ISSN No. 0976-5697

Volume 2, No. 5, Sept-Oct 2011

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Huddle based Harmonic K means Clustering Algorithm

S. Adaekalavan*
Asst. Professor,
Department of Information Technology,
J.J. College of Arts and Science,
Pudukkottai, Tamilnadu, India
kingsmakers@gmail.com

Dr. C. Chandrasekar
Reader,
Department of Computer Science,
Periyar University,
Salem, Tamilnadu, India

Abstract: Clustering method is one of the important methods in data mining. This method will weight the clustering result directly. This paper discusses, the traditional k-means clustering algorithm, analyzing its shortcomings clustering algorithm and measuring the harmonic based distance between each data object and cluster centers. This efficient method avoids the need to compute the distance of each data object to the cluster center. It saves the running time. The experimental results show that this efficient method can effectively improve the speed of clustering and accuracy, reducing the computational time.

Keywords: Data Mining; Clustering analysis; k means algorithm

I. INTRODUCTION

With the advances Information Technology and Computer Science, the computer is being applied today all walks of life. The efficiency and effectiveness of computer software has increased with time. People's capacity to produce and collect data and the size of the database has expanded a lot. Business enterprises, public departments and research institutions make use of clustering techniques in many application areas such as artificial intelligence, earthquake [9], biology, customer relationship management, data compression , data mining, information retrieval, image processing , machine learning, marketing, medicine, pattern recognition, psychology, statistics and so on. In the past several years massive amounts of data have accumulated stored in different forms, because these data are very difficult to store. So, to get valuable information or knowledge from them and to achieve this purpose to enable decision-making has become a very difficult task.[1]. As a result data mining techniques have emerged, and thus gradually become a hot spot, attracting a lot of researchers. Clustering analysis is a very important data mining technology, It is one of the very important techniques in data mining.

This paper includes four parts The Second part shows the shortcomings of the standard k means algorithm. The third part presents the efficient huddle based Harmonic K Means Clustering algorithm and the last part describes the experimental results and conclusions through experiments with earthquake[9] data sets.

II. THE PROCESS OF CLUSTERING ANALYSIS IN DATA MINING

This part briefly describes the standard k-means algorithm. K means is a typical clustering algorithm in data mining which is widely used for clustering large sets of data. In 1967, Mac Queen proposed the k-means algorithm. It was one of the most simple, non-supervised learning algorithms, applied to solve the problem of the well-known cluster [2]. It is a partitioning clustering algorithm. This method is used to classify the given data objects into k different clusters through iterative converging to a local minimum. So the resultants 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[3]. Euclidean distance is generally considered to determine the distance between each data object and the cluster centers. When all the data objects 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, xi indicates the average of cluster Ci, criterion function is defined as follows:

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

E is the sum of the squared error of all objects in the database. The distance of the criterion function is Euclidean distance, which is used for determining the nearest distances 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:

$$D(x_i\, y_i) = \left(\sum_{i=1}^{n} (x_i - y_i)^2 \right)^{\frac{1}{2}}$$

The steps of k means algorithm as follows
Input : Number of desired clusters, k , and a database D= $\{d_1, d_2, \ldots\ldots d_n\}$ containing n data objects
Output : A set of K Clusters
Steps

a. Randomly select k data objects from dataset D as initial cluster centers.
b. Repeat;
c. 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.
d. For each cluster j (1<=j<=k), recalculate the cluster center
e. until there no change in the center of the clusters.

The k-means clustering algorithm always converges to a 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 [4]. The distribution of data points has a relationship with the new clustering center, so the computational time, complexity of the kmeans 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.

III. PROPOSED EFFICIENT HUDDLE BASED ARMONIC K MEANS CLUSTERING ALGORITHM (HKM).

The traditional K means algorithms need to calculate the distance from the each data object to all the centers of k clusters when it executes the iteration each time. This takes up a lot of execution time especially for large capacity of datasets. To overcome this shortcoming of the k means algorithm, this paper presents a new harmonic based distance (HKD). K-Means algorithm needs to give the clustering number to be built, which will first create an initial division, and then use an iterative relocation technique to try to improve the division through an object moving along it. The K-harmonic means does not rely on the initial point to calculate clustering time and clustering results. So we can add some improved conditions and factors based on this algorithm, to further enhance the effectiveness of this algorithm. This paper will apply a new distance measurement to the K-Harmonic means, and test the effectiveness of this combination through numerical experiments. The new measurement function is as follows:

$$d(x_i\, y_i) = \sum_{i=1}^{n} \left(1 - \exp\left(-\beta \,\|\, x_i - y_i \,\|^2\right)\right)$$

In which β is a positive constant, from this distance function it can be seen that d(x, y) is a monotonically increasing function || x - y || , namely d (x, y)increases with the increase of || x - y ||.

The steps of huddle based harmonic K means algorithm (HKM) is described as follows

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

Output : A set of K Clusters

A. Randomly select k objects from dataset D as initial cluster centers.
B. Repeat
C. Calculate the distance between each data object di (1 <= i<=n) and all k cluster centers c_j (1<=j<=k) as harmonic based distance measure function $d(d_i , c_j)$ and assign data object di to the nearest cluster. The harmonic based distance function is as follows

$$d(x_i\, y_i) = \sum_{i=1}^{n} \left(1 - \exp\left(-\beta \,\|\, x_i - y_i \,\|^2\right)\right)$$

In which β is a positive constant. From this distance function it can be seen that $d(x_i \quad y_i)$ is a monotonically increasing function $\| x_i - y_i \|$.

D. For each data object d_i, find the closest center c_j and assign d_i to cluster center j;
E. Repeat
F. Until the convergence criteria is met.
G. Output the clustering results;

As we all know, to make a point weight and more robust, it should meet the weight of the abnormal points and noise points should be smaller, and the weight of the compact point with data concentration should be greater. This new measurement precisely meets this requirement.

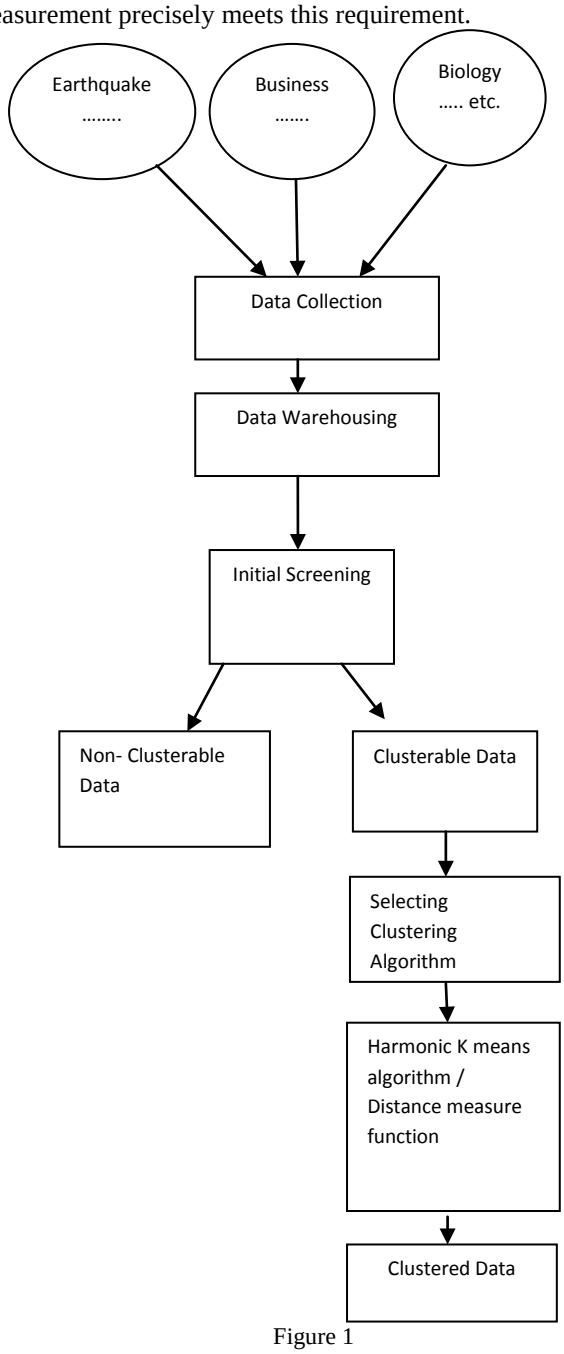

Figure 1

Flow of the proposed huddle based harmonic K means algorithm

IV. NUMERICAL EXPERIMENTAL RESULTS AND ANALYSIS

In this paper, numeric data is used for numerical experiments with different β selections. Table I is gives clustering results.

Table I: Clustering results

β Value	Number of Records	Running Time for HKM algorithm (Sec)	Running Time for K Means algorithm (Sec)
30	5000	5.942	5.991
30	10000	22.81	22.945
50	15000	49.354	50.038
50	20000	88.479	94.373
70	25000	158.615	163.711

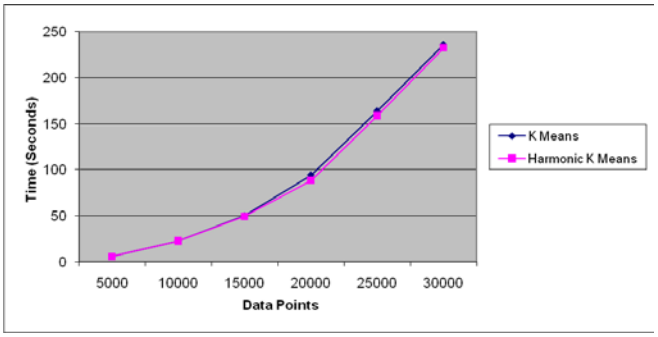

Figure 2

As can be seen from the Table, when the β Value is not the same, HKM running time is low compared to traditional K means algorithm. The efficient harmonic K means clustering algorithm can generate the final clustering results in a relatively short period of time and so it can enhance the speed of clustering.

V. CONCLUSION

The traditional K means is a typical clustering algorithm and it is extensively used for clustering large sets of data. This paper details k means algorithm and a new harmonic K means measurement algorithm. Through the regulation of distance metric parameters we can achieve better clustering effects than the traditional k means algorithm, and the new algorithm has an advantage in execution time.

VI. REFERENCES

[1]. 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

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

[3]. 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.

[4]. 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

[5]. 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.

[6]. 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.

[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]. Data Collected from : http://infochimps.com/datasets/ disasters-wordwide-from-1900-2008