International Journal of Engineering and Techniques - Volume 3 Issue 3, May-June 2017

RESEARCH ARTICLE OPEN ACCESS

Modified K-Means Clustering Algorithm for Disease Prediction

Dr.Sandeep Kumar[1], Simarpreet Kaur[2]

[1]Associate professor & HOD, Department of Computer Science & Engineering, Galaxy Global Group of Institutions, Dinarpur, Ambala, Haryana, India

[2]PG Student, Department of Computer Science & Engineering, Galaxy Global Group of Institutions, Dinarpur, Ambala, Haryana, India

Abstract:

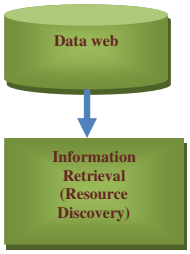
This work presents the outline of K-means clustering algorithm and enhanced technique applied on K-means clustering. The K-means clustering is the basic algorithm to find the groups of data or clusters in the dataset. To find the similar groups of data the initial selection of cetroid is done and the Euclidean distance is calculated from cetroid to all other data points, and based on the smaller Euclidean distance the data points are assigned to that centroid. 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 are improved. So, to improve the performance of clusters the Normalization which is a pre-processing stage is used to enhance the Euclidean distance by calculating more nearer centers, which result in reduced number of iterations which will reduce the computational time as compared to k-means clustering. By applying this enhanced technique one can build a new proposed algorithm which will be more efficient, supports faster data retrieval from databases, makes the data suitable for analysis and prediction, accurate and less time consuming than previous work.

Keywords — **K-means clustering, Euclidean distance, Normalization, centroid.**

. I. INTRODUCTION

Data Mining is known as the process of extraction of knowledge or useful patterns from the unorganized and huge data. The goal of data mining is to analyze different type of data by using available data mining tools. The steps in data mining are: Data cleaning, Data Integration, data selection, data transformation, data mining, pattern evaluation and knowledge representation.

Data mining plays an important role in medical for prediction of disease. There are high risky consequences because of doctor's assumptions and lack of knowledge in particular area. Data mining here plays an important role in discovering or deriving out useful patterns from the historic data of patients, from these patterns prediction analysis for future aspects can be done [1]. Data Mining is the process of analyzing the huge amount of data and encapsulating the relevant information from it. In other words, we can say that data mining is the procedure of mining knowledge from data. The information extracted can be used for any of the following fields of application –



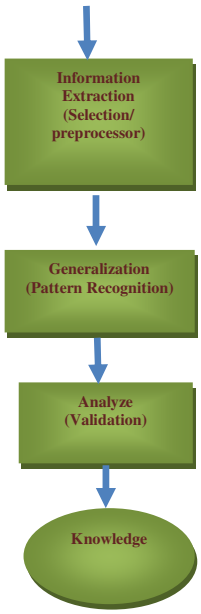


Fig.1 Data Mining Process

- Market Analysis
- Intrusion Detection
- Bio-Informatics
- Production Control
- Telecommunication industry
- CRM(Customer Relationship Management)
- Fraud Detection
- Scientific Applications

II. KDD PROCES

KDD is the automatic extraction of non-obvious, hidden knowledge from large volumes of data [2]. It is the process of finding the knowledge in data, discovering useful knowledge from the collection of data. It mainly emphasizes on the high level application of particular data mining methods. Some people don't differentiate data mining from knowledge discovery while others view data mining as an essential step in the process of knowledge discovery. KDD refers to the overall process of discovering useful knowledge from data. It involves the evaluation and possibly interpretation of the patterns to make the decision of what qualifies as knowledge. It also includes the choice of encoding schemes, preprocessing, sampling, and projections of the data prior to the data mining step.

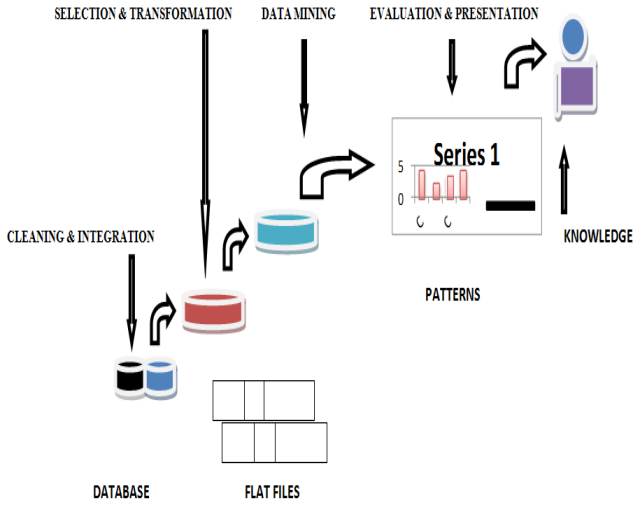


Fig.2 KDD Process

III. CLASSIFICATION AND PREDICTION IN DATA MINING

A. Classification: It is the process of finding a model that describes the data classes or concepts. It describes labeled data and unlabeled data, labeled data (training data) which works on given historical attributes and class labels, from the result obtained from training data, prediction of unlabeled data is done and its class label is predicted. The simplest type of classification problem is binary classification. In binary classification, the target attribute has only two possible values: for example, high credit rating or low credit rating. Multiclass targets have more than two values: for example, low, medium, high, or unknown credit rating. The data classification is a two step process:

- Building the classifier or Model
- Using the Classifier for Classification

1. Building the classifier:

- The training data is given which consist of database tuples and their associated class labels.
- Each tuple that constitute the training set is also referred to as class.
- By analyzing the training data, the classifier rules are built by classification algorithm.

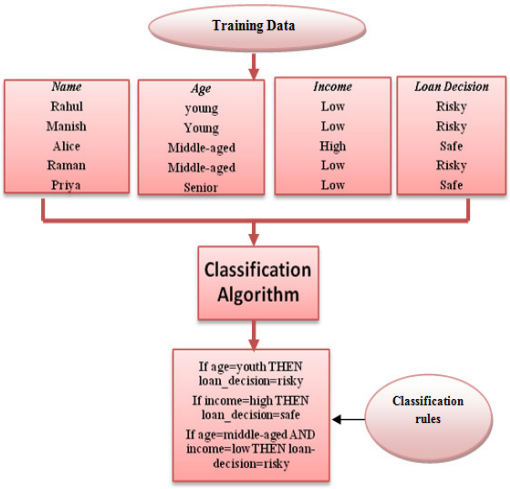


Fig.3 Building the Classifier

2. Using the classifier for classification: In this the accuracy of classification rules is estimated by using the test data. The classification rules are now applied on the new data tuples and here class label will be predicted by analyzing the classification rules.

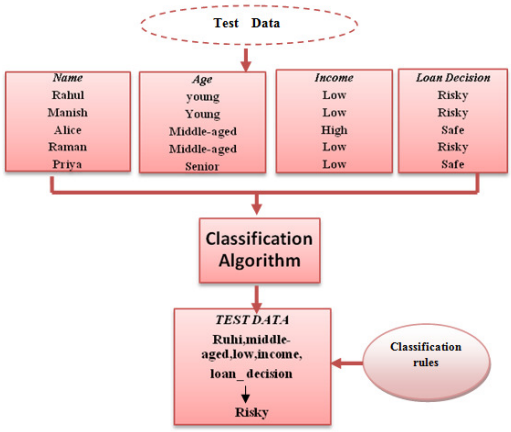


Fig.4 Using the Classifier

B. Prediction: Prediction model is used to predict the expenditures in dollars of potential customers on computer equipment by giving their income and occupation and predict the expected expenditures of different occupations. Prediction Analytics brings the management, skills, information technology and modeling. Predictive Analytics is a science multidisciplinary to the skills set essential for non-profit organizations and for the success in business and government organizations. Marketing forecasting sales or market share, good retail site opportunity has been found. It also identifies consumer segments and target marketing strategies and risks associated with existing products; predictive analytics provides the key note for it.

IV. K-MEANS CLUSTERING

K-means is an unsupervised learning algorithm which is used to classify the given dataset that is unlabeled. The goal of this algorithm is to find similar groups represented by variable k. Here k is the number of clusters, so k centroids are defined one for each cluster. Now, the Euclidean distance is calculated from each data point to the centroid, assignment of data points to the centroid depends upon the minimum Euclidean from that centroid. When no point is left un-assigned, an early grouping is done. Now, k new centroids are re-calculated, as a result iteration continues until the k centroids stop changing their position.

Let $Y = \{x_1, x_2, x_3, \ldots\ldots, x_n\}$ be the set of data points and $Z = \{z1, z2, \ldots\ldots, z_c\}$ be the set of centres.

1. Randomly select *'c'* cluster centres.
2. Calculate the distance between each data point and cluster centres.
3. Assign the data point to the cluster centre whose distance from the cluster centre is minimum of all the cluster centres.
4. Recalculate the new cluster centre using:

$$Z_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

5. Where, 'c_i' represents the number of data points in i^{th} cluster.
6. Recalculate the distance between each data point and new obtained cluster centres.
7. If no data point was reassigned then stop, otherwise repeat from step 3.

Advantages

1. This algorithm is fast, scalable and efficient in processing large datasets.
2. Results are easily understandable [3].

Disadvantages

1. Unable to handle outliers and noisy data.
2. Depends upon the initial cluster centroid.
3. Better results are not found due to randomly chosen number of clusters k.

III. LITERATURE REVIEW

In this paper [4], authors proposed the accuracy of k-means clustering technique for predicting the heart disease with real and artificial dataset with the existing method. Clustering is the method of cluster analysis which aims to group the partition into k clusters and each group or cluster has its observations with nearest mean. Each cluster assigned to the cluster k and started from random initialization. The above mentioned technique further divided into k groups. The grouping is done by minimizing the sum of squares of distances between data using Euclidean distance formula and the corresponding cluster centroid. The research result show that the integration of clustering gives promising results with highest accuracy rate and robustness.

In this paper [5], authors proposed the possibility for fastest searching or reading single paper in less time. As users have to spend more time in reading and searching for a single paper so, in this paper author's use enhanced search engine which is based on fastest reading algorithms, consumes less time and give best output and results. The Enhanced K-means algorithm with enhanced architecture is proposed for

making the algorithm more efficient and effective; to get good clusters with reduced complexity. The proposed method will search the base keyword content from the huge database. The search engine will be based on text mining and clustering.

In this paper [6], authors proposed k-means clustering algorithm for predicting Myocardial Infarction (MI). Myocardial Infarction is common disease in the world. This system extracts hidden information from historical datasets of heart disease. In this paper, number of input attributes is used for analyzing prediction systems for heart disease. Efficiency of output is increased by using k-means clustering. This method for predicting patients with heart disease is the most effective one.

In this paper [7], authors proposed the algorithm for analysing the behaviour of internet users based on log data network which is big data at an educational institution. The algorithm which is used is k-means clustering to determine the internet user's behaviour. The K-means is used for clustering based on the no. of visitors which are divided into low, medium and high amount of data. The results of the educational institution show that each of these clusters produces websites that are searched by the sequence: website search, social media, news, and information. This paper also results in determining which website have high amount of traffic in the sequence of searching.

In this paper [8], authors proposed the possibility of the student performance of high learning. 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 combined with deterministic model to analyze student's performance of the system.

In this paper [9], authors proposed explored the K-means clustering algorithm for prediction of various diseases(heart, liver, diabetics and cancer).The comparison of disease, data mining techniques and accuracy is done, in which k-means algorithm predict more accurate than all other techniques. The k-means clustering method for prediction reduces the human effects and is cost effective one.

IV. PROPOSED RESEARCH METHDOLOGY

In k-mean clustering algorithm, the goal is to find groups of data (data is unlabeled) and after that functions are clustered based on feature similarity by using Euclidian distance formula.

In this paper, the quality of clusters is increased by enhancing the Euclidian distance formula. The enhancement that has to done will be based on normalization. Normalization which is a pre-processing technique will enhance the accuracy and efficiency of clusters by calculating best distances from the dataset which will result in more accurate center points and as a result best clusters are formed, the feature which is added is for calculating normal distance metrics on the basis of normalization. The proposed technique is implemented in MATLAB.

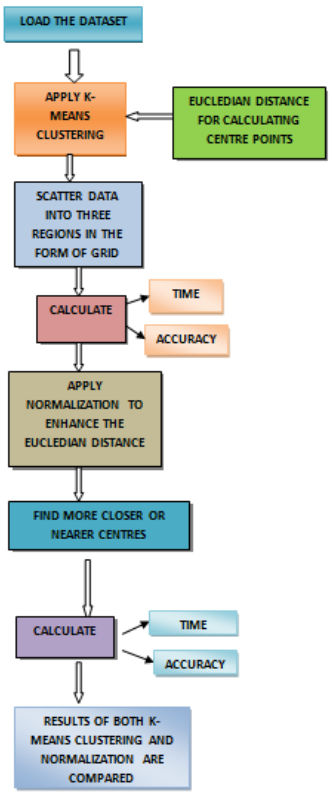


Fig.5 Flowchart of Research Methodology

Working:

1. Firstly we have generated or loaded the user-defined dataset.
2. Dataset is scattered and plotted.
3. Now, k-means clustering is applied on the generated dataset in which Euclidean distance formula is used for calculating centroids.
4. After applying k-means clustering time and accuracy of centers is calculated.
5. For classifying the dataset SVM is applied.
6. Now, after applying Normalization on the data in which iterations process started, and more nearest and accurate centers are calculated.

7. **Normalization:** It is scaling technique or a pre-processing stage. Where, we can find new range from an existing one range. It can be helpful for the prediction or forecasting purpose.
   Min-Max Normalization: Min-Max Normalization transforms a value A to B which fits in the range [C,D]. It is given by the formula below:
   B = ((A – minimum value of A) / maximum value of A- minimum value of A)*(D-C) + C
   A=Original data point
   B=Normalized data point
   [C, D]= specified range.
8. Data is plotted after applying Normalization.
9. Again time and accuracy is calculated in which we got better accuracy of clusters than k-means clustering.

V. EXPERIMENTAL RESULTS

In this paper HIV dataset is used for research process and prediction analysis. The dataset is plotted between two attributes of HIV that is: CD4 count and Viral Load. CD4 cells are a type of white blood cell. They are specialized cells of the immune system that are destroyed by HIV.

A CD4 count measures how many CD4 cells are in your blood. The higher your CD4 cell count, the healthier your immune system. Viral Load Test that measures the number of HIV virus particles in a milliliter of our blood. These are called' copies' it helps prior information on health status and how well antiretroviras therapy is working.

A. Dataset is plotted:

The scattered HIV dataset is loaded using MATLAB as shown in Fig.6.

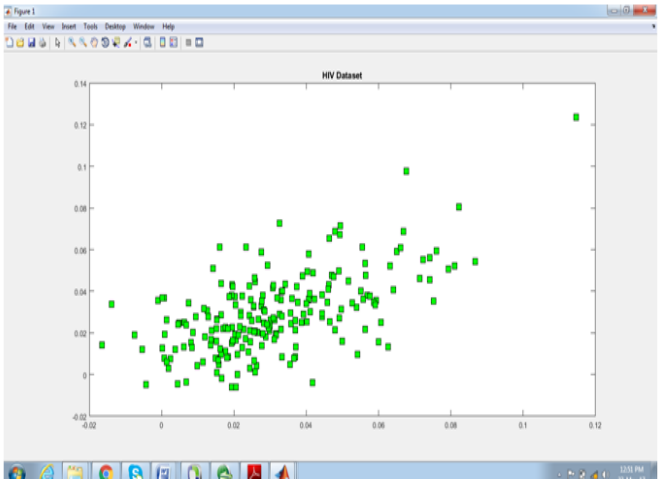


Fig.6 Dataset is plotted

The 2-d dataset is plotted in MATLAB.

B. K-means clustering is applied on the dataset:

Now the k-means clustering is applied on the given dataset with k=3. The dataset is divided into 3 regions. The regions are in the form of grid, and are divided with the existing k-means clustering algorithm. In k-means we randomly initialize centroids from the dataset and then Euclidean distance is calculated of each data point from centroids and depending upon the minimum distance between the centroids and data points, that data point are assigned to that centroids, and repeat again these steps till we get the same centroids that is no change in the centroids. In this way, three clusters are formed and in the Fig.7 the clusters are divided in the form of regions. Each region depicts different nature.

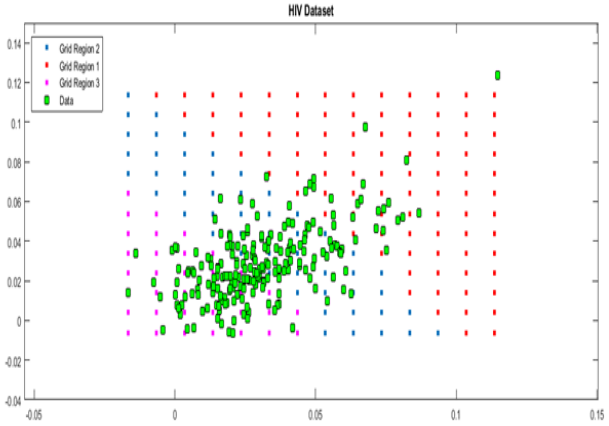


Fig.7 K-means clustering on data

C. Classifier is applied on the dataset:

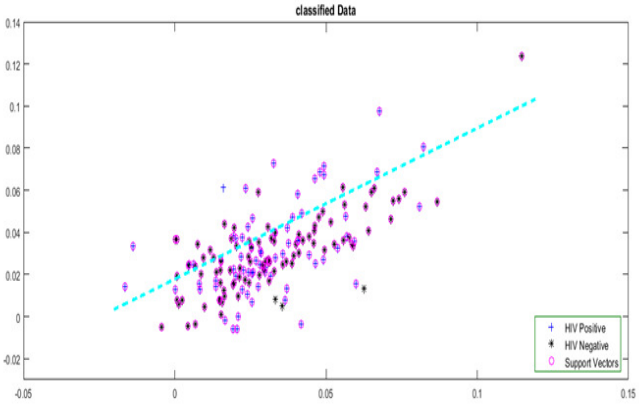


Fig.8 Dataset is classified

The SVM is applied on the dataset in a 2-d plot. The support vector machine is better because when you get a new sample or new data points, the line that is separation line(hyper plane) has already made that keeps the two groups(HIV Positive & HIV Negative) or clusters far away from each other as possible. So it will be easy place data-point in the particular cluster.

D. Normalization is applied on the dataset:

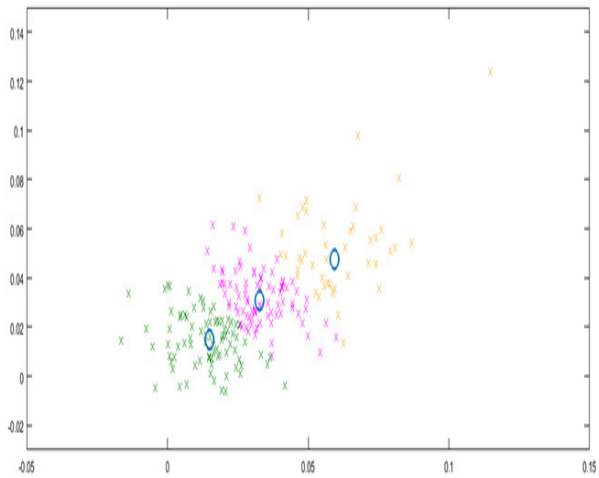


Fig.9 Normalized dataset

Normalization is applied on the dataset. While performing K-means clustering Euclidean distance is calculated for forming similar groups or clusters. But the clusters which we get are not accurate, here the Euclidian distance lacks , so Normalization(Min-Max) in which some range is specified is applied on the dataset for calculating best distance for more nearer centers and best clusters and best cost is calculated according to the number of iterations. We have taken here 3 clusters, which are dataset is divided into three groups. It depends upon us how many clusters we want in our output.

E. Graph is plotted between Best cost and Iterations:

Bar graph has been plotted on a 2D plane in Fig. 10, between best cost and iterations; best cost is calculated using normalization for analysis of Cluster Quality Analysis. From the graph we can see that best cost is declining and then become constant, it means that when there will be no change in the centres during normalization i.e. best centres has been calculated. After analyzing the graph, we can say that the quality of the third cluster s low.

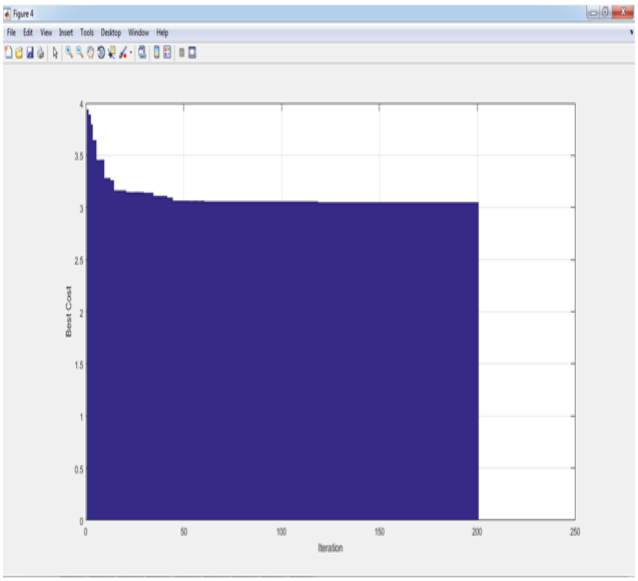


Fig.10 Calculation of Best Cost

VI. COMPARISON OF EXISTING ALGORITHM WITH PROPOSED ALGORITHM

Table: 1 Comparison with existing algorithm

Algorithm	Existing	Proposed
Efficiency	74.3%	95.6%
Time	8.4 sec	10.5sec

VII. CONCLUSION

The prediction analysis is the technique in which user predicts the future on the basis of current situations. The prediction analysis consists of two steps. The first step is of clustering, which will cluster the similar and dissimilar type of data. The second step consists of classification which will classify the clustered data for the prediction analysis. In this work, K-mean clustering is used for the clustering. The SVM classifier is used to classify the dataset for predicting the complex data. The k-mean clustering consists of two steps. In the first step, the arithmetic mean of the loaded dataset is calculated which will be the centroid point. In the second step, Euclidean distance from the centroid point is calculated which defines similarity between the data points. The accuracy of clustering and classification is reduced when

some points remain unclustered or wrongly clustered. In this work, technique of Normalization is being applied which will reduce the complexity of large datasets, will calculate Euclidean distance in the dynamic manner and retain maximum accuracy as normalization works better when dataset are complex in nature. Normalized results make the data suitable for specific analysis and prediction to be performed. . The proposed improvement leads to increase accuracy of classification. The proposed improvement and existing technique is being implemented in MATLAB and it is being analyzed that accuracy is increased, execution time is reduced.

VIII. FUTURE SCOPE

Following are the various possibilities which can be done.

1. The proposed scheme can be compared with other techniques of prediction analysis to check reliability.
2. In the proposed scheme SVM classifier is used for classification which has high complexity. It can be replaced with some other classifier like naive biased, to achieve maximum accuracy with minimum complexity.

REFERENCES

[1] Purvashi Mahajan, Abhishek Sharma, "Role Of K-means Algorithm in Disease Prediction", International Journal of Engineering And Computer Science, ISSN: 2319-7242, Vol.5, Issue 4, 2016, pp.16216-16217.

[2] Nainja Rikhi, "Data Mining and Knowledge Discovery in Database", International Journal of Engineering Trends and Technology, Vol.23, No.2, 2015.

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

[4] Bala Sundar V,T Devi, N Savan,"Development of a Data Clustering Algorithm for Predicting Heart", International Journal of Computer Applications (0975 – 888) Vol.48–No.7,2012.

[5] Sachin Shinde, Bharat Tidke, "Improved K-means Algorithm for searching Research Papers", International Journal of Computer Science & Communication networks, ISSN: 2249-5789, Vol.4 (6), 197-202.

[6] M.Umamaheswari, Dr. P. Isakki @Devi, "Myocardial Infarction Prediction using K-means Clustering Algorithm", International Journal of Innovative Research in Computer and Communication, Vol. 5, Special Issue 1, March 2017.

[7] Muhammad Zulfadhilah, Imam Riadi, Yudi Prayudi, "Log Classification using K-means Clustering for Identify Internet User Behaviours", International Journal of Compiler Applications, (0975-8887), Vol.154-No.3,November 2016.

[8] 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.

[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), Vol. 4, Issue 7, July 2015.

[10] BV Sumana, T.Santhanam,"Prediction of diseases by Cascading Clustering and Classification",International Conference on Advances in Electronics , Computers and Communication(ICAECC),2014.

[11] Abhay Kumar, Ramnish Sinha, Vandana Bhattacherjee, Daya Shankar Verma, Satinder Singh,'Modelling using K-means clustering algorithm", Ist Int'1 Conf. on Recent Advances in Information Technology(RAIT), 2012.

[12] G. Kesavaraj, Dr. S.Sukumaran,"A Study on Classification Techniques in Data Mining",IEEE-31661.

[13] F.U.Siddiqui, N.A.Mat Isa, "Optimized k-means clustering algorithm for image segmentation", School of Electrical and electronic engineering, university Sains Malaysia, 14300, Nibong Tebel, Penang, Malaysia, 2012.

[14] 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.

[15] Tapas kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, Angela Y.Wu,"An Efficient k-Means Clustering Algorithm: Analysis and Implememtation", IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol.24, No.7,2002.

[16] Kapil Joshi, Himanshu Gupta, Prashant Chaudhary, Punit Sharma," Survey on Different enhanced K-means Clustering Algorithm", International Journal Of Engineering Trends And Technology, Vol. 27 ,No. 4-September 2015.

[17] Adil Fahad, Najlaa Alshatri, Zahir Tari, Addullah Alamri, Ibrahim Khalil, albert Y.Zomaya, Sebti Foufou, Abdelaziz Bouras, "A survey of Clustering Algorithms for Big Data: Taxonomy And empirical Analysis", IEEE TTANSACTION ON Emerging Topics in Computing.

[18] Jiawei Han, Micheline kamber, Jian pei, "Data Mining Concepts and Techniques", Third Edition, © 2012, Elsevier Inc.