Research on K-means clustering algorithm and its implementation

Jianming Cui
Guilin University of Technology
School or College of Information science and engineering Guilin, China
ljm0112@163.com

Jianming Liu
Guilin University of Technology
School or College of Information science and engineering Guilin, China
ljm0112@163.com

Zhouyu Liao
Guilin University of Technology
School or College of Information science and engineering Guilin, China
ljm0112@163.com

***Abstract*—K-means algorithm is a kind of clustering analysis based on partition algorithm, it through constant iteration to clustering, when algorithm converges to an end conditions, and the output iterative process termination clustering results. Because its algorithm is simple, and easy to realize thoughts of large-scale data clustering, so k-means algorithm has become one of the most commonly used one of the clustering algorithm. K-means algorithm can find about clustering error local optimal solution, be applied in many clustering on the question of the rapid iteration algorithm. In this paper, we deeply research and analysis of the K-means clustering algorithm in the cluster analysis and analysis of its advantages and disadvantage, finally, we implement the K-means and do an experiment for application.**

Keywords; K-means; Clustering center; Data mining; algorithm.

I. INTRODUCTION

Clustering is one of three main areas of data mining (association rules, clustering, and classification). Clustering is a means of the analysis of the data and discovers the useful knowledge. It grouped the collection of data objects into a plurality of clusters of similar objects. The same objects clusters are similar to each other or similar objects to each other differ in different clusters. According to the object's properties we can calculate The Similarity of the data object. The distance is one of the often used metric, from the perspective of the machine learning, clustering belongs to unsupervised learning, different the classification, clustering and unsupervised learning does not depend on the pre-defined classes and a labeled training examples of a class. As a branch of statistics and a supervised learning method, clustering from the point of view of the mathematical analysis provides an accurate, detailed analysis tool. Therefore, the clustering analysis has a wide range of applications, such as market segmentation, pattern recognition, biological studies, spatial data analysis, and web document classification. In addition, clustering analysis can also be used as a stand-alone data mining tool to understand the data release, or a preprocessing step for other data mining algorithms (such as association rules, classification, etc.). Clustering has been extensively studied for many years; so far, researchers have proposed many clustering algorithms. In general, these algorithms can be divided into a division-based approach, a method of hierarchical, the density-based method, the grid the method and a model-based approach. K-means is a basic division in the clustering analysis and error square and guidelines function is often used as a clustering criterion. The main advantage of K-means algorithm is simple, fast, and able to efficiently handle large data sets. In many applications, the cluster analysis as a data pre-processing, is the basis for further analysis and processing of data. A high-quality clustering algorithm must meet the following two conditions: the similarity of the same cluster of the data must be the most be less, while the kinds of the different cluster distance should be as long as possible to the description of the firmness class between data separation described. The quality of clustering usually depends on the level of a method and implementation of the clustering algorithm used by the similarity measure, and also depends on whether the algorithm could find the pattern of the part or all of the hidden data.

II. K-MEANS CLUSTERING ALGORITHM

K-means algorithm with K input parameters, N objects ware distributed into K clusters, that makes a similar high similarity in the one cluster, low similarity between clusters. K-means algorithm process as follows.

First, randomly select K objects, each object represents a cluster of initial mean or center. Remaining for each object, according to the average distance with each cluster, it is assigned to the most similar cluster. And then calculate a new mean value for each cluster. This process is continuously repeated. Until criterion function is convergence. Typically, the squared error criterion, which is defined as follows:

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} \left| p - m_i \right|^2 \quad (1)$$

Whereas, E is the sum of square error of all the objects in the dataset, and P is a point in space, representing the given object is a cluster means (P and are multidimensional). As for each object, the requested object to the square of the distance from the cluster center, and then do a sum. The process of K-means clustering algorithm [1] described in Figure 1:

K-means algorithm, the mean represents the center of each cluster.

Input:

K: the number of the clusters

D: contain N object in data set

Output:

K clusters collection。

Method:

(1) Choose K objects as initial cluster centers; from D

(2) Repeat

(3) Each object is assigned to the most similar clusters based on the mean value of the object in the cluster

(4) Profile of the mean of each cluster, and calculating the mean of each cluster;

(5) Until no change

From the K-means step initial cluster centers have a larger impact on the selection of the clustering results, because K-means algorithm randomly selected K points as initial cluster centers, which representative the initial one cluster. Having prior knowledge, we can select a representative point. Figure 2 demonstrates the K-means clustering algorithm clustering process. Assume that the data distribution in Figure 2 (a) shown, let k = 3, that is the data set is divided into three class (clustering)[2].

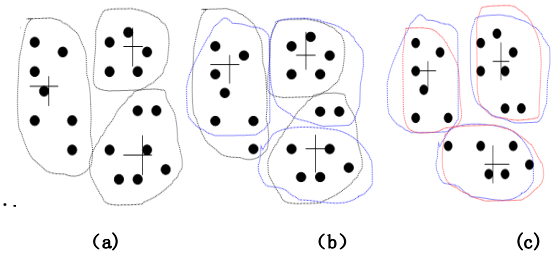

Figure 1. the process of K-means clustering algorithm clustering

Class distance is based on the distance between the points defined [3]: For instance, the nearest point between the two types of distance between the distances between these two types can be used as, The distance between the furthest point of the two categories can be used as the distance between these two; Of course, can use the distance of the longest point of the two between the centers between the classes. When calculate the distance between various points between the distance and class selection is achieved by the statistical software option. Different from the results of the selection will be different, but not too much. The degree of similarity between the objects is based on the distance between the object.

The most commonly used distance measurement method:

Let $X = (X_1, \cdots, X_p)$ and $X = (y_1, \cdots, y_p)$ is the n-dimensional space of two vectors, the distance of x and y or similar coefficients have three definitions formula:

Absolute distance formula:

$$d(x,y) = \sum_{i=1}^{p} | x_i - y_i |$$

Euclidean distance formula:

$$d(x,y) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$$

Chebyshev distance formula:

$$d(x,y) = \max_i | x_i - y_i |$$

III. EXPERIMENT RESULTS

In this paper, we use MTATLAB as the experimental platform to do the simulation experiment. The data come from the 1st graduate of Mathematical Contest in Modeling, which is about the problem for the Colleges bursaries grading in Guangxi.The experimental data is a 234 * 3 matrix the quantitative after pretreatment. Each row of the matrix which is assessed the bursaries level basis indicate a student index.

The data .But we sometime found some data in a border state when calculating and it is inevitable to make some errors. So, using the K-means clustering to assess the level of the student grants in the boundary is a good .Using the clustering algorithm is aimed at judging whether the grade boundary of students should be divided on one level or another level. Moreover, could avoid making an inaccurate result for the students' grant grade.

Let:

m : The number of students

n_1 : The number of first-class stipend

n_2 : The number of second-class stipend

n_3 : The number of third-class stipend

(l) Initial classification: According to the extent of poverty, student dataset will be divided into the three classes:1to n_1 , n_1+1 to n_1+n_2 , n_1+n_2+1 to $n_1+n_2+n_3$, $n_1+n_2+n_3+1$ to m .

(2) Modify classification: Calculate the center of the initial class, and then find the each student's distance to the initial K class by poverty. If the student dataset is closest to where he was before the classification, the student is still in the original class, otherwise he moves and his closest class, recalculate lost the center of the student's class, and the center of the students class which accept it.

(3) Repeat: Repeat steps (2).Modify the classification and check the every student's data in each class until no change, finally, we will get what we want.

In this Experiment, we get a satisfactory result through many times experiments with test. Table 1 compares the result for the number of student's stipend.

TABLE I. COMPARES DATA EXPERIMENT

	the number of one Stipend	*the number of two Stipend*	*the number of three Stipend*
Before experiment	***49***	***25***	***160***
After experiment	***57***	***80***	***97***

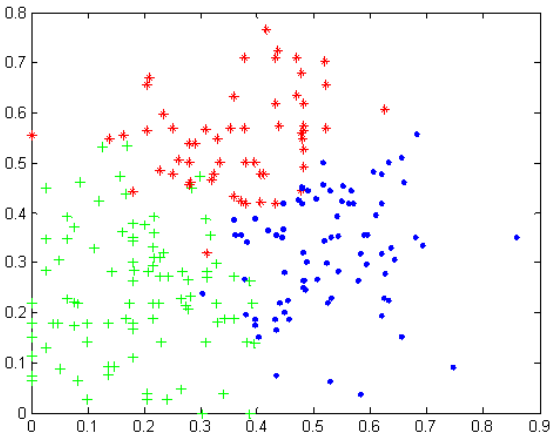

Figure 2. Clustering analysis

From Fig 2,* is the number of first-class stipend, is the number of second-class stipend. + is the number of third-class stipend, we can conclude the result is better than it before from the Fig 2.The K-means makes a great improvement the experiment. But in this paper, initial cluster centers are selected randomly, So when the initial cluster centers, Search path will be different, The objective function is easy to fall into local optimal solution [4].Thus it causes the clustering results instabilities. To solve this problem, the K-means algorithm can make some improvements [5]: such as a good data preprocessing, the initial cluster centers choose, the iterative process of clustering seed selection and so on.

IV. CONCLUSION AND FUTURE WORK

In this paper, we firstly introduces data mining clustering algorithm and describe the process of the k-means algorithm, then overview of clustering analysis algorithm and the criterion function [6]from the simulation experiment ,we can get a better results and solve a application problem of the clustering analysis. According to the advantages of the K-means algorithm, we give some improved methods. The future work is to propose new methods to improve the K-means.

REFERENCES

[1] A.K.Jain, M.N.Murty, P.J.Flynn. Data Clustering A Review[J].ACM Computing Surveys.1999, 31(3):264 — 323.

[2] Jiawei Han. Data mining concepts and technology [M] Beijing: China Machine Press, 2007

[3] Juanying Xie generals an improved global K-means clustering algorithm [J]. computer application software, 2008.3,3 (25)

[4] T.Kanungo, D.M Mount, N.S.Netanyahu, et al. An efficient K-means clustering algorithm: analysis and implementation. IEEE Transactions on Pattern Ana1ysis and Machine Intelligence,2002,24(7):881-892

[5] Shuai jiang. K-means clustering algorithm. Shaanxi Normal University, a master's degree thesis .2010

[6] Xiaorong Wu.Initial center of K-means clustering algorithm to select the relevant issues. Hunan University, a master's degree thesis 2008