Early Prediction of Heart Attacks using Data Mining Techniques

Data Mining Project by
Avichal Jain
Varun Gupta
Rigvita Sharma
Aastha Kataria
Aditi Agarwal
Sahil Ranadive

Introduction

- Classification is one of the most important techniques used in data mining.
- Implementation of the paper “Early Prediction of Heart Disease using Data Mining techniques” is done using a set of classification techniques namely,

1. Classification and Regression Tree (CART)
2. Iterative Dichotomized 3 (ID 3)
3. Random Forests Ensemble Classifier

DATASET AND ATTRIBUTES:

Dataset taken from:
http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/processed.cleveland.data

ATTRIBUTES:

The 14 attributes used in the dataset are:

1. Age
2. sex
3. Cp
4. trestbps
5. Chol: serum cholesterol in mg/dl
6. Fbs
7. Restecg
8. thalach
9. Exang
10. Oldpeak
11. Slope
12. ca
13. Thal
14. num (the predicted attribute)

Preprocessing

1. Data Cleaning - Data is cleaned by filling the missing values using mode of the feature.
2. Data Transformation - Data is normalized using zero-mean normalization.
3. PCA(Principal Component Analysis): PCA has been used for feature selection and reducing the variance.
4. Constructing separate input and target lists.

CART

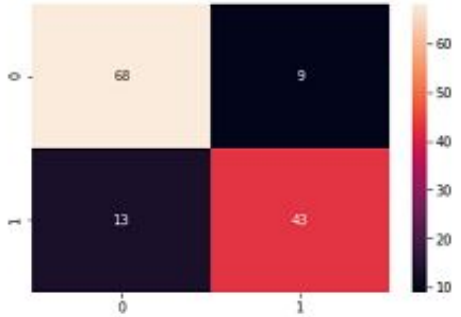
- CART algorithm uses decision trees to predict the target value of data.
- CART uses binary trees to predict outcomes.
- It used simple yes/no answers to queries on attributes to select a specific branch.

Accuracy of the CART is: 83.45864661654136

	precision	recall	f1-score	support
0.0	0.84	0.88	0.86	77
1.0	0.83	0.77	0.80	56
avg / total	0.83	0.83	0.83	133

Confusion Matrix
==========

Given in the figure is the confusion matrix for the CART algorithm along with its accuracy, Precision, recall, f1-score, and support.



ID3

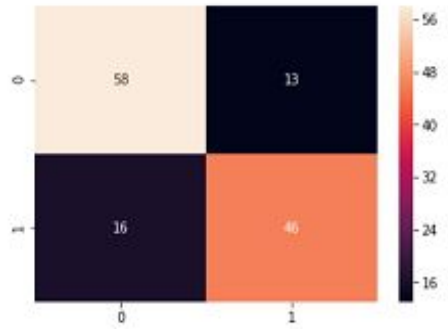
- ID3 also uses decision tree to predict the target value.
- It is used for the classification of the objects with the iterative inductive approach.
- Top to down, greedy search approach is used. It’s traversing from root node to leaf nodes.
- It uses information gain to help it decide which attribute goes into a decision node.It Does not handle numeric attributes and missing values.

Given in the figure is the confusion matrix for the ID3 algorithm along with its accuracy, Precision, recall, f1-score, and support.

Accuracy of the ID3 is: 78.19548872180451

	precision	recall	f1-score	support
0.0	0.52	0.59	0.55	71
1.0	0.44	0.37	0.40	62
avg / total	0.48	0.49	0.48	133

Confusion Matrix



Random Forest Ensemble Classifier

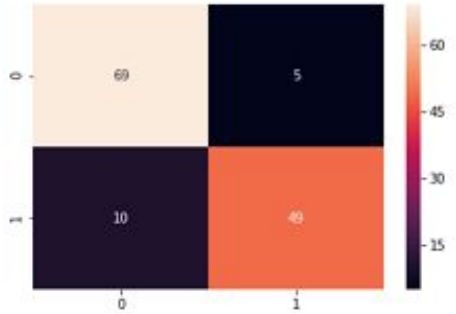
- Random forests adds an additional layer of randomness to bagging.
- It constructs a number of decision trees at training time and outputting the class that is the average of the classes output by individual trees.

Figure shows the accuracy and confusion matrix along with the precision, recall, f1-score and support for the Random Forest ensemble classifier.

Accuracy of the Random Forest Classifier is: 88.7218045112782

	precision	recall	f1-score	support
0.0	0.87	0.93	0.90	74
1.0	0.91	0.83	0.87	59
avg / total	0.89	0.89	0.89	133

Confusion Matrix



Algorithm Comparison

Evaluation Criteria	CART	ID3	Random Forests
Cohen Kappa Score	0.657	0.560	0.769
Mean Absolute Error	0.165	0.218	0.112
Root Mean Squared Error	0.406	0.466	0.335

An analysis of the Kappa Score, Mean Absolute Error and the Root Mean Squared Error show that Random Forest Ensemble Classifier gives the optimum result since the error in this technique is the least.

Conclusion

1. The best algorithm based on the patient's data is Random Forest Classification with accuracy of 88.72% and lowest average error at 0.112 compared to others.

2. Average error for ID3 and CART is 0.218 and 0.165 respectively.

3. Out of the two classifiers CART(83.45%) has more accuracy than ID3(78.14%).

Thank You.