Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques

Chaitrali S. Dangare
Student, M.E. (CSE)
Walchand Institute of Technology
Solapur, Maharashtra, India

Sulabha S. Apte, PhD.
Professor, Dept. CSE
Walchand Institute of Technology
Solapur, Maharashtra, India

ABSTRACT

The Healthcare industry is generally “information rich”, but unfortunately not all the data are mined which is required for discovering hidden patterns & effective decision making. Advanced data mining techniques are used to discover knowledge in database and for medical research, particularly in Heart disease prediction. This paper has analysed prediction systems for Heart disease using more number of input attributes. The system uses medical terms such as sex, blood pressure, cholesterol like 13 attributes to predict the likelihood of patient getting a Heart disease. Until now, 13 attributes are used for prediction. This research paper added two more attributes i.e. obesity and smoking. The data mining classification techniques, namely Decision Trees, Naive Bayes, and Neural Networks are analyzed on Heart disease database. The performance of these techniques is compared, based on accuracy. As per our results accuracy of Neural Networks, Decision Trees, and Naive Bayes are 100%, 99.62%, and 90.74% respectively. Our analysis shows that out of these three classification models Neural Networks predicts Heart disease with highest accuracy.

Keywords

Data Mining, Heart Disease, Neural Networks, Decision Trees, Naive Bayes.

1. 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.

According to survey of WHO, 17 million total global deaths are due to heart attacks and strokes. The deaths due to heart disease in many countries occur due to work overload, mental stress and many other problems. On the whole it is found as primary reason behind death in adults. Diagnosis is complicated and important task that needs to be executed accurately and efficiently. The diagnosis is often made, based on doctor’s experience & knowledge. This leads to unwanted results & excessive medical costs of treatments provided to patients.

Therefore an automatic medical diagnosis system is designed that take advantage of collected data base and decision support system. This system can help in diagnosing disease with less medical tests & effective treatments.

2. HEART DISEASE

The heart is important organ of human body part. 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.

A number of factors have been shown that increases the risk of Heart disease [2]:

- Family history
- Smoking
- Poor diet
- High blood pressure
- High blood cholesterol
- Obesity
- Physical inactivity
- Hyper tension

Factors like these are used to analyze the Heart disease. In many cases, diagnosis is generally based on patient’s current test results & doctor’s experience. Thus the diagnosis is a complex task that requires much experience & high skill.

3. 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 Sellappan Palaniappan 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 Heon Gyu 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.
- Latha Parthiban 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.

4. PROPOSED PREDICTION SYSTEM

Today, many hospitals manage healthcare data using healthcare information system; as the system contains huge amount of data, used to extract hidden information for making intelligent medical diagnosis. The main objective of this research is to build Intelligent Heart Disease Prediction System that gives diagnosis of heart disease using historical heart database. To develop this 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 are used, as these attributes are considered as important attributes for heart disease. The data mining classification techniques viz. Neural Networks, Decision Trees, and Naive Bayes are used.

5. DATA SOURCE

The publicly available heart disease database is used. The Cleveland Heart Disease database [11] consists of 303 records & Statlog Heart Disease database consists of 270 records [12].

The data set consists of 3 types of attributes: Input, Key & Predictable attribute which are listed below.

5.1. Input attributes

Table 1. Description of 13 input attributes

Sr. no	Attribute	Description	Values
1	age	Age in years	Continuous
2	sex	Male or female	1 = male 0 = female
3	cp	Chest pain type	1 = typical type 1 2 = typical type agina 3 = non-agina pain 4 = asymptomatic
4	thestbps	Resting blood pressure	Continuous value in mm hg
5	chol	Serum cholesterol	Continuous value in mm/dl
6	Restecg	Resting electrographic results	0 = normal 1 = having_ST_T wave abnormal 2 = left ventricular hypertrophy
7	fbs	Fasting blood sugar	1 ≥ 120 mg/dl 0 ≤ 120 mg/dl
8	thalach	Maximum heart rate achieved	Continuous value
9	exang	Exercise induced agina	0= no 1 = yes
10	oldpeak	ST depression induced by exercise relative to rest	Continuous value
11	solpe	Slope of the peak exercise ST segment	1 = unsloping 2 = flat 3 = downsloping
12	ca	Number of major vessels colored by floursopy	0-3 value
13	thal	Defect type	3 = normal 6 = fixed 7 = reversible defect

All the research papers referred above have used 13 input attributes for prediction of Heart disease. To get more appropriate results two important attributes i.e. obesity and smoking are added to input attributes.

Table 2. Description of newly added attributes

Sr. no	Attribute	Description	Values
14	obes	obesity	1 = yes 0 = no
15	smoke	smoking	1= past 2 = current 3 = never

5.2. Key attribute

PatientID: Patient's Identification Number

5.3. Predictable attribute

Diagnosis: Value 1 = < 50 % (no heart disease)

Value 0 = > 50 % (has heart disease)

6. 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.

6.1. 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. The structure of MLPNN is as shown in Figure 1.

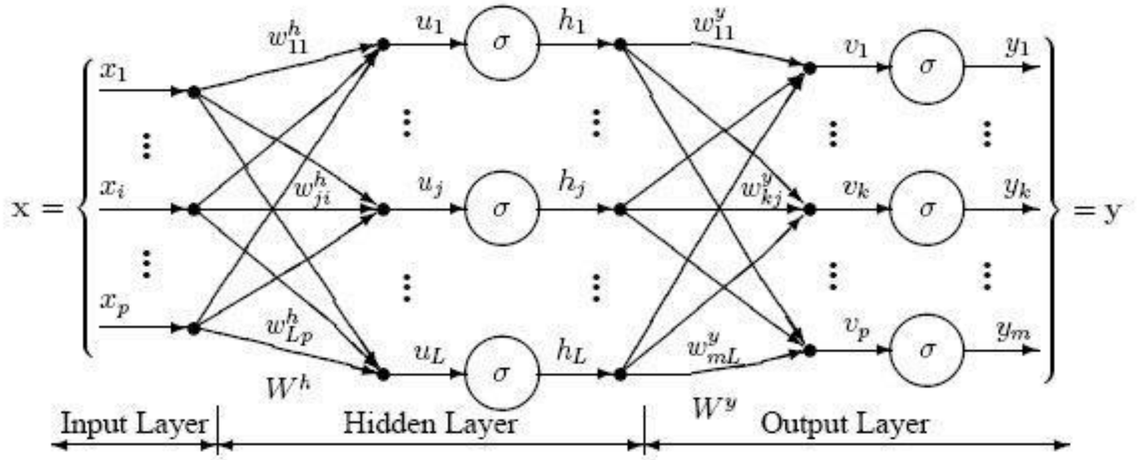

Figure 1: Structure of Multi Layer Perceptron Neural Network

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 x_i into neurons in hidden layer. Neuron of hidden layer adds input signal x_i with weights w_{ji} of respective connections from input layer. The output Y_j is function of

$$Y_j = f (\sum w_{ji} x_i)$$

Where f is a simple threshold function such as sigmoid or hyperbolic tangent function.

6.2. 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 overfitting 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.

6.3. 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={x_1, x_2, ..., x_n} 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

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

7. RESULTS

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. The data mining tool Weka 3.6.6 is used for experiment.

Initially dataset contained some fields, in which some value in the records was missing. These were identified and replaced with most appropriate values using ReplaceMissingValues filter from Weka 3.6.6. The ReplaceMissingValues filter scans all records & replaces missing values with mean mode method. This process is known as Data PreProcessing. After pre-processing the data, data mining classification techniques such as Neural Networks, Decision Trees, & Naive Bayes were applied.

A confusion matrix is obtained to calculate the accuracy of classification. A confusion matrix shows how many instances have been assigned to each class. In our experiment we have two classes, and therefore we have a 2x2 confusion matrix.

Class a = YES (has heart disease)

Class b = NO (no heart disease)

Table 3. A confusion matrix

	a (has heart disease)	**b (no heart disease)**
a (has heart disease)	TP	FN
b (no heart disease)	FP	TN

TP (True Positive): It denotes the number of records classified as true while they were actually true.

FN (False Negative): It denotes the number of records classified as false while they were actually true.

FP (False Positive): It denotes the number of records classified as true while they were actually false.

TN (True Negative): It denotes the number of records classified as false while they were actually false.

Results obtained with 13 attributes are specified below

Table 4. Confusion matrix obtained for three classification methods with 13 attributes

Confusion matrix for Naive Bayes:

	a	**b**
a	110	5
b	10	145

Confusion matrix for Decision Trees:

	a	**b**
a	123	4
b	5	138

Confusion matrix for Neural Networks:

	a	**b**
a	117	0
b	2	151

Results obtained by adding two more attributes i.e. total 15 attributes are specified below.

Table 5. Confusion matrix obtained for three classification methods with 15 attributes

Confusion matrix for Naive Bayes:

	a	**b**
a	100	7
b	18	145

Confusion matrix for Decision Trees:

	a	**b**
a	85	0
b	1	184

Confusion matrix for Neural Networks:

	a	**b**
a	106	0
b	0	164

Table 3 shows accuracy for different classification methods with 13 input attributes & 15 input attributes values.

Table 6. Comparison of data mining techniques

Classification Techniques	**Accuracy with**	
	13 attributes	**15 attributes**
Naive Bayes	94.44	90.74
Decision Trees	96.66	99.62
Neural Networks	99.25	100

The accuracy is of each method is plotted on a graph as below:

Where blue line represents accuracy for 15 attribute dataset & brown line represents accuracy for 13 attribute dataset.

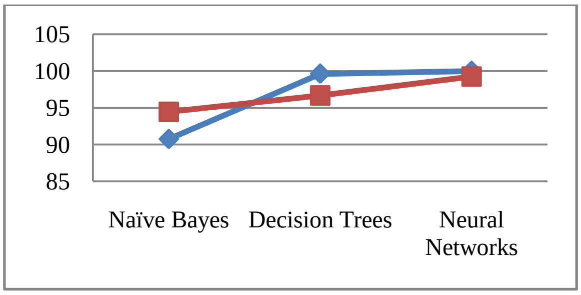

Figure 2: Graphical representation of accuracy for each method.

8. CONCLUSION

The overall objective of our work is to predict more accurately the presence of heart disease. In this paper, two more input attributes obesity and smoking are used to get more accurate results. Three data mining classification techniques were applied namely Decision trees, Naive Bayes & Neural Networks. From results it has been seen that Neural Networks provides accurate results as compare to Decision trees & Naive Bayes.

This system can be further expanded. It can use more number of input attributes listed above in table 1 and 2. Other data mining techniques can also be used for predication e.g. Clustering, Time series, Association rules. The text mining can be used to mine huge amount of unstructured data available in healthcare industry database.

9. 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] Sellappan Palaniappan, Rafiah Awang, "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, Navin Rajpal, "Decision Support System for Heart Disease Diagnosis Using Neural Network", Delhi Business Review, Vol. 8, No. 1 (January - June 2007).

[5] Heon Gyu Lee, Ki Yong Noh, Keun Ho Ryu, “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] Shantakumar B.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, Heon Gyu Lee, Ho-Sun Shon, Bum Ju Lee, and Keun Ho Ryu, "Associative Classification Approach for Diagnosing Cardiovascular Disease", Springer, Vol:345, pp: 721- 727, 2006.

[9] Franck Le Duff, Cristian Munteanb, 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] Latha Parthiban 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.

[11] Cleveland database: http://archive.ics.uci.edu/ml/datasets/Heart+Disease

[12] Statlog database: http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/heart/

[13] Dr. Yashpal Singh, Alok Singh chauhan “Neural Networks in data mining” Journal of Theoretical and Applied Information Technology , 2005 - 2009 JATIT.