Chapter 6

Case Study on Heart Disease

6.1 Introduction

Heart disease has become more prevalent in recent years, prompting scholars to devote more attention to its risk factors. Early diagnosis and treatment is one of the best approaches to reduce the disease death rate. For such reasons, Knowledge Discovery techniques may be a practical and effective solution for generating and proving new hypothesis, mining and generalizing new medical knowledge directly from pertinent real examples.


In this chapter, an approach to apply a real time dataset on our new model CURE is presented. Section 6.2 describes the dataset used in the example for construction of the tree. Section 6.3 discusses the construction of decision tree for C4.5 and CURE model. Section 6.4 gives the parameters on which the evaluation of the results will be done. Section 6.5 compares the experimental results and discusses the obtained results. Section 6.6 discusses the observation found by the above experiments. Section 6.7 describes the related work for heart disease. In Section 6.8 a summary of the chapter and a conclusion is provided.
6.2 The Data Set Used

6.2.1 Collection of the Data

In this case study, the goal is to identify the risk factors for heart disease. The present data analysis was carried out on a data collection of heart disease cases for men from several hospitals ranging from simple general hospital to specialized cardiac centers in and around Machilipatnam, Andhra Pradesh, India. There are roughly two controls per case of Chronic Heart Disease (CHD). Many of the CHD positive men have undergone blood pressure reduction treatment and other programs to reduce their risk factors after their CHD event. In some cases the measurements were made after these treatments.

The number of instances in the data set is 462 (302 CHD absent, 160 CHD present) and the number of attributes is 11 and the number of classes is 2 (CHD absent and CHD present). There are no missing values in the data set. The main attributes which are used in the decision tree are shown in Table 6.1.

Table 6.1: Attribute information of the heart disease dataset collected in and around machilipatnam for the purpose of study.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>Identity of the patient</td>
</tr>
<tr>
<td>SBP</td>
<td>systolic blood pressure</td>
</tr>
<tr>
<td>Tobacco</td>
<td>cumulative tobacco (kg)</td>
</tr>
<tr>
<td>Ldl</td>
<td>low density lipoprotein cholesterol</td>
</tr>
<tr>
<td>Adiposity</td>
<td>Adiposity of the patient</td>
</tr>
<tr>
<td>Famhist</td>
<td>family history of heart disease (Present, Absent)</td>
</tr>
<tr>
<td>Typea</td>
<td>type-A behavior</td>
</tr>
<tr>
<td>Obesity</td>
<td>Whether patient is fat or not</td>
</tr>
<tr>
<td>Alcohol</td>
<td>current alcohol consumption</td>
</tr>
<tr>
<td>Age</td>
<td>age at onset</td>
</tr>
<tr>
<td>CHD</td>
<td>response, coronary heart disease</td>
</tr>
</tbody>
</table>
6.2.2 Dividing the Data Set
We run our new algorithm with 20 runs of 10 fold cross validation. In 462 instances we have taken 415 cases as training set and 47 cases as test set. The data specific pruning value for this dataset is set as 0.280.

6.3 Construction of the decision tree

6.3.1 Constructing the C4.5 decision tree

We have constructed the C4.5 decision tree with the collected datasets. Figure 6.1 shows the tree which was constructed using the 415 cases in the training set. The circles represent test nodes which are used to split the data and the squares or rectangles represent the leaf nodes of the tree.

![Figure 6.1: Decision tree generated by C4.5 algorithm.](image-url)
6.3.2 New Decision Tree construction with CURE

We have constructed the decision tree for the same dataset by using our new CURE method. Figure 6.2 shows the new decision tree constructed by using 415 cases in the training set.

\[
\text{Accuracy} = \frac{(TP) + (TN)}{Total\ Samples} \quad (6.1)
\]

Figure 6.2: Decision tree generated by Classifier Using Randomization and Expert knowledge.

6.4 Evaluating the algorithm

One method of judging the performance of a classifier is to compare the accuracy of all classifications. This number is calculated by using equation (6.1),
Where, TP is the total number of correct positive classifications and TN is the total number of correct negative classifications, or correct rejections.

Another common method of evaluating classifier performances is to look at the sensitivity of a classification model. Sensitivity is equivalent to the true positive rate, and is calculated as the number of true positive classifications divided by all positive classifications as given in equation (6.2),

\[ Sensitivity = \frac{TP}{(TP) + (FN)} \quad (6.2) \]

Another common metric in biomedical literature is the specificity of a classification model. Specificity, also known as the correct rejection rate, is defined as the number of true negative classifications divided by all negative classifications as given in equation (6.3),

\[ Specificity = \frac{TN}{(TN) + (FP)} \quad (6.3) \]

As a classifier becomes more sensitive it will identify a greater proportion of true positive instances, however, the number of false negative classifications will consequently rise. Similarly, as a classification model becomes more specific, i.e. correctly rejecting greater proportion of true negative instances, then the number of false positive classifications will also rise.

6.5 Experimental Results

Table 6.2: The evaluation results of C4.5 and Classifier Using Randomization and Expert knowledge (CURE) for accuracy, sensitivity, specificity, tree size, depth and number of leaves.

<table>
<thead>
<tr>
<th>Evaluation</th>
<th>C4.5</th>
<th>CURE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>70.16</td>
<td>73.10</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.80</td>
<td>0.82</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.48</td>
<td>0.50</td>
</tr>
<tr>
<td>Tree Size</td>
<td>29.00</td>
<td>14.40</td>
</tr>
<tr>
<td>Depth</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>Leaves</td>
<td>15.00</td>
<td>7.70</td>
</tr>
</tbody>
</table>
From the experimental results reported in Table 6.2, Figure 6.1 and Figure 6.2, it is clear that our new algorithm CURE can perform very well, especially if the objective is known. If simplicity is the main quality we are looking for then small tree size is required. Our new algorithm has greatly increased simplicity; since the tree size of our new algorithm over C4.5 is reduced from 29.0 to 14.4. When a good model of existing collected data is sought, then high accuracy is the objective. Our algorithm CURE has improved the accuracy over C4.5 from 70.16 to 73.10. In case of accuracy also our algorithm CURE is the best performing algorithm.

For achieving good generalization over C4.5 our new algorithm tree depth is reduced from 7 to 6. But due to this both algorithms avoid over-fitting the data. The numbers of leaves generated are equal to number of rules in the decision tree; less the number of rules generated easy the way to predict outcome in the clinical domain. Our algorithm have produced only around 8 rules where as C4.5 have produced around 15 rules. The best performing algorithm to predict outcome easily is again our new algorithm CURE. In case of sensitivity and specificity our algorithm has slightly improved over C4.5.

6.6 Observations

Figure 6.2 shows the complete decision tree generated by using CURE model to predict patients output – Present or Absent. The root node of the tree is a test on age. The age emerges as most important and significant predictor for conforming heart disease.
Four Simple Rules

1. A case with AGE greater than 31 and TYPEA value greater than 68 may have a heart disease. This rule covers 19 cases in our dataset among which 2 cases are misclassified.

2. A case with AGE greater than 50, TYPEA value less than or equal to 68 and FAMHIST is present may have a heart disease. This rule covers 84 cases in our dataset among which 27 cases are misclassified.

3. A case with AGE greater than 50, TYPEA value less than or equal to 68, FAMHIST is absent and TOBACO greater than 7.6, may have a heart disease. This rule covers 24 cases in our dataset among which 7 cases are misclassified.

4. A case with AGE greater than 50, TYPEA value less than or equal to 68, FAMHIST is absent and TOBACO greater than 3.96 and less than or equal to 4.82 may have a heart disease. This rule covers 6 cases in our dataset among which 1 case is misclassified.

Early Detection:

From the clinical experts we have confirmed the main parameters for indicating the presence of heart disease are – age, famhist, tobacco and typea. The same parameters are revealed by our decision tree CURE.

If a patient makes number of regular visits to the clinic then in the first one or two visits itself we can make an early detection of heart disease by using our four simple rules. For example a patient has these symptom age > 50, famhist-absent and tobacco > 7.6. From the third rule, it says that

\[ \text{If } \text{age} > 50 \land \text{TypeA} \leq 6.8 \land \text{famhist} = \text{absent} \land \text{tobacco} > 7.6 \text{ Then heart disease.} \]

In this above discussed case, the patient has three out of four symptoms present and one symptom is absent. Therefore in this case according to the number of cases misclassified in our dataset there is a 70 % chance of having a heart disease. In this early stage we should detect heart disease and should start the treatment to see that TYPEA value should
not be less than 6.8, which is a critical value and we should try to reduce the modifiable factors from critical values. In this case, out of three factors age, family history and tobacco the modifiable factor is only tobacco. We should try to reduce the tobacco value from 7.6. By doing so, we can slowly cure the third factor also, thereby making a future heart patient healthier.

Finally, we may conclude that only decision tree rules are not enough for early detection but the physicians experience and interpretive skills are also must. Our decision trees rules can serve as a clinical decision support system. In fact, by this early detection and treatment in 100 cases if we are going to cure the heart disease for at least 20 cases then also it is a great achievement for our early detection and treatment thereby reduce the heart disease death rate.

However, one should not forget that even though data mining can provide assistance in making the diagnosis or prescribing the treatment. It cannot replace the physician’s intuition, experience and interpretive skills [96],[97]. In short, data mining is not aiming to replace medical professionals and researchers, but to complement and support their efforts to save human life [98].

6.7 Conclusion

Early detection of heart disease is a challenging problem in medical domain for avoiding high death rates due to heart failures. We solve this problem in a scalable way by using our new classifier CURE which gives a simplified tree with high accuracy. Experiments conducted using a heart disease dataset collected during the study suggests that our algorithm have attained better results by showing supremacy in all the departments to generate a feasible solution for the early heart disease detection. This study suggests that (Age), age of the patient is the most important predictor for confirming the presence of heart disease, (TYPEA) type-A behavior, (FAMHIST) family history of heart disease (present, absent) and (TOBACO) cumulative tobacco as biomarkers of heart disease.