

Agenda

- **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



- 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

Explorer: pre-processing the data



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

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 only deals with “flat” files



@relation heart-disease-simplified

@attribute age numeric

@attribute sex { female, male }

@attribute chest_pain_type { typ_angina, asympt, non_anginal, atyp_angina }

@attribute cholesterol numeric

@attribute exercise_induced_angina { no, yes }

@attribute class { present, not_present }

@data

63,male,typ_angina,233,no,not_present

67,male,asympt,286,yes,present

67,male,asympt,229,yes,present

38,female,non_anginal,?,no,not_present

numeric attribute

nominal attribute

WEKA with “flat” files (Sparse format)

@relation example

@attribute att1 numeric

@attribute att2 numeric

@attribute att3 numeric

@attribute att4 numeric

@attribute class {"class A", "class B"}



@data

0, X, 0, Y, "class A"

0, 0, W, 0, "class B"

@relation example

@attribute att1 numeric

@attribute att2 numeric

@attribute att3 numeric

@attribute att4 numeric

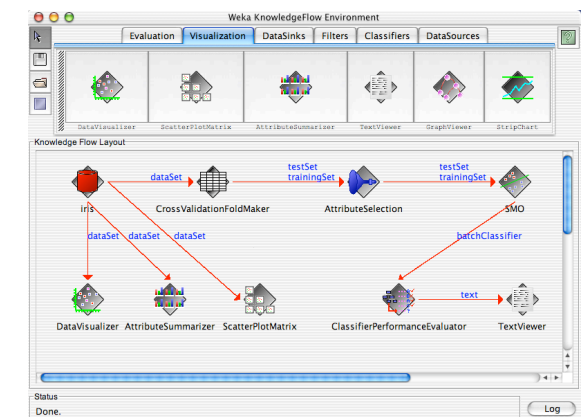
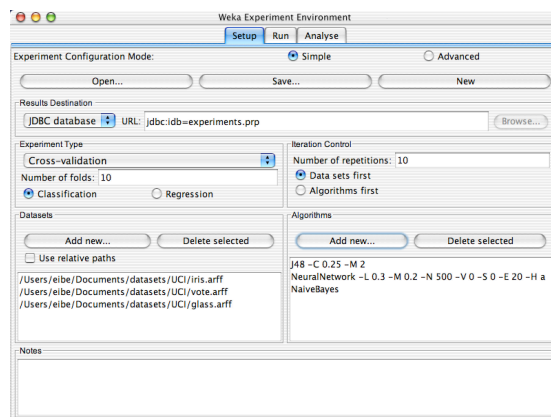
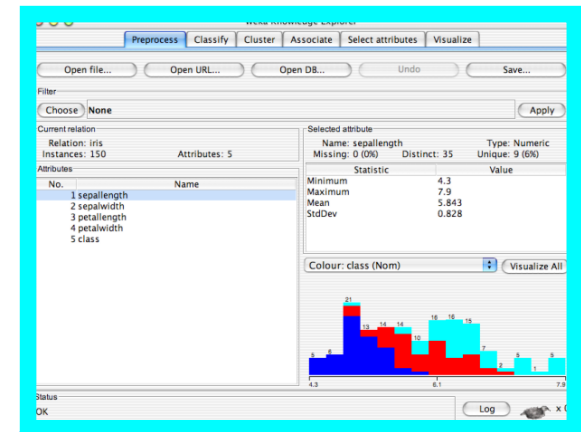
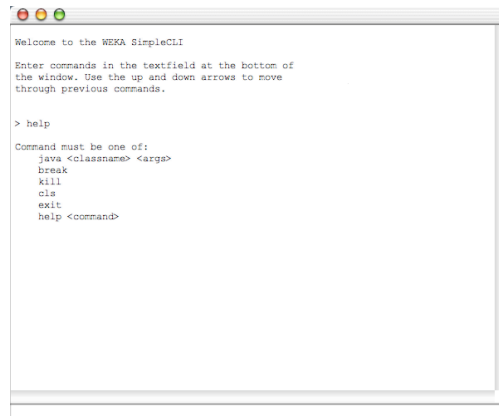
@attribute class {"class A", "class B"}

@data

{1 X, 3 Y, 4 "class A"}

{2 W, 4 "class B"}

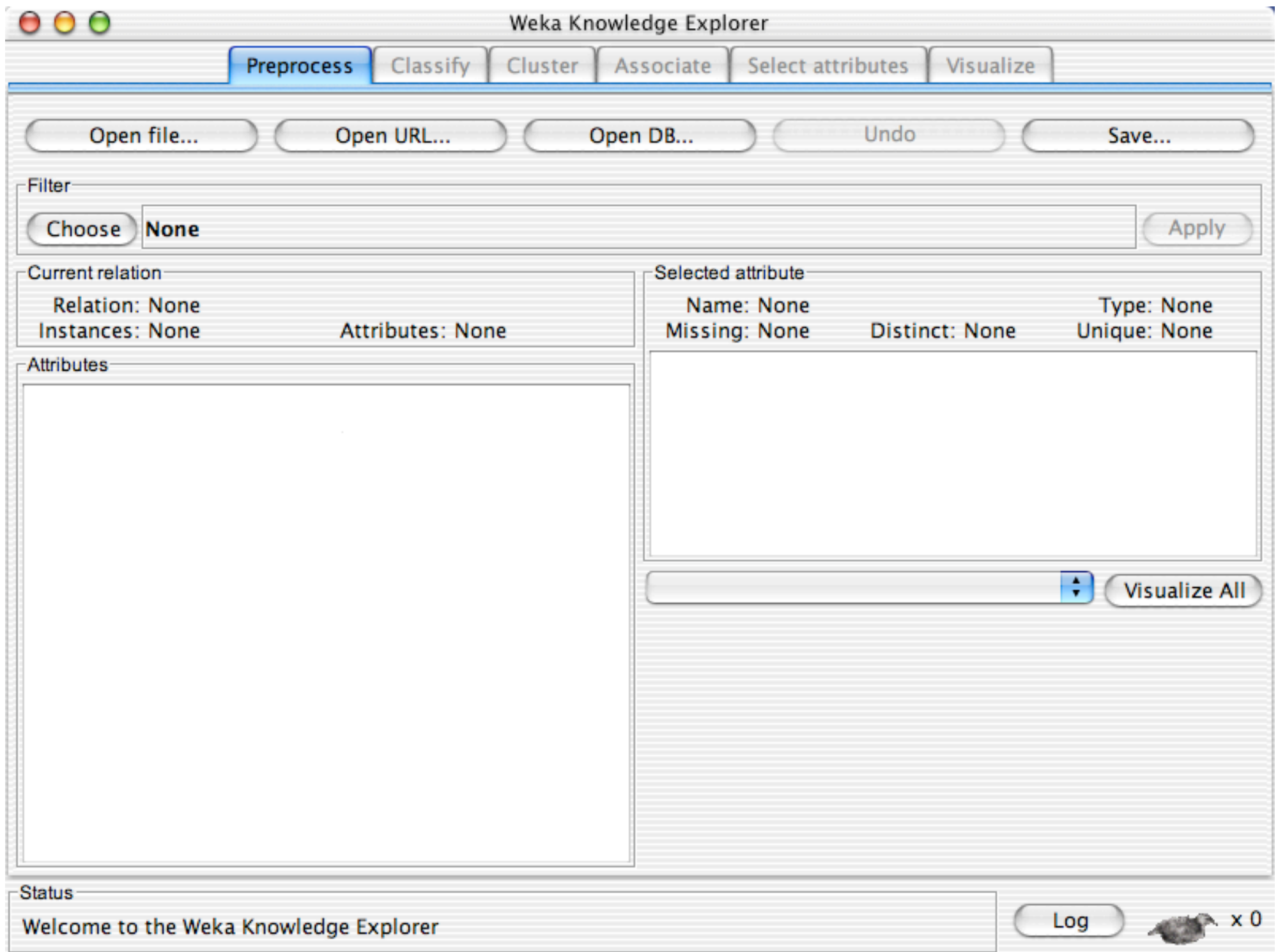
- Considering only the non 0 values
- Represent each values with: POSITION VALUE information
- Each couple (POSITION - VALUE) is separated with a comma
- Useful for documents representation in ARFF format

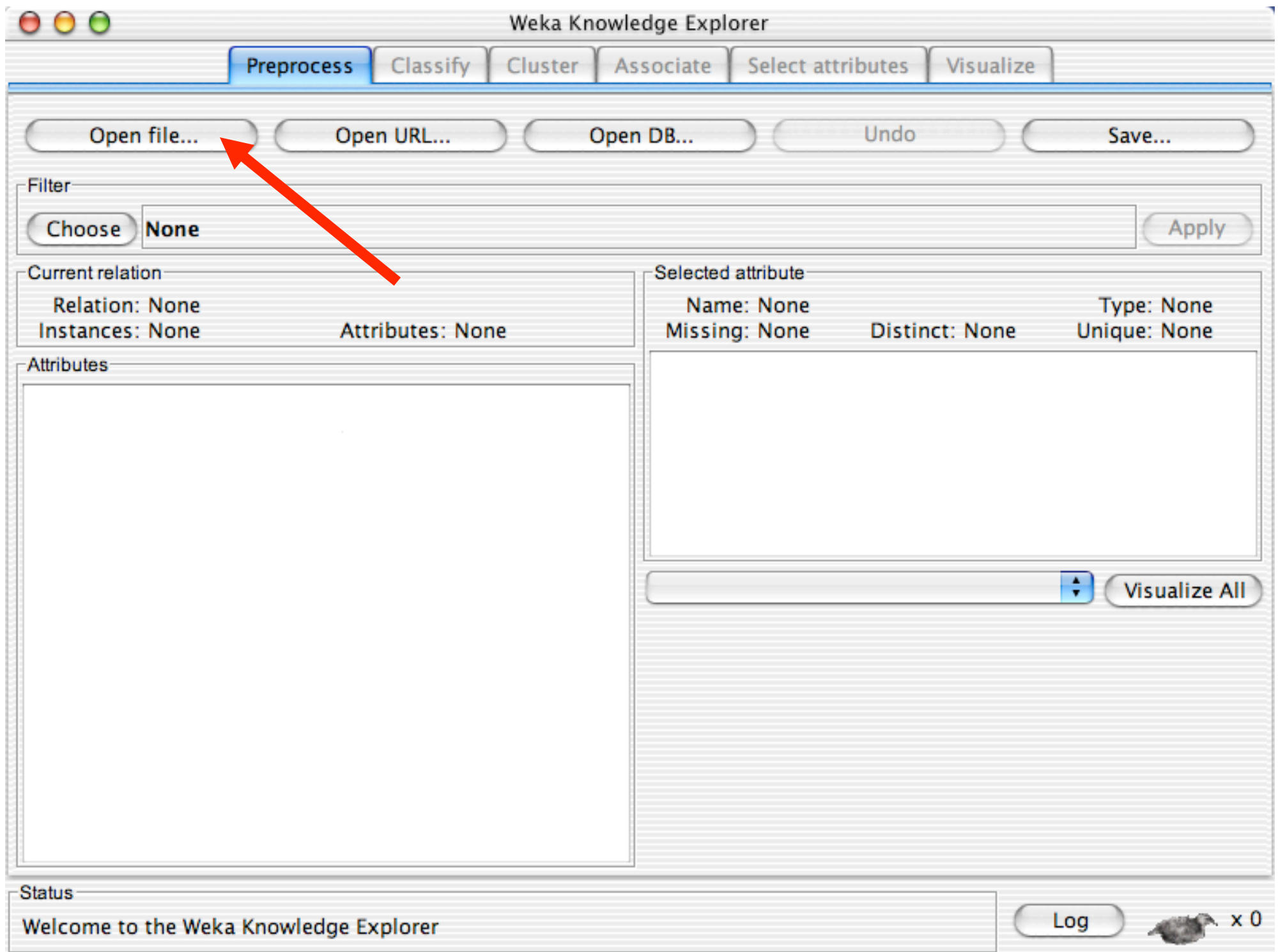


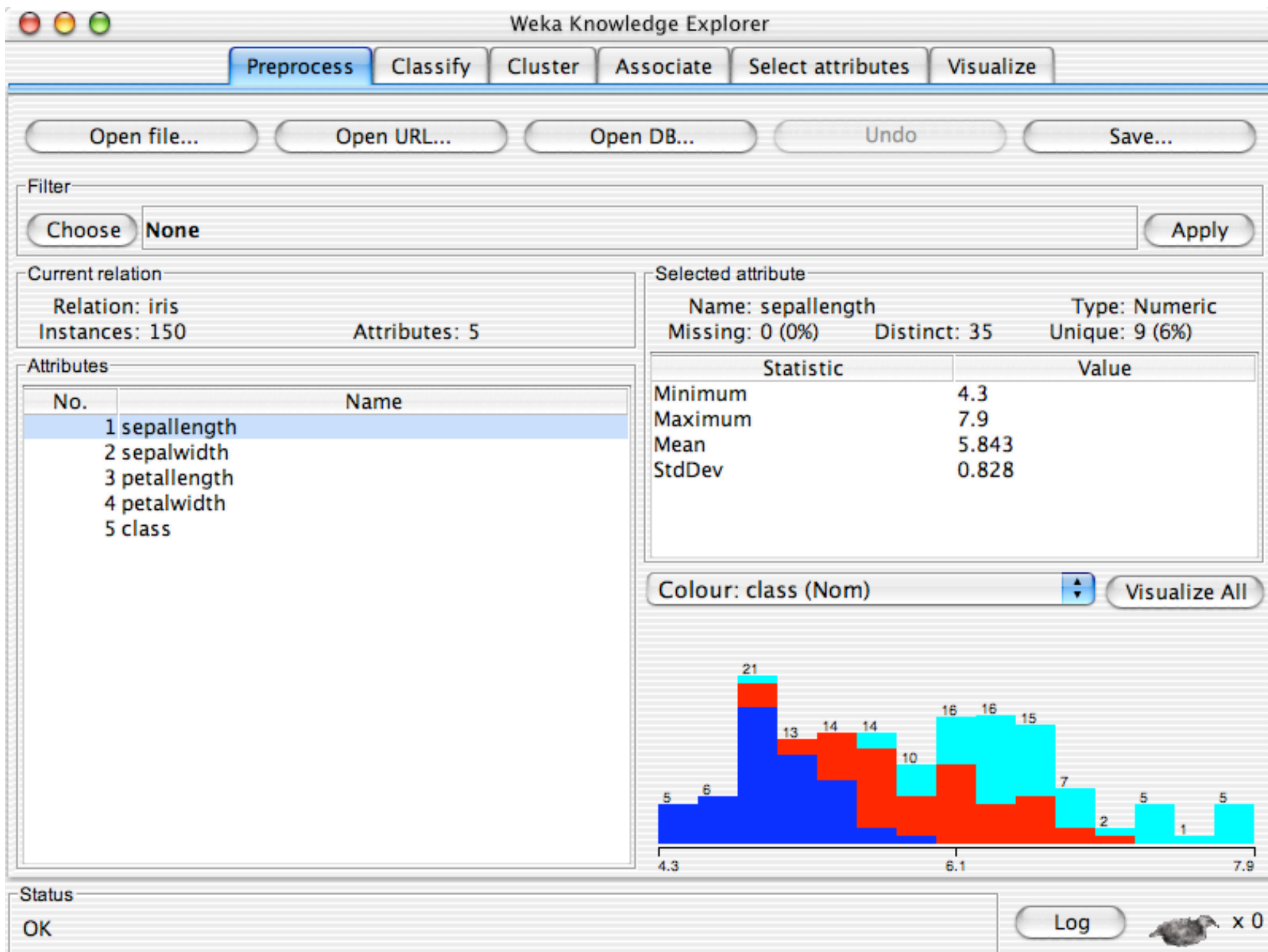
Explorer: Pre-processing

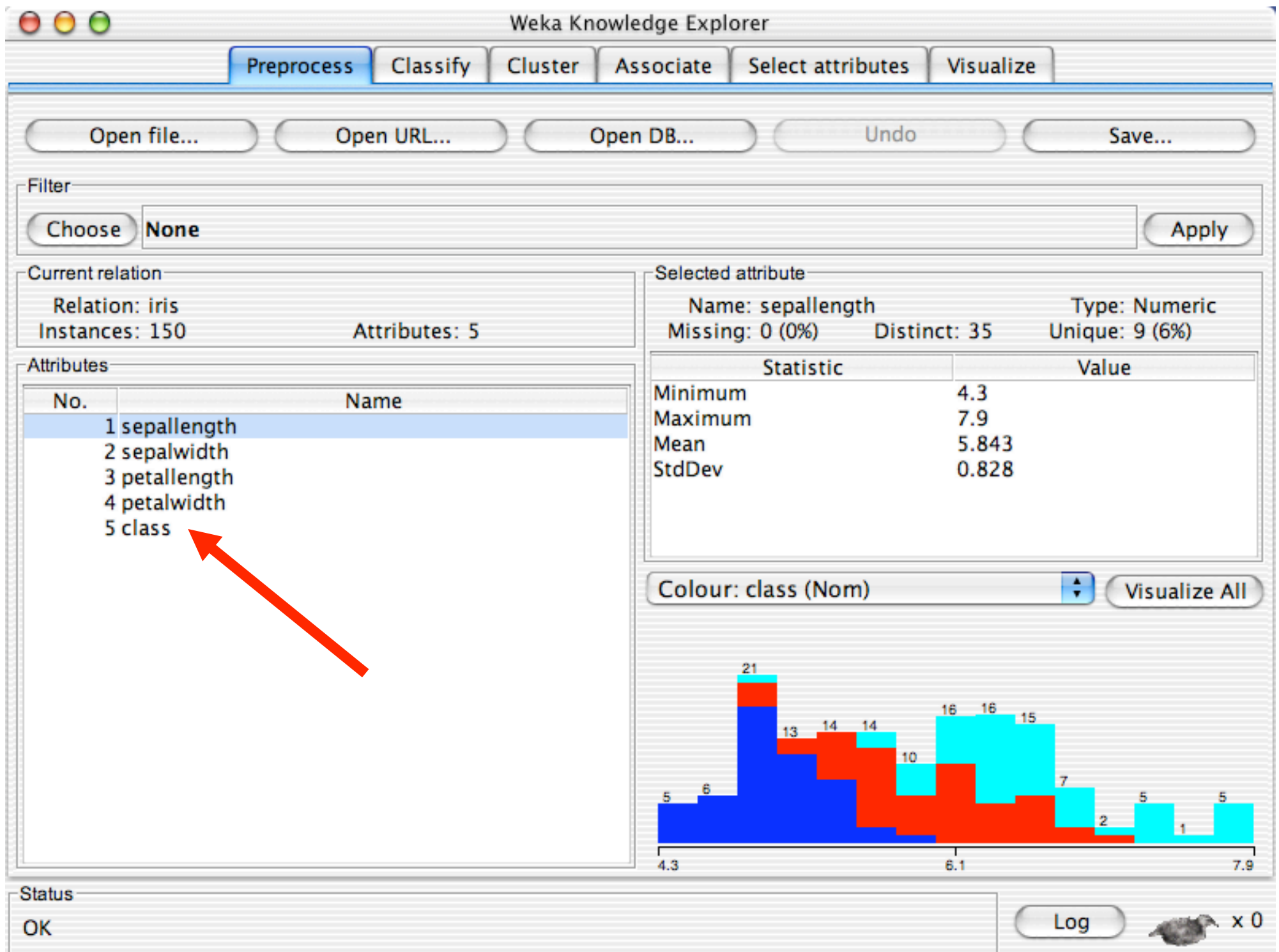


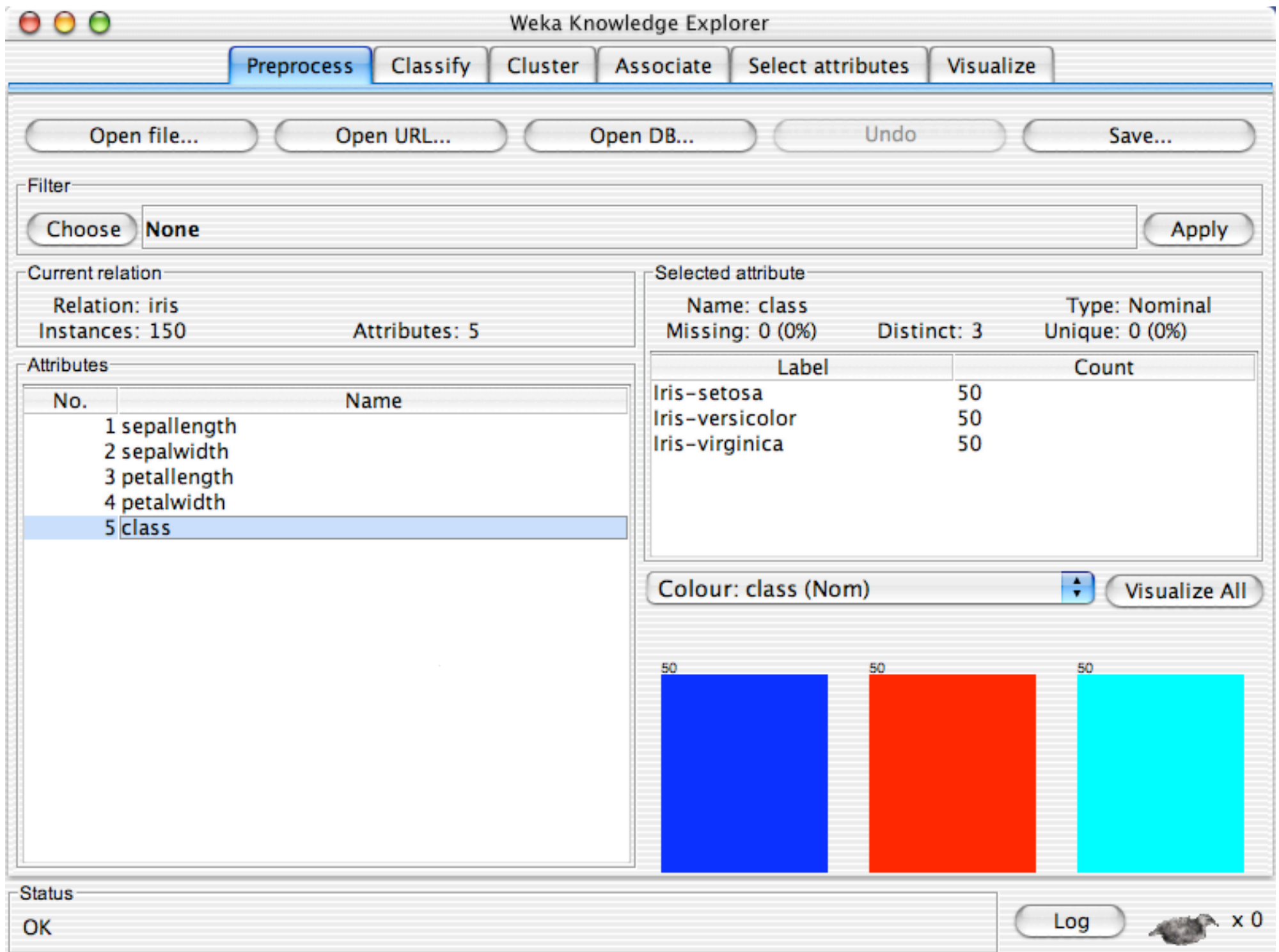
- Data can be imported from a file in various formats:
 - ARFF, CSV, C4.5, binary
- Data can also be read from a **URL** or from an **SQL** database (using JDBC)
- Pre-processing tools in WEKA are called “**filters**”
- WEKA contains filters for:
 - Discretization, normalization, resampling, attribute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

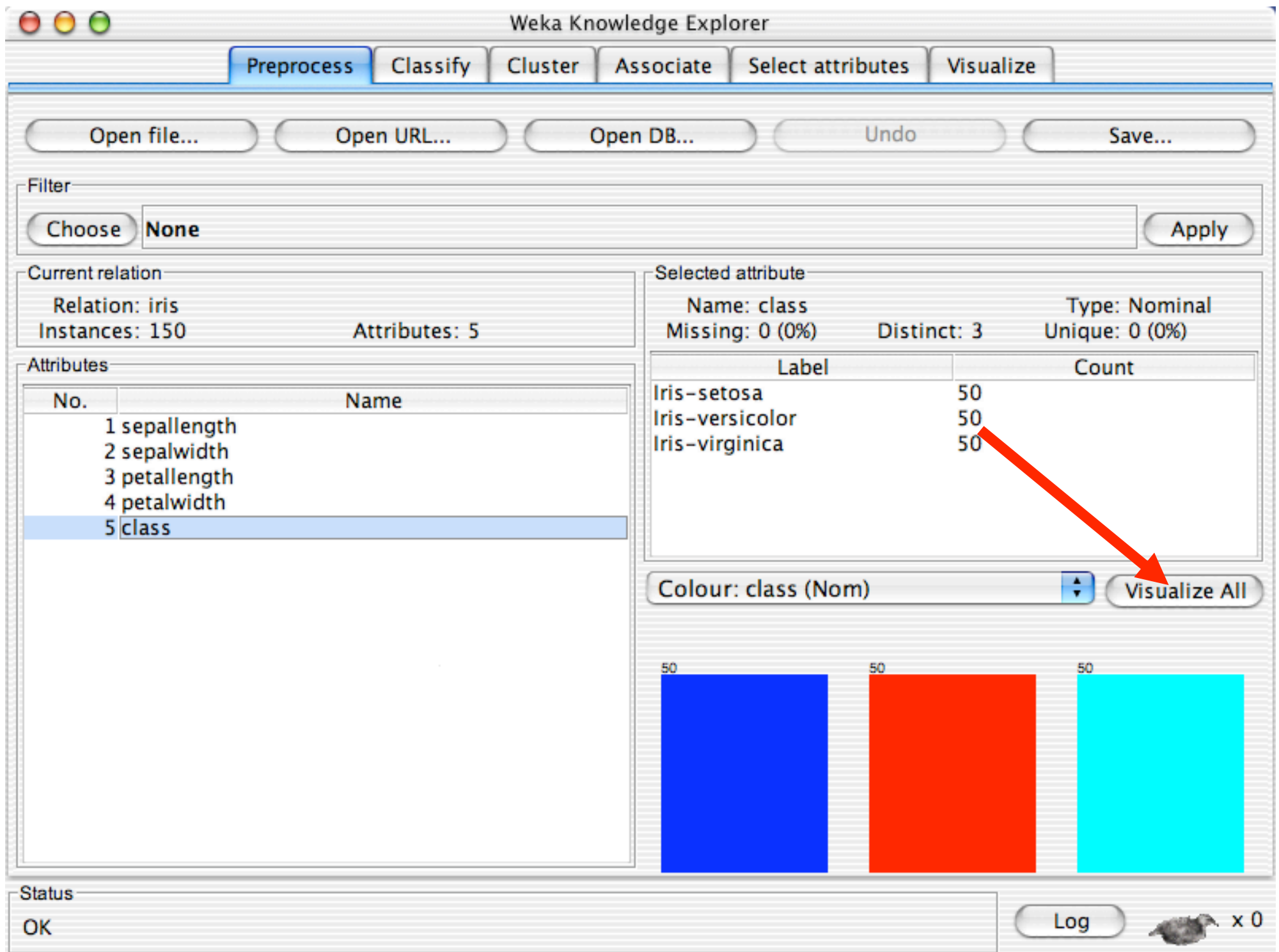


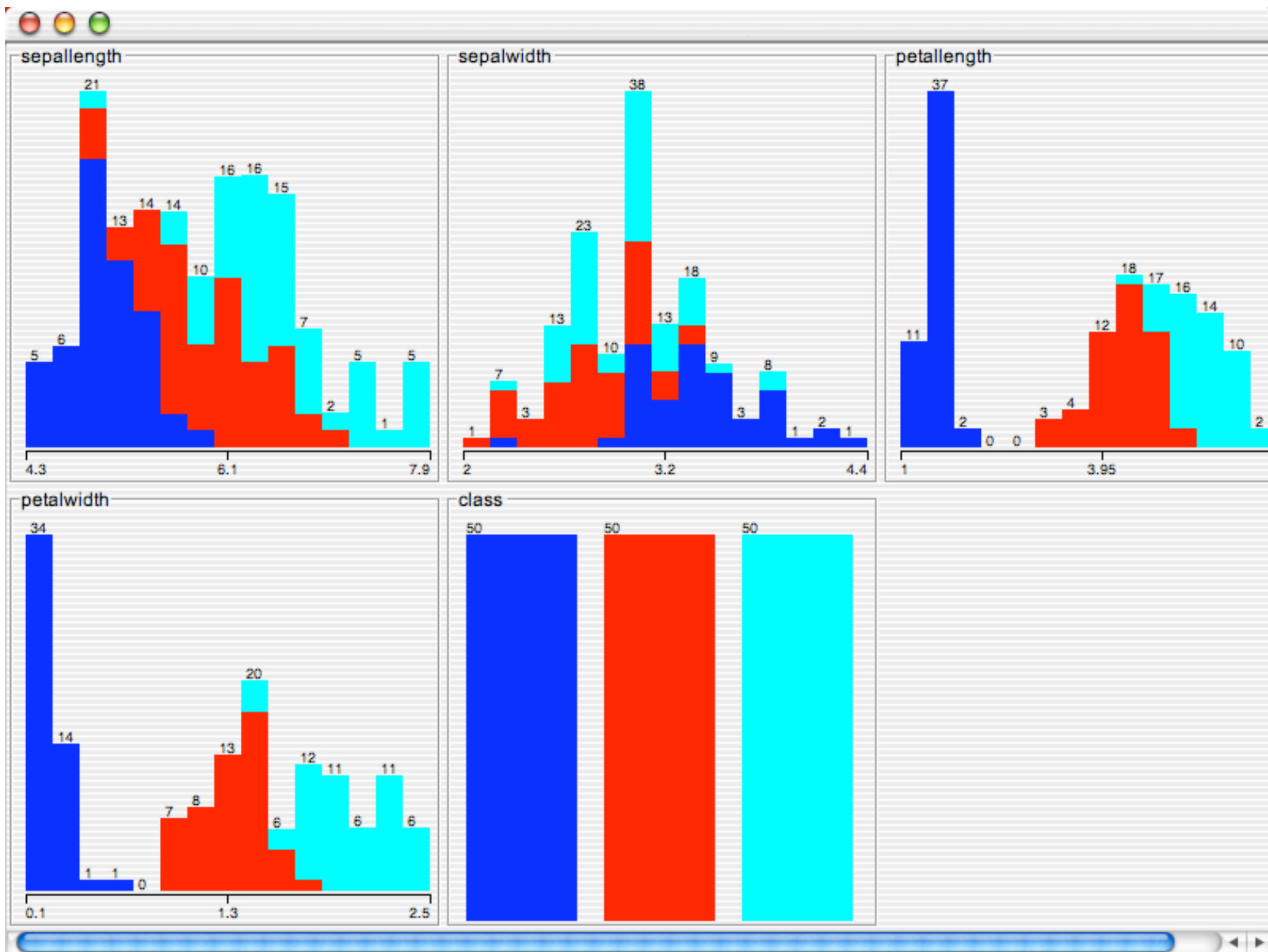


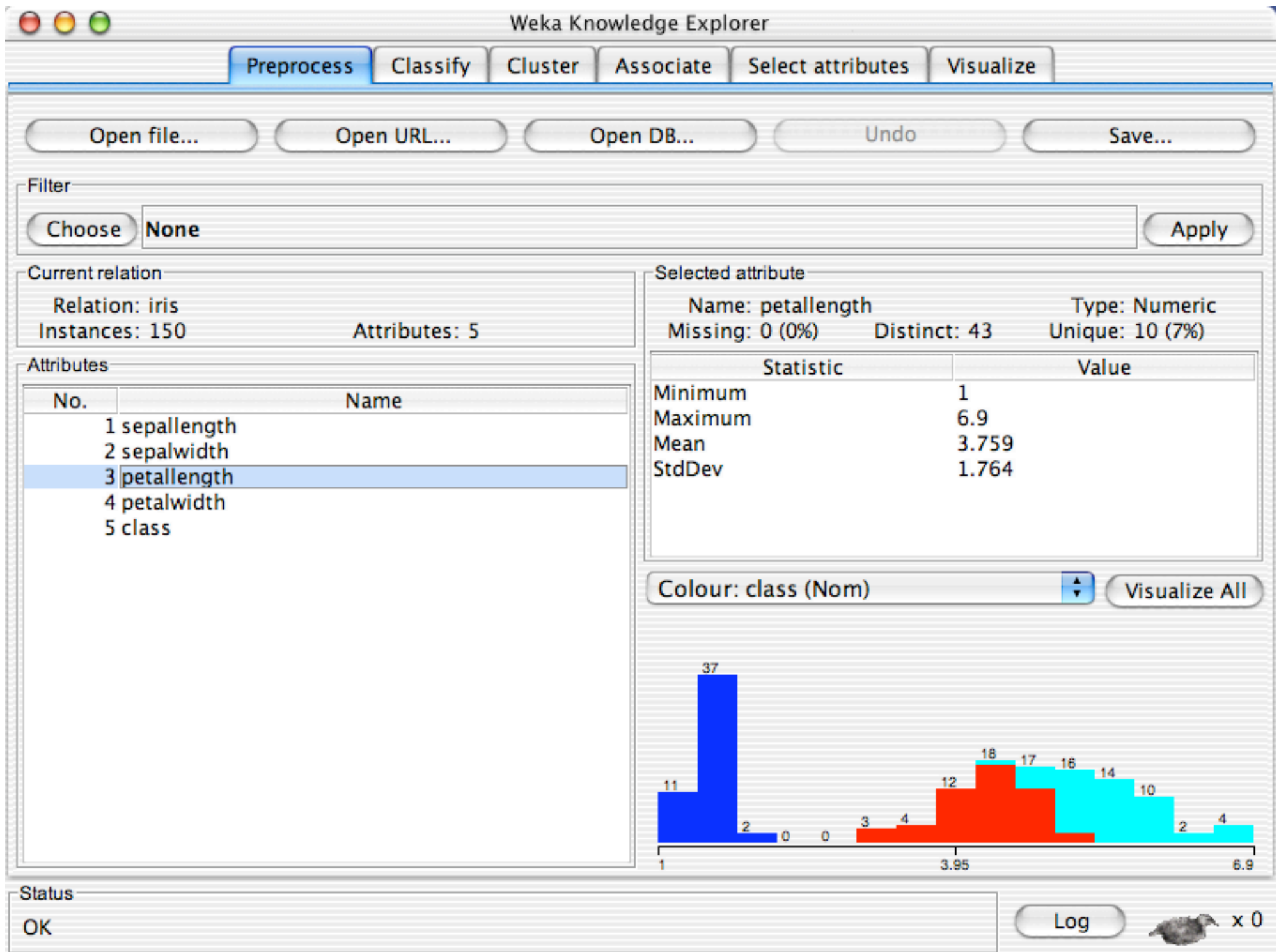


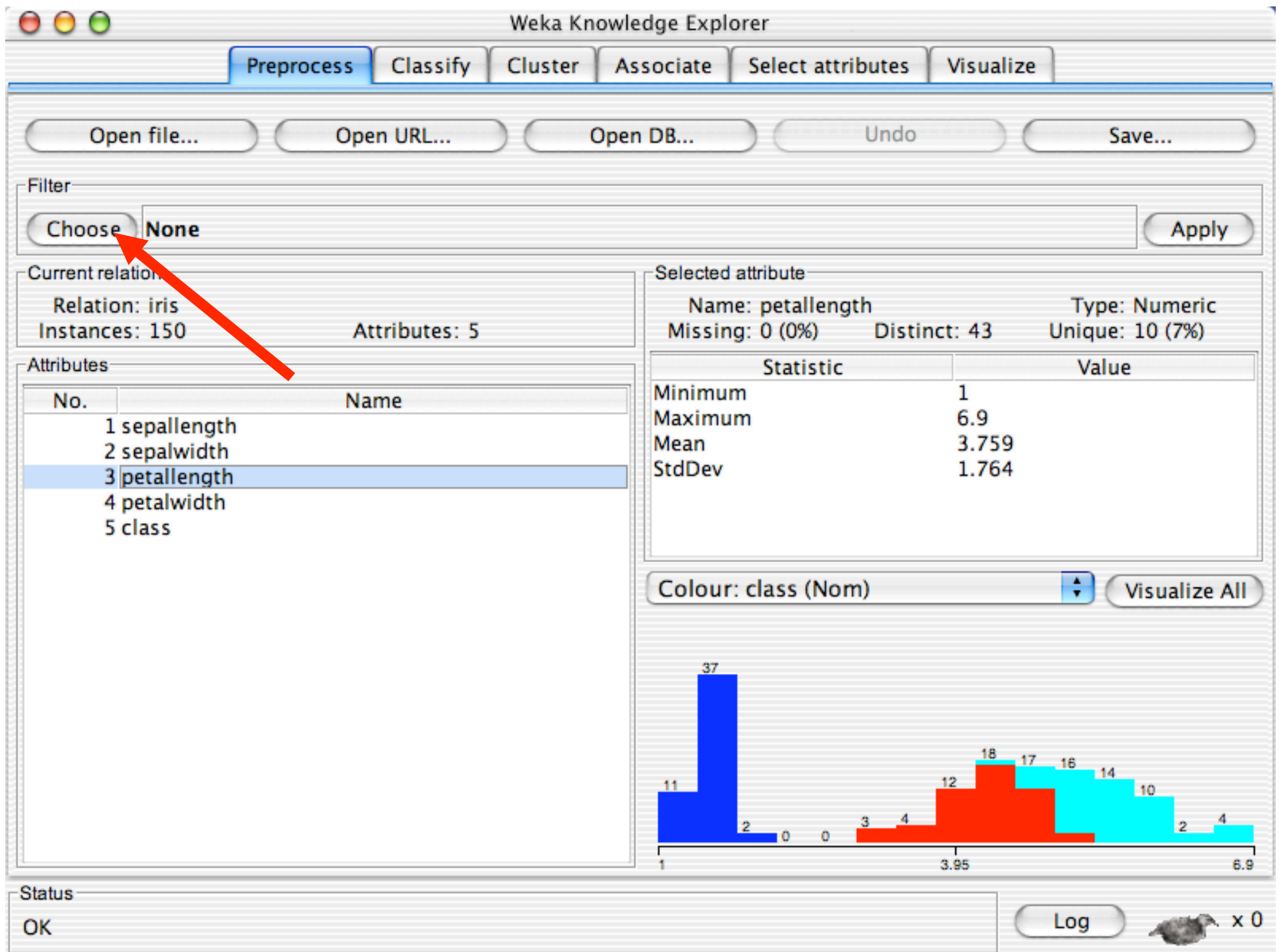


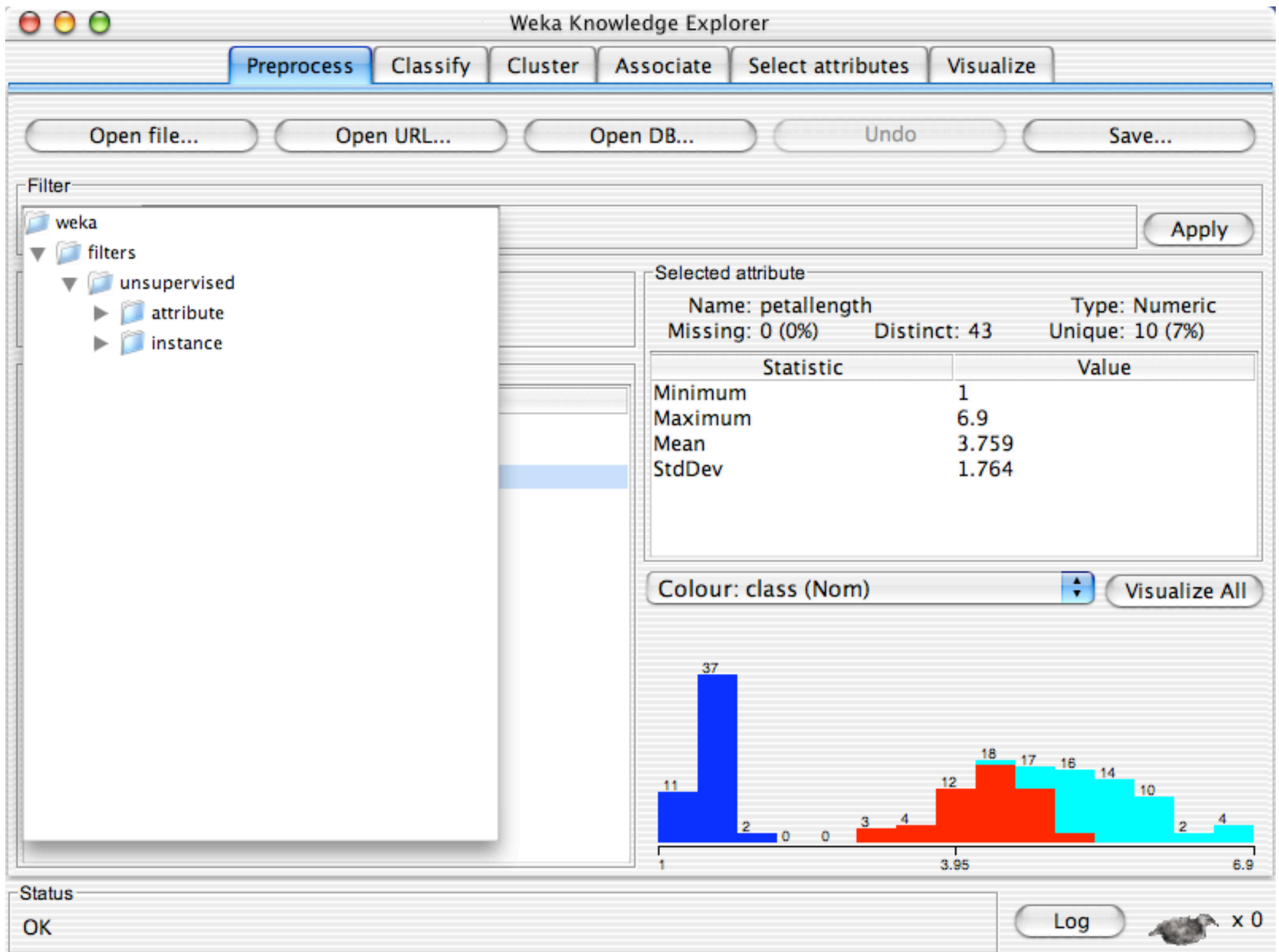


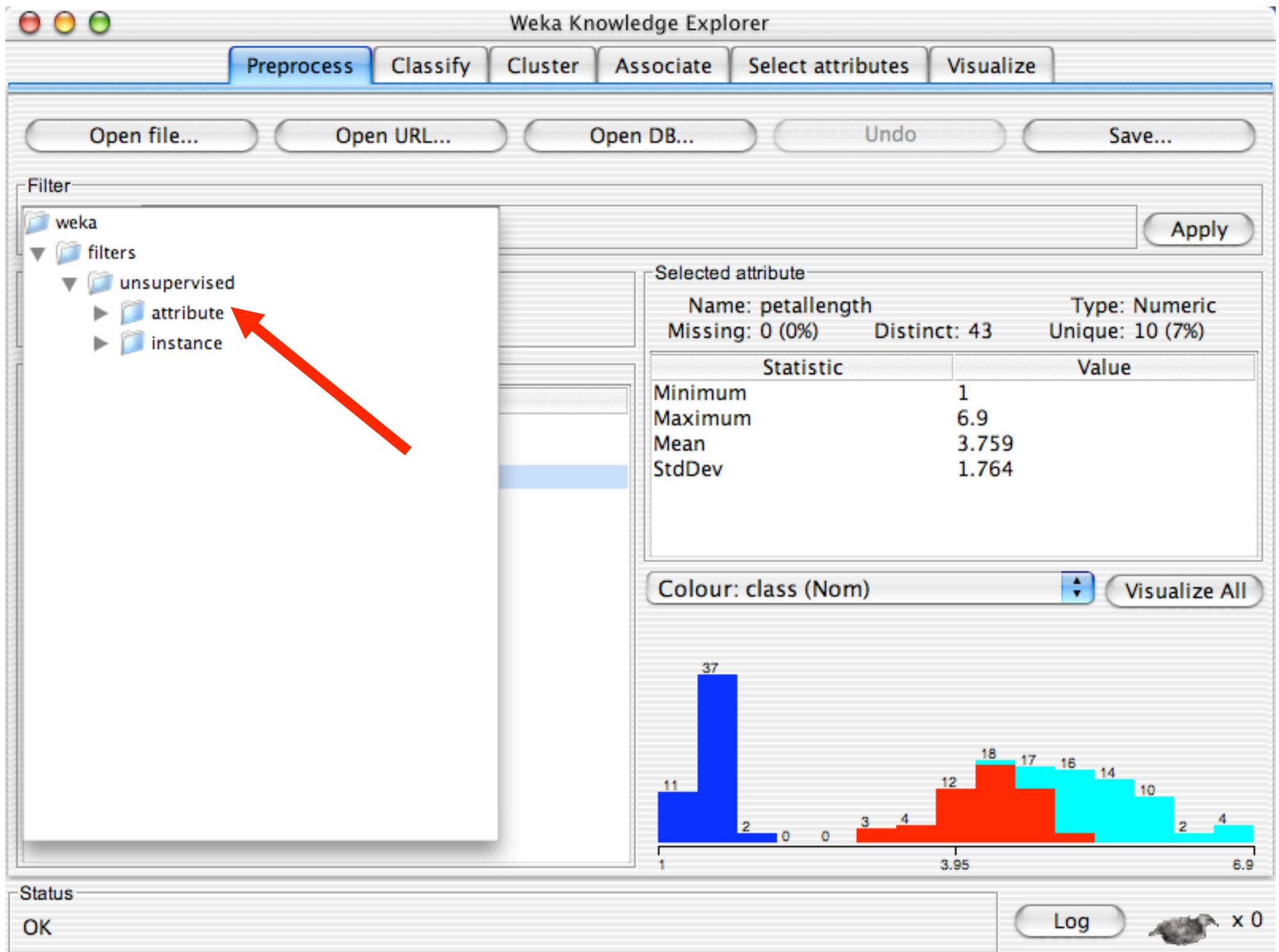


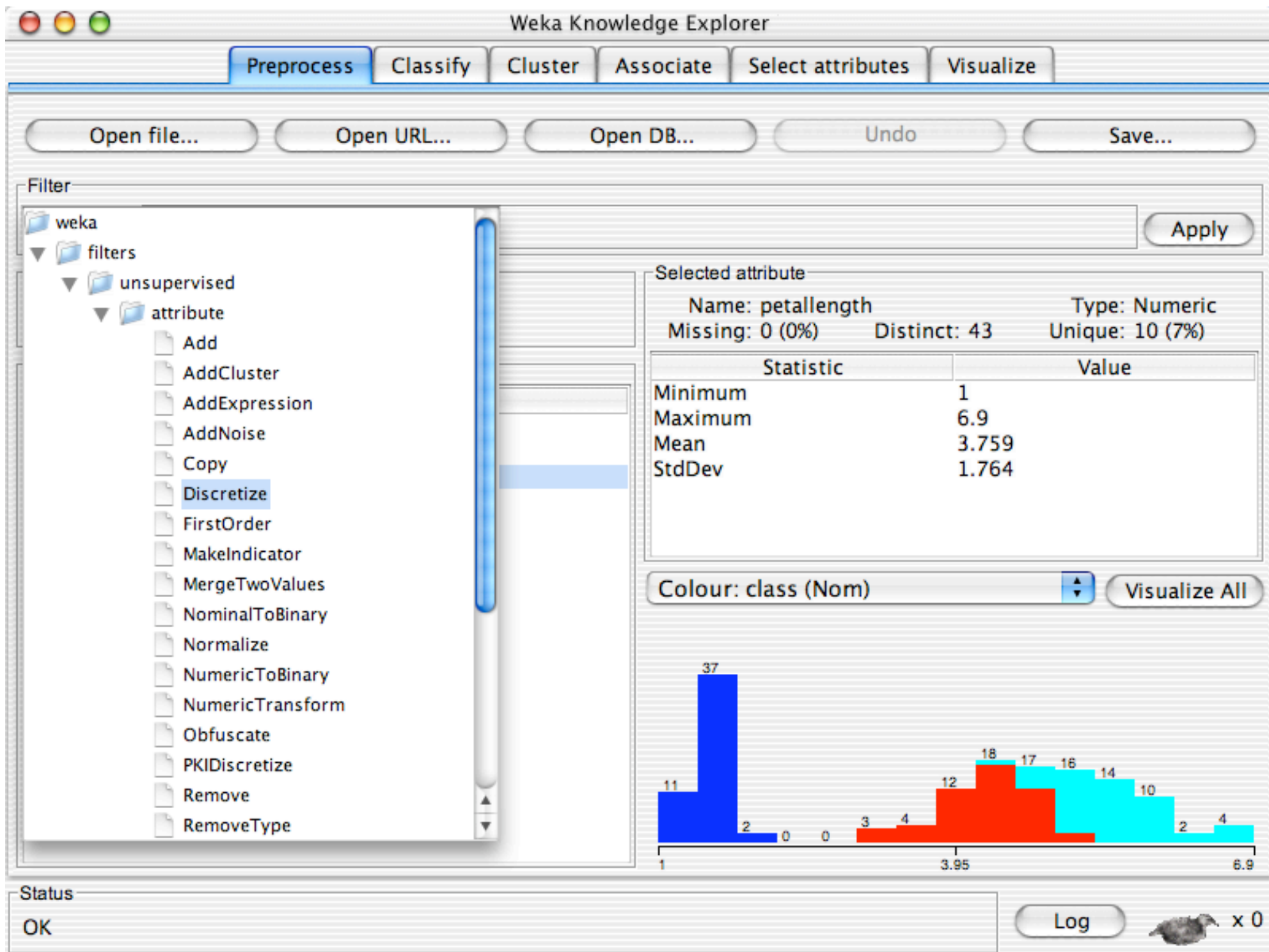


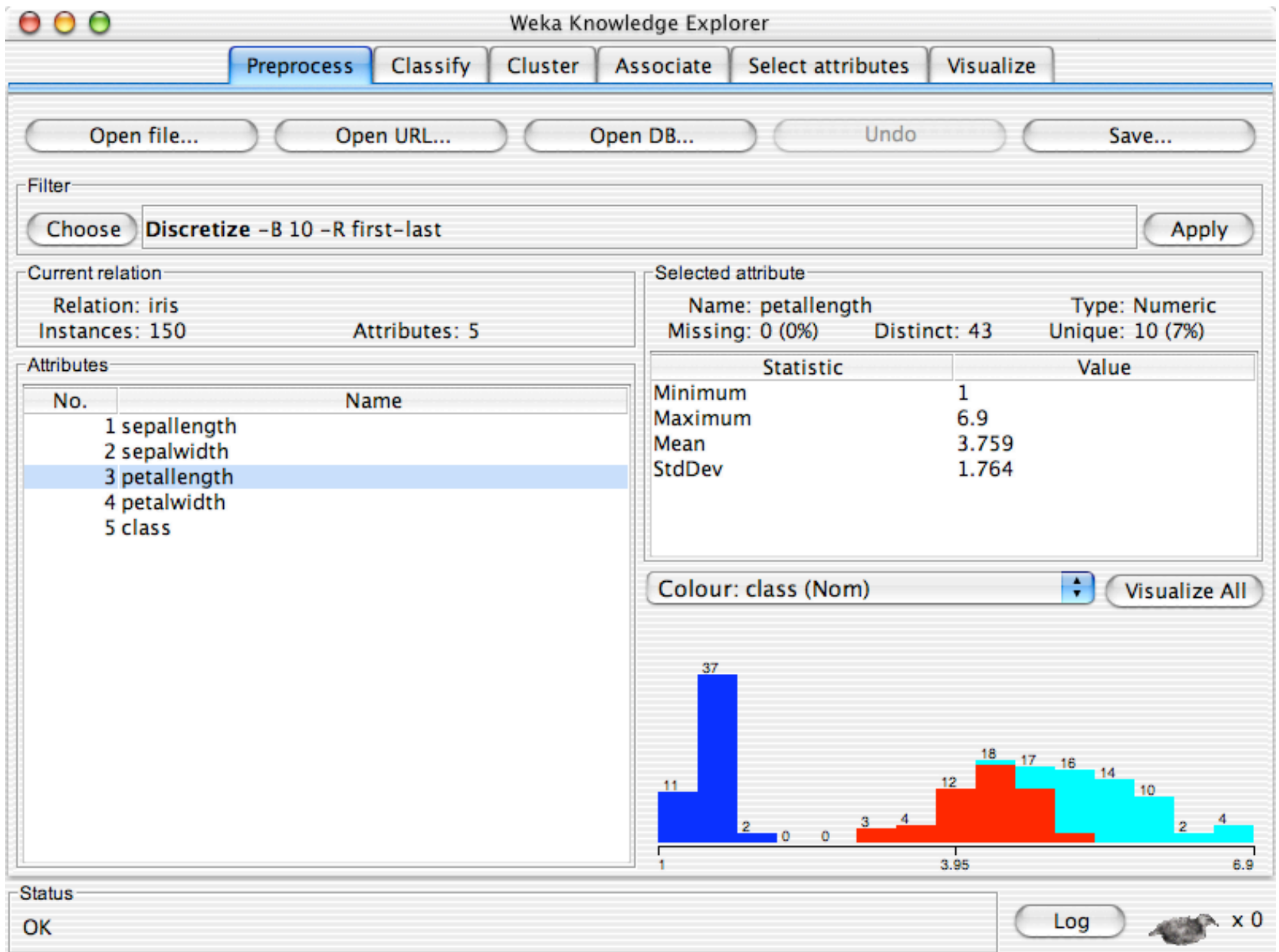


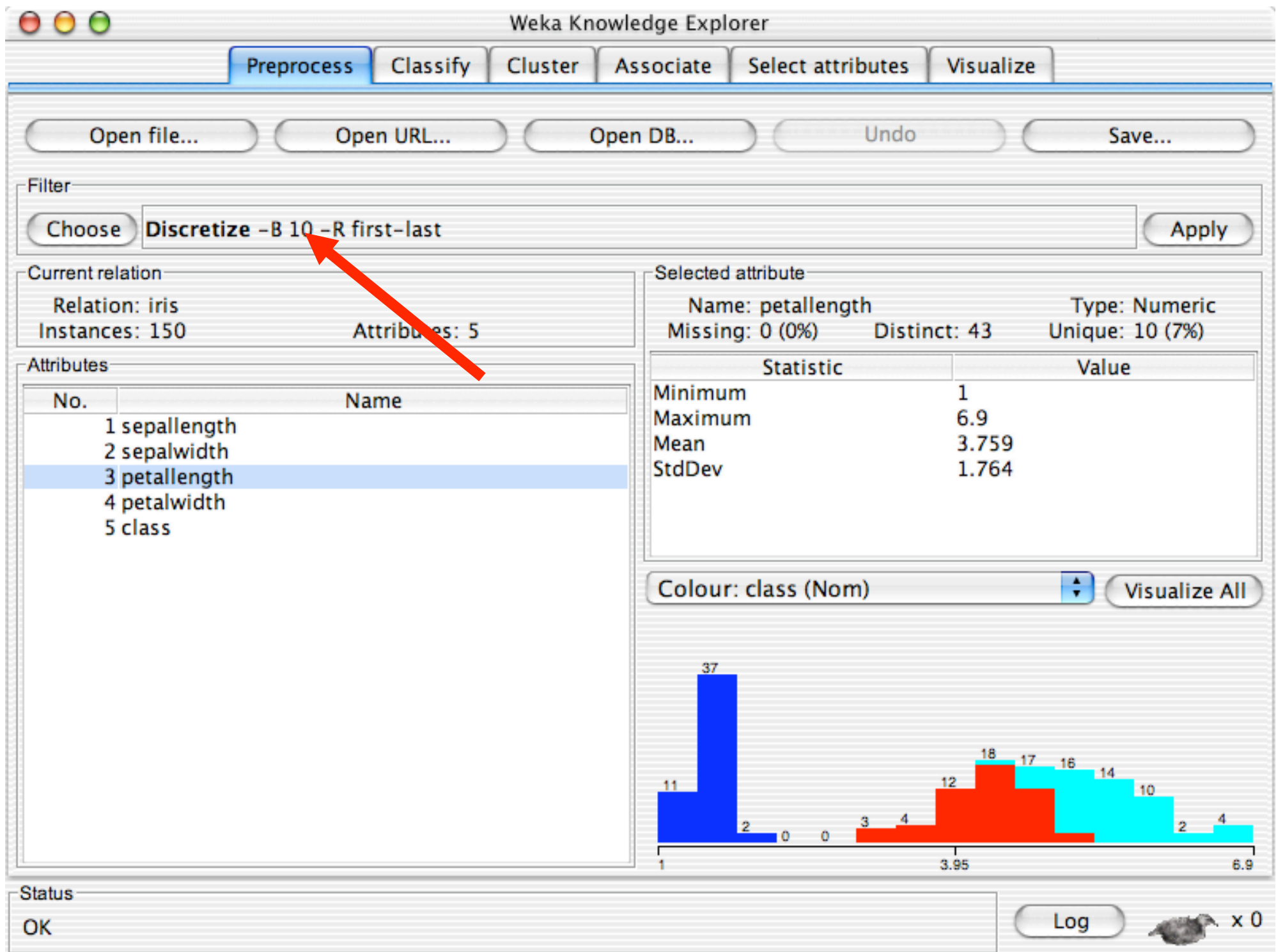












Weka Knowledge Explorer

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Open file...

Open URL...

Open DB...

Undo

Save...

Filter

Choose

Discretize -B 10 -R first-last

Apply

Current relation

Relation: iris

Instances: 150

Attributes:

Attributes

No.	Name
1	sepalength
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.

More

attributeIndices

first-last

bins

10

findNumBins

False

invertSelection

False

makeBinary

False

useEqualFrequency

False

Open...

Save...

OK

Cancel

Visualize All

Frequency distribution of petallength attribute:

Bin Range	Frequency
1.0 - 1.5	11
1.5 - 2.0	2
2.0 - 2.5	0
2.5 - 3.0	0
3.0 - 3.5	3
3.5 - 4.0	4
4.0 - 4.5	12
4.5 - 5.0	10
5.0 - 5.5	2
5.5 - 6.0	4

Status

OK

Log

x 0

Weka Knowledge Explorer

PreprocessClassifyClusterAssociateSelect attributesVisualize

Open file...Open URL...Open DB...UndoSave...

Filter

ChooseDiscretize -B 10 -R first-last

Current relation

Relation: iris

Instances: 150

Attributes:

Attributes

No.	Name
1	sepal.length
2	sepal.width
3	petal.length
4	petal.width
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.

More

attributeIndicesfirst-last

bins10

findNumBinsFalse

invertSelectionFalse

makeBinaryFalse

useEqualFrequencyFalse

Open...Save...OKCancel

Visualize All

11

2

0

0

3

4

12

10

2

4

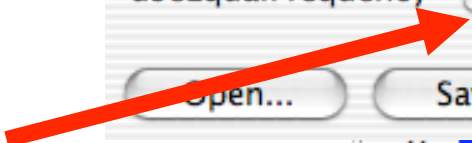
13.956.9

Status

OK

Log

x 0



Weka Knowledge Explorer

PreprocessClassifyClusterAssociateSelect attributesVisualize

Open file...Open URL...Open DB...UndoSave...

Filter

ChooseDiscretize -B 10 -R first-last

Current relation

Relation: iris

Instances: 150

Attributes:

Attributes

No.	Name
1	sepal.length
2	sepal.width
3	petal.length
4	petal.width
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.

More

attributeIndicesfirst-last

bins10

findNumBinsFalse

invertSelectionFalse

makeBinaryFalse

useEqualFrequencyTrue

Open...Save...OKCancel

Visualize All

11

2

0

0

3

4

12

10

2

4

13.956.9

Status

OK

Log

x 0

Weka Knowledge Explorer

PreprocessClassifyClusterAssociateSelect attributesVisualize

Open file...Open URL...Open DB...UndoSave...

Filter

ChooseDiscretize -B 10 -R first-last

Current relation

Relation: iris

Instances: 150

Attributes:

Attributes

No.	Name
1	sepal.length
2	sepal.width
3	petal.length
4	petal.width
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.

More

attributeIndicesfirst-last

bins10

findNumBinsFalse

invertSelectionFalse

makeBinaryFalse

useEqualFrequencyTrue

Open...Save...OKCancel

Visualize All

11

2

0

0

3

4

12

10

2

4

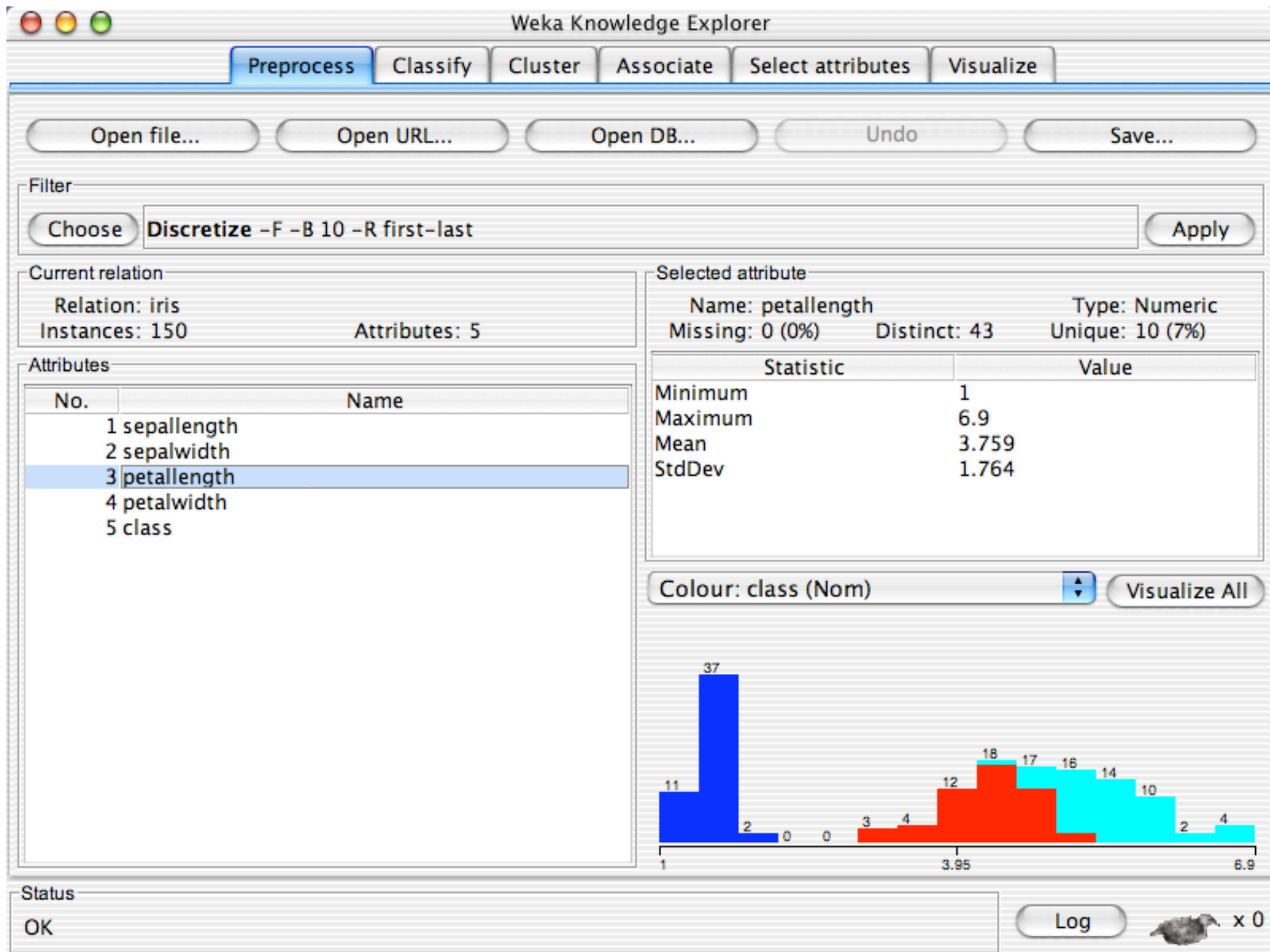
13.956.9

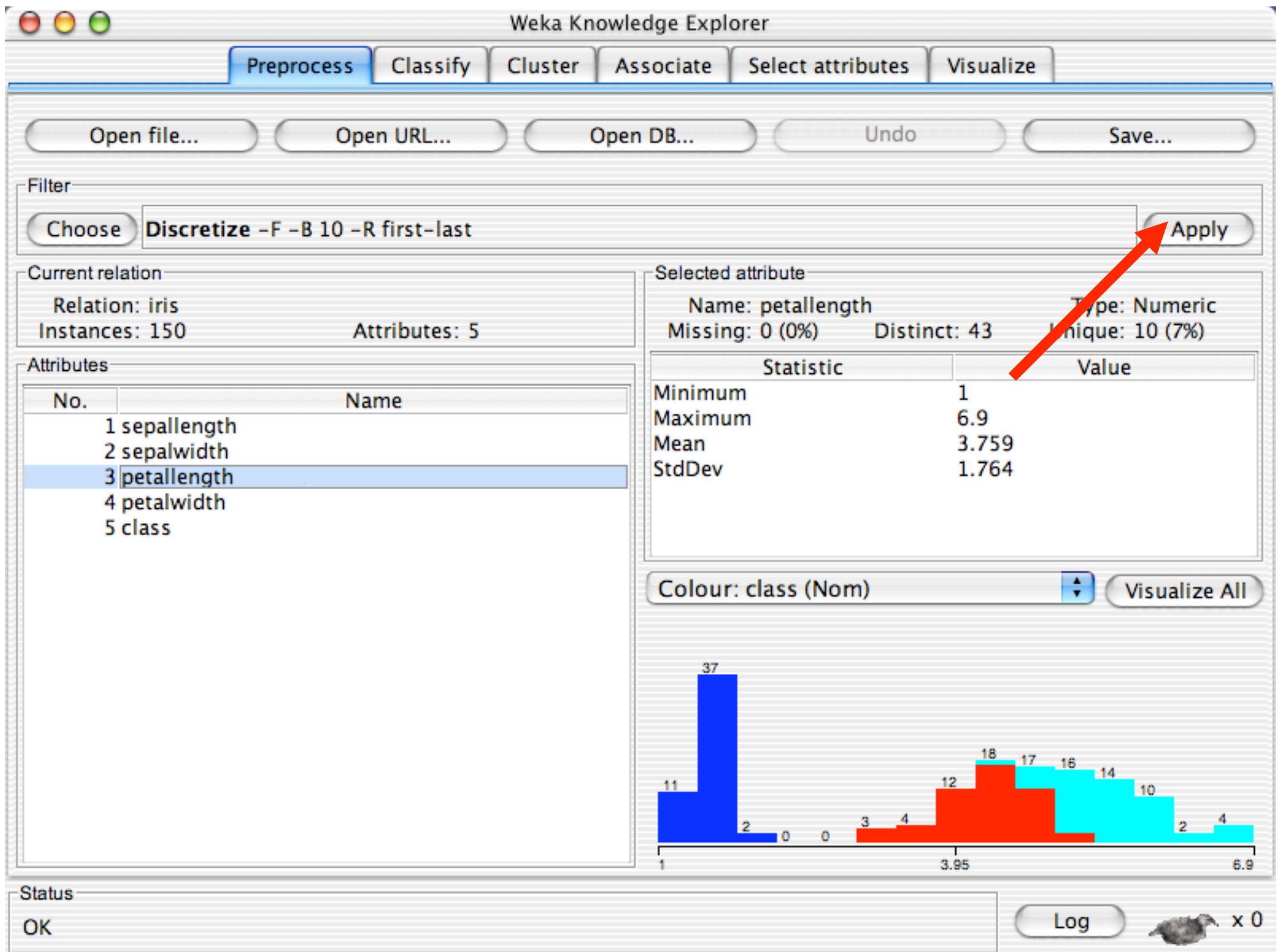
Status

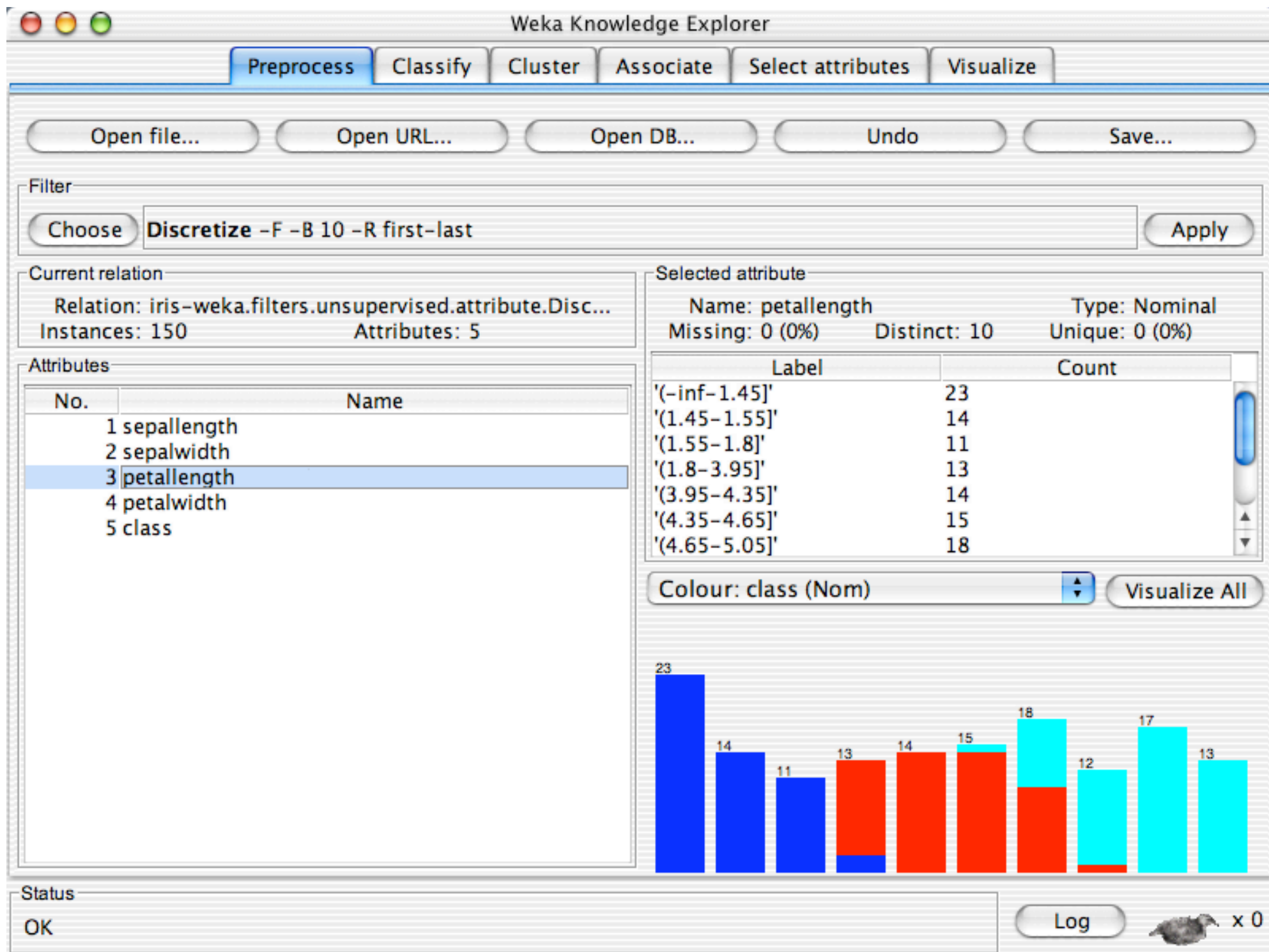
OK

Log

x 0



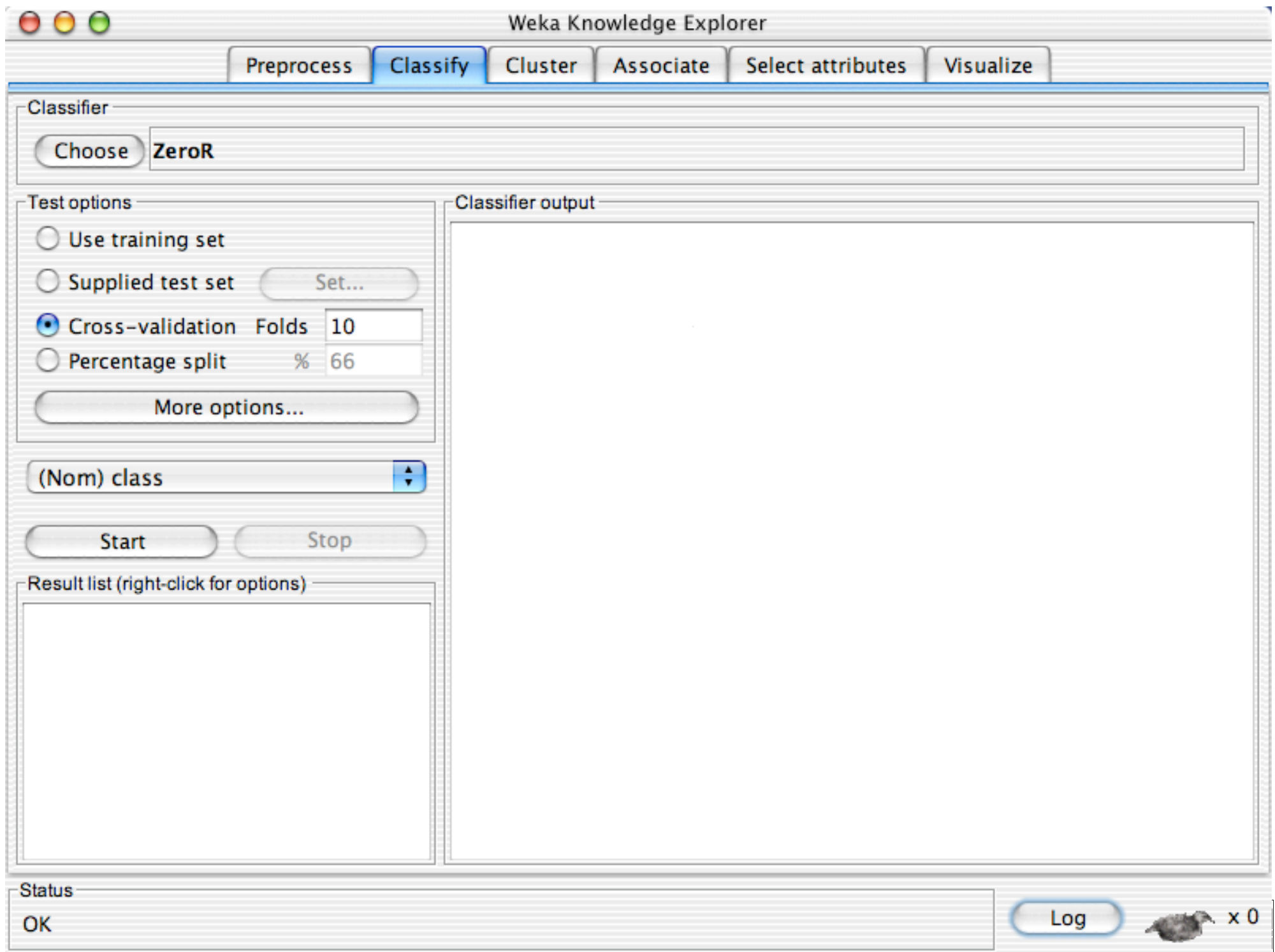


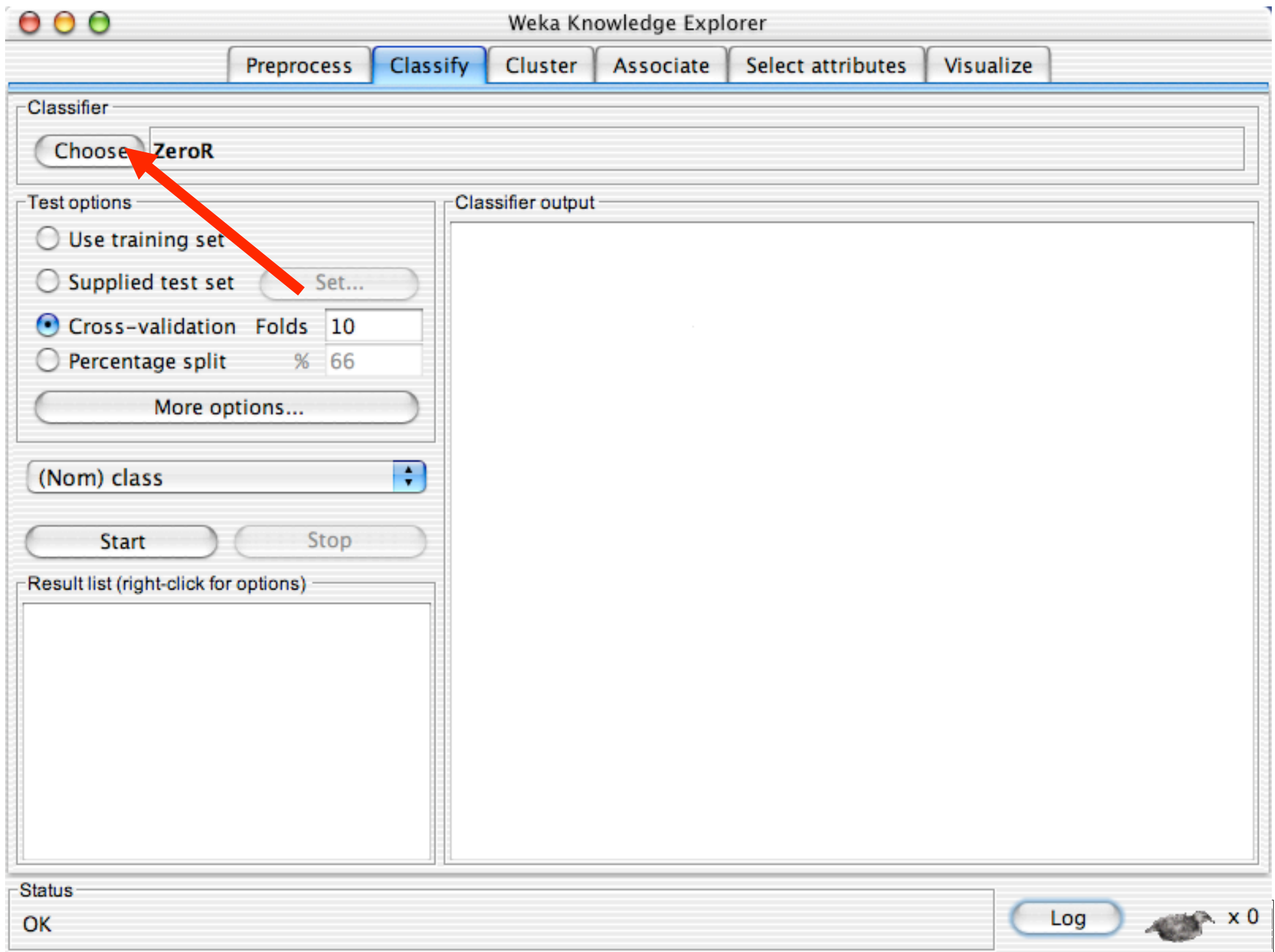


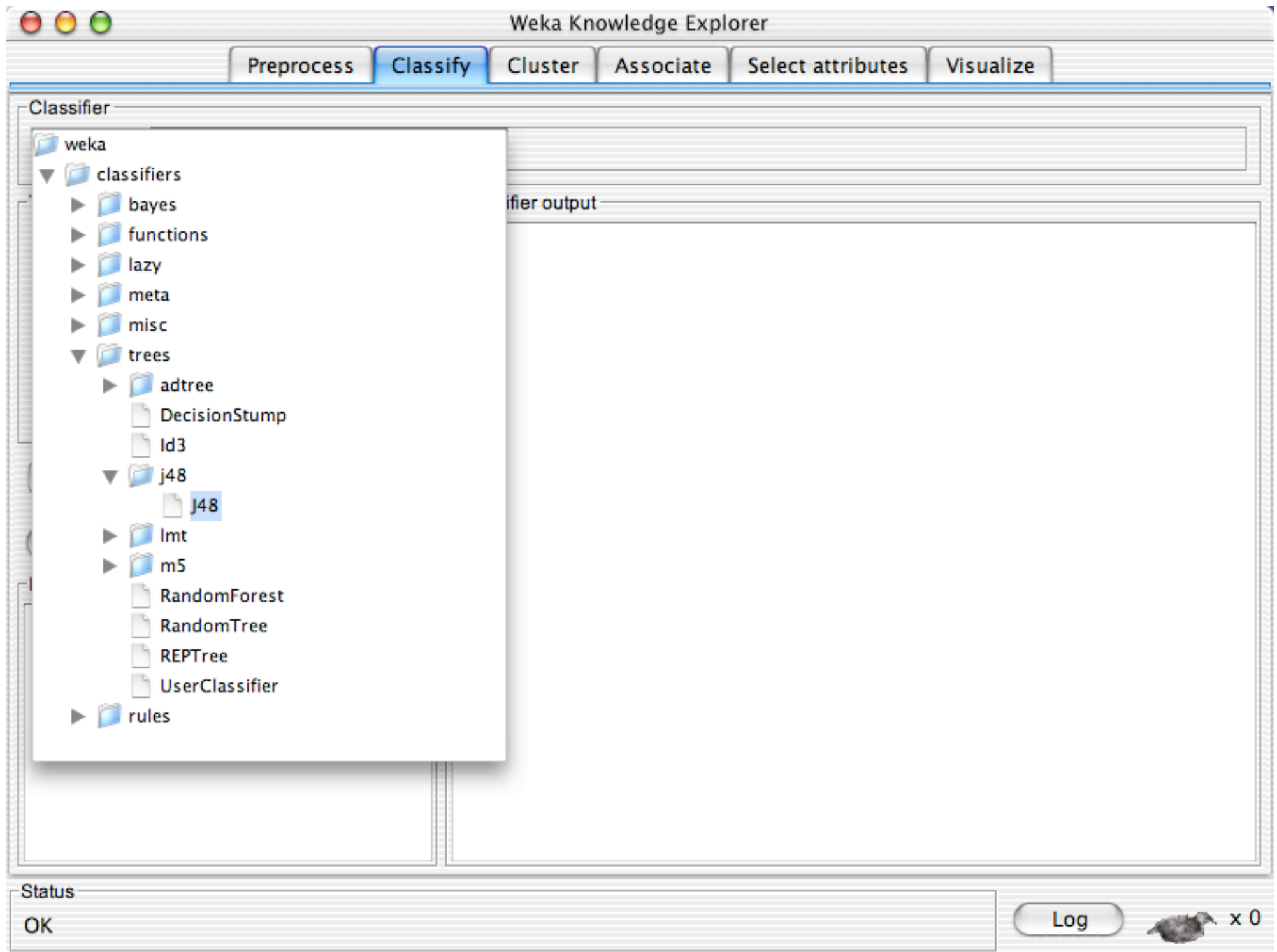
Explorer: Building “Classifiers”

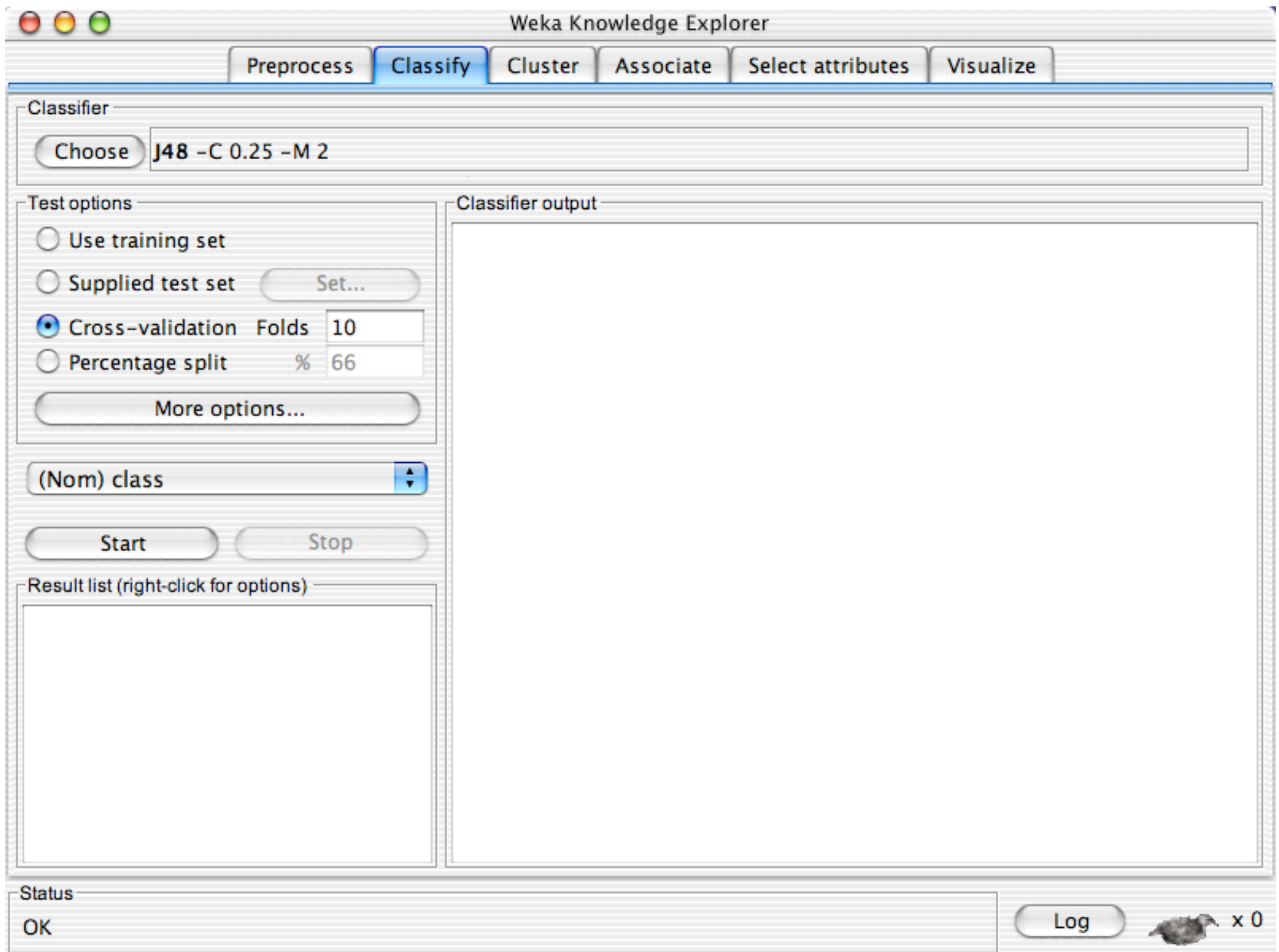


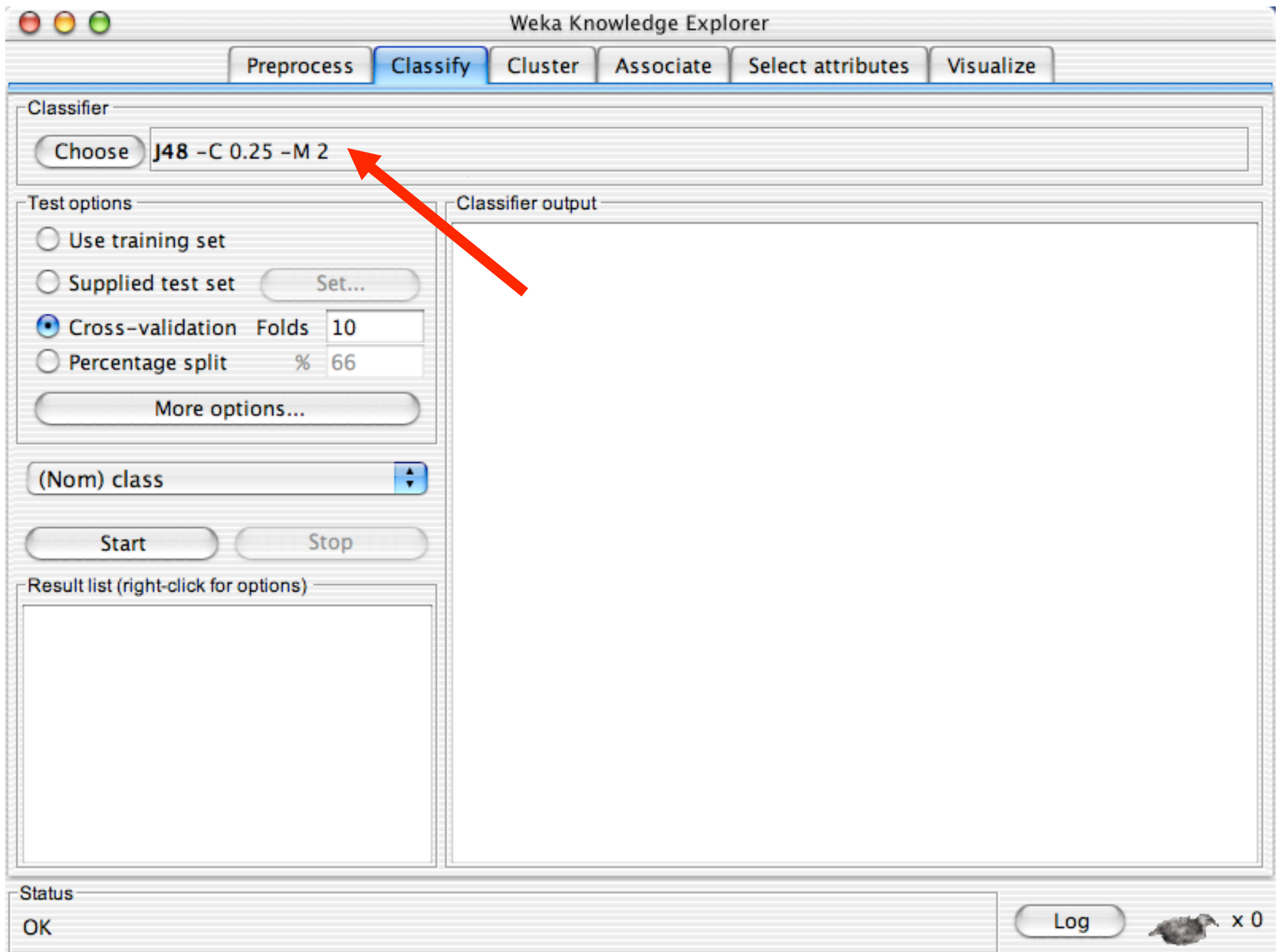
- Classifiers in WEKA are models for predicting nominal or numeric quantities
- Implemented learning schemes include:
 - Decision trees and lists, instance-based classifiers, support vector machines, multi-layer perceptrons, logistic regression, Bayes' nets, ...
- “Meta”-classifiers include:
 - Bagging, boosting, stacking, error-correcting output codes, locally weighted learning, ...

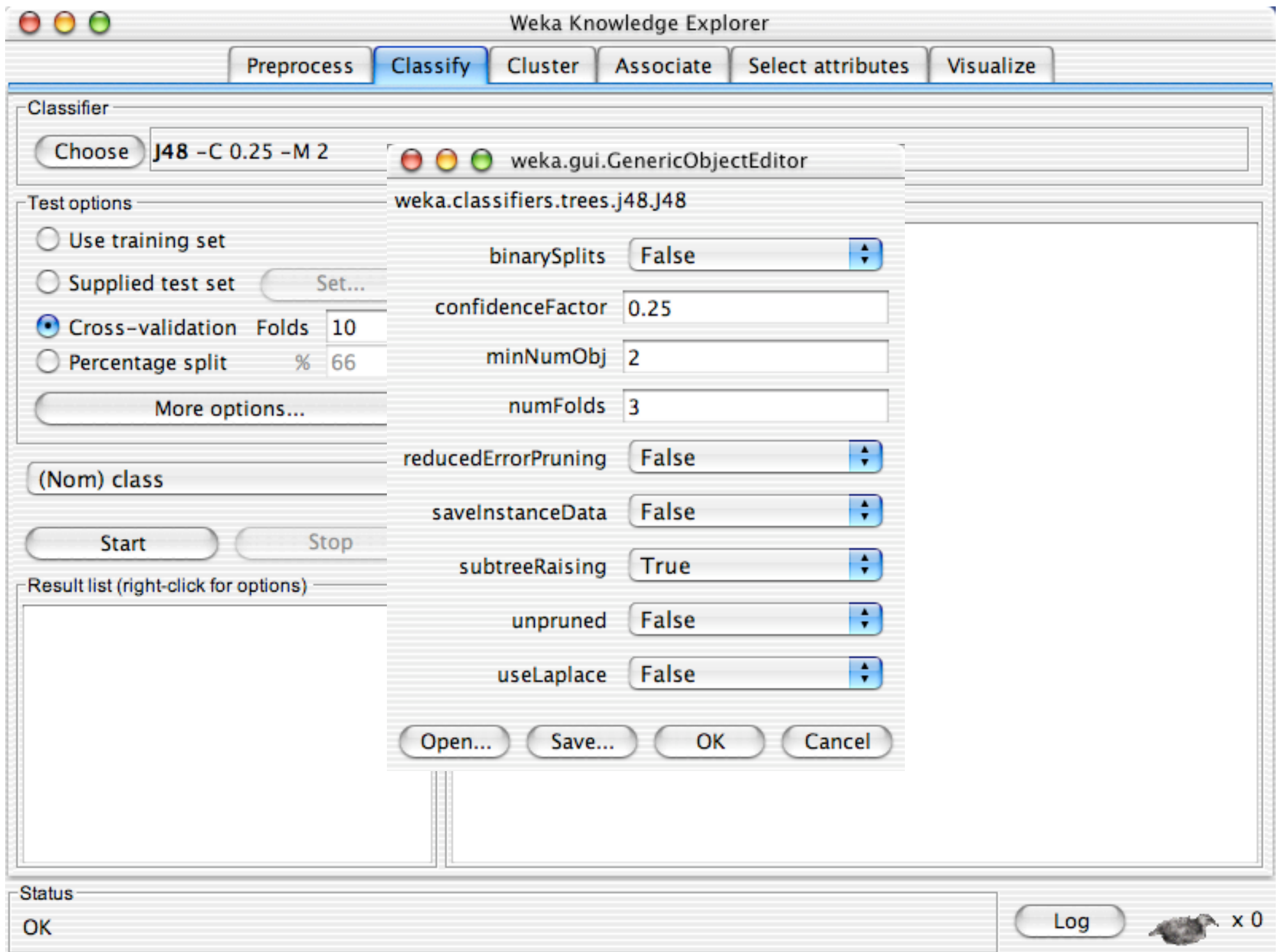


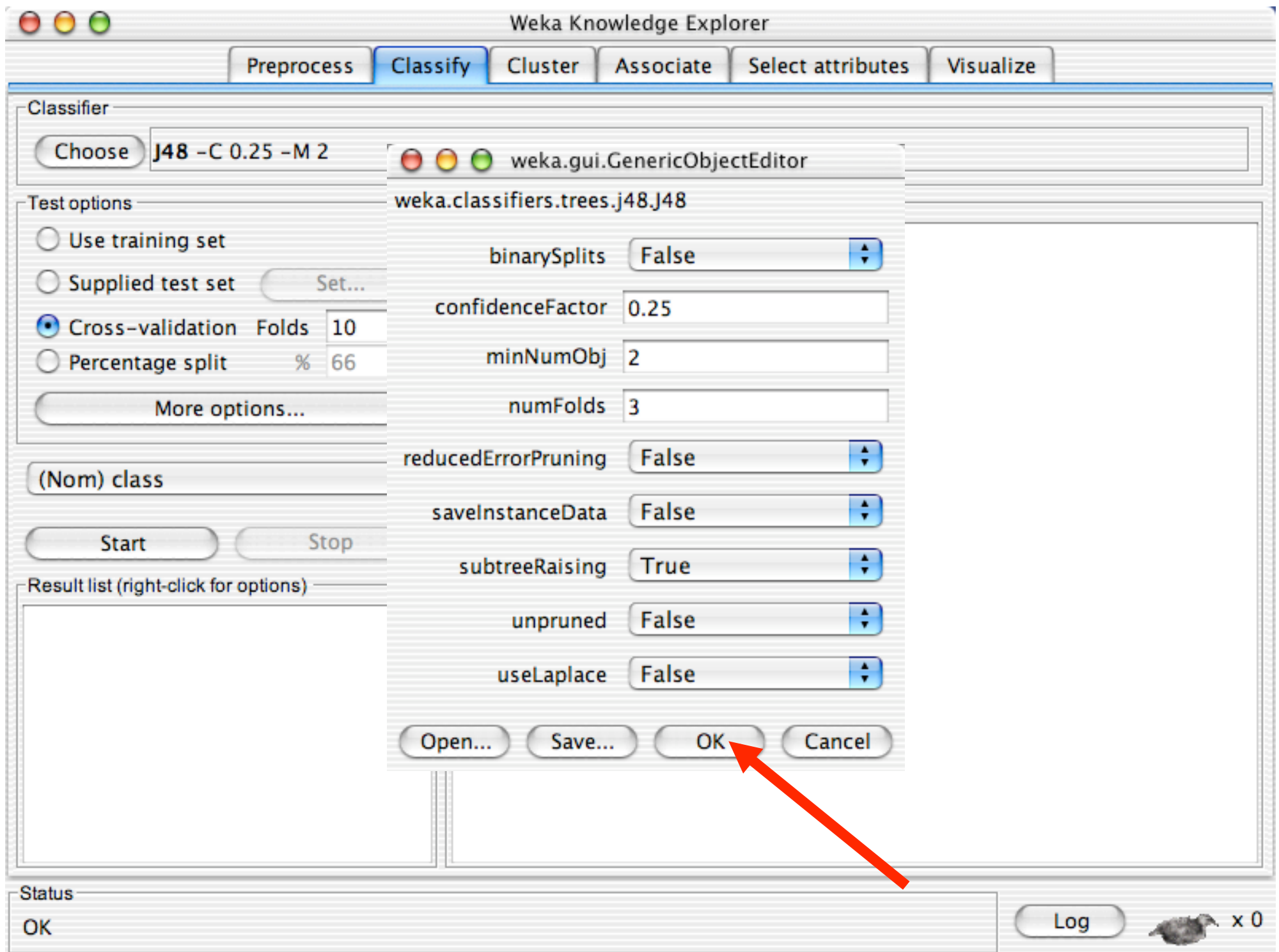


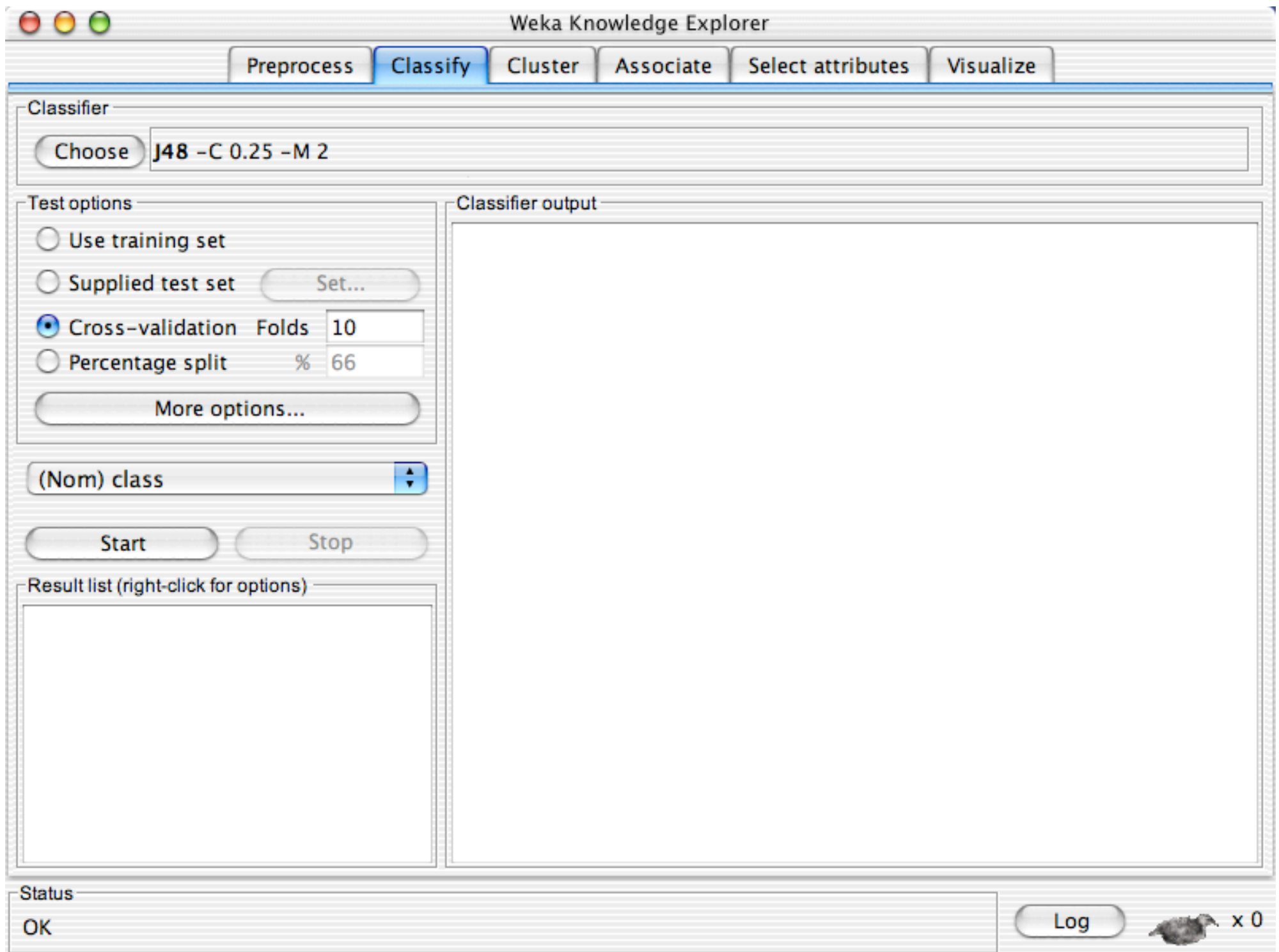


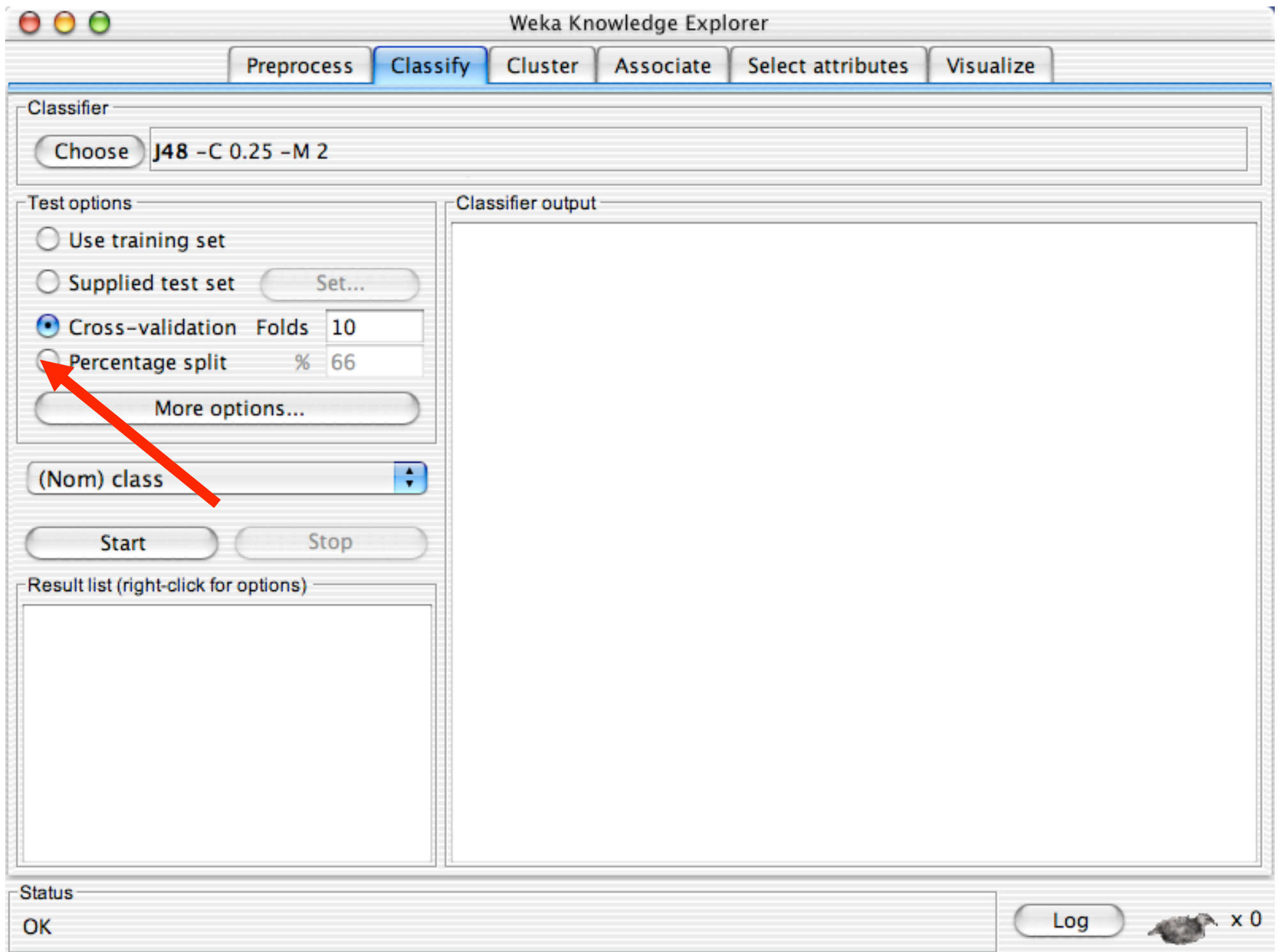


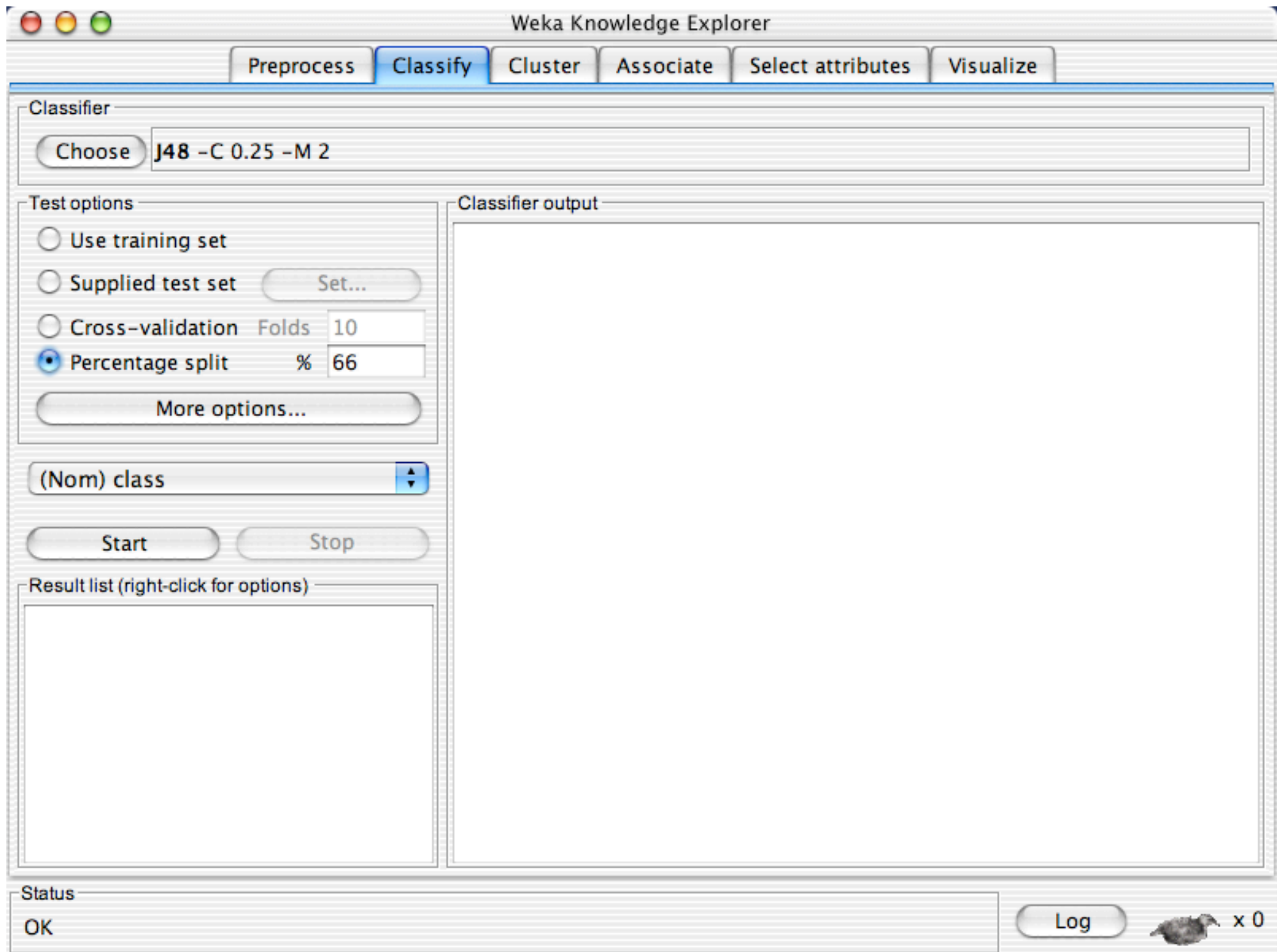


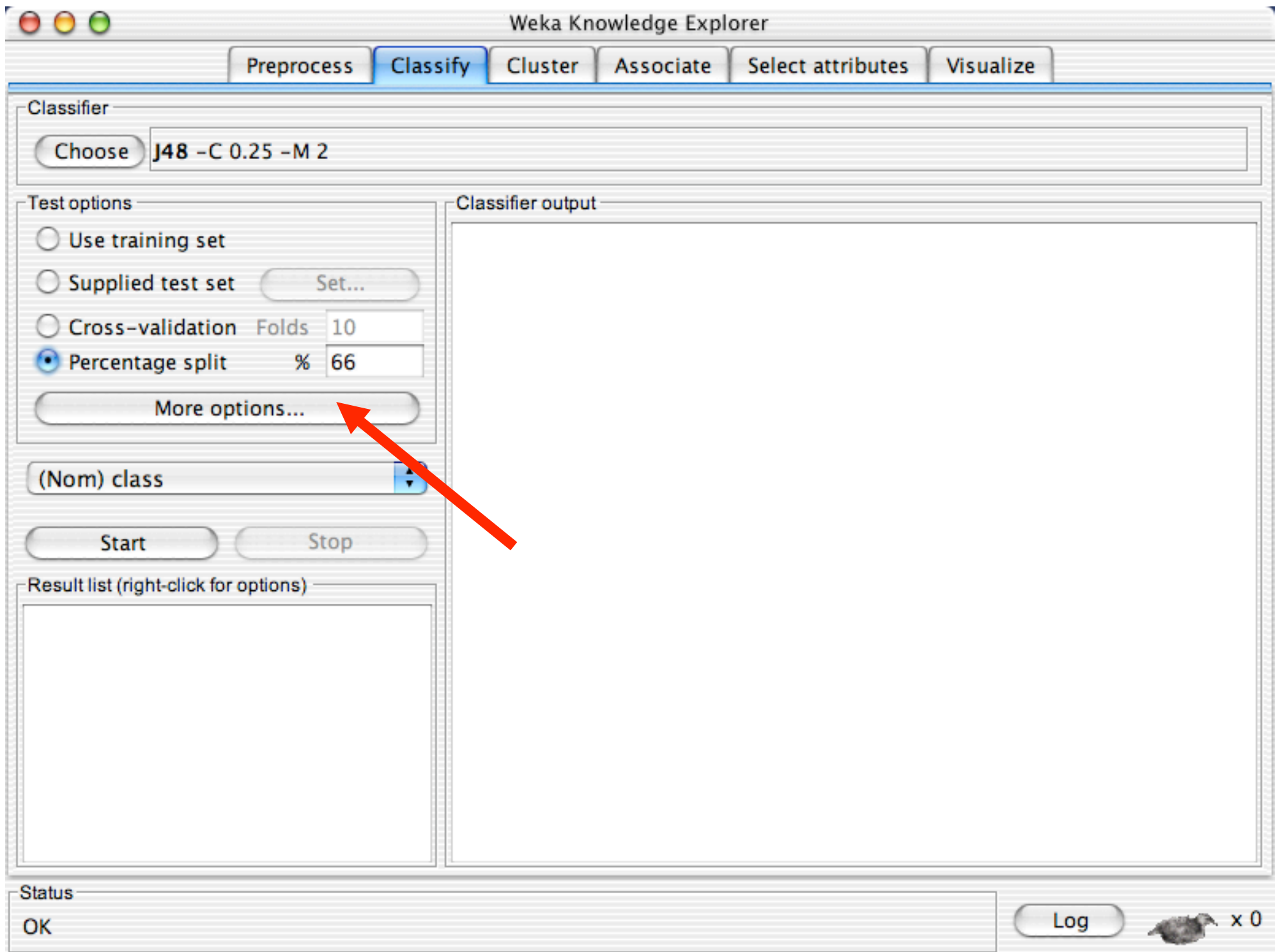


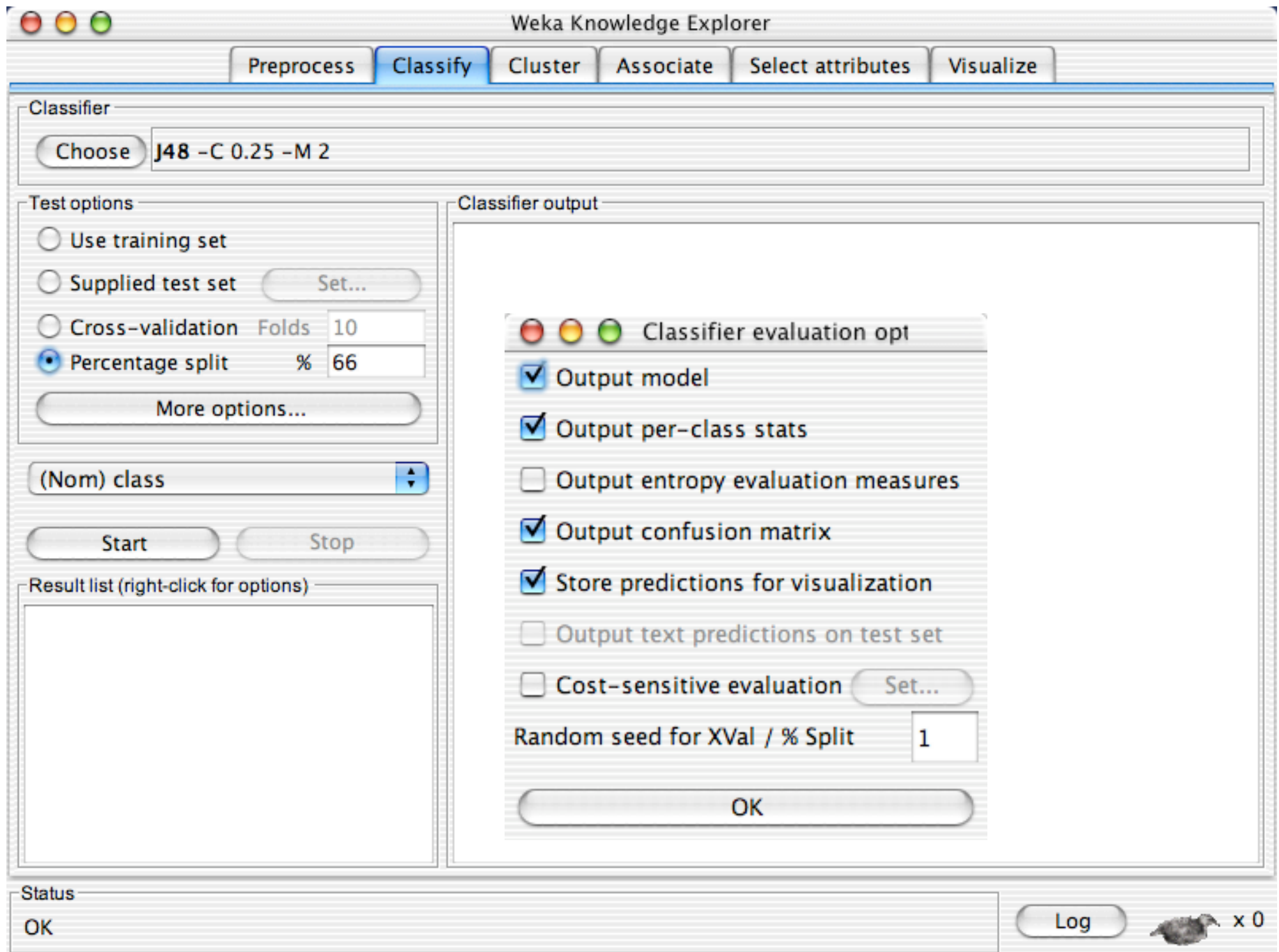


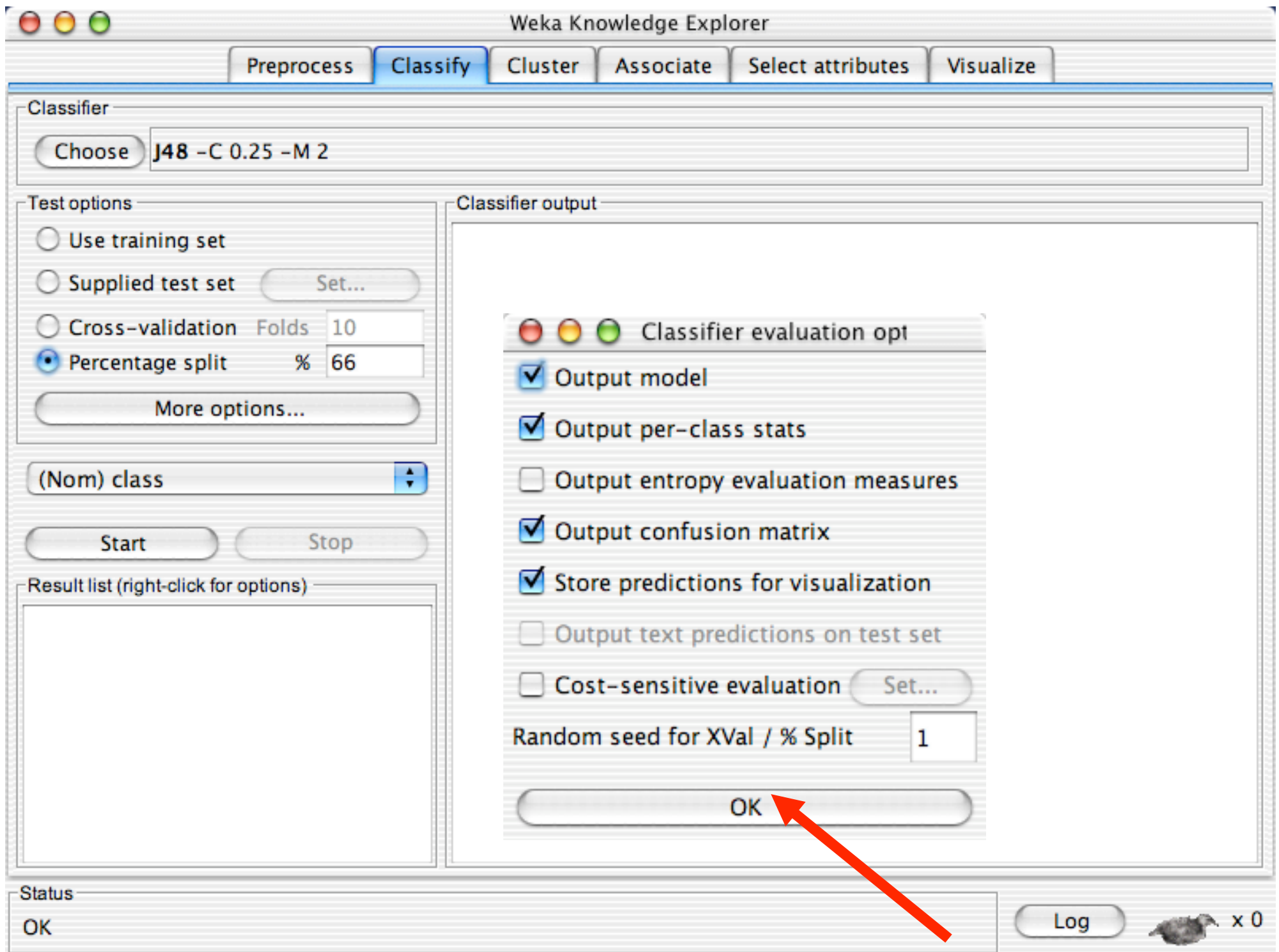


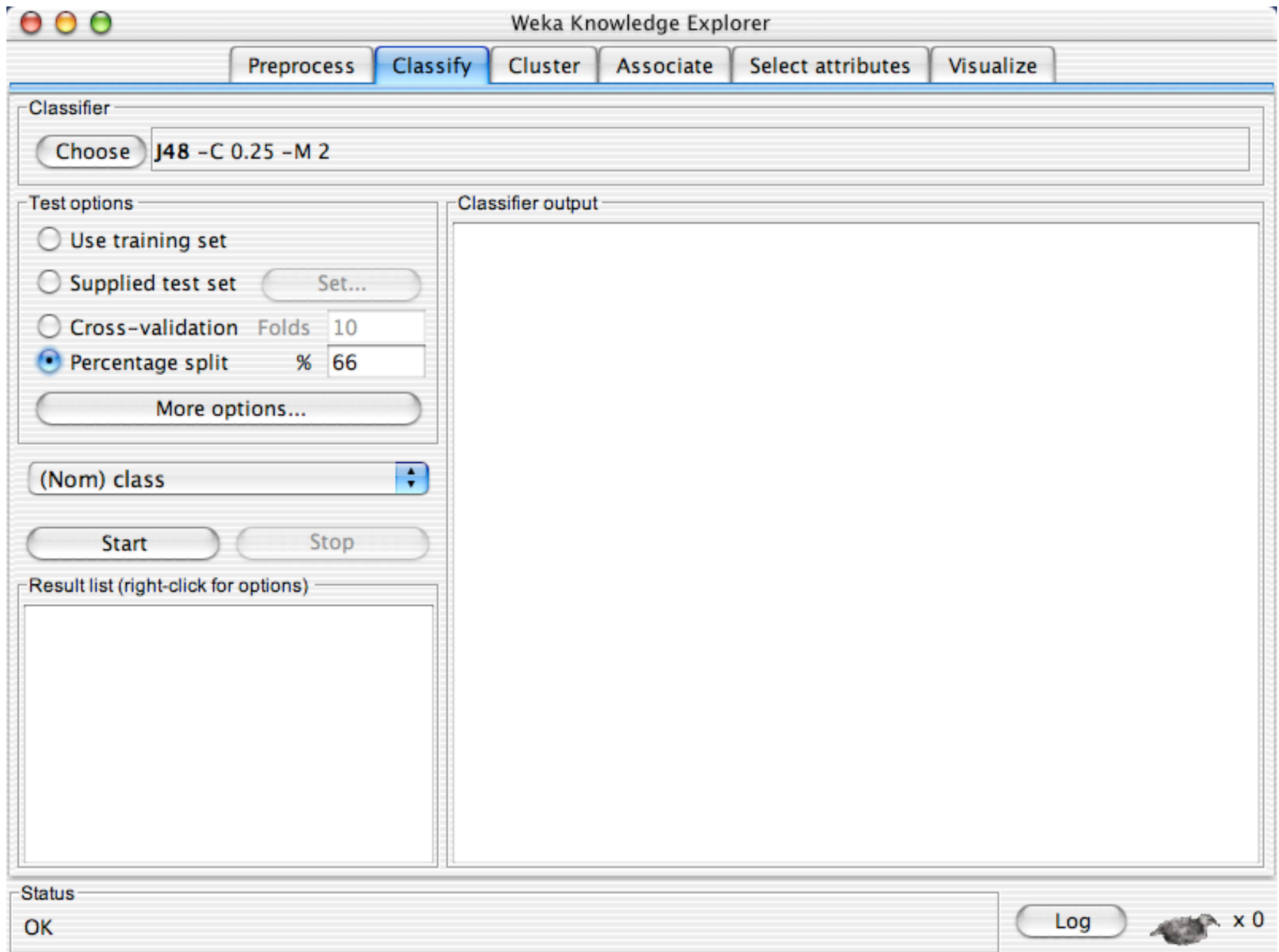


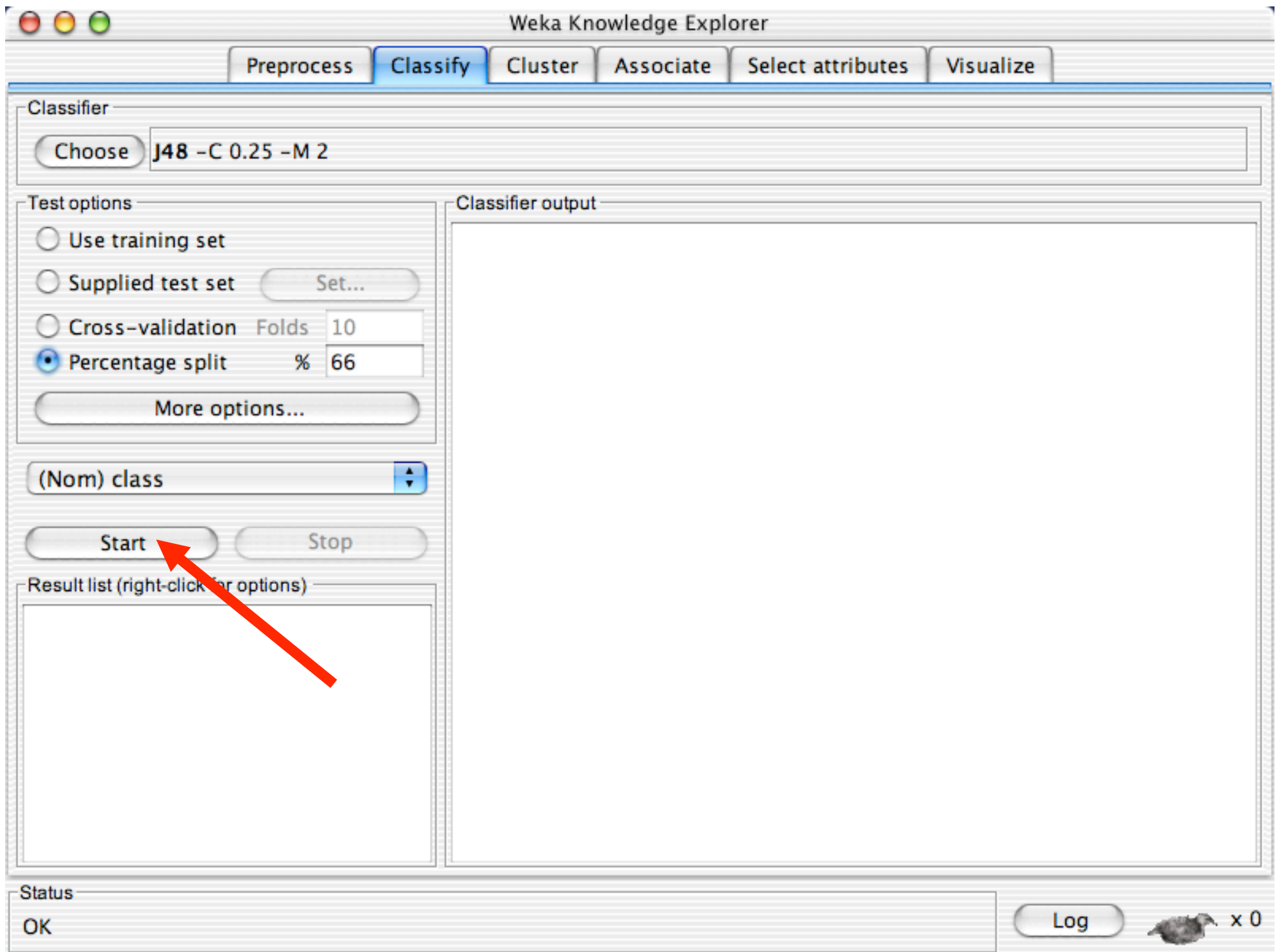


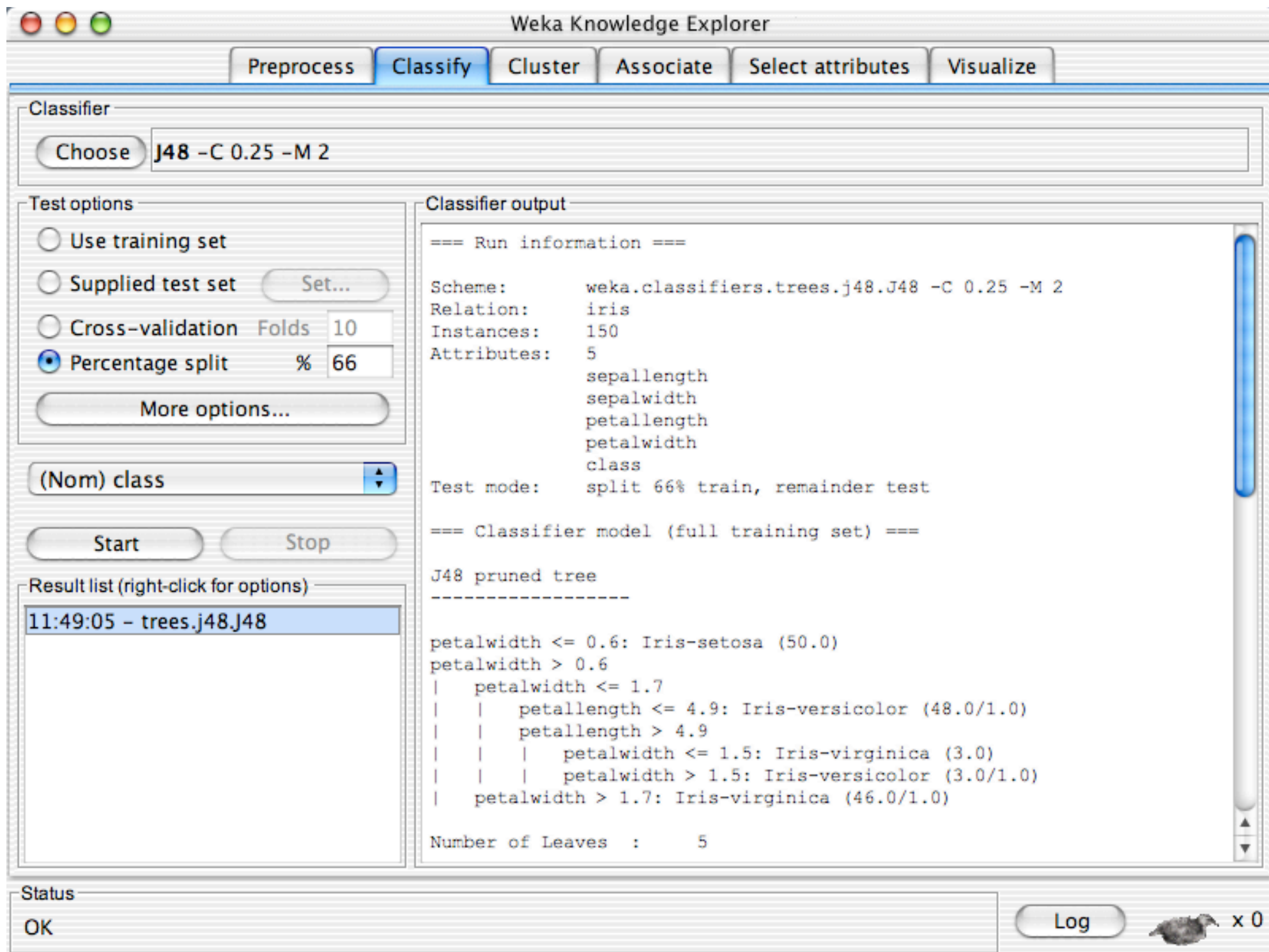


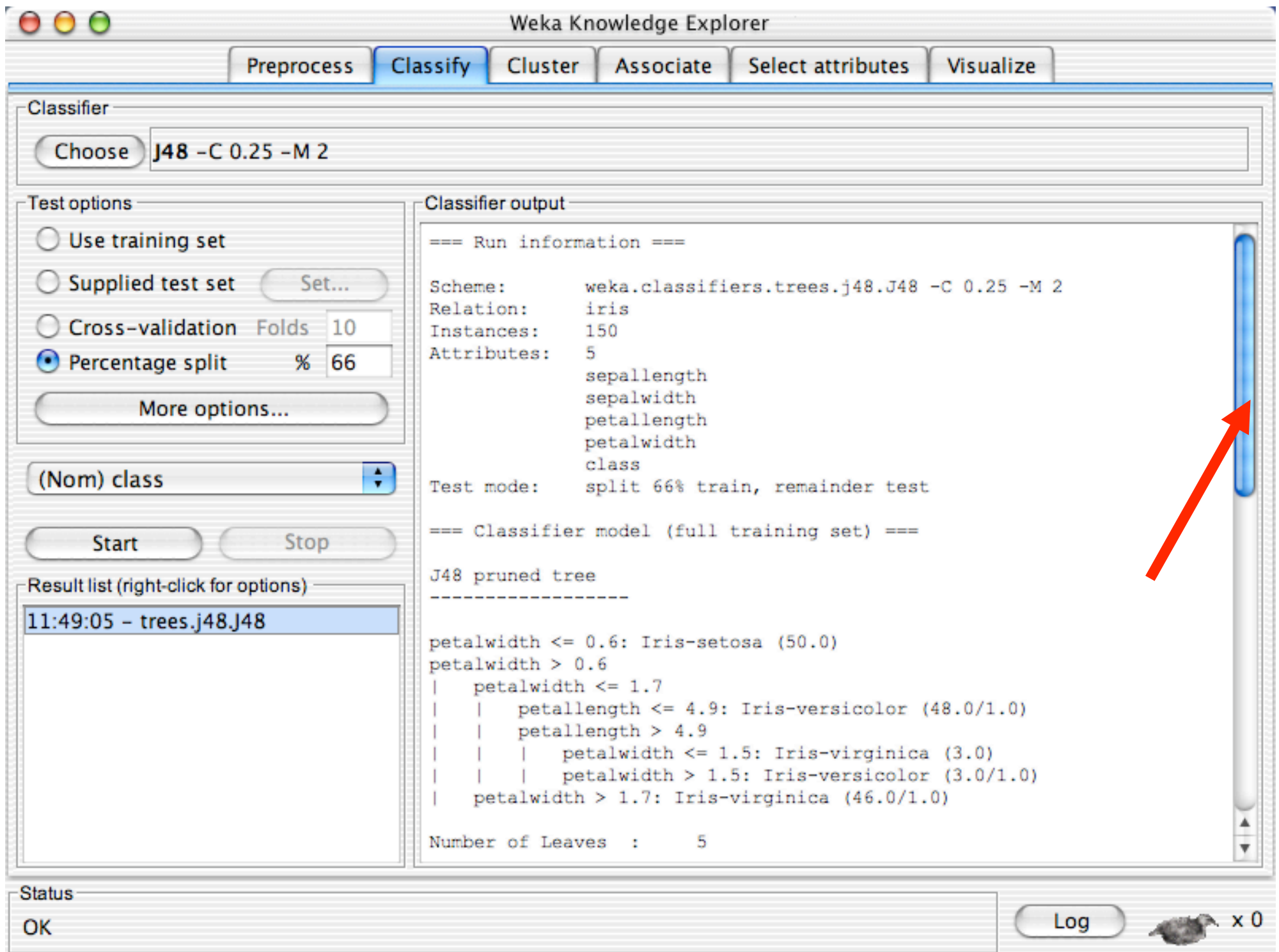













Weka Knowledge Explorer

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)

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

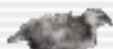
0 19 0 | b = Iris-versicolor

0 2 15 | c = Iris-virginica

Status

OK

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)

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

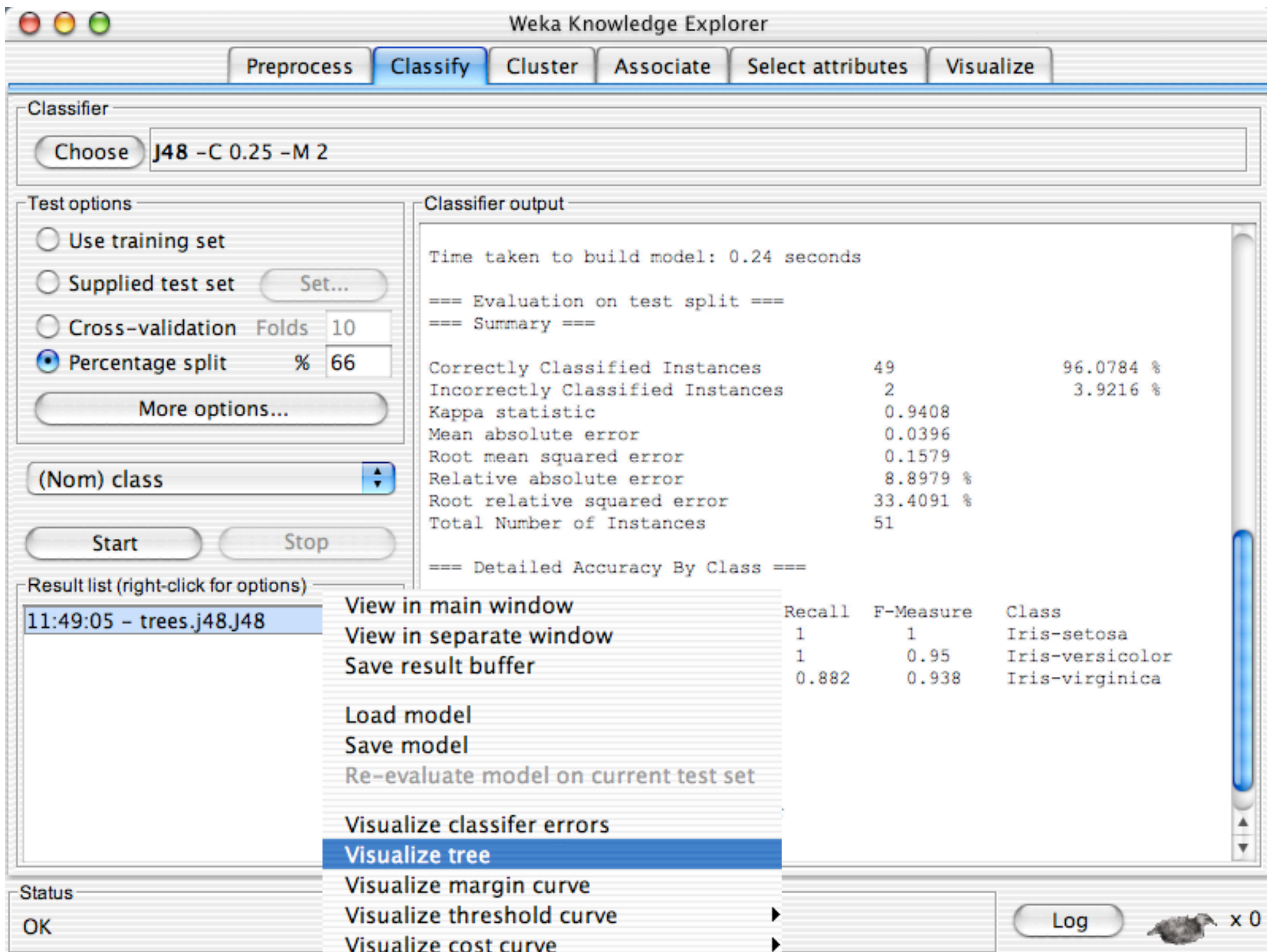
=== Confusion Matrix ===

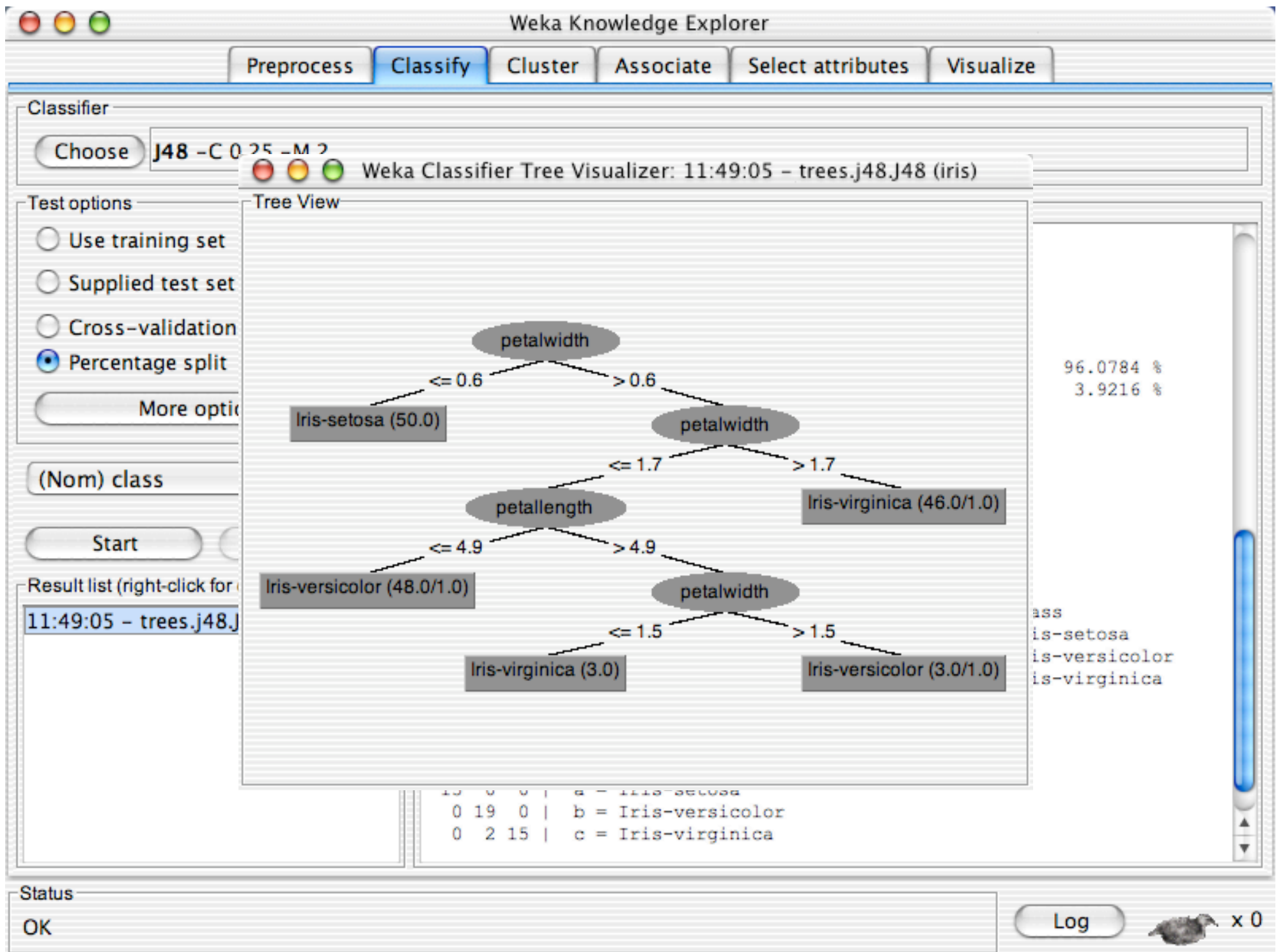
a	b	c	<-- classified as
15	0	0	a = Iris-setosa
0	19	0	b = Iris-versicolor
0	2	15	c = Iris-virginica

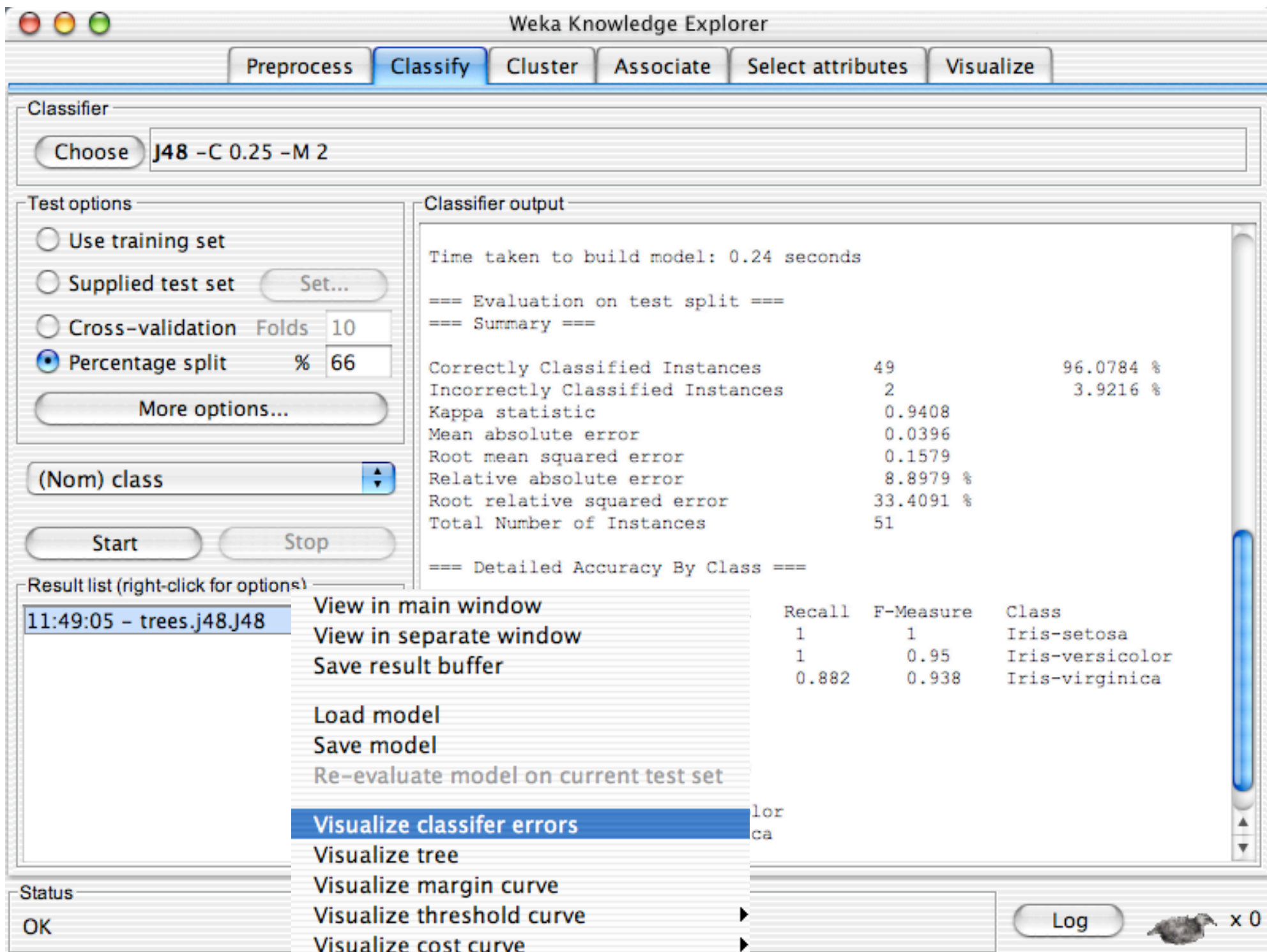
Status

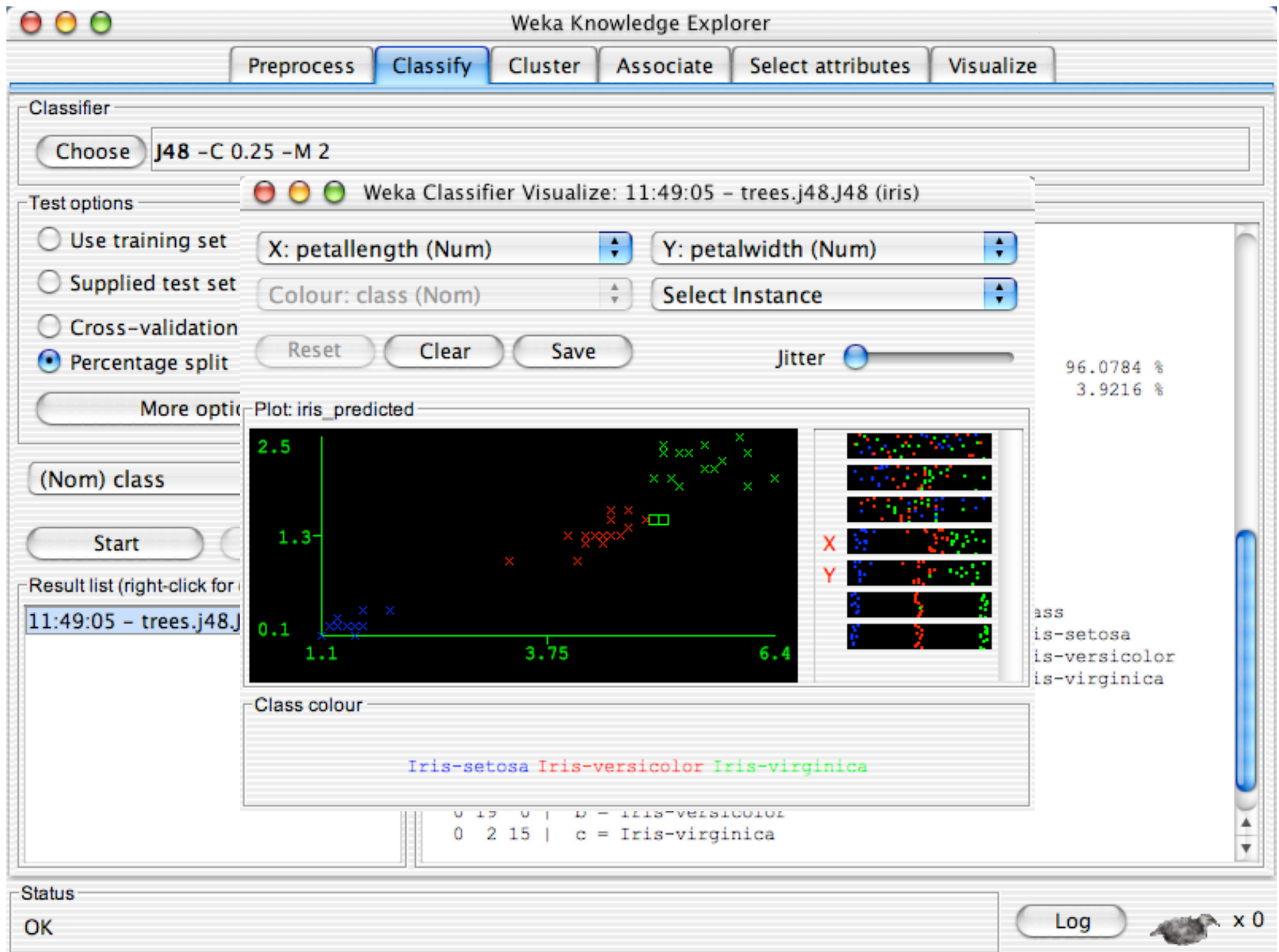
OK

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)

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

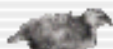
=== Confusion Matrix ===

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

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)

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

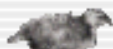
0 19 0 | b = Iris-versicolor

0 2 15 | c = Iris-virginica

Status

OK

Log

 x 0

Weka Knowledge Explorer

PreprocessClassifyClusterAssociateSelect attributesVisualize

Classifier

weka

classifiers

bayes

AODE

BayesNetK2

BayesNetB

NaiveBayes

NaiveBayesMultinomial

NaiveBayesSimple

NaiveBayesUpdateable

functions

lazy

meta

misc

trees

rules

Classifier output

== Evaluation on test split ==

== Summary ==

Correctly Classified Instances

Incorrectly Classified Instances

Gamma statistic

Mean absolute error

Root mean squared error

Relative absolute error

Root relative squared error

Total Number of Instances

50

1

0.9704

0.0239

0.1101

5.3594 %

23.2952 %

51

98.0392 %

1.9608 %

== 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

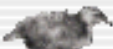
0

1

16

c = Iris-virginica

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

Log x 0


Weka Knowledge Explorer

Preprocess
Classify
Cluster
Associate
Select attributes
Visualize

Classifier

Choose
NaiveBayes

Test options

☐ Use training set
☐ Supplied test set

Set...

☐ Cross-validation Folds

10

☒ Percentage split %

66

More options...

(Nom) class

Start

Stop

Result list (right-click for options)

11:49:05 - trees.j48.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

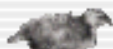
=== Confusion Matrix ===

 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 picture

Log

 x 0

Weka Knowledge Explorer

Preprocess **Classify** Cluster Associate Select attributes Visualize

Classifier: Choose **NaiveBayes**

Test options

☐ Use training set

☐ Supplied test set Set...

☐ Cross-validation Folds 10

☒ Percentage split % 66

More options...

(Nom) class

Start Stop

Result list (right-click for options)

11:49:05 - trees.j48.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

Log x 0

Weka Knowledge Explorer

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Classifier

Choose

NaiveBayes

Test options

☐ Use training set

☐ Supplied test set

☐ Cross-validation

☒ Percentage split

Set...

Folds

%

10

66

More options...

(Nom) class

Start

Stop

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

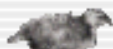
0 18 1 | b = Iris-versicolor

0 2 15 | c = Iris-virginica

Status

OK

Log

 x 0

Weka Knowledge Explorer

Preprocess **Classify** Cluster Associate Select attributes Visualize

Classifier

Choose **NaiveBayes**

Test options

☐ Use training set

☐ Supplied test set Set...

☐ Cross-validation Folds 10

☒ Percentage split % 66

More options...

(Nom) class

Start Stop

Result list (right-click for options)

- 11:49:05 - trees.j48.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

Log x 0


Weka Knowledge Explorer

Preprocess
Classify
Cluster
Associate
Select attributes
Visualize

Classifier

Choose
NaiveBayes

Test options

☐ Use training set
☐ Supplied test set

Set...

☐ Cross-validation Folds

10

☒ Percentage split %

66

More options...

(Nom) class

Start

Result list (right-click for)

11:49:05 - trees.j48.J
14:34:28 - functions.
14:48:05 - bayes.Nai

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

Classifier 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 ===

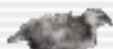
classified as
Iris-setosa
Iris-versicolor
Iris-virginica

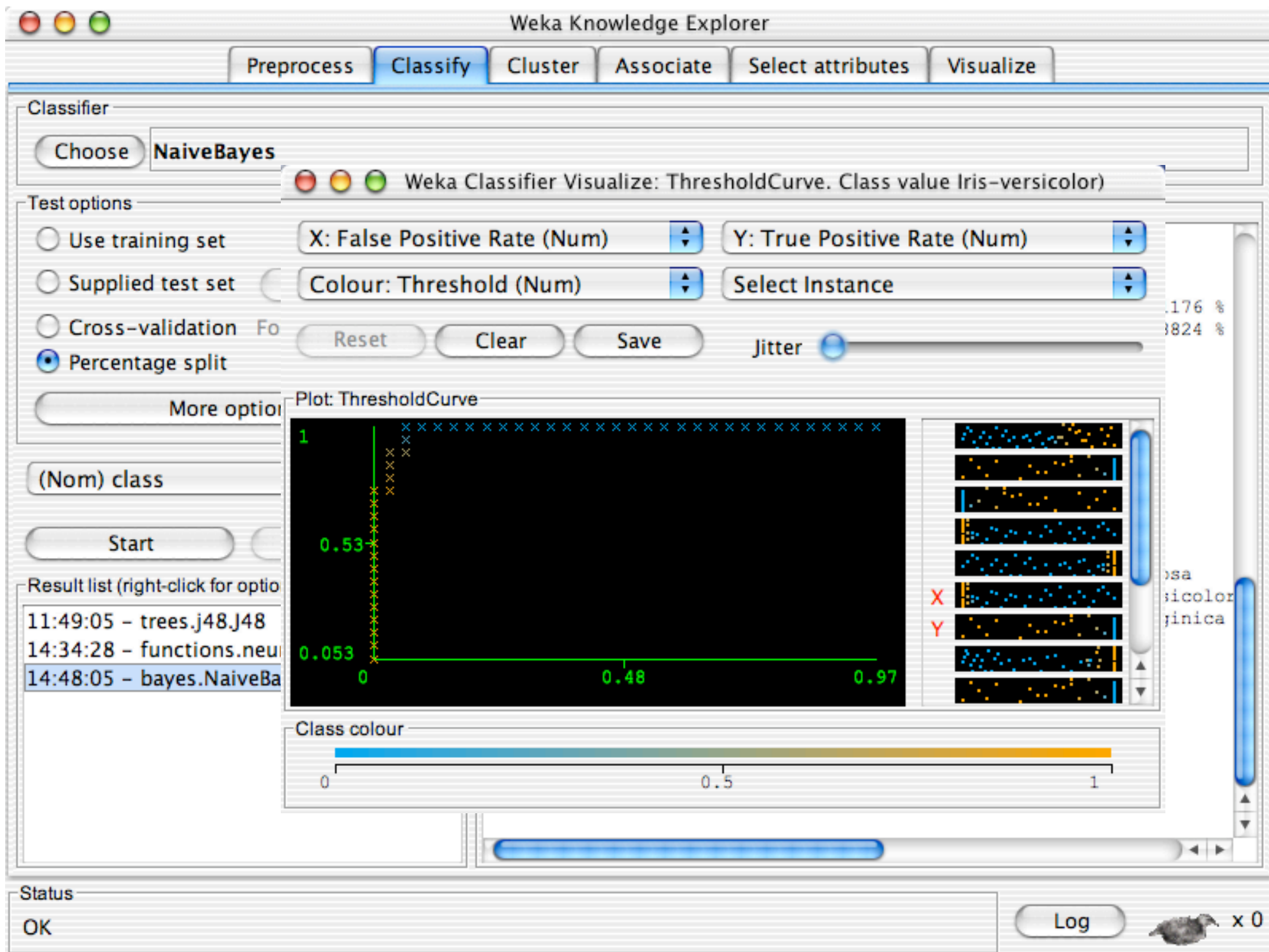
```

Iris-setosa
Iris-versicolor
Iris-virginica

Status

OK

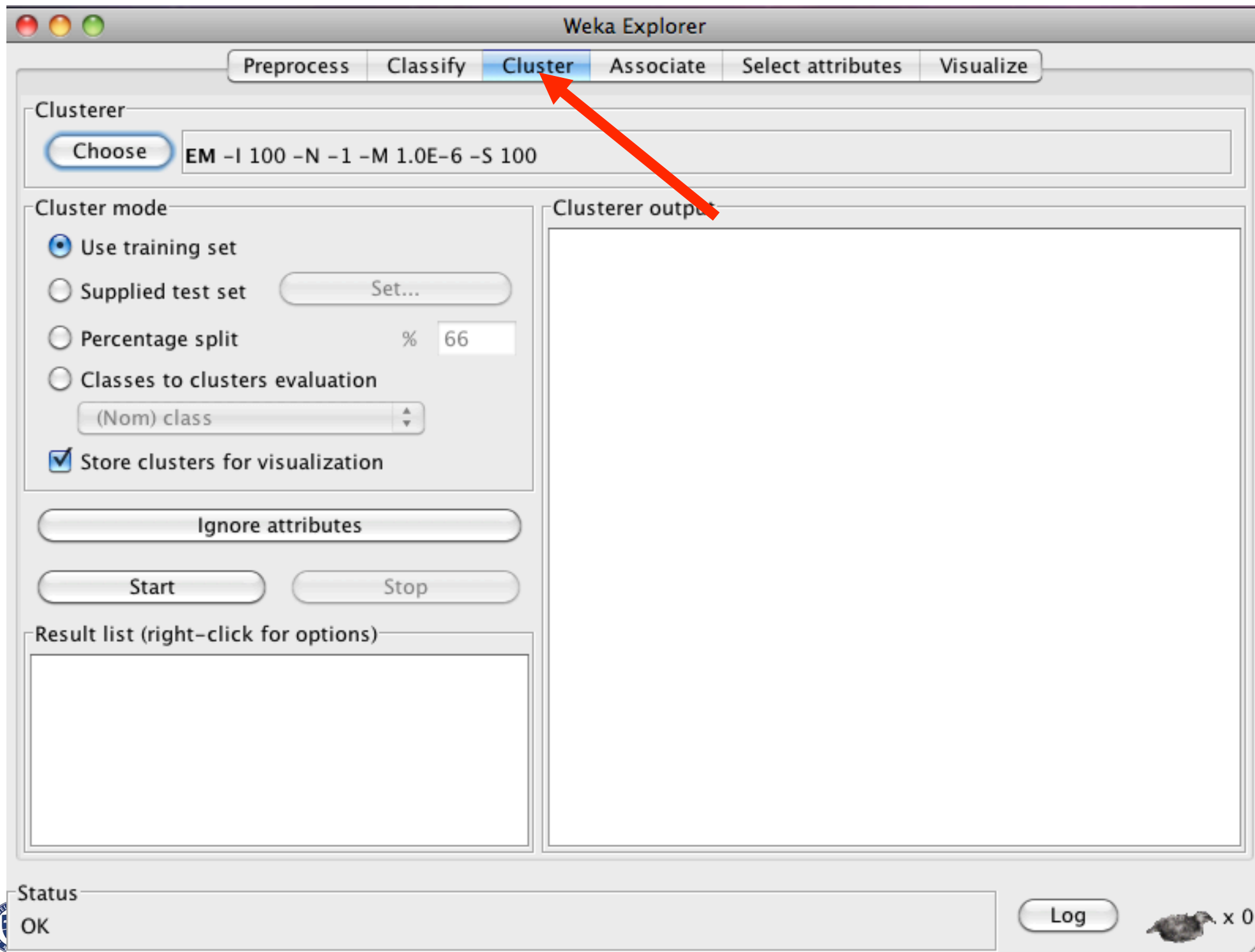
Log  x 0

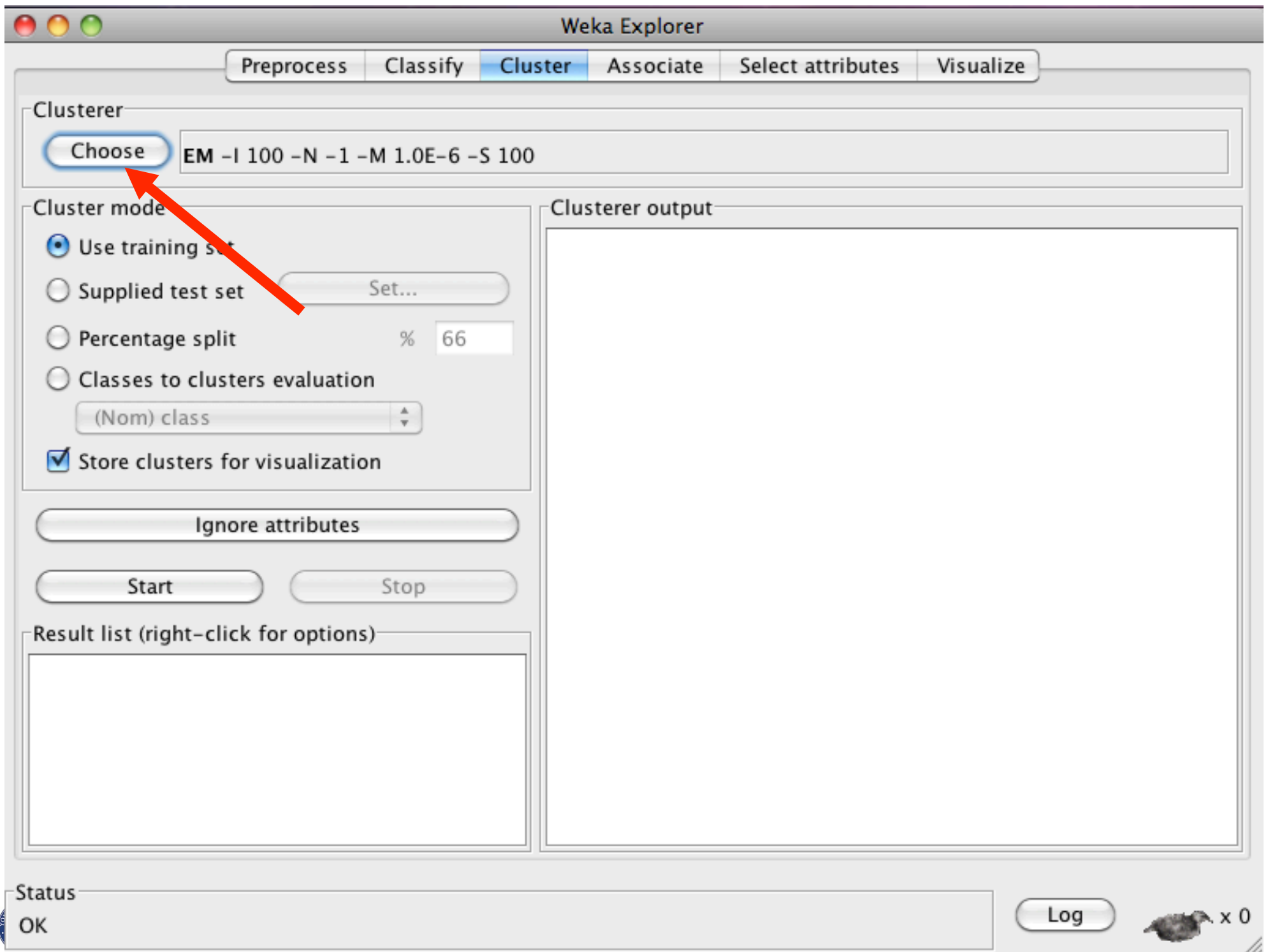


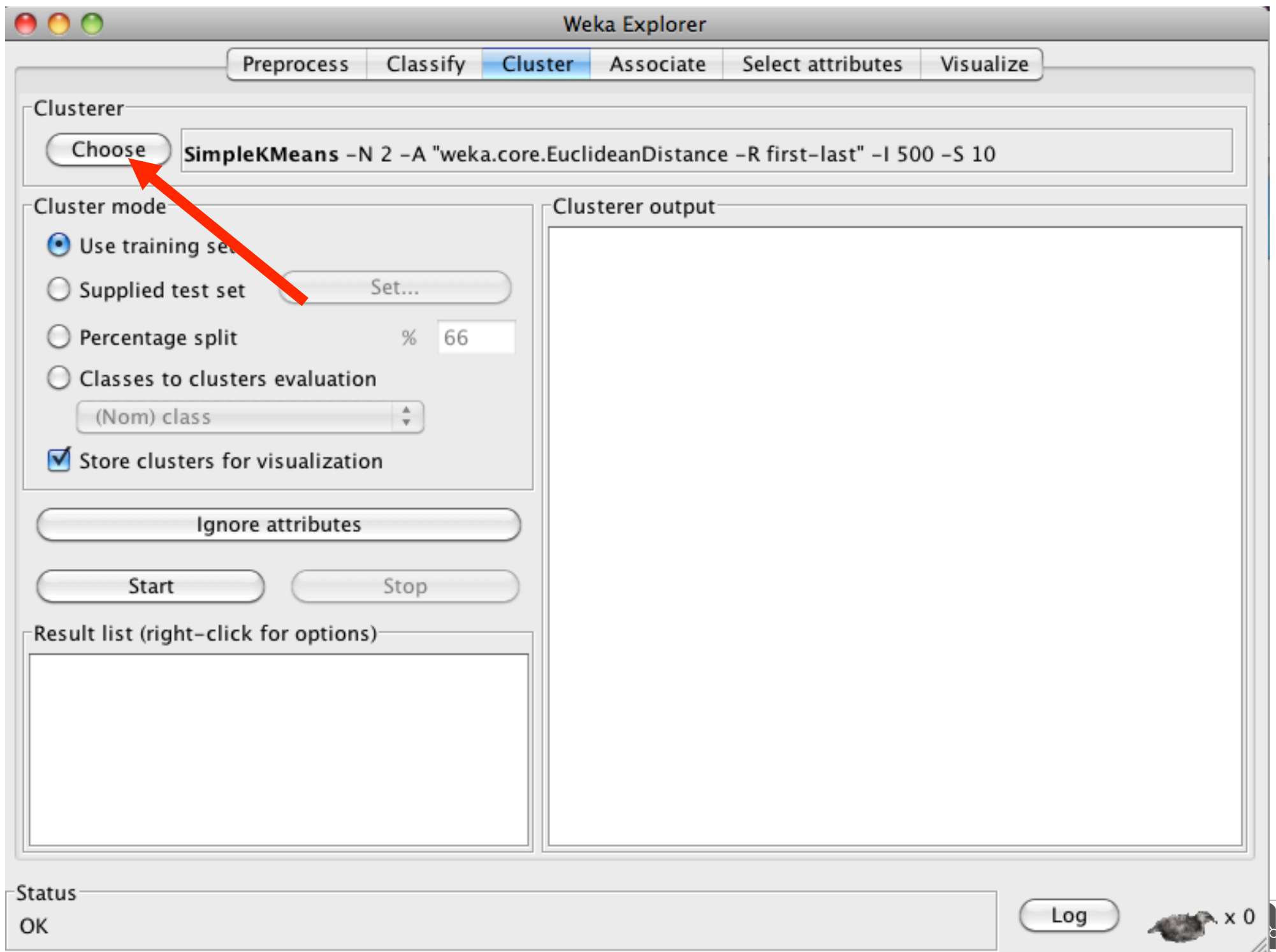
Explorer: Clustering



- WEKA contains many clustering implementations:
 - Works with both discrete and numerical data
- Example of K-means







Clusterer

Choose

Cluster mode

☒ Use training data☐ Supplied test data☐ Percentage of data☐ Classes to cluster

(Nom) cl

☒ Store cluster results

Start

Result list (right)

15:11:12 - SimpleKMeans

weka.gui.GenericObjectEditor

weka.clusterers.SimpleKMeans

About

Cluster data using the k means algorithm.

More

Capabilities

displayStdDevs False

distanceFunction Choose EuclideanDistance -R first-last

dontReplaceMissingValues False

maxIterations 500

numClusters 3

preserveInstancesOrder False

seed 10

Open... Save... OK Cancel

-S 10

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

g set ===

rs: 62.1436882815797
with 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%)

Status

OK

Log

x 0

Weka Explorer

Preprocess Classify **Cluster** Associate Select attributes Visualize

Clusterer

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

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)

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

Clusterer output

Relation: iris
Instances: 150
Attributes: 5
sepalwidth
sepalwidth
petalwidth
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:

Attribute	Full Data (150)	Cluster# 0 (50)	1 (50)	2 (50)
sepalwidth	5.8433	5.936	5.006	6.588
sepalwidth	3.054	2.77	3.418	2.974
petalwidth	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

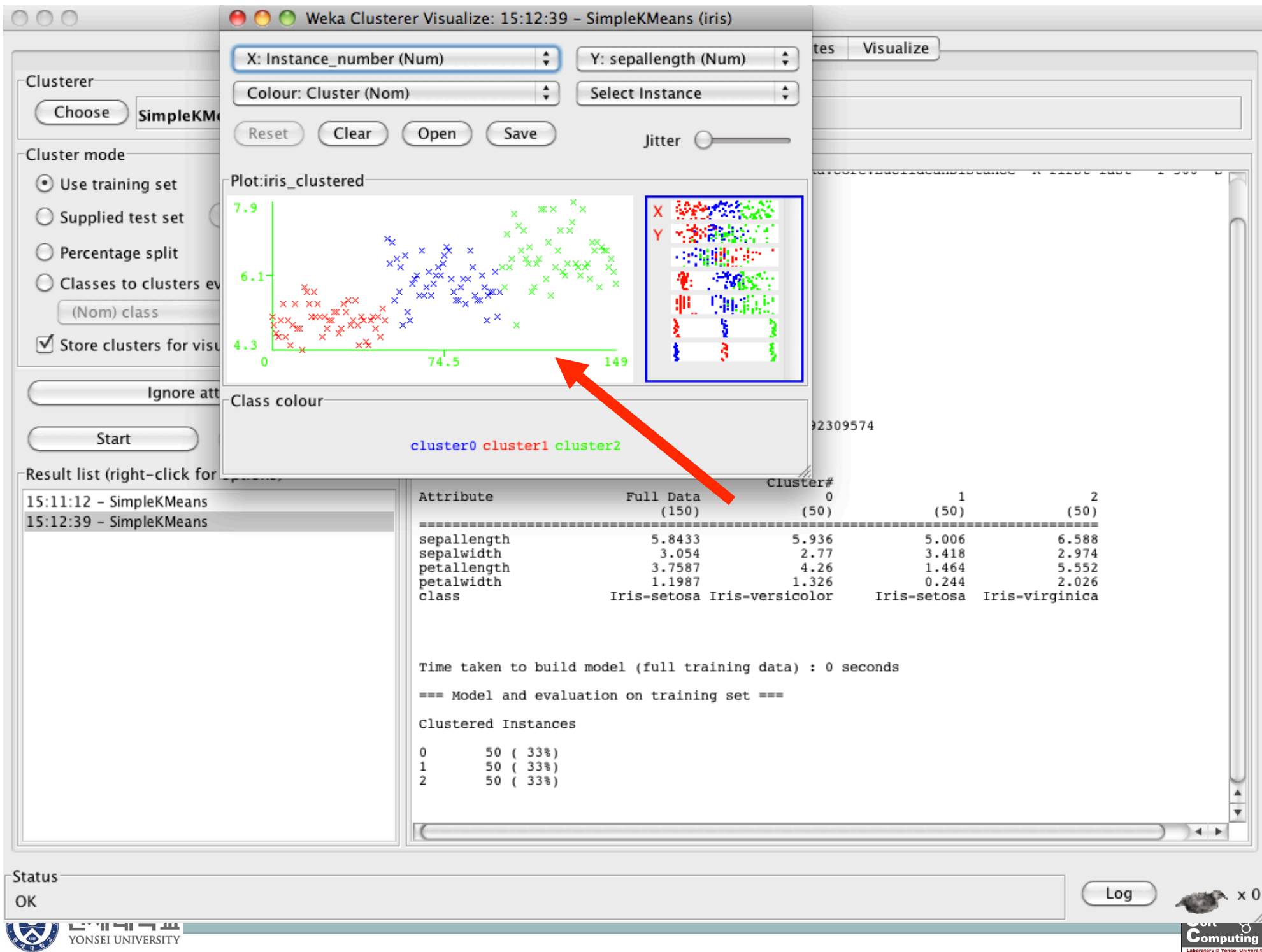
0	50 (33%)
1	50 (33%)
2	50 (33%)

Status

OK

Log

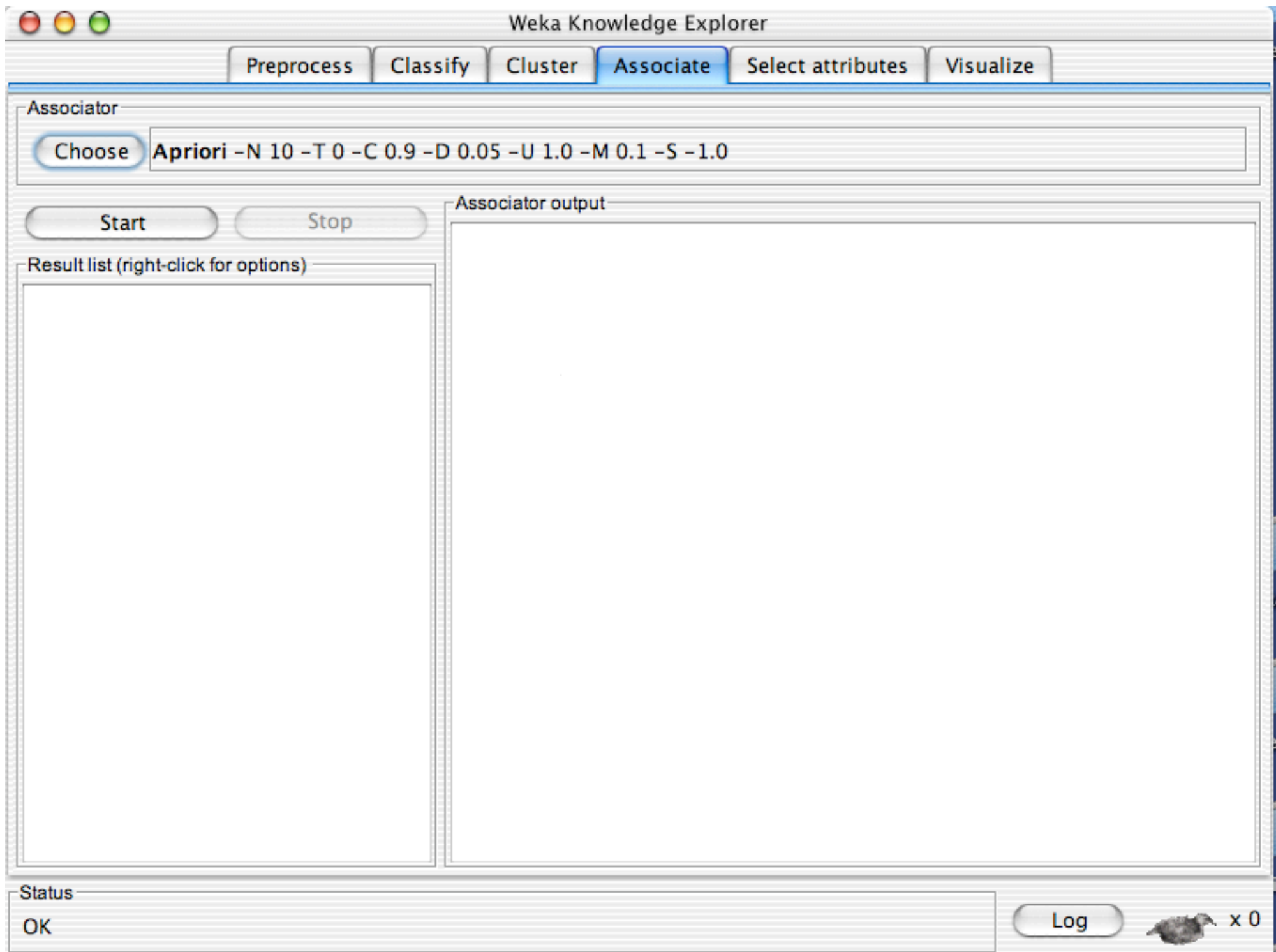
x 0

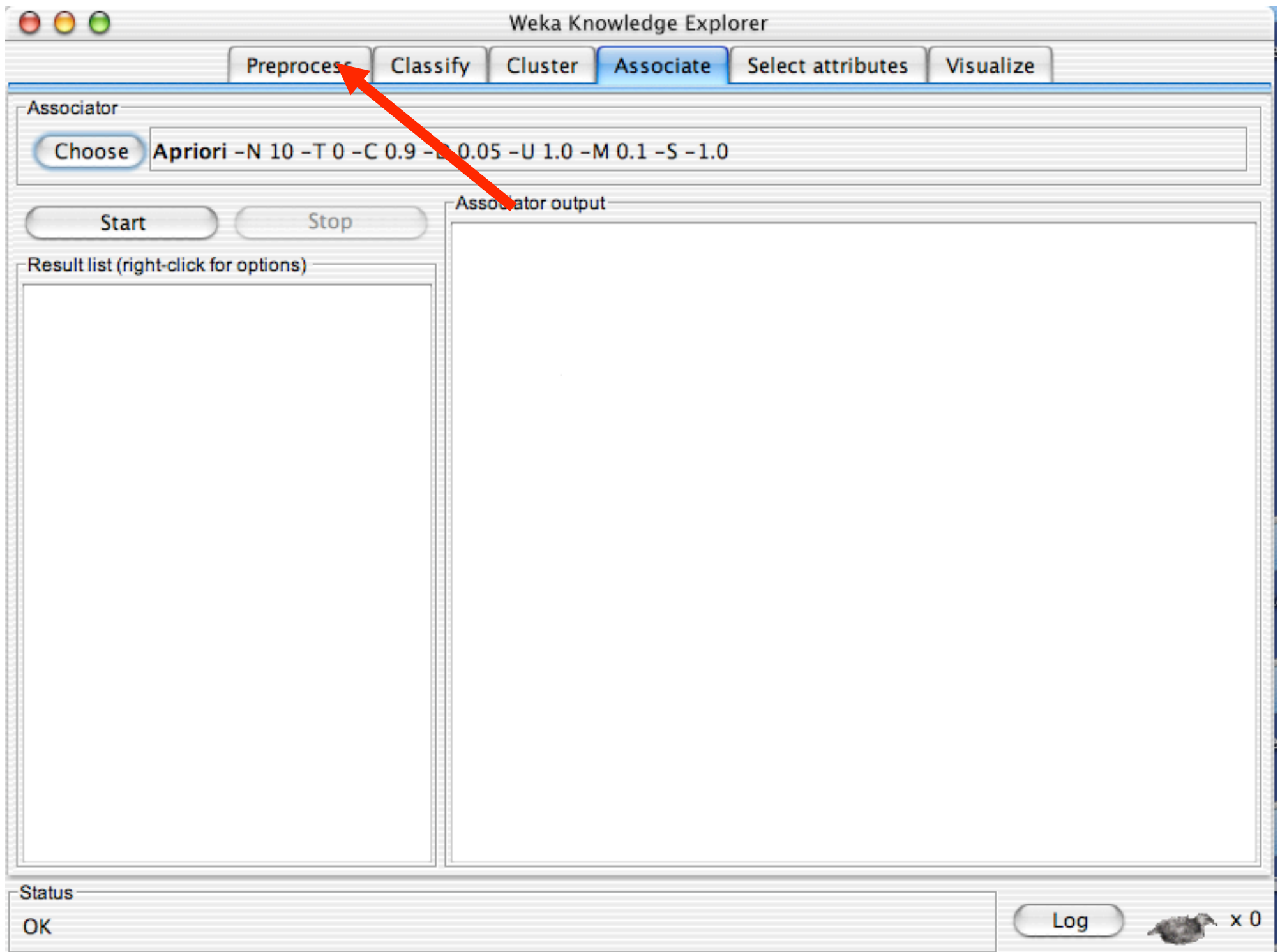


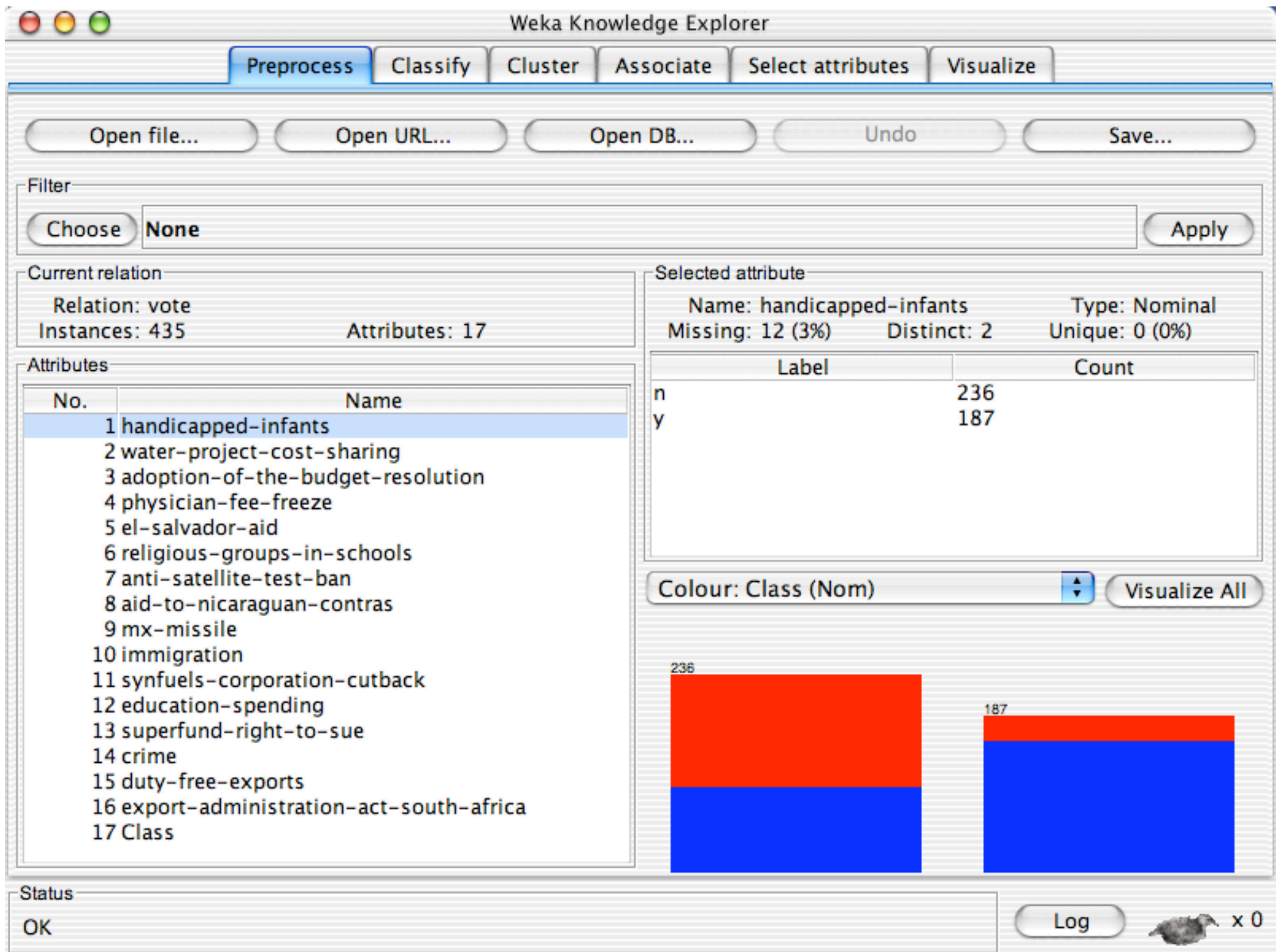
Explorer: Finding Associations



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







Weka Knowledge Explorer

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Open file...

Open URL...

Open DB...

Undo

Save...

Filter

Choose

None

Apply

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)

Visualize All

236

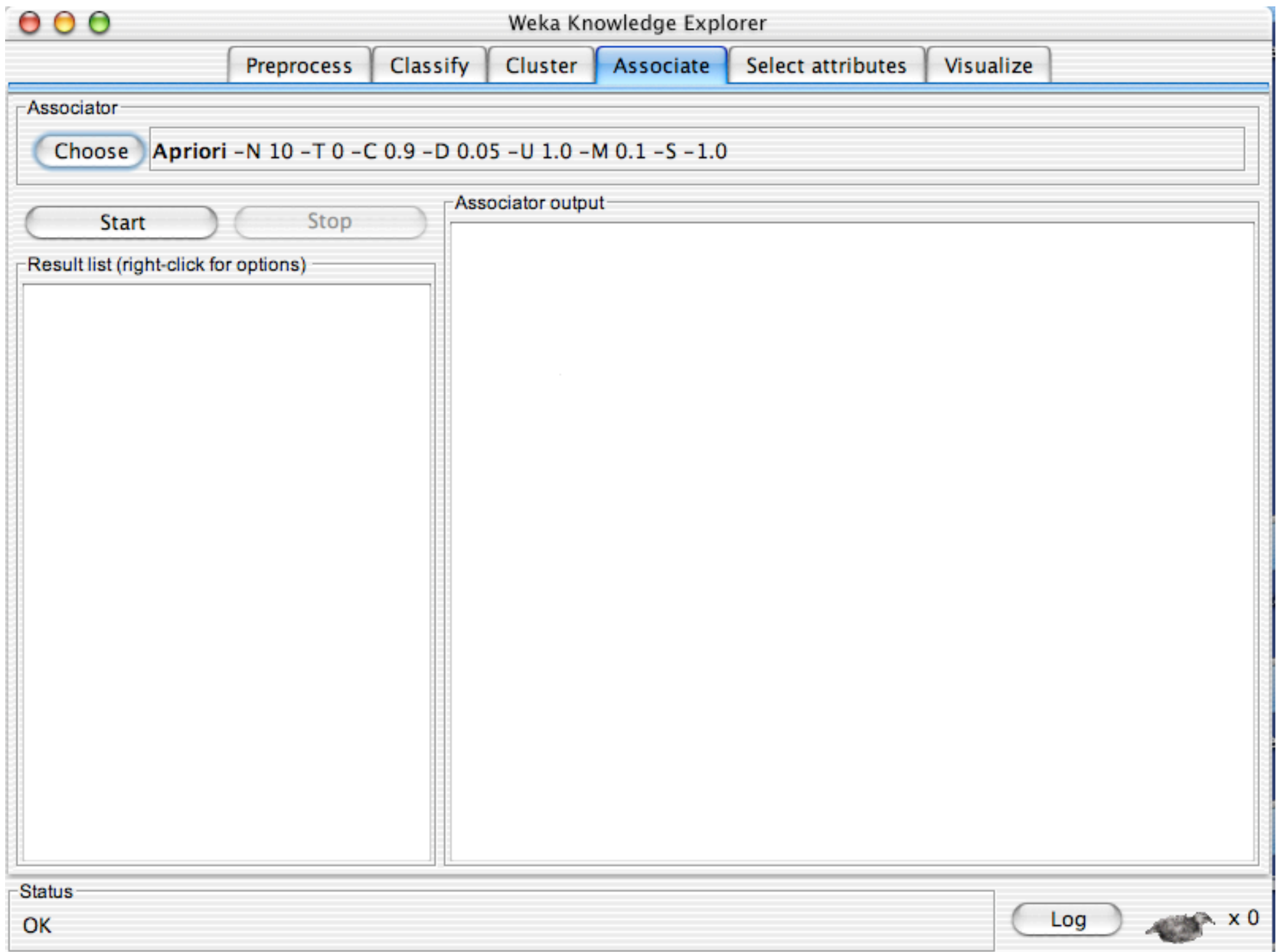
187

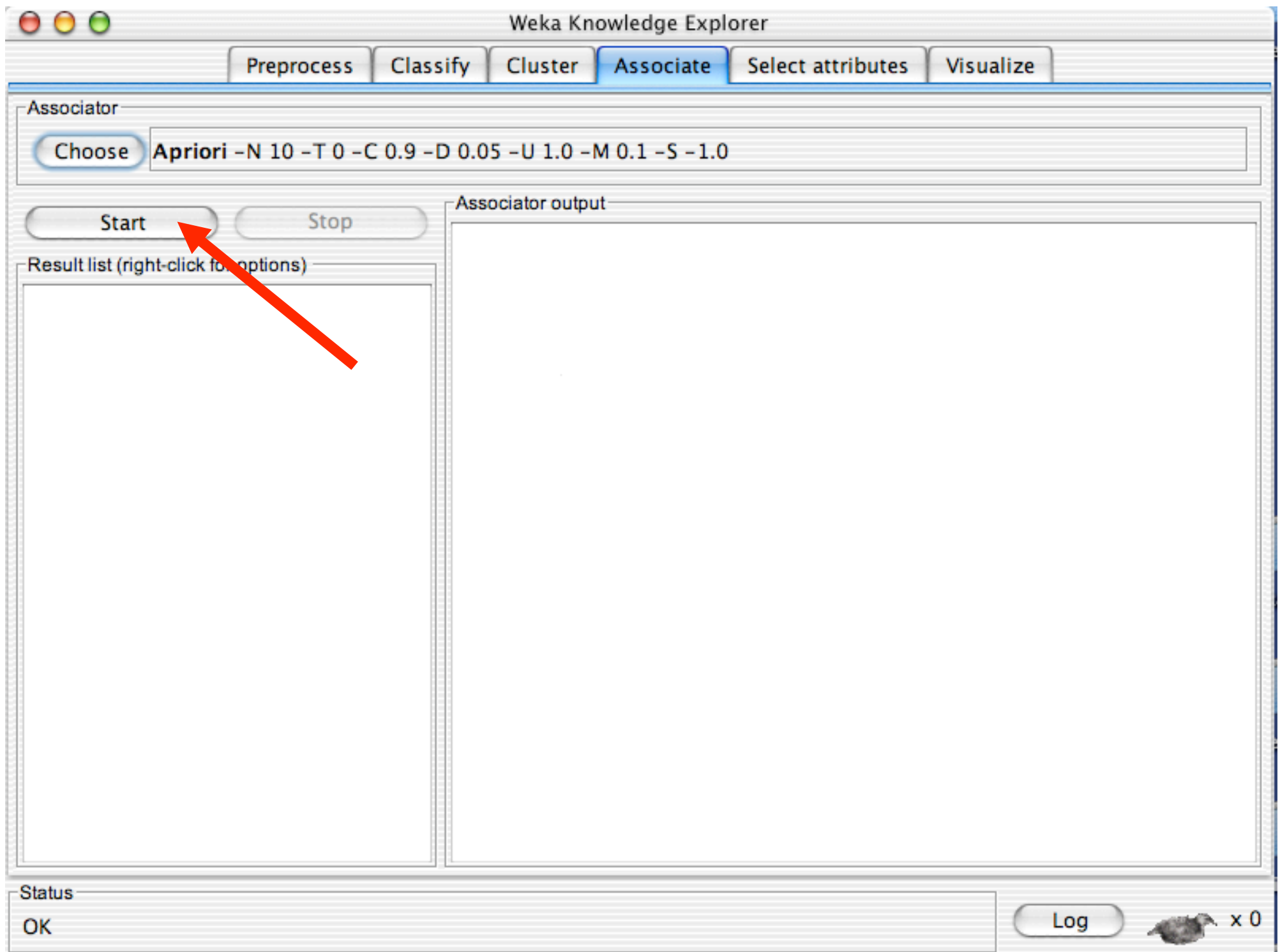
Status

OK

Log

x 0





Weka Knowledge Explorer

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Associator

Choose

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

Start

Stop

Result list (right-click for options)

16:29:37 - Apriori

Associator output

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

Generated sets of large itemsets:

Size of set of large itemsets L(1): 20

Size of set of large itemsets L(2): 17

Size of set of large itemsets L(3): 6

Size of set of large itemsets L(4): 1

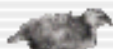
Best rules found:

1. adoption-of-the-budget-resolution=y physician-fee-freeze=n 219 ==> Class=democrat 210
2. adoption-of-the-budget-resolution=y physician-fee-freeze=n aid-to-nicaraguan-contras=y 211 ==> Class=democrat 210
3. physician-fee-freeze=n aid-to-nicaraguan-contras=y 211 ==> Class=democrat 210
4. physician-fee-freeze=n education-spending=n 202 ==> Class=democrat 201 conf:(0.98)
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:(0.98)
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 204
9. el-salvador-aid=n aid-to-nicaraguan-contras=y 204 ==> Class=democrat 197 conf:(0.98)
10. aid-to-nicaraguan-contras=y Class=democrat 218 ==> physician-fee-freeze=n 210

Status

OK

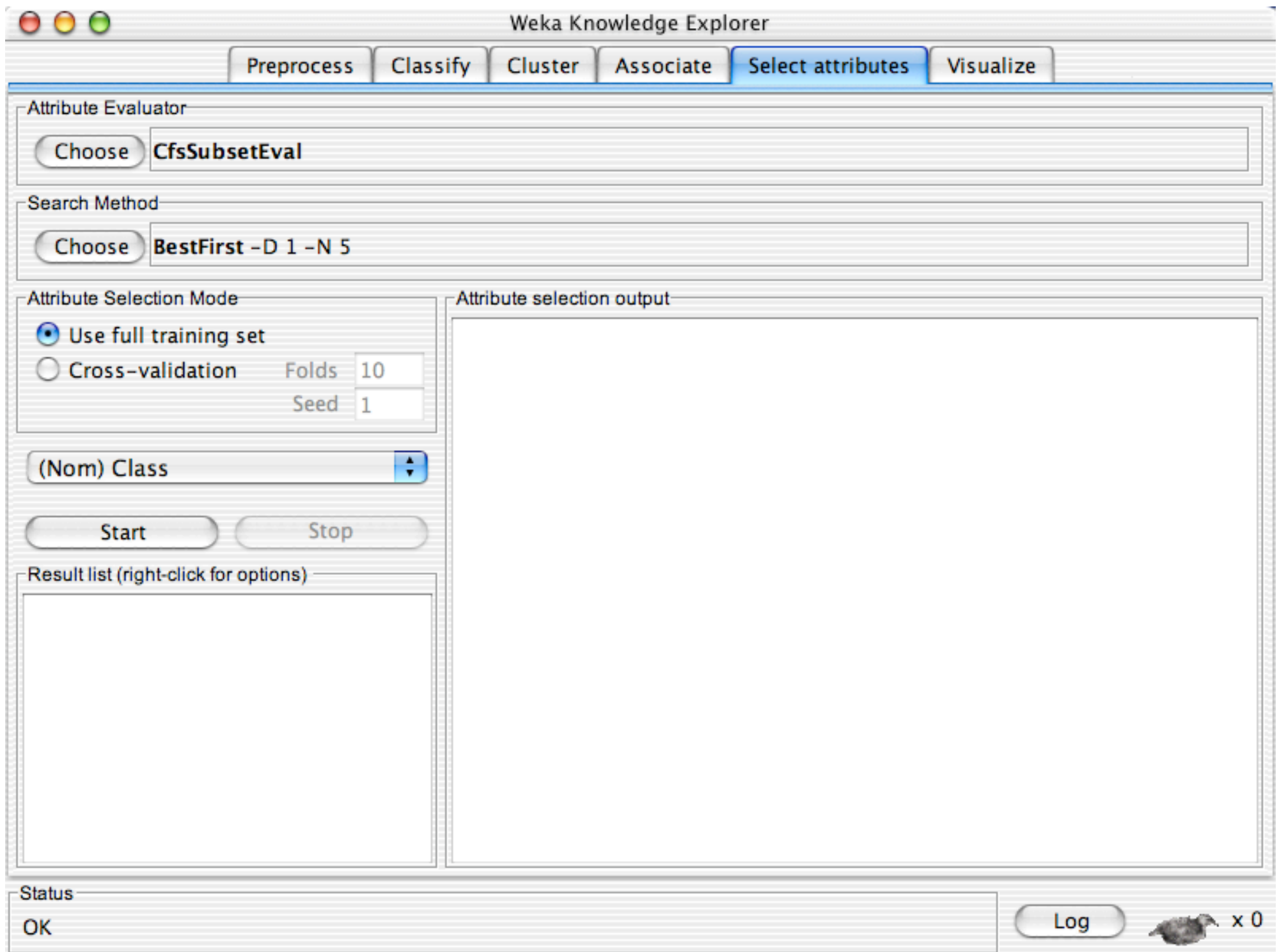
Log

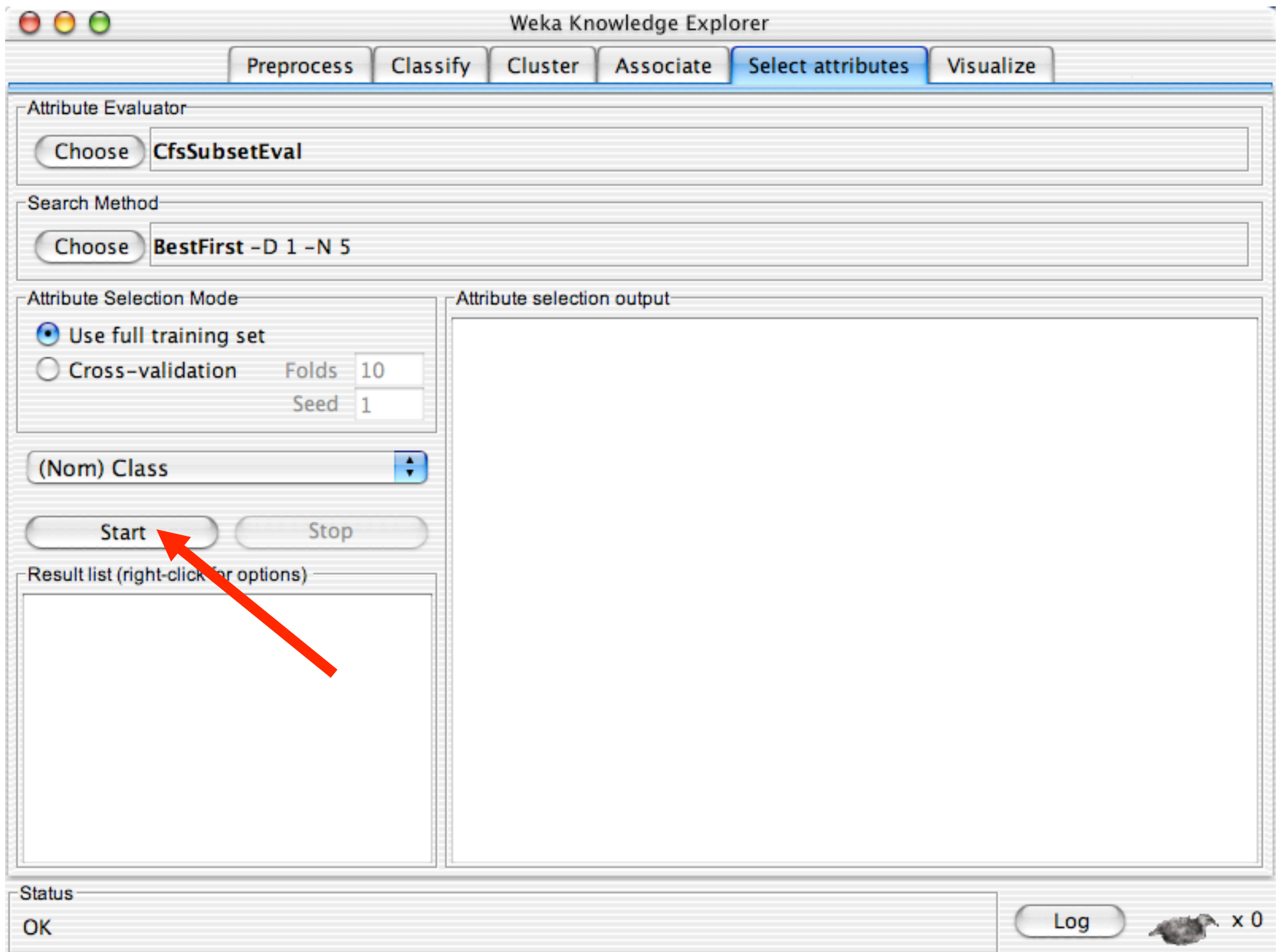
 x 0

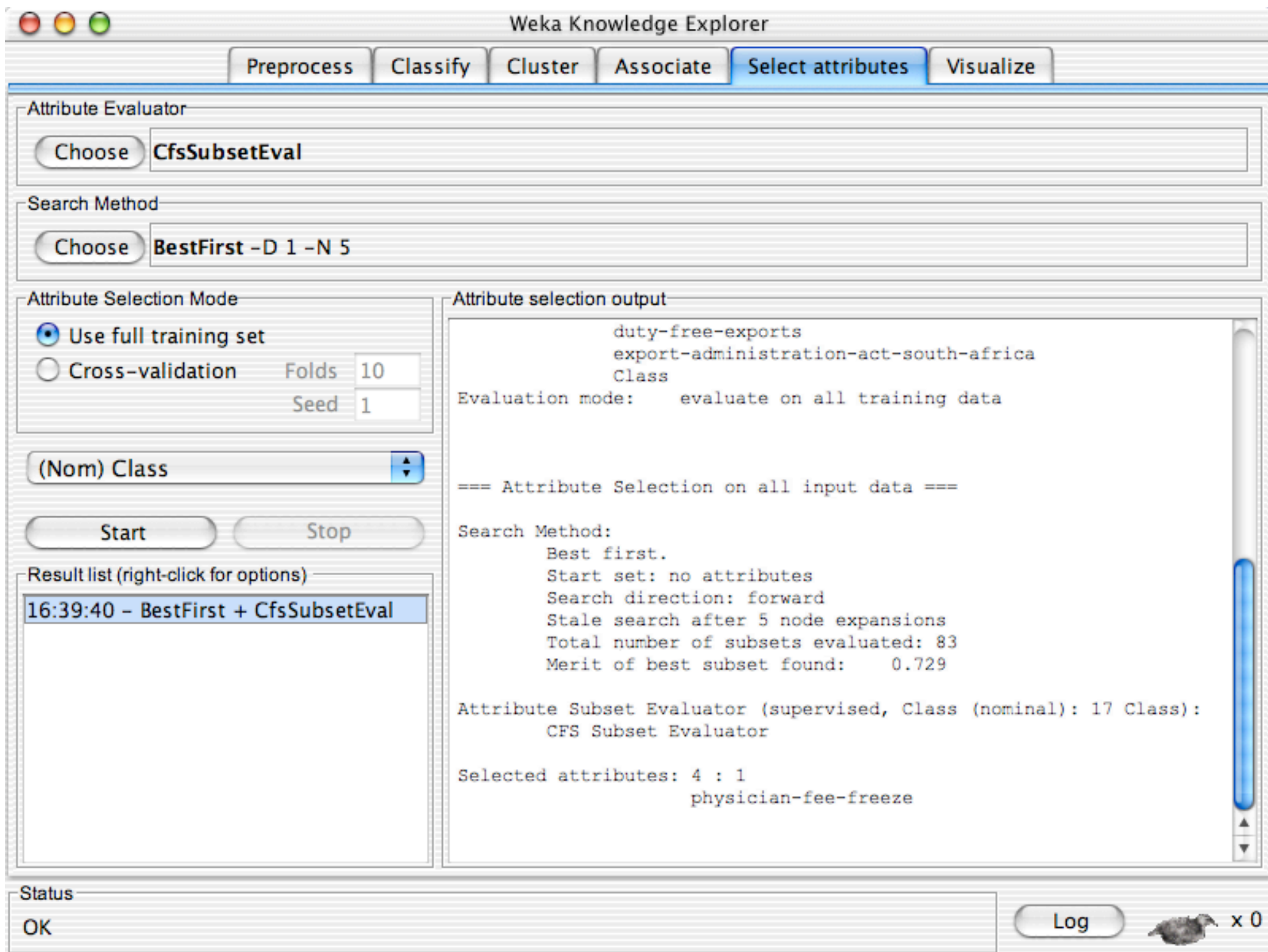
Explorer: Attribute Selection

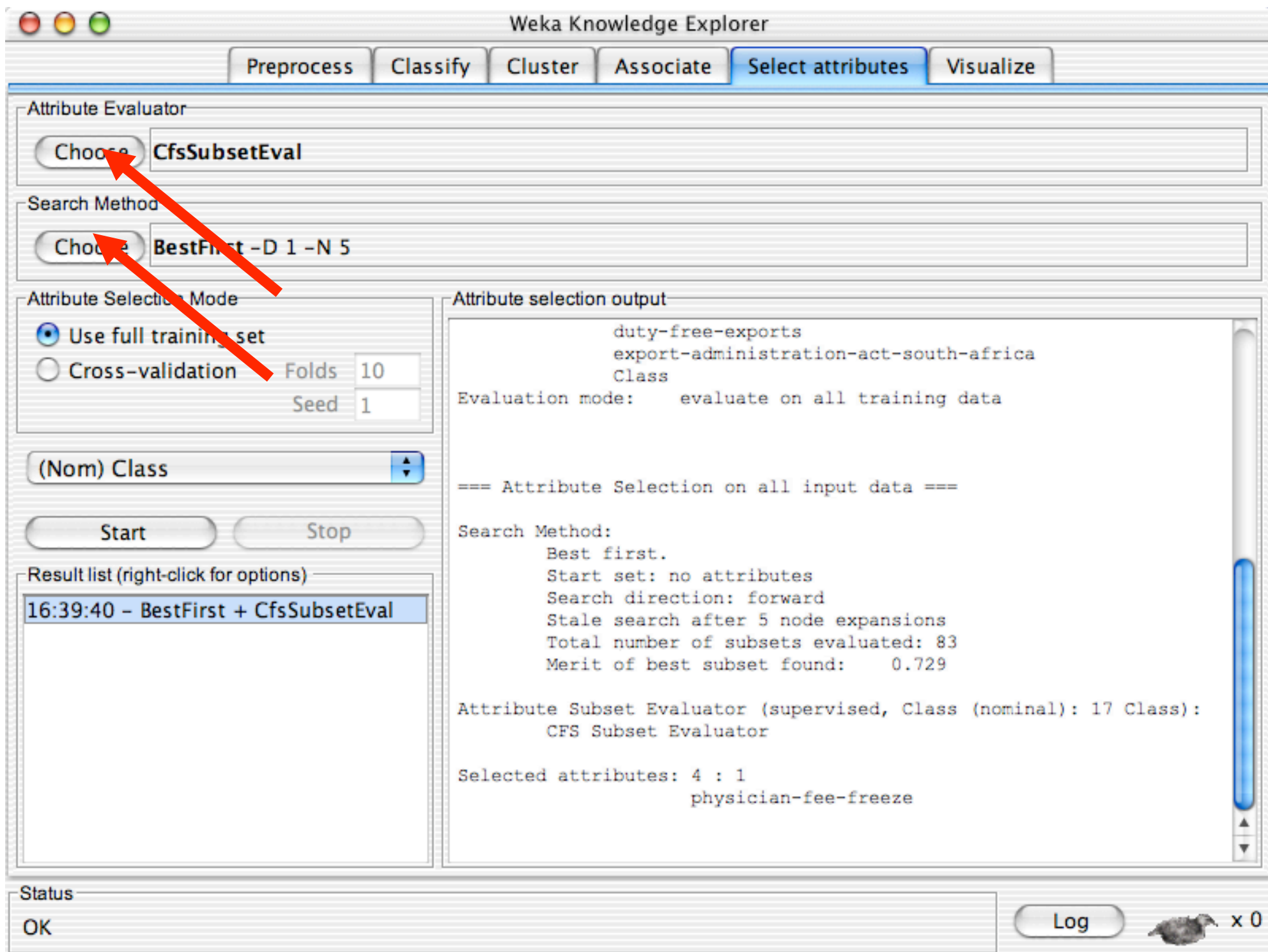


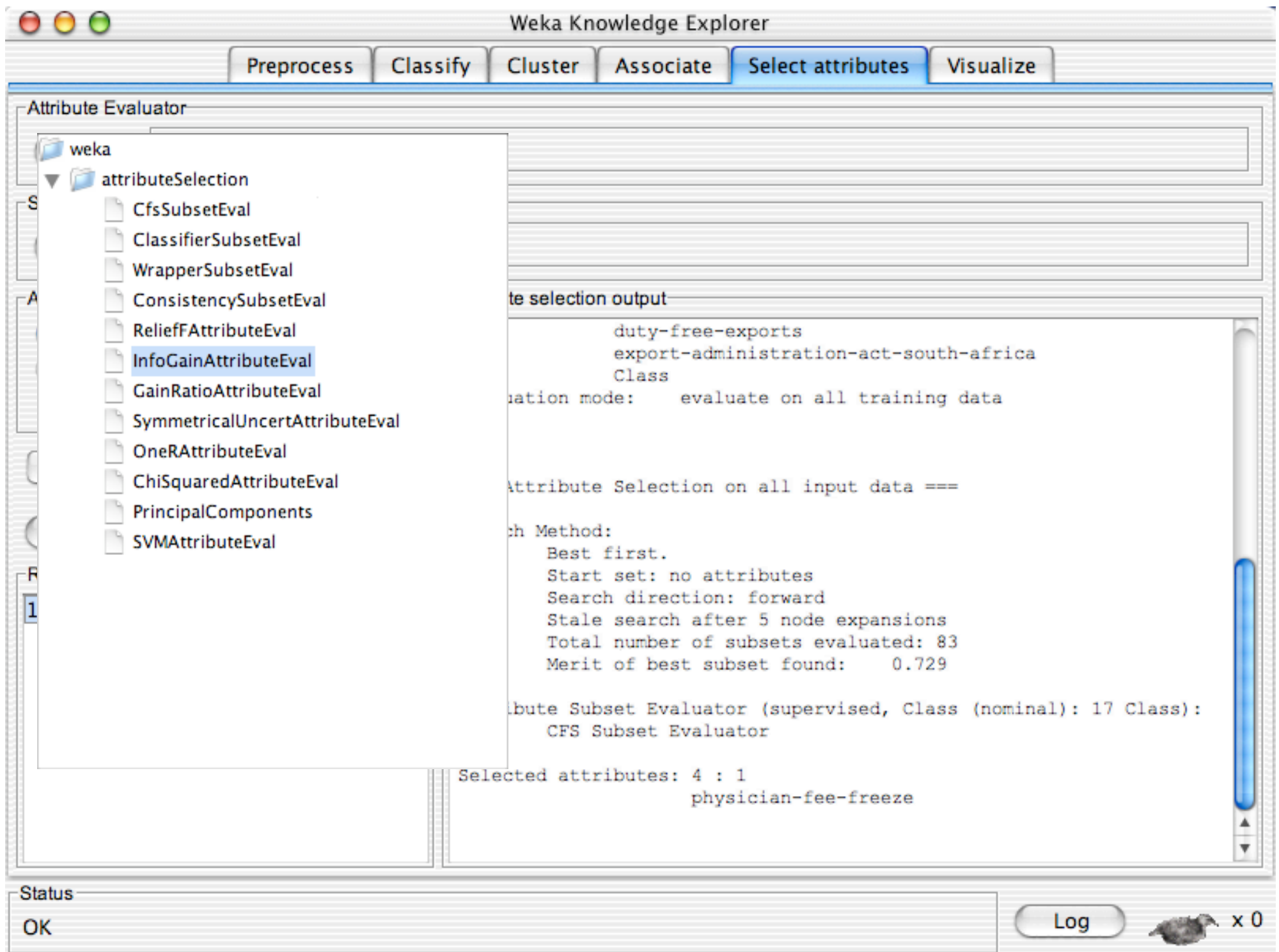
- Panel that can be used to investigate which (subsets of) attributes are the most predictive ones
- Attribute selection methods contain two parts:
 - A search method:
 - ✦ best-first, forward selection, random, exhaustive, genetic algorithm, ranking
 - An evaluation method:
 - ✦ correlation-based, wrapper,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











Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Attribute Evaluator

Choose

InfoGainAttributeEval

Search Method

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

E308 -N -1

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
```

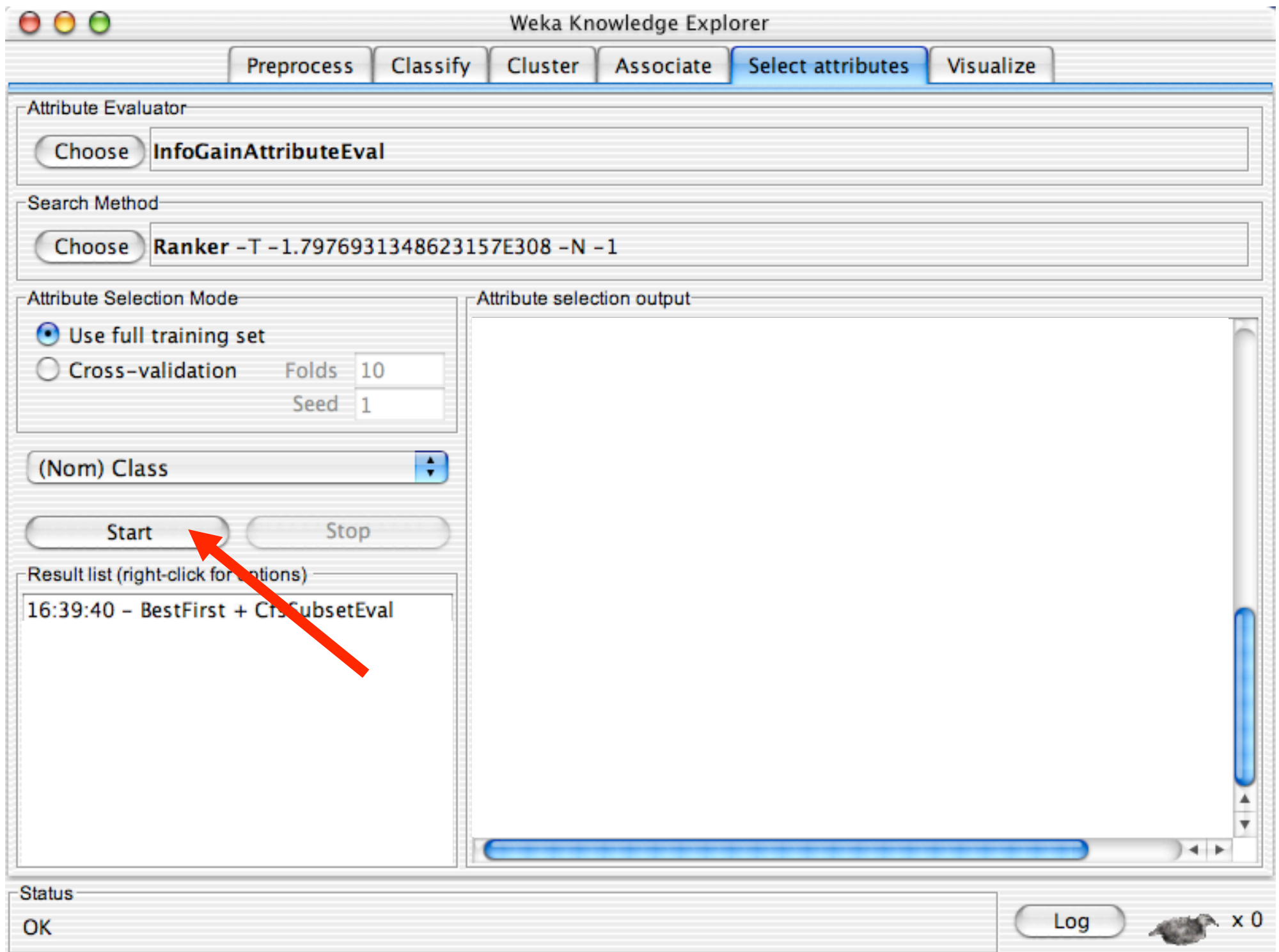
```
Selected attributes: 4 : 1
                    physician-fee-freeze
```

Status

OK

Log

 x 0



Weka Knowledge Explorer

Preprocess Classify Cluster Associate **Select attributes** Visualize

Attribute Evaluator

Choose **InfoGainAttributeEval**

Search Method

Choose **Ranker -T -1.7976931348623157E308 -N -1**

Attribute Selection Mode

☒ Use full training set
☐ Cross-validation Folds 10 Seed 1

(Nom) Class

Start Stop

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

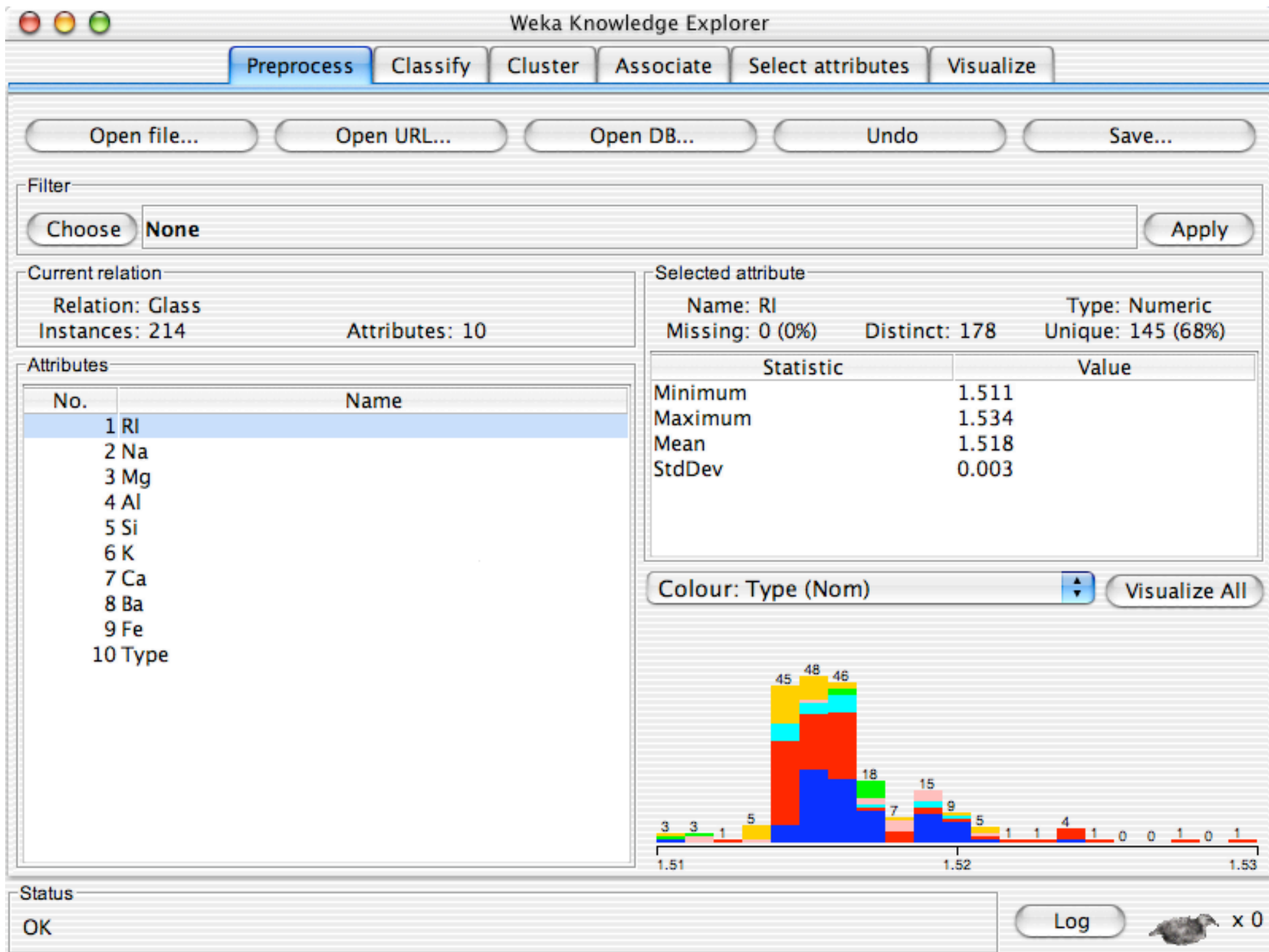
OK

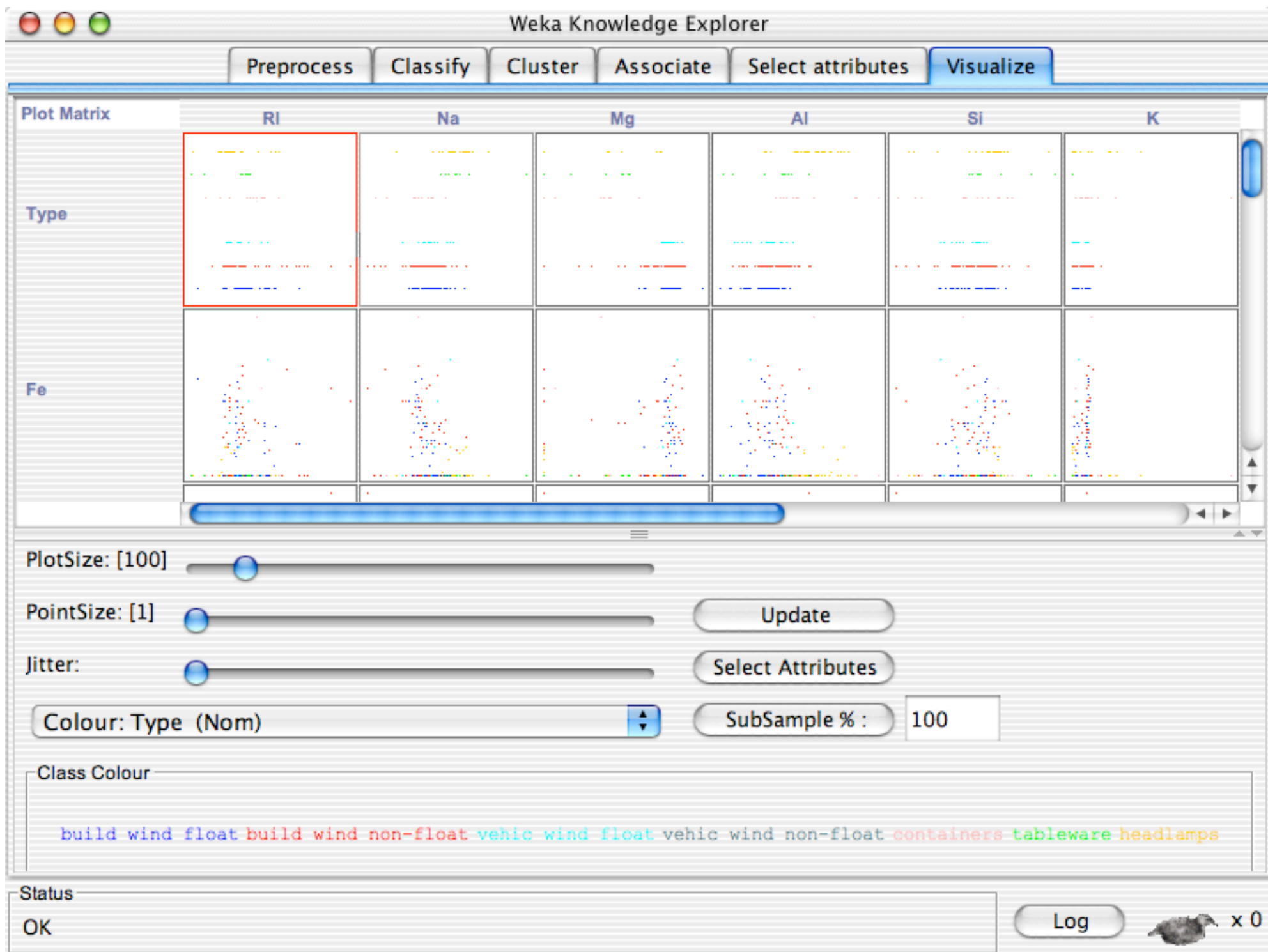
Log x 0

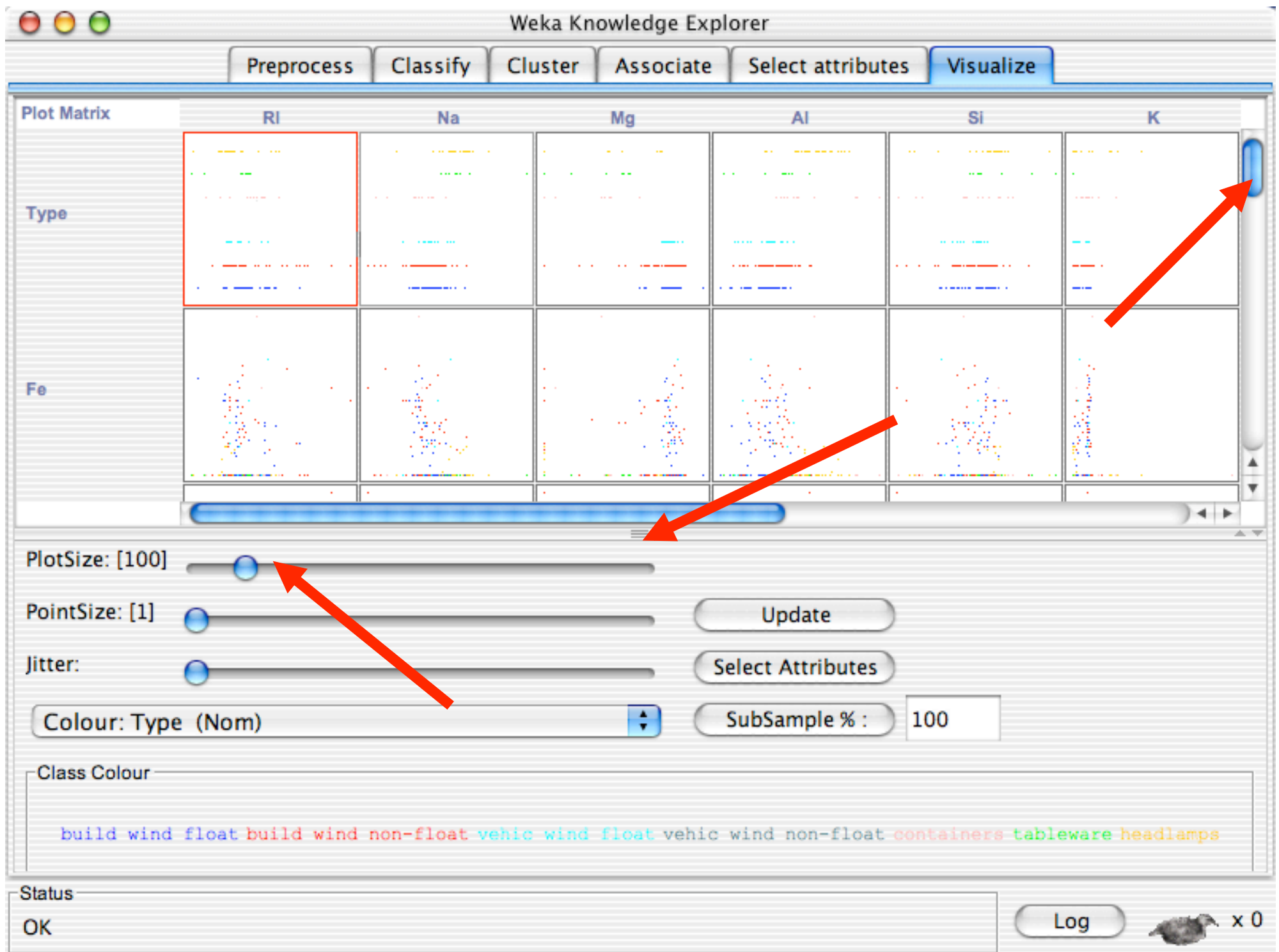
Explorer: Data Visualization

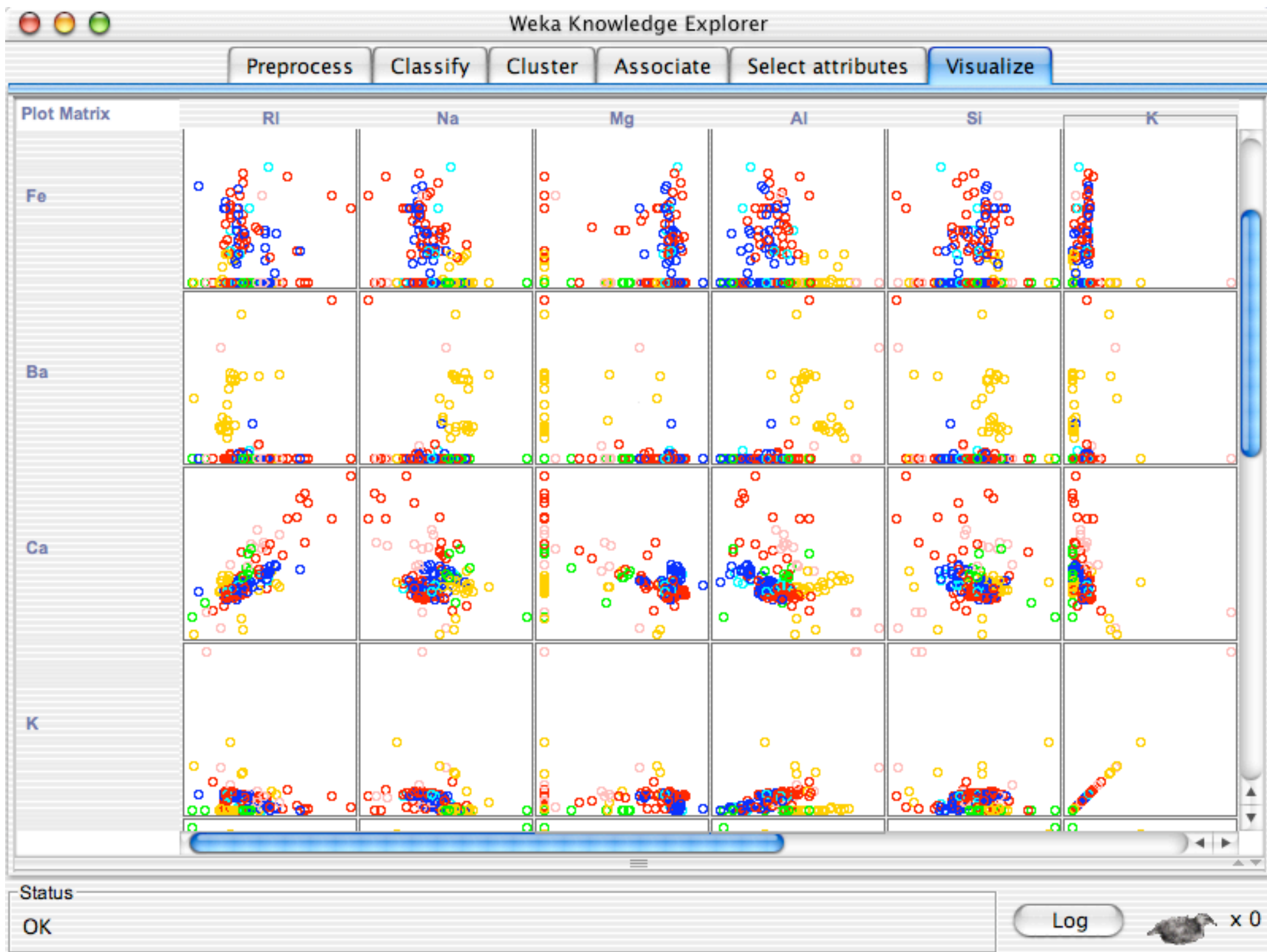


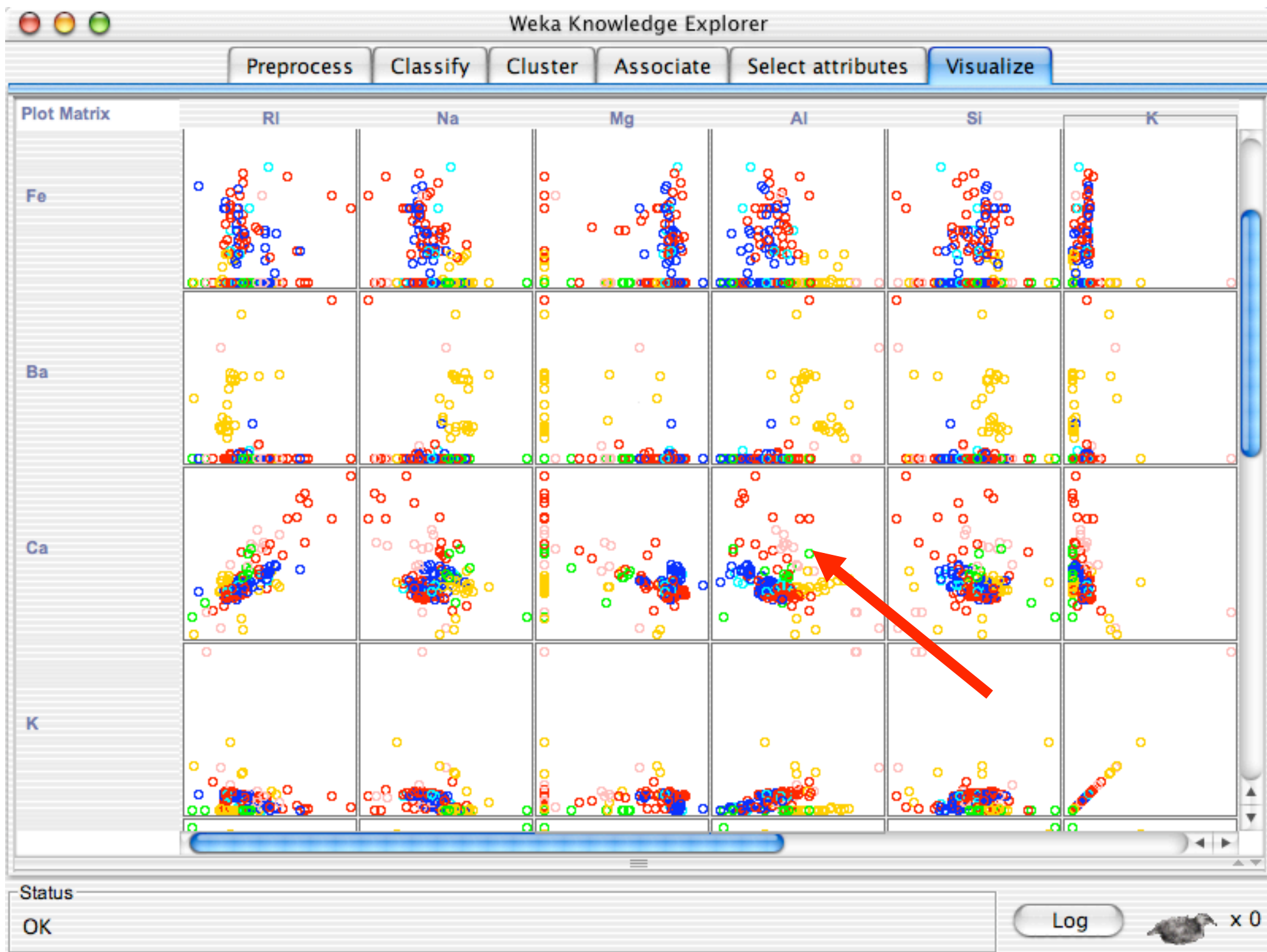
- Visualization very useful in practice:
 - e.g. helps to determine difficulty of the learning problem
- WEKA can visualize single attributes and pairs of attributes
 - To do: rotating 3-d visualizations (Xgobi-style)
- Color-coded class values
- “Jitter” option to deal with nominal attributes (and to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







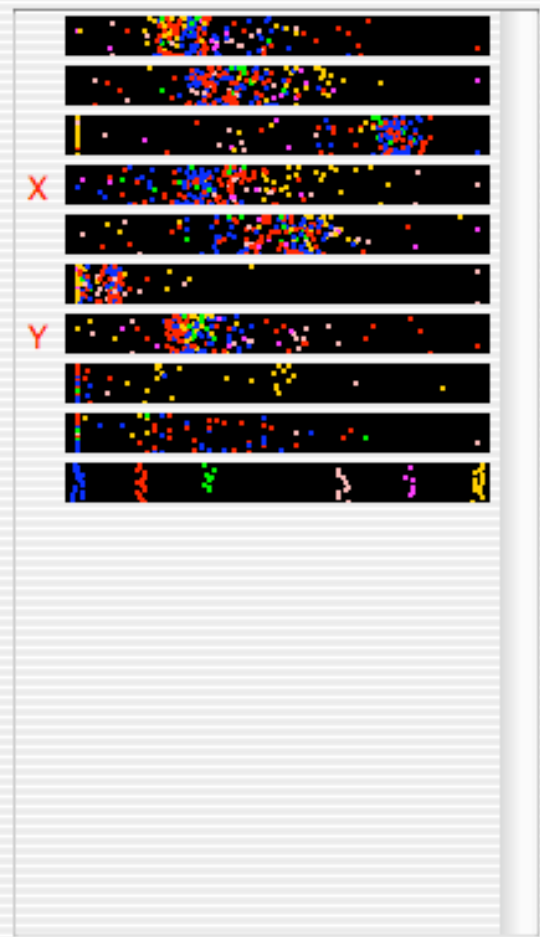
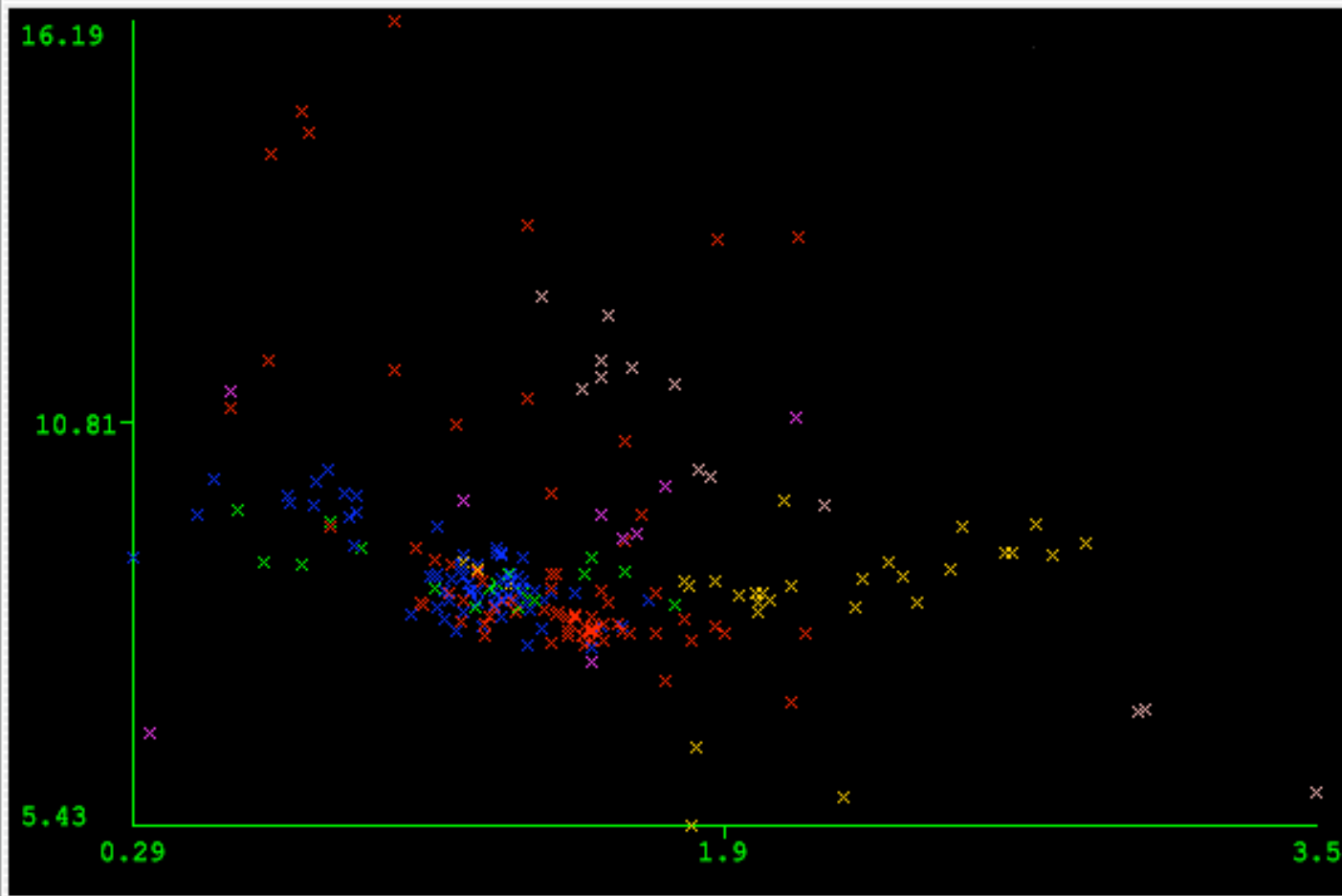




X: Al (Num) Y: Ca (Num)
 Colour: Type (Nom) Select Instance

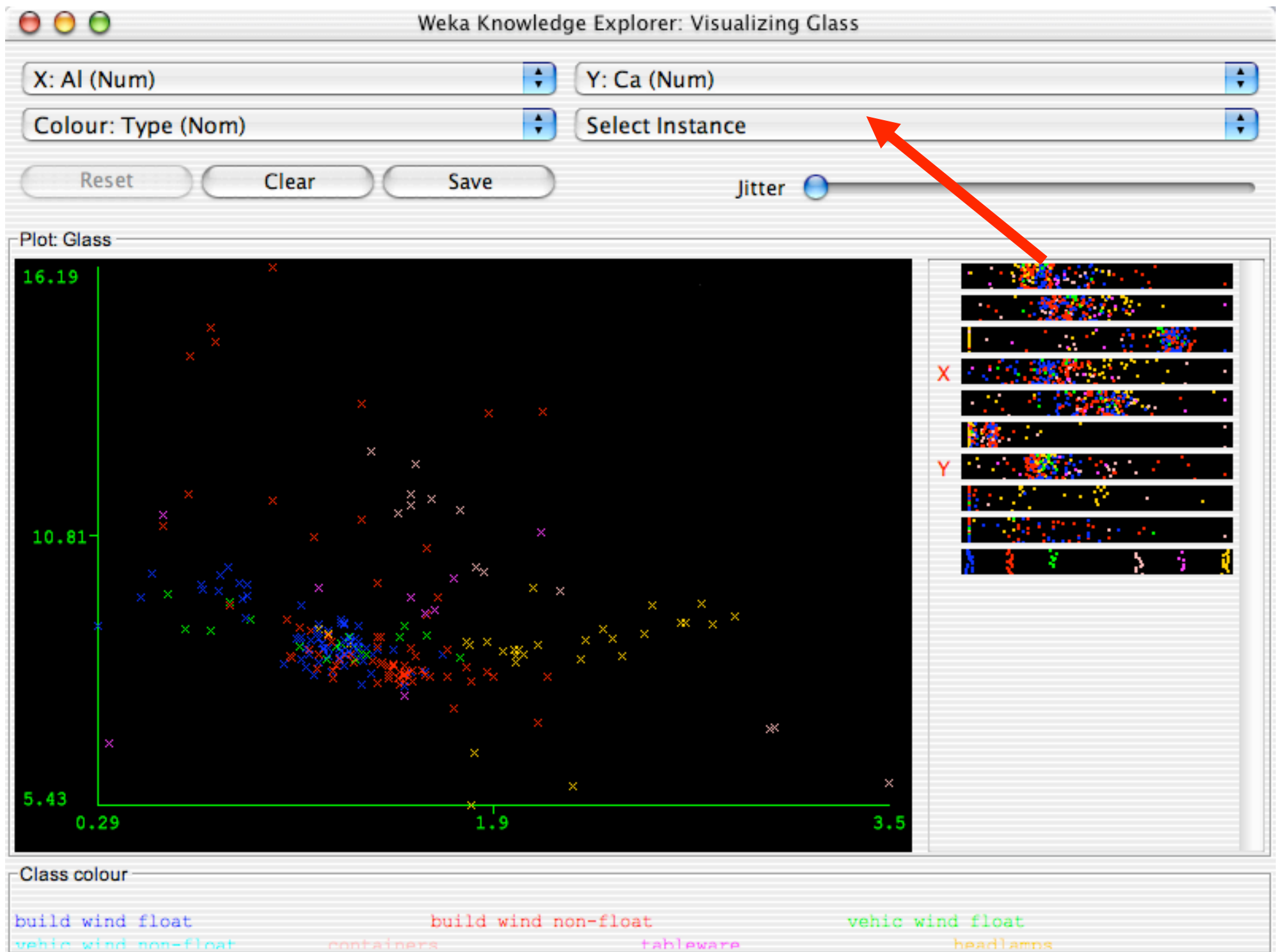
Reset Clear Save Jitter

Plot: Glass



Class colour

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



Weka Knowledge Explorer: Visualizing Glass

X: Al (Num)

Y: Ca (Num)

Colour: Type (Nom)

Rectangle

Submit

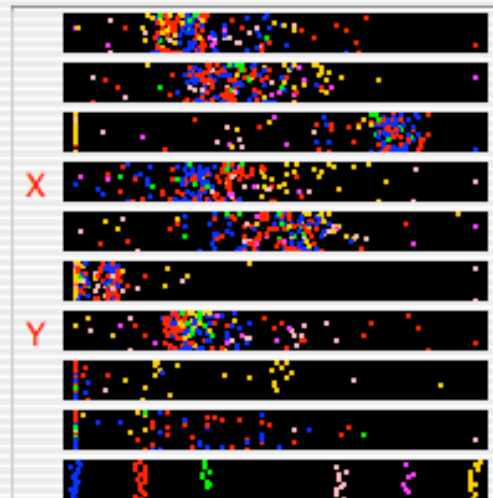
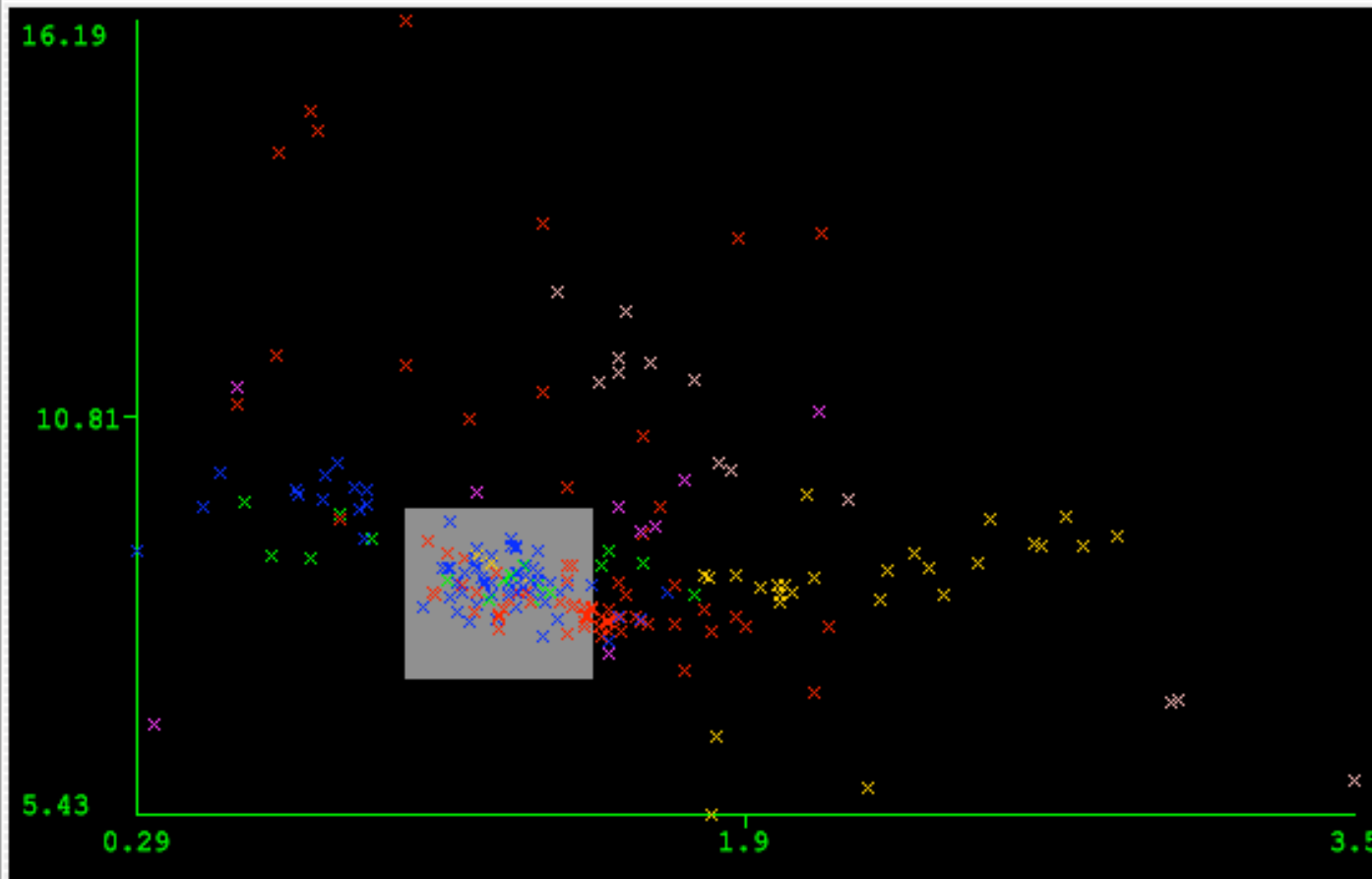
Clear

Save

Jitter



Plot: Glass



Class colour

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

Weka Knowledge Explorer: Visualizing Glass

X: Al (Num)

Y: Ca (Num)

Colour: Type (Nom)

Rectangle

Submit

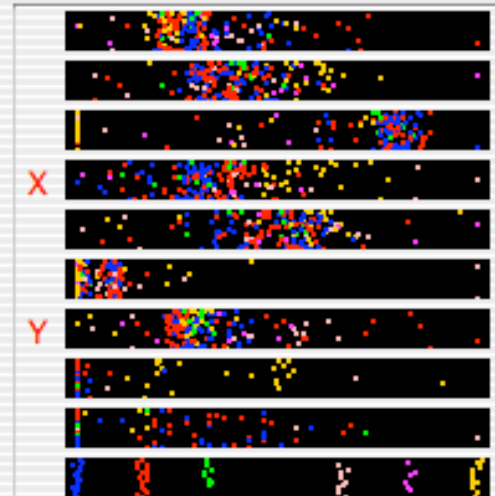
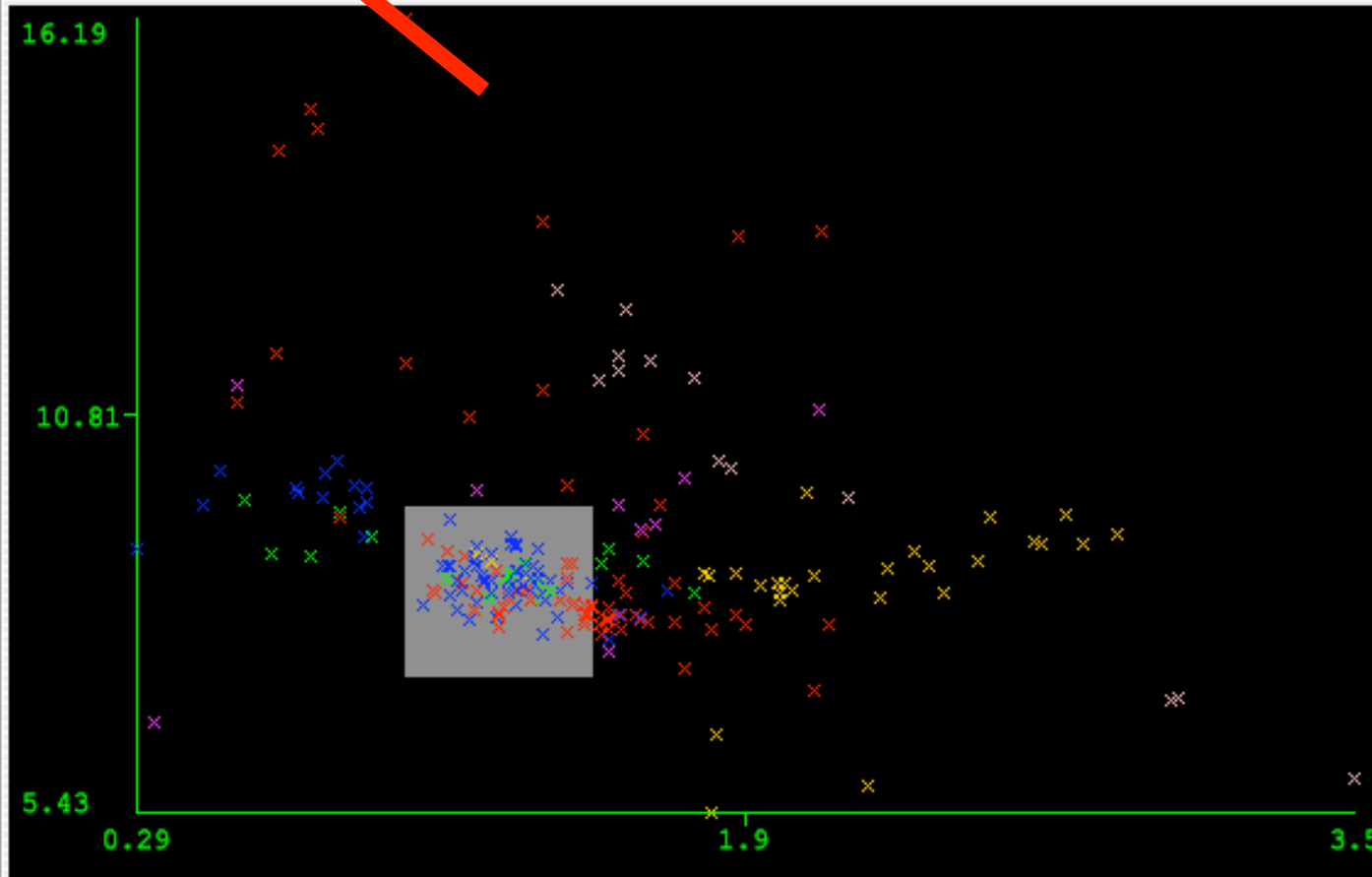
Clear

Save

Jitter



Plot: Glass



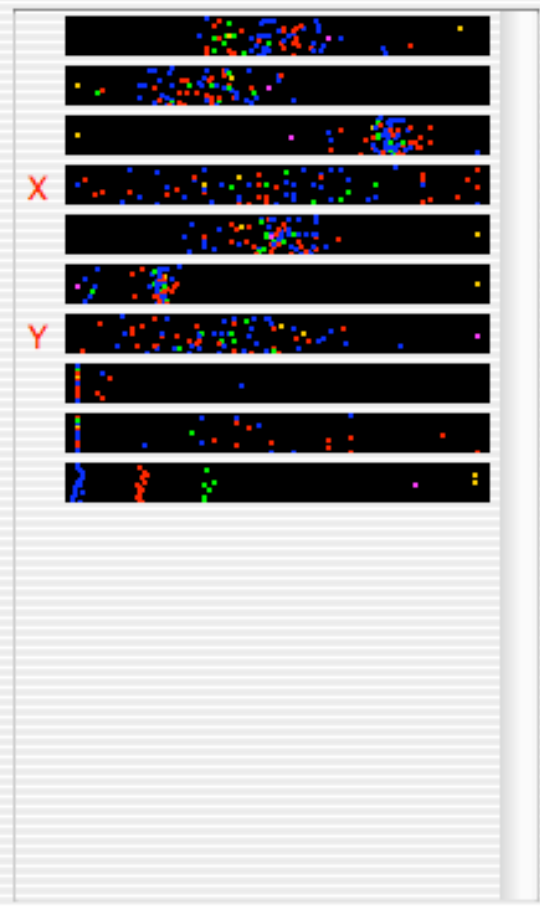
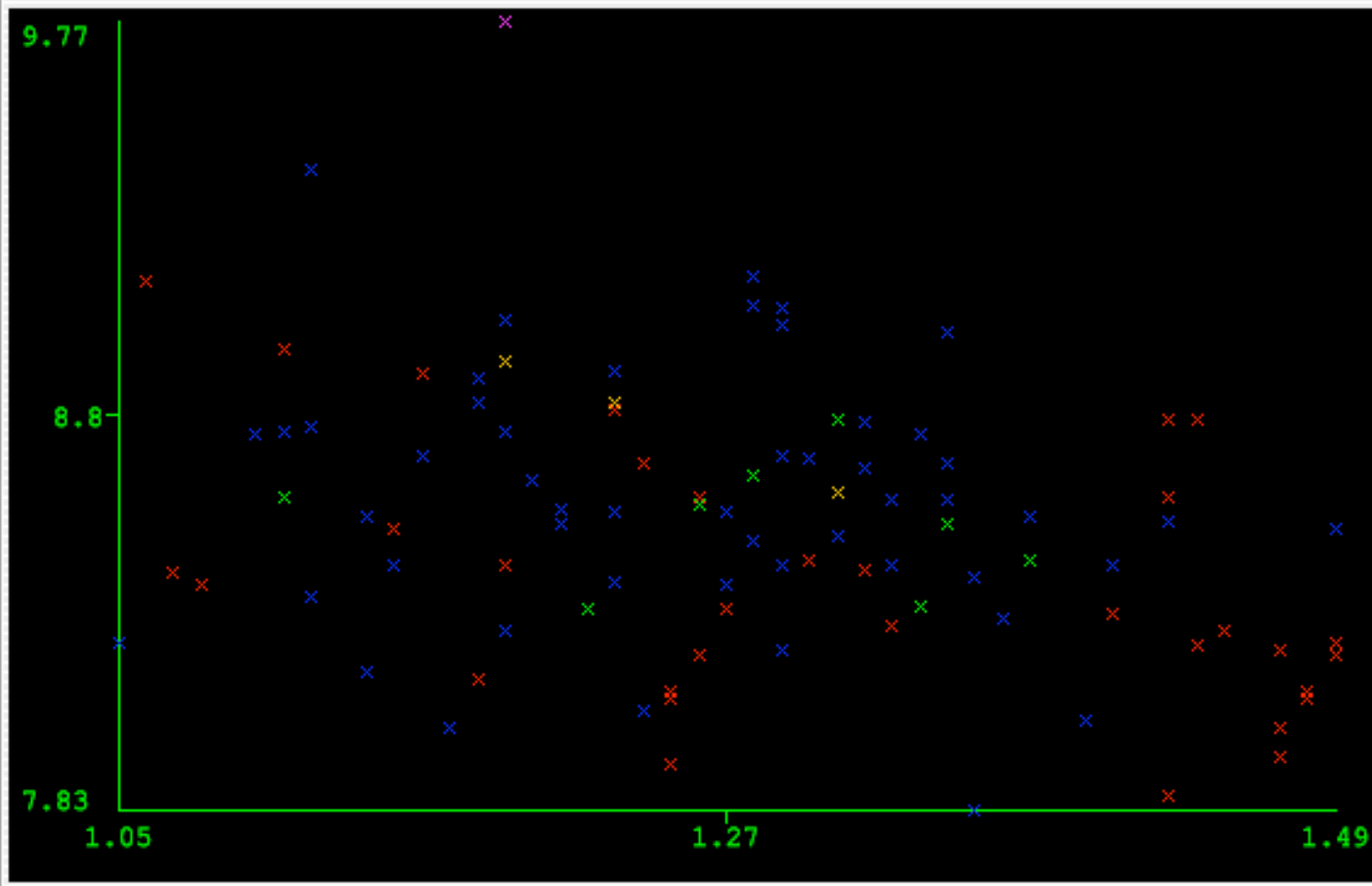
Class colour

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

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

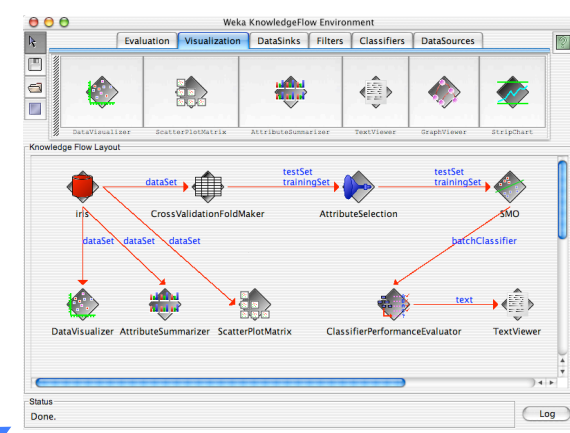
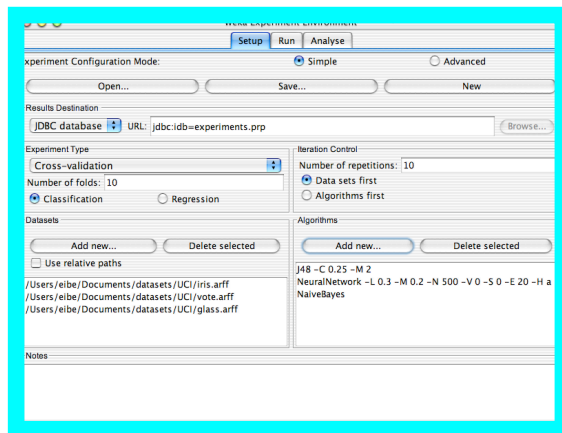
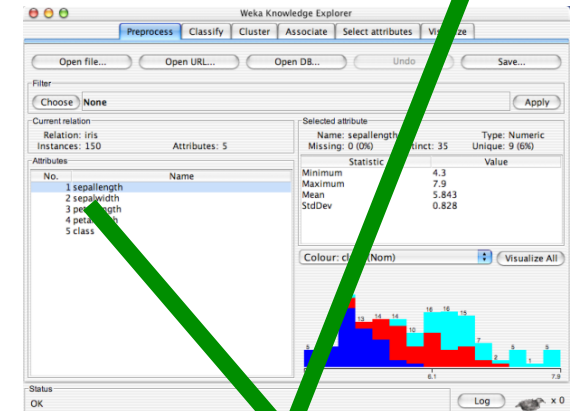
Reset Clear Save Jitter

Plot: Glass



Class colour

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



Performing Experiments



- Experimenter makes it easy to compare the performance of different learning schemes
- For classification and regression problems
- Results can be written into file or database
- Evaluation options: cross-validation, learning curve, hold-out
- Can also iterate over different parameter settings
- Significance-testing built in!

Weka Experiment Environment

Setup

Run

Analyse

Experiment Configuration Mode: ☒ Simple ☐ Advanced

Open...

Save...

New

Results Destination

JDBC database

Filename:

Browse...

Experiment Type

Cross-validation

Number of folds:

☒ Classification ☐ Regression

Iteration Control

Number of repetitions:

☒ Data sets first ☐ Algorithms first

Datasets

Add new...

Delete selected

☐ Use relative paths

Algorithms

Add new...

Delete selected

Notes

Weka Experiment Environment

Setup Run Analyse

Experiment Configuration Mode: ☒ Simple ☐ Advanced

Open... Save... New

Results Destination

JDBC database ▾ Filename: Browse...

Experiment Type

Cross-validation ▾

Number of folds:

☒ Classification ☐ Regression

Iteration Control

Number of repetitions:

☒ Data sets first ☐ Algorithms first

Datasets

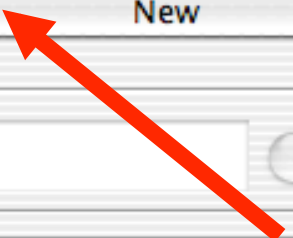
Add new... Delete selected

☐ Use relative paths

Algorithms

Add new... Delete selected

Notes



Weka Experiment Environment

SetupRunAnalyse

Experiment Configuration Mode: ☒ Simple ☐ Advanced

Open...Save...New

Results Destination

JDBC database URL: jdbc:ids=experiments.prp Browse...

Experiment Type

Cross-validation

Number of folds: 10

☒ Classification ☐ Regression

Iteration Control

Number of repetitions: 10

☒ Data sets first ☐ Algorithms first

Datasets

Add new...Delete selected

☐ Use relative paths

/Users/eibe/Documents/datasets/UCI/iris.arff
/Users/eibe/Documents/datasets/UCI/vote.arff
/Users/eibe/Documents/datasets/UCI/glass.arff

Algorithms

Add new...Delete selected

J48 -C 0.25 -M 2
NeuralNetwork -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a
NaiveBayes

Notes

Weka Experiment Environment

Setup Run Analyse

Experiment Configuration Mode: ☒ Simple ☐ Advanced

Open... Save... New

Results Destination

JDBC database URL: jdbc:ide=experiments.prp Browse...

Experiment Type

Cross-validation

Number of folds: 10

☒ Classification ☐ Regression

Iteration Control

Number of repetitions: 10

☒ Data sets first ☐ Algorithms first

Datasets

Add new... Delete selected

☐ Use relative paths

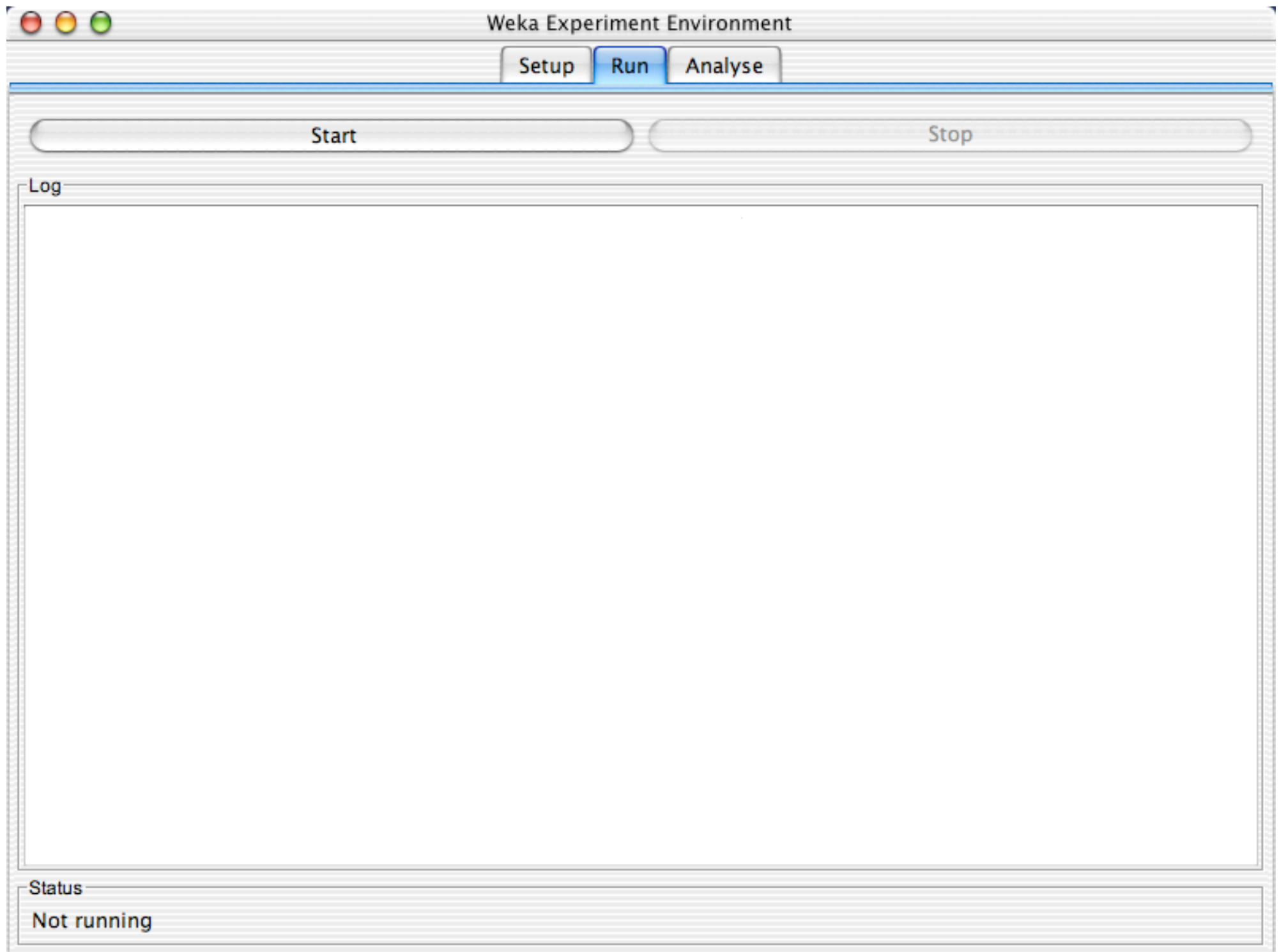
/Users/eibe/Documents/datasets/UCI/iris.arff
/Users/eibe/Documents/datasets/UCI/vote.arff
/Users/eibe/Documents/datasets/UCI/glass.arff

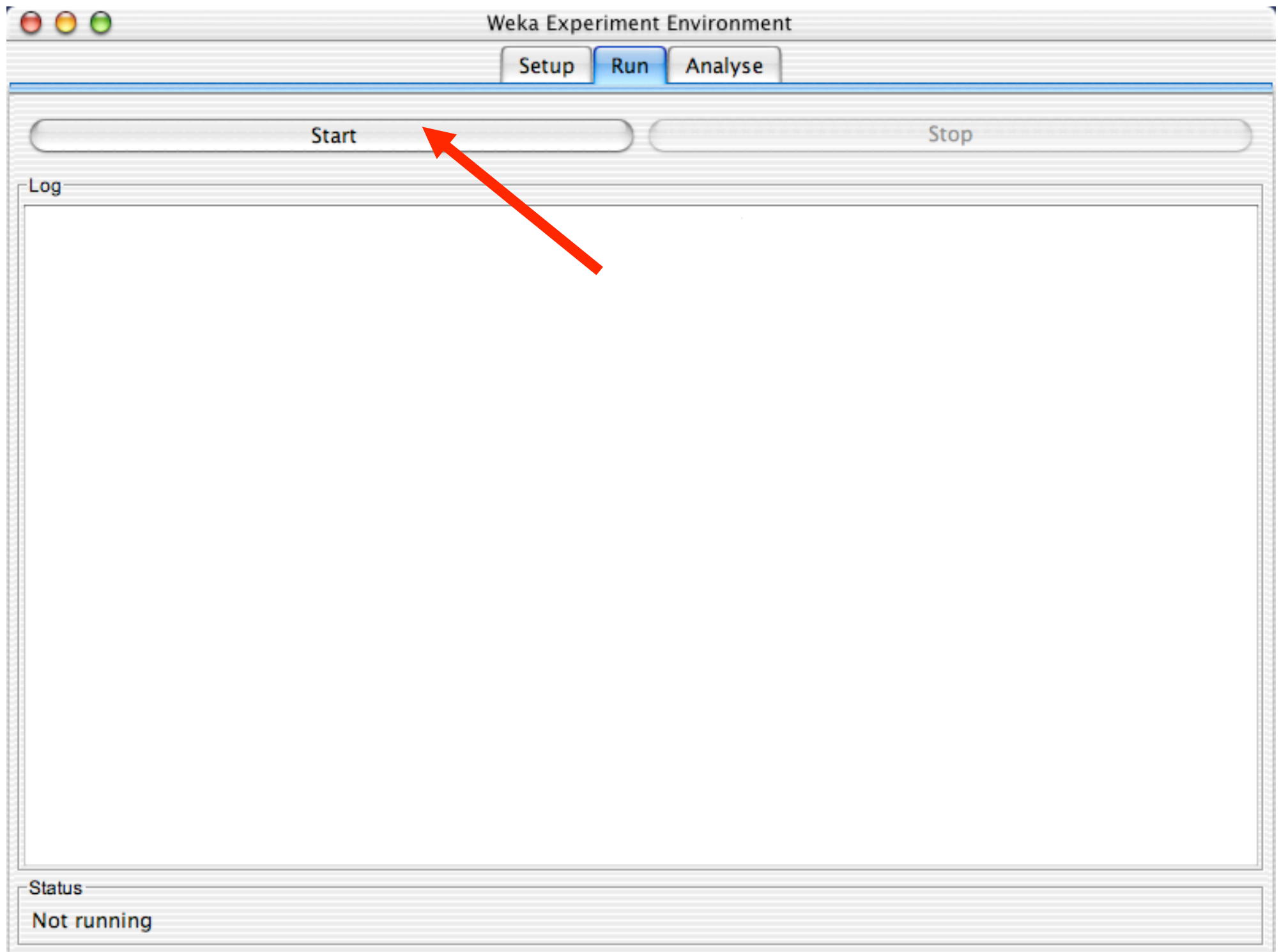
Algorithms

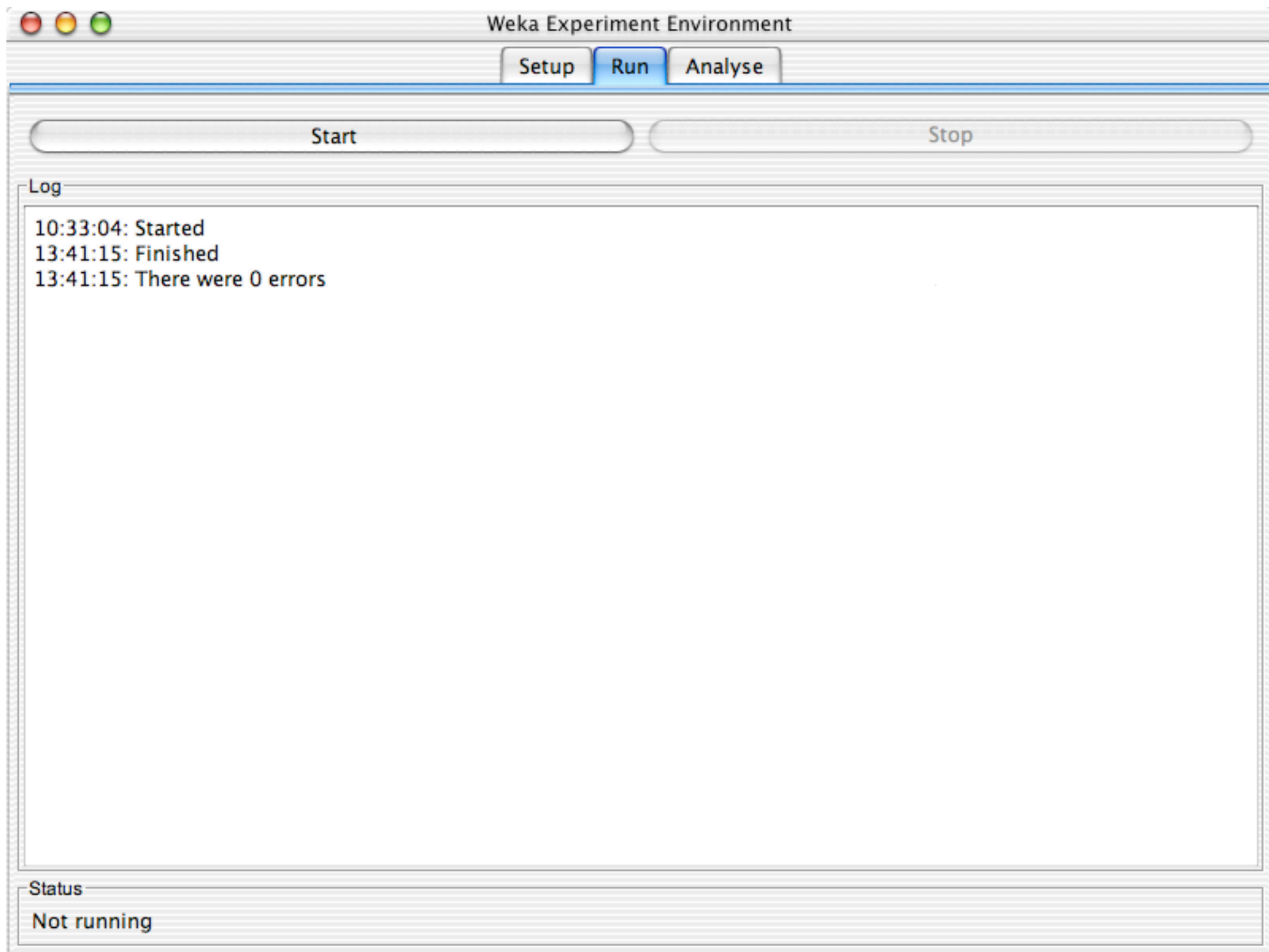
Add new... Delete selected

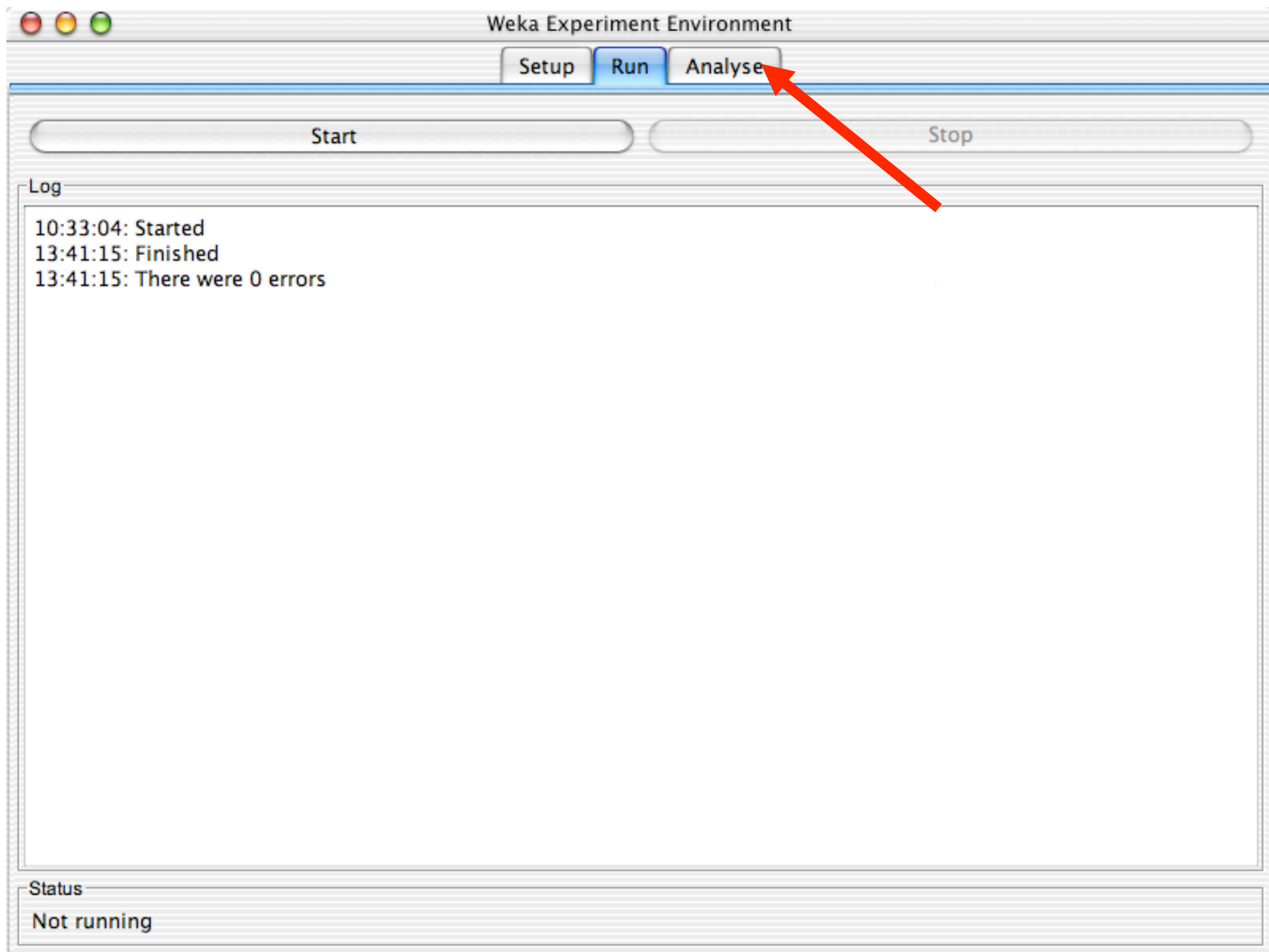
J48 -C 0.25 -M 2
NeuralNetwork -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a
NaiveBayes

Notes









Weka Experiment Environment

Setup

Run

Analyse

Source

No source

File...

Database...

Experiment

Configure test

Row key fields

Select keys...

Run field

Column key fields

Select keys...

Comparison field

Significance

0.05

Test base

Select base...

Show std. deviations

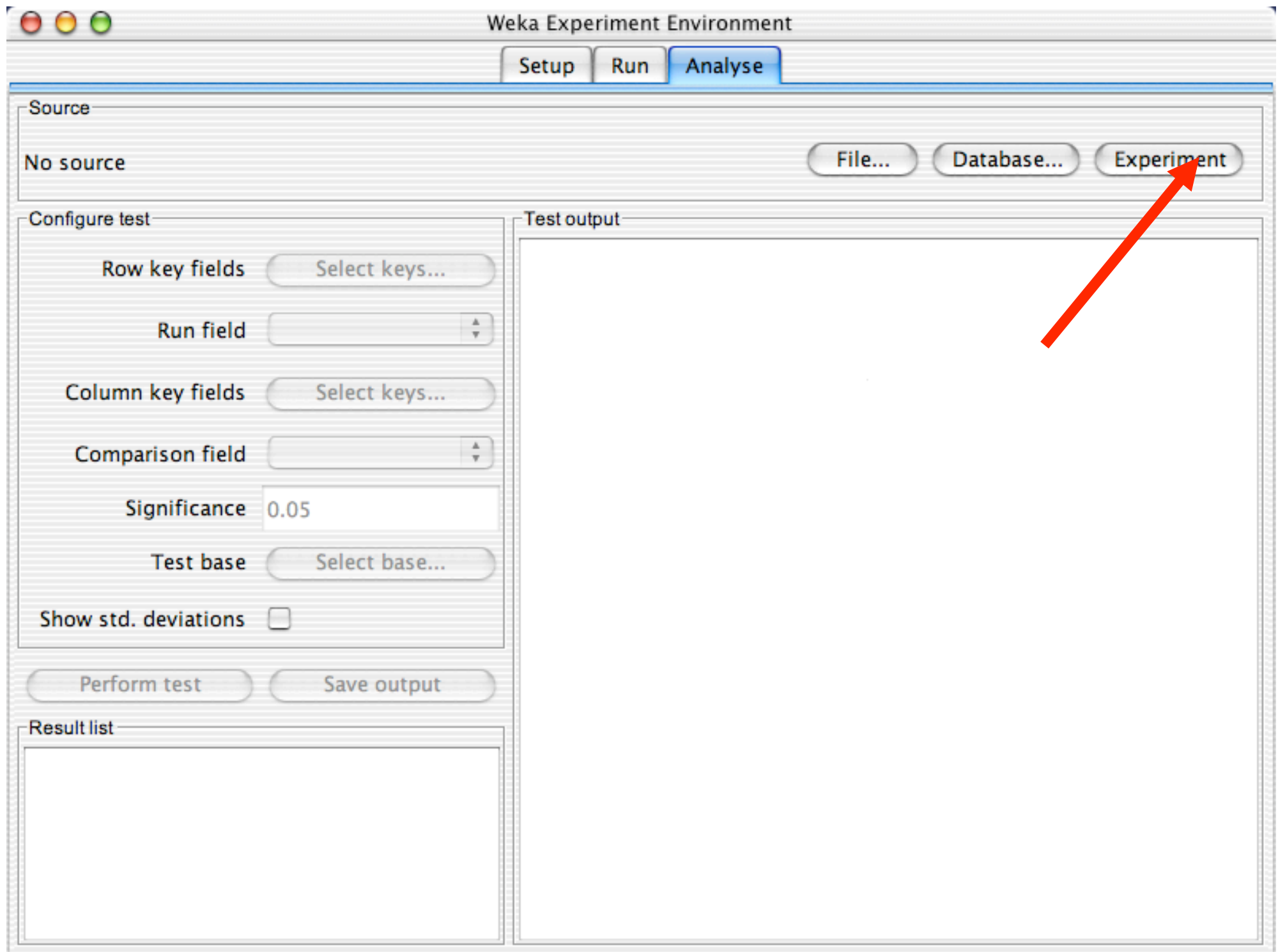
☐

Perform test

Save output

Result list

Test output



Weka Experiment Environment

SetupRunAnalyse

Source

Got 900 results

File...Database...Experiment

Configure test

Row key fields

Select keys...

Run field

Key_Run

Column key fields

Select keys...

Comparison field

Percent_correct

Significance

0.05

Test base

Select base...

Show std. deviations

☐

Perform test

Save output

Result list

13:44:17 - Available resultsets

13:44:55 - Percent_correct - trees.j48.J48 '-C 0

Test output

Analysing: Percent_correct

Datasets: 3

Resultsets: 3

Confidence: 0.05 (two tailed)

Date: 9/9/03 1:44 PM

Dataset	(1) trees.j4	(2) funct	(3) bayes
iris	(100) 94.73	96.4	95.53
vote	(100) 96.57	94.71 *	90.02 *
Glass	(100) 67.63	66.78	49.45 *

(v/ /*) | (0/2/1) (0/1/2)

Skipped:

Key:

(1) trees.j48.J48 '-C 0.25 -M 2' -217733168393644444

(2) functions.neural.NeuralNetwork '-L 0.3 -M 0.2 -N 500 -V 0 -S 0

(3) bayes.NaiveBayes '' 2029074699749330519

Conclusion: Try It Yourself!



- WEKA is available at
<http://www.cs.waikato.ac.nz/ml/weka>
- Launch WEKA:
 - Enter in the weka directory within a terminal
 - `java -Xmx2048m -jar weka.jar`