Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 7, Issue, 12, pp.24098-24103, December, 2015

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

ISSN: 0975-833X

RESEARCH ARTICLE

PREDICTING THE MISSING VALUE IN A KNOWLEDGE BASED SYSTEM USING BAYESIAN CLASSIFICATION TECHNIQUE

***,[1]Sayak Konar, [2]Md. Abdur Rahaman, [2]Debasrita Roy and [3]Shameek Mukhopadhyay**

[1]Assistant Professor (CSE), BIEMS, India
[2]M.Tech (CSE), VIT University, India
[3]Assistant Professor (BCA), The Heritage Academy, India

ARTICLE INFO

Article History:
Received 16th September, 2015
Received in revised form
24th October, 2015
Accepted 19th November, 2015
Published online 30th December, 2015

Key words:

Predictions, Missing data,
Data mining, Bayesian Classification.

ABSTRACT

When machine learning algorithms are applied to data collected from the huge amount of data in the universe, it is generally accepted that the data has not been consistently collected. The absence of expected data elements is common and the mechanism through which a data element is missing often involves the informative relevance of that data element in a specific purpose. Therefore, the absence of data may have information value of its own. In the process of designing an application intended to support a heart diseases system where we can predict the probability of heart attack of a patient on basis upon certain condition. Bayesian Classification is commonly used for presenting uncertainty and covariate interactions in an easily interpretable way. Because of their efficient inference and ability to predict the missing value in a database, it is an excellent choice for medical decision support systems in diagnosis, treatment, and prognosis. In applying this we will be able to predict whether the data is present in the database or not and give some idea about the probability of heart-attack to the patient.

Citation: **Sayak Konar, Md. Abdur Rahaman, Debasrita Roy and Shameek Mukhopadhyay, 2015.** "Predicting the missing value in a knowledge based system using Bayesian classification technique", *International Journal of Current Research,* 7, (12), 24098-24103.

INTRODUCTION

There is a big amount of data being collected across a wide variety of fields today and it is beyond our ability to reduce and analyze those data without the use of some kind of automated analysis techniques. There is much information hidden in the various fields of data. It is very difficult to obtain this information. So, it is essential for new types of computational techniques and tools to extract the knowledge for the benefit of human from the rapidly growing voluminous digital data. Knowledge discovery in databases (KDD) is the field that is gradually develops into an important and active area of research because there are certain kinds of challenges associated with the problem of discovering decision making solutions from the huge data. Knowledge discovery and data mining is the rapidly growing research field which merges together database management, probability theory, statistics, computational intelligence and related areas. The basic aim of all these is at extracting useful knowledge and information from voluminous data. 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.

***Corresponding author: Sayak Konar,**
Assistant Professor (CSE), BIEMS, India.

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.

Missing data is a common problem in knowledge discovery, data mining and statistical inference. Several approaches to missing data have been used in developing trained decision systems. [Tang and MacLennan, 2005] Little and Rubin (1976, 2002) has studied and categorized missing data into three types: missing completely at random, missing at random, and not missing at random. The easiest way to missing values is to discard the cases with missing values and do the analysis based only on the complete data. However, the absence of association rule and missing data may have information value to predict the decision for our own business or to setup a new business.

In this paper we develop a model where we apply Bayesian Classification technique to predict the probability of heart-attack i.e. based upon some certain condition of a patient we can predict the heart-attack probability.

MATERIALS AND METHODS

Knowledge based Systems are rich with hidden information that can be used for intelligent decision making. Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. Such analysis can help provide us with a better understanding of the data at large. Whereas classification predicts categorical (discrete, unordered) labels, prediction models continuous-valued functions *Han, Kamber* (Ho, 2005).

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class. Bayesian classification is based on Bayes' theorem, described below. Studies comparing classification algorithms have found a simple Bayesian classifier known as the naïve Bayesian classifier to be comparable in performance with decision tree and selected neural network classifiers. Bayesian classifiers have also exhibited high accuracy and speed when applied to large databases.

Bayes' theorem is named after Thomas Bayes, a nonconformist English clergyman who did early work in probability and decision theory during the 18th century. Let X be a data tuple. In Bayesian terms, X is considered "evidence." As usual, it is described by measurements made on a set of n attributes. Let H be some hypothesis, such as that the data tuple X belongs to a specified class C. For classification problems, we want to determine $P(H \mid X)$, the probability that the hypothesis H holds given the "evidence" or observed data tuple X. In other words, we are looking for the probability that tuple X belongs to class C, given that we know the attribute description of X.

$P(H \mid X)$ is the posterior probability, or a posterior probability, of H conditioned on X. For example, suppose our world of data tuples are confined to customers described by the attributes age and income, respectively, and that X is a 35-year-old customer with an income of Rs.40, 000. Suppose that H is the hypothesis that our customer will buy a computer. Then $P(H \mid X)$ reflects the probability that customer X will buy a computer given that we know the customer's age and income.

In contrast, $P(H)$ is the prior probability, or a priori probability, of H. For our example, this is the probability that any given customer will buy a computer, regardless of age, income, or any other information, for that matter. The posterior probability, $P(H \mid X)$, is based on more information (e.g., customer information) than the prior probability, $P(H)$, which is independent of X.

$P(X)$ is the prior probability of X. Using our example, it is the probability that a person from our set of customers is 35 years old and earns Rs.40,000. "How are these probabilities estimated?" $P(H)$, $P(X \mid H)$, and $P(X)$ may be estimated from the given data, as we shall see below. Bayes' theorem is useful in that it provides a way of calculating the posterior probability, $P(H \mid X)$, from $P(H)$, $P(X \mid H)$, and $P(X)$.

Bayes' Theorem is

$$P(H \mid X) = \frac{P(X \mid H)P(H)}{P(X)} \quad \ldots\ldots\ldots\ldots\ldots (1)$$

Now that we've got that out of the way, we will look at how Bayes' theorem is used in the Bayesian classification. The naïve 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, $X = (x_1, x_2, \ldots\ldots, x_n)$, depicting n measurements made on the tuple from n attributes, respectively, $A_1, A_2, A_3, \ldots\ldots\ldots, A_n$.

2. Suppose that there are m classes $C_1, C_2, \ldots\ldots, C_m$. 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

$$P(C_i \mid X) > P(C_j \mid X) \quad \text{for } 1 \leq j \leq m;\ j \neq i$$

Thus we maximize $P(C_i \mid X)$. The class *Ci* for which is $P(C_i \mid X)$ maximized is called the maximum posteriori hypothesis. By Bayes' theorem (Equation (1))

$$P(C_i \mid X) = \frac{P(X \mid C_i)P(C_i)}{P(X)}$$

3. As P(X) is constant for all classes, only $P(X \mid C_i)\ P(C_i)$ need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, $P(C_1) = P(C_2) = \ldots\ldots = P(C_m)$, and we would therefore maximize $P(X \mid C_i)$. Otherwise, we maximize P (X|Ci) P (Ci). 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 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 naive 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,

$P(X|Ci) = \prod_{k=1}^{n} P(Xk|Ci) \quad = P(X1|Ci) \times P(X2|Ci) \times \ldots\ldots \times P(Xn|Ci).$

We can easily estimate the probabilities $P(X_1 \mid C_i), P(X_2 \mid C_i), \ldots\ldots P(X_n \mid C_i)$ from the training tuples. Recall that here X_k refers to the value of attribute A_k for tuple X.

Implementation

Table 1. Knowledge based database: This is the total database on which the missing value is to be predicted.

ID	blood_sugar	blood_pressure	cholestrol	smoking	obesity	age	gender	heart_attack
1	Yes	low	normal	past	Yes	40	male	No
2	Yes	normal	high	current	Yes	35	male	No
3	Yes	high	high	current	No	45	female	Yes
4	Yes	high	low	never	No	38	female	Yes
5	Yes	normal	normal	never	Yes	41	male	No
6	Yes	low	low	current	No	31	male	No
7	Yes	normal	high	past	No	50	male	Yes
8	Yes	normal	normal	past	Yes	55	female	Yes
9	Yes	normal	high	never	No	60	female	Yes
10	Yes	high	low	current	Yes	43	male	Yes
11	Yes	low	normal	past	Yes	37	female	No
12	Yes	low	low	never	No	34	male	No
13	Yes	low	high	current	Yes	65	male	Yes
14	Yes	high	high	past	No	48	female	Yes
15	Yes	high	normal	current	No	30	male	No

Here we develop a heart-attack decision support system based upon certain condition of a patient. In table 1 there are 9 attributes,

i.e. $patient\_id$, $blood\_sugar = \{Yes\}$,
$blood\_pressure = \{high, normal, low\}$,
$cholestrol = \{high, normal, low\}$, $smoking = \{past, current, never\}$,
$obesity = \{Yes, No\}$, $age = \{\leq 40, > 40 and < 50, \geq 50\}$,
$gender = \{male, female\}$, $heart\_attack = \{Yes, No\}$.

In the above table the heart attack depends only one value of blood_sugar attributes i.e. Yes, blood_pressure may be high, low or normal, cholesterol may be high, low or normal, the patient may be a smoker in past, currently or never, the patient can have problem of obesity or not, the age is chosen of the patient in three category i.e. below and equal to 40, greater than 40 but less than 50 and greater than equal to 50. a patient can be male or female and the last one is heart_attack attribute which tells that whether it is possible for a heart attack or not.

We implement the Bayesian Classification technique by using Java code. First we make the above database in Microsoft Access, then we connect the java code to this database through the sun.jdbc.odbc.JdbcOdbcDriver.

Some small snippets of the different classes are shown in Algorithm format down below.

At first we compute the probability of each value of the heart_attack attribute of the above database i.e. $P(C_i)$.

Algorithm

County and countn are the total count of yes/no present in database pcounty and pcountn are the total count of probability present (Yes and No)

Step 1 :- Initialize county,countn and total to 0
Step 2:- Initialize pcounty and pcountn to 0
Step 3:- When class= "Yes" Execute query by selecting all from heart_attack where heart_attack is total class SET
Step 4:- While result set value is present Increment county
Step 5:- When class="No" Execute query by selecting all from heart_attack where heart_attack is total class SET
Step 6:- While result set value is present Increment countn
Step 7:- Total=county+countn;
Step 8:- Pcounty=county/total;
Step 9:- Pcountn=countn/total;

Then each value of the each attribute corresponding to the each value of heart_attack is computed i.e. $P(X \mid C_i)$.(Algorithm for *blood pressure* is shown below).

Algorithm

countyh_bp is for high blood pressure people present in Knowledge based system, countynor_bp is for normal blood pressure and countyl_bp is for low blood pressure

```
Step 1: - If attribute value equals("high")
Increment countyh_bp
        else If temp equals("normal")
            Increment countynor_bp
                else If temp equals("low")
                    Increment countyl_bp
```

Step 2:- pcountyh_bp=countyh_bp/county

Step 3:- pcountynor_bp=countynor_bp/county

Step 4:- pcountyl_bp=countyl_bp/county

After that we calculate every combination that can be possible for the seven attributes i.e.

$blood\_sugar, blood\_pressure, cholestrol, smoking, obesity, age, gender$

Here total combination will be $(1 \times 3 \times 3 \times 3 \times 2 \times 3 \times 2) = 324$.

Algorithm

pgiven_y and pgiven_n are equal to yes and no of probabilities found in database. pcountny_bs is for probability of blood sugar, pcountnh_bp is for probability for blood pressure ,pcountnh_cho is for probability of cholestrol, pcountnpst_smo is for smoking, pcountny is for obesity and so on for the seven attributes

```
If blood_sugar equals("Yes"))
        If  blood_pressure equals("high")
            If cholestrol equals("high")
                If smoking equals("past")
                    If obesity equals("Yes")
                        If age<=40
If gender equals("male")
```

pgiven_y is equal to pcountyy_bs*pcountyh_bp*pcountyh_cho *pcountypst_smo*pcountyy_ob
* pcounty1_age*pcountym_gen

pgiven_n is equal to pcountny_bs*pcountnh_bp *pcountnh_cho*pcountnpst_smo*pcountny_ob* pcountn1_age*pcountnm_gen

Then we take a combination from the total combination – the combination presented in the database already i.e. an unseen combination or missing combination.

At last we calculate the probability that the unseen combination is belonged to which value of the heart_attack attributes i.e. Yes or No.

Algorithm

pgiven_ty is for probability of Yes seen in database while pgiven_tn is for probability of NO seen in database. pcounty and pcountn are probability of counted yes and no present in database respectively

Step 1:- Initialize pgiven_ty,pgiven_tn equal to 0

Step 2: - pgiven_ty is equal to pgiveny*pcounty

Step 3:- pgiven_tn is equal to pgivenn*pcountn

Step 4:- If pgiven_ty > pgiven_tn classn="Yes"

else classn="No"

RESULTS

Let us take an unseen or missing sample i.e. the sample is not presented in the database, that is

$< blood\_sugar = Yes, blood\_pressure = high, cholestrol = normal, smoking = past,$

$obesity = Yes, age = 60, gender = female >$

Now we have to calculate the probability of each sample attribute corresponding to each value of heart_attack attribute.

i.e. for heart_attack attribute

$$P(Yes) = \frac{7}{15}$$

$$P(No) = \frac{8}{15}$$

Now, for $heart\_attack = Yes$

for $blood\_sugar: P(Yes \mid Yes) = \frac{7}{7}$

for $blood\_pressure: P(high \mid Yes) = \frac{4}{7}$

for $cholestrol: P(normal \mid Yes) = \frac{1}{7}$

for $smoking: P(past \mid Yes) = \frac{2}{7}$

for $obesity: P(Yes \mid Yes) = \frac{3}{7}$

for $age: P(60 \mid Yes) = \frac{3}{7}$

for $gender: P(female \mid Yes) = \frac{5}{7}$

Similarly for $heart\_attack = No$

for $blood\_sugar: P(Yes \mid No) = \frac{8}{8}$

for $blood\_pressure: P(high \mid No) = \frac{1}{8}$

for $cholestrol: P(normal \mid No) = \frac{4}{8}$

for $smoking: P(past \mid No) = \frac{3}{8}$

for $obesity: P(Yes \mid No) = \frac{4}{8}$

for $age: P(60 \mid No) = \frac{1}{8}$

for $gender: P(female \mid No) = \frac{1}{8}$

now,

$$P(heart\_attack = Yes \mid blood\_sugar = Yes, blood\_pressure = high, cholestrol = normal, smoking = past, obesity = Yes, age = 60, gender = female)$$

$$= (\frac{7}{7} \times \frac{4}{7} \times \frac{1}{7} \times \frac{2}{7} \times \frac{3}{7} \times \frac{3}{7} \times \frac{5}{7}) \times (\frac{7}{15}) = 0.001427$$

Similarly,

$$P(heart\_attack = No \mid blood\_sugar = Yes, blood\_pressure = high, cholestrol = normal, smoking = past, obesity = Yes, age = 60, gender = female) =$$

$$(\frac{8}{8} \times \frac{1}{8} \times \frac{4}{8} \times \frac{3}{8} \times \frac{4}{8} \times \frac{1}{8} \times \frac{1}{8}) \times (\frac{8}{15}) = 0.00009765$$

Clearly, 0.001427>0.00009765

So, the above unseen sample is belonged to the Yes category of the heart_attack attribute i.e. there is a probability of heart attack if the above unseen condition is satisfied.

Now, let look at snapshots for the above unseen sample after running the code, discussed above.

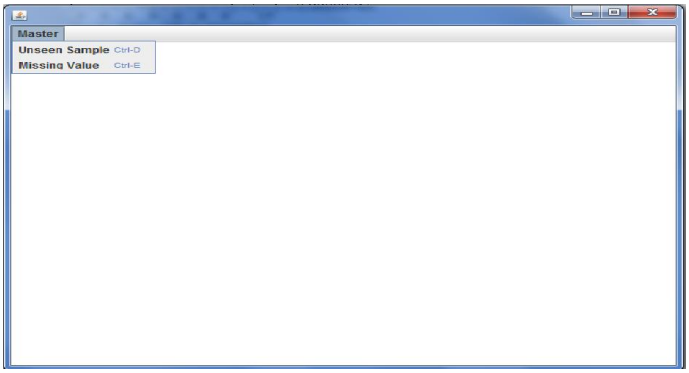

Fig. 1. Master Window:-It depicts the GUI interface that shows how the main window looks like

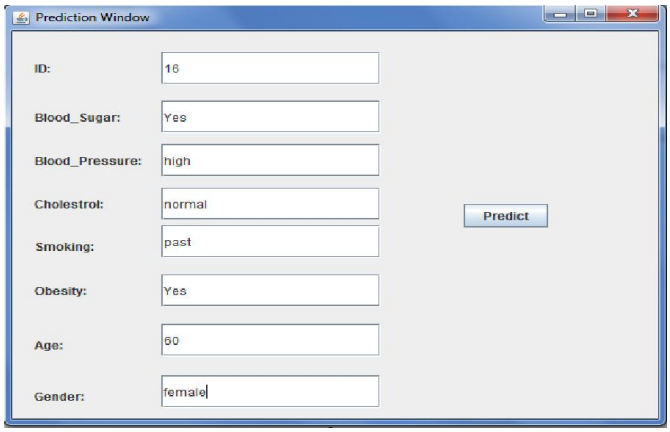

Fig. 2. Updation Form:- It gives us the portal to enter the details of a person

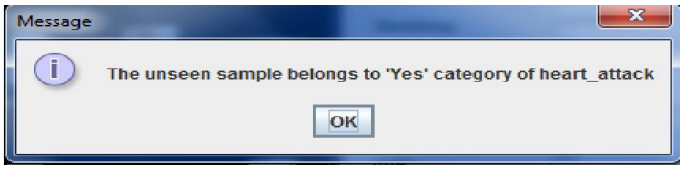

Fig.3. Dialog assertion box:- It gives the confirmation of the status of the patient at risk

The creation of the database :

heart_attack : Table

ID	blood_sugar	blood_pressure	cholestrol	smoking	obesity	age	gender	heart_attack
1	Yes	low	normal	past	Yes	40	male	No
2	Yes	normal	high	current	Yes	35	male	No
3	Yes	high	high	current	No	45	female	Yes
4	Yes	high	low	never	No	38	female	Yes
5	Yes	normal	normal	never	Yes	41	male	No
6	Yes	low	low	current	No	31	male	No
7	Yes	normal	high	past	No	50	male	No
8	Yes	normal	normal	past	Yes	55	female	Yes
9	Yes	normal	high	never	No	60	female	Yes
10	Yes	high	low	current	Yes	43	male	Yes
11	Yes	low	normal	past	Yes	37	female	No
12	Yes	low	low	never	No	34	male	No
13	Yes	low	high	current	Yes	65	male	Yes
14	Yes	high	high	past	No	48	female	Yes
15	Yes	high	normal	current	No	30	male	No
16	Yes	high	normal	past	Yes	60	female	Yes
(AutoNumber)								

Fig.4. Heart Attack decision support system: It gives the whole database of heart attack decision support GUI

The system also predicts the missing value of the attributes. Let us see how this is work:

heart_attack : Table

ID	blood_sugar	blood_pressure	cholestrol	smoking	obesity	age	gender	heart_attack
1	Yes	high	normal	past	Yes	40	male	No
2	Yes	low	normal	never	Yes	35	male	No
3	Yes	low	high	current	No	45	female	Yes
4	Yes	low	low	never	No	40	male	Yes
5	Yes	low	normal	never	Yes	41	male	No
6	Yes	null	normal	current	No	31	male	No
7	Yes	normal	high	never	No	40	male	No
8	Yes	high	normal	never	Yes	55	male	Yes
9	Yes	normal	normal	never	No	60	female	Yes
10	Yes	high	low	current	Yes	40	male	Yes
11	Yes	low	normal	past	Yes	37	female	No
12	Yes	high	normal	never	No	34	male	No
13	Yes	low	high	current	Yes	40	male	Yes
14	Yes	normal	normal	past	No	40	female	Yes
15	Yes	low	low	current	Yes	30	male	No
(AutoNumber)								

Fig.5. The missing attribute: It depicts the missing attribute in the database

We can see from table 3 that in ID=6 there is a missing value in blood_pressure attribute. Now we are going to predict this missing value by Bayesian Classification technique.

For blood_pressure attribute:

$$P(high) = \frac{4}{14}$$

$$P(normal) = \frac{3}{14}$$

$$P(low) = \frac{7}{14}$$

Now, for blood_sugar=Yes

$$P(Yes \mid high) = \frac{4}{4}$$

$$P(Yes \mid normal) = \frac{3}{3}$$

$$P(Yes \mid low) = \frac{7}{7}$$

For cholesterol=normal

$$P(normal \mid high) = \frac{3}{4}$$

$$P(normal \mid normal) = \frac{2}{3}$$

$$P(normal \mid low) = \frac{3}{7}$$

For smoking=current

$$P(current \mid high) = \frac{1}{4}$$

$$P(current \mid normal) = \frac{0}{3}$$

$$P(current \mid low) = \frac{3}{7}$$

For obesity=No

$$P(No \mid high) = \frac{1}{4}$$

$$P(No \mid normal) = \frac{3}{3}$$

$$P(No \mid low) = \frac{2}{7}$$

For age=31 i.e. <=40

$$P(age \le 40 \mid high) = \frac{2}{4}$$

$$P(age \le 40 \mid normal) = \frac{2}{3}$$

$$P(age \le 40 \mid low) = \frac{5}{7}$$

For gender=male

$$P(male \mid high) = \frac{4}{4}$$

$$P(male \mid normal) = \frac{1}{3}$$

$$P(male \mid low) = \frac{5}{7}$$

For heart_attack=No

$$P(No \mid high) = \frac{2}{4}$$

$$P(No \mid normal) = \frac{1}{3}$$

$$P(No \mid low) = \frac{4}{7}$$

Now,

$$P(blood\_pressure = high \mid blood\_sugar = Yes, cholestrol = normal, smoking = current, obesity = No, age \le 40, gender = male, heart\_attck = No) = (\frac{4}{4}\times\frac{3}{4}\times\frac{1}{4}\times\frac{1}{4}\times\frac{2}{4}\times\frac{4}{4}\times\frac{2}{4})\times(\frac{4}{14}) = 0.0034$$

$$P(blood\_pressure = normal \mid blood\_sugar = Yes, cholestrol = normal, smoking = current, obesity = No, age \le 40, gender = male, heart\_attck = No) = (\frac{3}{3}\times\frac{2}{3}\times\frac{0}{3}\times\frac{3}{3}\times\frac{2}{3}\times\frac{1}{3}\times\frac{1}{3})\times(\frac{3}{14}) = 0$$

and

$$P(blood\_pressure = low \mid blood\_sugar = Yes, cholestrol = normal, smoking = current, obesity = No, age \le 40, gender = male, heart\_attck = No) = (\frac{7}{7}\times\frac{3}{7}\times\frac{3}{7}\times\frac{2}{7}\times\frac{5}{7}\times\frac{5}{7}\times\frac{4}{7})\times(\frac{7}{14}) = 0.00764$$

so, easily we can see that 0.0034>0.00764>0

Therefore the missing value will be replaced by 'high' in the blood_pressure attribute.

After running the code the table is:

heart_attack : Table

	ID	blood_sugar	blood_pressure	cholestrol	smoking	obesity	age	gender	heart_attack
	1	Yes	high	normal	past	Yes	40	male	No
	2	Yes	low	normal	never	Yes	35	male	No
	3	Yes	low	high	current	No	45	female	Yes
	4	Yes	low	low	never	No	40	male	Yes
	5	Yes	low	normal	never	Yes	41	male	No
//	6	Yes	high	normal	current	No	31	male	No
	7	Yes	normal	high	never	No	40	male	No
	8	Yes	high	normal	never	Yes	55	male	Yes
	9	Yes	normal	normal	never	No	60	female	Yes
	10	Yes	high	low	current	Yes	40	male	Yes
	11	Yes	low	normal	past	Yes	37	female	No
	12	Yes	high	normal	never	No	34	male	No
	13	Yes	low	high	current	Yes	40	male	Yes
	14	Yes	normal	normal	past	No	40	female	Yes
	15	Yes	low	low	current	Yes	30	male	No
*	(AutoNumber)								

Fig.6. Prediction of the missing value. This table shows the prediction of the missing attribute

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. But applying Bayesian Classification in a large data base the complexity of the code is very much higher. So, we reduce some attributes related to the heart attack for simplicity of the program.

Acknowledgement

We are highly grateful to Prof. Venkatesan M., VIT Vellore, for his support and helpful suggestions in this project. We are also grateful to the esteemed institutions like Vellore Institute of Technology, Techno Group of Institutions and The Heritage Academy for their constant support and encouragement in every respect. Lastly, we are ever grateful to our parents, our better half's and all other well wishers without whose support and good wishes this project would not have seen the light of the earth.

REFERENCES

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

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.

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

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

Kaur, H., Wasan, S. K.: "Empirical Study on Applications of Data Mining Techniques in Healthcare", *Journal of Computer Science*, 2(2), 194-200, 2006.

Little RJA, Rubin DB. 2000. Statistical analysis with missing data, 2nd ed. Wiley-Interscience.

Rubin DB 1976. Inference and missing data. Biometrika; 63:581-92.

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

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.

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 © *Euro Journals Publishing*, Inc. 2009.

Tang, Z. H., MacLennan, J. 2005. "Data Mining with SQL Server 2005", Indianapolis: Wiley, 2005.

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.

****