International Journal of Advanced Technology in Engineering and Science
Vol. No.4, Issue No. 08, August 2016
www.ijates.com
ijates
ISSN 2348 - 7550

NORMALIZATION INDEXING BASED ENHANCED GROUPING K-MEAN ALGORITHM

Saroj[1], Ms. Kavita2

[1]Student of Masters of Technology, [2]Assistant Professor

Department of Computer Science and Engineering

JCDM college of Engineering, SIRSA, GJU, Hisar, Haryana, (India)

ABSTRACT

This paper explains the concept of K-Mean Clustering. Today almost work is done on Internet. So, data mining becomes necessary for easy searching of data. Clustering is an important technique of data mining. Clustering is that technique of data mining which divides the data into similar and dissimilar groups. The performance of k mean clustering depends upon centroids selection and frequency of nearest data. This paper proposed the modified and efficient method of Simple k mean clustering using WEKA tool and modified approach of K-Means clustering. In Modified approach of K mean clustering, the entire data will be reduced by normalized and indexing method. Thus fast clustering process will reduce the system resources and gives the efficient technique to generate the clusters in minimum time, contains less no. of squared errors and having minimum no. of iterations.

Keywords*: *Data mining, Modified K-mean Clustering, normalization and Indexing, Partitioning clustering, Simple k mean clustering

I. INTRODUCTION

Data Mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data Mining represents a set of specific methods and algorithms aimed solely at extracting patterns from raw data. The various techniques of data mining are classification, clustering, prediction, Association Rule. Clustering is particularly useful in those cases where the most common categories within the data set are not known in advance. The objective of clustering analysis is to find the segments or clusters and to examine their attributes and values. A simple explanation of clustering can be made from the following scenario. In a pharmacy, there is a wide range of medicines available. The challenge is how to keep those medicines in a way that pharmacist can take from several medicines in a particular order without hassle. By using clustering technique, we can keep medicines that have some kinds of similarities or alphabetical order in one cluster or in one shelf and label it with alphabets. If pharmacists want to take medicines in that alphabet, they would only have to go to that shelf instead of looking for entire pharmacy. Numerous algorithms are used for clustering. A partition clustering algorithm partition the data set into desired number of sets in a single step [1].

1.1 Nature of Problem

The first step is to analyze the simple k mean clustering using WEKA tool. Limitations in Simple k mean has to be removed Simple K-means is a widely used partition clustering method in the industries. This algorithm is the most commonly used partitioning clustering algorithm. Here's shown how k-mean algorithm works:

Input: k= the number of clusters. D= a data set that contains n objects.

Output: Set of k clusters.

Method:

1. Arbitrarily choose k objects from D as the initial cluster centres.
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. Untill no change.

Simple K-Means algorithm is implemented in different phases. In first phase k centers are chosen randomly, and the second phase consists of finding the nearest center for each data object that is done by calculating Euclidean distance. Completion of first step is included 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 repeated until the objective function is minimized. This algorithm takes much more time to execute the cluster. So we have to design such an algorithm which takes less time to execute the cluster and provides accuracy by minimizing the sum of squared errors and no. of iterations.

1.2 Related Work

[2]This paper gives a brief introduction to cluster analysis with an emphasis on the challenge of clustering high dimensional data. The principal challenge in extending cluster analysis to high dimensional data is to overcome the "curse of dimensionality," and we described, in some detail that in what way high dimensional data is different from low dimensional data, and how these differences might affect the process of cluster analysis.. In particular, there is no reason to expect that one type of clustering approach will be suitable for all types of data, even all high dimensional data. Statisticians and other data analysts are aware of the need to apply different tools for different types of data, and clustering is no different. Finally, high dimensional data is only one issue that needs to be considered when performing cluster analysis.

[3]Author represents that K-mean algorithm has biggest advantage of clustering the large data sets and its performance increases as number of clusters increases. But use of k mean is limited to numeric values. So Agglomerative and Divisive Hierarchical algorithm was adopted for categorical data, but because of its complexity a new approach for assigning rank value to each categorical attribute using K- means can be used in which categorical data is first converted into numeric by assigning rank. Hence performance of K- mean algorithm is better than Hierarchical Clustering Algorithm. Density Based methods OPTICS, DBSCAN are used to find clusters of arbitrary shape whereas partitioning and hierarchical methods are designed to find the spherical shaped clusters.

[4]Author says that the biggest advantage of using K-Means algorithm is its efficiency in handling numeric data sets for clustering but when it comes to categorical data set, the proposed algorithm which is an extension of traditional K Mode has been shown to work efficiently. The proposed Modified K Mode algorithm has been tested against 4 sets of data with the parameters of accuracy, error rate, and computation time and memory usage. It was found that the proposed K Mode algorithm yields better results in terms of all the above said parameter particularly giving better computation time under different conditions set.

1.3 Purpose

In this research work main focus is to improve the efficiency and accuracy of existed simple k mean clustering algorithm so that we can generate such clusters so that it contains high intra-class similarity and low inter-class similarity. This grouping is performed by minimizing the sum of squares of distances between data and the corresponding cluster centroid. To improve the simple k mean algorithm we have to make such an algorithm which uses the normalization based indexing technique. After that we analyze the results of both clustering technique that is simple and modified k mean clustering technique.

II. PROPOSED ALGORITHM FOR MODIFIED K MEANS CLUSTERING

Simple k mean clustering algorithm has been improved by using no. of parameters like instances in cluster 1, instances in cluster 2, time taken to make the clusters, no. of iterations, sum of squared errors and this improved algorithm is named as modified k mean algorithm. This algorithm has been implemented by using C#.NET. In this algorithm normalization and indexing method has been used for reducing the large amount of data set. The Modified k-means algorithm is given as follows:

1. Set up Data Points (Breast, Aids, Etc).
2. Select Row Data as Input for Cluster/Group Generation
3. Enter the Number of Clusters Setup
4. Implement Equation 1 on Data Points as Preprocessing Setup.
5. If Group Converged, move to Next Step

Else

6. Move to Initial Step
7. Euclidian and Indexing Based Update Clusters
8. Generate Results.

$$\text{Output} = \frac{(\text{Input}` - (Index((Max(I1)) < Index(Min(I2))?))}{(\text{Max-Min})} \quad (1)$$

2.1 Module for Preparation of Data

This module works in two parts. First, data preprocessing technique apply on the dataset received by transform module. The clean data is then passed to second part; Data Normalization which transform the clean raw data into specific range using different techniques.

2.2 Data Pre-processing

This is a very important step since it can affect the result of a clustering algorithm. This module calculates tuples using different options like maximum, minimum, constant, average and standard deviation before we apply normalization approach on the dataset. This process gives the indexed value of data set after that it applies to the second part (data normalization) of data preparation.

2.3 Normalization based Indexing Approach

Data Mining can generate effective result if proper and effective data mining technique can apply to the dataset. According to author [9], normalization is used to standardize all the features of the dataset into a specified predefined criterion so that redundant objects can be removed and use made of valid and reliable data which can effect and improve accuracy of the result. Indexing is done before applying the normalization. Indexing is used to arrange the data values in either ascending order or descending order. The importance of normalization is that it increases the accuracy of the results that are obtained while clustering. The Min-Max normalization technique involves the linear transformation on raw data. Min I1 and Max I2 are minimum and maximum value for the different attributes. This technique maps the value of attribute A into range [0, 1].

III. EXPERIMENTAL RESULTS OF COMPARISON OF SIMPLE AND MODIFIED K MEAN ALGORITHM

3.1 Comparison according to time taken to build the model

Time taken by modified k mean clustering for data sets Pima, Aids and Breast cancer is 0.005, 0.013and 0.004respectively. Time taken by simple k mean clustering for data sets Pima, Aids and Breast cancer is 0.2, 0.2and 0.05 sec. respectively. Fig 1 shows that Modified k mean algorithm takes less time to build the cluster.

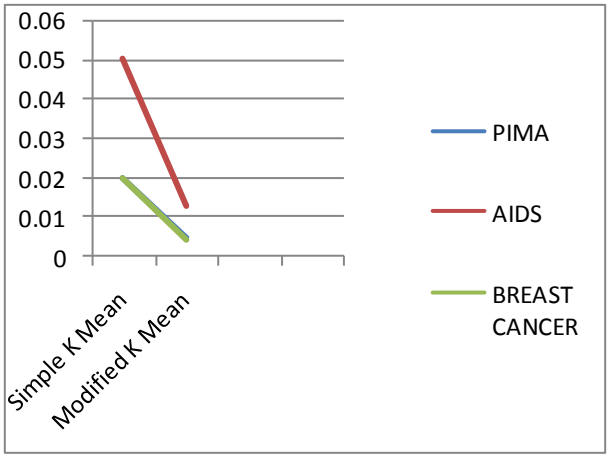

Fig 1 Comparison according to time taken

3.2 Comparison according to Instances in Cluster 1

Instances in cluster 1 are formed by Modified K Mean clustering are 186, 383 and 152 for the data sets Aids, Breast Cancer and Pima. Instances in cluster 1 are formed by Simple K Mean clustering are 112, 546 and 107 for the data sets Aids, Breast Cancer and Pima. So we conclude that Modified K Mean Clustering forms instances in cluster 1 with more accuracy than simple k means clustering.

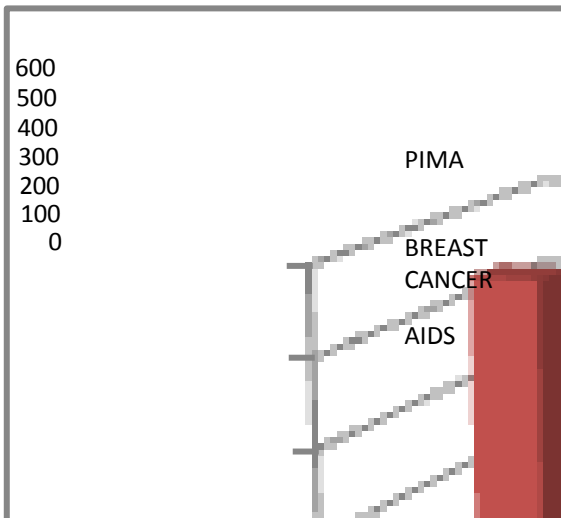

Fig 1 Comparison according to instances in cluster 1

3.3 Comparison according to Instances in cluster 2

Fig 3 shows that Instances in cluster 2 are formed by Modified K Mean clustering are 109, 316 and 180 for the data sets Aids, Breast Cancer and Pima. Instances in cluster 2 are formed by Simple K Mean clustering are 182,152 and 224 for the data sets Aids, Breast Cancer and Pima.

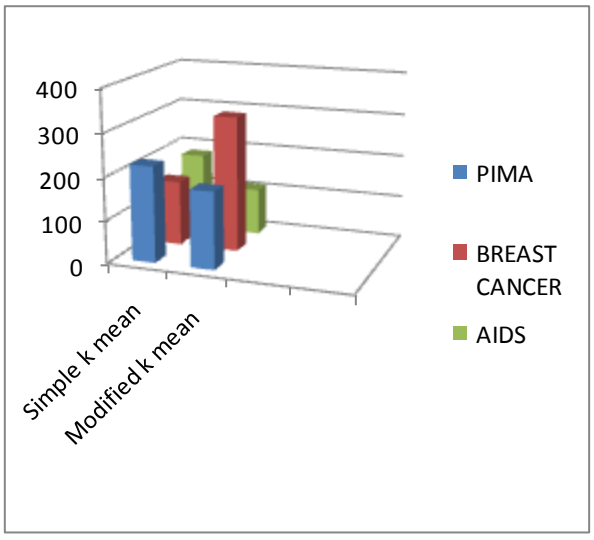

Fig 3 Comparison according to instance in cluster 2

3.4 Comparison according to Sum of squared errors

Sum of squared errors for modified k mean clustering for data sets Pima, Aids and Breast cancer is 0.0004, 0.006 and 0.02 respectively and for simple k mean clustering for data sets Pima, Aids and Breast cancer is 12.11, 10.64 and 22.67 respectively From the above table and figure, we conclude that Modified k mean method provides less Sum of squared errors to complete the process than Simple k mean method.

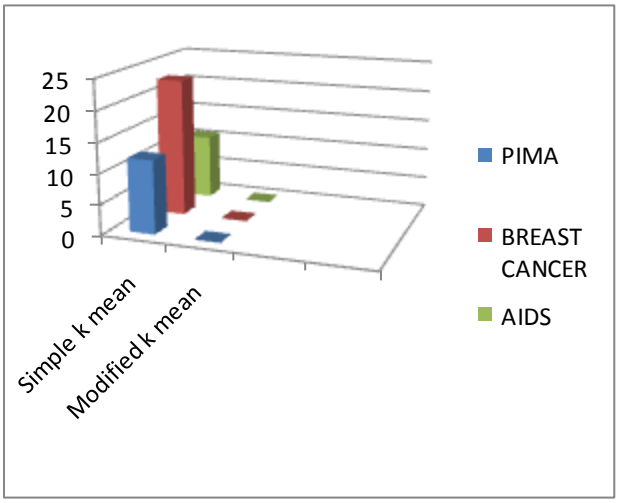

Fig 2 Comparison according to sum of squared errors

3.5 Comparison according to No. of Iterations

No. of Iterations for modified k mean clustering for data sets Pima, Aids and Breast cancer is 1, 0 and 0 respectively and for simple k mean clustering for data sets Pima, Aids and Breast cancer is 10, 6 and 3 respectively.

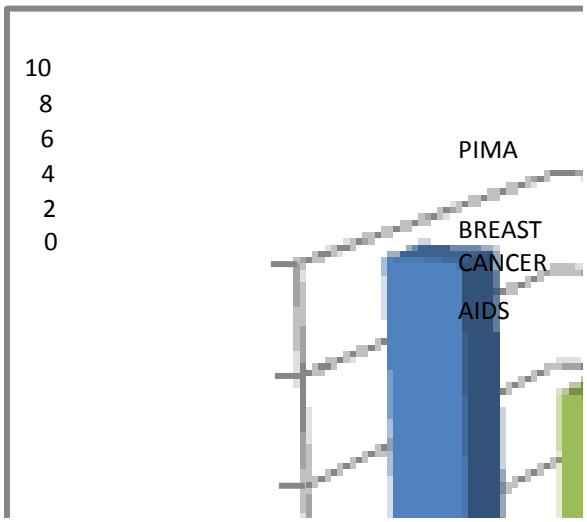

Fig 3 Comparison according to no. of iterations

IV. CONCLUSION

The clustering involves partitioning a given dataset into some groups of data whose members are similar in some way. The usability of cluster analysis has been used widely in data recovery, text and web mining, pattern recognition, image segmentation and software reverse engineering. Simple k mean algorithm using WEKA tool has the problem of time complexity and complex in calculating the Sum of squared errors, no. of iterations. These complexities have been resolved by implementing the modified k mean algorithm in C#.NET. This algorithm uses the concept of indexed normalization technique to reduce the time taken to build the clusters, no. of iterations, Sum of squared errors etc. All the comparisons have been shown through the graphs in this paper

which show that Modified k mean clustering gives the better result. Limitation of Modified k mean clustering algorithm on is that it works only numeric values. In future we can add the feature of implementing the text values on this algorithm.

V. ACKNOWLEDGEMENT

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

REFERENCES

1. Madhu Yedla, Srinivasa Rao Pathakota, T M Srinivasa, Enhancing K-means Clustering Algorithm with Improved Initial Centre, International Journal of Computer Science and Information Technologies, Volume 1, Issue 2 , 2010,
2. Arpit Gupta, Ankit Gupta, Amit Mishra, Research Paper on Cluster Techniques of Data Variations, International Journal of Advance Technology & Engineering Research, Volume 1, Issue 1, November 2011.
3. Sudesh Kumar, Nancy, Efficient K-Mean Clustering Algorithm for Large Datasets using Data Mining Standard Score Normalization, International Journal on Recent and Innovation Trends in Computing and Communication, Volume 2, Issue 10, October 2014.
4. Rishi Syal, Dr V.Vijaya Kumar, Innovative Modified K-Mode Clustering Algorithm, International Journal of Engineering Research and Applications, Volume 2, Issue 4, July-August 2012.
5. Taeho Jo, Inverted Index based Modified Version of K-Means Algorithm for Text Clustering, Journal of Information Processing Systems, Volume 4, Issue 2, June 2008.
6. B. F. Momin, P. M. Yelmar, Modifications in K-Means Clustering Algorithm, International Journal of Soft Computing and Engineering, Volume 2, Issue 3, July 2012.
7. Sonal Miglani, Kanwal Garg, Factors Affecting Efficiency of K-means Algorithm, International Journal of Advancements in Research & Technology, Volume 2, Issue5, May-2013.
8. Vaishali R. Patel1 and Rupa G. Mehta, Impact of Outlier Removal and Normalization Approach in Modified k-Means Clustering Algorithm, International Journal of Computer Science Issues, Vol. 8, Issue 5, No 2, September 2011.
9. N. Karthikeyani Visalakshi and K. Thangavel, Distributed Data Clustering: A Comparative Analysis, Foundations of Computational Intelligence (6)", 2009.