K-means Algorithm Based on Fitting Function

SiYong Chu[1,a],YanNi Deng[2,b],LinLi Tu[3,c]

WuHan university of technology,WuHan,430000,china

WuHan university of technology,WuHan,430000,china

WuHan university of technology,WuHan,430000,china

Keywords: Density,Optimal distance, Fitting function,K-means

Abstract. The K-means algorithm has the shortcomings of being sensitive to the initial clustering center, and in order to overcome this drawback, in this paper ,on the basis of the combination of data density and the optimal distance , a new definition of fitting function is made and then a kind of K-means algorithm based on fitting function is proposed. By utilizing the fitting function to select the initial clustering center, the selection of the initial cluster centers can be made as much close to the real sample clustering centers as possible. The experiments proved that, the K-means algorithm based on fitting function reduces the number of iterations and enhances the stability of the algorithm, as well as improves the efficiency of the algorithm.

Introduction

K-means algorithm [1] as a kind of classic clustering algorithm, has been widely used in many field such as data mining, image processing and pattern recognition.It finds the center of the cluster via several iterations, and assigns all points that need to be addressed to the clusters where the centers of the clusters are nearest to the specific points, according to the principle of proximity distribution. The purpose of multiple iterations is to make the degree of similarity lowest between clusters and highest within the clusters. Since the initial cluster centers k-means algorithm is randomly selected, although it can reduce the complexity of the algorithm and improve the scalability of the algorithm, it leads to the shortcoming of an unstable algorithm performance as well.

In order to improve the k-means algorithm, researchers have proposed various methods. In literature [2] [10], an improved method based on Huffman tree thinking is proposed. it uses the distances between each data points as the weights to build Huffman tree, and then chooses the cluster center according to the value of k in the reverse order of Huffman tree .In literature[3][4][5],in order to make the selected initial cluster centers far away from each other, the maximum distance method and the maximum and minimum distance method are utilized respectively. In literature[11], using this method effectively overcomes the issue of the algorithm being sensitive to the initial cluster centers, the method ensures the efficiency of the algorithm even in the situations where the center of initial cluster chosen randomly is bad. In literature[6][7][12],combined with the method of minimum spanning tree, based on the distance between the data points, a minimum spanning tree is constructe. And then the initial clustering center point can be obtained by cropping the minimum spanning tree according to the value of k. The idea of the density is combined with maximum distance method in [8][9] , to select the farthest distance and highest density points as the initial cluster centers. In literature[13], the concept of parallel is introduced to significantly reduce the running time of the algorithm and better adapt to the large-scale data.

The above related research of the improvement strategy, although overcomes some of the shortcomings of the traditional k-means algorithm to some extent, the efficiency and the results of the clustering algorithms still need to be further improved. Towards this end, this paper defines a fitting function of k-means algorithm (K-means algorithm based on fitting function, referred to as FKM).

FKM algorithm

The core idea of FKM algorithm is using the density method and the optimal distance method together to make the selected initial cluster centers not only strong representative, but also far away from each other. First, the density method can rule out the possibility of selecting the isolated points as the initial cluster centers, which improves the algorithm resistance of isolated points. Then, by using the optimal distance method, the possibility that the selected initial cluster centers are too close to each other can be effectively avoided, thus avoiding the algorithm becoming a local optimal solution.

Density function. In the field of data mining, the more points there are around a point, i.e. the more densely are the points distributed around a point, the more important the point is. In order to better reflect the importance of each point, the density function is defined as follows:

$$V(i)=\frac{\sum_{j=1}^{n} d_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}} \tag{1}$$

Where dij is :

$$d_{ij}=\sqrt{\sum_{r=1}^{m}\left(X_{ir}-X_{jr}\right)^2},\quad i,j=1,2...n \tag{2}$$

where Xi represents any one point of the data set, n is the size of the data set, m is the number of Xi properties. By equation(1) we can know the smaller V(i) is, the more significant the corresponding Xi is, and vice versa.

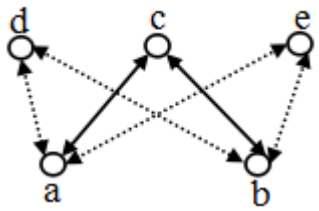
Optimal distance. During the implementation of k-means algorithm, we always want the randomly chosen initial cluster centers to be scattered, so as the results of the algorithm are more favorable. The selected initial cluster centers distance between each other as far as possible, i.e., in the hope that all the distances between the center points are as much equal as possible, that is, to achieve the optimal distance between each other. Towards this end, we use the average value inequality method to select the clustering center to obtain the optimal distance. Mathematical deduction as follows:

$$(d_1-d_2)^2\geq 0$$

$$\Downarrow$$

$$\frac{d_1^{\,2}+d_2^{\,2}}{2}\geq d_1d_2 \tag{3}$$

(3) is a two-dimensional space of the mean inequality, d1d2 has the maximum value, if and only if d1=d2. To have a better understanding, see the image of Figure.2.1,choosing the point C as the third center point is most appropriate.



To find the optimal distance diagram of Figure.2.1

To the n-dimensional space is derived:

$$\left(\frac{d_1+d_2+\cdots+d_n}{n}\right)^n\geq d_1d_2\cdots d_n \tag{4}$$

If and only if $d_1=d_2=...=d_n$, d1d2..dn has the maximum value.

Fitting function and the initial cluster center point selection method. If a random point or the point of maximum density is selected for the first initial cluster centers, or the most remote two points are selected for the start of the initial clustering center, the initial clustering centers selected this way tend to be affected by isolated points. In order to effectively avoid the defect of a separate method for selecting initial cluster centers and embody the advantages of the two algorithms, we define a new fitting function which takes the form:

$$F_1(i,j)=\frac{d_{ij}}{V(i)+V(j)},i,j=1,2...n \tag{5}$$

$$F_2(i)=\frac{\prod_{s=1}^{m} d_{is}}{V(i)} \tag{6}$$

where m represents the number of initial cluster centers.

Through analyzing (5) and (6) we can see that the initial clustering centers selected by the proposed fitting function not only have high density, but also are scattered apart. From another perspective, one can also know that, the higher the value of the fitting function of data points is, the smaller is the difference between the calculated value and value of the true cluster center. Therefore, the clustering results produced by the chosen initial clustering center are preferable. When selecting the initial cluster center point , we utilize equation (5) to select the initial two initial cluster centers, subsequent initial cluster centers are chosen by equation (6).

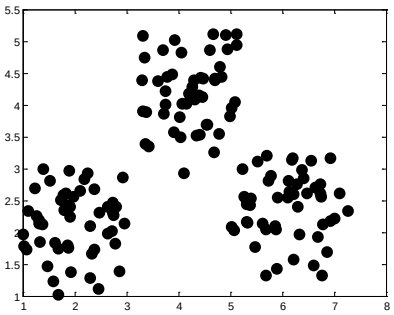


Fig.2.2 (a) the data set

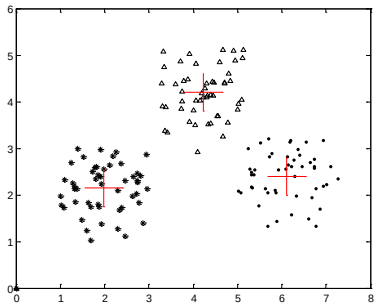


Fig.2.2 (b) correct classification

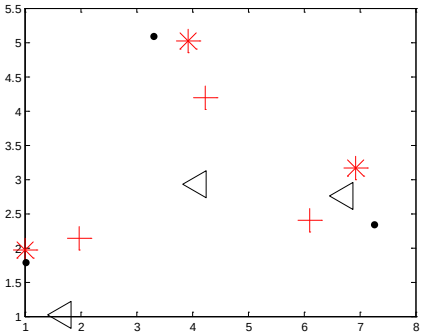


Fig.2.2 (c) the contrast of the initial cluster centers

Figure2.2 (a) shows a sample data set, figure2.2(b) shows the actual classification of the sample, "+" indicates the cluster center of each cluster. For each sample in figure2.2(a), we use a method based solely on data density, and another method solely on the optimal distance method and the proposed method based on fitting function approach to select the initial cluster centers and the results are represented by " △",and "*" respectively. Figure c depicts the comparison between the selected cluster center points and the actual cluster centers.

By comparing the points in figure c one can clearly see that the initial cluster centers selected by the proposed method based on the fitting function is closest to the actual cluster centers. The second close is the initial cluster centers selected by the method based on the optimal distance, the farthest

is selected by the method based solely on density. In other words, the initial cluster center selected by the method based on fitting function has the highest quality.

Algorithm processes.

Specific steps are as follows:

Input: Data set U and the value of clusters number k;

Output: k clusters and each cluster center point;

Step 1: calculate the distance between any two points in the data set U according to equation (2);

Step 2: calculate the density value of each point according to equation (1) ;

Step 3: find the start of the two initial cluster centers via (5), set k1 = 2;

Step 4: if (k1>k), then the initialization of the center of the cluster is completed, turn to step 5;

If(k1<k), the use equation (6) to find the next cluster center, set k1 = k1 + 1, repeat step 4;

Step 5: Using the initial cluster centers that have been found to conduct the k-means algorithm.

Experiment

In order to prove the effectiveness of the proposed FKM algorithm, we conduct a comparison test between the FKM algorithm and the k-means algorithm. Select the iris data set in prestigious UCI[14] machine learning database and the seed data set for the test. We utilize three indicators to evaluate the quality of clustering results, that is, accuracy (the number of correct classification data point / the size of data set), stability and the time consumed. In order to better reflect the performance of the two algorithms, each data set on the two algorithms are running 10 times. In order to ensure the authenticity of the experimental results, we do not do anything with the data set in advance, the two data sets are described in Table 3.1.

Table 3.1 Description of the two data sets

Data Set	Size	Attribute	Classes
iris	150	4	3
seed	210	7	3

Table 3-2 Accuracy rate comparison table

Data Set	Algorithm	Precision (The number of correct classification / The size of data set)		
		M aximum (%)	M inimu m (%)	Mean (%)
iris	k-means	89.33	51.33	81.40
	FKM	89.33	89.33	89.33
seed	k-means	89.52	66.19	75.76
	FKM	89.52	89.52	89.52

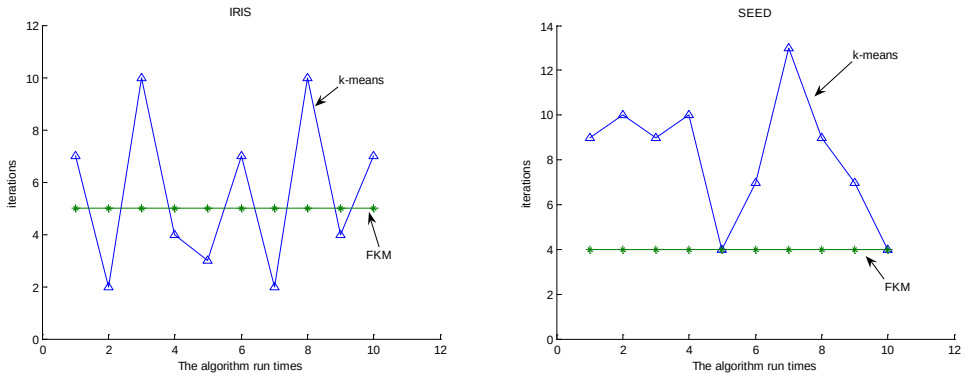


Fig.3.1 algorithm iterations comparison chart

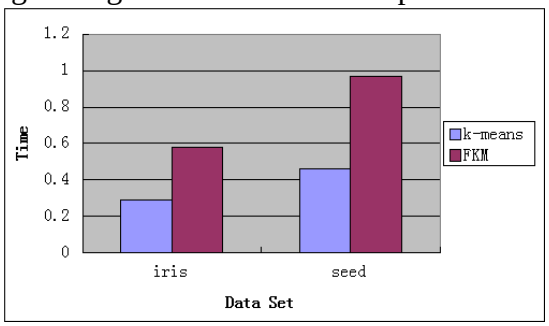


Fig.3.2 algorithm time-consuming comparison chart

According to Table 3.2, one can clearly know that, on the basis of the two data sets, the highest accuracy of FKM k-means algorithm is the same as k-means algorithm, which is a constant, but the lowest accuracy of k-means algorithm is very low, which lead to the fact that the average accuracy rate of k-means algorithm is considerably lower than the average accuracy of FKM algorithm. So from the aspect of accuracy, FKM algorithm is superior to k-means algorithm.

In contrast to the trend of the curve in Figure 3.1, both in the iris data and seed data set,can clearly see k-means algorithm in the process of experiment the number of iteration to fluctuate in large range, while the of FKM algorithm is stable and unchanging, and of k-means algorithm is also greater than the average value of iterative number FKM algorithm .So from the stability index,FKM algorithm is also superior to the k-means algorithm.

The comparison of the time-consuming characteristic of the two algorithms can be obtained from Figure 3.2. Due to the fact that the complexity of the FKM algorithm is slightly higher than the k-means algorithm, we can see that the time consumed by the FKM algorithm is approximately 2 times than that of k-means algorithm. However, since the accuracy and stability of FKM algorithm is better than k-means algorithm, so by taking the comparison of the three indicators into consideration, FKM algorithm performance is better than that of k-means algorithm.

Summary

In view of the problem that the k-means algorithm is sensitive to the initial cluster centers, the FKM algorithm proposed in this paper finds the initial cluster centers by a novel fitting function, in order to guarantee that the initial cluster centers found not only strongly representative, but also relatively scattered. The experimental results show that FKM algorithm improves the accuracy of the algorithm ,as well as strengthened the stability of the algorithm, although the time consumption still have defects, but within a tolerable range. This problem will be further investigated issue in future research.

References

[1]MabBao qiu,Lian Cui-ling,Zhao Xu.Application of K-Means partition type clustering algorithm based on distance.Journal of the Hebei Academy of Sciences,2013,30(4):17-22.

[2]Wu Xiaorong ,Yang Sheng.An Improved K-means Algorithm.Science & Technology magazine online.

[3]ZHAI Dong-hai,YU Jiang,GAO Fei,YU Lei,D1NG Feng.K-means text clustering algorithm based on initial cluster centers selection according to maximum distance.Application Research of Computer,2014,31(3):713-719.

[4]Tian tenghao.K-Means algorithm of optimized Initial clustering centers.Network Security Technology & Application ,2014,(6):42-44.

[5]ZHOU Juan,XIONG Zhong-Yang,ZHANG Yu-Fang,REN Fang.Multiseed clustering algorithm based on max-min distance means.Computer Application,2006,26(6):1425-1427.

[6]FENG Bo,HAO Wenning,CHEN Gang,ZHAN Donghui.Optimization to K-means initial cluster centers.Computer Engineering and Applications,2013,49(14):182-185.

[7]OUYANG Hao,CHEN Bo,HUANG Zhen-jin,WANG Meng,WANG Zhi-wen.MST Clustering Algorithm Based on K-means.Modular Machine Tool&Automatic Manufacturing Technique,2014,(4):14-45.

[8]WAN Guang-tong,WANG Xing-feng.Weighted K-Means based on density,2013.38(4):146-148.

[9]FU Baolong,ZHANG Aike.K-means clustering text mining method using center estimation based on mean density.Journal of Chongqing University of posts and Telecommunications(natural Science Edition),2014,26(1):111-116.

[10]Shunye Wang.An Improved K-means Clustering Algorithm Based on Dissimilarity. International Conference on Mechatronic Sciences, Electric Engineering and Computer (MEC).2013(11).

[11]Grigorios Tzortzis,Aristidis Likas.The MinMax k-Means clustering algorithm.Pattern Recognition.47(2014)2505-2516.

[12]Caiming Zhong,Mikko Malinen,Duoqian Miao,Pasi Fränti.A fast minimum spanning tree algorithm based on K-means.Information Sciences.295(2015)1-17.

[13]You Li,Kaiyong Zhao,Xiaowen Chu, Jiming Liu.Speeding up k-Means algorithm by GPUs.Journal of Computer and System Sciences.79(2013)216-229.

[14]UCI Machine Learning Repository[EB/OL].http://ar-chive.ics.u ci.edu/ml/ datasets/Iris.