Machine Learning the WEKA directions

Week-7

WEKA IS

- ✓ Machine Learning toolkit
- ✓ Open Source (Java)
- ✓ Well documented + huge community
- ✓ Provides API, command line and GUI-swing tools
- ✓ Relatively easy to learn

- ✓ Runnable on a remote server so you can “dumb terminal” your laptop and keep your data in one place!

WEKA is not…

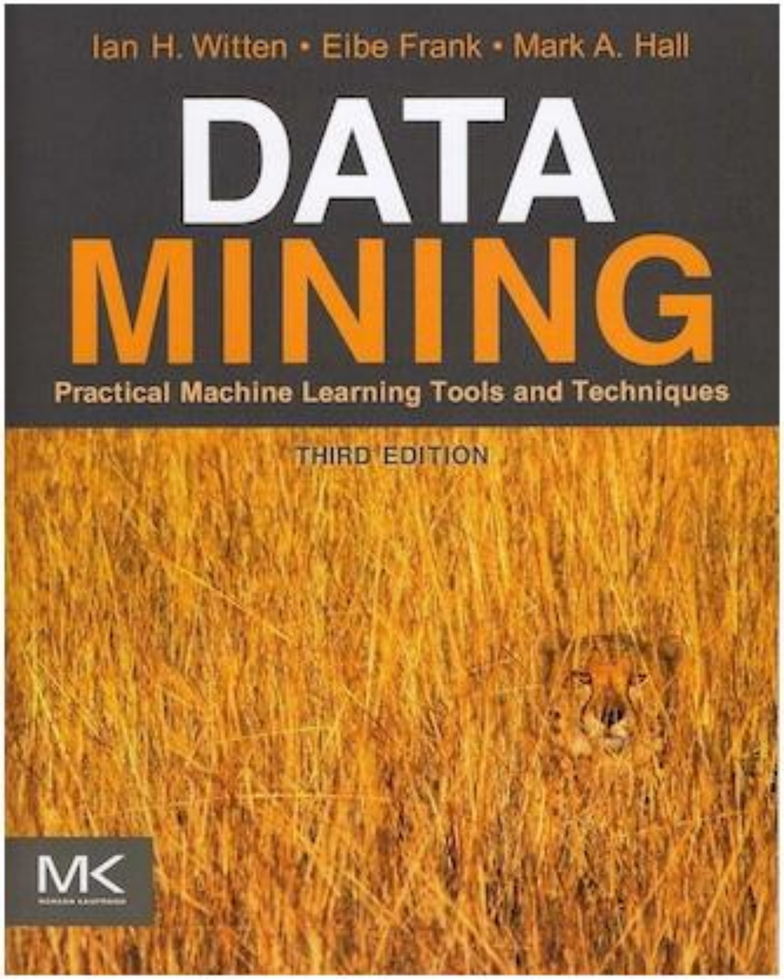
- A complete replacement for R/Matlab
- Optimized out of the box for multiple CPUs / compute farms
- An excuse to ignore the method details of a clustering/classification algorithm

Data Mining: Practical Machine Learning Tools and Techniques (Third Edition)

Ian H. Witten, Eibe Frank, Mark A. Hall

Morgan Kaufmann
January 2011
629 pages
Paper
ISBN
978-0-12-374856-0

Eibe Frank and Ian Witten



Ian H. Witten • Eibe Frank • Mark A. Hall
DATA MINING
Practical Machine Learning Tools and Techniques
THIRD EDITION
MK

Click here to order from Amazon.com

Personal WEKA dataset

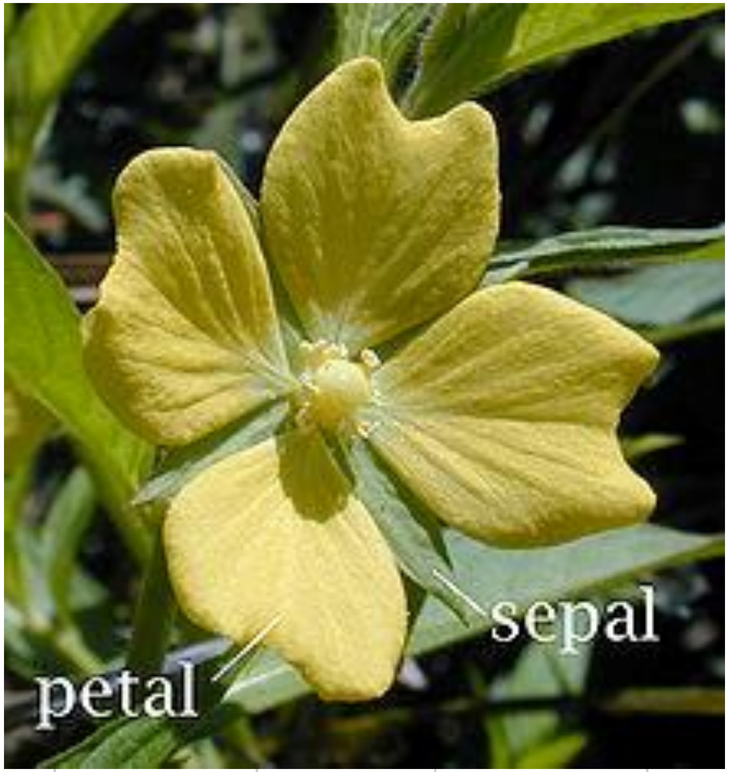
- ✓ 21k+ variables
- ✓ 14k+ subjects phenotyped
- ✓ 9k+ subjects genotyped 500k Affymetrix
- ✓ 54M recorded phenotype values of widely varying types

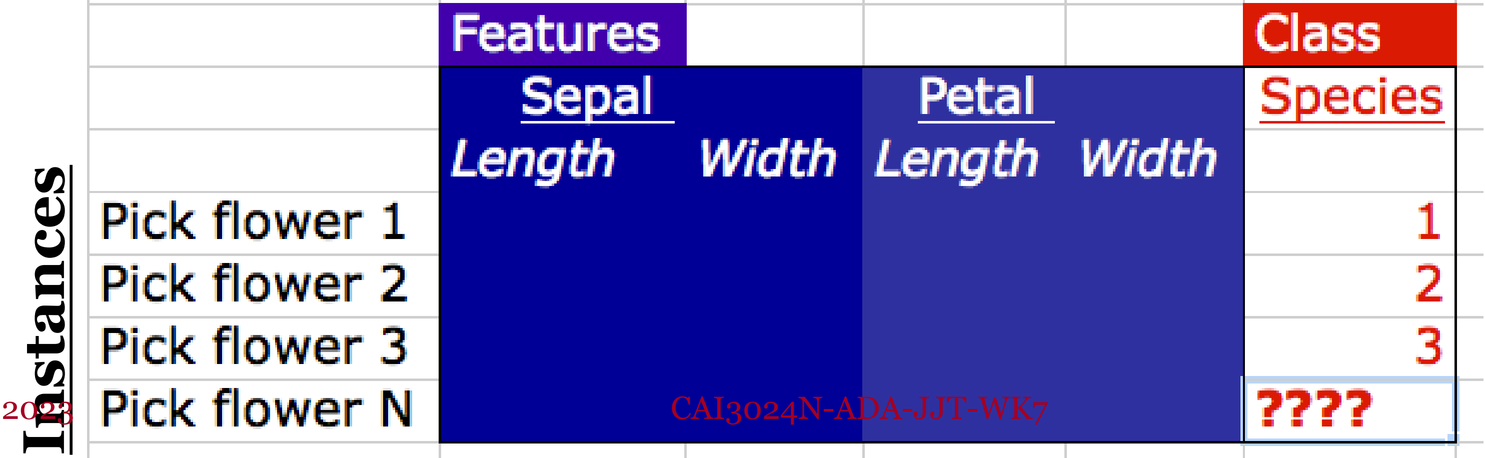
- ✓ Even the simplest correlation matrix
  - ✓ 20k * 20k = **400M** comparisons *before* including SNPs

WEKA basics

NOTE: GUI will be Slightly differ due to versions

Iris Example Data Set





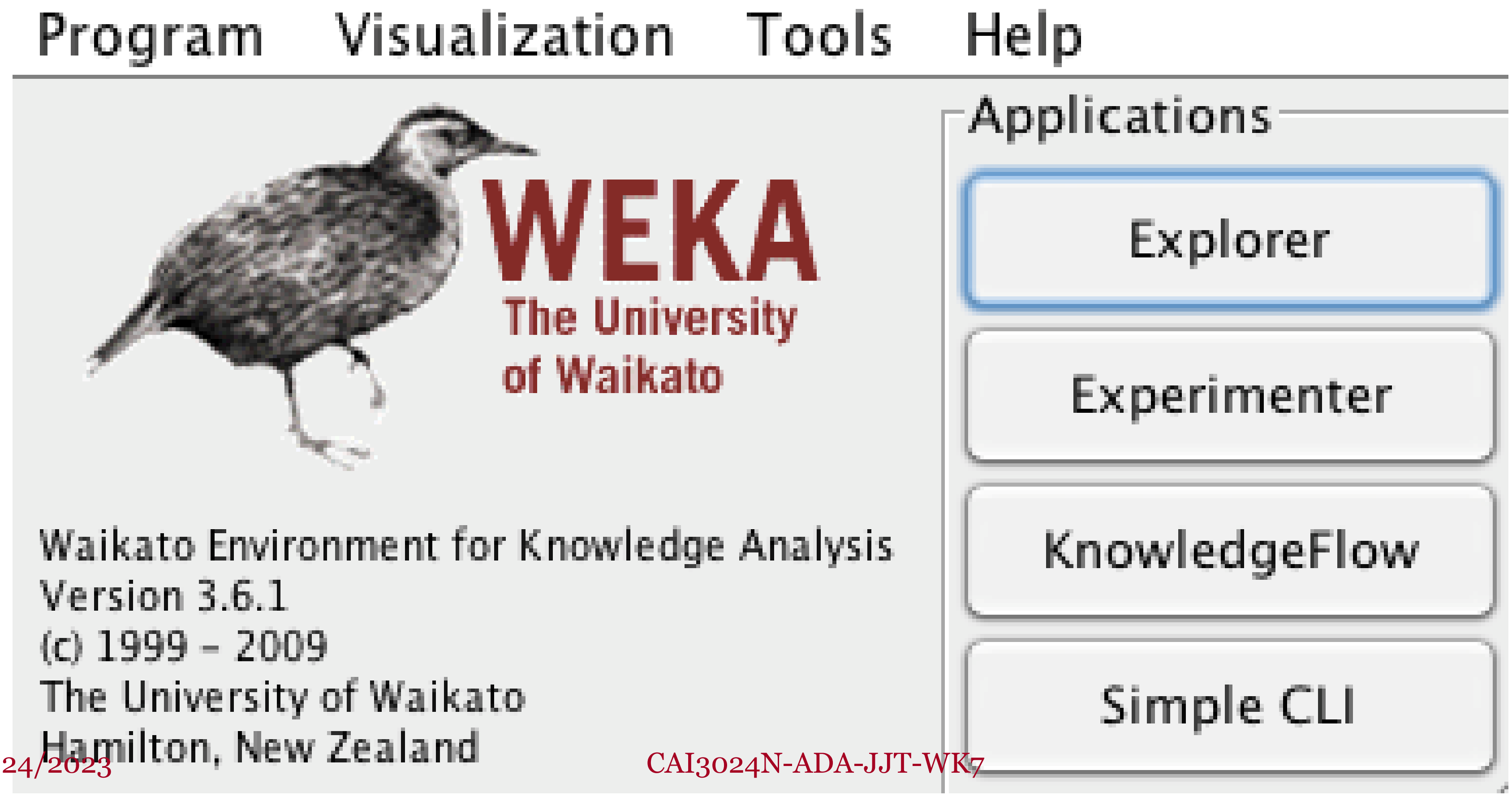
Instances

	Features				Class
	Sepal		Petal		Species
	Length	*Width*	*Length*	*Width*	
Pick flower 1					1
Pick flower 2					2
Pick flower 3					3
Pick flower N					**????**

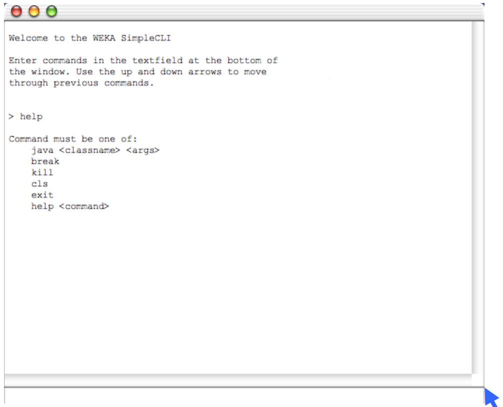
WEKA basics

NOTE: GUI will be Slightly differ due to versions

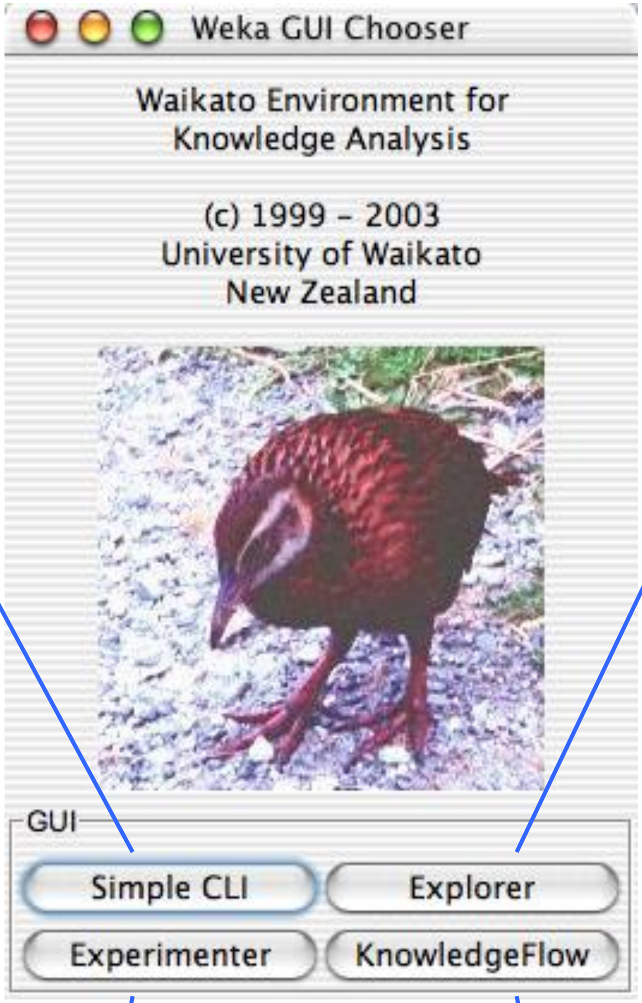
- API backs all functions of the CLI/GUI interfaces, can be easily used for your own project.



Waikato Environment for Knowledge Analysis
Version 3.6.1
(c) 1999 - 2009
The University of Waikato
Hamilton, New Zealand



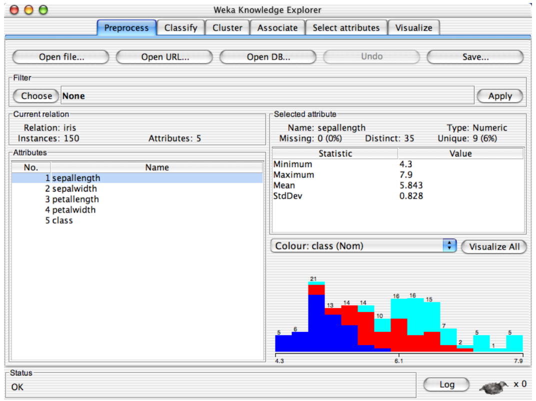
```
Welcome to the WEKA SimpleCLI

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


> help

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



Weka GUI Chooser

Waikato Environment for Knowledge Analysis

(c) 1999 - 2003
University of Waikato
New Zealand



Weka Knowledge Explorer

Filter: None

Current relation
Relation: iris
Instances: 150 Attributes: 5

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

Attributes

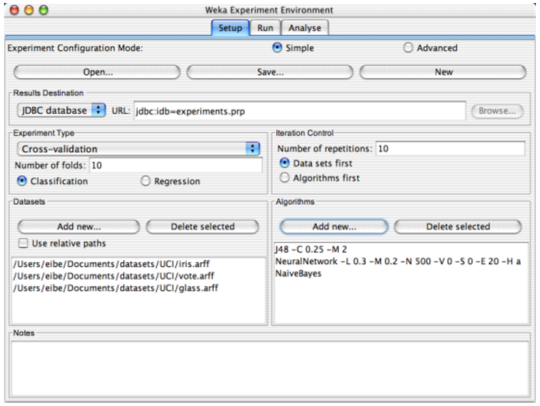
No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

Statistic	Value
Minimum	4.3
Maximum	7.9
Mean	5.843
StdDev	0.828

Colour: class (Nom)

Status
OK

x 0



Weka Experiment Environment

Experiment Configuration Mode: ● Simple ○ Advanced

Results Destination
JDBC database URL: jdbc:idb=experiments.prp

Experiment Type
Cross-validation
Number of folds: 10
● Classification ○ Regression

Iteration Control
Number of repetitions: 10
● Data sets first
○ Algorithms first

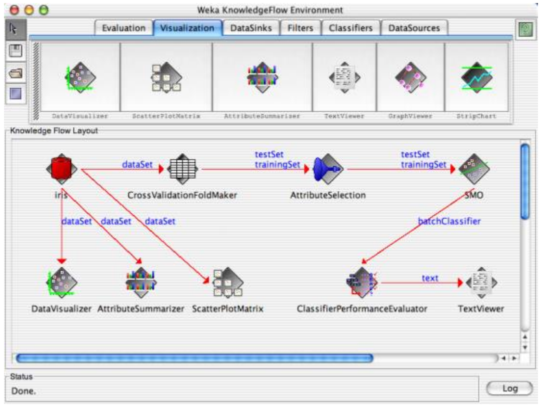
Datasets
□ Use relative paths

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

Algorithms

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 KnowledgeFlow Environment

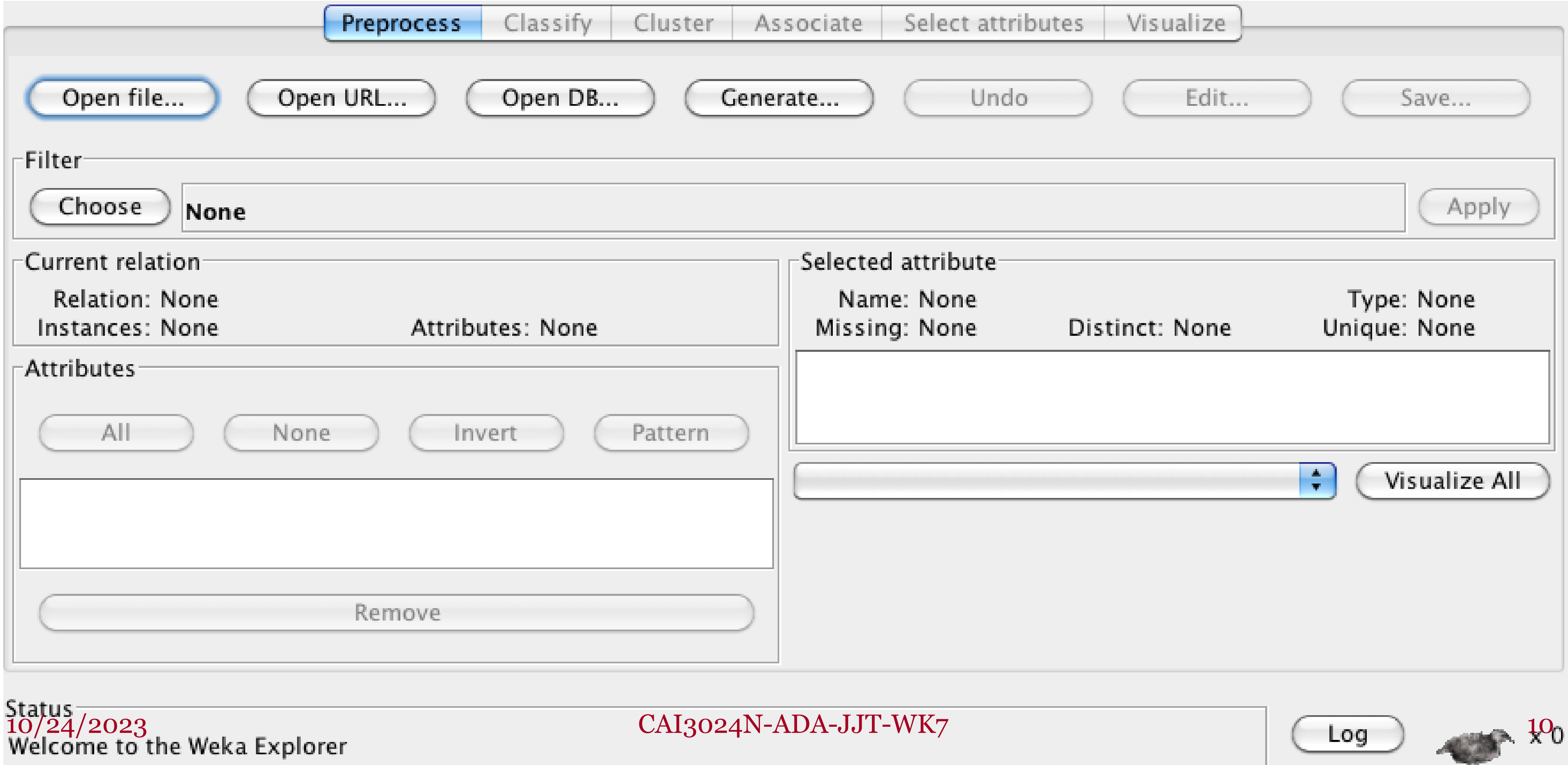
Status
Done.

WEKA Explorer Tutorial Examples

- **Preprocess**
  - **Instance and Attribute Filters (Supervised and Unsupervised)**
- Classify
  - Bayes
- Cluster
  - Expectation Maximization
  - Hierarchical Clustering
- Associate
  - Apriori
- Select Attributes
  - Via clustering

Preprocess

- File: CSV, ARFF*, ….
- Database: direct SQL access (useful)



Filter

None

Current relation

Relation: None
Instances: None
Attributes: None

Attributes

Selected attribute

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

Status

Welcome to the Weka Explorer

WEKA “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
...
```

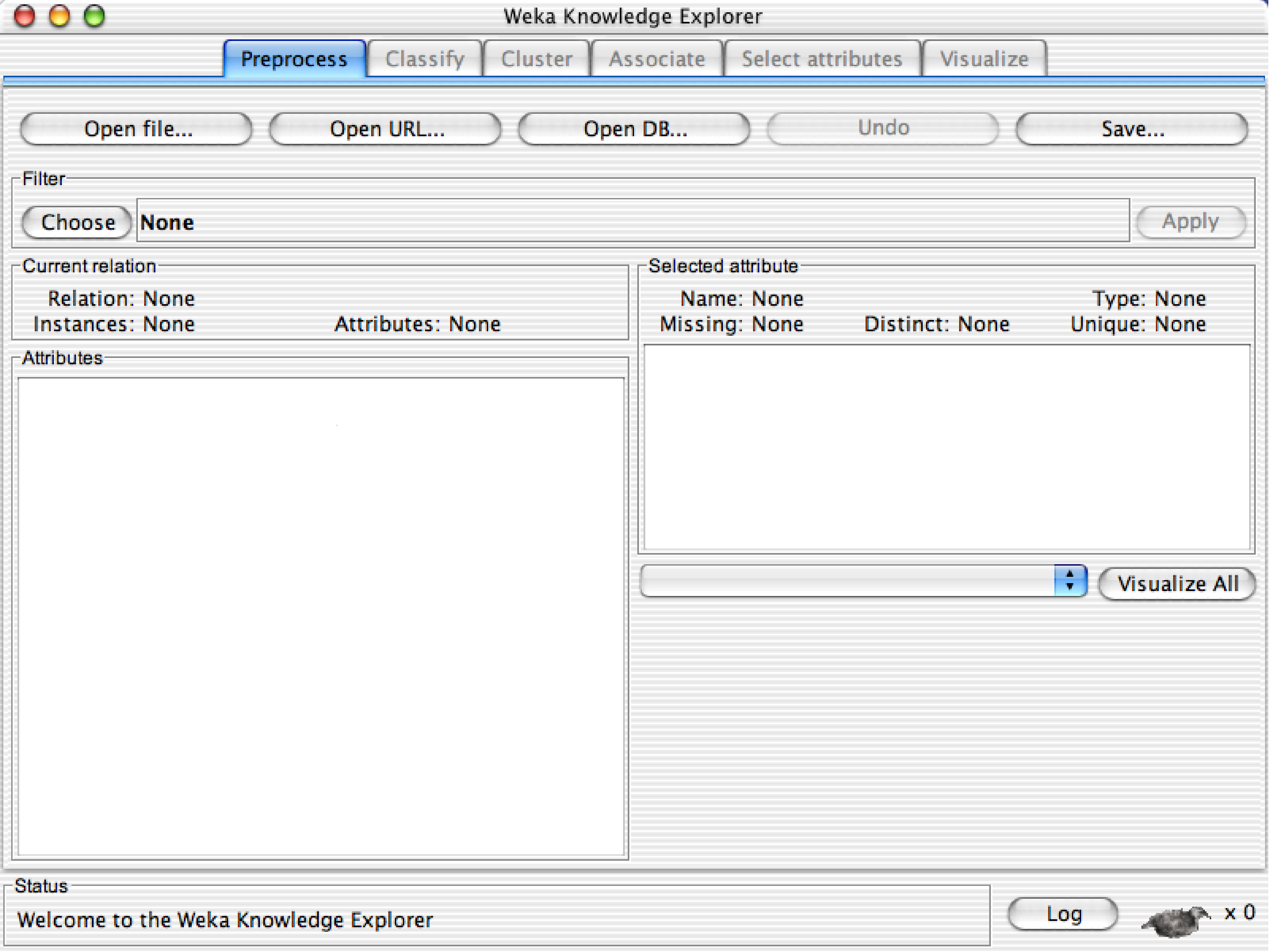
Flat file in ARFF format

WEKA “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

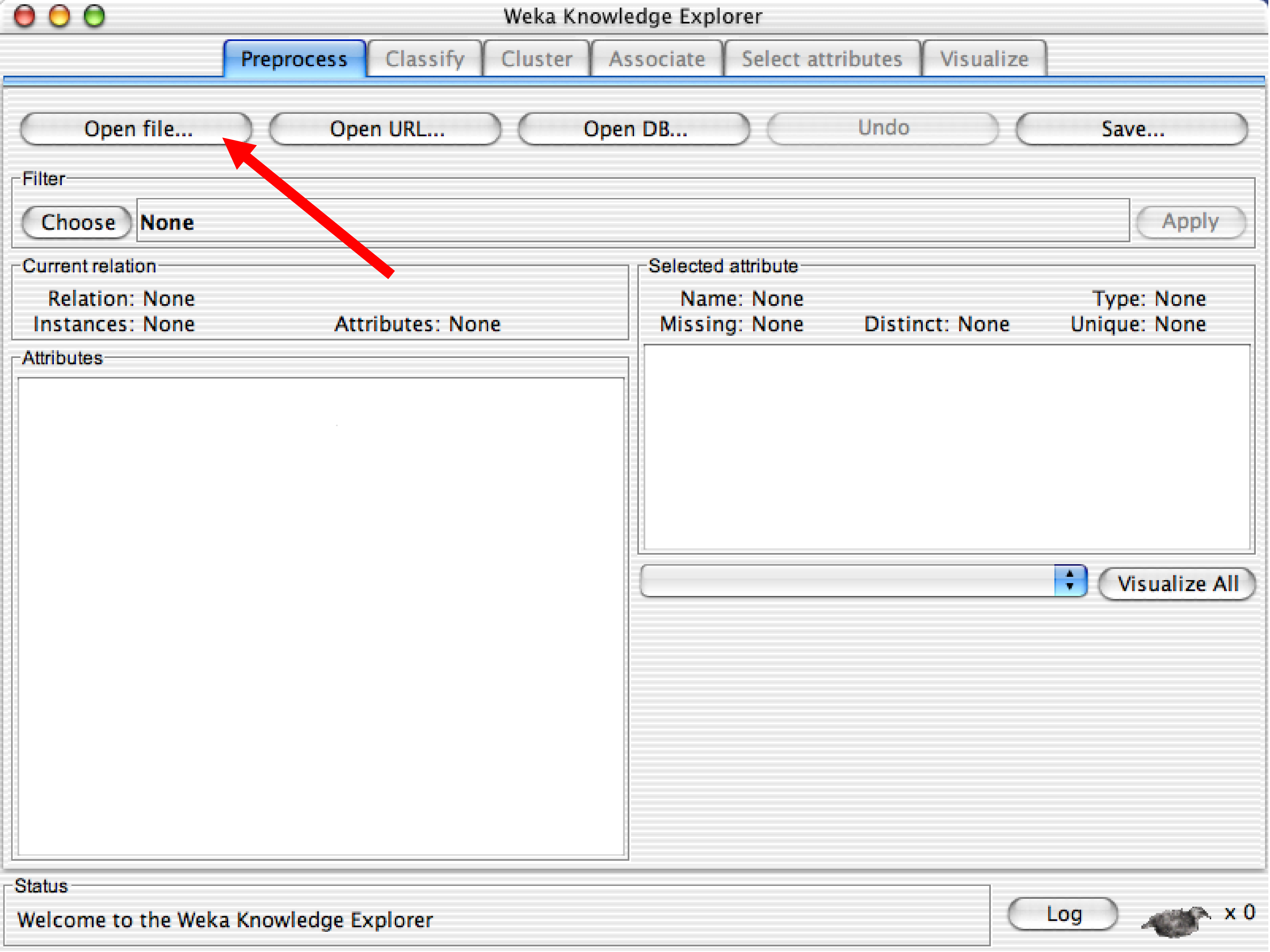


None

Relation: None
Instances: None
Attributes: None

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

Welcome to the Weka Knowledge Explorer

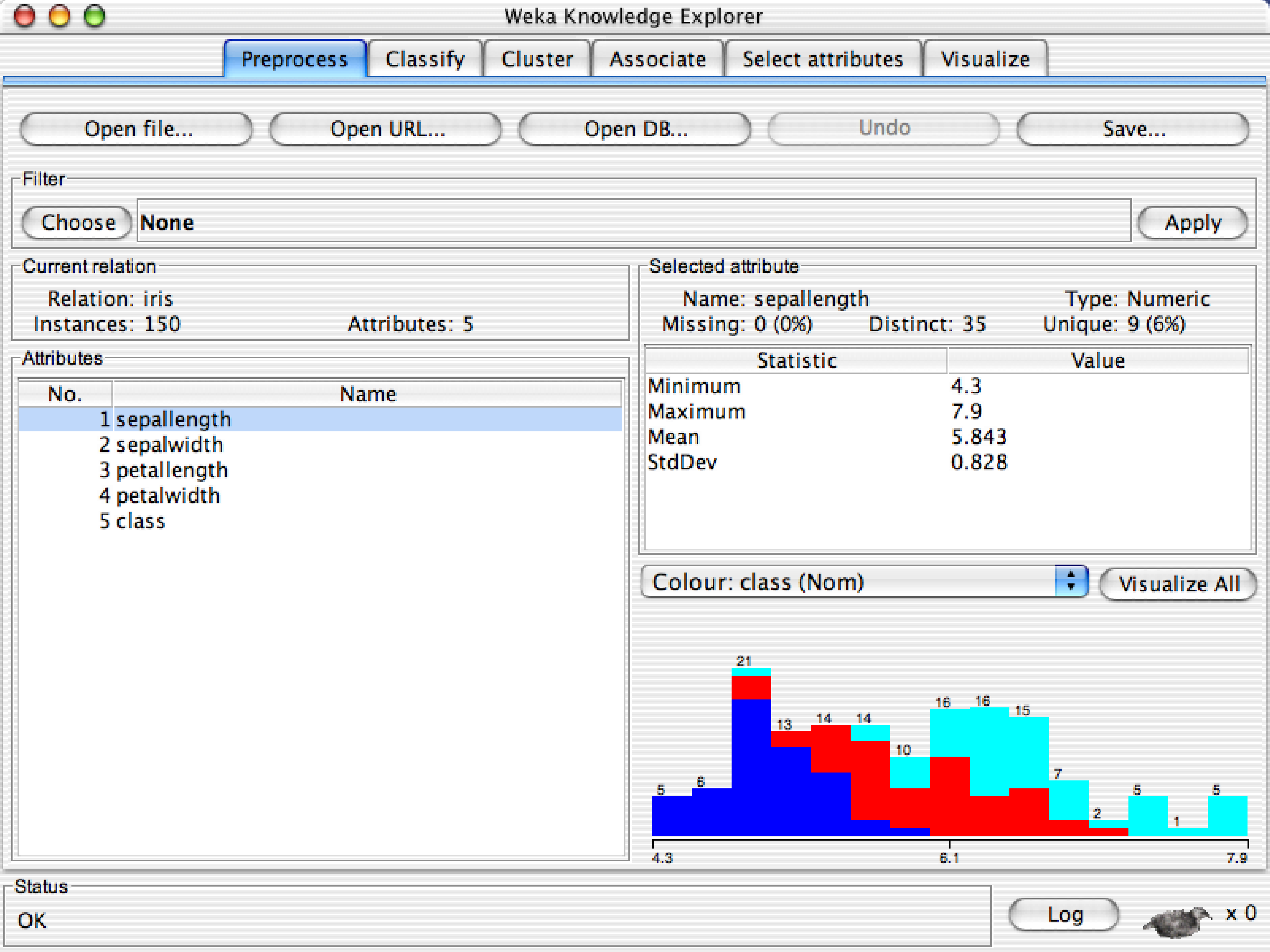


Filter: None

Relation: None
Instances: None
Attributes: None

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

Welcome to the Weka Knowledge Explorer



Filter: None

Relation: iris
Instances: 150 Attributes: 5

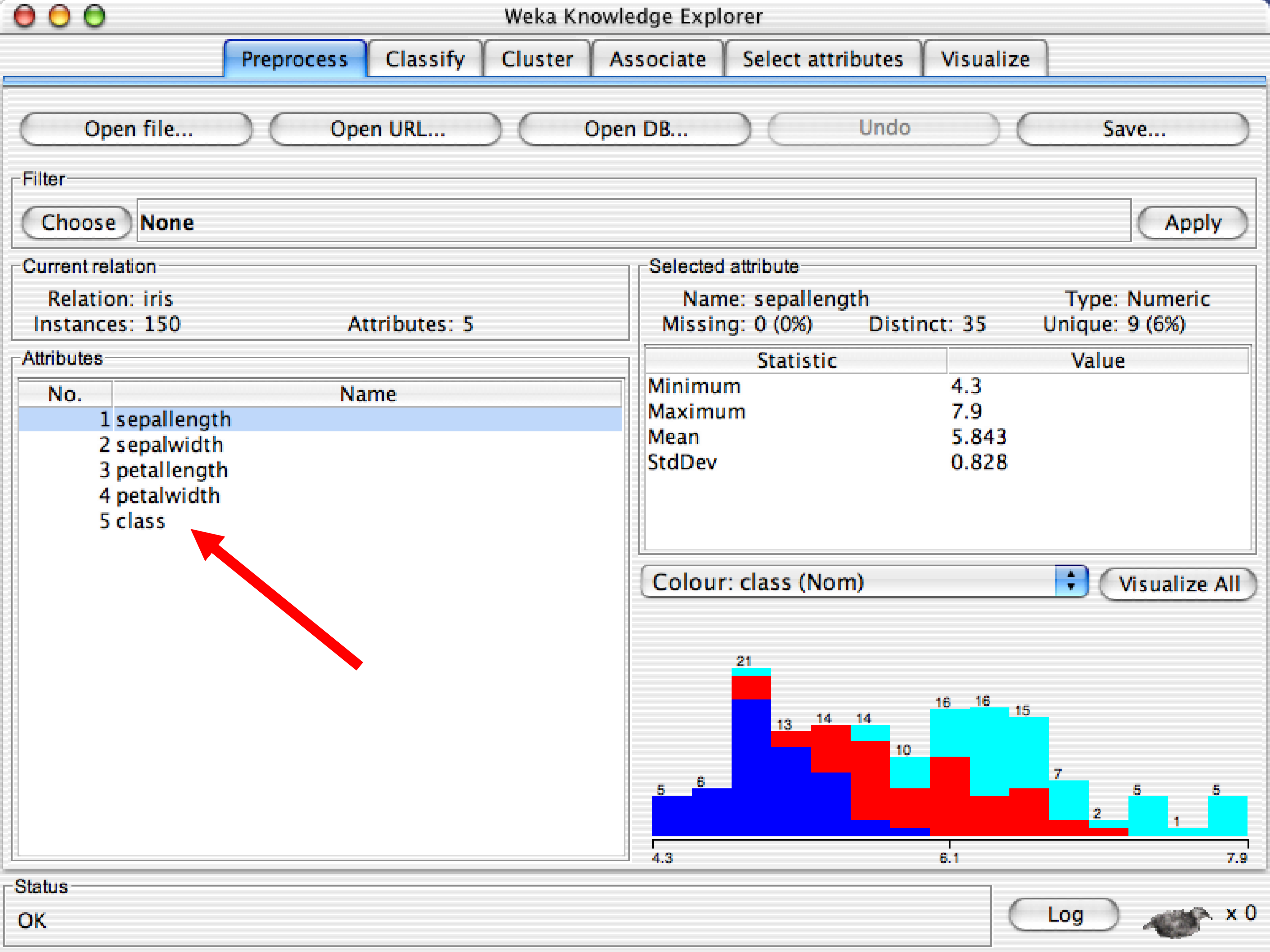
No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

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

Statistic	Value
Minimum	4.3
Maximum	7.9
Mean	5.843
StdDev	0.828

Colour: class (Nom)

OK



Current relation

Relation: iris
Instances: 150 Attributes: 5

Attributes

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

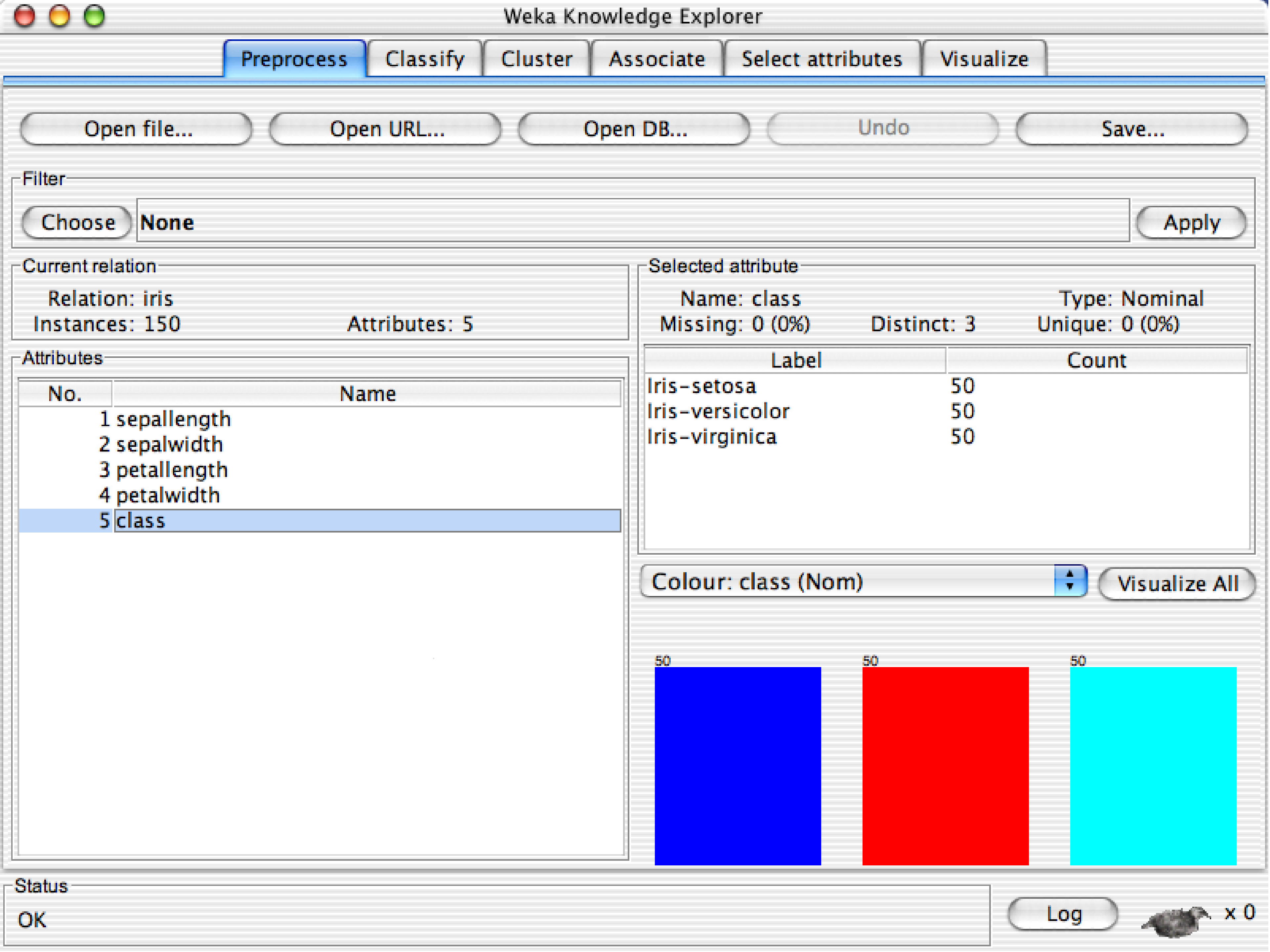
Selected attribute

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

Status

OK



Filter

None

Current relation

Relation: iris
Instances: 150 Attributes: 5

Attributes

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

Selected attribute

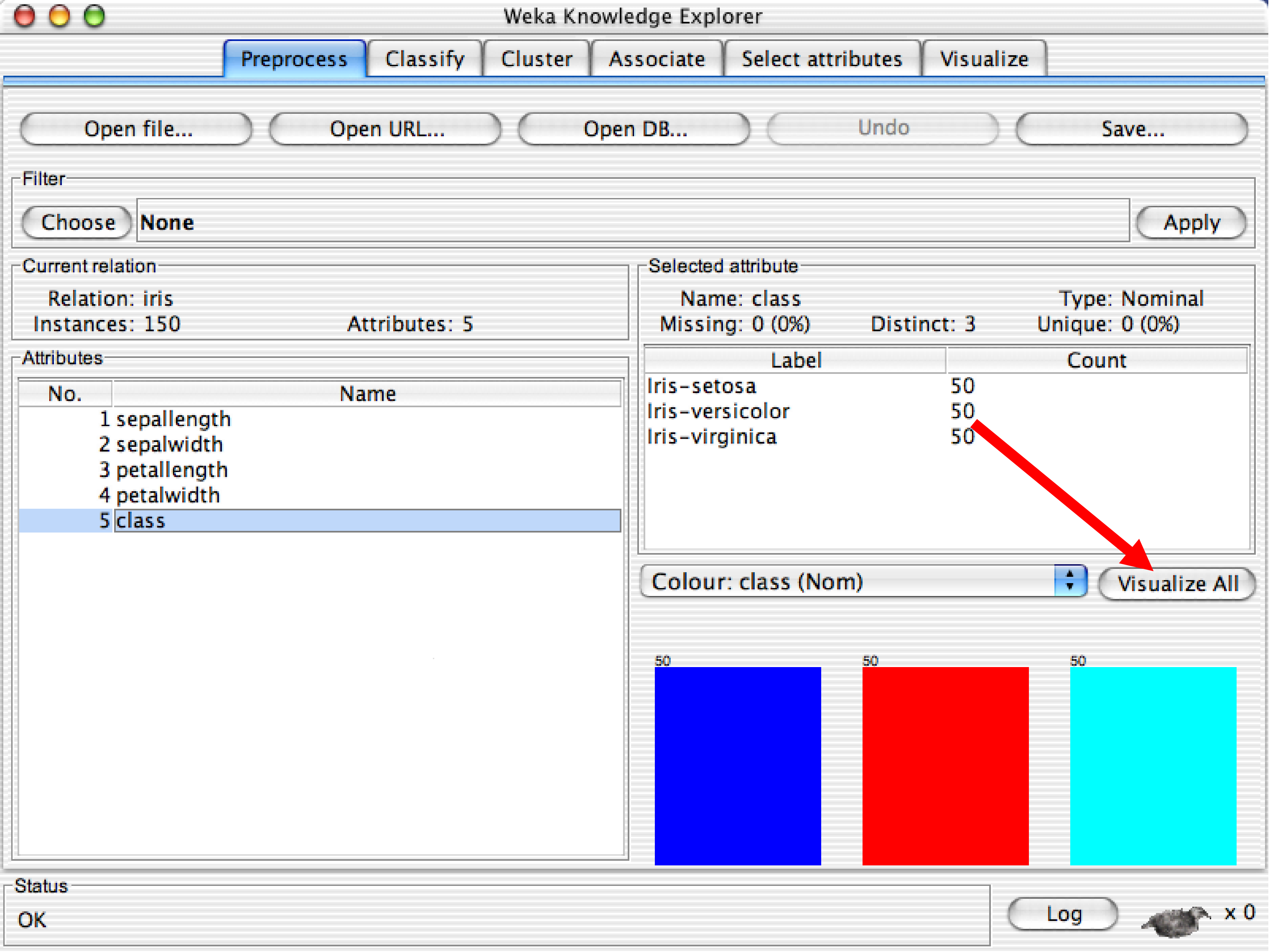
Name: class Type: Nominal
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

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

Colour: class (Nom)

Status

OK



Filter

None

Current relation

Relation: iris

Instances: 150 Attributes: 5

Attributes

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

Selected attribute

Name: class Type: Nominal

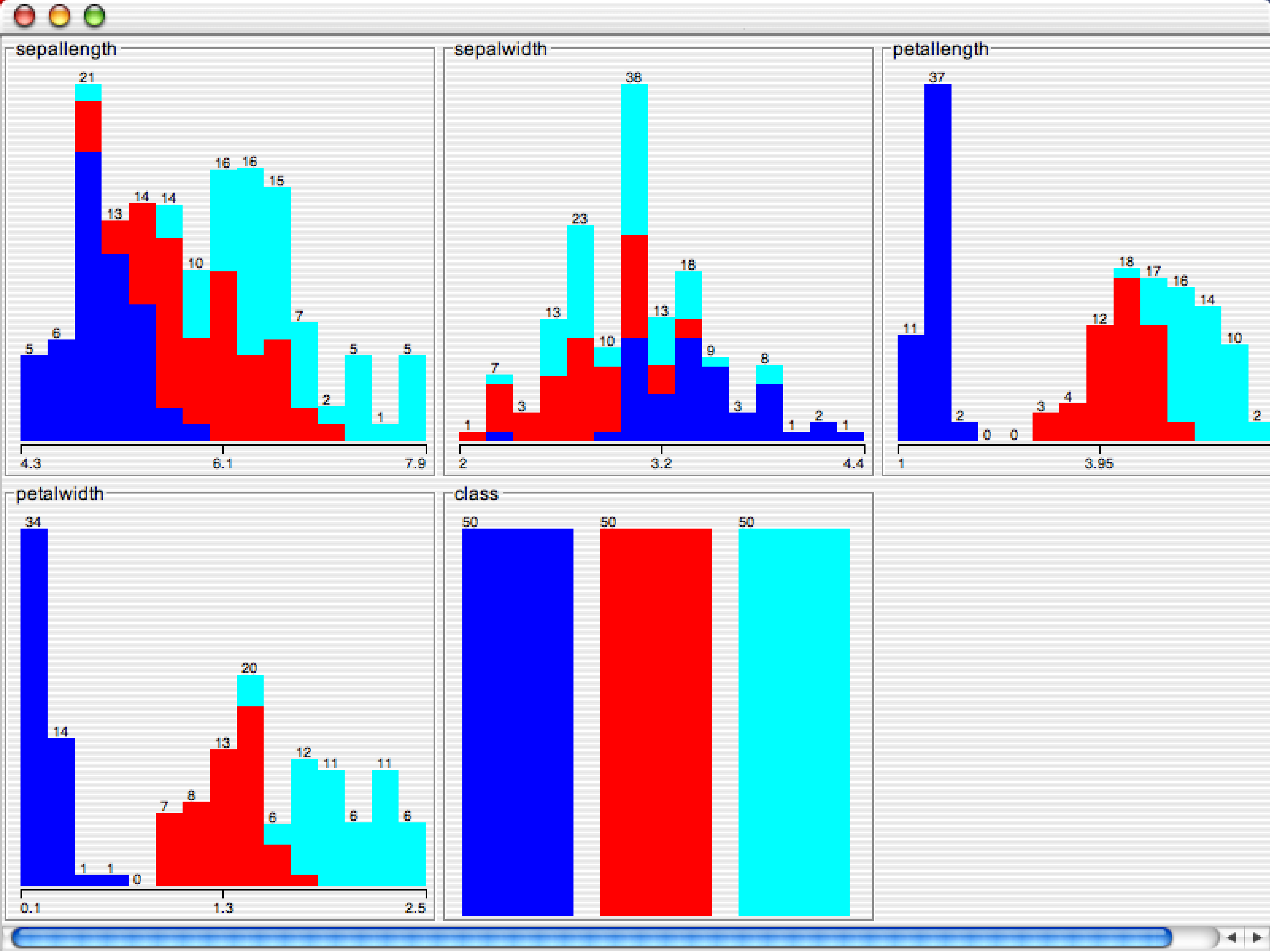
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

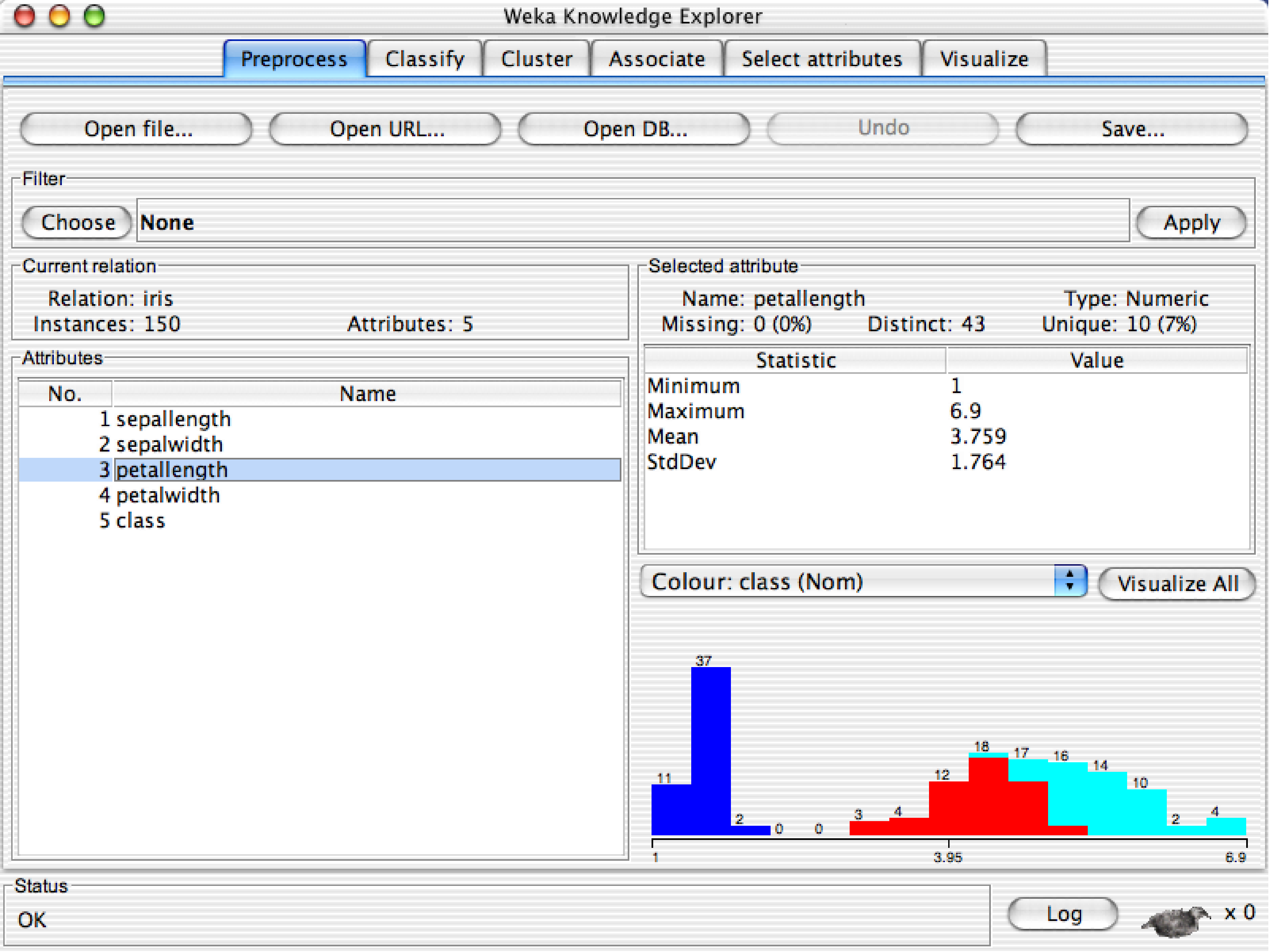
Label	Count
Iris-setosa	50
Iris-versicolor	50
Iris-virginica	50

Colour: class (Nom)

Status

OK


sepallength
21
13
14
14
16
16
15
10
7
6
5
5
5
2
1
4.3
6.1
7.9
sepalwidth
38
23
13
18
13
10
9
8
7
3
3
1
2
1
1
2
3.2
4.4
petallength
37
11
2
0
0
3
4
12
18
17
16
14
10
2
1
3.95
petalwidth
34
14
1
1
0
7
8
13
20
6
12
11
6
11
6
0.1
1.3
2.5
class
50
50
50



Filter

None

Current relation

Relation: iris

Instances: 150 Attributes: 5

Attributes

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

Selected attribute

Name: petallength Type: Numeric

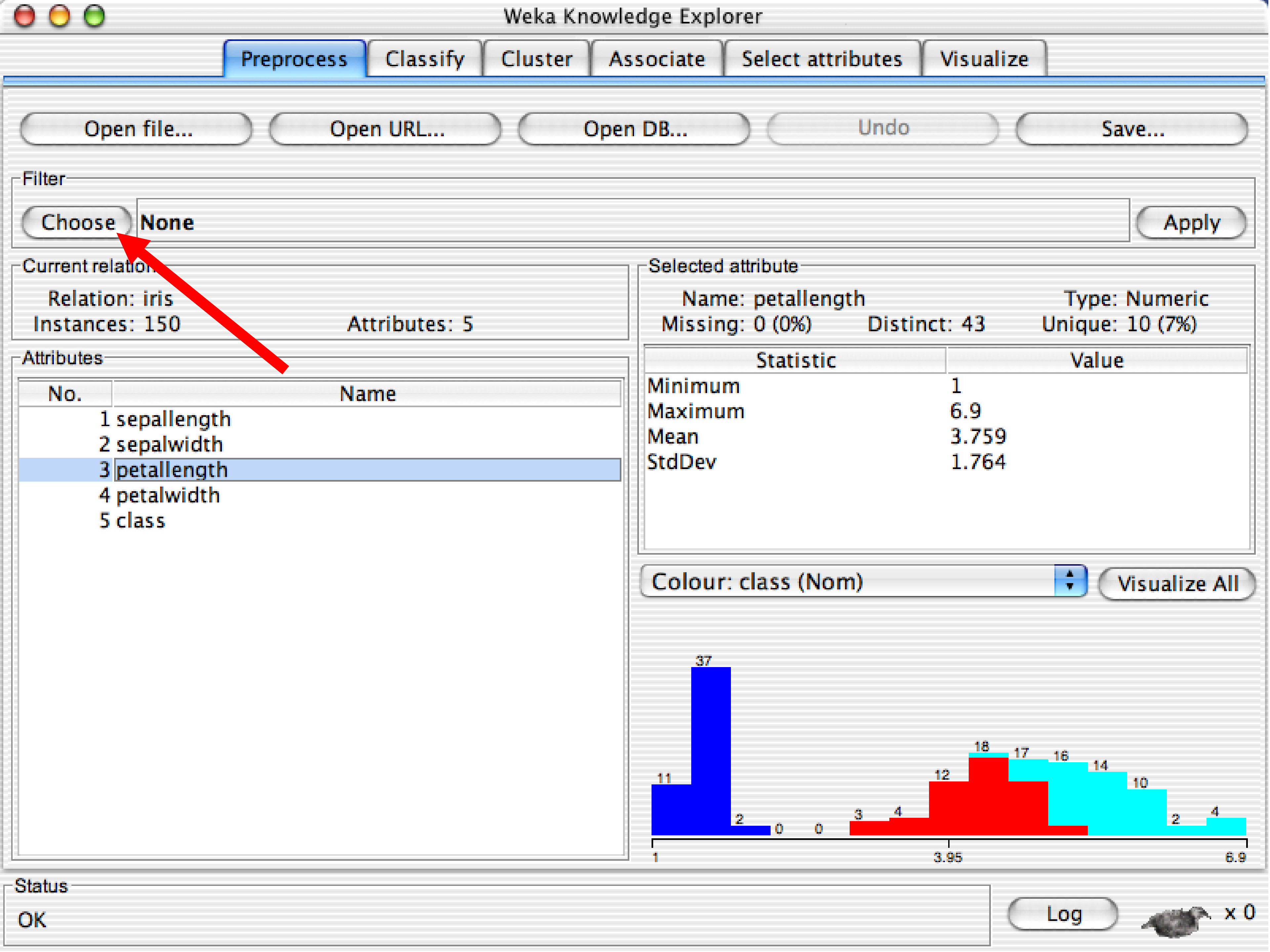
Missing: 0 (0%) Distinct: 43 Unique: 10 (7%)

Statistic	Value
Minimum	1
Maximum	6.9
Mean	3.759
StdDev	1.764

Colour: class (Nom)

Status

OK



Filter: None

Relation: iris
Instances: 150 Attributes: 5

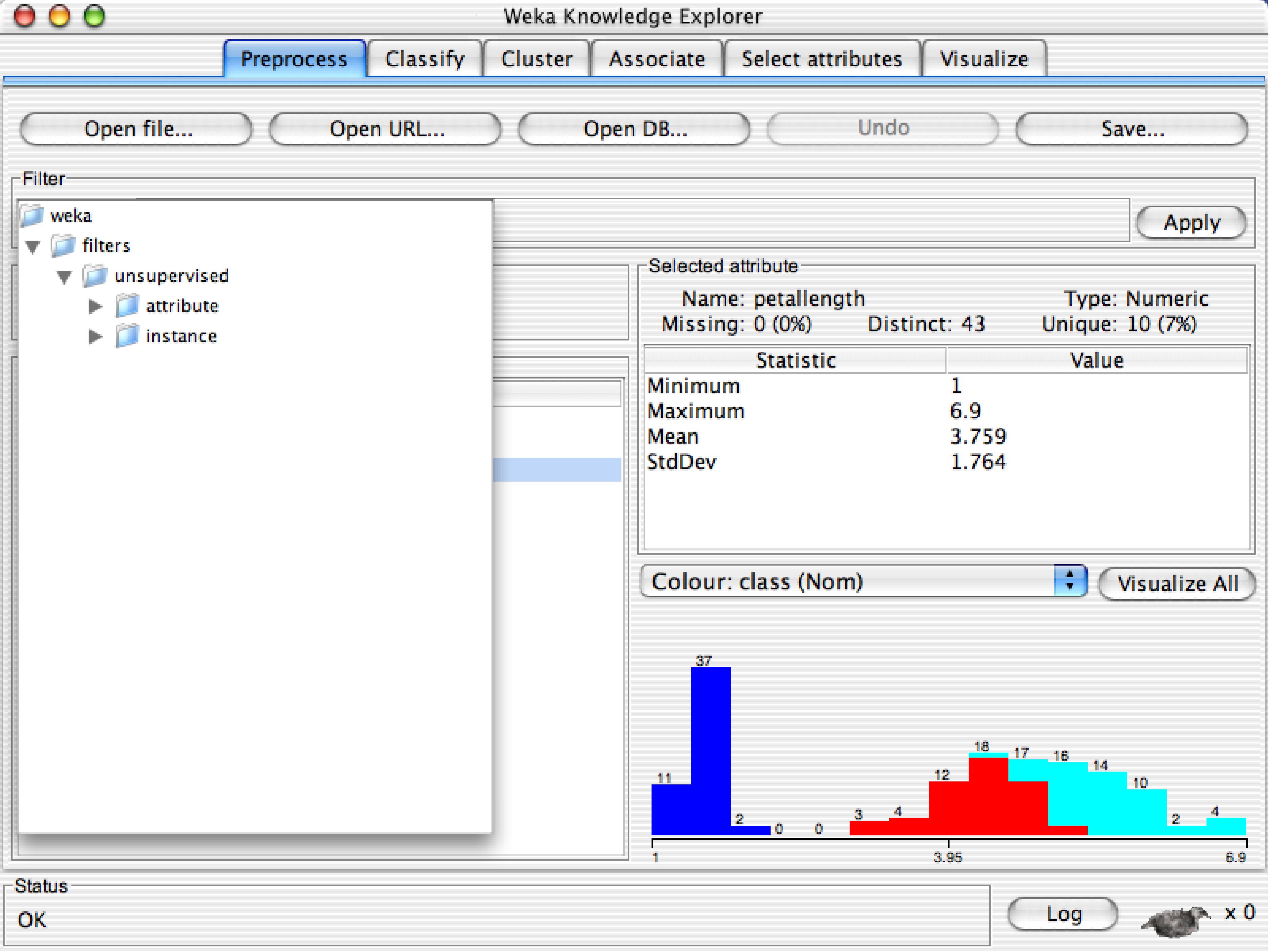
No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

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

Statistic	Value
Minimum	1
Maximum	6.9
Mean	3.759
StdDev	1.764

Colour: class (Nom)

OK



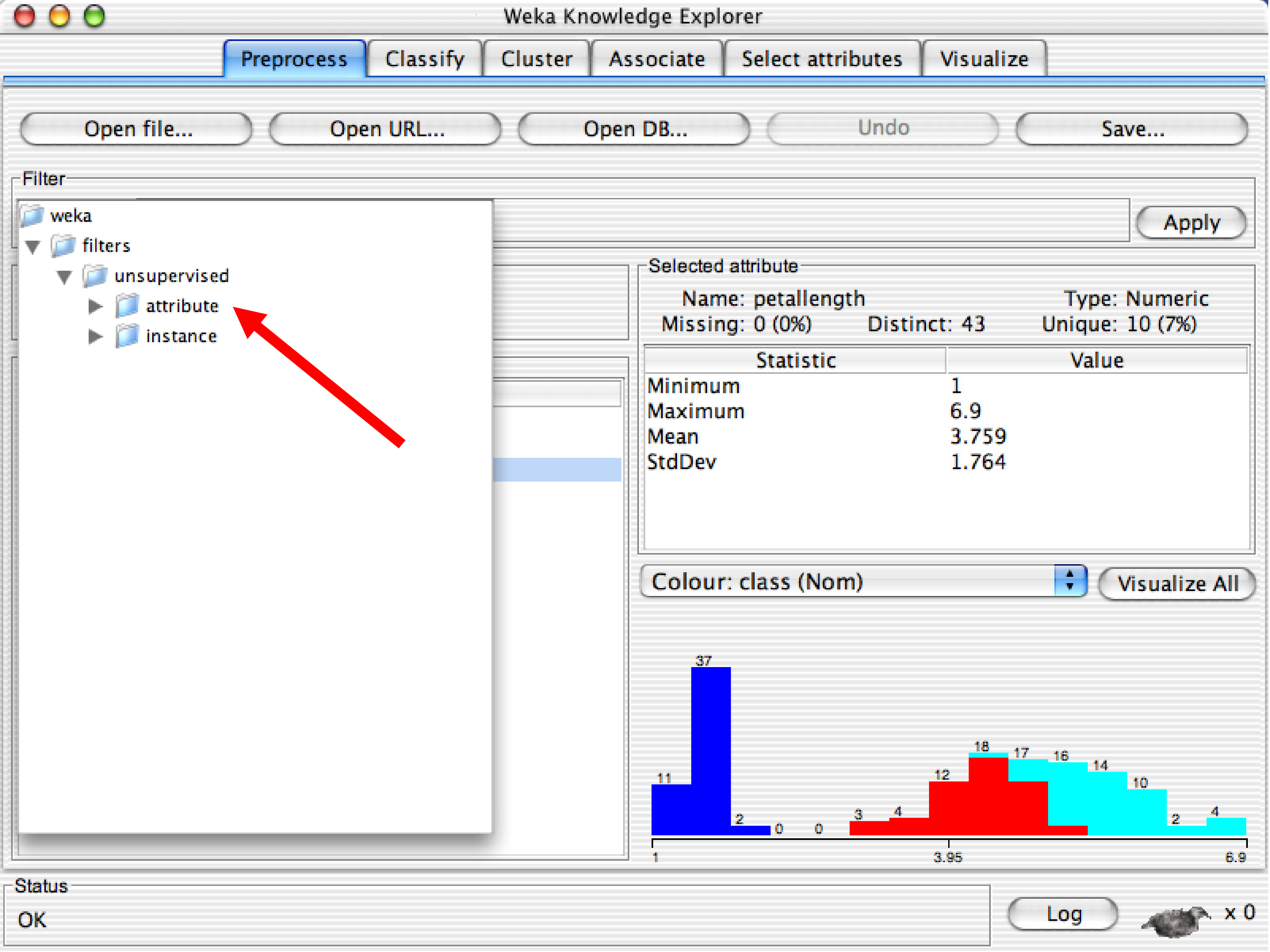
Selected attribute

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

Statistic	Value
Minimum	1
Maximum	6.9
Mean	3.759
StdDev	1.764

Status

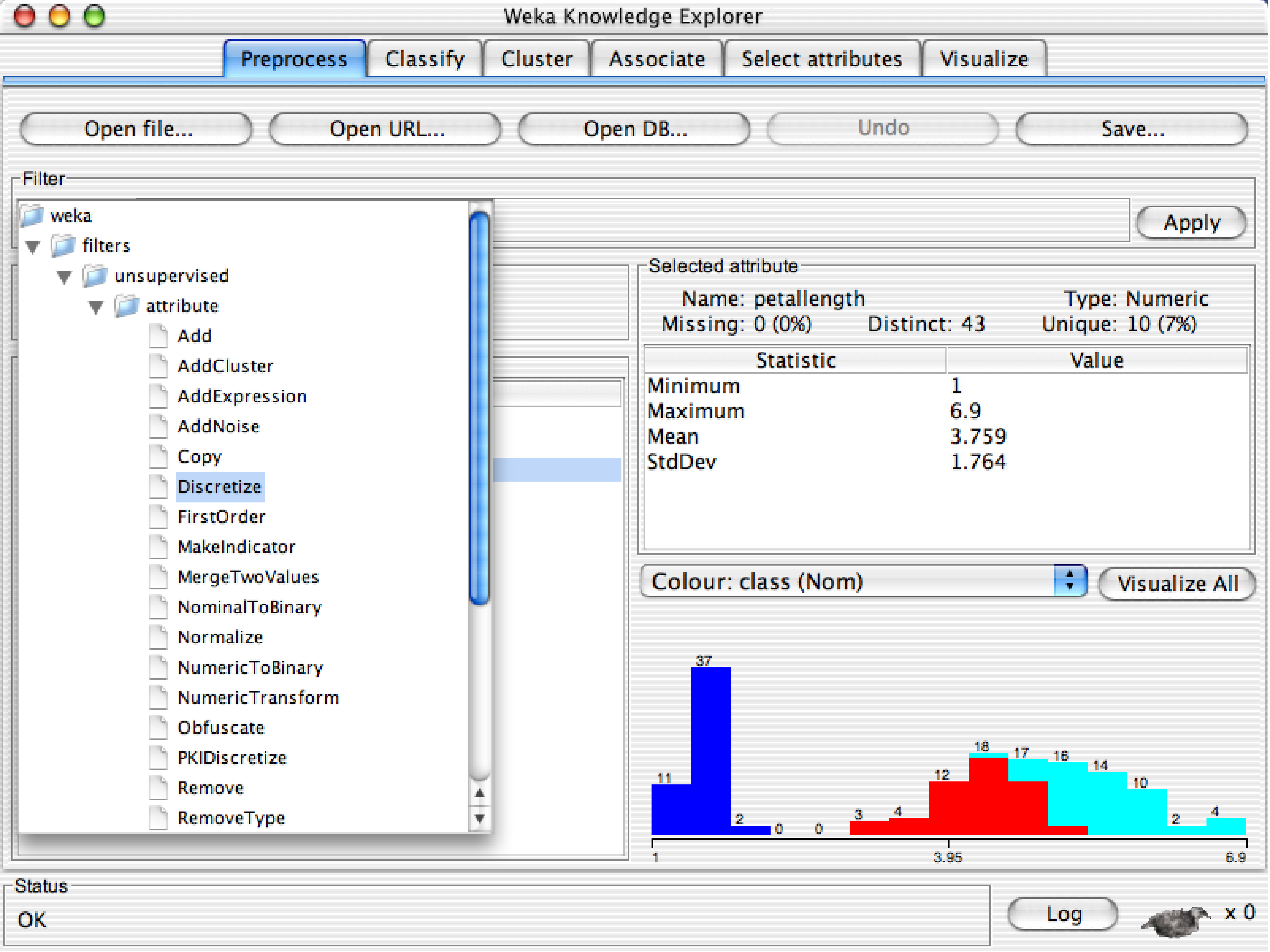
OK



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

OK



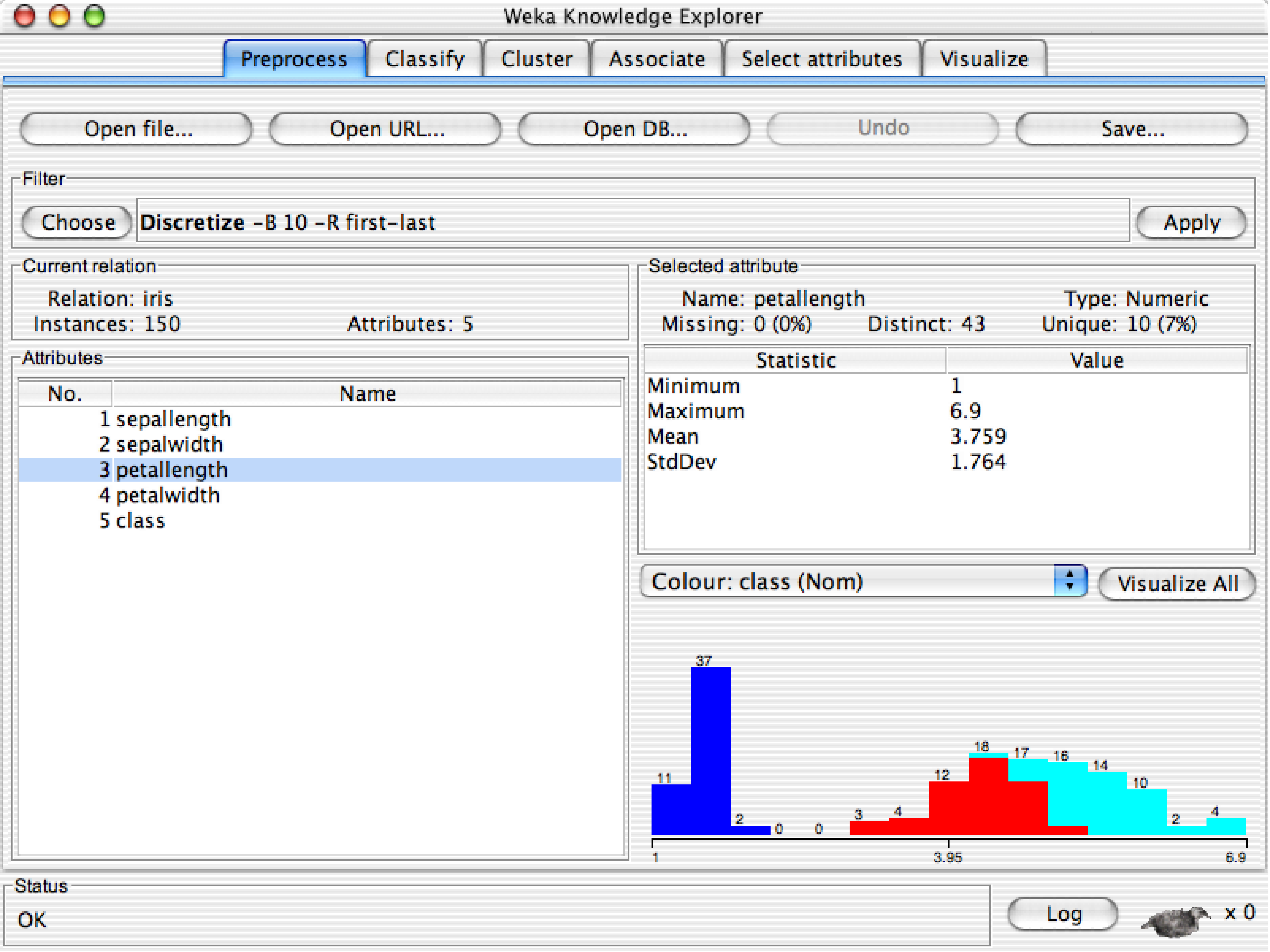
Selected attribute

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

Statistic	Value
Minimum	1
Maximum	6.9
Mean	3.759
StdDev	1.764

Colour: class (Nom)

Status
OK



Discretize -B 10 -R first-last

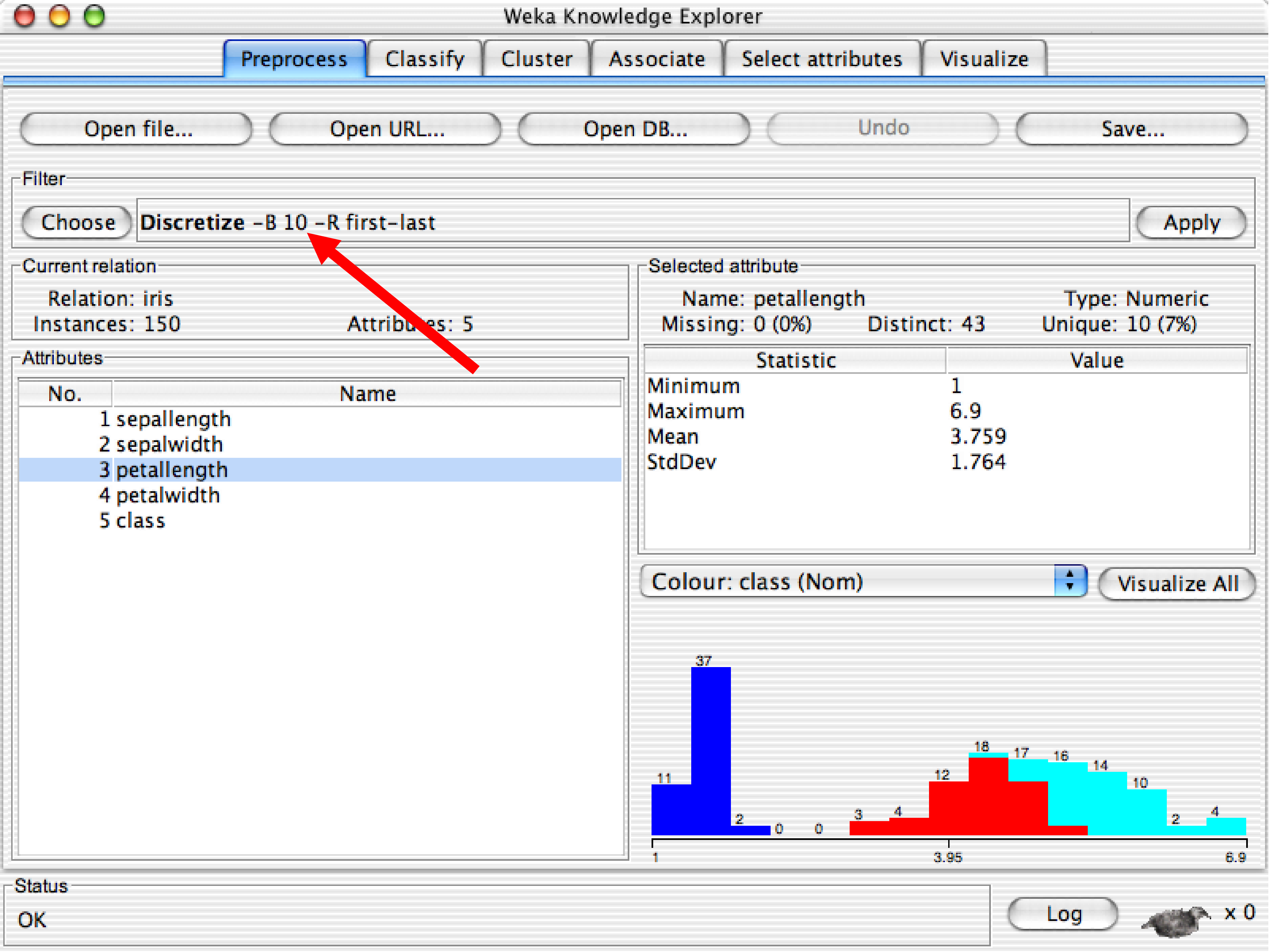
Relation: iris
Instances: 150
Attributes: 5

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

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

OK



Filter: Discretize -B 10 -R first-last

Current relation

Relation: iris
Instances: 150 Attributes: 5

Attributes

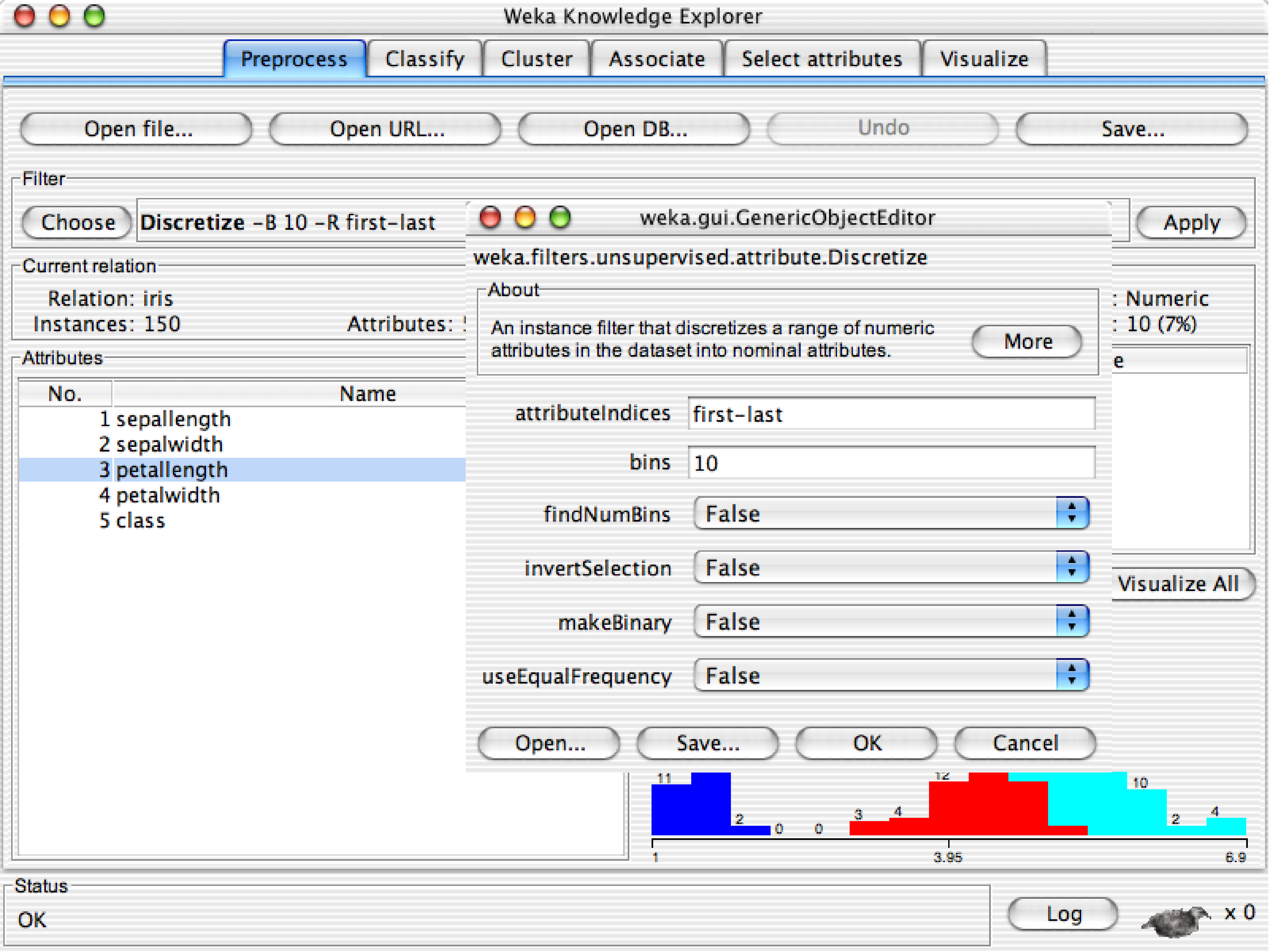
No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

Selected attribute

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

Statistic	Value
Minimum	1
Maximum	6.9
Mean	3.759
StdDev	1.764

Status
OK



Weka Knowledge Explorer

Filter

Discretize -B 10 -R first-last

weka.gui.GenericObjectEditor

weka.filters.unsupervised.attribute.Discretize

About

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

attributeIndices first-last

bins 10

findNumBins False

invertSelection False

makeBinary False

useEqualFrequency False

Current relation

Relation: iris
Instances: 150 Attributes: 5

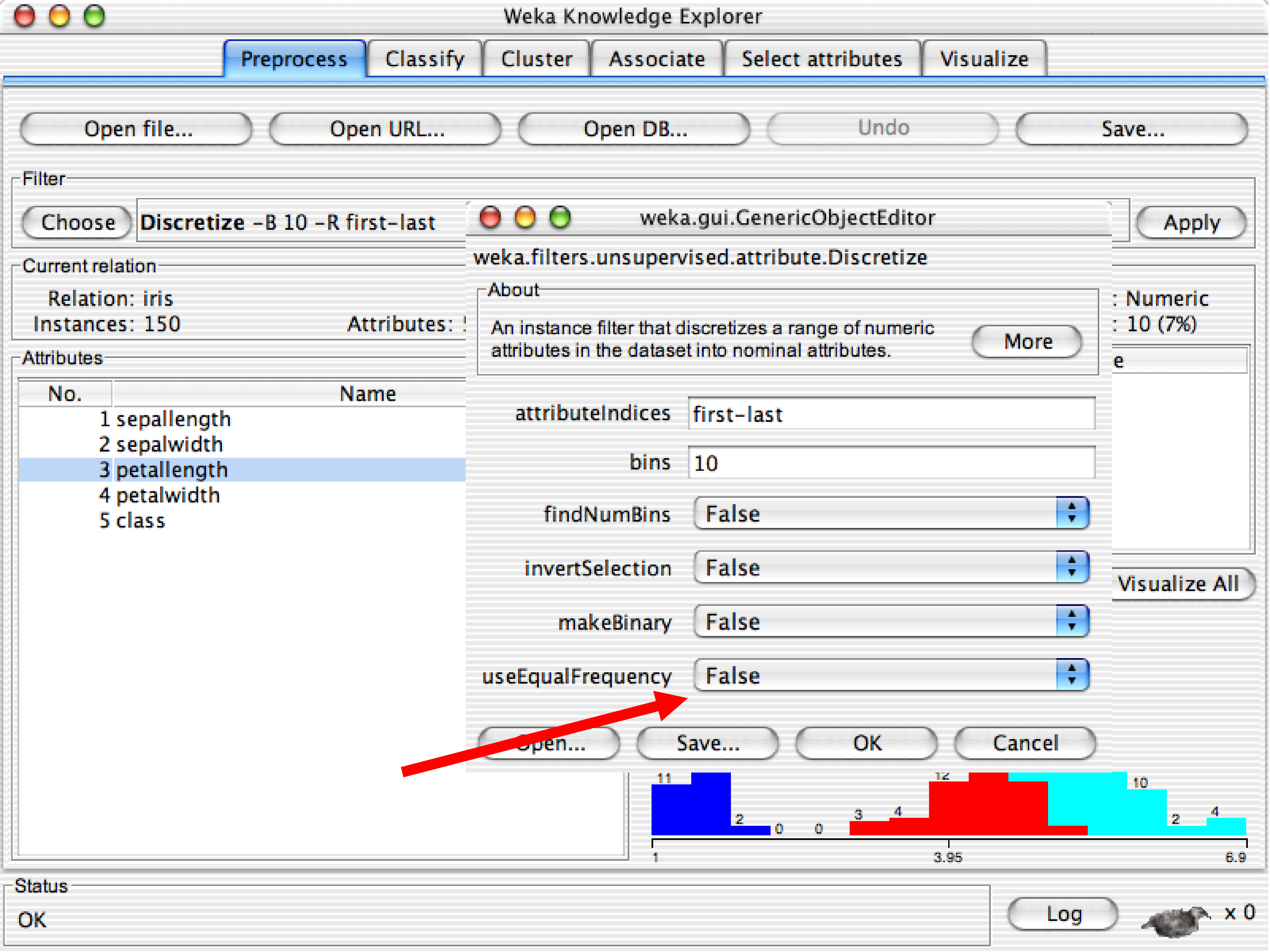
: Numeric
: 10 (7%)

Attributes

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

Status

OK



Weka Knowledge Explorer

Filter

Discretize -B 10 -R first-last

Current relation

Relation: iris
Instances: 150
Attributes: 5

: Numeric
: 10 (7%)

Attributes

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

weka.gui.GenericObjectEditor

weka.filters.unsupervised.attribute.Discretize

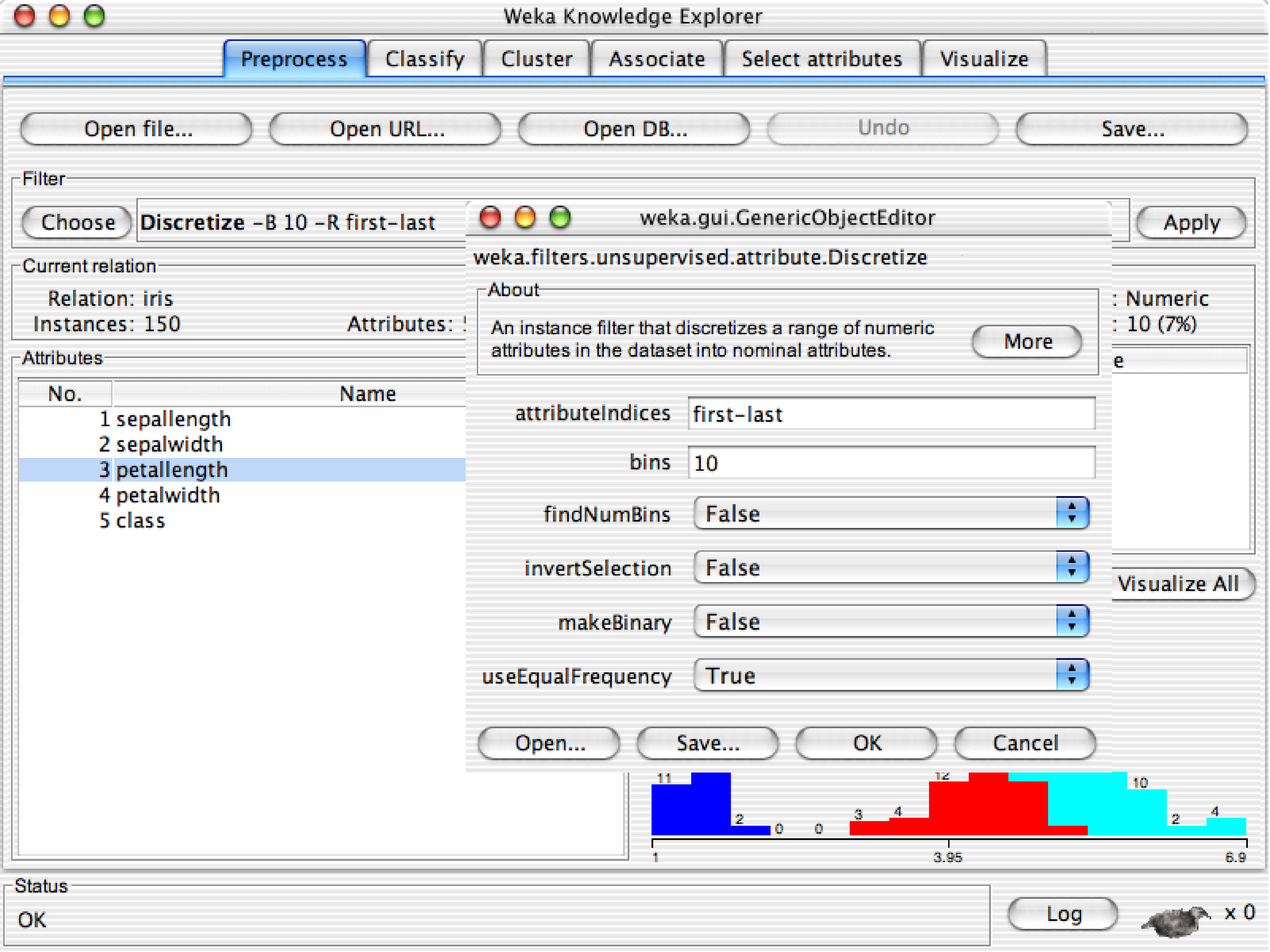
About

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

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

Status

OK



Discretize -B 10 -R first-last

weka.filters.unsupervised.attribute.Discretize

About

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

attributeIndices: first-last

bins: 10

findNumBins: False

invertSelection: False

makeBinary: False

useEqualFrequency: True

Current relation

Relation: iris

Instances: 150

Attributes: 5

: Numeric

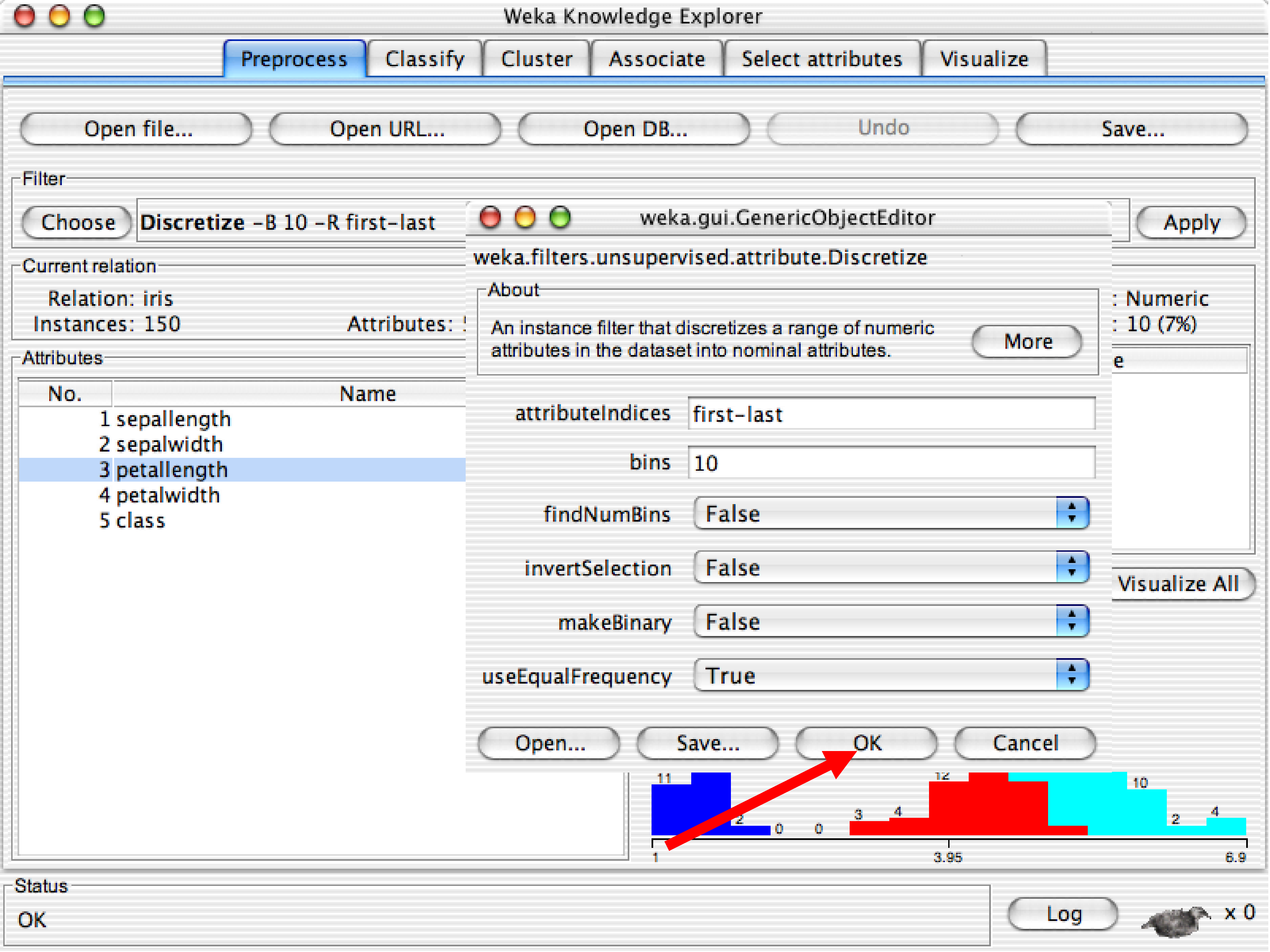
: 10 (7%)

Attributes

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

Status

OK



Discretize -B 10 -R first-last

Relation: iris
Instances: 150
Attributes: 5

: Numeric
: 10 (7%)

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

weka.gui.GenericObjectEditor

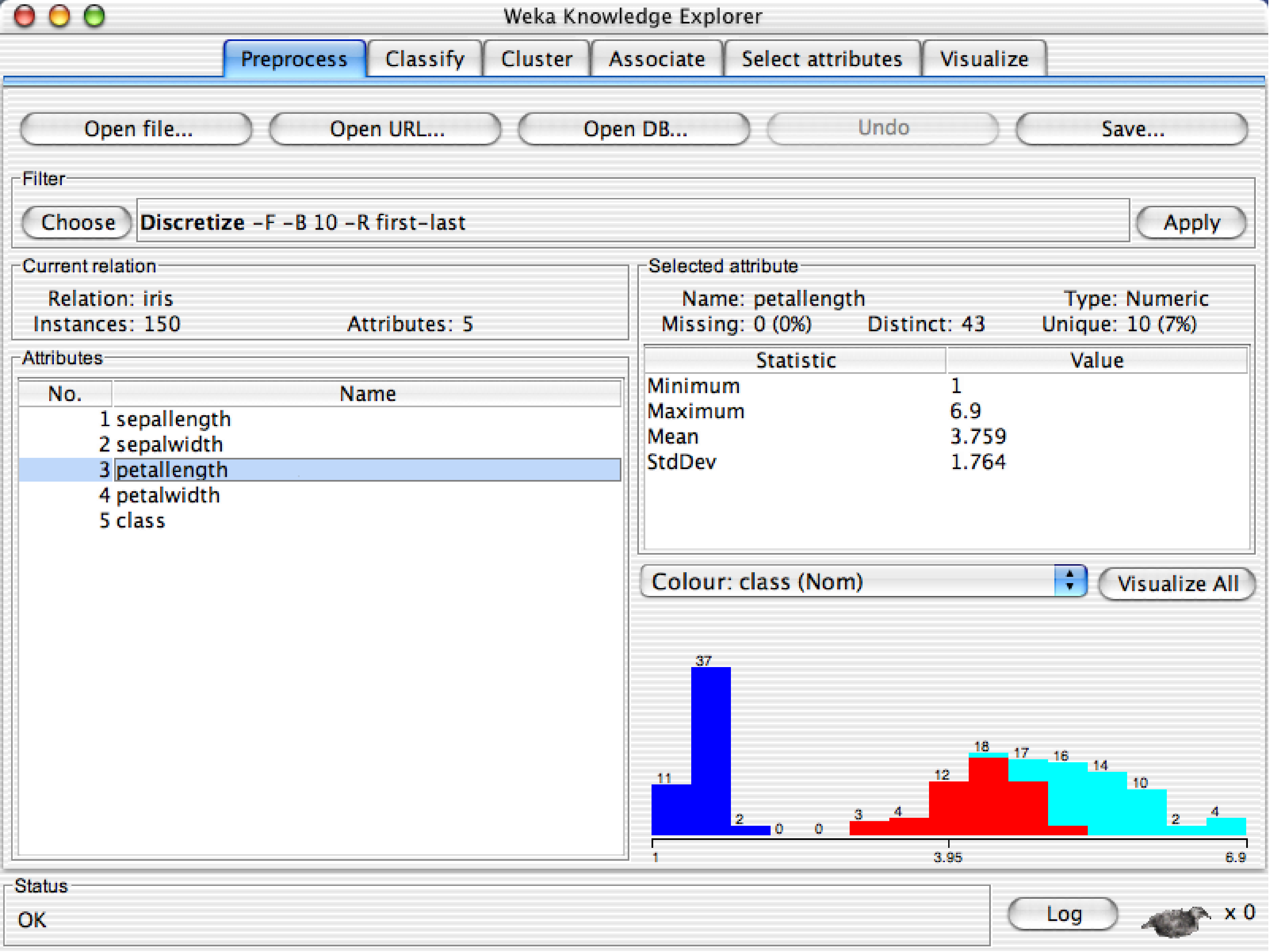
weka.filters.unsupervised.attribute.Discretize

About

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

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

OK



Weka Knowledge Explorer

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

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

Filter

Choose **Discretize** -F -B 10 -R first-last 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: petallength Type: Numeric
Missing: 0 (0%) Distinct: 43 Unique: 10 (7%)

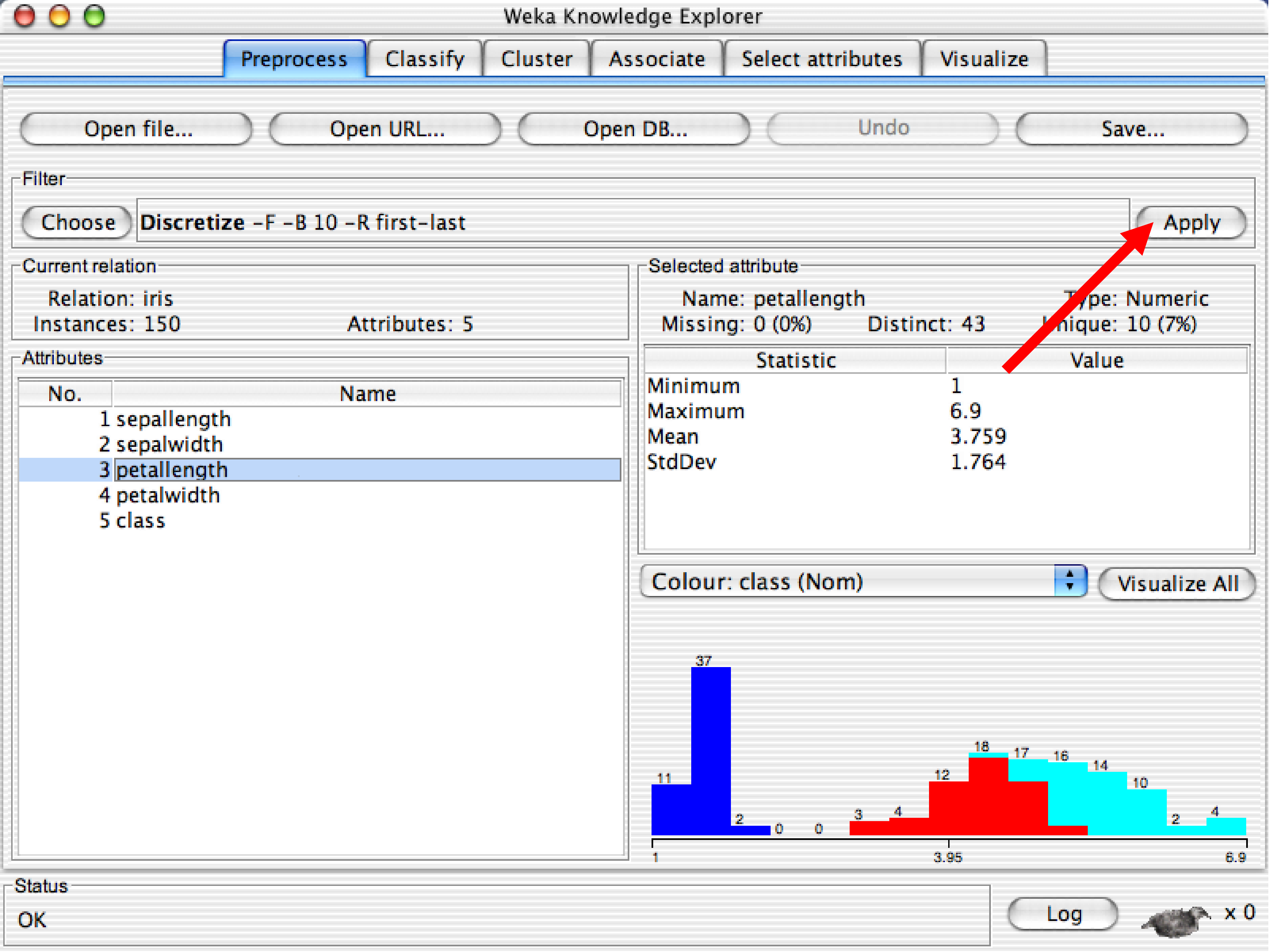
Statistic	Value
Minimum	1
Maximum	6.9
Mean	3.759
StdDev	1.764

Colour: class (Nom) Visualize All

Status

OK

Log x 0



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

Current relation

Relation: iris
Instances: 150 Attributes: 5

Attributes

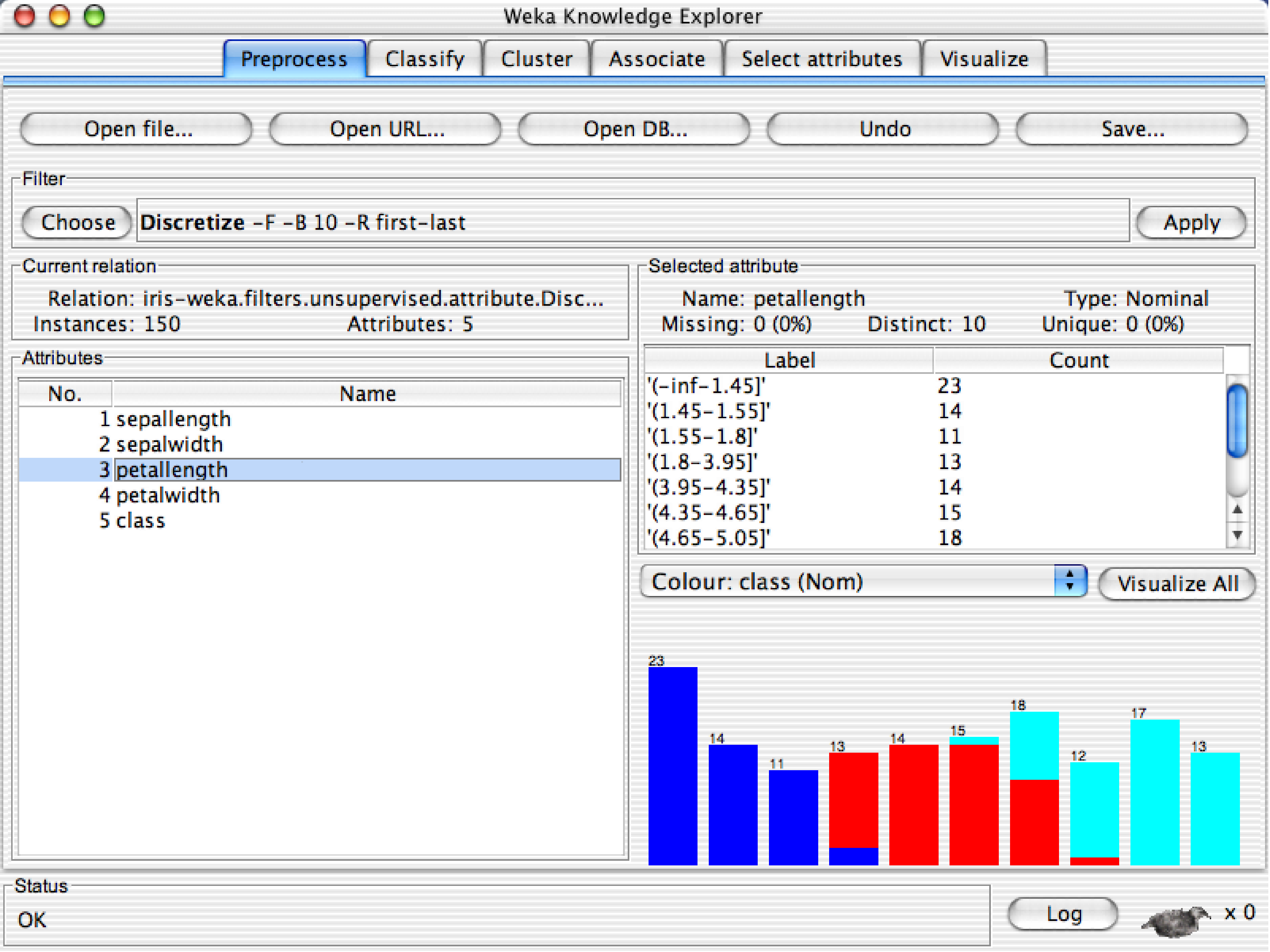
No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

Selected attribute

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

Statistic	Value
Minimum	1
Maximum	6.9
Mean	3.759
StdDev	1.764

Status
OK



Filter

Discretize -F -B 10 -R first-last

Current relation

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

Attributes

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

Selected attribute

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

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

Colour: class (Nom)

Status

OK

Preprocess: filters

- Choosing a filter
  - Supervised vs. Unsupervised
  - Attribute vs. Instance

- Supervised filters “require a class attribute” ; unsupervised filters do not

- “ meta-filters ” can filter results from clustering and classification steps

Preprocess: filter examples

	Instance	Attribute
UNsupervised	Resample	Discretize
Supervised	Resample	Discretize

Preprocess: instance filter example

- **Resampling**
  - Unsupervised
    - random % of the dataset
  - Supervised
    - takes the *class distribution* into account when generating a random sample
    - Can add bias towards a specific class value
    - Can specify maximum spread for rare/common class values

Preprocess: attribute filter example

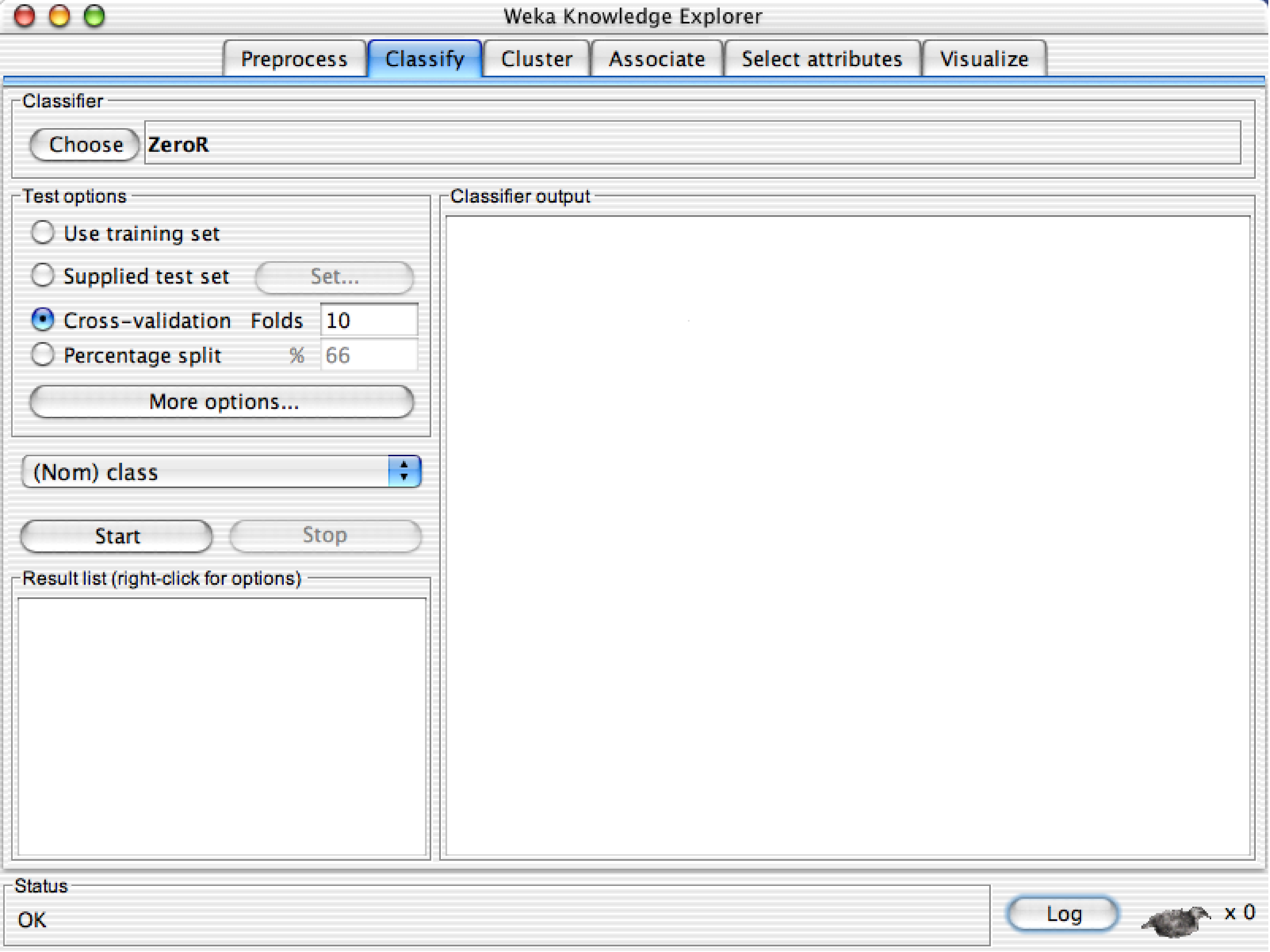
- **Discretize**
  - Unsupervised
    - K-Interval : simplest, can ensure small bin sizes
    - Proportional K-Interval : optimized for classification (Naïve Bayes)
  - Supervised
    - Entropy based
      - state of the art
      - computationally expensive
  - see *Chapter 7 of Data Mining by I. H. Witten and E. Frank*

Preprocess: attribute filter favorites

- Finding and Discarding variables
  - RemoveUseless : cut using variation threshold
- Datatype Transforms
  - NumericToNominal & NominalToBinary
  - StringToWord : NLP
- Value transforms
  - Normalize
  - ReplaceMissingValues : with mean value from training data
  - AddExpression : any math expression (think R)

WEKA Explorer Tutorial Examples

- Preprocess
  - Instance and Attribute Filters (Supervised and Unsupervised)
- **Classify**
  - ZeroR
  - Bayes
- Cluster
  - Expectation Maximization
  - Hierarchical Clustering
- Associate
  - Apriori
- Select Attributes
  - Via clustering



Weka Knowledge Explorer

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

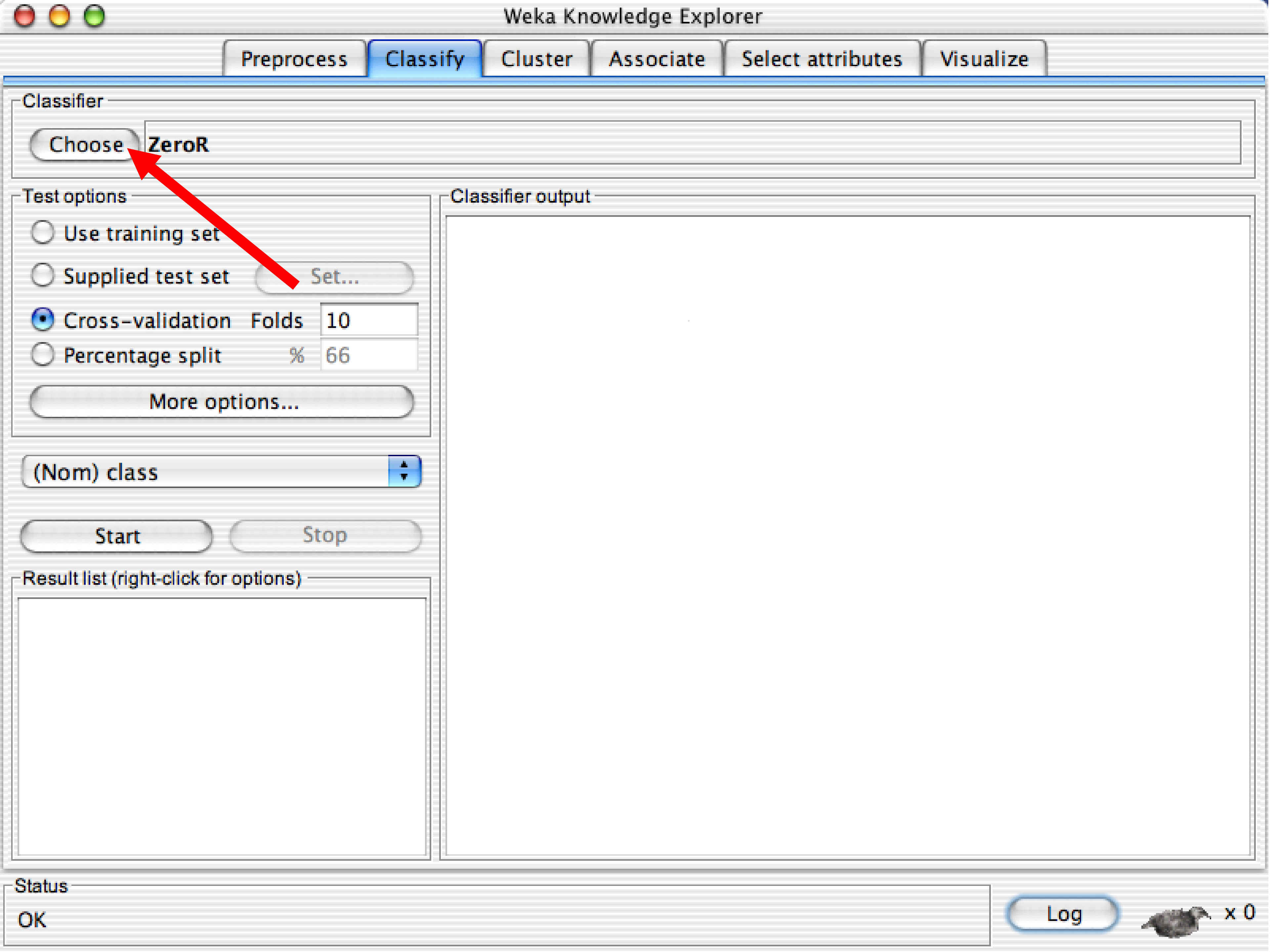
Result list (right-click for options)

Classifier output

Status

OK

Log x 0



Weka Knowledge Explorer

Classifier

ZeroR

Test options

○ Use training set

○ Supplied test set

● Cross-validation Folds 10

○ Percentage split % 66

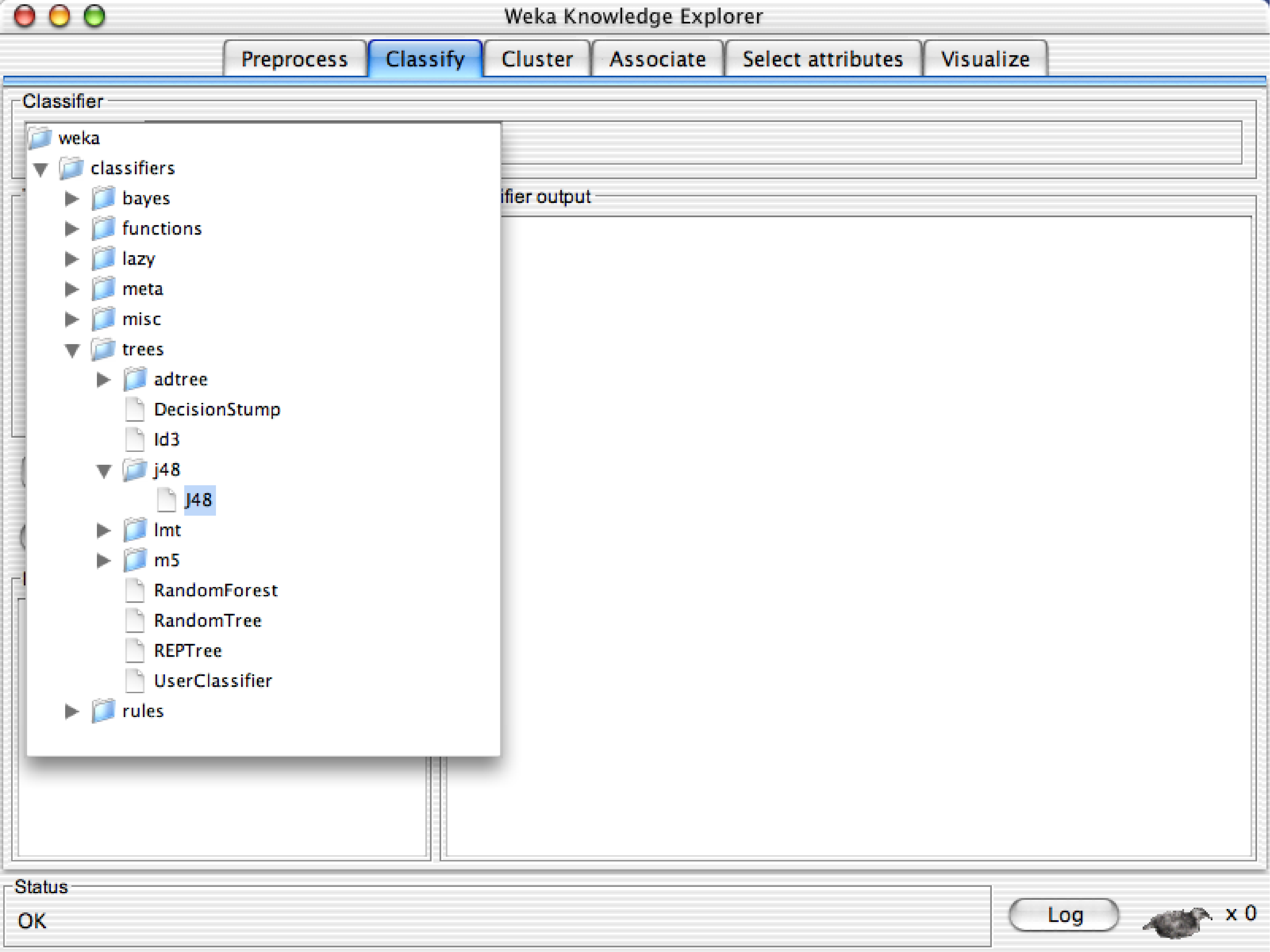
(Nom) class

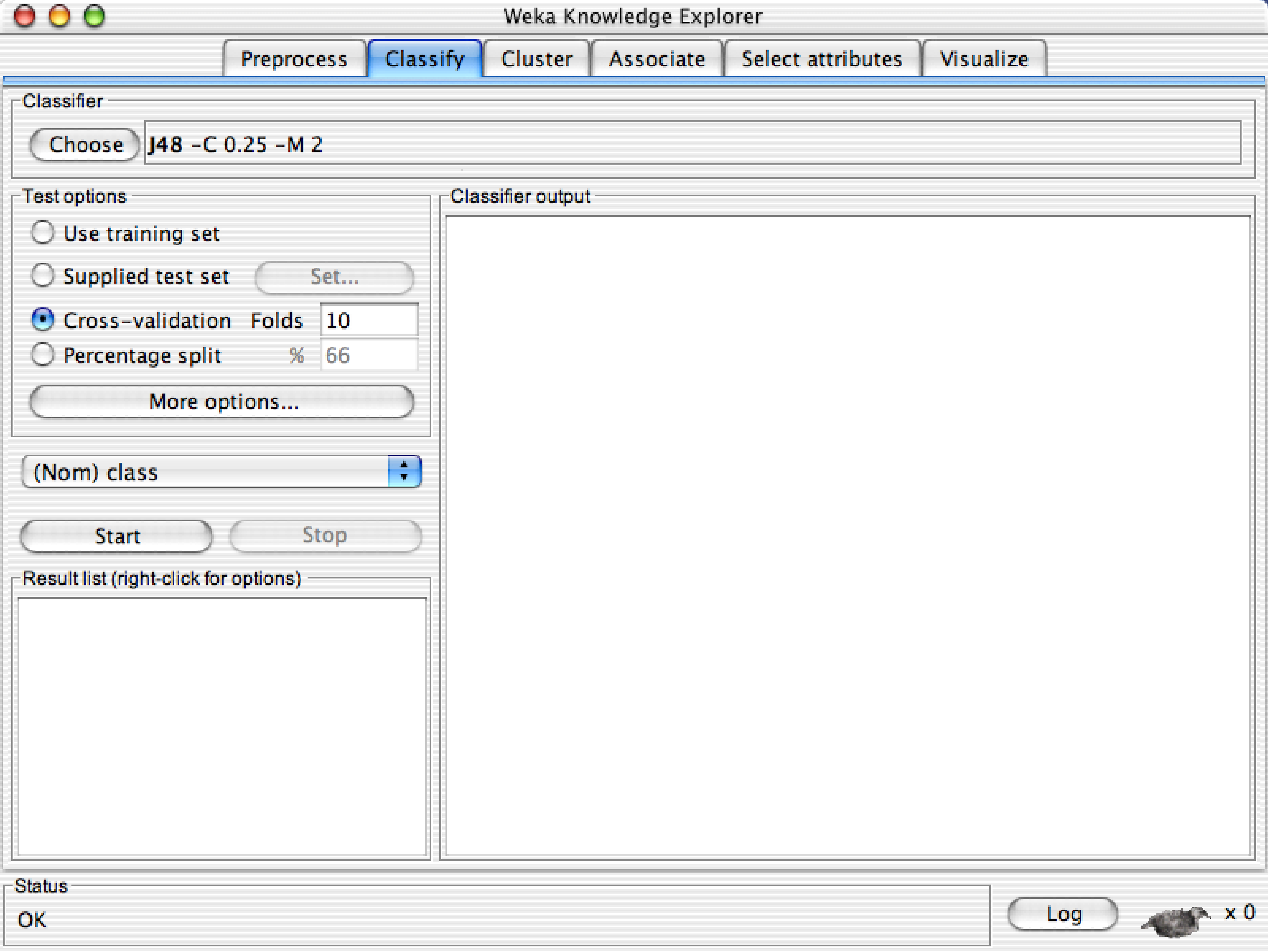
Result list (right-click for options)

Classifier output

Status

OK


Weka Knowledge Explorer
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
ifier output
Status
OK
Log
x 0



Classifier

J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set

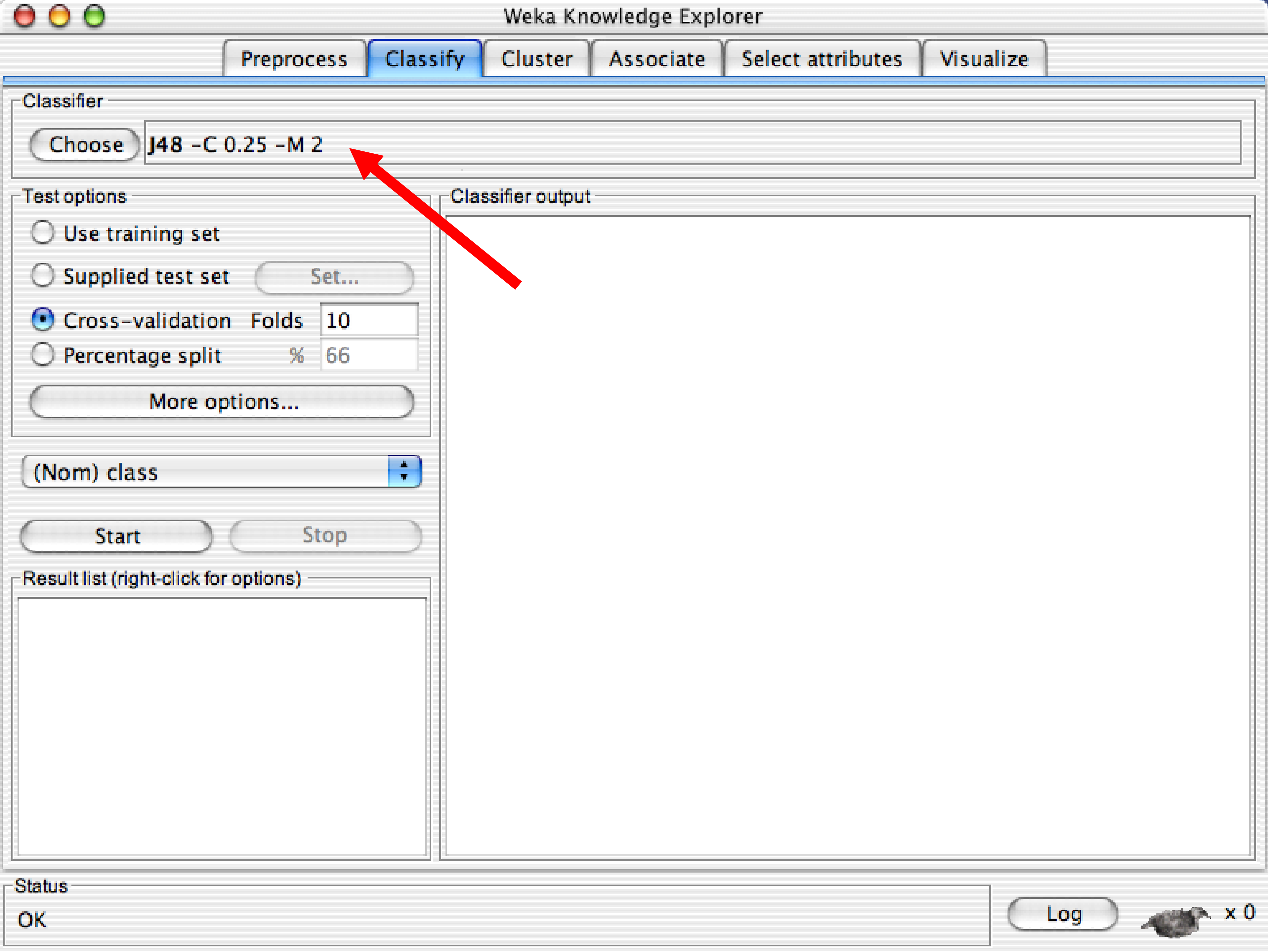
● Cross-validation Folds 10

○ Percentage split % 66

(Nom) class

Status

OK



Classifier

J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set

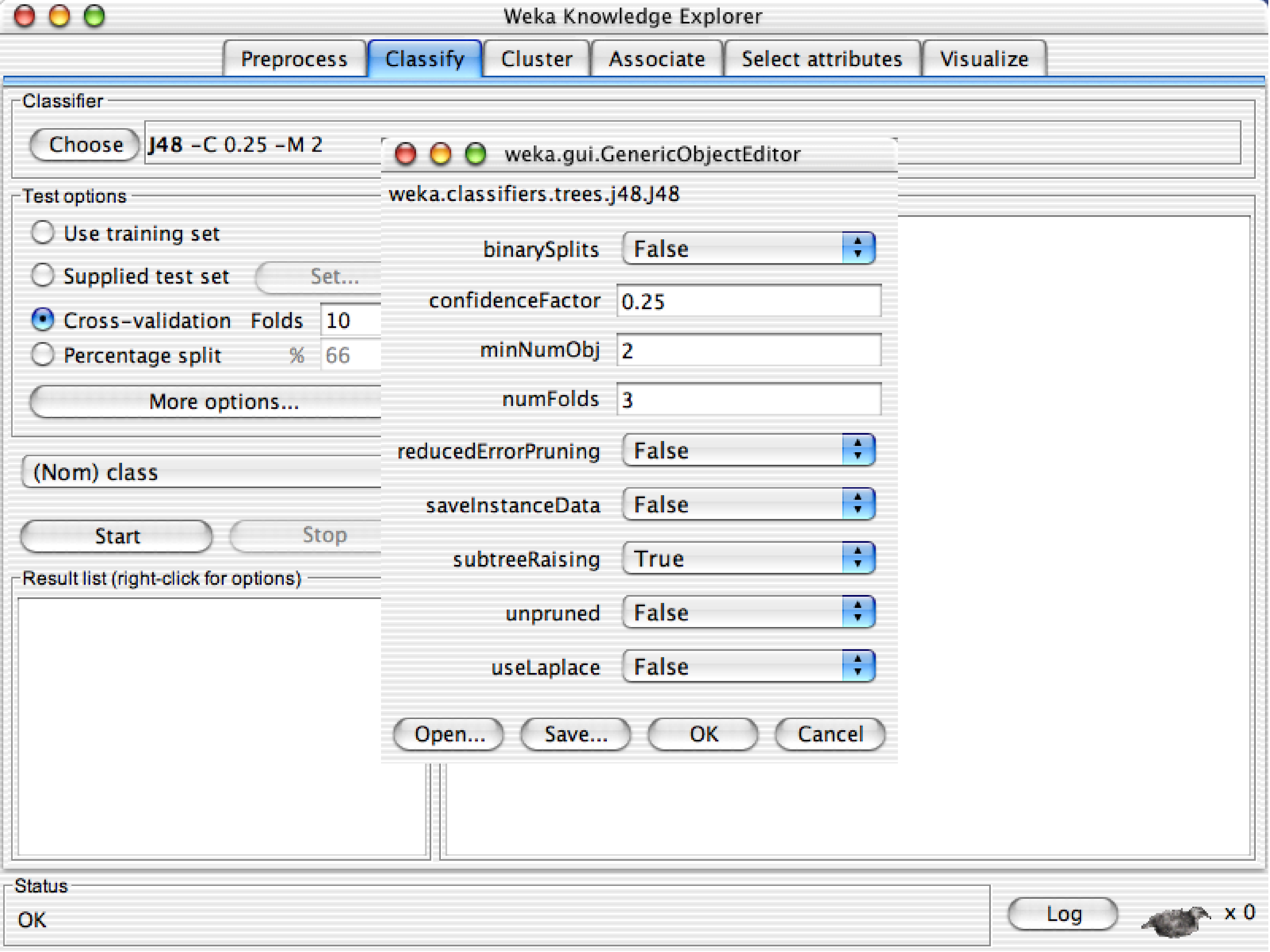
● Cross-validation Folds 10

○ Percentage split % 66

(Nom) class

Status

OK



J48 -C 0.25 -M 2

○ Use training set

○ Supplied test set

● Cross-validation Folds 10

○ Percentage split % 66

(Nom) class

weka.classifiers.trees.j48.J48

binarySplits False

confidenceFactor 0.25

minNumObj 2

numFolds 3

reducedErrorPruning False

saveInstanceData False

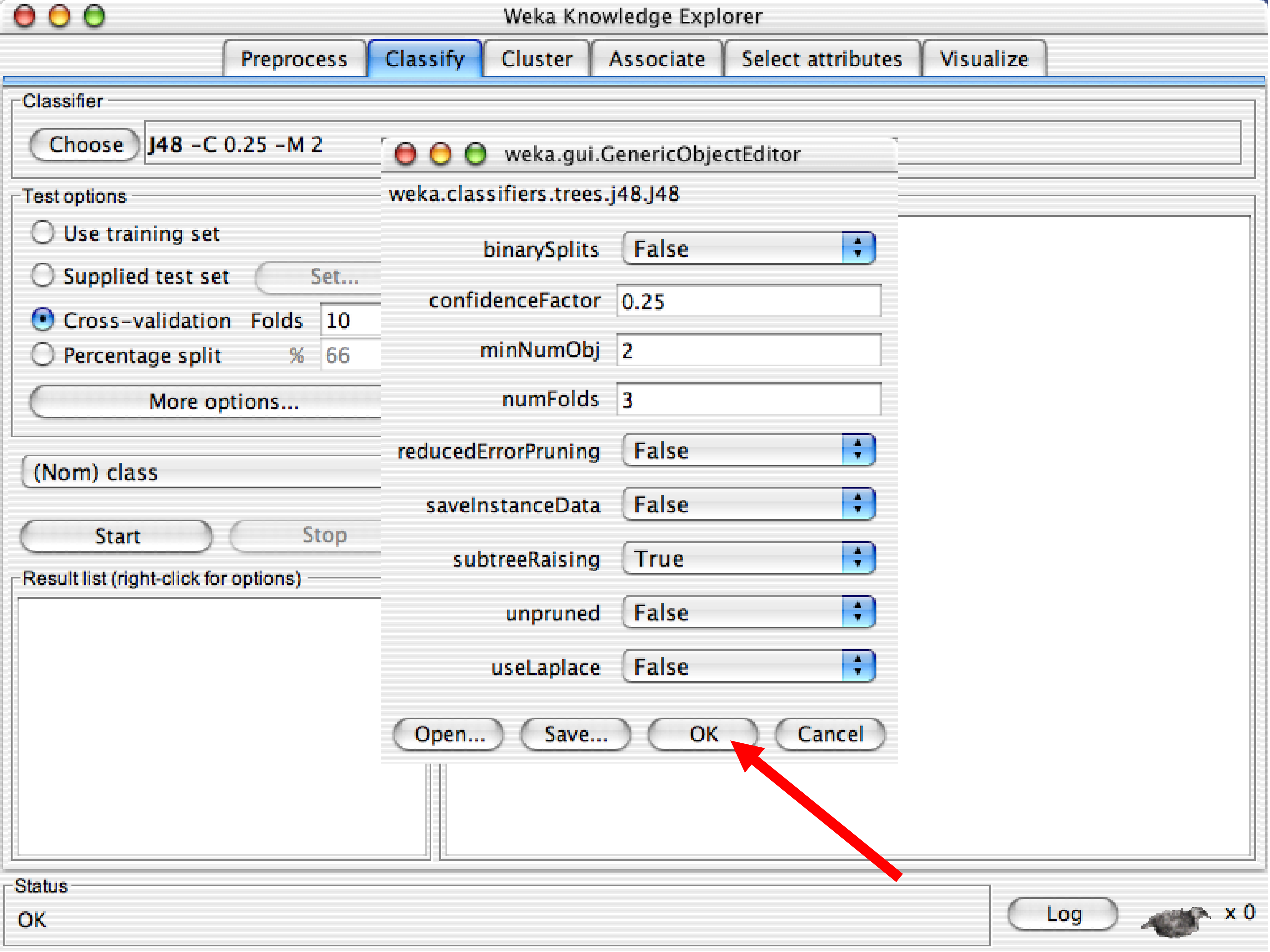
subtreeRaising True

unpruned False

useLaplace False

Status

OK


Weka Knowledge Explorer

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

Classifier

Choose J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set Set...

● Cross-validation Folds 10

○ Percentage split % 66

More options...

(Nom) class

Start Stop

Result list (right-click for options)

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

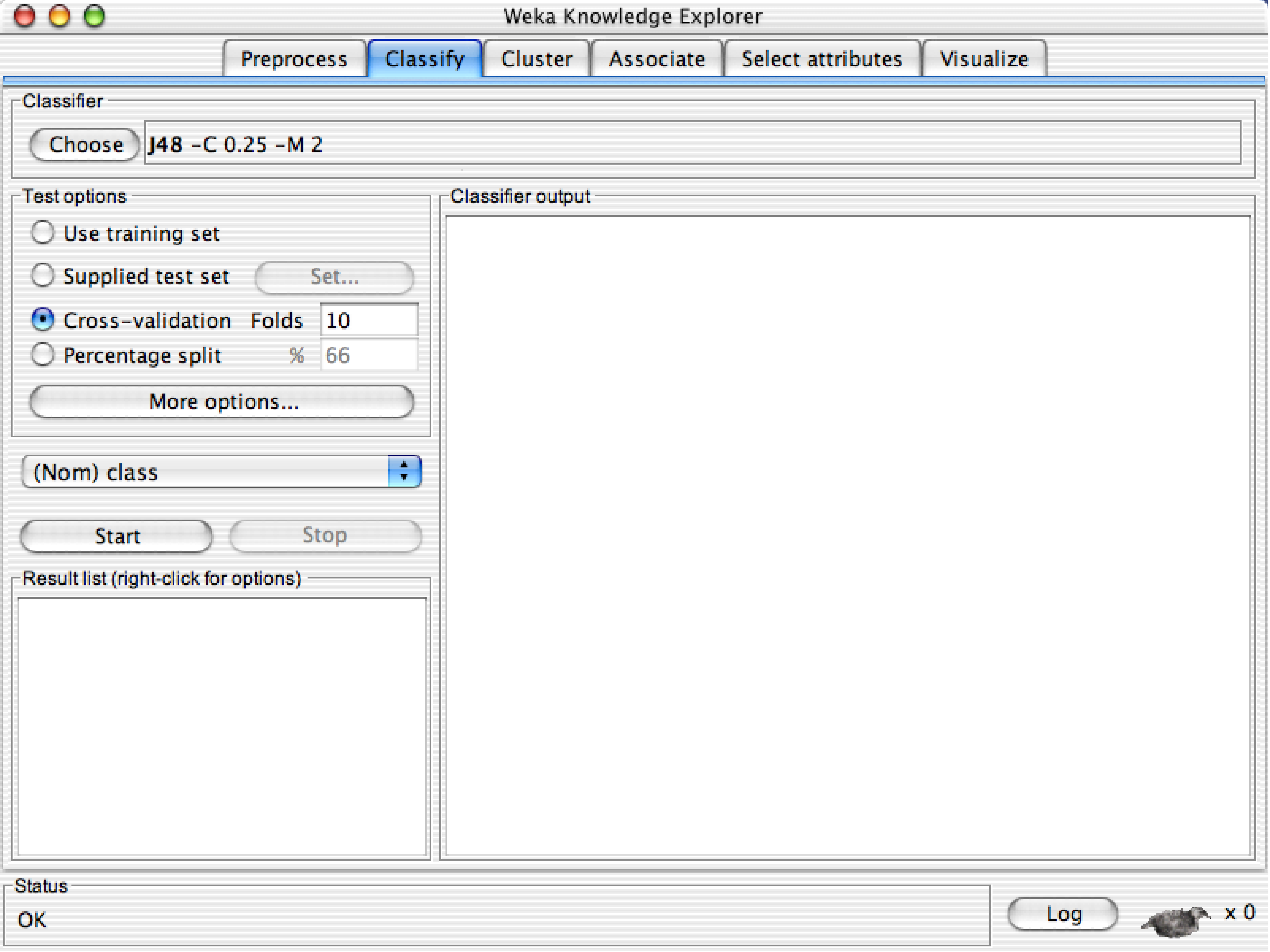
useLaplace: False

Open... | Save... | OK | Cancel

Status

OK

Log x 0



J48 -C 0.25 -M 2

○ Use training set

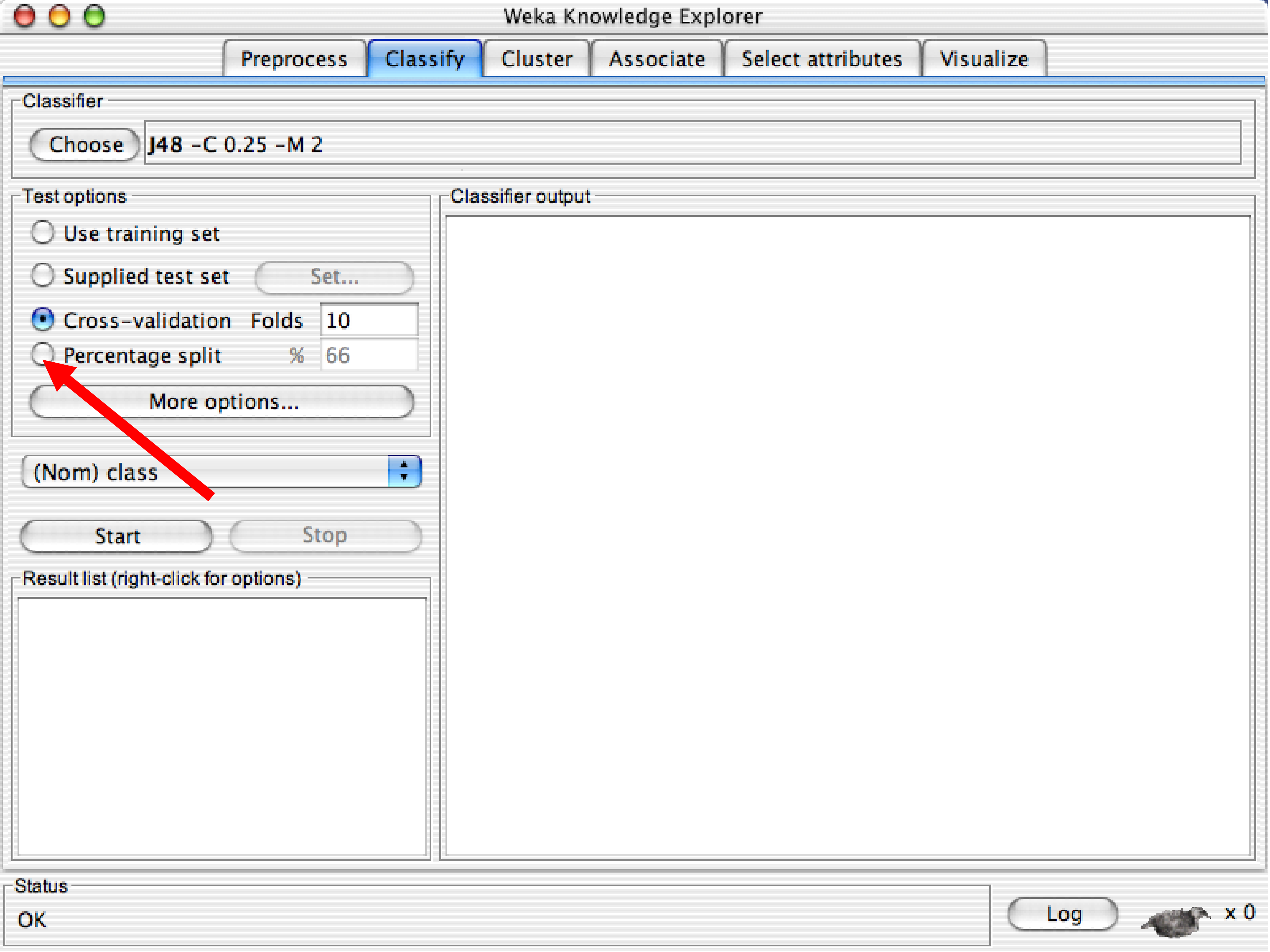
○ Supplied test set

● Cross-validation Folds 10

○ Percentage split % 66

(Nom) class

OK



Classifier

Choose J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set

● Cross-validation Folds 10

○ Percentage split % 66

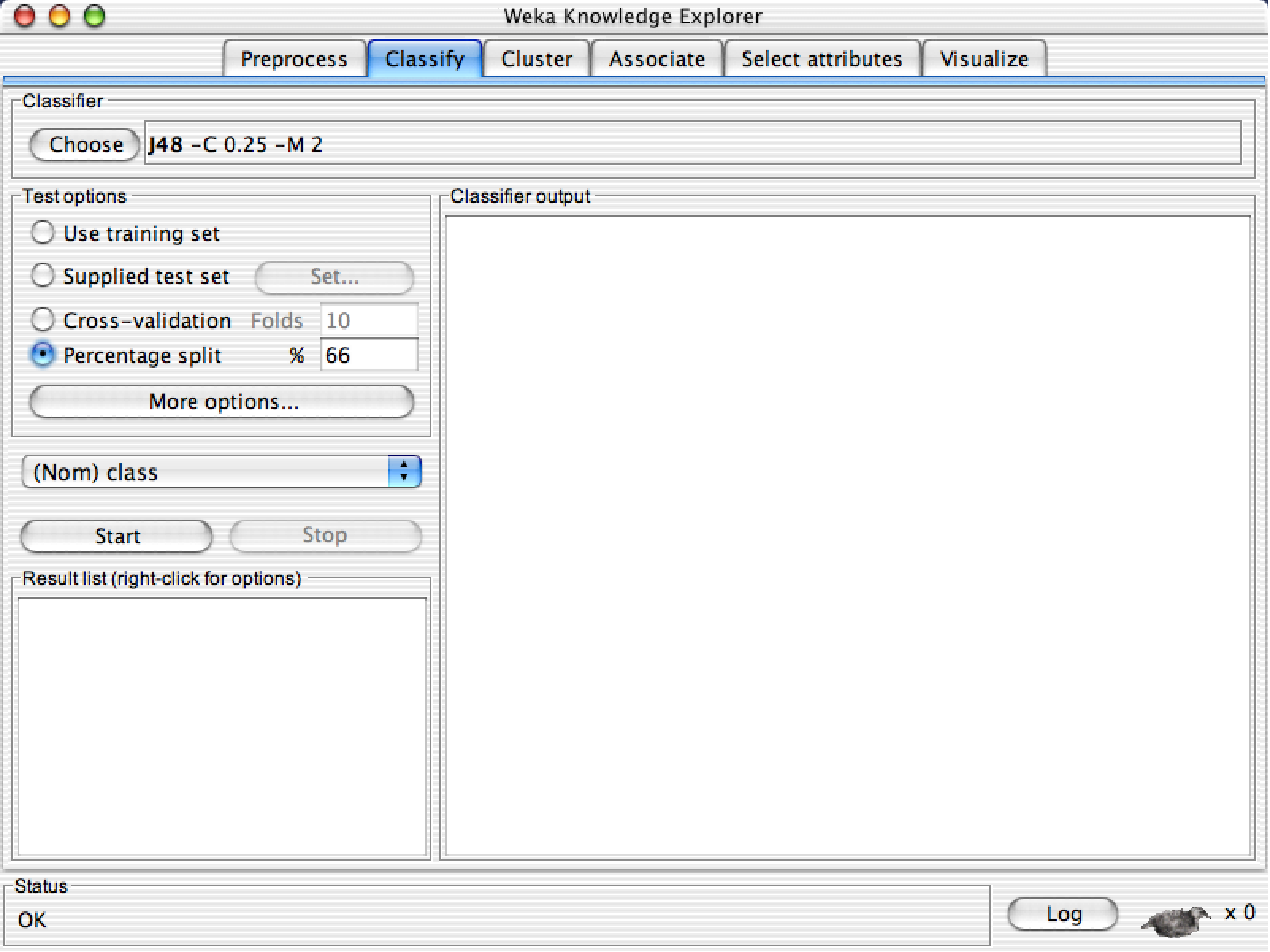
(Nom) class

Result list (right-click for options)

Classifier output

Status

OK



J48 -C 0.25 -M 2

○ Use training set

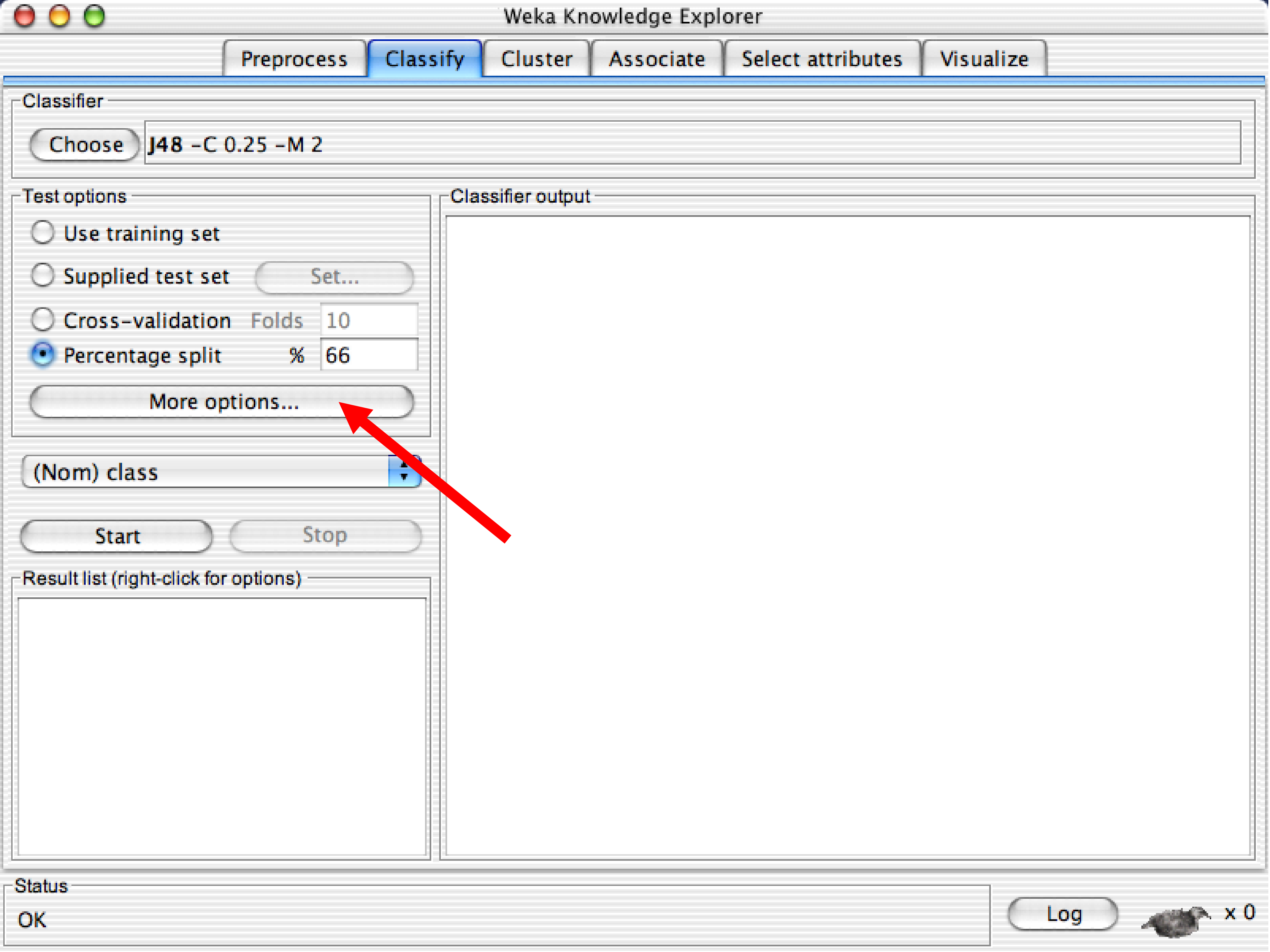
○ Supplied test set

○ Cross-validation Folds 10

● Percentage split % 66

(Nom) class

OK

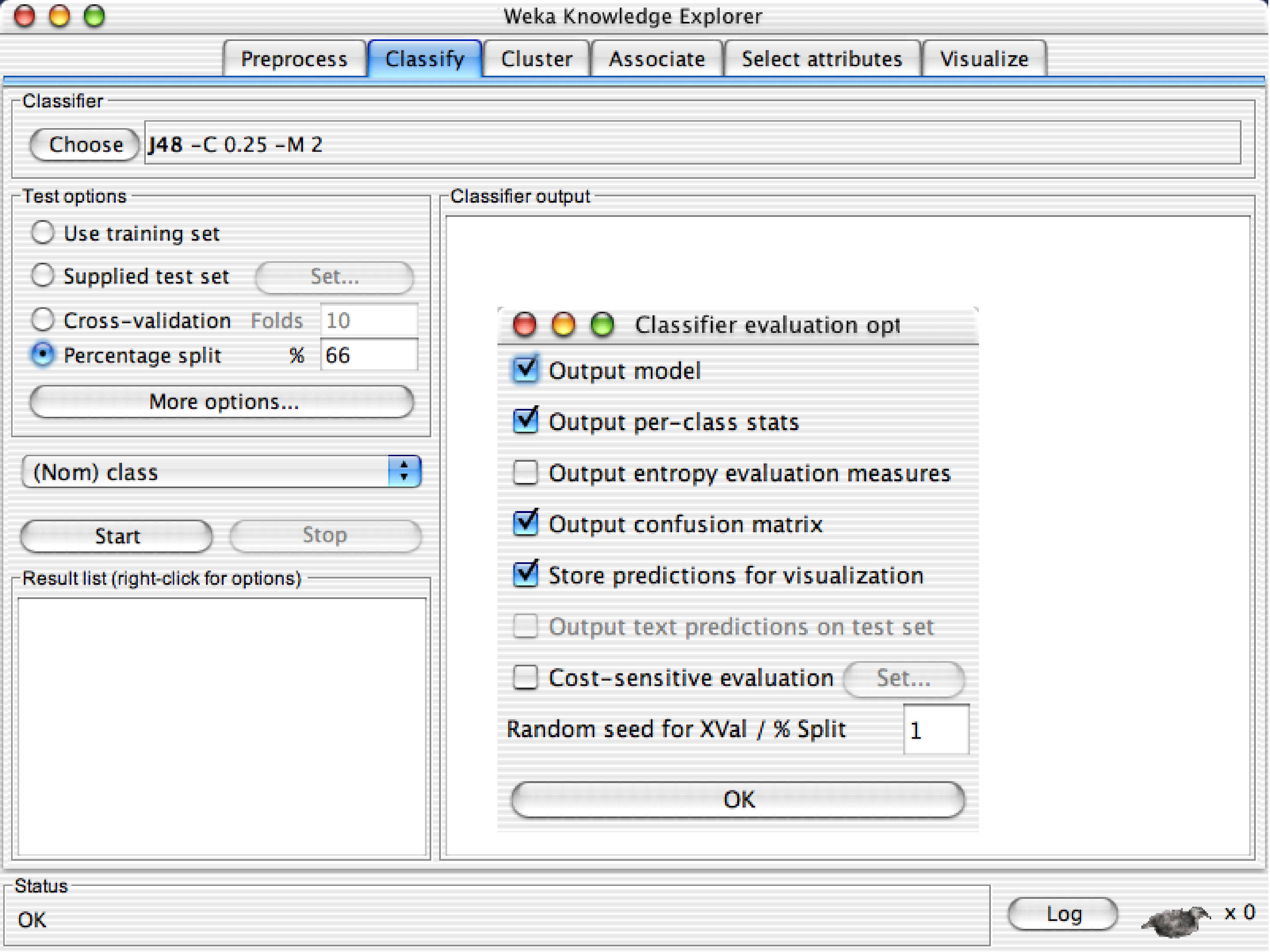


Classifier: J48 -C 0.25 -M 2

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

(Nom) class

Status
OK



Classifier

J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set

○ Cross-validation Folds 10

● Percentage split % 66

(Nom) class

Result list (right-click for options)

Classifier output

Classifier evaluation opt

☑ Output model

☑ Output per-class stats

□ Output entropy evaluation measures

☑ Output confusion matrix

☑ Store predictions for visualization

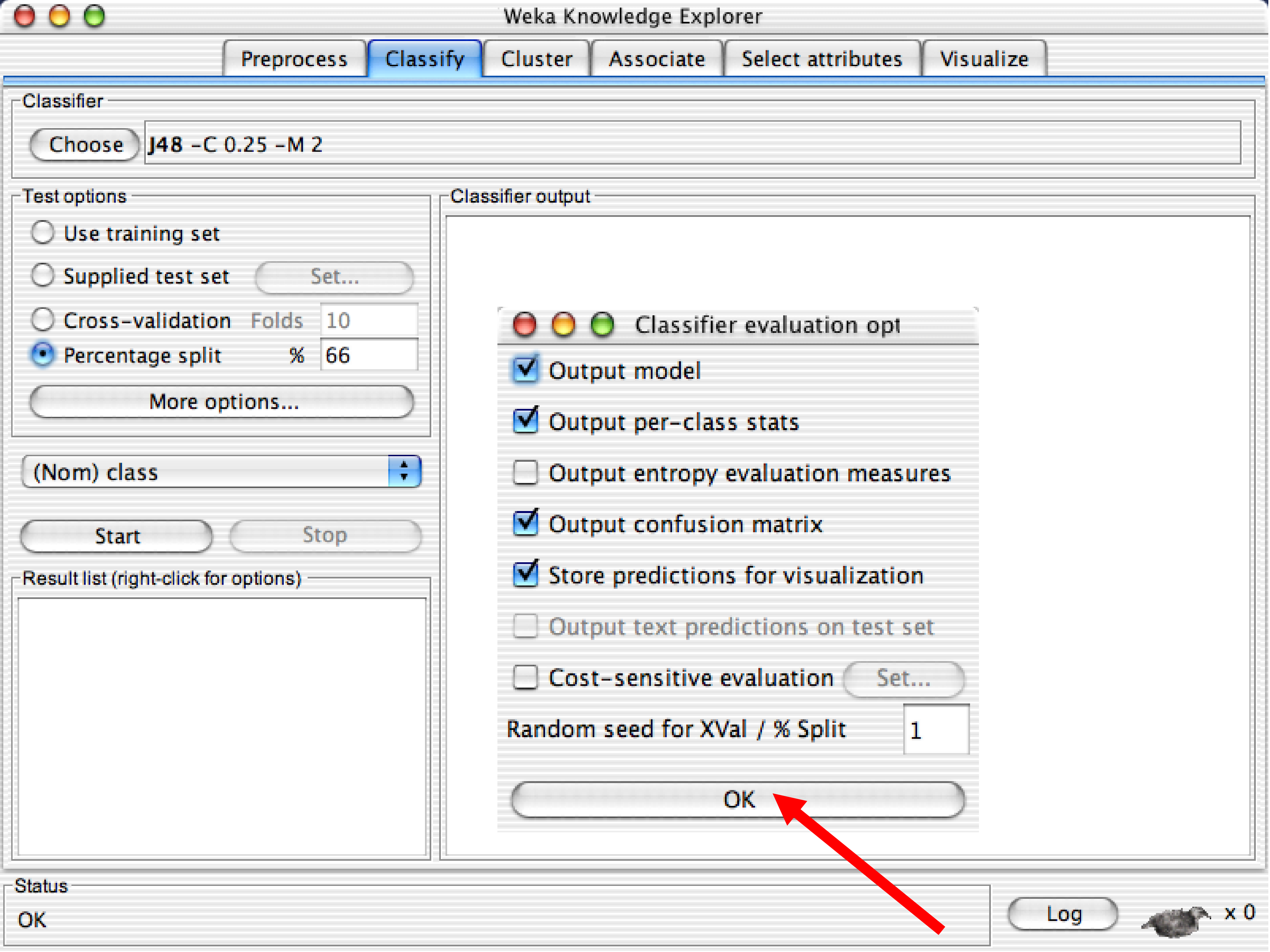
□ Output text predictions on test set

□ Cost-sensitive evaluation

Random seed for XVal / % Split 1

Status

OK



Classifier

J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set

○ Cross-validation Folds 10

● Percentage split % 66

(Nom) class

Result list (right-click for options)

Classifier output

Classifier evaluation opt

☑ Output model

☑ Output per-class stats

□ Output entropy evaluation measures

☑ Output confusion matrix

☑ Store predictions for visualization

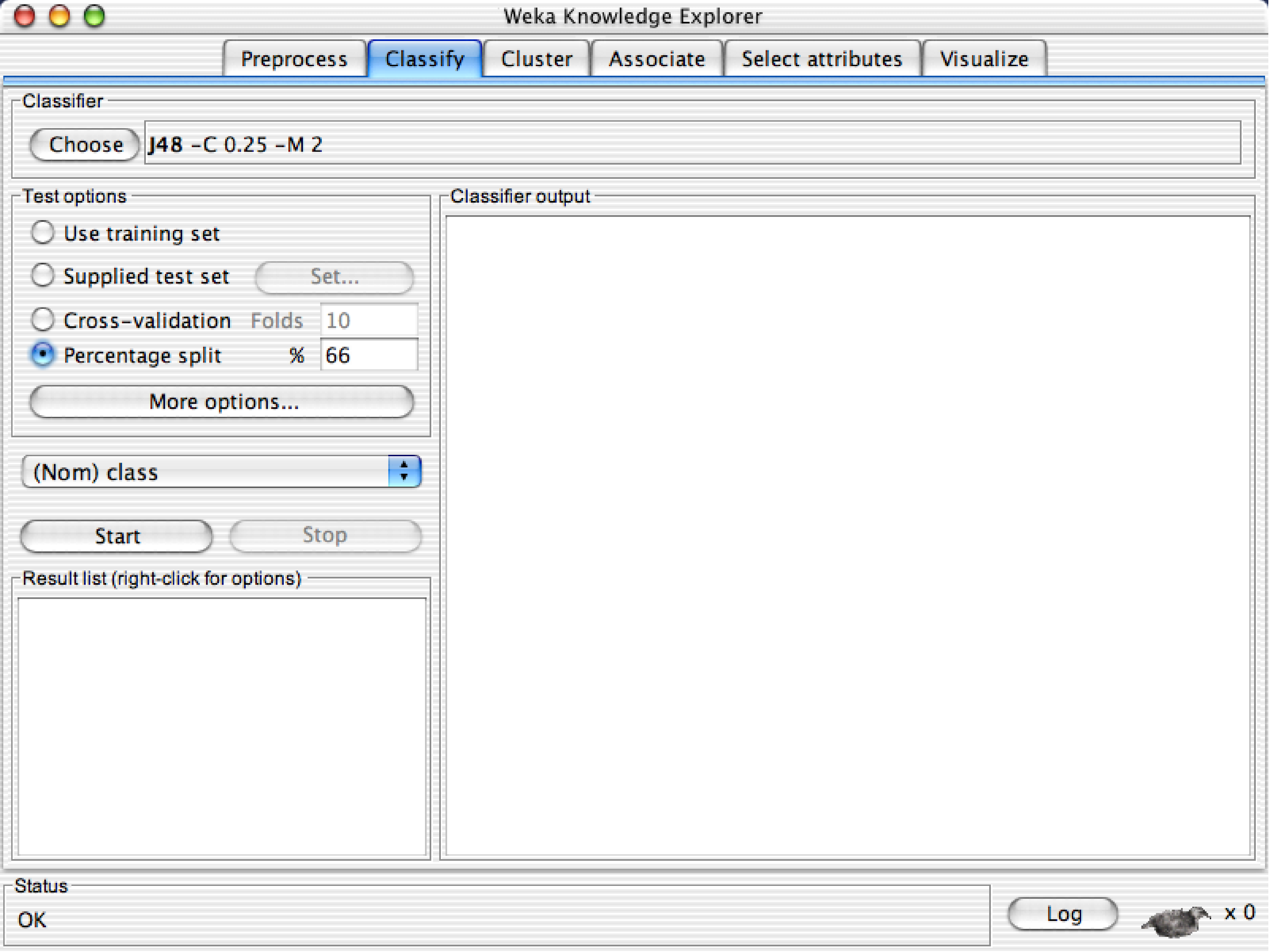
□ Output text predictions on test set

□ Cost-sensitive evaluation

Random seed for XVal / % Split 1

Status

OK



Classifier

J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set

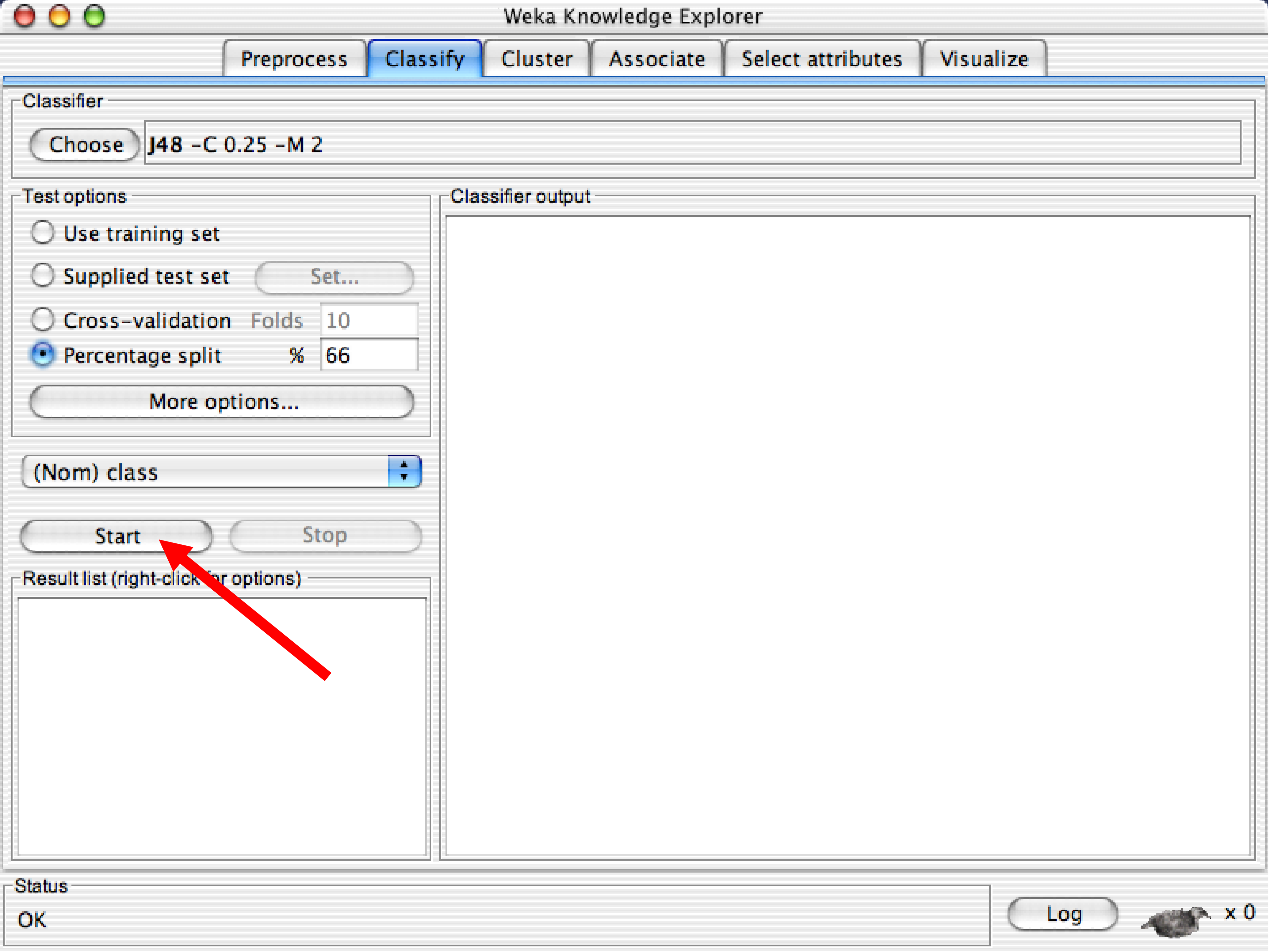
○ Cross-validation Folds 10

● Percentage split % 66

(Nom) class

Status

OK



Classifier

J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set

○ Cross-validation Folds 10

● Percentage split % 66

(Nom) class

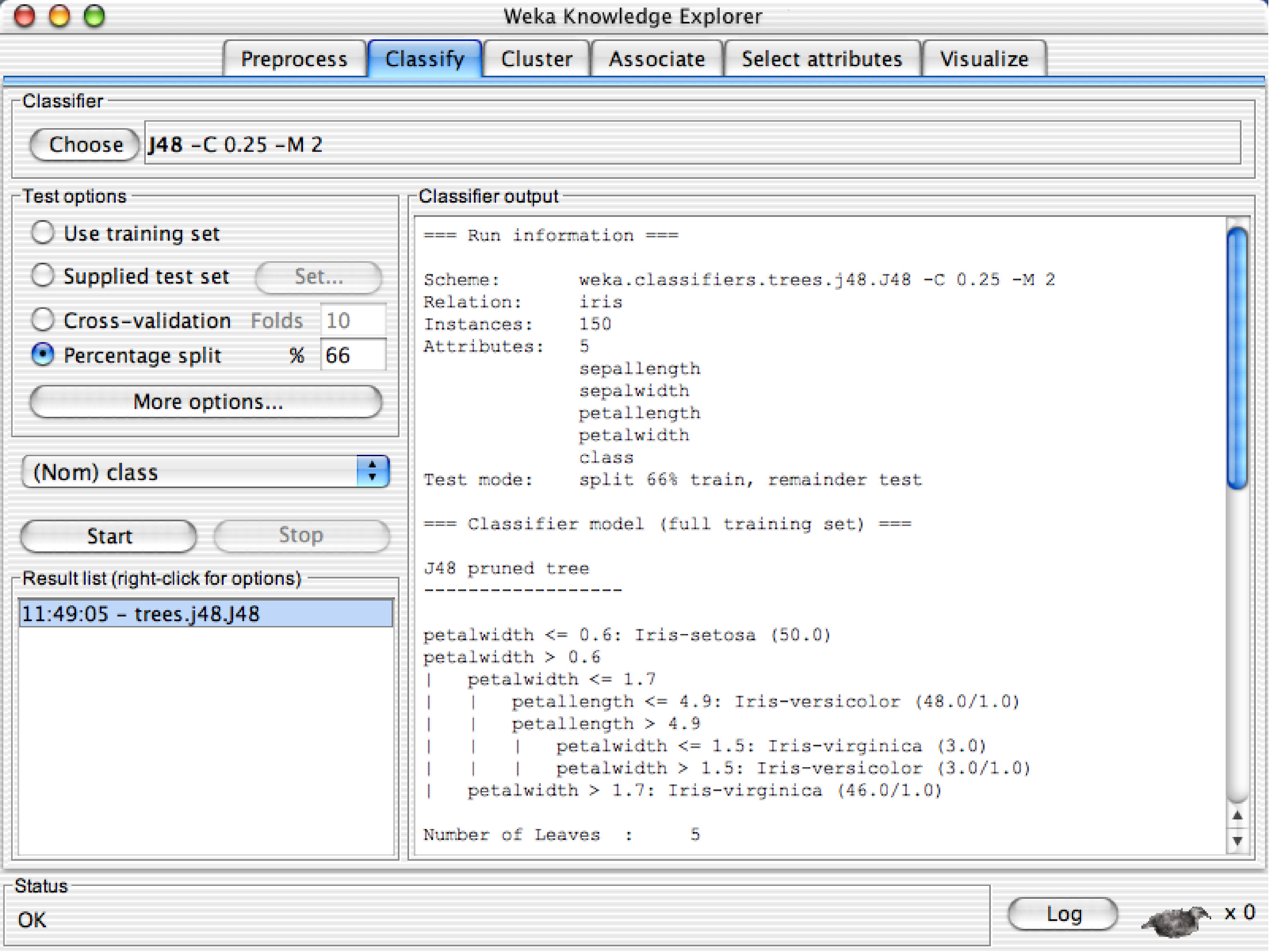
Result list (right-click for options)

Classifier output

Status

OK

x 0



Classifier

J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set

○ Cross-validation Folds 10

● Percentage split % 66

(Nom) class

Result list (right-click for options)

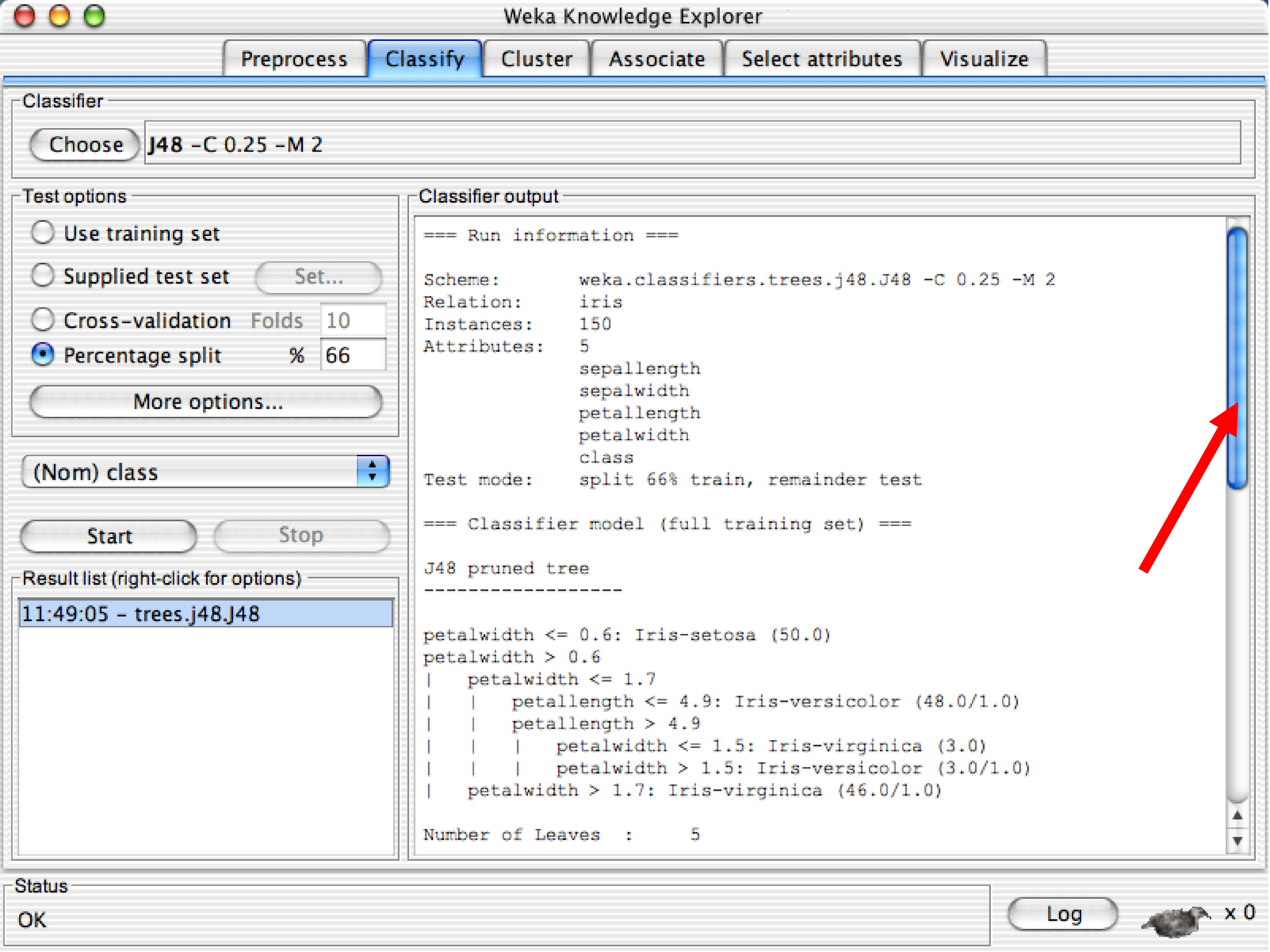
11:49:05 - trees.j48.J48

Status

OK

Classifier output

```
=== Run information ===

Scheme:       weka.classifiers.trees.j48.J48 -C 0.25 -M 2
Relation:     iris
Instances:    150
Attributes:   5
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
Test mode:    split 66% train, remainder test

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

J48 pruned tree
------------------

petalwidth <= 0.6: Iris-setosa (50.0)
petalwidth > 0.6
|   petalwidth <= 1.7
|   |   petallength <= 4.9: Iris-versicolor (48.0/1.0)
|   |   petallength > 4.9
|   |   |   petalwidth <= 1.5: Iris-virginica (3.0)
|   |   |   petalwidth > 1.5: Iris-versicolor (3.0/1.0)
|   petalwidth > 1.7: Iris-virginica (46.0/1.0)

Number of Leaves  :     5
```



Classifier

J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set

○ Cross-validation Folds 10

● Percentage split % 66

(Nom) class

Result list (right-click for options)

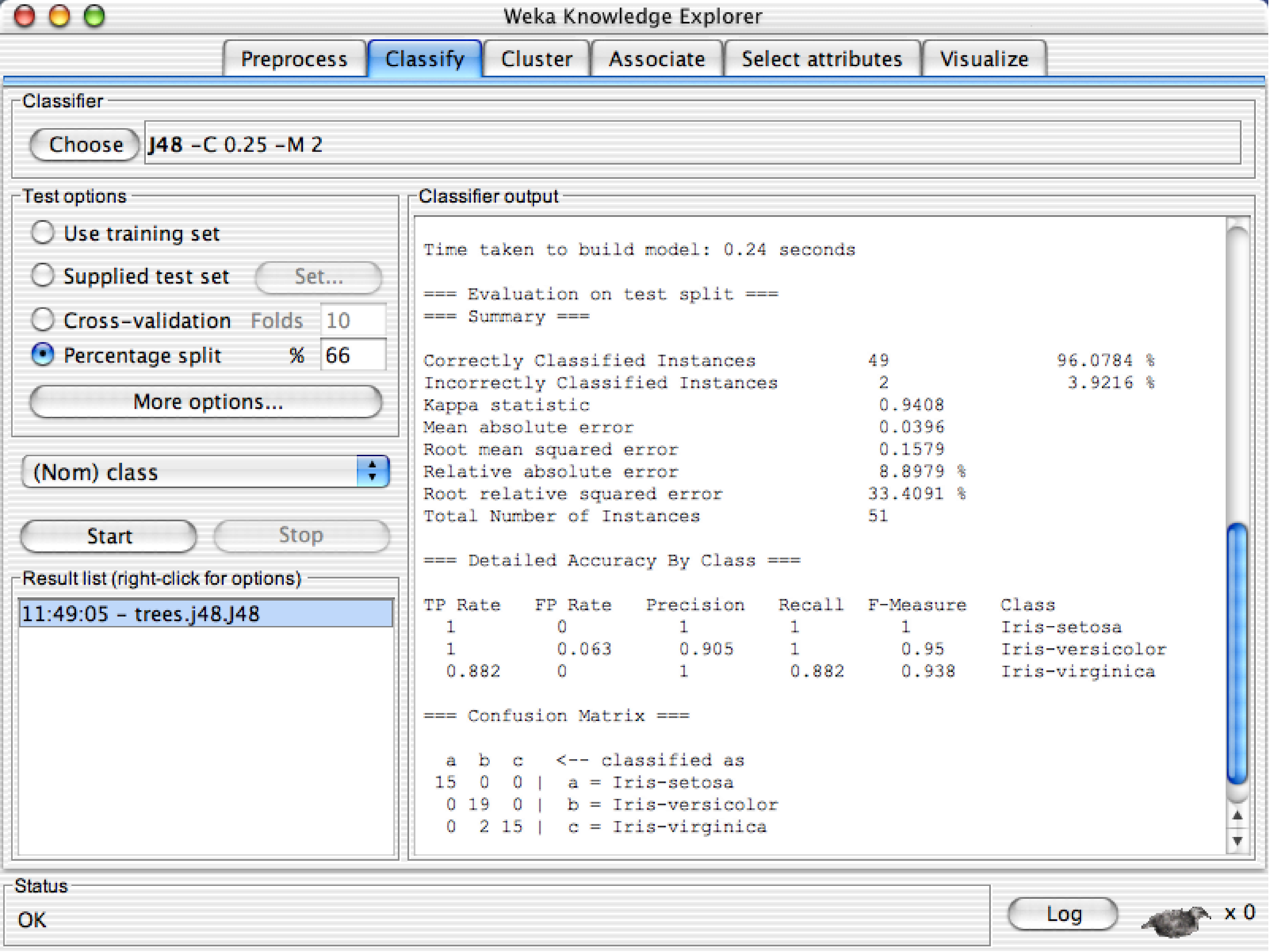
11:49:05 - trees.j48.J48

Classifier output

```
=== Run information ===

Scheme:       weka.classifiers.trees.j48.J48 -C 0.25 -M 2
Relation:     iris
Instances:    150
Attributes:   5
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
Test mode:    split 66% train, remainder test

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

J48 pruned tree
------------------

petalwidth <= 0.6: Iris-setosa (50.0)
petalwidth > 0.6
|   petalwidth <= 1.7
|   |   petallength <= 4.9: Iris-versicolor (48.0/1.0)
|   |   petallength > 4.9
|   |   |   petalwidth <= 1.5: Iris-virginica (3.0)
|   |   |   petalwidth > 1.5: Iris-versicolor (3.0/1.0)
|   petalwidth > 1.7: Iris-virginica (46.0/1.0)

Number of Leaves  :     5
```

Status

OK

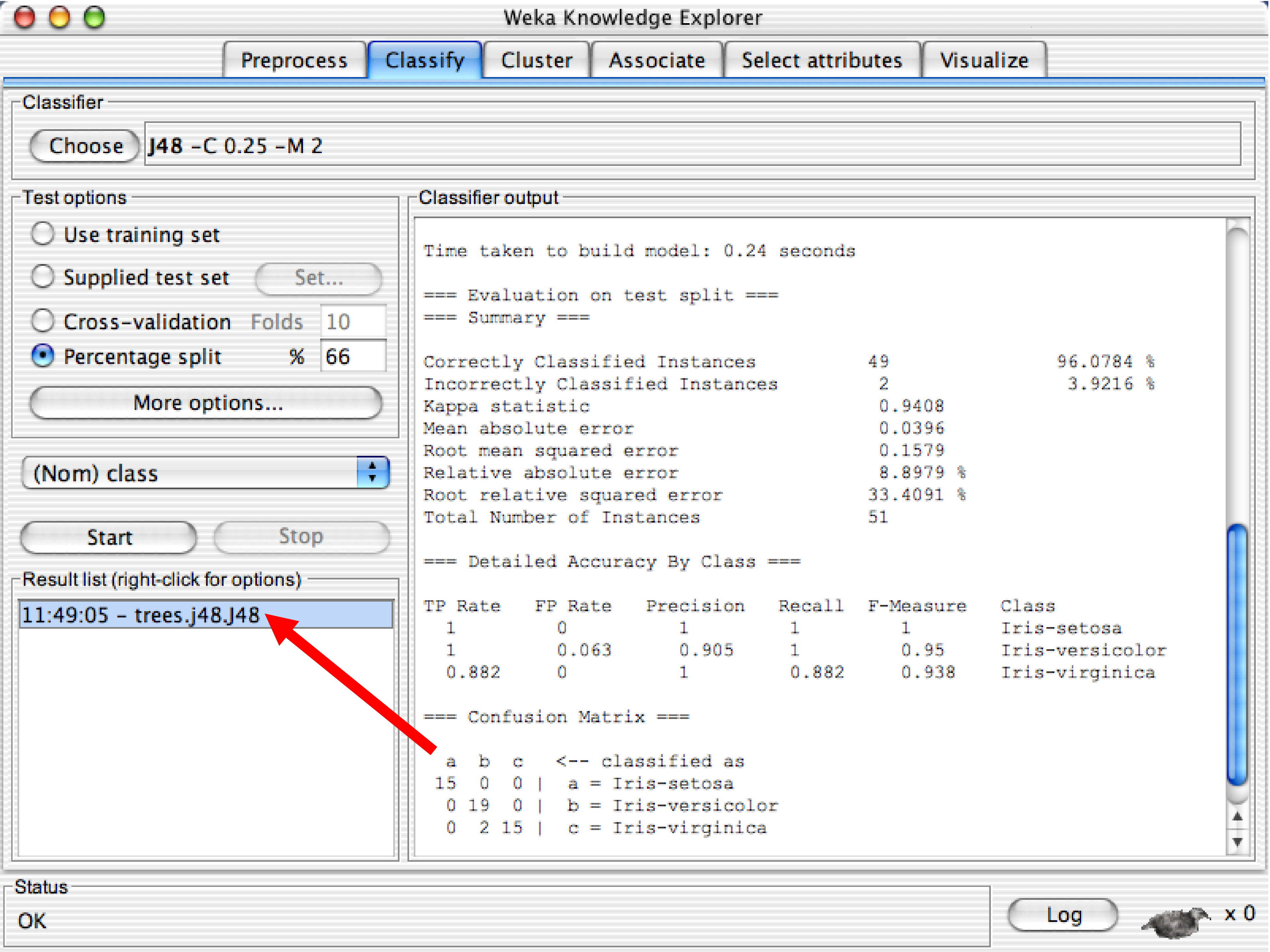


J48 -C 0.25 -M 2

11:49:05 - trees.j48.J48

```
Time taken to build model: 0.24 seconds

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

Correctly Classified Instances          49               96.0784 %
Incorrectly Classified Instances         2                3.9216 %
Kappa statistic                          0.9408
Mean absolute error                      0.0396
Root mean squared error                  0.1579
Relative absolute error                  8.8979 %
Root relative squared error             33.4091 %
Total Number of Instances               51

=== Detailed Accuracy By Class ===

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

=== Confusion Matrix ===

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

OK



J48 -C 0.25 -M 2

○ Use training set

○ Supplied test set

○ Cross-validation Folds 10

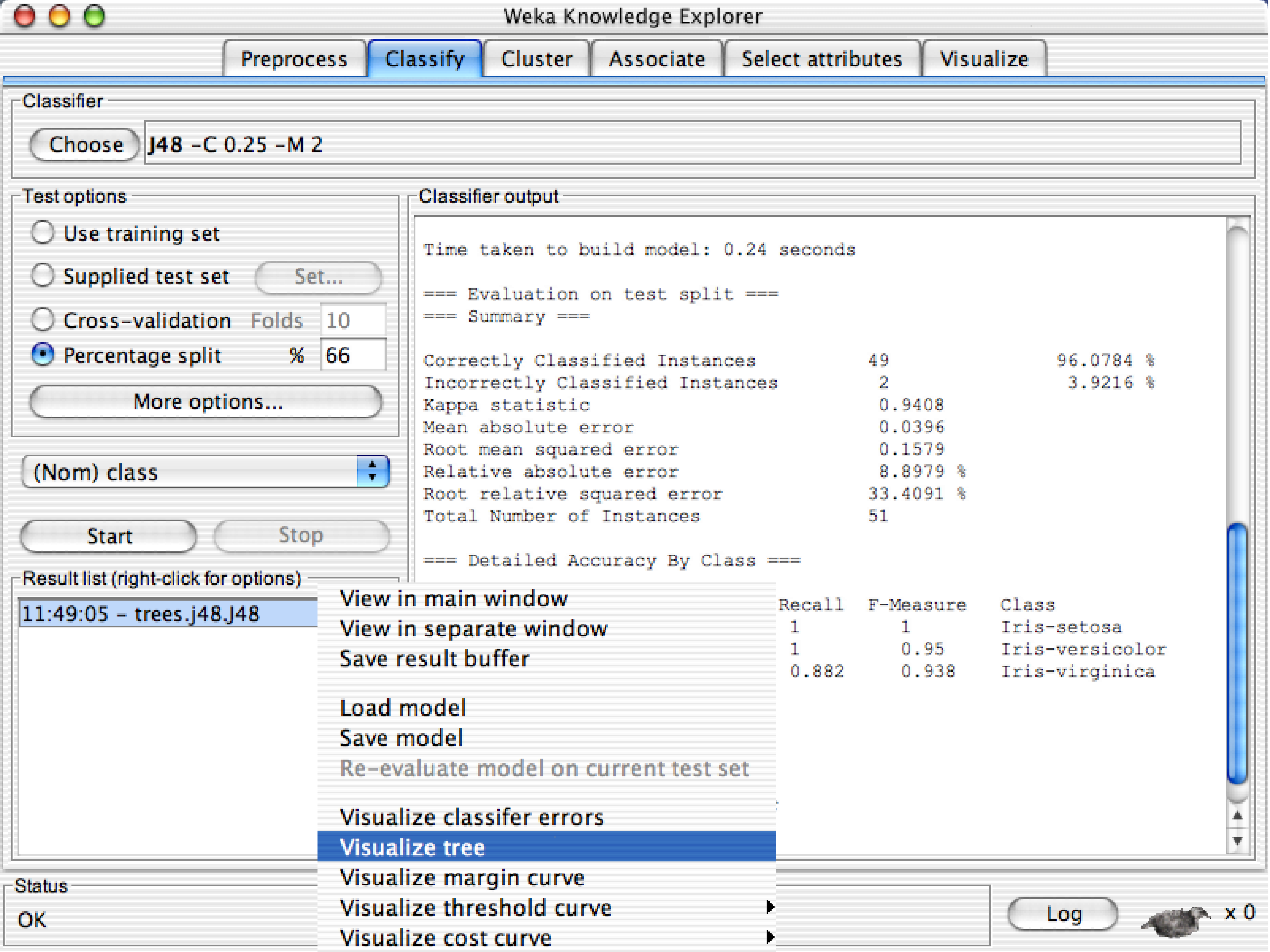
● Percentage split % 66

(Nom) class

11:49:05 - trees.j48.J48

```
Time taken to build model: 0.24 seconds

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

Correctly Classified Instances          49               96.0784 %
Incorrectly Classified Instances         2                3.9216 %
Kappa statistic                          0.9408
Mean absolute error                      0.0396
Root mean squared error                  0.1579
Relative absolute error                  8.8979 %
Root relative squared error             33.4091 %
Total Number of Instances               51

=== Detailed Accuracy By Class ===

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

=== Confusion Matrix ===

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

OK



J48 -C 0.25 -M 2

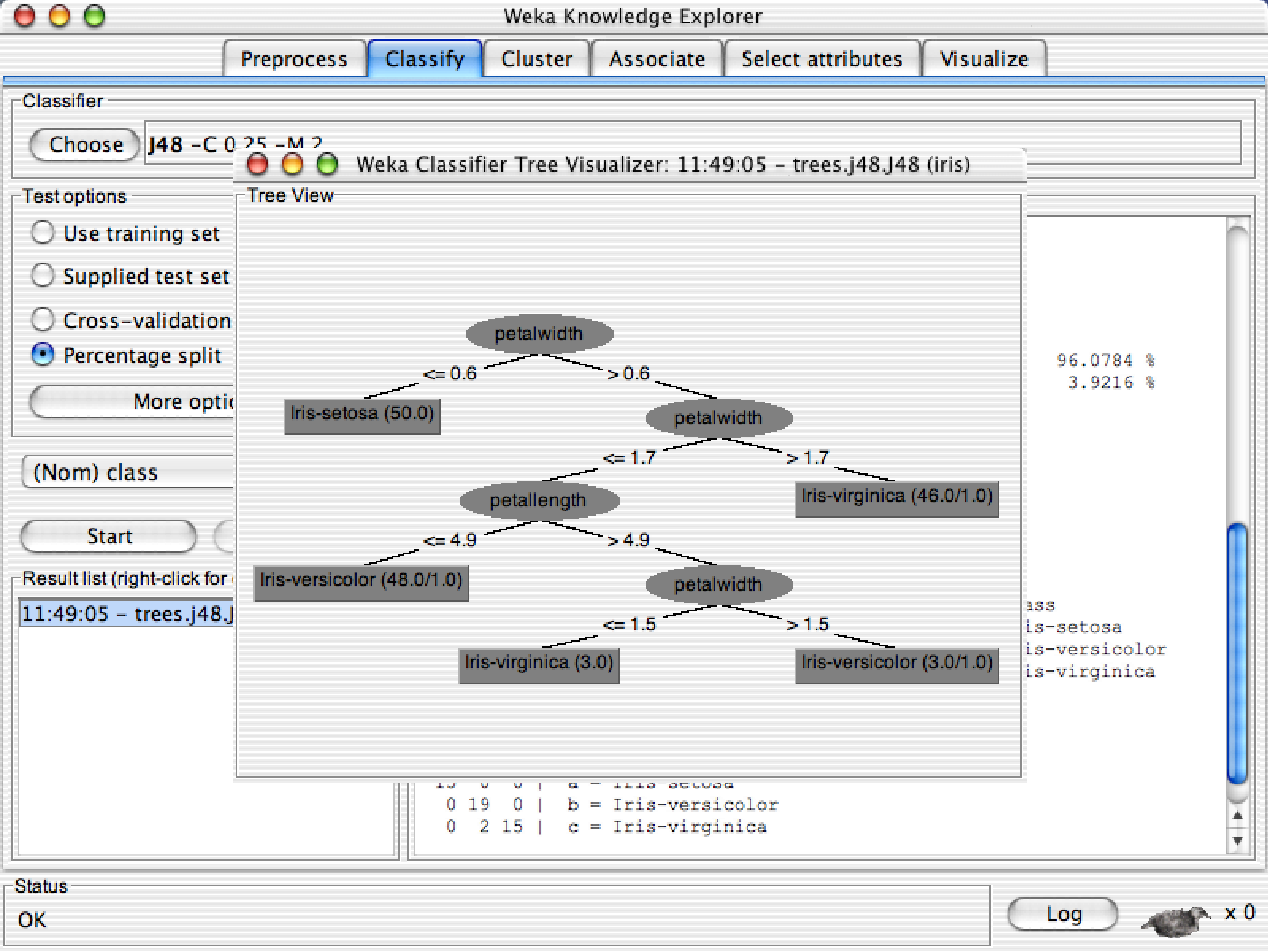
○ Use training set
○ Supplied test set
○ Cross-validation Folds 10
● Percentage split % 66

(Nom) class

11:49:05 - trees.j48.J48

```
Time taken to build model: 0.24 seconds

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

Correctly Classified Instances          49               96.0784 %
Incorrectly Classified Instances         2                3.9216 %
Kappa statistic                          0.9408
Mean absolute error                      0.0396
Root mean squared error                  0.1579
Relative absolute error                  8.8979 %
Root relative squared error             33.4091 %
Total Number of Instances               51

=== Detailed Accuracy By Class ===

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

Status
OK



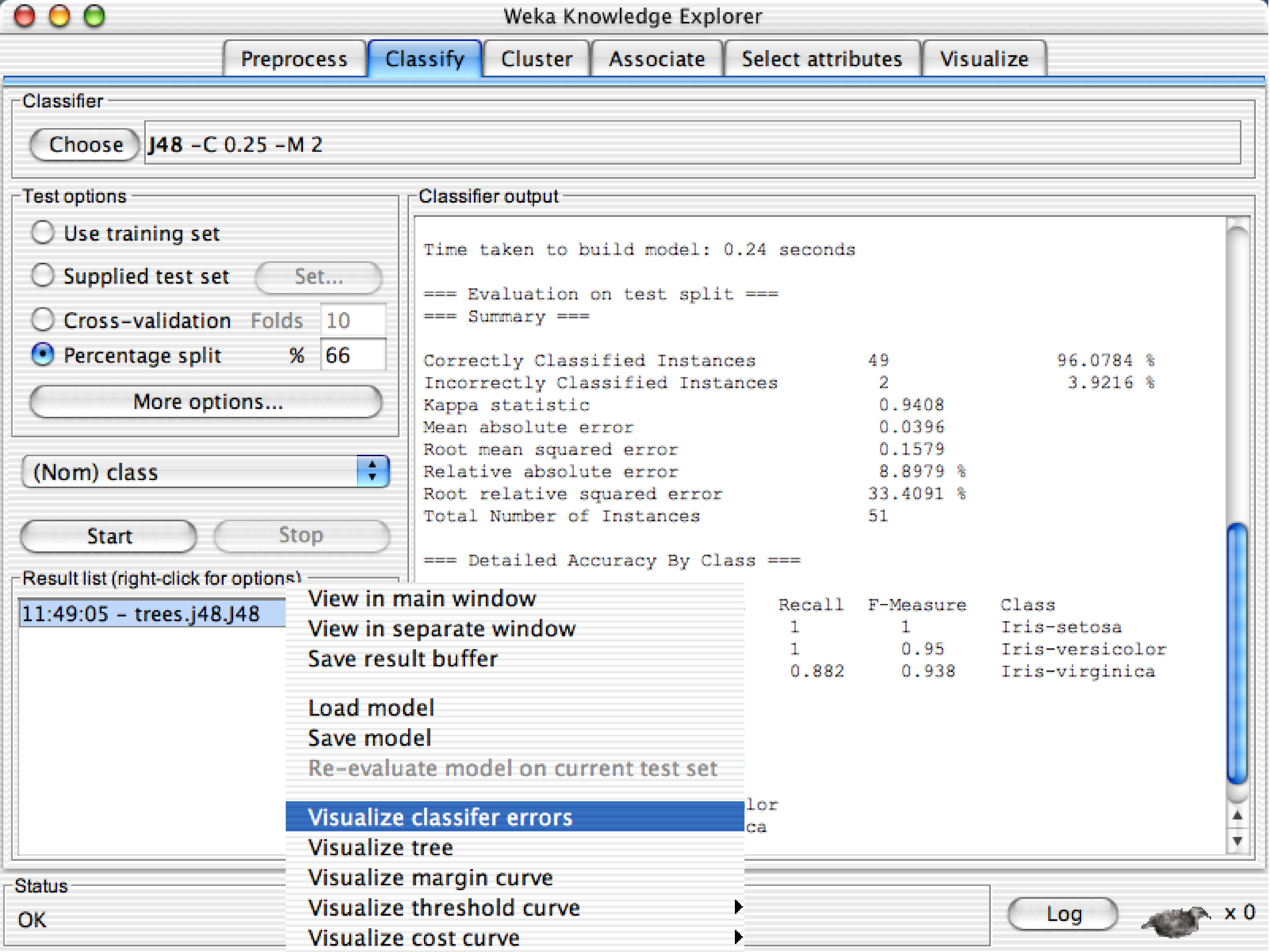
J48 -C 0.25 -M 2

(Nom) class

11:49:05 - trees.j48.J

```
96.0784 %
 3.9216 %

ass
is-setosa
is-versicolor
is-virginica

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

OK



Classifier

J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set

○ Cross-validation Folds 10

● Percentage split % 66

(Nom) class

Result list (right-click for options)

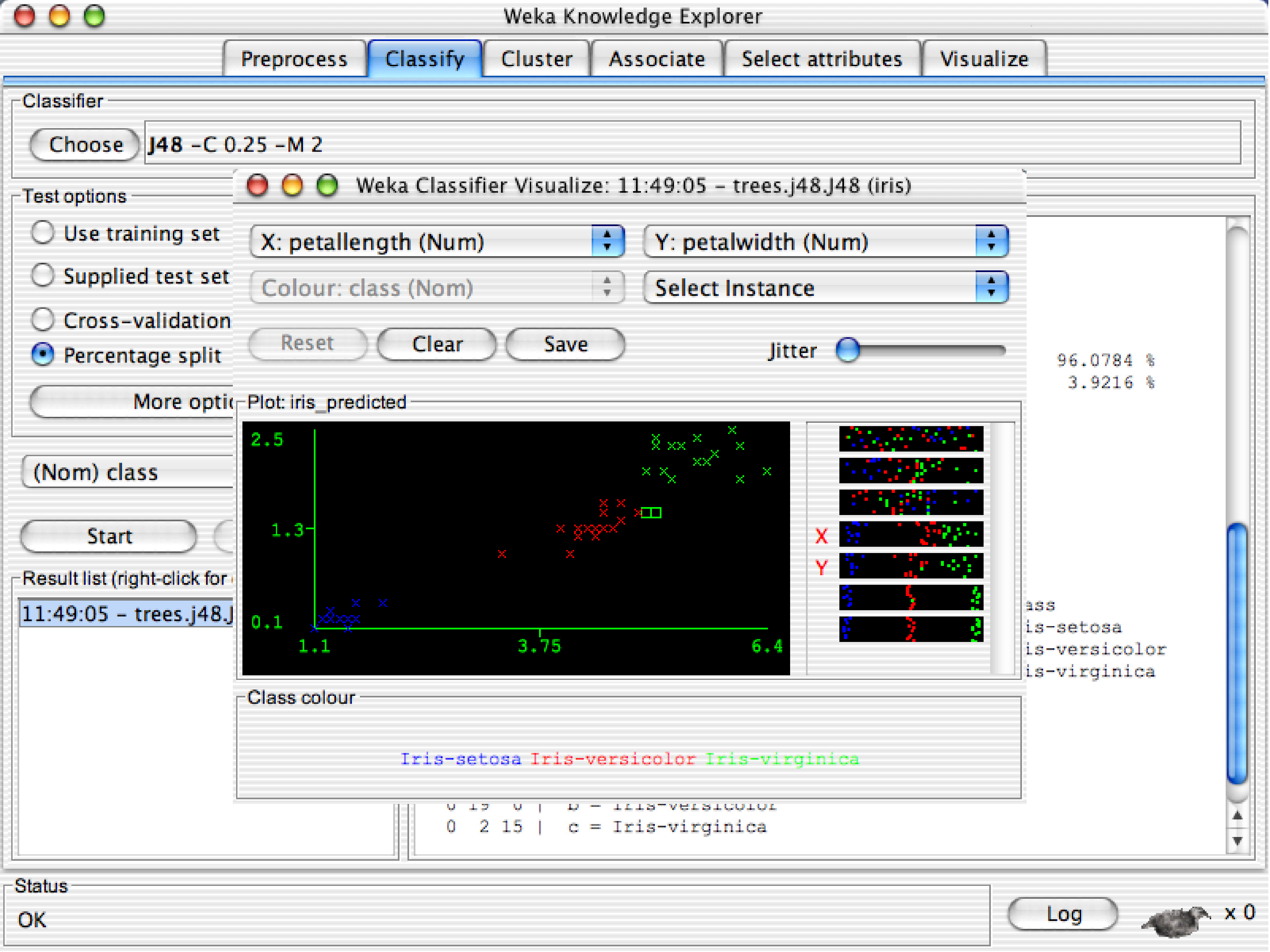
11:49:05 - trees.j48.J48

Classifier output

```
Time taken to build model: 0.24 seconds

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

Correctly Classified Instances          49               96.0784 %
Incorrectly Classified Instances         2                3.9216 %
Kappa statistic                          0.9408
Mean absolute error                      0.0396
Root mean squared error                  0.1579
Relative absolute error                  8.8979 %
Root relative squared error             33.4091 %
Total Number of Instances               51

=== Detailed Accuracy By Class ===

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




                  lor
                  ca
```

Status

OK



J48 -C 0.25 -M 2

(Nom) class

11:49:05 - trees.j48.J

X: petallength (Num)

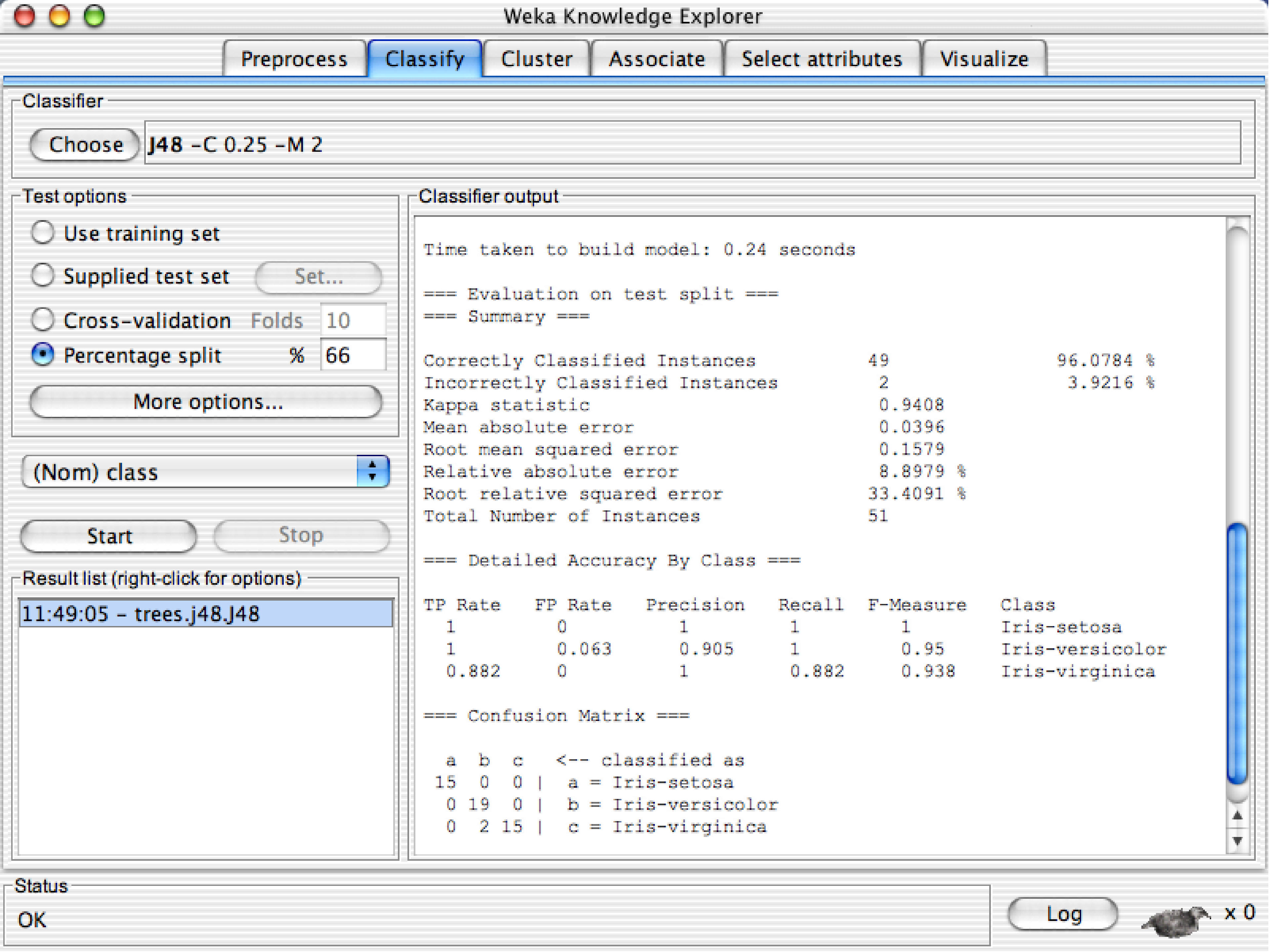
Y: petalwidth (Num)

Colour: class (Nom)

Select Instance

```
96.0784 %
 3.9216 %

ass
is-setosa
is-versicolor
is-virginica

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

OK



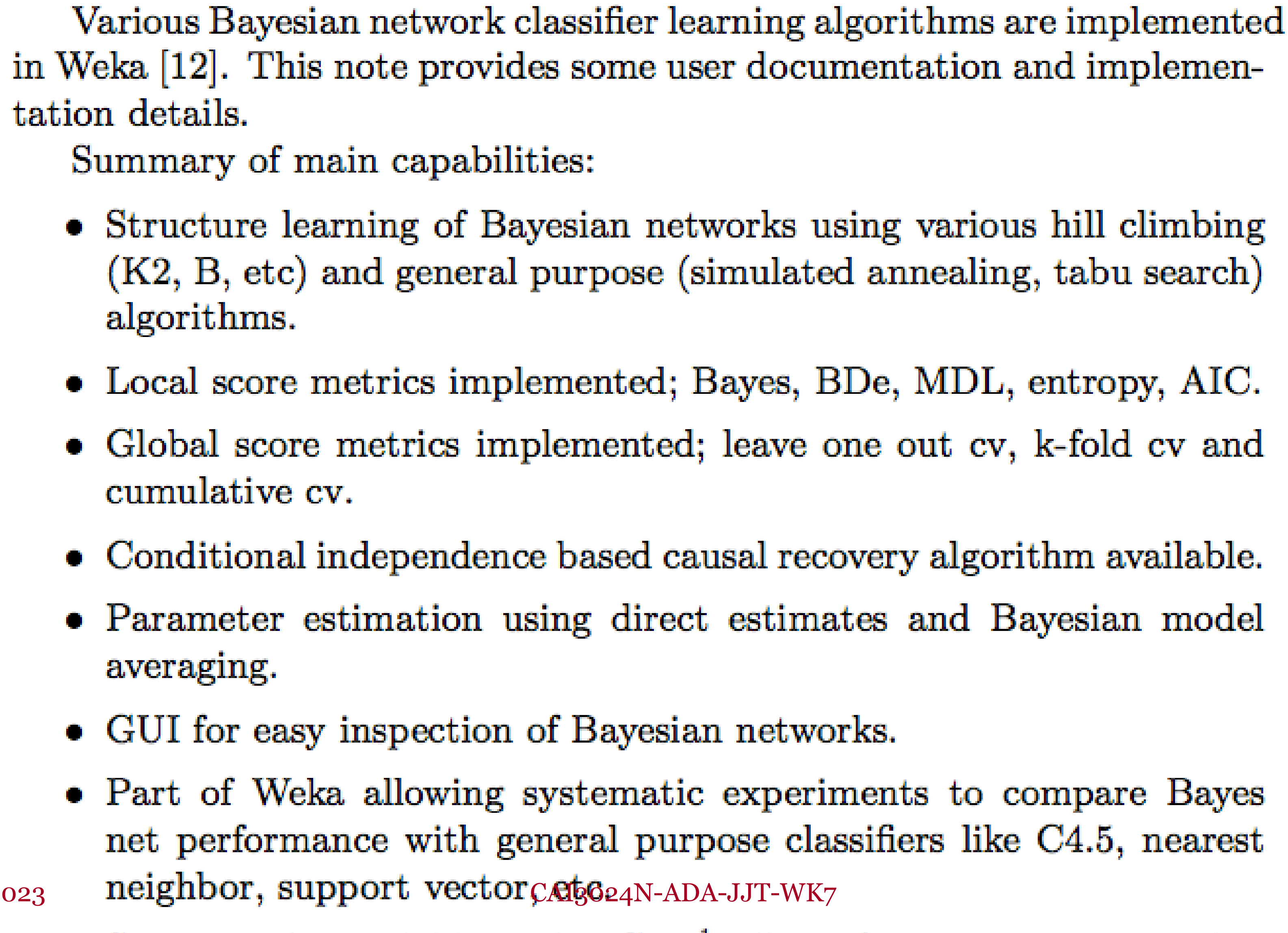
Classifier: J48 -C 0.25 -M 2

Result list: 11:49:05 - trees.j48.J48

```
Time taken to build model: 0.24 seconds

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

Correctly Classified Instances          49               96.0784 %
Incorrectly Classified Instances         2                3.9216 %
Kappa statistic                          0.9408
Mean absolute error                      0.0396
Root mean squared error                  0.1579
Relative absolute error                  8.8979 %
Root relative squared error             33.4091 %
Total Number of Instances               51

=== Detailed Accuracy By Class ===

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

Classify : Bayes

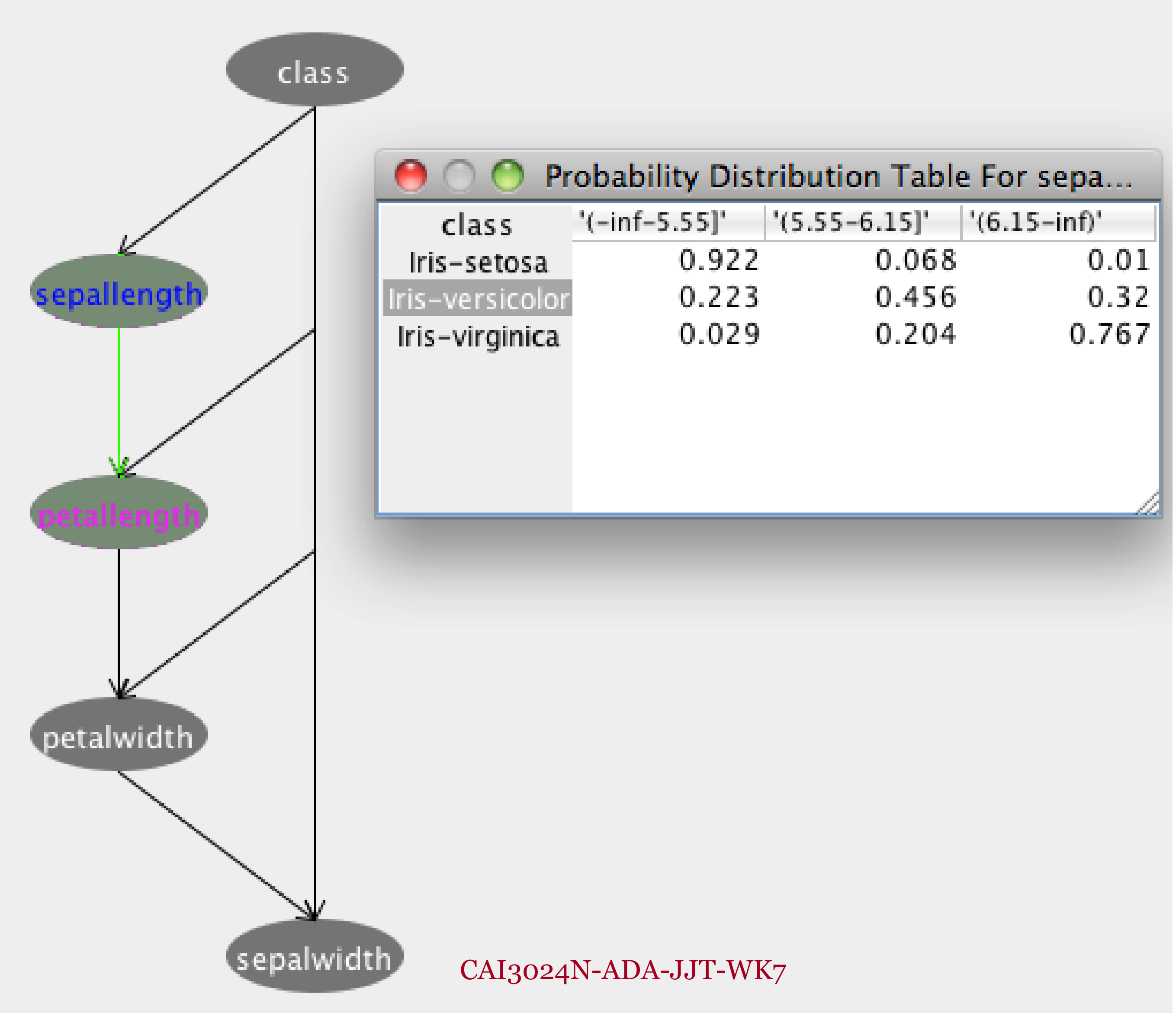


Various Bayesian network classifier learning algorithms are implemented in Weka [12]. This note provides some user documentation and implementation details.

Summary of main capabilities:

- Structure learning of Bayesian networks using various hill climbing (K2, B, etc) and general purpose (simulated annealing, tabu search) algorithms.
- Local score metrics implemented; Bayes, BDe, MDL, entropy, AIC.
- Global score metrics implemented; leave one out cv, k-fold cv and cumulative cv.
- Conditional independence based causal recovery algorithm available.
- Parameter estimation using direct estimates and Bayesian model averaging.
- GUI for easy inspection of Bayesian networks.
- Part of Weka allowing systematic experiments to compare Bayes net performance with general purpose classifiers like C4.5, nearest neighbor, support vector, etc.

Classify: Learning a Bayes Network Structure



Probability Distribution Table For sepa...

class	'(-inf-5.55]'	'(5.55-6.15]'	'(6.15-inf)'
Iris-setosa	0.922	0.068	0.01
Iris-versicolor	0.223	0.456	0.32
Iris-virginica	0.029	0.204	0.767

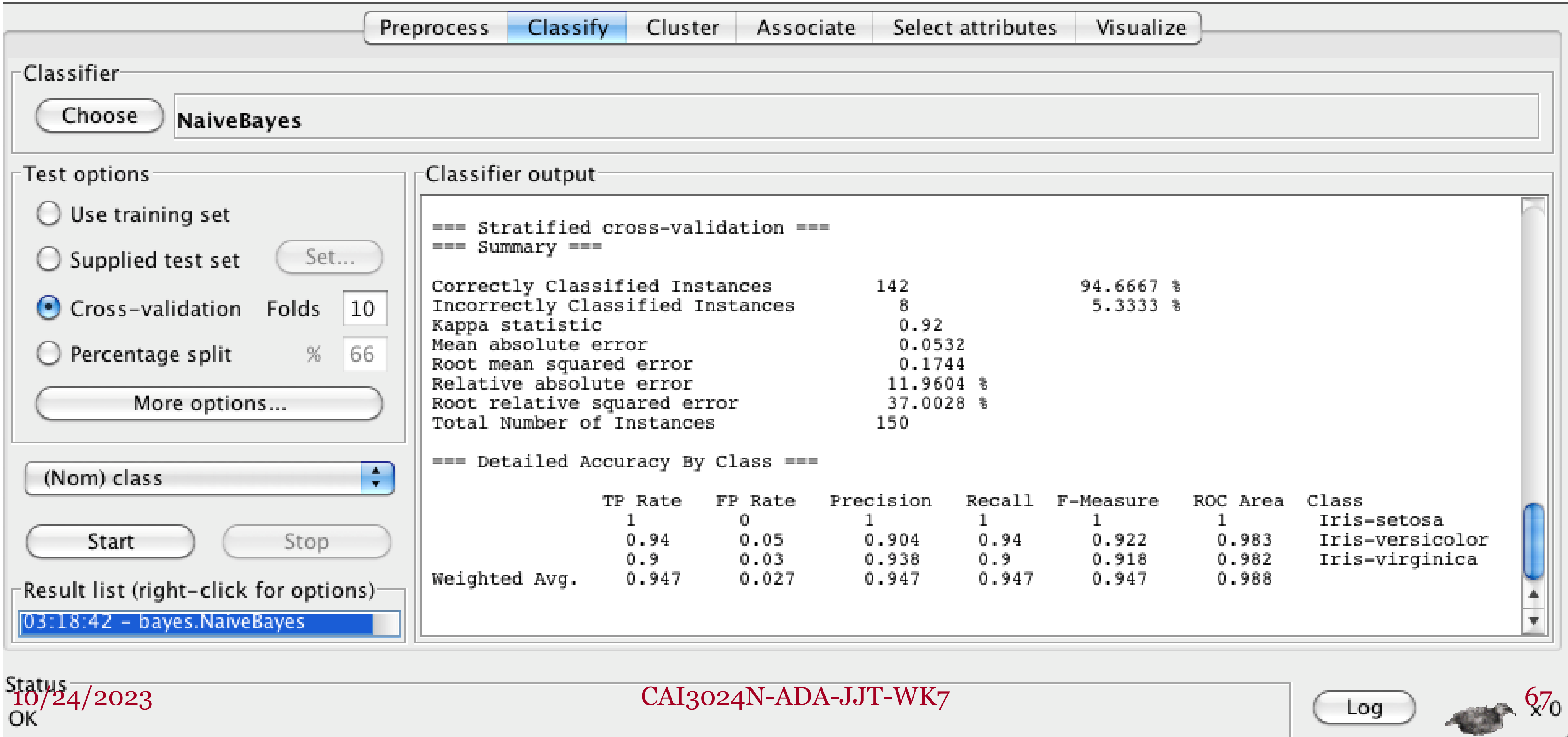
Classify using Naïve Bayes

- Naïve Bayes (standard)
- Multinominal Naïve Bayes (text classification)

- Hidden Naïve Bayes, others, ….

Classify using Naïve Bayes

1. Pick Discretization method* (hardest part)
2. Pick a class to predict
3. Run the classifier



Classifier: NaiveBayes

● Cross-validation Folds 10

(Nom) class

OK

03:18:42 - bayes.NaiveBayes

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         142               94.6667 %
Incorrectly Classified Instances         8                5.3333 %
Kappa statistic                          0.92
Mean absolute error                      0.0532
Root mean squared error                  0.1744
Relative absolute error                 11.9604 %
Root relative squared error             37.0028 %
Total Number of Instances              150

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0         1         1         1         1        Iris-setosa
                 0.94      0.05      0.904     0.94      0.922     0.983    Iris-versicolor
                 0.9       0.03      0.938     0.9       0.918     0.982    Iris-virginica
Weighted Avg.    0.947     0.027     0.947     0.947     0.947     0.988
```

Classify Text using Naïve Bayes Multinominal

- Framingham dataset contains text descriptions for each of the 21k+ variables
- I wrote a parallelized NLP program to calculate inverse word frequencies and score variable pairs (2 days)
- Highest scoring pairs were suggested for merger to reduce the variable space (curse of dimensionality)

Classify Text using Naïve Bayes Multinominal

What I should have done….

① Train an NBC to learn from small set of labeled cases

② Apply NBC to unlabeled data using Expectation Maximization with class probabilities (expectation step)

③ Retrain NBC using the labels for all the data

④ Repeat until convergence

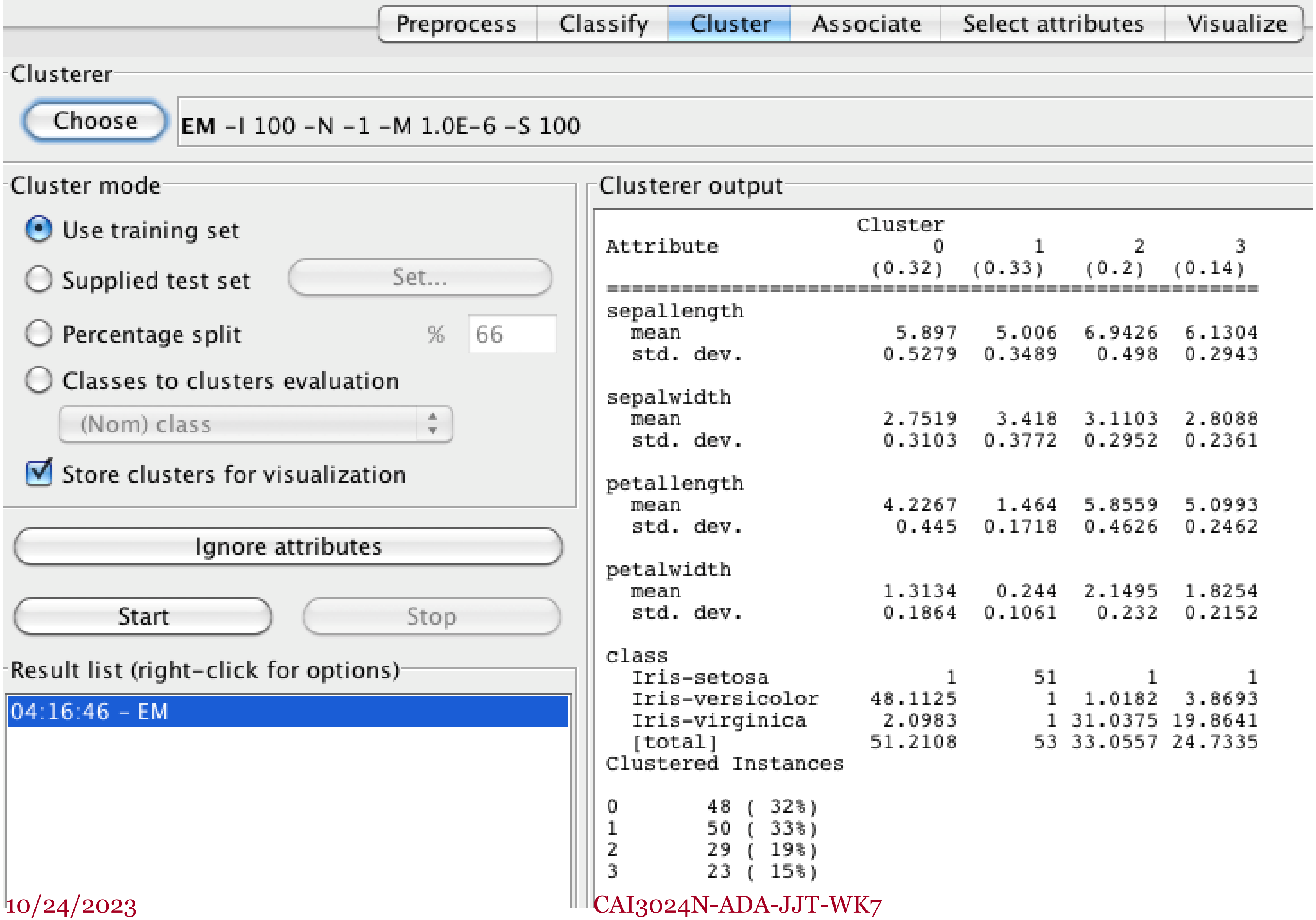
WEKA Explorer Tutorial Examples

- Preprocess
  - Instance and Attribute Filters (Supervised and Unsupervised)
- Classify
  - Bayes
- **Cluster**
  - **Expectation Maximization**
  - **Hierarchical Clustering**
- Associate
  - Apriori
- Select Attributes
  - Via clustering

Cluster Algorithm Examples

- Expectation Maximization (EM)
- Hierarchical Clustering (cobweb)
- *Note: Weka provides many more clustering methods*

Cluster with Expectation Maximization



Clusterer

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

Cluster mode

● Use training set

○ Supplied test set

○ Percentage split % 66

○ Classes to clusters evaluation

(Nom) class

☑ Store clusters for visualization

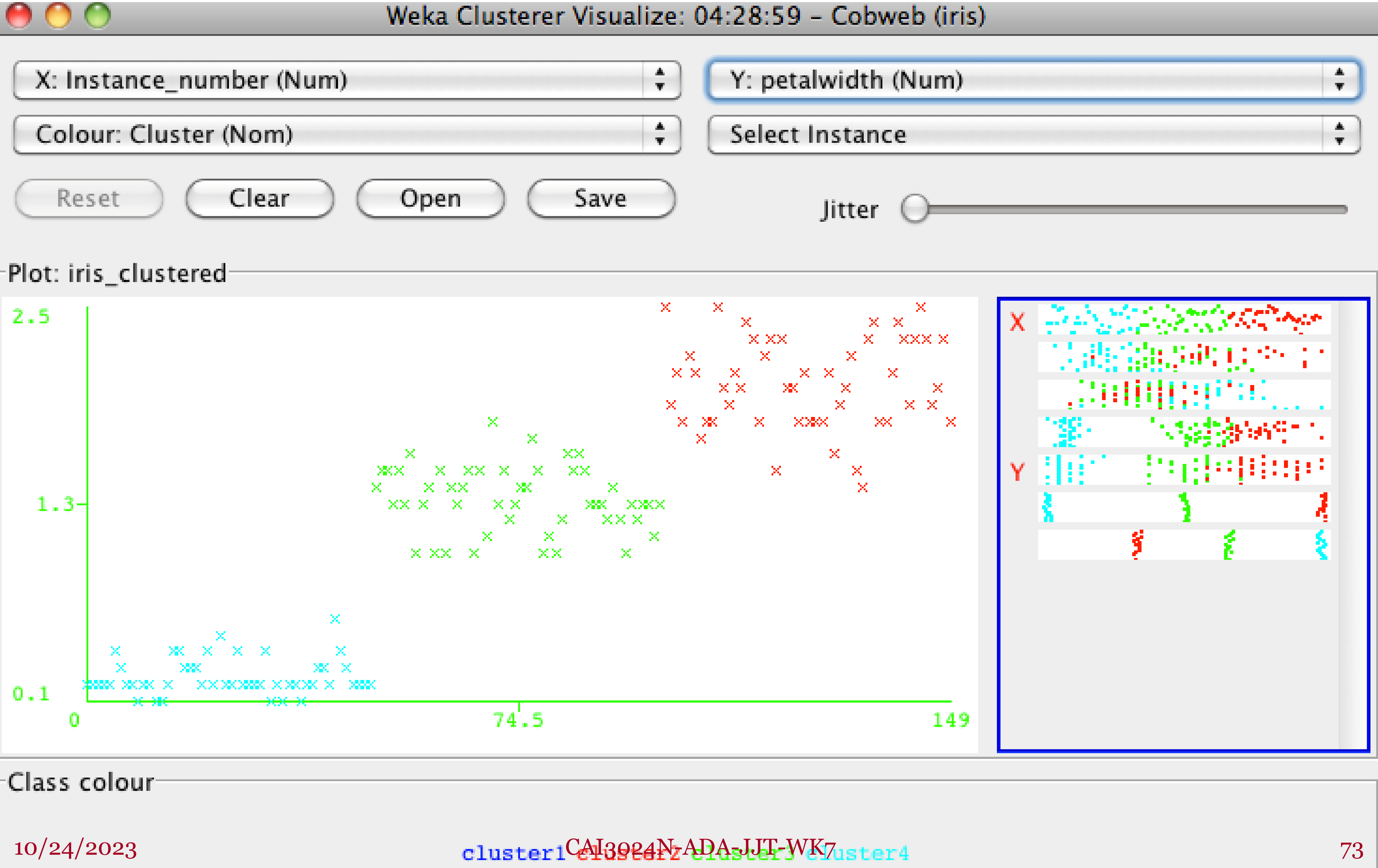
Result list (right-click for options)

04:16:46 - EM

Clusterer output

```
                        Cluster
Attribute                     0        1        2        3
                         (0.32)   (0.33)    (0.2)   (0.14)
===========================================================
sepallength
  mean                    5.897    5.006   6.9426   6.1304
  std. dev.              0.5279   0.3489    0.498   0.2943

sepalwidth
  mean                   2.7519    3.418   3.1103   2.8088
  std. dev.              0.3103   0.3772   0.2952   0.2361

petallength
  mean                   4.2267    1.464   5.8559   5.0993
  std. dev.               0.445   0.1718   0.4626   0.2462

petalwidth
  mean                   1.3134    0.244   2.1495   1.8254
  std. dev.              0.1864   0.1061    0.232   0.2152

class
  Iris-setosa                 1       51        1        1
  Iris-versicolor       48.1125        1   1.0182   3.8693
  Iris-virginica         2.0983        1  31.0375  19.8641
  [total]               51.2108       53  33.0557  24.7335
Clustered Instances

0       48 ( 32%)
1       50 ( 33%)
2       29 ( 19%)
3       23 ( 15%)
```

Cluster with cobweb (hierarchical clustering)



Weka Clusterer Visualize: 04:28:59 - Cobweb (iris)

X: Instance_number (Num)

Y: petalwidth (Num)

Colour: Cluster (Nom)

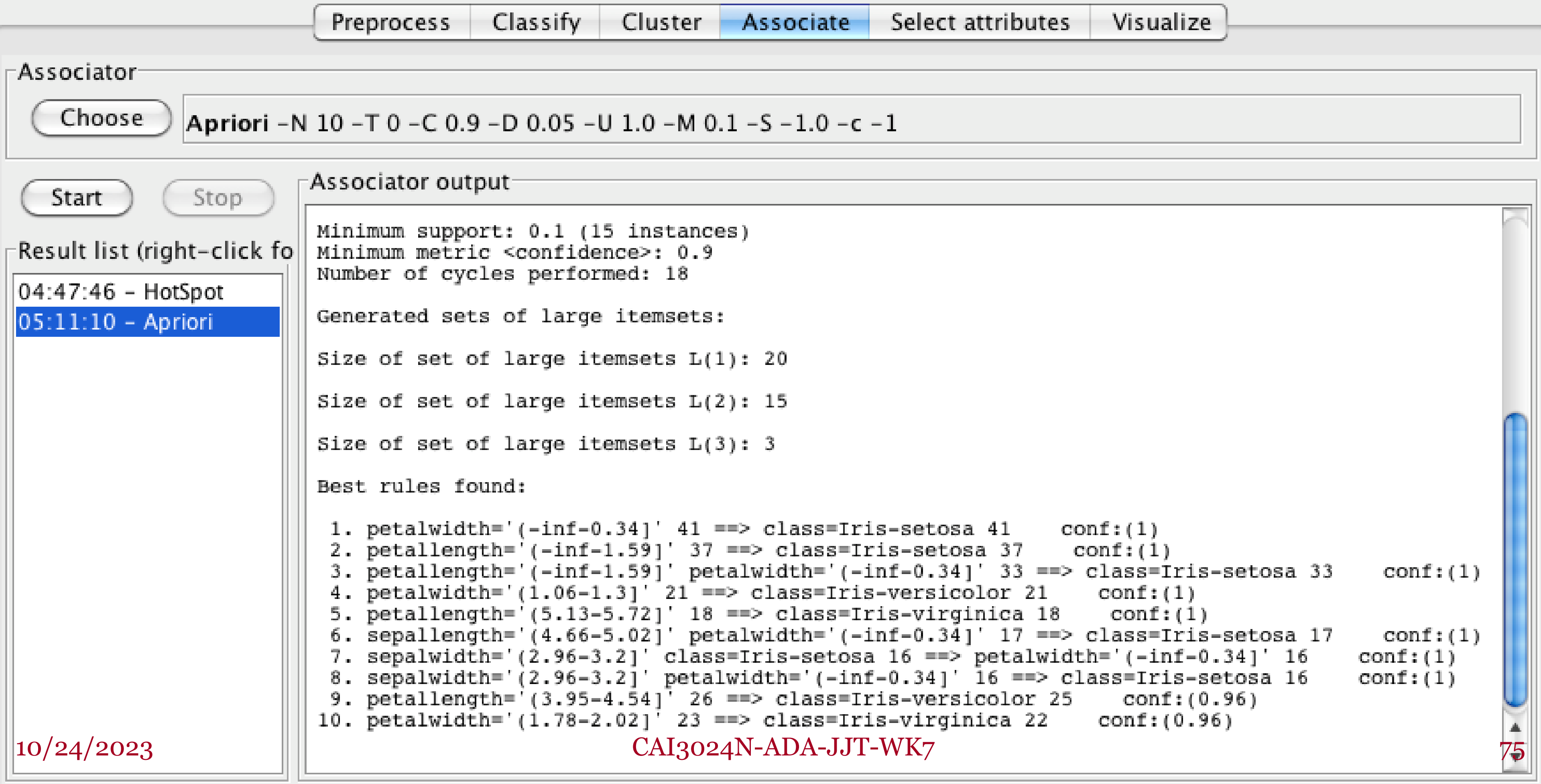
Plot: iris_clustered

WEKA Explorer Tutorial Examples

- Preprocess
  - Instance and Attribute Filters (Supervised and Unsupervised)
- Classify
  - Bayes
- Cluster
  - Expectation Maximization
  - Hierarchical Clustering
- **Associate**
  - **Apriori**
- Select Attributes
  - Via clustering

Associate

- Quick scan for association rules
- see "Fast Algorithms for Mining Association Rules in Large Databases"



Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

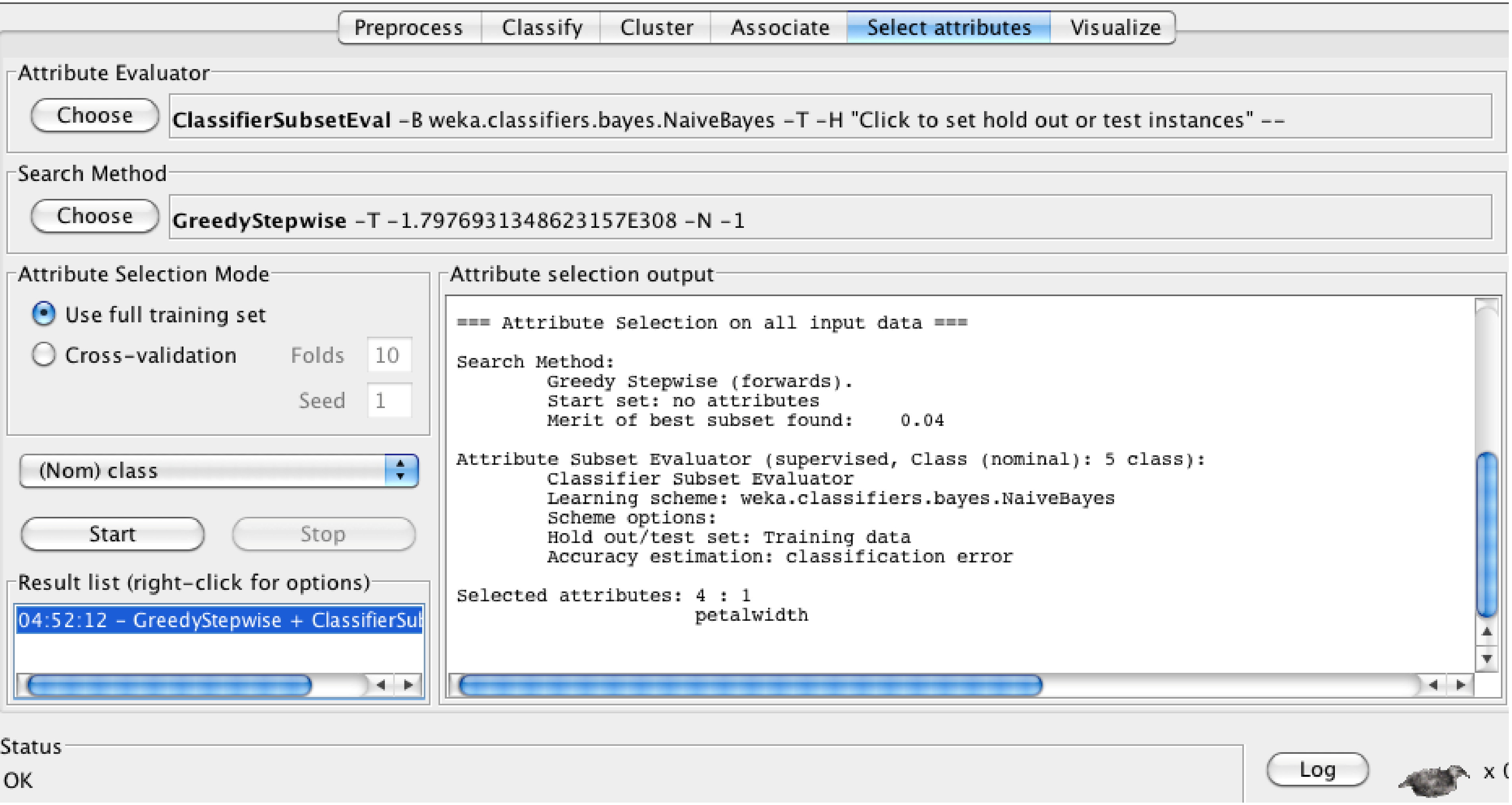
04:47:46 - HotSpot
05:11:10 - Apriori

```
Minimum support: 0.1 (15 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 18

Generated sets of large itemsets:

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

Size of set of large itemsets L(2): 15

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

Best rules found:

 1. petalwidth='(-inf-0.34]' 41 ==> class=Iris-setosa 41    conf:(1)
 2. petallength='(-inf-1.59]' 37 ==> class=Iris-setosa 37    conf:(1)
 3. petallength='(-inf-1.59]' petalwidth='(-inf-0.34]' 33 ==> class=Iris-setosa 33    conf:(1)
 4. petalwidth='(1.06-1.3]' 21 ==> class=Iris-versicolor 21    conf:(1)
 5. petallength='(5.13-5.72]' 18 ==> class=Iris-virginica 18    conf:(1)
 6. sepallength='(4.66-5.02]' petalwidth='(-inf-0.34]' 17 ==> class=Iris-setosa 17    conf:(1)
 7. sepalwidth='(2.96-3.2]' class=Iris-setosa 16 ==> petalwidth='(-inf-0.34]' 16    conf:(1)
 8. sepalwidth='(2.96-3.2]' petalwidth='(-inf-0.34]' 16 ==> class=Iris-setosa 16    conf:(1)
 9. petallength='(3.95-4.54]' 26 ==> class=Iris-versicolor 25    conf:(0.96)
10. petalwidth='(1.78-2.02]' 23 ==> class=Iris-virginica 22    conf:(0.96)
```

WEKA Explorer Tutorial Examples

- Preprocess
  - Instance and Attribute Filters (Supervised and Unsupervised)
- Classify
  - Bayes
- Cluster
  - Expectation Maximization
  - Hierarchical Clustering
- Associate
  - Apriori
- **Select Attributes**
  - **Via clustering**

Select Attributes Using a Classifier



Attribute Evaluator

ClassifierSubsetEval -B weka.classifiers.bayes.NaiveBayes -T -H "Click to set hold out or test instances" --

Search Method

GreedyStepwise -T -1.7976931348623157E308 -N -1

Attribute Selection Mode

● Use full training set

○ Cross-validation Folds 10

Seed 1

(Nom) class

Result list (right-click for options)

04:52:12 - GreedyStepwise + ClassifierSul

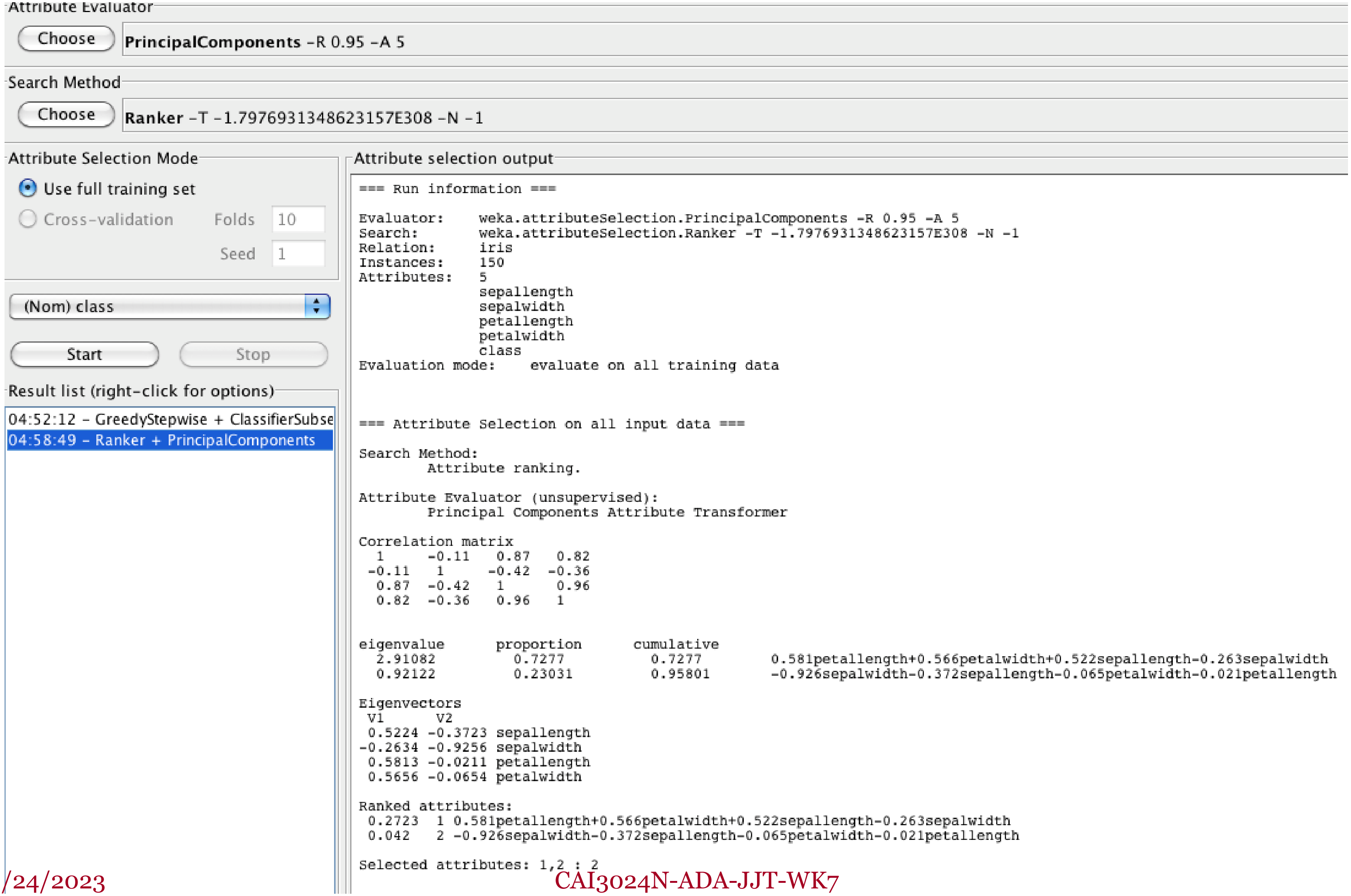
Attribute selection output

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

Search Method:
        Greedy Stepwise (forwards).
        Start set: no attributes
        Merit of best subset found:    0.04

Attribute Subset Evaluator (supervised, Class (nominal): 5 class):
        Classifier Subset Evaluator
        Learning scheme: weka.classifiers.bayes.NaiveBayes
        Scheme options:
        Hold out/test set: Training data
        Accuracy estimation: classification error

Selected attributes: 4 : 1
                     petalwidth
```

Status

OK

Select Attributes using PCA



Attribute Evaluator

PrincipalComponents -R 0.95 -A 5

Search Method

Ranker -T -1.7976931348623157E308 -N -1

Attribute Selection Mode

● Use full training set

○ Cross-validation Folds 10

Seed 1

(Nom) class

04:52:12 - GreedyStepwise + ClassifierSubse

04:58:49 - Ranker + PrincipalComponents

Attribute selection output

```
=== Run information ===

Evaluator:    weka.attributeSelection.PrincipalComponents -R 0.95 -A 5
Search:       weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1
Relation:     iris
Instances:    150
Attributes:   5
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
Evaluation mode:    evaluate on all training data



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

Search Method:
        Attribute ranking.

Attribute Evaluator (unsupervised):
        Principal Components Attribute Transformer

Correlation matrix
  1      -0.11   0.87   0.82
 -0.11   1      -0.42  -0.36
  0.87  -0.42   1       0.96
  0.82  -0.36   0.96    1


eigenvalue      proportion      cumulative
  2.91082         0.7277          0.7277        0.581petallength+0.566petalwidth+0.522sepallength-0.263sepalwidth
  0.92122         0.23031         0.95801       -0.926sepalwidth-0.372sepallength-0.065petalwidth-0.021petallength

Eigenvectors
 V1       V2
 0.5224 -0.3723 sepallength
-0.2634 -0.9256 sepalwidth
 0.5813 -0.0211 petallength
 0.5656 -0.0654 petalwidth

Ranked attributes:
 0.2723   1 0.581petallength+0.566petalwidth+0.522sepallength-0.263sepalwidth
 0.042    2 -0.926sepalwidth-0.372sepallength-0.065petalwidth-0.021petallength

Selected attributes: 1,2 : 2
```

WEKA Tutorial Summary

- Preprocess
  - Prepare datasets instances and attributes before analysis
- Classify
  - Pick a instance and predict the class
    - Iris : Pick a flower and use the attributes to predict species
    - Medicine: pick a patient and use the genes to predict cancer status
- Cluster
  - Group instances together (flowers, breast cancer cases, etc)
- Associate
  - Discover relationships between variables in your dataset

References

- Data Mining: Practical Machine Learning Tools and Techniques
- Data Mining (I. H. Witten and E. Frank)
- WEKA Exploratory Tool for Data Mining
- Bayesian Network Classifiers in Weka (Bouckaert)
- COC131 Data Mining – Clustering (Sykora)
- Fast, Correct Multithreaded Programs in Java (Gilbert)
- R. Agrawal, R. Srikant: Fast Algorithms for Mining Association Rules in Large Databases. In: 20th International Conference on Very Large Data Bases, 478-499, 1994.
- WEKA Wiki
  http://weka.wikispaces.com/
- Graphical User Interface
  http://prdownloads.sourceforge.net/weka/weka.ppt

References

- **OL- Ebook**