International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-2 Issue-5, June 2013

Experimental Study of an Improvedk-Means Algorithm and Its Comparison with Standardk-Means

Sonal Miglani, KanwalGarg

Abstract: ***K-means algorithm is a popular, unsupervised and iterative clustering algorithmwell known for its efficiency in clustering large datasets.It is used in a variety of scientific applications such as knowledge discovery, Data Mining, data compression, medical imaging and vector quantization. This paper aims at studying the standard k-means clustering algorithm, analyzing its shortcomings and its comparison with an improved k-means algorithm. Experimental results show that the improved method can effectively increase the speed of clustering and accuracy, reducing the computational complexity of the k-means.***

Keywords - Clustering, Data Mining, K-Means Clustering

I. INTRODUCTION

The data mining functionalities include clustering analysis, characterization and discrimination, the mining of frequent patterns, associations and correlations, outlier analysis and classification and regression etc. Clustering is a preprocessing step in all data mining algorithms in which the data objects are classified into several subclasses such that it contains high intra-class similarity and low inter-class similarity.K-means algorithm is a centroid-based clustering technique in which clusters are represented by a central vector which may not necessarily be a member of dataset. It's a very simple and fast algorithm that follows an unsupervised, non-deterministic and iterative approach towards clustering. This method is proved to be very effective and can generate good clustering results in the process of data mining. However, the performance of the standard k-means algorithm can be enhanced further by considering several factors that plays a crucial role in deciding its functionality. 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.

Manuscript published on 30 June 2013.
***** Correspondence Author (s)
Sonal Miglani, Research Scholar, M.Tech.(CSE) Dept. of Computer Science & Applications Kurukshetra University, Kurukshetra, India
KanwalGarg, Assistant Professor Dept. of Computer Science & Applications Kurukshetra University, Kurukshetra, India

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC-BY-NC-ND license http://creativecommons.org/licenses/by-nc-nd/4.0/

It is observed that k means algorithm unnecessarily calculates the distance between each data object to all the other cluster centers, consuming a long execution time and therefore affecting the efficiency of algorithm. 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.
This paper includes four parts: The second part details the kmeans algorithm and shows the shortcomings of the standard k-means algorithm. The third part presents the improved k-means clustering algorithm, the last part of this paper describes the experimental results and conclusions through experimenting using MATLAB.

II. STANDARD K-MEANS ALGORITHM

The process classifies raw data based on the attributes/features into K number of group. K is positive integer number. The grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid. Conceptually, the centroid of a cluster is its center point. According to Jiawaei (2006) [1], the centroid in K mean algorithm can be defined by the mean of the objects assigned to the cluster. Suppose a data set, D, contains n objects. Partitioning methods distribute the objects in D into k clusters C_1 …..C_k, that is $C_i \cap C_j = \phi$ for (1<=i, j<=k). An objective function is used to assess the partitioning quality so that objects within a cluster are similar to one another but dissimilar to objects in other clusters. This is the objective function aims for high intra-cluster similarity and low inter-cluster similarity. The quality of cluster can be measured by the within cluster variation, which is sum of squared error between all objects in Ci and the centroid c_i ,defined as

$$E = \sum \sum \text{dist}(p, c_i)^2.$$

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.

Algorithm:

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:

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

- 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. Untilno change

The standard k-means clustering algorithm always converges to 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 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.

III. THE SHORTCOMINGS OF STANDARD K-MEANS ALGORITHM

The first shortcoming of the standard k-means algorithm is the method used for selection of initial centroids[4]. In the technique, 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.

The second shortcoming involved the process of distance calculations. The algorithm has to calculate the distance from each data object to every cluster center in each iteration. However, by experiments we find that it is not necessary for us to calculate that distance each time. Assuming that cluster C formed after the first j iterations, thedata object x is assigned to cluster C, but in a few iterations, the data object x is still assigned to the cluster C. In this process, after several iterations, we calculate the distance from data object x to each cluster center and find that the distance to the cluster C is the smallest. So in the course of several iterations, k-means algorithm is to calculate the distance between data object x to the other cluster center, which takes up a long execution time thus affecting the efficiency of clustering.

This shortcoming of K-means is being improved by Shi Na et al. [14] 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.

IV. PREVIOUS RESEARCH WORK

MadhuYelda et al. [4] proposed an enhanced K-means algorithm with the reduced time complexity O(nlogn).P.S Bradley &Usama M Fayyad [8] presented a fast and efficient algorithm for refining an initial starting point for a general class of clustering algorithms. Koheriet. al. [2], M. R. Khammar& M. H. Marhaban[5]in their papers proposed and improved K-means algorithm using different strategies ,that intends to remove the loophole of random selection of initial centroids. S. Sujatha& A. ShanthiSona [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.Neha et.al [12] proposed a mid-point based K-means clustering algorithm with improved accuracy. An efficient k-means algorithm is presented by Elkan [6] that is intended to remove a large number of distance calculations between data objects and cluster centers. Hamerly [10] proposed an algorithm which is a modified and simplified version of Elkan's k-means algorithm. . Zhexue[9] presented two algorithms which extend the k-means algorithm to categorical domains and domains with mixed numeric & categorical values using a simple matching dissimilarity measure. 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.

V. IMPROVED K-MEANS CLUSTERING ALGORITHM

The process of the improved K-means Algorithm[14] is described as follows:

Input:

The number of desired clusters k, and a database D={d1,d2,…dn} containing n data objects.

Output:

A set of k clusters

Steps:

1) Randomly select k objects from dataset D as initial cluster centers.
2) Calculate the distance between each data object di (1 <=i<=n) and all k cluster centers cj(1<=j<=k) as Euclideandistance d(di , cj) and assign data object di to the nearest cluster.
3) For each data object di, find the closest center cjand assign di to cluster center j;
4) Store the label of cluster center in which data object di is and the distance of data object di to the nearest cluster and storethem in array Cluster[ ] and the Dist[ ] separately.

Set Cluster[i]=j, j is the label of nearest cluster. Set Dist[i]=d(di, cj), d(di, cj) is the nearest Euclideandistance to the closest center.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

5) For each cluster j (1<=j<=k), recalculate the cluster center;
6) Repeat
7) For each data object di

Compute it's distance to the center of the present nearest cluster;

a) If this distance is less than or equal to Dist[i], the data
b) Else

For every cluster center cj(1<=j<=k), compute the distance d(di, cj) of each data object to all the center, assign the data object di to the nearest center cj. Set Cluster[i]=j; Set Dist[i]= d(di, cj);

8) For each cluster center j(1<=j<=k), recalculate the centers;
9) Until the convergence criteria is met.
10) Output the clustering results;

The time complexity of the improved k-means algorithm is O(nk). Here some data points remain in the original clusters, while the others move to another cluster. If the data point retains in the original cluster, this needs O(1), else O(k). With the convergence of clustering algorithm, the number of data points moved from their cluster will reduce. If half of the data points move from their cluster, the time complexity is O(nk/2). Hence the total time complexity is O(nk). While the standard k-means clustering algorithm require O(nkt).

VI. EXPERIMENTAL RESULTS

The author of this paper selects random datasets to test the efficiency of the improved k-means algorithm and the standard k-means. Two simulated experiments have been carried out to demonstrate the comparison between two algorithms. The same data set of 50 points is given as input to the standard k-means algorithm as well as to the improved algorithm for different values of K. Experiments compare improved k-means algorithm with the standard k-means algorithm in terms of the total execution time of clusters and number of iterations performed. Experimental operating system is Window 7,Programming Platform is MATLABversion 7.9.0.529 (R2009b). The results proved that the execution time and number of iterations are reduced in Improved K-means as compared to Standard K-means Algorithm. Author has prepared the following Table& Graph showing the experimental results using different values of K (Number of clusters).

Value of K	Number of Iterations in Standard K-means	Number of Iterations in Improved K-means
3	5	3
4	12	5
5	9	6
6	8	4

Fig. 1 (a) Comparison on the basis of Number of Iterations Performed for different values of K

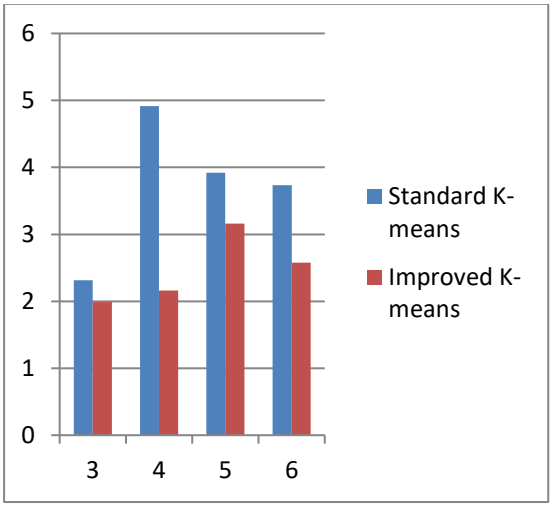

Fig. 1(b) Comparison on the basis of Execution Time (seconds) for different values of K

VII. 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. Various improvements have been carried out on the standard k-means algorithm by different researchers considering different shortcomings. The author of this paper performed a comparative study of k-means with its one of the improved algorithms using experimental approach. The results show a significant difference in the performance of two algorithms. The execution time as well as number of iterations is reduced resulting into increase of its use in various applications. The author assumed that the present research paper will definitely contribute a lot in the field of Data Mining.

REFERENCES

[1] Jiawei Han, Michelinekamber& Morgan Kauffman,"Data Mining: Concepts and Techniques", 2nd edition 2006.
[2] Koheriet. 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.
[3] 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.
[4] MadhuYedla, SrinivasaRaoPathakota& 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.
[5] 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.
[6] Charles Elkan. "Using the triangle inequality to accelerate k-means", In Tom Fawcett and Nina Mishra, editors, ICML, pages 147–153. AAAI Press, 2003.
[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.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

[8] Bradley & Fayyad, “Refining Initial Points for K-means Clustering”, International Conference of Machine Learning”, pp. 91-99, May 1998.
[9] 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.
[10] Hamerly, “Making K-means Even Faster”, Department Of computer Science, Baylor University.
[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] 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.
[13] 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.
[14] 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.

© Copyright: All rights reserved.