ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

Prediction of Heart Disease using Modified K-means and by using Naive Bayes

Sairabi H. Mujawar[1], P. R. Devale[2]

P.G. Student, Dept. of Information Technology, B. V. D. U. College of Engineering, Pune, India[1]

Professor, Pursuing Ph.D., Dept. of Information Technology, B. V. D. U. College of Engineering, Pune, India[2]

ABSTRACT: In medical sciences prediction of Heart disease is most difficult task. In India, main causes of Death is due to Heart Diseases. The deaths due to heart disease in many countries occur due to work overload, mental stress and many other problems. It is found as main reason in adults is due to heart disease. Thus, for detecting heart disease of a patient, there arises a need to develop a decision support system. Data Mining classification techniques, namely Naive Bayes and Modified K-means are analyzed on Heart Disease is proposed in this Paper.

KEYWORDS: Data Mining, Heart Disease, Naive Bayes, Decision Support, Modified K-Means

I. INTRODUCTION

In this world people want to live a very luxurious life so they work like a machine in order to earn lot of wealth. At very young age, this type of lifestyle doesn't take rest for themselves, which results in diabetics and blood pressure. It is a world known fact that heart is the most essential part in human body if that heart gets affected then it also affects the other parts of the body. Therefore it is essential for people to go for a heart disease diagnosis. People go to healthcare checkup but the prediction made by them is not 100% accurate.

Today, healthcare industry generates large amount of data about patients, disease diagnosis etc. Diagnosis is important task and complicated that needs to be executed accurately and efficiently. Based on doctor's experience & knowledge, the diagnosis is often made. This leads to unwanted results & excessive medical costs of treatments provided to patients. Quality of service is a major challenge facing Healthcare industry. Quality of service guarantee diagnosing disease correctly & to provides effective treatments to patients.

II. RELATED wORK

To have focus on diagnosis of heart disease different studies have been done. Different data mining techniques has been used by them for diagnosis & achieved different probabilities for different methods. Using data mining techniques an Intelligent Heart Disease Prediction System (IHDPS) is developed. Sellappan Palaniappan et al [14] proposed Naive Bayes, Neural Network, and Decision Trees. For appropriate results each method has its own strength. Hidden patterns and relationship between them is used to build this system. It is user friendly, expandable & web-based.

Niti Guru et al [7] proposed the prediction of Blood Pressure, Sugar and Heart disease with the aid of neural networks. The records of 13 attributes in each was used in the dataset. For training and testing of data, the supervised networks i.e. Neural Network with back propagation algorithm is used.

Heon Gyu Lee et al. [5] proposed a novel technique, to develop the multi-parametric feature with linear and nonlinear characteristics of HRV (Heart Rate Variability) Several classifiers e.g. Bayesian Classifiers, CMAR (Classification based on Multiple Association Rules), C4.5 (Decision Tree) and SVM (Support Vector Machine) has been used by them.

To measure the impurity of a partition or set of training tuples [2], CART uses Gini index. High dimensional categorical data can handled by it.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

III. HEART DISEASE

The heart is important part of our body. Our life is totally dependent on efficient working of heart. If operation of heart is not proper, it will affect the other parts of body such as kidney, brain, etc. It is nothing more than a pump, which pumps blood through the body. Death occurs within minutes, if circulation of blood is inefficient. The term Heart disease refers to blood vessel system within it and disease of heart.

There are number of factors which increase the risk of Heart disease[4].

- *Smoking*
- *Family history of heart disease*
- *Poor diet*
- *High Blood Pressure*
- *Physical inactivity*
- *Hyper tension*
- *Obesity*
- *Cholesterol*

Factors like these are used to analyze the Heart disease. In many cases, diagnosis is generally based on doctor's experience and patient's current test results. Thus the diagnosis is a complex task that requires much experience & high skill.

IV. DATA SOURCE

Dataset with input attributes is obtained from Cleveland Heart Disease database. With the help of recordset, the heart attack prediction with significant patterns are extracted. The attribute "Diagnosis" with value "1" is identified as Heart Disease prediction and value "0"is identified as no Heart disease prediction for patients. Here key attribute is "PatientId" and other attributes are used as input.

Predictable attribute

1. Diagnosis (value 0: <50% diameter narrowing (no heart disease); value 1: >50% diameter narrowing (has heart disease))

Key attribute

1. PatientId – Patient's identification number

Input attributes

1. Sex (value 1: Male; value 0: Female)
2. Age in Year
3. Oldpeak – ST depression induced by exercise
4. Restecg – resting electrographic results (value 0:normal; value 1: having ST-T wave abnormality; value 2: showing probable or definite left ventricular hypertrophy)
5. Fasting Blood Sugar (value 1: >120 mg/dl; value 0: <120 mg/dl)
6. Slope – the slope of the peak exercise ST segment (value 1:unsloping; value 2: flat; value 3: downsloping)
7. Exang - exercise induced angina (value 1: yes; value 0: no)
8. Serum Cholesterol (mg/dl)
9. Trest Blood Pressure (mm Hg on admission to the hospital)
10. Thal (value 3: normal; value 6: fixed defect; value 7: reversible defect)
11. CA – number of major vessels colored by fluoroscopy (value 0-3)
12. Thalach – maximum heart rate achieved
13. Chest Pain Type (value 1:typical type 1 angina, value 2: typical type angina, value 3:non-angina pain; value 4: asymptomatic)

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

IV. PROPOSED ALGORITHM

Today, many hospitals manage healthcare data using healthcare information system; as this system contains huge amount of data, and it is used to extract hidden information for medical diagnosis. The main objective of this system is to build Heart Disease Prediction System using historical heart database that gives diagnosis of heart disease. To build this system, medical terms such as blood pressure ,sex, cholesterol, sugar etc 13 input attributes are used. Data Mining techniques such as clustering, Classification are used in extracting knowledge from database.

A. *Modified K-means*:

The proposed modified algorithm proves to be a better method to determine the initial centroids and it is easy to implement. By eliminating one of its drawbacks, this modified K-means tries to enhance the k means clustering algorithm. K-means was used to apply on numerical data only. But, we encounter both numerical and categorical combination data values.

This algorithm does not require number of clusters(k) as input is described below. By choosing two initial centroids, two clusters are created initially, which are farthest apart in the datasets. It can create two clusters with the data members at the initial steps, which are most dissimilar ones.

Input:
D: The set of n tuples with attributes Al,A2, . . . , Am. All attributes are numeric, (where m = no. of attributes)

Output:
With n tuples suitable number of clusters distributed properly

Method:
1) To find the points in the data set which are farthest apart, compute sum of the attribute values of each tuple
2) As initial centroids take tuples with maximum and minimum values of the sum.
3) Using Euclidean distance create initial partitions (clusters) between the initial centroids and every tuple
4) From the centroid find distance of every tuple in both the initial partitions. Take other than zero. d=minimum of all distances.
5)) For the partitions created in step 3,compute new means (centroids)
6) From the new means (cluster centers) compute Euclidean distance of every tuple.
and depending on the following objective function, find the outliers: If Distance of the tuple from the cluster mean>d then only it is an Outlier.
7) New centroids of the clusters can be computed
8) From the new cluster centroids ,calculate Euclidean distance of every outlier and find the objective function in step 6. outliers is not satisfying
9) Let the set of outliers obtained in step 8 is B={ Yl,Y2,. ....Y p} (Where value of k is depends on number of outliers).
10)Repeat the steps until I (B==<D)
  a) By taking mean value of its members as centroid, create a new cluster for the set B,
  b) Depending on the objective function in step 6, find the outliers of this cluster,
  c) Check if no.of outliers = p then
    i) Test every other outlier for the objective function as in step 6 after creation of a new cluster with one of the outliers as its member
    ii) If there is any outfliers find it
  d) From the centroid of the existing clusters, calculate the distance of every outlier. If the existing clusters which satisfy the objective function in step 6. then adjust the outliers
  e) The new set of outliers be B={ ZI,Z2 .... Zq}.(Where value of q is depends on number of outliers)

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

B. ***Naive Bayes Algorithm:***

Naive Bayes which is also called as Bayes' Rule is the basis for data mining methods and machine-learning.It creates a model with predictive capabilities. It provides new ways of understanding data and exploring it.

Bayes Rule

A conditional probability is the likelihood of some conclusion, *C*, given some evidence/observation, *E*, where a dependence relationship exists between *C* and *E*.
This probability is denoted as P*(C |E)* where

$$P(C \mid E) = \frac{P(E \mid C)P(C)}{P(E)}$$

Naive Bayesian Algorithm

1. Consider D be a training set of tuples and their associated class labels. Each tuple is represented by an n-dimensional attribute vector, X=(x1, x2,…, xn), depicting n measurements made on the tuple from n attributes, respectively, A1, A2,.., An.
2. maximum posterior hypothesis. By Bayes' theorem :
Let there are m classes, C1, C2,…, Cm. Given a tuple, X, Conditioned on X, the classifier will predict that X belongs to the class having the highest posterior probability.
The prediction of naïve Bayesian classifier that if and only if
P (Ci|X)>P (Cj|X) for $1\leq j\leq m$, $j \neq i$ then tuple x belongs to the class Ci.
As we maximize P(Ci|X), The class Ci for which P(Ci|X) is maximized is called the maximum posterior hypothesis. By Bayes' theorem

$$P(Ci \mid X) = \frac{P(X \mid Ci)P(Ci)}{P(X)}$$

3. Only P (X|Ci) P (Ci) need be maximized, As P(X) is constant for all classes, If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, P(C1)=P(C2) =…=P(Cm), and we would therefore maximize P(X|Ci). Otherwise, we maximize P(X|Ci)P(Ci).
Recall that the class prior probabilities may be estimated by P(Ci)=|Ci,D|/|D|**,** where |Ci,D| is the number of training tuples of class Ci in D.
4. It would be expensive to compute P(X|Ci) for given data sets with many attributes.
The naïve assumption of class conditional independence is made in order to reduce computation in evaluating P(X|Ci).
This means that there are no dependence relationships among the attributes.
Thus,

$$P(X \mid Ci) = \prod_{k=1}^{n} P(xk \mid Ci)$$

=P(x1|Ci)x P(x2|Ci)x… P(xm|Ci).
From the training tuples, we can easily estimate the probabilities P(x1|Ci), P(x2|Ci),… ,P(xm|Ci)
where xk refers to the value of attribute Ak for tuple X.

Suppose to compute P(X|Ci), we consider the following:
(a) If Ak is categorical, then P(Xk|Ci) is the number of tuples of class Ci in D having the value xk for Ak, divided by |Ci,D|**,** the number of tuples of class Ci in D.
(b) If Ak is continuous valued, then it is typically assumed to have a Gaussian distribution with a mean μ and standard deviation σ, defined by

$$g(x,\mu,\sigma) = \frac{1}{\sqrt{2\Pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

So that
$P(xk|Ci)=g(xk, \mu ci, \sigma ci)$
We need to compute μci and σci, which are the mean and standard deviation, of the values of attribute Ak for training tuples of class Ci.
5. P(X|Ci)P(Ci) is evaluated for each class Ci. in order to predict the class label of X.
if and only if
$P(X|Ci)P(Ci)>P(X|Cj)P(Cj)$ for $1 \le j \le m, j \ne i$
Then the classifier predicts that the class label of tuple X is the class Ci
In other words, for P(X|Ci)P(Ci) is the maximum if the predicted class label is the class Ci.

V. EXPERIMENT RESULTS

A total of 300 records with 13 attributes were used from the Cleveland Heart database[1].User enter values in medical attributes like sex, age, slope, blood pressure, thal, chest pain, resting ECG etc. This model predicts that patient is having heart disease or not depending on this values, doctors would recommend to go for further heart examination.

Fig 1.are used to load UCI repository dataset for training purpose.

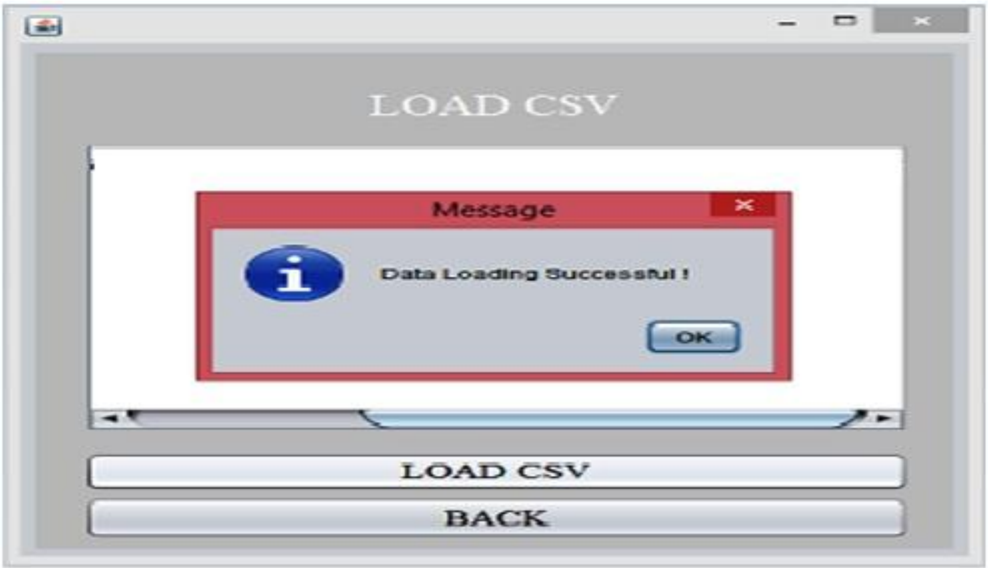

Fig 1.Load CSV File

In Fig 2.we can view the dataset which is loaded from UCI repository dataset. We can also add or delete the record from the dataset.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

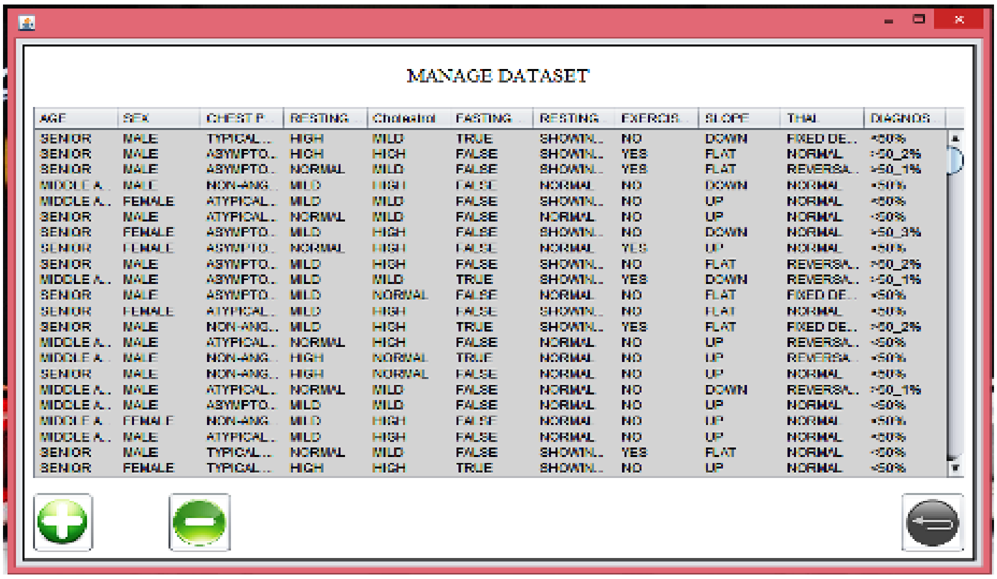

Fig 2.Manage Dataset

In Fig 3.here we can enter the single data using following attributes such as age, sex, chest pain, fasting blood sugar, slope, thal, cholesterol, resting ECG etc

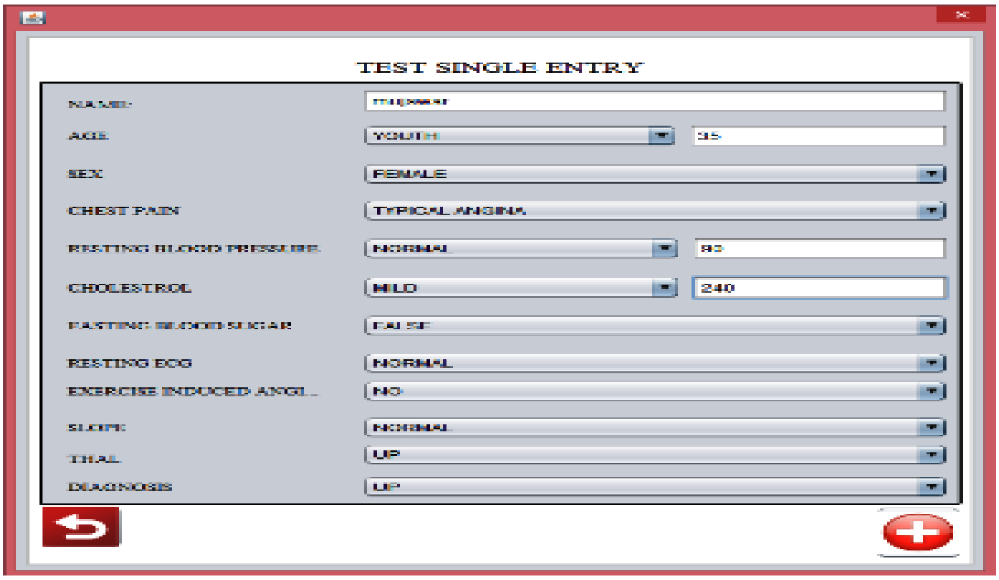

Fig 3.Single Entry Test

In Fig 4. we can calculate single and multiple entry and analyze them. Also it describes about the statistics of the patient against the current recording with reference to the pre-recorded data with maximum accuracy.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

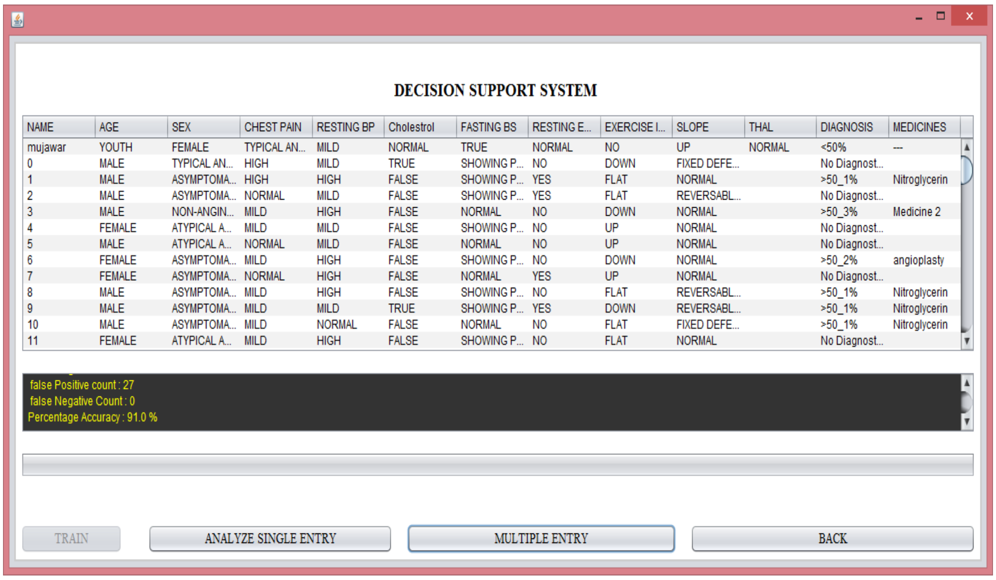

Fig 4.Prediction Result.

Naive Bayes model identifies patient's characteristics with Heart Disease. For each input attributes it shows probability for the predictable state.

Classification Matrix: Classification matrix displays the frequency of incorrect and correct predictions of Heart Disease. It compares the actual values in the test dataset with the predicted values in the trained model.

DataSet Name	Actual	Total	Correct
Heart Disease Detected	120	100	93
Heart Disease Not Detected	120	100	89

Fig 5.Classifiaction matrix

Fig. 5 show Classification matrix, in which row represent heart disease detected and heart disease not detected, where actual entry is 120 and 100 entry is used for testing purpose, in which 93 are heart disease detected and 89 are heart disease not detected.

After applying naïve bayes on training dataset the results obtained is shown in the matrix form called as confusion matrix in the form of 2 dimensional matrixes. The confusion matrix is easy to understand as the incorrect and correct classification is displayed in the table. The confusion matrix is shown in table below with Heart disease detected and heart disease not detected.

Confusion Matrix			
		Heart Disease Detected	**Heart Disease Not Detected**
	Heart Disease Detected	93	11
	Heart Disease Not Detected	7	89

Fig 6.Confusion matrix

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

Precision	Heart Disease Detected	(Relevant Intersect Retrieved) / Retrieved	Correct Retrieved Object / Retrieved Objects	0.93
Precision	Heart Disease Not Detected	(Relevant Intersect Retrieved) / Retrieved	Correct Retrieved Object / Retrieved Objects	0.89
Recall	Heart Disease Detected	(Relevant Intersect Retrieved) / Relevant	Correct Retrieved Object / Actual Objects	0.775
Recall	Heart Disease Not Detected	(Relevant Intersect Retrieved) / Relevant	Correct Retrieved Object / Actual Objects	0.741666667

Fig 7.Precision and Recall calculation

Fig 7 explain about the precision and recall calculation and fig 8 gives the accuracy of heart Disease detection with 13 attributes.

	Precision	Recall
Heart Disease Detected	0.93	0.775
Heart Disease Not Detected	0.89	0.741666667
Total	0.91	0.758333333

Fig 8.Precision and Recall table accuracy

Fig 9 gives graphical representation of accuracy result. Here blue line represent accuracy of heart disease detected and red line represents accuracy of heart disease not detected.

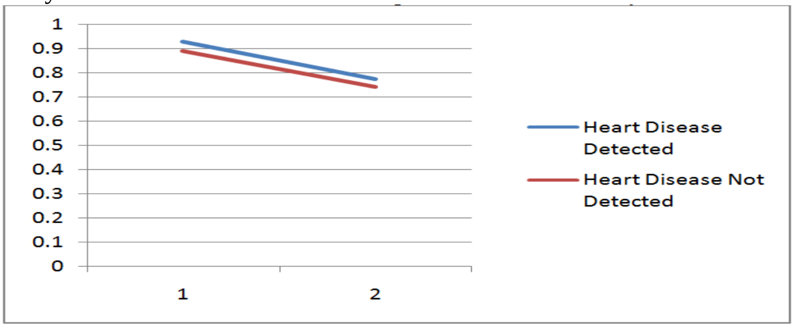

Fig 9 Graphical representation of Prediction Result

VI CONCLUSION AND FUTURE WORK

The main aim of our project is to predict more accurately the presence of heart disease., With less number of attributes is a challenging task in Data Mining, Instead of going for a number of tests. Two data classification techniques were applied namely modified K-means and Naïve Bayes. In this paper a modified K means algorithm is proposed which tries to remove one of the major limitations of basic K-means algorithm, which requires number of clusters as input.

This system can be further used in Future work as, For eg. it can incorporate other medical attributes besides the above list. To mine large amount of unstructured data, the Text mining can be used, available in healthcare industry database.

REFERENCES

[1] Blake, C.L., Mertz, C.J.: "UCI Machine Learning Databases", http://mlearn.ics.uci.edu/databases/heartdisease/,2004.
[2] Han, J., Kamber, M.: "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 2006.
[3] Mrs.G.Subbalakshmi, "Decision Support in Heart Disease Prediction System using Naive Bayes", Indian Journal of Computer Science and Engineering
[4] Yanwei, X.; Wang, J.; Zhao, Z.; Gao, Y., *"Combination data mining models with new medical data to predict outcome of coronary heart disease".* Proceedings International Conference on Convergence Information Technology 2007, pp. 868 – 872.
[5] Heon Gyu Lee, Ki Yong Noh, Keun Ho Ryu, "Mining Biosignal Data: Coronary Artery Disease Diagnosis using Linear and Nonlinear Features of HRV," LNAI 4819: Emerging Technologies in Knowledge Discovery and Data Mining, pp. 56-66, May 2007
[6] Statlog database: http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/heart
[7] Niti Guru, Anil Dahiya, NavinRajpal, *"Decision Support System for Heart Disease Diagnosis Using Neural Network"*, Delhi Business Review, Vol. 8, No. 1 (January - June 2007).

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

[8] B M Ahamed Shafeeq, K S Hareesha, " Dynamic Clustering of Data with Modified K-Means Algorithm",International Conference on Information and Computer Networks (ICICN 2012), IPCSIT, vol. 27,pages 221-225,2012

[9] Mohamed Abubaker, Wesam Ashour, "Efficient Data Clustering Algorithms: Improvements over K-means",International Journal of Intelligent Systems and Applications, vol. 5,issue 3, pages 37-49, 2013

[10] Mohammed EI Agha, Wesam M. Ashour, " Efficient and Fast Initializtion Algorithm for K-means Clustering" ,1.1. Intelligent Systems and Applications, vol. 4, issue 1, pages 21-31, 2012.

[11] Kaur, H., Wasan, S. K.: "Empirical Study on Applications of Data Mining Techniques in Healthcare", Journal of Computer Science 2(2), 194-200, 2006.

[12] MA.Jabbar,B.L Deekshatulu,Priti chandra,"Prediction of Risk Score for Heart Disease using Associative classification and Hybrid Feature Subset Selection",*In .Conf ISDA,pp 628-634,IEEE(2013)*

[13] MA.Jabbar,B.L Deekshatulu,Priti chandra,"Knowledge discovery from mining association rules for heart disease prediction" *pp45-53,vol 41,no 2 ,JATIT(2013*

[14] SellappanPalaniappan, RafiahAwang, *"Intelligent Heart Disease Prediction System Using Data Mining Techniques"*, IJCSNS International Journal of Computer Science and Network Security, Vol.8 No.8, August 2008

BIOGRAPHY

Sairabi Hajrat Mujawar is a student of M.Tech. Information Technology Of Bharati Vidyapeeth Deemed University, Pune. She had completed graduation in BE in Computer Engineering from Bharati Vidyapeeth college of Engineering, Mumbai University, Navi Mumbai, Maharashtra, India in 2007. Her area of interests are Data Mining, real time applications, database management system, web development.