Heart Disease Prediction System Using Data Mining and Hybrid Intelligent Techniques: A Review

J.Vijayashree and N.Ch.SrimanNarayanalyengar*
School of Computer Engineering, VIT University, Vellore-632014, TN, India
vijayashree.j@vit.ac.in; nchsniyr@vit.ac.in

Abstract

Heart disease is one of the main sources of demise around the world and it is imperative to predict the disease at a premature phase. The computer aided systems help the doctor as a tool for predicting and diagnosing heart disease. The objective of this review is to widespread about Heart related cardiovascular disease and to brief about existing decision support systems for the prediction and diagnosis of heart disease supported by data mining and hybrid intelligent techniques.

Keywords: Cardiovascular Disease, Decision Support System, Data Mining, Hybrid Intelligent System

1. Introduction

The heart is a most significant muscular organ in humans, which pumps blood through the blood vessels of the circulatory system[1]. Human life is dependent on the proper functioning of heart. Improper functioning of heart will influence other parts of human body like brain, kidney etc. If the blood circulation in body is inefficient, it affects both heart and brain. Generally blood arrest in heart is called as attack and blood arrest in brain is called as stroke. Human life is absolutely reliant on the efficient working of the heart and brain. The rest of this paper is presented as follows: Section 2 describes the cardiovascular disease and its pervasiveness. Section 3 describes the advantage of decision support system for the prediction of heart disease. Section 4 and 5 describes various data mining and hybrid intelligent techniques used for the prediction of heart disease.

2. Cardiovascular Disease

Cardiovascular heart disease is one of the principal reasons of death for both men and women. The term heart disease relates to a number of medical conditions related to heart which define the irregular health conditions that directly stimulate the heart and all its parts. Different types of heart related cardiovascular diseases along with description are given in Table1.

<table>
<thead>
<tr>
<th>Heart-related cardiovascular diseases</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acute coronary syndromes</td>
<td>Blood-supply to the heart muscle is swiftly obstructed</td>
</tr>
<tr>
<td>Angina</td>
<td>Chest pain due to a lack of blood to the heart muscle</td>
</tr>
<tr>
<td>Arrhythmia</td>
<td>Atypical heart rhythm</td>
</tr>
<tr>
<td>Cardiomyopathy</td>
<td>Heart muscle disease</td>
</tr>
<tr>
<td>Congenital heart disease</td>
<td>Heart disfigurements that are present at birth</td>
</tr>
<tr>
<td>Coronary heart disease</td>
<td>Arteries supplying blood to heart muscle becomes</td>
</tr>
</tbody>
</table>

ISSN: 2233–7849 IJBSBT
Copyright © 2016 SERSC
Table 2. Risk Factors and Symptoms of Heart Attack

<table>
<thead>
<tr>
<th>Risk factors</th>
<th>Symptoms of Heart Attack</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Chest Discomfort</td>
</tr>
<tr>
<td>Angina</td>
<td>Coughing</td>
</tr>
<tr>
<td>Blood cholesterol levels</td>
<td>Nausea</td>
</tr>
<tr>
<td>Diabetes</td>
<td>Vomiting</td>
</tr>
<tr>
<td>Diet</td>
<td>Crushing chest pain</td>
</tr>
<tr>
<td>Genes</td>
<td>Dizziness</td>
</tr>
<tr>
<td>Hypertension</td>
<td>Dyspnoea (shortness of breath)</td>
</tr>
<tr>
<td>Obesity</td>
<td>Restlessness</td>
</tr>
<tr>
<td>Physical Inactivity</td>
<td></td>
</tr>
<tr>
<td>Smoking</td>
<td></td>
</tr>
<tr>
<td>Work stress</td>
<td></td>
</tr>
</tbody>
</table>

2.1. Prevalence of Heart Disease

According to World life expectancy, India ranked 39th position of all the countries in the world suffering from coronary heart disease. In India the death rate per 100,000 is 138.36. Population suffering from Coronary heart disease in India by age, gender and region is presented in table 3.

Table 3. Coronary Heart Disease in India by Age, Gender and Region

<table>
<thead>
<tr>
<th>Year/Age</th>
<th>20-29</th>
<th>30-39</th>
<th>40-49</th>
<th>50-59</th>
<th>Male</th>
<th>Female</th>
<th>Urban</th>
<th>Rural</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>4.51</td>
<td>5.48</td>
<td>6.11</td>
<td>5.81</td>
<td>12.9</td>
<td>14</td>
<td>12.3</td>
<td>14.7</td>
</tr>
<tr>
<td>2005</td>
<td>6.15</td>
<td>7.25</td>
<td>8.33</td>
<td>7.71</td>
<td>17.1</td>
<td>18.7</td>
<td>17.8</td>
<td>18</td>
</tr>
<tr>
<td>2010</td>
<td>8.31</td>
<td>9.09</td>
<td>10.9</td>
<td>10</td>
<td>22.4</td>
<td>24.5</td>
<td>24.6</td>
<td>22.2</td>
</tr>
<tr>
<td>2015</td>
<td>10.4</td>
<td>12.4</td>
<td>14.3</td>
<td>13</td>
<td>28.7</td>
<td>32.7</td>
<td>36</td>
<td>25.4</td>
</tr>
</tbody>
</table>

From the above table it is obvious that in country like India female suffer from coronary heart disease more than men. Till the year 2010 the population suffering from coronary heart disease in the rural areas of India is more compared to urban population, whereas from the year 2010 onwards it is vice versa. In 2015 there is a drastic variance in the population suffering from CHD in rural and urban regions.

In India population of age group between 40 and 49 suffer profoundly from the heart disease. The population suffering from heart disease in all age groups has doubled in last fifteen years. The following figure illustrates Coronary heart disease in India by age, gender and region.
3. Decision Support System for Heart Disease Prediction and Diagnosis

Medical diagnosis can be improved by the use of computer-based systems and algorithms taking decisions at the appropriate stages. Such systems are called decision support systems (DSSs). Intelligence also plays a role here. These systems help to predict and diagnose the disease based on the patient information and domain knowledge. DSS
helps in improving the quality of healthcare by providing an effective and reliable diagnosis. DSS can decrease the cost of treatment by providing a more specific and faster diagnosis efficiently and also the time is reduced compare to traditional procedures. Once placed in cloud any health organization can utilize these services.

3.1. Knowledge Discovery in Database

Decision support systems was developed using a knowledge base. Knowledge discovery in database uses data mining process which extracts useful information from data set and transforms it into a reasonable structure for further use. Data mining combines statistical analysis, machine learning and database technology to extract hidden patterns and relationships from large databases [22]. [19] Defines data mining as “a process of nontrivial extraction of implicit, previously unknown and potentially useful information from the data stored in a database”. Data mining uses two strategies: supervised and unsupervised learning. A training set is used to learn the model parameters in supervised learning whereas no training set is used in unsupervised learning.

![Figure 4. Basic Methodology for Knowledge Discovery in Database](image)

3.2 Data Mining Algorithms

3.2.1. Neural Networks (NN)

Neural network is a parallel, distributed information processing structure consisting of numerous quantities of processing elements called node, they are interconnected via unidirectional signal channels called connections. Each processing element has a single output connection that branches into many connections and each conveys the equivalent signal. The NN can be classified in two main groups according to the way they learn. They are supervised learning and unsupervised learning.

In supervised learning the network compute a response to each input and then compares it with the target value. If the computed response differs from the target value, the weights of the network are adapted according to a learning rule. Examples of supervised learning are Single-layer perceptron and Multi-layer perceptron. In unsupervised learning the networks learn by identifying special features in the problems they are exposed to. Example for unsupervised learning is self-organizing feature maps.

3.2.2. Naïve Bayesian Classifier

Naïve Bayes [1] is a classification algorithm based on Bayes theorem, which calculates a probability by counting the frequency of values and combination of values n historical data. Bayes theorem finds the probability of an event occurring given the probability of another event that has already occurred.

\[
\text{Pro}(B \text{ given } A) = \frac{\text{Pro}(A \text{ and } B)}{\text{Pro}(A)}
\]
Advantage of this algorithm is it requires only a small amount of training data for estimating the parameters essential for classification.

3.2.3. Decision Tree

Berry and Linoff defined decision tree as “a structure that can be used to divide up a large collection of records into successive smaller sets of records by applying a sequence of simple decision rules. With each successive division, the members of the resulting sets become more and more similar to one another”. ID3(Iterative Dichotomiser 3) is one of the decision tree models which builds a decision tree from fixed set of training instances. C4.5 is the latest version of ID3 induction algorithm, C5.0 is an extension of C4.5 decision tree algorithm and J48 decision tree is the implementation of ID3 algorithm.

3.3. Genetic Algorithm

In Artificial Intelligence, Genetic Algorithm is a search technique which uses the process of natural selection. Genetic algorithms provide solutions to optimization and search problems by using techniques such as inheritance, mutation, selection, and crossover. A typical genetic algorithm requires genetic representation of the solution domain and a fitness function to evaluate the solution domain.

4. Surviving Techniques for Heart Disease Prediction using Data Mining Techniques

A Web based clinical decision support system which uses medical profiles like age, blood pressure, etc. is proposed to predict the prospect of patients attaining heart disease [1]. Naïve Bayes Data Mining algorithm answers complex what-if queries. The system is implemented on PHP platform which is ascendable, steadfast and expandable.

A survey [3] on different data mining techniques used for the prediction of heart disease and found hybrid approaches as best prediction model compared to single model by comparing previous researcher’s findings.

Using accuracy and sensitivity [4] as measures, author evaluated three data mining techniques such as Decision Tress (C4.5), neural networks (MLP) and Naïve Bayes. Results illustrate that with more number of attributes neural networks and Naïve Bayes perform very responding than Decision tree. Results also illustrate that neural networks perform the best with a classification accuracy of 0.897 and with sensitivity of 0.9017.

The efficacy of different decision tree algorithms used for classification and prediction of heart disease such as ID3, C4.5, C5.0 and J48 has been scrutinized [5]. ID3 is one of the decision tree models which build a decision tree from fixed set of training instances. C4.5 is the latest version of ID3 induction algorithm, C5.0 is an extension of C4.5 decision tree algorithm and J48 decision tree is the implementation of ID3 algorithm. Author also conferred attribute selection measures or the split criteria: Information gain, Gini Index & Gain ratio and performance evaluation measures: Sensitivity, specificity & accuracy.

The heart dataset contains large volumes of which consumes more time for classification so by using attribute selection methods the dimensionality of data is reduced. In both cases Naïve Bayes classification technique produced enhanced results. This is observed when the performance of four classification algorithms: Naïve Bayes, Decision Tree, K-NN and Neural Network are investigated on complete heart disease dataset and reduced dataset [7].
Table 4. Comparison Heart Disease Prediction System using Data Mining Classification Techniques

<table>
<thead>
<tr>
<th>Reference</th>
<th>Data Mining Techniques Compared</th>
<th>Accuracy Obtained</th>
<th>Number of Attributes used</th>
<th>Result: Best technique</th>
</tr>
</thead>
<tbody>
<tr>
<td>Srinivas K et al (2010)</td>
<td>Decision Trees (C4.5), Neural networks (MLP), Naïve Bayes, SVM</td>
<td>82.5%</td>
<td>15</td>
<td>Neural networks (MLP)</td>
</tr>
<tr>
<td>Chaitrali S et al (2012)</td>
<td>Decision Trees, Naïve Bayes, Neural Networks</td>
<td>96.66%</td>
<td>13 &amp; 15</td>
<td>Neural networks</td>
</tr>
<tr>
<td>John Peter T et al (2012)</td>
<td>Naïve Bayes, Decision Tree, K-NN, Neural Network</td>
<td>83.70%</td>
<td>13</td>
<td>Naïve Bayes</td>
</tr>
<tr>
<td>Anbarasi M et al (2010)</td>
<td>Naïve Bayes, Classification by clustering, Decision Tree</td>
<td>96.55%</td>
<td>6</td>
<td>Decision Tree</td>
</tr>
</tbody>
</table>

Five data mining algorithms: J48, Bayes Net, and Naïve Bayes, Simple Cart, and REPTREE are considered [8] for diagnosing heart disease, 11 attributes and Waikato Environment for Knowledge Analysis (WEKA) tool is used for prediction. J48, REPTREE and SIMPLE CART algorithm are evidenced to be best.
A survey four data mining classifiers: Naive Bayes, Neural network, Weighted Associative Classifier (WAC) algorithm and Support Vector Machine (SVM) was made [9] and used the same for predicting the heart disease with condensed number of attributes. Results demonstrate that WAC affords more accurate results in predicting the heart disease.

A three dimension survey conducted depending upon the type of dataset [10] for dataset consisting of labelled features classification model is appropriate and for unlabelled features clustering model is appropriate and to increase the performance of dataset with more optimization, bio-inspired based techniques are appropriate. Amongst the four classification models decision tree, ID3, SMV and neural networks, support vector machine is exceedingly used by researchers. Amongst the three clustering models K-means, Fuzzy C-means and hierarchical clustering, K-means is exceedingly used by researchers. Amongst the three bio-inspired based techniques Ant Colony Optimization, Artificial Immune system and Particle swarm optimization, Particle swarm optimization is exceedingly used by researchers.

Classification techniques such as logistic regression (LR), decision trees, and Artificial neural networks (ANNs) performance is compared to predict the patient’s attainment heart disease[11] Using lift chart and error chart performance is compared. Results demonstrate that artificial neural networks have the least of error rate and have the highest accuracy. Thus artificial neural networks are the best method to classify and predict the heart disease.

The presence of heart disease identified using three classifiers like Naive Bayes, Classification by clustering and Decision Tree [12]. Author reduced the number of attributes from 13 to 6 using genetic algorithm. Interpretations demonstrate that decision tree outperforms Naive Bayes, Classification by clustering technique after integrating feature subset selection with moderately high model construction time. Naïve Bayes achieves consistently before and after reduction of attributes with the same model construction time. Classification via clustering achieves poor compared to other two methods. Author intends to extend the work for predicting the intensity of the disease using fuzzy methods.

Data mining techniques such as classification, clustering, fuzzy system and association rules are studied and investigated [13] for the prediction of heart disease.

5. Surviving Techniques for Heart Disease Prediction using Hybrid Intelligent Techniques

Genetic algorithm and Fuzzy logic techniques are used for predicting heart disease where feature selection is done by Genetic algorithm and Classification and prediction is done by fuzzy logic [2]. Authors compares the performance of proposed method (GAFL system) with fuzzy entropy based method (NNTS) using the metrics like accuracy, specificity and sensitivity. Number of features is reduced from 13 to 7 and the accuracy of the proposed method is 86%.

Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm is presented in [14]. This model combines neural network, fuzzy logic and genetic algorithm. The proposed model improves training performance and classification accuracy.

Authors introduced a new classification approach for the classification of heart disease, which uses Artificial Neural Network and feature subset selection in [15]. Feature subset selection reduces the number of attributes. Pre-processing is through using Principal Component Analysis (PCA). Results demonstrate that the proposed approach indicates enhanced accuracy over traditional classification techniques.

The objective of [16] is to used genetic algorithm for determining the weights of neural networks and to evaluate two types of learning algorithms namely Feed forward neural
network algorithm and Fitting algorithm. Author validated the accuracy to be 97.75% and improvement of Feed-forward and Fitting neural network to be 1.29% and 1.37% by applying datasets on Feed Forward ANN model, Fitting ANN model and GA trained Feed Forward ANN model, GA trained Fitting ANN model.

[17] Implemented a system for predicting the heart disease using Data mining techniques: K Means and Weighted Association rule. Results demonstrate that K-means with decision tree technique make the system more accurate and efficient compared to the weighted association rule with Apriori algorithm.

Heart disease prediction system using three data mining classification techniques (Decision Trees, Naive Bayes, Neural Networks), two more attributes: Obesity and smoking along with consistent 13 attributes are used in [6]. J48 decision tree algorithm which uses pruning method for building a tree and data mining tool Weka 3.6.6 are used. Results show that neural networks produce accurate results by comparing the accuracy of classification techniques with 13 and 15 input attributes. A Multi-layer Perceptron Neural Networks (MLPNN) is used for improving the accuracy.

Artificial neural networks with back propagation error method are used to classify the cerebrovascular disease [18]. The neural network was trained with 16 input attributes using back propagation algorithm with sigmoid function on one hidden layer which improves the accuracy.

6. Conclusion

Many DSS exists to predict the heart disease with various techniques. The World life expectancy statistics implies that heart disease is prevailing more in number. So it is necessary to build an efficient intelligent trusted automated system which predicts the heart disease accurately based on the symptoms according to gender/age and domain knowledge of experts in the field at the lowest cost.

References


[21] www.worldlifeexpectancy.com

[22] www.world-heart-federation.org


Authors

J. Vijayashree, She is working as Assistant Professor in Computer Application Division at VIT University, Vellore, Tamil Nadu, India. Her area of interest includes Data Mining and Artificial Intelligence.

N. Ch. S. N. Iyengar, He is a Professor, SCS Engineering at VIT University, Vellore, TN, India. His research interests include Distributed Computing, Information Security, Intelligent Computing, and Fluid Dynamics (Porous Media). He had much teaching and research experience with a good number of publications in reputed International Journals & Conferences. He chaired many Intl. Conf. delivered Key note lectures, served as PC Member/Reviewer. He is Editorial Board member for many Int’l Journals like Int. J. of Advances in Science and Technology, of SERSC, Cybernetics and Information Technologies (CIT)- Bulgaria, Egyptian Computer Science Journal -Egypt, IJCA & IJConvC of Inderscience -China, etc., Also Editor in Chief for International Journal of Software Engineering and Applications(IJSEA) of AIRCC, Advances in Computer Science (ASC) of PPH and Many more.