Weka: Preprocessing

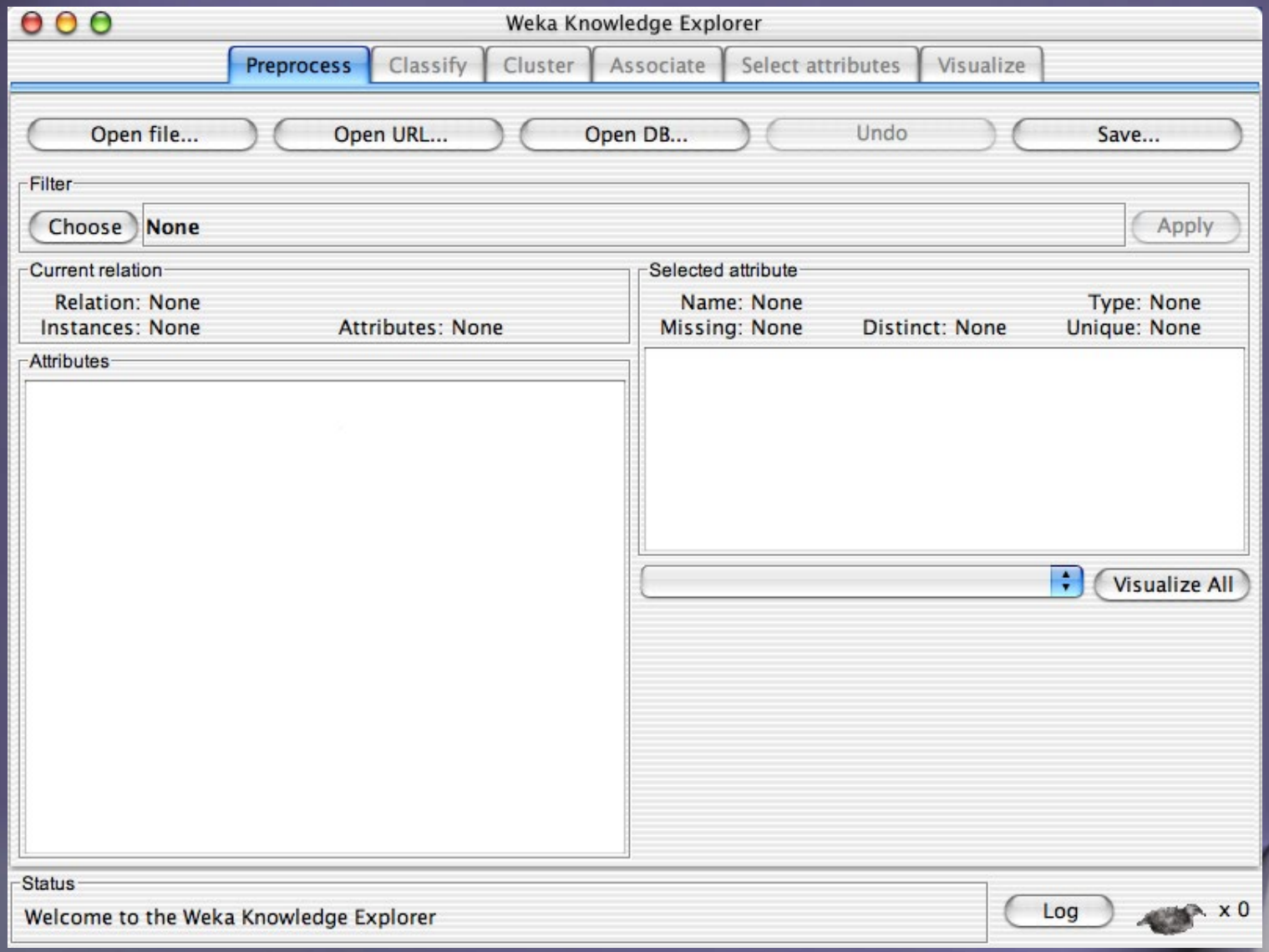
Cristóbal Romero,
Córdoba University, Campus Universitario de Rabanales, 14071, Córdoba, Spain
cromero@uco.es,

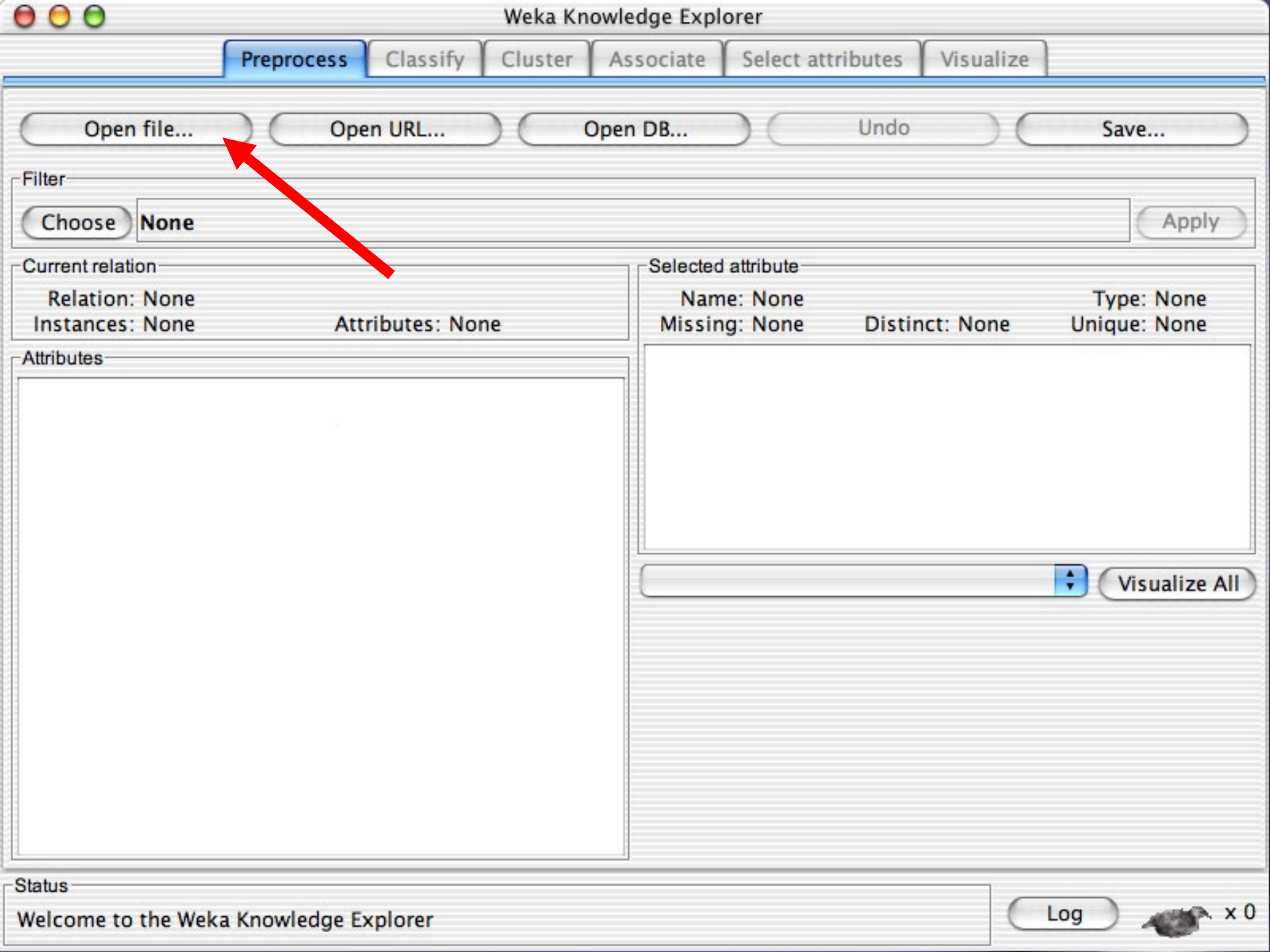
KDIS Lab
KNOWLEDGE DISCOVERY AND INTELLIGENT SYSTEMS

Preprocess Panel

- 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, …

Preprocess Panel: Visualize




Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
None
Apply
Current relation
Relation: None
Instances: None
Attributes: None
Selected attribute
Name: None
Type: None
Missing: None
Distinct: None
Unique: None
Attributes
Visualize All
Status
Welcome to the Weka Knowledge Explorer
Log
x 0

Weka Knowledge Explorer

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

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

Filter

Choose **None** | Apply

Current relation

Relation: iris
Instances: 150 Attributes: 5

Attributes

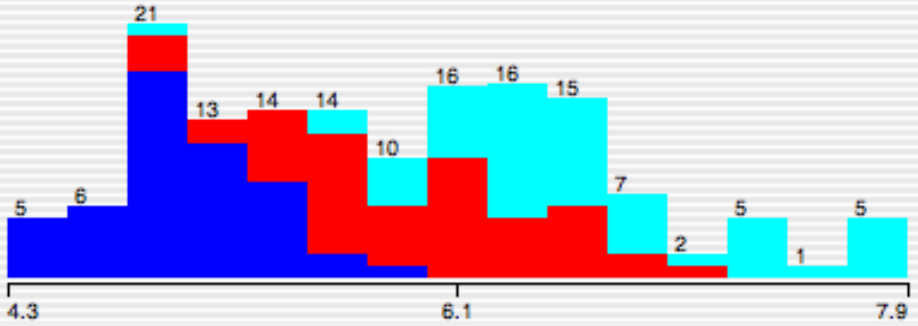
No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

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

Weka Knowledge Explorer

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

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

Filter

Choose **None** | Apply

Current relation

Relation: iris
Instances: 150 Attributes: 5

Attributes

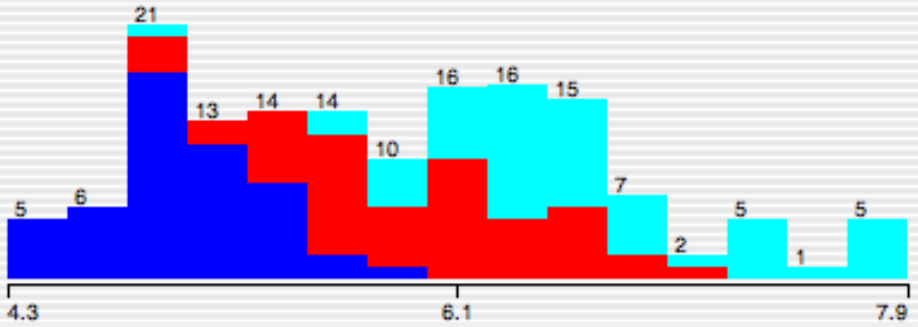
No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

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

Weka Knowledge Explorer

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

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

Filter

Choose **None** | Apply

Current relation

Relation: 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%)

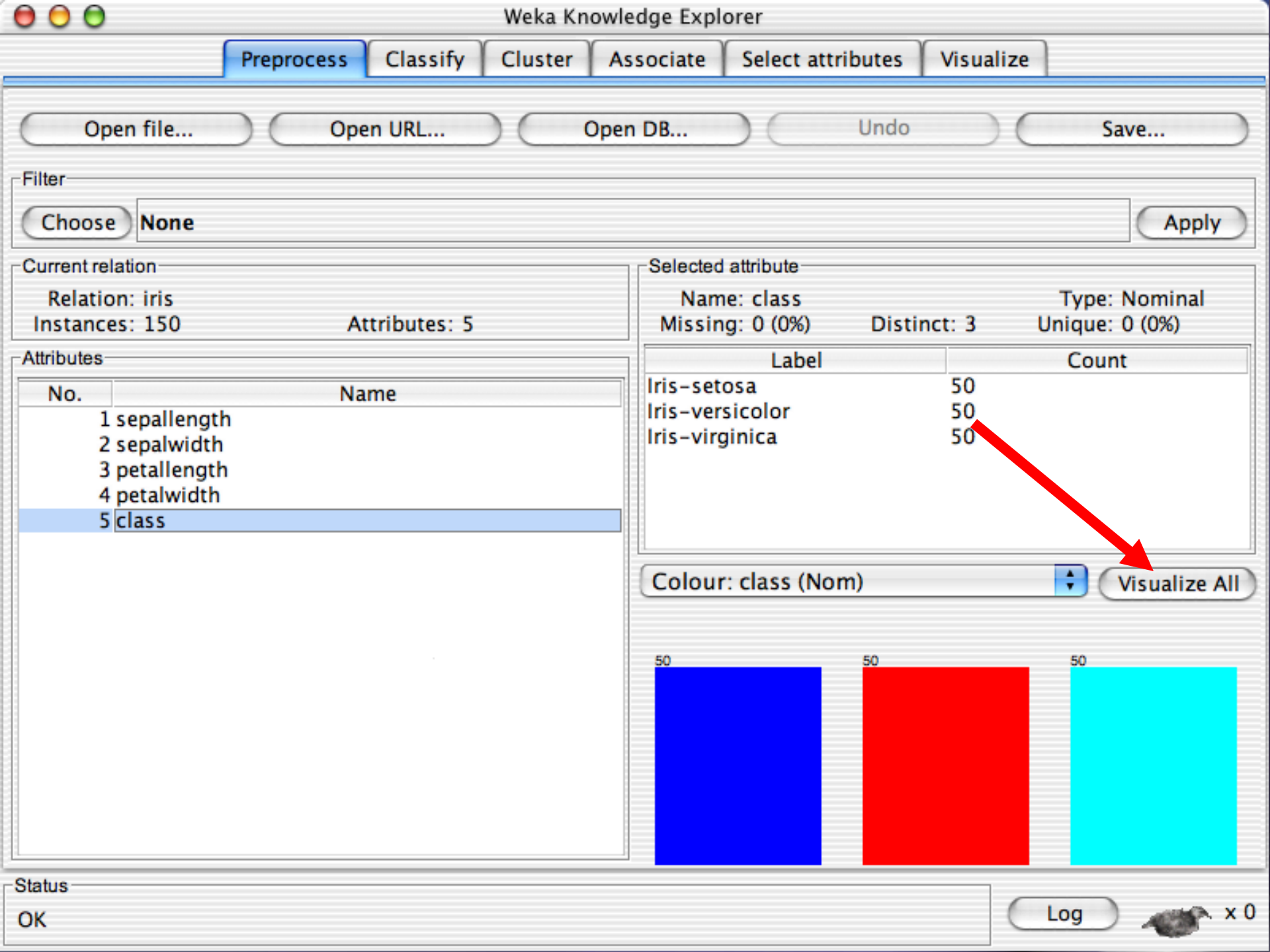
Label	Count
Iris-setosa	50
Iris-versicolor	50
Iris-virginica	50

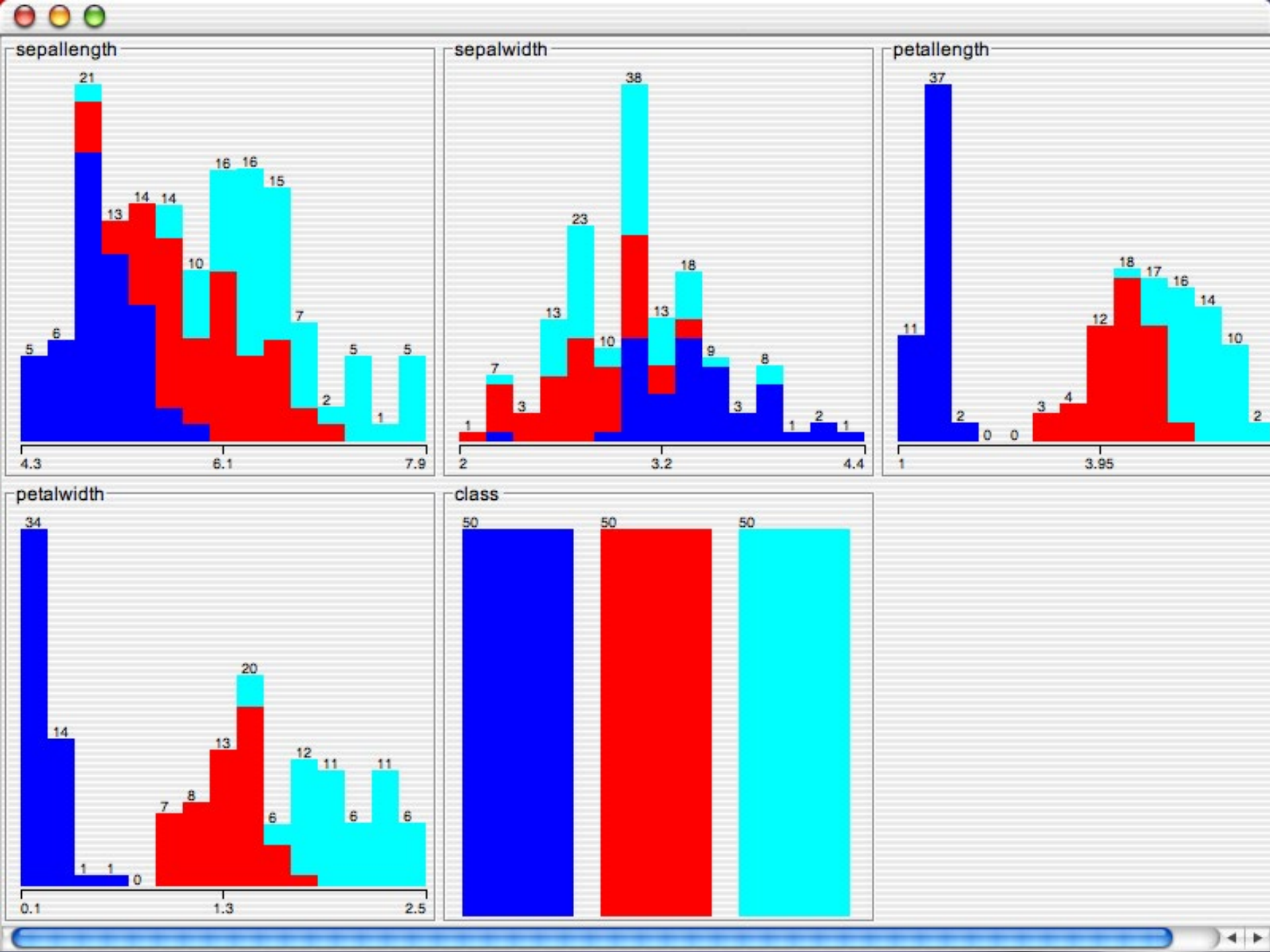
Colour: class (Nom) | Visualize All

Status

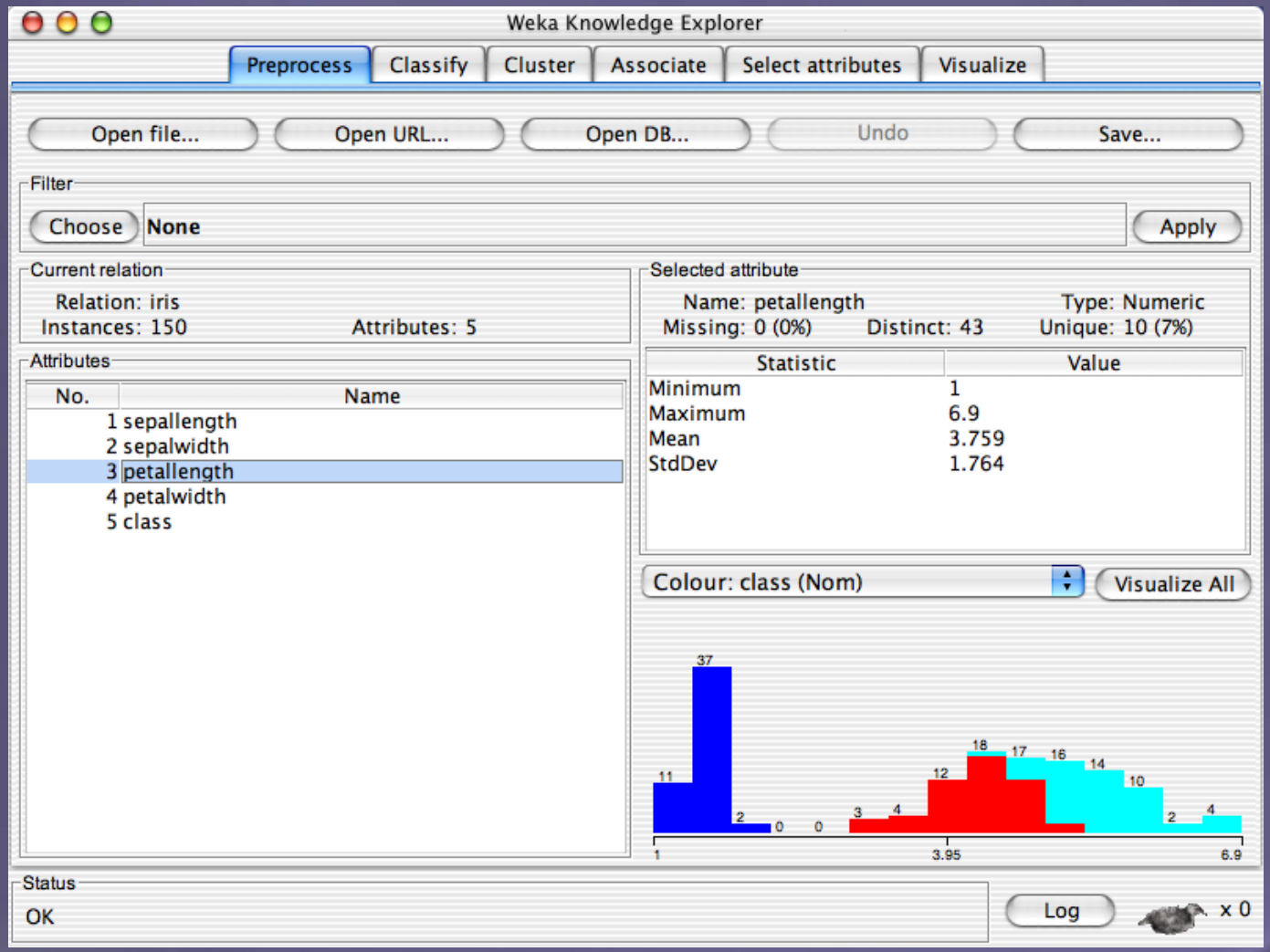
OK

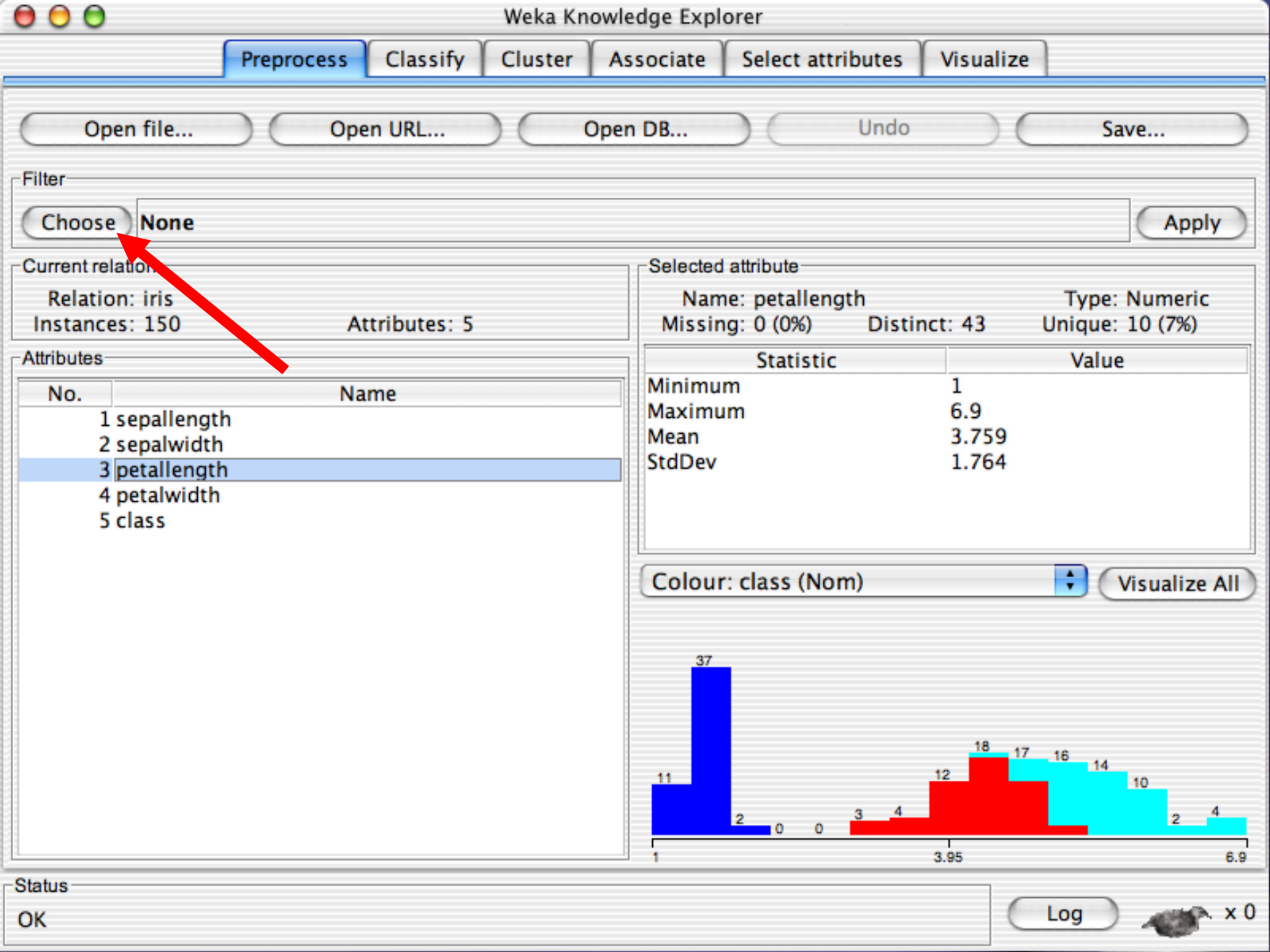
Log x 0


Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
None
Apply
Current relation
Relation: 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)
Visualize All
50
50
50
Status
OK
Log
x 0


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

Preprocess Panel: Discretize




Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
None
Apply
Current relation
Relation: 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
37
11
2
0
0
3
4
12
18
17
16
14
10
2
4
1
3.95
6.9
Status
OK
Log
x 0

Weka Knowledge Explorer

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

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

Filter

- weka
  - filters
    - unsupervised
      - attribute
      - instance

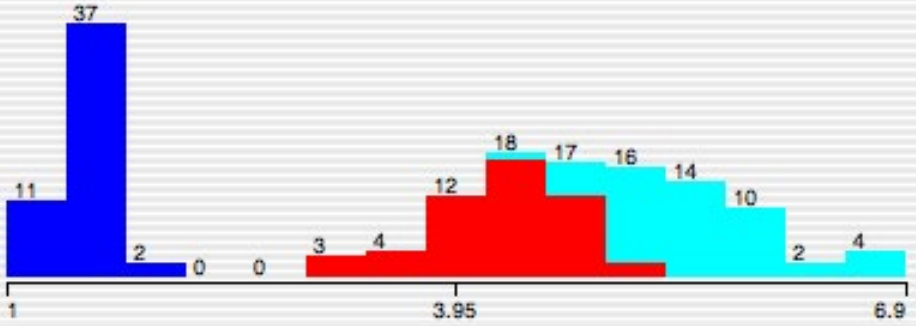
Apply

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

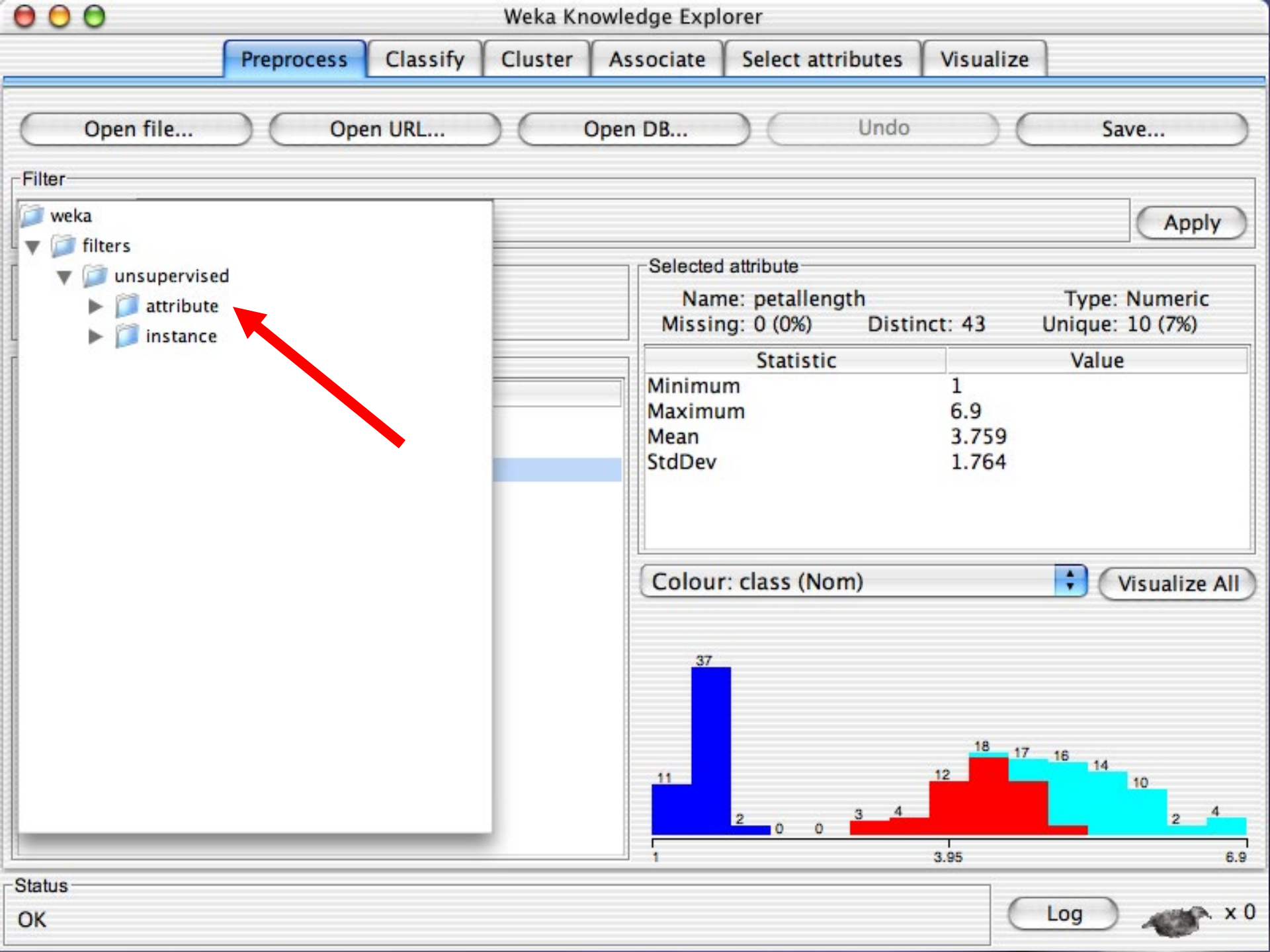
Colour: class (Nom) | Visualize All

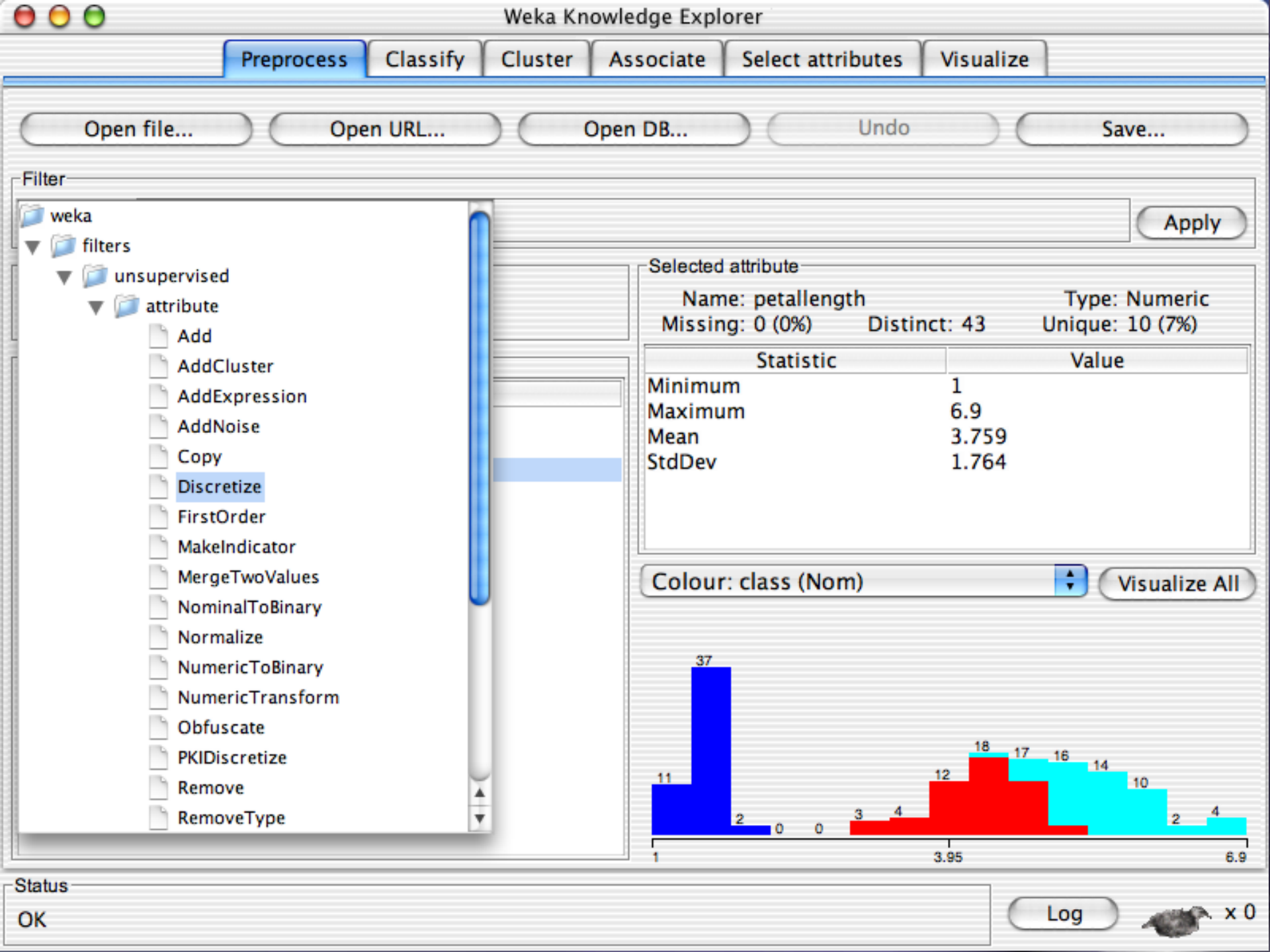


Status

OK

Log x 0


Weka Knowledge Explorer
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
37
11
2
0
0
3
4
12
18
17
16
14
10
2
4
1
3.95
6.9
Status
OK
Log
x 0


Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
weka
filters
unsupervised
attribute
Add
AddCluster
AddExpression
AddNoise
Copy
Discretize
FirstOrder
MakeIndicator
MergeTwoValues
NominalToBinary
Normalize
NumericToBinary
NumericTransform
Obfuscate
PKIDiscretize
Remove
RemoveType
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
37
11
2
0
0
3
4
12
18
17
16
14
10
2
4
1
3.95
6.9
Status
OK
Log
x 0

Weka Knowledge Explorer

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

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

Filter

Choose **Discretize** -B 10 -R first-last | Apply

Current relation

Relation: iris
Instances: 150 Attributes: 5

Attributes

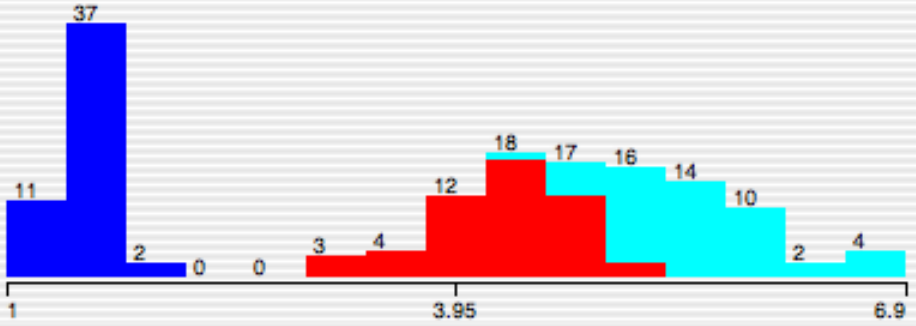
No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

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

Weka Knowledge Explorer

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

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

Filter

Choose | **Discretize** -B 10 -R first-last | Apply

Current relation

Relation: iris
Instances: 150 Attributes: 5

Attributes

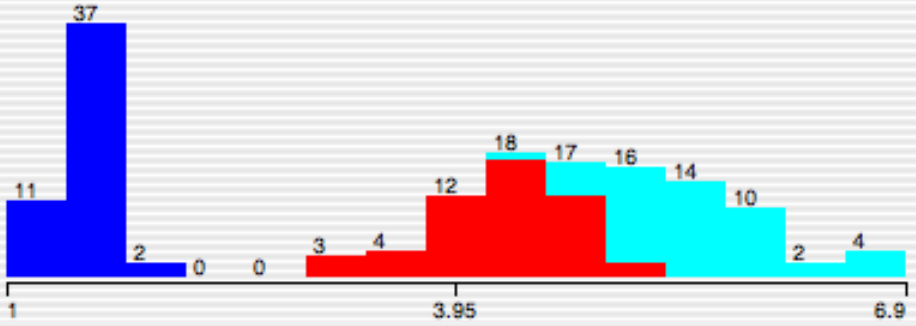
No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

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

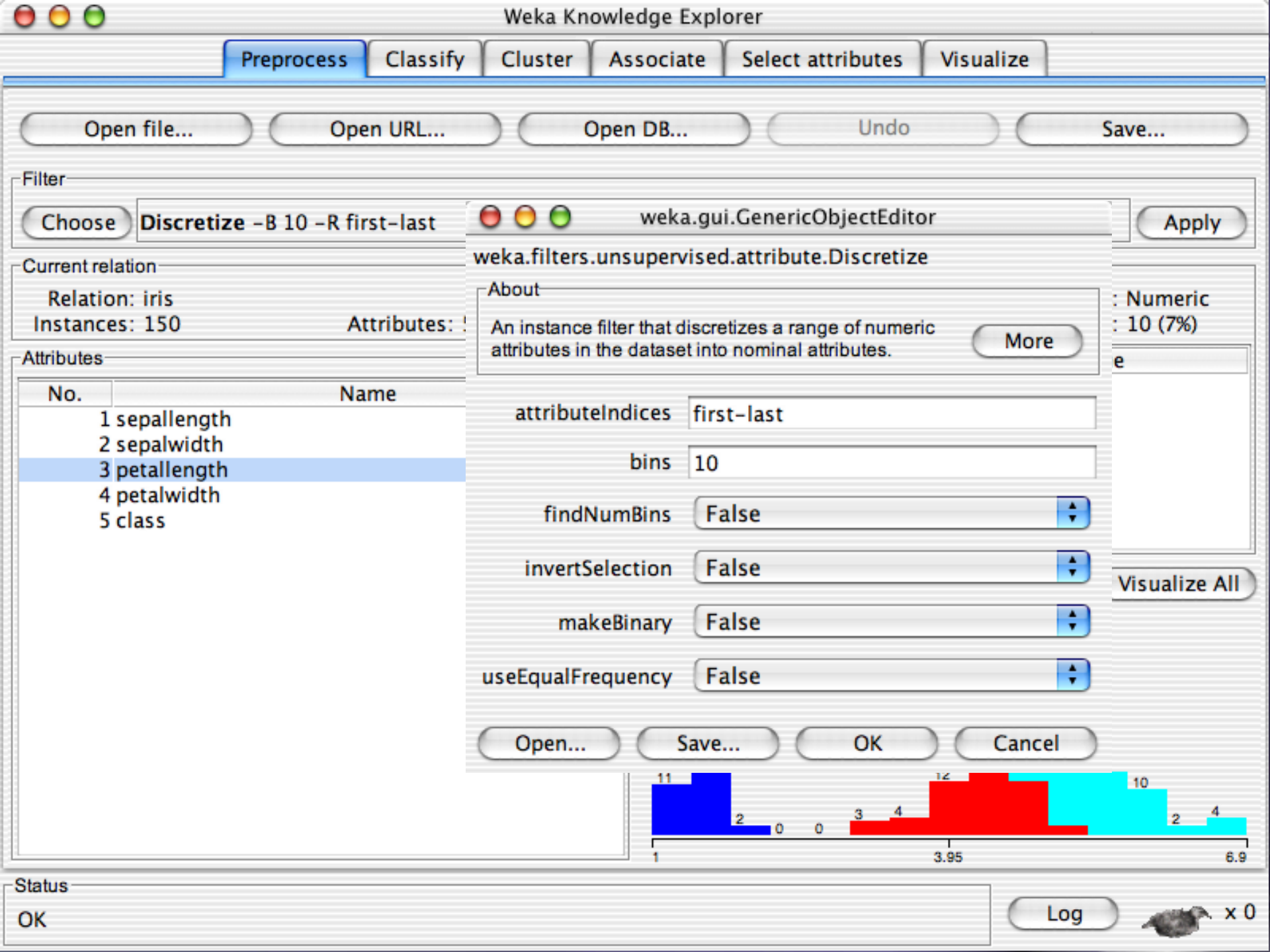
Colour: class (Nom) | Visualize All

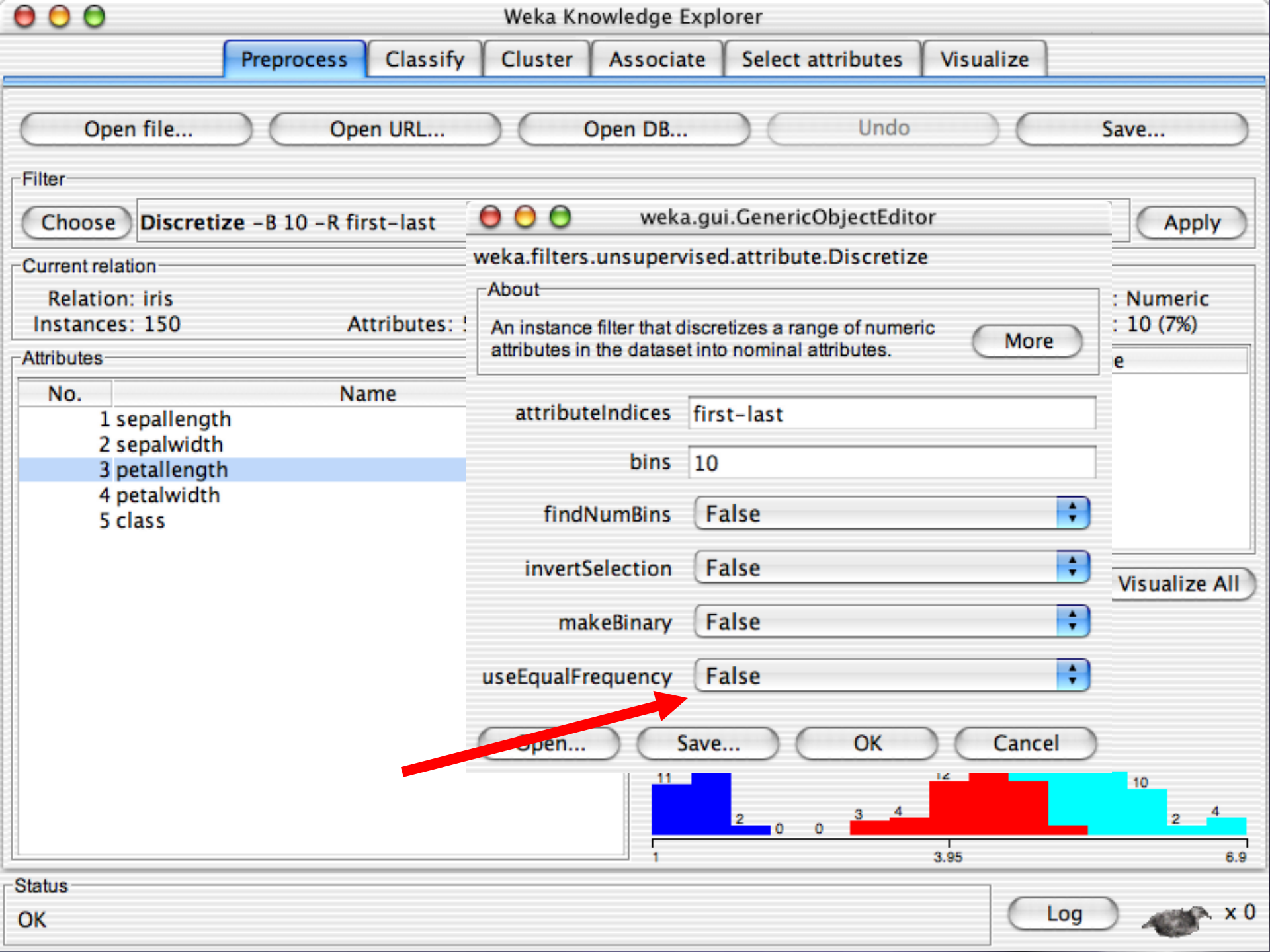


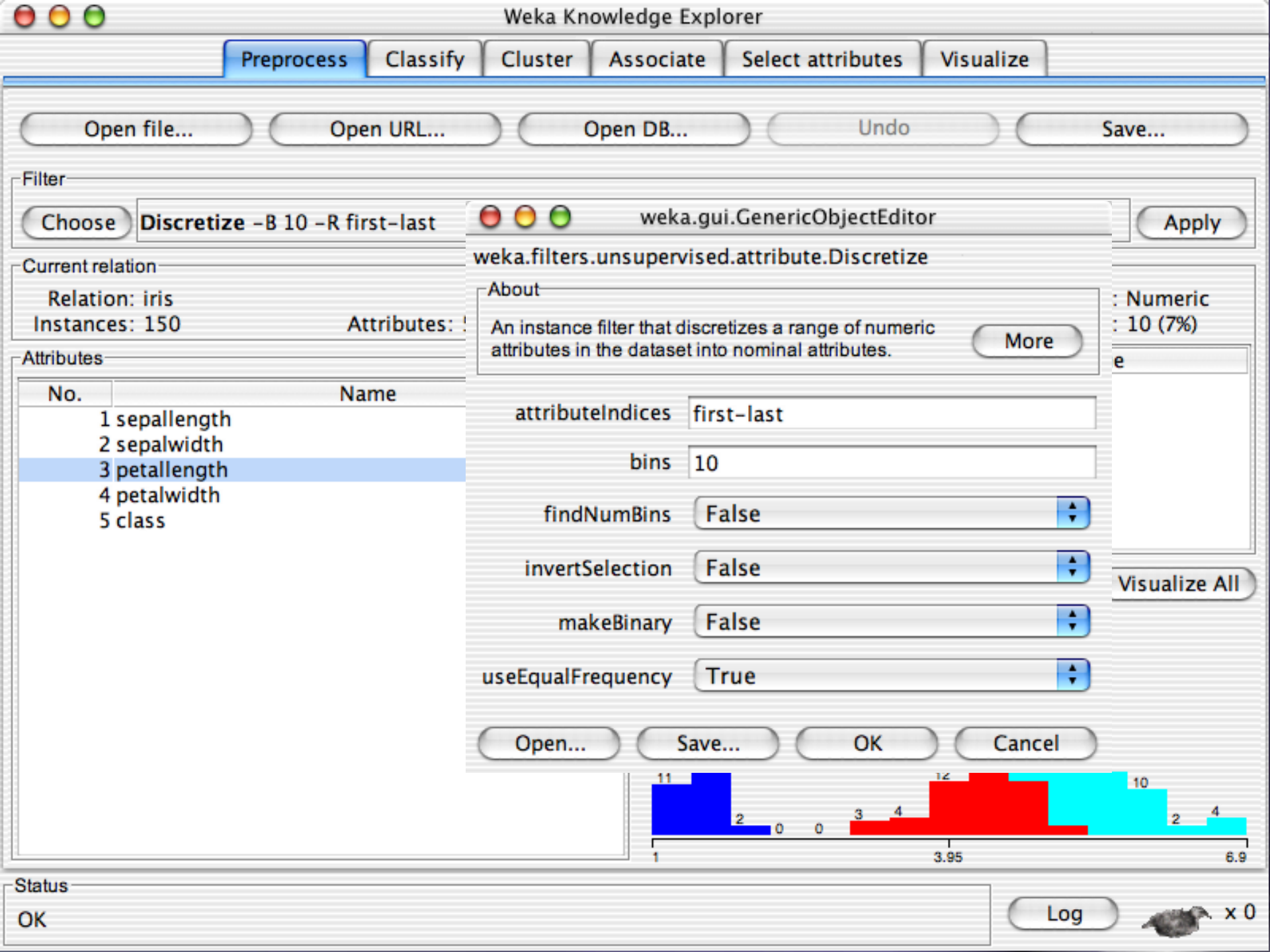
Status

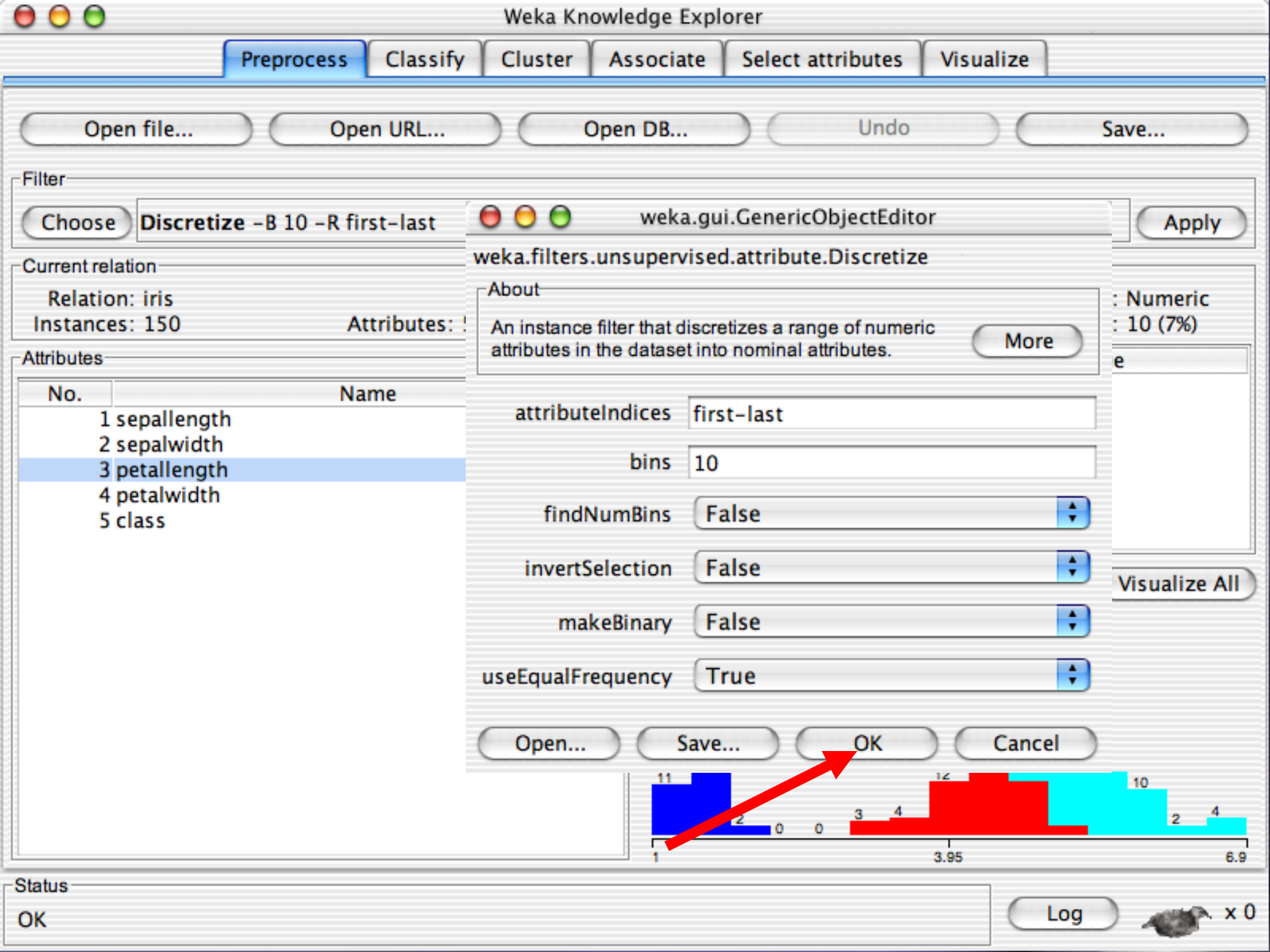
OK

Log | x 0


Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
Discretize -B 10 -R first-last
Apply
Current relation
Relation: iris
Instances: 150
Attributes:
Attributes
No.
Name
1 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.
More
attributeIndices
first-last
bins
10
findNumBins
False
invertSelection
False
makeBinary
False
useEqualFrequency
False
Open...
Save...
OK
Cancel
: Numeric
: 10 (7%)
e
Visualize All
11
2
0
0
3
4
12
10
2
4
1
3.95
6.9
Status
OK
Log
x 0


Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
Discretize -B 10 -R first-last
Apply
Current relation
Relation: iris
Instances: 150
Attributes:
Attributes
No.
Name
1 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.
More
attributeIndices
first-last
bins
10
findNumBins
False
invertSelection
False
makeBinary
False
useEqualFrequency
False
Open...
Save...
OK
Cancel
: Numeric
: 10 (7%)
Visualize All
11
2
0
0
3
4
12
10
2
4
1
3.95
6.9
Status
OK
Log
x 0


Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
Discretize -B 10 -R first-last
Apply
Current relation
Relation: iris
Instances: 150
Attributes:
Attributes
No.
Name
1 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.
More
attributeIndices
first-last
bins
10
findNumBins
False
invertSelection
False
makeBinary
False
useEqualFrequency
True
Open...
Save...
OK
Cancel
: Numeric
: 10 (7%)
e
Visualize All
11
2
0
0
3
4
12
10
2
4
1
3.95
6.9
Status
OK
Log
x 0


Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
Discretize -B 10 -R first-last
Apply
Current relation
Relation: iris
Instances: 150
Attributes:
Attributes
No.
Name
1 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.
More
attributeIndices
first-last
bins
10
findNumBins
False
invertSelection
False
makeBinary
False
useEqualFrequency
True
Open...
Save...
OK
Cancel
: Numeric
: 10 (7%)
Visualize All
11
2
0
0
3
4
12
10
2
4
1
3.95
6.9
Status
OK
Log
x 0

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

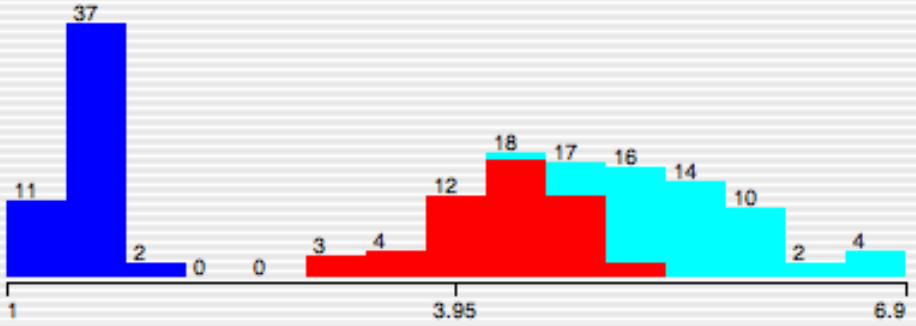
No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

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

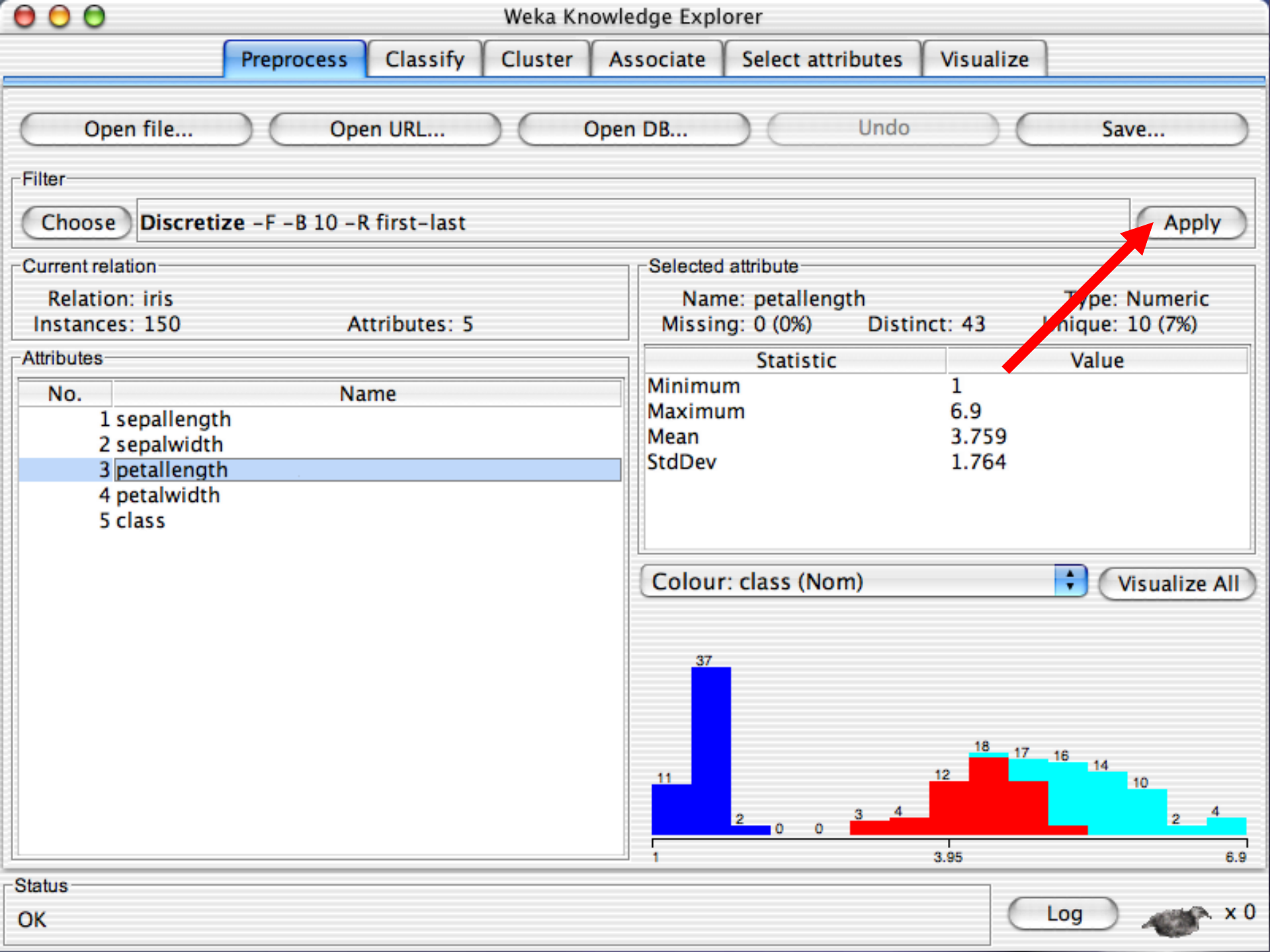
Colour: class (Nom) | Visualize All



Status

OK

Log | x 0


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
Selected attribute
Name: petallength
Type: Numeric
Missing: 0 (0%)
Distinct: 43
Unique: 10 (7%)
Attributes
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Statistic
Value
Minimum
1
Maximum
6.9
Mean
3.759
StdDev
1.764
Colour: class (Nom)
Visualize All
37
11
2
0
0
3
4
12
18
17
16
14
10
2
4
1
3.95
6.9
Status
OK
Log
x 0

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

Attributes

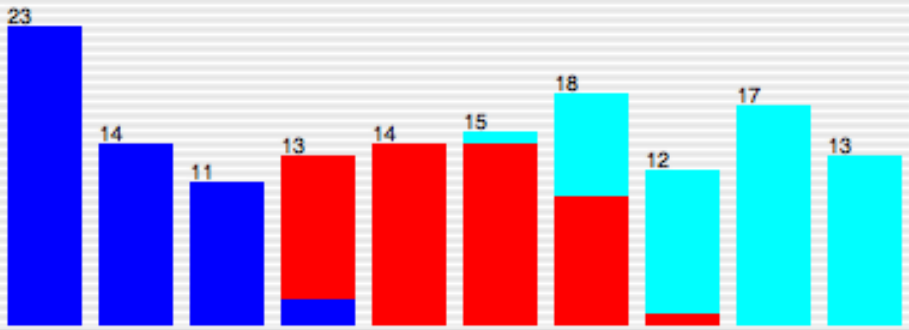
No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

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) | Visualize All



Status

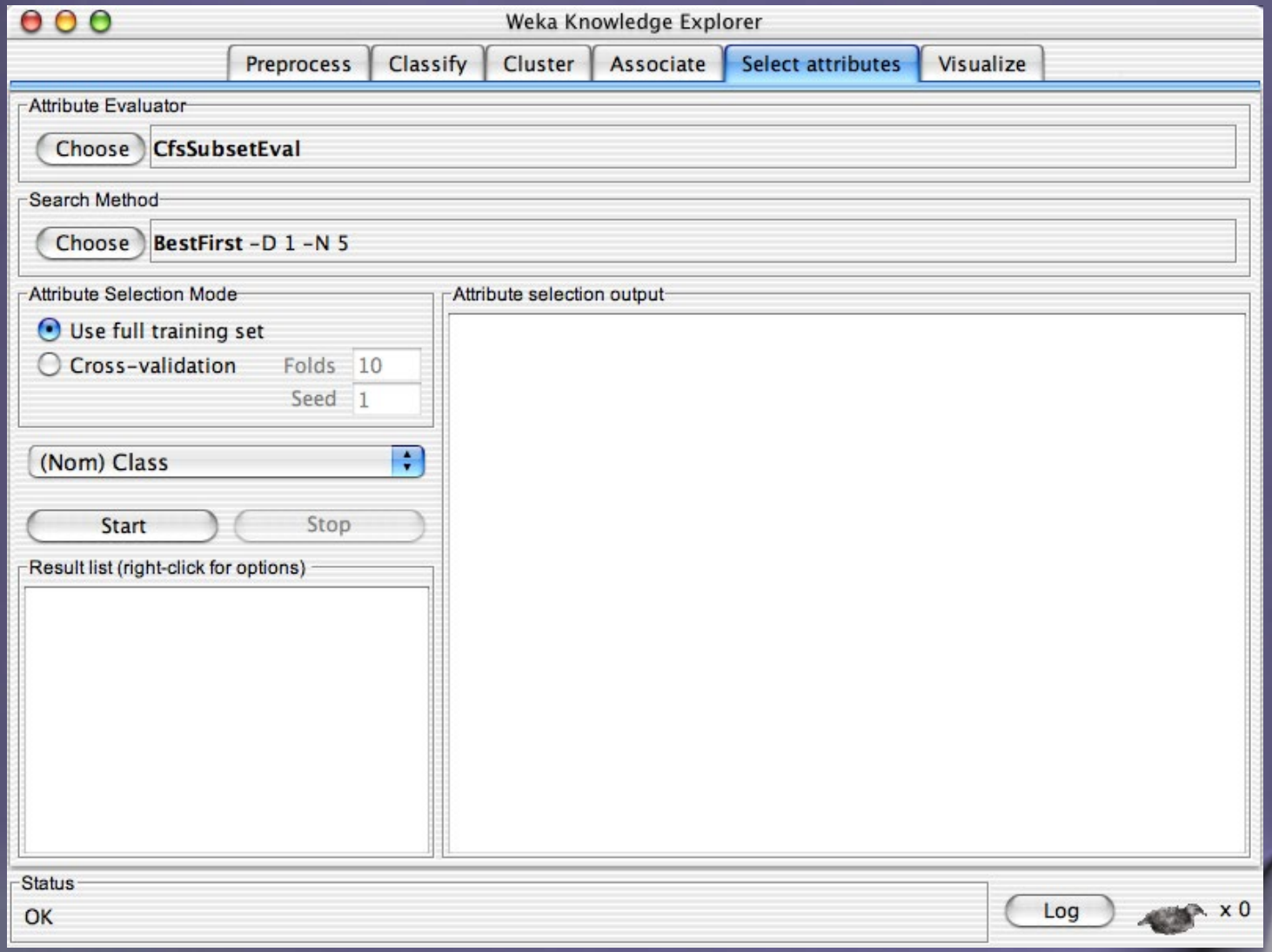
OK

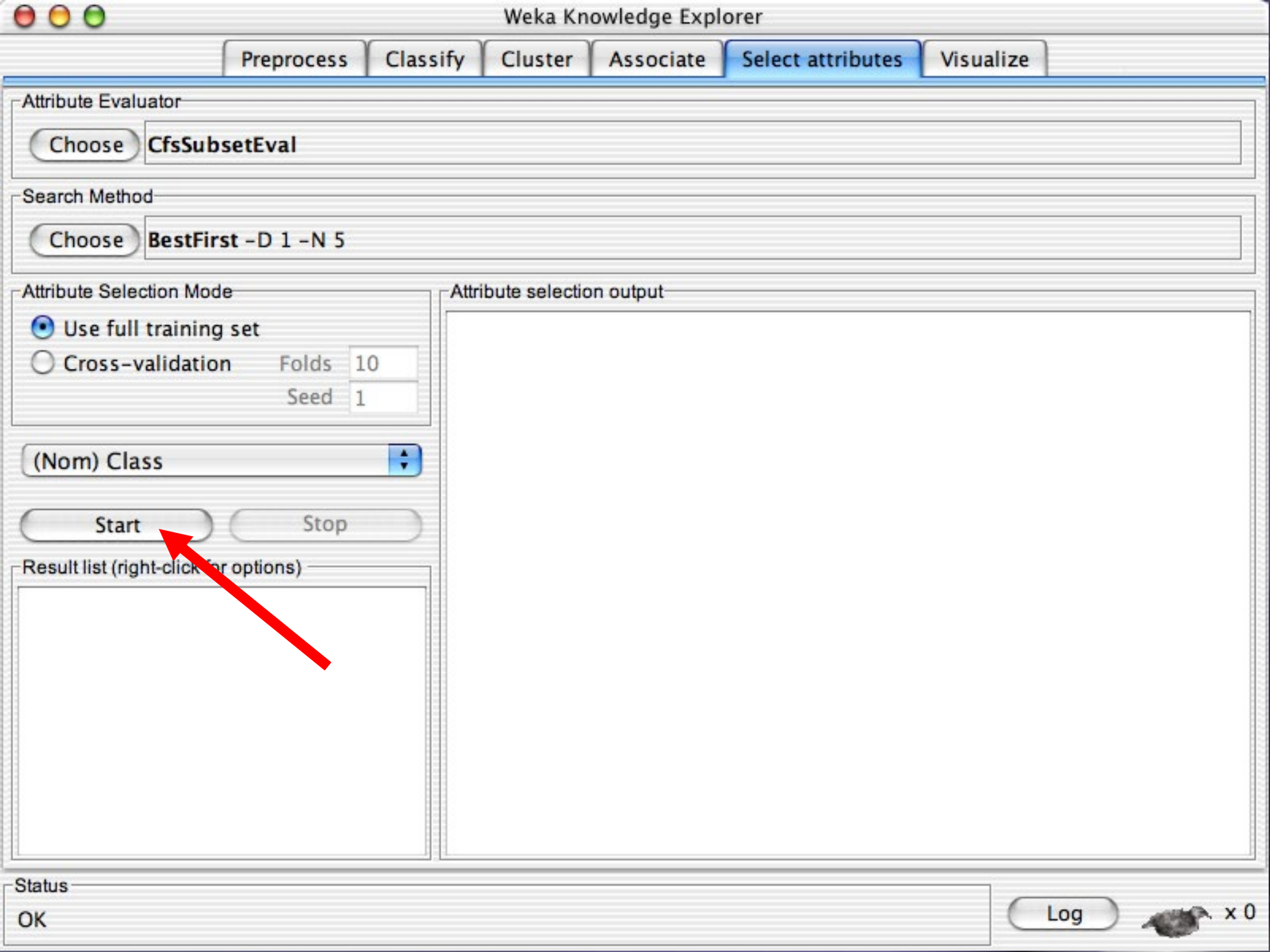
Log | x 0

Select attributes Panel

- It 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

Select attributes Panel




Weka Knowledge Explorer
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
Attribute selection output
(Nom) Class
Start
Stop
Result list (right-click for options)
Status
OK
Log
x 0

Weka Knowledge Explorer

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)

16:39:40 - BestFirst + CfsSubsetEval

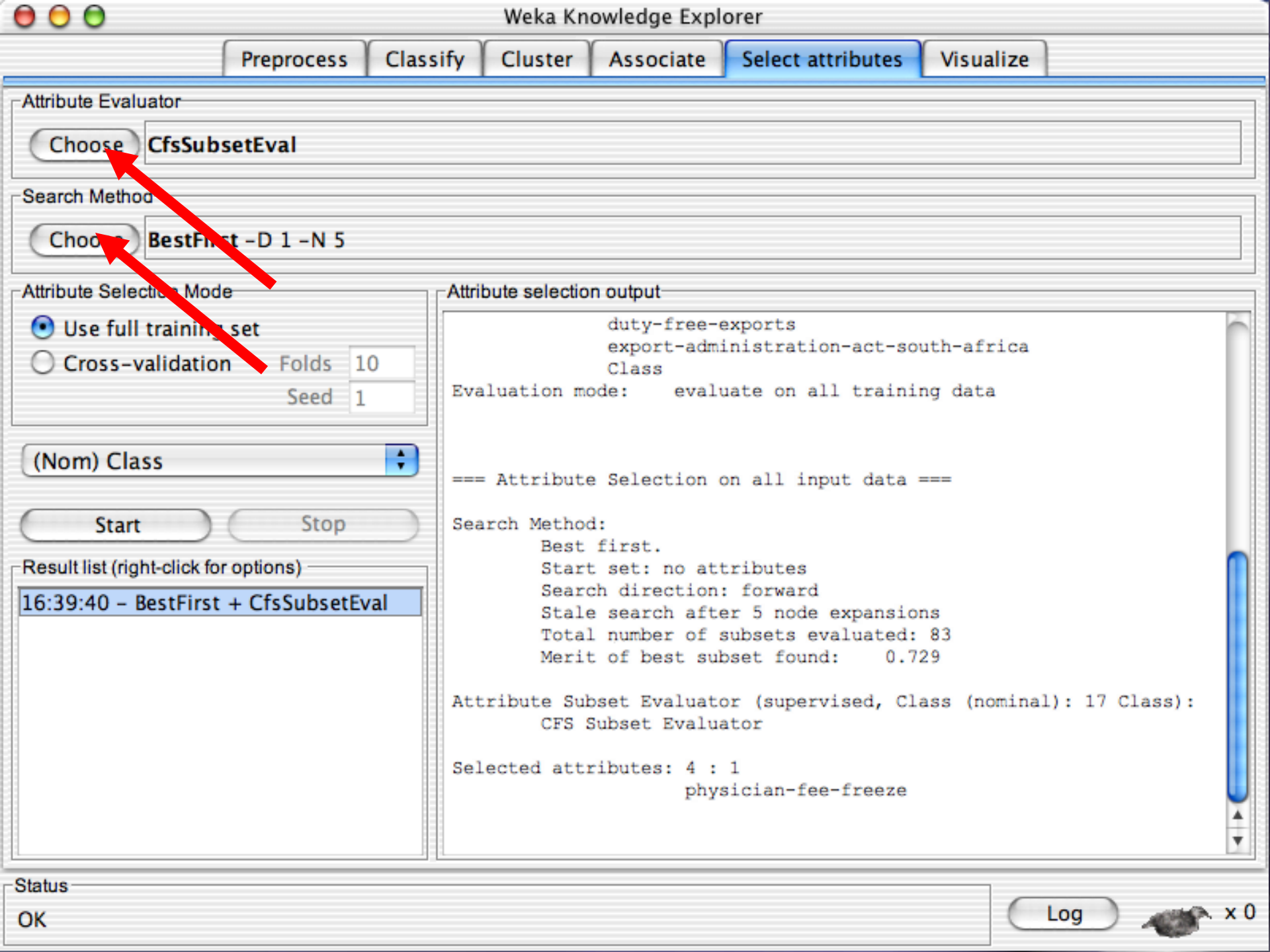
Attribute selection output

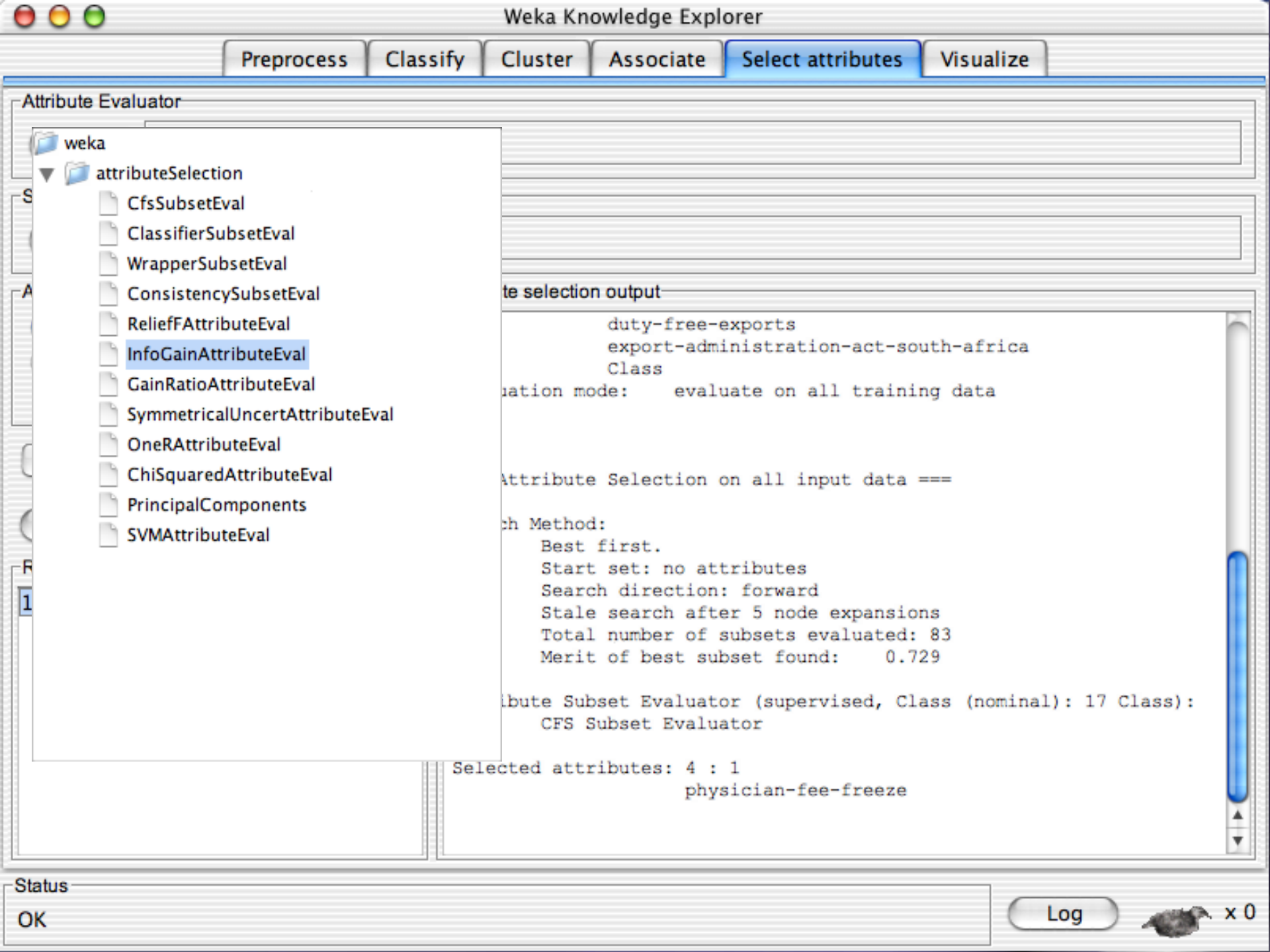
```
                duty-free-exports
                export-administration-act-south-africa
                Class
Evaluation mode:    evaluate on all training data



=== 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
```

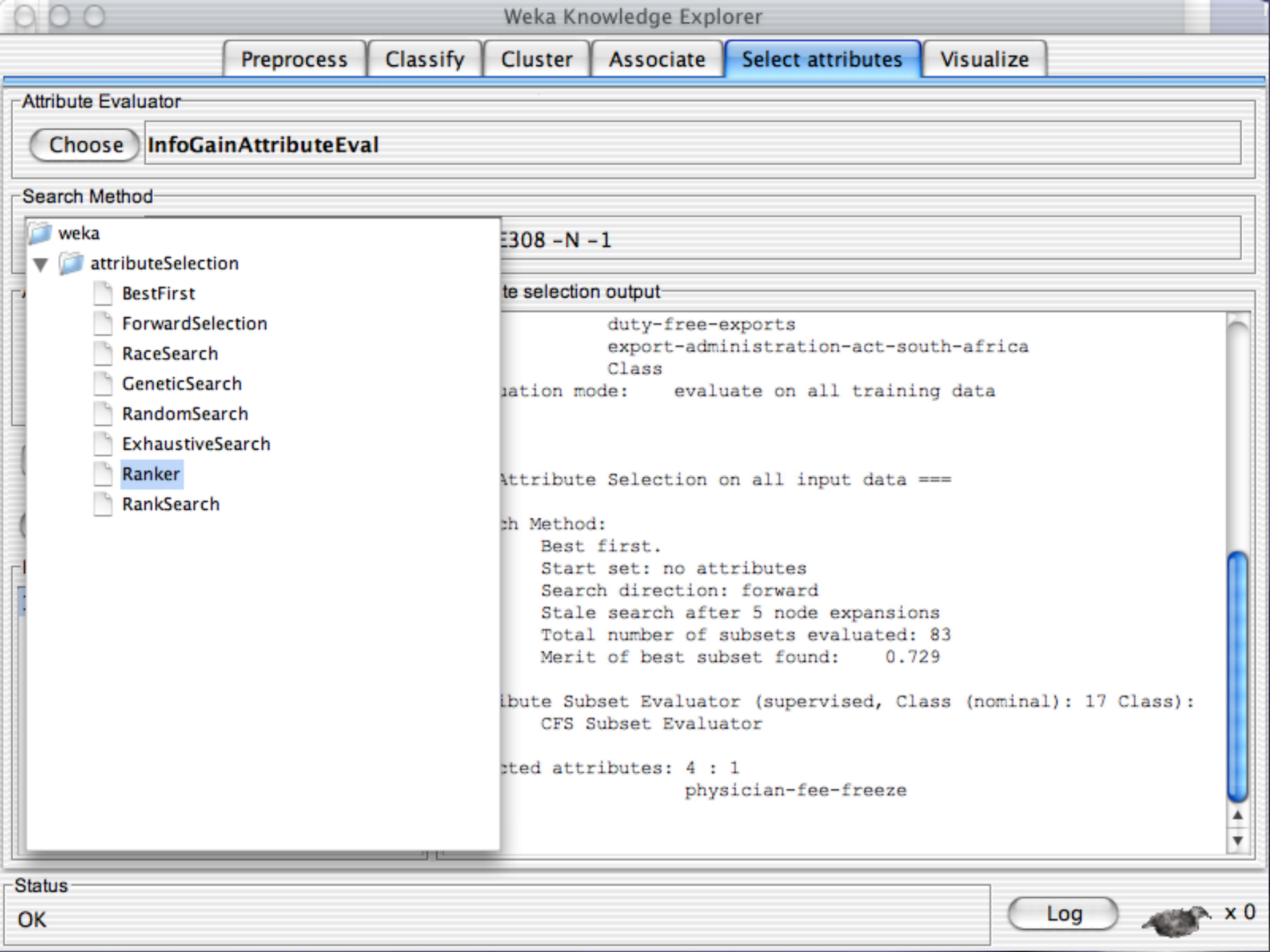
Status

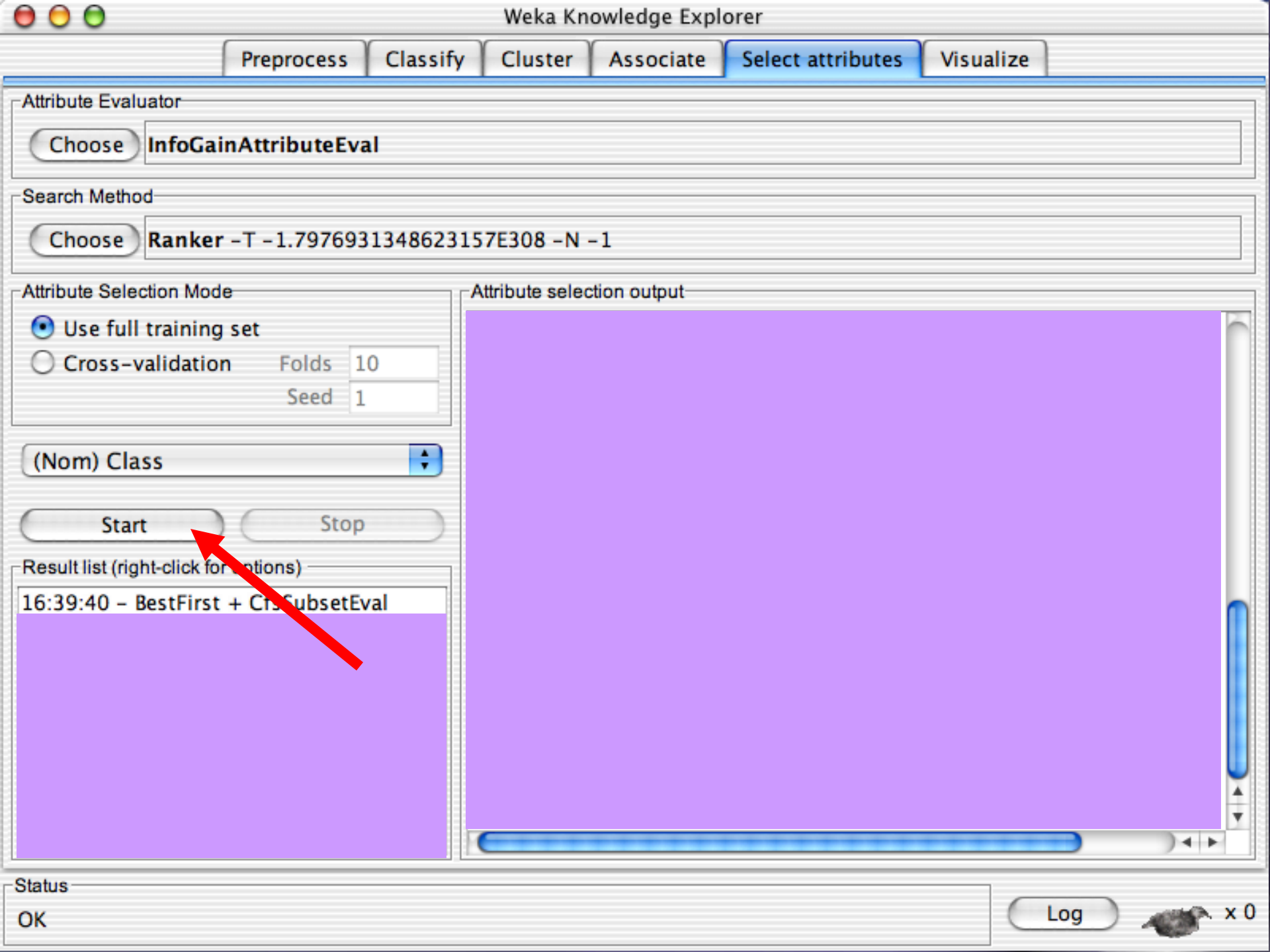
OK

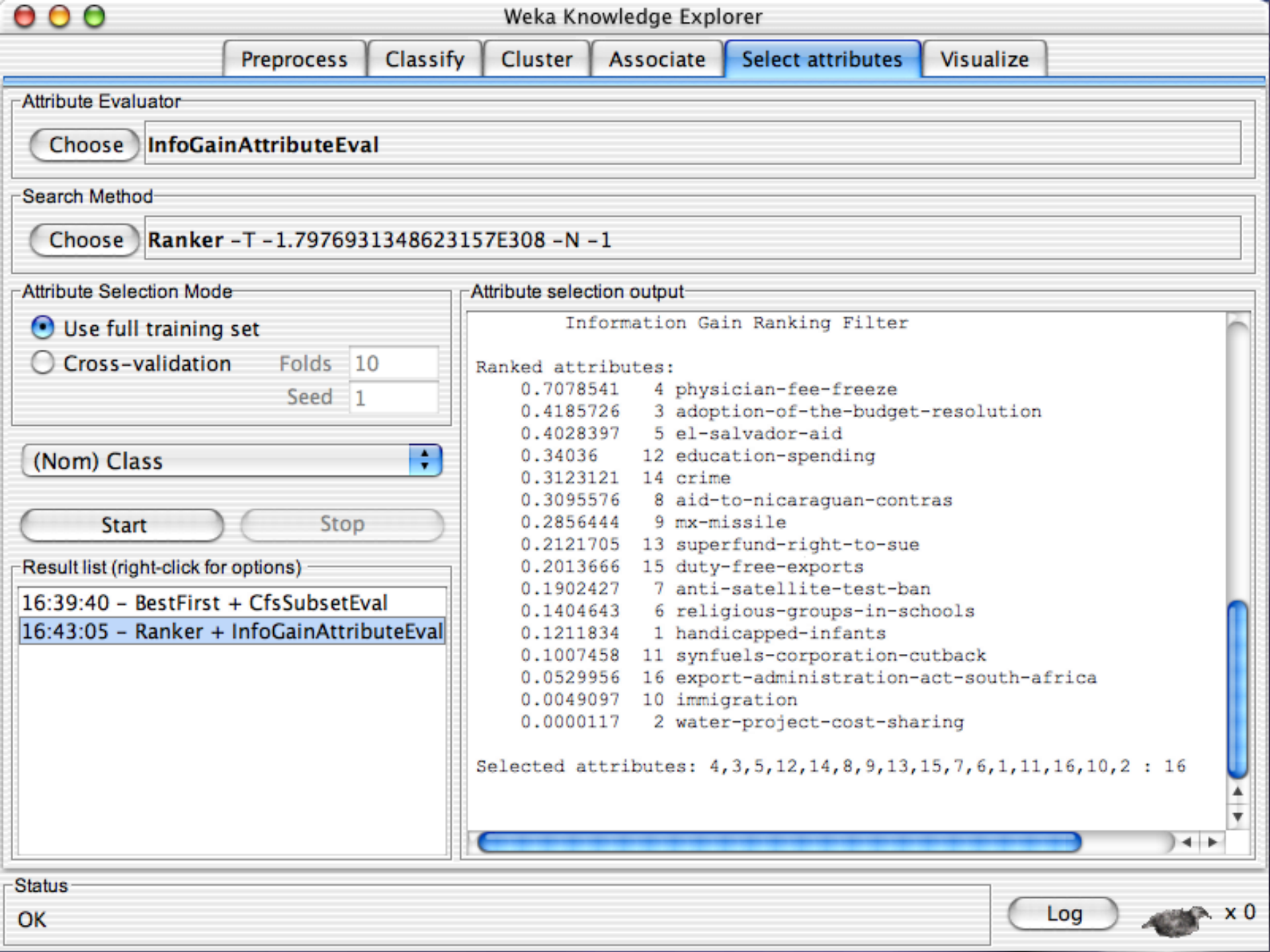
Log x 0


Weka Knowledge Explorer
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)
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


Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Attribute Evaluator
weka
attributeSelection
CfsSubsetEval
ClassifierSubsetEval
WrapperSubsetEval
ConsistencySubsetEval
ReliefFAttributeEval
InfoGainAttributeEval
GainRatioAttributeEval
SymmetricalUncertAttributeEval
OneRAttributeEval
ChiSquaredAttributeEval
PrincipalComponents
SVMAttributeEval
te selection output
duty-free-exports
export-administration-act-south-africa
Class
uation mode: evaluate on all training data
Attribute Selection on all input data ===
ch Method:
Best first.
Start set: no attributes
Search direction: forward
Stale search after 5 node expansions
Total number of subsets evaluated: 83
Merit of best subset found: 0.729
ibute Subset Evaluator (supervised, Class (nominal): 17 Class):
CFS Subset Evaluator
Selected attributes: 4 : 1
physician-fee-freeze
Status
OK
Log
x 0


Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Attribute Evaluator
Choose
InfoGainAttributeEval
Search Method
E308 -N -1
te selection output
weka
attributeSelection
BestFirst
ForwardSelection
RaceSearch
GeneticSearch
RandomSearch
ExhaustiveSearch
Ranker
RankSearch
duty-free-exports
export-administration-act-south-africa
Class
uation mode: evaluate on all training data
Attribute Selection on all input data ===
ch Method:
Best first.
Start set: no attributes
Search direction: forward
Stale search after 5 node expansions
Total number of subsets evaluated: 83
Merit of best subset found: 0.729
ibute Subset Evaluator (supervised, Class (nominal): 17 Class):
CFS Subset Evaluator
cted attributes: 4 : 1
physician-fee-freeze
Status
OK
Log
x 0


Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Attribute Evaluator
Choose
InfoGainAttributeEval
Search Method
Choose
Ranker -T -1.7976931348623157E308 -N -1
Attribute Selection Mode
Use full training set
Cross-validation
Folds
10
Seed
1
Attribute selection output
(Nom) Class
Start
Stop
Result list (right-click for options)
16:39:40 - BestFirst + CfsSubsetEval
Status
OK
Log
x 0


Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Attribute Evaluator
Choose
InfoGainAttributeEval
Search Method
Choose
Ranker -T -1.7976931348623157E308 -N -1
Attribute Selection Mode
Use full training set
Cross-validation
Folds
10
Seed
1
(Nom) Class
Start
Stop
Result list (right-click for options)
16:39:40 - BestFirst + CfsSubsetEval
16:43:05 - Ranker + InfoGainAttributeEval
Attribute selection output
Information Gain Ranking Filter
Ranked attributes:
0.7078541 4 physician-fee-freeze
0.4185726 3 adoption-of-the-budget-resolution
0.4028397 5 el-salvador-aid
0.34036 12 education-spending
0.3123121 14 crime
0.3095576 8 aid-to-nicaraguan-contras
0.2856444 9 mx-missile
0.2121705 13 superfund-right-to-sue
0.2013666 15 duty-free-exports
0.1902427 7 anti-satellite-test-ban
0.1404643 6 religious-groups-in-schools
0.1211834 1 handicapped-infants
0.1007458 11 synfuels-corporation-cutback
0.0529956 16 export-administration-act-south-africa
0.0049097 10 immigration
0.0000117 2 water-project-cost-sharing
Selected attributes: 4,3,5,12,14,8,9,13,15,7,6,1,11,16,10,2 : 16
Status
OK
Log
x 0