I J C T A, 9(11) 2016, pp. 5169-5180
© International Science Press

IMPROVED K-MEANS CLUSTERING ALGORITHM USING BACK PROPAGATION METHOD

Shaweta Bhardwaj[*] and Vikas Verma[**]

Abstract: Clustering is procedure which is utilized to examine the data in proficient way and to get mined data. The various clustering has been proposed so far which can cluster the datasets. Among all the proposed algorithms, K-mean is the efficient algorithm in the terms of complexity and execution time. The k-means algorithm used the concept of central point and Euclidian Distance to cluster the dataset. When the dataset is of complex in nature, K-mean clustering is unable to drive relationship between the attribute of the dataset which leads to reduction in cluster quality. In this paper, improvement has been proposed in K-means clustering to increase the cluster quality of complex datasets.

Key Words: Clustering, Prediction Analysis, K-Means, Back Propagation, Data Mining

1. INTRODUCTION

In today's world, every single second, enormous data is stored for processing and used for analysis and prediction. The process named Data Mining is used for analyzing data to extract interesting patterns and knowledge. To deal with this much of data, data is classified or grouped into a set of parts named as clusters. The cluster is applied to analyze data in more efficient manner. In this paper, we have taken prediction analysis in which final clustered data will be analyzed and concluded according to requirement. There are number of algorithms have been proposed and tested. The objective of this research paper is to review limitations which left in different algorithms especially K-means and there will be a comparison in terms of accuracy of those clusters.

Data Mining is a technique of knowledge discovery which helps in predicting hidden details from data. It works for analysis purpose and according to that analyze different type of data with the use of available data mining tools. Data mining is used these days in every field of technology either it is medical or computer. The extracted information is used for wide range of applications like customer retention, education system, production control, healthcare, scientific discovery and decision making etc. Data mining [1] is studied for different databases like relational database, multimedia databases, and object-relational databases, etc. The steps of knowledge discovery from database are depicted in Figure 1.

There are two types of data analytic:

1. Prediction   2. Classification

[*] School of Computer Science and Engineering Lovely Professional University Phagwara, Punjab shaweta2bhardwaj@gmail.com

[**] School of Computer Science and Engineering Lovely Professional University Phagwara, Punjab vikas.verma@lpu.co.in

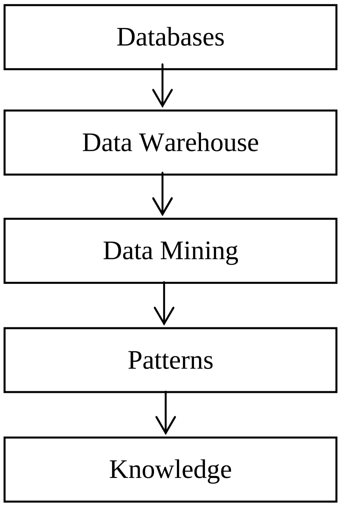


Figure 1: Knowledge discovery process

Classification: Classification models forecast definite class labels and also the models of prediction which predict uninterrupted valued functions.

Prediction: Prediction model is basically to make a prediction of the costs in bucks of prospective customers on the basis of computer value given by the revenue and profession.

Data mining has never been a laidback task because the algorithms which are used become very complex and also we never get the whole data at one place. Data has to be taken from different data sources in various forms, which results into some issues. The major issues are following:

- Data Mining Approach and User Interface
- Performance
- Dissimilar Types of Data

Figure 2 shows the categorization of different issues. It includes set of some functional modules which explains the following functions:

(a) Classification

(b) Cluster analysis

(c) Outlier analysis

(d) Prediction analysis

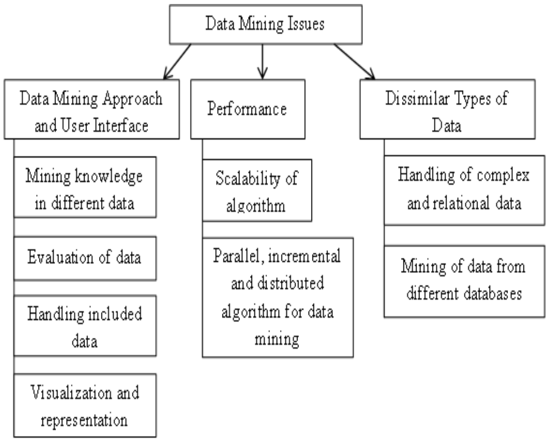

Data Mining Issues
Data Mining Approach and User Interface
Performance
Dissimilar Types of Data
Mining knowledge in different data
Evaluation of data
Handling included data
Visualization and representation
Scalability of algorithm
Parallel, incremental and distributed algorithm for data mining
Handling of complex and relational data
Mining of data from different databases

Figure 2: Data mining major issues

2. DATA MINING CLUSTERING

Cluster analysis [3] is widely used in number of applications which include pattern recognition, market research, and the image processing. If we talk about business, clustering is helpful to marketers to find out the interests of the customers who are distinct on the basis of their purchasing arrangements and then characterizing groups of the customers. In biology, it can be used to obtain classify genes, plants and animal taxonomies and gain vision into structures intrinsic in populations. In geology, specialist can employ grouping to identify similar lands in a place. Data clustering can also be useful to categorize forms on the webpages to discover and process the information.

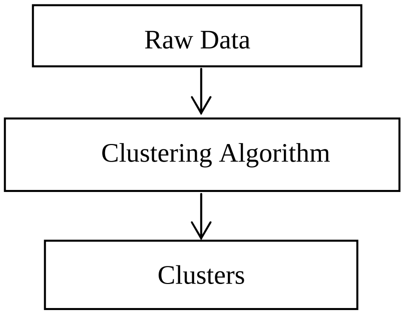


Figure 3: Data mining clustering steps

Data collecting [4] (or clustering), is an denounce classification method whose purpose is to create groups of objects, or clusters, in such a way that objects in the same cluster are having same properties and those objects in different clusters are quite distinguishable. The same is depicted in Figure 3. This is a traditional topic in the field of mining of the data. It is the initial step in the direction of exciting knowledge discovery. It is the procedure of selecting data objects and grouping them into set of disjoint classes as shown in Figure 4.

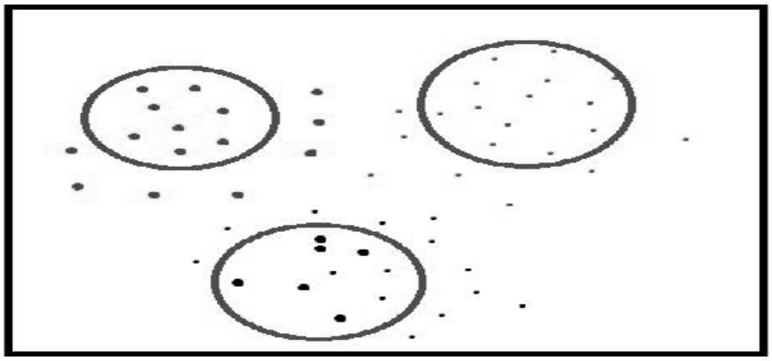


Figure 4: Data clustering

This method is used to make groups of similar kind of documents, but it is dissimilar from classification of documents which get clustered on the fly rather through the usage of some topics. There is one more advantage of clustering, the documents can appear in multiple topics and ensure that useful document does not get misplaced from search results. A fundamental clustering algorithm forms a vector of topics for each of document which measures the weights of - how healthy the document can fit into every cluster [5]. Clustering comes under un-supervised classification. It refers to a method that allocates data objects in the set of classes. Here there is one more clustering named unsupervised clustering which is completely different from pattern reorganization in the area of statistics, called as discriminate analysis and decision analysis, which classify the data from a set of objects.

2.1 Methods of Clustering

There are number of algorithms which are used for clustering. The most important fundamental methods which are categorized below [8]:

(a) Density-based method

(b) Hierarchical method

(c) Partitioning method

(d) Grid-based method

Density Based Methods: In this method, objects of the cluster are based on objects distance. Spherical shaped clusters can be discovered by these methods and meet difficulty in finding the clusters of random shapes. So for arbitrary shapes new methods are used known as density based methods which are based upon the view of density. It helps to discover arbitrary shape clusters. It also handles noise in the data. It is one time scan. It requires density parameters also [9]. Figure 5 depicts density based modeling looks like.

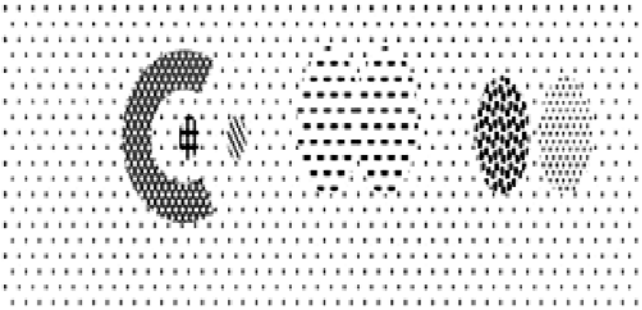


Figure 5: Destiny Based Clustering

Hierarchical Method: It explains the splitting of available set of data objects. It can be classified as being either divisive or agglomerative, based upon the order of decomposition thus formed. In this type of clustering, it is possible to view partitions at various levels of granularities using various types of K. e.g. Flat Clustering [10] of granularities using different types of K. Agglomerative approach works in bottom to up approach that starts with each object which forms an independent group. It then combines groups close to each another up to all sets are combined into single one. Hierarchical clustering approach is shown in Figure 6.

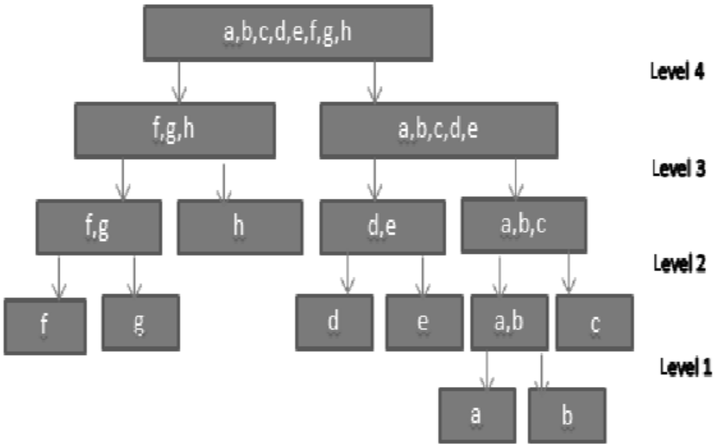


Figure 6: Hierarchical Based Clustering

Partitioning Methods: The fundamental basis is a mixture of high comparability of the samples of groups with high refinement between various clusters. The vast majority of the partitioning strategies are distance-based. By taking the consider of partitions "k" to build, which makes an opening point part which then uses an repetitive procedure that tries to enhance the apportioning by the moving objects starting with one, then onto the next gathering.

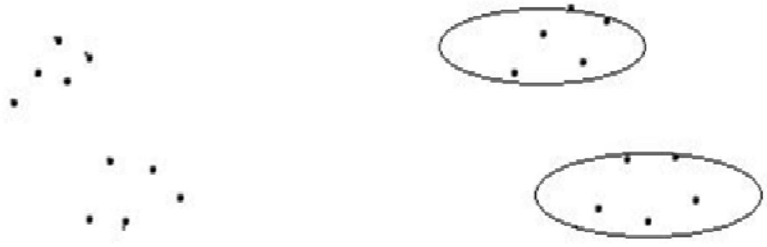


Figure 7: Partitioning Clustering

Grid Based Methods: It is a quick strategy and is independent of the quantity of data objects and just reliant on the quantity of cells in every measurement as in quantized space. These strategies do quantize the object space into a limited number of cells which custom a grid structure. In this strategy, objects together structures the framework which is named as grid. This calculation quantizes the space into grids which are in constrained in number and which do each operation on this quantized space. Those sort of methodologies have their favorable circumstances of handling in quick way and time autonomy of the extent of information set and depends just in every measurement of number of fragments in the quantized space [11].

3. K-MEANS ALGORITHM FOR CLUSTERING

The k-means algorithm is a partitioning method which is used for many clustering tasks with less dimension datasets. Using k as cluster parameter, divide n objects into k clusters so that the objects in the identical group are similar to each other. [12] The algorithm tries to search for the cluster centers from C_1 to C_k, such that the sum of the taking square of the distances of all data point, x_i , $1 \leq i \leq n$, *to* its nearest *Cj (which is a cluster center)*, $1 \leq j \leq k$, is reduced. Here initially the algorithm selects the random k objects, and in which there will be a representation of a cluster mean or centre. Then, each object *xi* in the data set which is allocated to the adjacent cluster centre i.e. to the similar centre. The calculation then registers the new mean for each of one group and reassigns every object to the closest new centre. This procedure emphasizes until no progressions happen in the allocating of objects. This sort of joining results in backing off the sum of-squares error which is characterized as the summation of the squared distances from every object to its clustered centres [3, 9]. The accompanying strategy outlines the k-means algorithm [1]:

For partitioning, in this calculation, every center of the cluster is represented by the objects mean value in the cluster [16].

Input

k: Here k is the number of clusters which are going to use.

D: A data set which is containing n objects.

Output:

We get a set of ‘k’ clusters.

Method:

(i) First select k objects in arbitrary way from data sets by taking it as initial cluster center.

(ii) Now simply rehash the strides.

(iii) Based upon the objects in the group's mean worth, again dole out objects to the cluster, to which the article is the comparative in greatest way.

(iv) By redesigning the cluster means value; simply figure the mean estimation of the objects for every cluster.

(v) Do it until there is no change.

Drawbacks

Despite being used in a wide array of applications, the k-means algorithm has following drawbacks:

(i) As many clustering methods, this algorithm says that the clusters k in the database is called as beforehand which are not completely right in real-world applications [17].

(ii) By saying as an iterative technique, this algorithm is mainly important for initial centers selection.

(iii) The k-means algorithm may converge to local minima.

(iv) K-means does not work for density based clusters, which is actually disadvantage of using this algorithm and we can see this in Figure 8.

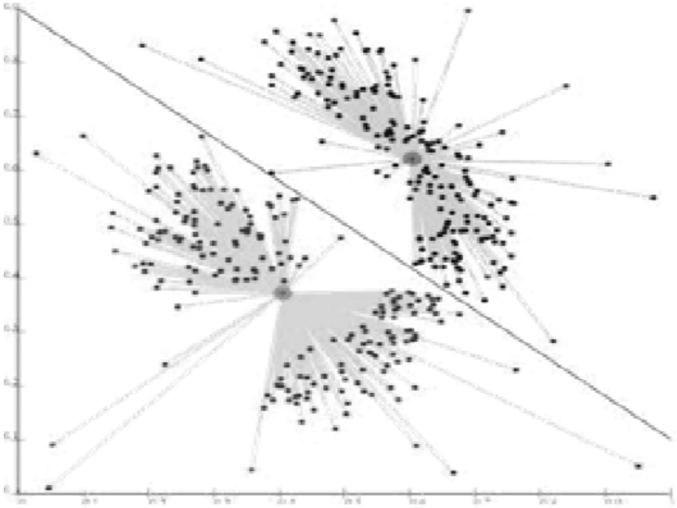


Figure 8: Density based clustering Issue

Enhancements in K-Mean algorithm

There are various types of enhancements which have done till now. These enhancements are as follow:

- First enhancement in the areas of reduce number of repetition and Time complexity [3]: In this algorithm input is entered in same manners as it is entered in basic algorithm. The whole process is divided in two phases. In phase 1, cluster size is fixed and output of first phase helps to form initial clusters. Array of elements is broken into sub parts that represent initial clusters. In phase 2 cluster sizes vary and its output forms finalized clusters. The entire process is calculated, until there is no change in the movements. As in this algorithm, the initial centroid is chosen randomly from the input data, so clusters vary from one another. The number of iterations and total elapsed me also changes in each run of the same data. In proposed K-mean algorithm initial centroid are calculated. If the data is same, it results in same calculations, the count of iterations remains constant and elapsed time is also improved. That is the main cause that proposed K-mean clustering algorithm is efficient from basic K-mean algorithm.

- Second enhancement as far as enhanced initial centre [13]: k-means clustering algorithm is a prevalent clustering algorithm; however the nature of the last clusters in this strategy depends for the most part on the beginning centroid, which is chosen haphazardly. In addition, the k-means is computationally high in expense as well. The proposed algorithm has been observed to be more productive and exact and by contrasting this and genuine algorithm. The given method of finding the better initial centroid which gives an accurate way to assign the data points to the clusters. This method ensures the total mechanism of clustering in O (nlogn) time without any loss of accuracy. This technique does not need input like threshold value. Here the given technique gives the more admired results. The value of k, desired number of clusters is still required to be given as an input to the algorithm.
- Third enhancement [21] in terms of improve accuracy and efficiency: This algorithm is mainly used for clustering of large datasets. In normal algorithm, there is no assurance of good results for the correctness of the last clusters by depending upon the selection of initial centroid. Moreover, the computational complexity of the standard algorithm is objectionable higher owing to the need to assign new values to the data points in numbers and with the process of iteration of the loop. This paper represents an improved version of k-means algorithm: that adds a systematic method for finding initial centroid and an efficient way of giving data values to clusters. This method tells that the full process of clustering in O (n_2) time without sacrificing the correctness of clusters. In the last improvements of the k-means algorithm, it compromises on either accuracy or efficiency. There is a limitation of this proposed algorithm which is - the value of k^{th} number of desired clusters is still required to be given as an input, as without giving thought to the distribution of the data points. Developing some statistical methods to compute the value of k, which depends upon the data distribution, is suggested for future research. A method for refining the computation of initial centroid is worth investigating.

4. RELATED WORK

(i) Medical field: The work [1] is done based on the fact that huge data is available in medical field to extract information from large data sets using analytic tool. In this paper a real data set has been taken from SGPGI. The clustering is the alternate solution for data analytics. The main focus of this paper is to bring a novel method which is based upon foggy k-mean clustering. Here result of the experiment depicts that this algorithm has excellent result in case of datasets which are real when comparing to simple k-means and provides an enhanced result to the problems of real world.

In the same field, prediction analysis approach [15] is defined for survivability rate of one of decease breast cancer using data mining techniques. SEER public datasets have been used in this project. This preprocessed dataset consists of 151,886 records which have SEER's 16 fields. After that they have analyzed some of the data mining techniques: Naive Bayes, C4.5 decision tree algorithm and back propagated neural network. Many experiments have been implemented using above mentioned experimental. At the end existing techniques have been compared with the achieved prediction performance. Later on it is concluded that C4.5 algorithm is more efficient in the case of accuracy than those other two techniques.

(ii) Weather forecasting field [2]: They explained that clustering is the influential tool; used in various prediction tools. In this paper generic methodology of incremental K-mean clustering is proposed for weather forecasting. This research has been done on air pollution of west Bengal dataset. This paper generally makes use of typical K-means clustering on a list of weather category and the main air pollution database will be developed based on the peak clusters mean values. Thus it is able to predict weather information of future. This prediction database is completely rely on the weather and the forecasting which is prepared to mitigate the consequences of pollutions and to launch focused modeling computations for prediction and forecasts of weather events. Here accuracy of this approach is also kept in notice.

(iii) Academic field [14]: By this paper, they proposed a system named Student Performance Analysis System which actually tells the performance of student and then analyzed to keep path of student's result in a particular university. The proposed project offers a system which predicts performance of the students on the basis of their result on the basis of analysis and design. The proposed system offers the prediction of student performance by the rules which are generated via data mining technique. The approach used in this area is all about classification that actually classifies the students based on students' grade.

Another work in the same field [11] was done to define the ability of the student performance of high learning. An approach was designed to analyze student result based on cluster analysis and use standard statistical algorithm to arrange their score according to the level of their performance. In this paper K-mean clustering is implemented to analyze student result. The model was added up with deterministic model to check student's performance of the system. Another orthogonal search based rule extraction [12] was also defined.

(iv) Stock Market [15]: In this paper, it is concluded that forecasting stock return is one of the important subjects to learn for prediction from data analysis. It is analyzed that past investigations help to predict future in data analysis. This paper helped investors in stock market better timing for the buying and selling stocks on the basis of knowledge of past historical experiments. In this paper they define decision tree classifier which is one of the best data mining techniques.

On other side, approaches can be defined to evaluate Service Development Approaches (SDA) used for migrating towards Service Oriented Architecture (SOA) [20].

We summarize the clustering algorithms used for prediction analysis in various areas in Table 1.

Table 1
Clustering Algorithm used in Prediction Analysis

Problem and Objective	*Algorithm and Method*	*Dataset and Data Source*
Explained how huge data is available in medical field to extract information [1]	Compared to simple k-means and provides an enhanced result to the problems of real world	A real data set has been taken from SGPGI
Making data analysis for forecasting stock [15] and e-commerce [6] items to make prediction for future	Decision tree classifier technique is used	Taken historical data from stock history
To outline the ways to deal with showing profiles of educators in advanced education on the premise of their scores on the ATI abbreviated for approaches to teaching Inventory (ATI) [7]	A hierarchical cluster analysis was performed on the survey information	Questionnaires and Interviews conducted in University X with 30 academics and response rate of 20%
Analyzed various disease predictions techniques in the extremely fast growing field of medical [9]	K-means algorithm	Diseases like breast cancer, heart diseases data has taken from hospitals
Proposed generic methodology of incremental K-mean clustering for weather forecasting [10]	Used K-means clustering for making results based on the peak clusters mean values	Air pollution of west Bengal has taken as a dataset
Analyze the ability of the student performance of high learning [11]	Cluster analysis and use standard statistical algorithm like K-means	Student academic record file, not mention from where it is obtained
Proposed a system named SPAS (student performance analysis system) which predicts the performance of the students on the basis of their result on the basis of analysis and design [14]	Utilized classification that really arranges the students in view of student' evaluation	Student academic random record
Evaluating undergraduate student academic performance [18]	Using a combination of DM methods like ANN (Artificial Neural Network), Farthest First method based on k-means clustering and Decision Tree as a classification approach	Computer Science Student data of the department at National Defence university of Malaysia (NUDM), Faculty of Science and Defence Technology

5. PROPOSED METHODOLOGY

The prediction based analysis is the technique to analyze the clusters which are formed with hierarchal clustering algorithm. In the hierarchal clustering algorithm dataset will be taken as input and relationship will be maintained among various attributes of the dataset. On the basis of derived relationship, central point will be considered. From that central point, Euclidian distance will be calculated. Based on Euclidian distance, final clusters are formed which are analyzed according to application. But in the complex datasets it is very difficult to derive a relationship among attributes of the dataset. In this proposed work, Back Propagation algorithm will be applied which can derive relationship between various attributes of the complex dataset and after that final clusters will be easy to analyze. It will overcome the overlapping of clusters which have closure values.

Step 1: Here we will be taking dataset that can be from any application area either some disease, performance of student or Google trends.

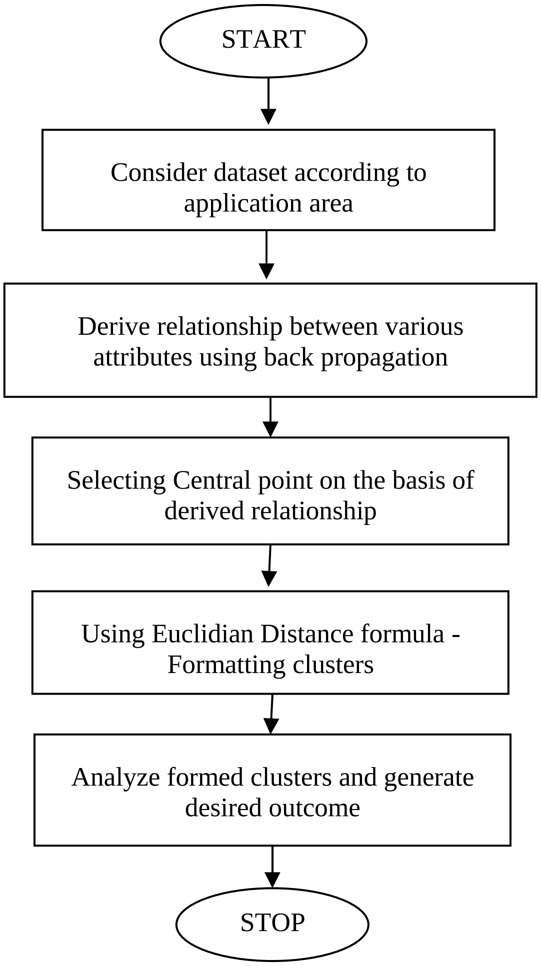
Step 2: Here, we will be deriving the relationship between various attributes which we have taken in our dataset by using algorithm ‘Back Propagation’.

Step 3: After deriving relationship, now select the central point and which will decide the further process.

Step 4: In this part, the central point which we have found in last step will be used to formulate different clusters. These clusters will be showing that there will be less overlapping among clusters.

Step 5: On the basis of these formulated clusters, we will be generating the outcomes.

The following flow chart depicts the major steps.


START
Consider dataset according to application area
Derive relationship between various attributes using back propagation
Selecting Central point on the basis of derived relationship
Using Euclidian Distance formula - Formatting clusters
Analyze formed clusters and generate desired outcome
STOP

So, if we have huge data set with us that need to be clustered to do prediction analysis, the best approach will be to get the value of number of clusters to be defined depending on the complexity as well as size of data set. This includes considering the evaluation of contribution of different parameters involved in the data set. Hence the result is the improvement in accuracy and efficiency.

6. CONCLUSION

In this work, it has been proposed that k-mean clustering is an efficient technique which can cluster dataset points. When dataset is loaded and from that 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. Because of client characterized group values, a few purposes of the dataset are remained un-clustered which decreases the precision of clusters. So change has proposed in k-means clustering which will automatically define number of clusters and generate final clustered data.

References

[1] K. A. Yadav, D. Tomar, and S. Agarwal, "Clustering of Lung Cancer Data Using Foggy K-Means," International Conference On Recent Trends In Information Technology (ICRTIT-2013).

[2] K. A. Abdul Nazeer, M. P. Sebastian, “Improving the Accuracy and Efficiency of the k- means Clustering Algorithm,” Proceedings of the World Congress on Engineering, Vol IWCE 2009, July 1 - 3, 2009, London, U.K.

[3] Azhar Rauf, 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 12 (7): 959-963, 2012 ISSN 1990-92332012.

[4] Osamor VC, Adebiyi EF, Oyelade JO and Doumbia S, “Reducing the Time Requirement of K-Means Algorithm,” PLoS ONE, Volume 7, Issue 12, pp-56-62, 2012.

[5] Qasem A. Al-Radaideh, Adel Abu Assaf Eman Alnagi, “Predicting Stock Prices Using Data Mining Techniques,” The International Arab Conference on Information Technology (ACIT’2013).

[6] Luhach, A.K., Dwivedi, S.K. and Jha, C.K., 2016. Implementing the Logical Security Framework for E-Commerce Based on Service-Oriented Architecture. In Proceedings of International Conference on ICT for Sustainable Development (pp. 1-13), Springer Singapore.

[7] P. Golding and O. Donaldson, “Predicting academic performance,” in Proc. Front. Educ.36th Annu. Conf., 2006, pp. 21-26.

[8] Madhu Yedla, T M Srinivasa, “Enhancing K-means Clustering Algorithm with Improved Initial Center,” International Journal of Computer Science and Information Technologies, Vol. 1 (2) 2010, page 121-125.

[9] K. Rajalakshmi, Dr. S. S. Dhenakaran, N. Roobin, “Comparative Analysis of K-Means Algorithm in Disease Prediction,” International Journal of Science, Engineering and Technology Research (IJSETR), Volume 4, Issue 7, July 2015.

[10] Sanjay Chakraborty, N.K.Nagwani, Lopamudra Dey, “Weather Forecasting using Incremental K-Means clustering,” CIIT International Journal of Data Mining & Knowledge Engineering, May 2012.

[11] Oyelade, O. J, Oladipupo, O. O and Obagbuwa, I. C, “Application of k-Means Clustering algorithm for prediction of Students’ Academic Performance,” International Journal of Computer Science and Information Security, Vol. 7, No. 1, 2010.

[12] T. Etchells, À. Nebot, A. Vellido, P. Lisboa, and F. Mugica, “Learning what is important: Feature selection and rule extraction in a virtual course,” ESANN, pp. 26-28, 2006.

[13] Kajal C. Agrawal and MeghanaNagori, “Clusters of Ayurvedic Medicines Using Improved K-means Algorithm,” International Conf. on Advances in Computer Science and Electronics Engineering, 2013.

[14] Chew Li Sa; BtAbang Ibrahim, D.H.; DahlianaHossain, E.; bin Hossin, M., "Student performance analysis system (SPAS)," in Information and Communication Technology for The Muslim World (ICT4M), 2014 The 5th International Conference on , vol., no., pp.1-6, 17-18 Nov. 2014.

[15] Abdelghani Bellaachia, ErhanGuven, “Predicting Breast Cancer Survivability Using Data Mining Techniques,” Washington DC 20052, 2006.

[16] Daljit Kaur and Kiran Jyoti, “Enhancement in the Performance of K-means Algorithm,” International Journal of Computer Science and Communication Engineering, Volume 2 Issue 1, 2013.

[17] Siddheswar Ray and Rose H. Turi, “Determination of Number of Clusters in K-Means Clustering and Application in Colour Image Segmentation,” School of Computer Science and Software Engineering Monash University, Wellington Road, Clayton, Victoria, 3168, Australia, 1999.

[18] S. Chen and X. Liu, “An integrated approach for modeling learning patterns of students in web-based instruction: A cognitive style perspective,” ACM Trans. Comput. Interact., Vol. 15, No. 1, 2008.

[19] Rajee, A.M., and F. Sagayaraj Francis, “A Study on Outlier distance and SSE with multidimensional datasets in K-means clustering,” 2013 Fifth International Conference on Advanced Computing (ICoAC), 2013.

[20] Luhach, A.K., Dwivedi, S.K. and Jha, C.K., 2014, December. Applying SOA to an E-commerce system and designing a logical security framework for small and medium sized E-commerce based on SOA. In Computational Intelligence and Computing Research (ICCIC), 2014 IEEE International Conference on (pp. 1-6), IEEE.

[21] K. A. Abdul Nazeer, “Clustering Biological Data Using Enhanced k-Means Algorithm,” Lecture Notes in Electrical Engineering, 2010.