ISSN 2278-7763

Factors Affecting Efficiency of K-means Algorithm

Sonal Miglani[1], Kanwal Garg[2]
[1]Research Scholar, M.Tech (CSE), Dept. of Computer Science & Applications, Kurukshetra University, Kurukshetra, India.
snlmiglani@gmail.com.
[2]Assistant Professor, Dept. of Computer Science & Applications, Kurukshetra University, Kurukshetra, India.
gargkanwal@gmail.com.

ABSTRACT

K-means algorithm is a simple technique that partitions a dataset into groups of sensible patterns. It is well known for clustering large datasets and generating effective results that are used in a variety of scientific applications such as Data Mining, knowledge discovery, data compression, vector quantization and medical imaging. The performance of this algorithm can be improved further by studying all those factors that plays a crucial role in its functionality. The aim of this research paper is to uncover the significant factors in order to enhance the efficiency as well as reducing the complexity of K-means algorithm.

Keywords: Clustering, Data Mining, Initial Centroids, K-means.

1. INTRODUCTION

In the process of data mining, meaningful patterns are discovered from large datasets with an intention to support efficient decision making. Clustering is an important step in all data mining algorithms in which the data objects are classified into number of different subclasses. The purpose is to generate clusters such that it contains high intra-class similarity and low inter-class similarity. The algorithm classifies raw data into K number of group on the basis of defined attributes. This grouping is performed by minimizing the sum of squares of distances between data and the corresponding cluster centroid. Conceptually, the center point is considered as the centroid of a cluster. According to Jiawaei [1], the quality of cluster can be measured by within cluster variation, which is sum of squared error between all objects in Ci and the centroid c_i, defined as

$$E= \sum \sum \text{dist}\,(p,c_{i)}^{2.} \qquad (1)$$

Where E is the sum of the squared error for all objects in the data set; p is the point in space representing a given object; and c_i is the centroid of cluster C_i. In other words, for each object in each cluster, the distance from the object to its cluster center is squared, and the distances are summed. This objective function tries to make the resulting k clusters as compact and as separate as possible.

1.1 Algorithm [1]:

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

Input:
- k: the number of clusters,
- D: a dataset 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 the most similar, based on the mean value of the objects in the cluster;
4. Update the cluster means, that is, calculate the mean value of the objects for each cluster;
5. **Until** no change.

One of the important properties of this algorithm is its linear time complexity given by O(nkt) where n is the number of data objects, k is the number of clusters and t is the number of iterations performed to complete the process which precisely depends on the initial starting cluster centers.
To attain the objective; this paper includes five sections: the first section introduces the process of K-means clustering, the second section presents the literature review in this field, the third section elaborates the factors affecting the efficiency of this algorithm, fourth section presents a case study of an improved k-means algorithm, and finally the last section concludes the work presented in this research work.

2. LITERATURE REVIEW

Koheriet. al. [3], M. R. Khammar & M. H. Marhaban[4]in their papers proposed and improved K-means algorithm using different strategies ,that intends to remove the loophole of random selection of initial centroids. MadhuYelda et al. [5] proposed an enhanced K-means algorithm with the reduced time complexity O(nlogn).P.S Bradley & Usama M Fayyad [6] presented a fast and efficient algorithm for refining an initial starting point for a general class of clustering algorithms. S. Sujatha & A. Shanthi Sona [7] presented a novel initialization technique proving that the clustering accuracy of the proposed initialization technique using Spectral Constraint Prototype is very high as compared to the Standard K-Means, DPDA K-Means and K-Means using CSC. Zhexue[8] presented two algorithms which extend the k-means algorithm to categorical domains and domains with mixed numeric & categorical values

ISSN 2278-7763

using a simple matching dissimilarity measure. An efficient k-means algorithm is presented by Elkan [10] that is intended to remove a large number of distance calculations between data objects and cluster centers. Hamerly [9] proposed an algorithm which is a modified and simplified version of Elkan's k-means algorithm. Dheebet. al. [11] designed, implemented and evaluated an image-processing based software solution for automatic detection and classification of plant leaf disease using K-means Clustering Algorithm. Neha et.al [14] proposed a mid-point based K-means clustering algorithm with improved accuracy.

3. FACTORS THAT INFLUENCES THE ALGORITHM'S EFFICIENCY

After reviewing the literature available from year 1998 to 2012, the author found that there are basically two important factors that influence the efficiency of K-means algorithm. The following section gives the detail description of how these factors play a significant role in determining the efficiency of K-Means Algorithm.

3.1 Selection of Initial Centroids

In K-means algorithm, clusters are formed with the help of centroids. First of all, K data elements are chosen as initial centers and then the distances of all data elements are calculated using Euclidean distance formula. Data elements that have less distance to centroids are moved to the appropriate cluster. By taking the mean of the data points of each cluster, the centroid of each cluster is updated and this process is continued until no more changes occur in clusters [12]. The time complexity of the K-Means algorithm and the quality of the final clustering results highly depends on the random selection of the initial centroids. In the original K-Means algorithm, the initial centroids are chosen randomly and hence different clusters are obtained for different runs for the same input data. Moreover, the computation will run the chance of converging to a local minimum rather than the global minimum solution if the initial centers are not chosen carefully. Hence, choosing the proper initial centroids is the key step of the basic K-Means procedure [7]. One solution is to run the algorithm several times with different initializations. If the results converge to the same partition then it can be assumed that a global minimum has been reached. But this strategy has a drawback of being computationally expensive as well as very time consuming. It is observed that there is a great requirement of establishing an appropriate strategy to select initial centroids. If this selection is on the basis of some heuristics instead of random selection, there is a high probability of improvement in functionality of this algorithm.

3.2 Distance Calculation

K-Means algorithm is implemented in two different phases. In first phase k centers are selected randomly, and the second phase consists of finding the nearest center for each data object which is done by calculating Euclidean distance. The first step is completed when all the data objects are included in some clusters. The average value of each cluster is then recalculated which is now consider as the new centroid of corresponding clusters. This iterative process continues repeatedly until the objective function is minimized.

It is observed that in each iteration, calculations have to be performed to find the distance from each data object to every cluster center before the algorithm converges to global minima. Suppose a cluster 'X' is formed after the first n iterations, the data object 'a' is assigned to cluster 'X', but in next few iterations, the data object 'a' is still assigned to the same cluster 'X'. After several iterations, when we calculate the distance from data object 'a' to each cluster center, it is found that the distance to the cluster 'X' is the smallest. Thus it shows that k means algorithm unnecessarily calculates the distance between data object 'a' to the other cluster centers, consuming a long execution time and therefore affecting the efficiency of algorithm.

After the above examination; it is observed that if such insignificant calculations are removed using an appropriate strategy; the complexity of this algorithm will be decreased that may result in increasing its use in various applications.

4. CASE STUDY

The shortcoming of K-means is being improved by Shi Na et al. [13] in their proposed method. The author has created two simple data structures which are used to retain the labels of cluster and the distance of all the date objects to the nearest cluster during the each iteration. The data stored can be used in next iteration and the distance between the current date object and the new cluster center is then calculated. If the distance computed is less than or equal to the distance to the old center, the data object remains in it is 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- 1 clustering centers, which saves the calculative time to rest of the k-1 cluster centers. Else, the distance from the current data object to all k cluster centers needs to be calculated, nearest cluster center is discovered and this point is assigned to the nearest cluster center. The label of nearest cluster center and the distance to its center are recorded separately. Due to lesser number of distance calculation, the time complexity is reduced that ultimately enhances the efficiency of k-means. This improved algorithm is implemented using datasets of iris and glass because they are fit to clustering analysis and their clustering results are much reliable. Experiment is performed on Window XP Operating system and program language used is VC++ 6.0.

After analyzing the experimental results, the author of this paper has prepared the following charts (Fig. 1 and Fig. 2) for glass dataset with an intention to make the observations more clear.

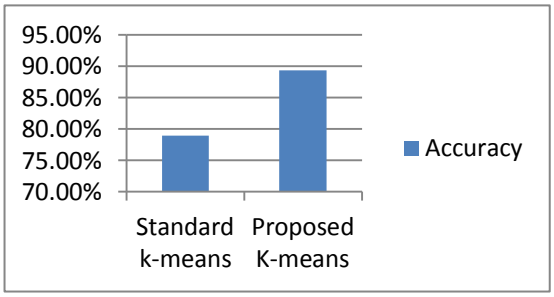

Fig-1 Comparison of standard K-means and Proposed K-means on the basis of Accuracy.

ISSN 2278-7763

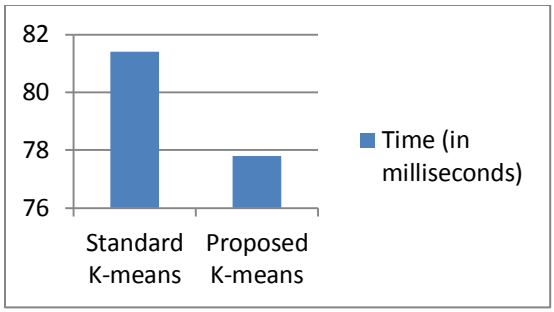

Fig-2 Comparison of standard K-means and Proposed K-means on the basis of Execution time.

The above research work studied by author proves that the proposed algorithm has lesser execution time as compared to original K-means. Along with it, the accuracy of the proposed method is also higher.
This case study emphasizes on the fact that if some factors of standard k-means algorithm are reconsidered, its performance will definitely rise up.

5. CONCLUSION

K-means is one of the most popular and an effective method to cluster large datasets which is used in number of scientific and commercial applications. However, this method has several loopholes such as getting trapped in local minima and large number of distance calculations that ultimately leads to high time complexity. The accuracy as well as complexity of this algorithm is based on certain criteria which includes the selection of initial centroid and the strategy used in performing calculations from each data object to different cluster centers. Determining the factors that highly influences the performance of this algorithm can play a significant role in making improvements in original k-means. The author assumed that the present research paper will definitely contribute a lot in making an improved K-means Clustering Algorithm.

6. REFERENCES

[1] Jiawei Han, Micheline kamber& Morgan Kauffman,"Data Mining: Concepts and Techniques", 2nd edition 2006.
[2] Sumathi & Kirubakaran, "Enhanced Weighted K-means Clustering based risk level prediction for Coronary heart disease", European Journal of Scientific research, ISSN 1450-216X, No. 4, pp. 490-500, 2012.
[3] Koheiet. al, "Hierarchical K-means: an algorithm for centroids initialization for K-means", Reports of the Faculty of Science and Engineering, Saga University, Vol. 36, No.1, 2007.
[4] Khammar & Marhaban, "Obtaining the initial centroids based on the most dense colonies in the k-means Algorithm", Reasearch Journal of Computer Systems & Engineering, ISSN: 2230-8563, vol. 03, issue. 01, July 2012.
[5] Madhu Yedla, Srinivasa Rao Pathakota& T M Srinivasa, "Enhancing K-means ClusteringAlgorithm with Improved Initial Center". International Journal of Computer Science and Information Technologies, Vol. 1 (2), 121-125, 2010.
[6]Bradley & Fayyad, "Refining Initial Points for K-means Clustering", International Conference of Machine Learning", pp. 91-99, May 1998.
[7] Sujatha & ShanthiSona, "Novel Initialization Technique for K-means Clustering using spectural Constraint Prototype", published in Journal of Global Research in Computer Science, Vol. 3 No. 6, ISSN-2229-371X, June 2012.
[8] Zhexue, "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values", Data Mining and Knowledge Discovery, Vol. 2, Issue. 3,pp. 283-304, 1998.
[9] Hamerly, "Making K-means Even Faster", Department Of computer Science, Baylor University.
[10]Charles Elkan. "Using the triangle inequality to accelerate k-means", In Tom Fawcett and Nina Mishra, editors, ICML, pages 147–153. AAAI Press, 2003.
[11]Dheeb, Malik & Sulieman, "Detection and Classification of Leaf Diseases using K-means-based Segmentation and Neural-networks-based Classification", Information Technology Journal, Vol. 10, Issue. 2, pp. 267-275, 2011.
[12] Azharet. al., "Enhanced K-means Clustering Algorithm To reduce number of iterations and time complexity", Middle east Journal of Scientific Research 12 (7): 959-963, 2012.
[13] Shi Na, Liu & Guan, "Research on K-Means Clustering Algorithm", Third International Symposium on Intelligent Information Technology and Security Informatics, ISBN- 978-0-7695-4020, 2010.
[14] Neha & Kirti, "A mid-point based k-Means Clustering Algorithm", International Journal of Computer Science and Engineering, ISSN 0975-3397, Vol. 4, No. 6, June 2012.

