**Assignment 4: Build a Classifier**

This assignment lets you practice with data classification in R. We’ll use the rpart package to build decision trees and the randomForest package as an implementation of the random forest ensemble method.

I use Fisher’s classic iris data for illustration.

> library(rpart)

> dt=rpart(Species~.,iris)

> plot(dt);text(dt);

> table(predict(dt,type="class"),iris$Species)

Try these statements in R. Explain their meanings. Explain the result of the contingency table.

> dt=rpart(Species~.,iris,control=rpart.control(cp=0.0,minsplit=0))

> plot(dt);text(dt);

> table(predict(dt,type="class"),iris$Species)

Explain the meaning of the first statement. Explain the result of the contingency table and its difference with the previous result.

> train\_index = c(sample(50,30), sample(50,30)+50, sample(50,30)+100)

> iris\_train=iris[train\_index,]

> iris\_test=iris[-train\_index,]

> dt=rpart(Species~.,iris\_train)

> plot(dt);text(dt);

> table(predict(dt,newdata=iris\_test,type="class"),iris\_test$Species)

Explain the meaning of these statements.

> library(randomForest)

> rf=randomForest(Species~., iris, ntree=1000, proximity=TRUE)

> table(predict(rf,type="class"),iris$Species)

Explain the meaning of these statements.

You are asked to build data classifiers on the following datasets using both decision tree and the ensemble method, and evaluate the quality of the classifiers.

1. balance.scale.data

URL: <http://archive.ics.uci.edu/ml/datasets/Balance+Scale>

Number of attributes: 4

Number of records 625

Suggested split for tree classifier: 400 for training and the rest for testing

1. Nursery

URL: <http://archive.ics.uci.edu/ml/datasets/Nursery>

Number of attributes: 8

Number of records: 12960

Suggested split for tree classifier: 8000 for training and the rest for testing

1. LED

URL: <http://archive.ics.uci.edu/ml/datasets/LED+Display+Domain>

Number of attributes: 7

Number of records: 10000 (generated with 10% noise level)

Suggested split for tree classifier: 8000 for training and the rest for testing

These datasets are downloaded/generated for your convenience. But you need to visit the URLs and read carefully to get the meanings of attributes and labels.

Split each dataset into a training data set and a testing data set. Use the training set to build a classifier and test its performance on the test set. When build your classifiers, be careful about which is an attribute and which is the class label.

Include your R scripts and contingency table of each classifier you build in your report. For overall performance of each classifier, furthermore, report accuracy (sum on the diagonal of the contingency table / total number of test records). For each class, report accuracy, specificity, precision and recall. For each data set, you should give and explain your conclusion on whether the ensemble method improves the performance of the decision tree method.

Submit your report to e-learning by the due date.