A Highly Efficient Fast Global K-Means Clustering Algorithm

Xian Liang [1], Fuheng Qu[1] , Yong Yang[1]
[1]College of Computer Science and Technology,
Changchun University of Science and Technology,
Changchun 130022, China

Hua Cai[2]
[2] College of electronic and information engineering,
Changchun University of Science and Technology,
Changchun 130022, China

***Abstract*—To improve clustering effects of fast global K-means and reduce time complexity, a highly efficient fast global K-means algorithm is proposed in this paper. Which, maximal point of density in data sets is chosen as the first initial clustering center, then finding the next initial clustering center, firstly we can exclude a certain number of clusters around the given clustering center ,and narrow selection range of the next initial clustering center, further utilize the related theorem of triangle inequality ,reduce computational amount ,choose the sample which have great contribution in reducing error sum of squares and are apart from the given clustering center as the next initial clustering center, then the modified fast global K-means algorithm reassigns sample to cluster, which will collect sample to the weighted distance of cluster center ,and partition the sample to cluster when weighted distance is minimum. The modified algorithm can select more reasonable initial cluster center, obtain more objective ,true clustering results and shorten the clustering time . The experiment results shows the algorithm is valid.**

Keywords-clustering algorithm ; k-means; global k-means; fast global k-means

I. INTRODUCTION

K-means algorithm[1,2,3]has randomness to choose initial clustering center, resulting in variety of clustering results due to the different initial clustering center and different initial data input sequence , thus, make the results only converge to a local optimum.

To these drawbacks of the K-means algorithm, many scholars have researched and modified. The literature[4]select the farthest points in high density region as the initial cluster center,overcomes the sensitivity to the initial clustering center of K- means algorithm. The literature[5]explain how to define the sample point density,choose the larger density point In the sample ,greedy algorithm is used to search out scatter the larger point as the initial cluster center.Entitled “the fast global K-means clustering algorithm ” introduced by Liklas[6], this algorithm insert bn and select its corresponding sample when bn is maximum as optimal clustering center. However, when optimal clustering center is selected, due to no consideration of distance between the clusters, the computational amount is too large. Thus, an efficient fast global K-means algorithm is discussed, as a result of enhancing clustering mass and decreasing time complexity .

II. FAST GLOBAL K-MEANS ALGORITHM

Global K-means algorithm can get a better clustering results except that the amount of computation is too large. To this, this algorithm is modified by Likas and other scholars ,that is called fast global K-means algorithm. This algorithm, through computing bn, selects its corresponding sample when bn is maximum as optimal clustering center, further reduces computational amount.

bn is defined as :

$$b_n = \sum_{j=1}^{N} \max(d_{K-1}^{j} - ||x_n - x_j||^2, 0) \tag{1}$$

Where, d_{k-1}^{j} denotes distance between x_j to its nearest clustering center. Experiments show that compared fast global K-means algorithm with global K-means algorithm, it can decrease clustering execution time without the affects to clustering mass.

III. AN EFFECTIVE FAST GLOBAL K-MEANS

In order to enhance clustering effect of the fast global k-means and reduce time complexity, we proposed an efficient fast global K-means algorithm. The definition

$$M = \min\{\frac{\sum_{j=1}^{n} d(a_j, a_i)}{\sum_{x=1}^{n}\sum_{y=1}^{n} d(a_x, a_y)}\} \tag{2}$$

selects the most intensive sample from data set as the first initial cluster center, and ensures its correctness, and saves the calculated distance of samples in the matrix D, avoiding repeated calculation when it would be used later, and reducing the quantity of calculation. We calculate the average value of samples' distance d , find the samples' distance less than d , based on this, and calculate the average value $\bar{d}$ as the threshold value. We make a circle whose center is the existing cluster center and whose radius is $\bar{d}$. Except for the samples in circle, we select the next cluster center from the rest samples. In data set, we exclude the sample, which is the most impossible to become the next cluster center. In this way, the quantity of calculation is reduced.The definition

$$G_i = \max\{(1-\frac{1}{d_{\min}}) + \frac{\sum_{j=1}^{N}\max\{\| x_j - v_{k-1}\|^2 - \| x_j - x_i \|^2,0\}}{\sum_{s=1}^{k-1}\sum_{t=1}^{N}\| x_t - v_s \|^2}\} \tag{3}$$

selects the next cluster center, in which $d_{\min}$ is the minimum distance value between candidate cluster center and existing's. Meanwhile, we have to consider the distance between candidate cluster center and existing's, and the ratio of sum of squares error, which is reduced by candidate cluster center. In order to choose a more reasonable cluster center, corresponded with the maximum G_i , a candidate cluster center x_i would be the next cluster center.We can use triangle inequality theorem to calculate the value of $\sum_{j=1}^{N}\max\{\| x_j - v_{k-1}\|^2 - \| x_j - x_i \|^2,0\}$ in G_i . If 2AB≤BC, then AB≤AC. This way reduces the quantity of calculation. We can inquire about the value of $d(v_{k-1}, x_i)$ and $d(v_{k-1}, x_{\mathrm{j}})$ in the matrix D. If $2d(v_{k-1}, x_{\mathrm{j}}) \le d(v_{k-1}, x_{\mathrm{i}})$, then $d(v_{k-1}, x_{\mathrm{j}}) \le d(x_i, x_{\mathrm{j}})$. In this way, we don't have to inquire about the value of $d(x_i, x_{\mathrm{j}})$ in the matrix D, however, we can quickly calculate the value of $\sum_{j=1}^{N}\max\{\| x_j - v_{k-1}\|^2 - \| x_j - x_i \|^2,0\}$, avoiding wasting time and reducing time complexity. This method of adjusting fast global K-means algorithm reassigns samples to cluster. In the iterative process samples is assigned to the cluster of minimum weighted distance, further improving the cluster quality. The calculation method of weighted distance: In the iterative process samples is assigned to the cluster whose average error increasing quantity is minimum. Suppose x is assigned to the kth cluster, and average error increasing quantity of cluster K is μ_k , in which, c_k is the number of cluster K, v_k is cluster center.

$$\mu_k = \frac{\sqrt{\sum_{i=1}^{c_k}\| x_{ki} - v_k \|^2 + \| \mathrm{x} - v_k \|^2}}{c_k} - \frac{\sqrt{\sum_{i=1}^{c_k}\| x_{ki} - v_k \|^2}}{c_k}$$

$$= \frac{1}{c_k} \cdot \frac{\| x - v_k \|^2}{\sqrt{\sum_{i=1}^{c_k}\| x_{ik} - v_k \|^2 + \| x - v_k \|^2} + \sqrt{\sum_{i=1}^{c_k}\| x_{ik} - v_k \|^2}}$$

Since: $\sum_{i=1}^{c_k}\| x_{ik} - v_k \|^2 >> \| x - v_k \|^2$

Hence:

$$\sqrt{\sum_{i=1}^{c_k}\| x_{ik} - v_k \|^2 + \| x - v_k \|^2} \approx \sqrt{\sum_{i=1}^{c_k}\| x_{ik} - v_k \|^2}$$

i.e.:

$$\mu_k = \frac{1}{c_k} \cdot \frac{\| \mathrm{x} - v_k \|^2}{2\sqrt{\sum_{i=1}^{c_k}\| x_{ki} - v_k \|^2}} = \frac{1}{2} \cdot \frac{1}{c_k} \cdot \frac{1}{\sqrt{\sum_{i=1}^{c_k}\| x_{ki} - v_k \|^2}} \cdot \| \mathrm{x} - v_k \|^2$$

Suppose weight value $\delta_k = \frac{1}{c_k} \cdot \frac{1}{\sqrt{\sum_{i=1}^{c_k}\| x_{ki} - v_k \|^2}}$, then weighted distance is:

$$\nabla_k = \delta_k \cdot \| x - v_k \|^2 \tag{4}$$

In every iterative samples allocation process, samples are always assigned to the cluster whose weighted distance ∇_k is minimum.

IV. Analysis Of Experimental Results

The experiment uses UCI standard data set. In order to validate the performance of improved algorithm,we Run fast global K-means（FGKM）, modified global K-means（MGKM）,simple fast global K-means（SaFGKM）and efficient fast global K-means（EFGKM）60 times respectively. We compare average value of error sum of squares in clustering results. The comparison is shown in Table 1. For each test data set,the clustering time as shown in figure 1, figure 2, figure 3, figure 4.

TABLE 1 THE COMPARISON OF CLUSTERING ERROR IN TEST DATA SET

Test data	FGKM	MGKM	SaFGKM	EFGKM	Test data	FGKM	MGKM	SaFGKM	EFGKM
Wine	29.8	28.8	28.9	27.2	pixel averages	1143	1074	1093	1085
glass	15.0	11.2	14.3	11.1	Pendigits	60243	60098	60117	60142

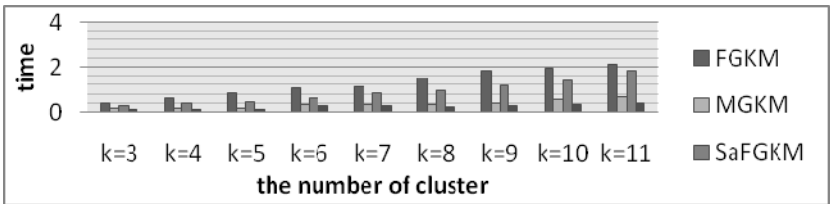

Figure 1. The comparison of clustering time in the Wine test data set

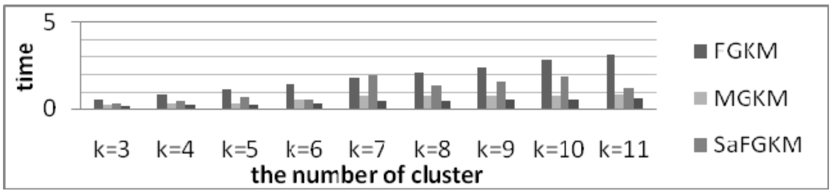

Figure 2. The comparison of clustering time in the glass test data set

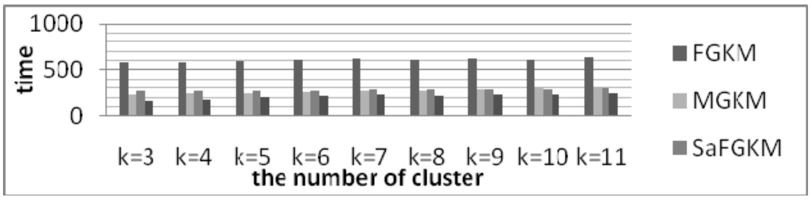

Figure 3. The comparison of clustering time in the pixel averages test data set

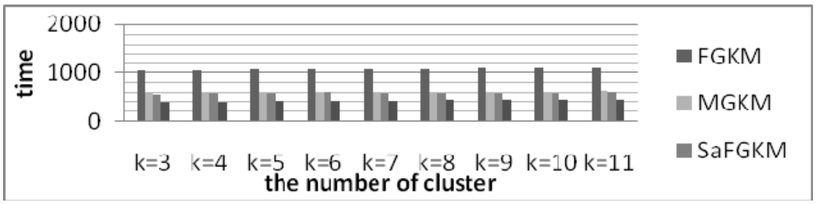

Figure 4. The comparison of clustering time in the Pendigits test data set

In test data set, the comparison of average clustering error is summarized in table 1. The value of EFGKM is less than FGKM in every test data set. The value of EFGKM is a little greater than SaFGKM in the Pendigits test data set. However, the value of EFGKM is less than SaFGKM in the rest data set. In the Wine and glass data set, the value of EFGKM is less than MGKM, and in the rest data set, the value of EFGKM is a little greater than MGKM. By comparing the results of average running time, we can see that: the running time of FGKM algorithm is the longest, and the value of EFGKM is less than MGKM's and SaFGKM's.Experiments show that: In reducing the clustering error, EFGKM algorithm is better than FGKM, and a little better than MGKM and SaFGKM; In reducing the running time, EFGKM algorithm is better than FGKM, MGKM and SaFGKM. It proves that EFGKM algorithm is effective.

V. CONCLUSIONS

This paper proposes an efficient fast global K-means algorithm， which selects the most intensive sample of data set as the first initial cluster center, and ensures its correctness. Standards of the next reasonable cluster center include: reducing error of squares and far distance away existing cluster center. We exclude the sample, which is the most impossible to become the next cluster center. In this way, the quantity of calculation is reduced by using triangle inequality theorem. This method of adjusting fast global K-means algorithm reassigns samples to cluster. In the iterative process, samples are assigned to the cluster of minimum weighted distance, improving the cluster quality and reducing the time complexity.

ACKNOWLEDGEMENTS

It is a project supported by the natural science foundation of Jilin Province(20130101179JC-13) and the natural science foundation of Jilin Province(201215145) the Twelfth Five-Science and Technology Research Projects of Education Department of Jilin Province (Grant No. 2013-420). The corresponding author is Fuheng Qu.

REFERENCES

[1] Han Jiawei,Kamber M.Data Mining:Concepts and Techniques[M].2nd ed.Beijing, China:China Machine Press,2011.

[2] Yu H,Li Z,Yao N.Research on optimization method for K-Means clustering algorithm[J].Journal of Chinese Computer Systems,2012,33(10):2273-2277.

[3] Han Zhongming,Cheng Ni,Zhang Hui.A Hierarchical Clustering Algorithm Based on Asymmetric Distance[J], Pattern recognition and artificial intelligence, 2014,27(5):410-416.

[4] Fu Desheng,Zhouchen.Improved K-means algorithm and its implementation based on density[J],Journal of Computer Application,2011,31(2):432-434.

[5] Wang Sifei,Huang Fei.Application of K-means Clustering-Based KPCA in Fault Diagnosis[J],Computer Applications and Software,2013,30(4):120-123.

[6] Likas A,Vlassis M,Verbeek J.The global K-means clustering algorothm[J].Pattern Recognition,2003,36(2):451-461.