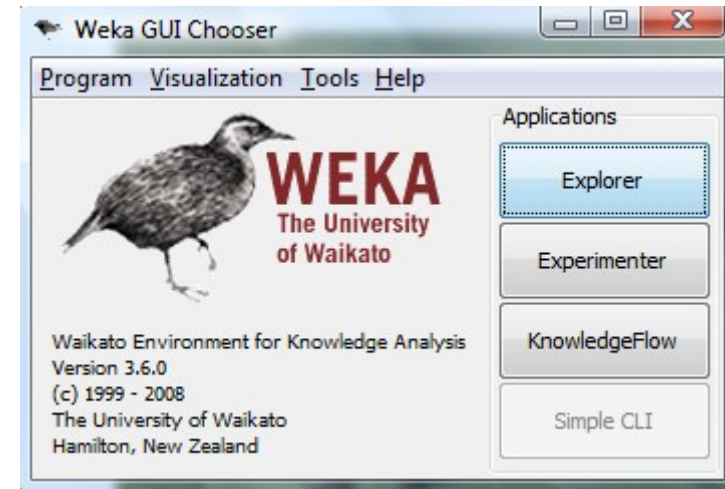


Introduction to Weka



Overview

- What is Weka?
- Where to find Weka?
- Command Line Vs GUI
- Datasets in Weka
- ARFF Files
- Classifiers in Weka
- Filters

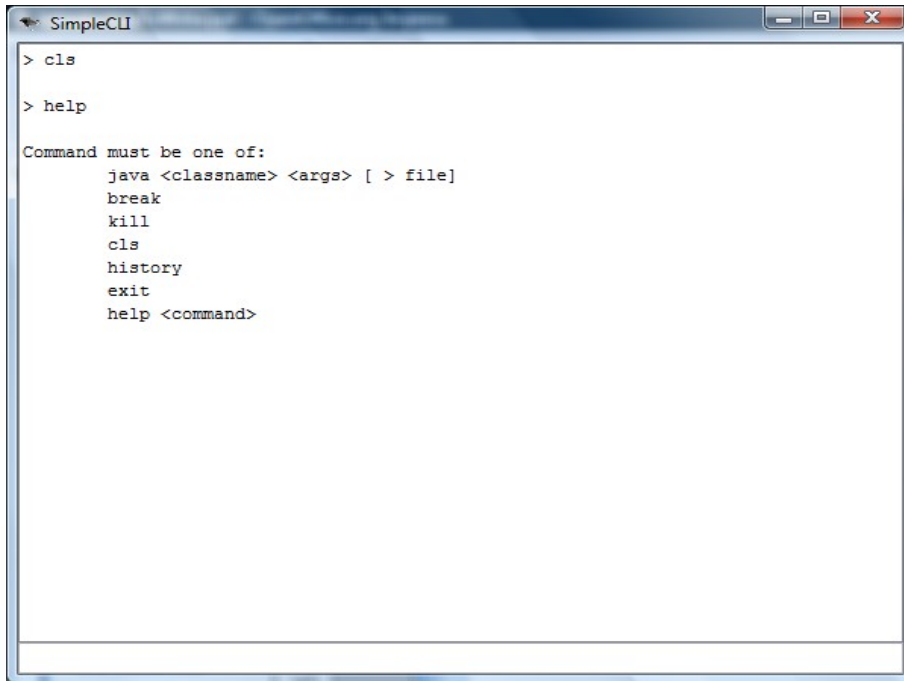
What is Weka?

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

Where to find Weka

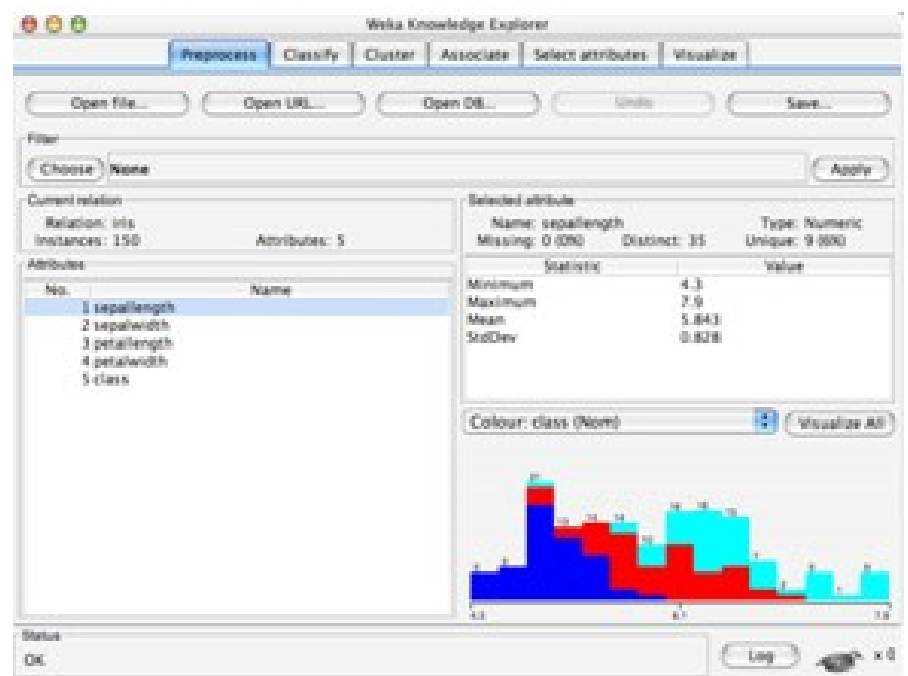
- Weka website (Latest version 3.6):
 - <http://www.cs.waikato.ac.nz/ml/weka/>
- Weka Manual:
 - <http://transact.dl.sourceforge.net/sourceforge/weka/WekaManual-3.6.0.pdf>

CLI Vs GUI



```
> cls
> help
Command must be one of:
  java <classname> <args> [ > file]
  break
  kill
  cls
  history
  exit
  help <command>
```

- Recommended for in-depth usage
- Offers some functionality not available via the GUI



- Explorer
- Experimenter
- Knowledge Flow

Datasets in Weka

- Each entry in a dataset is an instance of the java class:
 - `weka.core.Instance`
- Each instance consists of a number of attributes

Attributes

- *Nominal*: one of a predefined list of values
 - e.g. red, green, blue
- *Numeric*: A real or integer number
- *String*: Enclosed in “double quotes”
- *Date*
- *Relational*

ARFF Files

- The external representation of an Instances class
- Consists of:
 - A header: Describes the attribute types
 - Data section: Comma separated list of data

ARFF File Example

```
% This is a toy example, the UCI weather dataset.  
% Any relation to real weather is purely coincidental
```

```
@relation weather
```

Dataset name

```
@attribute outlook {sunny, overcast, rainy}
```

```
@attribute temperature real
```

```
@attribute humidity real
```

```
@attribute windy {TRUE, FALSE}
```

```
@attribute play {yes, no}
```

Attributes

```
@data
```

```
sunny, 85, 85, FALSE, no
```

```
sunny, 80, 90, TRUE, no
```

```
overcast, 83, 86, FALSE, yes
```

```
rainy, 70, 96, FALSE, yes
```

```
rainy, 68, 80, FALSE, yes
```

```
rainy, 65, 70, TRUE, no
```

```
overcast, 64, 65, TRUE, yes
```

```
sunny, 72, 95, FALSE, no
```

```
sunny, 69, 70, FALSE, yes
```

```
rainy, 75, 80, FALSE, yes
```

```
sunny, 75, 70, TRUE, yes
```

```
overcast, 72, 90, TRUE, yes
```

```
overcast, 81, 75, FALSE, yes
```

```
rainy, 71, 91, TRUE, no
```

Target / Class variable

Data Values

Comment

Assignment ARFF Files

- Credit-g
 - Heart-c
 - Hepatitis
 - Vowel
 - Zoo
-
- <http://www.cs.auckland.ac.nz/~pat/weka/>

ARFF Files

- Basic statistics and validation by running:
 - `java weka.core.Instances data/soybean.arff`

Classifiers in Weka

- Learning algorithms in Weka are derived from the abstract class:
 - `weka.classifiers.Classifier`
- Simple classifier: ZeroR
 - Just determines the most common class
 - Or the median (in the case of numeric values)
 - Tests how well the class can be predicted without considering other attributes
 - Can be used as a Lower Bound on Performance.

Classifiers in Weka

- Simple Classifier Example
 - `java weka.classifiers.rules.ZeroR -t data/weather.arff`
 - `java weka.classifiers.trees.J48 -t data/weather.arff`
- Help Command
 - `java weka.classifiers.trees.J48 -h`

Classifiers in Weka

- **Soybean.arff** split into train and test set

- Soybean-train.arff
- Soybean-test.arff

- Input command:

- `java weka.classifiers.trees.J48 -t soybean-train.arff -T soybean-test.arff -i`

Training data



Test data



Provides more detailed output



Soybean Results

=== Error on test data ===

Correctly Classified Instances	151	88.3041 %
Incorrectly Classified Instances	20	11.6959 %
Kappa statistic	0.8719	
Mean absolute error	0.0146	
Root mean squared error	0.0909	
Relative absolute error	15.157 %	
Root relative squared error	41.5116 %	
Total Number of Instances	171	

Soybean Results (cont...)

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.6	0.012	0.6	0.6	0.6	0.992	diaporthe-stem-canker
1	0	1	1	1	1	charcoal-rot
1	0	1	1	1	1	rhizoctonia-root-rot
1	0.007	0.957	1	0.978	0.995	phytophthora-rot
1	0	1	1	1	1	brown-stem-rot
1	0	1	1	1	1	powdery-mildew
1	0	1	1	1	1	downy-mildew
0.913	0.007	0.955	0.913	0.933	0.999	brown-spot
1	0	1	1	1	1	bacterial-blight
1	0	1	1	1	1	bacterial-pustule
1	0	1	1	1	1	purple-seed-stain
0.727	0.013	0.8	0.727	0.762	0.861	anthracnose
1	0.012	0.714	1	0.833	0.999	phyllosticta-leaf-spot
0.739	0.02	0.85	0.739	0.791	0.991	alternaria-leaf-spot
0.826	0.041	0.76	0.826	0.792	0.988	frog-eye-leaf-spot
1	0	1	1	1	1	diaporthe-pod-&-stem-blight
1	0	1	1	1	1	cyst-nematode
0.25	0	1	0.25	0.4	0.996	2-4-d-injury
1	0.018	0.4	1	0.571	1	herbicide-injury
0.883	0.012	0.896	0.883	0.881	0.987	Weighted Avg.

Soybean Results (cont...)

- True Positive (*TP*)
 - Proportion classified as class x / Actual total in class x
 - Equivalent to Recall
- False Positive (*FP*)
 - Proportion incorrectly classified as class x / Actual total of all classes, except x

Soybean Results (cont...)

- Precision:
 - Proportion of the examples which truly have class x / Total classified as class x
- F-measure:
 - $2 * \text{Precision} * \text{Recall} / (\text{Precision} + \text{Recall})$
 - i.e. A combined measure for precision and recall

Soybean Results (cont...)

=== Confusion Matrix ===

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s
a	3	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0
b	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
c	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d	0	0	0	22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
e	0	0	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0
f	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0
g	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0
h	0	0	0	0	0	0	0	21	0	0	0	0	2	0	0	0	0	0	0
i	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0
j	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0
k	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0
l	2	0	0	1	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0
m	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0
n	0	0	0	0	0	0	0	0	0	0	0	0	0	17	6	0	0	0	0
o	0	0	0	0	0	0	0	1	0	0	0	0	0	2	19	0	0	0	0
p	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0
q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0
r	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3
s	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Total Actual h

<-- classified as
a = diaporthe-stem-canker
b = charcoal-rot
c = rhizoctonia-root-rot
d = phytophthora-rot
e = brown-stem-rot
f = powdery-mildew
g = downy-mildew
h = brown-spot
i = bacterial-blight
j = bacterial-pustule
k = purple-seed-stain
l = anthracnose
m = phyllosticta-leaf-spot
n = alternaria-leaf-spot
o = frog-eye-leaf-spot
p = diaporthe-pod-&-stem-blight
q = cyst-nematode
r = 2-4-d-injury
s = herbicide-injury

Total Classified as h

Total Correct

Filters

- weka.filters package
- Transform datasets
- Support for data preprocessing
 - e.g. Removing/Adding Attributes
 - e.g. Discretize numeric attributes into nominal ones
- More info in Weka Manual p. 15 & 16.

More Classifiers

- `trees.J48` A clone of the C4.5 decision tree learner
- `bayes.NaiveBayes` A Naive Bayesian learner. `-K` switches on kernel density estimation for numerical attributes which often improves performance.
- `meta.ClassificationViaRegression -W functions.LinearRegression` Multi-response linear regression.
- `functions.Logistic` Logistic Regression.
- `functions.SMO` Support Vector Machine (linear, polynomial and RBF kernel) with Sequential Minimal Optimization Algorithm due to [3]. Defaults to SVM with linear kernel, `-E 5 -C 10` gives an SVM with polynomial kernel of degree 5 and lambda of 10.
- `lazy.KStar` Instance-Based learner. `-E` sets the blend entropy automatically, which is usually preferable.
- `lazy.IBk` Instance-Based learner with fixed neighborhood. `-K` sets the number of neighbors to use. `IB1` is equivalent to `IBk -K 1`
- `rules.JRip` A clone of the RIPPER rule learner.

Explorer

- Preprocess
- Classify
- Cluster
- Associate
- Select attributes
- Visualize

Preprocess

- Load Data
- Preprocess Data
- Analyse Attributes

Weka Explorer

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Open file...

Open URL...

Open DB...

Generate

Filter

Choose

None

Current relation

Relation: weather

Instances: 14

Attributes: 5

Attributes

All

None

Invert

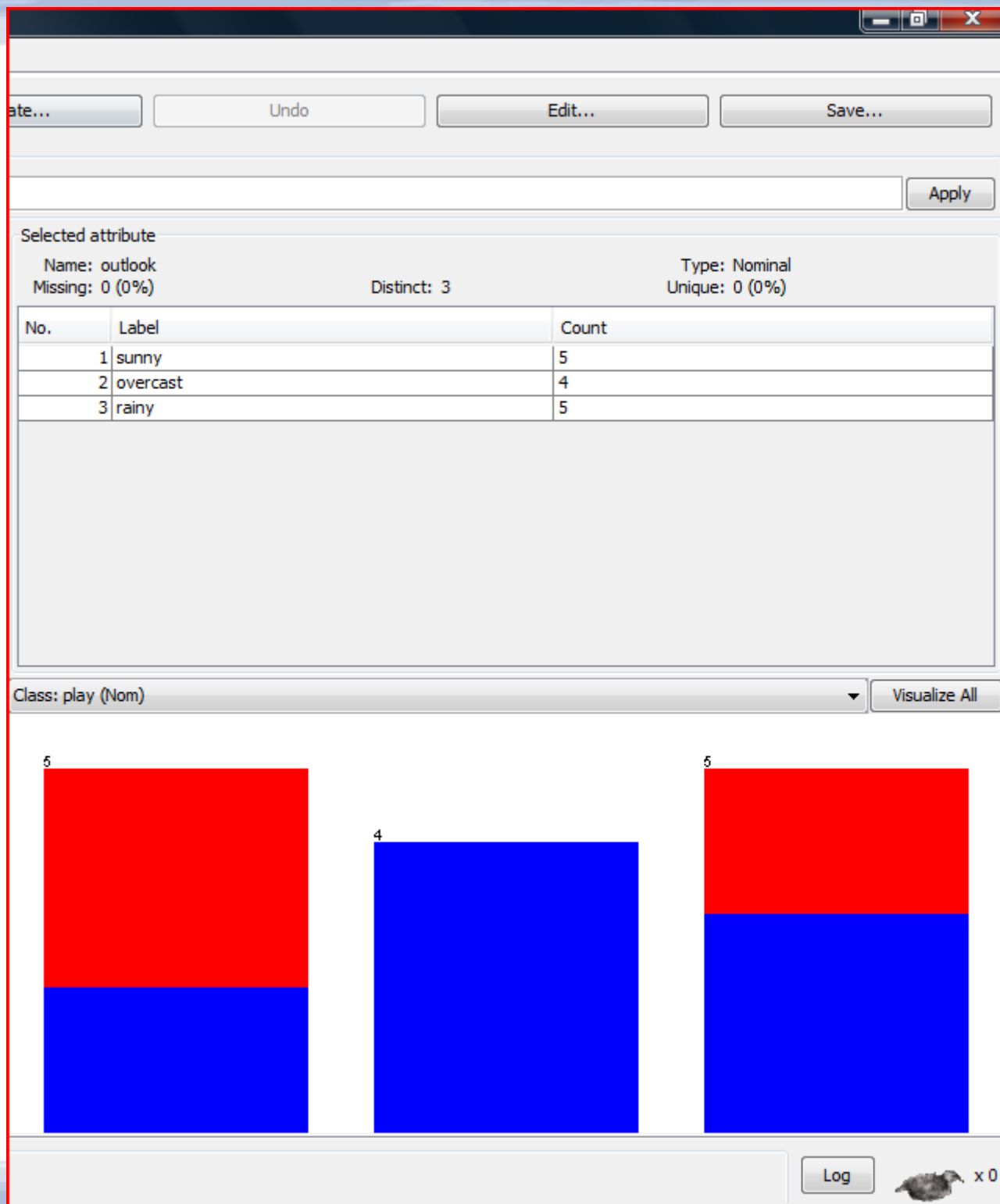
Pattern

No.	Name
1	<input checked="" type="checkbox"/> outlook
2	<input type="checkbox"/> temperature
3	<input type="checkbox"/> humidity
4	<input type="checkbox"/> windy
5	<input type="checkbox"/> play

Remove

Status

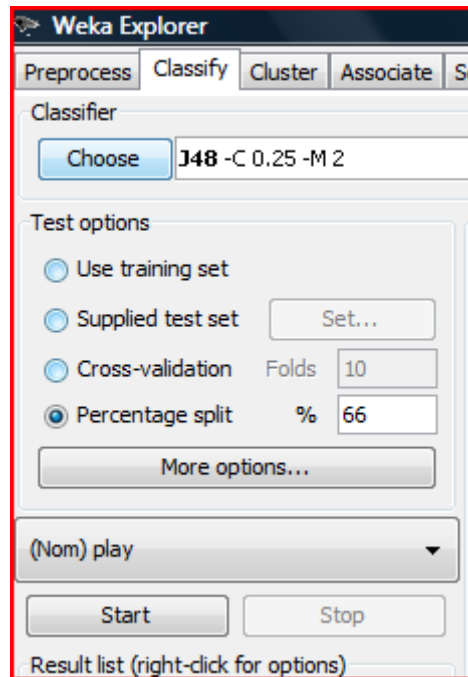
OK

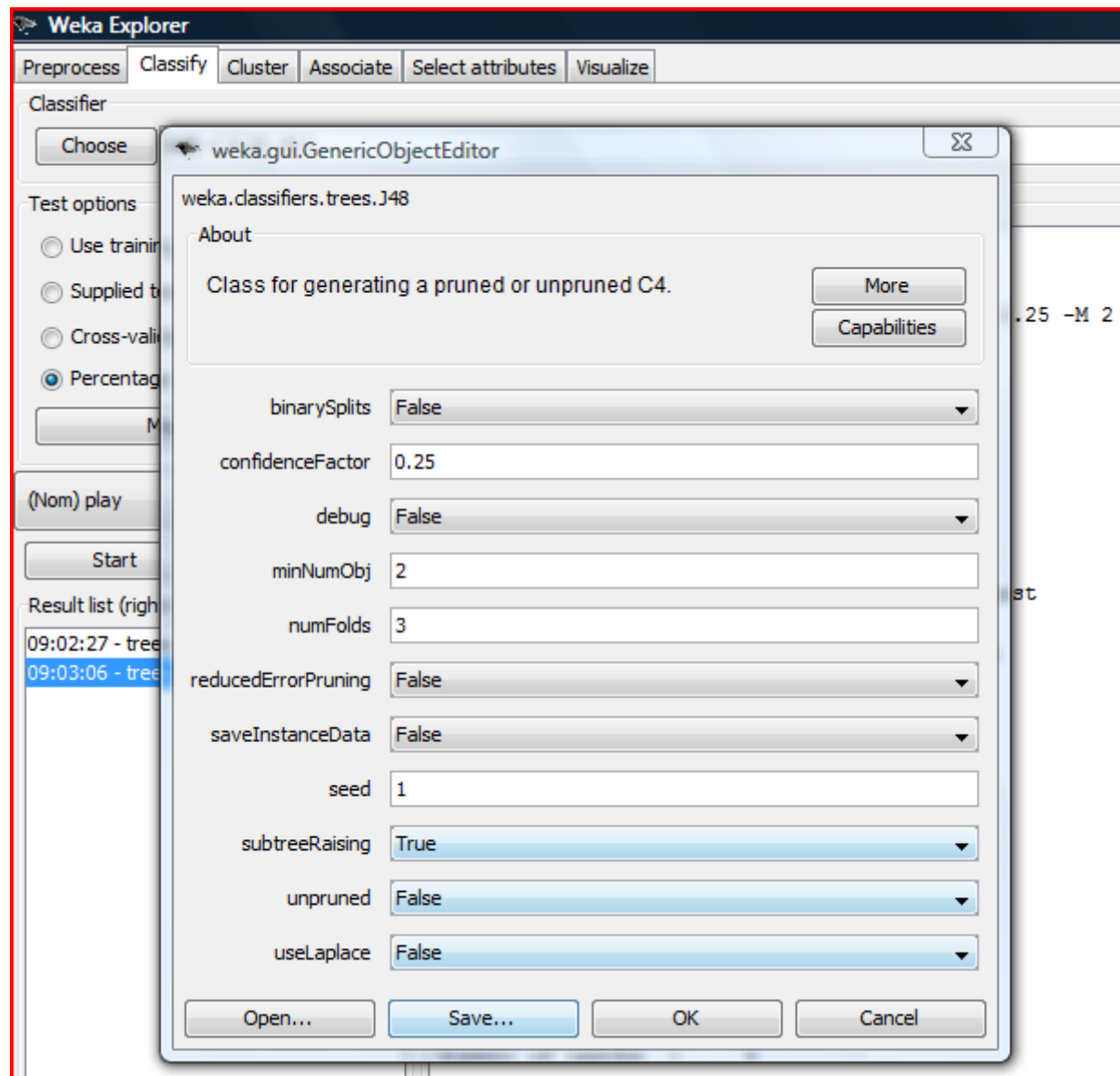


Classify

- Select Test Options e.g:
 - Use Training Set
 - % Split,
 - Cross Validation...
- Run classifiers
- View results

Classify





Classifier output

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: weather
Instances: 14
Attributes: 5
outlook
temperature
humidity
windy
play
Test mode: split 66.0% train, remainder test

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

J48 pruned tree

outlook = sunny
| humidity <= 75: yes (2.0)
| humidity > 75: no (3.0)
outlook = overcast: yes (4.0)
outlook = rainy
| windy = TRUE: no (2.0)
| windy = FALSE: yes (3.0)

Number of Leaves : 5

Size of the tree : 8

Time taken to build model: 0 seconds

← Results

Weka Explorer

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

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) play

Start Stop

Result list (right-click for options)

09:02:27 - trees.J48
09:03:06 - trees.J48

View in main window
View in separate window
Save result buffer
Delete result buffer
Load model
Save model
Re-evaluate model on current test set
Visualize classifier errors
Visualize tree
Visualize margin curve

Classifier output

=== Run information ===

Scheme: weka.classifiers.trees.J48
Relation: weather
Instances: 14
Attributes: 5
outlook
temperature
humidity
windy
play

Test mode: split 66.0% train, 34.0% test

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

Tree View

```
graph TD
    outlook((outlook)) -- "= sunny" --> humidity((humidity))
    outlook -- "= overcast" --> yes4[yes (4.0)]
    outlook -- "= rainy" --> windy((windy))
    humidity -- "<= 75" --> yes2[yes (2.0)]
    humidity -- "> 75" --> no3[no (3.0)]
    windy -- "= TRUE" --> no2[no (2.0)]
    windy -- "= FALSE" --> yes3[yes (3.0)]
```

Experimenter

- Allows users to create, run, modify and analyse experiments in a more convenient manner than when processing individually.
 - Setup
 - Run
 - Analyse

Experimenter: Setup

- Simple/Advanced
- Results Destinations
 - ARFF
 - CSV
 - JDBC Database

Weka Experiment Environment

Setup Run Analyse

Experiment Configuration Mode:

Open...

Results Destination

CSV file Filename: Experiments1.csv

Experiment Type

Cross-validation

Number of folds: 10

☒ Classification ☐ Regression

Datasets

Add new... Edit selected... Delete selected

☐ Use relative paths

C:\Program Files\Weka-3-6\data\iris.arff

Up Down

10-fold
Cross
Validation

Datasets

☒ Simple ☐ Advanced

Save... New

Browse...

Iteration Control

Number of repetitions: 1

☒ Data sets first
☐ Algorithms first

Algorithms

Add new... Edit selected... Delete selected

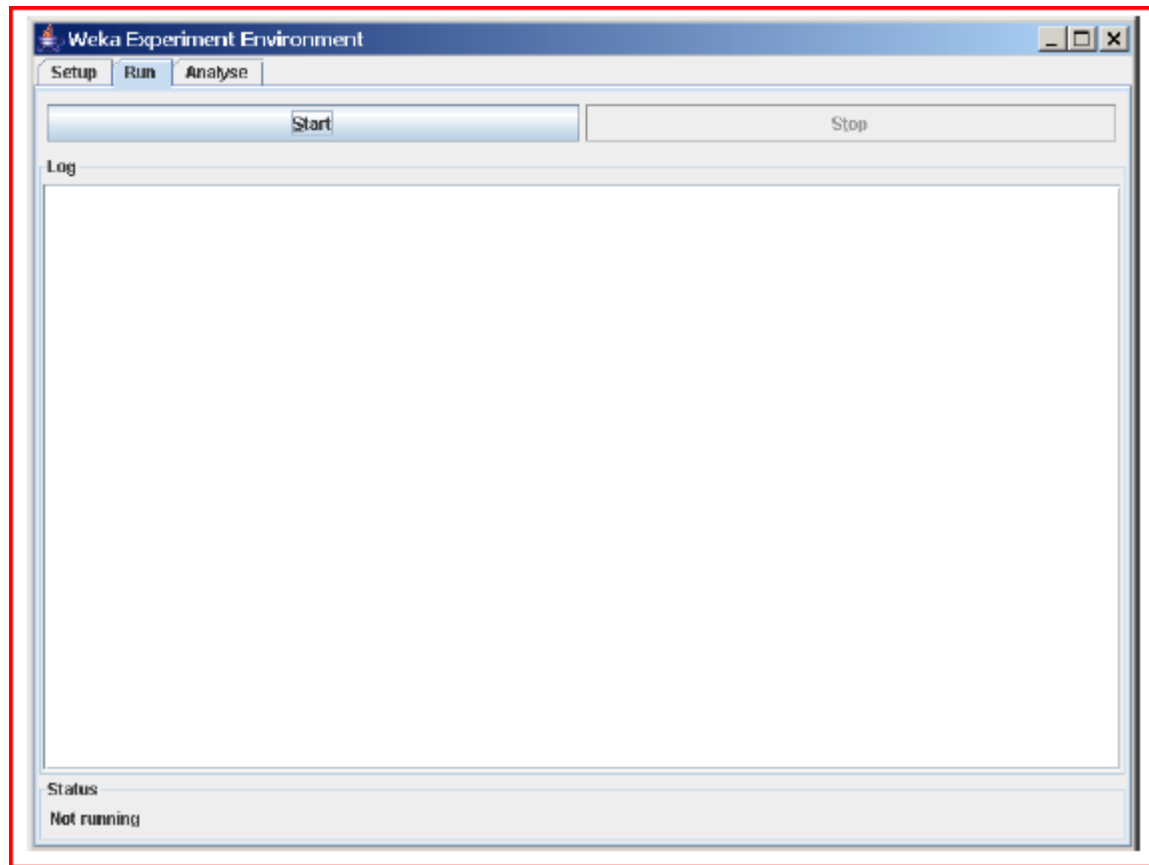
ZeroR
J48 -C 0.25 -M 2

Load options... Save options... Up Down

Num of
runs

Classifiers

Run Simple Experiment



Results

	A	B	C	D	E	F	G	H	I	N
1	Key_Dataset	Key_Run	Key_Fold	Key_Scheme	Key_Scheme_options	Key_Scheme_version_ID	Date_time	Number_of_training_instances	Number_of_testing_instances	
2	iris	1	1	weka.classifiers.rules.ZeroR	'	4.81E+016	2.01E+007	135	15	
3	iris	1	2	weka.classifiers.rules.ZeroR	'	4.81E+016	2.01E+007	135	15	
4	iris	1	3	weka.classifiers.rules.ZeroR	'	4.81E+016	2.01E+007	135	15	
5	iris	1	4	weka.classifiers.rules.ZeroR	'	4.81E+016	2.01E+007	135	15	
6	iris	1	5	weka.classifiers.rules.ZeroR	'	4.81E+016	2.01E+007	135	15	
7	iris	1	6	weka.classifiers.rules.ZeroR	'	4.81E+016	2.01E+007	135	15	
8	iris	1	7	weka.classifiers.rules.ZeroR	'	4.81E+016	2.01E+007	135	15	
9	iris	1	8	weka.classifiers.rules.ZeroR	'	4.81E+016	2.01E+007	135	15	
10	iris	1	9	weka.classifiers.rules.ZeroR	'	4.81E+016	2.01E+007	135	15	
11	iris	1	10	weka.classifiers.rules.ZeroR	'	4.81E+016	2.01E+007	135	15	
12	iris	1	1	weka.classifiers.trees.J48	-C 0.25 -M 2'	-2.18E+017	2.01E+007	135	15	
13	iris	1	2	weka.classifiers.trees.J48	-C 0.25 -M 2'	-2.18E+017	2.01E+007	135	15	
14	iris	1	3	weka.classifiers.trees.J48	-C 0.25 -M 2'	-2.18E+017	2.01E+007	135	15	
15	iris	1	4	weka.classifiers.trees.J48	-C 0.25 -M 2'	-2.18E+017	2.01E+007	135	15	
16	iris	1	5	weka.classifiers.trees.J48	-C 0.25 -M 2'	-2.18E+017	2.01E+007	135	15	
17	iris	1	6	weka.classifiers.trees.J48	-C 0.25 -M 2'	-2.18E+017	2.01E+007	135	15	
18	iris	1	7	weka.classifiers.trees.J48	-C 0.25 -M 2'	-2.18E+017	2.01E+007	135	15	
19	iris	1	8	weka.classifiers.trees.J48	-C 0.25 -M 2'	-2.18E+017	2.01E+007	135	15	
20	iris	1	9	weka.classifiers.trees.J48	-C 0.25 -M 2'	-2.18E+017	2.01E+007	135	15	
21	iris	1	10	weka.classifiers.trees.J48	-C 0.25 -M 2'	-2.18E+017	2.01E+007	135	15	

Advanced Example

Weka Experiment Environment

Setup | Run | Analyse

Experiment Configuration Mode: ☐ Simple ☒ Advanced

Open... Save... New

Destination
Choose CSVResultListener -O Tut1Exp1.csv

Result generator
Choose CrossValidationResultProducer -X 10 -O splitEvaluatorOut.zip -W weka.experiment.ClassifierSplitEvaluator -- -W weka.classifiers.rules.ZeroR -I 0 -C 1 --

Runs
From: 1 To: 10

Distribute experiment
☐ Hosts
☒ By data set ☐ By run

Generator properties
Enabled Select property...

Iteration control
☒ Data sets first ☐ Custom generator first

Datasets
Add new... Edit selected... Delete selected
☐ Use relative paths

C:\Program Files\Weka-3-6\data\iris.arff

Up Down

Choose J48 -C 0.25 -M 2 Add

ZeroR
OneR -B 6
J48 -C 0.25 -M 2

Delete Edit Up Down

Multiple Classifiers

Advanced Example

Weka Experiment Environment

Setup Run Analyse

Source

Got 300 results

Configure test

Testing with: Paired T-Tester (correc... ▼

Row: Select

Column: Select

Comparison field: Percent_correct ▼

Significance: 0.05

Sorting (asc.) by: <default> ▼

Test base: Select

Displayed Columns: Select

Show std. deviations: ☐

Output Format: Select

Perform test Save output

Result list

12:12:57 - Available resultsets

12:12:59 - Percent_correct - rules.ZeroR " 4.80555414

Test output

Tester: weka.experiment.PairedCorrectedTTester

Analysing: Percent_correct

Datasets: 1

Resultsets: 3

Confidence: 0.05 (two tailed)

Sorted by: -

Date: 28/04/09 12:12 PM

Dataset	(1) rules.ZeroR	(2) rules	(3) trees
iris	(100) 33.33	93.53 v	94.73 v

(v/ /*) | (1/0/0) (1/0/0)

Key:

(1) rules.ZeroR " 4.8055541465867952E16

(2) rules.OneR "-B 6' -2.4594270021478615E18

(3) trees.J48 "-C 0.25 -M 2' -2.17733168393644448E17

