ATLANTIS PRESS

Advances in Computer Science Research, volume 79

2nd International Conference on Sensor Network and Computer Engineering (ICSNCE 2018)

Improved K-means Algorithm Based on Optimizing Initial Cluster Centers and Its Application

Xue Linyao
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
e-mail: xuelinyaoyao@foxmail.com

Wang Jianguo
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China

***Abstract*—Data mining is a process of data grouping or partitioning from the large and complex data, and the clustering analysis is an important research field in data mining. 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. By constantly iterating, the distance between each data object and the center of its subclass is minimized. Because K-means algorithm is simple and efficient, it is applied to data mining, knowledge discovery and other fields. However, the algorithm has its inherent shortcomings, such as the K value in the K-means algorithm needs to be given in advance; clustering results are highly dependent on the selection of initial clustering centers and so on. In order to adapt to the historical data clustering of the geological disaster monitoring system, this paper presents a method to optimize the initial clustering center and the method of isolating points. The experimental results show that the improved k-means algorithm is better than the traditional clustering in terms of accuracy and stability, and the experimental results are closer to the actual data distribution.**

Keywords-Clustering Analysis; Improved K-means 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. The research on the deficiency of K-means algorithm is divided into two branches: 1) the number of initial clustering centers K; 2) the choice of initial clustering center. In this paper, we mainly study the latter, and propose a new initial clustering center 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 algorithm is better than the traditional clustering in terms of accuracy and stability, and the experimental results are closer to the actual data distribution.

II. BRIEF AND RESEARCH STATUS OF K-MEANS ALGORITHM

A. Overview of K-means Algorithm

The K-means algorithm is a classical unsupervised clustering algorithm. The purpose is to divide a given data set 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: Firstly, k data objects are randomly selected from the sample set X as the initial clustering center; Secondly, according to the degree of similarity between each data object and k clustering centers, it is allocated to the most similar clusters; Then recalculate the average of each new cluster and use it as the next iteration of the clustering center, and repeat the process until the updated cluster center is consistent with the update, that is, the criterion function E converges.

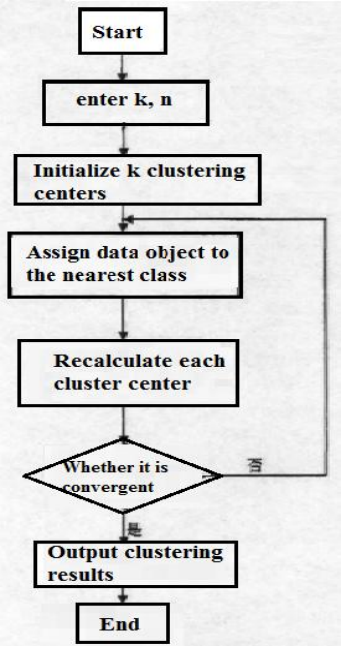

Figure 1. K-means flow

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:

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

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

$$E =\sum_{i=0}^{n} \sum_{x \in c_i} |x - \overline{x}_i|^2 \quad (2)$$

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

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 Analysis of Shortcomings of K-means Algorithm

1) *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.

2) *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.

3) Sensitive to noise and isolated points. The traditional K-means algorithm requires constant adjustment of the cluster center, and the noise and isolation points usually affect the quality of the clustering results. Therefore, if the noise and isolation points are selected, the resulting clustering center of data is likely to deviate from the actual data-intensive area, resulting in inaccurate clustering results.

III. IMPROVEMENT OF K-MEANS ALGORITHM AND ITS APPLICATION

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 2015 to 2016, 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 3D 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

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. The algorithm is based on the small cluster with the largest variance and can be divided into two clusters with different variance. The algorithm of initializing center is proposed. In addition, a method of isolating points has been proposed. The idea of this algorithm is to first find out the two points furthest from the sample point as the initial center point, and then divide the other sample points into the cluster to which the nearest center point belongs, and determine the number of points within the cluster And whether the corresponding initial clustering center is an isolated point, and finally select the next object to be split according to the variance within the cluster and update the initial cluster center according to certain rules. The above steps are repeated until the number of cluster centers is satisfied.

1) Initial clustering center selection algorithm

X={x1,x2,x3…xn}, n is the number of samples. d (xi , xj) (i, j ∈ {1,2⋯n}) is the Euclidean distance between the data points xi and xj, ci (iϵ1,2 ⋯ n} is the clustering center, Q is the data object that will be spited, S is the number of clustering centers.

The initial clustering center selection algorithm is as follows:

Input: data set X, number of clusters k, threshold u

Output: cluster center set C and isolated point set D

a) Let Q=X={x1,x2,x3,……xn};S=0;

b) Calculate the Euclidean distance d (xi,xj) between the two data points in W, and find the two points xi,xj, which are marked as ci, cj, and

let:

S = S + 2;

Qi = {xp | d(xp,xi)< d(xp,xj),xp ϵ Q},

Qj = {xp | d(xp,xj)< d(xp,xi),xp ϵ Q},

which means Q is divided by the cluster ,and Qi and Qj become the split clusters.

c) *If the number of data objects in Qi or Qj is less than u, the selected initial center xi or xj is an isolated point.* Remove xi or xj from Q, remove ci or cj in set C, and add xi or xj to D, return to step 1;

d) If the number of data objects in the set C is less than k, find the set Qp with the largest variance in the splitting cluster and let Q=QpQp ， S=S-1, then remove cpcp the set C;

e) Calculate the mean of all the objects in the split cluster, and the resulting mean is k initial clustering centers.

2) Improved K-means algorithm

Data set X={x1,x2,x3,……xn}, there are n objects. Cold,i represents the i-th cluster center of the previous round, Cnew,,i represents the new cluster center calculated in current time, and the algorithm is described as follows:

Input: data set X, number of clusters k, threshold u

Output: k clusters and the number of bands

a) Call the improved initialization center selection algorithm to get the initialization center, if there is an isolated point will be isolated points alone in a class, do not participate in the follow-up clustering algorithm;

b) Calculate the distance between all data objects and k cluster centers, and assign the text to the nearest cluster;

c) Calculate the mean of each cluster to obtain a new round of cluster center;

d) If $E' = \sum_{i=1}^{k} \sum_{x \in c_i} \left| C_{o,i} - C_{n,i} \right|^2$ *<10-10E'* $= \sum_{i=1}^{k} \sum_{x \in c_i} \left| C_{o,i} - C_{n,i} \right|^2$ *<10-10, then the iteration is*

terminated, otherwise it returns to 2). (note: E' is the measure function)

IV. 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 improved algorithm, the original k-means algorithm, the algorithm in literature [4]and the improved algorithm are analyzed and compared under the rainfall data set. In order to further verify the superiority of the algorithm in dealing with isolated points, the algorithm is compared with other algorithms on the rainfall data set after adding noise. The clustering results are clustered and criterion function changes and the clustering time are used to evaluate the clustering results.

B. *Experimental Results and Analysis*

The clustering criterion function of the two algorithms will decrease with the increase of the number of the adjustment of the cluster until the final convergence, and the more compact the two curves, the higher the accuracy of the corresponding clustering results The vice versa. Figure 2 is the comparison of the traditional k-means algorithm and the improved algorithm standard function values with the clustering cancroids adjustment and constantly changing the comparison chart.; In order to test the speed of the improved algorithm in this paper, three samples were randomly selected from the historical data of the geological hazard system, and the sample capacity was 5000，10000 and 18000 respectively. The experimental results are shown in Figure 3.

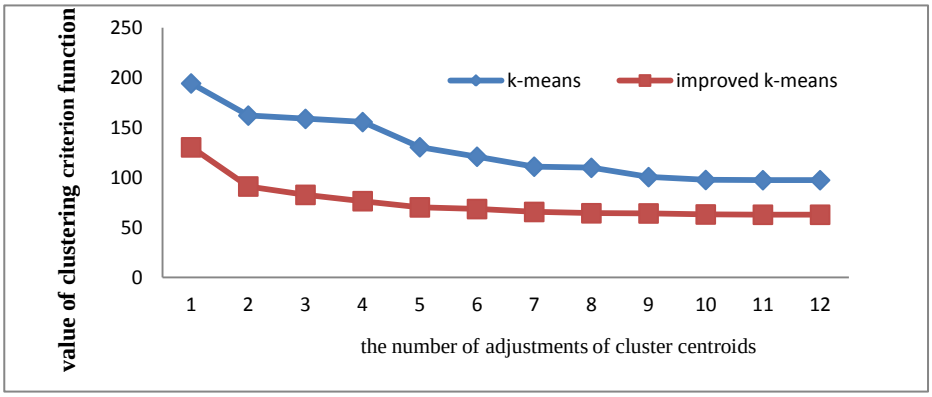

Figure 2. The comparison of criterion function changes trend graph

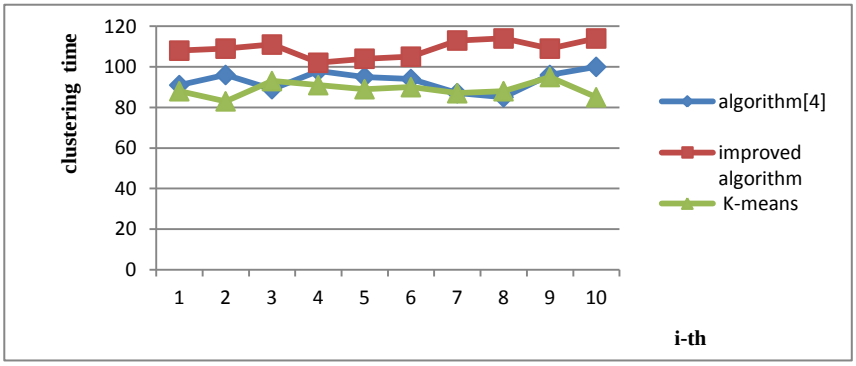

Figure 3. The comparison of clustering times in artificial data sets

According to the comparison of criterion function change trend graph in the rainfall data set in Figure 2, the clustering criterion function of the improved algorithm is superior to the clustering criterion function value of the traditional k-means algorithm because the data objects in the optimized cluster are more compact and independent in each iteration process, the criterion function value is significantly lower than the traditional k-means algorithm, which also further validates the superiority of this algorithm; Figure 3 shows that the traditional k-means algorithm is running at the fastest speed, and the speed of this algorithm is slightly lower than that of algorithm[4].

V. CONCLUSION

Aiming at the instability of the clustering results caused by the random clustering of the traditional k-means algorithm and the effect of the isolated points on the clustering results, the authors of this paper have the advantages of small distance from the large sample points to the same cluster The clustering algorithm with the largest variance of variance can be split into two clusters with relatively small variance, a k-means clustering algorithm is proposed to optimize the initial clustering center. Simulation experiments in geological hazard systems and artificial data sets with the same proportion of noise show that the proposed algorithm improves the accuracy and clustering error compared with the traditional k-means algorithm and the other two optimization initial center algorithms. However, the initial algorithm of the algorithm is somewhat complicated, and it takes too much time in the selection of the central problem. In the future work, it will be further improved, and it will be tried in all respects.

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 Applications . 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. Machine learning techniques for data mining:A survey[C].Computational 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.

[17] Optimal variable weighting for ultrametric and additive tree clustering[J]. Geert Soete. Quality and Quantity. 1986 (2).