Improved K-mean Clustering Algorithm for Prediction Analysis using Classification Technique in Data Mining

Arpit Bansal
Department of Computer Science & Engineering
MIET Meerut, India

Mayur Sharma
Software Engineer
Meerut, Uttar Pradesh
India

Shalini Goel
Department of Computer Science & Engineering
MIET Meerut, India

ABSTRACT
Clustering is technique which is used to analyze the data in efficient manner and generate required information. To cluster the dataset, there is a technique named k-mean, is applied which is based on central point selection and calculation of Euclidian Distance. Here in k-mean, dataset will be loaded and from the dataset. Central points are selected using the formulae Euclidian distance and on the basis of Euclidian distance points are assigned to the clusters. The main disadvantage of k-mean is of accuracy, as in k-mean clustering user needs to define number of clusters. Because of user defined number of clusters, some points of the dataset are remained un-clustered. In this work, improvement in the k-mean clustering algorithm will be proposed which can define number of clusters automatically and assign required cluster to un-clustered points. The proposed improvement will leads to improvement in accuracy and reduce clustering time by the member assigned to the cluster to predict cancer.

General Terms
Clustering Algorithm, Predictive Analysis, Data Mining

Keywords
K-mean clustering, Prediction, clustering, Classification, Hierarchal clustering

1. INTRODUCTION
Data Mining is known as the process of analyzing data to extract interesting patterns and knowledge. Data mining is used for analysis purpose to analyze different type of data by using available data mining

tools. This information is currently used for wide range of applications like customer retention, education system, production control, healthcare, market basket analysis, manufacturing engineering, scientific discovery and decision making etc [1]. Data mining is studied for different databases like object-relational databases, relational database, data ware houses and multimedia databases etc.

Data mining is playing a vital role in many applications like market-basket analysis, etc. Frequent item sets have significant role in data mining which is used to find out the correlations between the fields of database [2]. Association

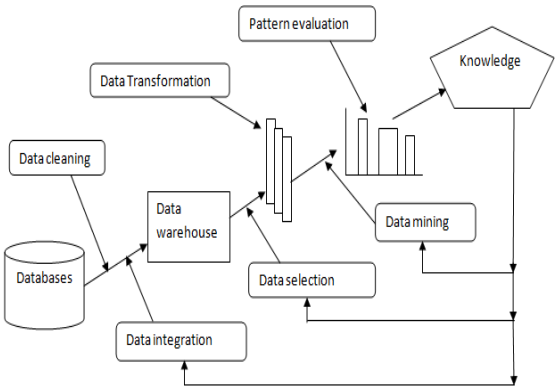

Figure 1: Data Mining Process

rule is based on discovering frequent item sets and frequently used by retail stores. Mining data in other words, named as Discovery of new knowledge in Databases which further moves to the nontrivial extraction of indirect, new and much required information from data in databases [3].

1.1 Clustering in Data Mining
Cluster analysis has been widely used in numerous applications, including market research, pattern recognition, data analysis, and image processing. In business, clustering can help marketers discover interests of their customers based on purchasing patterns and characterize groups of the customers. In biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionality, and gain insight into structures inherent in populations. In geology, specialist can employ clustering to identify areas of similar lands, similar houses in a city and etc. data clustering can also be helpful in classifying documents on the Web for information discovery.

Data clustering [4] 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. Now objects within a class have high resemblance to each other in the meantime objects in separate classes are more unlike.

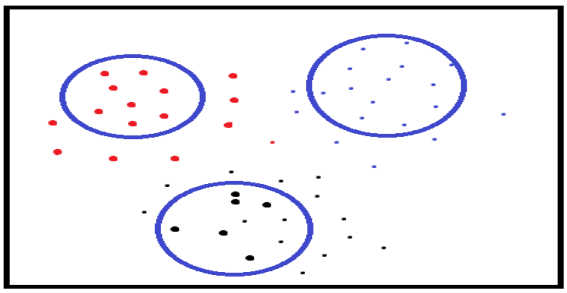

Figure 2: Clustering

1.2 Types of Clustering

There are various types of clustering in data mining.

1.2.1 Partitioning Clustering

The general criterion for partitioning is a combination of high similarity of the samples inside of clusters with high dissimilarity between distinct clusters. Most partitioning methods are distance-based. These clustering methods are work well for finding spherical –shaped clusters in small to medium size databases [5].

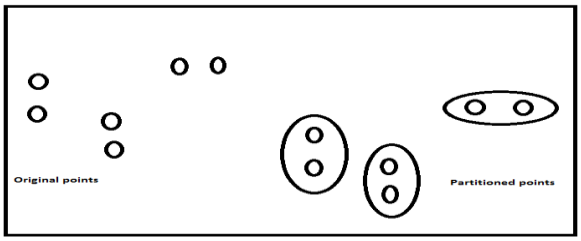

Figure 3: Partitioning Clustering

1.2.2 Density Based Clustering

Most partitioning methods cluster objects based on distance between objects. In these methods the cluster is continue to grow as long as the density in the neighborhood exceeds some threshold [6].

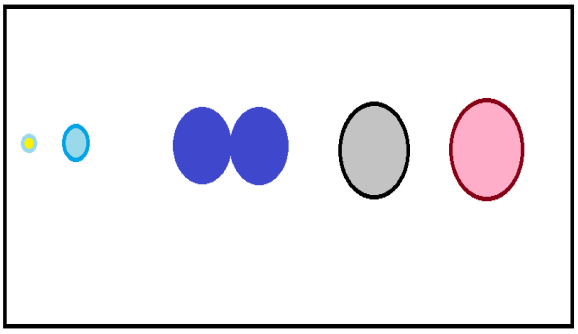

Figure 4: Density based Cluster

Grid Based Clustering

Grid based methods quantize the object space into a finite 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 [7].

1.2.3 Hierarchical Methods

In this method hierarchical decomposition of the given set of data objects is created. It can be classified as being either agglomerative or divisive based on how hierarchical decomposition is formed. Agglomerative approach is the bottom up approach starts with each object forming a separate group. Hierarchical algorithms create a hierarchical decomposition of the given data set of data objects. The hierarchical decomposition is represented by a tree structure, called dendrogram. It does not need clusters as inputs. In this type of clustering it is possible to view partitions at different level of granularities using different types of K. E.g. Flat Clustering [8].

It then merges groups close to one another until all the groups are merged into one. Divisive approach is top down approach starts with all the clusters in the same cluster and then in each iteration step a cluster is split into smaller clusters until each object is in one cluster.

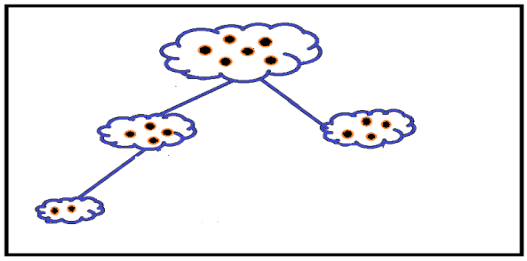

Figure 5: Hierarchal Clustering

2. REVIEW OF LITERATURE

In this paper [9] they explained 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. Real time data sets are always interlinked with some challenges like missing values, high dimensional values and noise etc which is not efficient for all the classification. Therefore clustering is the alternate solution for data analytics. The main focus of this paper is to develop a novel technique based upon foggy k-mean clustering. The result of the experiment depicts that foggy k-means clustering algorithm has excellent result on datasets which are real as compared to simple k-means clustering algorithm and provides a enhanced result to the real world problem.

In this paper [10] they explained that clustering is the powerful tool which used in various forecasting 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 uses typical K-means clustering on the main air pollution database and a list of weather category will be developed based on the peak mean values of the clusters. Whenever new data are coming, the incremental K-means is used to group data into those clusters where weather category has been already defined. Thus it is able to predict weather information of future. This forecasting database is totally based on the weather of west Bengal and this forecasting methodology is prepared to mitigate the consequences of air pollutions and launch focused modeling computations for prediction and forecasts of weather events. Here correctness of this approach is also measured.

In this paper [11] they proposed a system named Student Performance Analysis System (SPAS) to keep track 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 student performance prediction through the rules generated via data mining technique. The data mining technique used in this project is

classification, which classifies the students based on students' grade.

In this paper [12] they presented an analysis of the prediction of survivability rate of breast cancer patients using data mining techniques. SEER public datasets has been used in this project. This preprocessed dataset consists of 151, 886 records which have 16 fields from the SEER. After that they have investigated three data mining techniques: Naïve Bayes, back propagated neural network and C4.5 decision tree algorithm. Several 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 has a much better performance than other two techniques.

In this paper [13] they explained that forecasting stock return is one the important subject to be learn for prediction for data analysis. It is analysis that past investigations help to predict future in data analysis. In this paper they try to help 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.

3. BASIC THEORY

The k-means clustering algorithm [14] 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 ,$1 \leq i \leq$ n, to its nearest cluster centre Cj, $1 \leq j \leq k$, is minimized. First, the algorithm randomly selects the k objects, each of which initially represents a cluster mean or centre. Then, each object xi in the data set is assigned to the nearest cluster centre i.e. to the most similar centre [4]. The algorithm then computes the new mean for each cluster and reassigns each object to the nearest new centre. This process iterates until no changes occur to the assignment of objects.

3.1 Data Analytics

Data Analytics [15] is the science of examining raw data with the purpose of extract useful information to draw conclusions. Data Analytics can be used by many industries and organizations to get better business decision. Data analytics focuses on inference, the process of deriving a conclusion based solely on what is already known by the researcher. There are two types of data analytics: These are:

1. Classification

2. Prediction

1. Classification: Classification models predict categorical class labels; and prediction models predict continuous valued functions.

2. Prediction: Prediction model to predict the expenditures in dollars of potential customers on computer equipment given their income and occupation.

Prediction Analytics brings management, skills, information technology and modeling. Predictive Analytics is a data science, multidisciplinary skills set essential for non-profit organizations, success in business and government. Marketing forecasting sales or market share, good retail site opportunity has been found. It also identify consumer segments and target marketing and risks associated with existing products, predictive analytics provides the key for it.

Classification is a data mining technique which comes under machine learning technique to predict group membership for data instances. Classification technique is capable of processing a wider variety of data than regression and is growing in popularity.

4. PROPOSED METHODOLOGY

The k-mean clustering algorithm is used to cluster the similar type of data for prediction analysis. In k-mean clustering algorithm, probability of the most relevant function is calculated and using Euclidian distance formula the functions are clustered. In this work, we will enhance the Euclidian distance formula to increase the cluster quality. The enhancement will be based on normalization. In the enhancement two new features will be added. The first point is to calculate normal distance metrics on the basis of normalization. In second point the functions will be clustered on the basis of majority voting. The proposed technique will be implemented in MATLAB.

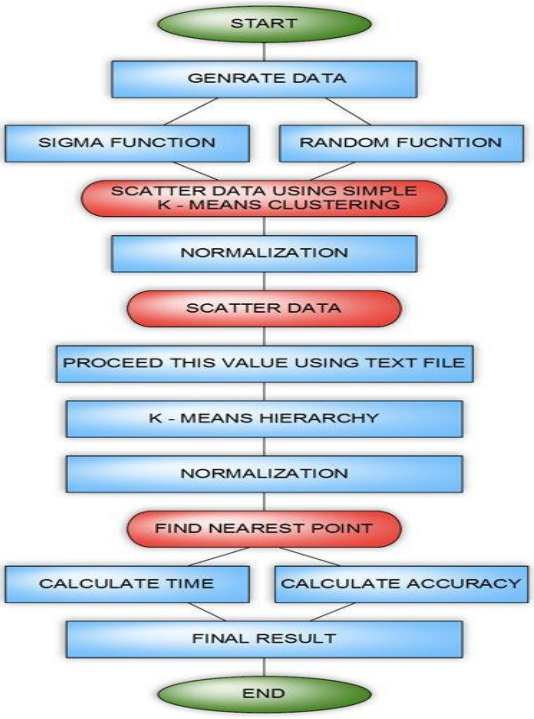

Figure 6 : Flowchart of Methodology

The steps of proposed work -

- First of all, we have started process in which at initial stage we generated data from user end in which we give number of data inputs, which are generated by sigma and random functions.
- When all data has been generated then Simple k-means applied and got result in subplot.
- After applying normalization on that data, we gave scatter data in second subplot.
- Now we applied normalization, in precede in which we read text file data of that generate data after that applied hierarchy k-means before normalization in which we got result in different form rather than first subplot.

- After this process normalization on that process is done in which iterations process started.
- This process is continued until we don't get a nearest point to accurate position with generating data.
- At last calculated their total time in which we got results which shows betterment in accuracy of cluster.

5. EXPERIMENTAL RESULTS

The Cancer Dataset is used for the research process and to find the results. The original data were highly dimensional, but only 5 attributes has been finally considered on the basis of requirements. The dataset is loaded using MATLAB.

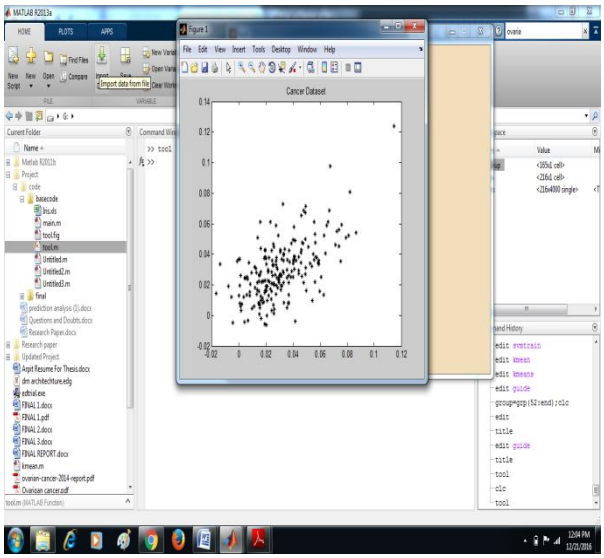

Figure 7: Dataset is Plotted

As shown in figure 7, The dataset is plotted on a 2D plane with according to existing K-means algorithm. Now the entire dataset can be analyzed in one single view. We can check how outliers does this dataset have . Thus we can put some controls on the data to minimize or reduce their effect for a better analysis and prediction.

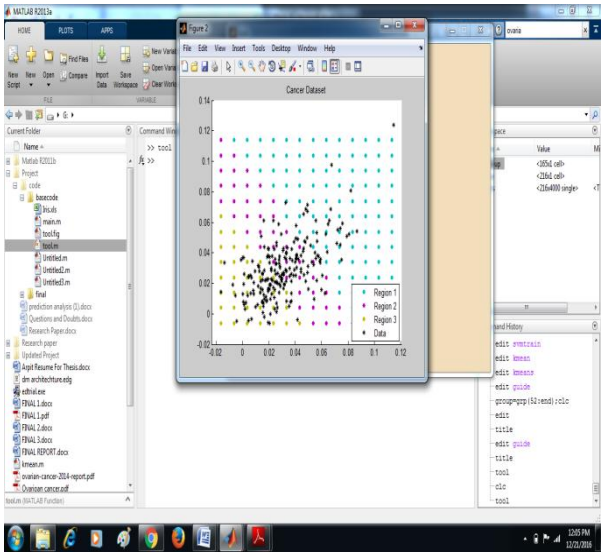

Figure 8 : Cluster using k-mean algorithm

As shown in figure 8, The regions have been divided with the existing k-mean clustering algorithm. Here the dataset is divided in three regions. All these regions are different from each other as they all depict something different from the other one.

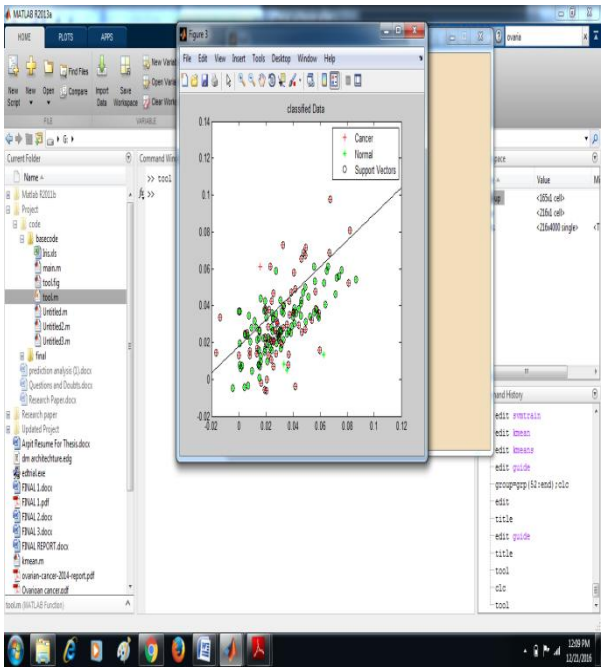

Figure 9: SVM is applied on the Dataset

As shown in figure 9, here in this the SVM is applied on the data for classifying the dataset.

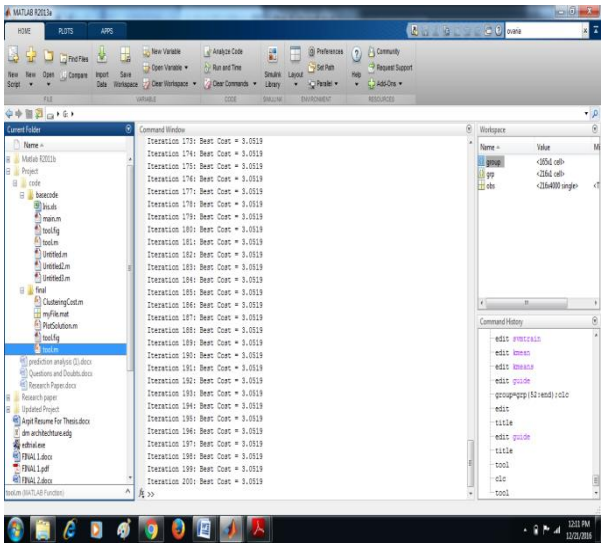

Figure 10 : Calculating Best cost using Normalization

As shown in figure 10, the k-mean clustering is improvement to improve cluster quality using the technique of normalization. The dataset is loaded is and it is shown on the command window. The dataset which is loaded is plotted on the 2-D plane for analysis. The plotted data will be clustered using the algorithm of k-mean clustering. The central points are marked in each cluster. The normalization technique is applied to calculate best distance to make high quality clusters.

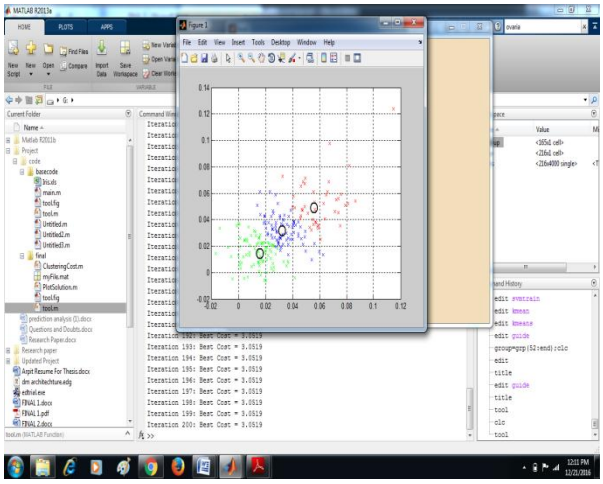

Figure 11: Central Point of Cluster

As shown in figure 11, The best calculated clusters have been formed along with their central points. All these 3 clusters are different from each other somehow. It depends on us that how many clusters we want to have in our output data. In this case we have define to divide the dataset in 3 separate clusters.

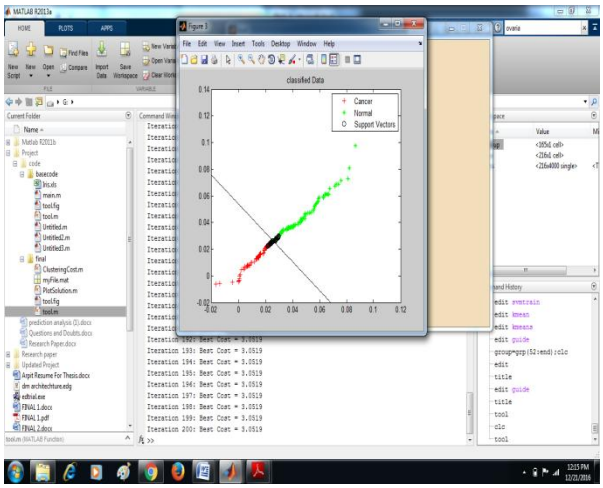

Figure 12 : Final output of k-mean algorithm

As shown in figure 12, The dataset has been divided into two separate regions. This graph plotted on a 2D plane, can be described easily as our technique has made it very easy to understand and predict after seeing the graph that the red dots that are below the line are indicting the cancerous and the green dots that above the line are non cancerous data.

5.1 Result

As it can be seen in the Table 1 shown below that there is a big variation between execution time of existing algorithm and proposed algorithm. And with this proposed algorithm the accuracy level gets almost double. So it can be said that the proposed modification in the existing k-means algorithm will give a huge improvement to the clustering techniques. The problem of accuracy level and execution time do not matter much, but when it comes to large dataset when there are millions of record, then it becomes enormously big problem. Because then the entire study of the dataset may move to a wrong direction as there was less accuracy level.

Table 1: Comparison with Existing Algorithm

Parameters	Existing Algorithm	Proposed Algorithm
Execution Time	6.90047 e-02	8.59160 e-02
Accuracy	57.14%	92.86%

6. CONCLUSION

In this paper, it is concluded that clustering is technique by which large datasets are divide in to small data collections that are called clusters. The cluster is a collection of the data that are turned into information. The data clusters are different from each others as they are possessing some different values from each other. There are number of algorithms that work well for clustering the data that can divide a dataset into clusters. In this paper we have proposed technique for a modification in K-Means Clustering Algorithm. Here in this proposed modification, the K-Means clustering will vanish off the two major drawbacks of K-Means clustering, that are accuracy level and calculation time consumed in clustering the dataset. Although when we use small datasets these two factors accuracy level and calculation time may not matter much but when we use large datasets that have trillions of records, then little dispersion in accuracy level will matter a lot and can lead to a disastrous situation, if not handled properly. So in last considering the whole of the situation, it can be said that this proposed modification can be more extended to achieve the full accuracy level upto 100%, with very little time and with more quality clusters.

7. ACKNOWLDEMENT

I am extremely grateful and indebted to my parents and my colleague for being pillars of strength , for their unfailing moral support, and encouragement. I treasure their blessings and good wishes and dedicate this study to them.

I thank one and all who have been instrumental in helping me to complete this dissertation work.

8. REFERENCES

[1] 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

[2] 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, o. 1, 2010

[3] Shital A. Raut and S. R. Sathe, "A Modified Fastmap K-Means Clustering Algorithm for Large Scale Gene Expression Datasets", *International Journal of Bioscience, Biochemistry and Bioinformatics*, Vol. 1, No. 4, page 120-124, November 2011.

[4] Daljit Kaur and Kiran Jyot, "Enhancement in the Performance of K-means Algorithm", International Journal of Computer Science and Communication Engineering, Volume 2 Issue 1, 2013

[5] 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

[6] 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

[7] 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

[8] 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.

[9] Akhilesh Kumar Yadav, Divya Tomar, Sonali Agarwal, "Clustering of Lung Cancer Data Using Foggy K-Means", International Conference on Recent Trends in Information Technology (ICRTIT) 2013

[10] Sanjay Chakrabotry, Prof. N.K Nigwani and Lop Dey "Weather Forecasting using Incremental K-means Clustering", 2014

[11] Chew Li Sa; Bt Abang Ibrahim, D.H.; Dahliana Hossain, 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

[12] Abdelghani Bellaachia, Erhan Guven, "Predicting Breast Cancer Survivability Using Data Mining Techniques", Washington DC 20052, 2010

[13] Qasem a. Al-Radaideh, Adel Abu Assaf 3eman Alnagi, " Predictiong Stock Prices Using Data Mining Techniques", The International Arab Conference on Information Technology (ACIT'2013)

[14] K. A. Abdul Nazeer, M. P. Sebastian, "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm, Vol IWCE 2009, July 1 - 3, 2009, London, U.K