Development of a Data Clustering Algorithm for Predicting Heart

Bala Sundar V
Dept. of Computer Applications
Bharathiar University
Coimbatore, India

T DEVI
Dept. of Computer Applications
Bharathiar University
Coimbatore, India

N SARAVANAN
Dept. of Computer Applications
Bharathiar University
Coimbatore, India

ABSTRACT
This research paper proposes the findings of the accuracy of the result by using the K-Means clustering technique in prediction of heart disease diagnosis with real and artificial datasets. K-Means Clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. Each cluster is assigned a random target number of clusters-k and started from a random initialization. The proposed technique classifies the group of the objects based on attributes into K number of 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 shows that the integration of clustering gives promising results with highest accuracy rate and robustness.

General Terms
Data mining, Clustering algorithm, Heart disease.

Keywords
Decision Tree, Naive Bayes, Neural Network, K-Means Clustering.

1. INTRODUCTION
Data Mining refers to the process of finding interesting hidden patterns. The process of data mining is composed of selecting, analyzing, preparing, applying, interpreting and evaluating the results [2, 4, 7]. Many clinical diagnoses accomplished in data mining techniques for classification and prediction. The healthcare industry collects huge amounts of healthcare data which are not ―mined‖ to discover hidden information for effective decision making. The depart research developed Intelligent Heart Disease Prediction System using data mining techniques such as Decision Tree, Naive Baye's and Neural Network, K-Means Clustering Using medical profiles such as age, sex, blood pressure and blood sugar it can predict the performance analysis and comparison between those techniques of patients getting heart disease or not [19]. It enables significant knowledge, e.g. patterns, relationships between medical factors related to heart disease, to be established. It is a scalable, reliable and expandable. It is implemented on the .Net platform. The existing research mainly focused in performance analysis and comparison. The quality of the service is very essential in analyzing field that means the accuracy result. This research is mainly focuses on Prediction of Heart Disease using K-Means Clustering in the context of data mining. The cost function like predictive stable accuracy evaluated in prediction of heart disease using K-Means Clustering technique in data mining. Similarly complementary measures like compactness and connectedness of clusters are treated as two objectives for cluster analysis.

This research carried out an extensive prediction for the task of using different real life and artificially created datasets. Experimental results shows that K-Means Clustering technique fetches a clear edge, whereas clustering task produce comparable results with existing techniques. In K-Means Clustering technique, each cluster is assigned a random target number of clusters-k and started from a random initialization. The technique classifies the group of the objects based on attributes or features into K number of group. The grouping is done by minimizing the sum of squares of distances between data using Euclidean distance formula and the corresponding cluster centroid. Thus the purpose of K-Means Clustering is to classify the data with predictive stable accuracy.

2. DATA MINING
Data Mining is defined in many ways in different situations. Major definitions used in literature are refers to the finding of relevant and useful information from databases [2] and also deals with finding of patterns and hidden information from a large database [21]. Data Mining is also known as Knowledge Discovery in Databases (KDD) which is defined as the non–trivial extraction of implicit, previously unknown and potentially useful information from the data [21].The term Knowledge Discovery in Databases or KDD refers to the broad process of finding knowledge in data and emphasizes the "high-level" application of particular data mining methods. It is of interest to researchers in machine learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for expert systems and data visualization. The unifying goal of the KDD process is to extract knowledge from data in the context of large databases [7]. Using data mining methods or algorithms, the techniques which extract and identify the deemed knowledge, dealing to the specifications of measures and thresholds, using a database along with any required preprocessing, sub sampling, and transformations of that database. An outline of the steps of the KDD Process and the overall process of finding and interpreting patterns from data involves the repeated application of the following steps are developing an understanding of the application domain and the relevant prior knowledge and identifying the goal of the process from the customer's viewpoint. Creating a target data set which performs selecting a data set or focusing on a subset of variables or data samples on which discovery is to be performed. Data cleaning and preprocessing includes removal of noise or outliers for collecting necessary information to model or account for noise and deciding on strategies for handling missing data fields and according for time sequence information and known changes. Data reduction and projection consist of finding useful features to represent the data depending on the goal of the task. Using dimensionality reduction or transformation methods, the method has to

reduce the effective number of variables or to find invariant representations for the data. The data mining task consist of the KDD process such as classification, regression, clustering and so on [2,7]. Exploratory analysis, model and hypothesis selection to be composed of the data mining algorithms and the selecting methods to be used for searching for patterns in the data, deciding which models and parameters may be appropriate and matching a particular data mining method with the overall criteria of the KDD process.

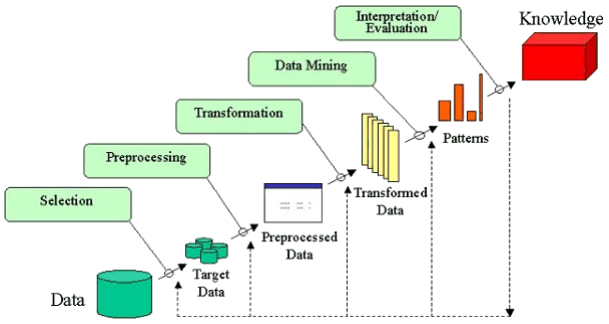

Fig 1: KDD Process of Data Mining

Fig 1. Shows the KDD Process of Data Mining the mechanism works and searches for patterns of interest in a particular representational form or a set of such representations as classification rules or trees, regression and clustering. Interpreting mined patterns possibly returning to iterations. This can involve visualization of the extracted patterns. Consolidating discovered knowledge using the knowledge directly and incorporating the knowledge into another system for further action or simply documentation and reporting it to interested parties. 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 the quality as knowledge. It also includes the choice of encoding schemes, preprocessing, sampling, and projections of the data prior to the data mining step. Data mining refers to the application of algorithms for extracting patterns from data without the additional steps of the KDD process [7, 35].

3. CLUSTERING USING K-MEANS ALGORITHMS

The categorization of objects into various groups or the partitioning of data set into subsets so that the data in each of the subset share a general feature, frequently the proximity with regard to some defined distance measure [36], is known as Clustering. The clustering problem has been identified in numerous contexts and addressed being proven beneficial in many medical applications. Clustering the medical data into small with meaningful data can aid in the discovery of patterns by supporting the extraction of numerous appropriate features from each of the clusters thereby introducing structure into the data and aiding the application of conventional data mining techniques [20, 37]. Numerous methods are available in the literature for clustering and employed the renowned K-Means clustering algorithm in this approach.

The k-means algorithm [11, 21, 27, 37, 38] is one of the widely recognized clustering tools that are applied in a variety of scientific and industrial applications. k-means groups the data in accordance with their characteristic values into k distinct clusters [20]. Data categorized into the same cluster have identical feature values. k, the positive integer denoting the number of clusters, needs to be provided in advance. The steps involved in a k-means algorithm are given subsequently: Prediction of heart disease using K – Means clustering technique

1. K points denoting the data to be clustered are placed into the space. These points denote the primary group centroids.
2. The data are assigned to the group that is adjacent to the centroid.
3. The positions of all the K centroids are recalculated as soon as all the data are assigned.
4. Steps 2 and 3 are reiterated until the centroids stop moving any further. This results in the segregation of data into groups from which the metric to be minimized can be deliberated [20].

The preprocessed heart disease data is clustered using the K-means algorithm with the K values. Clustering is a type of multivariate statistical analysis also known as cluster analysis, unsupervised classification analysis, or numerical taxonomy. K-Means clustering generates a specific number of disjoint, flat (non-hierarchical) clusters. It is well suited to generating globular clusters. The K-Means method is numerical, unsupervised, non-deterministic and iterative.

3.1 K-Means and derivatives

The k-Means algorithm assigns each point to the cluster whose center (also called centroid) is nearest. The center is the average of all the points in the cluster — that is, its coordinates are the arithmetic mean for each dimension separately over all the points in the cluster. Example: The data set has three dimensions and the cluster has two points X = (x1, x2, x3) and Y = (y1, y2, y3). Then the centroid Z becomes Z = (z1, z2, z3), where

$$z1 = \frac{x1+y1}{2}, \quad z2 = \frac{x2+y2}{2} \quad \text{and} \quad z3 = \frac{x3+y3}{2}$$

3.2 Advantages to using this technique

a. The main advantages of this algorithm are its simplicity and speed which allows it to run on large datasets.

b. With a large number of variables, K-Means may be computationally faster than hierarchical clustering (if K is small).

c. K-Means may produce tighter clusters than hierarchical clustering, especially if the clusters are globular.

4. DATA PREPROCESSING

Cleaning and filtering of the data might be necessarily carried out with respect to the data and data mining algorithm employed so as to avoid the creation of deceptive or inappropriate rules or patterns[8]. The steps involved in the pre-processing of a dataset are the removal of duplicate records, normalizing the values used to represent information in the database, accounting for missing data points and removing unneeded data fields. To make data appropriate for the mining process it needs to be transformed. The raw data is changed into data sets with a few appropriate characteristics. Moreover it might be essential to combine the data so as to reduce the number of data sets besides minimizing the memory and processing resources required by the data mining algorithm [30]. This leads to removal of duplicate records and

supplying the missing values in the heart disease data warehouse. In addition, it is also transformed to a new form which is appropriate for clustering [20].

Clinical databases have accumulated large quantities of information about patients and their medical conditions. Heart disease is the major cause of casualties in the world. The term Heart disease encompasses the diverse disease that affects the heart. Heart disease kills one person every 34 seconds in the United States. Coronary heart disease, Cardiomyopathy and Cardiovascular disease are some categories of heart diseases. The term ―cardiovascular disease‖ includes a wide range of conditions that affect the heart and the blood vessels and the manner in which blood is pumped and circulated through the body. Cardiovascular disease (CVD) results in severe illness, disability, and death.

A total of 454 records with 13 medical attributes (factors) were obtained from the Cleveland Heart Disease database [3, 32]. Fig 5.1 lists the attributes. The records were split equally into two datasets such as training dataset (227 records) and testing dataset (227 records). To avoid bias, the records for each set were selected randomly. The attribute ―Diagnosis‖ is identified as the predictable attribute with value 2‖ for patients with heart disease and value ―1‖ for patients with no heart disease. The key used as ―PatientId‖, the rest are input attributes. It is assumed that problems such as missing data, inconsistent data, and duplicate data had all been resolved.

Predictable Attribute

1. Diagnosis (value 1: <50% diameter narrowing (no heart disease); value 2: >50% diameter narrowing (has heart disease))

Key Attribute

1. PatientId – Patient's identification number

Input Attributes

1. Age in Year

2. Sex (value 1: Male; value 0: Female)

3. Chest Pain Type (value 1: typical type 1 angina, value 2: typical type angina, value

3: non-angina pain, value 4: Asymptomatic)

4. Fasting Blood Sugar (value 1: >120 mg/dl; value 0: <120 mg/dl)

5. Serum Cholesterol (mg/dl)

6. Restecg – resting electrographic results (value 0: normal; value 1: having ST-T wave

Abnormality; value 2: showing probable or definite left ventricular hypertrophy)

7. Maximum Heart Rate Archieved; value (0.0) :> 0.0 and <=80, value (1.0) : >81 and

<119, value (2.0):=120;

8. Fasting Blood Sugar; 120

9. Exang - exercise induced angina (value 1: yes; value 0: no)

10. Old peak – ST depression induced by exercise

11. Slope – the slope of the peak exercise ST segment (value 1: unsloping; value 2: flat; value 3: down sloping)

12. CA – number of major vessels colored by floursopy (value 0-3)

13. Thal (value 3: normal; value 6: fixed defect; value 7: reversible defect)

Fig 2: Predictable Input Attributes

5. EXPERIMENTAL RESULTS

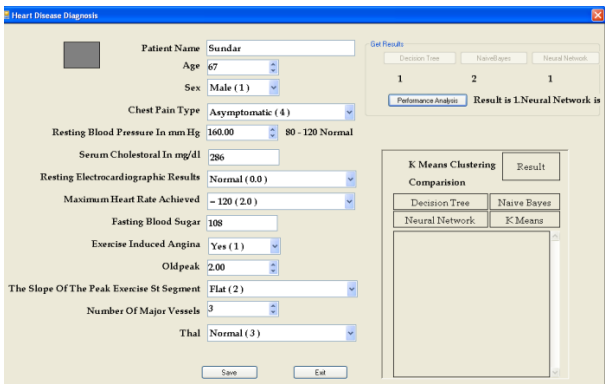

Fig 3: Performance Analysis of Decision Tree, Naive Baye's and Neural Network

Fig 3. Shows the existing techniques of the performance analysis of Decision Tree, Naive Baye's and Neural Network and which associates the values in numerical representations as 1 represents person not having heart disease and 2 represents having disease. Fig. 3 shows one person's information's, the existing techniques which hit upon the performance analysis of whether the patient getting heart disease or not based on his health information. Here, the Neural Network performance assured the person having heart disease which represents the value as 2 and Decision Tree and Naïve Baye's shows the value as 1. So Neural Network performance is best which evaluates to the Decision Tree and Naïve Baye's techniques.

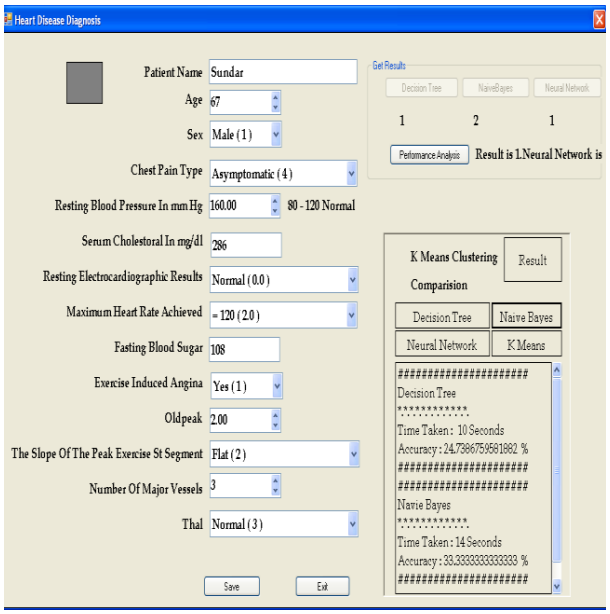

Fig 4: Comparison of Decision tree and Naive Baye's

Fig 4. Shows comparing the time and accuracy results of Decision Tree and Naive Baye's Techniques. The comparison performs the derived units of seconds.

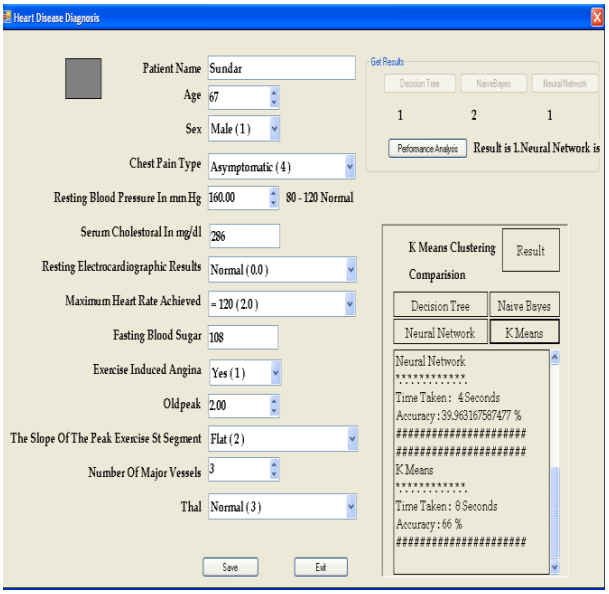

Fig 5: Comparison of Neural Network and K-Means Clustering Technique

Fig 5. Shows comparing the time and accuracy results of Neural Network and K-Means Clustering Techniques. The comparison performs the derived units of seconds.

5.1 Graph Result

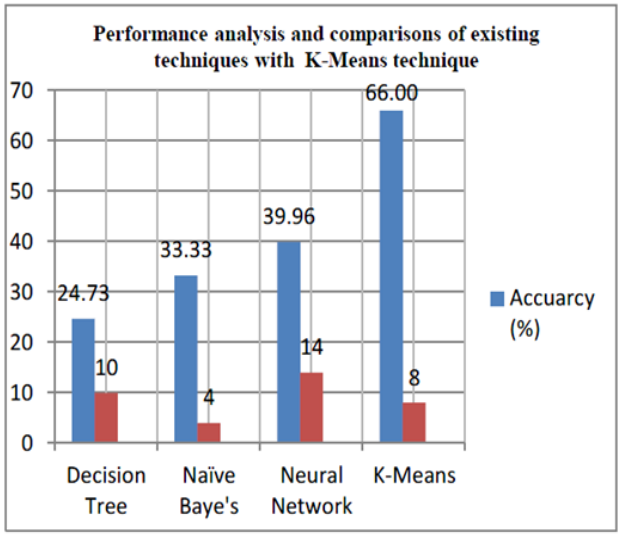

Fig 6: Graphical Representation of K-Means with Existing Techniques

The k-means algorithm assigns each point to the cluster whose center (also called centroid) is nearest. The center is the average of all the points in the cluster — that is, its coordinates are the arithmetic mean for each dimension separately over all the points in the cluster. Fig. 6 Shows the graphical representation of time and accuracy results of prediction of heart disease diagnosis, based on the heart dataset it predicts the time and accuracy result.

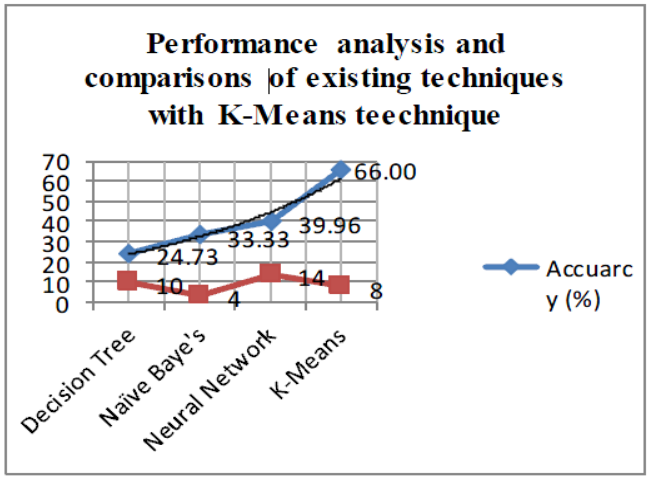

Fig 7: Exponential Accuracy of K-Means with Existing Techniques

The graph shows the exponent accuracy values of Decision Tree, Naïve Baye's, Neural Network and K-Means Clustering. Compared with other three techniques as Decision Tree, Naïve Baye's and Neural Network, the point hit upon the K-Means Clustering has the highest range. X-Axis takes the range of points as 0 to 5 and Y- Axis takes the accuracy values of 0 to 70. The highest exponential value of K-Means cluster is 66.00%.

Table 1. Performance of K-Means Clustering

Algorithm	Time taken	Accuracy
Decision Tree	10 sec	24.73%
Naive Baye's	14 sec	33.33%
Neural Network	4 sec	39.96%
K – Means	8 sec	66.00%

Table 1. Shows the performance of K-Means clustering technique with existing techniques of Decision tree, Naive Baye's and Neural Network. The results are acquired using the trained dataset. The representation shows the time and accuracy of heart disease diagnosis and implementation of K-Means clustering for getting accurate result. The accuracy of decision tree is 24.73%, Naive Baye's is 33.33%, Neural Network is 39.96% and K-Means is 66.00%. K-Means clustering technique performs the accurate convergence, compared with existing techniques.

6. RESULTS AND DISCUSSIONS

Data Mining is the process of extracting hidden information in the database. The Modern trend of the data mining process is to mine the huge amount of the data in the data set. Many clinical diagnosis, examined the data mining techniques for classification and prediction. Many researches completed in prediction of heart disease using decision tree, Naive Baye's and Neural Network. In those techniques, the Naive Baye's processing time is too low when compared to other techniques such as Decision Tree and Neural Network. The technique of Neural Network acquired the accuracy result compared to Decision tree and Naive Baye's. In Neural Network technique, it achieved the accuracy result based on more number of iterations. Iterations based on micro processes in back ground terminal execution. So, always the network bandwidth capacity should be high otherwise some problems occur in the overall process like loop termination, process termination and so on. The Health diagnosis always needs the full accuracy of prediction of heart disease.

In this research implemented the K-Means Clustering algorithms has been implemented. This performs certain number of iterations randomly which access the nearest n-observations into k, so as to attain the high speed time consumption and offers stability of the accurate result. Here, this research approaches the Compactness and Connectedness for accuracy result. Using Compactness, which minimizes the sum of squares by using Euclidean distance and Connectedness, which performs the nearest observations in the random process then identifies the mean and centroid value of the objects.

The training dataset contains the real valued heart patient details and testing dataset contains the real and artificial valued dataset. When the new patient's record has been appended, it would accumulate the record in the user data set and also it would find the accurate result for that record. Once the execution takes place, it checks the conditions and processing with the existing research techniques such as Decision tree, Naive Baye's and Neural Network. The process find out, the patient who have heart disease or not and performance analysis takes place in those techniques. Subsequently, comparing the moment of processing time and accuracy result analyzed with the K-Means Clustering technique. This technique examines the prediction of high speed and accuracy result compare to those existing techniques.

7. CONCLUSION

This research is concerned with the study and analysis of Data Mining and Data Clustering algorithms, analysing the existing methods for Predicting Heart Disease and to design and develop an efficient and effective method for predicting heart disease. The existing methods for Heart Disease Prediction are Decision Tree, Naive Bayes and Neural Network. In this research K-Means Clustering techniques has been used. The compactness and connectedness for complementary measures are used and it is found that the efficiency and effectiveness of the method for predicting Heart Disease is better than the other three techniques through software prototype

This research work can be further extended to mine the huge amount of unstructured data available in the form of health care databases. The work can also be extended to include images such as Electro Cardio Gram (ECG) scanned images.

8. REFERENCES

[1] Alexander Rakhlin, Andrea Caponnetto, ―Stability of K-Means Clustering‖, 2006.

[2] Arun K. Pujari, ―Data Mining Techniques‖, Universities Press (India) Ltd, 2001.

[3] Blake, C.L., Mertz, C.J. ―UCI Machine Learning Databases‖, 2004.

[4] Cipolla, Emil T. ―Data Mining: Techniques to Gain Insight into Your Data Enterprise Systems Journal‖, pp.18-24, 64 December, 1995.

[5] Clifton, Christopher, "Encyclopedia Britannica: Definition of Data Mining", 2010.

[6] Duda RO, Hart PE, Stork DG. ―Pattern Classification‖. New York: Wiley-Interscience, 2000

[7] Fayyad, Piatetsky-Shapiro, Smyth, "From Data Mining to Knowledge Discovery: An Overview", in, Advances in Knowledge Discovery and Data Mining, AAAI Press / The MIT Press, Men, 1996.

[8] Gerhard Münz, Sa Li, and Georg Carle, "Traffic anomaly detection using k-means clustering", In Proc. of performance, reliability and dependability evaluation of communication networks and distributed systems, 4 GI / ITG Workshop MMBnet 2007, Hamburg, Germany, September 2007.

[9] Guthrie L, Walker E, Guthrie J. Document classification by machine: theory and practice. Proceedings of the 15th International Conference on Computational Linguistics: Association for Computational Linguistics, Morristown, NJ. pp. 1059-1063, 1994.

[10] Harleen Kaur,Siri Krishan Wasan,‖Empirical Study on Applications of Data Mining Techniques in Healthcare, Journal of Computer Science 2 (2): 194-200, ISSN 1549-3636, 2006.

[11] Jain A.K, Murthy M.N and Flynn P.J. Data Clustering: A Review, ACM Computing Reviews, November 1999.

[12] Jyoti Soni, Ujma Ansari, Dipesh Sharma, ―Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction‖, International Journal of Computer Applications (0975 – 8887) Volume 17– No.8, March 2011.

[13] Kavitha K.S, K.V.Ramakrishnan, Manoj Kumar Singh, ―Modeling and design of evolutionary neural network for heart disease detection‖, IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 5, September 2010, ISSN (Online): 1694-0814

[14] Khaled Hammouda, Prof. Fakhreddine Karray ―A Comparative Study of Data Clustering Techniques‖. University of Waterloo, Ontario, Canada, Volume 13, Issues 2-3, pp. 149-159, November 1997.

[15] Latha Parthiban and R.Subramanian, ― Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm‖, International Journal of Biological and Life Sciences 3:3 2007.

[16] Meira Jr., Zaki, M. Fundamentals of Data Mining Algorithms, 2009.

[17] Miller, A., Blott, B., & Hames, T. ― Review of Neural Network Applications in Medical Imaging and Signal Processing‖. Medical and Biological Engineering and Computing, 30(5), pp: 449-464, 1992.

[18] Rennie JDM, Shih L, Teevan J, Karger Dr. Tackling the poor assumptions of Naive Bayes text classifiers. Proceedings of the Twentieth International Conference on Machine Learning pp. 616-23, 2003.

[19] Sellappan Palaniappan, Rafiah Awang,. ― Intelligent Heart Disease Prediction System Using Data Mining Techniques. Computer Systems and Applications, 2008. AICCSA 2008. IEEE / ACS International Conference on, pp. 108-115, March 31-April 4 2008.

[20] Shantakumar, B. Patil and Y.S Kumaraswamy., ―Intelligent and Effective Heart attack Prediction System Using Data Mining and Artificial Neural Network, Eurp Journals Publishing Inc. ISSN 1450-216X Vol.31 No.4 2009, pp.642-656, 2009.

[21] Shawkat Ali A B M,Saleh A.Wasimi, ―Data Mining : Methods and Techniques ,Cengage Learning Indis Ltd , 2009.

[22] Shearer C. ―The CRISP-DM model: the new blueprint for data mining, Data Warehousing Vol. 5, pp.13-22, 2000.

[23] Siri Krishan Wasan, Vasudha Bhatnagar and Harleen Kaur, ―The Impact of Data mining techniques on Medical Diagnostics‖, Data Science Journal, Volume 5, 19, October 2006.

[24] Srinivas K, B. Kavihta Rani and Dr. A.Govrdhan., ―Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks, International Journal on Computer Science and Engineering,Vol. 02, No. 02,pp. 250-255, 2010.

[25] Subbalakshmi Mrs.G, E. Anupriya, N.CH.S.N.Iyengar, ―Enhanced Prediction of Heart Disease with feature Subset Selection using Genetic Algorithm, International Journal of Engineering Science and Technology Vol. 2(10), pp. 5370-5376, 2010.

[26] Tahseen A. Jilani, Huda Yasin, Madiha Yasin, Cemal Ardil, ―Acute Coronary Syndrome Prediction Using Data Mining Techniques-an Application‖ International Journal of Information and Mathematical Sciences 5:4, 2009

[27] Tapas Kanungo, David M. Mount, Nathan. An Efficient k-Means Clustering Algorithm: Analysis and Implementation IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, July 2002

[28] Mitchell Tom M, ―Machine Learning, Singapore, McGraw- Hill.1997.

[29] Weinstein, J.N., Kohn, K.W. Neural computing in Cancer Drug Development: Predicting Mechanisms of Action. Science. pp. 447-451, 1992.

[30] Wynne Hsu, Mong-Li Lee, Bing Liu, Tok Wang Ling, ―Exploration Mining in Diabetic Patients Databases: Findings and Conclusions, KDD 2000: pp: 430-436, 2000.

[31] http://cbcl.mit.edu/publications/ps/rakhlin-stability-clustering.pdf.K-Means Clustering.05-03-11.

[32] http://mlearn.ics.uci.edu/databases/heartdisease/,.Heart Disease Dataset.09-09-10.

[33] http://www.anderson.ucla.edu/faculty/jason.frand/teacher /technologies/palace/datamining.htm. Data mining techniques. 07-12-10.

[34] http://www.britannica.com/EBchecked/topic/1056150 /data-mining. Data Mining.19-11-10.

[35] http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd. html.Data mining. 18-10-10.

[36] Zakaria Nouir, Berna Sayrac, Benoît Fourestié, Walid Tabbara, and Françoise Brouaye, "Generalization Capabilities Enhancement of a Learning System by Fuzzy Space Clustering," Journal of Communications, Vol. 2, No. 6, pp. 30-37, November 2007.

[37] F. H. Saad, B. de la Iglesia, and G. D. Bell, ― A Comparison of Two Document Clustering Approaches for Clustering Medical Documents, Proceedings of the 2006 International Conference on Data Mining (DMIN-06), 2006.

[38] C. Ordonez, ― Programming the K-Means Clustering Algorithm in SQL, Proc. ACM Int'l Conf. Knowledge Discovery and Data Mining, pp. 823-828, 2004.