

WEKA

A Data Mining Tool

By Susan L. Miertschin

Data Mining

Task Types

- Classification
- Clustering
- Discovering Association Rules
- Discovering Sequential Patterns – Sequence Analysis
- Regression
- Detecting Deviations from Normal

Numerous Algorithms

- C4.5 Decision Tree
- K-Means Clustering

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

WEKA can be freely downloaded by visiting the Web site

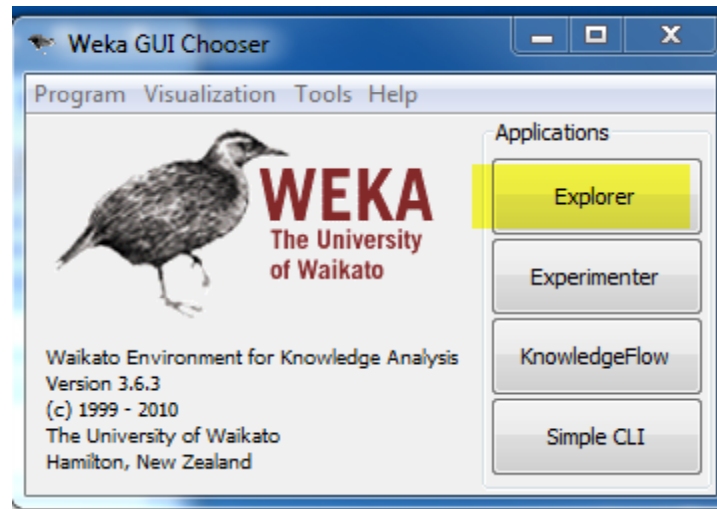
WEKA – Data Mining Software

- Developed by the Machine Learning Group, University of Waikato , New Zealand
- Vision: Build state-of-the-art software for developing machine learning (ML) techniques and apply them to real-world data-mining problems
- Developed in Java

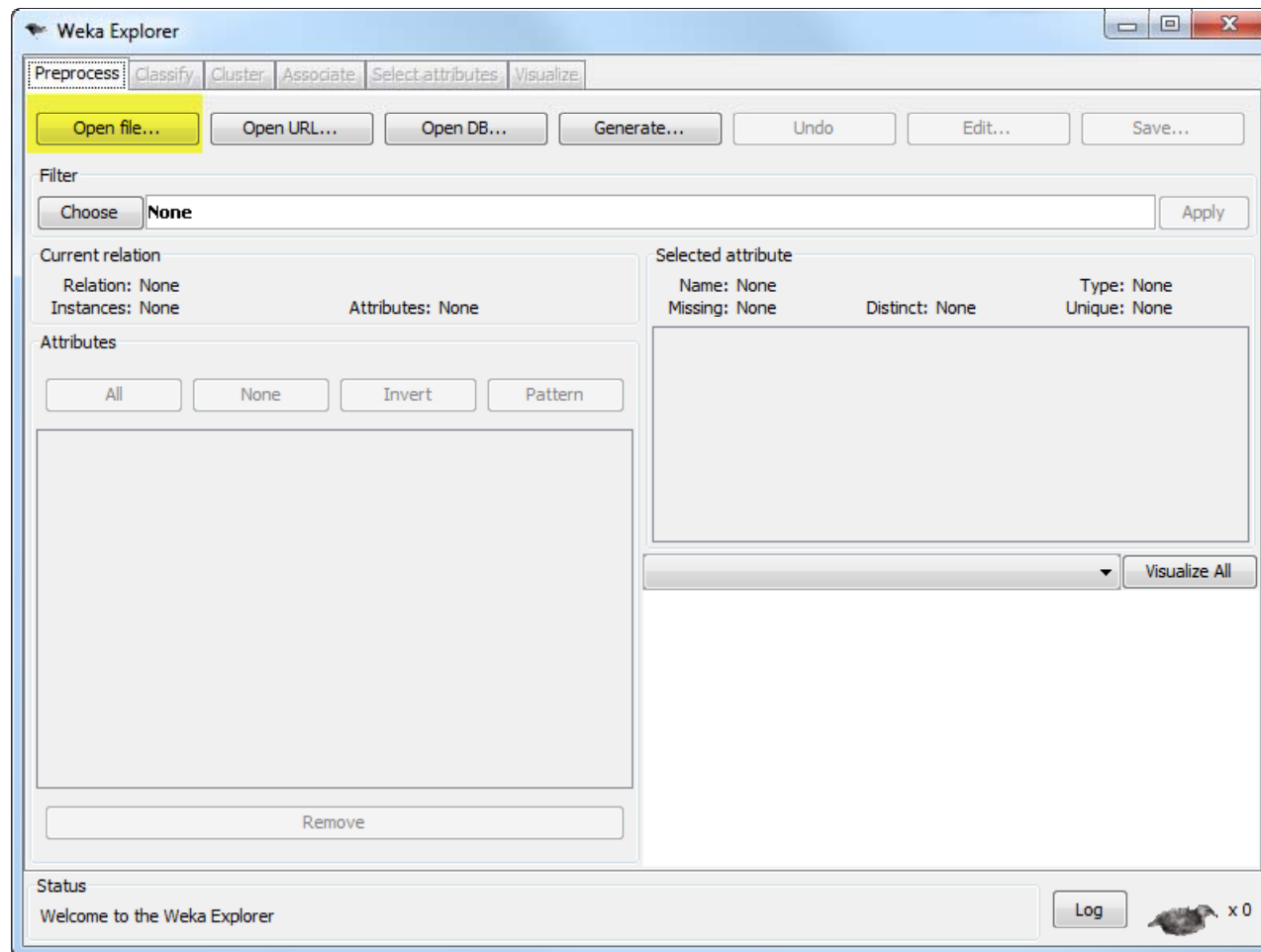
WEKA's Collection of Machine Learning Algorithms

- Algorithms for data mining tasks
- WEKA is open source software issued under the [GNU General Public License](#)
- Tools for:
 - Data pre-processing
 - Classification
 - Regression
 - Clustering
 - Association rules
 - Visualization

After Installing - Start WEKA



WEKA Main Interface



WEKA Sample Files

- C:\Program Files\weka\data
- WEKA formatted files (.arff)
- Open the contact-lenses file

Example – Contact Lens Data

How many data instances are in the file?

How many attributes?

Numerical attributes?

Categorical attributes?

Weka Explorer

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

Open file... | Open URL... | Open DB... | Generate... | Undo | Edit... | Save...

Filter: Choose None Apply

Current relation: Relation: contact-lenses, Instances: 24, Attributes: 5

Attributes: All | None | Invert | Pattern

| No. | Name |
|-----|---|
| 1 | <input checked="" type="checkbox"/> age |
| 2 | <input type="checkbox"/> spectacle-prescrip |
| 3 | <input type="checkbox"/> astigmatism |
| 4 | <input type="checkbox"/> tear-prod-rate |
| 5 | <input type="checkbox"/> contact-lenses |

Remove

Selected attribute: Name: age, Type: Nominal, Missing: 0 (0%), Distinct: 3, Unique: 0 (0%)

| No. | Label | Count |
|-----|----------------|-------|
| 1 | young | 8 |
| 2 | pre-presbyopic | 8 |
| 3 | presbyopic | 8 |

Class: contact-lenses (Nom) Visualize All

Status: OK Log x 0

Example – Contact Lens Data

Can you think of problems that might be solved with this data?

The screenshot shows the Weka Explorer interface with the 'contact-lenses' dataset loaded. The 'age' attribute is selected, and its distribution is visualized as three stacked bar charts. The distribution is as follows:

| Label | Count |
|----------------|-------|
| young | 8 |
| pre-presbyopic | 8 |
| presbyopic | 8 |

The status bar at the bottom indicates 'OK' and 'Log'.

Example – Contact Lens Data

If supervised learning were to be done, which would be the output attribute, do you think?

The screenshot shows the Weka Explorer interface with the 'contact-lenses' dataset loaded. The 'age' attribute is selected, and its distribution is visualized as three stacked bar charts. The distribution is as follows:

| Label | Count |
|----------------|-------|
| young | 8 |
| pre-presbyopic | 8 |
| presbyopic | 8 |

The status bar at the bottom indicates 'Status OK' and 'Log'.

Example – Contact Lens Data

Weka Explorer

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

Open file... | Open URL... | Open DB... | Generate... | Undo | Edit... | Save...

Filter: Choose **None** Apply

Current relation: Relation: contact-lenses, Instances: 24, Attributes: 5

Attributes: All | None | Invert | Pattern

| No. | Name |
|-----|--|
| 1 | <input type="checkbox"/> age |
| 2 | <input type="checkbox"/> spectacle-prescrip |
| 3 | <input type="checkbox"/> astigmatism |
| 4 | <input type="checkbox"/> tear-prod-rate |
| 5 | <input checked="" type="checkbox"/> contact-lenses |

Remove

Selected attribute: Name: contact-lenses, Missing: 0 (0%), Distinct: 3, Type: Nominal, Unique: 0 (0%)

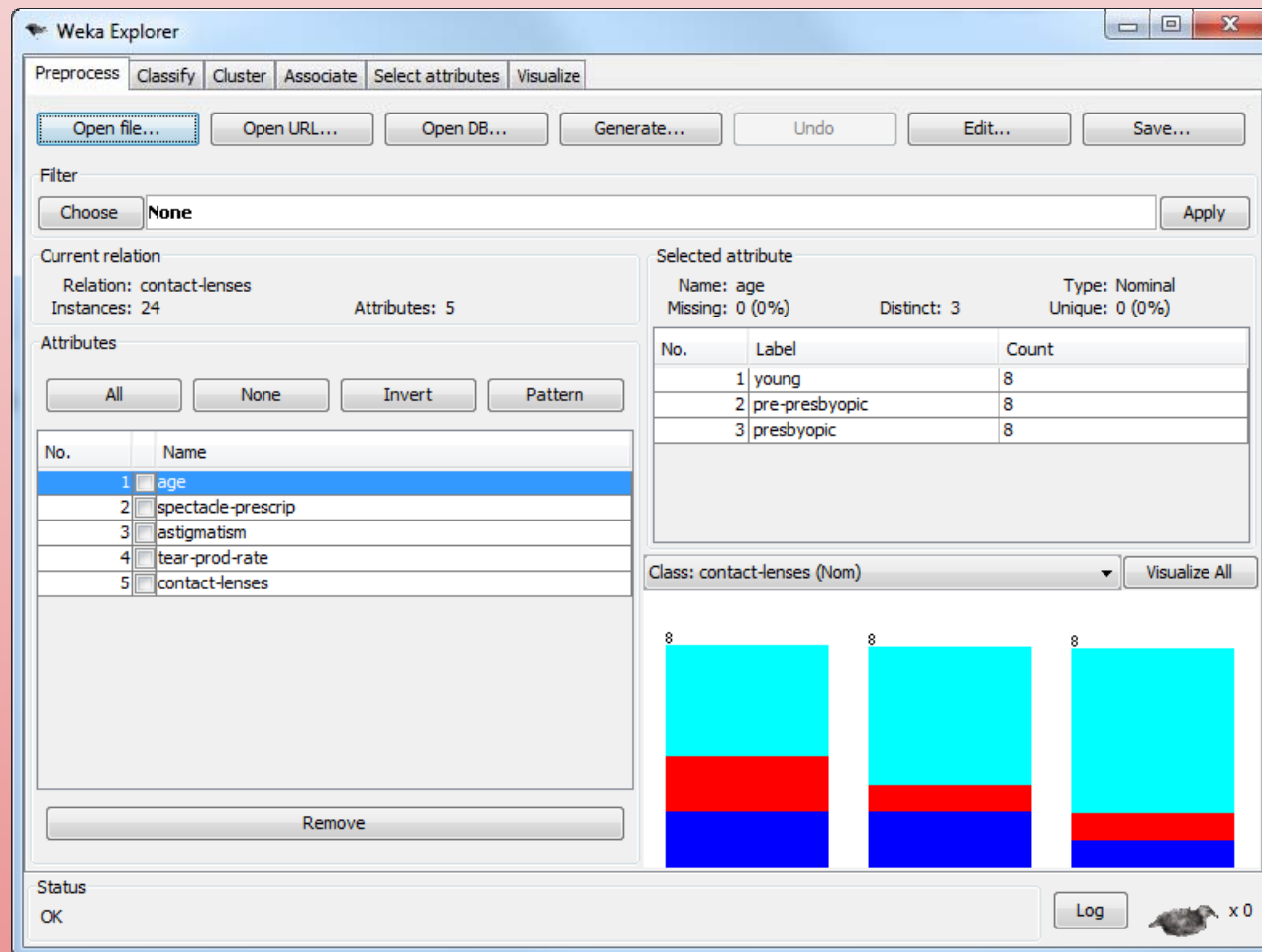
| No. | Label | Count |
|-----|-------|-------|
| 1 | soft | 5 |
| 2 | hard | 4 |
| 3 | none | 15 |

Class: contact-lenses (Nom) Visualize All

5 SOFT, 4 HARD, 15 NONE

Status: OK Log x 0

Example – Contact Lens Data



Weka Explorer

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

Open file... | Open URL... | Open DB... | Generate... | Undo | Edit... | Save...

Filter: Choose **None** Apply

Current relation: Relation: contact-lenses, Instances: 24, Attributes: 5

Attributes: All | None | Invert | Pattern

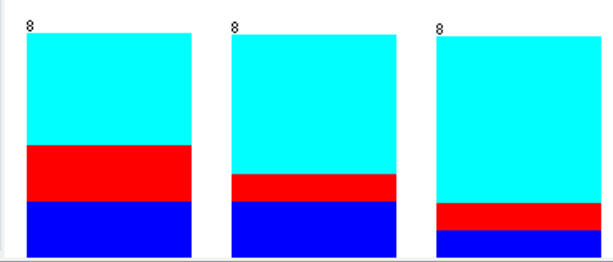
| No. | Name |
|-----|---|
| 1 | <input checked="" type="checkbox"/> age |
| 2 | <input type="checkbox"/> spectacle-prescrip |
| 3 | <input type="checkbox"/> astigmatism |
| 4 | <input type="checkbox"/> tear-prod-rate |
| 5 | <input type="checkbox"/> contact-lenses |

Remove

Selected attribute: Name: age, Missing: 0 (0%), Distinct: 3, Type: Nominal, Unique: 0 (0%)

| No. | Label | Count |
|-----|----------------|-------|
| 1 | young | 8 |
| 2 | pre-presbyopic | 8 |
| 3 | presbyopic | 8 |

Class: contact-lenses (Nom) Visualize All



Status: OK Log x 0

Example – Contact Lens Data

Weka Explorer

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

Open file... | Open URL... | Open DB... | Generate... | Undo | Edit... | Save...

Filter: Choose **None** Apply

Current relation: Relation: contact-lenses, Instances: 24, Attributes: 5

Selected attribute: Name: age, Missing: 0 (0%), Distinct: 3, Type: Nominal, Unique: 0 (0%)

| No. | Label | Count |
|-----|----------------|-------|
| 1 | young | 8 |
| 2 | pre-presbyopic | 8 |
| 3 | presbyopic | 8 |

Attributes: All | None | Invert | Pattern

| No. | Name |
|-------------------------------------|----------------------|
| <input checked="" type="checkbox"/> | 1 age |
| <input type="checkbox"/> | 2 spectacle-prescrip |
| <input type="checkbox"/> | 3 astigmatism |
| <input type="checkbox"/> | 4 tear-prod-rate |
| <input type="checkbox"/> | 5 contact-lenses |

Class: contact-lenses (Nom) Visualize All

8 ← None → 8 8

← Hard →

← Soft →

Status: OK Log x 0

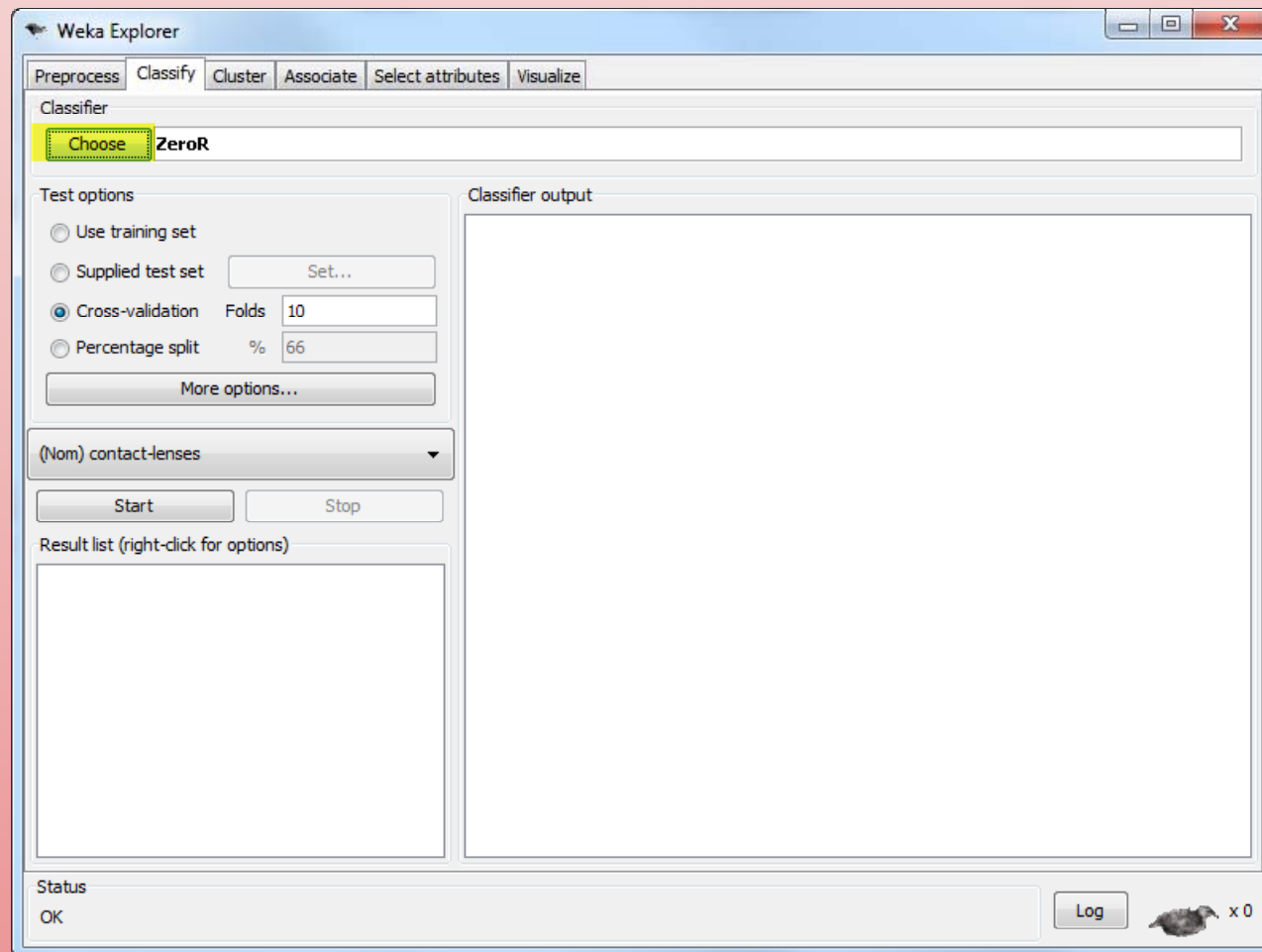
Example – Classify - Contact Lens Data

The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. The 'Current relation' is 'contact-lenses' with 24 instances and 5 attributes. The 'Attributes' list includes 'age', 'spectacle-prescrip', 'astigmatism', 'tear-prod-rate', and 'contact-lenses'. The 'Selected attribute' table shows the distribution of the 'contact-lenses' attribute:

| No. | Label | Count |
|-----|-------|-------|
| 1 | soft | 5 |
| 2 | hard | 4 |
| 3 | none | 15 |

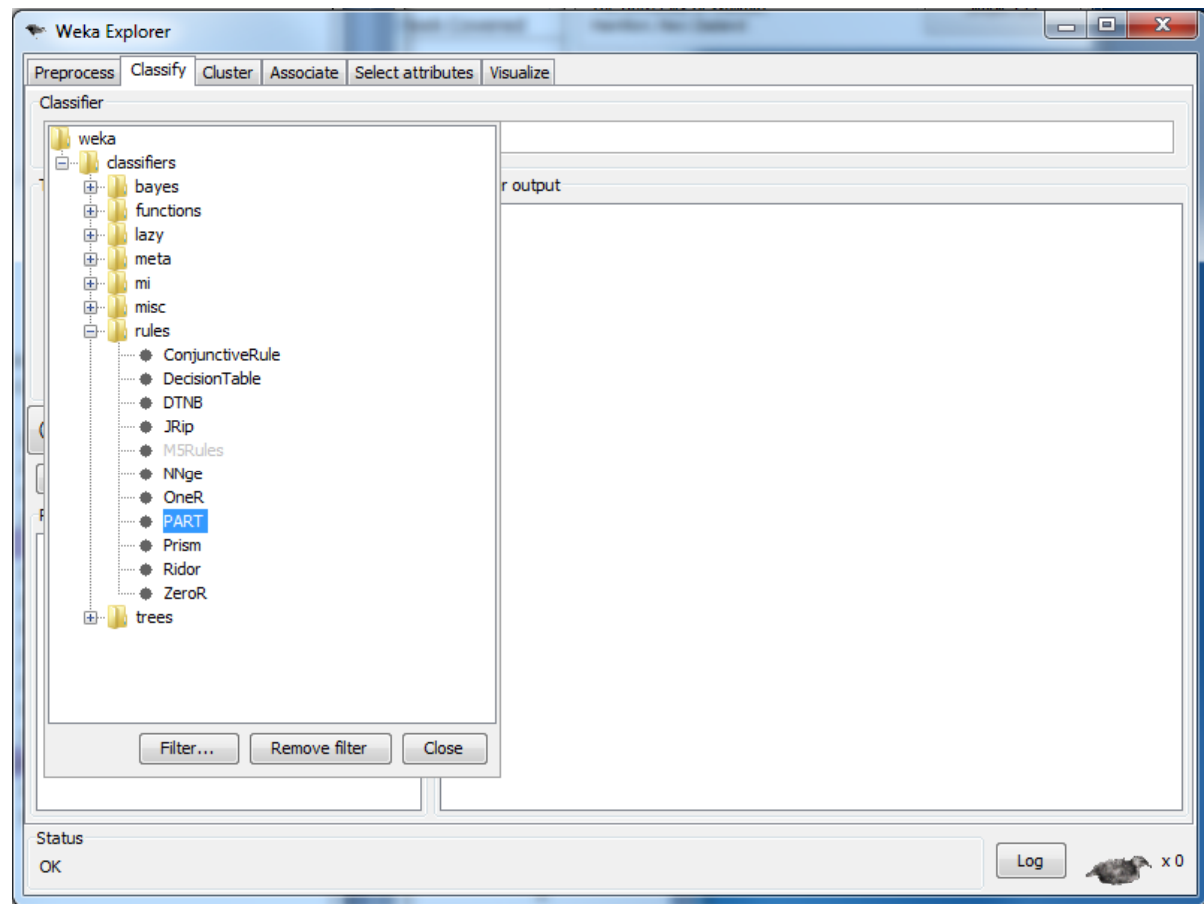
A bar chart below the table visualizes this distribution with three bars: a blue bar for 'soft' (count 5), a red bar for 'hard' (count 4), and a cyan bar for 'none' (count 15). The status bar at the bottom shows 'OK' and a 'Log' button.

Example – Classify - Contact Lens Data

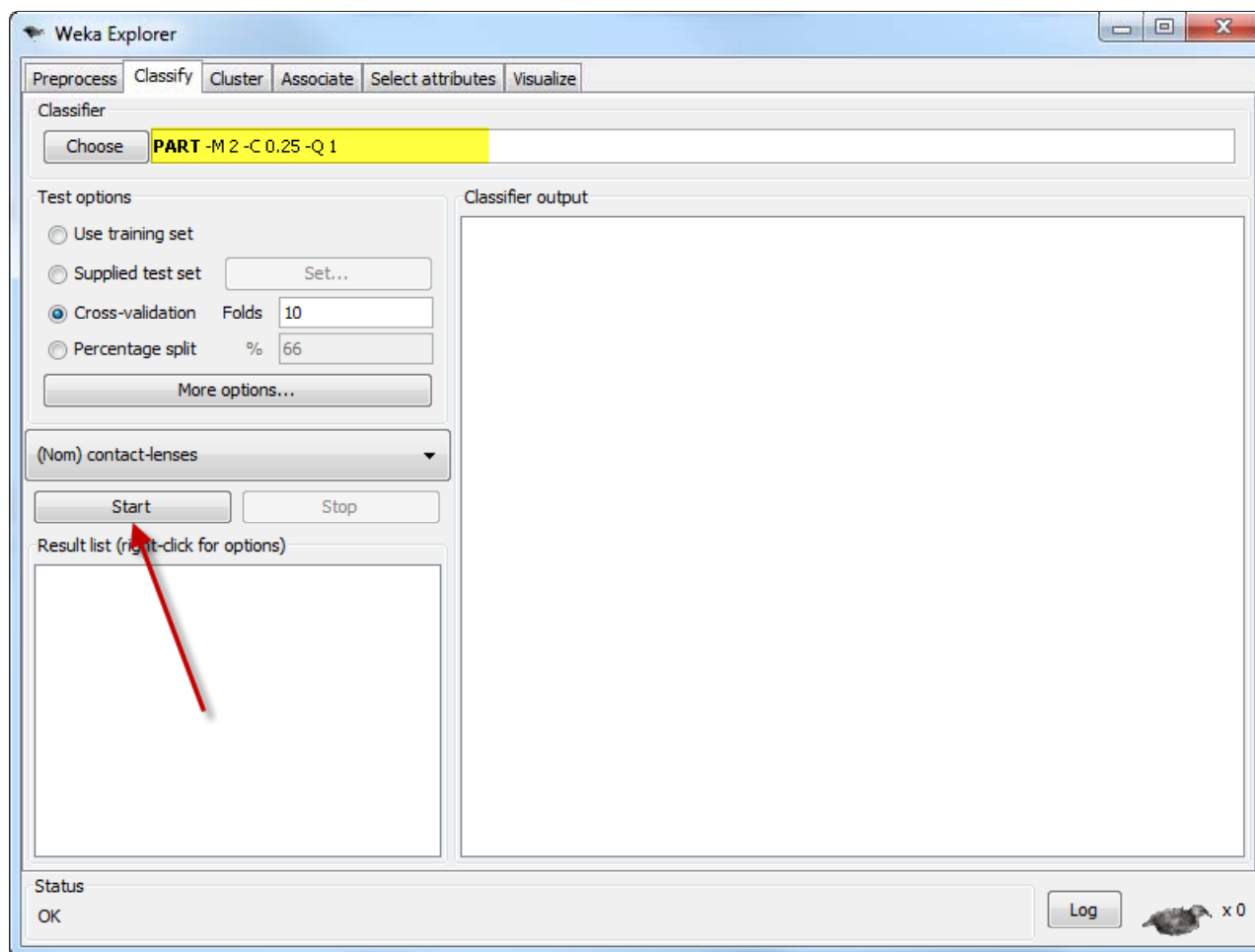


Example – Classify - Contact Lens Data

Select the
rule
generator
named
PART from
the list that
shows up
after you
select
Choose



Example – Classify - Contact Lens Data



Example – Classify - Contact Lens Data

The screenshot shows the Weka Explorer interface with the following details:

- Classifier:** PART -M 2 -C 0.25 -Q 1
- Test options:** Cross-validation (10 folds), Percentage split (66%)
- Classifier output:**

```
=== Run information ===  
Scheme:      weka.classifiers.rules.PART -M 2 -C 0.25 -Q 1  
Relation:    contact-lenses  
Instances:   24  
Attributes:  5  
              age  
              spectacle-prescrip  
              astigmatism  
              tear-prod-rate  
              contact-lenses  
Test mode:   10-fold cross-validation  
  
=== Classifier model (full training set) ===  
  
PART decision list  
-----  
  
tear-prod-rate = reduced: none (12.0)  
  
astigmatism = no: soft (6.0/1.0)  
  
spectacle-prescrip = myope: hard (3.0)  
  
: none (3.0/1.0)  
  
Number of Rules :      4
```

10-Fold Cross-Validation

- Data is partitioned into 10 equally (or nearly equally) sized segments or folds
- 10 iterations of training and validation are completed
- In each iteration a different fold of the data is held out for validation, with the remaining 9 folds used for learning

Example – Classify - Contact Lens Data

The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. The classifier chosen is 'PART -M 2 -C 0.25 -Q 1'. The test options are set to 'Cross-validation' with 10 folds. The classifier output window displays the following information:

```
=== Run information ===  
Scheme:      weka.classifiers.rules.PART -M 2 -C 0.25 -Q 1  
Relation:    contact-lenses  
Instances:   24  
Attributes:  5  
             age  
             spectacle-prescrip  
             astigmatism  
             tear-prod-rate  
             contact-lenses  
Test mode:   10-fold cross-validation  
  
=== Classifier model (full training set) ===  
  
PART decision list  
-----  
  
tear-prod-rate = reduced: none (12.0)  
  
astigmatism = no: soft (6.0/1.0)  
  
spectacle-prescrip = myope: hard (3.0)  
  
: none (3.0/1.0)  
  
Number of Rules :      4
```

The 'Result list' on the left shows a single entry: '17:43:50 - rules.PART'. The status bar at the bottom indicates 'Status OK' and a 'Log' button.

Example – Classify - Contact Lens Data

IF tear-prod-
rate = reduced
THEN contact-
lenses = none

IF astigmatism
= no THEN
contact-lenses
= soft

The screenshot shows the Weka Explorer interface. The classifier selected is PART -M 2 -C 0.25 -Q 1. The test mode is 10-fold cross-validation. The classifier output is as follows:

```
=== Run information ===  
Scheme:      weka.classifiers.rules.PART -M 2 -C 0.25 -Q 1  
Relation:    contact-lenses  
Instances:   24  
Attributes:  5  
age  
spectacle-prescrip  
astigmatism  
tear-prod-rate  
contact-lenses  
Test mode:   10-fold cross-validation  
  
=== Classifier model (full training set) ===  
PART decision list  
-----  
tear-prod-rate = reduced: none (12.0)  
astigmatism = no: soft (6.0/1.0)  
spectacle-prescrip = myope: hard (3.0)  
: none (3.0/1.0)  
Number of Rules :      4
```

A red callout box points to the value 12.0 in the first rule, with the text "coverage = 12".

Example – Classify - Contact Lens Data

Coverage =
12

The screenshot shows the Weka Explorer interface with the 'tear-prod-rate' attribute selected. The 'Selected attribute' table shows the following data:

| No. | Label | Count |
|-----|---------|-------|
| 1 | reduced | 12 |
| 2 | normal | 12 |

The visualization shows two bars, each with a count of 12. The left bar is cyan, and the right bar is divided into three segments: cyan (top), red (middle), and blue (bottom). A red callout box points to the cyan bar with the text 'coverage = 12'.

Example – Classify - Contact Lens Data

IF tear-prod-
rate = reduced
THEN contact-
lenses = none

IF astigmatism
= no THEN
contact-lenses
= soft

The screenshot shows the Weka Explorer interface. The classifier selected is PART -M 2 -C 0.25 -Q 1. The test options are set to Cross-validation with 10 folds. The classifier output window displays the following information:

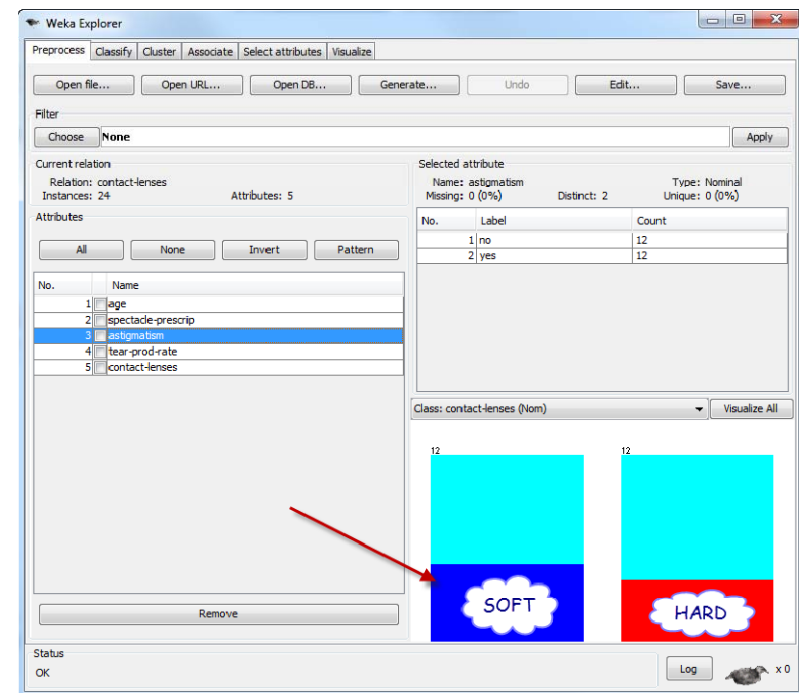
```
=== Run information ===  
Scheme: weka.classifiers.rules.PART -M 2 -C 0.25 -Q 1  
Relation: contact-lenses  
Instances: 24  
Attributes: 5  
age  
spectacle-prescrip  
astigmatism  
tear-prod-rate  
contact-lenses  
Test mode: 10-fold cross-validation  
  
=== Classifier model (full training set) ===  
PART decision list  
-----  
tear-prod-rate = reduced: none (12.0)  
astigmatism = no: soft (6.0/1.0)  
spectacle-prescrip = myope: hard (3.0)  
: none (3.0/1.0)  
Number of Rules : 4
```

A red callout box highlights the decision list entry for "astigmatism = no: soft (6.0/1.0)" with the following text:

coverage = 6
misclassification = 1
accuracy = 5/6 = 83.3%

Example – Classify - Contact Lens Data

- Coverage = 6
- Misclassification = 1
- Accuracy = $5/6 = 83.3\%$



Example – Classify - Contact Lens Data

The screenshot shows the Weka Explorer interface with the PART classifier selected. The classifier output is displayed in the right pane, showing stratified cross-validation results and a confusion matrix. The results are highlighted in yellow.

Classifier output

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      20      83.3333 %
Incorrectly Classified Instances    4       16.6667 %
Kappa statistic                    0.71
Mean absolute error                 0.15
Root mean squared error             0.3249
Relative absolute error             39.7059 %
Root relative squared error         74.3898 %
Total Number of Instances          24

--- Detailed Accuracy By Class ---

      TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area
      1       0.053   0.833     1      0.909     0.947
      0.75    0.1     0.6       0.75   0.667     0.813
      0.8     0.111   0.923     0.8    0.857     0.811
Weighted Avg. 0.833   0.097    0.851   0.833   0.836     0.84

=== Confusion Matrix ===

 a  b  c  <-- classified as
5  0  0 | a = soft
0  3  1 | b = hard
1  2 12 | c = none
```

Test options

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

Result list (right-click for options)

- 17:43:50 - rules.PART

WEKA

A Data Mining Tool

By Susan L. Miertschin