Yonghyun_Lab3

March 29, 2020

1 Data Import and Load

[1]:

```
import numpy as np
import pandas as pd

import arff

#sudo pip install python-weka-wrapper3
#sudo pip install javabridge
#Go to https://fracpete.github.io/python-weka-wrapper/install.html for more␣
 ↪information
import weka
import weka.core.jvm as jvm
from weka.core.dataset import create_instances_from_lists, Instances
import weka.core.converters as converters
from weka.core.converters import Loader
from weka.classifiers import Classifier, Evaluation
from weka.core.classes import Random
import weka.plot.graph as graph

print("Numpy version = %s" % np.__version__)
print("Pandas version = %s" % pd.__version__)
print("arff version = %s" % arff.__version__)
print("python-weka-wrapper3 version = %s" % "0.1.12")
print("javabridge version = %s" % "1.0.18")

jvm.start()
```

```
DEBUG:weka.core.jvm:Adding bundled jars
DEBUG:weka.core.jvm:Classpath=['C:\\Users\\ghkfk\\Anaconda3\\lib\\site-
packages\\javabridge\\jars\\rhino-1.7R4.jar',
'C:\\Users\\ghkfk\\Anaconda3\\lib\\site-
packages\\javabridge\\jars\\runnablequeue.jar',
'C:\\Users\\ghkfk\\Anaconda3\\lib\\site-
packages\\javabridge\\jars\\cpython.jar',
'C:\\Users\\ghkfk\\Anaconda3\\lib\\site-packages\\weka\\lib\\python-weka-
wrapper.jar', 'C:\\Users\\ghkfk\\Anaconda3\\lib\\site-
packages\\weka\\lib\\weka.jar']
DEBUG:weka.core.jvm:MaxHeapSize=default
DEBUG:weka.core.jvm:Package support disabled
```

```
Numpy version = 1.16.5
Pandas version = 0.25.1
arff version = 2.4.0
python-weka-wrapper3 version = 0.1.12
javabridge version = 1.0.18
```

[2]:

```
# import data
data_dir = "https://archive.ics.uci.edu/ml/machine-learning-databases/
 ↪voting-records/"
data1 = pd.read_csv(data_dir + "house-votes-84.data", header=None)
data1 = data1.reindex(columns=[1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,0])
data1 = np.array(data1)

# create arff file
obj = {
   'description': u'',
   "relation": "vote",
   'attributes': [
       ("handicapped-infants", ["n", "y"]),
       ('water-project-cost-sharing', ['n', 'y']),
       ('adoption-of-the-budget-resolution', ['n', 'y']),
       ('physician-fee-freeze', ['n', 'y']),

       ('el-salvador-aid', ['n', 'y']),
       ('religious-groups-in-schools', ['n', 'y']),
       ('anti-satellite-test-ban', ['n', 'y']),
       ('aid-to-nicaraguan-contras', ['n', 'y']),

       ('mx-missile', ['n', 'y']),
       ('immigration', ['n', 'y']),
       ('synfuels-corporation-cutback', ['n', 'y']),
       ('education-spending', ['n', 'y']),

       ('superfund-right-to-sue', ['n', 'y']),
       ('crime', ['n', 'y']),
       ('duty-free-exports', ['n', 'y']),
       ('export-administration-act-south-africa', ['n', 'y']),

       ('\'Class\'', ['democrat', 'republican']),
   ],
   'data': data1,
}
fp = open("vote2.arff", "w")
arff.dump(obj, fp)
```

```
fp.close()

# load data
data = converters.load_any_file("vote2.arff")
data.class_is_last()
```

2 Decision Tree modeling and analysis

2.1 Learning a Decision Tree Classifier

In order to learn a decision tree on the vote data set, we use *J48* in *Weka*.

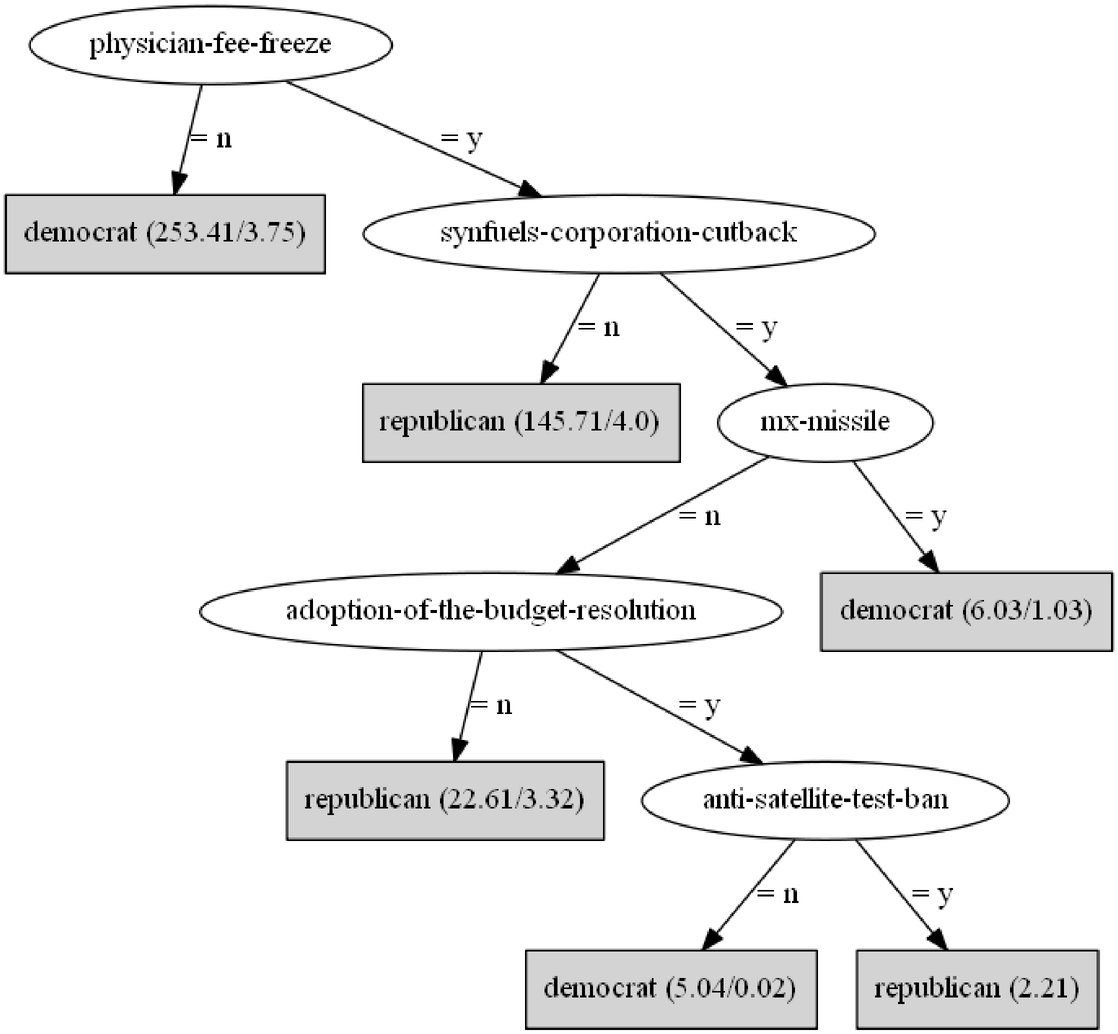
[3]:

```
cls = Classifier(classname="weka.classifiers.trees.J48")
cls.build_classifier(data)
#print(cls.to_help())
print(cls)
```

```
J48 pruned tree
------------------

physician-fee-freeze = n: democrat (253.41/3.75)
physician-fee-freeze = y
|   synfuels-corporation-cutback = n: republican (145.71/4.0)
|   synfuels-corporation-cutback = y
|   |   mx-missile = n
|   |   |   adoption-of-the-budget-resolution = n: republican (22.61/3.32)
|   |   |   adoption-of-the-budget-resolution = y
|   |   |   |   anti-satellite-test-ban = n: democrat (5.04/0.02)
|   |   |   |   anti-satellite-test-ban = y: republican (2.21)
|   |   mx-missile = y: democrat (6.03/1.03)

Number of Leaves  :     6

Size of the tree :      11
```

[4]:

```
graph.plot_dot_graph(cls.graph, "Tree.png")
```



One of the most important attribute to classify voting records is *physician-fee-freeze*. If one voted against or announced against freezing physizian fee, it highly implies that that person affiliates democrat. This makes sense because democrats insists on more expenditure on health care and medical service. The next important attribute is *synfuels-corporation-cutback*. Republicans are about 30000% more likely to be opposed to cutback on Synthetic Fuels Corporation.

Also, *mx-missile* and *education-spending* was the following important attributes in the tree. People who think that the nation should spend more on education-spending seemed to affiliate republicans. The other attributes were not considered important when classifying party affiliation.

2.2 k-fold cross-validation

[5]:

```
n = 5
evaluation = Evaluation(data)                           # initialize with priors
evaluation.crossvalidate_model(cls, data, n, Random(1))  # 5-fold CV
print("Accuracy = %g" % evaluation.percent_correct + "%")
```

```
z = 1.96
accuracy = evaluation.percent_correct/100
margin = z * np.sqrt( (accuracy * (1 - accuracy)) / n)
print("95% "+"Confidence Interval = (%g, %g)" % (accuracy - margin, accuracy +␣
 ↪margin))

print(evaluation.summary())

#print("Number of incorrect = %g" % evaluation.incorrect)
print(evaluation.class_details())
```

```
Accuracy = 96.5517%
95% Confidence Interval = (0.805579, 1.12546)

Correctly Classified Instances         420               96.5517 %
Incorrectly Classified Instances        15                3.4483 %
Kappa statistic                          0.9275
Mean absolute error                      0.059
Root mean squared error                  0.1731
Relative absolute error                 12.4478 %
Root relative squared error             35.5458 %
Total Number of Instances              435

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC
Area  PRC Area  Class
                 0.966    0.036    0.977      0.966    0.972      0.928    0.967
0.967     democrat
                 0.964    0.034    0.947      0.964    0.956      0.928    0.967
0.932     republican
Weighted Avg.    0.966    0.035    0.966      0.966    0.966      0.928    0.967
0.953
```

It turns out that the accuracy of the decision tree is 96.5517%, so we can use this model to classify one's political view. We can see that almost all of the elements in the above table is close to 1.

The accuracy is based on the whole data. We trained the model based on the whole data and obtain accuracy using the same data, but we can split training and test data to compute more accurate accuracy(as we will see in the next section).

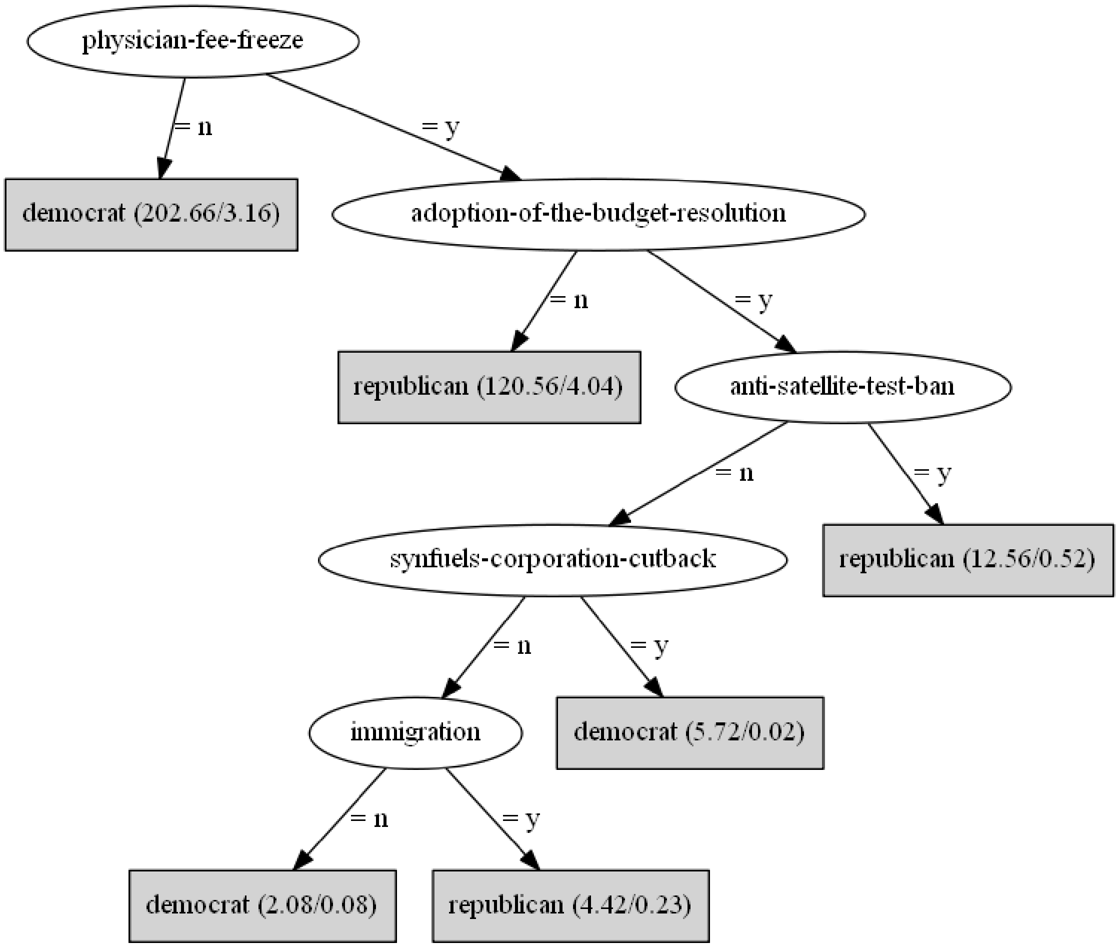
Note that the 95% confidence interval includes 1. Since accuracy is strictly less than 1, one can imporve this interval by increasing the number of iterations(in this experiment, 5)

2.3 Stability of decision tree

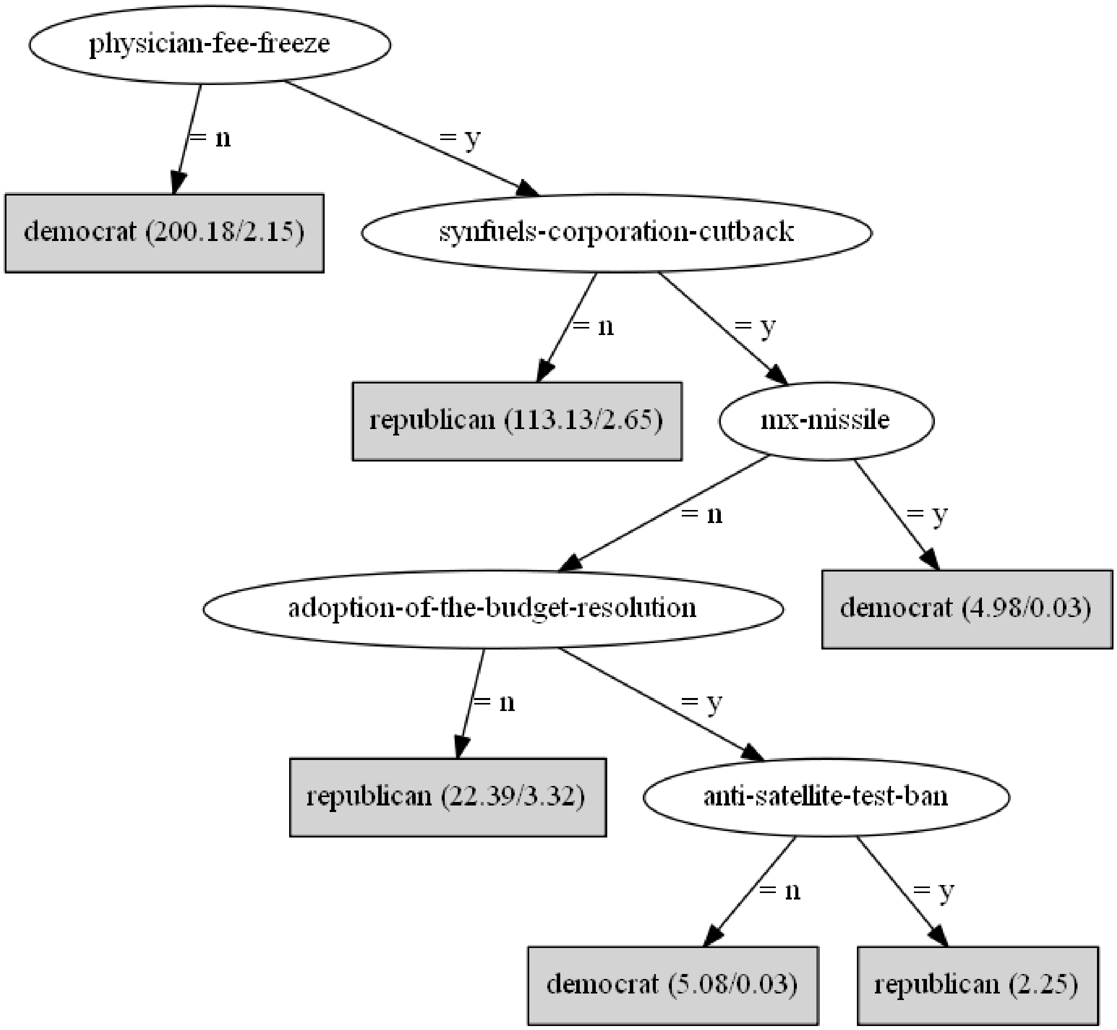
[6]:

```
n = 5
seed = 1
rnd = Random(seed)
rand_data = Instances.copy_instances(data)
rand_data.randomize(rnd)
classifier = Classifier(classname="weka.classifiers.trees.J48")

for i in range(n):
    # randomely splilt the dataset
    train = rand_data.train_cv(n, i)
    test = rand_data.test_cv(n, i)

    # split train and test datasetand measure accuracy
    cls = Classifier.make_copy(classifier)
    cls.build_classifier(train)
    evaluation = Evaluation(rand_data)
    evaluation.test_model(cls, train)

    print("-------------%g-th fold-------------" % i)
    print("Accuracy for training data = %g" % evaluation.percent_correct + "%")

    evaluation = Evaluation(rand_data)
    evaluation.test_model(cls, test)
    print("Accuracy for test data = %g" % evaluation.percent_correct + "%")

    #Visualize five trees constructed
    graph.plot_dot_graph(cls.graph, ("Tree" + str(i) + ".png"))
```

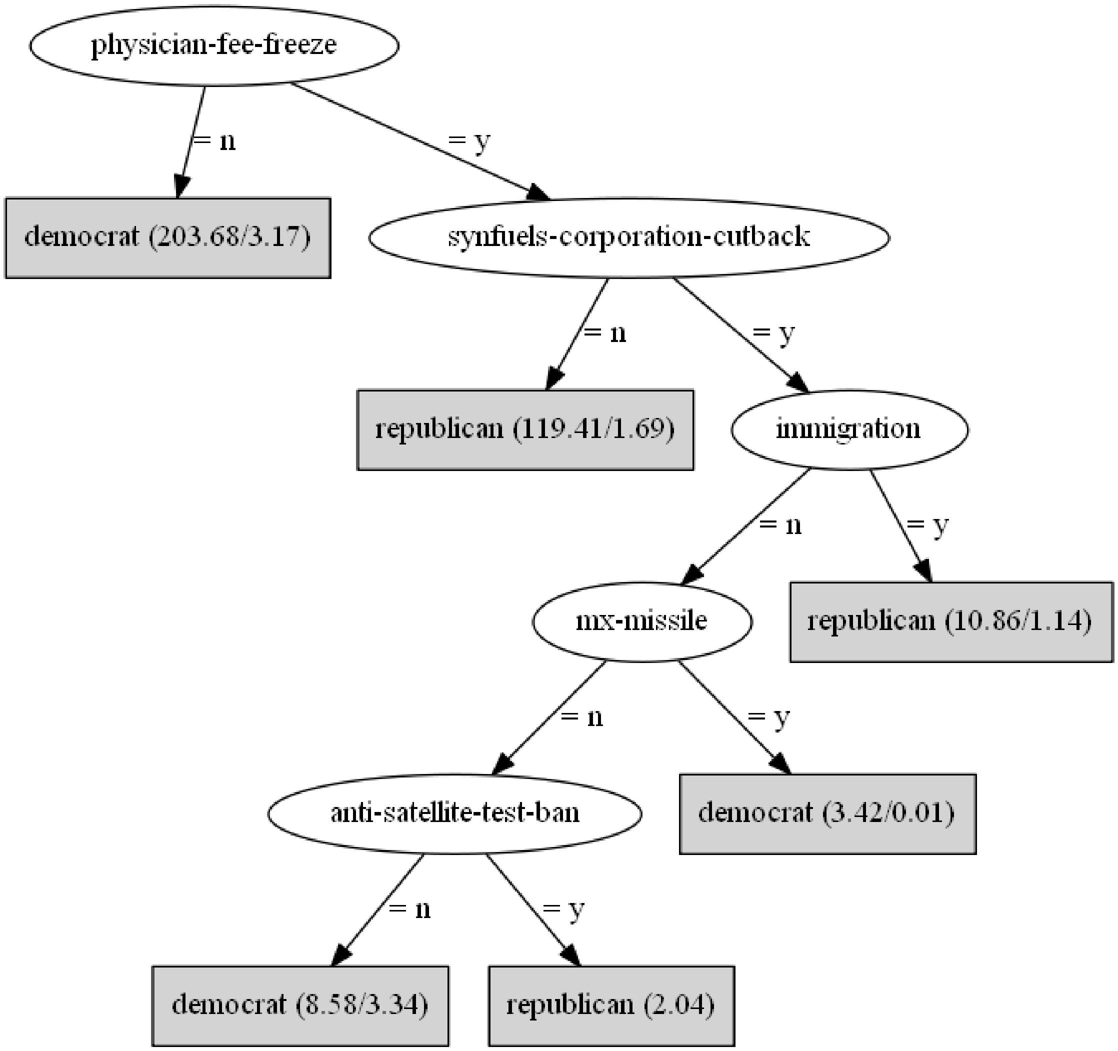
```
-------------0-th fold-------------
Accuracy for training data = 97.9885%
Accuracy for test data = 94.2529%
-------------1-th fold-------------
Accuracy for training data = 97.7011%
Accuracy for test data = 95.4023%
-------------2-th fold-------------
Accuracy for training data = 97.4138%
Accuracy for test data = 88.5057%
-------------3-th fold-------------
Accuracy for training data = 96.2644%
Accuracy for test data = 98.8506%
-------------4-th fold-------------
Accuracy for training data = 96.8391%
Accuracy for test data = 94.2529%
```


physician-fee-freeze
= n
= y
democrat (202.66/3.16)
adoption-of-the-budget-resolution
= n
= y
republican (120.56/4.04)
anti-satellite-test-ban
= n
= y
synfuels-corporation-cutback
republican (12.56/0.52)
= n
= y
immigration
democrat (5.72/0.02)
= n
= y
democrat (2.08/0.08)
republican (4.42/0.23)

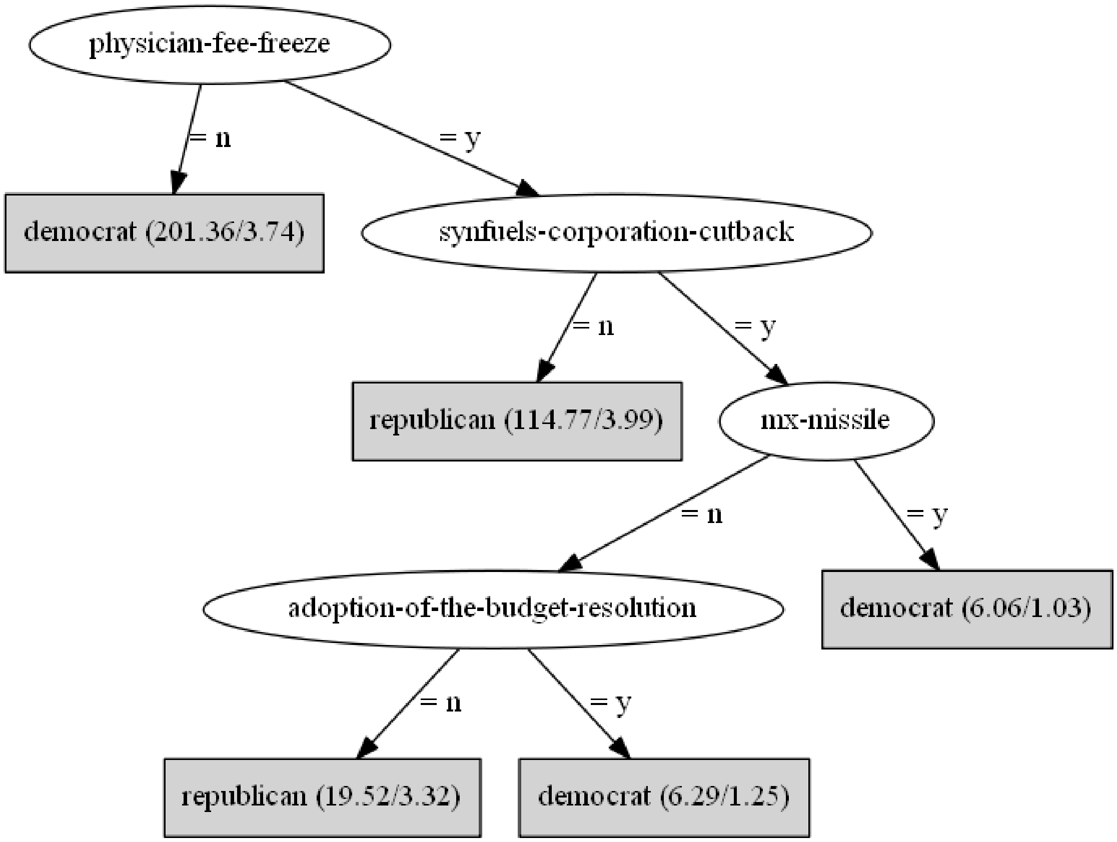
Tree0


physician-fee-freeze
= n
= y
democrat (200.18/2.15)
synfuels-corporation-cutback
= n
= y
republican (113.13/2.65)
mx-missile
= n
= y
adoption-of-the-budget-resolution
democrat (4.98/0.03)
= n
= y
republican (22.39/3.32)
anti-satellite-test-ban
= n
= y
democrat (5.08/0.03)
republican (2.25)

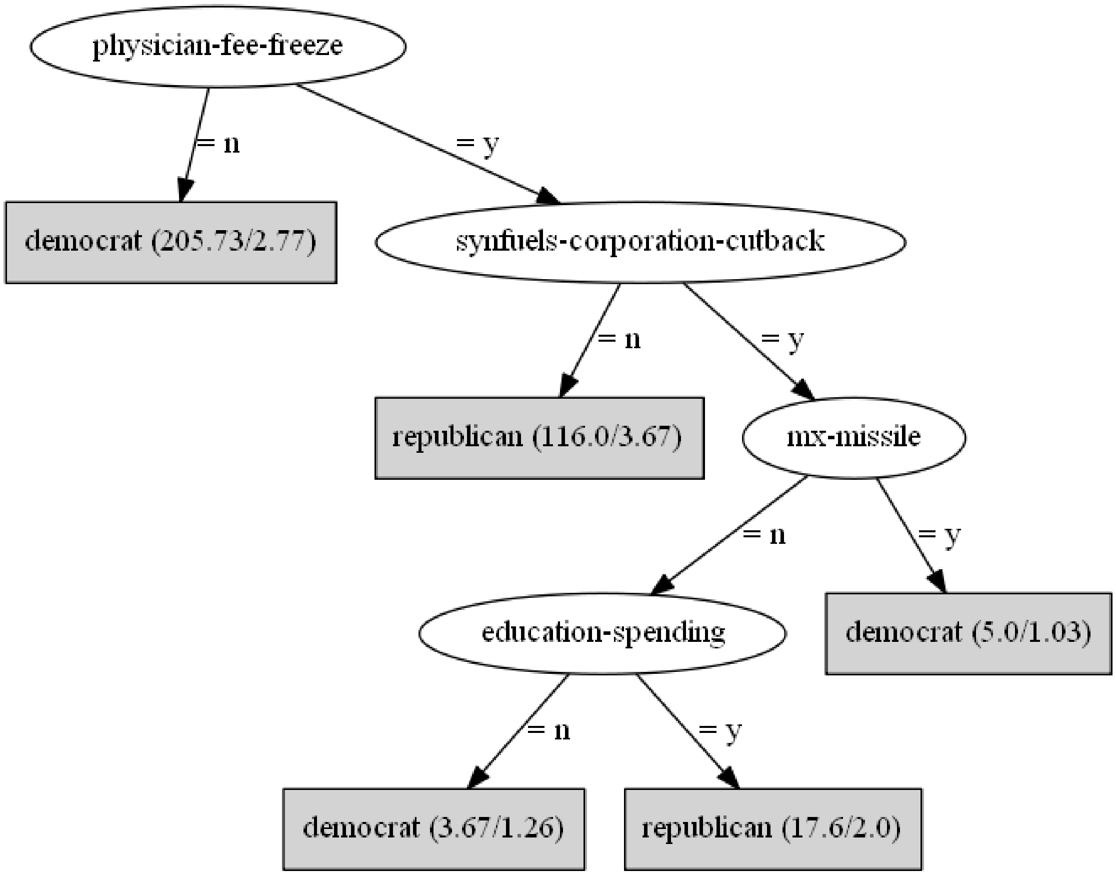
Tree1


physician-fee-freeze
= n
= y
democrat (203.68/3.17)
synfuels-corporation-cutback
= n
= y
republican (119.41/1.69)
immigration
= n
= y
mx-missile
republican (10.86/1.14)
= n
= y
anti-satellite-test-ban
democrat (3.42/0.01)
= n
= y
democrat (8.58/3.34)
republican (2.04)

Tree2


physician-fee-freeze
= n
= y
democrat (201.36/3.74)
synfuels-corporation-cutback
= n
= y
republican (114.77/3.99)
mx-missile
= n
= y
adoption-of-the-budget-resolution
democrat (6.06/1.03)
= n
= y
republican (19.52/3.32)
democrat (6.29/1.25)

Tree3



Tree4 One can observe that all of the five trees are similar and they resemble the tree that is constructed using all the data in Task1. K-fold cross validation result shows that decision tree learning algorithm is stable enough, because using different training and test data does not bring out significantly different trees. We can check that the first node is always *physician-fee-freeze* but the second node varies depending on which dataset is used as test data. Putting all five experiments together, we can conclude that *physician-fee-freeze*, *synfuels-corporation-cutback* and *mx-missile* are important factors(or attribute) for classifying two different groups.

Also, 5 experiments show consistent accuracy except accuracy for test data in third experiment. Outliers of test data in the third experiment may have caused lower accuracy of the model.