CSE 5800 Mining/Learning and the Internet HW1
Due Sep 9, Wed, 6:30pm
Submit Server: Course=ml-internet , Assignment=hw1

Implement and evaluate the RIPPER algorithm:

1. FOIL gain: `http://jmvidal.cse.sc.edu/talks/learningrules/foilgain.xml`
2. Allow continuous-valued attributes
3. Allow more than two classes
4. Allow the option of no pruning (default is with pruning)
5. Allow "optimizations" and a parameter k for the number of "optimizations"
6. Three data sets:
   (a) Restaurant in the handout and on the course web site
   (b) Intrusion detection on the course web site
   (c) your own data set with more than two classes [or from Resources on the course web site]
7. Separate the data set into a training set and a test set, report the accuracy on the two disjoint sets (with and without pruning).
8. A report (in pdf) that discusses the following, for the second data set:
   (a) corrupt the class labels of randomly selected training examples from 0% to 20% (2% increment), by changing from the correct class to another class.
   (b) calculate accuracy on the (corrupted) training and (non-corrupted) test sets
   (c) plot accuracy vs. noise percentage in the training and test sets.
   (d) compare the training and test accuracy of the rules with and without pruning
   (e) vary k = 0, 1, and 2
9. Implementation:
   (a) preferrably use one of these programming languages: C, C++, Java, Python, or LISP.
   (b) input files: attributes description, training data, test data
   (c) Suggestion: you might have two (or three) executables:
      i. Miner/Learner: input training examples/instances, output ruleset
      ii. Classifier/predictor: input ruleset and labeled instances, output the classifications/predictions and how accurate the tree is with respect to the correct labels (% of correct classifications).
      iii. ruleset printer: if the output from the learner is human-readable, no need for a ruleset printer; otherwise, build a ruleset printer so that we can see the learned ruleset.
10. Submission:
    (a) source code
    (b) your data set
    (c) report in pdf
    (d) README.txt (how to compile and run your program/experiments on code.fit.edu or hopper.cs.fit.edu)