K-Mean Evaluation in Weka Tool and Modifying It using Standard Score Method

Sudesh kumar
Computer science and engineering
BRCM CET Bahal Bhiwani (Haryana)
Bhiwani, India
ksudesh@brcm.edu

Nancy
Computer science and engineering
BRCM CET Bahal Bhiwani (Haryana)
Bhiwani, India
nancypubreja09@gmail.com

***Abstract*—** This paper introduces the concept of Data mining with K-Mean Clustering. The Data mining is the main object that finds the useful patterns among the large amount of data. Today most of the work is done on Internet. So, mining of data becomes essential thing for easy searching of data. Mining of data is done using multiple clustering techniques. Cluster Analysis seeks to identify homogeneous groups of objects based on the values of their attributes. The performance of clustering depends upon centroids selection and frequency of nearest data. The existing K-Means technique is complex in terms of time and calculation. This paper proposed the modified approach of K-Means clustering and algorithm has been designed. The entire data will be normalized using standard score method which is also called z score and then cluster will be formed using Euclidean distance. The fast clustering process will reduce the system resources and provides the efficient technique to generate the clusters.

.

Keywords- *Modified K-Mean , Data Mining, Clustering, Centroids, standard score*

*****

I. INTRODUCTION

Data mining technology is used to give the user an ability to extract meaningful patterns from large database. Information retrieval systems have made large quantities of textual data available. Extraction of meaningful patterns from this data is very difficult. Current tools for mining structured data are inappropriate for free text. The problems involved in this are knowledge discovery in text and present architecture for extracting patterns that hold across multiple documents. Data mining technology has created a new opportunity for exploiting the information from the databases. Patterns in the data, such as associations among similar items purchases, enables target marketing to focus on what things the customers are likely to purchase. Data mining, the extraction of hidden predictive information from large databases, is a powerful, new technology with great potential to help many companies to focus on the most important information in the information storehouse. Data mining tools calculate upcoming trends and activities, allowing company to make practical, knowledge-driven decisions.

Cluster group study or clustering is the task of collection a set of objects in such a way that items in the same group known as cluster and extra similar to each other than to those in other clusters groups. Data modeling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis. Clustering is a division of data into groups of similar objects. Representing the data by the fewer clusters necessarily loses certain fine details, but achieves simplification. As a machine learning perspective clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represent a data concept. It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics

K-mean algorithm**:** It accepts the number of clusters to group data into, and the dataset to cluster as input values. Then it creates the first K initial clusters and K means the no. of clusters required from the dataset by selecting K rows of records arbitrarily from the dataset. Clustering Algorithm K-Mean compute the Mean of the every group or Cluster formed in the dataset. The mathematics Mean of a group or cluster is the mean of all the individual records in the cluster. In each of the initial K groups or clusters, there will be only one record. The Mean of a cluster with single record is the set of values that make up that record. The K-Means Clustering Algorithm allocates every record in the dataset to the one record set or clusters. Every record is allocated to the nearby group using calculation of distance or correspondence like the Euclidean Distance calculation. The algorithm K-Means re-assigns every record in the dataset to the majority related group and re-calculates the mathematics mean of the entire group in the dataset. The mathematics mean of a group is the sums mean of all the records in that group. The Algorithm re-allocates every record in the dataset to only one of the new group created. A data point or record is allocated to the nearby group by means of a compute of distance or similarity. The earlier steps will be repetitive until constant groups are created and the K-Means clustering process is concluded. Constant clusters are shaped after new iterations or repetitions of the K-Means clustering algorithm and it does not create fresh group as the Arithmetic Mean or cluster center of each group formed is the similar as the older cluster center [9].

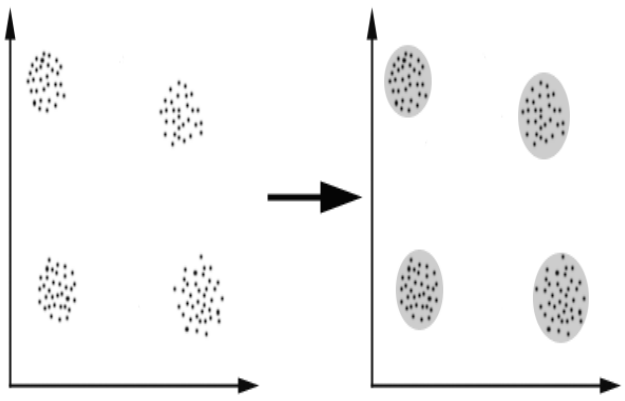


Fig 1: Clustering Concept

II. LITERATURE REVIEW

The author has proposed an efficient, modified K-mean clustering algorithm to cluster large data-sets whose objective is to find out the cluster centers which are very close to the final solution for each iterative steps. Clustering is often done as a prelude to some other form of data mining or modeling. Performance of iterative clustering algorithms depends highly on the choice of cluster centers in each step. The algorithm in this paper is based on the optimization formulation of the problem and a novel iterative method. The cluster centers computed using this methodology are found to be very close to the desired cluster centers. The experimental results using the proposed algorithm with a group of randomly constructed data sets are very promising. The best algorithm in each category was found out based on the performance [1]

The author has been assimilated the Knowledge about 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 they described the way in which high dimensional data is different from low dimensional data, and how these differences might affect the process of cluster analysis and also described several recent approaches to clustering high dimensional data, including our own work on concept-based clustering. All of these approaches have been successfully applied in a number of areas, although there is a need for more extensive study to compare these different techniques and better understand their strengths and limitations. Cluster analysis divides data into groups (clusters) for the purposes of summarization or improved understanding. For example, cluster analysis has been used to group related documents for browsing, to find genes and proteins that have similar functionality, or as a means of data compression [2].

The author explained the knowledge about Weka tool and demonstrated with taking example of clustering method and introduce WEKA workbench, reviews the history of the project and k-means clustering execution in WEKA 3.7. There are number of clustering techniques from which the K-Means clustering is explained by the author and working of Weka. WEKA's support for clustering tasks is as extensive as its support for classification and regression and it has more techniques for clustering than for association rule mining, which has up to this point been somewhat neglected. WEKA support various clustering algorithms execution in Java which gives a platform for data mining research process. Releasing WEKA as open source software and implementing it in Java has played no small part in its success. This paper provides a comprehensive review of K-means clustering techniques in WEKA 3.7.More than twelve years have elapsed since the first public release of WEKA. In that time, the software has been rewritten entirely from scratch, evolved downloaded more than 1.4 million times since being placed on Source-Forge in April 2000[3].

The author has been assimilated the knowledge about clustering, techniques and perform comparative study on the techniques. Author also explained the clustering; Clustering is a process of putting similar data into groups. Clustering can be considered the most important unsupervised learning technique so as every other problem of this kind; it deals with finding a structure in a collection of unlabeled data. This paper reviews six types of clustering techniques- k-Means Clustering, Hierarchical Clustering, DBSCAN clustering, OPTICS, STING. K-mean algorithm has biggest advantage of clustering large data sets and its performance increases as number of clusters increases. But its use is limited to numeric values. Therefore Agglomerative and Divisive Hierarchical algorithm was adopted for categorical data, but due to 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, author concluded the performance of K- mean algorithm which is better than Hierarchical Clustering Algorithm [4].

In this paper, Author has been explained the concept of K-Means clustering, its advantages and disadvantages. Author has also explained the biggest advantage of the k-means algorithm in data mining applications and its efficiency in clustering large data sets. However, its use is limited to numeric values. Due to filtering capacity of K-mean, this algorithm is only used in case of numeric data sets. The Agglomerative and Divisive Hierarchical Clustering algorithm was adopted the dataset of categorical nature initially. Due to complexity in both of the above algorithm, this paper has presented a new approach to assign rank value to each categorical attribute for K-mean Clustering. The categorical data have been converted into numeric by assigning rank value. It is a categorical dataset can be made clustering as numeric datasets. It is observed that implementation of this logic, k- mean yield same performance as used in numeric datasets [5].

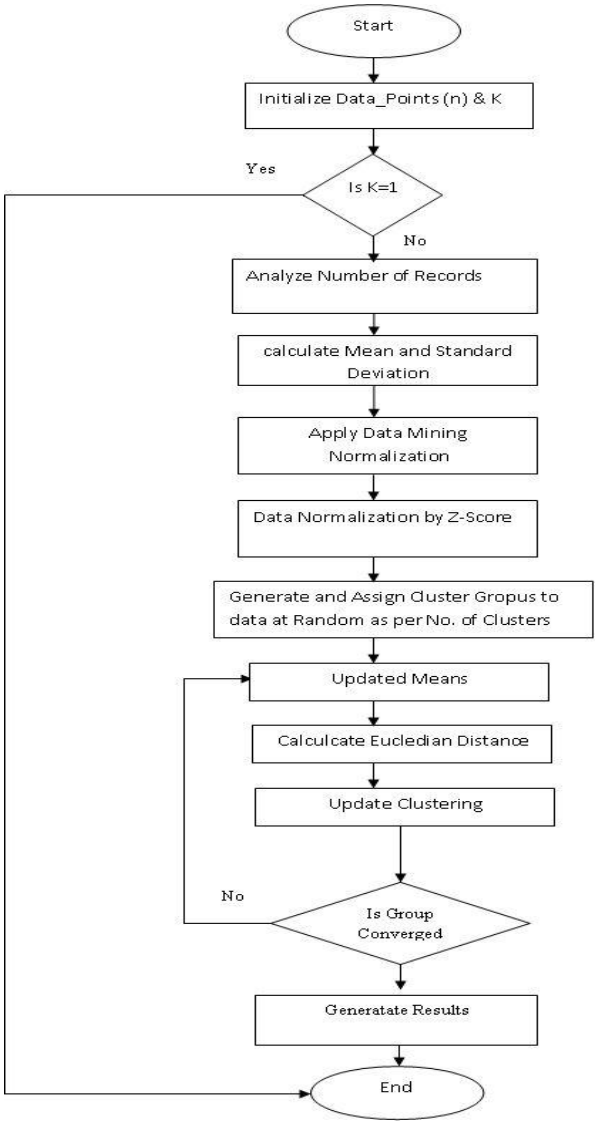
In this paper, Author has explained the concept of clustering in data mining and techniques. Author has been studied the six techniques introduced previously, and testing each one of them using Weka Clustering Tool on a set of banking data related to customer information. The whole data set consists of 11 attributes and 600 entries. Clustering of the data set is done with each of the clustering algorithm using Weka tool. Author analyzed the results of testing the algorithms and running them under different factors and situations [6].

In this paper, Author has given the introduction about K-means clustering and its algorithm. The experimental result of K-means clustering and its performance in case of execution time is discussed here. But there are certain limitations in K-means clustering algorithm such as it takes more time for execution. So in order to reduce the execution, time we are using the Ranking Method. And also shown that how clustering is performed in less execution time as compared to the traditional method. This work makes an attempt at studying the feasibility of K-means clustering algorithm in data mining using the Ranking Method. Modifications in hard K-means algorithm such that algorithm can be used for clustering data with categorical attributes. To use the algorithm for categorical data modifications in distance and prototype calculation are proposed. To use the algorithm on numerical attribute values, means is calculated to represent centre, and Euclidean distance is used to calculate distance [7]

In this paper, less similarity based clustering method is proposed for finding the better initial centroids and to provide an efficient way of assigning the data points to suitable clusters with reduced time complexity. They mainly classified into three main categories: text-based, link-based and hybrid. This method also reduces time complexity. In this less similarity based clustering method the initial cluster centers will not be selected randomly so accuracy will be high. The experimental results show that proposed algorithm provides better results for various datasets .The value of k; desired number of clusters is still required to be given as an input to the proposed algorithm [8].

III. PROPOSED METHODOLOGY

The existing Clustering algorithm will be analyzed through WEKA Tool and result will be generated. The duplicate, irrelevant information will be cleaned in relational database.
Flow Chart:



The result will be analyzed by WEKA tool with backend relational database. WEKA Explorer is an application that provides the functionality of Dataset Management, loading data, feeding them to classifiers, filters, storing the results of classification, apportioning data between training and testing subsets.

1. Study of K-Mean Algorithm.
2. Collection of Data.
3. Analysis of K-Mean algorithm using WEKA Explorer Tool.
4. Analyze Clusters with WEKA Tool.
5. Study K-Mean Advantages and Disadvantages.
6. Design a new Approach algorithm.
7. Implement new Algorithm in VC# and Result Analysis using WEKA and Proposed Algorithm.

The flow chart has been designed for this approach [9].

IV. OBJECTIVE

There are some problems in existing system that can degrade the efficiency of clustering described as follows:

1. The availability of outliers contained in the data collection which means no belonging group.
2. Observations out of the scope in Clusters.
3. Complex calculations for generate centroids.
4. Difficult to work on large datasets because of time consumption.
5. Discover relevant knowledge in data Collections.

Data mining may be viewed as the extraction of patterns and models from observed data or a method used for analytical process designed to explore data. There are many different methods, which may be used to predict the appropriate class for the objects. The majority of data mining techniques can deal with different data types. There are number of techniques and many variations of the methods, one of the techniques from the mentioned group is almost always used in real world deployments of data mining systems. But existing techniques are complex in terms of calculation. For this, there is needed to be design a new algorithm.

1. To Study of existing data analysis clustering technique.
2. To analyze complexity and outlier issue in Algorithm.
3. To find point of complexity in Algorithm.
4. To Develop the Clustering Algorithm for large and small datasets.
5. To Study the proposed algorithm and its advantage.
6. To Implement the algorithm and perform analysis.
7. Generate Results [9].

V. CONCLUSION AND FUTURE WORK

In this paper, the Modified K-Mean algorithm has been proposed for generate the clusters with less complexity and in less time. The modification has been done as internal level that means the complete data has been standardized using Z score method and then Euclidean distance is calculated .This method will take less time compared to original k mean algorithm. This will be implemented in future using VC#

REFERENCES

[1] Anwiti Jain, Anand Rajava,t Rupali Bhartiya(2012) "Design Analysis and Implementation of Modified KMean Algorithm for Large Data-set to Increase Scalability and Efficiency ",IEEE

[2] Er. Arpit Gupta ,Er.Ankit Gupta (2011) "Research Paper on Cluster Techniques of Data Variations", IJATER.

[3] Sapna Jain, M Afshar Aalam(2010),"K-Means Clustering Using Weka Interface", Proceedings of the 4th National Conference; INDIACom..

[4] Aastha Joshi, Rajneet Kaur (2013),"A Review: Comparative Study of Various Clustering Techniques in Data Mining", IJARCSSE. Volume 3, Issue 3.

[5] Sovan Kumar Patnaik, Soumya Sahoo(2012), "Clustering of Categorical Data by Assigning Rank through Statistical Approach", International Journal of Computer Applications (0975 – 8887)Volume 43– No.2, April 2012.

[6] Manish Verma, Mauly Srivastava (2012),"A Comparative Study of Various Clustering Algorithms in Data Mining", International Journal of Engineering Research and Applications (IJERA), Vol. 2, Issue 3, pp.1379-1384.

[7] G. Singh and N. Kaur(2013), "Hybrid Clustering Algorithm with Modified Enhanced K-Mean and Hierarchical Clustering", International Journal of Advanced Research in Computer Science and Software Engineering 2013.

[8] Kaur , N. Kaur (2013), "Web Document Clustering Approaches Using K-Means Algorithm ", International Journal of Advanced Research in Computer Science and Software Engineering 2013