Improvement in Symmetric Hybrid K-Mean Clustering For the Prediction Analysis Using Normalization Techniques

1Er. Shilpy (Author), 2Er Rashim Rana
1,2Department of Computer Science and Engineering
Maharishi Ved Vyas Engineering College – Jagadhri

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

Keywords: Clustering, K-Mean Clustering, Euclidian distance, central point

I. INTRODUCTION

Data mining is the technology of discovering interesting patterns from large amount of data. It is extraction of implicit previously unknown and potentially useful information from data. Data mining is also called as extraction of hidden patterns. It also known as knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology and data dredging. It may fully automate or semi-automated process to discover knowledge that is useful for user. [1]

 ![Fig 1.1 KDD Process in data mining](image)

It also a process of finding a hidden information in data base this process may use one or more computer learning techniques to automatically analyses and extract knowledge from data contain within the database, it is part of knowledge discovery process.

Data mining applies algorithms to large data to produce models or patterns interesting to the user and will extract the hidden patterns.
1.1 Clustering in data mining: Clustering is an unsupervised learning technique. It is a process of partitioning a set of data into a set of significant sub-classes, called cluster [1]. Data is organized into clusters such that there is high intra-cluster similarity and low inter-cluster similarity. It is a main task of exploratory data mining and a common technique for statistical data analysis [3]. It is implemented in many fields including machine learning, pattern recognition, image analysis, information retrieval and bioinformatics.

There are several application of clustering [9] such as data/text mining, image processing, web mining, voice mining. There are several methods of clustering [10]. The major primary clustering methods can be classified into following categories [11]:

Partitioning Methods: The general criterion for partitioning is a combination of high similarity of the samples inside of clusters with high dissimilarity between separate clusters. Most partitioning methods are distance-based.

Hierarchical Methods: In this method hierarchical breakdown of the given set of data objects is created. It can be classified into two approaches agglomerative and divisive. Agglomerative approach is the bottom up approach. This approach starts with each object forming a separate group [7]. Divisive approach is top down approach starts with all the clusters in the same cluster and then each repetitive step a cluster is split into smaller cluster untilled each object is in one cluster.

Density Based Methods: Generally partitioning methods cluster objects based on distance between objects. Spherical shaped clusters can be discovered by these methods and come across difficulty in discovering clusters of random or arbitrary shapes. So for arbitrary shapes new methods are used known as density-based methods which are based on the notion of density. In these methods the cluster is continue to produce as long as the density in the neighborhood cross some threshold [6].

Grid Based Methods: Grid based methods quantize the object space into a fixed number of cells that form a grid structure. It is a fast method and is independent of the number of data objects and depends only on the number of cells in each dimension in the quantized space [8].

II. REVIEW OF LITERATURE

In Paper [1], proposed non metric distance measure which is based on live symmetry is used to measure cluster
soundness. For this thresholding technique is applied first to extracts object from the original image the object pixels are transferred to be the data patterns. Object pixels are labeled by applying the fuzzy clustering algorithm and number of objects are determining by applying proposed validity measure. To define performance of proposed measure simulation results are used.

In Paper [2] they described comparison between various clustering techniques like partitioning method hierarchical method, density based method, grid method. Clustering algorithm are mainly used to manage data, categorized data for data compression, model creation and also used for outlier discovery etc. Main motive each clustering technique is to find cluster center that represent each cluster. Then input data is compared with each cluster center, and then based on these cluster centers defined which cluster is nearest or similar one. Partitioning method like k-mean clustering algorithm is used for large datasets, as number clusters is increased its performance is also increased. But its use is limited to numeric values. Hierarchical algorithms are used for categorical data. DBSCAN is adopted to find cluster of arbitrary shapes.

In Paper [3] provides new method to improve accuracy and performance k-mean clustering that is a ranking based method. Analysis done on existing k-mean clustering approach which is fit in with some threshold value and ranking method which is weighted page ranking applied on k-mean algorithm. In this in links and out links are used to compare performance in the form of execution time of clustering. Weighted page rank algorithm with k-mean provides better result than existing k-mean algorithm. It takes less computational time then existing k-mean algorithm.

In Paper [4] they provide a new method to improve accuracy and performance k-mean clustering that is a ranking based method. Analysis done on existing k-mean clustering approach which is fit in with some threshold value and ranking method which is weighted page ranking applied on k-mean algorithm. In this in links and out links are used to compare performance in the form of execution time of clustering. Weighted page rank algorithm with k-mean provides better result than existing k-mean. It takes less computational time then existing k-mean algorithm [3].

In Paper [5] they implemented K-means along with Genetic algorithm for dimensionality reduction and support vector machine to classify the data set. K-means algorithm is used to remove outliers and the noisy data. The optimal features are selected by using the genetic algorithm and then Support vector machine classifies the reduced data space using 10 fold cross validation technique. Genetic algorithm selects different features from original set of feature during each run. To obtain consistent results, the experiment was performed 50 times. The result shows that the purposed model achieves the accuracy of 98.82%.

In Paper [6] proposed a prediction model for medical data with missing value imputation techniques, then analyzing these techniques by using K-means algorithm and choosing the best among them. Thus this model improves the quality of data by using the best imputation technique. Methods such as case deletion, most common method, concept most common, K-means clustering imputation, k-nearest neighbor etc. are applied to fill the missing data values in the data. The efficiency is calculated on three data sets namely Hepatitis, Wisconsin Breast Cancer and Pima Indians Diabetes from the UCI repository. This model achieved accuracy of 99.82% for Diabetes data set, 99.39% for Breast Cancer and 99.08% for Hepatitis data set. For Diabetes and Hepatitis data sets Concept Most Common (CMC) is chosen as the best method, and for Breast Cancer Case deletion is selected as best missing value imputation method.

In Paper [7] they proposed an optimized version of k-mean that reduces the problem of re-distribution of the data elements that will remain part of the same cluster during the next iteration. After a number of iterations only a few number of data elements change their cluster. While assigning the data element to the cluster there is no need to visit the entire data set, but just a small list of data objects. The implementation showed up to 70% reduction of the running time.

In paper [8] algorithms like genetic algorithm, PSO, ANN that can be used in predicting heart disease. Combining these algorithms with the data mining techniques such as clustering, classification etc. or by combining these algorithms with one another will give better performance and accuracy.
III. K-MEAN CLUSTERING

The K-mean clustering algorithm is the indispensable algorithm based on partitioning method which is used for multiple clustering chores (jobs) especially with low dimension datasets. K-mean clustering algorithm has some properties that are specified below:

- There should be always k cluster.
- Each cluster always contains at least one item.
- Non-hierarchical clusters are formed and they do not overlie.

**Algorithm:** K-means:- The k-means algorithm is one of partitioning algorithm, in which each cluster’s center is represented by the mean value of the objects in the cluster [1].

**Input:**

K: represent number of clusters,

D: specify a data set contain n objects.

**Output:**

A set of k clusters are generated.

**Method:**

- Choose k data objects representing the cluster centroids.
- Assign each data object of the entire data set to the cluster having the closest centroid.
- Compute new centroid for each cluster, by averaging the data objects belonging to the cluster.
- If at least one of the centroids has changed, go to step 2, otherwise go to step 5.
- Output the clusters.

**Advantages of K-means:**

- Simple and easier to understand.
- Fast and robust.
- Relatively efficient and give better result.

**Challenges in K-means:**

- 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 centers selection.
- K-means algorithm may converge to local minima.

**KNN Approach:** It 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 nonsupport vectors like SVM [10]. It makes decision on the basis of entire training data set.

![Fig.3.1 Classification of KNN](image)
IV. PROPOSED METHODOLOGY

The k-mean clustering algorithm is used to cluster various data in data mining. The main problem exists in k-mean clustering algorithm is of efficiency. The time required for data cluster is very high and cluster quality is not so good. To improve cluster quality and to reduce escape time of the algorithm ranking algorithm will be used in the enhancement. The ranking algorithm will rank data in the dataset and according to ranking of the data.

V. EXPERIMENTAL RESULTS

The proposed idea will be implemented in MATLAB which is widely used in all areas of research universities, and also in the industry.

As figure 5.1 illustrates that the K-mean is the algorithm in which the data will be clustered according to Euclidian distance. The random center points had been selected from the data. The Euclidian distance will be
calculated from the data centers to other points and points will be clustered accordingly. The output of the clustered will be shown in the 2D plane. When the data will be shown in 2D plan, some points which are very close to each other cannot be shown which reduce the cluster quality.

As illustrated in figure 5.2, the dataset which is read with xls read commands of MATLAB will be plotted on the 2-D plane.

As shown in figure 5.3, to analyze the performance of the algorithm the K-mean algorithm will be applied on another dataset. In this dataset, various figures have been shown for data clustering.
As shown in Figure 5.4, first select points that are used for clustering the data. The dataset which is used for clustering is been clustered and each cluster will be marked with different colors. In this figure various iteration runs which means at every iteration new centered point is selected on the basis of that centered point is cluster assignment procedure will be done.

As shown in figure 5.5, 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.
As illustrated in figure 5.6, 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.

As illustrated in figure 5.7, the final clustering result is shown on the 2-D pane. 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 then the final clusters are procreated on the bases of Euclidian distance.
Fig. 5.8 final output

As shown in the figure 5.8, to improve accuracy of k-mean clustering the points which are remained un-clustered are clustered using the technique of normalization. The normalization calculate Euclidian distance in the iterative manner. When high accuracy of clustering is achieved the output is shown in the form of clusters.

VI. CONCLUSION

In this paper, it has been concluded that k-mean clustering the efficient technique which can cluster dataset points. In k-mean clustering dataset is loaded and from the loaded dataset central points are selected according to defined number of clusters. The central point acts as reference point and from which Euclidian distance is calculated and according to Euclidian distance members are assigned to each cluster. Due to user defined cluster values, some points of the dataset are remained un-clustered which reduced accuracy of clustering. In this paper enhanced K-Mean clustering is defined to improve accuracy. The technique of normalization is applied which is used to calculate Euclidian distance in the iterative manner to improve its accuracy.

REFERENCES
