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
...
Weka GUI Chooser

Waikato Environment for Knowledge Analysis
(c) 1999 – 2003
University of Waikato
New Zealand

Simple CLI
Explorer
Experimenter
KnowledgeFlow
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, …
Welcome to the Weka Knowledge Explorer
Relation: iris  
Instances: 150  
Attributes: 5

Selected attribute
Name: sepal length
Missing: 0 (0%)  
Distinct: 35  
Unique: 9 (6%)

Statistic  |  Value
---|---
Minimum  | 4.3
Maximum  | 7.9
Mean  | 5.843
StdDev  | 0.828

Status OK
The Weka Knowledge Explorer interface is shown, with a focus on the Iris dataset. The dataset contains 150 instances with 5 attributes. The selected attribute is 'class', which has 3 distinct values: Iris-setosa, Iris-versicolor, and Iris-virginica, each occurring 50 times. The visualization shows the distribution of the 'class' attribute, with one color for each class.
Weka Knowledge Explorer

Filter
Choose: None

Current relation
Relation: iris
Instances: 150
Attributes: 5

Attributes
No. | Name
--- | ---
1 | sepal length
2 | sepal width
3 | petal length
4 | petal width
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)

Status
OK
Weka Knowledge Explorer

Current relation
- Relation: iris
- Instances: 150
- Attributes:

<table>
<thead>
<tr>
<th>No.</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>sepal length</td>
</tr>
<tr>
<td>2</td>
<td>sepal width</td>
</tr>
<tr>
<td>3</td>
<td>petal length</td>
</tr>
<tr>
<td>4</td>
<td>petal width</td>
</tr>
<tr>
<td>5</td>
<td>class</td>
</tr>
</tbody>
</table>

Filter
- Choose: Discretize - B 10 - R first-last
- weka.filters.unsupervised.attribute.Discretize

About
An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes.

- attributeIndices: first-last
- bins: 10
- findNumBins: False
- invertSelection: False
- makeBinary: False
- useEqualFrequency: False

Visualize All

Status
OK
### Weka Knowledge Explorer

#### Preprocess

- **Filter**
  - Choose: **Discretize** -B 10 –R first–last
  - Description: An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes.

#### Current relation

- **Relation:** iris
- **Instances:** 150

#### Attributes

<table>
<thead>
<tr>
<th>No.</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>sepal length</td>
</tr>
<tr>
<td>2</td>
<td>sepal width</td>
</tr>
<tr>
<td>3</td>
<td>petal length</td>
</tr>
<tr>
<td>4</td>
<td>petal width</td>
</tr>
<tr>
<td>5</td>
<td>class</td>
</tr>
</tbody>
</table>

#### Discretize Settings

- **attributeIndices:** first–last
- **bins:** 10
- **findNumBins:** False
- **invertSelection:** False
- **makeBinary:** False
- **useEqualFrequency:** True

#### 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, …
Classifier: J48 - C 0.25 - M 2

Test options:
- Cross-validation: Folds 10
- Confidence Factor: 0.25
- Minimum number of instances: 2
- Number of folds: 3
- Reduced error pruning: False
- Save instance data: False
- Subtree raising: True
- Unpruned tree: False
- Use Laplace: False

Status: OK
Classifier
Choose J48 -C 0.25 -M 2

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

(Nom) class

Start  Stop

Result list (right-click for options)

Status
OK
Choose J48 -C 0.25 -M 2

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

Test mode: split 66% train, remainder test

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 Set...
☐ Cross-validation Folds 10
☐ Percentage split % 66

More options...

(Nom) class

Start Stop

Result list (right-click for options)

11:49:05 - trees.j48

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>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

Status

OK
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:

- Iris-setosa: Recall 1, F-Measure 1
- Iris-versicolor: Recall 0.95, F-Measure 0.938
- Iris-virginica: Recall 0.882, F-Measure 0.938

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
Visualize margin curve
Visualize threshold curve
Visualize cost curve
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:

- Iris-setosa: Recall 1, F-Measure 1, Class Iris-setosa
- Iris-versicolor: Recall 1, F-Measure 0.95, Class Iris-versicolor
- Iris-virginica: Recall 0.882, F-Measure 0.938, Class Iris-virginica

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
Visualize margin curve
Visualize threshold curve
Visualize cost curve

Status: OK

Log
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>
</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 ===

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

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</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</td>
<td>1</td>
<td>1</td>
<td>0.882</td>
<td>Iris-versicolor</td>
</tr>
<tr>
<td>Iris-virginica</td>
<td>0.905</td>
<td>0.063</td>
<td>0.905</td>
<td>1</td>
<td>0.95</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
```
Evaluation on test split ===

Summary ===

<table>
<thead>
<tr>
<th>Class</th>
<th>Exactly Classified Instances</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iris-setosa</td>
<td>47</td>
<td>96.0784 %</td>
</tr>
<tr>
<td>Iris-versicolor</td>
<td>2</td>
<td>3.9216 %</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Absolute error</th>
<th>Mean squared error</th>
<th>Relative squared error</th>
<th>Number of Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9408</td>
<td>0.0396</td>
<td>0.1579</td>
<td>8.8979 %</td>
<td>51</td>
</tr>
<tr>
<td>33.4091 %</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Detailed Accuracy By Class ===

<table>
<thead>
<tr>
<th>Class</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</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Iris-versicolor</td>
<td>0.063</td>
<td>0.905</td>
<td>1</td>
<td>0.95</td>
</tr>
<tr>
<td>Iris-virginica</td>
<td>0</td>
<td>1</td>
<td>0.882</td>
<td>0.938</td>
</tr>
</tbody>
</table>

Confusion Matrix ===

<table>
<thead>
<tr>
<th>Iris-virginica</th>
<th>Iris-versicolor</th>
<th>Iris-setosa</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>19</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>15</td>
</tr>
</tbody>
</table>

a = Iris-setosa
b = Iris-versicolor
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 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 ---

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

Status
OK
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

More options...

(Nom) class

Start

Result list (right-click for options)

11:49:05 - trees.j48.j48

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 0.0396
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></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>1</td>
<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>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>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
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>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>0.941</td>
<td>0.031</td>
<td>0.95</td>
<td>1</td>
<td>0.974</td>
</tr>
<tr>
<td>Iris-virginica</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.941</td>
<td>0.97</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
```
Choose NeuralNetwork -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a -G -R

Correctly Classified Instances 50 98.0392 %
Incorrectly Classified Instances 1 1.9608 %

Accuracy: 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>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>0.941</td>
<td>0.031</td>
<td>0.95</td>
<td>1</td>
<td>0.974</td>
<td>Iris-versicolor</td>
</tr>
<tr>
<td>Iris-virginica</td>
<td>0</td>
<td>1</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 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>Class</th>
<th>Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Iris-setosa</td>
<td>1</td>
<td>0.031</td>
<td>0.95</td>
<td>1</td>
<td>0.974</td>
<td>1</td>
</tr>
<tr>
<td>Iris-versicolor</td>
<td>0.941</td>
<td>0.047</td>
<td>0.941</td>
<td>0.97</td>
<td>0.97</td>
<td>0</td>
</tr>
</tbody>
</table>

=== Confusion Matrix ===

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

Choose NaiveBayes

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 ---

TP Rate FP Rate Precision Recall F-Measure Class
1 0 1 1 1 1
1 0.031 0.95 1 0.974 Iris-versicolor
0.941 0 1 0.941 0.97 Iris-virginica

--- 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
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>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>0.063</td>
<td>0.9</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Iris-versicolor</td>
<td>0.947</td>
<td>0.029</td>
<td>0.938</td>
<td>0.947</td>
<td>0.923</td>
</tr>
<tr>
<td>Iris-virginica</td>
<td>0.882</td>
<td>0.029</td>
<td>0.938</td>
<td>0.882</td>
<td>0.909</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
```
== Evaluation on test split ==
== Summary ==

correctly Classified Instances  48  94.1176 %
incorrectly Classified Instances  3  5.8824 %
apparent statistic  0.9113
mean absolute error  0.0447
oot mean squared error  0.1722
elevation absolute error  10.0365 %
oot relative squared error  36.4196 %
Total Number of Instances  51

== Detailed Accuracy By Class ==

P Rate  FP Rate  Precision  Recall  F-Measure  Class
1  0.947  0.063  1  1  1  Iris-setosa
0.947  Iris-setosa
0.882  0.029  0.938  0.947  0.923  Iris-versicolor
0.923  Iris-versicolor
0.882  0.029  0.938  0.947  0.923  Iris-virginica
0.909  Iris-virginica

== Confusion Matrix ==

a  b  c  <-- classified as
15  0  0  a = Iris-setosa
18  1  0  b = Iris-versicolor
15  2  1  c = Iris-virginica
Weka Knowledge Explorer

Classifier: UserClassifier

Test options:
- Use training set
- Supplied test set
- Cross-validation
- Percentage split

(X) class

More options...

Start

Result list (right-click for details):
- 11:49:05 - trees.j48
- 14:34:28 - functions
- 14:48:05 - bayes.Naive
- 15:26:57 - trees.Use

Log: Building model on training data...

Plot: iris

Class colour:
- Iris-setosa
- Iris-versicolor
- Iris-virginica

X: petal length (Num)  Y: petal width (Num)
Colour: class (Nom)  Polyline

Jitter

Submit  Clear  Save
Split on petallength AND petalwidth

True
[ Iris-versicolor, 1.0]
[ Iris-virginica, 48.0]

False
[ Iris-setosa, 50.0]
[ Iris-versicolor, 49.0]
[ Iris-virginica, 2.0]
<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 ---

```
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: UserClassifier

Test options

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

Classifiers:

- (Num) sepal length
- (Num) sepal width
- (Num) petal length
- (Num) petal width
- (Nom) class

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>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>0.063</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Iris-versicolor</td>
<td>1</td>
<td>0.882</td>
<td>0.905</td>
<td>0.882</td>
<td>0.938</td>
</tr>
<tr>
<td>Iris-virginica</td>
<td>0.882</td>
<td>0</td>
<td>1.000</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
```
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>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>0</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

a b c <-- classified as
=== Run information ===

Scheme: weka.classifiers.trees.m5.M5P -M 4.0
Relation: iris
Instances: 150
Attributes: 5
  sepallength
  sepalwidth
  petallength
  petalwidth
  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 =
  0.4052 + petalwidth
Classifier

Choose M5P - M 4.0

Test options

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

Classifier output

1. class=iris-virginica > 0.5 : LM3 (50/17.598%)

LM num: 1
Linear Regression Model

petallength =
0.4957 \times petalwidth + 1.343

LM num: 2
Linear Regression Model

petallength =
0.4208 \times sepalwidth + 1.2692 \times petalwidth + 0.0795

LM num: 3
Linear Regression Model

petallength =
0.7501 \times sepalwidth + 0.6105

Number of Rules : 3

Status

OK
Classifier output

Lambda: 1.0

Linear Regression Model

petallength =

0.7501 * sepalwidth + 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
Classifier: M5P

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

Classifiers:
- Linear Regression Model
- Rule-based Model

Classifier output:

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.4208</td>
<td>sepallength</td>
</tr>
<tr>
<td>0.2692</td>
<td>petalwidth</td>
</tr>
<tr>
<td>0.0795</td>
<td></td>
</tr>
</tbody>
</table>

Linear Regression Model:

petallength =

0.7501 * sepallength +
0.6105

Number of Rules: 3

Time taken to build model: 1.31 seconds

Evaluation on test split:

- 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
Instance info:

Instance number: 31.0
sepal length: 6.9
sepal width: 3.1
predicted petal length: 5.892812341943582
petal length: 5.1
petal width: 2.3
class: Iris-virginica
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 EM -l 100 -N -1 -S 100 -M 1.0E-6

Use training set

Supplied test set

Percentage split

Classes to clusters evaluation

Store clusters for visualization

Ignore attributes

Start

Stop

Clusterer output

Result list (right-click for options)

Status

OK
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

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
- 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
  sepal length
  sepal width
  petal length
  petal width

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 ===
Cobweb - A 1.0 - C 0.0028209479177387815

Scheme: weka.clusterers.Cobweb -A 1.0 -C 0.002820947917
Relation: iris
Instances: 150
Attributes: 5
  sepal length
  sepal width
  petal length
  petal width
Ignored: class
Test mode: Classes to clusters evaluation on training data

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
Clusterer

Choose Cobweb-A 1.0 -C 0.0028209479177387815

Cluster mode

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

Clusterer output

Number of clusters: 3

node 0 [150]
  | leaf 1 [96]
  node 0 [150]
  | 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 %

Status

OK
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

Incorrectly clustered instances: 50.0 33.3333 %
Cluster mode:
- Use training set
- Supplied test set
- Percentage split
- Classes to clusters evaluation
  (Nom) class
- Store clusters for visualization

Clusterer output:

--- Run information ---
Scheme: weka.clusterers.Cobweb -A 1.0 -C 0.0028209479177387815
Relation: iris
Instances: 150
Attributes: 5
  - sepal length
  - sepal width
  - petal length
  - petal width
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
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
<table>
<thead>
<tr>
<th>No.</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>handicapped-infants</td>
</tr>
<tr>
<td>2</td>
<td>water-project-cost-sharing</td>
</tr>
<tr>
<td>3</td>
<td>adoption-of-the-budget-resolution</td>
</tr>
<tr>
<td>4</td>
<td>physician-fee-freeze</td>
</tr>
<tr>
<td>5</td>
<td>el-salvador-aid</td>
</tr>
<tr>
<td>6</td>
<td>religious-groups-in-schools</td>
</tr>
<tr>
<td>7</td>
<td>anti-satellite-test-ban</td>
</tr>
<tr>
<td>8</td>
<td>aid-to-nicaraguan-contras</td>
</tr>
<tr>
<td>9</td>
<td>mx-missile</td>
</tr>
<tr>
<td>10</td>
<td>immigration</td>
</tr>
<tr>
<td>11</td>
<td>synfuels-corporation-cutback</td>
</tr>
<tr>
<td>12</td>
<td>education-spending</td>
</tr>
<tr>
<td>13</td>
<td>superfund-right-to-sue</td>
</tr>
<tr>
<td>14</td>
<td>crime</td>
</tr>
<tr>
<td>15</td>
<td>duty-free-exports</td>
</tr>
<tr>
<td>16</td>
<td>export-administration-act-south-africa</td>
</tr>
<tr>
<td>17</td>
<td>Class</td>
</tr>
</tbody>
</table>

Selected attribute:
- Name: handicapped-infants
- Type: Nominal
- Missing: 12 (3%)
- Distinct: 2
- Unique: 0 (0%)

<table>
<thead>
<tr>
<th>Label</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>236</td>
</tr>
<tr>
<td>y</td>
<td>187</td>
</tr>
</tbody>
</table>
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
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 204 => Class=democrat 197
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: CfsSubsetEval

Search Method: BestFirst -D 1 -N 5

Attribute Selection Mode: Use full training set

(Nom) Class

Start

Result list (right-click for options):

16:39:40 - BestFirst + CfsSubsetEval

Attribute selection output:

duty-free-exports
eexport-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

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

- weka
  - attributeSelection
    - CfsSubsetEval
    - ClassifierSubsetEval
    - WrapperSubsetEval
    - ConsistencySubsetEval
    - ReliefFAttributeEval
    - InfoGainAttributeEval
    - GainRatioAttributeEval
    - SymmetricalUncertAttributeEval
    - OneRAttributeEval
    - ChiSquaredAttributeEval
    - PrincipalComponents
    - SVMAttributeEval

Search selection output

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

Search 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

weka

attributeSelection

BestFirst
ForwardSelection
RaceSearch
GeneticSearch
RandomSearch
ExhaustiveSearch
Ranker
RankSearch

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

Attribute selection output
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
- 0.0049097 10 immigration
- 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

Status
OK
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
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
Conclusion: try it yourself!

- WEKA is available at
  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