IMPROVEMENT IN K-MEANS CLUSTERING USING VARIANT TECHNIQUES

[1]Rawinder Kaur,[2]Prabhjot Kaur
[1]Student,[2]Assistant Professor
[1]Department of Computer Science Engineering
[1]Patiala Institute Of Engineering & Technology,Patiala,Punjab

Abstract: This work presents an overview of the K-means clustering algorithm & various enhanced variations done on K-means clustering algorithm. K-means is the basic algorithm used for discovering clusters within a dataset. The initial point selection effects on the results of the algorithm, both in the number of clusters found and their centroids. Methods to enhance the k-means clustering algorithm are discussed. With the help of these methods efficiency, accuracy, performance and computational time is improved. Some enhanced variations improves the efficiency and accuracy of algorithm. Basically in all the methods the main aim is to reduce the number of iterations which will decrease the computational time. Studies shows that K-means algorithm in clustering is widely used technique. Various enhancements done on K-mean are collected, so by using these enhancements one can build a new hybrid algorithm which will be more efficient, accurate and less time consuming than the previous work.

Keywords- Min-Max, K-mean, Clustering

I. INTRODUCTION

Data mining can be viewed as a result of the natural evolution of information technology. The database and data management industry evolved in the development of several critical functionalities: data collection and database creation, data management (including data storage and retrieval and database transaction processing), and advanced data analysis (involving data warehousing and data mining). The early development of data collection and database creation mechanisms served as a prerequisite for the later development of effective mechanisms for data storage and retrieval, as well as query and transaction processing. Emerging data repository architecture is the data warehouse. This is a repository of multiple heterogeneous data sources organized under a unified schema at a single site to facilitate management decision making [1]. Data warehouse technology includes data cleaning, data integration, and online analytical processing (OLAP)-that is, analysis techniques with functionalities such as summarization, consolidation, and aggregation, as well as the ability to view information from different angles. The effective and efficient analysis of data from such different forms of data by integration of information retrieval, data mining, and information network analysis technologies is a challenging task [2]. This section explains the various steps in knowledge discovery from data or KDD process. The term Knowledge Discovery from data or KDD, refers to the broad process of finding knowledge in data, and emphasizes the "high-level" application of particular data mining methods. The role of data mining (KDD) is very important in many of the field such as the analysis of market basket, classification, etc. if talk about data mining, the most important role presented by frequent item set which is used to find out the correlation between the various type of the field that is display in the database . Another name of the data mining is KDD (Knowledge discovery from the database).discovery of frequent item set is done by association rule. Retail store also used the concept of association rule for managing marketing, advertising, and errors that are presented in the telecommunication network. Data mining is an interdisciplinary subfield of computer science which is the process of discovering insightful, interesting and novel patterns from large-scale data sets. Data mining [3] is an essential step of the "Knowledge Discovery in Databases" process, or KDD. Data mining is often treated as synonym for another popularly used term, Knowledge Discovery in Databases (KDD). Data clustering (or just clustering), is an unsupervised classification method aims at creating groups of objects, or clusters, in such a way that objects in the same cluster are very similar and objects in different clusters are quite distinct. Cluster analysis is one of the traditional topics in the data mining field. It is the first step in the direction of exciting knowledge discovery. Clustering is the procedure of grouping data objects into a set of disjoint classes, called clusters [4]. Now objects within a class have high resemblance to each other in the meantime objects in separate classes are more unlike. The k-means clustering algorithm is the basic algorithm which is based on partitioning method which is used for many clustering tasks especially with low dimension datasets. It uses k as a parameter, divide n objects into k clusters so that the objects in the same cluster are similar to each other but dissimilar to other objects in other clusters. The algorithm attempts to find the cluster centers, (*C1 …… Ck*), such that the sum of the squared distances of each data point, *xi* , $1 \le i \le n$, *to* its nearest cluster center *Cj*, $1 \le j \le k$, is minimized. First, the algorithm randomly selects the k objects, each of which initially represents a cluster mean or center. Then, each object *xi* in the data set is assigned to the nearest cluster center i.e. to the most similar center. The algorithm then computes the new mean for each cluster and reassigns each object to the nearest new center [5]. This process iterates until no changes occur to the assignment of objects. The convergence results in minimizing the sum-of-squares error that is defined as the summation of the squared distances from each object to its cluster center. KNN Approach is an enhancement of K-mean clustering. It is based upon normalization.KNN is a non parametric lazy learning algorithm. It is very easy to understand but hard to implement. Non-parametric statement means that it does not make any assumptions on the underlying data distribution. Most of the algorithm doesn't obey theoretical assumptions. It is also a lazy algorithm that does not use the training data points to do any generalization. It does not discard non support vectors like SVM [6]. It makes decision on the basis of entire training data set. It has minimal training phase but a costly testing phase. Cost is in terms of memory and time. It requires more time to access all the data training sets. It also requires more memory to store all the data.

II. LITERATURE REVIEW

Neha Aggarwal, et.al (2012) proposed [7] K-Means Clustering is an immensely popular clustering algorithm for data mining which partitions data into different clusters on the basis of similarity between the data points and aims at maximizing the intra-class similarity and minimizing the inter-class similarity. This paper shows the comparison of Basic K-Means and Enhanced K-Means algorithm which shows that Enhanced K- Means is more efficient than Basic K-Means Algorithm. K-Means This Algorithm suffers from the limitation of being time consuming and producing different results with different centroids chosen randomly. The first limitation is solved using the Enhanced K-Means algorithm. This paper shows the comparison of Basic K-Means and Enhanced K-Means algorithm which shows that Enhanced K-Means is more efficient than Basic K-Means Algorithm.

Ahamed Shafeeq B M et.al (2012) proposed [8] K-means is a widely used partitional clustering method. While there are considerable research efforts to characterize the key features of K-means clustering, further investigation is needed to reveal whether the optimal number of clusters can be found on the run based on the cluster quality measure. This paper presents a modified Kmeans algorithm with the intension of improving cluster qualityand to fix the optimal number of cluster. The K-means algorithm takes number of clusters (K) as input from the user. It is shown that how the modified k-mean algorithm will increase the quality of clusters compared to the K-means algorithm. It assigns the data point to their appropriate class or cluster more effectively.

Manpreet Kaur et.al (2013) introduced [9] Query redirection provides a mechanism for BI Server to determine the set of logical table sources (LTS) applicable to a logical request whenever a request can be satisfied by more than one LTS. The Oracle BI repository shipped in Oracle Fusion applications contains metadata content for real-time reporting analysis (using Transactional Business Intelligence) and historical reporting (using BI Applications).The proposed work represents query redirection method that improved K-means clustering algorithm performance and accuracy in distributed environment. In this paper we have done analysis on k-mean and hierarchical algorithm by applying validation measures like entropy, f-measure, coefficient of variance and time. The experimental results show that k-mean algorithm performs better as compared to hierarchical algorithm and takes less time for execution.

Amar Singh et.al (2013) proposed [10] work represents ranking based method that improved K-means clustering algorithm performance and accuracy. In this we have also done analysis of K-means clustering algorithm, one is the existing K-means clustering approach which is incorporated with some threshold value and second one is ranking method which is weighted page ranking applied on K-means algorithm, in weighted page rank algorithm mainly in links and out links are used and also compared the performance in terms of execution time of clustering. Proposed ranking based K-means algorithm produces better results than that of the existing k-means algorithm.

Vijay Jumb et.al, (2014) proposed an approach for color image segmentation. In this method foreground objects are distinguished clearly from the background [11]. As the HSV color space is similar to the way human eyes perceive color, hence in this method, first RGB image is converted to HSV (Hue, Saturation, Value) color model and V (Value) channel is extracted, as Value corresponds directly to the concept of intensity/brightness in the color basics section. The result of Otsu's multi-thresholding may consist of over segmented regions, hence K-means clustering is applied to merge the over segmented regions. The proposed method is compared with three different types of segmentation algorithms that ensure accuracy and quality of different types of color images. The experimental results are obtained using metrics such as PSNR and MSE, which proves the proposed algorithm, produces better results as compared to other algorithms.

III. RESERCH METHODOLOGY

K-Means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. Despite being used in a wide array of applications, the k-means algorithm is not exempt of drawbacks, mainly: - As many clustering methods, the k-means algorithm assumes that the number of clusters k in the database is known beforehand which, obviously, is not necessarily true in real-world applications.

- As an iterative technique, the k-means algorithm is especially sensitive to initial centres selection.
- The k-means algorithm may converge to local minima.

K-Means algorithm has many disadvantages so in this work a new hybrid K-Means algorithm will be implemented to solve the problem of efficiency and accuracy. Accuracy and efficiency are related to each other in other words there is a trade off between these two. So a proper method is needed that will balance between both the accuracy and efficiency of the k-means clustering algorithm. By doing so clustering quality improves.

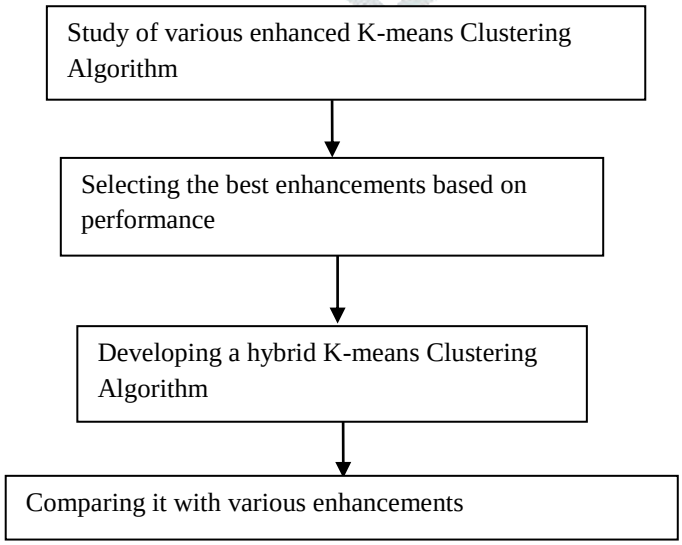

Figure. 1: Design of Proposed Work

IV. EXPERIMENTAL RESULTS

As shown in figure 2, the data set which is used for clustering is been clustered and each cluster will be marked with different colors. In this figure, various iterations run, means at every iteration new centered point is selected and on the basis of that centered point, cluster assignment procedure will be done.

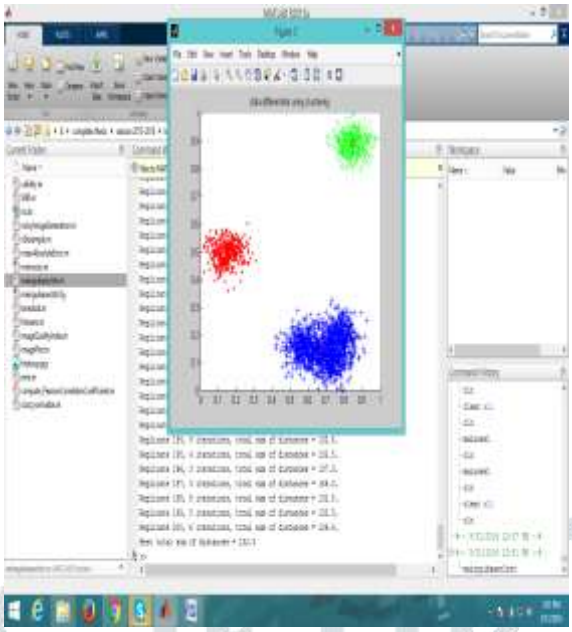

Figure 2: Clustering of Data

As shown in figure 3, the dataset which is used in the previous figure will be clustered using the hybrid type of k-mean clustering algorithm. When the dataset will be clustered using hybrid algorithm cluster quality will be improved and each point in the dataset will be shown on voronlie plane for better analysis of dataset.

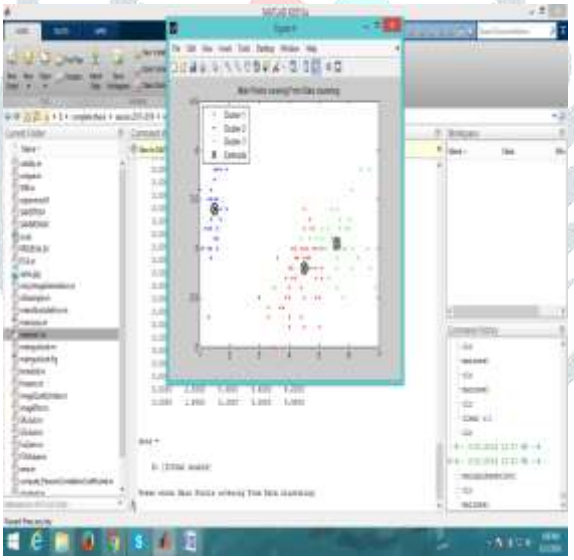

Figure 3: Vornoalie Representation

As illustrated in figure 4, the final clustering result is shown on the 2-D plane. The data which is read from the excel file. The data which is in the excel file are given as input to k-mean clustering and that final clusters are generated on the basis of Euclidian Distance.

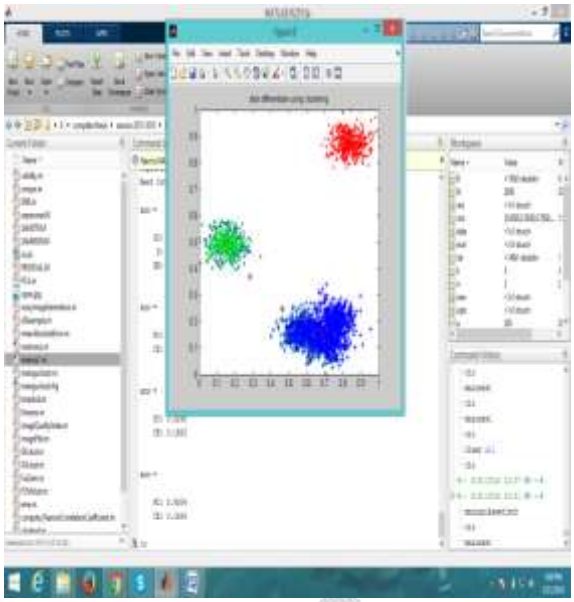

Figure 4: Clustering of Dataset

V. CONCLUSION

The clustering is the technique which can cluster similar and dissimilar type of data. The k-mean clustering is the partitioned based clustering algorithm in which central point is calculated and from the point Euclidian distance is calculated to all other points in the dataset. The data points which have similar type of Euclidian distance is clustered in one cluster and other in the second cluster. In this research work, k-mean clustering is improved using the back propagation algorithm to increase accuracy of clustering. In the technique of back propagation system learns from the previous experience and drive new values. The simulation results show that proposed technique performs well in terms of accuracy and execution.

REFERENCES

[1] Amar Singh and Navot Kaur, "To Improve the Convergence Rate of K-Means Clustering Over K-Means with Weighted Page Rank Algorithm," International journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 8, August 2012.

[2] Harpreet Kaur and Jaspreet Kaur Sahiwal, "Image Compression with Improved K-Means Algorithm for Performance Enhancement," International Journal of Computer Science and Management Research, Volume 2, Issue 6, June 2013.

[3] Osamor VC, Adebiyi EF, Oyelade JO and Doumbia S "Reducing the Time Requirement of K-Means Algorithm" PLoS ONE, Volume 7, Issue 12, 2012.

[4] Azhar Rauf, Sheeba, Saeed Mahfooz, Shah Khusro and Huma Javed, "Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity," Middle-East Journal of Scientific Research, pages 959-963, 2012.

[5] Kajal C. Agrawal and Meghana Nagori, "Clusters of Ayurvedic Medicines Using Improved K-means Algorithm," International Conf. on Advances in Computer Science and Electronics Engineering, 2013.

[6] Chieh-Yuan Tsai and Chuang-Cheng Chiu, " Developing a feature weight self-adjustment mechanism for a K-means clustering Analysis, pages 4658-4672, Volume 52, 2008.

[7] Neha Aggarwal, Kirti Aggarwal and Kanika Gupta, "Comparative Analysis of k-means and Enhanced K-means clustering algorithm for data mining," International Journal of Scientific & Engineering Research, Volume 3, Issue 3, August-2012.

[8] Ahamed Shafeeq B M and Hareesha K S, "Dynamic Clustering of Data with Modified K- Means Algorithm," International Conference on Information and Computer Networks, Volume 27, 2012.

[9] Manpreet Kaur and Usvir Kaur, "Comparison Between K-Mean and Hierarchical Algorithm Using Query Redirection", International Journal of Advanced Research in Computer Science and Social , Volume 3, Issue 7, July 2013 ISSN: 2277 128X

[10] Amar Singh and Navot Kaur, "To Improve the Convergence Rate of K-Means Clustering Over K-Means with Weighted Page Rank Algorithm," International journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 8, August 2012.

[11] Vijay Jumb Mandar Sohani Avinash Shrivas, "Color Image Segmentation Using K-Means Clustering and Otsu's Adaptive Thresholding, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-3, Issue-9, February 2014