COMP30120 Tutorial

Ensembles

Q1

(a) *Bagging* (bootstrap aggregation) has a mechanism for achieving diversity for ensemble classifiers. Explain how it works.

(b) In Weka, load the *Wine* data set using the ARFF file provided, and evaluate a decision tree classifier (J48) using 10-fold cross-validation. What percentage of instances are correctly classified?

(c) Now, apply ensemble classification using bagging to achieve diversity and with a decision tree classifier. What percentage of instances are now correctly classified with an ensemble of size 10?

(Note: Bagging is available in Weka by clicking the "Classifier" button and choosing *classifiers* → *meta* → *Bagging*).

(d) Repeat (c), but increase the ensemble size to 100, 200, then 300 classifiers. What level of improvement does this provide, in terms of percentage of instances correctly classified?

(e) Why does the level of improvement in accuracy often "level off" after an ensemble has been increased to a certain size?

Q2

(a) Explain what differentiates the ensemble members in a *boosting* classifier ensemble.

(b) In Weka, load the *Glass* data set using the ARFF file provided, and evaluate a decision tree classifier (J48) using 10-fold cross-validation. What percentage of instances are correctly classified?

(c) Apply *bagging* with a decision tree classifier for an ensemble size of 100. What is the improvement over a single tree?

(d) Now apply *boosting* with a decision tree classifier for an ensemble size of 100. How does it compare to the results from (c)? How do you explain this difference?

(Note: Boosting is available in Weka by clicking the “Classifier” button and choosing *classifiers* → *meta* → *AdaBoostM1*).

Q3

(a) Applying bagging to a “stable” classifier is generally not a good idea. Why is this?

(b) Explain how diversity is generated using a *random subspacing* classifier ensemble.

(c) In Weka, load the *Glass* data set. Evaluate a k-NN classifier with *k=5* neighbours using 10-fold cross-validation. What percentage of instances are correctly classified?

(d) Apply *bagging* with a k-NN classifier (*k=5*) for an ensemble size of 100. What is the improvement in terms of percentage of instances are correctly classified?

(e) Now apply *random subspacing* with a k-NN classifier (*k=5*) for an ensemble size of 100. How does it compare to the results from (d)? How do you explain this difference?

(Note: Random subspacing is available in Weka by clicking “Classifier” and choosing *classifiers* → *meta* → *RandomSubSpace*).