Machine Learning with WEKA

Eibe Frank

Department of Computer Science, University of Waikato, New Zealand

- WEKA: A Machine Learning Toolkit
- The Explorer
  - Classification and Regression
  - Clustering
  - Association Rules
  - Attribute Selection
  - Data Visualization
- The Experimenter
- The Knowledge Flow GUI
- Conclusions
WEKA: the bird

Copyright: Martin Kramer (mkramer@wxs.nl)
WEKA: the software

- Machine learning/data mining software written in Java (distributed under the GNU Public License)
- Used for research, education, and applications
- Complements “Data Mining” by Witten & Frank
- Main features:
  - Comprehensive set of data pre-processing tools, learning algorithms and evaluation methods
  - Graphical user interfaces (incl. data visualization)
  - Environment for comparing learning algorithms
WEKA: versions

- There are several versions of WEKA:
  - WEKA 3.2: “GUI version” adds graphical user interfaces (book version is command-line only)
  - WEKA 3.3: “development version” with lots of improvements

- This talk is based on the latest snapshot of WEKA 3.3 (soon to be WEKA 3.4)
WEKA only deals with “flat” files

@relation heart-disease-simplified

@attribute age numeric
@attribute sex { female, male}
@attribute chest_pain_type { typ_angina, asympt, non_anginal, atyp_angina}
@attribute cholesterol numeric
@attribute exercise_induced_angina { no, yes}
@attribute class { present, not_present}

@data
63,male,typ_angina,233,no,not_present
67,male,asympt,286,yes,present
67,male,asympt,229,yes,present
38,female,non_anginal,?,no,not_present
...
WEKA only deals with “flat” files

@relation heart-disease-simplified

@attribute age numeric
@attribute sex { female, male}
@attribute chest_pain_type { typ_angina, asympt, non_anginal, atyp_angina}
@attribute cholesterol numeric
@attribute exercise_induced_angina { no, yes}
@attribute class { present, not_present}

@data
63,male,typ_angina,233,no,not_present
67,male,asympt,286,yes,present
67,male,asympt,229,yes,present
38,female,non_anginal,?,no,not_present
...
Explorer: pre-processing the data

- Data can be imported from a file in various formats: ARFF, CSV, C4.5, binary
- Data can also be read from a URL or from an SQL database (using JDBC)
- Pre-processing tools in WEKA are called “filters”
- WEKA contains filters for:
  - Discretization, normalization, resampling, attribute selection, transforming and combining attributes, …
**Current relation**

Relation: iris  
Instances: 150  
Attributes: 5

**Attributes**

<table>
<thead>
<tr>
<th>No.</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>sepalwidth</td>
</tr>
<tr>
<td>2</td>
<td>sepalwidth</td>
</tr>
<tr>
<td>3</td>
<td>petallength</td>
</tr>
<tr>
<td>4</td>
<td>petallength</td>
</tr>
<tr>
<td>5</td>
<td>class</td>
</tr>
</tbody>
</table>

**Selected attribute**

Name: petallength  
Missing: 0 (0%)  
Distinct: 43  
Unique: 10 (7%)

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum</td>
<td>1</td>
</tr>
<tr>
<td>Maximum</td>
<td>6.9</td>
</tr>
<tr>
<td>Mean</td>
<td>3.759</td>
</tr>
<tr>
<td>StdDev</td>
<td>1.764</td>
</tr>
</tbody>
</table>

**Colour: class (Nom)**

![Bar chart showing the distribution of class values]
Weka Knowledge Explorer

Filter
Choose Discretize -B 10 -R first-last
weka.gui.GenericObjectEditor
weka.filters.unsupervised.attribute.Discretize
About
An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes.

Attributes

<table>
<thead>
<tr>
<th>No.</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>sepallength</td>
</tr>
<tr>
<td>2</td>
<td>sepalwidth</td>
</tr>
<tr>
<td>3</td>
<td>petallength</td>
</tr>
<tr>
<td>4</td>
<td>petalwidth</td>
</tr>
<tr>
<td>5</td>
<td>class</td>
</tr>
</tbody>
</table>

Attribute Indices: first–last
Bins: 10
Find Num Bins: False
Invert Selection: False
Make Binary: False
Use Equal Frequency: False

Visualize All

Status
OK
Weka Knowledge Explorer

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

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

Filter
Choose Discretize -F -B 10 -R first-last

Apply

Current relation
Relation: iris
Instances: 150
Attributes: 5

Attributes
No. | Name
--- | ---
1 | sepalwidth
2 | sepalwidth
3 | petallength
4 | petalwidth
5 | class

Selected attribute
Name: petallength
Missing: 0 (0%) Distinct: 43 Unique: 10 (7%)

Statistic | Value
--- | ---
Minimum | 1
Maximum | 6.9
Mean | 3.759
StdDev | 1.764

Colour: class (Nom)

Visualize All

Status
OK

Log

x 0
Weka Knowledge Explorer

Filter
Choose Discretize -F -B 10 -R first-last

Current relation
Relation: iris-weka.filters.unsupervised.attribute.Discretize
Instances: 150 Attributes: 5

Attributes
<table>
<thead>
<tr>
<th>No.</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>sepalwidth</td>
</tr>
<tr>
<td>2</td>
<td>petallength</td>
</tr>
<tr>
<td>3</td>
<td>sepallength</td>
</tr>
<tr>
<td>4</td>
<td>petalwidth</td>
</tr>
<tr>
<td>5</td>
<td>class</td>
</tr>
</tbody>
</table>

Selected attribute
Name: petallength
Type: Nominal
Missing: 0 (0%) Distinct: 10 Unique: 0 (0%)

<table>
<thead>
<tr>
<th>Label</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>'(inf-1.45]'</td>
<td>23</td>
</tr>
<tr>
<td>'(1.45-1.55]'</td>
<td>14</td>
</tr>
<tr>
<td>'(1.55-1.8]'</td>
<td>13</td>
</tr>
<tr>
<td>'(1.8-3.95]'</td>
<td>11</td>
</tr>
<tr>
<td>'(3.95-4.35]'</td>
<td>14</td>
</tr>
<tr>
<td>'(4.35-4.65]'</td>
<td>18</td>
</tr>
<tr>
<td>'(4.65-5.05]'</td>
<td>15</td>
</tr>
</tbody>
</table>

Colour: class (Nom)

Status
OK
Explorer: building “classifiers”

- Classifiers in WEKA are models for predicting nominal or numeric quantities

- Implemented learning schemes include:
  - Decision trees and lists, instance-based classifiers, support vector machines, multi-layer perceptrons, logistic regression, Bayes’ nets, ...

- “Meta”-classifiers include:
  - Bagging, boosting, stacking, error-correcting output codes, locally weighted learning, ...
Choose J48 -C 0.25 -M 2

Cross-validation Folds 10

(Nom) class

Start

Status OK
Choose J48 -C 0.25 -M 2

Cross-validation Folds 10

Start

Status OK
Choose J48 -C 0.25 -M 2

Percentage split % 66

Output model
Output per-class stats
Output confusion matrix
Store predictions for visualization

Random seed for XVal / % Split 1

OK
Choose J48 -C 0.25 -M 2

Test options
- Use training set
- Supplied test set
- Cross-validation Folds 10
- Percentage split % 66

Classifier output

Start

Result list (right-click for options)
Weka Knowledge Explorer

Classifier

Choose J48 -C 0.25 -M 2

Test options

- Use training set
- Supplied test set
- Cross-validation Folds 10
- Percentage split % 66

Classifier output

=== Run information ===
Scheme: weka.classifiers.trees.J48 J48 -C 0.25 -M 2
Relation: iris
Instances: 150
Attributes: 5
  sepalwidth
  petalwidth
  class
Test mode: split 66% train, remainder test

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

J48 pruned tree

petalwidth <= 0.6: Iris-setosa (50.0)
petalwidth > 0.6
  |  petalwidth <= 1.7
  |   |  petalwidth <= 4.9: Iris-versicolor (48.0/1.0)
  |   |  petalwidth > 4.9
  |   |   |  petalwidth <= 1.5: Iris-virginica (3.0)
  |   |   |  petalwidth > 1.5: Iris-versicolor (3.0/1.0)
  |   |  petalwidth > 1.7: Iris-virginica (46.0/1.0)

Number of Leaves :  5
Classifier

Choose: J48 -C 0.25 -M 2

Test options
- Use training set
- Supplied test set
- Cross-validation Folds 10
- Percentage split % 66

Classifier output

--- Run information ---

Scheme: weka.classifiers.trees.j48.J48 -C 0.25 -M 2
Relation: iris
Instances: 150
Attributes: 5
  sepal length
  sepal width
  petal length
  petal width
  class

Test mode: split 66% train, remainder test

--- Classifier model (full training set) ---

J48 pruned tree

-------------

petal width <= 0.6: Iris-setosa (50.0)
petal width > 0.6
  | petal width <= 0.6: Iris-setosa (50.0)
  | petal width > 0.6
  |   | petal length <= 4.9: Iris-versicolor (48.0/1.0)
  |   | petal length > 4.9
  |   |   | petal width <= 1.5: Iris-virginica (3.0)
  |   |   | petal width > 1.5: Iris-versicolor (3.0/1.0)
  |   |   | petal width > 1.7: Iris-virginica (46.0/1.0)

Number of Leaves : 5
 Classifier:

Choose J48 -C 0.25 -M 2

Test options:
- Use training set
- Supplied test set
- Cross-validation Folds 10
- Percentage split % 66

Classifier output:

Time taken to build model: 0.24 seconds

--- Evaluation on test split ---

--- Summary ---

Correctly Classified Instances 49 96.0784%
Incorrectly Classified Instances 2 3.9216%
Kappa statistic 0.9408
Mean absolute error 0.0396
Root mean squared error 0.1579
Relative absolute error 8.8979%
Root relative squared error 33.4091%
Total Number of Instances 51

--- Detailed Accuracy By Class ---

<table>
<thead>
<tr>
<th>Class</th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iris-setosa</td>
<td>1</td>
<td>0.063</td>
<td>0.905</td>
<td>1</td>
<td>0.95</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>Iris-versicolor</td>
<td>0.882</td>
<td>0.025</td>
<td>0.882</td>
<td>0.882</td>
<td>0.938</td>
<td>Iris-versicolor</td>
</tr>
</tbody>
</table>

--- Confusion Matrix ---

a b c   <-- classified as
15 0 0   a = Iris-setosa
0 19 0   b = Iris-versicolor
0 2 15   c = Iris-virginica
Weka Knowledge Explorer

Classifier
Choose J48 -C 0.25 -M 2

Test options
- Use training set
- Supplied test set
- Cross-validation Folds 10
- Percentage split % 66

Classifier output
Time taken to build model: 0.24 seconds

=== Evaluation on test split ===
=== Summary ===
Correctly Classified Instances 49 96.0784 %
Incorrectly Classified Instances 2 3.9216 %
Kappa statistic 0.9408
Mean absolute error 0.0396
Root mean squared error 0.1579
Relative absolute error 8.8979 %
Root relative squared error 33.4091 %
Total Number of Instances 51

=== Detailed Accuracy By Class ===

<table>
<thead>
<tr>
<th>Class</th>
<th>Recall</th>
<th>F-Measure</th>
<th>1st</th>
<th>Iris-setosa</th>
<th>0.882</th>
<th>0.938</th>
<th>Iris-virginica</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Iris-setosa</td>
<td>0.882</td>
<td>0.938</td>
<td>Iris-virginica</td>
<td></td>
</tr>
<tr>
<td>Iris-versicolor</td>
<td>0.95</td>
<td></td>
<td>1</td>
<td>Iris-versicolor</td>
<td>0.882</td>
<td>0.938</td>
<td>Iris-virginica</td>
</tr>
</tbody>
</table>

Result list (right-click for options)
11:49:05 - trees.j48.j48

View in main window
View in separate window
Save result buffer

Load model
Save model
Re-evaluate model on current test set

Visualize classifier errors
Visualize tree

Status OK

Log x 0
Time taken to build model: 0.24 seconds

Correctly Classified Instances 49 96.0784 %
Incorrectly Classified Instances 2 3.9216 %
Kappa statistic 0.9408
Mean absolute error 0.0396
Root mean squared error 0.1579
Relative absolute error 8.8979 %
Root relative squared error 33.4091 %
Total Number of Instances 51

Detailed Accuracy By Class

<table>
<thead>
<tr>
<th>Class</th>
<th>Recall</th>
<th>F-Measure</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iris-setosa</td>
<td>1</td>
<td>1</td>
<td>Iris-versicolor</td>
</tr>
<tr>
<td>Iris-versicolor</td>
<td>0.882</td>
<td>0.938</td>
<td>Iris-virginica</td>
</tr>
<tr>
<td>Iris-virginica</td>
<td>1</td>
<td>1</td>
<td>Iris-setosa</td>
</tr>
</tbody>
</table>
J48 -C 0.25 -M 2

Time taken to build model: 0.24 seconds

Correctly Classified Instances 49 96.0784 %
Incorrectly Classified Instances 2 3.9216 %
Kappa statistic 0.9408
Mean absolute error 0.0396
Root mean squared error 0.1579
Relative absolute error 8.8979 %
Root root mean squared error 33.4091 %
Total Number of Instances 51

Detailed Accuracy By Class

<table>
<thead>
<tr>
<th>Class</th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iris-setosa</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>Iris-versicolor</td>
<td>0.882</td>
<td>0.063</td>
<td>0.905</td>
<td>1</td>
<td>0.95</td>
<td>Iris-versicolor</td>
</tr>
<tr>
<td>Iris-virginica</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.882</td>
<td>0.938</td>
<td>Iris-virginica</td>
</tr>
</tbody>
</table>

Confusion Matrix

```
a  b  c  <-- classified as
15  0  0  | a = Iris-setosa
  0 19  0  | b = Iris-versicolor
  0  2 15 | c = Iris-virginica
```
Choose J48 -C 0.25 -M 2

Time taken to build model: 0.24 seconds

--- Evaluation on test split ---
--- Summary ---

Correctly Classified Instances 49 96.0784 %
Incorrectly Classified Instances 2 3.9216 %
Kappa statistic 0.9408
Mean absolute error 0.0396
Root mean squared error 0.1579
Relative absolute error 8.8979 %
Root relative squared error 33.4091 %
Total Number of Instances 51

--- Detailed Accuracy By Class ---

<table>
<thead>
<tr>
<th>Class</th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iris-setosa</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Iris-versicolor</td>
<td>1</td>
<td>0.063</td>
<td>0.905</td>
<td>1</td>
<td>0.95</td>
</tr>
<tr>
<td>Iris-virginica</td>
<td>0.882</td>
<td>0</td>
<td>1</td>
<td>0.882</td>
<td>0.938</td>
</tr>
</tbody>
</table>

--- Confusion Matrix ---

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>--- classified as</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>b</td>
<td>0</td>
<td>19</td>
<td>0</td>
<td>Iris-versicolor</td>
</tr>
<tr>
<td>c</td>
<td>0</td>
<td>2</td>
<td>15</td>
<td>Iris-virginica</td>
</tr>
</tbody>
</table>
Evaluation on test split ===

Summary ===

<table>
<thead>
<tr>
<th>Class</th>
<th>Correctly Classified Instances</th>
<th>% Correctly Classified Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iris-setosa</td>
<td>49</td>
<td>96.0784%</td>
</tr>
<tr>
<td>Iris-versicolor</td>
<td>2</td>
<td>3.9216%</td>
</tr>
</tbody>
</table>

Confusion Matrix ===

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>15</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>b</td>
<td>0</td>
<td>19</td>
<td>0</td>
</tr>
<tr>
<td>c</td>
<td>0</td>
<td>2</td>
<td>15</td>
</tr>
</tbody>
</table>

FP Rate  | Precision | Recall | F-Measure | Class
----------|------------|--------|-----------|-------
0.063     | 0.905      | 1      | 0.95      | Iris-setosa
0          | 1          | 0.882  | 0.938     | Iris-versicolor
0          | 2          | 15     |           | Iris-virginica

Time taken to build model: 0.24 seconds
NeuralNetwork -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a

Evaluation on test split
Summary
Correctly Classified Instances 49 96.0784 %
Incorrectly Classified Instances 2 3.9216 %
Kappa statistic 0.9408
Mean absolute error 0.0396
Root mean squared error 0.1579
Relative absolute error 8.8979 %
Root relative squared error 33.4091 %
Total Number of Instances 51

Detailed Accuracy By Class

<table>
<thead>
<tr>
<th>Class</th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iris-setosa</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Iris-versicolor</td>
<td>0.63</td>
<td>0</td>
<td>0.905</td>
<td>1</td>
<td>0.95</td>
</tr>
<tr>
<td>Iris-virginica</td>
<td>0.882</td>
<td>0</td>
<td>1</td>
<td>0.882</td>
<td>0.938</td>
</tr>
</tbody>
</table>

Confusion Matrix

a b c  <-- classified as
15 0 0 | a = Iris-setosa
0 19 0 | b = Iris-versicolor
0 2 15 | c = Iris-virginica
Classifier:

Choose: NeuralNetwork -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a

Test options:
- Use training set
- Supplied test set
- Cross-validation Folds 10
- Percentage split % 66

Classifier output:

=== Evaluation on test split ===

Summary:

Correctly Classified Instances 49
Incorrectly Classified Instances 2
Kappa statistic 0.9408
Mean absolute error 0.0396
Root mean squared error 0.1579
Relative absolute error 8.8979%
Root relative squared error 33.4091%
Total Number of Instances 51

Detailed Accuracy By Class:

<table>
<thead>
<tr>
<th>Class</th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iris-setosa</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>Iris-versicolor</td>
<td>1</td>
<td>0.063</td>
<td>0.905</td>
<td>1</td>
<td>0.95</td>
<td>Iris-versicolor</td>
</tr>
<tr>
<td>Iris-virginica</td>
<td>0.882</td>
<td>0</td>
<td>1</td>
<td>0.882</td>
<td>0.938</td>
<td>Iris-virginica</td>
</tr>
</tbody>
</table>

Confusion Matrix:

a b c  <-- classified as
15 0 0 | a = Iris-setosa
0 19 0 | b = Iris-versicolor
0 2 15 | c = Iris-virginica
Neural Network

Sepal length
Sepal width
Petal length
Petal width
Iris-setosa
Iris-versicolor
Iris-virginica

Start
Epoch 0
Num Of Epochs 500
Error per Epoch = 0

Learning Rate = 0.3
Momentum = 0.2

Building model on training data...
Classifier

Choose: NeuralNetwork -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a -G -R

Test options
- Use training set
- Supplied test set
- Cross-validation Folds 10
- Percentage split % 66

Classifier output

=== Evaluation on test split ===

=== Summary ===

Correctly Classified Instances 50 98.0392 %
Incorrectly Classified Instances 1 1.9608 %
Kappa statistic 0.9704
Mean absolute error 0.0239
Root mean squared error 0.1101
Relative absolute error 5.3594 %
Root relative squared error 23.2952 %
Total Number of Instances 51

=== Detailed Accuracy By Class ===

<table>
<thead>
<tr>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>1</td>
<td>0.031</td>
<td>0.95</td>
<td>1</td>
<td>0.974</td>
<td>Iris-versicolor</td>
</tr>
<tr>
<td>0.941</td>
<td>0</td>
<td>1</td>
<td>0.941</td>
<td>0.97</td>
<td>Iris-virginica</td>
</tr>
</tbody>
</table>

=== Confusion Matrix ===

a b c <-- classified as
15 0 0 | a = Iris-setosa
0 19 0 | b = Iris-versicolor
0 1 16 | c = Iris-virginica
Classifier output

=== Evaluation on test split ===

=== Summary ===

Correctly Classified Instances 48 94.1176 %
Incorrectly Classified Instances 3 5.8824 %
Kappa statistic 0.9113
Mean absolute error 0.0447
Root mean squared error 0.1722
Relative absolute error 10.0365 %
Root relative squared error 36.4196 %
Total Number of Instances 51

=== Detailed Accuracy By Class ===

<table>
<thead>
<tr>
<th></th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iris-setosa</td>
<td>1.00</td>
<td>0.063</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>Iris-versicolor</td>
<td>0.947</td>
<td>0.029</td>
<td>0.937</td>
<td>0.947</td>
<td>0.923</td>
<td>Iris-versicolor</td>
</tr>
<tr>
<td>Iris-virginica</td>
<td>0.882</td>
<td>0.015</td>
<td>0.938</td>
<td>0.882</td>
<td>0.909</td>
<td>Iris-virginica</td>
</tr>
</tbody>
</table>

=== Confusion Matrix ===

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>---classified as---</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>0</td>
<td>18</td>
<td>0</td>
<td>18</td>
<td>Iris-versicolor</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>15</td>
<td>17</td>
<td>Iris-virginica</td>
</tr>
</tbody>
</table>
NaiveBayes

Correctly Classified Instances 48  94.1176 %
Incorrectly Classified Instances 3  5.8824 %
Kappa statistic 0.9113
Mean absolute error 0.0447
Root mean squared error 0.1722
Relative absolute error 10.0365 %
Root relative squared error 36.4196 %
Total Number of Instances 51

Detailed Accuracy By Class

<table>
<thead>
<tr>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.063</td>
<td>0.9</td>
<td>1</td>
<td>0.947</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>0.882</td>
<td>0.029</td>
<td>0.938</td>
<td>0.882</td>
<td>0.909</td>
<td>Iris-versicolor</td>
</tr>
</tbody>
</table>

Confusion Matrix

a b c  --- classified as
15 0 15 | a = Iris-setosa
 0 18 1 | b = Iris-versicolor
 0  2 15 | c = Iris-virginica
Classifier

Choose NaiveBayes

Test options

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

More options...

(Nom) class

Start Stop

Classifier output

=== Evaluation on test split ===

=== Summary ===

Correctly Classified Instances 48 94.1176 %
Incorrectly Classified Instances 3 5.8824 %
Kappa statistic 0.9113
Mean absolute error 0.0447
Root mean squared error 0.1722
Relative absolute error 10.0365 %
Root relative squared error 36.4196 %
Total Number of Instances 51

=== Detailed Accuracy By Class ===

<table>
<thead>
<tr>
<th>Class</th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iris-setosa</td>
<td>1</td>
<td>0.063</td>
<td>1</td>
<td>1</td>
<td>0.909</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>Iris-versicolor</td>
<td>0.947</td>
<td>0.029</td>
<td>0.938</td>
<td>0.882</td>
<td>0.909</td>
<td>Iris-versicolor</td>
</tr>
<tr>
<td>Iris-virginica</td>
<td>0.882</td>
<td>0.029</td>
<td>0.938</td>
<td>0.882</td>
<td>0.909</td>
<td>Iris-virginica</td>
</tr>
</tbody>
</table>

=== Confusion Matrix ===

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>c</th>
<th>--- classified as</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>0</td>
<td>0</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>0</td>
<td>18</td>
<td>1</td>
<td>Iris-versicolor</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>15</td>
<td>Iris-virginica</td>
</tr>
</tbody>
</table>

Status

OK
== Evaluation on test split ==
== Summary ==

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>correctly Classified Instances</td>
<td>48 94.1176 %</td>
</tr>
<tr>
<td>incorrectly Classified Instances</td>
<td>3 5.8824 %</td>
</tr>
<tr>
<td>appa statistic</td>
<td>0.9113</td>
</tr>
<tr>
<td>mean absolute error</td>
<td>0.0447</td>
</tr>
<tr>
<td>oot mean squared error</td>
<td>0.1722</td>
</tr>
<tr>
<td>relative absolute error</td>
<td>10.0365 %</td>
</tr>
<tr>
<td>oot relative squared error</td>
<td>36.4196 %</td>
</tr>
<tr>
<td>oot Number of Instances</td>
<td>51</td>
</tr>
</tbody>
</table>

== Detailed Accuracy By Class ==

<table>
<thead>
<tr>
<th>Class</th>
<th>Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iris-setosa</td>
<td>0.947</td>
<td>0.063</td>
<td>0.9</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Iris-versicolor</td>
<td>0.882</td>
<td>0.029</td>
<td>0.938</td>
<td>0.947</td>
<td>0.923</td>
</tr>
<tr>
<td>Iris-virginica</td>
<td>0.947</td>
<td>0.029</td>
<td>0.938</td>
<td>0.882</td>
<td>0.909</td>
</tr>
</tbody>
</table>

== Confusion Matrix ==

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>c</th>
<th>&lt;-- classified as</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>0</td>
<td>0</td>
<td>a = Iris-setosa</td>
</tr>
<tr>
<td>0</td>
<td>18</td>
<td>1</td>
<td>b = Iris-versicolor</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>15</td>
<td>c = Iris-virginica</td>
</tr>
</tbody>
</table>
Correctly Classified Instances 48  94.1176 %
Incorrectly Classified Instances 3   5.8824 %
Kappa statistic 0.9113
Mean absolute error 0.0447
Root mean squared error 0.1722
Relative absolute error 10.0365 %
Root relative squared error 36.4196 %
Total Number of Instances 51

=== Detailed Accuracy By Class ===

<table>
<thead>
<tr>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.947</td>
<td>0.063</td>
<td>1</td>
<td>1</td>
<td>0.947</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>0.882</td>
<td>0.029</td>
<td>0.938</td>
<td>1</td>
<td>0.882</td>
<td>Iris-versicolor</td>
</tr>
</tbody>
</table>

=== Confusion Matrix ===

a b c  <-- classified as
15 0 0 | a = Iris-setosa
0 18 1 | b = Iris-versicolor
0 2 15 | c = Iris-virginica
QuickTime™ and a TIFF (LZW) decompressor are needed to see this picture.
QuickTime™ and a TIFF (LZW) decompressor are needed to see this picture.
== Evaluation on test split ==
== Summary ==

Correctly Classified Instances   49   96.0784%
Incorrectly Classified Instances  2   3.9216%
Kappa statistic                   0.9408
Mean absolute error               0.0319
Root mean squared error           0.1622
Relative absolute error           7.1634%
Root relative squared error       34.312%
Total Number of Instances         51

== Detailed Accuracy By Class ==

<table>
<thead>
<tr>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>1</td>
<td>0.063</td>
<td>0.905</td>
<td>1</td>
<td>0.95</td>
<td>Iris-versicolor</td>
</tr>
<tr>
<td>0.882</td>
<td>0</td>
<td>1</td>
<td>0.882</td>
<td>0.938</td>
<td>Iris-virginica</td>
</tr>
</tbody>
</table>

== Confusion Matrix ==

a  b  c <- classified as
15 0 0 | a = Iris-setosa
0 19 0 | b = Iris-versicolor
0 2 15 | c = Iris-virginica
QuickTime™ and a TIFF (LZW) decompressor are needed to see this picture.
Classifier output

== Evaluation on test split ==

== Summary ==

| Correctly Classified Instances | 49 | 96.0784 % |
| Incorrectly Classified Instances | 2 | 3.9216 % |
| Kappa statistic | 0.9408 |
| Mean absolute error | 0.0319 |
| Root mean squared error | 0.1622 |
| Relative absolute error | 7.1634 % |
| Root relative squared error | 34.312 % |
| Total Number of Instances | 51 |

== Detailed Accuracy By Class ==

<table>
<thead>
<tr>
<th>Class</th>
<th>Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iris-setosa</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Iris-versicolor</td>
<td>1</td>
<td>0.063</td>
<td>0.905</td>
<td>1</td>
<td>0.95</td>
</tr>
<tr>
<td>Iris-virginica</td>
<td>0.882</td>
<td>1</td>
<td>0.882</td>
<td>0.938</td>
<td></td>
</tr>
</tbody>
</table>

== Confusion Matrix ==

```
15 0 0  a = Iris-setosa
0 19 0  b = Iris-versicolor
0 2 15  c = Iris-virginica
```
Weka Knowledge Explorer

Classifier

Choose M5P -M 4.0

Test options

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

More options...

(Num) petallength

Start  Stop

Result list (right-click for options)

14:34:28 - functions.neural.NeuralNetwork
14:48:05 - bayes.NaiveBayes
15:44:32 - trees.UserClassifier
15:49:03 - trees.m5.M5P

Classifier output

=== Run information ===

Scheme: weka.classifiers.trees.m5.M5P -M 4.0
Relation: iris
Instances: 150
Attributes: 5
  sepalwidth
  petallength
  class

Test mode: split 66% train, remainder test

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

M5 pruned model tree:
(using smoothed predictions)

petalwidth <= 0.8 : LM1 (50/10.469%)
petalwidth > 0.8 :
  | class=Iris-virginica <= 0.5 : LM2 (50/14.325%)
  | class=Iris-virginica > 0.5 : LM3 (50/17.598%)

LM num: 1

Linear Regression Model

petallength =
0.4057 + petalwidth
Classifier output:

1. class=iris-virginica > 0.5 : LM3 (50/17.598%)
   LM num: 1
   Linear Regression Model
   petallength =
   0.4957 * petalwidth + 1.343

2. LM num: 2
   Linear Regression Model
   petallength =
   0.4208 * sepallength + 1.2692 * petalwidth + 0.0795

3. LM num: 3
   Linear Regression Model
   petallength =
   0.7501 * sepallength + 0.6105

Number of Rules : 3
Classifier output:

```
0.4208 * sepal\text{length} +
1.2692 * petal\text{width} +
0.0795
```

LM num: 3
Linear Regression Model
petallength =

```
0.7501 * sepal\text{length} +
0.6105
```

Number of Rules : 3
Time taken to build model: 1.31 seconds
Classifier output

```
0.4208 * sepal_length +
1.2692 * petal_width +
0.0795
```

LM num: 3
Linear Regression Model

```
petal_length =
0.7501 * sepal_length +
0.6105
```

Number of Rules : 3

Time taken to build model: 1.31 seconds

--- Evaluation on test split ---

--- Summary ---

Correlation coefficient 0.9889
Mean absolute error 0.1861
Root mean squared error 0.255
Relative absolute error 11.9578 %
Root relative squared error 14.9153 %
Total Number of Instances 51
Explorer: clustering data

- WEKA contains “clusterers” for finding groups of similar instances in a dataset
- Implemented schemes are:
  - k-Means, EM, Cobweb, X-means, FarthestFirst
- Clusters can be visualized and compared to “true” clusters (if given)
- Evaluation based on loglikelihood if clustering scheme produces a probability distribution
Choose Cobweb - A 1.0 - C 0.0028209479177387815

Cluster mode:
- Use training set
- Supplied test set
- Percentage split (% 66)
- Classes to clusters evaluation
- Store clusters for visualization

Ignore attributes

Status: OK
Choose Cobweb - A 1.0 - C 0.0028209479177387815

Cluster mode
- Use training set
- Supplied test set: Set...
- Percentage split: 66%
- Classes to clusters evaluation
  - (Nom) class
- Store clusters for visualization

Ignore attributes
Start Stop

Result list (right-click for options)

Status: OK
Clusterer

Choose: Cobweb - A 1.0 - C 0.0028209479177387815

Cluster mode

- Use training set
- Supplied test set
- Percentage split % 66
- Classes to clusters evaluation
  (Nom) class
- Store clusters for visualization

Ignore attributes

Start  Stop

Result list (right-click for options)

16:05:58 - Cobweb

Clusterer output

=== Run information ===

Scheme: weka.clusterers.Cobweb -A 1.0 -C 0.0028209479177387815
Relation: iris
Instances: 150
Attributes: 5
  sepallength
  sepalwidth
  petallength
  petalwidth
Ignored:
  class
Test mode: Classes to clusters evaluation on training data

=== Clustering model (full training set) ===

Number of merges: 0
Number of splits: 0
Number of clusters: 3

node 0 [150]
  | leaf 1 [96]
node 0 [150]
  | leaf 2 [54]

=== Evaluation on training set ===

Status
OK
Cluster mode:
- Classes to clusters evaluation

Clusterer output:

Number of clusters: 3

Clustered Instances:
1 100 (67%)
2 50 (33%)

Class attribute: class
Classes to Clusters:
1 2 <--- assigned to cluster
0 50 | Iris-setosa
50 0 | Iris-versicolor
50 0 | Iris-virginica

Cluster 1 <--- Iris-versicolor
Cluster 2 <--- Iris-setosa

Incorrectly clustered instances: 50.0 33.3333 %
Clusterer

Choose Cobweb - A 1.0 - C 0.0028209479177387815

Cluster mode
- Use training set
- Supplied test set
- Percentage split % 66
- Classes to clusters evaluation

(Nom) class

Store clusters for visualization

Clusterer output

Number of clusters: 3

node 0 [150]
  | leaf 1 [96]
  | leaf 2 [54]

Clustered Instances
1  100 (67%)
2   50 (33%)

Class attribute: class

Classes to Clusters:
1 2  -- assigned to cluster
0 50  Iris-setosa
50 0  Iris-versicolor
50 0  Iris-virginica

Cluster 1  -- Iris-versicolor
Cluster 2  -- Iris-setosa

Incorrectly clustered instances: 50.0 33.3333 %
Choose Cobweb - A 1.0 - C 0.028209479177387815

Cluster mode
- Use training set
- Supplied test set
- Percentage split
- Classes to cluster
- (Nom) class
- Store clusters for future runs

Start

Result list (right-click for options)
16:05:58 - Cobweb

Clusterer output
Weka Classifier Tree Visualizer: 16:05:58 - Cobweb (iris)

Node 0 (150)

Leaf 1 (96)

Leaf 2 (54)
Clusterer

Choose Cobweb - A 1.0 - C 0.0028209479177387815

Cluster mode

- Use training set
- Supplied test set
- Percentage split % 66
- Classes to clusters evaluation (Nom) class
- Store clusters for visualization

Ignore attributes

Start Stop

Result list (right-click for options)

16:05:58 - Cobweb

View in main window
View in separate window
Save result buffer

Load model
Save model
Re-evaluate model on current test set

Visualize cluster assignments
Visualize tree
Explorer: finding associations

- WEKA contains an implementation of the Apriori algorithm for learning association rules
  - Works only with discrete data
- Can identify statistical dependencies between groups of attributes:
  - milk, butter $\Rightarrow$ bread, eggs (with confidence 0.9 and support 2000)
- Apriori can compute all rules that have a given minimum support and exceed a given confidence
Weka Knowledge Explorer

Associated

Choose: Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0

Start

Stop

Result list (right-click for options)

16:29:37 - Apriori

Associated output

Minimum metric <confidence>: 0.9
Number of cycles performed: 11

Generated sets of large itemsets:

Size of set of large itemsets L(1): 20
Size of set of large itemsets L(2): 17
Size of set of large itemsets L(3): 6
Size of set of large itemsets L(4): 1

Best rules found:

1. adoption-of-the-budget-resolution=y physician-fee-freeze=n 219 ==> Class=democrat
2. adoption-of-the-budget-resolution=y physician-fee-freeze=n aid-to-nicaraguan-contras=y 211 ==> Class=democrat 210
3. physician-fee-freeze=n aid-to-nicaraguan-contras=y 211 ==> Class=democrat 210
4. physician-fee-freeze=n education-spending=n 202 ==> Class=democrat 201 conf: (0.99)
5. physician-fee-freeze=n 247 ==> Class=democrat 245 conf: (0.98)
6. el-salvador-aid=n Class=democrat 200 ==> aid-to-nicaraguan-contras=y 197 conf: (0.9)
7. el-salvador-aid=n 208 ==> aid-to-nicaraguan-contras=y 204 conf: (0.98)
8. adoption-of-the-budget-resolution=y aid-to-nicaraguan-contras=y Class=democrat 200
9. el-salvador-aid=n aid-to-nicaraguan-contras=y Class=democrat 200
10. aid-to-nicaraguan-contras=y Class=democrat 218 ==> physician-fee-freeze=n 210
Explorer: attribute selection

- Panel that can be used to investigate which (subsets of) attributes are the most predictive ones
- Attribute selection methods contain two parts:
  - A search method: best-first, forward selection, random, exhaustive, genetic algorithm, ranking
  - An evaluation method: correlation-based, wrapper, information gain, chi-squared, …
- Very flexible: WEKA allows (almost) arbitrary combinations of these two
Attribute Evaluator
Choose CfsSubsetEval

Search Method
Choose BestFirst - D 1 - N 5

Attribute Selection Mode
○ Use full training set
  ○ Cross-validation Folds 10
    Seed 1

(Nom) Class

Start Stop

Result list (right-click for options)
16:39:40 - BestFirst + CfsSubsetEval

Attribute selection output

duty-free-exports
export-administration-act-south-africa
Class
Evaluation mode: evaluate on all training data

=== Attribute Selection on all input data ===

Search Method:
Best first.
Start set: no attributes
Search direction: forward
Stale search after 5 node expansions
Total number of subsets evaluated: 83
Merit of best subset found: 0.729

Attribute Subset Evaluator (supervised, Class (nominal): 17 Class):
CFS Subset Evaluator

Selected attributes: 4 : 1
physician-fee-freeze

Status
OK
Attribute Evaluator

Choose: CfsSubsetEval

Search Method

Choose: BestFirst -D 1 -N 5

Attribute Selection Mode

- Use full training set
- Cross-validation

(Nom) Class

Start list (right-click for options)

16:39:40 - BestFirst + CfsSubsetEval

Attribute selection output

duty-free-exports
export-administration-act-south-africa
Class
Evaluation mode: evaluate on all training data

=== Attribute Selection on all input data ===

Search Method:
Best first.
Start set: no attributes
Search direction: forward
Stale search after 5 node expansions
Total number of subsets evaluated: 83
Merit of best subset found: 0.729

Attribute Subset Evaluator (supervised, Class (nominal): 17 Class):
CFS Subset Evaluator

Selected attributes: 4 : 1
physician-fee-freeze
Attribute Selection on all input data ===

Search Method:
Best first.
Start set: no attributes
Search direction: forward
Stale search after 5 node expansions
Total number of subsets evaluated: 83
Merit of best subset found: 0.729

Attribute Subset Evaluator (supervised, Class (nominal): 17 Class):
  CFS Subset Evaluator

Selected attributes: 4 : 1
  physician-fee-freeze
Attribute Evaluator

Choose InfoGainAttributeEval

Search Method

- BestFirst
- ForwardSelection
- RaceSearch
- GeneticSearch
- RandomSearch
- ExhaustiveSearch
- Ranker
- RankSearch

Attribute selection output

duty-free-exports
export-administration-act-south-africa
Class

Evaluation mode: evaluate on all training data

Attribute Selection on all input data

Search Method:
Best first.
Start set: no attributes
Search direction: forward
Stale search after 5 node expansions
Total number of subsets evaluated: 83
Merit of best subset found: 0.729

Attribute Subset Evaluator (supervised, Class (nominal): 17 Class):
CFS Subset Evaluator

Selected attributes: 4 : 1
physician-fee-freeze
Attribute Evaluator
Choose InfoGainAttributeEval

Search Method
Choose Ranker -T -1.7976931348623157E308 -N -1

Attribute Selection Mode
- Use full training set
- Cross-validation
  Folds 10
  Seed 1

(Nom) Class

Result list (right-click for options)
16:39:40 - BestFirst + CfsSubsetEval
16:43:05 - Ranker + InfoGainAttributeEval

Information Gain Ranking Filter

Ranked attributes:
0.7078541  4  physician-fee-freeze
0.4185726  3  adoption-of-the-budget-resolution
0.4028397  5  el-salvador-aid
0.34036  12  education-spending
0.3123121  14  crime
0.3095576  8  aid-to-nicaraguan-contras
0.2856444  9  mx-missile
0.2121705  13  superfund-right-to-sue
0.2013666  15  duty-free-exports
0.1902427  7  anti-satellite-test-ban
0.1404643  6  religious-groups-in-schools
0.1211834  1  handicapped-infants
0.1007458  11  synfuels-corporation-cutback
0.0529956  16  export-administration-act-south-africa
0.0490970  10  immigration
0.0049097  16  export-administration-act-south-africa
0.0000117  2  water-project-cost-sharing

Selected attributes: 4,3,5,12,14,8,9,13,15,7,6,1,11,16,10,2 : 16
Explorer: data visualization

- Visualization very useful in practice: e.g. helps to determine difficulty of the learning problem
- WEKA can visualize single attributes (1-d) and pairs of attributes (2-d)
  - To do: rotating 3-d visualizations (Xgobi-style)
- Color-coded class values
- “Jitter” option to deal with nominal attributes (and to detect “hidden” data points)
- “Zoom-in” function
### Attributes

<table>
<thead>
<tr>
<th>No.</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>RI</td>
</tr>
<tr>
<td>2</td>
<td>Na</td>
</tr>
<tr>
<td>3</td>
<td>Mg</td>
</tr>
<tr>
<td>4</td>
<td>Al</td>
</tr>
<tr>
<td>5</td>
<td>Si</td>
</tr>
<tr>
<td>6</td>
<td>K</td>
</tr>
<tr>
<td>7</td>
<td>Ca</td>
</tr>
<tr>
<td>8</td>
<td>Ba</td>
</tr>
<tr>
<td>9</td>
<td>Fe</td>
</tr>
<tr>
<td>10</td>
<td>Type</td>
</tr>
</tbody>
</table>

### Selected Attribute

**Name:** RI  
**Missing:** 0 (0%)  
**Distinct:** 178  
**Unique:** 145 (68%)

### Statistics

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum</td>
<td>1.511</td>
</tr>
<tr>
<td>Maximum</td>
<td>1.534</td>
</tr>
<tr>
<td>Mean</td>
<td>1.518</td>
</tr>
<tr>
<td>StdDev</td>
<td>0.003</td>
</tr>
</tbody>
</table>

### Visualize

**Colour:** Type (Nom)
Weka Knowledge Explorer: Visualizing Glass

Plot: Glass

X: Al (Num)
Y: Ca (Num)
Colour: Type (Nom)
Rectangle

Submit  Clear  Save

Jitter

Class colour

build wind float build wind non-float vehic wind float vehic wind non-float containers tableware headlamps
Weka Knowledge Explorer: Visualizing Glass

X: Al (Num)
Y: Ca (Num)
Colour: Type (Nom)
Rectangle

Plot Glass

Class colour
build wind float  build wind non-float  vehic wind float
vehic wind non-float containers  tableware  headlamps
Performing experiments

- Experimenter makes it easy to compare the performance of different learning schemes
- For classification and regression problems
- Results can be written into file or database
- Evaluation options: cross-validation, learning curve, hold-out
- Can also iterate over different parameter settings
- Significance-testing built in!
10:33:04: Started
13:41:15: Finished
13:41:15: There were 0 errors
10:33:04: Started
13:41:15: Finished
13:41:15: There were 0 errors
Weka Experiment Environment

Configure test
- Row key fields: Select keys...
- Run field: Key_Run
- Column key fields: Select keys...
- Comparison field: Percent_correct
- Significance: 0.05
- Test base: Select base...
- Show std. deviations

Perform test
- Save output

Test output
- Analysing: Percent_correct
- Datasets: 3
- Resultsets: 3
- Confidence: 0.05 (two tailed)
- Date: 9/9/03 1:44 PM

Dataset
- (1) trees.j44 | (2) funct | (3) bayes
- iris
  - (100) 94.73 | 96.4 | 95.53
- vote
  - (100) 96.57 | 94.71 * | 90.02 *
- Glass
  - (100) 67.63 | 66.78 | 49.45 *

Skipped:
- (v/ *) | (0/2/1) (0/1/2)

Key:
- (1) trees.j48.J48 '-C 0.25 -M 2' -2177331683936444444
- (2) functions.neural.NeuralNetwork '-L 0.3 -M 0.2 -N 500 -V 0 -S 0.00001
- (3) bayes.NaiveBayes '' 2029074699749330519
The Knowledge Flow GUI

- New graphical user interface for WEKA
- Java-Beans-based interface for setting up and running machine learning experiments
- Data sources, classifiers, etc. are beans and can be connected graphically
- Data “flows” through components: e.g., “data source” -> “filter” -> “classifier” -> “evaluator”
- Layouts can be saved and loaded again later
Weka KnowledgeFlow Environment

Knowledge Flow Layout

Status
Done.
Weka KnowledgeFlow Environment

Knowledge Flow Layout

ArffLoader
Knowledge Flow Layout

ArffLoader

DataVisualizer

Status
Done.

Log
Knowledge Flow Layout

- Edit
- Delete
- Configure...
- Connections
- dataSet
- instance
- Actions
- Start loading

DataVisualizer

Status
Done.

Log
Knowledge Flow Layout

ArffLoader

dataSet

DataVisualizer

Done.
Can continue this...
Correctly Classified Instances 144 96%
Incorrectly Classified Instances 6 4%
Kappa statistic 0.94
Mean absolute error 0.2311
Root mean squared error 0.288
Relative absolute error 52%
Root relative squared error 58.704%
Total Number of Instances 150
Conclusion: try it yourself!

- WEKA is available at [http://www.cs.waikato.ac.nz/ml/weka](http://www.cs.waikato.ac.nz/ml/weka)
- Also has a list of projects based on WEKA
- WEKA contributors:

  Abdelaziz Mahoui, Alexander K. Seewald, Ashraf M. Kibriya, Bernhard Pfahringer, Brent Martin, Peter Flach, Eibe Frank, Gabi Schmidberger, Ian H. Witten, J. Lindgren, Janice Boughton, Jason Wells, Len Trigg, Lucio de Souza Coelho, Malcolm Ware, Mark Hall, Remco Bouckaert, Richard Kirkby, Shane Butler, Shane Legg, Stuart Inglis, Sylvain Roy, Tony Voyle, Xin Xu, Yong Wang, Zhihai Wang