

An Introduction to WEKA

Contributed by Yizhou Sun
2008

Content

- What is WEKA?
- The Explorer:
 - Preprocess data
 - Classification
 - Clustering
 - Association Rules
 - Attribute Selection
 - Data Visualization
- References and Resources

What is WEKA?

- **Waikato Environment for Knowledge Analysis**
 - It's a data mining/machine learning tool developed by Department of Computer Science, University of Waikato, New Zealand.
 - Weka is also a bird found only on the islands of New Zealand.



Download and Install WEKA

- Website:
<http://www.cs.waikato.ac.nz/~ml/weka/index.html>
- Support multiple platforms (written in java):
 - Windows, Mac OS X and Linux

Main Features

- 49 data preprocessing tools
- 76 classification/regression algorithms
- 8 clustering algorithms
- 3 algorithms for finding association rules
- 15 attribute/subset evaluators + 10 search algorithms for feature selection

Main GUI

- Three graphical user interfaces
 - “The Explorer” (exploratory data analysis)
 - “The Experimenter” (experimental environment)
 - “The KnowledgeFlow” (new process model inspired interface)



Content

- What is WEKA?
- **The Explorer:**
 - Preprocess data
 - Classification
 - Clustering
 - Association Rules
 - Attribute Selection
 - Data Visualization
- References and Resources

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 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

...



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



Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Open file...

Open URL...

Open DB...

Undo

Save...

Filter

Choose

None

Apply

Current relation

Relation: None

Instances: None

Attributes: None

Selected attribute

Name: None

Missing: None

Type: None

Distinct: None

Unique: None

Attributes

Visualize All

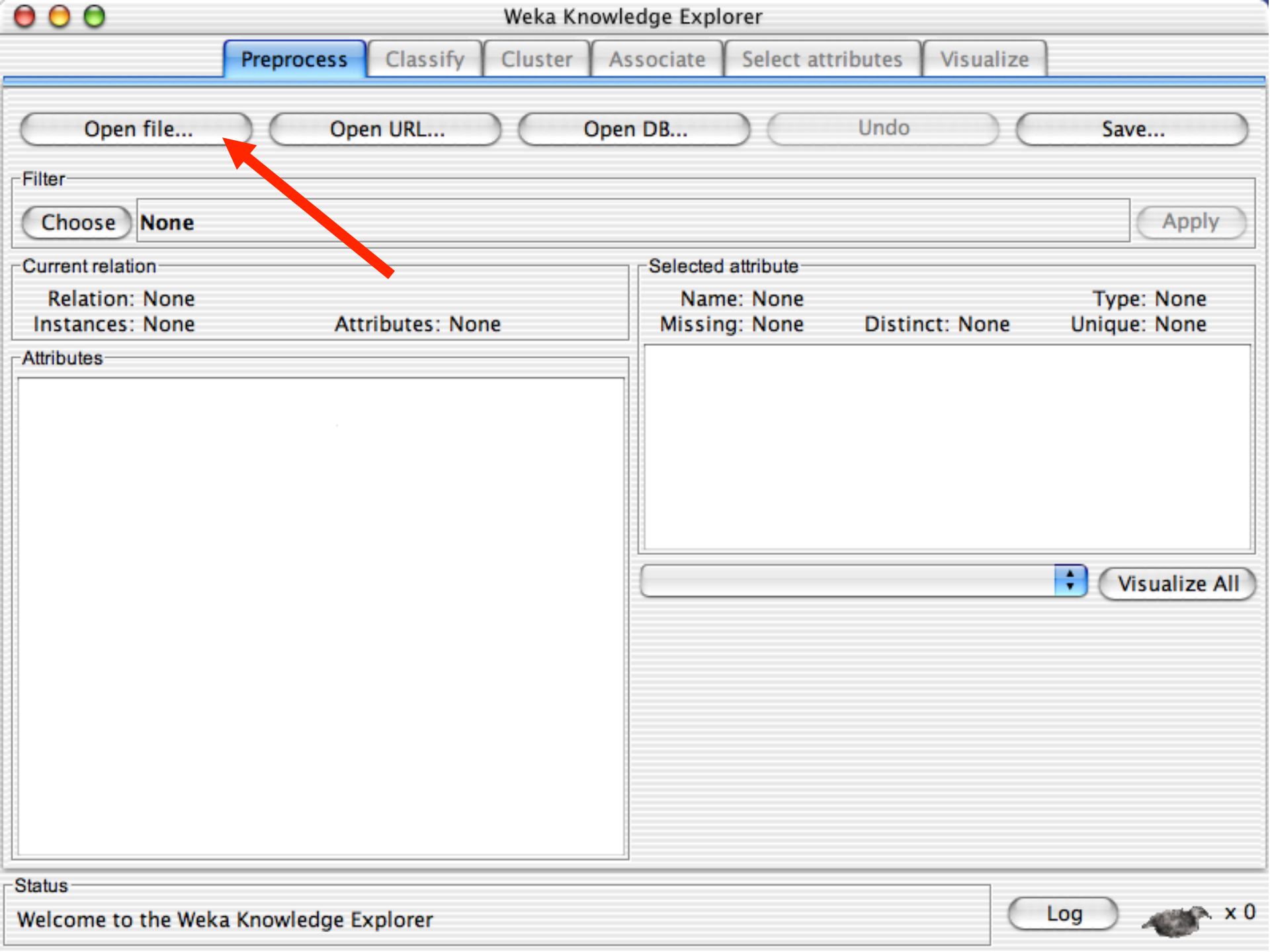
Status

Welcome to the Weka Knowledge Explorer

Log



x 0





Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Open file...

Open URL...

Open DB...

Undo

Save...

Filter

Choose

None

Apply

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%)

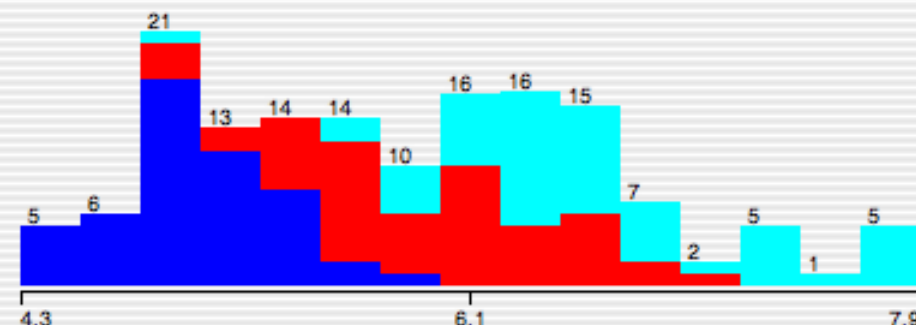
Distinct: 35

Unique: 9 (6%)

Statistic	Value
Minimum	4.3
Maximum	7.9
Mean	5.843
StdDev	0.828

Colour: class (Nom)

Visualize All

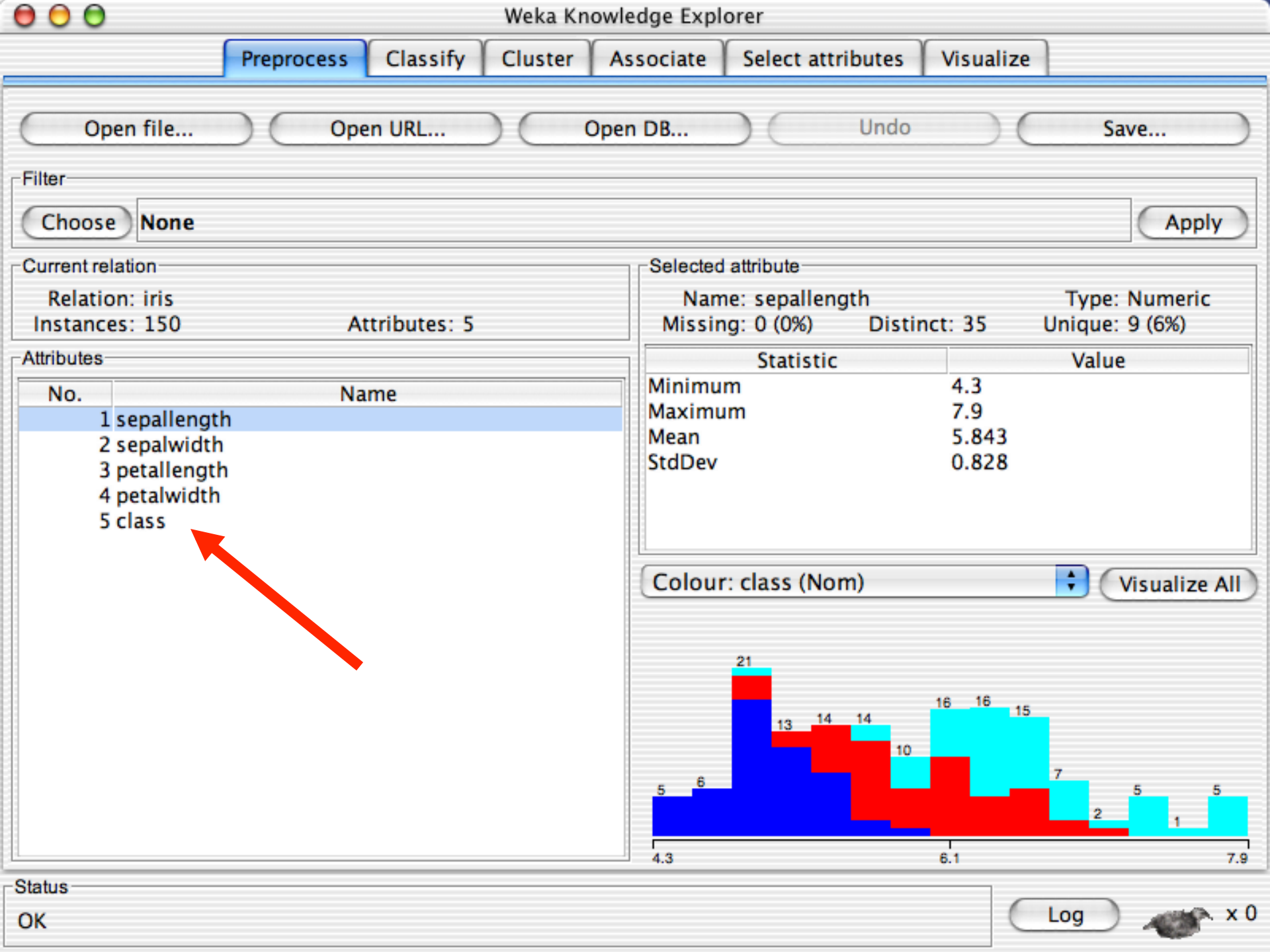


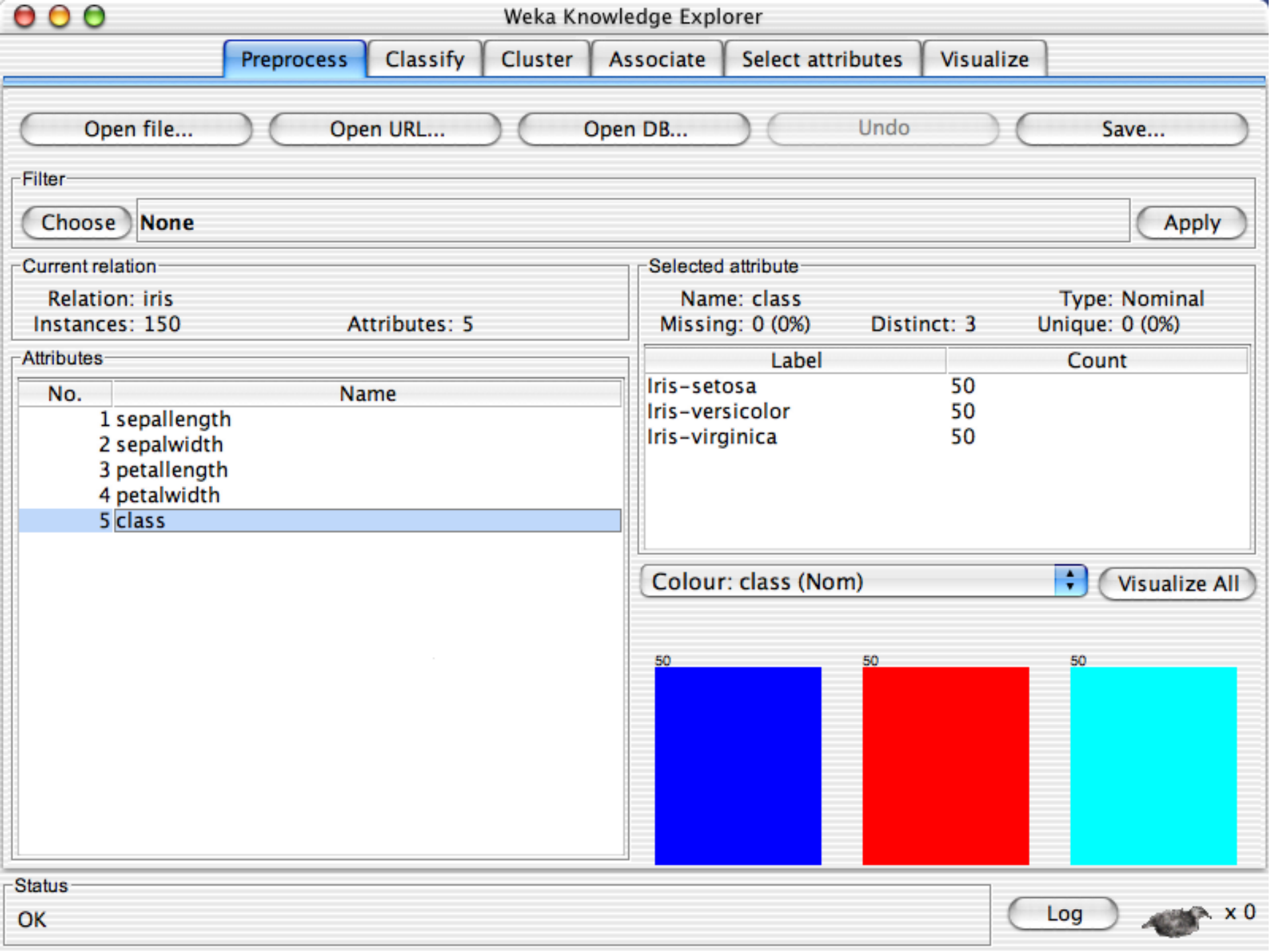
Status

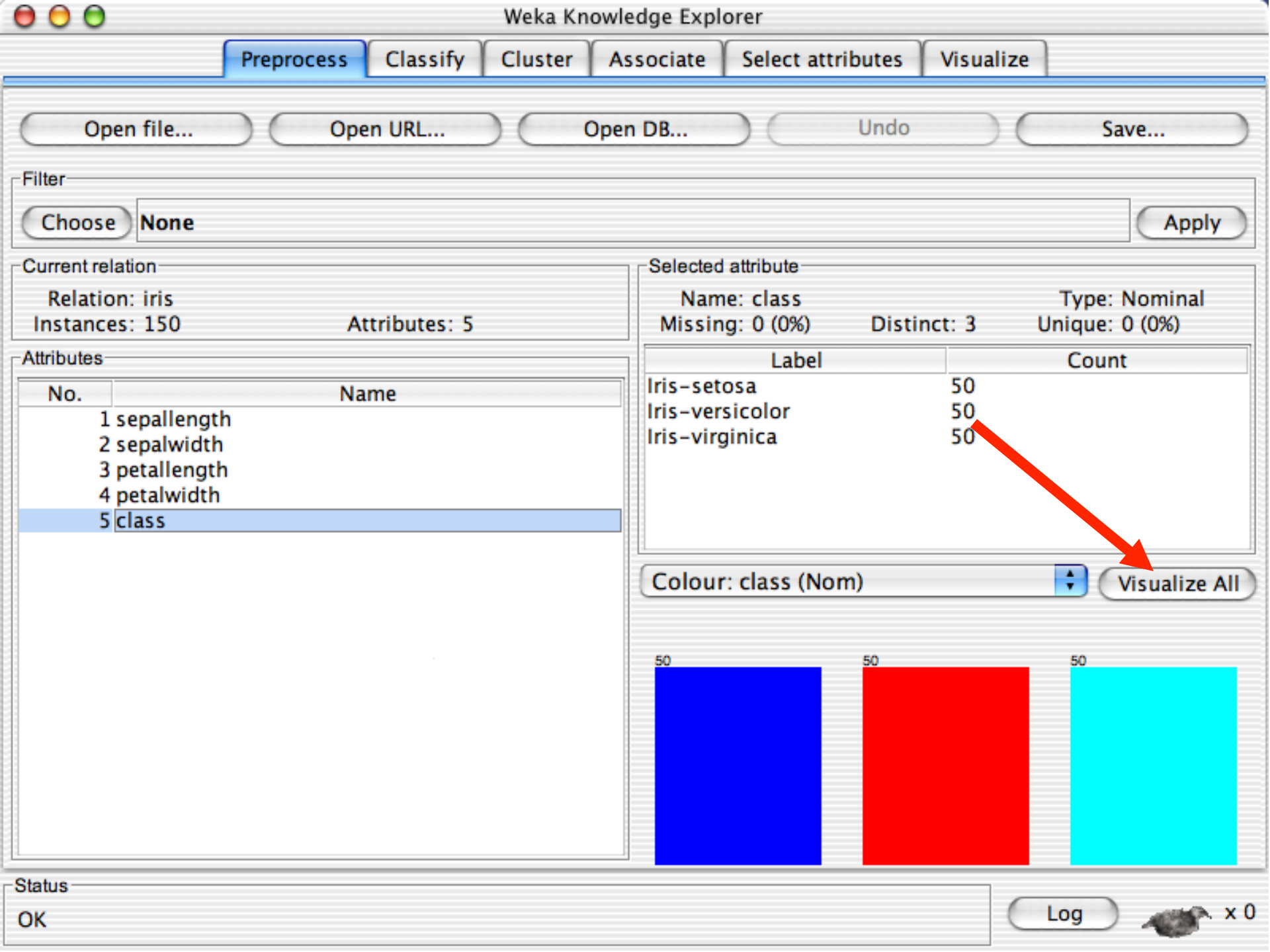
OK

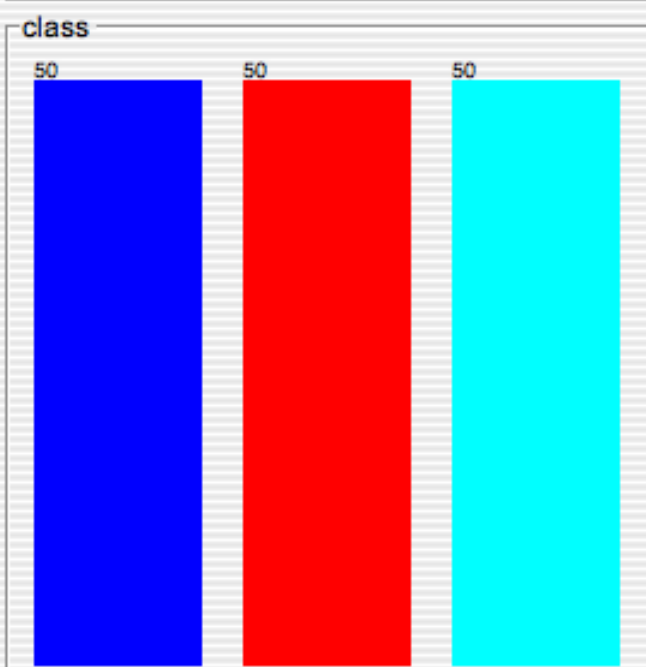
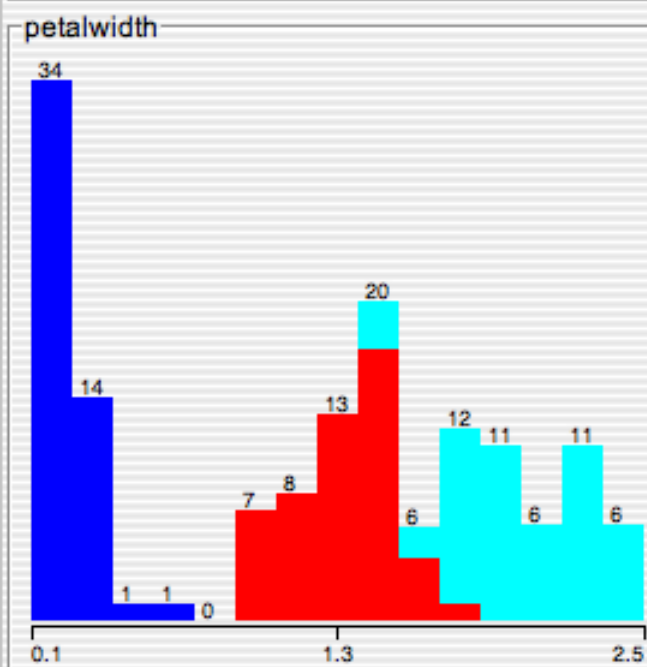
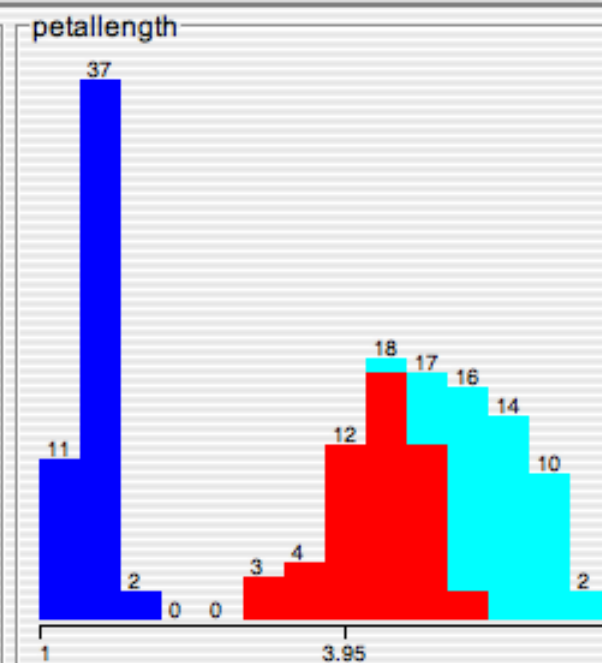
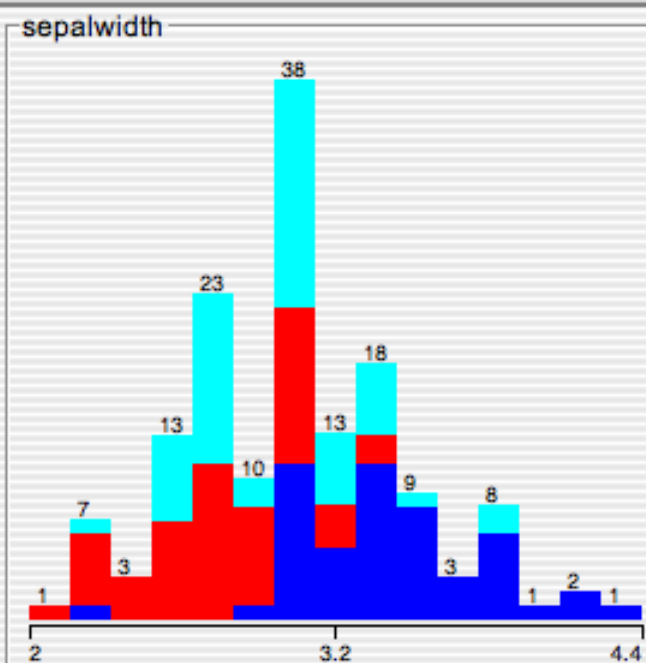
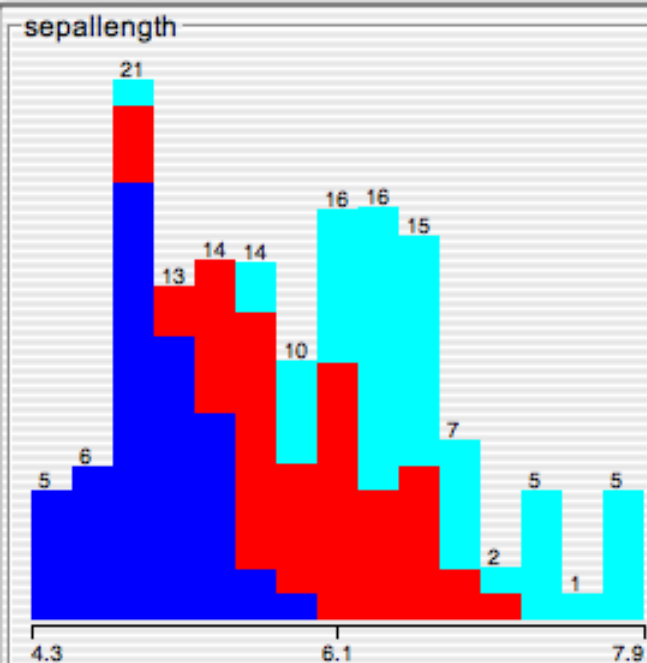
Log

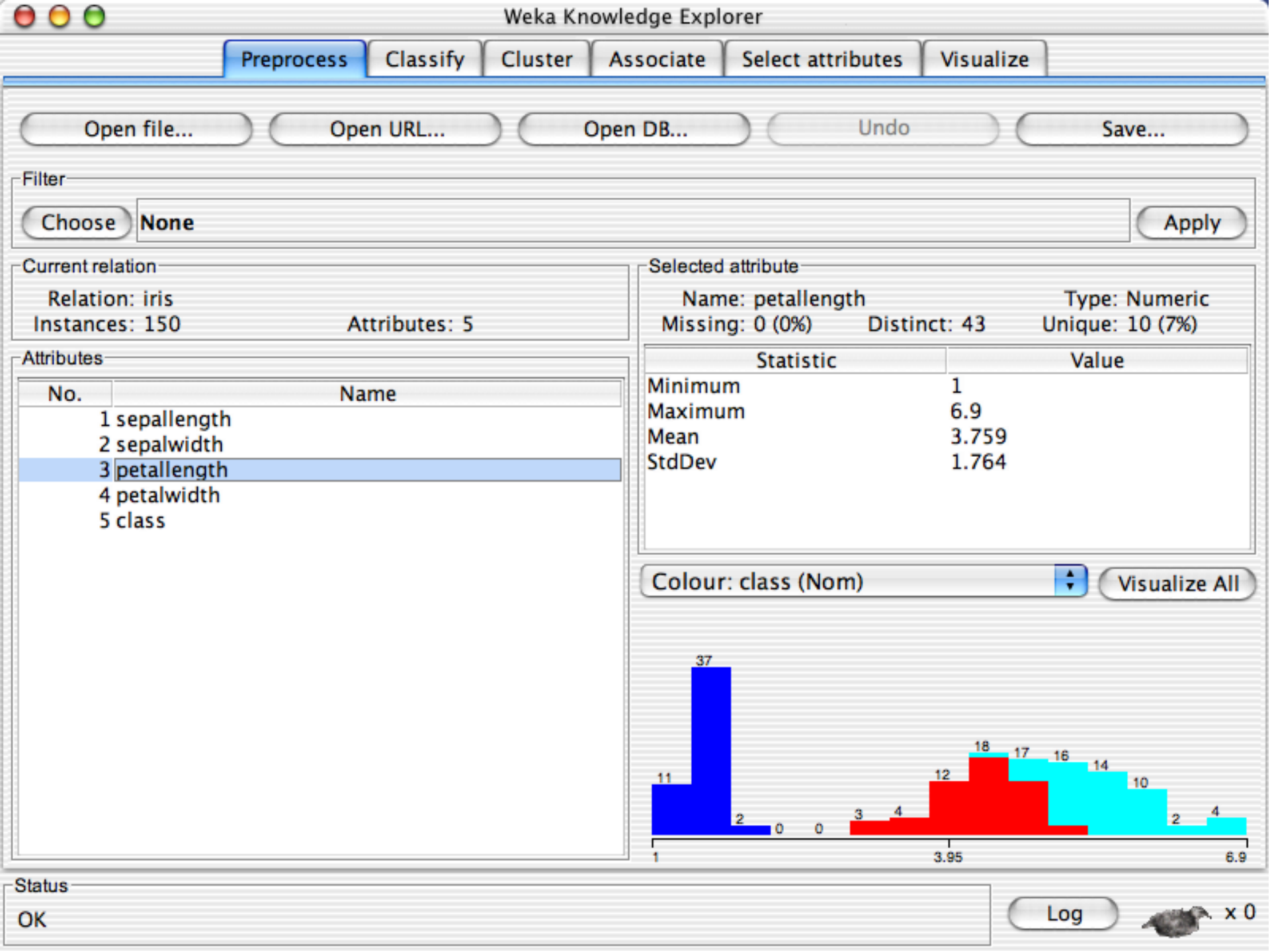
 x 0

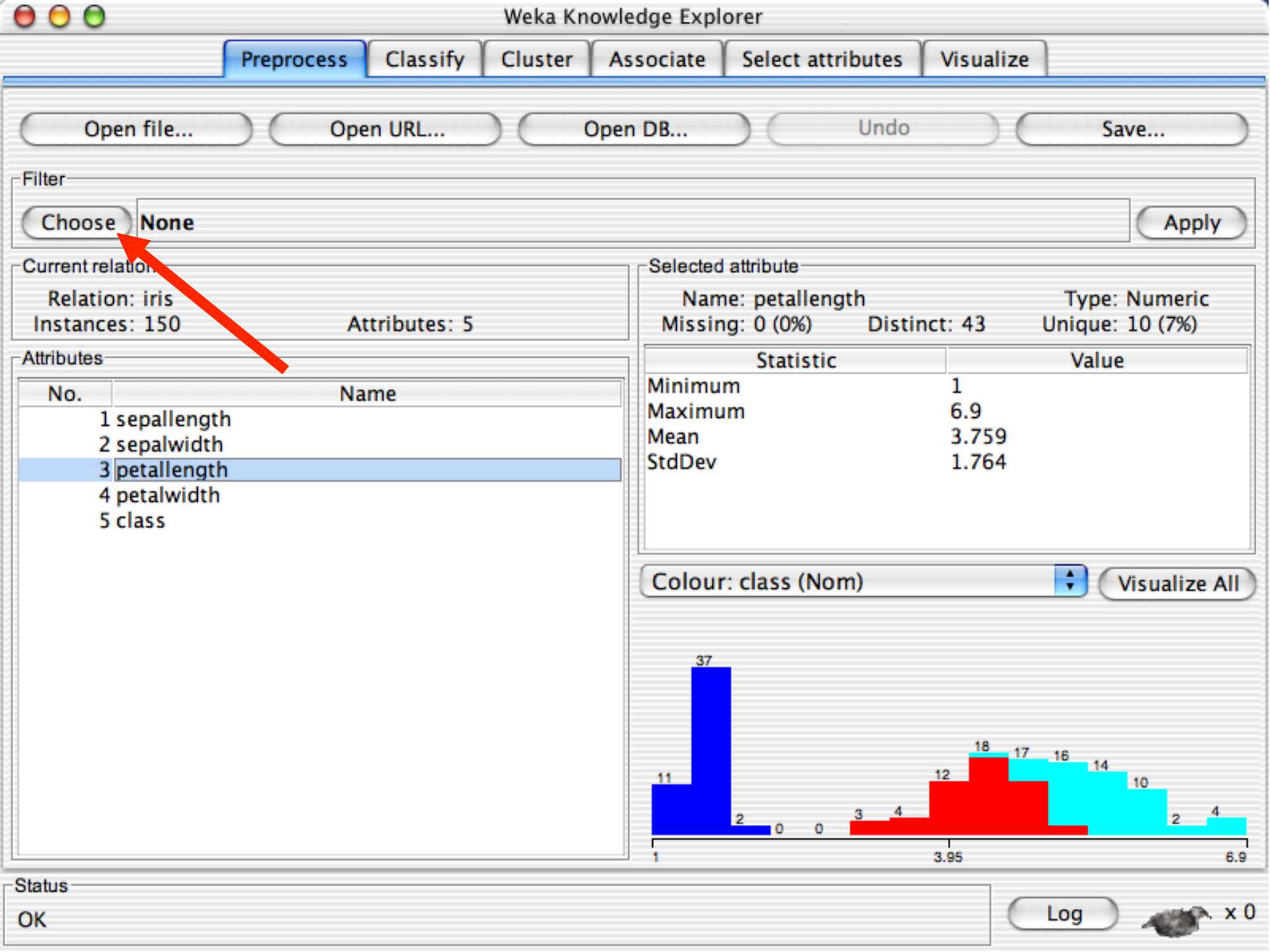














Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Open file...

Open URL...

Open DB...

Undo

Save...

Filter

weka

- filters
 - unsupervised
 - attribute
 - instance

Apply

Selected attribute

Name: petallength

Type: Numeric

Missing: 0 (0%)

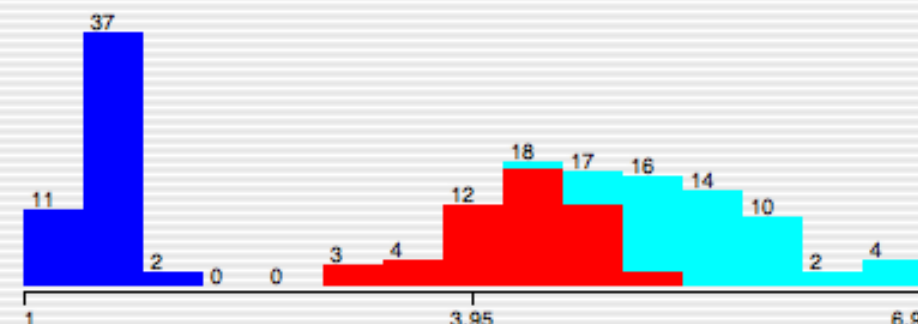
Distinct: 43

Unique: 10 (7%)

Statistic	Value
Minimum	1
Maximum	6.9
Mean	3.759
StdDev	1.764

Colour: class (Nom)

Visualize All



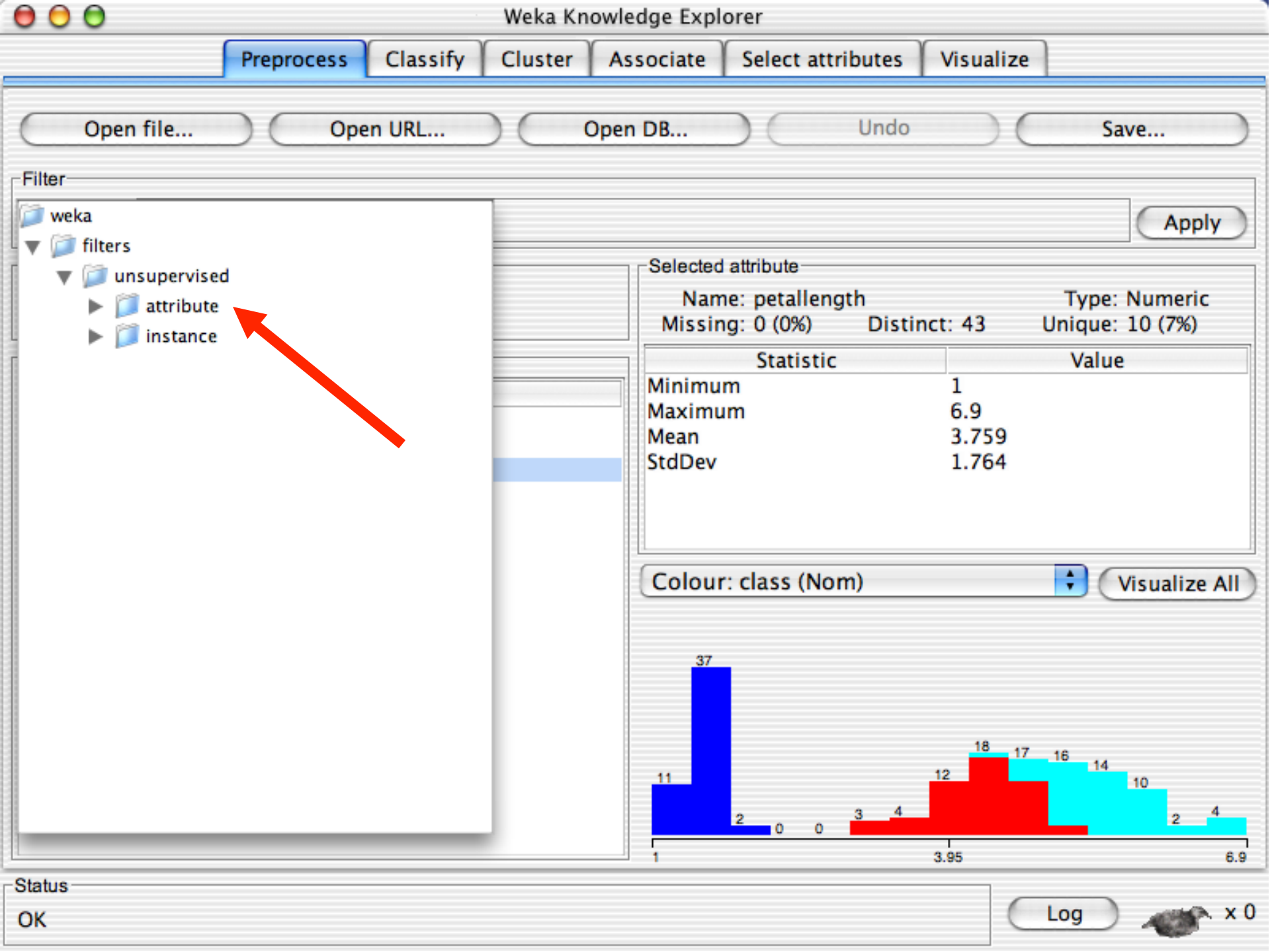
Status

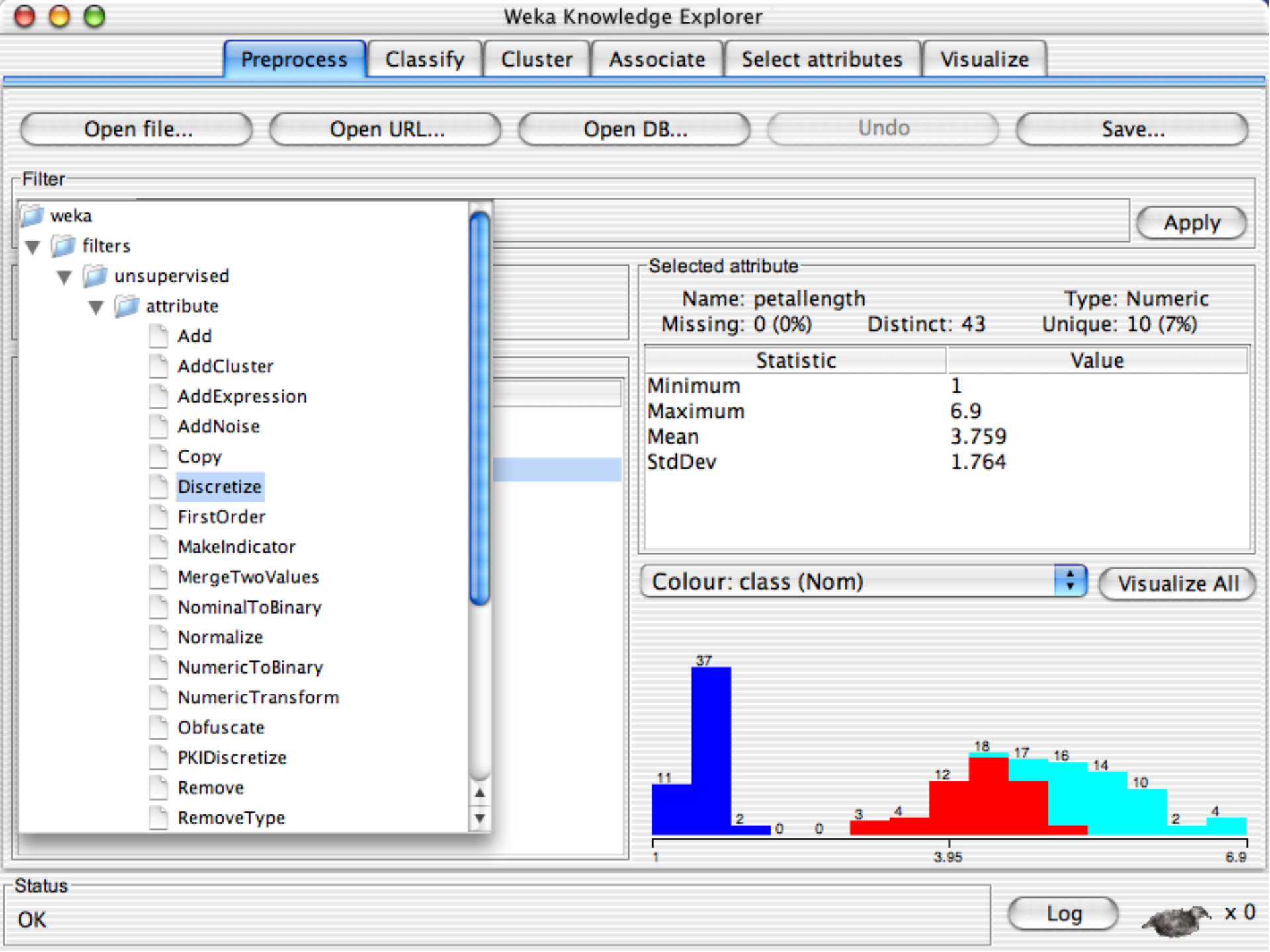
OK

Log



x 0







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: 5

Attributes

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

Selected attribute

Name: petal.length

Type: Numeric

Missing: 0 (0%)

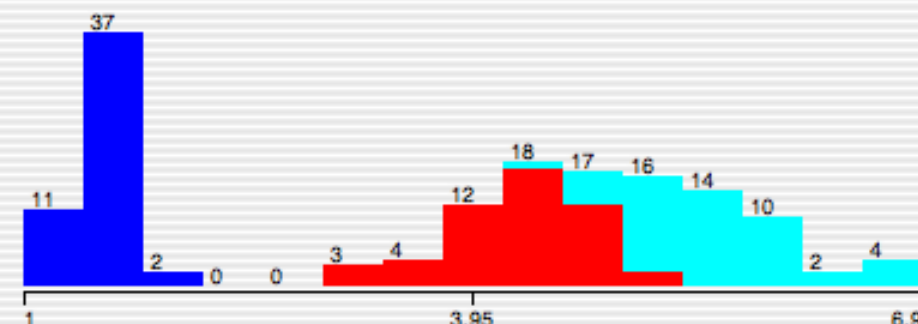
Distinct: 43

Unique: 10 (7%)

Statistic	Value
Minimum	1
Maximum	6.9
Mean	3.759
StdDev	1.764

Colour: class (Nom)

Visualize All



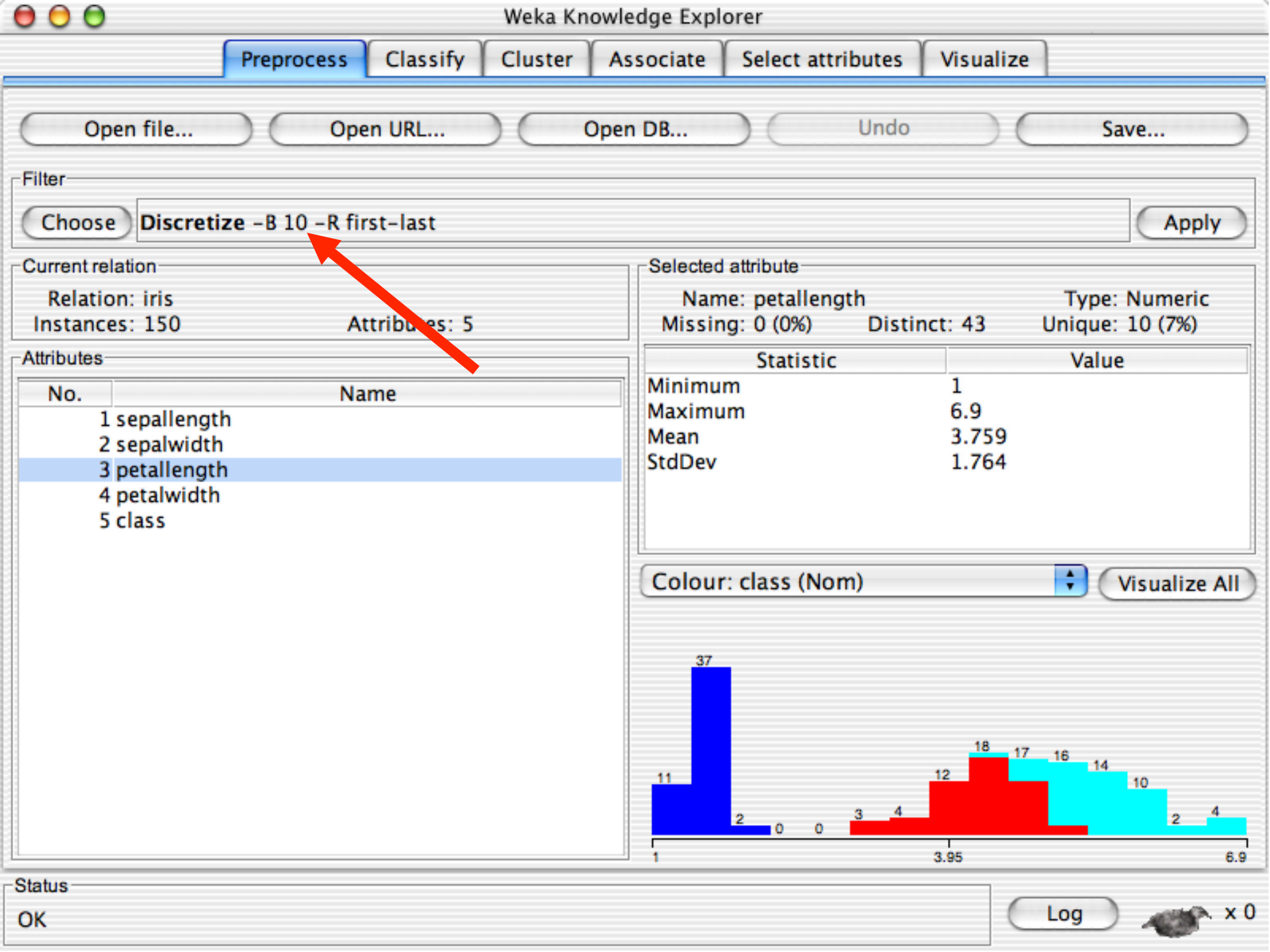
Status

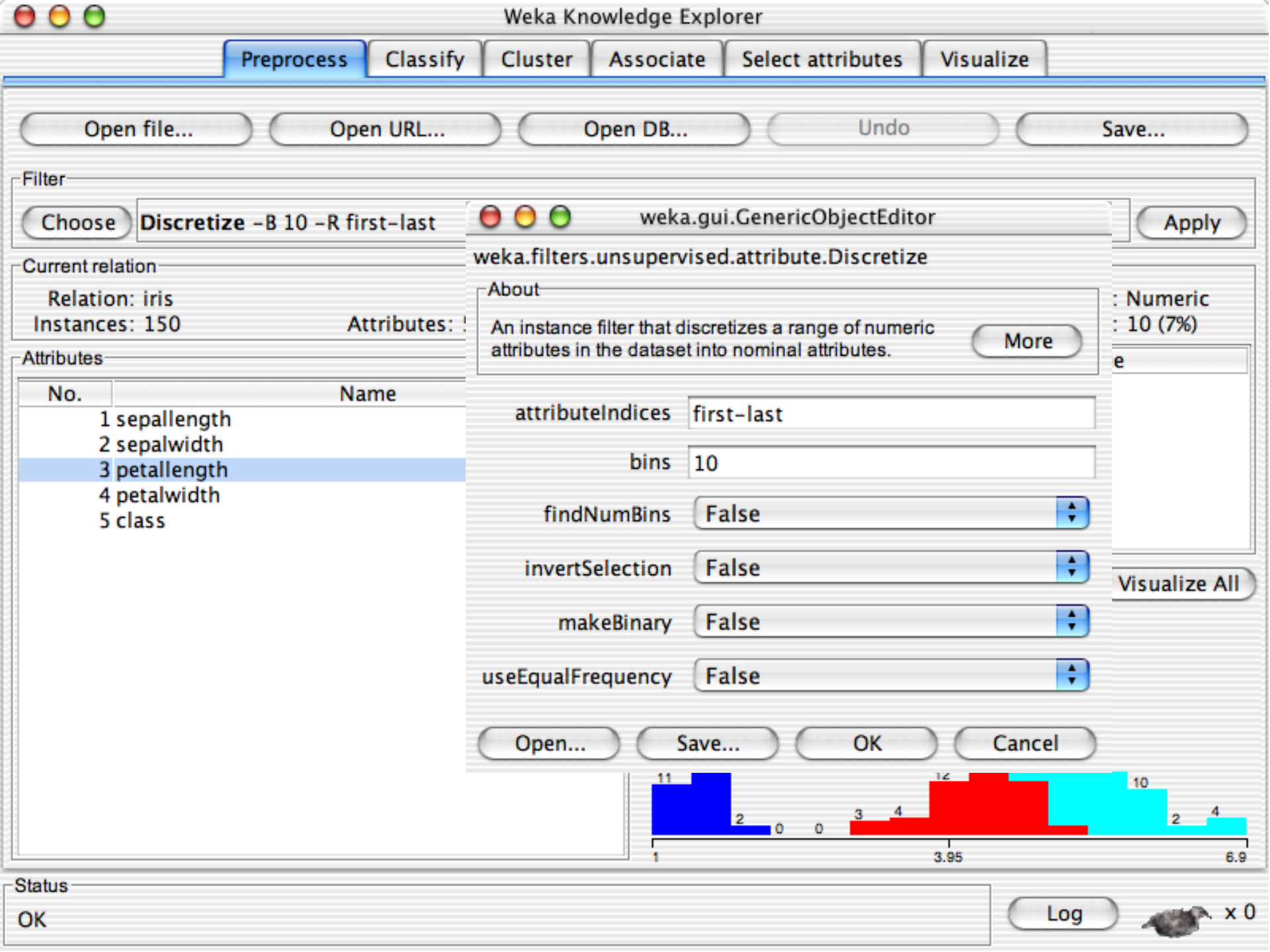
OK

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: 5

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

attributeIndices

first-last

bins

10

findNumBins

False

invertSelection

False

makeBinary

False

useEqualFrequency

False

Open...

Save...

OK

Cancel

Visualize All

1120034161024

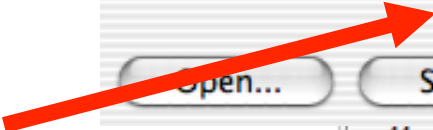
13.956.9

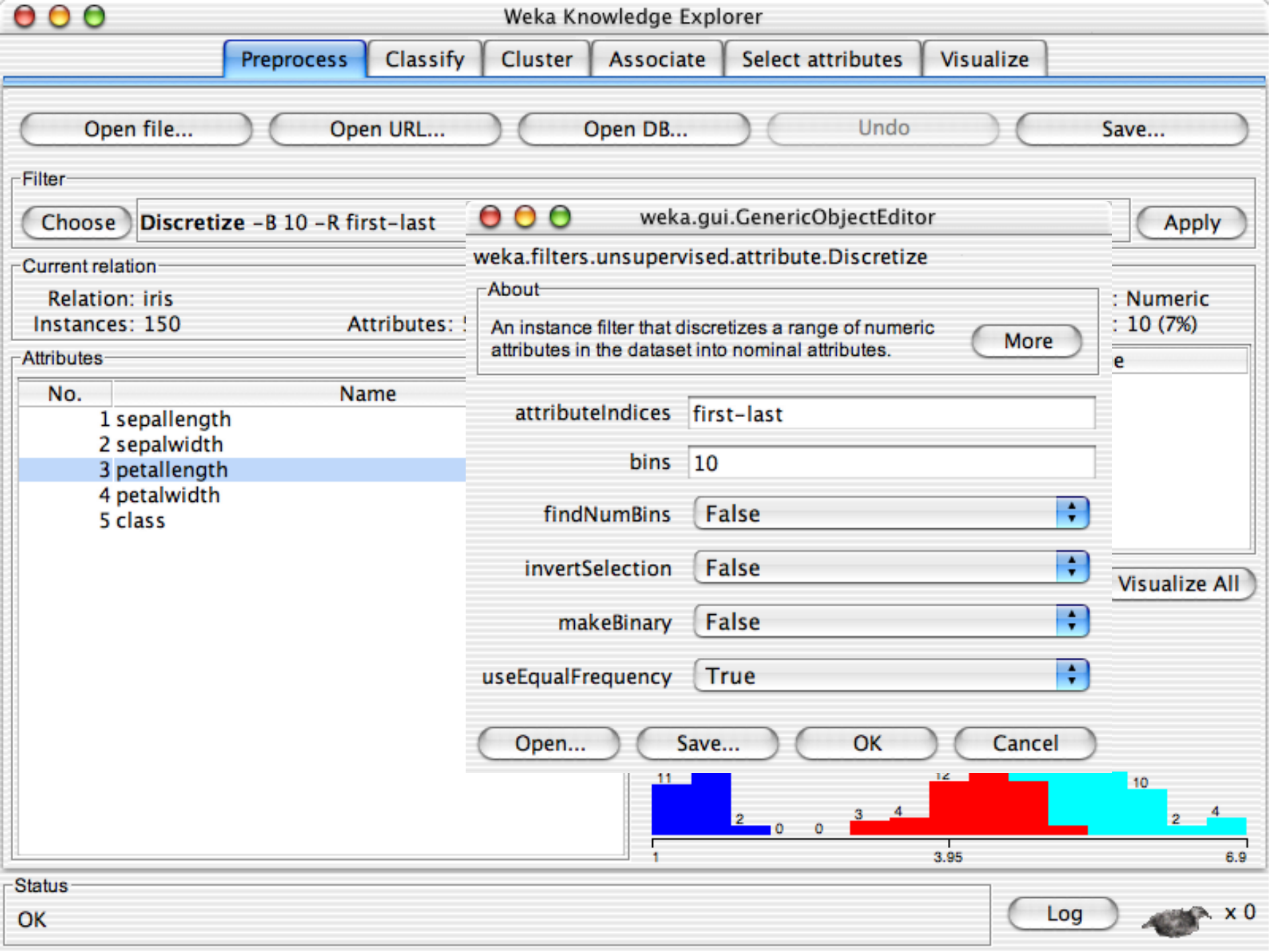
Status

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: irisInstances: 150Attributes: !

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

attributeIndices

first-last

bins

10

findNumBins

False

invertSelection

False

makeBinary

False

useEqualFrequency

True

Open...

Save...

OK

Cancel

Visualize All

1120034161024

13.956.9

Status

OK

Log

x 0

Filter

Current relation

Attributes

Status

OK

Relation: iris
Instances: 150

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

weka.filters.unsupervised.attribute.Discretize

About

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

attributeIndicesfirst-last

bins10

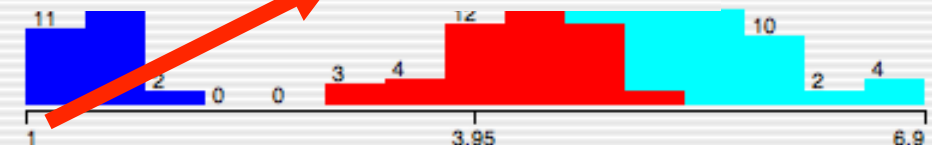
findNumBinsFalse

invertSelectionFalse

makeBinaryFalse

useEqualFrequencyTrue

Open...Save...OKCancel

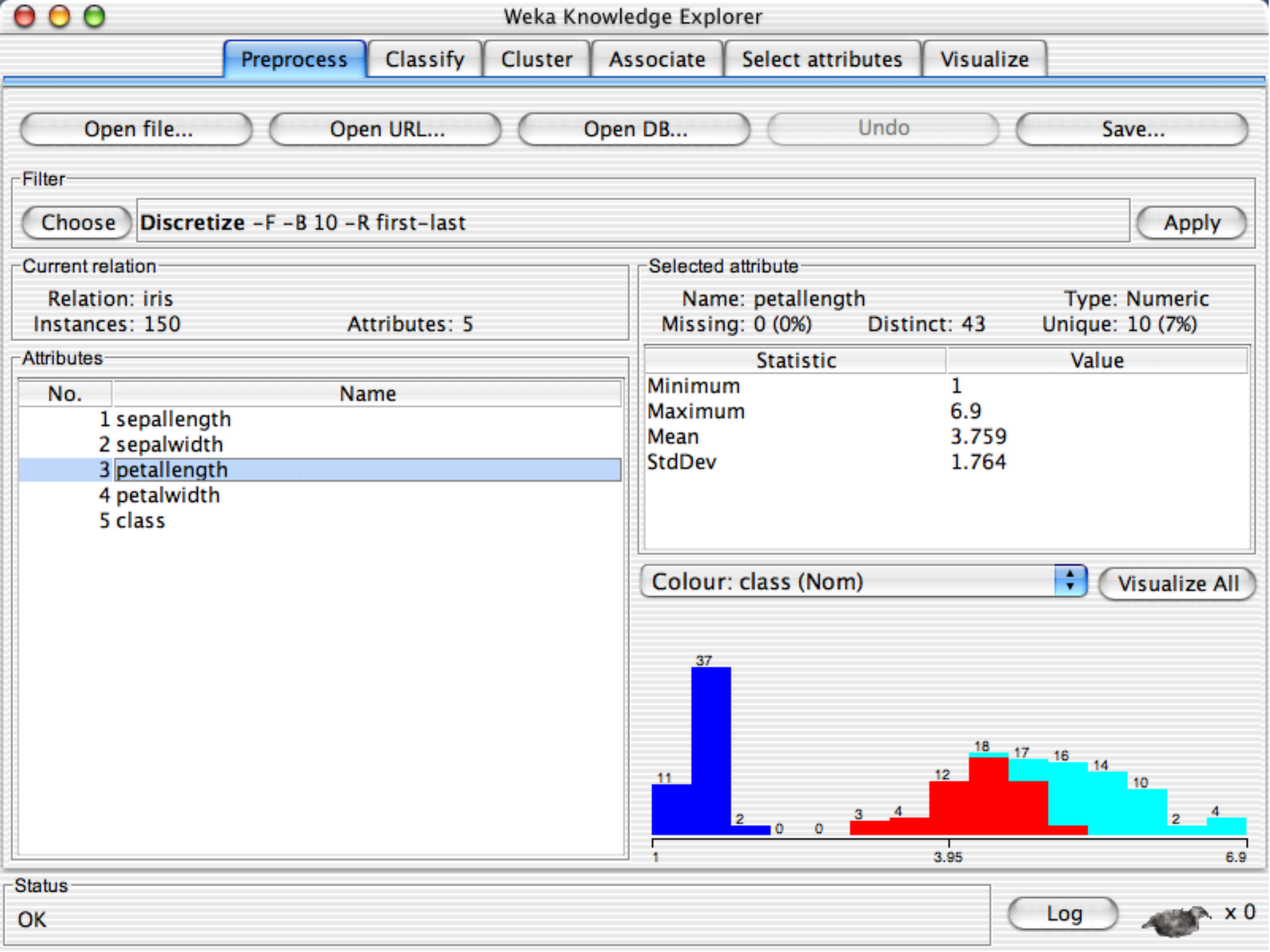


: Numeric
: 10 (7%)

Visualize All

Log

x 0



Selected attribute

Name: petal.length

Missing: 0 (0%)

Distinct: 43

Type: Numeric

Unique: 10 (7%)

Statistic	Value
Minimum	1
Maximum	6.9
Mean	3.759
StdDev	1.764

Colour: class (Nom)

Visualize All

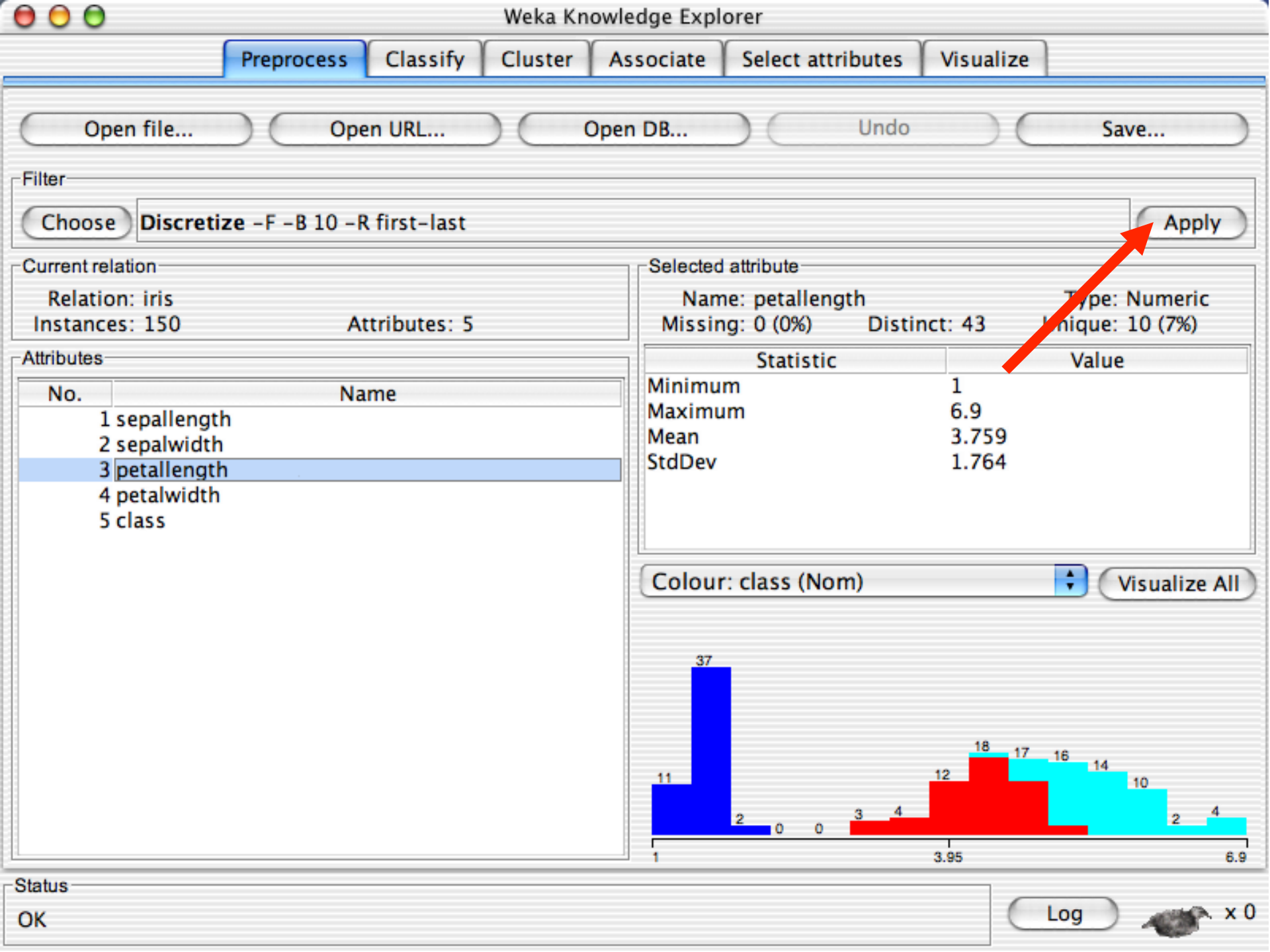
Bin Range	Frequency	Class
1.0 - 2.0	11	1
2.0 - 3.0	37	1
3.0 - 4.0	2	2
4.0 - 5.0	0	2
5.0 - 6.0	3	2
6.0 - 7.0	4	2
7.0 - 8.0	12	2
8.0 - 9.0	18	2
9.0 - 10.0	17	3
10.0 - 11.0	16	3
11.0 - 12.0	14	3
12.0 - 13.0	10	3
13.0 - 14.0	2	3
14.0 - 15.0	4	3

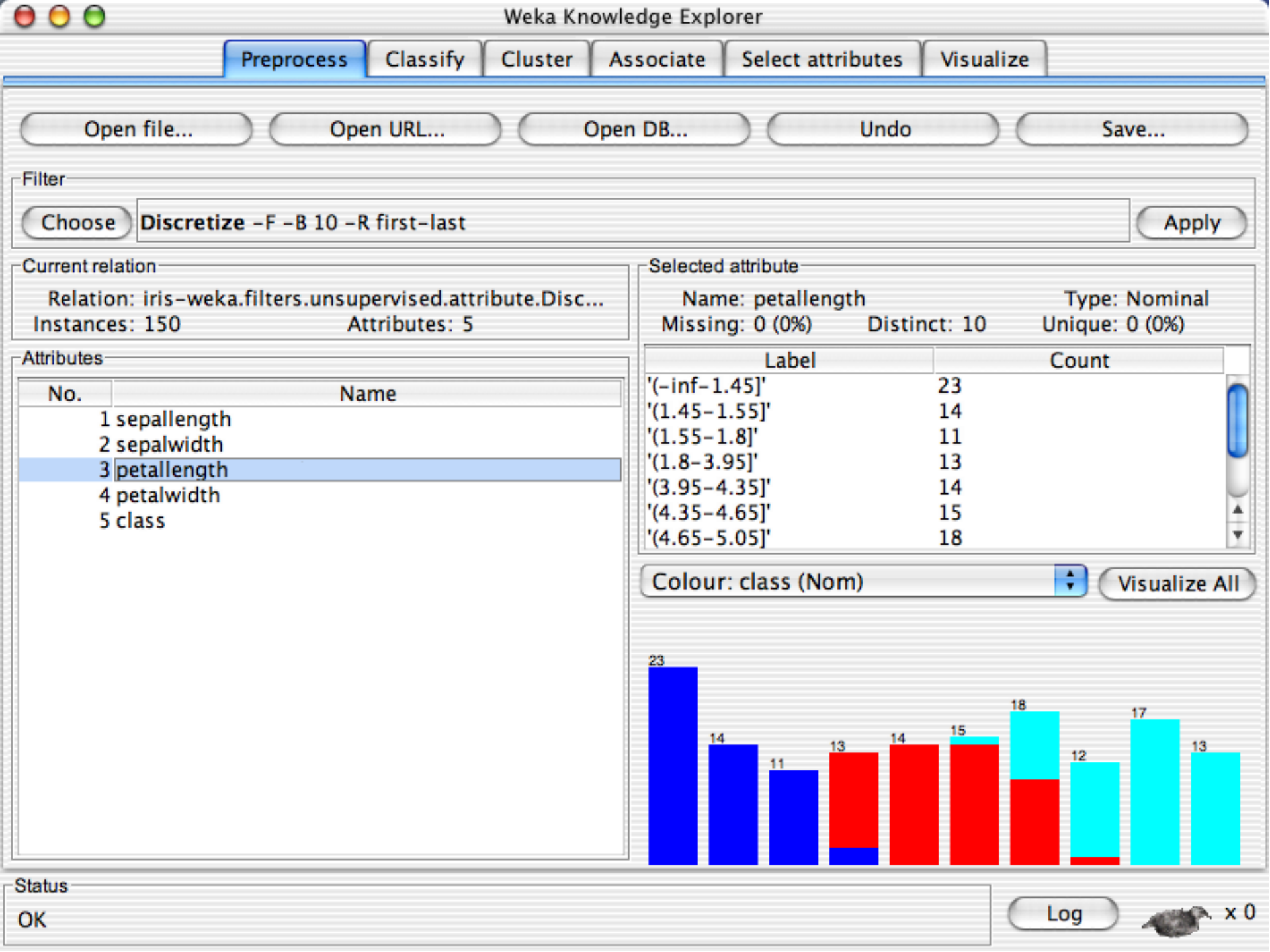
Status

OK

Log

x 0





Explorer: building “classifiers”

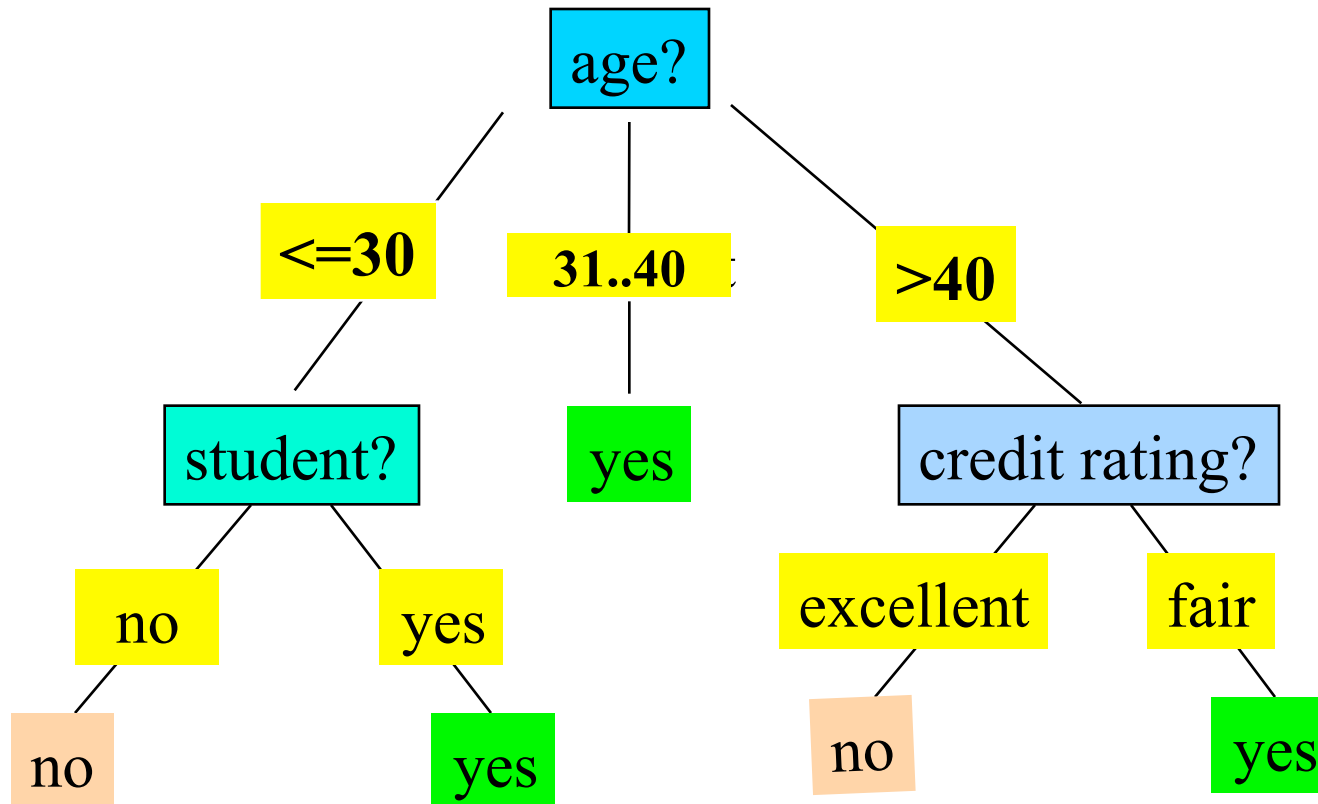
- Classifiers in WEKA are models for predicting nominal or numeric quantities
- Implemented learning schemes include:
 - **Decision trees** and lists, instance-based classifiers, support vector machines, multi-layer perceptrons, logistic regression, Bayes' nets, ...

Decision Tree Induction: Training Dataset

This follows
an example
of Quinlan's
ID3 (Playing
Tennis)

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31...40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31...40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
>40	medium	no	excellent	no

Output: A Decision Tree for “buys_computer”



Algorithm for Decision Tree Induction

- Basic algorithm (a greedy algorithm)
 - Tree is constructed in a **top-down recursive divide-and-conquer manner**
 - At start, all the training examples are at the root
 - Attributes are categorical (if continuous-valued, they are discretized in advance)
 - Examples are partitioned recursively based on selected attributes
 - Test attributes are selected on the basis of a heuristic or statistical measure (e.g., **information gain**)



Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Classifier

Choose

ZeroR

Test options

☐ Use training set☐ Supplied test set

Set...

☒ Cross-validation Folds 10☐ Percentage split % 66

More options...

(Nom) class



Start

Stop

Result list (right-click for options)

Classifier output

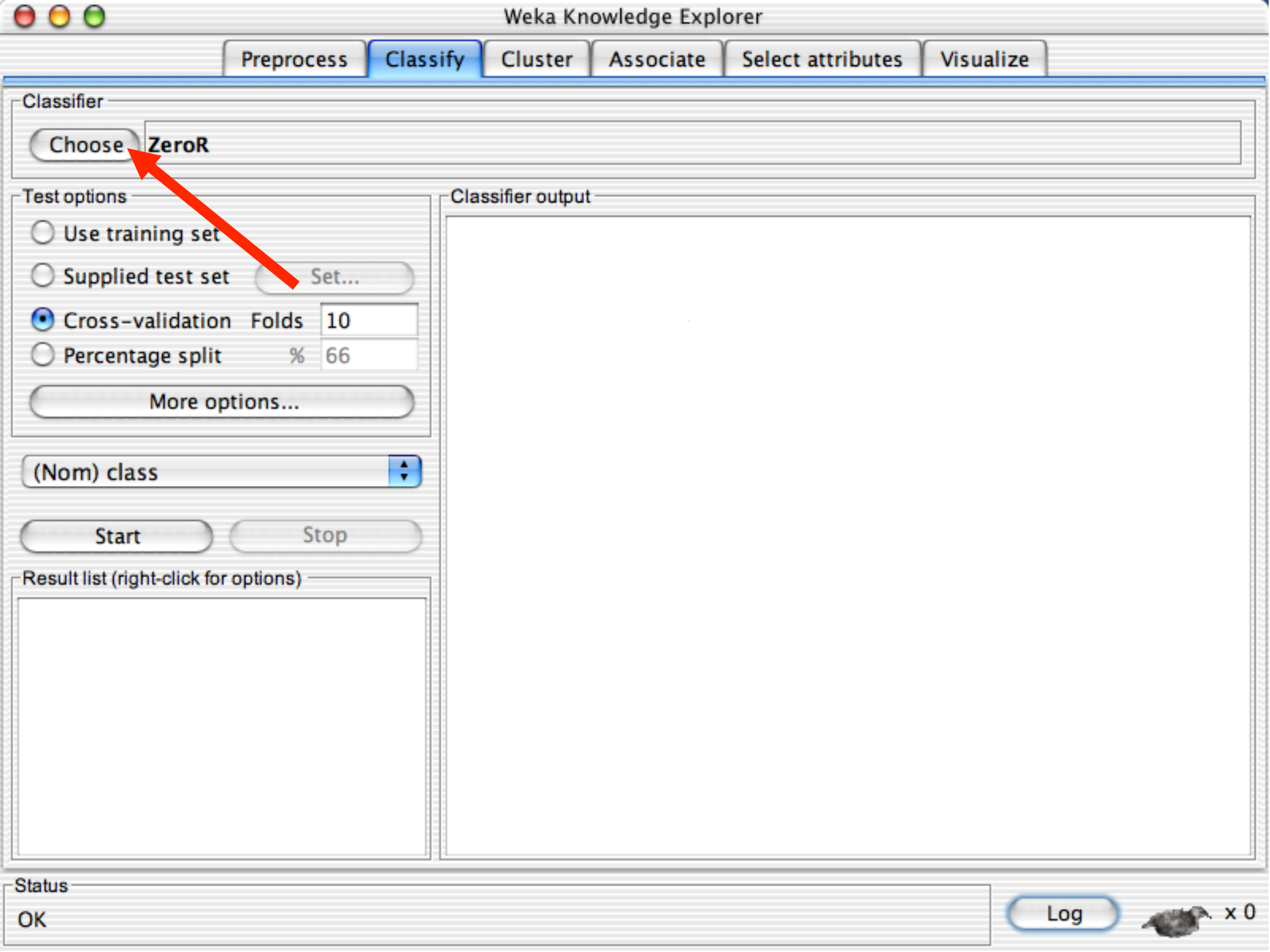
Status

OK

Log



x 0



Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Classifier

Choose

ZeroR

Test options

☐ Use training set

☐ Supplied test set

☒ Cross-validation

☐ Percentage split

Folds

10

%

66

More options...

(Nom) class

Start

Stop

Result list (right-click for options)

Classifier output

Status

OK

Log



x 0



Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Classifier

- weka
 - classifiers
 - bayes
 - functions
 - lazy
 - meta
 - misc
 - trees
 - adtree
 - DecisionStump
 - Id3
 - j48
 - J48
 - lmt
 - m5
 - RandomForest
 - RandomTree
 - REPTree
 - UserClassifier
 - rules

Classifier output

Status

OK

Log



x 0



Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Classifier

Choose

J48 -C 0.25 -M 2

Test options

☐ Use training set☐ Supplied test set

Set...

☒ Cross-validation Folds 10☐ Percentage split % 66

More options...

(Nom) class



Start

Stop

Result list (right-click for options)

Classifier output

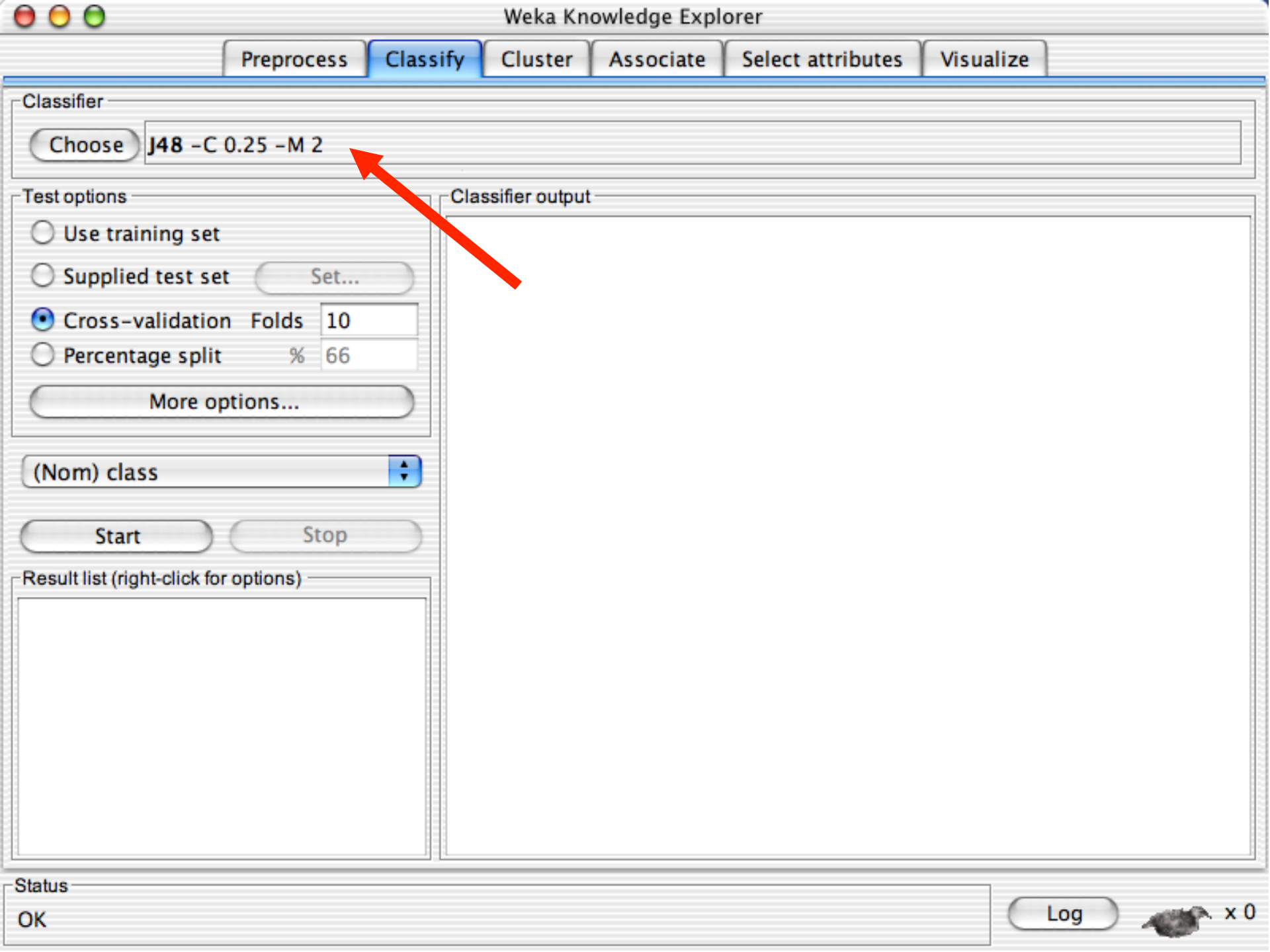
Status

OK

Log



x 0



Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Classifier

Choose

J48 -C 0.25 -M 2

Test options

☐ Use training set

☐ Supplied test set

Set...

☒ Cross-validation Folds 10

☐ Percentage split % 66

More options...

(Nom) class

Start

Stop

Result list (right-click for options)

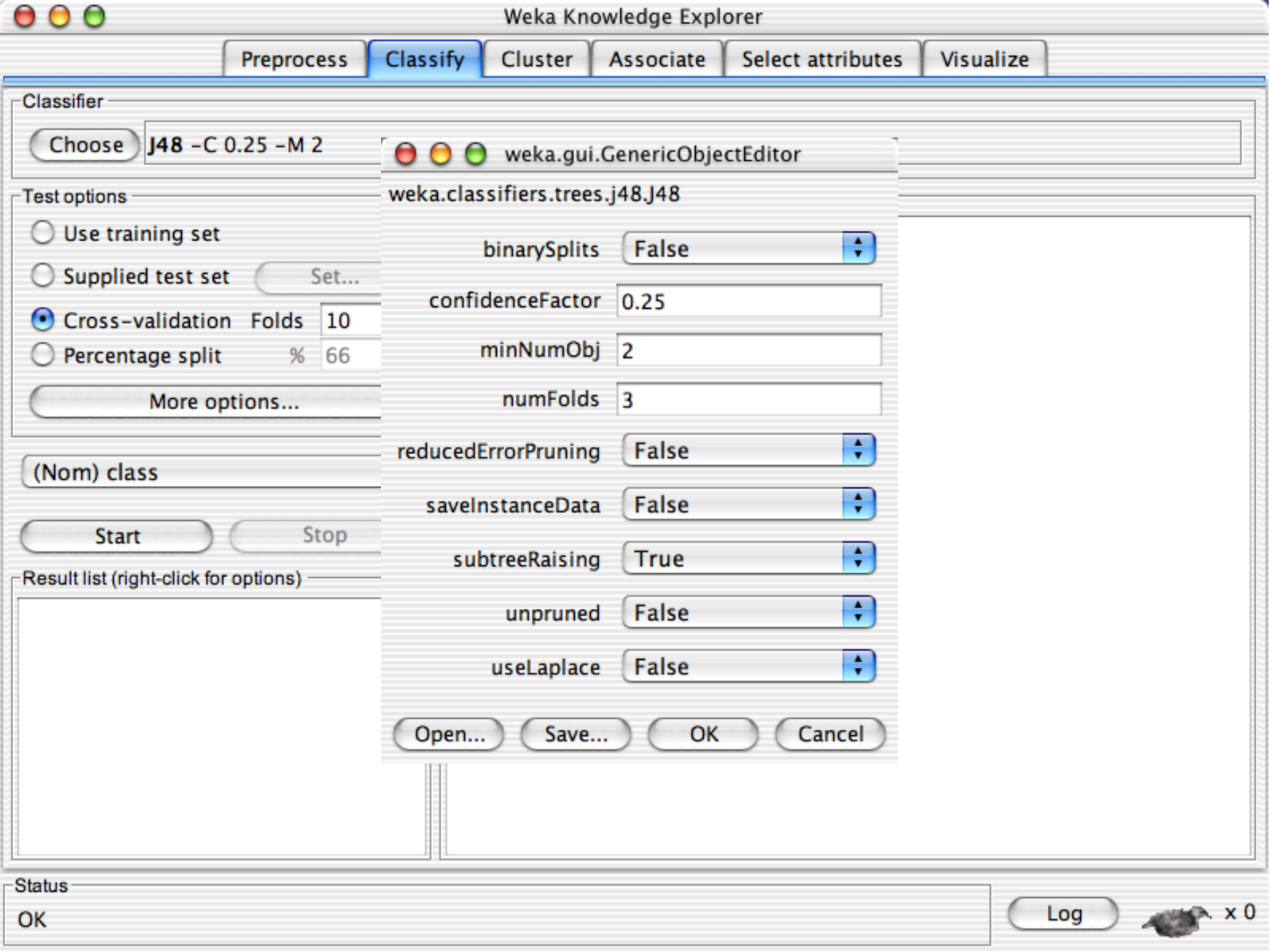
Classifier output

Status

OK

Log

x 0



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)



weka.gui.GenericObjectEditor

weka.classifiers.trees.J48.J48

binarySplits False

confidenceFactor 0.25

minNumObj 2

numFolds 3

reducedErrorPruning False

saveInstanceData False

subtreeRaising True

unpruned False

useLaplace False

Open...

Save...

OK

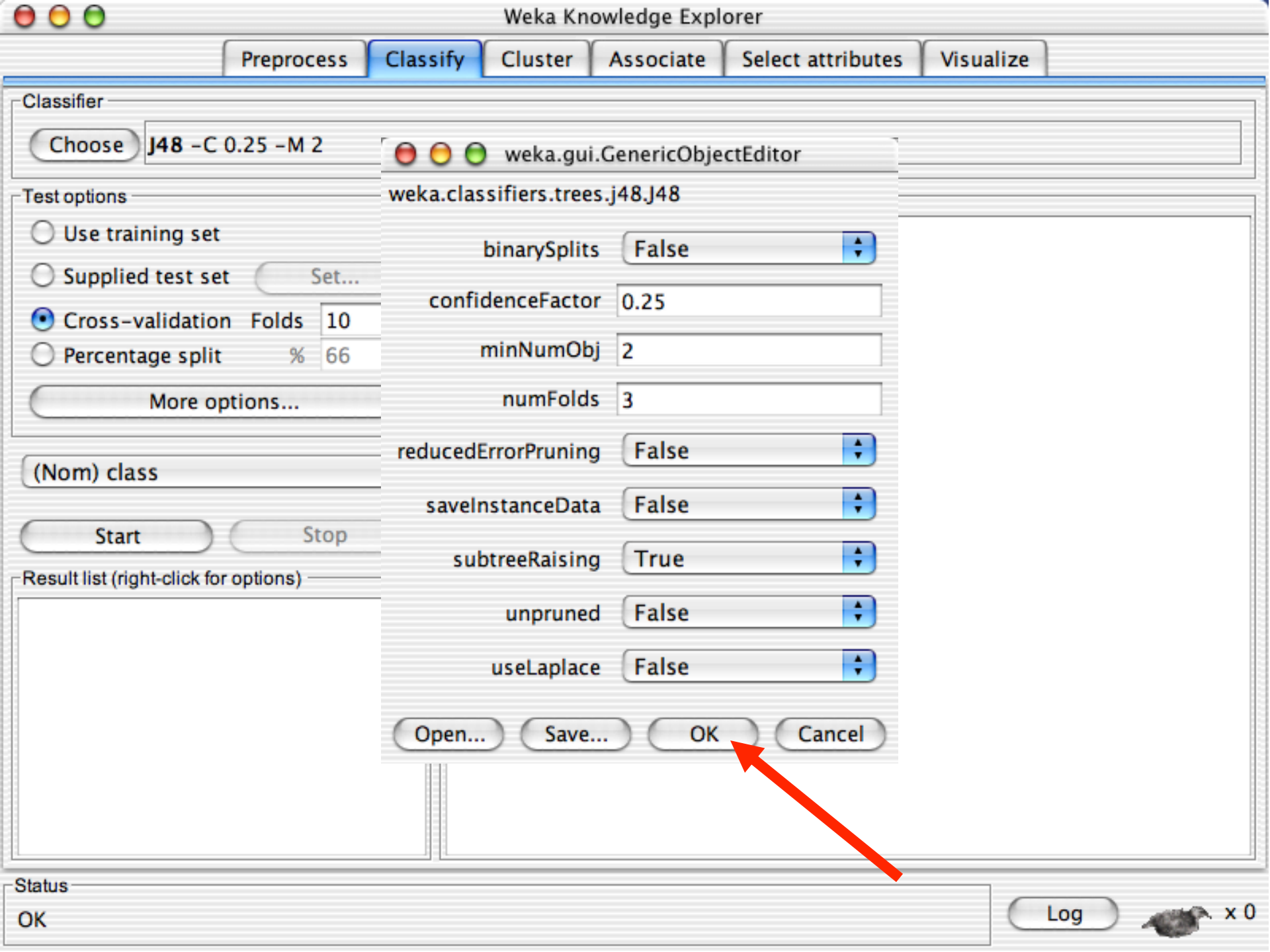
Cancel

Status

OK

Log

x 0



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)



weka.gui.GenericObjectEditor

weka.classifiers.trees.j48.J48

binarySplits False

confidenceFactor 0.25

minNumObj 2

numFolds 3

reducedErrorPruning False

saveInstanceData False

subtreeRaising True

unpruned False

useLaplace False

Open...

Save...

OK

Cancel

Status

OK

Log

x 0



Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Classifier

Choose

J48 -C 0.25 -M 2

Test options

☐ Use training set☐ Supplied test set

Set...

☒ Cross-validation Folds 10☐ Percentage split % 66

More options...

(Nom) class



Start

Stop

Result list (right-click for options)

Classifier output

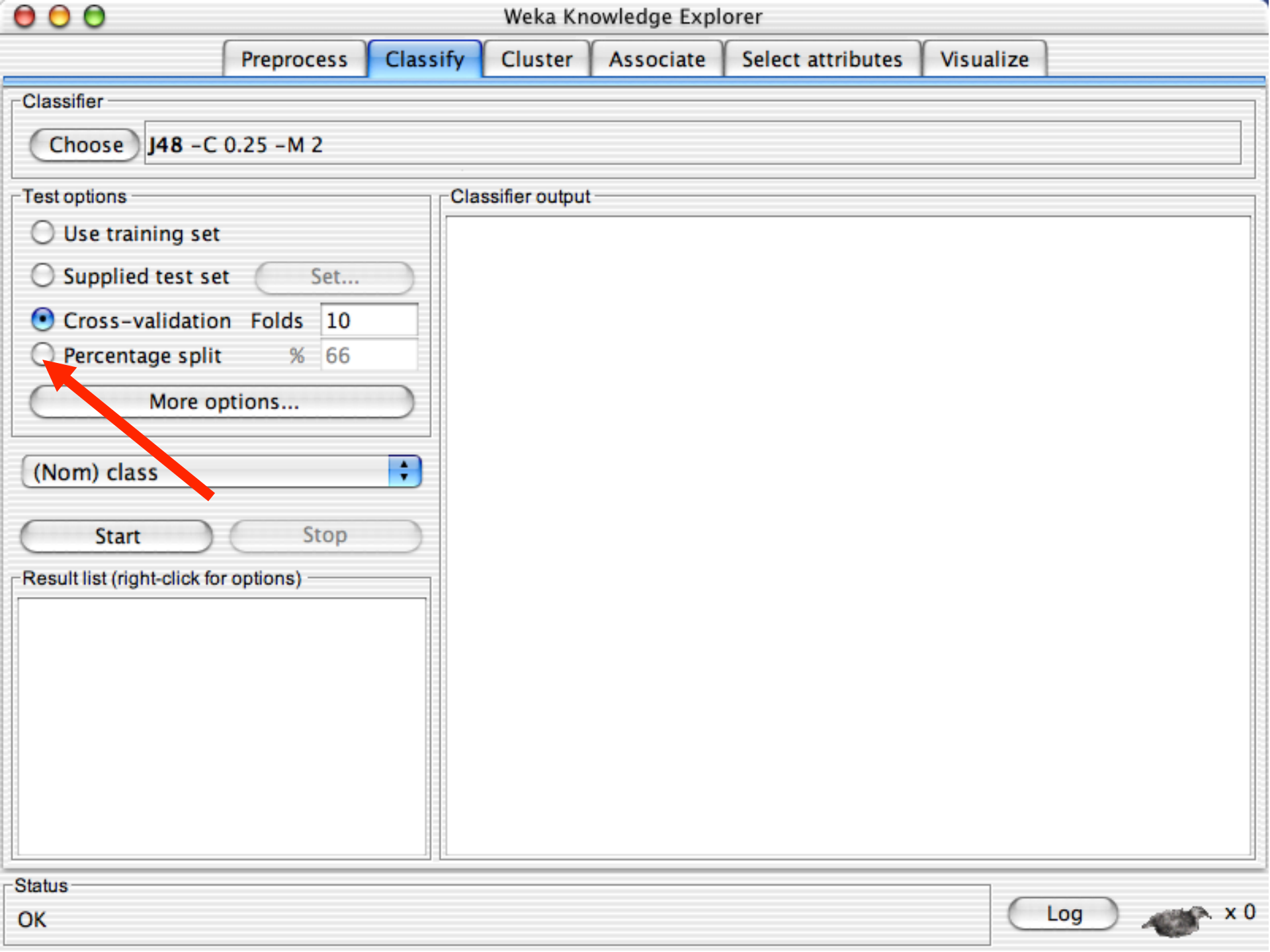
Status

OK

Log



x 0



Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Classifier

Choose

J48 -C 0.25 -M 2

Test options

☐ Use training set☐ Supplied test set

Set...

☒ Cross-validation Folds 10☐ Percentage split % 66

More options...

(Nom) class

Start

Stop

Result list (right-click for options)

Classifier output

Status

OK

Log



x 0



Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Classifier

Choose

J48 -C 0.25 -M 2

Test options

☐ Use training set☐ Supplied test set

Set...

☐ Cross-validation Folds 10☒ Percentage split % 66

More options...

(Nom) class

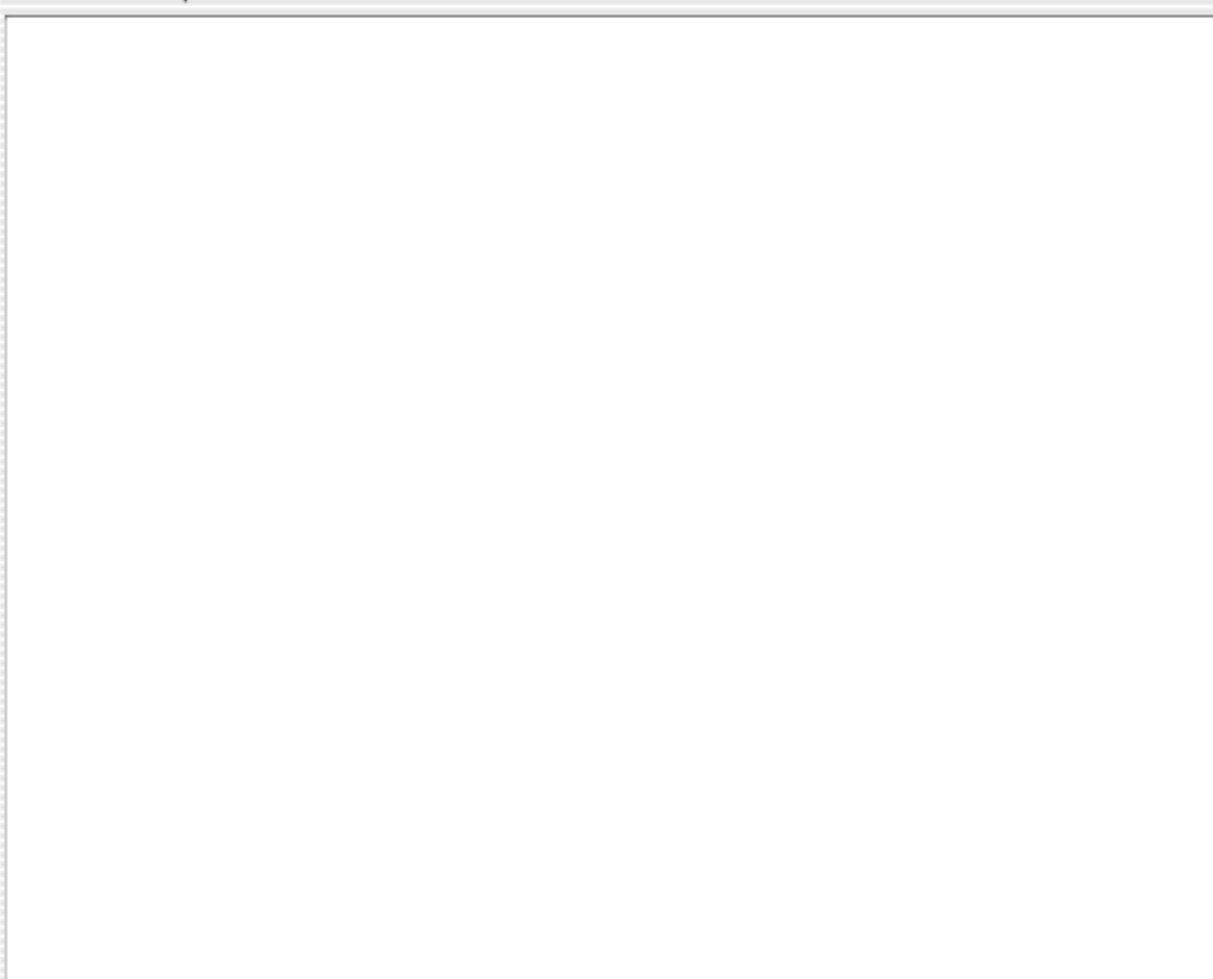


Start

Stop

Result list (right-click for options)

Classifier output



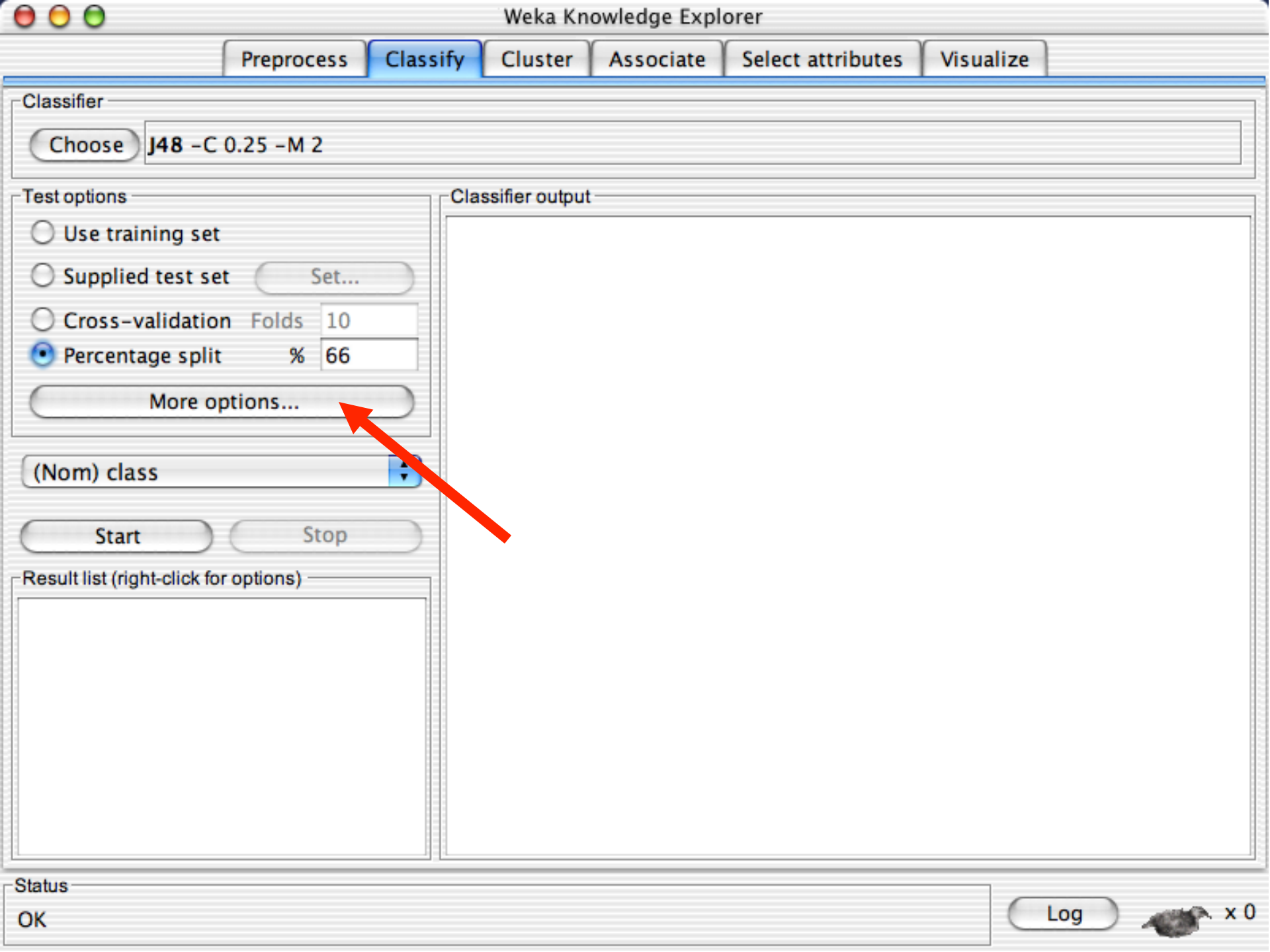
Status

OK

Log



x 0



Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Classifier

Choose

J48 -C 0.25 -M 2

Test options

☐ Use training set

☐ Supplied test set

Set...

☐ Cross-validation Folds 10

☒ Percentage split % 66

More options...

(Nom) class

Start

Stop

Result list (right-click for options)

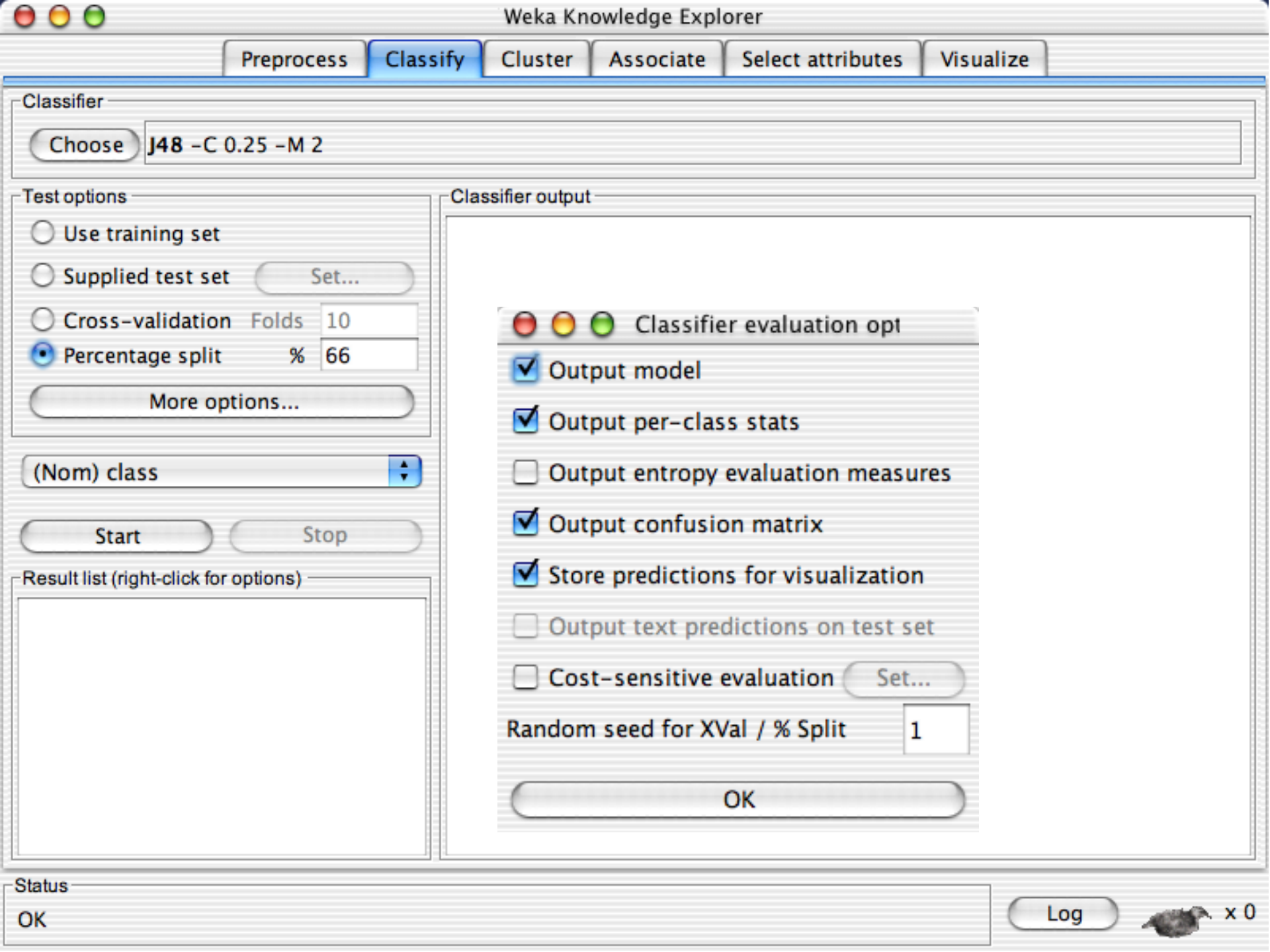
Classifier output

Status

OK

Log

x 0



Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Classifier

Choose

J48 -C 0.25 -M 2

Test options

☐ Use training set☐ Supplied test set

Set...

☐ Cross-validation Folds 10☒ Percentage split % 66

More options...

(Nom) class

Start

Stop

Result list (right-click for options)

Classifier output

Classifier evaluation opt

☒ Output model☒ Output per-class stats☐ Output entropy evaluation measures☒ Output confusion matrix☒ Store predictions for visualization☐ Output text predictions on test set☐ Cost-sensitive evaluation

Set...

Random seed for XVal / % Split

1

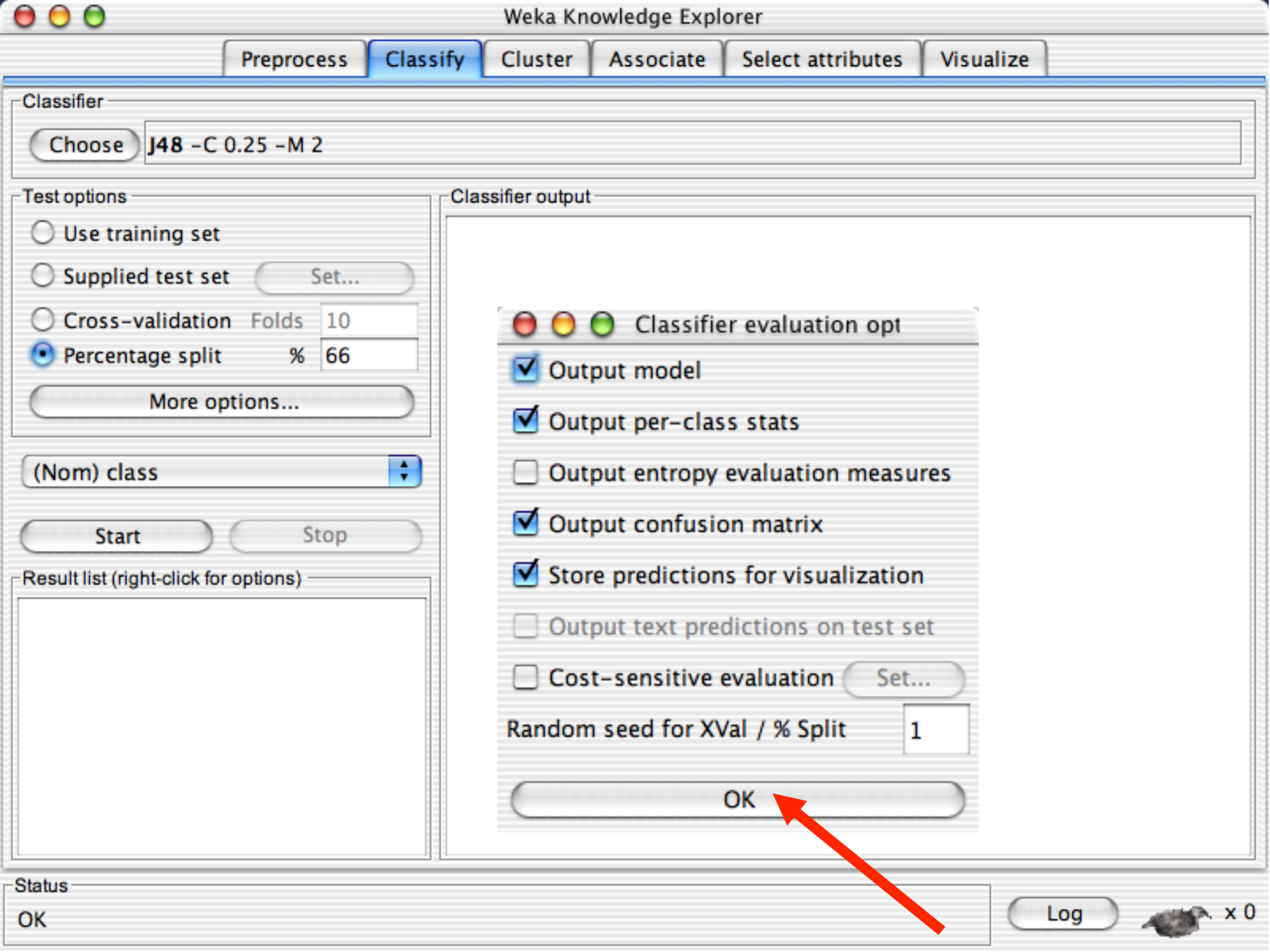
OK

Status

OK

Log

x 0



Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Classifier

Choose

J48 -C 0.25 -M 2

Test options

☐ Use training set

☐ Supplied test set

Set...

☐ Cross-validation Folds 10

☒ Percentage split % 66

More options...

(Nom) class



Start

Stop

Result list (right-click for options)

Classifier output



Classifier evaluation opt

☒ Output model

☒ Output per-class stats

☐ Output entropy evaluation measures

☒ Output confusion matrix

☒ Store predictions for visualization

☐ Output text predictions on test set

☐ Cost-sensitive evaluation

Set...

Random seed for XVal / % Split

1

OK



Status

OK

Log



x 0



Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Classifier

Choose

J48 -C 0.25 -M 2

Test options

☐ Use training set☐ Supplied test set

Set...

☐ Cross-validation Folds 10☒ Percentage split % 66

More options...

(Nom) class

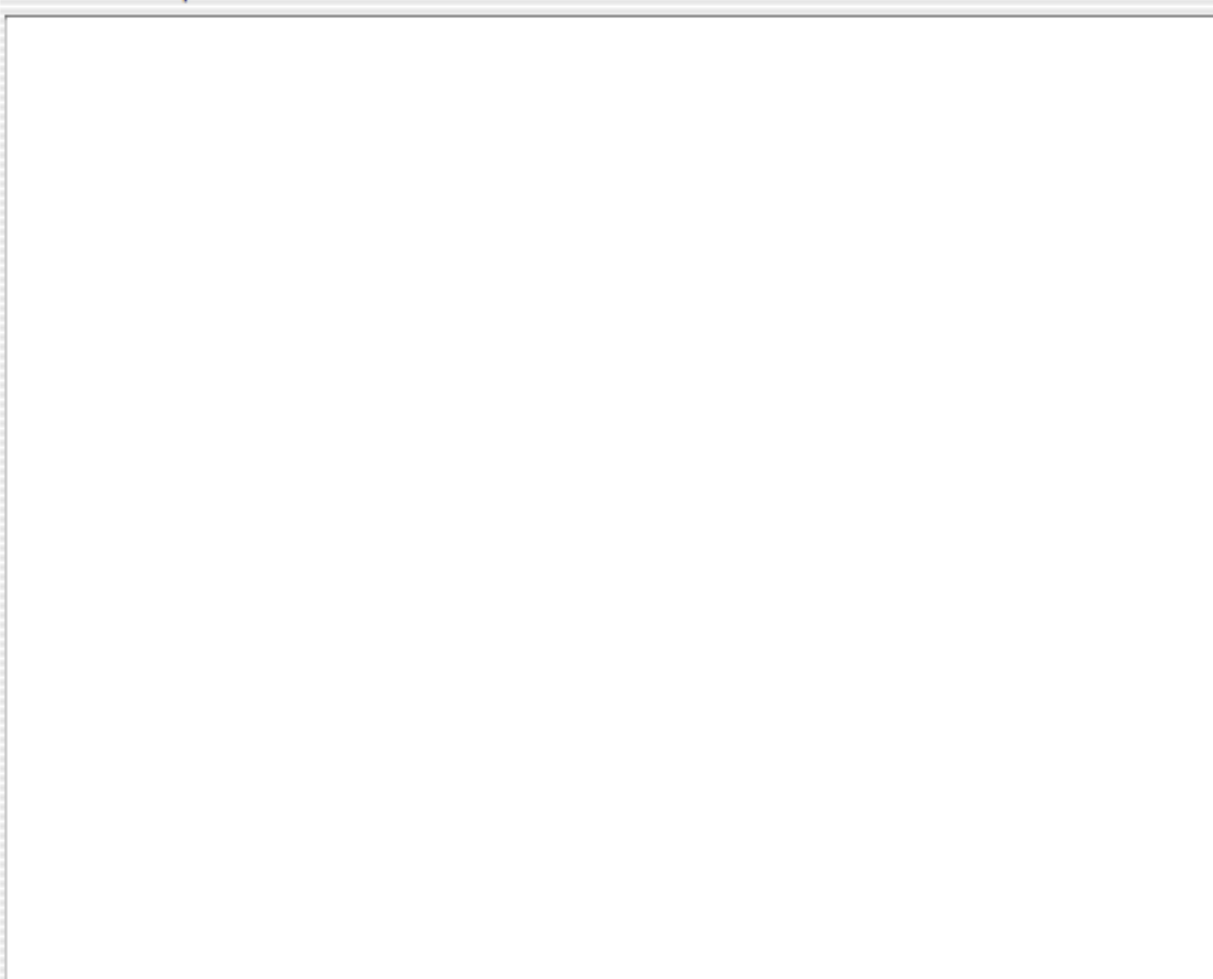


Start

Stop

Result list (right-click for options)

Classifier output



Status

OK

Log



x 0



Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Classifier

Choose

J48 -C 0.25 -M 2

Test options

☐ Use training set☐ Supplied test set

Set...

☐ Cross-validation Folds 10☒ Percentage split % 66

More options...

(Nom) class

Start

Stop

Result list (right-click for options)

Classifier output

Status

OK

Log



x 0



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

=== Run information ===

Scheme: weka.classifiers.trees.j48.J48 -C 0.25 -M 2

Relation: iris

Instances: 150

Attributes: 5

sepalength

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

Log



x 0

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

=== Run information ===

Scheme: weka.classifiers.trees.j48.J48 -C 0.25 -M 2
Relation: iris
Instances: 150
Attributes: 5

sepalength
sepalwidth
petallength
petalwidth
class

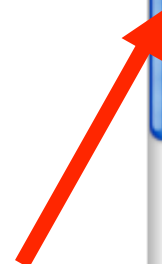
Test mode: split 66% train, remainder test

=== 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

Log

 x 0

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



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



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

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

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

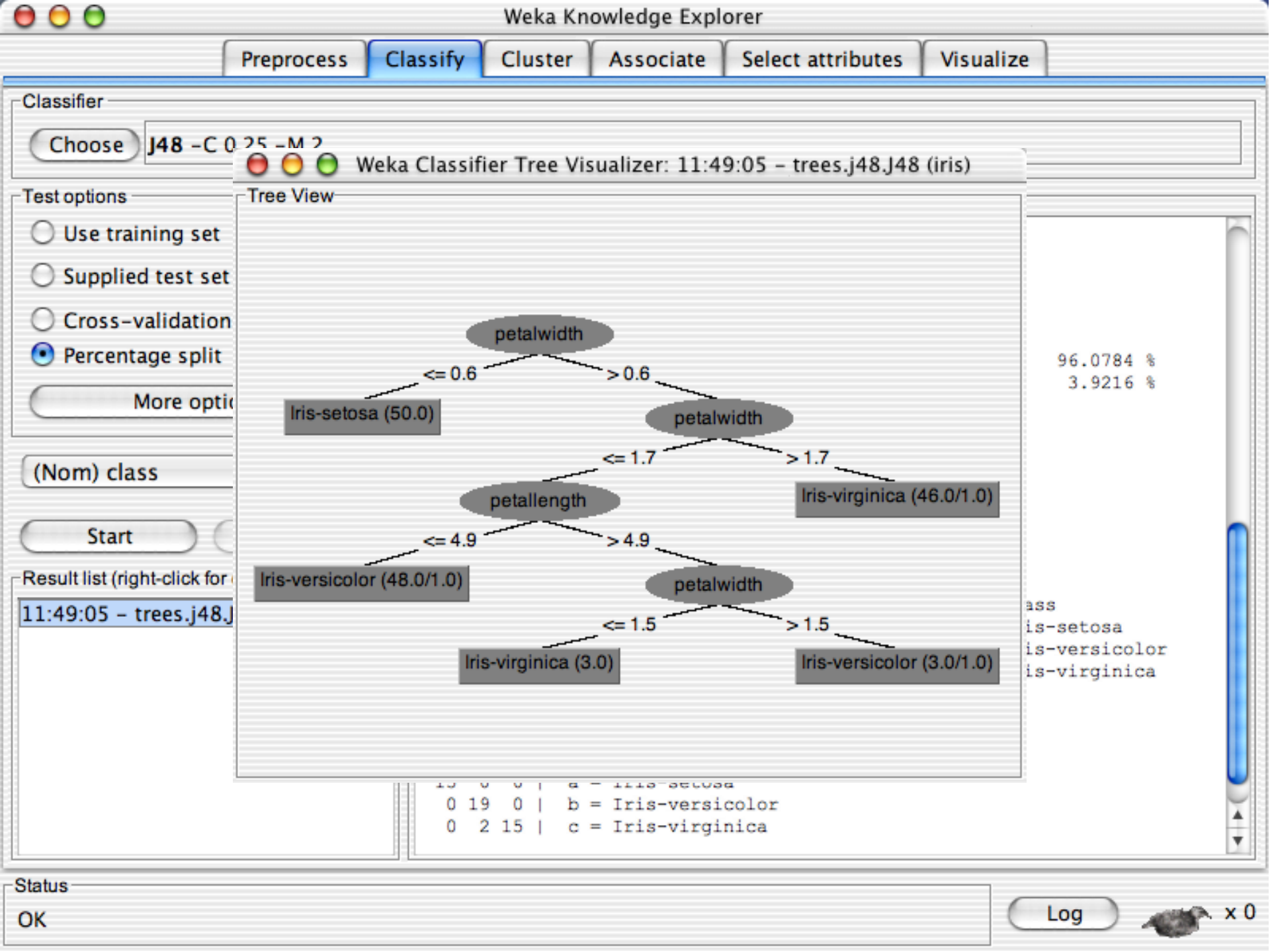
Status

OK

Log



x 0



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

Explorer: clustering data

- WEKA contains “clusterers” for finding groups of similar instances in a dataset
- Implemented schemes are:
 - *k-Means*, EM, Cobweb, X-means, FarthestFirst
- Clusters can be visualized and compared to “true” clusters (if given)
- Evaluation based on loglikelihood if clustering scheme produces a probability distribution

The K-Means Clustering Method

- Given k , the *k-means* algorithm is implemented in four steps:
 - Partition objects into k nonempty subsets
 - Compute seed points as the centroids of the clusters of the current partition (the centroid is the center, i.e., *mean point*, of the cluster)
 - Assign each object to the cluster with the nearest seed point
 - Go back to Step 2, stop when no more new assignment

- Demo Now. (Demo Online)

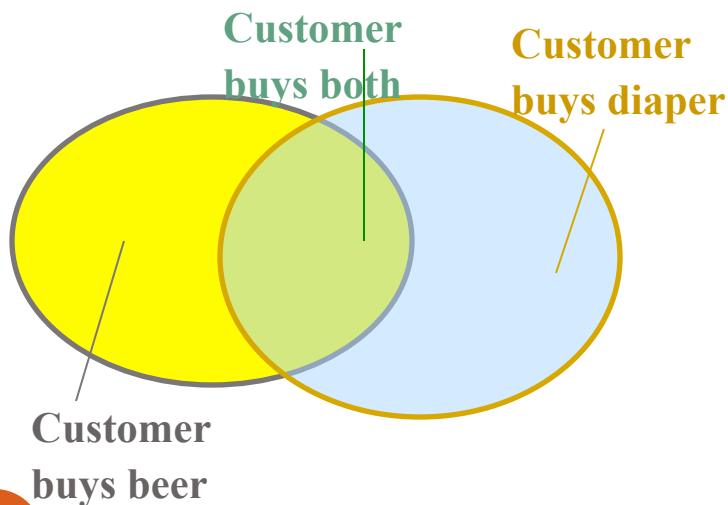
Explorer: finding associations

- 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

Basic Concepts: Frequent Patterns

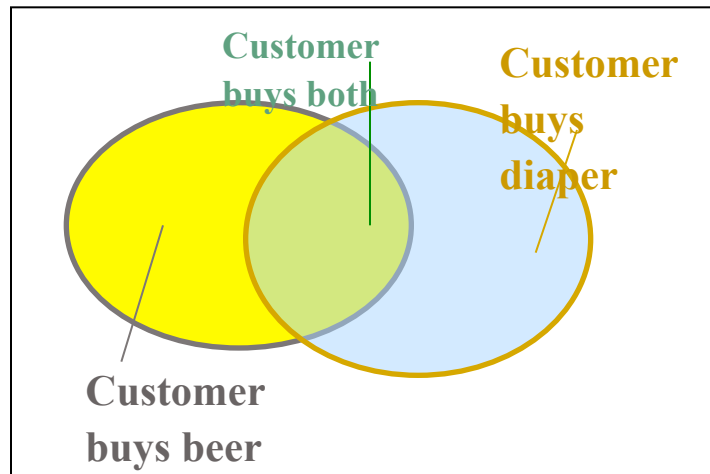
Tid	Items bought
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk

- **itemset**: A set of one or more items
- **k-itemset** $X = \{x_1, \dots, x_k\}$
- **(absolute) support**, or, **support count** of X : Frequency or occurrence of an itemset X
- **(relative) support**, s , is the fraction of transactions that contains X (i.e., the probability that a transaction contains X)
- An itemset X is **frequent** if X 's support is no less than a *minsup* threshold



Basic Concepts: Association Rules

Tid	Items bought
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk



- Find all the rules $X \rightarrow Y$ with minimum support and confidence
- support**, s , probability that a transaction contains $X \cup Y$
- confidence**, c , conditional probability that a transaction having X also contains Y

Let $\text{minsup} = 50\%$, $\text{minconf} = 50\%$

Freq. Pat.: Beer:3, Nuts:3, Diaper:4, Eggs:3, {Beer, Diaper}:3

- Association rules: (many more!)
 - $\text{Beer} \rightarrow \text{Diaper}$ (60%, 100%)
 - $\text{Diaper} \rightarrow \text{Beer}$ (60%, 75%)



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

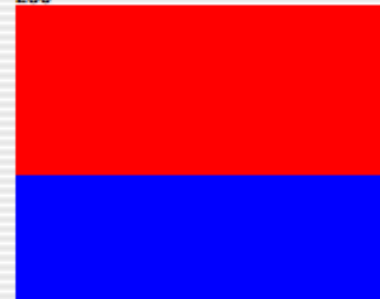
Unique: 0 (0%)

Label	Count
n	236
y	187

Colour: Class (Nom)

Visualize All

236



187



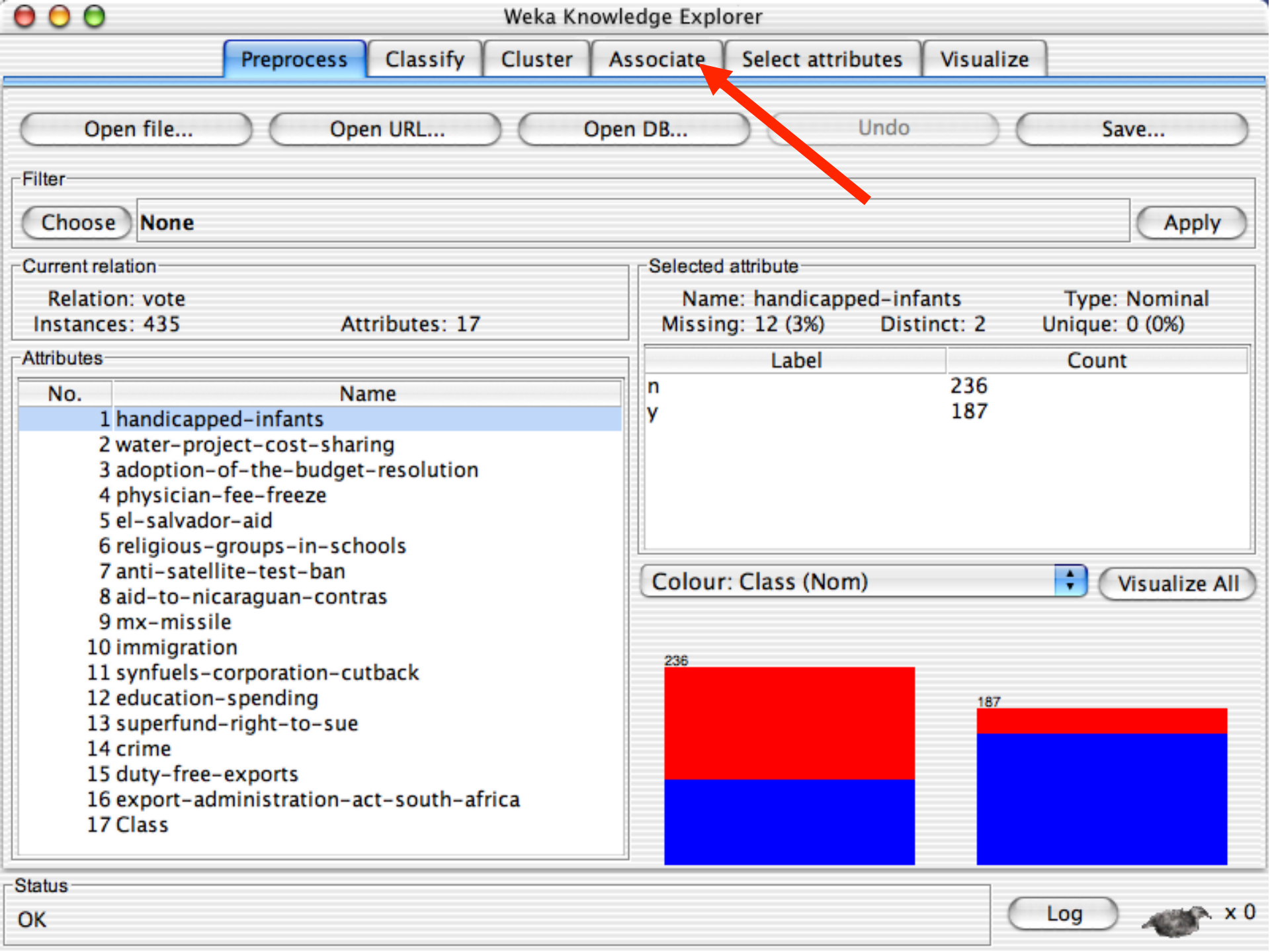
Status

OK

Log



x 0





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)

Associator output

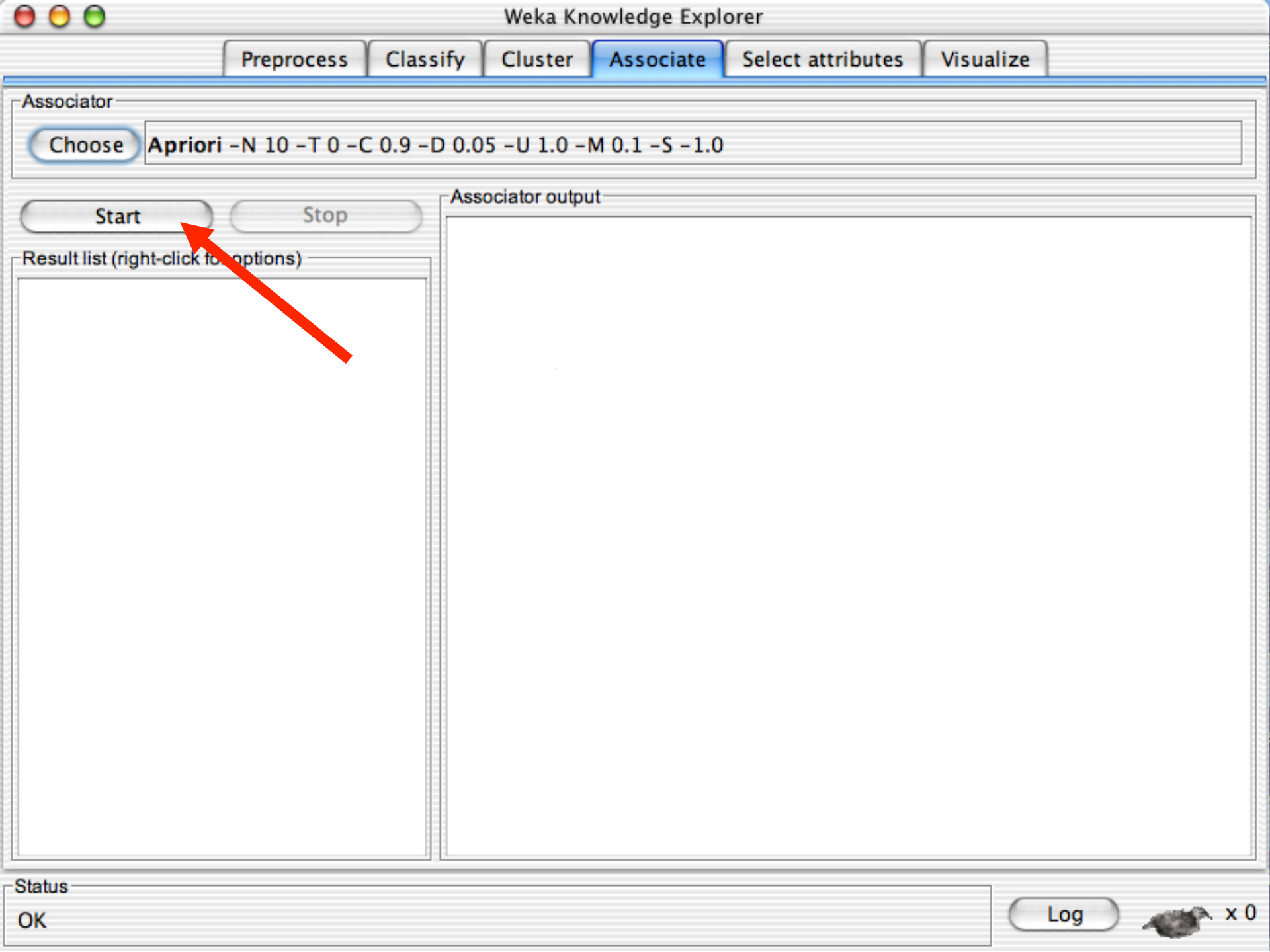
Status

OK

Log



x 0



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

Associator output

Result list (right-click for options)

Status

OK

Log



x 0

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 219
2. adoption-of-the-budget-resolution=y physician-fee-freeze=n aid-to-nicaraguan-contras=y 210 ==> Class=democrat 210
3. physician-fee-freeze=n aid-to-nicaraguan-contras=y 211 ==> Class=democrat 210
4. physician-fee-freeze=n education-spending=n 202 ==> Class=democrat 201 conf: (0.99)
5. physician-fee-freeze=n 247 ==> Class=democrat 245 conf: (0.99)
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: WEKA allows (almost) arbitrary combinations of these two



Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Attribute Evaluator

Choose

CfsSubsetEval

Search Method

Choose

BestFirst -D 1 -N 5

Attribute Selection Mode



Use full training set



Cross-validation

Folds

10

Seed

1

(Nom) Class

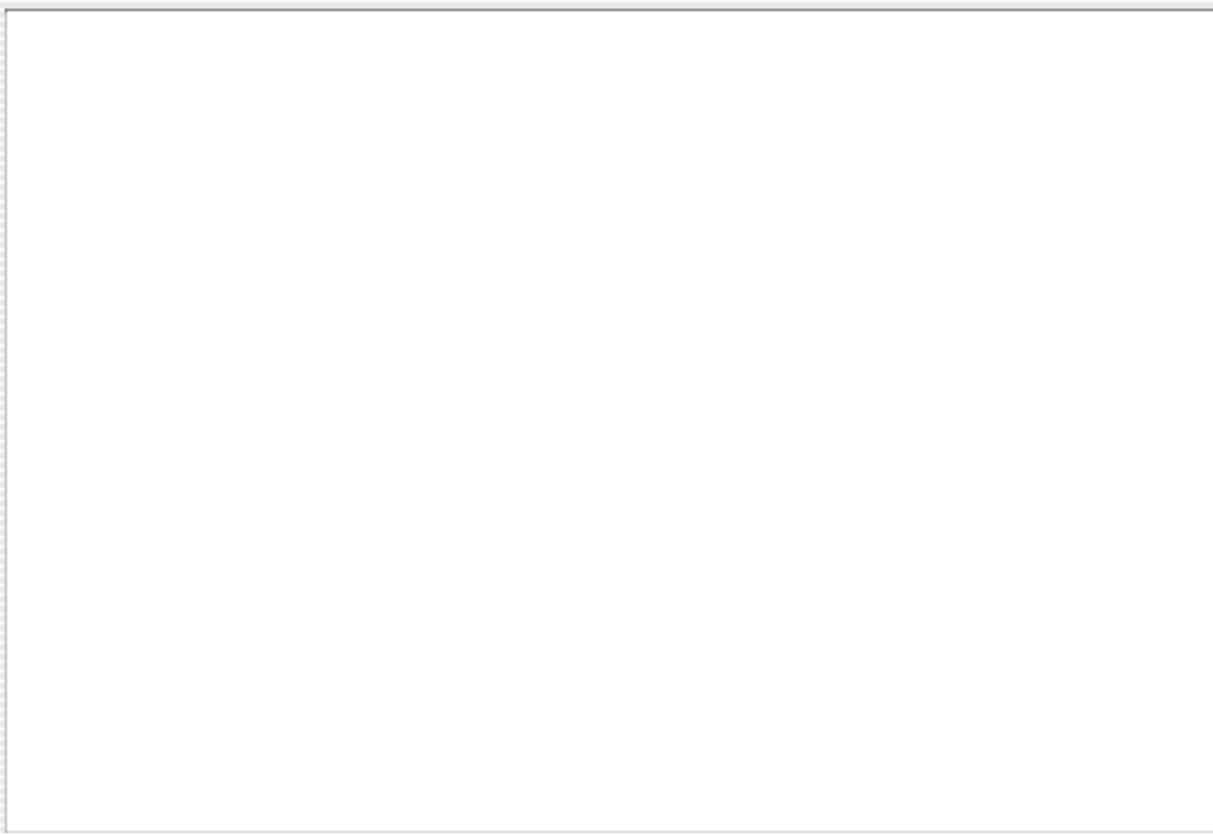


Start

Stop

Result list (right-click for options)

Attribute selection output



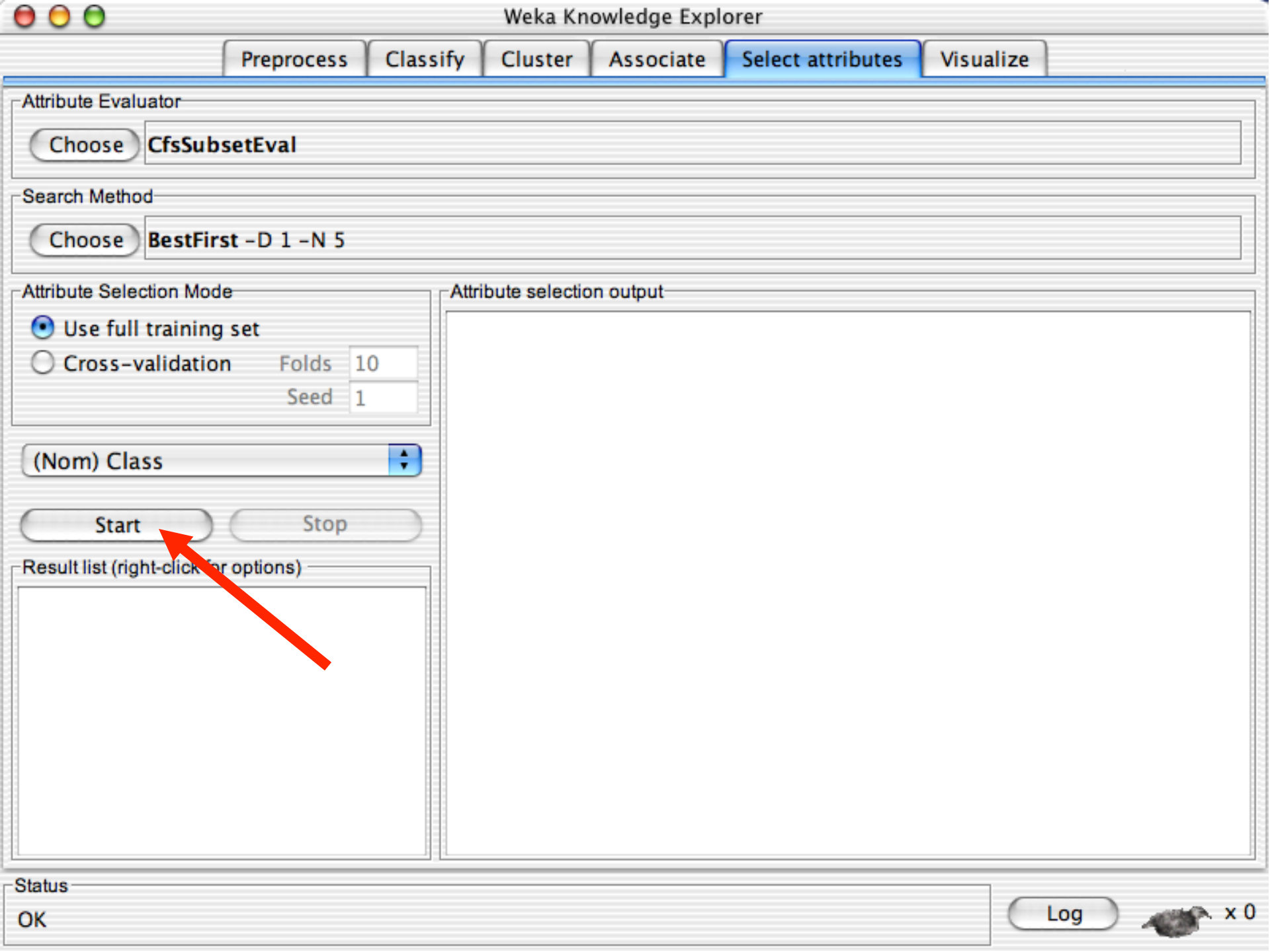
Status

OK

Log



x 0



Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Attribute Evaluator

Choose

CfsSubsetEval

Search Method

Choose

BestFirst -D 1 -N 5

Attribute Selection Mode



Use full training set



Cross-validation

Folds

10

Seed

1

(Nom) Class



Start

Stop

Result list (right-click for options)

Attribute selection output

Status

OK

Log



x 0

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Attribute Evaluator

 CfsSubsetEval

Search Method

 BestFirst -D 1 -N 5

Attribute Selection Mode

☒ Use full training set☐ Cross-validation

Folds

10

Seed

1

(Nom) Class

Result list (right-click for options)

16:39:40 - BestFirst + CfsSubsetEval

Attribute selection output

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

=== Attribute Selection on all input data ===

Search Method:
  Best first.
  Start set: no attributes
  Search direction: forward
  Stale search after 5 node expansions
  Total number of subsets evaluated: 83
  Merit of best subset found:    0.729

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

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

Status

OK

 x 0

Preprocess

Classify

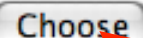
Cluster

Associate

Select attributes

Visualize

Attribute Evaluator

 CfsSubsetEval

Search Method

 BestFirst -D 1 -N 5

Attribute Selection Mode

☒ Use full training set☐ Cross-validation

Folds 10

Seed 1

(Nom) Class

Start

Stop

Result list (right-click for options)

16:39:40 - BestFirst + CfsSubsetEval

Attribute selection output

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

=== Attribute Selection on all input data ===

Search Method:
  Best first.
  Start set: no attributes
  Search direction: forward
  Stale search after 5 node expansions
  Total number of subsets evaluated: 83
  Merit of best subset found:    0.729

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

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

Status

OK

Log

 x 0



Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Attribute Evaluator

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

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

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

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



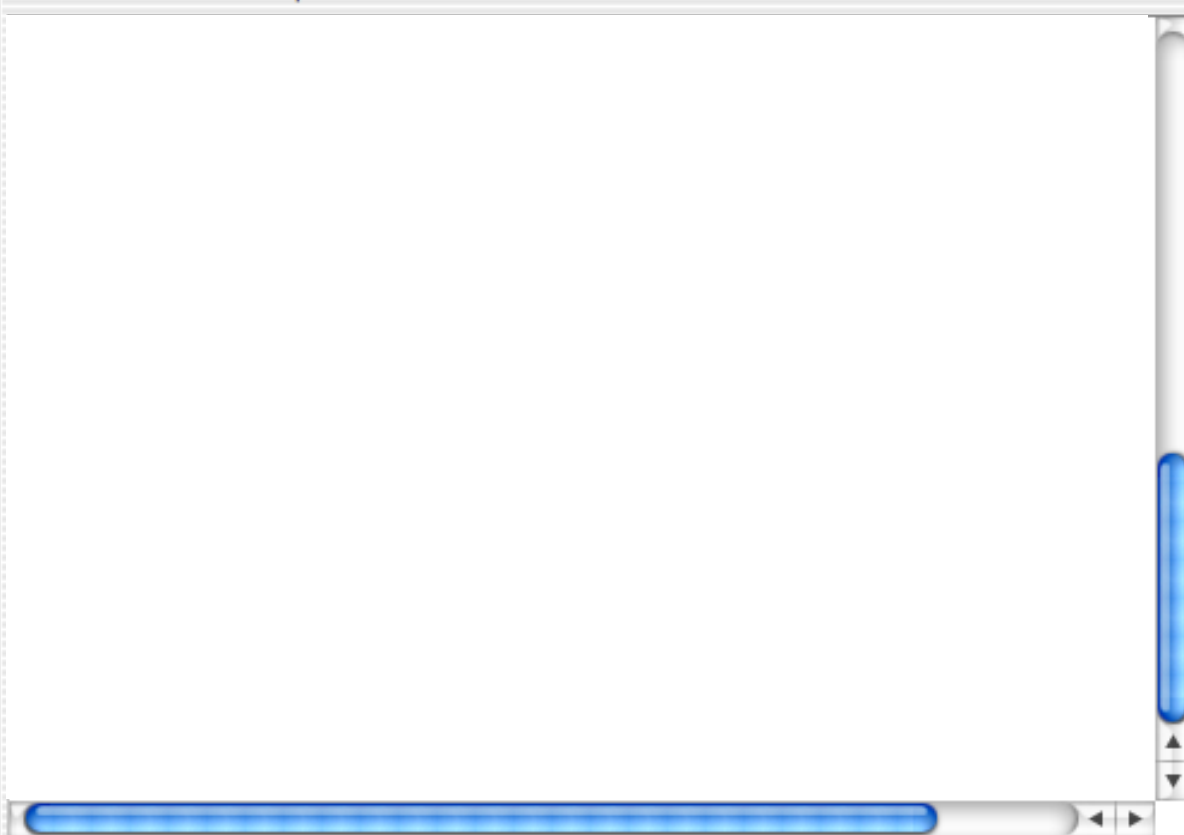
Start

Stop

Result list (right-click for options)

16:39:40 - BestFirst + CrossSubsetEval

Attribute selection output



Status

OK

Log



x 0

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 (1-d) and pairs of attributes (2-d)
 - To do: rotating 3-d visualizations (Xgobi-style)
- Color-coded class values
- “Jitter” option to deal with nominal attributes (and to detect “hidden” data points)
- “Zoom-in” function



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

Missing: 0 (0%)

Distinct: 178

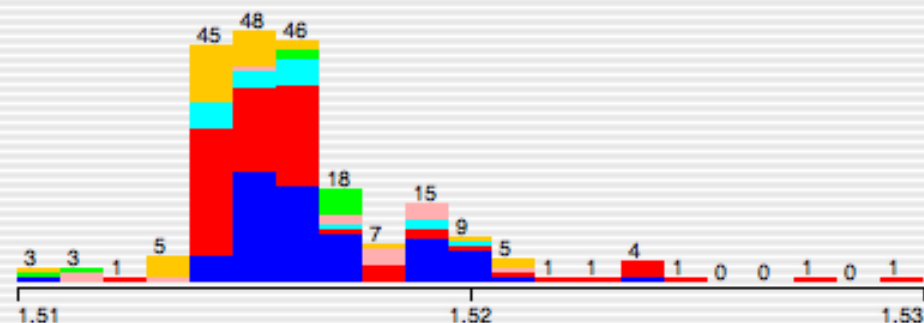
Type: Numeric

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

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Plot Matrix

RI

Na

Mg

Al

Si

K

Type

Fe

PlotSize: [100]

PointSize: [1]

Jitter:

Colour: Type (Nom)

Update

Select Attributes

SubSample % :

100

Class Colour

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

Status

OK

Log

x 0

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Plot Matrix

Ri

Na

Mg

Al

Si

K

Type

Fe

PlotSize: [100]

PointSize: [1]

Jitter:

Colour: Type (Nom)

Update

Select Attributes

SubSample % : 100

Class Colour

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

Status

OK

Log

x 0



Preprocess

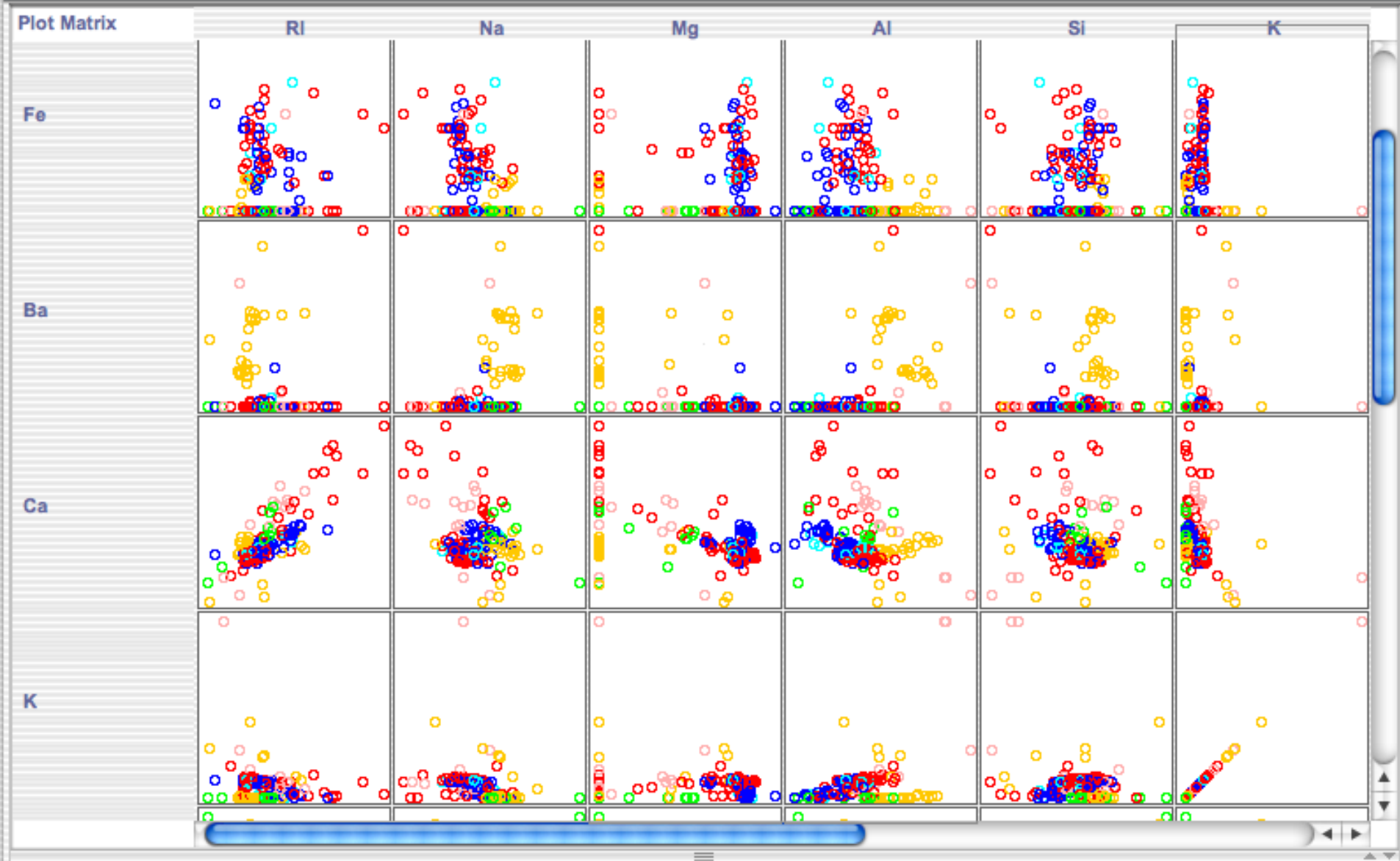
Classify

Cluster

Associate

Select attributes

Visualize



Status

OK

Log

x 0



Preprocess

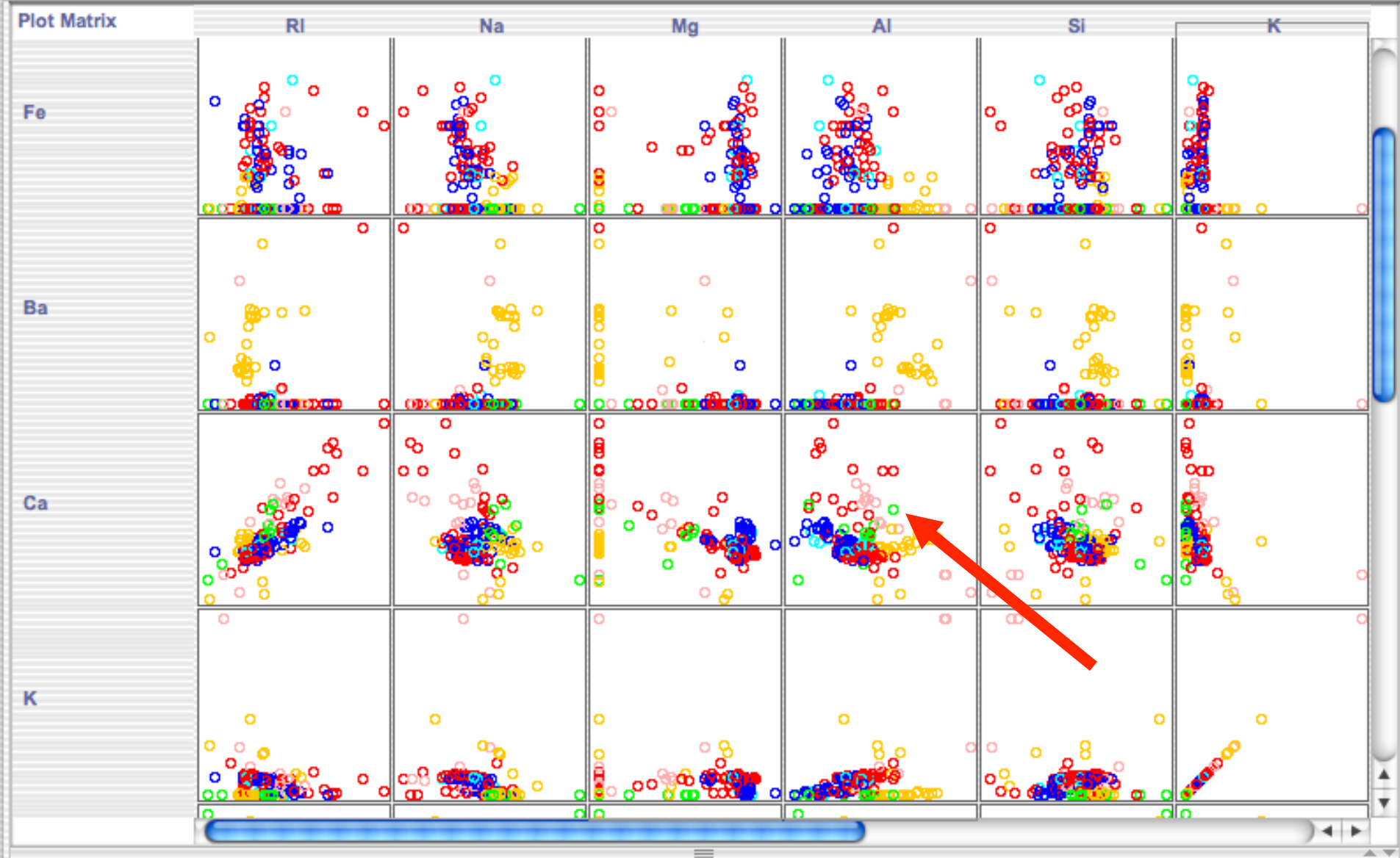
Classify

Cluster

Associate

Select attributes

Visualize



Status

OK

Log

x 0

X: Al (Num)

Y: Ca (Num)

Colour: Type (Nom)

Select Instance

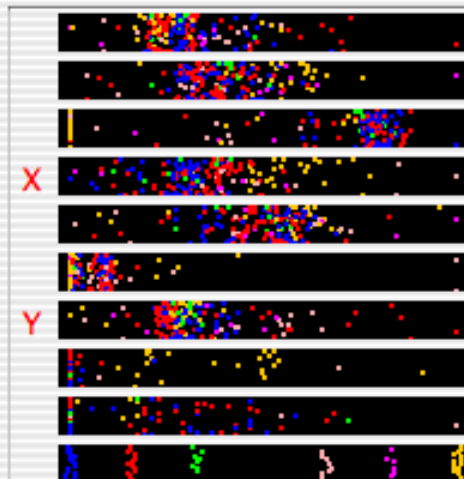
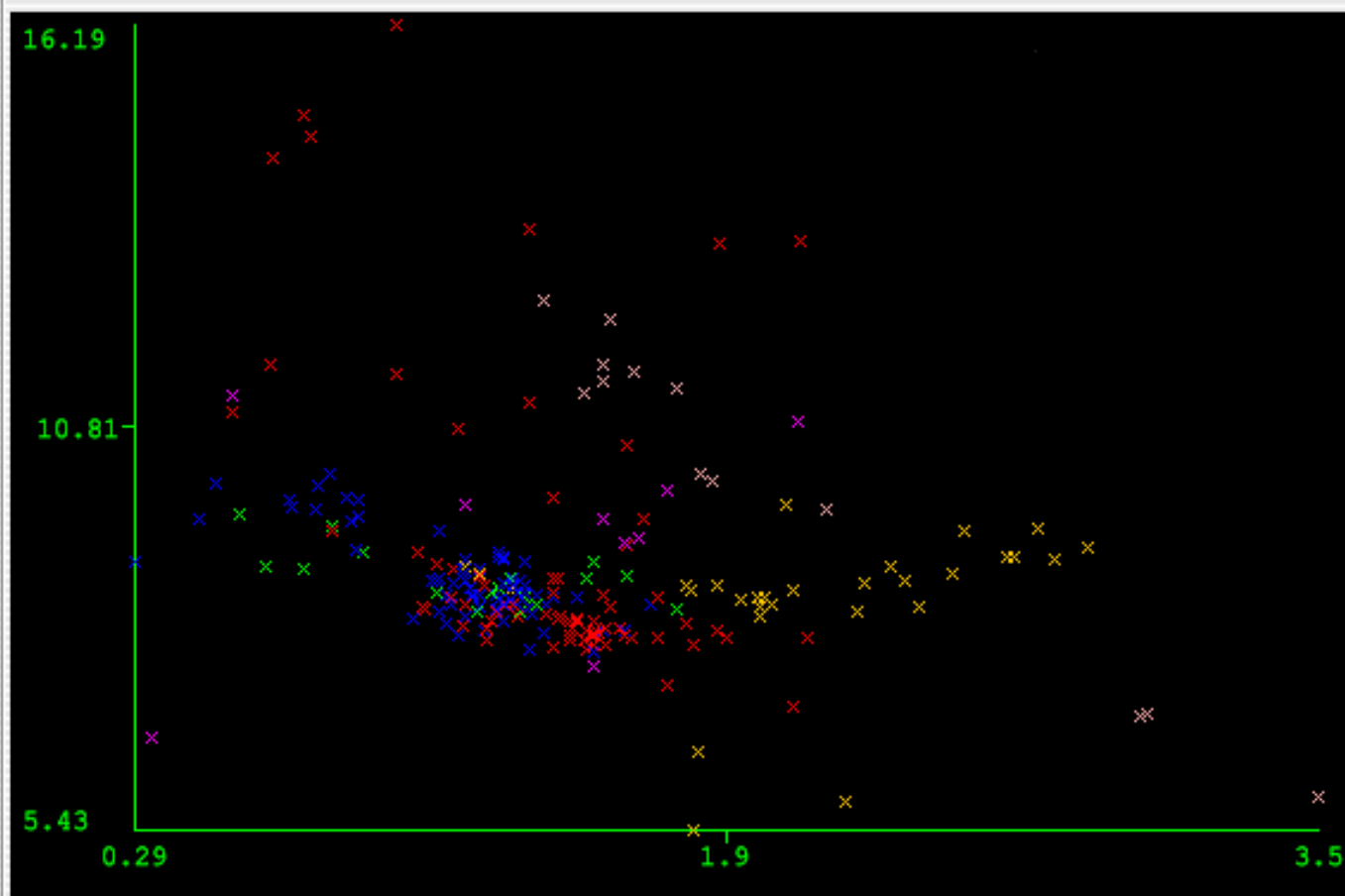
Reset

Clear

Save

Jitter

Plot: Glass



Class colour

build wind float

build wind non-float

vehic wind float

vehic wind non-float

containers

tableware

headlamps

X: Al (Num)

Y: Ca (Num)

Colour: Type (Nom)

Select Instance

Reset

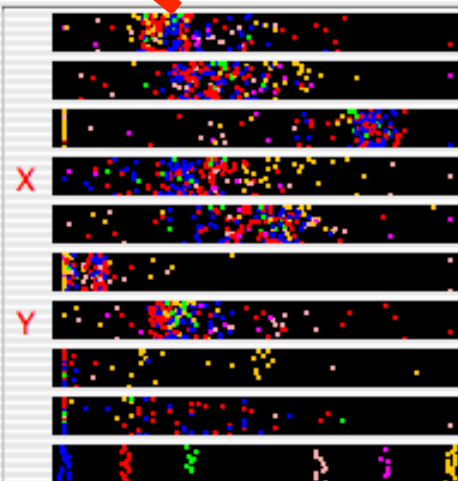
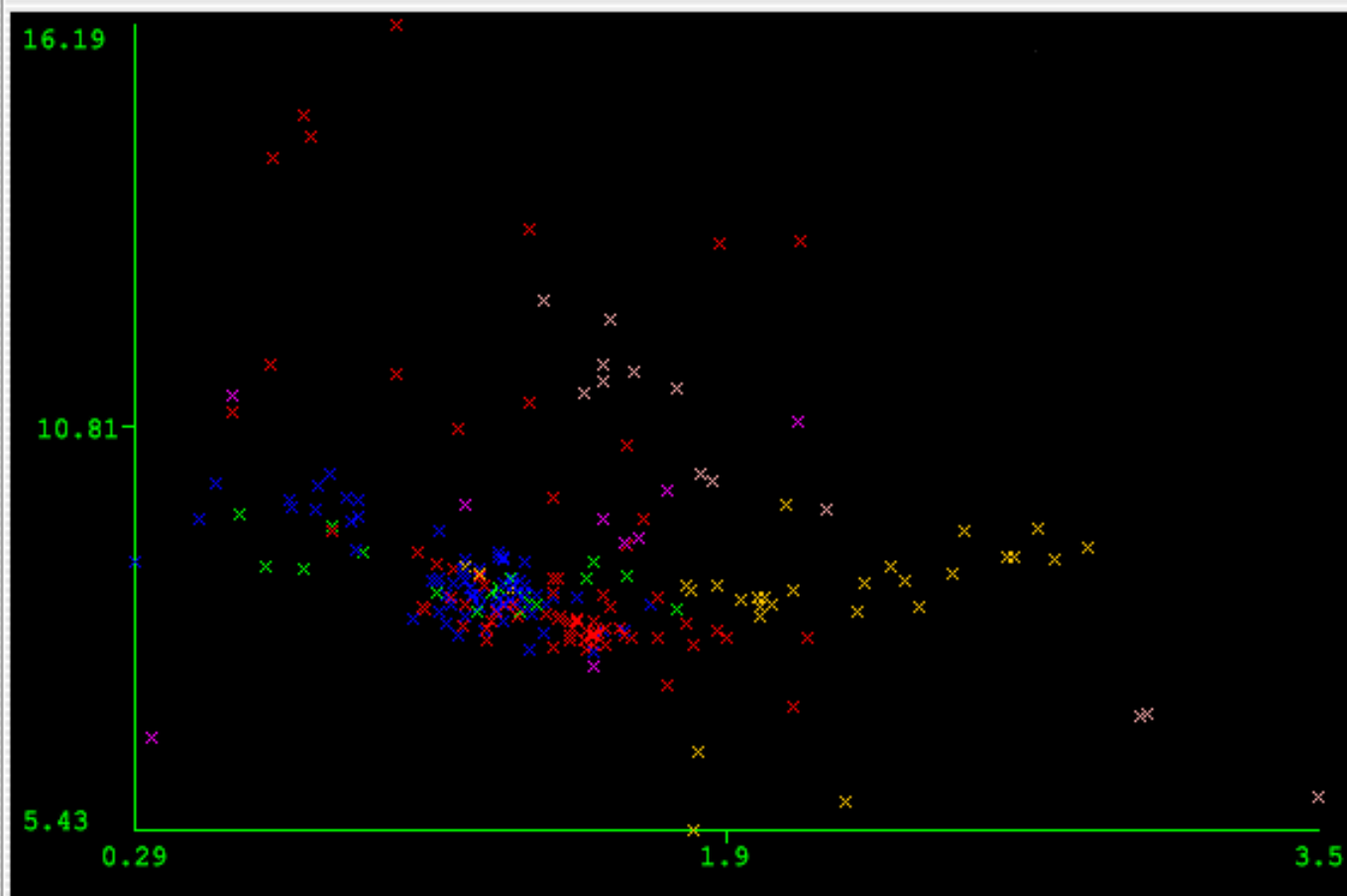
Clear

Save

Jitter



Plot: Glass



Class colour

build wind float

build wind non-float

vehic wind float

vehic wind non-float

containers

tableware

headlamps

X: Al (Num)

Y: Ca (Num)

Colour: Type (Nom)

Rectangle

Submit

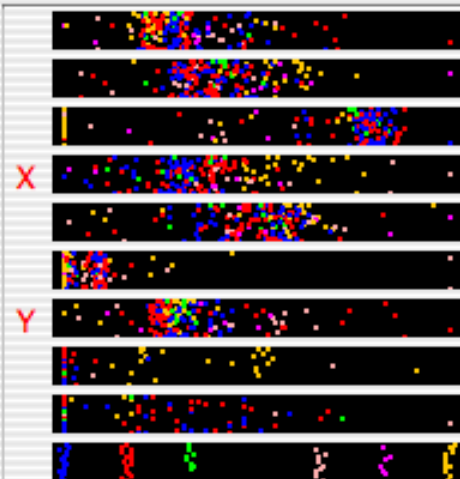
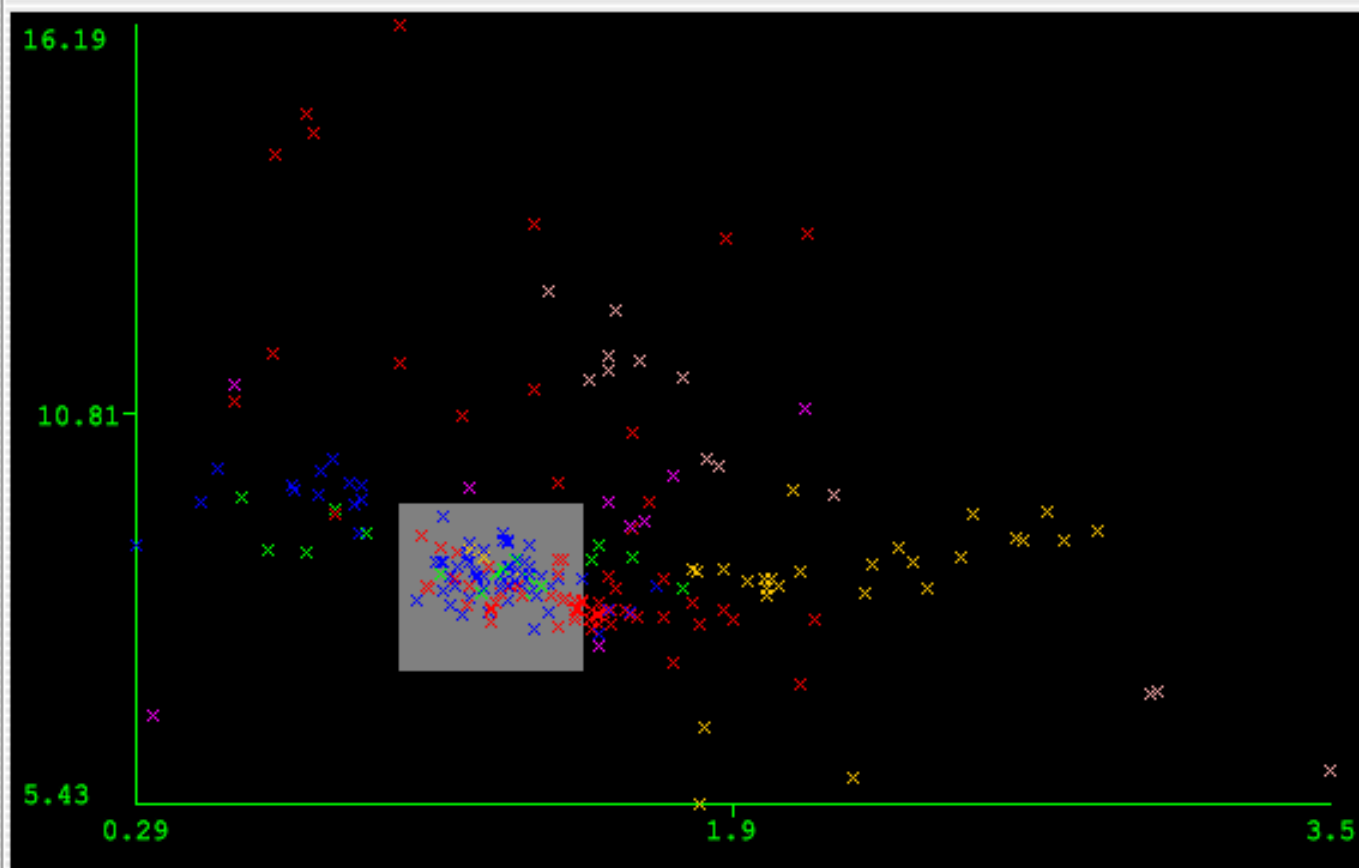
Clear

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

Submit

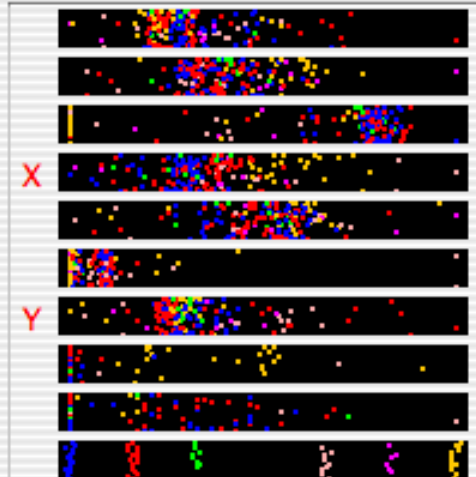
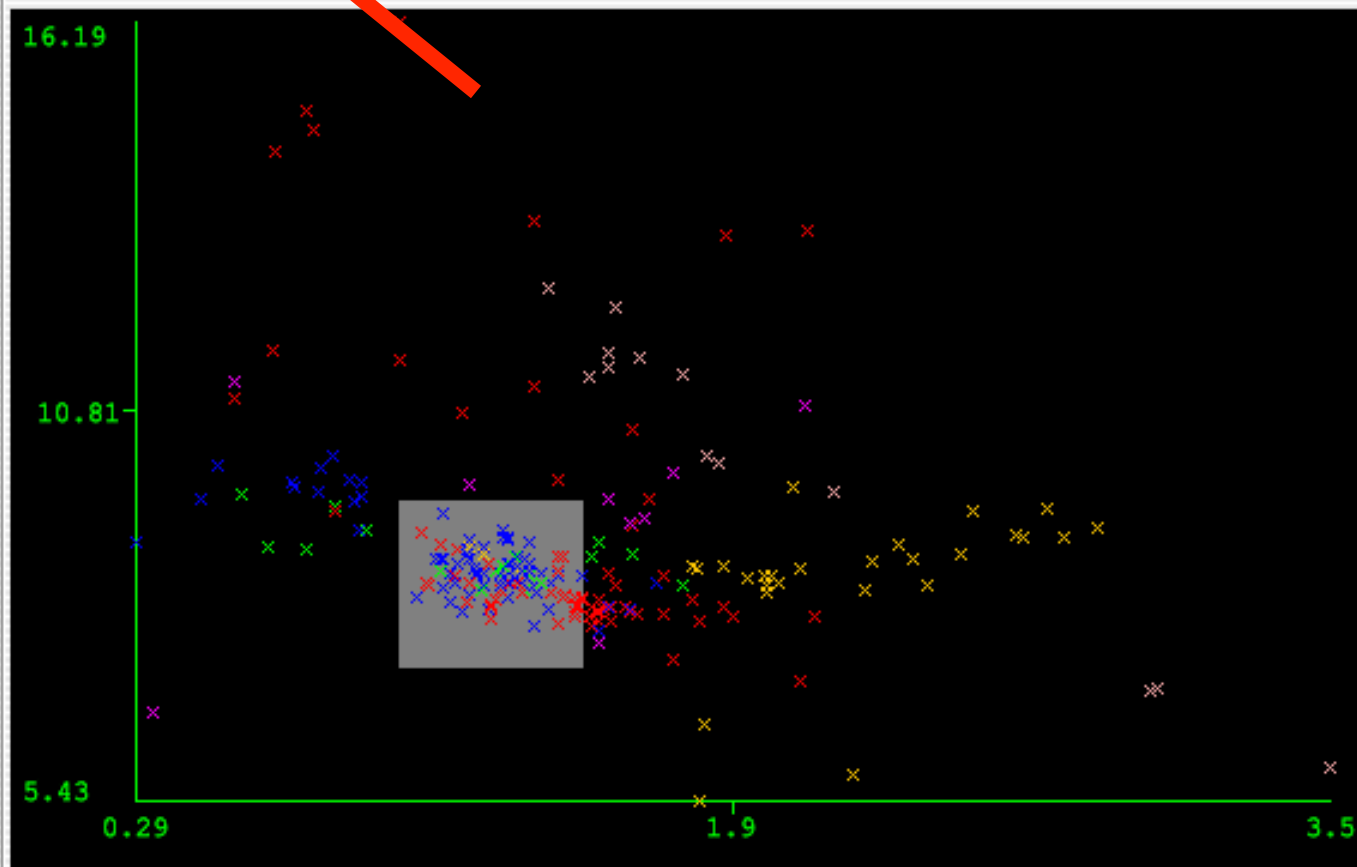
Clear

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

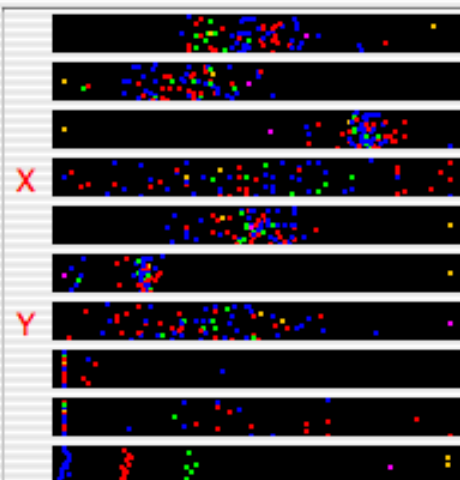
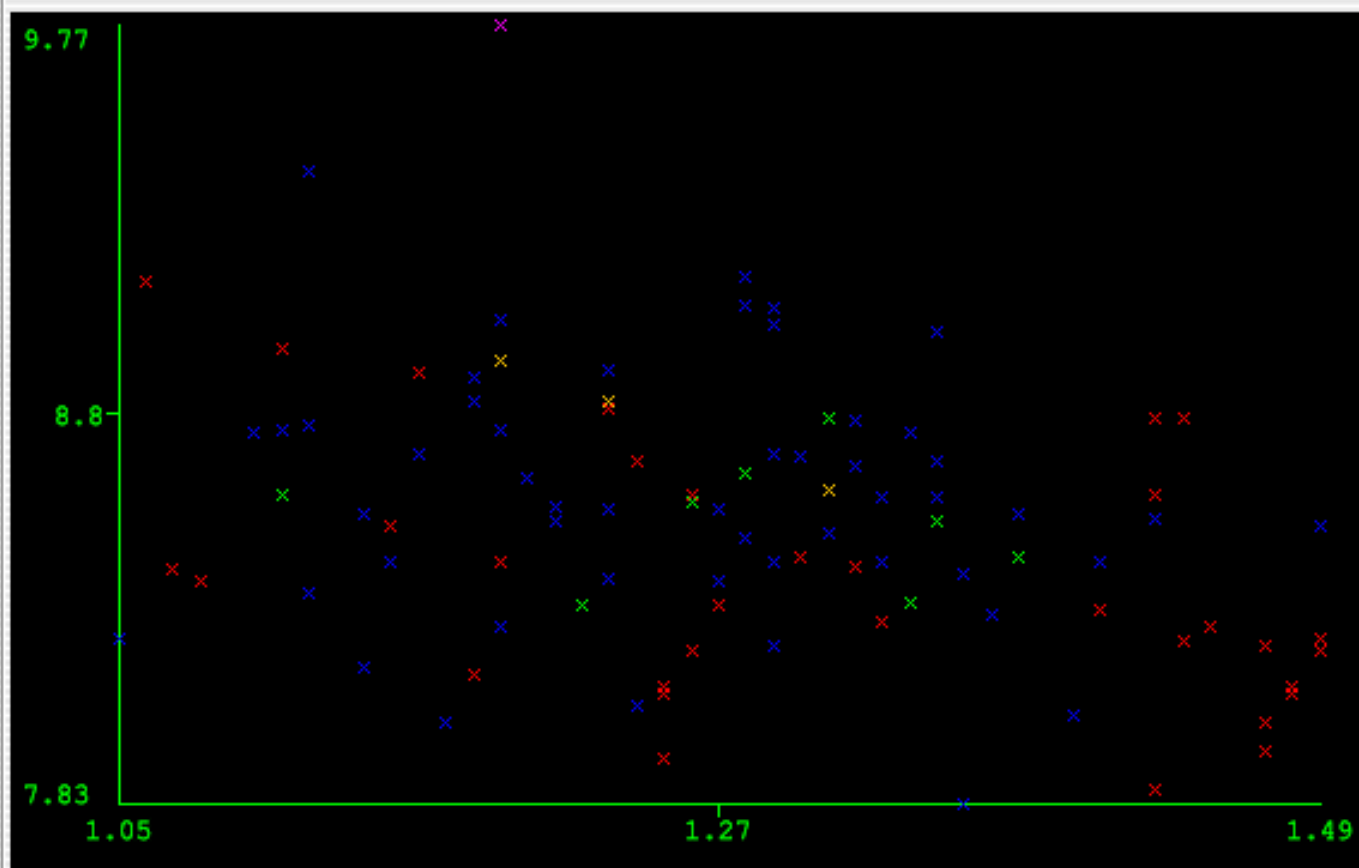
Clear

Save

Jitter



Plot: Glass



Class colour

build wind float

vehic wind non-float

build wind non-float

containers

vehic wind float

tableware

headlamps

References and Resources

- References:
 - WEKA website:
<http://www.cs.waikato.ac.nz/~ml/weka/index.html>
 - WEKA Tutorial:
 - Machine Learning with WEKA: A [presentation](#) demonstrating all graphical user interfaces (GUI) in Weka.
 - A [presentation](#) which explains how to use Weka for exploratory data mining.
 - WEKA Data Mining Book:
 - Ian H. Witten and Eibe Frank, Data Mining: Practical Machine Learning Tools and Techniques (Second Edition)
 - WEKA Wiki: http://weka.sourceforge.net/wiki/index.php/Main_Page
 - Others:
 - Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, 2nd ed.