Introduction to data mining – Spring 2010

Assignment 1 (100 points)

Due: 11:55pm, Feb 18, 10 (submit to TRACS)

Overview: In this assignment, you are going to implement two decision tree induction methods, and use stratified 10-fold cross-validation on UCI datasets to verify some assumptions.

Decision trees:

1. C4.5-simplified: Similar to C4.5, use the Gain Ratio criterion for attribute selection. However, do not need to handle missing values and continuous-valued attributes. Do not need to do pruning either.

2. Random: Similar to C4.5-simplified, but select attributes randomly.

   For both, you do not need output the actual inducted trees, but only the heights of the trees. Also, calssification accuracies need to be reported.

Stratified 10-fold cross-validation:

If you feel the stratified version is hard to implement, you can also use 10-fold cross-validation. This part of implement can be integrated together with the implementation of decision trees, or separated. If separated, for each data set, you may need to run the decision tree 10 times and manually calculate the average accuracy over the 10 testing sets. If together, this process can be automated and you only need to output the averaged accuracy.

Data sets:

From the UCI repository, http://archive.ics.uci.edu/ml/, choose all the usable datasets. They are the ones that for the task of classifcation, with categorical attributes only, and no missing values.

To simplify your imlementation, you can manually preprocess the datasets, e.g., change the class labels to integers of 1, 2, 3 … You may also want to change the format of the datasets so that your program can read them easily. Preprocessing is not necessary, you choose to do so only if you feel it reduces your workload.

Assumptions to verify:

1. C4.5-simplified generates shorter trees than Random.
2. C4.5-simplified trees are more accurate than Random trees.

   If it is the case that C4.5-simplified actually generates shorter trees, then this is indeed to verify the validity of Occam's Razor applied to decision tree induction, i.e., shorter trees are more accurate.

C4.5-simplified is a deterministic method, so only one tree will be generated for each training set. On the contrary, the random method can generate many different trees. To verify the assumptions, you need to use the random method to generate multiple trees, say, 3 (to make it simple), for the same training set. Then, you take the average for height and accuracy when compared to the C4.5-simplified tree.

Report:

Use a table to present your experiment results. The table should include the following information: name of the dataset, number of instances, number of attributes, number of classes, C4.5-simplified height, C4.5-simplified accuracy, random height, random accuracy.

Then, for both C4.5-simplified and Random, report the average height and average accuracy, based on which, draw your conclusion regarding the assumptions.

Submission:

Submit a single zip file to TRACS including your code and report. In the zip file, please also include 2 (preprocessed) datasets that you used in your experiments, as well as a short description on how to run your program on these datasets.