Assignment 1. Introduction to WEKA and YALE.

Note: WEKA and YALE are downloaded to C:\MyTemp. Install them to the same location.

Understanding basic input data formats and basic operations with WEKA and YALE:

data loading, filtering (try to reduce the number of attributes, and the number of instances). Try to apply simply classification techniques: 1R, decision tree (ID3 or J48 that is C4.5), 1NN, Naïve Bayes and others.

One way to proceed:

1) From the UCI Repository (http://www.ics.uci.edu/~mlearn/MLRepository.html) download the Adult or Weather or Lenses Dataset. Format the data and test sets as required by ARFF specification. (attributes description is located in the file "filename".names)
2) In Weka Explorer evaluate the test set using One Rule and Decision Trees.
a) Show rule and at least part of tree learned. Explain their meaning.
b) Compare performance and confusion tables of both algorithms.

Analysis of rules.

- What are your top five rules?
- What is their support/confidence?
- Are they intuitive from the domain perspective or completely surprising?
- How do the association rules change if you increase/decrease the number of bins (levels) in the attributes?

Please note that rule-based algorithms can not process numerical attributes. So, you will need to use discretization filter.

Comparing different algorithms on different data sets

- Try to analyze the accuracy estimates on (1) train data, (2) cross-validation, (3) leave-on-out and (4) train/test split. Report major findings.
- Select 2 data sets: 1 – with approximately equal distribution of instances among classes, 2 – with unbalanced data (e.g. 90% of instances belong to one class and 10% to the other).

If you can not find such data set – apply filter to any data set to remove most of the instances of one class and save this data set under different name for further use.
Compare the confusion matrices with total accuracy estimates for both dataset. Report the findings.

In Weka Experimenter compare the performance of the previous algorithms using 10-fold cross-validation. (In your report make reference to the statistical significance of the results.)

Feature Selection

Use Simple Bayes, C4.5 (J48 in Weka), and IBk classifiers for

- Few UCI data sets.
- Find out which Feature Selection (FS) techniques are available in WEKA
- Compare performance of classifiers
  - without any FS
  - with FS that estimates goodness of features individually
  - with Feature Subset Selection
    - Filter paradigm
    - Wrapper paradigm
- Report the results of comparison. What are the interesting findings?

Compare WEKA and YALE

- Compare data exploration possibilities, experimentation with different DM techniques, and their parameters turning, prototyping of KDD solutions.
- Which software was easier to start with for you? Which one would you start using if you need to?
- Compare data visualization and visualization of models and other results in WEKA and YALE (we will partly address this during Visualization lecture on Friday, Nov 17)

Additional resources to help you start with WEKA and YALE

WEKA

- GUI in WEKA http://prdownloads.sourceforge.net/weka/weka.ppt
- Command line http://weka.sourceforge.net/wekadoc/index.php/en%3APrimer
- WEKA tutorial (under the same directory where WEKA is installed - Tutorial.pdf file)
- Another online tutorial http://maya.cs.depaul.edu/~classes/ect584/WEKA/index.html

YALE

- When YALE starts choose Start the YALE online tutorial from "Welcome to YALE!" window. Pass this tutorial.
- YALE Tutorials are downloadable from http://rapid-i.com/component/option,com_weblinks/catid,24/Itemid,60/
- The first chapters of the YALE tutorial together with the GUI manual should be enough for you to start playing with YALE