COC131 Data Mining - Feature Selection

Martin D. Sykora m.d.sykora@lboro.ac.uk
Tutorial 04, Friday 13th March 2009

1. Fire up Weka (Waikako Environment for Knowledge Analysis) software, launch the explorer window and select the "Preprocess" tab. Open the iris data-set ("iris.arff", this should be in the ./data/ directory of the Weka install).

2. In week 2 (Tutorial 2) we briefly looked at the "AttributeSelection" filter, before we carry on with this tutorial, it is worth recapping what we did. The purpose of feature selection is to select a subset of most relevant features for building robust classifiers. This is usually approached by keeping features that discriminate best between classes in the dataset, and at the same time removing features that are redundant, Minimum-Redundancy-Maximum-Relevance.

   (a) At the bottom right of the "Preprocess" window there is a graph that visualises the dataset. Make sure "Class: class (Nom)" is selected in the drop-down box and click "Visualize All". What can you interpret from these graphs, which attribute(s) discriminate best between the classes in the data-set and why? Make a mental note of the attributes you have picked.

   (b) Now lets have a look at how the attributes we picked in the previous point compare to the attributes picked by Weka's "AttributeSelection" filter. Under "Filter" *(filters/supervised/attribute/...)* choose the "AttributeSelection" filter. And hit "Apply", do not change any of the filter parameters. What are the attributes it has thrown out, are the attributes it selects the same as the ones you chose as being discriminatory?

   (c) Select the "Visualize" tab. This shows you 2D scatter plots of each attribute against each other attribute (you could also generate a correlation matrix in excel as we did in tutorial 1). Make sure the drop-down box at the bottom says "Color: class (Nom)". Generally speaking we can identify redundant attributes, when some

features are highly correlated. In our case, which attribute(s) do you think can be removed without harming the classification?

3. Select the "Classify" tab to get into the Classification perspective of Weka. Click on "Choose" and pick the IBk (kNN) Lazy Learner Classifier, set its "K" *("kNN")* parameter to 3, in "Test options" select 10-fold cross validation and hit Start. Make a note of the Classification accuracy *("Correctky Classified Instances")*. Now go back to the "Preprocess" tab. In the "Attributes" section, select *petallength, petalwidth* and click "Remove". Now go back to the "Classify" tab and run the IKb classifier. Now repeat the process by removing *sepallength, sepalwidth* and run the classifier (you can click "Undo" in the "Preprocess" tab to undo any attribute removals). Finally remove petallength, sepallength and sepalwidth and run the classifier again. How do the prediction accuracies compare? Think about what happens when you use all the features, when you only use the two 'worst' features, the two best ones, and when you only use petalwidth?

4. Weka provides a separate tab for performing feature reduction, access it by selecting "Select attributes" tab. Feature selection algorithms typically fall into two categories, feature ranking and subset selection. Feature ranking ranks all the features according to some metric and eliminates all features that do not achieve a specific treshold score. Subset Selection searches the set of all possible features in order to find the best features. In Weka, we have algorithms for both types of feature selection. Let us have a look at how we can perform subset selection based attribute reduction using a decision tree.

   (a) We first need to select the measure based on which we will judge the 'usefullness' of features and subsets of features.

   (b) In the "AttributeEvaluator" section choose the "ClassifierSubsetEval" then click on the text box next to the "Choose" button - a parameter dialog box will appear, select the J48 decision tree classifier.

   (c) As the search method, use the default "BestFirst" search and run the attribute search.

   (d) What attributes did the search return? Were these in accordance with your expectation? Try a number of other search methods (e.g. GeneticSearch, RandomSearch), you can get more information in the parameter window by clicking on "More".

5. In your lectures this thursday you have covered evolutionary techniques such as GA and ES. Go to Ingo Rechenbergs' website (a pioneer in ES algorithms) at http://www.bionik.tu-berlin.de/institut/xstart.htm, click on ES "Animations". Here you can play with a number of example problem scenarious to which ES systems are applied to work out hypothetical optimal solutions. In particular look at "Evolution of an optical lens" and "Evolution of a cantilever girder". Press "init" to initialise the problem and then press "start" and or "stop" to kick of, halt/pause the execution of the animations respectively.