CS3600 - Project 4 Analysis

Saumya Jain

Question 6

Test Dummy Set 1 has 20 examples with a tree size of 3. The classification rate for this set is 1.0, or 100%. This expected classification rate was also 1.0 or 100%. The reason for such a high classification rate is that the tree only has a size of 3, which is very small. The state does not depend on several tree features in such a small tree. Hence, a test case can be classified even after evaluating only some of its features. The testing data is also drawn from a similar population as the training data and has appropriate representation of the classes in the training data. The data is also binary.

Test Dummy Set 2 has 20 examples with a tree size of 11. The classification rate for this set is 0.65, or 65%. The expected classification rate was 0.55 or 55%. The actual rate we attained is off by 0.1 or 10%, because the data couldn't be narrowed down into separate differing attributes, leading it to be randomly placed. We expected this outcome because the tree is much larger in this set than the previous case (tree size = 11 vs tree size = 3). The number of features in Test Dummy Set 1 and 2 are the same and both trees are binary. The reason why the classification rate is lower is because we need to ask more questions when classifying each test case.

Connect4 has 67557 examples with a tree size of 41521. The classification rate for this set is 0.75385 or 75.385%. The expected classification rate was 0.75 or 75%. The actual rate is only slightly off of the expected rate. The classification rate is very accurate as the test was run 10 times with a test size of 2000 each time, leading to a better evaluation of data. Here we have a tree with a very large size. This tree has several examples used along with a very high number of attributes. Some questions were left out due to logic.

Cars has 1728 examples with a tree size of 408. The classification rate for this set is 0.945, or 94.5%. We expected a classification rate of around 0.95 or 95%. The rates are only slightly off. We can see that the classification rate is very high and close to our expected classification rate, even though the tree size is very big. This proves that we can attain high classification rates even with large trees. There exist more examples in this set to train on. There are many features, each with greater than 2 options, along with a non-binary final classification. Having a large data set and lots of features complements having a large tree and aids to the large classification rate.

Question 7

When thinking about the cars dataset, an online shopping website comes to mind. We can take Dell as an example. Along with the UI, the decision tree can be used to give a user exactly the product they desire. The dataset will include attributes such as type (laptop, 2-in-1, tablet, desktops), color (gray, black, white, red, blue), processor (i3, i5, i7), RAM (8GB, 12GB, 16GB, 32GB), etc. The largest split attribute would be at the very top of the tree (probably the device type) and each node down would decrease the split factor.

We can improve the connect4 dataset to make a better playing bot. If we use particle filtering, we would be able to compute which holes are filled and which are unfilled. This would help us narrow the decision-making capacity of the bot and cause it to make smarter decisions. If a single hole has a particle in a branch, the bot can just skip going into that branch and make better decisions.