candidates for possible candidates. Therefore, the algorithm can be used to ensure that each new clustering center is far from the distance of the existing clustering center.

b) The analysis of advantages and disadvantages of improved algorithm of Canopy

The improved Canopy algorithm can use the k data objects farthest from each other in the data set as the initial clustering center, so as to avoid the situation that the initial clustering center distribution is too concentrated and intensive. But on the one hand, it is possible to select the noise data and the edge data, making the algorithm easy to fall into the local optimal solution, it is difficult to get the global optimal solution.

On the other hand, if the sample size of the whole data set is n, we need to scan the database first if we want to find a new cluster center each time; After finding the nearest distance from each object to the existing cluster center, we need scan the database to get the maximum-minimum distance. so we need a total of 2n distance calculation. The time complexity of the improved Canopy algorithm is: O (nk) .if the k clustering centers need to be found in the end of algorithm. Therefore, the computational complexity of the improved Canopy algorithm depends on the size of n, and there are thousands of objects in large databases usually, if we treat the original data set with the improved Canopy algorithm directly, the implementation efficiency is low and the required storage space will be significantly increased.

3) MCU-kmeans algorithm

Generally, in order to ensure fully reflecting the distribution of data in the entire data set, every cluster center should be distributed in the high density area of the data set and dispersed as much as possible. Based on the above considerations, this paper proposes the MCU-kmeans algorithm, which combines the improved UPGMA algorithm and the improved Canopy algorithm to obtain the optimized initial clustering center, and then apply these optimized initial clustering centers to the k-means algorithm to enhance the clustering effect. Among them, the improved UPGMA algorithm is used to find the high density region, so that the selected initial clustering center candidate point away from the noise data and edge data; And the improved Canopy algorithm is used to avoid that the initial clustering center distribution is too concentrated and dense to ensure that the distances between the cluster center points are far away, which fully reflect the overall distribution of the data set. Therefore, the improved UPGMA algorithm and the improved Canopy algorithm complement each other so that the initial clustering centers selected by the algorithm are far apart from each other and all are located in the high density region of the data set. To sum up, the CMU-kmeans algorithm is as follows.

a) The initialization of the cluster center;

- Improved UPGMA algorithm: obtain the initial clustering center candidate point;
- Improved Canopy algorithm: obtain the appropriate initial clustering center;

b) K-means algorithm iteration;

c) The assessment of clustering results.

It can be seen that the framework of the CMU-kmeans algorithm is divided into three phases, as shown in Figure 2, the first stage of the algorithm is the initial optimization algorithm, which is the most important part of the improvement. The purpose is to intelligently capture the original The optimal initial clustering seed and the optimal initial clustering number of the data set distribution. The second stage is the main body of the algorithm, and the K-means algorithm is used to cluster on the whole data set and get the clustering result. The third stage is experiment and evaluated to verify the validity of the proposed CMU-kmeans algorithm.

It can be seen that the framework of the CMU-kmeans algorithm is divided into three phases, as shown in Fig.2 , the first stage of the algorithm is the initial optimization algorithm, which is the most important part of the improvement. The purpose is to intelligently capture the optimal initial clustering seed and the number of the data set distribution. The second stage is the main body of the algorithm, and the K-means algorithm is used to cluster on the whole data set and get the clustering result. The third stage is experiment and evaluated to verify the validity of the proposed CMU-kmeans algorithm.

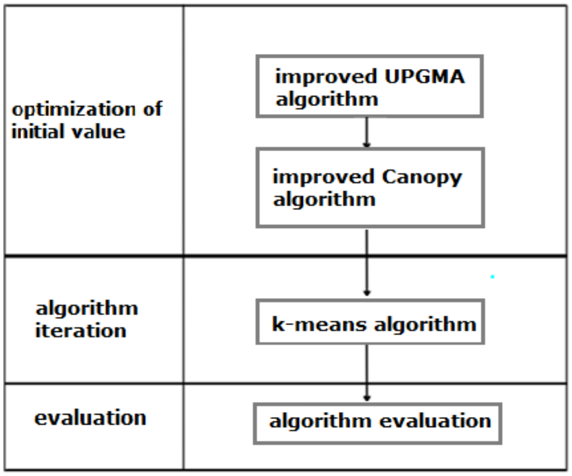

Figure 2. CMU-kmeans algorithm framework

The CMU-kmeans algorithm proposed in this paper can effectively reduce the dependency of the k-means algorithm on the initial clustering center selection. For the data set with uneven data distribution, on the one hand, it avoids the idea that the initial clustering center is too dense; On the other hand, it avoids the fact that the selected initial clustering centers are too scattered and even select noise data and edge data is happening, which can improve the stability and validity of the algorithm. At the same time, the number k of the initial clustering center can be automatically determined without the pre-set and the simplicity and maneuverability of the algorithm can be improved.

III. EXPERIMENT ANALYSIS

A. Experimental description

The data set selected from the experiment comes from the rainfall data collected in the geological hazard detection system and the rainfall data set after the artificial noise is added. The experimental environment is: Inter(R)Core(TM)i3-2330M,4G RAM，250G hard disk，Win7 operating system.

In order to verify the validity and stability of the algorithm, the traditional K-means clustering algorithm, the improved Canopy algorithm and the CMU-kmeans algorithm are compared under the rainfall data set. The clustering result of the traditional k-means algorithm is an average of 10 executions. Evaluate the performance of the algorithm according to the accuracy of the clustering results and the recall rate.

B. Performance evaluation criteria

The traditional k-means algorithm, the improved Canopy algorithm and the clustering effect of CMU-kmeans algorithm proposed in this paper are evaluated by the commonly used evaluation method to evaluate the quality of clustering effect, namely, precision and recall. The accuracy and recall rate are defined as follows:

$$P(i, j) = precision(i, j) = Ni,j / Ni \quad (5)$$

R（i, j）= recall(i, j) = Ni,j / Nj (6)

Here, Ni, j represents the number of classes i in cluster j; Ni is the number of all objects in class i; Nj is the number of all objects in cluster j.

C. Experimental content and structure analysis

Table3 below shows the detailed experimental results of the three algorithms on the geo-disaster monitoring system rainfall data set.

TABLE III. DETAILED EXPERIMENTAL RESULTS ON THE RAINFALL DATASET

Rainf all set	**k-means algorithm**		**Improved Canopy algorithm**		**CMU-kmeans algorithm**	
	precision	recall	Precision	recall	precision	recall
1st	25.423	26.125	50.799	61.078	56.939	65.783
2nd	24.287	25.365	52.975	63.288	56.423	64.921
3rd	25.61	18.864	48.895	58.887	57.413	66.174
4th	27.143	26.143	53.425	63.683	57.682	68.108
5th	22.365	25.31	50.073	58.404	56.163	65.063
6th	18.102	26.421	50.444	64.65	56.468	65.224
7th	25.326	24.623	49.362	57.338	58.921	67.638
8th	28.325	26.852	49.975	60.075	56.239	66.405
9th	26.562	28.154	54.267	62.392	58.341	66.423
10th	23.985	26.523	51.445	60.651	57.267	65.392
average	25.013	25.938	51.666	61.045	57.186	66.113

As can be seen from the above table, in ten experiments, values of the two performance evaluation criteria (precision and recall)vary greatly based on the traditional k-means algorithm ,showing a very unstable state. To precision as an example, the minimum value of the ten experimental results is 18.102, and the maximum is 28.325, the difference is 10.323, and the recall is different from 9.290. The result of the improved Canopy algorithm has improved, the precision is 6.549, and the difference is 6.345.

In the CMU-kmeans algorithm, the values of the two performance evaluation criteria are obviously improved and are still stable. The precision of the ten results is 56.163, the maximum is 58.921, the difference is 2.758, and the recovery value is 3.187.

In order to make the experimental results more straightforward, the above 10 experimental results with the wave diagram shown in order to compare the stability of the two algorithms and accuracy.

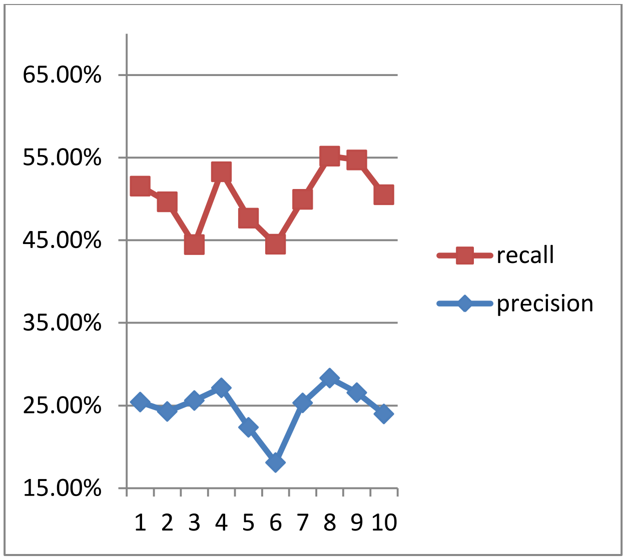

Figure 3. The precision and recall values of the traditional k-means algorithm

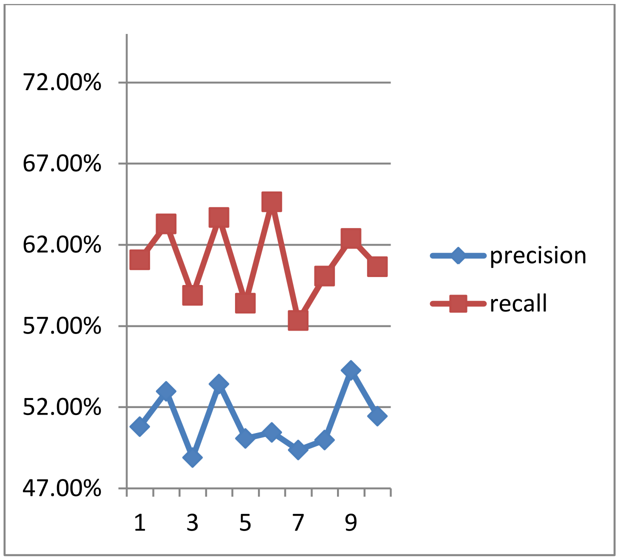

Figure 4. The precision and recall values of the improved Cannopy algorithm

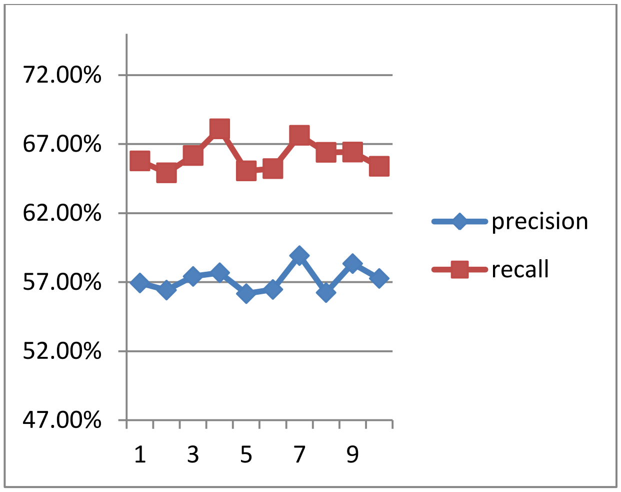

Figure 5. The precision and recall values of the CMU-kmeans algorithm

It can be seen from Fig.3 , Fig.4 and Fig.5 that the improved Canopy algorithm clustering effect is improved on the basis of the traditional k-means algorithm. The improved CMU-kmeans algorithm can improve the state of the improved Canopy algorithm. The precision and recall value of the CMU-kmeans algorithm are small and obviously improved, and the lifting effect is significant.

IV. CONCLUSION

The CMU-kmeans algorithm improves the clustering effect, make the performance tend to be stable, and the computational complexity of the calculation is obviously reduced compared with the traditional k-means algorithm and the improved Canopy algorithm. Also, the algorithm can adaptively determine the number k of the initial clustering center, avoid the influence of the noise data and the edge data and random selection of initial clustering center, and also well reflect the actual distribution of clustering center in the dataset.

REFERENCES

[1] Zhai D H, Yu J, Gao F, et al. k-means text clustering algorithm based oninitial cluster centers selection according to maximum distance [J]. Application Research of Computers, 2014, 31(3):379 – 382.

[2] Baolin Yi, Haiquan Qiao, Fan Yang, Chenwei Xu. An Improved Initialization Center Algorithm for K-Means Clustering[C]. Computational Intelligence and Software Engineering, 2010, pp:1-4.

[3] Redmond S J, Heneghan C.A method for initializing the K-means clustering algorithm using kd-trees[J]. Pattern recognition letters, 2007, 28(8):965-973.

[4] Liu J X, Zhu G H, Xi M. A k-means Algorithm based on the radius [J]. Journal of Guilin University of Electronic Technology,2013,33(2):134-138.

[5] Habibpour R, Khalipour K. A new k-means and K-nearest-neighbor algorithms for text document clustering [J]. International Journal of Academic Research Part A, 2014,6(3) : 79 — 84.

[6] Data mining techniques and applications - A decade review from 2000 to 2011[J]. Shu-Hsien Liao, Pei-Hui Chu, Pei-Yuan Hsiao. Expert Systems With Applicati--ons . 2012 (12).

[7] Application of Improved K-Means Clustering Algorithm in Transit Data Collection. Ying Wu, Chun long Yao. 20103rd International Conference on Biomedical Engineering and Informatics (BMET）. 2010.

[8] Zhou A W, Yu Y F. The research about clustering algorithm of K-means [J]. Computer Technology and Development, 2011,21(2):62-65.

[9] Duan G Q. Auto generation cloud optimization based on genetic algorithm [J]. Computer and Digital Engineering, 2015,43(3):379-382.

[10] Wang C L, Zhang J X. Improved k-means algorithm based on latent Dirichlet allocation for text clustering [J]. Journal of Computer Applications,2014,34(1):249-254.

[11] Deepa V K,Geetha J R R. Rapid development of applications in data mining[C]. Green High Performance Computing (ICGHPC),2013,pp:1-4.

[12] Sharma S, Agrawal J, Agarwal S, et al. Machinelearn-ing techniques for data mining: A survey[C]. Com-putational Intelligence and Computing Research (ICCIC),2013,pp:1-6.

[13] Efficient Data Clustering Algorithms: Improvements over Kmeans[J]. Mohamed Abubaker, Wesam Ashour. International Journal of Intelligent Systems and Applications(IJISA). 2013 (3).

[14] Fahad A, Alshatri N, Tari Z, Alamri A. A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis[C]. Emerging Topics in Computing.2014:267-279.

[15] Abubaker M, Ashour Wesam. Efficient data clustering algorithm algorithms:improvemwnts over k-means[J]. I nternational Journal of Intelligent Systems and Applications.2013(3):37-49.

[16] Tang Zhaoxia, Zhang Hui. Improved K-means Clustering Algorithm Based on Genetic Algorithm[C], Telkomnika Indonesian Journal of Electrical Engineering.2014, pp:1917-1923.