ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

Study of Heart Disease Prediction Using Data Mining

Geetika Sunil Agrawal, Prof. Dinesh D. Patil

Department of Computer Science and Engineering, Shrisantgadge Baba College of Engineering and Technology, Maharashtra, 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: classification Techniques, Decision Tree Algorithm, heart disease, KNN, Naïve Bayes, Neural Network, Risk level.

I.INTRODUCTION

Heart diseases are the number one cause of death globally:more people die annually from Heart diseases than from anyother cause. An estimated 17.3 million people died from Heartdiseases in 2008, representing 30% of all global deaths. Ofthese deaths, an estimated 7.3 million were due to coronaryheart disease and 6.2 million were due to stroke [1]. Recentresearch in the field of medicine has been able to identify riskfactors that may contribute toward the development of heartdisease but more research is needed to use this knowledge inreducing the occurrence of heart diseases. Diabetes,hypertension, and high blood cholesterol have beenestablished as the major risk factors of heart diseases. Lifestyle risk factors which include eating habits, physicalinactivity, smoking, alcohol intake, obesity are also associatedwith the major heart disease risk factors and heart disease[2,3].There are studies showing that reducing these riskfactors for heart disease can actually help in preventing heartdiseases [4]. There are many studies and researches on theprevention of heart disease risk. Data from studies ofpopulation has helped in prediction of heart diseases, based onblood pressure, smoking habit, cholesterol and blood pressurelevels, diabetes. Researchers have used these predictionalgorithms in adapted form of simplified score sheets thatallow patients to calculate the risk of heart diseases [6]. TheFramingham Risk Score (FRS) is a popular risk predictioncriterion which is used in algorithms for heart diseaseprediction [7].This study aimed at developing an intelligent data miningsystem based on genetic algorithm optimized neural networksfor the prediction of heart disease based on risk factors'categories. The system was implemented using MATLABR2012a.

II.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 .[1]. 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

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

technique was proposed by HeonGyu Lee et al. [3].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. [2]. 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 [5]. The resultant dataset contains records of patients having heart disease. Three constraints were introduced to decrease the number of patterns [4].

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. [7] builds a decision tree with database of patient for a medical problem. Data mining techniques are used to explore, analyze and extract medical data using complex algorithms in order to discover unknown patterns. Researchers are using data mining techniques for the diagnosis of many diseases such as heart disease [6], diabetes [7], stroke [8] and cancer and many data mining techniques have been used in the diagnosis of heart disease with good accuracy. Researchers have been applying different data mining techniques such as naïve bayes, neural network, decision tree, bagging, kernel density, and support vector machine for prediction and diagnosis of heart diseases

III. 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 :

A.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

- 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

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

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

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

V.FLOWCHART

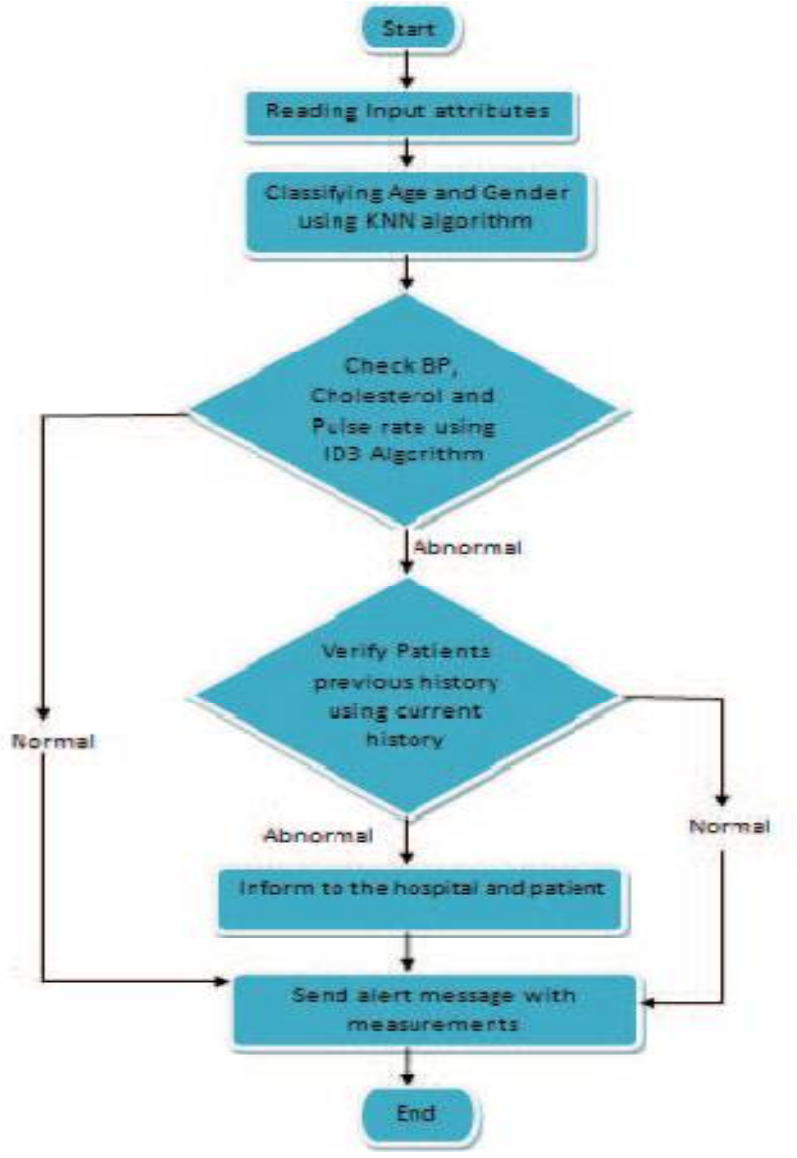

Figure1: Flowchart of the risk level prediction system

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

VI. RESULT AND DISCUSSION

Figure.1 shown below plots the range of minimum and maximum systolic, risk levels of each class, modeled.

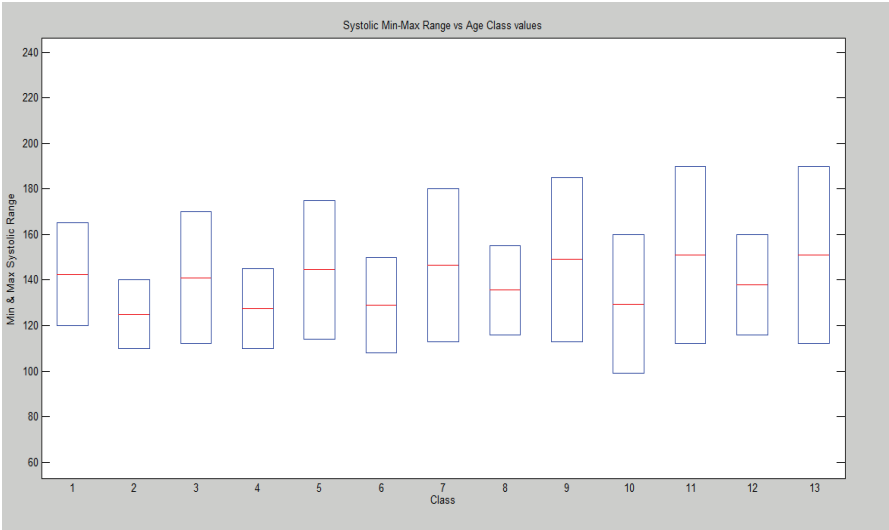

Figure2: various risk levels of systolic for input attribute.

The age range was grouped by the K-Nearest Neighbor algorithm into class. The Risk level of each class was identified with the help of ID3 algorithm. In this Model each class has maximum, minimum and average risk values plotted. These risk values were obtained from the test data compared with the model generated for the systolic input factor.

Figure.3 shown below plots the range of minimum and maximum diastolic, risk levels of each class, modeled. The age range was grouped by the K-Nearest Neighbor algorithm into class. The Risk level of each class was identified with the help of ID3 algorithm. In this Model each class has maximum, minimum and average risk values plotted. These risk values were obtained from the test data compared with the model generated for the diastolic input factor.

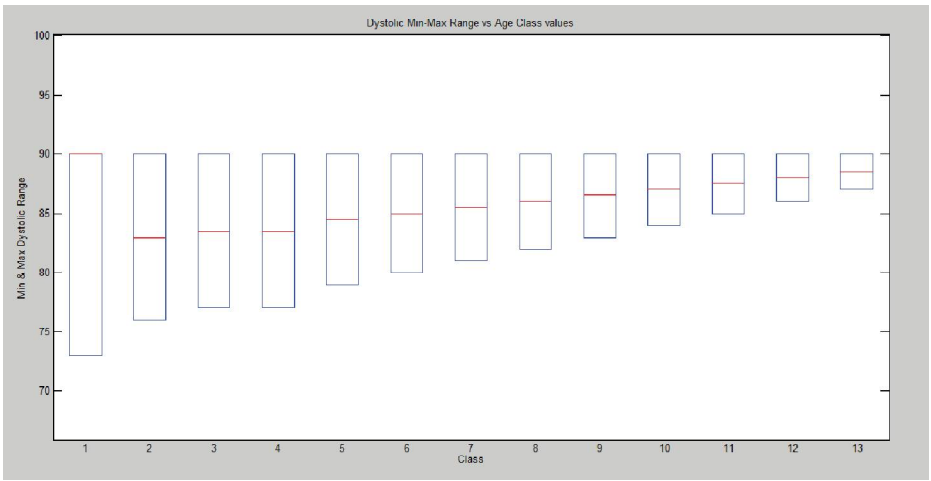

Figure3: various risk levels of diastolic for input attributes

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

Figure.4 shown below plots the range of minimum and maximum pulse rates, risk levels of each class, modeled. The age range was grouped by the K-Nearest Neighbor algorithm into class. The Risk level of each class was identified with the help of ID3 algorithm. In this Model each class has maximum, minimum and average risk values plotted. These risk values were obtained from the test data compared with the model generated for the pulse rate input factor.

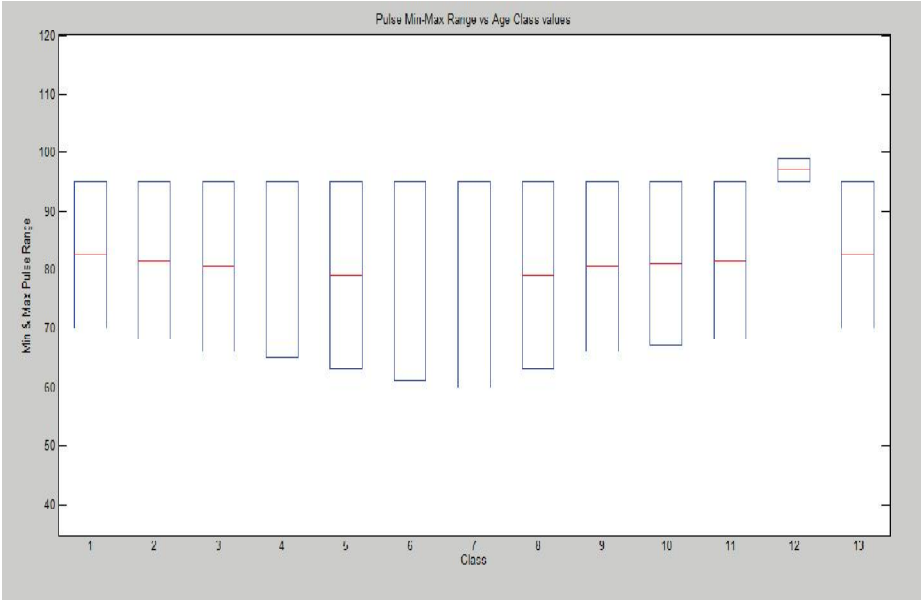

Figure4: various risk levels of pulse rate for input attributes.

VII. CONCLUSION

Data mining techniques and methods applied in patient medical dataset has resulted in innovations, standards and decision support system that have significant success in improving the health of patients and the overall quality of medical services. But we still need systems which could predict heart diseases in early stages.In this study, a new hybrid model of Neural Networks and Genetic Algorithm to optimize the connection weights of ANN so as to improve the performance of the Artificial Neural Network. The system uses identified important risk factors for the prediction of heart disease and it does not require costly medical tests.

REFERENCES

[1] 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
[2] 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).
[3] 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.
[4] 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.
[5] Carlos Ordonez, "Improving Heart Disease Prediction Using Constrained Association Rules," Seminar Presentation at University of Tokyo, 2004.
[6] R. Das, I. Turkoglu, and A. Sengur, Effective diagnosis of heart disease through neural networks ensembles, Expert Systems with Applications, Elsevier, pp. 7675–7680, 2009.
[7] T. Porter and B. Green, "Identifying Diabetic Patients: A Data Mining Approach," Americas Conference on Information Systems, 2009.
[8] S. Panzarasa et al, "Data mining techniques for analyzing stroke care processes," in Proc. of the 13th World Congress on Medical Informatics, 2010