Homework 4 Report

1. Overall accuracy:

	DecisionTree	RandomForest
balance.scale	0.74	0.76
nursery	0.99	0.98
led	0.86	0.86
synthetic.social	0.48	0.78

2. Introduction of the classification methods:

Decision tree: I used gini-index decision tree as my model. One unique scheme I used is recursive binary splitting. Given a node, my code will loop through all attributes. For each attribute, my algorithm will again loop through all possible values and partition data points into two groups: that specific value vs. all other values. For example, if there are three values 1,2 and 3 for a certain attribute “x”, all data will be partitioned into three possible scenarios: x=1 vs. x=2 or 3, x=2 vs. x=1 or 3 and x=3 vs. x=1 or 2. The partition with the smallest Gini index among all attributes and all attribute values will be selected. The tree is pruned by max. depth and min. node size that are hard coded.

Random forest: I used Forest-RI as my basic model. I followed the algorithm above to build a single tree. For each node, the amount of attributes randomly selected to split the data is the square root of the total number of attributes. Number of trees is hard coded in my program. During prediction procedure, the testing data will be run through all trees and the classification with most votes will be chosen as the prediction.

3. Evaluation measures

			balance.scale	nursery	led	synthetic.social
DecisionTree	Accuracy	Training	0.91	1.00	0.86	0.93
		Testing	0.74	0.99	0.86	0.48
	F1 (In ascending order: 1,2,3,...)	Training	0.45,0.94,0.94	1.00,1.00,1.00,1.00,1.00	0.77,0.90	0.93,0.93,0.94,0.94
		Testing	0.05,0.82,0.77	0.99,0.98,0.99,1.00,0.00	0.78,0.90	0.49,0.42,0.51,0.51
RandomForest	Accuracy	Training	0.99	1.00	0.86	1.00
		Testing	0.76	0.98	0.86	0.78
	F1 (In ascending order: 1,2,3,...)	Training	0.94,0.99,0.99	0.99,0.97,1.00,1.00,0.00	0.77,0.90	1.00,1.00,1.00,1.00
		Testing	0.00,0.83,0.81	0.97,0.94,0.97,1.00,0.00	0.78,0.90	0.81,0.81,0.81,0.79

4. Parameters

All parameters were selected based on many rounds of testing. These parameters combined give most promising results.

Decision tree:

	balance.scale	nursery	led	synthetic.social
Max. depth	15	15	8	20

Min. node size	6	2	8	5

Random forest:

Num. of randomly selected attributes: sqrt(num. of attributes)

	balance.scale	nursery	led	synthetic.social
Num. of trees	100	15	30	100

5. Whether the ensemble method improves the performance

According to the results I presented above, I think ensemble method will dramatically increase the accuracy only when the number of attributes are large and accuracy for decision tree method is relatively low. For synthetic.social, ensemble method helped to increase the accuracy from 0.48 to 0.78. However, for the other three datasets, which have fewer attributes and achieve very high accuracy with decision tree alone, ensemble method doesn’t make much a difference.