International Journal of Advanced Computer and Mathematical Sciences
ISSN 2230-9624. Vol 3, Issue 3, 2012, pp 290-294
http://bipublication.com

PREDICTION SYSTEM FOR HEART DISEASE USING NAIVE BAYES

***Shadab Adam Pattekari and Asma Parveen**

Department of Computer Science and Enginering Khaja Banda Nawaz College of Engineering,
Rouza Buzurg, Gulbarga-585 104, Karnataka, India.
*shadabpattekari@gmail.com, asma.aleem4a@yahoo.com

[Received-15/05/2012, Accepted-12/06/2012]

ABSTRACT

The main objective of this research is to develop an Intelligent System using data mining modeling technique, namely, Naive Bayes. It is implemented as web based application in this user answers the predefined questions. It retrieves hidden data from stored database and compares the user values with trained data set. It can answer complex queries for diagnosing heart disease and thus assist healthcare practitioners to make intelligent clinical decisions which traditional decision support systems cannot. By providing effective treatments, it also helps to reduce treatment costs.

Keyword: Data mining Naive bayes, heart disease, prediction

I. INTRODUCTION

In this fast moving world people want to live a very luxurious life so they work like a machine in order to earn lot of money and live a comfortable life therefore in this race they forget to take care of themselves, because of this there food habits change their entire lifestyle change, in this type of lifestyle they are more tensed they have blood pressure, sugar at a very young age and they don't give enough rest for themselves and eat what they get and they even don't bother about the quality of the food if sick the go for their own medication as a result of all these small negligence it leads to a major threat that is the heart disease. It is a world known fact that heart is the most [6] essential organ in human body if that organ gets affected then it also affects the other vital parts of the body. Therefore it is very important for people to go for a heart disease diagnosis [1].

As a result of this people go to healthcare practitioners but the prediction made by them is not 100% accurate [3]. Quality service implies diagnosing patients correctly and administering treatments that are effective. Poor clinical decisions can lead to disastrous consequences which are therefore unacceptable. Hospitals must also

minimize the cost of clinical tests. They can achieve these results by employing appropriate computer-based information and/or decision support systems [5]. The healthcare industry collects huge amounts of healthcare data which, unfortunately, are not "mined" to discover hidden information for effective decision making. Discovery of hidden patterns and relationships often goes unexploited. Advanced data mining techniques can help remedy this situation. This research has developed a prototype Heart Disease Prediction System (HDPS) using data mining techniques, namely, Decision Trees, Naïve Bayes and Neural Network. Results show that each technique has its unique strength in realizing the objectives of the defined mining goals [10]. HDPS can answer complex what if queries which traditional decision support systems cannot. Using medical profiles such as age, sex, blood pressure and blood sugar it can predict the likelihood of patients getting a heart disease. It enables significant knowledge, e.g. patterns, relationships between medical factors related to heart disease, to be established. HDPS is Web-based, user-friendly, scalable, reliable and expandable.

II. RESEARCH OBJECTIVE

The main objective of this research is to develop a prototype Health Care Prediction System using, Naive Bayes .The System can discover and extract hidden knowledge associated with diseases (heart attack, cancer and diabetes) from a historical heart disease database. It can answer complex queries for diagnosing disease and thus assist healthcare practitioners to make intelligent clinical decisions which traditional decision support systems cannot. By providing effective treatments, it also helps to reduce treatment costs. To enhance visualization and ease of interpretation, it displays the results in tabular and PDF forms.

III. SCOPE OF THE PROJECT

Here the scope of the project is 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 [9]. 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 main objective of this research is to develop a prototype Heart Disease Prediction System (HDPS) using three data mining modeling techniques, namely, Decision Trees, Naïve Bayes and Neural Network [8]. So it provides effective treatments, it also helps to reduce treatment costs and also enhances visualization and ease of interpretation.

With immense knowledge and accurate data in that field. Large corporations invest heavily in this kind of activity to help focus attention on possible events and risks that are involved. Such work brings together all available past and current data, as a basis on which to develop reasonable expectations about the future.

IV. DATA SOURCES

Questionnaires have advantages over some other types of medical symptoms that they are cheap, do not require as much effort from the questioner as verbal or telephone surveys, and often have standardized answers that make it simple to compile data [8]. However, such standardized answers may frustrate users. Questionnaires are also sharply limited by the fact that respondents must be able to read the questions and respond to them. Here our questionnaire is based on the attribute given in the data set, so the questionnaire contains:

A. Input attributes

Sr.No	Attribute	Description
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 induced by exercise relative to rest
13	Age	In year
14	Height	In cms
15	Weight	In kgs

Table 1 Data set description

V. IMPLEMENTATION OF NAÏVE BAYES CLASSIFIER

A. Classifier

A classifier is a process of mapping from a (discrete or continuous) feature space *X* to a discrete set of labels *Y*. Here we are dealing about learning classifiers, and learning classifiers are divided into supervised and unsupervised learning classifiers [2]. The applications of classifiers are wide-ranging. They find use in medicine, finance, mobile phones, computer vision (face recognition, target tracking), voice recognition, data mining and uncountable other areas.

An example is a classifier that accepts a person's details, such as age, marital status, home address and medical history and classifies the person with respect to the conditions of the project.

B. Naïve Bayes

In probability theory, Bayes' theorem (often called Bayes' law after Thomas Bayes) relates the conditional and marginal probabilities of two random events. It is often used to compute posterior probabilities given observations [2].

For example, a patient may be observed to have certain symptoms. Bayes' theorem can be used to compute the probability that a proposed diagnosis is correct, given that observation.

A naive Bayes classifier is a term dealing with a simple probabilistic classification based on applying Bayes' theorem.In simple terms, a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 4" in diameter. Even though these features depend on the existence of the other features, a naive Bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple.

Depending on the precise nature of the probability model, naive Bayes classifiers can be trained very efficiently in a supervised learning setting [7]. Naive Bayes classifiers often work much better in many complex real-world situations than one might expect. Here independent variables are considered for the purpose of prediction or occurrence of the event.

In spite of their naive design and apparently over-simplified assumptions, naive Bayes classifiers often work much better in many complex real-world situations than one might expect. Recently, careful analysis of the Bayesian classification problem has shown that there are some theoretical reasons for the apparently unreasonable efficacy of naive Bayes classifiers [4].

An advantage of the naive Bayes classifier is that it requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix [8].

C. Theorem

This is a simple probabilistic classifier based on the Bayes theorem, from the Wikipedia article. This project contains source files that can be included in any C# project.

$$P(W|Q) = \frac{P(Q|W)P(W)}{P(Q)} = \frac{P(Q|W)P(W)}{P(Q|W)P(W) + P(Q|M)P(M)}$$

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}.$$

The Bayesian Classifier is capable of calculating the most probable output depending on the input. It is possible to add new raw data at runtime and have a better probabilistic classifier. A naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 4" in diameter. Even if these features depend on each other or upon the existence of other features, a naive Bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple.

D. Bayesian interpretation

In the Bayesian (or epistemological) interpretation, probability measures a *degree of belief*. Bayes' theorem then links the degree of belief in a proposition before and after accounting for evidence [4]. For example, suppose somebody proposes that a biased coin is twice as likely to land heads as tails. Degree of belief in this might initially be 50%. The coin is then flipped a number of times to collect evidence. Belief may rise to 70% if the evidence supports the proposition [2].

For proposition *A* and evidence *B*,

P(*A*), the *prior*, is the initial degree of belief in *A*.

P(*A* | *B*), the *posterior*, is the degree of belief having accounted for *B*.

P(*B* | *A*) / *P*(*B*) represents the support *B* provides for *A*.

F. Sex classification

Problem: classify whether a given person is a male or a female based on the measured features. The features include height, weight, and foot size.

E. Training

Example training set is shown below.

sex	height (feet)	weight (lbs)	foot size (inches)
male	6	180	12
male	5.92 (5'11")	190	11
male	5.58 (5'7")	170	12
male	5.92 (5'11")	165	10
female	5	100	6
female	5.5 (5'6")	150	8
female	5.42 (5'5")	130	7
female	5.75 (5'9")	150	9

Table2. Data set

The classifier created from the training set using a Gaussian distribution assumption would be:

sex	mean (h)	variance (h)	mean (w)	variance (w)	mean (foot size)	variance (foot size)
male	5.855	3.5033e-02	176.25	1.2292e+02	11.25	9.1667e-01
female	5.4175	9.7225e-02	132.5	5.5833e+02	7.5	1.6667e+00

Table 3. Values calculated by Gaussian distribution

Let's say we have equiprobable classes so P(male)= P(female) = 0.5. There was no identified reason for making this assumption so it may have been a bad idea. If we determine P(C) based on frequency in the training set, we happen to get the same answer.

Below is a sample to be classified as a male or female.

sex	height (feet)	weight (lbs)	foot size (inches)
sample	6	130	8

Table 4. Sample for classification

We wish to determine which posterior is greater, male or female. For the classification as male, the posterior is given by:

$$posterior(male) = \frac{P(male)\,p(height|male)\,p(weight|male)\,p(footsize|male)}{evidence}$$

For the classification as female, the posterior is given by:

$$posterior(female) = \frac{P(female)\,p(height|female)\,p(weight|female)\,p(footsize|female)}{evidence}$$

The evidence may be ignored since it is a positive constant. (Normal distributions are always positive.) We now determine the sex of the sample.

P (male) = 0.5

$$p(height|male) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(6-\mu)^2}{2\sigma^2}\right) \approx 1.5789$$

Where $\mu = 5.855$ and $\sigma^2 = 3.5033e-02$ are the parameters of normal distribution which have been previously determined from the training set. Note that a value greater than 1 is OK here – it is a probability density rather the probability, because height is a continuous variable.

P(weight | male) = 5.9881e-06

P(foot size | male) = 1.3112e-3

Posterior numerator (male) = their product = 6.1984e-09

P(female) = 0.5
p(height | female) = 2.2346e-1
p(weight | female) = 1.6789e-2
p(foot size | female) = 2.8669e-1
Posterior numerator (female) = their product = 5.3778e-04
Since posterior numerator is greater in the female case, we predict the sample is female.

VI. ISSUES AND CHALLENGES

Medical diagnosis is considered as a significant yet intricate task that needs to be carried out precisely and efficiently. The automation of the same would be highly beneficial. Clinical decisions are often made based on doctor's 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. Data mining have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions.

VII. 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. HDPS 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.

HDPS can be further enhanced and expanded. For example, it can incorporate other medical attributes besides the 15 listed in Figure 1. It can also incorporate other data mining techniques, e.g., Time Series, Clustering and Association Rules. Continuous data can also be used instead of just categorical data. Another area is to use Text Mining to mine the vast amount of unstructured data available in healthcare databases. Another challenge would be to integrate data mining and text mining

VIII. 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]. Mrs.G.Subbalakshmi, "Decision Support in Heart Disease Prediction System using Naive Bayes", Indian Journal of Computer Science and Engineering.
[4]. Fayyad, U: "Data Mining and Knowledge Discovery in Databases: Implications for scientific databases", Proc. of the 9th Int. Conf. on Scientific and Statistical Database Management, Olympia, Washington, USA, 2-11, 1997.
[5]. Giudici, P.: "Applied Data Mining: Statistical Methods for Business and Industry", New York: John Wiley, 2003.
[6]. Han, J., Kamber, M.: "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 2006.
[7]. Ho, T. J.: "Data Mining and Data Warehousing", Prentice Hall, 2005.
[8]. Intelligent Heart Disease Prediction System Using Data Mining Techniques-Sellappan Palaniappan, Rafiah Awang 978-1-4244-1968-5/08/ ©2008 IEEE
[9]. Obenshain, M.K: "Application of Data Mining Techniques to Healthcare Data", Infection Control and Hospital Epidemiology, 25(8), 690–695, 2004.
[10]. 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.