Research Paper

Volume : 3 | Issue : 9 | September 2014 • ISSN No 2277 - 8179

# Different Data Mining Techniques Involved in Heart Disease Prediction: A Survey

Computer Science

KEYWORDS : Data Mining, Classification, Neural Network, Decision Tree, Naïve Bayes.

| | |
|---|---|
| **K. THENMOZHI** | Asst.Professor, Dr.N.G.P. Arts and Science College, Bharathiar University, Coimbatore. |
| **P. DEEPIKA** | Research Scholar, Dr.N.G.P. Arts and Science College, Bharathiar University, Coimbatore. |
| **M.MEIYAPPASAMY** | Research Scholar, Dr.N.G.P. Arts and Science College, Bharathiar University, Coimbatore. |

## ABSTRACT

*Data mining is used to convert through very large amount of data for useful information. In Data Mining, some of the most important and popular techniques are used. Those are Classification, Clustering, Prediction, Sequential patterns and Association Rules. Heart disease is the term that assigns to a large number of medical conditions related to heart. These medical conditions are used to describe the abnormal conditions of health that directly influence the heart and all its parts. This paper aims at analyzing the various data mining techniques introduced in recent years for heart disease prediction.*

## 1. INTRODUCTION:

"Data mining is a non-trivial extraction of implicit, previously unknown and potential useful information about data" [1]. Data mining is the process of finding previously unknown patterns and trends in database and using that information to built predictive models. Today health care industry generates large amount of complex data about patients, hospitals, resources, diseases and their records etc. this large amount of data is an essential resource to be processed and analyzed for knowledge extraction that supports for cost saving and decision-making. Heart disease prediction system can assist medical professionals in predicting heart disease status based on the clinical data of the patient. Data mining in health care is an emerging field of high importance for providing prognosis and a deeper understanding of medical data. In Biomedical diagnosis, the information provided by the patients may include redundant and interrelated symptoms and signs especially when the patient suffers from more than one type of disease of the same category.

## 2. HEART DISEASE:

Heart is the important part of every human body. It is nothing more than a pump, which pumps blood through the body. If circulation of blood in body is insufficient the organs like a brain suffer and if heart stops working altogether, death occurs within a minutes. The term heart disease refers to disease of heart and blood vessel system within it.

**The factors that are included in heart disease are:**

- ★ Age
- ★ Smoking
- ★ Family hereditary
- ★ Poor diet
- ★ High blood pressure
- ★ High blood cholesterol
- ★ Obesity
- ★ Physical inactivity
- ★ Hyper tension

Medical data mining has high potential for exploring the hidden patterns in the data sets of the medical domain. These data should be collected in an organized form. Data mining technology provides a user oriented approach to novel and hidden patterns in the data. From the analysis of World Health Organization, they estimated 12 million deaths occur worldwide, every year due to Heart disease.

## 3. RESEARCH THECHNIQUES IN SURVEY

This paper surveys various classification algorithms in data mining is used to predict the heart disease using various algorithms.

Nidhi Bhatala et al. developed a model that compared the classification techniques to find the highest accuracy [2]. This heart disease prediction system used 15 attributes [3]. Earlier 13 attributes were used for prediction but this research work incorporated 2 more attributes (obesity and smoking) for efficient diagnosis of heart disease.

Table-1show the outcome of this research work and it shows that neural network outplayed over other data mining techniques.

| Classification Techniques | Accuracy |
|---|---|
| Naïve Bayes | 90.74% |
| Decision Trees | 99.62% |
| Neural Networks | 100% |

**Table1: Various data mining Techniques Comparison**

K.Srinivas et al. presented Application of data mining techniques in Healthcare and Prediction of heart Attacks [4]. The potential use of classification based data mining techniques such as Rule based, Decision tree, Naïve Bayes and Artificial Neural Network to massive Volume of healthcare data. Tanagra data mining tool was used for exploratory data analysis, machine learning and statistical learning algorithms. The training dataset consists of 3000 instances with 14 different attributes. The instances in the dataset are representing the result of different types of testing to predict the accuracy of Heart disease. The performance of the classifier is evaluated and their results are analyzed. The results of comparison are based on 10 tenfold cross validations. The comparison made among these classification algorithms out of which the Naïve Bayes algorithm considered as the best performance algorithm.

**Table2. Performance Study of Data Mining Algorithms**

| Algorithm Used | Accuracy | Time Taken |
|---|---|---|
| Naïve Bayes | 52.33% | 609ms |
| Decision tree | 52% | 719ms |
| k-NN | 45.67% | 1000ms |

Jyoti soni et.al proposed three different supervised machine learning algorithms [5]. They are Naïve Bayes, K-NN and Decision Tree algorithms. These algorithms have been used for analyzing the heart disease dataset. Tanagra data mining tool is

used for classifying these data. These classified data is evaluated using 10 fold cross validation and the results are compared. Decision tree is one of the popular and important classifier which is easy and simple to implement. It handle huge amount of dimensional data. It is more suitable for exploratory knowledge discovery. The result given by the Decision Tree is easier to interpret and read. Naïve Bayes is a statistical classifier which assigns no dependency between attributes. To determine the class the posterior probability should be maximized. The advantages are one can work with the Naïve Bayes model without using any Bayesian methods. K-Nearest Neighbor's algorithm is one of the most important methods for classifying objects based on closest training data in the feature space. It is simplest among all machine learning algorithms but, the accuracy of K-NN algorithm can be degraded by presence of noisy features. This observation is performed using training to consist 3000 instances with 14 different attributes. The dataset is divided into two testing and training i.e. 70% of data are used for training and 30% is used for testing. The authors concluded that Naïve Bayes algorithm performs well when compared to other algorithms.

Jafreen Hossian et.al proposed three classification algorithms are chosen for the purpose of accuracy benchmarking in clinical data, which are the Naïve Bayes, Multilayer Perception (MLP) and Decision Tree. A Naïve Bayes classifier is an important classifier based on applying Bayesian theorem with strong independence assumptions. In simple terms, a Naïve Bayes classifier assumes that the presence of a particular feature of a class is unrelated to the presence (or absence) of any other feature, given the class variable. A Multilayer Perception is a feed-forward artificial neural network model that maps sets of input data onto a set of appropriate output. An MLP model consists of multiple layers of nodes in a directed graph, with each layer is fully connected to the next one. The Multilayer Perception utilizes a supervised learning technique called back-propagation for the training network. MLP is the modification of the standard linear perceptions and can distinguish data that is not linearly separable. C4.5 is an algorithm used to generate a Decision Tree developed by Quinlan (1993). C4.5 is an extension of Quinlan's earlier ID3 algorithm. The dataset consists 13 attributes. From the observation, the average of ROC area using Naïve Bayes was 88.8%, where as 82.4% and 78.7% for Multilayer Perception and Decision Tree respectively. After feature selection, the accuracy for Naïve Bayes is 88%, 86% and 79% for Multilayer Perception and decision Tree respectively. Though Naïve Bayes shows better results in their experiment, they suggest Multilayer perception since Naïve Bayes algorithm assumes independency among variables whereby in real-life situation the variables are inter-dependent among each other. They also suggest to use Multilayer Perception classification algorithm together with the filtering method of attribute select classifier in WEKA, which produced result to increase the accuracy from 82.4% to 86%.

Chaitrali S.Dangare et.al proposed three classification algorithms namely Naïve Bayes, Decision Tree and Neural Network [3]. The performances of these techniques are compared based on the accuracy. The main objective of this research is to build an Intelligent Heart Disease Prediction System that is used to diagnosis the heart disease using historical data related to heart. Basically all the heart diseases prediction methods use 13 attributes. But in this research paper they include two more attributes that are obesity and smoking. In this research data are collected from publicly. The Cleveland Heart Disease dataset consists of 303 records and Statlog Heart Disease dataset consists of 270 records. The dataset consists three types of attributes that are Input, Key and predictable attribute. An Artificial Neural Network is a mathematical model or a computational model based on biological neural network. In this research work a Multilayer Perception Neural Network is used. Decision Tree is the most important classification technique which is used widely. There are many decision tree algorithms: CART, ID3, C4.5, J48 and CHAID. In this research J48 decision tree algorithm is used. Naïve Bayes Classifier is based on Bayesian Theorem. This classifier assumes that an attribute value on a given class is independent of values of other attributes. In this research, a dataset which consist 573 records in Heart Disease database. The overall objective of this research is to find the more accuracy in Heart Disease. This paper shows that Neural Network provides more accurate results than the other two classification techniques.

## 4. CONCLUSION

In this survey paper, the different data mining algorithms are used in the field of Heart Disease is discussed. Mainly the three different classification algorithms namely Naïve Bayes, Decision Tree and Neural Networks are focused in the field of Heart Disease Prediction. This paper provides various results of using these three classification techniques. But the target of each research work is to find the better accuracy in Heart Disease. Most of this Research work shows that Neural Network provides more accuracy compared to Naïve Bayes and Decision Tree. Some Research paper provides additional points that in case of changing the number of attributes the performance of these three algorithms are varying.

## REFERENCE

1. Frawley and G.Piatetsky-Shapiro, "Knowledge Discovery in Database: An Overview", Published by the AAAJ Press/ the MIT Press, Menlo Park, C.A 1996. | 2. Nidhi Bhatala, KiranJyothi, "An Analysis of Heart Disease Prediction using Different Data mining Techniques" International Journal of Engineering Research Technology(IJERT) ISSN:2278-0181 vol.1, Issue-8, October-2012. | 3. Chaitrali S.Dangare,Sulabha S.Apte, "Improved study of Heart Disease Prediction system Using Data mining Classification Techniques", International Journal of Computer Applications(0975-888) Vol-47-No.10, June 2012. | 4. K.Srinivas, B.Kavitha Rani and Dr.A.Govardhan," 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. | 5. Jyoti Soni, Sunita Soni et al., "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. | 6. Jafreen Hossain, Nor FazlidaMohdSani , Aida Mustapha and Lilly SurianiAffendey, " Using Feature Selection as accuracy benchmarking in Clinical Data Mining"; Journal of Computer Science, ISSN:1549-3636, 2013. |