Decision Support in Heart Disease Prediction System using Naive Bayes

Mrs.G.Subbalakshmi (M.Tech),
Kakinada Institute of Engineering & Technology
(Affiliated to JNTU-Kakinada),
Yanam Road, Korangi-533461,
E.G.Dist., A.P., India.

Mr. K. Ramesh M.Tech, Asst. Professor,
KIET, Korangi-533461
E.G.Dist., A.P., India.

Mr. M. Chinna Rao M.Tech,(Ph.D.) Asst. Professor,
KIET, Korangi-533461
E.G.Dist., A.P., India.

Abstract

Data Mining refers to using a variety of techniques to identify suggest of 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 support, predictions, forecasting and estimation. The healthcare industry collects huge amounts of healthcare data which, unfortunately, are not “mined” to discover hidden information for effective decision making. Discovering relations that connect variables in a database is the subject of data mining. This research has developed a Decision Support in Heart Disease Prediction System (DSHDPS) using data mining modeling technique, namely, Naïve Bayes. Using medical profiles such as age, sex, blood pressure and blood sugar it can predict the likelihood of patients getting a heart disease. It is implemented as web based questionnaire application. It can serve a training tool to train nurses and medical students to diagnose patients with heart disease.

Keywords: data mining, decision support, heart disease, Naïve Bayes.

1. Introduction

Data Mining is the nontrivial process of identifying valid, novel, potentially useful and ultimately 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 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 cleaning, data integration, data selection, data mining pattern recognition and knowledge presentation. Data mining is the search for the relationships and global patterns that exist in large databases bur are hidden among large amounts of data.

Many hospital information systems are designed to support patient billing, inventory management and generation of simple statistics. Some hospitals use decision support systems, but are largely limited. They can answer simple queries like “What is the average age of patients who have heart disease?” , “How many surgeries had resulted in hospital stays longer than 10 days?”, “Identify the female patients who are single, above 30 years old, and who have been treated for cancer.” However they cannot answer complex queries like “Given patient records, predict the probability of patients getting a heart disease.” Clinical decisions are often made based on doctors’ intuition and experience rather than on the knowledge rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. The proposed system that integration of clinical decision support with computer-based patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome. This suggestion is promising as data modeling and analysis tools, e.g., data mining, have the potential to generate a knowledge rich environment which can help to significantly improve the quality of clinical decisions.

2. Research objectives

Most hospitals today employ sort of hospital information systems to manage their healthcare or patient data. These systems typically generate huge amounts of data. There is a wealth of hidden information in these data that is largely untapped. How data is turned into useful information that can enable healthcare practitioners to make intelligent clinical decisions. The main objective of this research is to develop a Decision Support in Heart Disease Prediction System (DSHDPS) using one data mining modeling technique, namely, Naïve Bayes. DSHDPS is implemented as web based questionnaire application. Based on user answers, 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 in two ways using chart and pdf which indicates whether that particular patient having the heart disease or not. This suggestion is promising as data modeling and analysis tools, e.g., data mining, have the potential 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 tedious task in medicine. The detection of heart disease from various factors or symptoms is a multi-layered issue which is not free from false presumptions often accompanied by unpredictable effects. Thus the effort to utilize knowledge and experience of numerous specialists and clinical screening data of patients collected in databases to facilitate the diagnosis process is considered a valuable option. Providing precious services at 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 efficient treatment. Poor clinical decisions may lead to disasters and hence are seldom entertained. Besides, it is essential that the hospitals decrease the cost of clinical test. Appropriate 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 capabilities. It provides new ways 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.

3. Data source

Clinical databases have accumulated large quantities of information about patients and their medical conditions. The term Heart disease encompasses the diverse diseases that affect the heart. Heart disease is the major cause of casualties in the world. The term Heart disease encompasses the diverse diseases that affect the heart. Heart disease kills one person every 34 seconds in the United States. Coronary heart disease, Cardiomyopathy and Cardiovascular disease 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 help of the dataset, the patterns significant to the heart attack prediction are extracted. The records were split equally into two datasets: training dataset and testing dataset. To avoid bias, the records for each set were selected randomly.

The attribute "Diagnosis" is identified as the predictable attribute with value "1" for patients with heart disease and value "0" for patients with no heart disease. "PatientId" is used as the key; the rest are input attributes. It is assumed that problems such as missing data, inconsistent data, and duplicate 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))

Key attribute

1. PatientId – Patient's identification number

Input attributes

1. Age in Year
2. Sex (value 1: Male; value 0: Female)
3. Chest Pain Type (value 1:typical type 1 angina, value 2: typical type 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: downsloping)
8. CA – number of major vessels colored by floursopy (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. Oldpeak – ST depression induced by exercise
14. Smoking – (value 1: past; value 2: current; value 3: never)
15. Obesity – (value 1: yes; value 0: no)

4. Implementation of Bayesian Classification

The Naïve Bayes Classifier technique is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods. Naïve Bayes model identifies the characteristics of patients with heart disease. It shows the probability of each input attribute for the predictable state.

4.1. 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 preferred naive bayes implementation:

1) When the data is high.

2) When the attributes are independent of each other.

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

4.2. 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

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

4. 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 tuple is represented by an n-dimensional attribute vector, X=(x_1, x_2,…, x_n), depicting n measurements made on the tuple from n attributes, respectively, A_1, A2,.., An.

2. Suppose that there are m classes, C1, C2,…, Cm. Given a tuple, 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 $1\leq j \leq m$, $j \neq i$

   Thus we maximize P(Ci|X). The class Ci for which P(Ci|X) is maximized is called the maximum posteriori 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|C_i) P (C_i) 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(C_1)=P(C_2) =…=P(C_m), and we would therefore maximize P(X|C_i). Otherwise, we maximize P(X|C_i)P(C_i). Note that the class prior probabilities may be estimated by P(C_i)=|$C_{i,D}$|/|D|, where |$C_{i,D}$| is the number of training tuples of class C_i in D.

4. Given data sets with many attributes, it would be extremely computationally expensive to compute P(X|C_i). In order to reduce computation in evaluating P(X|C_i), 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,

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

=P(x_1|C_i)x P(x_2|C_i)x… P(x_m|C_i).

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

(a) If A_k is categorical, then P(X_k|C_i) is the number of tuples of class C_i in D having the value xk for A_k, divided by |$C_{i,D}$|, the number of tuples of class C_i in D.

(b) If A_k is continuous valued, then we need to do a bit more work, 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

$$g(x,\mu,\sigma)=\frac{1}{\sqrt{2\Pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

So that

$P(x_k|C_i)=g(x_k, \mu_{C_i}, \sigma_{C_i})$

We need to compute μ_{C_i} and σ_{C_i}, which are the mean and standard deviation, of the values of attribute Ak for training tuples of class Ci. We then plug these two quantities into the above equation.

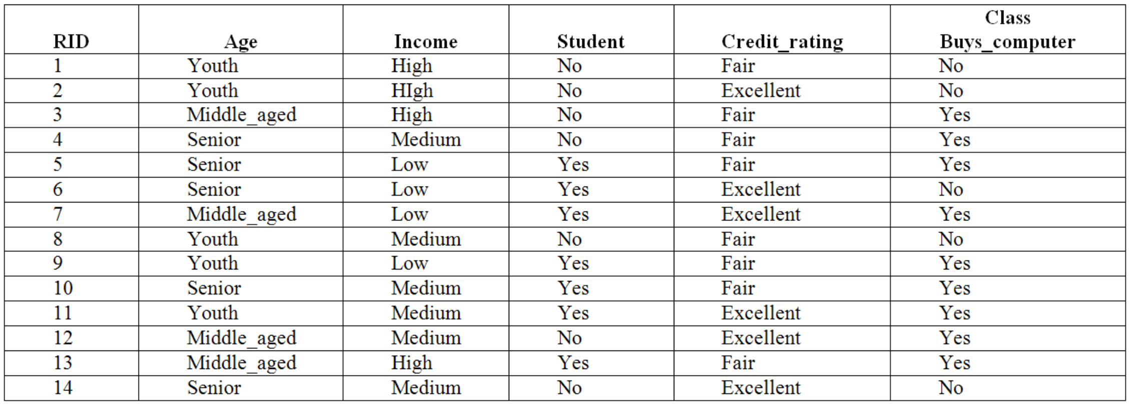
5. In order to predict the class label of X, $P(X|C_i)P(C_i)$ is evaluated for each class C_i. The classifier predicts that the class label of tuple X is the class Ci if and only if

$P(X|C_i)P(C_i)>P(X|C_j)P(C_j)$ for $1 \le j \le m, j \ne i$

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

4.3. Naive Bayesian Classifiers with an example

The following example is a simple demonstration of applying the Naïve Bayes Classifier. This example shows how to calculate the probability using Naïve Bayes classification algorithm.



RID	Age	Income	Student	Credit_rating	Class Buys_computer
1	Youth	High	No	Fair	No
2	Youth	HIgh	No	Excellent	No
3	Middle_aged	High	No	Fair	Yes
4	Senior	Medium	No	Fair	Yes
5	Senior	Low	Yes	Fair	Yes
6	Senior	Low	Yes	Excellent	No
7	Middle_aged	Low	Yes	Excellent	Yes
8	Youth	Medium	No	Fair	No
9	Youth	Low	Yes	Fair	Yes
10	Senior	Medium	Yes	Fair	Yes
11	Youth	Medium	Yes	Excellent	Yes
12	Middle_aged	Medium	No	Excellent	Yes
13	Middle_aged	High	Yes	Fair	Yes
14	Senior	Medium	No	Excellent	No

Table: Class-labeled training tuples from the All Electronics customer database

Predicting a class label using naïve Bayesian classification, we wish to predict the class label of a tuple using naive Bayesian classification from the training data as in the above table. The data tuples are described by the attributes age, income, student and credit rating. The class label attribute, buys_computer, has two distinct values (namely, {yes, no}). Let C_1 correspond to the class buys_computer=yes and C_2 correspond to buys_computer=no. The tuple we wish to classify is

X = (age=youth, income=medium, student=yes, credit_rating=fair)

We need to maximize P(X|Ci)P(Ci), for i=1, 2. P(Ci), the prior probability of each class, can be computed based on the training tuples:

P(buys_computer=yes) = 9/14=0.643

P(buys_computer=no) = 5/14=0.357

To compute P(X|Ci), for i=1, 2, we compute the following conditional probabilities:

P(age=youth|buys_computer=yes) =2/9=0.222

P(age=youth|buys_computer=no) =3/5=0.600

P(income=medium|buys_computer=yes) =4/9=0.444

P(income=medium|buys_computer=no) =2/5=0.400

P(student=yes|buys_computer=yes) =6/9=0.667

P(student=yes|buys_computer=no) =1/5=0.200

P(credit_rating=fair|buys_computer=yes) =6/9=0.667

P(credit_rating=fair|buys_computer=no) =2/5=0.400

Using the above probabilities, we obtain

P(X|buys_computer=yes) = P(age=youth|buys_computer=yes) x
P(income=medium|buys_computer=yes) x
P(student=yes|buys_computer=yes) x
P(credit_rating=fair|buys_computer=yes)
=0.222 x 0.444 x 0.667 x 0.667=0.044

Similarly, P(X|buys_computer=no) = 0.600 x 0.400 x 0.200 x 0.400 = 0.019.

To find the class, Ci, that maximizes P(X|Ci)P(Ci), we compute

P(X|buys_computer=yes) P(buys_computer=yes)=0.044 x 0.643 = 0.028

P(X|buys_computer=no) P(buys_computer=no) =0.019 x 0.357 = 0.007

Therefore, the naïve Bayesian classifier predicts buys_computer = yes for tuple X.

4.4 Implementation on patient data

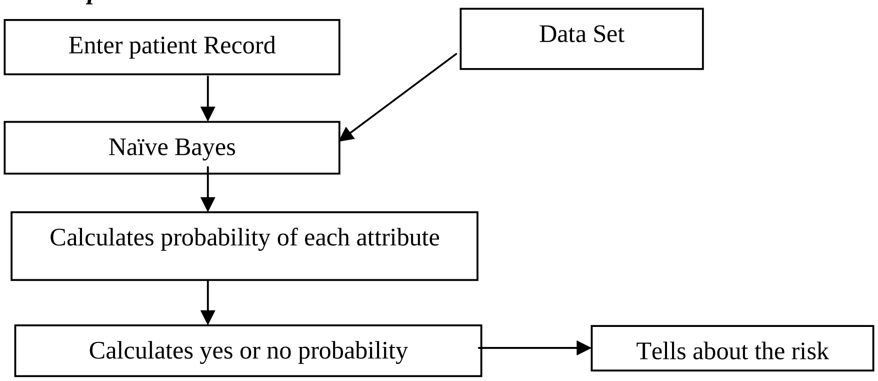

Enter patient Record
Data Set
Naïve Bayes
Calculates probability of each attribute
Calculates yes or no probability
Tells about the risk

Fig. Implementation of Naïve Bayes algorithm on the patient data.

It learns from the “evidence” by calculating the correlation between the target (i.e., dependent) and other (i.e., independent) variables. The Naïve Bayes Classifier technique is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods.

5. Conclusion

Decision Support in Heart Disease Prediction System is developed using Naive Bayesian Classification technique. The system extracts hidden knowledge from a historical heart disease database. This is the most effective model to predict patients with heart disease. This model could answer complex queries, each with its own strength with respect to ease of model interpretation, access to detailed information and accuracy. DSHDPS can be further enhanced and expanded. For, example it can incorporate other medical attributes besides the above list. It can also incorporate other data mining techniques. Continuous data can be used instead of just categorical data.

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] Tang, Z. H., MacLennan, J.: “Data Mining with SQL Server 2005”, Indianapolis: Wiley, 2005.
[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] Kaur, H., Wasan, S. K.: “Empirical Study on Applications of Data Mining Techniques in Healthcare”, Journal of Computer Science 2(2), 194-200, 2006.