1. Zhongliu Li zlr8947
Xiaoyang Wang xwd2178

2. Yes. We add a field called "modeClass" to store the mode class passing the node in the training process. This field will be used in the pruning process to determine the pruned node's class.

3. Split the unknown attributes into different values base on the ratio of the known values.

Assign the '?' to v_i with a probability of p_i. $p_i = \frac{num(v_i)}{\sum_i num(v_i)}$

For example: if we have {'y': 3, 'n': 1, '?' : 10}, for each '?'-example, it will be assigned to 'y' with a probability of 75% and be assigned to 'n' with probability of 25%.

For each attributes, '?' only take a very small portion and these unknown values will have little influence on the distribution of the attribute values. Thus we can predict the missing value base on attribute values' distribution.

4. Post order DFS + Reduce Error Pruning

Traverse the tree from bottom up.

if the node is leaf, then return

if the node is non-leaf and all its children are leaf, then replace the node by a leaf node whose label is the 'modeCLass' of the original node. If the accuracy does not reduce, then keep the pruning. Otherwise undo the pruning.

We take this strategy because of its simplicity and high speed. This strategy traverses the tree one time and achieve a O(n) time complexity.

5. a. The accuracy tends to increase as the training set size increase.

Because, as training set increase, we can get more information about the model thus the model will be more accurate.

b. The advantage tends to increase.

As the data set size increase, the tree will be more complicate. And thus more node can be pruned to make the tree more general and overfitting error will be reduced. Thus the advantage will increase.