ISSN:2231-0711

Saroj et al | IJCSET(www.ijcset.net) | July 2016 | Vol 6, Issue 7, 279-281

Review: Study on Simple k Mean and Modified K Mean Clustering Technique

Saroj
Student of Masters of Technology,
Department of Computer Science and Engineering
JCDM college of Engineering, SIRSA, GJU, Hisar, Haryana, India

Kavita
Assistant Professor,
Department of Computer Science and Engineering
JCDM college of Engineering, SIRSA, GJU, Hisar, Haryana, India

Abstract: The main aim of this review paper is to provide comprehensive review of simple k mean clustering and modified k mean clustering techniques. Clustering is used as active research in various fields like statistics, pattern recognition and machine learning etc. Cluster Analysis is data mining tool for a large and multivariate database. Clustering is the one of data mining techniques in which data is divided into the groups of similar objects and dissimilar objects into another group. Clustering is a suited example of unsupervised classification.

Keywords: Clustering, clustering Technique: Simple K Mean and Modified k mean clustering.

I. INTRODUCTION

Due to the increased availability of computer hardware and software and the rapid computerization of business, large amount of data has been collected and stored in databases. Researchers have estimated that amount of information in the world doubles for every 20 months [1]. However raw data cannot be used directly. Its real value is predicted by extracting information useful for decision support. In most areas, data analysis was traditionally a manual process. When the size of data manipulation and exploration goes beyond human capabilities, people look for computing technologies to automate the process. Data mining is one of the youngest research activities in the field of computing science and is defined as extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data. Data mining is the process of analyzing data from different perspectives and summarizing it into useful information [2]. Data mining consists of extract, transform, and load transaction data onto the data warehouse system, Data mining includes the anomaly detection, association rule learning, classification, regression, summarization and clustering. Data mining is one of the most important research fields which are due to the expansion of both computer hardware and software technologies, which has imposed organizations to depend heavily on these technologies. Data mining concepts and methods can be applied in various fields like marketing, medicine, real estate, customer relationship management, engineering, web mining, etc. Various clustering algorithms according to different techniques have been designed and applied to various data mining problems successfully. In this paper, clustering analysis is done by using simple k mean clustering and Modified k mean clustering. Normalization and Indexing is an important preprocessing step in to standardize the values of all variables from dynamic range into specific range. Cluster analysis is type data mining technique which is used to find data segmentation and pattern information. By clustering the data people get the data distribution, observe the character of each cluster, and make further study on particular clusters. The aim of cluster analysis is that the objects in a group should be similar to one another and different from the objects in other groups. Clustering is much better when there is greater similarity within a group and greater the difference between the groups. So we can say that raw data has to be used with the algorithm to extract useful information from it. Various clustering algorithms according to different techniques have been designed and applied to various data mining problems successfully. The most commonly used algorithms in Clustering are Hierarchical, Partitioning, Density Based and Grid based algorithms. The popular clustering techniques which have been suggested so far are either partition based clustering or hierarchical clustering but both approaches have their own advantages and disadvantages in terms of the number of clusters, shape of clusters, and cluster overlapping. When any clustering algorithm is applied to the raw data, only then we can get clusters which are useful as shown in fig. [2].

Fig 1: Stages of Clustering [3]

II. PARTITIONING CLUSTERING

Data objects are partitioned into non overlapping clusters so that each and every object is in exactly in one subset. The reason of division of the data into several subsets is that checking of the all possible subset systems is computationally not feasible; there are certain greedy heuristics schemes which are used in the form of iterative optimization. This means different relocation schemes that iteratively reassign points between the k clusters [3].

Fig 2 before and after partitioning

A. Simple K-Means Clustering

It is a partitioning method which finds mutual exclusive clusters of spherical shape. It generates a specific number of disjoint, flat (non-hierarchical) clusters. K-Means algorithm recognizes objects into k – partitions where each partition represents a cluster. We start out with initial set of means and classify cases based on their distances to their centers. Next, we compute the cluster means again, using the cases that are assigned to the clusters; then, we reclassify all cases based on the new set of means. We keep repeating this step until cluster means don't change between successive steps. Finally, we calculate the means of cluster once again and assign the cases to their permanent clusters. [4]

1) Method for Simple K means clustering

1 Input: k= no. of clusters. D= data set that contains n objects.
2 Output: Set of k clusters.

Method:

1. Randomly choose k objects from D as the initial cluster centre.
2. Repeat.
3. Reassign 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.

2) Dataset

The dataset is set of data items and this is a very basic concept of machine learning. A dataset is equivalent to a two-dimensional spreadsheet or database table. In WEKA, dataset is implemented by the weka.core.Instances class. A dataset is a collection of examples; each example can be taken from one of class weka.core.Instance. Each Instance made of a number of attributes, any of which can be nominal, numeric or a string. Data set is used here medical data.

3) Overview of WEKA Tool

For working with WEKA we do not have any need the deep knowledge of data mining because it is very popular data mining tool. WEKA is open source and freely available as well as platform-independent user and provides many facilities. It also provides the graphical user interface to the user. WEKA is a landmark system in the history of the data mining and machine learning research communities, because it is the only toolkit that has gained such widespread adoption and survived for an extended period of time. It provides different algorithms for data mining and machine learning. Simple k mean clustering is done by using this tool. We have to give the different datasets as input to the WEKA tool. After that it will process the input and gives the output. In this way clusters will be formed.

Fig 3 Front view of WEKA tool

III. MODIFIED K MEAN CLUSTERING

Modified K Mean approach is designed to improve the time, no. of iterations and sum of squared errors and this provides much better result as compare to Simple K Mean clustering done by using K Mean tool. This is simple to use and it also provides graphical user interface to the user. In Modified K Mean clustering data is reduced by normalized method and then the parameters such as time taken, no. of iterations, Sum of squared errors are improved by using normalized technique. Normalization index method used for modification. In this method input data has to be minimized by using indexing so that we can get the raw data in sequence after that we calculate the Euclidean distance between different clusters. After that normalization is done to minimize the sum of squared errors and no. of iteration resulting in less execution time to make the grouping of clusters.

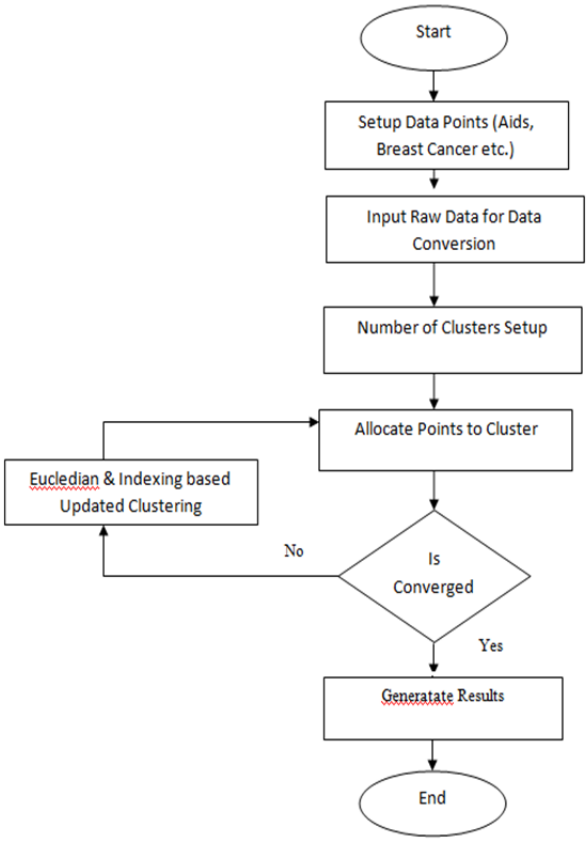

Fig 4 Flow chart of Modified k means clustering

First step according to the given flow chart is that we have to provide the raw data for conversion of data using indexing and calculation of Euclidean distance is done so that grouping of clusters is done in similar and dissimilar categories.

IV. Conclusion

K Mean clustering is the most important type of Partitioning clustering. Partitioning clustering is the one in which clusters are partitioned according to their distances. K mean clustering is that technique in which K cluster is chosen and cluster which are at the less distance from k cluster is selected in one group and others which are farthest from the K cluster is placed in different group. In this paper simple k mean clustering has been described by using the WEKA tool and taking the medical data set i.e Pima, Aids, Breast Cancer. On the other hand Modified k mean clustering has been described based on normalization and indexing approach using .NET which takes less time with minimum no. of sum of squared errors to execute the cluster.

Acknowledgment

I would like to thank Computer Science Engineering department of JCDMCOE, SIRSA for the support and providing an environment for this research work.

References

[1] Shraddha Shukla and Naganna S. “A Review on K-means Data Clustering Approach” *International Journal of Information & Computation Technology, Volume 4, 2014*

[2] S.Anupama Kumar and M. N. Vijayalakshmi “Relevance of data mining techniques in editification sector”, *International Journal of Machine Learning and Computing, Volume 3, Issue 1, February 2013.*

[3] Saroj, Tripti Chaudhary, “Study on Various Clustering Techniques”, *International Journal of Computer Science and Information Technologies, Volume 6, Issue 3, 2015.*

[4] Narender Kumar, Vishal Verma, Vipin Saxena “Cluster analysis in data mining using k-means method”, *International Journal of Computer Applications, Volume 76, Issue.12, August 2013.*

[5] Aastha Joshi, Rajneet Kaur “A Review: Comparative Study of Various Clustering Techniques in Data Mining” *International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 3, March 2013.*

[6] Amandeep Kaur Mann & Navneet Kaur “Review paper on Clustering Techniques”, *Global Journal of Computer Science and Technology Software and Data Engineering, Volume 13, Issue 5, Year 2013.*

[7] Bharat Chaudhary, Manan Parikh “A Comparative Study of Clustering Algorithm using WEKA Tool”, *International Journal of Application or Innovation in Engineering and Management, Volume 1, Issue 2, October 2012.*

[8] Vaishali R. Patel, Rupa G. Mehta, “Clustering Algorithms: A Comprehensive Survey”, *International Conference on Electronics, Information and Communication Systems Engineering,* 2011.