Research on k-means Clustering Algorithm

An Improved k-means Clustering Algorithm

Shi Na
College of Information Engineering, Capital Normal University
CNU
Beijing, China
shina8237140@126.com

Liu Xumin
College of Information Engineering, Capital Normal University
CNU
Beijing, China
hellosn@126.com

Guan yong
College of Information Engineering, Capital Normal University
CNU
Beijing, China
whwqd@126.com

***Abstract*—Clustering analysis method is one of the main analytical methods in data mining,the method of clustering algorithm will influence the clustering results directly. This paper discusses the standard k-means clustering algorithm and analyzes the shortcomings of standard k-means algorithm, such as the k-means clustering algorithm has to calculate the distance between each data object and all cluster centers in each iteration, which makes the efficiency of clustering is not high. This paper proposes an improved k-means algorithm in order to solve this question, requring a simple data structure to store some information in every iteration,which is to be used in the next interation.The improved method avoids computing the distance of each data object to the cluster centers repeatly, saving the running time. Experimental results show that the improved method can effectively improve the speed of clustering and accuracy, reducing the computational complexity of the k-means.**

Keywords-clustering analysis; k-means algorithm; distance; computational complexity

I. INTRODUCTION

Clustering is a way that classify the raw data reasonably and searchs the hidden patterns that may exist in datasets [7]. It is a process of grouping data objects into disjointed clusters so that the datas in the same cluster are similar, yet datas belonging to different cluster differ.The demand for organizing the sharp increasing datas and learning valuable information from data, which makes clustering techniques are widely applied in many application areas such as artificial intelligence, biology, customer relationship management,data compression, data mining, information retrieval, image processing, machine learning, marketing, medicine, patternrecognition, psychology, statistics [3] and so on.

K-means is a numerical, unsupervised, non-deterministic, iterative method. It is simple and very fast , so in many practical applications, the method is proved to be a very effective way that can produce good clustering results. But it is very suitable for producing globular clusters. Several attemps were made by researchers to improve efficiency of the k-means algorithms [5]. In the literature [3], there is an improved k-means algorithm based on weights. This is a new partitioning clustering algorithm, which can handle the datas of numerical attribute, and it also can handle the datas of symbol attribute. Meanwhile, this method reduces the impact of isolated points and the “noise”, so it enhances the efficiency of clustering. However, this method has no improvement on the complexity of time. In the literature [1], it proposed a systematic method to find the initial cluster centers, This centers obtained by this method are consistant with the distribution of datas. Hence this method can produce more accurate clustering results than the standard k-means algorithm, but this method does not have any improvements on the executive time and the time complexity of algorithm. This paper presents an improved k-means algorithm. Although this algorithm can generate the same clustering results as that of the standard k-means algorithm, the algorithm of this paper proposed is superior to the standard k-means method on running time and accuracy, thus enhancing the speed of clustering and improving the time complexity of algorithm. By comparing the experimental results of the standard k-means and the improved k-means, it shows that the improved method can effectively shorten the running time.

This paper includes four parts: The second part details the k-means algorithm and shows the shortcomings of the standard k-means algorithm. The third part presents the improved k-means clustering algorithm, the last part of this paper describes the

978-0-7695-4020-7/10 $26.00 © 2010 IEEE
DOI 10.1109/IITSI.2010.74

experimental results and conclusions through experimenting with UCI data sets.

II. THE K-MEANS CLUSTERING ALGORITHM

A. The process of k-means algorithm

This part briefly describes the standard k-means algorithm. k-means is a typical clustering algorithm in data mining and which is widely used for clustering large set of datas. In 1967, MacQueen firstly proposed the k-means algorithm, it was one of the most simple, non-supervised learning algorithms, which was applied to solve the problem of the well-known cluster [2]. It is a partitioning clustering algorithm, this method is to classify the given date objects into k different clusters through the iterative, converging to a local minimum. So the results of generated clusters are compact and independent.

The algorithm consists of two separate phases. The first phase selects k centers randomly, where the value k is fixed in advance. The next phase is to take each data object to the nearest center[5]. Euclidean distance is generally considered to determine the distance between each data object and the cluster centers. When all the data objcets are included in some clusters, the first step is completed and an early grouping is done. Recalculating the average of the early formed clusters. This iterative process continues repeatedly until the criterion function becomes the minimum.

Supposing that the target object is x, x_i indicates the average of cluster C_i, criterion function is defined as follows:

$$E=\sum_{i=1}^{k}\sum_{x\in C_i}\left|x-x_i\right|^2$$

E is the sum of the squared error of all objects in database. The distance of criterion function is Euclidean distance, which is used for determining the nearest distance between each data object and cluster center. The Euclidean distance between one vector $x=(x_1, x_2, \ldots x_n)$ and another vector $y=(y_1, y_2, \ldots y_n)$, The Euclidean distance $d(x_i, y_i)$ can be obtained as follow:

$$d(x_i, y_i)=\left[\sum_{i=1}^{n}(x_i-y_i)^2\right]^{1/2}$$

The process of k-means algorithm as follow:

Input:

Number of desired clusters, k, and a database $D=\{d_1, d_2, \ldots d_n\}$ containing n data objects.

Output:

A set of k clusters

Steps:

1) Randomly select k data objects from dataset D as initial cluster centers.

2) Repeat;

3) Calculate the distance between each data object d_i ($1 <= i<=n$) and all k cluster centers c_j($1<=j<=k$) and assign data object d_i to the nearest cluster.

4) For each cluster j ($1<=j<=k$), recalculate the cluster center.

5) until no changing in the center of clusters.

The k-means clustering algorithm always converges to local minimum. Before the k-means algorithm converges, calculations of distance and cluster centers are done while loops are executed a number of times, where the positive integer t is known as the number of k-means iterations. The precise value of t varies depending on the initial starting cluster centers [8]. The distribution of data points has a relationship with the new clustering center, so the computational time complexity of the k-means algorithm is O(nkt). n is the number of all data objects, k is the number of clusters, t is the iterations of algorithm. Usually requiring $k <<n$ and $t <<n$.

B. The shortcomings of k-means algorithm

We can see from the above analysis of algorithms, the algorithm has to calculate the distance from each data object to every cluster center in each iteration. However, by experiments we find that it is not necessary for us to calculate that distance each time. Assuming that cluster C formed after the first j iterations, the data object x is assigned to cluster C, but in a few iterations, the data object x is still assigned to the cluster C. In this process, after several iterations, we calculate the distance from data object x to each cluster center and find that the distance to the cluster C is the smallest. So in the course of several iterations, k-means algorithm is to calculate the distance between data object x to the other cluster center, which takes up a long execution time thus affecting the efficiency of clustering.

III. IMPROVED K-MEANS CLUSTERING ALGORITHM

The standard k-means algorithm needs to calculate the distance from the each date object to all the centers of k clusters when it executes the iteration each time, which takes up a lot of execution time especially for large-capacity databases. For the shortcomings of the above k-means algorithm, this paper presents an improved k-means method. The main idea of algorithm is to set two simple data structures to retain the labels of cluster and the distance of all the date objects to the nearest cluster during the each iteration, that can be used in next iteration, we calculate the distance between the current date object and the new cluster center, if the computed distance is smaller than or equal to the distance to the old center, the data object stays in it's cluster that was assigned to in previous iteration. Therefore, there is no need to calculate the distance from this data object to the other k-1clustering centers, saving the calculative time to the k-1 cluster centers. Otherwise, we must calculate the distance from the current data object to all k cluster centers, and find the nearest cluster center and assign this point to the nearest cluster center. And then we seperately record the label of nearest cluster center

and the distance to it's center. Because in each interation some data points still remain in the original cluster, it means that some parts of the data points will not be calculated, saving a total time of calculating the distance, thereby enhancing the efficiency of the algorithm.

The process of the improved algorithm is described as follows:

Input:

The number of desired clusters k, and a database D={d_1, d_2,…d_n} containing n data objects.

Output:

A set of k clusters

Steps:

1) Randomly select k objects from dataset D as initial cluster centers.

2) Calculate the distance between each data object d_i (1 <= i<=n) and all k cluster centers c_j (1<=j<=k) as Euclidean distance d(d_i , c_j) and assign data object d_i to the nearest cluster.

3) For each data object d_i, find the closest center c_j and assign d_i to cluster center j;

4) Store the label of cluster center in which data object d_i is and the distance of data object d_i to the nearest cluster and store them in array Cluster[ ] and the Dist[ ] separately.

Set Cluster[i]=j, j is the label of nearest cluster.

Set Dist[i]=d(di, cj), d(d_i, c_j) is the nearest Euclidean distance to the closest center.

5) For each cluster j (1<=j<=k), recalculate the cluster center;

6) Repeat

7) For each data object d_i

Compute it's distance to the center of the present nearest cluster;

a) If this distance is less than or equal to Dist[i], the data object stays in the initial cluster;

b) Else

For every cluster center c_j(1<=j<=k), compute the distance d(d_i, c_j) of each data object to all the center, assign the data object d_i to the nearest center c_j.

Set Cluster[i]=j;

Set Dist[i]= d(d_i, c_j);

8) For each cluster center j(1<=j<=k), recalculate the centers;

9) Until the convergence criteria is met.

10) Output the clustering results;

The improved algorithm requires two data structure (Cluster[ ]and Dist[ ])to keep the some information in each iteration which is used in the next interation. Array cluster[ ] is used for keep the label of the closest center while data structure Dist[ ] stores the Euclidean distance of data object to the closest center. The information in data structure allows this function to reduce the number of distance calculation required to assign each data object to the nearest cluster, and this method makes the improved k-means algorithm faster than the standard k-means algorithm.

This paper propses an improved k-means algorithm, to obtain the initial cluster, time complexity of the improved k-means algorithm is O(nk). Here some data points remain in the original clusters, while the others move to another clusters. If the data point retains in the original cluster, this needs O(1), else O(k). With the convergence of clustering algorithm, the number of data points moved from their cluster will reduce. If half of the data points move from their cluster, the time complexity is O(nk/2). Hence the total time complexity is O(nk). While the standard k-means clustering algorithm require O(nkt). So the proposed k-meands algorithm in this paper can effectively improve the speed of clustering and reduce the computational complexity. But the improved k-means algorithm requires the preestimated the number of clusters, k, which is the same to the standard k-means algorithm. If you want to get to the optimal solution, you must test the different value of k.

IV. EXPERIMENTAL RESULTS

This paper selects three different data sets from the UCI [4] repository of machine learning databases to test the efficiency of the improved k-means algorithm and the standard k-means. Two simulated experiments have been carried out to demonstrate the performance of the improved k-means algorithm in this paper. This algorithm has also been applied to the clustering of real datasets. In two experiments, time taken for each experiment is computed. The same data set is given as input to the standard k-means algorithm and the improved algorithm. Experiments compare improved k-means algorithm with the standard k-means algorithm in terms of the total execution time of clusters and their accuracy. Experimental operating system is Window XP, program language is VC++ 6.0.

This paper uses iris, glass, letter [4] as the test datasets and gives a brief description of the datasets used in experiment evaluation. Table 1 shows some characteristics of the datasets.

TABLE I. CHARACTERISTIC OF THE DATASETS

Dataset	Number of attributes	Number of records
Iris	4	150
Glass	9	214
letter	16	20000

A. Experiment 1

In experiment 1, datasets of Iiris and glass are selected because they are fit to clustering analysis and their clustering results are reliable. The number of cluster k sets 3. Clustering

results for the standard k-means algorithm and the improved k-means algorithm proposed in this paper are listed in Table Ⅱ.

TABLE II. CLUSTERING RESULTS FOR IRIS, GLASS ON THE STANDARD K-MEANS AND THE IMPROVED K-MEANS

Dataset	k-means Running time (s)	Improved k-means Running time(s)	k-means Accuracy %	Improved k-means Accuracy %
Iris	0.0586	0.0559	84.3	91.6
glass	0.0814	0.0778	78.9	89.3

The results of experiment 1 show that the improved k-means algorithm can produce the final cluster results in shorter time than the standard k-means. At the same time the improved k-means can enhance the accuracy of algorithm.

B. Experiment 2

In the Experiment 2, the same dataset is used on different k values.

This experiment uses dataset letter containing 20000 samples for testing, k sets respectively 40,60,80,100. Figure 1 depicts the performance of the improved k-means algorithm and the standard k-means algorithm in terms of the total execution time of clusters.

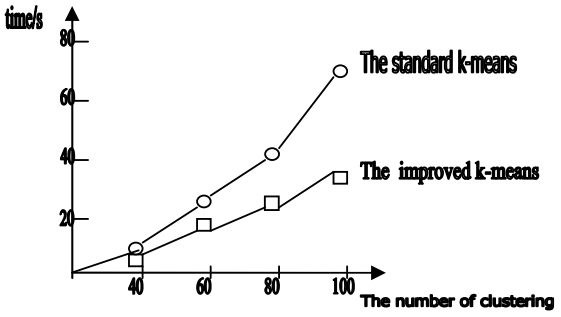

Figure 1. Excution time comparison of the improved k-means algorithm and the standard k-means algorithm

The results of Experiment 2 shows that, the improved k-means algorithm is able to fully demonstrate it's superiority on the running time comparing to the standard k-means algorithm, especially when it facs large-capacity database. The improved algorithm can generate the final clustering results in relatively short period of time, so it can enhance the speed of clustering.

Results of two simulated experiments show that the improved k-means clustering algorithm significantly outperforms the standard k-means algorithm in the overall execution time. So the improved algorithm proposed in this paper is feasible.

CONCLUSION

K-means is a typical clustering algorithm and it is widely used for clustering large sets of data. This paper elaborates k-means algorithm and analyses the shortcomings of the standard k-means clustering algorithm. Because the computational complexity of the standard k-means algorithm is objectionably high owing to the need to reassign the data points a number of times during every iteration, which maks the efficiency of standard k-means clustering is not high. This paper presents a simple and efficient way for assigning data points to clusters. The proposed method in this paper ensures the entire process of clustering in O(nk) time without sacrificing the accuracy of clusters. Experimental results shows the improved algorithm can improve the execution time of k-means algorithm. So the proposed k-means method is feasible.

ACKNOWLEDGMENT

This research was supported by the AST3 real-time data processing key technology and system (grant number: 10978016).

REFERENCES

[1] Yuan F, Meng Z. H, Zhang H. X and Dong C. R, "A New Algorithm to Get the Initial Centroids," Proc. of the 3rd International Conference on Machine Learning and Cybernetics, pp. 26–29, August 2004.

[2] Sun Jigui, Liu Jie, Zhao Lianyu, "Clustering algorithms Research",Journal of Software ,Vol 19,No 1, pp.48-61,January 2008.

[3] Sun Shibao, Qin Keyun,"Research on Modified k-means Data Cluster Algorithm"I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," Computer Engineering, vol.33, No.13, pp.200–201,July 2007.

[4] Merz C and Murphy P, UCI Repository of Machine Learning Databases, Available:

ftp://ftp.ics.uci.edu/pub/machine-learning-databases

[5] Fahim A M,Salem A M,Torkey F A, "An efficient enhanced k-means clustering algorithm" Journal of Zhejiang University Science A, Vol.10, pp:1626-1633,July 2006.

[6] Zhao YC, Song J. GDILC: A grid-based density isoline clustering algorithm. In: Zhong YX, Cui S, Yang Y, eds. Proc. of the Internet Conf. on Info-Net. Beijing: IEEE Press, 2001. 140−145. http://ieeexplore.ieee.org/iel5/7719/21161/00982709.pdf

[7] Huang Z, "Extensions to the k-means algorithm for clustering large data sets with categorical values," Data Mining and Knowledge Discovery, Vol.2, pp:283–304, 1998.

[8] K.A.Abdul Nazeer, M.P.Sebastian, "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm",Proceeding of the World Congress on Engineering, vol 1,london, July 2009.

[9] Fred ALN, Leitão JMN. Partitional vs hierarchical clustering using a minimum grammar complexity approach. In: Proc. of the SSPR & SPR 2000. LNCS 1876, 2000. 193−202. http://www.sigmod.org/dblp/db/conf/sspr/sspr2000.htm

[10] Gelbard R, Spiegler I. Hempel's raven paradox: A positive approach to cluster analysis. Computers and Operations Research, 2000,27(4):305−320.

[11] Huang Z. A fast clustering algorithm to cluster very large categorical data sets in data mining. In: Proc. of the SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery. Tucson, 1997. 146−151.

http://www.informatik.uni-trier.de/~ley/db/conf/sigmod/sigmod97.html

[12] Ding C, He X. K-Nearest-Neighbor in data clustering: Incorporating local information into global optimization. In: Proc. of the ACM Symp. on Applied Computing. Nicosia: ACM Press, 2004. 584−589. http://www.acm.org/conferences/sac/sac2004/

[13] Hinneburg A,Keim D.An efficient approach to clustering in large multimedia databases with noise.In:Agrawal R,Stolorz PE,Piatetsky-Shapiro G,eds.Proc.of the 4th Int'l Conf.on Knowledge Discovery and Data Mining(KDD'98).New York:AAAIPress,1998.58~65.

[14] Zhang T,Ramakrishnan R,Livny M.BIRCH:An efficient data clustering method for very large databases.In:Jagadish HV,Mumick IS,eds.Proc.of the 1996 ACM SIGMOD Int'l Conf.on Management of Data.Montreal:ACM Press,1996.103~114.

[15] Birant D, Kut A. ST-DBSCAN: An algorithm for clustering spatial-temporal data. Data & Knowledge Engineering, 2007,60(1): 208-221.