Prediction of Heart Disease Using Machine Learning Algorithms

Sonam Nikhar[1], A.M. Karandikar[2]

[1]M.Tech Student, Department of CSE, Ramdeobaba College of Engineering and Management, Nagpur, India
[2]Assistant Professor, Department of CSE, Ramdeobaba College of Engineering and Management, Nagpur, India

Abstract— *The successful experiment of data mining in highly visible fields like marketing, e-business, and retail has led to its application in other sectors and industries. Healthcare is being discovered among these areas. There is an opulence of data available within the healthcare systems. However, there is a scarcity of useful analysis tool to find hidden relationships in data. This research intends to provide a detailed description of Naïve Bayes and decision tree classifier that are applied in our research particularly in the prediction of Heart Disease. Some experiment has been conducted to compare the execution of predictive data mining technique on the same dataset, and the consequence reveals that Decision Tree outperforms over Bayesian classification.*

Keywords— Data mining, Heart Disease Prediction, Naïve Bayes Classifier, Decision tree Classifier.

I. INTRODUCTION

Data mining is the computer based process of extracting useful information from enormous sets of databases. Data mining is most helpful in an explorative analysis because of nontrivial information from large volumes of evidence. Medical data mining has great potential for exploring the cryptic patterns in the data sets of the clinical domain. These patterns can be utilized for healthcare diagnosis. However, the available raw medical data are widely distributed, voluminous and heterogeneous in nature. These data need to be collected in an organized form. This collected data can be then integrated to form an medical information system. Data mining provides a user-oriented approach to novel and hidden patterns in the data

The data mining tools are useful for answering business questions and techniques for predicting the various diseases in the healthcare field. Disease prediction plays a significant role in data mining.

This paper analyzes the heart disease predictions using classification algorithms. These invisible patterns can be utilized for health diagnosis in healthcare data. Data mining technology affords an efficient approach to the latest and indefinite patterns in the data. The information which is identified can be used by the healthcare administrators to get better services. Heart disease was the most crucial reason for victims in the countries like India, United States. Data mining techniques like clustering, Association Rule Mining, Classification algorithms such as Decision Tree [2], C4.5 algorithm, Naive Bayes [4] are used to explore the different kinds of heart - based problems. These algorithms can be used to enhance the data storage for practical and legal purposes.

II. RELATED WORK

Numerous works in literature related to the diagnosis of Heart disease using data mining techniques have motivated this work. A brief literature survey is presented here.

A model Intelligent Heart Disease Prediction System built with the assistance of data mining techniques namely, Neural Network, Naïve Bayes, and Decision Tree. Results show that each technique has its infrequent strength in realizing the objectives of the defined mining goals. IHDPS can answer complex “what if” queries which conventional decision support systems cannot be proposed by Sellappan Palaniappan et al. [2]. The results illustrated the uncouth strength of each of the methodologies in comprehending the goal of the specified mining objectives. IHDPS was capable of responding queries that the traditional decision support systems were not able to. It facilitated the installation of crucial knowledge such as patterns, relationships amid medical factors connected with heart disease. IHDPS remains well-being web-based, user-friendly, reliable, scalable and expandable.

The diagnosis of Heart Disease, Blood Pressure and diabetes with the aid of neural networks was introduced by Niti Guru et al. [7]. Experiments were carried out on a sampled data set of patient’s records. The Neural Network is trained and tested with 13 input variables such as Blood Pressure, Age, Angiography's report and the like. The supervised network has been advised for diagnosis of heart diseases. Training was carried out with the help of back propagation algorithm. Whenever unfamiliar data was inserted by the doctor, the system identified the unknown data from comparisons with the trained data and produced a catalog of probable diseases that the patient is vulnerable to.

In 2014, M.A.Nishara BanuB.Gomathy Professor, Department of Computer Science and Engineering has published a research paper "Disease Forecasting System Using Data Mining Methods"[8].In this article, the pre-processed data is clustered using clustering algorithms as K-means to gather relevant data in a database. Maximal Frequent Item set Algorithm (MAFIA) is applied for mining maximal frequent model in heart disease database. The regular patterns can be classified into different classes using the C4.5 algorithm as training algorithm using the concept of information entropy. The result demonstrates that the designed prediction system is capable of predicting the heart attack successfully.

In 2012, T.John Peter and K. Somasundaram Professor, Dept of CSE presented a paper, "An Empirical Study on Prediction of Heart Disease using classification data mining technique"[5]. In this research paper, the use of pattern recognition and data mining techniques are used for prediction of risk in the medical domain of heart disease medicine is proposed here. Some of the limitations of the traditional medical scoring systems are that there is a presence of intrinsic linear combinations of variables in the input set, and hence they are not skilled at modeling nonlinear complex interactions in medical domains. This limitation is handled in this research by use of classification models which can implicitly detect complex nonlinear relationships between independent and dependent variables as well as the ability to identify all possible interactions between predictor variables.

In 2013, Shamsher Bahadur Patel, Pramod Kumar Yadav, and Dr. D. P.Shukla presented a research paper,"Predict the Diagnosis of Heart Disease Patients Using Classification Mining Techniques "[6].In this research paper, the health care industry, the data mining is mainly utilized for the prediction of heart disease. The objective of our works to predict the diagnosis of heart disease with a reduced number of attributes using Naïve Bayes, Decision Tree.

III. DATA SOURCE

Clinical databases have collected a significant amount of information about patients and their medical conditions. The term Heart disease encompasses the diverse conditions that affect the heart. Cardiovascular disease is the leading cause of casualties in the world. The term "cardiovascular disease" comprises a broad range of conditions that affect the heart and the blood vein and the way in which blood is pumped and circulated through the body.

Records set with medical attributes were obtained from the Cleveland Heart Disease database. With the help of the dataset, the patterns significant to the heart attack diagnosis are extracted. The records were split equally into two datasets: training dataset and testing dataset.

A total of 303 records with 76 medical attribute were obtained. All attributes are numeric-valued. We are working on a reduced set of attributes, i.e. only 19 attributes. The following table shows the list of attributes on which we are working.

Id: Identification Number
Age: Age in year
Sex: Sex (value 1: Male; value 0 :Female)
CP(Chest Pain): Value 1:Yes; Value 0:No
CPT(chest pain type): value 1:typical type 1 angina; value 2 : typical type angina; value 3: non-angina pain; value 4 : asymptomatic
trestbps: resting blood pressure (in mm Hg on admission to the hospital)
chol: serum cholesterol in mg/dl
fbs: (fasting blood sugar > 120 mg/dl)
restecg: resting electrocardiographic results
thalach: maximum heart rate achieved
exang: exercise induced angina
oldpeak: ST depression induced by exercise relative to rest
slope: the slope of the peak exercise ST segment Value 1: upsloping; Value 2: flat; Value 3: downsloping
ca: number of major vessels (0-3) colored by flourosopy
thal: 3 = normal; 6 = fixed defect
Smoke: I believe this is 1=Yes; 0=No (is or is not smoker).
dm: Value 1: history of diabetes; Value 0 : No such history of diabetes.
famhist: family history of coronary artery disease (1=yes; 0=No).
num: diagnosis of heart disease Value 0: No Risk ;Value 1:Low Risk ;Value 2:Risk Value 3:High Risk; Value 4:Higher Risk

Fig 1: Selected Cleveland Heart Disease Data Set Attributes.

IV. PROPOSED SYSTEM

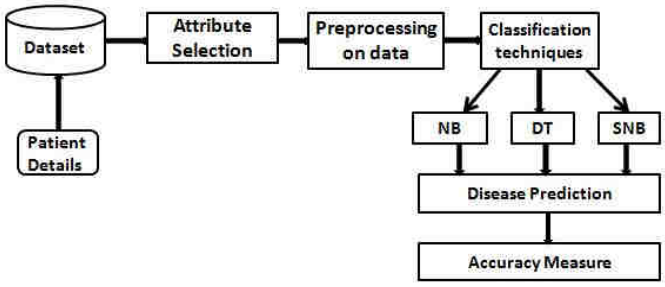

Fig 2: Block diagram of proposed system.

The working of this system is described in a step by step:

1. Dataset collection which contains patient details.
2. Attributes selection process selects the useful attributes for the prediction of heart disease.
3. After identifying the available data resources, they are further selected, cleaned, made into the desired form.
4. Different classification techniques as stated will be applied on preprocessed data to predict the accuracy of heart disease.
5. Accuracy measure compares the accuracy of different classifiers.

V. RESEARCH METHODOLOGY

In this section, we are introducing methods for a new proposed system.

5.1 Naïve Bayesian Classifier

In data mining, naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong independence assumptions between the features. Naive Bayes classifiers are intensely scalable, requiring some parameters linear in the number of variables (features/predictors) in a learning problem.

The Bayesian Classification depicts a supervised learning method as well as a statistical method for classification. Assumes a fundamental probabilistic model and it allows us to capture uncertainty about the model in a principled way by determining probabilities of the outcomes. It can solve diagnostic and predictive problems.

5.1.1 Implementation of Bayesian Classification

The Naïve Bayes Classifier technique is particularly suited when the amplitude of the inputs is high. Naïve Bayes model identifies the specialty of patients with heart disease. It shows the probability of each input attribute for the predictable state.

5.1.2. Bayes Rule

A conditional probability is the likelihood of some conclusion, C, given some evidence/observation, E, where a dependence relationship exists between C and E. This probability is denoted as P(C |E) where

$$P(C|E) = \frac{P(E|C)P(C)}{P(E)}$$

5.1.3 Naive Bayesian Classification Algorithm

5.1.3 Naive Bayesian Classification Algorithm

The naive Bayesian classifier, or simple Bayesian classifier, works as follows:

1. Let D be a training set of tuples and their associated class labels. As usual, each row is represented by an n-dimensional attribute vector, X=(x1, x2,…, xn), depicting n measurements made on the tuple from n attributes, respectively, A1, A2,.., An.

2. Suppose that there are m classes, C1, C2,…, Cm. Given a tuple, X, the classifier will foreshow that X belongs to the class having the highest posterior probability, conditioned on X. That is; the naïve Bayesian classifier predicts that tuple x belongs to the class Ci if and only if;

P (Ci|X)>P (Cj|X) for $1 \leq j \leq m, j \neq i$

Thus, we maximize P(Ci|X). The class Ci for which P(Ci|X) is maximized is called the maximum posterior hypothesis. By Bayes' theorem

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

3. As P(X) is constant for all classes, only P (X|Ci) P (Ci) need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, P(C1)=P(C2) =…=P(Cm), and we would, therefore, maximize P(X|Ci). Otherwise, we maximize P(X|Ci)P(Ci). Note that the class prior probabilities may be estimated by P(Ci)=|Ci,D|/|D|, where |Ci,D| is the number of training tuples of class Ci in D.

4. Given data sets with many attributes, it would be extremely computationally precious to compute P(X|Ci). To decrease computation in evaluating P(X|Ci), the naïve assumption of class conditional independence is made. This reckons that the values of the attributes are conditionally independent of one another, given the class label of the tuple (i.e., that there are no dependence relationships among the attributes). Thus;

$$P(X|C_i) = \prod_{k=1}^{n} P(xk|C_i)$$

=P(x1|Ci)x P(x2|Ci)x… P(xm|Ci).

We can easily estimate the probabilities P(x1|Ci), P(x2|Ci),…, P(xm|Ci) from the training tuples. Recall that here xk refers to the value of attribute Ak for tuple X.

5. To predict the class label of X, P(X|Ci)P(Ci) is evaluated for each class Ci. The classifier predicts that the class label of tuple X is the class Ci if and only if

P(X|Ci)P(Ci)>P(X|Cj)P(Cj) for $1 \leq j \leq m, j \neq i$

In other words, the predicted class label is the class Ci for which P(X|Ci)P(Ci) is the maximum.

5.2 Decision Tree

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible outcomes, comprising chance event outcomes, resource costs, and utility. It is one way to display an algorithm.

Decision trees are generally used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach an objective, but are also a popular tool for machine learning.

A decision tree is a flowchart-like structure in which each internal node depicts a "test" on an attribute, each branch represents the outcome of the test and each leaf node accounts for a class label. The path from the root to leaf depicts classification rules.

The basic algorithm for decision tree induction is a greedy algorithm that builds decision trees in a top-down recursive divide-and-conquer manner. The algorithm starts with the entire set of rows in the Training set, selects the best attribute that yields maximum information for classification, and originates a test node for this attribute. Then, top-down induction of decision trees

divides the current set of tuples according to their values of the current test attribute. Classifier generation stops, if all tuples in a subset pertain to the same class, or if it is not worth to proceed with an additional separation into further subsets, i.e. if further attribute tests produce only information for classification below a pre-specified threshold. The decision tree algorithm commonly uses an entropy-based measure known as "information gain" as a heuristic for selecting the attribute that will best split the training data into separate classes. The algorithm computes the information gain of each attribute, and in each round, the one with the highest information gain will be chosen as the test attribute for the given set of training data. A well-chosen split point should help in dividing the data to the best possible limit. After all, a primary criterion in the greedy decision tree approach is to build shorter trees. The best split point can be quickly evaluated by considering each unique value for that feature in the given data as a possible split point and calculating the associated information gain.

5.2.1 Information Gain

The critical step in decision trees is the selection of the best test attribute. The information gain measure is used to select the test attribute at each node in the tree.

First, another related term called entropy needs to be introduced. In general, entropy is a measure of the purity in an arbitrary collection of examples. Let S be a set consisting of s data samples. Suppose the class label attribute has m distinct values defining m different categories, Ck. Let si be the number of samples of S in class Ck. The expected information needed to classify a given sample is provided by;

$$\boldsymbol{I(S_1, S_2, \dots, S_m) = -\sum_{k=1}^{m} p_k \, log_2 \, (p_k)}$$

Where, p_k is the probability that an arbitrary sample belongs to class C_k and is estimated by s_k / s. Let attribute A have *v* distinct values, {a1,a2,…,av}.Attribute A can be used to partition S into *v* subsets, {S1,S2,…,Sv}, where S_j contains those samples in S that have value aj of A. Let s_{kj} be the number of samples of class Ck in a subset Sj. The entropy, or expected information based on the partitioning into subsets by A, is given by;

$$\boldsymbol{E(A) = \sum_{j=1}^{v} \frac{s_{1j} + \dots + s_{mj}}{s} I(s_{1j}, \dots, s_{mj})}$$

The term $\frac{s_{1j}+\dots+s_{mj}}{s}$ acts as the weight of the j^{th} subset and is the number of samples in the subset divided by the total number of samples in S. For a given subset S_j;

$$\boldsymbol{I(S_{1j}, S_{2j}, \dots, S_{mj}) = -\sum_{k=1}^{m} p_{kj} \, log_2 \, (p_{kj})}$$

Where;$p_{kj} = s_{kj}/|s_j|$ and is the probability that a sample in S_j belongs to class C_k. The entropy is zero when the sample is pure, i.e. when all the examples in the sample S belong to one class. Entropy has a maximum value of 1 when the sample is maximally impure, i.e. there are same proportions of positive and negative examples in the sample S.

The encoding information would be gained by branching on A is;

$$\boldsymbol{Information\ Gain(A) = I(s_1, s_2, \dots, s_m) - E(A)}$$

The attribute with the highest information gain is chosen as the test attribute for the current node. Such approach minimizes the expected number of tests needed to classify an object and guarantees that a simple (but may not be the simplest) tree is found.

VI. EXPERIMENTAL RESULTS AND DISCUSSION

From these results it is concluded that although most researchers are using different classifier techniques such as Neural network, SVM, KNN and binary discretization with Gain Ratio Decision Tree in the diagnosis of heart disease, applying Naïve Bayes and Decision tree with information gain calculations provides better results in the diagnosis of heart disease and better accuracy as compared to other classifiers . We surmise that the improvement in accuracy arises from the increased attributes. We have also observed that decision tree outperforms over Naïve Bayes. The decision tree classifier has better accuracy as compared to Naïve Bayes classifier.

VII. CONCLUSION AND FUTURE SCOPE

In this paper, we introduce about the heart disease prediction system with different classifier techniques for the prediction of heart disease. The techniques are Naïve Bayes classifier and decision tree classifier; we have analyzed that the decision tree has better accuracy as compared to naïve Bayes classifier. To increase the performance of the classifier in future, we will be working on Selective naïve Bayes classifier; It is known that Naïve Bayesian classifier (NB) works very well on some domains, and poorly on some. Our purpose is to improve the performance of the Naïve Bayesian classifier by removing unnecessary and irrelevant attributes from the dataset and only picking those that are most informative for the classification task. To achieve this, we

use the trees that are constructed by C4.5 and this technique is known as Selective Naïve Bayes classifier.

REFERENCES

[1] V. Manikantan and S. Latha, "Predicting the analysis of heart disease symptoms using medicinal data mining methods", International Journal of Advanced Computer Theory and Engineering, vol. 2, pp.46-51, 2013.

[2] Sellappan Palaniappan and Rafiah Awang, "Intelligent heart disease prediction system using data mining techniques", International Journal of Computer Science and Network Security, vol.8, no.8, pp. 343-350,2008.

[3] K.Srinivas, Dr.G.Ragavendra and Dr. A. Govardhan," A Survey on prediction of heart morbidity using data mining techniques",International Journal of Data Mining & Knowledge Management Process (IJDKP) vol.1, no.3, pp.14-34, May 2011.

[4] G.Subbalakshmi, K.Ramesh and N.Chinna Rao," Decision support in heart disease prediction system using Naïve Bayes", ISSN: 0976-5166, vol. 2, no. 2.pp.170-176, 2011.

[5] T.John Peter , K. Somasundaram, "An Empirical Study on Prediction of Heart Disease using classification data mining technique" IEEE-International Conference On Advances In Engineering, Science And Management (ICAESM - 2012) March 30, 31, 2012.

[6] Shamsher Bahadur Patel, Pramod Kumar Yadav and Dr. D. P.Shukla, "Predict the Diagnosis of Heart Disease Patients Using Classification Mining Techniques",IOSR Journal of Agriculture and Veterinary Science (IOSR-JAVS),Volume 4, Issue 2 (Jul. - Aug. 2013.

[7] Niti Guru, Anil Dahiya, Navin Rajpal, "Decision Support System for Heart Disease Diagnosis Using Neural Network", Delhi Business Review, Vol. 8, No. I (January - June 2007.

[8] M.A.Nishara Banu, B.Gomathy, "Disease Forecasting System Using Data Mining Methods," International Conference on Intelligent Computing Applications,2014

[9] Feixiang Huang, Shengyong Wang, and Chien-Chung Chan, "Predicting Disease By Using Data Mining Based on Healthcare Information System", IEEE International Conference on Granular Computing,2012.

[10] Chotirat "Ann" Ratanamahatana and Dimitrios Gunopulos,"Scaling up the Naive Bayesian Classifier:Using Decision Trees for Feature Selection", Computer Science Department University of California Riverside, CA 92521 1-909-787-5190

[11] Blake, C.L., Mertz, C.J.: "UCI MachineLearningDataset", http://mlearn.ics.uci.edu/databases/heartdisease/, 2004.