International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Comparative Study on Heart Disease Prediction System Using Data Mining Techniques

T. Revathi[1], S. Jeevitha[2]

[1]Associate Professor, Dept of CS, PSG college of Arts & Science, Coimbatore, Tamilnadu, India

[2]Research scholar, Dept of CS, PSG college of Arts & Science, Coimbatore, Tamilnadu, India

Abstract: ***Healthcare industry has huge amount of data that contains hidden information. This information supports decision making process on related area. In this research paper, we discussed various approaches of data mining which are useful in predicting the heart disease. One of the complex tasks in healthcare industry is predicting of heart disease and it requires more experience and knowledge. Some of the ways of predicting heart diseases are ECG, stress test and heart MRI etc. Here the system uses 14 parameters for predicting heart disease that include blood pressure, cholesterol, chest pain and heart rate. These parameters are used to improve an accuracy level. The main aim of this paper is to provide an analysis of data mining techniques on diagnosing heart disease.***

Keywords: Heart disease, Neural Network, Naïve bayes and Decision tree.

1. Introduction

Heart is an important part of human body. Life depends on an efficient working of heart. If working of heart is not good then it will affects the other parts of our human body like kidney and brain. Heart disease is predicted based on the performance of heart. Some of the factors that are used to predict heart diseases are:

- Cholesterol
- High blood pressure
- Lack of physical exercise
- Smoking
- Obesity
- Family history of heart disease

Heart disease is the major cause of human deaths. Predictions should be taken to reduce the risk of heart disease. Generally, doctors will diagnosis heart disease based on the symptoms and physical examination of the patient body. Heart disease prediction is a difficult task in healthcare industry. Nowadays, healthcare industry contains huge amount of data of patients, disease diagnosis, electronic patient records and medical devices. It is a key resource that needs to be processed while knowledge extraction and it will support decision making process.

Figure 1, shows the difficulties that will happen while diagnosing that leads to the negative presumptions and unpredictable effects.

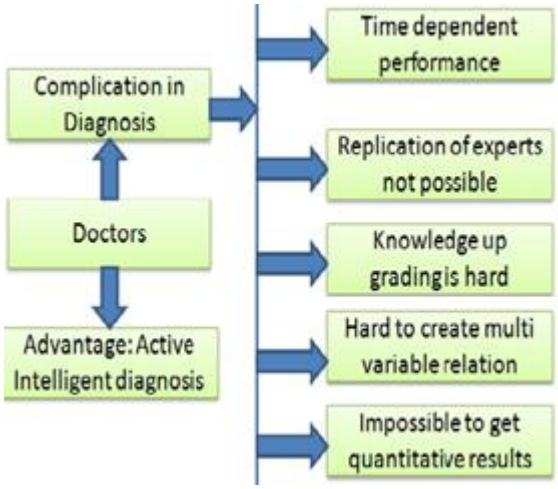

Figure 1: Complexity in Diagnosis of Heart Diseases

The main aim of this paper is to create a prototype for heart disease prediction system by using neural networks, naïve bayes and decision tree techniques of data mining. So that it provides an effective treatment. And it improves visualization and ease of interpretation.

2. Data Mining Approaches

Data mining is used to analyze data from different dimensions and summarizes the data into useful information. Data mining software allows the users to analyze the data from different dimensions and categorizes it. Normally data mining is the process of identifying interesting patterns on various fields in large relational databases. In this study, we discussed the data mining techniques namely neural networks, naïve bayes and decision tree algorithms which are used to predict heart disease.

2.1. Back Propagation Network

A neural network is a technique which consist more number of processing elements (node) and each processing elements are interconnected by unidirectional signals (connections). Each and every processing element calculates a weighted sum of its input signals and it computes an output signal that

is passed to other processing elements. During the training phase of neural network, the weights are adjusted to provide the target output.

The most accepted neural network algorithm is back-propagation algorithm. This algorithm mainly focuses on the feed forward multi-layer networks. And it is most admired, effective algorithm that is used to learn the complex model in an easy manner. One of the advantages of back propagation network is non-linear answers to non-specific problems. Generally, the back-propagation network consists three layers namely input layer, hidden layer and an output layer. The hidden layers can be added based on the problem. At most the problem can use five layers that is one for input layer, one for output layer and three for hidden layer to solve any complex problems.

Back-propagation algorithm performs two processes called forward pass and backward pass. The layers of back-propagation network are connected to the next successive layer. The inputs are given to the input layer. Based on the inputs, the output is computed and it is matched with the desired output. This process is known as forward pass. During the backward pass, the weights are adjusted based on the error correction rule. The real output of the network is take away from the desired output to provide an error signal. This signal is given backward to all layers of the network. So that the weights are adjusted to proceed the actual output of the network that goes closer to the desired output in a statistical manner.
Process of back-propagation network is described in steps as follows:
Step 1: Grant training data to the network.
Step 2: Evaluate the actual output and desired output.
Step 3: Compute the error in each layer.
Step 4: Estimate what output should be for each layer and how much lower or higher output should be adjusted for desired output.
Step 5: After that adjust the weights.

The clinical data related to heart disease are derived from Cleveland database which is publicly available dataset on internet. This dataset contains 76 attributes. But most important attributes are defined in above table 1. Here, the goal attribute is "num" which refers to the prediction of heart disease. Value of "num" attribute is 0, it represents the patient is not affected by heart disease. If it is 1, then it refers to the presence of heart disease. Table 1 show the attributes and corresponding classes are shown in table 2.

Table 1: Input attributes

Attribute	*Description*	*Values*
Age	Age in years	Continuous
Sex	Male or female	1=male 0=female
Cp	Chest pain type	1=typical type 1 2=typical type angina 3=non-angina pain 4=asymptomatic
thestbps	Resting blood pressure	Continuous value in mm hg
Chol	Serum cholesterol	Continuous value in mm/dl
restecg	Resting electrographic results	0=normal 1=having ST_T wave abnormal 2=left ventricular hypertrophy
Fbs	Fasting blood pressure	1>= 120 mg/dl 0<=120 mg/dl
thalach	Maximum heart rate achieved	Continuous value
exang	Exercise induced angina	0=no 1=yes
oldpeak	ST depression induced by exercise relative to rest	Continuous value
slope	Slope of the peak exercise ST segment	1=unsloping 2=flat 3=downsloping
ca	Number of major vessels colored by floursopy	0-3 value
thal	Defect type	3=normal 6=fixed 7=reversible defect
num	Diagnosis of heart disease	1= >50% diameter narrowing 0= <50% diameter narrowing

Table 2: Dataset classes

Class	**Description**
Class 0	Normal Person.
Class 1	First Stroke
Class 2	Second Stroke
Class 3	End of Life

The back- propagation network algorithm can be assessed by confusion matrix which is shown in table 3. Confusion matrix is a detailed layout which represents the performance of an algorithm. Each row of this matrix represents the predicted class instances while each column of the matrix represents the actual class instances. This matrix is used to show the correct and incorrect instances.

Table 3: Confusion Matrix

Predicted class	Actual class		
		Positive	Negative
	True	True Positive	True Negative
	False	False Positive	False Negative

True Positive: This instance indicates the number of records categorizes as true even as they were really true.
False Negative: It indicates the number of records categorizes as false even as they were really false.
False Positive: This instance shows the number of records categorizes as false even as they were really true.
True Negative: It specifies the number of records categorizes as true even as they were really false.
Back-propagation network provides better results and it helps the doctors to plan for better prediction and the system predicts heart disease patients accurately.

2.2. Naïve Bayes Algorithm

The Bayesian classification is a statistical technique which follows supervised learning method. It determines the probabilities of the outputs. It solves the predictive problems. It provides a model for understanding many learning algorithms. And this algorithm computes the probabilities for the problems.

Bayes theorem relates the conditional and marginal probabilities of two random events. It applies a simple probabilistic classification. And naïve bayes classifier thinks that the presence or absence of a particular feature of a class is not related to the presence or absence of other feature. For instance, an apple can be judged based on its color, shape and about 4" diameter. Even if these features rely on the occurrence of other features, a naïve bayes classifier uses all of these properties to independently provide the probability that "it's an apple".

Naïve bayes theorem works well in many intricate real world situations. The independent variables are used for predicting the event. One of the advantages of naïve bayes classifier is that it needs a small amount of training data to evaluate the parameters which is necessary for classification process. Here, the independent variables are only the difference between the variables for each and every class that is required to be determined.

Naïve bayes algorithm is given below:

For example, the training data is X, posterior probability of a hypothesis H then P(H|X) can be calculated by,

$$P(H|X)=P(X|H)P(H)/P(X) \quad (1.1)$$

Algorithm

The naïve bayes algorithm is based on equation (1.1) is given below:

Step 1: Each data sample is characterized by an "n" dimensional feature vector, $X = (x_1, x_2 \ldots x_n)$, representing n measurements made on the sample from n attributes, respectively $A_1, A_2, \ldots, A_n$.

Step 2: Assume that there are "m" classes, $C_1, C_2 \ldots C_m$. Given an unknown data sample X (having no class label), the classifier will envisage that X belongs to the class having the highest posterior probability, conditioned if and only if:

$$P(C_i/X)>P(C_j/X) \text{ for all } 1 <= j <= m \text{ and } j!= i$$

Thus we can maximize $P(C_i|X)$. The class C_i for which $P(C_i|X)$ is maximized which is named by Bayes theorem as the maximum posteriori hypothesis.

Step 3: As P(X) is constant for all classes, only $P(X|C_i)P(C_i)$ have to be maximized. If the class prior probabilities are not well-known, then it is generally assumed that the classes are evenly like, $P(C_1) = P(C_2) = \ldots = P(C_m)$, and we have to maximize $P(X|C_i)$. Else, we can maximize $P(X|C_i)P(C_i)$.

Note that the class prior probabilities can be computed by $P(C_i) = S_i/s$, where S_i is the number of training samples of class C_i, and s is the total number of training samples on X. i.e., the naive probability allocates an unknown sample X to the class C_i.

Heart disease prediction system using naïve Bayesian classification mines unknown knowledge from historical medical database. This system identifies the heart disease patients. And this system can answer difficult queries in an effective manner.

2.3. Decision Tree Algorithm

Decision tree is used to separate a large collection of records into small number of records by using decision rules. Decision tree is same as flowchart in which all non-leaf nodes represent a test on specific attribute and each branch represents an output of that test and the leaf nodes has its class label. The top most labels of a node in the decision tree are known as root node. Decision makers will choose the best option by using the decision tree. Decision trees are used to recognize the various methods of dividing a dataset into smaller segments. And these segments create a reverse of decision tree. Decision tree can be constructed by using various techniques. In this paper, we are using C4.5 algorithm to originate a decision tree. C4.5 algorithm is given below:

Algorithm

Step 1: Verify the base cases.
Step 2: For each and every attribute *a*, discover the normalized information gain ratio from splitting on *a*.
Step 3: Assign *a_best* as the attribute with the maximum normalized information gain.
Step 4: Make a decision node which divides on *a_best.*
Step 5: Return the sublists found by dividing on *a_best* and attach the nodes as child node.

3. Comparison of Data Mining Techniques

We have calculated the accuracy of back-propagation network, naïve bayes and decision tree algorithms. Table 5 specifies the accuracy of these algorithms while predicting the heart disease.

Table 5: Comparison of data mining techniques

Techniques	Accuracy
Back-propagation network	100%
Naïve Bayes	90.74%
Decision tree	99.62%

4. Conclusion

In this paper we have discussed data mining approaches while predicting heart disease. Back-propagation algorithm, Naïve bayes algorithm and decision tree algorithm have been compared and the results of those approaches have been analyzed. By comparing all the techniques of data mining, neural network works well in predicting heart disease and it provides 100% of accuracy. All of these algorithms can answer difficult queries and each of these algorithms has its strength based on the information and accuracy.

References

[1] Andrea D'Souza, *Heart Disease Prediction Using Data Mining Techniques*, IJRES, Vol 3, Num 3, 2015.
[2] Chaitrali S. Dangare**,** Sulabha S. Apte, *A Data Mining Approach for Prediction of Heart Disease using Neural Network*, IJCET, Vol 3, Num 3, 2012.

[3] Dhanashree S. Medhekar, Mayur P. Bote, Shruti D. Deshmukh, *Heart Disease Prediction System using Naïve Bayes*, IJERSTE, Vol 2, Num 3, March.-2013.
[4] Ishtake S.H, Sanap S.A. *Intelligent Heart Disease Prediction System Using Data Mining Techniques*, IJHBR, Vol 1, Num 3, April 2013.
[5] Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni. *Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction*, IJCA, Vol 17, Num 8, 2011.
[6] Manikandan V, Latha S. *Predicting the Analysis of Heart Disease Symptoms using Medicinal Data Mining Method*, IJACTE, Vol 2, Num 2, 2013.
[7] Nabeel Al-Milli, *Back Propagation Neural Network for Prediction of Heart Disease*, JATIT, Vol 56, Num 1, 2013.
[8] Radhimeenakshi S, Nasira G.M. *Prediction of Heart Disease using Neural Network with Back Propagation*, IJCOA, Vol 4, 2015.
[9] Shadab Adam Pattekari and Asma Parveen. *Prediction System for Heart Disease using Naïve Bayes*, IJACMS, Vol 3, Num 3, 2012.
[10] Subbalakshmi G, Ramesh K, Chinna Rao M. *Decision Support in Heart DiseasePrediction System using Naïve Bayes*, IJCSE, Vol 2 Num 2, Apr-May 2011.
[11] Thenmozhi K, Deepika P. *Heart Disease Prediction Using Classification with Different Decision Tree Techniques*, IJERGS, Vol 2, Num 6, October-November, 2014.
[12] Usha Rani K. *Analysis of Heart Disease Dataset using Neural Network Approach*, IJDKP, Vol 1, Num 5, 2011.

Author Profile

T. Revathi , Associate Professor, Dept of CS, PSG college of Arts & Science, Coimbatore, Tamilnadu.

S. Jeevitha , Research scholar, Dept of CS, PSG college of Arts & Science, Coimbatore, Tamilnadu.