



De La Salle University • College of Computer Studies

Introduction to Artificial Intelligence
INTROAI

Assignment # 2 (Group):
Comparing Decision Tree Learning (C4.5/ID3) and Multilayer Network (Backprop)
SY 2009-10 Term 1

Instructions

Your overall task is to compare the learning curves1 of a decision tree learner and a multilayer neural network learner on a published data set.

There is no need to implement the learning algorithms; you can use Weka, a suite of machine learning algorithms implemented in Java, available at http://www.cs.waikato.ac.nz/ml/weka/. Extensive documentation is also available from the said site. For this assignment, use Weka's J48 (which implements C4.5, a more robust version of ID3) and MultilayerPerceptrons (which implements Backprop for multilayer neural networks.)

Get the data set from the UC Irvine Machine Learning Repository at http://archive.ics.uci.edu/ml/datasets.html. Many datasets in the UCI repository are in C4.5 data format, a brief description of which is available at http://www.cs.washington.edu/dm/vfml/appendixes/c45.htm. Weka accepts C4.5 format as well as ARFF, Weka's own format, explained in detail in http://www.cs.waikato.ac.nz/~ml/weka/arff.html.

No 2 groups may work on the same dataset, so obtain dataset approval though raymund.sison@delasalle.ph.

Grading and Deadline

This assignment, due on August 26, 2009, is worth 10 points, broken down as follows.

Table with 2 columns: Description of the experiments, Decision tree and neural network model, Learning curves for ID3 and MLN, Analysis of the learning curves. Corresponding values: 1, 2, 2, 5.

1 Assuming you are using cross-validation with N=10 folds, for each of the 10 folds you normally use 9/10 of the data for training (TRAINi, i=1..10), and 1/10 for testing (TESTi, i=1..10). To plot a learning curve, you also consider subsets of the training data. That is, for each fold you repeat the experiment by using 10%, 20%,..., 100% of TRAINi for training, while still using the entire test set (TESTi) for testing. The x-axis of the learning curve will be the number of training instances, while the y-axis will be the percentage of test items that are correct. Below are sample learning curves from (Russell & Norvig, 2003, p. 747).

