**CSI5387: Data Mining and Concept Learning, Winter 2012**

**Assignment 2**

**Due Date: Tuesday March 6, 2012**

Here is a list of new themes that will be explored by this assignment:

* **Learning Paradigms:** Naive Bayes, k-Nearest Neighbours, Support Vector Machines, Ripper
* **Techniques:** Feature Selection approaches, SMOTE (technique for dealing with Class-Imbalances)
* **Evaluation:**
* **Evaluation metrics/method:** ROC-Analysis, Cost-Curves, G-Mean
* **Statistical Testing:** Wilcoxon signed rank test, ANOVA, Friedman’s test

1. Select 3 domains of interest to you from the [UCI Repository for Machine Learning](http://archive.ics.uci.edu/ml/). Experiment with various feature-selection approaches provided in [WEKA](http://www.cs.waikato.ac.nz/ml/weka/) on these domains with k-NN as the base classifier. Select the feature-selection approach that seems most appropriate to you, overall, from these experiments. Select another 8 domains from the UCI Repository and run a) k-NN without feature selection and b) k-NN with the feature selection approach that you identified as most appropriate in your previous experiment, on these 8 domains. Is there a difference in the performance of the two learning schemes? Verify the significance of your results using Wilcoxon’s Signed Rank Test. (Note: you can use WEKA’s default 10x10-fold CV for error estimation)
2. Repeat 1) using SVMs as the base classifier instead of k-NN.
3. Select a domain of interest to you from the UCI Repository that happens to have a large class imbalance (note: you can choose a multi-class domain and select one class as the positive class of interest and treating all the other classes together as a single negative class). Run Naive Bayes on that data set. Compare the Accuracy and AUC results you obtain. Is there any evidence that Naive Bayes suffers from the class imbalance problem on this data set? Explain your answer. If you didn’t find that the class imbalance problem was an issue for Naive Bayes on this domain, look for another domain where it is. On that domain run both Naive Bayes and [SMOTE](http://nd.edu/~dial/software/SMOTE.java) followed by Naive Bayes. Draw the two ROC Curves obtained by these two classifiers on the same ROC Graph and discuss the results you obtain. Repeat this experiment using Cost-Curves rather than ROC Curves.
4. Run J48, Naive Bayes, k-NN and JRip on the 8 domains used in Questions 1 and 2. Do you see any evidence that these classifiers are not equivalent on these domains? Run ANOVA followed by Tukey’s post-hoc test to see if these observations are statistically significant. Repeat the statistical significance testing portion of your experiments using Friedman’s test followed by Nemenyi’s post-hoc test. Do you obtain the same results in both cases? Explain the meaning of these results.