CS573 - Lab3 Report

Phuong Nguyen

April 2020

1 Overview

We used Weka for this lab. For Decision Tree classifiers, we used J48 in Weka.

2 Tasks

1. Learn a Decision Tree classifiers on the data set. Visualize the tree constructed by the decision tree algorithm.(Food for thought: Are there some interesting rules that make sense based on what you understand about the data?)

   The visualized tree is shown in figure 1.

   There are some interesting rules that makes sense based on our understanding:

   - If someone does not support "physician-free-freeze", then he or she is most likely a democrat.
   - If someone supports "physician-free-freeze", and does not support "synfuels-corporation-cutback", then he/she is most likely a republican.
   - Democrats do not want "anti-satellite-test-ban".
   - Republicans do not want "adoption-of-the-budget-resolution".

2. Report the accuracy of the Decision Tree classifier using 5-fold cross-validation. Report 95% confidence interval.

   The accuracy is: 96.5517%.

   Number of correctly classified instances: 420.
   Total number of instances: 435.
   The accuracy rate:

$$\hat{p} = \frac{a}{n} = \frac{420}{435} = \frac{28}{29}$$

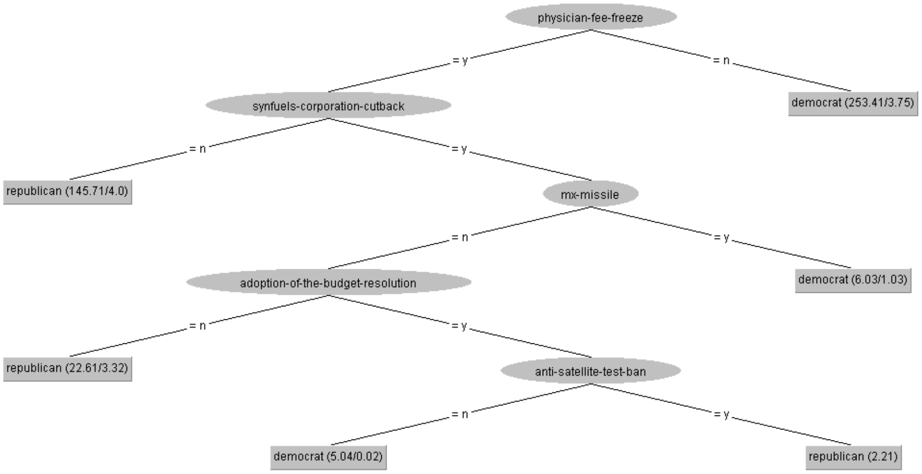

Figure 1: Tree constructed by 5-fold cross validation.

The standard deviation $\sigma_{\hat{p}}$:

$$\sigma_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{\frac{28}{29}(1-\frac{28}{29})}{435}} = 8.7486 \times 10^{-3}$$

With 95% confidence interval $\Rightarrow z_N = 1.96$.

So, the estimated accuracy of the classifier is p in:

$$[\hat{p} - z_N \cdot \sigma_{\hat{p}},\ \hat{p} + z_N \cdot \sigma_{\hat{p}}] = [0.9484, 0.9827]$$

3. Perform the following experiments to study the stability of decision tree learning algorithm over the variability of data samples.

   (a) Randomly split the data set into 5 data sets of (roughly) equal size $D_1, D_2, ..., D_5$.

   Run the python file lab3.py, we will get:

   - 5 csv files of equal size (87 samples each) in the current working folder: testing_1.csv, testing_2.csv, testing_3.csv, testing_4.csv, testing_5.csv. These files will be used as testing set.
   - 5 csv files of equal size (348 samples each) in the current working folder: training_1.csv, training_2.csv, training_3.csv, training_4.csv, training_5.csv. These files will be used as testing set. We have: {training_x} = {all data set} - {testing_x}.

(b) For $i = 1, 2, ..., 5$, each time use D_i as test data and the rest as training data to learn a decision tree and measure its accuracy p_i.

- $i = 1$: Use training_1.csv for training and testing_1.csv for testing. The accuracy is: 95.4023%.
- $i = 2$: Use training_2.csv for training and testing_2.csv for testing. The accuracy is: 96.5517%.
- $i = 3$: Use training_3.csv for training and testing_3.csv for testing. The accuracy is: 98.8506%.
- $i = 4$: Use training_4.csv for training and testing_4.csv for testing. The accuracy is: 94.2529%.
- $i = 5$: Use training_5.csv for training and testing_5.csv for testing. The accuracy is: 95.4023%.

(c) Visualize the five trees constructed. Do the five trees differ with each other and with the tree constructed using all the data (in Task 1)? How much do their accuracy differ?

The visualized tree is shown in figures below.

Yes, five trees different with each others and with the tree constructed using all the data but the first two nodes (physician-fee-freeze and synfuels-corporation-cutback) are similar throughout. Their accuracy are almost the same, around 95%.

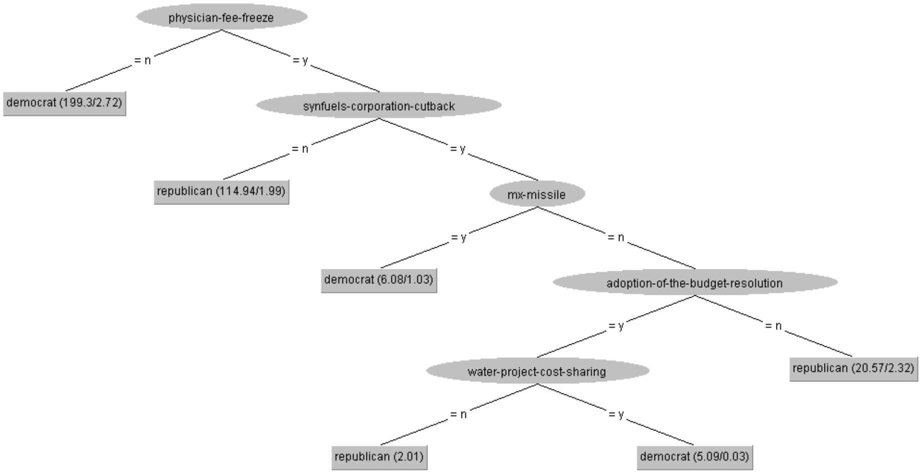

Figure 2: Tree constructed by using D_1 as testing data.

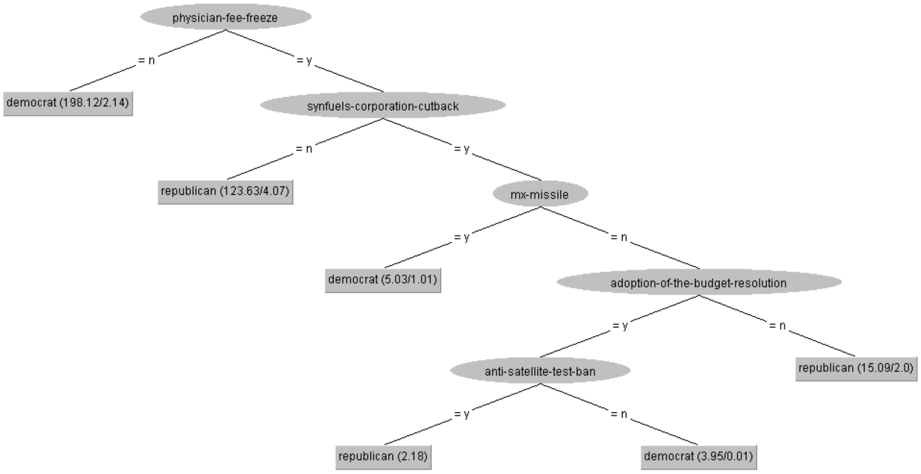

Figure 3: Tree constructed by using D_2 as testing data.

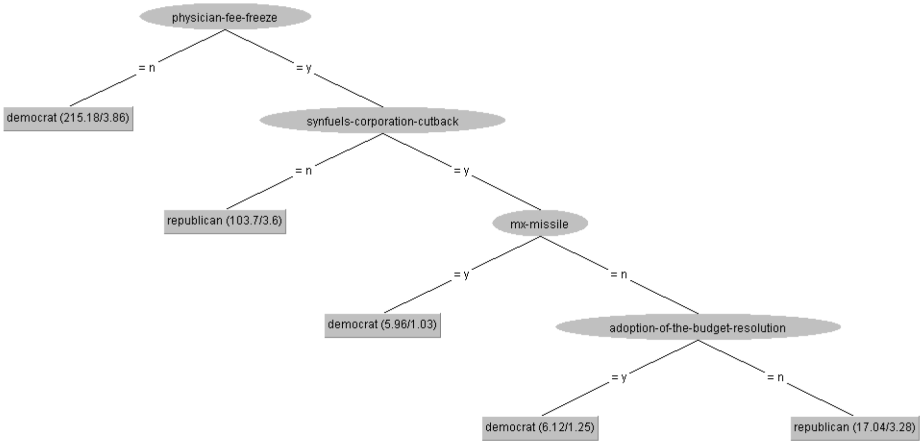

Figure 4: Tree constructed by using D_3 as testing data.

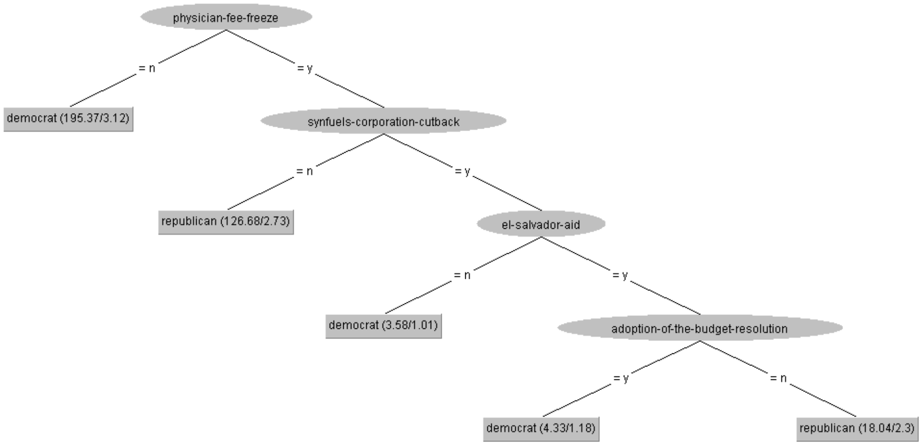

Figure 5: Tree constructed by using D_4 as testing data.

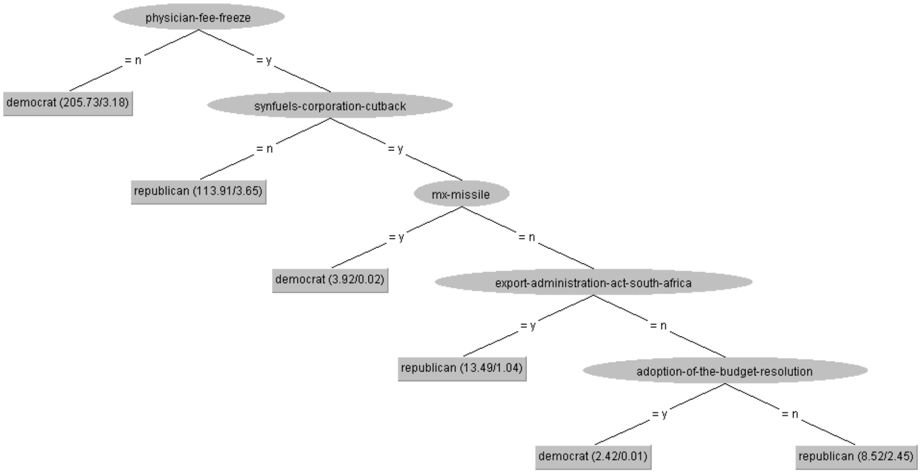

Figure 6: Tree constructed by using D_5 as testing data.