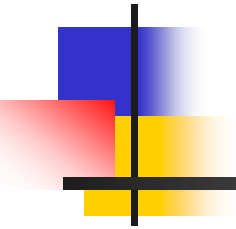


Machine Learning with WEKA
An Introduction

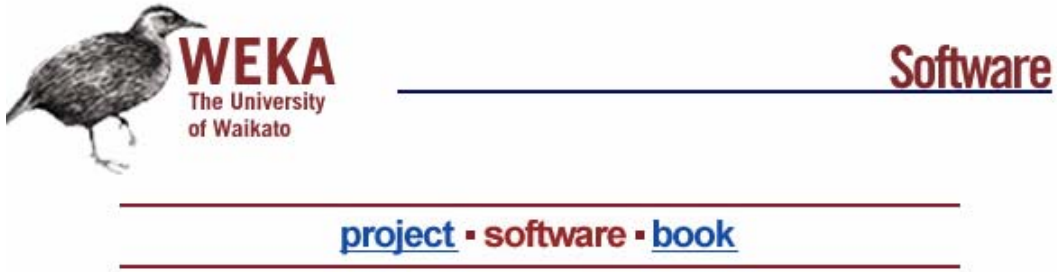
林 松 江

2005/9/30

WEKA: the bird

WEKA is available at

- http://www.cs.waikato.ac.nz/ml/weka/



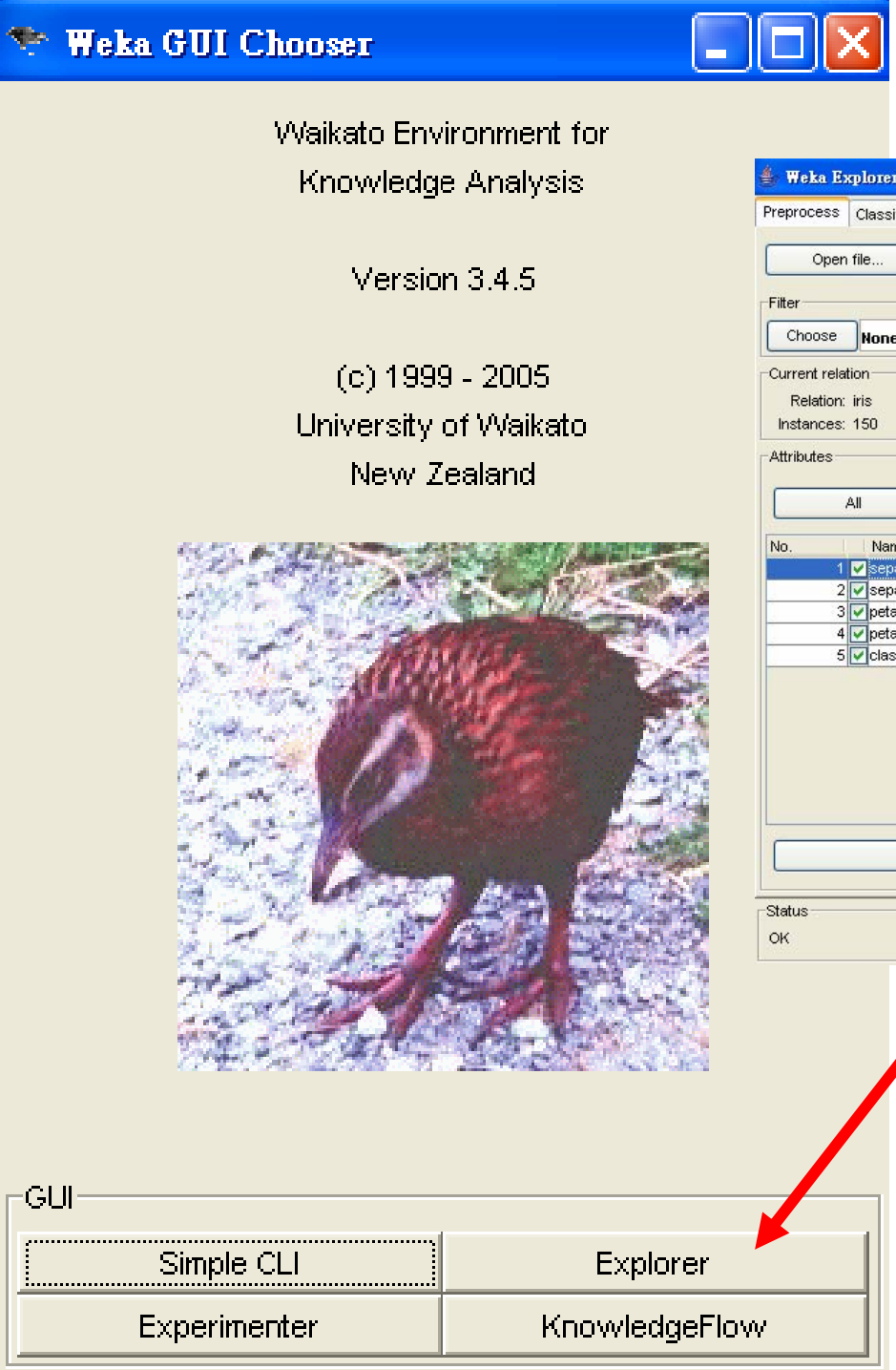
The format of Dataset in WEKA(1)

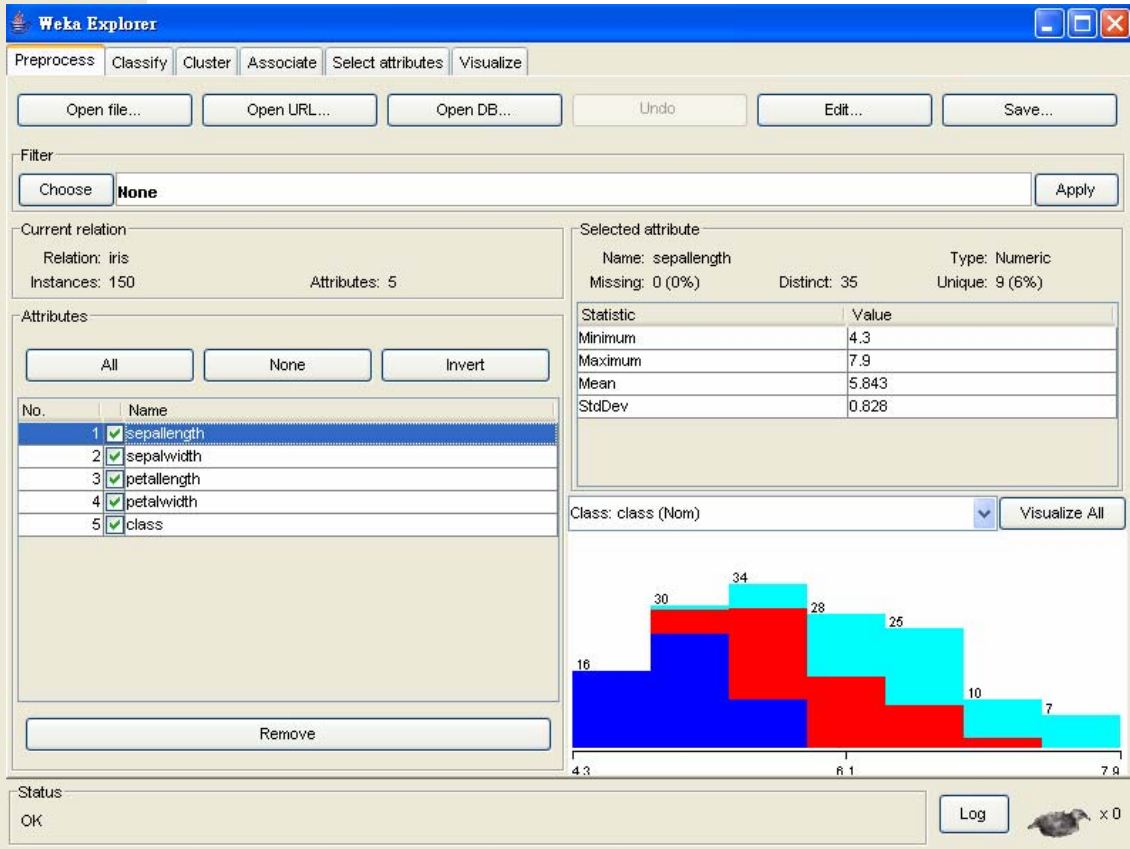
```
@relation heart-disease-simplified

@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
...
```

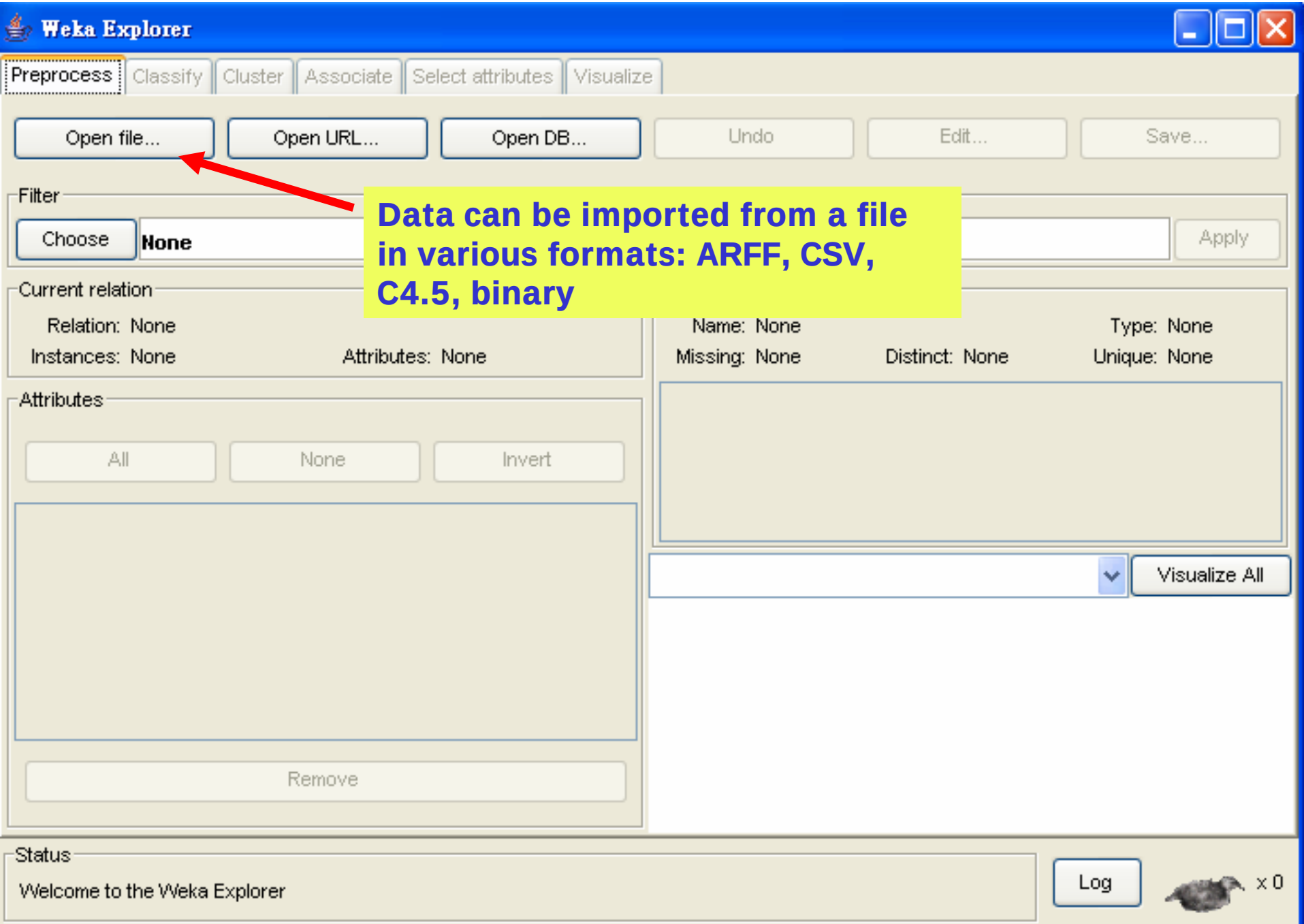
Flat file in ARFF format

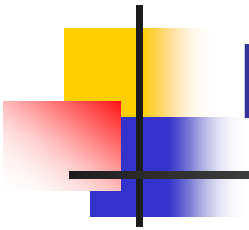
The format of Dataset in WEKA(2)


```
@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
...
```


Weka GUI Chooser
Waikato Environment for
Knowledge Analysis
Version 3.4.5
(c) 1999 - 2005
University of Waikato
New Zealand
GUI
Simple CLI
Explorer
Experimenter
KnowledgeFlow

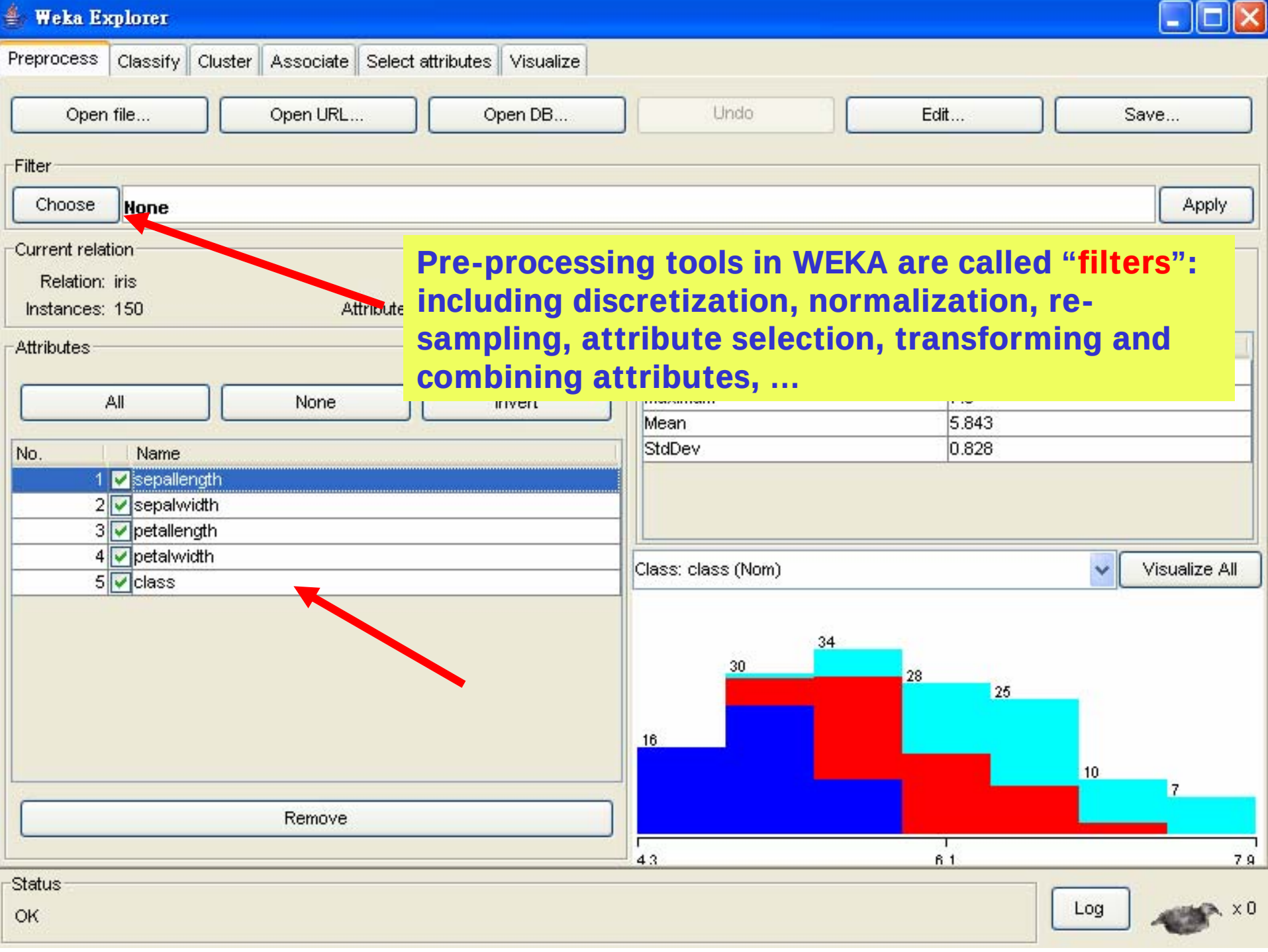

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: iris
Instances: 150
Attributes: 5
Selected attribute
Name: sepallength
Type: Numeric
Missing: 0 (0%)
Distinct: 35
Unique: 9 (6%)
Statistic
Value
Minimum
4.3
Maximum
7.9
Mean
5.843
StdDev
0.828
Attributes
All
None
Invert
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Class: class (Nom)
Visualize All
34
30
28
25
16
10
7
4.3
6.1
7.9
Remove
Status
OK
Log
x 0

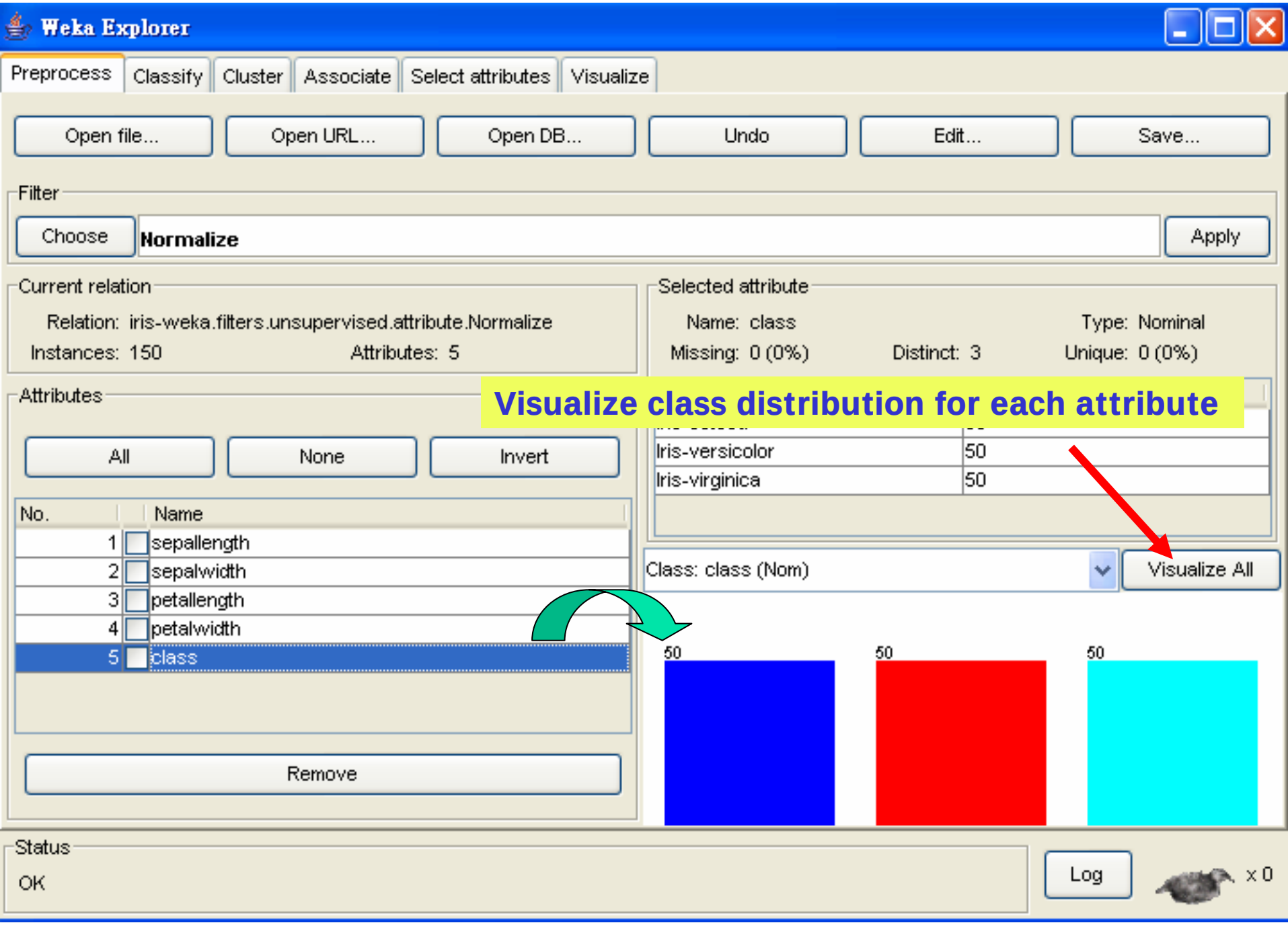

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: None
Instances: None
Attributes: None
Name: None
Type: None
Missing: None
Distinct: None
Unique: None
Attributes
All
None
Invert
Remove
Visualize All
Status
Welcome to the Weka Explorer
Log
x 0
Data can be imported from a file in various formats: ARFF, CSV, C4.5, binary

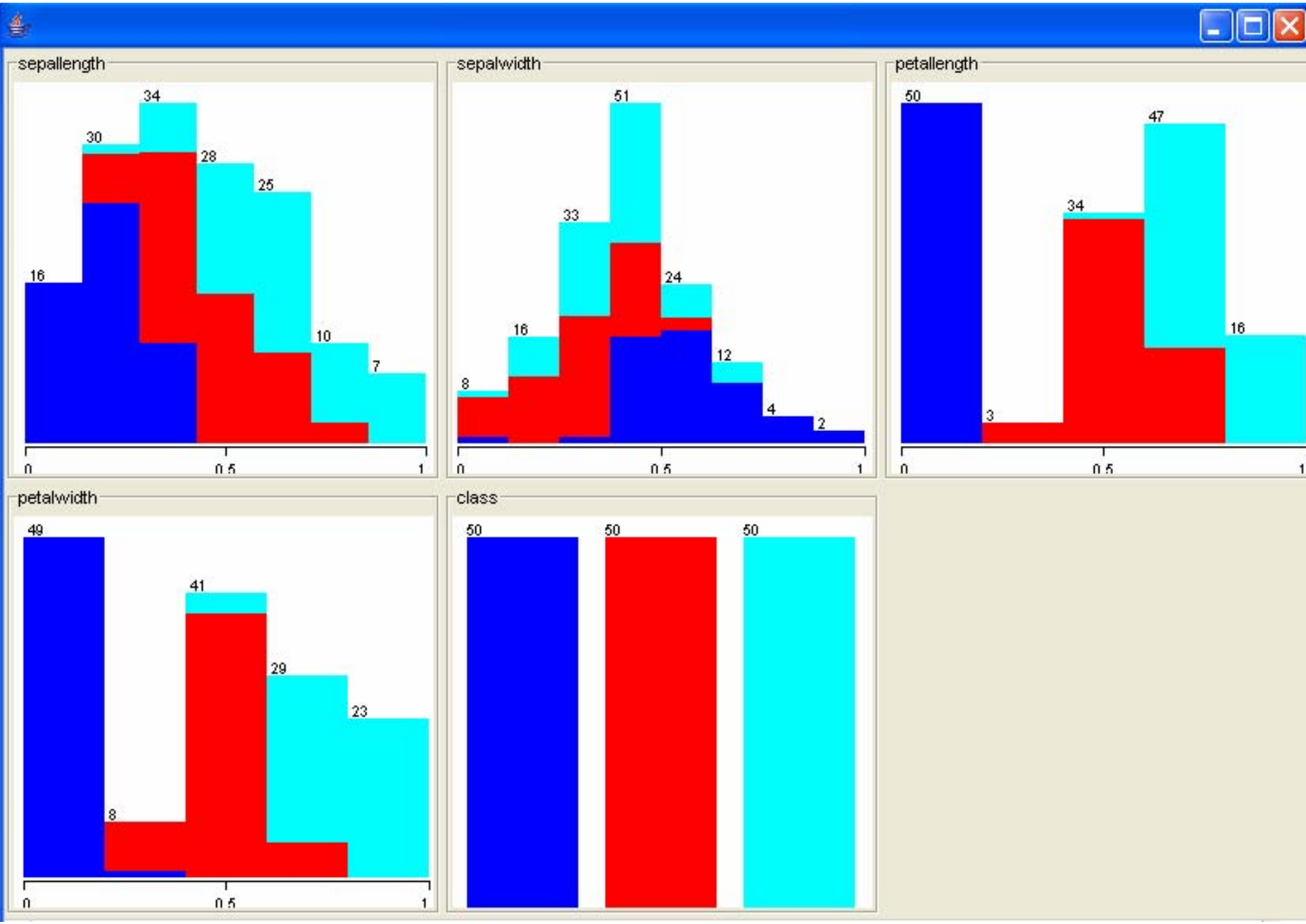


Explorer: pre-processing the data

- 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 Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: iris
Instances: 150
Attribute
Pre-processing tools in WEKA are called "filters": including discretization, normalization, re-sampling, attribute selection, transforming and combining attributes, ...
Attributes
All
None
Invert
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Remove
Mean 5.843
StdDev 0.828
Class: class (Nom)
Visualize All
34
30
28
25
16
10
7
4.3
6.1
7.9
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Edit...
Save...
Filter
Choose
Normalize
Apply
Current relation
Relation: iris-weka.filters.unsupervised.attribute.Normalize
Instances: 150
Attributes: 5
Selected attribute
Name: class
Type: Nominal
Missing: 0 (0%)
Distinct: 3
Unique: 0 (0%)
Visualize class distribution for each attribute
Attributes
All
None
Invert
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Iris-versicolor 50
Iris-virginica 50
Class: class (Nom)
Visualize All
50
50
50
Remove
Status
OK
Log
x 0


sepallength
34
30
28
25
16
10
7
0
0.5
1
sepalwidth
51
33
24
16
12
8
4
2
0
0.5
1
petallength
50
47
34
16
3
0
0.5
1
petalwidth
49
41
29
23
8
0
0.5
1
class
50
50
50

Weka Explorer

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

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

Filter

Choose **None** | Apply

Current relation

Relation: iris

Instances: 150 Attributes: 5

Attributes

All | None | Invert

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

Remove

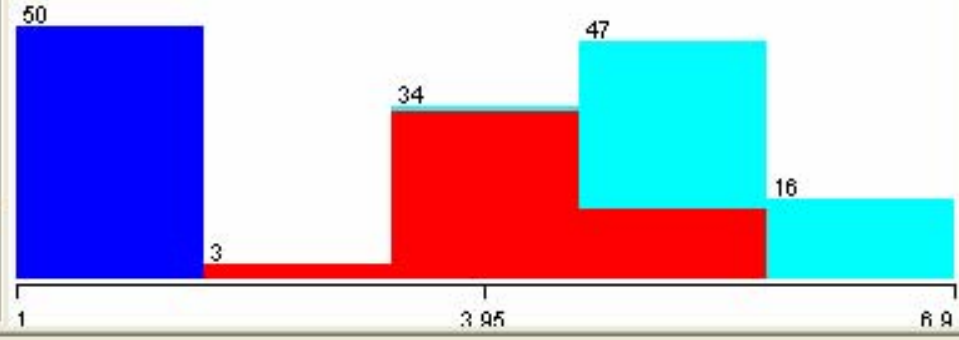
Selected attribute

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

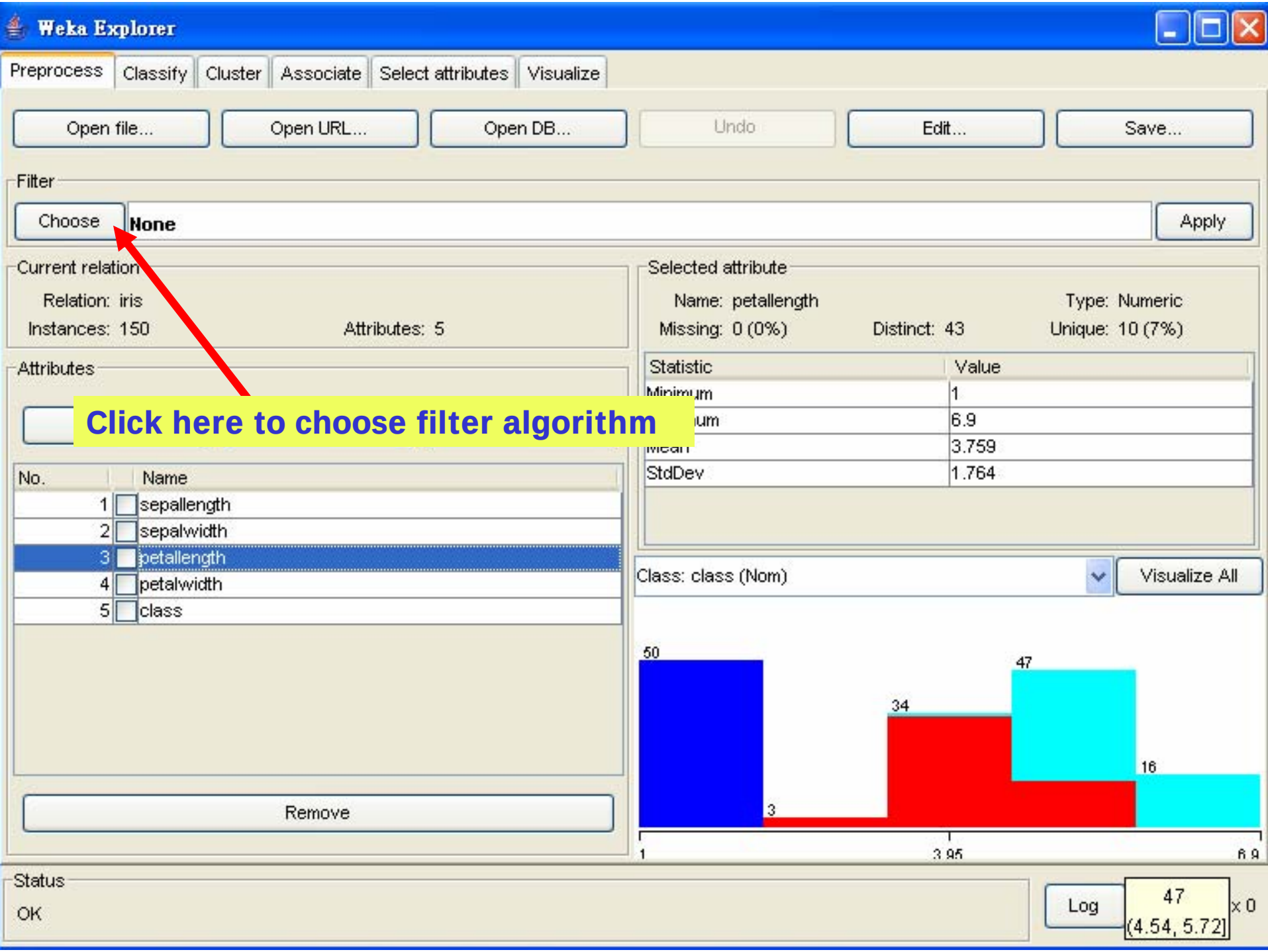
Class: class (Nom) | Visualize All

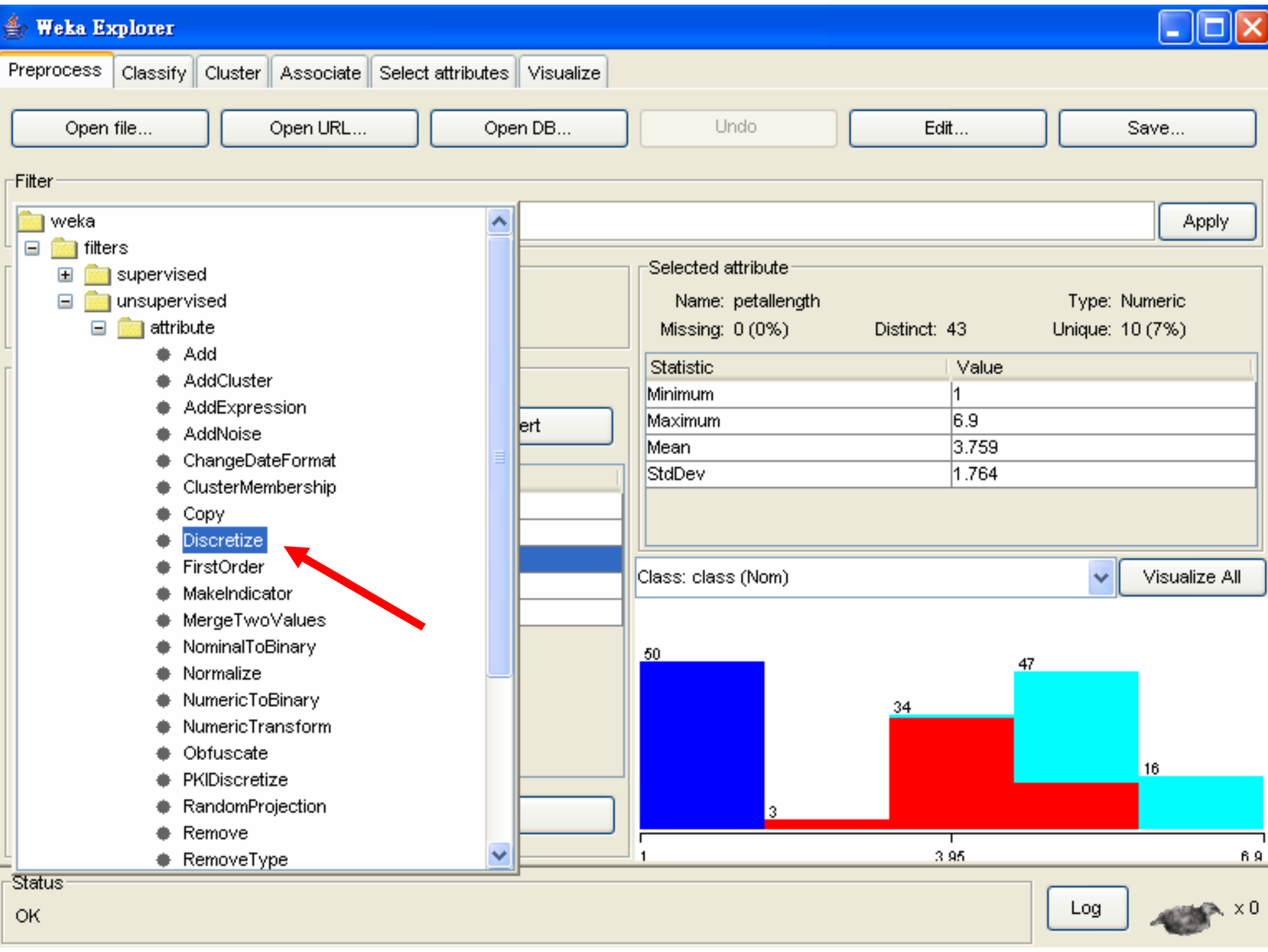


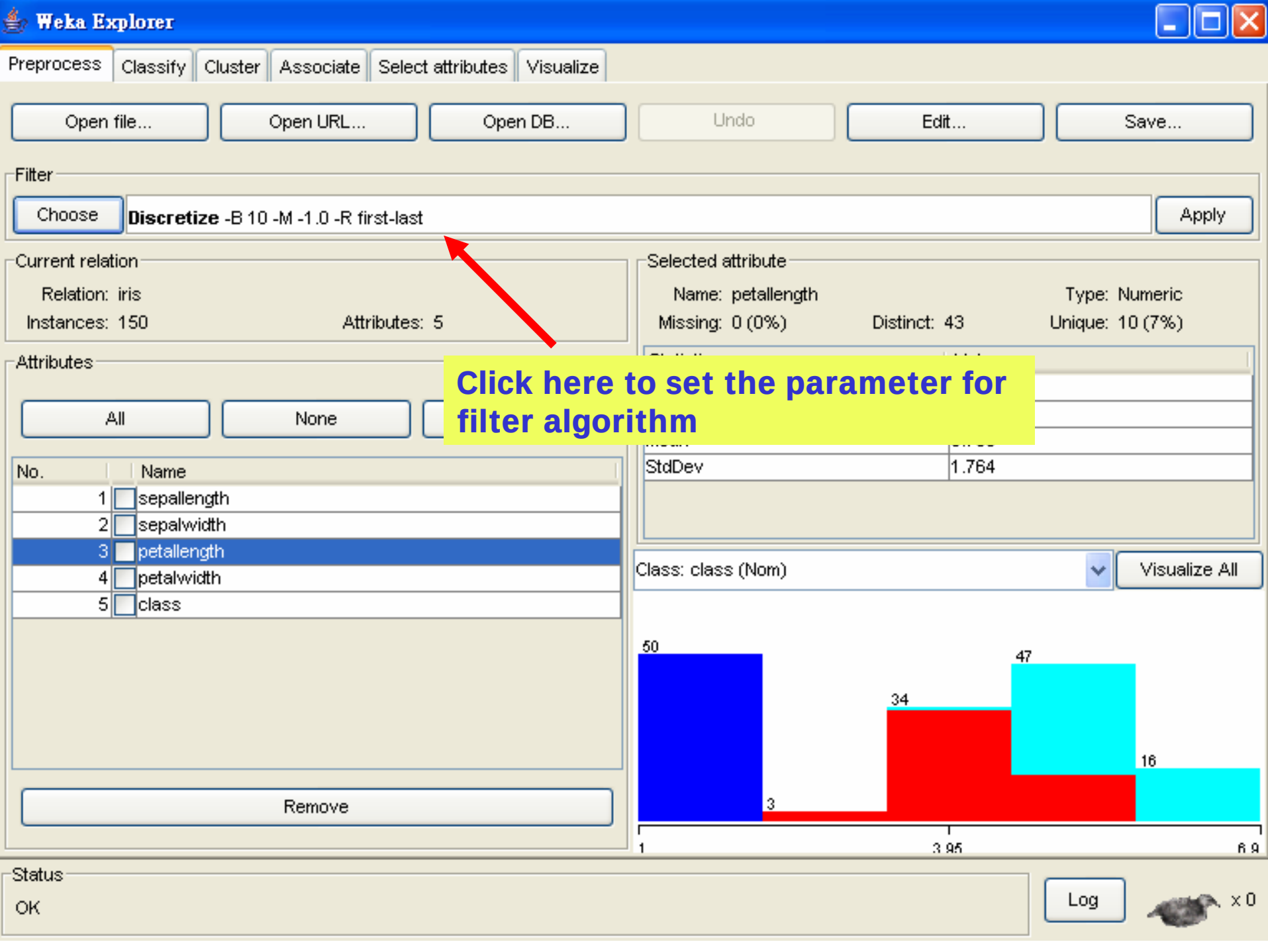
Status

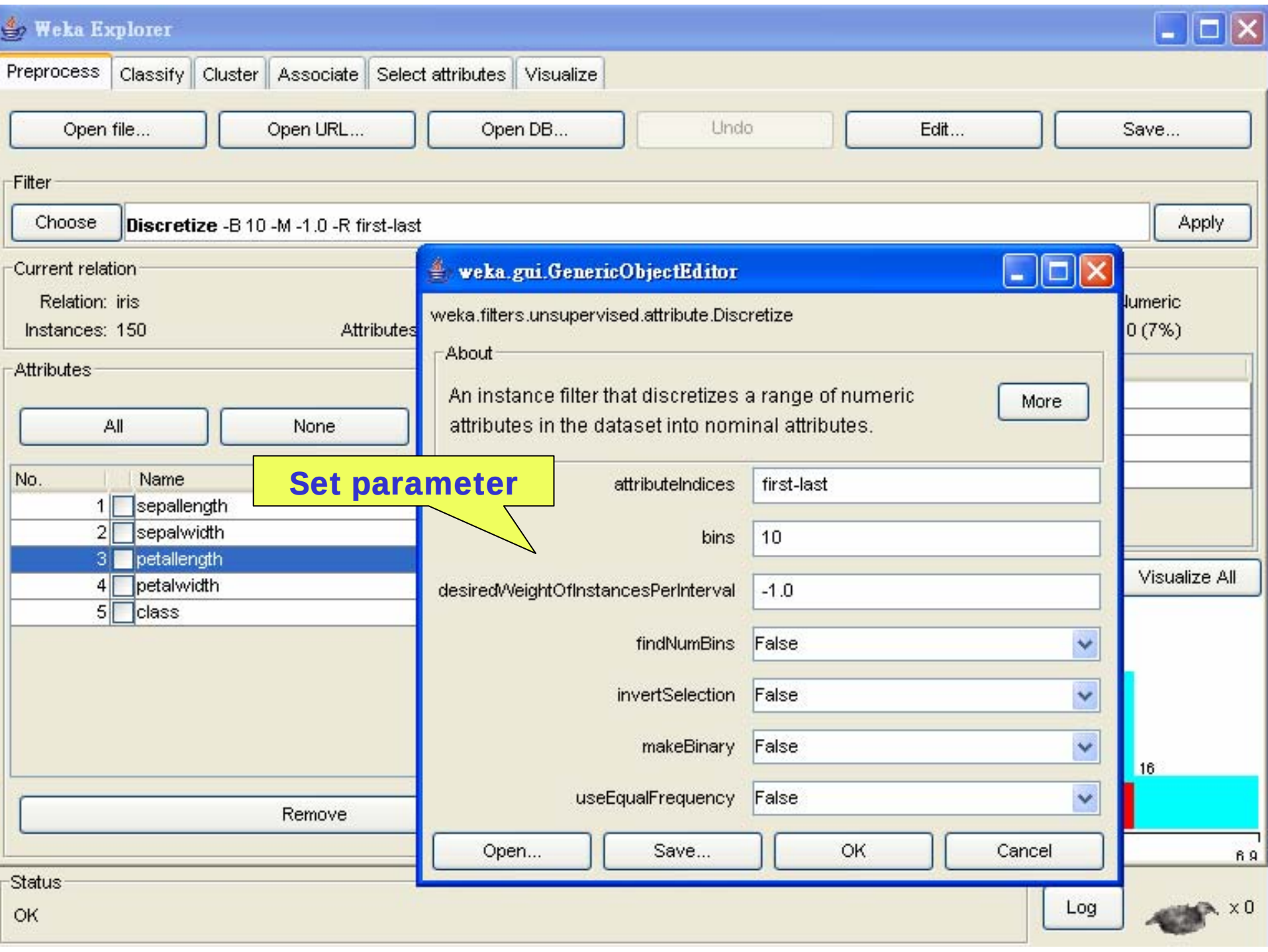
OK

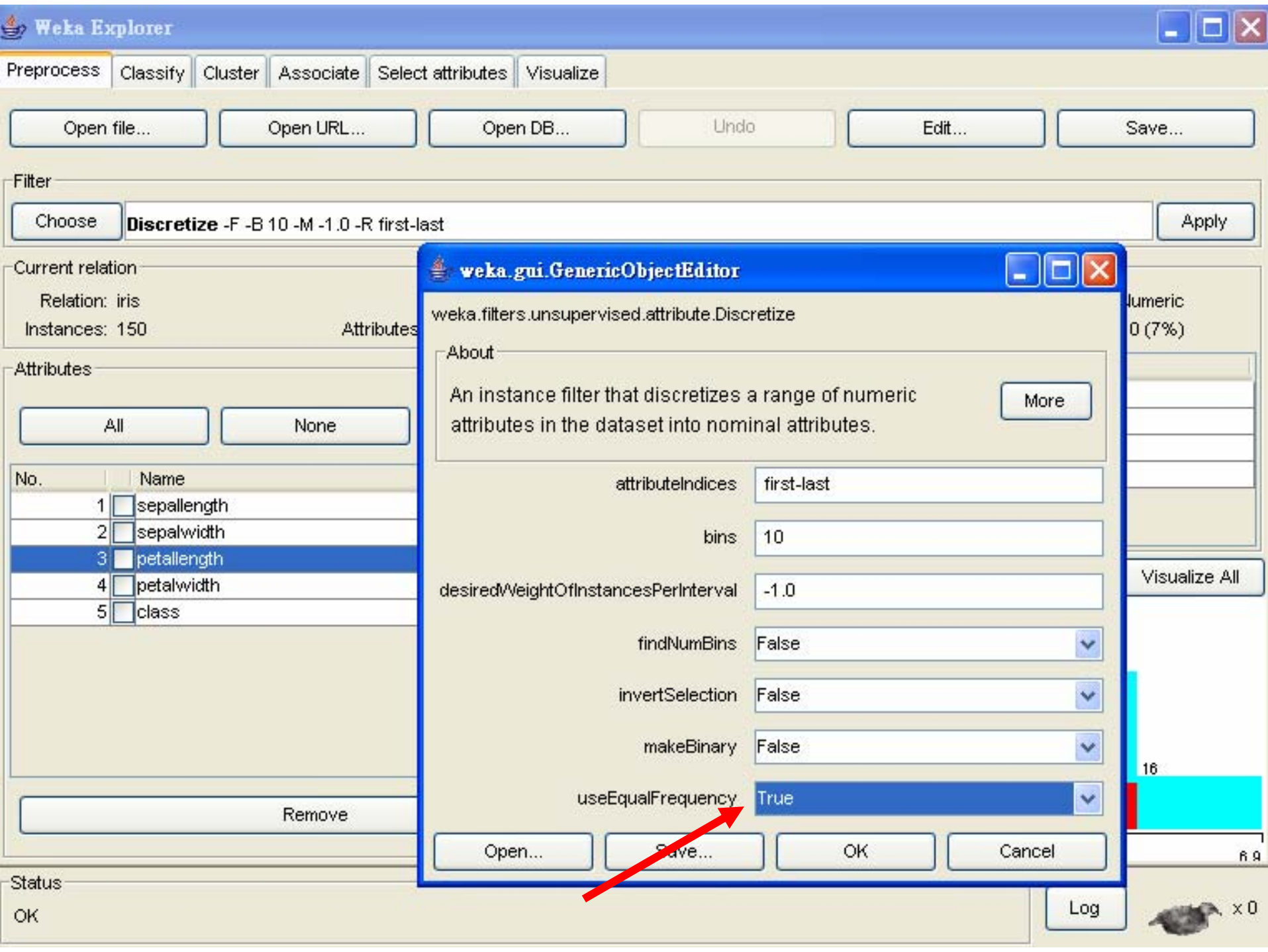
Log | 47 (4.54, 5.72) x 0

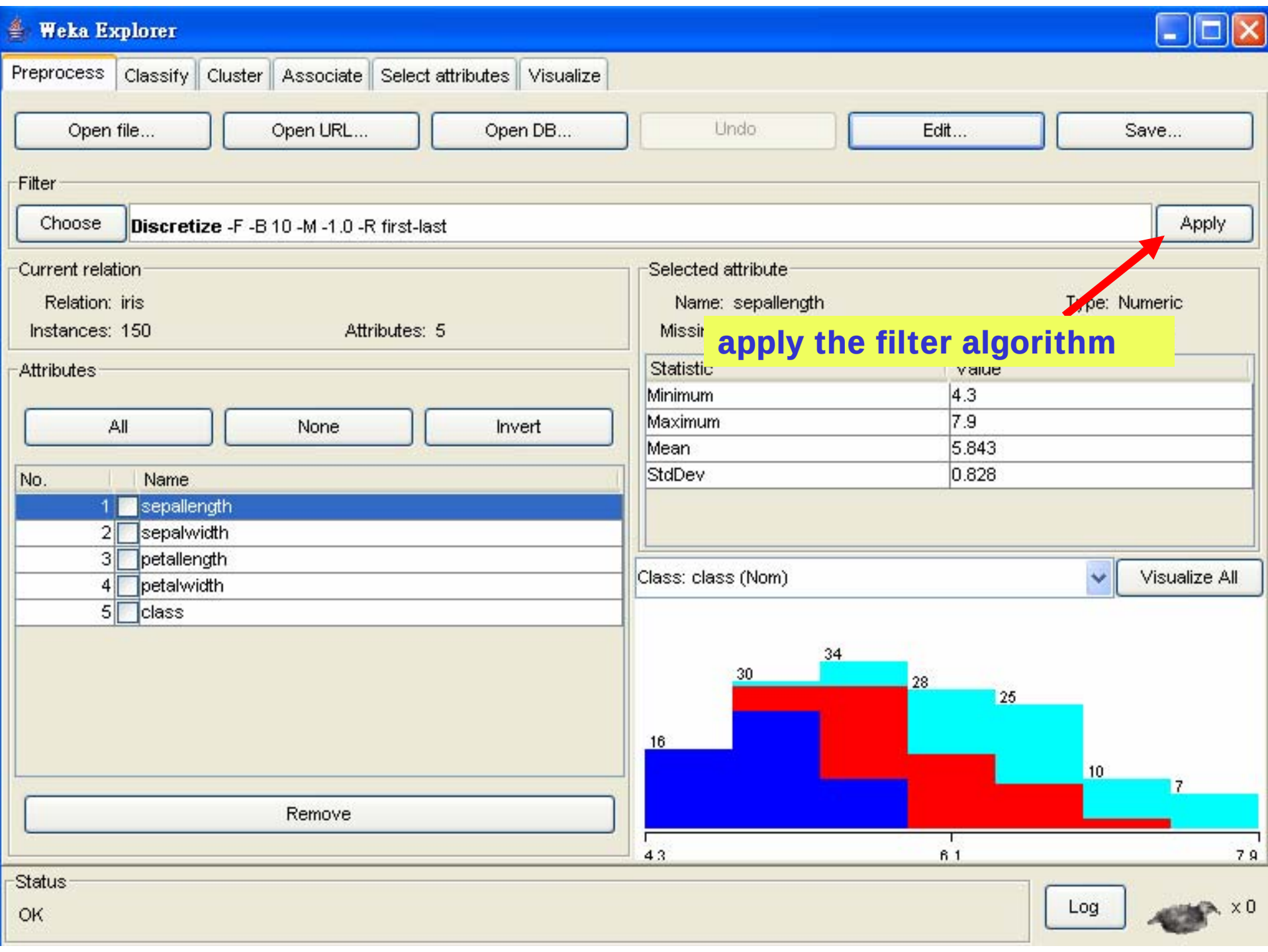

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: iris
Instances: 150
Attributes: 5
Selected attribute
Name: petallength
Type: Numeric
Missing: 0 (0%)
Distinct: 43
Unique: 10 (7%)
Statistic
Value
1
6.9
3.759
StdDev
1.764
Attributes
Click here to choose filter algorithm
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Remove
Class: class (Nom)
Visualize All
50
3
34
47
16
1
3.95
6.9
Status
OK
Log
47
(4.54, 5.72]
x 0

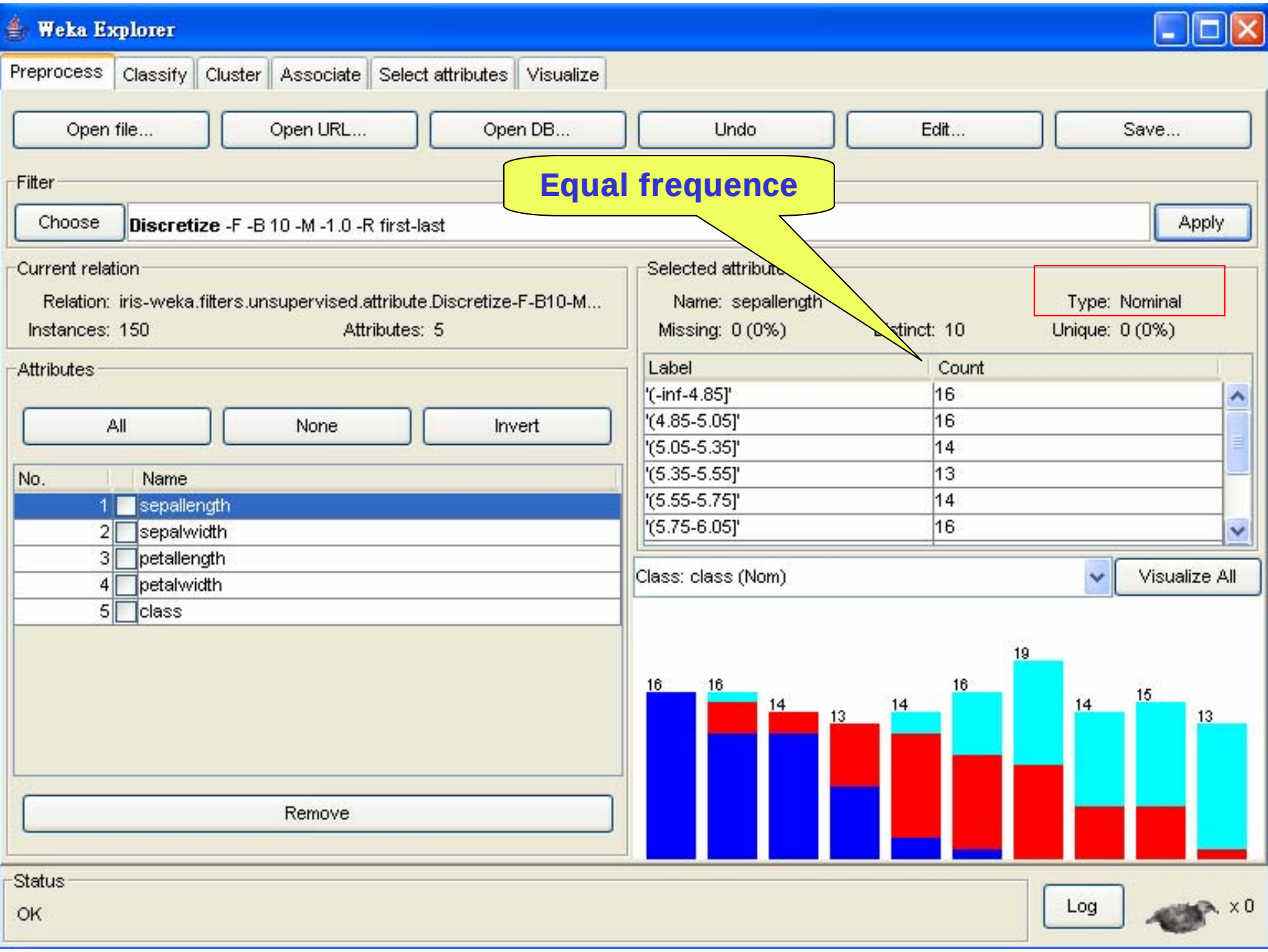

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Edit...
Save...
Filter
weka
filters
supervised
unsupervised
attribute
Add
AddCluster
AddExpression
AddNoise
ChangeDateFormat
ClusterMembership
Copy
Discretize
FirstOrder
MakeIndicator
MergeTwoValues
NominalToBinary
Normalize
NumericToBinary
NumericTransform
Obfuscate
PKIDiscretize
RandomProjection
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
Class: class (Nom)
Visualize All
50
3
34
47
16
1
3.95
6.9
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Edit...
Save...
Filter
Choose
Discretize -B 10 -M -1.0 -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%)
Click here to set the parameter for filter algorithm
Attributes
All
None
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
StdDev
1.764
Class: class (Nom)
Visualize All
50
3
34
47
16
1
3.95
6.9
Remove
Status
OK
Log
x 0

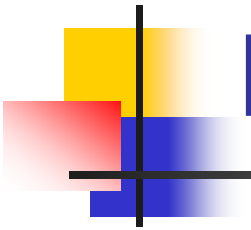

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Edit...
Save...
Filter
Choose
Discretize -B 10 -M -1.0 -R first-last
Apply
Current relation
Relation: iris
Instances: 150
Attributes
Attributes
All
None
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Remove
Set parameter
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
desiredWeightOfInstancesPerInterval
-1.0
findNumBins
False
invertSelection
False
makeBinary
False
useEqualFrequency
False
Open...
Save...
OK
Cancel
Visualize All
16
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Edit...
Save...
Filter
Choose
Discretize -F -B 10 -M -1.0 -R first-last
Apply
Current relation
Relation: iris
Instances: 150
Attributes
Attributes
All
None
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Remove
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
desiredWeightOfInstancesPerInterval
-1.0
findNumBins
False
invertSelection
False
makeBinary
False
useEqualFrequency
True
Open...
Save...
OK
Cancel
Visualize All
16
Status
OK
Log
x 0

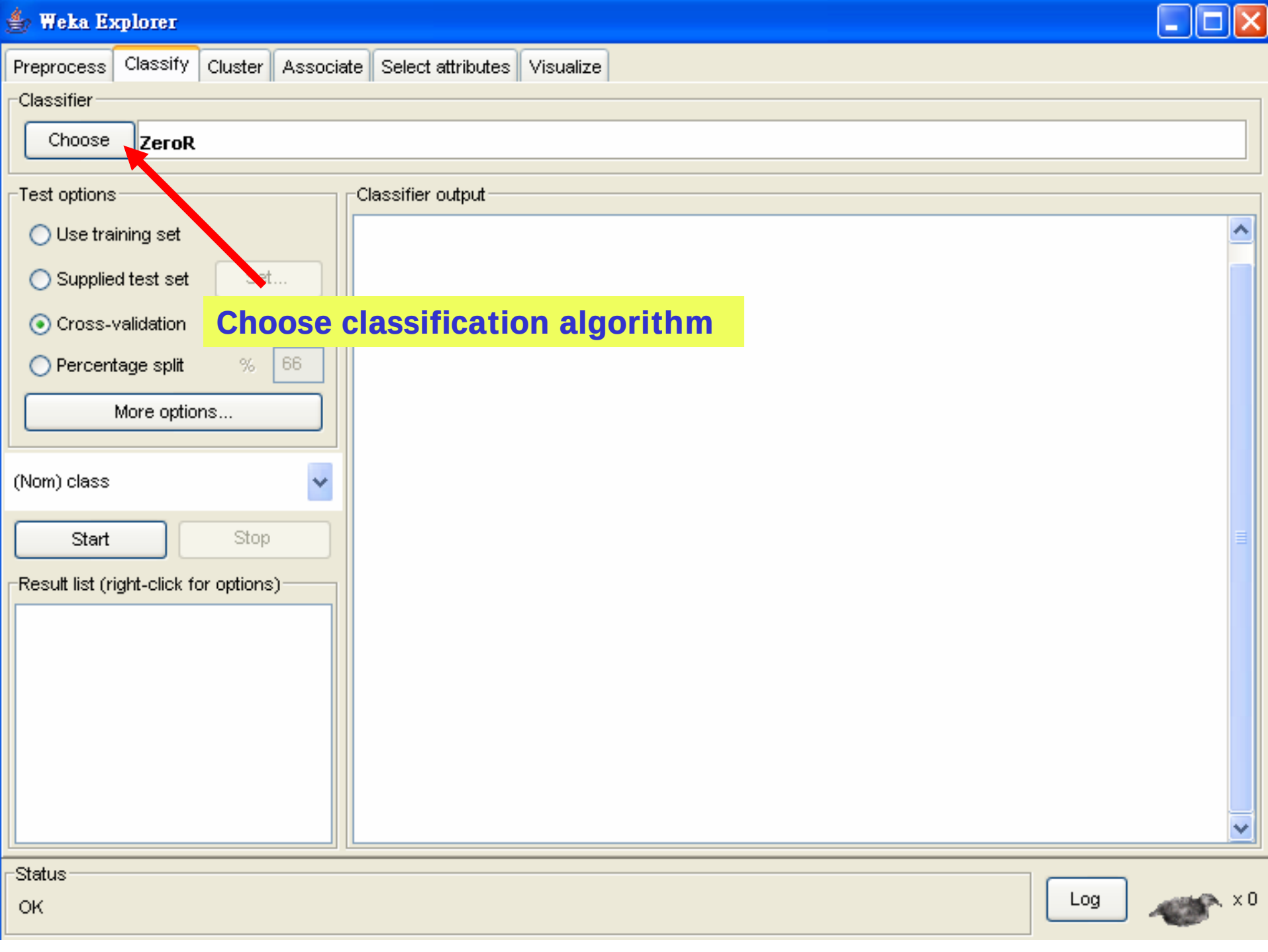

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Edit...
Save...
Filter
Choose
Discretize -F -B 10 -M -1.0 -R first-last
Apply
Current relation
Relation: iris
Instances: 150
Attributes: 5
Selected attribute
Name: sepallength
Type: Numeric
apply the filter algorithm
Statistic
Value
Minimum 4.3
Maximum 7.9
Mean 5.843
StdDev 0.828
Attributes
All
None
Invert
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Class: class (Nom)
Visualize All
34
30
28
25
16
10
7
4.3
6.1
7.9
Remove
Status
OK
Log
x 0

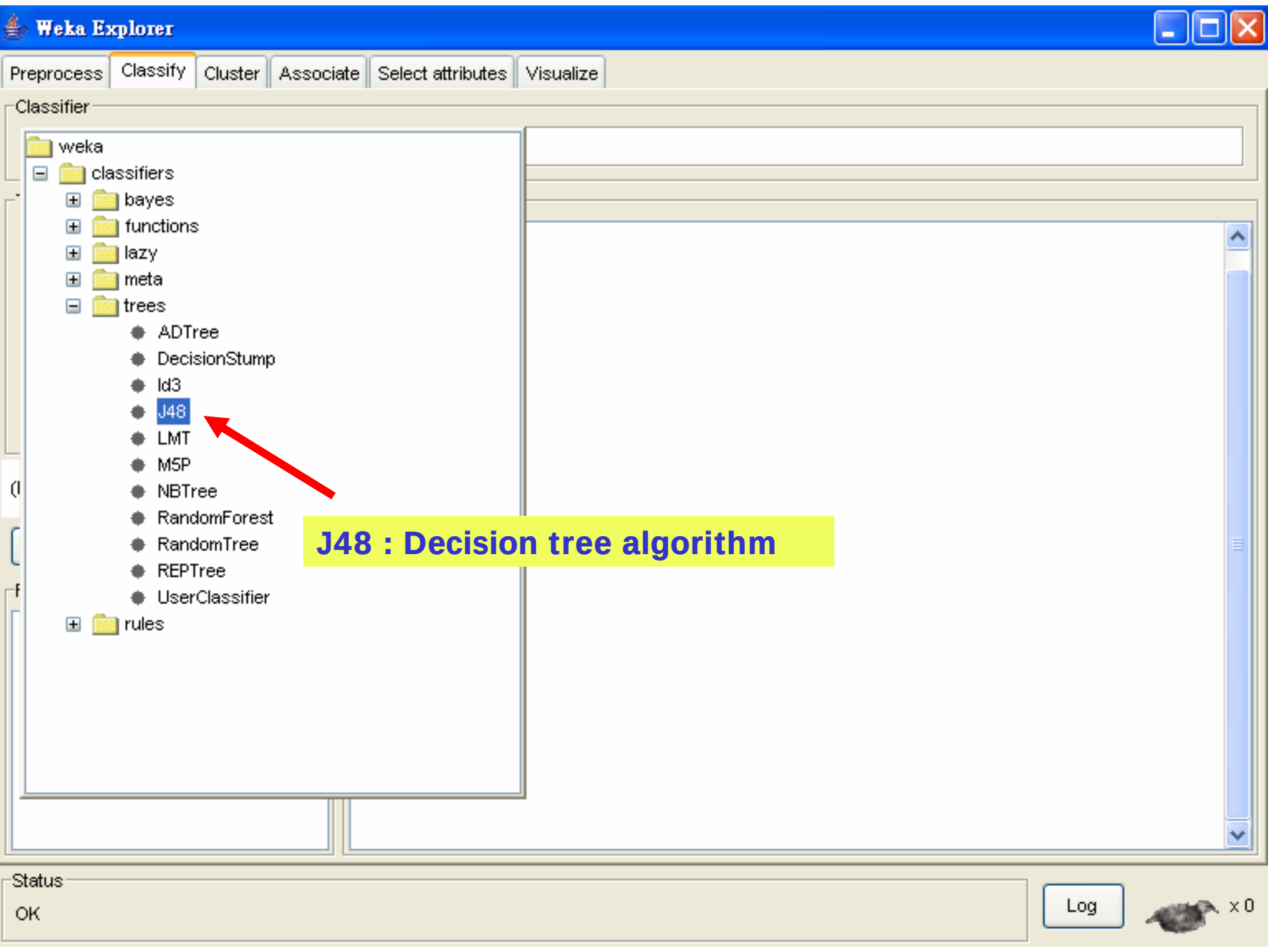

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Edit...
Save...
Equal frequence
Filter
Choose
Discretize -F -B 10 -M -1.0 -R first-last
Apply
Current relation
Relation: iris-weka.filters.unsupervised.attribute.Discretize-F-B10-M...
Instances: 150
Attributes: 5
Selected attribute
Name: sepallength
Type: Nominal
Missing: 0 (0%)
Distinct: 10
Unique: 0 (0%)
Label
Count
'(-inf-4.85]' 16
'(4.85-5.05]' 16
'(5.05-5.35]' 14
'(5.35-5.55]' 13
'(5.55-5.75]' 14
'(5.75-6.05]' 16
Attributes
All
None
Invert
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Remove
Class: class (Nom)
Visualize All
16
16
14
13
14
16
19
14
15
13
Status
OK
Log
x 0

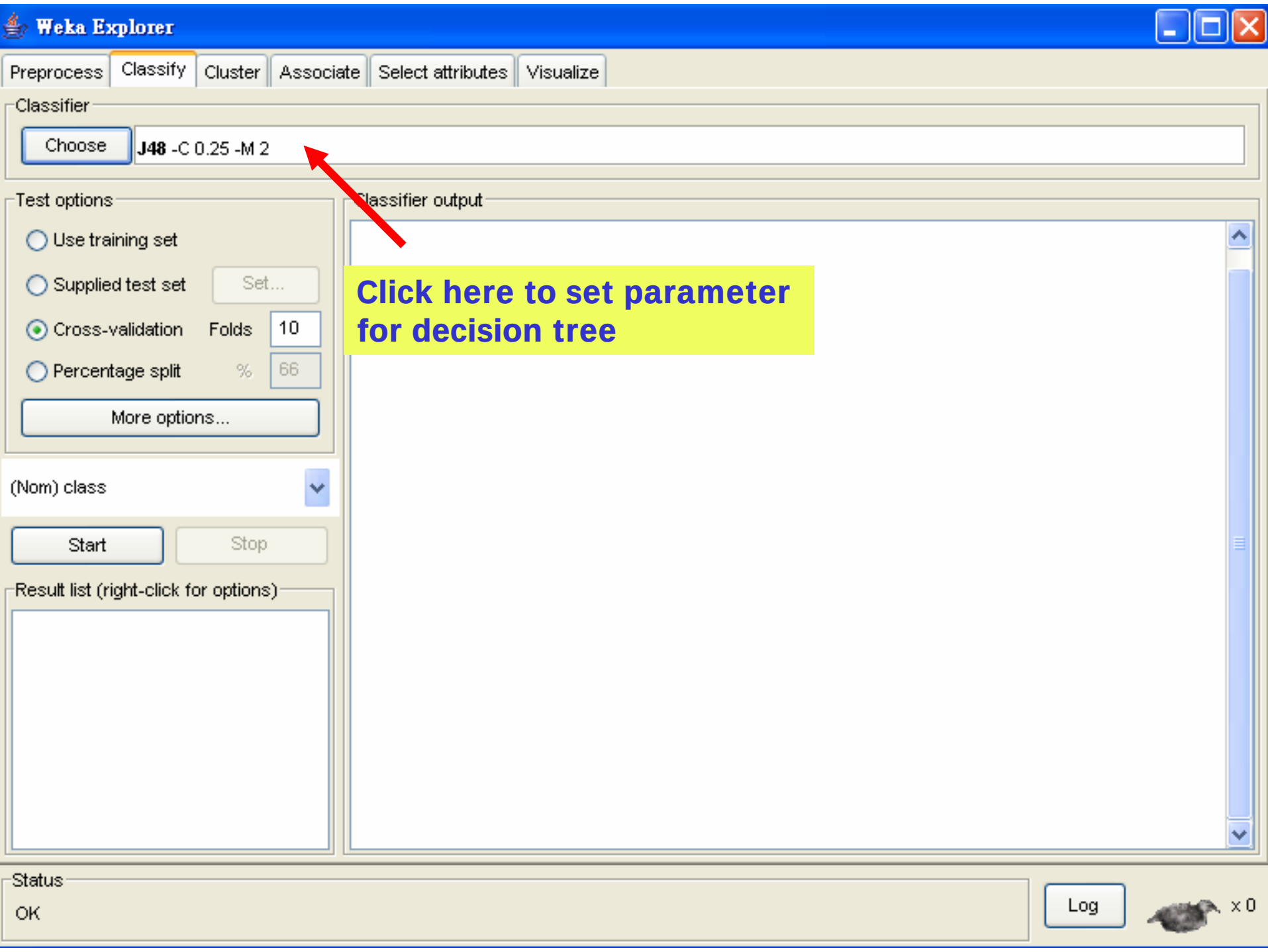
Building "classifiers"

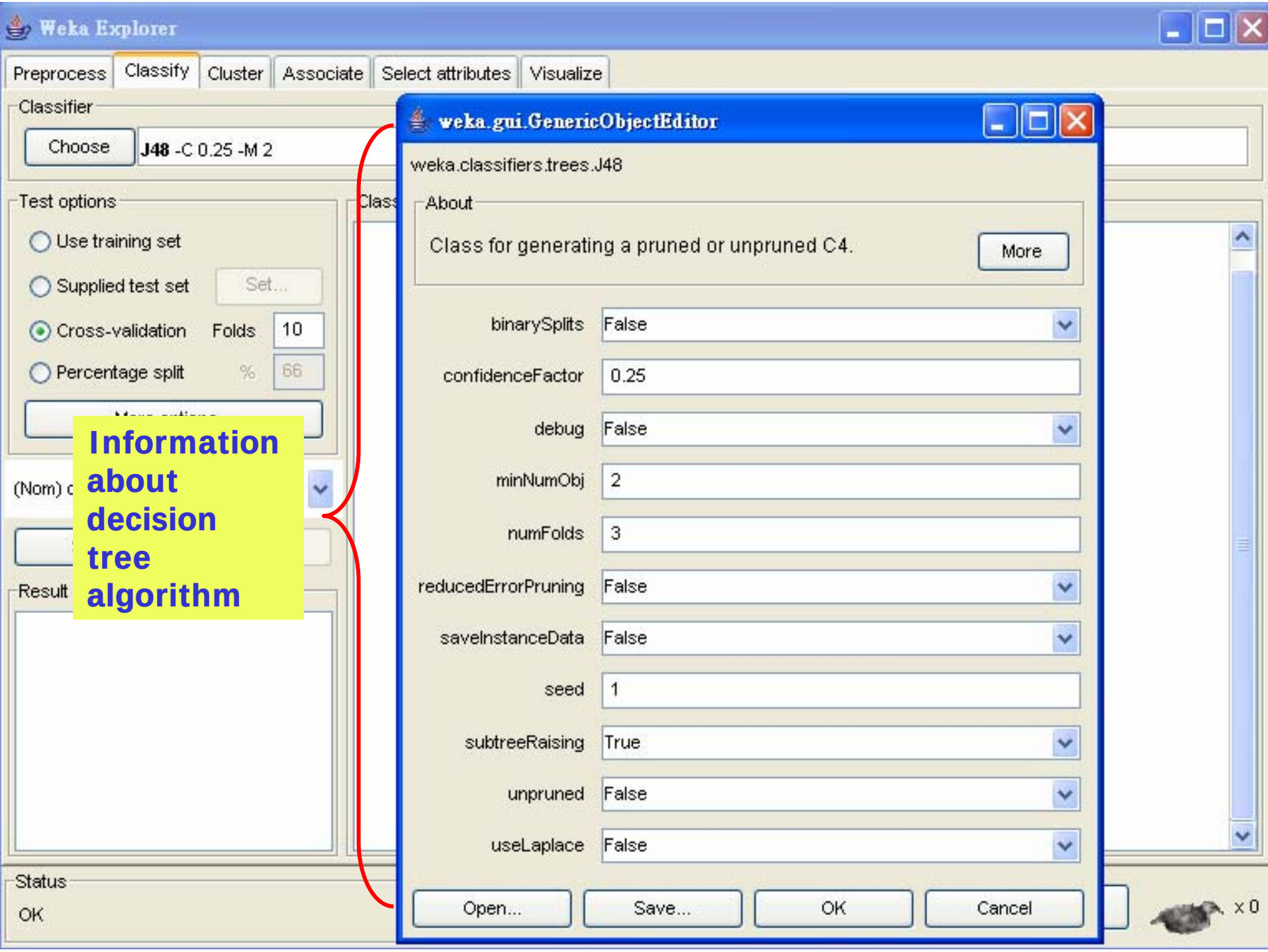


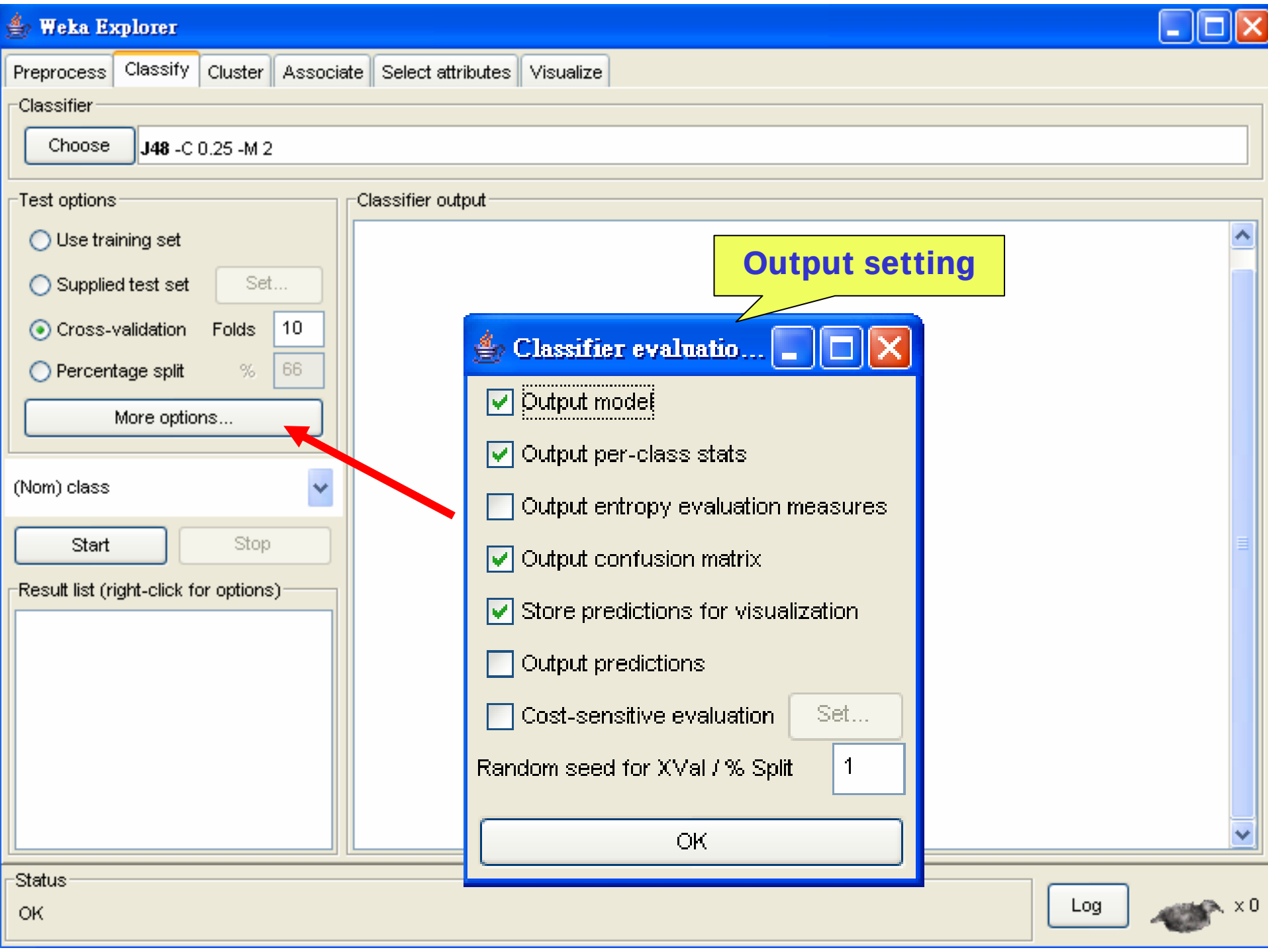
- Classifiers in WEKA are models for predicting nominal or numeric quantities
- Implemented learning schemes include:
  - **Decision trees** and lists, instance-based classifiers, **support vector machines**, multi-layer perceptrons, logistic regression, Bayes' nets, …
- "Meta"-classifiers include:
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

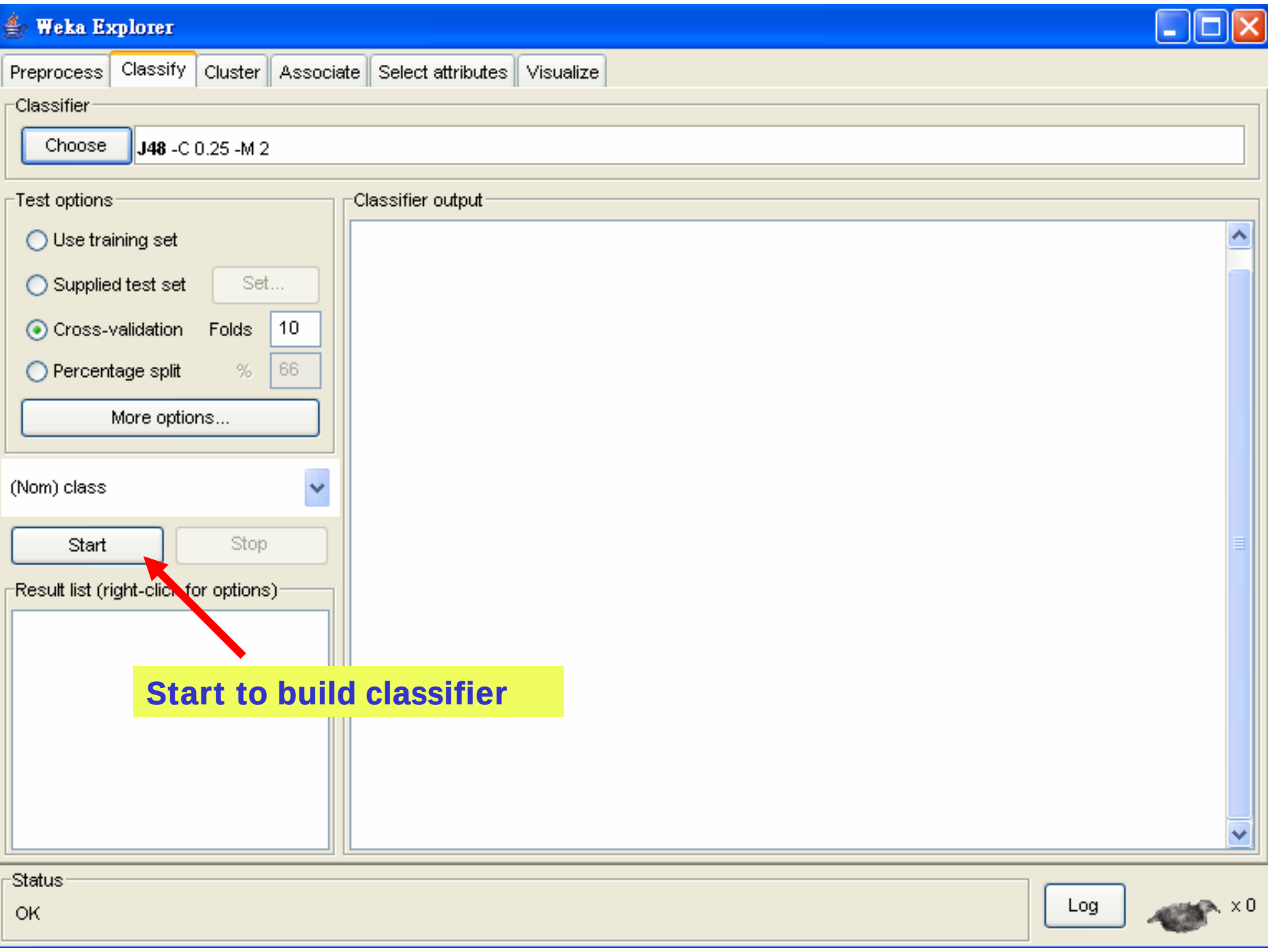

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
ZeroR
Test options
Use training set
Supplied test set
Cross-validation
Percentage split
%
66
More options...
Classifier output
Choose classification algorithm
(Nom) class
Start
Stop
Result list (right-click for options)
Status
OK
Log
x 0

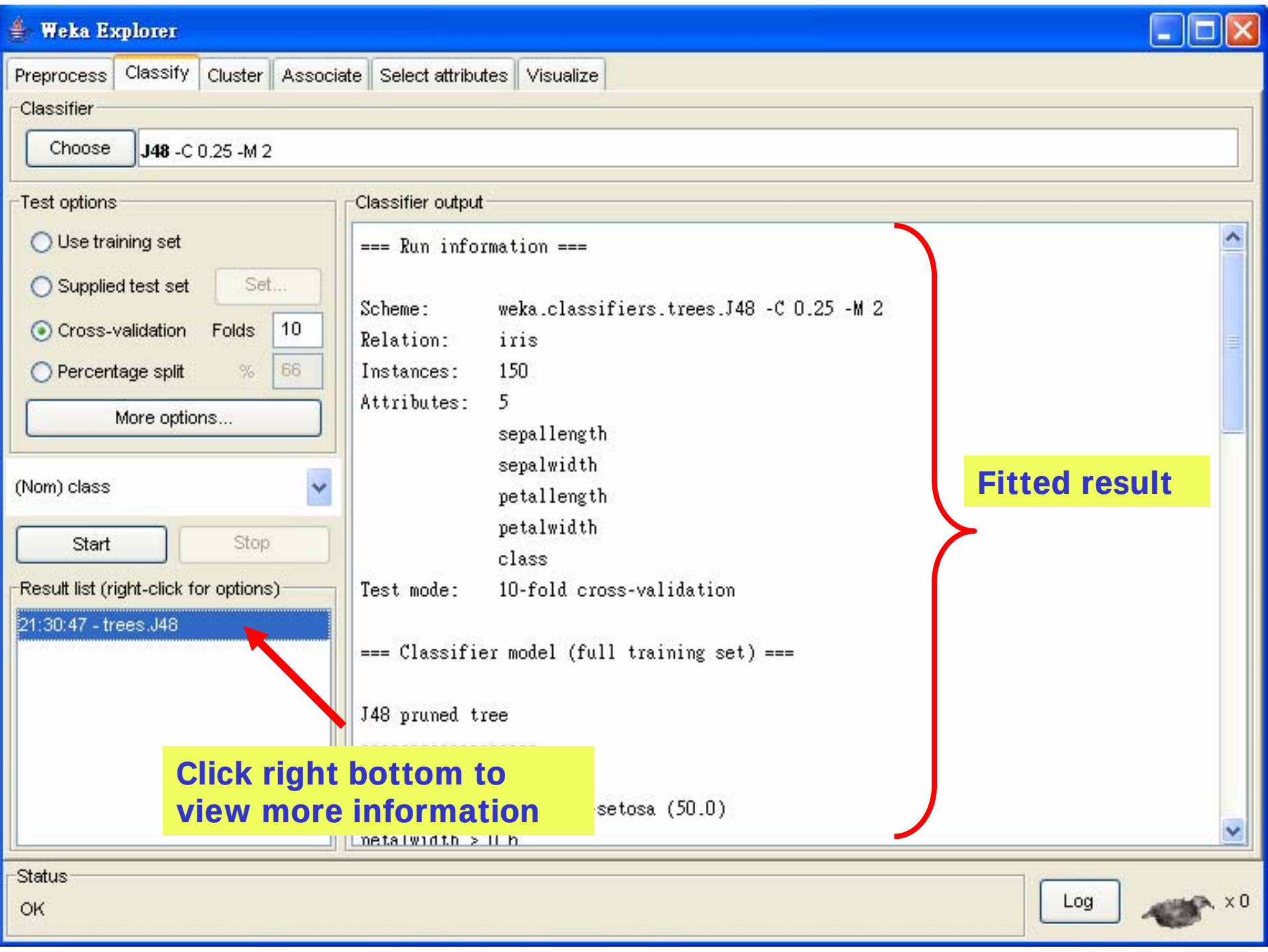

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
weka
classifiers
bayes
functions
lazy
meta
trees
ADTree
DecisionStump
Id3
J48
LMT
M5P
NBTree
RandomForest
RandomTree
REPTree
UserClassifier
rules
J48 : Decision tree algorithm
Status
OK
Log
x 0

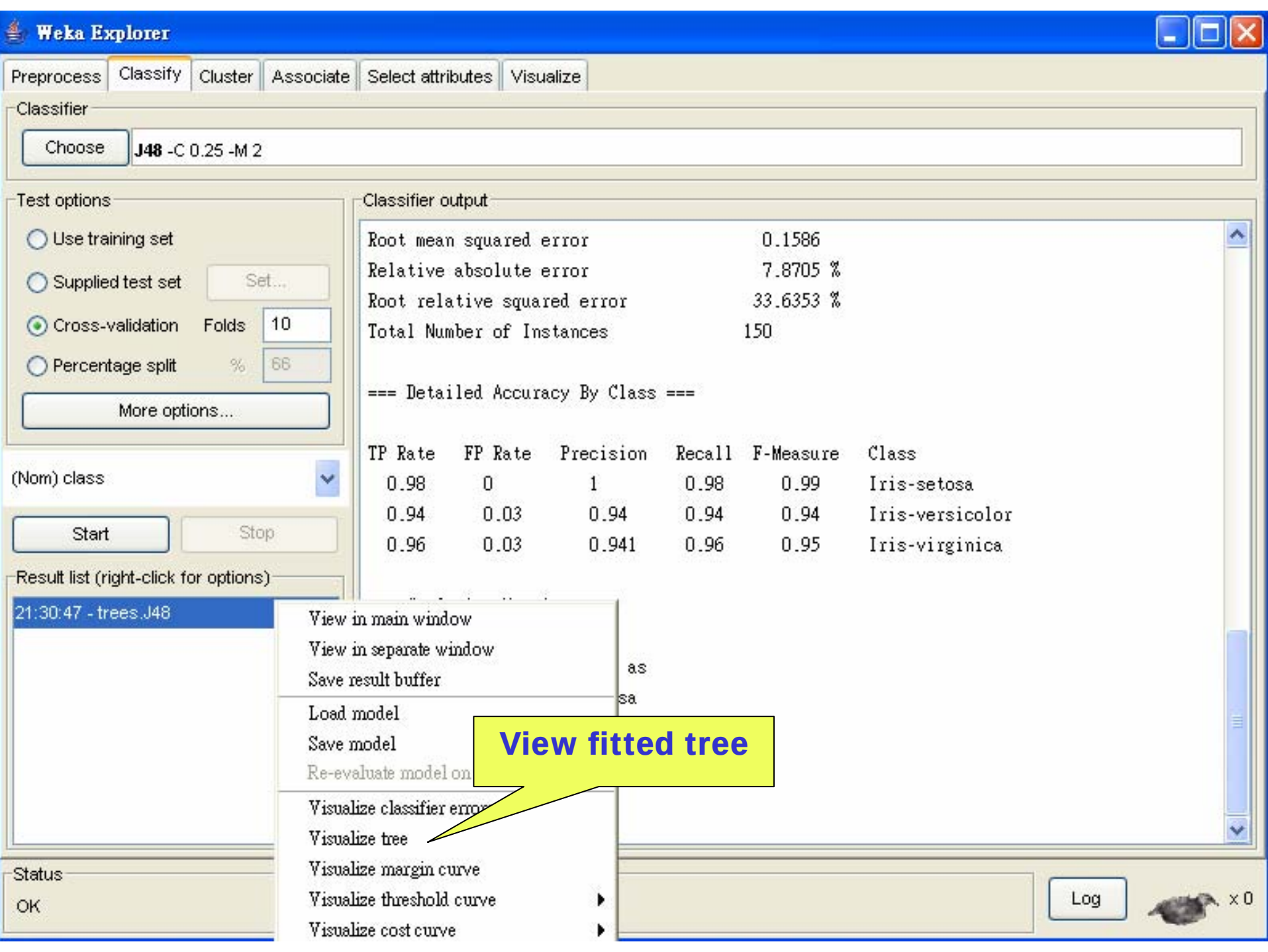

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
Classifier output
Click here to set parameter for decision tree
(Nom) class
Start
Stop
Result list (right-click for options)
Status
OK
Log
x 0

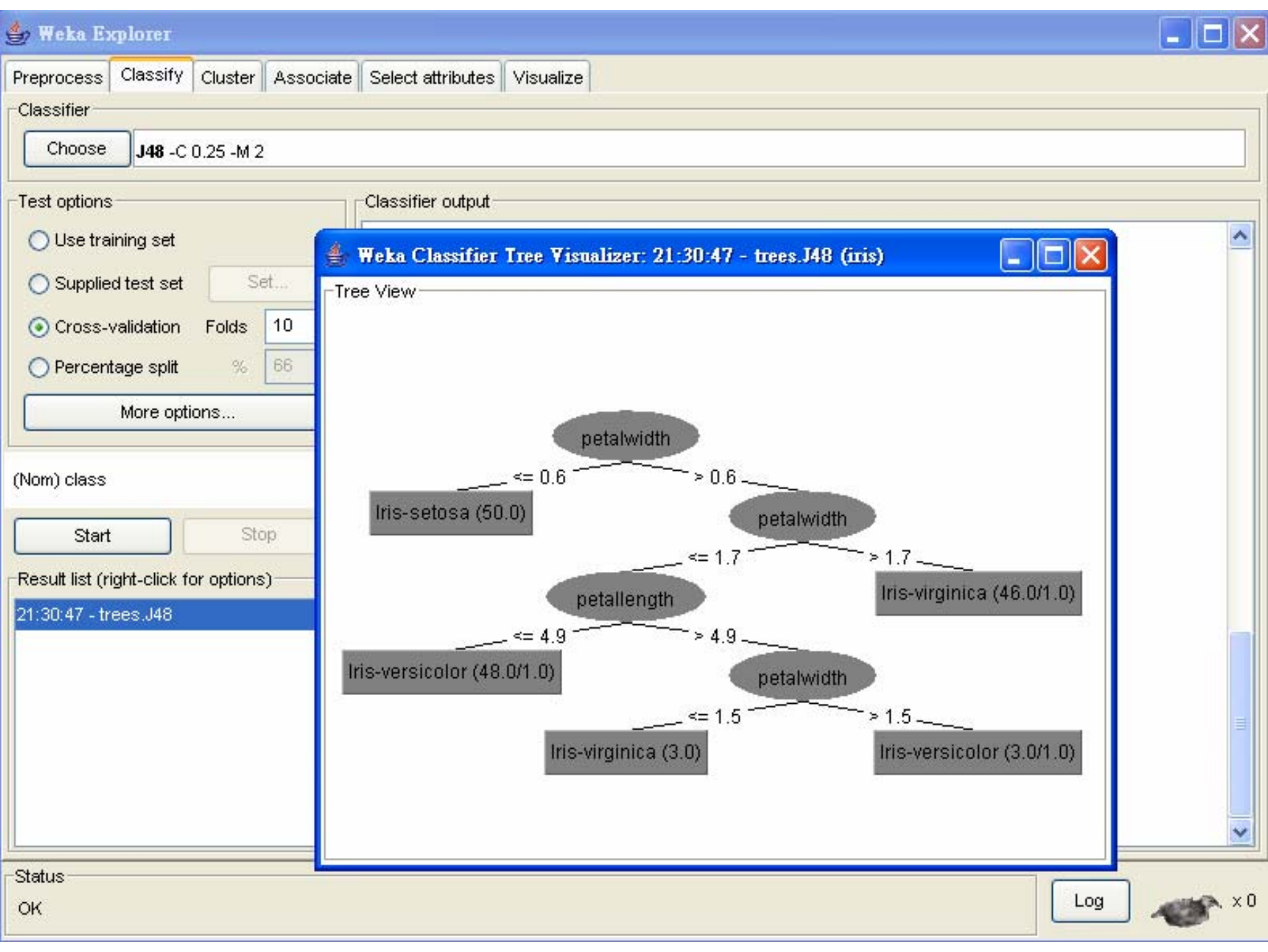

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
Result
Status
OK
weka.gui.GenericObjectEditor
weka.classifiers.trees.J48
About
Class for generating a pruned or unpruned C4.
More
binarySplits
False
confidenceFactor
0.25
debug
False
minNumObj
2
numFolds
3
reducedErrorPruning
False
saveInstanceData
False
seed
1
subtreeRaising
True
unpruned
False
useLaplace
False
Open...
Save...
OK
Cancel
Information about decision tree algorithm

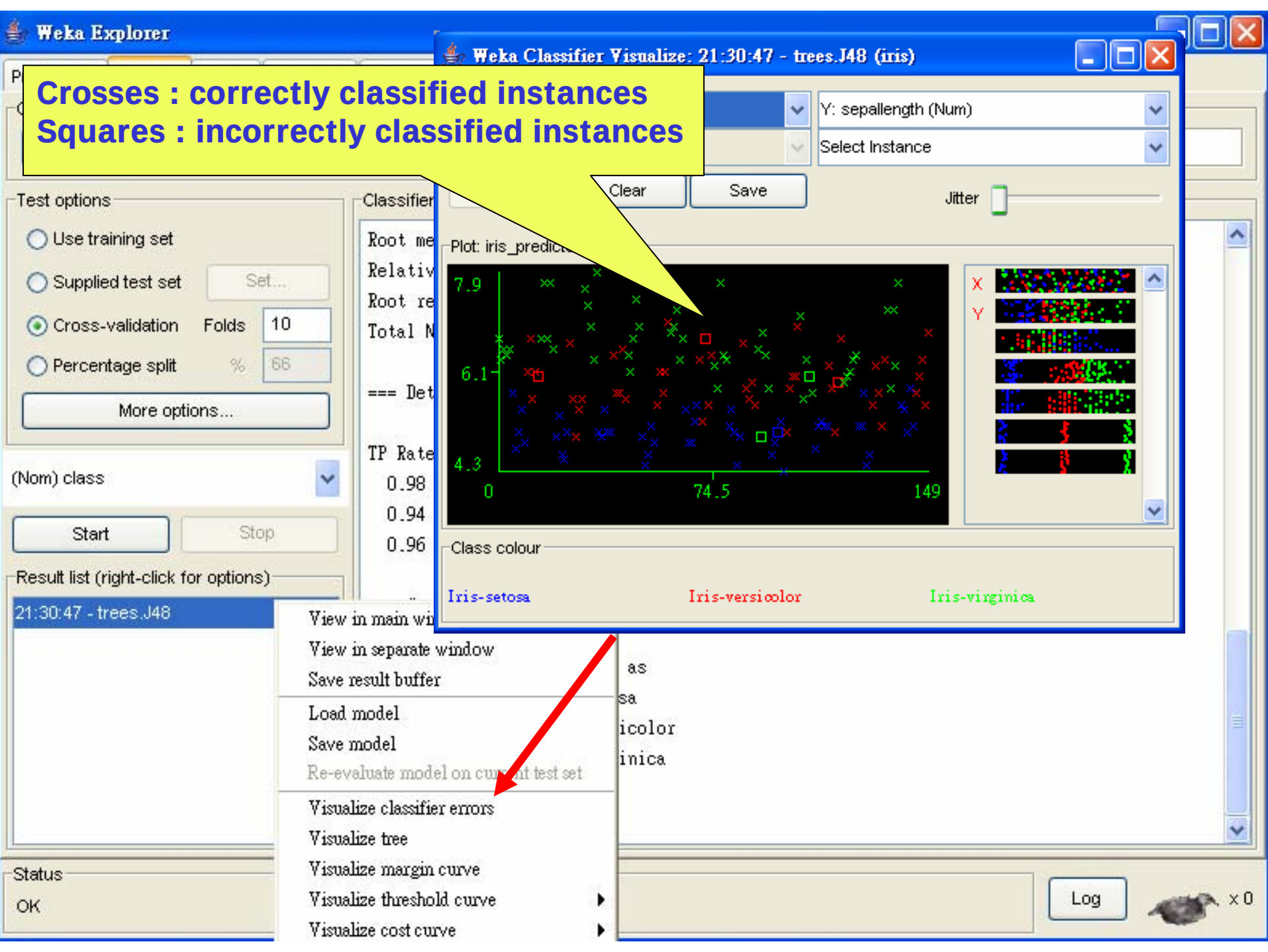

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
Classifier output
Output setting
Classifier evaluatio...
Output model
Output per-class stats
Output entropy evaluation measures
Output confusion matrix
Store predictions for visualization
Output predictions
Cost-sensitive evaluation
Set...
Random seed for XVal / % Split
1
OK
(Nom) class
Start
Stop
Result list (right-click for options)
Status
OK
Log
x 0

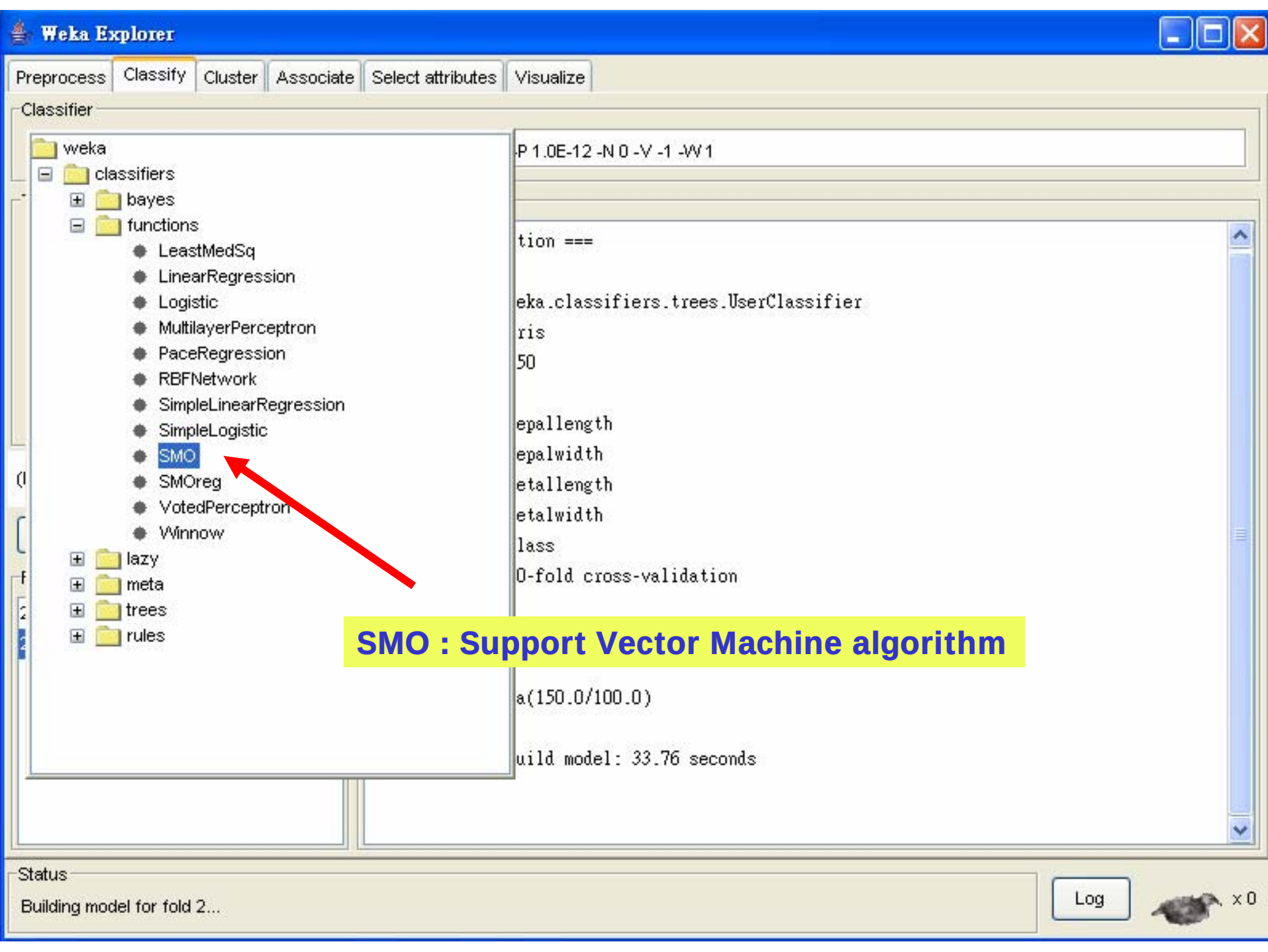

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
Classifier output
(Nom) class
Start
Stop
Result list (right-click for options)
Start to build classifier
Status
OK
Log
x 0

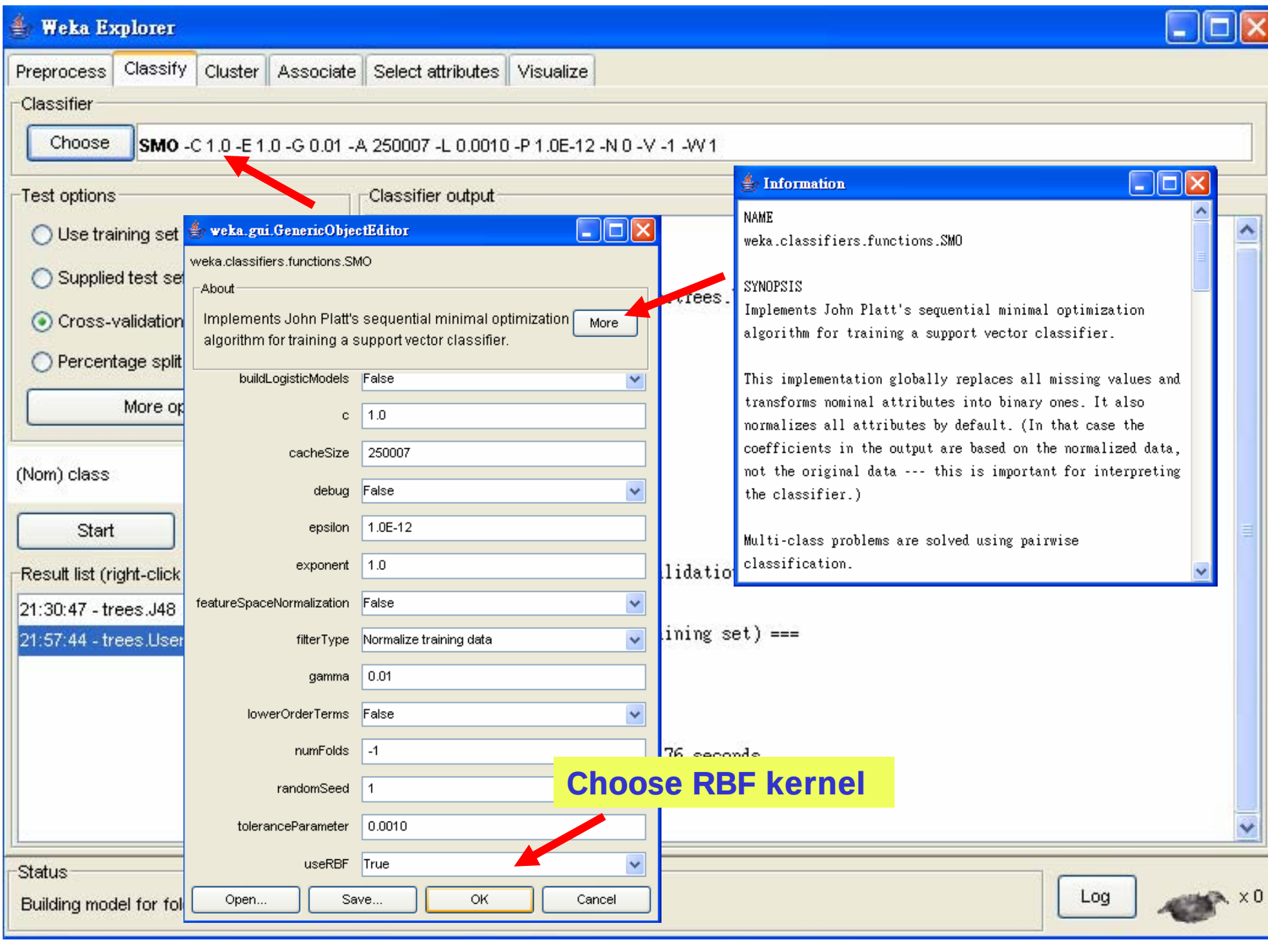

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
21:30:47 - trees.J48
Classifier output
=== Run information ===
Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: iris
Instances: 150
Attributes: 5
sepallength
sepalwidth
petallength
petalwidth
class
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree
setosa (50.0)
Fitted result
Click right bottom to view more information
Status
OK
Log
x 0

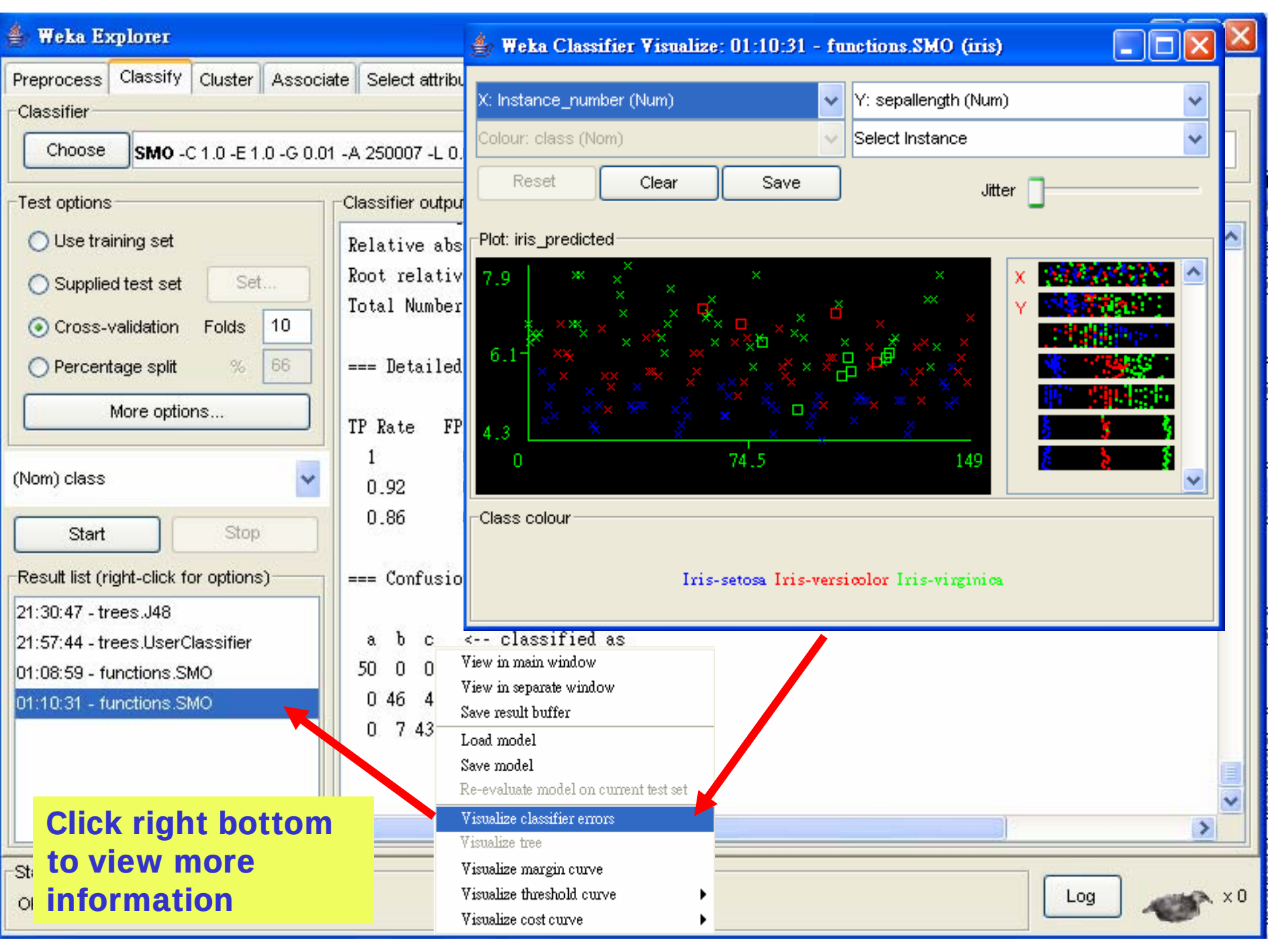

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
Classifier output
Root mean squared error 0.1586
Relative absolute error 7.8705 %
Root relative squared error 33.6353 %
Total Number of Instances 150
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure Class
0.98 0 1 0.98 0.99 Iris-setosa
0.94 0.03 0.94 0.94 0.94 Iris-versicolor
0.96 0.03 0.941 0.96 0.95 Iris-virginica
(Nom) class
Start
Stop
Result list (right-click for options)
21:30:47 - trees.J48
View in main window
View in separate window
Save result buffer
Load model
Save model
Re-evaluate model on
Visualize classifier error
Visualize tree
Visualize margin curve
Visualize threshold curve
Visualize cost curve
View fitted tree
Status
OK
Log
x 0

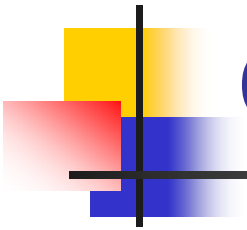

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
Classifier output
(Nom) class
Start
Stop
Result list (right-click for options)
21:30:47 - trees.J48
Status
OK
Log
x 0
Weka Classifier Tree Visualizer: 21:30:47 - trees.J48 (iris)
Tree View
petalwidth
<= 0.6
> 0.6
Iris-setosa (50.0)
petalwidth
<= 1.7
> 1.7
petallength
Iris-virginica (46.0/1.0)
<= 4.9
> 4.9
Iris-versicolor (48.0/1.0)
petalwidth
<= 1.5
> 1.5
Iris-virginica (3.0)
Iris-versicolor (3.0/1.0)


Weka Explorer
Weka Classifier Visualize: 21:30:47 - trees.J48 (iris)
Crosses : correctly classified instances
Squares : incorrectly classified instances
Y: sepallength (Num)
Select Instance
Clear
Save
Jitter
Plot: iris_predicted
7.9
6.1
4.3
0
74.5
149
X
Y
Class colour
Iris-setosa
Iris-versicolor
Iris-virginica
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)
21:30:47 - trees.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
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
weka
classifiers
bayes
functions
LeastMedSq
LinearRegression
Logistic
MultilayerPerceptron
PaceRegression
RBFNetwork
SimpleLinearRegression
SimpleLogistic
SMO
SMOreg
VotedPerceptron
Winnow
lazy
meta
trees
rules
-P 1.0E-12 -N 0 -V -1 -W 1
tion ===
eka.classifiers.trees.UserClassifier
ris
50
epallength
epalwidth
etallength
etalwidth
lass
0-fold cross-validation
SMO : Support Vector Machine algorithm
a(150.0/100.0)
uild model: 33.76 seconds
Status
Building model for fold 2...
Log
x 0

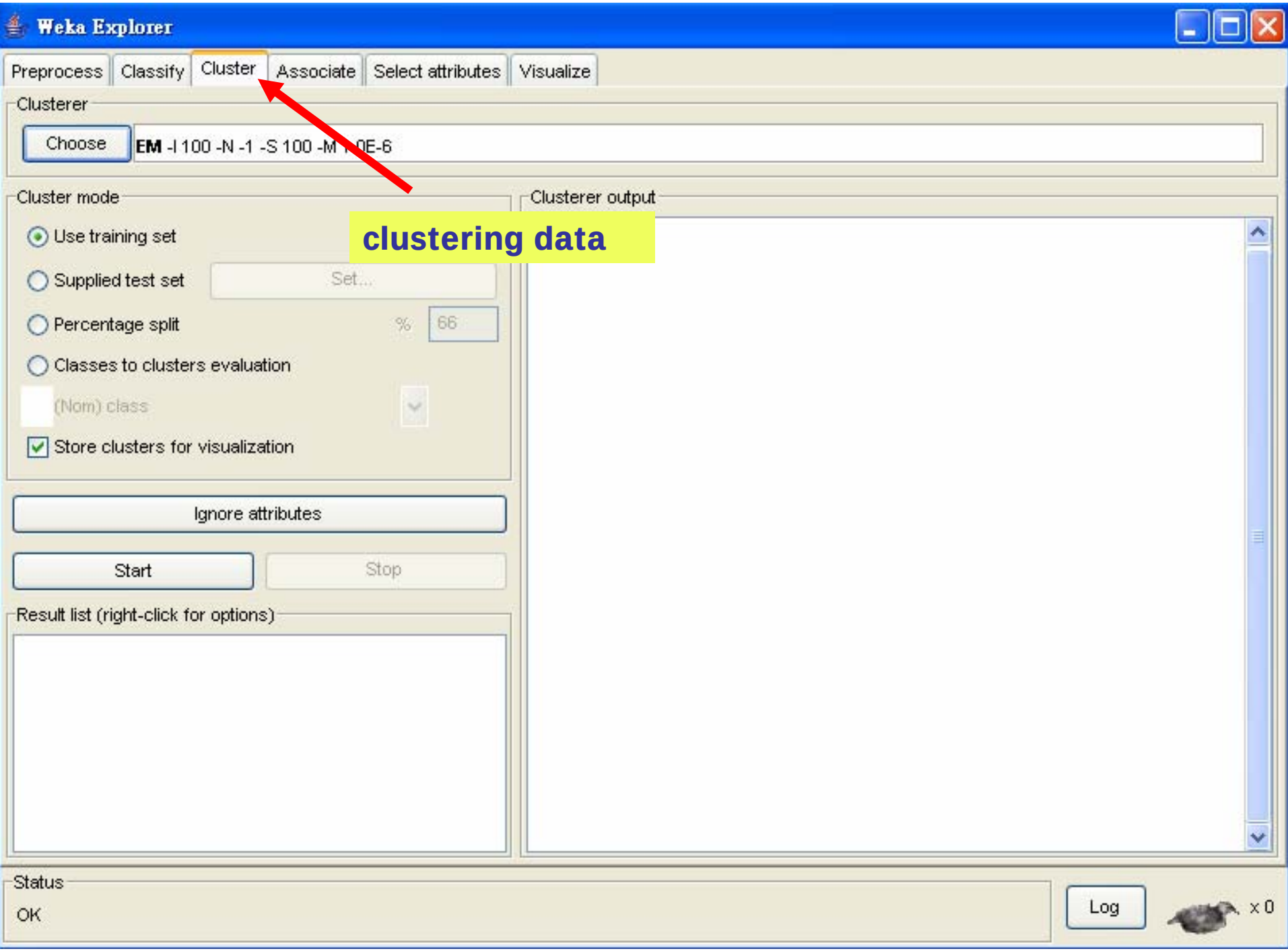

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
SMO -C 1.0 -E 1.0 -G 0.01 -A 250007 -L 0.0010 -P 1.0E-12 -N 0 -V -1 -W 1
Test options
Use training set
Supplied test set
Cross-validation
Percentage split
More options
Classifier output
(Nom) class
Start
Result list (right-click
21:30:47 - trees.J48
21:57:44 - trees.User
Status
Building model for fol
Log
x 0
weka.gui.GenericObjectEditor
weka.classifiers.functions.SMO
About
Implements John Platt's sequential minimal optimization algorithm for training a support vector classifier.
More
buildLogisticModels False
c 1.0
cacheSize 250007
debug False
epsilon 1.0E-12
exponent 1.0
featureSpaceNormalization False
filterType Normalize training data
gamma 0.01
lowerOrderTerms False
numFolds -1
randomSeed 1
toleranceParameter 0.0010
useRBF True
Open...
Save...
OK
Cancel
Information
NAME
weka.classifiers.functions.SMO
SYNOPSIS
Implements John Platt's sequential minimal optimization algorithm for training a support vector classifier.
This implementation globally replaces all missing values and transforms nominal attributes into binary ones. It also normalizes all attributes by default. (In that case the coefficients in the output are based on the normalized data, not the original data --- this is important for interpreting the classifier.)
Multi-class problems are solved using pairwise classification.
Choose RBF kernel

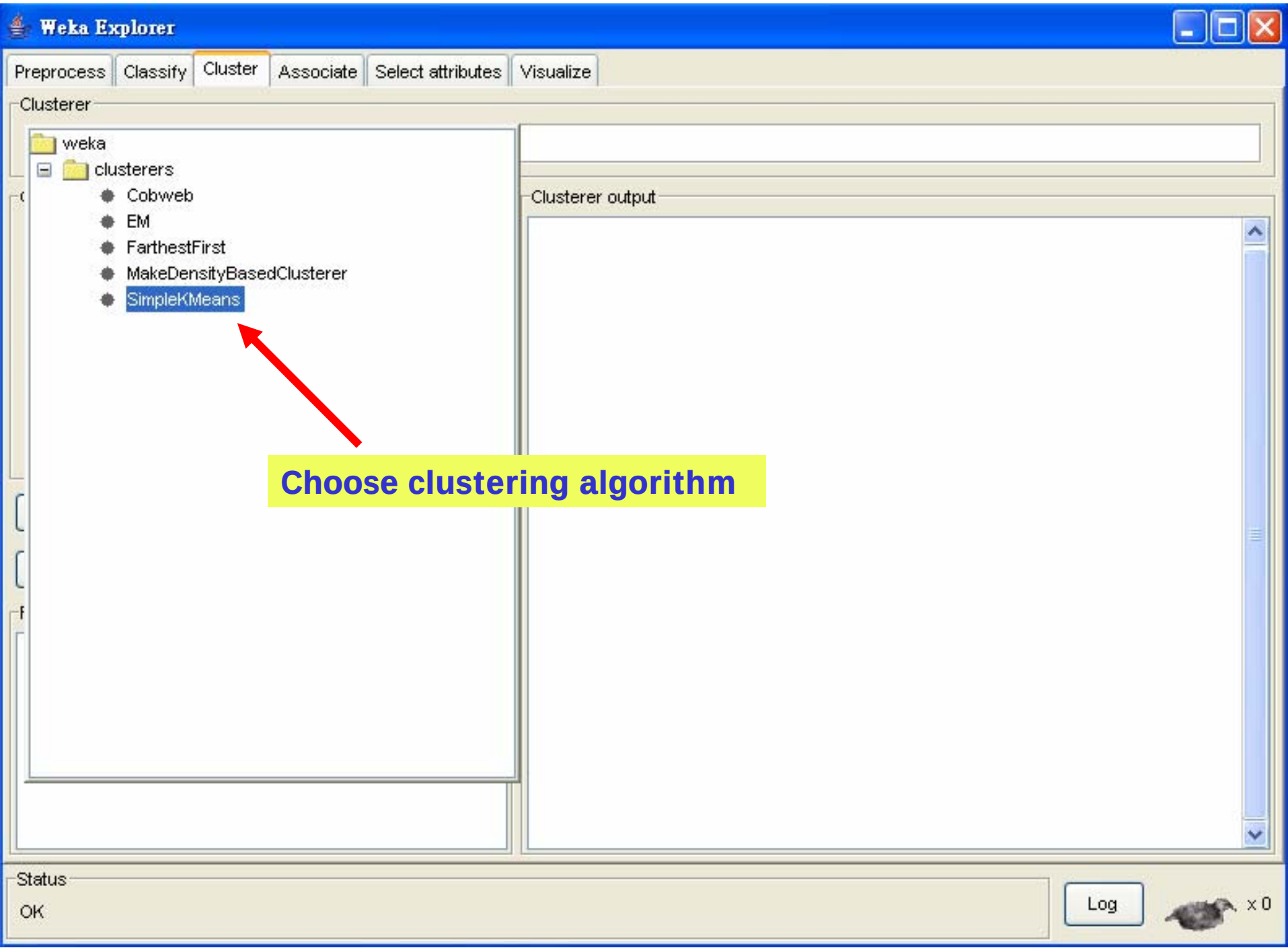

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attribu
Classifier
Choose
SMO -C 1.0 -E 1.0 -G 0.01 -A 250007 -L 0.
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
Classifier outpu
Relative abs
Root relativ
Total Number
=== Detailed
TP Rate FP
1
0.92
0.86
=== Confusio
a b c <-- classified as
50 0 0
0 46 4
0 7 43
(Nom) class
Start
Stop
Result list (right-click for options)
21:30:47 - trees.J48
21:57:44 - trees.UserClassifier
01:08:59 - functions.SMO
01:10:31 - functions.SMO
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
Log
x 0
Weka Classifier Visualize: 01:10:31 - functions.SMO (iris)
X: Instance_number (Num)
Y: sepallength (Num)
Colour: class (Nom)
Select Instance
Reset
Clear
Save
Jitter
Plot: iris_predicted
7.9
6.1
4.3
0
74.5
149
X
Y
Class colour
Iris-setosa Iris-versicolor Iris-virginica
Click right bottom to view more information

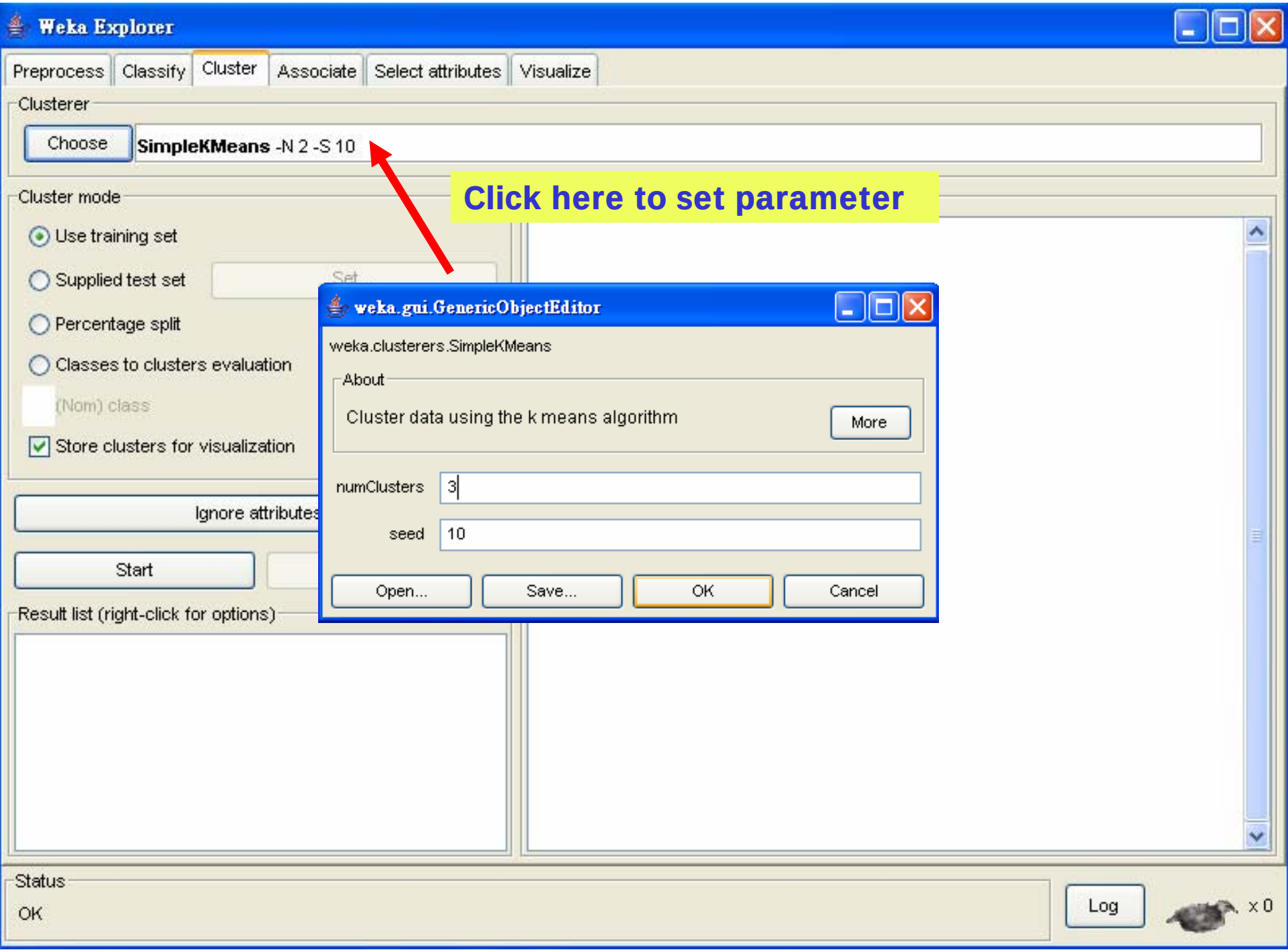


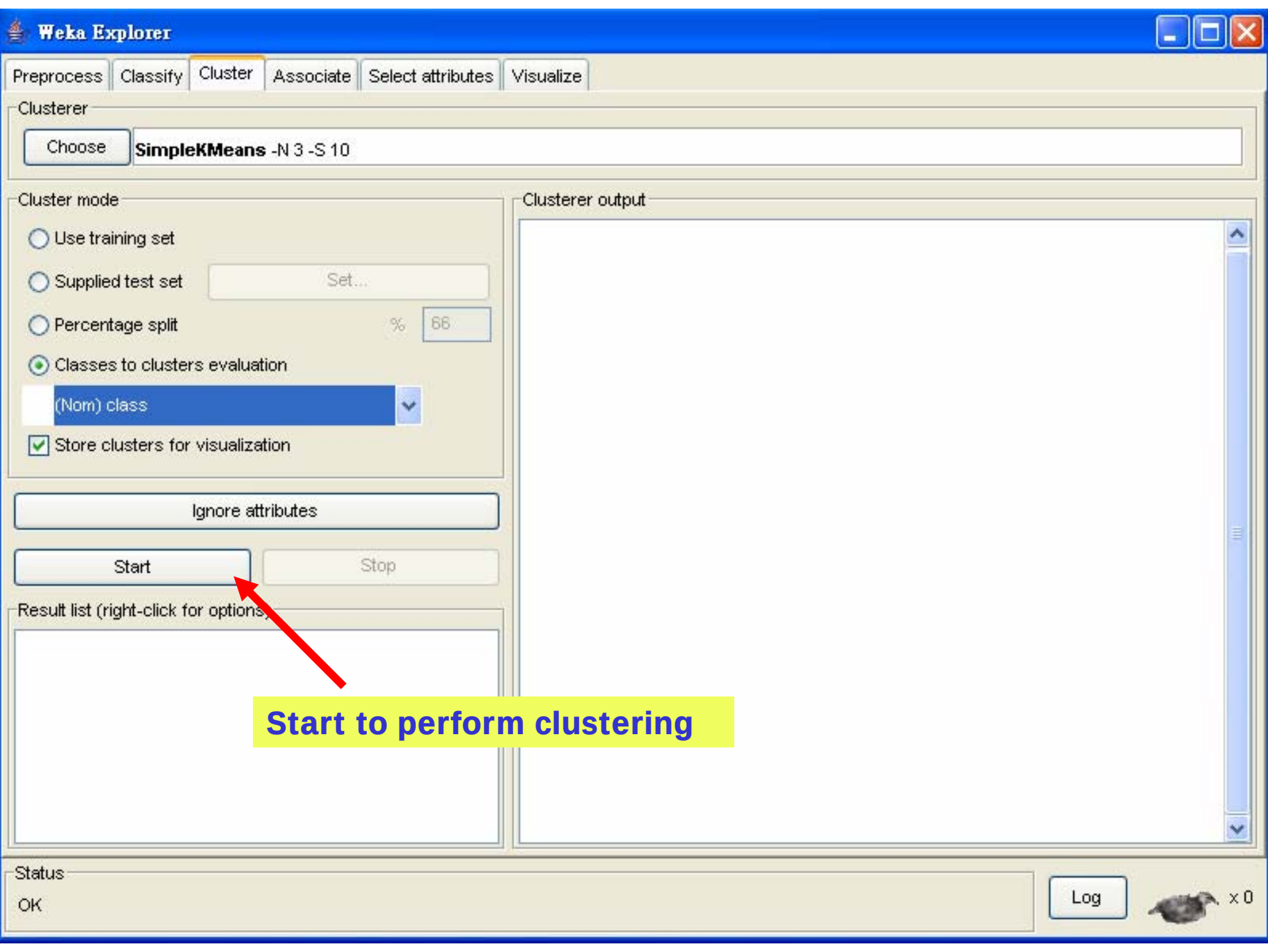
clustering data

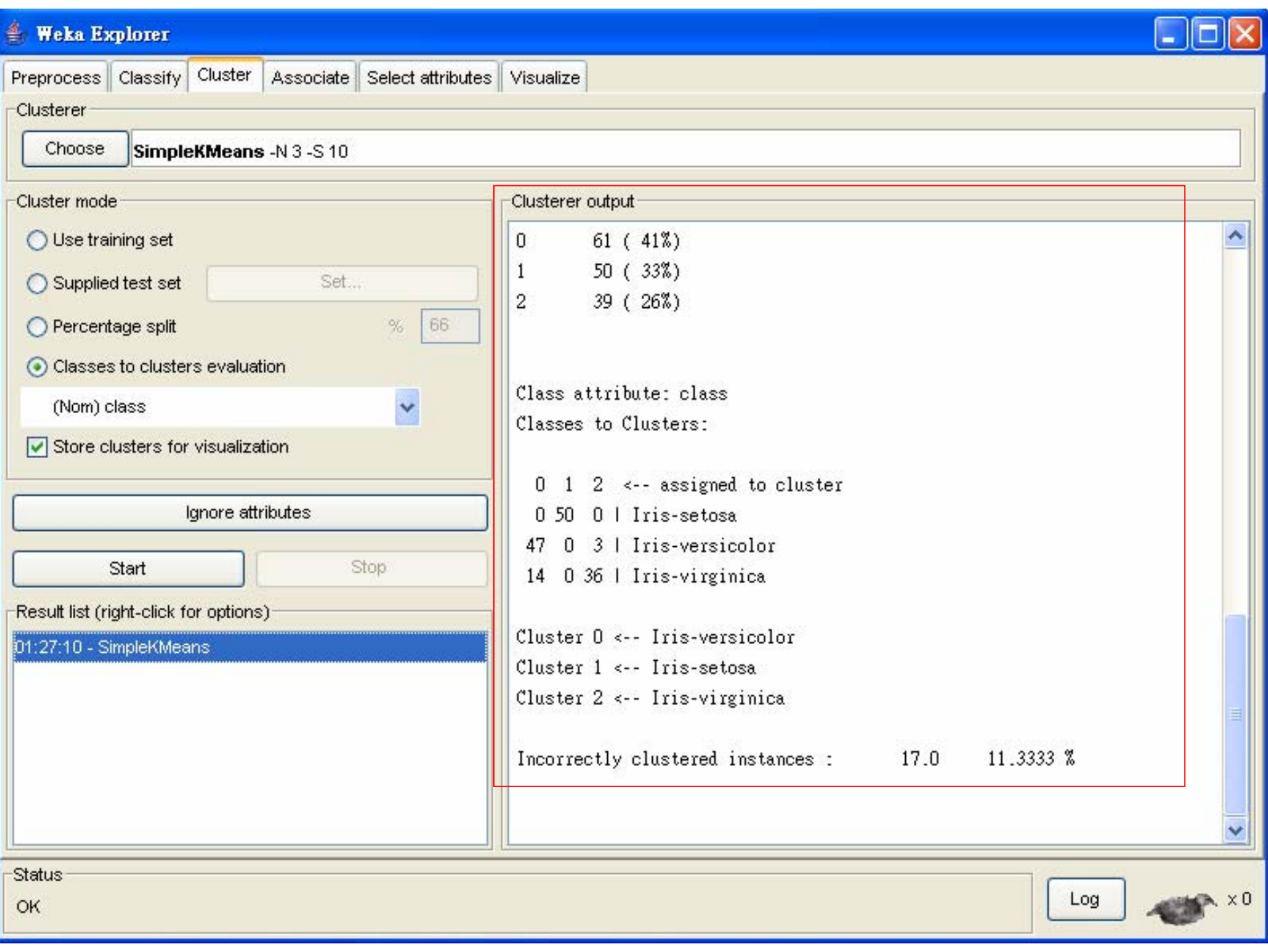
- WEKA contains “clusterers” for finding groups of similar instances in a dataset
- Implemented schemes are:
  - *k*-Means, EM, Cobweb, *X*-means, FarthestFirst
- Clusters can be visualized and compared to “true” clusters (if given)
- Evaluation based on loglikelihood if clustering scheme produces a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

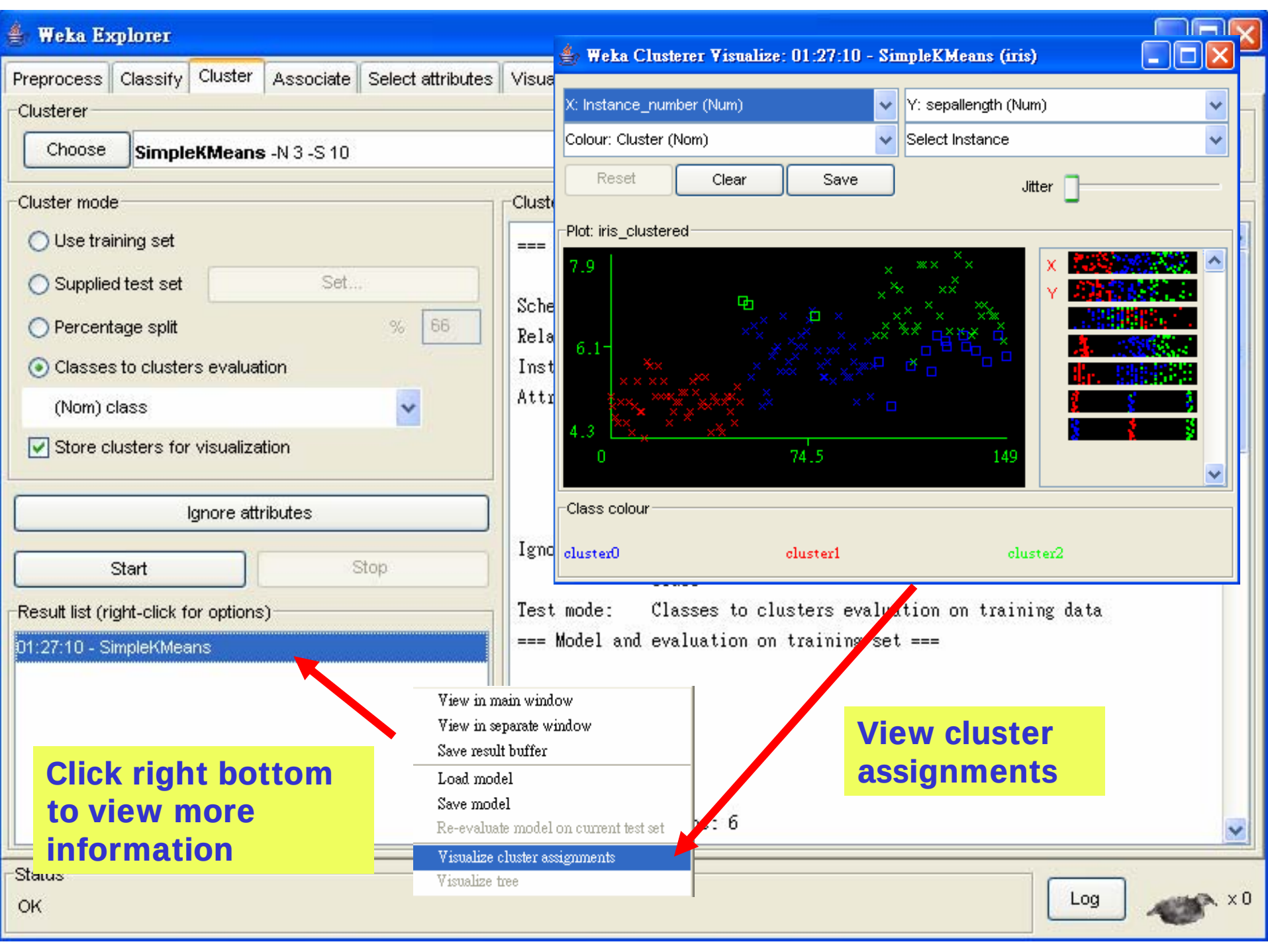

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Clusterer
Choose
EM -I 100 -N -1 -S 100 -M 1.0E-6
clustering data
Cluster mode
Use training set
Supplied test set
Set...
Percentage split
%
66
Classes to clusters evaluation
(Nom) class
Store clusters for visualization
Ignore attributes
Start
Stop
Result list (right-click for options)
Clusterer output
Status
OK
Log
x 0

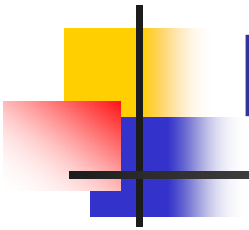

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Clusterer
weka
clusterers
Cobweb
EM
FarthestFirst
MakeDensityBasedClusterer
SimpleKMeans
Clusterer output
Choose clustering algorithm
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Clusterer
Choose
SimpleKMeans -N 2 -S 10
Click here to set parameter
Cluster mode
Use training set
Supplied test set
Set...
Percentage split
Classes to clusters evaluation
(Nom) class
Store clusters for visualization
Ignore attributes
Start
Result list (right-click for options)
weka.gui.GenericObjectEditor
weka.clusterers.SimpleKMeans
About
Cluster data using the k means algorithm
More
numClusters
3
seed
10
Open...
Save...
OK
Cancel
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Clusterer
Choose
SimpleKMeans -N 3 -S 10
Cluster mode
Use training set
Supplied test set
Set...
Percentage split
%
66
Classes to clusters evaluation
(Nom) class
Store clusters for visualization
Ignore attributes
Start
Stop
Result list (right-click for options)
Clusterer output
Start to perform clustering
Status
OK
Log
x 0

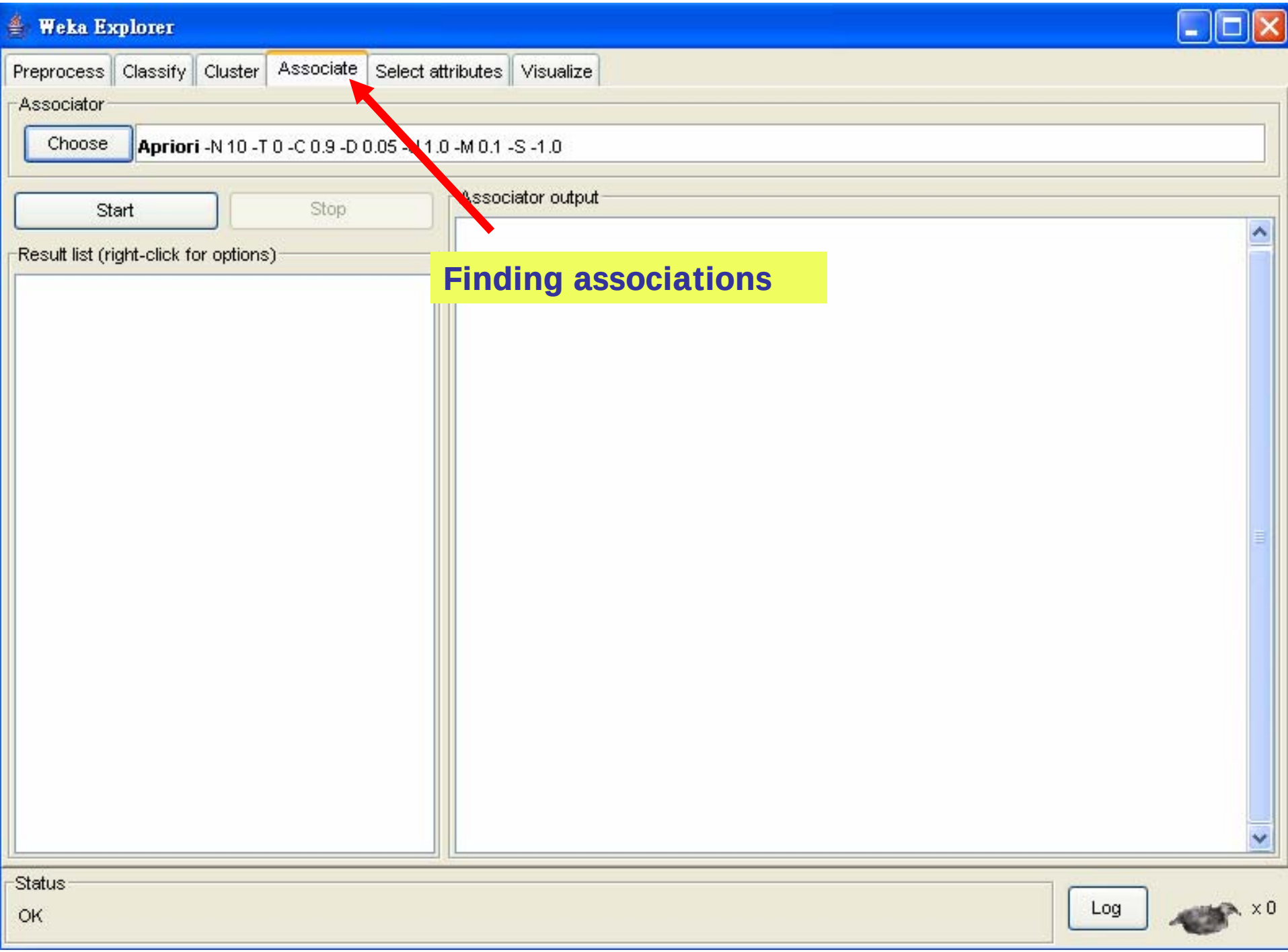

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Clusterer
Choose
SimpleKMeans -N 3 -S 10
Cluster mode
Use training set
Supplied test set
Set...
Percentage split
%
66
Classes to clusters evaluation
(Nom) class
Store clusters for visualization
Ignore attributes
Start
Stop
Result list (right-click for options)
01:27:10 - SimpleKMeans
Clusterer output
0 61 (41%)
1 50 (33%)
2 39 (26%)
Class attribute: class
Classes to Clusters:
0 1 2 <-- assigned to cluster
0 50 0 | Iris-setosa
47 0 3 | Iris-versicolor
14 0 36 | Iris-virginica
Cluster 0 <-- Iris-versicolor
Cluster 1 <-- Iris-setosa
Cluster 2 <-- Iris-virginica
Incorrectly clustered instances : 17.0 11.3333 %
Status
OK
Log
x 0

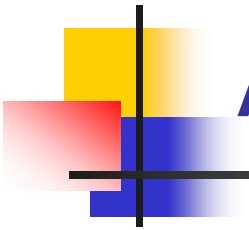

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Clusterer
Choose
SimpleKMeans -N 3 -S 10
Cluster mode
Use training set
Supplied test set
Set...
Percentage split
%
66
Classes to clusters evaluation
(Nom) class
Store clusters for visualization
Ignore attributes
Start
Stop
Result list (right-click for options)
01:27:10 - SimpleKMeans
Test mode: Classes to clusters evaluation on training data
=== Model and evaluation on training set ===
Weka Clusterer Visualize: 01:27:10 - SimpleKMeans (iris)
X: Instance_number (Num)
Y: sepallength (Num)
Colour: Cluster (Nom)
Select Instance
Reset
Clear
Save
Jitter
Plot: iris_clustered
7.9
6.1
4.3
0
74.5
149
X
Y
Class colour
cluster0
cluster1
cluster2
View in main window
View in separate window
Save result buffer
Load model
Save model
Re-evaluate model on current test set
Visualize cluster assignments
Visualize tree
Click right bottom to view more information
View cluster assignments
Status
OK
Log
x 0



Finding associations

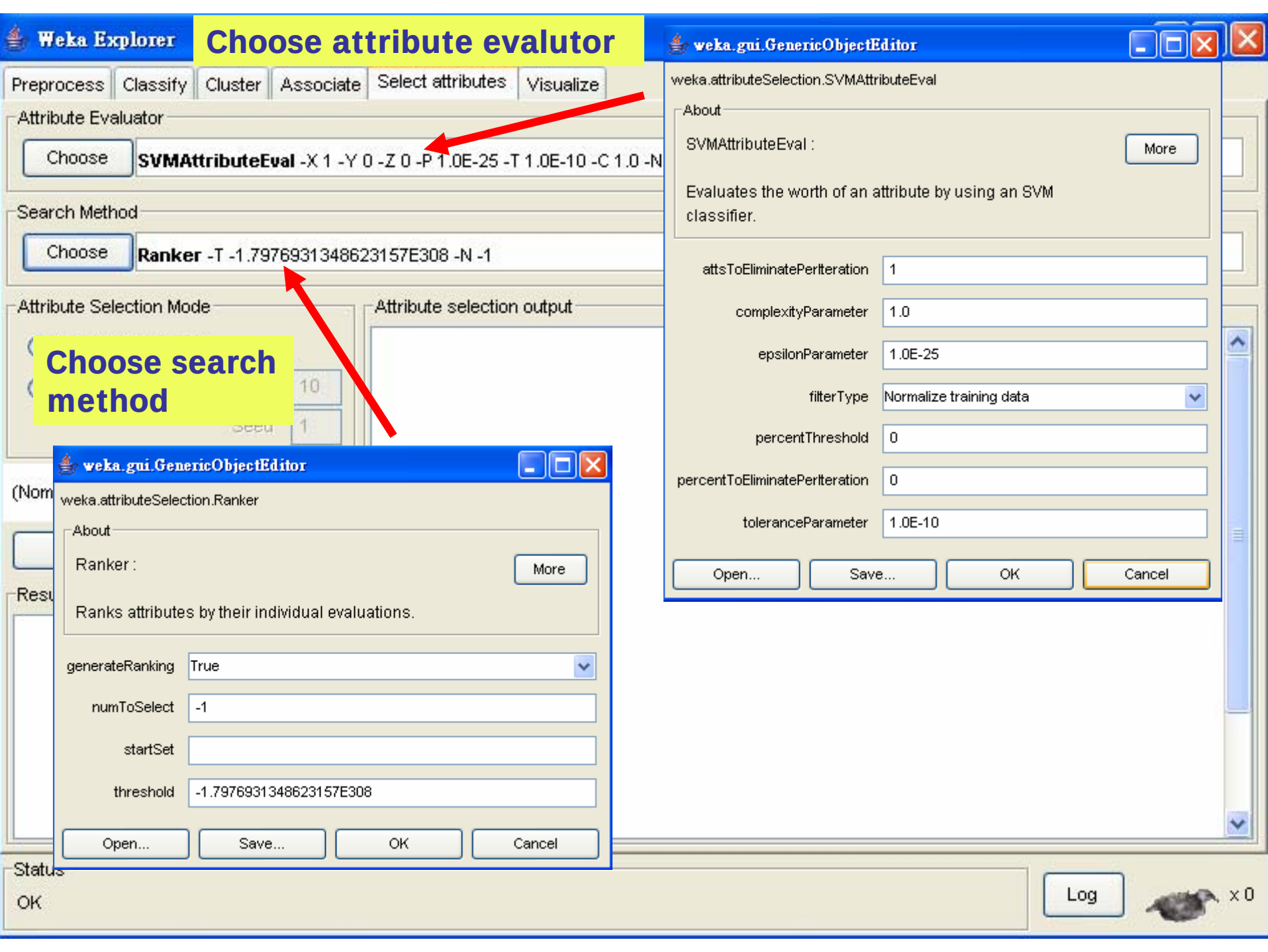
- WEKA contains an implementation of the Apriori algorithm for learning association rules
  - Works only with discrete data
- Can identify statistical dependencies between groups of attributes:
  - milk, butter $\Rightarrow$ bread, eggs (with confidence 0.9 and support 2000)
- Apriori can compute all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

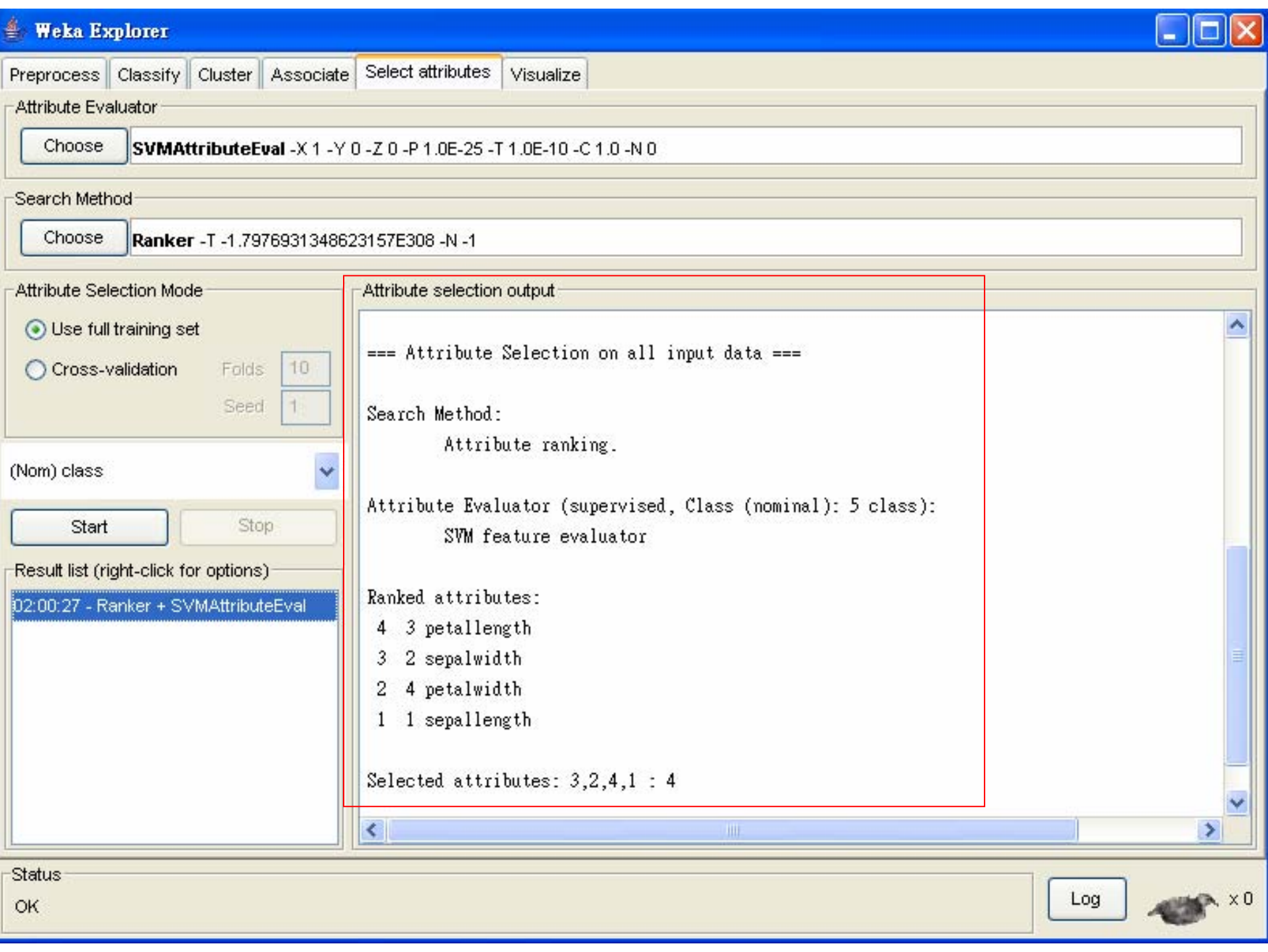

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Associator
Choose
Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0
Start
Stop
Result list (right-click for options)
Associator output
Finding associations
Status
OK
Log
x 0

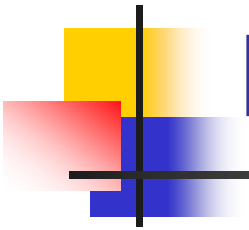


Attribute selection

- 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

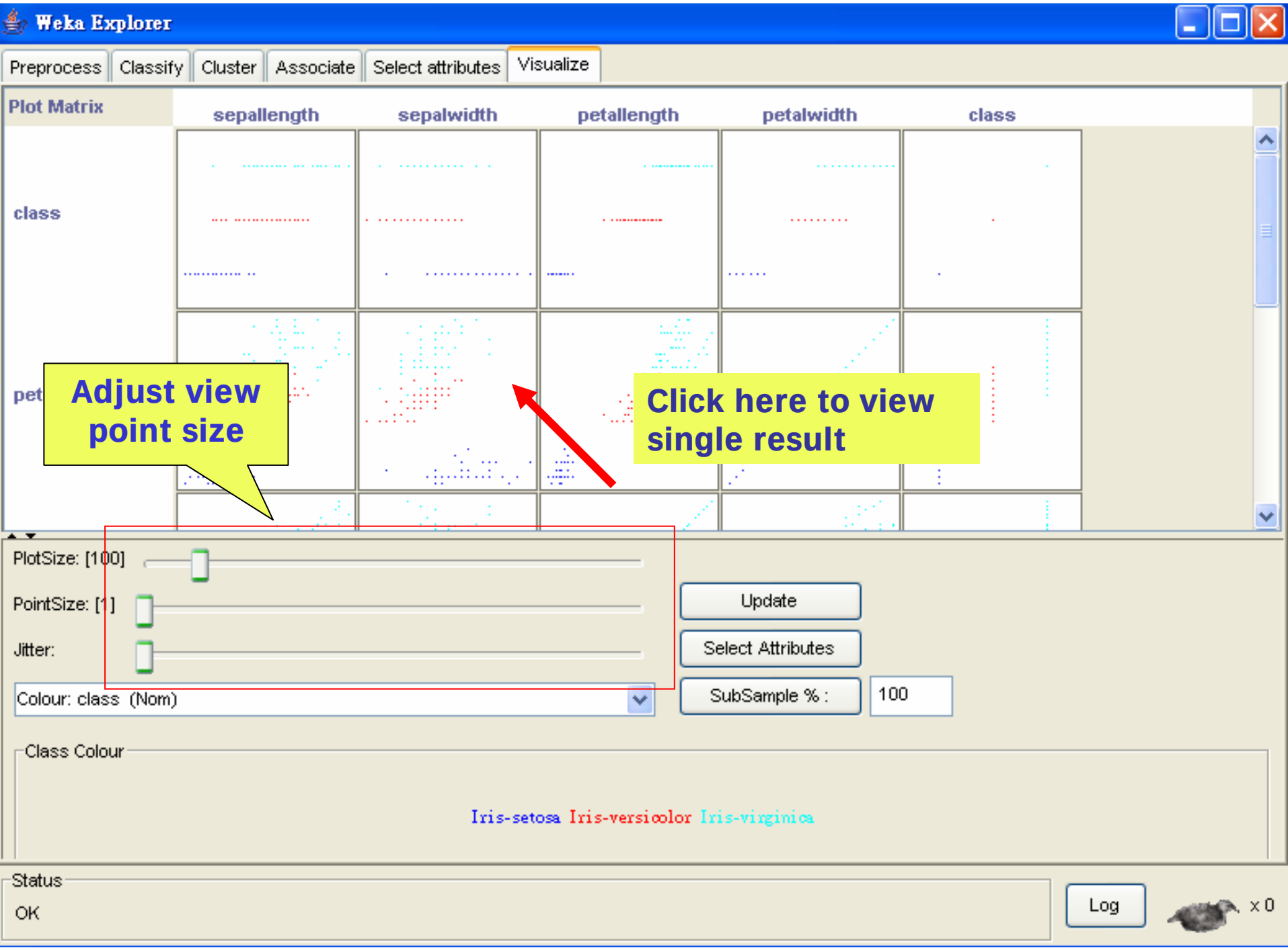

Weka Explorer
Choose attribute evalutor
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Attribute Evaluator
Choose
SVMAttributeEval -X 1 -Y 0 -Z 0 -P 1.0E-25 -T 1.0E-10 -C 1.0 -N
Search Method
Choose
Ranker -T -1.7976931348623157E308 -N -1
Attribute Selection Mode
Attribute selection output
Choose search method
weka.gui.GenericObjectEditor
weka.attributeSelection.SVMAttributeEval
About
SVMAttributeEval :
More
Evaluates the worth of an attribute by using an SVM classifier.
attsToEliminatePerIteration
1
complexityParameter
1.0
epsilonParameter
1.0E-25
filterType
Normalize training data
percentThreshold
0
percentToEliminatePerIteration
0
toleranceParameter
1.0E-10
Open...
Save...
OK
Cancel
weka.gui.GenericObjectEditor
weka.attributeSelection.Ranker
About
Ranker :
More
Ranks attributes by their individual evaluations.
generateRanking
True
numToSelect
-1
startSet
threshold
-1.7976931348623157E308
Open...
Save...
OK
Cancel
Status
OK
Log
x 0

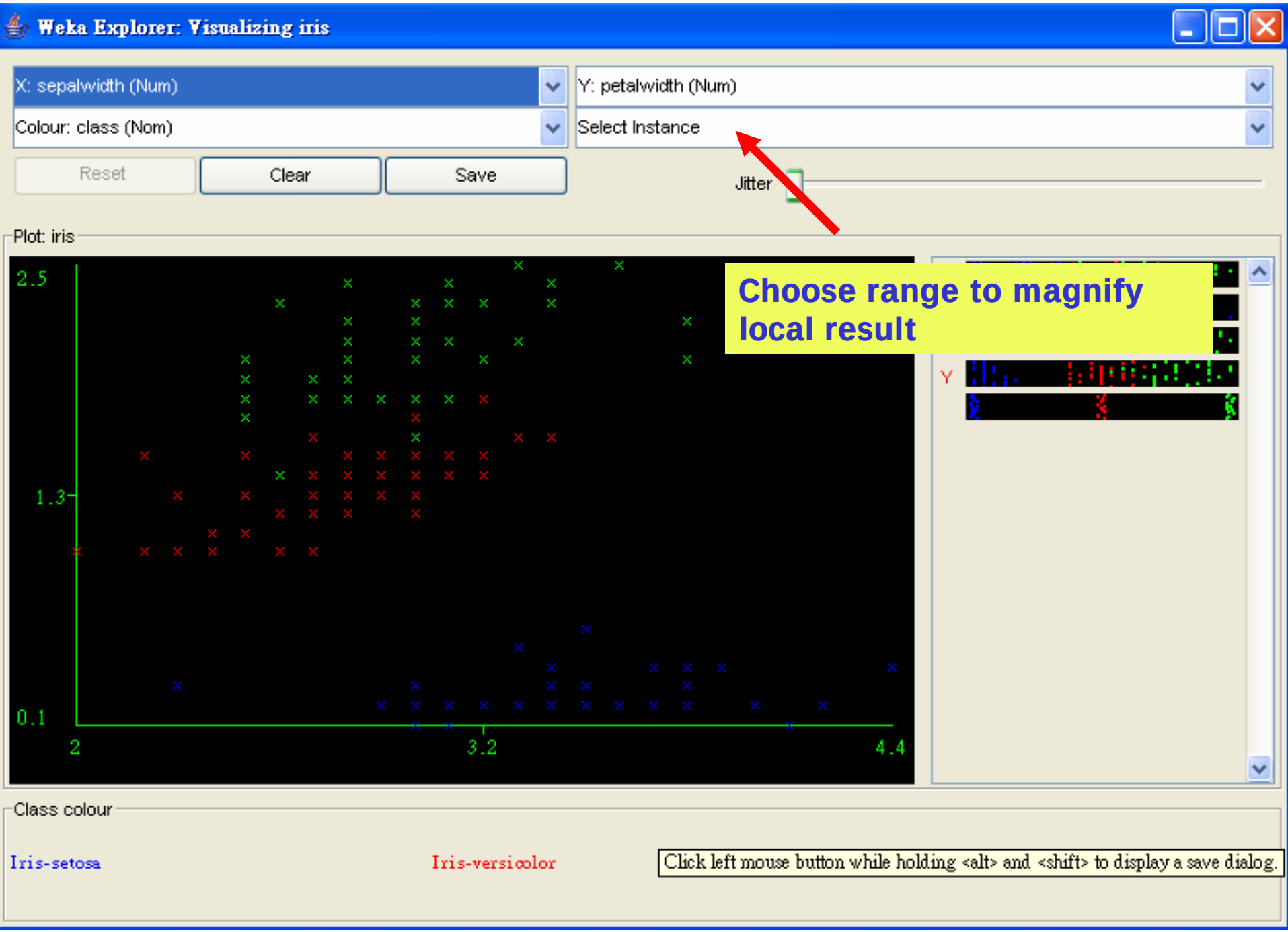

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Attribute Evaluator
Choose
SVMAttributeEval -X 1 -Y 0 -Z 0 -P 1.0E-25 -T 1.0E-10 -C 1.0 -N 0
Search Method
Choose
Ranker -T -1.7976931348623157E308 -N -1
Attribute Selection Mode
Use full training set
Cross-validation
Folds
10
Seed
1
(Nom) class
Start
Stop
Result list (right-click for options)
02:00:27 - Ranker + SVMAttributeEval
Attribute selection output
=== Attribute Selection on all input data ===
Search Method:
Attribute ranking.
Attribute Evaluator (supervised, Class (nominal): 5 class):
SVM feature evaluator
Ranked attributes:
4 3 petallength
3 2 sepalwidth
2 4 petalwidth
1 1 sepallength
Selected attributes: 3,2,4,1 : 4
Status
OK
Log
x 0

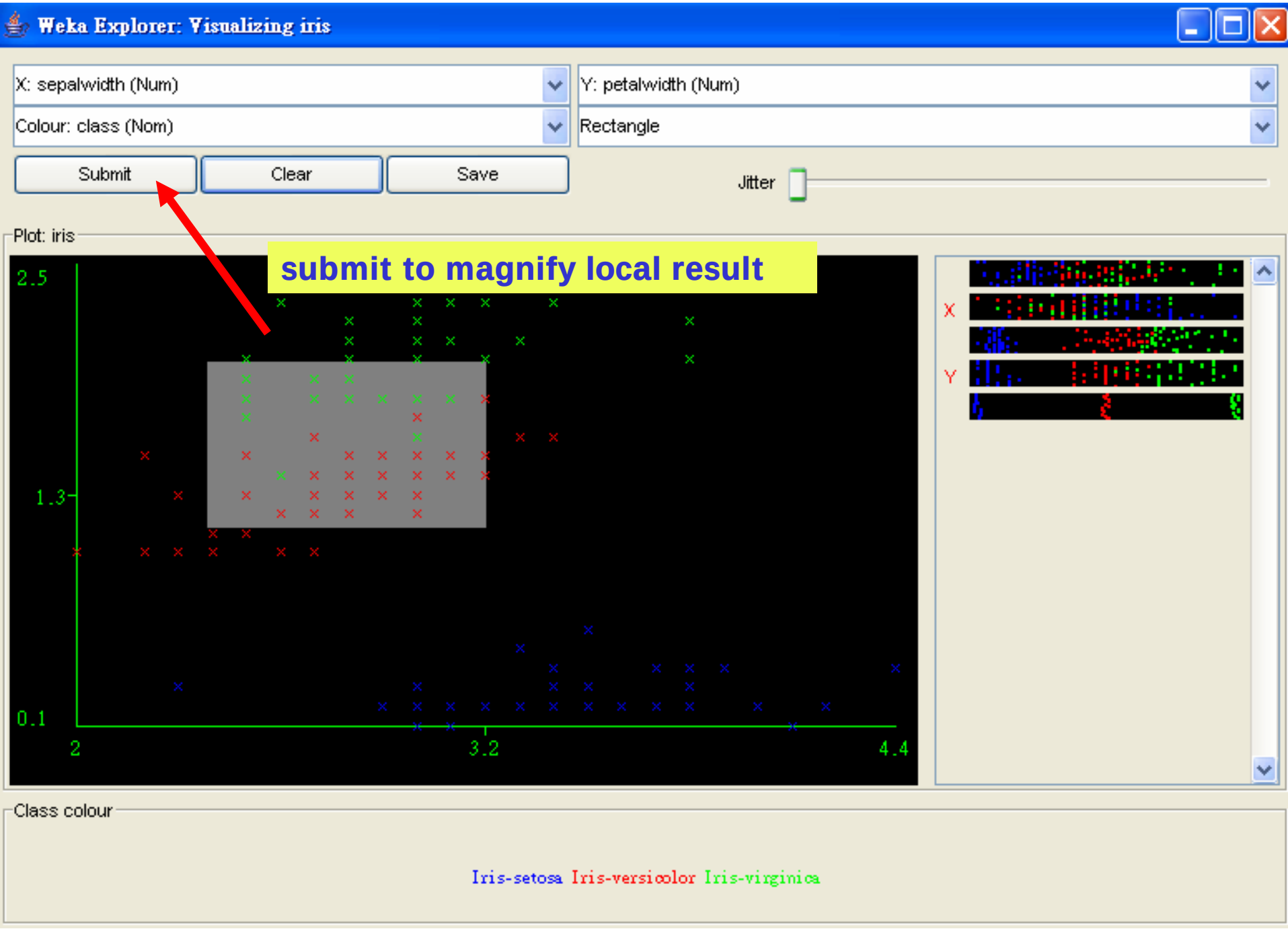


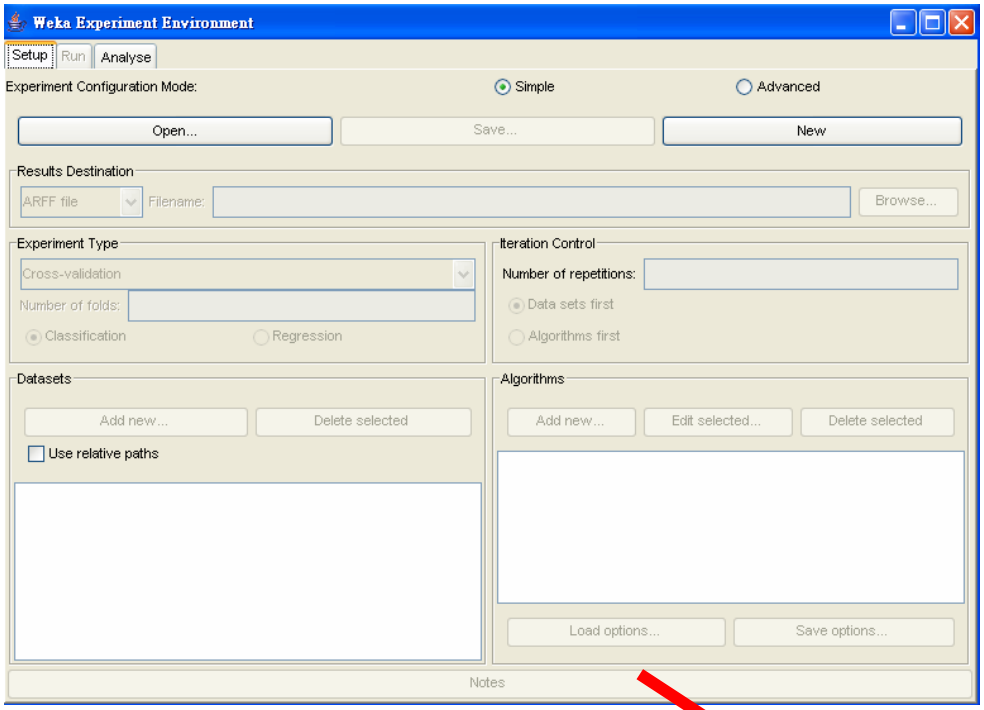
Data visualization

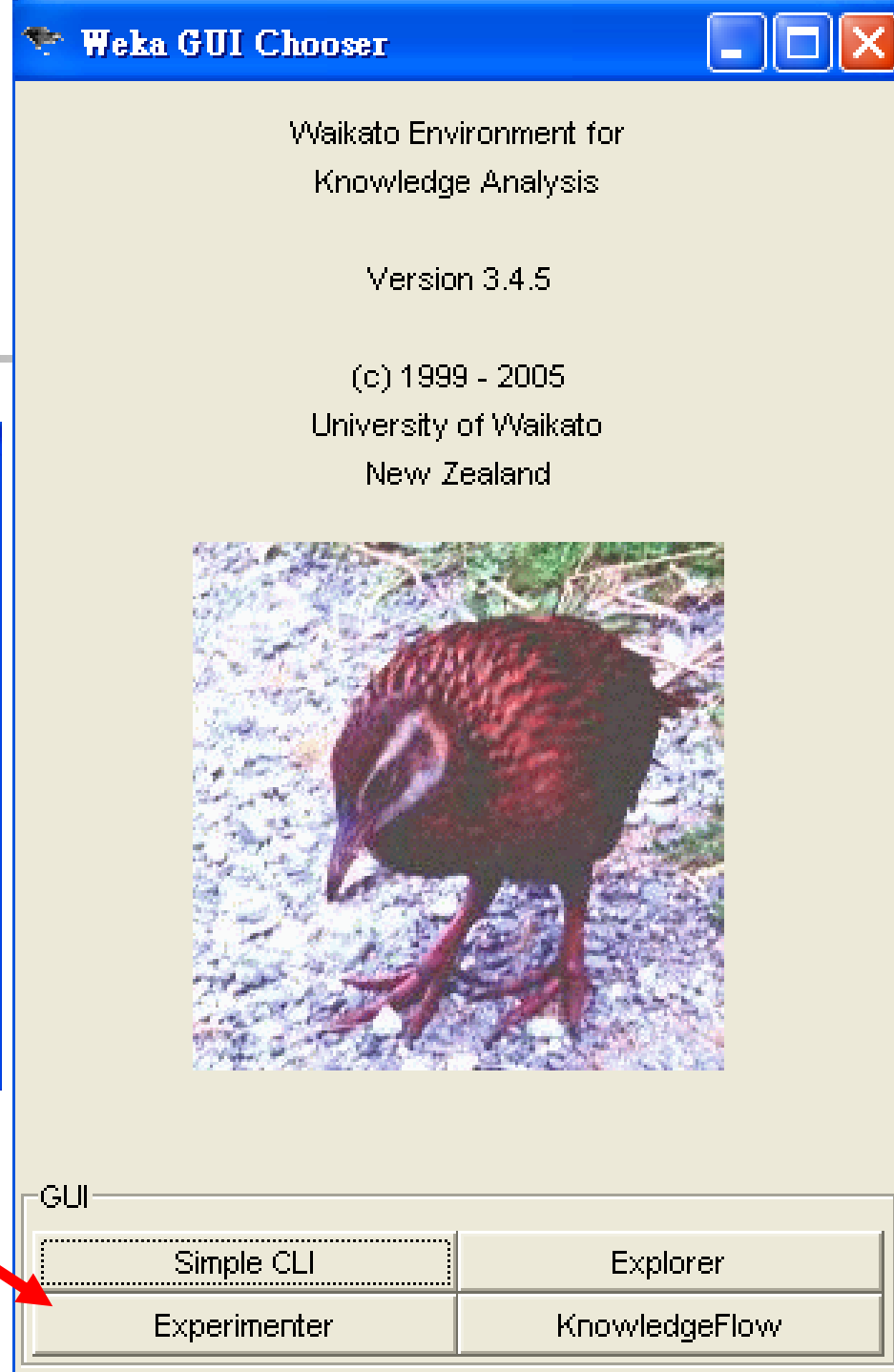
- Visualization very useful in practice: e.g. helps to determine difficulty of the learning problem
- WEKA can visualize single attributes (1-d) and pairs of attributes (2-d)
  - To do: rotating 3-d visualizations (Xgobi-style)
- Color-coded class values
- “Jitter” option to deal with nominal attributes (and to detect “hidden” data points)
- “Zoom-in” function

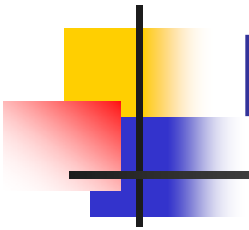

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Plot Matrix
sepallength
sepalwidth
petallength
petalwidth
class
class
pet
Adjust view point size
Click here to view single result
PlotSize: [100]
PointSize: [1]
Jitter:
Update
Select Attributes
SubSample % :
100
Colour: class (Nom)
Class Colour
Iris-setosa Iris-versicolor Iris-virginica
Status
OK
Log
x 0


Weka Explorer: Visualizing iris
X: sepalwidth (Num)
Y: petalwidth (Num)
Colour: class (Nom)
Select Instance
Reset
Clear
Save
Jitter
Plot: iris
2.5
1.3
0.1
2
3.2
4.4
Y
Choose range to magnify local result
Class colour
Iris-setosa
Iris-versicolor
Click left mouse button while holding <alt> and <shift> to display a save dialog.


Weka Explorer: Visualizing iris
X: sepalwidth (Num)
Y: petalwidth (Num)
Colour: class (Nom)
Rectangle
Submit
Clear
Save
Jitter
Plot: iris
submit to magnify local result
2.5
1.3
0.1
2
3.2
4.4
X
Y
Class colour
Iris-setosa Iris-versicolor Iris-virginica

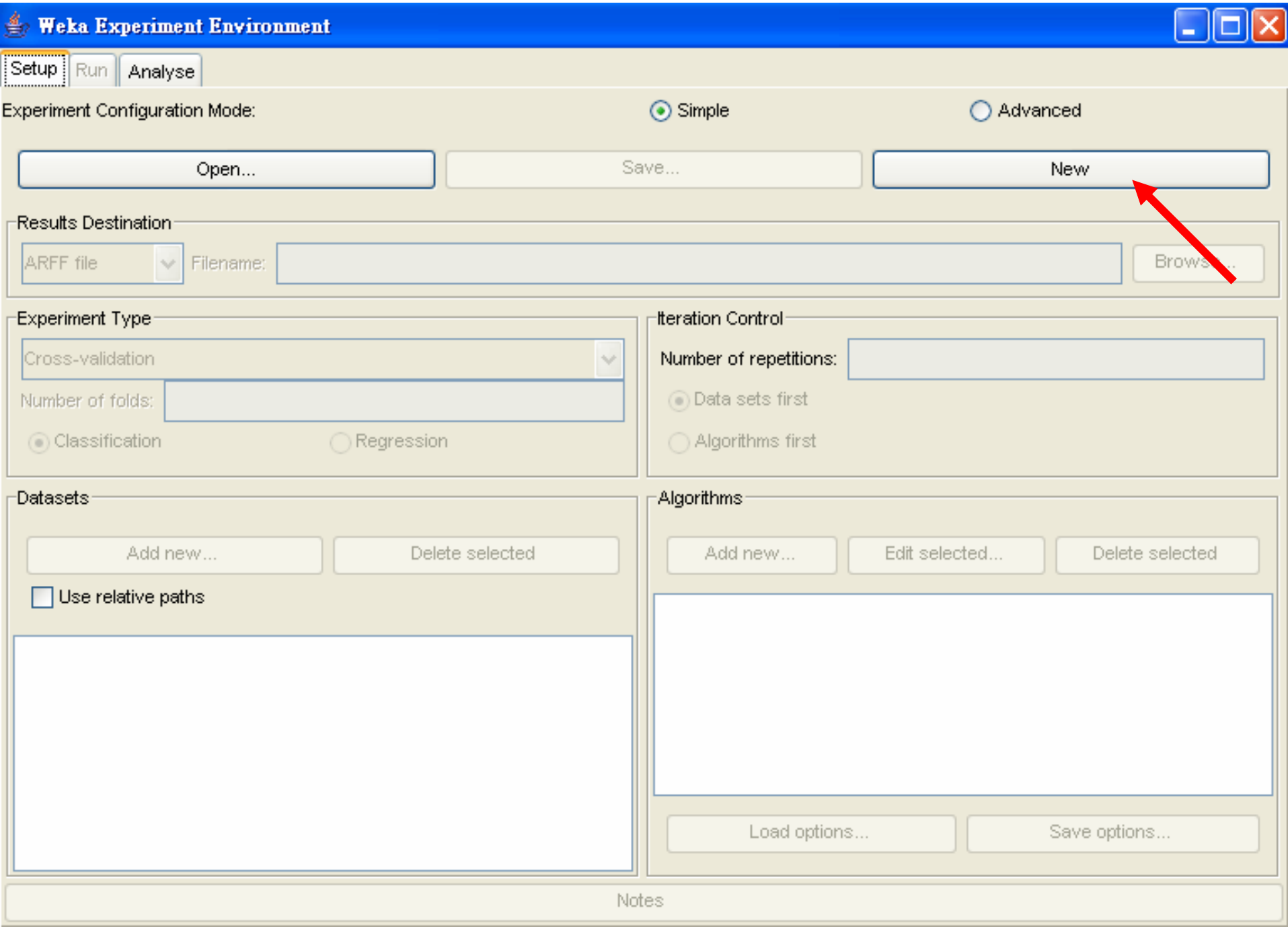

Weka Experiment Environment
Setup
Run
Analyse
Experiment Configuration Mode:
Simple
Advanced
Open...
Save...
New
Results Destination
ARFF file
Filename:
Browse...
Experiment Type
Cross-validation
Number of folds:
Classification
Regression
Iteration Control
Number of repetitions:
Data sets first
Algorithms first
Datasets
Add new...
Delete selected
Use relative paths
Algorithms
Add new...
Edit selected...
Delete selected
Load options...
Save options...
Notes

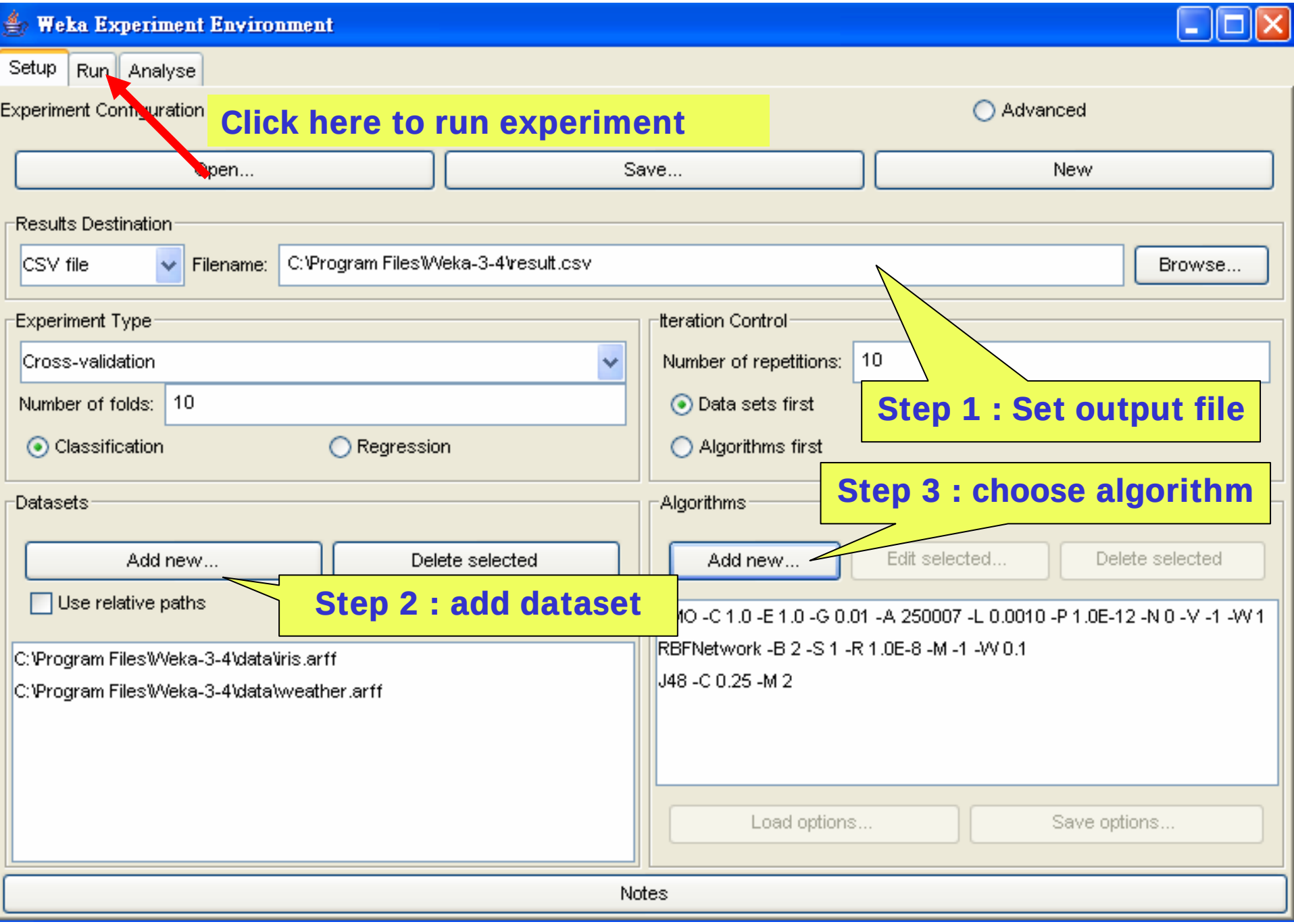

Weka GUI Chooser
Waikato Environment for
Knowledge Analysis
Version 3.4.5
(c) 1999 - 2005
University of Waikato
New Zealand
GUI
Simple CLI
Explorer
Experimenter
KnowledgeFlow

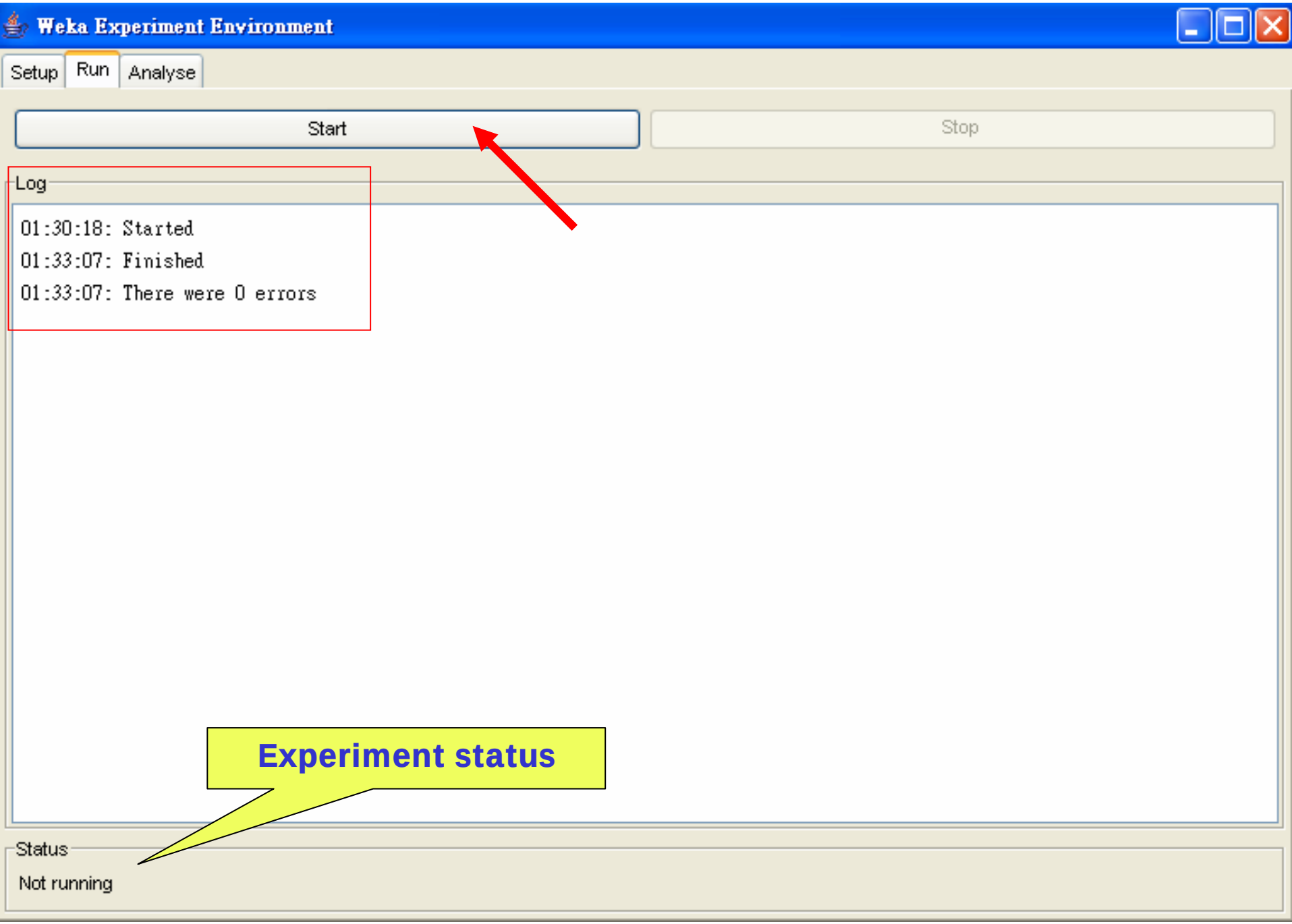


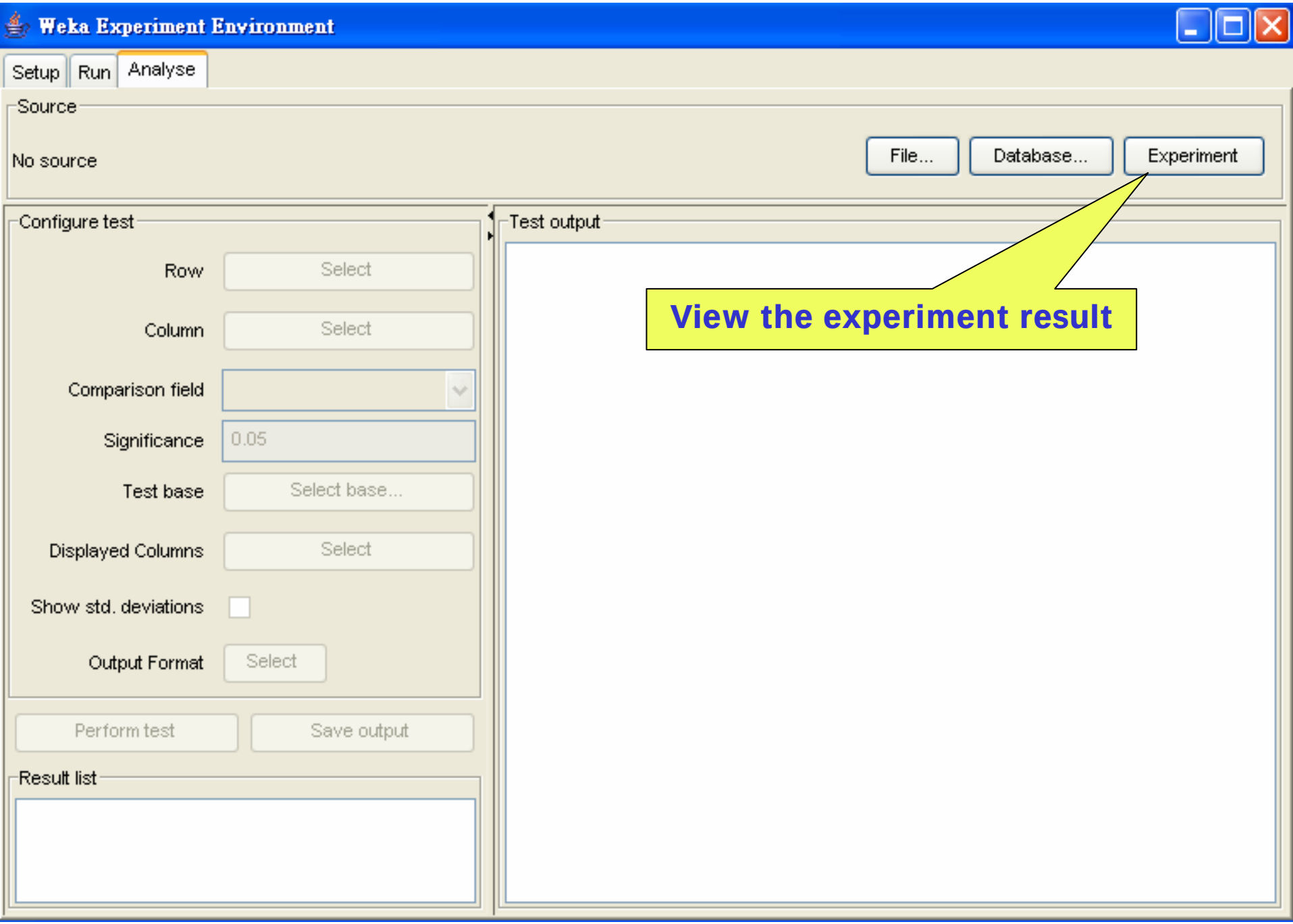
Performing experiments

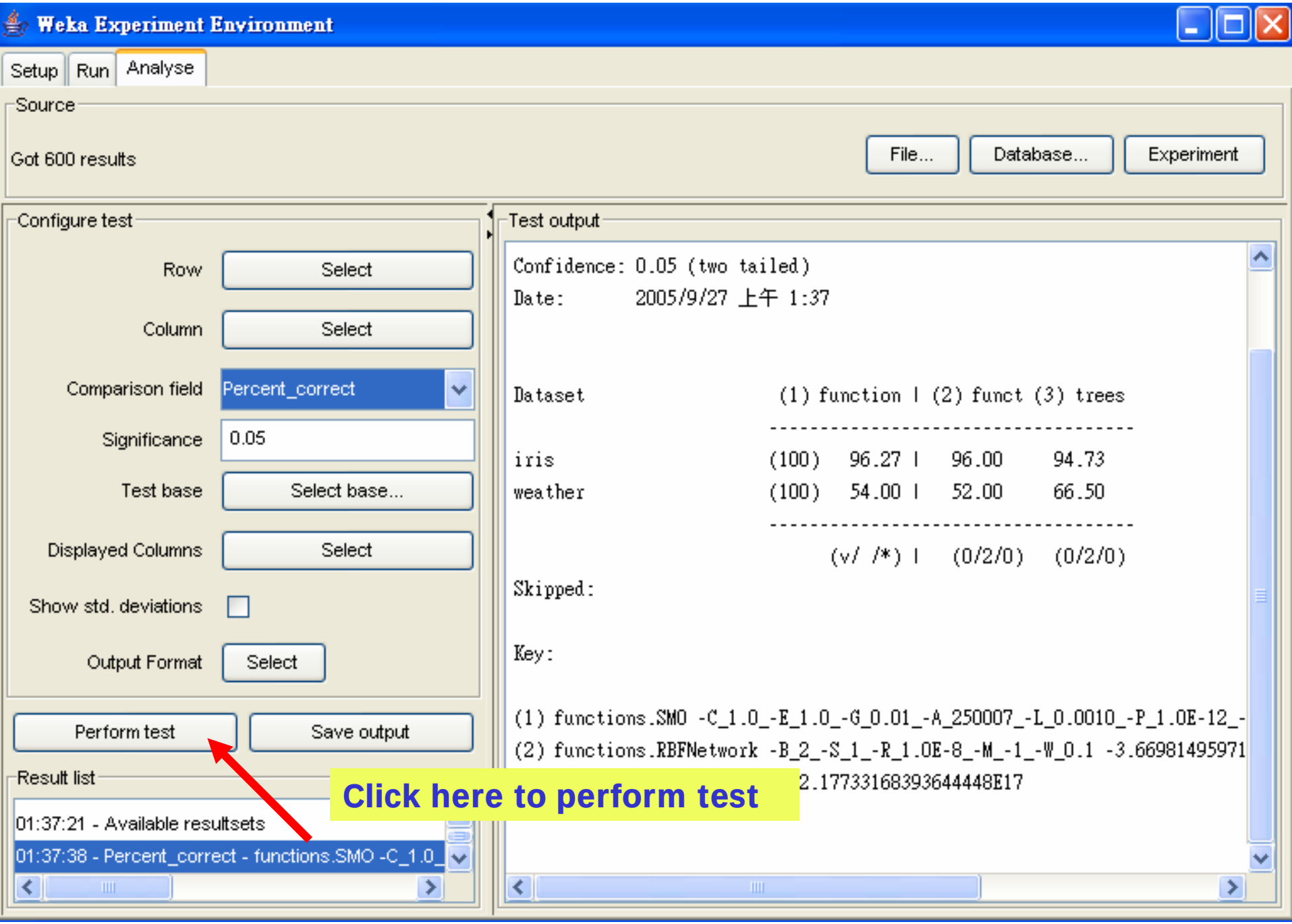
- Experimenter makes it easy to compare the performance of different learning schemes
- For classification and regression problems
- Results can be written into file or database
- Evaluation options: cross-validation, learning curve, hold-out
- Can also iterate over different parameter settings
- Significance-testing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


Weka Experiment Environment
Setup
Run
Analyse
Experiment Configuration Mode:
Simple
Advanced
Open...
Save...
New
Results Destination
ARFF file
Filename:
Browse...
Experiment Type
Cross-validation
Number of folds:
Classification
Regression
Iteration Control
Number of repetitions:
Data sets first
Algorithms first
Datasets
Add new...
Delete selected
Use relative paths
Algorithms
Add new...
Edit selected...
Delete selected
Load options...
Save options...
Notes


Weka Experiment Environment
Setup
Run
Analyse
Experiment Configuration
Click here to run experiment
Advanced
Open...
Save...
New
Results Destination
CSV file
Filename:
C:\Program Files\Weka-3-4\result.csv
Browse...
Experiment Type
Cross-validation
Number of folds:
10
Classification
Regression
Iteration Control
Number of repetitions:
10
Data sets first
Algorithms first
Step 1 : Set output file
Step 3 : choose algorithm
Datasets
Add new...
Delete selected
Use relative paths
Step 2 : add dataset
C:\Program Files\Weka-3-4\data\iris.arff
C:\Program Files\Weka-3-4\data\weather.arff
Algorithms
Add new...
Edit selected...
Delete selected
MO -C 1.0 -E 1.0 -G 0.01 -A 250007 -L 0.0010 -P 1.0E-12 -N 0 -V -1 -W 1
RBFNetwork -B 2 -S 1 -R 1.0E-8 -M -1 -W 0.1
J48 -C 0.25 -M 2
Load options...
Save options...
Notes


Weka Experiment Environment
Setup
Run
Analyse
Start
Stop
Log
01:30:18: Started
01:33:07: Finished
01:33:07: There were 0 errors
Experiment status
Status
Not running


Weka Experiment Environment
Setup
Run
Analyse
Source
No source
File...
Database...
Experiment
Configure test
Row
Select
Column
Select
Comparison field
Significance
0.05
Test base
Select base...
Displayed Columns
Select
Show std. deviations
Output Format
Select
Perform test
Save output
Result list
Test output
View the experiment result


Weka Experiment Environment
Setup
Run
Analyse
Source
Got 600 results
File...
Database...
Experiment
Configure test
Row
Select
Column
Select
Comparison field
Percent_correct
Significance
0.05
Test base
Select base...
Displayed Columns
Select
Show std. deviations
Output Format
Select
Perform test
Save output
Result list
01:37:21 - Available resultsets
01:37:38 - Percent_correct - functions.SMO -C_1.0_
Click here to perform test
Test output
Confidence: 0.05 (two tailed)
Date: 2005/9/27 上午 1:37
Dataset (1) function | (2) funct (3) trees
iris (100) 96.27 | 96.00 94.73
weather (100) 54.00 | 52.00 66.50
(v/ /*) | (0/2/0) (0/2/0)
Skipped:
Key:
(1) functions.SMO -C_1.0_-E_1.0_-G_0.01_-A_250007_-L_0.0010_-P_1.0E-12_-
(2) functions.RBFNetwork -B_2_-S_1_-R_1.0E-8_-M_-1_-W_0.1 -3.66981495971
2.17733168393644448E17

Weka GUI Chooser

Waikato Environment for
Knowledge Analysis

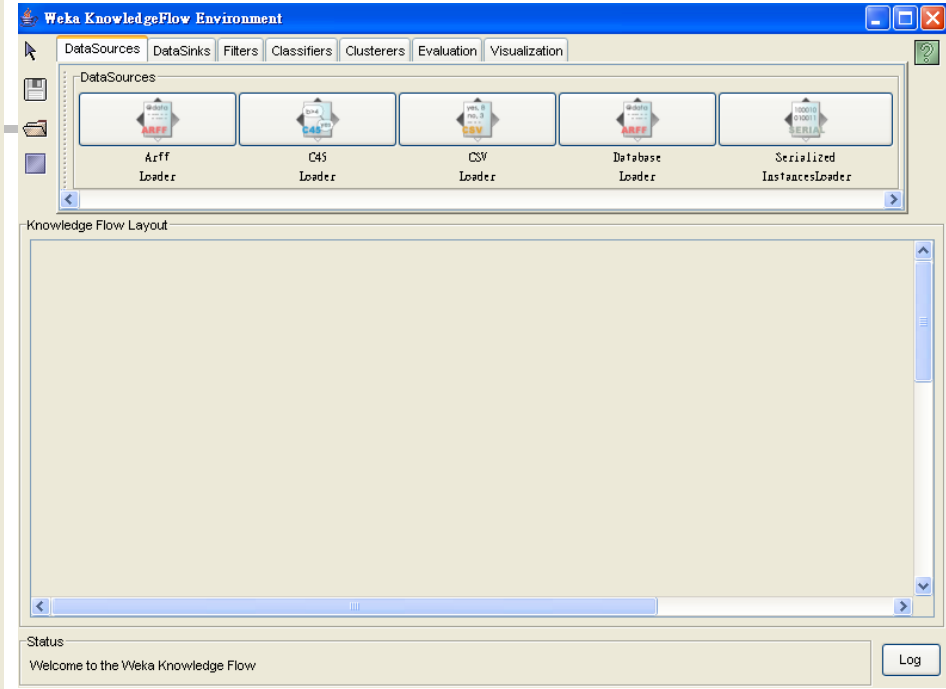
Version 3.4.5

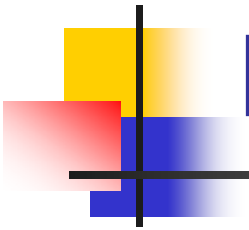
(c) 1999 - 2005
University of Waikato
New Zealand

GUI

Simple CLI | Explorer

Experimenter | KnowledgeFlow

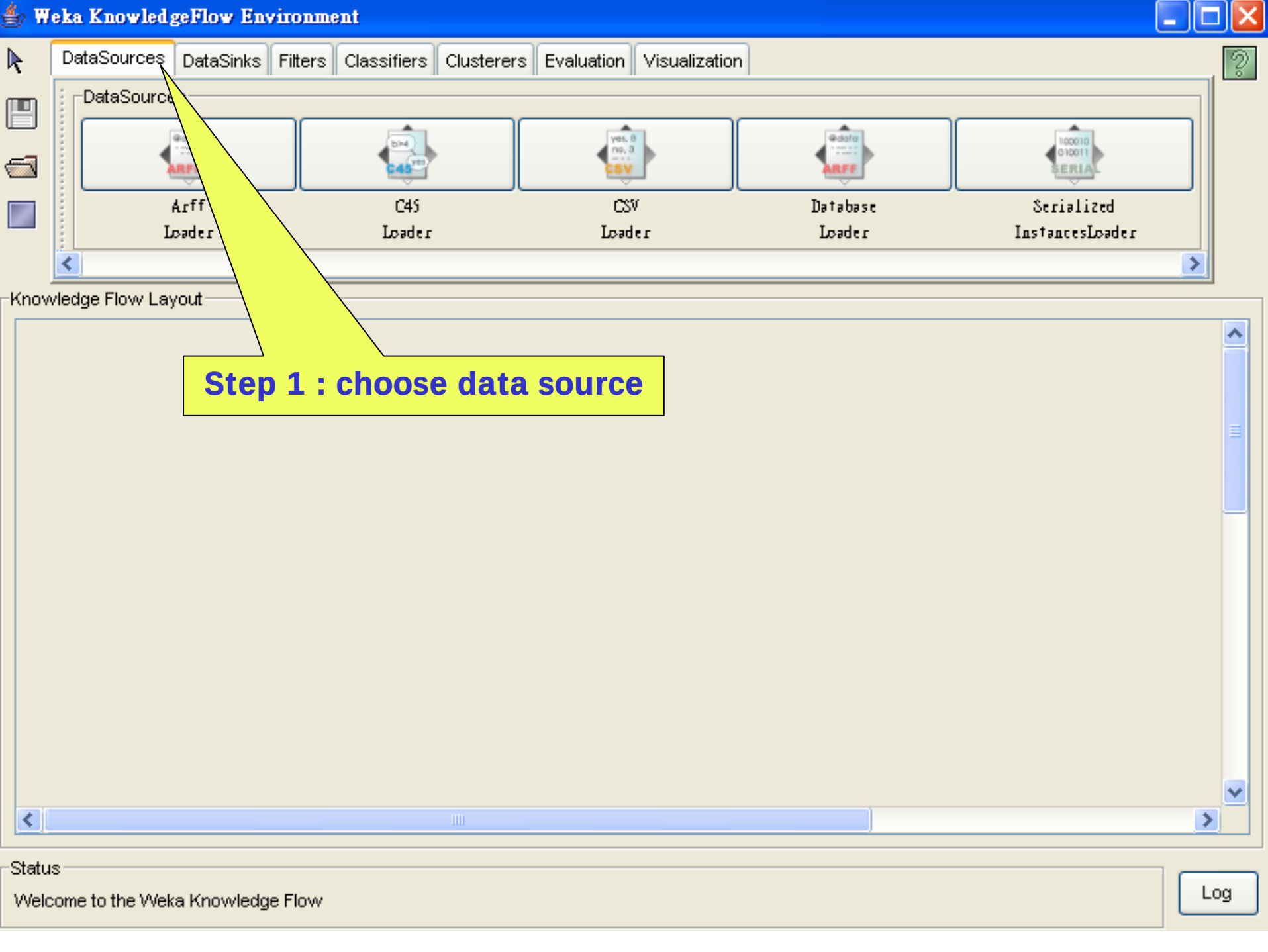


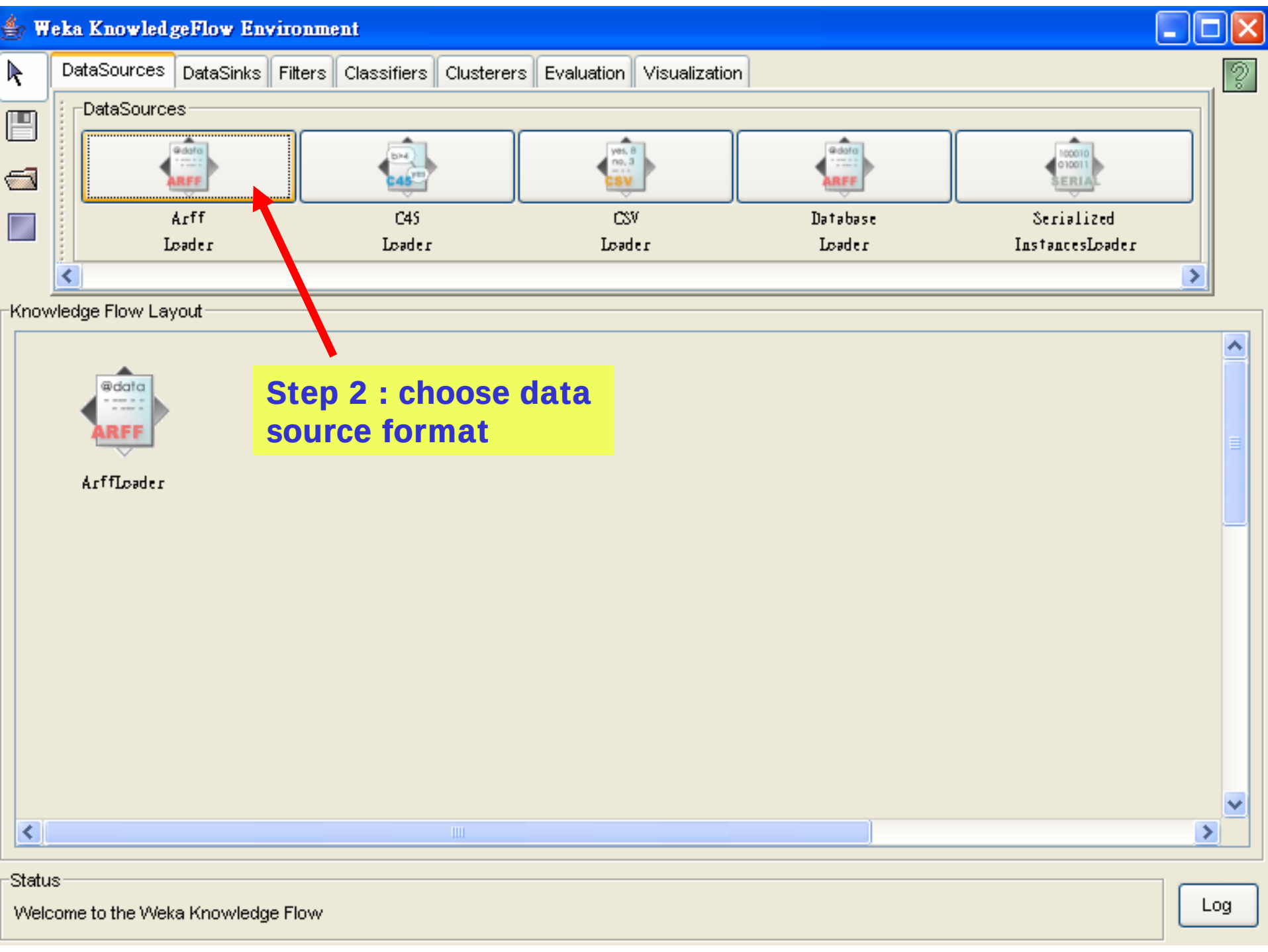


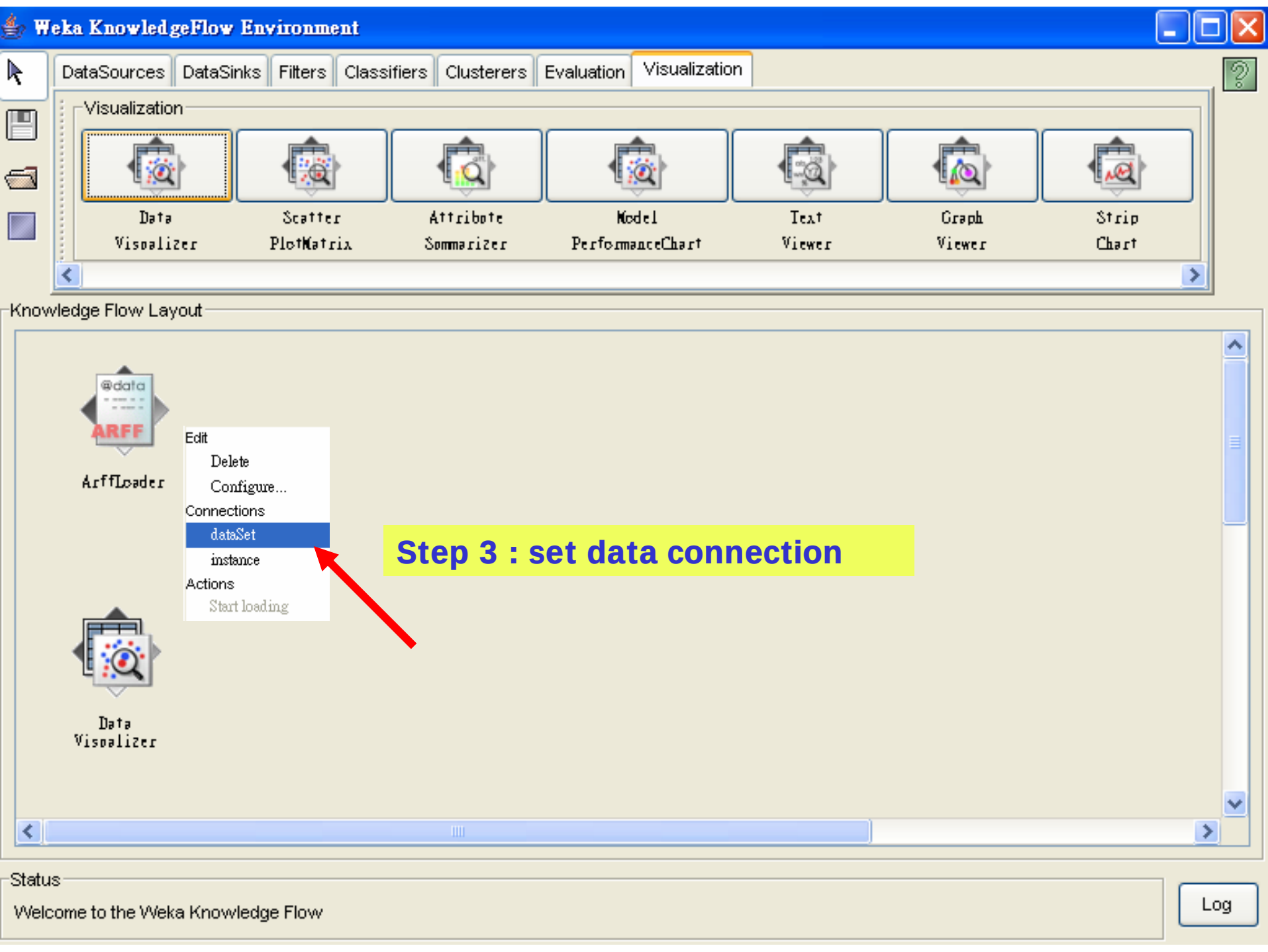
Knowledge Flow GUI

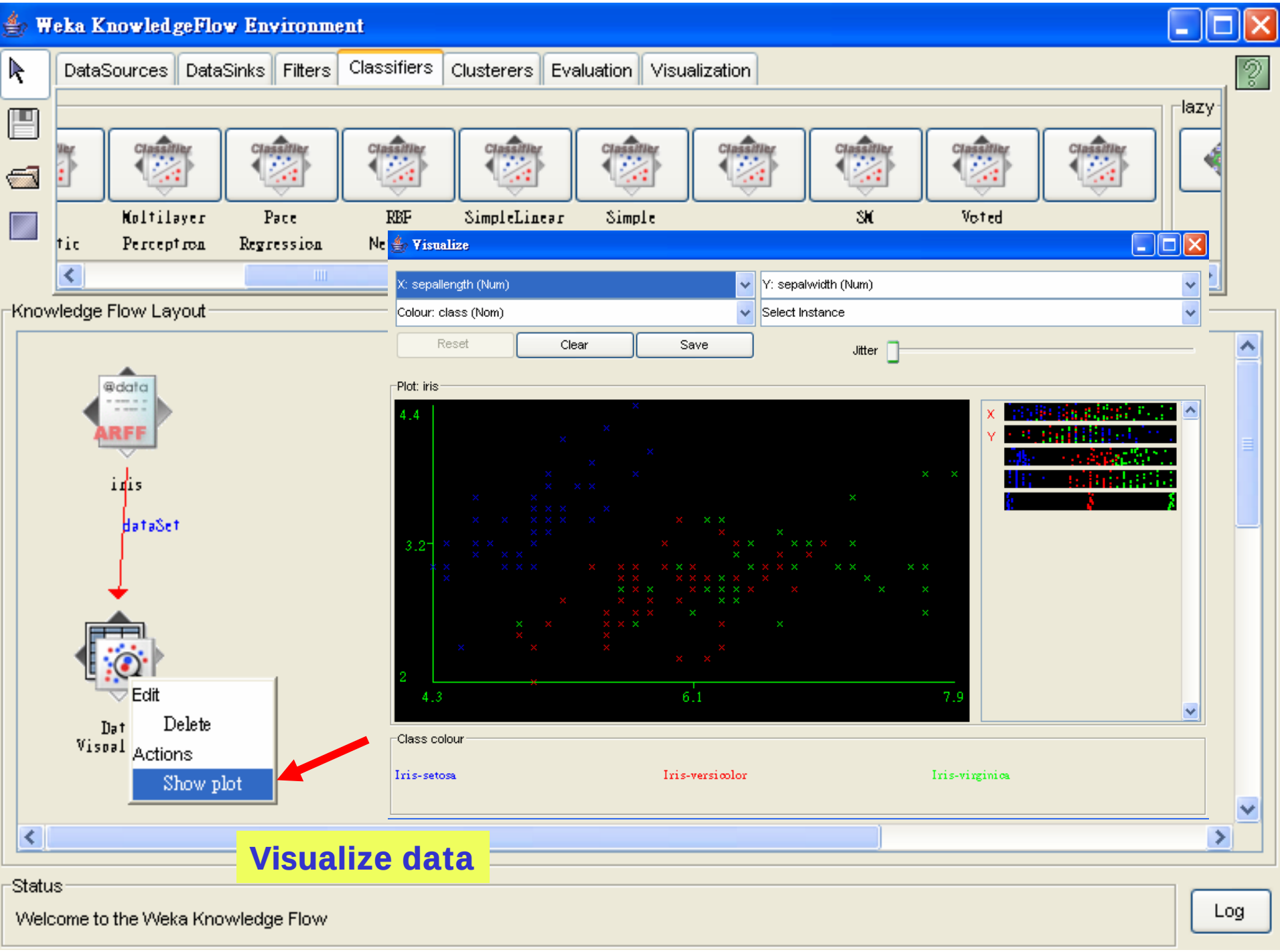
- New graphical user interface for WEKA
- Java-Beans-based interface for setting up and running machine learning experiments
- Data sources, classifiers, etc. are beans and can be connected graphically
- Data “flows” through components: e.g.,

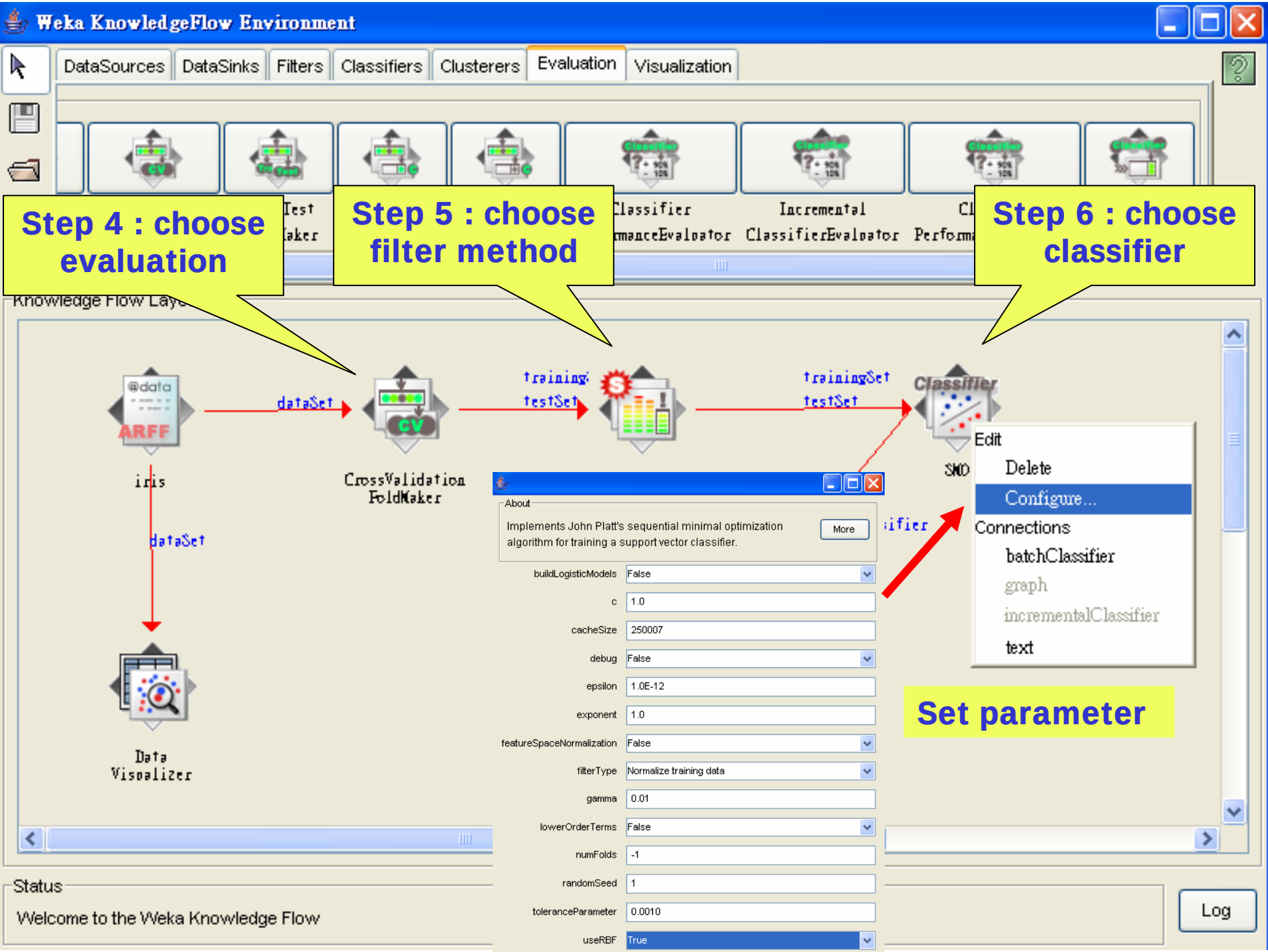
  “data source” -> “filter” -> “classifier” -> “evaluator”
- Layouts can be saved and loaded again later

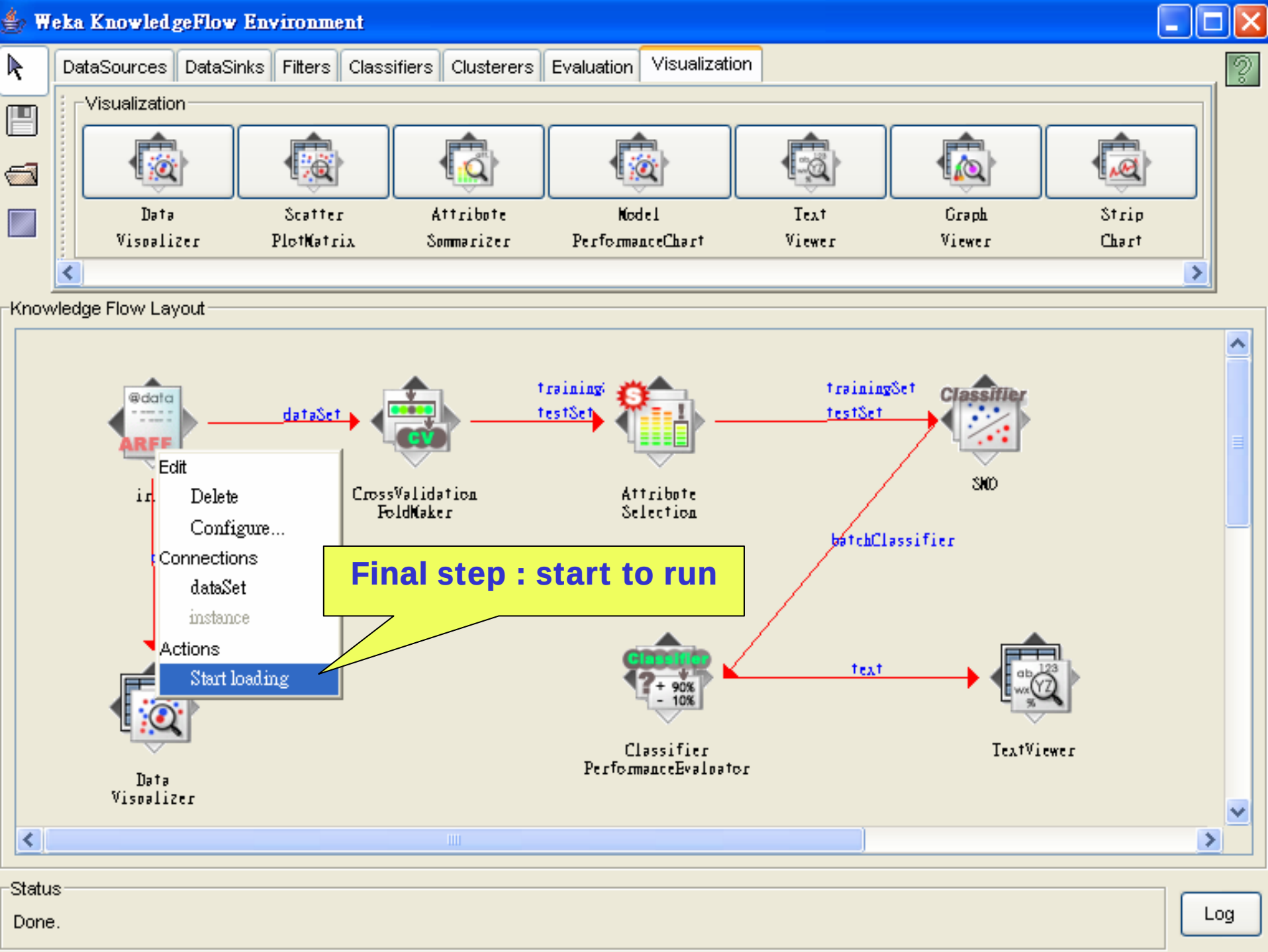

Weka KnowledgeFlow Environment
DataSources
DataSinks
Filters
Classifiers
Clusterers
Evaluation
Visualization
DataSource
Arff Loader
C45 Loader
CSV Loader
Database Loader
Serialized InstancesLoader
Knowledge Flow Layout
Step 1 : choose data source
Status
Welcome to the Weka Knowledge Flow
Log

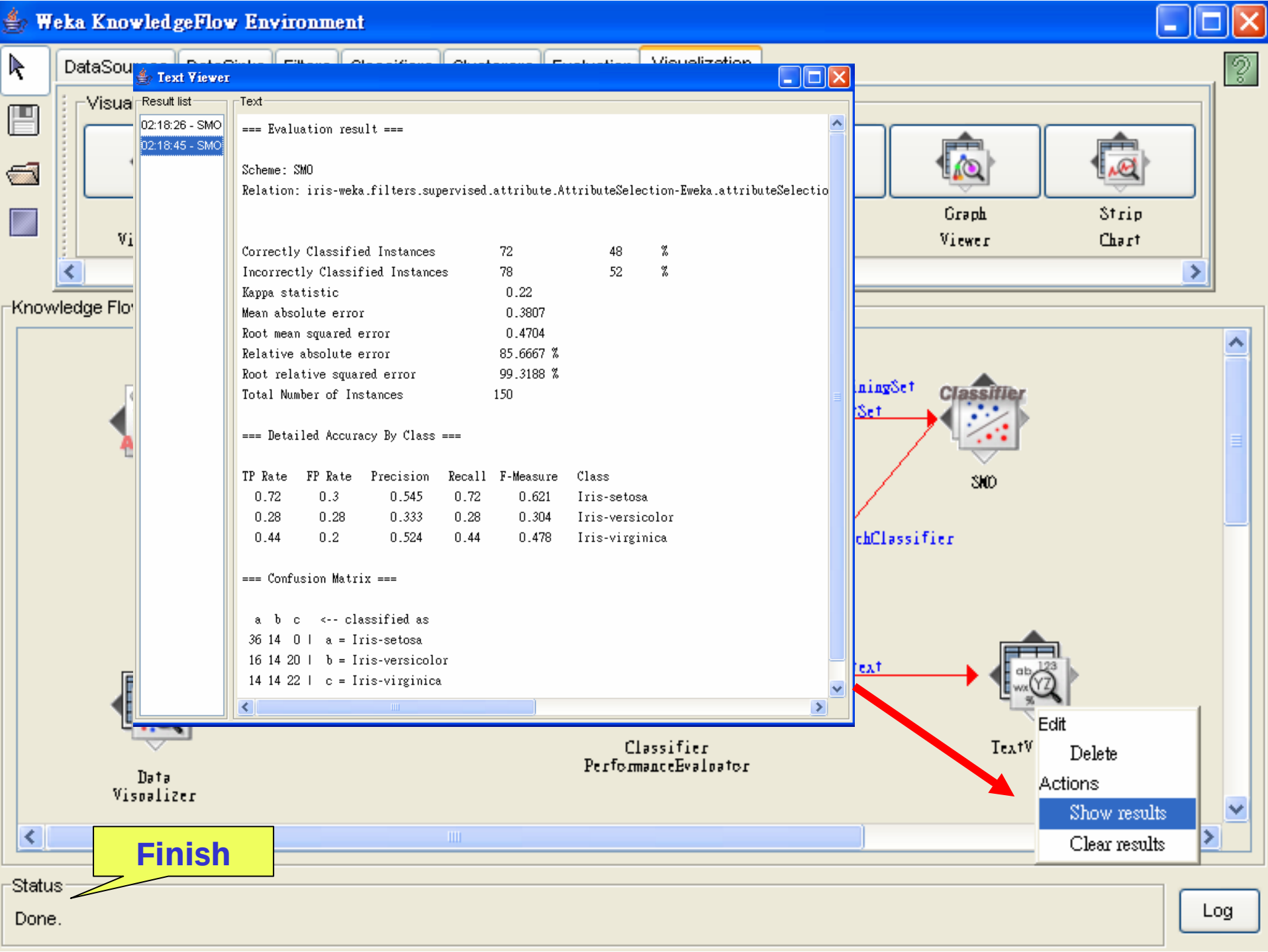

Weka KnowledgeFlow Environment
DataSources
DataSinks
Filters
Classifiers
Clusterers
Evaluation
Visualization
DataSources
Arff Loader
C45 Loader
CSV Loader
Database Loader
Serialized InstancesLoader
Knowledge Flow Layout
ArffLoader
Step 2 : choose data source format
Status
Welcome to the Weka Knowledge Flow
Log

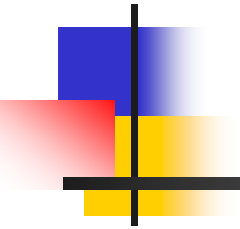

Weka KnowledgeFlow Environment
DataSources
DataSinks
Filters
Classifiers
Clusterers
Evaluation
Visualization
Visualization
Data Visualizer
Scatter PlotMatrix
Attribute Summarizer
Model PerformanceChart
Text Viewer
Graph Viewer
Strip Chart
Knowledge Flow Layout
ArffLoader
Edit
Delete
Configure...
Connections
dataSet
instance
Actions
Start loading
Step 3 : set data connection
Data Visualizer
Status
Welcome to the Weka Knowledge Flow
Log


Weka KnowledgeFlow Environment
DataSources
DataSinks
Filters
Classifiers
Clusterers
Evaluation
Visualization
lazy
Multilayer Perceptron
Pace Regression
RBF
SimpleLinear
Simple
SM
Voted
Visualize
X: sepallength (Num)
Y: sepalwidth (Num)
Colour: class (Nom)
Select Instance
Reset
Clear
Save
Jitter
Plot: iris
4.4
3.2
2
4.3
6.1
7.9
X
Y
Class colour
Iris-setosa
Iris-versicolor
Iris-virginica
Knowledge Flow Layout
@data
ARFF
iris
dataSet
Edit
Delete
Actions
Show plot
Visualize data
Status
Welcome to the Weka Knowledge Flow
Log


Weka KnowledgeFlow Environment
DataSources
DataSinks
Filters
Classifiers
Clusterers
Evaluation
Visualization
Step 4 : choose evaluation
Step 5 : choose filter method
Step 6 : choose classifier
Knowledge Flow Layout
@data
ARFF
iris
dataSet
CV
CrossValidation FoldMaker
trainingSet
testSet
Classifier
SMO
Edit
Delete
Configure...
Connections
batchClassifier
graph
incrementalClassifier
text
Set parameter
About
Implements John Platt's sequential minimal optimization algorithm for training a support vector classifier.
More
buildLogisticModels False
c 1.0
cacheSize 250007
debug False
epsilon 1.0E-12
exponent 1.0
featureSpaceNormalization False
filterType Normalize training data
gamma 0.01
lowerOrderTerms False
numFolds -1
randomSeed 1
toleranceParameter 0.0010
useRBF True
Data Visualizer
Status
Welcome to the Weka Knowledge Flow
Log


Weka KnowledgeFlow Environment
DataSources
DataSinks
Filters
Classifiers
Clusterers
Evaluation
Visualization
Visualization
Data Visualizer
Scatter PlotMatrix
Attribute Summarizer
Model PerformanceChart
Text Viewer
Graph Viewer
Strip Chart
Knowledge Flow Layout
@data
ARFF
dataSet
CV
CrossValidation FoldMaker
trainingSet
testSet
Attribute Selection
trainingSet
testSet
Classifier
SMO
batchClassifier
Edit
Delete
Configure...
Connections
dataSet
instance
Actions
Start loading
Final step : start to run
Classifier
+ 90%
- 10%
Classifier PerformanceEvaluator
text
TextViewer
Data Visualizer
Status
Done.
Log


Weka KnowledgeFlow Environment
Text Viewer
Result list
02:18:26 - SMO
02:18:45 - SMO
Text
=== Evaluation result ===
Scheme: SMO
Relation: iris-weka.filters.supervised.attribute.AttributeSelection-Eweka.attributeSelectio
Correctly Classified Instances 72 48 %
Incorrectly Classified Instances 78 52 %
Kappa statistic 0.22
Mean absolute error 0.3807
Root mean squared error 0.4704
Relative absolute error 85.6667 %
Root relative squared error 99.3188 %
Total Number of Instances 150
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure Class
0.72 0.3 0.545 0.72 0.621 Iris-setosa
0.28 0.28 0.333 0.28 0.304 Iris-versicolor
0.44 0.2 0.524 0.44 0.478 Iris-virginica
=== Confusion Matrix ===
a b c <-- classified as
36 14 0 | a = Iris-setosa
16 14 20 | b = Iris-versicolor
14 14 22 | c = Iris-virginica
Graph Viewer
Strip Chart
Classifier
SMO
Data Visualizer
Classifier PerformanceEvaluator
Edit
Delete
Actions
Show results
Clear results
Finish
Status
Done.
Log



Thank you