ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

An Optimizing K-Means Algorithm

[I]Bhawna Aggarwal, [II]Puneet Rani
[I]M.Tech Scholar of computer science & Engineering, SRCEM, Palwal
[II]Dept. of computer science & Engineering, SRCEM, Palwal

Abstract
Data mining is a new technology, developing with database and artificial intelligence. It is a processing procedure of extracting credible, novel, effective and understandable patterns from database. Cluster analysis is an important data mining technique used to find data segmentation and pattern information. By clustering the data, people can obtain the data distribution, observe the character of each cluster, and make further study on particular clusters. In addition, cluster analysis usually acts as the preprocessing of other data mining operations. Therefore, cluster analysis has become a very active research topic in data mining. As the development of data mining, a number of clustering methods have been founded, The study of clustering technique from the perspective of statistics, based on the statistical theories, our paper make effort to combine statistical method with the computer algorithm technique, and introduce the existing excellent statistical methods, including factor analysis, correspondence analysis, and functional data analysis, into data mining.

Keywords
Data mining, Clustering, K-mean clustering algorithm

I. Introduction

Data clustering is a data exploration technique that allows objects with similar characteristics to be grouped together in order to facilitate their further processing. Data clustering has many engineering applications including the identification of part families for cellular manufacture. The K -means algorithm is a popular data- clustering algorithm. To use it requires the number of clusters in the data to be pre-specified. Finding the appropriate number of clusters for a given data set is generally a trial-and-error process made more difficult by the subjective nature of deciding what constitutes 'correct' clustering.

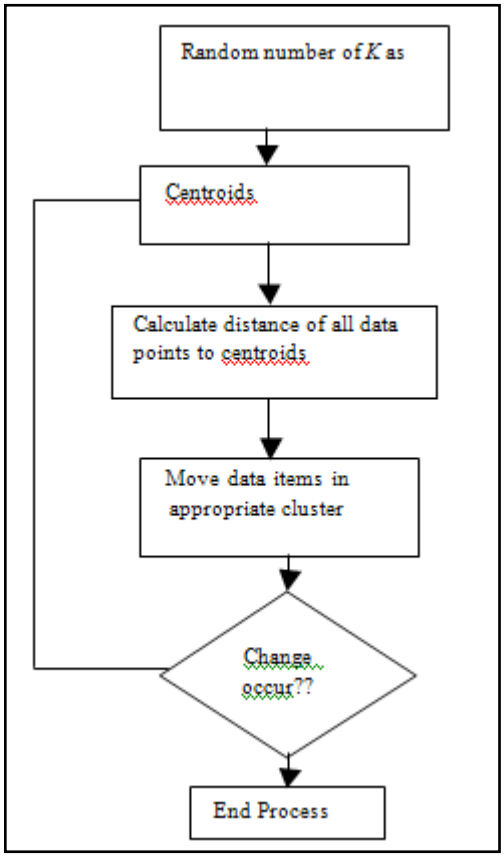
II. Basic K-mean Clustering Algorithm

According to the basic K-mean clustering algorithm, clusters are fully dependent on the selection of the initial clusters centroids. K data elements are selected as initial centers; then distances of all data elements are calculated by Euclidean distance formula. Data elements having less distance to centroids are moved to the appropriate cluster. The process is continued until no more changes occur in clusters.
Following are the algorithmic steps for basic K-mean algorithm

INPUT: Number of desired clusters K ,Data objects D= {d 1, d2…dn}
OUTPUT: A set of K clusters
Steps:

- Randomly select k data objects from data set D as initial centers.
  Repeat;
  Calculate the distance between each data object di (1 <= i<=n) and all k clusters C j(1 <= j<=k) and assign data object di to the nearest cluster.
- For each cluster j (1 <= j<=k), recalculate the cluster center.
- Until no change in the center of clusters.

Time complexity of K-mean Clustering is represented by O(nkt). Where n is the number of objects, k is the number of clusters and t is the number of iterations.
The following figure shows steps of the basic K- mean clustering algorithm .



III. Proposed Algorithm

The process and algorithmic steps of proposed algorithm are given.

A. Clustering Process

In proposed algorithm, the input remains in the same order in which data items are entered. The whole process is divided into two phases.
Phase-I: In phase-I, the cluster size is fixed and the output of the first phase forms initial clusters. Here, the input array of elements is scanned and split up into sub-arrays, which represent the initial

clusters.

Phase-II: In phase-II, the cluster sizes vary and the output of this phase are the finalized clusters. Initial clusters are inputs for this phase. The centroids of these initial clusters are computed first, on the basis of which distance from other data elements are calculated. Furthermore the data elements having less or equal distance remains in the same cluster otherwise they are moved to appropriate clusters. The entire process continues until no changes in the clusters are detected.

B. Steps of Algorithm

Algorithm is divided into two Phases.

In Phase-I, we find the initial clusters, while in Phase-II, data elements are moved in appropriate clusters.

Phase-I:

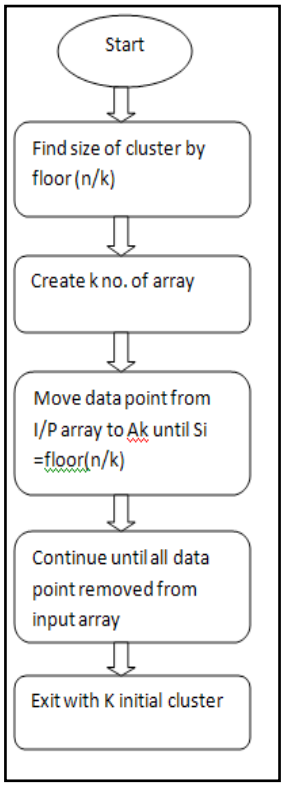
Algorithm 2.1- To find the initial clusters

INPUT: Array {a1, a2, a3,…… an}, K //Number of Required clusters

OUTPUT: A set of Initial Clusters.

Steps:

- Find the size of cluster Si (1= i = k) byFloor (n/k).
- Where n= number of data points Dp (a1, a2, a3, …… an) K= number of clusters.
- Create K number of Arrays Ak
- Move data points (Dp) from Input Array to Ak untill Si =Floor (n/k).
- Continue Step 3 untill all Dp removedfrom input array
- Exit with having k initial clusters.



Phase-II:

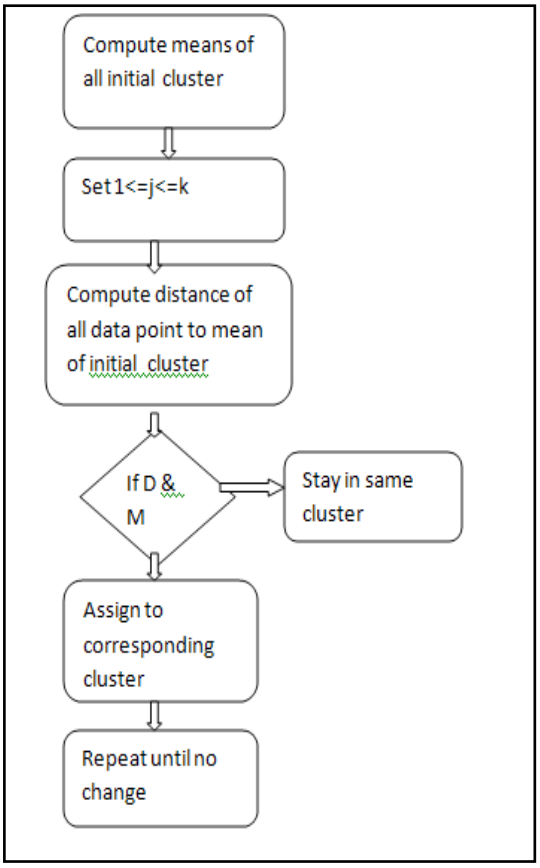
Algorithm 2.2- To find the final clusters

INPUT: A set of Initial Clusters.

OUTPUT: A set of k Clusters.

Steps:

- Compute the Arithmetic Mean M of all initial clusters CI
- Set 1<= j<= k
- Compute the distance D of all Dp to M of Initial Clusters Cj
- If D of Dp and M is less than or equal to other distances of Mi (1 i k) then Dp stay in same cluster
- Else Dp having less D is assigned to Corresponding Ci
- For each cluster Cj (1<= j<= k), Recompute the M and move Dp until no change in clusters.



IV. Experimental Work

Experimental work was designed to compare the performance of proposed K-mean algorithm. Number of data elements selected was 1000. And for the sake of experiment, 8 numbers of clusters (k) were entered at run time. The process was repeated 10 times for different data sets generated by MATLAB. The proposed K-mean algorithm is efficient because of less number of iterations and improved cluster quality, as well as reduced elapsed time. In Figure 2, Basic and proposed K-mean clustering algorithms are compared in terms of different data sets.For each run different data sets are generated by MATLAB and entered, to observe the number of iterations.

In Figure 3, Basic and proposed K-mean clustering algorithms are compared in terms of same data set. For each run same data set is entered, to observe that at each time numbers of iterations are different in basic K-mean clustering algorithm. The numbers of iterations are fixed in proposed K-mean clustering algorithm because initial centroids are not selected randomly.Basic K-mean clustering algorithm gives different clusters, as well as clusters size differs in different runs.

Table 1 shows different results for same data set as well as elapsed time. Proposed K-mean clustering algorithm gives same clusters, as well as clusters size is same in different runs. Table 2 shows same number of iterations and cluster size.

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

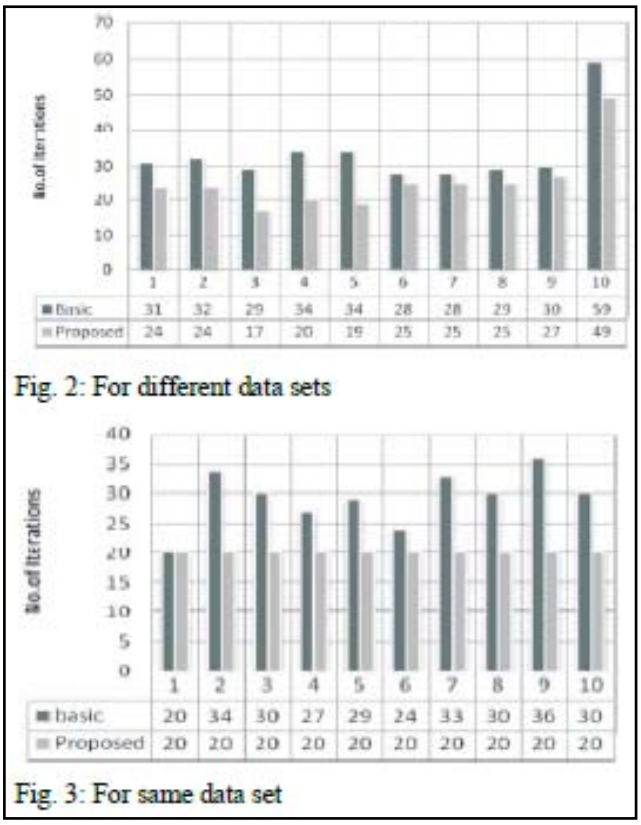


Fig. 2: For different data sets

Fig. 3: For same data set

V. Comparison of Basic and Proposed K-Mean Clustering Algorithm

Proposed K-mean algorithm is efficient from basic K-mean algorithm in terms of iterations, cluster quality as well as elapsed time. As in basic K-mean algorithm, initial centroids are selected randomly from the input data, so clusters vary from one another, because of which the number of iterations and total elapsed time also changes in each run of the same data. In proposed K-mean algorithm initial centroids are calculated and as the data is same, it results in same calculations, so the number of iterations remains constant and elapsed time is also improved. This is the reason that proposed K-mean clustering algorithm is efficient from basic K-mean algorithm.

Comparison of Proposed K-Mean Clustering Algorithm with Other Enhanced K-Mean Clustering Algorithm:

In [2], initial clusters are based on the searching mechanism. First two smallest elements are searched and those elements are then deleted from the input array and moved to the new sub arrays. The threshold value is set to fix the size of initial cluster and the process is continued to find initial cluster. In proposed K-Mean algorithm , there is no searching mechanism so the running time of the proposed algorithm is improved as compare to other techniques.

VI. Conclusion

One of the partitional clustering algorithms is K-mean clustering algorithm which depends on initial clusters.

In basic K-mean clustering, initial clusters are based on randomly selected centroids.

In this paper, an enhanced K-mean algorithm is introduced and compared with the basic K-mean algorithm. In enhanced K-mean clustering algorithm any type of integer data is used. The performance of basic K-mean clustering algorithm in terms of number of iterations and time complexity is improved. In future, this idea can be tested on text based clustering

References

[1]. Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurasamy, "Advances in Knowledge Discovery and Data Mining", AAAI Press/ The MIT Press, 1996.

[2]. J. Ross Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, 1993.

[3]. Michael Berry and Gordon Linoff, "Data Mining Techniques (For Marketing, Sales, and Customer Support), John Wiley & Sons, 1997.

[4]. Sholom M. Weiss and Nitin Indurkhya, "Predictive Data Mining: A Practical Guide", Morgan Kaufmann Publishers, 1998.

[5]. Alex Freitas and Simon Lavington, "Mining Very Large Databases with Parallel Processing", Kluwer Academic Publishers, 1998.

[6]. A. K. Jain and R. C. Dubes, "Algorithms for Clustering Data", Prentice Hall, 1988.

[7]. V. Cherkassky and F. Mulier, "Learning From Data", John Wiley & Sons, 1998.

Author's Profile

I (Bhawna aggarwal) is a scholar of M.Tech (C.S.E) in Shri Ram College of Engg And Management Palwal. I have done my B.Tech from B.S.A.I.T.M Faridabad with honors degree in I.T.