Jordan Shartar
Jshartar6

Q6: For each data set, describe why you think the decision tree learning performed as it did in terms of accuracy statistics and tree size. Please be specific and justify your claims (ie: if you say there was not enough training data for data set X, or the nature of training data made Decision Tree learning less effective, what makes you think so?). Note that you can inspect the two fake data sets by looking into DataInteface.py and finding the lists that contain them, and that you can inspect the real datasets in the datasets folder. To get full credit you should report all the information asked for (classification rate and tree size) for each data set, and briefly give some justification for why each dataset got the results that it did regarding both the classification rate and tree size.

ANSWER: (view "Testing Output Q6.txt") to see outputs of my code

Dummy dataset 1: The decision tree learning performed well. It gave accurate results for the dataset. For the fifth attribute the value is equal to one or zero thus resulting in a tree of size three. That is a root with two leaf nodes. The fifth attribute was chosen because it has the largest information gain thus is used to split the data. Now when testing data the outputs are all correct making the classification rate 1.0.

Dummy dataset 2: The decision tree learning performed badly for dummy dataset 2. The best attribute was the second attribute (most information gain attribute). Which gave a classification rate of 0.65 and a tree size of 11. These result in the decision tree performing badly. Because attribute 2 performed badly, it is not the best attribute for the testing data. Therefore, additional attributes may need to be added to further classify the data in the testing set more accurately. The training dataset was of size 20. Increasing the amount of data available for training can also improve accuracy as it will have more information and general trend to classify with its attributes.

Connect 4 dataset: The decision tree learning did alright, not very well slightly better in comparison to dummy dataset 2. There are 67,557 number of examples which equates to 42 attributes that have 3 values each. Each attribute has a minimal effect on classifying the data set since there are so many (such as first attribute not classifying data at all). Due to the minimal effect each attribute has, the tree size is 41,521 (a very large tree). The average classification rate over 10 runs with test size of 2000 is 0.75965 (slightly better than dummy dataset 2 but still not that good). This could be caused by different things. One obvious reason is the number of examples being so large (67,557). Another one is the noise of the dataset during training. Since there is so much data a lot of noise will skew the accuracy of the data. A suggested fix I have to this issue would be to prune the tree to get rid of any unnecessary attributes (like what needs to be done in the extra credit part of the assignment), this will help the tree more accurately predict the outcomes for the testing data.

Car dataset: The decision tree learning performed well. There are 1728 number of examples. The reason is that there were six attributes with half having four values and the second half having values, giving the combination of 1728. Each attribute helps classify the data into a tree of size 408. Because each splits the data it creates a small tree in comparison to the data inputs. The average classification rate of the 20 runs is 0.944750. This is a high rate because the data set

is small for the number of attributes given. When testing because of this high accuracy and small data sets, the chance of overfitting is minimal.