International Journal of Modern Computer Science (IJMCS)
Volume 4, Issue 1, February, 2016
ISSN: 2320-7868 (Online)

Analysis of Cardiovascular Heart Disease Prediction Using Data Mining Techniques

Sharan Monica L [1]
M.Phil Scholar
Department of Computer Science
Bishop Heber College
Trichy, India
E-mail: monica.sharanAL@gmail.com

Sathees Kumar B[2]
Assistant Professor
Department of Computer Science
Bishop Heber College
Trichy, India
E-mail: satbhc@yahoo.com

Abstract: Hear diseases are the number one cause of death. The health industry is generally "Information rich" but "Knowledge poor" which is not possible to handle manually. The data mining is used to predict the disease from the datasets. Knowledge discovery in databases can help organizations turn their data into information. This research paper intends to provide a survey of current techniques of knowledge discovery in databases using data mining techniques which will be useful for medical practitioners to take important decisions. Data mining algorithms such as J48, NB Tree and Simple CART are used to predict the heart disease. The objective of this research work is to predict heart disease more accurately with reduced number of attributes.

Keywords: Data Mining, Heart disease, J48, NB Tree, Simple CART

I. INTRODUCTION

Data mining is the extraction of useful information from the large database. The extracted information from the dataset is transformed it into an understandable structure. Classification is the technique used in data mining used to classify the item according to the features of the items in the dataset.

Medical data mining has great potential for exploring the hidden patterns in the datasets of the medical domain. Quality of service is a major problem and challenge faced by healthcare industry. This leads to the development in healthcare for accurate diagnosis and prediction of disease at the earlier stage.

Heart disease is the blockage of the arteries and vessels that provide oxygen and nutrient-rich blood to the heart. Cardiovascular disease is a class of disease that involves the heart or the vessels. It occurs when the coronary arteries become partially blocked. Cholesterols and other fatty substances gathered on the inner wall of the arteries. Major risk factors such as high blood pressure, smoking, diabetes, lack of exercise, obesity, high blood cholesterol, poor diet and excessive alcohol.

II. LITERATURE SURVEY

Jayashri S.Sonawane, Dharmaraj R.Patil and Vishal S.Thakare [5] have proposed "Survey on Decision Support System for Heart disease". In this paper Decision Support System is used to diagnose the heart disease at earlier stage. The neural network is used in Decision Support System to detect the heart disease.

"Survey on data mining techniques for the diagnosis of diseases in medical domain" was proposed by Parvathi I and Siddharth Rautaray [8]. The proposal of hybrid data mining model is used to extract classification knowledgein clinical decision system. Different types of Algorithm and various tools are used to diagnose the disease. As the result they conclude that single algorithm may not be accurate for weakly classified datasets.

"A Survey paper of data mining in medical diagnosis" was proposed by Sona Baby and Ariya T.K [3]. Here they compared various data mining techniques to get accurate result.

"A Survey of data mining techniques on medical data for finding temporarily frequent diseases" was proposed by Mohammed Abdul Khaleel, Sateesh Kumar Pradhan, G.N Dash and F.A Mazarbhuiya [1]. Temporal data mining on medical datasets is used to predict the disease.

"A Survey on data mining approaches for healthcare" was proposed by Divya Tomar and Sonali Agarwal [13]. The advantages and disadvantages of the data mining techniques are explained here to find out which is the best technique to diagnose the disease.

"Prediction of Heart disease using classification algorithms" was proposed by Hlaudi Daniel Masethe and Mosima Anna Masethe [14]. In this paper different algorithms were used to predict the heart disease at the earlier stage. The algorithms were compared to see which algorithm gives more accuracy to predict the heart disease.

N.Aditya Sundar et al proposed a work of "Performance analysis of classification data mining techniques over heart disease database" [15]. This extracts hidden knowledge from a historical heart disease database. Classification Matrix methods are used to evaluate the effectiveness of models.

Tina R.Patil and Mrs S.S.Sherekar proposed a research of "Performance analysis of Naive Bayes and J48 Classification algorithm for data classification" [1]. This paper compares the

ISSN: 2320-7868 (Online)

performance of Naive Bayes and J48 and finally concludes that J48 gives better performance than Naive Bayes.

The datasets used in the above papers are collection of heart patients record collected from various medical and healthcare organizations.

III. DATASET DESCRIPTIONS

The data are collected from real time database from UCI repository. The objective of this dataset is to predict the heart disease based on the given attributes. The dataset consists of 14 attributes that are used to predict the heart disease. The detail description of the attributes are given as,

S.No	ATTRIBUTES	DESCRIPTION
1	Age	Age of the patient in years
2	Gender	Gender of the patient (1=male; 0=female)
3	cp	Defines the type of chest pain 1.Typical Angina 2.Atypical Angina 3.Non-Anginal Pain 4.Asymptomatic
4	trestbps	Resting blood pressure (in mm Hg on admission to the hospital)
5	chol	Serum cholesterol in mg/dl
6	fbs	Fasting blood sugar>120 mg/dl (1=true; 0=false)
7	restecg	Resting electrocardiographic results (0=normal; 1=having ST-T; 2=hypertrophy)
8	thalach	Maximum heart rate achieved
9	exang	Exercise induced angina (1=yes; 0=no)
10	oldpeak	ST depression induced by exercise relative to rest
11	slope	The slope of the peak exercise ST segment (1=up sloping; 2=flat; 3=down sloping)
12	Ca	Number of major vessels (0-3) colored by fluoroscopy
13	Thal	3=normal; 6=fixed defect; 7= reversible defect
14	Num	The predicted attribute – diagnosis of heart disease (angiographic disease status) (Value 0=<50 % diameter narrowing ; value 1=>50% diameter narrowing

The attributes are given based on data types. The dataset is based on the numeric and nominal data type.

IV. PROPOSED MODEL

In the proposed method mainly decision tree is used for predicting the cardiovascular heart disease from the given dataset instances. Here the framework can be given as below,

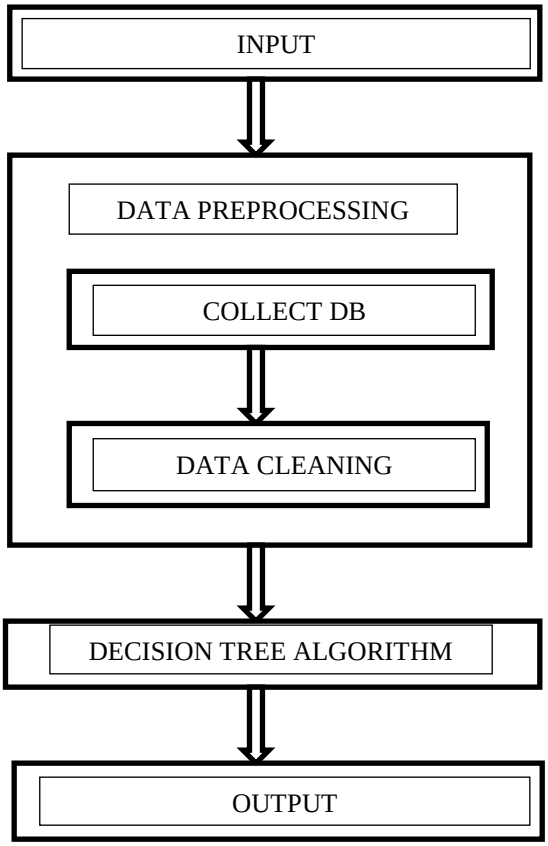

Figure 4.1: Proposed Framework

In the proposed model three different types of decision tree algorithms such as J48, NB Tree and simple CART are applied on cardiovascular heart disease dataset in the WEKA tool and the performance is calculated.

4.1 J48 Decision Tree

J48 is an open source java implementation of the C4.5 algorithm in the weka data mining tool. Information gain is used in C4.5 to collect the information in data sets. And this information is used to take the decisions. It is the mathematical tool that algorithm J48 has used to decide, in each tree node, which variable fits better in terms of target variable prediction.

4.2 Naive Bayes Classifier

NB classifier is derived from Bayesian theorem. It is a simple technique for constructing classifiers. The common principle of naive bayes algorithm is all naive bayes classifiers assume that the value of a particular feature is independent of

ISSN: 2320-7868 (Online)

the value of any other feature. It can be trained very efficiently in a supervised learning.

This system will extract hidden knowledge from a heart disease database. Minimum error rate will be produced by Bayesian classifier. It provides creative ways of exploring and understanding data. It is also used to create models with predictive capabilities. It is one of the effective algorithms to predict the heart disease. Continuous data set is used rather than the categorical data to predict the heart disease easily.

4.3 Simple CART

Classification and Regression Trees (CART). It is used to display important data relationships that could remain hidden using other analytical tools very quickly. It supports high speed deployment.

Classification trees are where the target variable is categorical and the tree is used to identify the "class" within which a target variable would likely fall into. Regressions are where the target variable is continuous and tree is used to predict its value.

The rules for working with CART are: Split the data at a node based on the value, Stop the rules when the branch is terminated, a prediction for the target variable.

The advantages such as: It is nonparametric. It is not significantly impacted by outliers. It incorporates both testing with a test data set and cross-validation. It can use the same variable more than once in different parts of the tree. It can be used in conjunction with other prediction methods. These three algorithms are applied on the given data set and the performance is given in the experimental results.

V. EXPERIMENTAL RESULTS

The given decision tree algorithms like J48, NB Tree and Simple CART are applied on the cardiovascular heart disease data set in WEKA and the performance of the algorithm are given based on various factors. The performance can be obtained based on the time taken to build the decision tree and correctly classified instances to build the decision tree and correctly classified instances.

Name of the Algorithm	Time taken to build the decision tree
J48	0.08 seconds
NB Tree	1.03 seconds
Simple CART	0.13 seconds

Table 5.1: Time taken by the algorithms to build the decision tree

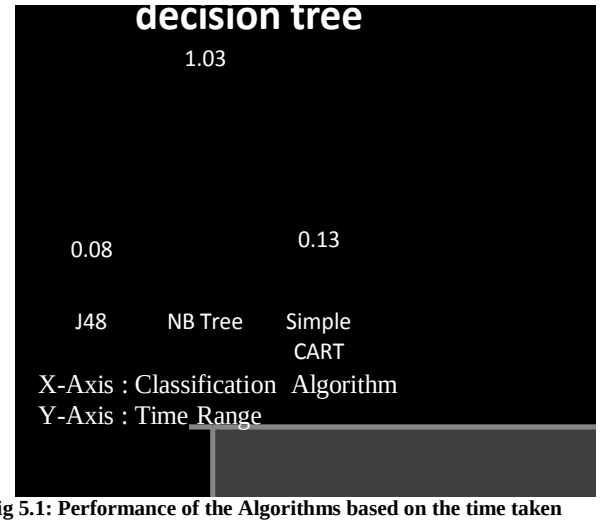

Fig 5.1: Performance of the Algorithms based on the time taken

The data set consists of 10 instances and they are applied as a test case in the classification algorithm. The performance of the algorithms can be known from the instances that are correctly classified. The instances which are correctly classified and the accuracy using the WEKA tool can be given as below,

Name of the Algorithm	Accuracy
J48	91.4 %
NB Tree	88.5 %
Simple CART	92.2 %

Table 5.2: Number of instances correctly classified and their accuracy

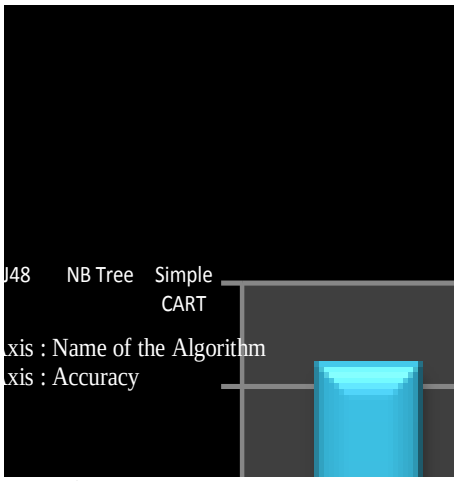

Fig 5.2: Performance of the algorithm based on the instances

A Confusion Matrix is a useful visualization tool for analyzing the classifier accuracy. Structure of the confusion matrix can be given as follows:

 ISSN: 2320-7868 (Online)

TP	TN
FP	FN

Table 6.1: Structure of the confusion matrix

Where,

- ✓ **TP** is True Positive: Heart patients correctly diagnosed as cardiovascular heart disease.
- ✓ **FP** is False Positive: Healthy people incorrectly identified as heart disease.
- ✓ **TN** is True Negative: Healthy people correctly identified as healthy.
- ✓ **FN** is False Negative: Heart patients incorrectly identified as healthy.

The Confusion Matrix for the decision tree classification algorithms such as J48, NB Tree and Simple CART can be given as follows based on the execution of the algorithm using WEKA tool.

TP:141	TN:9
FP:14	FN:106

Table 6.2: Confusion Matrix for J48

TP:136	TN:14
FP:17	FN:103

Table 6.3: Confusion Matrix for NB Tree

TP:138	TN:12
FP:9	FN:111

Table 6.4: Confusion Matrix for CART

From the above confusion matrix we can conclude that J48 decision tree algorithm gives better performance accuracy in the given data set.

VI. CONCLUSION

Medical Industry is a broad area that cannot handle manually. Data Mining plays a major role in mining the useful information from the large medical database. Data preprocessing is used to improve the quality of the data. This model is developed based on the real time data set. The performance and accuracy of the algorithms are calculated using WEKA tool.

The experiment has been successfully performed with several data mining decision tree mechanism and it is found that the J48 algorithm gives a better performance over the supplied data set with the accuracy of 92.2 %. Data Mining can significantly helpful in heart disease research and may improve the quality of healthcare of heart patients. It can also be implemented using several classification techniques.

REFERENCES

[1]. "Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification" Tina R. Patil and Mrs. S.S. Sherekar , International journal of computer science and Applications, Volume 6, No. 2, April 2013.

[2]. S.J Gnanasoundhari et al, "A Survey on heart disease prediction system using Data mining techniques", International journal of computer science and mobile applications, Vol.2 Issue 2, February 2014.

[3].Sona Baby and Ariya T.K "A Survey paper of data mining in medical diagnosis" published by IJRCCT in 2014.

[4]. Beant Kaur and Williamjeet Singh "Review on Heart disease Prediction System Using Data mining techniques", International journal on recent and innovation trends in computing and communication, volume: 2, Issue: 10.

[5].Jayshri S.Sonawane et al, "Survey on Decision Support System for Heart disease", International journal of advancements in technology, ISSN 0976-4860.

[6].RajKumar and Dr.Rajesh Verma "Classification Algorithms for data mining: A survey", International journal of innovations in engineering and technology, Vol.1 Issue 2 August 2012.

[7]. Mohammed Abdul Khaleel et al, "A Survey of Data mining techniques on medical data for finding temporarily frequent diseases", International journal of advanced research in computer and communication engineering ,Vol.2, Issue 12, December 2013.

[8]. Parvathi I and Siddharth Rautaray , "Survey on data mining techniques for the diagnosis of diseases in medical domain", (IJCSIT) vol.5(1),2014.

[9]. M. Kumari and S. Godara, "Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction", IJCST ISSN: 2229- 4333, vol. 2, no. 2, (2011) June.

[10]. J. Soni, U. Ansari, D. Sharma and S. Soni, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", (2011).

[11]. Shamsher Bahadur Patel et al, "Predict the diagnosis of heart disease patients using classification mining techniques", IOSR Journal of Agriculture and Veterinary Science Volume 4, Issue 2, 2013.

[12] .Xindong Wu and Vipin Kumar "The top ten algorithms in data mining", 2009 by Taylor & Francis Group,LLC.

[13]. Divya Tomar and Sonali Agarwal proposed "A survey on Data mining approaches for healthcare", International journal of bio-science and bio – technology, Vol.5, 2013.

[14]. Hlaudi Daniel Masethe and Mosima Anna Masethe proposed "Prediction of heart disease using classification algorithms" , Proceedings of the world congress on engineering and computer science, Vol II, October 2014, San Francisco, USA.

[15]. N. Aditya Sundar, P.Pushpa Latha and M.Rama Chandra, " Performance analysis of classification data mining techniques over heart disease database", International Journal of Engineering Science and Advanced Technology, Vol 2, Issue 3, 2014.

[16]. Milan Kumari and Sunila Godara , " Comparative study of data mining classification methods in cardiovascular disease prediction", IJCST, Vol. 2, Issue 2, June 2011.

[17]. Trilok Chand Sharma and Manoj Jain, " WEKA Approach for comparative study of Classification Algorithm", International Journal of Advanced Research in Computer and Communication Engineering, Vol 2, Issue 4, April 2013.