Predict Sympathy Infection using Naive Bayesian Algorithm

M. Jai Ganesh
TRP Engineering College,
Tiruchirappalli , India.

S. Madhuranjini
TRP Engineering College,
Tiruchirappalli , India.

J. Merlin Monica
TRP Engineering College,
Tiruchirappalli , India.

P. Renuka
TRP Engineering College,
Tiruchirappalli , India.

C. Priyanka
TRP Engineering College,
Tiruchirappalli , India.

***Abstract* - Life is dependent on competent functioning of heart, because heart is necessary part of our body. If function of heart is not suitable, it will affect the other body parts of human such as brain, kidney etc. Heart disease is a disease that effects on the function of heart. There are number of factors which increases risk of heart disease. At the present days, in the world heart disease is the main cause of deaths. The World Health Organization (WHO) has expected that 12 million deaths occur worldwide, every year due to the heart diseases. Prediction by using data mining techniques gives us accurate result of disease. IHDPS (Intelligent Heart Disease Prediction System) is used to extract the hidden data from a historical database. It can also support for diagnosing the heart disease for health care departments. A few kinds of heart disease are cardiovascular diseases, heart attack, coronary heart disease and Stroke. Stroke is a type of heart disease; it is caused by narrowing, blocking, or hardening of the blood vessels that go to the brain or by high blood pressure. System based on the risk factors would not only help medical professionals but also it would give patients a warning about the probable presence of heart disease even before he visits a hospital or goes for costly medical Checkups. Hence this system presents a technique for prediction of heart disease. These techniques involve one successful data mining technique named Naïve Bayesian algorithm.**

Keywords-Data mining; Clustering; Classification; Naïve Bayesian algorithm; Semi supervised learning. 1.

INTRODUCTION

1.1 DATA MINING:

Data mining (the analysis step of the "Knowledge Discovery in Databases" process), a field at the intersection of computer science and statistics, is the process that attempts to discover patterns in large data sets. Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

2.RELATED WORK

2.1 Data Mining Techniques

There are several major *data mining techniques* have been developed and used in data mining projects recently including association, classification, clustering, prediction and sequential patterns.

a.Association

Association is one of the best known data mining technique. In association, a pattern is discovered based on a relationship of a particular item on other items in the same transaction. For example, the association technique is used in *market basket analysis* to identify what products that customers frequently purchase together. Based on this data businesses can have corresponding marketing campaign to sell more products to make more profit.

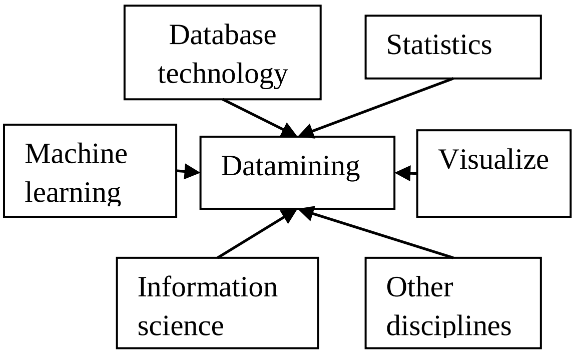

Fig 2.1 Techniques of data mining

b. Classification

Classification is an one of the data mining technique. This technique is used to classify an each item in a set of data into a predefined set of classes (or) groups. Goal of this technique is to accurately predict the target class for each case in the data. Classification technique is based on "Supervised learning".

c. Clustering

Clustering is a data mining technique that makes meaningful or useful cluster of objects that have similar characteristic using automatic technique. Is a process of partitioning a set of meaningful sub classes, called “clusters”. To make the concept clearer, can take book shop as an example. In a book shop, books have a wide range of topics available. The challenge is how to keep those books in a way that readers can take several books in a specific topic without hassle. Clustering technique is based on “Unsupervised learning”.

3. IMPLEMENATION DETAILS

A. System Framework

The system framework diagram is shown below, and by looking that it gives as the clear idea about the system and it’s working. In the system there are the main key terms are given below:

- ✓ Data Set Acquisition
- ✓ Preprocessing
- ✓ Clustering
- ✓ Feature Selection
- ✓ Classification

Module 1: Prediction Phase

- Dataset Acquisition

In this module, upload the datasets. Gather the data from hospitals, data centers and cancer research centers. The collected data is pre-processed and stored in the knowledge base to build the model. The „Diagnosis‟ attribute is used to predict the heart disease with value “2” for patient having heart disease and “1” for patient having no heart disease. The „patient ID‟ attribute is used as a key and others are input attributes.

- Preprocessing

Data pre-processing is an important step in the data mining process. The phrase "garbage in, garbage out" is particularly applicable to data mining and machine projects. Data-gathering methods are often loosely controlled, resulting in out-of-range values, impossible data combinations, missing values, etc. Analyzing data that has not been carefully screened for such problems can produce misleading results.

- Clustering

Clustering is a technique in data mining to find interesting patterns in a given dataset .The K-means algorithm is an evolutionary algorithm that gains its name from its method of operation. The algorithm clusters information’s into k groups, where k is considered as an input parameter. It then assigns each information’s to clusters based upon the observation’s proximity to the mean of the cluster. The cluster’s mean is then more computed and the process begins again. The k-means algorithm is one of the simplest clustering techniques and it is commonly used in medical data and related fields. K-Means algorithm is a divisive, unordered method of defining clusters.

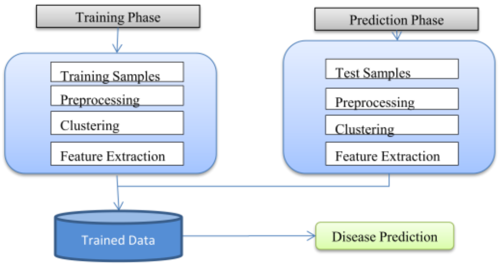

Fig 3: System Framework

- Feature Selection

In this module is used to select the features of the given dataset. Attribute selection was performed to determine the subset of features that were highly correlated with the class while having low inter correlation.

- Classification

In our project we apply Naïve Bayesian alogoithm for classification phase. Naïve Bayesian algorithm is derived from “Bayes Theorem”. This algorithm gives an “Probabilistic” output. Assumes a probabilistic model which allows us to solve the diagnostic and predictive problems. Bayes classification has been proposed which is based on Bayes rule of conditional probability. Naïve Bayesian rule is a technique used to estimate the likelihood of a property from the given data set. The approach is called “naïve” because it assumes the independence between the various attribute values. Bayesian classification can be seen as both a descriptive and a predictive type of algorithm. The probabilities are descriptive and used to predict the class membership for a target tuple.

Module 2: Prediction Phase

The prediction phase consisting with the following main operations those are listed below:

- In this phase user enter our details of clinical parameters.suchas age,weight,cholesterol,restingBP,restingECG,old peak+slope,exercise,inducedangina,thalac,thal,
- gender,chest pain,ca,fasting Bs.
- Next steps are done with the help of training phase.

Module 3: Disease Prediction

- Making database proper in to required numerical format.
- Apply the Naïve Bayesian algorithm to get an output.

B. Algorithms Used

There various algorithms are available for the data mining we can choose as per the requirement of our database, and requiring approach.

Naive Bayesian Algorithm

It is a classification technique based on Bayes’ Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Naive Bayes model is easy to build and particularly useful for very large data sets. Bayes theorem is given below,

$$P(A/B) = \frac{P(B/A)\ P(A)}{P(B)}$$

Above

- ✓ P(A/B) : Output
- ✓ P(B/A) : Likelihood of patient.
- ✓ P(A) : Prior probability
- ✓ P(B) : Training set

Algorithm:

Input : Parameters entered by users.

Output: Prediction of Sympathy Infection.

Steps:

Step 1: Initially preprocess the training sets.

Step 2: Then apply clustering phase. In clustering phase use K-means algorithm. In this step we get cluster values for each instance.

Step 1 &2 are comes under the "Training phase".

Step 3: Apply feature selection concept.

Step 4: This is a classification phase. Here use "Naïve Bayesian" algorithm.

Step 3 &4 are comes under the "Testing phase".

Other Concept:

In our project other concept is used to predict heart disease. i.e Semi Supervisied Learning.

Semi supervised learning means, it deals about "limited" amount of labeled data and "unlimited" amount of unlabeled data. Semi supervised learnhing is placed between of supervised and unsupervised learning. It may refer to either Transductive learning (or) Inductive learning.

4. DATASET USED IN THE SYSTEM

The database which is used in the system in that 13 clinical parameters are used for the prediction of Heart Disease.

Dataset parameters are listed in Table 4.1

5.CONCLUSION

Decision support in heart disease prediction system developed using Naïve Bayesian Classification. The system is expandable in the sense that more number of records are attributes can be incorporated and new significant rules can be generated using underlying data mining technique. Naïve Bayesian is the algorithm which is showing the "High Accuracy" because it is a "Probabilistic classifier".

S.N0	Parameters Name	Description
1.	Age	In years
2.	Gender	Value 1 = male, Value 0 = female
3.	Chest pain	Value 1: typical angina Value 2:non-angnal pain Value 3:asymtomatic
4.	Resting Blood Pressure	Mm Hg on admission to hospital
5.	Fasting Bs	Value 1:>120 mg/dl Value 0:<120 mg/dl
6.	Cholesterol	Serum cholesterol in mg/dl
7.	Resting ECG	Value 0: normal Value 1: having ST-T wave abnormality Value 2: showing probable
8.	Thalac	Maximum heart rate achived in number
9.	Exercise Induced Angina	Value 1: yes, Value 0: no
10.	Oldpeak + slope	Value 1: unsloping Value 2: flat Value 3: downsloping
11.	Thal	Value 3: normal Value 6: fixed defeat Value 7: reversible defeat
12.	Ca	Number of major vesels (0-3) colored by flursopy.
13.	Diagnosis	Value 0: normal Value 1: <50% affected Value 2:>50% affected Value 3:<75% affected Value 4: 100% affected

Table 4.1 Diagnosis Parameters

6. REFERENCES

[1] Swati A sonawale & Roshani Ade: Dimensionality Reduction: An Effective Technique for Feature Selection : International Journal of Computer Applications, Volume 117-No 3, May (2015).

[2] Tarun Kumar Gupta, Chanchal Kumar, Shiv Prakash and Mukesh Prasad: Dimensionality Reduction Techniques and its Applications: Computer Science Systems Biology, (2015).

[3] K.Sudhakar & Dr.M.Manimekalai : Study of Heart disease prodiction using data mining: International journal of Advanced Research in Computer Science and Software Engineering, Vol 4, Issues 1, ISSN: 2277 i28x, pp 1157-1160, January (2014**).**

[4] N. Aaditya Sunder, P. PushpaLatha, “Performance analysis of classification data mining techniques over heart disease database” Inernational Journal Of Engineering Science and Advance Technology”-vol-2 issue-3,470-478,MayJune 2012.

[5] Shadab Adam Pattekari and Asma Parveen” PREDICTION SYSTEM FOR HEART DISEASE USING NAIVE BAYES” International Journal of Advanced Computer and Mathematical Sciences ISSN 2230-9624. Vol 3, Issue 3, 2012.