Second Opinion Decision Support System for Cardiovascular Disease Using Data Mining Techniques (SODSS)

1Miss. Poonam N. Mane  
Department of CSE,  
Dr. Babasaheb Ambedkar Technological University,  
Lonere, Mangaon, Raigad, Maharashtra, India.

2Prof. Yogesh N. Patil  
Department of CSE,  
Dr. Babasaheb Ambedkar Technological University,  
Lonere, Mangaon, Raigad, Maharashtra, India.

Abstract--The healthcare industry collects huge amounts of healthcare data which, unfortunately, are not “mined” to discover hidden information for effective decision making by healthcare practitioners. Discovery of hidden patterns and relationships often goes unexploited. Advanced data mining modeling techniques can help overcome this situation. Models developed from these techniques will be useful for medical practitioners to take effective decision. In this research paper, one of the data mining classification technique Decision Tree algorithm ID3, CART and Support Vector Machine are analyzed on cardiovascular disease dataset. Performances of these algorithms are compared through sensitivity, specificity, accuracy, error rate, True Positive Rate and False Positive Rate. In our studies 10-fold cross validation method was used to measure the unbiased estimate of these prediction models. As per our results, error rates for Decision Tree algorithms- ID3, CART and SVM are 0.2756, 0.2755 and 0.2248 respectively. Accuracy of Decision Tree algorithm ID3, CART and Support Vector Machine are 80.06%, 81.08% and 86.12% respectively. Our analysis shows that SVM predicts cardiovascular disease with least error rate and highest accuracy.

Keywords: Data mining, medical engineering, ID3, CART and SVM.

I. INTRODUCTION

The heart is the organ that pumps blood, with its life giving oxygen and nutrients, to all tissues of the body. If the pumping action of the heart becomes inefficient, vital organs like the brain and kidneys suffer and if the heart stops working altogether, death occurs within minutes. Life itself is completely dependent on the efficient operation of the heart. Cardiovascular disease is not contagious; you can’t catch it like you can the flu or a cold. Instead, there are certain things that increase a person’s chances of getting cardiovascular disease. Cardiovascular disease (CVD) refers to any condition that affects the heart. Many CVD patients have symptoms such as chest pain (angina) and fatigue, which occur when the heart isn’t receiving adequate oxygen. As per a survey nearly 50 percent of patients, however, have no symptoms until a heart attack occurs. A number of factors have been shown to increase the risk of developing CVD. Some of these are:

• Family history of cardiovascular disease
• High levels of LDL (bad) cholesterol
  - Low level of HDL (good) cholesterol
• Hypertension
• High fat diet
• Lack of regular exercise
• Obesity

With so many factors to analyze for a diagnosis of cardiovascular disease, physicians generally make a diagnosis by evaluating a patient’s current test results. Previous diagnoses made on other patients with the same results are also examined by physicians. These complex procedures are not easy. Therefore, a physician must be experienced and highly skilled to diagnose cardiovascular disease in a patient. Data mining has been heavily used in the medical field, to include patient diagnosis records to help identify best practices. The difficulties posed by prediction problems have resulted in a variety of problem-solving techniques. It is difficult, however, to compare the accuracy of the techniques and determine the best one because their performance is data dependent. A few studies have compared data mining and statistical approaches to solve prediction problems. The comparison studies have mainly considered a specific data set or the distribution of the dependent variable.
II. BACKGROUND

Up to now, several studies have been reported that have focused on cardiovascular disease diagnosis. These studies have been applied for different approaches to the given problem and achieved high classification accuracies of 77% or higher.

Here are some examples:

a. Robert Detrano’s experimental results showed correct classification accuracy of approximately 77% with logistic regression derived discriminant function [3].
b. Zheng Yao applied a new model called R-C4.5 which is based on C4.5 and improved the efficiency of attribution selection and partitioning models. An experiment showed that the rules created by R-C4.5s can give health care experts clear and useful explanations [4].
c. Resul Das introduced a methodology that uses SAS basesoftware 9.13 for diagnosing heart disease. A neural networks ensemble method is at the center of this system [5].
d. Colombet et al. evaluated implementation and performance of CART and artificial neural networks comparatively with a LR model, in order to predict the risk of cardiovascular disease in a real database [6].

The difficult of recognizing constrained association rules for heart illness prediction was studied by Carlos Ordonez. The data mining techniques have been engaged by various works in the works to analyze various diseases, for instance: Hepatitis, Cancer, Diabetes, Heart diseases. Frequent Itemset Mining (FIM) is measured to be one of the basic data mining difficulties that expects to discern collections of items or values or forms that co-occur regularly in a dataset. The term Heart illness covers the various diseases that affect the heart. Heart disease kills one in every 32 seconds in the United States of America. This technique is used while prescribing the patient and this system predicts which remedy in the form of medicines and medical test suits best.

III. PROBLEM STATEMENT

Many hospital information systems are designed to support patient billing, inventory management and generation of simple statistics. Some hospitals use decision support systems, but they are largely limited. They can answer simple queries like “What is the average age of patients who have heart disease?”, “How many surgeries had resulted in hospital stays longer than 10 days?”. However, they cannot answer complex queries like “Identify the important Preoperative predictors that increase the length of hospital stay and “Given patient records, predict the probability of patients getting a heart disease.” 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 affect the quality of service provided to patients. Wu et al. proposed 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. 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.

IV. CLASSIFICATION MODELS

Under this section we will discuss following data mining Classification Models Decision Tree algorithm ID3, CART and Support Vector Machine to predict cardiovascular disease:

I. Decision Tree Algorithm

a) ID3 (Itemized Dichotomize 3)

Itemized Dichotomize 3 algorithm or better known as ID3 algorithm was first introduced by J.R Quinlan in the late 1970s [7]. It is a greedy algorithm that selects the next attributes based on the information gain associated with the attributes. Throughout the algorithm, the decision tree is constructed with each non-terminal node representing the selected attribute on which the data was split, and terminal nodes representing the class label of the final subset of this branch [7].

Algorithm of the ID3:

i. Create a root node for the tree
ii. If all examples are positive, Return the single-node tree Root, with label = +.
iii. If all examples are negative, Return the single-node tree Root, with label = -.
iv. If number of predicting attributes is empty, then Return the single node tree Root, with label = most common value of the target attribute in the examples.

b) CART

The term Classification And Regression Tree (CART) analysis is an umbrella term used to refer to both of the above procedures, first introduced by Breiman et al. Trees used for regression and trees used for classification have some similarities - but also some differences, such as the procedure used to determine where to split. It uses Gini impurities and information gain to calculate decision tree.

CART innovations include:

1. Solving the “how big to grow the tree”- problem,
2. Using strictly two-way (binary) splitting,
3. Incorporating automatic testing and tree validation.
2. Support Vector Machine

The SVM is a state-of-the-art maximum margin classification algorithm rooted in statistical learning theory. SVM is method for classification of both linear and non-linear data. It uses a non-linear mapping to transform the original training data into a higher dimension. Within this new dimension it searches for linear optimal separating hyperplane. With an appropriate nonlinear mapping to a sufficiently high dimension, data from two classes can always be separated by a hyperplane. The SVM finds this hyperplane using support vectors and margins. SVM performs classification tasks by maximizing the margin separating both classes while minimizing the classification errors. Fig 1. Shows SVM topology in hyperspace:

Fig 1. Shows SVM topology

V. DATASET

To compare these data mining classification techniques Cleveland cardiovascular disease dataset from UCI repository was used\[10\]. The dataset has 14 attributes and 303 records.

**Key attribute**

1. Patient_id – Patient’s identification number.

Attribute value to be taken into the project for heart disease is as follows:

**Heart disease dataset:**

1. Sex (value 1: Male; value 0: Female)
2. Chest Pain Type (value 1: typical type 1 angina, value 2: 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: down sloping)
7. CA – number of major vessels colored by fluoroscopy (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. Num Class (0 = healthy, 1 = have heart disease)

VI. RESULTS

These data mining classification model were developed using datamining classification tool Weka version 3.6. Initially dataset had 14 attributes and 303 records. Algorithm for attribute selection was applied on dataset to preprocess the dataset. After attribute selection missing values records were identified and were deleted from dataset. After deleting records with missing values we were left with 296 records. On these 296 records data mining Decision Tree algorithms ID3, CART and SVM were applied. A distinguished confusion matrix was obtained to calculatesensitivity, specificity and accuracy.

<table>
<thead>
<tr>
<th></th>
<th>Classified as Healthy</th>
<th>Classified as Unhealthy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual Healthy</td>
<td>TP</td>
<td>FN</td>
</tr>
<tr>
<td>Actual not Healthy</td>
<td>FP</td>
<td>TN</td>
</tr>
</tbody>
</table>

The upper left cell denotes the number of samples classifies as true while they were true (i.e., TP), and the lower right cell denotes the number of samples classified as false while they were actually false (i.e., TN). The other two cells (lower left cell and upper right cell) denote the number of samples misclassified. Specifically, conclusion the upper right cell denoting the number of samples classified as false while they actually were true (i.e., FN), and the lower left cell denoting
the number of samples classified as true while they actually were false (i.e., FP).

**Below formulae were used to calculate sensitivity, specificity and accuracy:**

1. Sensitivity = TP / (TP + FN)
2. Specificity = TN / (TN + FP)
3. Accuracy = (TP + TN) / (TP + FP + TN + FN)

Table 2. Shows sensitivity, specificity and accuracy for different classification algorithms.

<table>
<thead>
<tr>
<th>Classification models</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID3</td>
<td>83.75%</td>
<td>75.73%</td>
<td>80.06%</td>
</tr>
<tr>
<td>CART</td>
<td>86.25%</td>
<td>75.82%</td>
<td>81.08%</td>
</tr>
<tr>
<td>SVM</td>
<td>92.0%</td>
<td>77.20%</td>
<td>86.12%</td>
</tr>
</tbody>
</table>

![Fig 2. Shows sensitivity, specificity, accuracy and error rate for different classification techniques in 3-D Column chart format.](image)

As per results, error rates for Decision Tree algorithm ID3, CART and SVM are 0.2756, 0.2575 and 0.2248 respectively. Accuracy of Decision Tree algorithm ID3, CART and SVM are 80.06%, 81.08% and 86.12% respectively.

1. True Positive Rate = TP / (TP + FN)
2. False Positive Rate = FP / (FP + TN)

Table 3. Shows True Positive Rate and False Positive Rate for Classification Models Decision Tree algorithms-ID3, CART and Support Vector Machine. This will represent 100% True Positive Rate and no False Positive Rate which will be ideal case.

<table>
<thead>
<tr>
<th>Classification models</th>
<th>True Positive Rate</th>
<th>False Positive Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID3</td>
<td>0.8375</td>
<td>0.2526</td>
</tr>
<tr>
<td>CART</td>
<td>0.8625</td>
<td>0.2410</td>
</tr>
<tr>
<td>SVM</td>
<td>0.9000</td>
<td>0.2279</td>
</tr>
</tbody>
</table>

**VII. CONCLUSION**

There are different data mining techniques that can be used for the identification and prevention of cardiovascular disease among patients. In this paper, four classification techniques in data mining to predict cardiovascular disease in patients are compared: Decision Tree algorithms ID3, CART and SVM. These techniques are compared on basis of Sensitivity, Specificity, Accuracy, Error Rate, True Positive Rate and False Positive Rate. Our studies showed that SVM classification model turned out to be the best classifier for cardiovascular disease prediction. In future, we intend to improve performance of these basic classification techniques by creating meta model which will be used to predict cardiovascular disease in patients.

**ACKNOWLEDGMENT**

I take this opportunity to express my sincere thanks and deep sense of gratitude to my guide, Prof. Y. N. Patil for his constant support, motivation, valuable guidance and immense help during the entire course of this work. Without his constant encouragement, timely advice and valuable discussions, it would have been difficult to complete this work.

I am also grateful to Dr. A. W. Kiwlekar, Head of Department and the entire staff members for providing me the necessary facilities.

I am equally thankful to my batch mates Miss. Afroz Momin and my friends for their valuable guidance, timely help for completion of the work successfully. I also express my sincere thanks and gratitude to my father, mother, entire family and my friends for their constant help and support during the entire project work. Lastly, I thank everyone and express my apology that I could not mention the names one by one, who have been related directly or indirectly in this successful journey.

**REFERENCES**


[10] N. AdityaSundar, P. PushpaLatha, M. Rama Chandra “Performance Analysis Of Classification Data Mining Techniques Over Heart Disease Data Base”.
