Dynamic Clustering Case Study using K-Mean Clustering Algorithm

Charanjit Singh,
Research Scholar, Guru Kashi University, Talwandi Sabo
Punjab, India

Vijay Laxmi
Professor and Dean, Guru Kashi University, Talwandi Sabo
Punjab, India

Arvinder Singh Kang
Professor and Dean, Chandigarh University, Gharaun
Punjab, India

Abstract: **A Partitional Clustering Algorithm called K-mean Algorithm is widely used in the research. A large number research considering on the features of k-mean clustering have resultant that it should be search that whether the number of cluster can be originated during the execution phase, on the same time having the cluster features itself. This paper presents the part of researcher's research which states various consideration during the dynamic clustering with the k-mean clustering itself, improving the clustering of data. K-mean algorithm initializes the number of clusters to be taken as input by the user, but in practical it is difficult to fix the number of clusters at initial stage. The study states both the cases i.e. for known number of clusters at initial state and unknown number of clusters. This study paper shows that how the dynamic clustering works using k-mean algorithm further it will be used by the researcher in his research work.**

Keywords- K-means clustering; cluster quality; dynamic clustering

I Introduction

Clustering is the major problem arises frequently in the field of knowledge discovery, data mining and pattern classification [1]. The importance of data mining is increasing exponentially since last decade and in recent time where there is very tough competition in the market where the quality of information and information on time play a very crucial role in decision making of policy has attracted a great deal of attention in the information industry and in society as a whole.

To provide the information, needed with in the time and in required pattern from the huge database is very difficult. There is very huge amount of data available in real world which has to find out from the database. So tool called data mining is use for extracting information from these defined databases and represent in the required format. It is a very useful and helpful trend or application to know customer feedback, create new application on the bases of feedback, fraud identification, science exploration, production and casting unit etc. Conclude that in one sentence data mining is mining of knowledge from the huge amount of data.

Using data mining we can predict the nature or behaviour of any pattern. Cluster analysis of data is an important task in knowledge discovery and data mining.

II Clustering

Clustering or cluster analysis [2] is the process of grouping a set of objects that are meaningful, useful or both. However, the groups are not predefined. Clustering can be used in many application domains like marketing, medicine, bioinformatics, economics and anthropology. Clustering can be sometimes referred to as unsupervised learning. An unsupervised learning finds some kind of structure in the data. A clustering is a set of clusters which contains all objects in the data set.

To group the data on the basis of similarities and dissimilarities is the basic aim of cluster analysis. The data can be divided in a supervised, semi-supervised and unsupervised manner, different algorithm perform differently on the basis of input given to the cluster or the nature of data [3].

Most of the algorithms take the number of clusters (K) as an input and it is fixed. In the real-world application, it is very difficult predict the number of clusters for the unknown domain data set. If the fixed number of cluster is very small, then there is a chance of putting dissimilar objects into same group and suppose the number of fixed cluster is large then the more similar objects will be put into different groups.

In this study paper, we say that a Dynamic clustering of data can be happened and could been resulted out with the help of k-mean algorithm. Number of Cluster (k) takes by the algorithm as input from the user end and the user has to mention the number of cluster is fixed or not on the time of input. If the number of clusters are fixed on the time of input, then it will work same as the k-mean algorithm works.

Proposed study can conclude that if the clusters is not fixed then the study says that clustering can be done dynamically at run time by giving the least number of possible clusters. The k-means algorithm procedure will be repeats itself and incrementing the cluster by one until the cluster reaches validity threshold. Validity could have various aspects i.e quality, quantity etc.

III History of the K-means Algorithm

The term "k-means" was first used by James MacQueen in 1967 [4], though the idea goes back to 1957 [5]. The standard algorithm was first proposed by Stuart Lloyd in 1957 as a technique for pulse-code modulation, though it wasn't published until 1982. K-means is a widely used partitional clustering method in the industries. Most commonly used partitional clustering algorithm is called k-mean algorithm having the popularity among all the algorithms because it is an algorithm which can be easily implemented and on the execution time it is the very efficient.

The major problem with this algorithm is that it is sensitive to the selection of the initial partition and may converge to local optima [6 7 8]. The partitioning method constructs k partitions of the data, where each partition represents a cluster and k ≤ n (data objects) [7]. It classifies the data into k groups, which together satisfy the following requirements: i) each group must contain at least one object, and ii) each object must belong to exactly one group. The researchers have investigated K-means clustering from various perspectives. Many data factors which may strongly affect the performance of K-means, have been identified in the literature [8 9 10 11 12].

IV K-means Clustering

K-means (KM) clustering is a heuristic algorithm that can minimize sum of squares of the distance from all samples emerging in clustering domain to clustering centers to seek for the minimum k clustering on the basis of objective function [13]. First and foremost, the k as input is accepted, and then data objects which are belonging to clustering domain (including n data objects, n>k) are divided into k types. As a result, the similarity between same cluster samples of is higher, but lower between hetero-cluster samples.

K data objects, as original clustering centers, are randomly selected from clustering domain by KM algorithm. K-means clustering is a data mining/machine learning algorithm used to cluster observations into groups of related observations without any prior knowledge of those relationships. The k-means algorithm is one of the simplest clustering techniques and it is commonly used in medical imaging, biometrics and related fields.

The k-means algorithm is an evolutionary algorithm that gains its name from its method of operation. The algorithm clusters observations into k groups, where k is provided as an input parameter. It then assigns each observation to clusters based upon the observation's proximity to the mean of the cluster [Eq. 1]. The cluster's mean is then recomputed and the process begins again. Here's how the algorithm works [7]:

A. Algorithm for K-Means

The algorithm for partitioning, where each cluster's center is represented by mean value of objects in the cluster.

Input: k: the number of clusters. D: a data set containing n objects.
Output: A set of k clusters.

Method:
1. Arbitrarily choose k objects from D as the initial cluster centers.
2. Repeat.
3. (re)assign each object to the cluster to which the object is most similar using Eq. 1, based on the mean value of the objects in the cluster.
4. Update the cluster means, i.e. calculate the mean value of the objects for each cluster.
5. until no change.

$$j = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \qquad \text{Eq. 1}$$

where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance(intra) measure between a data point $x_i^{(j)}$ and the cluster centre c_j, is an indicator of the distance of the n data points from their respective cluster centers. The term intra is used to measure the compactness of the clusters. The inter term is the minimum distance between the cluster centroids which is defined as

$$\text{Inter=min}\{\, m_k - m_{kk} \,\} \qquad \text{Eq. 2}$$

Whereas k =1, 2……k-1 and kk=k+1………., k
This term is used to measure the separation of the clusters. The standard deviation is used to check the closeness of the data points in each cluster and computed as:

$$\sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - X_m)^2} \qquad \text{Eq. 3}$$

One of the main disadvantages of k-means is the fact that you must specify the number of clusters as an input to the algorithm. As designed, the algorithm is not capable of determining the appropriate number of clusters and depends upon the user to identify this in advance.

V Dynamic Clustering Done

B.M Ahamed Shafeeq and K S Hareesha [14] created an application which used to create the clusters dynamically. They described that the user has the option to select the number of cluster fix at the input or by the input the minimum cluster required. Former case works same as the k-means algorithm by input the number of cluster to formed. They also developed the case where the algorithm computes the new cluster by the increment the cluster by one until the it satisfies the validity of cluster quality threshold.

They also defined algorithm is as follows:

Input: *k: number of clusters (for dynamic clustering initialize k=2) Fixed number of clusters = yes or no (Boolean). D: a data set containing n objects.*
Output: *A set of k clusters.*

Method:

```
1. Arbitrarily choose k objects from D as the initial cluster
   centers.
2. Repeat.
3. (re)assign each object to the cluster to which the object is
   most similar, based on the mean value of the objects in the
   cluster.
4. Update the cluster means, i.e. calculate the mean value of
   the objects for each cluster.
5. until no change.
6. If fixed_no_of_clusters =yes goto 12.
7. Compute inter-cluster distance using Eq.2
8. Compute intra-cluster distance using Eq. 3.
9. If new intra-cluster distance < old_intra_cluster distance
   and new_intercluster >old_inter_cluster distance goto 10
   else goto 11.
10. k= k + 1 goto step 1.
11. STOP
```

The algorithm was developed and tested for efficiency of different data points in C language. The algorithm took more computational time compared to the K-means algorithm for large dataset in some cases. The algorithm worked same as K-means for the fixed number of clusters. For the unknown data set it starts with the minimum number of cluster given by the user and after the completion of every set of iteration, the algorithm checks for efficiency and it repeats by incrementing the number of cluster by 1 until it reaches the termination condition.

VI Conclude the Study and Future Scope

As per earlier studies say that the K-means algorithm creates the clusters as well as have the information that how many clusters have to be taken out at the initial level. But at the run time, find out the number of cluster of unknown dataset in practical scenario is very essential. Poor quality of clusters might be find out by the fixing the number of clusters at the initial state. Proposed study may use to find good quality clusters at the run time. This method may work on both the cases i.e for fixed number of clusters at initial stage and create clusters dynamically at the run time stage.

We can conclude the study and says that improved data clustering can be done for unknown data set. To retain the simplicity and improved modification can be done on K-mean algorithm. K-mean algorithm takes K number of clusters as input from the user but this can be assumed as the problem to fix the number of clusters in advance. In practical it is so hard to fix the cluster in advance. There are major chances to putting the dissimilar objects into the same group if the fixed number of cluster is very small on the other hand fixed cluster is large, then the more similar objects will be put into different group/cluster. The already described algorithm might be overcome this problem by find the optimal number of cluster at run time.

The researcher would like to use the proposed case study developed by B.M Ahamed Shafeeq and K S Hareesha [14] in his research work where researcher wants to create the application of search engine followed the concept of dynamic clustering. Using the concept of dynamic clustering, search engine will able to create the cluster dynamically when the user's keyword/phrase is not founded in the database.

The main limitation of described algorithm is that it took more computational time than the k-mean for large data set. Future work can be focused on how to reduce the time without compromising the quality of cluster.

Acknowledgement

I would like to express my deep thankfulness and respect to B.M Ahamed Shafeeq and K S Hareesha[14] whose case study will help in my research. Further I also thanks to Dr. Vijay laxmi whose advices and insight was valuable to me. For all I learned from her. I would also like to thank her for being an open person to ideas, and for encouraging and helping me to shape my interest and ideas and for his continuous help and support in all stages of this research work.

REFERENCES

[1] Wei Li, "Modified K-means clustering algorithm", IEEE computer society Congress on Image and Signal Processing, 2008, pp. 618-621.

[2] Patil, K.M. and Bakal, J.W "A Survey of Clustering Algorithms", International Journal of Computer Engineering and Applications, ISSN 2321-3469 Volume IX, Issue IV, April 15

[3] Ran Vijay Singh and M.P.S Bhatia, "Data Clustering with Modified K means Algorithm", *IEEE InternationalConference on Recent Trends in Information Technology*, ICRTIT 2011, pp 717-721.

[4] MacQueen, J. B. "Some Methods for classification and Analysis of Multivariate Observations", *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press. 1967, pp. 281–297.

[5]. Lloyd, S. P. "Least square quantization in PCM". *IEEE Transactions on Information Theory* 28, 1982, pp. 129–137.

[6] Ye Yingchun, Zhang Laibin, Liang Wei, Yu Dongliang , and Wang Zhaohui, "Oil Pipeline Work Conditions Clustering Based on Simulated Annealing K-Means algorithm", World Congress on Computer Science and Information Engineering,2009, pp. 646-650.

[7] Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, second Edition, (2006).

[8] Khan, S.S., Ahmad, A., "Cluster center initialization algorithm for kmeans clustering", Pattern Recognition Letter. 25, 2004, pp. 1293–1302.

[9] Grigorios F. Tztzis and Aristidis C. Likas, "The Global Kernel k-Means Algorithm for Clustering in Feature Space", IEEE Trans. On Neural Networks, Vol. 20, No. 7, July 2009, pp. 1181-1194.

[10] R. Xu and D. Wunsch, II, "Survey of clustering algorithms", IEEE Trans. Neural Networks., vol. 16, no. 3, 2005, pp. 645– 678.

[11] Shi Na., Liu Xumin, Guan Yon , "Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm", Third International Symposium on Intelligent Information Technology and Security Informatics(IITSI), pp.63-67, 2-4 April 2010.

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

[13] S. Prakash kumar and K. S. Ramaswami, "Efficient Cluster Validation with K-Family Clusters on Quality Assessment", European Journal of Scientific Research, 2011, pp.25-36.

[14] Ahamed Shafeeq B M and Hareesha K S "Dynamic Clustering of Data with Modified K-Means Algorithm" *2012 International Conference on Information and Computer Networks (ICICN 2012) vol. 27 (2012), IACSIT Press, Singapore,* pp.221-225.