Abstract

Heart attack occurs when the blood flow to a part of the heart is blocked by a blood clot. If this clot cuts off the blood flow completely, the part of the heart muscle supplied by that artery begins to die. Currently there is no cure for heart attack but it can be controlled by quitting smoking, lowering cholesterol, controlling high blood pressure, maintaining a healthy weight, and exercising. Generally many tests are done that involve clustering or classification of large scale data. In this work, prediction of heart disease using support vector machine ensembles using Matlab is done. The aim of this paper is to analyze the performance of Support Vector Machine (SVM) classifier and Ensemble classifier methods such as Bagging, Boosting and Random subspace for heart disease prediction. Accuracies of different classification algorithms are compared to bring out the best and effective algorithm suitable for heart disease prediction.

Keywords: Data Mining, Statlog heart dataset, Support Vector Machine, Ensemble Classifiers.

1. Introduction

Data mining is the process of extracting information from data. It is also called as knowledge discovery. Data mining has become more and more popular to analyze large amount of data in the past few years. The most important and popular data mining techniques are classification and clustering. In this paper SVM classifier and ensemble methods such as Bagging, Boosting and Random subspace are investigated for heart disease prediction. The heart disease is often used interchangeably with the term cardiovascular disease. Cardiovascular disease generally refers to conditions that involve narrowed or blocked blood vessels that can lead to a heart attack. Symptoms of a heart attack includes discomfort, pressure, heaviness, or pain in the chest, arm, or below the breastbone, discomfort radiating to the back, jaw, throat, or arm, indigestion, sweating, nausea, vomiting, extreme weakness, anxiety, or shortness of breath, rapid or irregular heartbeats. Initial symptoms may start as a mild discomfort that progress to significant pain. In general to detect a disease numerous tests must be conducted in a patient. The usage of data mining techniques in disease prediction is to reduce the test and increase the accuracy rate of detection.
2. Related work


3. Dataset description

In this work, Statlog Heart Dataset [11] is used from the UCI machine learning repository. The dataset contains 270 instances and 14 attributes with 2 class attributes. This dataset contains information concerning heart disease diagnosis.

<table>
<thead>
<tr>
<th>Sl. No.</th>
<th>Attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Age</td>
</tr>
<tr>
<td>2</td>
<td>Gender</td>
</tr>
<tr>
<td>3</td>
<td>Chest Pain(CP)</td>
</tr>
<tr>
<td>4</td>
<td>trestbps: resting blood pressure</td>
</tr>
<tr>
<td>5</td>
<td>Cholesterol</td>
</tr>
<tr>
<td>6</td>
<td>fbs: fasting blood sugar</td>
</tr>
<tr>
<td>7</td>
<td>restecg: resting electrocardiographic</td>
</tr>
<tr>
<td>8</td>
<td>thalach : maximum heart rate</td>
</tr>
<tr>
<td>9</td>
<td>exang : exercise induced angina</td>
</tr>
<tr>
<td>10</td>
<td>Oldpeak</td>
</tr>
<tr>
<td>11</td>
<td>Slope</td>
</tr>
<tr>
<td>12</td>
<td>ca: no. of major vessels</td>
</tr>
<tr>
<td>13</td>
<td>Thal</td>
</tr>
<tr>
<td>14</td>
<td>Class variable</td>
</tr>
</tbody>
</table>

The 14 attributes are as follows;

10 fold Cross validation technique is used to split the data. In a 10 fold cross validation, the data is divided into 10 parts where each part is approximately same to form the full dataset. Learning procedure executes 10 times on training sets and finally the accuracy rates for 10 sets are averaged to yield an overall accuracy rate. Confusion matrix is used to present the accuracy of classifiers obtained through classification.

4. Methodology

4.1 SVM Classifier

SVM is a commonly used technique for data classification. SVM produces a model which predicts target value of data instances in the testing set. Support Vector Machine (SVM) is used when data has exactly two classes. An SVM classifier classifies data by finding the best hyperplane that separates all data points of one class from the other class. The hyperplane with the largest margin between the two classes is the best one. Margin means the
maximal width of the slab parallel to the hyperplane that has no interior data points.

The formulae for Support Vector Machine as follows:

Set of training data-\{(a_1,b_1),\ldots,(a_l,b_l)\},

where,

Each \(a_i \subseteq S^n\) denotes the input space of the sample
\(b_i \subseteq S\) denotes the target value
\(i=1,2,\ldots,l\)

\(l\)- size of training data

\[\min J(D, \xi) = \frac{1}{2} |D|^2 + C \sum_{i=1}^{l} \xi_i\]

where,

\(C\) - constant of capacity control
\(\xi_i\) - slack factor

Optimization problem can be rewritten as follows:

\[\max M(\alpha) = -\frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j b_i b_j K(a_i, a_j) + \sum_{i=1}^{l} \alpha_i \sum_{i=1}^{l} \alpha_i b_i\]

\(\alpha_i \in [0,C], i = 1, 2, \ldots, l\).

\(K(a_i, a_j)\) - kernel function

Optimal hyperplane with maximal margin:

\[\sum_{xv} \alpha_i b_i K(a_i, a_i) + b = 0\]

SVM for nonlinear classification in the input space is:

\[f(x) = \text{sgn} \left[\sum_{xv} b_i \alpha_i K(a_i, a) + b\right]\]

4.2 Ensemble Classifier (EC)

Ensemble Data Mining Methods, also known as Committee Methods or Model Combiners are machine learning methods that use the power of many models to achieve better accuracy than the individual models.

The following ensemble methods are used in this work:

- Bagging
- Boosting
- Random Subspace

Bagging, boosting and random subspace method have been used commonly for combining weak classifiers.

4.2.1 Bagging

Bootstrap aggregation, or bagging, is a technique that can be used with many classification methods and regression methods to improve the prediction process by reducing the variance associated with prediction. It is a simple technique that many bootstrap samples are drawn from the available data and some prediction method is applied to each bootstrap sample, and then the results are combined, by averaging for regression and simple voting for classification, to obtain the overall prediction, with the variance being reduced due to the averaging.

4.2.2 Boosting

The AdaBoost family of algorithms also known as boosting is another category of powerful ensemble method. It changes the distribution of weights. Initially the weights are uniform for all the training samples. The weights are adjusted after training of each classifier is completed. For misclassified samples the weights are increased while for correctly classified samples they are decreased. The final ensemble is constructed by combining individual classifiers according to their own accuracies.

4.2.3 Random Subspace

In random subspace feature subspaces are picked at random from the original feature space and individual classifiers are created only based on those attributes in the chosen feature subspaces using the original training set. The outputs from different individual classifiers are combined by the uniform majority voting to give the final prediction.

5. Result Analysis

Different metrics can be used for evaluating the performance of classifiers. In this work, the performance metrics such as Accuracy, Sensitivity, Specificity, Positive Predictive Value (PPV), Negative Predictive Value (NPV) are used for evaluating the classifiers.

The formulas for these metrics are given below:

\[\text{Sensitivity(\%)} = \frac{TP}{TP + FN} \times 100\]
Specificity(%) = \frac{TN}{TN + FP} \times 100

PPV(%) = \frac{TP}{TP + FP} \times 100

NPV(%) = \frac{TN}{TN + FN} \times 100

Accuracy(%) = \frac{TP + TN}{TP + TN + FP + FN} \times 100

where,

TP - Total number of correctly classified true data.
TN - Total number of misclassified true data.
FP - Total number of correctly classified false data.
FN - Total number of misclassified false data.

5.1 Experimental Results

The performance measures of Support Vector Machine (SVM) and ensemble classifiers are given in table 2. The results are also shown graphically in figure 2.

Table 2. Performance Analysis of SVM and Ensemble Classifier Methods

<table>
<thead>
<tr>
<th></th>
<th>SVM</th>
<th>Bagging</th>
<th>Boosting</th>
<th>Random Subspace</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>73.7</td>
<td>81.85</td>
<td>83.22</td>
<td>80.00</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>73.78</td>
<td>81.49</td>
<td>83.00</td>
<td>77.9</td>
</tr>
<tr>
<td>Specificity</td>
<td>73.7</td>
<td>81.02</td>
<td>82.12</td>
<td>77.2</td>
</tr>
<tr>
<td>PPV</td>
<td>74.05</td>
<td>81.67</td>
<td>82.4</td>
<td>80.08</td>
</tr>
<tr>
<td>NPV</td>
<td>73.5</td>
<td>80.56</td>
<td>84.00</td>
<td>80.00</td>
</tr>
</tbody>
</table>

Figure 2. Performance Analysis of SVM and Ensemble Classifier Methods

The experimental results shows that SVM classifier achieved classification accuracy of 73.7%, Bagging achieved a classification accuracy of 81.85% and Random Subspace achieved a classification accuracy of 80% and Boosting achieved a classification accuracy of 83.22%.

It is clear that, Boosting method performs better when compared to the other techniques in terms of accuracy, sensitivity, specificity, PPV and NPV.

6. Conclusion

In this work, SVM and three ensemble methods such as bagging, boosting and random subspace have been implemented and tested with statlog heart dataset. 10-fold cross-validation evaluation was used to measure the accuracy of the three algorithms. The final comparative analysis shows that the Boosting ensemble method performed better than other methods. This work can be further extended using other datasets. Ensemble methods with other base classifiers can be applied for classification in future. Feature selection technique can also be adopted so that the performance may be improved.
7. References


