Predictions in Heart Disease Using Techniques of Data Mining

Monika Gandhi
Computer Science and Engineering Department
ASET, Amity University, Noida, India
mgandhi9404@gmail.com

Dr. Shailendra Narayan Singh
Computer Science and Engineering Department
ASET, Amity University, Noida, India
snsingh36@amity.edu
sns2033@gmail.com

Abstract: **As huge amount of information is produced in medical associations (healing facilities, therapeutic focuses) yet this information is not properly utilized. The health care system is "data rich" however "knowledge poor ". There is an absence of successful analysis methods to find connections and patterns in health care data. Data mining methods can help as remedy in this circumstance. For this reason, different data mining techniques can be utilized. The paper intends to give details about various techniques of knowledge abstraction by using data mining methods that are being used in today's research for prediction of heart disease. In this paper, data mining methods namely, Naive Bayes, Neural network, Decision tree algorithm are analyzed on medical data sets using algorithms.**

Keywords - Data mining, Heart disease, classification, prediction, Decision tree, Neural network, Naïve Bayes.

I. INTRODUCTION

The headway of information technology, framework coordination and additionally programming advancement, systems have formed an imaginative era of multifaceted computer framework. Information technology specialists have been offered few challenges by these frameworks. A case of such framework is the healthcare services framework. Recently, there has been an augmented attention to make utilization of the headway of data mining advances in healthcare frameworks. Thus, the target of the present effort is to find out the aspects of use of healthcare data for aid of people by method of machine learning furthermore data mining procedures. The main aim is to suggest an automated system for diagnosing heart diseases by taking into account earlier information and data.

978-1-4799-8433-6/15/$31.00©2015 IEEE

A major challenge confronting healthcare associations i.e. hospitals, medicinal focuses are the procurement of quality services at reasonable expenses. Quality services suggest diagnosing patients accurately and overseeing medicines that are more effective. Poor clinical decisions can prompt to poor outcomes which are therefore unsatisfactory. Healthcare organizations can reduce costs by accomplishment of computer based data and/or decision support systems. Healthcare services data is very huge as it incorporates patient records, resource management information and updated information. Human services associations must have capacity to break down information. Treatment records of many patients can be stored away in computerized way; furthermore data mining methods may help in finding out a few vital and basic inquiries related with healthcare organizations.

Clinical choices are frequently made focused around doctors' instinct and experience instead of on the knowledge rich information covered up in the database. This practice prompts undesirable biases, blunders and unnecessary medicinal expenses which influence the quality of services given to the patients. Wu, et al proposed that combination of clinical choice backing with computer based patient records could decrease medical errors, enhance safety of patients, lessening undesirable practice variety, and enhance patient outcome [4]. This suggestion is guaranteeing as the data demonstrating and analysis tools for example data mining, have the possibility to create a knowledge rich environment which can help to essentially enhance the nature of clinical decisions.

Data mining is an important step of KDD i.e. knowledge discovery from database. KDD

comprises of an iterative sequence of data cleaning, data integration, data choice, data mining pattern recognition furthermore data presentation. In particulars, data mining may be accomplished using classification, clustering, prediction, association and time series analysis.

II. HEART DISEASE

Heart is vital part or an organ of the body. Life is subject to proficient working of heart. In the event that operation of heart is not proper, it will influence the other body parts of human, for example, mind, kidney, etc. Heart is simply a pump, which pumps the blood through the body. In the event that if blood in body is insufficient then many organs like cerebrum suffer and if heart quits working by, death happens inside minutes. Life is totally subject to effective working of the heart. The term Heart sickness alludes to illness of heart & vessel framework inside it.

There are number of elements which build the danger of Heart infection:

- family history of coronary illness
- smoking
- Poor eating methodology
- high pulse
- cholesterol
- high blood cholesterol
- obesity
- Physical inertia

Symptoms of a Heart Attack

Manifestations of a heart assault can include:

- Discomfort, weight, largeness, or agony in the midsection, arm, or beneath the breastbone.
- Discomfort emanating to the back, jaw, throat, or arm.
- Fullness, heartburn, or stifling feeling (may feel like indigestion).
- Sweating, queasiness, heaving, or unsteadiness.
- Extreme shortcoming, nervousness, or shortness of breath.
- Rapid or not regular heart beats

Types of heart Disease

Heart illness is a wide term that incorporates different sorts of sicknesses influencing diverse segments of the heart. Heart signifies "cardio." Therefore, all heart sicknesses fit in with the class of cardiovascular ailments.

A few sorts of Heart illnesses are

a). Coronary illness: It otherwise called coronary supply route malady (CAD), it is the most well-known kind of coronary illness over the world. It is a condition in which plaque stores obstruct the coronary veins prompting a lessened supply of blood and oxygen to the heart.

b) Angina pectoris: It is a therapeutic term for midsection torment that happens because of deficient supply of blood to the heart. Otherwise called angina, it is a cautioning sign for heart assault. The midsection torment is at interims running for few seconds or minutes.

c). Congestive heart disappointment: It is a condition where the heart can't pump enough blood to whatever is left of the body. It is generally known as heart disappointment.

d). Cardiomyopathy: It is the debilitating of the heart muscle or a change in the structure of the muscle because of lacking heart pumping. A portion of the normal reasons for Cardiomyopathy are hypertension, liquor utilization, viral diseases, and hereditary imperfections.

e). Innate coronary illness: It alludes to the development of an irregular heart because of a deformity in the structure of the heart or its working. It is additionally a sort of innate ailment that kids are conceived with.

f). Arrhythmias: It is connected with an issue in the musical development of the pulse. The pulse can be abating, quick, or unpredictable. These unusual heartbeats are brought about by a short out in the heart's electrical framework.

g). Myocarditis: It is an aggravation of the heart muscle normally brought on by popular, parasitic, and bacterial contaminations influencing the heart. It is an exceptional malady with few indications like joins agony, leg swelling or fever that can't be

specifically identified with the heart.

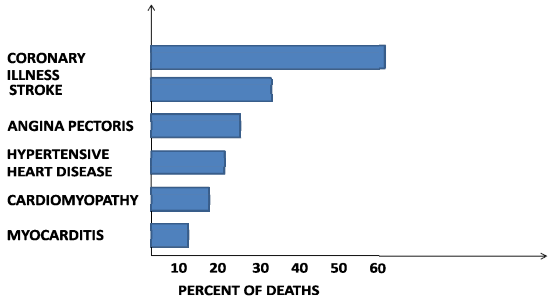

Fig 1: A Bar chart showing percent of deaths caused due to different types of heart diseases.

III. DATA MINING IN PREDICTION OF HEART DISEASE

Despite the fact that data mining has been around for more than one decade, its potential is just been realized now. Data mining joins factual examination, machine learning and database engineering to extract hidden patterns and connections from substantial databases. Fayyad characterizes data mining as "a procedure of nontrivial extraction of implied, lastly not known and possibly valuable data from the information that is stored in a database" [10]. Giudici characterizes data mining as "a procedure of choice, investigation and demonstrating of vast amounts of information to find regularities or relations with the point of getting clear and helpful results for the manager of database".

Data mining mainly uses two methodologies: supervised and unsupervised learning. In supervised learning, a training set is utilized to learn model parameters though in unsupervised learning no training set is utilized like in k-means clustering. The two most normal modeling goals of data mining are classification and prediction. Classification models classify discrete, unordered values or data whereas prediction models predicts about continuous valued. Decision trees and Neural Networks are examples of classification models while Regression, Association Rules and Clustering are examples of prediction algorithm [2].

In this prediction of heart disease, we will use the following classification models of data mining are analyzed:

A. Decision trees
B. Neural networks
C. Naive Bayes Classifier

A. Decision trees :

The decision tree approach is one of the most powerful techniques in classification in data mining. It builds the models in the form of tree structure. Mainly, dataset breaks in small sets and concurrently, an associated decision tree is formed. Decision trees can handle both numerical data and categorical data. For medical purpose, decision trees determine order in different attributes and a decision is then taken based on the attribute.

There are various decision tree algorithms that are used. Most preferred algorithm is ID3 i.e. Iterative Dichotomized 3 by J. R. Quinlan. ID3 uses information gain and entropy to classify data in tree structure.

Iterative Dichotomized 3 Algorithm: The algorithm produces decision trees using Shannon Entropy.

Steps:

a) Build Classification Attribute (from the table).

b) Compute Classification Entropy

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i)$$

Where, X is current data set for which entropy can be calculated, n is set of classes in X and p(x) is proportion of the number of elements in class n to the number of elements in set X.

Entropy is figured for each one remaining quality. The property with the littlest entropy is utilized to part the set on this iteration. The higher the entropy, the higher the possibility to enhance classification.

c) For each one attribute in table, compute

Information Gain utilizing classification attribute.

$$IG(A,X) = H(X) - \sum_{t \in T} p(t)H(t)$$

Where H(X) - Entropy of set X, T is subset created from splitting set X, p(t) is proportion of elements in **t** to the number of elements in X.

The attribute which have highest information gain

is used to split the set X on particular iteration.

d) Select Attribute with the highest gain to following Node in the tree (beginning from the Root hub).

e) Remove Node Attribute, making decreased table.

f) Repeat steps 3-5 until all attributes have been utilized, or the same classification values stays in rows of reduced table. At that point, smallest tree is preferred.

Id3 attempt in making short decision tree out of set of learning data, shortest is not generally the best classification. Due to limitation, it is succeeded by Quinlan's C4.5 and C5.0 calculations.

Advantages:

- easy to understand, interpret
- Rules are easily generated.
- Implicit perform feature selection
- Allows addition of number of new data.

Disadvantages:

- Can suffer from over fitting.
- Non numeric data is difficult to handle.
- Tress with many branches is difficult to understand.
- Time consuming.

B. Neural networks:

An artificial neural network is information processing method encouraged by biological nervous system. Dr. Robert Hecht-Nielsen. He defines a neural network as:"a computing system made up of a number of simple, highly interconnected processing elements, which process information by their dynamic state response to external inputs".(In "Neural Network Primer: Part I" by Maureen Caudill, AI Expert, Feb. 1989)

Neural network is organized into number of layers consisting of huge number of elements that are highly interconnected i.e. neurons that have an activation function. Different patterns are generated with input layer that communicates with one or more hidden layers and finally output layer is generated. Mostly, ANN consists of 'Learning rule' that modify the weight of connections. Learning in neutral net can be of both types i.e. supervised learning and unsupervised learning. An artificial neural network consists of three layers i.e. input layer, hidden layer and output layer. The principal layer is the input layer and last layer is output layer. Between input and output layer, there may be extra layer i.e. hidden layer. a neural network can easily be trained to perform different functions by adjustments in values of weight among elements

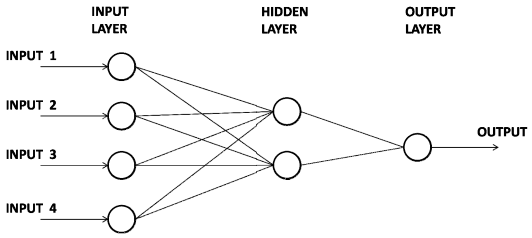

Fig1: Basic Model of an artificial neural network with four inputs and one output.

There are different classes of network architectures i.e. single layer feed forward network, multi layer feed forward network, recurrent network. Artificial neural networks make a useful tool to help doctors to analyze, model complex clinical data. Most Das, Turkoglu and Sengurn constructed a neural networks based methodology for diagnosing of the heart disease. (7)

Advantages:

- Neural networks can easily handle missing or noise data.
- Once trained, does not need to reprogram.
- It can easily work with large number of datasets.

Disadvantages:

- Neural network needs training to operate well.
- High processing time is required for large networks.
- Neural networks cannot be retrained i.e. if there is any modification in data, it is

almost impossible to add to an existing network.

C. Naive Bayes Classifier:

A Naive Bayes classifier is a simple probabilistic classifier that depends on Bayes' theorem with strong i.e. naive independence assumptions. It is also be called as "independent feature model". In general terms, a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. Naive Bayes classifiers are trained to work in supervised learning.

Naive Bayes classifier mainly pre assumes the effect of a variable value on predefined class that is not dependent on value of other variable. This is called as property of class conditional independence. It is particularly suited when the dimensionality of the inputs is high. Naïve Bayesian is mainly used to form models with predictive capabilities.

Bayes' Theorem:

Probability (B given A) = (Probability (A and B)/ Probability (A))

Assume X as a data tuple. Let H be any hypothesis. P (H|X) be posterior probability of the H that is conditioned on X. In the same way, P (X|H) is the posterior probability of X condition on H.

P (H|X) = (P(X|H)P(H)/ P (X))
P(H) is prior probability of H.

Naive Bayes Algorithm:

1. Assume D to be training set of tuple. Every record can be represented by n-dimensional attribute vector i.e. $X=(x_1, x_2\ldots, x_n)$,predicting n measurements on tuple from n attributes, i.e. A1 to An.
2. Let m number of class for prediction (C_1, C_2……., C_m). As for record X, the classifier predict that X will belong to the class with maximum posterior probability that is conditioned on X. Naïve Bayes predict that the tuple x will belong to class Ci only if P (Ci|X)>P (Cj|X). Therefore we have to maximize P(Ci|X). By Bayes' theorem:

$$P(Ci|X) = \frac{P(X|Ci)*P(Ci)}{P(X)}$$

3. Because P(X) is constant in all classes, therefore P (X|Ci)* P (Ci) need be maximized.
4. As then assumption of class conditional independence is done. Therefore it is pre assumed that value of attributes are conditionally independent of each other.
   Thus,
   $$P(X|Ci) = \prod_{K=1}^{m} P(x_k|Ci)$$
   $$= P(x_1|Ci)* P(x_2|Ci)\ldots\ldots P(x_m|Ci)$$
5. To predict class of X, P(X|Ci)P(Ci) is calculated for each class Ci. Naive Bayes predict that class label of X is Ci class if
   $$P(X|Ci)P(Ci) > P(X|Cj)P(Cj)$$
   for $1 \le j \le m, j \ne i$

In medical DM, Naïve Bayes classifier plays a crucial role. It shows high performance as if attributes are not dependent on one other, one can easily use it in medical diagnosis. As in medical data, there are missing values and this classifier can easily handle missing values. [9]

Advantage:

- Easy handle of large amount of data.
- It mainly require small amount of training set to estimate the parameters i.e. mean and variance needed for classification.
- Fast to train and Fast to classify
- Not sensitive to irrelevant features
- Handles real and discrete data
- Handles streaming data well

Disadvantages:

- Loss of accuracy
- Practically, there are dependencies among variables, but these dependencies are not handled by the classifier.
- Assumes independence of features

IV. CONCLUSION

In this paper we studied how data mining techniques brings with set of techniques to find out

hidden patterns for making decision in healthcare organizations. We focussed on classification methods of data mining used in data discovery. Different classification techniques of data mining have merits and demerits for data classification and knowledge extraction.

Furthermore, neural networks, decision trees or naïve Bayes can be studied in more detail to implement an algorithm that is helpful in healthcare organizations.

REFERENCES

[1] K Raj Mohan, Ilango Paramasivam ,Subhashini, SathyaNarayan " Prediction and Diagnosis of Cardio Vascular Disease – A Critical Survey", 2014 World Congress on Computing and Communication Technologies.

[2] Ms. Priti V. Wadal, Dr. S. R. Gupta, "Predictive Data Mining For Medical Diagnosis: An Overview Of Heart Disease Prediction "International Conference on Industrial Automation and Computing (ICIAC- 12-13th April 2014)

[3] Aqueel Ahmed, Shaikh Abdul Hannan, **"**Data Mining Techniques to Find Out Heart Diseases: An Overview", International Journal of Innovative Technology and Exploring Engineering (IJITEE), September 2012.

[4] K.Srinivas B.Kavihta Rani Dr. A.Govrdhan, "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks", (IJCSE) International Journal on Computer Science and Engineering, 2010.

[5] Deepali Chandna, "Diagnosis of Heart Disease Using Data Mining Algorithm", 1678-1680, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014.

[6] K.Sudhakar, Dr. M. Manimekalai, " Study of Heart Disease Prediction using Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering Volume 4, Issue 1, January 2014.

[7] Qeethara Kadhim Al-Shayea, "Artificial Neural Networks in Medical Diagnosis", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 2, March 2011.

[8]Ms.Rupali R.Patil," Heart Disease Prediction System using Naive Bayes and Jelinek-mercer smoothing", International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 5, May 2014.