CS795/895: Topics in Data Mining and Security

Summer 2011

Homework #2: Due: June 1, 2011 (Wednesday)

Using the Weka software, with the training data provided in sick.arff (not the same as in HW1), determine the effect of different stratification techniques of generating training data and validation data in evaluating and comparing the classifiers. Use the following 4 methods. For each method, summarize the results (as given by Weka).

1. NaiveBayes
2. Multilayerperceptron
3. ConjunctiveRule
4. DecisionTable

Evaluate the credibility of the classifiers using the following techniques:

(i) Stratified hold out (2/3 and 1/3 rule). Repeat this random choice 3 times and find the average estimation error in the validation data.

(ii) Use 5-fold cross-validation method. Repeat it 5 times.

(iii) Assuming that the cost of false positives is 10.0 and false negatives is 2.0, determine a way to generate training data and validation data. Evaluate the 4 techniques based on the resulting sets.

Finally, provide a conclusive summary as to which one (in your opinion) is the most accurate.