COMP30120 Assignment 3: Ensembles in Weka

Deadline: Monday November 30th 2015

Submission: Submit your report as a single PDF file via the COMP30120 CS Moodle page. Include your full name and student ID number in the report.

Overview:

The objective of this assignment is to use the ensemble learning functionality in Weka to identify the extent to which classification performance can be improved through the combination of multiple classifiers.

For your assignment, you will need to make use of the dataset described below. When downloading the dataset, please ensure your student number is correct. Submissions using an incorrect dataset will receive a 0 grade.

Dataset:

The data describes a set of patients undergoing voice rehabilitation treatment. The problem is a binary classification task, predicting whether treatment will be successful (class=yes) or unsuccessful (class=no), based on a large number of speech signal features.

You should download your personal dataset from the URL:

http://mlg.ucd.ie/datasets/comp30120/voice/<STUDENT_NUMBER>.arff

For example, if your student number is 126023491, your dataset is at the URL:

http://mlg.ucd.ie/datasets/comp30120/voice/126023491.arff

Tasks:

Write a report which addresses the following 5 tasks for the voice dataset:

1. Evaluate the performance of two individual classifiers on the data: KNN (k=2) and Decision Trees (J48).
2. Apply ensembles with *bagging* using each of the classifiers from Task 1. Investigate the performance of both classifiers as the ensemble size increases, in steps from 10 members to 100 members. Also investigate how changing the number of instances in the bootstrap samples affects classification performance (i.e. the "bag size").
3. Apply ensembles with *boosting* using each of the classifiers from Task 1. Investigate the performance of both classifiers as the ensemble size increases, in steps from 10 members to 100 members.
4. Apply ensembles with *random subspacing* using each of the classifiers from Task 1. Investigate the performance of both classifiers as the ensemble size increases, in steps from 10 members to 100 members. Also investigate how changing the number of features used when applying random subspacing affects classification performance (i.e. the "subspace size").
5. Summarise the differences in the classification performance results from Tasks 1-4, discussing the factors which might explain these differences.