NORWEGIAN UNIVERSITY OF SCIENCE AND TECHNOLOGY
TDT4171 – ARTIFICIAL INTELLIGENCE METHODS

Assignment 4

Per Magnus Veierland
permve@stud.ntnu.no

April 17, 2016

1 RANDOM ATTRIBUTE SELECTION

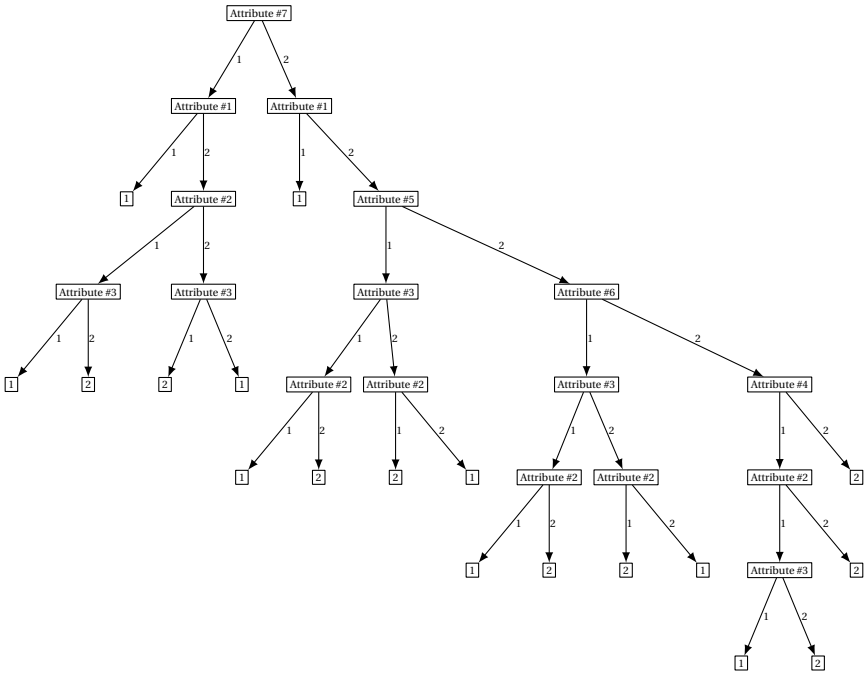

Figure 1: Decision tree based on random attribute selection.

Figure 1 shows a decision tree created from the `training.txt` examples using uniform random numbers in the range $[0.0, 1.0)$ as the IMPORTANCE function with the DECISION-TREE-LEARNING algorithm.

When evaluated using the examples from `test.txt`, the decision tree created with random attribute selection correctly classified $^{25}/_{28}$ examples (89.29% accuracy).

2 INFORMATION GAIN BASED ATTRIBUTE SELECTION

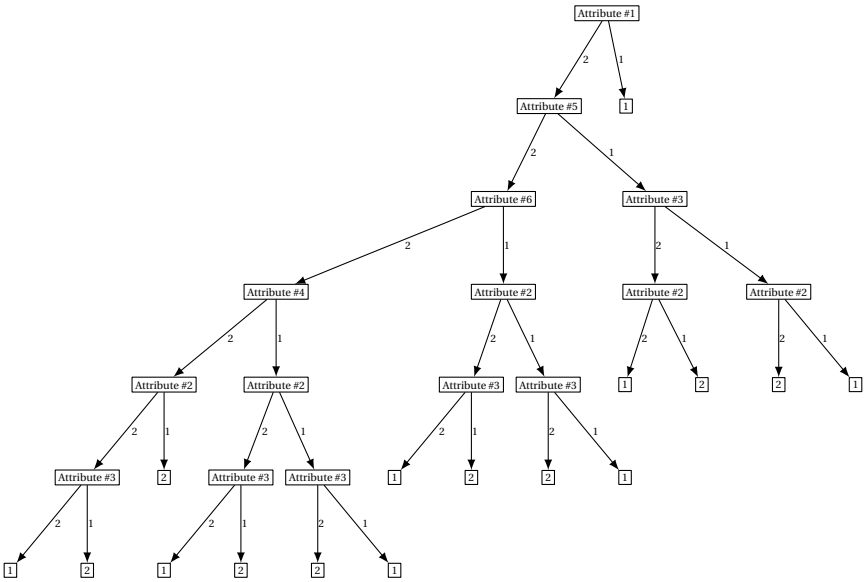

Figure 2: Decision tree based on information gain attribute selection.

Figure 2 shows a decision tree created from the `training.txt` examples using an information gain based IMPORTANCE function with the DECISION-TREE-LEARNING algorithm.

When evaluated using the examples from `test.txt`, the decision tree created with information gain based attribute selection correctly classified $^{26}/_{28}$ examples (92.86% accuracy).

3 FINDINGS

When using information gain based attribute selection, it is not possible to achieve 100% classification accuracy with the given training and test data. To verify that the data is consistent, a decision tree was trained with both the training and test data. This tree was constructed without the PLURALITY-VALUE function being invoked, and was able to correctly classify all training- and test data examples. This confirms that the training- and test data is consistent.

With random attribute selection, some trained decision trees were able to correctly classify all test examples. This suggests that there are enough examples in the training data to correctly classify all test examples, but that whether this can be achieved using decision trees

depends on the order of attribute comparison. Improving the stability of the average accuracy of the trained decision trees requires more training data.

Using random attribute selection can be better than information gain based attribute selection for the following reasons:

1. In the case of few training examples, random attribute selection can result in a decision tree which achieves a better test accuracy due to a more beneficial attribute selection order.

2. As the DECISION-TREE-LEARNING algorithm is based on greedy search, random attribute selection can result in more optimal decision trees than trees found with information gain based attribute selection.

However in the average case, the information gain based approach is likely to be better, as in the case of few training examples the resulting model is not likely to be good, even though selecting attributes at random may be able to correctly classify all test examples, and in the case where there is enough training data, selecting attributes based on information gain should result in smaller decision trees which can be evaluated faster, as their distance from root to answer in the average case will be shorter than in decision trees based on random attribute selection.

Running the DECISION-TREE-LEARNING algorithm multiple times with random attribute selection, the test classification results ranged from $^{19}/_{28}$ (67.86%) to $^{28}/_{28}$ (100%). The reason for the varying results is that the attribute selection order is random, and depending on the attribute evaluation order the training data will either be lacking or sufficient to correctly classify the test examples.

Running the DECISION-TREE-LEARNING algorithm multiple times with information gain based attribute selection, the test classification results ranged from $^{19}/_{28}$ (67.86%) to $^{26}/_{28}$ (92.86%). The reason for the varying results is that there are several scenarios where multiple attributes yields the same information gain, which results in one of the tied attributes being selected randomly.