

http://kij.ssupetra_kraj/dmki.html

Hand on Weka

2014/11/11

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



http://kij.ssupetra_kraj/dmki.html

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/full/181/196/>
- ClowdFlows <http://clowdflows.org/>



http://kij.ssupetra_kraj/dmki.html

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 License



http://kij.ssupetra_kraj/dmki.html

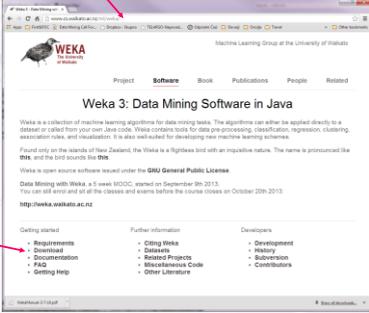
Exersercise1: ID3 in Weka

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



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

Weka: Install



Download version 3.6



http://kij.ssupetra_kraj/dmki.html

Weka: Run Explorer



Choose Explorer

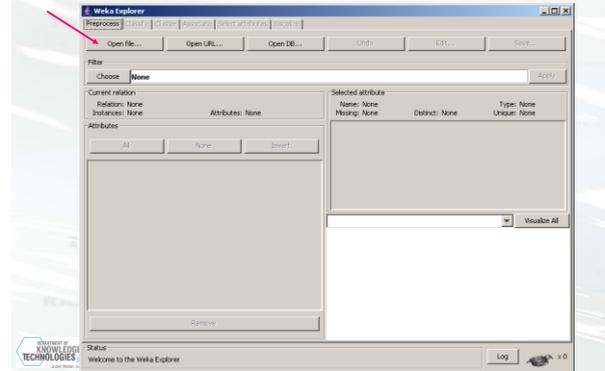


Exercise 1: ID3 in Weka

- 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.

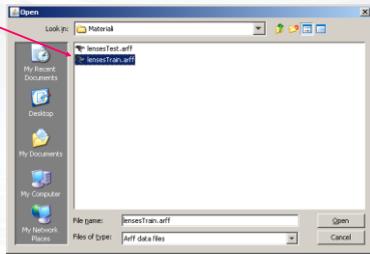


Load the data

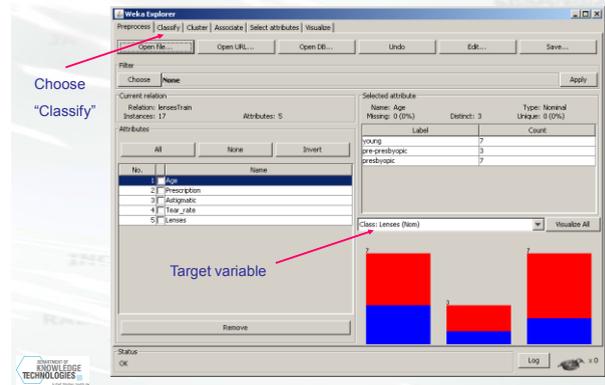


Load the data - 2

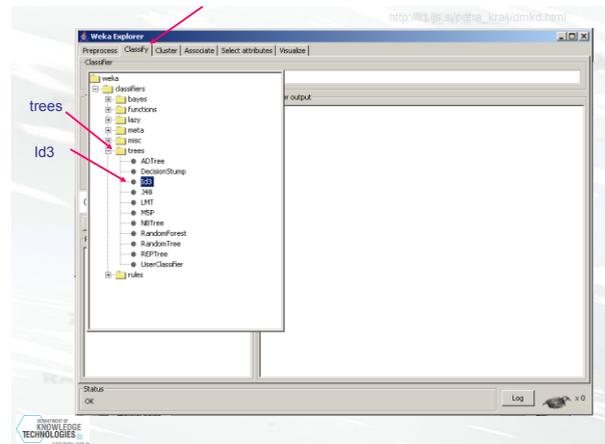
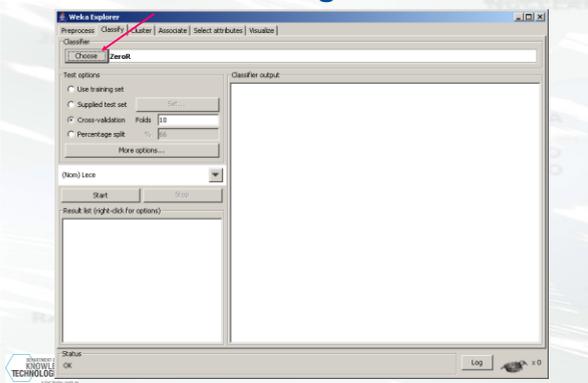
lensesTrain.arff



The data are loaded



Choose algorithm



Load the data

Car.csv

Weka Explorer interface showing the 'Load the data' step. The 'Car.csv' file is loaded into the 'Instances' list. A bar chart shows the distribution of the target variable 'y' with three classes: 'usacc' (130), 'acc' (89), and 'v-good' (18). The 'Target variable' label points to the 'y' column in the 'Selected attribute' table.

Choose algorithm J48

Weka Explorer interface showing the 'Choose algorithm J48' step. The 'J48' algorithm is selected from the 'Classifiers' tree. Red circles 1, 2, and 3 indicate the selection process.

Building and evaluating the tree

Weka Explorer interface showing the 'Building and evaluating the tree' step. The 'J48 - C.0.25 #12' classifier is selected, and the 'Start' button is clicked. Red circles 1 and 2 indicate the selection and execution steps.

Classification accuracy

Weka Explorer interface showing the 'Classification accuracy' step. The 'Classifier output' window displays performance metrics and a confusion matrix. Red circles and arrows highlight the 'Classification accuracy' (92.3611%) and the 'Confusion Matrix'.

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.962	0.064	0.972	0.962	0.967	usacc
0.887	0.047	0.941	0.887	0.954	acc
0.092	0.011	0.763	0.092	0.623	v-good
0.594	0.011	0.695	0.594	0.641	good

Tree pruning

Weka Explorer interface showing the 'Tree pruning' step. The 'Visualize tree' option is selected from the 'More options...' menu. A red arrow points to the 'Visualize tree' option.

Tree pruning

Weka Explorer interface showing the 'Tree pruning' step. The 'Parameters of the algorithm' dialog box is open, and the 'Minimal number of objects per leaf' is set to 15. Red circles 1 and 2 indicate the dialog box and the 'Minimal number of objects per leaf' setting.

Classifier output

Number of leaves : 19
 Size of the tree : 27

Reduced number of leaves and nodes
 Easier to interpret

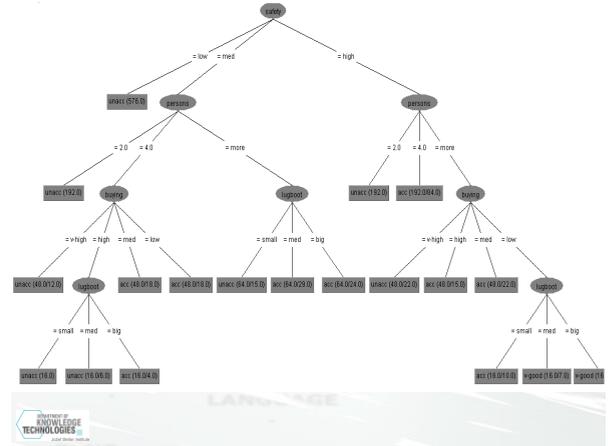
Lower classification accuracy

Summary

Correctly Classified Instances	1397
Incorrectly Classified Instances	331
Kappa statistic	0.789
Mean absolute error	0.12
Root mean squared error	0.3504
Relative absolute error	52.3999 %
Root relative squared error	74.026 %
Total Number of Instances	1728

Detailed Accuracy By Class

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.907	0.17	0.926	0.907	0.917	unacc
0.706	0.35	0.684	0.706	0.694	acc
0.323	0.533	0.5	0.323	0.393	v-good
0	0.004	0	0	0	good



Naïve Bayes classifier

Classifier output

Summary

Correctly Classified Instances	1478
Incorrectly Classified Instances	250
Kappa statistic	0.6665
Mean absolute error	0.1137
Root mean squared error	0.2262
Relative absolute error	49.6626 %
Root relative squared error	66.9048 %
Total Number of Instances	1728

Detailed Accuracy By Class

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.96	0.283	0.817	0.96	0.908	unacc
0.706	0.090	0.672	0.706	0.689	acc
0.415	0.001	0.931	0.415	0.574	v-good
0.275	0.067	0.633	0.275	0.384	good

Summary

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