Two-step Technique for Prediction Analysis using K-Means Clustering Algorithm

Shalu Saxena
Department of Computer Science & Engineering,
PG Scholar, SRMCEM
Lucknow

Pankaj Kumar, PhD
Department of Computer Science & Engineering,
Assistant Professor, SRMCEM
Lucknow

Raj Gaurang Tewari, PhD
Department of Computer Science & Engineering
Assistant Professor, SRMCEM

ABSTRACT
The technique that is utilized for analyzing the complex data is known as data mining technique. As per the input dataset provided, the predictions are made for the data with the help of prediction analysis method. There are various new techniques proposed for the execution of prediction analysis technique. In this paper, the k-mean algorithm is utilized for categorizing the data. Further, for the classification of this data, the SVM classifier is applied. For improving the performance of prediction analysis in terms of accuracy the back propagation algorithm is used along with the k-mean clustering algorithm. For executing this proposed technique, the MATLAB tool is used. As per the experimental results it is concluded that the accuracy of the clustering algorithm is improved as well as the execution time utilized for prediction analysis is decreased.

Keywords
Prediction, Classification, Back Propagation, K-mean, SVM

1. INTRODUCTION
The government, corporate, and industrial communities are confronted with a constantly increasing number of databases. These databases require to be managing as well as exploring. The principal requires secure access to distributed heterogeneous multimedia databases with rich metadata and meeting timing constraints. The second requires exploratory tools supporting the identification of domain and mission critical elements, for example, patterns in data access (e.g., security breach determinations), patterns in data (e.g., marketing and clustering), or for patterns in transactions (e.g., data compression), to site a couple. Knowledge Discovery in Databases is a moderately new research territory that employs a variety of tools to explore and identify structure and patterns in these huge databases [1]. On the basis of the properties of a specific set of objects, the division of these objects on the basis of their similarities is known as clustering process. A specific joint algorithm that can be applicable to almost all required information analysis is done with the help of the technique which partitions the data. The major fundamental clustering methods can be classified into following categories [2]:

1.1. Partitioning Methods
The highly similar samples present within a cluster are combined within this partitioning technique. The different clusters formed have high dissimilarity among themselves. The various partitioning methods are distance-based. In a system, if k is the given number of partitions for construction, the partitioning method helps in creating an initial partitioning [3]. Further, the iterative relocation method is used for improving the partitioning technique with the help of moving the objects from one group to another.

1.2. Hierarchical Methods
The hierarchical decomposition of given set of data objects in known as the hierarchical clustering technique. There are two broader classifications involved here which are the agglomerative and the divisive based methods [4].

1.3. Density Based Methods
The clusters which are of various arbitrary shapes can be encountered with certain difficulties. Hence, there are certain density-based methods which are utilized for the arbitrary shapes for the notion of density. The clusters keep growing as long as the density of the neighborhood exceeds certain threshold. The notion of density is the base of this method.

1.4. Grid Based Methods
The object space is quantized into a finite number of cells which create a grid structure [5]. This method is known as the grid based method. This method is really of high speed and does not depend on the number of data objects present.

1.5. k-means clustering algorithm
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 [6]. 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, ($C1 \ldots\ldots Ck$), 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. 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 [7].

2. LITERATURE REVIEW
Richa Sharma, et.al proposed in this paper [8], a study on the various methods in field of medical data mining utilizing different classification and clustering techniques. This survey study reveals the importance of research in area of life debilitating disease diagnosis. Promote the instance of wellbeing is talked about that one needs to achieve for the precision of cent percent various investigates approximately comes to their target yet disease diagnosis suffers from high false alarm so we have to propose novel approach to reduce this false alarm rate which would help in early diagnosis of disease.

Sonali Shankar, et.al discussed in this paper, [9] that the large volume of data in all fields over the globe must be managed and is utilized by the decision creators to get something productive out of it. The enormous data of 14000x5 of Harvard University online course is broke down to discover the performance metrics of registered students from different countries by means of K-mean clustering method. The attributes are subsequently compared with the average grades of students of respective countries and it is concluded that the grades are by all account not the only factor to represent the best possible understanding of the course. The analysis can likewise be extended to think about alternate attributes, for example, 'certified', "explored" and so forth.

Vadlana Baby, et.al proposed in this paper [10], an efficient distributed threshold privacy-preserving k-means clustering algorithm is proposed that uses the code based threshold secret sharing as a privacy-preserving instrument. This protocol takes less number of iterations compare with existing protocols and it don't require any trust among the servers or users. The experiment results are likewise furnished alongside comparison and security analysis of the proposed scheme. It permits gatherings to collaboratively perform clustering and hence avoiding trusted outsider. The protocol is compared with CRT based clustering proposed. This algorithm does not require any trust among the servers or users and it give idealize privacy preserving of client data.

Cheng-Fa Tsai, et.al explained in this paper, [11] that clustering is the unsupervised classification of patterns (data items, feature vectors, or perceptions) into groups (clusters). In this paper, another data clustering method is presented for data mining in large databases. These simulation results demonstrate that the proposed novel clustering method performs superior to anything the Fast SOM combines K-means approach (FSOM+K-means) and Genetic K-Means Algorithm (GKA). Furthermore, in every one of the cases studied, this method creates much smaller errors than both the FSOM+K-means approach and GKA.

Steve Russell, et.al proposed in this paper [12], four key fuzzy enhancements to traditional database marketing. To start with, customers open have significant membership values in more than one distinct fuzzy cluster and can be considered in a natural manner for hybrid or multiple contacts in a given marketing campaign. Second, fuzzy clustering outcomes are appeared to be dependent on the specific offer or marketing message. Third, there are contrasts in clustering outcomes after some time as various offers and treatments are successively presented to consumers, and as products and tastes change. Fourth, in the more extended run, formal procedures can be suggested including intuitive fuzzy-based clustering metrics for continuous process improvement, to support progressively flexible and opportunistic campaign management.

Vaibhav Kumar, et.al proposed in this paper, [13] K-mean clustering based unsupervised learning method has been adopted for the performance enhancement of cooperative spectrum sensing in generalized k-μ fading channels. The achieved results affirm that K = 2 gives the best performance characteristic compared to the cluster size of K = 4 and K = 7. Encourage concentrates on have as of now been initiated to exploit other learning methods, for example, graph discriminant analysis on multi-complex and restricted Boltzmann machines (RBM) for improved performance pick up in cooperative spectrum sensing and spectrum occupancy prediction in CR networks.

3. PROPOSED WORK

The prediction of situations as per the input dataset is known as prediction analysis. There are two phases involved within this technique. The clustering of similar as well as dissimilar data is done in the initial phase of this technique which includes k-mean clustering for this process. Further, for the purpose of classifying this data, the SVM classifier is utilized. There are three steps which execute the k-mean clustering amongst which the first one calculated the arithmetic mean of the complete dataset across the central point. Next, the second step of the process is the computation of Euclidian distance from the central point.

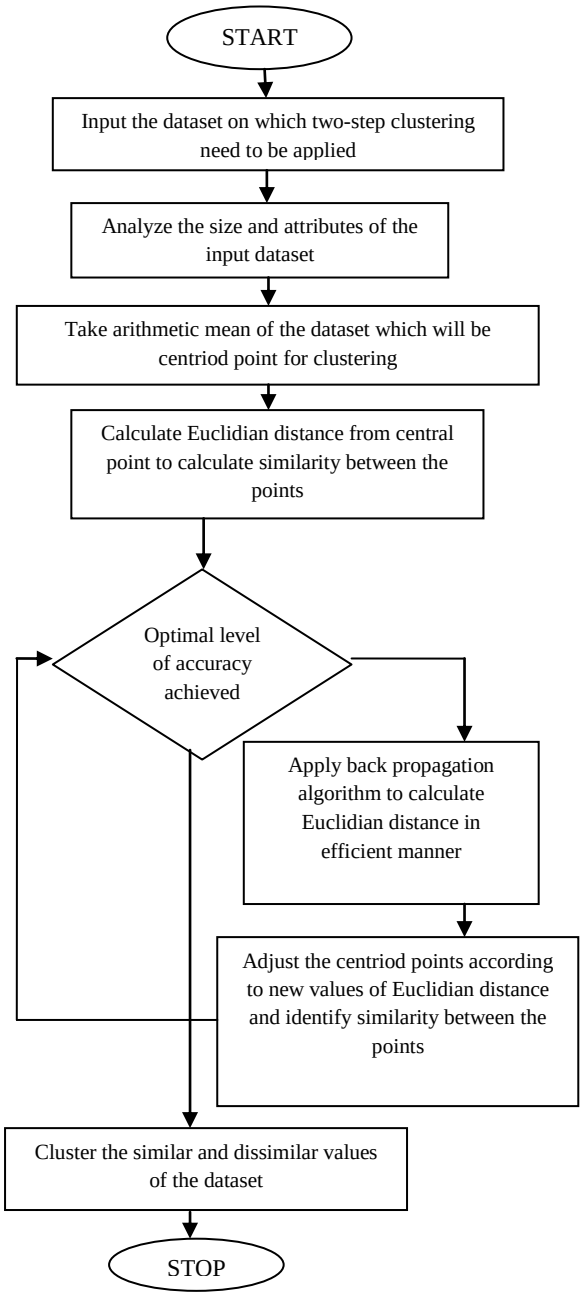

Fig. 1 Flowchart for Proposed Algorithm

The next step is the last step that clusters the data as per the similarity and provides the similar data for classification to the SVM classifier. On the basis of the quality of cluster the data is classified. The cluster quality is to be improved that

further enhances the classification quality. These enhancements are provided in this work by the enhancement of the k-mean clustering algorithm. The cluster quality is enhanced here by applying the back propagation algorithm with the k-mean clustering algorithm. The Euclidian distance is calculated in a dynamic manner through the back propagation algorithm. The final distance that is selected for data clustering is the Euclidian distance that has maximum accuracy.

4. RESULTS AND DISCUSSIONS

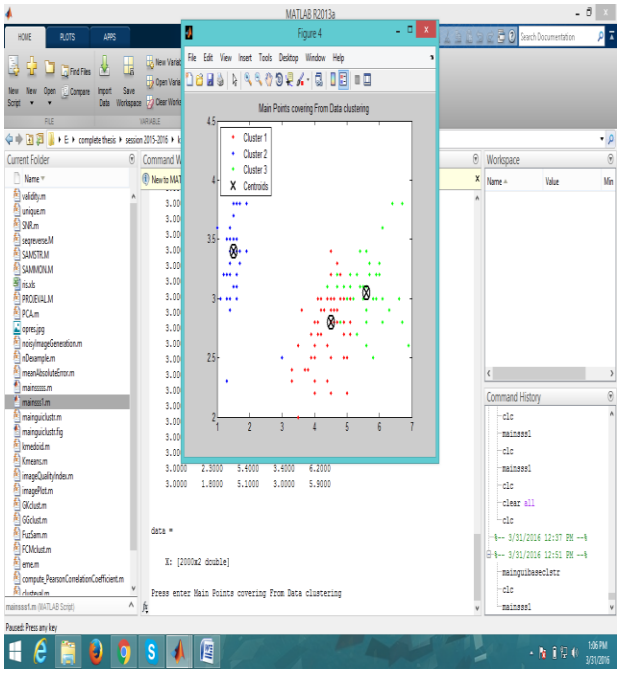

Fig 2: Voronoi Representation

As shown in figure 2, 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 voronoi plane for better analysis of dataset.

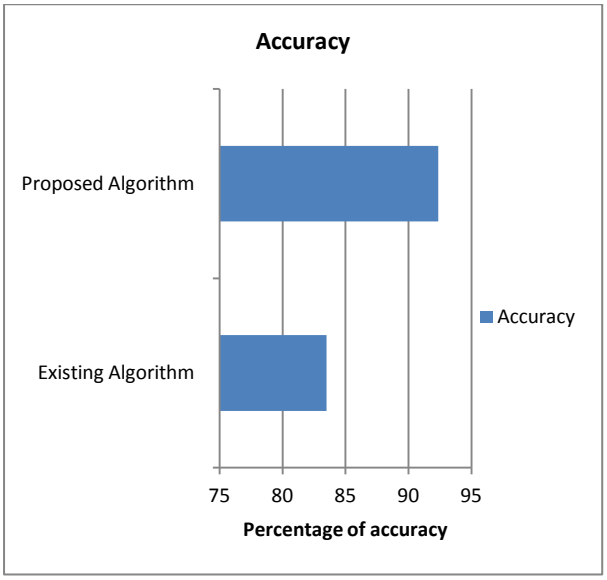

Fig 3: Accuracy Comparison

As shown in figure 3, the accuracy of proposed and existing algorithm is been compared and it is been analyzed that proposed algorithm has high accuracy due to clustering of uncluttered points from the dataset

Fig 4: Execution time

As illustrated in figure 4, the execution time of proposed and existing algorithm is been compared and due to used of back propagation algorithm execution time is due in the proposed work

5. CONCLUSION

As per the experimental results achieved in this paper, it is concluded that the proposed prediction analysis technique is efficient for analyzing the complex data. For enhancing the accuracy of data clustering technique, the back propagation algorithm is used along with the k-mean clustering algorithm. The clustered output provided is further classified with the help of SVM classifier. MATLAB tool is utilized for executing this algorithm. An improvement in accuracy has been observed here upt o 20 percent. Further there is 10 percent of decrement in the execution time with the help of this proposed novel technique.

6. REFERENCES

[1] Javeria Ayub, Jamil Ahmad, Jan Muhammad, Usman Akram, Imran Basit," Glaucoma Detection through Optic Disc and Cup Segmentation using K-mean Clustering", 2016, IEEE, 978-1-5090-1252-7

[2] Asmita Singh, Devendra Somwanshi," Offline Location Search using Reverse K-Mean Clustering & GSM Communication", 2015, IEEE, 978-1-4673-7910-6

[3] R. Kumari, Sheetanshu, M. K. Singh, R. Jha, N. K. Singh," Anomaly Detection in Network Traffic using K-mean clustering", 2016, IEEE, 978-1-4799-8579-1

[4] Vaibhav Kumar, Deep Chandra Kandpal, Monika Jain," K-mean Clustering based Cooperative Spectrum Sensing in Generalized *k*-μ Fading Channels", 2016, IEEE, 978-1-5090-2361-5

[5] Dweepna Garg, Khushboo Trivedi," Fuzzy K-mean Clustering in MapReduce on Cloud Based Hadoop", 2014, IEEE, ISBN No. 978-1-4799-3914-5

[6] Anjali Gautam, H.S. Bhadauria," White Blood Nucleus Extraction Using K-Mean Clustering and Mathematical Morphing", 2014, IEEE, 978-1-4799-4236-7

[7] P. Shanmugavadivu and R. Santhini Rajeswari," Identification of Microcalcifications in Digital Mammogram using Modified K-Mean Clustering", 2012, IEEE, ISBN: 978-1-4673-5144-7

[8] Richa Sharma, Dr. Shailendra Narayan Singh, Dr. Sujata Khatri," Medical Data Mining Using Different Classification and Clustering Techniques: A Critical Survey", 2016, IEEE, 978-1-5090-0210-8

[9] Sonali Shankar, Bishal Dey Sarkar, Sai Sabitha, Deepti Mehrotra," Performance Analysis of Student Learning Metric using K-Mean Clustering Approach", 2016, IEEE, 978-1-4673-8203-8

[10] Vadlana Baby, Dr. N. Subhash Chandra," Distributed threshold k-means clustering for privacy preserving data mining", 2016, IEEE, 978-1-5090-2029-4

[11] Cheng-Fa Tsai, Han-Chang Wu, and Chun-Wei Tsai," A New Data Clustering Approach for Data Mining in Large Databases", 2002, IEEE, 1087-4089

[12] Steve Russell, Steve Russell," Fuzzy Clustering in Data Mining for Telco Database Marketing Campaigns", 1999, IEEE, 0-7803-521 1-4

[13] Vaibhav Kumar, Deep Chandra Kandpal, Monika Jain," K-mean Clustering based Cooperative Spectrum Sensing in Generalized k-μ Fading Channels", 2016, IEEE, 978-1-5090-2361-5