



Data Analysis & Data Mining with WEKA

Lan Man

Department of Computer Science and Technology

East China Normal University

©2017 All rights reserved.

Synopsis

- About WEKA
- Data Format and Preprocessing
- WEKA Usage (version 3.8)
 - Explorer
 - Experimenter
 - Knowledge Flow
- Other Data Mining Tools
- Summary

About WEKA

- Developed at the machine learning group of CS Department, University of Waikato, New Zealand from 1993
 - Waikato Environment for Knowledge Analysis
 - Written in Java. Open Source, issued under the GNU General Public License.
 - Provides multiple implementations of learning algorithms
 - Classification
 - Clustering
 - Association Rule Mining
 - Attribute Selection (Feature Selection)
- which can be directly applied to the dataset

About WEKA (contd..)

- Main features
 - Comprehensive set of data pre-processing tools, learning algorithms and evaluation methods
 - Graphical user interfaces (incl. data visualization)
 - Environment for comparing learning algorithms
- Usage
 - Apply a learning method to a dataset & analyze the result
 - Use a learned model to make predictions on new instances
 - Apply different learners to a dataset & compare results
- WEKA downloaded from <http://www.cs.waikato.ac.nz/~ml/weka/>
- This lesson is based on the latest snapshot of WEKA 3.8

Launching Weka 3.8



Machine Learning with WEKA 3.8

- About WEKA
- Data Format and Preprocessing
- WEKA Usage (version 3.8)
 - Explorer
 - Experimenter
 - Knowledge Flow
- Other Data Mining Tools
- Summary

WEKA 数据格式

- 跟很多电子表格或数据分析软件一样，WEKA所处理的数据集是如图的一个二维的表格

Viewer

Relation weather.symbolic

| No. | 1: outlook | 2: temperature | 3: humidity | 4: windy | 5: play |
|-----|------------|----------------|-------------|----------|---------|
| | Nominal | Nominal | Nominal | Nominal | Nominal |
| 1 | sunny | hot | high | FALSE | no |
| 2 | sunny | hot | high | TRUE | no |
| 3 | overcast | hot | high | FALSE | yes |
| 4 | rainy | mild | high | FALSE | yes |
| 5 | rainy | cool | normal | FALSE | yes |
| 6 | rainy | cool | normal | TRUE | no |
| 7 | overcast | cool | normal | TRUE | yes |
| 8 | sunny | mild | high | FALSE | no |
| 9 | sunny | cool | normal | FALSE | yes |
| 10 | rainy | mild | normal | FALSE | yes |
| 11 | sunny | mild | normal | TRUE | yes |
| 12 | overcast | mild | high | TRUE | yes |
| 13 | overcast | hot | normal | FALSE | yes |
| 14 | rainy | mild | high | TRUE | no |

- WEKA中的术语：表格里的一行称作一个实例（Instance）；
- 一列称作一个属性（Attribute），相当于统计学中的一个变量，或者数据库中的一个字段。
- 这样一个表格，或者叫数据集(dataset)，呈现属性之间的一种关系(Relation)。
- 14个实例，5个属性，关系名称为“weather.symbolic”

WEKA 数据格式

```
% ARFF file for the weather data
@relation weather.symbolic

@attribute outlook {sunny, overcast, rainy}
@attribute temperature {hot, mild, cool}
@attribute humidity {high, normal}
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
% 14 instances

sunny,hot,high, FALSE,no
sunny,hot,high, TRUE,no
overcast,hot,high, FALSE, yes
rainy,mild,high, FALSE, yes
rainy,cool,normal, FALSE, yes
rainy,cool,normal, TRUE, no
overcast,cool,normal, TRUE, yes
sunny,mild,high, FALSE, no
sunny,cool,normal, FALSE, yes
rainy,mild,normal, FALSE, yes
sunny,mild,normal, TRUE, yes
overcast,mild,high, TRUE, yes
overcast,hot,normal, FALSE, yes
rainy,mild,high, TRUE, no
```

- WEKA存储数据的格式是ARFF (**Attribute-Relation File Format**)文件，这是一种ASCII文本文件。上页图中所示的二维表格存储在如左图示的ARFF文件中。这也就是WEKA自带的“weather.arff”文件，在WEKA安装目录的“data”子目录下可以找到。
 - 识别ARFF文件的重要依据是分行，因此不能在这种文件里随意的断行。空行（或全是空格的行）将被忽略。
 - 以“%”开始的行是注释，WEKA将忽略这些行。

WEKA 数据格式

```
% ARFF file for the weather data
@relation weather symbolic

@attribute outlook {sunny, overcast, rainy}
@attribute temperature {hot, mild, cool}
@attribute humidity {high, normal}
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
% 14 instances

sunny,hot,high,FALSE,no
sunny,hot,high,TRUE,no
overcast,hot,high,FALSE,yes
rainy,mild,high,FALSE,yes
rainy,cool,normal,TRUE,no
rainy,cool,normal,FALSE,yes
overcast,cool,normal,TRUE,yes
sunny,mild,high,FALSE,no
sunny,cool,normal,FALSE,yes
rainy,mild,normal,TRUE,yes
sunny,mild,normal,TRUE,yes
overcast,mild,high,TRUE,yes
overcast,hot,normal,FALSE,yes
rainy,mild,high,TRUE,no
```

- 头信息（Head information），包括了对关系的声明和对属性的声明
关系名称在ARFF文件的第一个有效行来定义；如果这个字符串包含空格，它必须加上引号（指英文标点的单引号或双引号）
- 属性声明用一列以“@attribute”开头的语句表示。数据集中的每一个属性都有它对应的“@attribute”语句，来定义它的属性名称和数据类型。
- 数据信息（Data information），即数据集中给出的数据。从“@data”标记开始，后面的就是数据信息了。

WEKA 数据格式

```
% ARFF file for the weather data
@relation weather.symbolic

@attribute outlook {sunny, overcast, rainy}
@attribute temperature {hot, mild, cool}
@attribute humidity {high, normal}
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
% 14 instances

sunny,hot,high,FALSE,no
sunny,hot,high,TRUE,no
overcast,hot,high,FALSE,yes
rainy,mild,high,FALSE,yes
rainy,cool,normal,TRUE,yes
rainy,cool,normal,FALSE,no
overcast,cool,normal,TRUE,no
sunny,mild,high,FALSE,no
sunny,cool,normal,FALSE,yes
rainy,mild,normal,FALSE,yes
sunny,mild,normal,TRUE,yes
overcast,mild,high,TRUE,yes
overcast,hot,normal,FALSE,yes
rainy,mild,high,TRUE,no
```

- 属性声明语句的顺序很重要。首先表明该项属性在数据部分的位置。其次，最后一个声明的属性被称作class属性，在分类或回归任务中，它是默认的目标变量。
- <attribute-name>是必须以字母开头的字符串。和关系名称一样，如果这个字符串包含空格，它必须加上引号。
- WEKA支持的<datatype>有四种：
numeric: 数值型
nominal-specification: 分类(nominal)
string: 字符串型
date [<date-format>]: 日期和时间型
默认的字符串是ISO-8601所给的日期时间组合格式“yyyy-MM-ddTHH:mm:ss”

WEKA 数据格式

```
% ARFF file for the weather data
@relation weather.symbolic

@attribute outlook {sunny, overcast, rainy}
@attribute temperature {hot, mild, cool}
@attribute humidity {high, normal}
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
% 14 instances

sunny,hot,high,FALSE,no
sunny,hot,high,TRUE,no
overcast,hot,high,FALSE,yes
rainy,mild,high,FALSE,yes
rainy,cool,normal,FALSE,yes
rainy,cool,normal,TRUE,no
overcast,cool,normal,TRUE,yes
sunny,mild,high,FALSE,no
sunny,cool,normal,FALSE,yes
rainy,mild,normal,FALSE,yes
sunny,mild,normal,TRUE,yes
overcast,mild,high,TRUE,yes
overcast,hot,normal,FALSE,yes
rainy,mild,high,TRUE,no
```

- 数据信息中 “@data” 标记独占一行，剩下的是各个实例的数据。每个实例占一行。实例的各属性值用逗号 “,” 隔开。如果某个属性的值是缺失值 (missing value) ，用问号 “?” 表示，不能省略。例如：
sunny,hot,high,? ,no
?,hot,high,?,yes
- 字符串属性和分类属性的值是区分大小写。若含有空格，必须被引号括起来。如：
AG5, 'Encyclopedias and dictionaries'
'Twentieth century'
- 日期属性的值必须与属性声明中给定的一致。例如：
@ATTRIBUTE timestamp DATE "yyyy-MM-
dd HH:mm:ss"
@DATA
"2001-04-03 12:12:12"

WEKA 数据格式-稀疏数据

- 有的数据集中含有大量的0值（比如超市购物篮数据），这时用稀疏格式的数据存贮更加省空间
- 稀疏格式是针对数据信息中某个实例的表示而言，不需要修改ARFF文件的其它部分。如下：

@data

0, X, 0, Y, 0, 0, "class A"
0, 0, W, 0, 0, 0, "class B"

用稀疏格式表达:

@data

{1 X, 3 Y, 6 "class A"}
{2 W, 6 "class B"}



每个实例用**花括号**括起来。实例中每一个非0的属性值用<index> <空格> <value>表示。<index>是属性的序号，**从0开始计**；<value>是属性值。属性值之间仍用逗号隔开。这里每个实例的数值**必须按属性的顺序来写**，如 {1 X, 3 Y, 6 "class A"}, 不能写成{3 Y, 1 X, 6 "class A"}

- 注意在稀疏格式中没有注明的属性值不是缺失值，而是0值。若要表示缺失值必须显式的用问号（?）表示出来。

数据预处理

- 使用WEKA作数据挖掘，面临的第一个问题往往是我们数据不是ARFF格式的。WEKA提供了对CSV文件的支持，而这种格式是被很多其他软件所支持的。此外，WEKA还提供了通过JDBC访问数据库的功能。
- 根据ARFF文件格式，我们也可以先用自己的预处理方法将原有的文件转化成ARFF文件，毕竟后者才是WEKA支持得最好的文件格式。
- 面对一个ARFF文件，我们仍有一些预处理要做，才能进行挖掘任务。

Machine Learning with WEKA 3.8

- About WEKA
- Data Format and Preprocessing
- WEKA Usage (version 3.8)
 - Explorer
 - Experimenter
 - Knowledge Flow
- Other Data Mining Tools
- Summary

The image shows the WEKA Experiment Environment interface. On the left, the SimpleCLI window displays a welcome message and command completion information. A red arrow points from the text "Enter commands in thetextfield at the bottom of the window. Use the up and down arrows to move through previous commands." to the command line area at the bottom of the window. Another red arrow points from the text "help <command>" to the same command line area. In the center, the Applications window lists several options: Explorer, Experimenter, KnowledgeFlow, Workbench, and Simple CLI. A red arrow points from the "Simple CLI" button in the Applications window to the "Simple CLI" tab in the bottom navigation bar of the main window. On the right, a large blue box highlights the Weka Explorer window, which contains tabs for Preprocess, Classify, Cluster, Associate, Select attributes, and Visualize. The "Preprocess" tab is selected. A red arrow points from the "Simple CLI" tab in the bottom navigation bar to the "Simple CLI" tab in the Weka Explorer window.

WEKA窗口

- Applications
 - Explorer
 - Experimenter
 - Knowledge Flow
 - Command line Interface (SimpleCLI)

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open UR... Open DB... Generate... Undo

Choose None

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

Status Welcome to the Weka Explorer

Simple CLI

Untitled1

Processes Attribute summary Scatter plot matrix SQL Viewer Simple CLI

Weka

Data Analysis & Data Mining

15

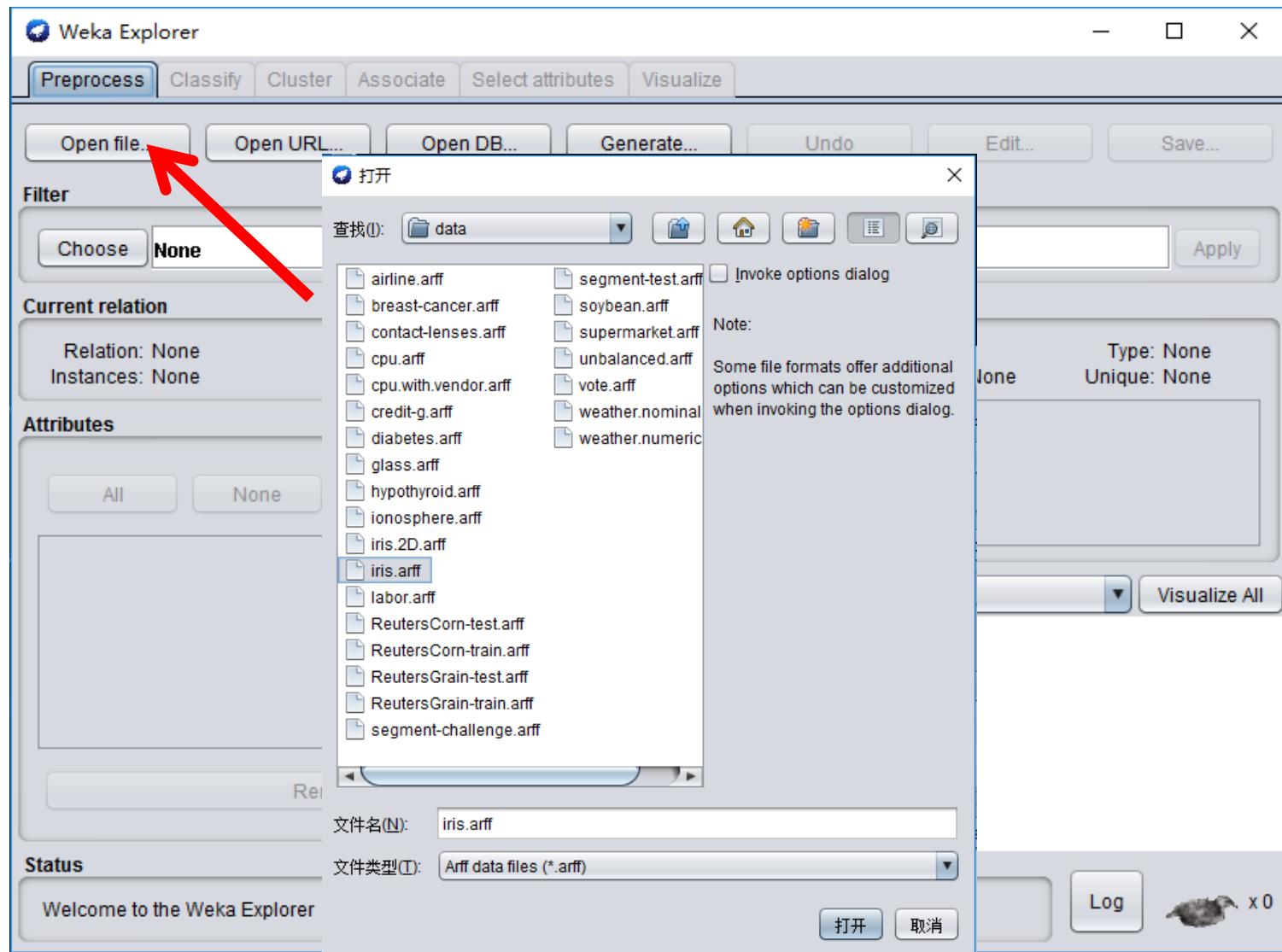
Roadmap: WEKA Usage

- WEKA Usage (version 3.8)
 - Explorer
 - Data preprocessing
 - Classification
 - Clustering
 - Association rule
 - Attribute selection (Feature selection)
 - Data visualization
 - Experimenter
 - Knowledge Flow

Explorer: Data Pre-processing

- 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, ...

Explorer: Data Pre-processing



Explorer: Data Pre-processing

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

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

Filter

Choose None Apply

Current relation

Relation: iris Instances: 150 Attributes: 5 Sum of weights: 150

Selected attribute

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

| Statistic | Value |
|-----------|-------|
| Minimum | 4.3 |
| Maximum | 7.9 |
| Mean | 5.843 |
| StdDev | 0.828 |

Attributes

All None Invert Pattern

| No. | Name |
|-----|---|
| 1 | <input checked="" type="checkbox"/> sepallength |
| 2 | <input type="checkbox"/> sepalwidth |
| 3 | <input type="checkbox"/> petallength |
| 4 | <input type="checkbox"/> petalwidth |
| 5 | <input type="checkbox"/> class |

Remove

Status

OK Log X 0

The image shows the Weka Data Explorer window with the 'Preprocess' tab selected. The 'Current relation' section displays 'Relation: iris', 'Instances: 150', 'Attributes: 5', and 'Sum of weights: 150'. The 'Selected attribute' section shows 'Name: sepallength', 'Missing: 0 (0%)', 'Distinct: 35', 'Type: Numeric', and 'Unique: 9 (6%)'. Below these are descriptive statistics for 'sepallength'. The 'Attributes' section lists all five attributes: sepallength, sepalwidth, petallength, petalwidth, and class. The 'sepallength' attribute is highlighted with a red circle and a red arrow points to its entry in the list. A histogram at the bottom visualizes the distribution of sepallength values, with three specific bins highlighted: one blue bin at 4.3 with value 16, one red bin at 5.1 with value 30, and one cyan bin at 7.9 with value 7. The status bar at the bottom indicates 'OK'.

Explorer: Data Pre-processing

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

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

Filter

Choose None Apply

Current relation

Relation: iris Attributes: 5
Instances: 150 Sum of weights: 150

Attributes

All None Invert Pattern

| No. | Name |
|-----|---|
| 1 | <input type="checkbox"/> sepallength |
| 2 | <input type="checkbox"/> sepalwidth |
| 3 | <input type="checkbox"/> petallength |
| 4 | <input type="checkbox"/> petalwidth |
| 5 | <input checked="" type="checkbox"/> class |

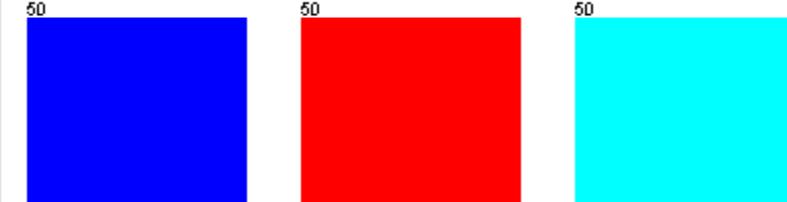
Remove

Selected attribute

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

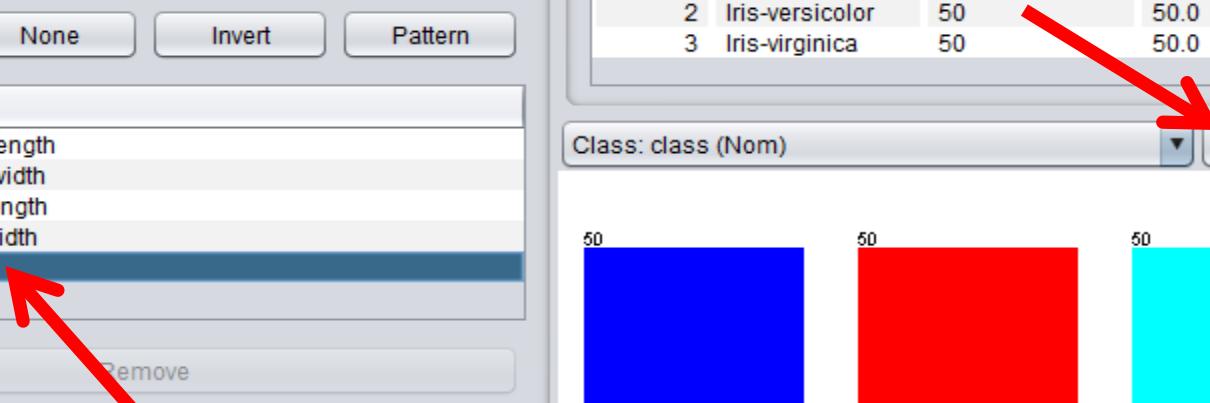
| No. | Label | Count | Weight |
|-----|-----------------|-------|--------|
| 1 | Iris-setosa | 50 | 50.0 |
| 2 | Iris-versicolor | 50 | 50.0 |
| 3 | Iris-virginica | 50 | 50.0 |

Class: class (Nom) ▾ Visualize All

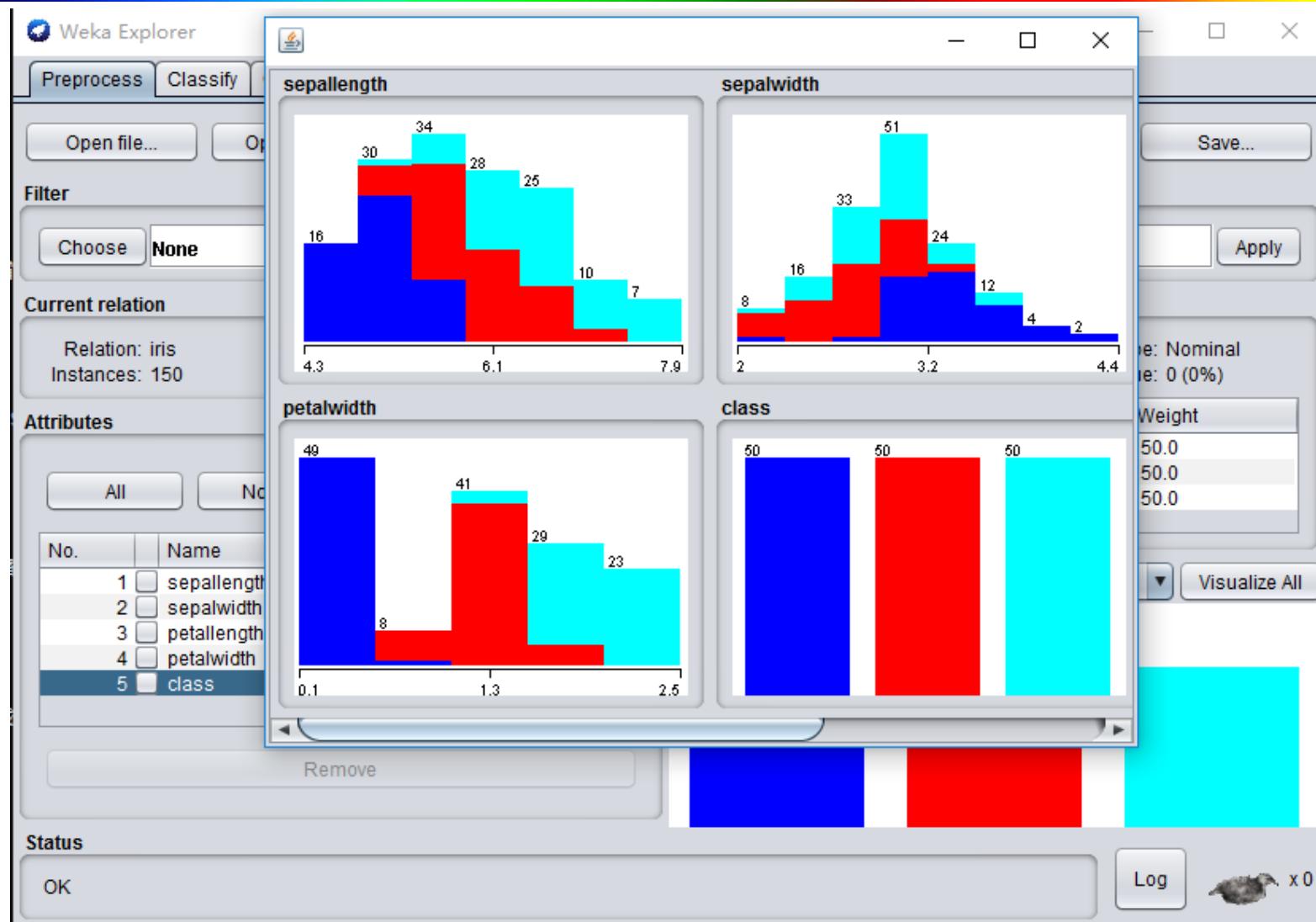


Status

OK Log x 0



Explorer: Data Pre-processing



Explorer: Data Pre-processing

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

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

Filter

Choose **None** Apply

Current relation

Relation: iris Instances: 150 Attributes: 5 Sum of weights: 150

Attributes

All None Invert Pattern

| No. | Name |
|-----|---|
| 1 | <input type="checkbox"/> sepalwidth |
| 2 | <input type="checkbox"/> petallength |
| 3 | <input type="checkbox"/> petalwidth |
| 4 | <input type="checkbox"/> class |
| 5 | <input checked="" type="checkbox"/> class |

Remove

Status

OK Log x 0

Selected attribute

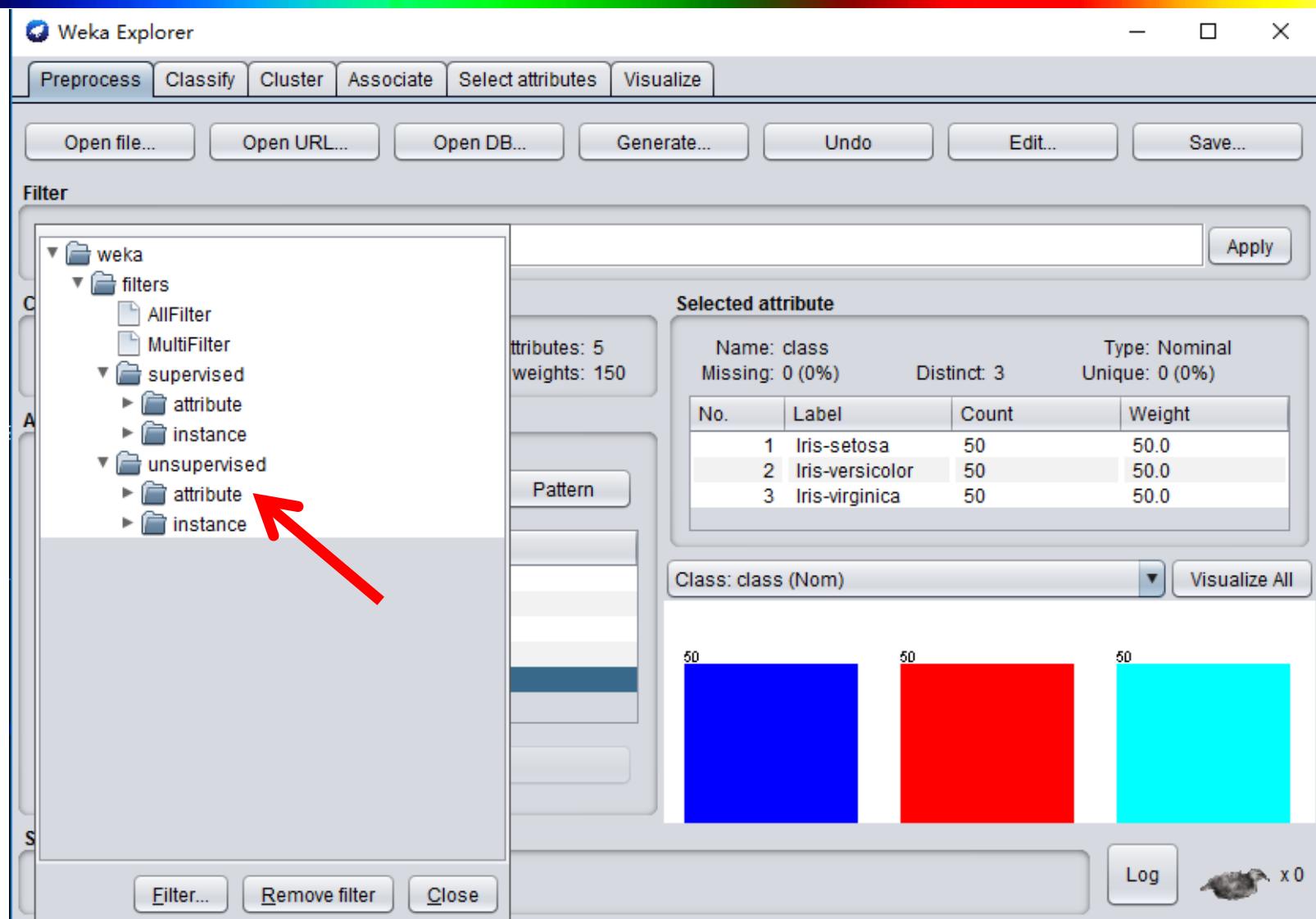
Name: class Type: Nominal
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

| No. | Label | Count | Weight |
|-----|-----------------|-------|--------|
| 1 | Iris-setosa | 50 | 50.0 |
| 2 | Iris-versicolor | 50 | 50.0 |
| 3 | Iris-virginica | 50 | 50.0 |

Class: class (Nom) Visualize All

50 50 50

Explorer: Data Pre-processing



Explorer: Data Pre-processing

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

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

Filter

Selected attribute

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

| Statistic | Value |
|-----------|-------|
| Minimum | 4.3 |
| Maximum | 7.9 |
| Mean | 5.843 |
| StdDev | 0.828 |

Class: class (Nom) Visualize All

Discretize

Log x 0

Filter... Remove filter Close

A red arrow points to the "Discretize" option in the list of filters.

Explorer: Data Pre-processing

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

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

Filter

Choose Discretize -B 10 -M -1.0 -R first-last Apply

Current relation

Relation: iris Instances: 150 Attributes: 5 Sum of weights: 150

Attributes

All None Invert Pattern

| No. | Name |
|-----|--|
| 1 | <input checked="" type="checkbox"/> sepal length |
| 2 | <input type="checkbox"/> sepal width |
| 3 | <input type="checkbox"/> petal length |
| 4 | <input type="checkbox"/> petal width |
| 5 | <input type="checkbox"/> class |

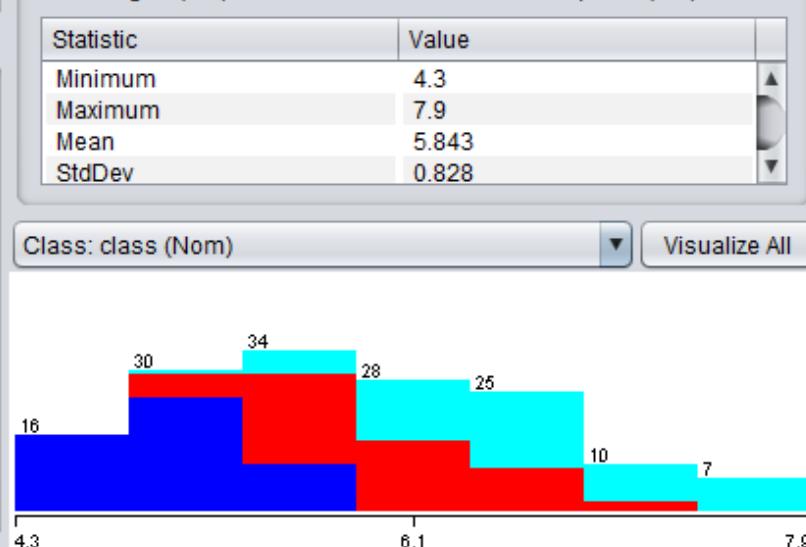
Remove

Selected attribute

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

| Statistic | Value |
|-----------|-------|
| Minimum | 4.3 |
| Maximum | 7.9 |
| Mean | 5.843 |
| StdDev | 0.828 |

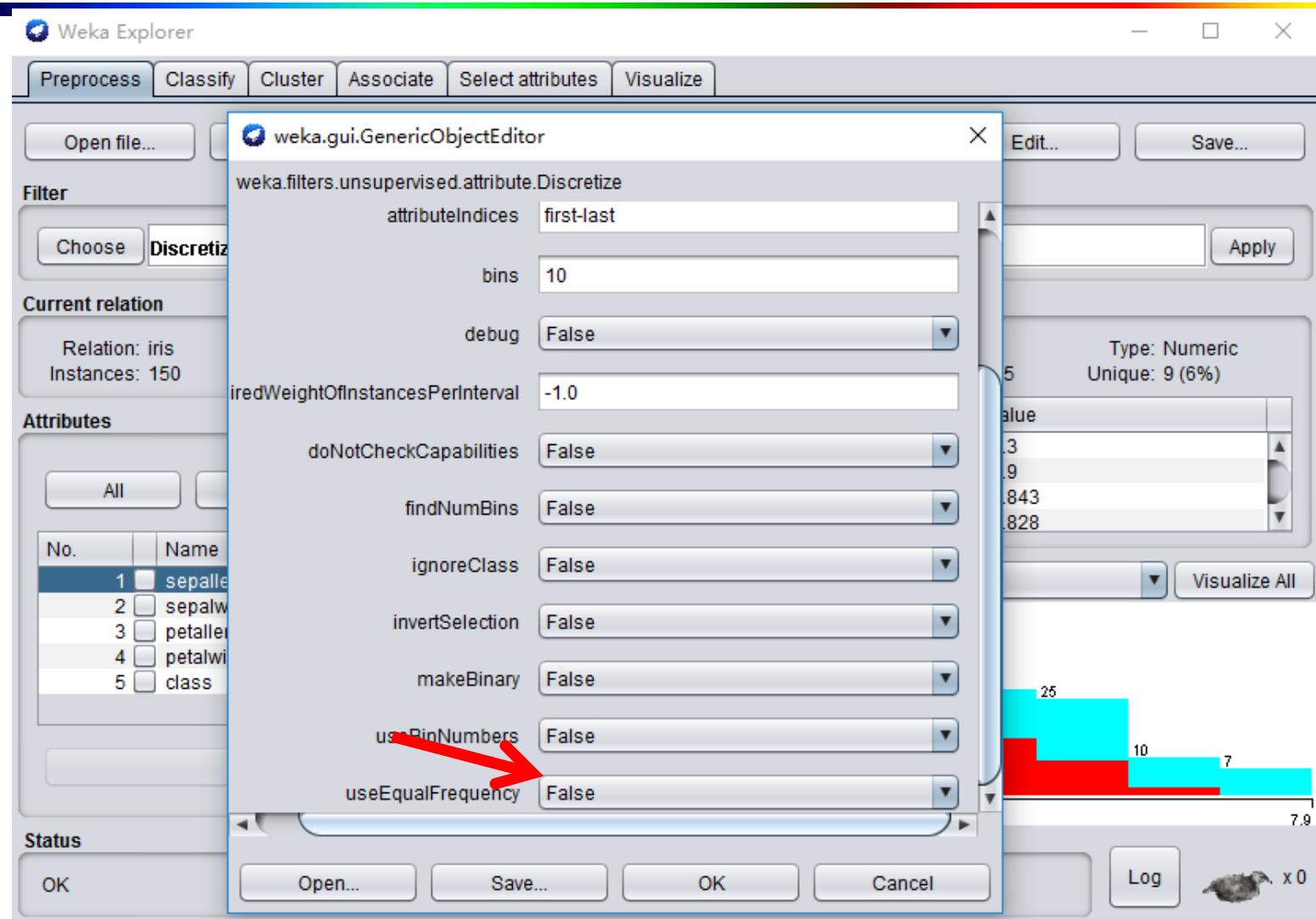
Class: class (Nom) Visualize All



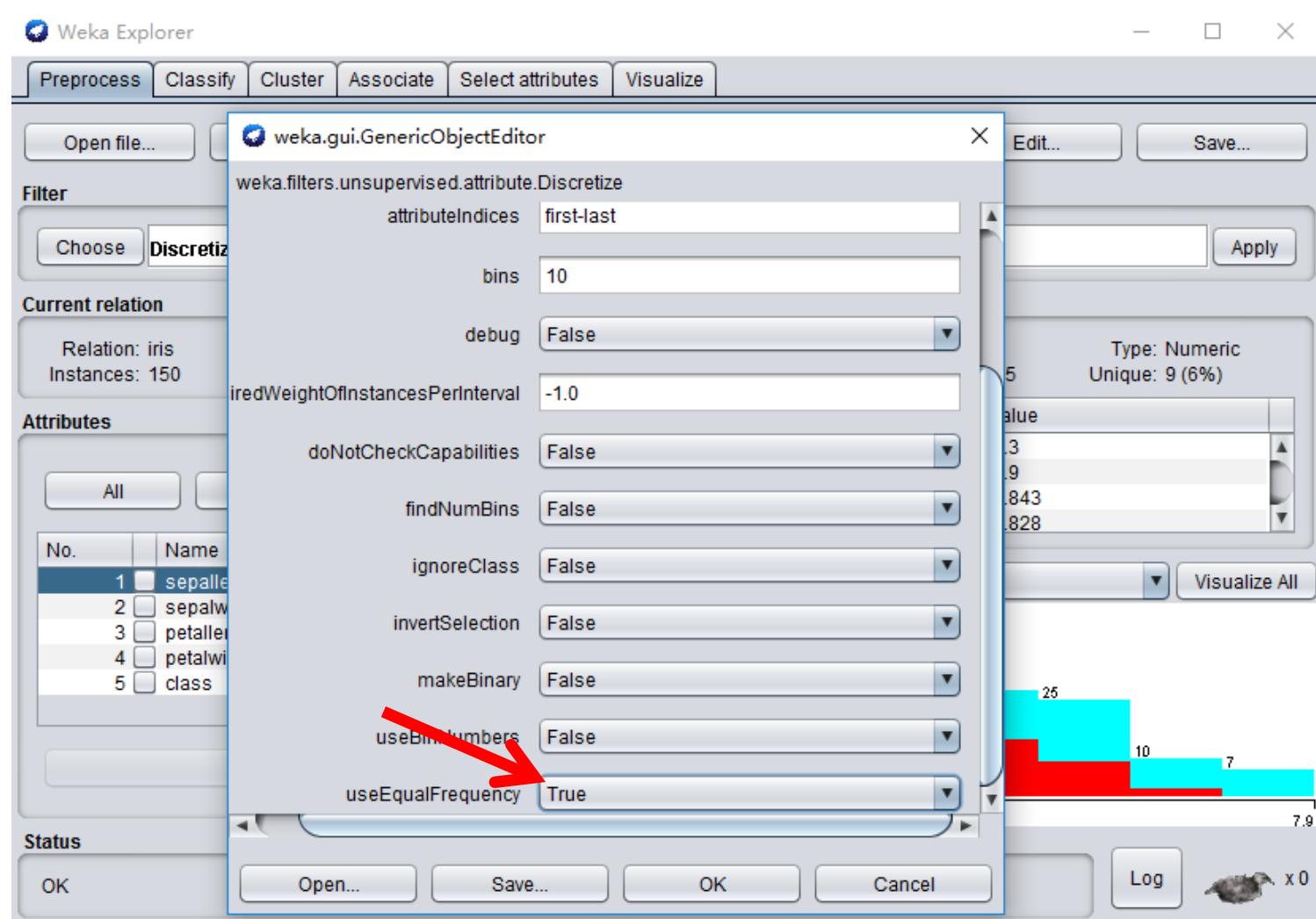
Status

OK Log x 0

Explorer: Data Pre-processing



Explorer: Data Pre-processing



Explorer: Data Pre-processing

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

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

Filter

Choose Discretize -F -B 10 -M 1.0 -R first-last Apply

Current relation

Relation: iris Attributes: 5
Instances: 150 Sum of weights: 150

Attributes

All None Invert Pattern

| No. | Name |
|-----|---|
| 1 | <input checked="" type="checkbox"/> sepallength |
| 2 | <input type="checkbox"/> sepalwidth |
| 3 | <input type="checkbox"/> petallength |
| 4 | <input type="checkbox"/> petalwidth |
| 5 | <input type="checkbox"/> class |

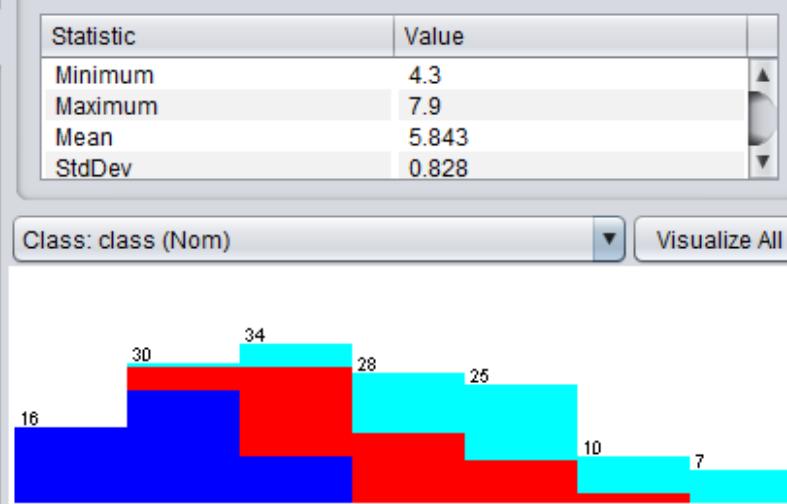
Remove

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 |

Class: class (Nom) Visualize All



Status

OK Log x 0

A red arrow points to the "Apply" button in the Filter section.

Explorer: Data Pre-processing

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

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

Filter

Choose Discretize -F -B 10 -M -1.0 -R first-last Apply

Current relation

Relation: iris-weka.filters.unsupervise... Attributes: 5
Instances: 150 Sum of weights: 150

Selected attribute

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

| No. | Label | Count | Weight |
|-----|---------------|-------|--------|
| 1 | '(-inf-4.85]' | 16 | 16.0 |
| 2 | '(4.85-5.05]' | 16 | 16.0 |
| 3 | '(5.05-5.35]' | 14 | 14.0 |
| 4 | '(5.35-5.55]' | 13 | 13.0 |

Attributes

All None Invert Pattern

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

Remove

Status

OK Log x 0

Class: class (Nom) Visualize All

| Bin | Blue (Sepal Length) | Red (Sepal Width) | Cyan (Petal Length) | Total Count |
|-----|---------------------|-------------------|---------------------|-------------|
| 1 | 16 | 0 | 0 | 16 |
| 2 | 16 | 0 | 0 | 16 |
| 3 | 14 | 0 | 0 | 14 |
| 4 | 13 | 0 | 0 | 13 |
| 5 | 14 | 0 | 0 | 14 |
| 6 | 16 | 0 | 0 | 16 |
| 7 | 19 | 0 | 0 | 19 |
| 8 | 14 | 0 | 0 | 14 |
| 9 | 15 | 0 | 0 | 15 |
| 10 | 13 | 0 | 0 | 13 |

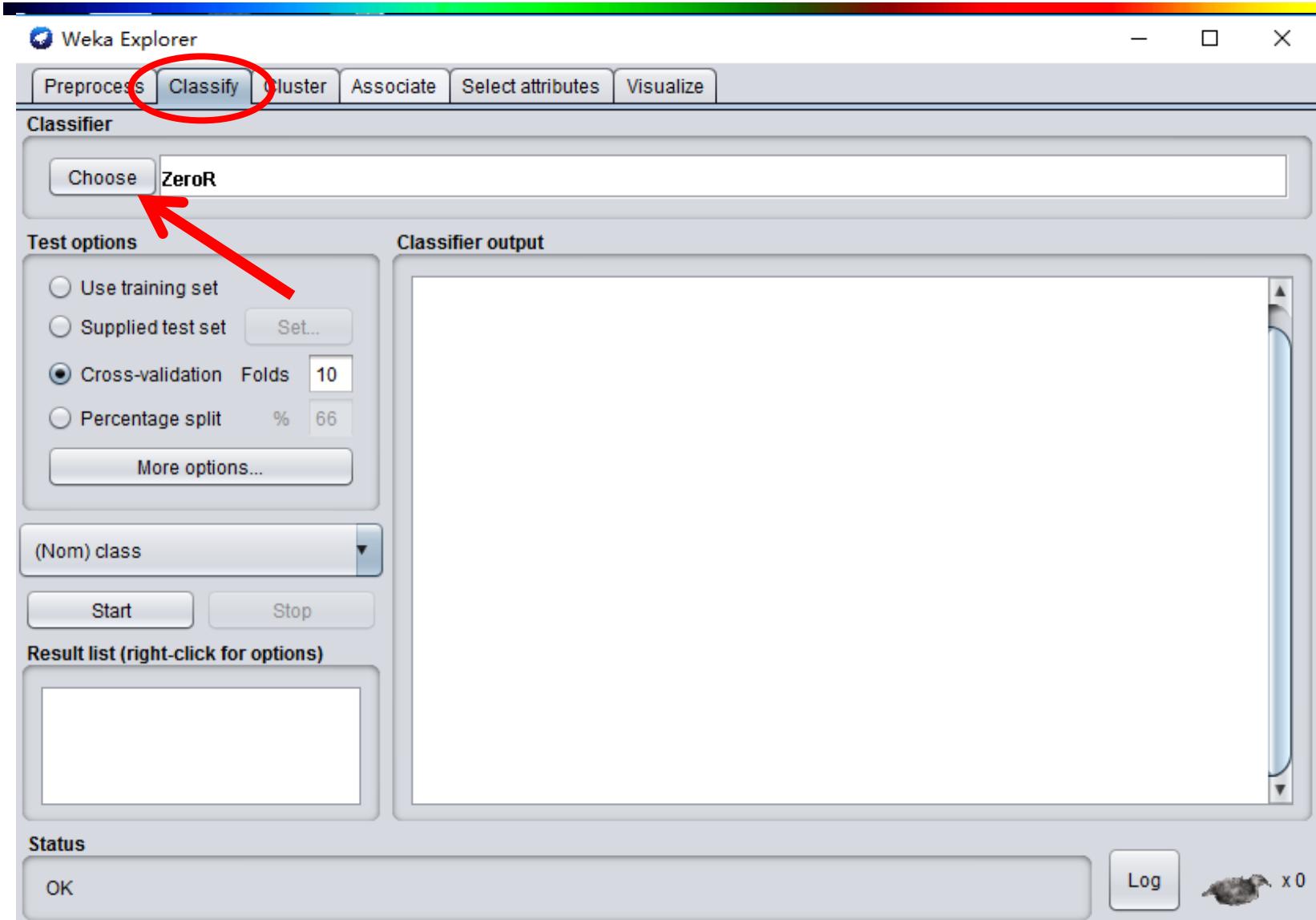
Roadmap: WEKA Usage

- WEKA Usage (version 3.8)
 - Explorer
 - Data preprocessing
 - Classification
 - Clustering
 - Association rule
 - Attribute selection (Feature selection)
 - Data visualization
 - Experimenter
 - Knowledge Flow

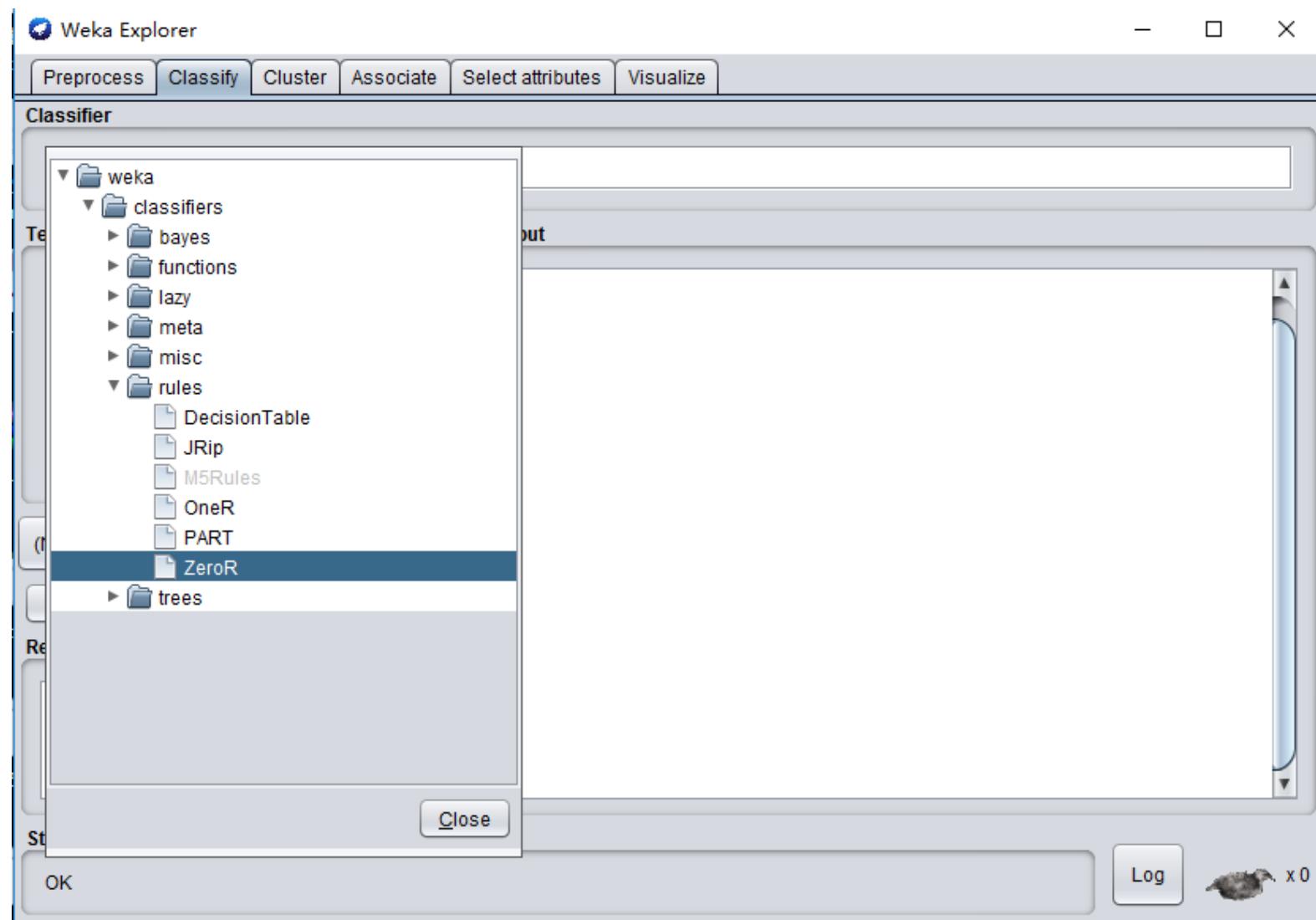
Explorer: Building Classifiers

- Classifiers in WEKA are models for predicting nominal or numeric quantities
- Implemented learning schemes include:
 - Decision trees, 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, ...

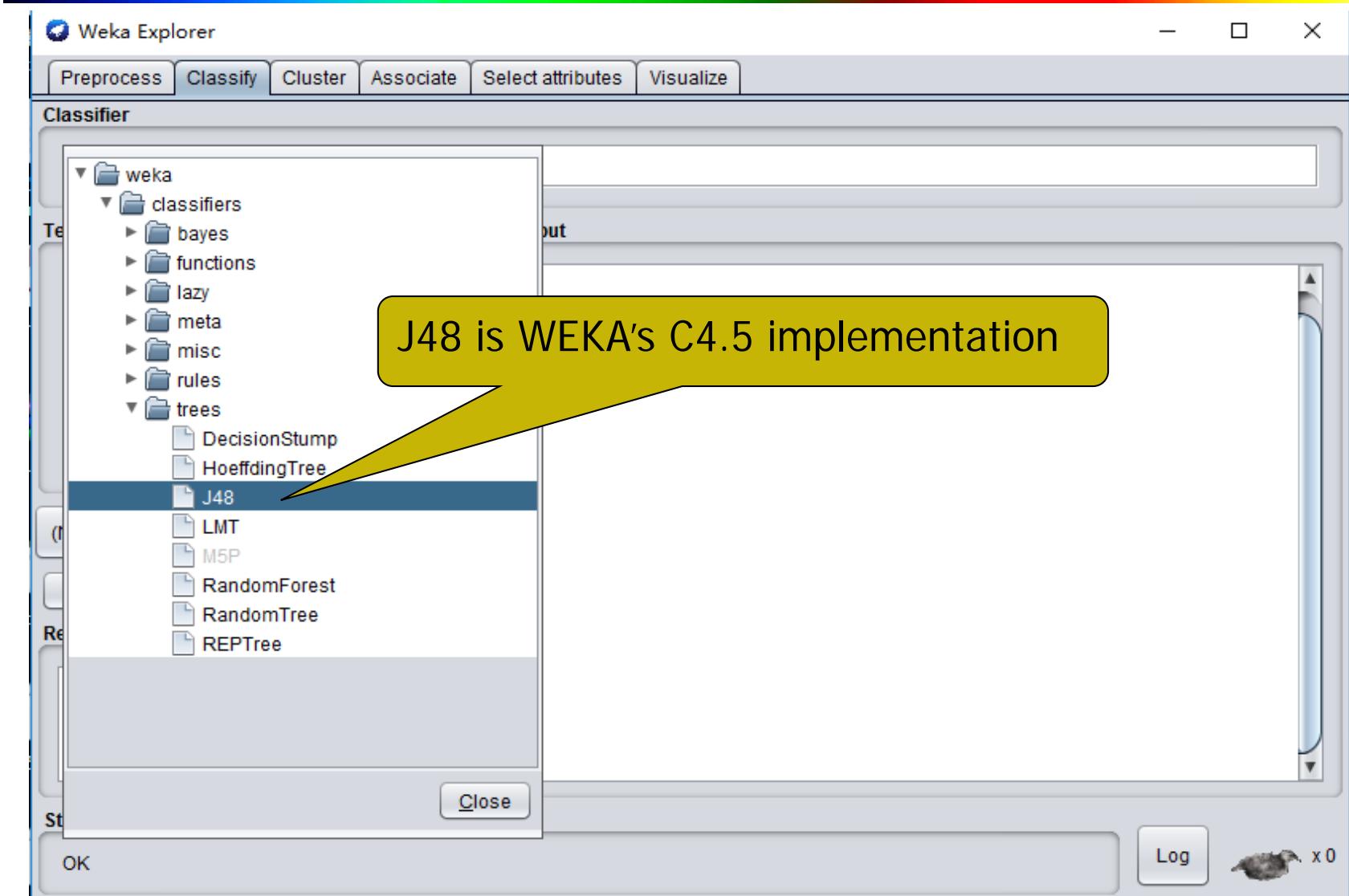
Explorer: Classification



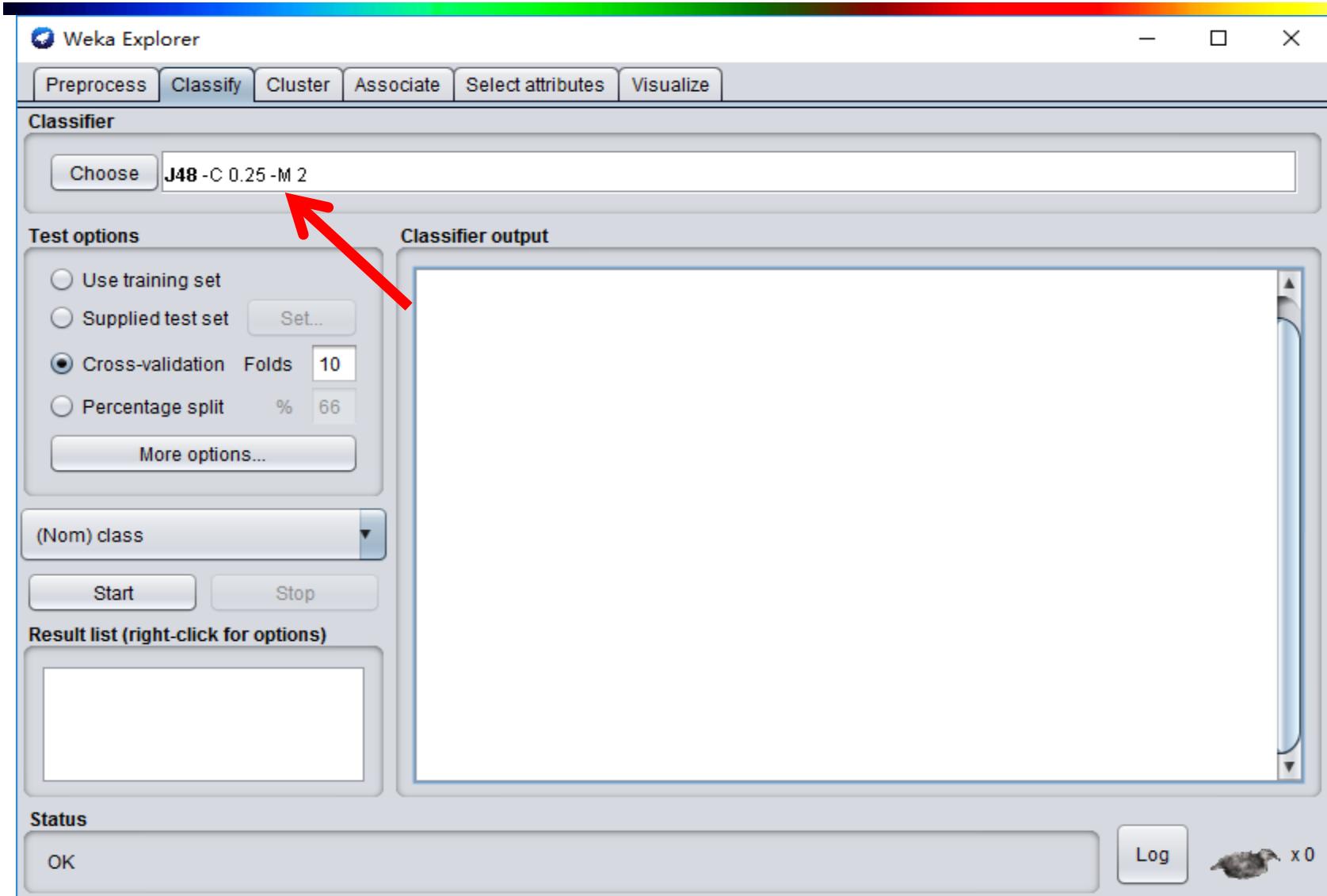
Explorer: Classification



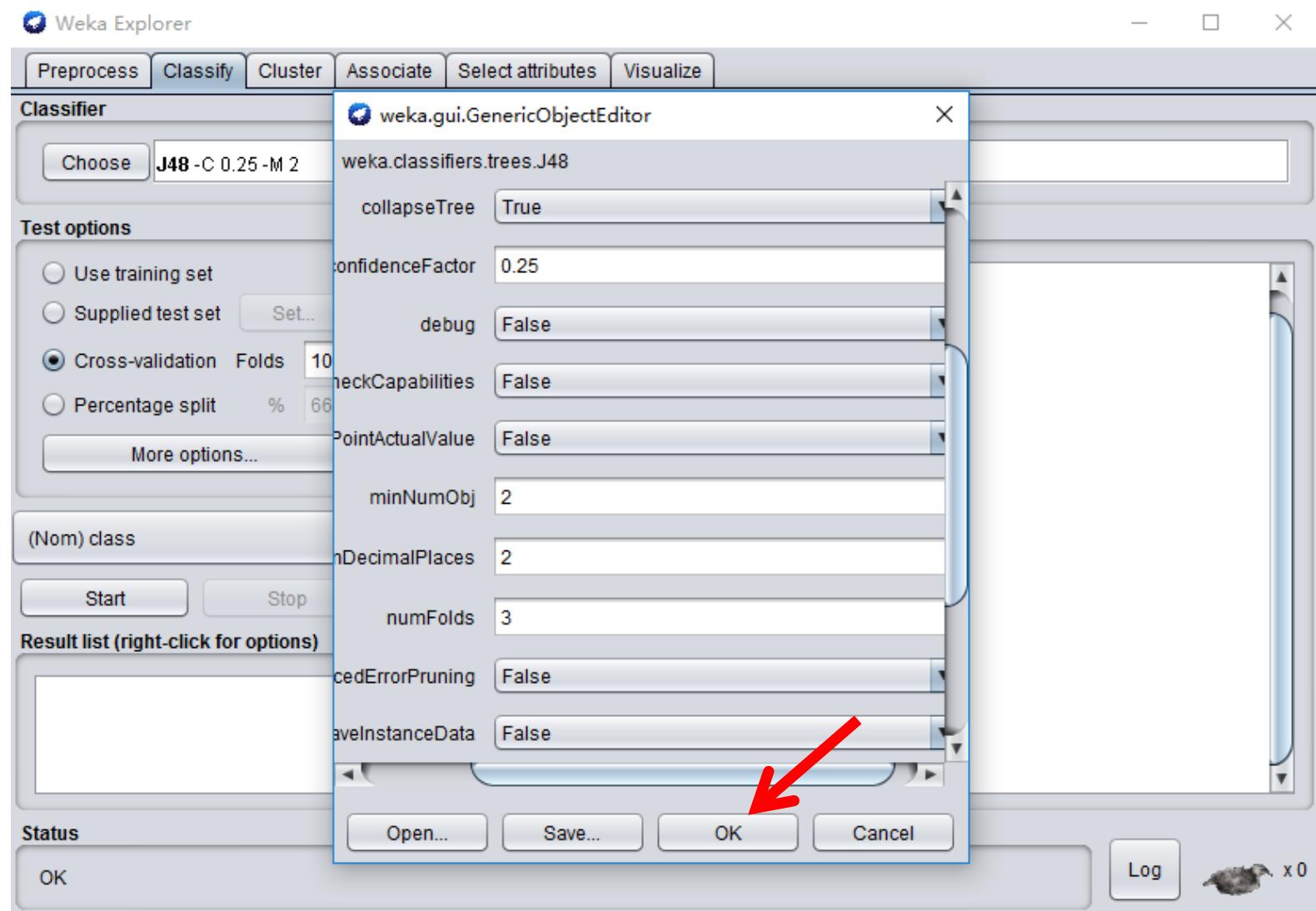
Explorer: Classification



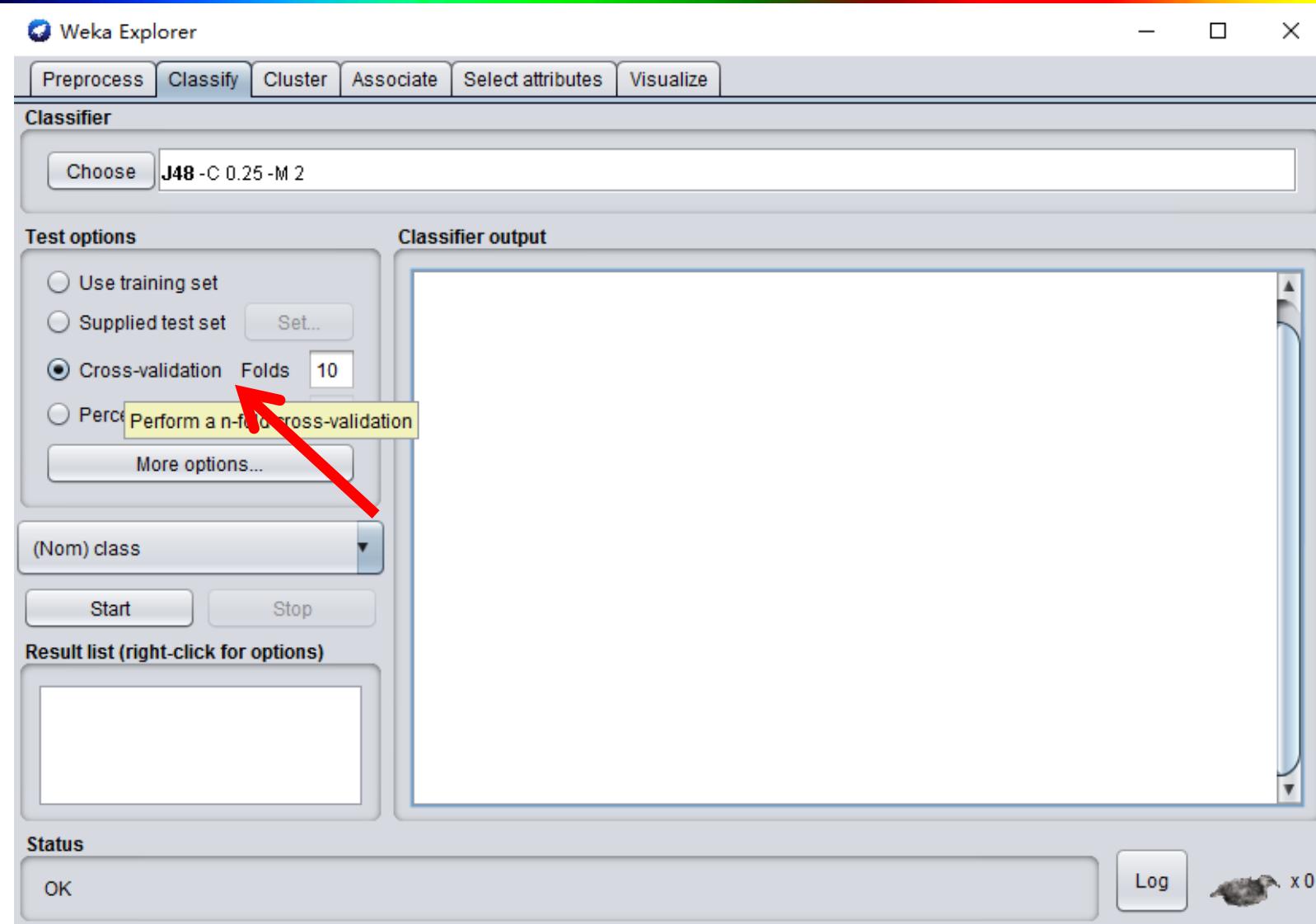
Explorer: Classification



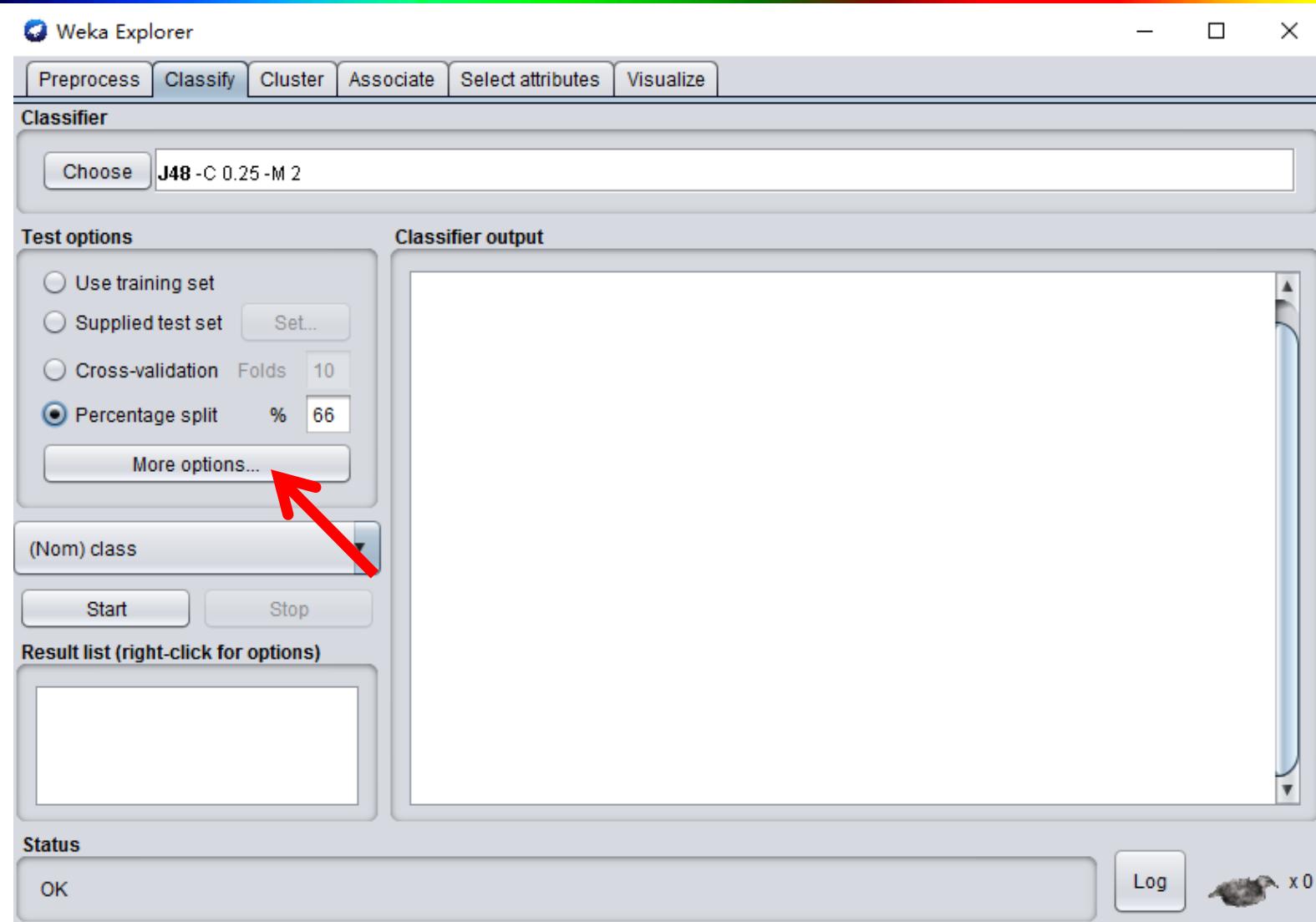
Explorer: Classification



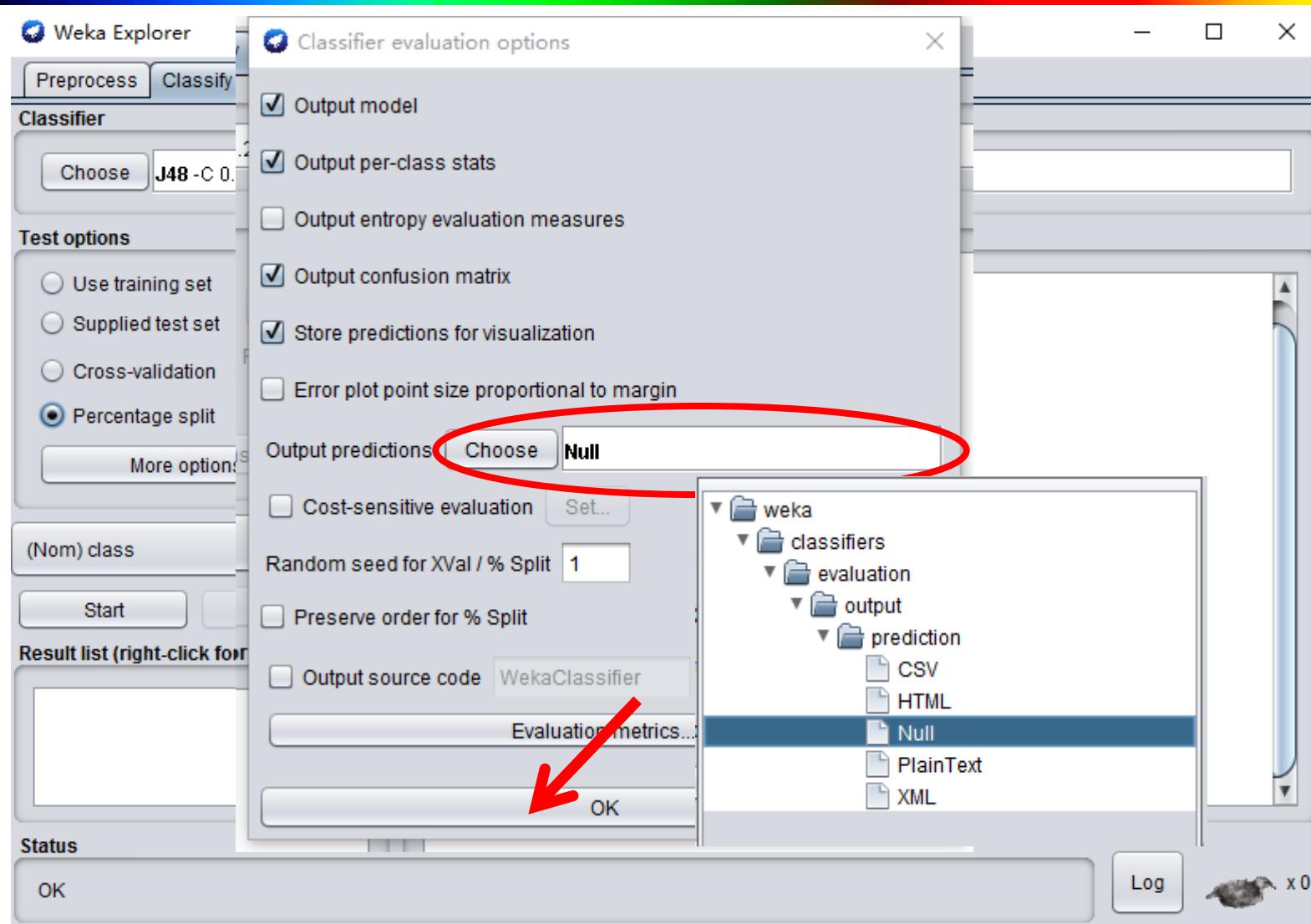
Explorer: Classification



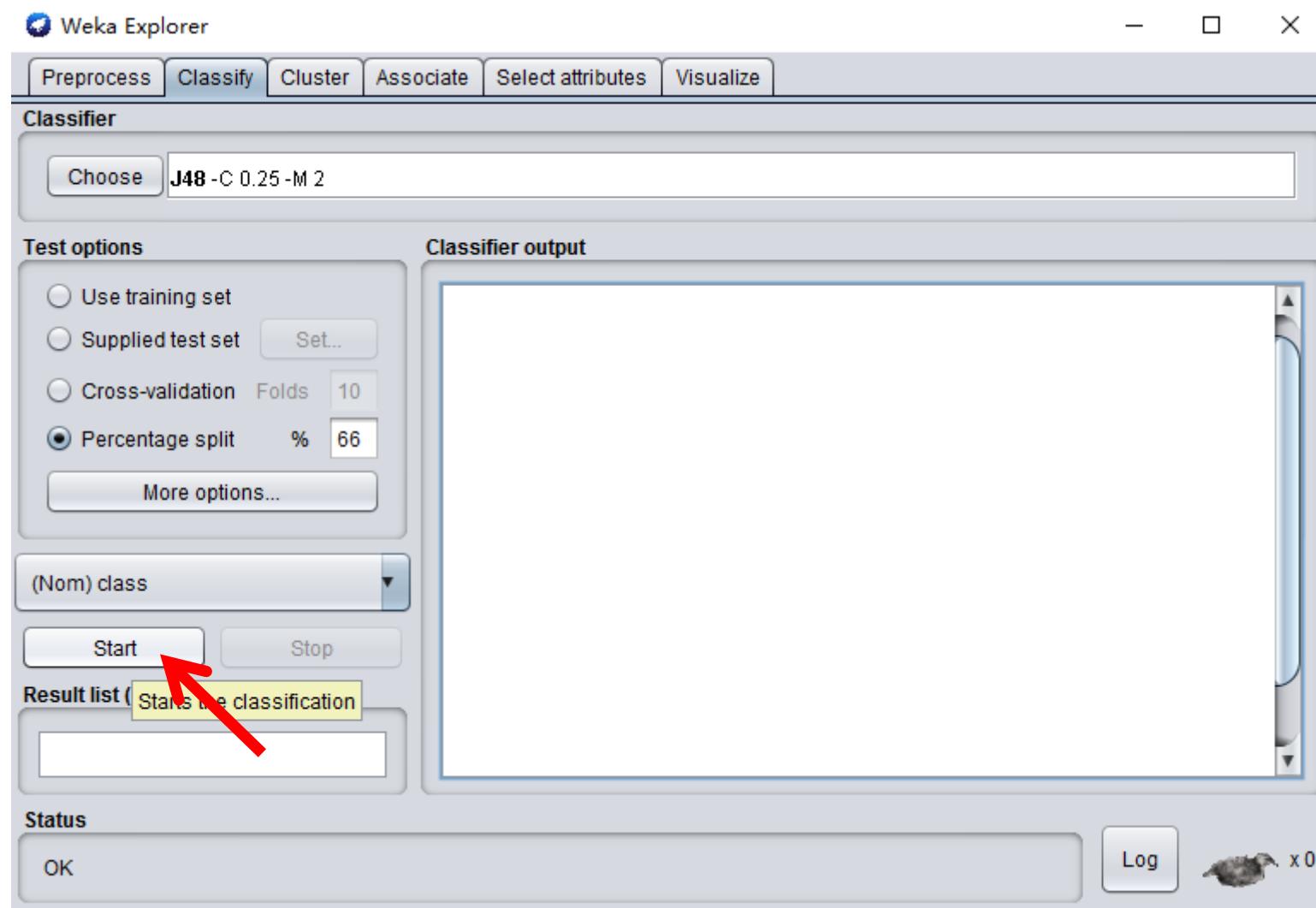
Explorer: Classification



Explorer: Classification



Explorer: Classification



Explorer: Classification

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)

21:41:50 - trees.J48

Classifier output

```
==== Run information ====
Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: iris-weka.filters.unsupervised.attribute.Discretize-F-B10-M-1.
Instances: 150
Attributes: 5
sepallength
sepalwidth
petallength
petalwidth
class
Test mode: split 66.0% train, remainder test
==== Classifier model (full training set) ====
J48 pruned tree
-----
petalwidth = '(-inf-0.15]': Iris-setosa (6.0)
```

Status

OK Log x 0



Explorer: Classification

The screenshot shows the Weka Explorer interface with the following details:

- Top Bar:** Weka Explorer, Preprocess, Classify (selected), Cluster, Associate, Select attributes, Visualize.
- Classifier Panel:** Choose J48 - C 0.25 - M 2.
- Test options:** Percentage split (66%).
- Result list (right-click for options):** 21:53:09 - trees.J48 (highlighted with a red arrow).
- Classifier output:**
 - Correctly Classified Instances: 49 (96.0784 %)
 - Incorrectly Classified Instances: 2 (3.9216 %)
 - Kappa statistic: 0.9408
 - Mean absolute error: 0.0396
 - Root mean squared error: 0.1579
 - Relative absolute error: 8.8979 %
 - Root relative squared error: 33.4091 %
 - Total Number of Instances: 51

==== Detailed Accuracy By Class ====

| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC |
|---------------|---------|---------|-----------|--------|-----------|-------|
| 1 | 1.000 | 0.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 2 | 1.000 | 0.063 | 0.905 | 1.000 | 0.950 | 0.921 |
| 3 | 0.882 | 0.000 | 1.000 | 0.882 | 0.938 | 0.913 |
| Weighted Avg. | 0.961 | 0.023 | 0.965 | 0.961 | 0.961 | 0.942 |

==== Confusion Matrix ====

| | a | b | c | classified as |
|---|----|----|----|---------------------|
| a | 15 | 0 | 0 | a = Iris-setosa |
| b | 0 | 19 | 0 | b = Iris-versicolor |
| c | 0 | 2 | 15 | c = Iris-virginica |
- Status:** OK.

Explorer: Classification

The screenshot shows the Weka Explorer interface for classification. The top menu bar includes tabs for Preprocess, Classify (which is selected), Cluster, Associate, Select attributes, and Visualize. The main area is titled "Classifier" and shows a "Choose" button followed by the selected model: "J48 -C 0.25 -M 2".

Test options:

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

(Nom) class dropdown menu.

Start and **Stop** buttons.

Result list (right-click for options)

- 21:53:09 - tree
 - View in main window
 - View in separate window
 - Save result buffer
 - Delete result buffer
 - Load model
 - Save model
 - Re-evaluate model on current test set
 - Re-apply this model's configuration
 - Visualize classifier errors
 - Visualize tree** (this option is highlighted)
 - Visualize margin curve
 - Visualize threshold curve
 - Cost/Benefit analysis
 - Visualize cost curve

Classifier output:

```
Evaluation on test split

Time taken to test model on training split: 0 seconds

==== Summary ====

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

==== Detailed Accuracy By Class ====

           TP Rate   FP Rate   Precision   Recall   F-Measure   MCC
           1.000     0.000     1.000      1.000     1.000     1.000
           1.000     0.063     0.905      1.000     0.950     0.921
           0.882     0.000     1.000      0.882     0.938     0.913
avg.       0.961     0.023     0.965      0.961     0.961     0.942

Confusion Matrix

<-- classified as
|  a = Iris-setosa
|  b = Iris-versicolor
|  c = Iris-virginica
```

Status: OK **Log:** x0

Explorer: Classification

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)

21:53:09 - trees.J48

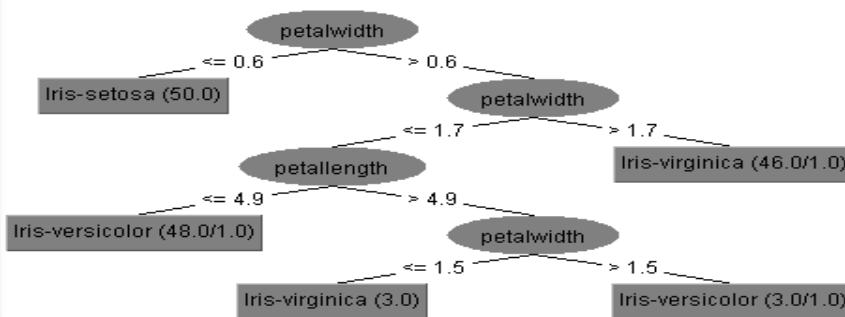
Classifier output

Evaluation on test split

Time taken to test model on training split: 0 seconds

Weka Classifier Tree Visualizer: 21:53:09 - trees.J48 (...)

Tree View



```
graph TD; Root(petalwidth <= 0.6) --> IrisSetosa[Iris-setosa (50.0)]; Root --> PetalWidth1(petalwidth > 0.6); PetalWidth1 --> PetalLength1(petallength <= 4.9); PetalWidth1 --> PetalWidth2(petalwidth > 1.7); PetalLength1 --> IrisVersicolor1[Iris-versicolor (48.0/1.0)]; PetalLength1 --> IrisVirginica1[Iris-virginica (46.0/1.0)]; PetalWidth2 --> PetalWidth3(petalwidth <= 1.5); PetalWidth2 --> PetalWidth4(petalwidth > 1.5); PetalWidth3 --> IrisVirginica2[Iris-virginica (3.0)]; PetalWidth3 --> IrisVersicolor2[Iris-versicolor (3.0/1.0)];
```

MCC

| |
|-------|
| 1.000 |
| 0.921 |
| 0.913 |
| 0.942 |

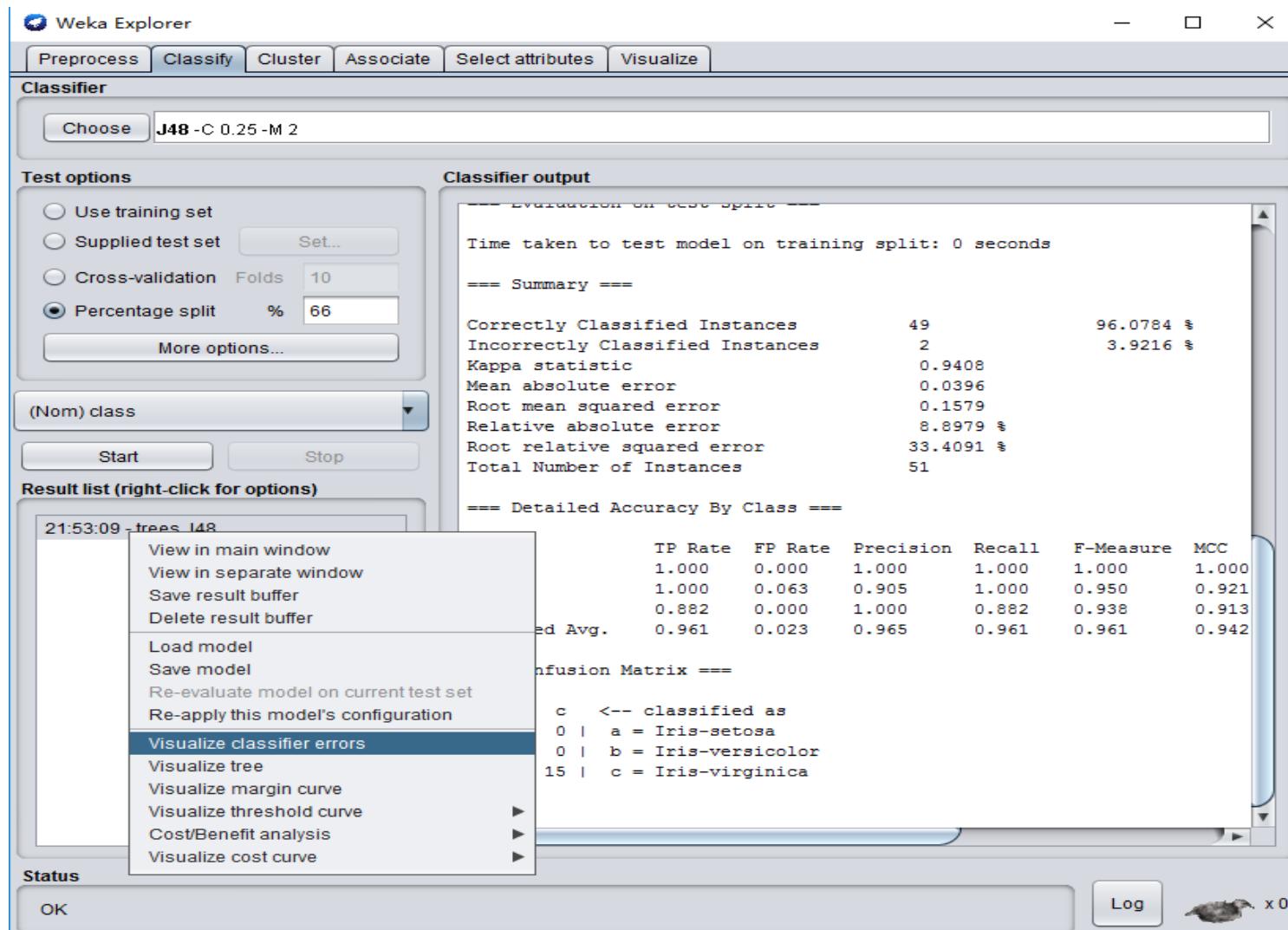
Status

OK

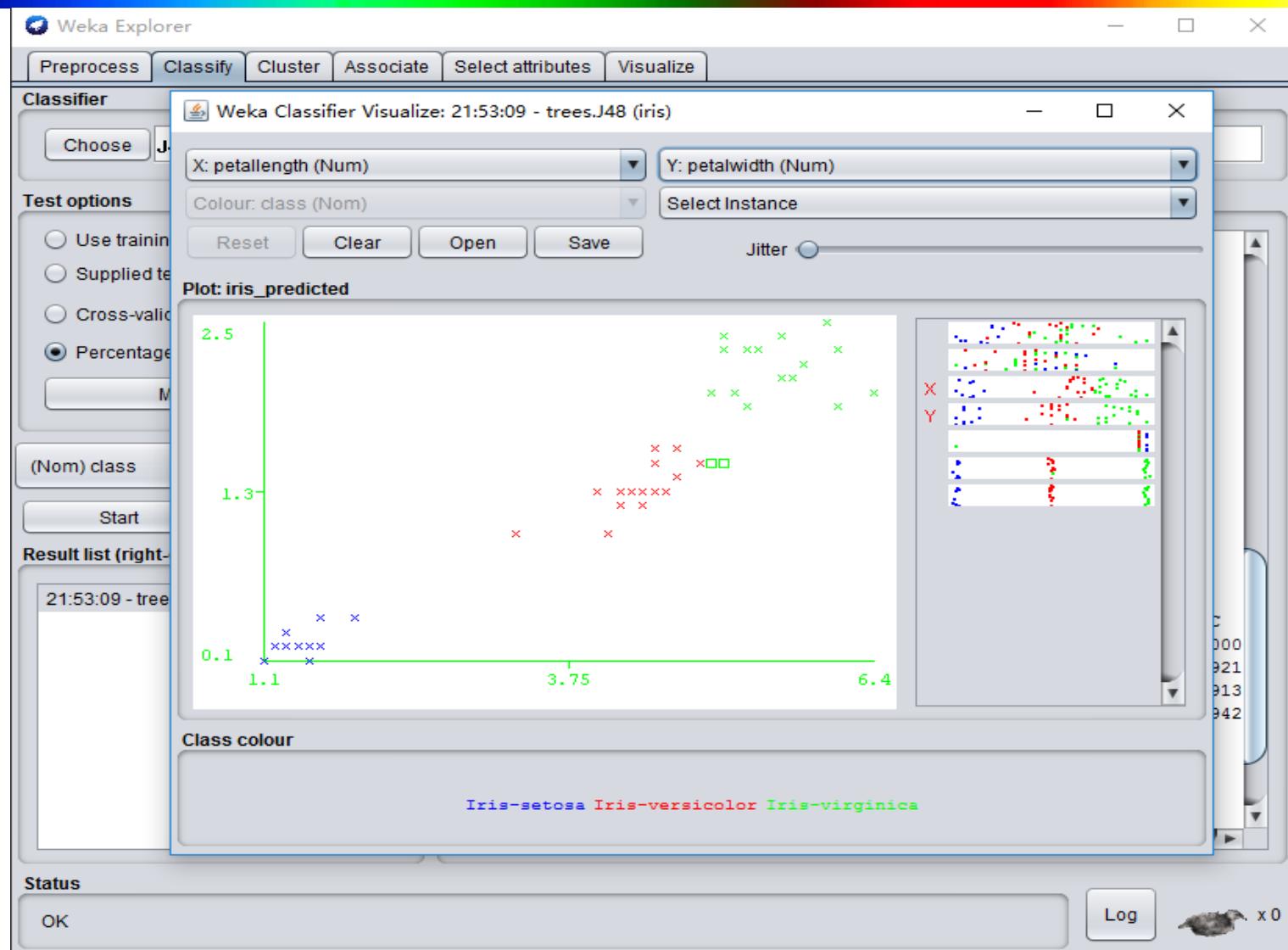
Log x 0

A small mouse icon is located in the bottom right corner.

Explorer: Classification



Explorer: Classification



Explorer: Classification

The screenshot shows the Weka Explorer interface for classification. A red arrow points to the 'Choose' button in the Classifier panel, which has 'J48 - C 0.25 - M 2' selected. The 'Test options' panel shows 'Percentage split' set to 66%. The 'Classifier output' panel displays evaluation metrics for the J48 model on a test split, including a summary table and detailed accuracy by class.

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)

21:53:09 - trees.J48

Classifier output

Time taken to test model on training split: 0 seconds

==== Summary ===

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

==== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC |
|---------------|---------|---------|-----------|--------|-----------|-------|
| 1.000 | 0.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 1.000 | 0.063 | 0.905 | 1.000 | 0.950 | 0.921 | 0.921 |
| 0.882 | 0.000 | 1.000 | 0.882 | 0.938 | 0.913 | 0.913 |
| Weighted Avg. | 0.961 | 0.023 | 0.965 | 0.961 | 0.961 | 0.942 |

==== Confusion Matrix ===

| | a | b | c | <-- classified as |
|---|----|---|---|--------------------|
| a | 15 | 0 | 0 | 15 - This dataset |
| b | 0 | 1 | 0 | 1 - Other datasets |
| c | 0 | 0 | 1 | 1 - Other datasets |

Status

OK Log x 0

Explorer: Classification

The screenshot shows the Weka Explorer interface with the title bar "Weka Explorer". The menu bar includes "Preprocess", "Classify" (which is selected), "Cluster", "Associate", "Select attributes", and "Visualize". The main window is titled "Classifier" and displays a tree view of available classifiers under the "weka.classifiers" package. The "MultilayerPerceptron" classifier is selected and highlighted in blue. To the right of the tree view is the "Test options" panel, which contains the command: "-V 0 -S 0 -E 20 -H a". Below this is the "Test output" pane, which shows the results of a classification run. The output includes a summary of correctly classified instances (49, 96.0784 %) and misclassified instances (2, 3.9216 %). It also provides statistics like a statistic (0.9408), absolute error (0.0396), mean squared error (0.1579), and relative squared error (8.8979 %). The total number of instances is 51. A detailed accuracy by class table is shown, with columns for TP Rate, FP Rate, Precision, Recall, F-Measure, and MCC. The table includes rows for each class and an overall "Tested Avg." row. At the bottom of the output pane is a "Confusion Matrix" section. The status bar at the bottom left shows "OK", and the bottom right has a "Log" button and a small icon.

-V 0 -S 0 -E 20 -H a

Test output

taken to test model on training split: 0 seconds

Summary ===

| | Correctly Classified Instances | % |
|---|--------------------------------|-----------|
| a | 49 | 96.0784 % |
| b | 2 | 3.9216 % |

a statistic 0.9408
absolute error 0.0396
mean squared error 0.1579
relative squared error 8.8979 %
Number of Instances 51

Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC |
|-------------|---------|---------|-----------|--------|-----------|-------|
| a | 1.000 | 0.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| b | 1.000 | 0.063 | 0.905 | 1.000 | 0.950 | 0.921 |
| c | 0.882 | 0.000 | 1.000 | 0.882 | 0.938 | 0.913 |
| Tested Avg. | 0.961 | 0.023 | 0.965 | 0.961 | 0.961 | 0.942 |

== Confusion Matrix ==

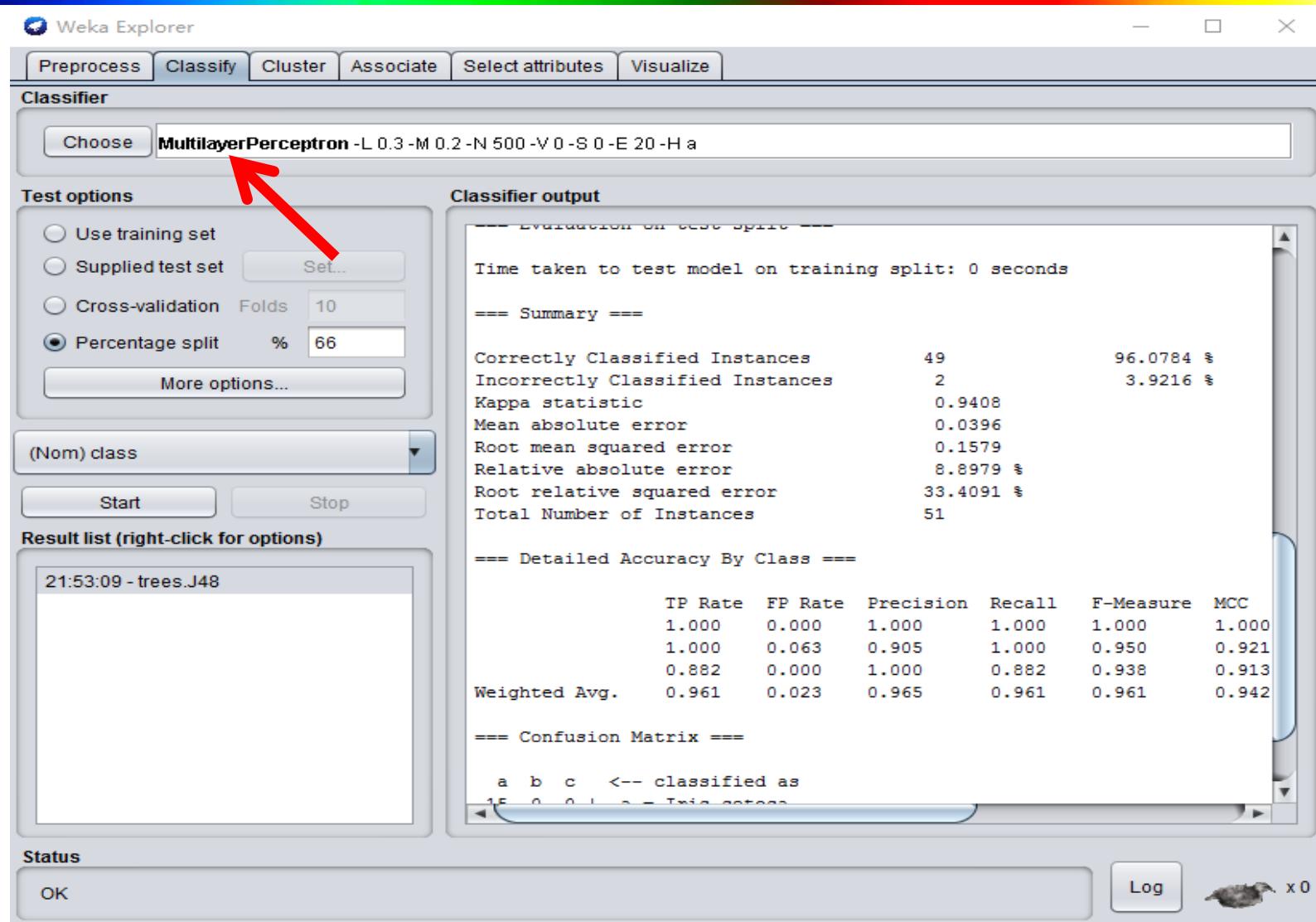
| | a | b | c | <-- classified as |
|---|----|---|---|-------------------|
| a | 49 | 0 | 1 | a |
| b | 0 | 2 | 0 | b |
| c | 1 | 0 | 0 | c |

Status

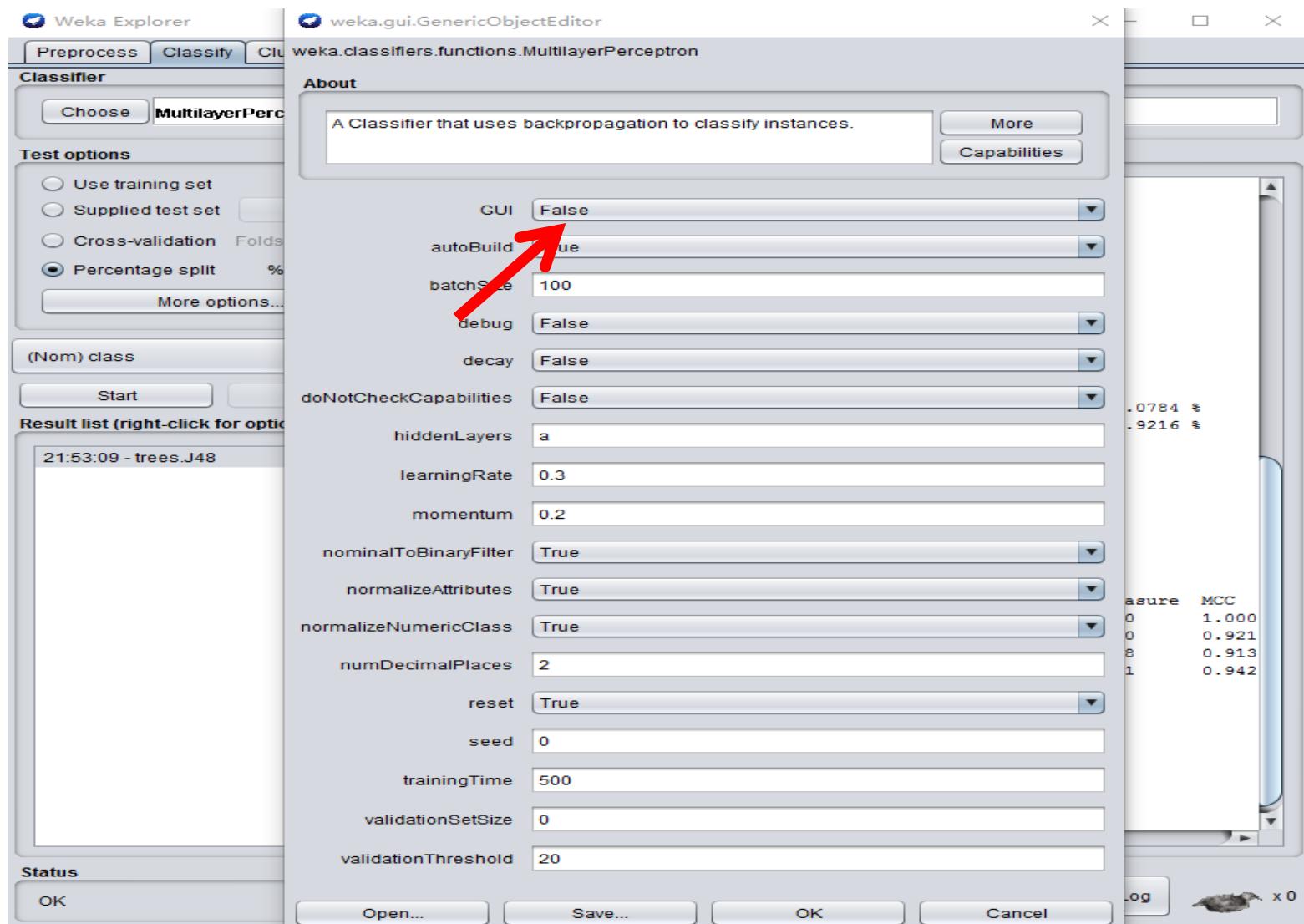
OK

Log x 0

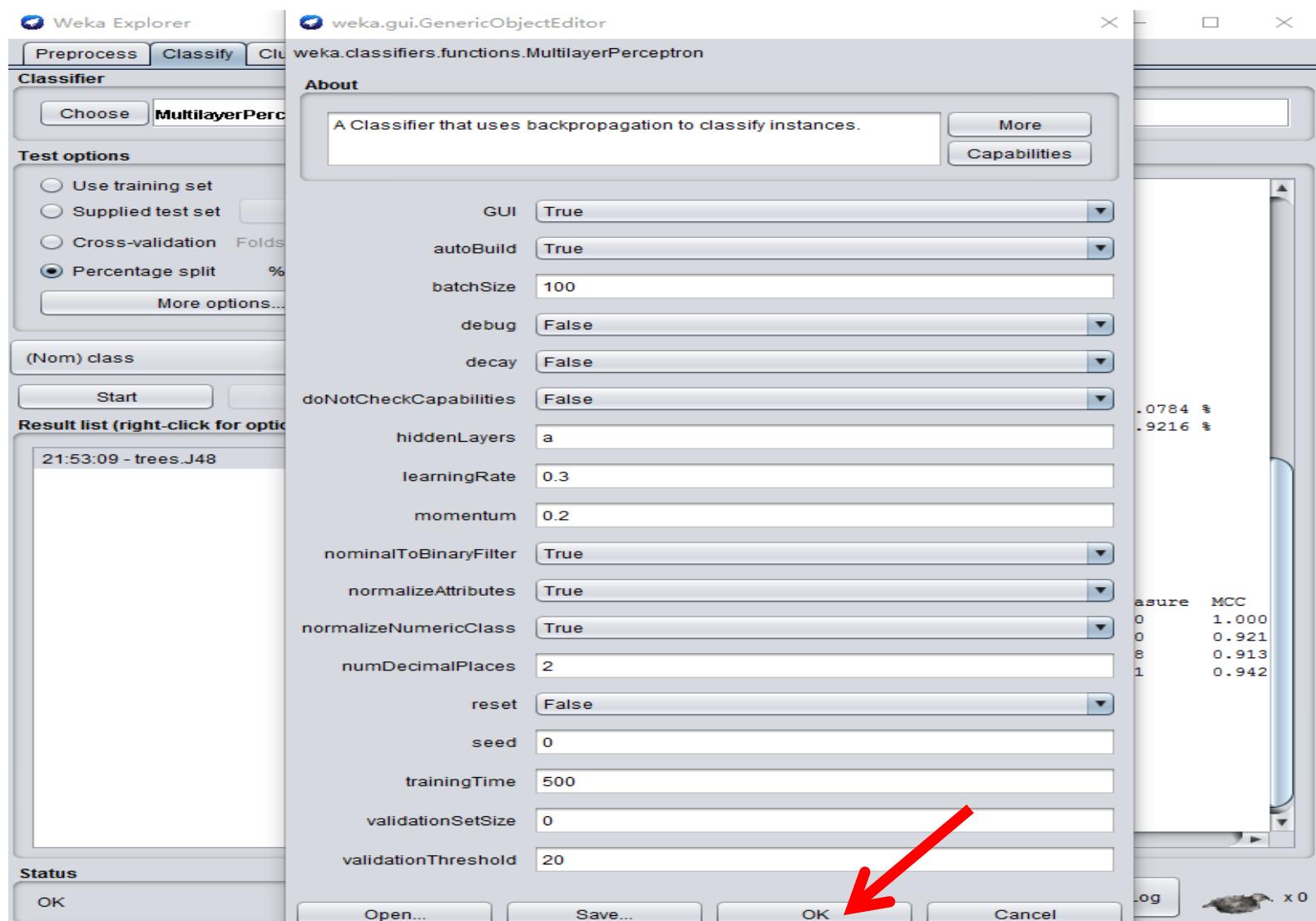
Explorer: Classification



Explorer: Classification



Explorer: Classification



Explorer: Classification

The screenshot shows the Weka Explorer interface with the following details:

- Top Bar:** Weka Explorer, Preprocess, Classify (highlighted), Cluster, Associate, Select attributes, Visualize.
- Classifier Panel:** Choose MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a -G -R.
- Test options Panel:** Percentage split (66%) selected. Other options include Use training set, Supplied test set, Cross-validation (Folds 10), and More options... button.
- Result list (right-click for options):** Displays "21:53:09 - trees.J4". A red arrow points to the "Start" button.
- Classifier output Panel:**
 - Number of Leaves : 5
 - Size of the tree : 9
 - Time taken to build model: 0 seconds
 - ==== Evaluation on test split ===
 - Time taken to test model on training split: 0 seconds
 - ==== Summary ===

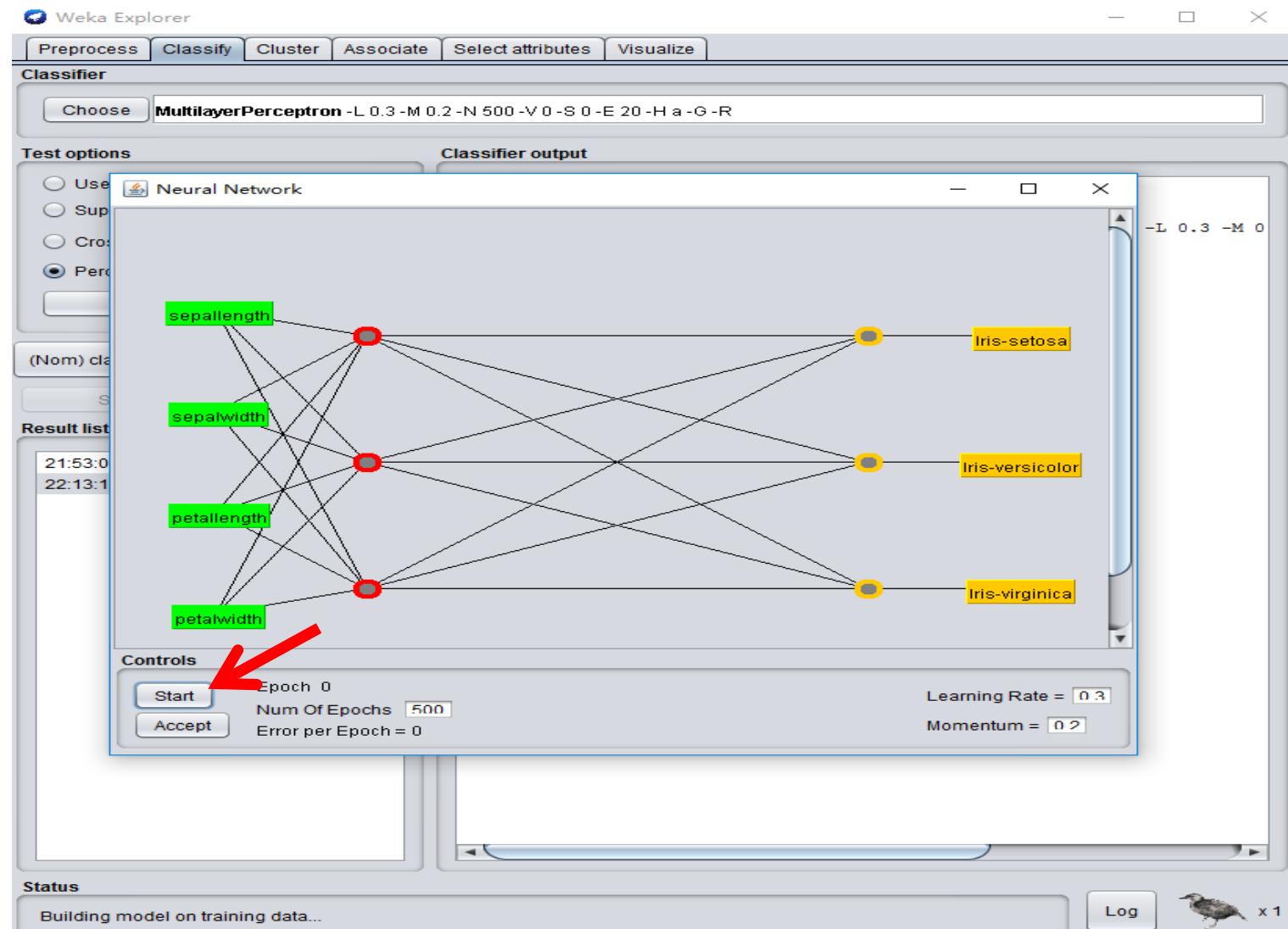
| | Correctly Classified Instances | 49 | 96.0784 % |
|----------------------------------|--------------------------------|----------|-----------|
| Incorrectly Classified Instances | 2 | 3.9216 % | |
| Kappa statistic | 0.9408 | | |
| Mean absolute error | 0.0396 | | |
| Root mean squared error | 0.1579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

 - ==== Detailed Accuracy By Class ====

| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC |
|---------------|---------|---------|-----------|--------|-----------|-------|
| 1 | 1.000 | 0.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 2 | 1.000 | 0.063 | 0.905 | 1.000 | 0.950 | 0.921 |
| 3 | 0.882 | 0.000 | 1.000 | 0.882 | 0.938 | 0.913 |
| Weighted Avg. | 0.961 | 0.023 | 0.965 | 0.961 | 0.961 | 0.942 |
 - ==== Confusion Matrix ====

| | a | b | c | <-- classified as |
|---|----|----|----|---------------------|
| 1 | 15 | 0 | 0 | a = Iris-setosa |
| 2 | 0 | 19 | 0 | b = Iris-versicolor |
| 3 | 0 | 2 | 15 | c = Iris-virginica |
- Status Panel:** OK, Log, and a small icon.

Explorer: Classification



Explorer: Classification

The screenshot shows the Weka Explorer interface for classification. The top menu bar includes tabs for Preprocess, Classify (which is selected), Cluster, Associate, Select attributes, and Visualize. Below the tabs, the title "Classifier" is displayed, followed by a "Choose" button and the classifier name "MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a -G -R".

The "Test options" panel on the left contains four radio button options: "Use training set", "Supplied test set", "Cross-validation", and "Percentage split" (which is selected with a value of 66). It also includes a "More options..." button and a dropdown menu for "(Nom) class". Below these are "Start" and "Stop" buttons.

The "Classifier output" panel on the right displays the results of the classification run. It starts with the message "==== Evaluation on test split ====" and "Time taken to test model on training split: 0 seconds".

Summary:

| | Correctly Classified Instances | 50 | 98.0392 % |
|----------------------------------|--------------------------------|----------|-----------|
| Incorrectly Classified Instances | 1 | 1.9608 % | |
| Kappa statistic | 0.9704 | | |
| Mean absolute error | 0.0239 | | |
| Root mean squared error | 0.1101 | | |
| Relative absolute error | 5.3594 % | | |
| Root relative squared error | 23.2952 % | | |
| Total Number of Instances | 51 | | |

Detailed Accuracy By Class:

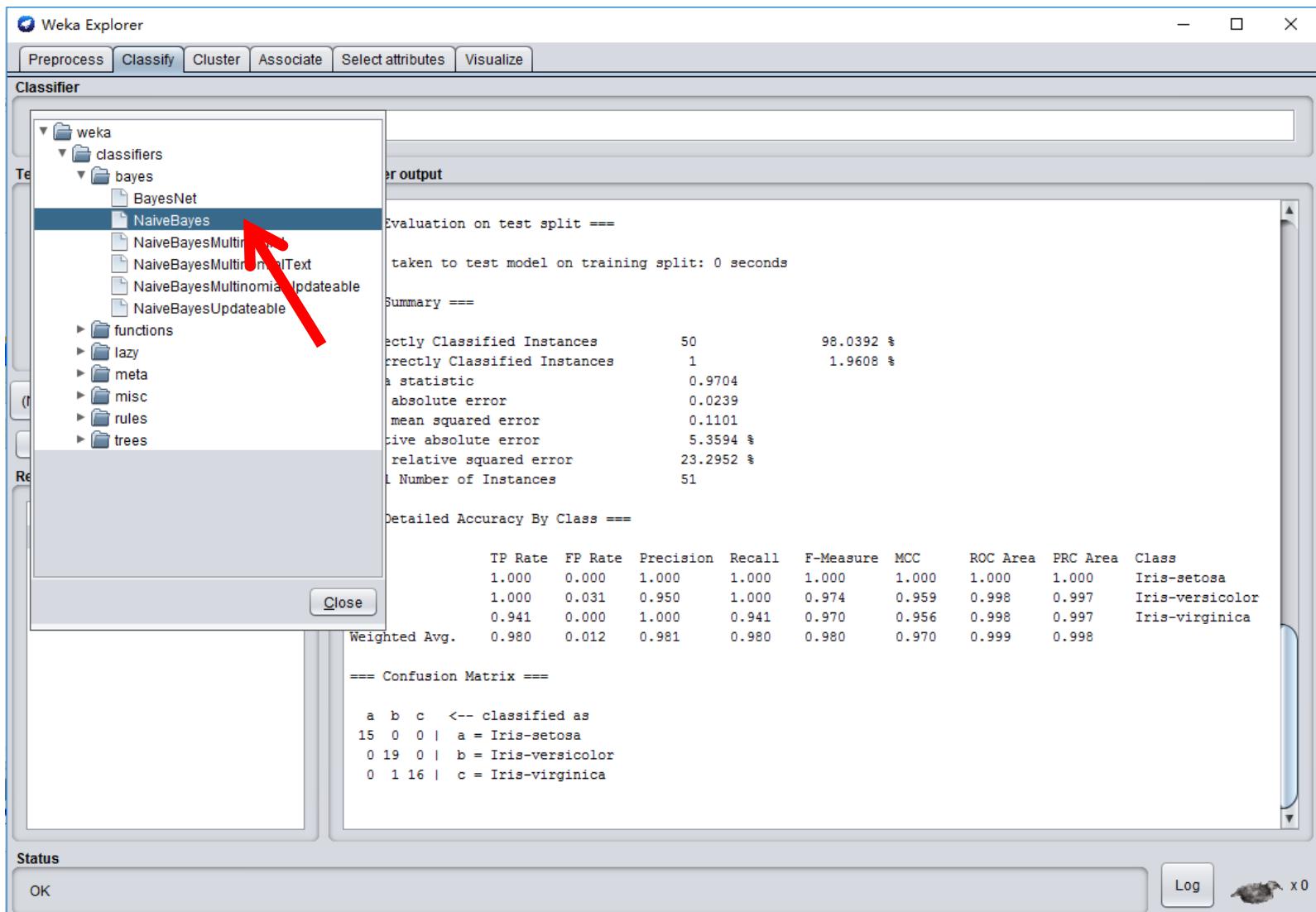
| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|-------|----------|----------|-----------------|
| 1.000 | 0.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | Iris-setosa |
| 1.000 | 0.031 | 0.950 | 1.000 | 0.974 | 0.959 | 0.998 | 0.997 | 0.997 | Iris-versicolor |
| 0.941 | 0.000 | 1.000 | 0.941 | 0.970 | 0.956 | 0.998 | 0.997 | 0.997 | Iris-virginica |
| Weighted Avg. | 0.980 | 0.012 | 0.981 | 0.980 | 0.980 | 0.970 | 0.999 | 0.998 | |

Confusion Matrix:

| | a | b | c | <-- classified as |
|----|----|----|---|---------------------|
| 15 | 0 | 0 | 1 | a = Iris-setosa |
| 0 | 19 | 0 | 1 | b = Iris-versicolor |
| 0 | 1 | 16 | 1 | c = Iris-virginica |

The "Status" panel at the bottom left shows "OK". The bottom right corner features a "Log" button and a small icon with the number "x 0".

Explorer: Classification



Explorer: Classification

The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. The 'Classifier' panel shows 'NaiveBayes' selected. In the 'Test options' panel, the 'Percentage split' radio button is selected with 66% chosen. The 'Start' button is highlighted with a red arrow. The 'Classifier output' panel displays evaluation results for the Naive Bayes model on the Iris dataset.

Classifier output:

```
==== Evaluation on test split ====
Time taken to test model on training split: 0 seconds
==== Summary ====
Correctly Classified Instances      50      98.0392 %
Incorrectly Classified Instances   1       1.9608 %
Kappa statistic                   0.9704
Mean absolute error               0.0239
Root mean squared error          0.1101
Relative absolute error           5.3594 %
Root relative squared error     23.2952 %
Total Number of Instances        51

==== Detailed Accuracy By Class ====

```

| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|-------|----------|----------|-----------------|
| 1.000 | 0.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | Iris-setosa |
| 1.000 | 0.031 | 0.950 | 1.000 | 0.974 | 0.959 | 0.998 | 0.997 | 0.997 | Iris-versicolor |
| 0.941 | 0.000 | 1.000 | 0.941 | 0.970 | 0.956 | 0.998 | 0.997 | 0.997 | Iris-virginica |
| Weighted Avg. | 0.980 | 0.012 | 0.981 | 0.980 | 0.980 | 0.970 | 0.999 | 0.998 | |

```
==== Confusion Matrix ====
a b c    <-- classified as
15 0 0 | a = Iris-setosa
0 19 0 | b = Iris-versicolor
0 1 16 | c = Iris-virginica
```

Status: Building model on training split (99 instances)... Log x 1

Explorer: Classification

The screenshot shows the Weka Explorer interface with the following details:

- Top Bar:** Weka Explorer, Preprocess, Classify (selected), Cluster, Associate, Select attributes, Visualize.
- Classifier Panel:** Choose NaiveBayes.
- Test options:** Percentage split (66%).
- Classifier output:**
 - petallength:

| | mean | 4.2452 | 5.5516 |
|------------|--------|--------|--------|
| std. dev. | 0.1782 | 0.4712 | 0.5529 |
| weight sum | 50 | 50 | 50 |
| precision | 0.1405 | 0.1405 | 0.1405 |
 - petalwidth:

| | mean | 1.3097 | 2.0343 |
|------------|--------|--------|--------|
| std. dev. | 0.1096 | 0.1915 | 0.2646 |
| weight sum | 50 | 50 | 50 |
| precision | 0.1143 | 0.1143 | 0.1143 |
- Result list:** 21:53:09 - trees.J48, 22:13:18 - functions.MultilayerPerceptron, 22:24:10 - bayes.NaiveBayes (highlighted).
- Status:** OK.
- Log:** x 0.

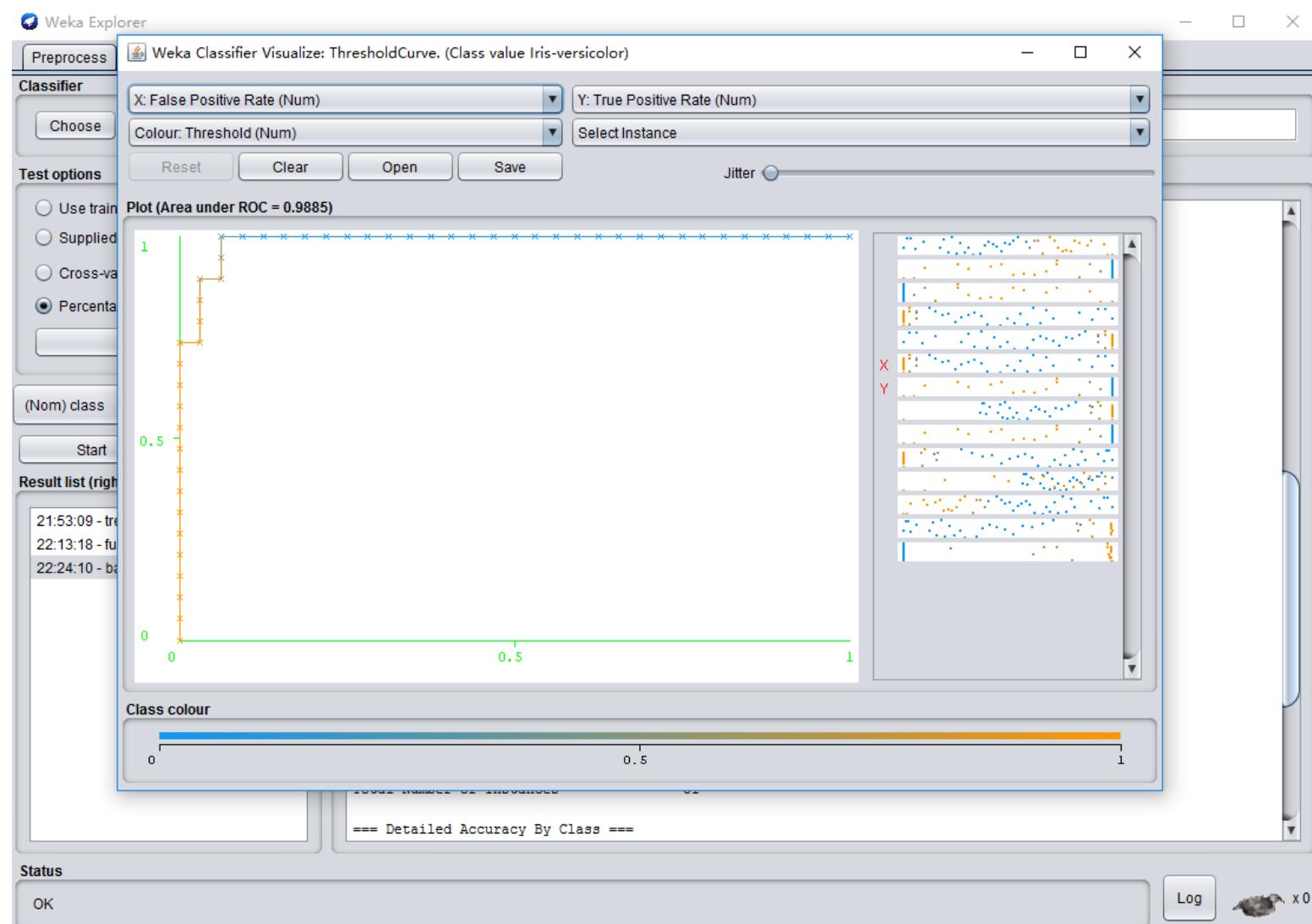
Explorer: Classification

The screenshot shows the Weka Explorer interface with the following details:

- Top Bar:** Weka Explorer, Preprocess, Classify, Cluster, Associate, Select attributes, Visualize.
- Classifier Panel:** Choose button set to NaiveBayes.
- Test options:** Use training set, Supplied test set, Cross-validation Folds 10, Percentage split % 66 (selected).
- Classifier output:** Displays descriptive statistics for petal length and petal width across four classes (mean, std. dev., weight sum, precision).

| | petal length | petal width | |
|------------|--------------|-------------|--------|
| mean | 1.4694 | 0.2743 | 4.2452 |
| std. dev. | 0.1782 | 0.1096 | 0.4712 |
| weight sum | 50 | 50 | 50 |
| precision | 0.1405 | 0.1143 | 0.1405 |
- Result list:** 21:53:09 - trees.J48, 22:13:18 - functions.MultilayerPerceptron, 22:24:10 - base.NaiveBayes. A context menu is open over the last item, listing options: View in main window, View in separate window, Save result buffer, Delete result buffer, Load model, Save model, Re-evaluate model on current test set, Re-apply this model's configuration, Visualize classifier errors, Visualize tree, Visualize margin curve, Visualize threshold curve, Cost/Benefit analysis, Visualize cost curve.
- Status:** OK.
- Log:** Shows log entries: Time taken to build model: 0 seconds, === Evaluation on test split ===, Time taken to test model on training split: 0 seconds, ===, Classified Instances 48 94.1176 %, by Class 1.0113 5.8824 %, etc.
- Bottom Right:** Log icon, Log x 0.

Explorer: Classification



Explorer: Classification

The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. On the left, a tree view displays various classifiers under the 'weka.classifiers' package, with 'M5P' highlighted by a red arrow. The main pane on the right shows the 'Model output' and 'Evaluation' sections. The 'Model output' section contains tables for 'Ilength', 'Iwidth', and 'Iheight'. The 'Evaluation' section shows the time taken to build the model (0 seconds), evaluation on the test split, and the time taken to test the model on the training split (0 seconds). The 'Summary' section provides a detailed accuracy report.

Model output

| | Ilength | Iwidth | Iheight |
|-----------|---------|--------|---------|
| mean | 1.4694 | 4.2452 | 5.5516 |
| i. dev. | 0.1782 | 0.4712 | 0.5529 |
| light sum | 50 | 50 | 50 |
| precision | 0.1405 | 0.1405 | 0.1405 |

| | Ilength | Iwidth | Iheight |
|-----------|---------|--------|---------|
| mean | 0.2743 | 1.3097 | 2.0343 |
| i. dev. | 0.1096 | 0.1915 | 0.2646 |
| light sum | 50 | 50 | 50 |
| precision | 0.1143 | 0.1143 | 0.1143 |

taken to build model: 0 seconds

Evaluation on test split ===

taken to test model on training split: 0 seconds

Summary ===

| | Correctly Classified Instances | 94.1176 % |
|----------------------------------|--------------------------------|-----------|
| Incorrectly Classified Instances | 3 | 5.8824 % |
| Kappa statistic | 0.9113 | |
| Mean absolute error | 0.0447 | |
| Root mean squared error | 0.1722 | |
| Relative absolute error | 10.0365 % | |
| Root relative squared error | 36.4196 % | |
| Total Number of Instances | 51 | |

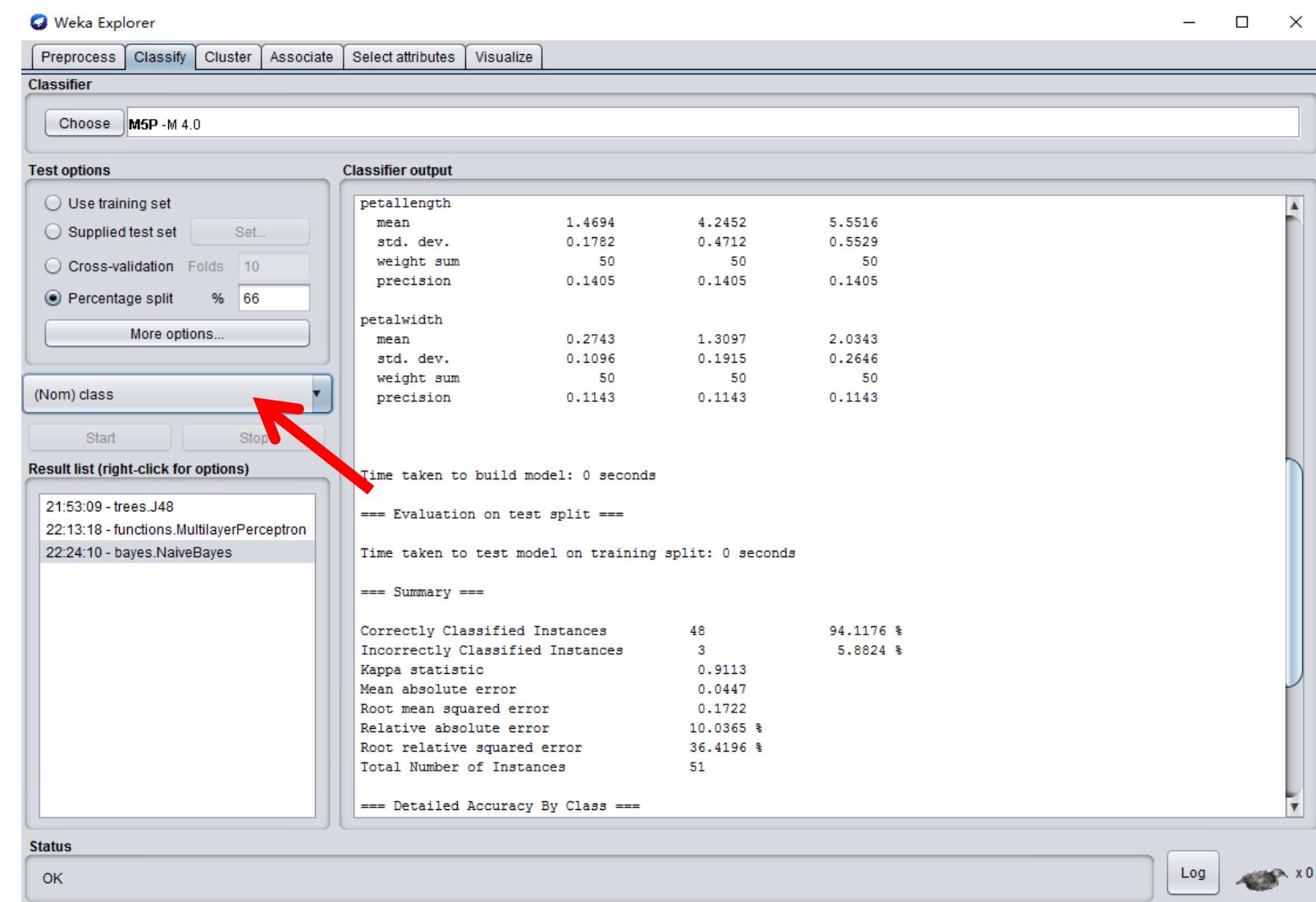
== Detailed Accuracy By Class ==

Status

OK

Log x 0

Explorer: Classification



Explorer: Classification

The screenshot shows the Weka Explorer interface for classification. The top menu bar includes Preprocess, Classify (selected), Cluster, Associate, Select attributes, and Visualize. The main area is titled 'Classifier' and shows 'M5P - M 4.0' selected under 'Choose'. The 'Test options' panel shows 'Percentage split' at 66%. The 'Classifier output' panel displays statistical data for petallength and petalwidth, followed by model build and evaluation logs.

Test options

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

(Nom) class

- (Num) sepallength
- (Num) sepalwidth
- (Num) petallength** (highlighted with a red arrow)
- (Num) petalwidth
- (Nom) class

22:24:10 - bayes.NaiveBayes

Classifier output

| | petallength | petalwidth | |
|------------|-------------|------------|--------|
| mean | 1.4694 | 0.2743 | 4.2452 |
| std. dev. | 0.1782 | 0.1096 | 0.4712 |
| weight sum | 50 | 50 | 50 |
| precision | 0.1405 | 0.1143 | 0.1405 |

Time taken to build model: 0 seconds

==== Evaluation on test split ===

Time taken to test model on training split: 0 seconds

==== Summary ===

| | Correctly Classified Instances | 94.1176 % |
|----------------------------------|--------------------------------|-----------|
| Incorrectly Classified Instances | 3 | 5.8824 % |
| Kappa statistic | 0.9113 | |
| Mean absolute error | 0.0447 | |
| Root mean squared error | 0.1722 | |
| Relative absolute error | 10.0365 % | |
| Root relative squared error | 36.4196 % | |
| Total Number of Instances | 51 | |

==== Detailed Accuracy By Class ===

Status

OK

Explorer: Classification

The screenshot shows the Weka Explorer interface with the following details:

- Weka Explorer** window title.
- Tab Bar:** Preprocess, Classify (selected), Cluster, Associate, Select attributes, Visualize.
- Classifier Panel:** Choose classifier set to **M5P -M 4.0**.
- Test options Panel:** Percentage split (66%) selected. Other options include Use training set, Supplied test set, Cross-validation (Folds 10), and More options... button.
- Result list Panel:** Shows log entries:
 - 21:53:09 - trees.J48
 - 22:13:18 - functions.MultilayerPerceptron
 - 22:24:10 - bayes.NaiveBayes
 - 22:33:47 - trees.M5P (highlighted)
- Classifier output Panel:** Displays run information and classifier model details.

```
=== Run information ===  
Scheme: weka.classifiers.trees.M5P -M 4.0  
Relation: iris  
Instances: 150  
Attributes: 5  
sepallength  
sepalwidth  
petallength  
petalwidth  
class  
Test mode: split 66.0% train, remainder test  
  
=== Classifier model (full training set) ===  
  
M5 pruned model tree:  
(using smoothed linear models)  
  
petalwidth <= 0.8 : LM1 (50/9.298%)  
petalwidth > 0.8 :  
| class=Iris-virginica <= 0.5 : LM2 (50/12.723%)  
| class=Iris-virginica > 0.5 : LM3 (50/15.631%)  
  
LM num: 1  
petallength =  
    0.1685 * sepallength  
    - 0.1503 * sepalwidth  
    + 0.715 * petalwidth  
    + 0.9748  
  
LM num: 2  
petallength =
```

A red arrow points to the scroll bar of this panel.
- Status Panel:** OK.
- Log Panel:** Shows a small icon and count of 0.

Explorer: Classification

The screenshot shows the Weka Explorer interface with the 'Classifier' tab selected. The 'Choose' dropdown is set to 'M5P - M 4.0'. In the 'Test options' panel, the 'Percentage split' option is selected with 66% chosen. The 'Classifier output' panel displays the generated regression equation and other evaluation details.

Test options

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

(Num) petallength

Result list (right-click for options)

- 21:53:09 - trees.J48
- 22:13:18 - functions.MultilayerPerceptron
- 22:24:10 - bayes.NaiveBayes
- 22:33:47 - trees.M5P**

Classifier output

```
0.5075 * sepallength  
- 0.085 * sepalwidth  
+ 1.1314 * petalwidth  
+ 0.1083 * class=Iris-virginica  
- 0.0257

LM num: 3
petallength =
    0.7278 * sepallength
    - 0.085 * sepalwidth
    + 0.2824 * petalwidth
    + 0.1083 * class=Iris-virginica
    + 0.3295

Number of Rules : 3

Time taken to build model: 0.1 seconds

==== Evaluation on test split ===

Time taken to test model on training split: 0 seconds

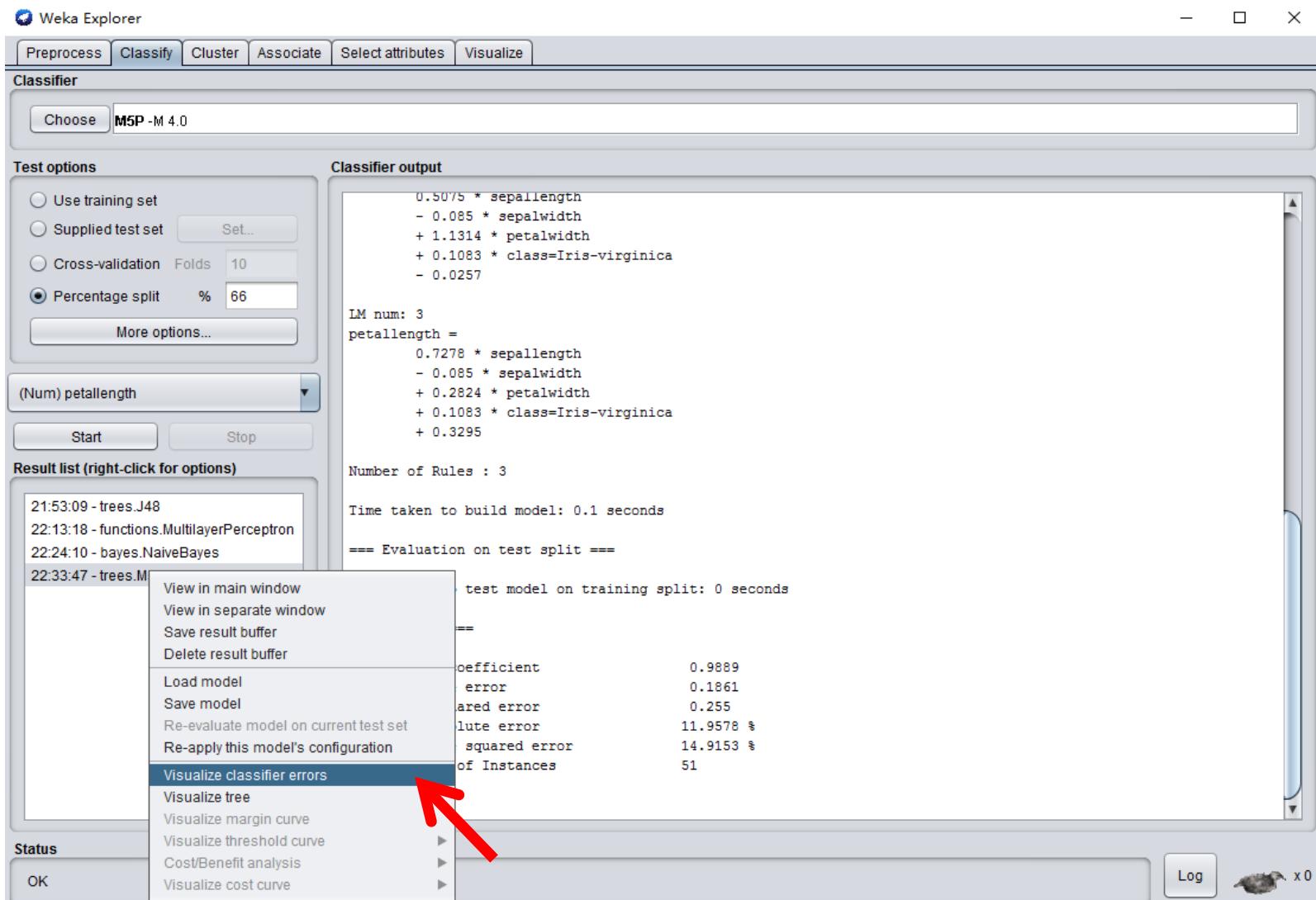
==== Summary ===

Correlation coefficient          0.9889
Mean absolute error             0.1861
Root mean squared error         0.255
Relative absolute error          11.9578 %
Root relative squared error     14.9153 %
Total Number of Instances       51
```

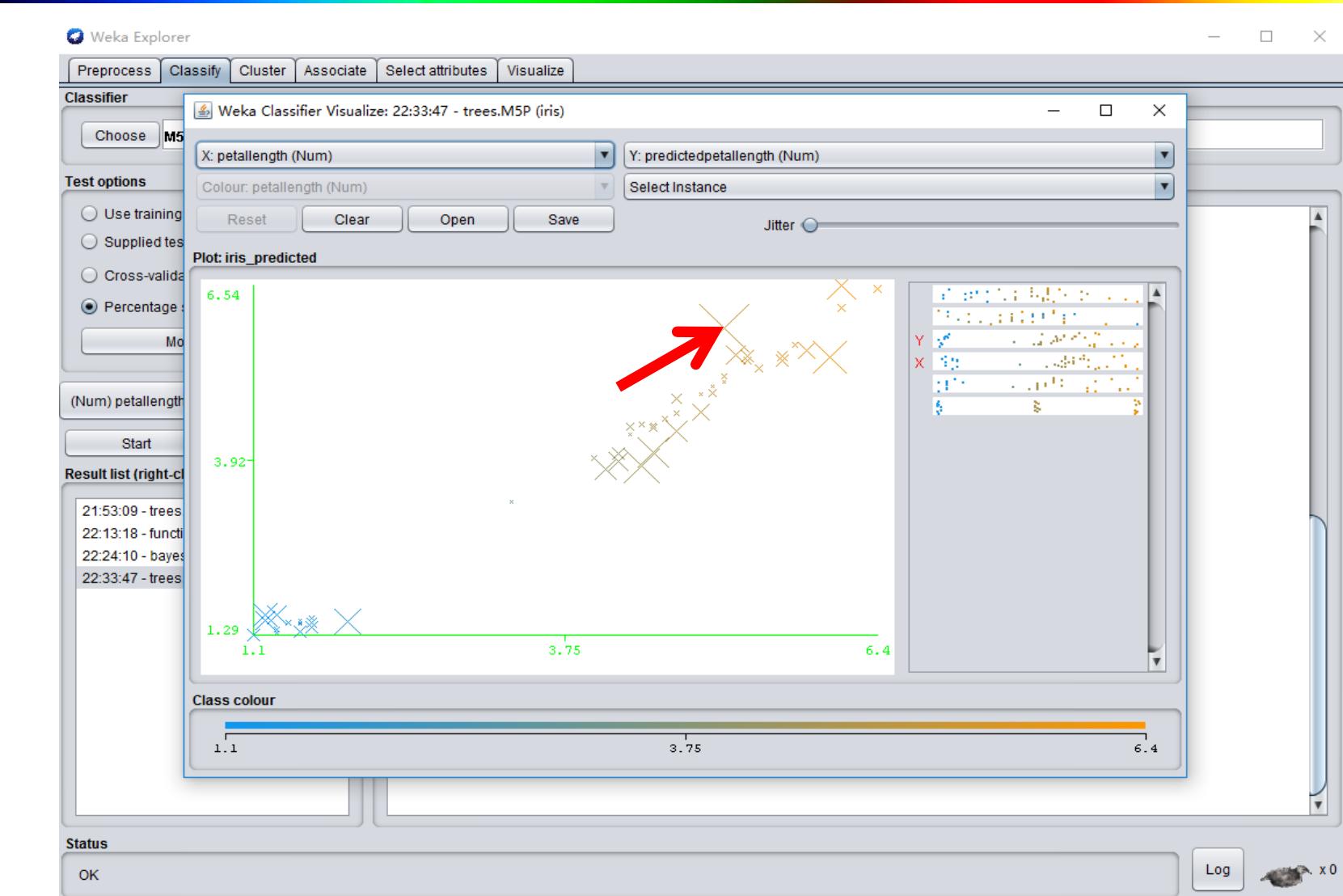
Status

OK

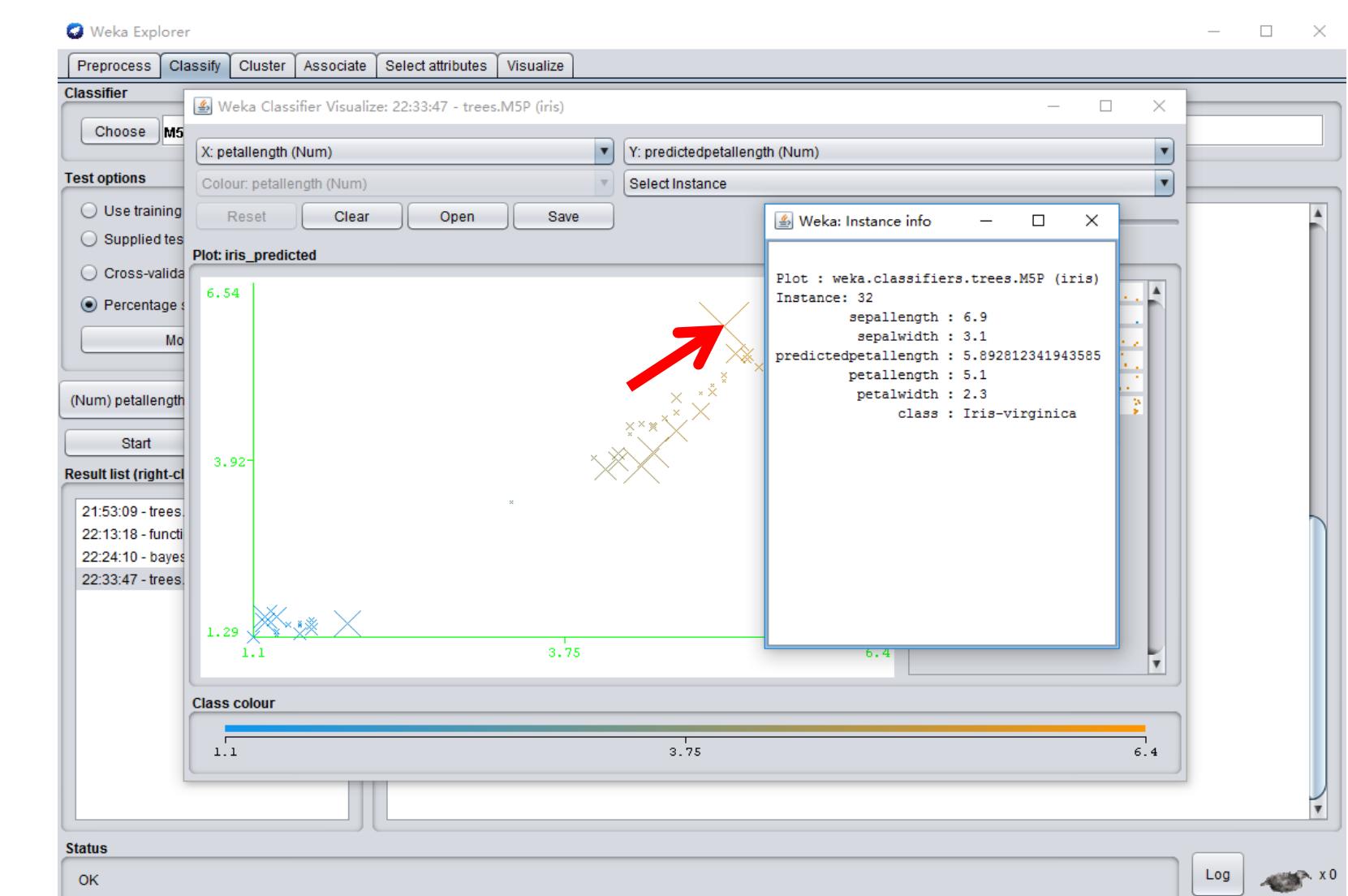
Explorer: Classification



Explorer: Classification



Explorer: Classification



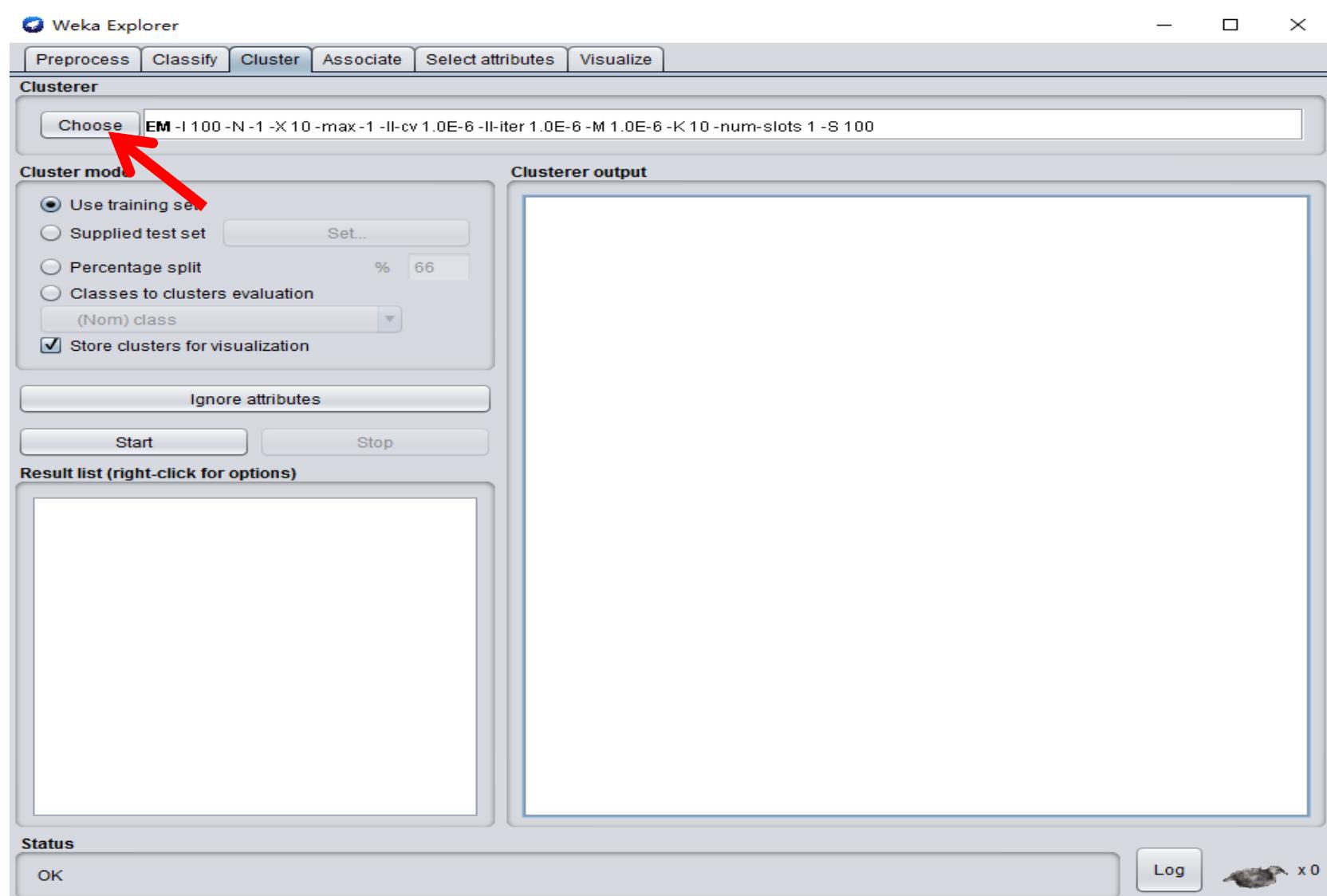
Roadmap: WEKA Usage

- WEKA Usage (version 3.8)
 - Explorer
 - Data preprocessing
 - Classification
 - Clustering
 - Association rule
 - Attribute selection (Feature selection)
 - Data visualization
 - Experimenter
 - Knowledge Flow

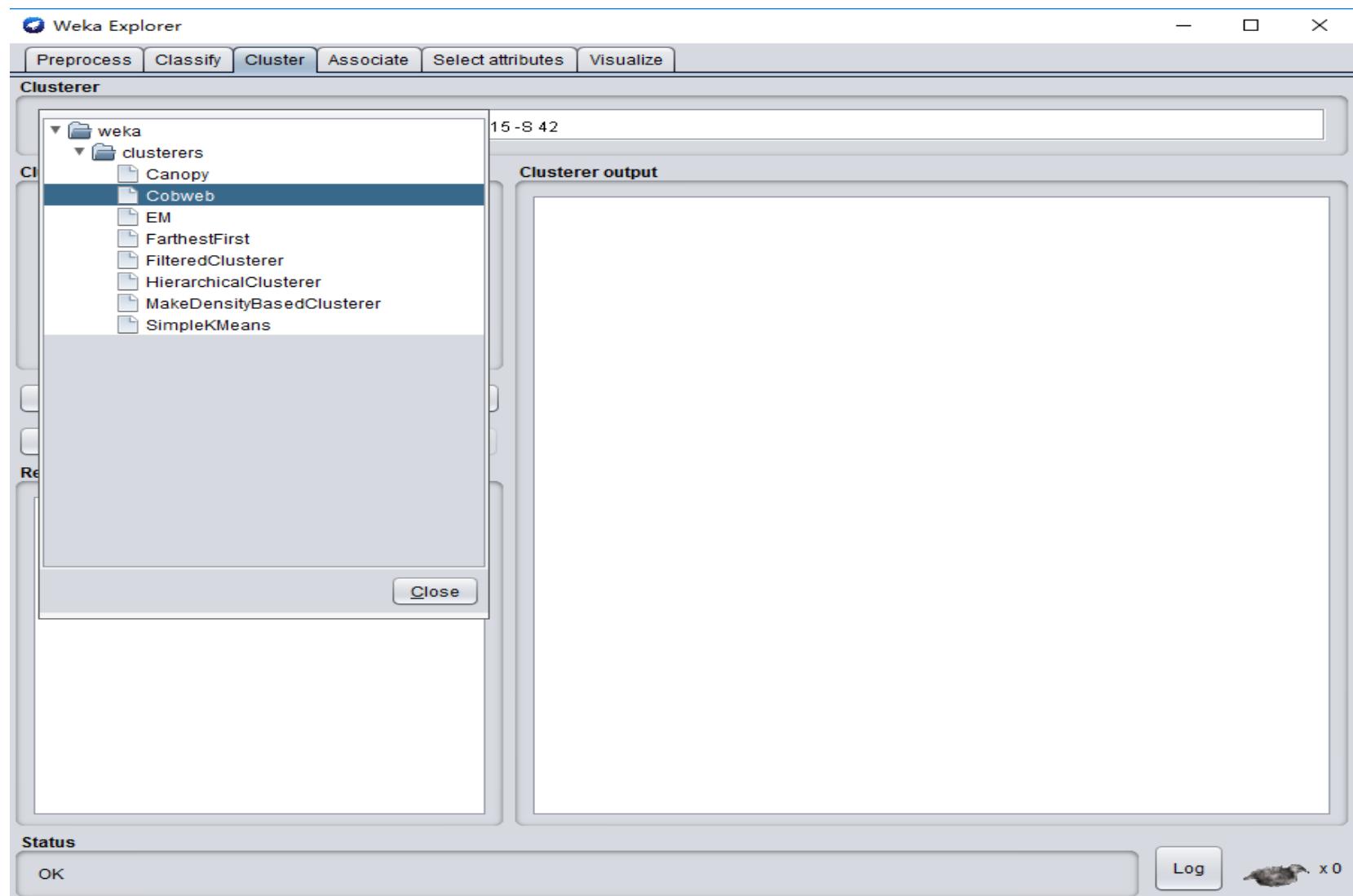
Explorer: Clustering Data

- WEKA contains “clusters” for finding groups of similar instances in a dataset
- Some 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

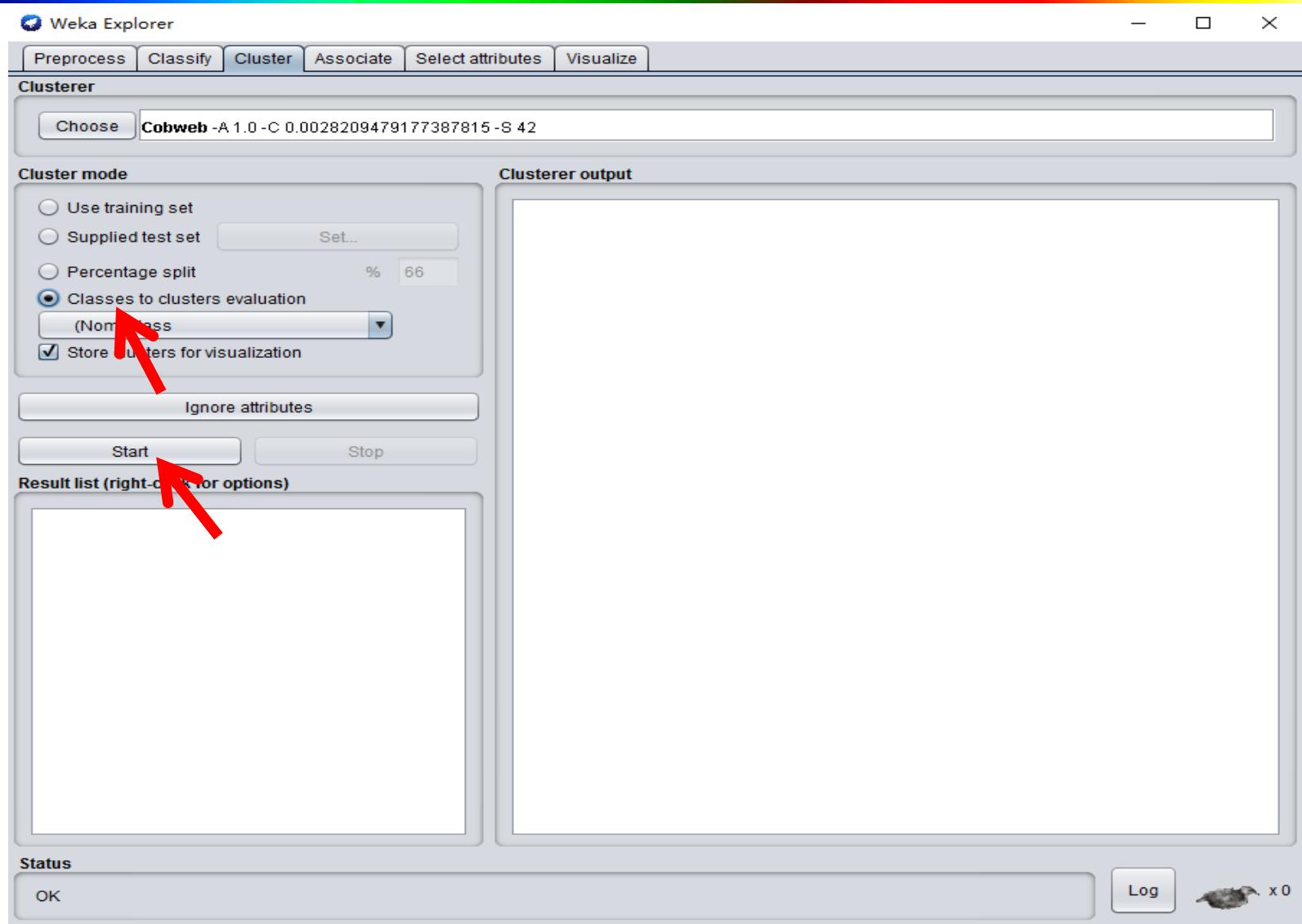
Explorer: Clustering



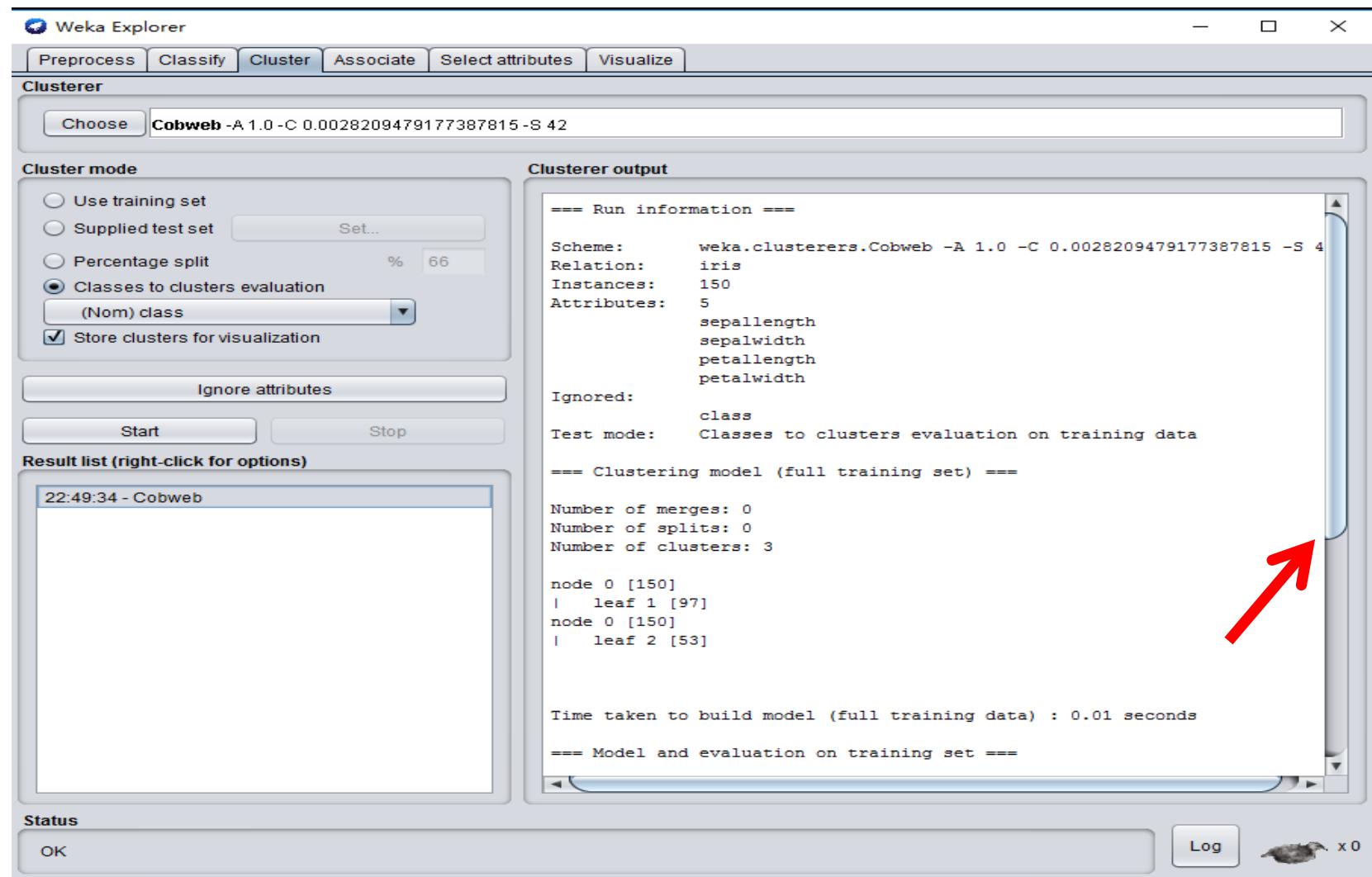
Explorer: Clustering



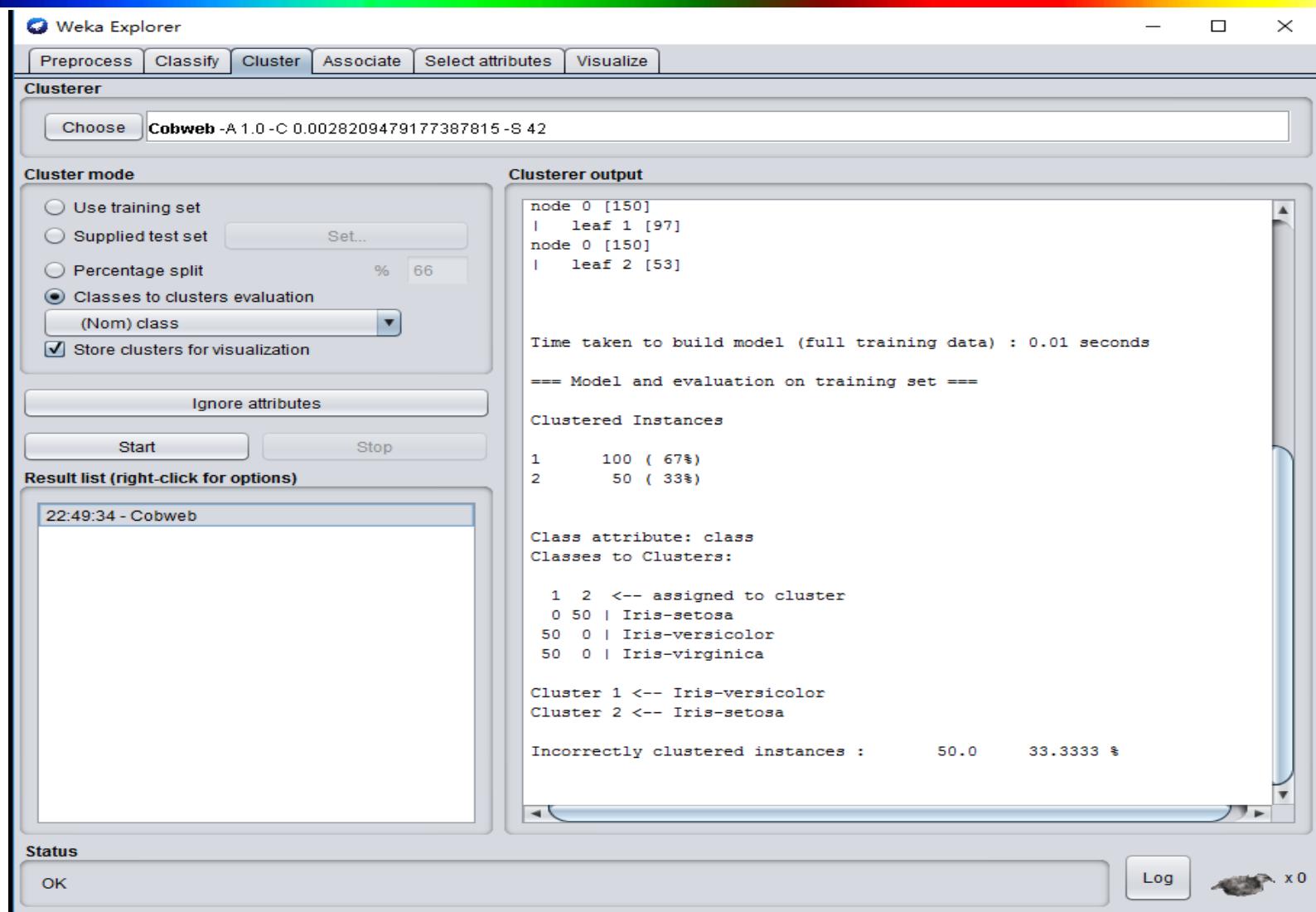
Explorer: Clustering



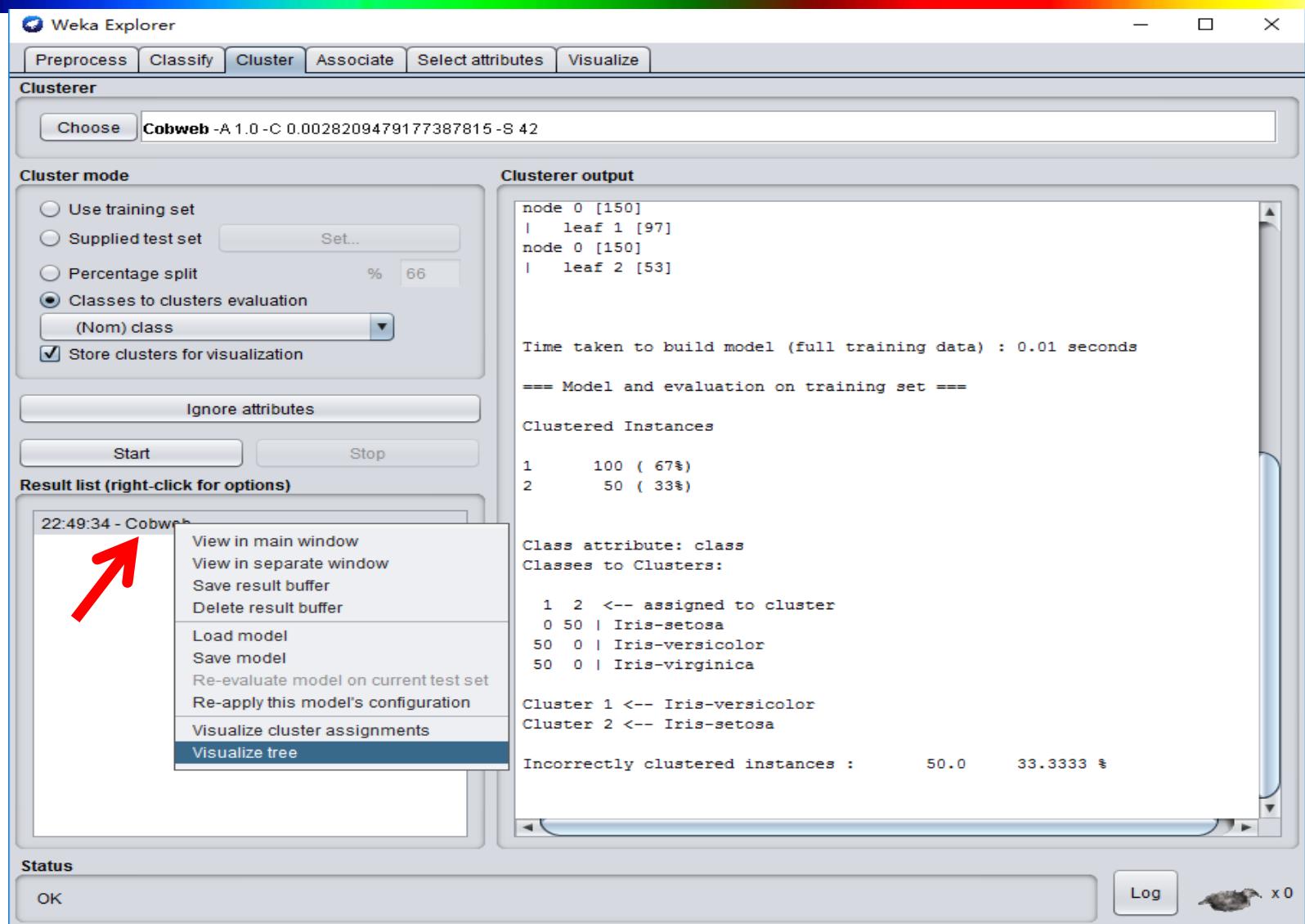
Explorer: Clustering



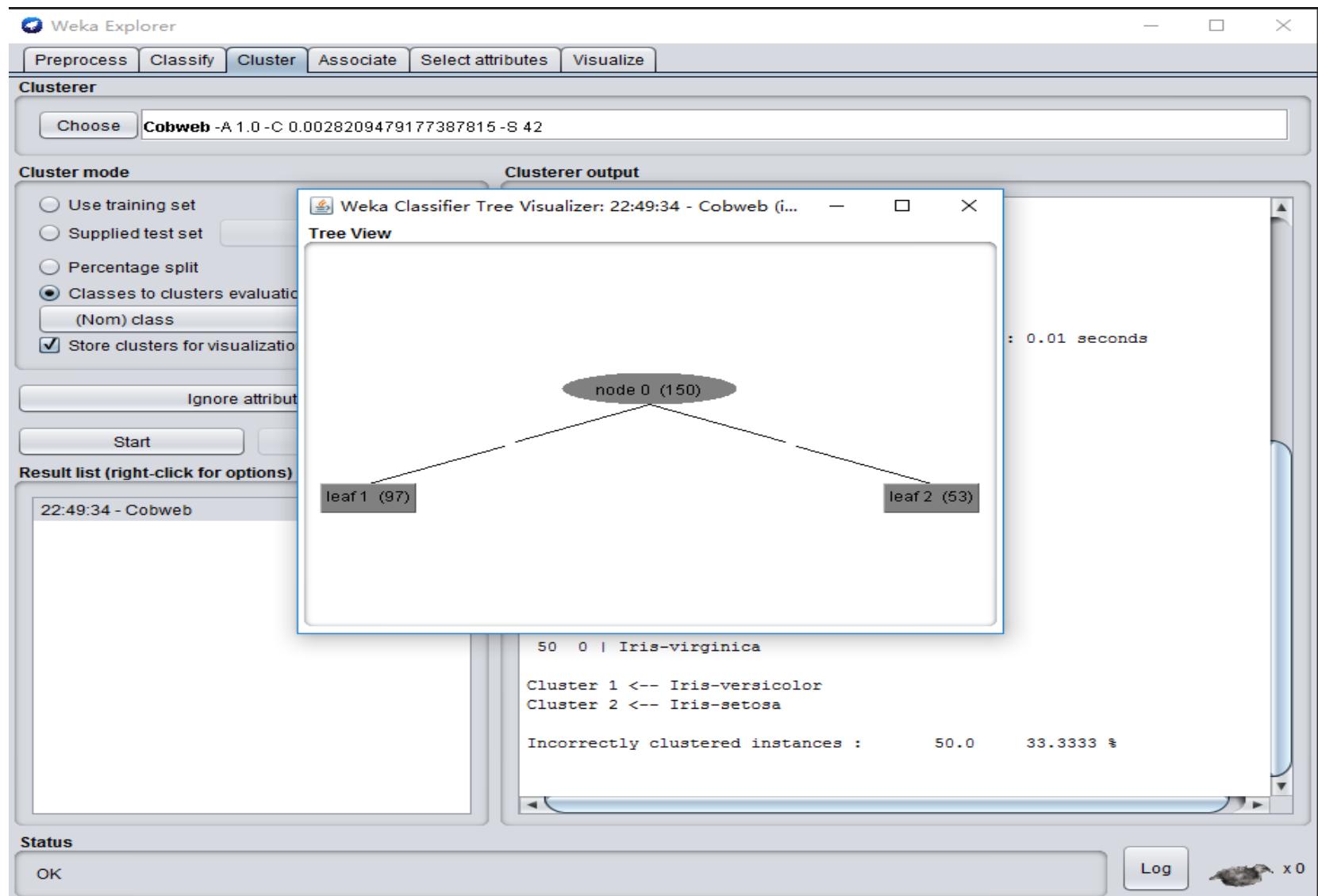
Explorer: Clustering



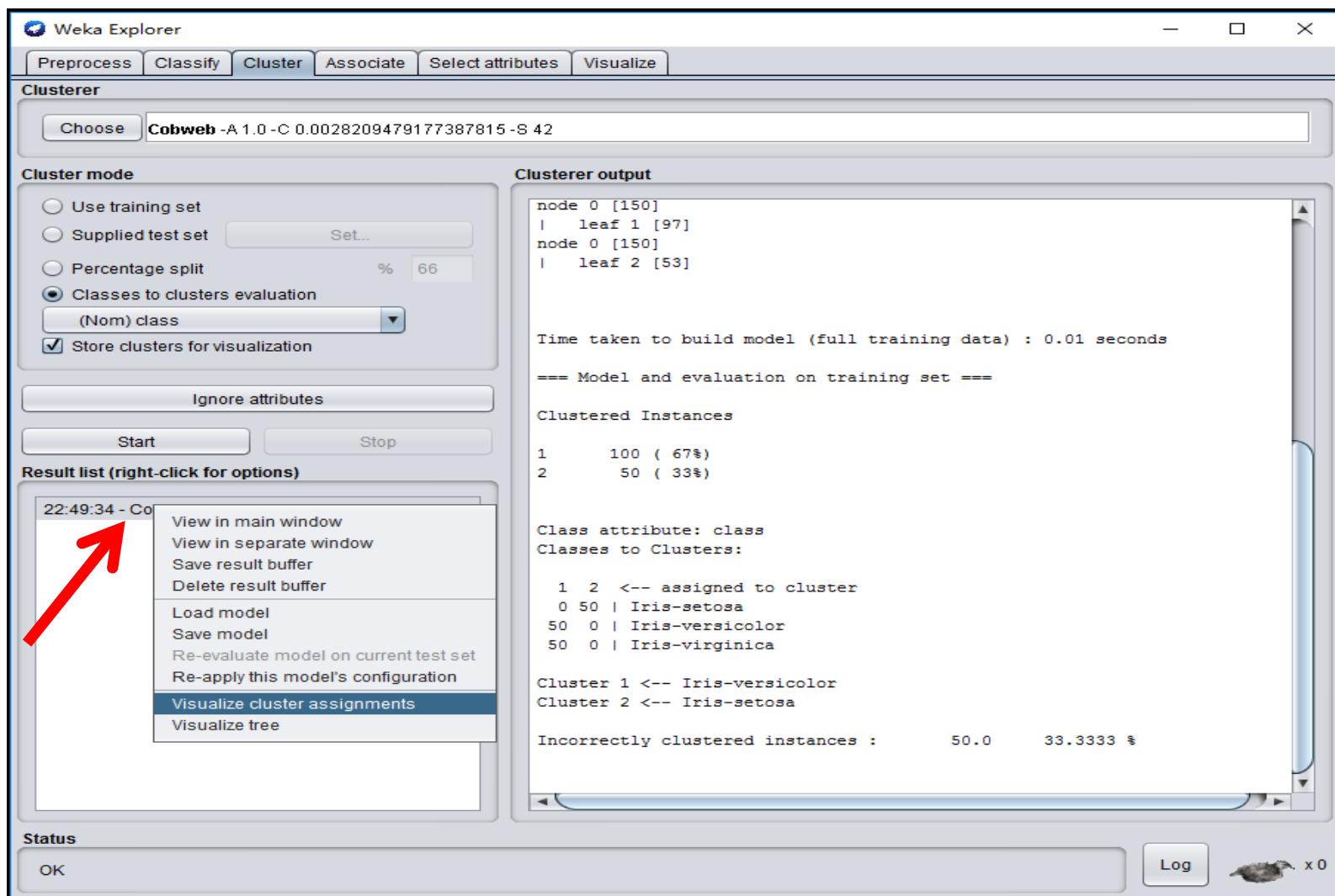
Explorer: Clustering



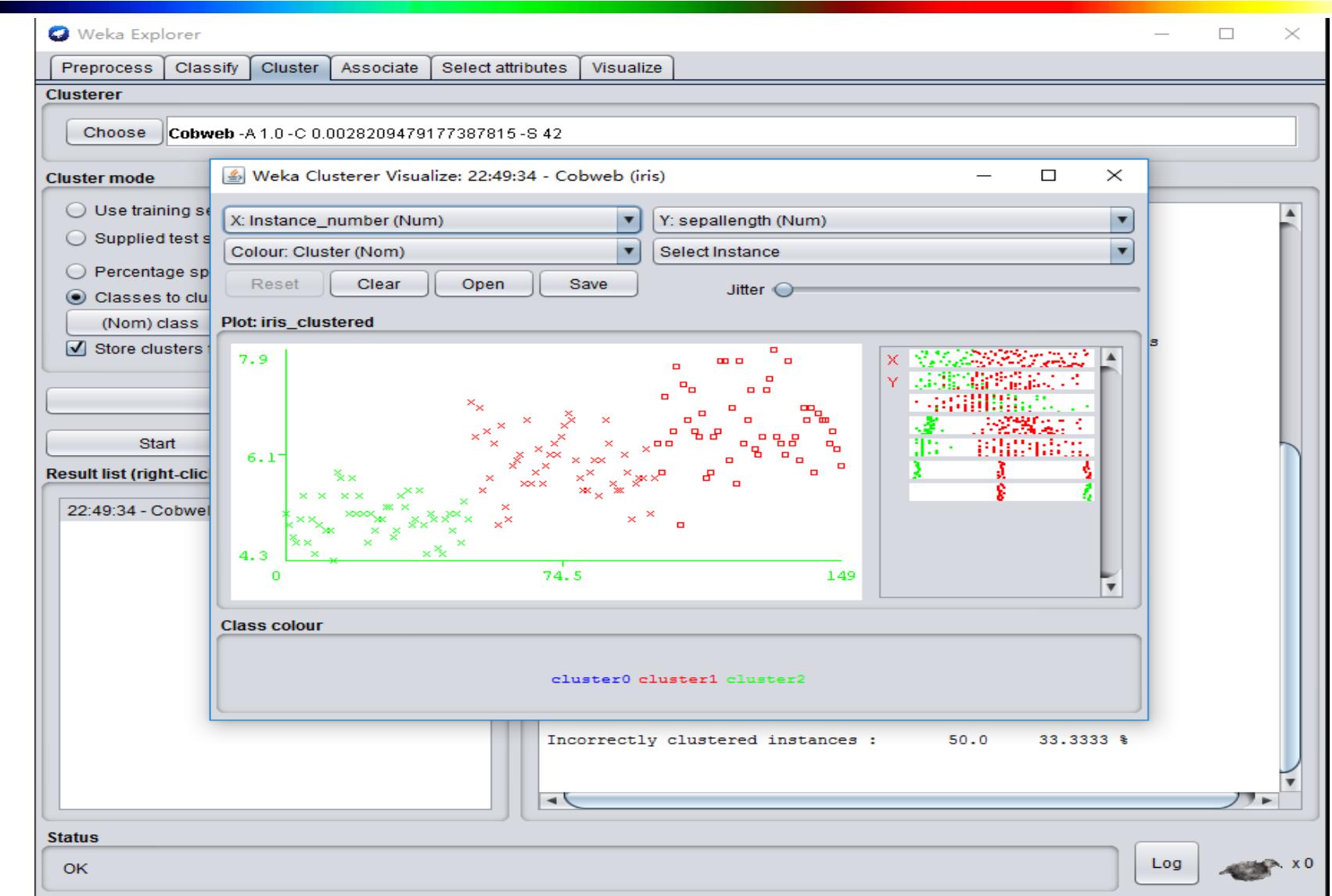
Explorer: Clustering



Explorer: Clustering



Explorer: Clustering



Roadmap: WEKA Usage

- WEKA Usage (version 3.8)
 - Explorer
 - Data preprocessing
 - Classification
 - Clustering
 - Association rule
 - Attribute selection (Feature selection)
 - Data visualization
 - Experimenter
 - Knowledge Flow

Explorer: Finding Associations

- WEKA contains the Apriori algorithm (among others) 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

Explorer: Association Rule

Weka Explorer

Preprocess Classify Cluster Associate **Select attributes** Visualize

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

Filter

Choose None Apply

Current relation

Relation: vote Instances: 435 Attributes: 17 Sum of weights: 435

Selected attribute

Name: handicapped-infants Type: Nominal
Missing: 12 (3%) Distinct: 2 Unique: 0 (0%)

| No. | Label | Count | Weight |
|-----|-------|-------|--------|
| 1 | n | 236 | 236.0 |
| 2 | y | 187 | 187.0 |

Attributes

All None Invert Pattern

| No. | Name |
|-----|-----------------------------------|
| 1 | handicapped-infants |
| 2 | water-project-cost-sharing |
| 3 | adoption-of-the-budget-resolution |
| 4 | physician-fee-freeze |
| 5 | el-salvador-aid |
| 6 | religious-groups-in-schools |
| 7 | anti-satellite-test-ban |

Remove

Status

OK Log x 0

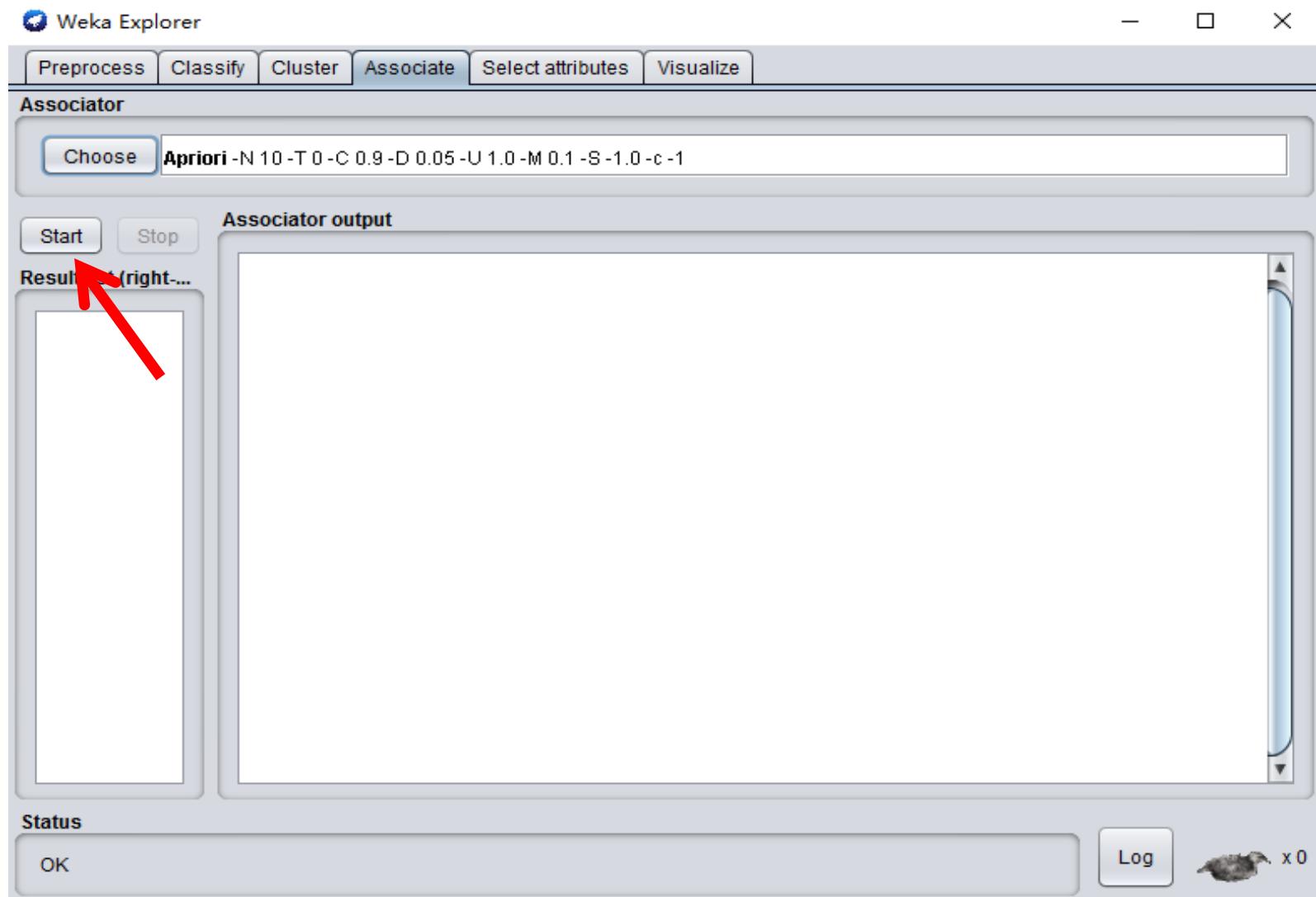
Class: Class (Nom) Visualize All

236

187

Figure showing two horizontal bar charts. The left chart represents the count of 'n' (236) and the right chart represents the count of 'y' (187). The bars are composed of red and blue segments.

Explorer: Association Rule



Explorer: Association Rule

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 -c -1**

Start Stop

Result list (right-click to copy)

22:49:12 - Apriori

Associator output

```
Minimum support: 0.45 (196 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 11

Generated sets of large itemsets:

Size of set of large itemsets L(1): 20
Size of set of large itemsets L(2): 17
Size of set of large itemsets L(3): 6
Size of set of large itemsets L(4): 1

Best rules found:

1. adoption-of-the-budget-resolution=n physician-fee-freeze=n 219 ==> Class=democrat 219
2. adoption-of-the-budget-resolution=n physician-fee-freeze=n aid-to-nicaraguan-contras=y 1
3. physician-fee-freeze=n aid-to-nicaraguan-contras=y 211 ==> Class=democrat 210    <conf:(0.99) lift:(1.62) lev:(0.21)
4. physician-fee-freeze=n education-spending=n 202 ==> Class=democrat 201    <conf:(1) lift:(1.62) lev:(0.21)
5. physician-fee-freeze=n 247 ==> Class=democrat 245    <conf:(0.99) lift:(1.62) lev:(0.21)
6. el-salvador-aid=n Class=democrat 200 ==> aid-to-nicaraguan-contras=y 197    <conf:(0.98)
7. el-salvador-aid=n 208 ==> aid-to-nicaraguan-contras=y 204    <conf:(0.98) lift:(1.76) 1
8. adoption-of-the-budget-resolution=n aid-to-nicaraguan-contras=y Class=democrat 203 ==> p
9. el-salvador-aid=n aid-to-nicaraguan-contras=y 204 ==> Class=democrat 197    <conf:(0.97)
10. aid-to-nicaraguan-contras=y Class=democrat 218 ==> physician-fee-freeze=n 210    <conf:(0.97)
```

Status

OK Log x0

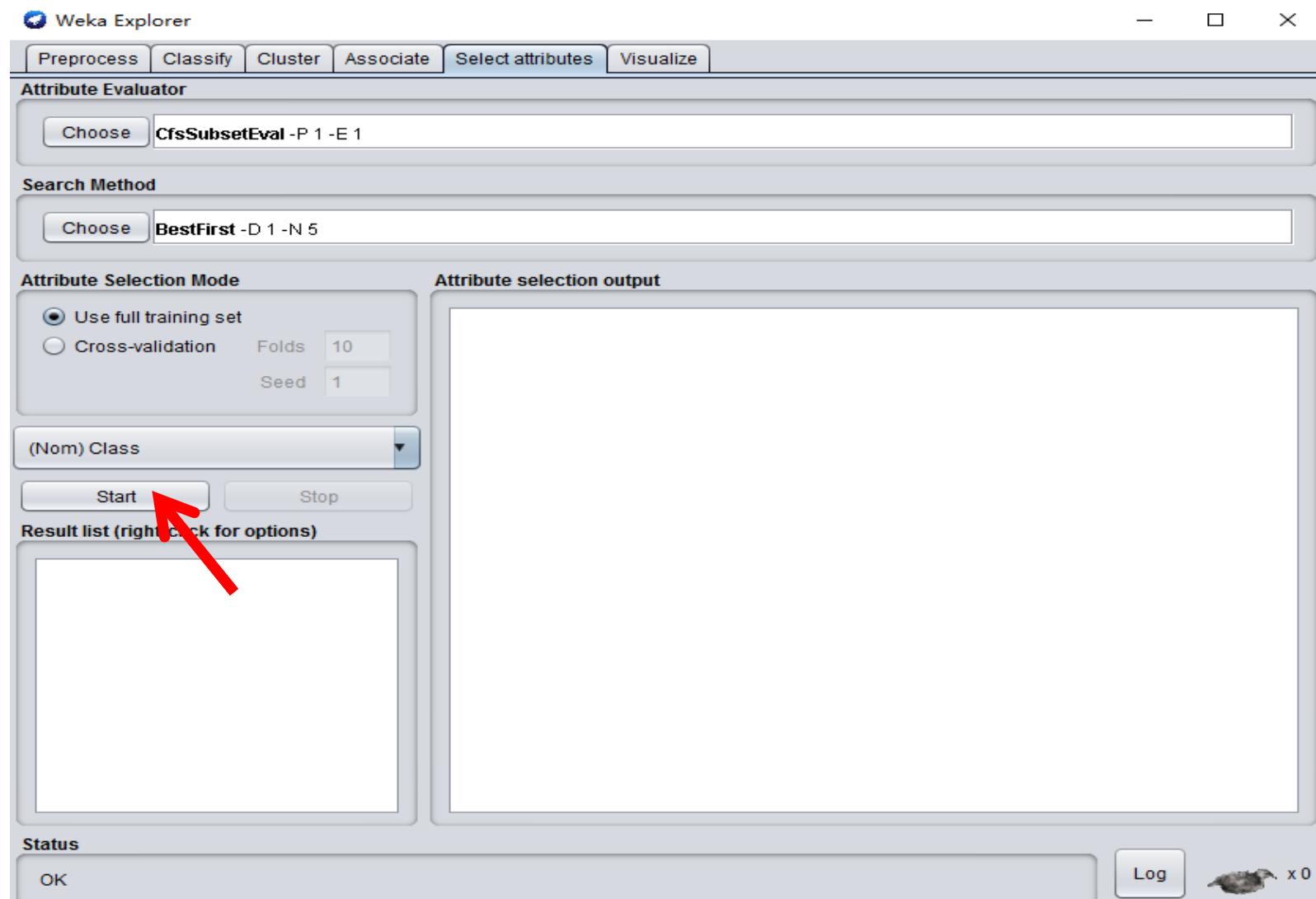
Roadmap: WEKA Usage

- WEKA Usage (version 3.8)
 - Explorer
 - Data preprocessing
 - Classification
 - Clustering
 - Association rule
 - Attribute selection (Feature selection)
 - Data visualization
 - Experimenter
 - Knowledge Flow

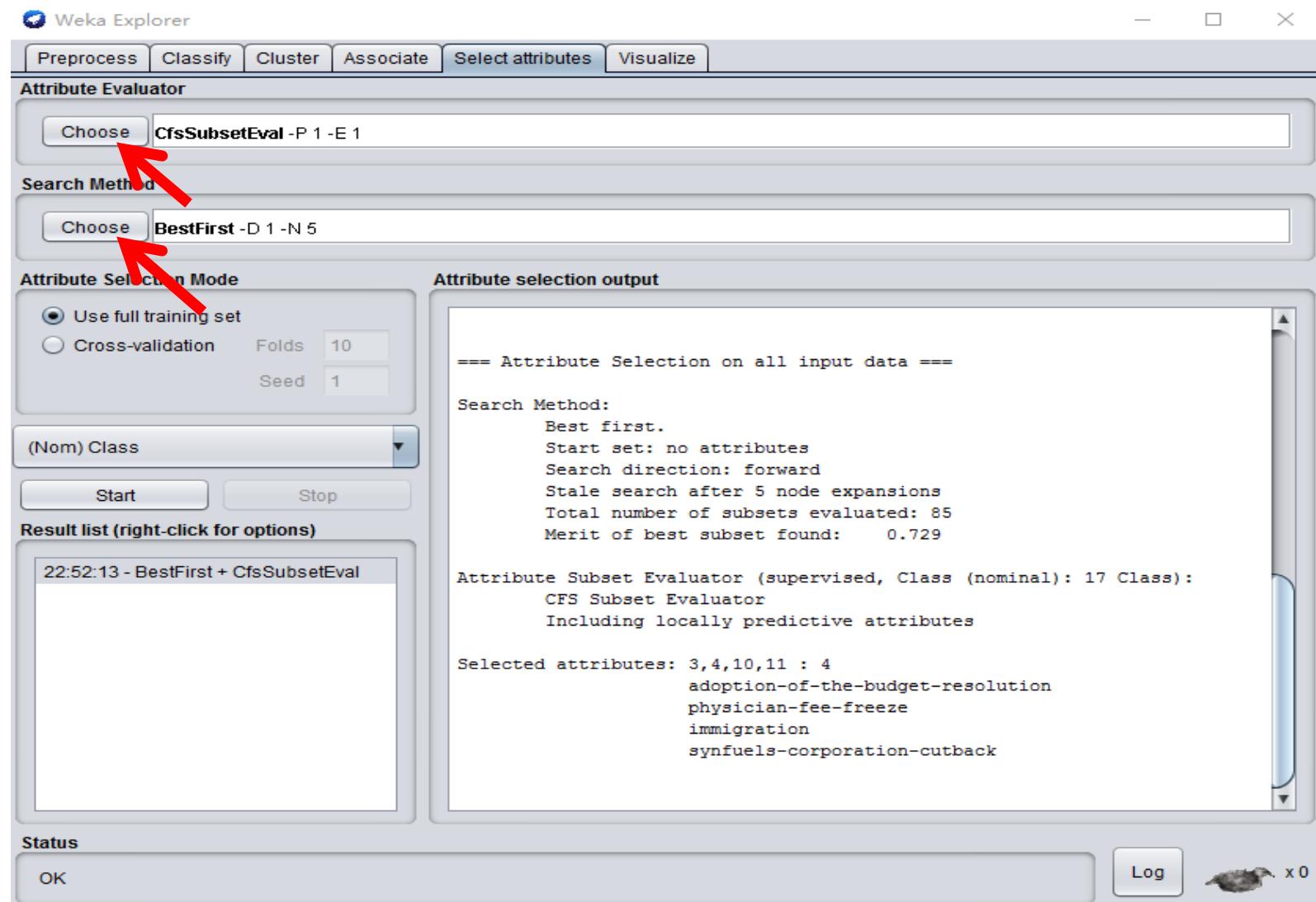
Explorer: Attribute Selection

- Panel that can be used to investigate which (subsets of) attributes are the most predictive ones
- Attribute selection methods contain two parts:
 - A [search](#) method: best-first, forward selection, random, exhaustive, genetic algorithm, ranking
 - An [evaluation](#) method: correlation-based, wrapper, information gain, chi-square, ...
- Very flexible: WEKA allows (almost) arbitrary combinations of these two

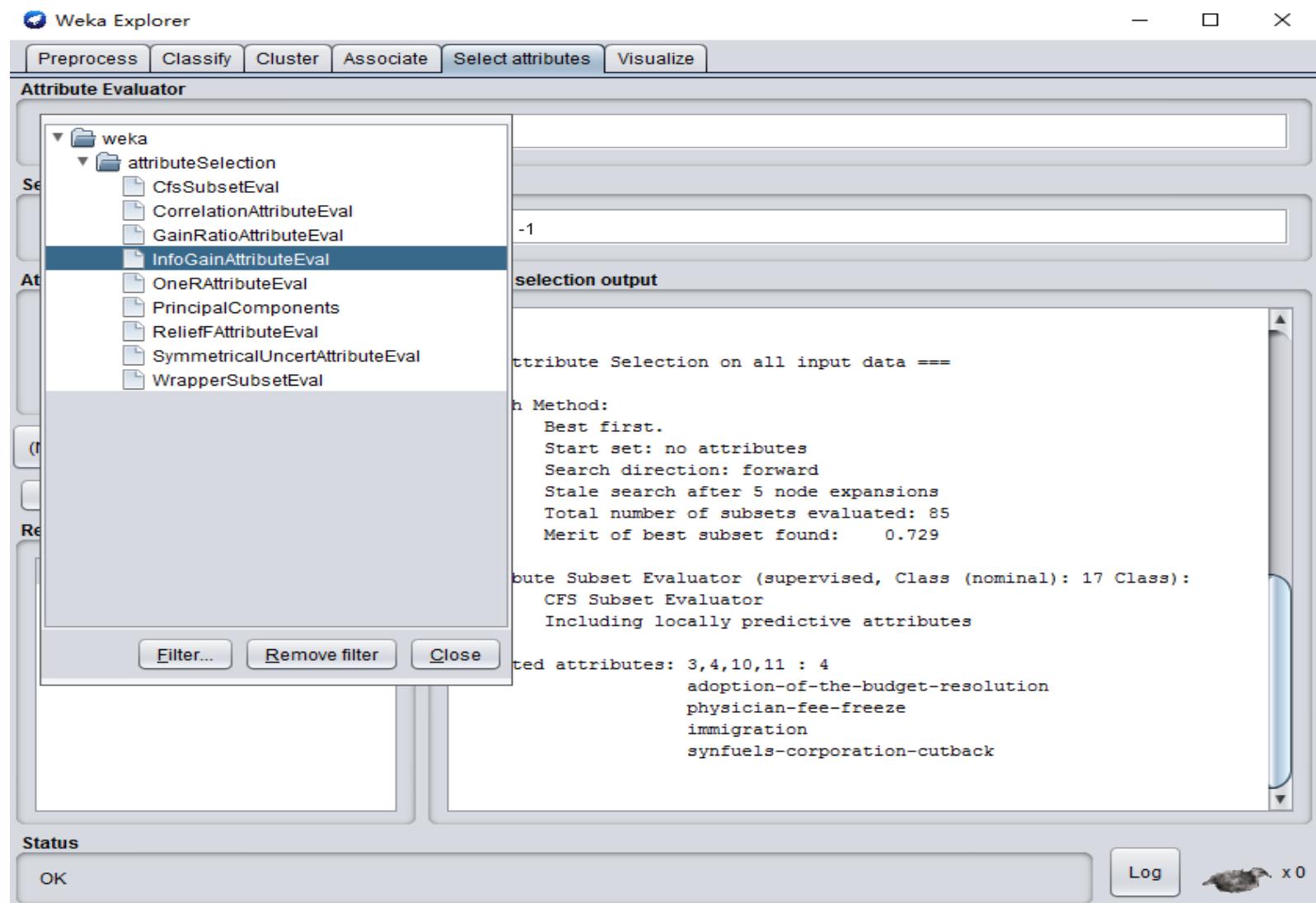
Explorer: Attribute Selection



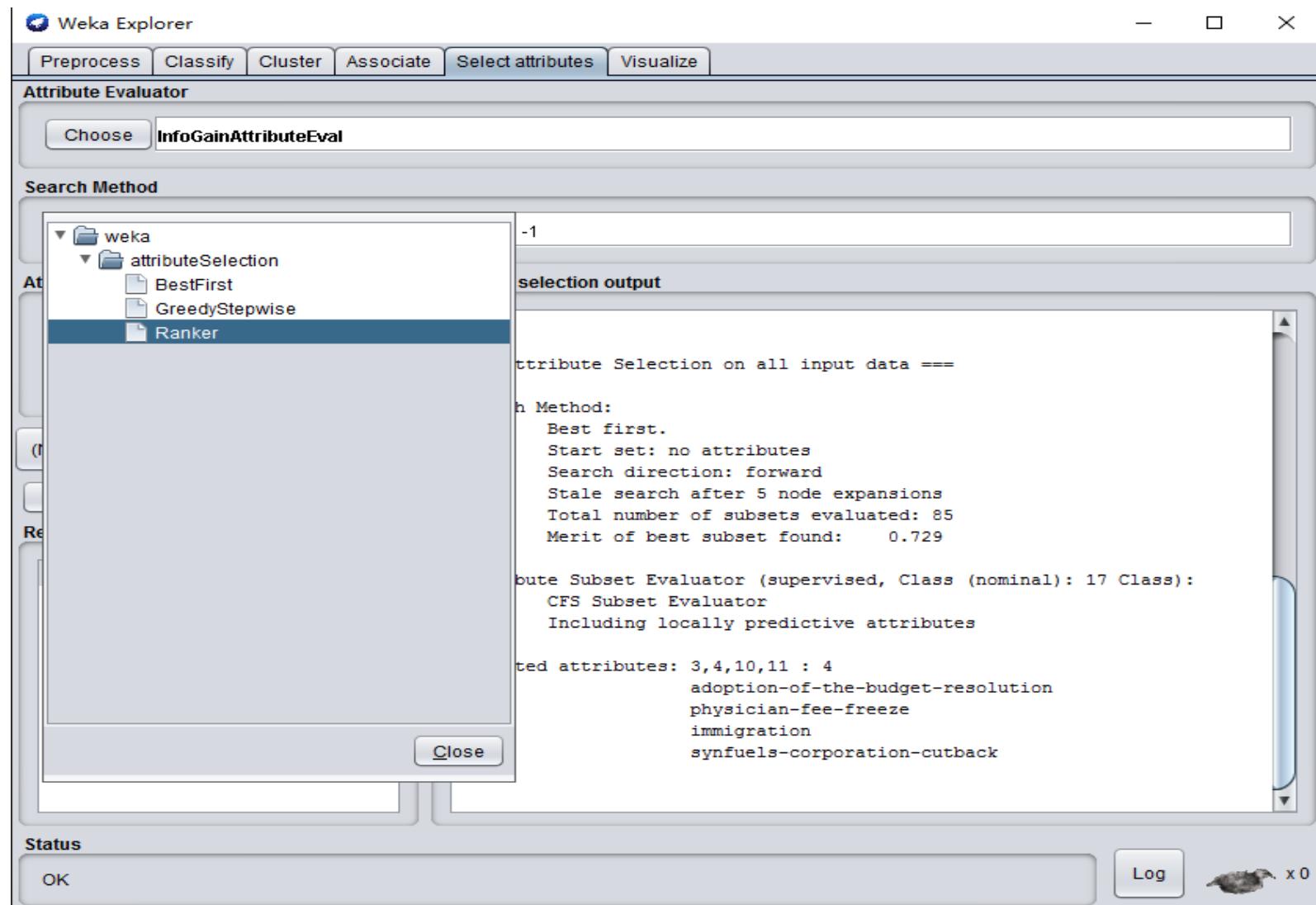
Explorer: Attribute Selection



Explorer: Attribute Selection



Explorer: Attribute Selection



Explorer: Attribute Selection

The screenshot shows the Weka Explorer interface with the following configuration:

- Attribute Evaluator:** Set to **InfoGainAttributeEval**.
- Search Method:** Set to **Ranker -T -1.7976931348623157E308 -N -1**.
- Attribute Selection Mode:** Set to **Use full training set**.
- (Nom) Class:** A dropdown menu.
- Result list (right-click for options):** Displays two entries:
 - 22:52:13 - BestFirst + CfsSubsetEval
 - 22:59:27 - Ranker + InfoGainAttributeEval
- Status:** Shows "OK".
- Log:** Shows "x 0".

Attribute selection output:

```
Search Method:
    Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 17 Class):
    Information Gain Ranking Filter

Ranked attributes:
0.7078541 4 physician-fee-freeze
0.4185726 3 adoption-of-the-budget-resolution
0.4028397 5 el-salvador-aid
0.34036 12 education-spending
0.3123121 14 crime
0.3095576 8 aid-to-nicaraguan-contras
0.2856444 9 mx-missile
0.2121705 13 superfund-right-to-sue
0.2013666 15 duty-free-exports
0.1902427 7 anti-satellite-test-ban
0.1404643 6 religious-groups-in-schools
0.1211834 1 handicapped-infants
0.1007458 11 synfuels-corporation-cutback
0.0529956 16 export-administration-act-south-africa
0.0049097 10 immigration
0.0000117 2 water-project-cost-sharing

Selected attributes: 4,3,5,12,14,8,9,13,15,7,6,1,11,16,10,2 : 16
```

Roadmap: WEKA Usage

- WEKA Usage (version 3.8)
 - Explorer
 - Data preprocessing
 - Classification
 - Clustering
 - Association rule
 - Attribute selection (Feature selection)
 - Data visualization
 - Experimenter
 - Knowledge Flow

Explorer: Data Visualization

- So how do visualization helps in mining ?
 - Displayed are scatter plots.
 - It helps answer the questions regarding the involved variables X & Y such as
 - Related?
 - Positively correlated?
 - Negatively correlated?
 - Does Y's variation depends X's variation?
 - Are there outliers ?

Explorer: Data Visualization

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

Explorer: Data Visualization

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

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

Filter Choose None Apply

Current relation
Relation: Glass Instances: 214 Attributes: 10 Sum of weights: 214

Attributes All None Invert Pattern

| No. | Name |
|-----|------|
| 1 | RI |
| 2 | Na |
| 3 | Mg |
| 4 | Al |
| 5 | Si |
| 6 | K |
| 7 | Ca |
| 8 | Ba |
| 9 | Fe |
| 10 | Type |

Remove

Status OK Log x 0

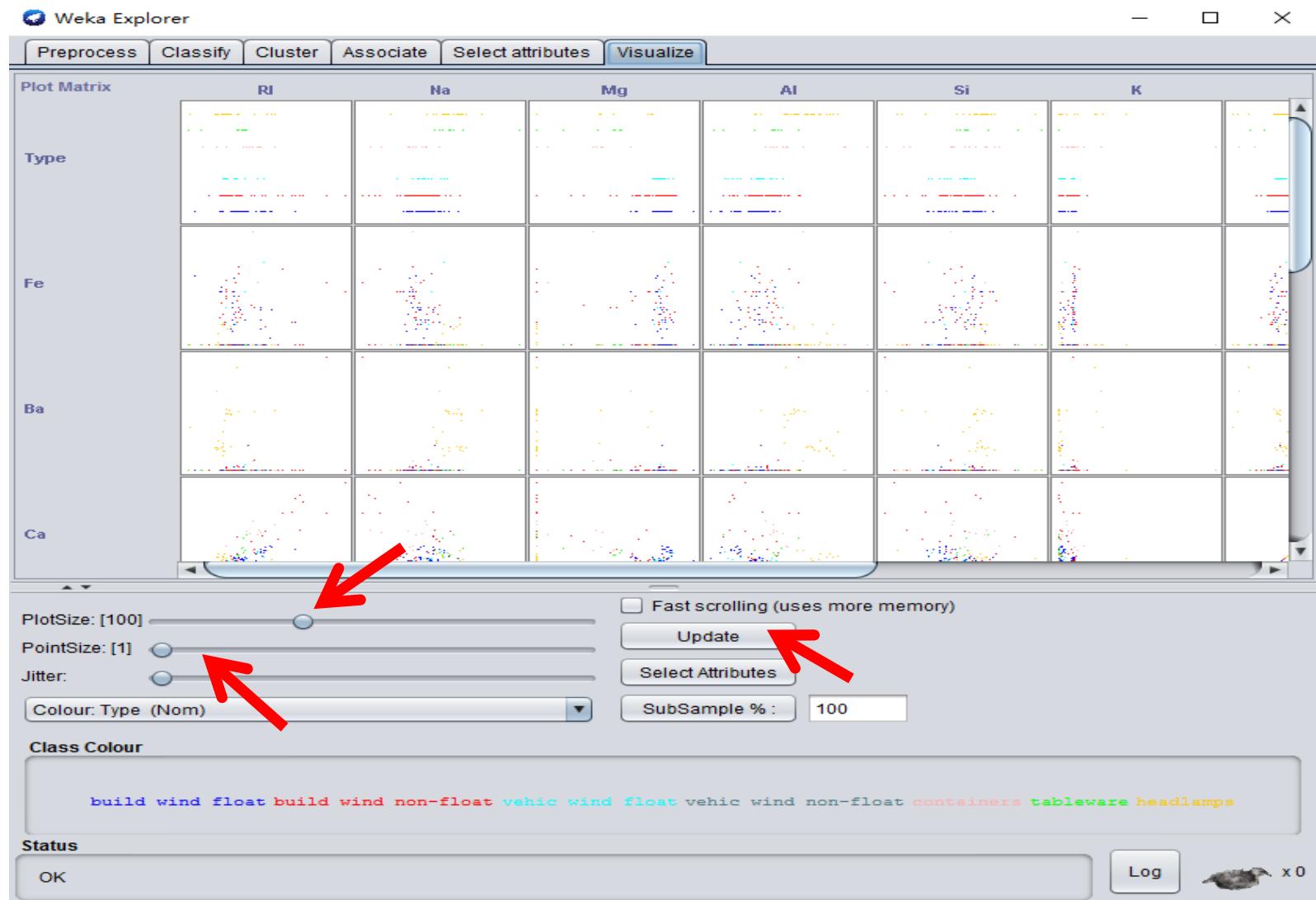
Selected attribute
Name: RI Missing: 0 (0%) Distinct: 178 Type: Numeric Unique: 145 (68%)

| Statistic | Value |
|-----------|-------|
| Minimum | 1.511 |
| Maximum | 1.534 |
| Mean | 1.518 |
| StdDev | 0.003 |

Class: Type (Nom) Visualize All

84
39
39
3
4
16
17
4
3
3
0
1
1

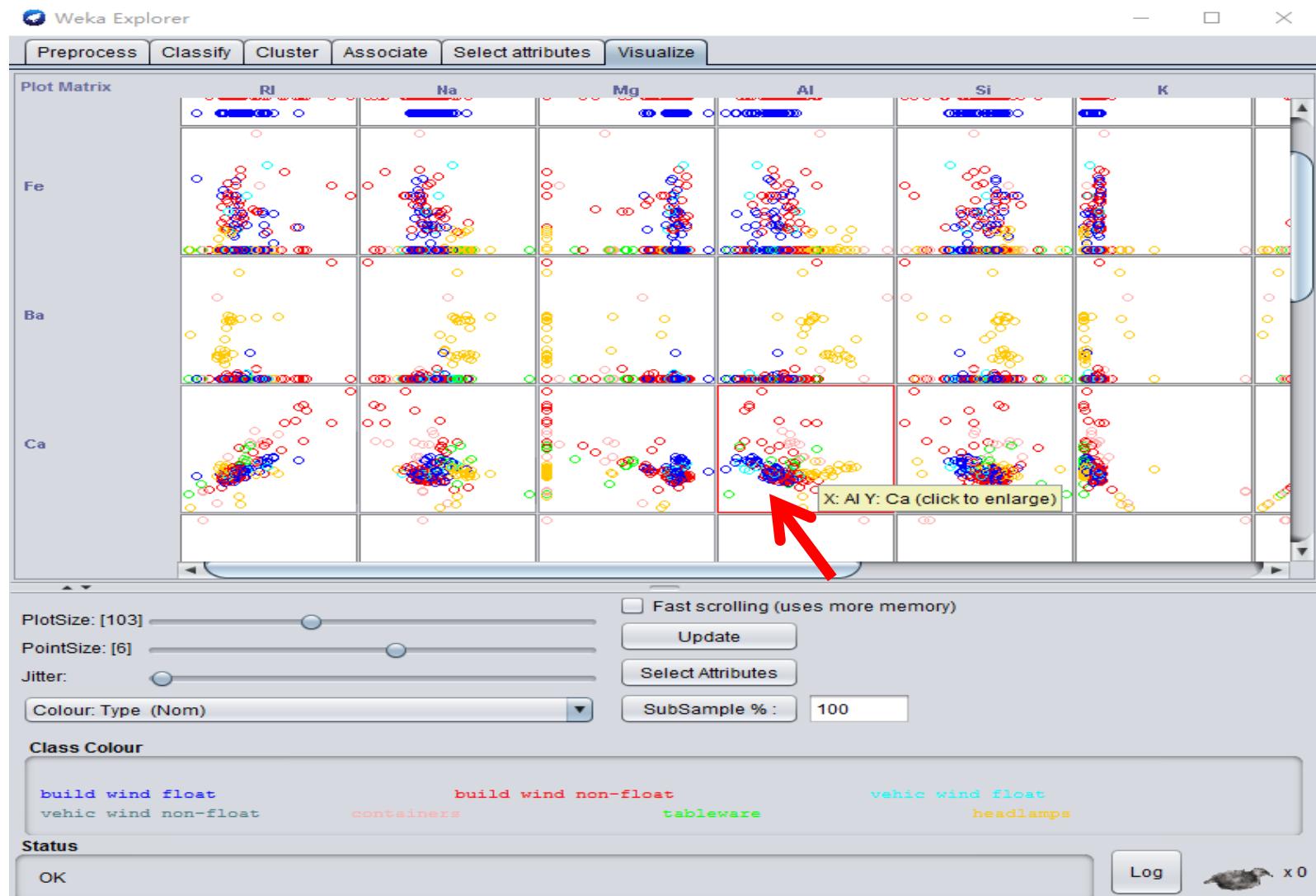
Explorer: Data Visualization



Explorer: Data Visualization



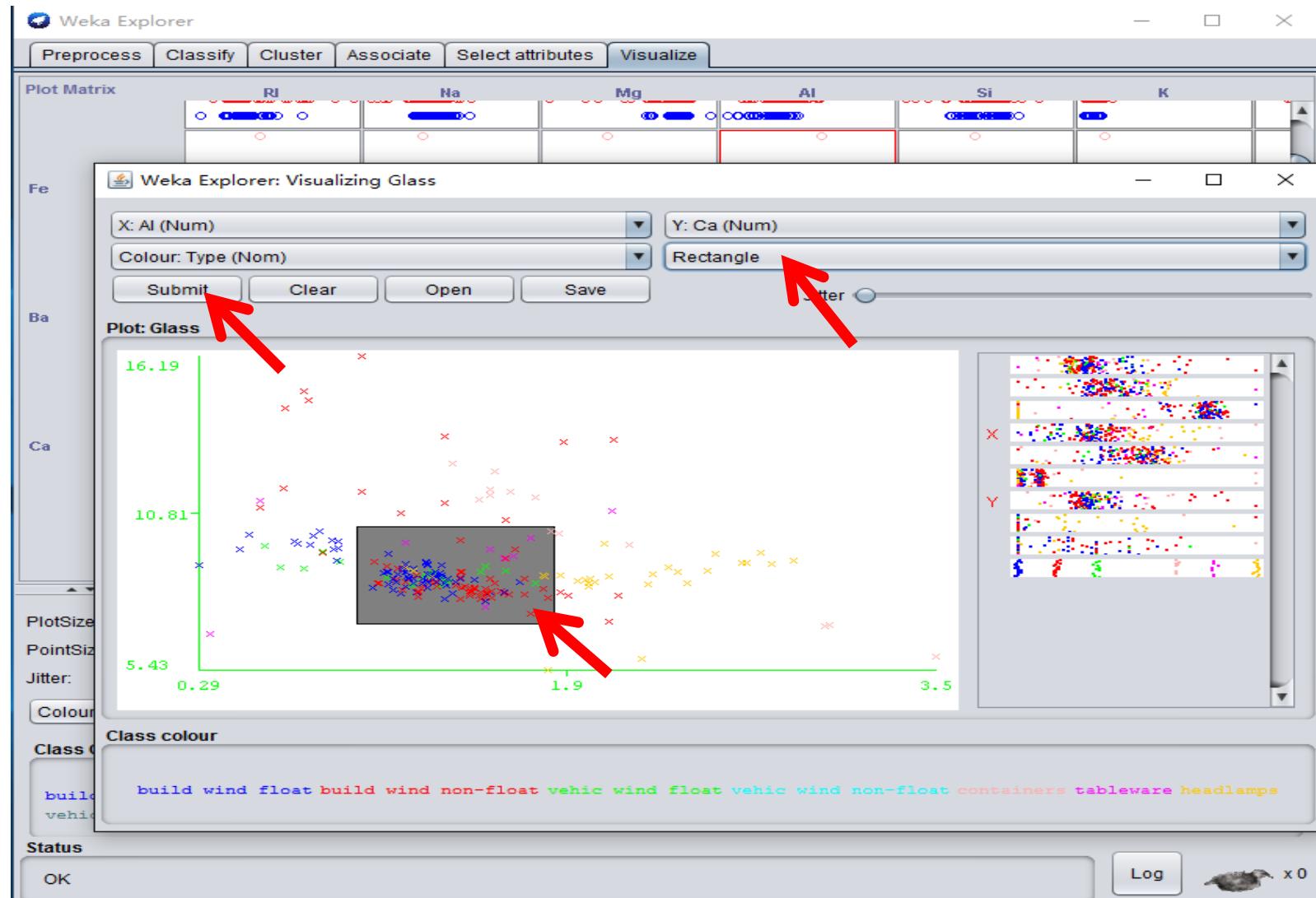
Explorer: Data Visualization



Explorer: Data Visualization



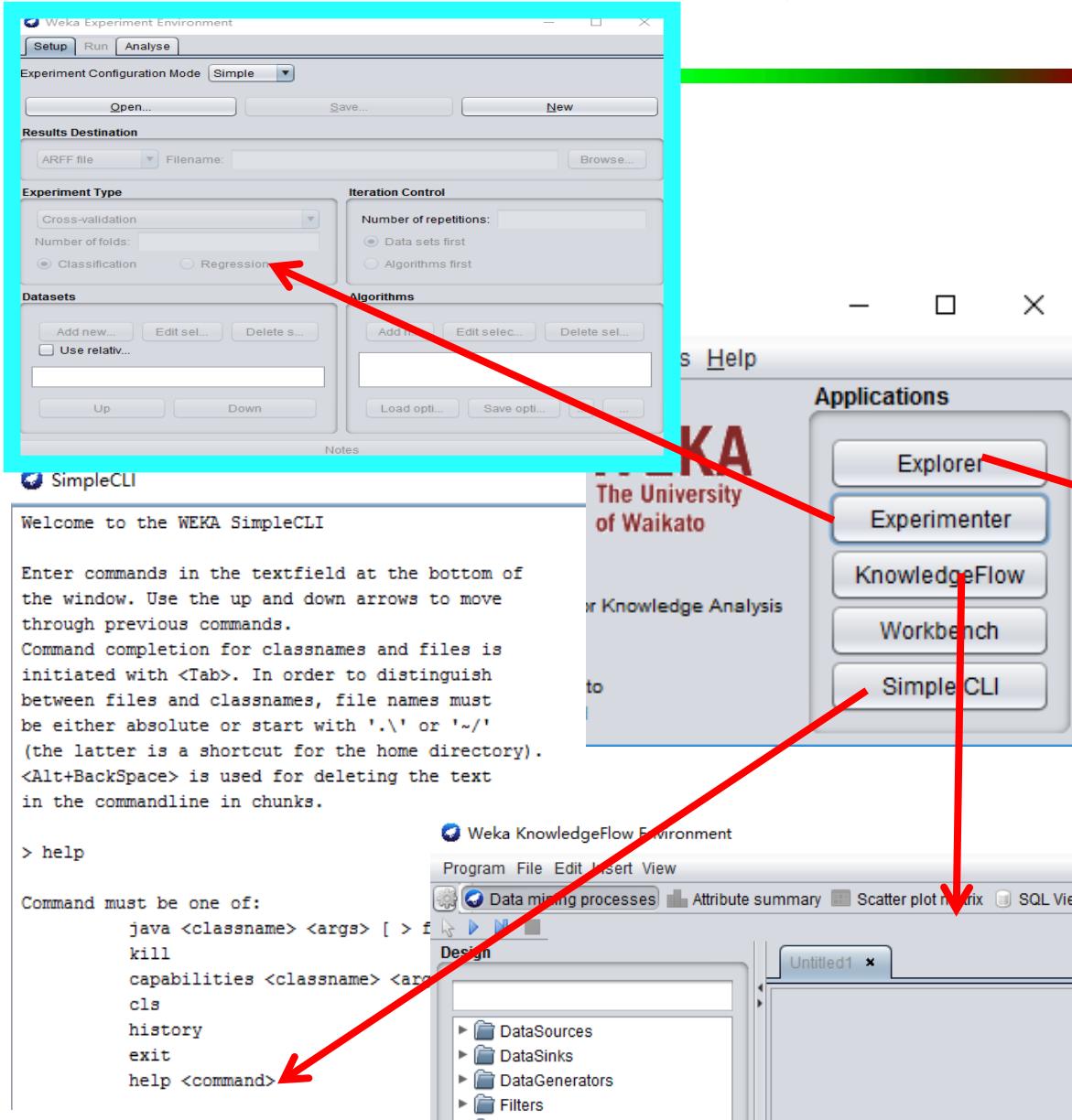
Explorer: Data Visualization



Explorer: Data Visualization



WEKA窗口



■ Applications

- Explorer
- **Experimenter**
- Knowledge Flow
- Command line Interface (SimpleCLI)

Roadmap: WEKA Usage

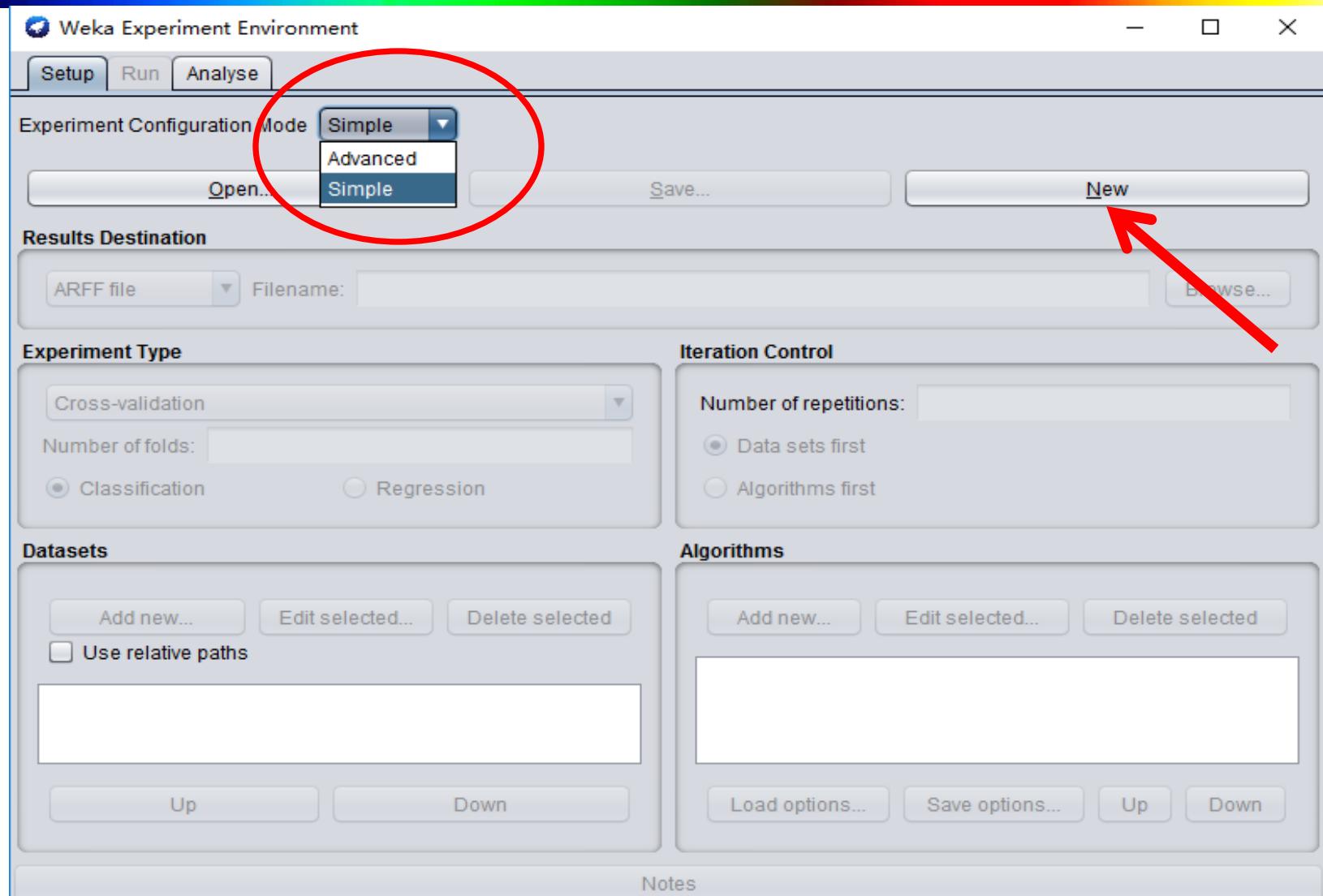
- WEKA Usage (version 3.8)

- Explorer
- Experimenter
- Knowledge Flow

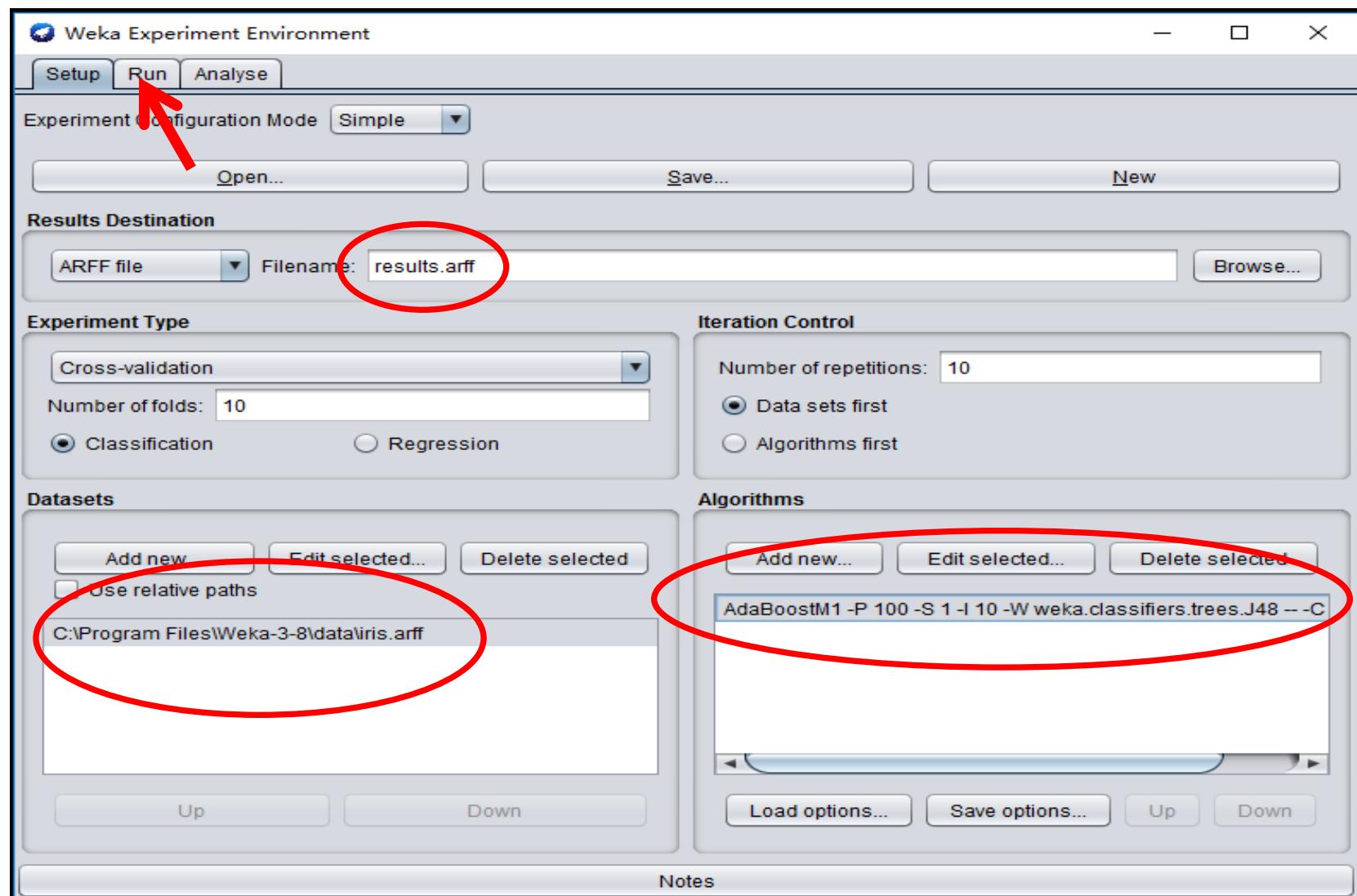
Performing Experiments

- 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!

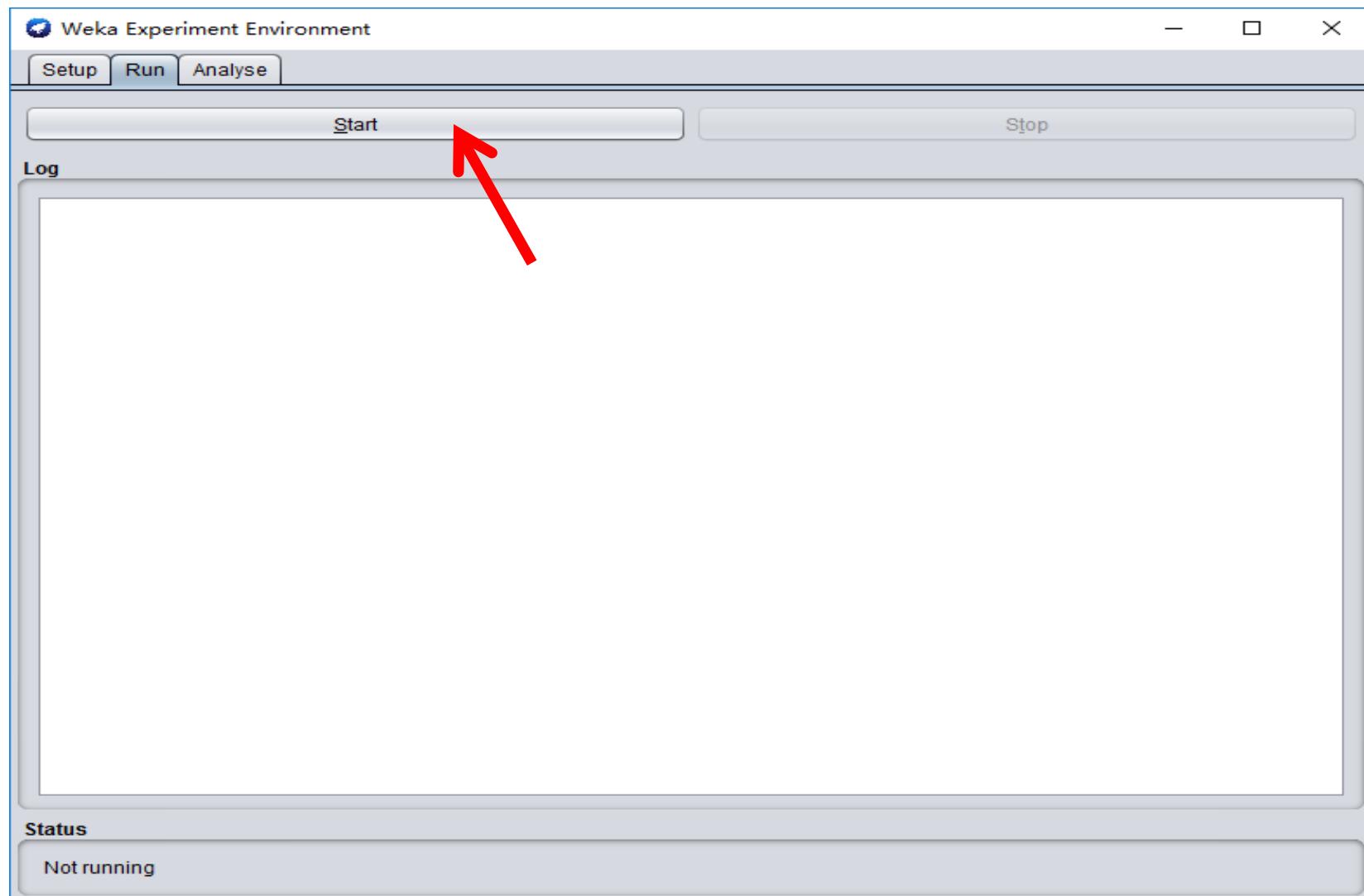
Experimenter



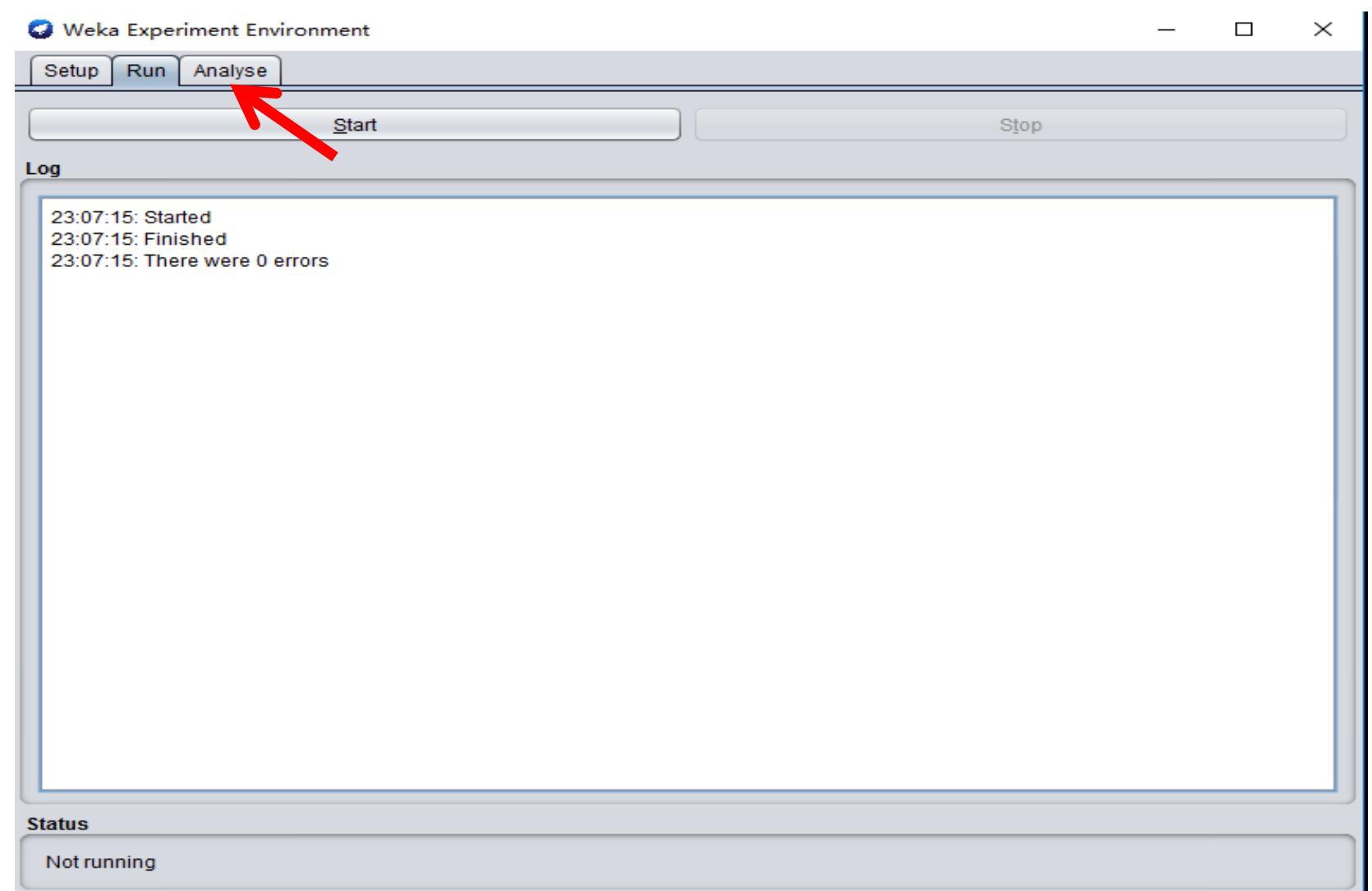
Experimenter



Experimenter



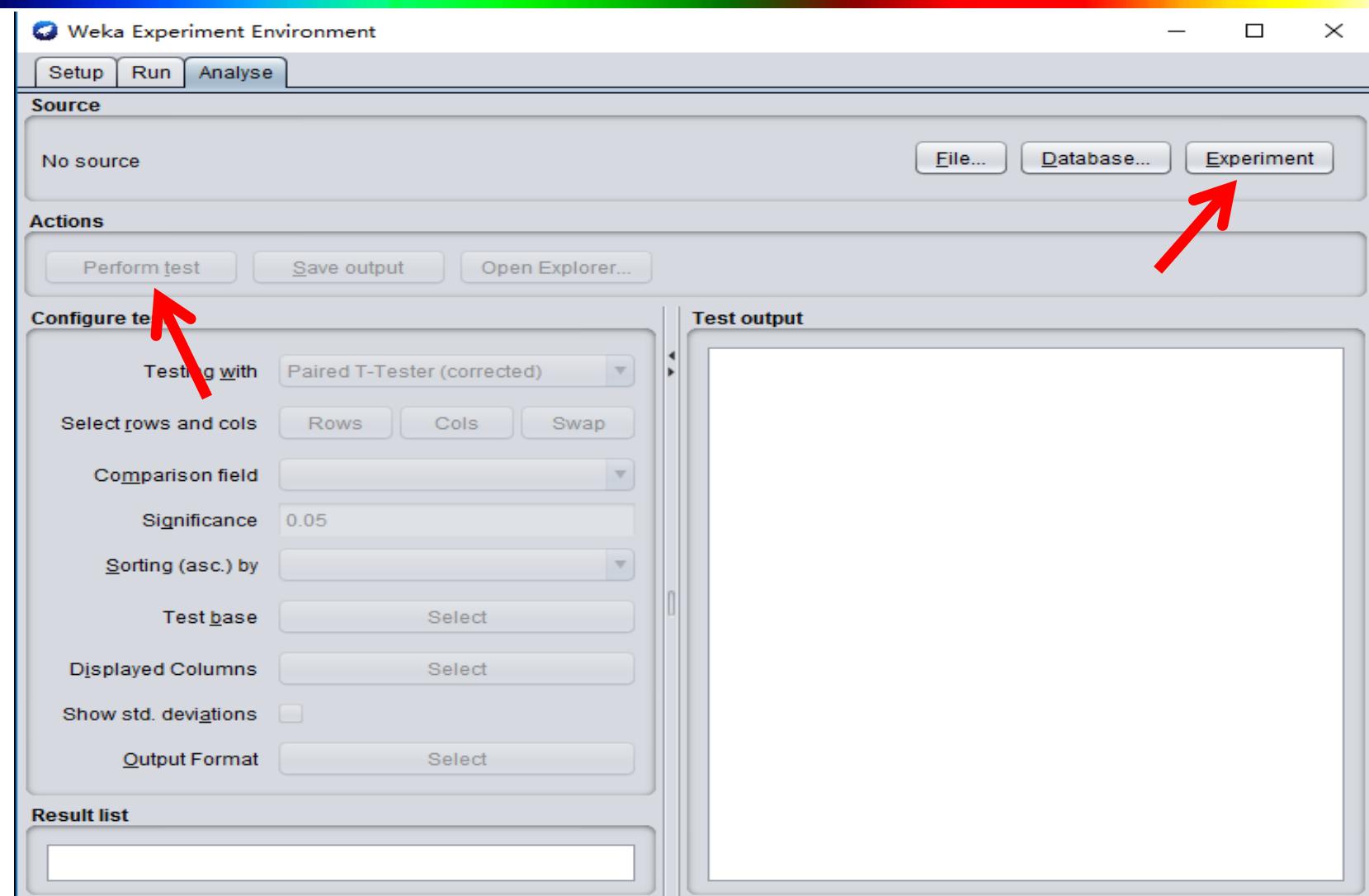
Experimenter



Status

Not running

Experimenter



Experimenter

Weka Experiment Environment

Setup Run Analyse

Source

Got 900 results

Actions

Perform test Save

Configure test

Testing with Paired T-Tester (corrected)

Select rows and cols Rows Cols Swap

Comparison field Percent_correct

Significance 0.05

Sorting (asc.) by <default>

Test base Select

Displayed Columns Select

Show std. deviations

Output Format Select

Result list

23.12.16 - Percent_Correct - meta AdaBoostM1 - P 100 - S 1 - I 10 - W trees.J48 -- - C 0.25 - M 2'

3 datasets
* 3 classifiers
* 10 cross-validation folds
* 10 iterations
= 900 runs

File... Database... Experiment

Dataset (1) meta.Ada | (2) trees (3) bayes

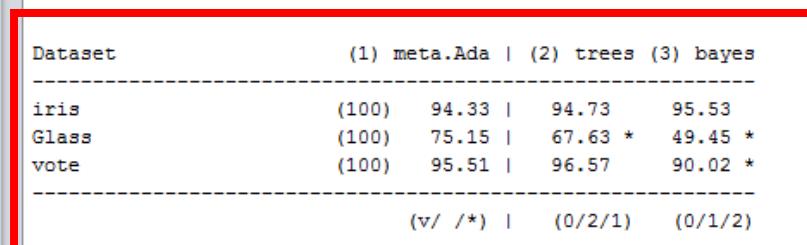
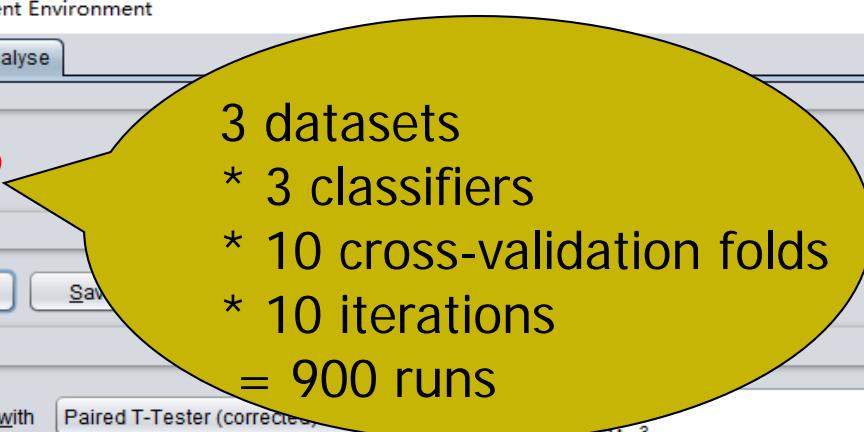
iris (100) 94.33 | 94.73 95.53

Glass (100) 75.15 | 67.63 * 49.45 *

vote (100) 95.51 | 96.57 90.02 *

(v/ /*) | (0/2/1) (0/1/2)

Key:
(1) meta.AdaBoostM1 '-P 100 -S 1 -I 10 -W trees.J48 -- -C 0.25 -M 2'
(2) trees.J48 '-C 0.25 -M 2' -2177331683936444444
(3) bayes.NaiveBayes '' 5995231201785697655



Close-Up: the result

| Dataset | (1) meta.Ada | | (2) trees | (3) bayes | |
|---------|--------------|-------|-----------|-----------|---------|
| iris | (100) | 94.33 | | 94.73 | 95.53 |
| Glass | (100) | 75.15 | | 67.63 * | 49.45 * |
| vote | (100) | 95.51 | | 96.57 | 90.02 * |
| | (v/ /*) | | (0/2/1) | (0/1/2) | |

- The first classifier acts as baseline classifier
- ✓ denotes the significantly better than baseline at the significance level $\alpha = 0.05$
- * denotes the significantly worse than baseline
- (0/1/2): significantly better in 0 dataset; 1 dataset has no significant difference; 2 datasets significantly worse

Experimenter

Weka Experiment Environment

Source

Got 900 results

Actions

Perform test Save output Open Explorer...

Configure test

Testing with: Paired T-Tester (corrected)

Select rows and cols: Rows Cols Swap

Comparison field: Percent_correct

Significance: 0.05

Sorting (asc.) by: <default>

Test base: Select

Displayed Columns: Select

Show std. deviations:

Output Format: Select

Result list

23:12:17 - Available resultsets
23:12:18 - Percent_correct - meta.AdaboostM1 '-P 100 -S 1'

Test output

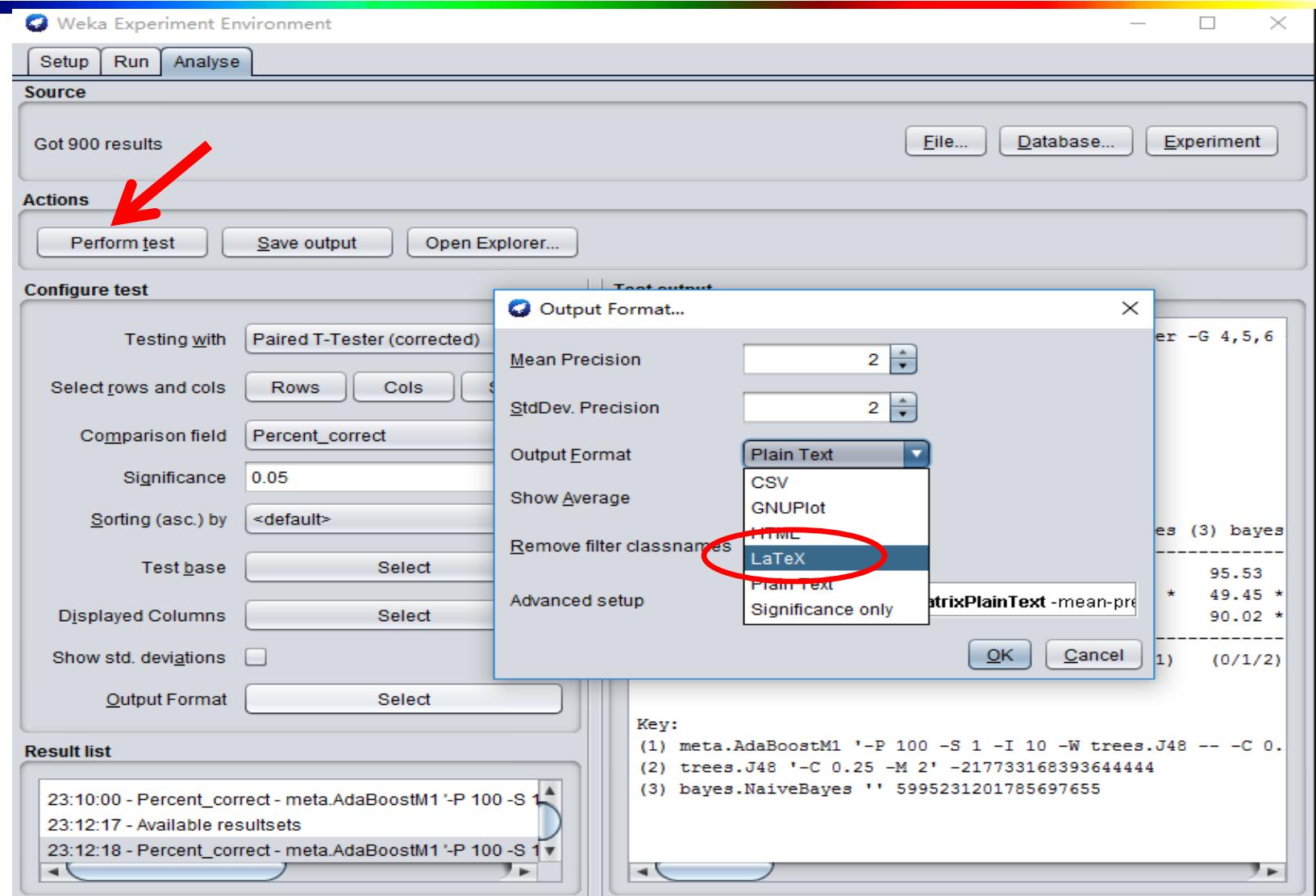
```
Analysing: Percent_correct
Datasets: 3
Resultsets: 3
Confidence: 0.05 (two tailed)
Sorted by: -
Date: 16-10-7 下午11:12

Dataset (1) meta.Ada | (2) trees (3) bay
-----
iris (100) 94.33 | 94.73 95.53
Glass (100) 75.15 | 67.63 * 49.45
vote (100) 95.51 | 96.57 90.02
-----
(vv /*) | (0/2/1) (0/1/)

Key:
(1) meta.AdaboostM1 '-P 100 -S 1 -I 10 -W trees.J48 -- -C
(2) trees.J48 '-C 0.25 -M 2' -217733168393644444
(3) bayes.NaiveBayes '' 5995231201785697655
```



Experimenter



Experimenter

Weka Experiment Environment

Setup Run Analyse

Source

Got 900 results

Actions

Perform test Save output Open Explorer...

Configure test

Testing with: Paired T-Tester (corrected)

Select rows and cols: Rows Cols Swap

Comparison field: Percent_correct

Significance: 0.05

Sorting (asc.) by: <default>

Test base: Select

Displayed Columns: Select

Show std. deviations:

Output Format: Select

Result list

- 23:10:00 - Percent_correct - meta.AdaboostM1 '-P 100 -S 1'
- 23:12:17 - Available resultsets
- 23:12:18 - Percent_correct - meta.AdaboostM1 '-P 100 -S 1'
- 23:22:52 - Percent_correct - meta.AdaboostM1 '-P 100 -S 1'

File... Database... Experiment

Test output

```
\hline
Dataset & (1) & (2) & (3) & \\
\hline
iris & 94.33 & 94.73 & 95.53 & \\
Glass & 75.15 & 67.63 & \bullets & 49.45 & \bullets\\
vote & 95.51 & 96.57 & 90.02 & \bullets\\
\hline
\multicolumn{6}{c}{\circles, \bullet statistically significant} \\
\end{tabular} \footnotesize \par
\end{table}

\begin{table}[thb]
\caption{\label{labelname}Table Caption (Key)}
\scriptsize
\centering
\begin{tabular}{cl}
(1) & meta.AdaboostM1 '-P 100 -S 1 -I 10 -W trees.J48 -- \\
(2) & trees.J48 '-C 0.25 -M 2' -2177331683936444444 \\
(3) & bayes.NaiveBayes '' 5995231201785697655 \\
\end{tabular}
\end{table}
```

WEKA窗口

The diagram illustrates the various components of the WEKA environment:

- SimpleCLI**: A terminal-like window where commands like `help <command>` can be entered.
- Applications**: A window listing the available applications: Explorer, Experimenter, KnowledgeFlow, Workbench, and Simple CLI.
- Weka Explorer**: A window for data preprocessing and analysis, showing tabs for Preprocess, Classify, Cluster, Associate, Select attributes, and Visualize.
- Weka KnowledgeFlow Environment**: A window for building data mining processes, showing tabs for Data mining processes, Attribute summary, Scatter plot matrix, SQL Viewer, and Simple CLI.

Applications

- Explorer
- Experimenter
- Knowledge Flow
- Command line Interface (SimpleCLI)

Roadmap: WEKA Usage

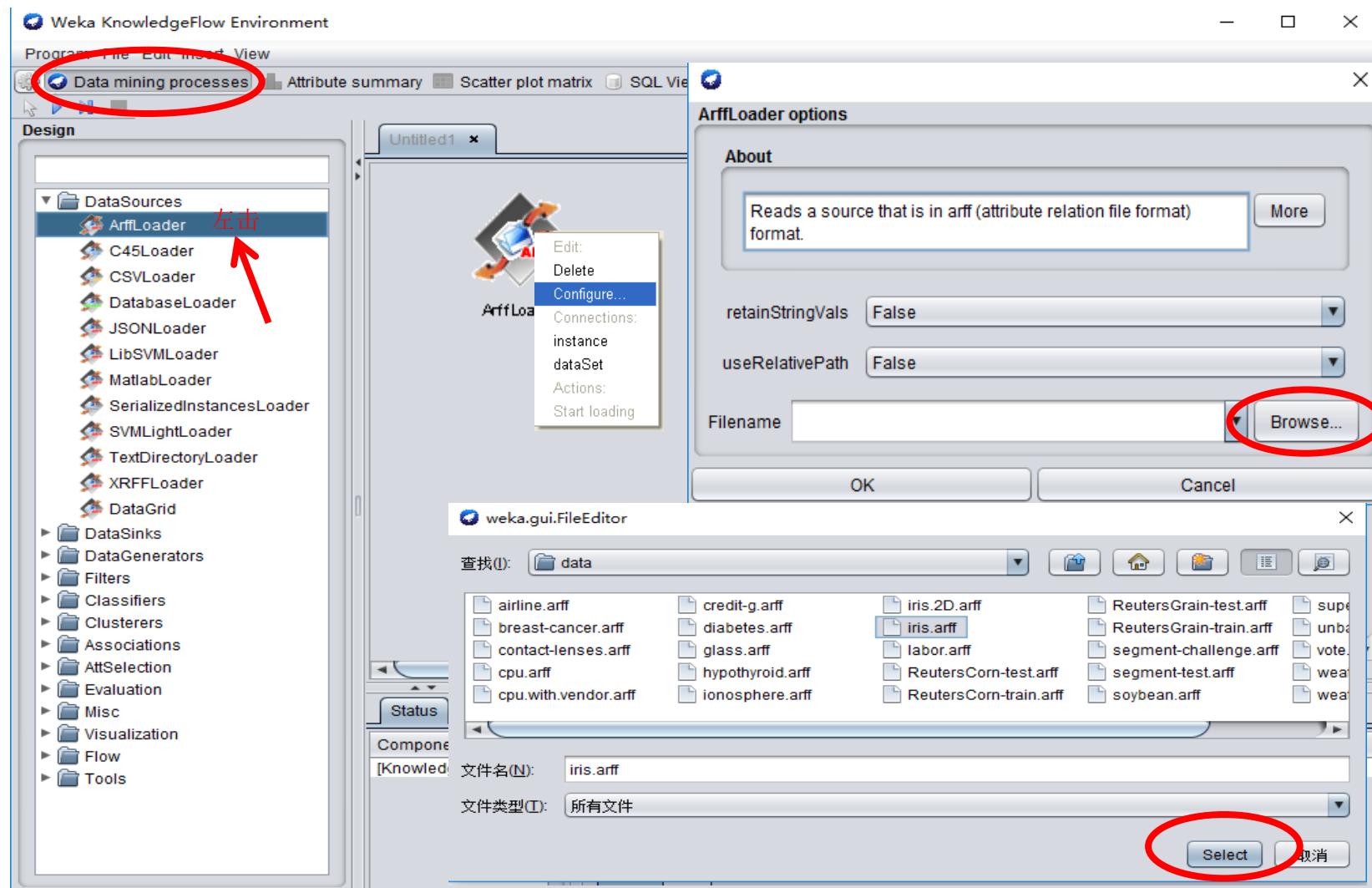
- WEKA Usage (version 3.8)

- Explorer
- Experimenter
- Knowledge Flow

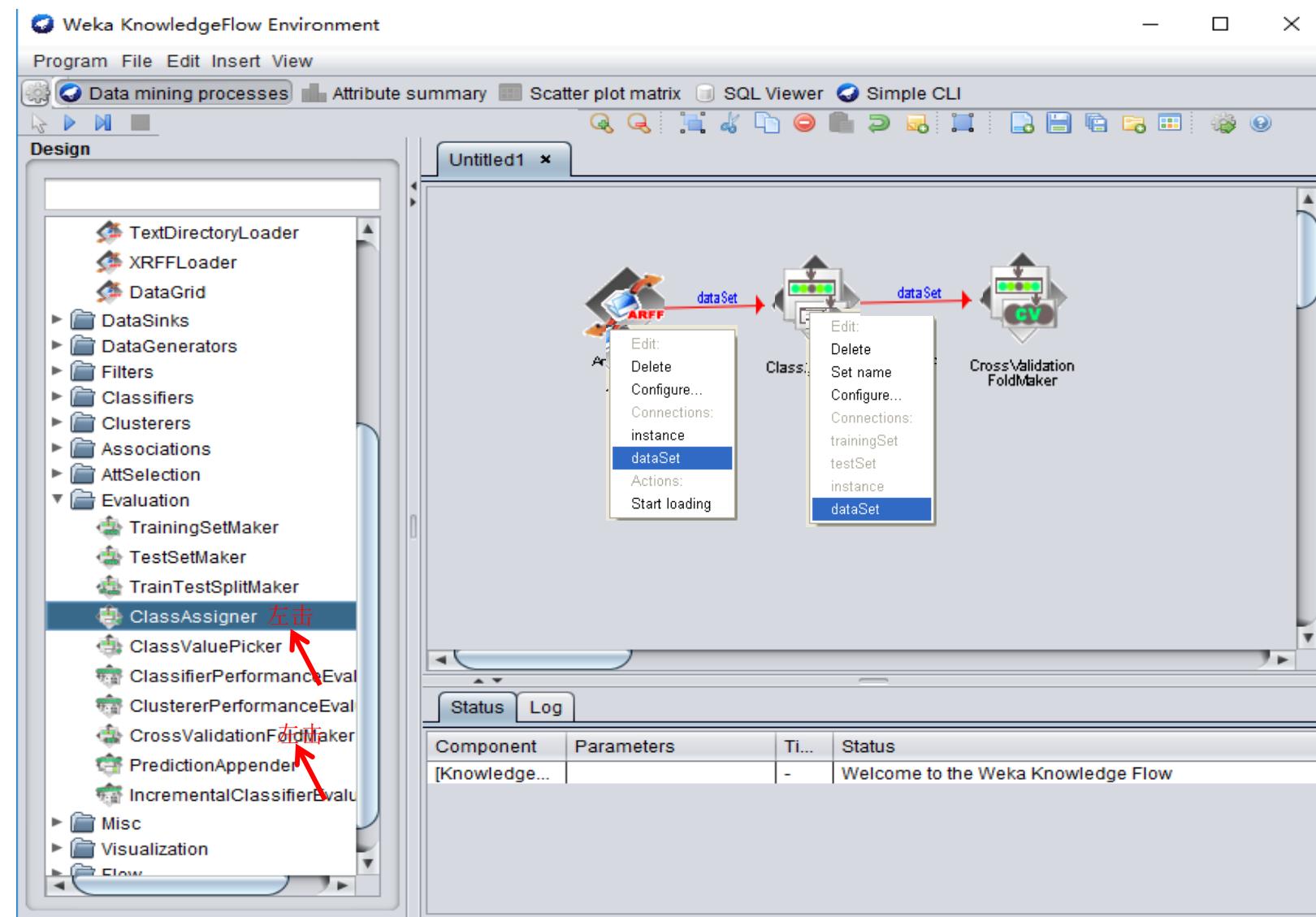
The Knowledge Flow GUI

- 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

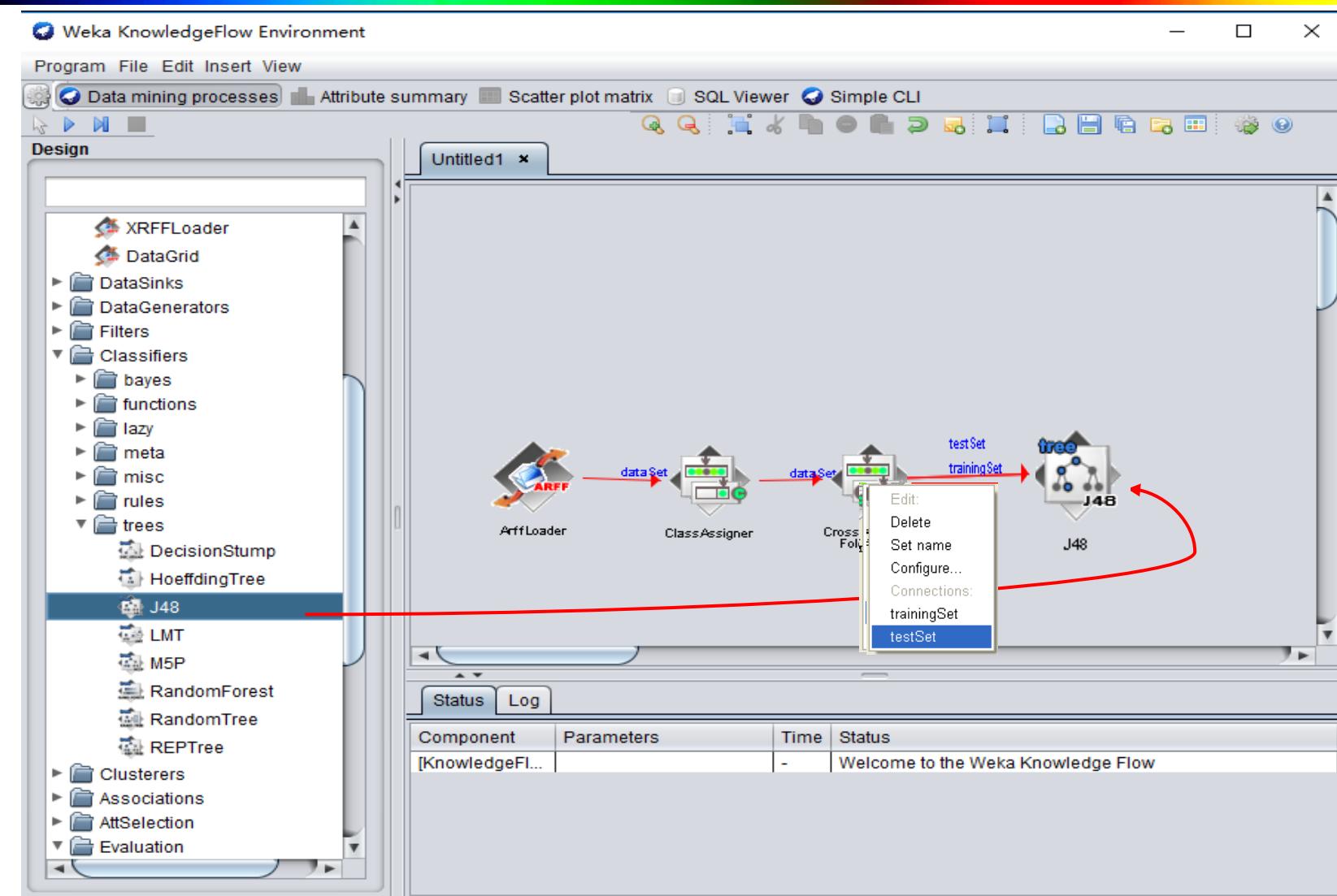
Ex1: Cross-validated J48 (C4.5)



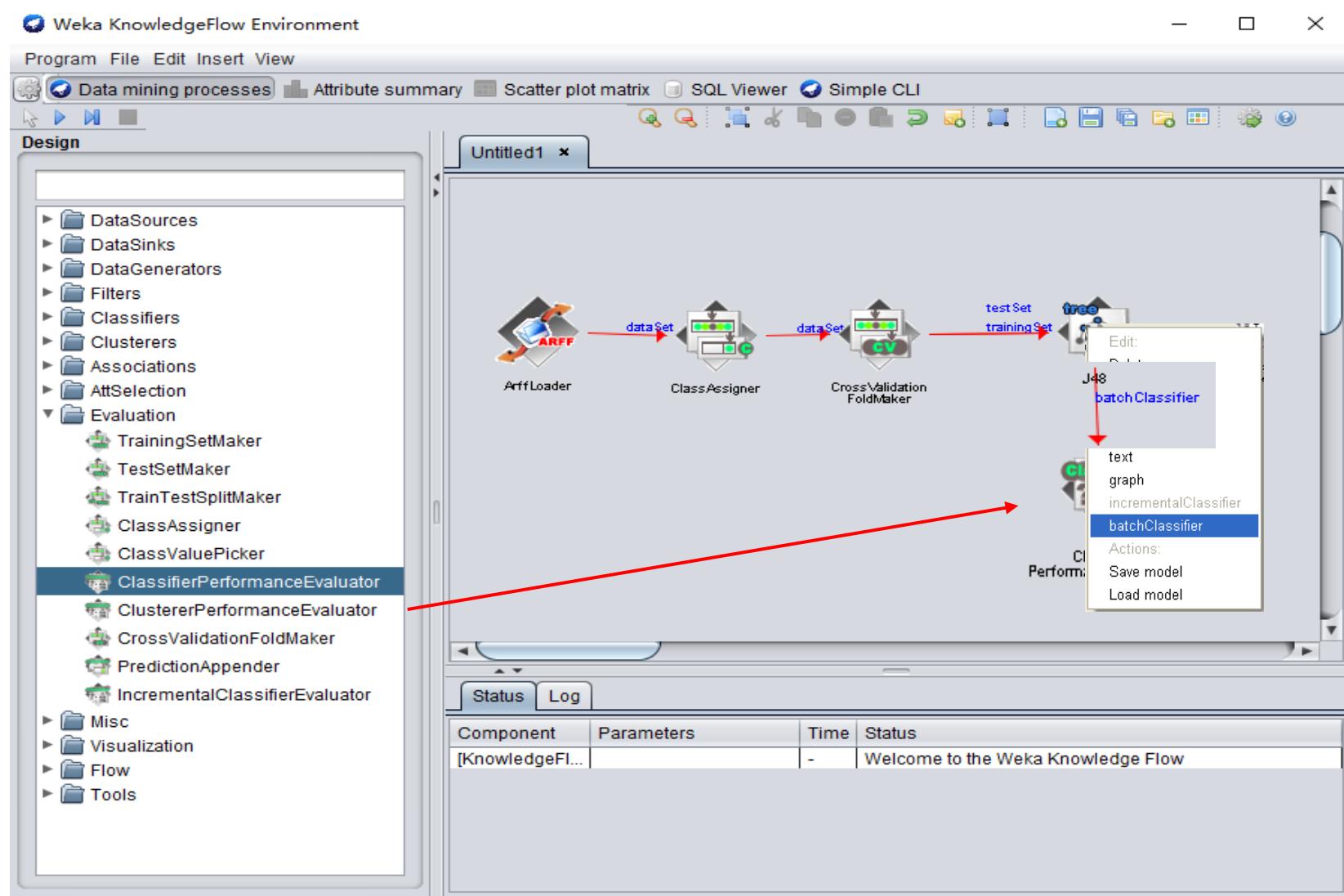
Ex1: Cross-validated J48 (C4.5)



Ex1: Cross-validated J48 (C4.5)



Ex1: Cross-validated J48 (C4.5)



Ex1: Cross-validated J48 (C4.5)

Weka KnowledgeFlow Environment

Program File Edit Insert View

Data mining processes Attribute summary Scatter plot matrix SQL Viewer Simple CLI

Design Untitled1

Text Viewer

Result list

13:56:32.588 - J48

Text

```
Mean absolute error          0.035
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  MCC  ROC Ar
0.980   0.000    1.000     0.980    0.990     0.985   0.990
0.940   0.030    0.940     0.940    0.940     0.910   0.952
0.960   0.030    0.941     0.960    0.950     0.925   0.961
Weighted Avg.    0.960   0.020    0.960     0.960    0.960     0.940   0.968

*** Confusion Matrix ***

  a  b  c  <-- classified as
49  1  0 |  a = Iris-setosa
0 47  3 |  b = Iris-versicolor
0  2 48 |  c = Iris-virginica
```

Close Settings Clear results

GraphViewer CostBenefitAnalysis

Flow Tools

Component Parameters Time Status

[KnowledgeFl...]

CV Validation Maker

testSet trainingSet

J48 batchClassifier

Classifier Performance Evaluator

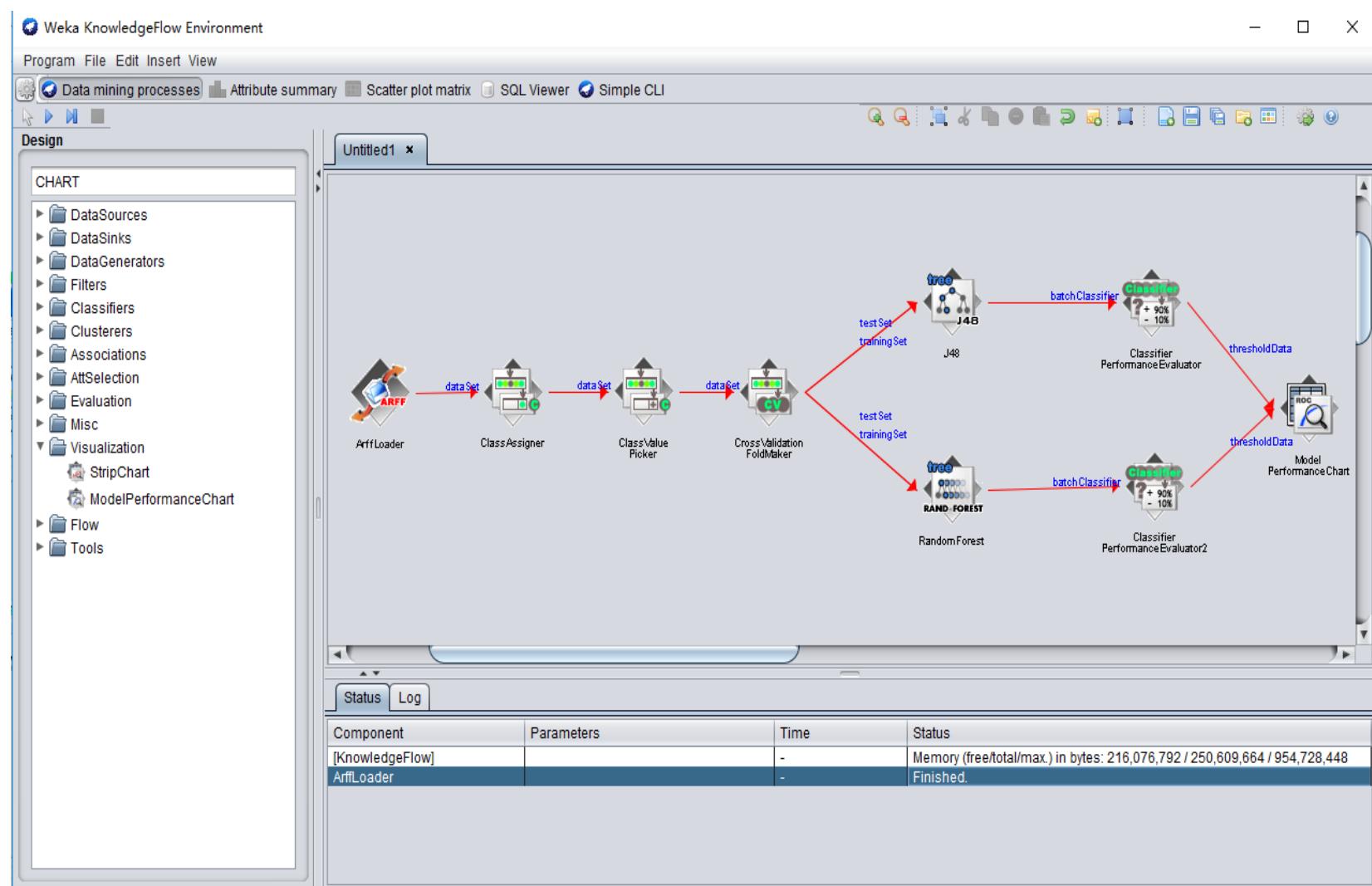
text

ab 123 wx Y7

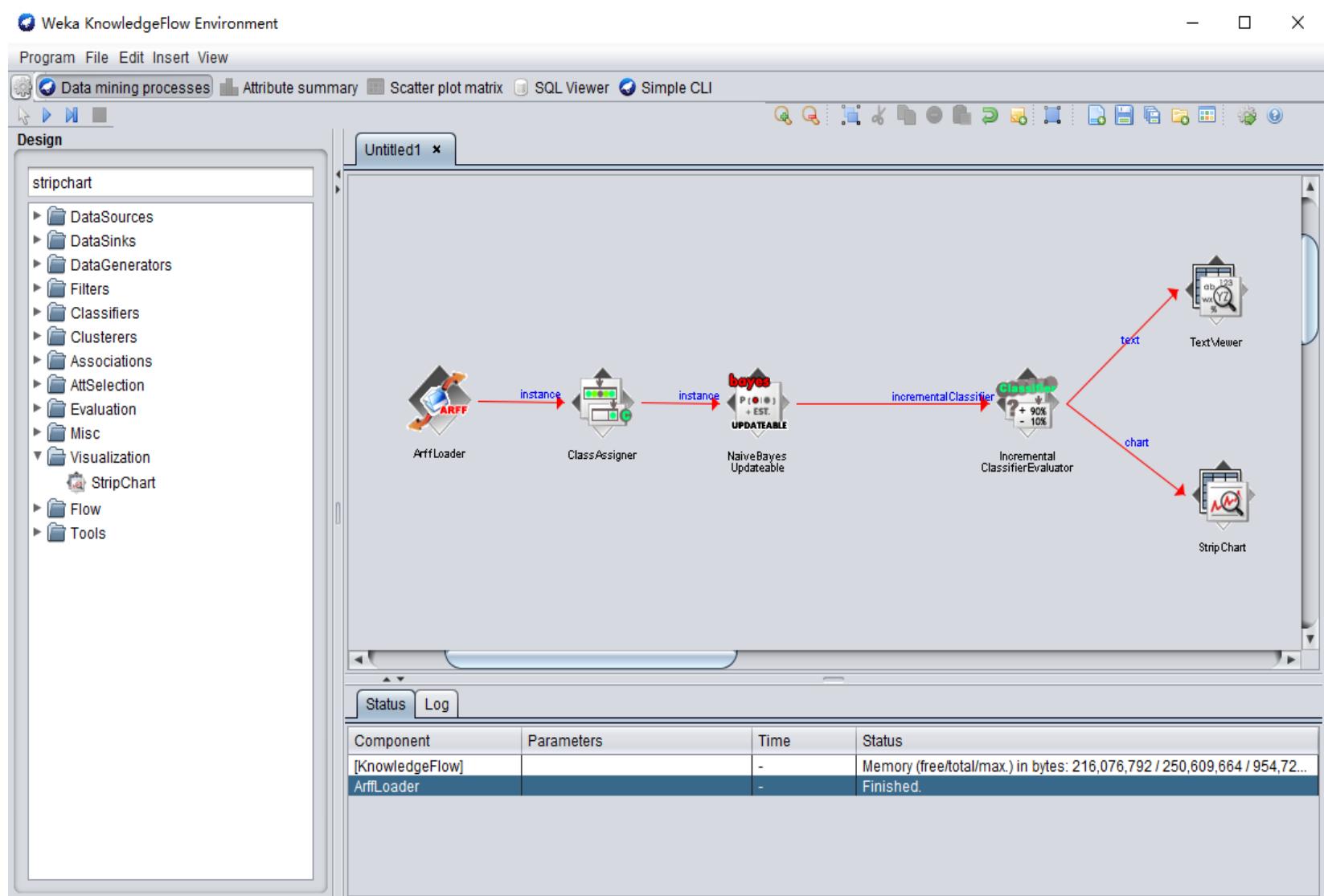
Edit: Delete Set name Connections: text Actions: Show results Clear results

```
graph TD
    CV[CV] --> JM[Validation Maker]
    JM -- testSet --> J48[J48]
    JM -- trainingSet --> BC[batchClassifier]
    J48 --> BC
    BC --> CE[Classifier Performance Evaluator]
    CE -- text --> CV
```

Ex2: Plotting multiple ROC curves



Ex3: Processing data incrementally



Machine Learning with WEKA 3.8

- About WEKA
- Data Format and Preprocessing
- WEKA Usage (version 3.8)
 - Explorer
 - Experimenter
 - Knowledge Flow
- Other Data Mining Tools
- Summary

Other Data Mining Tools

- R (<http://www.r-project.org>)
- Tanagra (<http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html>)
- YALE (Yet Another Learning Environment,
<http://rapid-i.com/>, new name: RapidMiner)
- KNIME (Konstanz Information Miner,
<http://www.knime.org>)
- Orange (<http://orange.biolab.si/>)
- GGobi (<http://www.ggobi.org/>)

Machine Learning with WEKA 3.8

- About WEKA
- Data Format and Preprocessing
- WEKA Usage (version 3.8)
 - Explorer
 - Experimenter
 - Knowledge Flow
- Other Data Mining Tools
- Summary

Summary

- Java-based open source
- Provide various data preprocessing, data mining functions and evaluation measures
- Incremental functions and components
- WEKA is available at

<http://www.cs.waikato.ac.nz/ml/weka>

Projects based on WEKA

- 45 projects currently (30/01/07) listed on the [WekaWiki](#)
- Incorporate/wrap WEKA
 - GRB Tool Shed - a tool to aid gamma ray burst research
 - YALE - facility for large scale ML experiments
 - GATE - NLP workbench with a WEKA interface
 - Judge - document clustering and classification
 - RWeka - an R interface to Weka
- Extend/modify WEKA
 - BioWeka - extension library for knowledge discovery in biology
 - WekaMetal - meta learning extension to WEKA
 - Weka-Parallel - parallel processing for WEKA
 - Grid Weka - grid computing using WEKA
 - Weka-CG - computational genetics tool library

Limitations of WEKA

- Traditional algorithms need to have all data in main memory
 - ==> big datasets are an issue
 - Solution:
 - Incremental schemes
 - Stream algorithms
- MOA “**M**assive **O**nline **A**nalysis”
- (not only a *flightless* bird, but also *extinct!*)

Conclusion: try it yourself!

- WEKA related reference
 - WekaWiki – <http://weka.sourceforge.net/wiki/>
 - WekaDoc – <http://weka.sourceforge.net/wekadoc/>
 - Ensemble Selection on WekaDoc –
 - [http://weka.sourceforge.net/wekadoc/index.php/en:Ensemble Selection](http://weka.sourceforge.net/wekadoc/index.php/en:Ensemble_Selection)
 - Extensions for Weka's main GUI on WekaWiki –
 - [http://weka.sourceforge.net/wiki/index.php/Extensions for Weka%27s main GUI](http://weka.sourceforge.net/wiki/index.php/Extensions_for_Weka%27s_main_GUI)
 - Adding tabs in the Explorer on WekaWiki –
 - [http://weka.sourceforge.net/wiki/index.php/Adding tabs in the Explorer](http://weka.sourceforge.net/wiki/index.php/Adding_tabs_in_the_Explorer)
 - Explorer visualization plugins on WekaWiki –
 - [http://weka.sourceforge.net/wiki/index.php/Explorer visualization plugins](http://weka.sourceforge.net/wiki/index.php/Explorer_visualization_plugins)
 - Other related reference
 - <http://www.itl.nist.gov/div898/handbook/index.htm>
 - <http://wiki.wekacn.org/index.php/%E6%95%B0%E6%8D%AE%E6%8C%96%E6%8E%98>
 - <http://forum.wekacn.org/>
 - <http://www.wekacn.org/>