Diagnosing of Heart Diseases using Average K-Nearest Neighbor Algorithm of Data mining

C. Kalaiselvi, PhD
Associate Professor and Head of Computer Applications,
Tiruppur Kumaran College for Women,
Tirupur, Tamilnadu, India
Email Id: kalaic29@gmail.com

Abstract- The most important and prevalent diseases that commonly occur in people and causes 80% of death in country is heart disease. It is estimated that by 2030, approximately 25 million people die because of heart diseases. Though many researchers have suggested and proposed methods for diagnosing the heart diseases from the enormous amount of heart disease data, there is no proper effective techniques and are not properly mined. To get improved classification accuracy and efficiency a new approach called average k-nearest neighbor algorithm is proposed in this paper. The dataset used for prediction is obtained and utilized from UCI machine learning repositories. The main objective of this research work is to diagnose heart disease with reduced number of attributes that are relevant to heart diseases.

Keywords— Average K- Nearest neighbor; Data mining; K- Means Clustering; Neural networks; Prediction

NOMENCLATURE

CMAR-Classification Based On Multiple Association Rules

ECG –Electro Cardio Graph

HRV- Heart Rate Variability

KNN- K-Nearest Neighbor

AKNN-*Average K-Nearest Neighbor*

SVM – Support Vector Machines

UCI- University of California Irvine repositories

WHO- World Health Organization

I. INTRODUCTION

The most innovative and emerging field of computer science is that uses various statistical techniques, classification and clustering algorithms and pattern recognition for problems is called as data mining. This methodology has the potential ability to find patterns and relationships from data and it is also applied in variety of tasks which includes medicine. Data mining proved its efficiency in most of the areas of medicine in which the results obtained were compared with other methodologies to achieve improved accuracy and performance.

As data mining derived from the name of searching for valuable information in large databases and it is also known as knowledge discovery in databases. In the era of computer science it is the process of discovering interesting and useful patterns for the real world problems. Classification is termed as one of the data mining techniques which are used to predict group membership for data instances. For example, it can also be applied in prediction [1]. Of these most popular classification techniques may include decision trees and neural networks.

Many new algorithms as well as techniques of data mining have been used for predicting heart diseases. The diagnosis using machine learning techniques is one of the existing techniques which have a transparent diagnostic knowledge which gives more accurate results. Machine learning is classified into two types which are connectionist learning and symbolic learning [4]. Of these techniques, the user can easily understand the rules of symbolic learning techniques and are considered as comprehensible techniques. The best example for the symbolic technique is rule induction which is extensively used for medical diagnosis.

As per, WHO report regarding cardiovascular disease prevention and control, of all the diseases, cardiovascular diseases is the leading causes of death and disability worldwide. Even though, mostly all the proportion of cardiovascular diseases is preventable, they continue to rise mainly because of inadequate preventive measures. Over 347 million people worldwide are having heart diseases. Some of the symptoms of heart failure are swelling in the legs, ankles, feet and abdomen, breathing problems and swelling veins in neck. If the condition is diagnosed earlier and if they follow their treatment plans regularly [5-7] then the people who have heart failure can live longer. The risk factors of heart disease are obesity, abnormal blood cholesterol levels, high blood pressure and other physical activities, smoking etc.

978-9-3805-4421-2/16/$31.00

The heart and blood vessel diseases which are called as cardiovascular diseases are of two types which are more common in people with diabetes.

They are,

i. Coronary Artery diseases
ii. Cerebral vascular diseases

High blood glucose levels may cause blockage of blood vessels result into irregularity of heart beats. It may cause poor blood circulation in legs and feet, heart attack or stroke. The symptoms of heart disease are pain in chest, shoulder, arms, jaw, breath shortness, sweating, giddiness, nausea and light-headedness. The major problems of coronary heart diseases are high blood pressure and also diabetes which may weaken the heart. Then the body may release some proteins and other substances into the blood [11, 13].

II. RELATED WORK

There is large number of research work have been undergoing in the most important field of medical data mining and it is also carried out in the prediction of heart diseases and some researchers have been analysed and investigated the uses of data mining techniques to help physicians to make the decision effective.

Naive Bayes is the important data mining techniques which is applied to major real world problems. Mr Chau Tu compares the bagging algorithm in addition to C4.5 algorithm, bagging algorithm in addition to Naive Bayes algorithm to diagnose the heart disease in patients. Raj Kumar and Reena analyzed and compared Naive Bayes, K-Nearest Neighbor, and decision trees in diagnosis of heart disease in patients. Cheung suggested Naive Bayes classifier for the heart disease prediction on the dataset.

An Intelligent Heart Disease Prediction System which is proposed and developed by Sellappan, Palaniappan et al [8]. He proposed data mining techniques such as Decision Trees, Naive Bayes and Neural Networks. Each and every method of data mining has its own nature of strength and weakness to predict and gave finite and appropriate results. The system was built using hidden patterns and its connected relationship with other patterns.

Heon Gyu Lee et al. suggested a technique for developing a multi parametric feature with linear and non linear features that are relevant to HRV. Heon and his team used multiple classifiers such as Bayesian classifiers, CMAR, C4.5 (Decision Trees) and SVM to predict heart disease and to achieve good results [9]. Niti Guru et al. proposed neural networks to predict heart diseases, Blood Pressure and Diabetes [10]. The dataset they have used contains 13 attributes. They suggested supervised networks i.e. Neural Network with back propagation algorithm is proposed for training and testing of datasets.

Association rules were used and also the problem of identifying the constraints for heart disease prediction was suggested by Carlos Ordonez [12]. Finally the resultant dataset obtained includes records of patients related to heart disease whether the presence or absence of heart disease. Latha Parthiban et al. [14] proposed a technique called Co active neuro fuzzy inference system to predict heart disease. This model uses neural network capabilities with fuzzy logic and genetic algorithm in combination for prediction.

Kiyong Noh et al. [15] proposed classification method for extracting multi parametric features to examine HRV obtained from Heart disease patterns of ECG which are data pre-processed. They have used the dataset that consists of 670 records of peoples with or without heart disease and was partitioned into two categories as normal patients and patients with heart disease. Shruti Ratnakar et al. proposed a genetic algorithm to reduce the set of attributes. An efficient associative classification algorithm with genetic algorithm was proposed by Akhil Jabbar et al. to predict heart disease which gives high predictive accuracy. This paper examines the various data mining algorithms and techniques used in heart disease prediction and suggested average K-nearest neighbor algorithm for heart disease prediction.

III. ALGORITHMS TO PREDICT HEART DISEASE

Algorithms and techniques such as Classification, Clustering, Regression, Association rules, Neural Networks, Decision Trees, Genetic Algorithms, and Algorithms were proposed by many researchers to heart disease prediction [16, 17].

A. Classification

The most commonly used technique for the real world entities is classification, which includes set of pre-classified examples for developing models that could classify the population of records at a very large scale. This approach frequently employs a decision tree algorithm or neural network-based classification algorithms. Here the data classification process which involves first learning then classification. In learning phase, the training data are analyzed by classification algorithm. In classification phase, test data are used to estimate the accuracy of the classification. Based on the accuracy achieved and is accepted then the rules are created which can be applied to the new set of data or tuples. Various types of classification models are

i. Bayesian Classification
ii. Classification by decision tree induction
iii. Neural Networks
iv. Classification Based on Associations

B. Clustering

Clustering can also be called as identification of similar classes of objects which are not known in prior. By using clustering techniques, the dense and sparse regions are also identified in object space and these techniques can discover

overall distribution pattern and correlations among data attributes.

Various types of clustering methods are

i. Hierarchical methods
ii. Partitioning methods
iii. Density based methods
iv. Grid based methods
v. Model based methods

C. *Prediction*

Regression techniques can be used for prediction of future values. Regression analysis can be used to model the relationship between more than one independent variables and dependent variables based on the criteria selected. The independent variables are viewed as attributes which are already known prior and the dependent variables are viewed as attributes what we want to predict. The techniques used for prediction such as decision trees, logistic regression or neural nets can be used necessarily for forecasting future values based on current values [18-20]. The model types used for both prediction as well as classification are same.

The types of regression methods are

i. Linear regression
ii. Multivariate Linear regression
iii. Nonlinear Regression
iv. Multivariate Nonlinear Regression

D. *Association rule*

Association and correlation are used to find frequent item set findings among large data sets based on the rules created. This type of finding can help in making business criteria and also to make effective decisions, such as customer shopping behaviour analysis, catalogue design, and marketing. Association Rule algorithms are often needed to generate rules with confidence values less than one [21-23]. The number of possible rules which are created for a given dataset is generally very large and a high proportion of the rules are usually a small value.

The types of association rule are

i) Multilevel association rule
ii) Multidimensional association rule
iii) Quantitative association rule

E. *Neural networks*

Neural networks are one of the most popular techniques of data mining. It consists of 3 layers input layer, hidden layer & output layer. These are set of connected input and output units; each connected units has an associated weight present with respective units. During the learning phase of the neural networks, network learns by adjusting weights so that it can be able to predict the correctly labelled class categories of the input tuples for the output prediction. The primary function of neurons of input layer is to divide input into neurons in hidden layer. It maps a set of input data onto a set of appropriate output data.

F. *K- Means Algorithms*

The k-means algorithm is one of the important clustering algorithms that are applied in a variety of real world applications. K-means algorithm groups the data in accordance with their characteristic values into k distinct cluster and the input data is taken for categorization into the same cluster having identical feature values. Let k be the positive integer denoting the number of clusters and it needs to be provided in advance. Then the pre processed heart disease data is clustered using the K Means algorithm with the K cluster of values.

G. *Genetic Algorithms*

Genetic algorithm is one of the frequently utilized technique of data mining used to solve variety of problems [4], here it is used in size reduction of the attribute data set and to get the optimal subset of attributes for heart disease prediction. In the prediction process, pairs of strings of the new generation are selected and techniques like crossover and mutation are applied to get better performance. With a known and certain probability, genes are mutated before all solutions are evaluated again. And this procedure can be repeated until a maximum number of generations are reached to achieve a good result.

IV. PROPOSED SYSTEM TO PREDICT HEART DISEASE

A. *KNN Algorithm*

KNN is a lazy supervised learning algorithm since it takes more time to train data until the data are used for classification. It is a method to classify the data using the training examples available in the feature space. It is used for classification and prediction. The classification phase divides the data into testing data and training data. The K nearest (Distance Calculation) training set data are found for each row of testing data and the classification is done by determining the majority vote by breaking the ties at random. Each neighbor is assigned a weight such that nearest neighbors contributes high to the average than the distant neighbors. The neighbors are taken based on the object dataset with reference to a known class. So no explicit time is required for training. If tie occurs for K_{th} nearest vector, all the candidates are included for voting. It is better to choose odd values for k in case of binary classification to avoid ties.

The main drawback of voting is more frequent classes dominate the prediction of new example.

B. *Average K nearest Neighbour Algorithm*

To remove the drawbacks and to make the KNN a faster algorithm AKNN is proposed here. In case of Average KNN, super sample is created for each class, which is the average of every training sample in that particular class, For example if there are n sample AKNN reduces the training sample size to n super samples. When the test samples are given the AKNN searches the sample data and find the closest to the input. The closest neighbor is identified by measuring the distance

between the neighbors. Three types of distance measures exist which are: Euclidean, Manhattan, Minkowski. All three types of distance measurements are valid for continuous variables. Euclidean is faster than other two functions. The Euclidean distance is measured by

Euclidean Distance D (x,y) = $(x_i - y_i)_{2ki=1}$

Where k- denotes the number of clusters,

x, y are co-ordinates of sample set.

The formula for Manhattan is given by,

Manhattan distance= $|X_i - Y_i|_{ni=1}$

Where X and Y are co-ordinates.

The minkowski distance is generally an Euclidean distance and is measured by

Min= $(|X_i - Y_i|_p)_{1/p}$

The algorithm for AKNN is as follows:

1. Let D be the samples used for training and k denotes the number of nearest neighbors

2. For each sample class create a super class (Average of every training set)

3. Compute Euclidean distance for every training sample

4. Classify the sample based on majority of class among the neighbors

5. End

The efficiency of the KNN algorithm depends upon the number of clusters chosen. A larger k value is precise and reduces overall noise. The main advantage in AKNN is that by grouping the samples based on super classes reduces the number for samples used for training, thus making the KNN a faster algorithm. KNN algorithm does not require time for the training phase and can adapt to changes in the training data. This works well if the data are well segregated and doesn't work if the data are noisy.

The heart disease database is obtained from the UCI archive is used for prediction [2]. This database consists of four data sets from the Cleveland Clinic Foundation, V.A. Medical Centre, Hungarian Institute of Cardiology and University Hospital of Switzerland as it provides a total of 920 records. Actually the database has 76 raw attributes with values that are related to heart disease [3]. The experiments carried out so far used and refer to 13 of these: We have taken 12 attributes from it to predict and classify the heart disease.

The Input attributes used are

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

2. CPT-Chest Pain Type (value: 1or 2 or 3) (value 1: angina; value 2: non-angina; value 3: asymptotic)

3. FBS- Fasting Blood Sugar (value 0 or 1)

4. RECG – Resting electrocardiographic (ECG) (value 0: normal; value 1: ST-T wave abnormality; value 2: definite left ventricular hypertrophy)

5. Ex-Ang – Exercise induced angina (value 0: no; value 1: yes)

6. Sl – slope of peak exercise ST segment (Value 1: Up sloping; Value 2: Flat; Value 3: down sloping)

7. Col-Ves– Number of major vessels colored by fluoroscopy (value 0 – 3)

8. Thal (Value 0: normal; Value 1: fixed defect; Value2: reversible defect)

9. Serum Cholesterol (mg/dl)

10. Thalach – maximum heart rate achieved

11. Old_Peak – ST depression induced by exercise relative to rest

12. Age in Years

Output Attribute is class which has value 0 for the category of no heart disease and value1 for no heart disease. The classification accuracy after attribute reduction is shown in Table 1 and Fig 1. The implementation part of this work is done using MATLAB12.

TABLE I. CLASSIFICATION ACCURACY AFTER ATTRIBUTE REDUCTION

Classification Techniques	Accuracy with	
	13 Attributes	12 Attributes
NaiveBayes	94.43	90.72
Decision Trees	96.1	96.62
Average K-Nearest Neighbor	96.5	97

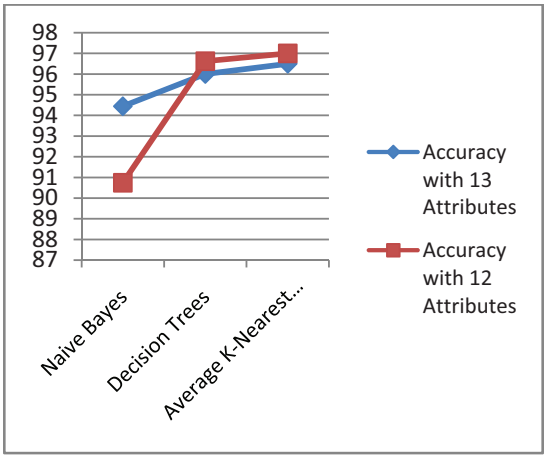

Fig. 1. Classification Accuracy of various data mining techniques

V. RESULTS AND DISCUSSIONS

The experimental result of heart diseases prediction gives the classification accuracy achieved is much better than the existing approaches and the implementation part have been done using MATLAB12. The data set taken is classified into two categories of heart disease and no heart disease. The classification accuracy achieved is higher than the previous approaches and the reduction of attributes gives more accuracy in predicting the results.

VI. CONCLUSION AND FUTURE SCOPE

The Proposed approach gives the higher efficiency and reduces complexity based on attribute reduction. The average K- nearest neighbor algorithm performs well and classifies the dataset of heart disease into two classes well when compared to traditional methods. The proposed work reduces the cost for different medical tests and helps the patients to take precautionary measures well in advance. In future the same method can also be applied in predicting and diagnosing other disease types.

ACKNOWLEDGEMENT

The author is thankful to her supervisor **Dr. G. M. Nasira**, Ph.D., Assistant Professor of Computer Science, Dept of Computer Applications, Chikkanna Govt. Arts College, Tiruppur for her valuable Guidance, Support and Mentorship. The author is also thankful to Tiruppur Kumaran College for women, Tiruppur.

REFERENCES

[1] Shantakumar B.Patil, Y.S.Kumaraswamy "Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network" *European Journal of scientific Research* ISSN 1450-216X Vol.31 No.4 pp.642-656, 2009.

[2] UCI machine learning repository and archive.ics.uci.edu/ml/datasets.html.

[3] ClevelandDatabase: http://archive.ics.uci.edu/ml/datasets/Heart+Disease

[4] Kalaiselvi. C, Dr. G.M. Nasira "Classification and Prediction of heart disease from diabetes patients using hybrid particle swarm optimization and library support vector machine algorithm" *International Journal of Computing Algorithm (IJCOA)*, ISSN. 2278-2397, volume 4 pp. 1403-1407, March 2015.

[5] Aqueel Ahmed, Shaikh Abdul Hannan "Data Mining Techniques to Find Out Heart Diseases: An Overview " *International Journal of Innovative Technology and Exploring Engineering (IJITEE),* ISSN: 2278-3075, Volume-1, Issue-4, pp. 18-23, September 2012.

[6] Sellappan Palaniappan, Rafiah Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", *International Journal of Computer Science and Network Security (IJCSNS),* Vol.8 No.8, pp. 343-350, August 2008.

[7] Heon Gyu Lee, Ki Yong Noh, Keun Ho Ryu, "Mining Bio signal Data: Coronary Artery Disease Diagnosis using Linear and Nonlinear Features of HRV," *Emerging Technologies in Knowledge Discovery and Data Mining*, LNAI 4819: pp. 56-66, May 2007.

[8] Niti Guru, Anil Dahiya, Navin Rajpal, "Decision Support System for Heart Disease Diagnosis Using Neural Network", *Delhi Business Review*, Vol. 8, No. 1, January - June 2007.

[9] Bakris GL. Preclinical diabetic cardiomyopathy: prevalence, screening and outcome. Internal Medicine University Department; pp. 1–27, 2009.

[10] Carlos Ordonez, "Improving Heart Disease Prediction Using Constrained Association Rules," Seminar Presentation at University of Tokyo, 2004.

[11] M.K. Ali, et al., "Diabetes and coronary heart disease: Current perspectives" *Indian journal of medical research*. 132(5):584–597, Nov 2010.

[12] Latha Parthiban and R.Subramanian, "Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm", *International Journal of Biological and Medical Sciences*, Vol. 3, No. 3, pp. 157-160, 2008.

[13] Kiyong Noh, Heon Gyu Lee, Ho-Sun Shon, Bum Ju Lee, and Keun Ho Ryu, "Associative Classification Approach for Diagnosing Cardiovascular Disease", *Springer,* Vol:345, pp: 721-727, 2006.

[14] Collins K, MS, RD, "The cancer, diabetes and heart disease Link" *Today's Dietitian*. 15(3):46 Mar 2013.

[15] Kalaiselvi C, Nasira.G M "Prediction of Heart Diseases and Cancer in Diabetic Patients Using Data Mining Techniques" *Indian Journal of Science and Technology* Vol 8(14), July 2015.

[16] M.Harris M "The role of primary health care in preventing the onset of chronic disease, with a particular focus on the lifestyle risk factors of obesity, tobacco and alcohol" *Centre for Primary Health Care and Equity, UNSW Canberra: National Preventive health taskforce,* pp. 1-21, 2008.

[17] Anderson KM. "Correlation of regional cardiovascular disease mortality in India with lifestyle and nutritional factors" *International J Cardiol*, 108: 291–300, 2006.

[18] Santhanam T, Ephzibah "Heart disease prediction using hybrid genetic fuzzy model" *Indian Journal of Science and Technology*, 8(9) May 2015.

[19] Kalaiselvi C, Nasira G M. "A New Approach for the diagnosis of diabetes and prediction of cancer using ANFIS" *WCCCT-14. IEEE Proceedings, International conference publications* Feb 2014.

[20] Olaniyi, Ebenezer Obaloluwa, Oyebade Kayode Oyedotun, and Khashman Adnan. "Heart Diseases Diagnosis Using Neural Networks Arbitration", *International Journal of Intelligent Systems and Applications*, IJISA, Vol. 7, No. 12, pp. 75-82, Nov 2015.

[21] Sivagowry, S., M. Durairaj, and A. Persia. "An empirical study on applying data mining techniques for the analysis and prediction of heart disease", *International Conference on Information Communication and Embedded Systems* (ICICES), 2013.

[22] Shouman, Mai, Tim Turner, and Rob Stocker."Using data mining techniques in heart disease diagnosis and treatment", *Japan-Egypt Conference on Electronics Communications and Computers*, 2012.

[23] Krisnaiah, V, M Srinivas, G Narsimha, and N Subhash Chandra "Diagnosis of heart disease patients using fuzzy classification technique", *International Conference on Computing and Communication Technologies*, 2014.