2012 International Conference on Information and Computer Networks (ICICN 2012)
IPCSIT vol. 27 (2012) © (2012) IACSIT Press, Singapore

Dynamic Clustering of Data with Modified K-Means Algorithm

Ahamed Shafeeq B M [1] and Hareesha K S [2]

[1] Department of Computer Science & Engineering, Manipal Institute Of Technology, Manipal University, Manipal -576104, India email : ahamed.shafeeq@manipal.edu

[2] Department of Computer Science & Engineering, Manipal Institute Of Technology, Manipal University, Manipal -576104, India email : hareesha.ks@manipal.edu

Abstract. K-means is a widely used partitional clustering method. While there are considerable research efforts to characterize the key features of K-means clustering, further investigation is needed to reveal whether the optimal number of clusters can be found on the run based on the cluster quality measure. This paper presents a modified K-means algorithm with the intension of improving cluster quality and to fix the optimal number of cluster. The K-means algorithm takes number of clusters (K) as input from the user. But in the practical scenario, it is very difficult to fix the number of clusters in advance. The proposed method works for both the cases i.e. for known number of clusters in advance as well as unknown number of clusters. The user has the flexibility either to fix the number of clusters or input the minimum number of clusters required. In the former case it works same as K-means algorithm. In the latter case the algorithm computes the new cluster centers by incrementing the cluster counter by one in each iteration until it satisfies the validity of cluster quality. It is shown that how the modified k-mean algorithm will increase the quality of clusters compared to the K-means algorithm. It assigns the data point to their appropriate class or cluster more effectively.

Keywords: K-means clustering, cluster quality, dynamic clustering.

1. Introduction

A fundamental problem that frequently arises in a great variety of fields such as data mining and knowledge discovery, and pattern classification is the clustering problem [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. There is very large amount of data availability in real world and it is very difficult to excess the useful information from this huge database and provide the information to which it is needed within time limit and in required pattern. So data mining is the tool for extracting the information from huge database and present it in the form in which it is needed for each specific task. The use of data mining is very vast. It is very helpful in application like to know the trend of market, fraud detection, and shopping pattern of customers, production control and science exploration etc. in one sentence data mining is mining of knowledge from huge amount of data. Using data mining we can predict the nature or behavior of any pattern.

Cluster analysis of data is an important task in knowledge discovery and data mining. Cluster analysis aims to group data on the basis of similarities and dissimilarities among the data elements. The process can be performed in a supervised, semi-supervised or unsupervised manner [2]. Different algorithms have been proposed which take into account the nature of the data and the input parameters in order to cluster the data. 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 paper we propose a dynamic clustering of data with modified k-means algorithm. The algorithm takes number of clusters (K) as the input from the user and the user has to mention whether the number of clusters is fixed or not. If the number of clusters fixed then it works same as K-means algorithm. Suppose the number of clusters is not fixed then the user has to give least possible number of clusters as an input. The K-means procedure repeated by incrementing the number of clusters by one in each iteration until it reaches the cluster quality validity threshold.

2. Background of the K-means Algorithm

The term "*k*-means" was first used by James MacQueen in 1967 [3], though the idea goes back to 1957 [4]. 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. The K-means algorithm is the most commonly used partitional clustering algorithm because it can be easily implemented and is the most efficient one in terms of the execution time. The major problem with this algorithm is that it is sensitive to the selection of the initial partition and may converge to local optima [5 6 7]. The partitioning method constructs k partitions of the data, where each partition represents a cluster and $k \leq n$ (data objects) [6]. 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 [7 8 9 10 11].

3. 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 [12]. 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 [6]:

K-Means Algorithm: 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 $\| x_i^j - c_j \|^2$ 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

Inter =min { $\| m_k - m_{kk} \|$ } $\forall$ k= 1 , 2 , ……K-1 and kk = k+1, ……,K **Eq. 2**

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:

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

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.

4. Proposed Method

The K-means algorithm finds the predefined number of clusters. In the practical scenario, it is very much essential to find the number of clusters for unknown dataset on the runtime. The fixing of number of clusters may lead to poor quality clustering. The proposed method finds the number of clusters on the run based on the cluster quality output. This method works for both the cases i.e. for known number of clusters in advance as well as unknown number of clusters. The user has the flexibility either to fix the number of clusters or by input the minimum number of clusters required. In the former case it works same as K-means algorithm. In the latter case the algorithm computes the new clusters by incrementing the cluster counter by one in each iteration until it satisfies the validity of cluster quality threshold. The modified 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_inter-cluster >old_inter_cluster distance goto 10 else goto 11.
10. k= k + 1 goto step 1.

11. STOP

Dynamic clustering of data with modified K-means Algorithm

5. Experimental Results and Analysis

In this section, in order to verify the effectiveness of the proposed algorithm, we take some random numbers of 300,500 and 1000 data points. The experimental results show that the proposed method outperforms K-means algorithm in quality and optimality for the unknown data set. The experiment is conducted on synthetic data set. The new algorithm works for fixed number of clusters as well as unknown number of clusters.

Data points	No. of clusters (K-means)	No. of clusters (Dynamic K-means)	Inter_cluster distance (K-means)	Inter_cluster Distance (Dynamic K-means)	Intra_cluster Distance(K-Means)	Intra_cluster Distance (Dynamic K-Means)
300 (Fig:1)	3	5	0.06961	0.07453	0.02814	0.01686
500 (Fig:2)	5	6	0.04837	0.08077	0.01376	0.01136
1000(Fig:3)	12	8	0.02068	0.02376	0.00718	0.00680

Table 1: Experimental Results

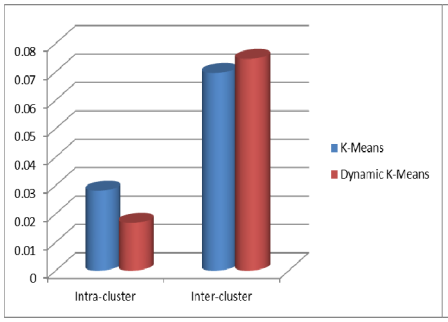

Figure : 1

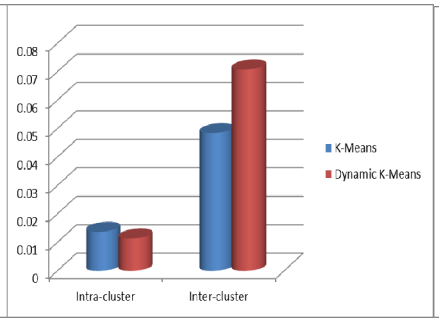

Figure : 2

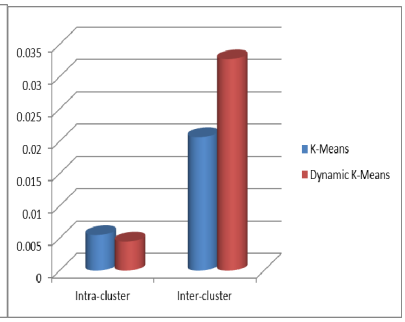

Figure : 3

The above experimental results on synthetic data show that the proposed method gives optimal number of clusters for the unknown data set. It is also observed that the time taken in the proposed method is almost same as K-Means algorithm for smaller data set. The algorithm is developed and tested for efficiency of different data points in C language. The algorithm takes more computational time compared to the K-means algorithm for large dataset in some cases. The algorithm works 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.

5. Conclusion and Further Research Direction

This paper proposed an improved data clustering for the unknown data set. The algorithm works well for the unknown data set with better results than K-means clustering. The k-means algorithm is well known for its simplicity and the modification is done in the proposed method with retention of simplicity. The K-means algorithm takes number of clusters (K) as input from the user. The major problem in K-means algorithm is fixing the number of clusters in advance. In the practical scenario, it is very difficult to fix the number of cluster in advance. 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. The proposed algorithm will overcome this problem by finding the optimal number of clusters on the run. The main drawback of the proposed approach is that it takes more computational time than the K-means for larger data sets. Future work can focus on how to reduce the time

complexity without compromising cluster quality and optimality. More experiments will be conducted with natural datasets with different features.

6. References

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

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

[3] 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.

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

[5] 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.

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

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

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.