Fiona Tay
6 Jan 2012
Evaluating two TDIDT algorithms on three data sets
Objective: To compare the performance of ID3 with and without reduced-error pruning.

Methodology: An ID3 learner and an ID3 with reduced error pruning were each trained on data sets from the UCI machine learning repository. The experiment was run with 5-fold cross validation and repeated over 10 trials. A brief description of the datasets follows:

- iris (150) - measurements of flowers and the target is the class of iris
- mushroom (8124) - discrete qualities of mushroom and the target is its edibility
- restaurant (12) - discrete qualities of restaurant and the target is its availability

Results:
The average accuracy over all ten trials and over each 5-fold validation is shown in Table 1.

	ID3	ID3 with pruning
Iris	0.8467	0.8513
Mushroom	0.9997	1.0000
Restaurant	0.5400	0.4500

Discussion: For iris and restaurant, the performance of the pruning algorithm was poorer. I suspected that reserving some data for the pruning stage reduced the quality of the decision trees generated during training, so I compared the sizes of the trees before and after pruning. For ID3, the average tree height was 4.94, while for ID3 with pruning, the average height was 2.84. This strongly suggests overfitting, caused by the small size of the validation set. Similarly, since restaurant is a very small and noisy data set (many attributes compared to number of data), we see overfitting.

For mushroom, the performance of both algorithms were very similar. I cross-validated mushroom across more trials and concluded that the differences in performances were not statistically significant. The trees generated by both algorithms had average height 5, suggesting that they came up with similar decision trees. That pruning performed as well as ID3 is because the data set is large and reserving ⅓ as validation did not deprove the training process, since there was sufficient data.

Conclusion: ID3 performed as well or better than ID3 with pruning in all of the examples. For the two small data sets, setting aside some data for validation severely decreased the quality of the training process and thus the tree before pruning captured less information. For mushroom, I did not observe any overfitting, so pruning had no benefit.