Name : Shishir Kumar Prasad

CS ID : skprasad

CS760, HW1 - Paper Submission (Decision Tree Learning)

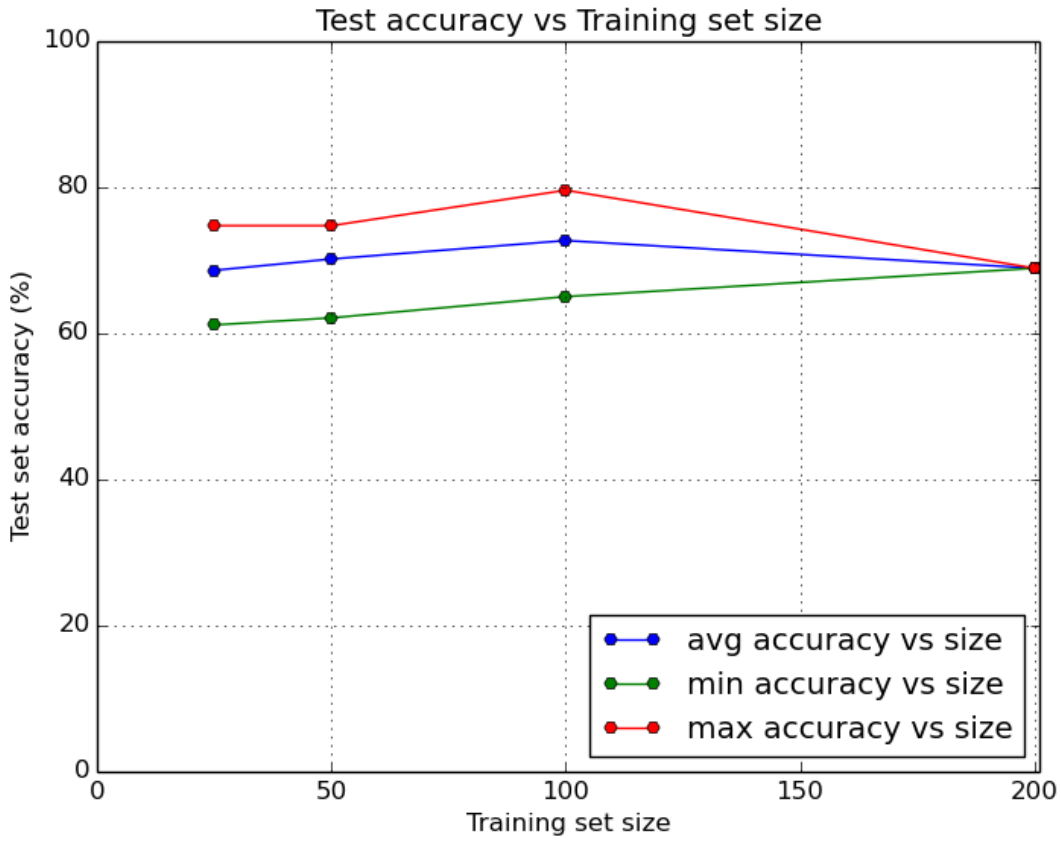
Part 2

For this part, you should use heart_train.arff as your training set and heart_test.arff as your test set.

Set the stopping criterion m=4.

Plot a learning curve for your decision-tree learner. You should plot points for training set sizes of 25, 50, 100, and 200 instances. For each training-set size (except the largest one), randomly draw 10 different training sets using stratified sampling and evaluate each resulting decision tree model on the test set. For each training set size, plot the average test-set accuracy and the minimum and maximum test-set accuracy. Be sure to label the axes of your plot.

Solution



Explanation

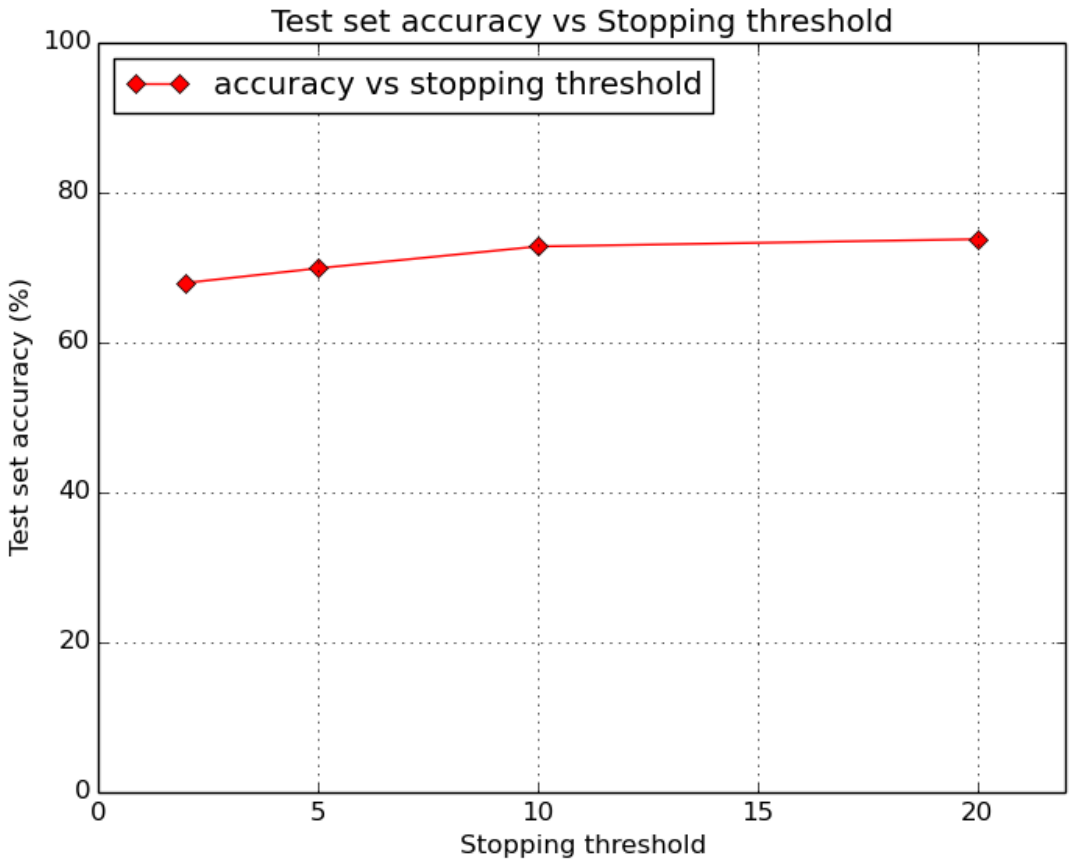
As we increase the size of training data set, test data accuracy increases because with the increase in the number of labelled examples, the decision tree is able to capture more real-world data patterns. But after a certain threshold of training dataset size, the test data accuracy starts decreasing because larger training data set might have some spurious examples which can lead to over-fitting.

Part 3

For this part, you should use heart_train.arff as your training set and heart_test.arff as your test set.

Using the entire training set, plot a curve showing how accuracy varies with the value m used in the stopping criteria. Show points for m = 2, 5, 10 and 20. Be sure to label the axes of your plot.

Solution



Explanation

As we increase the value of 'm' i.e. the stopping criteria threshold , the test data accuracy increases because it reduces the chances of over-fitting. But after a certain threshold value, the increase in test data accuracy becomes relatively lesser and might start further decreasing because of under-fitting the data.