COC131 Data Mining - Data Preprocessing in Weka

Martin D. Sykora m.d.sykora@lboro.ac.uk
Tutorial 02, Friday 22nd February 2008

1. Fire up the Weka (Waikato Environment for Knowledge Analysis) software, launch the explorer window and select the "Preprocess" tab.

2. Open the iris data-set ("iris.arff", this should be in the ./data/ directory of the Weka install). What information do you have about the data set (e.g. number of instances, attributes and classes)? What type of attributes does this data-set contain (nominal or numeric)? What are the classes in this data-set? Which attribute has the greatest standard deviation? What does this tell you about that attribute? (You might also find it useful to open "iris.arff" in a text editor).

3. Under "Filter" choose the "Standardize" filter and apply it to all attributes. What does it do? How does it afect the attributes' statistics? Click "Undo" to un-standardize the data and now apply the "Normalize" filter and apply it to all the attributes. What does it do? How does it affect the attributes' statistics? How does it differ from "Standardize"? Click "Undo" again to return the data to its original state.

4. At the bottom right of the window there should be a graph which visualizes the data-set, making sure "Class: class (Nom)" is selected in the drop-down box click "Visualize All". What can you interpret from these graphs? Which attribute(s) discriminate best between the classes in the data-set? How do the "Standardize" and "Normalize" filters affect these graphs?

5. Under "Filter" choose the "AttributeSelection" filter. What does it do? Are the attributes it selects the same as the ones you chose as discriminatory above? How does its behavior change as you alter its parameters?

6. Select the "Visualize" tab. This shows you 2D scatter plots of each attribute against each other attribute (similar to the F1 vs F2 plots from tutorial 1). Make sure the drop-down box at the bottom says "Color: class (Nom)". Pay close attention to the plots between attributes you think discriminate best between classes, and the plots between attributes selected by the "AttributeSelection" filter. Can you verify from these plots whether your thoughts and the "AttributeSelection" filter are correct? Which attributes are correlated?