Study and Analysis of Prediction Model for Heart Disease: An Optimization Approach using Genetic Algorithm

P. Suresh and M.D. Ananda Raj

Department of Computer Science, Sowdeswari College, Salem, TamilNadu, India.
sur_bhoo71@rediffmail.com

Bharathiyanar University, Coimbatore, TamilNadu, India
anandmd@rediffmail.com

Abstract

In the medical field, the diagnosis of heart disease is the most difficult task. The diagnosis of heart disease is difficult as a decision relied on grouping of large clinical and pathological data. Due to this complication, the interest increased in a significant amount between the researchers and clinical professionals about the efficient and accurate heart disease prediction. In case of heart disease, the correct diagnosis in early stage is important as time is the very important factor. Heart disease is the principal source of deaths widespread, and the prediction of Heart Disease is significant at an untimely phase. Machine learning in recent years has been the evolving, reliable and supporting tools in medical domain and has provided the greatest support for predicting disease with correct case of training and testing. The main idea behind this work is to study diverse prediction models for the heart disease and selecting important heart disease feature using genetic algorithm. The optimized prediction models using genetic algorithm performance is better than traditional prediction models. The performances of the different prediction models retested with different heart disease data sets and validated with real-time data sets.
1. Introduction

Heart disease (cardiovascular disease) is a disease involving heart or blood vessels. According to WHO, heart disease is one of the leading causes for the universal death. According to a recent study, An estimated 17.7 million people died from CVDs in 2015, representing 31% of all worldwide deaths and If current trends are allowed to continue, 23.6 million people will die from heart disease in coming 2030 [1]. The main risk factors of Heart disease include diabetes, family history of Heart disease, smoking, obesity, high LDL cholesterol and low HDL cholesterol. Effective and efficient automated heart disease prediction systems can be beneficial in the healthcare sector for heart disease prediction. This automation will also reduce the number of tests to be taken by a patient. Hence, it will not only save cost but also the time for both, Doctors and patients.

The diagnosis of Heart disease in most cases depends on a complex combination of clinical and pathological data. Because of this complication, there exists a significant amount of interest among clinical professionals and researchers regarding the efficient and accurate prediction of Heart disease. According to the statistical data from WHO, one third population worldwide died from Heart disease; Heart disease is found to be the leading cause of death globally.

Computational biology is often applied through the process of translating biological knowledge into clinical practice, as well as in the understanding of biological phenomena from the clinical data. This process involves the development of a predictive model and the integration of distinctive types of data and knowledge for diagnostic purposes. Furthermore, this process requires the design and combination of distinct methodologies from statistical analysis and data mining.

Heart disease prediction system can assist medical professionals in predicting the state of Heart, based on the clinical data of patients provided into the system. Doctors may sometimes fail to take accurate decisions while diagnosing the Heart disease of a patient, therefore, Heart disease prediction systems, which use machine learning algorithms assist in such cases to get accurate results. Early detection and correct diagnosis of heart disease are needed using appropriate counseling and medicines.

Machine learning approaches have the potential to generate a knowledge-rich environment which can help to significantly help in Medical diagnosis [14]. Various supervised machine learning approaches like Decision tree, SVM, Naive Bayes, Random forest and Neural network can be used for prediction of heart disease, and all need to be analyzed in terms of performance of heart disease prediction.
2. Literature Review

The “medicine of the future” for cardiovascular disease prevention and prediction was proposed using ubiquitous computing and machine learning techniques. The process done through 5 stages. In sensor stage, the ECG signal is passed through an ADC and record into memory. In the second-stage, preprocessing is done by noise removal and then passed to annotation stage. In Annotation stage it found the fiducial points from the ECG signal. Then features are selected from ECG and then applied machine learning algorithms. ANN gives 90% of sensitivity and 90% specificity [1]. The current lists of free mobile Personal Health Record (mPHR) applications are identified and assessed the specification of selected mPHR. An extraction form has been defined to assist mPHR stakeholders who want to monitor their HR to select the mPHR that best fits their needs [2]. Cleveland clinic database [3] was used for heart disease prediction. Particle Swarm Optimization (PSO) was used along with Feed forward Neural Network for feature selection and classification. It gave the accuracy up to 91.94%.

A mobile application was proposed to provide support and monitor for coronary artery disease patients named ”Mobile Heart”. It informs the patient about his/her cardiac disease, risk factor profile, therapy, patient-tailored exercise training program etc. [4]. Childhood obesity was the highly risk factors for cardiovascular diseases. Survey using regression analysis shows that children who are overweight or obese with high waist circumference are increased risk of having the abnormal lipid profile [5]. This leads to cardiovascular diseases in the future. Adult with pulmonary arterial hypertension due to congenital heart disease prognosis is very difficult. A study conducted in 59 PAH-CHD patients developed a conclusion that Cystine C is a novel biomarker for predicting long-term mortality among these patients[6].

The correlation between Acute Heart Failure (AHF) and miRNA [7] was proposed. The selective findings suggest a time-dependent effect of circulating miRNAs and highlight the susceptibility to individual patient characteristics influencing potential relations between miRNAs and biomarkers. A study is conducted on 30 congenital heart disease patients to assess whether low radiation dose cardiac CT is feasible to evaluate ventricular systolic function in adults with congenital heart disease. Then they reached a conclusion that CT with the median radiation dose less than 1mSV was successful in evaluating ventricular systolic function [8].

Dynamic Bayesian network [9] with temporal abstraction was presented for prognosis of coronary heart disease. This model used incomplete data values for prognosis of CHD. Four techniques named random oversampling, SMOTE-N, a combination of SMOTE-N with clustering under sampling, and a combination of random oversampling with clustering under sampling used for resampling of imbalance in training data. An application named Heart-Trend
was proposed, and it was easy-to-use and affordable and can provide effective cardiac management, specifically towards preventive health care systems. It reduces the false alarm rate by analyzing the signal through intelligent blend of statistical method and signal processing techniques. The noise in ECG signal was reduced by cancellation circuit of contact resistance imbalance by Cds photo resistors.

The complex wavelet sub-band bi-spectrum (CWSB) features were proposed for detection and classification of myocardial infarction (MI), heart muscle disease (HMD) and bundle branch block (BBB) from 12-lead ECG. SU score based feature selection method was used for MI detection and SVM and ELM classifiers are used for detection and classification of cardiac ailment. Disease risk was predicted using fast Fourier transform and machine learning ensemble model. ANN, least square SVM and Naïve Bayes classifier collectively used to create ensemble model for prediction.

Most of the papers have implemented several Machine learning techniques for prediction of heart disease such as classification trees, naive bayes, neural network, support vector machine and K-nearest neighbor algorithm. Some of the papers also implemented feature selection techniques such as wavelet transformation, principle component analysis and information gain module for identifying important attributes for efficient performance of classifier for prediction of heart disease. One of the bases on which the papers differ is the selection of parameters on which the methods have been used. Many authors have specified different parameters and databases for testing the accuracies.

In this paper, decision support system for heart disease prediction is described. Integer coded Genetic Algorithm is used to find optimized feature subset for maximizing classification accuracy with a reduced number of features. Author suggested to do more intensive testing using a larger heart disease database to get more efficient results. Naive bayes classifier is used for prediction of the heart disease.

The main contribution was made to develop an Intelligent System using predictive modeling using Naïve Bayes. Both the author has used weka tool for prediction of heart disease using different machine learning algorithms. They have used support vector machine, Naive bayes, Decision tree, adaboost, random forest and multilayer perception. They also used information gain and genetic algorithms for feature selection and further to enhance classifier's accuracy. In this paper, Author has used principle component analysis, which finds the minimum number of attributes required to increase the accuracy of various supervised machine learning algorithms. The motivation behind this research is to analyze supervised machine learning algorithms to predict heart disease and compare the accuracy of naive bayes, support vector machine and decision tree. State-of-the-art literature on heart disease prediction has been done using data mining and machine learning methodologies. The approaches for predictive modeling are focused in...
particular. In this work, we have used popular machine learning algorithms random forest, naive bayes and decision tree (DT) and support vector machine (SVM). These classification algorithms are selected because they are very often used for research purposes and have been potential to give efficient results for heart disease prediction. Moreover, they use different approaches for generating the classification models, which increases the chances for finding a prediction model with high classification accuracy.

3. Analysis of Prediction Model

Predictive analysis solutions are delivered by using data mining technologies that use analytical models to discover hidden patterns and apply them to predict future trends and behaviors. Interest in the predictive analysis within the healthcare domain is increasing. Rapid decision making will increase in speed and effectiveness in the future as tools and information become more easily accessible. Healthcare analytics enables hospital and doctors to predict the patient's disease level and provides necessary advice and proper medicine to the patients. Predictive analytics and data visualization are becoming mainstream in the medical field which helps the doctors can take better-informed decisions that are based directly on data.

Random Forest

Random forest is an ensemble learning method for classification, it operates by constructing a multitude of decision trees at training time and outputting the class that have mean prediction

Support Vector Machine

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyper plane. In other words, given labeled training set (supervised learning), the algorithm outputs an optimal hyper plane which categorizes new examples. Binary SVMs are classifiers which discriminate data points of two categories. Each data object (or data point) is represented by an n-dimensional vector. Each of these data points belongs to only one of two classes. A linear classifier separates them with a hyper plane. There exist many hyper planes which can separate the data samples. In order to achieve maximum separation between the two classes, SVM picks the hyper plane which has the largest margin. The margin is the summation of the shortest distance from the separating hyper plane to the nearest data point of both categories. Such a hyper plane is likely to generalize better, meaning that the hyper plane correctly classify unseen or testing data points. SVM does the mapping from input space to feature space to support nonlinear classification problems. The kernel trick is helpful for doing this which makes a linear classification in the feature space equivalent to nonlinear classification in the original space. SVMs do these by mapping input vectors to a higher dimensional space where a maximal separating hyper plane is constructed.
**Naive Bayes Classifier**

Naive Bayes is a classification algorithm based on Bayes' theorem. Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods. It assumes that the presence or absence of particular feature of a class is unrelated to the presence or absence of any other feature. It is based on conditional probabilities. In this model, classifiers assume that the value of a particular attribute is independent of the value of any other attribute, in the class variable. An advantage of the Naive Bayes classifier is that it requires only a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Since independent variables are assumed, only the variances of the variables for each class need to be determined. It can be used for both binary and multiclass classification problems.

**Decision Tree**

Decision tree learning is one of the most widely used and practical machine learning method for inductive inference. It is a method for approximating discrete valued functions that is robust to noisy data and capable of learning disjunctive expressions. This learned function is represented by a decision tree. Tree models where the target variable can take a finite set of values are called classification trees. In these tree structures, leaves represent classification results and branches represent conjunctions of features that point to those classification results. Decision tree is a recursive structure for expressing classification rules that makes use of the divide and conquer paradigm by splitting the subsets according to attribute values, thereby seeking the best quality of a subset so as to make it a tree branch, and proceeding in the same manner for the other subsets. The importance of decision trees lies in their ability to build interpretable models, which is a decisive factor for implementation.

**4. Optimized Prediction Model**

An evolutionary algorithm (EA) is a subset of evolutionary computation, a generic population-based metaheuristic optimization algorithm. Evolutionary algorithms often perform well approximating solutions to all types of problems because they ideally do not make any assumption about the underlying fitness landscape. The evolutionary algorithms use the three main principles of the natural evolution: reproduction, natural selection and diversity of the species, maintained by the differences of each generation with the previous. The evolutionary algorithms are used to optimize the model for better performance. Genetic Algorithm (GA) is one category of evolutionary algorithm. The GA algorithms are direct, parallel, stochastic method for global search and optimization, which imitates the evolution of the living beings. Genetic Algorithms works with a set of individuals, representing possible solutions of the task. The selection principle is applied by using a criterion, giving an
evaluation for the individual with respect to the desired solution. The best-suited individuals create the next generation. Genetic algorithms are commonly used to generate high-quality solutions to optimization and search problems by relying on bio-inspired operators such as mutation, crossover and selection.

In this research work various prediction models are studied and analyzed with heart disease data sets. The prediction models are optimized using genetic algorithm and it increases the performance of predictive models.

![Optimized Prediction Model for Heart Disease](image)

Fig. 1: Optimized Prediction Model for Heart Disease

The data preprocessing techniques have been applied to remove noisy and missing values present in the data set. In this research work, genetic algorithm is used to select important feature from the large amount of heart disease data set. The K–Cross validation methodology is applied to produce balanced training and testing data set. The various classification algorithms are developed to predict the heart disease from various data sets. The performance of the models are compared and validated with real time data. The figure 1 shows the overall flow of optimized prediction model for heart disease.

5. **Experiment and Result Analysis**

The idea of this research work is to study the various prediction models for heart disease and improve the performance of the model by using genetic algorithm. The performances of the models are tested with four region heart disease data sets such as Cleaveland, Hungarian, Lonf-beach and Switzerland datasets. The dataset used in this work is obtained from UCI machine learning repository [21].Heart Disease database contains 76 attributes, but mostly all published experiments refer to using a subset of 14 from the 76 attribute for the prediction. The proposed work considered risk factors hypertension and family history as predictor and used 16 attributes for efficient heart disease prediction. Risk factors such as smoking, obesity, diabetes are having missing values in
most of the data. Sample data set with the list of attributes and their values are shown in Table I and the number of instances available for each of the datasets is shown in Table II.

### Table I: Sample Heart Disease Data Set

<table>
<thead>
<tr>
<th>Sr No.</th>
<th>Description</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Age</td>
<td>Continuous Age in years</td>
</tr>
<tr>
<td>2</td>
<td>Sex</td>
<td>Discrete value 1: Male value 0: Female</td>
</tr>
<tr>
<td>3</td>
<td>Chest Pain (CP)</td>
<td>Discrete 1 = typical angina 2 = atypical angina 3 = non-anginal pain 4 = asymptomatic</td>
</tr>
<tr>
<td>4</td>
<td>Trestbps</td>
<td>Continuous Resting blood pressure (in mmHg)</td>
</tr>
<tr>
<td>5</td>
<td>htn</td>
<td>Discrete 1=presence 0=absence</td>
</tr>
<tr>
<td>6</td>
<td>chol</td>
<td>Continuous Serum cholesterol in mg/dl</td>
</tr>
<tr>
<td>7</td>
<td>Fbs</td>
<td>Discrete Fasting blood sugar ¿ 120 mg/dl : 1 = true 0 = false</td>
</tr>
<tr>
<td>8</td>
<td>Fmchis</td>
<td>Discrete 1=presence 0=absence</td>
</tr>
<tr>
<td>9</td>
<td>Restecg</td>
<td>Discrete Resting result: 0 = normal 1 = having ST-T abnormality 2 = showing probable or define left ventricular hypertrophy by Estecriteria</td>
</tr>
<tr>
<td>10</td>
<td>Thalach</td>
<td>Continuous Maximum heart rate achieved</td>
</tr>
<tr>
<td>11</td>
<td>Exang</td>
<td>Discrete Exercise induced angina: 1 = yes, 0 = no</td>
</tr>
<tr>
<td>12</td>
<td>Oldpeak</td>
<td>Continuous Depression induced by exercise relative to rest</td>
</tr>
<tr>
<td>13</td>
<td>slope</td>
<td>Discrete The slope of the peak exercise segment: 1 = up sloping 2 = flat 3 = down sloping</td>
</tr>
<tr>
<td>14</td>
<td>ca</td>
<td>Discrete Number of major vessels colored by fluoroscopy that ranged between 0 to 3</td>
</tr>
<tr>
<td>15</td>
<td>Thal</td>
<td>Discrete 3 = normal 6 = fixed defect 7 = reversible defect</td>
</tr>
<tr>
<td>16</td>
<td>num</td>
<td>Discrete 0 = healthy 1 = patient who is subject to possible heart disease(1,2,3,4)</td>
</tr>
</tbody>
</table>

### Table II: Number of Instance in Various Heart Disease Dataset

<table>
<thead>
<tr>
<th>Class</th>
<th>Cleveland</th>
<th>Hungarian</th>
<th>LongBeach</th>
<th>Switzerland</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>155</td>
<td>180</td>
<td>48</td>
<td>8</td>
</tr>
<tr>
<td>1</td>
<td>49</td>
<td>35</td>
<td>51</td>
<td>40</td>
</tr>
<tr>
<td>2</td>
<td>30</td>
<td>26</td>
<td>36</td>
<td>29</td>
</tr>
<tr>
<td>3</td>
<td>32</td>
<td>27</td>
<td>40</td>
<td>23</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
<td>15</td>
<td>9</td>
<td>5</td>
</tr>
</tbody>
</table>

The presence and absence of heart disease are predicted using various classification models. This experiment is done for all four regions. The performance of the prediction models are measured using various measures such as accuracy, sensitivity and specificity. All the prediction models
provide 80% and above accuracy above for cleavel and data set. Figure 2 shows the accuracy, sensitivity specificity of the all the prediction models used for experiments.

Fig. 2: Optimized Prediction Model for Heart Disease

The feature selection is performed using genetic algorithm. The attributes "age", "cp", "trestbps", "htn", "chol", "fbs", "thalach", "exang", "oldpeak" are selected as important attribute from the cleaveland data set. The incorporation of genetic algorithm increases the prediction model performances as shown in the figure 2.

Fig. 3: Performance of Prediction Model for Hungarian Dataset

Hungarian dataset. The attributes "htn, chol, restecg, thalach, exang, oldpeak" are selected as important attributes based on genetic algorithm. The naïve bayes prediction model provides better performance than other prediction models.
Figure 4: Performance of Prediction Model for Long_Beach Dataset

Figure 4 shows the accuracy, sensitivity, specificity and area under the curve for Long_beach dataset.

The attributes “cp”, ”chol”, ”fbs”, ”fmhis”, ”restecg”, ”oldpeak” are selected as important attributes based on genetic algorithm. The Naive bayes and random forest are the best among all four classifier in terms of accuracy and area under the curve.

Figure 5: Performance of Prediction Model for Switzerland dataset

Figure 5 shows the accuracy, sensitivity, specificity and area under the curve for Long_Beach dataset. The attributes ”trestbps”, ”restecg”, ”cp”, are selected as important attributes based on genetic algorithm. For Switzerland dataset all classifiers provides same accuracy but in terms of ROC curve area, random forest and naive bayes classifiers provides better performance than other prediction models. Since Long beach and Switzerland data sets have so many missing values, as a result we have very few instances of absence class of heart disease and got very less sensitivity.
6. Conclusion

The diagnosis of heart disease is difficult as a decision relied on grouping of large clinical and pathological data. The main idea behind this work is to study diverse prediction models for the heart disease and selecting important heart disease feature using genetic algorithm. In this work, different prediction models were studied and the experiments are conducted to find the best classifier for predicting the heart disease. Four classifiers Random Forest, NaiveBayes, Decision Tree, Support Vector Machine were used for prediction of patients with heart diseases. Observation shows that in most of the cases Naive bayes classifier performance is having more accuracy compared with other three classification methods with respect to all the dataset. The best dataset is cleaveland dataset as it has very less missing values and it is having all 16 attribute available as predictors. The other observation from this study is that, the genetic algorithm feature selection technique suggests that the most important attributes for heart diseases are cp (Chest pain), Chol (The slope of the peak exercise segment), trestbps (resting blood pressure), Exang (Exercise induced angina), and Restecg (Resting electrocardiographic). The result also shows that integration genetic algorithm with prediction models improves the performances of the models.

References


[3] Majid Ghonjifeshki and Omid Sojoodishijani, "Improving the Heart Disease Diagnosis by Evolutionary Algorithm of PSO and Feed Forward Neural Network"

[4] I. Frederix, S. Sankaran, K. Coninx, and P. Dendale, "MobileHeart, a mobile smartphone-based application that supports and monitors coronary artery disease patients during rehabilitation"


[7] Eline L. Vegter , Daniela Schmitter , Yanick Hagemeijer et.al "Use of biomarkers to establish potential role and function of circulating microRNAs in acute heart failure", International Journal of Cardiology


[12] Rajesh Kumar Tripathy, Samarendra Dandapat, "Automated detection of heart ailments from 12-lead ECG using complex wavelet sub-band bi-spectrum features"

[13] “Coupling a Fast Fourier Transformation With a Machine Learning Ensemble Model to Support Recommendations for Heart Disease Patients in a Telehealth Environment”, JI ZHANG1, RAID LUAIBI LAFTA et. al


