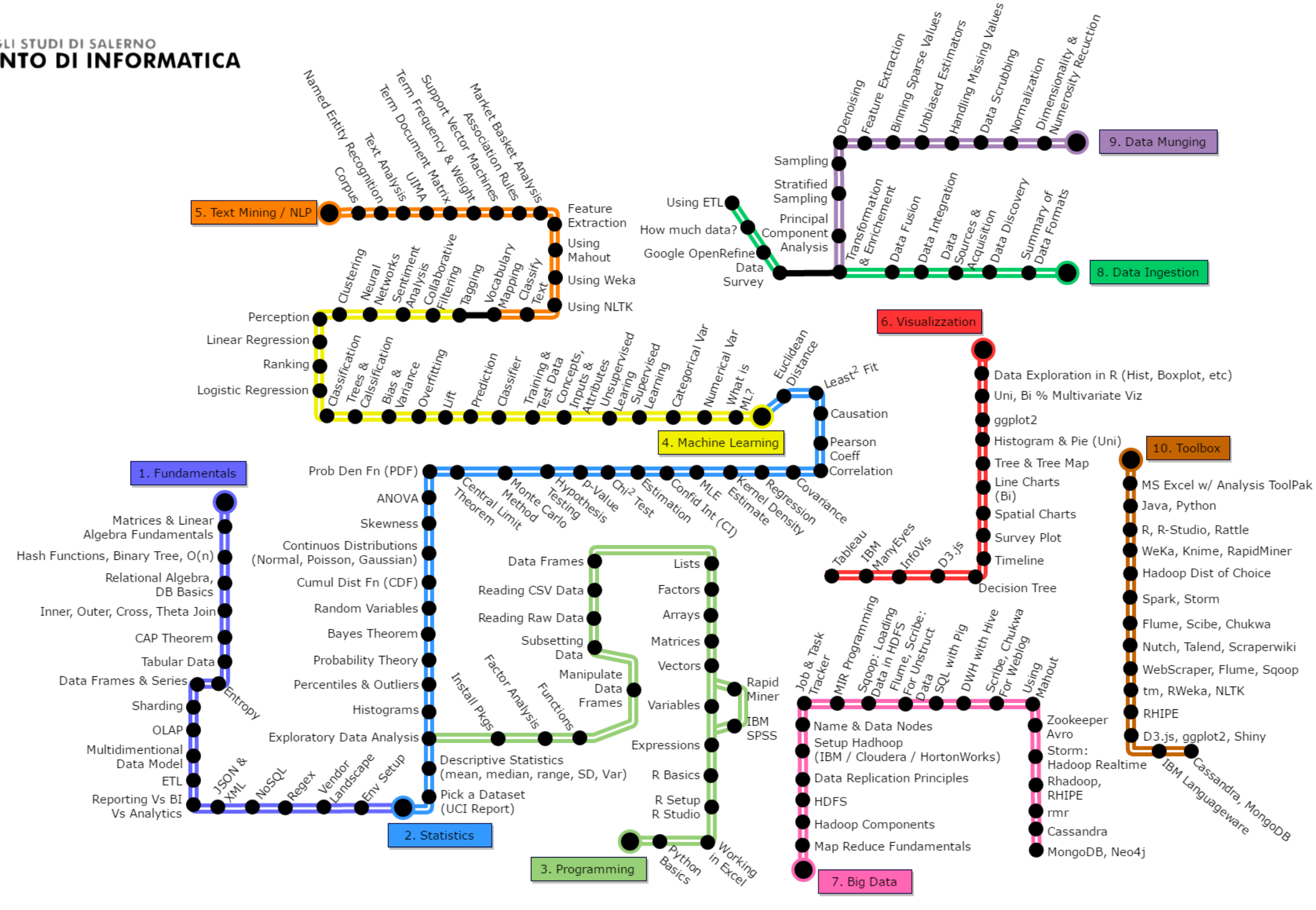
UNIVERSITÀ DEGLI STUDI DI SALERNO
DIPARTIMENTO DI INFORMATICA



1. Fundamentals
Matrices & Linear Algebra Fundamentals
Hash Functions, Binary Tree, O(n)
Relational Algebra, DB Basics
Inner, Outer, Cross, Theta Join
CAP Theorem
Tabular Data
Entropy
Data Frames & Series
Sharding
OLAP
Multidimentional Data Model
ETL
Reporting Vs BI Vs Analytics
JSON & XML
NoSQL
Regex
Vendor Landscape
Env Setup

2. Statistics
Pick a Dataset (UCI Report)
Descriptive Statistics (mean, median, range, SD, Var)
Exploratory Data Analysis
Histograms
Percentiles & Outliers
Probability Theory
Bayes Theorem
Random Variables
Cumul Dist Fn (CDF)
Continuos Distributions (Normal, Poisson, Gaussian)
Skewness
ANOVA
Prob Den Fn (PDF)
Central Limit Theorem
Monte Carlo Method
Hypothesis Testing
p-Value
Chi^2 Test
Estimation
Confid Int (CI)
MLE
Kernel Density Estimate
Regression
Covariance
Correlation
Pearson Coeff
Causation
Least^2 Fit
Euclidean Distance

3. Programming
Python Basics
Working in Excel
R Setup R Studio
R Basics
Expressions
Variables
IBM SPSS
Rapid Miner
Vectors
Matrices
Arrays
Factors
Lists
Data Frames
Reading CSV Data
Reading Raw Data
Subsetting Data
Manipulate Data Frames
Functions
Factor Analysis
Install Pkgs

4. Machine Learning
What is ML?
Numerical Var
Categorical Var
Supervised Learning
Unsupervised Learing
Concepts, Inputs & Attributes
Training & Test Data
Classifier
Prediction
Lift
Overfitting
Bias & Variance
Trees & Calssification
Classification
Logistic Regression
Ranking
Linear Regression
Perception
Clustering
Neural Networks
Sentiment Analysis
Collaborative Filtering
Tagging
Vocabulary Mapping
Classify Text

5. Text Mining / NLP
Corpus
Named Entity Recognition
Text Analysis
UIMA
Term Document Matrix
Term Frequency & Weight
Support Vector Machines
Association Rules
Market Basket Analysis
Feature Extraction
Using Mahout
Using Weka
Using NLTK

6. Visualizzation
Data Exploration in R (Hist, Boxplot, etc)
Uni, Bi % Multivariate Viz
ggplot2
Histogram & Pie (Uni)
Tree & Tree Map
Line Charts (Bi)
Spatial Charts
Survey Plot
Timeline
Decision Tree
D3.js
InfoVis
IBM ManyEyes
Tableau

7. Big Data
Map Reduce Fundamentals
Hadoop Components
HDFS
Data Replication Principles
Setup Hadhoop (IBM / Cloudera / HortonWorks)
Name & Data Nodes
Job & Task Tracker
M/R Programming
Sqoop: Loading Data in HDFS
Flume, Scribe: For Unstruct Data
SQL with Pig
DWH with Hive
Scribe, Chukwa For Weblog
Using Mahout
Zookeeper Avro
Storm: Hadoop Realtime
Rhadoop, RHIPE
rmr
Cassandra
MongoDB, Neo4j

8. Data Ingestion
Summary of Data Formats
Data Discovery
Data Sources & Acquisition
Data Integration
Data Fusion
Transformation & Enrichement
Data Survey
Google OpenRefine
How much data?
Using ETL

9. Data Munging
Dimensionality & Numerosity Recuction
Normalization
Data Scrubbing
Handling Missing Values
Unbiased Estimators
Binning Sparse Values
Feature Extraction
Denoising
Sampling
Stratified Sampling
Principal Component Analysis

10. Toolbox
MS Excel w/ Analysis ToolPak
Java, Python
R, R-Studio, Rattle
WeKa, Knime, RapidMiner
Hadoop Dist of Choice
Spark, Storm
Flume, Scibe, Chukwa
Nutch, Talend, Scraperwiki
WebScraper, Flume, Sqoop
tm, RWeka, NLTK
RHIPE
D3.js, ggplot2, Shiny
IBM Languageware
Cassandra, MongoDB

Fondamenti di Data Science e Machine Learning

WEKA

Outline

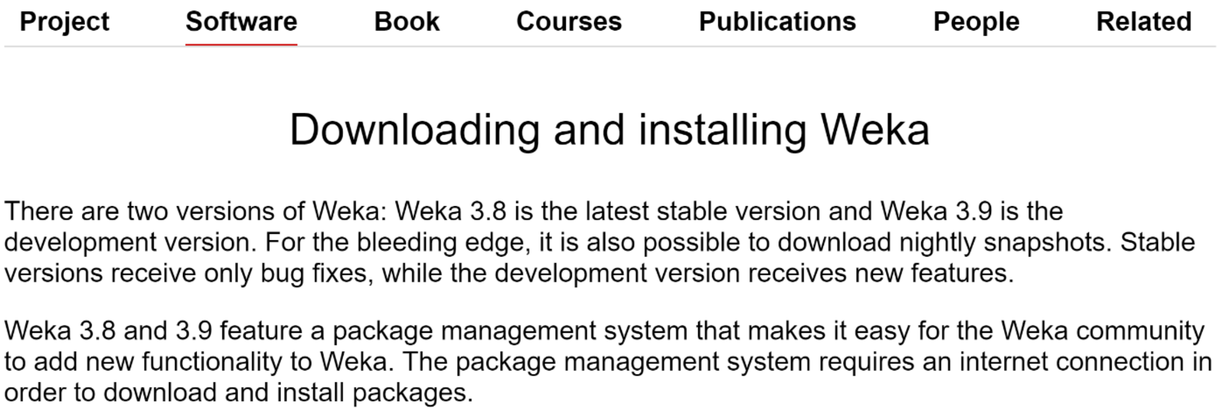
- Introduction to WEKA
- The Explorer
  - Process data
  - Classification
  - Clustering
  - Association Rules
  - Attribute Selection
  - Data Visualization
- References and Resources

WEKA

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

How to install WEKA

- Download and Install WEKA
  - Website:
  - https://waikato.github.io/weka-wiki/downloading_weka/
  - Support multiple platforms (written in java):
    - Windows, Mac OS X and Linux



Downloading and installing Weka

There are two versions of Weka: Weka 3.8 is the latest stable version and Weka 3.9 is the development version. For the bleeding edge, it is also possible to download nightly snapshots. Stable versions receive only bug fixes, while the development version receives new features.

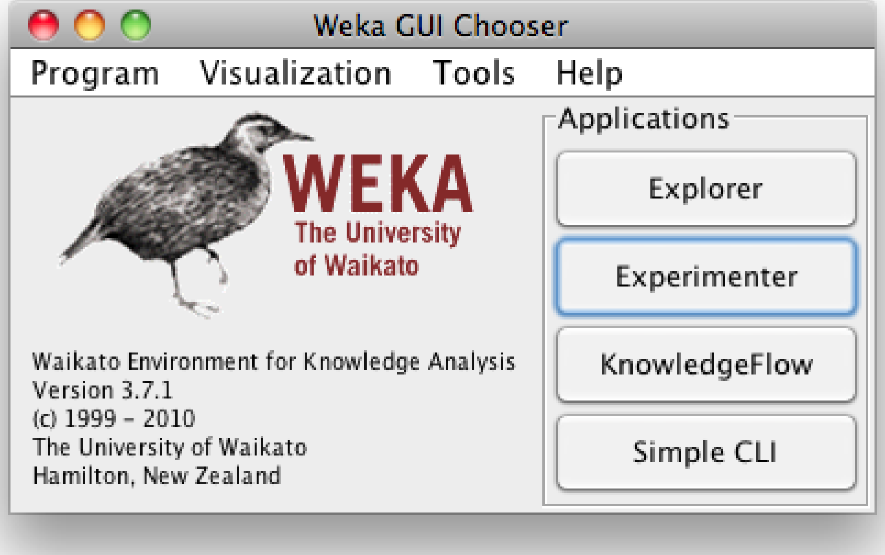
Weka 3.8 and 3.9 feature a package management system that makes it easy for the Weka community to add new functionality to Weka. The package management system requires an internet connection in order to download and install packages.

WEKA: Features

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

The Graphical User Interface of WEKA

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



Waikato Environment for Knowledge Analysis
Version 3.7.1
(c) 1999 - 2010
The University of Waikato
Hamilton, New Zealand

The Explorer

- 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, …

Data Preprocessing (1)

- WEKA only deals with “flat” files

```
@relation cancerTraining

@attribute  'ClumpThickness'           NUMERIC
@attribute  'UniformityofCellSize'     NUMERIC
@attribute  'UniformityofCellShape'    NUMERIC
@attribute  'MarginalAdhesion'         NUMERIC
@attribute  'SingleEpithelialCellSize' NUMERIC
@attribute  'BareNuclei' {1,2,3,4,5,6,7,8,9,10,'?'}
@attribute  'BlandChromatin'           NUMERIC
@attribute  'NormalNucleoli'           NUMERIC
@attribute  'Mitoses'     NUMERIC
@attribute  'Class'       {2,4}

@data
5,1,1,1,2,1,3,1,1,2
5,4,4,5,7,10,3,2,1,4
....
```

Flat file in ARFF format

Data Preprocessing (1)

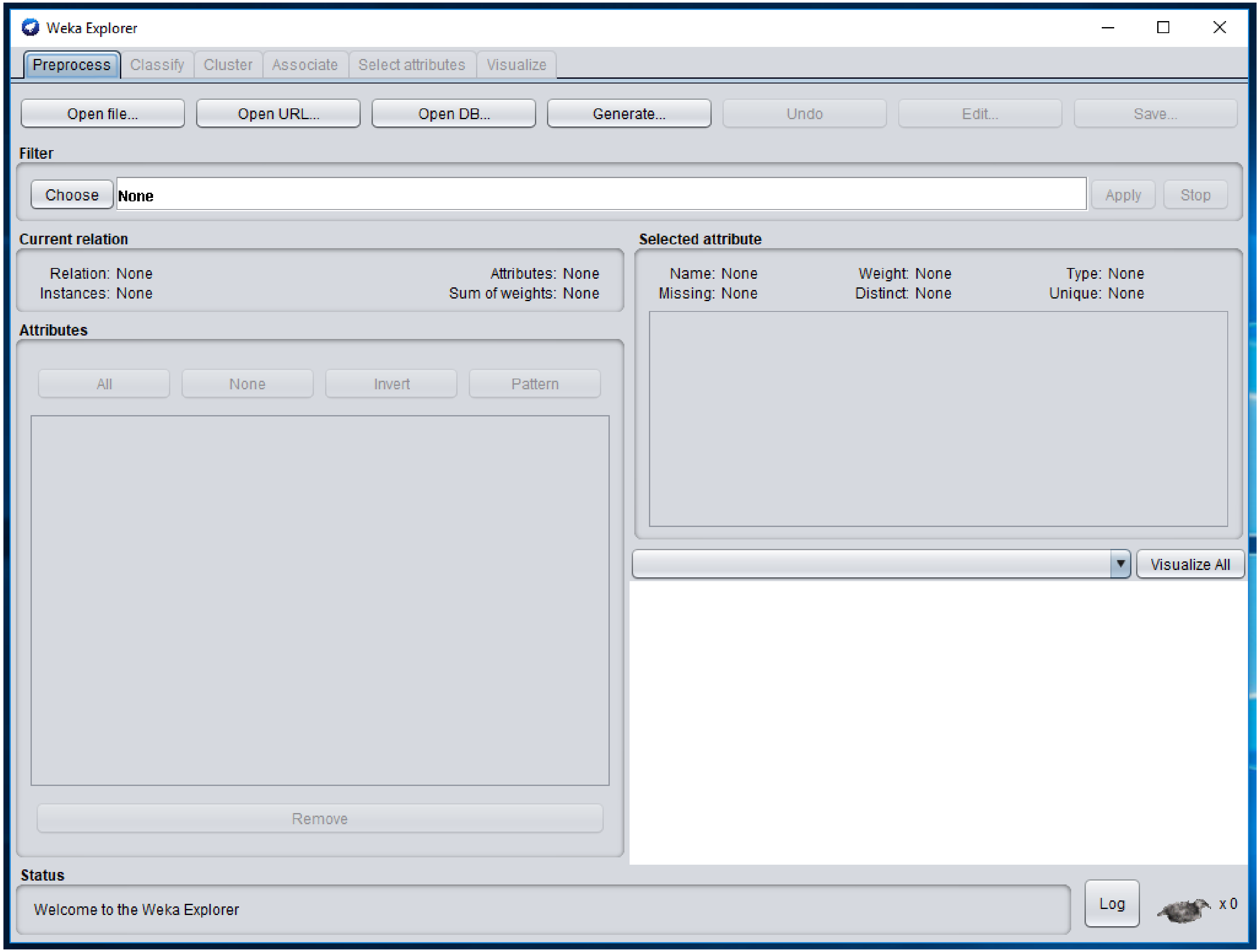
- WEKA only deals with “flat” files

```
@relation cancerTraining

@attribute  'ClumpThickness'           NUMERIC
@attribute  'UniformityofCellSize'     NUMERIC
@attribute  'UniformityofCellShape'    NUMERIC
@attribute  'MarginalAdhesion'         NUMERIC
@attribute  'SingleEpithelialCellSize' NUMERIC
@attribute  'BareNuclei' {1,2,3,4,5,6,7,8,9,10,'?'}
@attribute  'BlandChromatin'           NUMERIC
@attribute  'NormalNucleoli'           NUMERIC
@attribute  'Mitoses'     NUMERIC
@attribute  'Class'       {2,4}

@data
5,1,1,1,2,1,3,1,1,2
5,4,4,5,7,10,3,2,1,4
….
```

Numeric attribute

Nominal attribute

Data Preprocessing (2)



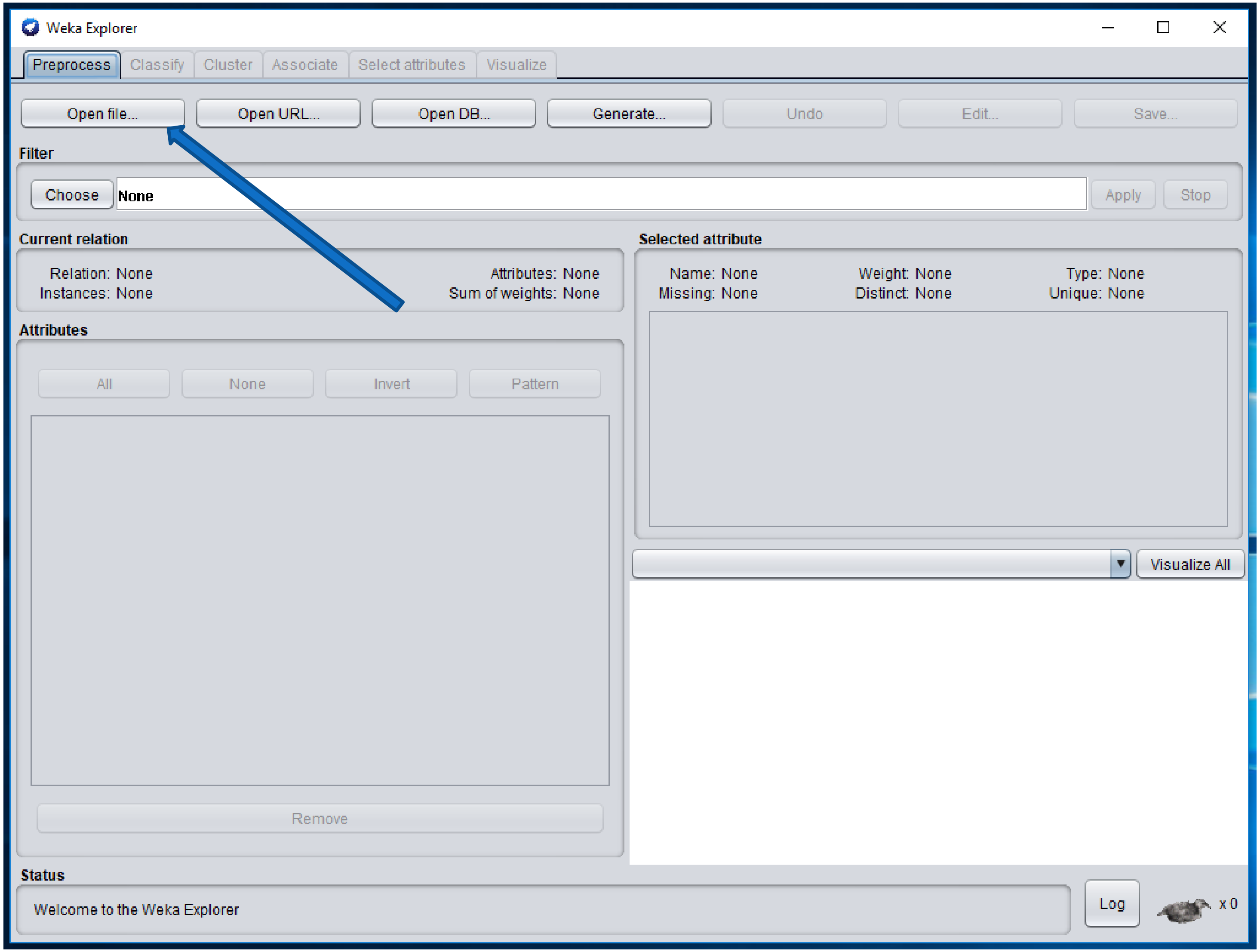
Filter: None

Relation: None
Instances: None
Attributes: None
Sum of weights: None

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

Welcome to the Weka Explorer

Data Preprocessing (2)



Filter

None

Current relation

Relation: None
Instances: None
Attributes: None
Sum of weights: None

Attributes

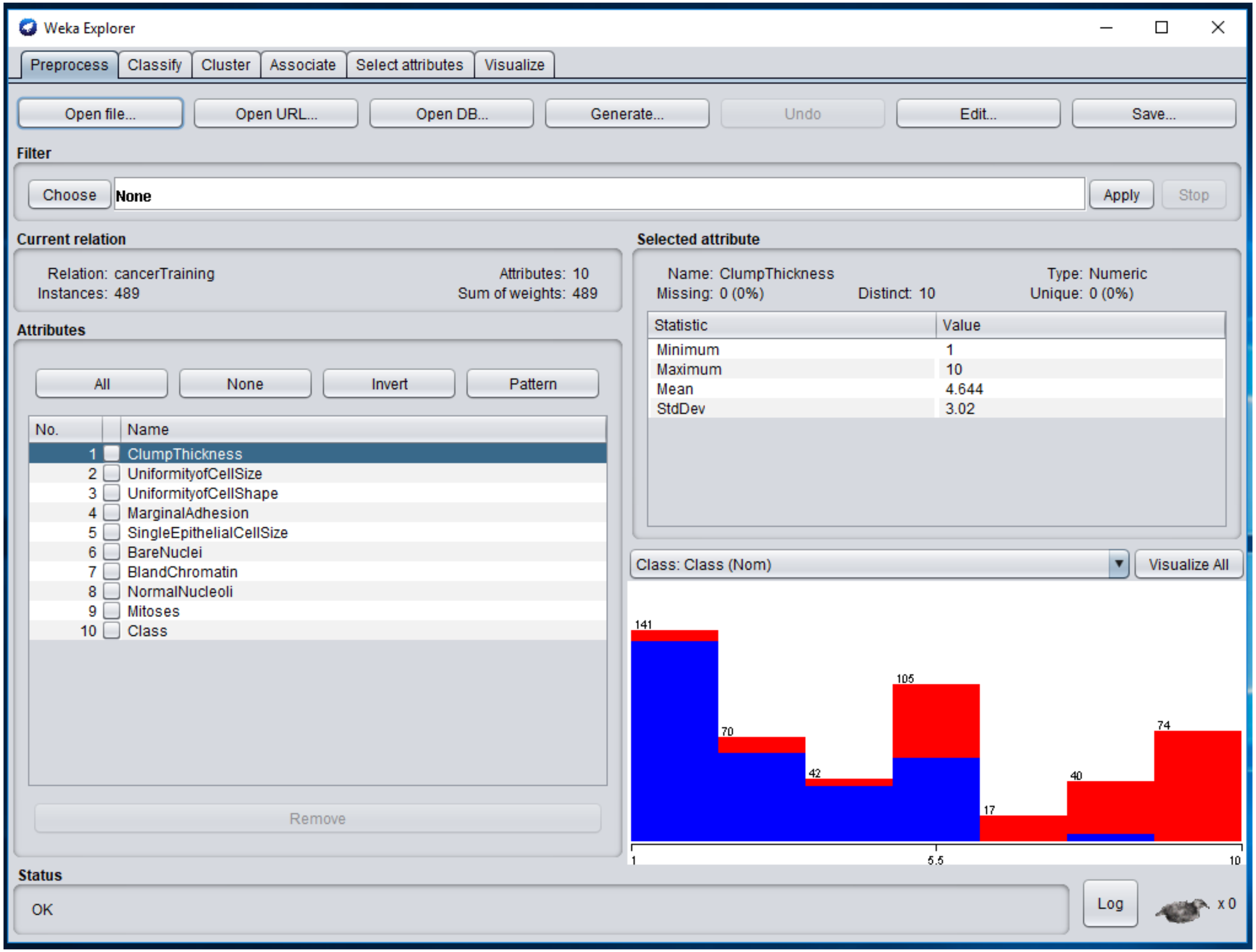
Selected attribute

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

Status

Welcome to the Weka Explorer

Data Preprocessing (3)



Filter

None

Current relation

Relation: cancerTraining
Instances: 489
Attributes: 10
Sum of weights: 489

Attributes

No.		Name
1	□	ClumpThickness
2	□	UniformityofCellSize
3	□	UniformityofCellShape
4	□	MarginalAdhesion
5	□	SingleEpithelialCellSize
6	□	BareNuclei
7	□	BlandChromatin
8	□	NormalNucleoli
9	□	Mitoses
10	□	Class

Selected attribute

Name: ClumpThickness
Type: Numeric
Missing: 0 (0%)
Distinct: 10
Unique: 0 (0%)

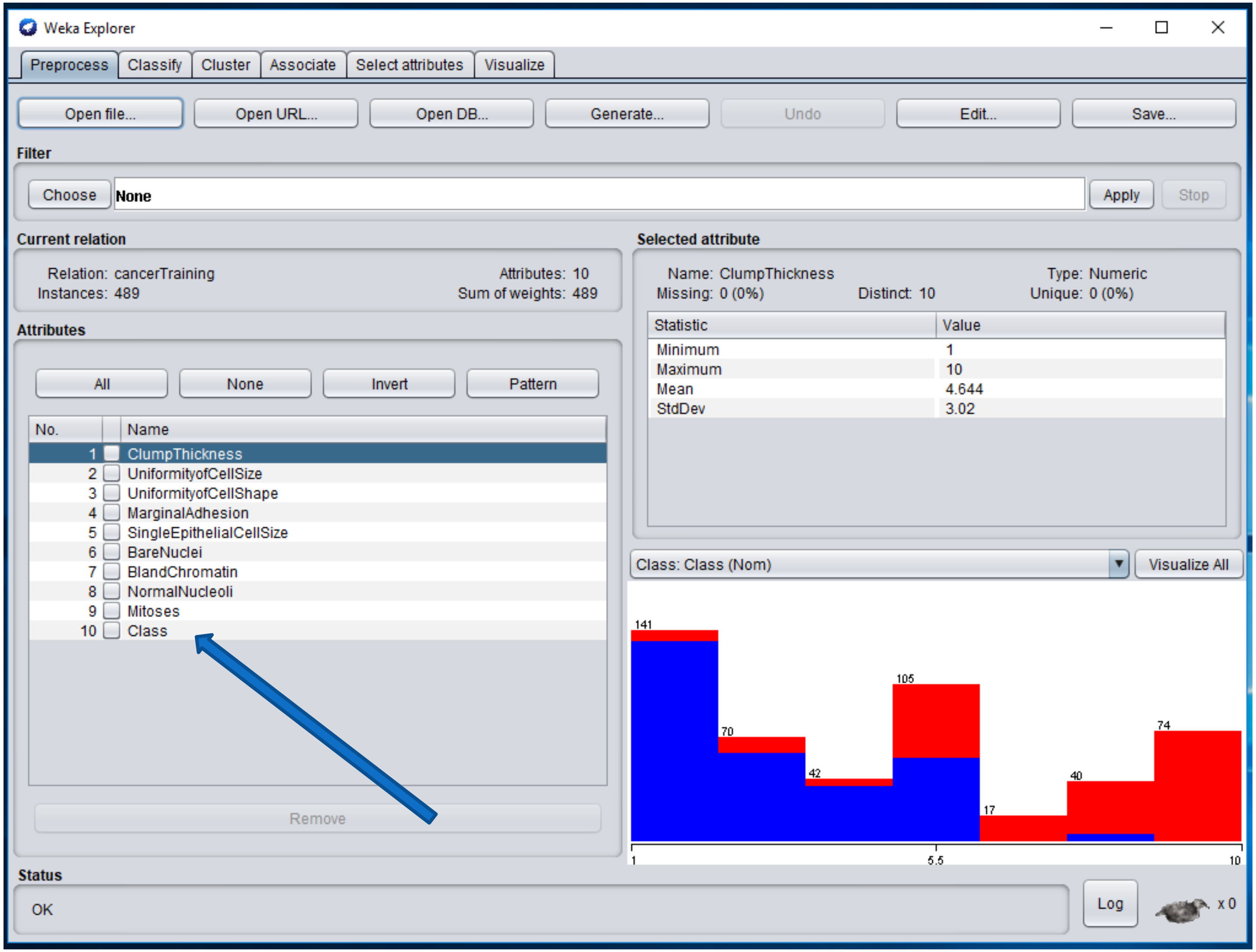
Statistic	Value
Minimum	1
Maximum	10
Mean	4.644
StdDev	3.02

Class: Class (Nom)

Status

OK

Data Preprocessing (3)



Filter

None

Current relation

Relation: cancerTraining
Instances: 489
Attributes: 10
Sum of weights: 489

Selected attribute

Name: ClumpThickness
Missing: 0 (0%)
Distinct: 10
Type: Numeric
Unique: 0 (0%)

Statistic	Value
Minimum	1
Maximum	10
Mean	4.644
StdDev	3.02

Attributes

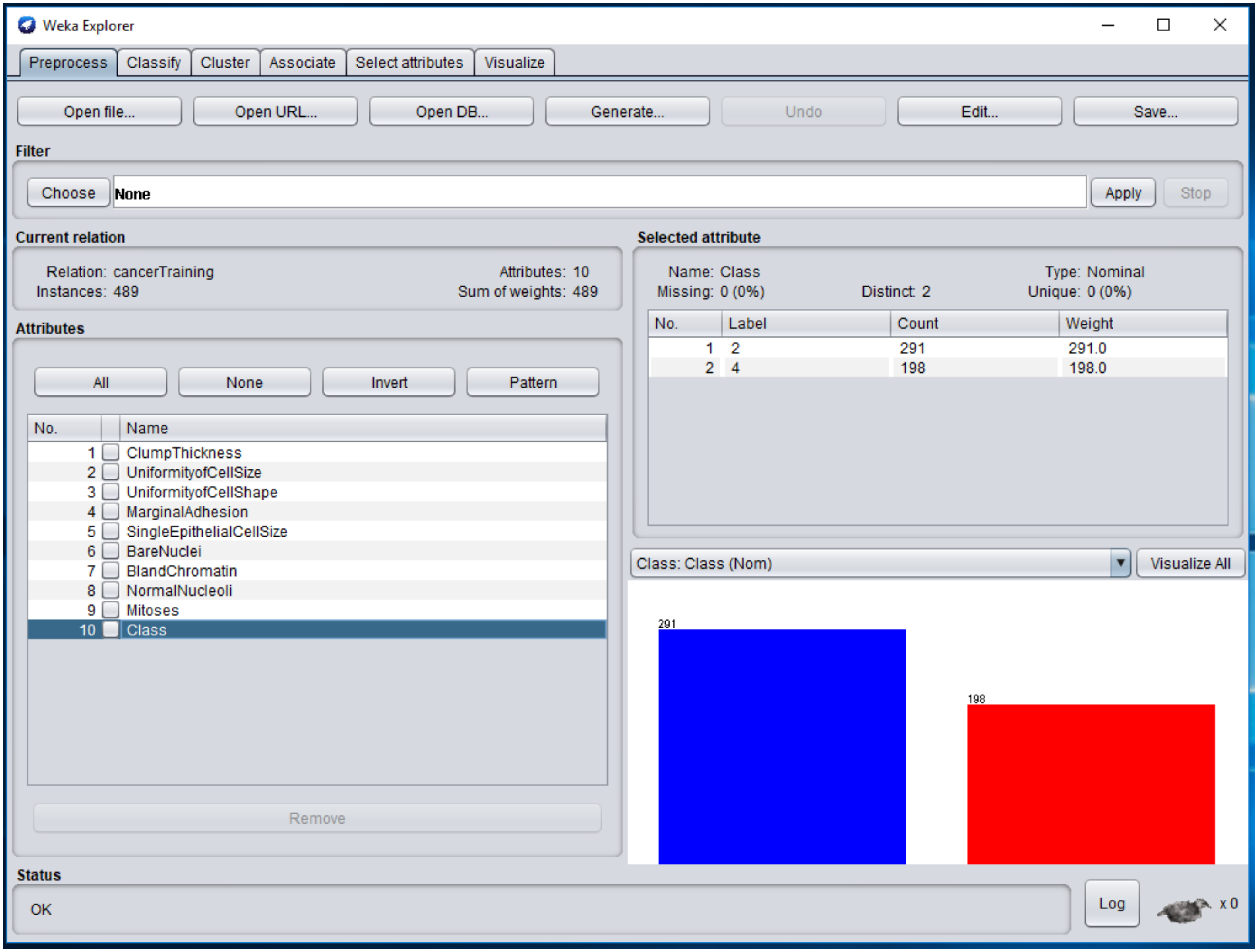
No.		Name
1	□	ClumpThickness
2	□	UniformityofCellSize
3	□	UniformityofCellShape
4	□	MarginalAdhesion
5	□	SingleEpithelialCellSize
6	□	BareNuclei
7	□	BlandChromatin
8	□	NormalNucleoli
9	□	Mitoses
10	□	Class

Class: Class (Nom)

Status

OK

Data Preprocessing (4)



Filter

None

Current relation

Relation: cancerTraining Attributes: 10
Instances: 489 Sum of weights: 489

Attributes

No.		Name
1	□	ClumpThickness
2	□	UniformityofCellSize
3	□	UniformityofCellShape
4	□	MarginalAdhesion
5	□	SingleEpithelialCellSize
6	□	BareNuclei
7	□	BlandChromatin
8	□	NormalNucleoli
9	□	Mitoses
10	□	Class

Selected attribute

Name: Class Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

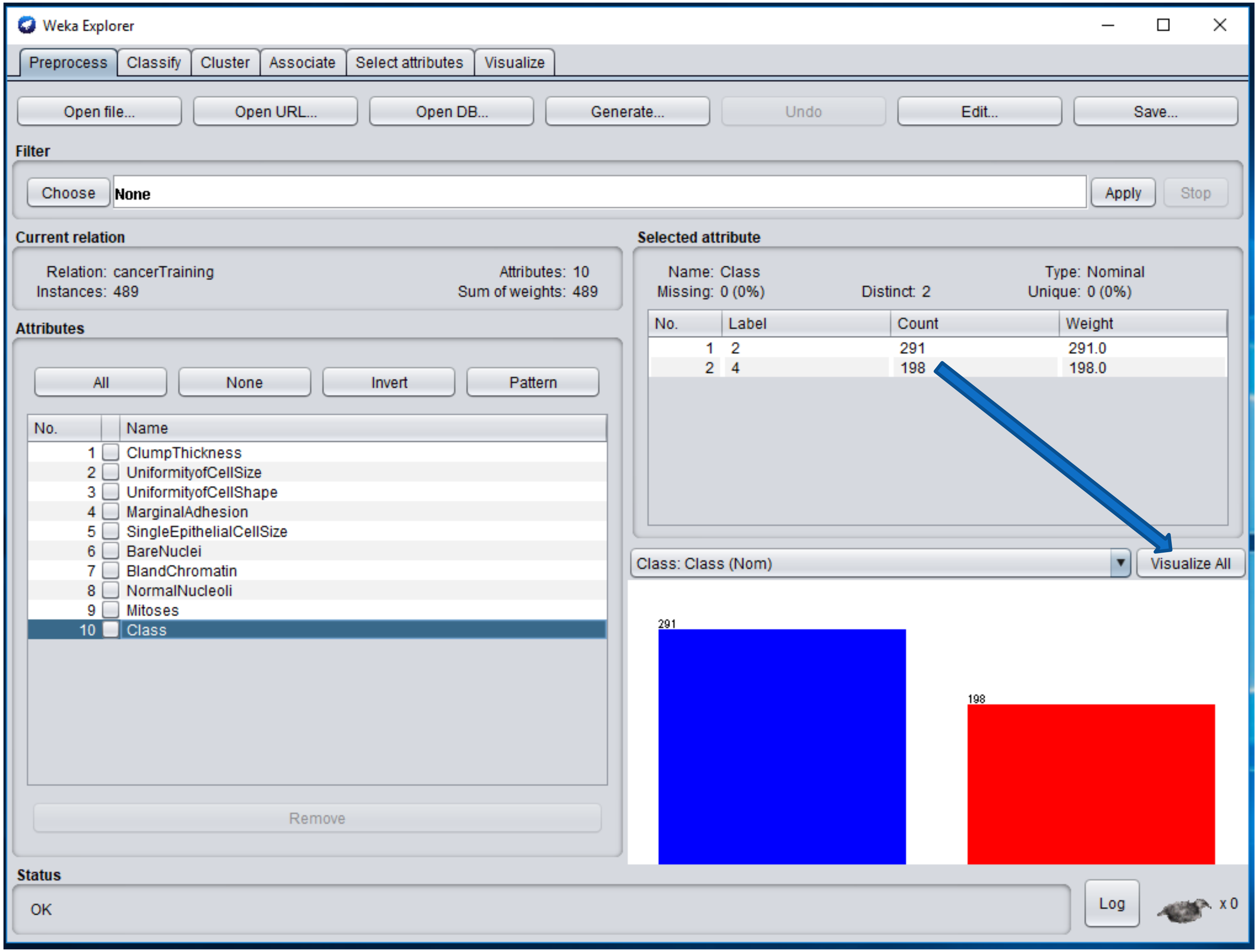
No.	Label	Count	Weight
1	2	291	291.0
2	4	198	198.0

Class: Class (Nom)

Status

OK

Data Preprocessing (4)



Filter: None

Current relation

Relation: cancerTraining
Instances: 489
Attributes: 10
Sum of weights: 489

Attributes

No.		Name
1	□	ClumpThickness
2	□	UniformityofCellSize
3	□	UniformityofCellShape
4	□	MarginalAdhesion
5	□	SingleEpithelialCellSize
6	□	BareNuclei
7	□	BlandChromatin
8	□	NormalNucleoli
9	□	Mitoses
10	□	Class

Selected attribute

Name: Class
Type: Nominal
Missing: 0 (0%)
Distinct: 2
Unique: 0 (0%)

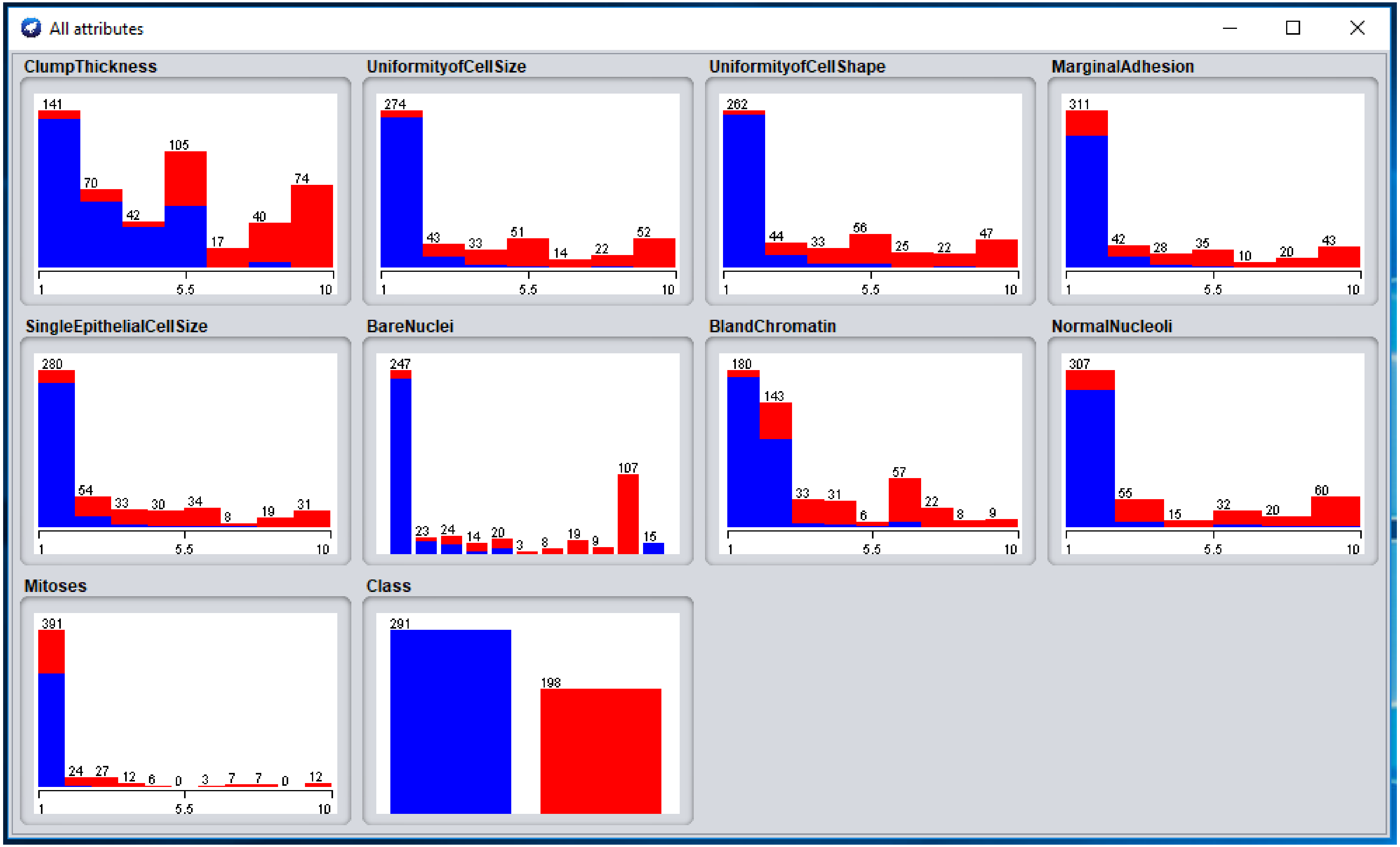
No.	Label	Count	Weight
1	2	291	291.0
2	4	198	198.0

Class: Class (Nom)

Status

OK

Data Preprocessing (5)



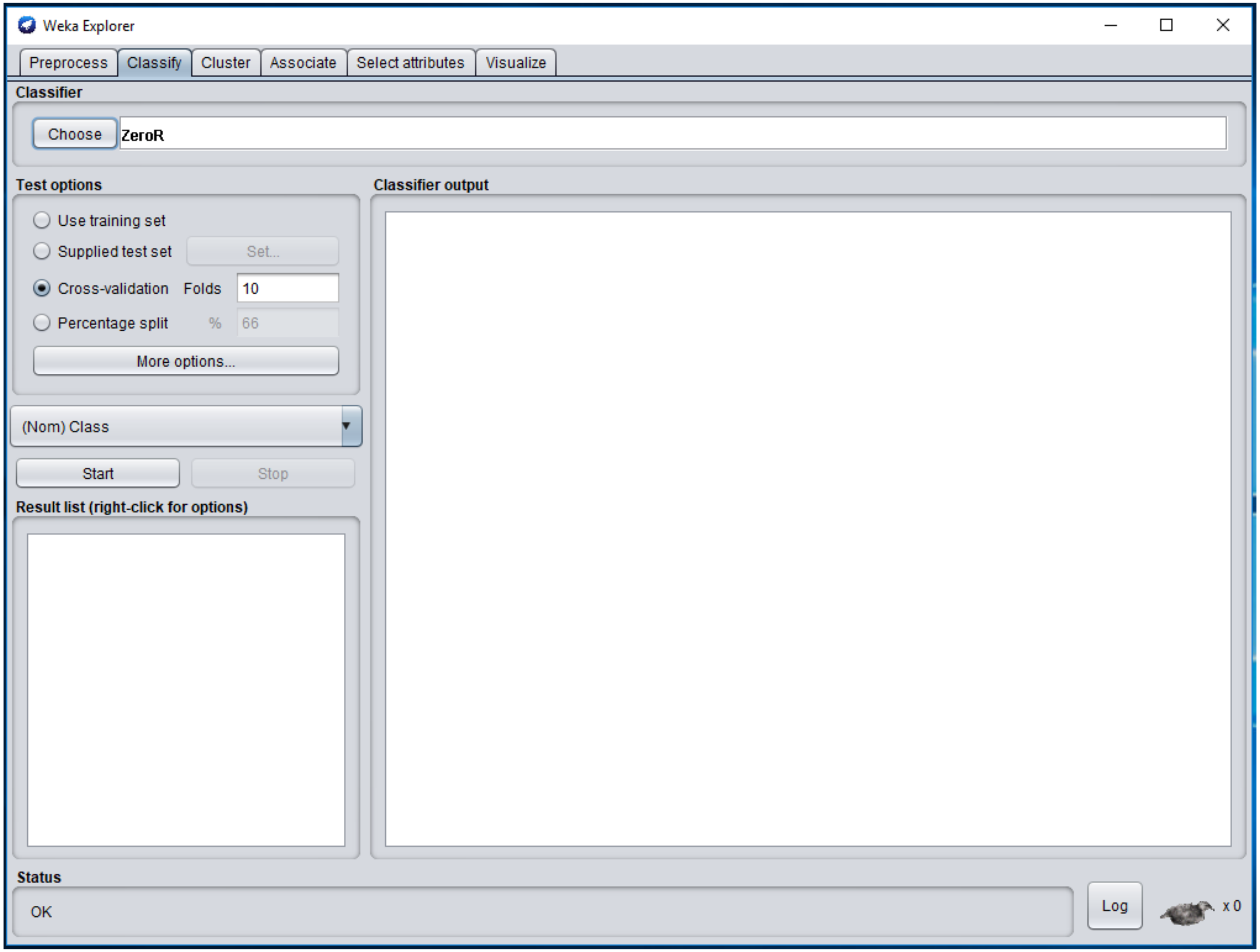
Classification

- Explorer: building “classifiers”
  - Classifiers in WEKA are models for predicting nominal or numeric quantities
  - Implemented learning schemes include:
    - Decision trees and lists
    - instance-based classifiers
    - support vector machines
    - multi-layer perceptrons
    - logistic regression
    - Bayes’ nets
    - etc

Decision Tree

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

Classification: An example (1)



Weka Explorer

Classifier

Choose ZeroR

Test options

○ Use training set

○ Supplied test set Set...

● Cross-validation Folds 10

○ Percentage split % 66

(Nom) Class

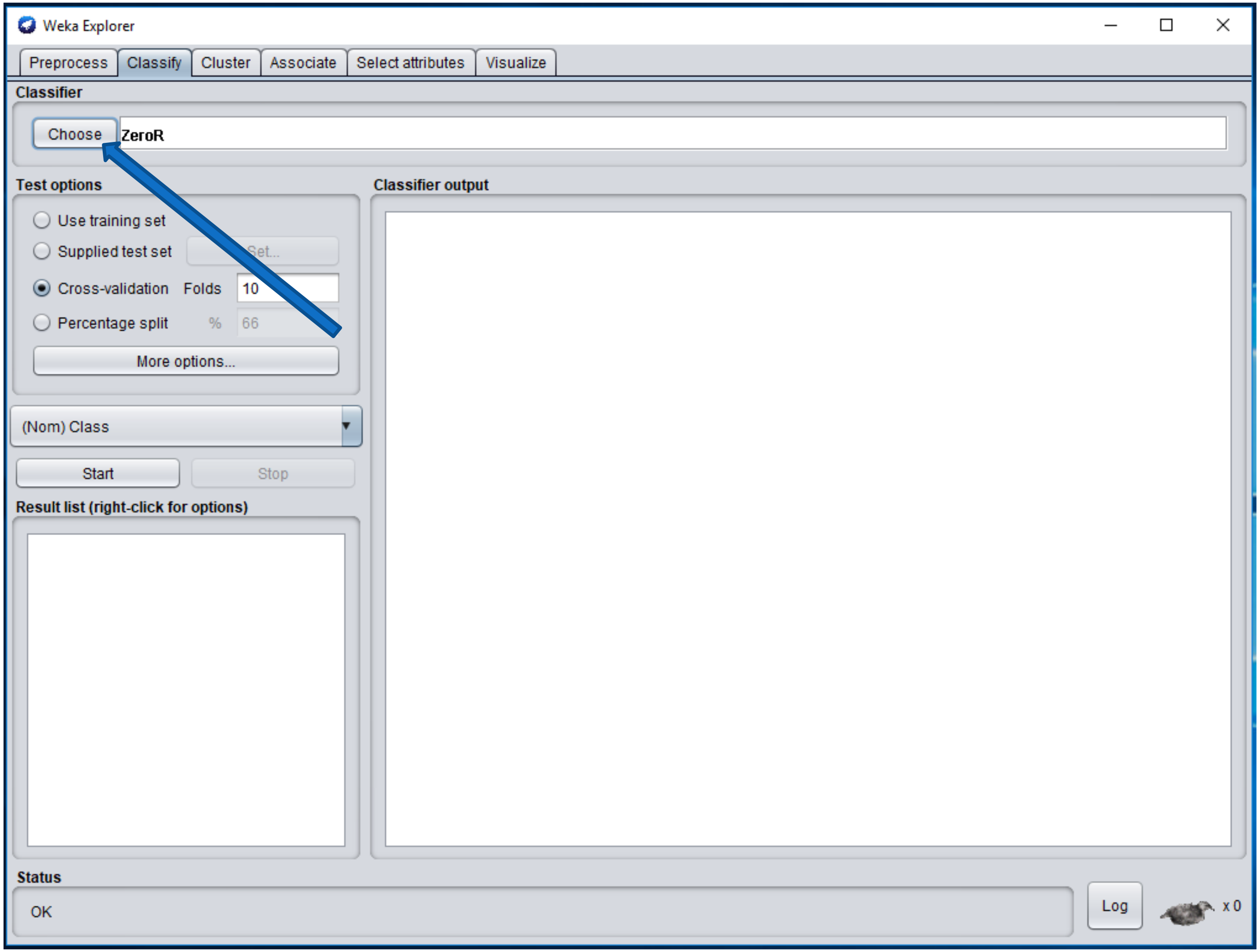
Result list (right-click for options)

Classifier output

Status

OK

Classification: An example (1)



Classifier

ZeroR

Test options

○ Use training set

○ Supplied test set

● Cross-validation Folds 10

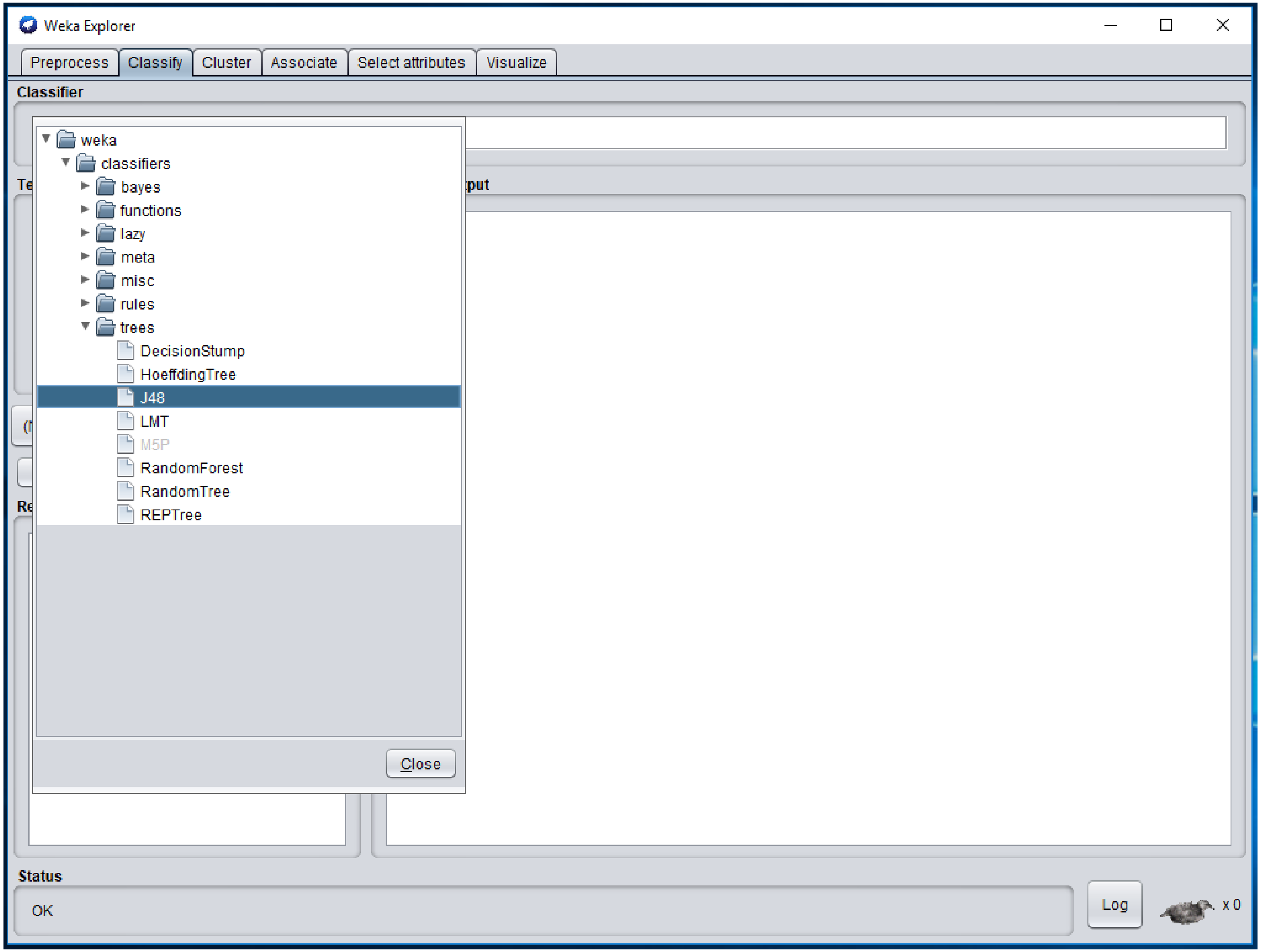
○ Percentage split % 66

(Nom) Class

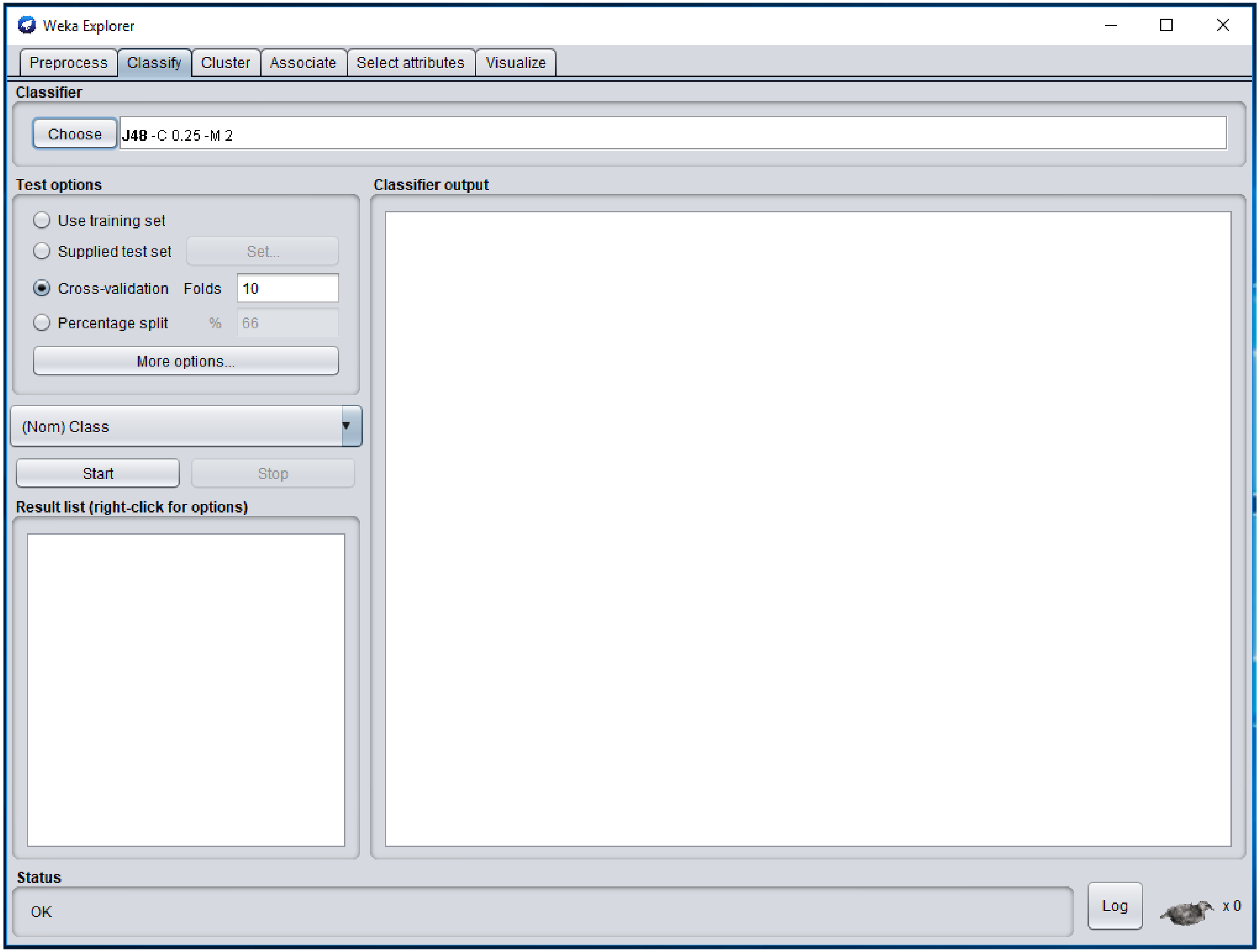
Status

OK

Classification: An example (1)



Classification: An example (1)



Classifier: J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set

● Cross-validation Folds 10

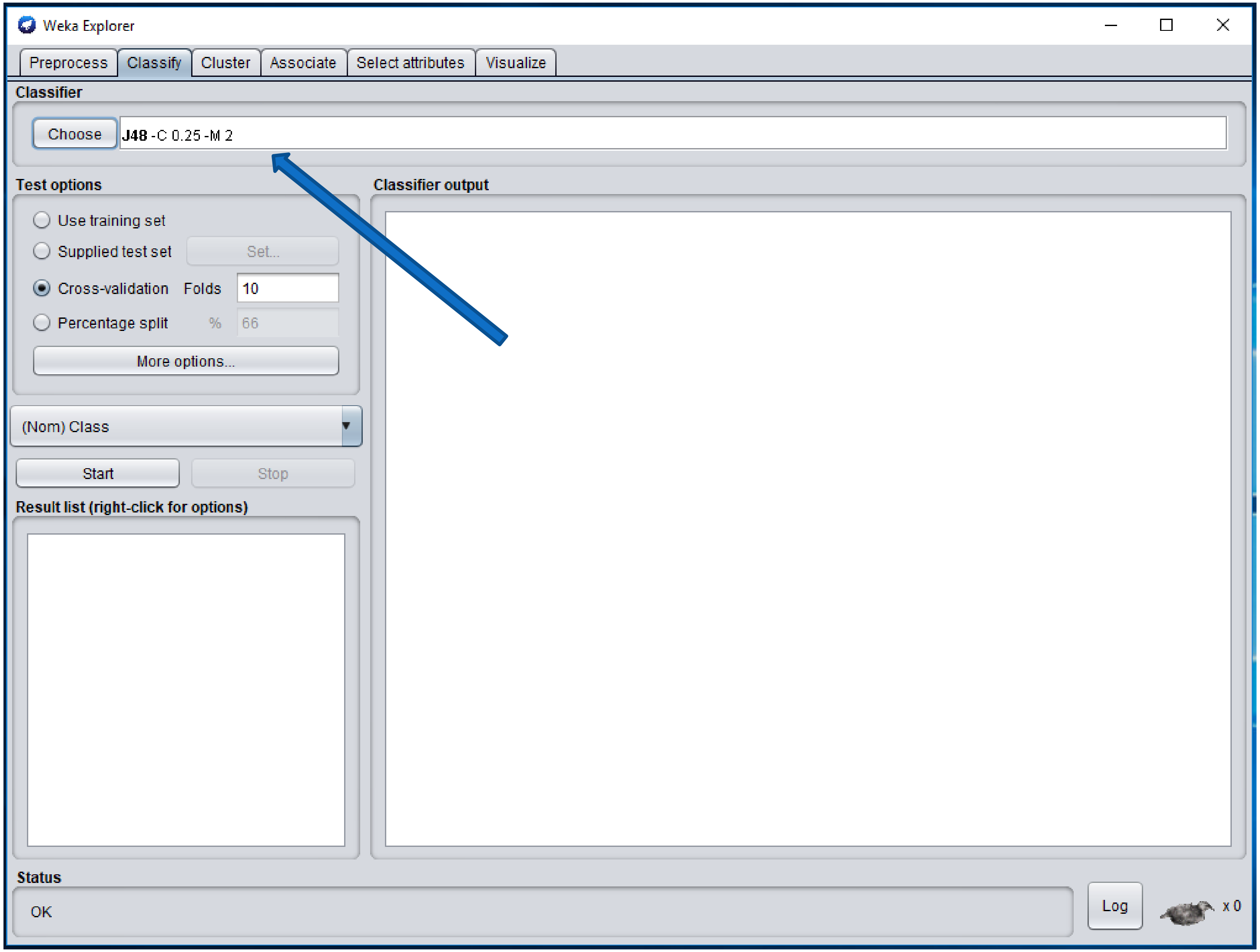
○ Percentage split % 66

(Nom) Class

Status

OK

Classification: An example (1)



Classifier

Choose J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set Set...

● Cross-validation Folds 10

○ Percentage split % 66

More options...

(Nom) Class

Start Stop

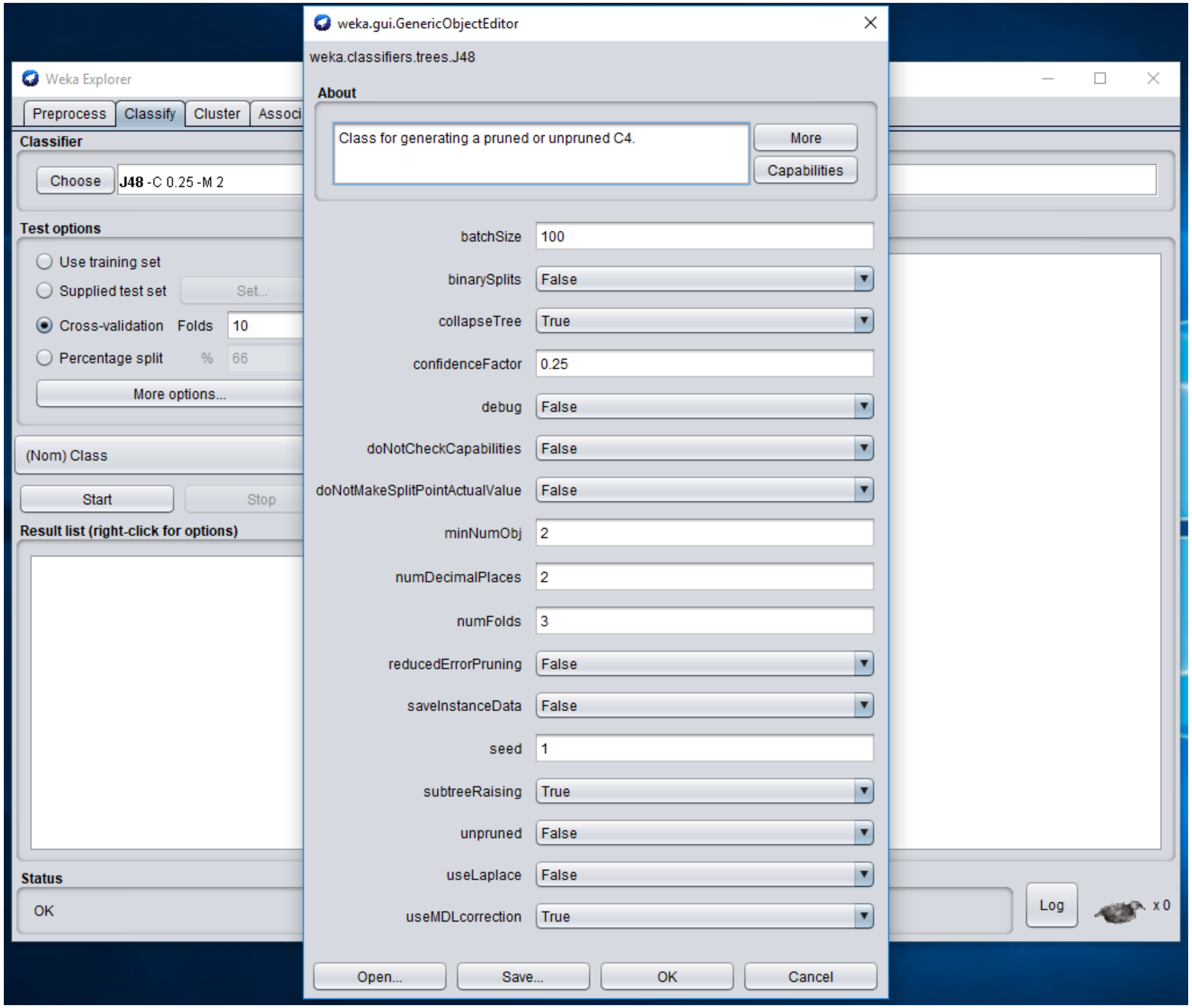
Result list (right-click for options)

Classifier output

Status

OK

Classification: An example (2)



weka.gui.GenericObjectEditor

weka.classifiers.trees.J48

About

Class for generating a pruned or unpruned C4.

batchSize	100
binarySplits	False
collapseTree	True
confidenceFactor	0.25
debug	False
doNotCheckCapabilities	False
doNotMakeSplitPointActualValue	False
minNumObj	2
numDecimalPlaces	2
numFolds	3
reducedErrorPruning	False
saveInstanceData	False
seed	1
subtreeRaising	True
unpruned	False
useLaplace	False
useMDLcorrection	True

Weka Explorer

Classifier: J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set

● Cross-validation Folds 10

○ Percentage split % 66

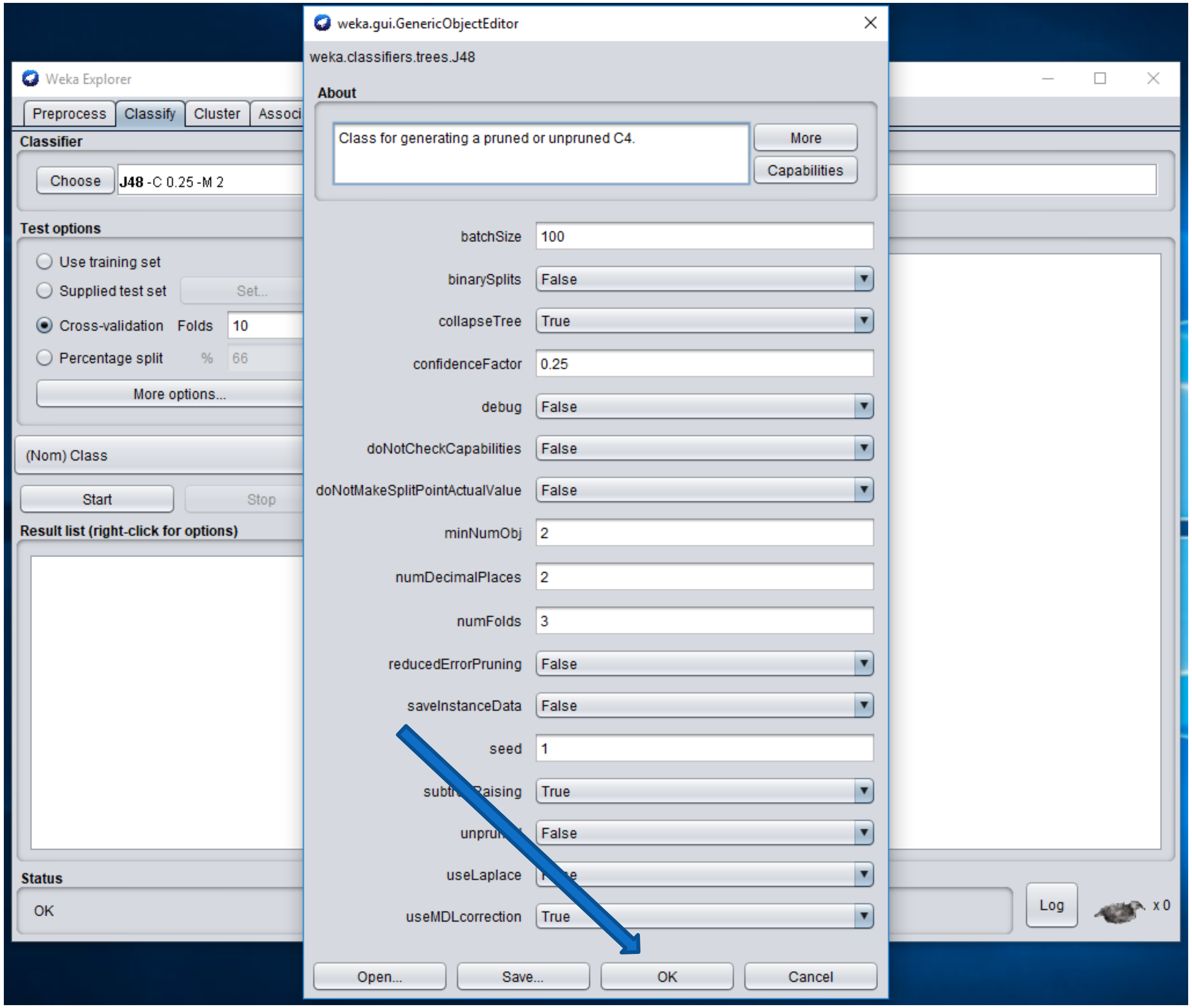
(Nom) Class

Result list (right-click for options)

Status

OK

Classification: An example (2)



weka.gui.GenericObjectEditor

weka.classifiers.trees.J48

About

Class for generating a pruned or unpruned C4.

batchSize	100
binarySplits	False
collapseTree	True
confidenceFactor	0.25
debug	False
doNotCheckCapabilities	False
doNotMakeSplitPointActualValue	False
minNumObj	2
numDecimalPlaces	2
numFolds	3
reducedErrorPruning	False
saveInstanceData	False
seed	1
subtreeRaising	True
unpruned	False
useLaplace	False
useMDLcorrection	True

Weka Explorer

Classifier: J48 -C 0.25 -M 2

Test options

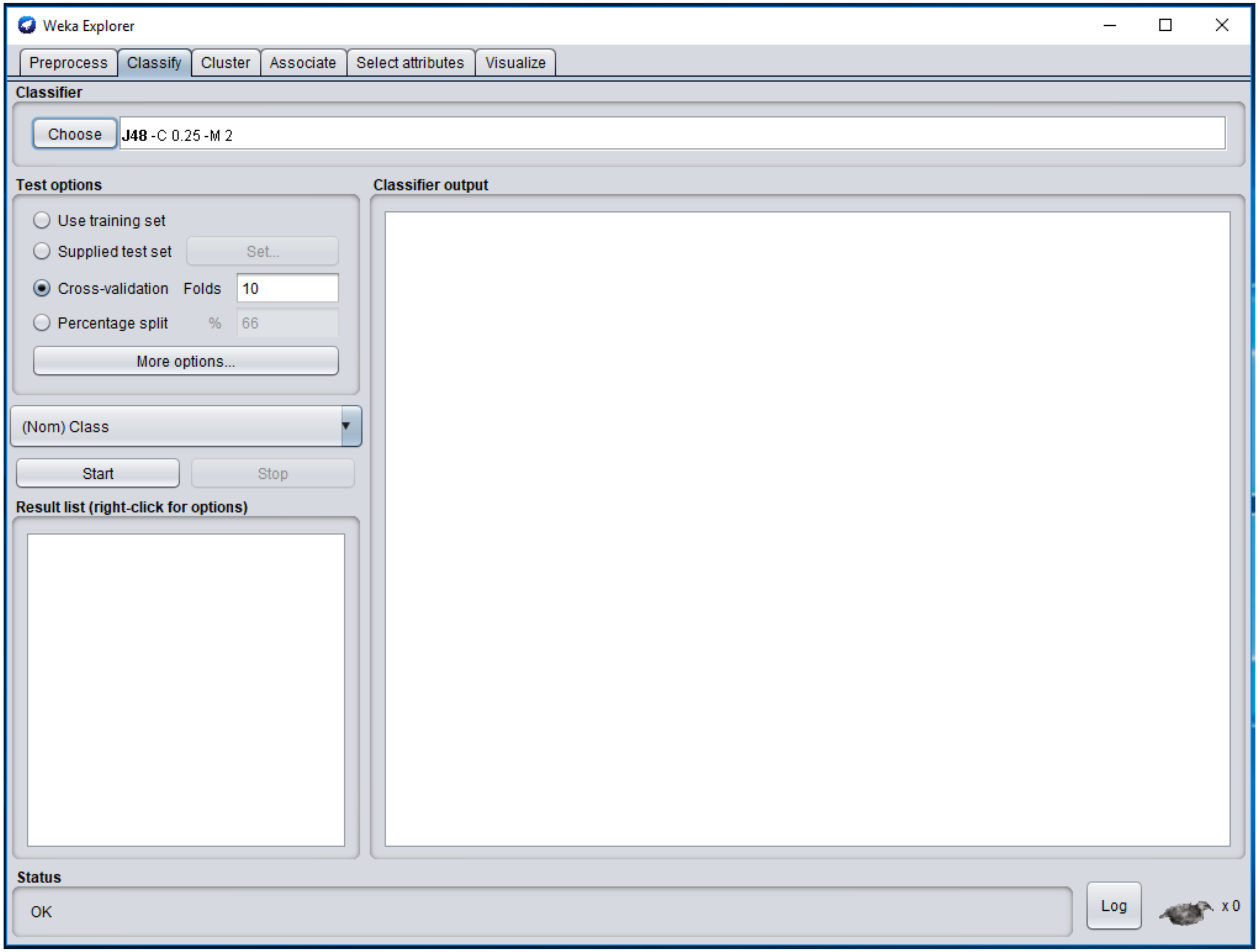
- ○ Use training set
- ○ Supplied test set
- ● Cross-validation Folds 10
- ○ Percentage split % 66

(Nom) Class

Result list (right-click for options)

Status: OK

Classification: An example (3)



J48 -C 0.25 -M 2

○ Use training set

○ Supplied test set   Set...

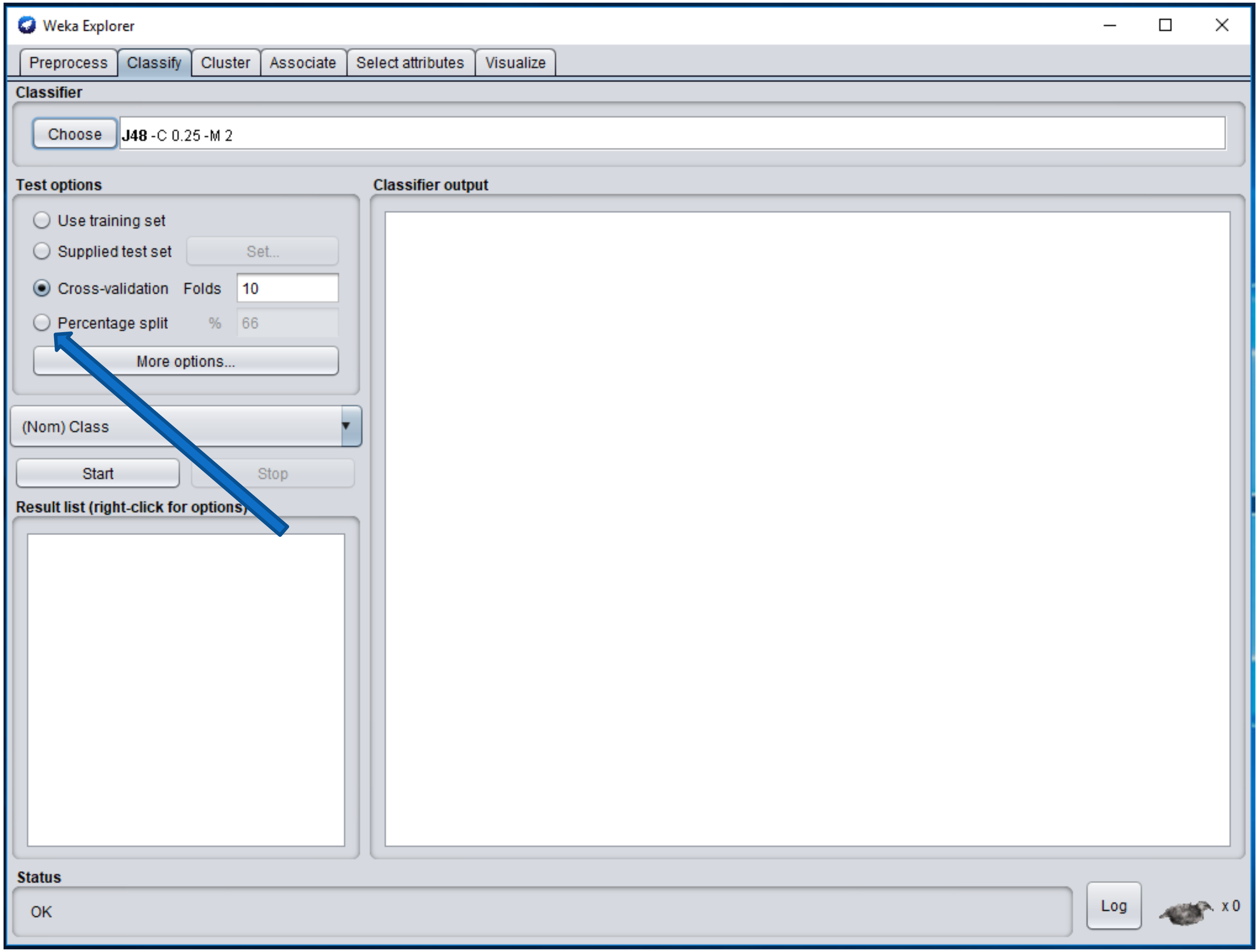
● Cross-validation   Folds 10

○ Percentage split   % 66

(Nom) Class

OK

Classification: An example (3)



Weka Explorer

J48 -C 0.25 -M 2

○ Use training set

○ Supplied test set

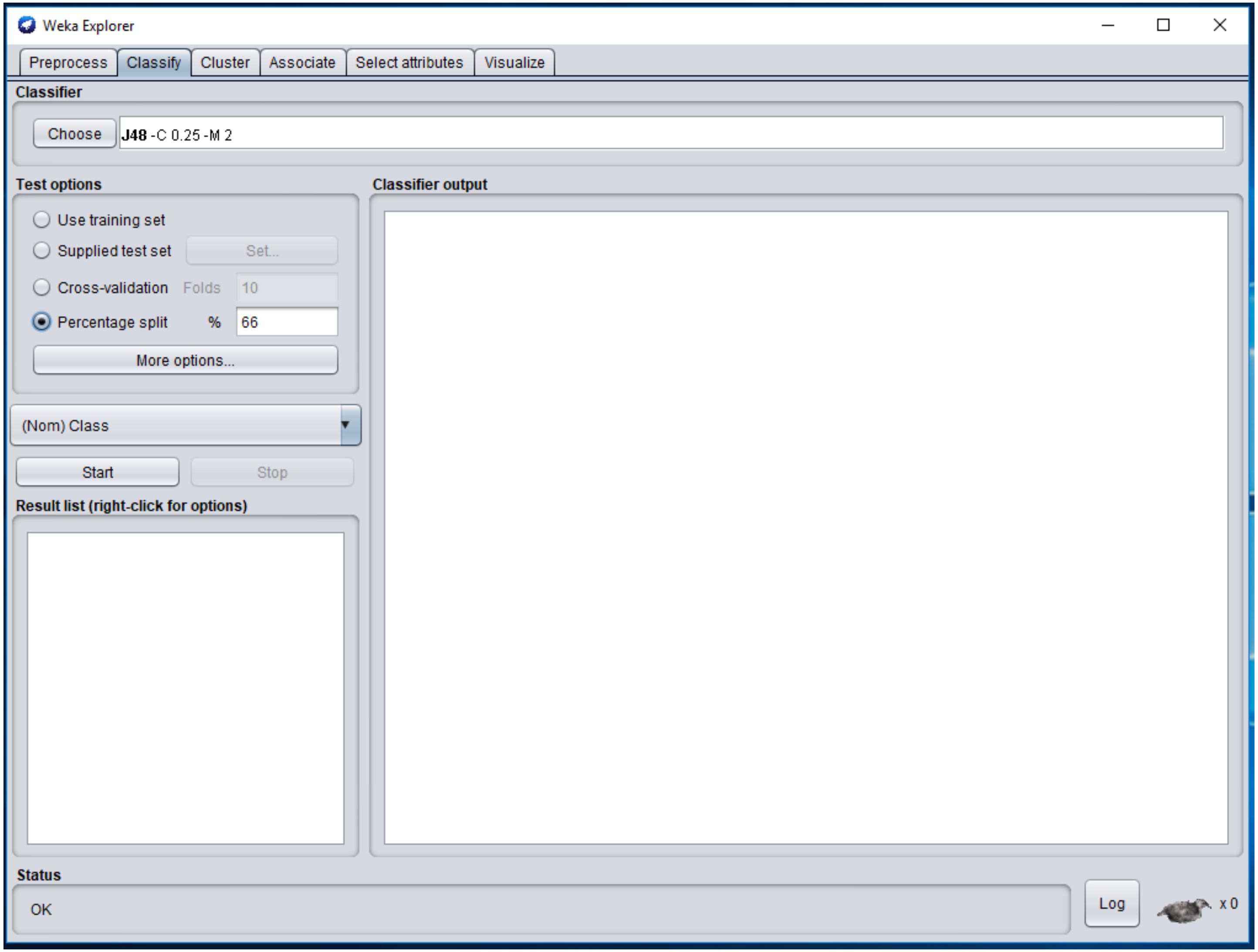
● Cross-validation Folds 10

○ Percentage split % 66

(Nom) Class

OK

Classification: An example (3)



Classifier

Choose J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set Set...

○ Cross-validation Folds 10

● Percentage split % 66

More options...

(Nom) Class

Start Stop

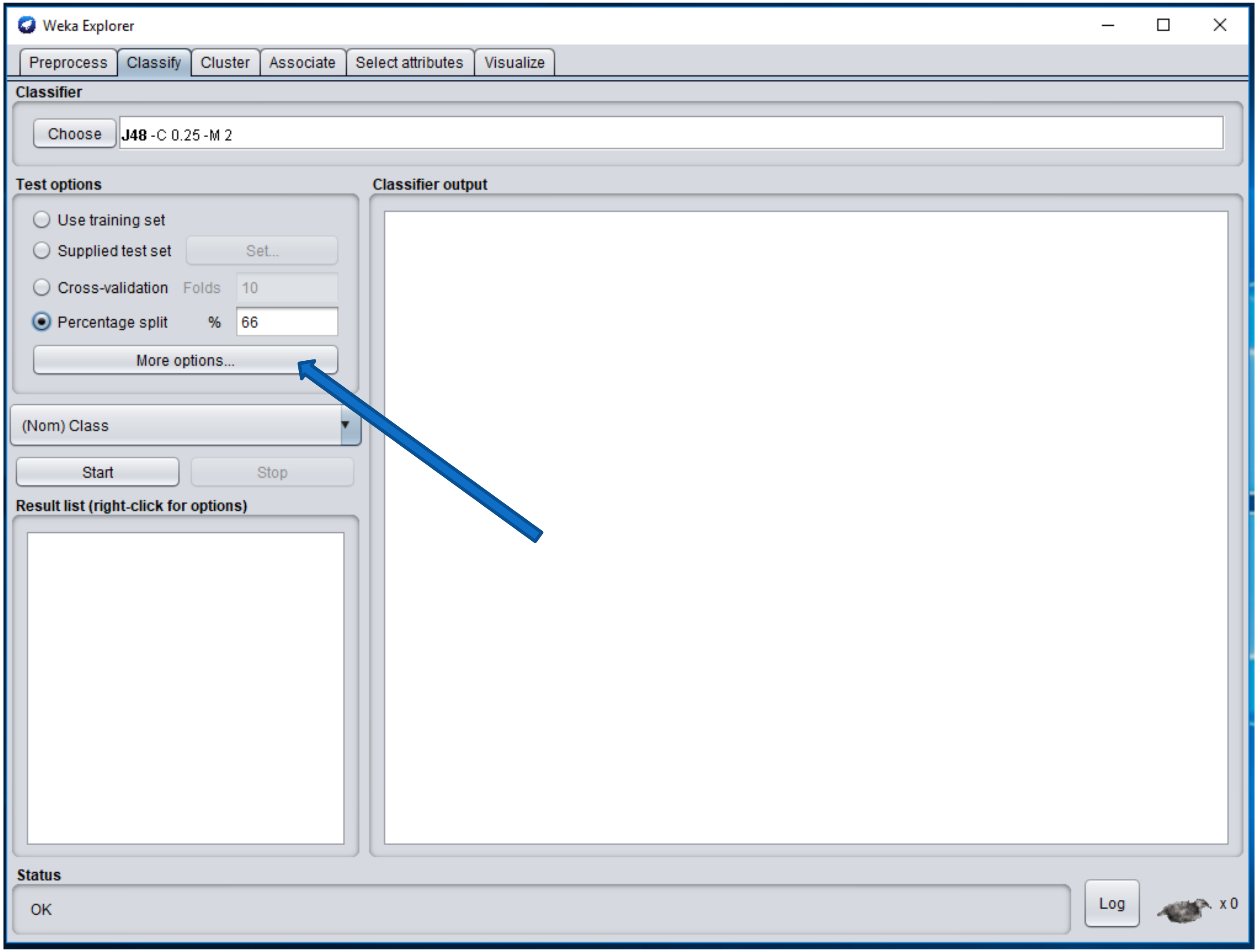
Result list (right-click for options)

Classifier output

Status

OK

Classification: An example (4)



Weka Explorer

Classifier

J48 -C 0.25 -M 2

Test options

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

(Nom) Class

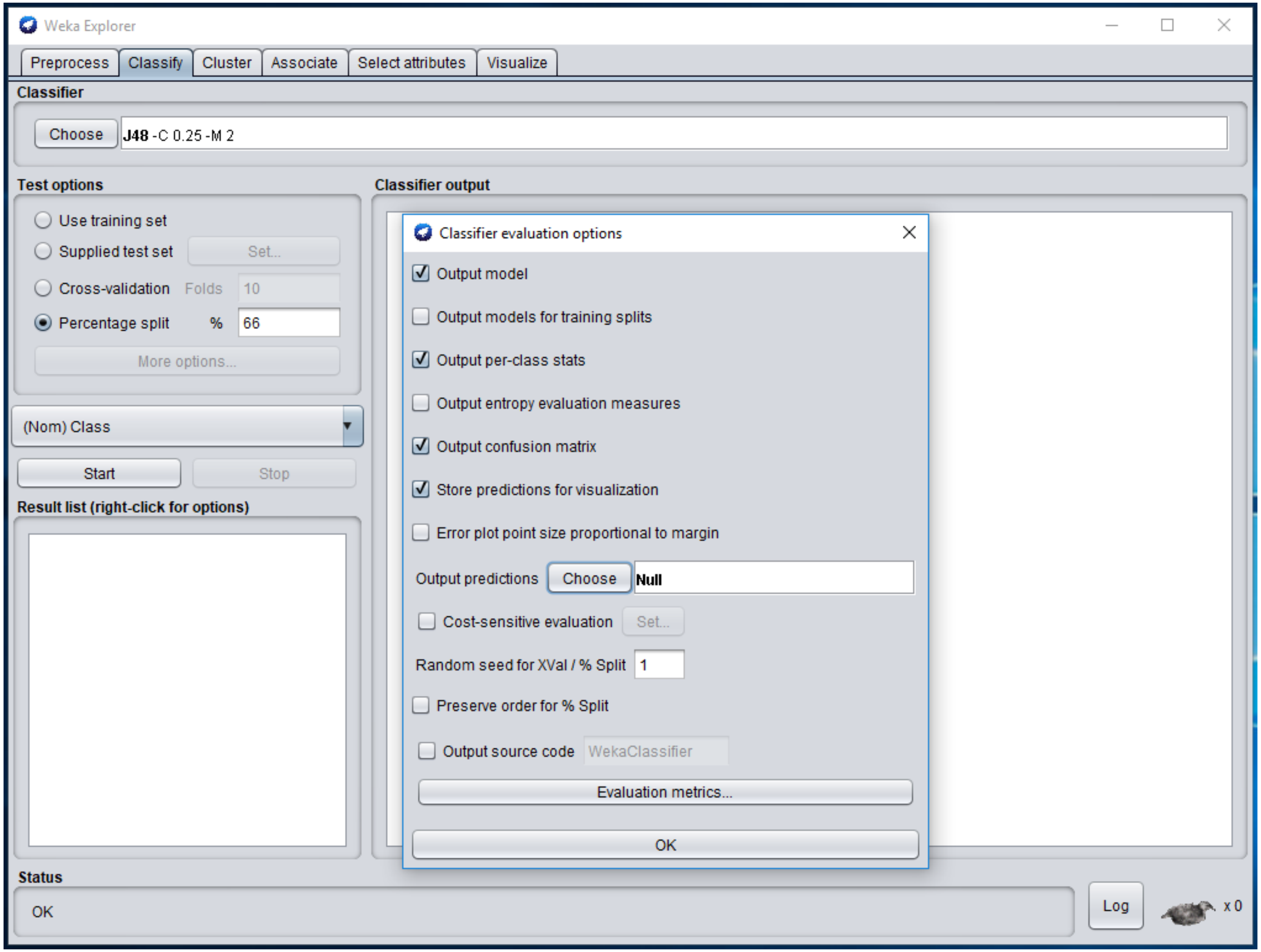
Result list (right-click for options)

Classifier output

Status

OK

Classification: An example (4)



Classifier

J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set

○ Cross-validation Folds 10

● Percentage split % 66

(Nom) Class

Classifier output

Result list (right-click for options)

Status

OK

Classifier evaluation options

☑ Output model

□ Output models for training splits

☑ Output per-class stats

□ Output entropy evaluation measures

☑ Output confusion matrix

☑ Store predictions for visualization

□ Error plot point size proportional to margin

Output predictions Choose Null

□ Cost-sensitive evaluation Set...

Random seed for XVal / % Split 1

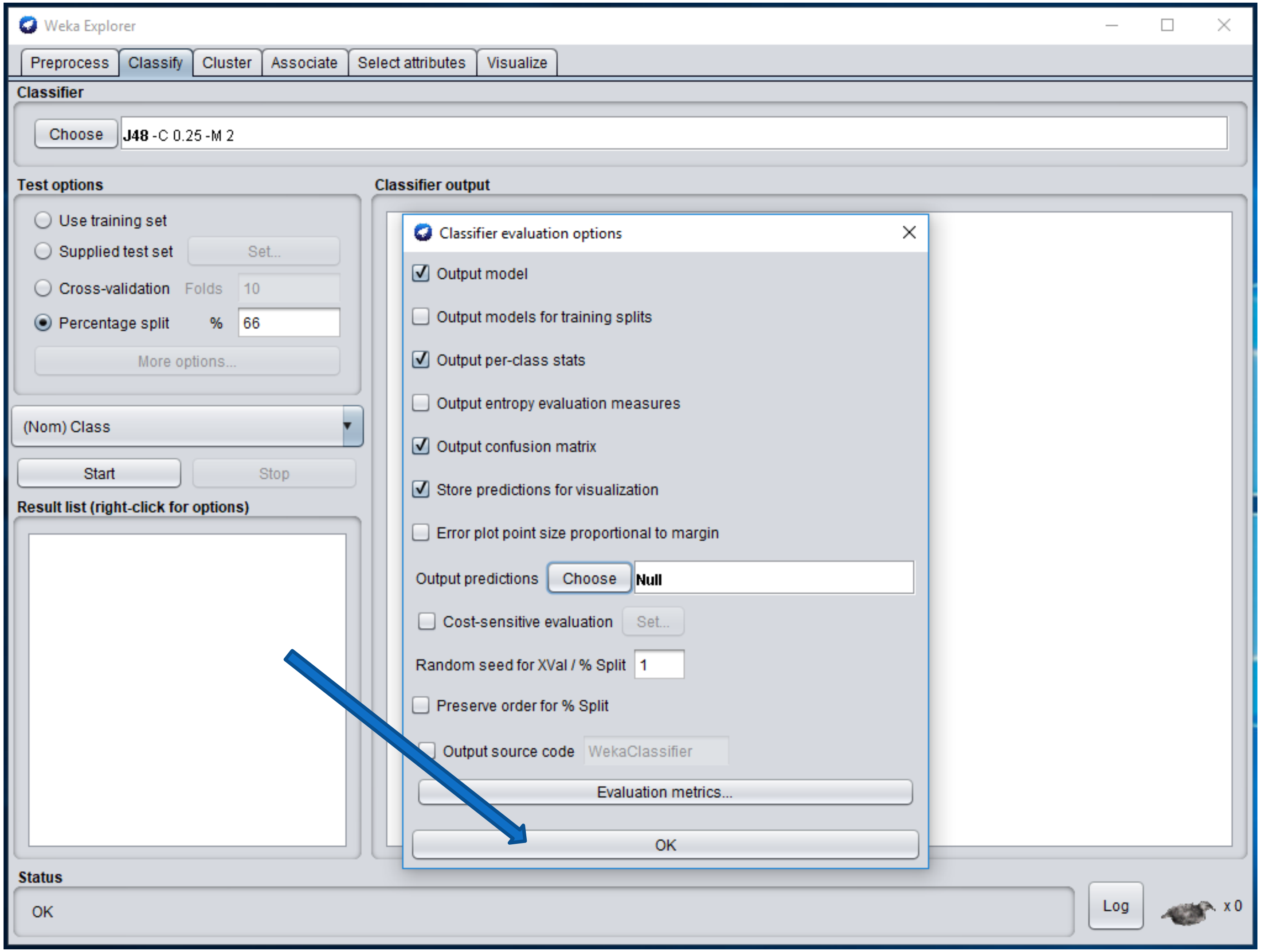
□ Preserve order for % Split

□ Output source code WekaClassifier

Evaluation metrics...

OK

Classification: An example (4)



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 list (right-click for options)

Classifier output

Classifier evaluation options

- ☑ Output model
- □ Output models for training splits
- ☑ Output per-class stats
- □ Output entropy evaluation measures
- ☑ Output confusion matrix
- ☑ Store predictions for visualization
- □ Error plot point size proportional to margin

Output predictions Choose Null

□ Cost-sensitive evaluation Set...

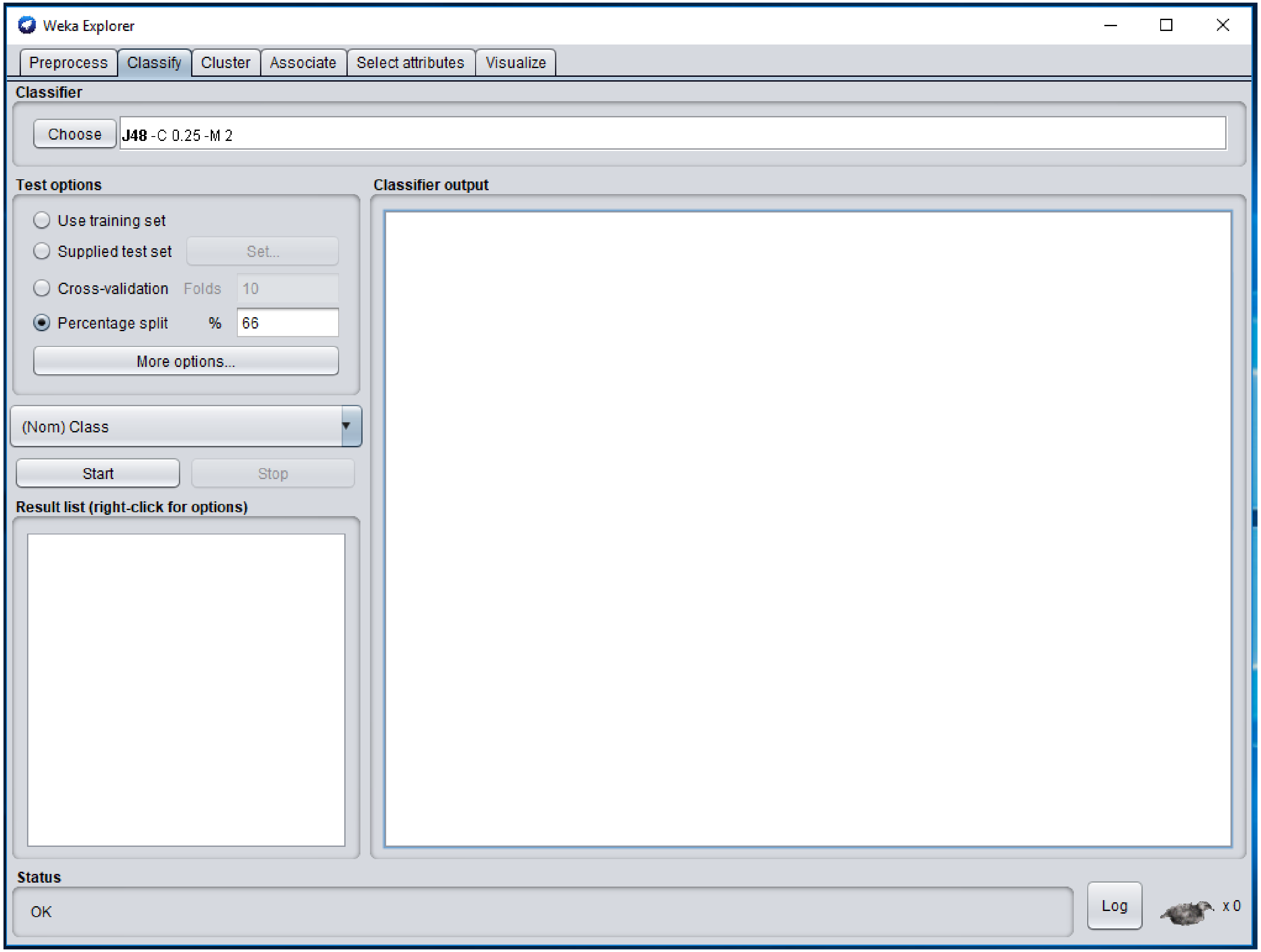
Random seed for XVal / % Split 1

- □ Preserve order for % Split
- □ Output source code WekaClassifier

Status

OK

Classification: An example (4)



Classifier: J48 -C 0.25 -M 2

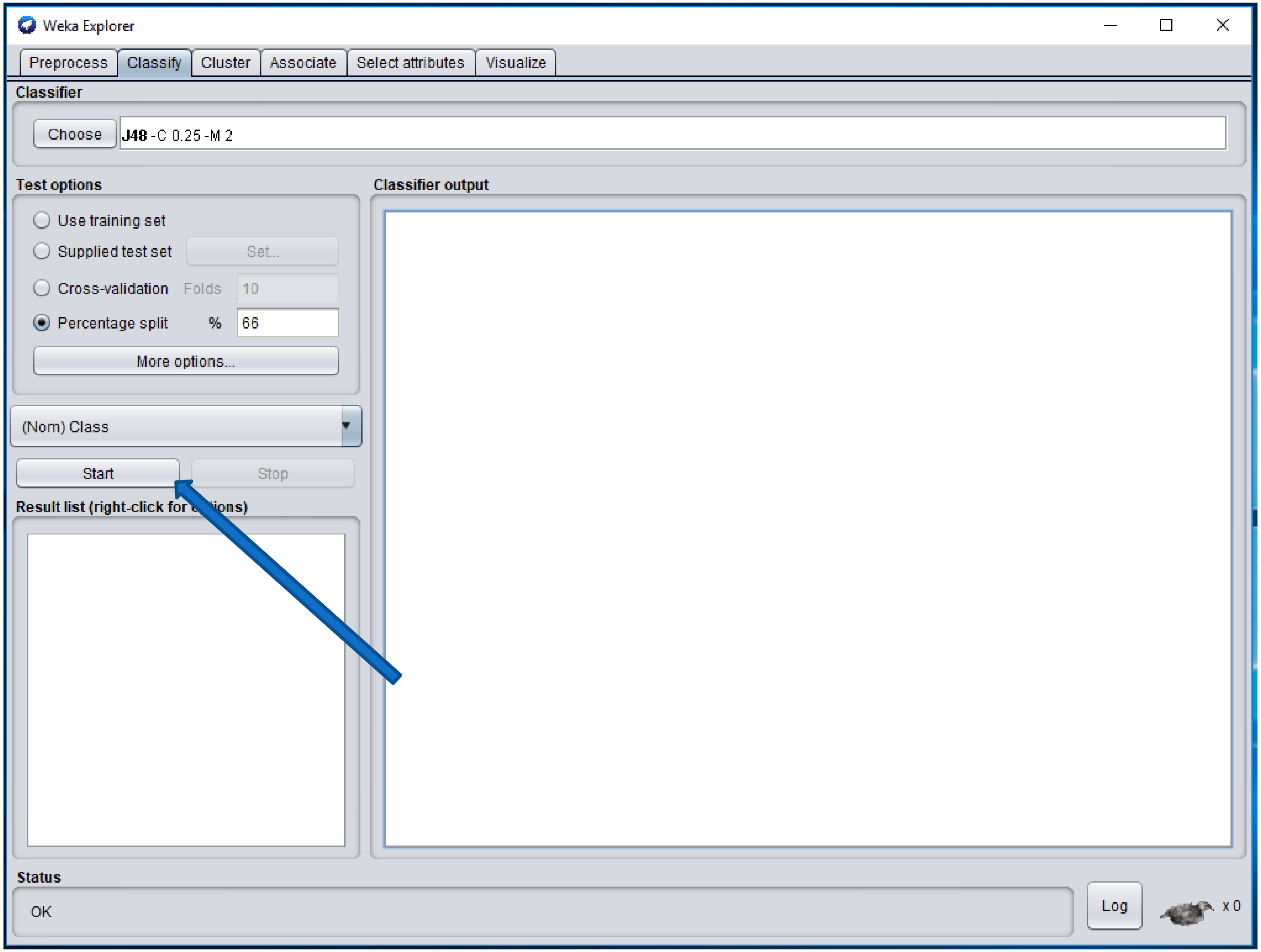
Test options:

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

(Nom) Class

Status: OK

Classification: An example (5)



J48 -C 0.25 -M 2

○ Use training set

○ Supplied test set

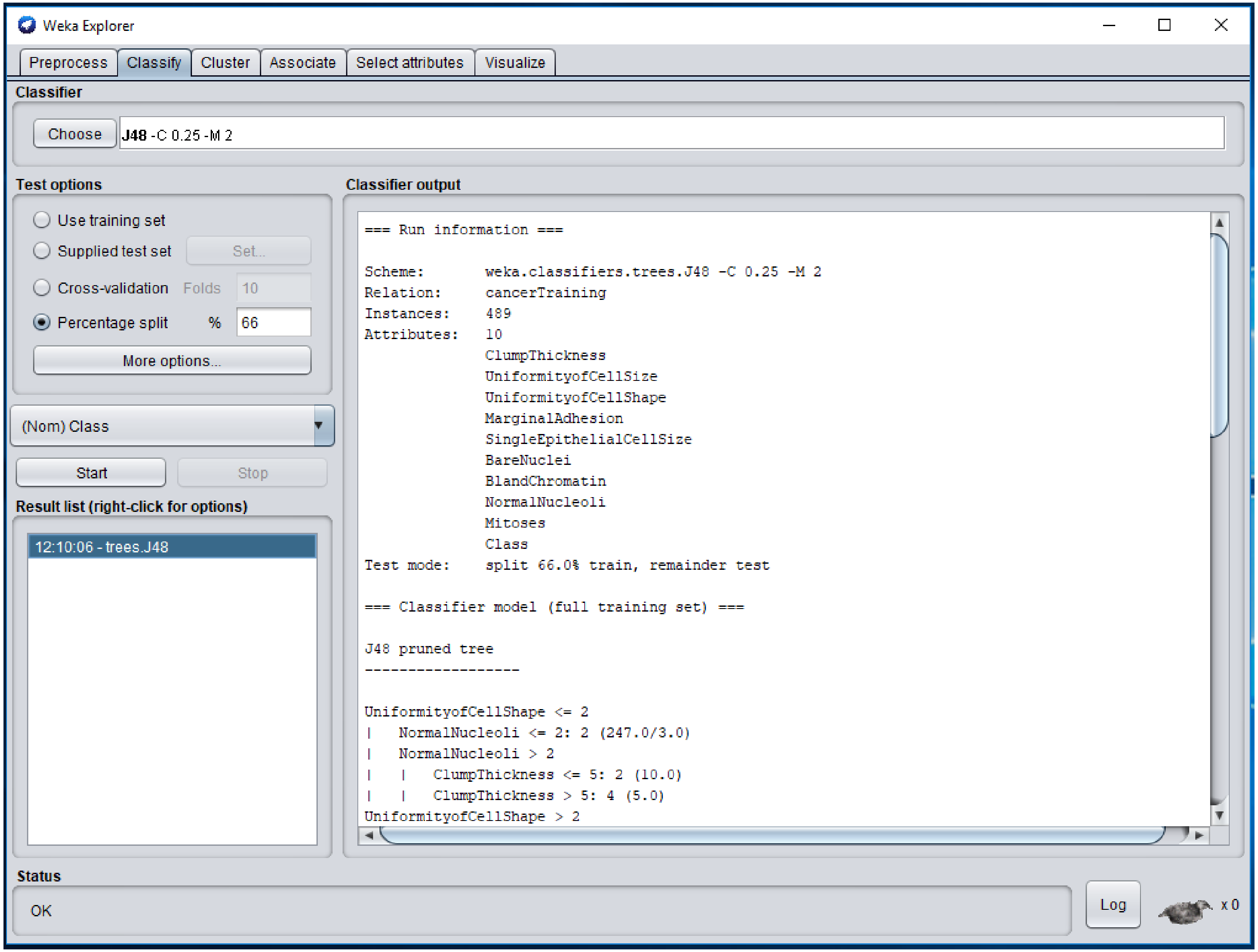
○ Cross-validation Folds 10

● Percentage split % 66

(Nom) Class

OK

Classification: An example (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)

12:10:06 - trees.J48

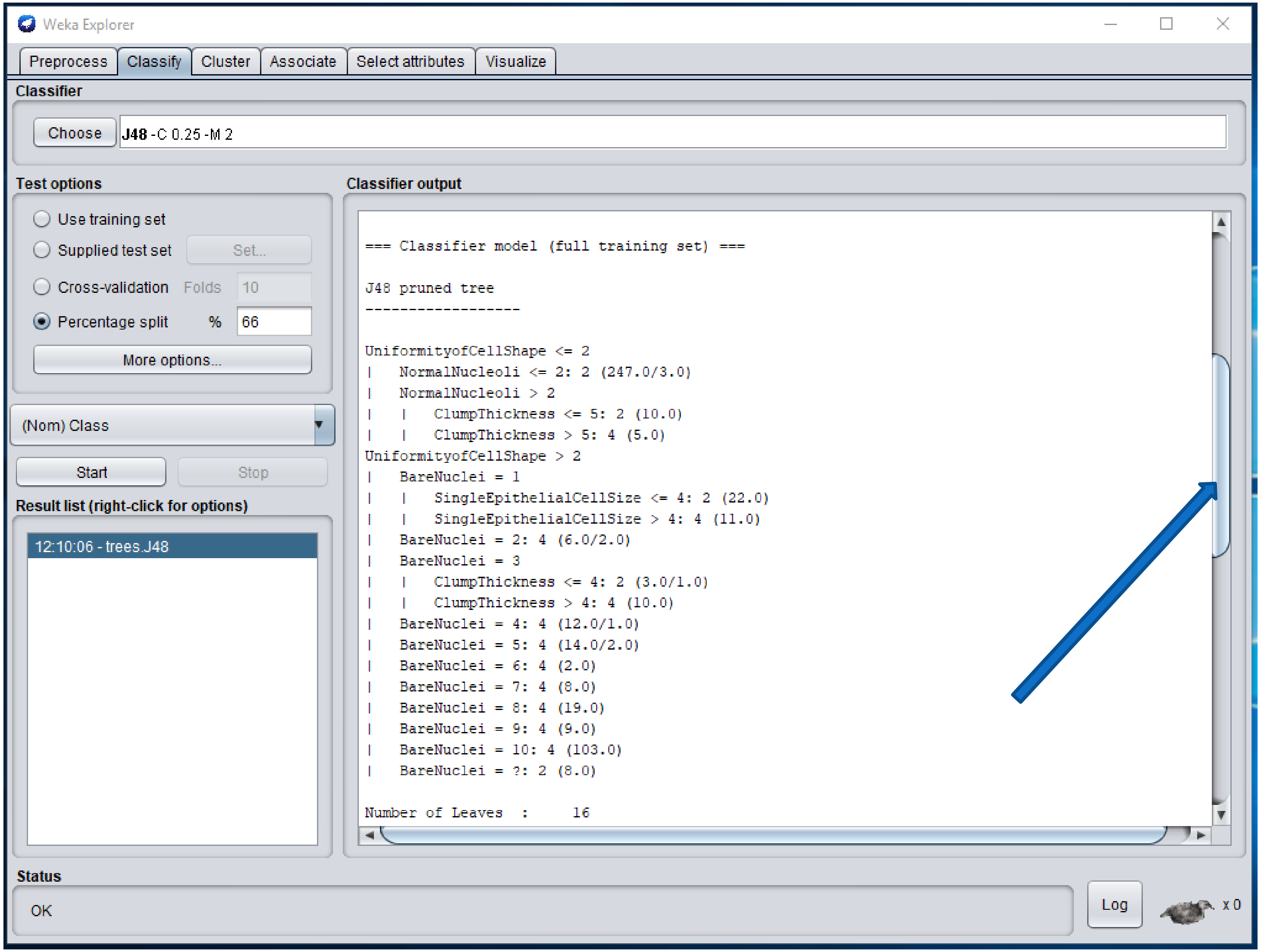
Classifier output

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     cancerTraining
Instances:    489
Attributes:   10
              ClumpThickness
              UniformityofCellSize
              UniformityofCellShape
              MarginalAdhesion
              SingleEpithelialCellSize
              BareNuclei
              BlandChromatin
              NormalNucleoli
              Mitoses
              Class
Test mode:    split 66.0% train, remainder test

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

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

UniformityofCellShape <= 2
|   NormalNucleoli <= 2: 2 (247.0/3.0)
|   NormalNucleoli > 2
|   |   ClumpThickness <= 5: 2 (10.0)
|   |   ClumpThickness > 5: 4 (5.0)
UniformityofCellShape > 2
```

Status

OK

Classification: An example (5)



J48 -C 0.25 -M 2

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

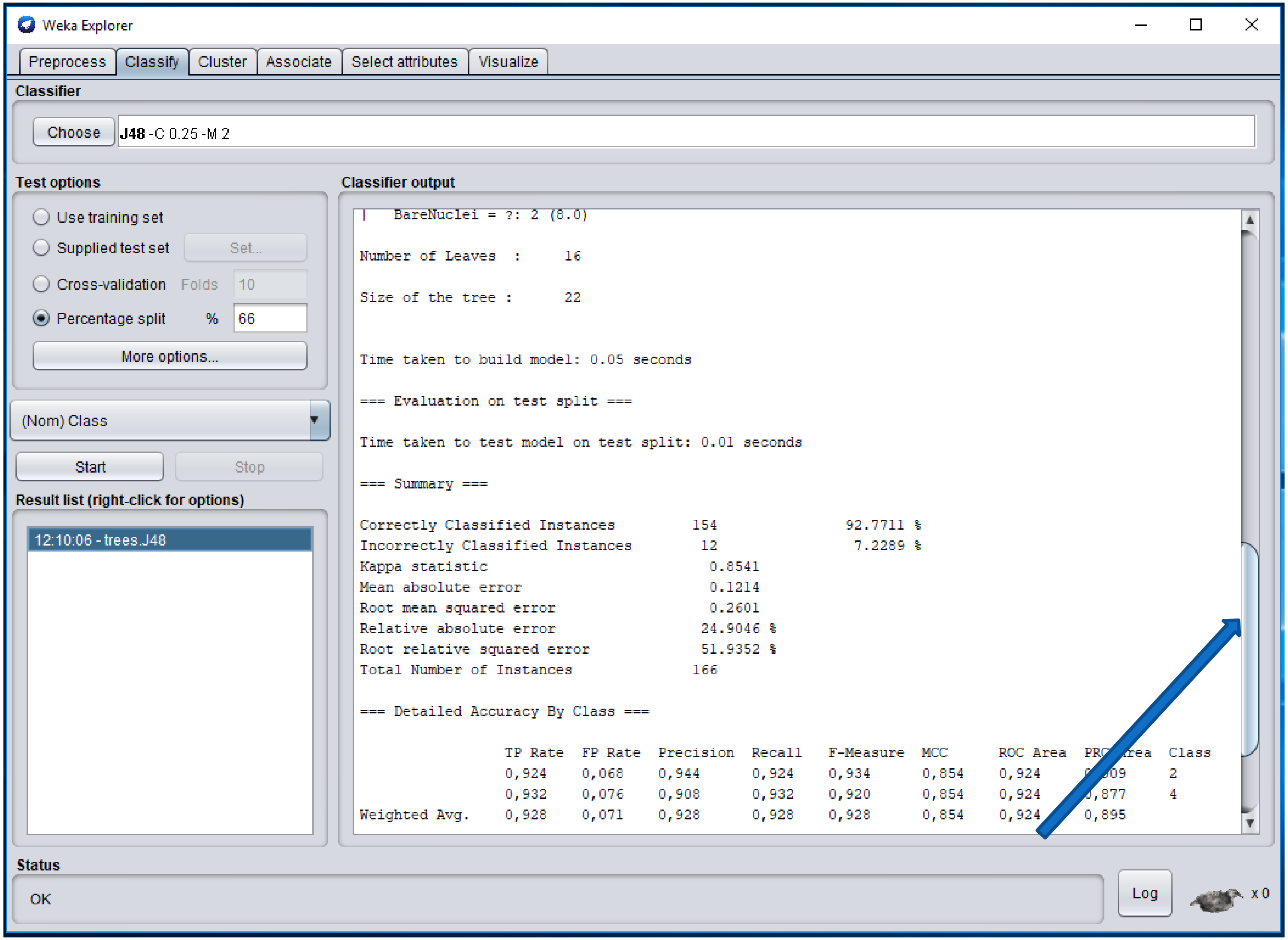
(Nom) Class

12:10:06 - trees.J48

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

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

UniformityofCellShape <= 2
|   NormalNucleoli <= 2: 2 (247.0/3.0)
|   NormalNucleoli > 2
|   |   ClumpThickness <= 5: 2 (10.0)
|   |   ClumpThickness > 5: 4 (5.0)
UniformityofCellShape > 2
|   BareNuclei = 1
|   |   SingleEpithelialCellSize <= 4: 2 (22.0)
|   |   SingleEpithelialCellSize > 4: 4 (11.0)
|   BareNuclei = 2: 4 (6.0/2.0)
|   BareNuclei = 3
|   |   ClumpThickness <= 4: 2 (3.0/1.0)
|   |   ClumpThickness > 4: 4 (10.0)
|   BareNuclei = 4: 4 (12.0/1.0)
|   BareNuclei = 5: 4 (14.0/2.0)
|   BareNuclei = 6: 4 (2.0)
|   BareNuclei = 7: 4 (8.0)
|   BareNuclei = 8: 4 (19.0)
|   BareNuclei = 9: 4 (9.0)
|   BareNuclei = 10: 4 (103.0)
|   BareNuclei = ?: 2 (8.0)

Number of Leaves  :     16
```

OK

Classification: An example (5)



Classifier: J48 -C 0.25 -M 2

Test options:

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

More options...

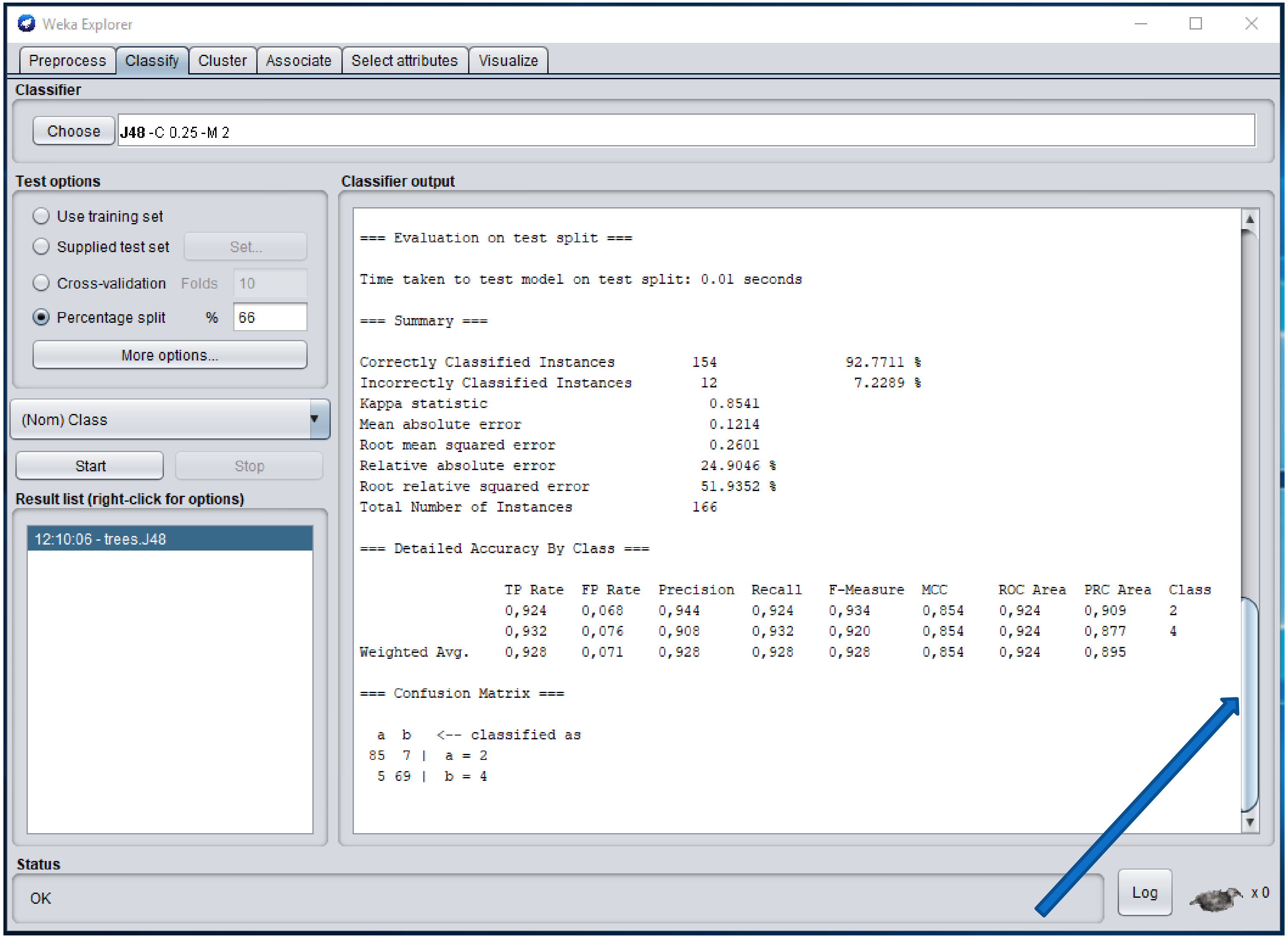
(Nom) Class

Result list: 12:10:06 - trees.J48

Status: OK

```
|   BareNuclei = ?: 2 (8.0)

Number of Leaves  :      16

Size of the tree :       22


Time taken to build model: 0.05 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.01 seconds

=== Summary ===

Correctly Classified Instances         154               92.7711 %
Incorrectly Classified Instances        12                7.2289 %
Kappa statistic                          0.8541
Mean absolute error                      0.1214
Root mean squared error                  0.2601
Relative absolute error                 24.9046 %
Root relative squared error             51.9352 %
Total Number of Instances              166

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,924    0,068    0,944      0,924    0,934      0,854    0,924     0,909     2
                 0,932    0,076    0,908      0,932    0,920      0,854    0,924     0,877     4
Weighted Avg.    0,928    0,071    0,928      0,928    0,928      0,854    0,924     0,895
```

Classification: An example (5)



J48 -C 0.25 -M 2

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

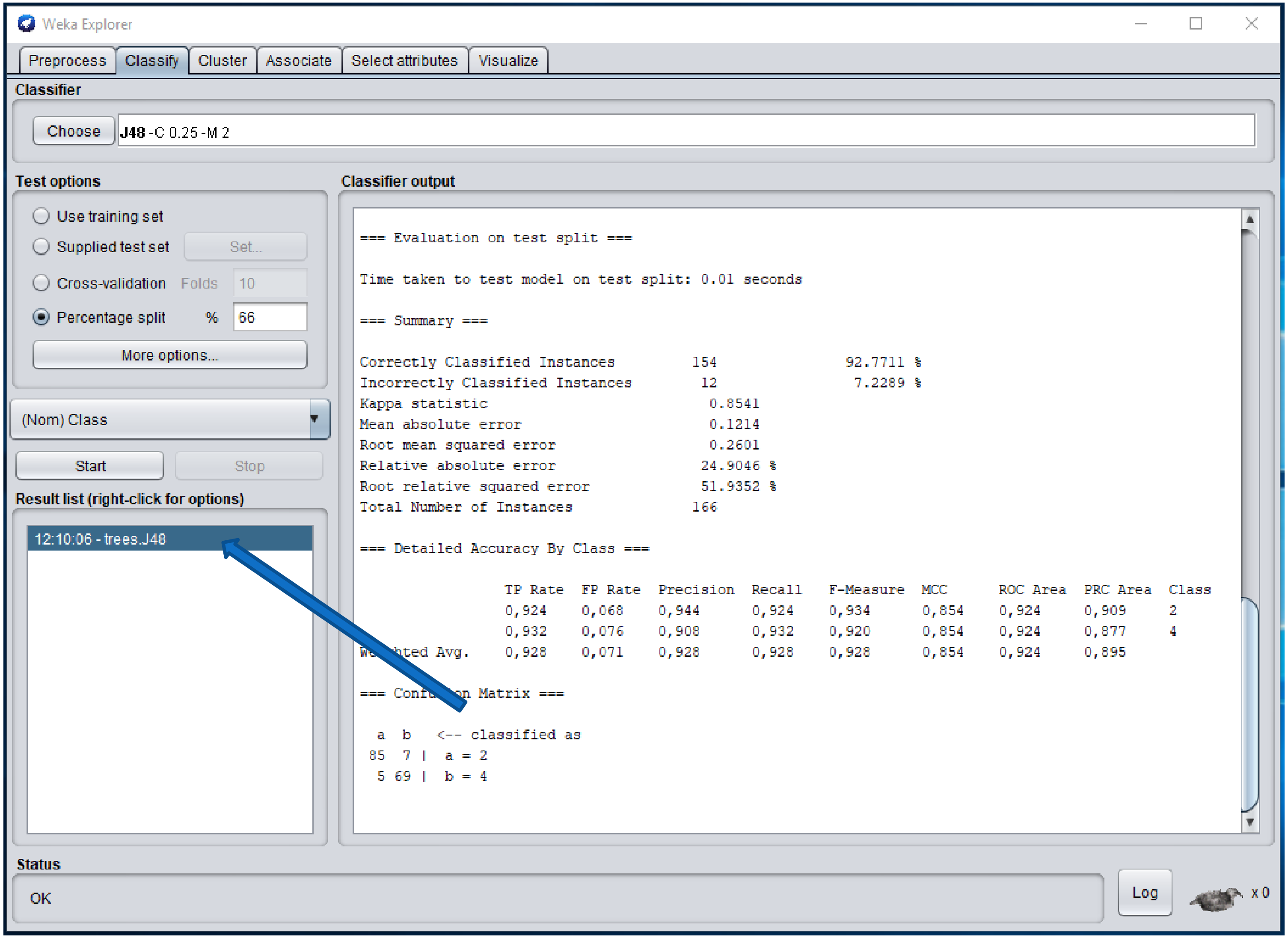
(Nom) Class

12:10:06 - trees.J48

```
=== Evaluation on test split ===

Time taken to test model on test split: 0.01 seconds

=== Summary ===

Correctly Classified Instances         154               92.7711 %
Incorrectly Classified Instances        12                7.2289 %
Kappa statistic                          0.8541
Mean absolute error                      0.1214
Root mean squared error                  0.2601
Relative absolute error                 24.9046 %
Root relative squared error             51.9352 %
Total Number of Instances              166

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,924    0,068    0,944      0,924    0,934      0,854    0,924     0,909     2
                 0,932    0,076    0,908      0,932    0,920      0,854    0,924     0,877     4
Weighted Avg.    0,928    0,071    0,928      0,928    0,928      0,854    0,924     0,895

=== Confusion Matrix ===

  a  b   <-- classified as
 85  7 |  a = 2
  5 69 |  b = 4
```

Status: OK

Classification: An example (6)



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)

12:10:06 - trees.J48

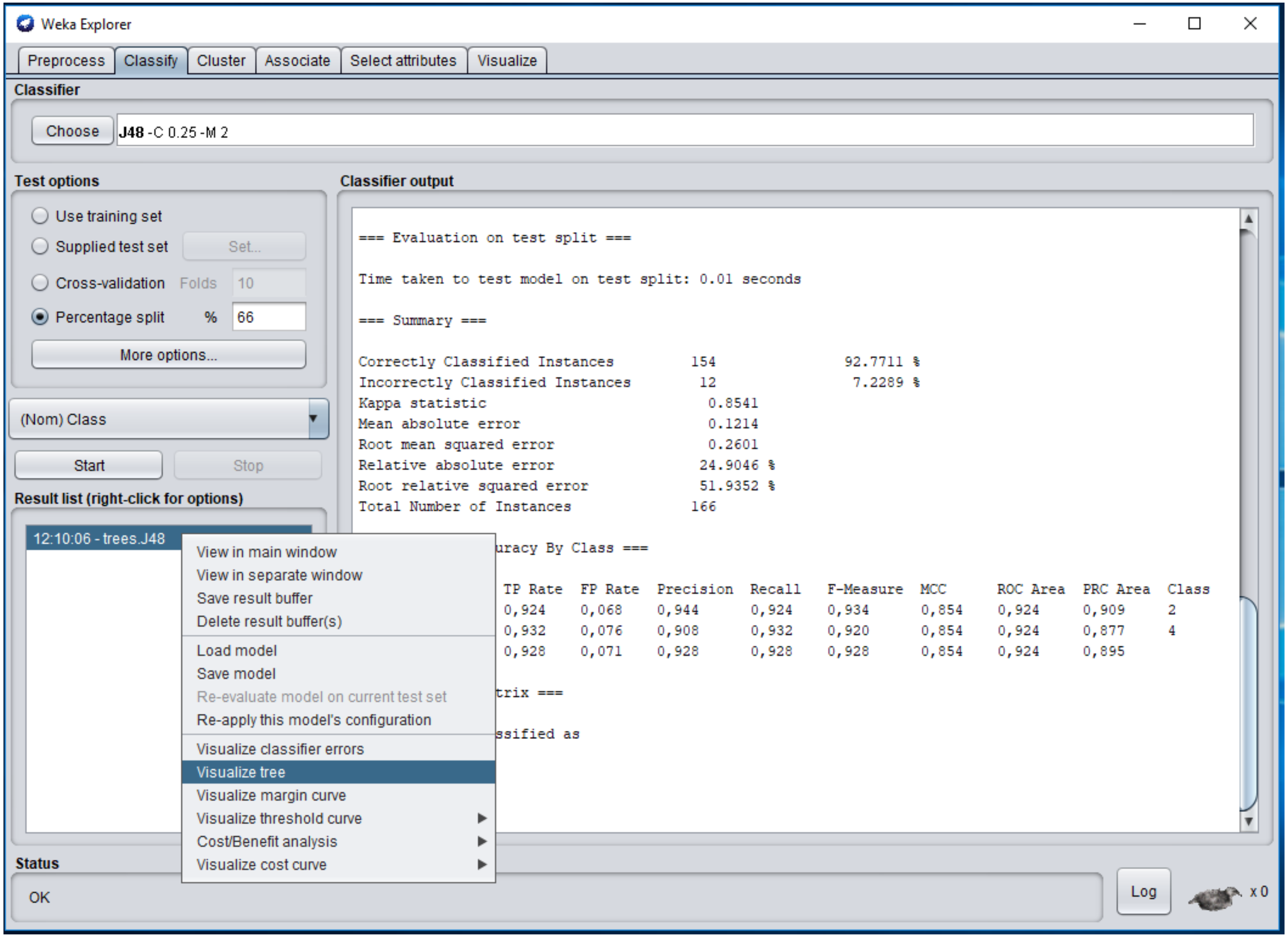
Classifier output

```
=== Evaluation on test split ===

Time taken to test model on test split: 0.01 seconds

=== Summary ===

Correctly Classified Instances         154               92.7711 %
Incorrectly Classified Instances        12                7.2289 %
Kappa statistic                          0.8541
Mean absolute error                      0.1214
Root mean squared error                  0.2601
Relative absolute error                 24.9046 %
Root relative squared error             51.9352 %
Total Number of Instances              166

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,924    0,068    0,944      0,924    0,934      0,854    0,924     0,909     2
                 0,932    0,076    0,908      0,932    0,920      0,854    0,924     0,877     4
Weighted Avg.    0,928    0,071    0,928      0,928    0,928      0,854    0,924     0,895

=== Confusion Matrix ===

  a  b   <-- classified as
 85  7 |  a = 2
  5 69 |  b = 4
```

Status

OK

Classification: An example (6)



Classifier: J48 -C 0.25 -M 2

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

(Nom) Class

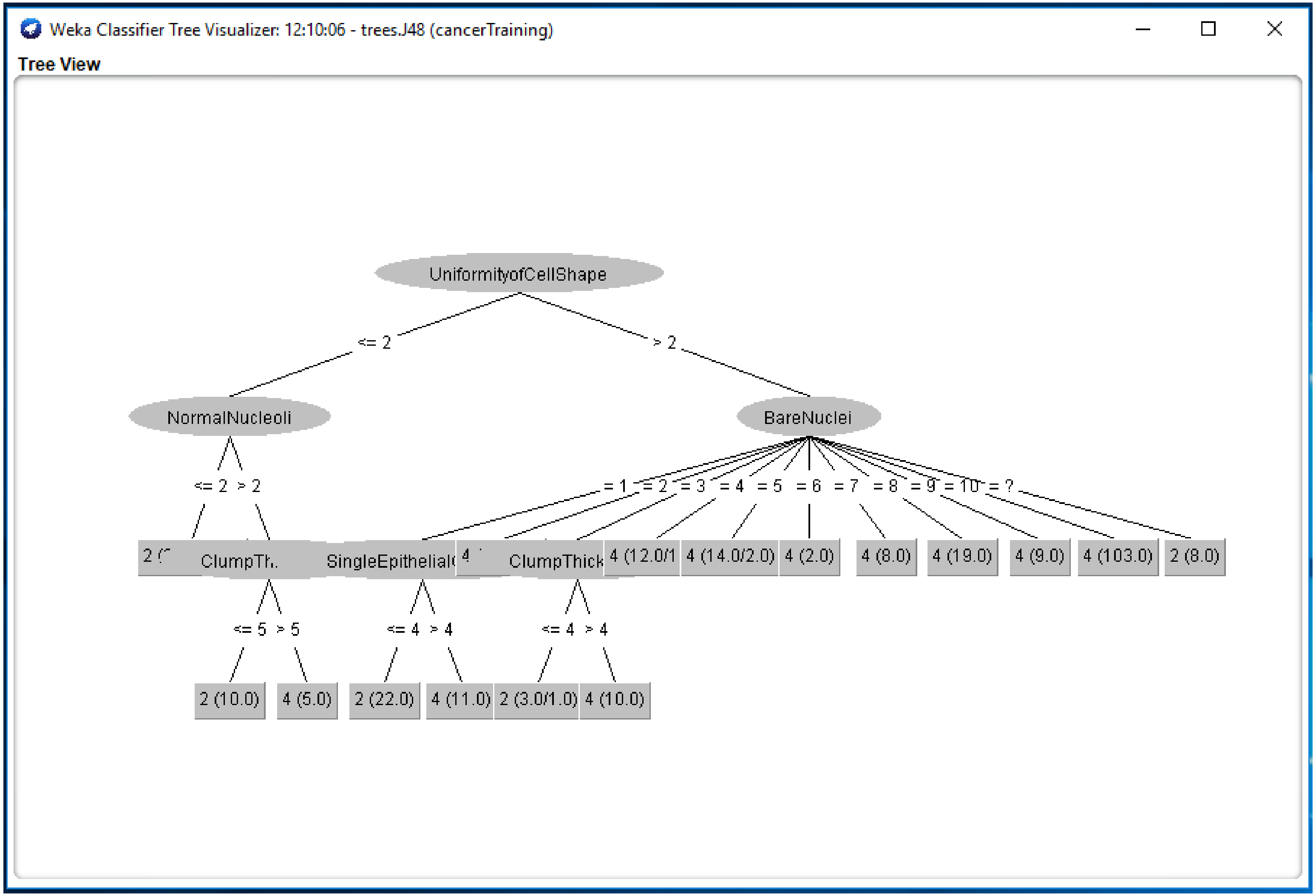
12:10:06 - trees.J48

Status: OK

Classifier output

```
=== Evaluation on test split ===

Time taken to test model on test split: 0.01 seconds

=== Summary ===

Correctly Classified Instances         154               92.7711 %
Incorrectly Classified Instances        12                7.2289 %
Kappa statistic                          0.8541
Mean absolute error                      0.1214
Root mean squared error                  0.2601
Relative absolute error                 24.9046 %
Root relative squared error             51.9352 %
Total Number of Instances              166

uracy By Class ===

TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
0,924    0,068    0,944      0,924    0,934      0,854    0,924     0,909     2
0,932    0,076    0,908      0,932    0,920      0,854    0,924     0,877     4
0,928    0,071    0,928      0,928    0,928      0,854    0,924     0,895

trix ===

ssified as
```

Classification: An example (7)



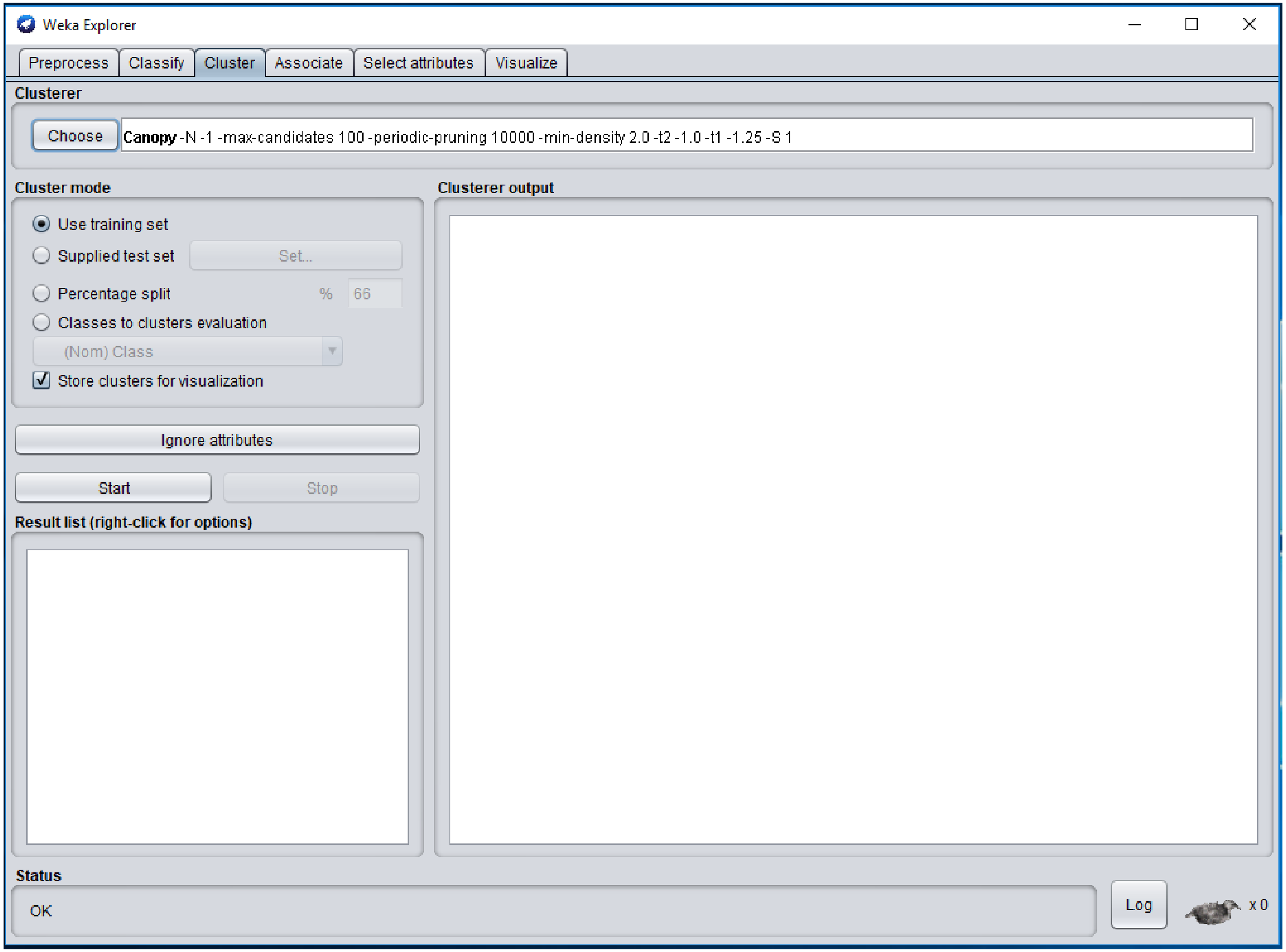
Clustering (1)

- WEKA contains “clusters” 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

Clustering (2)

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

Clustering: An example (1)



Clusterer

Canopy -N -1 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t2 -1.0 -t1 -1.25 -S 1

Cluster mode

● Use training set
○ Supplied test set
○ Percentage split % 66
○ Classes to clusters evaluation
(Nom) Class
☑ Store clusters for visualization

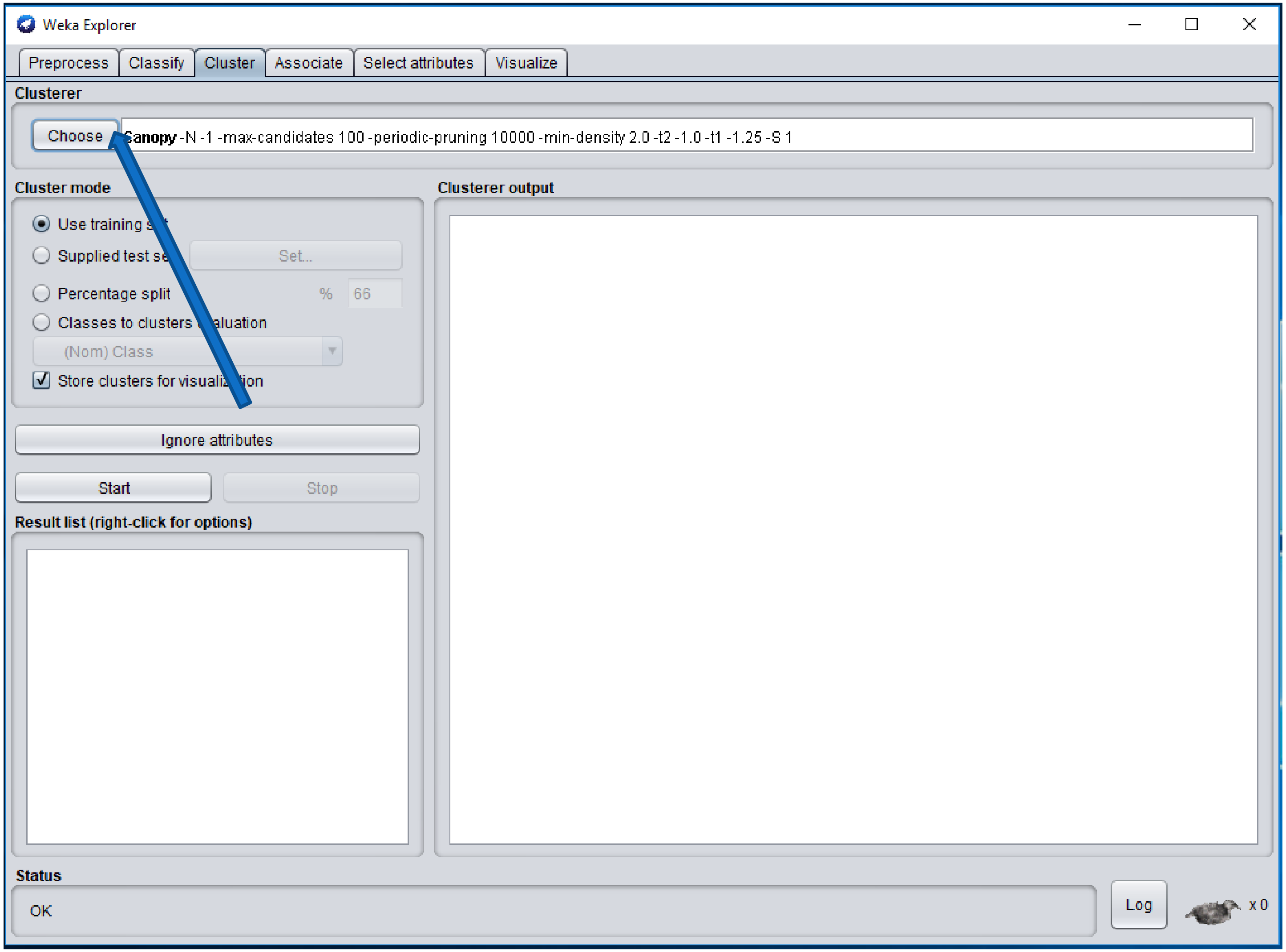
Clusterer output

Result list (right-click for options)

Status

OK

Clustering: An example (1)



Clusterer

Canopy -N -1 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t2 -1.0 -t1 -1.25 -S 1

Cluster mode

● Use training set

○ Supplied test set

○ Percentage split % 66

○ Classes to clusters evaluation

(Nom) Class

☑ Store clusters for visualization

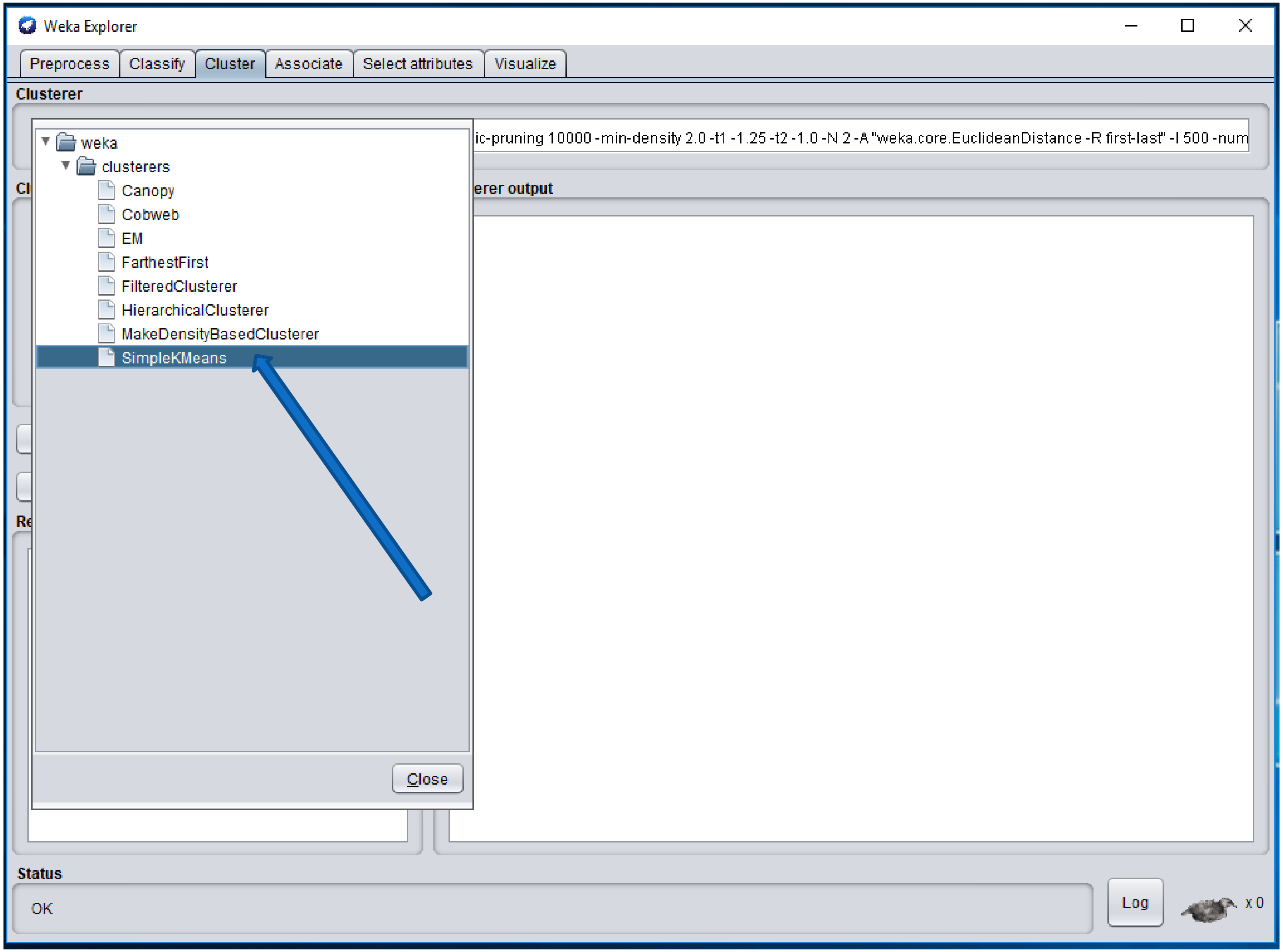
Clusterer output

Result list (right-click for options)

Status

OK

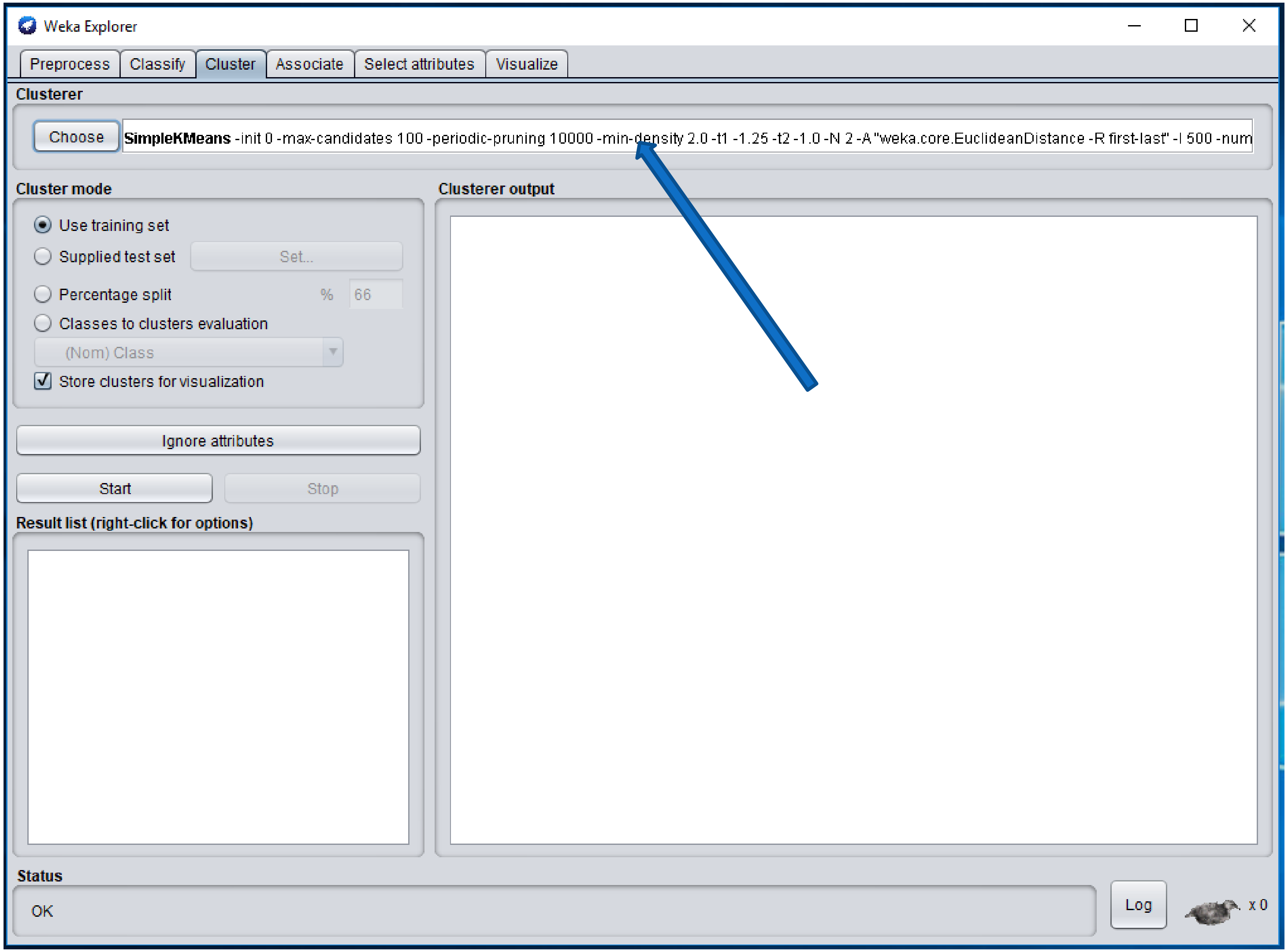
Clustering: An example (2)



ic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num

OK

Clustering: An example (2)

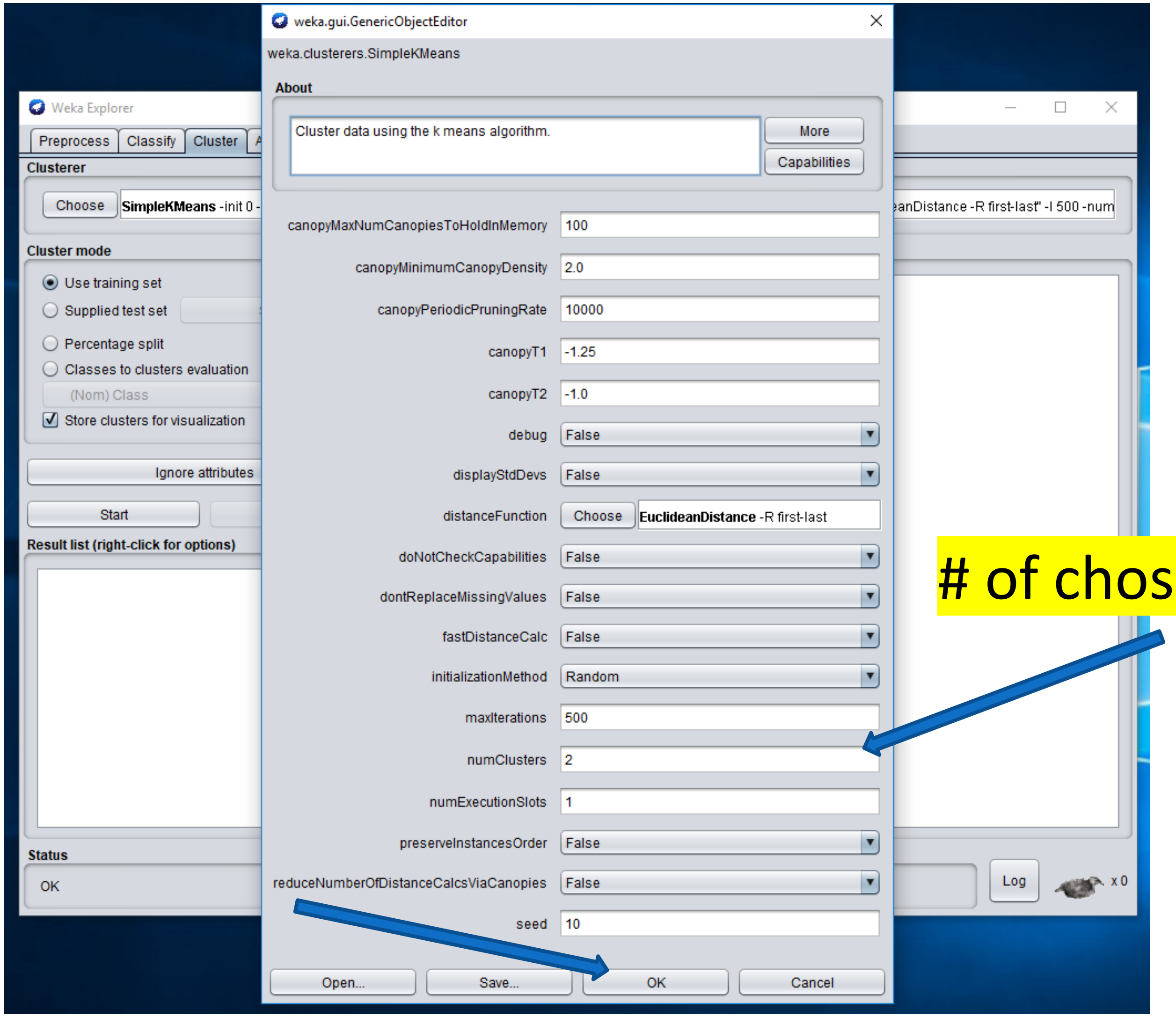


SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num

- ● Use training set
- ○ Supplied test set
- ○ Percentage split % 66
- ○ Classes to clusters evaluation
- (Nom) Class
- ☑ Store clusters for visualization

Status: OK

Clustering: An example (3)



weka.gui.GenericObjectEditor

weka.clusterers.SimpleKMeans

About

Cluster data using the k means algorithm.

canopyMaxNumCanopiesToHoldInMemory: 100

canopyMinimumCanopyDensity: 2.0

canopyPeriodicPruningRate: 10000

canopyT1: -1.25

canopyT2: -1.0

debug: False

displayStdDevs: False

distanceFunction: **EuclideanDistance** -R first-last

doNotCheckCapabilities: False

dontReplaceMissingValues: False

fastDistanceCalc: False

initializationMethod: Random

maxIterations: 500

numClusters: 2

numExecutionSlots: 1

preserveInstancesOrder: False

reduceNumberOfDistanceCalcsViaCanopies: False

seed: 10

Weka Explorer

Clusterer: **SimpleKMeans** -init 0 - … anDistance -R first-last" -I 500 -num

Cluster mode

● Use training set

○ Supplied test set

○ Percentage split

○ Classes to clusters evaluation

(Nom) Class

☑ Store clusters for visualization

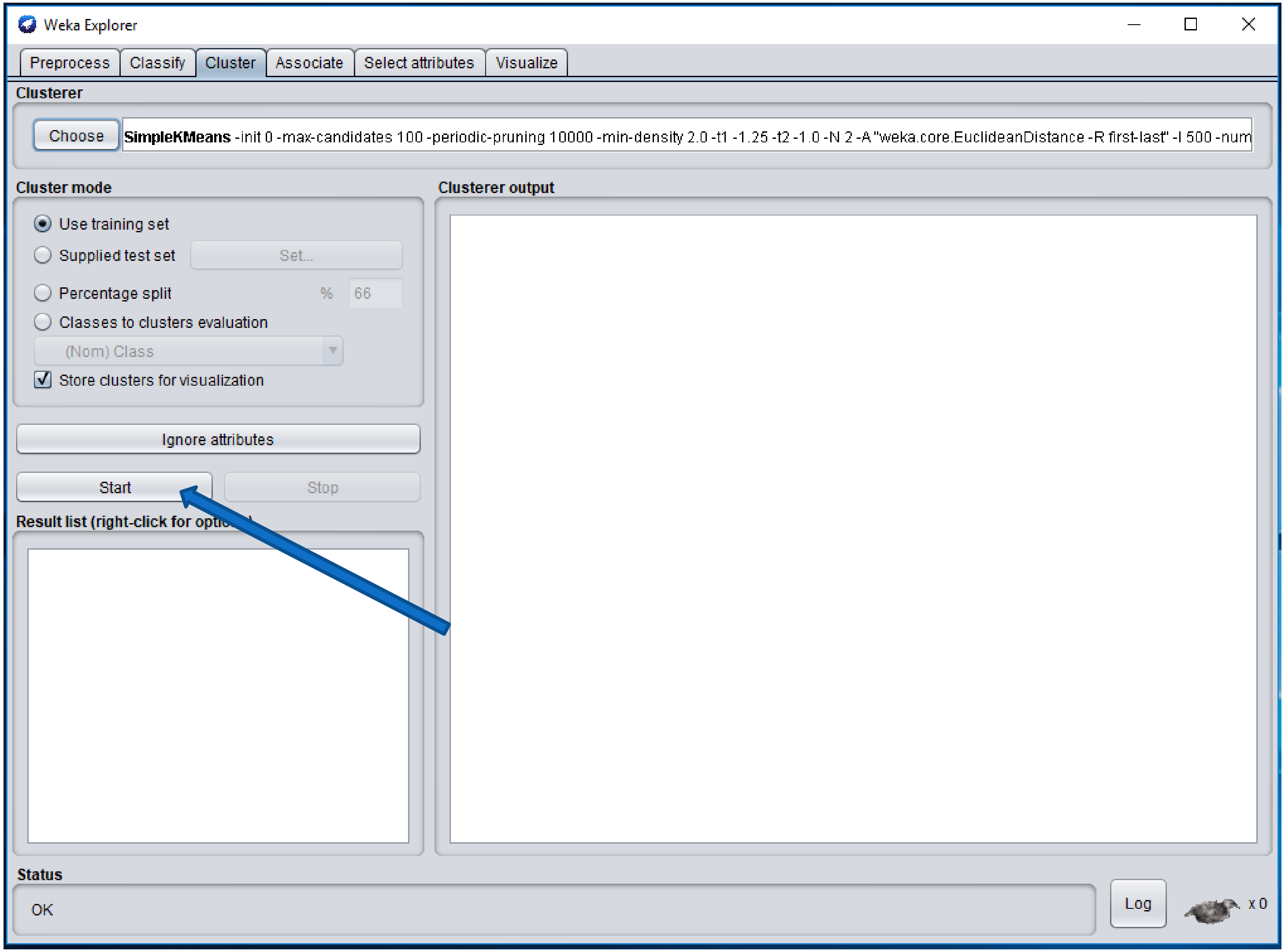
Result list (right-click for options)

Status

OK

# of chosen clusters

Clustering: An example (4)



Clusterer

SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num

Cluster mode

- ● Use training set
- ○ Supplied test set
- ○ Percentage split % 66
- ○ Classes to clusters evaluation
- (Nom) Class
- ☑ Store clusters for visualization

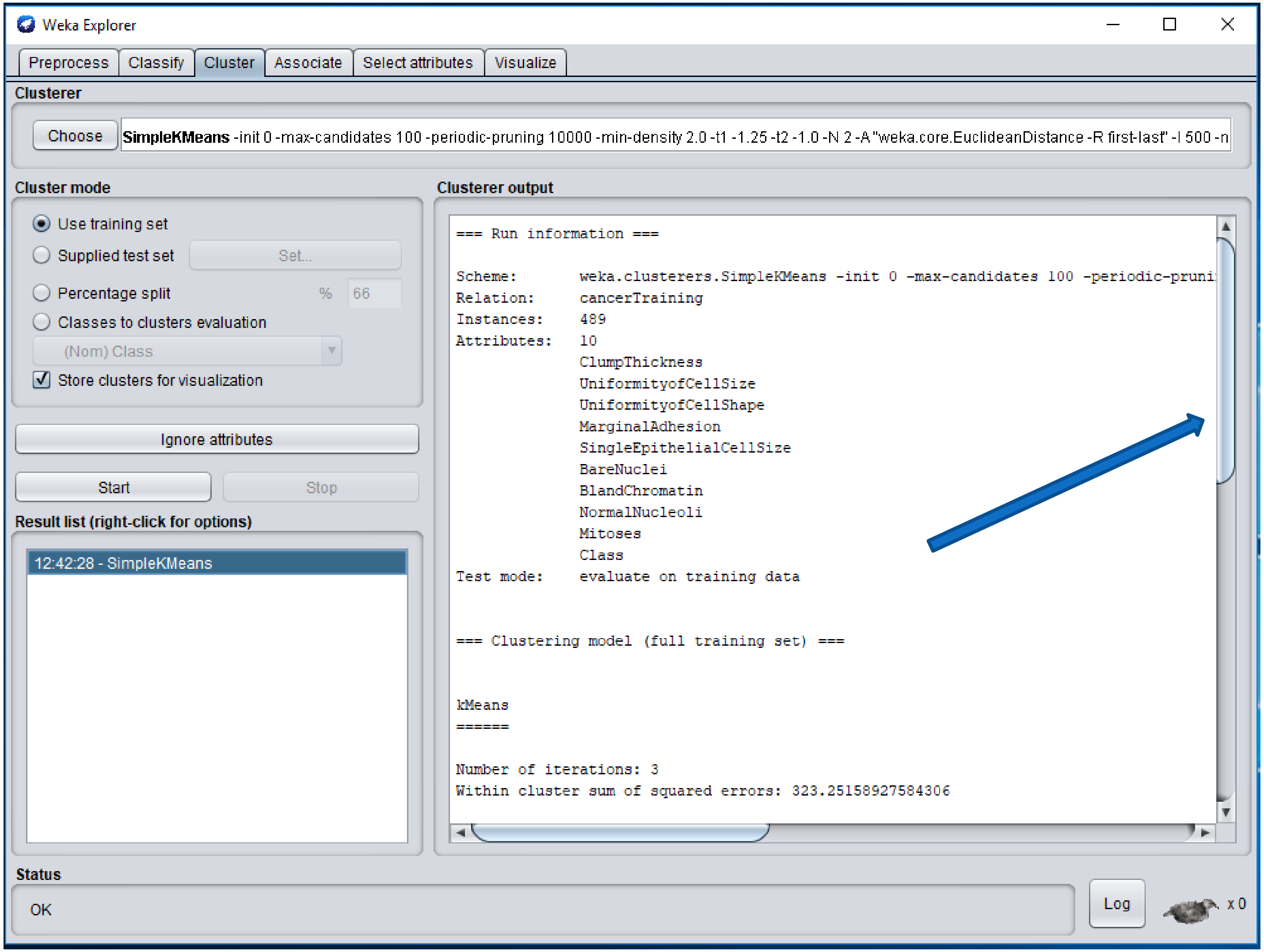
Clusterer output

Result list (right-click for options)

Status

OK

Clustering: An example (5)



SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -n

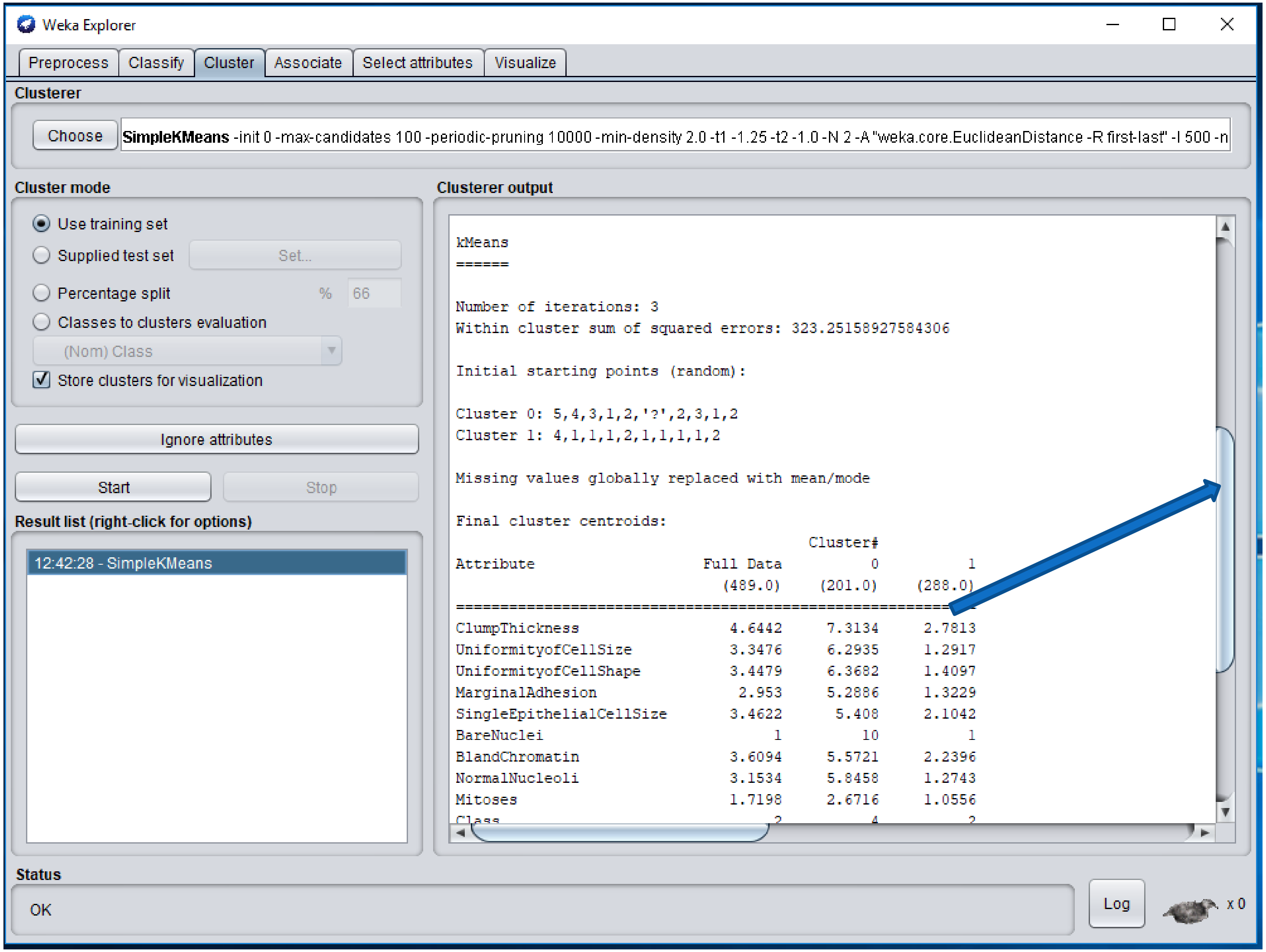
● Use training set
○ Supplied test set
○ Percentage split % 66
○ Classes to clusters evaluation
(Nom) Class
☑ Store clusters for visualization

12:42:28 - SimpleKMeans

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruni
Relation:     cancerTraining
Instances:    489
Attributes:   10
              ClumpThickness
              UniformityofCellSize
              UniformityofCellShape
              MarginalAdhesion
              SingleEpithelialCellSize
              BareNuclei
              BlandChromatin
              NormalNucleoli
              Mitoses
              Class
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 323.25158927584306
```

OK

Clustering: An example (5)



SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -n

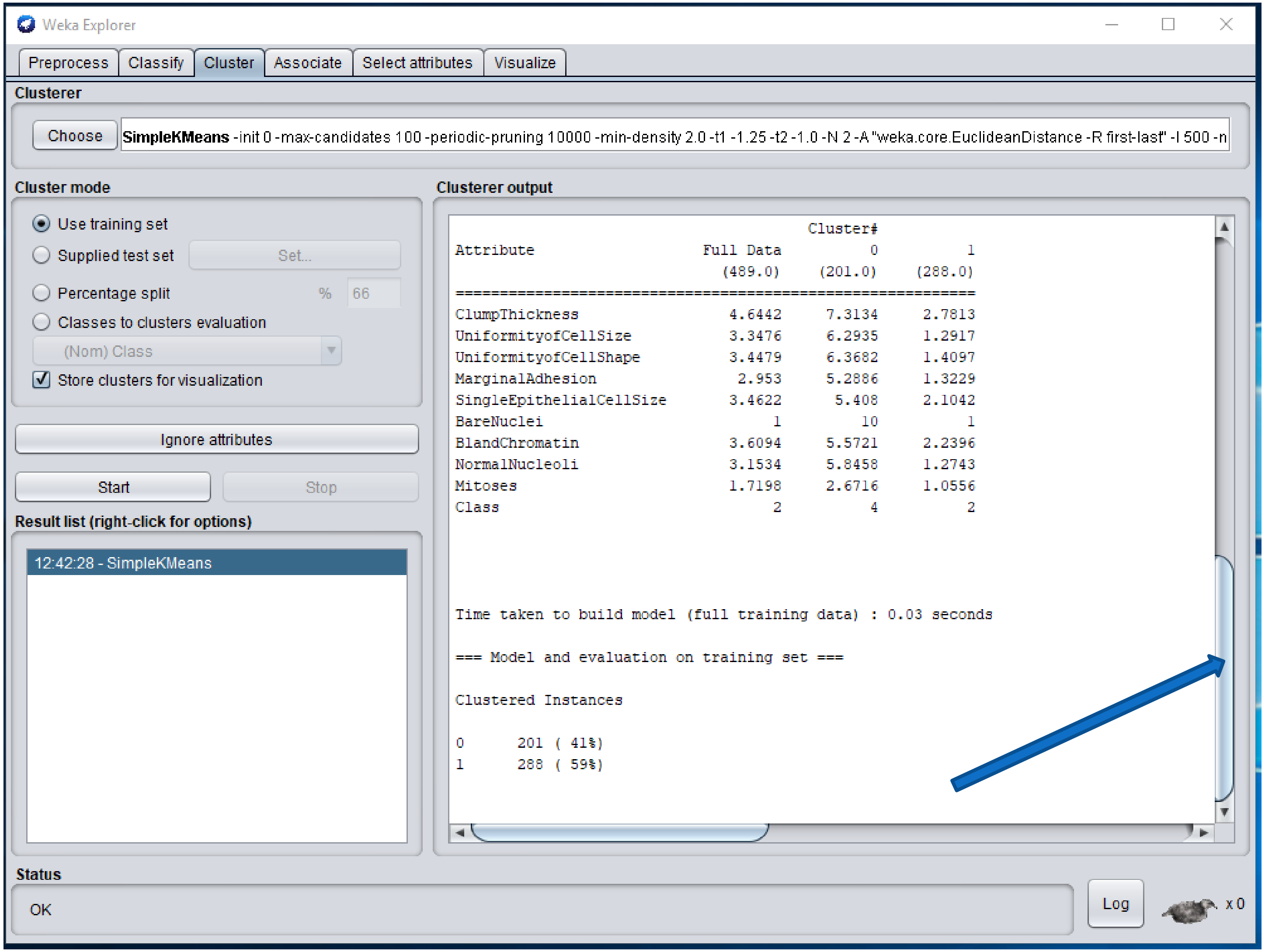
- ● Use training set
- ○ Supplied test set
- ○ Percentage split % 66
- ○ Classes to clusters evaluation
- ☑ Store clusters for visualization

12:42:28 - SimpleKMeans

```
kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 323.25158927584306

Initial starting points (random):

Cluster 0: 5,4,3,1,2,'?',2,3,1,2
Cluster 1: 4,1,1,1,2,1,1,1,1,2

Missing values globally replaced with mean/mode

Final cluster centroids:
                                        Cluster#
Attribute                   Full Data          0          1
                              (489.0)    (201.0)    (288.0)
===========================================================
ClumpThickness                 4.6442     7.3134     2.7813
UniformityofCellSize           3.3476     6.2935     1.2917
UniformityofCellShape          3.4479     6.3682     1.4097
MarginalAdhesion                2.953     5.2886     1.3229
SingleEpithelialCellSize       3.4622      5.408     2.1042
BareNuclei                          1         10          1
BlandChromatin                 3.6094     5.5721     2.2396
NormalNucleoli                 3.1534     5.8458     1.2743
Mitoses                        1.7198     2.6716     1.0556
Class                               2          4          2
```

OK

Clustering: An example (5)



Clusterer

Choose **SimpleKMeans** -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -n

Cluster mode

- ● Use training set
- ○ Supplied test set Set...
- ○ Percentage split % 66
- ○ Classes to clusters evaluation
- (Nom) Class
- ☑ Store clusters for visualization

Result list (right-click for options)

12:42:28 - SimpleKMeans

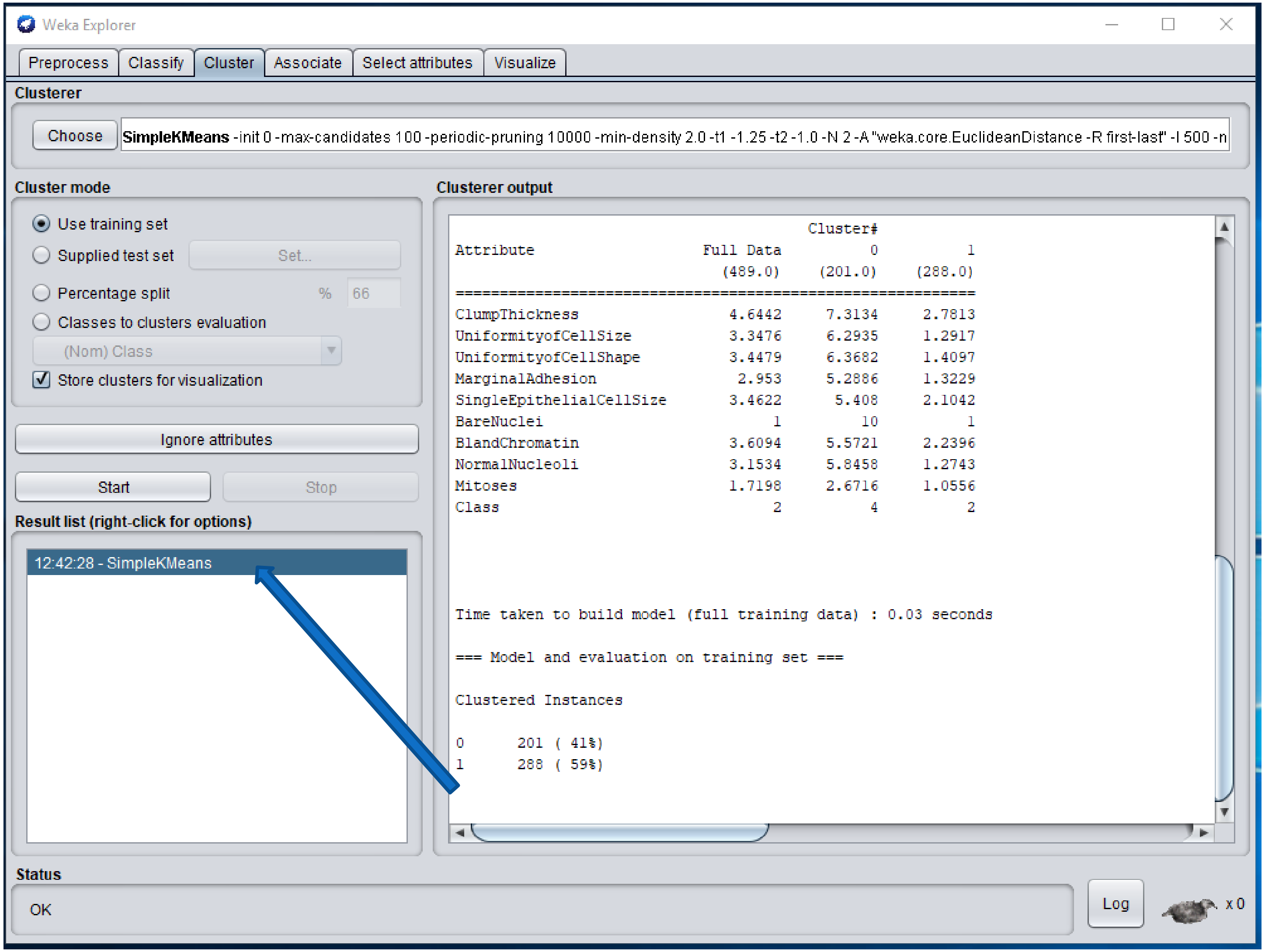
Clusterer output

```
                                        Cluster#
Attribute                   Full Data          0          1
                              (489.0)    (201.0)    (288.0)
============================================================
ClumpThickness                 4.6442     7.3134     2.7813
UniformityofCellSize           3.3476     6.2935     1.2917
UniformityofCellShape          3.4479     6.3682     1.4097
MarginalAdhesion                2.953     5.2886     1.3229
SingleEpithelialCellSize       3.4622      5.408     2.1042
BareNuclei                          1         10          1
BlandChromatin                 3.6094     5.5721     2.2396
NormalNucleoli                 3.1534     5.8458     1.2743
Mitoses                        1.7198     2.6716     1.0556
Class                               2          4          2




Time taken to build model (full training data) : 0.03 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      201 ( 41%)
1      288 ( 59%)
```

Status

OK

Clustering: An example (6)



Clusterer

SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -n

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)

12:42:28 - SimpleKMeans

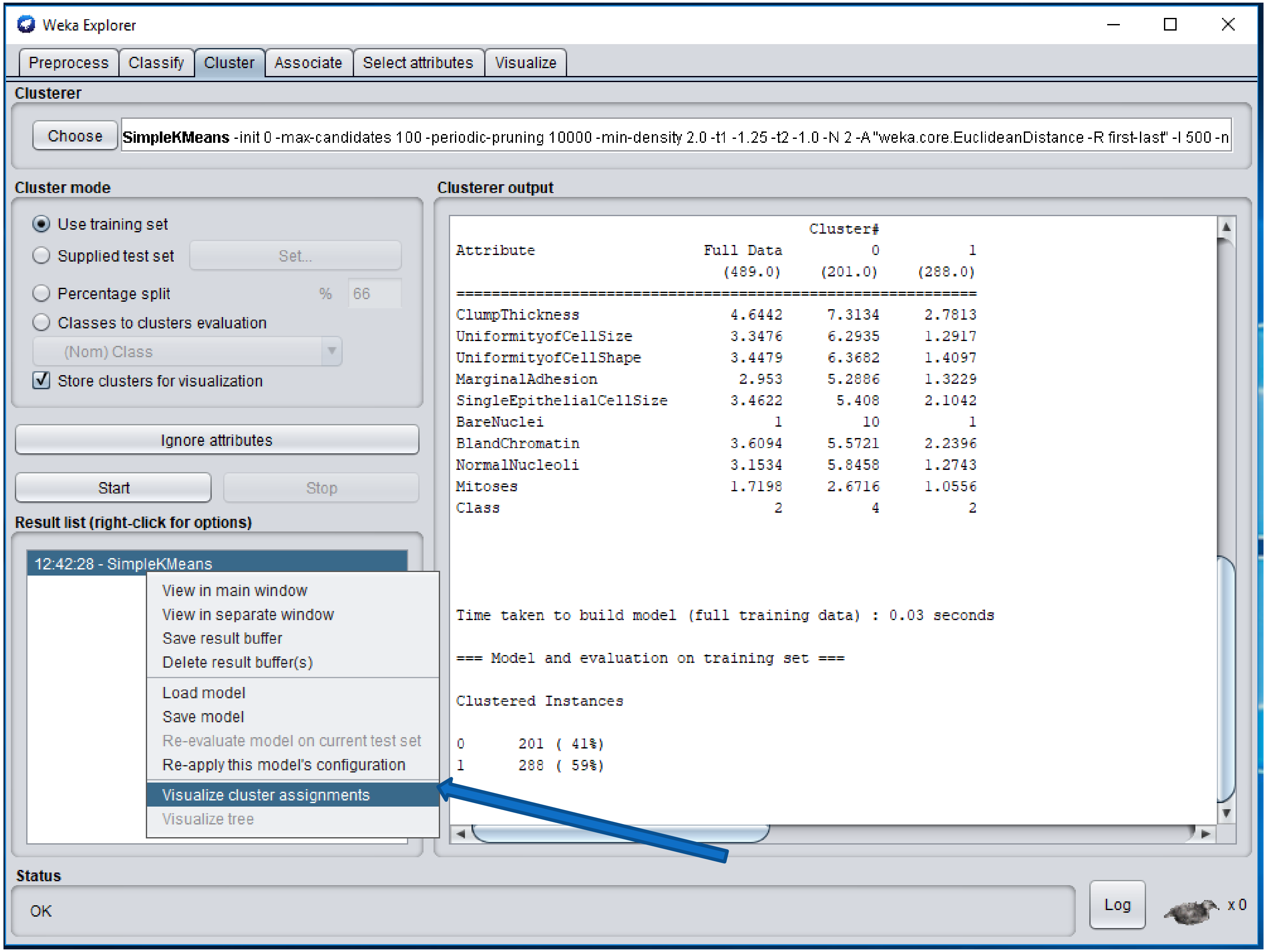
Clusterer output

```
                                            Cluster#
Attribute                    Full Data          0           1
                               (489.0)    (201.0)     (288.0)
============================================================
ClumpThickness                  4.6442     7.3134      2.7813
UniformityofCellSize            3.3476     6.2935      1.2917
UniformityofCellShape           3.4479     6.3682      1.4097
MarginalAdhesion                 2.953     5.2886      1.3229
SingleEpithelialCellSize        3.4622      5.408      2.1042
BareNuclei                           1         10           1
BlandChromatin                  3.6094     5.5721      2.2396
NormalNucleoli                  3.1534     5.8458      1.2743
Mitoses                         1.7198     2.6716      1.0556
Class                                2          4           2




Time taken to build model (full training data) : 0.03 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       201 ( 41%)
1       288 ( 59%)
```

Status

OK

Clustering: An example (6)



SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -n

Cluster mode

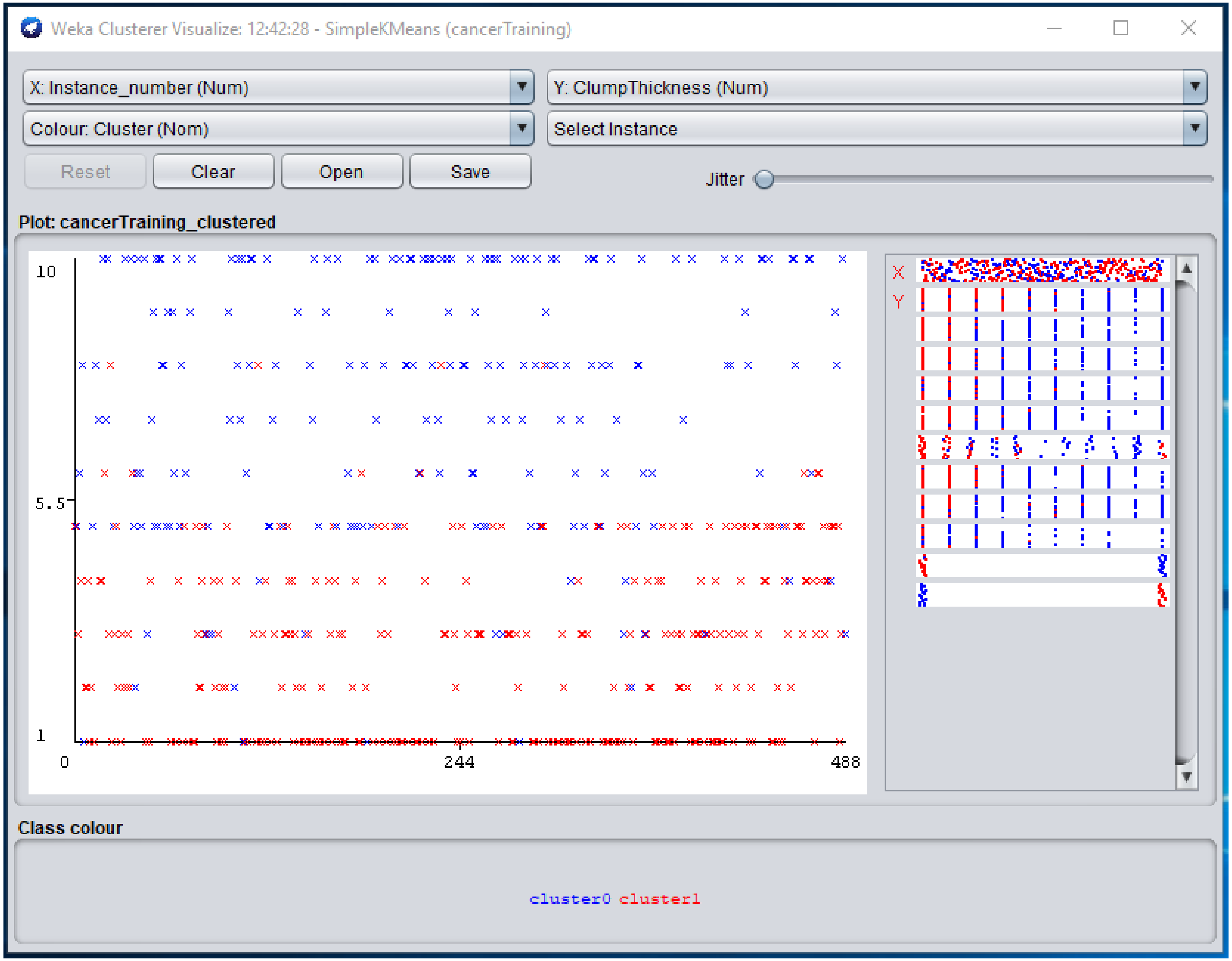
- ● Use training set
- ○ Supplied test set
- ○ Percentage split % 66
- ○ Classes to clusters evaluation
- (Nom) Class
- ☑ Store clusters for visualization

```
                                      Cluster#
Attribute                 Full Data          0          1
                            (489.0)    (201.0)    (288.0)
===========================================================
ClumpThickness               4.6442     7.3134     2.7813
UniformityofCellSize         3.3476     6.2935     1.2917
UniformityofCellShape        3.4479     6.3682     1.4097
MarginalAdhesion              2.953     5.2886     1.3229
SingleEpithelialCellSize     3.4622      5.408     2.1042
BareNuclei                        1         10          1
BlandChromatin               3.6094     5.5721     2.2396
NormalNucleoli               3.1534     5.8458     1.2743
Mitoses                      1.7198     2.6716     1.0556
Class                             2          4          2




Time taken to build model (full training data) : 0.03 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      201 ( 41%)
1      288 ( 59%)
```

Status

OK

Clustering: An example (7)



Plot: cancerTraining_clustered

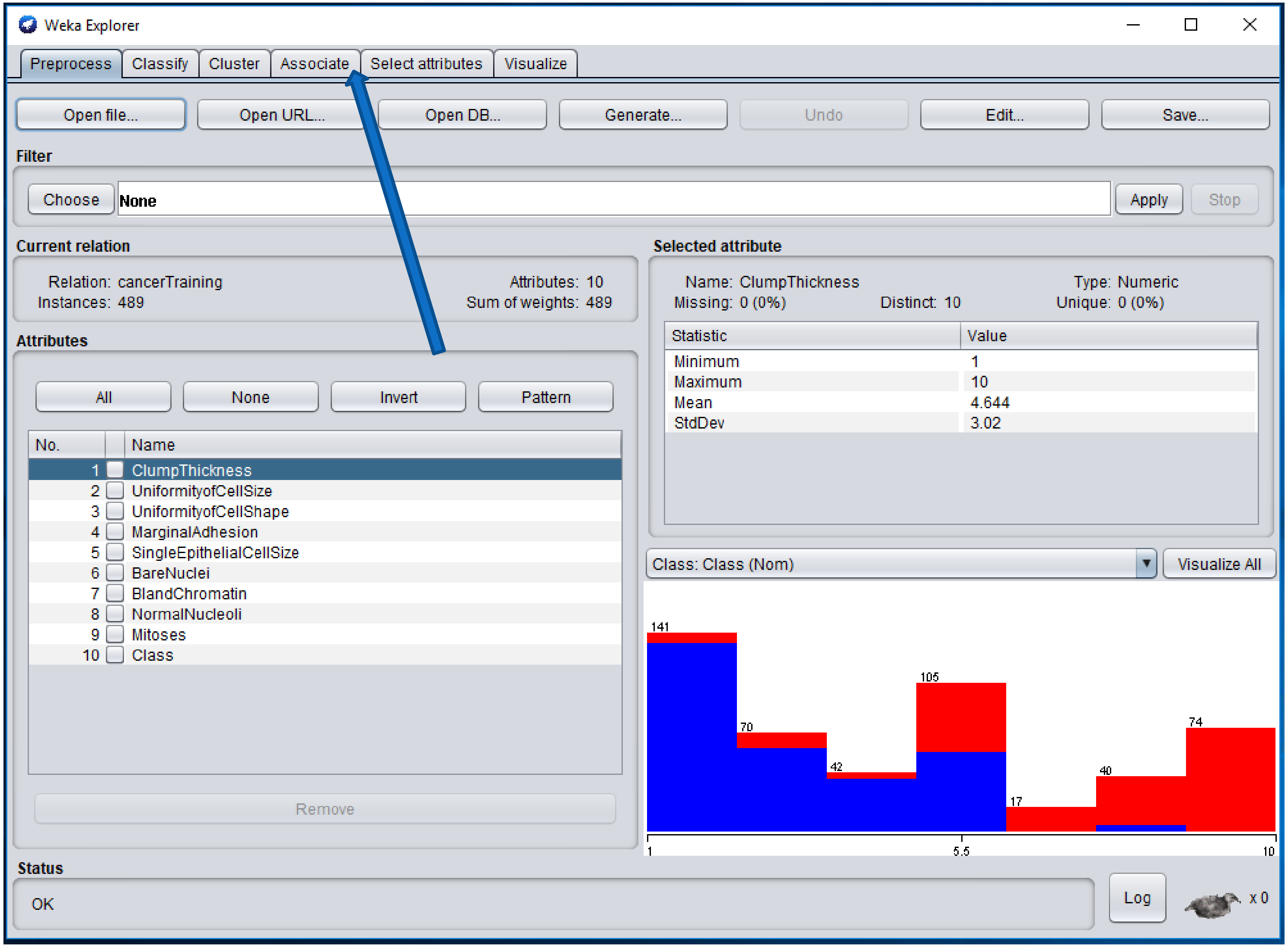
Class colour

cluster0 cluster1

Associations Rules (1)

- 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:
  - Es. milk, butter ⇒ bread, eggs (with confidence 0.9 and support 2000)
- Apriori can compute all rules that have given minimum support and exceed given confidence

Associations Rules: An example (1)



Filter: None

Relation: cancerTraining
Instances: 489
Attributes: 10
Sum of weights: 489

No.		Name
1	□	ClumpThickness
2	□	UniformityofCellSize
3	□	UniformityofCellShape
4	□	MarginalAdhesion
5	□	SingleEpithelialCellSize
6	□	BareNuclei
7	□	BlandChromatin
8	□	NormalNucleoli
9	□	Mitoses
10	□	Class

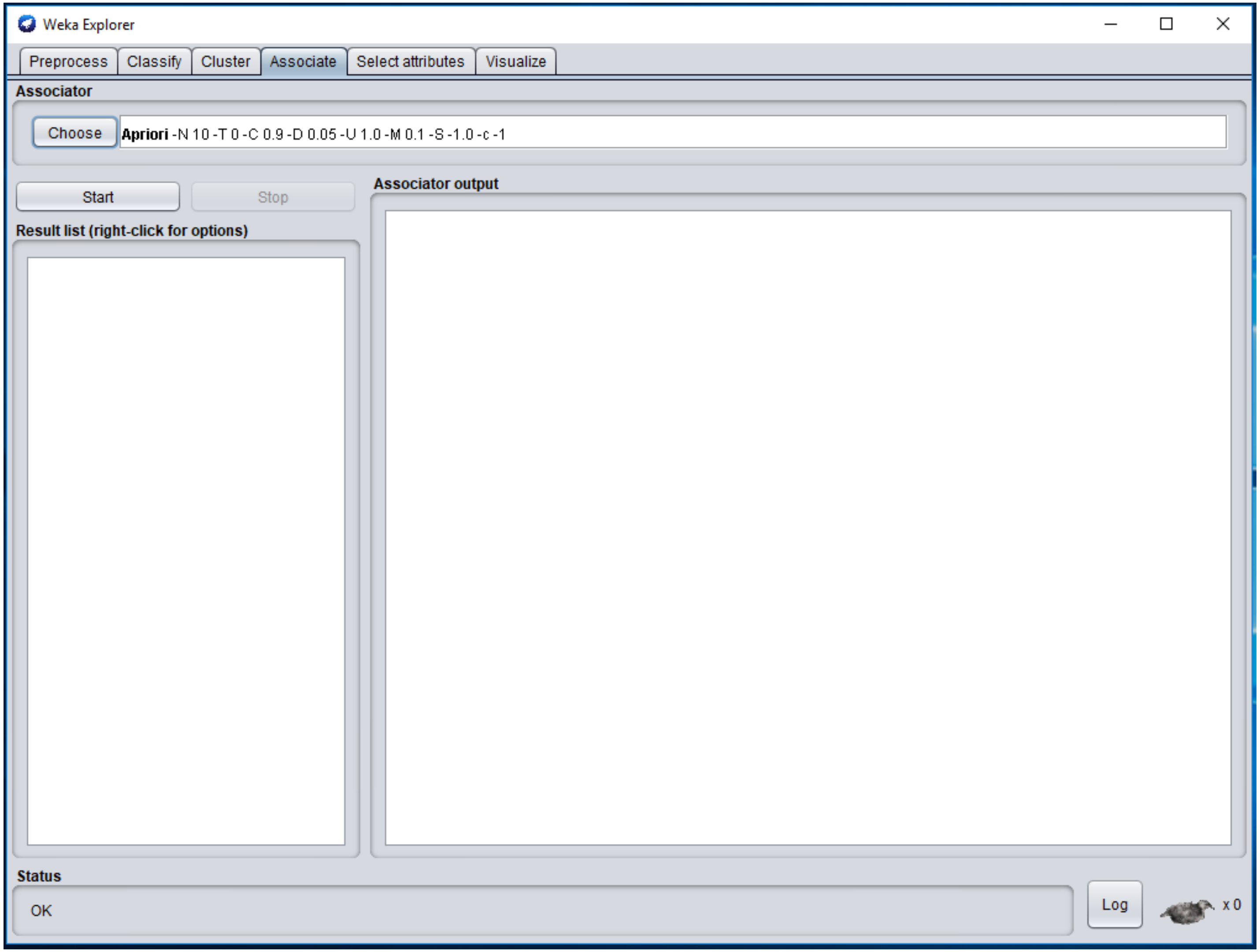
Name: ClumpThickness
Type: Numeric
Missing: 0 (0%)
Distinct: 10
Unique: 0 (0%)

Statistic	Value
Minimum	1
Maximum	10
Mean	4.644
StdDev	3.02

Class: Class (Nom)

OK

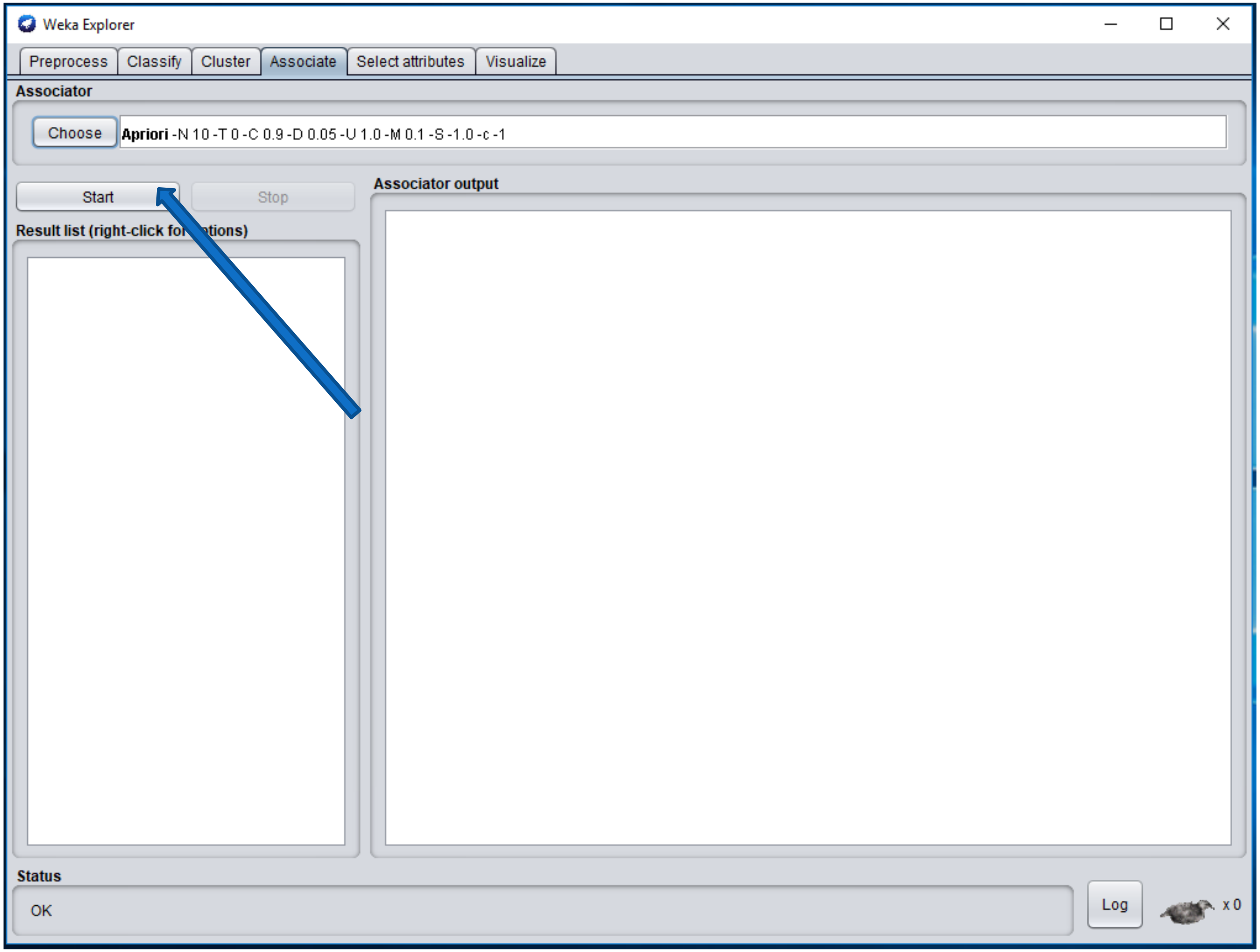
Associations Rules: An example (1)



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

OK

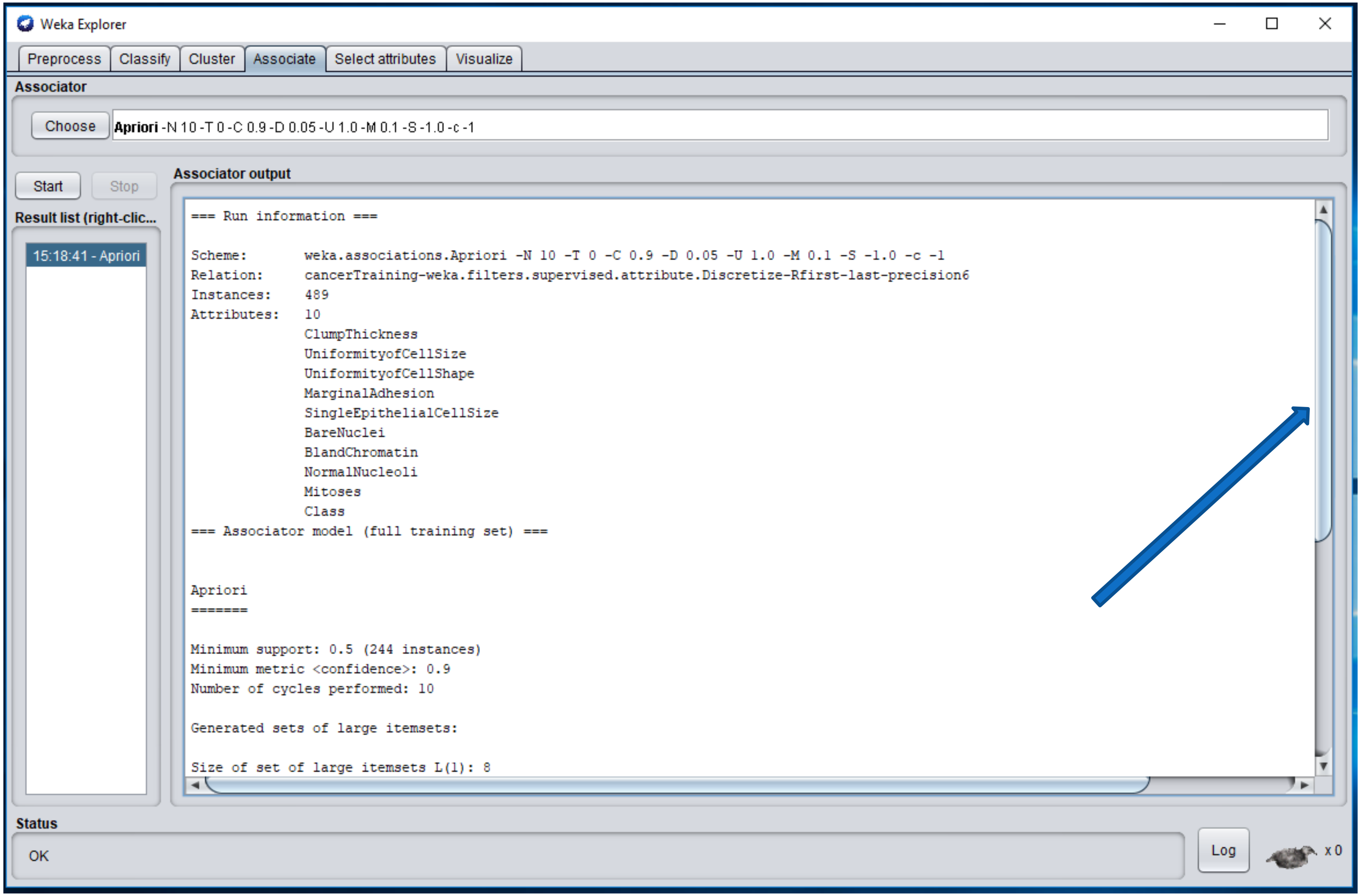
Associations Rules: An example (1)



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

Status: OK

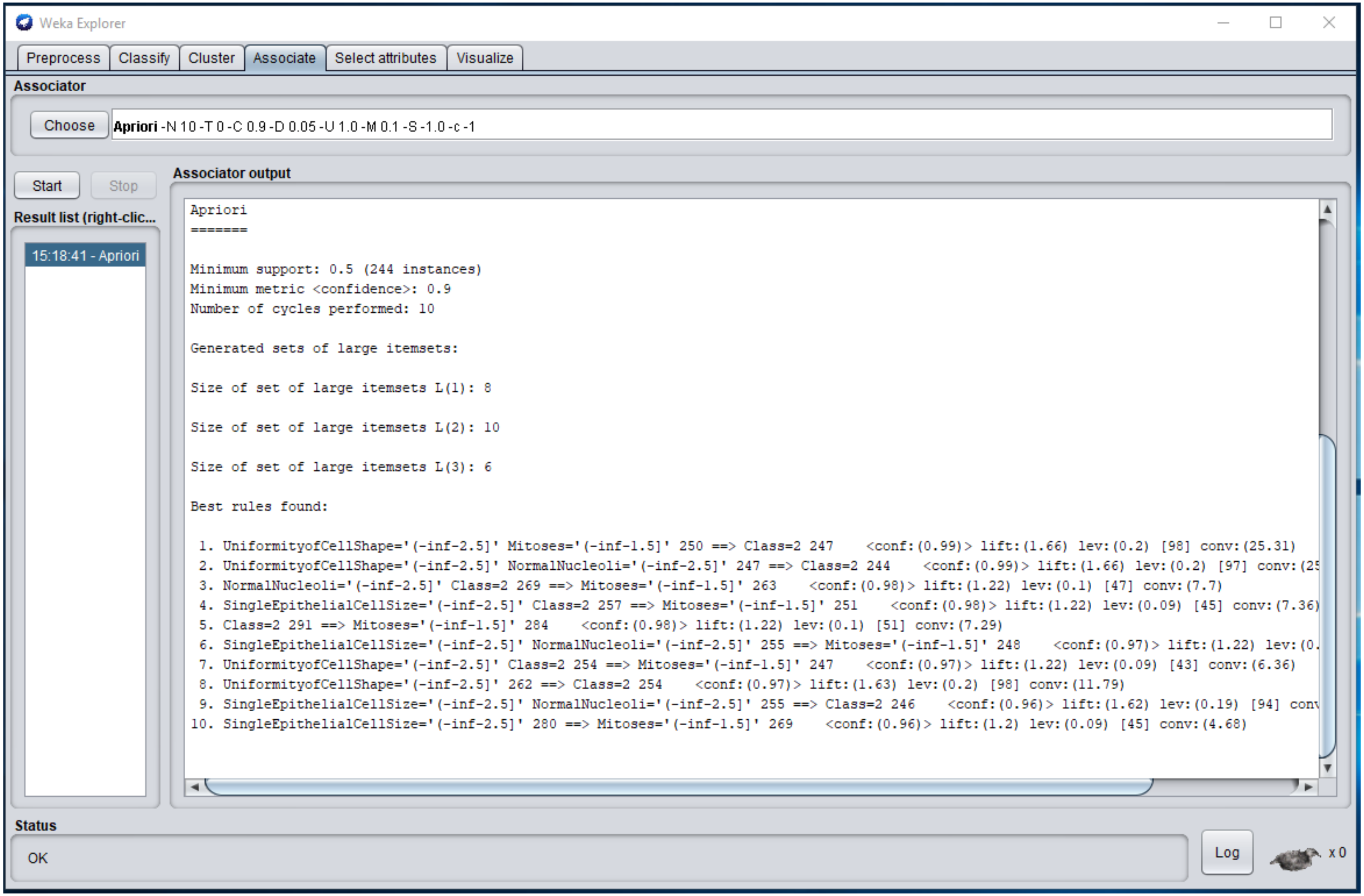
Associations Rules: An example (2)



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

```
=== Run information ===

Scheme:       weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:     cancerTraining-weka.filters.supervised.attribute.Discretize-Rfirst-last-precision6
Instances:    489
Attributes:   10
              ClumpThickness
              UniformityofCellSize
              UniformityofCellShape
              MarginalAdhesion
              SingleEpithelialCellSize
              BareNuclei
              BlandChromatin
              NormalNucleoli
              Mitoses
              Class
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.5 (244 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 10

Generated sets of large itemsets:

Size of set of large itemsets L(1): 8
```

Associations Rules: An example (2)



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

```
Apriori
=======

Minimum support: 0.5 (244 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 10

Generated sets of large itemsets:

Size of set of large itemsets L(1): 8

Size of set of large itemsets L(2): 10

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

Best rules found:

 1. UniformityofCellShape='(-inf-2.5]' Mitoses='(-inf-1.5]' 250 ==> Class=2 247    <conf:(0.99)> lift:(1.66) lev:(0.2) [98] conv:(25.31)
 2. UniformityofCellShape='(-inf-2.5]' NormalNucleoli='(-inf-2.5]' 247 ==> Class=2 244    <conf:(0.99)> lift:(1.66) lev:(0.2) [97] conv:(25
 3. NormalNucleoli='(-inf-2.5]' Class=2 269 ==> Mitoses='(-inf-1.5]' 263    <conf:(0.98)> lift:(1.22) lev:(0.1) [47] conv:(7.7)
 4. SingleEpithelialCellSize='(-inf-2.5]' Class=2 257 ==> Mitoses='(-inf-1.5]' 251    <conf:(0.98)> lift:(1.22) lev:(0.09) [45] conv:(7.36)
 5. Class=2 291 ==> Mitoses='(-inf-1.5]' 284    <conf:(0.98)> lift:(1.22) lev:(0.1) [51] conv:(7.29)
 6. SingleEpithelialCellSize='(-inf-2.5]' NormalNucleoli='(-inf-2.5]' 255 ==> Mitoses='(-inf-1.5]' 248    <conf:(0.97)> lift:(1.22) lev:(0.
 7. UniformityofCellShape='(-inf-2.5]' Class=2 254 ==> Mitoses='(-inf-1.5]' 247    <conf:(0.97)> lift:(1.22) lev:(0.09) [43] conv:(6.36)
 8. UniformityofCellShape='(-inf-2.5]' 262 ==> Class=2 254    <conf:(0.97)> lift:(1.63) lev:(0.2) [98] conv:(11.79)
 9. SingleEpithelialCellSize='(-inf-2.5]' NormalNucleoli='(-inf-2.5]' 255 ==> Class=2 246    <conf:(0.96)> lift:(1.62) lev:(0.19) [94] conv
10. SingleEpithelialCellSize='(-inf-2.5]' 280 ==> Mitoses='(-inf-1.5]' 269    <conf:(0.96)> lift:(1.2) lev:(0.09) [45] conv:(4.68)
```

OK

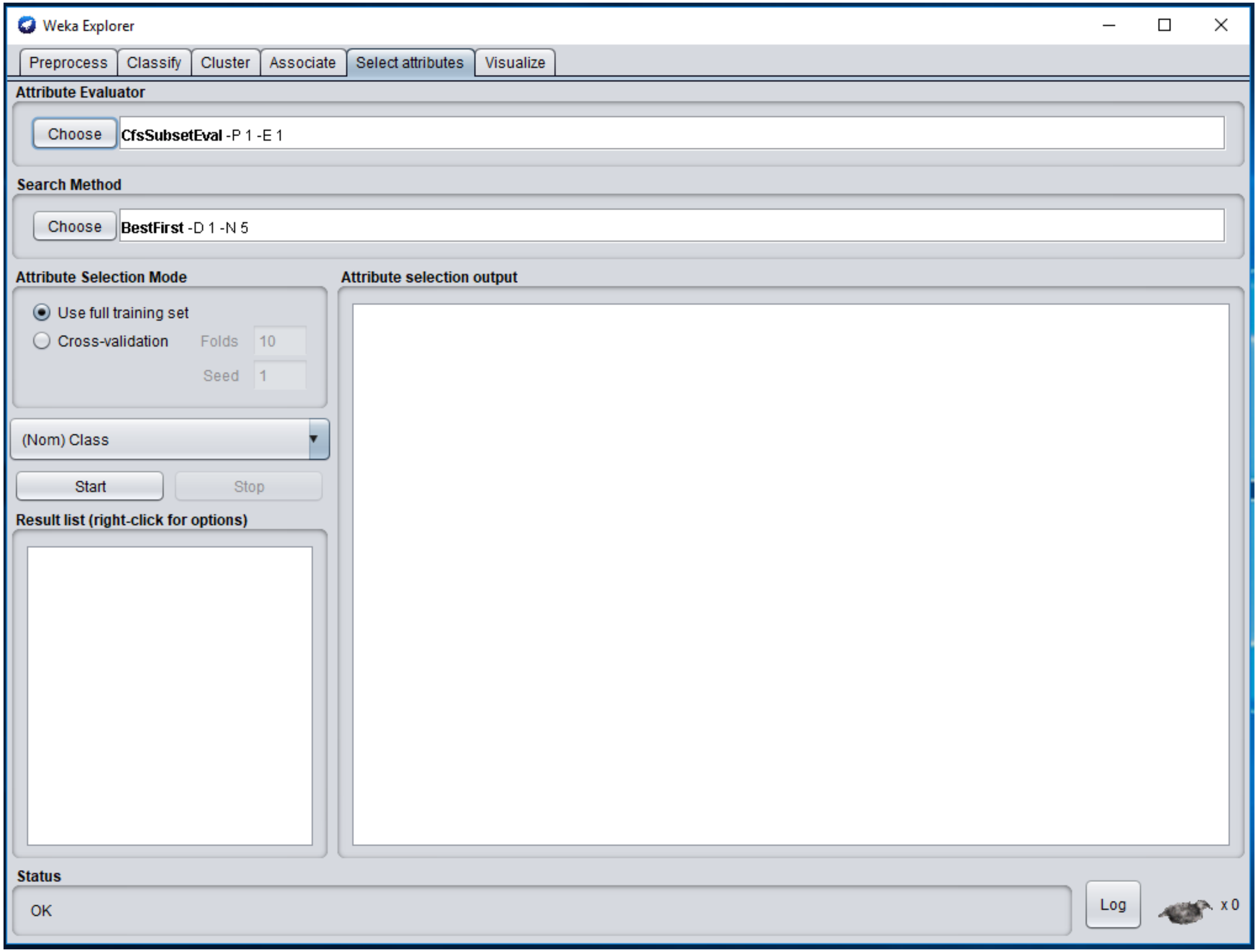
Attribute selection

- 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

Attribute selection: An example (1)



Attribute Evaluator

CfsSubsetEval -P 1 -E 1

Search Method

BestFirst -D 1 -N 5

Attribute Selection Mode

● Use full training set

○ Cross-validation Folds 10

Seed 1

(Nom) Class

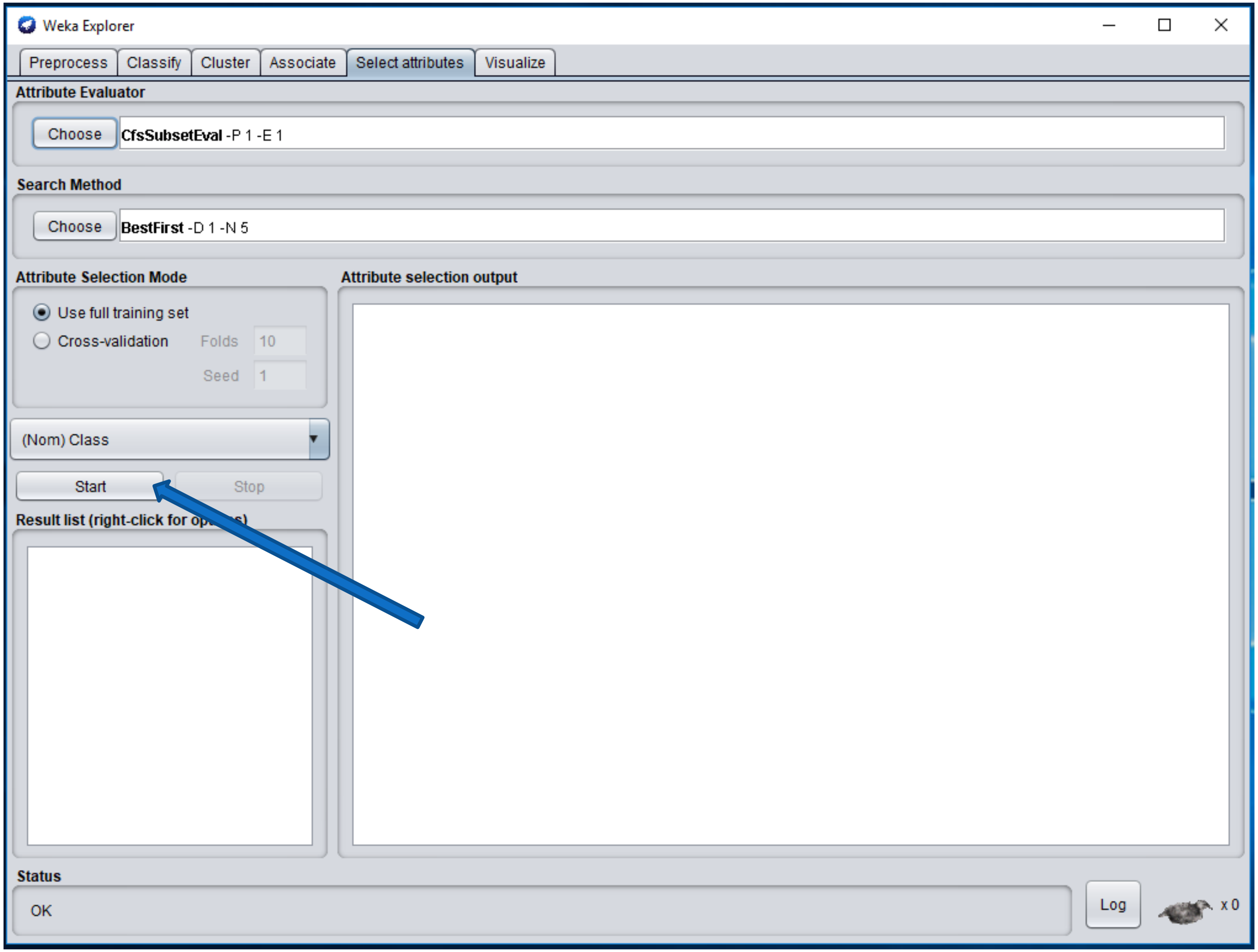
Result list (right-click for options)

Attribute selection output

Status

OK

Attribute selection: An example (1)



Attribute Evaluator

CfsSubsetEval -P 1 -E 1

Search Method

BestFirst -D 1 -N 5

Attribute Selection Mode

● Use full training set

○ Cross-validation Folds 10

Seed 1

(Nom) Class

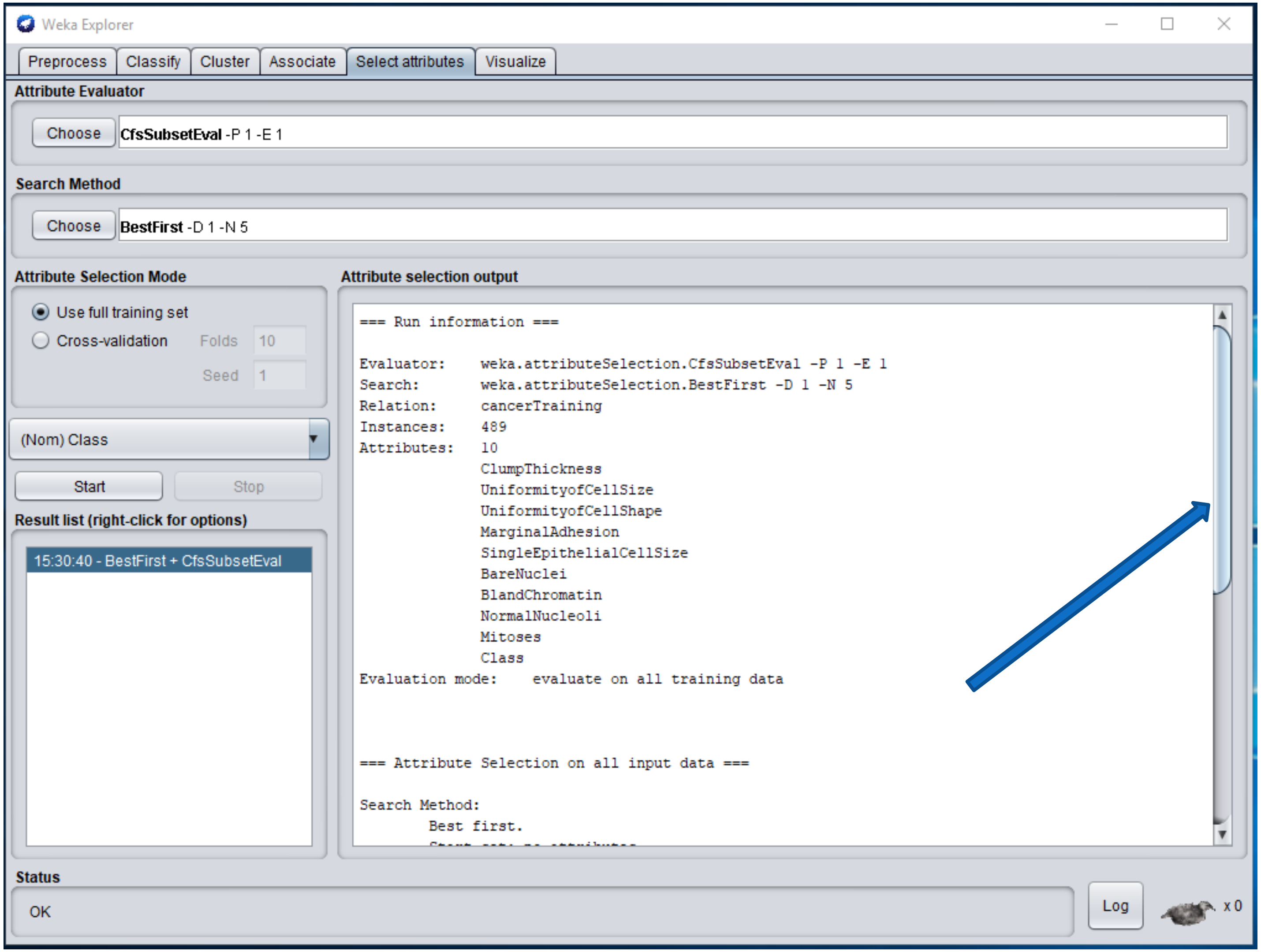
Attribute selection output

Result list (right-click for options)

Status

OK

Attribute selection: An example (2)



CfsSubsetEval -P 1 -E 1

BestFirst -D 1 -N 5

● Use full training set
○ Cross-validation Folds 10
Seed 1

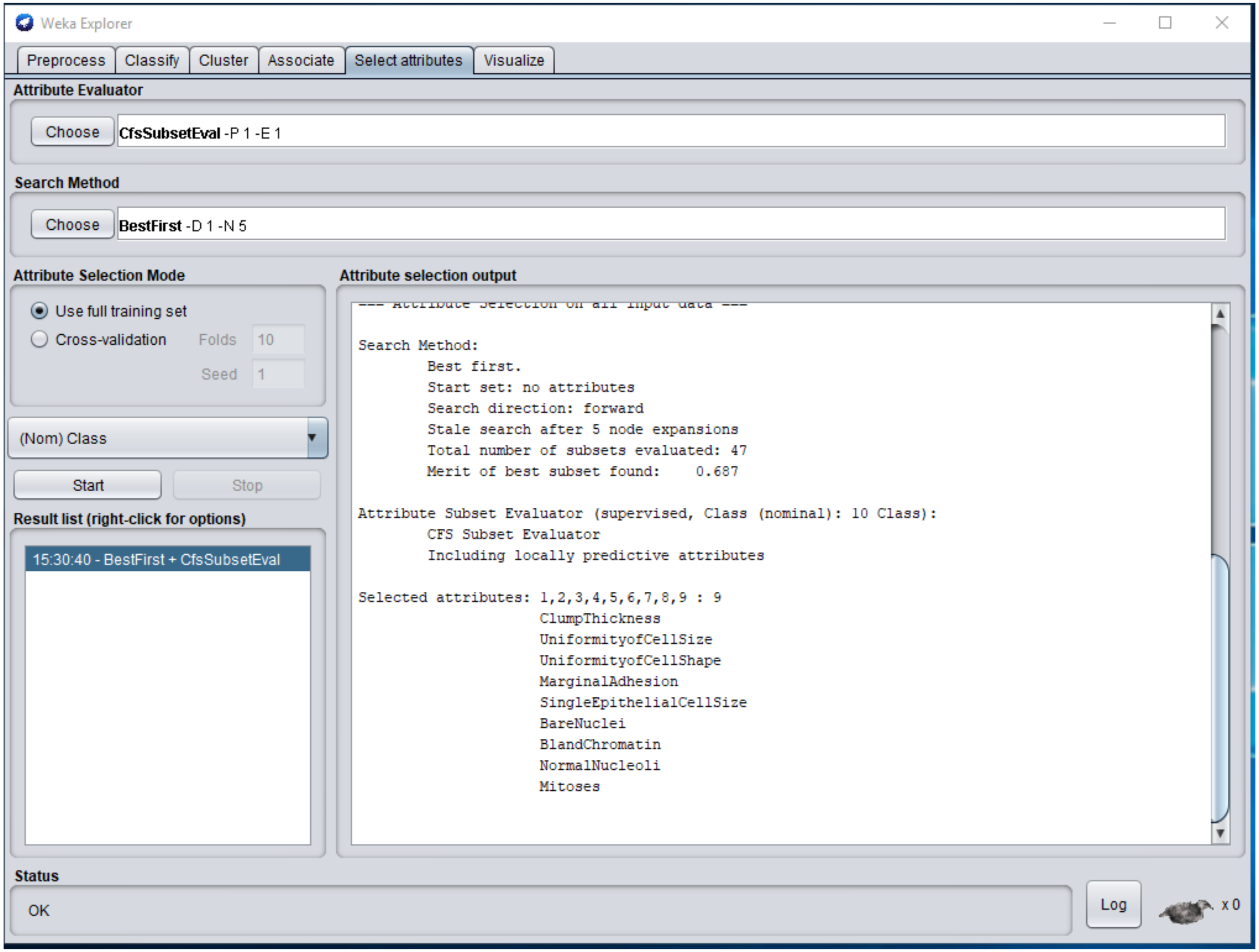
(Nom) Class

15:30:40 - BestFirst + CfsSubsetEval

```
=== Run information ===

Evaluator:    weka.attributeSelection.CfsSubsetEval -P 1 -E 1
Search:       weka.attributeSelection.BestFirst -D 1 -N 5
Relation:     cancerTraining
Instances:    489
Attributes:   10
              ClumpThickness
              UniformityofCellSize
              UniformityofCellShape
              MarginalAdhesion
              SingleEpithelialCellSize
              BareNuclei
              BlandChromatin
              NormalNucleoli
              Mitoses
              Class
Evaluation mode:    evaluate on all training data



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

Search Method:
        Best first.
```

Status: OK

Attribute selection: An example (2)



Attribute Evaluator: CfsSubsetEval -P 1 -E 1

Search Method: BestFirst -D 1 -N 5

Attribute Selection Mode: ● Use full training set ○ Cross-validation Folds 10 Seed 1

(Nom) Class

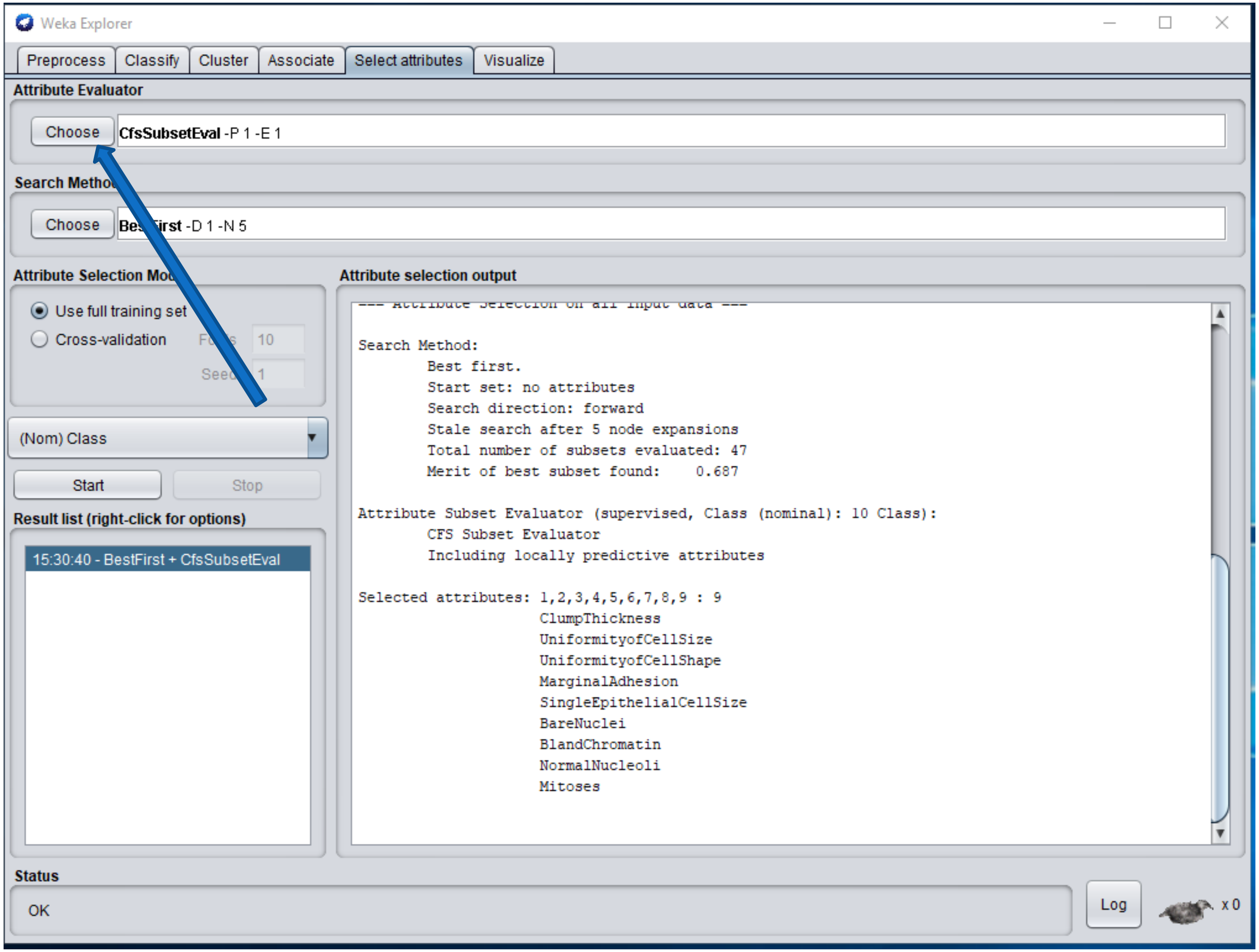
Result list (right-click for options): 15:30:40 - BestFirst + CfsSubsetEval

Attribute selection output:

```
=== 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: 47
        Merit of best subset found:    0.687

Attribute Subset Evaluator (supervised, Class (nominal): 10 Class):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 1,2,3,4,5,6,7,8,9 : 9
                     ClumpThickness
                     UniformityofCellSize
                     UniformityofCellShape
                     MarginalAdhesion
                     SingleEpithelialCellSize
                     BareNuclei
                     BlandChromatin
                     NormalNucleoli
                     Mitoses
```

Status: OK

Attribute selection: An example (2)



Attribute Evaluator

CfsSubsetEval -P 1 -E 1

Search Method

BestFirst -D 1 -N 5

Attribute Selection Mode

● Use full training set

○ Cross-validation Folds 10 Seed 1

(Nom) Class

Result list (right-click for options)

15:30:40 - BestFirst + CfsSubsetEval

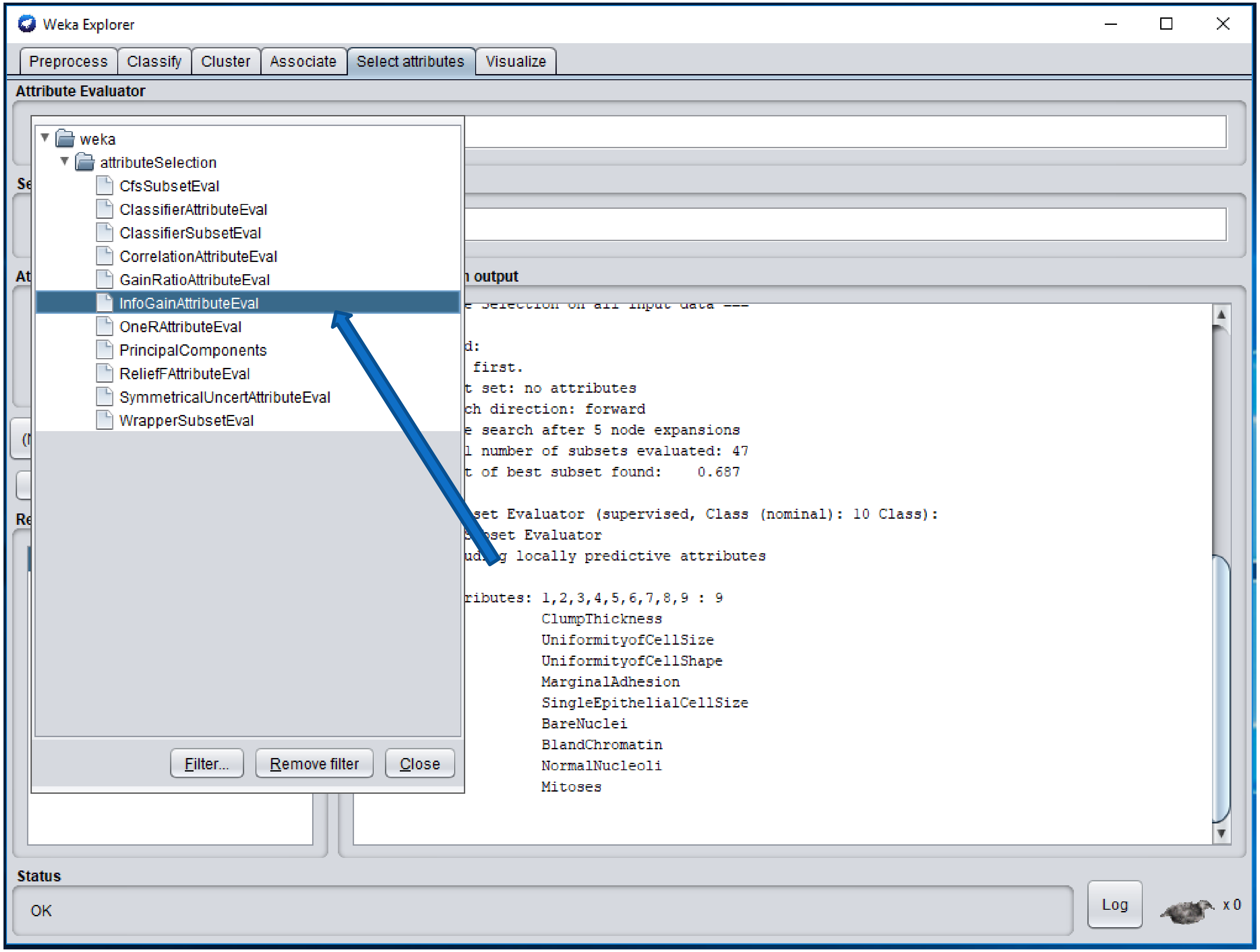
Attribute selection output

```
=== 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: 47
        Merit of best subset found:    0.687

Attribute Subset Evaluator (supervised, Class (nominal): 10 Class):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 1,2,3,4,5,6,7,8,9 : 9
                     ClumpThickness
                     UniformityofCellSize
                     UniformityofCellShape
                     MarginalAdhesion
                     SingleEpithelialCellSize
                     BareNuclei
                     BlandChromatin
                     NormalNucleoli
                     Mitoses
```

Status

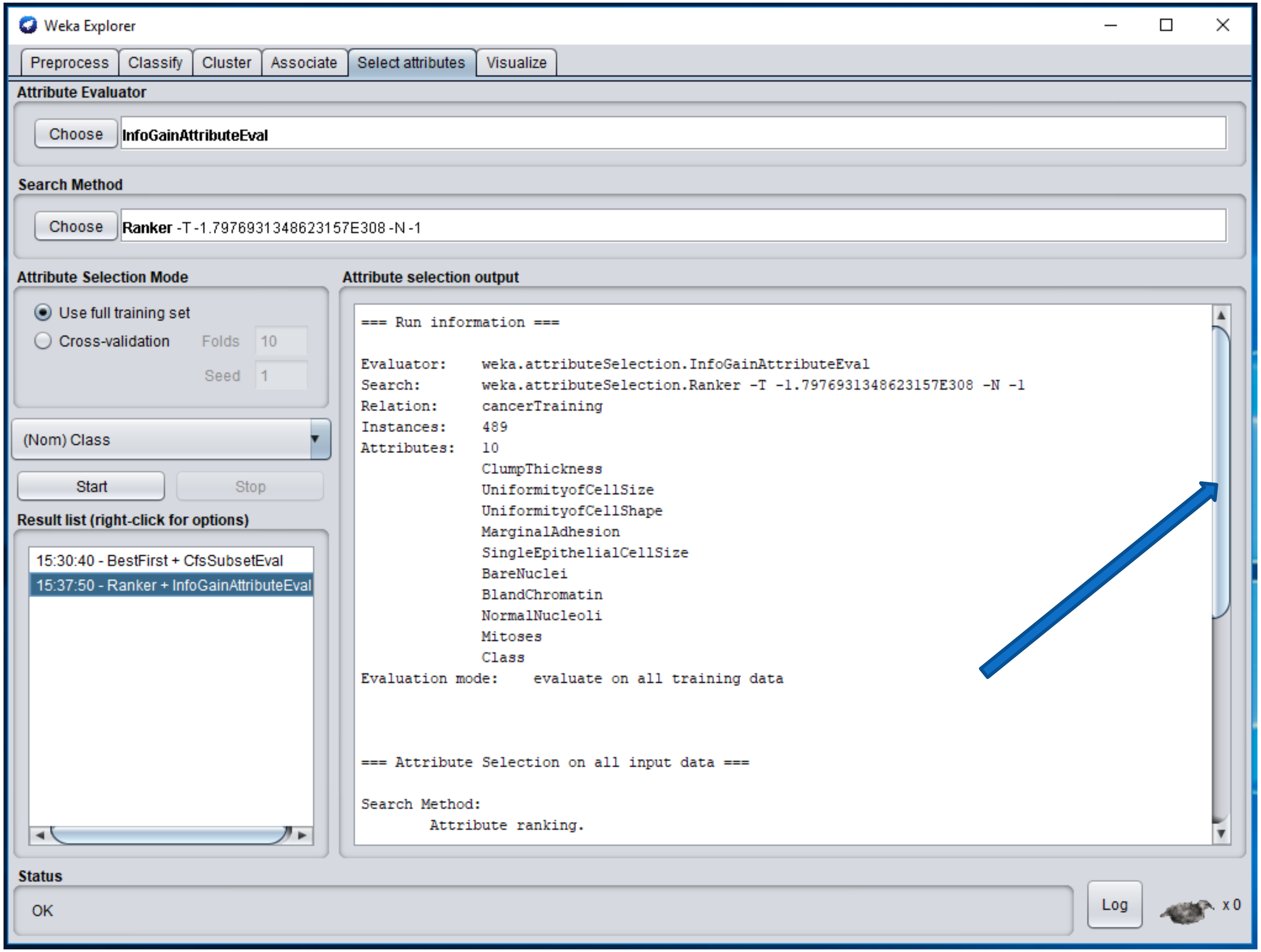
OK

Attribute selection: An example (3)



```
e Selection on all input data ===

d:
 first.
t set: no attributes
ch direction: forward
e search after 5 node expansions
l number of subsets evaluated: 47
t of best subset found:     0.687

set Evaluator (supervised, Class (nominal): 10 Class):
bset Evaluator
uding locally predictive attributes

ributes: 1,2,3,4,5,6,7,8,9 : 9
         ClumpThickness
         UniformityofCellSize
         UniformityofCellShape
         MarginalAdhesion
         SingleEpithelialCellSize
         BareNuclei
         BlandChromatin
         NormalNucleoli
         Mitoses
```

Attribute selection: An example (4)



Attribute Evaluator: InfoGainAttributeEval

Search Method: Ranker -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)

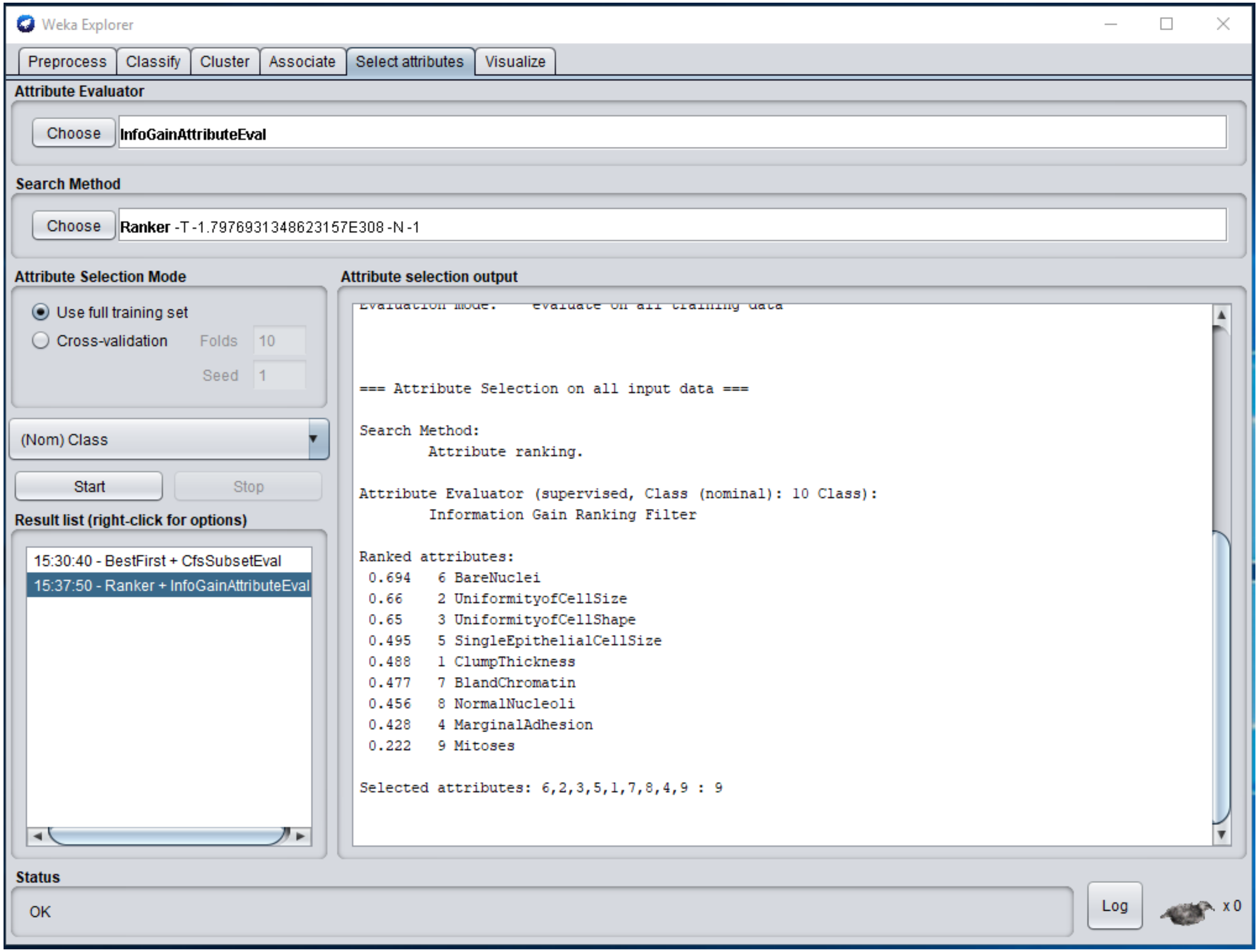
15:30:40 - BestFirst + CfsSubsetEval
15:37:50 - Ranker + InfoGainAttributeEval

Attribute selection output

```
=== Run information ===

Evaluator:    weka.attributeSelection.InfoGainAttributeEval
Search:       weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1
Relation:     cancerTraining
Instances:    489
Attributes:   10
              ClumpThickness
              UniformityofCellSize
              UniformityofCellShape
              MarginalAdhesion
              SingleEpithelialCellSize
              BareNuclei
              BlandChromatin
              NormalNucleoli
              Mitoses
              Class
Evaluation mode:    evaluate on all training data



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

Search Method:
        Attribute ranking.
```

Status: OK

Attribute selection: An example (4)



Attribute Evaluator: InfoGainAttributeEval

Search Method: Ranker -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)

- 15:30:40 - BestFirst + CfsSubsetEval
- 15:37:50 - Ranker + InfoGainAttributeEval

Attribute selection output

```
Evaluation mode:    evaluate on all training data



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

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 10 Class):
        Information Gain Ranking Filter

Ranked attributes:
 0.694   6 BareNuclei
 0.66    2 UniformityofCellSize
 0.65    3 UniformityofCellShape
 0.495   5 SingleEpithelialCellSize
 0.488   1 ClumpThickness
 0.477   7 BlandChromatin
 0.456   8 NormalNucleoli
 0.428   4 MarginalAdhesion
 0.222   9 Mitoses

Selected attributes: 6,2,3,5,1,7,8,4,9 : 9
```

Status: OK

Data Visualization (1)

- Visualization very useful in practice: e.g. helps to determine the difficulty of the learning problem

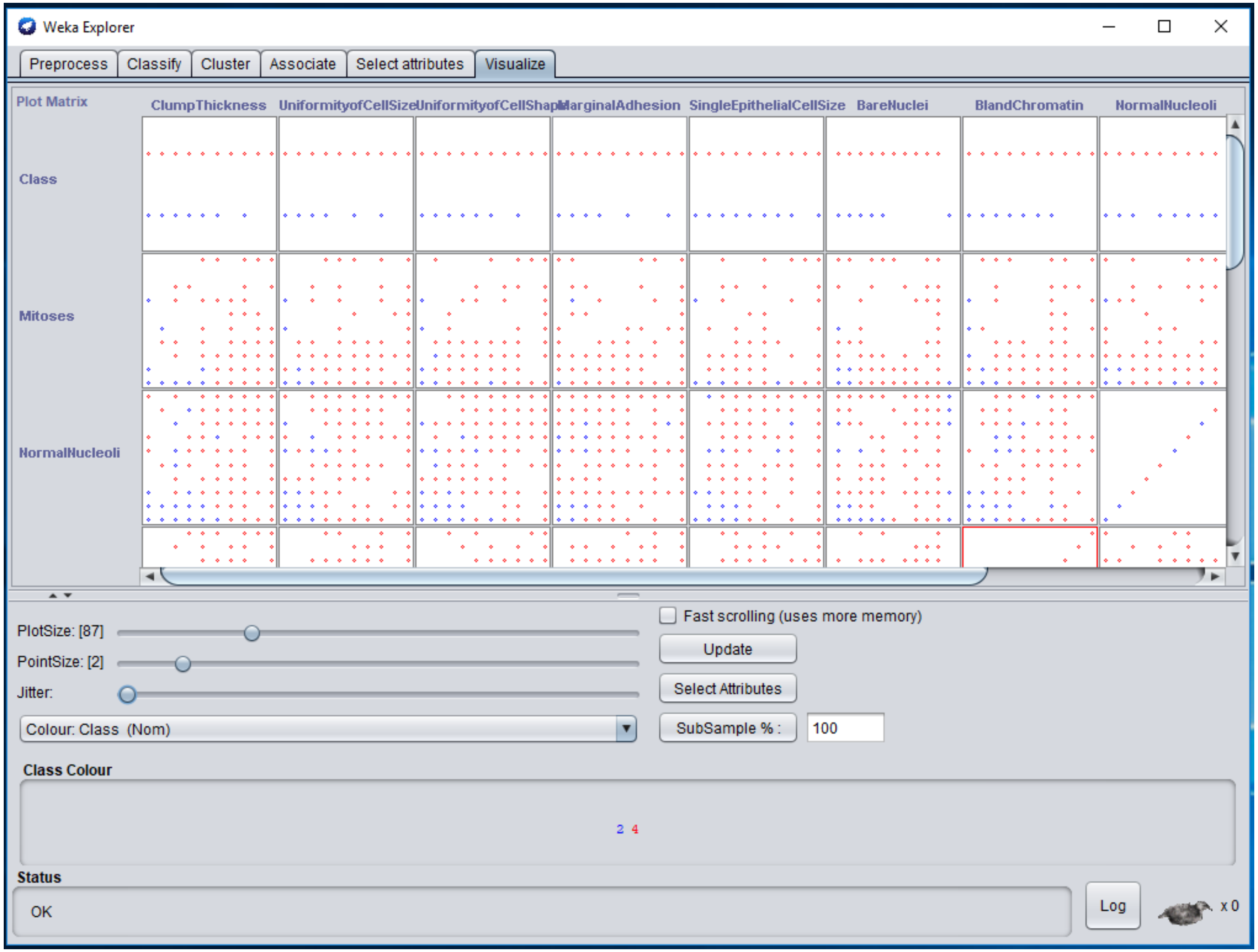
- WEKA can visualize single attributes (1-d) and pairs of attributes (2-d)
  - To do: rotating 3-d visualizations (Xgobi-style)

- Colour-coded class values

- “Jitter” option to deal with nominal attributes (and to detect “hidden” data points)

- “Zoom-in” function

Data Visualization (2)



☐ Fast scrolling (uses more memory)

PlotSize: [87]

PointSize: [2]

Jitter:

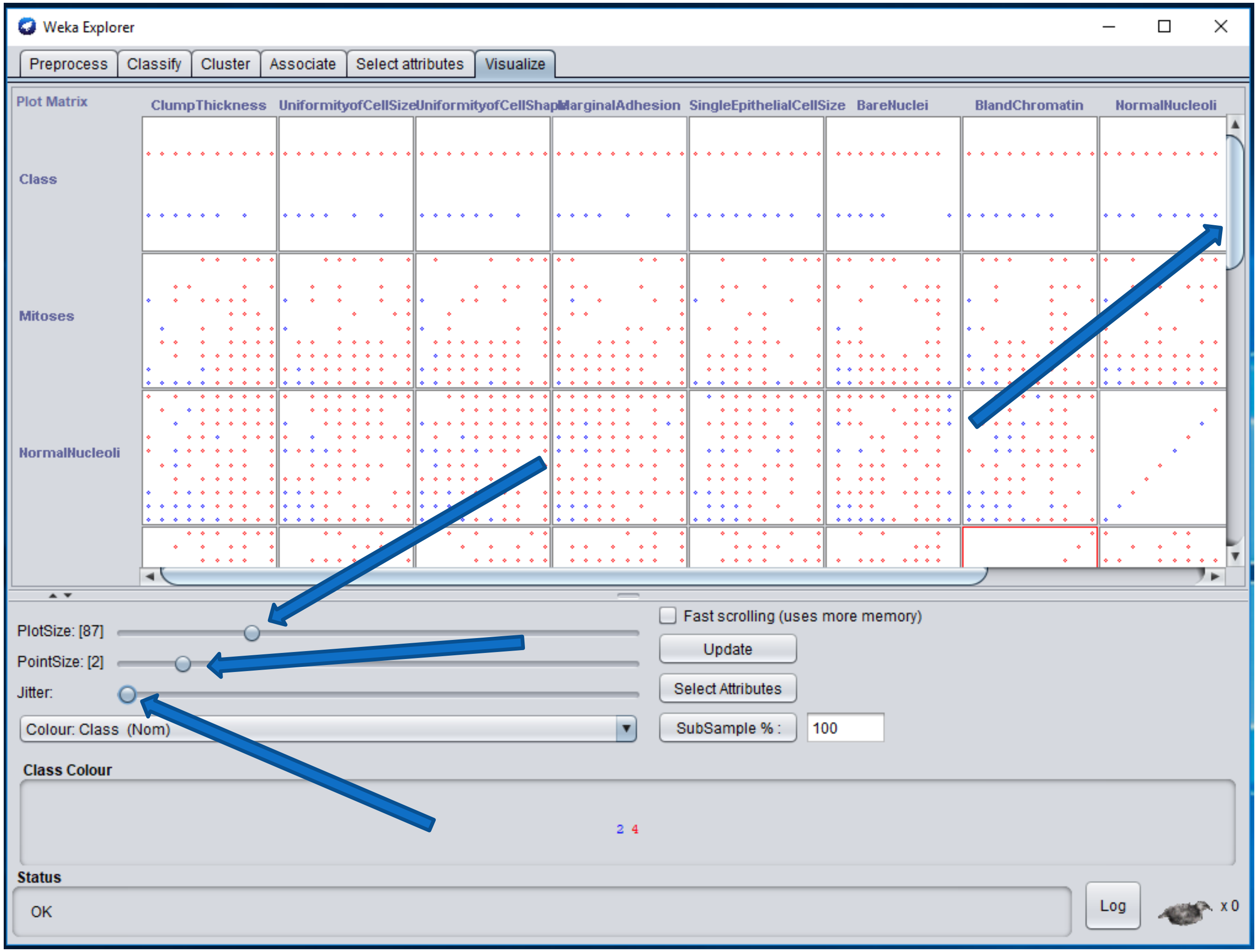
Colour: Class (Nom)

SubSample % : 100

Status

OK

Data Visualization (2)



PlotSize: [87]

PointSize: [2]

Jitter:

Colour: Class (Nom)

☐ Fast scrolling (uses more memory)

SubSample % : 100

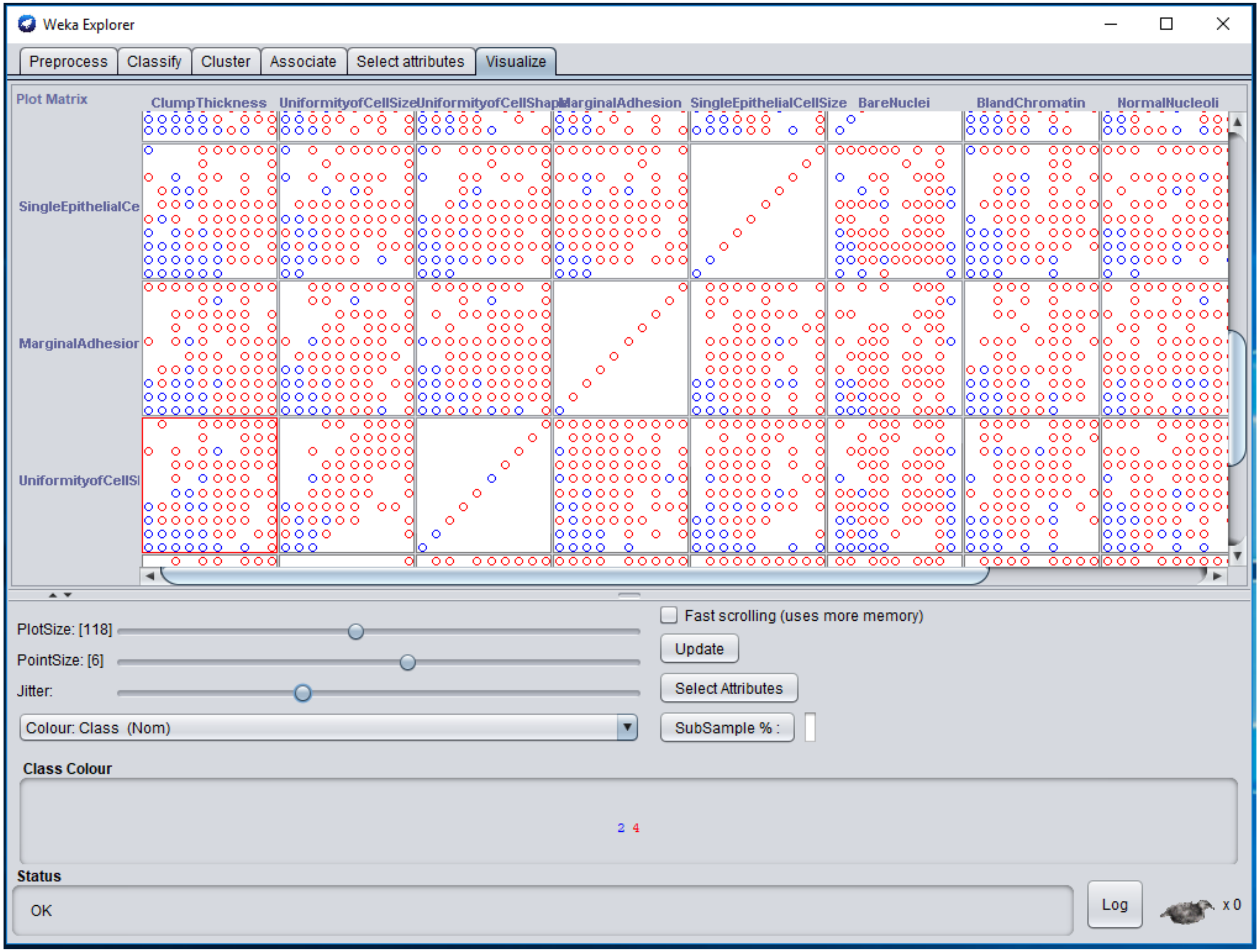
Class Colour

2 4

Status

OK

Data Visualization (3)



PlotSize: [118]

PointSize: [6]

Jitter:

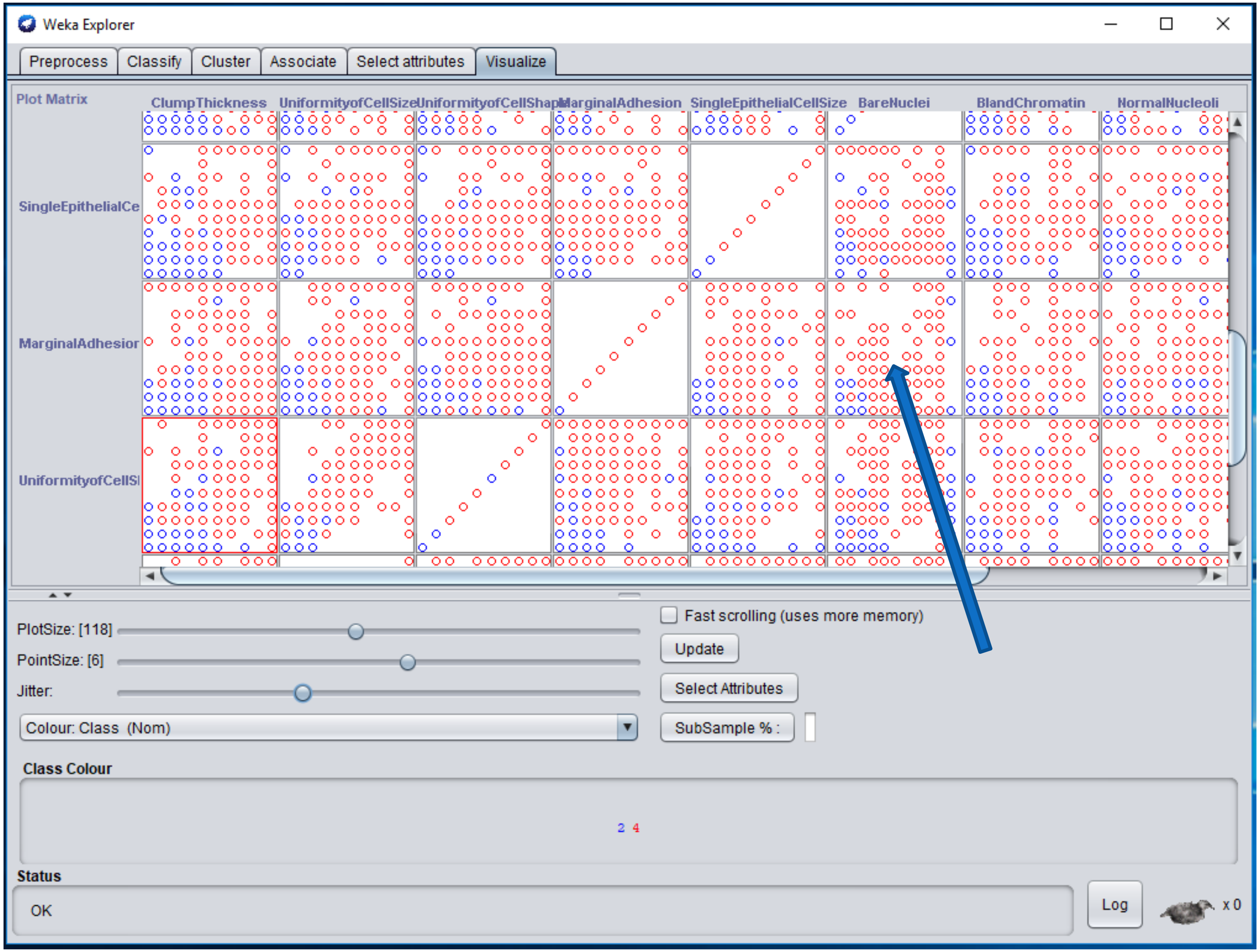
Colour: Class (Nom)

□ Fast scrolling (uses more memory)

Status

OK

Data Visualization (3)



Weka Explorer

□ Fast scrolling (uses more memory)

PlotSize: [118]

PointSize: [6]

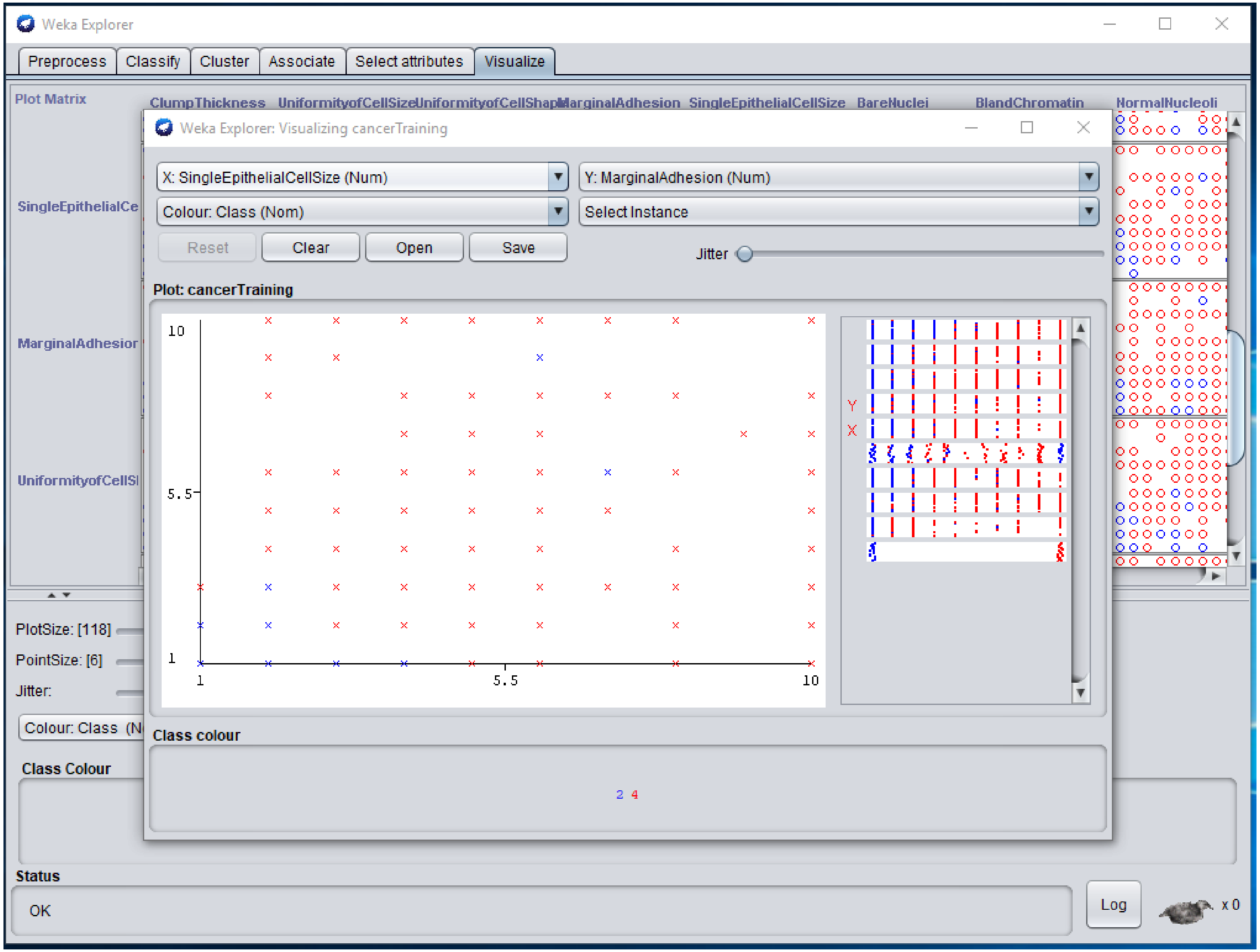
Jitter:

Colour: Class (Nom)

Status

OK

Data Visualization (4)



Weka Explorer

Weka Explorer: Visualizing cancerTraining

X: SingleEpithelialCellSize (Num)

Y: MarginalAdhesion (Num)

Colour: Class (Nom)

Select Instance

Plot: cancerTraining

Class colour

2 4

PlotSize: [118]

PointSize: [6]

Jitter:

Colour: Class (N

Class Colour

Status

OK

References and Resources

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