An automatic Clustering Method based on Maximum Distance

Hongbo Zhou, Yongqiang Feng, Juntao Gao

Northeast Petroleum University,Institute of Computer and Information Technology,Daqing,163318 Heilongjiang
China PetroChina Pipeline Company,Information Center,Langfang,065000 Hebei
Northeast Petroleum University,Institute of Computer and Information Technology,Daqing,163318 Heilongjiang

Summary
Traditional k-means clustering algorithm needs to determine cluster number and select initial cluster centers in advance,considering the shortcomings of k-means algorithm,the paper proposed an improved efficient clustering algorithm.The algorithm does not require pre-determined number of clusters and initial cluster centers,select cluster centers according to the principle of maximum distance,divide cluster according to the principle of minimum distance,and determine the optimal cluster partition according to the distance evaluation function.The improved algorithm avoids the choice of the value of cluster number and initial centers.Hence this method can produce more accurate clustering results than the standard k-means algorithm.Experimental results show that the improved algorithm has good performance and high time efficiency.

Key words:
Euclidean distance, cluster center, maximum distance principle, minimum distance principle, distance evaluation function

1 Introduction

The k-means[1] algorithm is one of the most popular partitional clustering used in variety of domains; nevertheless the k-means algorithm has two major problems.Firstly,the most difficult thing in using K-means algorithm to make clustering analysis is to determine the value of cluster number, which will directly affect the final clustering results.Secondly,K-means algorithm has a great dependence upon the selection of the initial cluster centers,K-means algorithm always falls into local minimum value and affects the final operating efficiency and effectiveness of the clustering algorithm.These problems have greatly limited its application.

For the shortcomings of the above k-means algorithm,this paper presents an automatic clustering method based on maximum distance,improved algorithm does not require pre-determined value of k,k only needs to pre-determined range of values.Without prior easily select several poly-point as the initial cluster center,only need to select farthest focal point as the initial cluster centers one by one according to Euclidean distance.Each cluster center corresponds to a cluster,the value of distance evaluation function is calculated corresponding to the different number of clusters.To determine the most appropriate value of k depending on the distance evaluation function,the most appropriate value of k correspond to the most appropriate k clustering centers,the divided cluster as the center of the k cluster centers is optimal partitioning clustering.Experimental results show that the algorithm is simple and practical,easy to operate,greatly improving the efficiency of clustering.

2 Related Concepts

2.1 Euclidean distance

Typically spatial clustering algorithm is built on the basis of various distances,such as the Euclidean distance,Manhattan distance, and Ming Cowes distance.Among them,the most commonly used is the Euclidean distance,we use Euclidean distance as a standard clustering to analyze the improved K-Means algorithm.

There are two points i=(xi1,xi2,…,xin) and j=(xj1,xj2,…,xjn),which are two n-dimensional data object, then the Euclidean distance between them is defined as follows:

$$d_{ij} = \sqrt{(X_{i1} - X_{j1})^2 + (X_{i2} - X_{j2})^2 + \ldots + (X_{in} - X_{jn})^2} \quad (1)$$

2.2 cluster center

The average of cluster data objects is called clustering center of the cluster, the nearer the cluster center between the two clusters, the more similar the two clusters, the farther the cluster center, the less these two clusters similar. Cluster center of one cluster is calculated as follows: Let n data object containing the data set X={x1,x2,x3,...,xn}, respectively, the cluster center z1,z2,…,zk.

$$Z_i = \frac{1}{C_i} \sum_{j \in n_i} j \quad i = 1,2,\ldots,k \quad (2)$$

2.3 maximum distance principle

It is a method of selecting the cluster centers.In the division process of clustering,the cluster center as a reference point to a cluster represents a local division of clustering.A good clustering division between the cluster having the largest difference,and the division in the cluster having the maximum similarity.In the K-means

Manuscript received April 5, 2015
Manuscript revised April 20, 2015

algorithm,the similarities and dissimilarities between the two clusters is evaluated by the value of Euclidean distance.Euclidean distance is smaller,the greater the similarity,dissimilarity smaller.Conversely,the greater the Euclidean distance,the smaller the similarity,the greater dissimilarity.In order to ensure maximum similarity in same clusters and maximum dissimilarity between different clusters,different cluster centers should be possible to maximize the distance.

2.4 minimum distance principle

It is a cluster partitioning method of clustering points.In the division of the clustering process,as a reference point to the cluster center of the cluster represents a local division of clustering,if a clustering is divided into multiple clusters,a distance between any cluster point to be divided and clustering cluster center is minimum,the clustering point will be divided into the cluster including its cluster centers.

For example:a cluster with n data points,X={x1,x2,..., xn},k∈[1,n],is divided into k clusters, cluster centers for the i-th cluster xi,i∈[1,n].In the process of division of clustering,Euclidean distance is compared in turn between data points in X and of cluster centers of k clusters,the point is divided into the cluster which nearest cluster center from the point.

2.5 distance evaluation function

The k-means algorithm works as follows.First,it randomly selects k objects from the dataset,and each object then represents an initial cluster center.For the remaining objects,each of them is assigned to the cluster, to which it is the most similar.The similarity is measured based on the distance between the object and the cluster center.New centers for each cluster are then calculated.This process is continued in itera-tive fashion until a criterion function converges.Typically,the squared-error criterion is used, defined as

$$S = \sum_{i=1}^{k} \sum_{p \in C_i} \| p - m_i \|^2 \qquad (3)$$

In the above formula,p is the point in space representing a given object,and mi is the mean of cluster Ci,k is the number of cluster,S is the square-error sum for all objects in the dataset.S is smaller, the clustering division is better. S is called the distance evaluation function,which can be used to evaluate the clustering division.This criterion tries to produce k clusters so that the objects in the same cluster are as compact as possible while the objects in different clusters are as separate as possible.Clustering can be selected best division according to the different values of the distance evaluation function.

3 Algorithm Principle

K-means algorithm is generally required in advance given number of clusters k,but in most cases,the number of clusters k cannot be determined in advance,so the need for the optimal number of clusters k to optimize. To determine the optimal number[2] of clusters is very important in the K-Means algorithm,a different number of clusters corresponding to different clusters division.Some tests have been proposed cluster validity function [3-6], it is using the clustering validity function to calculate the appropriate number of clusters k,that is the best number of clusters kopt.However,because these constructors own shortcomings,is generally difficult to directly find the optimal number of clusters kopt,therefore need to determine the scope of a search is to set a kmax, making kopt≤kmax for how to determine kmax,there is no clear theoretical guidance, rules of thumb used by most scholars[4] as:kmax≤$\sqrt{n}$.

In the K-means algorithm,usually by the cluster center represents a class,so to get a better clustering results,the distance between the center of each cluster as far as possible.In view of this,select the initial cluster centers, focusing on how to maximize the distance between them as possible.

Firstly,regarding all clustering points as a cluster,and calculate the cluster's cluster centers,Euclidean distance for each cluster point to the cluster center and the value of distance evaluation function when the value of k is one.Secondly,in all cluster points,calculate Euclidean distance between any two cluster points, two farthest data points apart as two initial cluster centers.The remaining clusters points are divided into clustering its nearest cluster center belongs to them by the minimum distance principle based on these two cluster centers.Calculating its new cluster center and Euclidean distance between any two data points after the division is completed,the value of distance evaluation function is calculated when cluster centers increased by one,and the two cluster centers are saved cluster centers set at the same time.Thirdly,in all clusters respectively identify the two most distant cluster point,remove the cluster center where the cluster centers from cluster centers set,and save the largest two cluster points to cluster center set corresponding to the distance.According to the principle of the smallest cluster,the remaining points are divided into different cluster center,and formed sub-clusters.Calculating every cluster centers of sub-clusters and distance evaluation function,And so on,continue to find new cluster center from maximum Euclidean distance of sub-cluster,new cluster centers and new value of distance evaluation function is recalculated,until k value exceeds $\sqrt{n}$.Finally,compare different value of distance evaluation function corresponding different values of k,the smallest

value of distance evaluation function corresponding the value of k is best division.According to the principle of the smallest cluster,these cluster points can be easily divided into the clustering with the center of k cluster centers in cluster center set.

4. Algorithm Process

Clustering data sets N={x1,x2,x3,x4,…,xn-2,xn-1,xn} has n data objects, each data object has attributes t,i-th data object can be expressed as xi={xi1,xi2,xi3,x4,…,xit}, algorithm is as follows：

(i) Initialize data set Y and W,distance evaluation function Si,i∈[1,$\sqrt{n}$],the number of clusters k,wherein Y is stored in the divided cluster centers,and W is stored no clustering data points.Y={ Φ },W={x1,x2,x3,x4,…,xn-2,xn-1,xn}, Si =0, i∈[1,$\sqrt{n}$].

(ii) Regard all clustering points as a cluster,and calculate the cluster's cluster centers C_0 and Euclidean distance for each cluster point to the cluster center C_0, and add them as S_1, set k=1,Y={ Φ }.

(iii) Calculate Euclidean distance between any two data points in W, and choose two data points p_1 and p_2 in W which has a maximum Euclidean distance.

(iv) Add p_1 and p_2 to Y, while p_1 and p_2 is removed from W.Y={p_1,p_2}.k=k+1,W={X_i|X_i≠p_1,X_i≠p_2},i ∈ [1,$\sqrt{n}$].

(v) Regard p_1 and p_2 as center points, data point in W is divided into the clusters p_1 or p_2 according to the principle of the minimum distance. For example s is a data point in W, a is the distance of s to p_1, the distance to p_2 is b, if a ≤ b, then s will be divided into clusters p_1, otherwise s is divided into the cluster p_2. So divided until all data points in W are divided into clusters p_1 or p_2. In this case W={ Φ }.

Ⅵ Calculate the cluster's cluster centers C_1 and C_2. Euclidean distance for each cluster of the clusering to its cluster point was calculated in two clustering, add them as S_1 and S_2, set S_2= S_1 + S_2.

Ⅶ For each clustering with the center of data point in Y, calculate Euclidean distance between any two data points.Find a data point p in Y,clustering of data points p resides exist maximum Euclidean distance of a and b, then a and b will be added to Y, the clustering of data points located in all the p data points (including p) are added to the W, and removed p from Y.At the mean time set k=k+1.

For example,set Y={p_1,p_2,p_3}, calculate the Euclidean distance between any two points in every clustering with three points of p_1、p_2 and p_3 as the cluster center, the maximum value obtained respectively d_1、d_2 and d_3. If $d_1 \geq d_2$, $d_1 \geq d_3$, and a and b is the two points corresponding to the maximum Euclidean distance of d_1, then the two data points a and b will be added to Y, all clustering points regarding p_1 as the cluster center are added to the W, and p_1 is removed from Y. At the mean time set k=k+1.

Ⅷ Regard every point in Y as cluster point,these cluster points in W will be divided into the clustering which is located closer to the cluster centers according to the principle of minimum distance,until the division is completed.Calculate the center of every cluster, respectively, Euclidean distance for each cluster of the clusering to its cluster point was calculated in the clusters,add them as S_1、S_2…S_k,set S_k= S_1+S_2+…+S_k

Ⅸ Compare k and $\sqrt{n}$,if k≤$\sqrt{n}$,cycle turn step 7, otherwise ends.

Ⅹ Compare S_k,k ∈ [1, $\sqrt{n}$],the smallest S_k(S_k≠0) corresponding the value of k is the optimal number of clusters,Y in the collection is a set of k cluster centers at this time.

XI The clusters is generated with the point in Y which is regard as middle point is the desired clustering,and thus can easily obtained the corresponding clustering division.Regard every point in Y as cluster point,these cluster points in W will be divided into the clustering which is located closer to the cluster centers according to the principle of minimum distance,until the division is completed.

5. Simulation Experiment

We manually configured by a set of data to test the effectiveness of the algorithm,table 1 is spatial coordinates of study object.Figure 1 is a spatial distribution of data points.Now using the k value of spatial clustering optimization algorithm presented above,according to the empirical formula described above previous k max≤$\sqrt{n}$,there should be:k max≤$\sqrt{9}$,and therefore the range of k can be reduced to k1=1,k2=2,k3=3,solving steps are as follows:

Table 1 Spatial coordinates of study Object

	P1	P2	P3	P4	P5	P6	P7	P8	P9
x	1	2	3	3	4	2	8	9	10
y	1	2	1	4	5	5	3	2	3

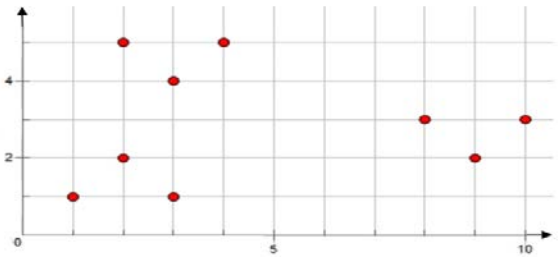

Fig 1 spatial distribution of data points

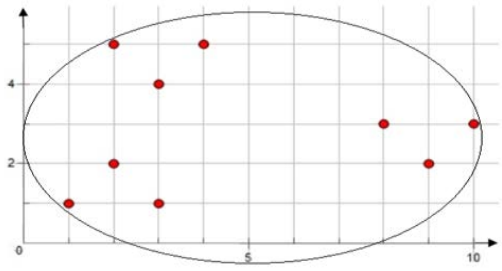

Fig 2 one cluster graphics

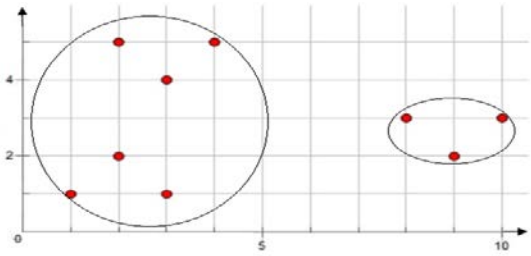

Fig 3 two clusters graphics

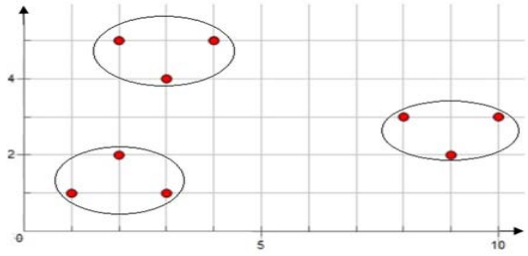

Fig 4 three clusters graphics

Table 1 is spatial coordinates of study object,Fig 1 is spatial distribution of data points.

Fig 2 is clustering result when the value of k is one,all cluster points is divided one cluster,the distance evaluation function is 30.1659.Fig 3 is clustering result when the value of k is two,all cluster points is divided two clusters,the distance evaluation function is 14.134.Fig 4 is clustering result when the value of k is 3,all cluster points is divided three clusters,the distance evaluation function is 8.3246.since 8.3246<14.134 and 8.3246<30.1659,the third divided cluster is best cluseter.

6 Conclusion

The typical spatial clustering K-means algorithm is required pre-determined value of k,but in practical applications,it is difficult to accurately determine the value of k,and k clustering centers is choosed inappropriate,it will lead to clustering divided into local optimum.therefore,in the application of the classic k-means spatial clustering algorithm needs to be further optimized and improved.

Improved algorithm does not require pre-determined value of k,k only needs to pre-determined range of values.Without prior easily select several poly-point as the initial cluster center,only need to select farthest focal point as the initial cluster centers one by one according to the Euclidean distance.

This approach avoids the choice of the value of k and initial centers.Hence this method can produce more accurate clustering results than the standard k-means algorithm.Experimental results show that the proposed algorithm has good performance and high time efficiency.

Acknowledgment

The paper is supported by the Education Department of Heilongjiang province science and technology research projects (No.1253G014).

References

[1] MacQueen J. Some methods for classification and analysis of multivariate observations[C].Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability.Berkeley :University of California Press,1967:281-297.

[2] Milligan G W, Cooper M C. An examination of procedures for determining the number of clusters in a data set.Psy-chometrika, 1985, 50: 159-179.

[3] Bezdek J C, Pal N R. Some new indexes of cluster validity [J].IEEE Transactions on Systems, Man, and Cybernetics Part B:Cybernetics, 1998, 28(3) : 301-315.

[4] Ramze R M, Lelieveldt B P F, Reiber J H C. A new cluster validity indexes for the fuzzy c mean[J] . Pattern Recognition Letters,1998, 19: 237- 246.

[5] Yu Jian, Chen Qiansheng. The range of optimal class number of fuzzy cluster[J]. Science of China (series E) , 2002, 32(2) : 274- 280.

[6] Fan Jiulun, Pei Jihong, Xie Weixin. Cluster validity function: Entropy formula[J] . Fuzzy Systems and Mathematics, 1998, 12(3) :68- 74.

Hongbo Zhou received the B.S.degrees in Computer Science from Northeast Petroleum University in 2006.His main research interests include data integration,and pattern recognition