Volume 4, Issue 1, January 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Study of Heart Disease Prediction using Data Mining

K.Sudhakar,
Research Scholar, Dept., of Computer Science
Shrimati Indira Gandhi College
Bharathidasan University
Tiruchirappalli, Tamil Nadu, India

Dr. M. Manimekalai
Director, Department of MCA
Shrimati Indira Gandhi College
Bharathidasan University
Tiruchirappalli, Tamil Nadu, India

Abstract— The Healthcare industry is generally "information rich", which is not feasible to handle manually. These large amounts of data are very important in the field of Data Mining to extract useful information and generate relationships amongst the attributes. The doctors and experts available are not in proportion with the population. Also, symptoms often be neglected. Heart disease diagnosis is a complex task which requires much experience and knowledge. Heart disease is a single largest cause of death in developed countries and one of the main contributors to disease burden in developing countries. In the health care industry the data mining is mainly used for predicting the diseases from the datasets. The Data Mining techniques, namely Decision Trees, Naive Bayes, Neural Networks, Associative classification, Genetic Algorithm are analyzed on Heart disease database.

Keywords: Heart disease, Data Mining, Associative Classification.

I. INTRODUCTION

"Data Mining is a non-trivial extraction of implicit, previously unknown and potential useful information about data [1]. In short, it is a process of analyzing data from different perspective and gathering the knowledge from it. The discovered knowledge can be used for different applications for example healthcare industry. Nowadays healthcare industry generates large amount of data about patients, disease diagnosis etc. Data mining provides a set of techniques to discover hidden patterns from data. A major challenge facing Healthcare industry is quality of service. Quality of service implies diagnosing disease correctly & provides effective treatments to patients. Poor diagnosis can lead to disastrous consequences which are unacceptable.

II. HEART DISEASE

The heart is important organ or part of our body. Life is itself dependent on efficient working of heart. If operation of heart is not proper, it will affect the other body parts of human such as brain, kidney etc. It is nothing more than a pump, which pumps blood through the body. If circulation of blood in body is inefficient the organs like brain suffer and if heart stops working altogether, death occurs within minutes. Life is completely dependent on efficient working of the heart. The term Heart disease refers to disease of heart & blood vessel system within it.

There are number of factors which increase the risk of Heart disease [2]:

- Family history of heart disease
- Smoking
- Cholesterol
- Poor diet
- High blood pressure
- High blood cholesterol
- Obesity
- Physical inactivity
- Hyper tension

Symptoms of a Heart Attack

Symptoms of a heart attack can include:

- Discomfort, pressure, heaviness, or pain in the chest, arm, or below the breastbone.
- Discomfort radiating to the back, jaw, throat, or arm.
- Fullness, indigestion, or choking feeling (may feel like heartburn).
- Sweating, nausea, vomiting, or dizziness.
- Extreme weakness, anxiety, or shortness of breath.
- Rapid or irregular heartbeats

Types of Heart diseases

Heart disease is a broad term that includes all types of diseases affecting different components of the heart. Heart means 'cardio.' Therefore, all heart diseases belong to the category of cardiovascular diseases.

Some types of Heart diseases are

1. Coronary heart disease

It also known as coronary artery disease (CAD), it is the most common type of heart disease across the world. It is a condition in which plaque deposits block the coronary blood vessels leading to a reduced supply of blood and oxygen to the heart.

2. Angina pectoris

It is a medical term for chest pain that occurs due to insufficient supply of blood to the heart. Also known as angina, it is a warning signal for heart attack. The chest pain is at intervals ranging for few seconds or minutes.

3. Congestive heart failure

It is a condition where the heart cannot pump enough blood to the rest of the body. It is commonly known as heart failure.

4. Cardiomyopathy

It is the weakening of the heart muscle or a change in the structure of the muscle due to inadequate heart pumping. Some of the common causes of cardiomyopathy are hypertension, alcohol consumption, viral infections, and genetic defects.

5. Congenital heart disease

It also known as congenital heart defect, it refers to the formation of an abnormal heart due to a defect in the structure of the heart or its functioning. It is also a type of congenital disease that children are born with.

6. Arrhythmias

It is associated with a disorder in the rhythmic movement of the heartbeat. The heartbeat can be slow, fast, or irregular. These abnormal heartbeats are caused by a short circuit in the heart's electrical system.

7. Myocarditis

It is an inflammation of the heart muscle usually caused by viral, fungal, and bacterial infections affecting the heart. It is an uncommon disease with few symptoms like joins pain, leg swelling or fever that cannot be directly related to the heart.

III. LITERATURE SURVEY

Numerous studies have been done that have focus on diagnosis of heart disease. They have applied different data mining techniques for diagnosis & achieved different probabilities for different methods.

- An Intelligent Heart Disease Prediction System (IHDPS) is developed by using data mining techniques Naive Bayes, Neural Network, and Decision Trees was proposed by SellappanPalaniappan et al .[3]. Each method has its own strength to get appropriate results. To build this system hidden patterns and relationship between them is used. It is web-based, user friendly & expandable.

- To develop the multi-parametric feature with linear and nonlinear characteristics of HRV (Heart Rate Variability) a novel technique was proposed by HeonGyu Lee et al. [5]. To achieve this, they have used several classifiers e.g. Bayesian Classifiers, CMAR (Classification based on Multiple Association Rules), C4.5 (Decision Tree) and SVM (Support Vector Machine).

- The prediction of Heart disease, Blood Pressure and Sugar with the aid of neural networks was proposed by Niti Guru et al. [4]. The dataset contains records with 13 attributes in each record. The supervised networks i.e. Neural Network with back propagation algorithm is used for training and testing of data.

- The problem of identifying constrained association rules for heart disease prediction was studied by Carlos Ordonez [7]. The resultant dataset contains records of patients having heart disease. Three constraints were introduced to decrease the number of patterns [6]. They are as follows:
  1. The attributes have to appear on only one side of the rule.
  2. Separate the attributes into groups. i.e. uninteresting groups.
  3. In a rule, there should be limited number of attributes.

  The result of this is two groups of rules, the presence or absence of heart disease.

- Franck Le Duff et al. [9] builds a decision tree with database of patient for a medical problem.

- LathaParthiban et al. [10] projected an approach on basis of coactive neuro-fuzzy inference system (CANFIS) for prediction of heart disease. The CANFIS model uses neural network capabilities with the fuzzy logic and genetic algorithm.

- Kiyong Noh et al. [8] uses a classification method for the extraction of multiparametric features by assessing HRV (Heart Rate Variability) from ECG, data pre-processing and heart disease pattern. The dataset consisting of 670 peoples, distributed into two groups, namely normal people and patients with heart disease, were employed to carry out the experiment for the associative classifier.

- ShrutiRatnakar et al. used genetic algorithm to reduce the set of attributes of Naïve Bayes generate relationship amongst the attributes.

- AkhilJabbar et al. proposes efficient associative classification algorithm using genetic approach for heart disease prediction. The main motivation for using genetic algorithm inthe discovery of high level prediction rules is that the discovered rules are highly comprehensible, having highpredictive accuracy and of high interestingness values.

VI. DATA MINING TECHNIQUES USED FOR PREDICTIONS

The three different data mining classification techniques, i.e. Neural Networks, Decision Trees, and Naive Bayes are used to analyze the dataset.

4.1. Neural Networks

An artificial neural network (ANN), often just called a "neural network" (NN), is a mathematical model or computational model based on biological neural network. In other words, it is an emulation of biological neural system [13]. A Multi-layer Perceptron Neural Networks (MLPNN) is used.

It maps a set of input data onto a set of appropriate output data. 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 xi into neurons in hidden layer. Neuron of hidden layer adds input signal xi with weights wji of respective connections from input layer. The output Yj is function of $Yj = f(\Sigma\ wji\ xi)$ Where f is a simple threshold function such as sigmoid or hyperbolic tangent function.

4.2. Decision Trees

The decision tree approach is more powerful for classification problems. There are two steps in this techniques building a tree & applying the tree to the dataset. There are many popular decision tree algorithms CART, ID3, C4.5, CHAID, and J48. From these J48 algorithm is used for this system. J48 algorithm uses pruning method to build a tree. Pruning is a technique that reduces size of tree by removing over fitting data, which leads to poor accuracy in predications. The J48 algorithm recursively classifies data until it has been categorized as perfectly as possible. This technique gives maximum accuracy on training data. The overall concept is to build a tree that provides balance of flexibility & accuracy.

4.3. Naive Bayes

Naive Bayes classifier is based on Bayes theorem. This classifier algorithm uses conditional independence, means it assumes that an attribute value on a given class is independent of the values of other attributes.

The Bayes theorem is as follows:

Let X={x1, x2, ...,xn} be a set of n attributes.

In Bayesian, X is considered as evidence and H be some hypothesis means, the data of X belongs to specific class C.

We have to determine P (H|X), the probability that the hypothesis H holds given evidence i.e. data sample X.

According to Bayes theorem the P (H|X) is expressed as

$P(H|X) = P(X|H)\ P(H) / P(X)$

V. ASSOCIATIVE CLASSIFICATION

Associative classification is a recent and rewarding technique which integrates association rule mining and classification to a model for prediction and achieves maximum accuracy. Associative classifiers are especially fit to applications where maximum accuracy is desired to a model for prediction.

Association rule mining and classification are two main functionalities of data mining. Association rule mining is used to find associations or correlations among the item sets. It is a unsupervised learning where no class attribute is involved in finding the association rule. On the other hand, classification is a supervised learning where class attribute is involved in the construction of the classifier and is used to classify or predict the data unknown sample.

Associative classification involves two stages.

1) Generate class based association rules from a training data set
2) Classify the test data set into predefined class labels.

VI. CONCLUSION

The overall objective is to study the various data mining techniques available to predict the heart disease and to compare them to find the best method of prediction.

REFERENCES

[1.] Frawley and G. Piatetsky -Shapiro, Knowledge Discovery in Databases: An Overview. Published by the AAAI Press/ The MIT Press, Menlo Park, C.A 1996.

[2.] 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.

[3.] 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

[4.] 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).

[5.] HeonGyu Lee, Ki Yong Noh, KeunHoRyu, "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.] ShantakumarB.Patil, Y.S.Kumaraswamy *"Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network"*. ISSN 1450-216X Vol.31 No.4 (2009), pp.642-656.

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

[8.] Kiyong Noh, HeonGyu Lee, Ho-Sun Shon, Bum Ju Lee, and KeunHoRyu, *"Associative Classification Approach for Diagnosing Cardiovascular Disease*", Springer, Vol:345, pp: 721- 727, 2006.

[9.] Franck Le Duff, CristianMunteanb, Marc Cuggiaa, Philippe Mabob, *"Predicting Survival Causes After Out of Hospital Cardiac Arrest using Data Mining Method"*, Studies in health technology and informatics, Vol. 107, No. Pt 2, pp. 1256-9, 2004.

[10.] LathaParthiban and R.Subramanian, *"Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm"*, International Journal of Biological, Biomedical and Medical Sciences, Vol. 3, No. 3, 2008.

