Review of Data Mining Classification Models in Cardiovascular Disease Diagnosis

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

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 review paper data mining classification techniques RIPPER classifier, Decision Tree, Artificial neural networks (ANNs), and Support Vector Machine (SVM) are reviewed. In our research work we will compare these techniques through lift chart, error rate and will determine sensitivity, specificity, and accuracy of these data mining techniques.

Keywords

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

1. 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 heart 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 heart 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.

2. BACKGROUND

Up to now, several studies have been reported that have focused on cardio vascular disease diagnosis. These studies have applied different approaches to the given problem and achieved high classification accuracies, of 77% or higher, using the dataset taken from the UCI machine learning repository. 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].

Cvd Prediction Models

Under this section following data mining classification models to predict cardiovascular disease are discussed:

H. 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

I. 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.

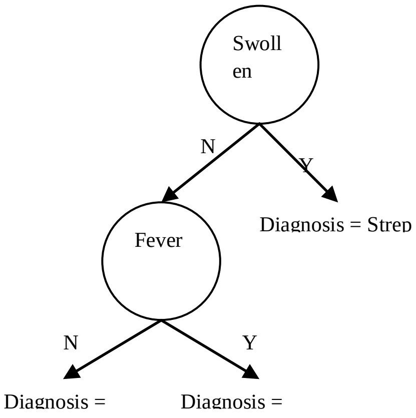

Fig 1: Decision Tree

J. 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].

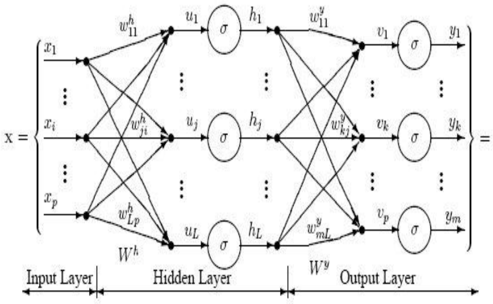

Fig 2: MLP

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.

K. 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:

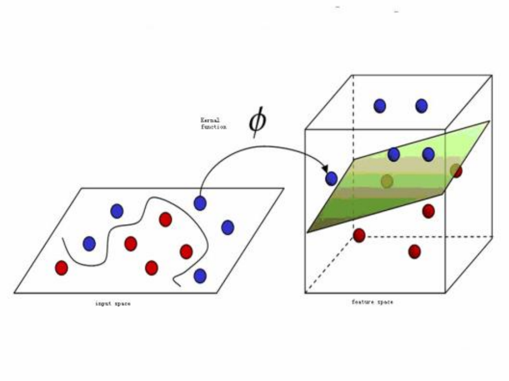

Fig 3: SVM topology

3. ANALYSIS

In this section comparative analysis of data mining classification techniques is explained.

Advantage of RIPPER is that it is fast rule based classifier because it is particularly capable of learning rules from multi-model data sets and it provides the result that are easy to interpret. Additional benefit of this method is that classification is computationally inexpensive. The major drawback of RIPPER is its greedy optimization algorithm and its tendency to overfit the training data at times.

Decision tree results are easier to read and interpret. The drill through feature to access detailed patients profiles is only available in decision trees. But when a decision tree is built many of the branches will reflect anomalies in training data to noise or outliers. This problem is called overfitting.

Advantage of neural networks is that it draws consistent conclusion and it can be built and evaluated using a large number of cases. Neural networks model the relationship between the possible signs and symptoms; and the diagnosis is the fact that this relationship does not have to be a linear one. However, despite the wide interest in the application of neural networks there are a number of limitations that make the introduction of these tools to daily practice difficult. First because of the black box nature of neural networks, it is difficult to explain. The relationship between attributes produced by Neural Network is more difficult to understand. Second problem is how to validate a trained neural network.

Advantage of support vector machine is that it is highly accurate, owning to their ability to model complex non-linear decision boundaries. Also, Support Vector Machine provides a compact description of a learned model. Support Vector Machine is much less prone to overfitting than other data mining classification methods. SVM can be used for prediction as well as classification. But major drawback of SVM is that training time of even the fastest SVM can be extremely slow.

4. 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 analyzed: rule based RIPPER techniques, decision tree, Artificial Neural Networks and Support Vector Machine. These techniques are analyzed on basis of their structure and efficiency. 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.

5. REFERENCES

[1] Y. Xing, J. Wang, Z. Zhao, and Y. Gao, "Combination data mining models with new medical data to predict outcome of coronary heart disease", in Proc of International Conference on Convergence Information Technology, 2007 p. 868 – 872.

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

[3] R. Detrano, W. Steinbrunn, M. Pfisterer, J. Schmid, and S. Sandhu, "International application of a new probability algorithm for the diagnosis of coronary artery disease", American Journal of Cardiology, Vol. 64, pp. 304-310, 1987.

[4] Z. Yao, L. Lei, and J. Yin, "R-C4.5 Decision tree model and its applications to health care dataset", in Proc of International Conference on Services Systems and Services Management, 2005, p. 1099-1103.

[5] R. Das and S. Abdulkadir, "Effective diagnosis of heart disease through neural networks ensembles", Elsevier, 2008.

[6] I. Colombet, A. Ruelland, G. Chatellier, and F. Gueyffier, "Models to predict cardiovascular risk: comparison of CART, multilayer perceptron and logistic regression", in Proc of AMIA Symp, 2000, p. 156-160.

[7] E. Avci and I. Turkoglu, "An intelligent diagnosis system based on principle component analysis and ANFIS for the heart valve diseases", Journal of Expert Systems with Application, Vol. 2, pp. 2873-2878, 2009.

[8] I. Kurt, M. Ture, and A. Kurum, "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, pp. 366-374, 2008.

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

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

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

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

[13] J. Han and M. Kamber, Data Mining Concepts and Techniques, 2nd Edition, Morgan Kaufmann, San Francisco, 2006.

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

6. AUTHOR'S PROFILE

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.

Milan Kumari received 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.