RESEARCH ARTICLE OPEN ACCESS

Heart Disease Prediction Using Naïve Bayes Algorithm and Laplace Smoothing Technique

Vincy Cherian [1], Bindu M.S [2]
Student [1], Reader [2]
Department of Computer Science
School Of Technology and Applied Sciences
India

ABSTRACT

This paper attempts to utilize the advantages of data mining technique for predicting the presence of heart disease when medical details of a patient are given. The medical details such as age, gender, fasting blood sugar, cholesterol, blood pressure etc. are used to predict the likelihood of having heart disease in a patient. The data mining technique used is a classification algorithm namely Naïve Bayes algorithm and a smoothing technique Laplace smoothing. This paper also compares the accuracy of prediction when the number of medical attributes used for prediction is decreased. The proposed decision support system is believed to avoid unnecessary diagnosis test conducted in a patient and the delay in starting appropriate treatment by quickly diagnosing heart disease in a patient. Thereby it saves both money and time. The decision support system helps doctors to diagnose patients without making unwanted practice variations caused due to doctor's intuition and inexperience. The system gives a second opinion regarding the patient's condition as from an experienced doctor since the prediction is made from a historical database containing large number of heart patient records.

***Keywords :*—** Decision support system, Data mining, Naïve Bayes classification algorithm, Laplace smoothing technique, heart disease.

I. INTRODUCTION

In current scenario, most of the patients complain about the various test conducted by hospitals for diagnosis, which cause them both money and time loss. In some cases, the delay in diagnosing the patient correctly results in the delay of starting the proper treatment. It may lead to disastrous consequences in case of deadly diseases. Sometimes, after conducting so many tests, the patient's results are negative and results in both money and time loss. All these are caused due to the doctor's wrong intuitions and inexperience. But we cannot blame doctors for this, since they suggest each diagnosis test based on their intuitions and experience after examining the patient which may go wrong.

This paper suggests a solution for overcoming these problems by utilizing the large amount of patient records collected by various hospitals or health care centers with the help of data mining techniques. The system thus developed can be used as a decision support system to seek a second opinion as from an experienced doctor. This will help to avoid unnecessary test conducted for diagnosis thereby saving both time and money. This helps hospitals to provide quality services at affordable cost. Doctors, medical students and nurses can use this system for second opinion. Patients can use this system if they have their test results.

Nowadays, some hospitals use decision support systems for simple queries, such as what is the average age of patients having a particular disease, whether it is more prevalent among men or women, whether it is more common among young or old people etc. They cannot run complex queries such as whether a patient is affected by a particular disease or not, which treatment is more effective among patients with deadly diseases after crossing a particular stage etc. Currently, the large amount of data collected from patients are simply stored in hospitals or health care centers and is not used for any other purpose. If we utilize the knowledge hidden in these databases, we can find solutions for many other problems that exist in health care field regarding the services provided to the patients.

This paper presents a solution for diagnosing patients with heart disease. This decision support system uses data mining technique Naïve Bayes algorithm for predicting whether a patient have heart disease or not and uses smoothing technique Laplace smoothing for increasing prediction accuracy.

II. LITERATURE REVIEW

The various studies were conducted regarding the diagnosis of heart disease and some are given below.

- An Intelligent Heart Disease Prediction System (IHDPS) [4] is developed by using data mining techniques Naive Bayes, Neural Network, and Decision Trees was proposed by Sellappan Palaniappan. 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 including linear and nonlinear characteristics of HRV (Heart Rate Variability), a novel technique was proposed for diagnosing cardiovascular disease by HeonGyu Lee, Ki Yong Noh and Keun Ho Ryu [9]. 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). In their experiment, SVM outperformed other classifiers.

- The prediction of Heart disease, Blood Pressure and Sugar with the aid of neural networks was proposed by Niti Guru, Anil Dahiya and Navin Rajpal [8]. The dataset contains records with 13 attributes. The supervised networks i.e. Neural Network with back propagation algorithm is used for training and testing of data.

- Latha Parthiban [7] proposed an approach on basis of coactive neuro-fuzzy inference system (CANFIS) for prediction of heart disease. The proposed CANFIS model combined the neural network adaptive capabilities and the fuzzy logic qualitative approach which is then integrated with genetic algorithm to diagnose the presence of the disease.

- Shruti Ratnakar, K. Rajeswari and Rose Jacob [10] used genetic algorithm to reduce the set of attributes and Naïve Bayes algorithm to develop a Heart Disease Prediction System which is implemented as web-based questionnaire application.

- Akhil Jabbar[6] proposes efficient associative classification algorithm using genetic approach for heart disease prediction. The main motivation for using genetic algorithm in the discovery of high level prediction rules is that the discovered rules are highly comprehensible, having high predictive accuracy and of high interestingness values.

III. DATA SOURCE

Record set with medical attributes was obtained from the Cleveland Heart Disease database [1]. The records were split equally into two datasets: training dataset and testing dataset. To avoid bias, records for every set were picked randomly. This database contains 76 attributes, but only 14 attributes including one predictive attribute is used. “Patient’s test” is employed as a record, one attribute as output and, the remaining are input attributes. It is assumed that issues like missing, inconsistent, and redundant data have all been resolved.

A. Predictable attribute

Diagnosis -- (Value 0: <50% diameter narrowing (no heart disease), Value 1: >50% diameter narrowing (has heart disease))

B. Input attributes

1. Age – in year
2. Gender (value 1: Male; value 0: Female)
3. Chest Pain Type – (Value 1: Typical angina, Value 2: Atypical angina, Value 3: Non-angina pain, Value 4: Asymptomatic)
4. Fasting Blood Sugar -- (Value 1: >120 mg/dl, Value 0: <120 mg/dl)
5. ECG – Resting electrographic results (Value 0: Normal, Value 1: Having ST-T wave abnormality, Value 2: Showing probable or definite left ventricular hypertrophy)
6. Exang - Exercise induced angina (Value 1: Yes, Value 0: No)
7. Slope – The slope of the peak exercise ST segment (Value 1: Up sloping, Value 2: Flat, Value 3: Down sloping)
8. CA – Number of major vessels colored by fluoroscopy (value 0-3)
9. Thal – Thallium test (Value 3: Normal, Value 6: Fixed defect, Value 7: Reversible defect)
10. Blood Pressure – mm Hg on admission to the hospital
11. Serum Cholesterol – mg/dl
12. Thalach – maximum heart rate achieved
13. Old peak – ST depression induced by exercise

IV. TECHNIQUES USED

Data mining is defined as “a process of nontrivial extraction of implicit, previously unknown and potentially useful information from the data stored in a database” or as “a process of selection, exploration and modeling of large quantities of data to discover regularities or relations that are at first unknown with the aim of obtaining clear and useful results for the owner of database”.

Data Mining is the process of discovering interesting patterns and knowledge from large amounts of data. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically.

Data mining uses two strategies: supervised and unsupervised learning. In supervised learning, a training set is used to learn model parameters whereas in unsupervised learning no training set is used (e.g., k means clustering is unsupervised). Each data mining technique serves a different purpose depending on the modeling objective. The two most common modeling objectives are classification and prediction. Classification models predict categorical labels (discrete, unordered) while prediction models predict continuous-valued functions. Decision Trees and Neural Networks use classification algorithms while Regression, Association Rules and Clustering use prediction algorithms.

This decision support system is developed using a classification algorithm. The data classification process includes two steps –

*1) **Building the classifier or model:*** This step is the learning step or the learning phase. In this step the classification algorithms build the classifier.

The classifier is built from the training set made up of database tuples and their associated class labels. Each tuple that constitutes the training set is referred to as a category or class. These tuples can also be referred to as sample, object or data points.

*2) **Using classifier for classification:*** In this step, the classifier is used for classification. Here the test data is used to estimate the accuracy of classification rules. The classification rules can be applied to the unlabeled data tuples if the accuracy is considered acceptable.

In this paper, the techniques used for predicting heart disease are Naïve Bayes classification algorithm and Laplace smoothing technique.

C. Naïve Bayes Algorithm

Naive Bayes or Bayes Rule [2] is the basis for many machine-learning and data mining methods. The rule (algorithm) is used to create models with predictive capabilities. It provides new ways of exploring and understanding data.

The Naïve Bayes Classifier technique is mainly applicable when the dimensionality of the inputs is high [2]. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods. Naïve Bayes model recognizes the characteristics of patients with heart disease. It shows the probability of each input attribute for the predictable state. Naive Bayes algorithm is preferred in the following cases.

- When the dimensionality of data is high.
- When the attributes are independent of each other. Otherwise, attributes are assumed to be independent in order to simplify the computations involved and, in this sense, is considered "naïve".
- When we expect more efficient output, as compared to other methods output.
- Exhibits high accuracy and speed when applied to large databases.

1) Bayes Rule:

A conditional probability is the likelihood of some conclusion say *C*, given some evidence/observation, *E*, where a dependence relationship exists between *C* and *E*.

This probability is denoted as P*(C |E)* where

$$P(C|E) = \frac{P(E|C)P(C)}{P(E)} \qquad (1)$$

2) Naive Bayesian Classification Algorithm :

The Naive Bayesian classifier, or simple Bayesian classifier [2], works as follows:

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

b) Suppose that there are m numbers of classes for prediction, C1, C2… Cm. Given a record, X, the classifier will predict 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

$$P\,(Ci|X) > P\,(Cj|X) \quad \text{for } 1 \leq j \leq m \text{ and } j \neq i \qquad (2)$$

Thus we maximize P(Ci|X). The class Ci for which P(Ci|X) is maximized is called the maximum posteriori hypothesis.

By Bayes theorem

$$P(Ci|X) = \frac{P(X|Ci)P(Ci)}{P(X)} \qquad (3)$$

c) 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 often assumed that the classes are equally likely, that is, P(C1) =P(C2) =…P(Cm-1) =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

$$P\,(Ci) = |Ci, D| \,/\, |D| \qquad (4)$$

where |Ci, D| is the number of training tuples of class Ci in D.

d) Given data sets with many attributes, it would be extremely computationally expensive to compute P(X|Ci). To reduce computation in evaluating P(X|Ci), the naïve assumption of class conditional independence is made. This presumes 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,

$$P(X|Ci) = \prod_{k=1}^{m} P(x_k\,|Ci)$$
$$= P(x_1|Ci) * P(x_2|Ci) * \ldots * P(x_n\,|Ci) \qquad (5)$$

We can easily estimate the probabilities P(x_1 |Ci), P(x_2|Ci)… P(x_n|Ci) from the database training tuples. Recall that here x_k refers to the value of attribute A_k for tuple X. For each attribute, we will see that whether the attribute is categorical or continuous-valued. For instance, to compute P(X|Ci), we consider the following:

i. If A_k is categorical, then P(x_k |Ci) is the number of tuples of class Ci in D having the value x_k for A_k, divided by |Ci,D|, the number of tuples of class Ci in D.

ii. If A_k is continuous-valued, then a bit more work is to be done, but the calculation is pretty straightforward. A continuous-valued attribute is typically assumed to have a Gaussian distribution with a mean μ and standard deviation σ, defined by

$$g(x,\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} \qquad (6)$$

So that

$$P(x_k|Ci) = g(x_k, \mu_{Ci}, \sigma_{Ci}). \qquad (7)$$

First, we compute μ_{Ci} and σ_{Ci} which are the mean (i.e, average) and standard deviation respectively, of the values of attribute A_k for training tuples of class Ci. Then substitute these values into the equation to estimate P(x_k|Ci).

e) In order 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

$$P(X|Ci)P(Ci) > P(X|Cj)P(Cj) \quad \text{for } 1 \le j \le m, j \ne i. \qquad (8)$$

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

D. Smoothing Technique

Smoothing is a technique to make an approximating function that attempts to capture important patterns in the data, while avoiding noise or other fine-scale structures/rapid phenomena.

1) Laplace smoothing:

Given an attribute d to be classified, Naive Bayes assumes that the attributes are conditionally independent and finds P(d|Ci). If we end up with a probability value of zero for some P(xk|Ci) where xk is the value for attribute Ak, it will return a zero probability for P(X|Ci).

To avoid this, we assume that training database D is so large that adding one to each count that we need would only make a negligible difference in the estimated probability value, so that it avoids the case of probability values of zero. This technique for probability estimation is known as the Laplacian correction or Laplacian estimator [2]. If we have 'q' counts to which we each add one, then we must add q to corresponding denominator used in the probability calculation. This technique makes prediction more accurate.

V. PROPOSED SYSTEM

The decision support system developed uses the above techniques to predict heart disease. Users can use either the prediction with 13 attribute if they have the test results of fluoroscopy, thallium test, ECG, ST depression etc. or the prediction with 6 attributes if they don't.

The 6 attributes used were age, gender, blood pressure, fasting blood sugar, cholesterol and exercise induced angina. These were selected since their results can be easily inputted by users without much help. Then the performance of the classifier thus formed, is evaluated. Then all 13 medical attributes selected from data source were used and its performance is evaluated.

VI. PERFORMANCE ANALYSIS

The following measures were used to analyze the performance of the prediction system.

E. Classifier Evaluation Measures

Neg are the negative tuples that were correctly labeled by the classifier. False positives (F_Pos) are the negative tuples that were incorrectly labeled by the classifier, while false negatives are the positive tuples that were incorrectly labeled by the classifier. The sensitivity [5] and specificity [5] measures can be used for calculating performance and precision is used for the percentage of samples labeled as "diseased" or "1".

1) *Sensitivity :*

It means recognition rate or true positive rate. It is used for measuring the percentage of sick people from the dataset.

$$\text{Sensitivity} = \frac{\text{TruePos}}{\text{Pos}} \qquad (9)$$

Where TruePos is the number of true positives (i.e. "Present" samples that were correctly classified) and Pos are the number of positive samples.

2) *Specificity :*

It means true negative rate. It is used for measuring the percentage of healthy people who are correctly identified from the dataset.

$$\text{Specificity} = \frac{\text{TrueNeg}}{\text{Neg}} \qquad (10)$$

TrueNeg is the number of true negatives (i.e." Absent" samples that were correctly classified) and Neg is the number of negative samples.

3) *Precision :*

It is used for the percentage of samples labeled as “diseased” or “1”. It is also known as positive predictive value. It is defined as the average probability of relevant retrieval [5].

$$\text{Precision} = \frac{\text{TruePos}}{\text{TruePos} + \text{FPos}} \qquad (11)$$

FPos is the number of false positives (“Absent” samples that were incorrectly labeled as “diseased” or “1”).

4) *Accuracy :*

Accuracy = Number of correctly classified samples/Total number of samples (12)

The true positives, true negatives, false positives and false negatives are also useful in assessing the costs and benefits (or risks and gains) associated with a classification model.

5) *Confusion Matrix :*

It is used for displaying the number of correct and incorrect predictions made by the model compared with the actual classifications in the test data. The matrix is represented in the form of n-by-n, where n is the number of classes. The accuracy of classification algorithm can be calculated using this matrix [5].

TABLE I
CONFUSION OR CLASSIFICATION MATRIX

Actual Class	**Predicted Class**	
	C1	***C2***
C1	True Positive	False Positive
C2	False Negative	True Negative

VII. RESULTS

To evaluate the classifier, I used 100 records as training dataset and 50 records as testing dataset.

F. Prediction with 6 attributes

Out of 50 testing records, 22 were true negatives and 15 were false negatives, 4 were false positive and 9 were true positives. 24 were positives and 26 were negatives. The values obtained are

- Sensitivity = 9/24 = 0.375
- Specificity = 22/26 =0.846
- Precision = 9/ (9+4) = 9/13 = 0.692
- Accuracy = (9 + 22)/50 = 0.62

TABLE II
CLASSIFICATION MATRIX FOR PREDICTION WITH 6 ATTRIBUTES

Actual Class	**Predicted Class**	
	With Heart Disease	***Without Heart Disease***
With Heart Disease	9	15
Without Heart Disease	4	22

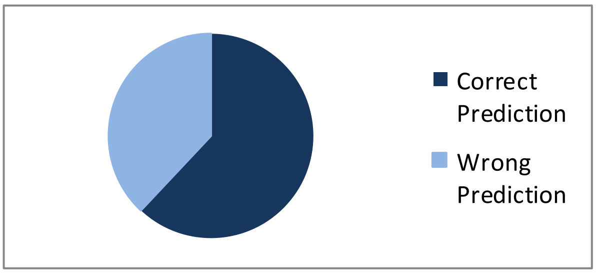

Fig. 1 Prediction with 6 attributes.

G. Prediction with 13 attributes

Out of 50 testing records, 25 were true negatives and 6 were false negatives, 1 was false positive and 18 were true positives. 24 were positives and 26 were negatives. The values obtained are

- Sensitivity = 18/24 = 0.75
- Specificity = 25/26 =0.961
- Precision = 18/(18+1) = 18/19 = 0.947
- Accuracy = (25 + 18)/50 = 0.86

TABLE III
CLASSIFICATION MATRIX FOR PREDICTION WITH 13 ATTRIBUTES

Actual Class	**Predicted Class**	
	With Heart Disease	***Without Heart Disease***
With Heart Disease	18	6
Without Heart Disease	1	25

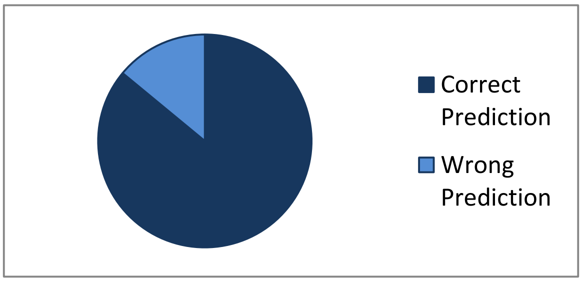

Fig. 2 Prediction with 13 attributes.

Therefore, the result is more accurate when 13 attributes were used. So it is better to use more number of relevant attributes whose value can be obtained through low cost diagnostic tests in order to reduce the money spend for diagnosis.

VIII. CONCLUSIONS

This decision support system is developed using Naive Bayesian classification algorithm and Laplace smoothing technique. The system predicts whether a patient have heart disease or not. Laplace smoothing technique makes more accurate results than Naive Bayes alone to predict patients with heart disease. The system is expandable in the sense that more number of records or attributes can be incorporated. Presently the users can use either 13 attributes prediction or 6 attributes prediction if they don't know the results of fluoroscopy test, thallium test etc. But the one with 13 attribute is more accurate since it has 86% accuracy. The system can also incorporate other data mining techniques for prediction. This system can serve as a training tool to train nurses and medical students to diagnose patients with heart disease. It can also provide decision support to assist doctors to make better clinical decisions or at least provide a "second opinion." A large dataset would definitely give better results. It is also necessary to test the system extensively with input from experienced cardiologists. The system can be expanded to include other areas or fields for disease prediction.

ACKNOWLEDGEMENT

First of all, I am grateful to GOD who guides me throughout this life to do good deeds and achieve things that I have never imagined in my life. I would also like to thank my family, all teachers and fellow students of my college who encouraged me to prepare this paper. I especially thank my teacher and guide Dr. Bindu M.S for her constant support and advice throughout the preparation of this paper.

REFERENCES

[1] Blake, C.L., Mertz, C.J.: "UCI Machine Learning Databases",http://mlearn.ics.uci.edu/databases/heartdisease/, 2004.

[2] Han, J., Kamber, M.: "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 2006.

[3] K.Sudhakar, Dr. M.Manimekalai, Study of Heart Disease Prediction System using Data Mining, ISSN: 2277 128X, 2014.

[4] Sellappan Palaniappan, Rafiah Awang, Intelligent Heart Disease Prediction System Using Data Mining Techniques, 978-1-4244-1968- 5/08/$25.00 ©2008 IEEE.

[5] B.Venkatalakshmi, M.V Shivsankar, "Heart Disease Diagnosis Using Predictive Data mining", International Journal of Innovative Research in Science, Engineering and Technology ISSN 2319-8753 Vol.3,Special Issue 3, pp. 1873-1877 ©2014 ICIET.

[6] M.Akhil jabbar, Dr.Priti Chandra, Dr.B.L Deekshatulu, Heart Disease Prediction System using Associative Classification and Genetic Algorithm, International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies, 2012.

[7] Latha Parthiban and R.Subramanian, Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm, International Journal of Biological and Medical Sciences, 2008.

[8] Niti Guru, Anil Dahiya, NavinRajpal,"Decision Support System for Heart Disease Diagnosis Using Neural Network", Delhi Business Review, Vol. 8, No. 1 (January - June 2007)

[9] HeonGyu Lee, Ki Yong Noh, KeunHoRyu, "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.

[10] Shruti Ratnakar, K. Rajeswari, Rose Jacob., Prediction Of Heart Disease Using Genetic Algorithm For Selection Of Optimal Reduced Set Of Attributes, International Journal of Advanced Computational Engineering and Networking, ISSN (PRINT): 2320-2106, Volume – 1, Issue – 2, 2013.