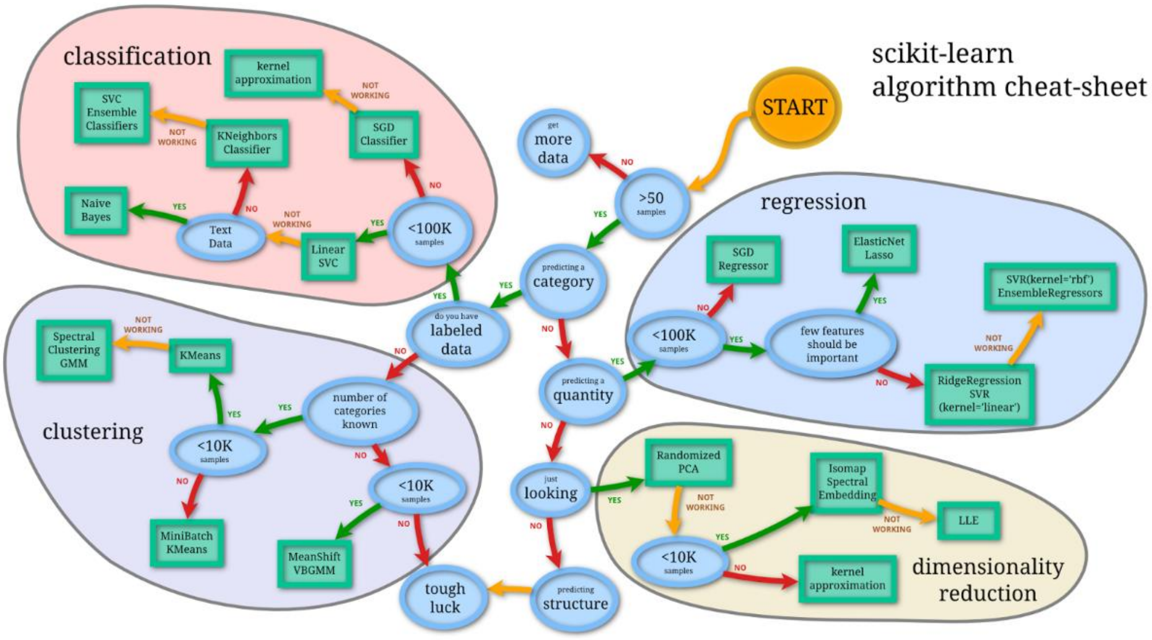
Overview: Data Mining Methods



WEKA Tutorial

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

WEKA - Introduction

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

Pre-processing the data

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

WEKA with “flat” files

```
@relation heart-disease-simplified

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

@data
63,male,typ_angina,233,no,not_present
67,male,asympt,286,yes,present
67,male,asympt,229,yes,present
38,female,non_anginal,?,no,not_present
...
```

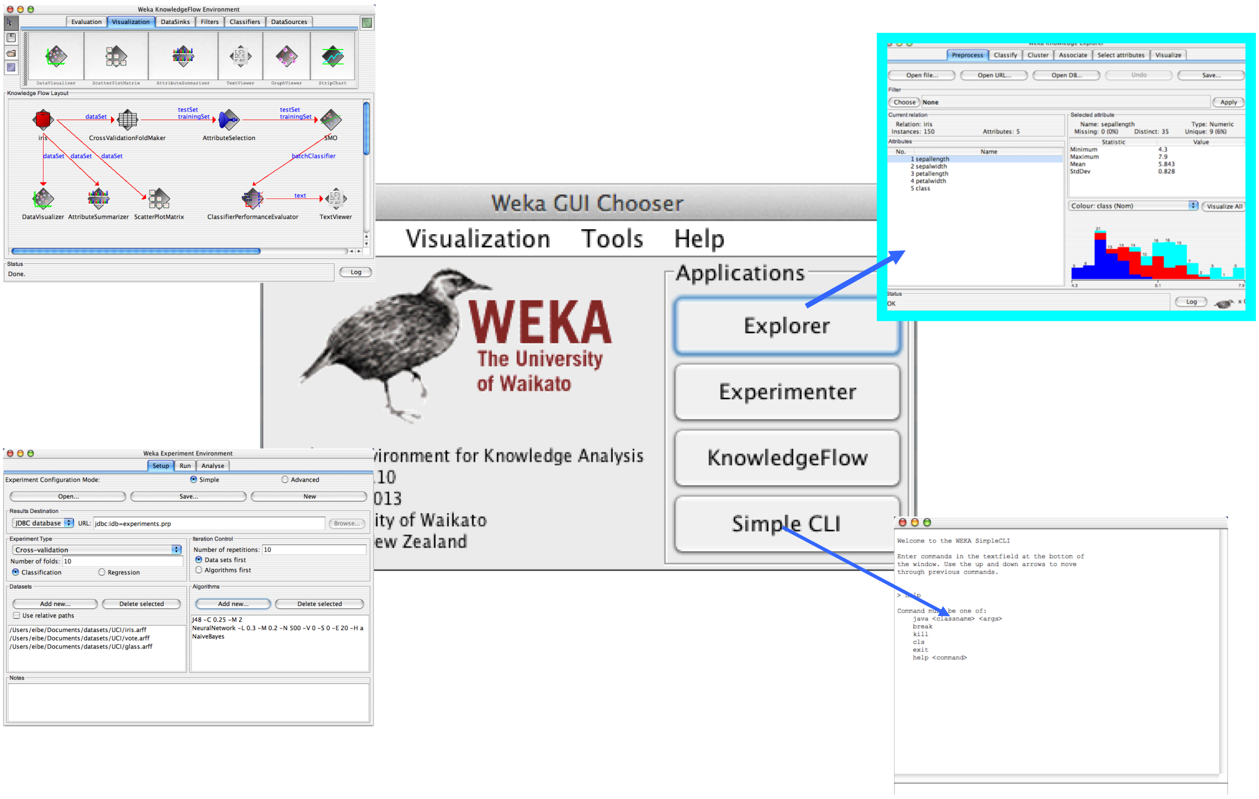
Flat file in ARFF format

WEKA with “flat” files

```
@relation heart-disease-simplified

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

@data
63,male,typ_angina,233,no,not_present
67,male,asympt,286,yes,present
67,male,asympt,229,yes,present
38,female,non_anginal,?,no,not_present
...
```

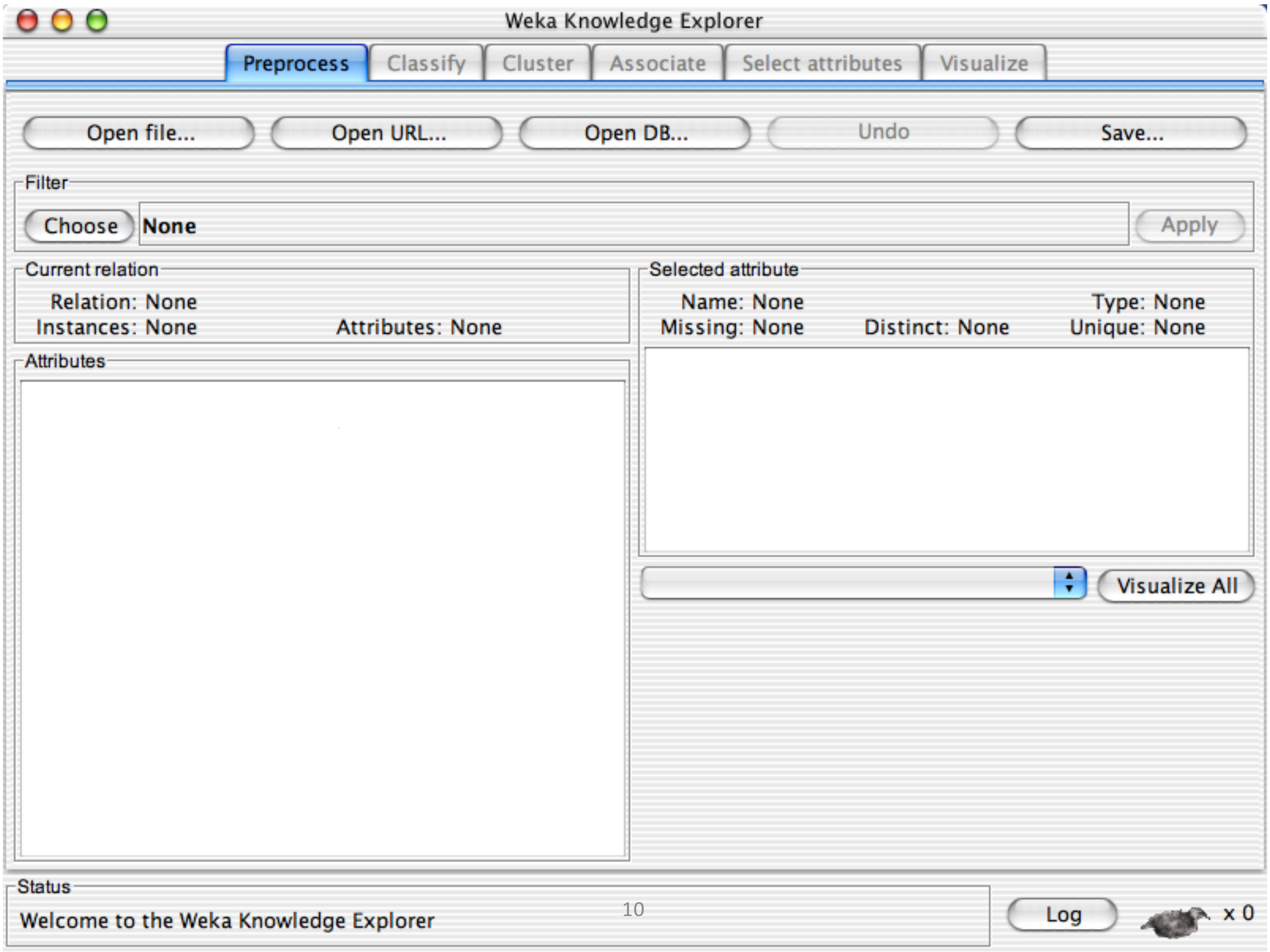
numeric attribute

nominal attribute



...vironment for Knowledge Analysis
...10
...013
...ity of Waikato
...ew Zealand

```
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>
```



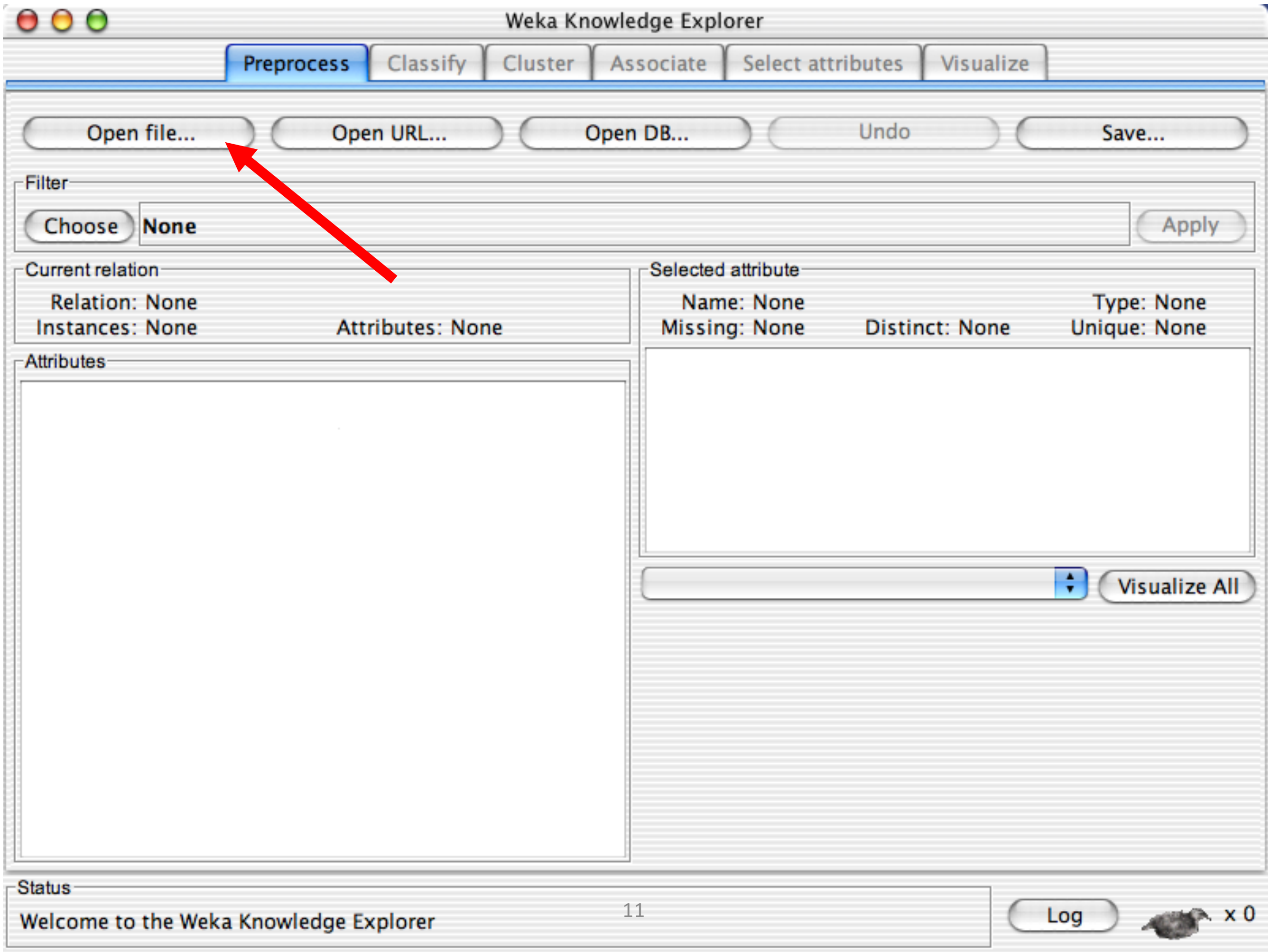
Filter: None

Relation: None
Instances: None Attributes: None

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

Welcome to the Weka Knowledge Explorer

x 0

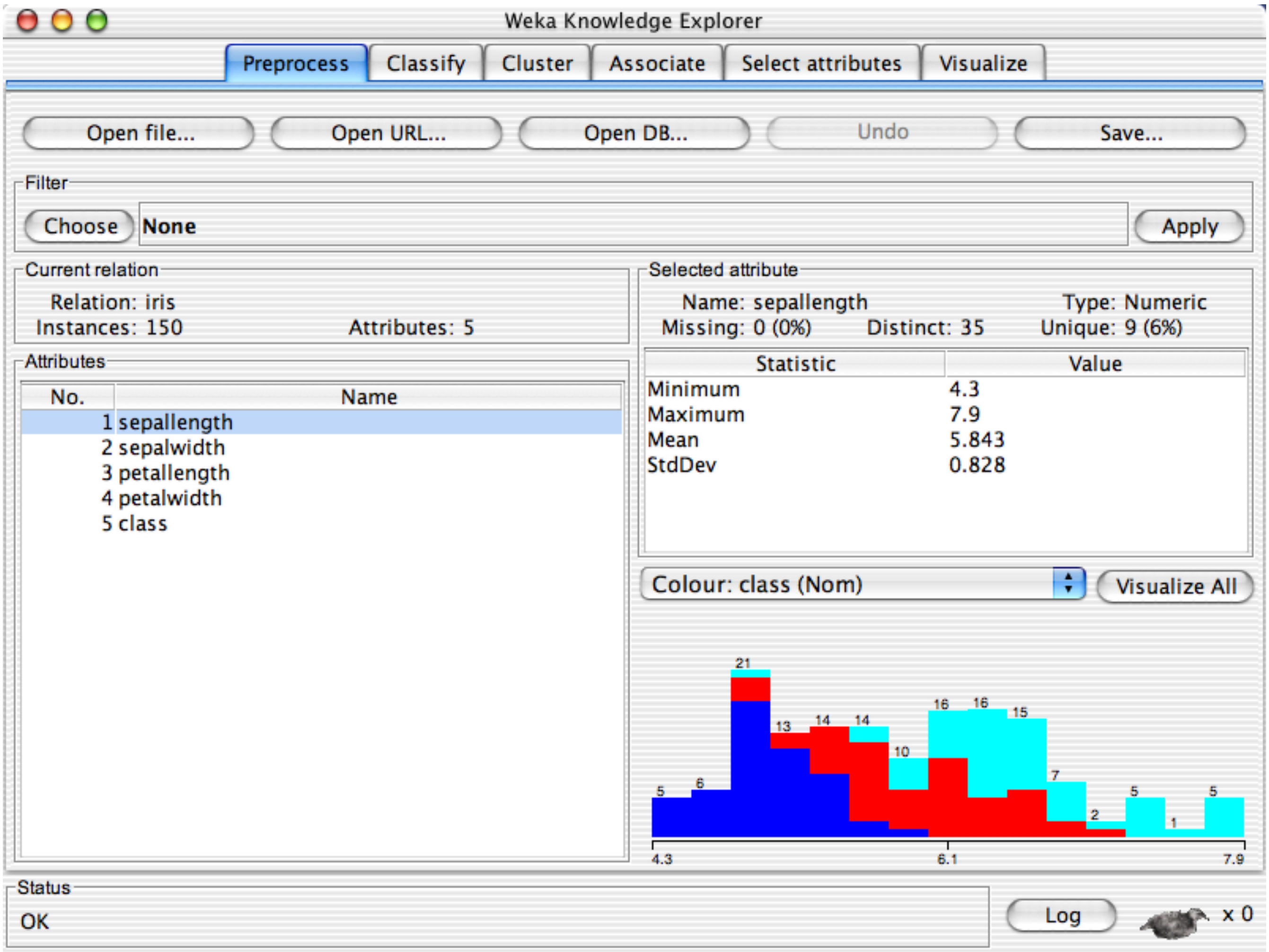


Filter: None

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

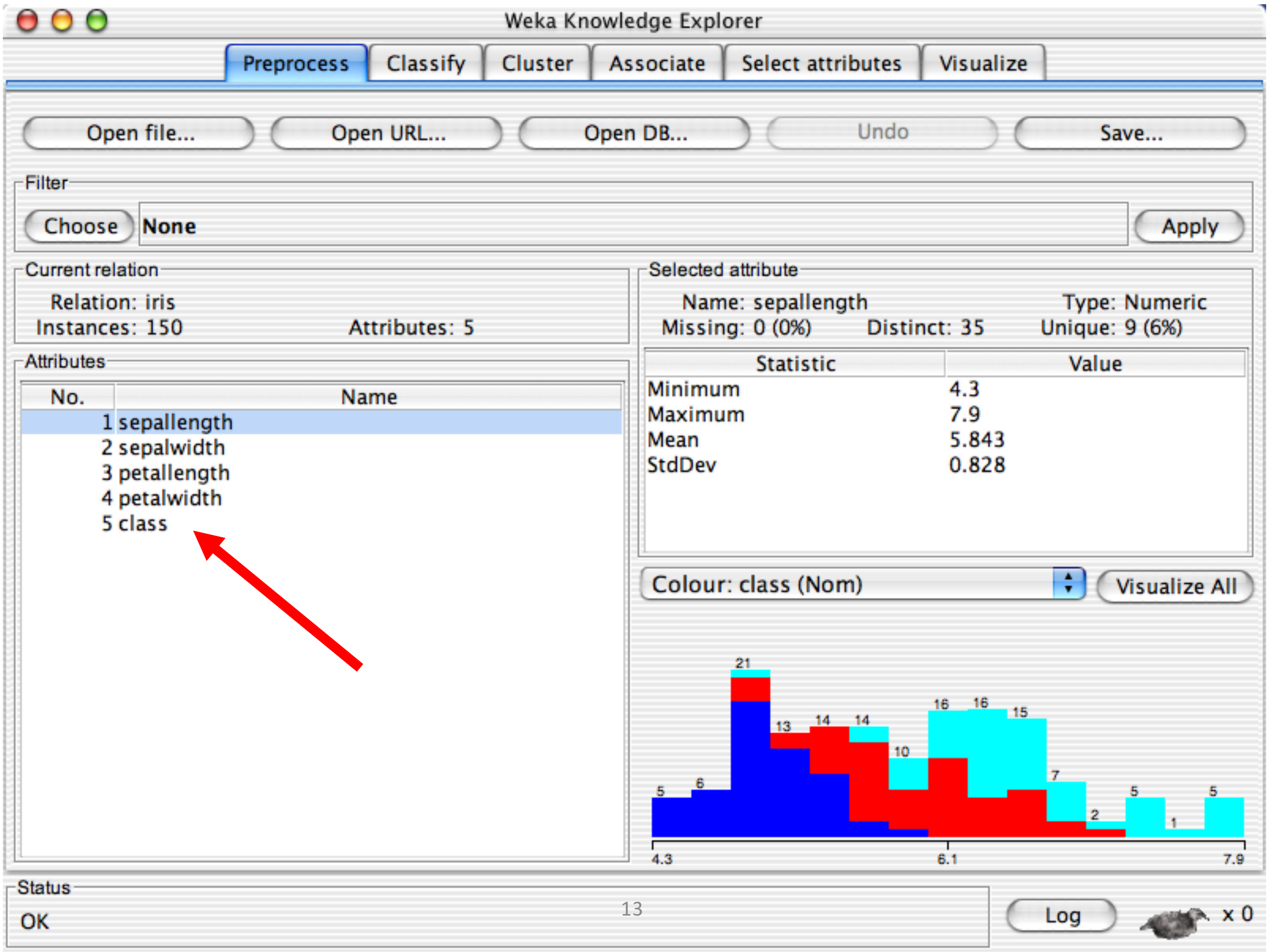
No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

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

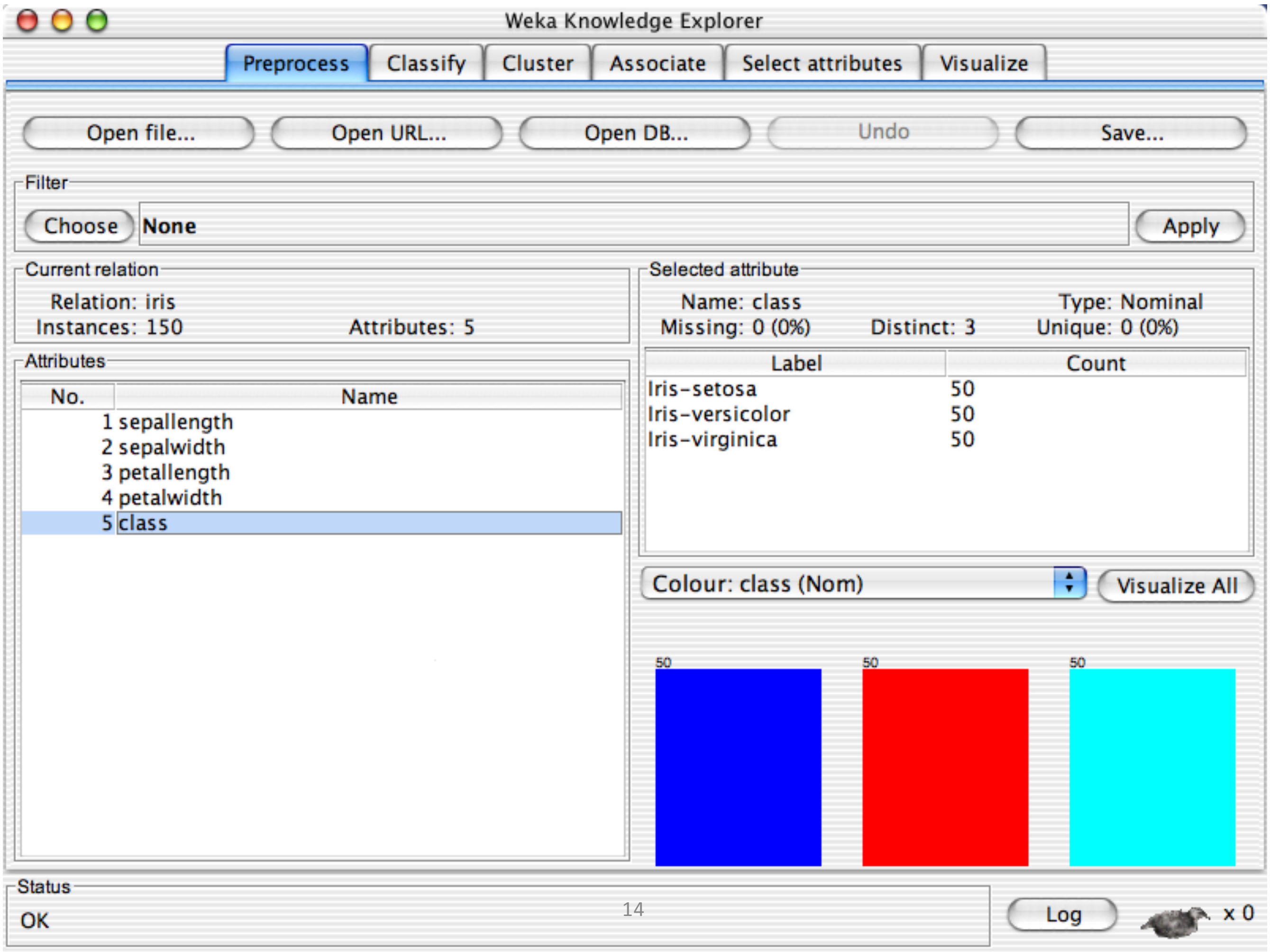
Name: sepallength Type: Numeric
Missing: 0 (0%) Distinct: 35 Unique: 9 (6%)

Statistic	Value
Minimum	4.3
Maximum	7.9
Mean	5.843
StdDev	0.828

Colour: class (Nom)

Status
OK

x 0



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

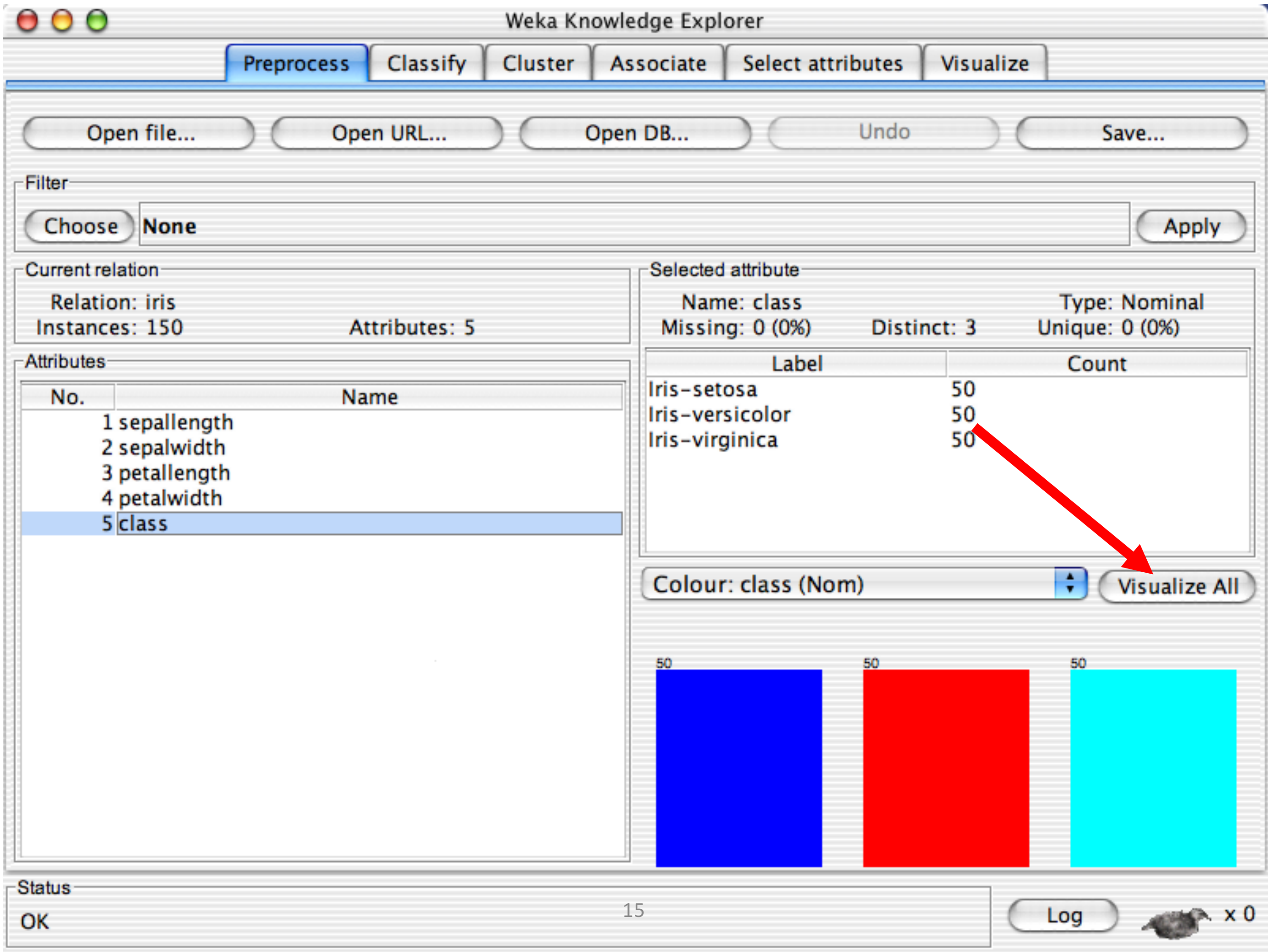
Name: class Type: Nominal
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

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

Colour: class (Nom)

Status
OK

x 0



Weka Knowledge Explorer

Filter: None

Current relation
Relation: iris
Instances: 150 Attributes: 5

Attributes

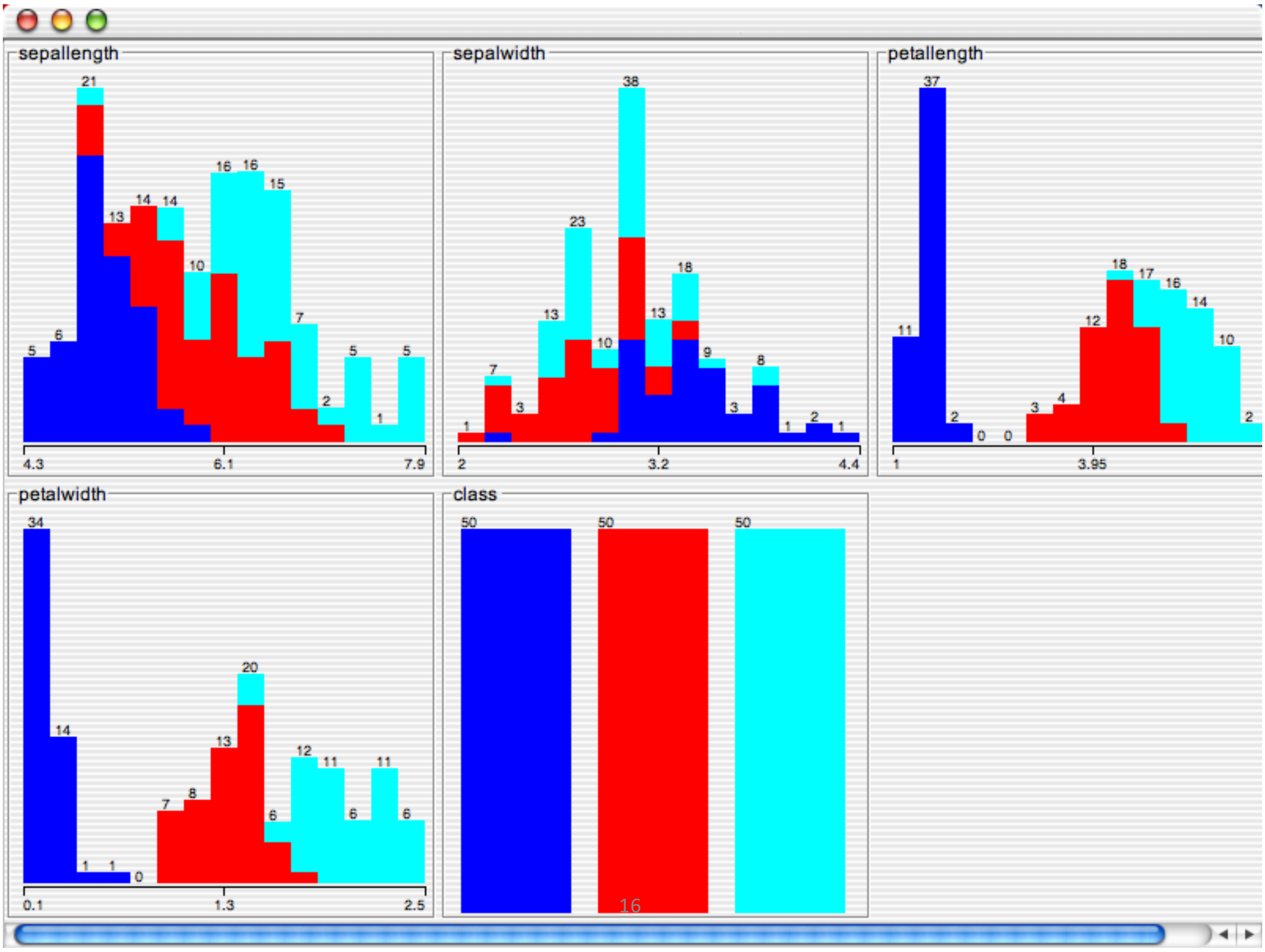
No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

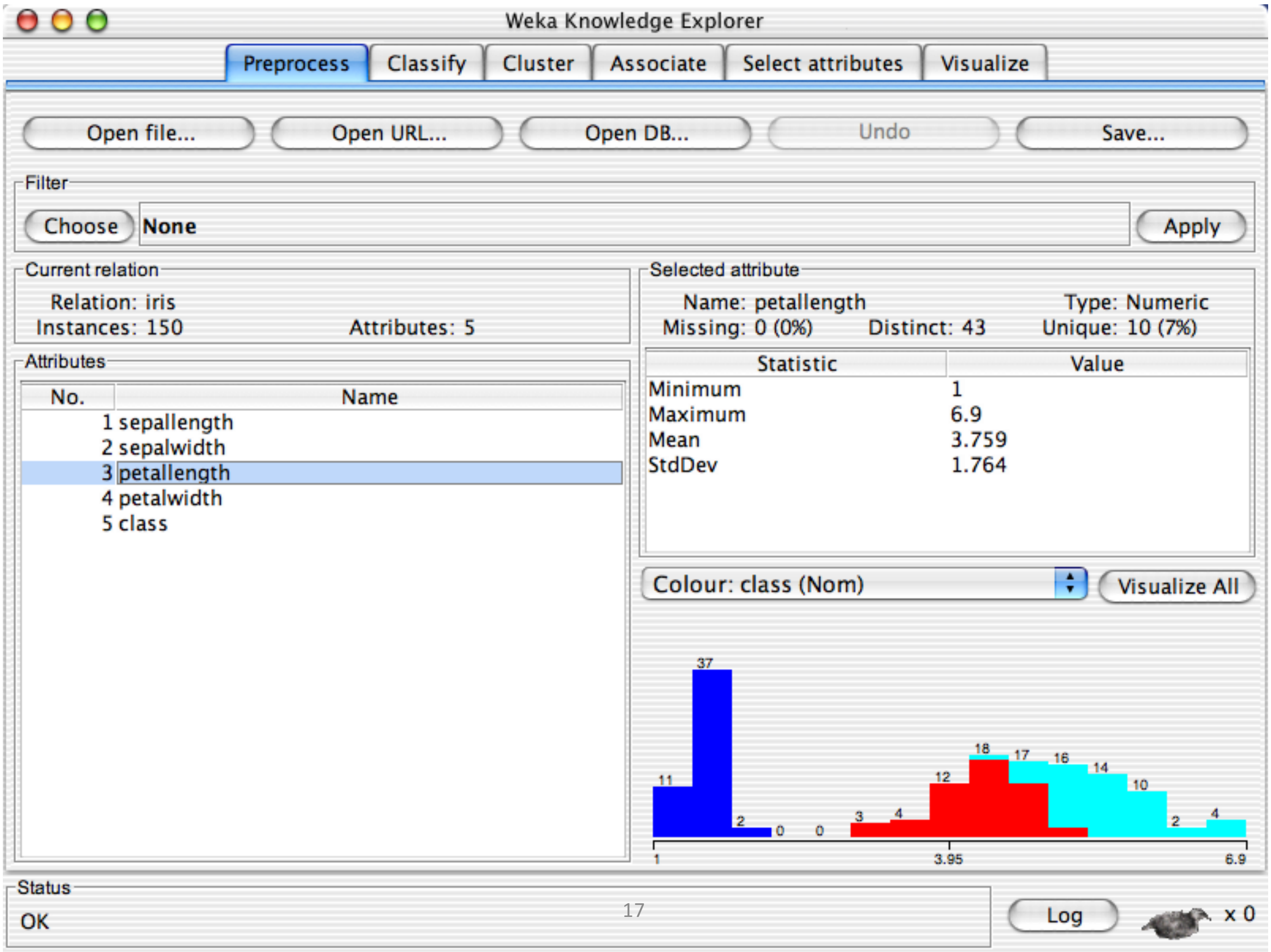
Selected attribute
Name: class Type: Nominal
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

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

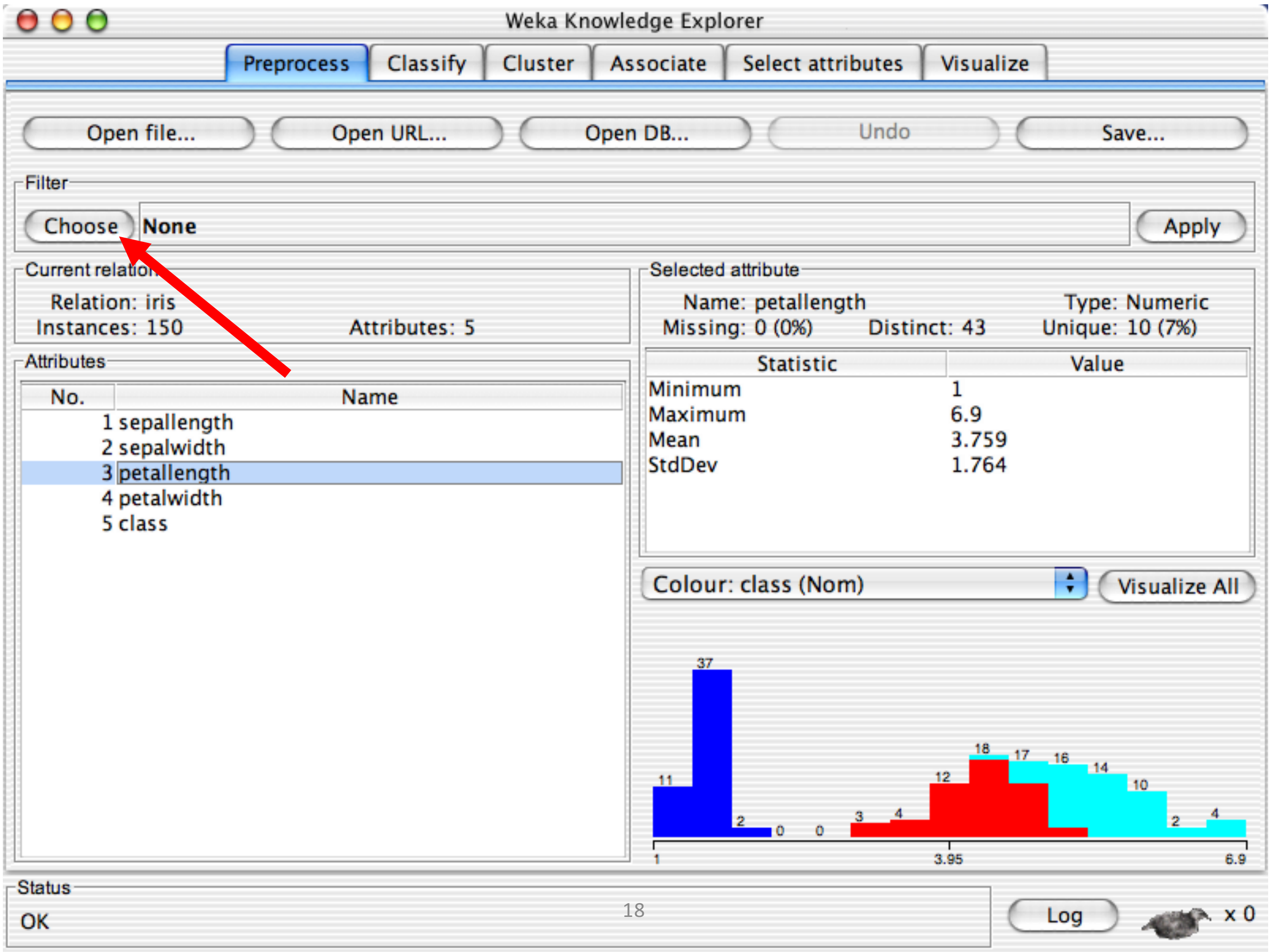
Name: petallength   Type: Numeric
Missing: 0 (0%)   Distinct: 43   Unique: 10 (7%)

Statistic	Value
Minimum	1
Maximum	6.9
Mean	3.759
StdDev	1.764

Colour: class (Nom)

Status

OK



Filter

None

Current relation

Relation: iris
Instances: 150   Attributes: 5

Attributes

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

Selected attribute

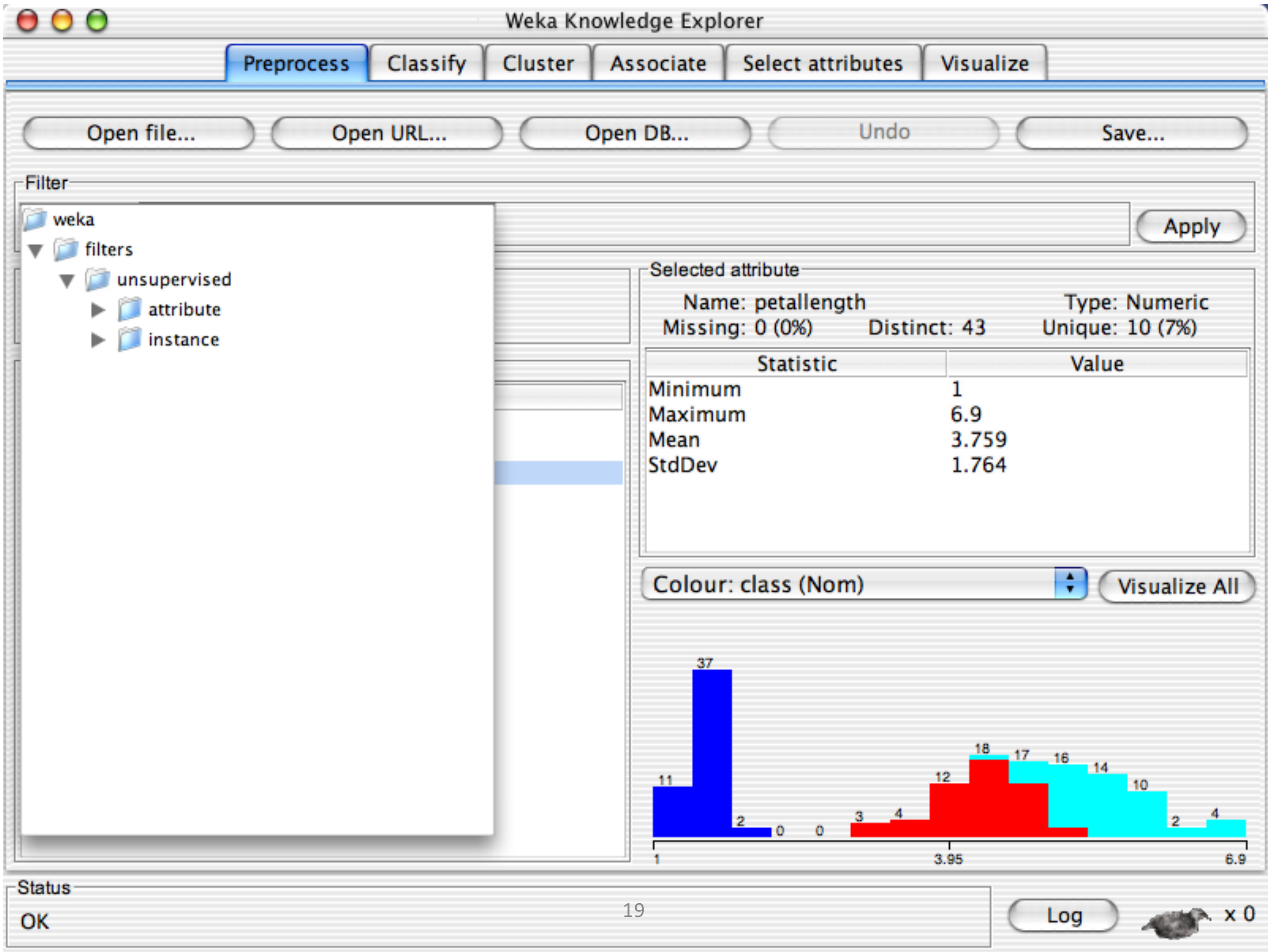
Name: petallength   Type: Numeric
Missing: 0 (0%)   Distinct: 43   Unique: 10 (7%)

Statistic	Value
Minimum	1
Maximum	6.9
Mean	3.759
StdDev	1.764

Colour: class (Nom)

Status

OK

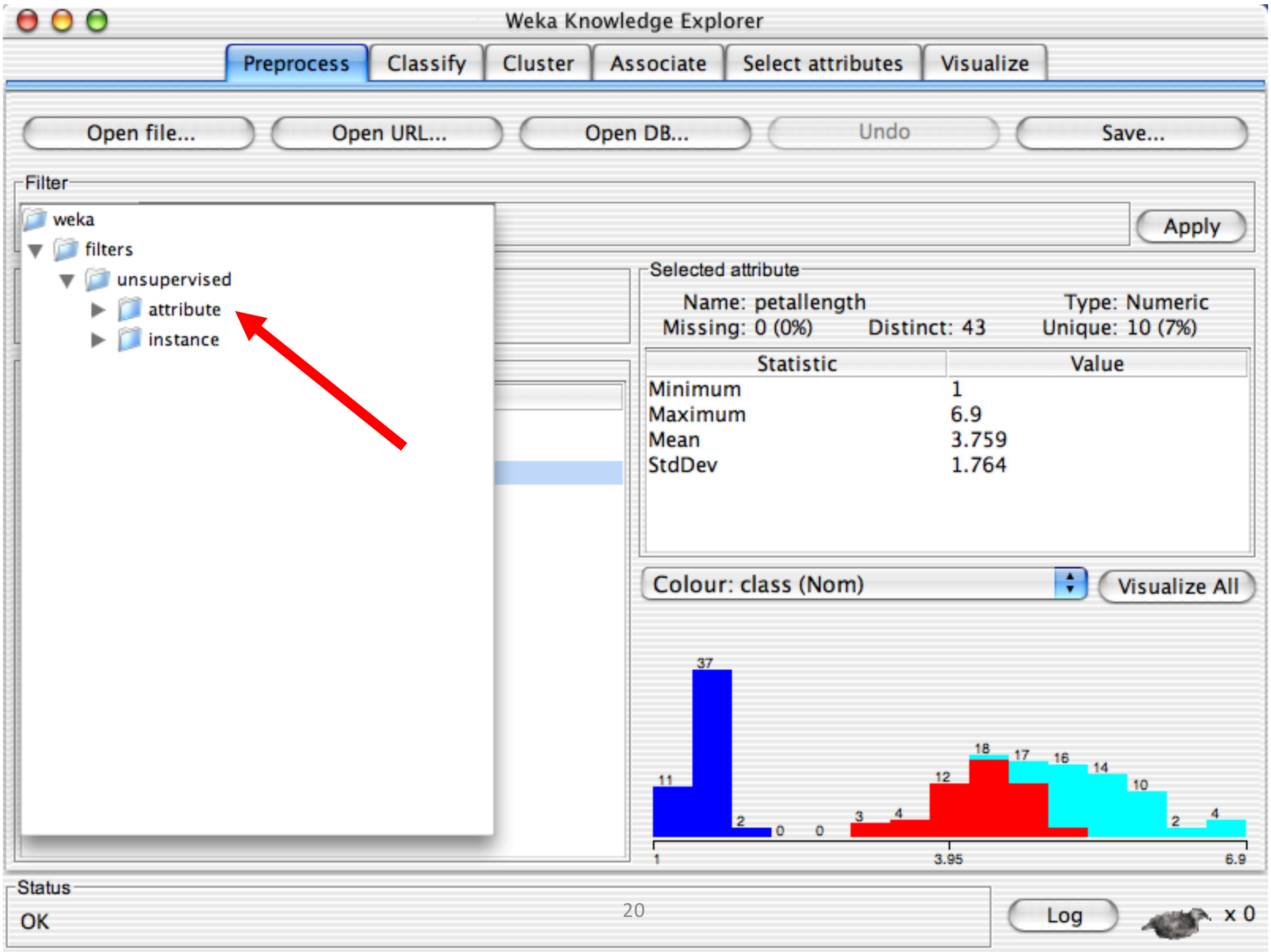


Selected attribute

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

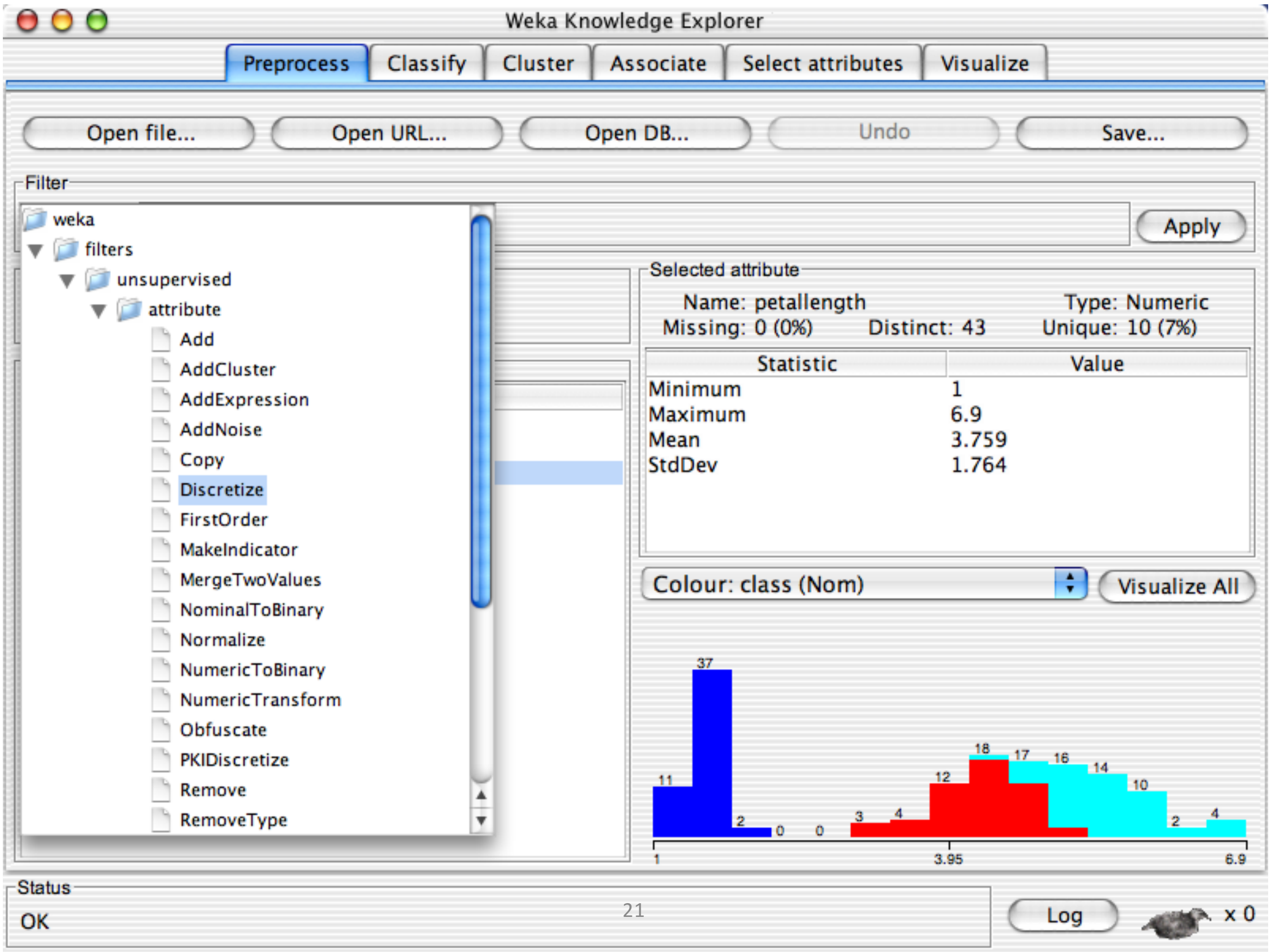


Selected attribute

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

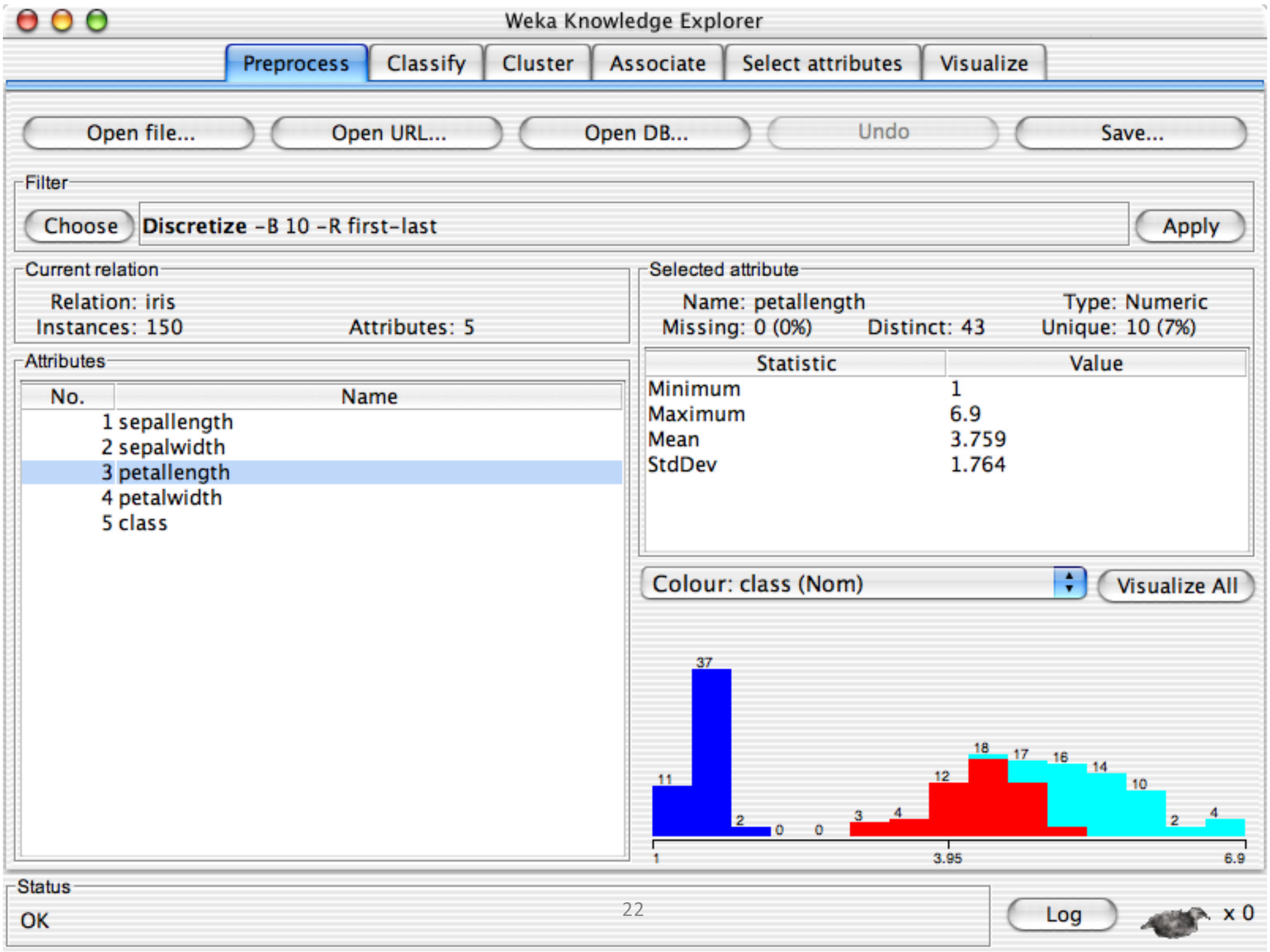


Selected attribute

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

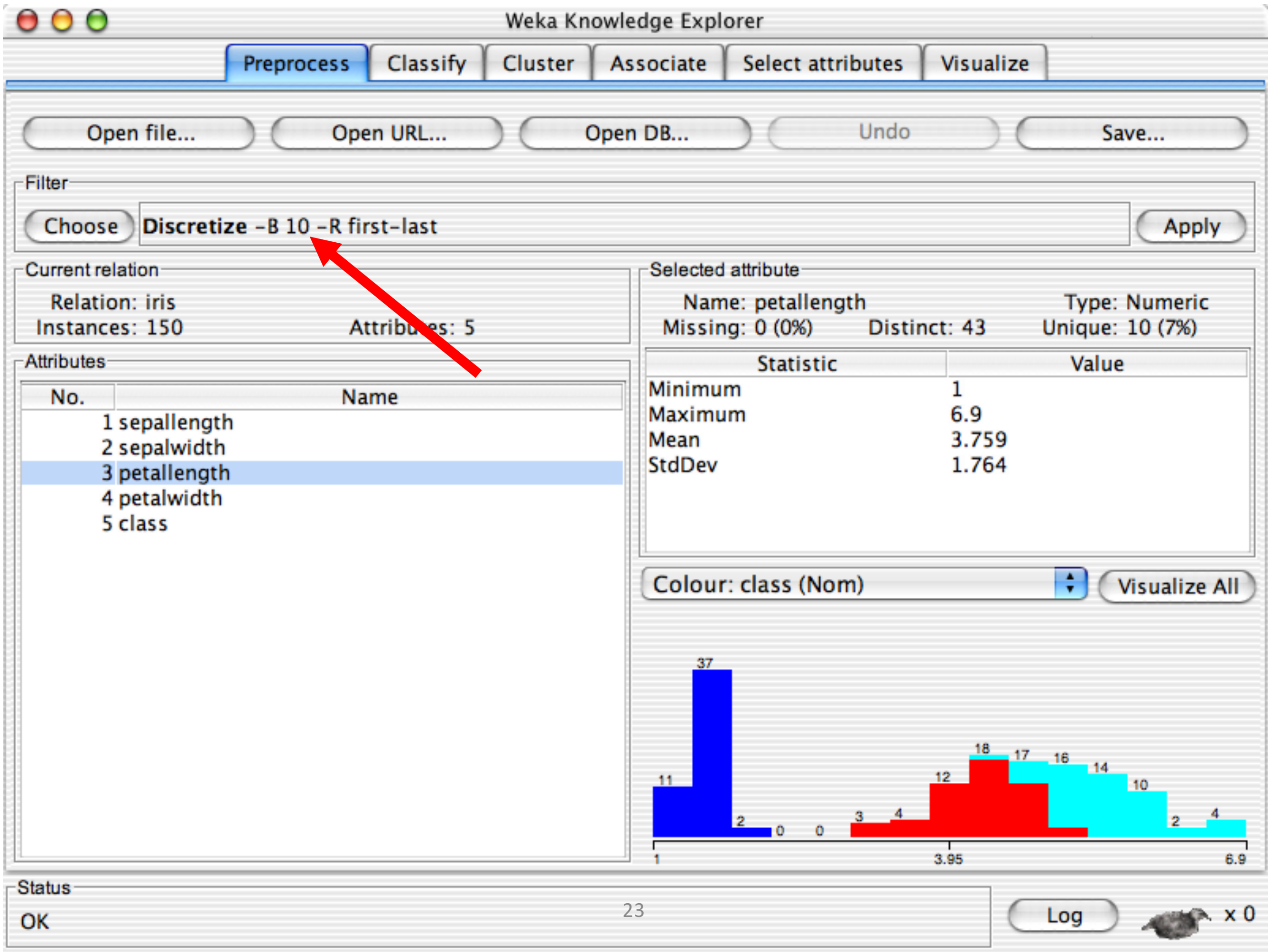
No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

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



Weka Knowledge Explorer

Filter: Discretize -B 10 -R first-last

Current relation
Relation: iris
Instances: 150
Attributes: 5

Attributes

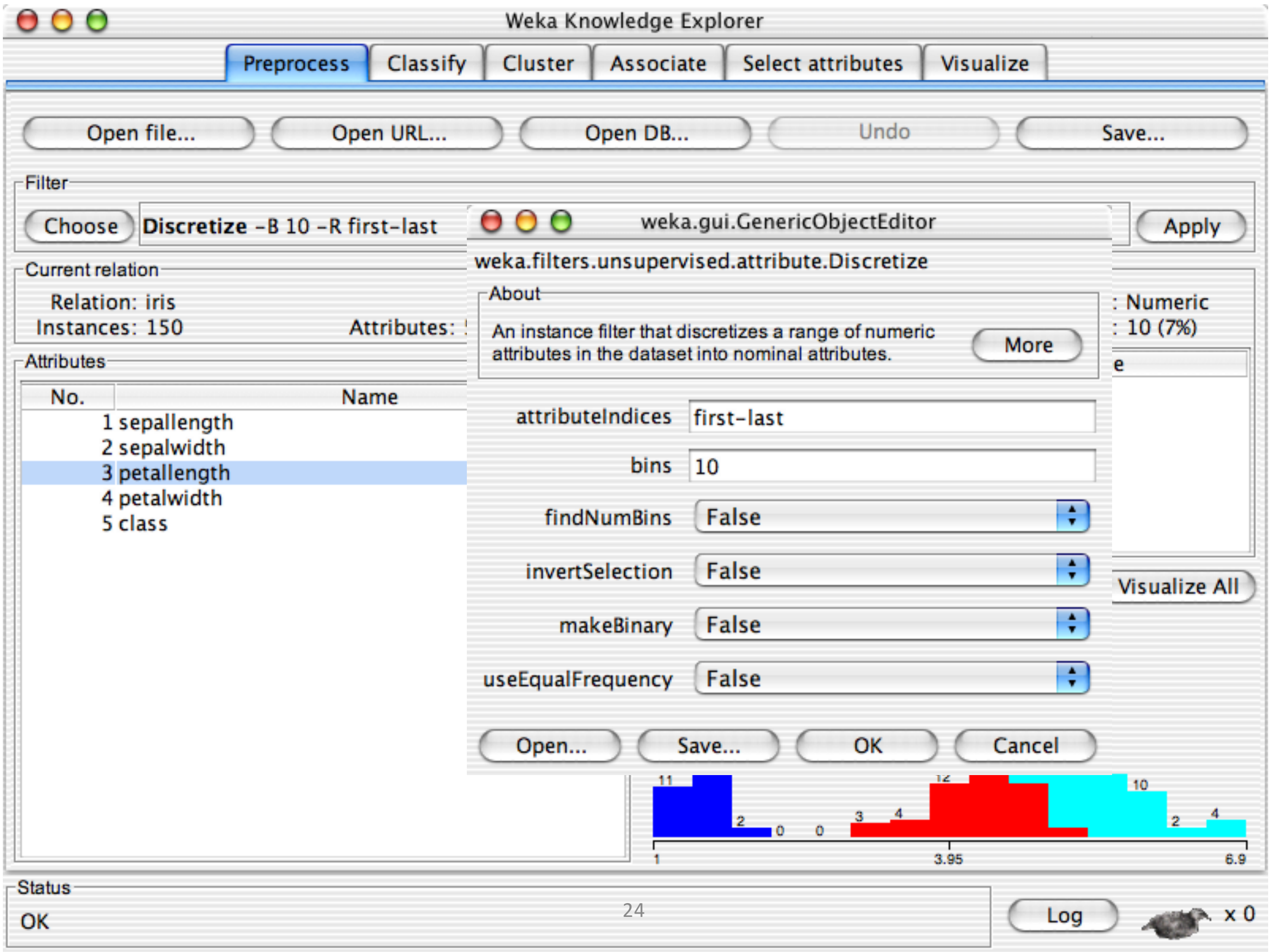
No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

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



Weka Knowledge Explorer

Filter: Discretize -B 10 -R first-last

Current relation
Relation: iris
Instances: 150

Attributes

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

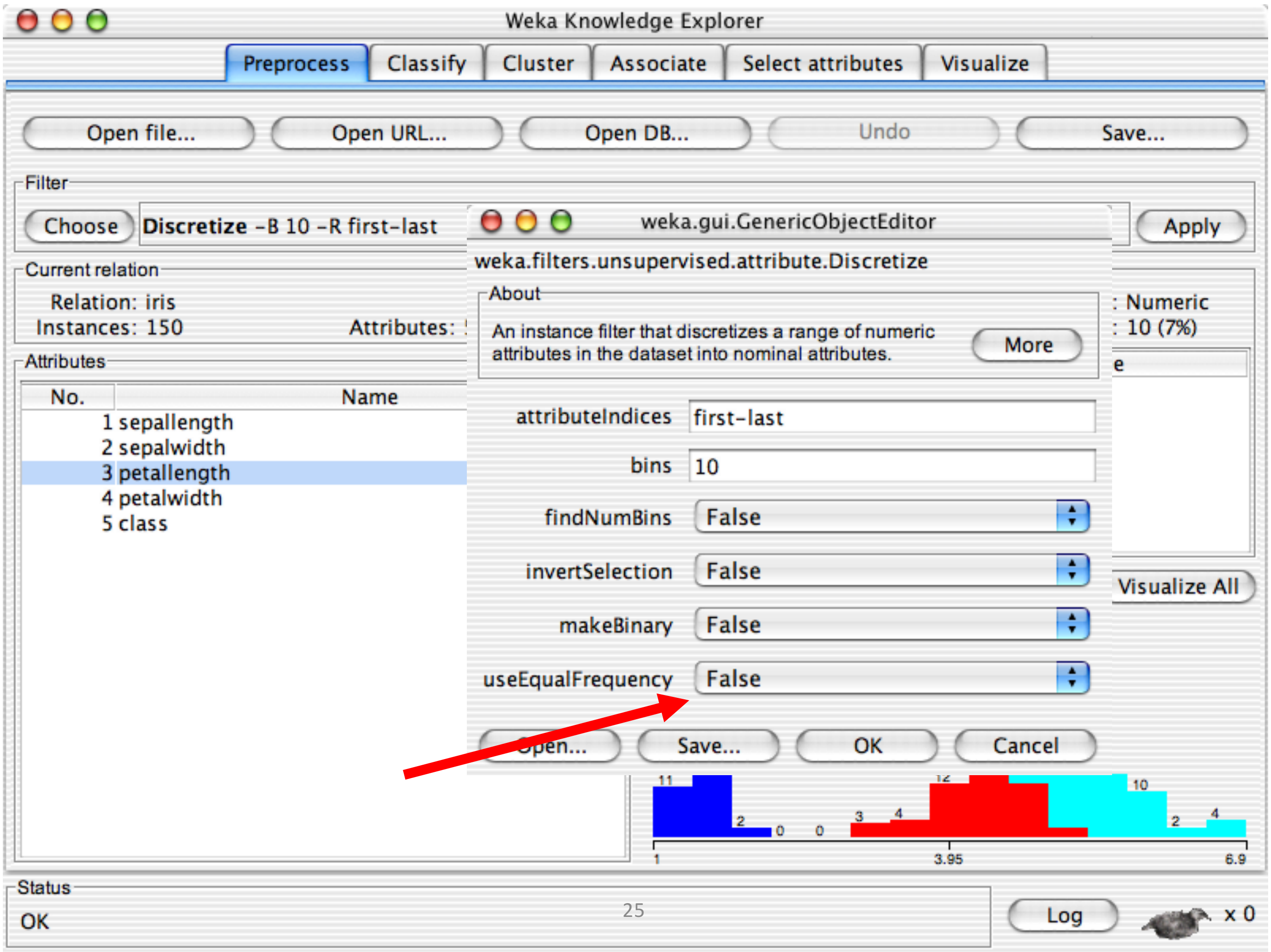
weka.gui.GenericObjectEditor

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

Status
OK



Discretize -B 10 -R first-last

weka.gui.GenericObjectEditor

weka.filters.unsupervised.attribute.Discretize

About

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

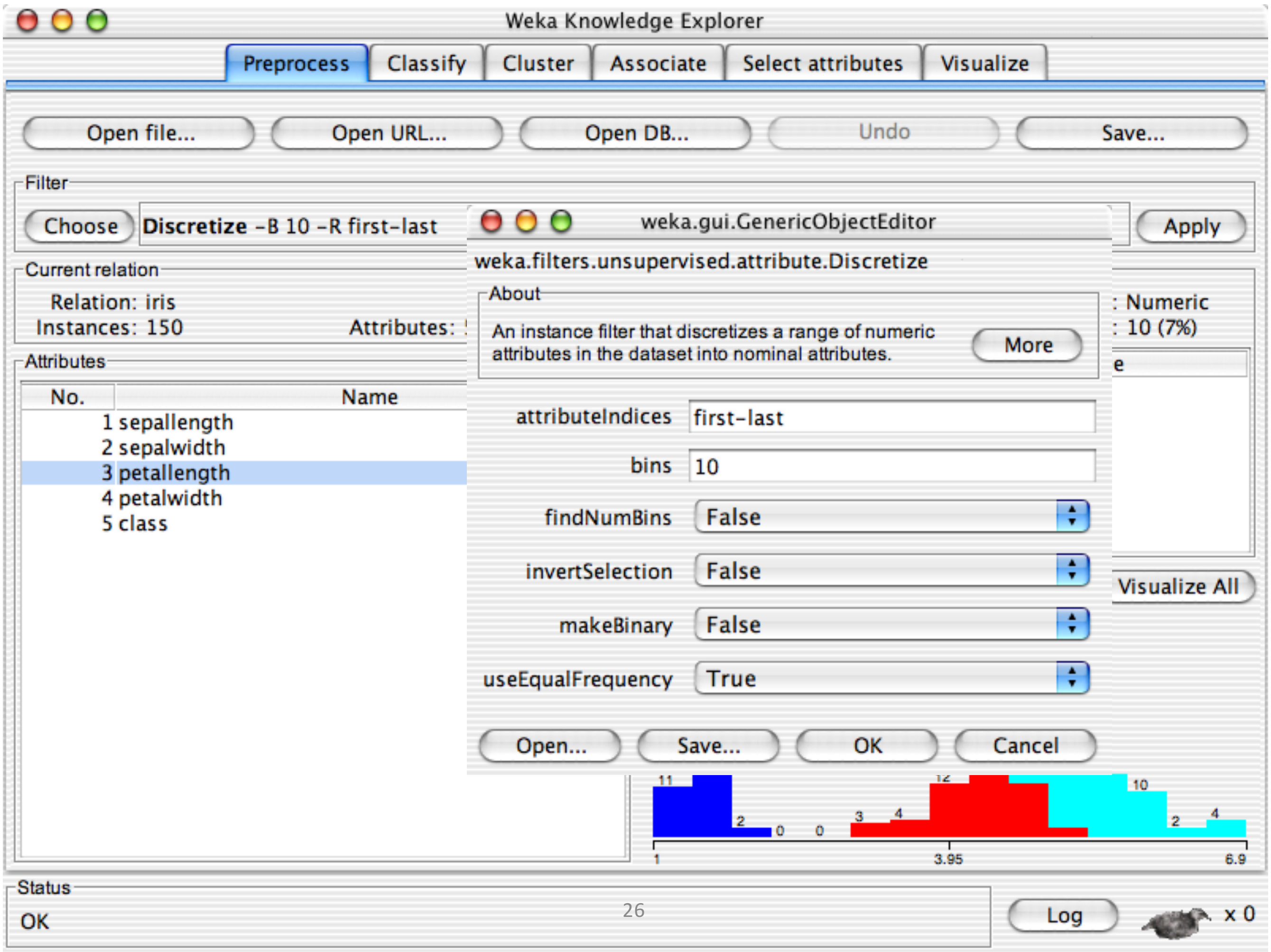
Relation: iris
Instances: 150
Attributes: 5

1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class

Numeric
10 (7%)

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

OK



Discretize -B 10 -R first-last

weka.gui.GenericObjectEditor

weka.filters.unsupervised.attribute.Discretize

About

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

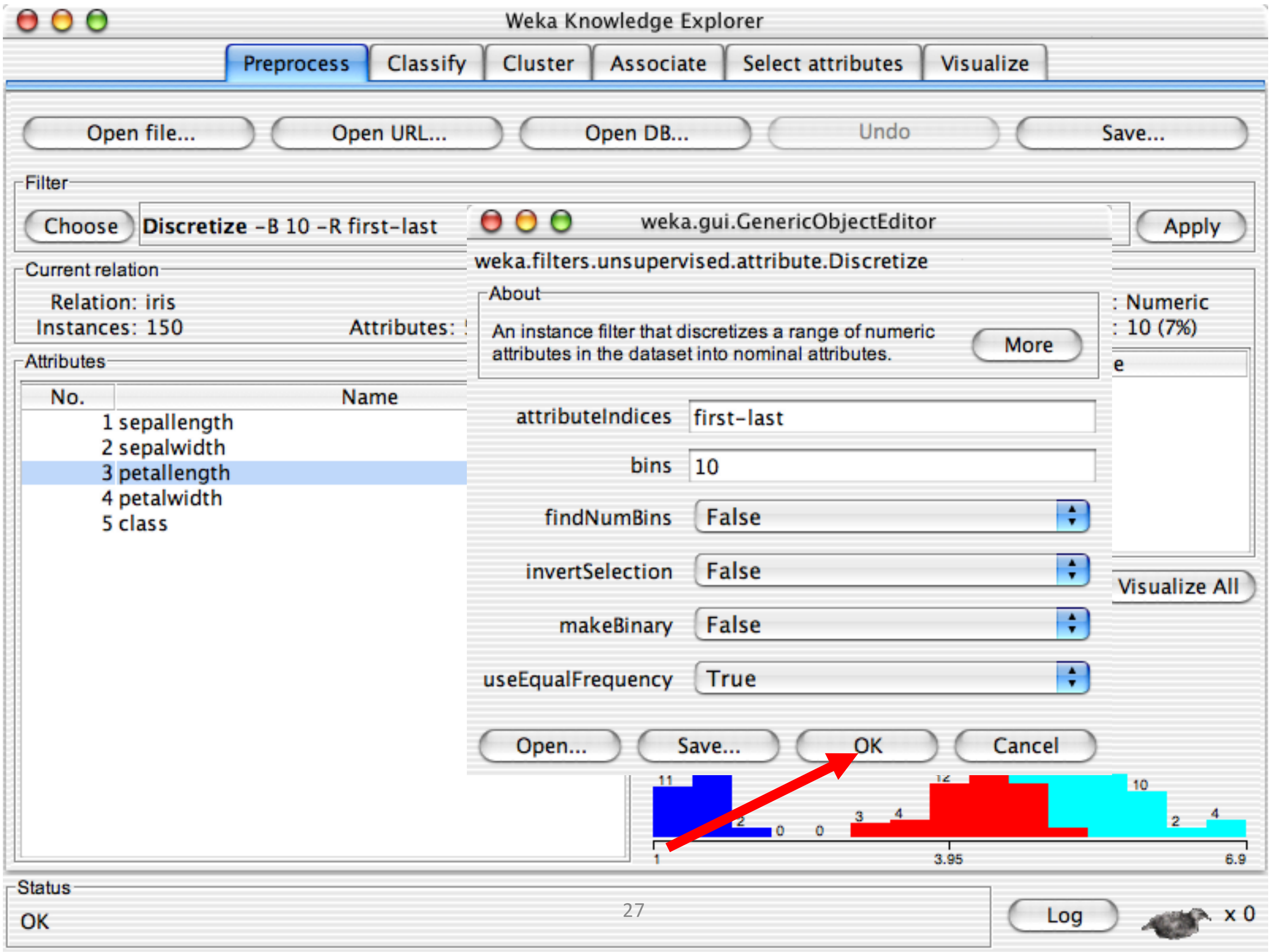
Relation: iris
Instances: 150
Attributes: 5

1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class

Numeric
10 (7%)

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

OK



Discretize -B 10 -R first-last

Relation: iris
Instances: 150
Attributes: 5

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

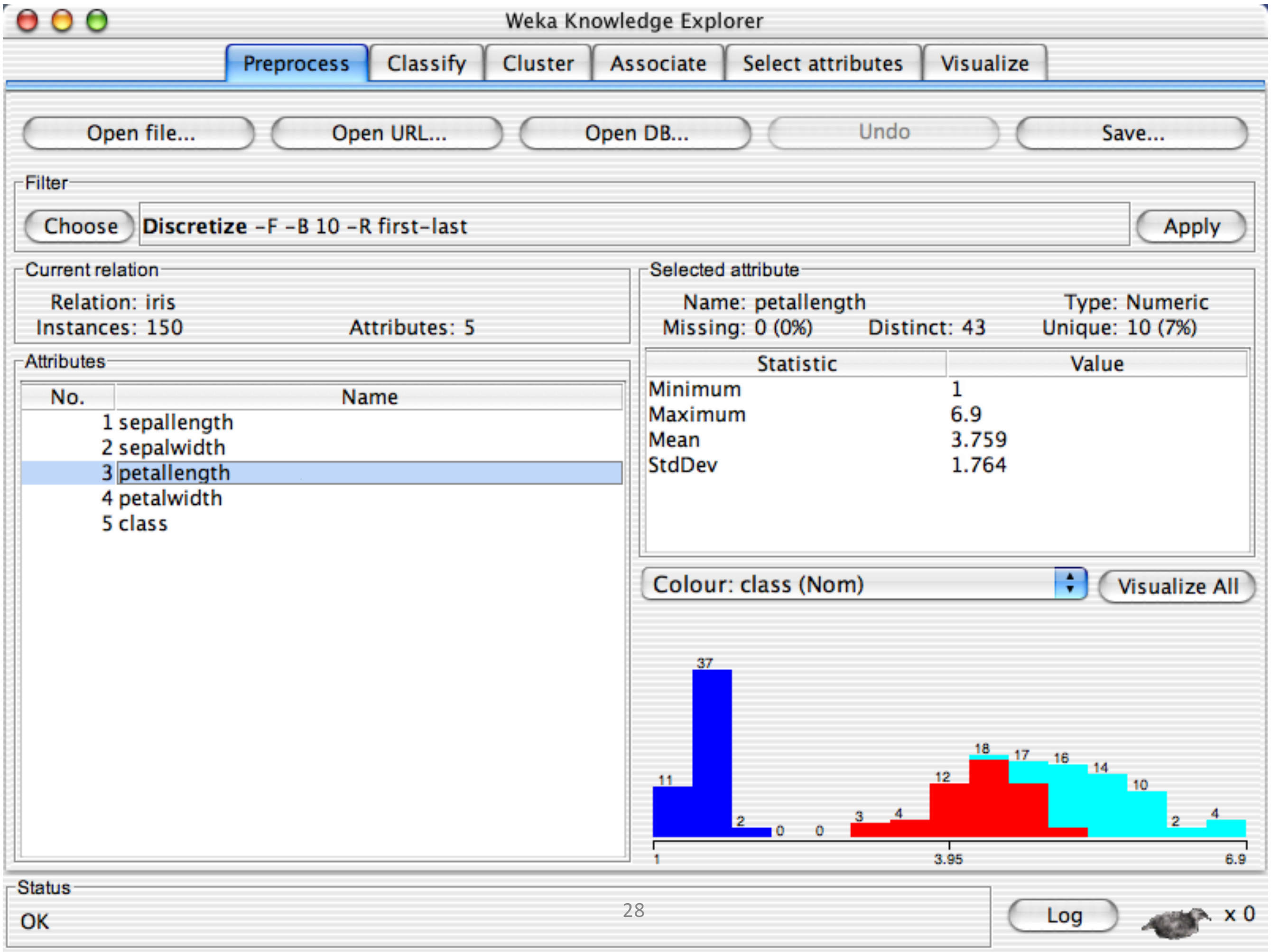
invertSelection False

makeBinary False

useEqualFrequency True

: Numeric
: 10 (7%)

OK



Discretize -F -B 10 -R first-last

Relation: iris
Instances: 150
Attributes: 5

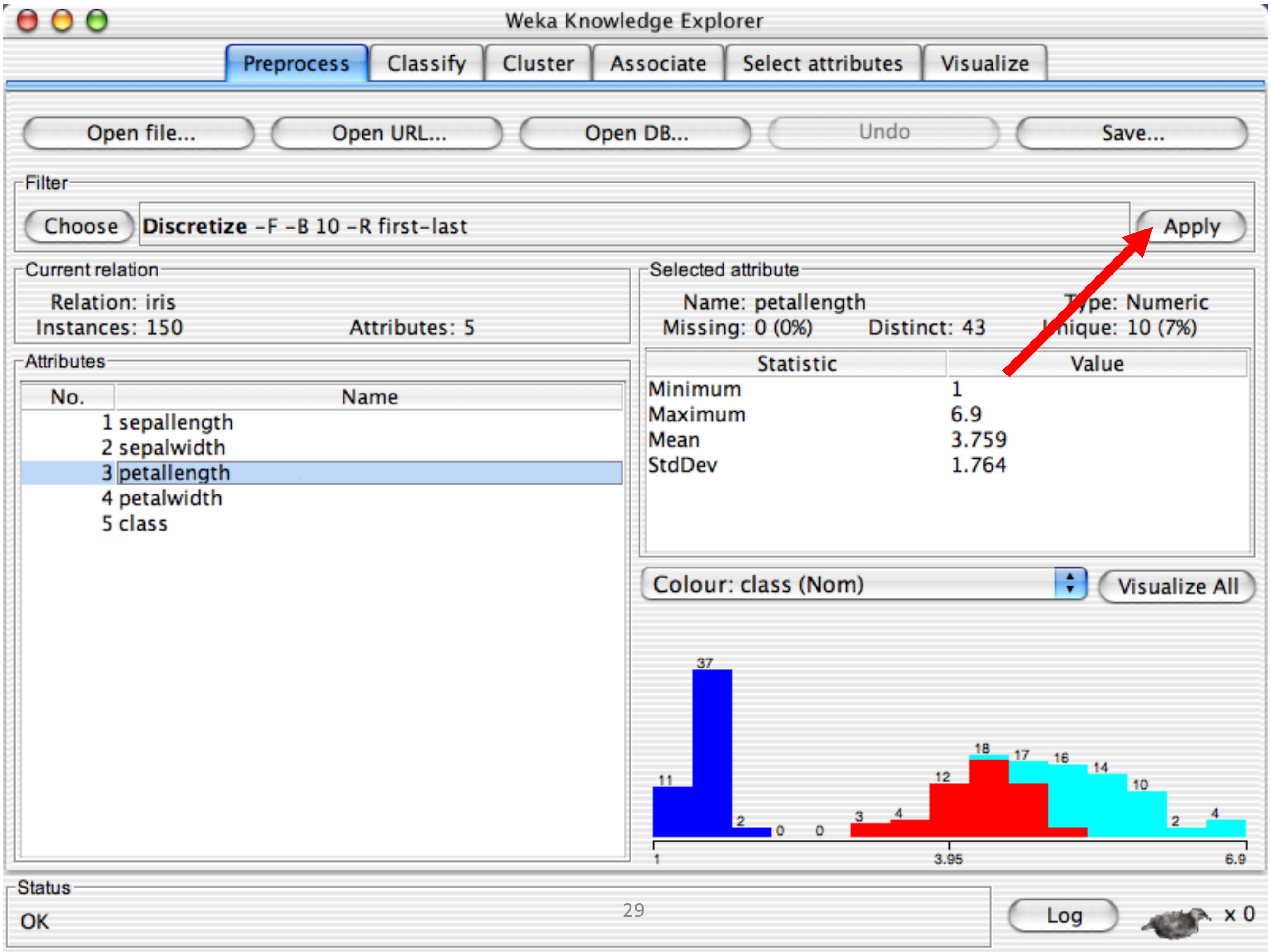
No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

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

Colour: class (Nom)

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

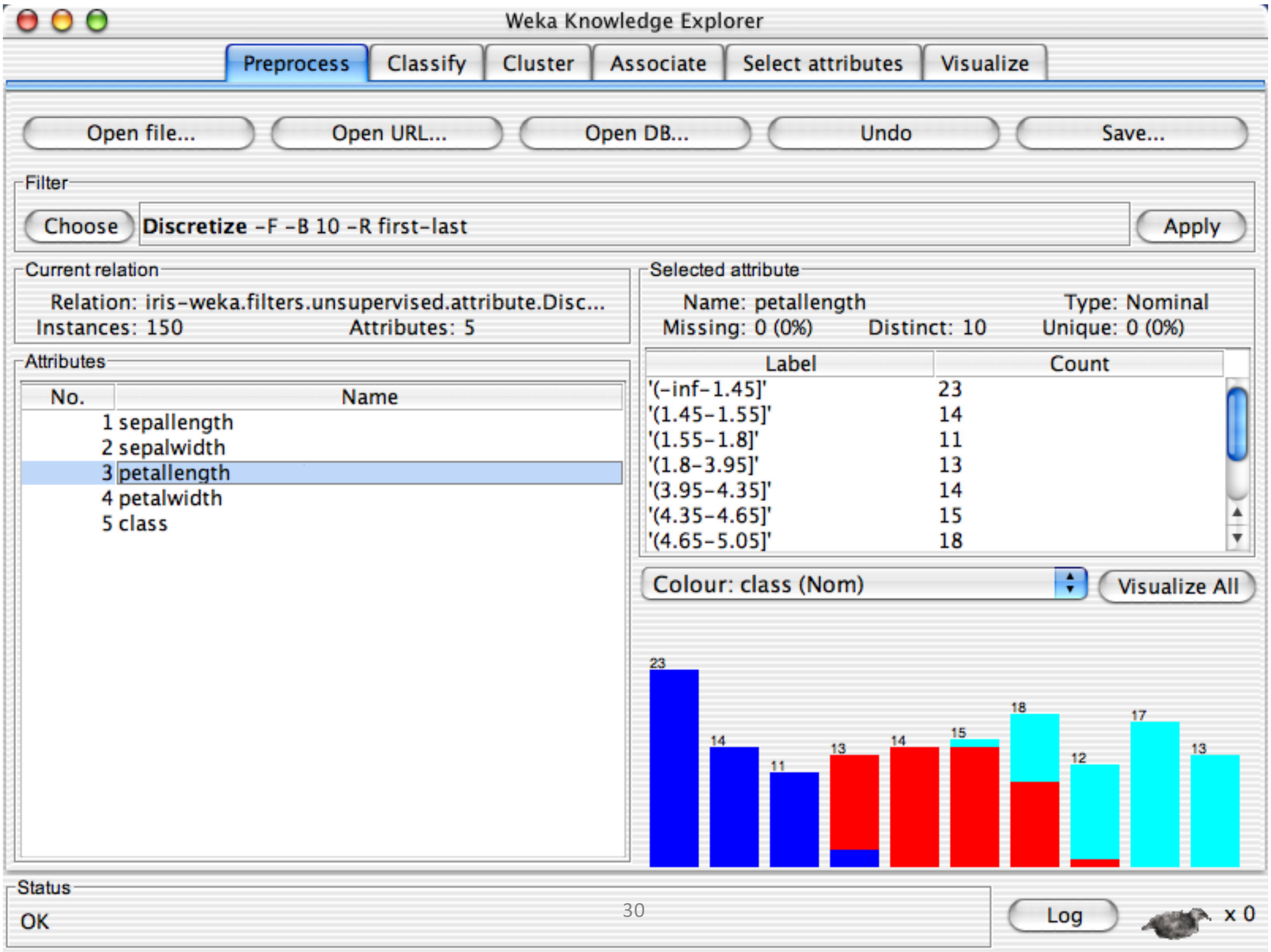
Selected attribute

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%)

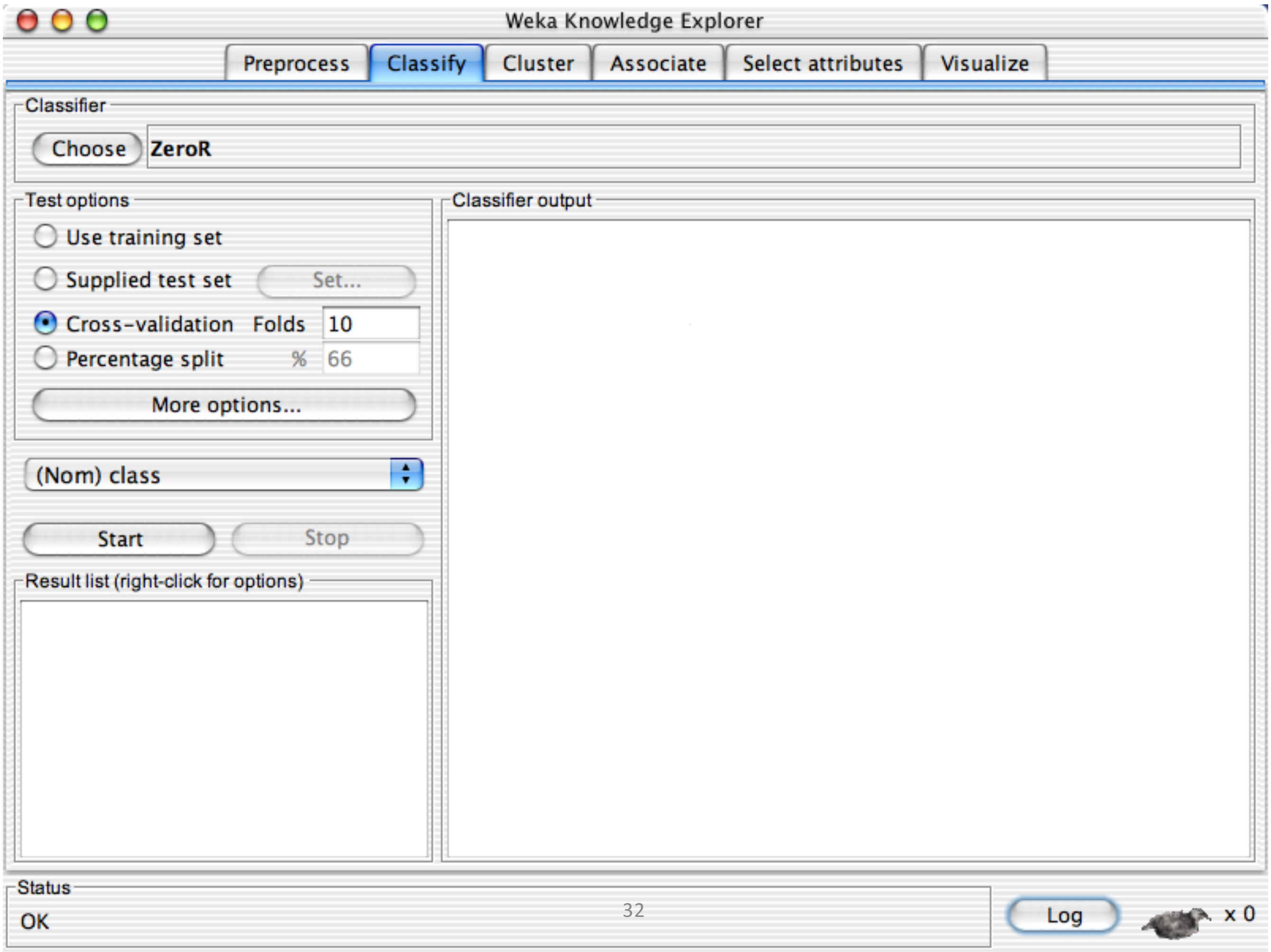
Label	Count
'(-inf-1.45]'	23
'(1.45-1.55]'	14
'(1.55-1.8]'	11
'(1.8-3.95]'	13
'(3.95-4.35]'	14
'(4.35-4.65]'	15
'(4.65-5.05]'	18

Colour: class (Nom)

Status
OK

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, …



Weka Knowledge Explorer

Classifier

ZeroR

Test options

○ Use training set

○ Supplied test set

● Cross-validation Folds 10

○ Percentage split % 66

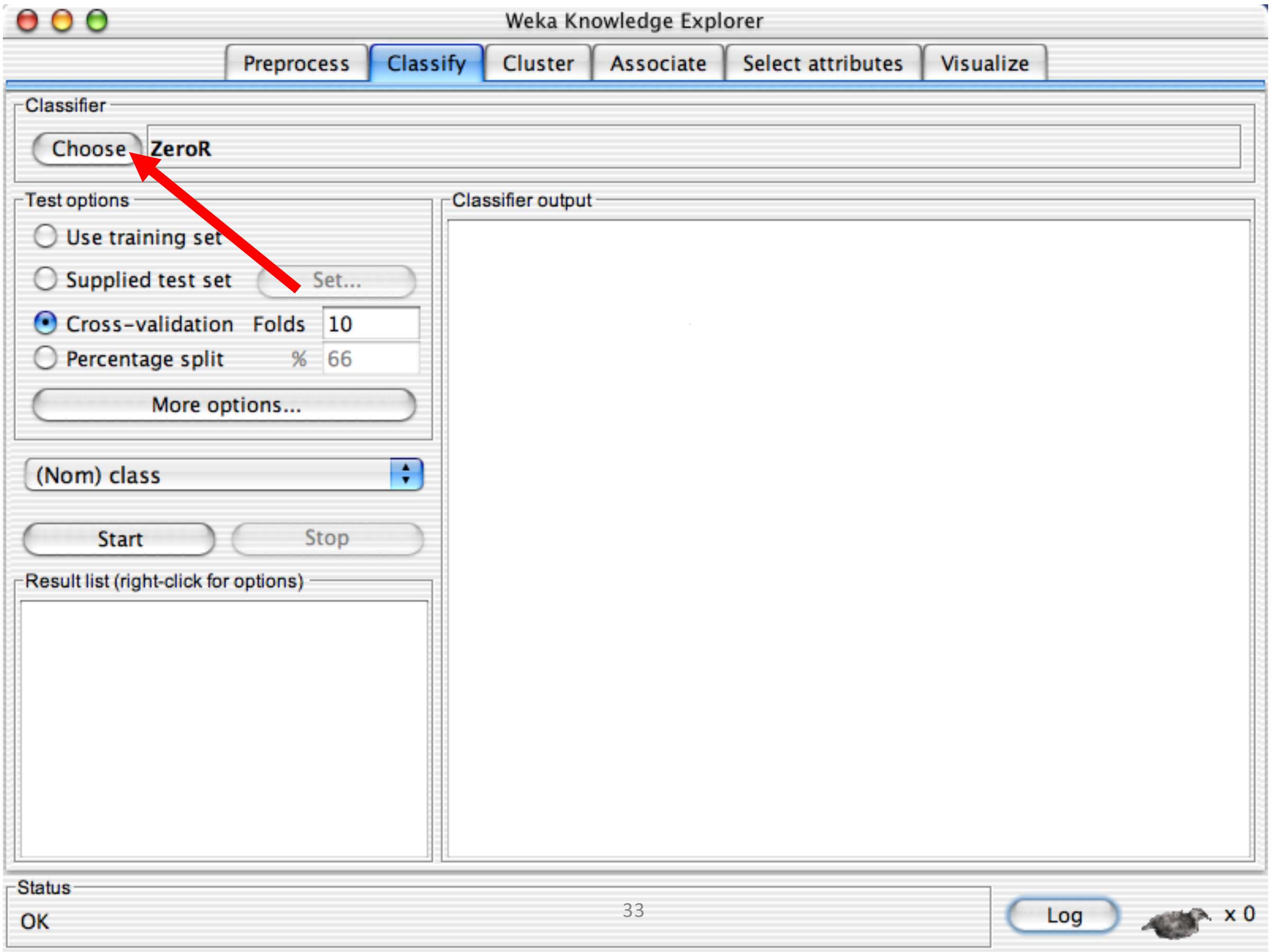
(Nom) class

Result list (right-click for options)

Classifier output

Status

OK



Classifier: ZeroR

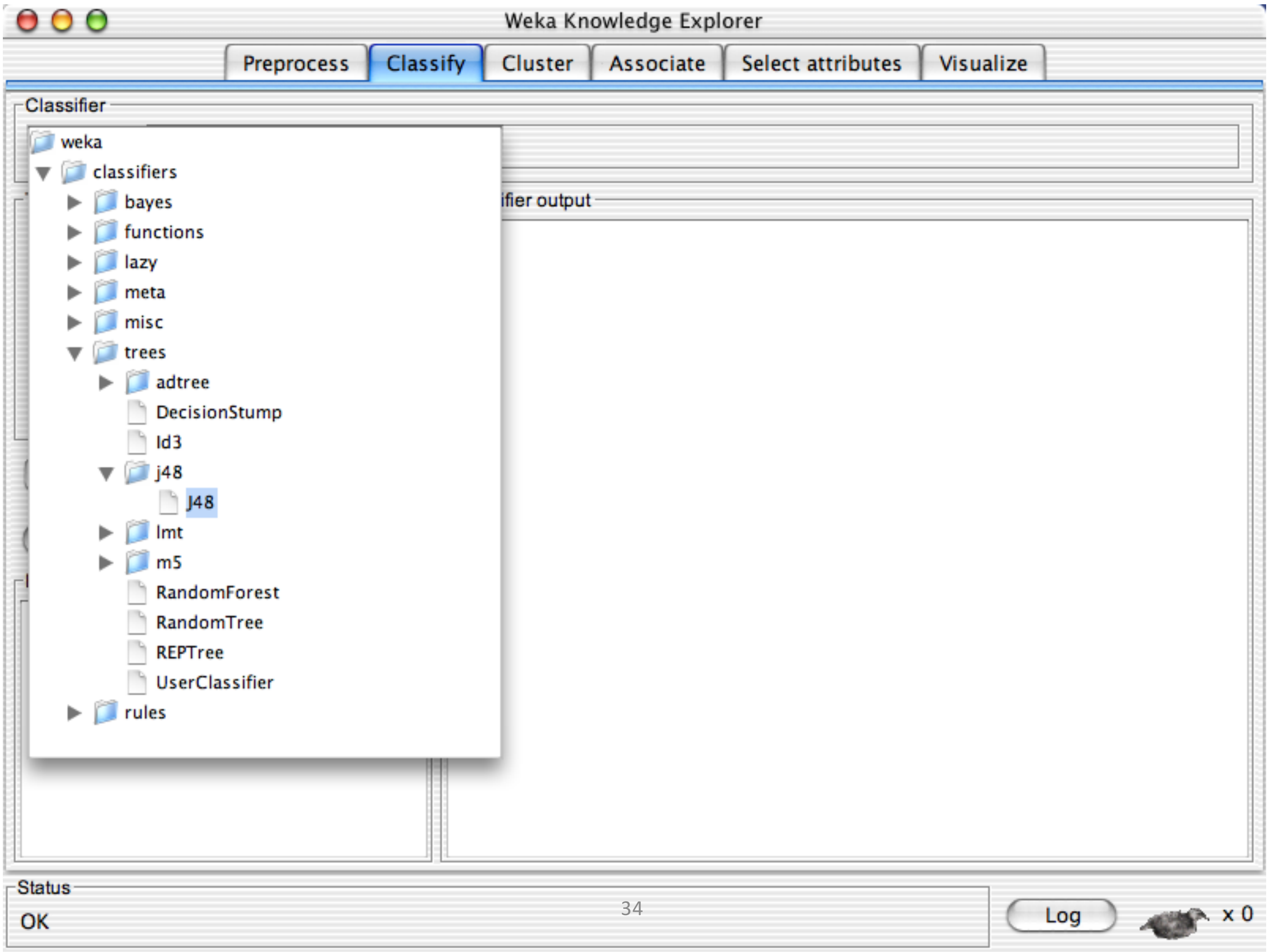
Test options
- ○ Use training set
- ○ Supplied test set
- ● Cross-validation Folds 10
- ○ Percentage split % 66

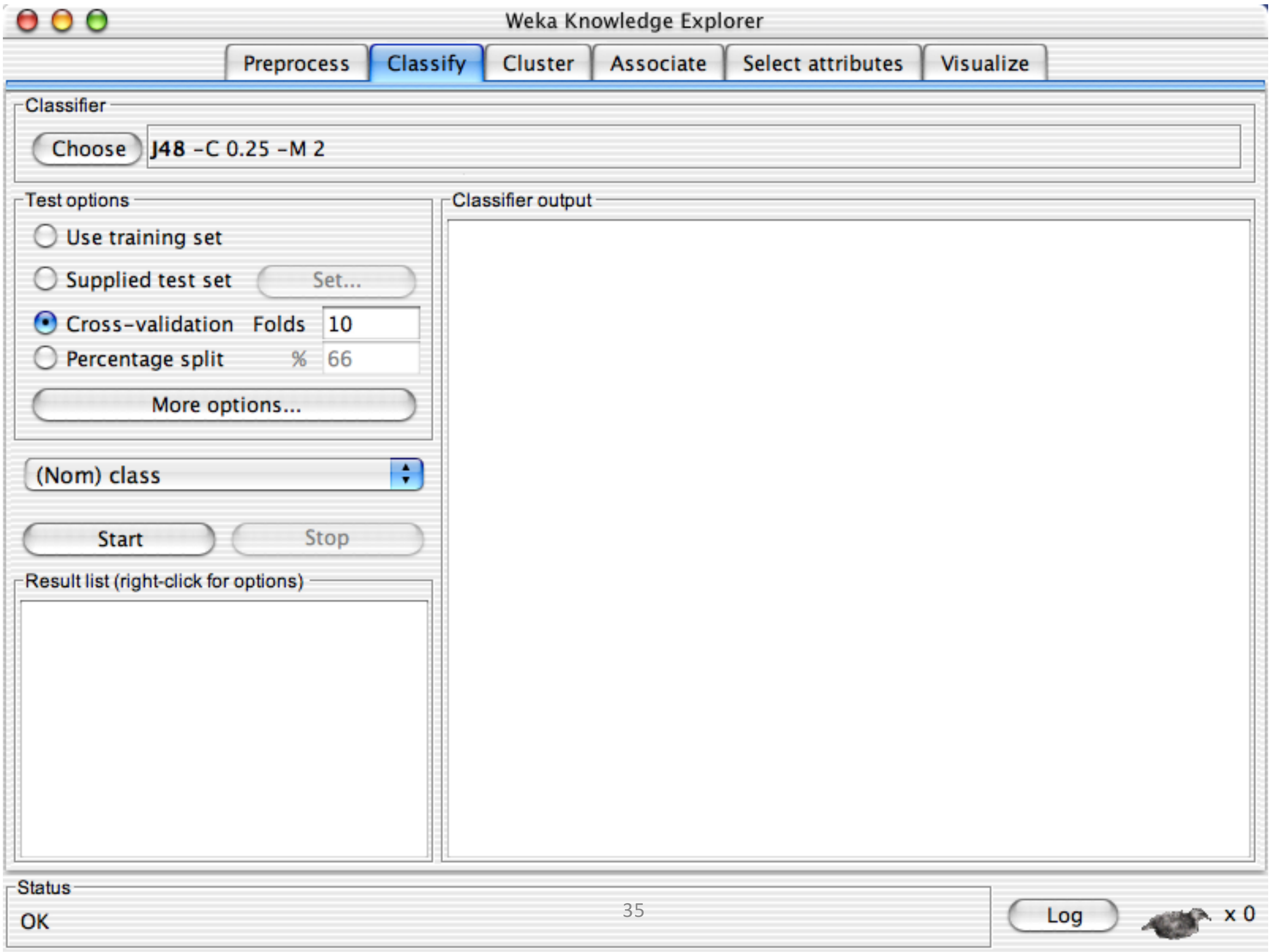
(Nom) class

Classifier output

Result list (right-click for options)

Status
OK





J48 -C 0.25 -M 2

○ Use training set

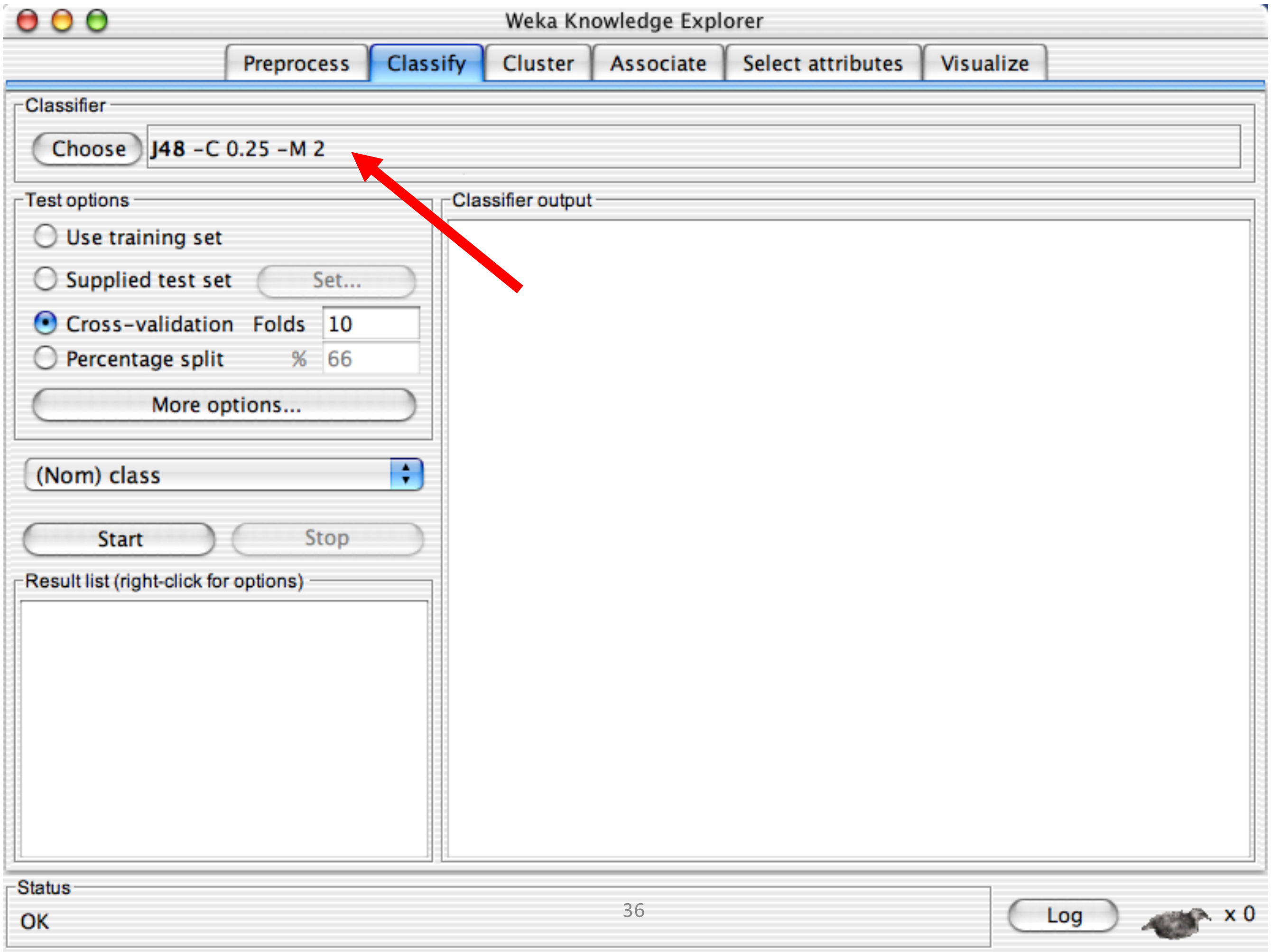
○ Supplied test set

● Cross-validation Folds 10

○ Percentage split % 66

(Nom) class

OK



J48 -C 0.25 -M 2

○ Use training set

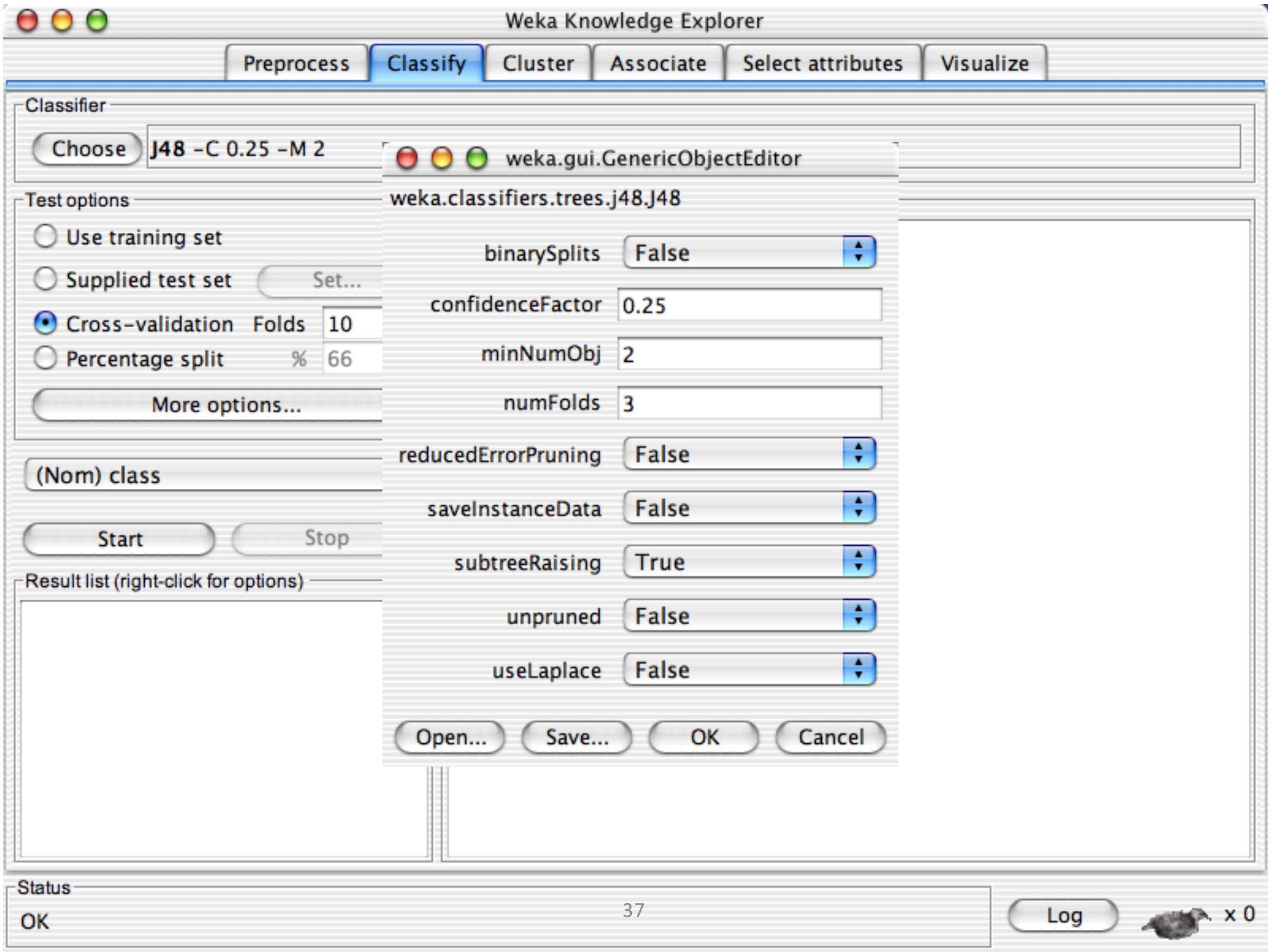
○ Supplied test set

● Cross-validation Folds 10

○ Percentage split % 66

(Nom) class

OK



J48 -C 0.25 -M 2

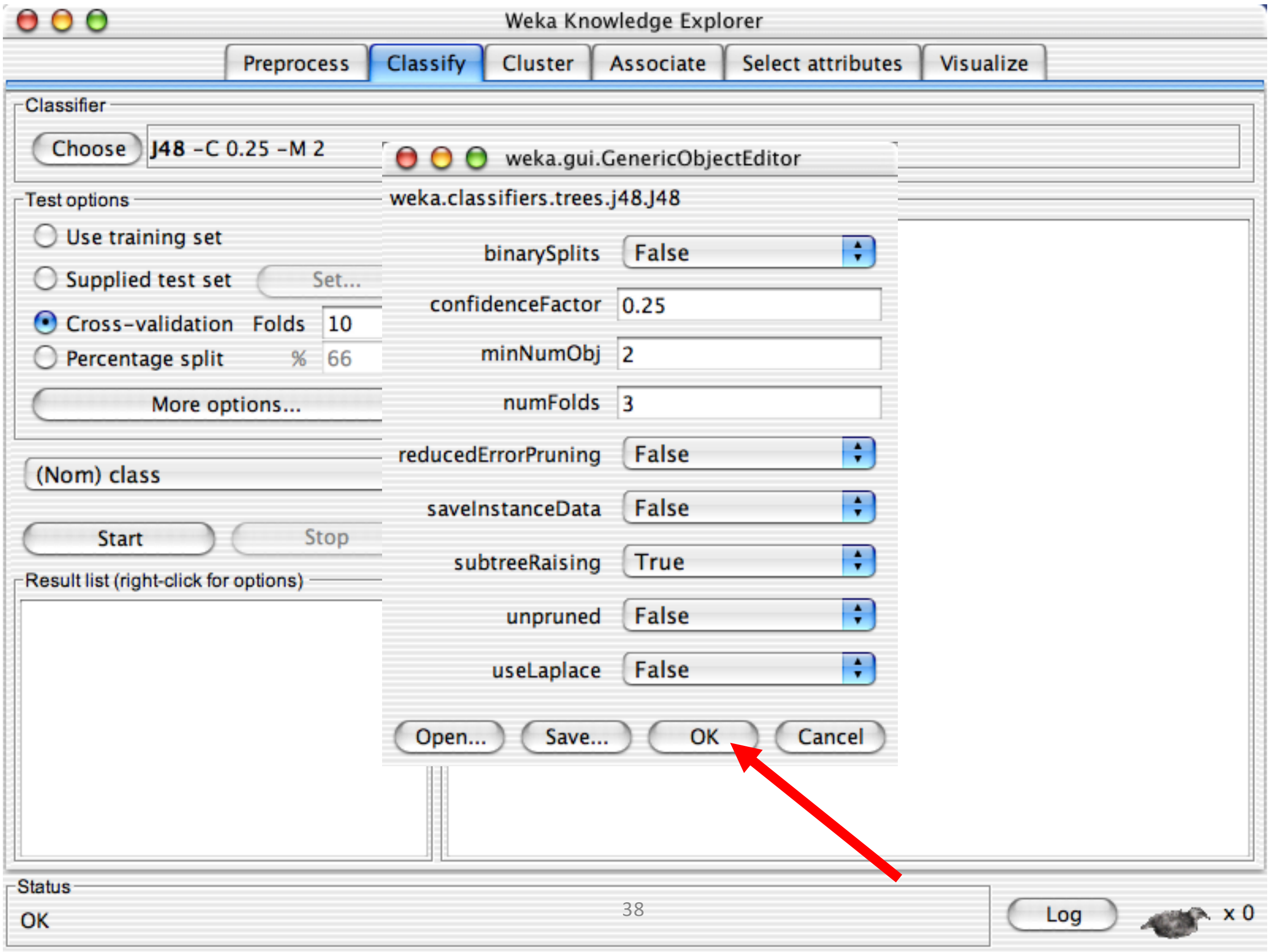
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(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

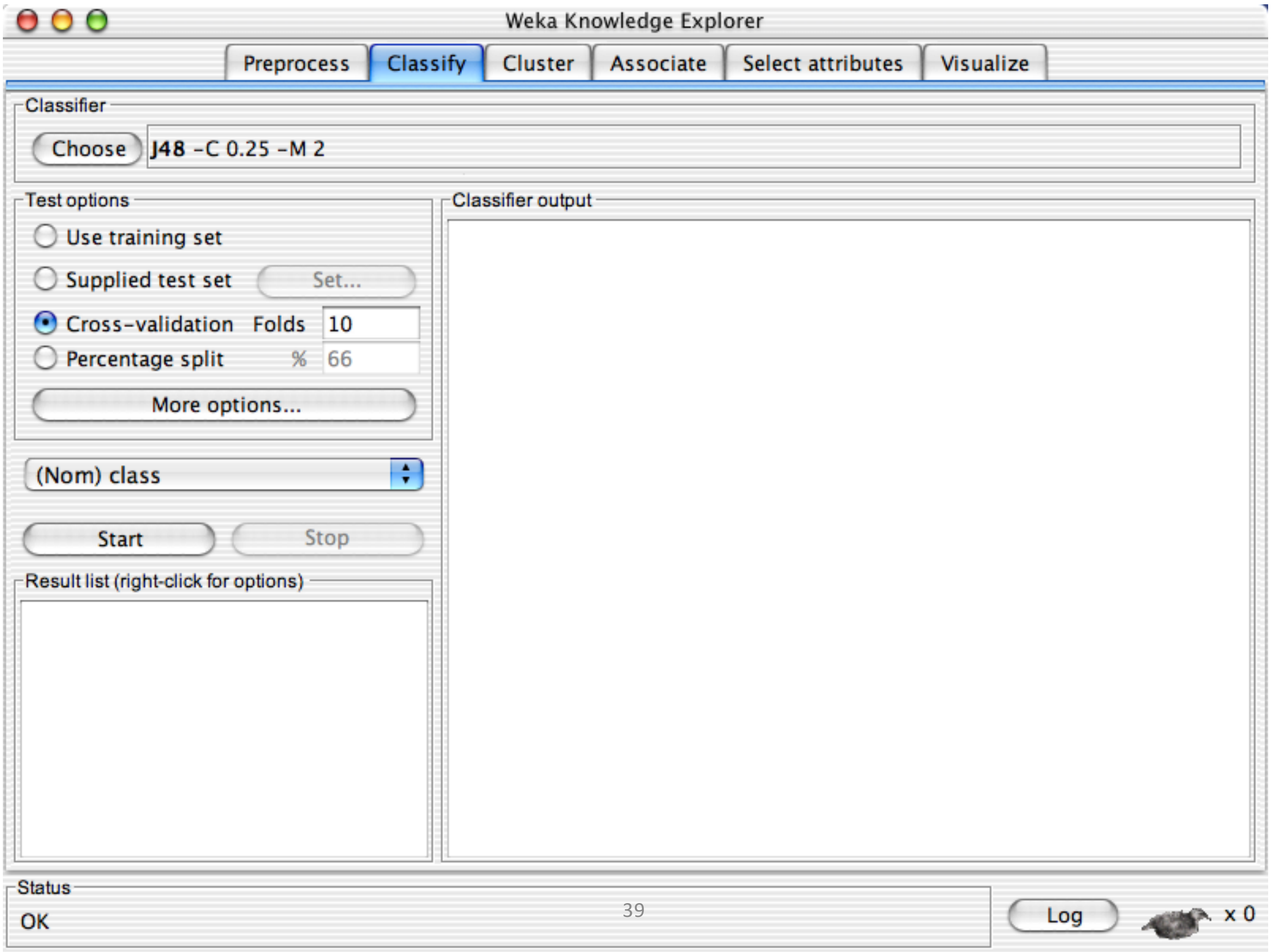
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(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



J48 -C 0.25 -M 2

○ Use training set

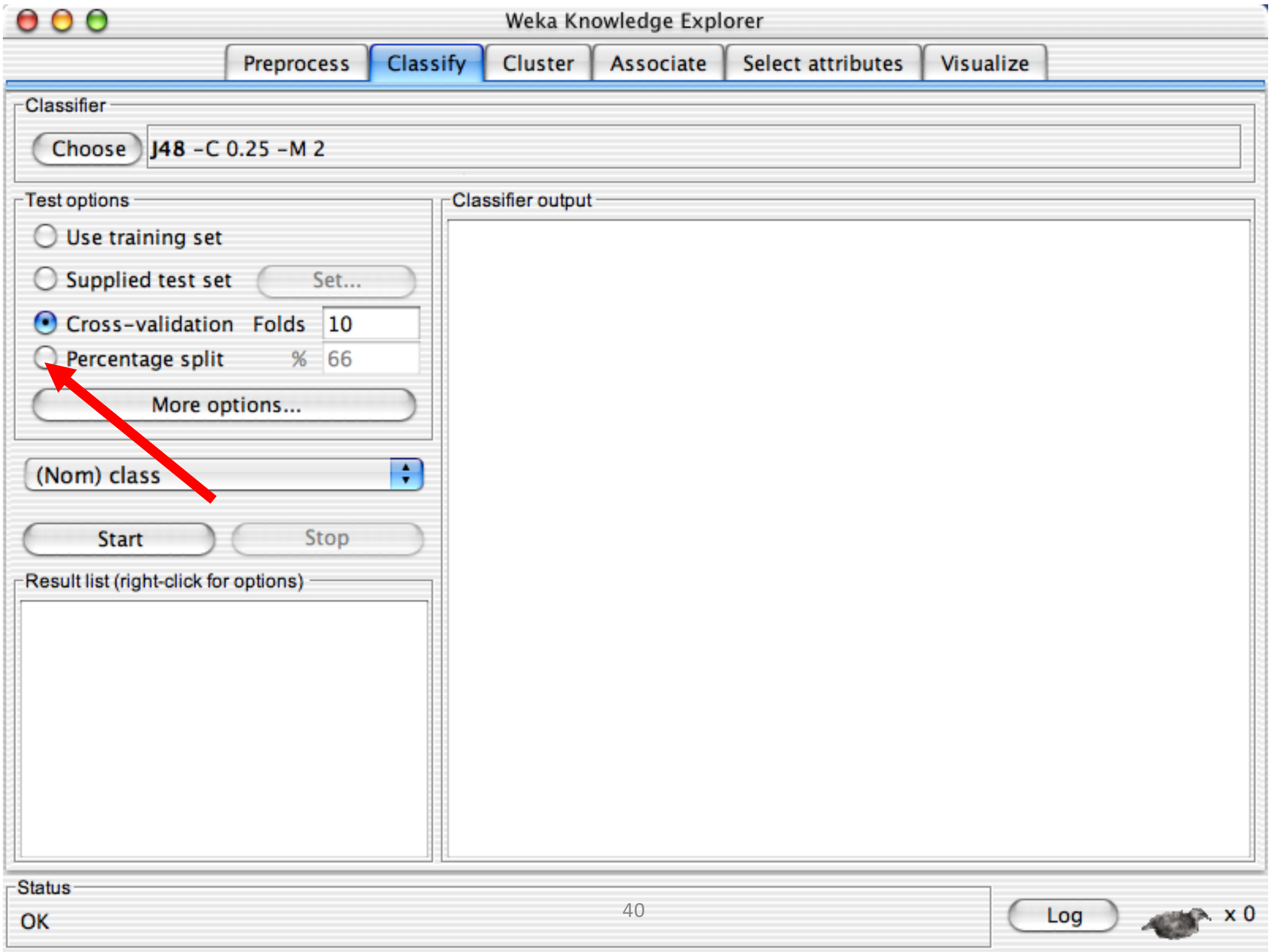
○ Supplied test set

● Cross-validation Folds 10

○ Percentage split % 66

(Nom) class

Status: OK



J48 -C 0.25 -M 2

○ Use training set

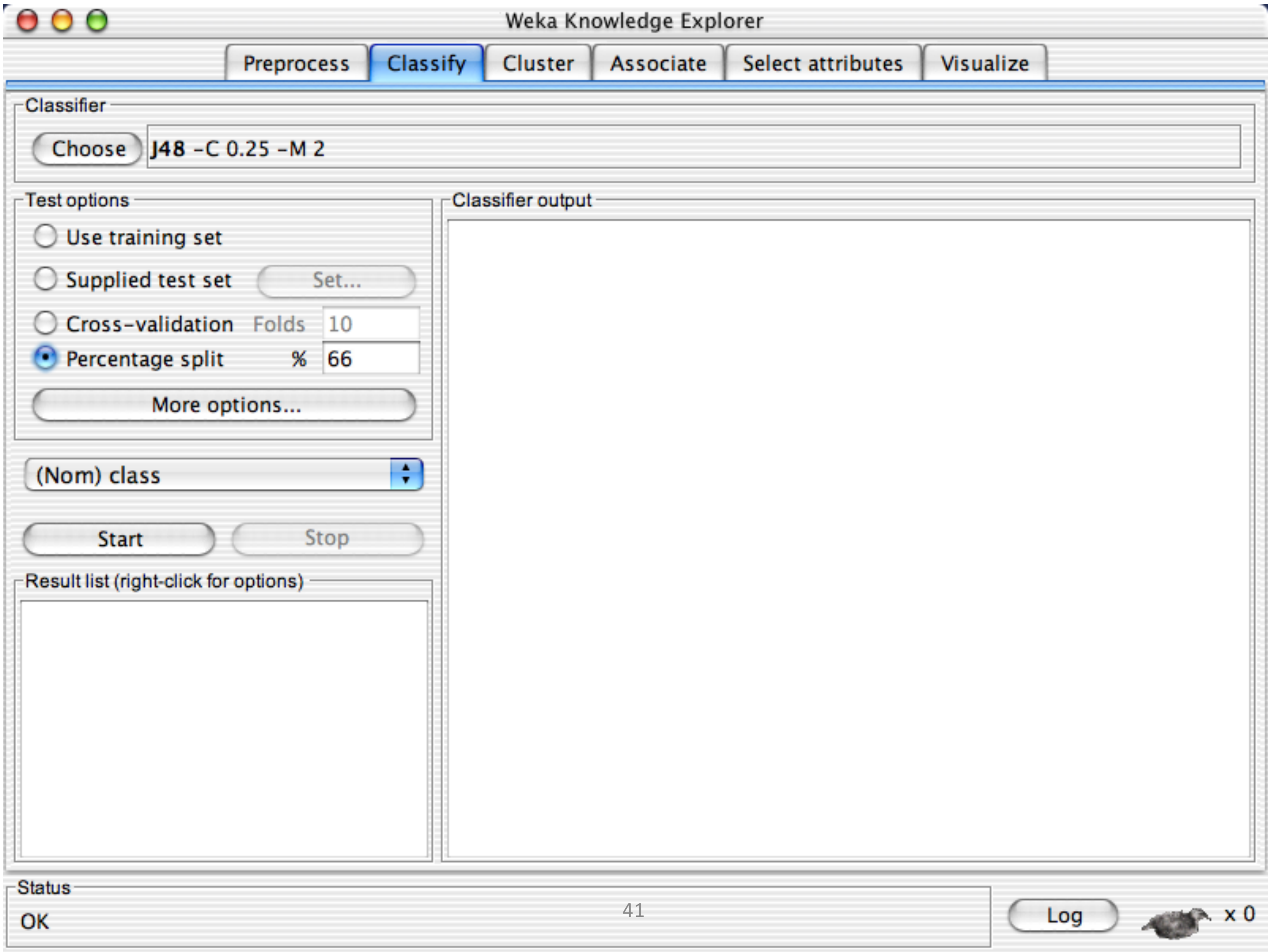
○ Supplied test set

● Cross-validation Folds 10

○ Percentage split % 66

(Nom) class

Status: OK

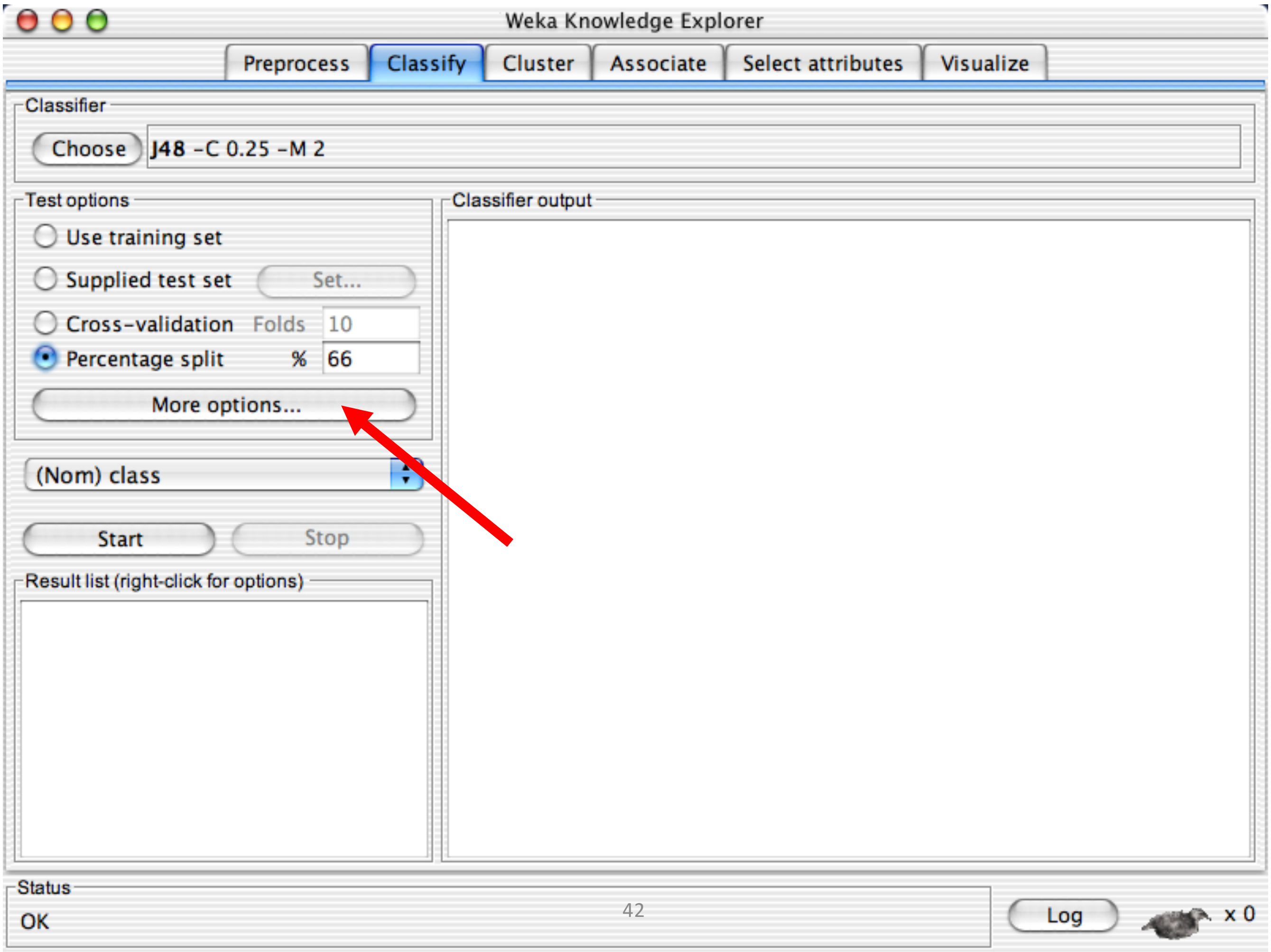


J48 -C 0.25 -M 2

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

(Nom) class

Status
OK

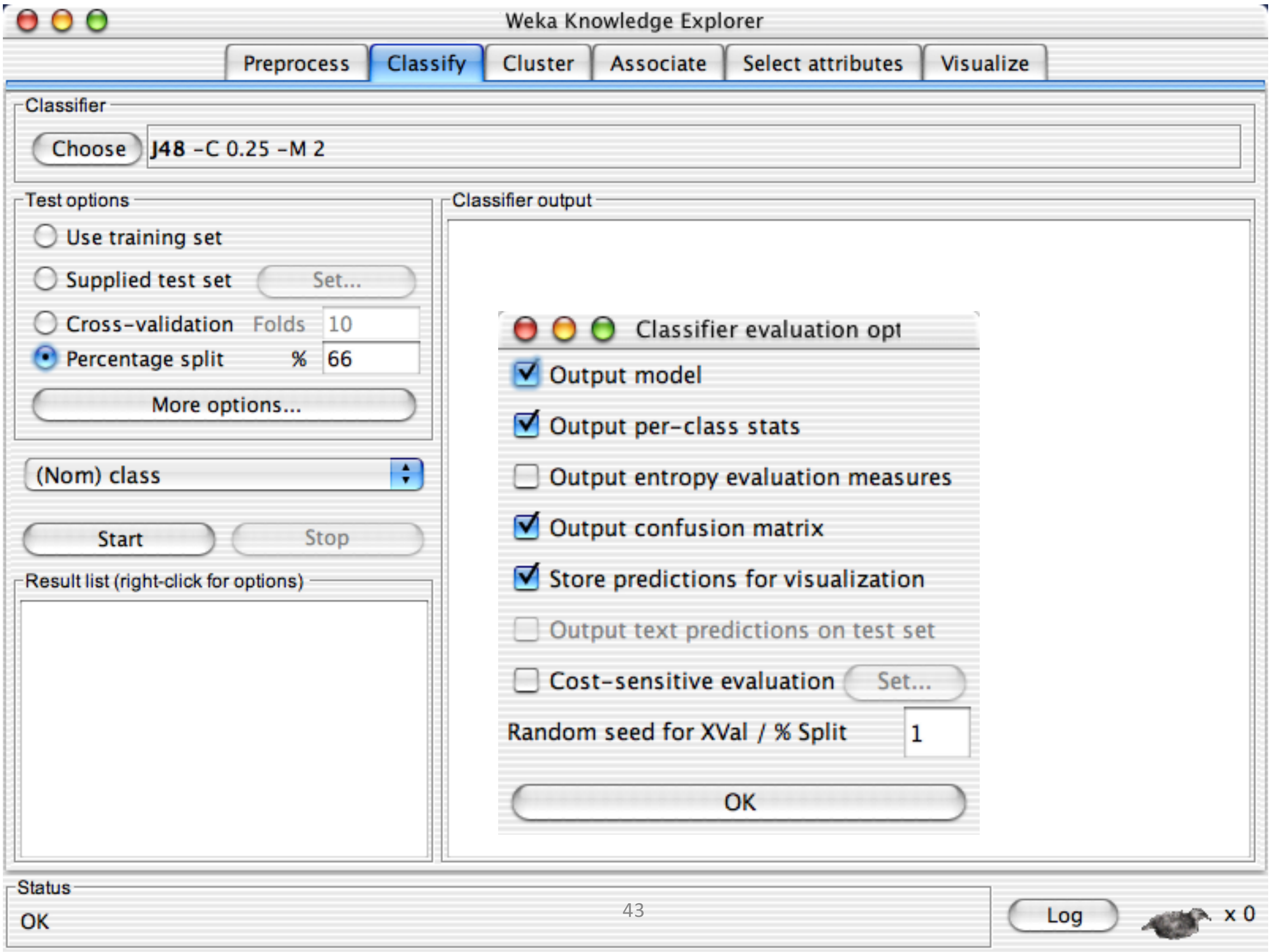


J48 -C 0.25 -M 2

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

(Nom) class

Status
OK



Weka Knowledge Explorer

Classifier

J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set Set...

○ Cross-validation Folds 10

● Percentage split % 66

(Nom) class

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

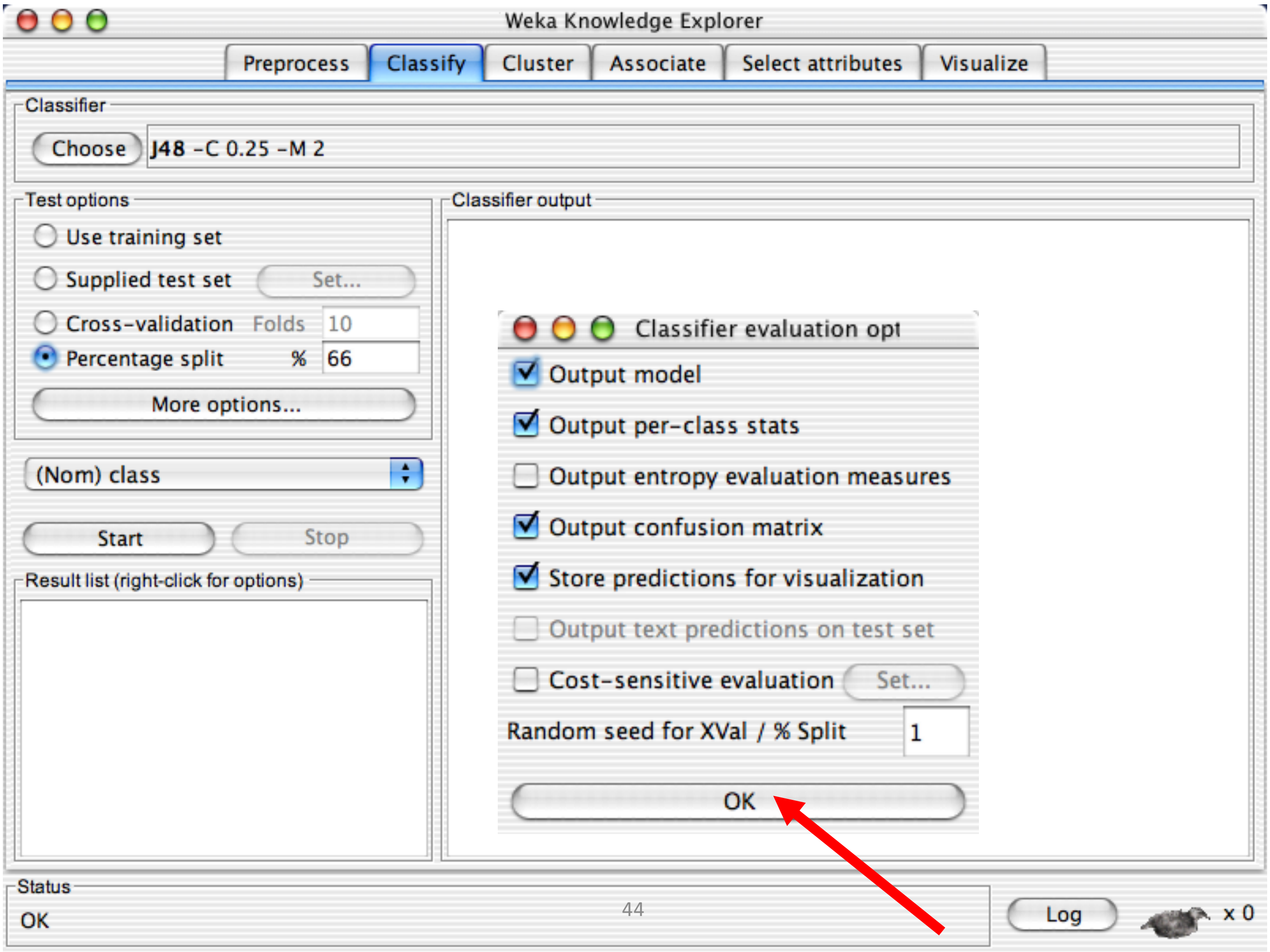
□ Output text predictions on test set

□ Cost-sensitive evaluation Set...

Random seed for XVal / % Split 1

Status

OK



Weka Knowledge Explorer

Classifier

J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set Set...

○ Cross-validation Folds 10

● Percentage split % 66

(Nom) class

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

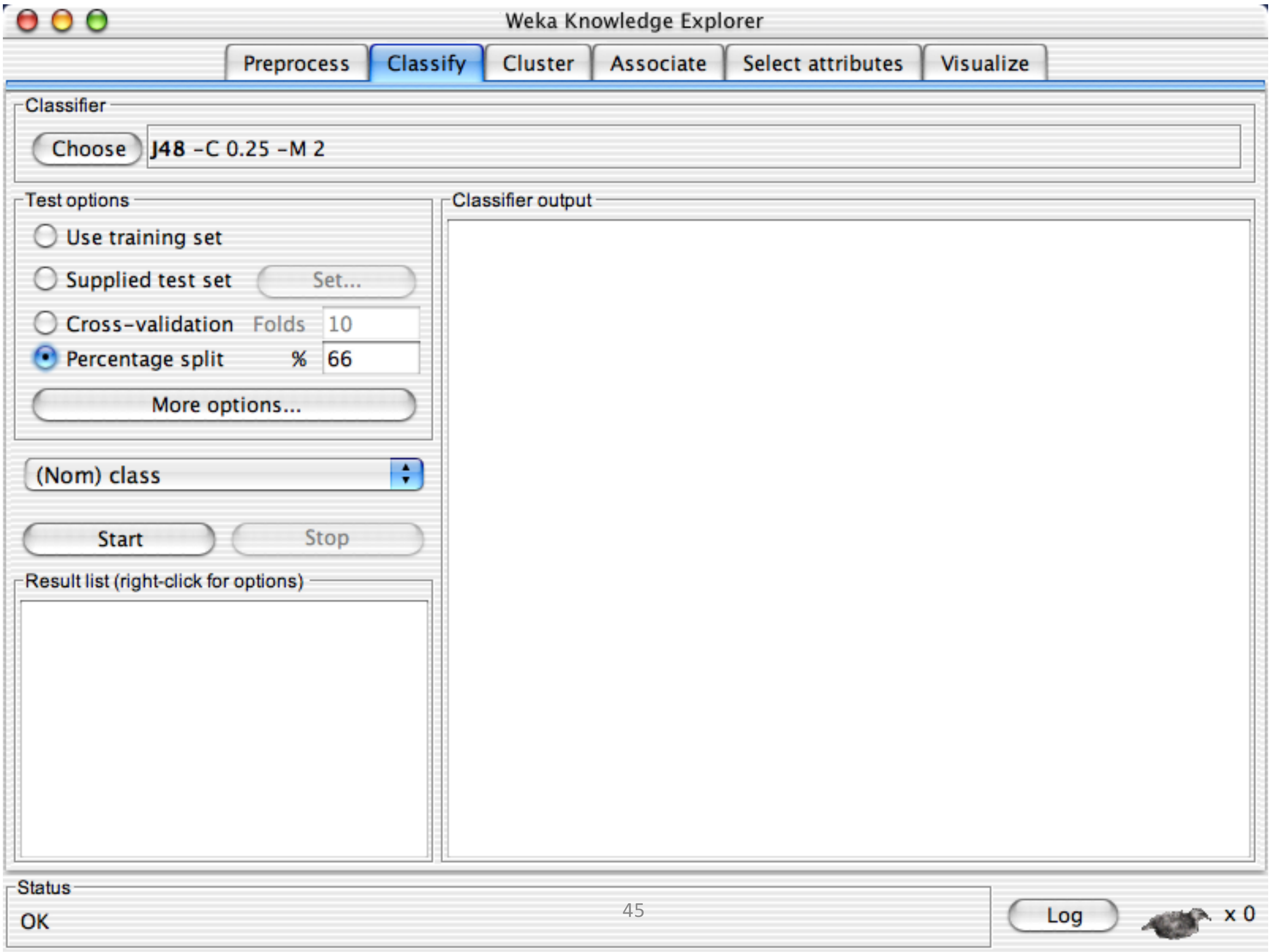
□ Output text predictions on test set

□ Cost-sensitive evaluation Set...

Random seed for XVal / % Split 1

Status

OK



Classifier

J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set

○ Cross-validation Folds 10

● Percentage split % 66

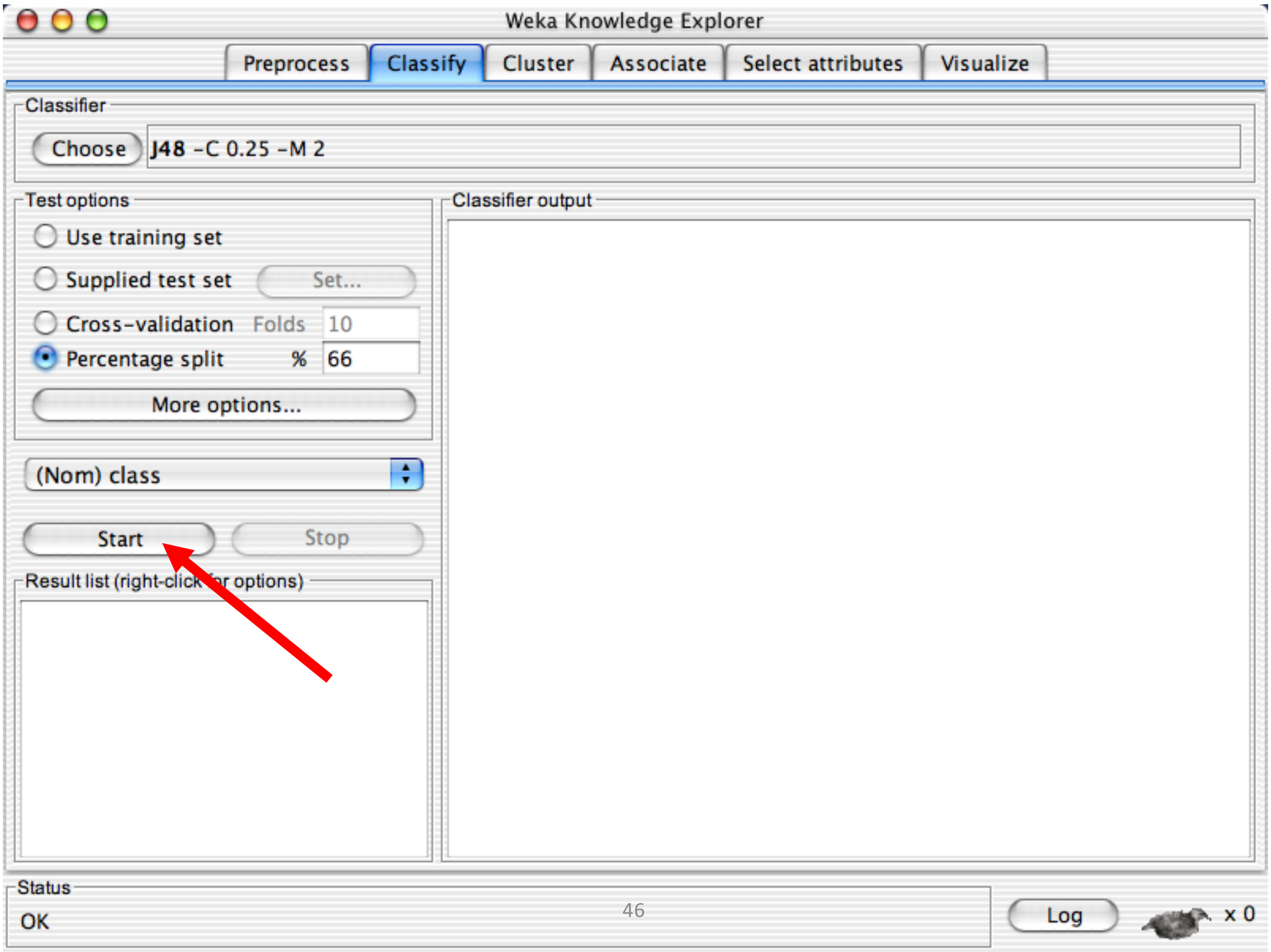
(Nom) class

Result list (right-click for options)

Classifier output

Status

OK



Classifier

J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set

○ Cross-validation Folds 10

● Percentage split % 66

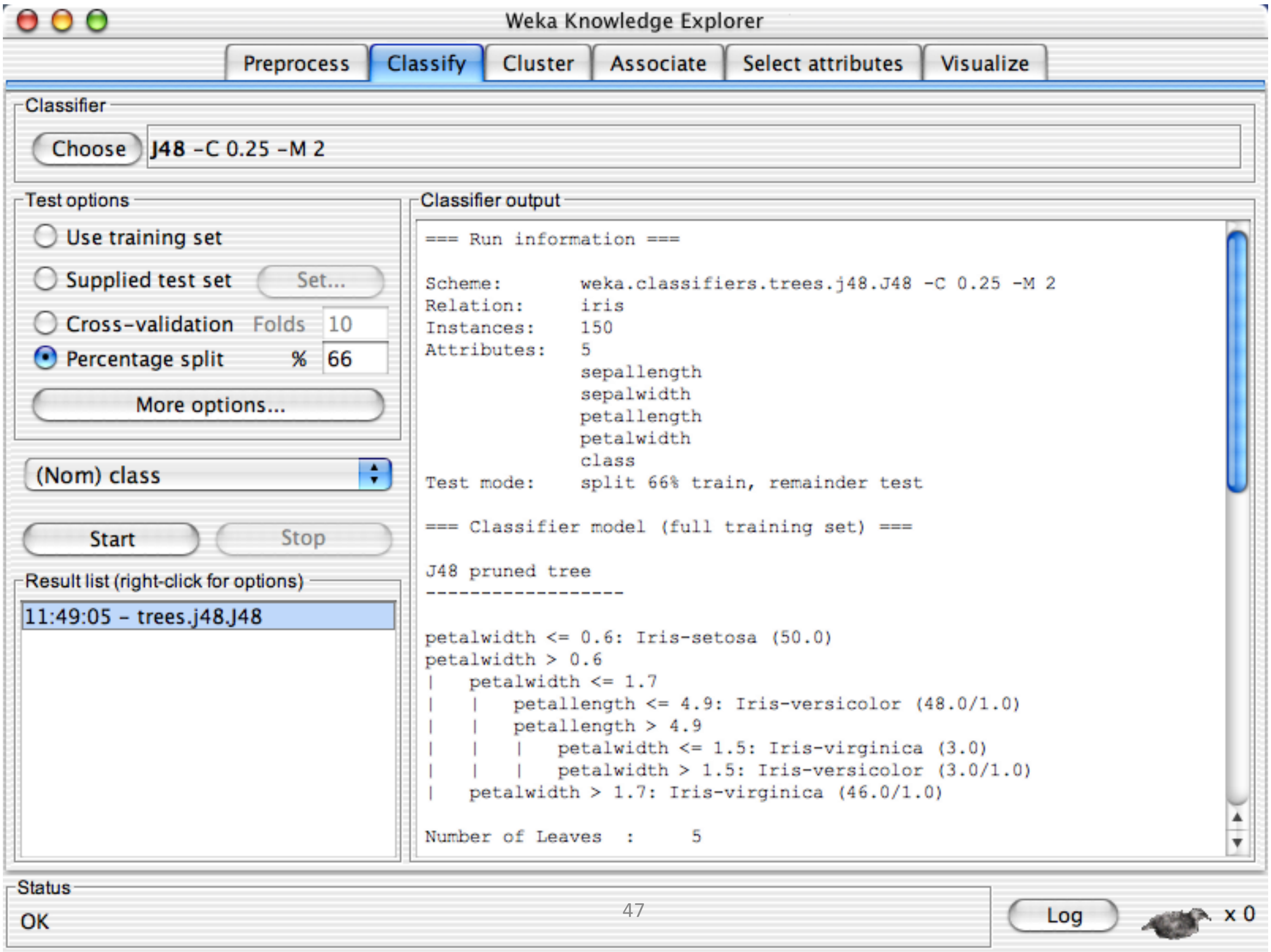
(Nom) class

Result list (right-click for options)

Classifier output

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)

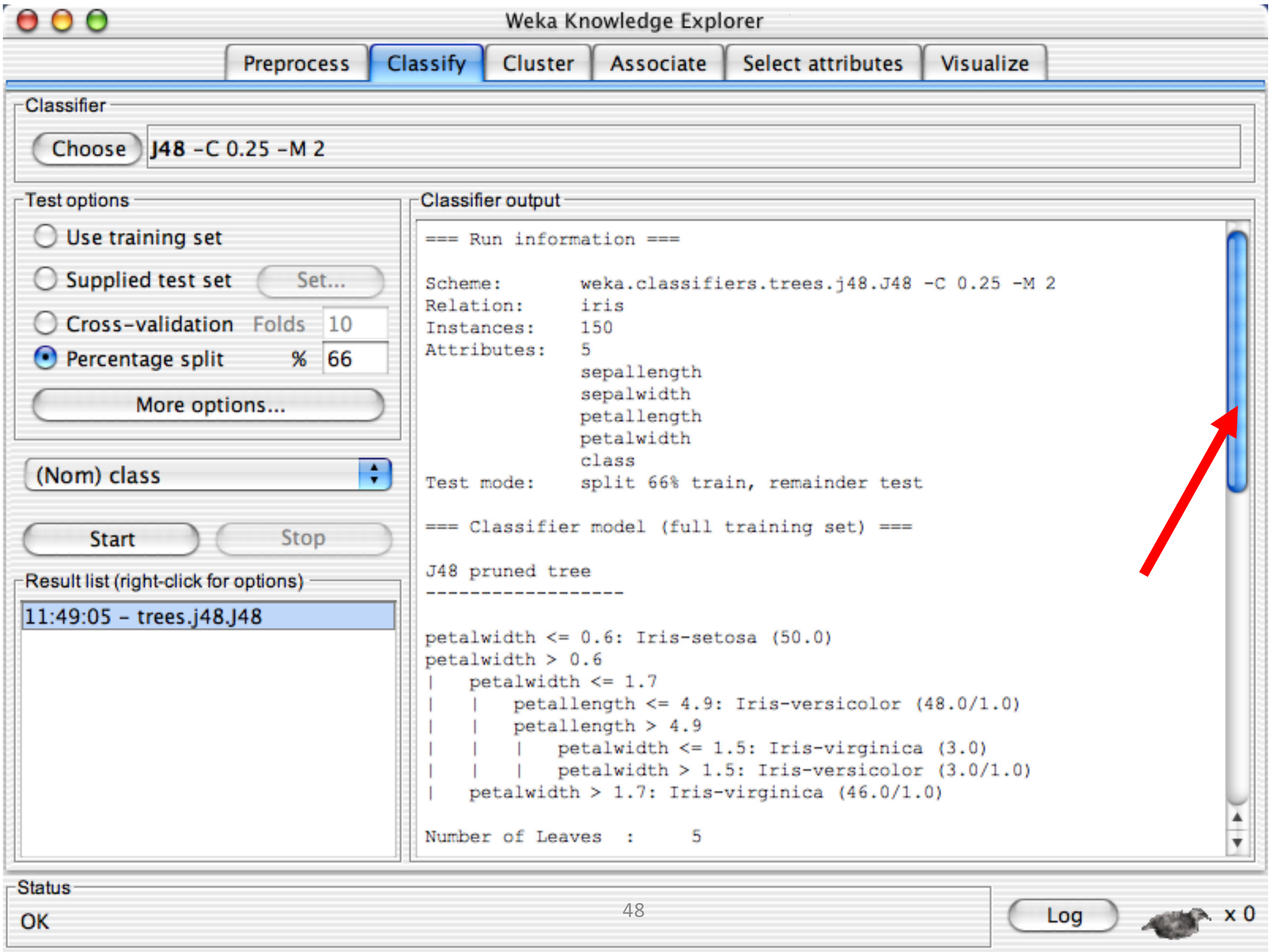
11:49:05 - trees.j48.J48

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



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)

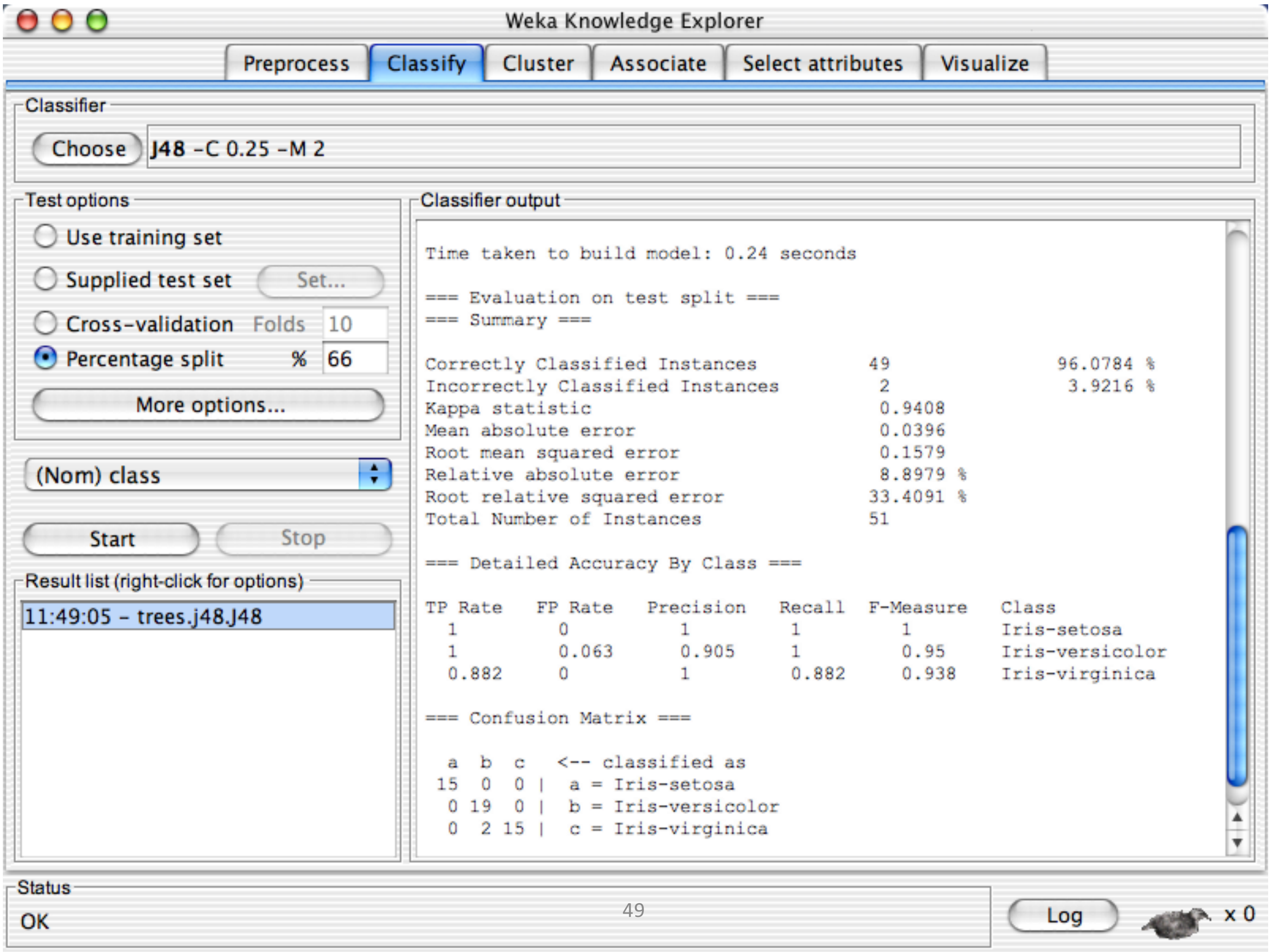
11:49:05 - trees.j48.J48

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

Choose 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)

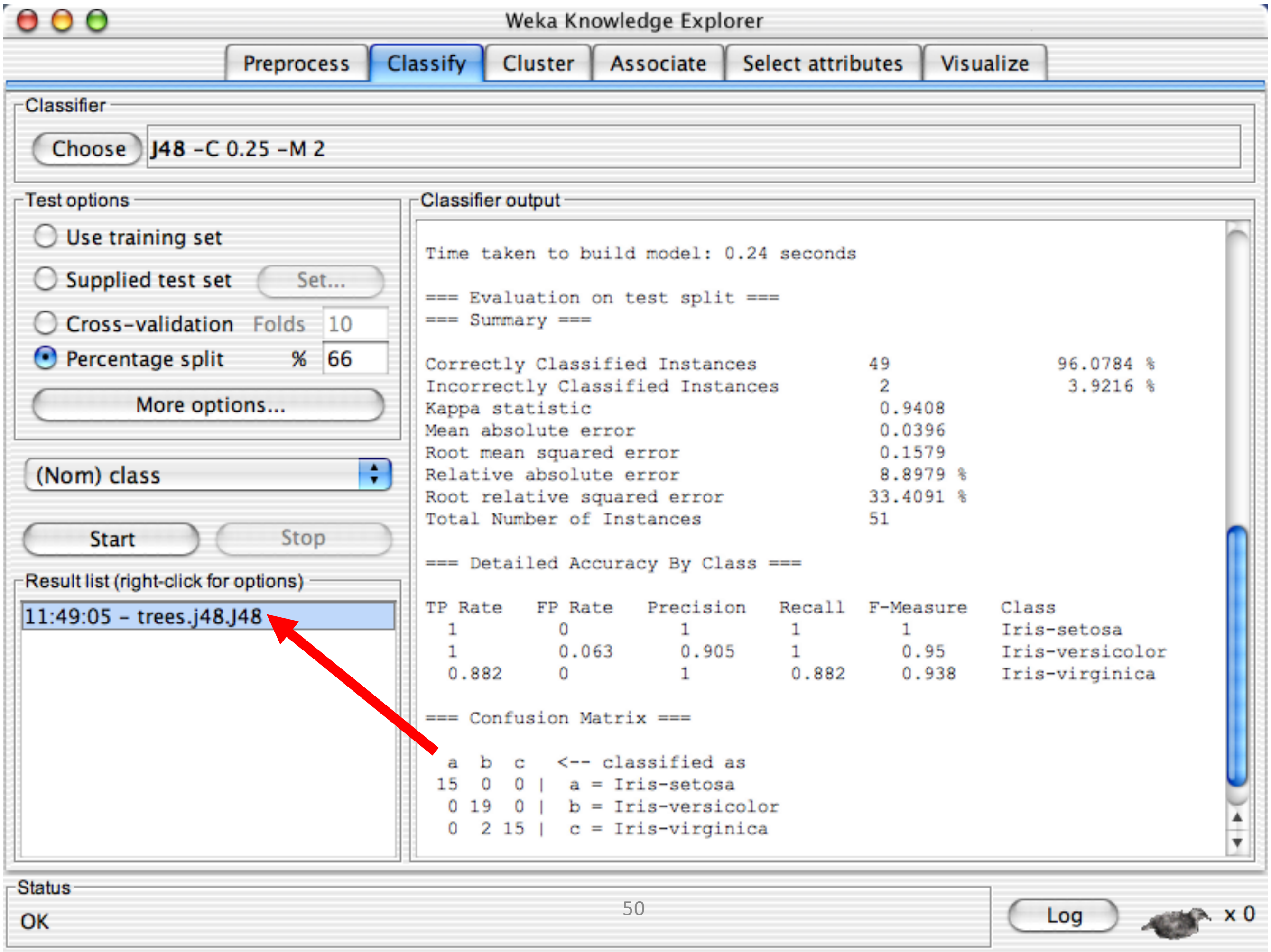
11:49:05 - trees.j48.J48

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



Classifier

Choose J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set

○ Cross-validation Folds 10

● Percentage split % 66

(Nom) class

Result list (right-click for options)

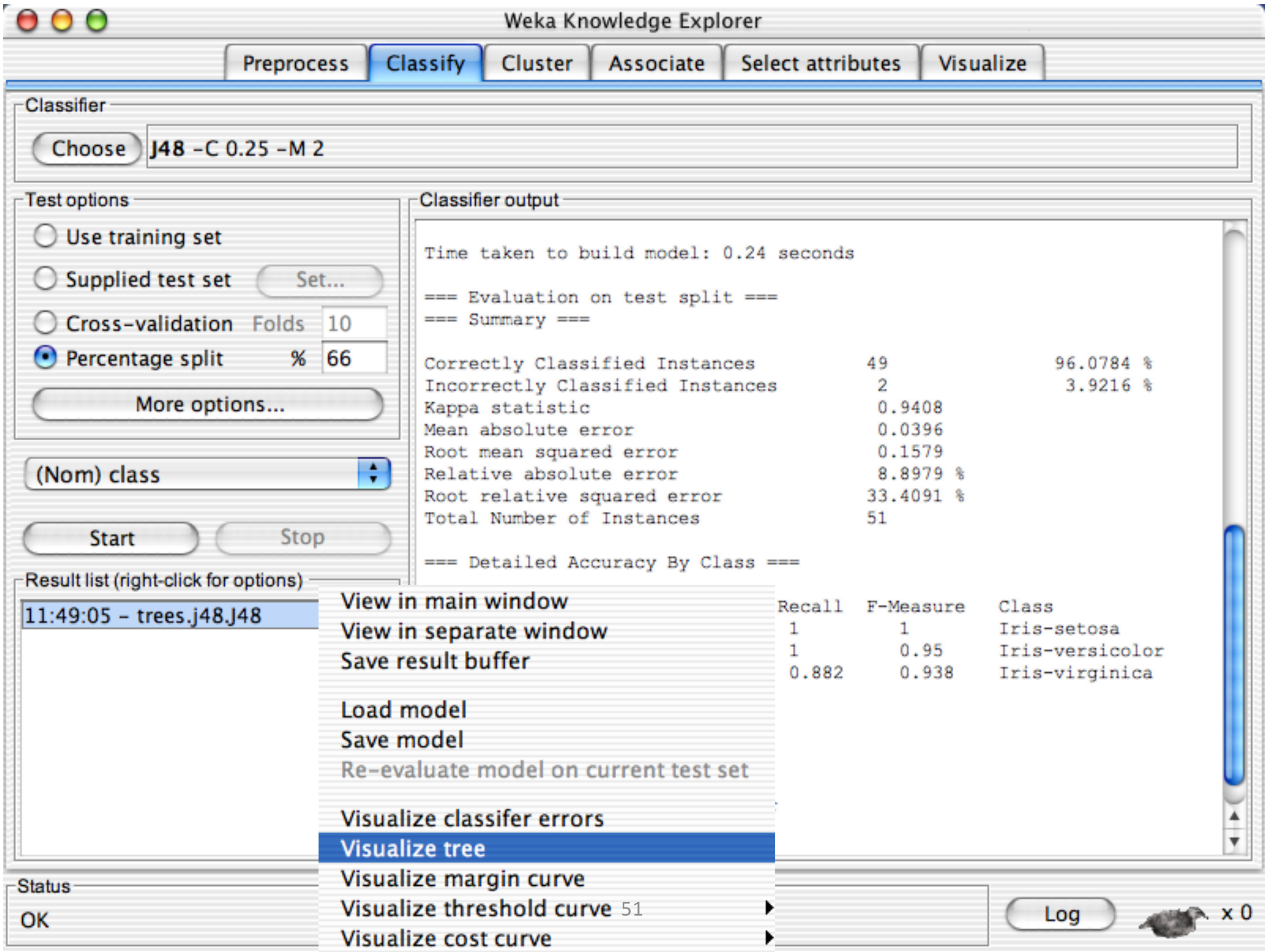
11:49:05 - trees.j48.J48

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



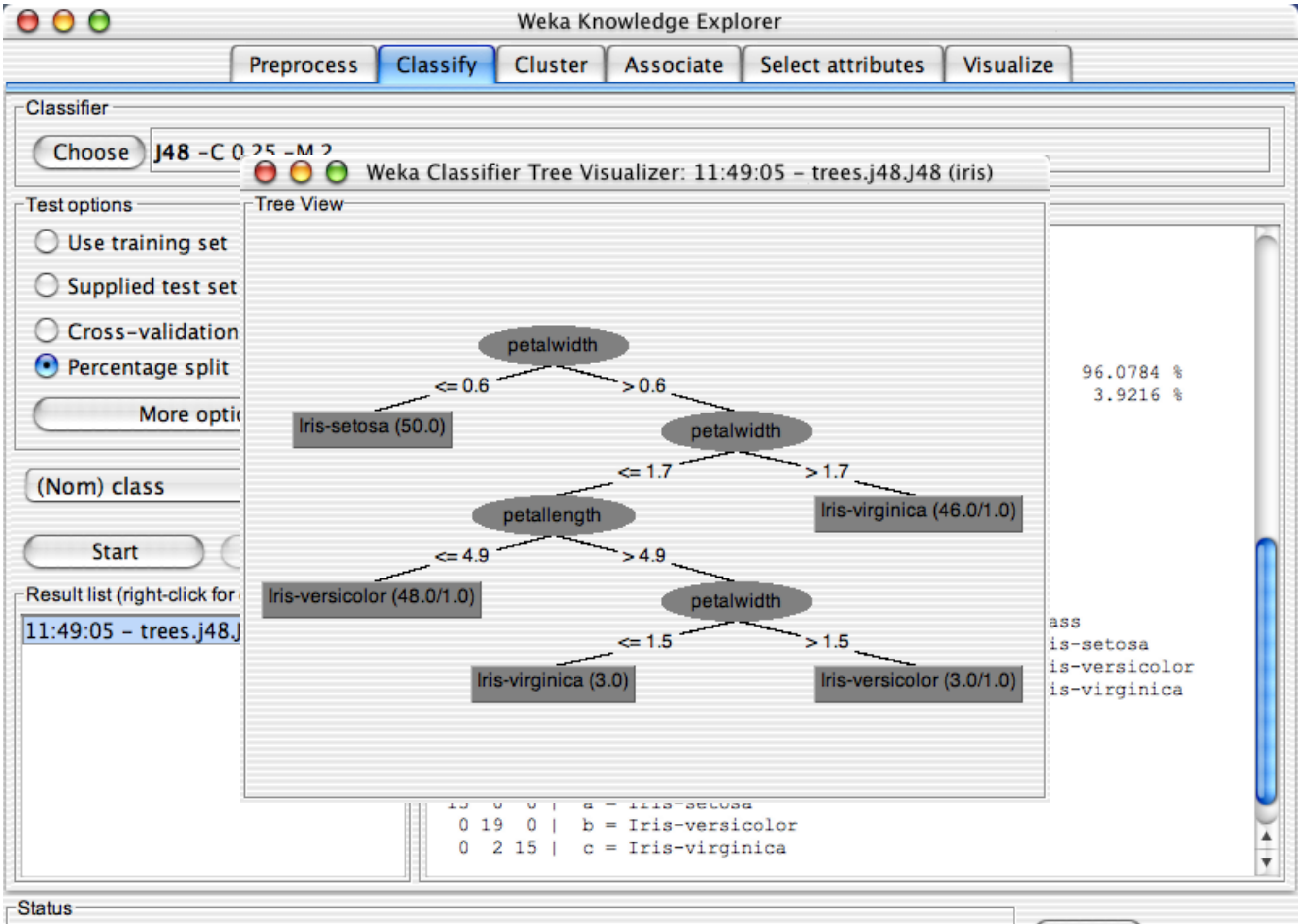
J48 -C 0.25 -M 2

(Nom) class

11:49:05 - trees.j48.J48

```
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
```

OK



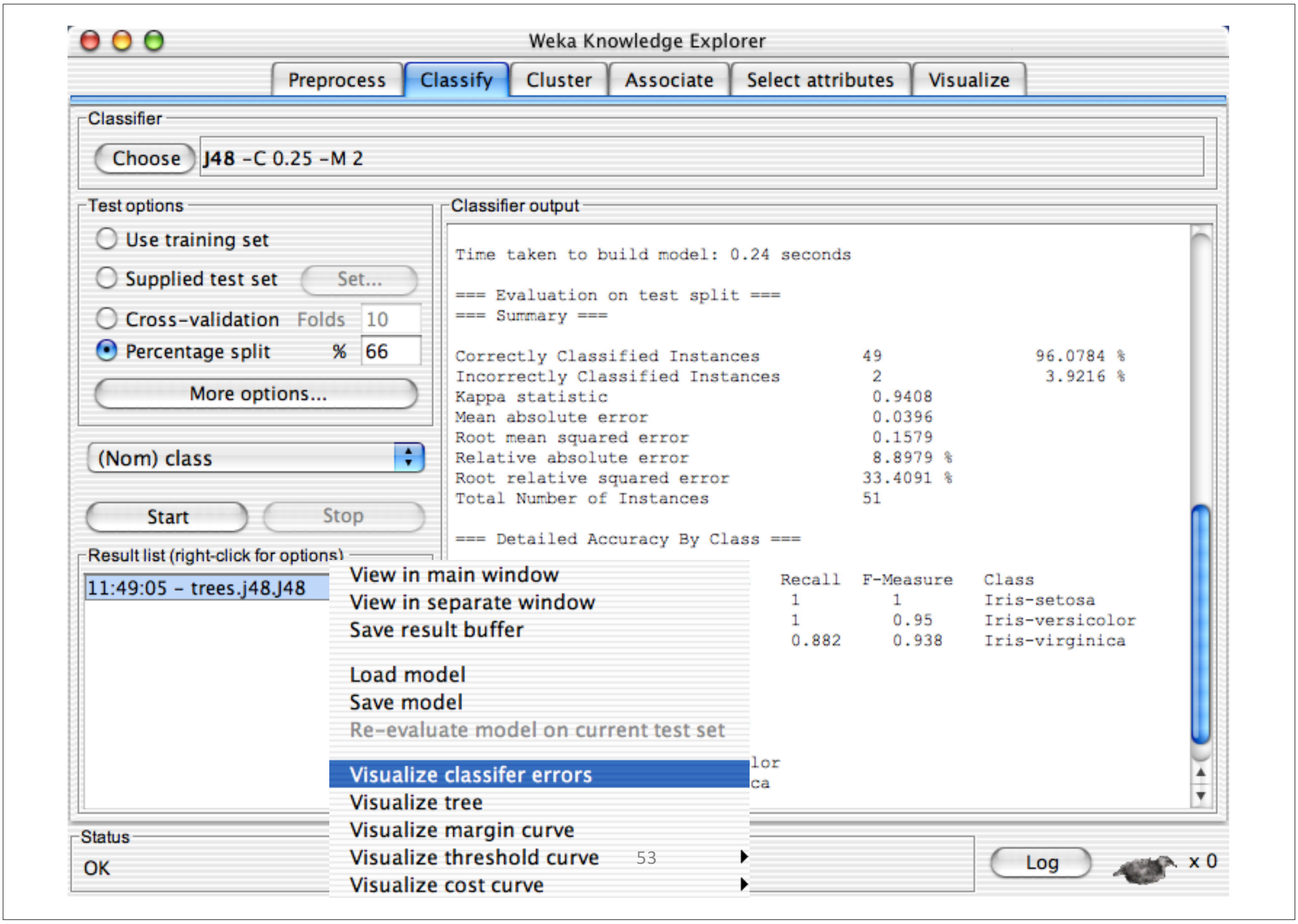
J48 -C 0.25 -M 2

(Nom) class

11:49:05 - trees.j48.J

```
96.0784 %
 3.9216 %

ass
is-setosa
is-versicolor
is-virginica

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

OK



J48 -C 0.25 -M 2

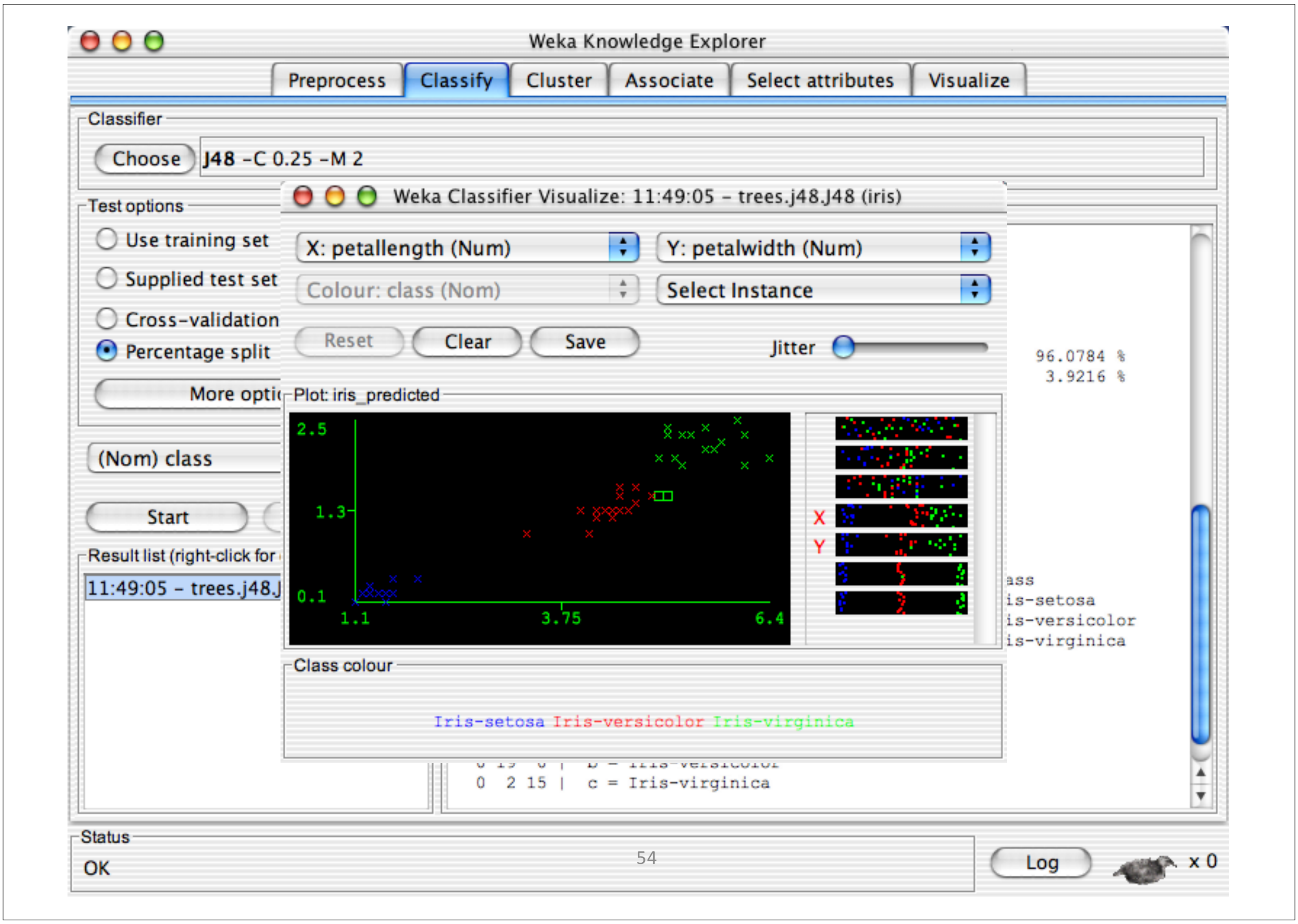
○ Use training set
○ Supplied test set Set...
○ Cross-validation Folds 10
● Percentage split % 66
More options...

(Nom) class

11:49:05 - trees.j48.J48

```
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
```

OK



J48 -C 0.25 -M 2

○ Use training set
○ Supplied test set
○ Cross-validation
● Percentage split
More opti

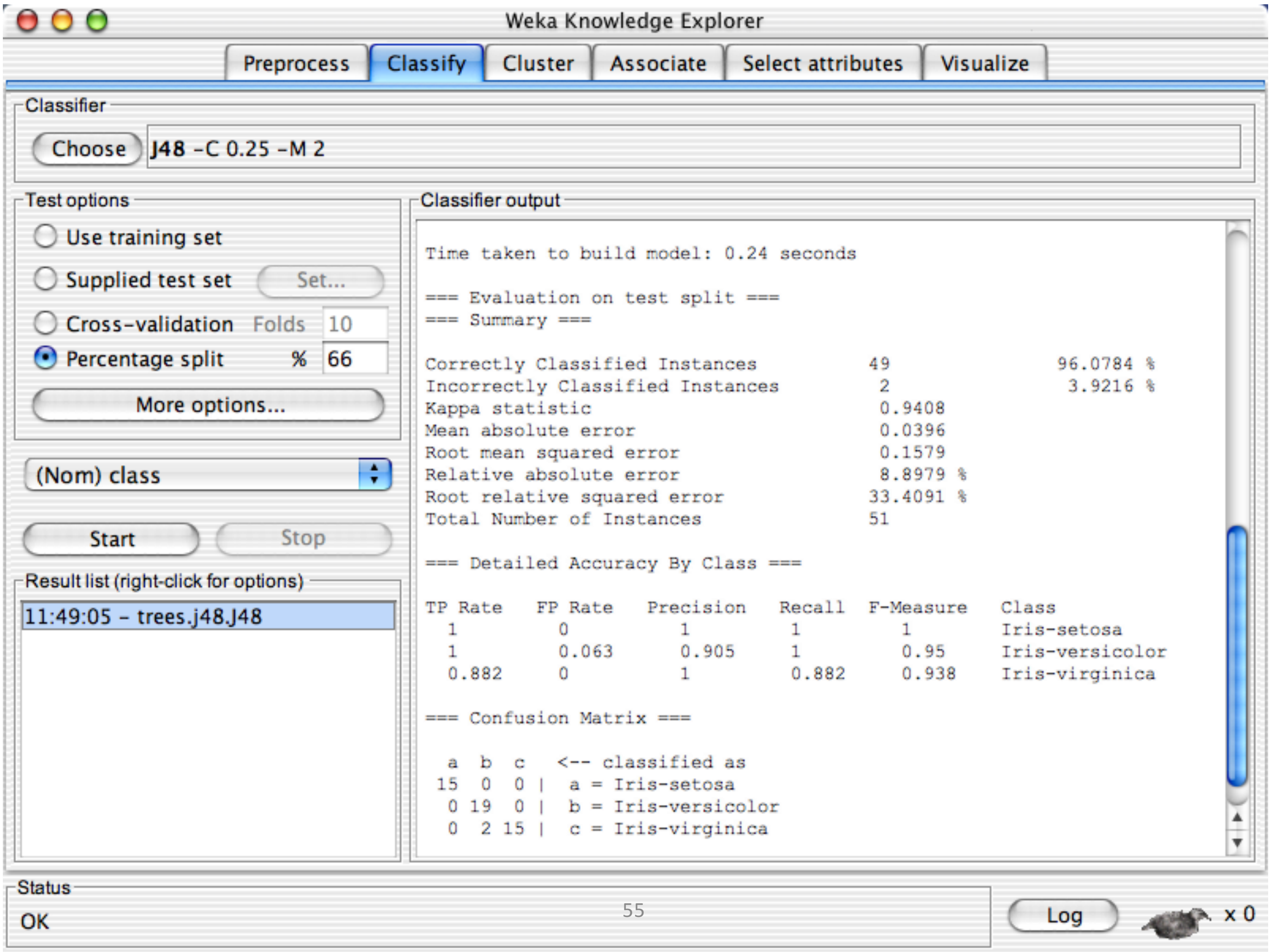
(Nom) class

11:49:05 - trees.j48.J

X: petallength (Num)
Y: petalwidth (Num)
Colour: class (Nom)
Select Instance

```
96.0784 %
 3.9216 %

ass
is-setosa
is-versicolor
is-virginica

 0  2 15 |  c = Iris-virginica
```

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)

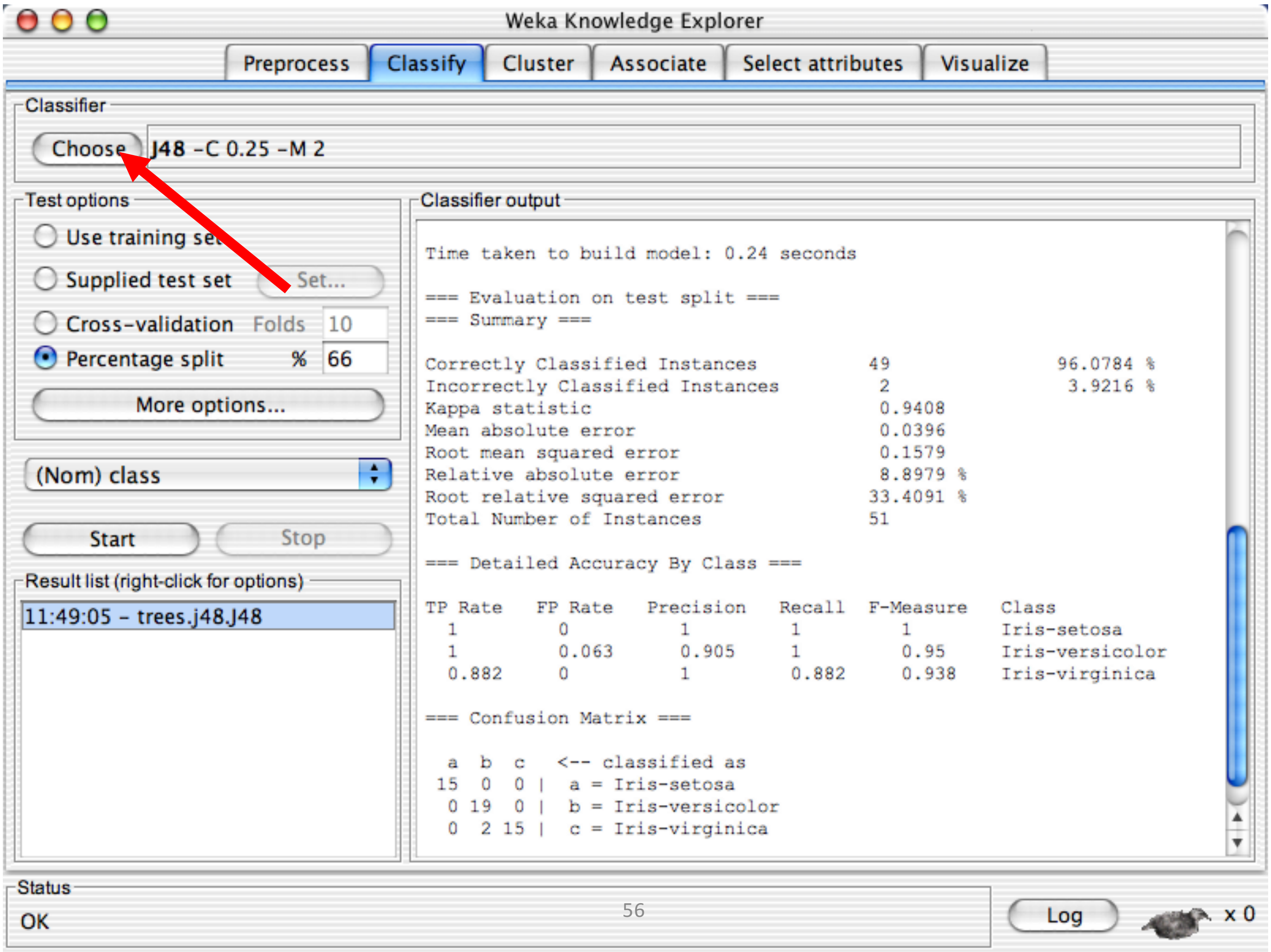
11:49:05 - trees.j48.J48

Classifier output

```
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



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)

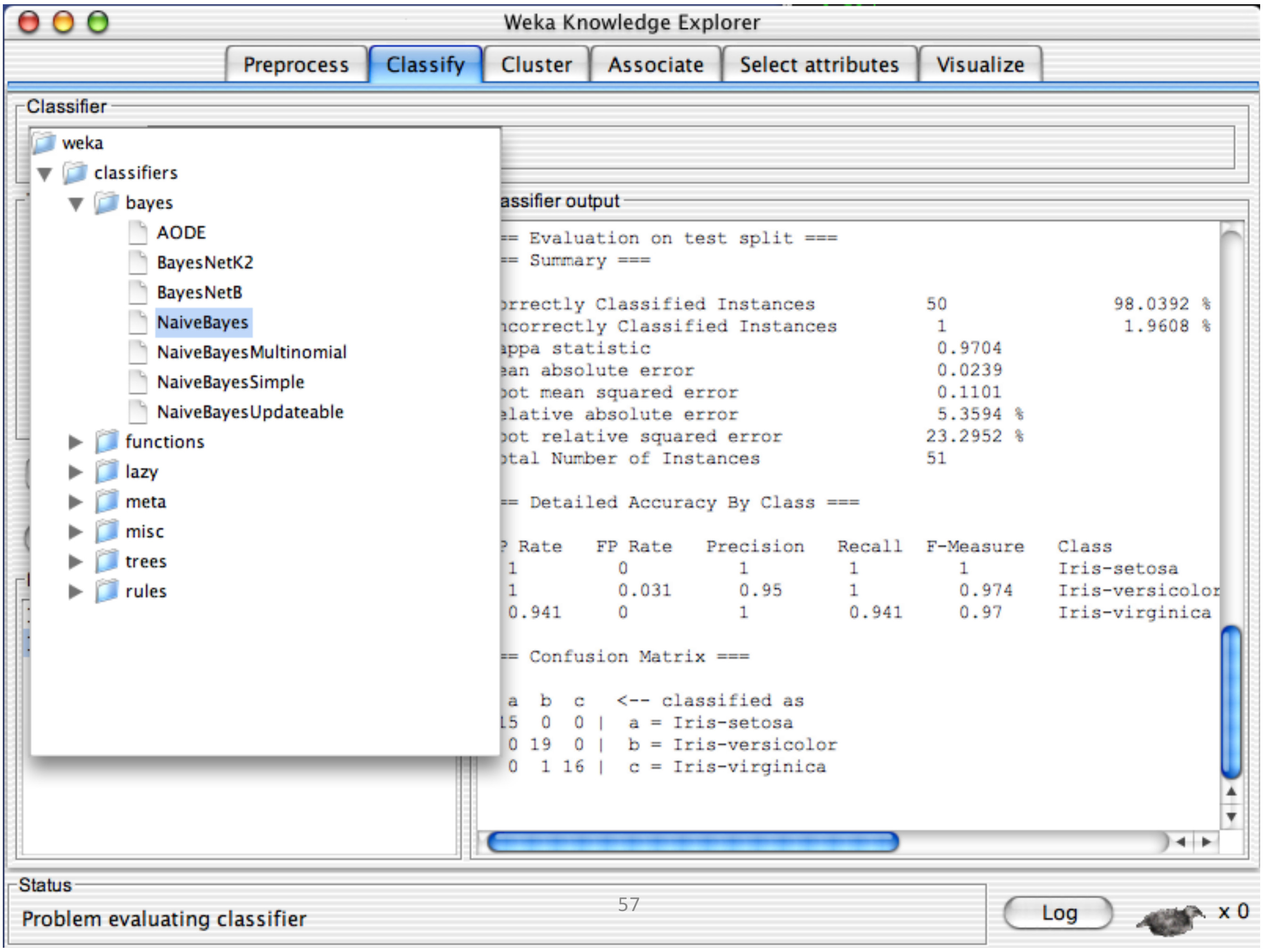
11:49:05 - trees.j48.J48

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
```

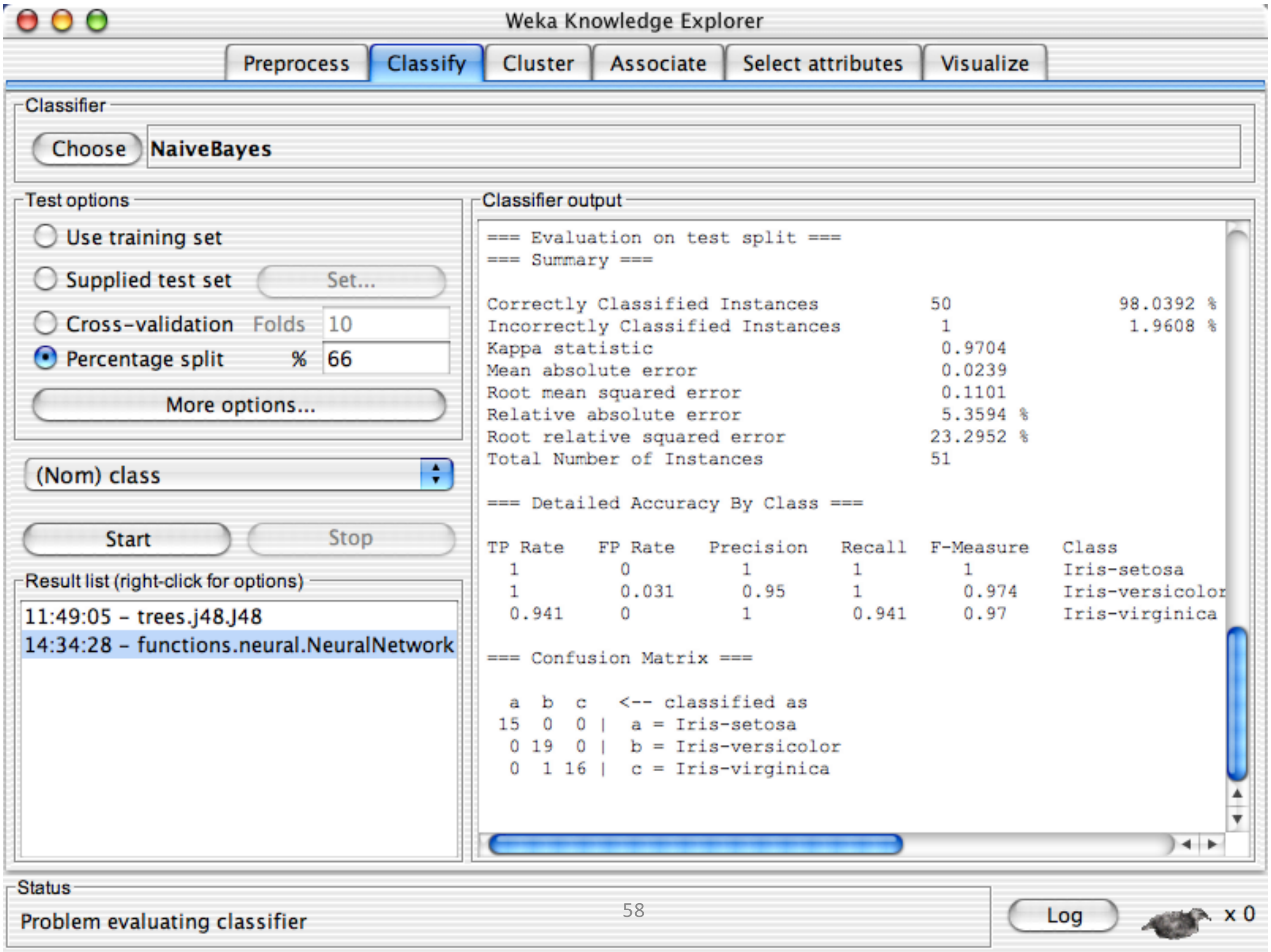
Status

OK



```
== 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
```

Problem evaluating classifier



NaiveBayes

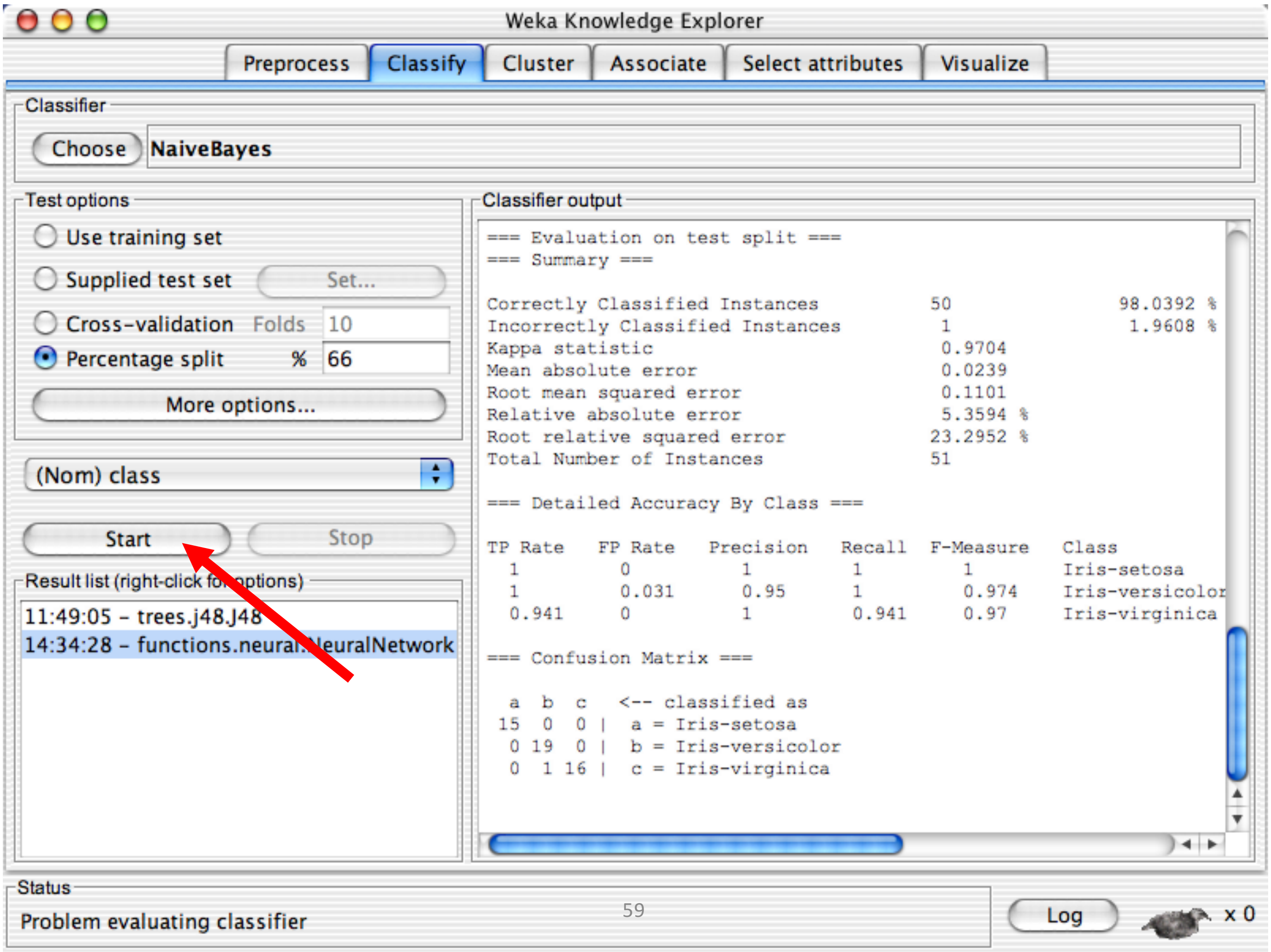
○ Use training set
○ Supplied test set
○ Cross-validation Folds 10
● Percentage split % 66

(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
```

Problem evaluating classifier



Classifier

Choose NaiveBayes

Test options

○ Use training set

○ Supplied test set   Set...

○ Cross-validation   Folds 10

● Percentage split   % 66

More options...

(Nom) class

Result list (right-click for options)

11:49:05 - trees.j48.J48

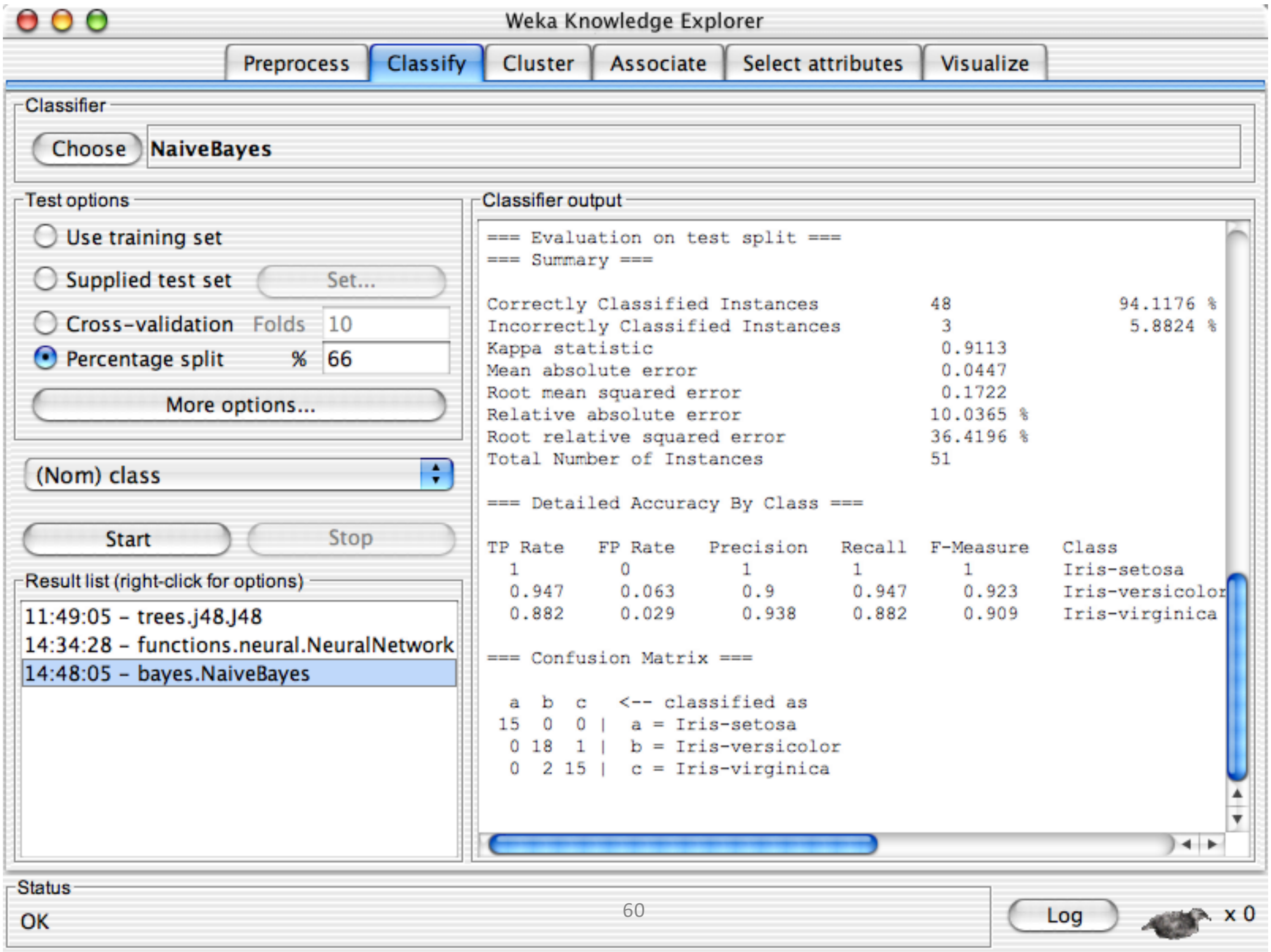
14:34:28 - functions.neural.NeuralNetwork

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

Problem evaluating classifier



Classifier

Choose NaiveBayes

Test options

○ Use training set

○ Supplied test set   Set...

○ Cross-validation   Folds 10

● Percentage split   % 66

More options...

(Nom) class

Result list (right-click for options)

11:49:05 - trees.j48.J48

14:34:28 - functions.neural.NeuralNetwork

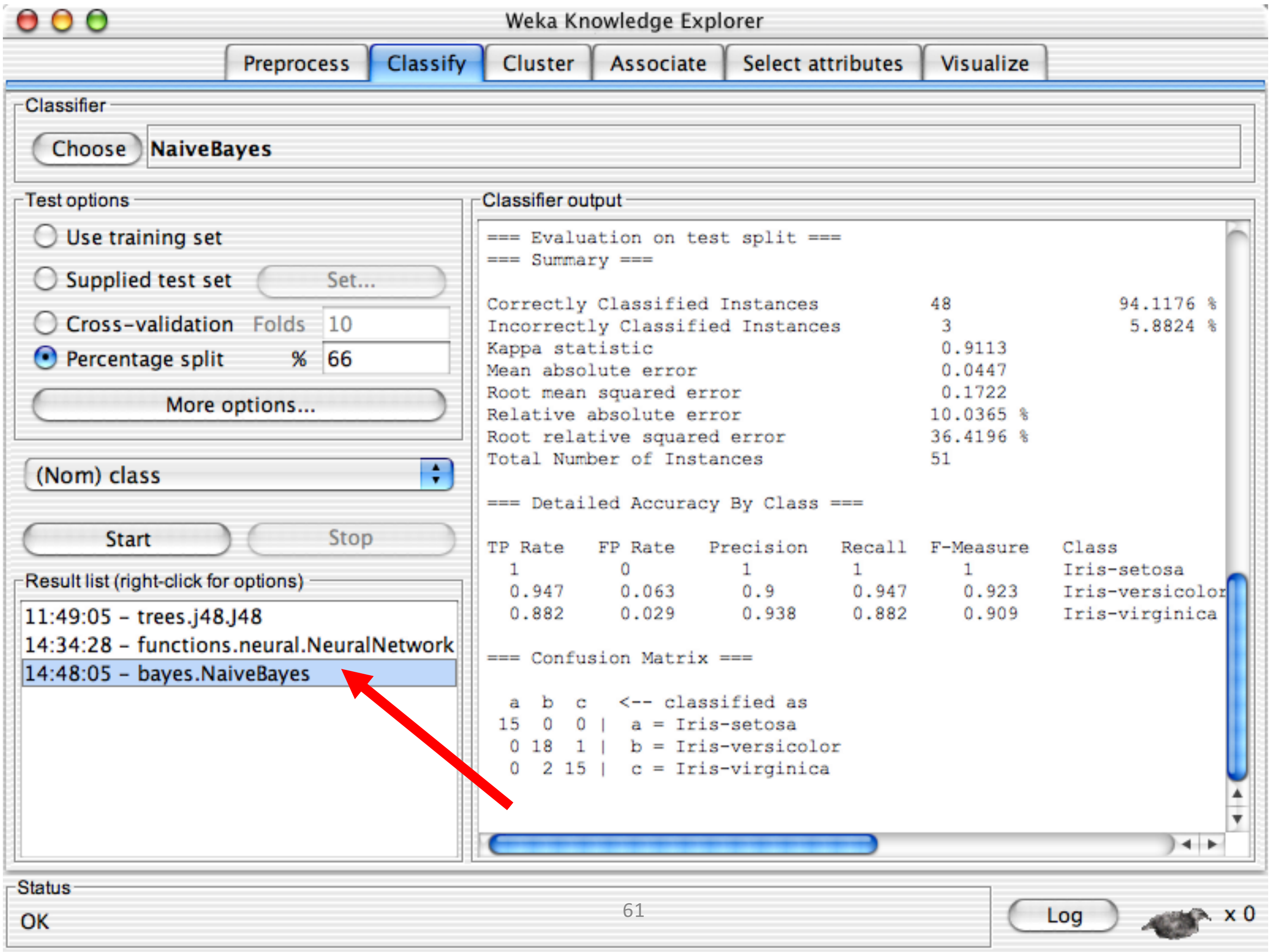
14:48:05 - bayes.NaiveBayes

Classifier output

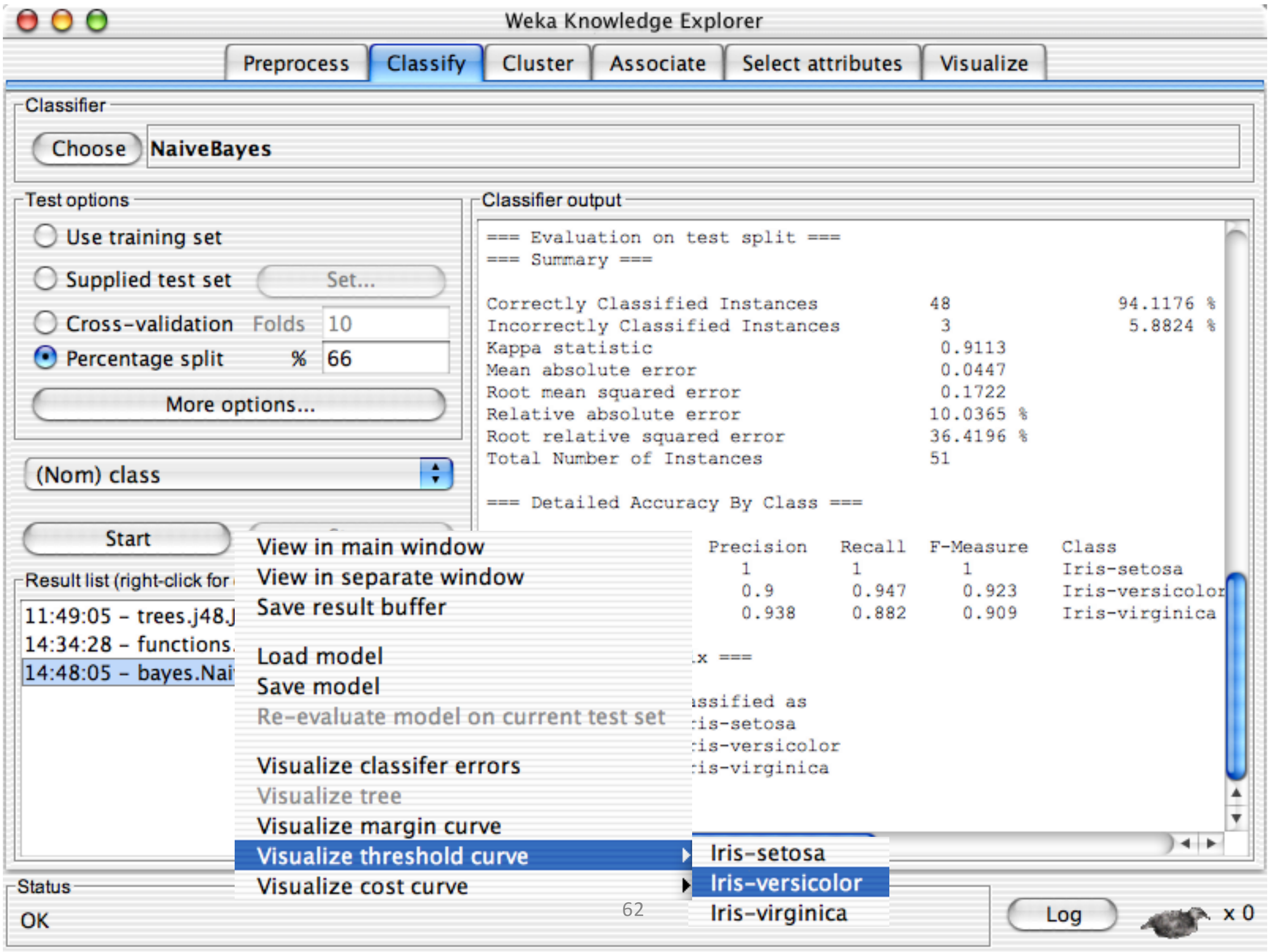
```
=== Evaluation on test split ===
=== Summary ===

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

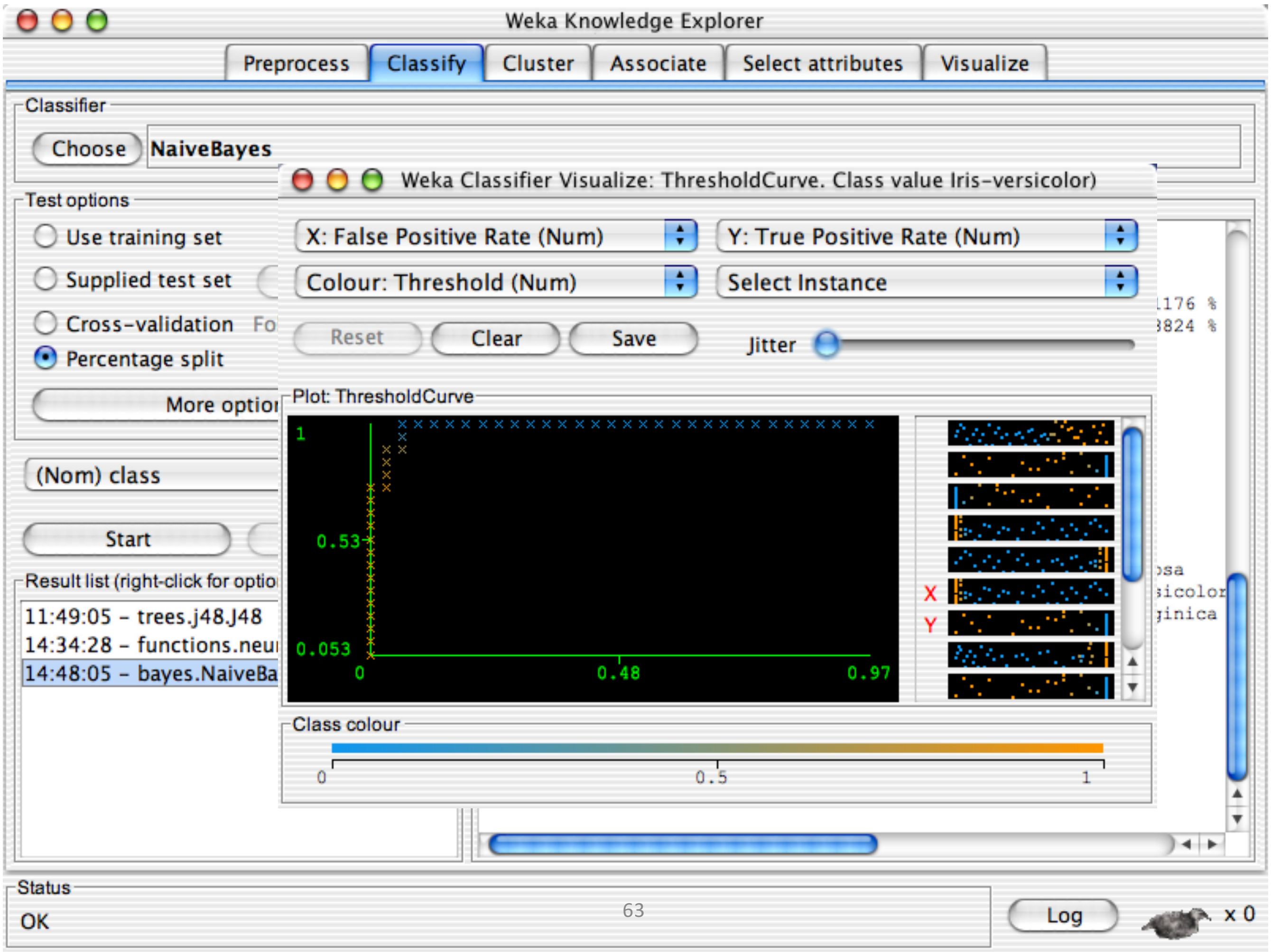
OK



```
=== 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
```



```
=== 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
```



Weka Knowledge Explorer

Classifier: NaiveBayes

Weka Classifier Visualize: ThresholdCurve. Class value Iris-versicolor

X: False Positive Rate (Num)

Y: True Positive Rate (Num)

Colour: Threshold (Num)

Test options: ● Percentage split

(Nom) class

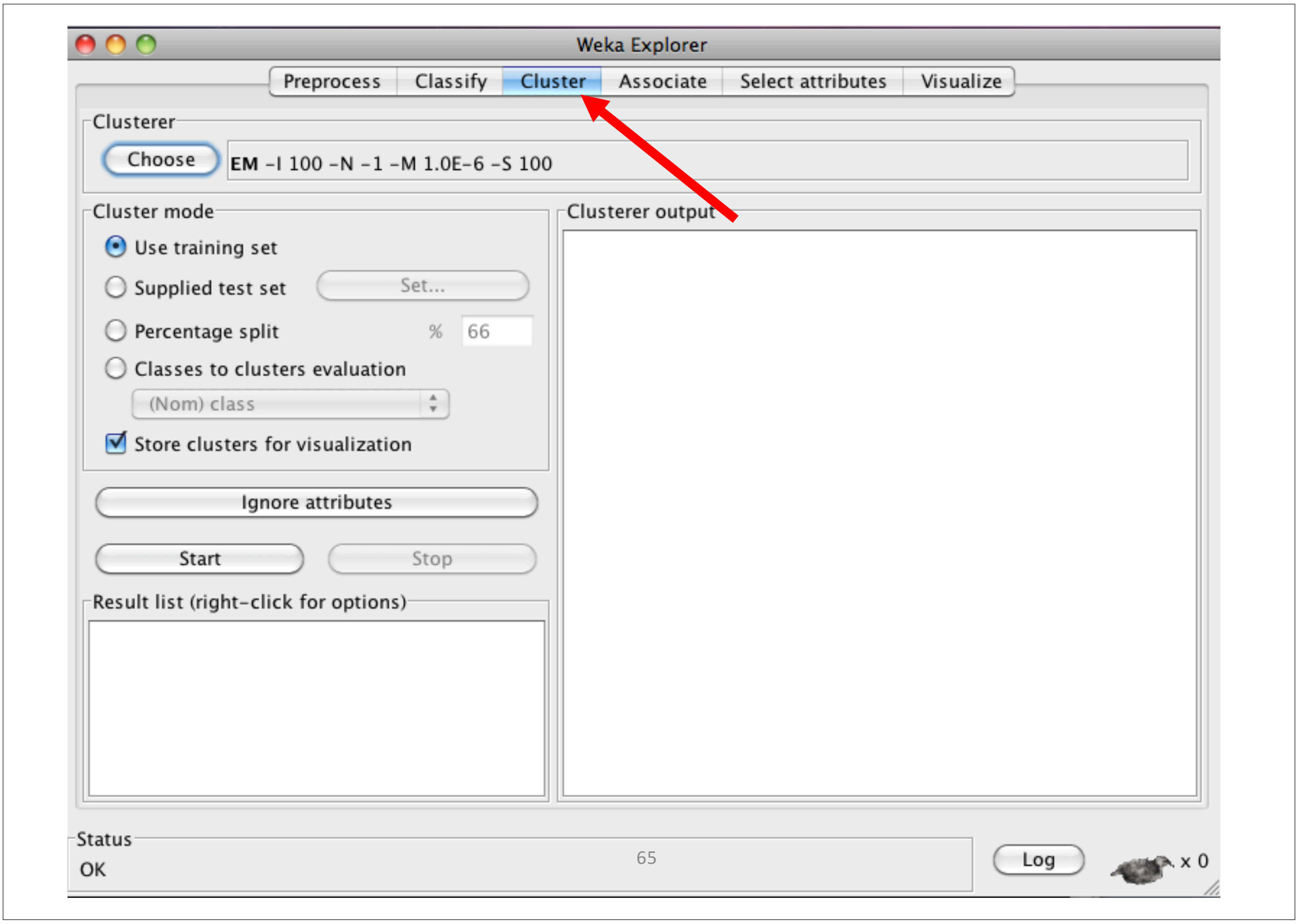
Result list (right-click for optio
11:49:05 - trees.j48.J48
14:34:28 - functions.neu
14:48:05 - bayes.NaiveBa

Status
OK

Clustering

- WEKA contains many clustering implementations:
  - Works with both discrete and numerical data

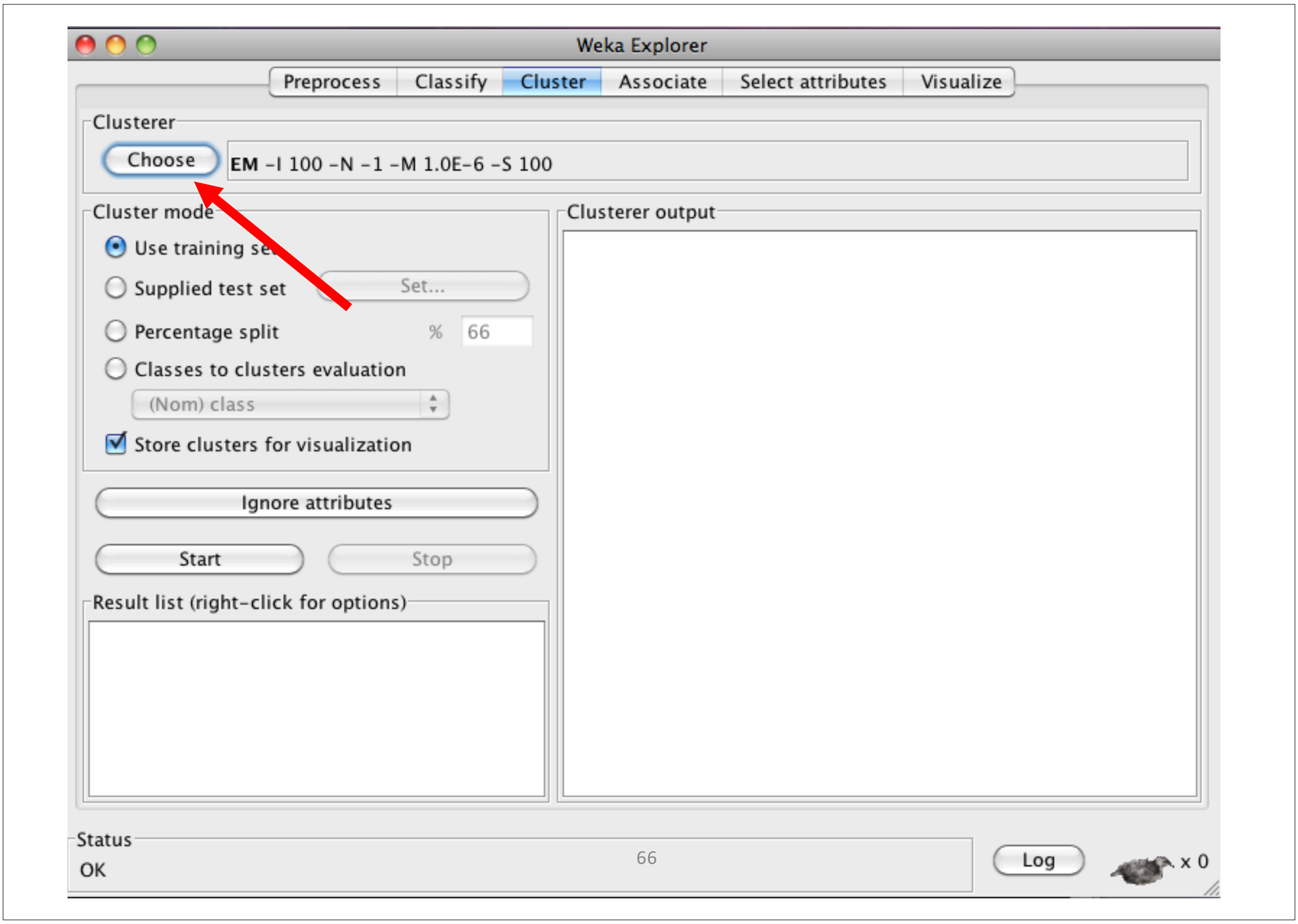
- Example of K-means



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

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

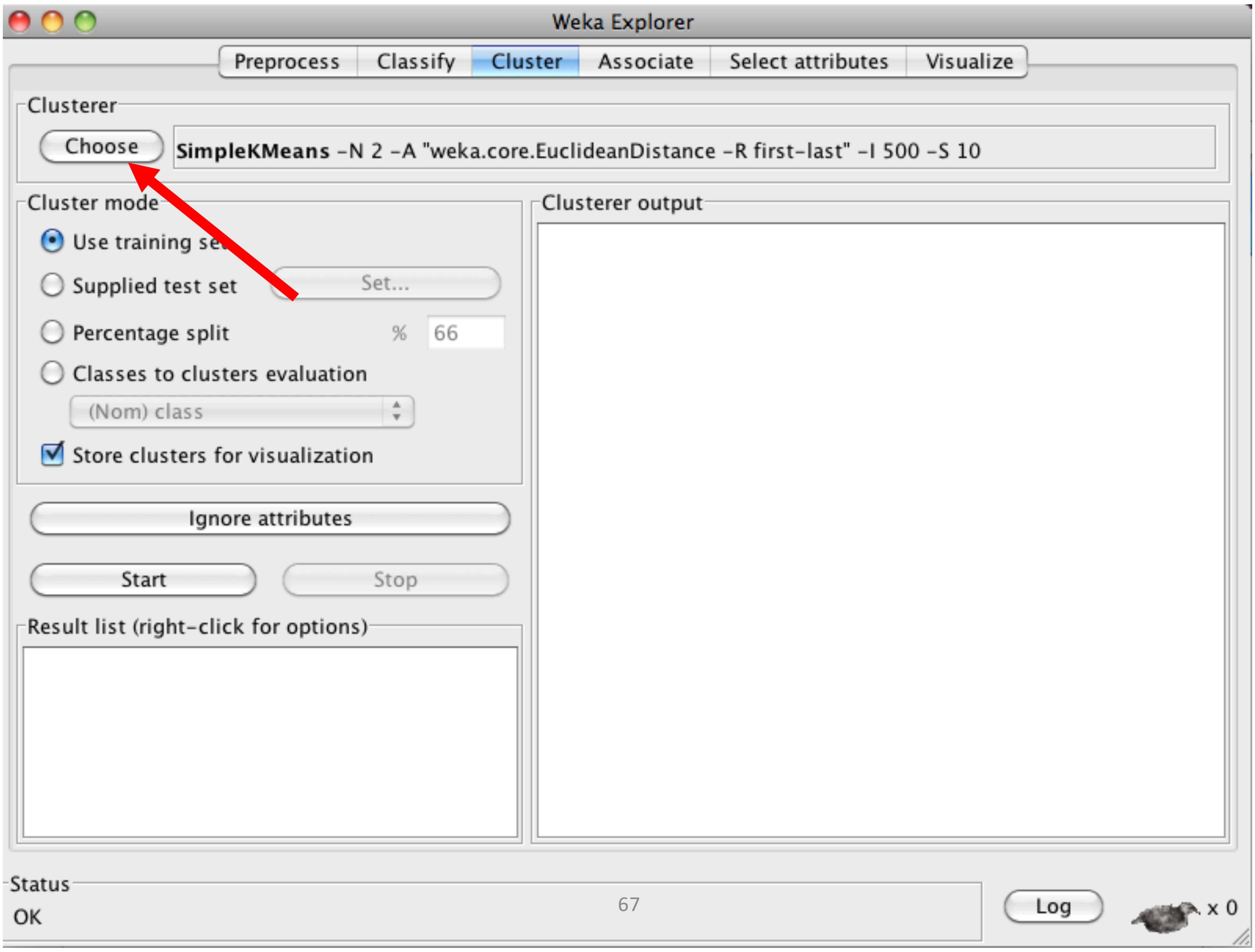
Status: OK



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

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

Status: OK



SimpleKMeans -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10

● Use training set

○ Supplied test set

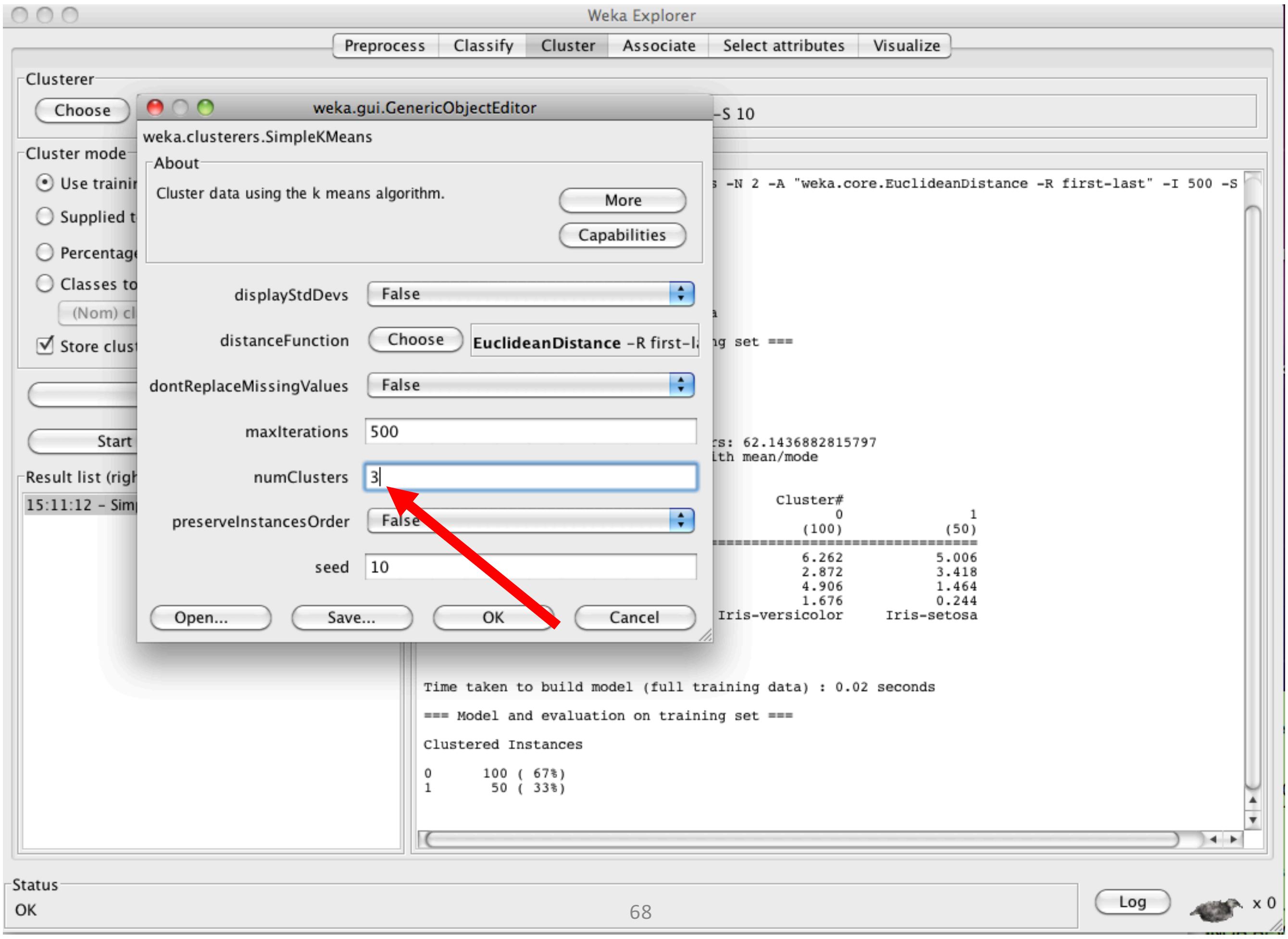
○ Percentage split % 66

○ Classes to clusters evaluation

(Nom) class

☑ Store clusters for visualization

OK



weka.clusterers.SimpleKMeans

Cluster data using the k means algorithm.

displayStdDevs False

distanceFunction EuclideanDistance -R first-l

dontReplaceMissingValues False

maxIterations 500

numClusters 3

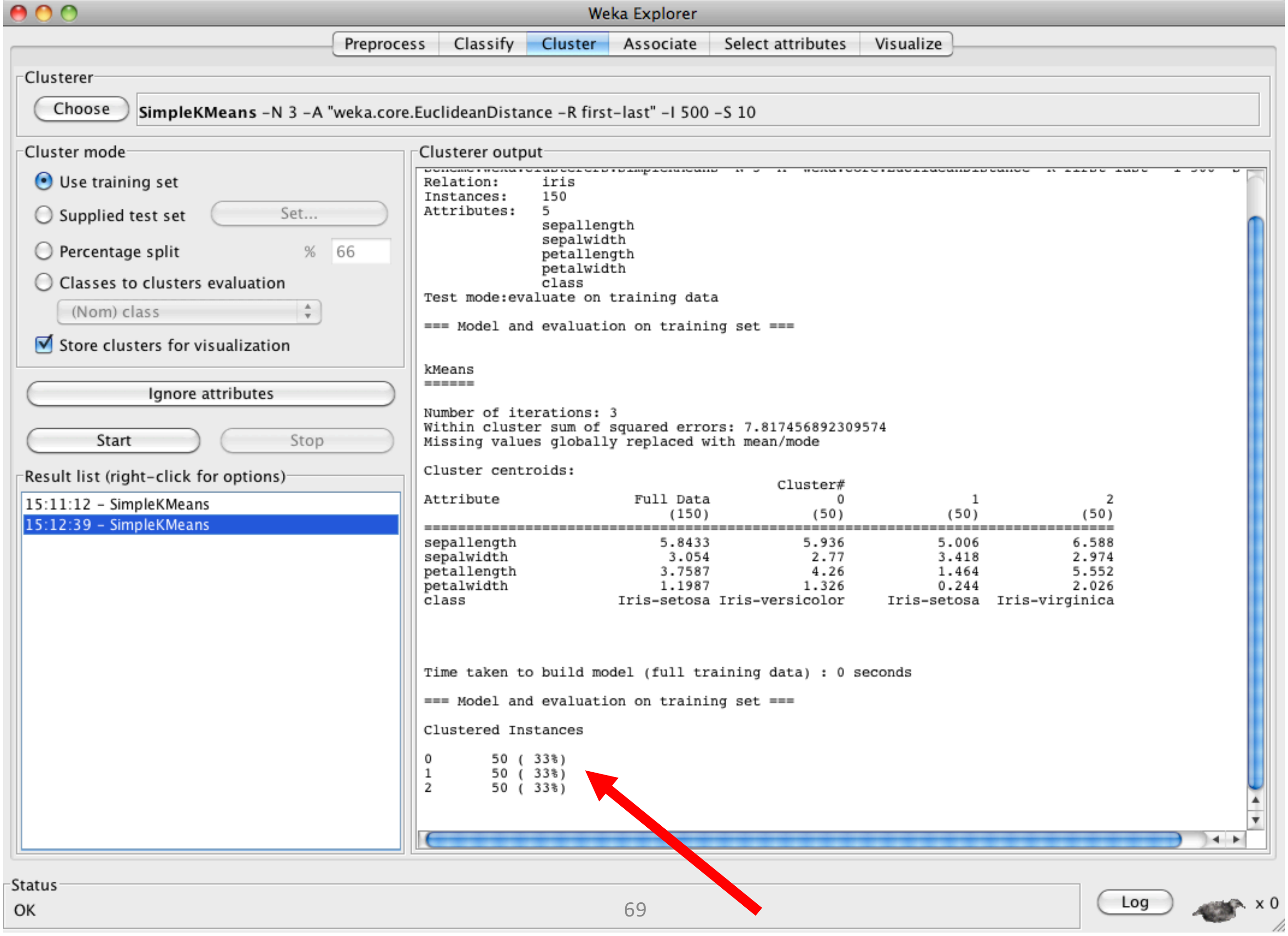
preserveInstancesOrder False

seed 10

15:11:12 - Sim

```
s -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S
a
ng set ===
rs: 62.1436882815797
ith mean/mode

                Cluster#
                       0                1
                   (100)             (50)
==================================
                   6.262            5.006
                   2.872            3.418
                   4.906            1.464
                   1.676            0.244
         Iris-versicolor      Iris-setosa

Time taken to build model (full training data) : 0.02 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      100 ( 67%)
1       50 ( 33%)
```

OK



Clusterer: SimpleKMeans -N 3 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10

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)

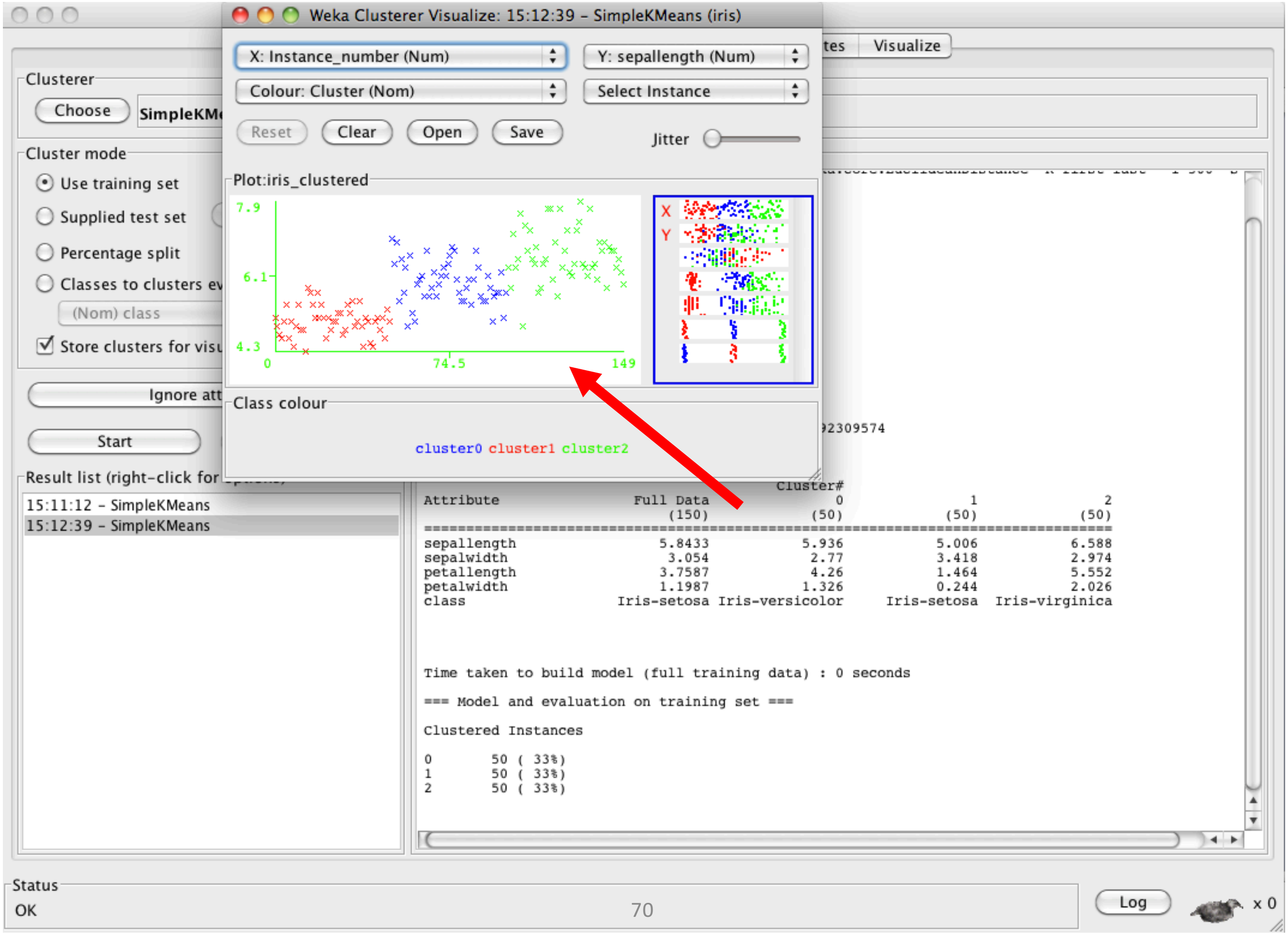
- 15:11:12 - SimpleKMeans
- 15:12:39 - SimpleKMeans

Clusterer output

```
Relation:     iris
Instances:    150
Attributes:   5
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
Test mode:evaluate on training data

=== Model and evaluation on training set ===


kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 7.817456892309574
Missing values globally replaced with mean/mode

Cluster centroids:
                                        Cluster#
Attribute               Full Data              0               1               2
                            (150)           (50)            (50)            (50)
================================================================================
sepallength                5.8433          5.936           5.006           6.588
sepalwidth                  3.054           2.77           3.418           2.974
petallength                3.7587           4.26           1.464           5.552
petalwidth                 1.1987          1.326           0.244           2.026
class                 Iris-setosa Iris-versicolor     Iris-setosa  Iris-virginica



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       50 ( 33%)
1       50 ( 33%)
2       50 ( 33%)
```

Status

OK



Clusterer: SimpleKMe

Cluster mode

- ● Use training set
- ○ Supplied test set
- ○ Percentage split
- ○ Classes to clusters ev
- (Nom) class
- ☑ Store clusters for visu

Result list (right-click for

- 15:11:12 - SimpleKMeans
- 15:12:39 - SimpleKMeans

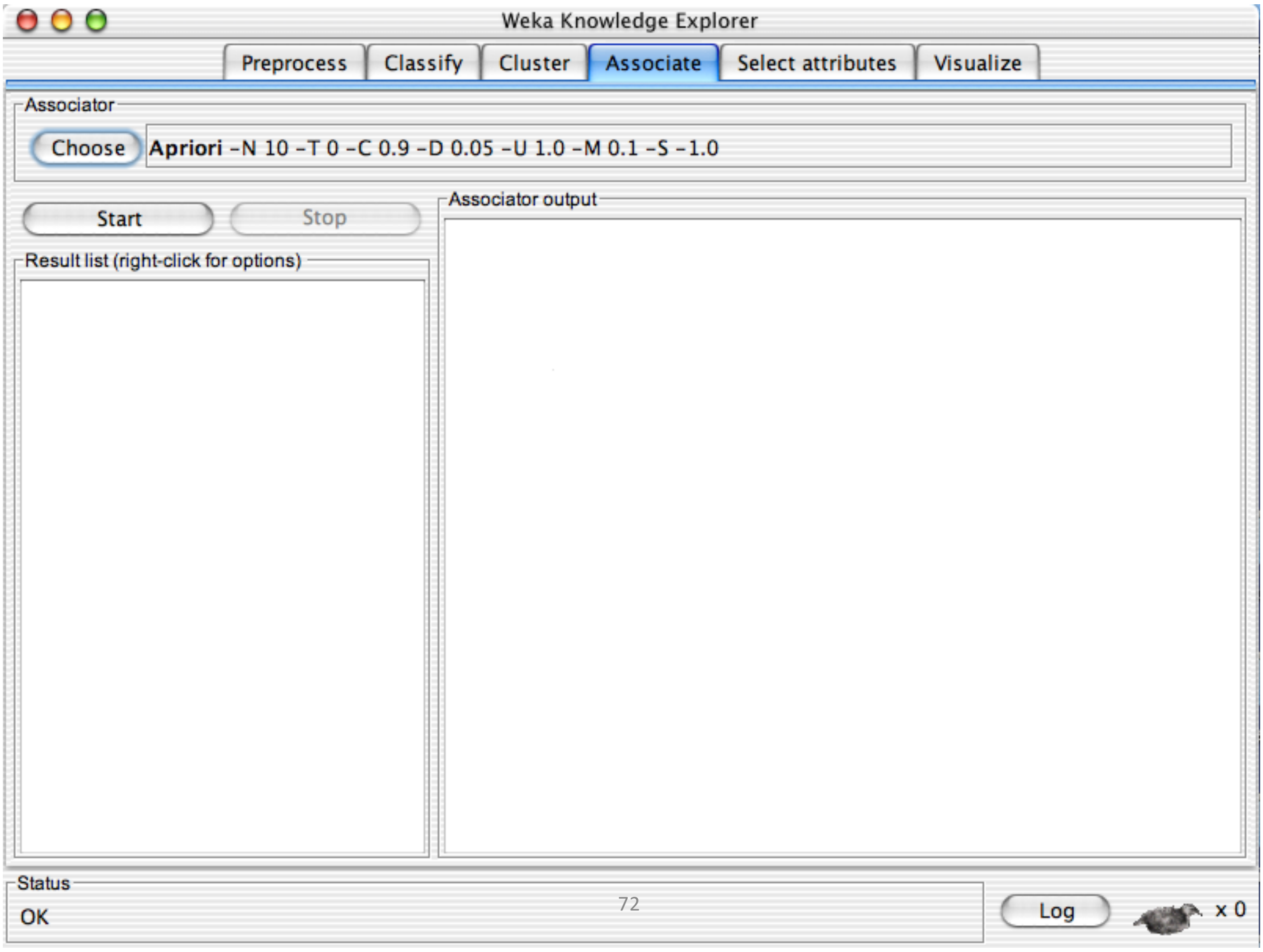
```
                                        Cluster#
Attribute               Full Data              0               1               2
                            (150)           (50)            (50)            (50)
================================================================================
sepallength                5.8433          5.936           5.006           6.588
sepalwidth                  3.054           2.77           3.418           2.974
petallength                3.7587           4.26           1.464           5.552
petalwidth                 1.1987          1.326           0.244           2.026
class                 Iris-setosa Iris-versicolor     Iris-setosa  Iris-virginica



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       50 ( 33%)
1       50 ( 33%)
2       50 ( 33%)
```

Status

OK

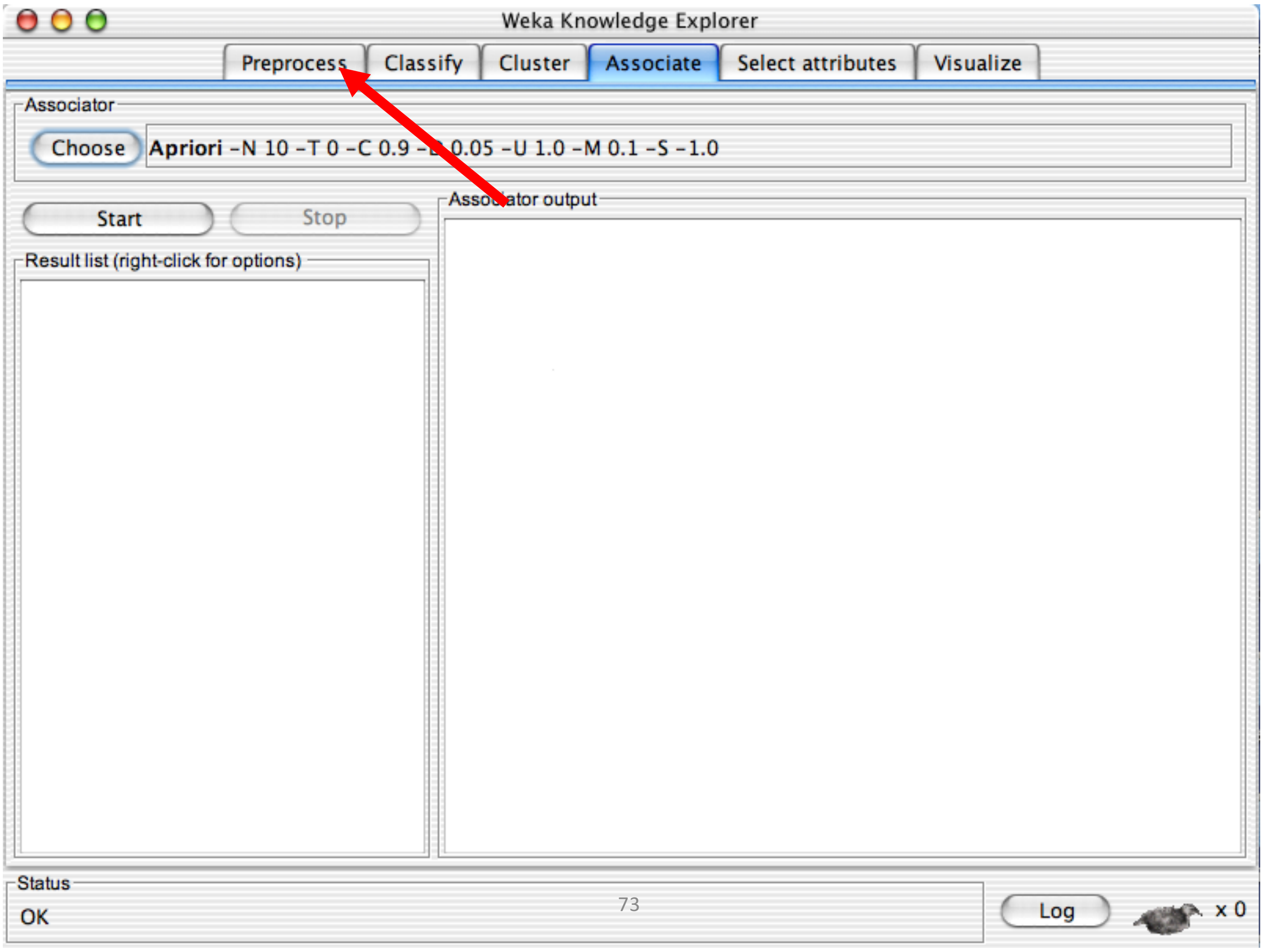
Finding Associations

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



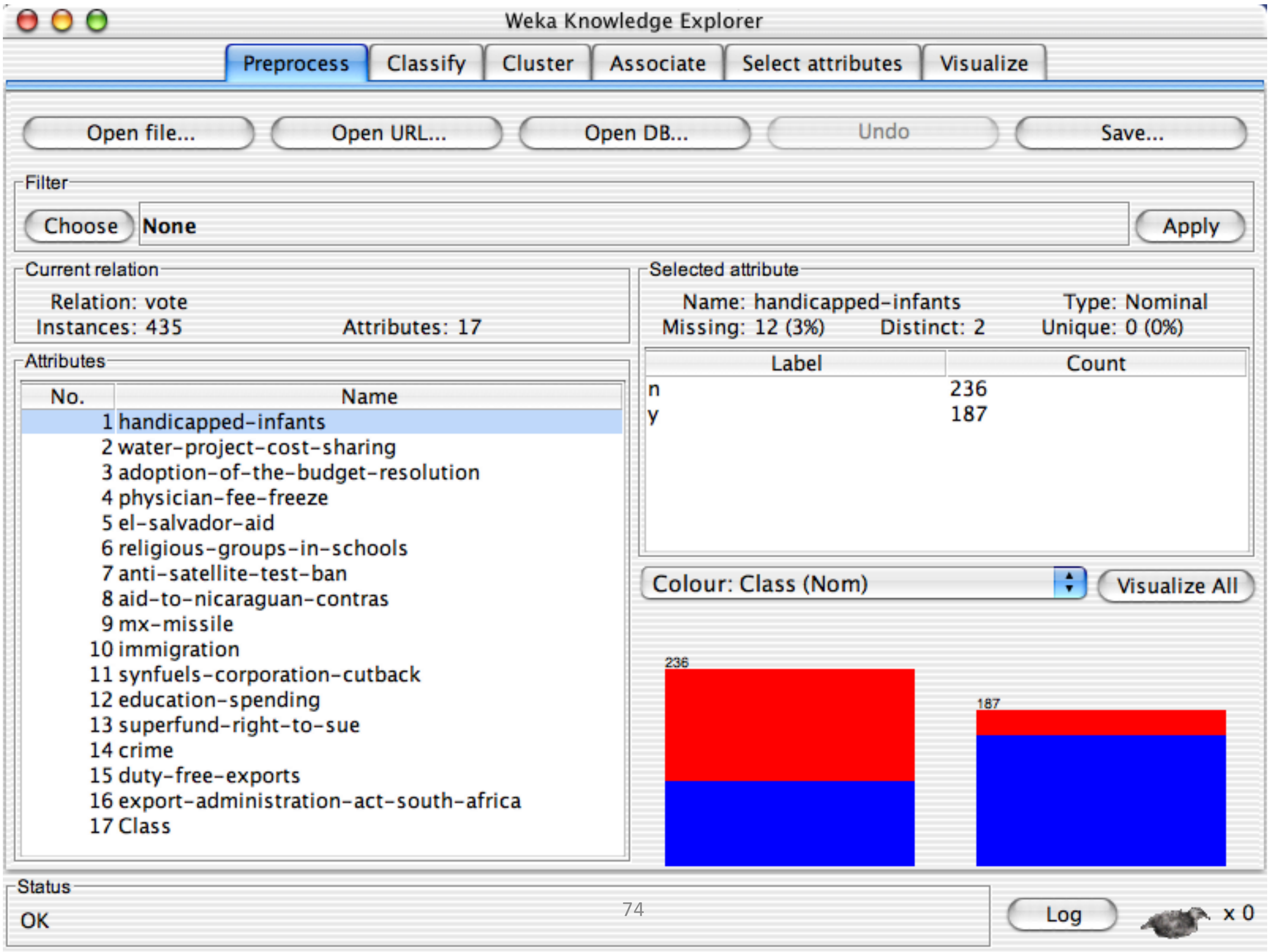
Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0

OK



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

Status: OK



Filter: None

Relation: vote
Instances: 435 Attributes: 17

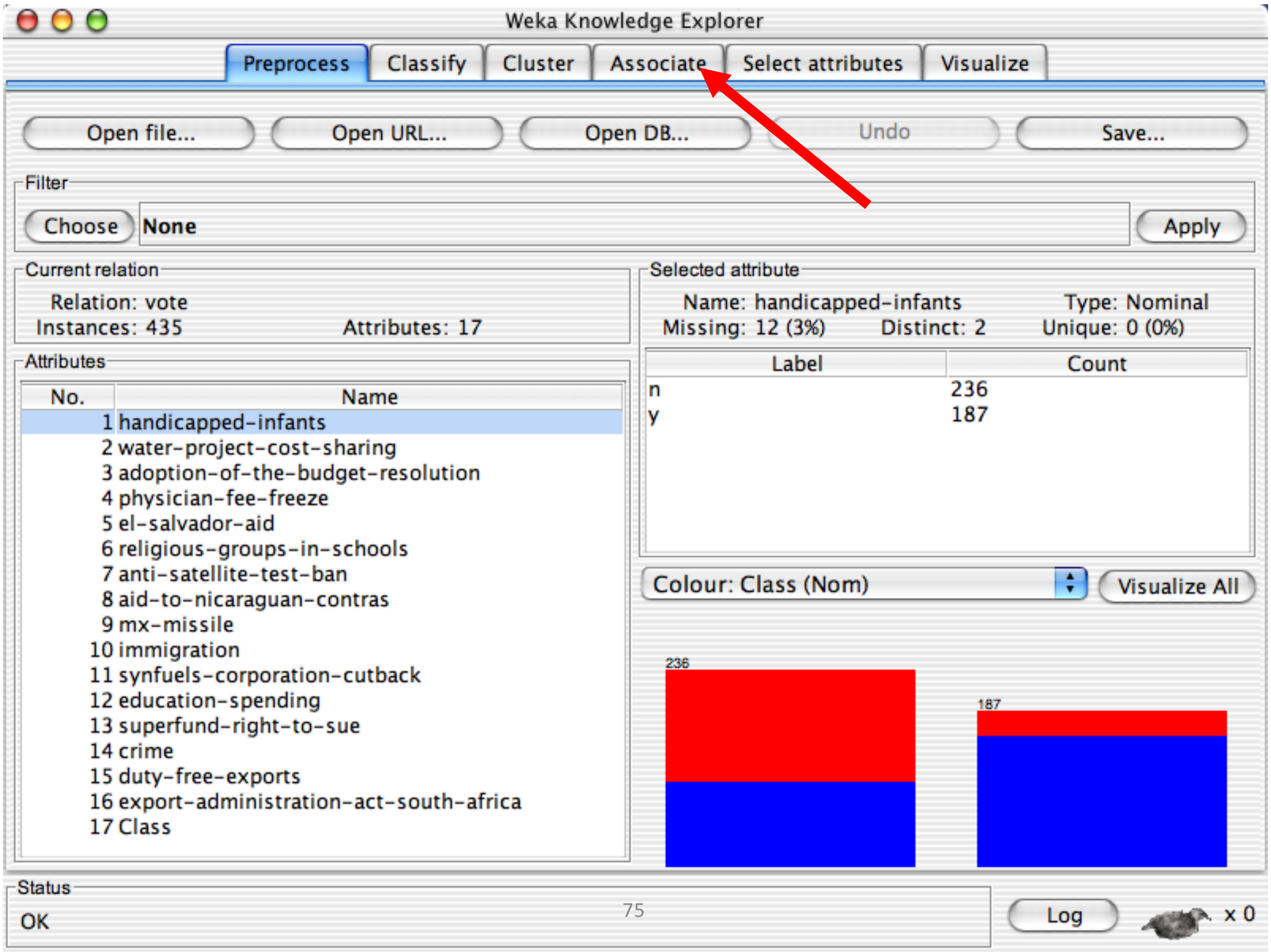
No.	Name
1	handicapped-infants
2	water-project-cost-sharing
3	adoption-of-the-budget-resolution
4	physician-fee-freeze
5	el-salvador-aid
6	religious-groups-in-schools
7	anti-satellite-test-ban
8	aid-to-nicaraguan-contras
9	mx-missile
10	immigration
11	synfuels-corporation-cutback
12	education-spending
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

Colour: Class (Nom)

Status: OK



Filter: None

Current relation

Relation: vote
Instances: 435 Attributes: 17

Attributes

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

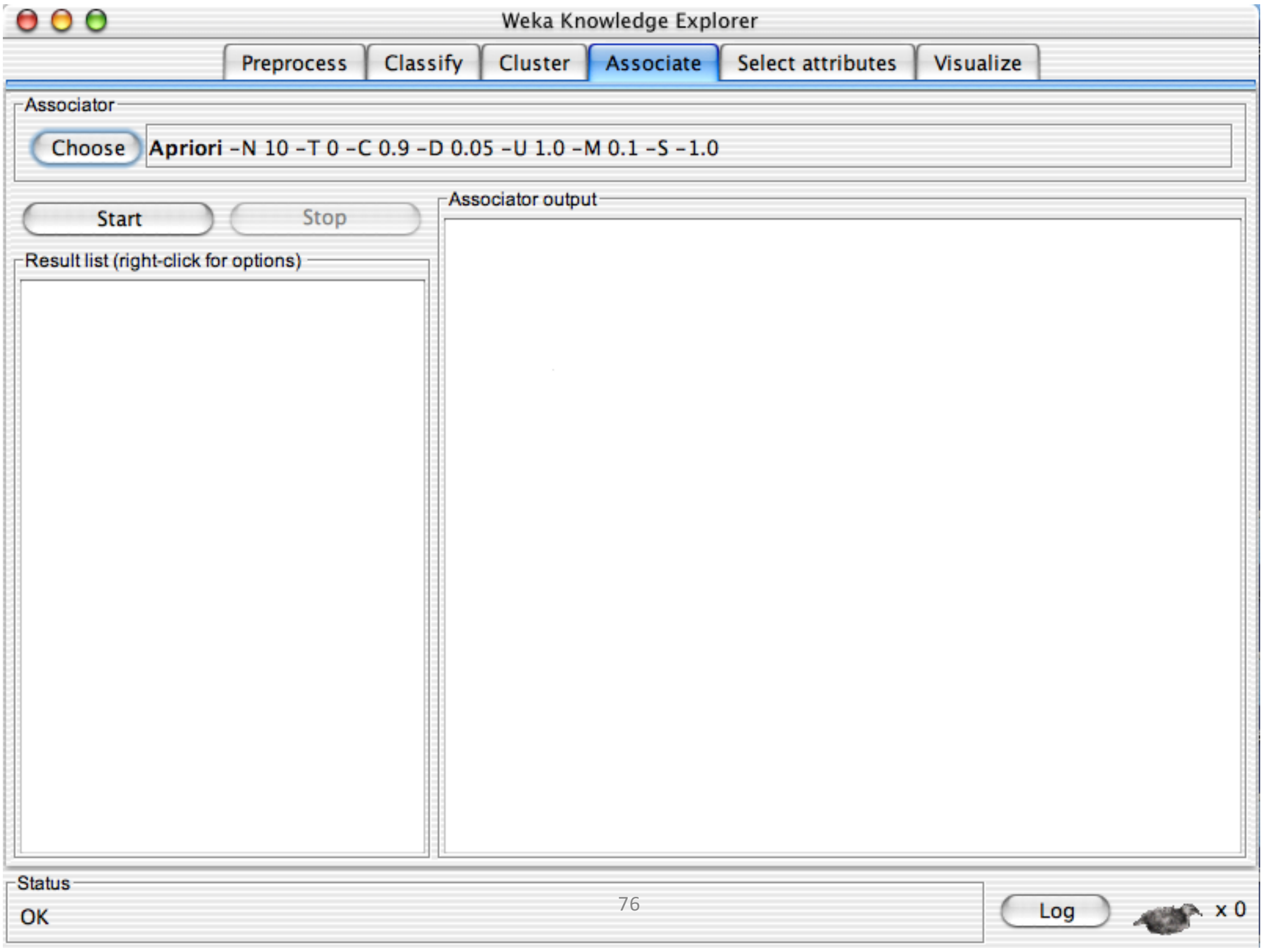
Selected attribute

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

Label	Count
n	236
y	187

Colour: Class (Nom)

Status
OK



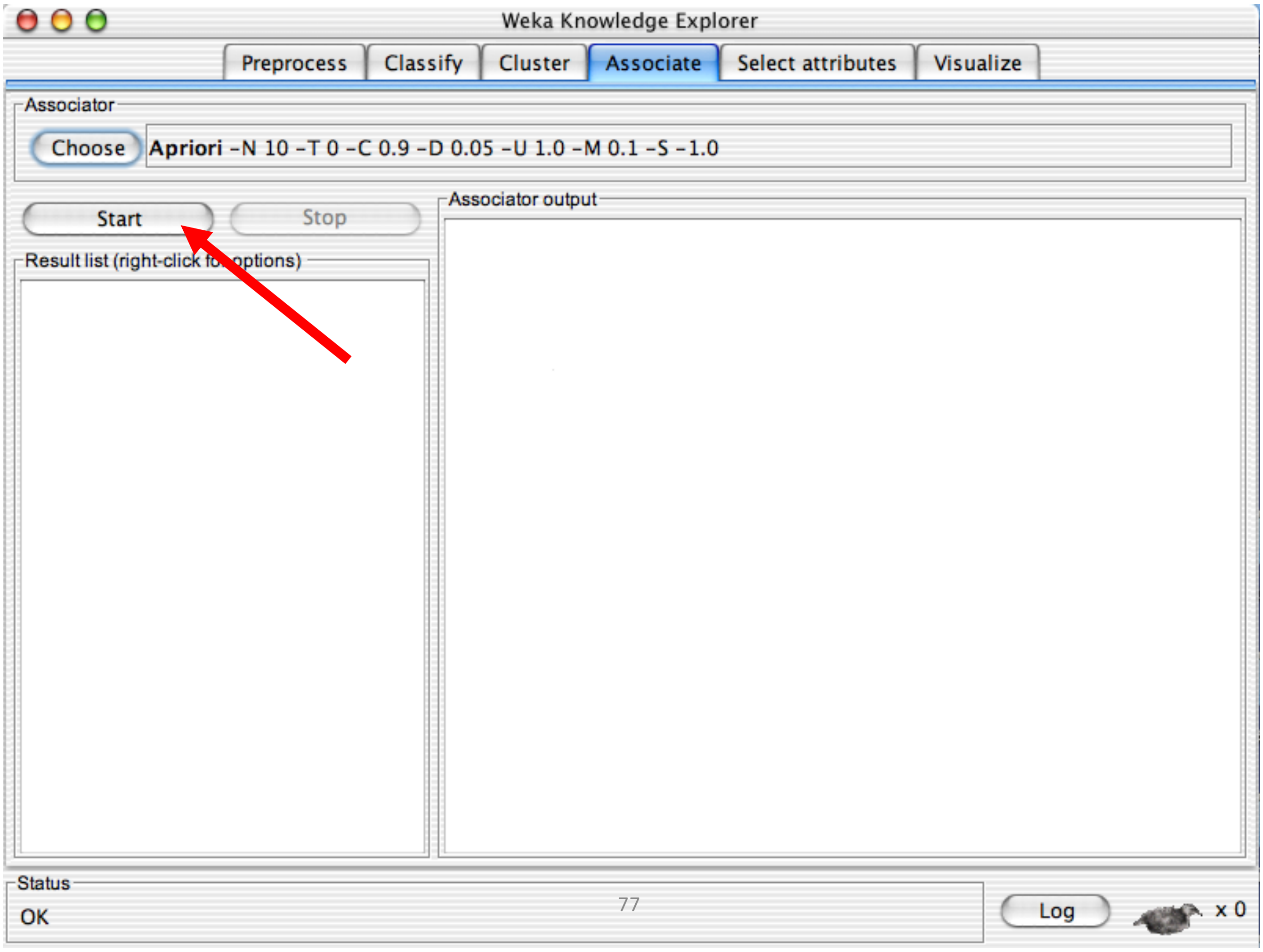
Associator

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

Result list (right-click for options)

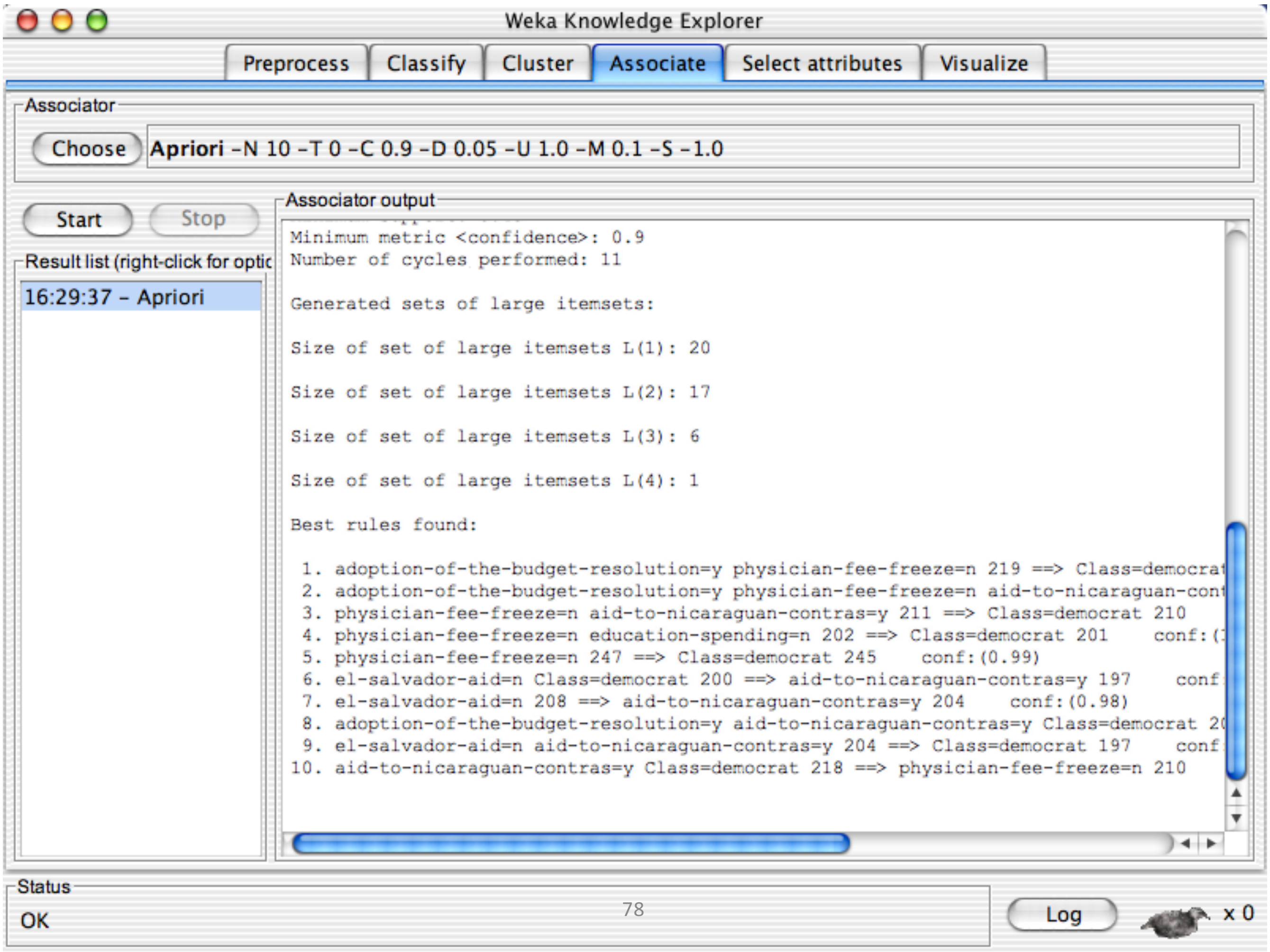
Associator output

Status
OK



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

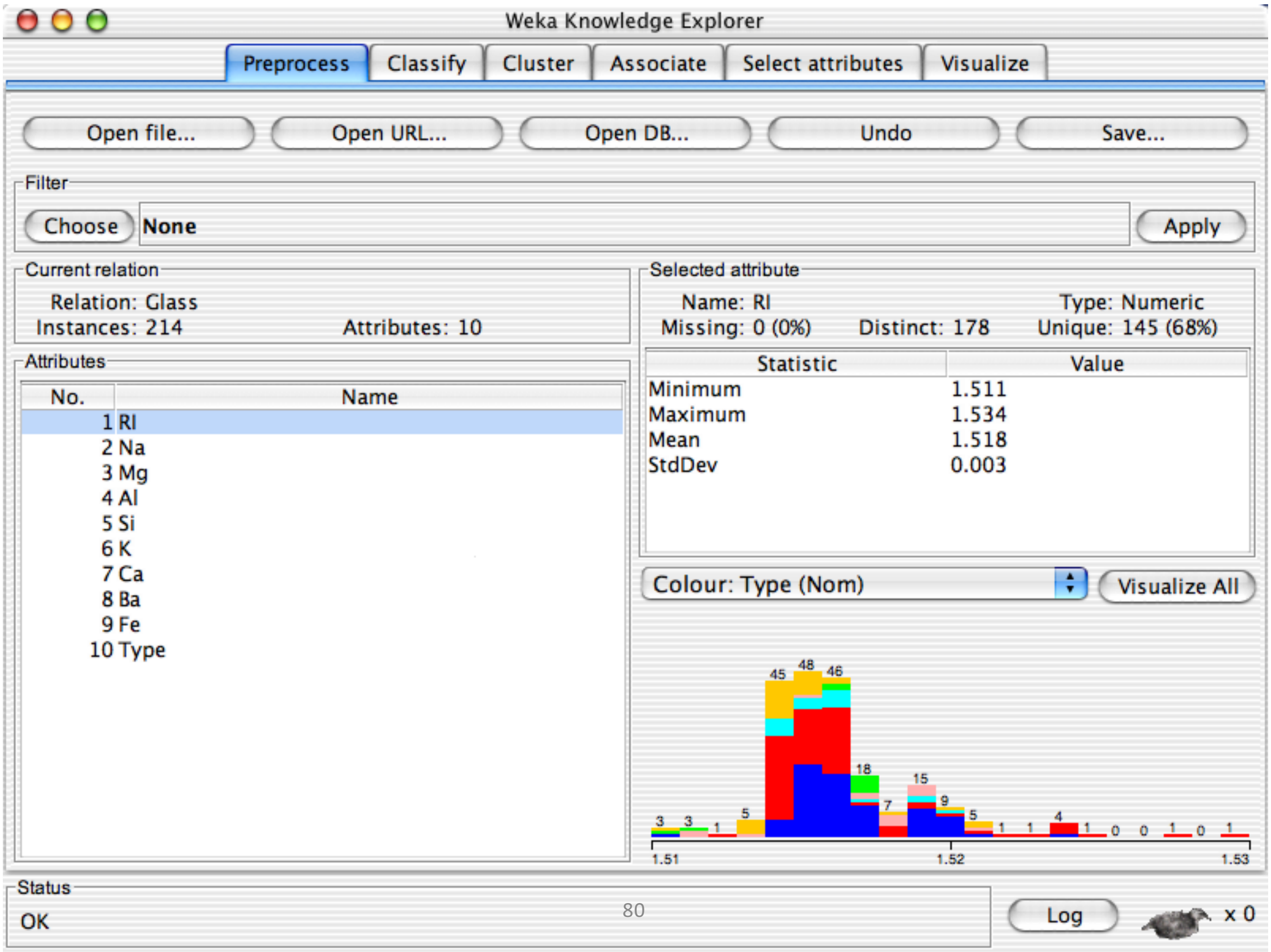
16:29:37 - Apriori

```
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
```

Status: OK

Data visualization

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



Weka Knowledge Explorer

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

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

Filter

Choose | None | Apply

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%)

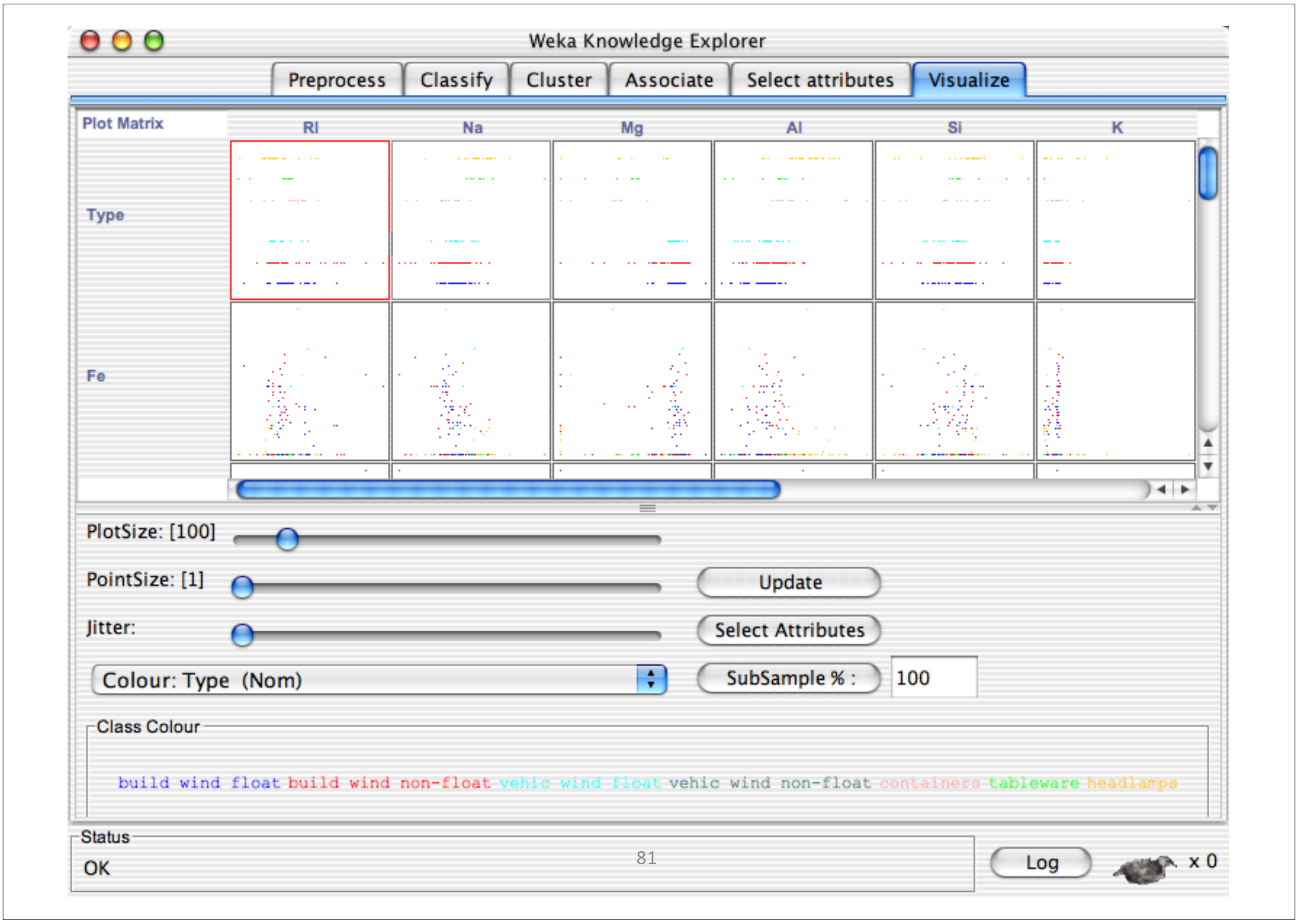
Statistic	Value
Minimum	1.511
Maximum	1.534
Mean	1.518
StdDev	0.003

Colour: Type (Nom) | Visualize All

Status

OK

Log | x 0



PlotSize: [100]

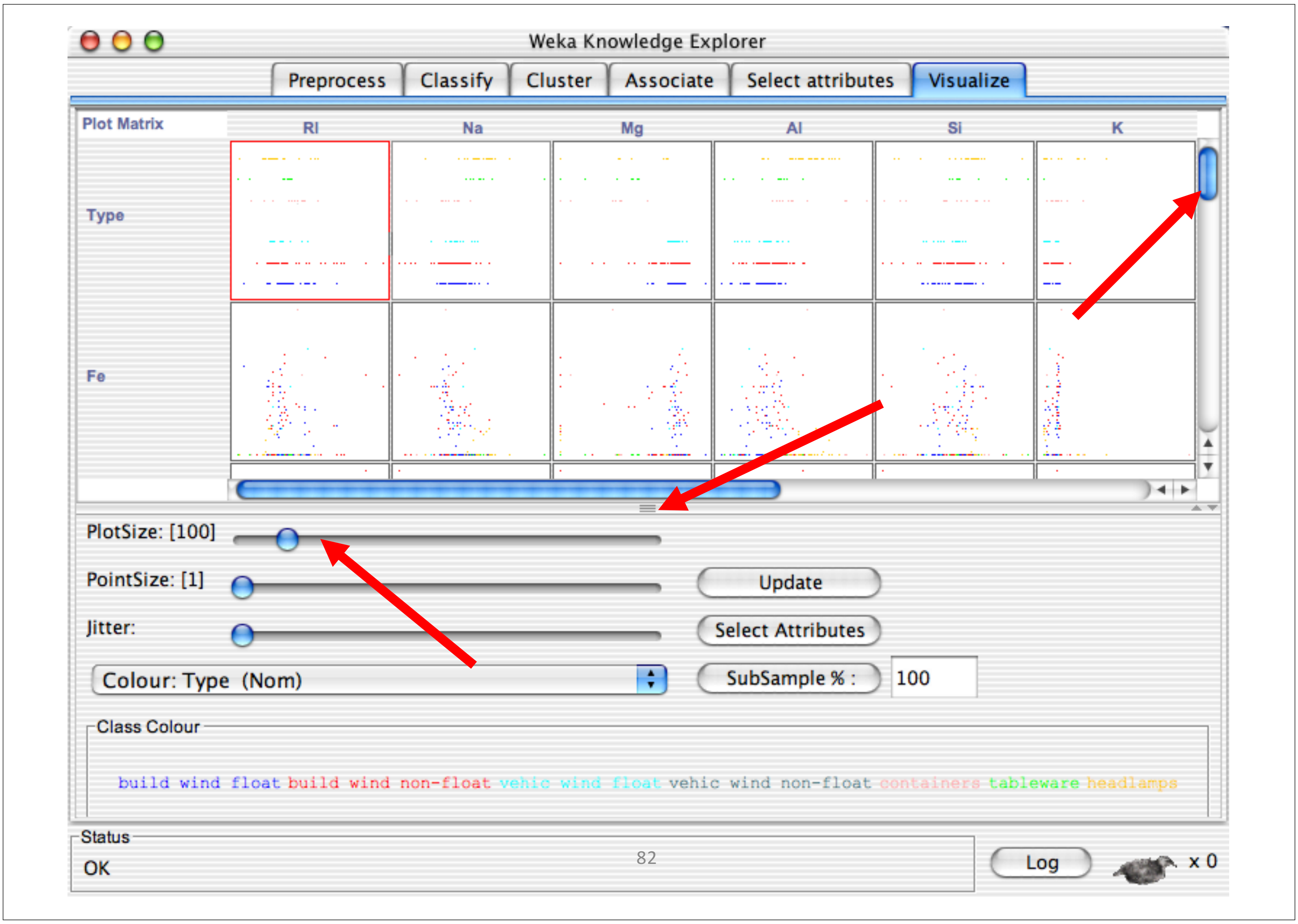
PointSize: [1]

Jitter:

Colour: Type (Nom)

SubSample % : 100

OK



PlotSize: [100]

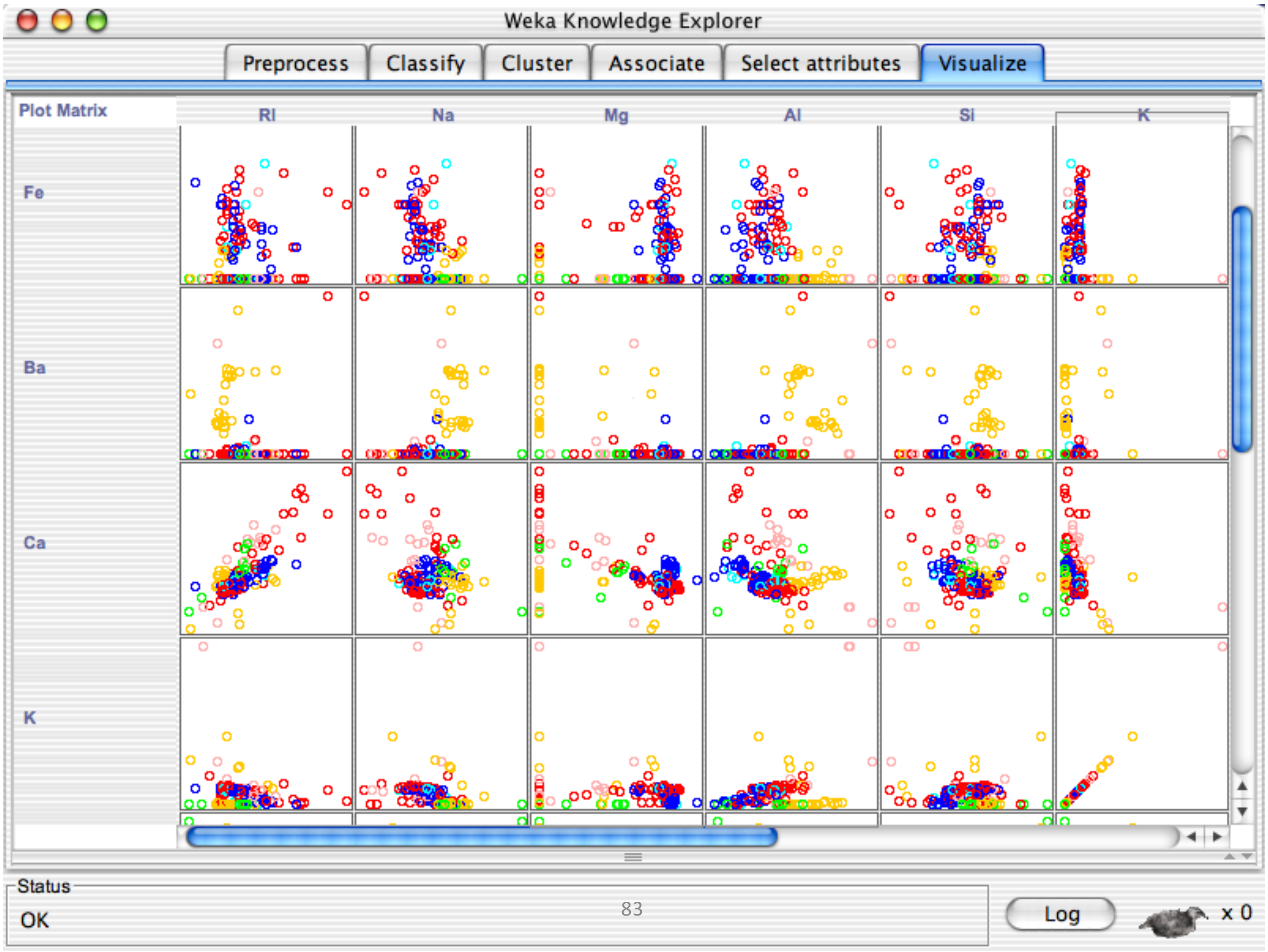
PointSize: [1]

Jitter:

Colour: Type (Nom)

SubSample % : 100

OK


Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Plot Matrix
RI
Na
Mg
Al
Si
K
Fe
Ba
Ca
K
83
Log
x 0

Status

OK