Prediction Heart Disease Using Naive Bayesian Classification

Khin Myo Aye, Yuzana
Computer University (Pyay)
khinmyoaye2009@gmail.com,yuzana.yzn@gmail.com

Abstract

Naive Bayes is one of the most efficient and effective inductive learning algorithms for machine learning and data mining. Its competitive performance in classification is surprising, because the conditional independence assumption on which it is based, rarely true in real-world applications. The healthcare industry collects huge amounts of healthcare data which, unfortunately, are not mined to discover of hidden information for effective decision making. Advanced data mining techniques can help medication. In this system, we developed a prototype that is Prediction Heart Disease Using Naive Bayesian Classification. We exploited medical profiles such as age, gender, blood pressure and blood sugar , it can predict the likelihood of patients getting a heart disease. This system is computer-based, user-friendly interface and the accuracy are reliable and expandable Moreover, we tested the train data of 326 and test data of 177 records and measured the performance with sensitivity and specificity So, the experimental result shows that the accuracy got 91.21%.

Keywords: Naive Bayesian Classifier, Diagnosis of Heart Disease, Probability, Accuracy, Holdout Method

1. Introduction

The healthcare environment is generally perceived as being information rich yet knowledge poor. There is a wealth of data available within the healthcare systems. However, there is a need of effective analysis tools to discover hidden relationships and trends in data. Medical diagnosis is regarded as an important yet complicated task that needs to be executed accurately and efficiently.

Numerous fields associated with medical services like prediction in effectiveness of surgical procedures, medical tests, medication, and the discovery of relationships among clinical and diagnosis data as well employ Data Mining methodologies. [7] Data mining has been defined as the nontrivial extraction of implicit, previously unknown, and potentially useful information from data [2]. In general, data mining tasks is classified into two broad categories: descriptive mining and predictive mining. Descriptive mining tasks characterize the general properties of the data in the database. Predictive mining tasks perform inference on the current data in order to make predictions.[3]

By applying data mining technique in medical organization, it can minimize the clinical cost. On the other hand, Naive Bayesian Classifiers have also exhibited high accuracy and speed when applied to large database. In health care organisation, treatment records of millions of patients can store. Data mining techniques may help in answering critical questions related to health care.

The purpose of this system is to assist the healthcare organisations, analyze the patient's data by using data mining technique, predict future events by using the knowledge of prior events and obtain good results in most of cases.

The main objective of prediction mining is to assign new data items into one of the few predefined categorical classes. [8]As classification is the most studied data mining and knowledge discovery task [5], there are many classification algorithms. In this paper, we present prediction heart disease using Naive Bayesian Classification. The classification algorithms are applied to our heart disease dataset comparisons of predictive accuracies will be performed.

However, there can be a concern of patient privacy. It is more than clear that the role of data mining is not to practice medicine, but to improve useful information and knowledge so that better treatment and health care provided. In this paper, describes the Related Work in section 2, Naive Bayesian Classifier in section 3, Proposed System Design in section 4, Estimating Classifier Accuracy in section 5, Implementation of the system in section 6 and Conclusion in section 7.

2. Related Work

Numerous works in literature related with heart disease diagnosis using data mining techniques have motivated our work. Naive Bayesian classifiers have proved to be powerful tools for solving classification problems in a variety of domains. There are many practical situations in which classification is of

immense use. Naive Bayesian classifiers have used for providing a diagnosis for a medical patient based on a set of test results.

Providing precious services at inexpensive costs is a major constraint encountered by the healthcare organization (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 tests at a reduced cost. [6]

A model Intelligent Heart Diseases Prediction System (IHDPS) built with the aid of data mining techniques like Decision Trees, Naive Bayes and Neural Network was proposed by Sellappan Palaniappan et al. [6] The results illustrated the peculiar strength of each of the methodologies in comprehending the objectives of the specified mining objectives. IHDPS was capable of answering queries that the conventional decision support systems were not able to. It facilitated the establishment of vital knowledge, e.g. patterns, relationships among medical factors connected with heart disease. IHDPS subsists well being, user friendly, scalable, reliable and expandable.

By using Bayesian posteriori probability classifier, it can be implemented in interpreting paddy distributions of three counties on Northern Taiwan during two crop seasons on year 2000 using multitemporal imageries together with cadastre GIS. In order to integrating Bayesian conditional probability, priori probabilities of paddy's attributes were estimated from photogrammettric interpretation results provided by the Food Bureau, and the spectrum reflectance from different growth stages was used. [1]

3. Naive Bayesian Classifier

Naive Bayesian classifier is a straightforward and frequently used method for supervised learning. It provides a flexible way for dealing with any number of attributes or classes, and is based on applying Bayes' theorem with strong (naive) independence assumptions. It is the asymptotically fastest learning algorithm that examines all its training input. It has been demonstrated to perform surprisingly well in a very wide variety of problems in spite of the simplistic nature of the model.

To be comparable in performance with decision tree and neural network classifier, a simple Bayesian classifier known as the Navie Bayesian classifiers have also exhibited high accuracy and speed when applied to large databases. Naïve Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence.

3.1. Step of Naive Bayesian Classification

There are five steps of naïve Bayesian classification. They are:

1. Each data sample is represented by an *n* dimensional feature vector, $X=(x_1, x_2, x_3,\ldots, x_n)$, depicting n measurements made on the sample from *n* attributes, respectively, $A_1, A_2, \ldots, A_n$.

2. Suppose that there are *m* classes, $C_1, C_2, \ldots, C_m$. The Naive Bayesian classifier assigns an unknown sample X to the class C_i if and only if

$$P(C_i \mid X) > P(C_j \mid X) \text{ for } 1 \le j \le m, j \ne i.$$

Thus, we maximize $P(C_i \mid X)$. The class C_i for which $P(C_i \mid X)$ is maximized is called the maximum posterior hypothesis. Bayes theorem,

$$P(C_i \mid X) = \frac{P(X \mid C_i)P(C_i)}{P(X)} \quad \textbf{(eq-1)}$$

3. The class prior probabilities may be estimated by

$$P(C_i) = s_i / s \quad \textbf{(eq-1.1)}$$

where s_i is the number of training samples of class C_i, and s is the total number of training samples.

4. Give data sets with many attributes. The naïve assumption of class conditional independence, that is, there are no dependence relationships among the attributes. Thus,

$$P(X \mid C_i) = \prod_{k=1}^{n} P(x_k \mid C_i) \quad \textbf{(eq-1.2)}$$

The probabilities $P(x_1|C_i),\ldots, P(x_n|C_i)$ can be estimated from the training samples, where

If A_k is categorical, then $P(X_k \mid C_i) = s_{ik} / s$, where s_{ik} is the number of training samples of class C_i having the values x_k for A_k, and s_i is the number of training samples belonging to C_i.

5. In order to classify an unknown sample X, $P(X|C_i)P(C_i)$ is evaluated for each class C_i. Sample X is then assigned to the class C_i if and only if

$P(X|C_i)P(C_i) > P(X|C_j)P(C_j)$ for $1 \le j \le m, j \ne i$. **(eq-1.3)**

In other words, it is assigned to the class C_i for which $P(X|C_i)P(C_i)$ is the maximum.

4. Proposed System Design

This system is to implement the prediction heart disease system based on Naive Bayesian classification method. It provides a sample prediction model to get secure data between given data and training data according to the user input. This system predict whether a person has heart disease or not by the input user. If the user input secure data, the system can give a completed result for heart disease prediction.

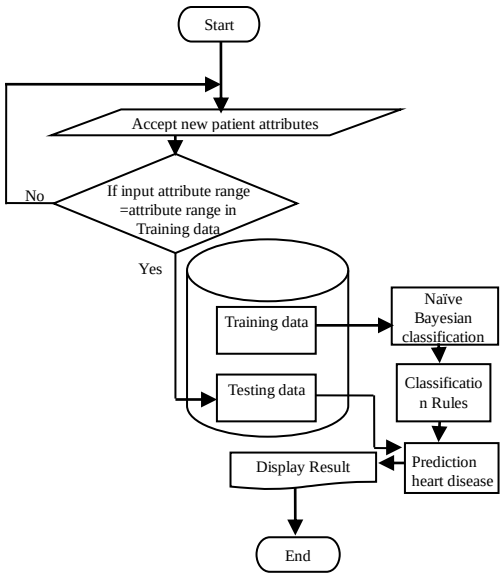

Figure 1. System Flow Diagram

For a good model, this system uses 503 records with 14 medical attributes and produces their relative classification rules according to the Naive Bayesian method. This system measures the performance with sensitivity and specificity. The experimental results show that the accuracy got 91.21%. The attribute diagnosis was identified as the predictable attribute with value "1" for patients with heart disease and value "0" for patients with non heart disease. The attribute "PatientID" was 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. Then the new data given by the user is tested and trained with the prepared classification rules.

The input data is fundamentally to get the secure prediction results. This system also provides accuracy rate, sensitivity and specificity performance of input data. That is why, this system is a good prototype for completed prediction system.

4.1. Attribute Information

Before training and classifying, a number of pre-processing decisions had to be made. This system used the dataset from Cleveland Heart Disease database taken from UCI (University of California. Irvine) Machine Learning Repository. This site describes the contents of the heart disease dataset. That database contains 76 raw attributes, but all published experiments refer to using a subset of 14 of them. [9]

The two classes are heart disease and non heart disease. There are 230 heart disease records and 273 non heart disease records.

Table 1. Name and Description of Attribute

No	Attribute Name	Description
1.	Gender	1=Male, 0=Female
2.	Age	Age in years
3.	Chest pain type	0=typical angina, 1=atypical angina, 2=non angina pain, 3=asymptomatic
4.	Blood Pressure	Resting blood pressure (90-119:normal, 120-139: prehypertension, 140-159: hypertension[satge-1]high, >160:hypertension[stage-2] high)
5.	Cholesterol	Serum cholesterol in mg/dl<200=normal, 200-239=borderline-high, >=240=high
6.	Blood Sugar	Fasting blood sugar >120 mg/dl : 1=true, 0=false
7.	ECG	Resting electrocardiographic result: 0=normal, 1=having ST-T wave abnormality, 2=showing probable
8.	Heart Rate	Maximum heart rate achieved: 50-100=normal, >100=high
9.	Angina	Exercise included angina:1=yes, 0=no
10.	Old Peak	ST depression included by exercise relative to rest: 0.6<=normal, >0.6= abnormal
11.	Slope	The slope of the peak exercise ST segment: 0=flat, 1=up sloping, -1= down slopping
12.	Vessel Count	No: of major vessels(0-3) colored by flourosopy
13.	Thal	0=normal, 1=fixed defect, -1=reversible defect
14.	Heart Disease	0=absence, 1=presence

4.2. NB model for Heart Disease Prediction

Consider the problem of prediction for heart disease by their attributes. The data samples are described by the attributes age, blood pressure, and blood sugar etc. The class label attribute is heart disease or non heart disease.

Let X = (age="50-60", Blood pressure=">=160", etc.)

The prior probability of each class, $P(C_i)$, can be computed based on the training data. According to eq-1.1 in section 3.1;

$$P(C_i) = \frac{s_i}{s}$$

Where s_i is the number of training data of class heart disease or not and s is the total number of training data. In Prediction Heart Disease System,

$$P(\text{heart disease="Yes"}) = \frac{230}{503} = 0.457$$

$$P(\text{heart disease="No"}) = \frac{273}{503} = 0.542$$

Given data sets with 14 attributes, it would be extremely computationally expensive to compute $P(X\backslash C_i)$. In order to reduce computation in evaluating $P(X\backslash C_i)$, the naive assumption of class conditional independence is made. According to eq-1.2 in section 3.1;

$$P(X\backslash C_i) = \prod_{k=1}^{n} P(x_k\backslash C_i)$$

where, $X = (A_1,\ldots,A_n)$ = 14 attributes
C_i = heart disease or non heart disease

The probabilities $P(x_1\backslash C_i)$, $P(x_2\backslash C_i)$, … , $P(x_n\backslash C_i)$ can be estimated from the training data set.

$$P(x_k\backslash C_i) = \frac{s_{ik}}{s_i}$$

where s_{ik} is the number of training samples of class C_i having the value x_k for A_k, and s_i is the number of training samples belonging to C_i. In this system,

$$P(\text{Age="50-60"} \mid \text{heart disease="Yes"}) = \frac{99}{230} = 0.43$$

$$P(\text{Age="50-60"} \mid \text{heart disease="No"}) = \frac{110}{273} = 0.40$$

$$P(\text{Blood pressure=">=160"} \mid \text{heart disease="Yes"}) = \frac{45}{230} = 0.195$$

$$P(\text{Blood pressure=">=160"} \mid \text{heart disease="No"}) = \frac{15}{273} = 0.054$$

P(X | heart disease="Yes")=0.43*0.195=0.08385
P(X| heart disease="No")=0.402*0.054=0.021708

In order to classify an unknown sample X, $P(X|C_i)P(C_i)$ is evaluated for each class C_i**.** In this system, if, According to eq-1.3 in section 3.1

P(X | heart disease="Yes") P(heart disease="Yes") =0.08385*0.457=0.03831
>P(X | heart disease="No")= P(heart disease="No") =0.021708*0.542=0.01176

then, data Simple X is defined by the system as heart disease.

5. Estimating Classifier Accuracy

Training and testing the data mining model requires the data to be split into two groups: one for model training (i.e., estimation of the model parameters) and one for model testing. If it doesn't use different training and test data, the accuracy of the model will be overestimated. [4] Estimating classifier accuracy is important in that it allows one to evaluate how accurately a given classifier will label future data, that is, data on which the classifier has not been trained. And accuracy estimates also help in the comparison of different classifiers.

After the model is generated using the training database, it is used to predict the test database, and the resulting accuracy rate is a good estimate of how the model will perform on future databases that are similar to the training and test databases. Holdout and Cross-Validation are two common techniques for assessing classifier accuracy, based on randomly sampled partitions of the given data.

5.1. Holdout Method

In the holdout method, the given data are randomly partitioned into two independent sets, a training seat and a test set. Typically, two thirds of the data are allocated to the training set, and the remaining one third is allocated to the test set. The training set is used to derive the classifier, whose accuracy is estimated with the test set. The estimate is pessimistic since only a portion of the initial data is used to derive the classifier. Random sub sampling is a variation of the holdout method in which the holdout method is repeated *k* times. The overall accuracy estimate is taken as the average of the accuracies obtained from each iteration. [3]

5.2. Experimental Result

In this system, the total 503 records with 14 medical attributes are used. According to holdout method, 326 records are used for training and 177 records are used for testing. The results of the experiments are:

Table 2. Experimental Result

Method	Sensitivity	Specificity	Accuracy
Naive Bayesian Classification	84.89%	97.75%	91.21%

where, t_pos is the number of true positives (“heart disease” samples that were correctly classified as such), pos is the number of positive (“heart disease”) samples, t_neg is the number of true negatives(“non heart disease” samples that were correctly classified as such), neg is the number of negative (“non heart disease”) samples, and f_pos is the number of false positives(“non heart disease” samples that were incorrectly labeled as “heart disease”). It can be shown that accuracy is a function of sensitivity and specificity:

Sensitivity = t_pos / pos
Specificity =t_neg / neg
Precision =t_pos / (t_pos + f_pos)

Accuracy = Sensitivity*(pos / (pos + neg)) + Specificity*(neg/(pos + neg))

6. Implementation of the system

This system implements the prediction of heart disease or not using Naive Bayesian Classification.

Figure 2. shows the training dataset of the system. These training data records are used Bayesian Classification to get the rules. There are 14 attributes and one result field. Each attribute has its own different value. For example, attribute (chest pain type) has attribute values such as typical angina, atypical angina, non-angina pain and asymptomatic.

Figure 3. shows training probability values generated from the Bayesian analysis for the system. We use this Bayesian analysis as a model to estimate the accuracy of the system and to diagnosis the new patient. This result support for heart disease prediction is obtained by using 14 attributes with different attribute values and two classes (heart disease or non heart disease) result.

Figure 3. shows the testing data for the system. These are used to test the system's accuracy after applying the rules.

The user can test by entering the input data (such as age, blood pressure, chest pain type, etc.,) in the entry form for new user. After entry input data then this system predict heart disease or not for the new user.

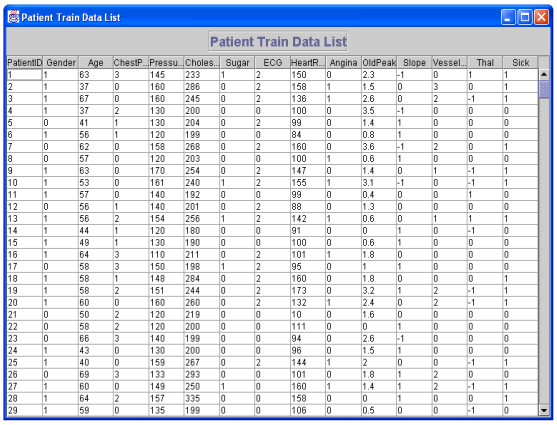

Figure 2. Training Dataset Form

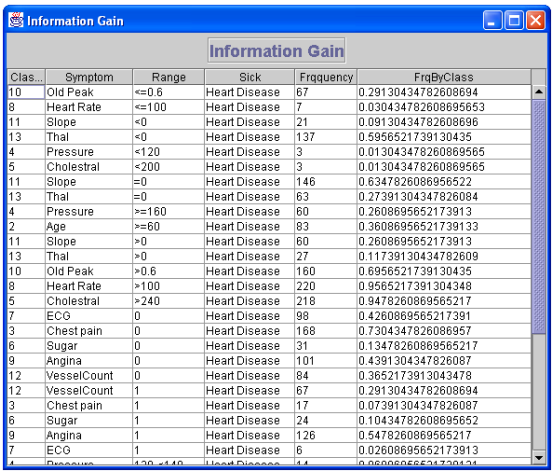

Figure 3. Training Probability Values Form

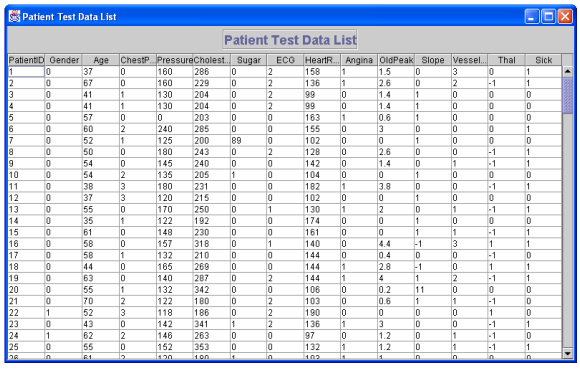

Figure 4. Testing Dataset Form

7. Conclusion

In this paper, the classifications based on patient' data with heart disease or not have been studied. This system support users in classifying heart disease diagnosis based on the symptoms of the patients.

This system will give user prediction into the most approximate outcome from the probability value of Naive Bayesian classification based on the attributes. The attribute "Diagnosis" was identified as the predictable attribute with value "1" for patients with heart disease and value "0" for patients with non heart disease.

From the above test one can find that the Bayesian decision method presents a high accuracy with a simple computation procedure. Naive Bayesian classifier may be the most commonly used because of its easy implementation and good results obtained in most cases. It is possible to predict the efficiency of medical treatments by building the data mining applications. This characteristic makes Bayesian classifier can easy be applied to health check region. Further experimentation is essential, and complementary experiments need to be carried out.

8. References

[1] Chi-Chung L., Kuo-Hsin H., “Bayesian Classification for Rice Paddy Interpretation”, Researcher, Energy and Resource Laboratories, Industrial Technology Research Institute.

[2] Frawley and Piatetsky-Shapiro, "Knowledge Discovery in Databases: An Overview". The AAAI/MIT Press, Menlo Park, C.A, 1996.

[3] Han .J., Kamber M. "Data Mining Concepts and Techniques", Academic Press, USA,2001.

[4] "Introduction to Data Mining and Knowledge Discovery", Two Crows Corporation

[5] Michie D., Spiegelhalter D., and Taylor C., "Machine Learning, Neural and Statistical Classification" New York: Ellis Horwood, 1994.

[6] Palaniappan S., Awang R., "Intelligent Heart Disease Prediction System Using Data Mining Techniques", IJCSNS International Journal of Computer Science and Network Security, Vol.8 No.8, August 2008

[7] Tzung-I Tang, Gang Zheng, Yalou Huang, Guangfu Shu, Pengtao Wang, "A comparative Study of Medical Data Classification Method Based on Decision Tree and System Reconstruction Analysis" , IEMS,Vol. 4, No. 1, pp. 102-108, June 2005.

[8] Weiss S. M. and Kulikowski C. A., "Computer Systems that Learn: Classification and Prediction Methods from Statistical, Neural Nets, Machine Learning, and Expert Systems" San Francisco: Morgan Kaufman, 1991.

[9] http://archive.ics.uci.edu/ml/datasets/Heart+Disease