

Hand on Weka: Just a Taste

2013/11/11

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 License

Practice with 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/>



Weka 3 - Data Mining with... x

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

Apps First&FOC Data Mining Call For... Dropbox - Skupno TELARGO-Napoved... Odpiralni Časi Slovarji Orodja Travel Other bookmarks

 **WEKA**
The University of Waikato

Machine Learning Group at the University of Waikato

Project Software Book Publications People Related

Weka 3: Data Mining Software in Java

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

Found only on the islands of New Zealand, the Weka is a flightless bird with an inquisitive nature. The name is pronounced like **this**, and the bird sounds like **this**.

Weka is open source software issued under the **GNU General Public License**.

Data Mining with Weka, a 5 week MOOC, started on September 9th 2013.
You can still enrol and sit all the classes and exams before the course closes on October 20th 2013:

<http://weka.waikato.ac.nz>

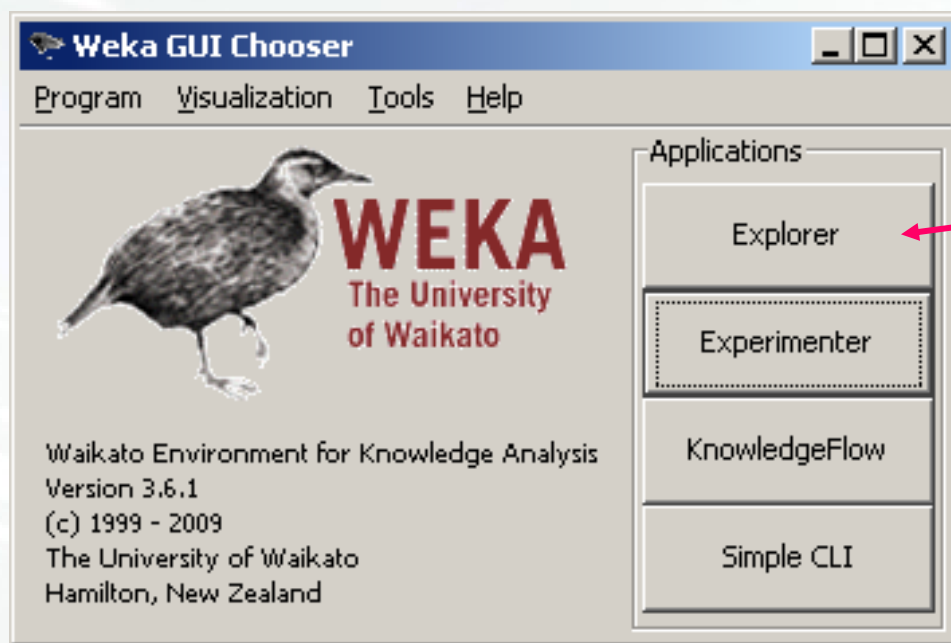
Getting started	Further information	Developers
<ul style="list-style-type: none"> • Requirements • Download • Documentation • FAQ • Getting Help 	<ul style="list-style-type: none"> • Citing Weka • Datasets • Related Projects • Miscellaneous Code • Other Literature 	<ul style="list-style-type: none"> • Development • History • Subversion • Contributors

WekaManual-3-7-10.pdf

Show all downloads...

Download
version
3.6

Weka: Run Explorer



Choose Explorer

Exercise: Lenses dataset

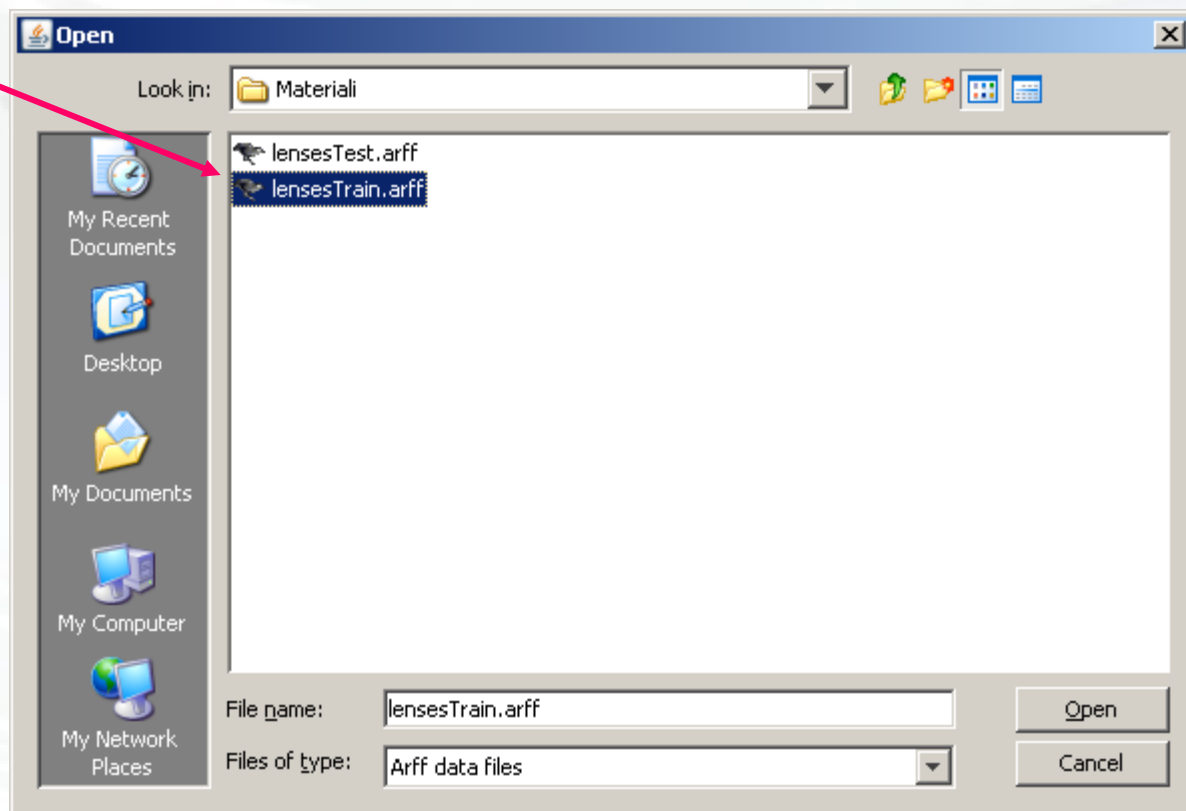
- In the Weka data mining tool, induce a decision tree for the lenses dataset with the ID3 algorithm.
- Data:
 - lensesTrain.arff
 - lensesTest.arff
- Compare the outcome with the manually obtained results.

Load the data

The screenshot shows the Weka Explorer application window. A red arrow points to the 'Open file...' button in the top toolbar. The interface includes several panels: 'Preprocess' (selected), 'Classify', 'Cluster', 'Associate', 'Select attributes', and 'Visualize'. Below the toolbar are buttons for 'Open file...', 'Open URL...', 'Open DB...', 'Undo', 'Edit...', and 'Save...'. The 'Filter' section has a 'Choose' button and a text field containing 'None', with an 'Apply' button. The 'Current relation' section shows 'Relation: None', 'Instances: None', and 'Attributes: None'. The 'Attributes' section has 'All', 'None', and 'Invert' buttons. The 'Selected attribute' section shows 'Name: None', 'Missing: None', 'Type: None', 'Distinct: None', and 'Unique: None'. A 'Visualize All' button is at the bottom right of the main area. The 'Status' bar at the bottom says 'Welcome to the Weka Explorer' and includes a 'Log' button and a small icon.

Load the data - 2

lensesTrain.arff



The data are loaded

Choose
"Classify"

The screenshot shows the Weka Explorer application window. The 'Classify' tab is selected in the top menu. A red arrow points to the 'Classify' tab. Below the menu, there are buttons for 'Open file...', 'Open URL...', 'Open DB...', 'Undo', 'Edit...', and 'Save...'. The 'Filter' section shows 'None' selected. The 'Current relation' section shows 'Relation: lensesTrain' and 'Instances: 17'. The 'Attributes' section shows a list of attributes: Age, Prescription, Astigmatic, Tear_rate, and Lenses. A red arrow points to the 'Lenses' attribute, which is labeled as the 'Target variable'. The 'Selected attribute' section shows 'Name: Age' and 'Type: Nominal'. Below this is a table showing the distribution of the 'Age' attribute:

Label	Count
young	7
pre-presbyopic	3
presbyopic	7

The 'Class: Lenses (Nom)' dropdown is also visible, with a 'Visualize All' button next to it. Below the dropdown are three stacked bar charts showing the distribution of the 'Lenses' attribute across the 'Age' categories. The first chart (young) has a total count of 7, the second (pre-presbyopic) has a total count of 3, and the third (presbyopic) has a total count of 7. Each bar is divided into red and blue segments.

Target variable

Choose algoritem

The screenshot shows the Weka Explorer application window. The title bar reads "Weka Explorer". The main menu includes "Preprocess", "Classify", "Cluster", "Associate", "Select attributes", and "Visualize". The "Classify" tab is active, and the "Classifier" section shows "ZeroR" selected in a list box. A red arrow points to the "Choose" button next to "ZeroR".


Test options

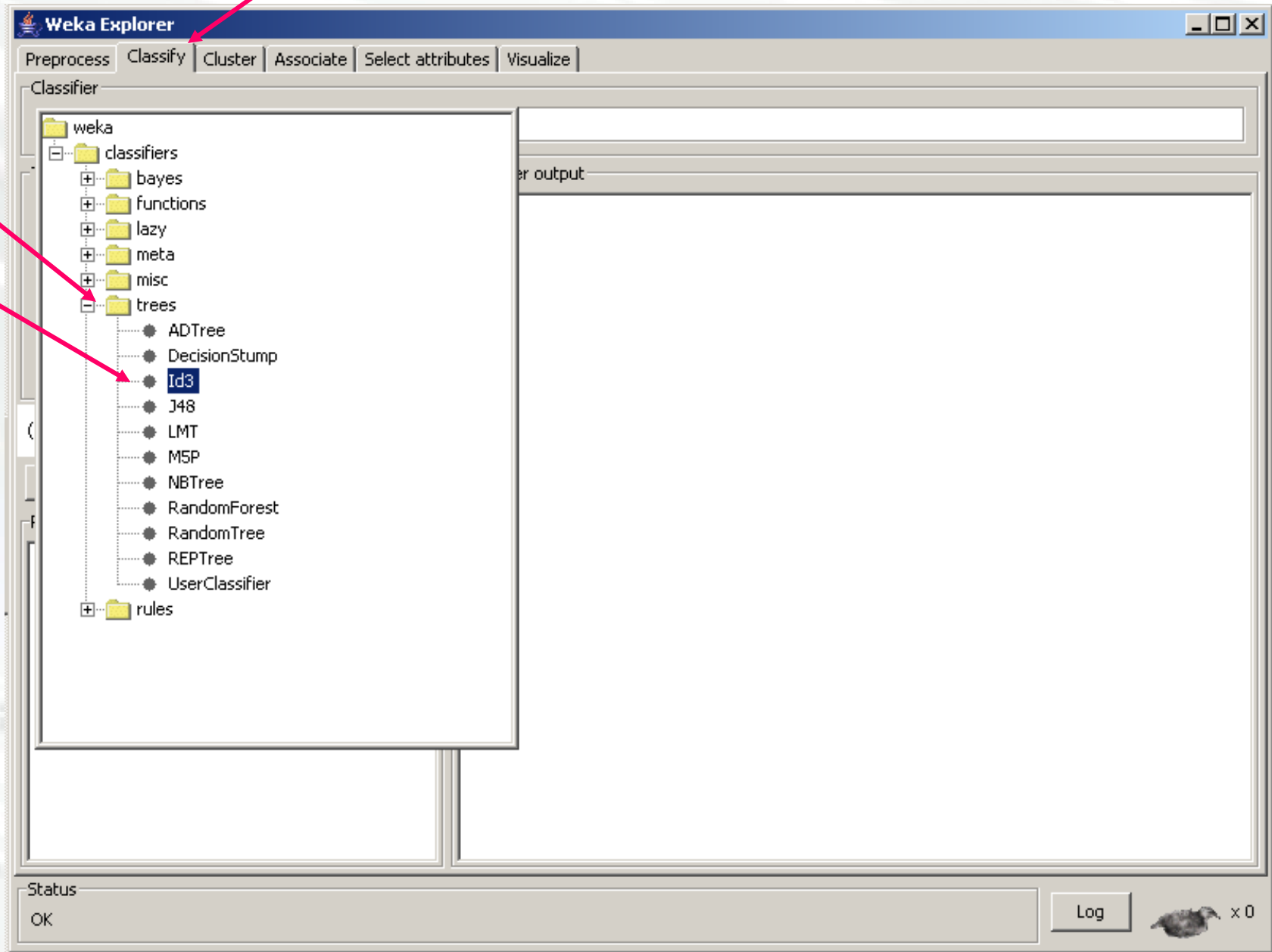
- Use training set
- Supplied test set
- Cross-validation Folds
- Percentage split %

(Nom) Lece

Result list (right-click for options)

Status: OK

 x 0



trees

Id3

The image shows the Weka Explorer software interface. At the top, there are tabs for Preprocess, Classify, Cluster, Associate, Select attributes, and Visualize. The 'Classify' tab is active, showing a 'Classifier' section with a 'Choose' button and 'Id3' selected. Below this is the 'Test options' section with radio buttons for 'Use training set', 'Supplied test set' (which is selected), 'Cross-validation', and 'Percentage split'. The 'Supplied test set' option has a 'Set...' button next to it. A 'Classifier output' window is open, showing 'Test Instances' with 'Relation: None', 'Instances: None', and 'Attributes: None', and buttons for 'Open file...' and 'Open URL...'. An 'Open' dialog box is also open, showing the 'Materiali' folder containing 'lensesTest.arff' and 'lensesTrain.arff'. The 'lensesTest.arff' file is selected and highlighted. The 'File name' field in the dialog contains 'lensesTest.arff' and the 'Files of type' dropdown is set to 'Arff data files'. Red arrows with numbers 1 through 5 point to various elements: 1 points to the 'Supplied test set' radio button, 2 points to the 'Set...' button, 3 points to the 'Open file...' button in the 'Test Instances' window, 4 points to the 'lensesTest.arff' file in the 'Open' dialog, and 5 points to the 'Start' button in the Weka Explorer interface.

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Classifier: Choose **Id3**

Test options

- Use training set
- Supplied test set
- Cross-validation Folds:
- Percentage split %:

(Nom) Lenses

Result list (right-click for options)

- 15:42:23 - trees.Id3

Classifier output

```
=== Run information ===
Scheme:      weka.classifiers.trees.Id3
Relation:    lensesTrain
Instances:   17
Attributes:  5
              Age
              Prescription
              Astigmatic
              Tear_rate
              Lenses
Test mode:   user supplied test set: 7 instances

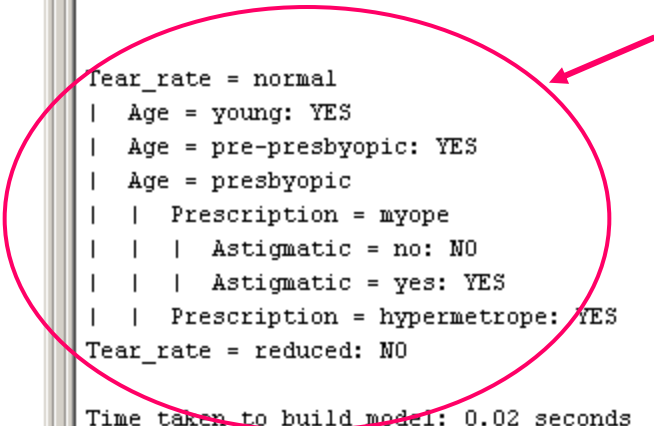
=== Classifier model (full training set) ===

Id3

Tear_rate = normal
| Age = young: YES
| Age = pre-presbyopic: YES
| Age = presbyopic
| | Prescription = myope
| | | Astigmatic = no: NO
| | | Astigmatic = yes: YES
| | Prescription = hypermetrope: YES
Tear_rate = reduced: NO

Time taken to build model: 0.02 seconds
```

Decision tree



Weka Explorer

Preprocess | **Classify** | Cluster | Associate | Select attributes | Visualize

Classifier: Choose **Id3**

Test options:
 Use training set
 Supplied test set (Set...)
 Cross-validation (Folds: 10)
 Percentage split (%: 66)
 More options...

(Nom) Lece

Start Stop

Result list (right-click for options):
 15:42:23 - trees.Id3
 15:45:48 - trees.Id3

Classifier output:

Time taken to build model: 0.02 seconds

=== Evaluation on test set ===
 === Summary ===

Correctly Classified Instances	5	71.4286 %
Incorrectly Classified Instances	2	28.5714 %
Kappa statistic	0.4615	
Mean absolute error	0.2857	
Root mean squared error	0.5345	
Relative absolute error	59.375 %	
Root relative squared error	107.2232 %	
Total Number of Instances	7	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
1	0.5	0.6	1	0.75	YES
0.5	0	1	0.5	0.667	NO

=== Confusion Matrix ===

```

a b <-- classified as
3 0 | a = YES
2 2 | b = NO
  
```

Status: OK

Log x 0

Classification accuracy



Confusion matrix

