CSI5387: Data Mining and Concept Learning, Winter 2014

Assignment 2

Due Date: Monday March 3, 2014

Question 1:

The Thoracic Surgery Data Data Set is a fairly new domain in the UCI Repository (http://archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data). It is imbalanced, so use SMOTE (http://www.jair.org/media/953/live-953-2037-jair.pdf) to deal with this issue. (You can run the classifiers without SMOTE, first, to see if SMOTE really makes a difference. Please comment on your observations regarding the class imbalance problem on this data set in your assignment).

I would like you to compare and contrast Naïve Bayes and k-Nearest neighbours on this domain, using the AUC and WEKA's default 10x10 fold cross-validation error estimation regimen.

Since you are working with a single domain, two statistical tests are appropriate to consider: the t-test and McNemar's test. Please use both. When using the t-test, if you found statistical significance in the difference in performance between the two classifiers, then verify that this difference is practically significant by computing the effect size, using Cohen's d statistics.

Question 2:

The Contact Lenses data set in the UCI Repository is quite small with 24 instance (http://archive.ics.uci.edu/ml/datasets/Lenses). It is not clear that 10-fold CV is the best error estimation approach in such cases. Use a single classifier on this data (e.g., a Decision Tree) and experiment with the different error estimation methods that we discussed: k-fold CV, n x k-fold CV, leave-one-out, bootstrapping, permutation test. Argue for or against each of these testing regimens in the case of this domain.

Question 3

The purpose of this question is to determine which of the classifiers we encountered so far is/are generally reliable. The classifiers we are dealing with are: Decision Trees, Multi-Layer Perceptrons (MLP), Naive Bayes, k-Nearest Neighbours, and Support Vector Machines. Select a number of domains (10 or 15) from the UCI Repository that have different characteristics from one another [we are looking at a *generally* reliable classifier, i.e., one to use when we don't know anything about the characteristics of the data]. Explain in one way the domains you chose differ from one another. Conduct your experiments using 10x10-fold CV or 10-fold CV (you can choose) and conduct Friedman's Test followed by Nemenyi's Test to verify whether or not your results are statistically significant.