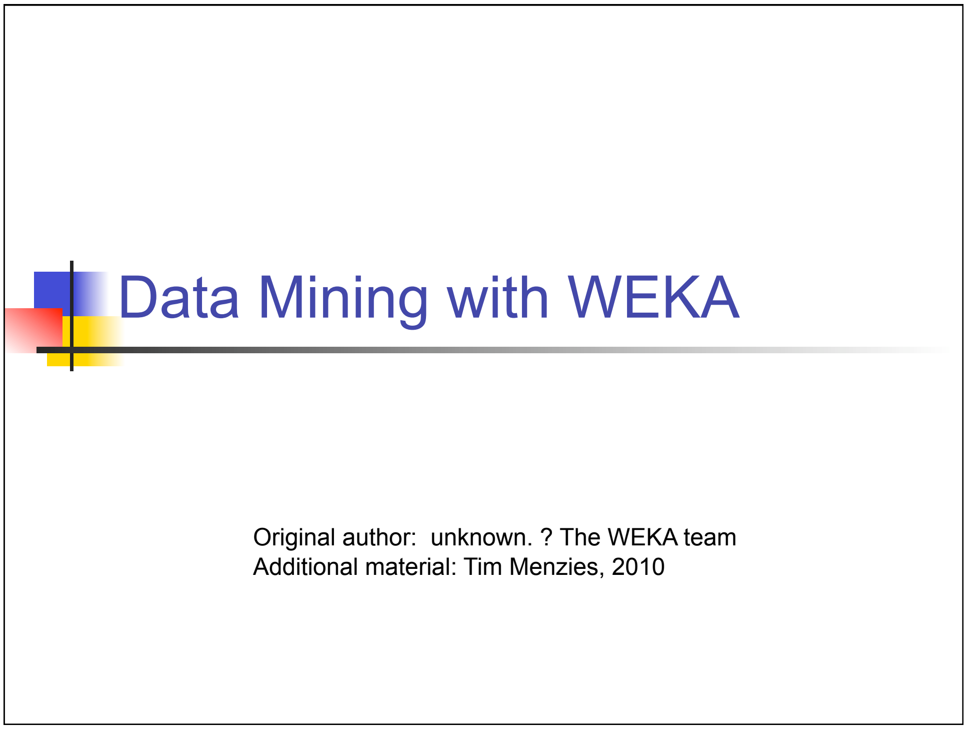
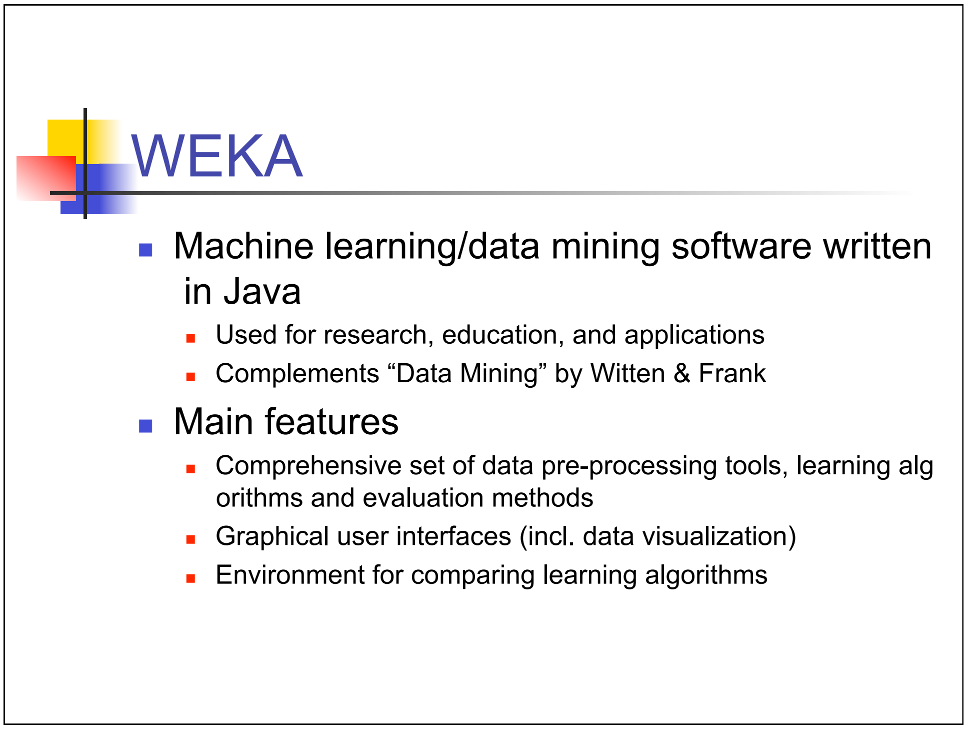


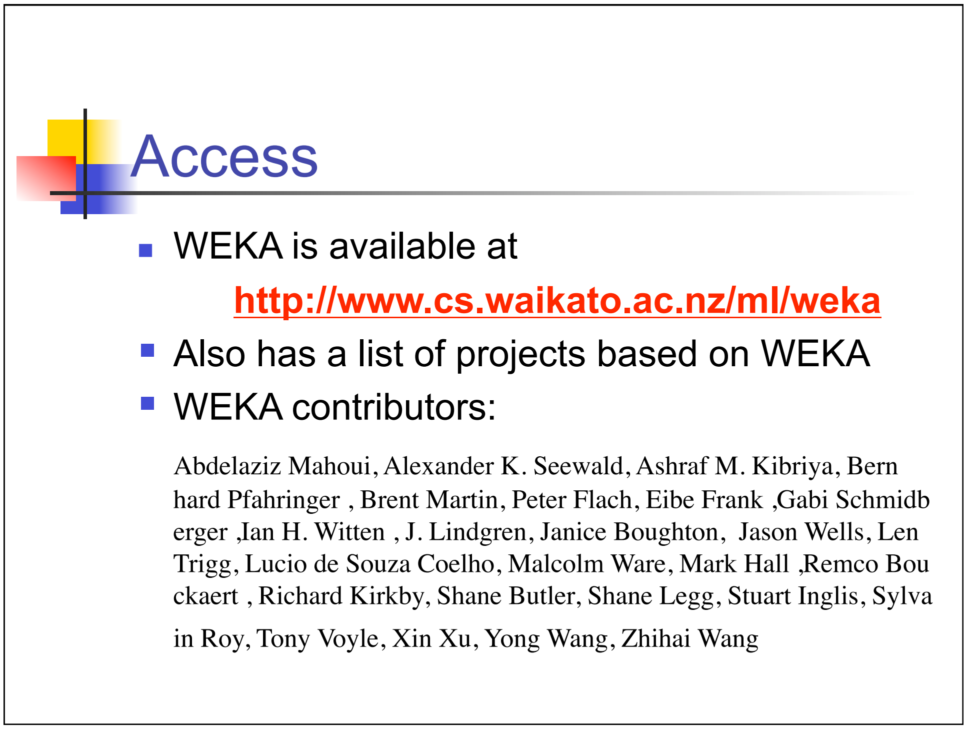
Data Mining with WEKA

Original author: unknown. ? The WEKA team
Additional material: Tim Menzies, 2010



WEKA

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

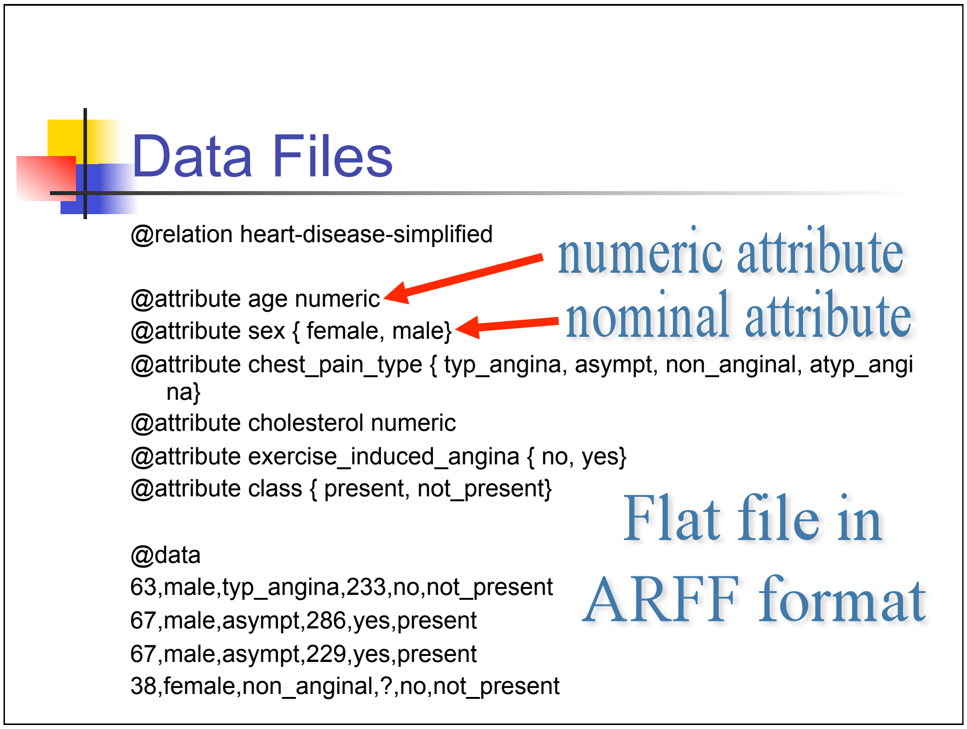


Access

- 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



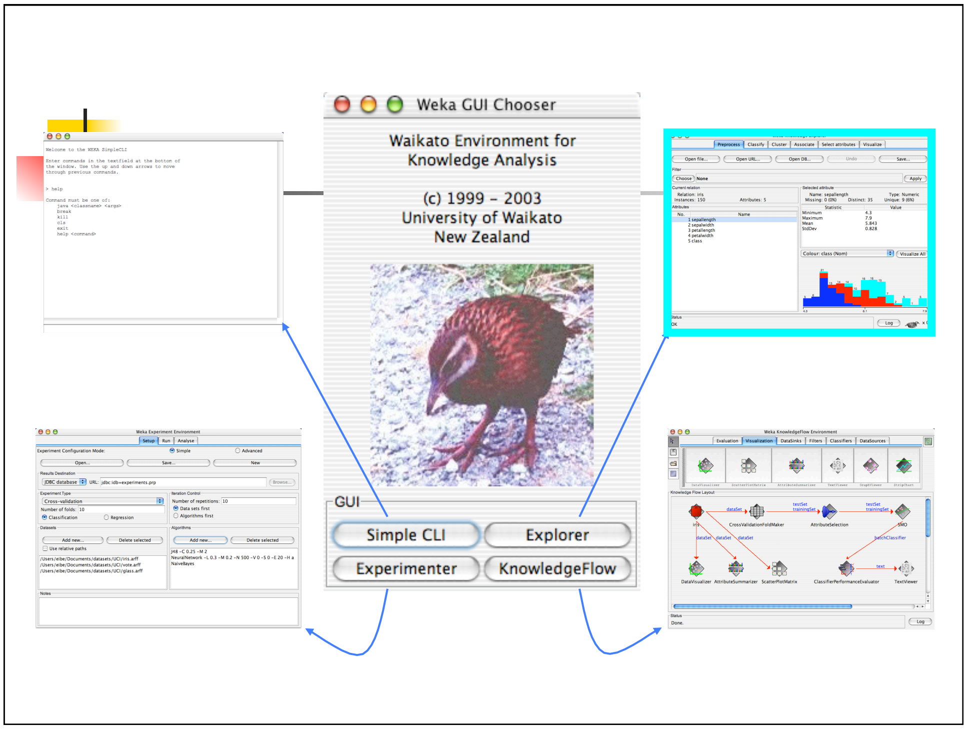
Data Files

```
@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
```

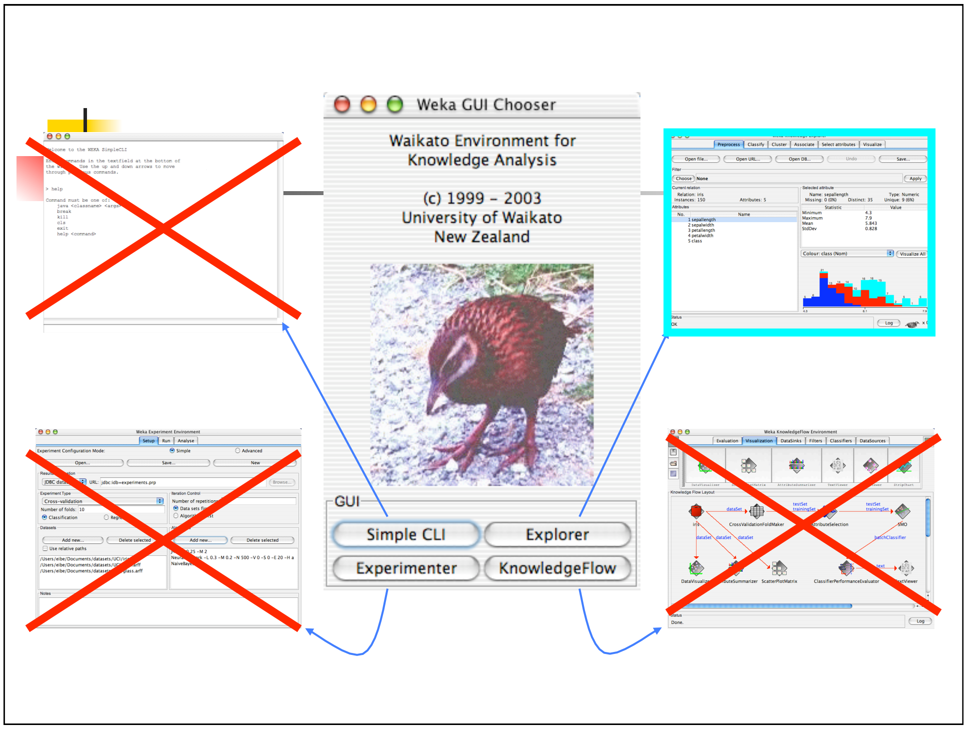
numeric attribute

nominal attribute

Flat file in ARFF format

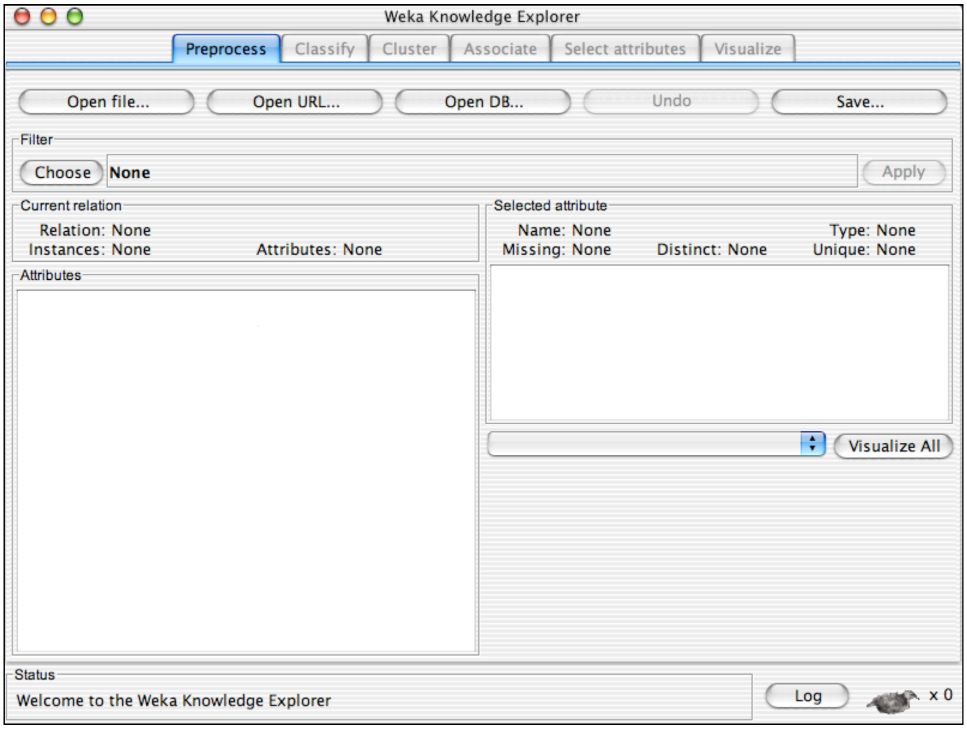


```
Welcome to the WEKA SimpleCLI

Enter commands in the textfield at the bottom of
the window. Use the up and down arrows to move
through previous commands.


> help

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


Weka GUI Chooser
Waikato Environment for Knowledge Analysis
(c) 1999 – 2003
University of Waikato
New Zealand
GUI
Simple CLI
Explorer
Experimenter
KnowledgeFlow

Explorer: pre-processing

- Source
  - 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
  - Called “filters”
  - Discretization, normalization, resampling, attribute selection, transforming and combining attributes, …



Filter: None

Current relation

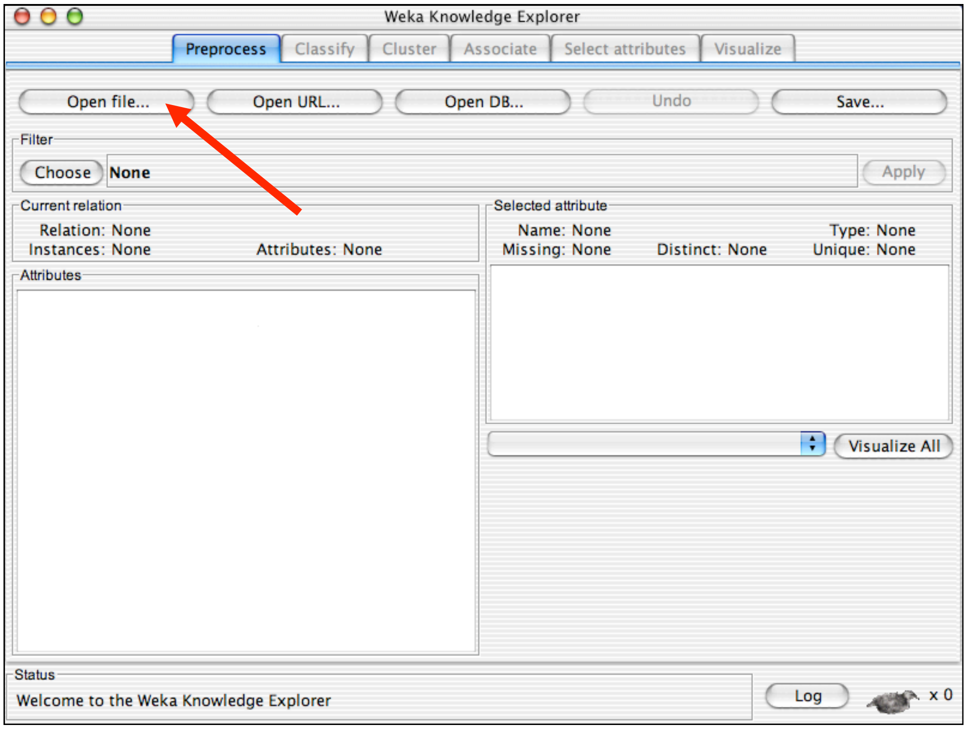
Relation: None
Instances: None
Attributes: None

Selected attribute

Name: None
Type: None
Missing: None
Distinct: None
Unique: None

Status

Welcome to the Weka Knowledge Explorer

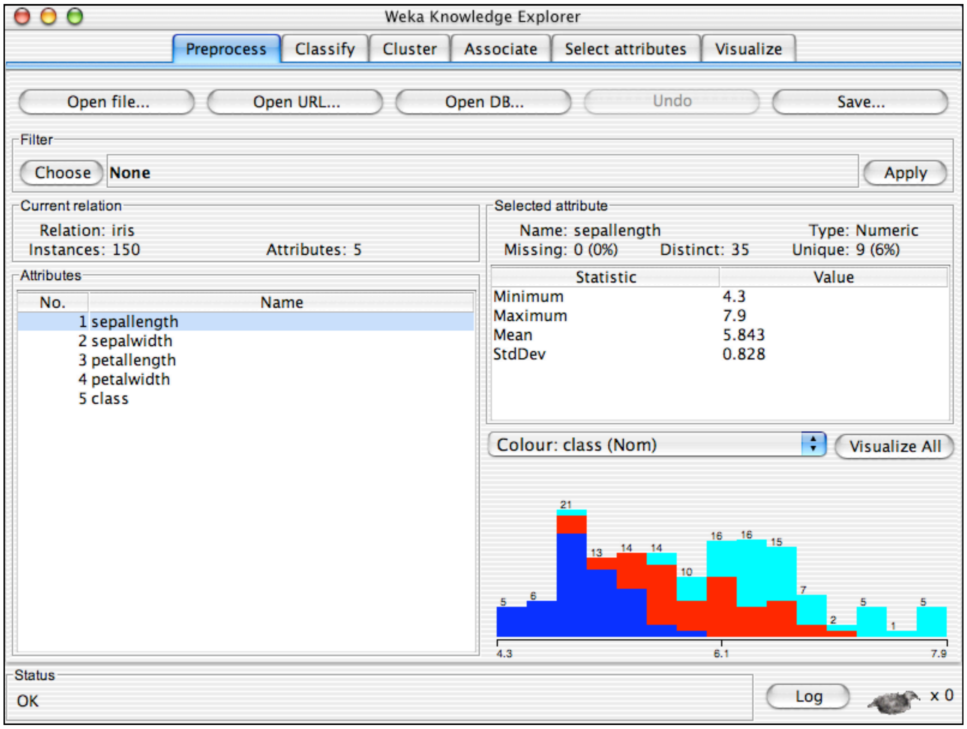


Filter: None

Relation: None
Instances: None Attributes: None

Name: None Type: None
Missing: None Distinct: None Unique: None

Welcome to the Weka Knowledge Explorer



Filter: None

Relation: iris
Instances: 150 Attributes: 5

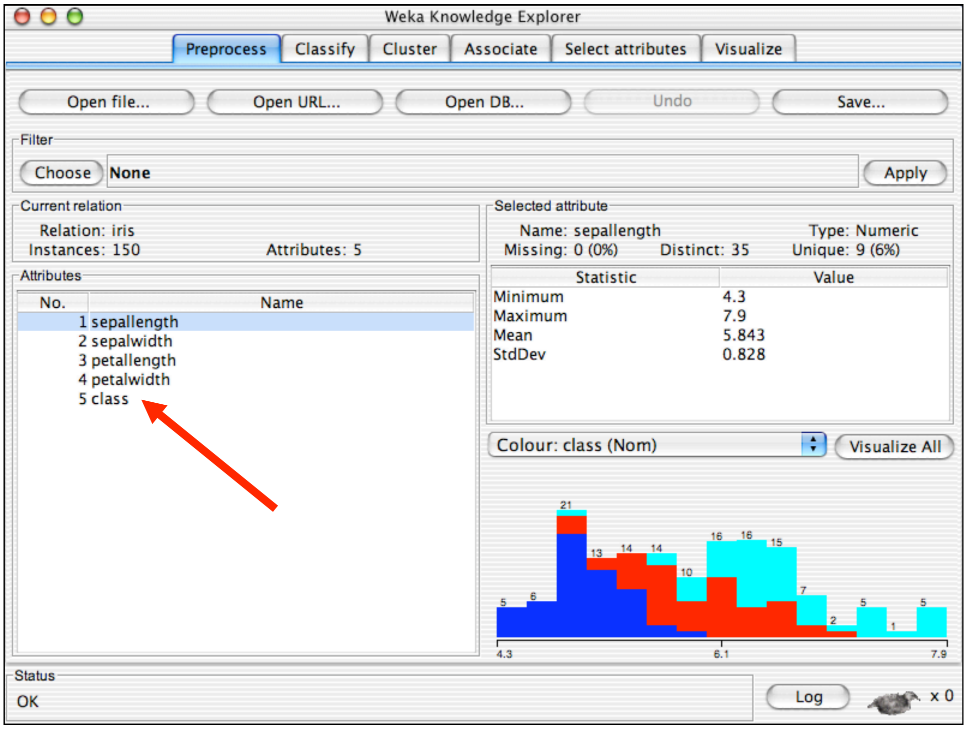
No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

Name: sepallength Type: Numeric
Missing: 0 (0%) Distinct: 35 Unique: 9 (6%)

Statistic	Value
Minimum	4.3
Maximum	7.9
Mean	5.843
StdDev	0.828

Colour: class (Nom)

OK



Filter: None

Current relation
Relation: iris
Instances: 150 Attributes: 5

Attributes

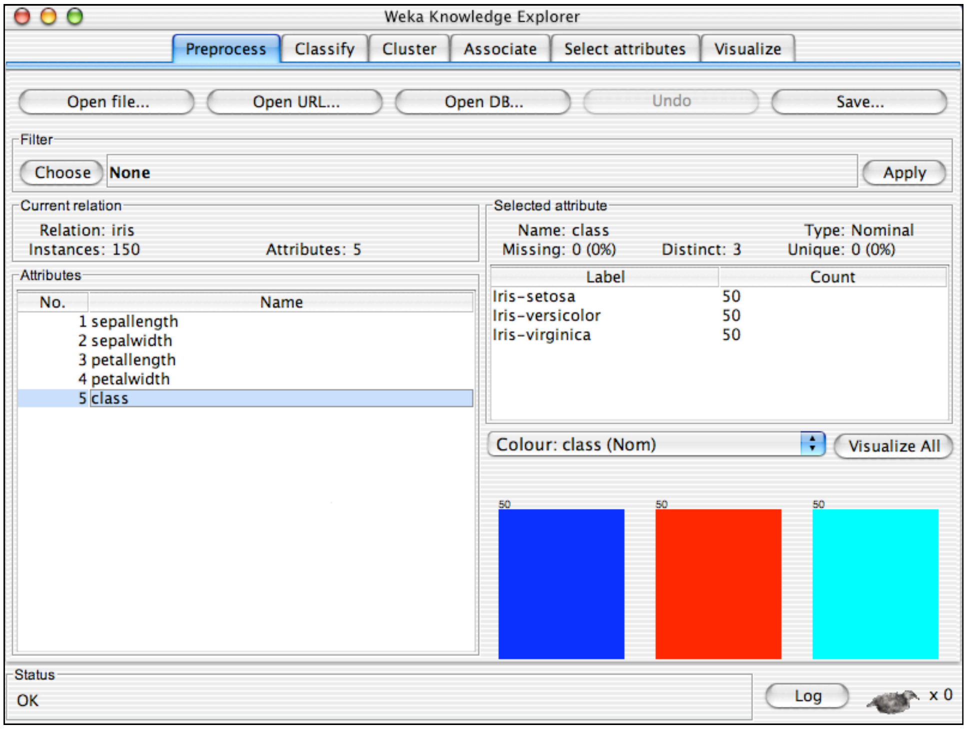
No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

Selected attribute
Name: sepallength Type: Numeric
Missing: 0 (0%) Distinct: 35 Unique: 9 (6%)

Statistic	Value
Minimum	4.3
Maximum	7.9
Mean	5.843
StdDev	0.828

Colour: class (Nom)

Status
OK



Filter: None

Current relation
Relation: iris
Instances: 150 Attributes: 5

Attributes

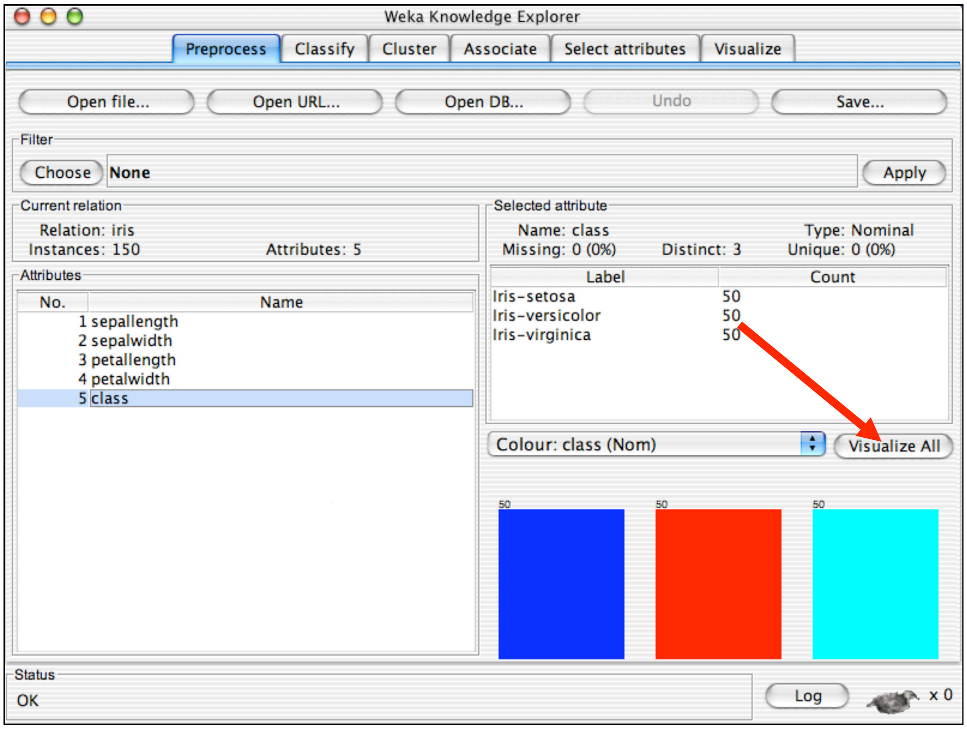
No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

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

Label	Count
Iris-setosa	50
Iris-versicolor	50
Iris-virginica	50

Colour: class (Nom)

Status
OK



Filter

None

Current relation

Relation: iris
Instances: 150 Attributes: 5

Attributes

No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

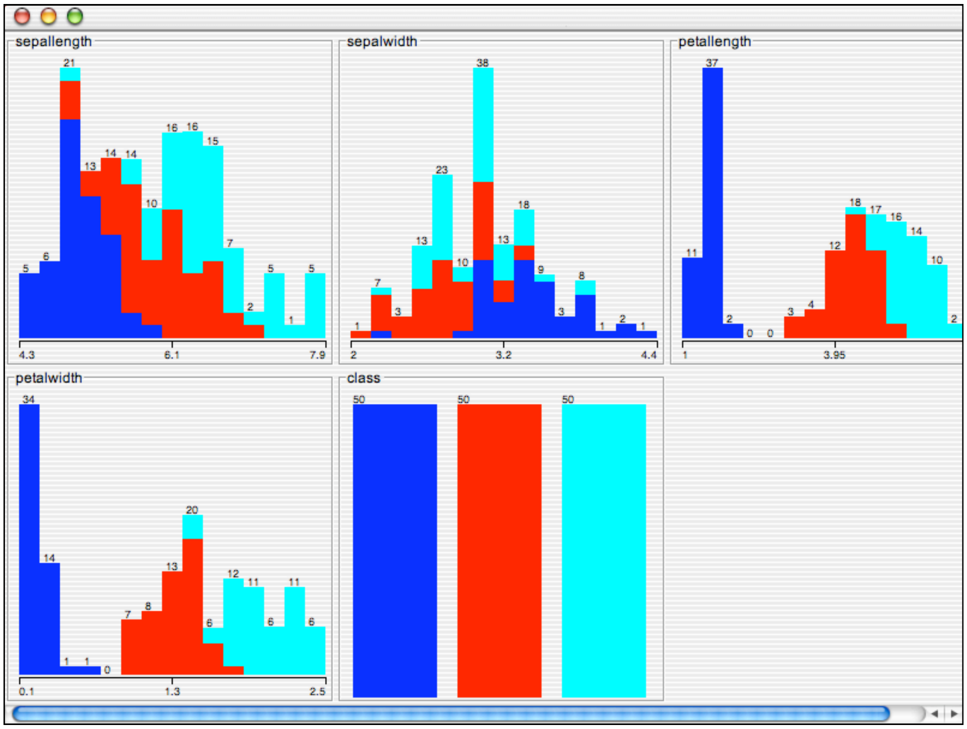
Selected attribute

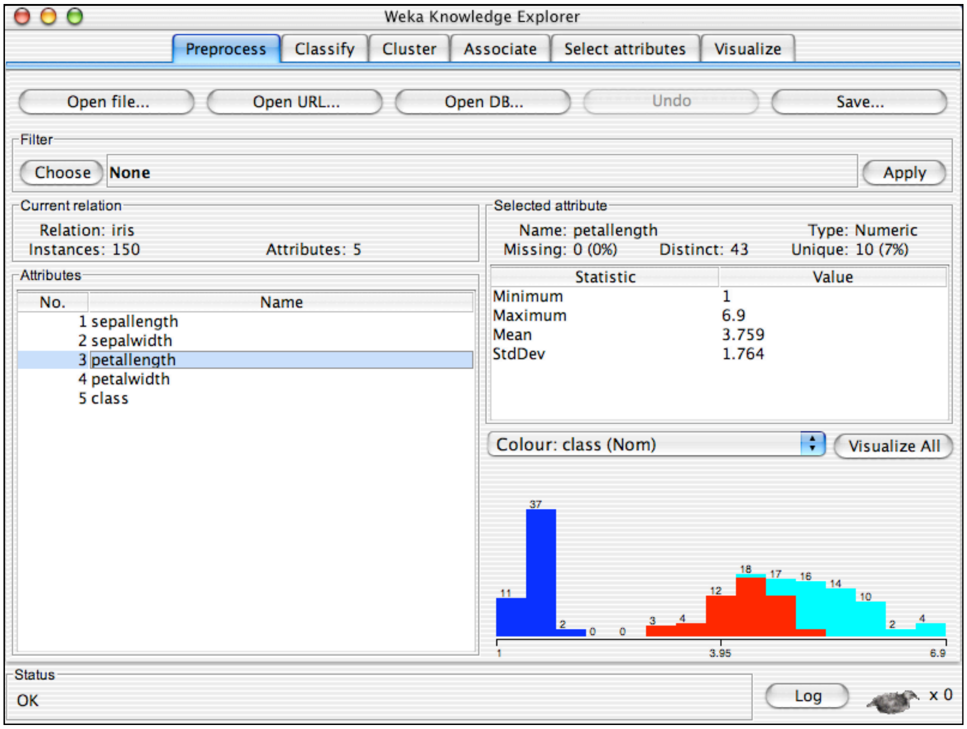
Name: class Type: Nominal
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

Label	Count
Iris-setosa	50
Iris-versicolor	50
Iris-virginica	50

Status

OK





Filter: None

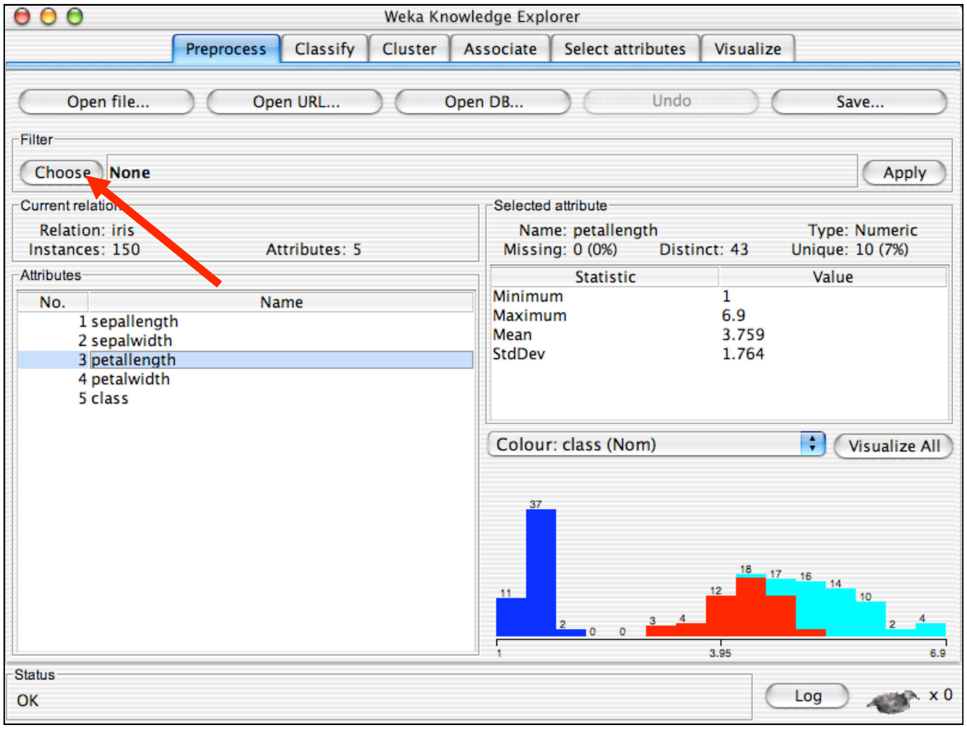
Relation: iris
Instances: 150 Attributes: 5

No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

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

Statistic	Value
Minimum	1
Maximum	6.9
Mean	3.759
StdDev	1.764

Status: OK



Filter: None

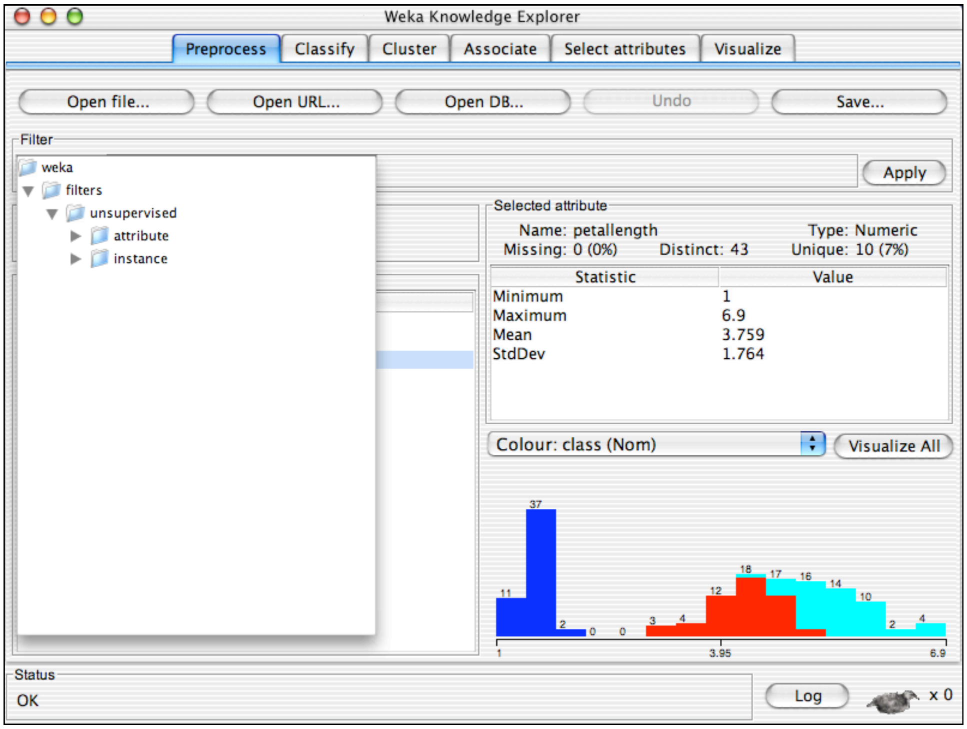
Relation: iris
Instances: 150 Attributes: 5

No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

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

Statistic	Value
Minimum	1
Maximum	6.9
Mean	3.759
StdDev	1.764

Status: OK

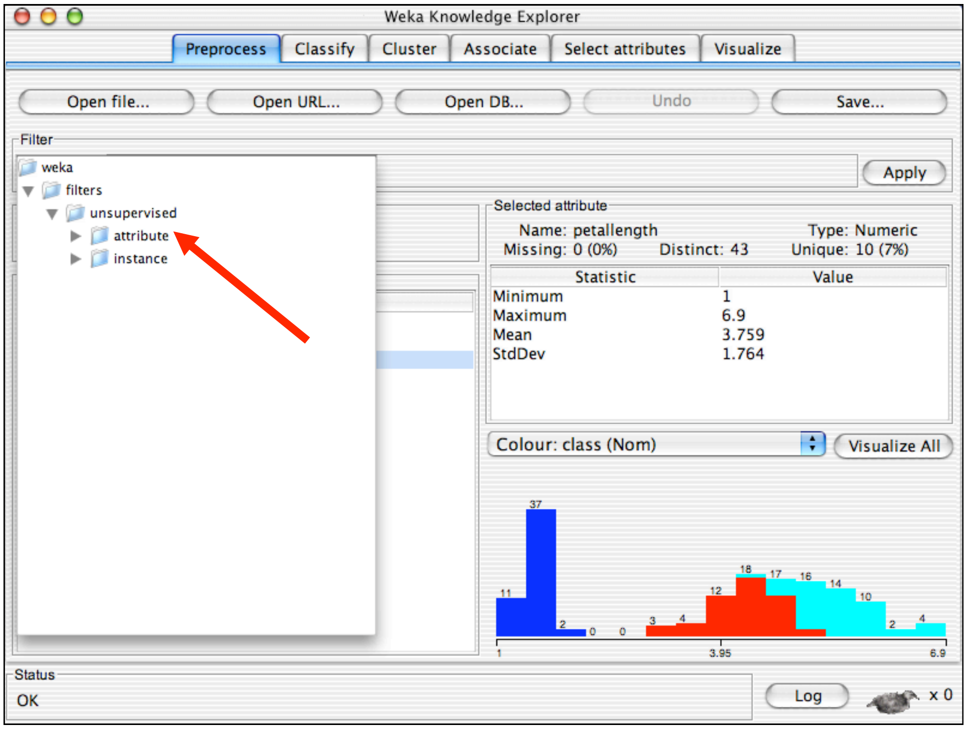


Selected attribute

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

Statistic	Value
Minimum	1
Maximum	6.9
Mean	3.759
StdDev	1.764

Status
OK

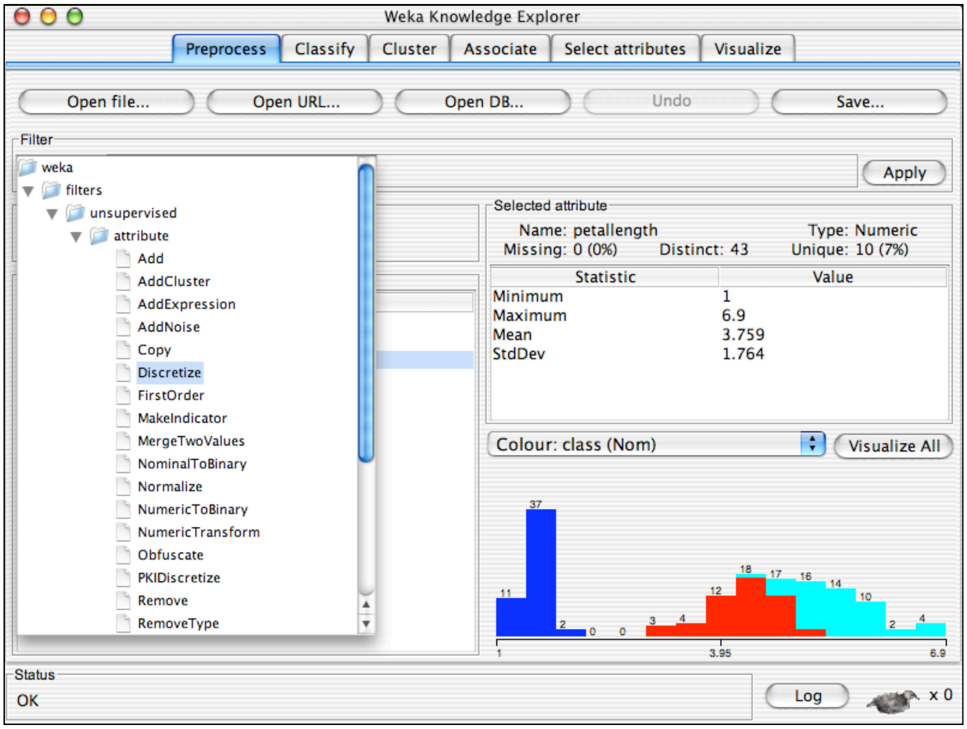


Selected attribute

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

Statistic	Value
Minimum	1
Maximum	6.9
Mean	3.759
StdDev	1.764

Status
OK



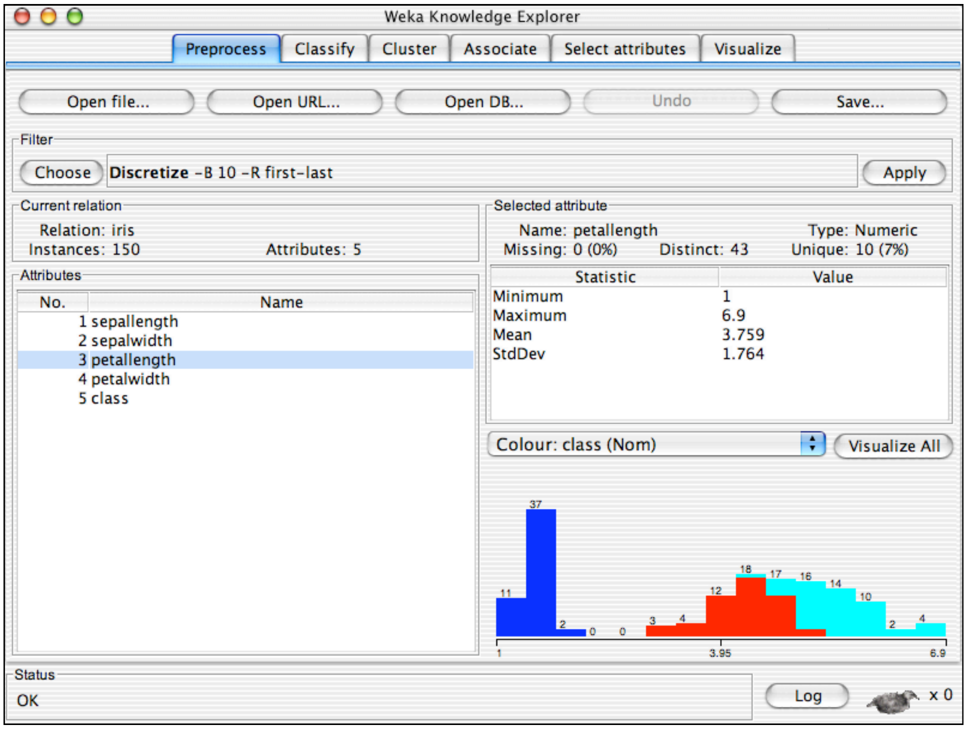
Selected attribute

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

Statistic	Value
Minimum	1
Maximum	6.9
Mean	3.759
StdDev	1.764

Status

OK



Filter

Discretize -B 10 -R first-last

Current relation

Relation: iris
Instances: 150 Attributes: 5

Attributes

No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

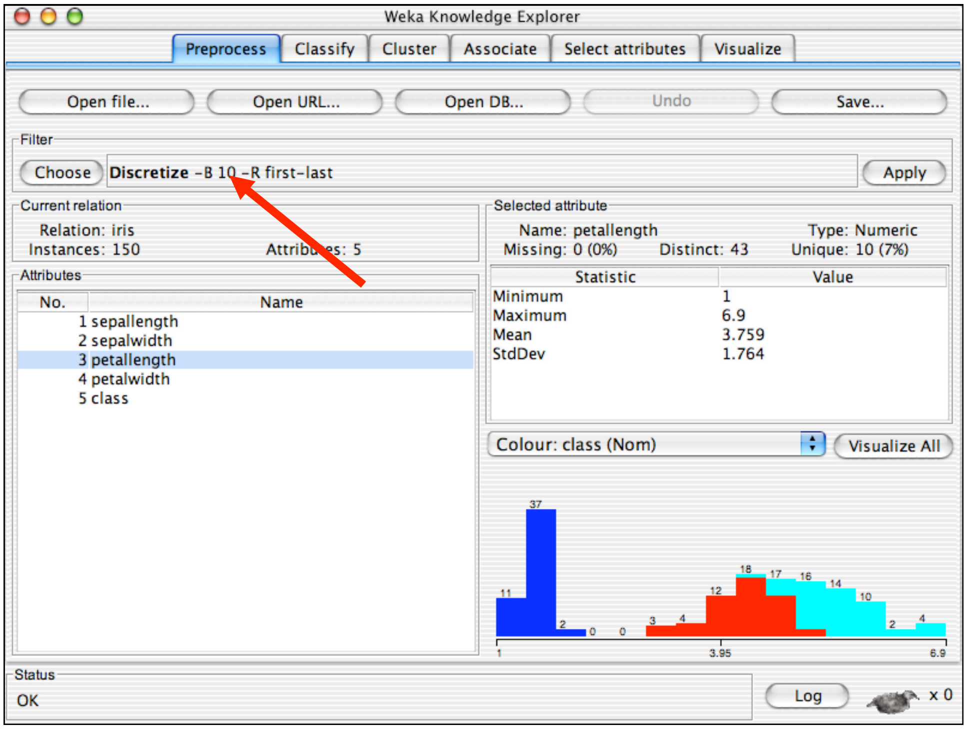
Selected attribute

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

Statistic	Value
Minimum	1
Maximum	6.9
Mean	3.759
StdDev	1.764

Status

OK



Filter

Discretize -B 10 -R first-last

Current relation

Relation: iris
Instances: 150 Attributes: 5

Attributes

No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

Selected attribute

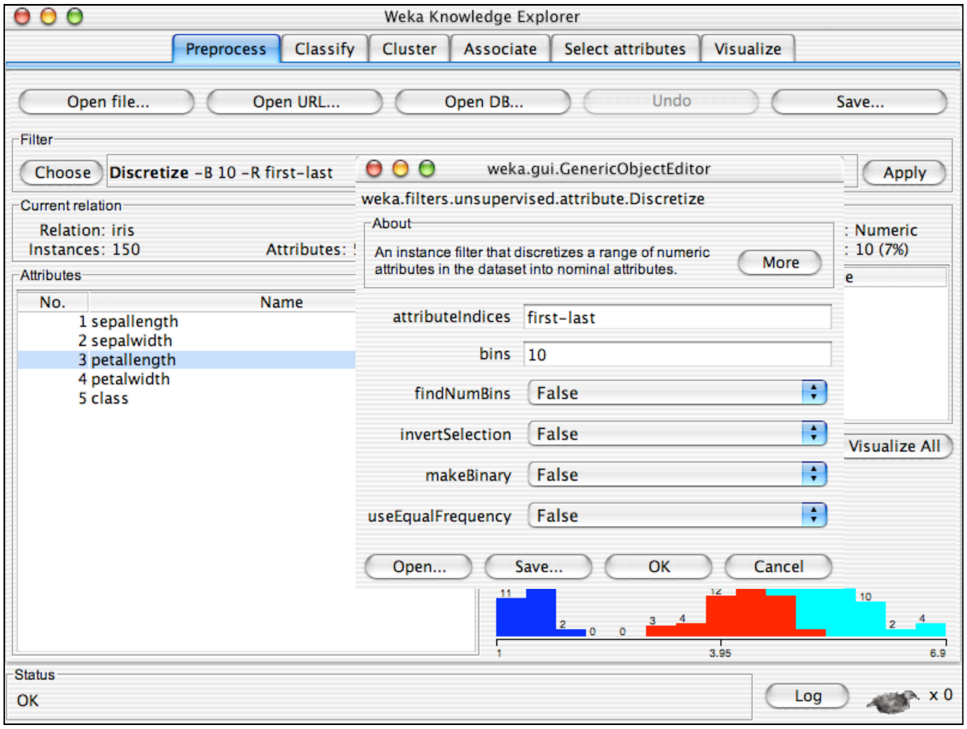
Name: petallength Type: Numeric
Missing: 0 (0%) Distinct: 43 Unique: 10 (7%)

Statistic	Value
Minimum	1
Maximum	6.9
Mean	3.759
StdDev	1.764

Colour: class (Nom)

Status

OK



Filter

Discretize -B 10 -R first-last

Current relation

Relation: iris
Instances: 150 Attributes: !

Attributes

No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

: Numeric
: 10 (7%)
e

weka.gui.GenericObjectEditor

weka.filters.unsupervised.attribute.Discretize

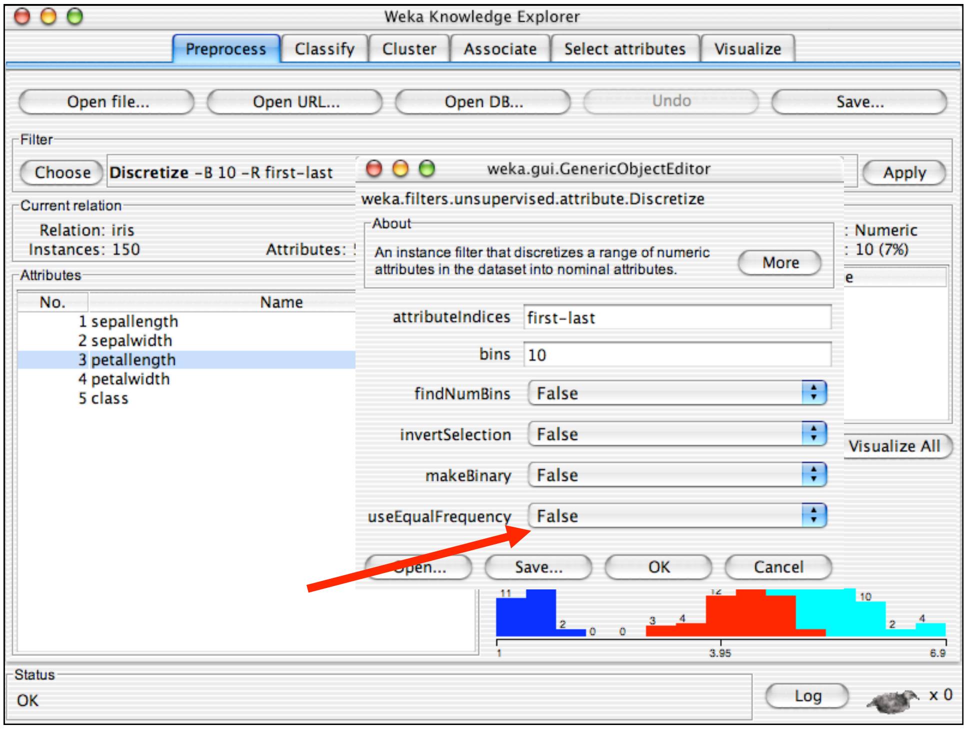
About

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

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

Status

OK



Discretize -B 10 -R first-last

Relation: iris
Instances: 150
Attributes: !

: Numeric
: 10 (7%)
e

No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

weka.gui.GenericObjectEditor

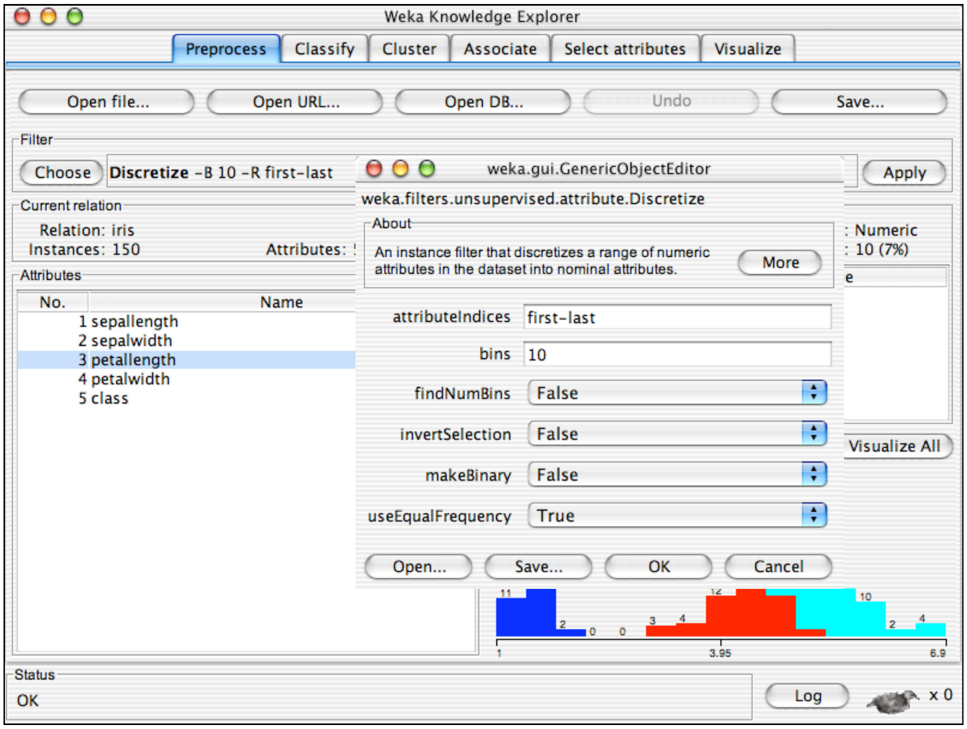
weka.filters.unsupervised.attribute.Discretize

About

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

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

OK



Discretize -B 10 -R first-last

Relation: iris
Instances: 150
Attributes: !

: Numeric
: 10 (7%)
e

No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

weka.gui.GenericObjectEditor

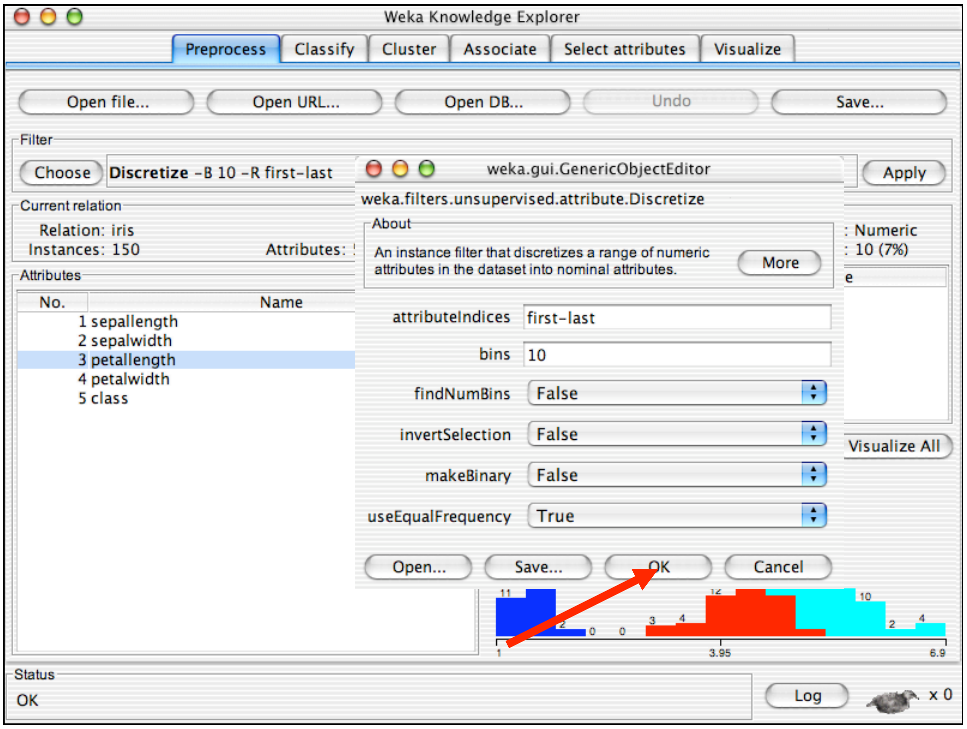
weka.filters.unsupervised.attribute.Discretize

About

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

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

OK



Filter: Discretize -B 10 -R first-last

Current relation
Relation: iris
Instances: 150
Attributes: 5

Attributes

No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

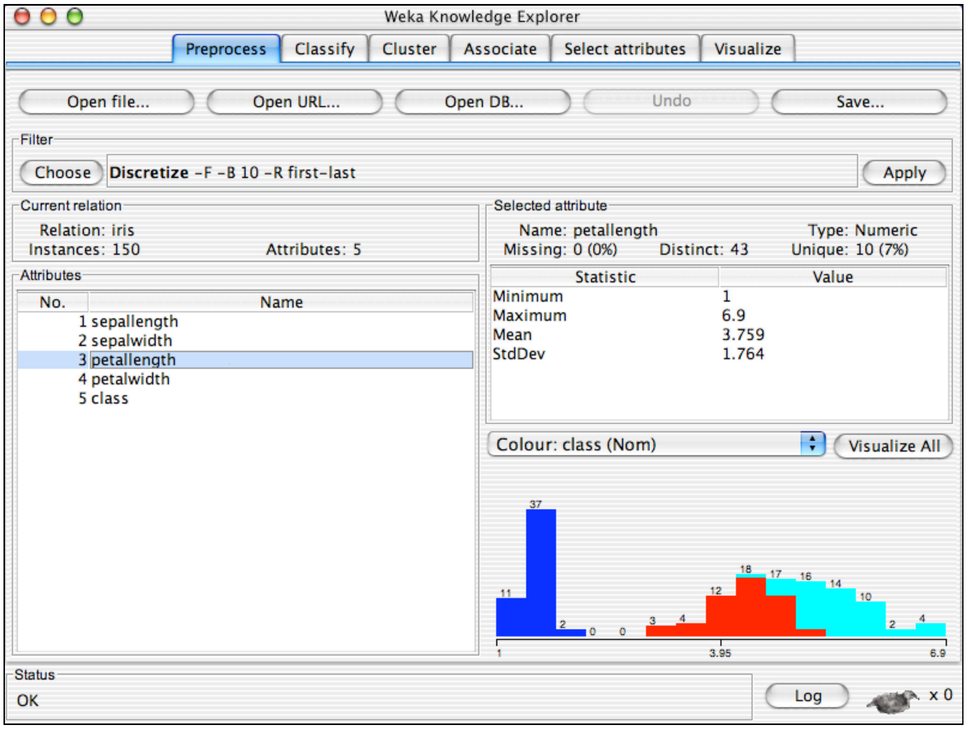
: Numeric
: 10 (7%)

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	True

Status
OK



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

Current relation
Relation: iris
Instances: 150
Attributes: 5

Attributes

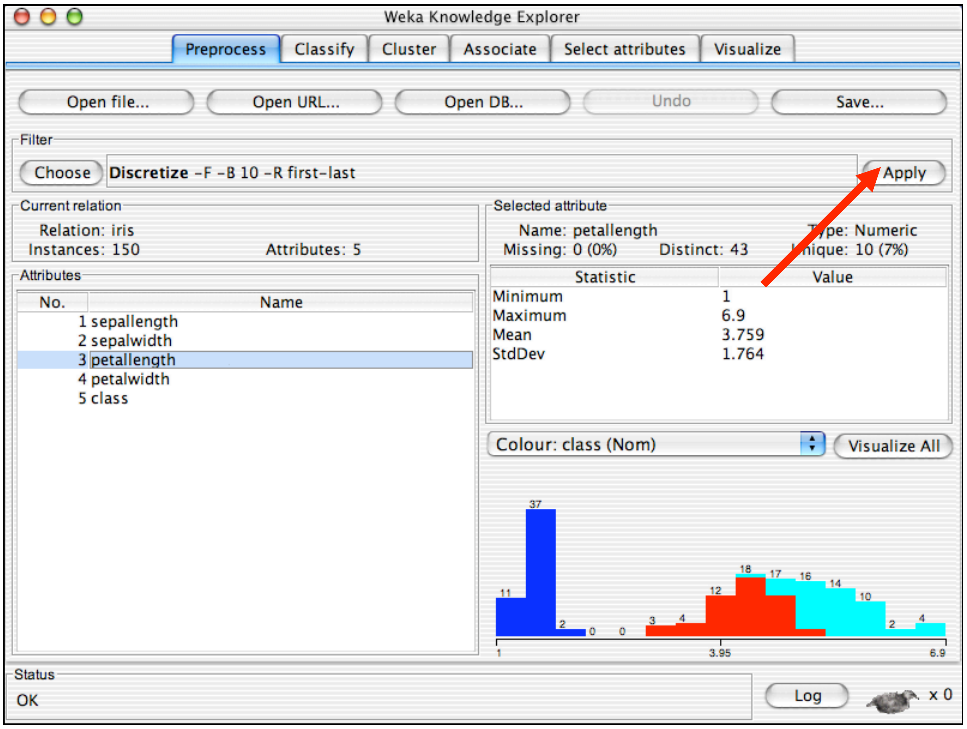
No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

Selected attribute
Name: petallength Type: Numeric
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



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

Current relation
Relation: iris
Instances: 150 Attributes: 5

Attributes

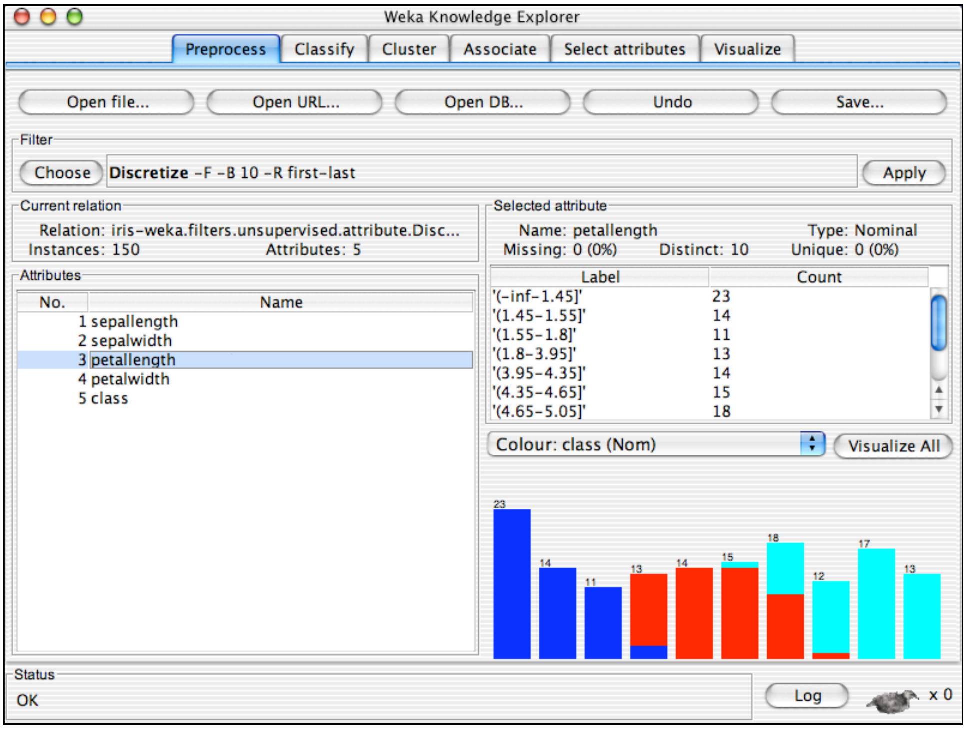
No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

Selected attribute
Name: petallength Type: Numeric
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



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

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

Attributes

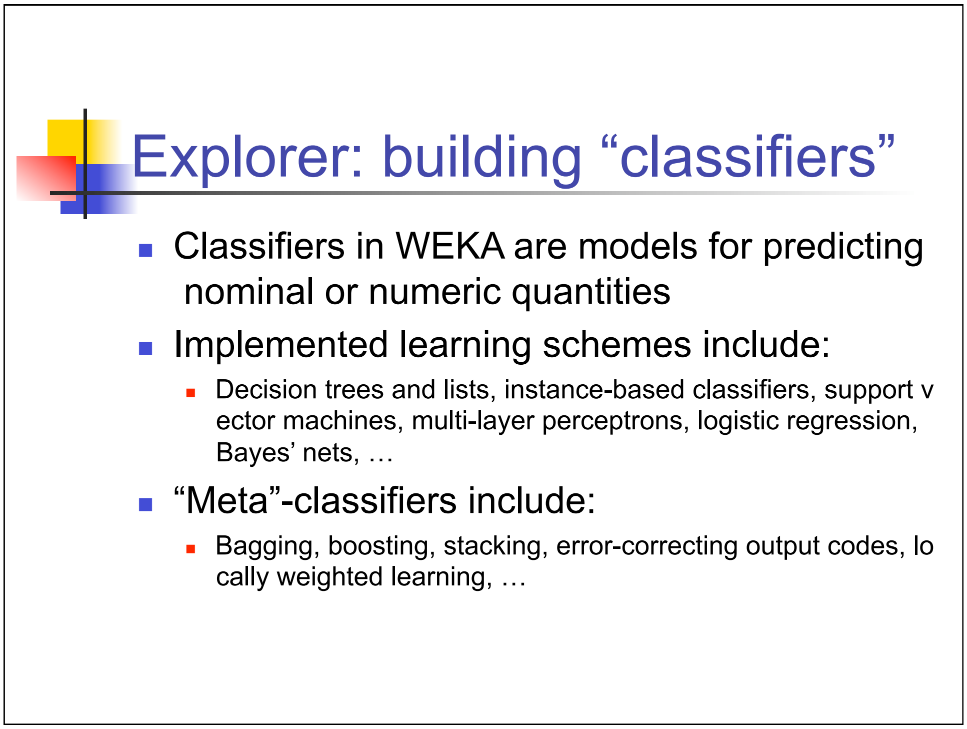
No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

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

Label	Count
'(-inf-1.45]'	23
'(1.45-1.55]'	14
'(1.55-1.8]'	11
'(1.8-3.95]'	13
'(3.95-4.35]'	14
'(4.35-4.65]'	15
'(4.65-5.05]'	18

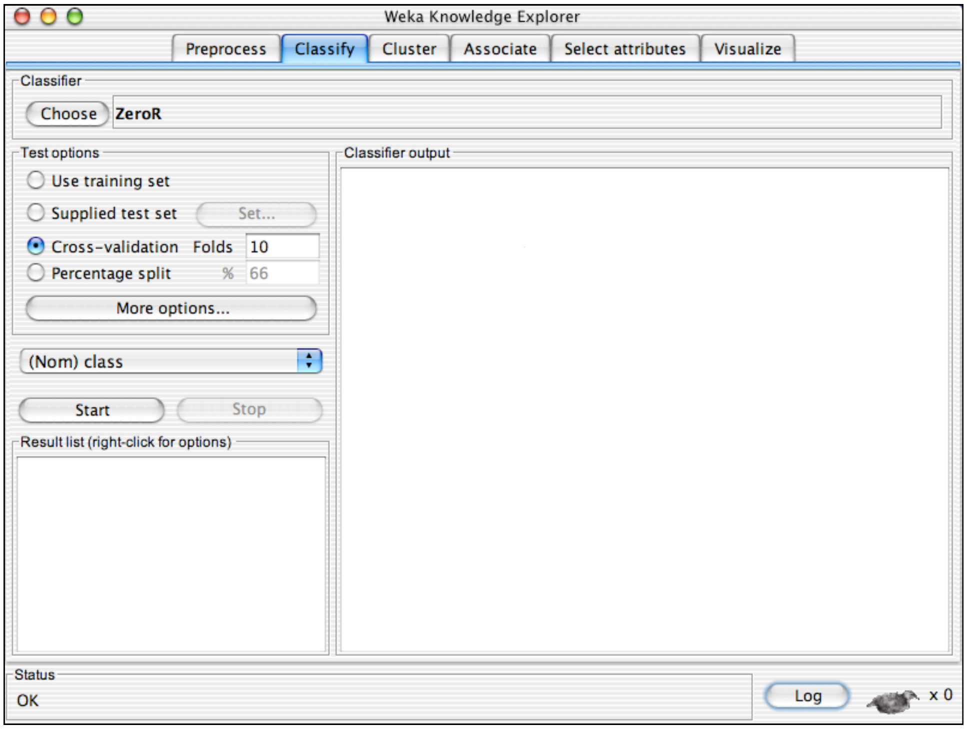
Colour: class (Nom)

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, …



Classifier: ZeroR

Test options

○ Use training set

○ Supplied test set

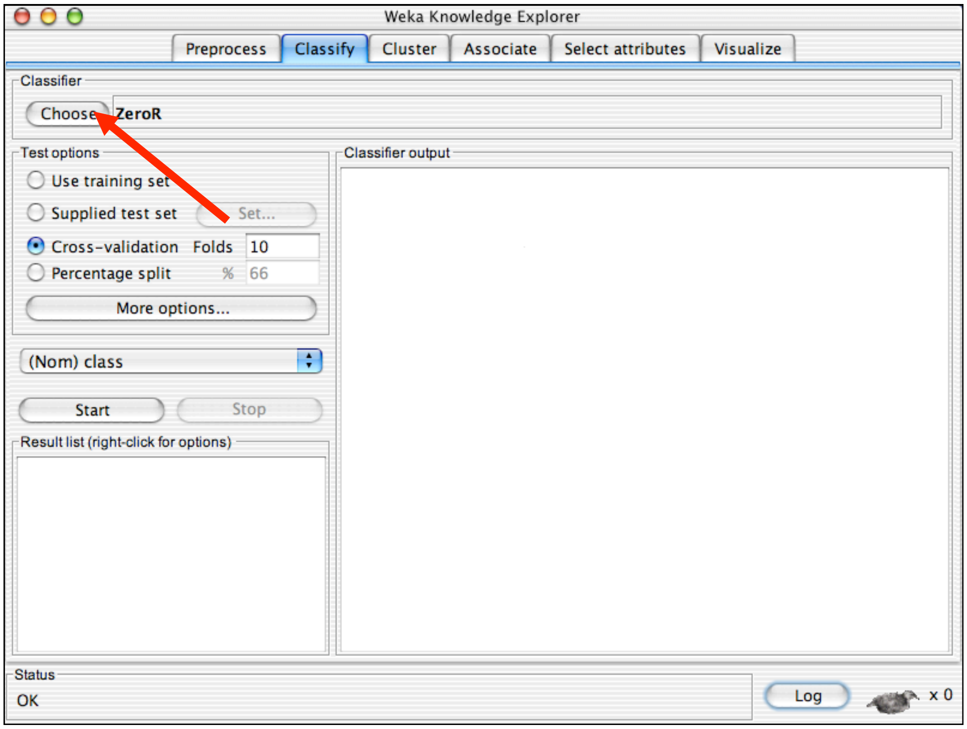
● Cross-validation Folds 10

○ Percentage split % 66

(Nom) class

Status

OK

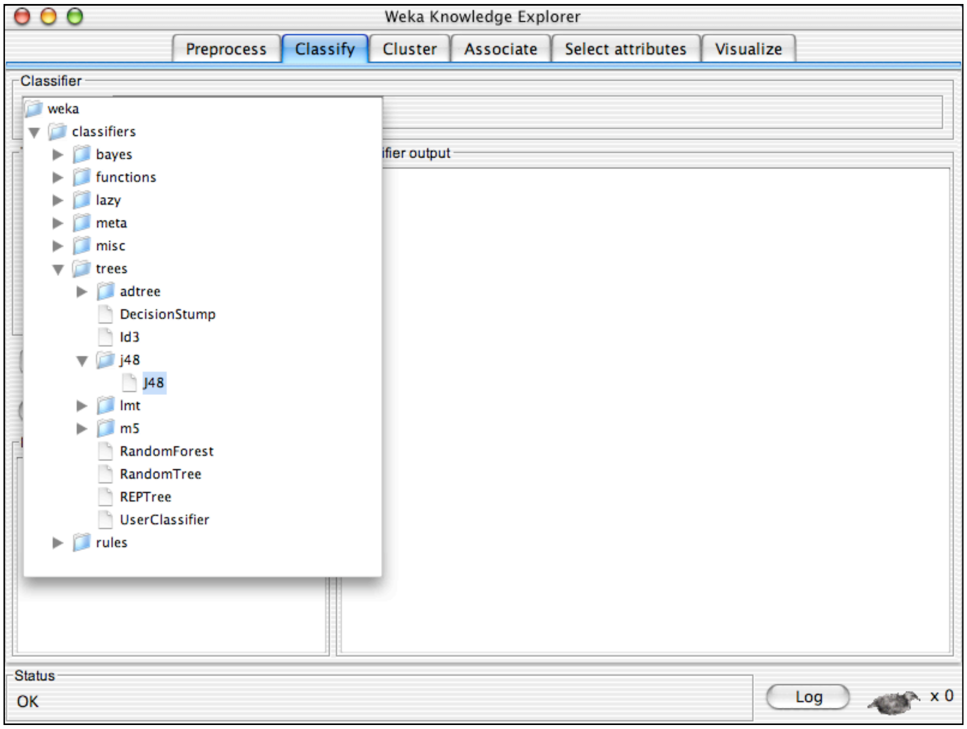


Classifier: ZeroR

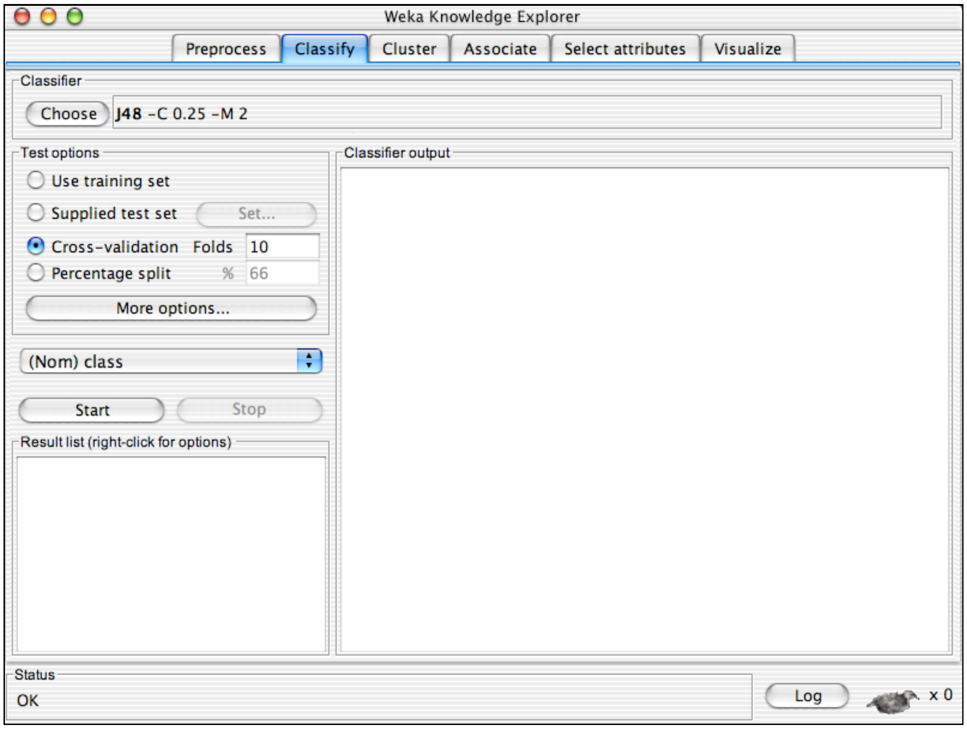
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) class

Status: OK



Status: OK



Classifier: J48 -C 0.25 -M 2

Test options

○ Use training set

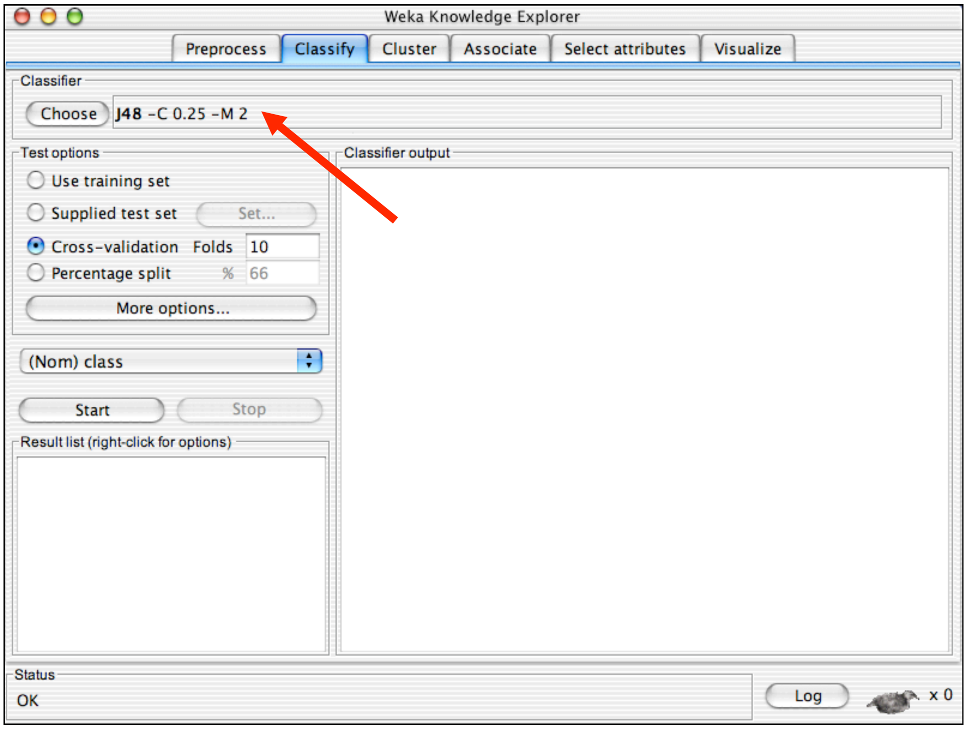
○ Supplied test set

● Cross-validation Folds 10

○ Percentage split % 66

(Nom) class

Status: OK



Classifier: J48 -C 0.25 -M 2

Test options

○ Use training set

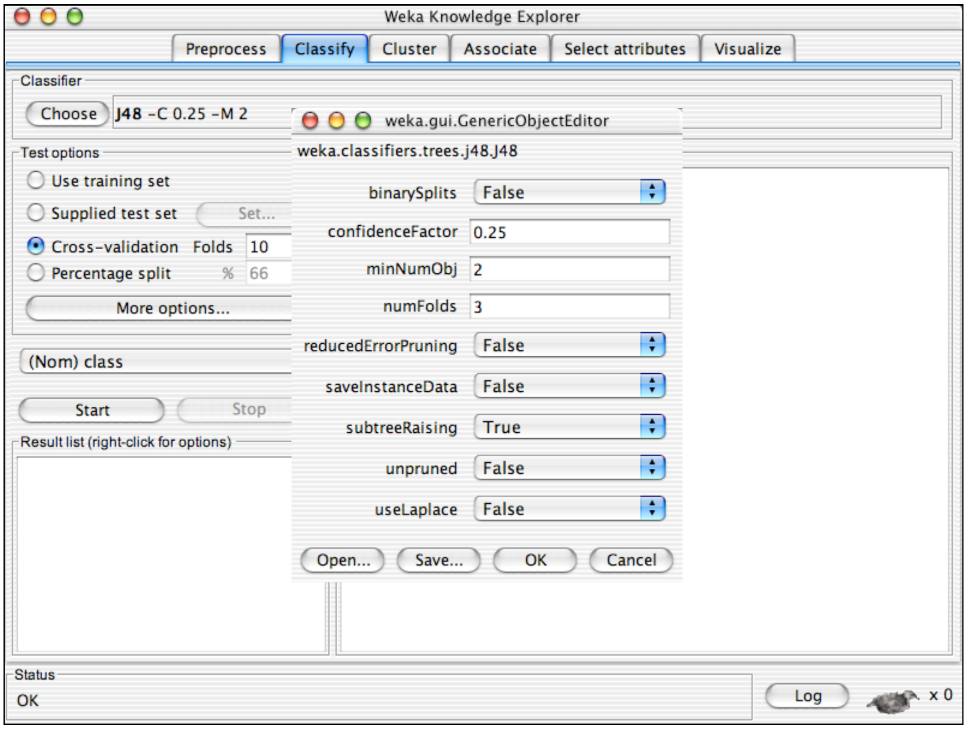
○ Supplied test set

● Cross-validation Folds 10

○ Percentage split % 66

(Nom) class

Status: OK



J48 -C 0.25 -M 2

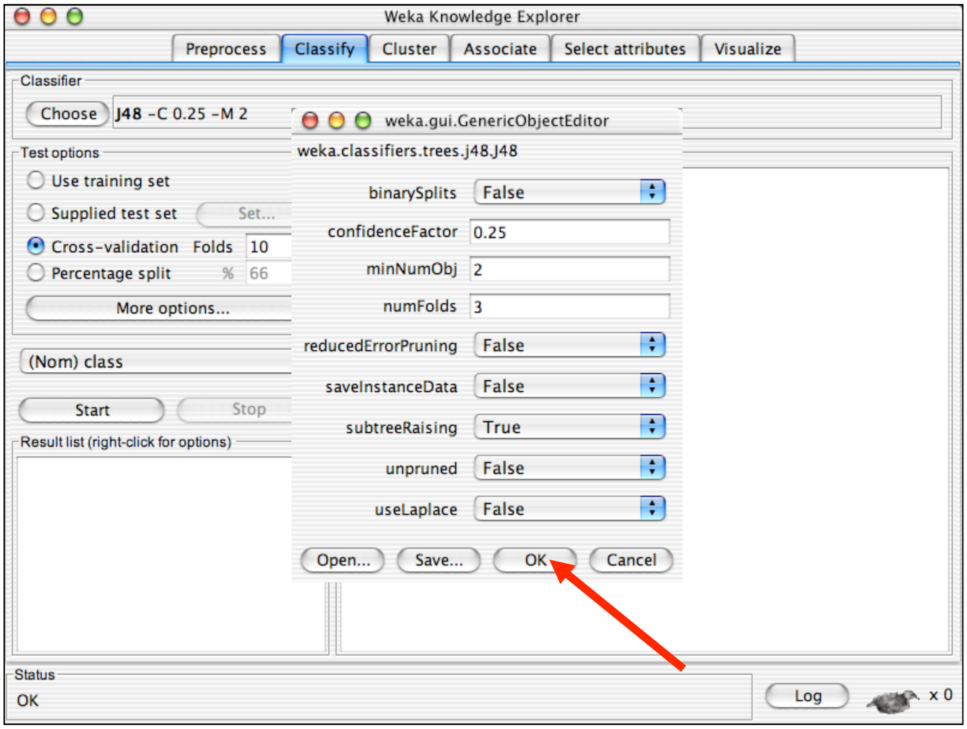
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) class

OK

weka.classifiers.trees.j48.J48

binarySplits: False
confidenceFactor: 0.25
minNumObj: 2
numFolds: 3
reducedErrorPruning: False
saveInstanceData: False
subtreeRaising: True
unpruned: False
useLaplace: False



J48 -C 0.25 -M 2

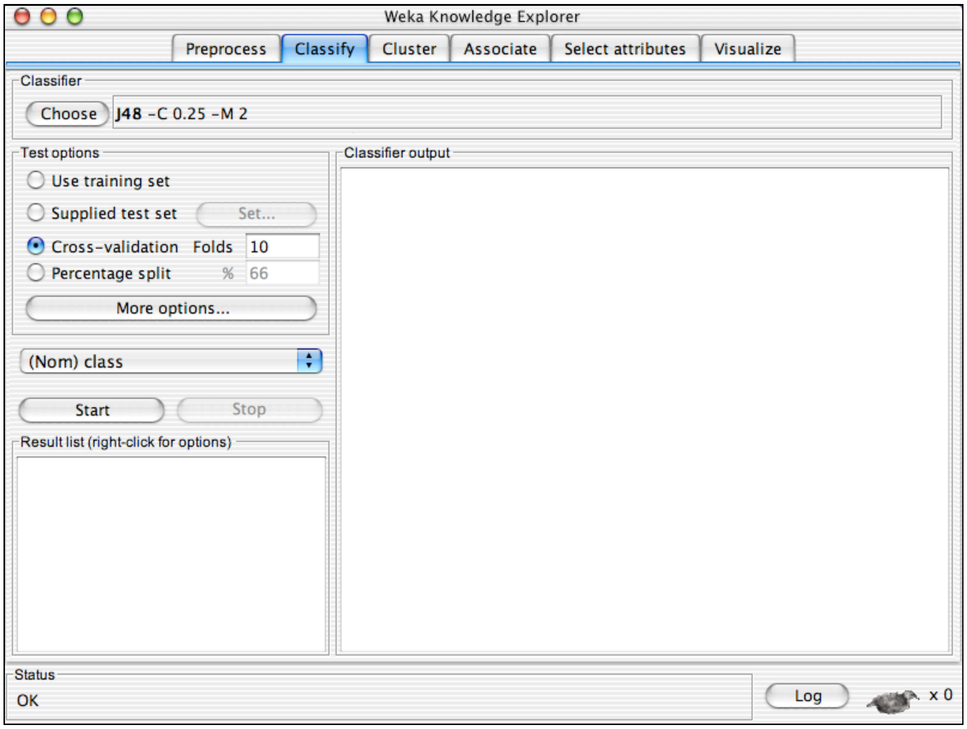
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) class

OK

weka.classifiers.trees.j48.J48

binarySplits: False
confidenceFactor: 0.25
minNumObj: 2
numFolds: 3
reducedErrorPruning: False
saveInstanceData: False
subtreeRaising: True
unpruned: False
useLaplace: False



Classifier: J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set

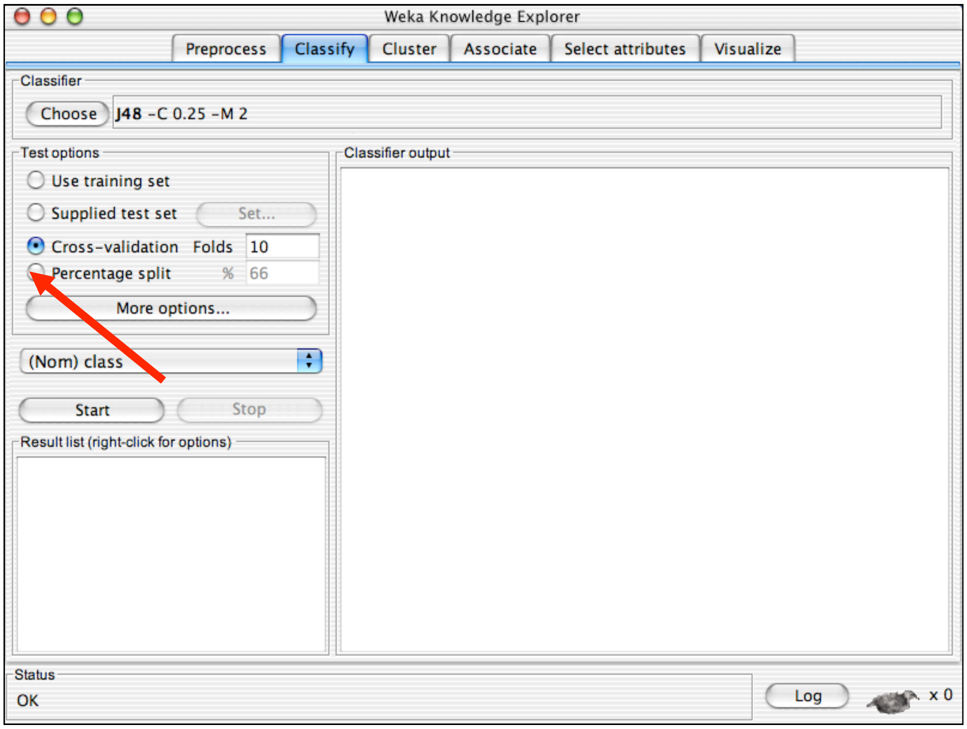
● Cross-validation Folds 10

○ Percentage split % 66

(Nom) class

Status

OK



Classifier: J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set

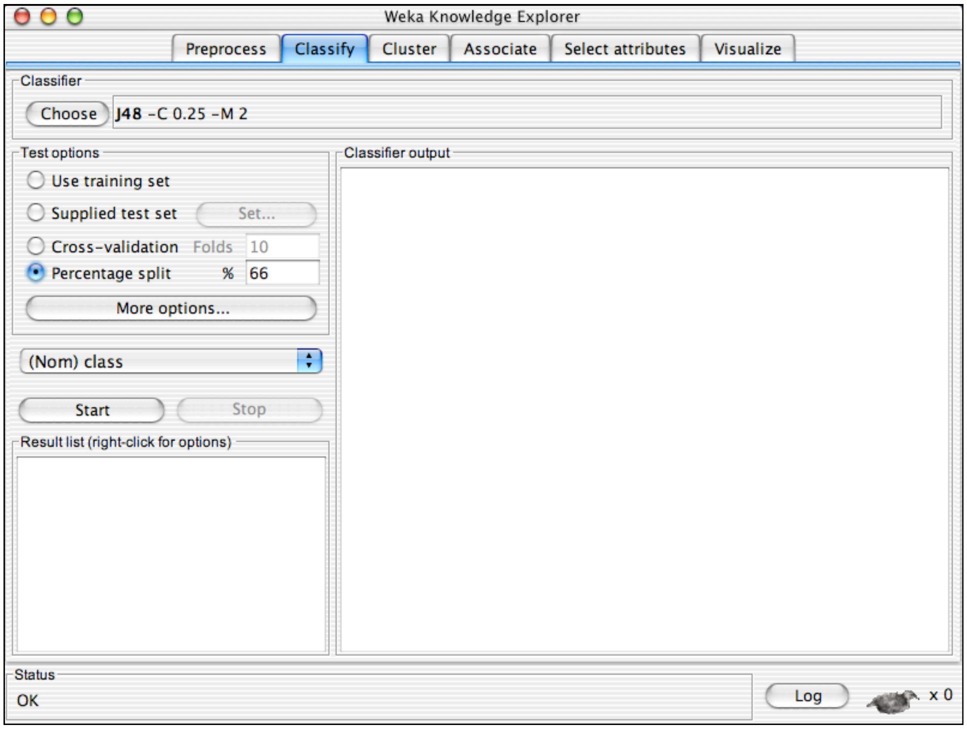
● Cross-validation Folds 10

○ Percentage split % 66

(Nom) class

Status

OK

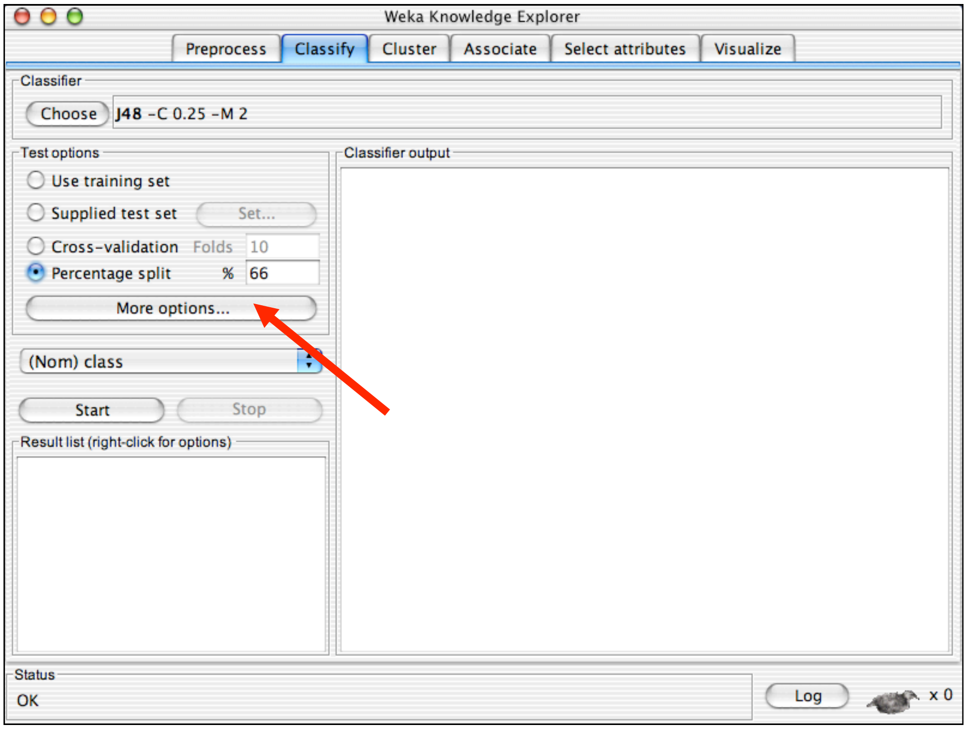


Classifier: J48 -C 0.25 -M 2

Cross-validation Folds 10

● Percentage split % 66

Status: OK

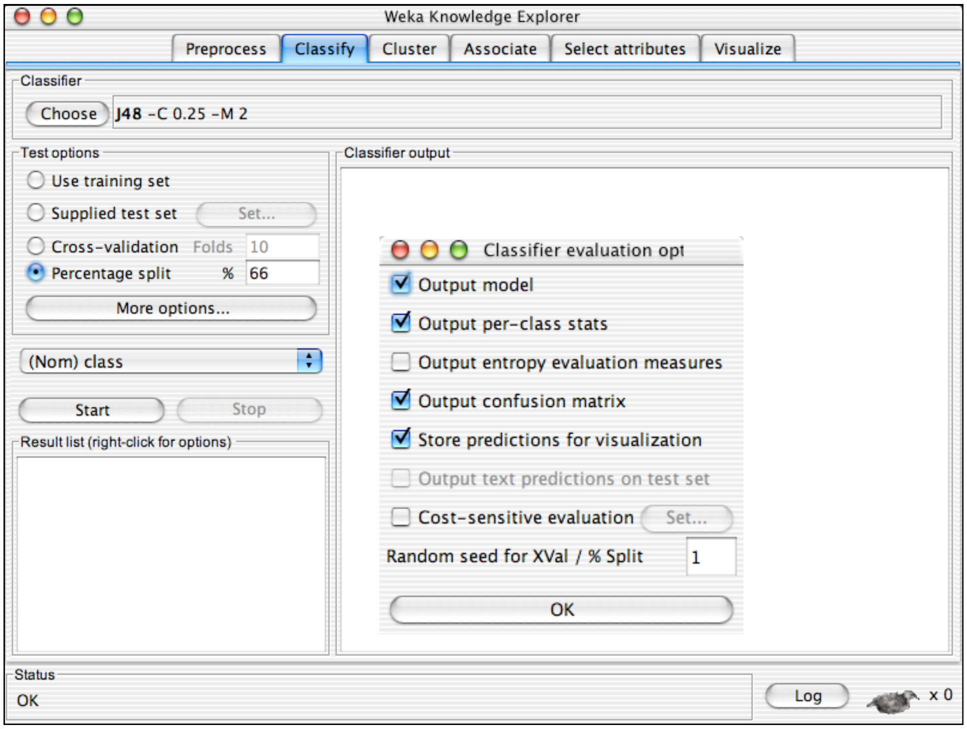


Classifier: J48 -C 0.25 -M 2

Cross-validation Folds 10

● Percentage split % 66

Status: OK



Weka Knowledge 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) class

Start Stop

Result list (right-click for options)

Classifier output

Classifier evaluation opt

☑ Output model

☑ Output per-class stats

□ Output entropy evaluation measures

☑ Output confusion matrix

☑ Store predictions for visualization

□ Output text predictions on test set

□ Cost-sensitive evaluation Set...

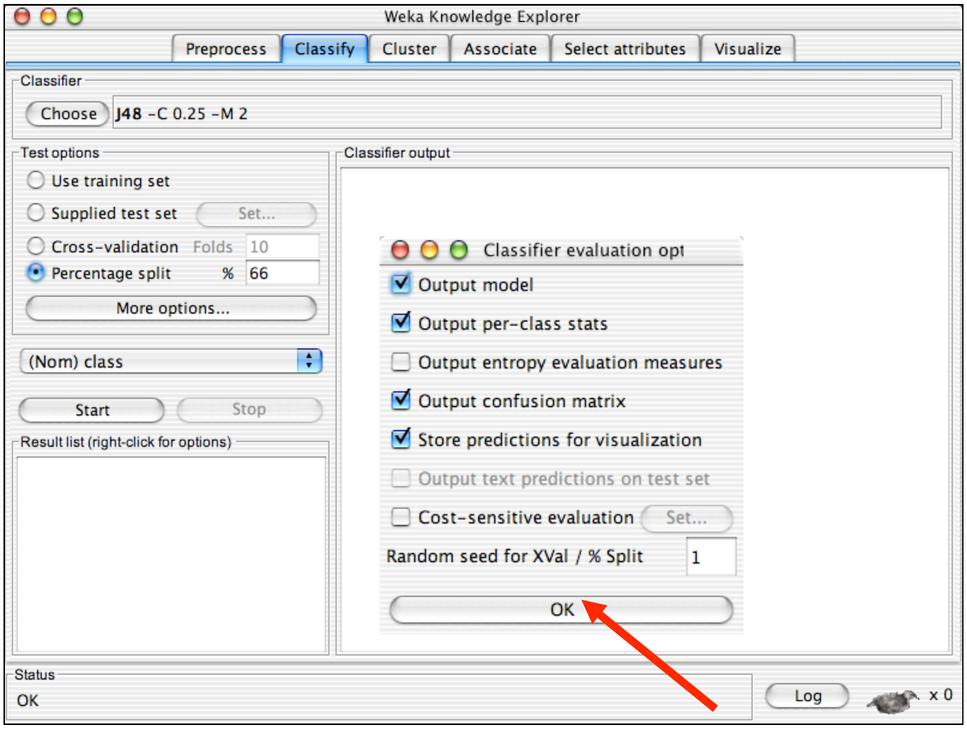
Random seed for XVal / % Split 1

OK

Status

OK

Log x 0



Weka Knowledge 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) class

Start Stop

Result list (right-click for options)

Classifier output

Classifier evaluation opt

☑ Output model

☑ Output per-class stats

□ Output entropy evaluation measures

☑ Output confusion matrix

☑ Store predictions for visualization

□ Output text predictions on test set

□ Cost-sensitive evaluation Set...

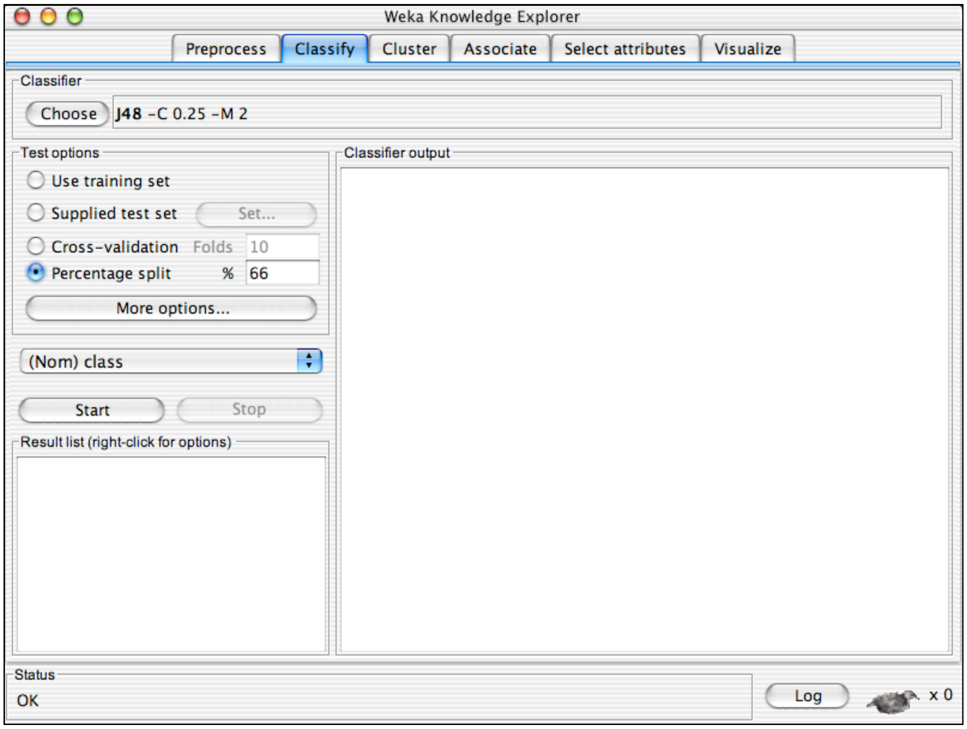
Random seed for XVal / % Split 1

OK

Status

OK

Log x 0



Classifier

J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set

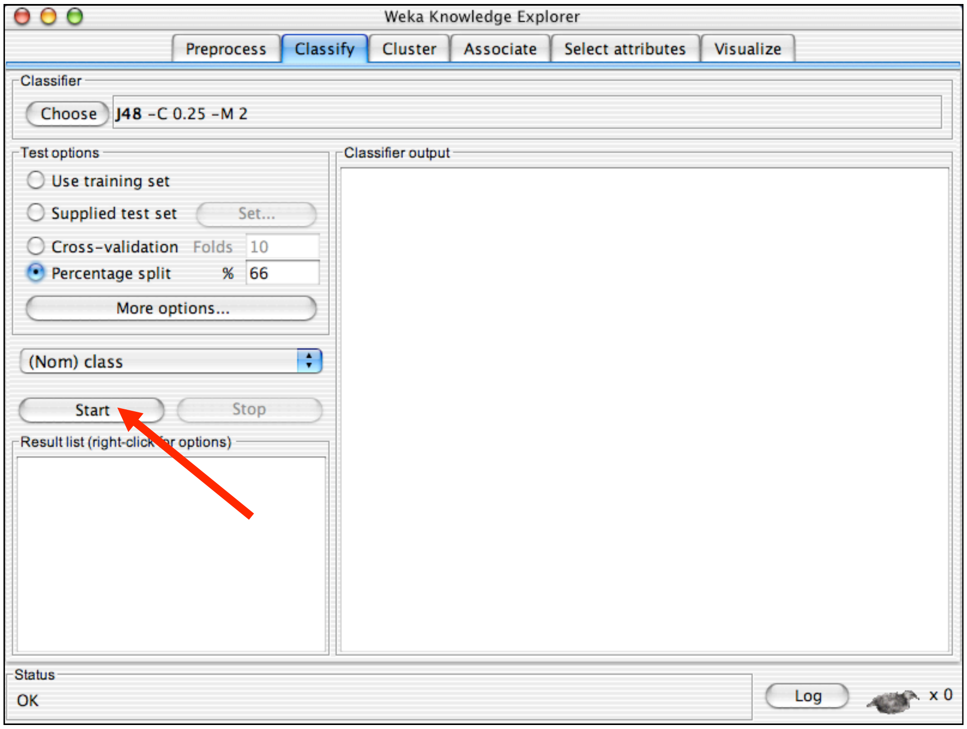
○ Cross-validation Folds 10

● Percentage split % 66

(Nom) class

Status

OK



Classifier

J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set

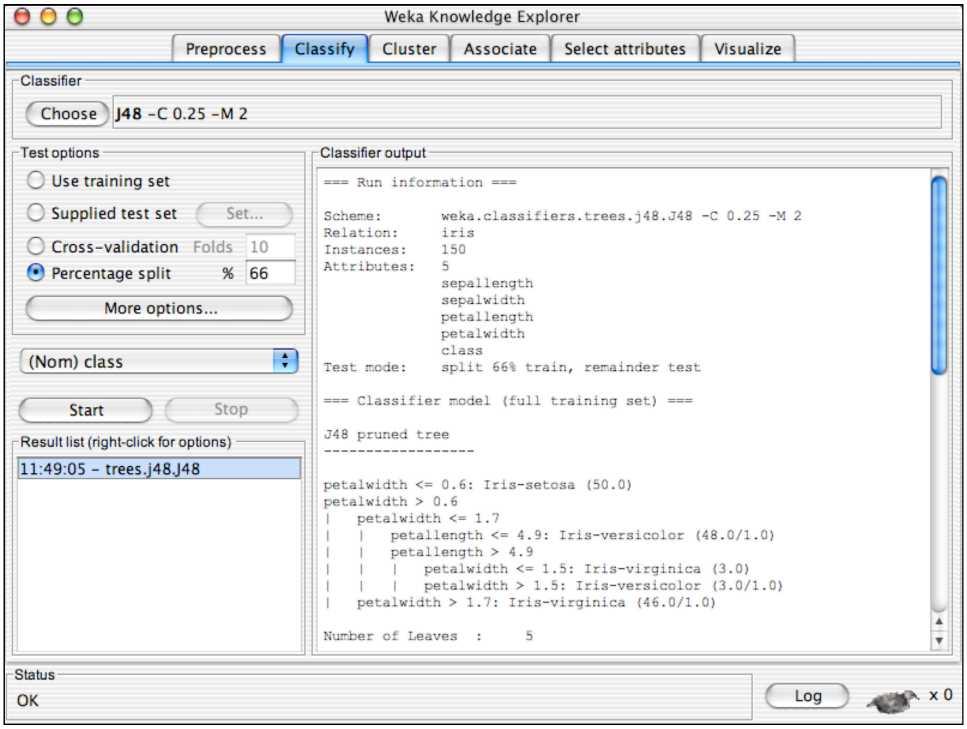
○ Cross-validation Folds 10

● Percentage split % 66

(Nom) class

Status

OK



Classifier: J48 -C 0.25 -M 2

Test options

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

(Nom) class

Result list (right-click for options)

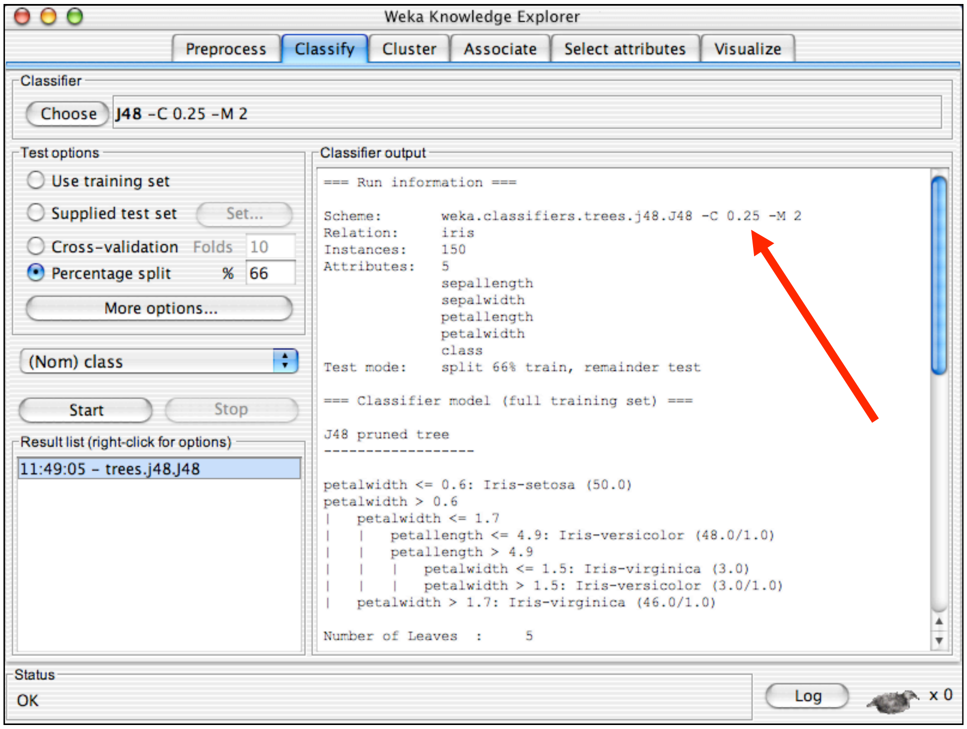
11:49:05 - trees.j48.J48

Classifier output

```
=== Run information ===

Scheme:       weka.classifiers.trees.j48.J48 -C 0.25 -M 2
Relation:     iris
Instances:    150
Attributes:   5
              sepallength
              sepalwidth
              petallength
              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
|   |   petallength <= 4.9: Iris-versicolor (48.0/1.0)
|   |   petallength > 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
```

Status

OK



Weka Knowledge Explorer

Classifier: J48 -C 0.25 -M 2

Test options

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

(Nom) class

Result list (right-click for options)

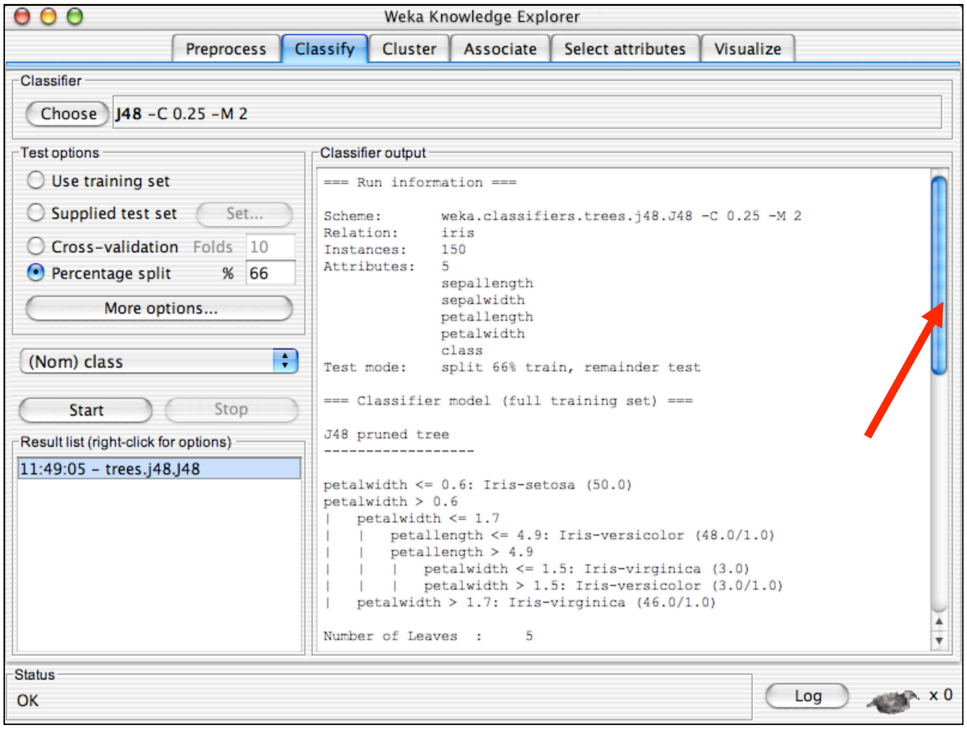
11:49:05 - trees.j48.J48

Classifier output

```
=== Run information ===

Scheme:       weka.classifiers.trees.j48.J48 -C 0.25 -M 2
Relation:     iris
Instances:    150
Attributes:   5
              sepallength
              sepalwidth
              petallength
              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
|   |   petallength <= 4.9: Iris-versicolor (48.0/1.0)
|   |   petallength > 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
```

Status

OK



Classifier: J48 -C 0.25 -M 2

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

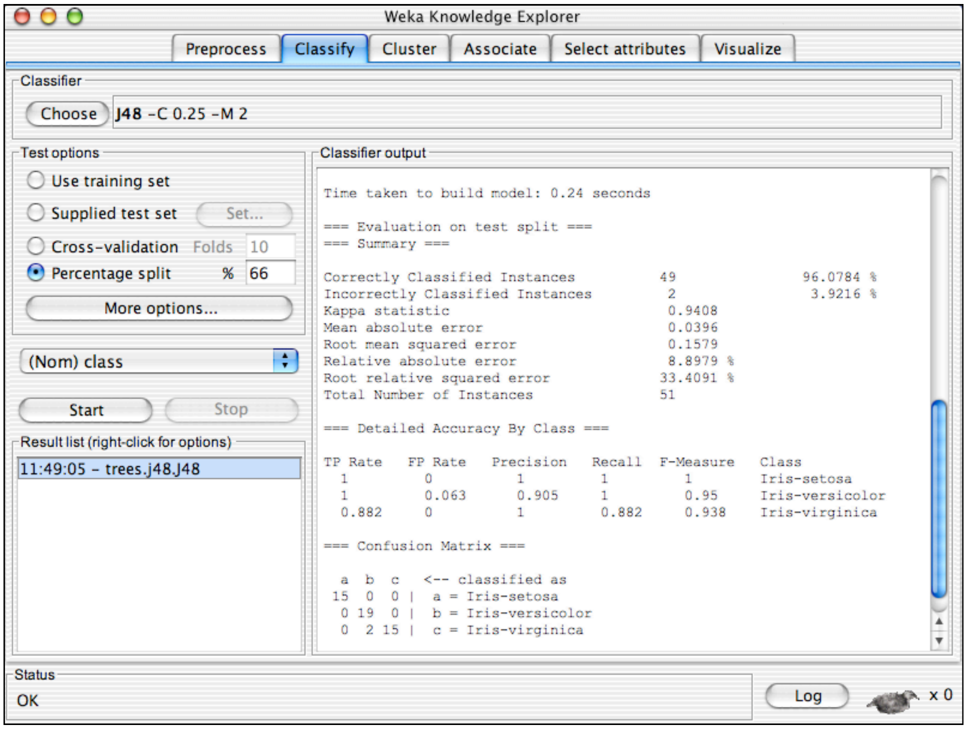
(Nom) class

Result list (right-click for options)
11:49:05 - trees.j48.J48

Status: OK

Classifier output

```
=== Run information ===

Scheme:       weka.classifiers.trees.j48.J48 -C 0.25 -M 2
Relation:     iris
Instances:    150
Attributes:   5
              sepallength
              sepalwidth
              petallength
              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
|   |   petallength <= 4.9: Iris-versicolor (48.0/1.0)
|   |   petallength > 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: J48 -C 0.25 -M 2

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

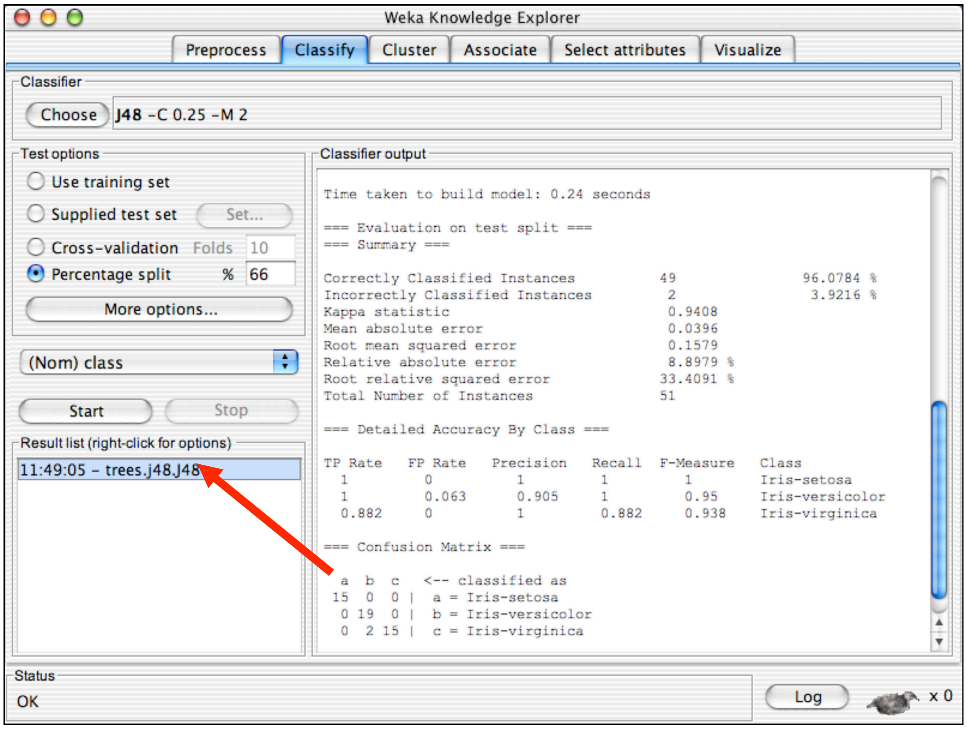
(Nom) class

Result list (right-click for options)
11:49:05 - trees.j48.J48

Status: OK

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 ===

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

=== 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

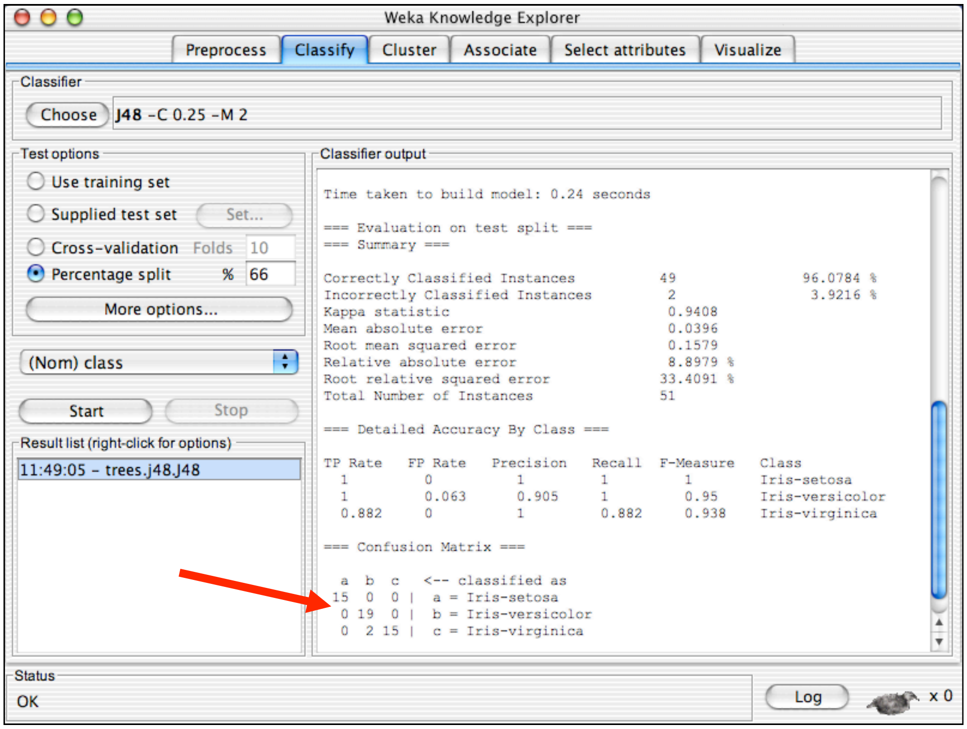
(Nom) class

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

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

=== 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



Weka Knowledge Explorer

Classifier: J48 -C 0.25 -M 2

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

(Nom) class

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

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

=== 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

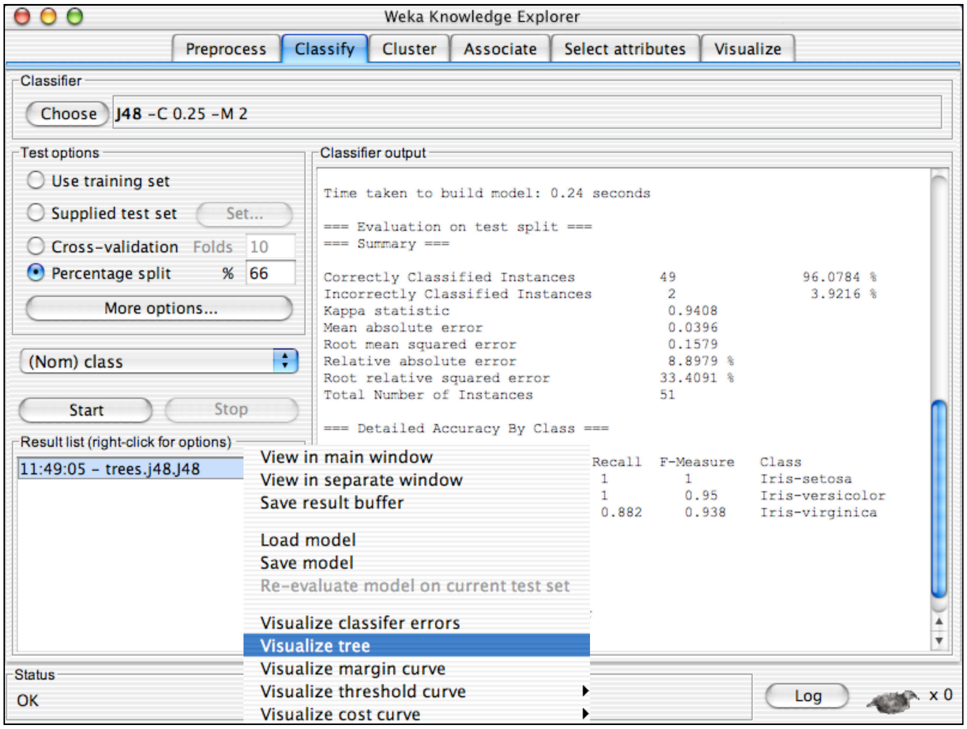
a	b	c	<-- classified as
15	0	0	a= Iris-setosa
0	19	0	b=Iris-versicolor
0	2	15	c=Iris-virginica

consider "TRUE"= iris-virginica and FALSE= everything else

	Ground truth	
	FALSE	TRUE
detector silent	A =34	B = 2
detector loud	C= 0	D = 15

accuracy	(A+D)/(A+B+C+D)	(34+15)/51	96%
recall (pd)	D/(B+D)	15/(2+15)	88%
false alarm (pf)	C/(A+C)	0/34	0%
precision	D/(C+D)	15/(15+0)	100%
f-measure	2*prec*pd/ (prec+pd)	2*1*0.88/ (1+0.88)	94%

Collect separately for each class.
Repeat 10 times (re-ordering data) * 10-way
Repeat for each learner * discretizer * x * y * ….



Classifier: J48 -C 0.25 -M 2

○ Cross-validation Folds 10
● Percentage split % 66

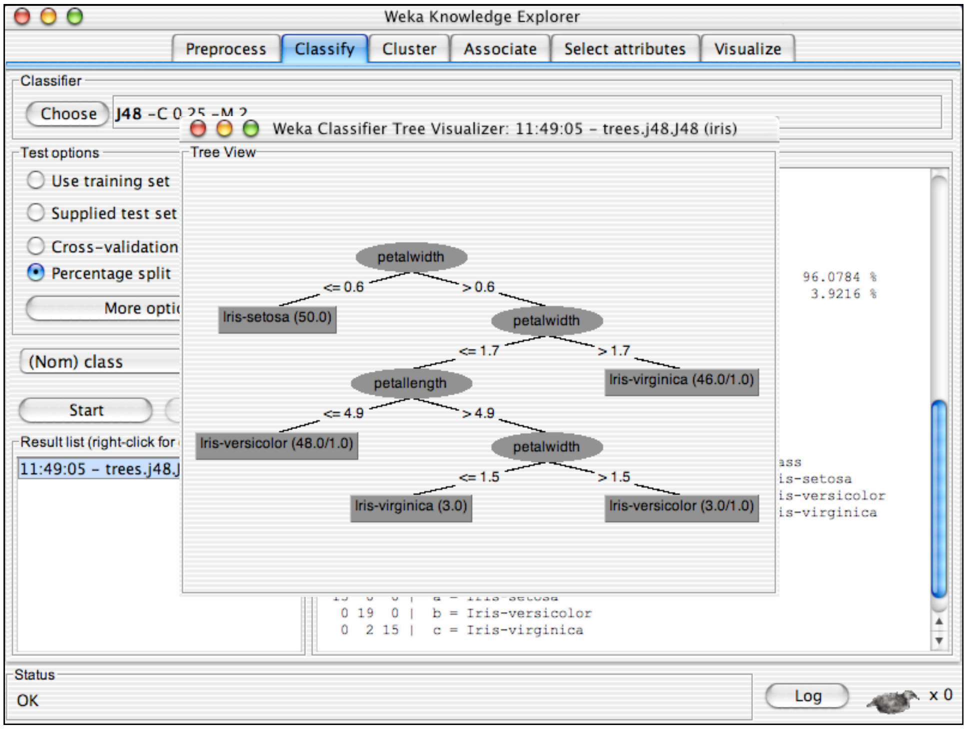
(Nom) class

11:49:05 - trees.j48.J48

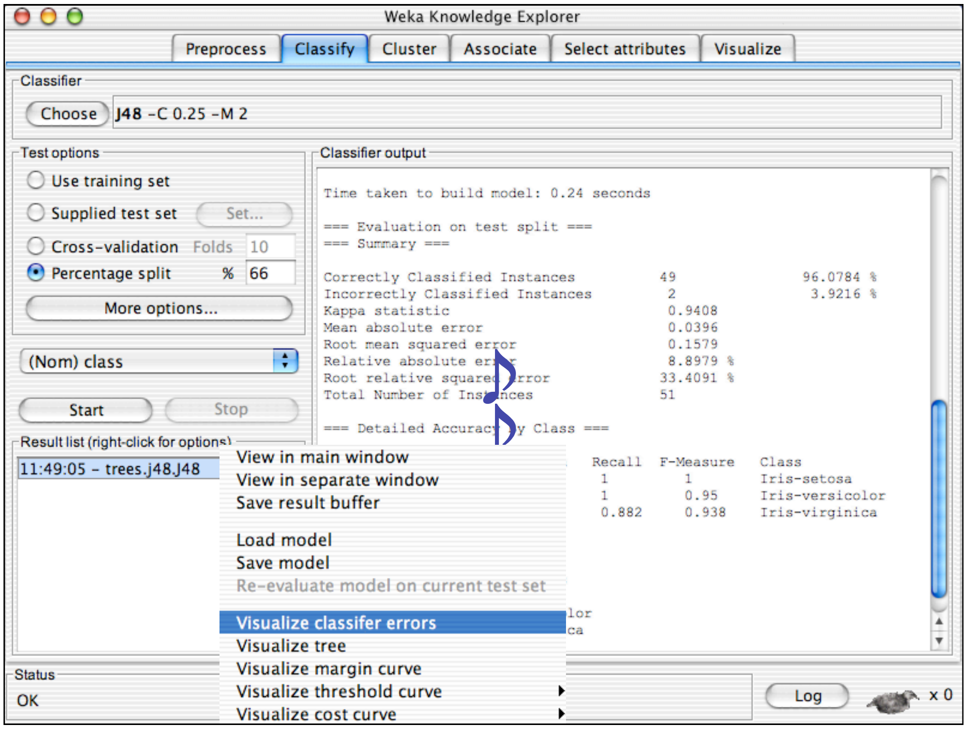
Status: OK

Classifier output

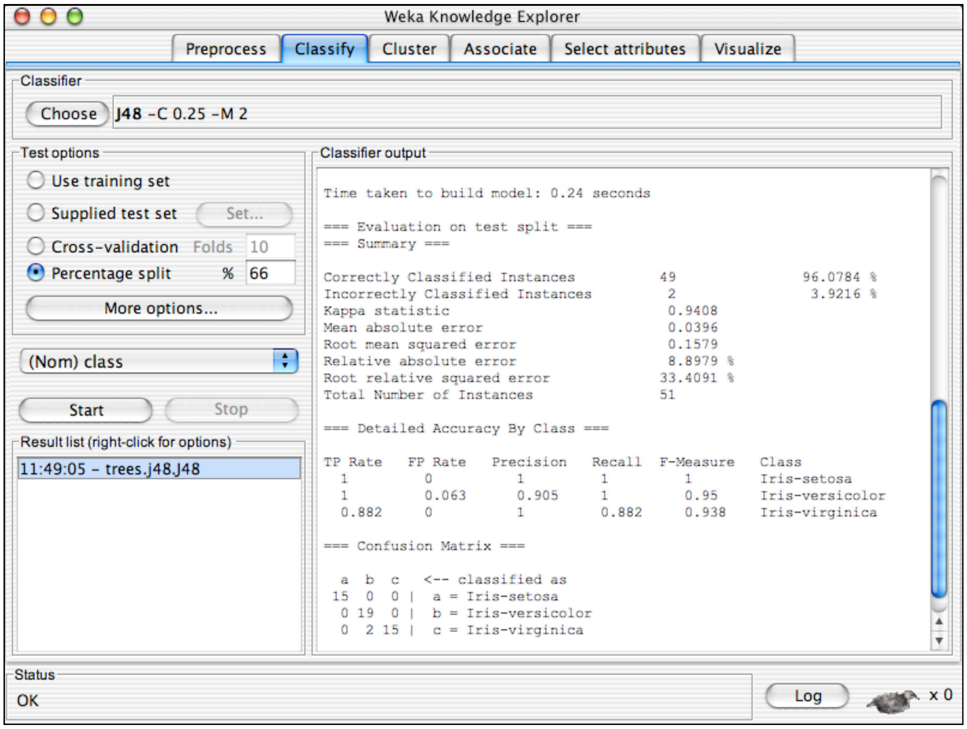
```
Time taken to build model: 0.24 seconds

=== 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 ===

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



```
96.0784 %
 3.9216 %

ass
is-setosa
is-versicolor
is-virginica

 0 19  0 |  b = Iris-versicolor
 0  2 15 |  c = Iris-virginica
```



```
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 ===

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

lor
ca
```



Classifier: J48 -C 0.25 -M 2

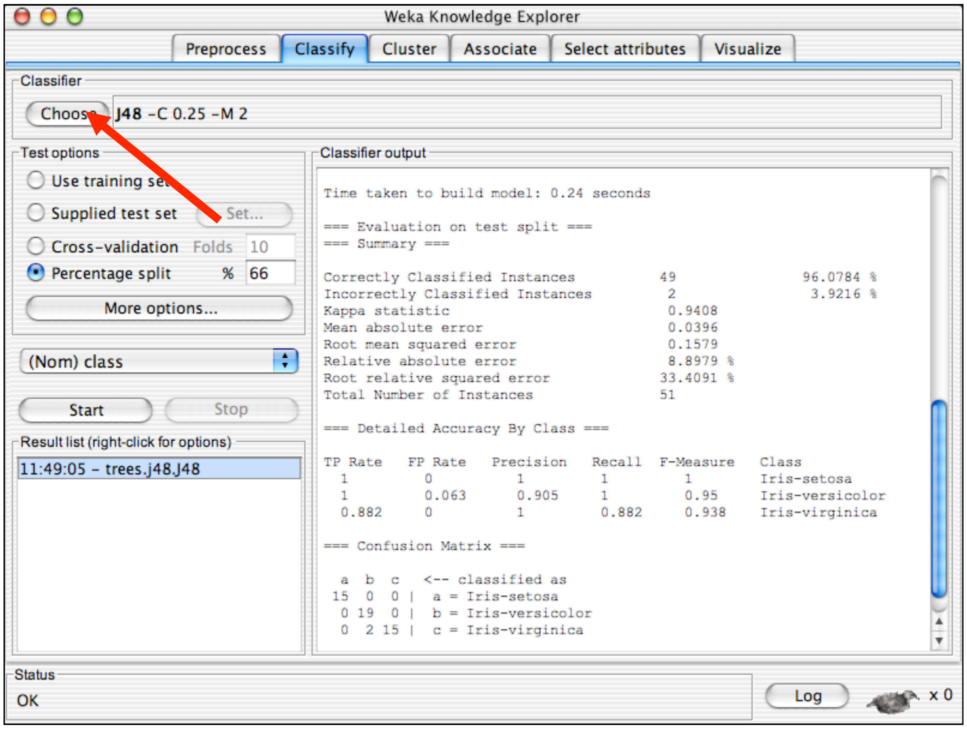
● Percentage split % 66

(Nom) class

11:49:05 - trees.j48.J48

Status: OK

```
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 ===

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

=== 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

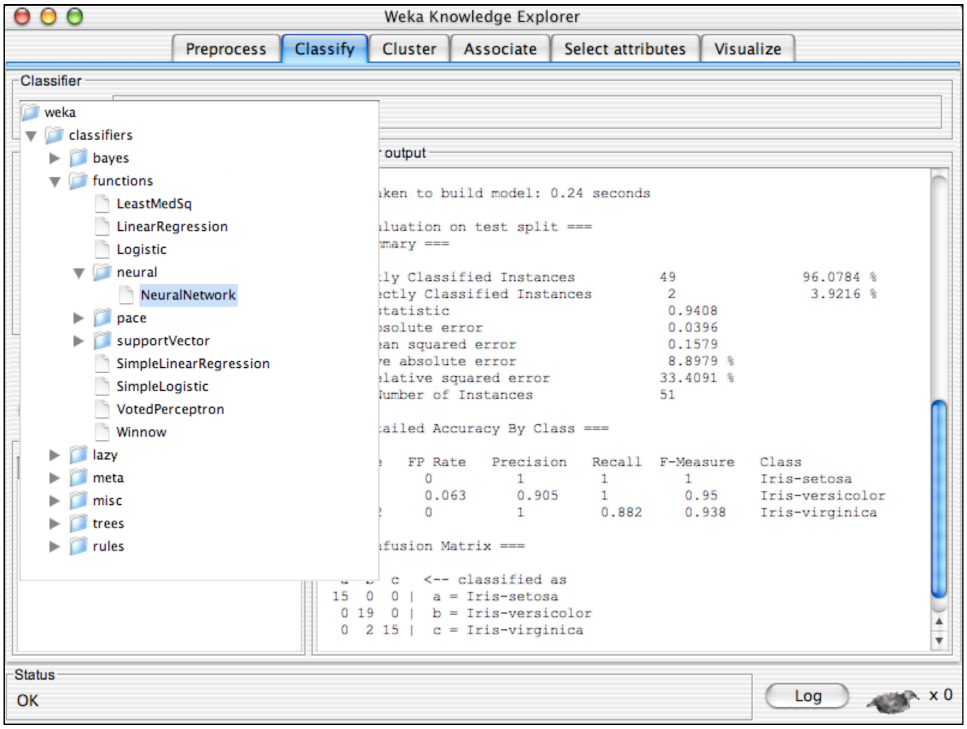
● Percentage split % 66

(Nom) class

11:49:05 - trees.j48.J48

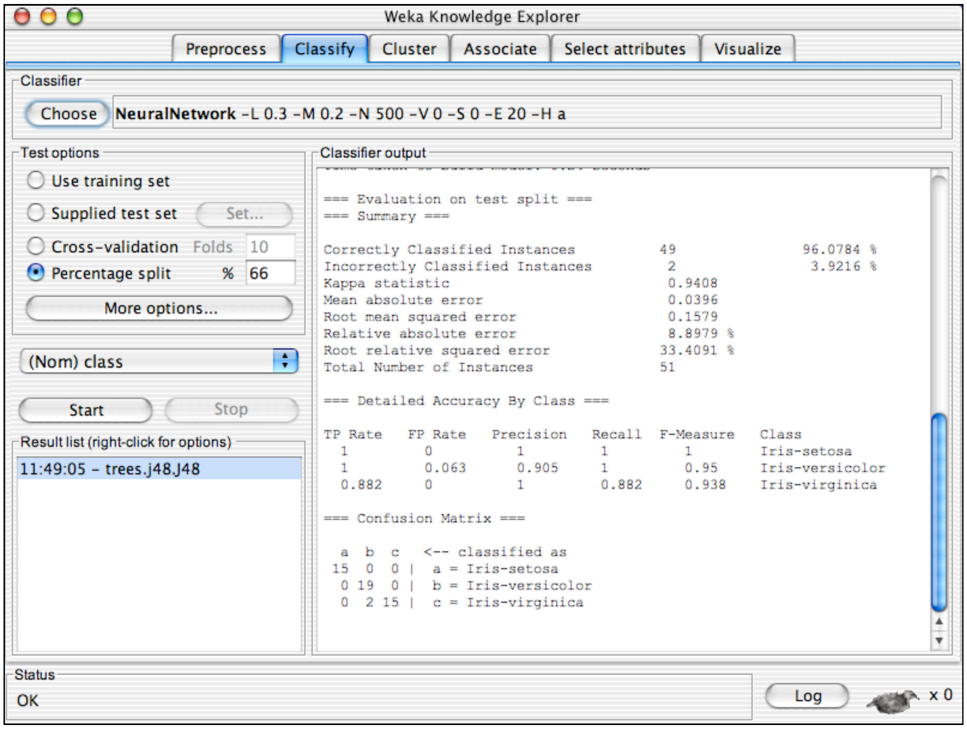
Status: OK

```
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 ===

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

=== 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
```



```
ken to build model: 0.24 seconds

luation on test split ===
mary ===

ly Classified Instances         49               96.0784 %
ctly Classified Instances        2                3.9216 %
tatistic                          0.9408
solute error                      0.0396
an squared error                  0.1579
e absolute error                  8.8979 %
lative squared error             33.4091 %
umber of Instances               51

ailed Accuracy By Class ===

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

fusion Matrix ===

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

OK



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

○ Use training set

○ Supplied test set

○ Cross-validation Folds 10

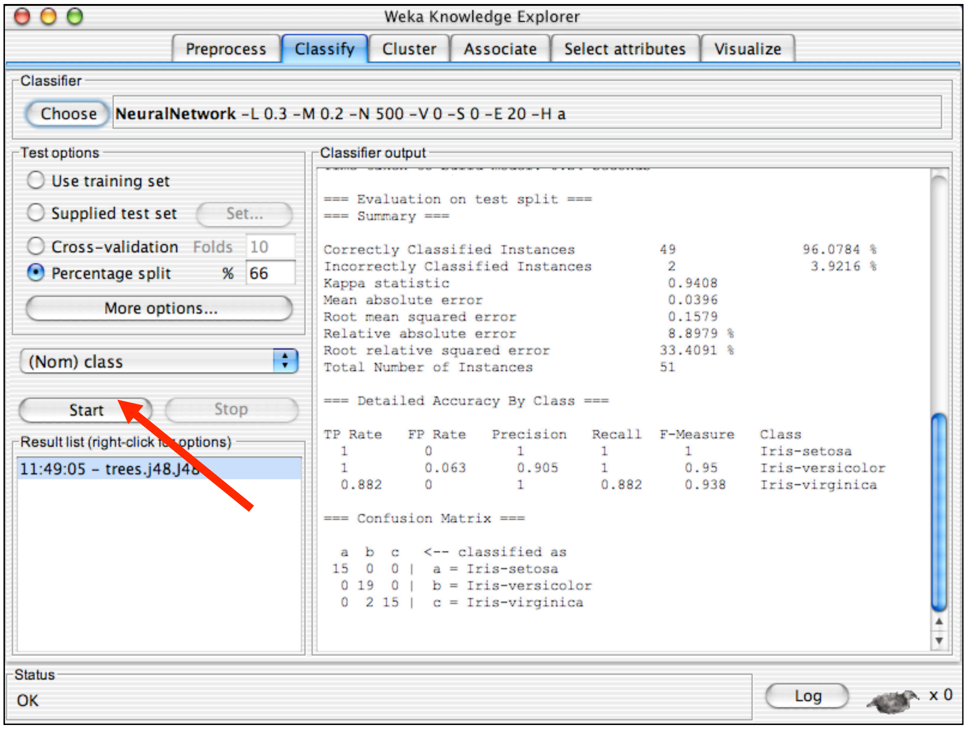
● Percentage split % 66

(Nom) class

11:49:05 - trees.j48.J48

```
=== 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 ===

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

=== 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
```

OK



Classifier

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 Set...

○ Cross-validation Folds 10

● Percentage split % 66

(Nom) class

Result list (right-click for options)

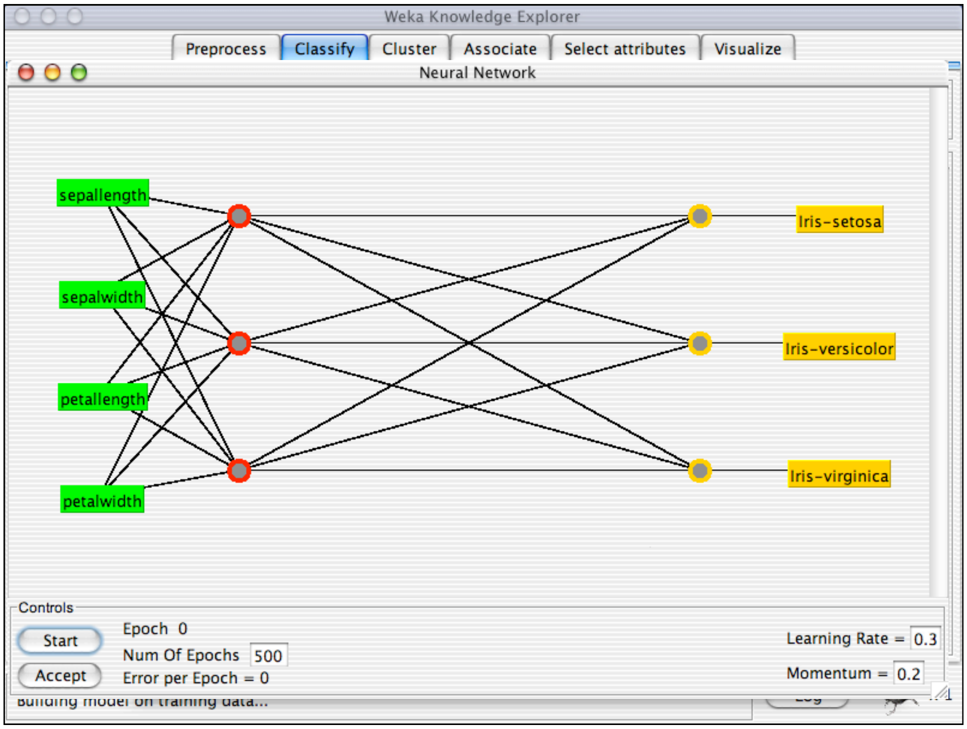
11:49:05 - trees.j48.J48

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 ===

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

=== 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



Epoch 0

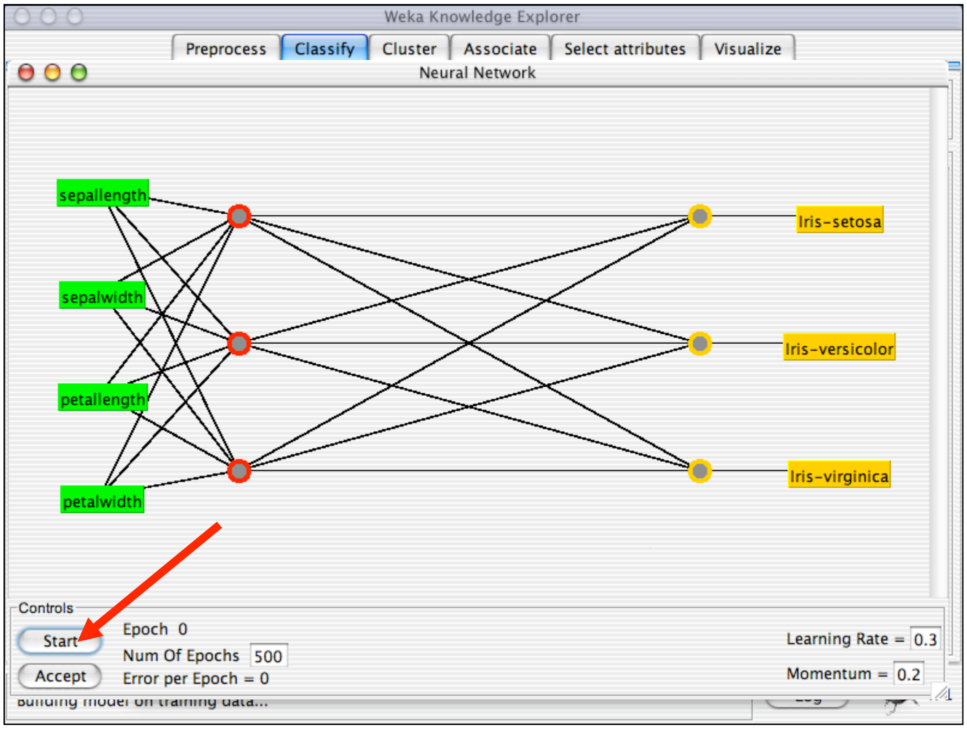
Num Of Epochs 500

Error per Epoch = 0

Learning Rate = 0.3

Momentum = 0.2

Building model on training data...



Epoch 0

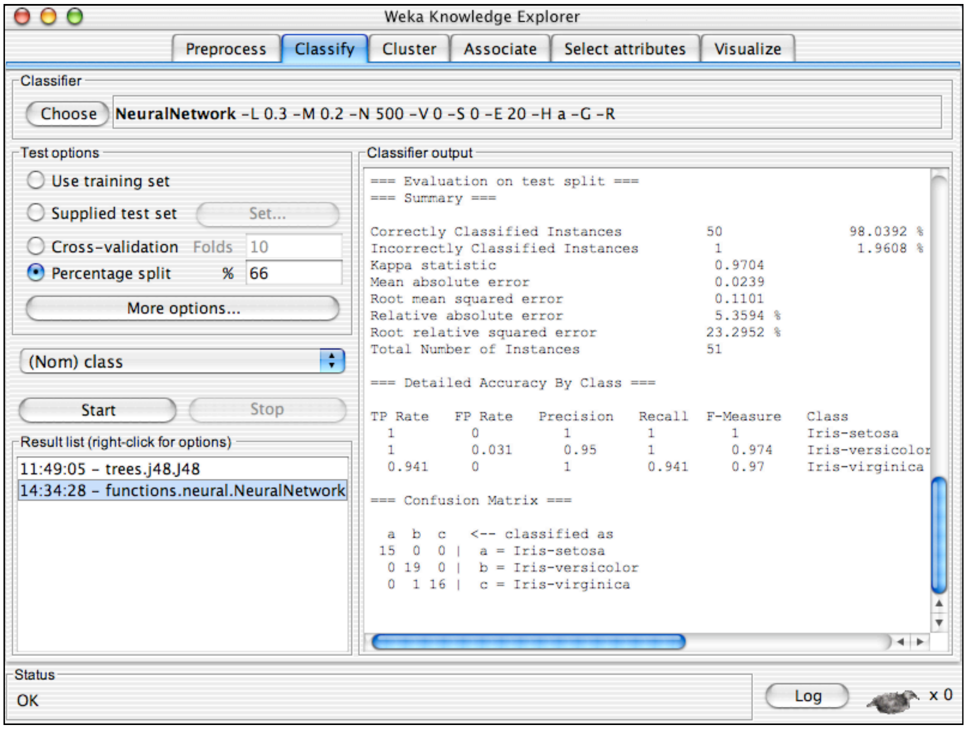
Num Of Epochs 500

Error per Epoch = 0

Learning Rate = 0.3

Momentum = 0.2

Building model on training data...



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

○ Use training set

○ Supplied test set

○ Cross-validation Folds 10

● Percentage split % 66

(Nom) class

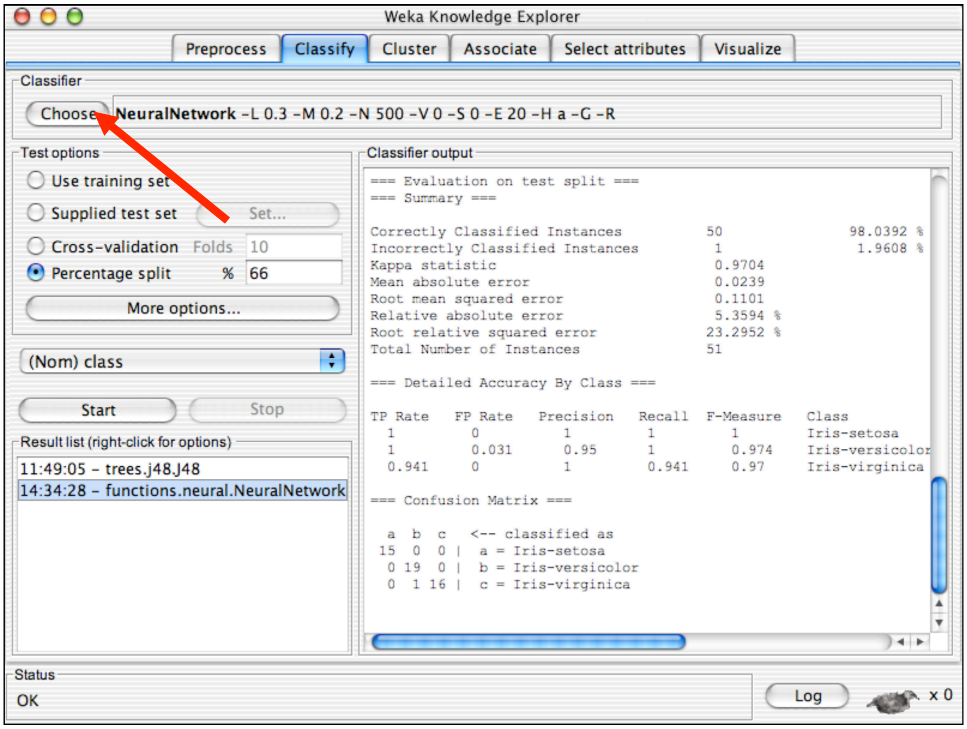
11:49:05 - trees.j48.J48

14:34:28 - functions.neural.NeuralNetwork

```
=== 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        Iris-setosa
  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
```

Status

OK



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 Set...
- ○ Cross-validation Folds 10
- ● Percentage split % 66

(Nom) class

Result list (right-click for options)

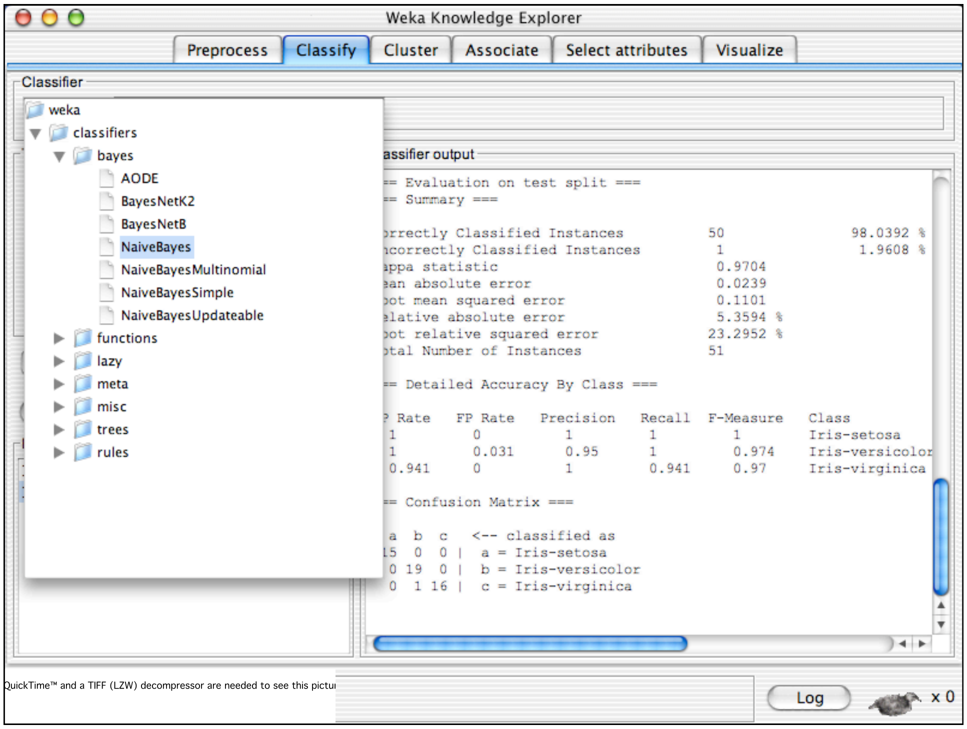
- 11:49:05 - trees.j48.J48
- 14:34:28 - functions.neural.NeuralNetwork

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       Iris-setosa
  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
```

Status

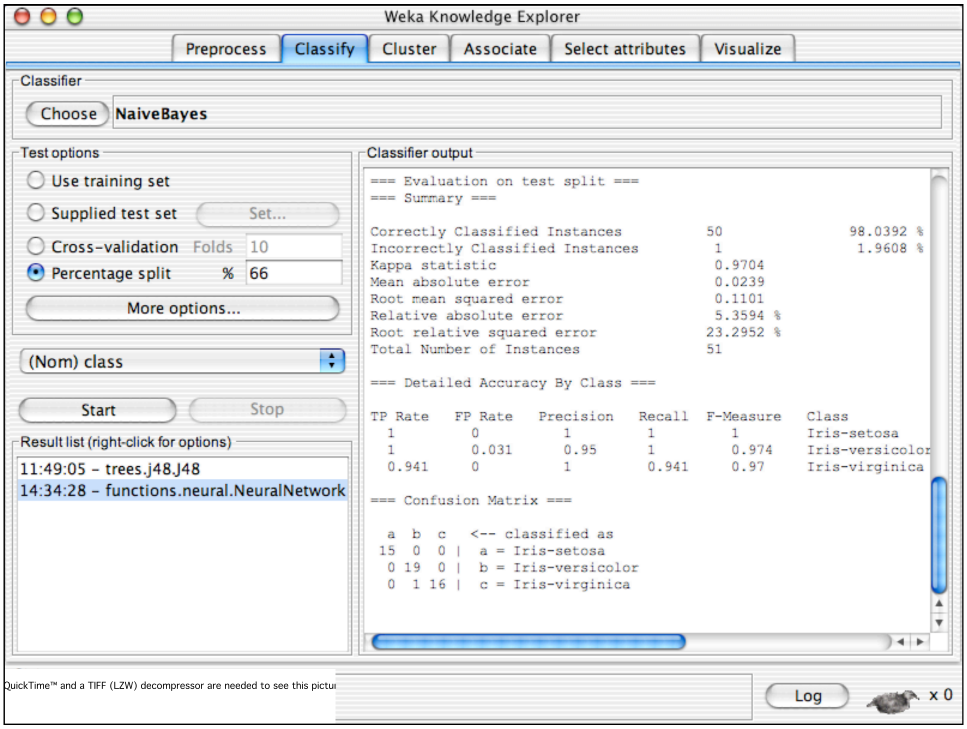
OK



assifier output

```
== Evaluation on test split ===
== Summary ===

orrectly Classified Instances          50               98.0392 %
ncorrectly Classified Instances         1                1.9608 %
appa statistic                          0.9704
ean absolute error                      0.0239
oot mean squared error                  0.1101
elative absolute error                  5.3594 %
oot relative squared error             23.2952 %
otal Number of Instances               51

== Detailed Accuracy By Class ===

 Rate   FP Rate   Precision   Recall  F-Measure   Class
1         0          1          1         1       Iris-setosa
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
```

QuickTime™ and a TIFF (LZW) decompressor are needed to see this pictu



Classifier: NaiveBayes

Test options

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

(Nom) class

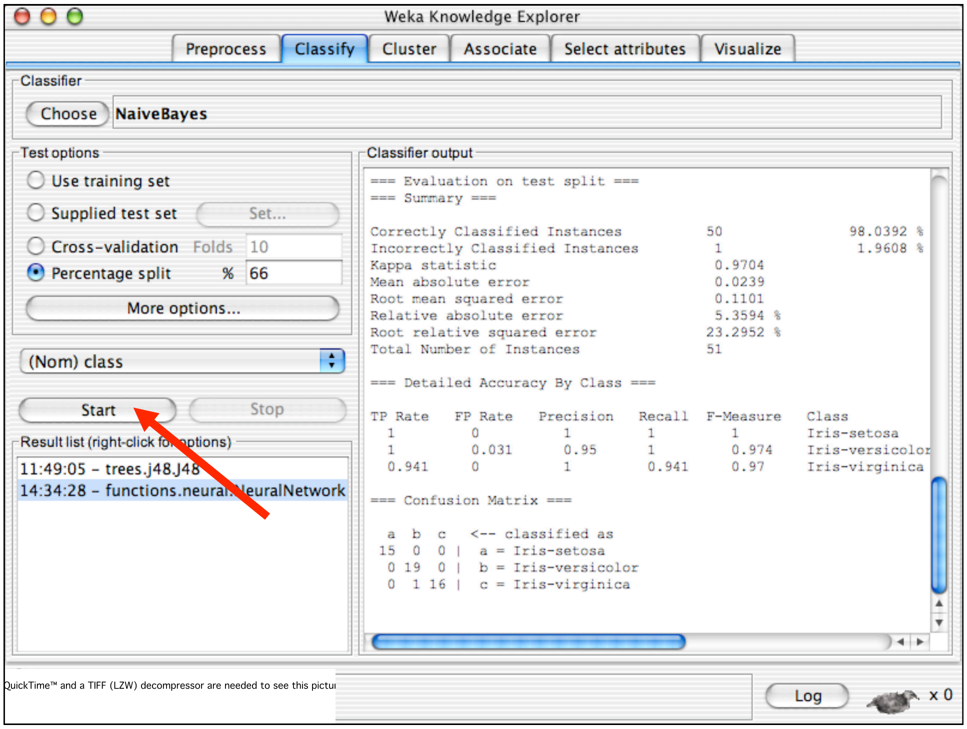
Result list (right-click for options)

- 11:49:05 - trees.j48.J48
- 14:34:28 - functions.neural.NeuralNetwork

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         Iris-setosa
  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
```

QuickTime™ and a TIFF (LZW) decompressor are needed to see this pictu


Classifier: NaiveBayes

Test options

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

(Nom) class

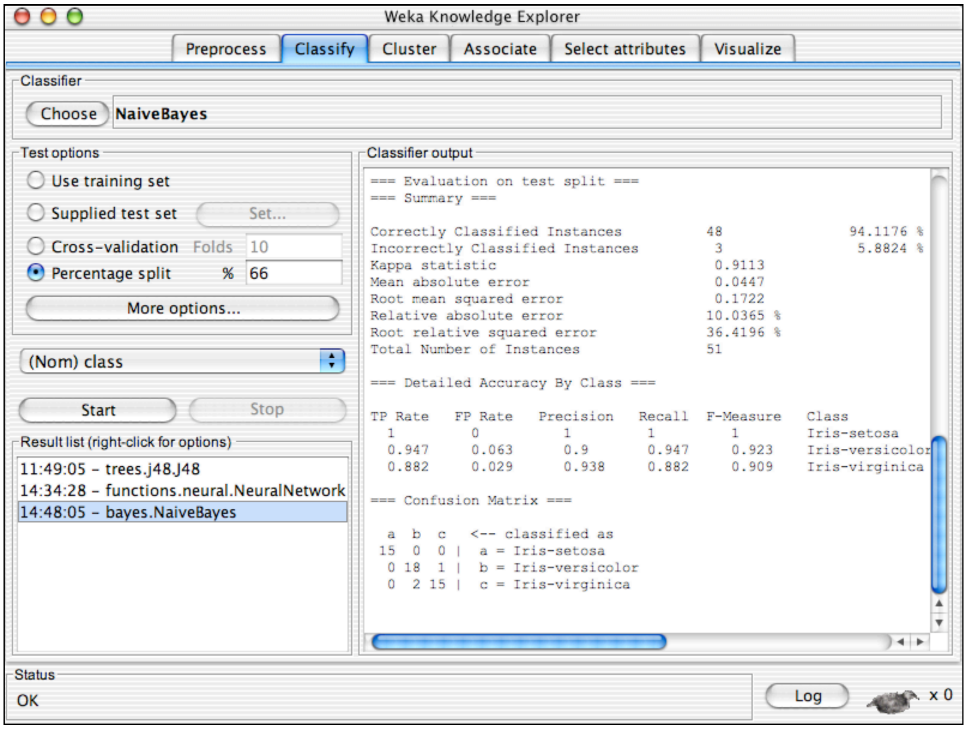
Result list (right-click for options)

- 11:49:05 - trees.j48.J48
- 14:34:28 - functions.neural.NeuralNetwork

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         Iris-setosa
  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
```

QuickTime™ and a TIFF (LZW) decompressor are needed to see this pictu



Classifier: NaiveBayes

Test options

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

(Nom) class

Result list (right-click for options)

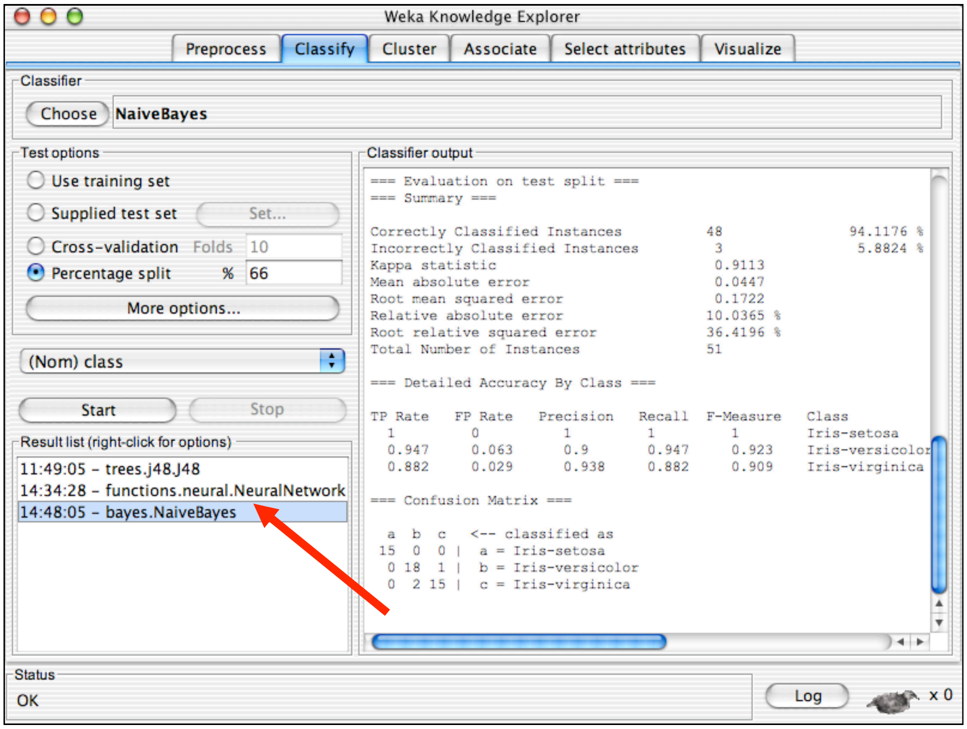
- 11:49:05 - trees.j48.J48
- 14:34:28 - functions.neural.NeuralNetwork
- 14:48:05 - bayes.NaiveBayes

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 ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  1         0          1          1         1        Iris-setosa
  0.947     0.063      0.9        0.947     0.923    Iris-versicolor
  0.882     0.029      0.938      0.882     0.909    Iris-virginica

=== 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
```

Status

OK


Classifier: NaiveBayes

Test options

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

(Nom) class

Result list (right-click for options)

- 11:49:05 - trees.j48.J48
- 14:34:28 - functions.neural.NeuralNetwork
- 14:48:05 - bayes.NaiveBayes

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 ===

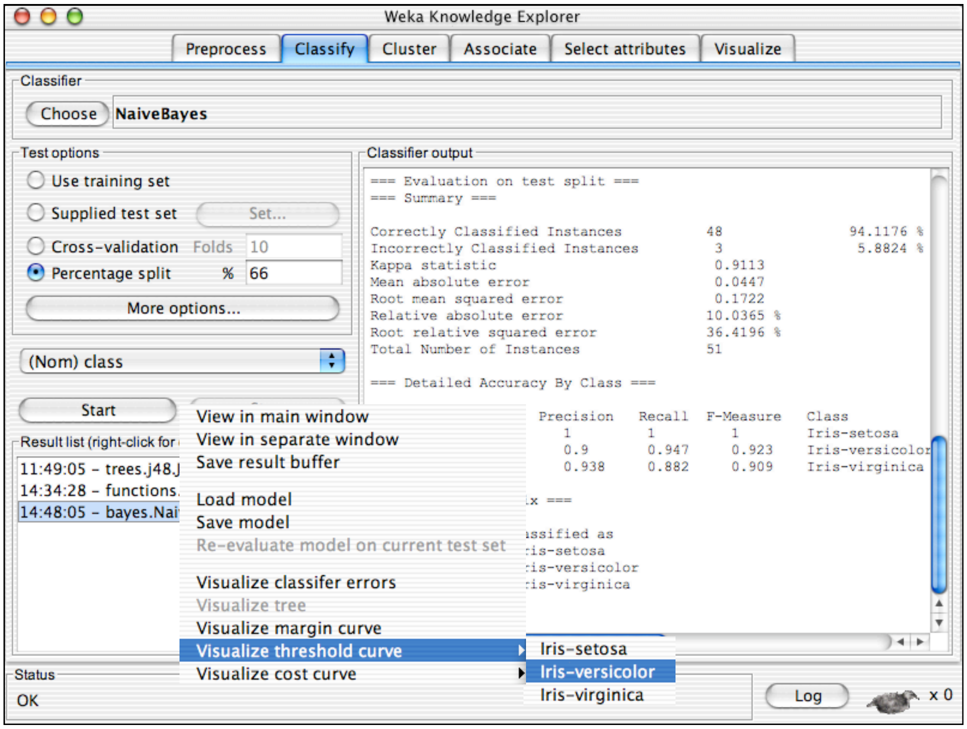
TP Rate FP Rate Precision Recall F-Measure Class
1 0 1 1 1 Iris-setosa
0.947 0.063 0.9 0.947 0.923 Iris-versicolor
0.882 0.029 0.938 0.882 0.909 Iris-virginica

=== 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

Status

OK



Classifier: NaiveBayes

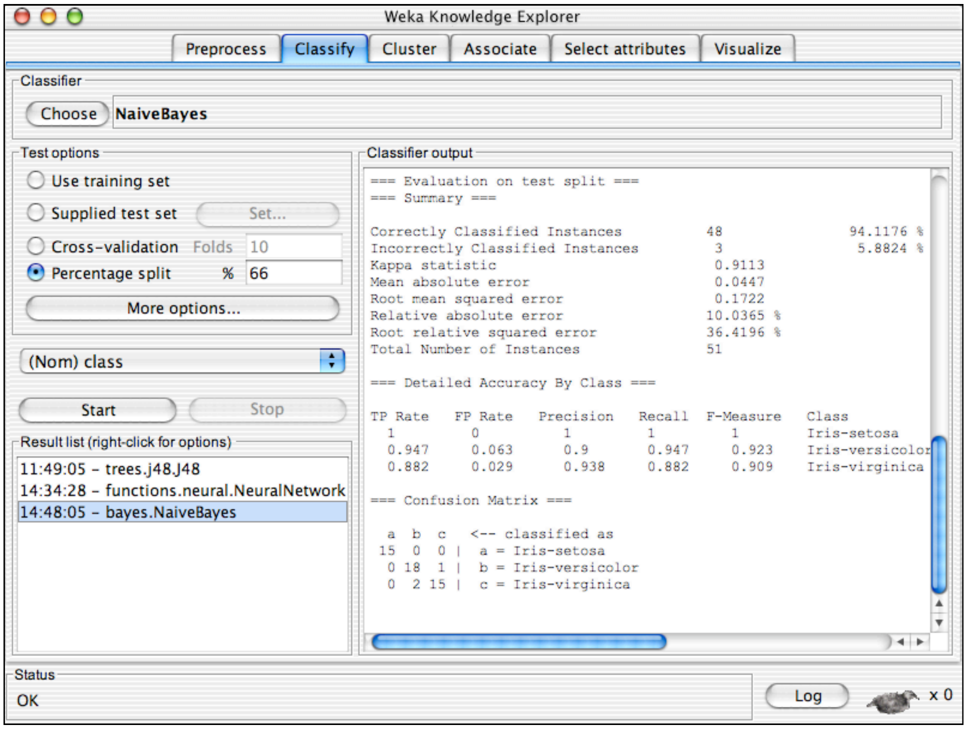
Test options
- ○ Use training set
- ○ Supplied test set
- ○ Cross-validation Folds 10
- ● Percentage split % 66

(Nom) class

Result list (right-click for
- 11:49:05 - trees.j48.J
- 14:34:28 - functions.
- 14:48:05 - bayes.Nai

Status: OK

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 ===

                     Precision   Recall  F-Measure   Class
                       1         1         1         Iris-setosa
                       0.9       0.947     0.923     Iris-versicolor
                       0.938     0.882     0.909     Iris-virginica

                     x ===

                     assified as
                     is-setosa
                     is-versicolor
                     is-virginica
```



Classifier: NaiveBayes

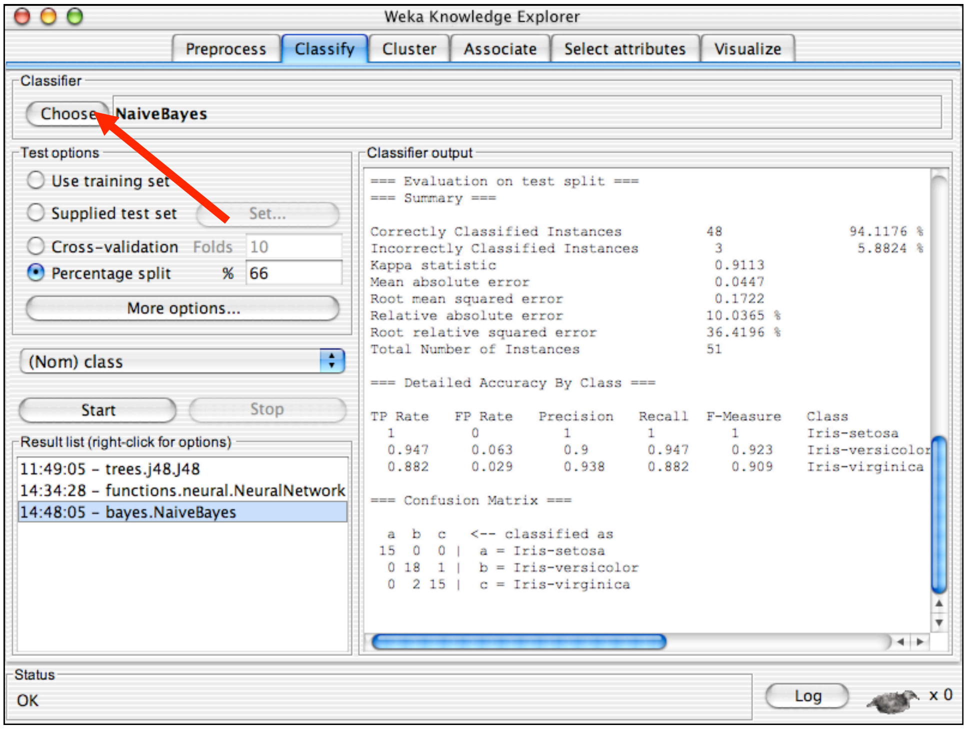
Test options
- ○ Use training set
- ○ Supplied test set
- ○ Cross-validation Folds 10
- ● Percentage split % 66

(Nom) class

Result list (right-click for options)
- 11:49:05 - trees.j48.J48
- 14:34:28 - functions.neural.NeuralNetwork
- 14:48:05 - bayes.NaiveBayes

Status: OK

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 ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  1         0         1         1         1         Iris-setosa
  0.947     0.063     0.9       0.947     0.923     Iris-versicolor
  0.882     0.029     0.938     0.882     0.909     Iris-virginica

=== 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
```



Classifier: NaiveBayes

Test options

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

(Nom) class

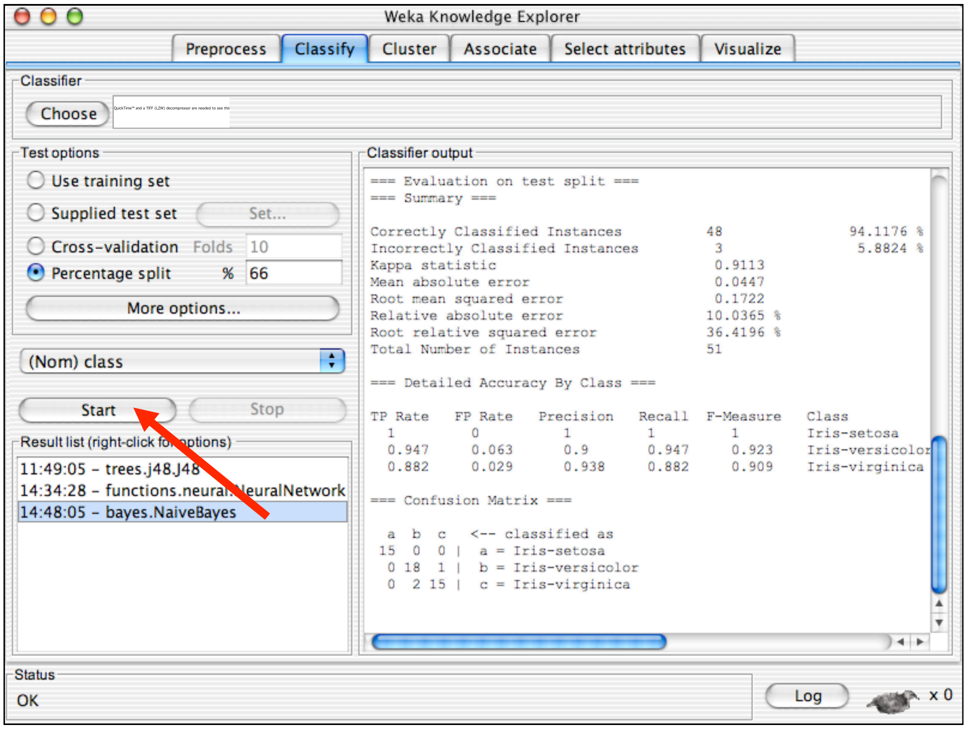
Result list (right-click for options)

- 11:49:05 - trees.j48.J48
- 14:34:28 - functions.neural.NeuralNetwork
- 14:48:05 - bayes.NaiveBayes

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 ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  1         0         1           1       1         Iris-setosa
  0.947     0.063     0.9         0.947   0.923     Iris-versicolor
  0.882     0.029     0.938       0.882   0.909     Iris-virginica

=== 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
```

Status: OK



Test options

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

(Nom) class

Result list (right-click for options)

- 11:49:05 - trees.j48.J48
- 14:34:28 - functions.neural.NeuralNetwork
- 14:48:05 - bayes.NaiveBayes

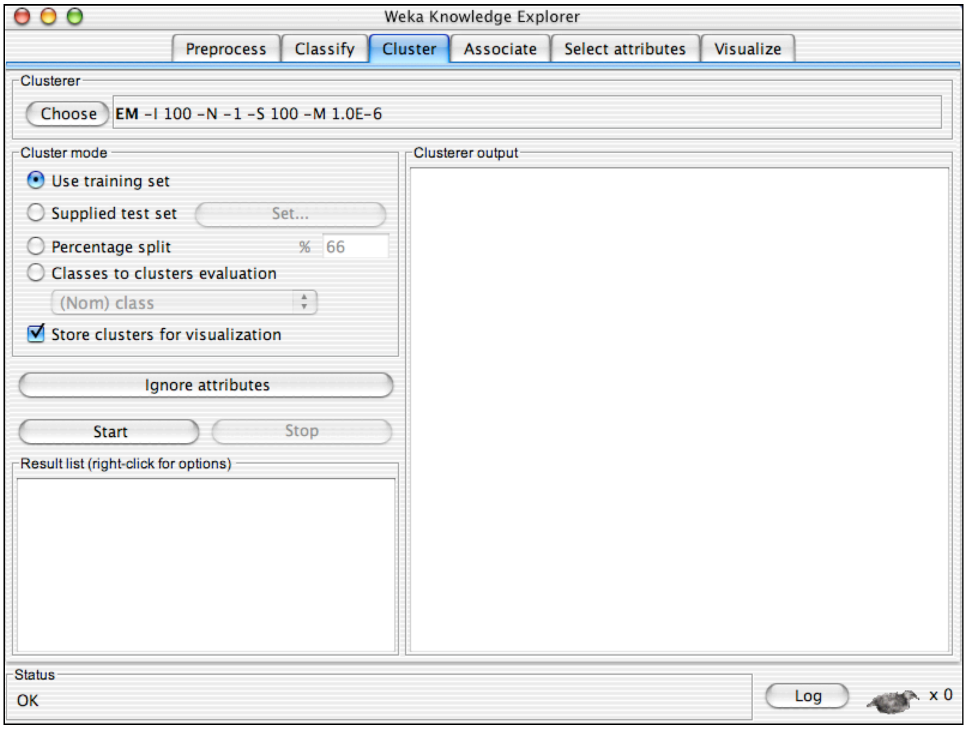
Classifier output

```
=== 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 ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  1         0         1           1       1         Iris-setosa
  0.947     0.063     0.9         0.947   0.923     Iris-versicolor
  0.882     0.029     0.938       0.882   0.909     Iris-virginica

=== 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
```

Status: OK

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 “t rue” clusters (if given)
- Evaluation based on loglikelihood if clustering scheme produces a probability distribution



Clusterer

EM -I 100 -N -1 -S 100 -M 1.0E-6

Cluster mode

● Use training set

○ Supplied test set

○ Percentage split % 66

○ Classes to clusters evaluation

(Nom) class

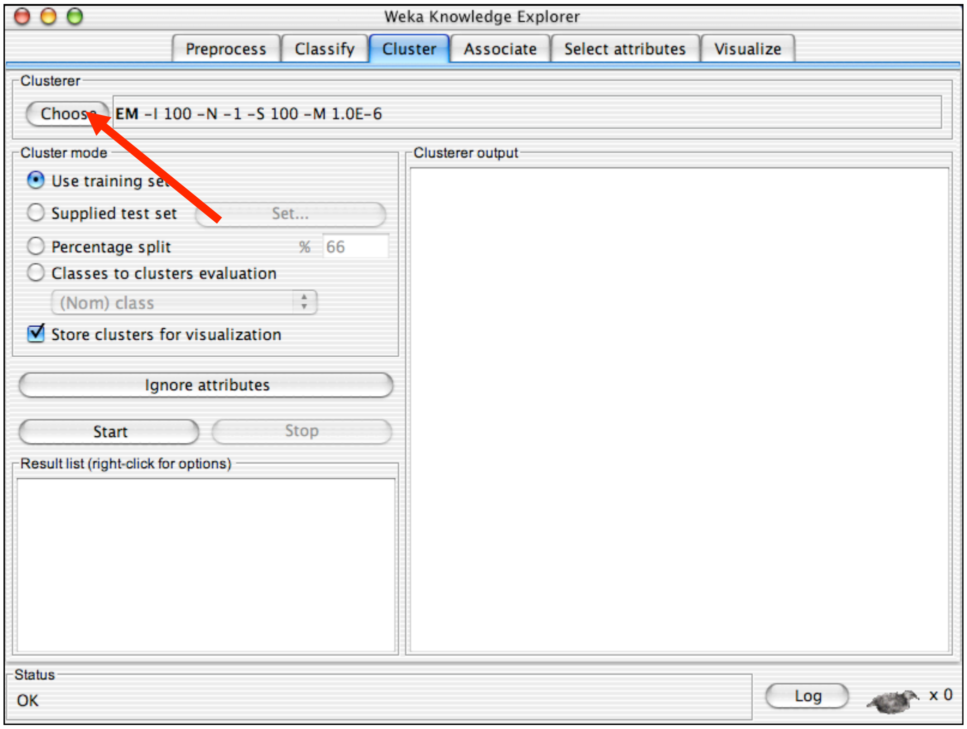
☑ Store clusters for visualization

Result list (right-click for options)

Clusterer output

Status

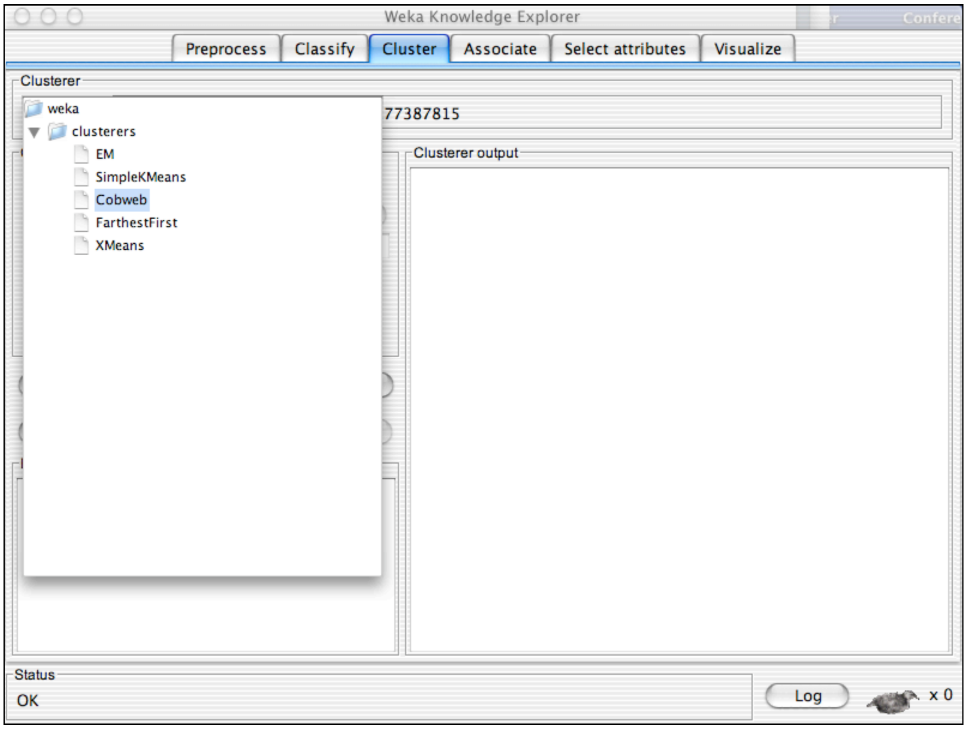
OK



EM -I 100 -N -1 -S 100 -M 1.0E-6

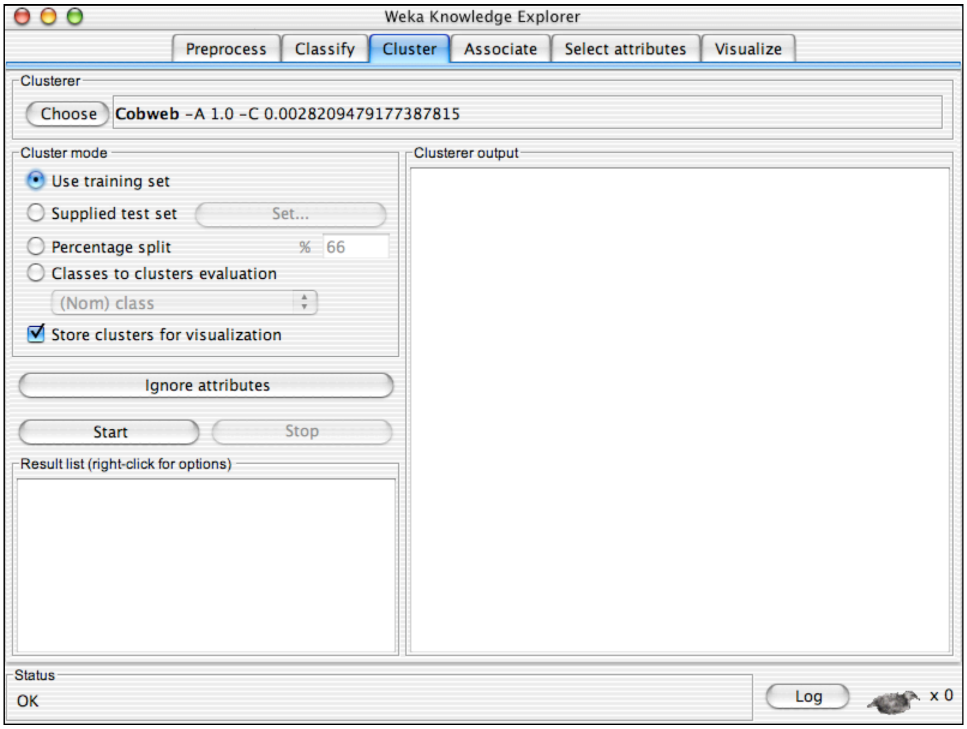
● Use training set
○ Supplied test set
○ Percentage split % 66
○ Classes to clusters evaluation
(Nom) class
☑ Store clusters for visualization

OK



77387815

OK



Clusterer

Cobweb -A 1.0 -C 0.0028209479177387815

Cluster mode

● Use training set

○ Supplied test set

○ Percentage split % 66

○ Classes to clusters evaluation

(Nom) class

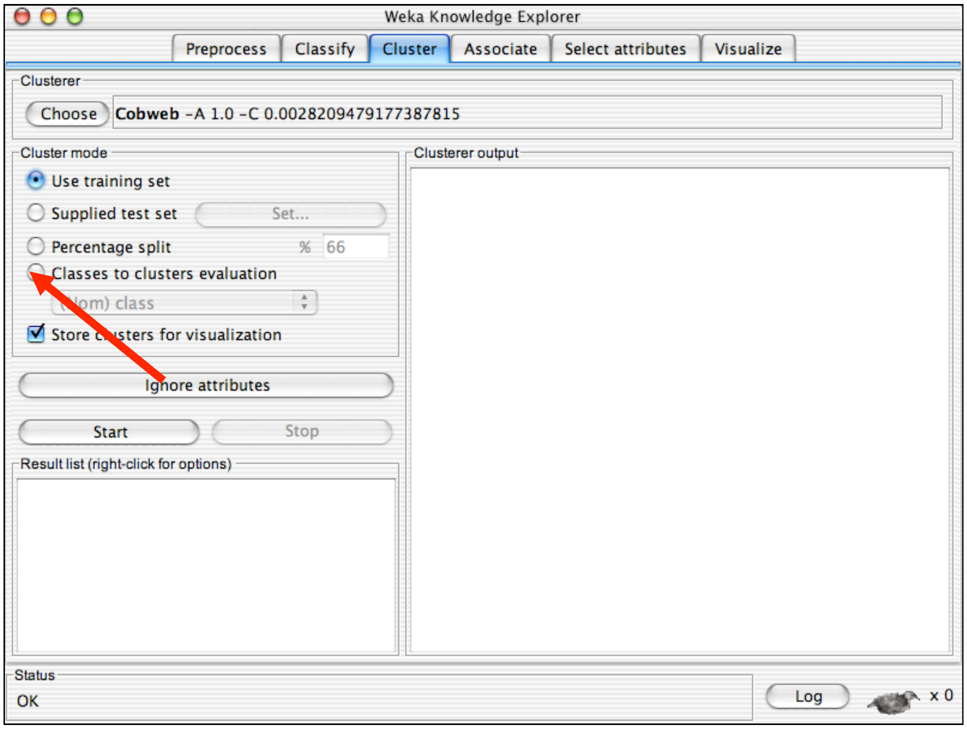
☑ Store clusters for visualization

Clusterer output

Result list (right-click for options)

Status

OK



Clusterer

Cobweb -A 1.0 -C 0.0028209479177387815

Cluster mode

● Use training set

○ Supplied test set

○ Percentage split % 66

○ Classes to clusters evaluation

(Nom) class

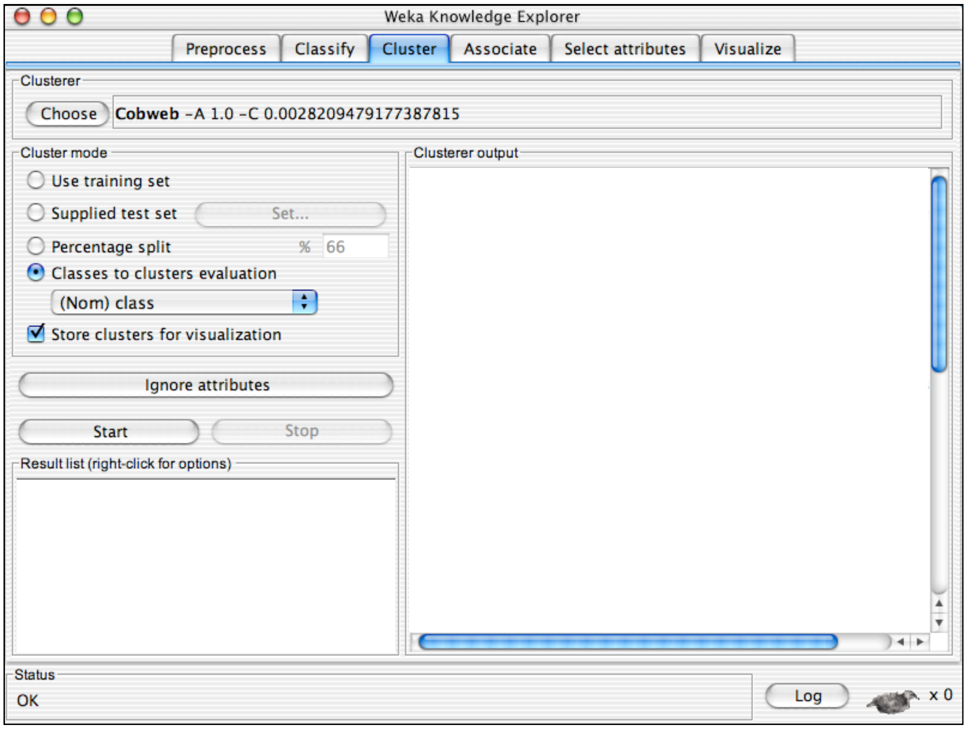
☑ Store clusters for visualization

Clusterer output

Result list (right-click for options)

Status

OK



Weka Knowledge Explorer

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

Clusterer

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

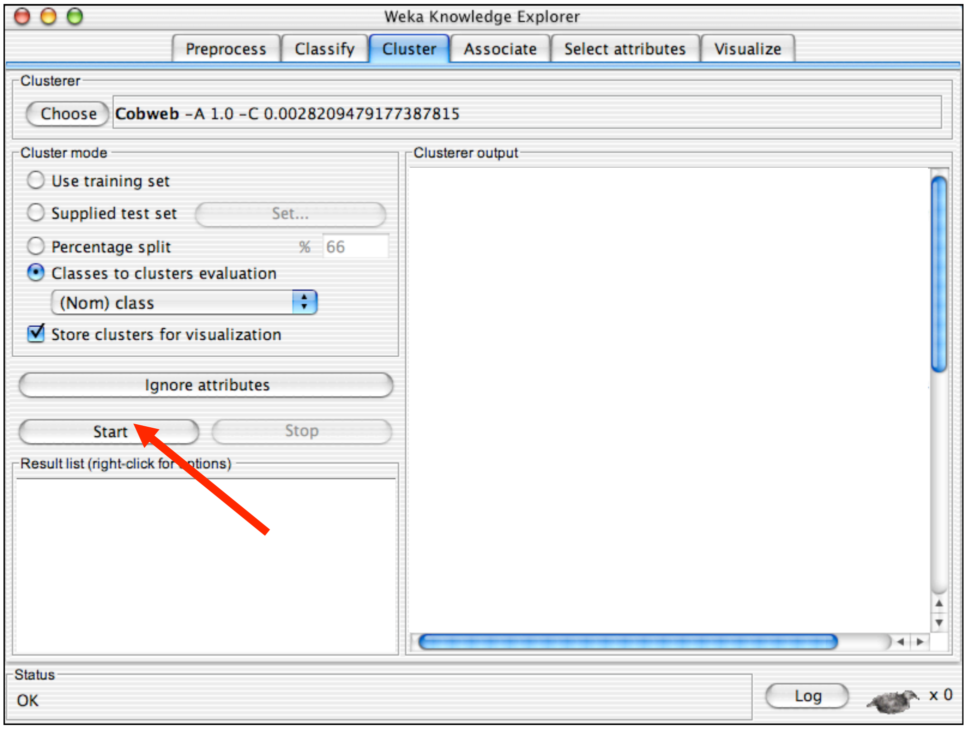
Start   Stop

Result list (right-click for options)

Clusterer output

Status

OK



Weka Knowledge Explorer

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

Clusterer

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

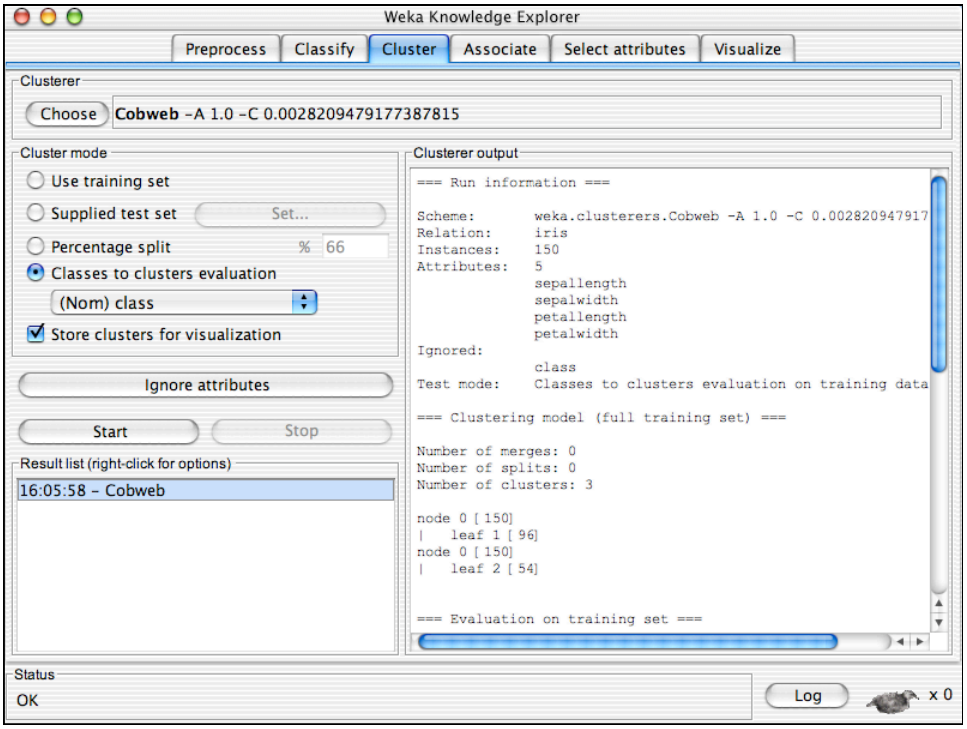
Start   Stop

Result list (right-click for options)

Clusterer output

Status

OK



Clusterer

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

Result list (right-click for options)

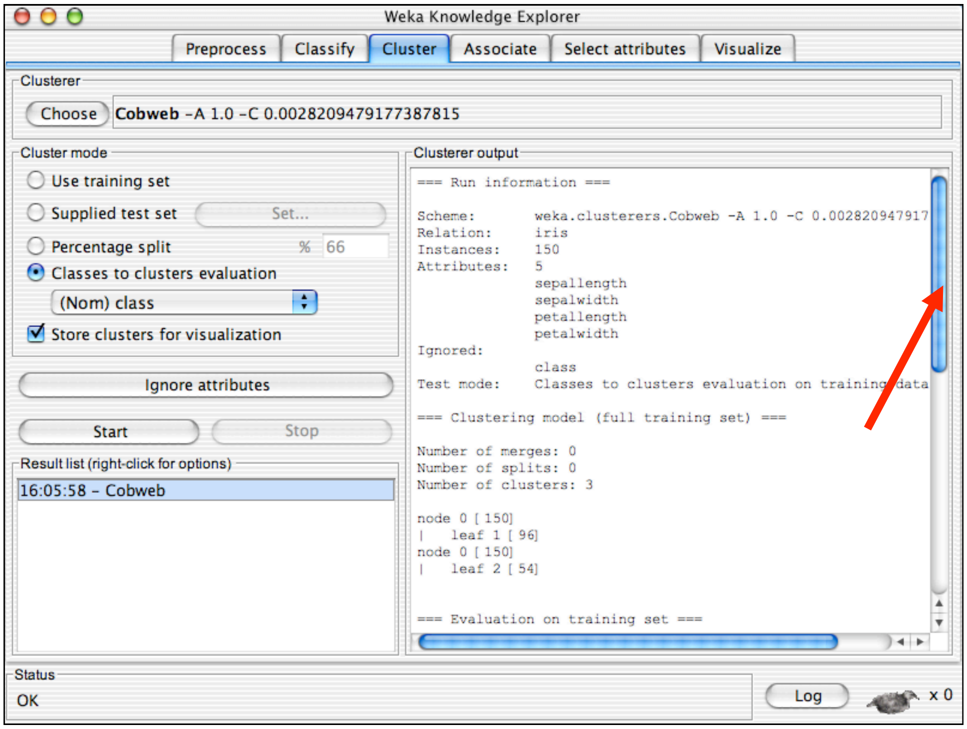
16:05:58 - Cobweb

Clusterer output

```
=== Run information ===

Scheme:       weka.clusterers.Cobweb -A 1.0 -C 0.002820947917
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



Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Choose
Set...
Ignore attributes
Start
Stop
Log
x 0

Clusterer

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

Result list (right-click for options)

16:05:58 - Cobweb

Clusterer output

=== Run information ===

Scheme: weka.clusterers.Cobweb -A 1.0 -C 0.002820947917
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) ===

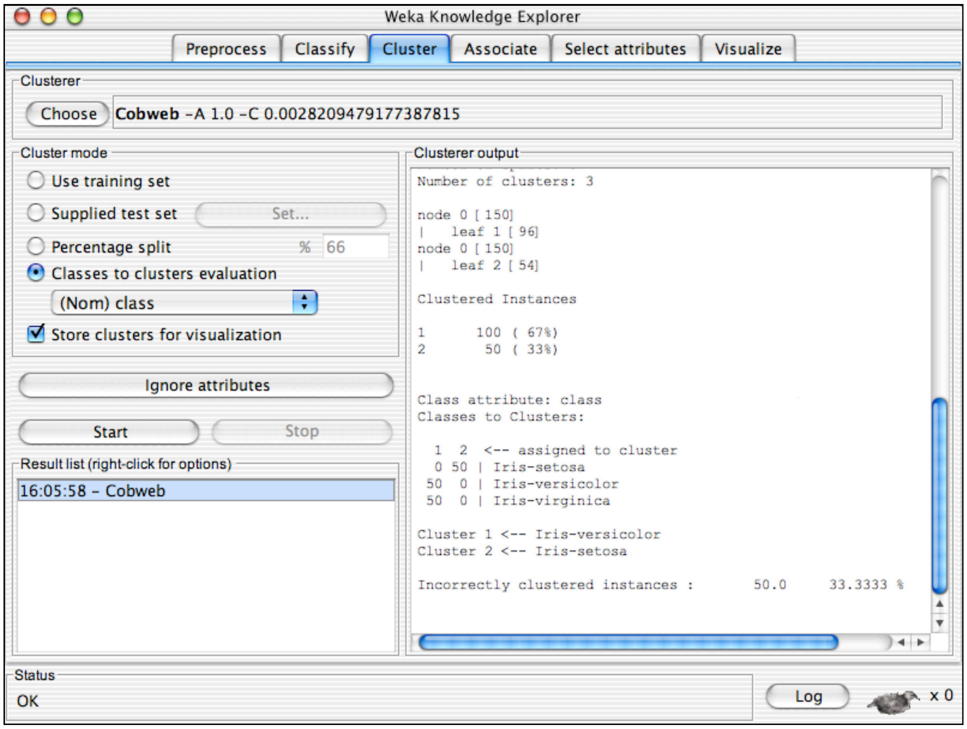
Number of merges: 0
Number of splits: 0
Number of clusters: 3

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

=== Evaluation on training set ===

Status

OK



Clusterer: 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

Result list (right-click for options)

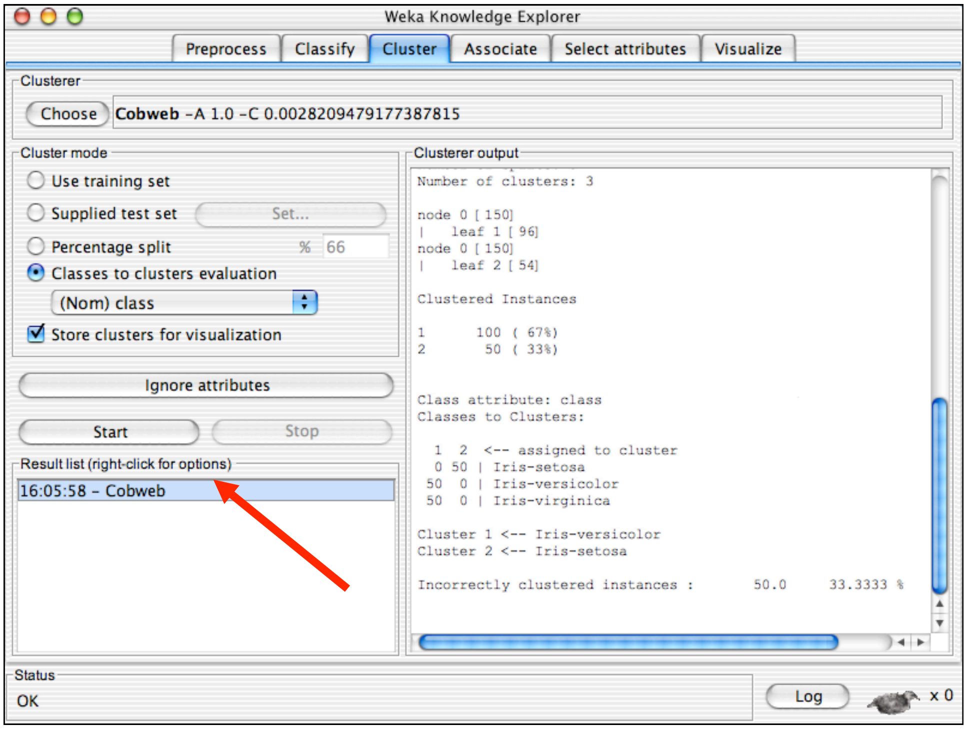
16:05:58 - Cobweb

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



Weka Knowledge Explorer

Clusterer: 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

Result list (right-click for options)

16:05:58 - Cobweb

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:

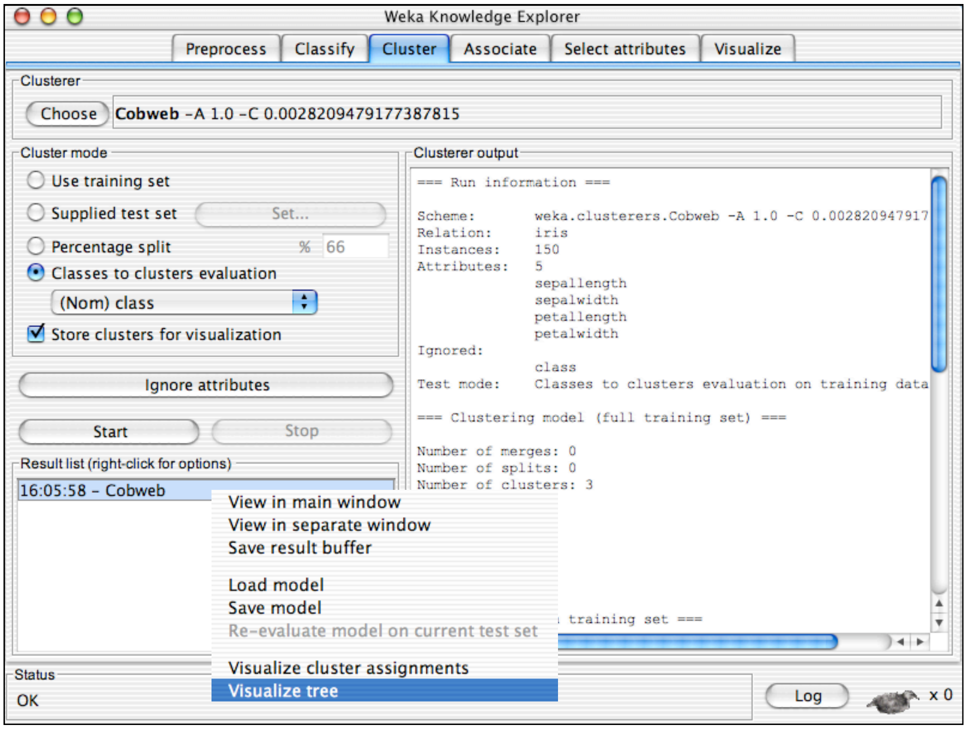
  1 2 <-- assigned to cluster
  0 50 | Iris-setosa
 50 0 | Iris-versicolor
 50 0 | Iris-virginica

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

Incorrectly clustered instances :   50.0   33.3333 %

Status

OK



Clusterer: **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

Result list (right-click for options)

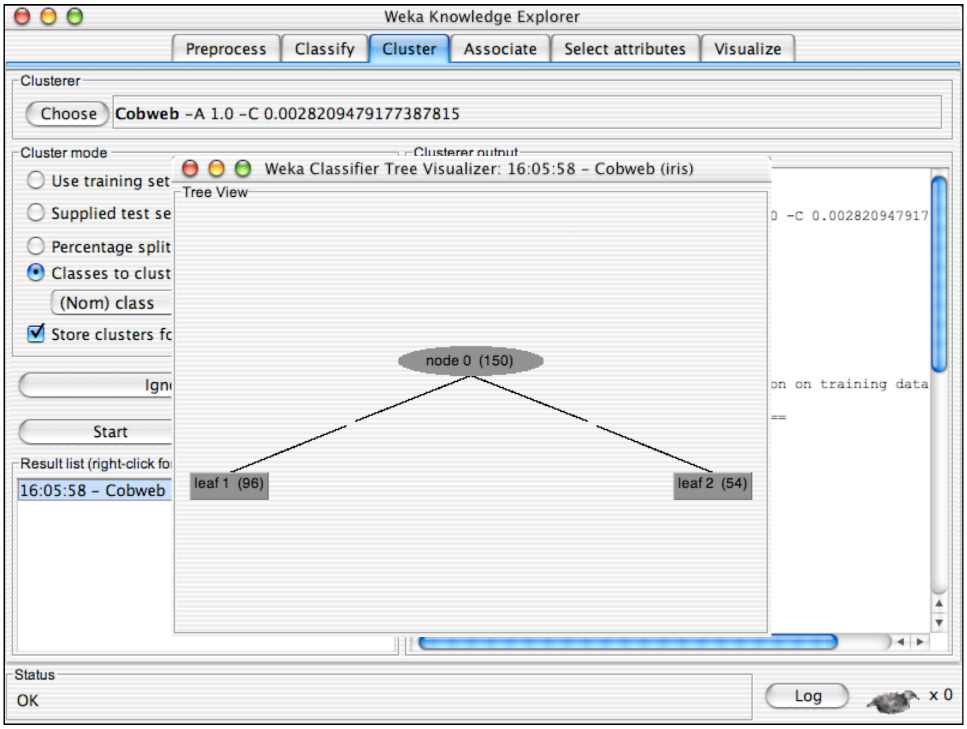
16:05:58 - Cobweb

Clusterer output

```
=== Run information ===

Scheme:       weka.clusterers.Cobweb -A 1.0 -C 0.002820947917
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



                  training set ===
```

Status

OK



Clusterer: **Cobweb** -A 1.0 -C 0.0028209479177387815

Cluster mode

○ Use training set

○ Supplied test se

○ Percentage split

● Classes to clust

(Nom) class

☑ Store clusters fc

Result list (right-click fo

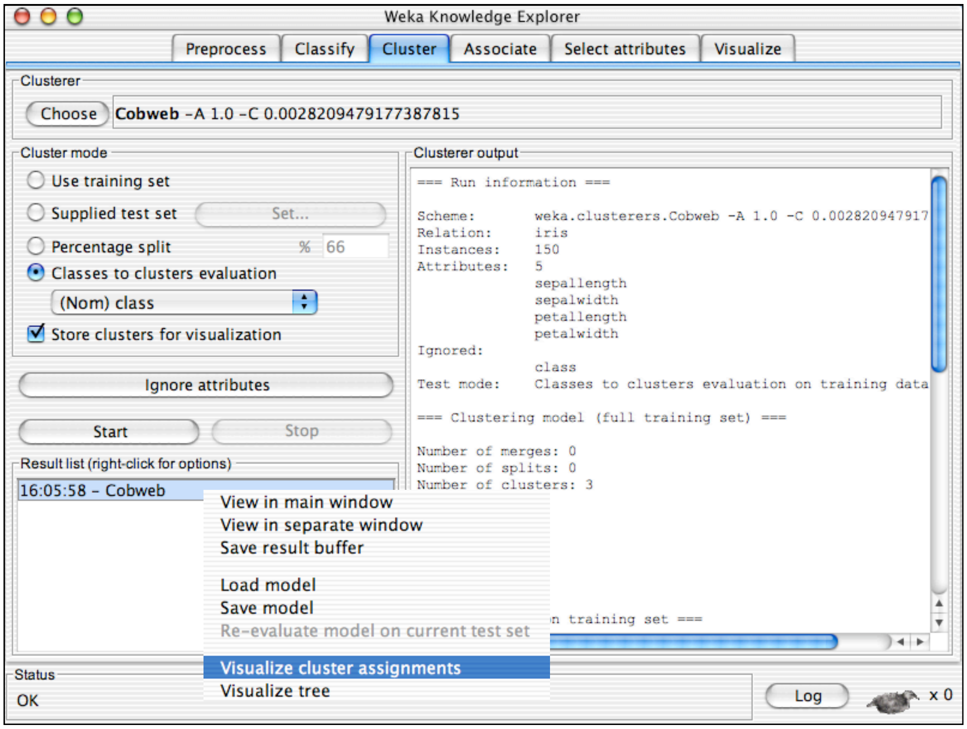
16:05:58 - Cobweb

Clusterer output

```
0 -C 0.002820947917

on on training data
==
```

Status

OK

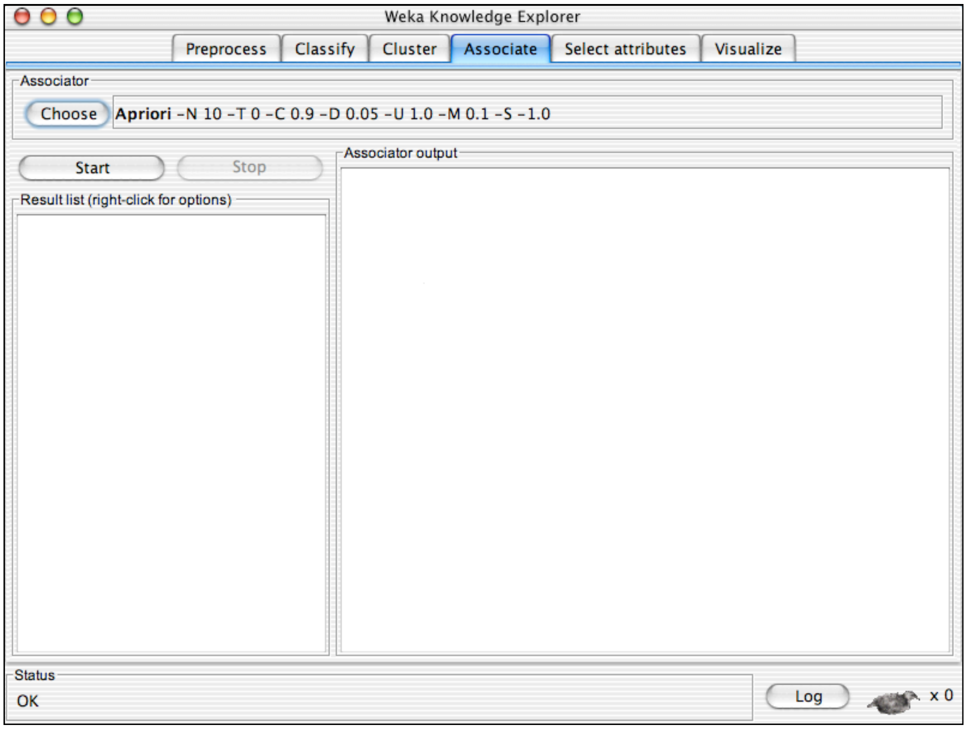


Cobweb -A 1.0 -C 0.0028209479177387815

```
=== Run information ===

Scheme:       weka.clusterers.Cobweb -A 1.0 -C 0.002820947917
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



n training set ===
```

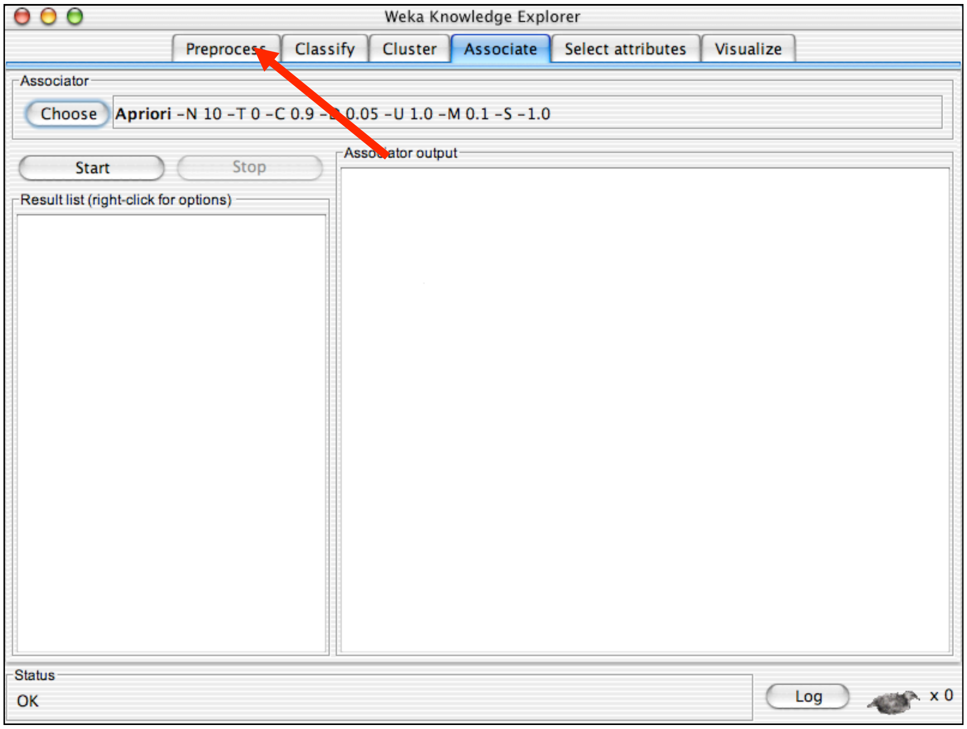
Explorer: finding associations

- WEKA contains an implementation of the Apri ori algorithm for learning association rules
  - Works only with discrete data
- Can identify statistical dependencies between groups of attributes:
  - milk, butter ⇒ bread, eggs (with confidence 0.9 and support 2000)
- Apriori can compute all rules that have a give n minimum support and exceed a given confi dence



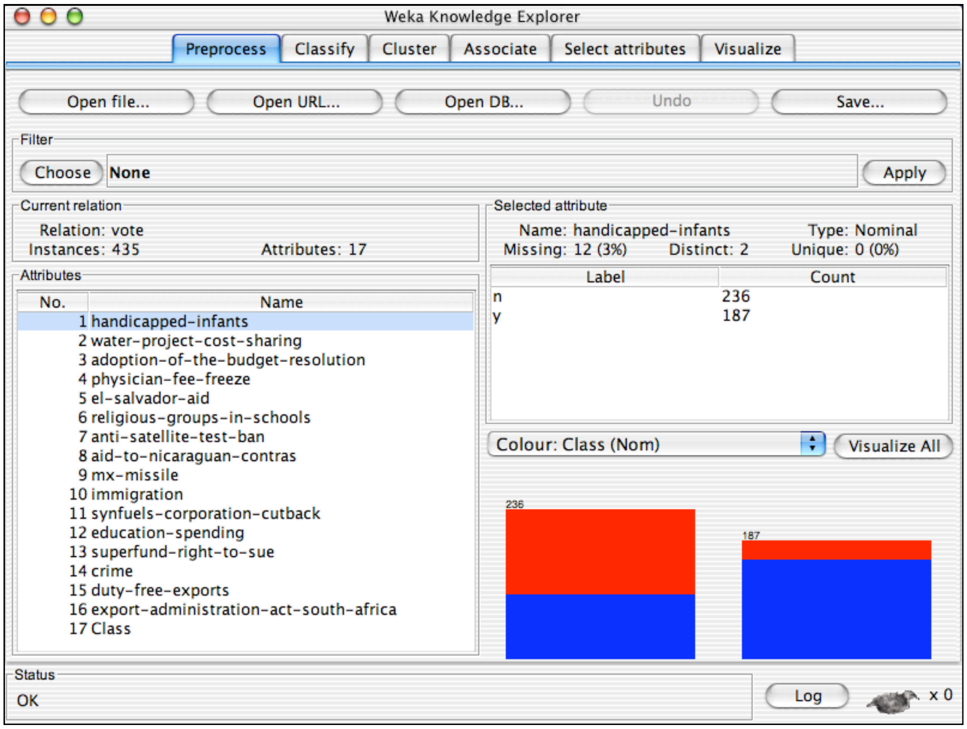
Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0

Status: OK



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

Status: OK



Filter: None

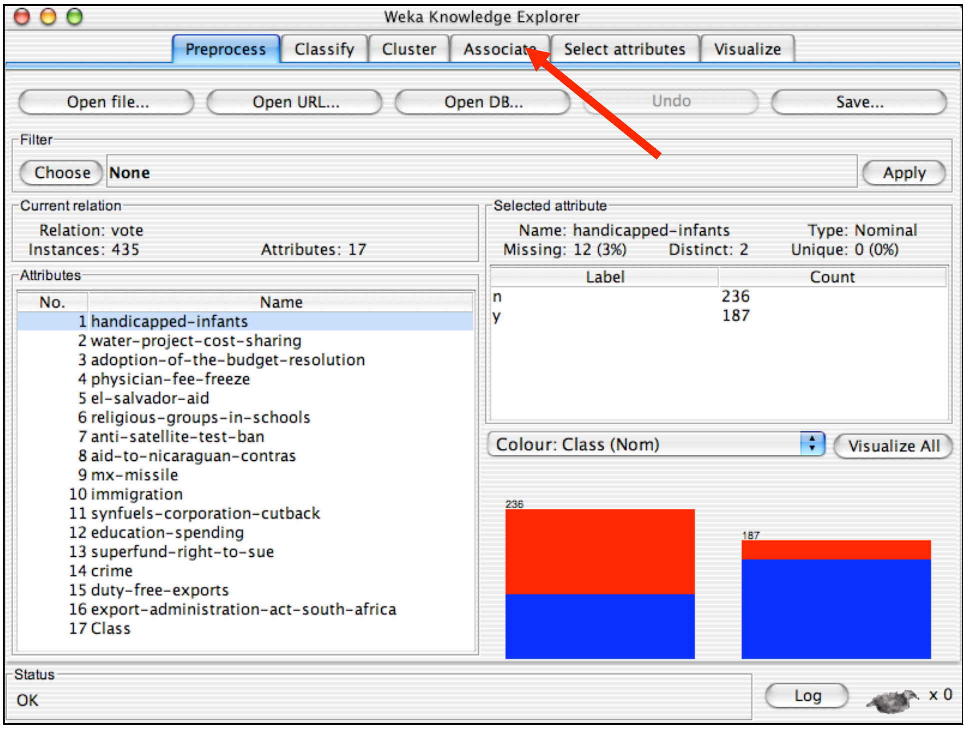
Relation: vote
Instances: 435 Attributes: 17

No.	Name
1	handicapped-infants
2	water-project-cost-sharing
3	adoption-of-the-budget-resolution
4	physician-fee-freeze
5	el-salvador-aid
6	religious-groups-in-schools
7	anti-satellite-test-ban
8	aid-to-nicaraguan-contras
9	mx-missile
10	immigration
11	synfuels-corporation-cutback
12	education-spending
13	superfund-right-to-sue
14	crime
15	duty-free-exports
16	export-administration-act-south-africa
17	Class

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

Label	Count
n	236
y	187

Status: OK


Weka Knowledge Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Filter: None

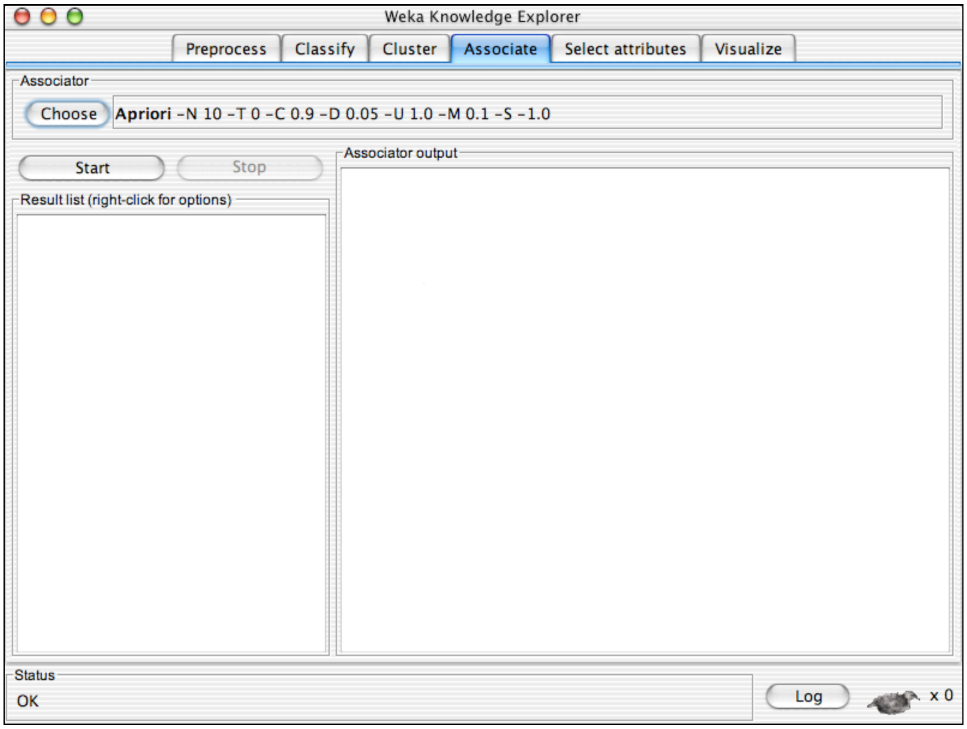
Relation: vote
Instances: 435 Attributes: 17

No.	Name
1	handicapped-infants
2	water-project-cost-sharing
3	adoption-of-the-budget-resolution
4	physician-fee-freeze
5	el-salvador-aid
6	religious-groups-in-schools
7	anti-satellite-test-ban
8	aid-to-nicaraguan-contras
9	mx-missile
10	immigration
11	synfuels-corporation-cutback
12	education-spending
13	superfund-right-to-sue
14	crime
15	duty-free-exports
16	export-administration-act-south-africa
17	Class

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

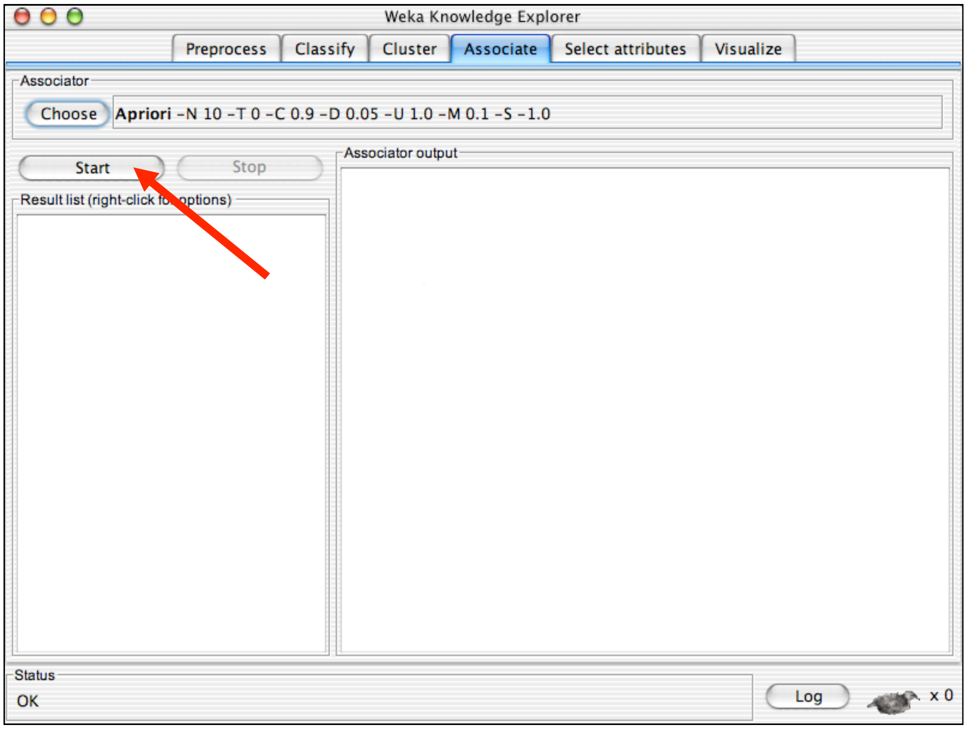
Label	Count
n	236
y	187

Status: OK



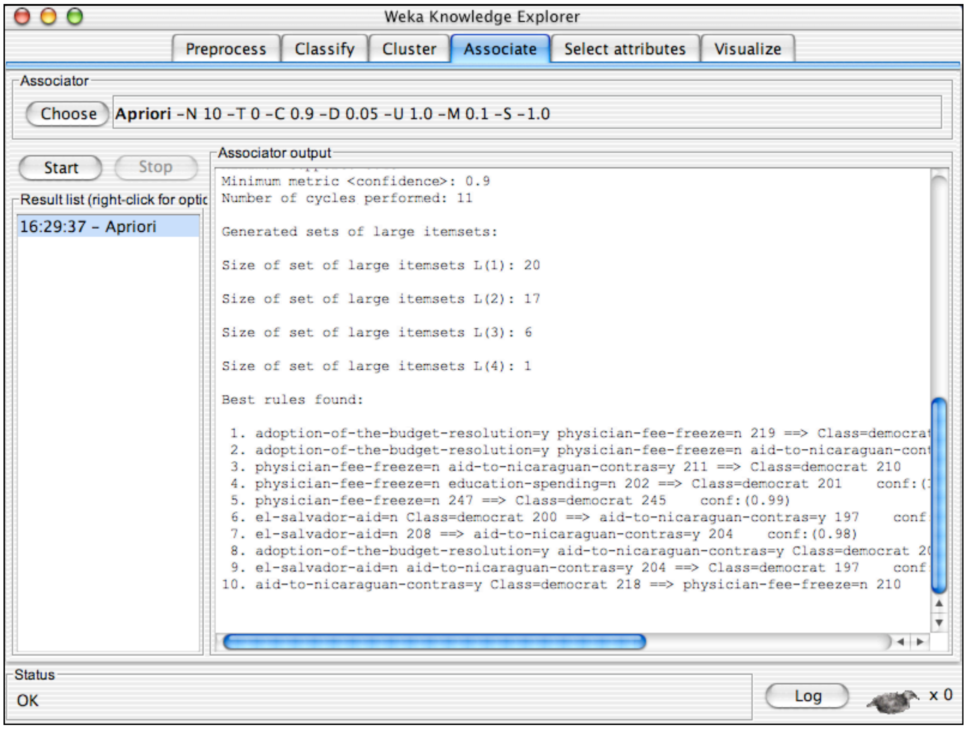
Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0

OK



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

OK

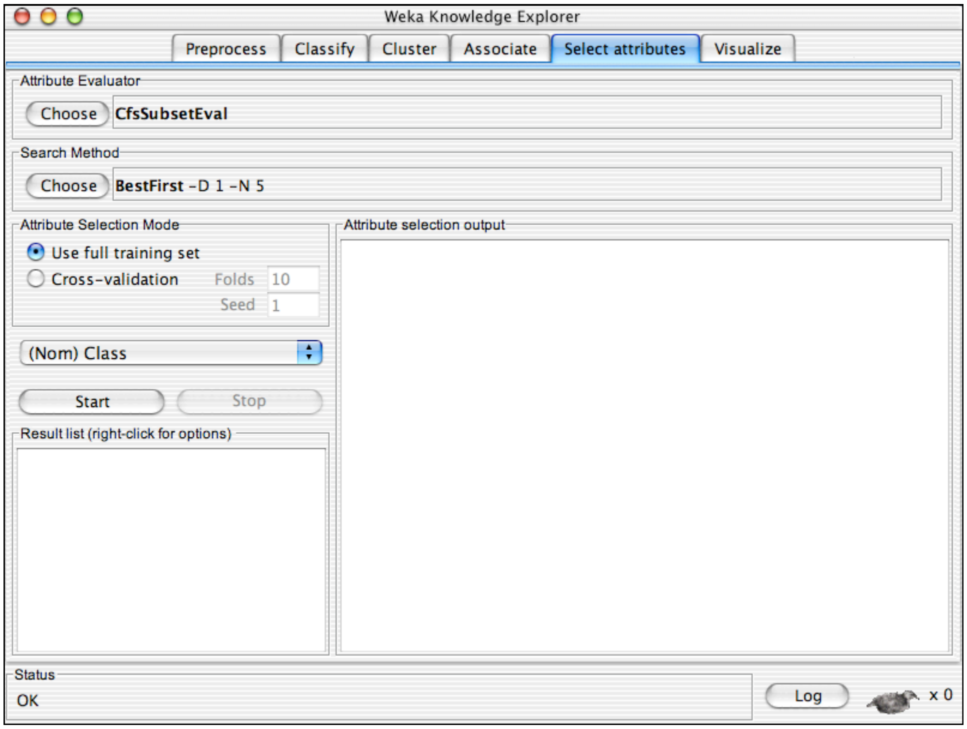


Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0

```
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-cont
 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:(
 5. physician-fee-freeze=n 247 ==> Class=democrat 245    conf:(0.99)
 6. el-salvador-aid=n Class=democrat 200 ==> aid-to-nicaraguan-contras=y 197    conf
 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 2
 9. el-salvador-aid=n aid-to-nicaraguan-contras=y 204 ==> Class=democrat 197    conf
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

○ Cross-validation Folds 10 Seed 1

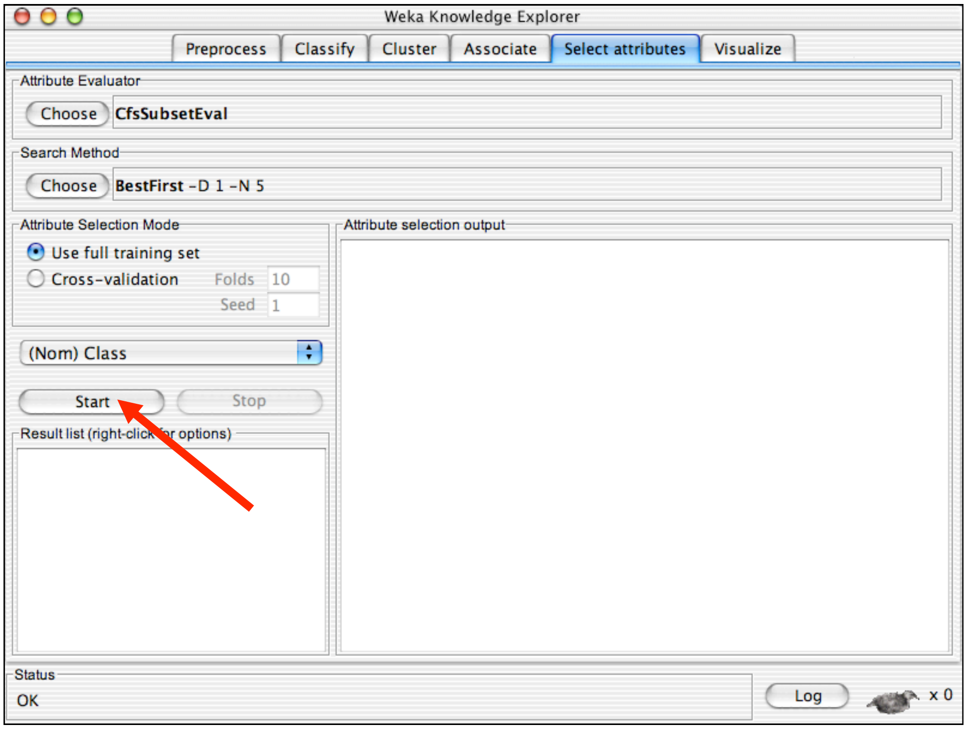
(Nom) Class

Result list (right-click for options)

Attribute selection output

Status

OK



Attribute Evaluator: CfsSubsetEval

Search Method: BestFirst -D 1 -N 5

Attribute Selection Mode

● Use full training set

○ Cross-validation Folds 10 Seed 1

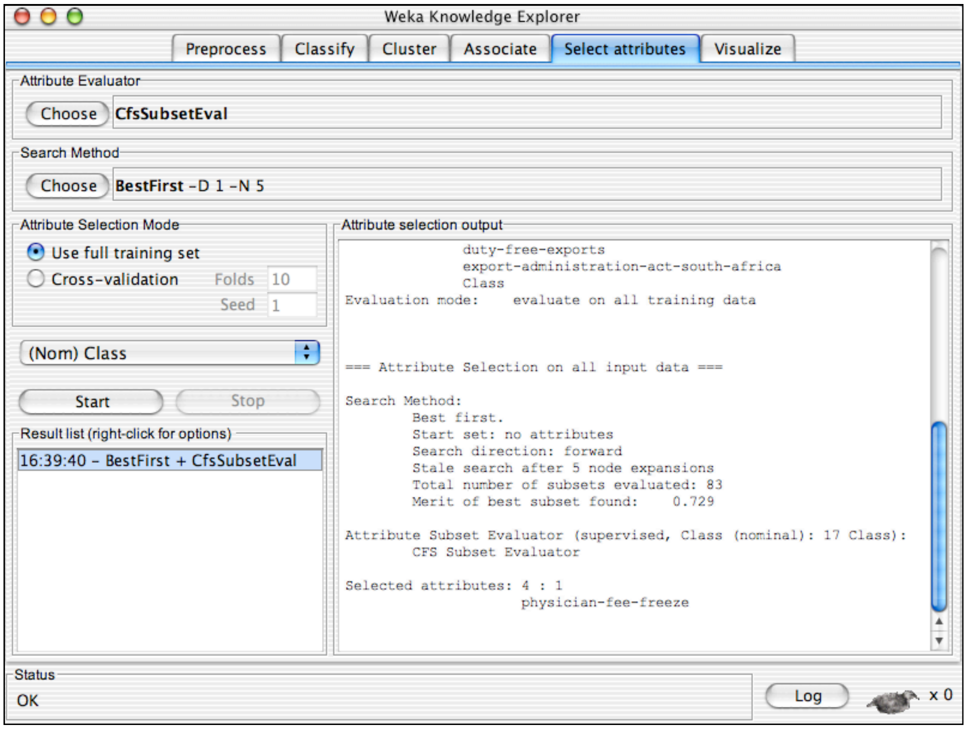
(Nom) Class

Result list (right-click for options)

Attribute selection output

Status

OK



Attribute Evaluator: CfsSubsetEval

Search Method: BestFirst -D 1 -N 5

Attribute Selection Mode

● Use full training set

○ Cross-validation Folds 10 Seed 1

(Nom) Class

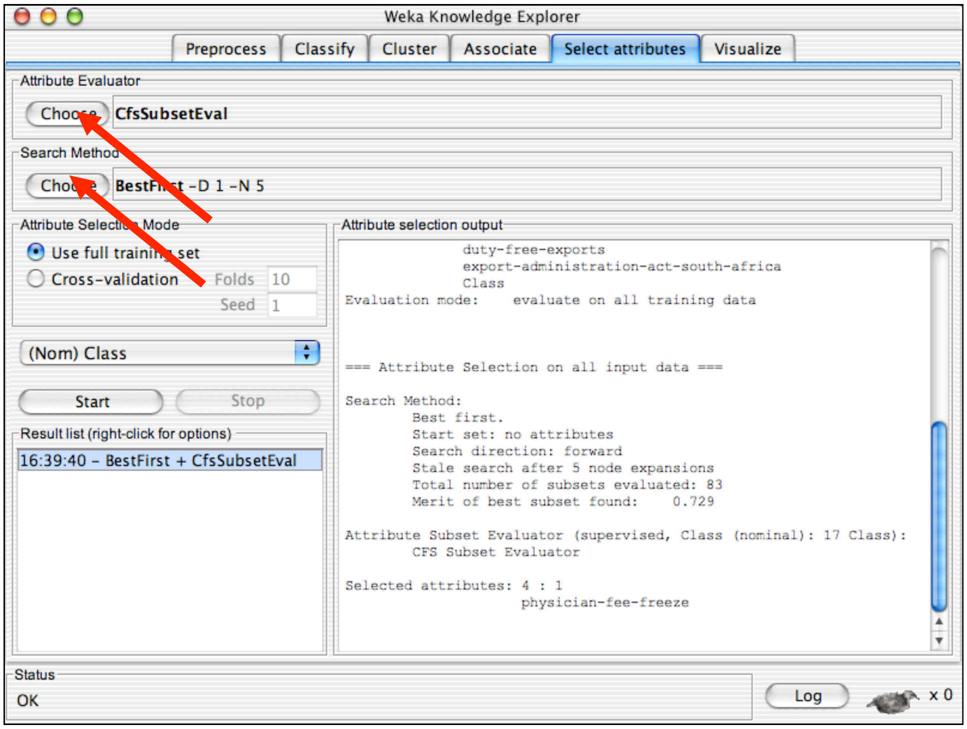
Result list (right-click for options)

16:39:40 - BestFirst + CfsSubsetEval

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: CfsSubsetEval

Search Method: BestFirst -D 1 -N 5

Attribute Selection Mode

● Use full training set

○ Cross-validation Folds 10 Seed 1

(Nom) Class

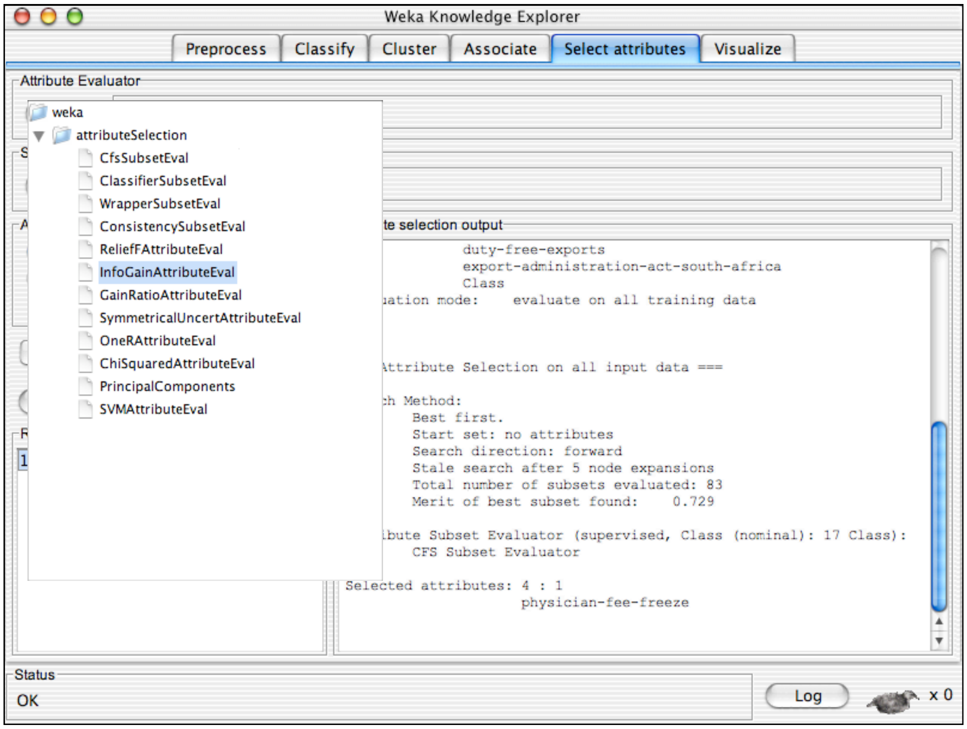
Result list (right-click for options)

16:39:40 - BestFirst + CfsSubsetEval

Attribute selection output

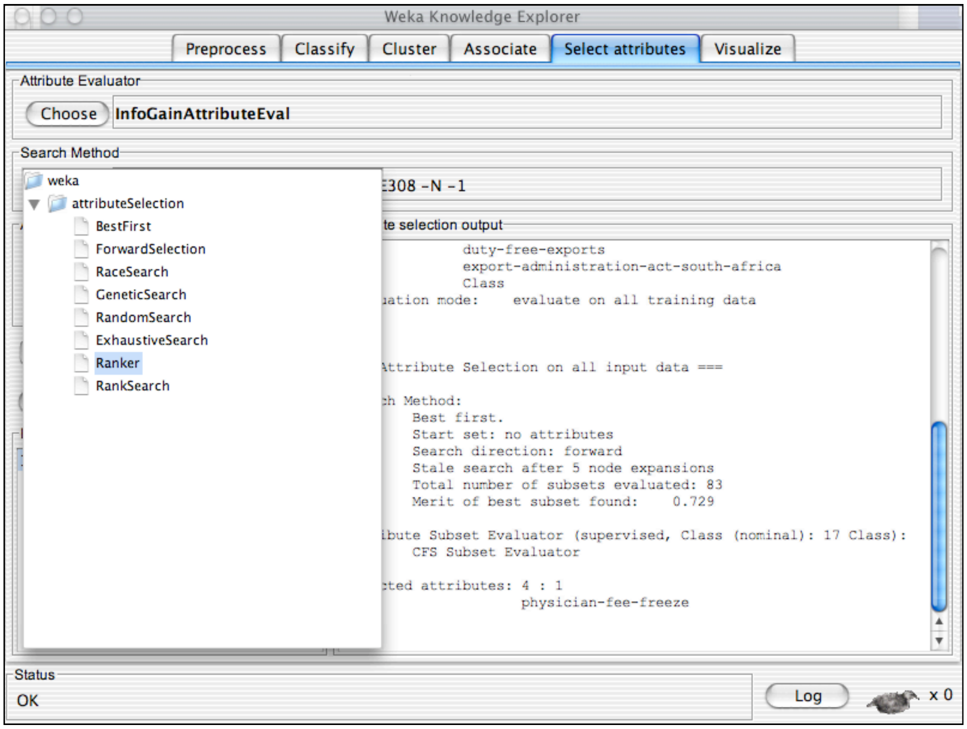
```
                 duty-free-exports
                 export-administration-act-south-africa
                 Class
Evaluation mode:    evaluate on all training data



=== 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



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


Attribute Selection on all input data ===

ch 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

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

Selected attributes: 4 : 1
                     physician-fee-freeze
```

OK

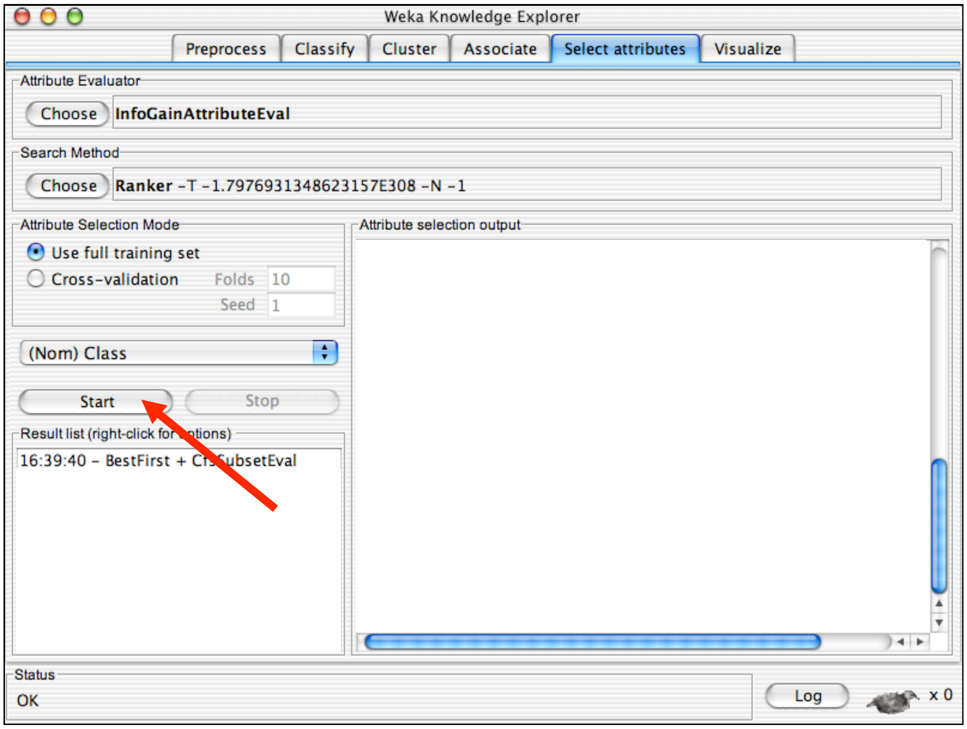


Attribute Evaluator: InfoGainAttributeEval

Search Method: E308 -N -1

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


Attribute Selection on all input data ===

ch 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

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

cted attributes: 4 : 1
                 physician-fee-freeze
```

OK



Attribute Evaluator: InfoGainAttributeEval

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

Attribute Selection Mode

● Use full training set

○ Cross-validation Folds 10 Seed 1

(Nom) Class

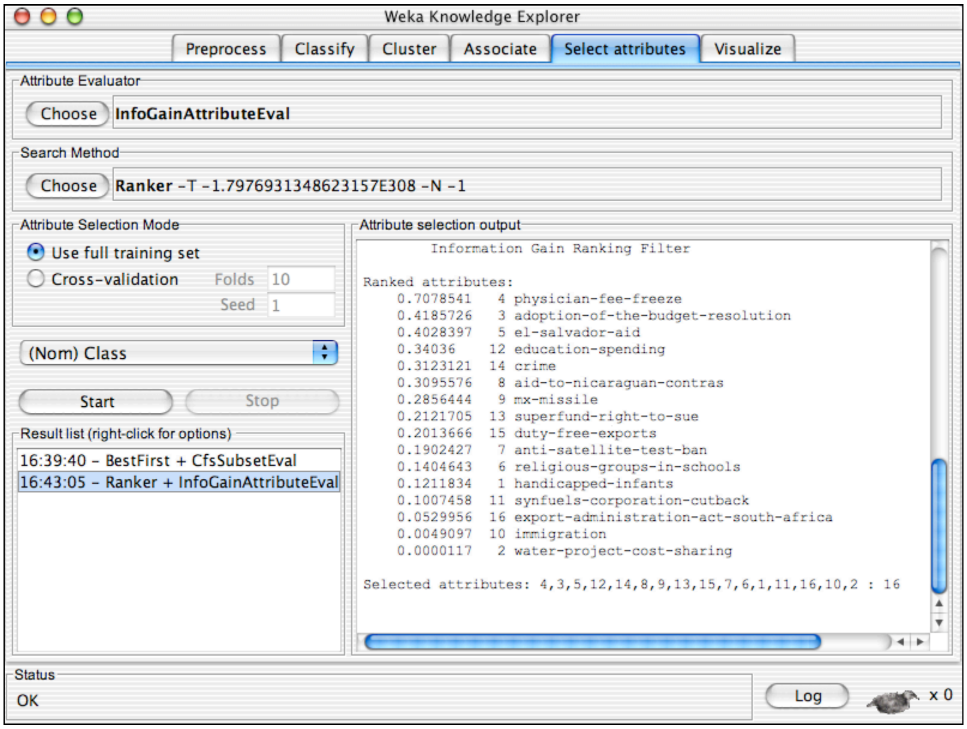
Result list (right-click for options)

16:39:40 - BestFirst + CfsSubsetEval

Attribute selection output

Status

OK



Attribute Evaluator: InfoGainAttributeEval

Search Method: 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

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

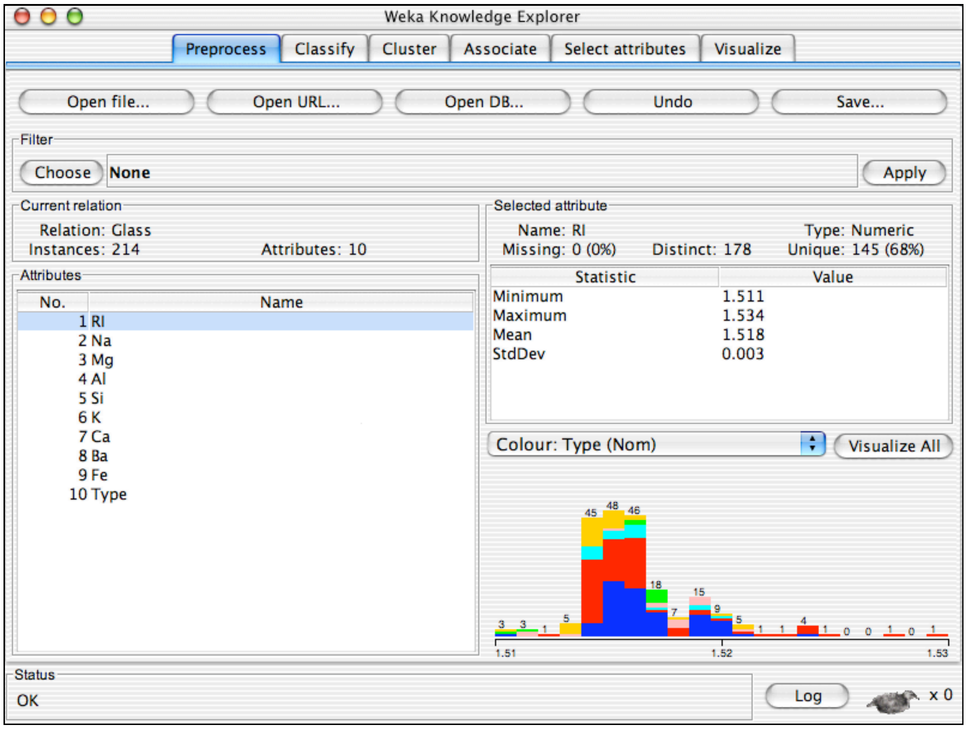
OK

Which attribute selector?

- Best: WRAPPER
  - Slow: O(2^N) search through all attribute combinations
  - The “wrapped” learner called to assess each combination
  - Some heuristics to prune the search; but does not scale
- If not WRAPPER
  - Use InfoGain / OneR for very big datasets
  - Use CFS otherwise
- Don’t use PCA
  - This is an unsupervised selector
  - So it is uninformed on how dimensions help classification

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



Filter
None

Current relation
Relation: Glass
Instances: 214
Attributes: 10

Attributes

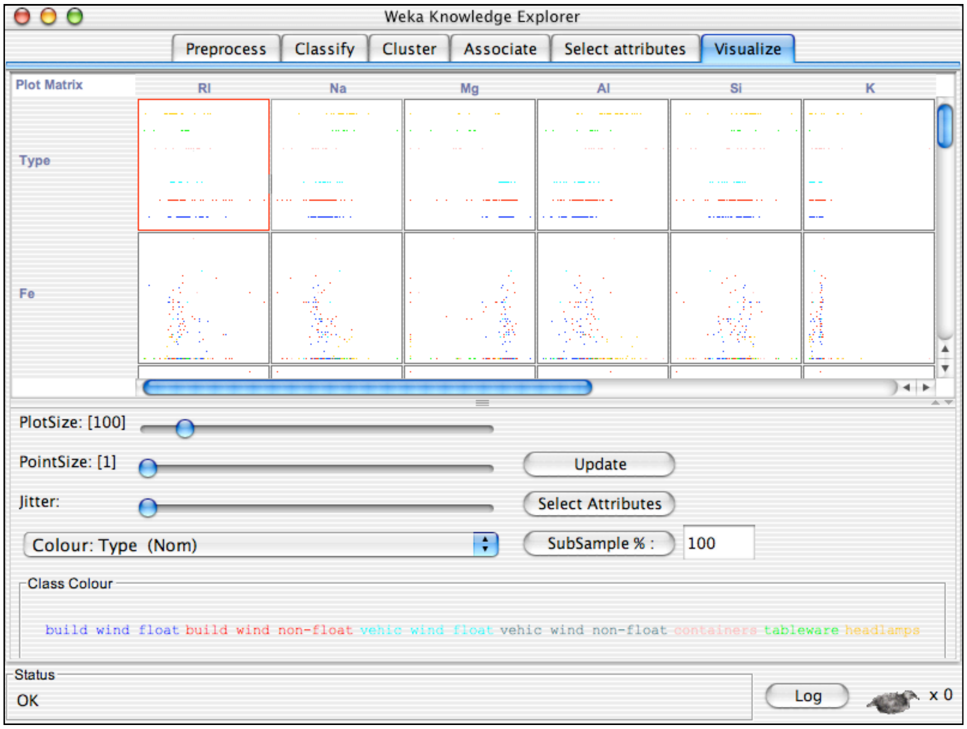
No.	Name
1	RI
2	Na
3	Mg
4	Al
5	Si
6	K
7	Ca
8	Ba
9	Fe
10	Type

Selected attribute
Name: RI
Type: Numeric
Missing: 0 (0%)
Distinct: 178
Unique: 145 (68%)

Statistic	Value
Minimum	1.511
Maximum	1.534
Mean	1.518
StdDev	0.003

Colour: Type (Nom)

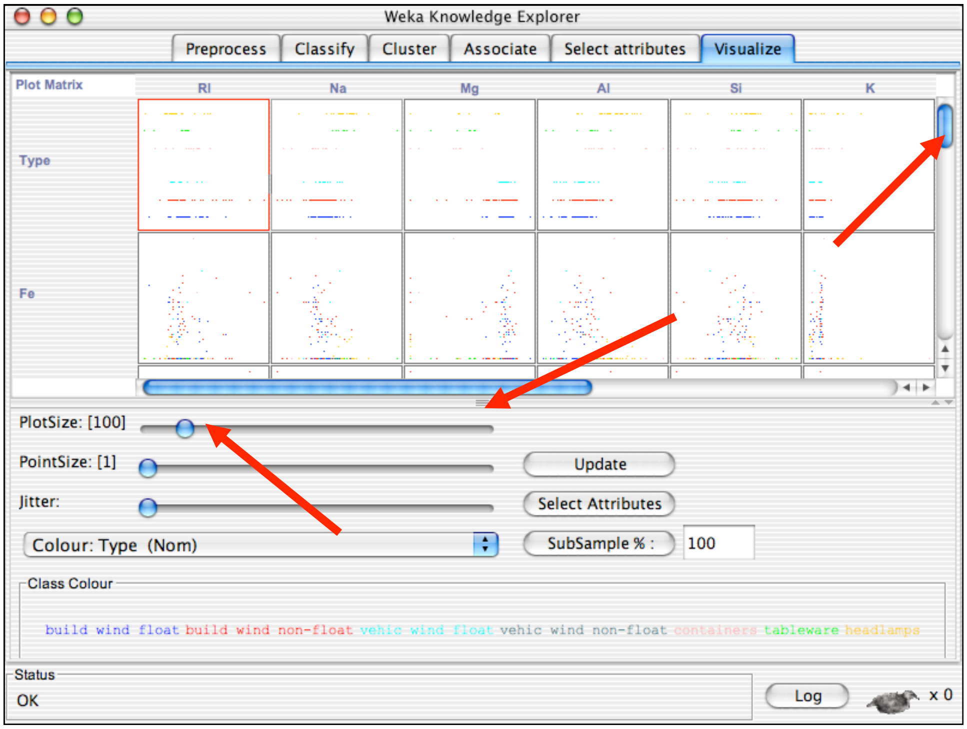
Status
OK



PlotSize: [100]
PointSize: [1]
Jitter:
Colour: Type (Nom)
SubSample % : 100

Class Colour
build wind float build wind non-float vehic wind float vehic wind non-float containers tableware headlamps

Status
OK



PlotSize: [100]

PointSize: [1]

Jitter:

Colour: Type (Nom)

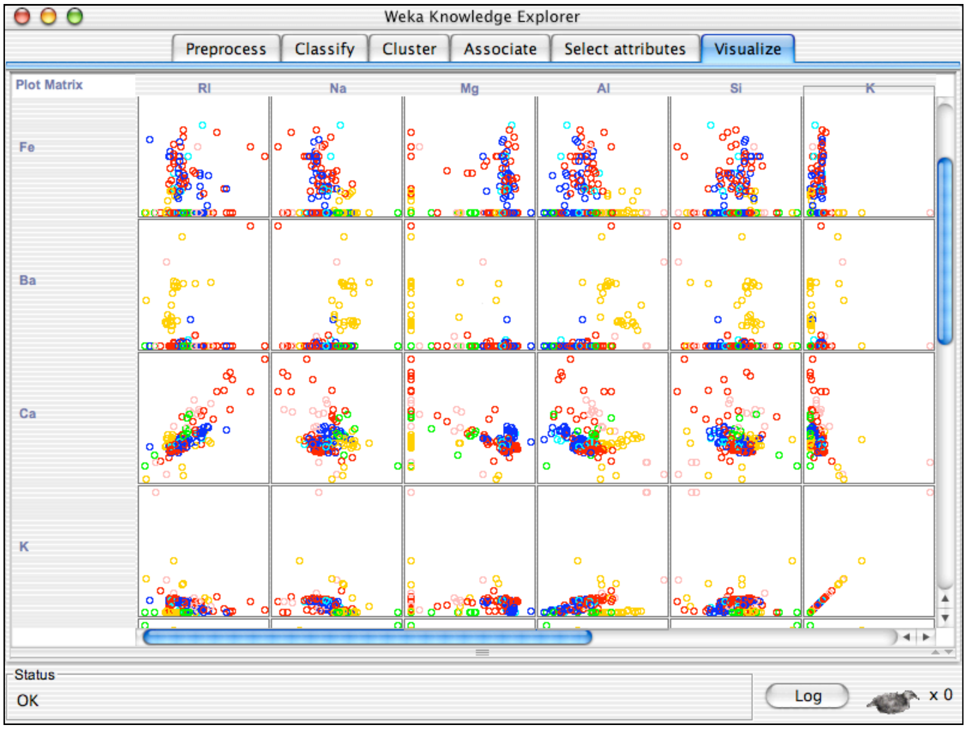
SubSample % : 100

Class Colour

build wind float build wind non-float vehic wind float vehic wind non-float containers tableware headlamps

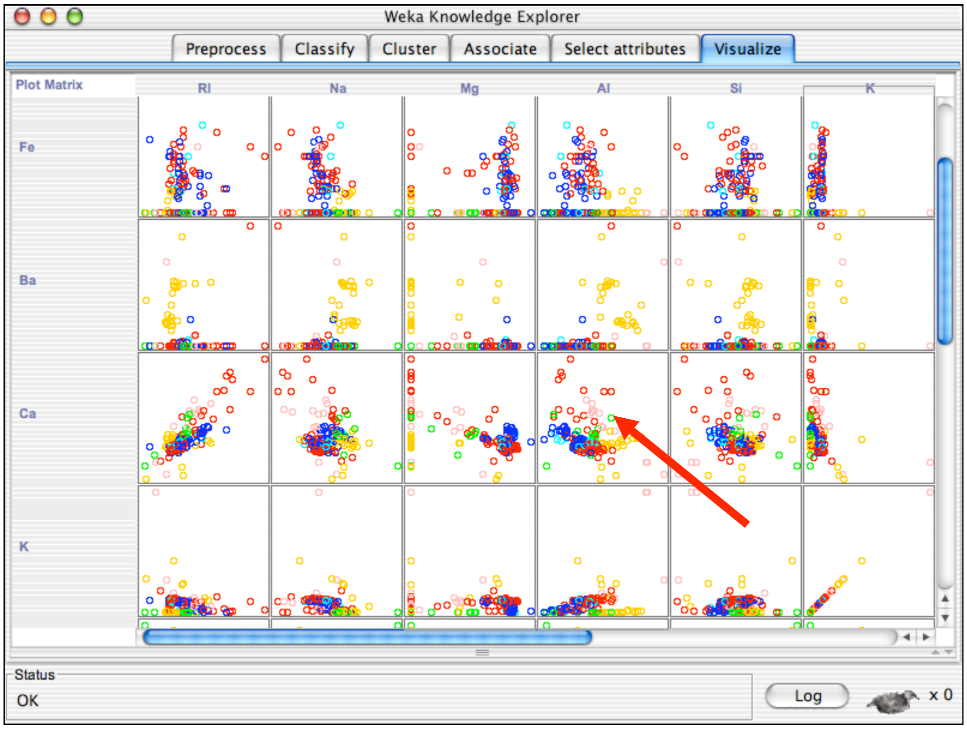
Status

OK



Status

OK

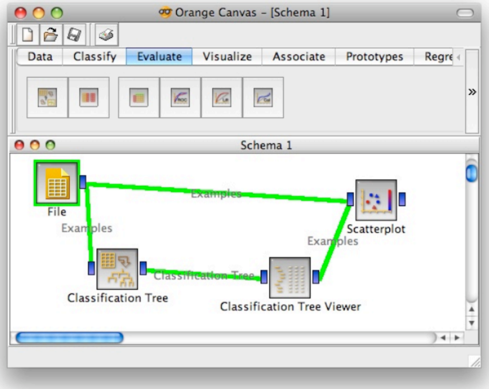

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Plot Matrix
RI
Na
Mg
Al
Si
K
Fe
Ba
Ca
K
Status
OK
Log
x 0

Evaluation

Limitations

- Loads all data into ram prior to learning
  - Problem for large data sets
- Not good for complex experiments
- IMHO, discourages experimentation with new learners
  - The "WEKA effect"
    - Try every learner till something works
- Still, very useful for
  - Initial investigations
  - Learning data mining
  - Or as a sub-routine of other systems

Alternate tools: Orange

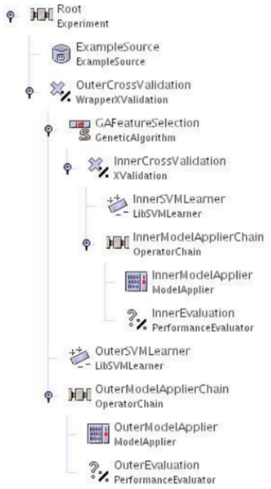


Written in Python

Simpler specification (but see WEKA's KnowledgeFlow Environment).

Also, less community support/ debugging. So sometimes frustrated by random bugs

Alternate tools: RapidMiner



```
Root
Experiment
    ExampleSource
    ExampleSource
    OuterCrossValidation
    WrapperXValidation
        GAFeatureSelection
        GeneticAlgorithm
            InnerCrossValidation
            XValidation
                InnerSVMLearner
                LibSVMLearner
                InnerModelApplierChain
                OperatorChain
                    InnerModelApplier
                    ModelApplier
                    InnerEvaluation
                    PerformanceEvaluator
        OuterSVMLearner
        LibSVMLearner
        OuterModelApplierChain
        OperatorChain
            OuterModelApplier
            ModelApplier
            OuterEvaluation
            PerformanceEvaluator
```

Experiments specified in an XML tree syntax

In theory, possible to share experimental descriptions

Alternate tools: OurMine

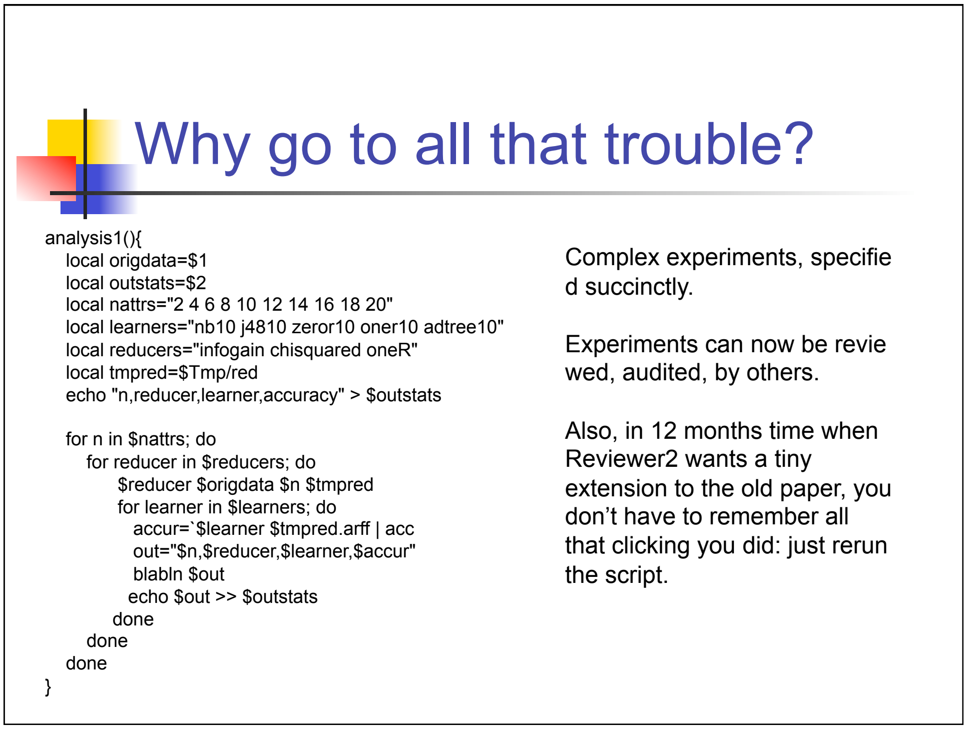
```
Java=$Base/lib/java
Weka="java -Xmx2048M -cp $Java/weka.jar ”
Clusterers="java -Xmx1024M -jar $Java/Clusterers.jar ”
Reducers="java -Xmx1024M -jar $Java/Reduce.jar ”

nb() {
   local learner=weka.classifiers.bayes.NaiveBayes
   $Weka $learner -p 0 -t $1 -T $2
}

nb10() {
   local learner=weka.classifiers.bayes.NaiveBayes
   $Weka $learner -i -t $1
}

j48() {
   local learner=weka.classifiers.trees.J48
   $Weka $learner -p 0 -C 0.25 -M 2 -t $1 -T $2
}
```

Forget the visuals.

Make WEKA a sub-routine inside Bash script

Now you can mix WEKA’s JAVA with learners written in your favorite language.

But how do you find the magic command strings?



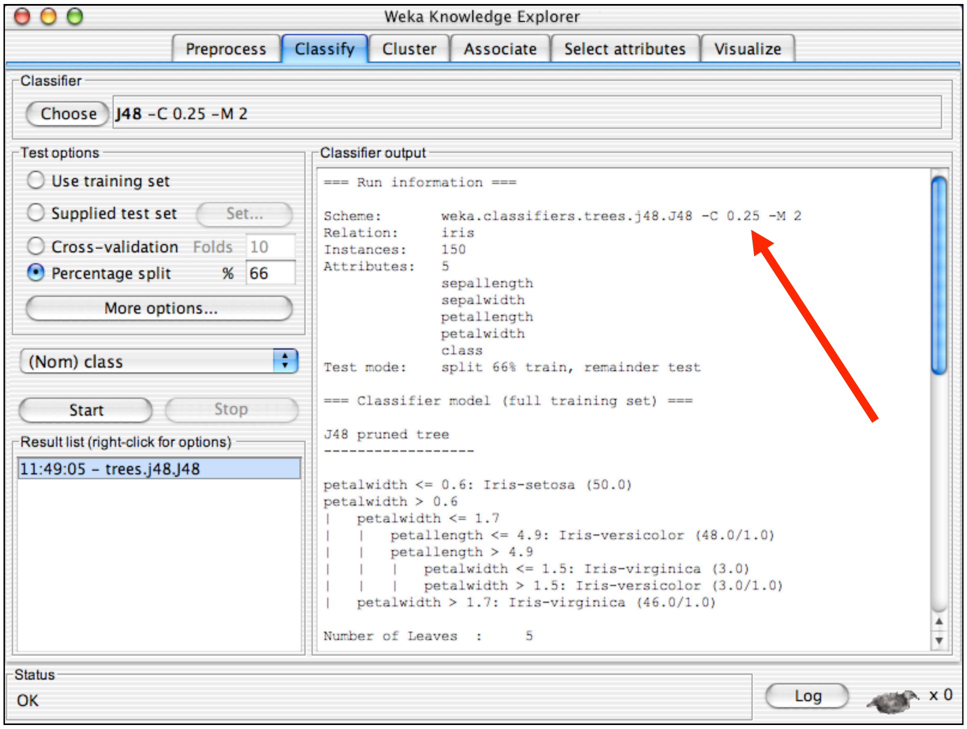
Why go to all that trouble?

```
analysis1(){
   local origdata=$1
   local outstats=$2
   local nattrs="2 4 6 8 10 12 14 16 18 20"
   local learners="nb10 j4810 zeror10 oner10 adtree10"
   local reducers="infogain chisquared oneR"
   local tmpred=$Tmp/red
   echo "n,reducer,learner,accuracy" > $outstats

   for n in $nattrs; do
      for reducer in $reducers; do
          $reducer $origdata $n $tmpred
          for learner in $learners; do
            accur=`$learner $tmpred.arff | acc
            out="$n,$reducer,$learner,$accur"
            blabln $out
           echo $out >> $outstats
         done
      done
   done
}
```

Complex experiments, specifie d succinctly.

Experiments can now be revie wed, audited, by others.

Also, in 12 months time when Reviewer2 wants a tiny extension to the old paper, you don't have to remember all that clicking you did: just rerun the script.



Classifier

J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set

○ Cross-validation Folds 10

● Percentage split % 66

(Nom) class

Result list (right-click for options)

11:49:05 - trees.j48.J48

Classifier output

```
=== Run information ===

Scheme:       weka.classifiers.trees.j48.J48 -C 0.25 -M 2
Relation:     iris
Instances:    150
Attributes:   5
              sepallength
              sepalwidth
              petallength
              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
|   |   petallength <= 4.9: Iris-versicolor (48.0/1.0)
|   |   petallength > 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
```

Status

OK