ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 5, May 2014

Heart Disease Prediction System using Naive Bayes and Jelinek-mercer smoothing

Ms.Rupali R.Patil

Asst. Professor, Jawaharlal Nehru College of Engineering, (Affiliated to BAMU,Aurangabad),Maharashtra, India.

Abstract: Data Mining is non trivial extraction of implicit data, previously not known, and imaginably useful information from data. Data mining is an essential process where intelligent methods are applied in order to extract data patterns. Using data mining we can evaluate patterns which we can use in future to take intelligent decisions and we can present the knowledge we extracted in better way. Data Mining refers to using a variety of techniques to identify information or decision making knowledge in the database and extracting these in a way that they can put to use in areas such as decision making, predictions, for valuable forecasting and computation. The healthcare industry collects huge amounts of healthcare data which, unfortunately, are not "mined" to discover hidden information, to take decisions effectively, to discover the relations that connect parameters in a database is the subject of data mining. This research work has developed a Decision Support in Heart Disease Prediction System (HDPS) using data mining modelling technique, namely, Naïve Bayes. Using medical profiles such as age, sex, blood pressure and blood sugar, chest pain, ECG graph etc it can predict the likelihood of patients getting a heart disease. It is implemented in matlab as an application which takes medical test's parameter as an input. It can be used as a training tool to train nurses and medical students to diagnose patients with heart disease.

Keywords: Data mining, Jelinek-mercer smoothing for Naive Bayes, heart disease, Naïve Bayes, decision support

1. INTRODUCTION

Data Mining is the nontrivial process of identifying true, novel, potentially useful and finally understandable pattern in data with the wide use of databases and the explosive growth in their sizes. Data mining refers to extracting or "mining" knowledge from large amounts of data. Data mining is the search for the relationships,association and global patterns that exist in large databases but are hidden among large amounts of data. The essential process of Knowledge Discovery is the conversion of data into knowledge in order to aid in decision making, referred to as data mining. Knowledge Discovery process consists of an iterative sequence of data pre-processing like cleaning, data integration, correct data selection, data mining pattern identification and knowledge presentation. Data mining is the search for the relationships and global patterns that exist in large databases but are hidden among large amounts of data.

Many hospital information systems are designed to support patient invoice generation, inventory management and generation of some statistics. Some hospitals already used decision support systems, but are largely limited. They can answer simple queries like "What is the average age of patients who have heart disease?" , "After surgery how many patients have to stay for more than a week?", "Identify according to gender -who are unmarried, above 35 years old, and who have been given treatment for Heart failure." However they cannot address complex queries like "Provided patient records, predict the probability of patients likely to have a heart disease." The decisions in the hospital are always made based on doctors' intuition and experience rather than on the knowledge rich data hidden in the database. This process leads to unwanted biases ,mistakes and excessive medical expenditure which affects the quality of service provided to patients. The proposed system uses analysis to integrate and make proper decision in the clinic with computer-based patient records .This system could reduce medical decision mistake , improve patient safety, reduces unwanted practice variation, and enhance patient outcome. This suggestion is promising as data modelling and analysis tools, e.g., data mining, have the capacity to generate a knowledge rich environment which can help to significantly improve the quality of clinical decisions.

Research objectives

Most hospitals today employ sort of hospital information systems to manage their healthcare or patient data. These systems typically produce large amounts of data. There is a wealth of unknown information in these data that is largely not accessed. So how this data can be converted into useful information that can enable healthcare systems and practitioners to make intelligent clinical decisions. The main objective of this research work is to develop a Decision Support system in Heart Disease Prediction System (HDPS) using one data mining modelling technique, namely, Naïve Bayes and another one is the smoothing to improve performance. HDPS is implemented as an application in matlab which can answer user queries, it can discover and extract hidden knowledge (patterns and relationships) associated with heart disease from a historical heart disease database. We provide the report of the patient which indicates whether that particular patient having the heart disease or not. This suggestion is promising as data modelling and analysis tools, e.g., data mining, have an ability to generate a knowledge rich environment which can help to significantly improve the quality of clinical decisions.

The diagnosis of diseases is a significant and complex task in medicine. To detect heart disease from various factors or symptoms is a multi -layered issue which is not free from false presumptions often accompanied by effects that are not predictable. Thus the attempt to utilize knowledge and experience of number of specialists and clinical screening data of patients which collected in databases to facilitate the diagnosis process is considered a valuable option. Providing quality services at an affordable costs is a major constraint encountered by the healthcare organizations (hospitals, medical centers). Valuable quality service denotes the accurate diagnosis of patients and providing better and appropriate treatment. Poor clinical decisions may lead to loss and hence are rarely entertained. Besides, it is important that the hospitals decrease the expenditure of clinical test that the patient have to do further. Essential computer-based information and/or decision support systems can aid in achieving clinical tests at a reduced cost. Naive Bayes or Bayes' Rule is the basis for many machine-learning and data mining methods. The rule (algorithm) is used to create models with predictive abilities. It provides new aspects of exploring and understanding data. It learns from the "evidence" by calculating the correlation between the target (i.e., dependent) and other (i.e., independent) variables. We apply several smoothing models to Naive Bayes for for improving results, and study their performance. The experimental results on a large database show that the smoothing methods are able to significantly improve performance of Naive Bayes.

2. DATA SOURCE

Clinical databases have accumulated giant quantities of information about patients and their medical conditions. The term cardiovascular disease encompasses the diverse diseases that affect the heart. Heart disease is the major cause of casualties in the world. cardiovascular disease kills one person each thirty four seconds within the United States. Coronary heart disease, Cardiomyopathy and Cardiovascular diseases are some categories of heart diseases. The term "cardiovascular disease" includes a wide range of conditions that affect the heart and the blood vessels and the manner in which blood is pumped and circulated through the body. Cardiovascular disease (CVD) results in severe illness, disability, and death.

Record set with medical attributes was obtained from the Cleveland Heart Disease database. With the assistance of the dataset, the patterns vital to the heart attack prediction are extracted. The records were split equally into two datasets: training dataset and testing dataset. To avoid bias, for every set were hand-picked haphazardly.

The attribute "Diagnosis" is known as the predictable attribute with value "1" for patients with heart disease and value "0" for patients with no heart disease. "Patient's test" is employed as a record,last attribute as output and, the remainingt are input attributes. It is assumed that issues like missing , inconsistent, and redundant data have all been resolved.

Predictable attribute

1. Diagnosis (value 0: <50% diameter narrowing (no heart disease); value 1: >50% diameter narrowing (has heart disease))

Input attributes

1. Age in Year
2. Sex (value 1: Male; value 0: Female)
3. Chest Pain Type (value 1:typical type1 angina, value 2: typical type 2 angina, value 3:non-angina pain; value 4: asymptomatic)
4. Fasting Blood Sugar (value 1: >120 mg/dl; value 0: <120 mg/dl)
5. Restecg – 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:unsloping; value 2: flat; value 3: down sloping)
8. CA – number of major vessels colored by fluoroscopy (value 0-3)
9. Thal (value 3: normal; value 6: fixed defect; value 7: reversible defect)
10. Trest Blood Pressure (mm Hg on admission to the hospital)
11. Serum Cholestrol (mg/dl)
12. Thalach – maximum heart rate achieved
13. Old peak – ST depression induced by exercise
14. Heart Disease Present - 0:No 1: Yes

3. Implementation of Bayesian Classification

The Naïve Bayes Classifier technique is mainly applicable when the dimensionality of the inputs is high. 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.

3.1. Why preferred Naive bayes algorithm

Naive Bayes or Bayes' Rule 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.

Why to prefer naive bayes implementation:

1) When the data is high.
2) When the attributes are independent of each other.

3) When we expect more efficient output, as compared to other methods output.

3.2 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

$$P(C \mid E) = \frac{\mathrm{P(E \mid C)P(C)}}{\mathrm{P(E)}}$$

3. 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 Ca and Cp. As usual, each record is represented by an n-dimensional attribute vector, X=(x_1, x_2…,x_{n-1}, x_n), depicting n measurements made on the tuple from n attributes, i.e. A1 to An.

2. Suppose that there are m number 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)   for 1≤ j≤m and j ≠ i

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

$$P(Ci \mid X) = \frac{\mathrm{P(x \mid Ci)\ P(Ci)}}{\mathrm{P(x)}} \qquad (1)$$

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 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|, 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 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,

$$P(X \mid Ci) = \prod_{k=1}^{m} P(xk \mid Ci)$$

=P (x1|Ci) * P (x2|Ci) *… P (xm|Ci).

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

(a) If Ak is categorical, then P (Xk|Ci) is the number of tuples of class Ci in D having the value xk for Ak, divided by |Ci, D|, the number of tuples of class Ci in D.

5. 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)   for 1 ≤ j ≤ m, j ≠ i

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

3.4 Smoothing Technique for Naive Bayes

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) Jelinek-Mercer (JM) smoothing:

Given an attribute d to be classified, Naive Bayes (NB) assumes that the features are conditionally independent and finds the class Ci that maximizes P(Ci)P(d|Ci).

$$P(Ci) = \frac{|Ci|}{|C|} \quad - \qquad P(x \mid Ci) = \prod_{k=1}^{|d|} P(xk \mid Ci)$$

where |Ci| is the range of attributes in Class and |C| is the total number of attributes in collection. For NB, likelihood P(xk|Ci) is calculated by Jelinek-mercer smoothing as follows:

$$P(x \mid Ci) = (1-\lambda)P(x \mid Ci) + \lambda P(x \mid C)$$

Where P(x|Ci) is the smoothened probability of a test given the patient record with existing tests .

$$P(x \mid Ci) = (1-\lambda)\frac{c(x, Ci)}{\sum_{x \in V} c(x', Ci)} + \lambda P(x \mid C)$$

Where λ is balancing parameter ranges from 0 to 1, V is the size of database collection and P(x|C) be the maximum likelihood estimation of attribute in class C.

Steps to implement Naïve Bayes with Jelinek-mercer smoothing:

- Enter patient record.
- The two classes in which we have to classify the data are.
  - 0:HD absent ,1:HD present
- Decide the probability of each attribute for both the classes using the database with result as training.
- Use Jelinek mercer smoothing for calculating smoothed probability of that attribute.
  - In Jelinek-mercer smoothing the smoothing agent is used whose value ranges from 0 to 1.
  - The value used in implementation is set optimally to give better performance for naïve bayes.
- Calculate the maximized probability from both the classes.
- Decide the class for patient record.

3.5 Flowchart for implementation of classification on patient data

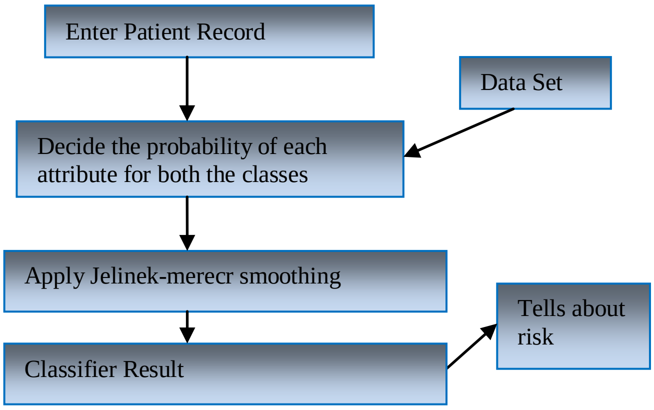

Enter Patient Record
Data Set
Decide the probability of each attribute for both the classes
Apply Jelinek-merecr smoothing
Tells about risk
Classifier Result

Fig. Implementation of Naïve Bayes with Jelinek-mercer smoothing on the patient data.

It learns from the “evidence” by calculating the classification by both the methods ,the probability calculated for each attribute differs in both the methods. The Naïve Bayes Classifier technique is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform by using more sophisticated classification methods and using smoothing techniques.

4.1 Performance Analysis:

Classifier Evaluation Measures:

Neg are the negative tuples that were correctly labelled by the classifier. False positives (F_Pos)are the negative tuples that were incorrectly labelled by the classifier, while false negatives are the positive tuples that were incorrectly labelled by the classifier. Sensitivity means Recognition rate or true positive rate The sensitivity and specificity measures can be used for calculating performance and precision is used for the percentage of samples labelled as “Yes”. These measures are defined as

$$\text{Sensitivity} = \backslash\ \frac{True\_Pos}{Pos}$$

Specificity means true negative rate .True_Pos is the number of true positives (i.e “Present” samples that were correctly classified) and Pos are the number of positive samples.

$$\text{Specificity} = \frac{True\_Neg}{Neg}$$

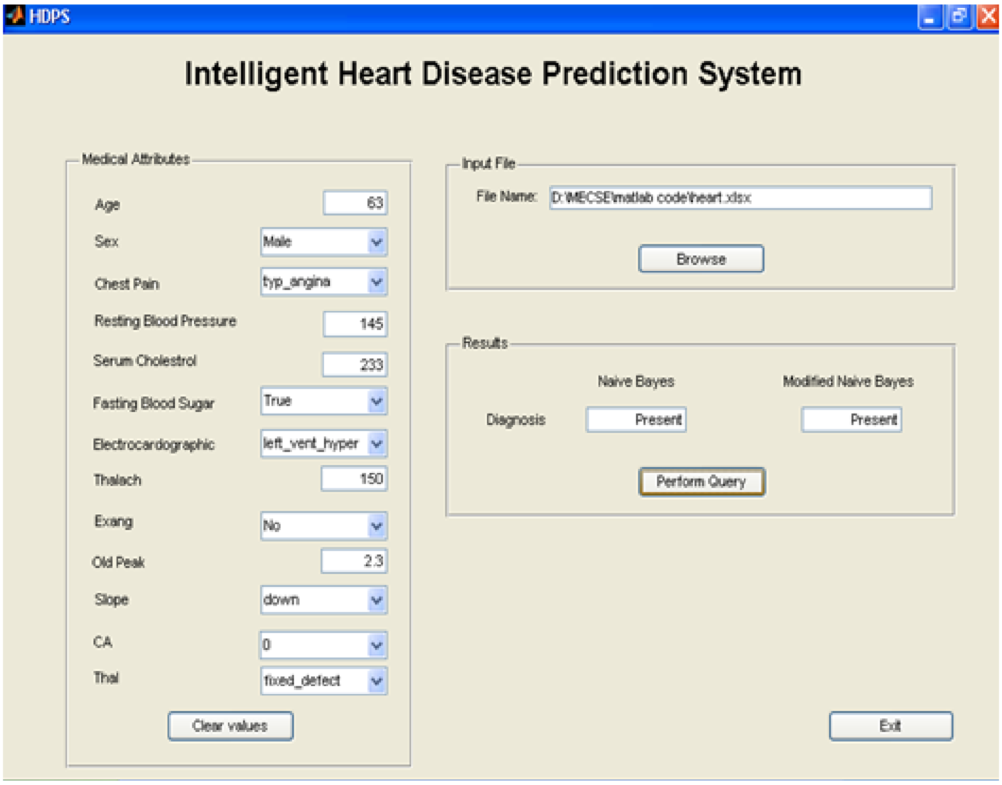
True_Neg is the number of true negatives (i.e.” Absent” samples that were correctly classified) and Neg is the number of negative samples and F_Pos is the number of false positives (“Absent” samples that were incorrectly labelled as “Yes”).

$$\text{Precision} = \frac{True\_Pos}{True\_Pos + F\_Pos}$$

$$\text{Accuracy} = \text{Sensitivity} \frac{Pos}{Pos + Neg} + \text{Specificity} \frac{Neg}{Pos + Neg}$$

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.

4.2 Experiments and Results



Intelligent Heart Disease Prediction System

Medical Attributes

Age	63
Sex	Male
Chest Pain	typ_angina
Resting Blood Pressure	145
Serum Cholestrol	233
Fasting Blood Sugar	True
Electrocardographic	left_vent_hyper
Thalach	150
Exang	No
Old Peak	2.3
Slope	down
CA	0
Thal	fixed_defect

Input File

File Name: D:\MECSE\matlab code\heart.xlsx

Results

	Naive Bayes	Modified Naive Bayes
Diagnosis	Present	Present

Fig. Input and Output Screen

	C1 (1)	C2 (0)
C1 (1)	True Positive	False Negative
C2 (0)	False Positive	True Negative

Fig .Confusion Matrix

Naïve Bayes for Medical Diagnosis

Classification Matrix for Simple Naïve Bayes

	HD Predicted to be Yes	**HD Predicted to be No**
HD Predicted to be Yes	34	3
HD Predicted to be No	4	9

Classification Matrix for prediction using Jelinek-mercer Smoothing

	HD Predicted to be Yes	**HD Predicted to be No**
HD Predicted to be Yes	35	6
HD Predicted to be No	5	4

Table 1 Confusion Matrix obtained from two classifiers

Method	Accuracy	Sensitivity	Specificity	Error Rate
Classification using Naïve Bayes	**78**	**85**	**44**	**22**
Classification using Laplace Smoothing	**86**	**91**	**69**	**14**

Table 2. Comparative Performance of the two classifiers

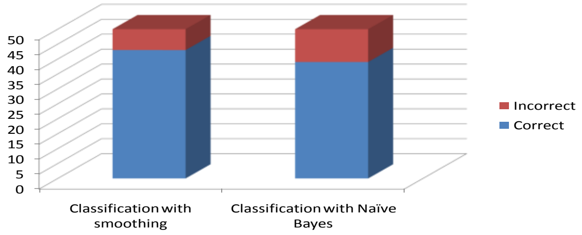


Table 3.Chart of Accuracy for both the classification

5. Conclusion

Decision Support in Heart Disease Prediction System is developed using both Naive Bayesian Classification and Jelinek-mercer smoothing technique. The system extracts hidden knowledge from a historical heart disease database. Jelinek-mercer smoothing technique is the more effective than naive bayes to predict patients with heart disease. This model could answer complex queries, each with its own strength with ease of model interpretation and an easy access to detailed information and accuracy. The system is expandable in the sense that more number of records or attributes can be incorporated and new significant rules can be generated using underlying Data Mining technique. Presently the system has been using 13 attributes of medical diagnosis. It can also incorporate other data mining techniques and additional attributes for prediction.

6. REFERENCES

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

[2] Chapman, P., Clinton, J., Kerber, R. Khabeza, T., Reinartz, T., Shearer, C., Wirth, R.: "CRISP-DM 1.0: Step by step data mining guide", SPSS, 1-78, 2000.

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

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

[5] Ho, T. J.: "Data Mining and Data Warehousing", Prentice Hall, 2005.

[6] Sellappan, P., Chua, S.L.: "Model-based Healthcare Decision Support System", Proc. Of Int. Conf. on Information Technology in Asia CITA'05, 45-50, Kuching, Sarawak, Malaysia, 2005

[7] Quan Yuan,Gao Cong,Nadia M. Thalmann" Enhancing Naive Bayes with Various Smoothing Methods for Short Text Classification "ACM 978-1-4503-1230-1/12/04. WWW 2012 Companion, April 16–20, 2012, Lyon, France.

[8] Shantakumar B.Patil, Y.S.Kumaraswamy, Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network, European Journal of Scientific Research ISSN 1450-216X Vol.31 No.4 (2009), pp.642-656 © EuroJournals Publishing, Inc. 2009.

[9] Wu, R., Peters, W., Morgan, M.W.: "The Next Generation Clinical Decision Support: Linking Evidence to Best Practice", Journal Healthcare Information Management. 16(4), 50-55, 2002.

[10] Clinical Data Mining: a Review J. Iavindrasana, G. Cohen, A. Depeursinge, H. Müller, R. Meyer, A.Geissbuhler University and Hospitals of Geneva, Switzerland