Analysis & Implementation of Clustering Data Mining Technique - An Approach to Efficient K-means Algorithm

Mr. Vilas Mahatme [1], Ms. Shital Radke [2]

[1]Head, CT Department, Kavikulguru Institute of Technology & Science, Ramtek, India
[2] M.Tech. Student, CSE Department, Kavikulguru Institute of Technology & Science, Ramtek, India

Abstract— *Thousands of techniques are emerging for collecting scientific and real life data on a large scale. The traditional database queering system is available to extract required information. Clustering is one of the important data analysis techniques in data mining. K means are mostly used for many applications. While using basic k means algorithm one may face difficulties in optimizing the results as it is computationally expensive in the terms of iterations. The result of k means is based on selection of initial seed and total number of clusters which depends on the data set. As K means gives variety of results in each run, there is no any thumb rule available for selection of initial seed and the number of clusters. This paper includes approaches to improve the efficiency of k means is mentioned, which provides a better way of selecting the value of k. This approach will result in better clustering with reduced complexity.*

Keywords— clustering, data analysis, k means, initial seed.

I. INTRODUCTION

Data mining is the process of identifying some patterns or knowledge of the large collection of data. Due to the increased speed of technology, data keeps on collecting in various ways. This collected data is not simply the raw data, but meaningful information is taken out to explore some knowledge out of it. Various data mining techniques are available for handling the data. Clustering is one of the important data mining techniques for handling uncertain data. Data is divided into different groups based on some similarity and dissimilarity context. A good clustering method produces quality cluster with more intra-class similarity and less inter-class similarity. Clustering is widely used in various applications such as artificial intelligence, data compression, image processing, machine learning, and pattern recognition. Various clustering techniques are available such as hierarchical, partition, greed based and density based. K means are mentioned in this paper is partially based clustering algorithm. A clustering algorithm separates a data set into different groups so that the similarity within a group is larger than among groups.[3] Clustering is a mechanism where a set of patterns (data), usually multidimensional data is classified into sections (clusters) so that data elements of one section are similar according to a predefined criterion. K means is popular clustering algorithm used for many applications like market analysis and fraud detection. This technique uses an iterative approach for partition based clustering. Uncertain data can be easily handled with k means portioning. This paper explains the difficulties faced by traditional k means and the next part of the paper provides an approach to improve efficiency. Efficiency is provided in the term of computation and accuracy in the term of time complexity of k means. Complexity of k means is the product of the number of data items, number of iterations and the number of clusters. K means are sensitive to the initial selection of clusters and calculation of centroids. As the numbers of clusters are increased, they may get overlapped with each other.

II. CLUSTERING WITH K MEANS

As disused above; this paper aims at finding the efficient value of the parameter k over data space x for grouping the data in appropriate number of clusters. The basic rule is to divide available data set into k number of different clusters where the value of k is not constant for individual cluster. Here the centroids of cluster need to define in the term of mean parameter values separate for each cluster. Mostly Euclidean distance is used to calculate the distance data items and centroids of clusters. K means aimed at minimizing an objective function known as the mean squared error function given by:

$$(Y)= \sum_{i=1,j=1}^{ci\ cj}\sum(\|X_i-Y_j\|)$$

Here c specifies the number of clusters on different data item i and j whose value is changed from 1 to me, j. X is the dataset as X=(x1, x2,…..xn) and Y is the centroids of clusters.

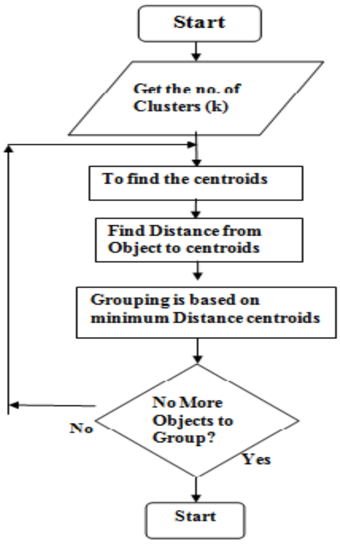
Start
Get the no. of Clusters (k)
To find the centroids
Find Distance from Object to centroids
Grouping is based on minimum Distance centroids
No More Objects to Group?
No
Yes
Start

Fig. 1: K-Means Flowchart

Above flowchart shows the process of k means. The k means algorithm is easy to implement. Overall idea of algorithm is explained in following steps

- Choose a number of desired clusters, *k*.
- Choose *k* starting points that can be used as initial cluster centroids.
- Examine each point in the data set.
- Assign it to the cluster whose centroid is nearest to it.
- When every point is assigned to a cluster, again calculate the new *k* centroids.
- Repeat steps 3 and 4 until all the points are covered.

III. ANALYSIS & CLUSTERING OF FOREST COVER TYPE DATASET WITH K MEANS

The multi variant forest cover type data set is taken from UCI repository of machine learning, database [14], is used for implementing k means. This dataset contains the information taken over 30*30 meter cell. No remotely sensed data is used. Total number of instances used is 5, 81,012 with 54 attributes. Various columns are Wilderness Area (4 binary columns), Soil Type (40 binary columns), Cover Type (7 types) and many more. Following figure shows output of k means algorithm implemented on forest dataset. It divides the data set into 7 clusters with seven different colors. Here, clusters are formed, but they are overlapped with each other with their centroids. This algorithm is iterative and it takes more time for computation of clusters.

Result: 22 iterations.

Total sum of distances = 1.3348e+006.

Elapsed time is 4.131132 seconds.

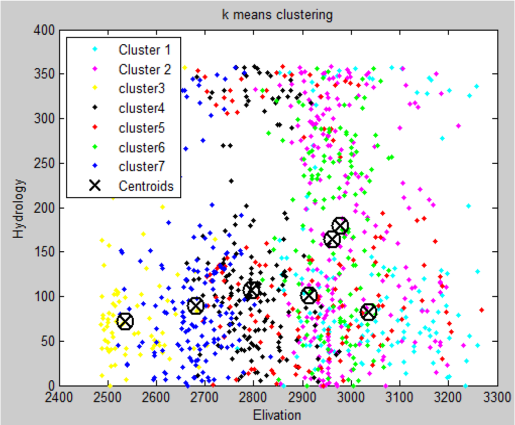

Fig. 2: Result of k Means

IV. PROPOSED APPROACH FOR IMPROVEMENTS IN K MEANS

To improve the accuracy, first need to work with the problem of selection of the initial seed value. Initially calculate the distance between each data point and rest of other data points in the data set. Next find out closet pairs of data sets and formulate new group as Z1 consisting of only closest data pairs. Then remove these data pairs from the original dataset. Continue the procedure until all the closet pairs are considered from the original dataset. Then go to original dataset again calculate the minimum distance with different attributes and formulate the new pairs. At this point form group Z2 and repeat the steps until Zn is obtained with all data pairs. Euclidean distance is used as a distance parameter to calculate closet distance. As a result of the above execution required initial centroids are obtained. Then these centroids can be used to calculate the appropriate number of clusters for the dataset.

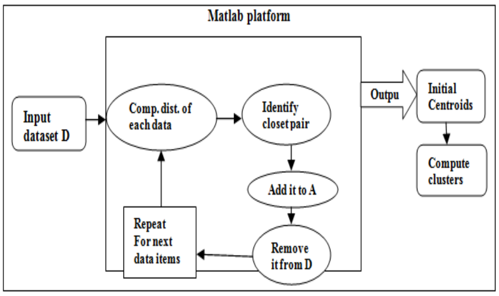

Fig. 3: Proposed Approach Plan

V. CONCLUSION

This paper discussed about k means algorithm which is partition based clustering and its primary values are depend on the initial seed value and centroids of the cluster. Experimental results shows that, it is difficult to predefine the value of k as cluster centers may get overlapped with each other. Mainly a problem in implementing k means is, as it increases the computational complexity with large number of iterations. Hence proposed approach of this paper can improve the problems of k means and try to improve the efficiency of k means up to a certain extent of minimizing the number of iterations. When the approach presented in this paper is implemented it will show the result which will overcome the problems of k means and provide more accuracy with the k means. In future the same proposed approach can be implemented with different dataset by using different preprocessing techniques. Various attributes can be tried for better result.

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, 2011, pp 717-721.

[3] Ahamed Shafeeq B M and Hareesha K S “Dynamic Clustering of Data with Modified K-Means Algorithm” International Conference on Information and Computer Networks, ICICN 2012, pp 221-225.

[4] D T Pham, S S Dimov, and C D Nguyen “Selection of K in K-means Clustering”, Mechanical Engineering Science, 2004, pp. 103-119.

[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] Khan, S.S., Ahmad, A., “Cluster Center Initialization Algorithm for K-means Clustering”, Pattern Recognition Letter. 25, 2004, pp. 1293–1302.

[7] Jiawei Han and Micheline Kamber, “Data Mining: Concepts and Techniques”, Morgan Kaufmann Publishers, Second Edition, 2006.

[8] Grigorios F. Tztzis and Aristidis C. Likas, “The Global Kernel K-means Algorithm for Clustering in Feature Space”, IEEE Transaction, On Neural Networks, Vol. 20, no. 7, July 2009, pp. 1181-1194.

[9] R. Xu and D. Wunsch, “Survey of Clustering Algorithms”, IEEE Transaction 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), April 2010, pp.63-67.

[11] Komarasamy G and Amitabh Wahi, “An Optimized K-means Clustering Technique Using Bat Algorithm”, European Journal of Scientific Research, ISSN 1450-216X Vol.84 no.2, August 2012, pp.263 – 273.

[12] K. A. Abdul Nazeer, M. P. Sebastian, “Improving the Accuracy and Efficiency of the K-means Clustering Algorithm”, Proceedings of the World Congress on Engineering 2009 Vol 1, WCE 2009, July 1 - 3, 2009, London, U.K.

[13] Mehmet Koyutu¨rk, Ananth Grama and Naren Ramakrishnan, “Compression, Clustering and Pattern Discovery in Very High-Dimensional Discrete-Attribute Data Sets,” IEEE Transactions on Knowledge and A Data Engineering”, Vol. 17, no. 4, Apr 2005, pp.447-461.

[14] Merz C and Murphy P, UCI Repository of Machine Learning Databases, Available ftp://ftp.ics.uci.edu/pub/machine-learning-databases.