Intelligent Heart Disease Prediction System using Naive Bayes Classifier

Amol P. Patil
amol7044@gmail.com

Ayodhyanadan P. Bhosale
ayodhyanadan@gmail.com

GokulP. Ambre
gokul.ambre91@gmail.com

Sumit P. Pawar
sumit@nevitus.com

Prof. M. S. Chaudhari
mschaudhari20@gmail.com

Department Of Computer Engineering
Sinhgad Institute Of Technology,
(Affiliated to University Of Pune)
Kusgaon(Bk), Lonavala, Pune-410401.

.

Abstract

Now a day each hospital maintains database that collects huge amount of data related to patients in the form of numbers, text, charts and images. Clinical decisions are often made based on doctors' intuition and experience rather than on the knowledge rich data hidden in the database. Variety of data mining techniques can be used to extract 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 of disease. This research has developed a decision support in Heart Disease Prediction System (HDPS) using data mining modeling technique, namely, Naïve Bayes Classifier. 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 standalone application. It can serve a training tool to train nurses and medical students to diagnose patients with heart disease.

Keywords: Data mining, Heart disease, Naïve Bayes classifier, Decision Support System

1. Introduction

Data mining is a process of analyzing data from different perspective and gathering the knowledge from it. The discovered knowledge can be used for different applications for example healthcare industry[4].

Nowadays healthcare industry generates large amount of data about patients, disease diagnosis etc. Data mining provides a set of techniques to discover hidden patterns from 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 [2]. Knowledge Discovery process

consists of an iterative sequence of data cleaning, data integration, data selection, data mining pattern recognition and knowledge presentation. 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 [1].

2. Research Objective

The main objective of this research is to build Intelligent Heart Disease Prediction System (IHDPS) using data mining modeling technique, namely, Naïve Bayes Classifier that gives diagnosis of heart disease using historical heart database.

IHDPS is implemented as 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 file which indicates whether that particular patient having the heart disease or not [5].

3. Data source

Record set with medical attributes is 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 [4].

The Cleveland Heart Disease database (303 Records,13 Attribute)[3] .The thirteen attributes are listed in figure 1 as below

Predictable attribute:

Diagnosis

Value 0: No heart disease

Value 1: Has Heart disease

Key attribute

1. PatientID – Patient's identification number

Input attributes

1. Sex (value 1: Male; value 0 : Female)
2. Chest Pain Type (value 1: typical type 1 angina, value 2: typical type angina, value 3: non-angina pain; value 4:asymptomatic)
3. Fasting Blood Sugar (value 1: > 120 mg/dl; value 0:< 120 mg/dl)
4. Restecg – resting electrographic results (value 0: normal; value 1: 1 having ST-T wave abnormality; value 2: showing probable or definite left ventricular hypertrophy.
5. Exang – exercise induced angina (value 1: yes; value 0: no)
6. Slope – the slope of the peak exercise ST segment(value 1: unsloping; value 2: flat; value 3: downsloping)
7. CA – number of major vessels colored by floursopy(value 0 – 3)
8. Thal (value 3: normal; value 6: fixed defect; value 7: reversible defect)
9. Trest Blood Pressure (mm Hg on admission to the hospital)
10. Serum Cholesterol (mg/dl)
11. Thalach – maximum heart rate achieved
12. Oldpeak – ST depression
13. Age in Year

Fig.1. Attributes list and description

4. Naïve Bayes Classifier Algorithm

Naïve Bayes is a statistical classifier which assumes no dependency between attributes. By theory, this classifier has minimum error rate but it may not be case always.Naïve Bayes model identifies the characteristics of patients with heart disease. It shows the probability of each input attribute for the predictable state.

4.1Naïve 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

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

4.2Naïve Bayes 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=(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 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|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 expensive to compute P(X|Ci). In order 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,

$P(X|Ci) = \prod_{k=1}^{n} P_{(xk|Ci)}$ =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. For each attribute, we look at 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.

(b) If Ak 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

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

$$P(xk|Ci)=g(xk, \mu ci, \sigma ci)$$

We need to compute μci and σci, 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|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 $1 \le j \le m, j \ne i$

In other words, the predicted class label is the class Ci for which P(X|Ci)P(Ci) is the maximum [4].

5. Implementation

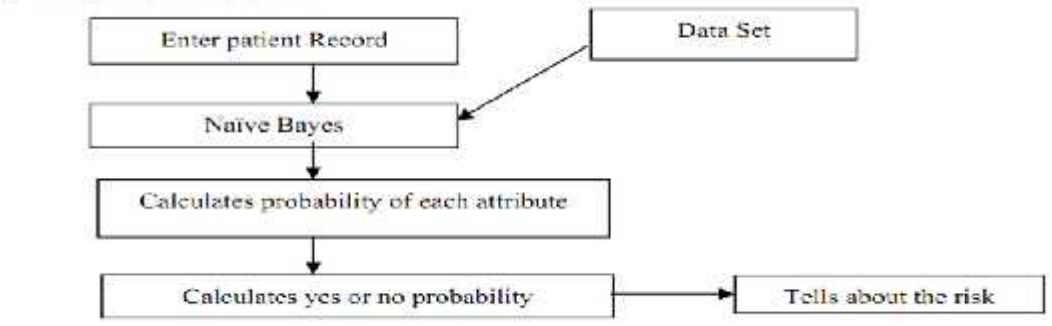

Fig.2. Implementation of Naïve Bayes algorithm on the patient data.

The Naïve Bayes Classifier technique is particularly suited when the dimensionality of the inputs is high. Naïve Bayes model identifies the characteristics of patients with heart disease.

6. Conclusion

A prototype heart disease prediction system is developed using the data mining classification modeling technique, namely, Naïve Bayes Classification. The system extracts hidden knowledge from a historical heart disease database. This model could answer complex queries, each with its own strength with respect to ease of model interpretation, access to detailed information and accuracy.

7. References

[1] Kaur, H., Wasan, S. K.: "Empirical Study on Applications of Data Mining Techniques in Healthcare", Journal ofComputer Science 2(2), 194-200, 2006. Enter patient Record Data Set Naïve Bayes Calculates probability of each attribute Calculates yes or no probability Tells about the risk G.Subbalakshmi et al. / Indian Journal of Computer Science and Engineering (IJCSE).2006

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

[3]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.

[4]Decision Support in Heart Disease Prediction System using Naïve Bayes: Mrs.G.Subbalaxmi, Mr.M.K.Ramesh Asst.Professor, Mr.Chinna Rao Asst. Professor,KIET Korangi-533461.E.G.Dist.,A.P.,India 2011

[5]Prediction System For Heart Disease Using Naïve Bayes:Shadab Adam Pattekari and Asma Parveen, Dept. of Computer Science and Engg. Khaja Banda Nawaz College of Engg. Rouza Buzurg, Gulbarga-585 104,Karnataka,India.2012.

