ISSN No. 0976-5697

Volume 2, No. 5, Sept-Oct 2011

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Intellectual Coronary Thrombosis Prediction using Naïve Bayes and Decision Trees

R.Priyadharsini*
Assistant Professor,
VLB Janakiammal College of Engineering and Technology,
Centre for Computer Applications,
Coimbatore, India.
riya_waves3@yahoo.co.in

P.Krishnakumari
Associate Professor,
Sri Ramakrishna College of Arts and Science for Women,
Centre for Computer Applications,
Coimbatore, India.
kkjagadeesh@yahoo.co.in

Abstract: The Diagnosis of disease is vital and intricate role in medicine. The recognition of heart disease from diverse feature is a multilayered problem that is not free from false assumptions and is frequently accompanied by impulsive effects. A proficient methodology for the extraction of significant patterns from the heart disease warehouse has been presented .Initially the data warehouse is preprocessed in order to make suitable for the mining process. In this research, the potential list of classification and prediction based data mining techniques has been presented. The conditional probability for having heart disease was estimated by Holdout Confusion method and compared with decision tress. The research work demonstrates better accuracy for Naïve Bayes than decision trees. The result of these evaluations shows the overall performance of naïve Bayes method can be applied successfully for predicting the heart attack effectively. The proposed model is implemented on the C Sharp dot net platform. In future, the work can be further expanded with other data mining techniques such as association rules and integrated with text mining.

Keywords: Data Mining, Disease Diagnosis, Naïve Bayes, Decision Trees, Holdout Confusion method.

I. INTRODUCTION

Data mining is the process of extracting hidden patterns from data. It can be applied to data sets of any size. The central process of Knowledge Discovery is the transformation of data into knowledge for decision making, known as Data Mining [3]. The approach aims to utilize the data mining techniques: Classification and Prediction. The heart disease data warehouse consists of mixed attributes containing both the numerical and categorical data [6]. Cleaned and filtered records are used with the intention that the irrelevant data from the warehouse would be removed before mining process occurs.

Classification analysis is the organization of data in given classes. The classification uses given class labels to order the objects in the data collection. It is not easy to differentiate Prediction from Classification. However, unlike a classification, the purpose of a predictive model is to determine future outcome rather than current behavior. The output attributes of a predictive model can be categorical or numeric.

Medical history data constitutes numerous tests necessary to diagnose a particular Disease. Data mining combines statistical analysis, machine learning and database technology to extract hidden patterns and relationships from large databases [1]. The proposed model was implemented by using holdout confusion method and decision tress. To evaluate each subset, it is easy to use holdout approach by adopting the classification base on data set for a prediction process. A decision tree is flowchart-like tree structure, where each internal node (no leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label.

A. Data Source

A **data set** (or **dataset**) is a collection of data, usually presented in tabular form. Each column represents a particular variable. Each row corresponds to a given member of the data set in question. Each value is known as a datum.

A total of 500 records with 13 medical attributes (factors) were obtained from Cleveland Clinic Foundation. The records were split equally into two datasets: training dataset (203 records) and testing dataset (297 records). To avoid bias, the records for each set were selected randomly. The Thirteen variables are shown in Table I.

Table I. Description of Attributes

Attribute name	Mixed values	Numeric values	Comments
Age	Numeric	Numeric	Age in Years
Sex	Male ,Female	1,0	Patient Gender
Chest Pain type	Angina and Abnormal	1-2	Test type for patients.
Blood Pressure	Numeric	Numeric	Resting bp upon admission
Cholesterol	Numeric	Numeric	Serum Cholesterol
Fasting Blood Sugar < 120	True, False	1,0	Is Fasting blood sugar less than 120?
Resting ECG	Normal, Abnormal, Hyp	0,1,2	Ventricular hypertrophy
Maximum Heart rate	Numeric	Numeric	Heart rate Archived
Induced Angina?	True, False	1,0	Patient experience in angina.
Old Peak	Numeric	Numeric	ST depression to test

Slope	Up, flat, down	1-3	Slope of the peak exercise ST segment
Number Colored Vessels	0,1,2,3	0,1,2,3	Major vessels by fluoroscopy
Thal	Nor, Fix, Rev	3,6,7	Thal value

II. RELATED WORK

Numerous works in literature related with heart disease diagnosis using data mining and artificial intelligence techniques have motivated the work. Some of the works are discussed below:

Fayyad. U [2] defined the basic notions on data mining and KDD, define the goals, present motivation, and give a high-level definition of the KDD Process and how it relates to Data Mining. The authors then focus on data mining methods. Basic coverage of a sampling of methods will be provided to illustrate the methods that are used. Fayyad cover a case study of a successful application in science data analysis: the classification of cataloging of a major astronomy sky survey covering 2 billion objects in the northern sky. The process can outperform human as well as classical computational analysis tools in astronomy on the task of recognizing faint stars and galaxies. The author also covers the problem of scaling a clustering problem to a large catalog database of billions of objects.

John C. Wood, Andrew J.Buda, and Daniel T.Barry [4] employed a new analytical tool, the Binomial joint time-frequency transform, to test the hypothesis that first heart sound frequency rises during the isovolumic contraction period. Cardiac vibrations were recorded from eight open chest of the dogs using an ultra light accelerometer cemented directly to the epicardium of the anterior left ventricle. The presence of rapid frequency dynamics limits the usefulness of stationary analysis techniques for the first heart sound. The Binomial transform provided much better resolution than the spectrograph or spectrogram, the two most common non-stationary signal analysis techniques. By revealing the onset and dynamics of first heart sound frequencies, time-frequency transforms may allow mechanical assessment of individual cardiac structures.

Jos´e A.Gamez ,Rafael Rum,and Antonio Salmeron [5] proposed a naive Bayes model for unsupervised data clustering, where the class variable is hidden. The feature variables can be discrete or continuous, as the conditional distributions are represented as mixtures of truncated exponentials (MTEs).The number of classes is determined using the data augmentation algorithm. The proposed model is compared with the conditional Gaussian model for some real world databases.

Onsy Abdel, Alini Nadder Hamdy and Mohammed. A.Hanjouri [7] proposed a heart diagnosis method using heart sounds. Heart Sound is one of the older for means for assessing the function of its valves it helps, together with Echocardiograms and Electrocardiographs, giving a clear and proper diagnostic of several heart diseases. In this research artificial neural networks are used to classify several valves related to heart disorders. A library of heart sound files, recorded via the traditional Stethoscope, are used to extract relevant features using several signal processing tools e.g. Discrete Wavelet Transfer (DWT) Fast Fourier Transform (FIT) and Linear Prediction Coding (LPC) .The achieved recognition rates were around 95.6%.

Thuraisingham. B [9] presented the empirical study on data mining and data warehousing. Data mining and datawarehousing are increasingly popular among IT Professional, academics and pupils. This analysis is about the need the value and the technological means of acquiring and using the information in a useful way and it has been structured as a self- learning guide. Thuraisingham have presented all the major functionalities and techniques of data mining and data warehousing in a comprehensible manner. The aim of this research is to provide the reader with sufficient information about data mining methods and algorithms so that they may use these methods for the research work.

III. METHODOLOGY

A. Existing System:

In the existing system, test model is performed on the same data by building the model on one portion of the data. The method is useful if entire portion of the data is tested. This is effectively applied if the amount of data is relatively small.

B. Proposed System:

Holdout confusion method contains the results about the actual and predicted classifications made by the naïve bayes on the test set. Such information is often displayed as a two dimensional matrix. To evaluate each subset, it is easy to use holdout approach by adopting the classification base on data set for a prediction process. The classification method used in this research is computed using the holdout confusion matrix. The classification is derived from the confusion matrix.

C. Purpose:

To illustrate how Naïve Bayes may be used to produce a holdout confusion method.

D. Problem:

A holdout confusion method contains information about actual and predicted classifications done by a classification algorithm. Performance of such algorithms is commonly evaluated using the data in the matrix.

E. Method:

The following is a holdout confusion method approach, a tabular representation of a rule.

Table II illustrates the arrangement of data into a 2 x 2 table for calculation of the indices of diagnostic problem.

D+ = presence of disease; D- = absence of disease; T+ = positive test; T- = negative test.

a = true positives (TP); b = false positive (FP); c = true negative (TN); d = false negative (FN).The total number of individuals who have the disease is (a + c), i.e. (TP + TN); the total number without the disease is (b + d), i.e. (FP + FN). The total number in the sample tested are (a + b + c + d).

a) TP = The test is positive and the patient has the disease.

b) FP = The test is positive and the patient does not have the disease.

c) TN= The test is negative and the patient have the disease.

d) FN = The test is negative and the patient does not have the disease.

Table II. Distribution of data

Test/Disease	D+	D-	Row Total
T+	TP (a)	FP(b)	TP+FP
T-	TN(c)	FN(d)	TN+FN
Column Total	TP+TN(a+c)	FP+FN(b+d)	

a. ***Holdout Confusion method Algorithm****:* Begin with the holdout confusion method of dimensions r by c (rows*columns). Although Bayes theorem provides a reliable method for modifying disease probabilities based on diagnostic test results.

i. Classify the data into situations a to d (as in Table II).
ii. Joint probabilities P (probability of 2 events occurring simultaneously), this involves building 2 x 2 probability table.
iii. Determine the relative amount of predictions; the probabilities (prevalence) of disease are entered into the appropriate equation. This yields the test probability for all the test cases.
iv. Compute Probability of TP, FP, and TN. i.e., *the* probability of failure occurrence gives a true positive, false positive or true negative prediction.
v. In order to estimate Probability FP or TP it must be known when a prediction is a false or true positive. In the false positive case, it must be proven that a failure would not have occurred if failure prediction and actions had not been in place, which seems infeasible. However, in the second case, it must be assured that a positive prediction is a true positive, which means that a failure is really non imminent.
vi. Calculate occurrence of failure only as Probability FN. The test situation can be caused by a false negative prediction where the execution of the failure prediction does not have actions performed upon negative predictions that caused the failure.
vii. The outcome of the failure prediction experiment yields a sequence of predictions (either positive or negative) and a sequence of failures, by which predictions can be classified as true or false.

Decision Tree Algorithm: Decision tree generates a Decision tree from the training tuples of data partition D.

Input: **a.** Data partition, D, Which is a set of training tuples and their associated class labels;

b. Attribute_list, the set of candidate attributes;

c. Attribute _selection _method, a procedure to determine the splitting criterion that"best"partitions the data tuples in to individual classes. This criterion consists of a splitting attribute and possibly, either a split point or splitting subset.

Output: A decision tree

Method: Each branch undergoes a test on an attributes.

i.Create a node N;
ii.If tuples in D are all of the same class,c then
iii.Return N as a leaf node labled with the class c;
iv.If attribute_list is empty then
v.Return N as a leaf node labeled with the majority class in D;
vi.Apply Attribute_selection_method(D,attribute list) to find the "best "splitting_critecerion;
vii.Label node N with splitting_criterion;
viii.If splitting_attribute is discrete-valued and multiway splits allowed then
ix.Attribute_list=attribute list splitting attribute;//remove splitting_attribute
x.For each outcome j of splitting_crition/*partition the tuples and grow subtress for each partition */
xi.Let D_j be the set of data tuples in D satisfying outcome j;//a partition
xii.If D_j is empty then
xiii.Attach a leaf labeled with the majority class in D to node N;
xiv.Else attach the node returned by Decision_Tree(D_j,attribute list) to node N;
xv.Return N.

IV. RESULTS AND DISCUSSION

The proposed method compares the actual values in the test dataset with the predicted values in the trained model. The method is tested with dataset contained 137 patients with heart disease and 160 patients without heart disease. The below Figure 1. shows the results of the holdout method for heart disease dataset.

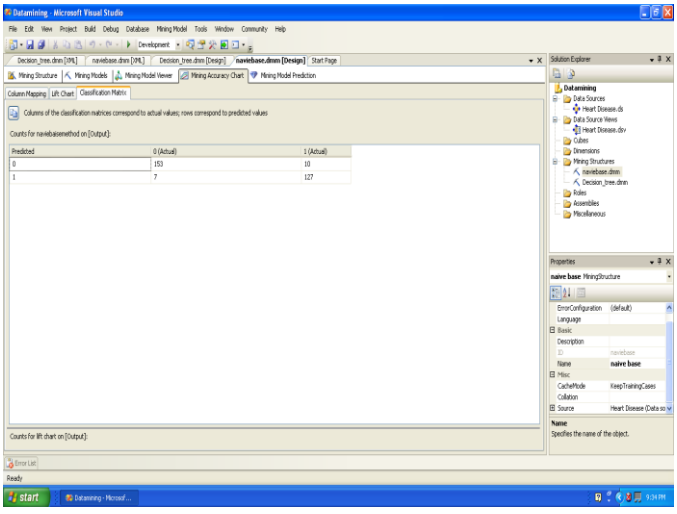

Figure 1. Holdout Confusion Method for Mining Structure

The rows represent predicted values while the columns represent actual values (1 for patients with heart disease, '0' for patients with no heart disease). The rows for each matrix

represent the predicted values for the model, whereas the columns represent the actual values.

A. Comparision of Mining Models:

Table III shows the Comparison of results that has been used to evaluate the performance on heart dataset collected from the data repository. In order to evaluate the performance of the proposed algorithm under different supporting Cases.

Table III. Comparision of Results

Method	Calculation of Accuracy
Naïve Bayes	294*100/297=98
Decision Tress	204*100/297=68.6

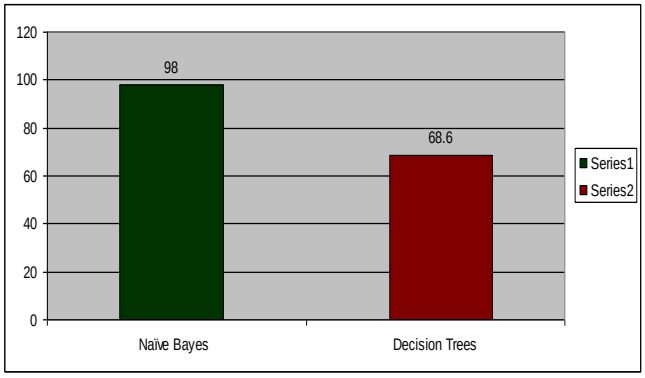

Figure 2.Chart representing the accuracy of database

The Chart (Figure.2) shows that Naïve Bayes has the overall better accuracy .The comparisons among both the Classification Algorithms is shown by calculating the Accuracy. Heart Dataset contains totally 297 Testing Data and total number of Supporting Cases for Naïve Bayes is 294.The Accuracy is calculated by using the formula.

Accuracy = Number of correct predictions that predicted as heart disease/ no .of Total Testing data*100.

V. CONCLUSION

A prototype heart disease prediction is developed using two data mining classification modeling techniques. The Process extracts hidden knowledge from a historical heart disease database [8]. Query language and functions are used to build and access the models [10]. The models are trained and validated against a test dataset. Proposed work uses two models that are able to extract patterns in response to the predictable state.

The most effective model to predict patients with heart disease appears to be Holdout Confusion Method followed by Decision Trees. The Prediction goals are evaluated against the trained models. Both two models could answer complex queries, each with its own strength with respect to ease of model interpretation, access to detailed information and accuracy. Although not the most effective model, Decision tree results are easier to read and interpret. The drill through feature to access detailed patients' profiles is only available in Decision Trees. Naïve Bayes fared better than Decision Trees as it could identify all the significant medical predictors.

VI. FUTURE WORK

The research can be further enhanced and expanded. 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. The future work can also be further expanded with applying fuzzy learning models to evaluate the intensity of cardiac disease.

VII. REFERENCES

[1] Chapman. P, Clinton. J, Kerber. R, Khabeza. T, Reinartz. T, Shearer.C, and Wirth.R, "CRISP- DM 1.0: Step by Step Data Mining Guide", 1-78, 2000.

[2] Fayyad.U: "Data Mining and Knowledge Discovery in Databases: Implications Scientific Databases", proc. of the 9th Int. Conf. on Scientific and Statistical Database Management, Olympia, 2 - 11, 1997.

[3] Han, J., Kamber, M, "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 2006.

[4] John C. Wood, Andrew J. Buda, and Daniel T. Barry,"Time-Frequency Transforms: A New Approach to First Heart Sound Frequency Dynamics", IEEE Transactions on Biomedical Engineering. Vol: 39, No: 7, July 1992.

[5] Jos´e A. G´amez, Rafael Rum and Antonio Salmeron,"Unsupervised Naive Bayes for Data Clustering with mixtures of Truncated Exponentials", Intelligent Systems and Data mining Group ,IEEE 2006.

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

[7] Onsy Abdel-AliniX, Nadder Hamdy and Mohammed. A. El-Hanjouri, "Heart Diseases Diagnosis Using Heart Sounds" Nineteenth National Radio Science Conference, March 19-21, 2002.

[8] Sandor Khoor, Ede Kekes**,** and Erno Berentey ,"Expert System for Automatic Phono- mechanocardiographic Diagnosis", IEEE 1990.

[9] Thuraisingham. B,"A Primer for Understanding and Applying Data Mining", Information Professional, 28-31, 2000.

[10] Winnie W. Hui, Ronald A. Pitt, John P. and Matonick, John Li,"Comparison of Heart Operations Recorded at the Chest and a Remote Arterial Site", IEEE 2002.