Hand on Weka

2014/11/11

Petra Kralj Novak
Petra.Kralj.Novak@ijs.si

Data Mining Tools

- Weka http://www.cs.waikato.ac.nz/ml/weka/
- Orange http://orange.biolab.si/
- Knime http://www.knime.org/
- Taverna http://www.taverna.org.uk/
- Rapid Miner http://rapid-i.com/content/view/181/196/
- ClowdFlows http://clowdflows.org/

Weka (Waikato Environment for Knowledge Analysis)

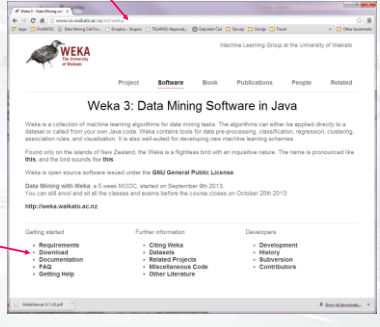
- Collection of machine learning algorithms for data mining tasks
- The algorithms
  - Can be applied directly to a dataset
  - Can be called from Java code (library)
- Weka contains tools for
  - Data pre-processing
  - Classification
  - Regression
  - Clustering
  - Association rules
  - Visualization
- Weka is open source software issued under the GNU General Public Licanse

Exsercise1: ID3 in Weka

1. Build a decision tree with the ID3 algorithm on the lenses dataset, evaluate on a separate test set

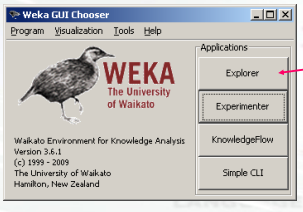
Weka: Install

http://www.cs.waikato.ac.nz/ml/weka/



Download version 3.6

Weka: Run Explorer

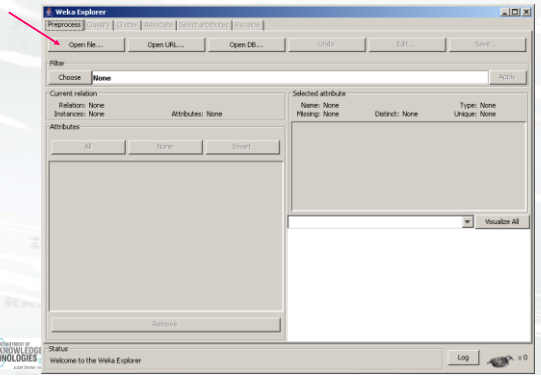


Choose Explorer

Exercise 1: ID3 in Weka

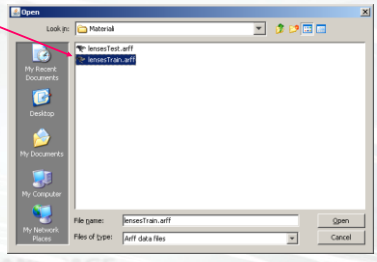
- In the Weka data mining tool, induce a decision tree for the **lenses** dataset with the ID3 algorithm.
- Data:
  - lensesTrain.arff
  - lensesTest.arff
- Compare the outcome with the manually obtained results.

Load the data



Load the data - 2

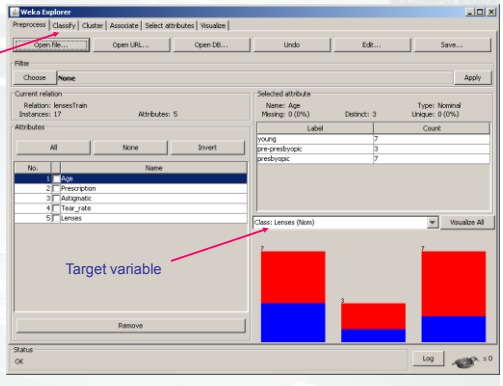
lensesTrain.arff



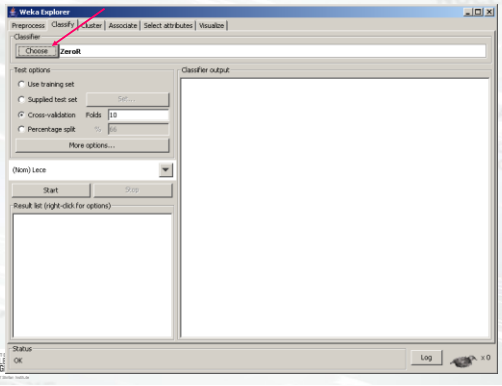
The data are loaded

Choose "Classify"

Target variable

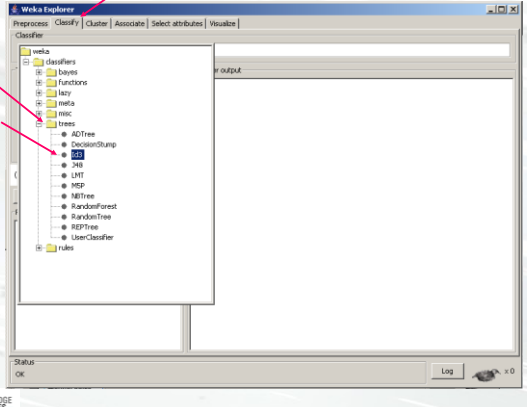


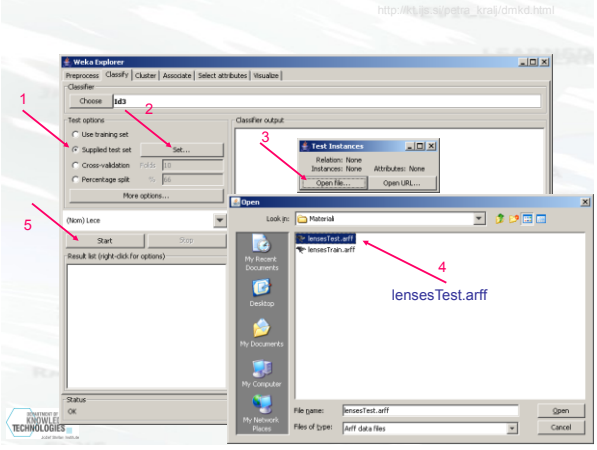
Choose algoritem

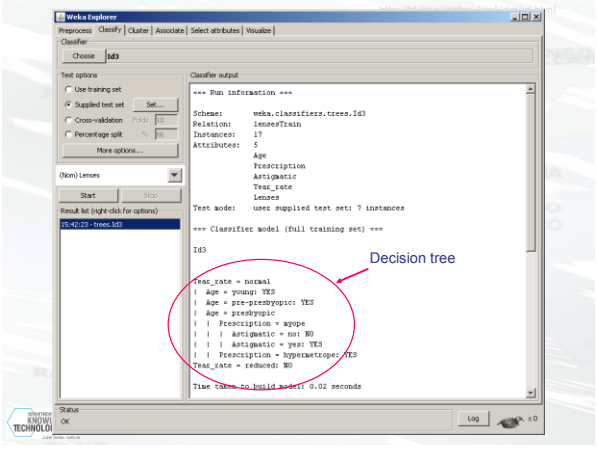


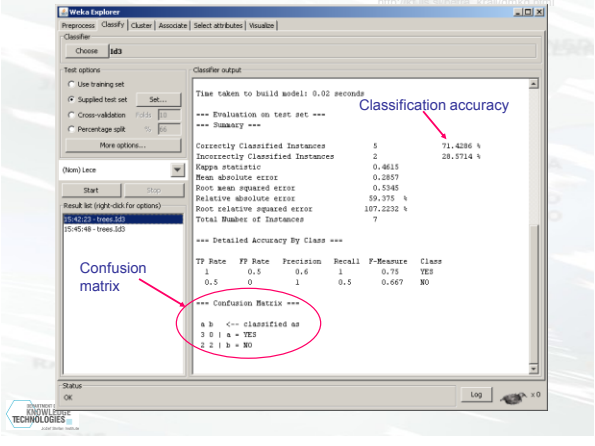
trees

Id3









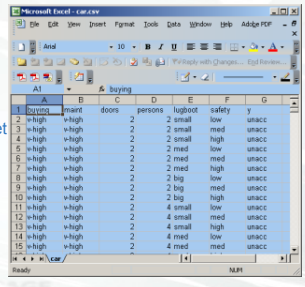
Exercise 2: CAR dataset

- 1728 examples
- 6 attributes
  - 6 nominal
  - 0 numeric
- Nominal target variable
  - 4 classes: unacc, acc, good, v-good
  - Distribution of classes
    - unacc (70%), acc (22%), good (4%), v-good (4%)
- No missing values

Preparing the data for WEKA - 1

Data in a spreadsheet (e.g. MS Excel)

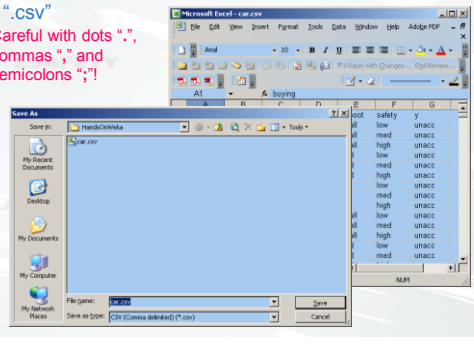
- Rows are examples
- Columns are attributes
- The last column is the target variable



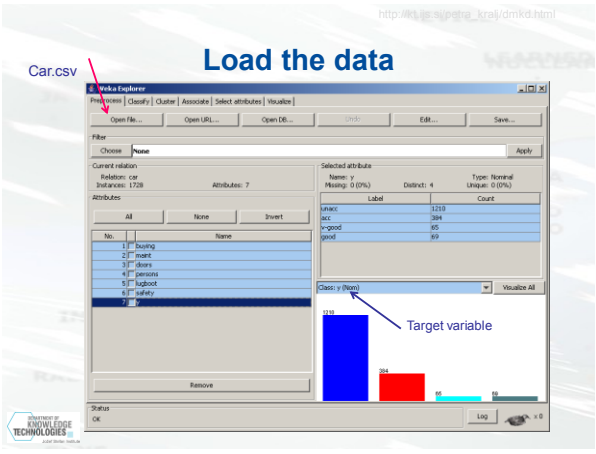
Preparing the data for WEKA - 2

Save as ".csv"

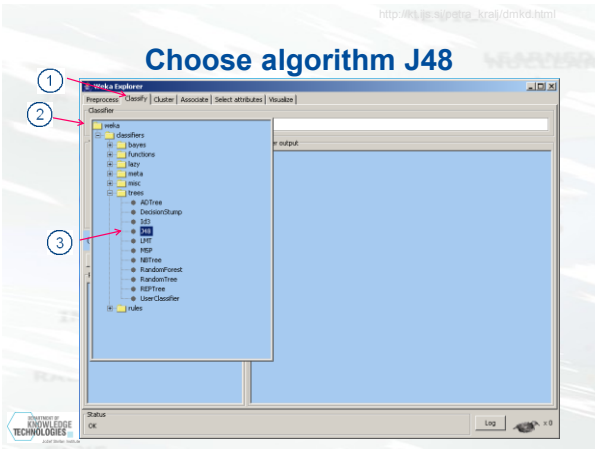
- Careful with dots ".", commas "," and semicolons ";"!



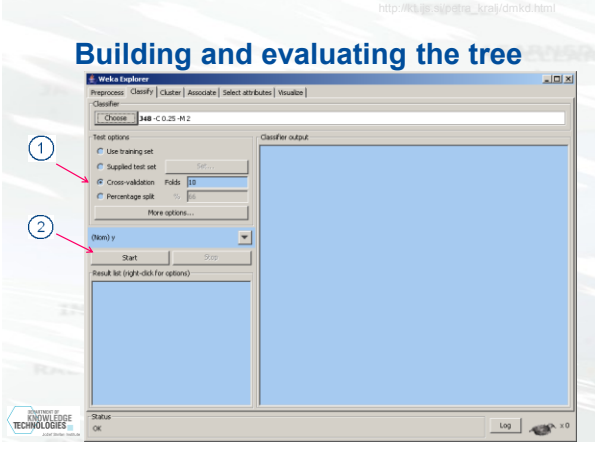
Load the data

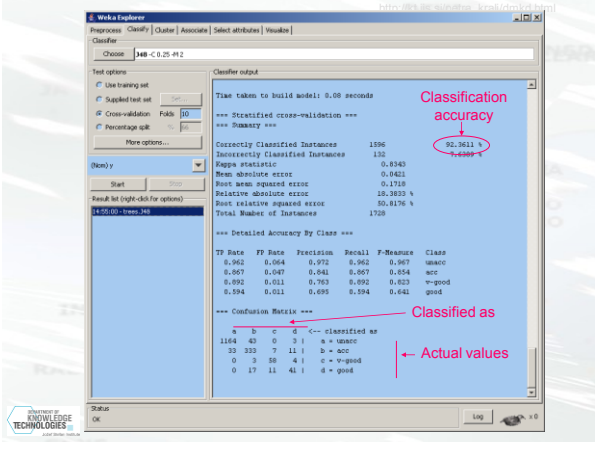


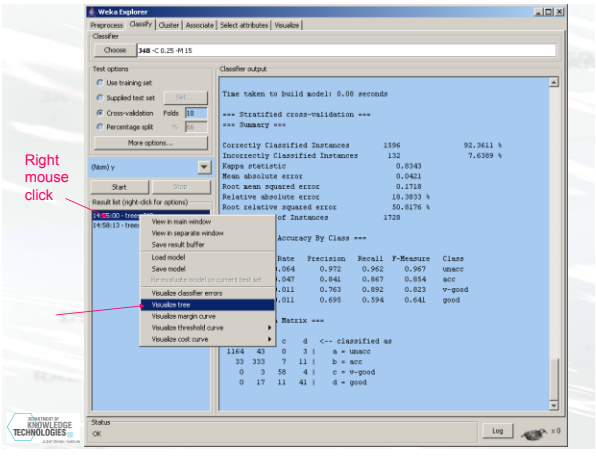
Choose algorithm J48



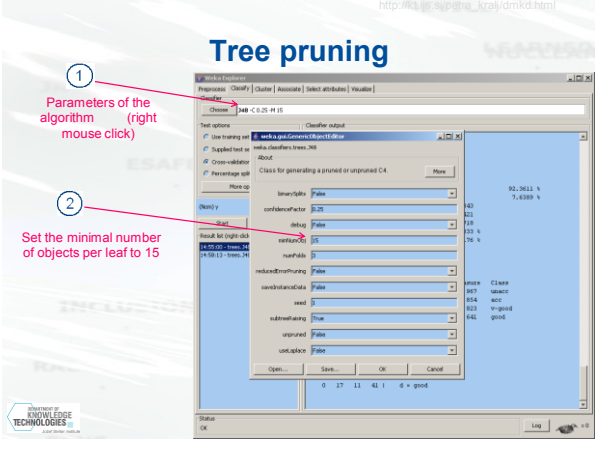
Building and evaluating the tree

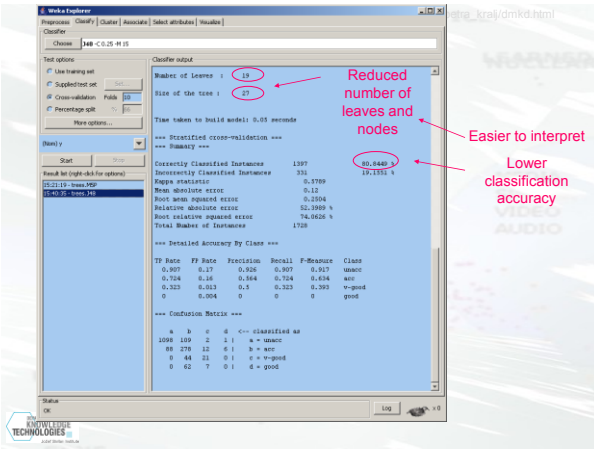


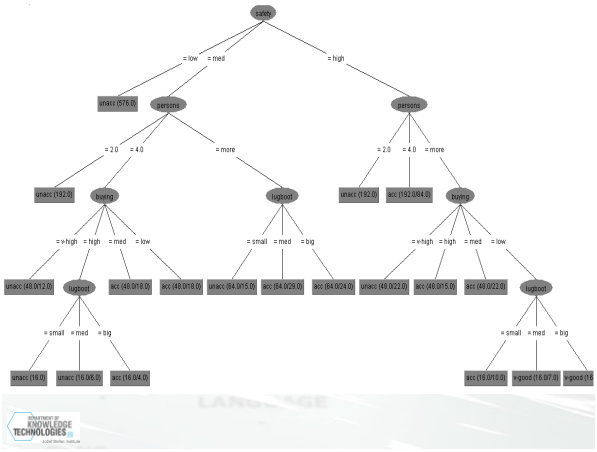




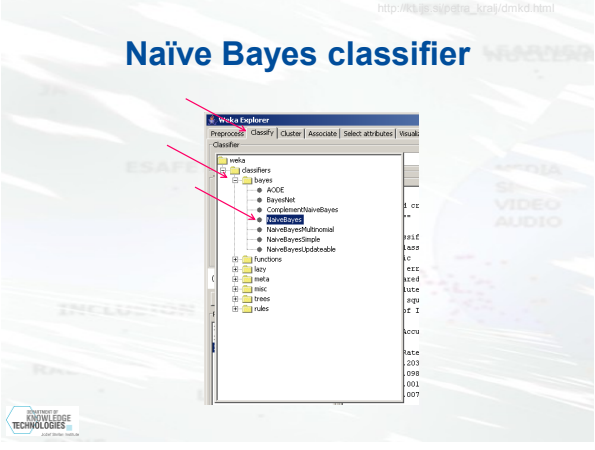
Tree pruning

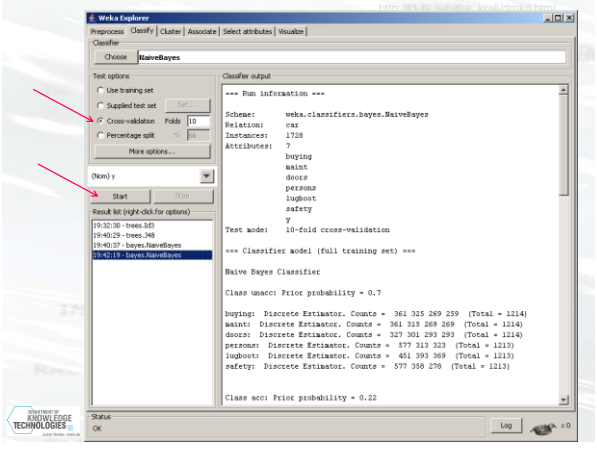


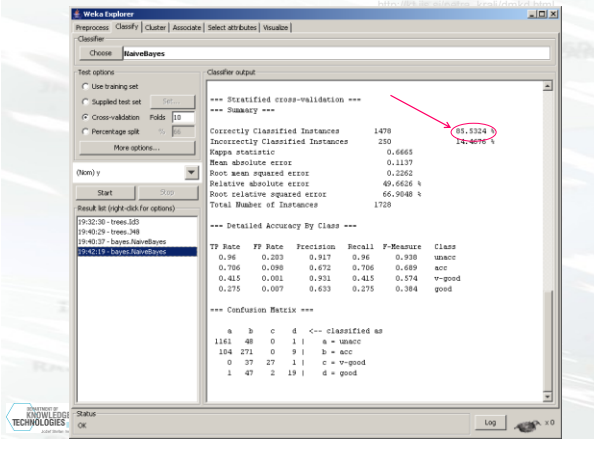




Naïve Bayes classifier







Summary

- Weka
- ID3, separate test set
- Data preparation
- J48 (C4.5), cross validation, tree prunning
- Naïve Bayes