A Data mining Model for predicting the Coronary Heart Disease using Random Forest Classifier

A. Sheik Abdullah
PG Scholar
Department of CSE
Kongu Engineering College, Erode, Tamil Nadu

R.R.Rajalaxmi
Professor
Department of CSE
Kongu Engineering College, Erode, Tamil Nadu

ABSTRACT

Coronary Heart Disease (CHD) is a common form of disease affecting the heart and an important cause for premature death. From the point of view of medical sciences, data mining is involved in discovering various sorts of metabolic syndromes. Classification techniques in data mining play a significant role in prediction and data exploration. Classification technique such as Decision Trees has been used in predicting the accuracy and events related to CHD. In this paper, a Data mining model has been developed using Random Forest classifier to improve the prediction accuracy and to investigate various events related to CHD. This model can help the medical practitioners for predicting CHD with its various events and how it might be related with different segments of the population. The events investigated are Angina, Acute Myocardial Infarction (AMI), Percutaneous Coronary Intervention (PCI), and Coronary Artery Bypass Graft surgery (CABG). Experimental results have shown that classification using Random Forest Classification algorithm can be successfully used in predicting the events and risk factors related to CHD.

Keywords

Coronary Heart Disease, Decision Trees, Random Forest.

1. INTRODUCTION

Data mining has become a fundamental methodology for computing applications in medical informatics. Progress in data mining applications and its implications are manifested in the areas of information management in healthcare organizations, health informatics, epidemiology, patient care and monitoring systems, assistive technology, large-scale image analysis to information extraction and automatic identification of unknown classes. Various algorithms associated with data mining have significantly helped to understand medical data more clearly, by distinguishing pathological data from normal data, for supporting decision-making as well as visualization and identification of hidden complex relationships between diagnostic features of different patient groups.

Coronary Heart Disease (CHD) is a major cause of disability in adults in and common cause of death in Europe, USA, South Asia, etc., It has been predicted that all the regions of the world will be affected due to CHD by the year 2020 [1]. Coronary Heart Disease refers to the failure of coronary circulation to supply adequate circulation to cardiac muscle and its surrounding tissue. This restricts the supply of blood and oxygen to the heart, particularly during exertion when the myocardial metabolic demands are increased. As the degree of coronary artery disease progresses, there may be near-complete obstruction of the lumen of the coronary artery, severely restricting the flow of oxygen-carrying blood to the myocardium. Individuals with this degree of coronary artery disease typically have suffered from one or more myocardial infarctions (heart attacks), and may have signs and symptoms of chronic coronary ischemia, including symptoms of angina at rest and flash pulmonary edema [2].

The effects produced due to CHD are constant fatigue, physical disability, mental stress and depression. This paper focuses on the creation of a data mining model using the Random forest classification algorithm for evaluating and predicting various events related to CHD. Some of these studies, has made with the implementation of data mining algorithm such as K-NN, Naïve bayes, K-means, ID3, and Apriori algorithms. The growing healthcare burden and suffering due to life threatening diseases such as heart disease and the escalating cost of drug development can be significantly reduced by design and development of novel methods in data mining technologies and allied medical informatics disciplines. In CHD, if the risk factors are predicted in advance two sorts of problem can be solved. First, various surgical treatments such as angioplasty, coronary stents, coronary artery bypass and heart transplant can be avoided to a great extent. Second, the associated cost with each risk factor can be reduced.

The rest of the paper is organized as follows. Section II discusses the literature review related to CHD. Section III describes the material and methods used to develop the model. Section IV concerns with the experimental results. Finally the paper is concluded in Section V.

2. LITERATURE REVIEW

Tsien et al [3] in their study indicated that classification trees, which have certain advantage over logistic regression models, with patients having Myocardial Infarction (MI). The results shown that the occurrence of MI has been noticed in male than the female. Age, Systolic blood pressure, smoking has been found to be the important risk factor in the patients with MI.

Rea et al [4] concluded that smoking has been associated with an elevated risk for recurrent coronary events such as Angina, Acute Myocardial Infarction (AMI). In some cases, smoking has been associated with cholesterol for AMI. The subjects can be extended with various other events related to CHD. Since the risk factors have different degrees of impact, the population-specific risk function is needed for the prediction of CHD.

Wang et al [5] used the risk factors such as age, sex, cholesterol, blood pressure, diabetes and smoking to predict CHD. They used the Framingham function and concluded that the traditional risk factors have different degrees of impact and the other factors that are contributing to the risk.

Karaolis et al [6] developed a data mining system for the assessment of heart related risk factors using association analysis based on Apriori algorithm. The results with 369 cases shown that smoking is one of the main risk factor that directly affect the coronary heart disease for all the events.

Kunc et al [7] presented simulation results which can be used for evaluation of patients with coronary heart disease, congestive heart failure, end-stage renal disease in Slovenia. At the same time also year treatment costs were calculated regarding each of observed diseases. The presented results enable the estimation of potential savings resulting from more intensive chronic diseases treatment.

Srinivas et al [8] focused on using different algorithms for predicting combinations of several target attributes and presented automated and effective heart attack prediction methods using data mining techniques such as Decision trees, Neural networks and Bayesian models. Firstly, they provided an efficient approach for the extraction of significant patterns from the heart disease data warehouses for the efficient prediction of heart attack Based on the calculated significant weightage, the frequent patterns having value greater than a predefined threshold were chosen for the valuable prediction of heart attack. Three mining goals are defined based on data exploration. For predicting heart attack significantly 15 attributes have been chosen. The future work signifies the usage of other attributes such as financial status, stress, pollution and previous medical history. Other data mining techniques, Time Series, Clustering and Association Rules can also be used to analyze patients' behavior.

Soni et al [9] provided a survey of current techniques in Data mining for heart disease prediction. Experiments has been conducted with various sorts of techniques using the same dataset out of which Decision tree shown high accuracy than that of the Bayesian classification, KNN, neural networks. The accuracy has been further improved by applying genetic algorithm with Decision trees. The work can be extended by using real dataset from health care organizations for the automation of Heart Disease prediction.

Rafiah et al [10] using Decision Trees, Naive Bayes, and Neural Network techniques developed a system for heart disease prediction using the Cleveland Heart disease database and shown that Naïve Bayes performs well followed by Neural Network and Decision Trees. The relationship between attributes produced by Neural Network is more difficult to understand than that of the other models used to predict heart disease. Continuous data can be used instead of categorical data and text mining methods can be incorporated to mine vast amount of unstructured data available in healthcare databases.

Karl berg and Elo [11] calculated the burden of Ischemic Heart Disease (IHD) and coronary risk factors in a defined population using data from all public providers of healthcare. Calculation of the actual burden of disease in the population showed that when hospital discharge data were combined with the outpatient data, there were no or slight difference in the age-specific rates of Acute Myocardial Infarction (AMI), while the rates of angina were between two-fold and four-fold higher, and unspecified IHD was between three-fold and ten-fold higher in individuals aged greater than 50 years compared with using hospital discharge data alone. These findings suggest that hospital discharge data should be combined with outpatient care data to provide a more comprehensive estimate of the burden of IHD and its risk factors.

Meanwhile, this paper deals with the investigation of various events with their impacts and its risk factors associated with CHD and to improve the overall prediction accuracy using the Random forest classification algorithm.

3. MATERIALS AND METHODS

3.1 Data Description

Data records of various CHD patients from the UCI machine learning repository has been taken for the evaluation of the model. The record related to each patient should have at least one of the following in their history: Angina, Acute Myocardial Infarction (AMI), Percutaneous Coronary Intervention (PCI), and Coronary Artery Bypass Graft (CABG). The predominant risk factors is of Clinical factors such as age, sex, trestbps, chest pain type and Biochemical factors such as serum cholesterol (TC) mg/dL, fasting blood sugar (FBS) mg/dL, thalach, exang, old peak, restecg, slope, the number of vessels colored by fluoroscopy and thal. The following flowchart depicts the evaluation and prediction of CHD and its associated events.

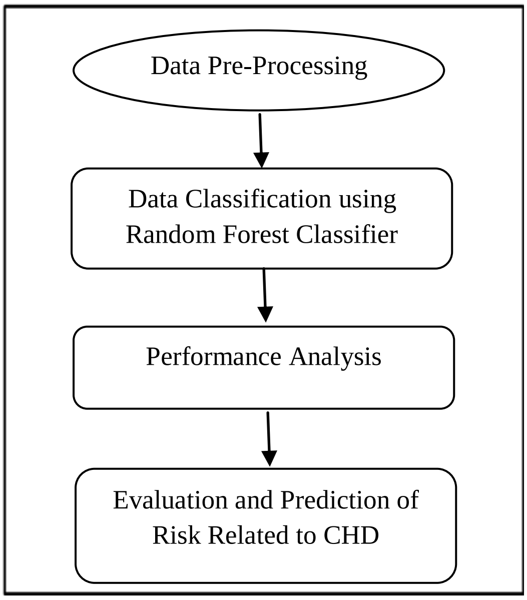
Data Pre-Processing
Data Classification using Random Forest Classifier
Performance Analysis
Evaluation and Prediction of Risk Related to CHD

Figure 1. Methodology used to construct the Model

The evaluation proceeds by normalizing the input data set by means of Min-Max normalization. This technique performs a linear transformation and preserves the relationship among the original data values. Min-Max normalization maps a value v, of an attribute A to v' in the range [new_min$_A$, new_max$_A$] by computing,

$$v' = \frac{v - min_A}{max_A - min_A}\left(new_{max_A} - new_{min_A}\right) + new_{min_A} \qquad (1)$$

It will encounter an "out of bounds" error if a future input case for normalization falls outside of the original data range for A.

3.2 Data Classification using Random Forest Classifier

Random forest (or random forests) is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the class's output by individual trees. It is one of the most accurate among the learning algorithms available.

For many data sets, it produces a highly accurate classifier. The algorithm for inducing a random forest was developed by Leo Breiman and Adele Cutler [12] the method combines Breiman's "bagging" idea and the random selection of features, in order to construct a collection of decision trees with controlled variation.

Algorithm: Random forest classifier

Input:

1. Training Dataset N, Which is a set of training observations and their associated class values.

Output: Generates Decision trees

Each tree is constructed based on the following steps

1. Let the number of training cases be *N*, and the number of variables in the classifier be *M*.
2. The number *m* of input variables to be used to determine the decision at a node of the tree; *m* should be much less than *M*.
3. Choose a training set for this tree by choosing *n* times with replacement from all *N* available training cases (i.e. take a bootstrap sample). Use the rest of the cases to estimate the error of the tree, by predicting their classes.
4. For each node of the tree, randomly choose *m* variables on which to base the decision at that node. Calculate the best split based on these *m* variables in the training set.
5. Each tree is fully grown and not pruned (as may be done in constructing a normal tree classifier).

For prediction a new sample is pushed down the tree. It is assigned the label of the training sample in the terminal node it ends up in. This procedure is iterated over all trees in the ensemble, and the average vote of all trees is reported as random forest prediction.

3.3 Models Investigated

1. AMI: Myocardial infarction (MI) or acute myocardial infarction (AMI), commonly known as a heart attack, results from the interruption of blood supply to a part of the heart, causing heart cells to die. This is most commonly due to occlusion (blockage) of a coronary artery following the rupture of a vulnerable atherosclerotic plaque, which is an unstable collection of lipids (cholesterol and fatty acids) and white blood cells (especially macrophages) in the wall of an artery.
2. PCI: Percutaneous coronary intervention (PCI), commonly known as coronary angioplasty or simply angioplasty, is one therapeutic procedure used to treat the stenotic (narrowed) coronary arteries of the heart found in coronary heart disease. These stenotic segments are due to the buildup of cholesterol-laden plaques that form due to atherosclerosis.
3. CABG: Coronary artery bypass surgery, also coronary artery bypass graft surgery, and colloquially heart bypass or bypass surgery is a surgical procedure performed to relieve angina and reduce the risk of death from coronary artery disease. Arteries or veins from elsewhere in the patient's body are grafted to the coronary arteries to bypass atherosclerotic narrowing's and improve the blood supply to the coronary circulation supplying the myocardium (heart muscle).

3.4 Performance Measures

In this approach, the classification accuracy rates for the datasets were measured. For example, in the classification problem with two-classes, positive and negative, an single prediction has four possibility. The True Positive rate (TP) and True Negative rate (TN) are correct classifications. A False Positive (FP) occurs when the outcome is incorrectly predicted as positive when it is actually negative. A False Negative (FN) occurs when the outcome is incorrectly predicted as negative when it is actually positive.

Table 1. Confusion Table

Prediction		Disease	
		+	-
Test	+	True Positive (TP)	False Positive (FP)
	-	False Negative (FN)	True Negative (TN)

1. Accuracy - It refers to the total number of records that are correctly classified by the classifier.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \qquad (2)$$

2. Classification error - This refers to the misclassified datasets from the correctly classified records.
3. True Positive Rate (TP) : It corresponds to the number of positive examples that have been correctly predicted by the classification model.
4. False Positive Rate (FP) : It corresponds to the number of negative examples that have been wrongly predicted by the classification model.
5. Kappa Statistics - A measure of the degree of nonrandom agreement between observers or measurements of the same categorical variable.
6. Precision - is the fraction of retrieved instances that are relevant.

$$Precision = \frac{TP}{TP + FP} \qquad (3)$$

7. Recall - is the fraction of relevant instances that are retrieved.

$$Recall = \frac{TP}{TP + FN} \qquad (4)$$

8. Root-Mean-Squared-error - It is a statistical measure of the magnitude of a varying quantity. It can be calculated for a series of discrete values or for a continuously varying function.

Since the class label prediction is of multi-class, the result on the test set will be displayed as a two-dimensional confusion matrix with a row and column for each class. Each matrix element shows the number of test cases for which the actual class is the row and the predicted class is the column.

4. RESULTS AND DISCUSSION

Random forest algorithm runs efficiently on large databases and has the capability of handling thousands of input variables. It generates an internal unbiased estimate of the generalization error as the forest building progresses and has an effective method for estimating missing data and maintains accuracy when large proportion of the data are missing. The tree forest that has been generated can be saved in order to make comparative study about the features of the attributes.

To measure the effectiveness of the approach experiments have been conducted using the UCI machine learning dataset with the Cleveland data consisting of 13 attributes. The attributes involved are age, sex, chest pain type, serum cholesterol, fasting blood sugar, resting electro cardio graphic results, maximum heart rate achieved, exercise induced angina, ST depression, the major vessels colored by fluoroscopy, and thal. For constructing a forest of random trees, the following parameters have been incorporated.

Table 2. Parameters Used

Parameter	Values
Maximum Depth	0
NumExecution Slots	2
Number of Trees	10
Seed	2

Meanwhile, Decision trees are constructed in a top-down recursive divide-and-conquer manner and the compatibility of Decision trees degrades because the output is limited to one attribute. Trees created from the numeric datasets seems to be more complex and also when the database is large the complexity of the tree increases. In comparison with the Random Forest algorithm the time complexity of Decision trees increases exponentially with the tree height. Hence shallow trees tend to have large number of leaves and high error rates.

As the tree size increases, training error decreases. However, as the tree size increases, testing error decreases at first since we expect the test data to be similar to the training data, but at a certain point, the training algorithm starts training to the noise in the data, becoming less accurate on the testing data. At this point we are no longer fitting the data and instead fitting the noise in the data. This is called over fitting to the data, in which the tree is fitted to spurious data. As the tree grows in size, it will fit the training data perfectly and not be of practical use for other data such as the testing set.

Table 3. Performance Analysis

Performance Measures	Decision Tree	Random Forest
Accuracy (%)	50.67	63.33
True Positive Rate	0.507	0.633
False Positive Rate	0.23	0.254
Precision (%)	0.478	0.570
Recall (%)	0.507	0.633
Classification Error (%)	49.32	36.66
Kappa Statistics	0.211	0.354
RMS Error	0.404	0.313

The performance obtained using Random forest classifier was found to be higher than the results obtained by Reena et al [13] as described in the Table 2 which depicts that Random forest algorithm performs better than that Decision trees.

5. CONCLUSION

Data mining plays an important role in the identification and prediction of various sort of metabolic syndromes and hence various sorts of diseases can be discovered. In the existing work, Decision tree classification algorithm has been used to assess the events related to CHD. The proposed work is mainly concerned with the development of a data mining model with the Random Forest classification algorithm. The developed model will have the functionalities such as predicting the occurrence of various events related to each patient record, prevention of risk factors with its associated cost metrics and an improvement in overall prediction accuracy. As a result, the causes and the symptoms related to each event will be made in accordance with the record related to each patient and thereby CHD can be reduced to a great extent.

6. REFERENCES

[1] British Heart Foundation. (2008, mar. 8). European Cardio-vascular Disease Statistics.

[2] Z.J Zhang and S.H Wang ‘Synopsis of Prescriptions of the Golden Chamber’ Beijing: People’s Medical Publishing House, pp.27–29, 2007.

[3] C.L. Tsien, H.S.F. Fraser, W.J. Long and R.L. Kennedy “Using classification trees and logistic regression methods to diagnose myocardial infarction” in Proc. 9th World Congr., Inf., vol. 52, pp. 483-497, 2001.

[4] T.D. Rea, S.R. Heckbert, R.C. Kaplan, N.L. Smith, R.N. Lemaitre, and B.M.Psaty, “Smoking status and risk for recurrent coronary events after myocardial infarction” Ann. Int. Med., vol. 137, pp. 494-500, 2002.

[5] Z. Wang and W.E. Hoy “Is the Framingham coronary heart disease absolute risk function applicable to Aboriginal people?” Med.J.Australia, Vol.182, No. 2, pp. 66-69, 2005.

[6] M. Karaolis, J.A. Moutiris, L. Pattichs “Assessment of the Risk Factors of Coronary Heart Events Based on Data Mining With Decision Trees”, IEEE Transactions on IT in Biomedicine, vol. 14, No. 3, 2010.

[7] J. Kunc, S. Drinovec, Rucigaj, and A. Mrhar, “Simulation analysis of coronary heart disease, congestive heart failure and end-stage renal disease economic burden." Mathematics and computers in simulation, 2010.

[8] K. Srinivas, G. Raghavendra, Rao, and A. Govardhan, “Analysis of Coronary Heart Disease and Prediction of Heart Attack in Coal Mining Regions Using Data Mining Techniques”, 5th IntConf on Computer Science & Education Hefei, China, pp. 1344-1349, 2010.

[9] J. Soni, U. Ansari, and D. Sharmaa “Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction” Int., Journal of Computer Applications vol. 17, No. 8, 2011.

[10] A.Rafiah and P.Sellappan “Intelligent Heart Disease Prediction System Using Data Mining Techniques”, 2008.

[11] I.H. Karlberg, and S.L. Elo “Validity and utilization of epidemiological data: A study of Ischaemic heart disease and coronary risk factors in a local population”, 2009.

[12] L. Breiman and A. Cutler “ www.stat.berkeley.edu”.

[13] A. Reena and S.G. Rajkumar, “Diagnosis of heart disease using Data mining algorithm”, Global Journal of Computer science and Technology, Vol. 10, No. 10, 2010.