ISSN: 2229-6956(ONLINE)
DOI: 10.21917/ijsc.2013.0087

REVIEW OF HEART DISEASE PREDICTION SYSTEM USING DATA MINING AND HYBRID INTELLIGENT TECHNIQUES

R. Chitra[1] and V. Seenivasagam[2]

[1]*Department of Computer Science and Engineering, Noorul Islam Centre for Higher Education, India*
E-mail: jesi_chit@yahoo.co.in
[2]*Department of Information Technology, National Engineering College, India*
E-mail: yespee1094@yahoo.com

Abstract

The Healthcare industry generally clinical diagnosis is done mostly by doctor's expertise and experience. Computer Aided Decision Support System plays a major role in medical field. With the growing research on heart disease predicting system, it has become important to categories the research outcomes and provides readers with an overview of the existing heart disease prediction techniques in each category. Neural Networks are one of many data mining analytical tools that can be utilized to make predictions for medical data. From the study it is observed that Hybrid Intelligent Algorithm improves the accuracy of the heart disease prediction system. The commonly used techniques for Heart Disease Prediction and their complexities are summarized in this paper.

Keyword:
Neural Network, Hybrid Intelligent Algorithm, Heart Disease Prediction, Computer Aided Decision Support System

1. INTRODUCTION

Heart Diseases remain the biggest cause of deaths for the last two decades. Recently computer technology and machine learning techniques to develop software to assist doctors in making decision of heart disease in the early stage. The diagnosis of heart disease depends on clinical and pathological data. Heart disease prediction system can assist medical professionals in predicting heart disease status based on the clinical data of patients. In biomedical field data mining plays an essential role for prediction of diseases In biomedical diagnosis, the information provided by the patients may include redundant and interrelated symptoms and signs especially when the patients suffer from more than one type of disease of the same category. The physicians may not able to diagnose it correctly.

Data mining with intelligent algorithms can be used to tackle the said problem of prediction in medical dataset involving multiple inputs. Now a day's Artificial neural network has been used for complex and difficult tasks. The neural network is trained from the historical data with the hope that it will discover hidden dependencies and that it will be able to use them for predicting. Feed forward neural networks trained by back-propagation have become a standard technique for classification and prediction tasks.

The healthcare industry collects huge amounts of healthcare data and that need to be mined to discover hidden information for effective decision making. Discover of hidden patterns and relationships often go unexploited [6].

Clinicians and patients need reliable information about an individual's risk of developing Heart Disease. Ideally, they would have entirely accurate data and would be able to use a perfect model to estimate risk. Such a model would be able to categorize people with heart disease and others. Indeed, the perfect model would even be able to predict the timing of the disease's onset. The risk factors for heart disease can be divided into modifiable and non modifiable. Modifiable risk factors include obesity, smoking, lack of physical activity and so on. The non modifiable risk factors for heart disease are like age, gender, and family history. Many people have at least one heart disease risk factor.

Some kinds of heart disease are cardiovascular diseases, heart attack, coronary heart disease and Stroke. Stroke is a type of heart disease it is caused by narrowing, blocking, or hardening of the blood vessels that go to the brain or by high blood pressure [12, 13]. The rest of the paper is organized as follows. Section 2 describes the heart disease prediction system using data mining techniques and the intelligent and hybrid technique with feature subset selection are discussed in section 3 and 4 respectively.

2. HEART DISEASE PREDICTION USING DATAMINING

In this section the demining systems used for the classification of heart disease is analyzed.

N. Deepika et al. proposed Association Rule for classification of Heart-attack patients [1]. The extraction of significant patterns from the heart disease data warehouse was presented. The heart disease data warehouse contains the screening clinical data of heart patients. Initially, the data warehouse preprocessed to make the mining process more efficient. The first stage of Association Rule used preprocessing in order to handle missing values. Later applied equal interval binning with approximate values based on medical expert advice on Pima Indian heart attack data. The significant items were calculated for all frequent patterns with the aid of the proposed approach. The frequent patterns with confidence greater than a predefined threshold were chosen and it was used in the design and development of the heart attack prediction system. The, Pima Indian Heart attack dataset used was obtained from the UCI machine learning repository. Characteristics of the patients like number of times of chest pain and age in years were recorded. The actions comprised in the preprocessing of a data set 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. 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 [15]. In the real world, data is not always complete and in the case of the medical data, it is always true. To remove the number of inconsistencies which are associated with data we use Data preprocessing.

K. Srinivas et al. presented Application of Data Mining Technique in Healthcare and Prediction of Heart Attacks [2]. The potential use of classification based data mining techniques such as Rule based, Decision tree, Naïve Bayes and Artificial Neural Network to the massive Volume of healthcare data. Tanagra data mining tool was used for exploratory data analysis, machine learning and statistical learning algorithms. The training data set consists of 3000 instances with 14 different attributes. The instances in the dataset are representing the results of different types of testing to predict the accuracy of heart disease. The performance of the classifiers is evaluated and their results are analyzed. The results of comparison are based on 10 tenfold cross-validations. According to the attributes the dataset is divided into two parts that is 70% of the data are used for training and 30% are used for testing. The comparison made among these classification algorithms out of which the naive Bayes algorithm considered as the best performance algorithm. The performance of various algorithms is listed below [2].

Table.1. Performance Study of Data mining Algorithms

The algorithm used	Accuracy	Time taken
Naïve Bayes	52.33%	609ms
Decision list	52%	719ms
K-NN	45.67%	1000ms

Diagnosis of heart disease was used Naïve Bayes, K-NN, Decision List in this Naïve Bayes has taken a time to run the data for accurate result when compared to other algorithms.

Sudha et al. [11] to propose the classification algorithm like Naïve Bayes, Decision tree and Neural Network for predicting the stroke diseases. The classification algorithm like decision trees, Bayesian classifier and back propagation neural network were adopted in this study. The records with irrelevant data were removed from data warehouse before mining process occurs. Data mining classification technology consists of classification model and evaluation model. The classification model makes use of training data set in order to build classification predictive model. The testing data set was used for testing the classification efficiency. Then the classification algorithm like decision tree, naive Bayes and neural network was used for stroke disease prediction. The performance evaluation was carried out based on three algorithms and compared with various models used and accuracy was measured. While comparing these classification algorithms, the observation shows the neural network performance was more than the other two algorithms.

M A. Jabbar et al. proposed Association Rule mining based on the sequence number and clustering for heart attack prediction [16]. The entire database is divided into partitions of equal size. The dataset with 14 attributes was used in that work and also each cluster is considered one at a time for calculating frequent item sets. This approach reduces main memory requirement. To predict the heart attack in an efficient way the patterns are extracted from the database with significant weight calculation. The frequent patterns having a value greater than a predefined threshold were chosen for the valuable prediction of heart attack. Three mining goals were defined based on data exploration and all those models could answer complex queries in predicting heart attack.18].

Mai Shouman, et al. [21] proposed k-means clustering with the decision tree method to predict the heart disease. In their work they suggested several centroid selection methods for k-means clustering to increase efficiency. The 13 input attributes were collected from Cleveland Clinic Foundation Heart disease data set. The sensitivity, specificity, and accuracy are calculated with different initial centroids selection methods and different numbers of clusters. For the random attribute and random row methods, ten runs were executed and the average and best for each method were calculated. When comparing integrating k-means clustering and decision tree with traditional decision tree applied previously on the same data set, integrating k-means clustering with decision tree could enhance the accuracy of decision tree in diagnosing heart disease patients. In Addition, integrating k-means clustering and decision tree could achieve higher accuracy than the paging algorithm in the diagnosis of heart disease patients. The accuracy achieved was 83.9% by the enabler method with two clusters.

3. HEART ATTACK PREDICTION USING INTELLIGENT TECHNIQUES

In this section the role of intelligent technique such as neural network for heart disease prediction is explained

Latha Parthiban and R. Subramanian presented Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm [4]. Adaptable based fuzzy inputs are adapted with a modular neural network to rapidly and accurately approximate complex functions. The CANFIS model combined the neural network adaptive capabilities and the fuzzy logic quantitative approach then integrated with genetic algorithm to diagnosis the presence of the disease. Coactive neuro-fuzzy inference system model has good training performance and classification accuracies. Dataset of heart disease was obtained from UCI Machine Learning Repository .Coactive Neuro-fuzzy modeling was proposed as a dependable and robust method developed to identify a nonlinear relationship and mapping between the different attributes.

Dangare et al. proposed Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques [6]. Prediction System for heart disease used system contains huge amount of data, used to extract hidden information for making intelligent medical diagnosis. The main objective of this research was to build Intelligent Heart Disease Prediction System that gives diagnosis of heart disease using historical heart database. To develop the system, medical terms such as sex, blood pressure, and cholesterol like 13 input attributes are used. To get more appropriate results, two more attributes i.e. obesity and smoking, as attributes were considered as important attributes for heart disease. A Multi-layer Perceptron Neural Networks (MLPNN) that maps a set of input data onto a set of appropriate. It consists of 3 layers input layer, hidden layer & output layer. There is connection between each layer & weights are assigned to each connection. The primary function of neurons of input layer is to divide input into neurons in hidden layer. The dataset consists of total 573 records in heart disease database. The total records are divided into two data sets one is used for training consists of 303 records & another for testing consists of 270 records. Initially dataset contained some fields, in which some value in the records was missing. These were identified and replaced with most appropriate values using Replace Missing Values filter. The Replace Missing Values filter scans all records & replaces missing values with mean mode method known as Data Preprocessing. After pre-processing the data, data mining classification techniques such as Neural Networks, is used for classification. Many problems in business, science, industry, and medicine can be treated as classification problems. Owing to the wide range of applicability of ANN and their ability to learn complex and nonlinear relationships including noisy or less precise information, neural networks are well suited to solve problems in biomedical engineering. So here use for the neural network technique is classification of medical dataset 14 attributes by considering the single and multilayer neural network models [19].

Olatubosun Olabode et al. [22] to classify the Cerebrovascular disease by using artificial neural network with back propagation error method. The Multi-layer perceptrons artificial neural networks with back-propagation error method were feed-forward nets with one or more layers of nodes between the input and output nodes. These additional layers contain hidden units or nodes that were not directly connected to both the input and output nodes. The neural network was trained using back propagation algorithm with sigmoid function on one hidden layer with the 16 input attributes. Predictive models were used in variety of domains for the diagnosis. Dataset for this work were collected 100 records (60 males and 40 females) from federal medical fields. The input values obtained from the records of the forms the input variables in the input layer with 16 nodes. The neural network weights were initialized randomly. This work range of the weights was between [-0.5 and 0.5] and the learning rate was set between 0.1 and 0.9. The training, validation, generalization accuracy was measured.

4. HEART DISEASE WITH FEATURE SUBSET SELECTION

Feature subset selection is one of the technique used for dimensionality reduction and thus to reduce the complexity of the algorithm. But the selection of appropriate feature is challenging one. In this section the heart disease prediction system with evolutionary feature selection is explained.

M. Anbarasi et al. proposed Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm [5]. Originally 13 attributes involved in prediction of heart disease, proposed enhanced prediction of heart disease with feature subset selection using genetic algorithm using 10 attributes for predicting and data mining techniques after incorporating feature subset selection with high model construction time. Classification techniques are Naïve Bayes, Decision Tree and Classification by clustering. The genetic search starts with zero attributes, and an initial population with randomly generated rules. Based on the idea of survival of the fittest, new population is constructed to comply with fittest rules in the current population, as well as offspring of these rules. Feature Extraction is the process of detecting and eliminating irrelevant, weakly relevant or redundant attributes or dimensions in a given data set. The goal of feature selection is to find the minimal subset of attributes such that the resulting probability distribution of data classes is close to original distribution obtained using all attributes. The reduced data set fed to three classification models. K fold cross validation method is used as the test mode. Genetic algorithm is used to determine the attributes which contribute more towards the diagnosis of heart ailments which indirectly reduces the number of tests which are needed to be taken by a patient. The 13 attributes are reduced to 6 attributes using genetic search. Subsequently, three classifiers like Naive Bayes.

D. Shanthi, et al. [7] proposed to functional model of ANN to aid existing diagnosis methods. The Back propagation algorithm was used to train the ANN architecture and the same had been tested for the various categories of stroke disease. The data for this study had been collected from 50 patients who had symptoms of stroke disease. The data had standardized so as to be error free in nature. All the fifty cases were analyzed after careful scrutiny with the help of the Physicians. Data were analyzed in the dataset to define column parameters and data anomalies. Data analysis information needed for correct data preprocessing. After data analysis, the values had been identified as missing, wrong type values or outliers and which columns were rejected as unconvertible for use with the neural network. In this study, backward stepwise method was used for input feature selection. The removal of insignificant inputs had improved the generalization performance of a neural network. This method begins with all inputs and it works by removing one input at each step. At each step, the algorithm finds an input that least deteriorates the network performance and becomes the candidate for removal from the input set. The architecture of the neural network used in this study was the multilayered feed-forward network architecture with 20 input nodes, 10 hidden nodes, and 10 output nodes. The number of input nodes was determined by the finalized data; the number of hidden nodes was determined through trial and error; and the numbers of output nodes were represented as a range showing the disease classification. The most widely used neural-network learning method is the BP algorithm. Learning in a neural network involves modifying the weights and biases of the network in order to minimize a cost function.

In this prediction of combinations of several targets attributes for intelligent and effective heart attack prediction using data mining. For predicting heart attack, significantly 15 attributes are listed and with basic data mining technique other approaches e.g. ANN, Time Series, Clustering and Association Rules, soft computing approaches etc. can also be incorporated and performance were analyzed. In compare to the performance of predictive data mining technique on the same dataset and the outcome reveals that Decision Tree outperforms and some time Bayesian classification is having similar accuracy as of decision tree but other predictive methods like KNN, Neural Networks, Classification based on clustering are not performing well. The second conclusion is that the accuracy of the Decision Tree and Bayesian Classification further improves after applying genetic algorithm to reduce the actual data size to get the optimal subset of attribute sufficient for heart disease prediction [17].

The Application of Artificial Neural Network (ANN) can be time-consuming due to the selection of input features for the Multi Layer Perceptron. The number of layers, number of neurons in each layer was also determined by the input attributes. Reducing the dimensionality, or selecting a good subset of features, without sacrificing accuracy, was of great importance for neural networks to be successfully applied to the area. D. Shanthi [5] propose a neuro-genetic approach to feature selection in disease classification.

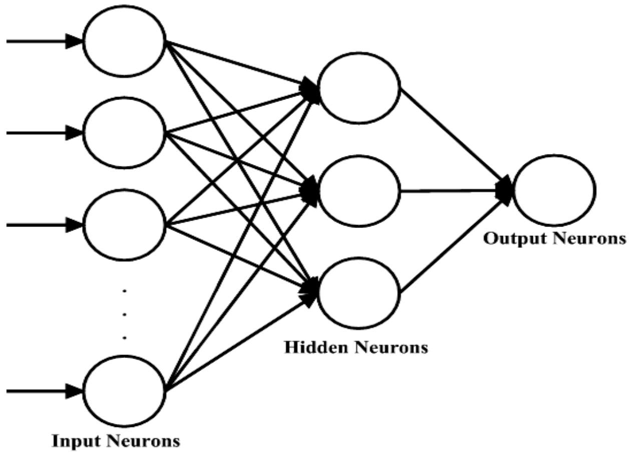

Fig.1. Architecture of MLP

The Fig.1 shows schematically a typical representation of a MLP with input neurons, four hidden neurons, and one output neuron. Each of the input neurons connects to each of the hidden neurons, and each of the hidden neurons connects to the output neuron. The Back propagation algorithm, in particular, adaptively changes the internal network free parameters based on external after trained, a neural network can make predictions about the membership of every test example. MLP is trained with the Back propagation algorithm suffers from the high number of parameters that need to be tuned, like learning rate, number of neurons, momentum rate, etc. However, the motivations to select this algorithm arise after observing that they had been used to solve problems in different domains, moreover, the output can be directly used for ranking purposes.

The Back propagation algorithm, in particular, adaptively changes the internal network free parameters based on external stimulus. After trained, a neural network can make predictions about the membership of every test example. MLP was trained with the Back propagation algorithm suffers from the high number of parameters that need to be tuned, like learning rate, number of neurons, momentum rate, etc. However, the motivations to select this algorithm arise after observing that they have been used to solve problems in different domains, moreover, the output can be directly used for ranking purposes.

The fuzzy neuro expert system with genetic feature reduction is used for diagnosis of heart disease [20].This system will help the doctors to arrive at a decision about the presence or absence of heart disease in patients.

5. DISCUSSION

The numerous heart attack predicting system techniques presented in this paper. In this paper Heart attack prediction system methodology is categorized in three types. At first type data mining technique (mainly classification technique) are analyzed. The second type intelligent techniques used for heart disease prediction are analyzed. The final type the role of feature subset in the heart disease prediction is discussed.

In data mining approach the heart disease data warehouse contains the screening clinical data of heart patients used for heart disease diagnosis. The classification technique is used in all proposed work.

In intelligent technique neural network is used for disease prediction. The MLFFNN with back propagation algorithm is proposed in many papers discussed in this section. The main drawback of this system is training time and complexity .The offline training of the neural network can be advised to reduce the time complexity. In the final model feature subset selection only the more significant attributed are extracted to predict accurate result. Data mining techniques combined with intelligent and evolutionary computation are discussed in the reviewed paper. From the result it is observed that the accuracy of the system improved with the Feature subset selection. But in this technique also time complexity is high. Selection of the algorithm for feature reduction is still challenging.

In many papers author used the dataset of Heart disease was obtained from UCI Machine Learning Repository, University of California. Hence it might be a good choice for training the network.

6. CONCLUSION

Heart disease is one of the leading causes of death worldwide and the early prediction of heart disease is important. The computer aided heart disease prediction system helps the physician as a tool for heart disease diagnosis. Some Heart Disease classification system is reviewed in this paper. From the analysis it is concluded that, data mining plays a major role in heart disease classification. Neural Network with offline training is a good for disease prediction in early stage and the good performance of the system can be obtained by preprocessed and normalized dataset. The classification accuracy can be improved by reduction in features.

REFERENCE

[1] N. Deepika and K. Chandra shekar, “Association rule for classification of Heart Attack Patients”, *International Journal of Advanced Engineering Science and Technologies*, Vol. 11, No. 2, pp. 253 – 257, 2011.

[2] K. Srinivas, B. Kavitha Rani and Dr. A. Govrdhan, “Application 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, 2011.

[3] Asha Rajkumar and B. Sophia Reena, “Diagnosis Of Heart Disease Using Data mining Algorithm” , *Global Journal of Computer Science and Technology*, Vol. 10, No. 10, pp. 38 - 43, 2010

[4] Latha Parthiban and R. Subramanian, “Intelligent Heart Disease Prediction System using CANFIS and Genetic

Algorithm", *International Journal of Biological and Life Science*, Vol. 15, pp. 157 - 160, 2007.

[5] M. Anbarasi, E. Anupriya and N.CH.S.N. Iyenga, "Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm", *International Journal of Engineering Science and Technology*, Vol. 2, No. 10, pp. 5370 - 5376, 2010

[6] Chaitrali S. Dangare and Sulabha S. Apte, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", *International Journal of Computer Applications*, Vol. 47, No. 10, pp. 0975 – 888, 2012

[7] D. Shanthi, G. Sahoo and Dr. N. Saravanan, "Designing an Artificial Neural Network Model for the Prediction of Thrombo-embolic Stroke", *International Journal of Biometric and Bioinformatics*, Vol. 3, No. 1, pp. 250 - 255, 2008.

[8] P. K. Anooj, "Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules", *Journal of King Saud University Computer and Information Sciences*, Vol. 11, pp. 309 - 314, 2011.

[9] J. C. Obi and A. A. Imainvan, "Decision Support System for the Intelligent Identification of Alzheimer using Neuro Fuzzy logic", *International Journal on Soft Computing* , Vol. 2, No. 2, pp. 25 - 38, 2011.

[10] D. Shanthi, G. Sahoo and Dr. N. Saravanan, "Evolving Connection Weights of Artificial Neural Network Using Genetic Algorithm With Application to the Prediction Stroke Diseases", *International Journal of Soft Computing*, Vol. 2, pp. 95 - 101, 2009.

[11] A. Sudha, P. Gayathiri and N. Jaisankar, "Effective Analysis and Predictive Model of Stroke Disease using Classification Methods", *International Journal of Computer Applications*, Vol. 43, No. 14, pp. 0975 – 8887, 2012.

[12] Tom Dent, "*Predicting the risk of coronary heart disease*", PHG foundation publisher, 2010.

[13] World Health Organization, "*Global status report on no communicable diseases*", 2010.

[14] World Health Organization, "*Pocket Guidelines for Assessment and Management of Cardiovascular Risk*", 2007.

[15] D. Shanthi, G. Sahoo and N. Saravanan "Input Feature Selection using Hybrid Neuro-Genetic Approach in the Diagnosis of Stroke Disease", *International Journal of Computer Science and Network Security*, Vol. 8, No.12, pp. 99 - 106, 2008.

[16] M A. Jabbar, Priti Chandra and B. L. Deekshatulu, "Cluster based association rule mining for heart attack prediction", *Journal of Theoretical and Applied Information Technology*, Vol. 32, No.2, pp. 197 - 201, 2011.

[17] Jyoti Soni, Ujma Ansari and Dipesh Sharma, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", *International Journal of Computer Applications*, Vol. 17, No. 8, pp. 43 – 48, 2011.

[18] K. Srinivas, G. Raghavendra Rao and A. Govardhan, "Survey on prediction of heart morbidity using data mining techniques", *International Journal of Data Mining & Knowledge Management Process*, Vol. 1, No. 3, pp. 14 - 34, 2011.

[19] K. Usha Rani, "Analysis of heart diseases dataset using neural network approach", *International Journal of Data Mining and Knowledge Management Process* ,Vol. 1, No. 5, pp. 1 - 8, 2011.

[20] E. P. Ephzibah and V. Sundarapandian, "A neuro fuzzy expert system for heart disease diagnosis", *An International Journal Computer Science & Engineering*, Vol. 2, No. 1, pp. 17 - 23, 2012.

[21] Mai Shouman, Tim Turner and Rob Stocker, "Integrating Decision Tree and K-Means Clustering with Different Initial Centroid Selection Methods in the Diagnosis of Heart Disease Patients", *Proceedings of the International Conference on Data Mining*, 2012.

[22] Olatubosun Olabode and Bola Titilayo Olabode, "Cerebrovascular Accident Attack Classification Using Multilayer Feed Forward Artificial Neural Network with Back Propagation Error", *Journal of Computer Science*, Vol. 8, No. 1, pp.18 - 25, 2012.