Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction

[1]Milan Kumari, [2]Sunila Godara
[1,2]Department of CSE, Guru Jambheshwar University of Science & Technology, Hisar, India

Abstract
Medical science industry has huge amount of data, but unfortunately most of this data is not mined to find out hidden information in data. Advanced data mining techniques can be used to discover hidden pattern in data. Models developed from these techniques will be useful for medical practitioners to take effective decision. In this research paper data mining classification techniques RIPPER classifier, Decision Tree, Artificial neural networks (ANNs), and Support Vector Machine (SVM) are analyzed on cardiovascular disease dataset. Performance of these techniques is 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 RIPPER, Decision Tree, ANN and SVM are 02.756, 0.2755, 0.2248 and 0.1588 respectively. Accuracy of RIPPER, Decision Tree, ANN and SVM are 81.08%, 79.05%, 80.06% and 84.12% respectively. Our analysis shows that out of these four classification models SVM predicts cardiovascular disease with least error rate and highest accuracy.

Keywords
heart disease, data mining techniques, RIPPER, decision tree, artificial neural networks, and support vector machine.

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 [1]:

- 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. For example, data mining methods comprise artificial neural networks and decision trees, and statistical techniques include linear regression and stepwise polynomial regression [2].

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 applied 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 base software 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].
E. Engin Avci and Ibrahim Turkoglu study an intelligent diagnosis system based on principle component analysis and ANFIS for the heart valve diseases [7].
F. Imran Kurt , Mevlut Ture , A. Turhan Kurum compare performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease [8].
G. The John Gennari's CLASSIT conceptual clustering system achieved a 78.9% accuracy on the Cleveland database [9].

III. CVD Prediction Models

Under this section we will discuss following data mining classification models to predict cardiovascular disease:

A. RIPPER

RIPPER stands for Repeated Incremental Pruning to Produce Error Reduction. This classification algorithm was proposed by William W Cohen.

It is based on association rules with reduced error pruning (REP), a very common and effective technique found in decision tree algorithms. In REP for rules algorithms, the training data is split into a growing set and a pruning set. First, an initial rule set is formed that is the growing set, using some heuristic method. This

overlarge rule set is then repeatedly simplified by applying one of a set of pruning operators typical pruning operators would be to delete any single condition or any single rule. At each stage of simplification, the pruning operator chosen is the one that yields the greatest reduction of error on the pruning set. Simplification ends when applying any pruning operator would increase error on the pruning set [10].

Here is algorithm:
Initialize RS = {}, and for each class from the less prevalent one to the more frequent one.

DO:

Building stage:
Repeat Grow phase and Prune phase until the description length(DL) of the ruleset and examples is 64 bits greater than the smallest DL met so far, or there are no positive examples, or the error rate >= 50%.

Grow phase:
Grow one rule by greedily adding antecedents (or conditions) to the rule until the rule is perfect (i.e. 100% accurate). The procedure tries every possible value of each attribute and selects the condition with highest information gain: p(log(p/t)-log(P/T)).

Prune phase:
Incrementally prune each rule and allow the pruning of any final sequences of the antecedents.

Optimization stage:
After generating the initial ruleset {Ri}, generate and prune two variants of each rule Ri from randomized data using procedure Grow phase and Prune phase. But one variant is generated from an empty rule while the other is generated by greedily adding antecedents to the original rule. Then the smallest possible DL for each variant and the original rule is computed. The variant with the minimal DL is selected as the final representative of Ri in the ruleset. After all the rules in {Ri} have been examined and if there are still residual positives, more rules are generated based on the residual positives using Building Stage again.

Delete the rules from the ruleset that would increase the DL of the whole ruleset if it were in it and add resultant ruleset to RS.

ENDDO

B. Decision Tree

Decision trees are powerful classification algorithms. Popular decision tree algorithms include Quinlan's ID3, C4.5, C5, and Breiman et al.'s CART. As the name implies, this technique recursively separates observations in branches to construct a tree for the purpose of improving the prediction accuracy. In doing so, they use mathematical algorithms to identify a variable and corresponding threshold for the variable that splits the input observation into two or more subgroups. This step is repeated at each leaf node until the complete tree is constructed. The objective of the splitting algorithm is to find a variable-threshold pair that maximizes the homogeneity of the resulting two or more subgroups of samples [11]. The most commonly used mathematical algorithm for splitting includes Entropy based information gain (used in ID3, C4.5, C5), Gini index (used in CART), and Chi-squared test (used in CHAID).
Below Fig. 1 shows an example of decision tree on patient diagnosis. Here non-terminal nodes represent tests on one or more attributes and terminal nodes reflect decision outcomes. Decision tree generalizes following data: If a patient has swollen glands, the diagnosis is strep throat. If a patient does not have swollen glands and has fever, the diagnosis is cold. If a patient does not have swollen glands and does not have fever, the diagnosis is allergy.

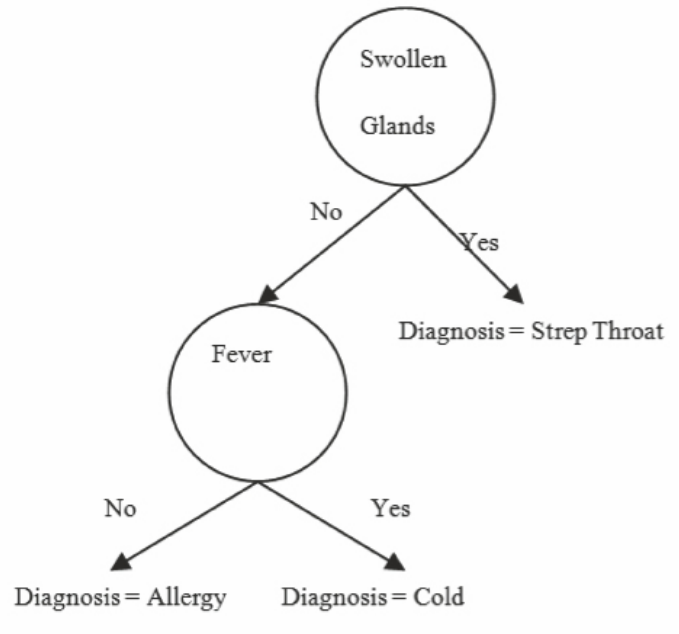

Fig. 1: Decision Tree

C. Artificial Neural Networks

Artificial neural networks (ANNs) are commonly known as biologically inspired, highly sophisticated analytical techniques, capable of modeling extremely complex non-linear functions. ANNs are analytic techniques modeled after the processes of learning in the cognitive system and the neurological functions of the brain and capable of predicting new observations from other observations (on the same or other variables) after executing a process of so-called learning from existing data. One of popular ANN architecture is called multi-layer perceptron (MLP) with back-propagation (a supervised learning algorithm). The MLP is known to be a powerful function approximator for prediction and classification problems. It is arguably the most commonly used and well-studied ANN architecture. Given the right size and the structure, MLP is capable of learning arbitrarily complex nonlinear functions to arbitrary accuracy levels. The MLP is essentially the collection of nonlinear neurons (perceptrons) organized and connected to each other in a feedforward multi-layer structure. Fig. 2 shows MLP feed forward Neural Network. This model is capable of mapping set of input data into a set of appropriate output data. The primary task of neurons in input layer is the division of input signal xi among neurons in hidden layer. Every neuron j in hidden layer adds up its input signals xi once it weights them with the strength of the respective connections wji from the input layer and determines its output yj as a function f of the sum, given as

$$Yj = f(\Sigma\ Wji\ Xi)$$

At this instant it is possible for f to be a simple threshold function such as a sigmoid, or a hyperbolic tangent function. The output of neurons in the output layer is determined in an identical fashion [12].

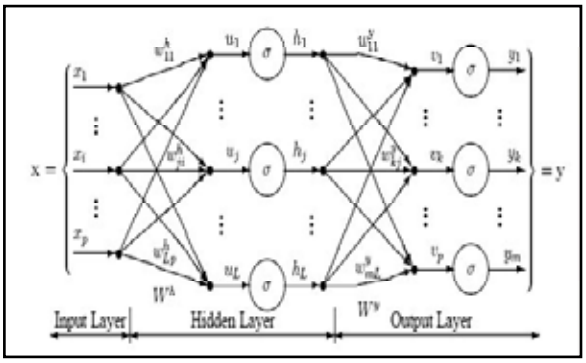

Fig. 2: MLP

The back-propagation algorithm can be employed effectively to train neural networks; it is widely recognized for applications to layered feed-forward networks, or multi-layer perceptrons. The back-propagation algorithm is capable of adjusting the network weights and biasing values to reduce the square sum of the difference between the given output (X) and an output values computed by the net (X ') with the aid of gradient decent method as follows:

$SSE = ½ N \Sigma (X-X')2$

Where N is the number of experimental data points utilized for the training.

D. 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 non-linear mapping to a sufficiently high dimension, data from two classes can always be separated by a hyperplane. The SVM find this hyperplane using support vectors and margins [13]. SVM performs classification tasks by maximizing the margin separating both classes while minimizing the classification errors. Fig 3 shows SVM topology in hyperspace:

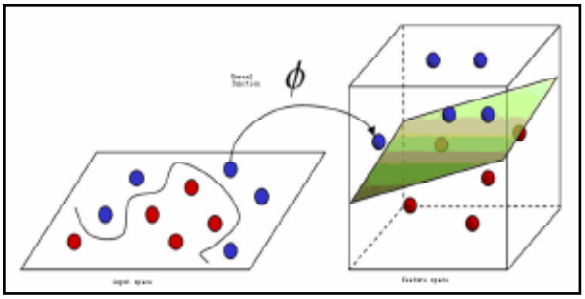

Fig. 3: SVM topology

IV. Data Source

To compare these data mining classification techniques Cleveland cardiovascular disease dataset from UCI repository was used. The dataset has 14 attributes and 303 records. Table 1 below lists these attributes:

Table 1: Attributes of Cardiovascular disease dataset

No.	Name	Description
1	Age	Age in years
2	Sex	1 = male, 0 = female
3	Cp	Chest pain type (1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptomatic)
4	Trestbps	Resting blood sugar (in mm Hg on admission to hospital)
5	Chol	Serum cholesterol in mg/dl
6	Fbs	Fasting blood sugar > 120 mg/dl (1 = true, 0 = false)
7	Restecg	Resting electrocardiographic results (0 = normal, 1 = having ST-T wave abnormality, 2 = left ventricular hypertrophy)
8	Thalach	Maximum heart rate
9	Exang	Exercise induced angina
10	Oldpeak	ST depression induced by exercise relative to rest
11	Slope	Slope of the peak exercise ST segment (1 = upsloping, 2 = flat, 3 = downsloping)
12	Ca	Number of major vessels colored by fluoroscopy
13	Thal	3 = normal, 6 = fixed defect, 7 = reversible defect
14	Num	Class (0 = healthy, 1 = have heart disease)

V. Results

These data mining classification model were developed using data mining 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 classification techniques RIPPER, Decision Tree, Artificial Neural Networks (ANNs) and Support Vector Machine (SVM) were applied.

A distinguished confusion matrix was obtained to calculate sensitivity, specificity and accuracy. Confusion matrix is a matrix representation of the classification results. Table 2 shows confusion matrix.

Table 2: Confusion Matrix

	Classified as Healthy	Classified as not healthy
Actual Healthy	TP	FN
Actual not healthy	FP	TN

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,

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:
Sensitivity = TP / (TP + FN)
Specificity = TN / (TN + FP)
Accuracy = (TP + TN) / (TP + FP + TN + FN)

Table 3 shows sensitivity, specificity and accuracy for different classification techniques. Table 4 shows sensitivity, specificity, accuracy and error rate for different classification techniques in graphical format.

Table 3: Comparison of Data Mining Models

	Sensitivity	Specificity	Accuracy
RIPPER	86.25%	75.82%	81.08%
Decision Tree C4.5	83.12%	74.26%	79.05%
ANN (MLP)	83.75%	75.73%	80.06%
SVM	90.0%	77.20%	84.12%

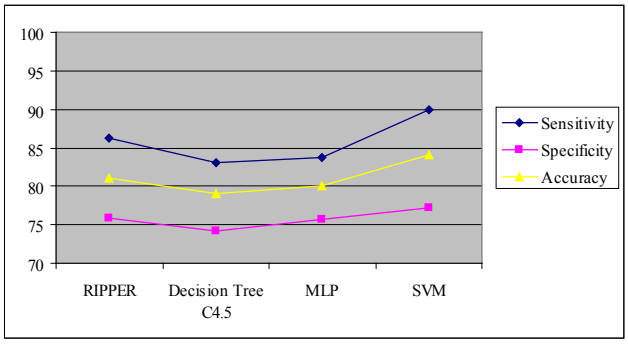

Fig. 4: Graphical representation of Sensitivity, Specificity and Accuracy

The error rate for RIPPER, Decision Tree, Artificial Neural Networks (ANNs) and Support Vector Machine (SVM) are 0.2756, 0.2755, 0.2248 and 0.1588 are respectively.

A Receiver Operating Characteristic (ROC) space is defined by False Positive Rate and True Positive Rate which shows relative trade-off between true positive and false positive.

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

Table 4 shows True Positive Rate and False Positive Rate for RIPPER, Decision Tree, Artificial Neural Networks (ANNs) and Support Vector Machine (SVM). Fig. 5 shows True Positive Rate and False Positive Rate for RIPPER, Decision Tree, Artificial Neural Networks (ANNs) and Support Vector Machine (SVM) in graphical format. The best possible prediction model will be at coordinate (0, 1) in graph on Fig. 5. This will represent 100% True Positive Rate and no False Positive Rate which will be ideal case.

Table 4: True Positive Rate and False Positive Rate

	True Positive Rate	False Positive Rate
RIPPER	0.8625	0.2410
Decision Tree C4.5	0.8312	0.2573
ANN (MLP)	0.8375	0.2426
SVM	0.9000	0.2279

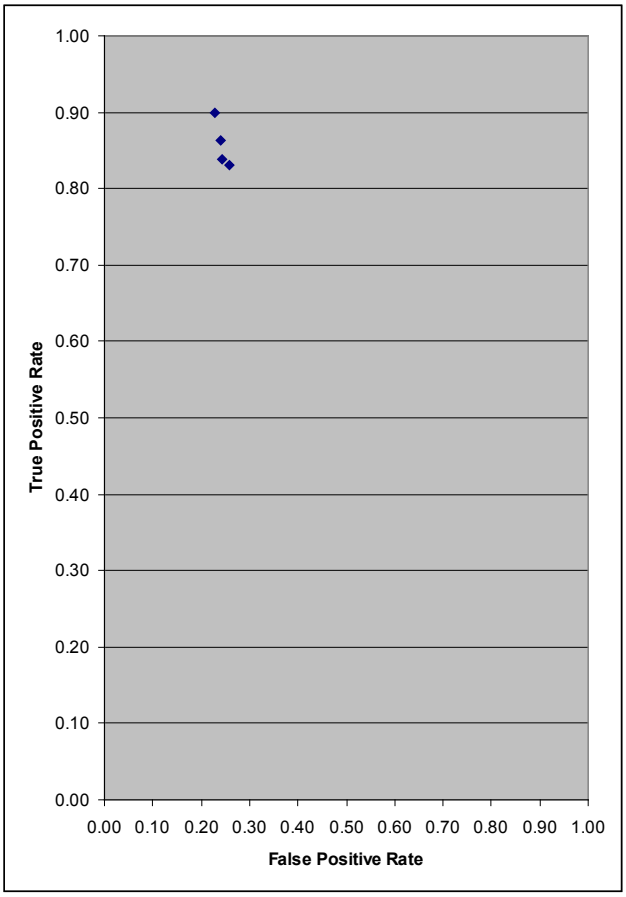

Fig. 5: Graphical representation of ROC space

Our results shows that out of RIPPER, Decision Tree, MLP and SVM models SVM outperforms others in all parameters Sensitivity, Specificity, Accuracy and Error Rates. Also, on ROC space point of SVM Model is closer to prefect point (0, 1) than other models which show SVM to be best predictor of cardiovascular disease.

VI. 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: rule based RIPPER techniques, decision tree, Artificial Neural Networks and Support Vector Machine. These techniques are compared on basis of Sensitivity, Specificity, Accuracy, Error Rate, True Positive Rate and False Positive Rate. Our studies showed that Support Vector Machine model turned out to be 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.

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] Khemphila, A.; Boonjing, V., "Comparing performance of logistic regression, decision trees and neural networks for classifying heart disease patients". Proceedings of International Conference on Computer Information System and Industrial Management Applications 2010, pp. 193 – 198.

[3] Detrano, R.; Steinbrunn, W.; Pfisterer, M., "International application of a new probability algorithm for the diagnosis of 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 and its applications to health care dataset". Proceedings of International Conference on Services Systems and Services Management 2005, pp. 1099-1103.

[5] Das, R.; Abdulkadir, S. (2008). "Effective diagnosis of heart disease 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] Avci, E.; Turkoglu, I., "An intelligent diagnosis system based on principle component analysis and ANFIS for the heart valve diseases". Journal of Expert Systems with Application, Vol. 2, No. 1, 2009, pp. 2873-2878.

[8] Kurt, I.; Ture, M.; Turhan, A., "Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease". Journal of Expert Systems with Application, Vol. 3, 2008, pp. 366-374.

[9] Gennari, J., "Models of incremental concept formation". Journal of Artificial Intelligence, Vol. 1, 1989, pp. 11-61.

[10] Cohen, W., "Fast effective rule induction". Proceedings of International Conference on machine Learning 1995, pp. 1-10.

[11] Chau, M.; Shin, D., "A Comparative Study of Medical Data Classification Methods Based on Decision Tree and Bagging Algorithms". Proceedings of IEEE International Conference on Dependable, Autonomic and Secure Computing 2009, pp. 183-187.

[12] Patil, S.; Kumaraswamy, Y., "Intelligent and effective Heart Attack prediction system using data mining and artificial neural networks". European Journal of Scientific Research, Vol. 31, 2009, pp. 642- 656.

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

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

Milan Kumari received her MCA degree from Guru Jambheshwar University of Science & Technology, HISAR. She is pursuing her MTech degree in Computer Science & Engineering from Guru Jambheshwar University of Science & Technology, HISAR. Her research areas are Data Mining and Database Management System.

Ms Sunila Godara received MSc and MTech degree in Computer Science & Engg from Guru Jambheshwar University of Science & Technology, HISAR. She is working as Assistant Professor in Deptt of Computer Sc. & Engg, Guru Jambheshwar University of Science & Technology, HISAR. She has published more then 15 papers in national and international journals and conferences. Her research areas are Data Mining and Database Management System.