CS771A Project Report : Assignment1

Siddhant Manocha
Roll 12714

January 20, 2015

1 Determining the Model for Classification

In the given assignment , we were asked to use Pima Indian Diabetes for prediction of diabetes among patients. The dataset has in total 768 data points and we use five fold cross validation to report our prediction results.
In order to build the model, we consider the following criterions

- Threshold the tree based on decrease in impurity value at the node
- Threshold on the number of nodes for a particular split or number of data vectors required at the node for the split
- Growing the complete tree and then pruning the tree
- Considering different impurity functions as gini and information(entropy)

We did 5 fold cross validation to determine the optimum value for various parameters.

- To set a threshold on the number of decrease in impurity, we iterated over values from 0 to 0.2 over a step size of 0.001 and found the threshold that maximizes the average accuracy over cross validation.
- To set threshold on number of nodes in split, we iterated over values from 0 to 150 to find its optimum value
- For pruning we iterated over values of the complexity parameter, which determines the extent upto which the tree will be pruned and determined its optimum value

2 Handling the Missing Data

In the given dataset, a lot of values were missing. In this case, missing values were not externally specified but the values were absurd for a given field. We considered the zero values in column 2,3,4,5,6 , namely amount of plasma glucose, blood pressure, triceps skin thickness, serum insulin, etc as the absurd values and replaced them by NA and handled them accordingly.

Steps taken for handling the missing values.
Test Set:

- Avoid missing attributes and perform impurity calculation by not considering the attributes if the value was missing
- Replace the missing values in positive examples by the mean of non missing samples having positive label and vice versa
- Replace the missing values in positive examples by the median of non missing samples having positive and vice versa
- Replace missing values by class based mean for the whole dataset

Training Set:

- Build surrogate splits at each node and make predictions on the basis the surrogate splits in case of missing attributes
- Replace the test data by the overall mean of the training data
- Replace the test data by the overall median of the training data
- Replace the missing values by class based means for the whole dataset

3 Results

3.1 Case1

Description: Do not replace the missing data in training. Use surrogate splits for testing.

- Threshold on number of nodes.Impurity function: information
  Accuracy: 76.95357 %
  Optimum Split: 23
- Threshold on decrease in impurity.Impurity function: information
  Accuracy: 76.16%
  Optimum threshold: 0.031 %
- Threshold on complexity parameter for pruning.Impurity function: information
  Accuracy: 73.57 %
  Optimum threshold:0.02
- Threshold on number of nodes.Impurity function: gini
  Accuracy: 76.17 %
  Optimum Split: 94
- Threshold on decrease in impurity.Impurity function: gini
  Accuracy: 75.12%
  Optimum threshold: 0.011 %
- Threshold on complexity parameter for pruning.Impurity function: gini
  Accuracy: 74.877 %
  Optimum threshold:0.04

3.2 Case2

Description:Replace the missing values by the class based(positive and negative) mean for the whole dataset .

- Threshold on number of nodes.Impurity function: gini
  Accuracy: 87.50 %
  Optimum Split: 36
- Threshold on decrease in impurity.Impurity function: gini
  Accuracy: 88.28 %
  Optimum threshold: 0.058
- Threshold on complexity parameter for pruning.Impurity function: gini
  Accuracy: 86.33 %
  Optimum threshold: 0.02

3.3 Case3

Description:Replace the missing values by the class based(positive and negative) medians for the whole dataset .

- Threshold on decrease in impurity.Impurity function: gini
  Accuracy: 87.88%
  Optimum Split: 24 %
- Threshold on complexity parameter for pruning.Impurity function: gini
  Accuracy: 86.84 %
  Optimum Split:0.016
- Threshold on number of nodes.Impurity function: gini
  Accuracy: 85.54 %
  Optimum Split: 0.02

3.4 Case4

Description:Replace the missing values in training by class based mean and testing set by mean of the training set

- Threshold on decrease in impurity.Impurity function: information
  Accuracy: 67.20%
  Optimum Split: 0.016 %
- Threshold on complexity parameter for pruning.Impurity function: information
  Accuracy: 68.61%
  Optimum Threshold:0.077
- Threshold on number of nodes.Impurity function: information
  Accuracy: 65.10 %
  Optimum Threshold: 0.58

4 Submission

For given assignment , I have submitted two files: hw_test.R and hw_test2.R
In hw_test.R :I have parameters: n_fold which specifies the fold for validation
thresholdtype : 1) for split size 2) decrease in impurity 3) pruning
impurity function : gini or information ,replace by : none , mean , median which specifies whether to replace data by mean , median or no change
Similarly for hwtest_2.R, except we can explicitly mention how to handle missing data for the test and the training

5 Comments

The best accuracy is achieved in case when the missing data is replaced by the mean of the overall data for the testing data as well as the training data. But such an approach may overfit our data and thus reports unexpectedly high accuracies.In case of using suurogate splits, results are reasonable and gives better results for missing data against using means to replace the data for training and testing seperately.

6 Tree Diagrams

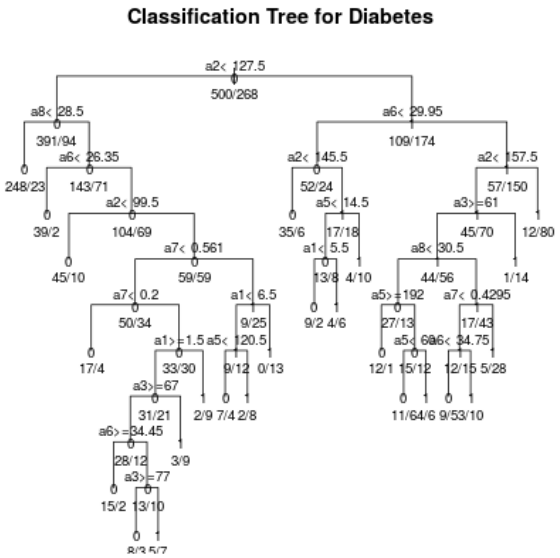

Figure 1: Full grown Tree

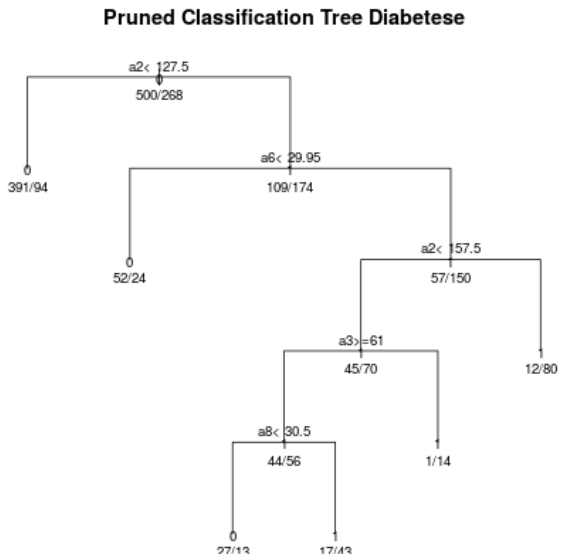

Figure 2: Pruned Tree at optimum complexity parameter