Project 1 Report

Joseph Allen

Overview

This program satisfies the requirements of Project 1, which state to build decision trees based on Quinlan's ID3 and C4.5 algorithms. Both implementations follow a decision tree learning algorithm template provided in class slides, specifically, the one in slide 5 of the slides titled "dt.pdf". What sets the two algorithms apart is the importance function they use to select attributes while building the decision tree. The ID3 implementation uses Information Gain as the importance measure and the C4.5 implementation uses the Information Gain Ratio, which is basically Information Gain normalized by attribute value prevalence.

Results

The results achieved show that both decision tree learning algorithms learned trees that perfectly fit the training data, as well as the testing data. The F1-score in all cases is 1.0, meaning that the decision tree has perfect precision and recall for both training and testing datasets. Upon inspection of the trees learned with both ID3 and C4.5, both algorithms chose 'odor' as the most important attribute for the tree with 'spore-print-color' as the second most important. The two trees are identical for these attributes, but the algorithms chose different attributes further on. Besides the first two attributes and their leaves, the ID3 tree has 3 more attribute nodes with 9 leaves and the C4.5 tree has 4 more attribute nodes with 5 leaves. These results can be seen in the image included at the end of this document (please excuse the fact that they are hand-drawn). Overall, the C4.5 algorithm produced a deeper but smaller tree in terms of total nodes and leaves. The perfect F1-score on this dataset for both trees indicates that it is perfectly separable with multiple partitions.