Intelligent Prediction of Heart Disease Using Risk Factors Based on Data Mining Techniques

Ashish P. Nanotkar¹, Prof. Mrs. Rahila Sheikh²
Department of Computer Science & Engineering#*
Rajiv Gandhi College of Engineering, Research & Technology, Chandrapur**
Gondwana University, Gadchiroli, MH, India.#*
ashishnanotkar@gmail.com#
rahila.patel@gmail.com*

Abstract- Data mining is the science of extracting useful information from large database. It is the process of extracting hidden knowledge from large volume of raw data. These techniques have been widely used in clinical decision support system for prediction of many diseases with good accuracy. Heart disease diagnosis is one of the most important applications of such system as it is one of the leading causes of deaths all over the world. Almost all system predicting heart disease use inputs from complex tests conducted in labs and none of the system predicts heart disease based on the risk factors such as tobacco smoking, alcohol intake, age, family history, diabetes, hypertension, high cholesterol, physical inactivity, obesity. These common risk factors can be used effectively for diagnosis of heart disease. System based on such risk factors would not only help medical professionals but it would give patients a warning about the probable presence of heart disease even before he/she visits a hospital or goes for costly medical checkups. Two most successful data mining tools, neural network and genetic algorithm are involved in this technique.

Keywords- Data mining, Decision trees, Genetic neural networks, Heart disease, Prediction, Risk factors

I. INTRODUCTION

Data mining is defined as shifting through very large amounts of data for useful information. Data Mining is the process of extracting hidden knowledge from large volumes of raw data. The knowledge must be new, and one must be able to use it. There are two primary goals of data mining tend to be prediction and description [27]. Prediction involves some variables or fields in the data set to predict unknown or future values of other variables of interest. On the other hand, description focuses on finding patterns describing the data that can be interpreted by humans. In health care industry, data mining plays an important role for predicting diseases.

The Disease Prediction plays an important role in data mining. There are several factors that contribute to the development of a heart event. These risk factors are classified into two categories, non-modifiable and modifiable [8]. The first category includes factors that cannot be altered by intervention such as age, gender, family history, and genetic attributes. Modifiable risk factors for Heart Disease include smoking, hypertension, diabetes, cholesterol, high-density lipoprotein, low-density lipoprotein, etc.

The issues in this paper are to decrease the number of rules for testing, to reduce the time and cost required for various excessive Medical Tests and to increase the accuracy of Performance.

The rest of this paper is organised as follows: Section II describes Heart Disease Prediction, Section III describes Classification by Decision Tree, Section IV describes Genetic algorithm, Section V describes Neural Network Architecture, Section VI describes Literature Review and Literature table and Section VII describes Proposed Work.

A. Heart Disease Prediction

From the analysis of World Health Organization, they estimated 12 million deaths occur worldwide, every year due to the Heart diseases. Half the deaths occur in United States and other developed countries due to cardiovascular diseases. On the above discussion, it is regarded as the primary reason behind deaths in adults. Heart disease kills one person every 34 seconds in the United States. An estimated 17.3 million people died from Heart diseases in 2008, representing 30% of all global deaths. Of these deaths, an estimated 7.3 million were due to coronary heart disease and 6.2 million were due to stroke. Diabetes, hypertension, and high blood cholesterol have been established as the major risk factors of heart diseases. There are studies showing that reducing these risk factors for heart disease can actually help in preventing heart diseases.

In health care industry, data mining plays an important role for predicting diseases. For detecting a disease number of tests should be required from the patient. But using data mining technique the number of test should be reduced. This reduced test plays an important role in time and performance. The most important risk factors, as extracted from the classification rules analysis were: 1) for MI, age, smoking, and history of hypertension; 2) for PCI, family history, history of hypertension, and history of diabetes; and 3) for CABG, age, history of hypertension, and smoking. Most of these risk factors were also extracted
by other investigators. The highest percentages of correct classifications achieved were 66%, 75%, and 75% for the MI, PCI, and CABG models, respectively.

![Heart Disease Prediction]

Fig. 1 Architecture of most important Risk Factors

B. **Classification By Decision Tree**

Decision tree (DT) is one of the popular and important classifier which is easy and simple to implement. It doesn’t have domain knowledge or parameter setting. It handle huge amount of dimensional data. It is more suitable for exploratory knowledge discovery. The results attained from Decision Tree are easier to interpret and read. DT are tree shaped structures that represent sets of decisions. It explains how a decision or diagnosis is reached. The DT approach can generate rules.

The C4.5 algorithm, which uses the divide-and-conquer approach to decision tree induction, was employed. The algorithm uses a selected criterion to build the tree. It works top–down, seeking at each stage an attribute to split on that which best separates the classes, and then recursively processing the sub problems that result from the split. The algorithm uses heuristics for pruning derived based on the statistical significance of splits.

**Algorithm to Generate Decision Tree:**

*Input:*

1. Training dataset $D$, which is a set of training observations and their associated class value.
2. Attribute list $A$, the set of candidate attributes.
3. Selected splitting criteria method.

*Output:* A decision tree.

*Method:*

1. Create a node $Nd$.
2. If all observations in the training dataset have the same class output value $C$, then return $Nd$ as a leaf node labeled with $C$.
3. If attribute list is empty, then return $Nd$ as leaf node labelled with majority class output value in training dataset.
4. Apply selected splitting criteria method to training dataset in order to find the “best” splitting criterion attribute.
5. Label node $Nd$ with the splitting criterion attribute.
6. Remove the splitting criterion attribute from the attribute list.
7. For each value $j$ in the splitting criterion attribute, 
   a) Let $D_j$ be the observations in training dataset satisfying attribute value $j$.
   b) If $D_j$ is empty (no observations), then attach a leaf node labeled with the majority class output value to node $Nd$.
   c) Else attach the node returned by generate decision tree ($D_j$, attribute list, selected splitting criteria method) to node $Nd$.
8. End for.
9. Return node $Nd$.

C. **Genetic Algorithm**

Genetic algorithms have been used to reduce the actual data size to get the optimal subset of attributed sufficient for heart disease prediction. In our proposed method we classify the heart disease data and prediction of disease by applying feature subset selection. Our approach prunes irrelevant, redundant attributes and generates compact rule set. These generated rules will be built as classifier. Accuracy of our method improves over other classification algorithms.

*Algorithm:*

1. Select the common attributes by applying feature subset selection measures like information gain, SU, and genetic search.
   a) $SU(X, Y) = 2[IG(X/Y)/H(X) H(Y)]$
   b) $IG(X/Y) = H(X) - H(X/Y)$
   c) Genetic search conditions
      i. Cross over probability: 0.6
      ii. Mutation probability: 0.033
      iii. Max. generation 20
      iv. Population: 20
      v. Method: 10 fold cross
2. Generate the class association rules using information centric attributes using training data set.
3. Build the classifier using generated class association rules.
4. Predict the rules on test data
5. Find the accuracy of the classifier.
6. Accuracy = Accuracy measures the ability of the classifier to correctly classify unlabelled data.

Accuracy = Number of objects correctly Classified Total No. of objects in the test set.
D. Neural Network Architecture

A neural network's ability to perform computations is based on the hope that we can reproduce some of the flexibility and power of the human brain by artificial means. Network computation is performed by dense mesh of computing nodes and connections. A three layer neural network having eight neurons in the input layer, nine neurons in the hidden layer, one neuron in the output layer is considered (shown in Fig. 3).

II. LITERATURE REVIEW

JyotiSoni [7] proposed three different supervised machine learning algorithms. These are Naïve Bayes, K-NN, and Decision List algorithm. These algorithms have been used for analyzing the heart disease dataset. Decision tree is one of the popular and important classifier which is easy and simple to implement. It does not have domain knowledge or parameter setting. It handle huge amount of dimensional data. It is more suitable for exploratory knowledge discovery. The results attained from Decision Tree are easier to interpret and read. Naïve Bayes is a statistical classifier which assigns no dependency between attributes. K-nearest neighbor’s algorithm (k-NN) is the one of the important method for classifying objects based on closest training data in the feature space. It is simplest among all machines learning algorithm but, the accuracy of k-NN algorithm can be degraded by presence of noisy features. The dataset is divided into two testing and training i.e. 70% of data are used for training and 30 % is used for testing. The authors concluded that Naïve Bayes algorithm performs well when compared to other algorithms.

In the survey of Naïve Bayes [7] have been used to predict attributes such as age, sex, blood pressure and blood sugar and the chances of a diabetic patient getting a heart disease. The heart disease is diagnosed for diabetic patients using Naïve Bayes technique. Of these the author concluded that Naïve Bayes classify 74% of input instances correctly.

Minas A. Karaolis [22] investigated the most important risk factors, as extracted from the classification rules analysis were: 1) for MI, age, smoking, and history of hypertension; 2) for PCI, family history, history of hypertension, and history of diabetes; and 3) for CABG, age, history of hypertension, and smoking. Most of these risk factors were also extracted by other investigators. The highest percentages of correct classifications achieved were 66%, 75%, and and 75% for the MI, PCI, and CABG models, respectively.

M. Akhil jabbar, Priti Chandra [5] discussed about prediction of risk score for heart disease in Andhra Pradesh. They generated class association rules using feature subset selection. These generated rules helped physicians to predict the heart disease of a patient. They used feature selection measure to determine the attributes which contributemore towards the prediction of heart disease which indirectly reduces the no. of diagnosis tests which are needed to be taken by a patient.

Markos G. Tsipouras [19] proposed a Fuzzy rule-based Decision Support System (DSS) for the diagnosis of coronary artery disease. The dataset used for the DSS generation and evaluation consists of 199 subjects, each one characterized by 19 features, including demographic and history data, as well as laboratory examinations. Tenfold cross validation is employed, and the average sensitivity and specificity obtained is 62% and 54%, respectively, using the set of rules extracted from the decision tree, while the average sensitivity and specificity increase to 80% and 65%, respectively, when the fuzzification and optimization stages are used.

AH Chen, SY Huang, PS Hong [4] developed a heart disease prediction system that can assist medical professionals in evaluating a patient’s heart disease based on the clinical data of the patient. Their approach included three steps. Firstly, they selected 13 important clinical features i.e., age, sex, chest pain type, cholesterol, fasting blood sugar, resting ECG, max heart rate, exercise induced angina, old peak, slope, number of vessels colored. Secondly, they developed an artificial neural network algorithm for
classifying heart disease based on these clinical features. The accuracy of prediction is near 80%.

The following TABLE I shows the survey on prediction of Heart Disease by different authors using various technologies and algorithms giving different accuracies.

<table>
<thead>
<tr>
<th>No.</th>
<th>Author Name</th>
<th>Year</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Erkki Vartiainen, Juha Pekkanen, Seppo Koskinen, Pekka Jousilahti</td>
<td>1998</td>
<td>Used mortality from IHD using logistic regression models with the risk factor levels assessed by cross sectional population surveys.</td>
</tr>
<tr>
<td>2.</td>
<td>Joanne T. Brindle, Henrik Antti, Elaine Holmes, G. Tranter</td>
<td>2002</td>
<td>They designed a technique capable of providing an accurate, non invasive and rapid diagnosis of coronary heart disease.</td>
</tr>
<tr>
<td>4.</td>
<td>Dragan Gamberger, Nada Lavrac, Goran Krstac</td>
<td>2003</td>
<td>They presented risk group about Coronary Heart Disease as well as decision rules which was used to help decision making at a level of one person.</td>
</tr>
<tr>
<td>5.</td>
<td>S. Kalaiarasi Anbananthen, G. Sainarayanan, Ali Chekima, Jason Teo</td>
<td>2005</td>
<td>ANNT approach was used on diabetes data set. It generated rules with strong generalization and comprehensibility ability.</td>
</tr>
<tr>
<td>6.</td>
<td>Carlos Ordonez</td>
<td>2006</td>
<td>Association rules were used to predict multiple related target attributes, reducing the number of discovered rules and improving running.</td>
</tr>
<tr>
<td>7.</td>
<td>Selappan Palaniappan, Rafiah Awang</td>
<td>2008</td>
<td>A heart disease prediction system was developed using three data mining classification techniques such as Naive Bayes, Neural Network and Decision Trees.</td>
</tr>
<tr>
<td>8.</td>
<td>Resul Das, Ibrahim Turkoglu, Abdulkadir Sengur</td>
<td>2008</td>
<td>They used a neural networks based methodology for diagnosis of the heart disease. The experimental results gained 89.01% classification accuracy.</td>
</tr>
<tr>
<td>9.</td>
<td>Markos G. Tspirouas, Themis P. Exarchos, Dimitrios I. Fotiadi</td>
<td>2009</td>
<td>The optimized fuzzy model DSS improved by 15% in terms of accuracy the results of the crisp rule-based classifier.</td>
</tr>
<tr>
<td>10.</td>
<td>Asil Oztekina, Dursun Delenb, Zhenyu Kong</td>
<td>2009</td>
<td>They made the use of voluminous datasets, which was useful to apply the data mining models.</td>
</tr>
<tr>
<td>11.</td>
<td>Shantakumar B. Patil, Dr. Y. S. Kumaraswamy</td>
<td>2009</td>
<td>They used K-means clustering algorithm and developed an efficient heart attack prediction system using artificial intelligence techniques.</td>
</tr>
<tr>
<td>12.</td>
<td>Tahseen A. Jilani, Huda Yasin, Madiha Yasin, Cemal Arbil</td>
<td>2010</td>
<td>Five mining goals were defined based on business intelligence and data exploration.</td>
</tr>
<tr>
<td>13.</td>
<td>K. Srinivas B. Kaviha Rani Dr. A. Govrdhan</td>
<td>2010</td>
<td>They applied fuzzy learning models for prediction of acute coronary syndrome.</td>
</tr>
<tr>
<td>14.</td>
<td>Asha Rajkumar, Mrs. G. Sophia Reena</td>
<td>2010</td>
<td>They used Naive Bayes, Decision list and K-NN algorithms; it gave 52.33% of accurate result.</td>
</tr>
<tr>
<td>15.</td>
<td>M. Anbarasi, E. Anupriya, N. Ch. S. N. Iyengar</td>
<td>2010</td>
<td>Genetic algorithm was used to determine the attributes which contribute more towards the diagnosis of heart ailments.</td>
</tr>
<tr>
<td>16.</td>
<td>Farhad Soleimanian Gharechopoghi, Zeynab Abbasi Khalifelu</td>
<td>2011</td>
<td>Results indicated that Neural Network system gave 85% of accuracy.</td>
</tr>
<tr>
<td>17.</td>
<td>AH Chen, SY Huang, PS Hong, CH Cheng, EJ Lin</td>
<td>2011</td>
<td>They developed an artificial neural network algorithm with accuracy of prediction near 80%.</td>
</tr>
</tbody>
</table>
III. PROPOSED WORK

Here we are designing a web based application that will be used to predict a heart disease using risk factors such as Sex, Age, Blood Cholesterol, Blood Pressure, Hereditary, Smoking, Alcohol Intake, Physical Activity, Diabetes, Diet, Obesity, Stress based on data mining techniques. The goal here is to predict Heart Disease with maximum of accuracy by using Genetic algorithm and Artificial Neural Network Techniques so that people will be aware of possibility the disease and can make efforts in reducing it. Syed Umar Amin and Kavita Agarwal [8] have proposed a system in the year 2013 by which Data mining techniques and methods applied in patient medical dataset has resulted in innovations, standards and decision support system that have significant success in improving the health of patients and the overall quality of medical services. But we still need systems which could predict heart diseases in early stages.

The objective here is to develop a data-mining system for the assessment of heart event-related risk factors targeting in the reduction of CHD events. The risk factors investigated are: 1) before the event: a) non modifiable-age, sex, and family history for premature CHD, b) modifiable-smoking before the event, history of hypertension, and history of diabetes; and 2) after the event: modifiable-smoking after the event, systolic blood pressure, diastolic blood pressure, total cholesterol, high-density lipoprotein, low-density lipoprotein, triglycerides, and glucose.

The issues here are- to decrease the number of rules for testing, to reduce the time and cost required for various excessive Medical Tests, and to increase the accuracy of performance. For detecting a disease, number of tests should be required from the patient. But using data mining technique, the number of test should be reduced. This reduced test plays an important role in time and performance.

<table>
<thead>
<tr>
<th>No.</th>
<th>Risk Factors</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sex</td>
<td>Male (1), Female (0)</td>
</tr>
<tr>
<td>2</td>
<td>Age (years)</td>
<td>20 - 34 (-2), 35 – 50 (-1), 51- 60 (0), 61 - 79 (1), &gt;79 (2)</td>
</tr>
<tr>
<td>3</td>
<td>Blood Cholesterol</td>
<td>Below200 mg/dL- Low(-1) 200-239mg/dL- Normal (0) 240mg/dLandabove- High(1)</td>
</tr>
<tr>
<td>4</td>
<td>Blood Pressure</td>
<td>Below 120 mm Hg – Low (-1) 120 o 139 mm Hg-Normal (0) Above139mmHg-High (-1)</td>
</tr>
<tr>
<td>5</td>
<td>Hereditary</td>
<td>Family Member diagnosed with HD – Yes (1) Otherwise – No (0)</td>
</tr>
<tr>
<td>6</td>
<td>Smoking</td>
<td>Yes (1) or No (0)</td>
</tr>
<tr>
<td>7</td>
<td>Alcohol Intake</td>
<td>Yes (1) or No (0)</td>
</tr>
<tr>
<td>8</td>
<td>Physical Activity</td>
<td>Low (-1),Normal (0), High (-1)</td>
</tr>
<tr>
<td>9</td>
<td>Diabetes</td>
<td>Yes (1) or No (0)</td>
</tr>
<tr>
<td>10</td>
<td>Diet</td>
<td>Poor (-1),Normal (0) or Good (1)</td>
</tr>
<tr>
<td>11</td>
<td>Obesity</td>
<td>Yes (1) or No (0)</td>
</tr>
<tr>
<td>12</td>
<td>Stress</td>
<td>Yes (1) or No (0)</td>
</tr>
<tr>
<td></td>
<td>Output</td>
<td>Heart Disease Yes (1) or No (0)</td>
</tr>
</tbody>
</table>
REFERENCES


