Friday, 7 April 2017

Decision Tree on Arrhythmia Dataset

Algorithm to continuous attributes:

To split the continuous attributes, we need to sort the training records with O(N logN) time. Then we can count the number of training records for each class with time complexity of O(N). The candidate split positions will be all the midpoints between to adjacent sorted values. So we can calculate the information gain for each candidate split positions. When computing the information gain, the time complexity is O(1), since we can matin two lists that records the numbers of records in each class for the two part of the data. Each time we shift our spilt point, we can update the lists in O(1) time can with these list, the information gain can be computed in constant time. Overall, the complexity is O(N logN).

Design for Decision Tree:

My implementation of decision trees obeys the following rules:

Missing Value: Filling the value with the most common value in that attribute

Splitting Rules:For all the attributes find the best splitting point and compute the information gain at the point, splitting the node with the attribute that has biggest information gain.

When to stop splitting:

- when the node hits the max depth of the tree
- when all the records in the tree are in the same class
- when we the biggest information gain for that node is 0.

The class label of the leaves:The most common class tag in that node.

Result:

The data set is small and strongly biased(245 of 452 records are normal). The decision stop growing at the depth of 12, since there's not enough data and good feature that allows the node to split. The optimal height of the decision tree is around 4, then the model is overfit and the accuracy is going down.