# Lab 04 – Pre-processing Revisited

In Lab 2 we briefly looked at the "AttributeSelection" filter, before we carry on with this lab, 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, i.e. the principle of minimum-redundancy and maximum- relevance. In this lab we are going to look at feature selection in more details and will also spend some time looking at the effect of discretization.

1. Fire up Weka, launch the explorer window, select the "Pre-process" tab and open the iris dataset.
2. Select the "Classify" tab. Click on "Choose" and select the IBk (kNN) Lazy Learner Classifier, set its "K" *("kNN")* parameter to 3, in "Test options" select 10-fold cross validation. Create classification models using the following attribute selections:

   a) pettallength, pettalwidth, sepallength and sepalwidth
   b) pettallength, pettalwidth and sepallength
   c) pettallength, pettalwidth and sepallwidth
   d) pettallength and pettalwidth
   e) pettallength and sepallength
   f) pettallength and sepallwidth
   g) pettalwidth and sepallength
   h) pettalwidth and sepallwidth
   i) pettlalength
   j) pettalwidth
   k) sepalwidth
   l) sepallength
   m) sepallength and sepalwidth
   n) sepallength, sepalwidth and pettallength
   o) sepallength, sepalwidth and pettallwidth

   Make a note of the Classification accuracy *("Correctly Classified Instances")* for each case. How do the prediction accuracies compare?

1

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?

3. 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 threshold score. Subset Selection searches the set of all possible features and returns the set with the best features. Weka has algorithms for both types of feature selection. We are going to try them using the diabetes data set so open the data set. Weka provides a separate tab for performing feature reduction. Access it by selecting the “Select attributes” tab.

Perform subset selection-based attribute reduction using a decision tree:

(a) 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.

(b) As the search method, use the default “BestFirst” search and click ‘Start’ to run the attribute search.

What attributes did the search return? Are these in accordance with your expectation? Try a number of other search methods such as ‘ExhaustiveSearch’ and ‘GreedyStepwise’.

Perform attribute ranking method:
(a) Under Attribute Evaluator choose ‘InfoGainAttributeEval’ and under Search Method choose ‘Ranker’ and click on ‘Start’. Does the ranking agree with the subset selection methods?
(b) Try another ranking method by choosing ‘ReliefAttributeEval’ under

Attribute Evaluator and choosing ‘Ranker’ under Search Method. Does the ranking agree with the results previously found?

4. Now try combining discretization with attribute selection to create a classification model and see whether it performs better than the classification model without attribute selection and discretization.

Run the classifier Multi-layer Perceptron on the diabetes data set without pre-processing and note the results.

Here are some of the options for combining discretization with attribute selection:

(a) Pre-process filter – supervised, attribute, discretize (remember to click ‘Apply’)

(b) Pre-process filter – AttributeSelection, Evaluator – ClassifierSubsetEval, click to choose MultilayerPerceptron. Search 2 – BestFirst (remember to click ‘Apply’)
(c) Classifier – MultilayerPerceptron
(d) Test options – Cross-validation 10 folds (remember to click ‘Start’)

Now compare the results and what is your conclusion?