International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 2, February - 2014

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

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

[2]Prof. Yogesh N. Patil
Department of CSE,
Dr.BabasahebAmbedkar 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 arecompared 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 02.756, 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 andhighest accuracy.**

Keywords:Data mining, medical engineeringID3, CART and SVM.

I. INTRODUCTION

The heart is the organ that pumps blood, with its life giving oxygenand nutrients, to all tissues of the body. If the pumping actionof the heart becomes inefficient, vital organs like the brain andkidneys suffer and if the heart stops working altogether, deathoccurs within minutes. Life itself is completely dependent onthe efficient operation of the heart. Cardiovascular disease is notcontagious; 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 gettingcardiovascular disease. Cardiovascular disease (CVD) refers toany condition that affects the heart. Many CVD patients havesymptoms such as chest pain (angina) and fatigue, which occurwhen the heart isn't receiving adequate oxygen. As per a surveynearly 50 percent of patients, however, have no symptoms untila heart attack occurs. A number of factors have been shown toincrease 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 cardiovasculardisease, physicians generally make a diagnosis by evaluating apatient's current test results. Previous diagnoses made on otherpatients with the same results are also examined by physicians. These complex procedures are not easy. Therefore, **a physicianmust be experienced and highly skilled to diagnose cardiovasculardisease in a patient.**Data mining has been heavily used in the medical field, toinclude patient diagnosis records to help identify best practices. The difficulties posed by prediction problems have resulted in avariety of problem-solving techniques.

It is difficult, however, to compare the accuracy of the techniquesand determine the best one because their performance is data dependent. A few studies have compared data mining and statisticalapproaches to solve prediction problems. The comparison studieshave mainly considered a specific data set or the distribution ofthe dependent variable.

II. BACKGROUND

Up to now, several studies have been reported that have focusedon cardiovascular disease diagnosis. These studies have been appliedfor different approaches to the given problem and achieved highclassification accuracies of 77% or higher.

Here are someexamples:

a. Robert Detrano's experimental results showed correctclassification accuracy of approximately 77% with logistic regression derived discriminant function [3] .
b. Zheng Yao applied a new model called R-C4.5 which is basedon C4.5 and improved the efficiency of attribution selectionand partitioning models. An experiment showed that the rulescreated by R-C4.5s can give health care experts clear anduseful explanations[4].
c. Resul Das introduced a methodology that uses SAS basesoftware 9.13 for diagnosing heart disease. A neural networksensemble method is at the center of this system[5].
d. Colombet et al. evaluated implementation and performanceof CART and artificial neural networks comparatively witha LR model, in order to predict the risk of cardiovasculardisease 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 Item set 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 miningClassification Models Decision Tree algorithm ID3, CART and Support Vector Machine
to predict cardiovascular disease:

1. 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 1970"s[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 intoa 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 findthis hyper plane using support vectors and margins . SVMperforms classification tasks by maximizing the margin separatingboth classes while minimizing the classification errors.Fig 1. Shows
SVM topology in hyperspace:

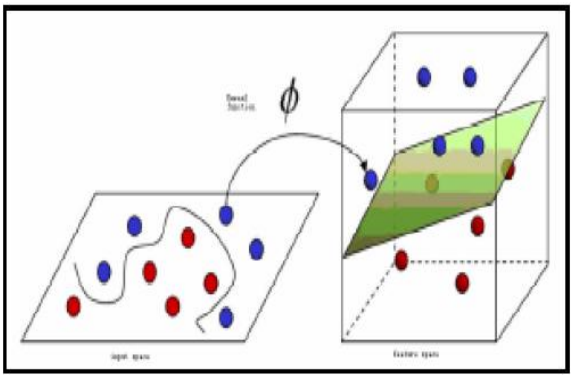

Fig 1.Shows SVM topology

V. DATASET

To compare these data mining classificationtechniques 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, value2: 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; value2: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 (value1: 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; value7: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 heartdisease)

VI. RESULTS

These data mining classification model were developed using datamining classification tool Weka version 3.6. Initially dataset had14 attributes and 303 records. Algorithm for attribute selection wasapplied on dataset to preprocess the dataset. After attribute selectionmissing values records were identified and were deleted fromdataset. After deleting records with missing values we were leftwith 296 records. On these 296 records data mining Decision Tree algorithms ID3, CARTand SVM were applied.
A distinguished confusion matrix was obtained to calculatesensitivity, specificity and accuracy.

Table 1.Shows confusion matrix.

	Classified as Healthy	**Classified as Unhealthy**
Actual Healthy	**TP**	**FN**
Actual not Healthy	**FP**	**TN**

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

the number of samples classified as true while theyactually were false (i.e., FP).

Below formulae were used to calculate sensitivity, specificityand 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 differentClassification algorithms.

Classification models	Sensitivity	Specificity	Accuracy
ID3	**83.75%**	**75.73%**	**80.06%**
CART	**86.25%**	**75.82%**	**81.08%**
SVM	**92.0%**	**77.20%**	**86.12%**

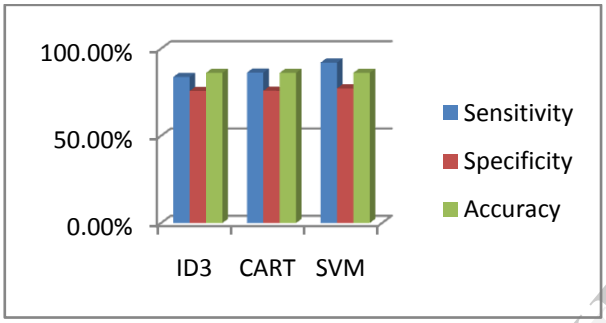

Fig 2.Shows sensitivity, specificity, accuracy and error rate for different classification techniques in 3-D Column chart format.

As per results , error rates for Decision Tree algorithm ID3 , CART and SVM are 02.756, 0.2755 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 idealcase.

Table 3.True Positive Rate and False Positive Rate

Classification models	True Positive Rate	False Positive Rate
ID3	**0.8375**	**0.2526**
CART	**0.8625**	**0.2410**
SVM	**0.9000**	**0.2279**

VII. CONCLUSION

There are different data mining techniques that can be usedfor the identification and prevention of cardiovascular diseaseamong patients. In this paper four classification techniques in datamining to predict cardiovascular disease in patients are compared: Decision Tree algorithms ID3, CART and SVM. These techniques arecompared on basis of Sensitivity, Specificity, Accuracy, ErrorRate, True Positive Rate and False Positive Rate. Our studiesshowed that SVM classification model turned out to bebest classifier for cardiovascular disease prediction.
In futurewe intend to improve performance of these basic classificationtechniques by creating meta model which will be used to predictcardiovascular 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 forproviding 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

[1] Yanwei, X.; Wang, J.; Zhao, Z.; Gao, Y., "Combination data mining models with new medical data to predict outcome of coronary heart disease". Proceedings International Conference on Convergence Information Technology 2007,pp. 868 – 872.
[2] L. Breiman, J. Friedman, R. Olshen, And C. Stone. Classification And Regression Trees. Wadsworth Int. Group, 1984.
[3] Detrano, R.; Steinbrunn, W.; Pfisterer, M., "Internationalapplication of a new probability algorithm for the diagnosisof coronary artery disease". American Journal of Cardiology,Vol. 64, No. 3, 1987, pp. 304-310.
[4] Yao, Z.; Lei, L.; Yin, J., "R-C4.5 Decision tree model andits applications to health care dataset". Proceedings ofInternational Conference on Services Systems and ServicesManagement 2005, pp. 1099-1103.
[5] Das, R.; Abdulkadir, S. (2008). "Effective diagnosis of heartdisease through neural networks ensembles". Elsevier,2008.
[6] Colombet, I.; Ruelland, A.; Chatellier, G.; Gueyffier, F.2000). "Models to predict cardiovascular risk: comparison of CART, multilayer perceptron and logistic regression".Proceedings of AMIA Symp 2000, p 156-160.
[7] Quinlan, J.R., "Induction Of Decision Trees," Machine Learning. Vol. 1. 1986. 81-1s06.
[8] K. Gang, P.Y., S. Yong, C. Zhengxin, "Privacy-Preserving Data Mining Of Medical Data Using Data Separation-Based Techniques," Data Science Journal, 2007. 6.

[9] L. Cao, "*Introduction To Domain Driven Data Mining*," Data Mining For Business Applications, Pp. 3-10, Springer, 2009.

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

[11] Han, J.; Kamber, M., "Data Mining Concepts and Techniques".2nd Edition, Morgan Kaufmann, San Francisco.

[12] Palaniappan, S.; Awang, R., "Intelligent Heart DiseasePrediction System Using Data Mining Techniques". Proceedings of IEEE/ACS International Conference onComputer Systems and Applications 2008, pp. 108-115.