Weka Practical I

Note:

- Briefly, OneR (one-attribute-rule) algorithm generates rules that only include one attribute plus the class. ID3 is an algorithm that generates a decision tree, which can be converted to several rules. It only accepts nominal values. J48 (C4.5) is an improved version of ID3. It accepts both numerical and nominal values and generates a pruned decision tree. For example, see the output by J48 on the weather data on page 5 in How to Use Weka. From the J48 pruned tree, we can generate a rule like this

  IF outlook is sunny and humidity is greater than 75 **THEN** don't play.

- Before you start this exercise please read the notes: *How to Use Weka* **first**.
- Don't change any parameters until you understand them.
- Study the data before you apply any algorithm on it.
- Scheme J48 in this exercise means **weak.classifiers.j48.J48**
- In general, you are expected to find patterns (rules) from models generated by weka. These patterns are normally not easily found by just studying the datasets manually.
- For all exercises, use different test options: Use training set, Cross Validation and Percentage split. What are the differences? You are expected to check the model evaluation as well.
- Datasets for this exercise can be found either under the weka's data directory or on the module's Moodle.
- If the dataset provided in .xlsx format, you need to convert it to .arff first.

1. Use OneR, ID3 and J48 to analyse weather.arff, weather.nominal.arff and contact-lenses.arff data under the data directory on the server.

2. Load weather.nominal.arff to Weka, select attribute *outlook* as the class attribute. Then apply ID3 and OneR classifiers. See what's the difference between the output generated here and the output generated using attribute *play* as the class attribute.

3. Use OneR, J48 to analyse the zoo data (in zoo.arff):

   How do the classifiers determine whether an animal is a mammal, bird, reptile, fish, amphibian, insect, or invertebrate? Do the decisions made by the classifiers make sense to you? Why does OneR perform so badly?

4. Filter out some of the attributes in each data set and try exercise 1, 2 and 3 again.

5. Use OneR, ID3 and J48 to analyse the titanic data set. The file titanic.xls contains data about the sinking of the `Titanic' on the night of April 15th, 1912. In particular, it contains information about who survived and who did not, among the

passengers and the crew. There is one line per person and are 5 attributes involved:

1) the case number, from a board of inquiry report
2) whether the person was travelling first, second or third class or was a crew member
3) whether the person was an adult or a child
4) whether the person was male or female
5) whether they survived or not.

The purpose of this exercise is to find out which type of passengers is likely to survive from the disaster. Some background history and research can be found at http://www.titanicfacts.net/. You may like to study the site, before you start this exercise.

6. Use OneR, ID3 and J48 to analyse the mushroom data (in mushroom.arff).

After you download the file, open it with MS word and read the explanation of the data set. There are some missing values in the file. Since ID3 doesn't accept files with missing values, you need to remove the 11th attribute before applying ID3 on it. The purpose of this exercise is to determine which kind of mushrooms is safe.

References

R. Kirkby and E. Frank: WEKA Explorer User Guide for Version 3.5.8, 2008