A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL)

Mythili T., Dev Mukherji, Nikita Padalia, and Abhiram Naidu
School of Computing Sciences and Engineering, VIT University
Vellore – 632014, Tamil Nadu, India

ABSTRACT

The early prognosis of cardiovascular diseases can aid in making decisions to lifestyle changes in high risk patients and in turn reduce their complications. Research has attempted to pinpoint the most influential factors of heart disease as well as accurately predict the overall risk using homogenous data mining techniques. Recent research has delved into amalgamating these techniques using approaches such as hybrid data mining algorithms. This paper proposes a rule based model to compare the accuracies of applying rules to the individual results of support vector machine, decision trees, and logistic regression on the Cleveland Heart Disease Database in order to present an accurate model of predicting heart disease.

Keywords

Heart disease, support vector machine (SVM), logistic regression, decision trees, rule based approach

1. INTRODUCTION

Data mining (DM) is the extraction of useful information from large data sets that results in predicting or describing the data using techniques such as classification, clustering, association, etc. Data mining has found extensive applicability in the healthcare industry such as in classifying optimum treatment methods, predicting disease risk factors, and finding efficient cost structures of patient care. Research using data mining models have been applied to diseases such as diabetes, asthma, cardiovascular diseases, AIDS, etc. Various techniques of data mining such as naïve Bayesian classification, artificial neural networks, support vector machines, decision trees, logistic regression, etc. have been used to develop models in healthcare research.

An estimated 17 million people die of cardiovascular diseases (CVD) every year [1]. Although such diseases are controllable, their early prognosis and a patient’s evaluated risk are necessary to curb the high mortality rates it presents. Common cardiovascular diseases include coronary heart disease, cardiomyopathy, hypertensive heard disease, heart failure, etc. Common causes of heart diseases include smoking, diabetes, lack of physical activity, hypertension, high cholesterol diet, etc.

Research in the field of cardiovascular diseases using data mining has been an ongoing effort involving prediction, treatment, and risk score analysis with high levels of accuracy. Multiple CVD surveys have been conducted with the most prominent one being the data set from the Cleveland Heart Clinic. The Cleveland Heart Disease Database (CHDD) [2] as such has been considered the de facto database for heart disease research. Recommending the parameters from this database, this paper proposes a framework to apply logistic regression, support vector machines, and decision trees to attain individual predictions which are in turn used in rule based algorithms. The result of each rule from this system is then compared on the basis of accuracy, sensitivity, and specificity.

The methodology aims to accomplish of two goals: the first is to primarily present a predictive framework for heart disease, and the second is to compare the efficiency of merging the outcomes of multiple models as opposed to using a single model.

2. LITERATURE SURVEY

Prediction of heart disease using data mining techniques has been an ongoing effort for the past two decades. Most of the papers have implemented techniques such as SVM, neural networks, regression, decision trees, naïve Bayesian classifiers, etc. on multiple databases of patients from around the world.

One of the bases on which the papers differ are the selection of parameters on which the methods have been applied. Many authors have specified different parameters and databases for testing the accuracies. Xing et al. [3] conducted a survey of 1000 patients, the results of which showed SVM to have 92.1% accuracy, artificial neural networks to have 91.0% and decision trees with 89.6% using TNF, IL6, IL8, HICRP, MPO1, TNI2, sex, age, smoke, hypertension, diabetes, and survival as the parameters. Similarly, Chen et al. [4] compared the accuracy of SVM, neural networks, Bayesian classification, decision tree and logistic regression. Considering 102 cases, SVM had the highest accuracy of 90.5%, neural networks 88.9%, Bayesian 82.2%, decision tree 77.9%, and logistic regression 73.9%.

Comparing the accuracies across multiple data sets with different parameters arrives at different results which do not provide a just basis for comparison. Realizing this, Soni et al. [5] listed most influential parameters as gender, smoking, overweight, alcohol intake, high salt diet, high saturated fat diet, exercise, sedentary lifestyle, hereditary, cholesterol, blood pressure, fasting blood sugar, and heart rate. More recently, Shouman et al. [6] cited the statistically identified risk factors to be age, blood pressure, cholesterol, smoking, total cholesterol, diabetes, hypertension, hereditary, obesity, lack of physical activity. The same paper also presented the Cleveland Heart Disease Database as the standard database for heart disease research as it has been widely accepted. As such, the CHDD has been employed for the method proposed in this paper and details of the parameters it involves have been elaborated in further sections.

Work on the CHDD can be dated since 1989, when Detrano et al. [7] used logistical regression to obtain 77% accuracy of prediction. The accuracies of different models on the CHDD have been tabulated in Table 1. A corresponding graph of the same has been presented in Figure 1.

What's noteworthy of these milestones is the improvement in the accuracy when using hybrid techniques such as that in Polat et al. [8], Ozsen and Gunes [9], Das et al. [10], and Muhammed's [11] CLIP4 ensemble. The improvement was also pointed out in the Shouman et al. [6], who also highlighted a necessity to conduct further research on hybrid techniques.

Table 1: Earlier methodologies and their accuracies applied to CHDD

Author (Ref #)	Technique	Accuracy
Detrano et al. [7]	Logistic regression	77%
Cheung [12]	C4.5 Naïve Bayes BNND BNNF	81.11% 81.48% 81.11% 80.96%
Polat et al. [13]	AIS	84.5%
Polat et al. [8]	Fuzzy-AIRS-KNN	87.0%
Ozsen and Gunes [9]	GA-AWAIS	87.0%
Tu et al. [14]	J4.8 Decision Tree Bagging Algorithm	78.9% 81.41%
Das et al. [10]	ANN ensembles	89.01%
Shouman et al. [15]	Nine Voting Equal Frequency Discretization Gain Ratio Decision Tree	84.1%
Muhammed [11]	CLIP3 CLIP4 CLIP4 ensemble	84.0% 86.1% 90.4%

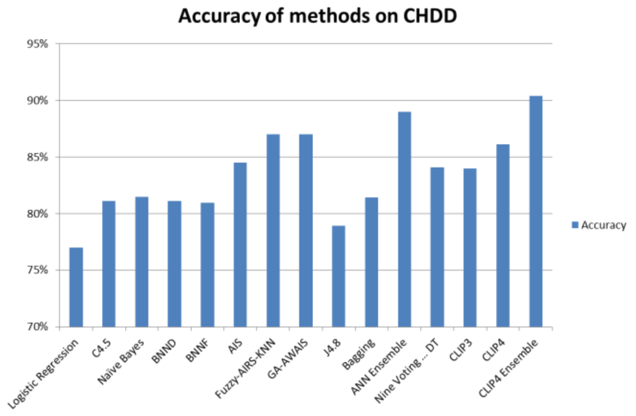

Figure 1: Performance Analysis

There is a need for prediction based on more sophisticated data mining models. A rule based approach is a commonly used technique that combines the results of multiple models. Rule based models such as C4.5 have been applied before, but never as a combination of multiple predictive models. As such, this paper presents a unique model to comparatively study the application of rule based algorithms to combinations of SVM, decision trees, and logistical regression.

3. PROPOSED FRAMEWORK

3.1 Introduction

For evaluation of risk of heart disease using a combination of models, this paper proposes the framework shown in Figure 2 on the next page. This approach is divided into six modules involving preprocessing, training, testing with individual models, application of rules, and finally, comparison of results and the prediction of heart disease. The modules have been described below.

3.2 Patient Database

Patient database is datasets collected from Cleveland Heart Disease Dataset (CHDD) available on the UCI Repository [11]. The 13 attributes considered are age: age, sex, chest pain type, trestbps (resting blood pressure), chol (serum cholesterol in mg/dl), FBS (fasting blood sugar > 120 mg/dl), restecg (resting electrocardiographic results), thalach (maximum heart rate achieved), exang (exercise induced angina), oldpeak (ST depression induced by exercise relative to rest), slope (the slope of the peak exercise ST segment), and CA (number of major vessels (0-3) colored by fluoroscopy). There are a total of 303 patient records in the database.

3.3 Data Preprocessing

This phase includes extraction of data from the Cleveland Heart Disease Dataset (CHDD) in a uniform format. The step involves transforming the data, which involves removal of missing fields, normalization of data, and removal of outliers. Out of the 303 available records, 6 tuples have missing attributes. These have been excluded from the data set. For SVM, data points were automatically centered at their mean and scaled to have unit standard deviation. No changes need be made to the data sets for decision trees or logistic regression.

3.4 Training the Models

Each of the three models has been trained using different methods. For decision trees, a node splitting criterion is required. The best split is one that splits the data into distinct groups. Purity is a measure used to evaluate a potential split. A split that divides an attribute into two distinct classes is the most pure. There are many different splitting criterions that can be used, such as the Gini Coefficient, or using statistical deviance. Gini Coefficient is the most commonly used splitting criterion which works on the population diversity of the attribute and thus splits it. As such, it is recommended that it be used as the splitting criterion for decision trees, although others may be used as well resulting in different accuracies.

For support vector machines, arguably the most efficient training method is through using K-fold cross validation, wherein the data can be trained set by dividing it into k blocks and averaging the results of the blocks. This method uses all the tuples to train the data, and then testing the data using one of the blocks. Generally, a 10-fold cross validation is used for training.

For logistic regression, the first step to training is to find the significant attributes by calculating their individual P-values. As a rule of thumb, if it is below 0.05, only then is the attribute significant. The Hosmer-Lemeshow test is also required to check for goodness fit of the model. The corresponding P-value must abide by a 5% level of significance in order to be a good fit model. Only after the model has been established is the data ready for testing.

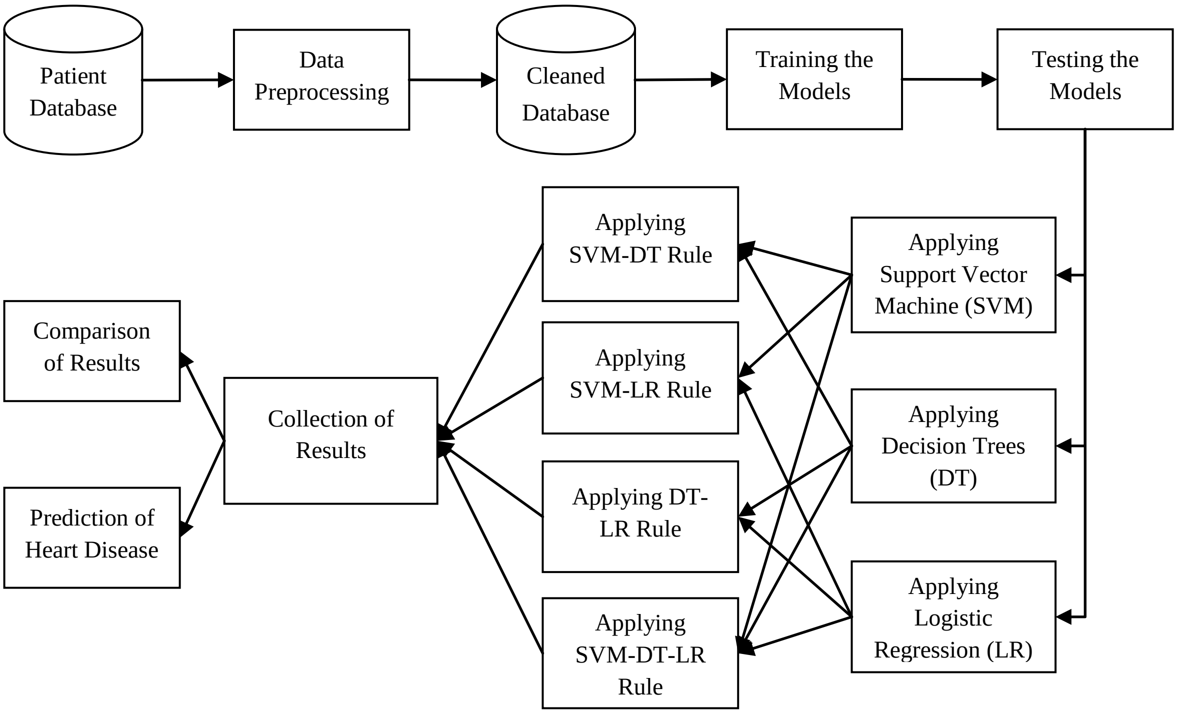

Figure 2: Proposed Framework

3.5 Testing the Models

3.5.1 Support Vector Machine

A support vector machine is a type of model used to analyze data and discover patters in classification and regression analysis. Support vector machine (SVM) is used when your data has exactly two classes. An SVM classifies data by finding the best hyper plane that separates all data points of one class from those of the other class. The larger margin between the two classes, the better the model is. A margin must have no points in its interior region. The support vectors are the data points that on the boundary of the margin. SVM is based on mathematical functions and used to model complex, and real world problems. SVM performs well on data sets that have many attributes, such as the CHDD.

Support Vector Machines map the training data into kernel space. There are many differently used kernel spaces – linear (uses dot product), quadratic, polynomial, Radial Basis Function kernel, Multilayer Perceptron kernel, etc. to name a few. In addition, there are multiple methods of implementing SVM, such as quadratic programming, sequential minimal optimization, and least squares. The challenging aspect of SVM is kernel selection and method selection such that your model is not over optimistic or pessimistic.

Considering that the CHDD has a large number of instances as well as features, it is arguable whether the kernel chosen is RBF or linear. Although the relation between the attributes and class labels are nonlinear, due to the large number of features, RBF kernel may not improve performance. It is recommended that both kernels be tested and the more efficient one be finally selected.

3.5.2 Decision Trees

A decision tree is a tool that uses classification or regression to predict a response to data. Classification is used when the features are grouped, and regression is used when the data is continuous. Decision tree is one of the main data mining methods. A decision tree is made of a root node, branches, and leaf nodes. To evaluate the data, follow the path from the root node to reach a leaf node.

Decision trees must be created using a purity index which will split the nodes as discussed in the training section. For the CHDD, each of the 297 tuples is evaluated down the decision tree and arrives at a positive or negative evaluation for heart disease. These are compared to the original decision parameter in the CHDD to check for false positives or false negatives, giving us the accuracy, specificity, and sensitivity of the model. The splitting criterion used is also indicative of the importance of each attribute.

3.5.3 Logistic Regression

Logistic regression is a type of regression analysis in statistics used for prediction of outcome of a categorical dependent variable (a dependent variable that can take a limited number of values) from a set of predictor or independent variables. In logistic regression the dependent variable is always binary (with two categories). Logistic regression is mainly used to for prediction and also calculating the probability of success. Logistic Regression involves fitting an equation of the form to the data:

$$\mathbf{Y = ß_0 + ß_1x_1 + ß_2x_2 + \ldots + ß_nx_n} - \textit{eq. 1}$$

The regression coefficients are usually estimated using maximum likelihood estimation. The maximum likelihood ratio helps to determine the statistical significance of independent variables on the dependent variables. The likelihood-ratio test assesses the contribution of individual predictors (independent variables). Then the probability (p) of each case is calculated using odds ratio,

$$P/(1-P) = e^{Y} \text{ – } \textit{eq. 2}$$

From this p–value is found out. This gives the probability or chance for the individual to have coronary heart disease.

3.6 Rule Based Algorithm

Rule based systems are essentially decision trees that use a small number of attributes for decision making. These are simple systems which are usually used to increase comprehension of knowledge patters. Rule based algorithms are indicative of trends in the features they consider and thus provides us with logical conclusions.

Rules are used to support decision making in classification, regression, and association tasks. Depending on the data, there are different types of rules that can be implemented such as classical prepositional logic (C-rules), association rules (A-rules), fuzzy logic (F-rules), M-of-N or threshold rules (T-rules), similarity or prototype-based rules (P-rules).

It is recommended that classification rule (C-rule) be used for this model. C-Rules are of the form of if-else ladders, and provide the simplest and most comprehensible way of expressing knowledge. These rules will account the result of each of the individual methods based on weight of the model which is dependent on the accuracy, specificity and sensitivity attained. It is hypothesized that a result with higher sensitivity and specificity but lower accuracy will be attained from the results of this model which is in itself, a highly efficient model.

3.7 Comparison of Results

The results obtained after applying the C-rule will be analyzed on the basis of sensitivity, specificity, and accuracy. From these, conclusions to the most effective model, the efficacy of conjoint models, and the final accuracy of the overall model can be drawn.

4. CONCLUSION AND FUTURE WORK

In conclusion, as identified through the literature review, there is a need for combinational and more complex models to increase the accuracy of predicting the early onset of cardiovascular diseases.

This paper proposes a framework using combinations of support vector machines, logistic regression, and decision trees to arrive at an accurate prediction of heart disease. Using the Cleveland Heart Disease database, this paper provides guidelines to train and test the system and thus attain the most efficient model of the multiple rule based combinations. Further, this paper proposes a comparative study of the multiple results, which include sensitivity, specificity, and accuracy. In addition, the most effective and most weighed model can be found. Further work involves development of the system using the mentioned methodologies and thus training and testing the system. Future work may also involve the development of a tool to predict the risk of disease of a prospective patient. The framework can also be extended for use on other models such as neural networks, ensemble algorithms, etc.

5. REFERENCES

[1] Mackay,J., Mensah,G. 2004 "Atlas of Heart Disease and Stroke" Nonserial Publication, ISBN-13 9789241562768 ISBN-10 9241562765.

[2] Robert Detrano 1989 "Cleveland Heart Disease Database" V.A. Medical Center, Long Beach and Cleveland Clinic Foundation.

[3] Yanwei Xing, Jie Wang and Zhihong Zhao Yonghong Gao 2007 "Combination data mining methods with new medical data to predicting outcome of Coronary Heart Disease" Convergence Information Technology, 2007. International Conference November 2007, pp 868-872.

[4] Jianxin Chen, Guangcheng Xi, Yanwei Xing, Jing Chen, and Jie Wang 2007 "Predicting Syndrome by NEI Specifications: A Comparison of Five Data Mining Algorithms in Coronary Heart Disease" Life System Modeling and Simulation Lecture Notes in Computer Science, pp 129-135.

[5] Jyoti Soni, Ujma Ansari, Dipesh Sharma 2011 "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction" International Journal of Computer Applications, doi 10.5120/2237-2860.

[6] Mai Shouman, Tim Turner, Rob Stocker 2012 "Using Data Mining Techniques In Heart Disease Diagnoses And Treatment" Electronics, Communications and Computers (JECECC), 2012 Japan-Egypt Conference March 2012, pp 173-177.

[7] Robert Detrano, Andras Janosi, Walter Steinbrunn, Matthias Pfisterer, Johann-Jakob Schmid, Sarbjit Sandhu, Kern H. Guppy, Stella Lee, Victor Froelicher 1989 "International application of a new probability algorithm for the diagnosis of coronary artery disease" The American Journal of Cardiology, pp 304-310.15

[8] Polat, K., S. Sahan, and S. Gunes 2007 "Automatic detection of heart disease using an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism and k-nn (nearest neighbour) based weighting preprocessing" Expert Systems with Applications 2007, pp 625-631.

[9] Ozsen, S., Gunes, S. 2009 "Attribute weighting via genetic algorithms for attribute weighted artificial immune system (AWAIS) and its application to heart disease and liver disorders problems" Expert Systems with Applications, pp 386-392.

[10] Resul Das, Ibrahim Turkoglub, and Abdulkadir Sengurb 2009 "Effective diagnosis of heart disease through neural networks ensembles" Expert Systems with Applications, pp 7675–7680.

[11] Lamia Abed Noor Muhammed 2012 "Using Data Mining technique to diagnosis heart disease" Electronics,

Communications and Computers (JEC-ECC), 2012 Japan-Egypt Conference on March 2012, pp 173-177.

[12] Cheung, N 2001 "Machine learning techniques for medical analysis" School of Information Technology and Electrical Engineering, B.Sc. Thesis, University of Queensland

[13] Polat, K., Sahan, S., Kodaz, H., Günes, S. 2005 "A new classification method to diagnose heart disease: Supervised artificial immune system". In Proceedings of the Turkish Symposium on Artificial Intelligence and Neural Networks (TAINN), pp 2186-2193.

[14] My Chau Tu, Dongil Shin, Dongkyoo Shin 2009 "Effective Diagnosis of Heart Disease through Bagging Approach" Biomedical Engineering and Informatics, 2009. BMEI '09. 2nd International Conference, pp 1- 4.

[15] Shouman, M., Turner, T. and Stocker.R 2011 "Using Decision Tree for Diagnosing Heart Disease Patients" Australasian Data Mining Conference (AusDM 11) Ballarat 2011, pp 23-30.