



THE UNIVERSITY OF  
**WAIKATO**  
*Te Whare Wānanga o Waikato*

# *More Data Mining with Weka*

Class 2 - Lesson 1

離散化數字屬性

*(Discretizing numeric attributes)*

Ian H. Witten

Department of Computer Science University of Waikato  
New Zealand

[weka.waikato.ac.nz](http://weka.waikato.ac.nz)

# Lesson 2.1: 離散化數字屬性

Class 1 探索Weka介面，處理大數據

Class 2 離散以及文本分類

Class 3 分類規則、關聯規則、聚類

Class 4 選擇屬性以及計算成本

Class 5 神經網路，學習曲線和表現優化

Lesson 2.1 離散化

Lesson 2.2 Supervised discretization

Lesson 2.3 Discretization in J48

Lesson 2.4 Document classification

Lesson 2.5 Evaluating 2-class classification

Lesson 2.6 Multinomial Naïve Bayes



## Lesson 2.1: 離散化數字屬性

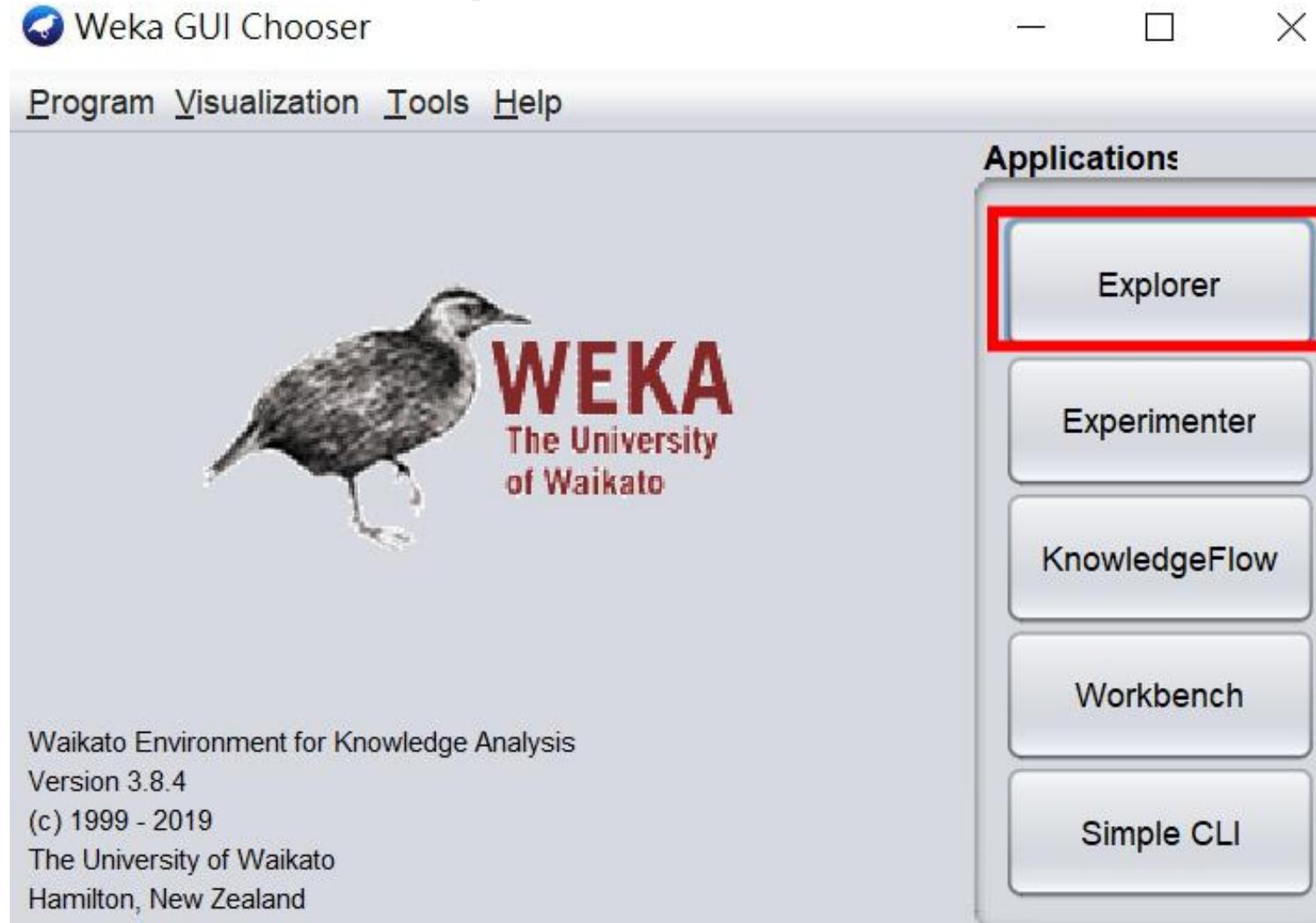
### 將數字屬性轉換為名詞屬性

- ❖ 等份裝箱 (Equal-width binning)
- ❖ 等頻裝箱(Equal-frequency binning) (“histogram equalization”)
- ❖ 應該選擇幾個箱子(bins)?
- ❖ 如何利用數值的排序信息?

## Lesson 2.1: 離散化數字屬性

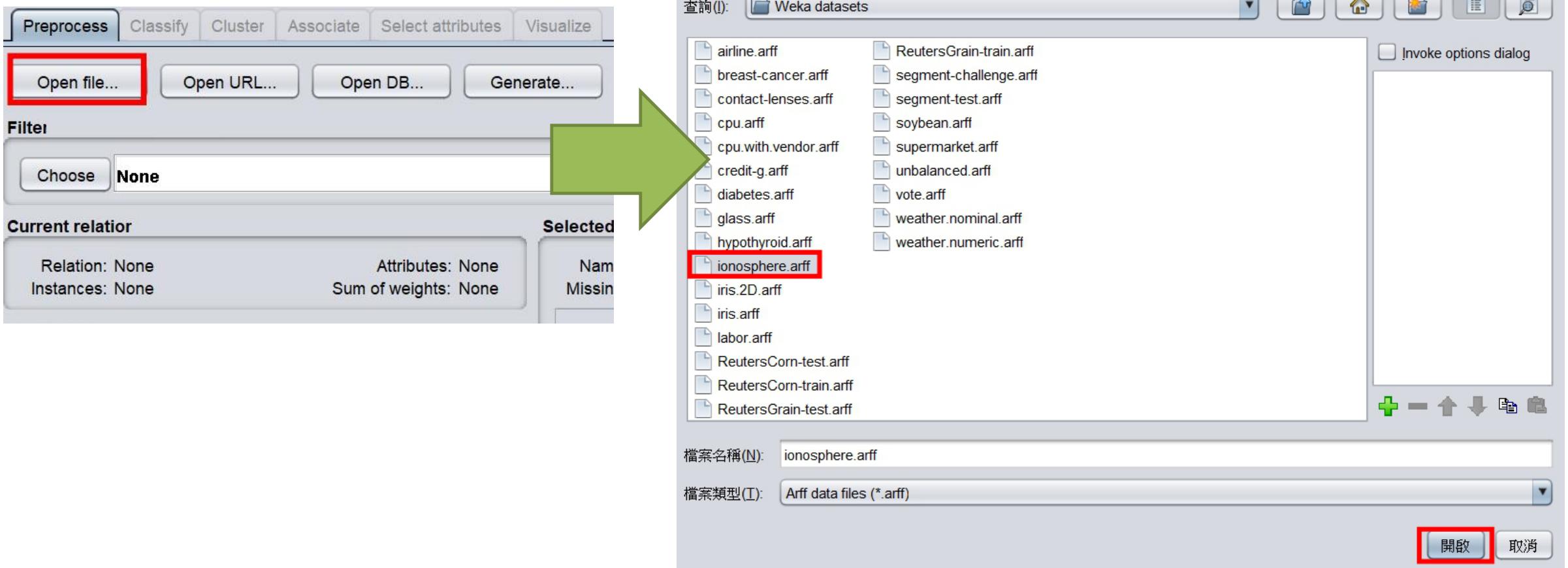
首先我們試試等份裝箱。

### 1. 開啟Weka的Explorer



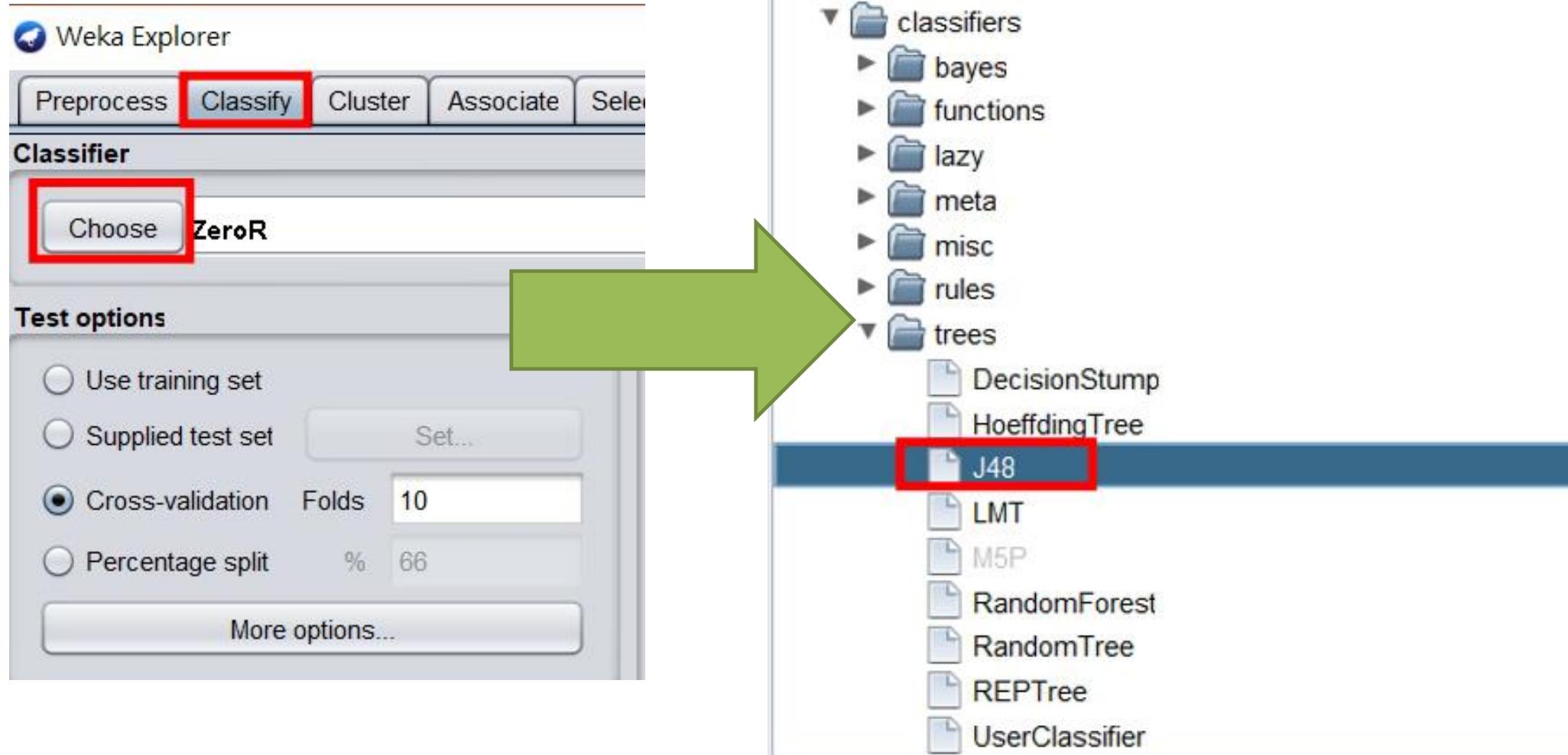
## Lesson 2.1: 離散化數字屬性

2. 左鍵點擊Open file...開啟右圖視窗，進入自行複製的Weka datasets，左鍵單擊**ionosphere.arff**的檔案後，再以左鍵單擊下方”開啟”以載入此檔案



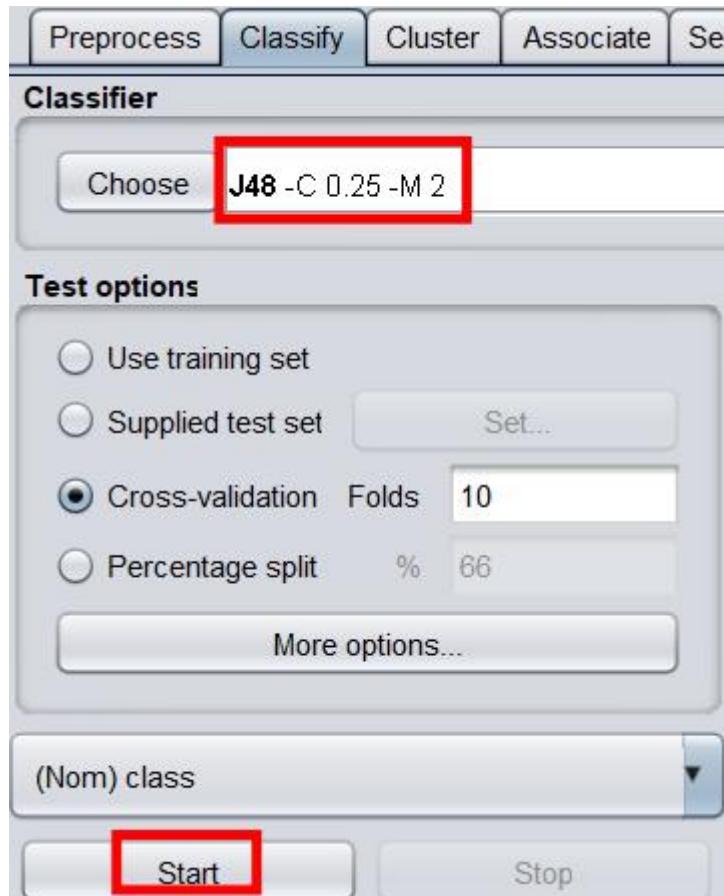
## Lesson 2.1: 離散化數字屬性

3. 切換到Classify界面點選Choose鈕，在出現的選單中左鍵單擊trees資料夾下的J48



## Lesson 2.1: 離散化數字屬性

4. 確認選擇了J48分類器後，左鍵單擊Start按鈕



## Lesson 2.1: 離散化數字屬性

▼執行結果：得到91.453%的準確率

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

- Weka Explorer** window title.
- Toolbar:** Preprocess, Classify (selected), Cluster, Associate, Select attributes, Visualize.
- Classifier Panel:** Choose dropdown set to **J48 -C 0.25 -M 2**.
- Test options:** Cross-validation (Folds 10) is selected.
- Classifier output:** Displays the following text:

```
Time taken to build model: 0.65 seconds
===
Stratified cross-validation ===
Summary ===

Correctly Classified Instances      321      91.453 %
Incorrectly Classified Instances   30       8.547 %
Kappa statistic                   0.8096
Mean absolute error               0.0938
Root mean squared error          0.2901
Relative absolute error           20.36 %
Root relative squared error     60.4599 %
Total Number of Instances        351
```

The accuracy value **91.453 %** is highlighted with a red box.

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

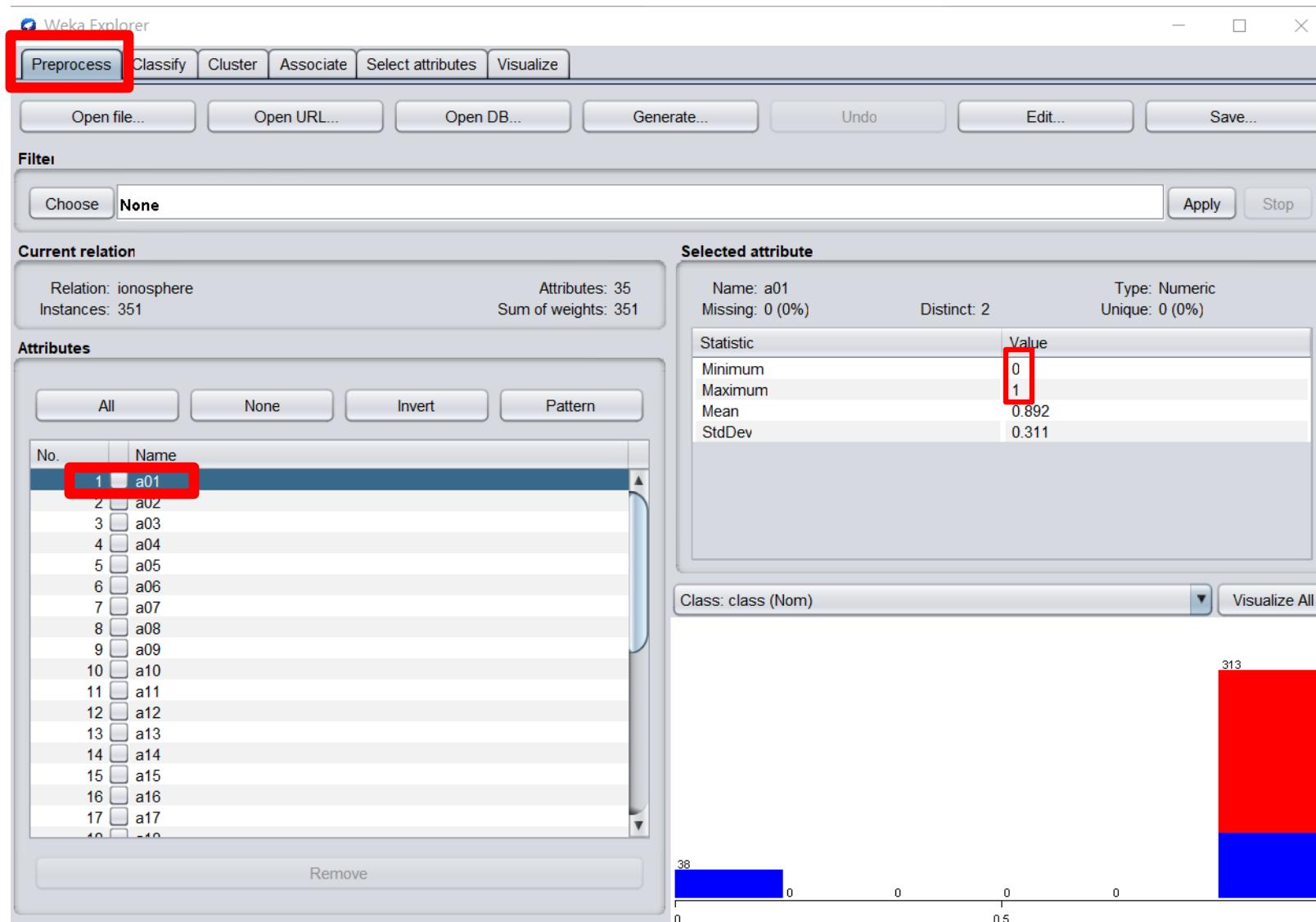
      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC  ROC Area  PRC Area  Cl
      0.825    0.036    0.929    0.825    0.874    0.813  0.892    0.855    b
      0.964    0.175    0.908    0.964    0.935    0.813  0.892    0.894    g
Weighted Avg.  0.915    0.125    0.915    0.915    0.913    0.813  0.892    0.880

===
Confusion Matrix ===

      a     b  <- classified as
104   22 |  a = b
     8 217 |  b = g
```

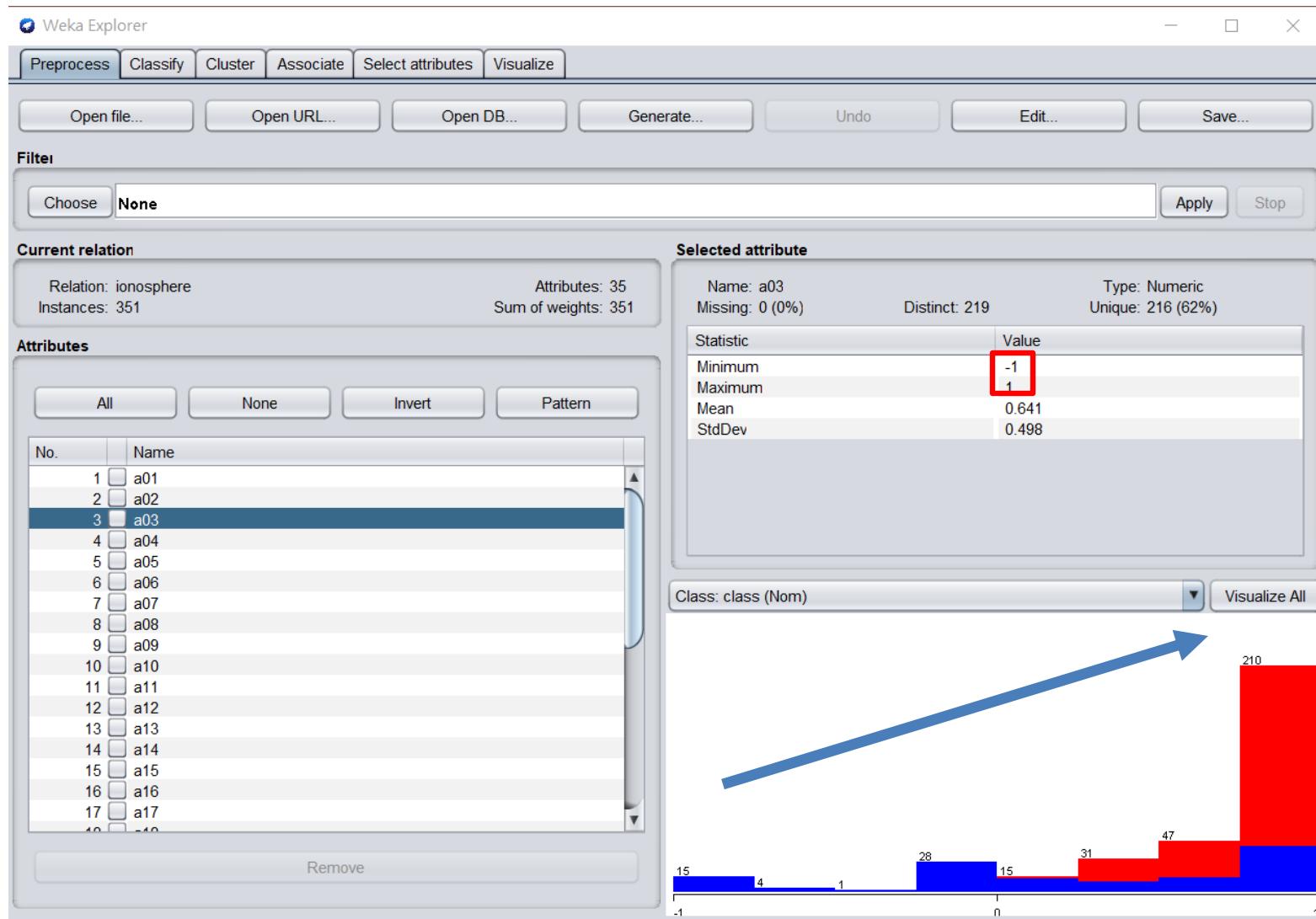
## Lesson 2.1: 離散化數字屬性

5. 切換到Preprocess面板，查看屬性a01：有兩個不同的數值0和1。



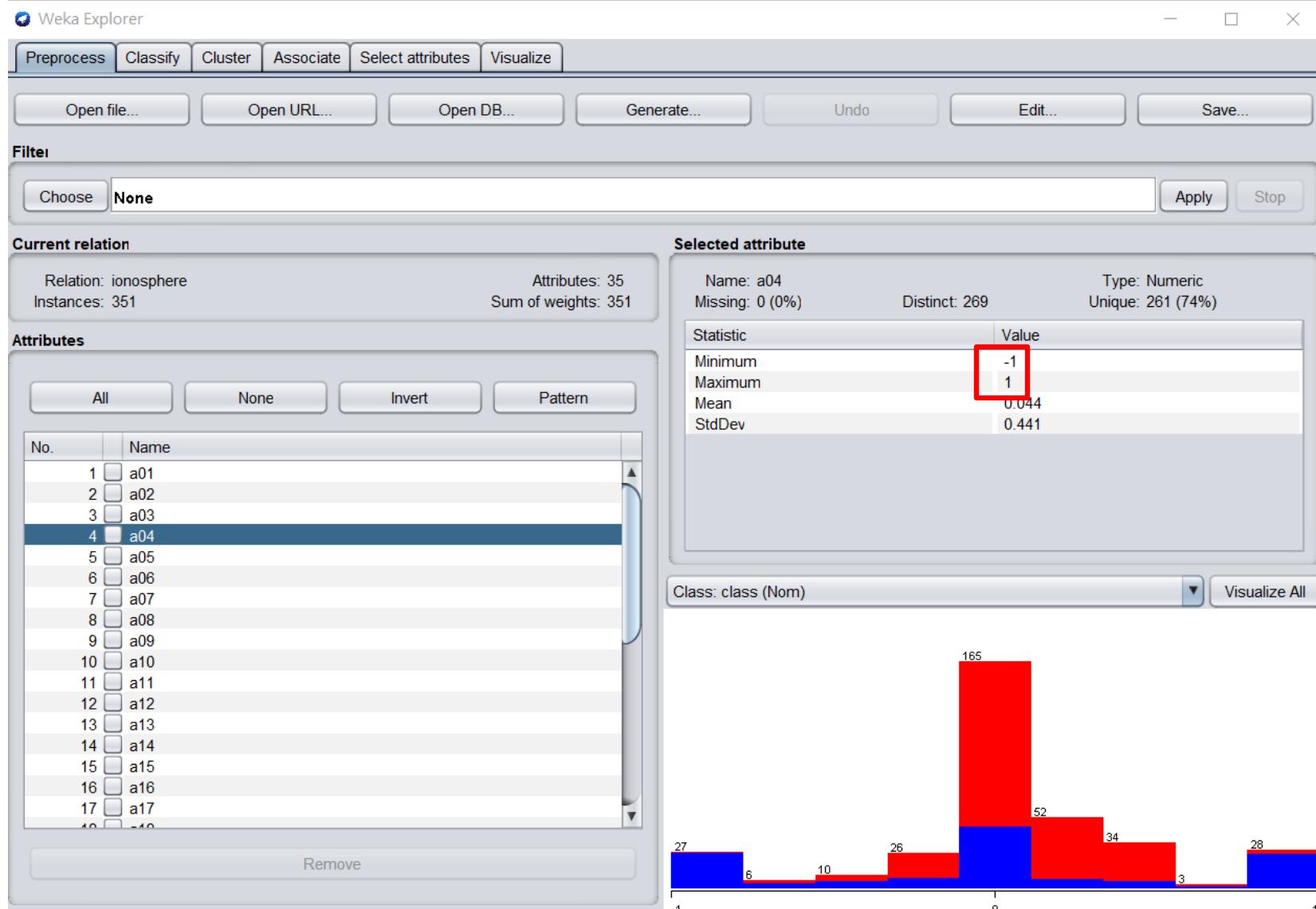
## Lesson 2.1: 離散化數字屬性

▼屬性a03包含從-1到+1的一組不同的數值，且逐漸增大。



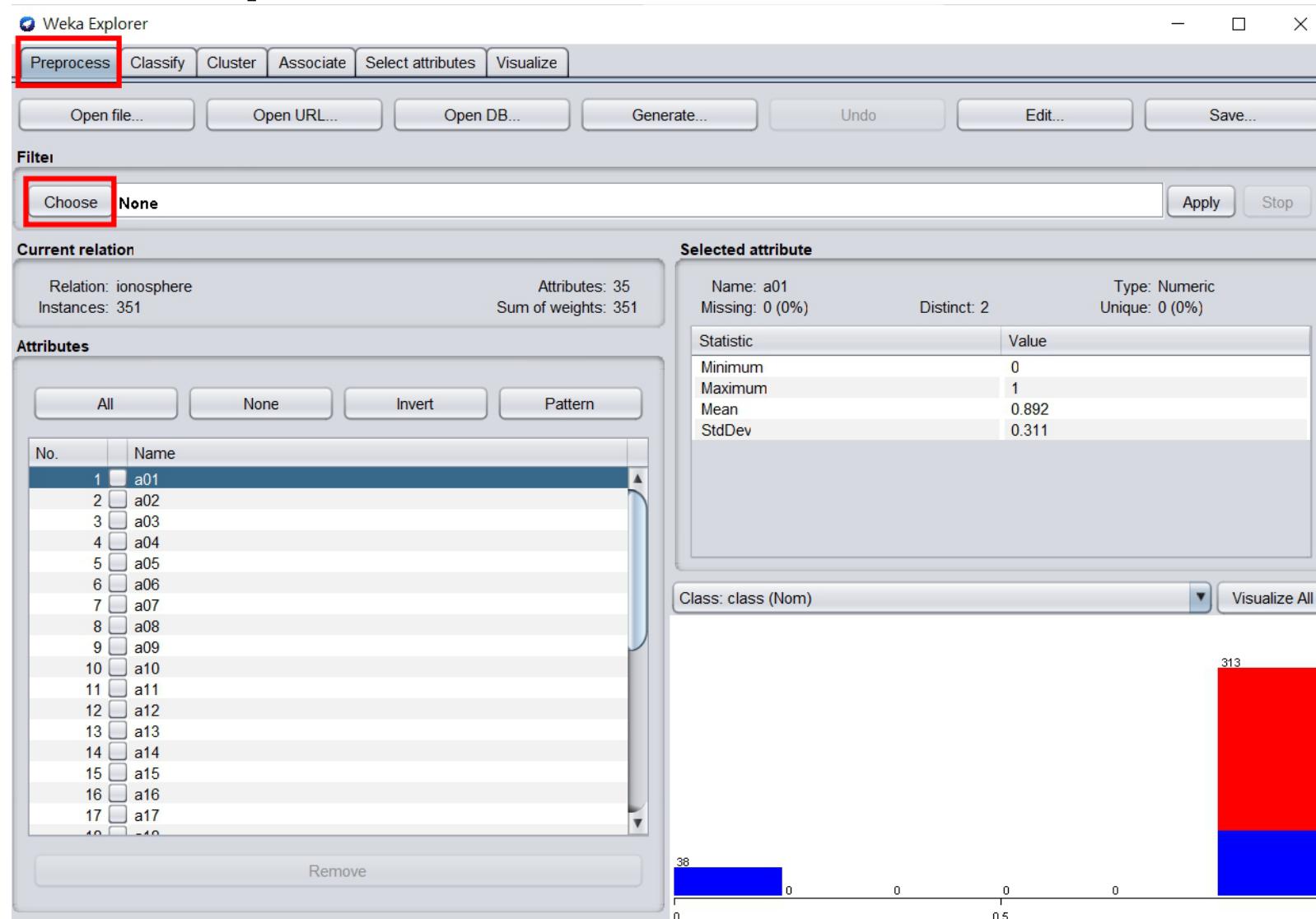
## Lesson 2.1: 離散化數字屬性

▼屬性a04也包含從-1到+1的不等數值，且圖表看上去很像常態分布。



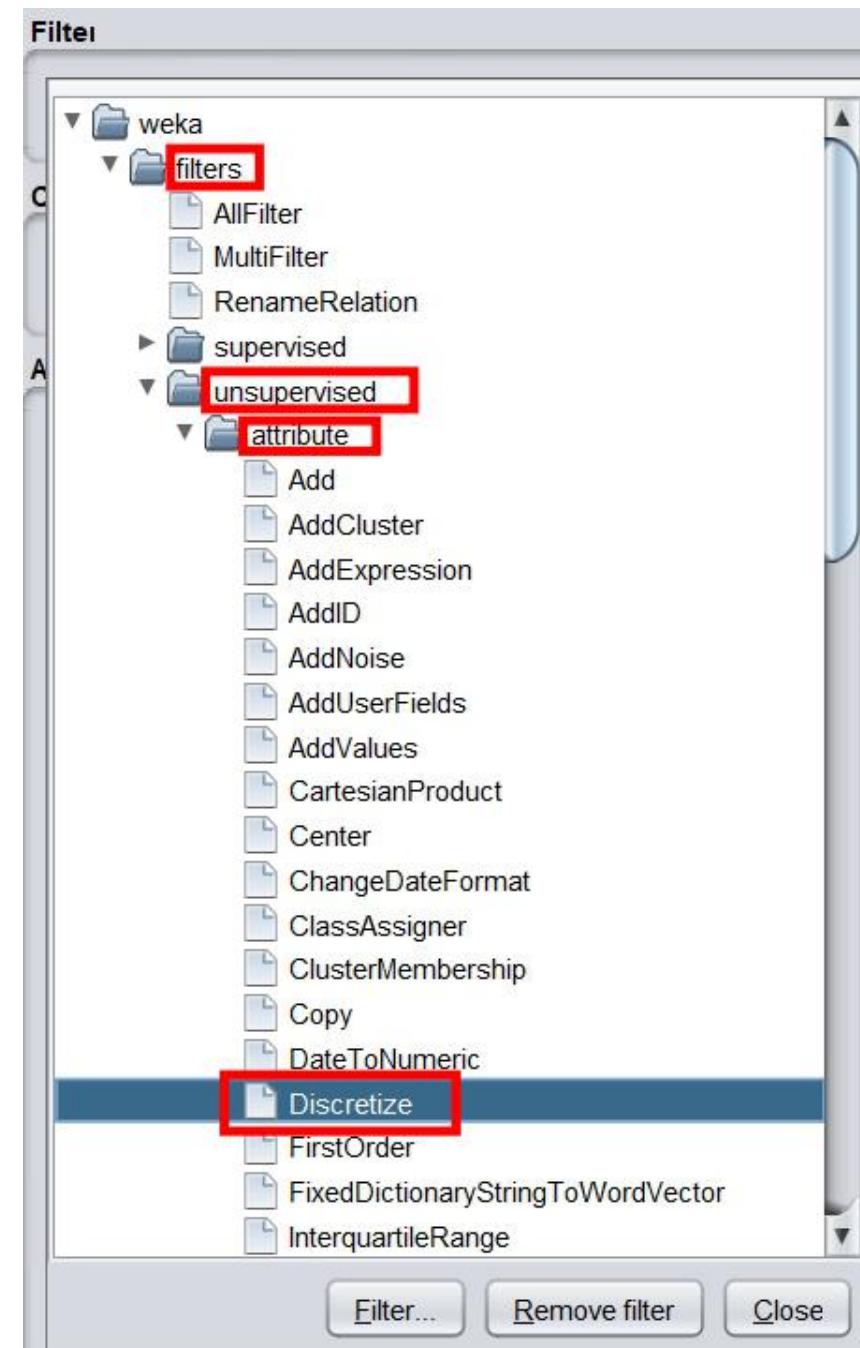
## Lesson 2.1: 離散化數字屬性

6. 在Preprocess面板左鍵單擊Choose按鈕選擇過濾器。



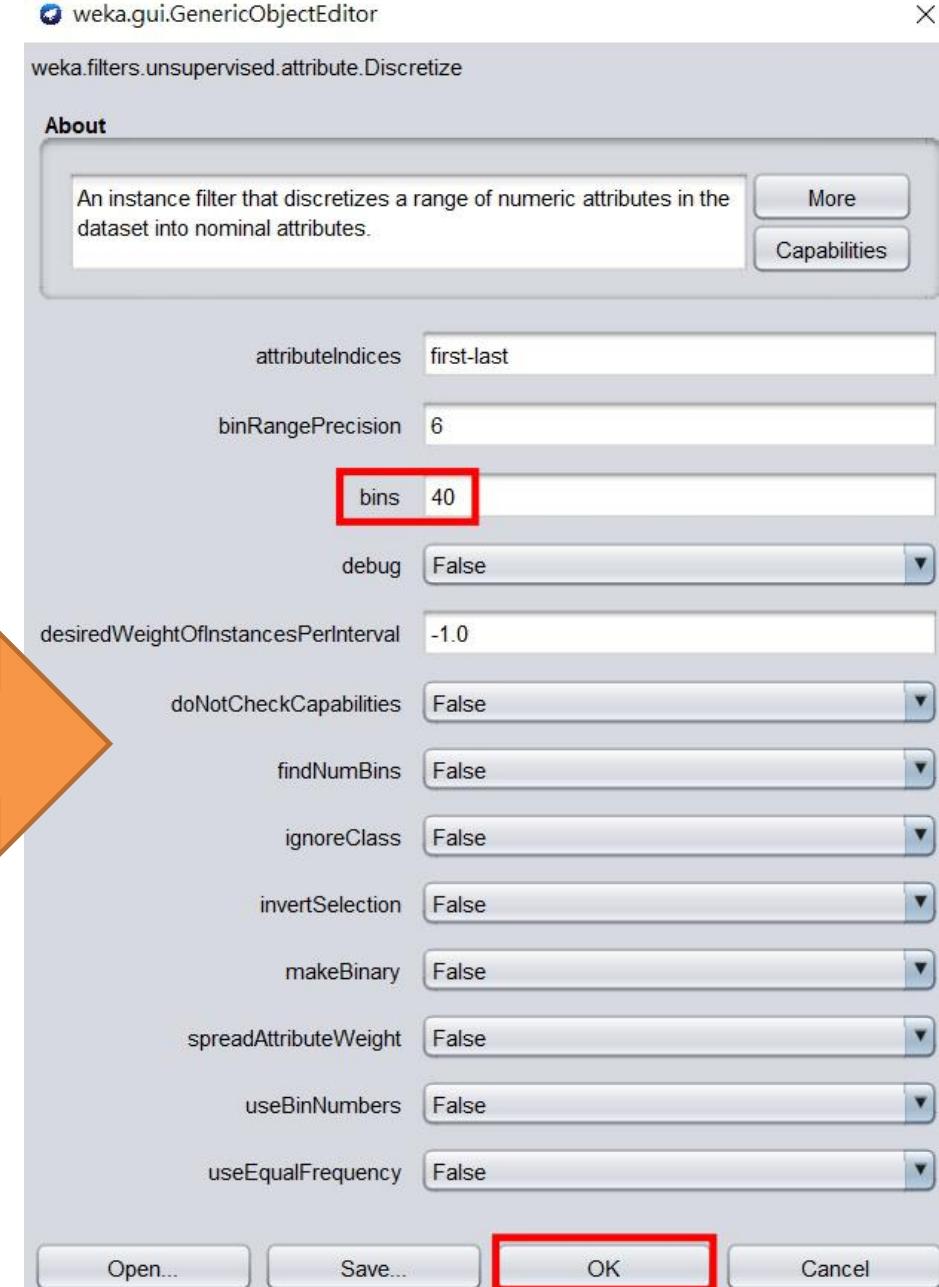
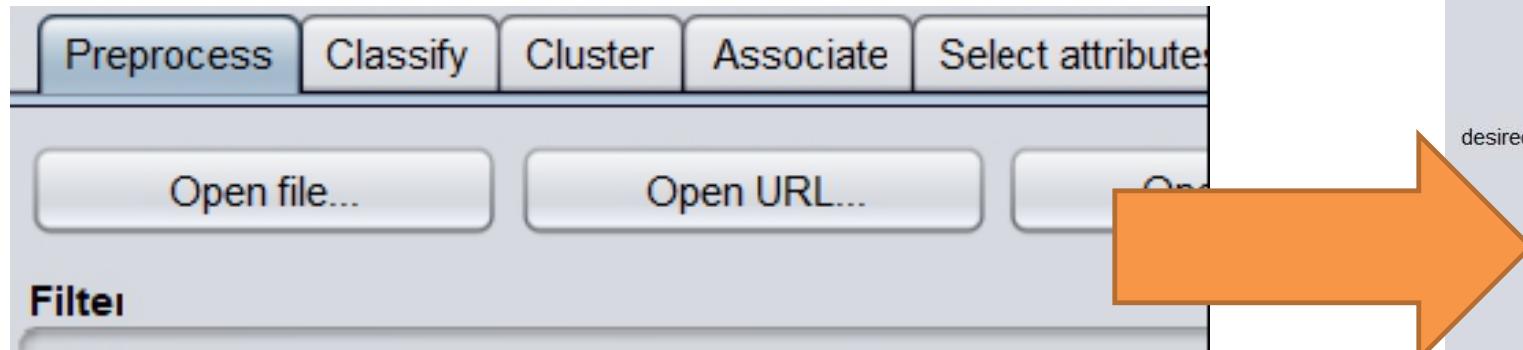
## Lesson 2.1: 離散化數字屬性

7. 左鍵單擊filters資料夾下的unsupervised資料夾下的attribute資料夾下的Dicretize過濾器。



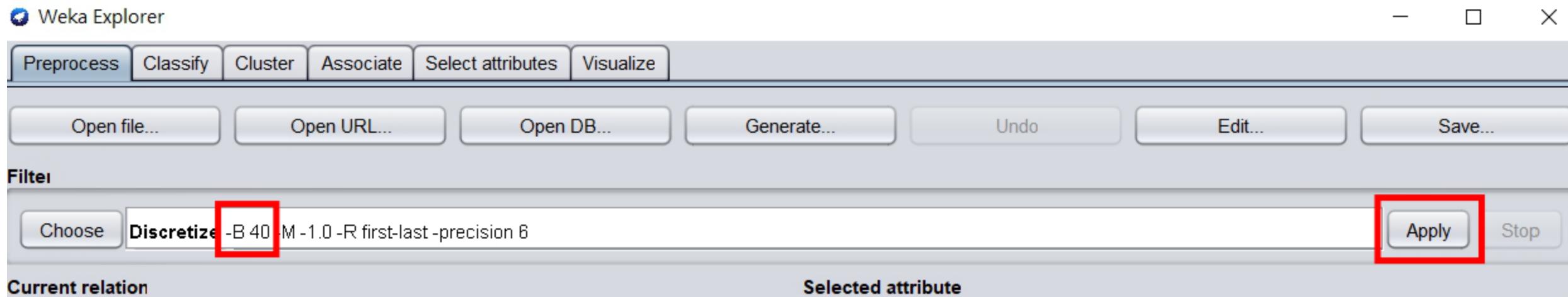
## Lesson 2.1: 離散化數字屬性

8. 左鍵單擊左圖中紅色方框處配置過濾器參數，在出現的視窗裡(右圖)將參數bins後面的輸入框中輸入40，接著按下下方OK按鈕。



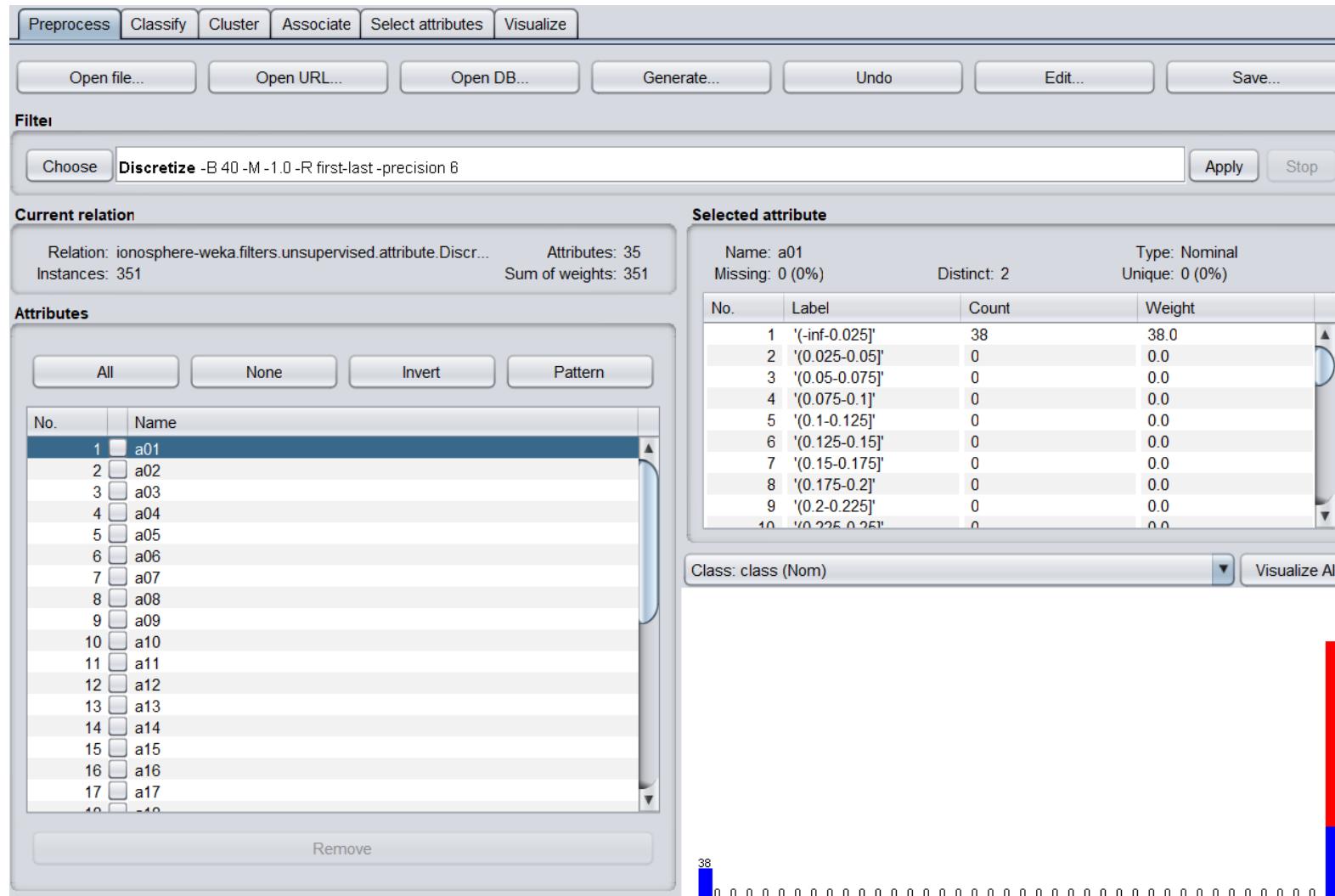
## Lesson 2.1: 離散化數字屬性

9.回到Preprocess面板，確認bins的數量變更為40後，左鍵單擊畫面右側的Apply按鈕套用過濾器。



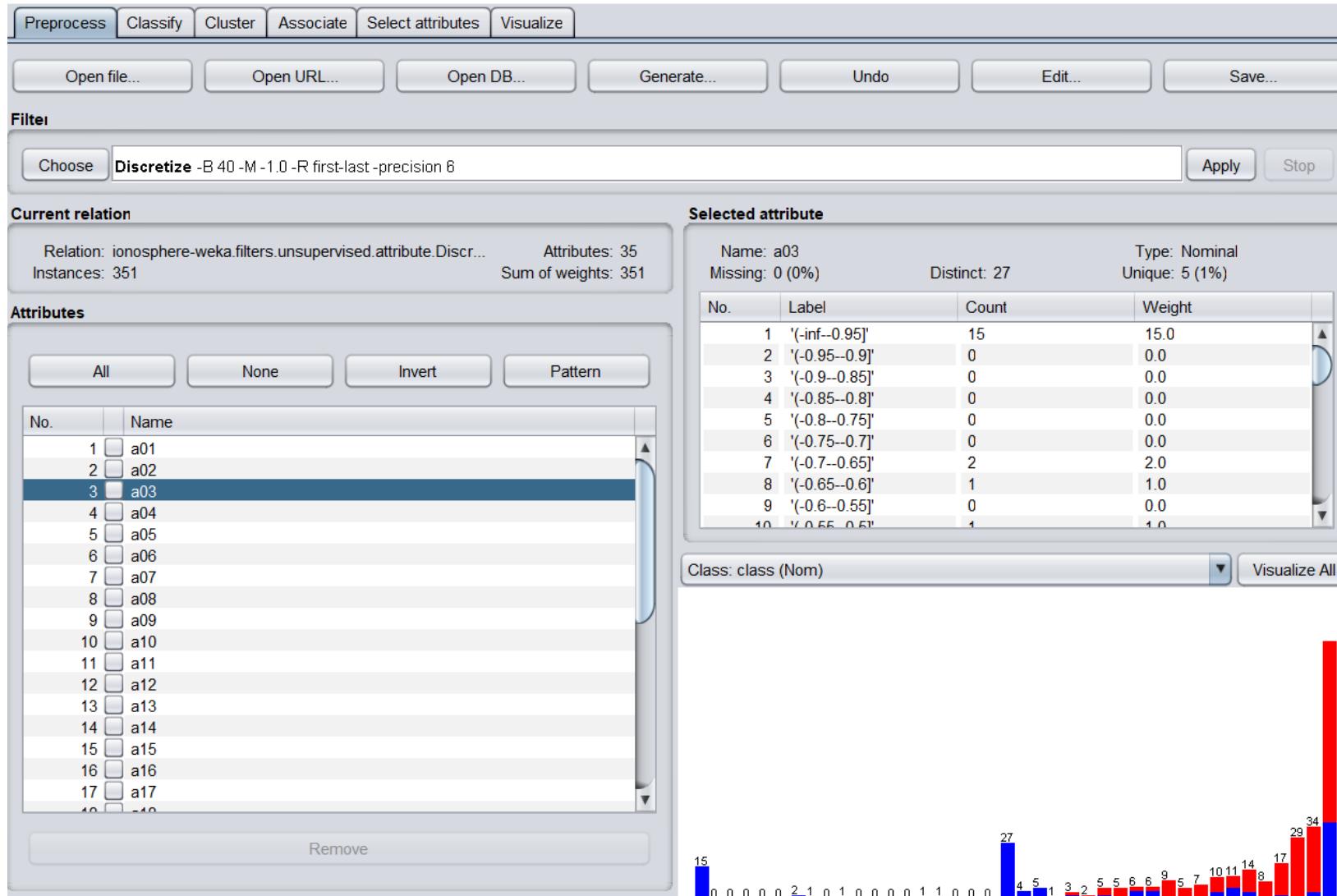
# Lesson 2.1: 離散化數字屬性

▼執行結果：a01。



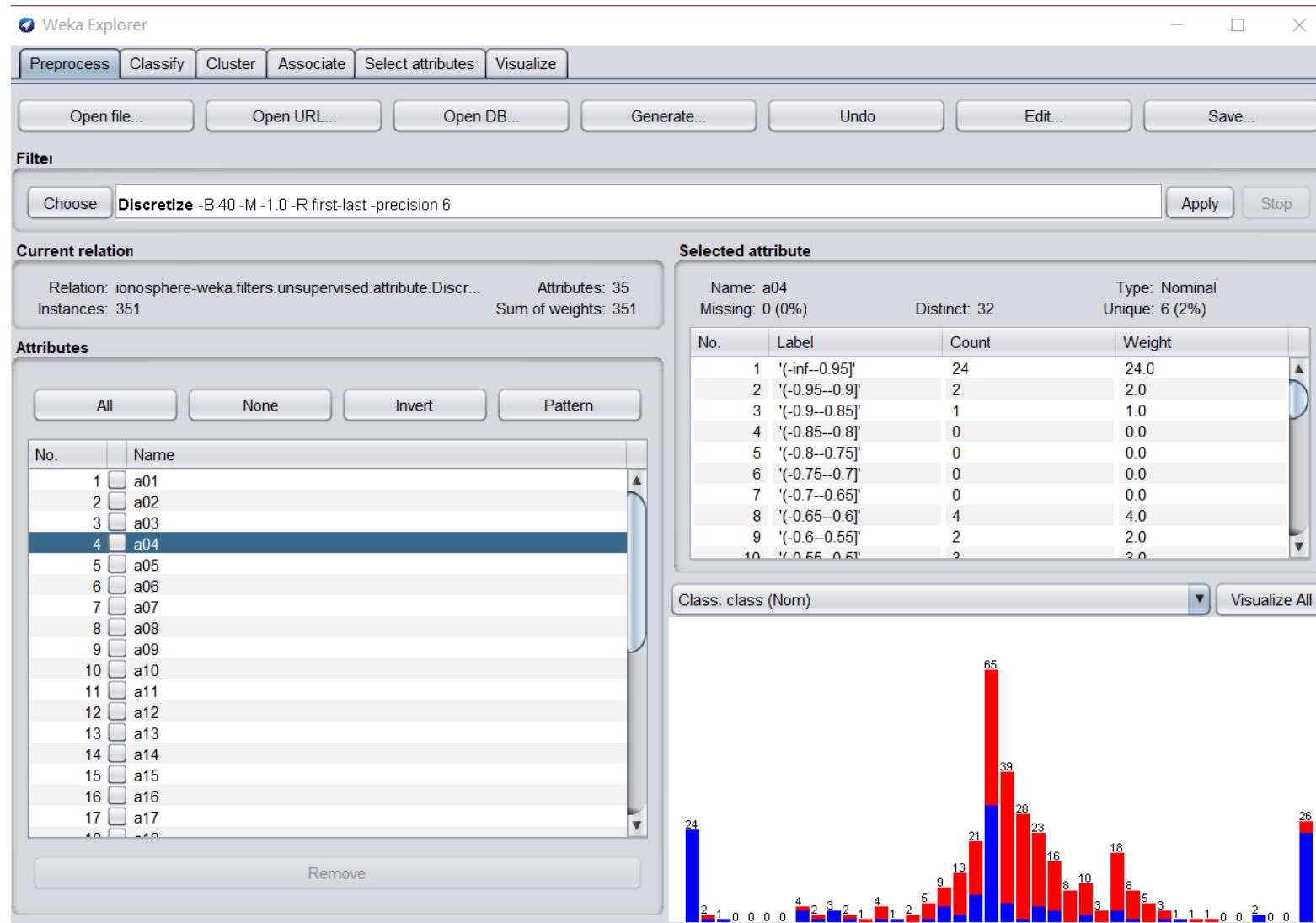
# Lesson 2.1: 離散化數字屬性

▼執行結果：a03。



# Lesson 2.1: 離散化數字屬性

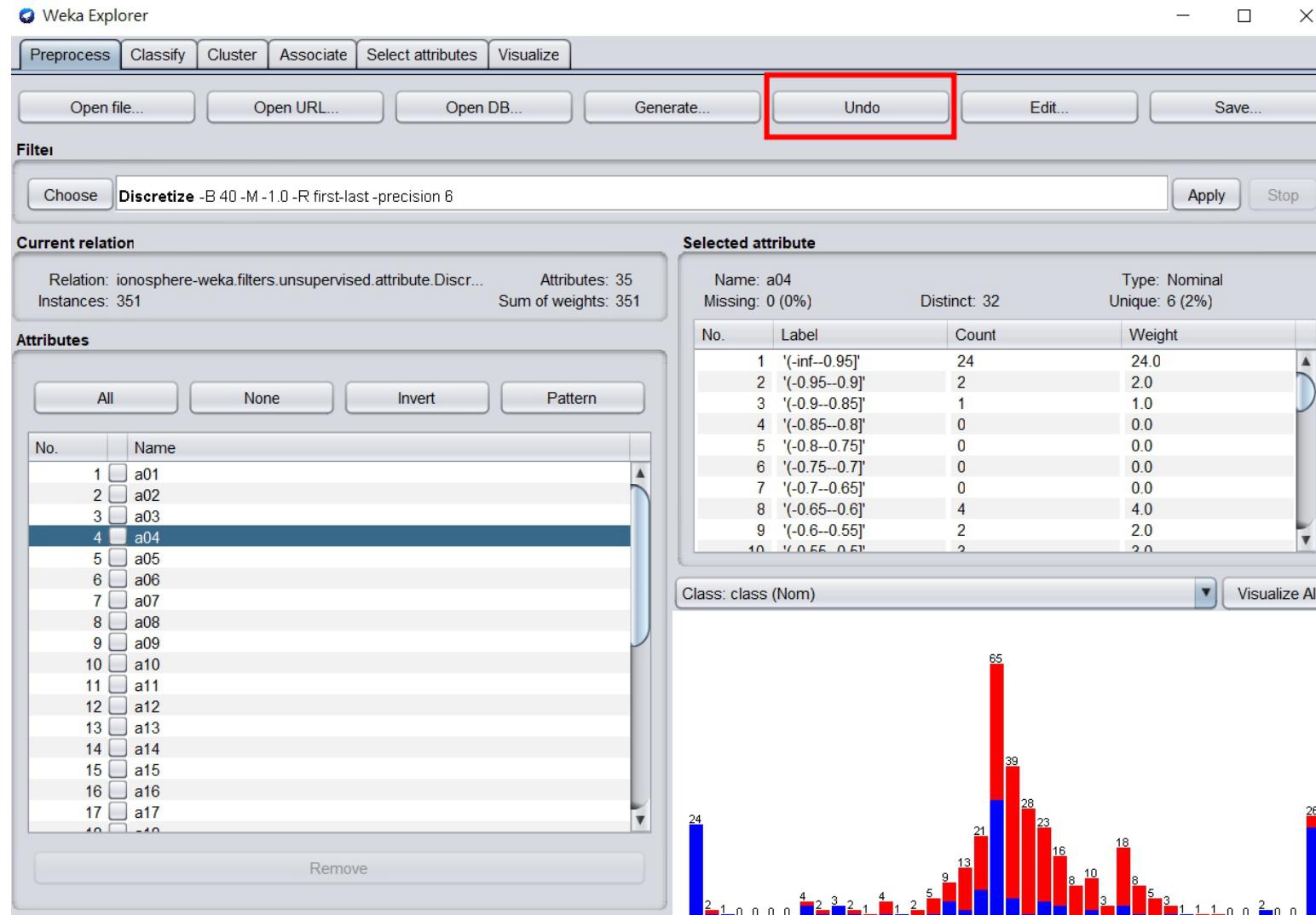
▼ 執行結果：a04。



## Lesson 2.1: 離散化數字屬性

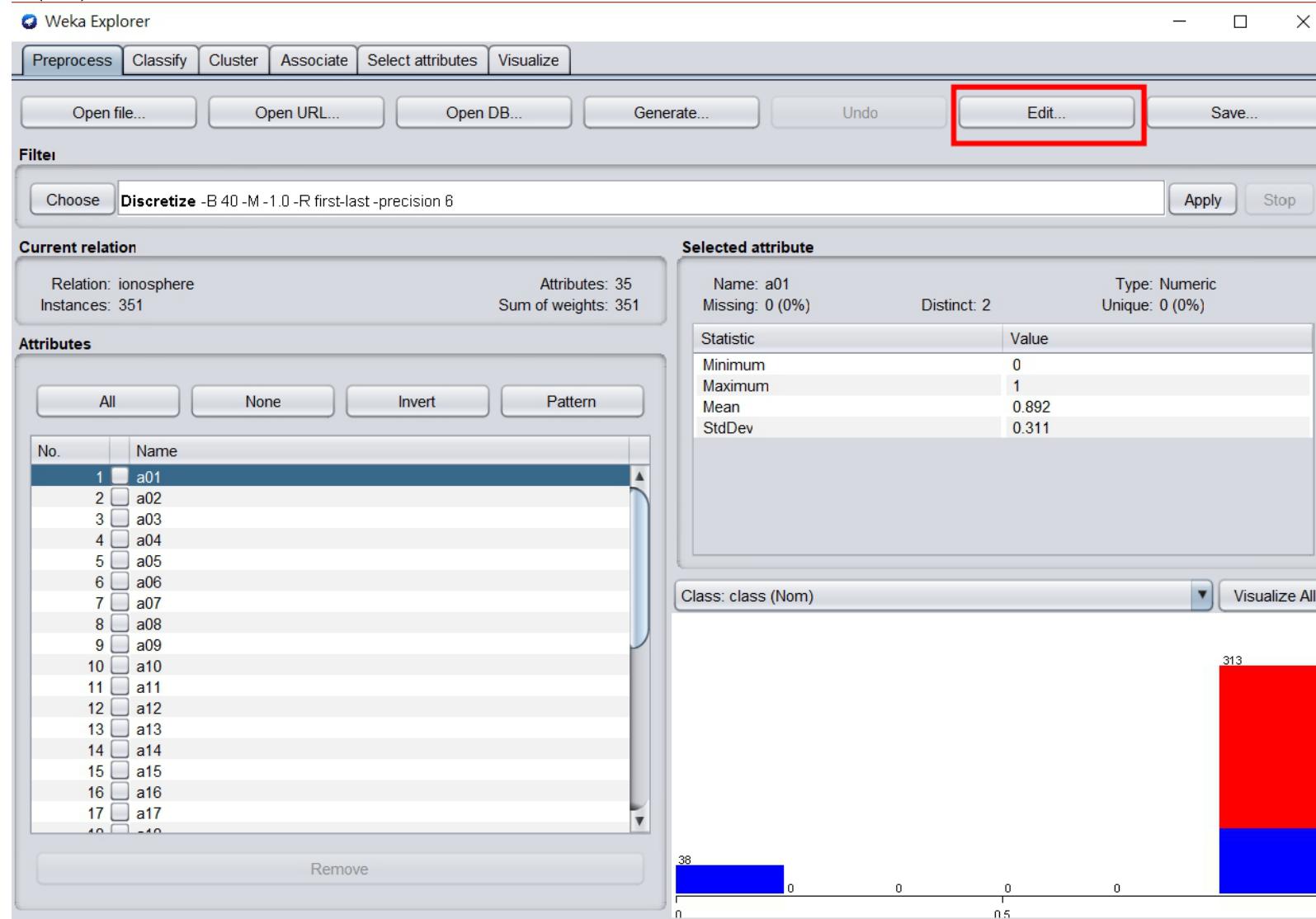
10.我們可以在Edit面板中查看數值。

(1)在Preprocess面板左鍵單擊視窗上方的Undo按鈕，撤銷過濾器的影響。



# Lesson 2.1: 離散化數字屬性

## (2) 左鍵單擊視窗上方的Edit按鈕



## Lesson 2.1: 離散化數字屬性

(3) 左鍵單擊左圖紅色方框處進行排序，可以得到右圖結果。接著，左鍵單擊視窗右上方的關閉按鈕。

Viewer

Relation: ionosphere

No.	1: a01	2: a02	3: a03	4: a04	5: a05	6: a06	7: a07	8: a08	9: a09	10: a10	11: a11	12: a12
	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric
1	1.0	0.0	0.99...	-0.05...	0.85...	0.02...	0.83...	-0.37...	1.0	0.0376	0.85...	-0.17...
2	1.0	0.0	1.0	-0.18...	0.93...	-0.36...	-0.10...	-0.93...	1.0	-0.04...	0.50...	-0.67...
3	1.0	0.0	1.0	-0.03...	1.0	0.00...	1.0	-0.12...	0.88...	0.01...	0.73...	0.05...
4	1.0	0.0	1.0	-0.45...	1.0	1.0	0.71...	-1.0	0.0	0.0	0.0	0.0
5	1.0	0.0	1.0	-0.02...	0.9414	0.06...	0.92...	-0.23...	0.77...	-0.16...	0.52...	-0.20...
6	1.0	0.0	0.02...	-0.00...	-0.09...	-0.11...	-0.00...	-0.11...	0.14...	0.06...	0.03...	-0.06...
7	1.0	0.0	0.97...	-0.10...	0.94...	-0.208	0.92...	-0.28...	0.85...	-0.27...	0.79...	-0.47...
8	0.0	0.0	0.0	0.0	0.0	1.0	-1.0	0.0	0.0	-1.0	-1.0	0.0
9	1.0	0.0	0.96...	-0.07...	1.0	-0.14...	1.0	-0.21...	1.0	-0.36...	0.9257	-0.43...
10	1.0	0.0	-0.01...	-0.08...	0.0	0.0	0.0	0.1147	-0.26...	-0.45...	-0.38...	0.0
11	1.0	0.0	1.0	0.06...	1.0	-0.18...	1.0	-0.27...	1.0	-0.43...	1.0	0.0
12	1.0	0.0	1.0	-0.54...	1.0	-1.0	1.0	-1.0	1.0	0.36...	1.0	0.0
13	1.0	0.0	1.0	-0.16...	1.0	-0.10...	0.99...	-0.15...	1.0	-0.19...	0.94...	0.0
14	1.0	0.0	1.0	-0.86...	1.0	0.2228	0.85...	-0.39...	1.0	-0.12...	1.0	-0.88...
15	1.0	0.0	1.0	0.0738	1.0	0.0342	1.0	-0.05...	1.0	0.08...	1.0	0.19...
16	1.0	0.0	0.50...	-0.93...	1.0	0.26...	-0.03...	-1.0	1.0	-1.0	0.43...	-1.0
17	1.0	0.0	0.99...	0.06...	1.0	-0.01...	0.97...	0.02...	0.96...	0.02...	0.99...	0.07...
18	0.0	0.0	0.0	-1.0	-1.0	1.0	1.0	-1.0	1.0	-1.0	1.0	1.0
19	1.0	0.0	0.67...	0.02...	0.66...	0.05...	0.57...	0.18...	0.08...	0.34...	0.63...	0.12...
20	0.0	0.0	1.0	-1.0	0.0	0.0	0.0	1.0	1.0	1.0	-1.0	0.0
21	1.0	0.0	1.0	-0.00...	1.0	-0.09...	1.0	-0.07...	1.0	-0.10...	1.0	-0.11...
22	0.0	0.0	1.0	1.0	0.0	0.0	0.0	-1.0	-1.0	0.0	0.0	0.0
23	1.0	0.0	0.96...	0.07...	1.0	0.04...	1.0	0.09...	0.90...	-0.05...	0.89...	0.0258
24	0.0	0.0	-1.0	1.0	0.0	0.0	0.0	-1.0	1.0	1.0	1.0	1.0
25	1.0	0.0	1.0	-0.06...	1.0	0.02...	1.0	-0.05...	1.0	-0.01...	1.0	-0.11...
26	1.0	0.0	1.0	0.5782	1.0	-1.0	1.0	-1.0	1.0	-1.0	1.0	-1.0
27	1.0	0.0	1.0	-0.08...	1.0	-0.17...	0.86...	-0.81...	0.94...	0.61...	0.95...	-0.41...
28	0.0	0.0	-1.0	-1.0	0.0	0.0	-1.0	1.0	1.0	-0.375	0.0	0.0

Add instance Undo OK Cancel

Viewer

Relation: ionosphere

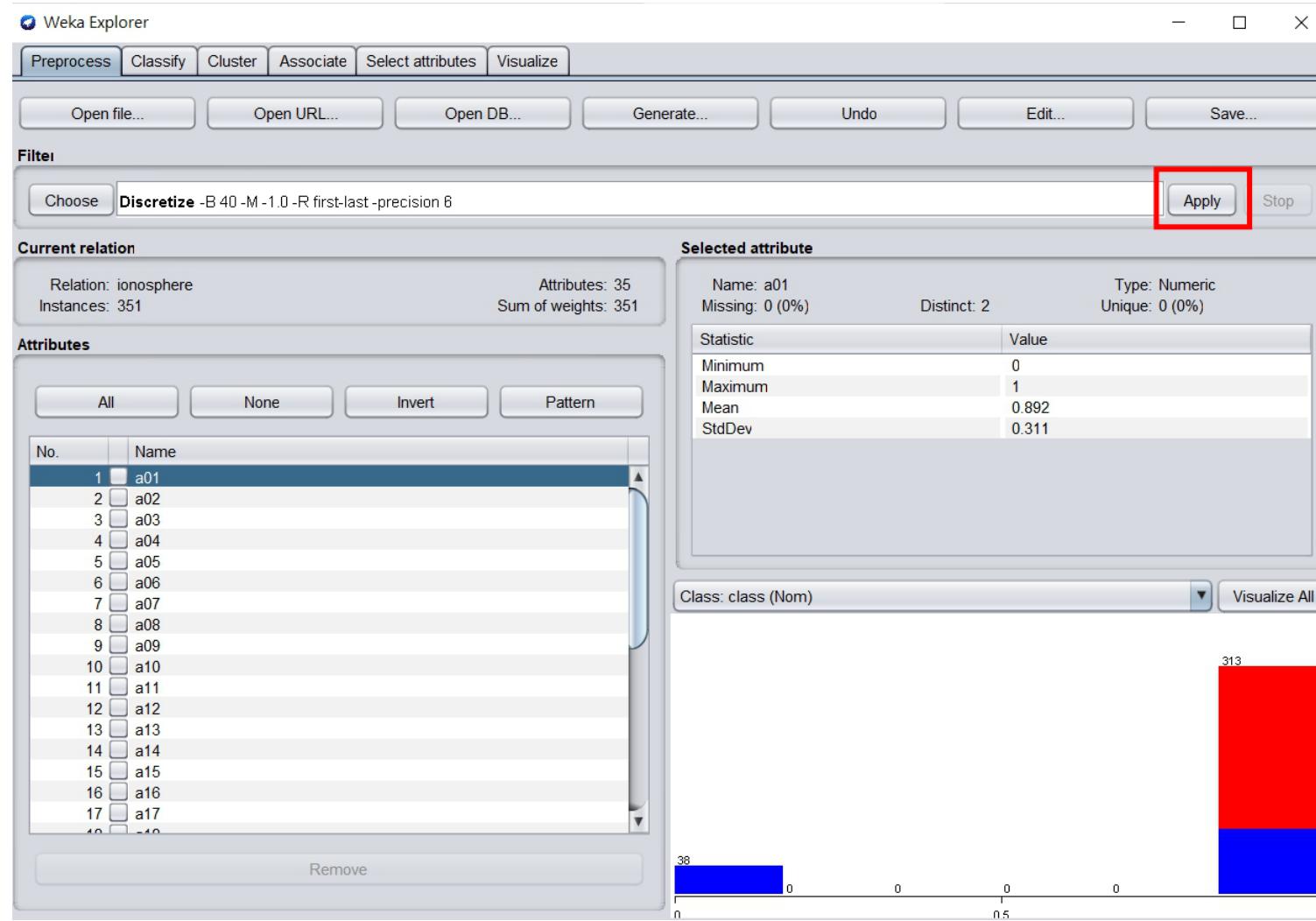
No.	1: a01	2: a02	3: a03	4: a04	5: a05	6: a06	7: a07	8: a08	9: a09	10: a10	11: a11	12: a12
	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric
1	0.0	0.0	1.0	-1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0
2	0.0	0.0	-1.0	-1.0	0.0	0.0	-1.0	1.0	1.0	-1.0	1.0	1.0
3	0.0	0.0	-1.0	-1.0	-1.0	-1.0	1.0	1.0	1.0	-1.0	-1.0	-1.0
4	1.0	0.0	1.0	-1.0	-1.0	-1.0	1.0	1.0	1.0	-1.0	-1.0	1.0
5	1.0	0.0	0.66...	-1.0	-1.0	1.0	1.0	1.0	1.0	-1.0	-0.67...	0.80...
6	1.0	0.0	-1.0	-1.0	-1.0	1.0	1.0	1.0	1.0	-0.14...	-0.14...	0.0
7	1.0	0.0	-0.64...	-1.0	-1.0	1.0	0.82...	1.0	1.0	-1.0	-1.0	1.0
8	1.0	0.0	-0.67...	-1.0	-1.0	1.0	-1.0	1.0	1.0	0.63...	0.03...	0.0
9	1.0	0.0	0.17...	-1.0	-1.0	1.0	-1.0	1.0	0.0	0.0	0.0	0.0
10	0.0	0.0	1.0	-1.0	-1.0	1.0	-1.0	1.0	-1.0	1.0	-1.0	1.0
11	0.0	0.0	1.0	-1.0	-1.0	1.0	1.0	1.0	-1.0	-1.0	-1.0	1.0
12	0.0	0.0	1.0	-1.0	-1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
13	1.0	0.0	-1.0	-1.0	-1.0	0.0	0.0	0.50...	-0.78...	0.60...	-0.78...	0.60...
14	1.0	0.0	1.0	-1.0	-1.0	0.0	0.0	0.77...	-0.99...	0.80...	-0.99...	0.80...
15	1.0	0.0	1.0	-1.0	-1.0	1.0	1.0	-1.0	1.0	-1.0	1.0	1.0
16	1.0	0.0	-1.0	-1.0	-1.0	-1.0	-0.50...	1.0	1.0	-1.0	-1.0	1.0
17	0.0	0.0	-1.0	-1.0	-1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
18	1.0	0.0	1.0	-1.0	-1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
19	1.0	0.0	0.36...	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	1.0
20	0.0	0.0	-1.0	-1.0	-1.0	1.0	-1.0	1.0	-1.0	-1.0	1.0	0.0
21	0.0	0.0	1.0	-1.0	-1.0	1.0	1.0	1.0	1.0	1.0	-1.0	1.0
22	0.0	0.0	1.0	-1.0	-1.0	-1.0	-1.0	1.0	-1.0	-1.0	1.0	1.0
23	0.0	0.0	1.0	-1.0	-1.0	1.0	1.0	1.0	1.0	-1.0	1.0	1.0
24	0.0	0.0	1.0	-1.0	-1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
25	1.0	0.0	0.50...	-0.93...	-1.0	-0.26...	-0.03...	1.0	0.26...	-0.03...	-1.0	1.0
26	1.0	0.0	1.0	-0.92...	-1.0	0.75...	0.49...	-0.05...	0.62...	0.62...	0.62...	1.0
27	1.0	0.0	1.0	-0.86...	-1.0	0.2228	0.85...	-0.39...	1.0	0.2228	0.85...	1.0
28	1.0	0.0	-1.0	-1.0	-1.0	-0.86	1.0	0.2228	0.85...	-0.39...	1.0	1.0

Add instance Undo OK Cancel



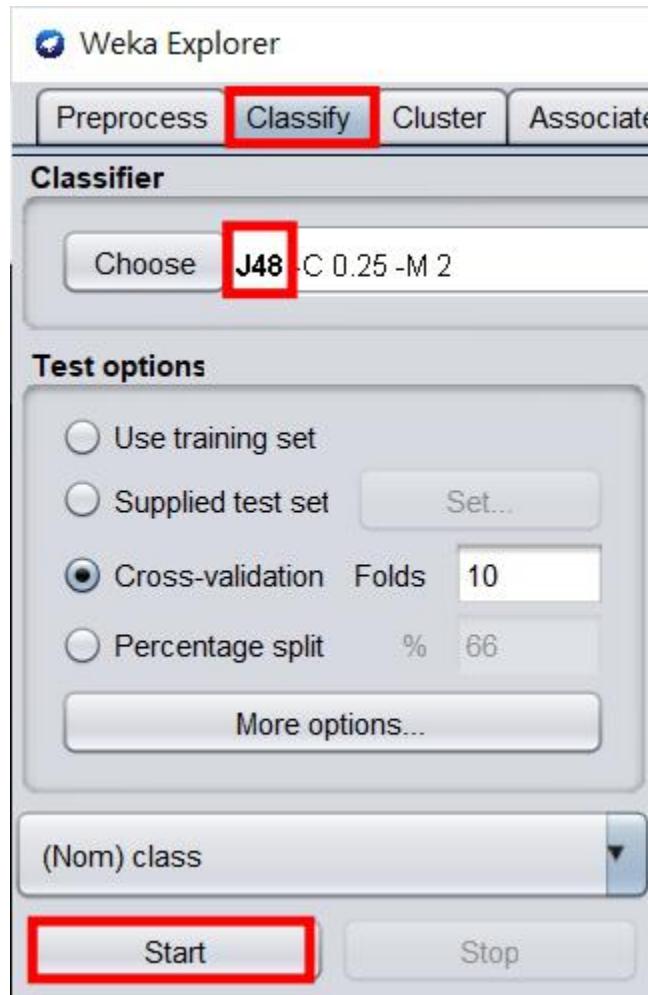
## Lesson 2.1: 離散化數字屬性

11.回到Preprocess界面左鍵單擊Apply按鈕，運行剛才為了查看資料而取消的分類器。



## Lesson 2.1: 離散化數字屬性

12. 切換到Classify界面，確定分類器為J48後左鍵單擊Start按鈕。



# Lesson 2.1: 離散化數字屬性

▼得到87.7493%準確率

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)**

16:07:08 - trees.J48  
16:11:39 - trees.J48

**Classifier output**

```
Time taken to build model: 0.01 seconds
===
Stratified cross-validation ===
===
Summary ===

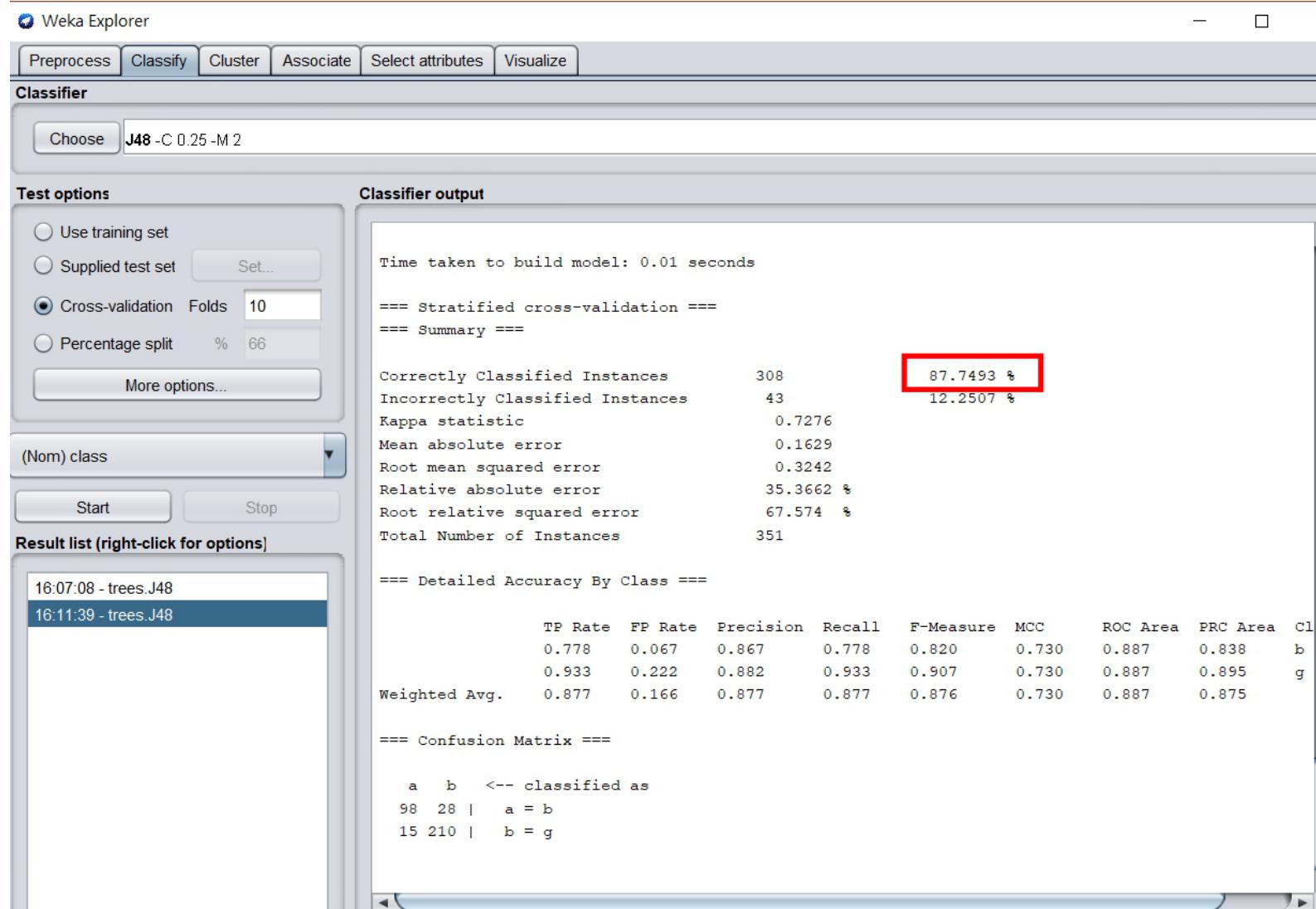
Correctly Classified Instances      308
Incorrectly Classified Instances   43
Kappa statistic                   0.7276
Mean absolute error               0.1629
Root mean squared error           0.3242
Relative absolute error           35.3662 %
Root relative squared error      67.574 %
Total Number of Instances         351

===
Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall   F-Measure  MCC    ROC Area  PRC Area  Cl
          0.778    0.067    0.867     0.778    0.820     0.730   0.887    0.838    b
          0.933    0.222    0.882     0.933    0.907     0.730   0.887    0.895    g
Weighted Avg.    0.877    0.166    0.877     0.877    0.876     0.730   0.887    0.875

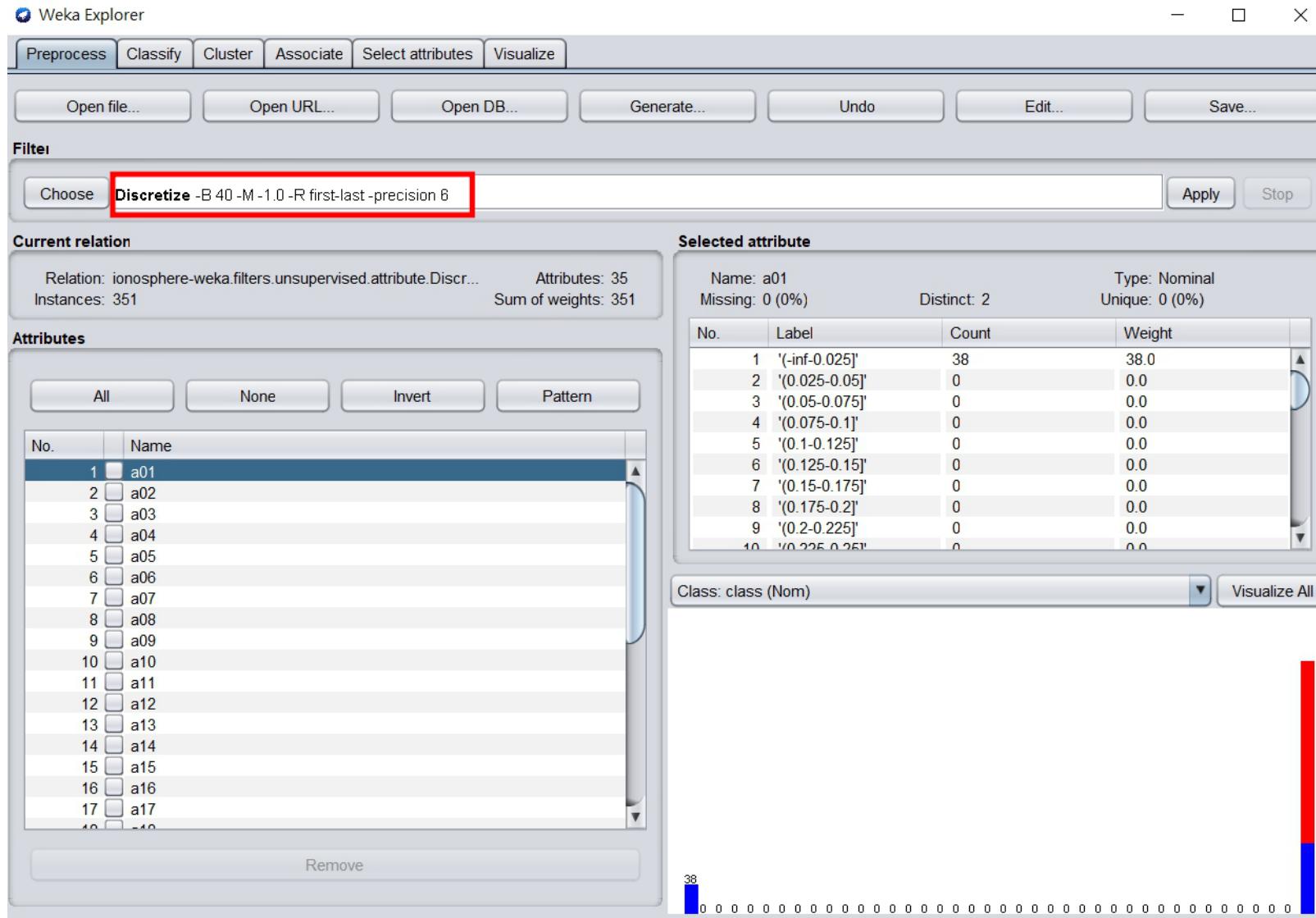
===
Confusion Matrix ===

  a   b   <-- classified as
98  28 |  a = b
15 210 |  b = g
```



## Lesson 2.1: 離散化數字屬性

13.回到Preprocess面板，左鍵單擊圖中紅色方框處配置分類器。



## Lesson 2.1: 離散化數字屬性

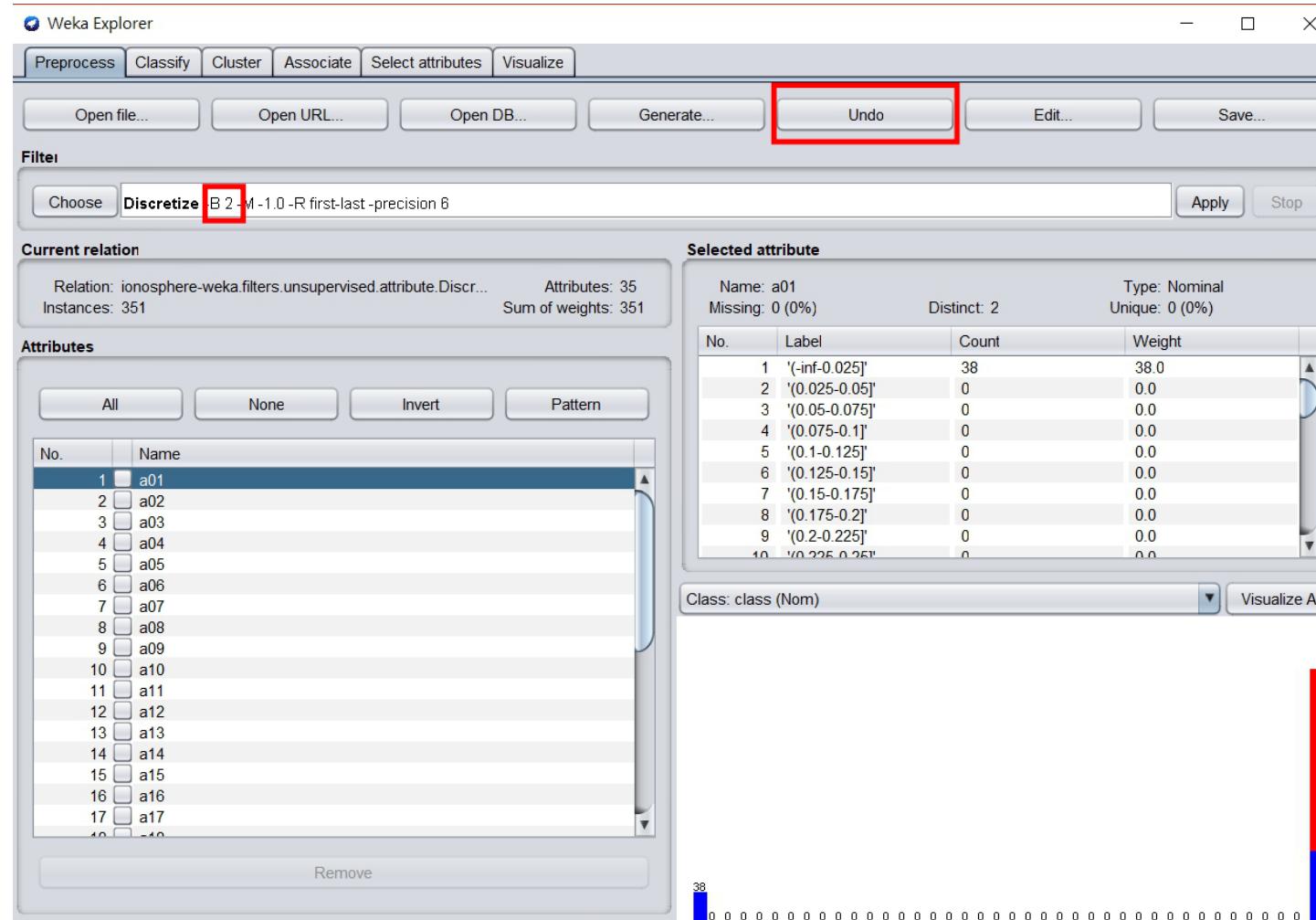
這次我們將bins的參數設定為2。

14.在bins參數後的輸入框中輸入2，  
然後左鍵單擊下方OK按鈕。



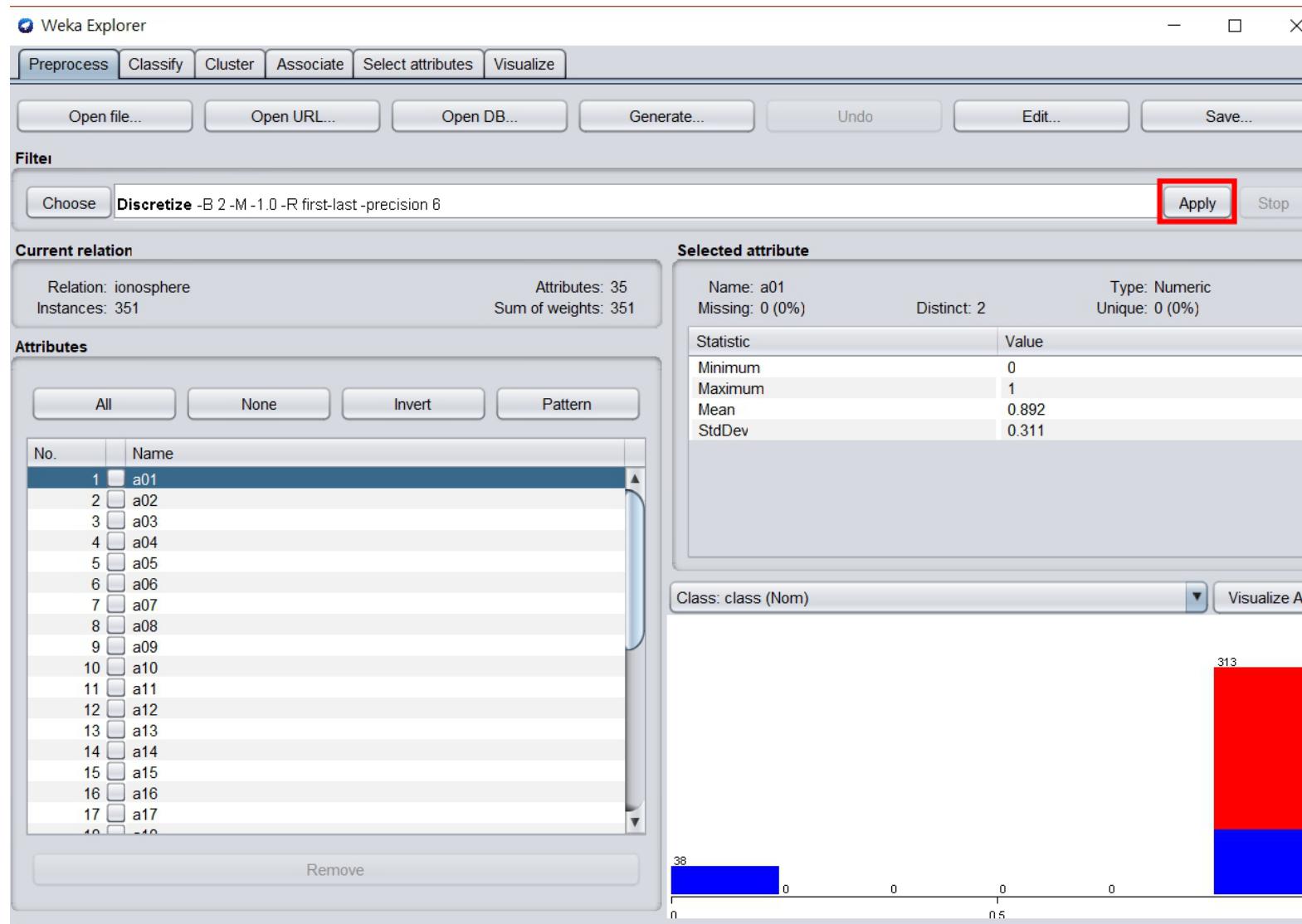
## Lesson 2.1: 離散化數字屬性

15. 確認bin參數值被設定為2後，按下Undo按鈕撤銷剛才bins參數為40的過濾器。



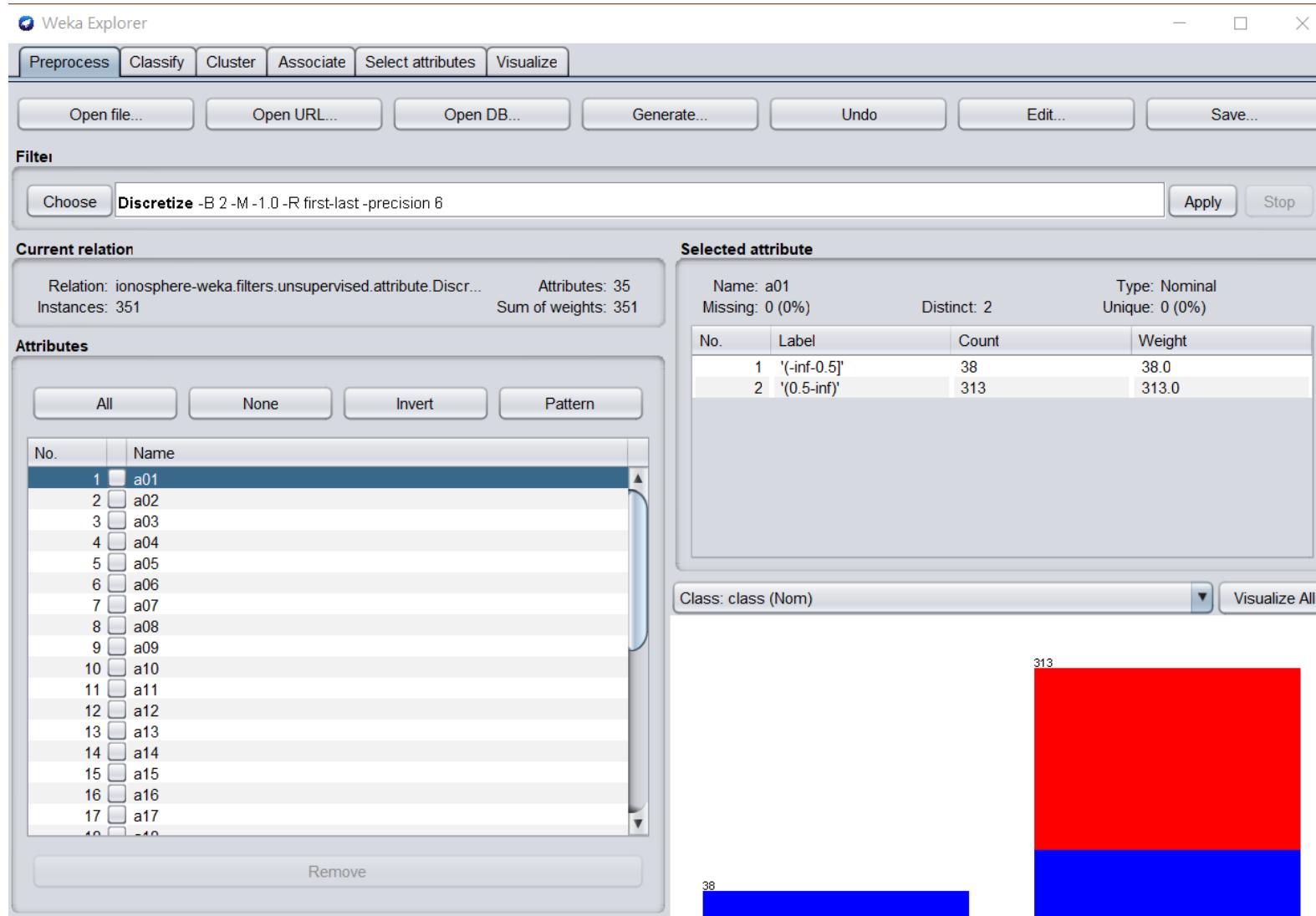
## Lesson 2.1: 離散化數字屬性

### 16. 左鍵單擊Apply按鈕套用剛才配置的過濾器。



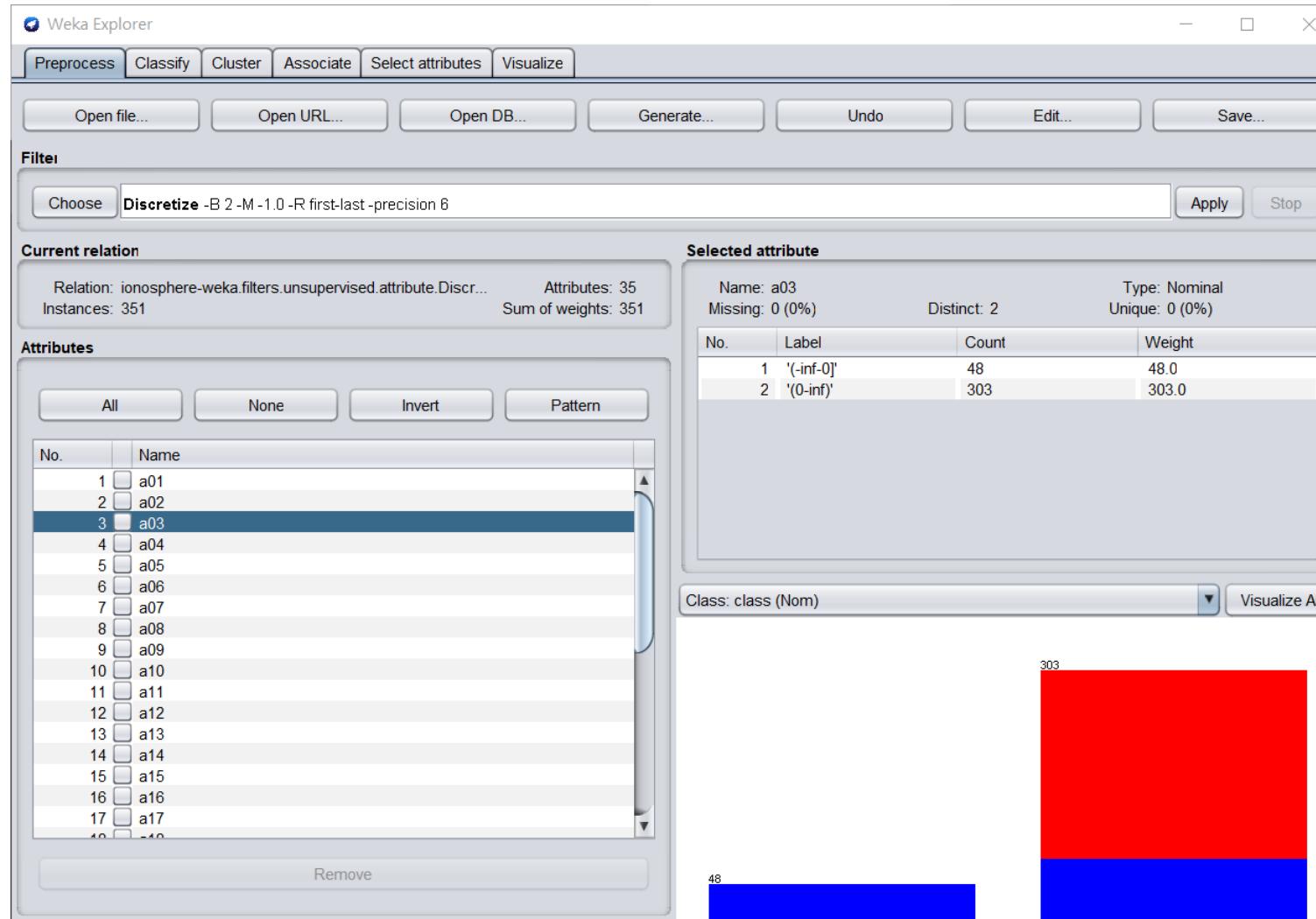
# Lesson 2.1: 離散化數字屬性

▼執行結果：a01。



# Lesson 2.1: 離散化數字屬性

▼執行結果：a03。



# Lesson 2.1: 離散化數字屬性

▼執行結果：a04。

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

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

Filter

Choose Discretize -B 2 -M -1.0 -R first-last -precision 6 Apply Stop

Current relation

Relation: ionosphere-weka.filters.unsupervised.attribute.Discretize Attributes: 35 Instances: 351 Sum of weights: 351

Selected attribute

Name: a04 Missing: 0 (0%) Distinct: 2 Unique: 0 (0%) Type: Nominal

No.	Label	Count	Weight
1	'(-inf-0]'	159	159.0
2	'(0-inf)'	192	192.0

Attributes

All None Invert Pattern

No.	Name
1	a01
2	a02
3	a03
4	a04
5	a05
6	a06
7	a07
8	a08
9	a09
10	a10
11	a11
12	a12
13	a13
14	a14
15	a15
16	a16
17	a17
18	-10

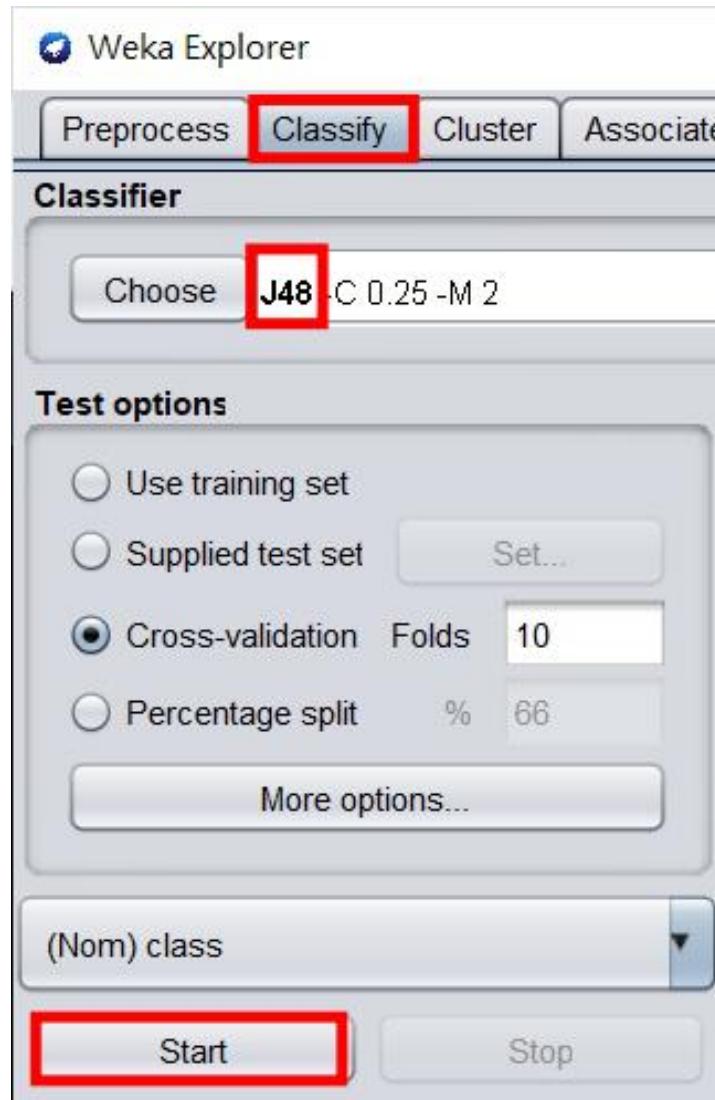
Remove

Class: class (Nom) Visualize All

The Weka Explorer interface shows the results of discretizing attribute a04. The 'Selected attribute' panel displays the distribution of values into two categories: '(-inf-0]' with 159 instances and '(0-inf)' with 192 instances. The 'Attributes' panel lists all attributes, with a04 selected. A bar chart at the bottom visualizes this distribution.

## Lesson 2.1: 離散化數字屬性

17.切換到Classify界面確定分類器為J48後，左鍵單擊Start按鈕。



## Lesson 2.1: 離散化數字屬性

▼執行結果：得到90.8832%準確率。

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)**

16:07:08 - trees.J48  
16:11:39 - trees.J48  
16:17:32 - trees.J48

**Classifier output**

```
Time taken to build model: 0.02 seconds

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

Correctly Classified Instances      319      90.8832 %
Incorrectly Classified Instances   32       9.1168 %
Kappa statistic                   0.7925
Mean absolute error               0.1493
Root mean squared error          0.2886
Relative absolute error          32.4268 %
Root relative squared error     60.1626 %
Total Number of Instances        351

==== Detailed Accuracy By Class ===

           TP Rate  FP Rate  Precision  Recall   F-Measure  MCC    ROC Area  PRC Area  Cl
           0.770    0.013    0.970     0.770    0.858     0.804    0.862    0.856    b
           0.987    0.230    0.884     0.987    0.933     0.804    0.862    0.859    g
Weighted Avg.    0.909    0.152    0.915     0.909    0.906     0.804    0.862    0.858

==== Confusion Matrix ===

      a   b  <- classified as
97  29 |  a = b
 3 222 |  b = g
```

## Lesson 2.1: 離散化數字屬性

### 等份裝箱 (Equal-width binning)

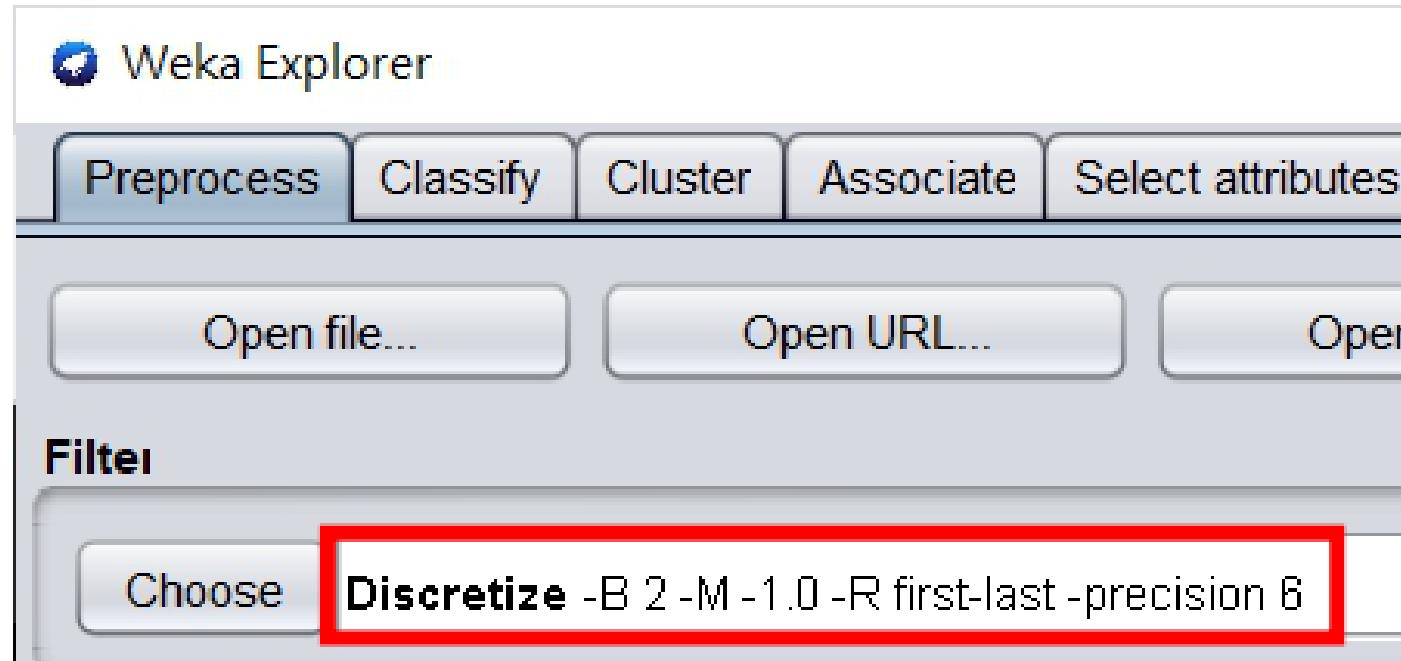
- ❖ 開啟`ionosphere.arff`; 使用 **J48**                            **91.5% (35個節點)**
  - *a01*: -1 值 (38個) 和 +1 值 (313個) [可以使用 [Edit... 查看](#)]
  - *a03*: 逐漸增大
  - *a04*: 常態分布?
- ❖ `unsupervised>attribute>discretize`: 檢驗參數
- ❖ **40 bins**; 所有屬性; 查看數值                            **87.7% (81個節點)**
  - *a01*:
  - *a03*:
  - *a04*: 查看有著一些 -1 和 +1's 的名詞
- ❖ **10 bins**    **86.6% (51個節點)**
- ❖ **5 bins**    **90.6% (46個節點)**
- ❖ **2 bins**    **90.9% (13個節點)**

可以看到最後一行 90.9% 的準確率，幾乎不亞於未離散的數值的準確率。而且，樹只有 13 個節點。比我們之前的樹，小得多也經濟得多，而且準確率也沒怎麼降低。

## Lesson 2.1: 離散化數字屬性

再來我們試試等頻裝箱。

1. 切換到Preprocess面板，左鍵單擊紅色方框處進行過濾器配置。



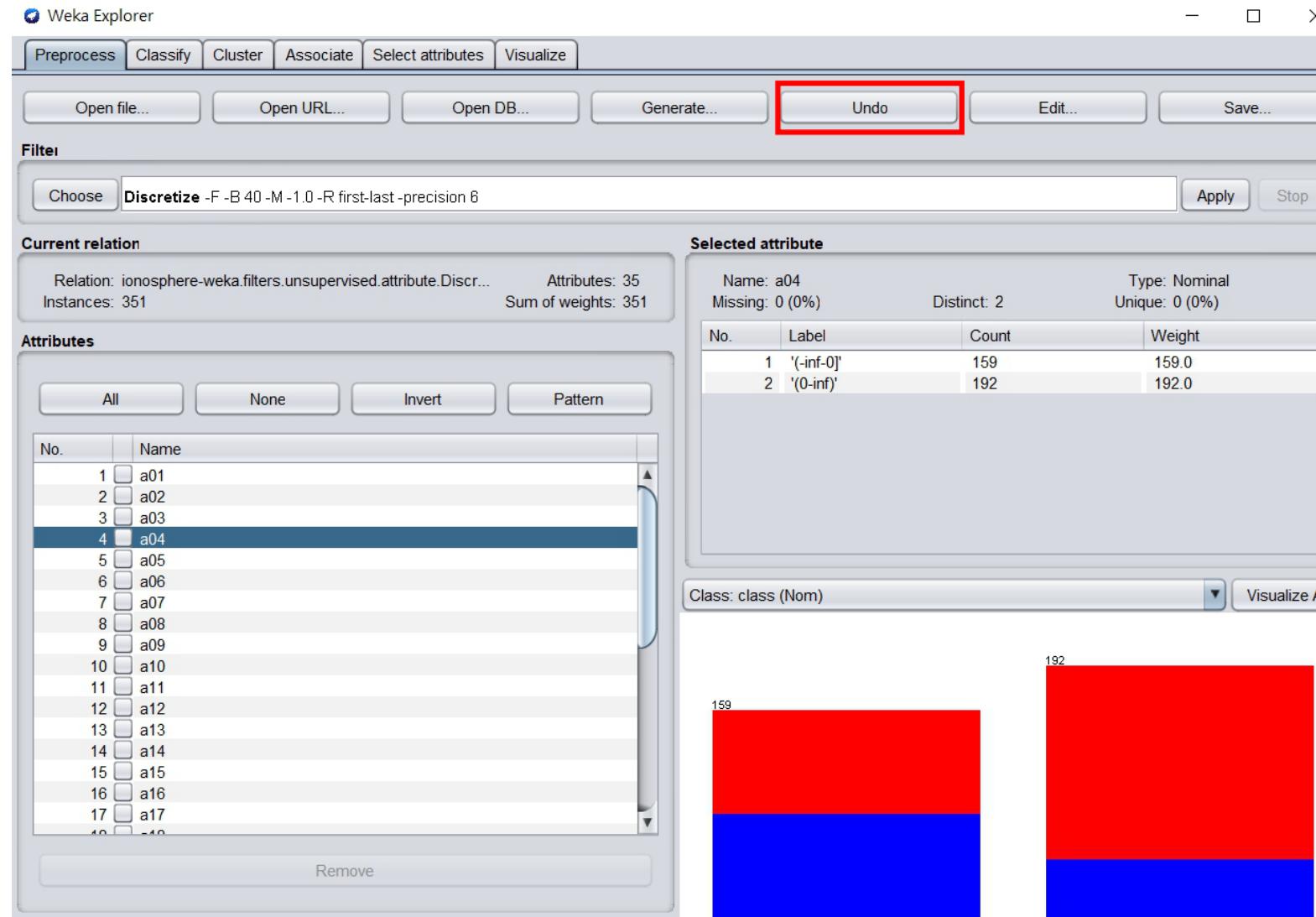
## Lesson 2.1: 離散化數字屬性

2. 在配置視窗中的**bins**參數後的輸入框輸入40，並將**useEqualFrequency**參數設定為True，然後左鍵單擊視窗下方OK按鈕。



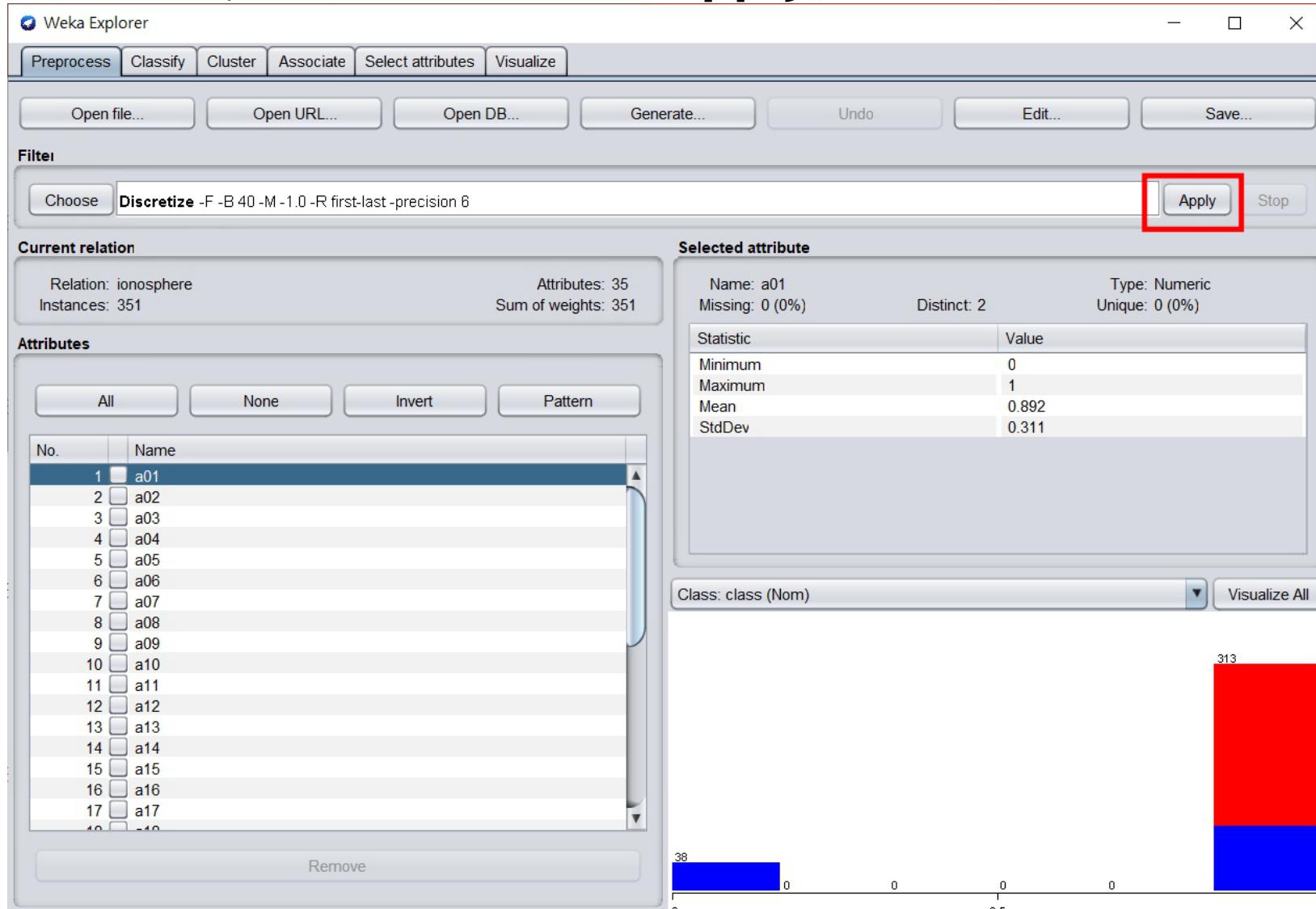
## Lesson 2.1: 離散化數字屬性

3. 左鍵單擊視窗上方 Undo 按鈕撤銷剛才的過濾器套用。



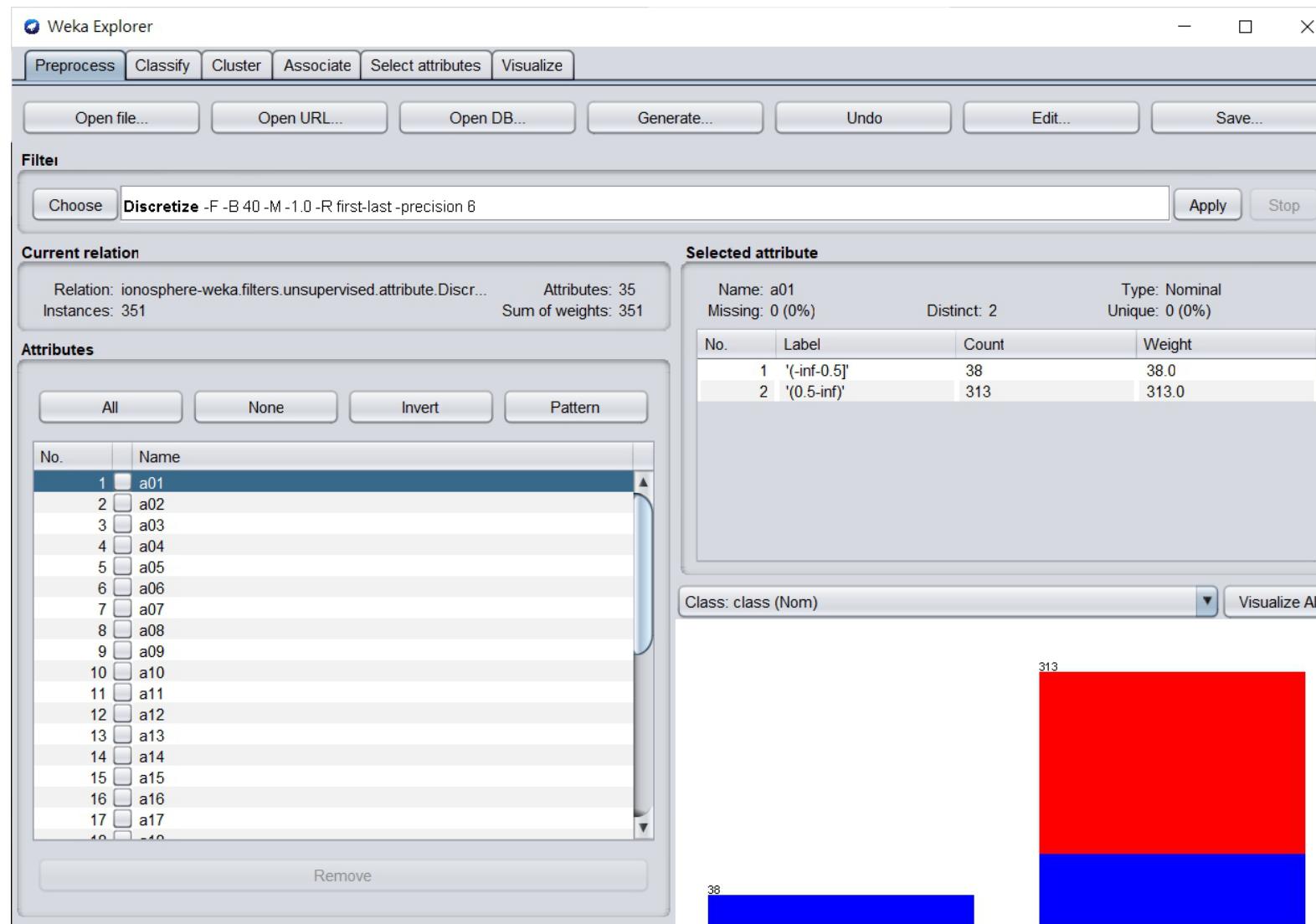
## Lesson 2.1: 離散化數字屬性

4. 左鍵單擊視窗右上方的Apply按鈕，套用剛配置的過濾器。



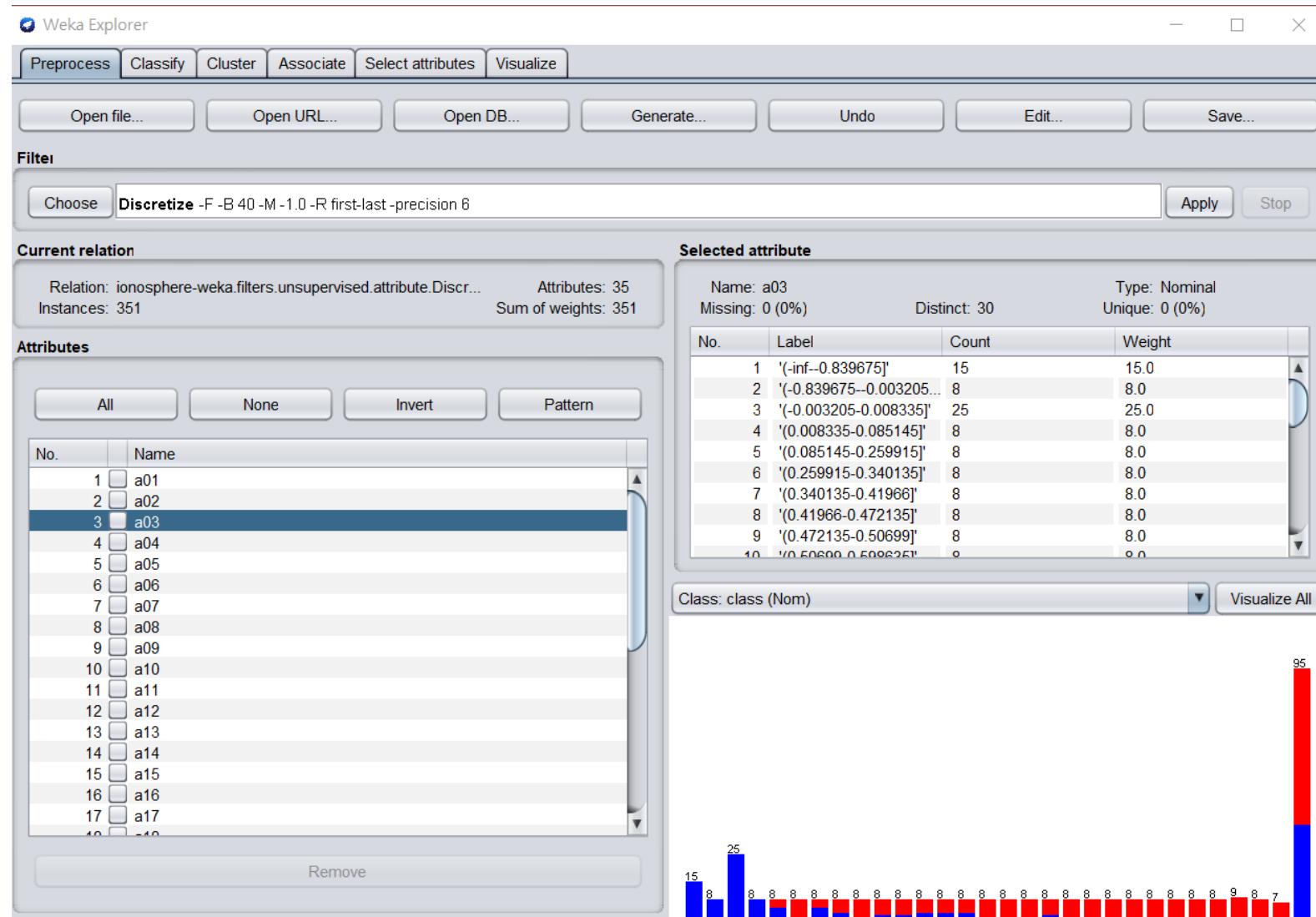
# Lesson 2.1: 離散化數字屬性

▼執行結果：a01。



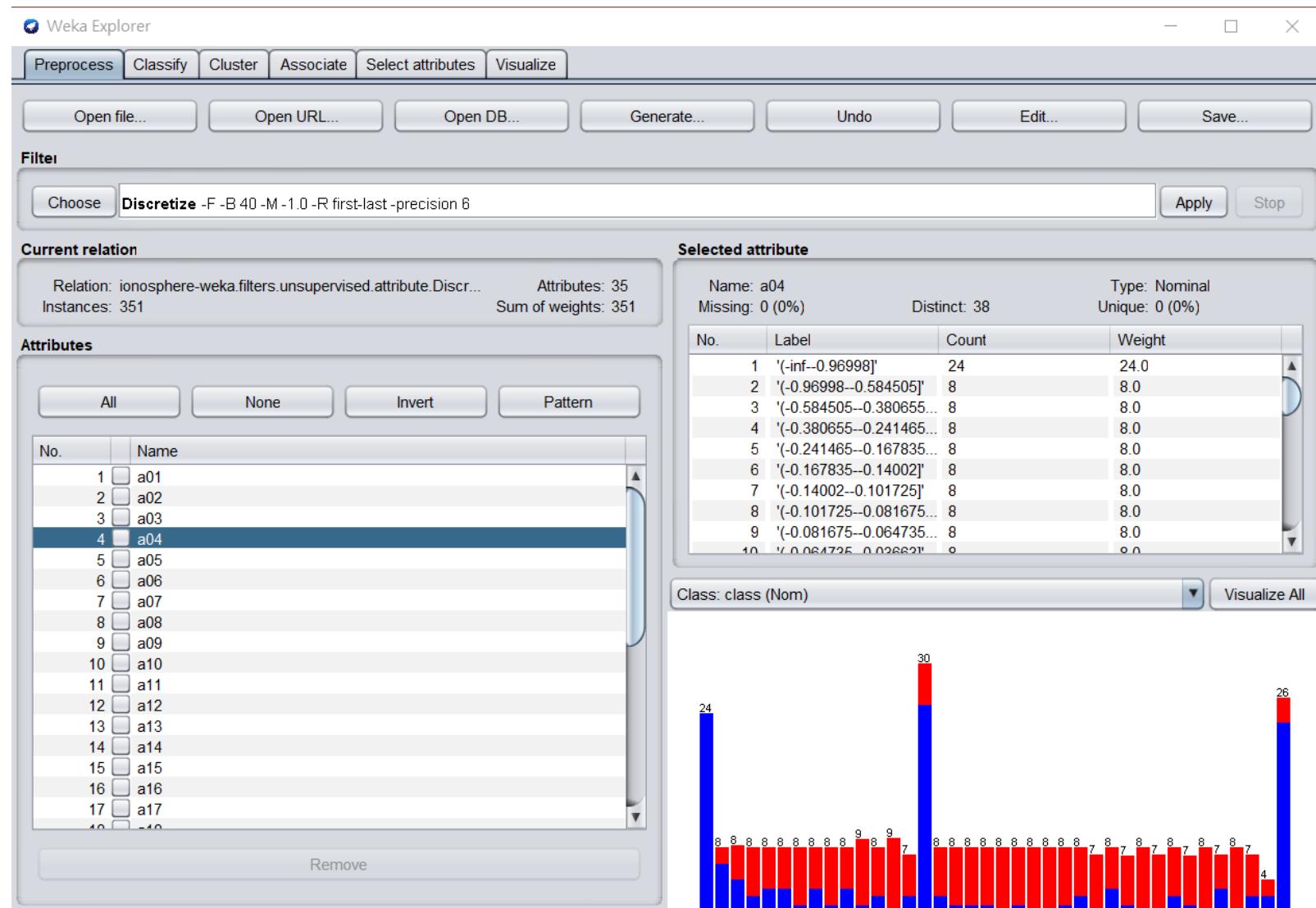
# Lesson 2.1: 離散化數字屬性

▼執行結果：a03。



# Lesson 2.1: 離散化數字屬性

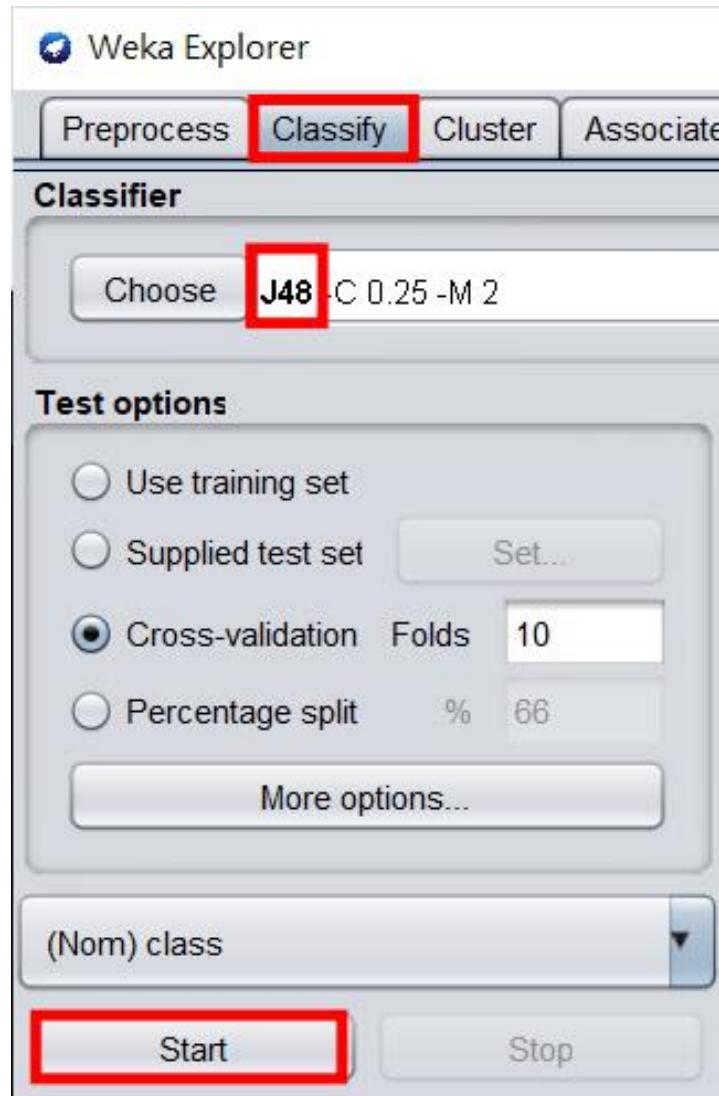
▼執行結果：a04。



這就是等頻裝箱法的原理：直方圖均衡化，在箱子中放入同等數量的數據。

## Lesson 2.1: 離散化數字屬性

5. 切換到Classify界面，確定分類器為J48後，左鍵單擊下方Start按鈕。



## Lesson 2.1: 離散化數字屬性

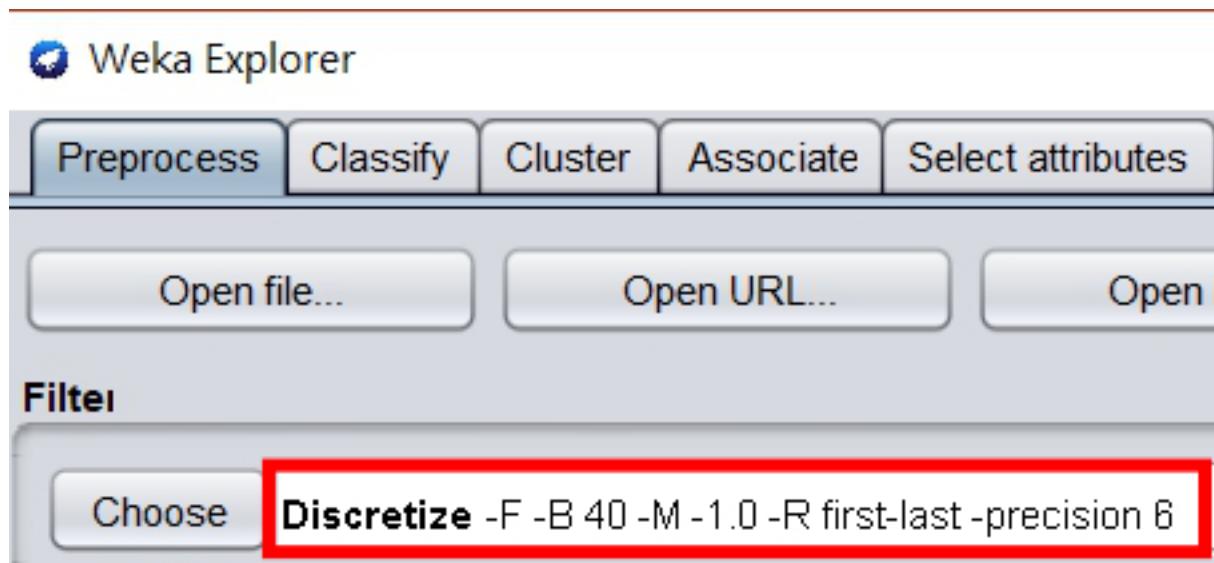
▼執行結果：得到87.1895%準確率。

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 Panel:** Cross-validation (Folds 10) is selected. Other options: Use training set, Supplied test set, Percentage split (66%).
- Classifier output Panel:**
  - Time taken to build model: 0.03 seconds
  - ==== Stratified cross-validation ===
  - ==== Summary ===
  - Correctly Classified Instances 306 87.1795 %
  - Incorrectly Classified Instances 45 12.8205 %
  - Kappa statistic 0.7219
  - Mean absolute error 0.1868
  - Root mean squared error 0.3387
  - Relative absolute error 40.56 %
  - Root relative squared error 70.6027 %
  - Total Number of Instances 351
- Result list Panel:** A list of recent runs:
  - 16:07:08 - trees.J48
  - 16:11:39 - trees.J48
  - 16:17:32 - trees.J48
  - 16:22:02 - trees.J48 (highlighted)

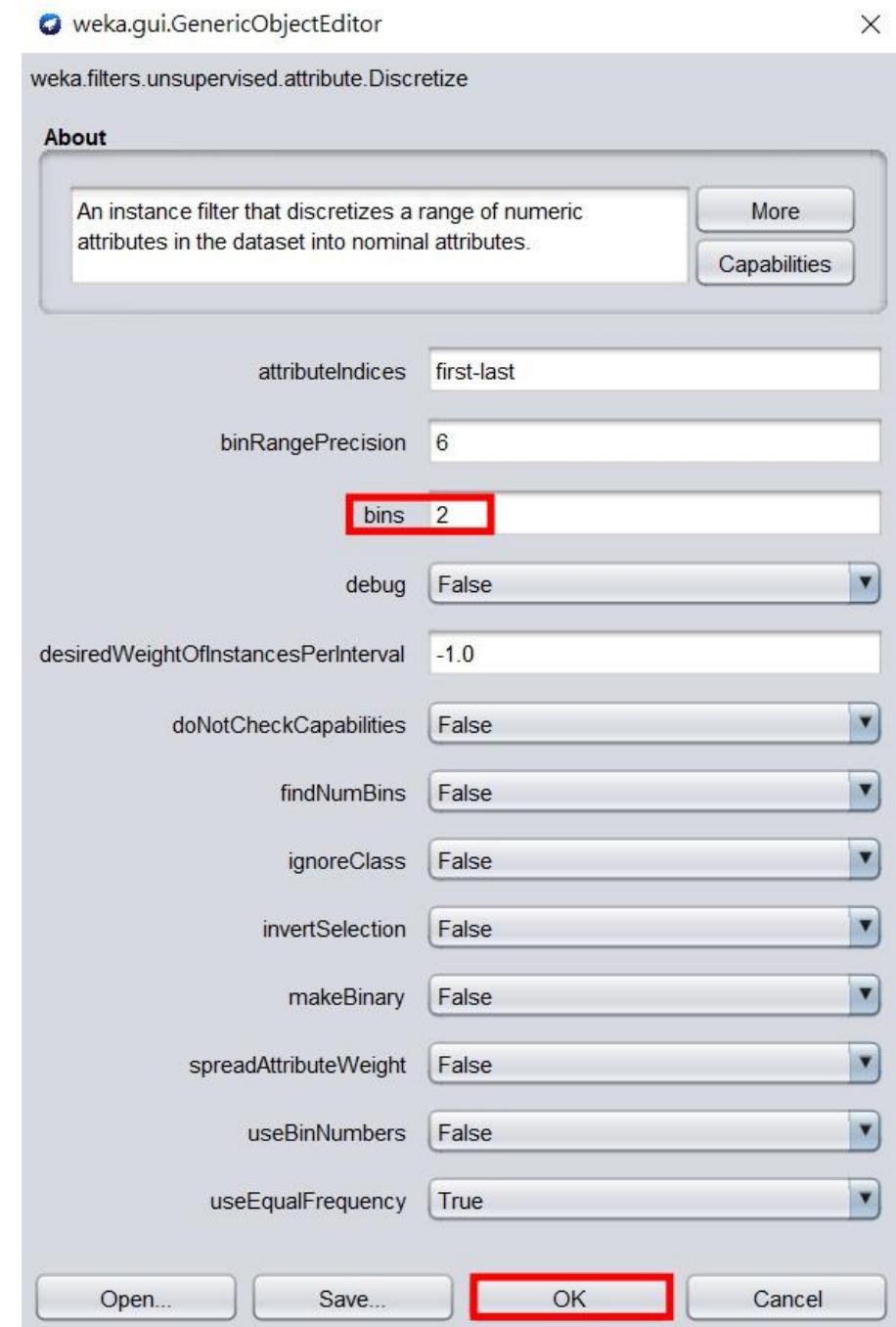
## Lesson 2.1: 離散化數字屬性

6. 切換到Preprocess面板，左鍵單擊紅色方框處進行過濾器配置。



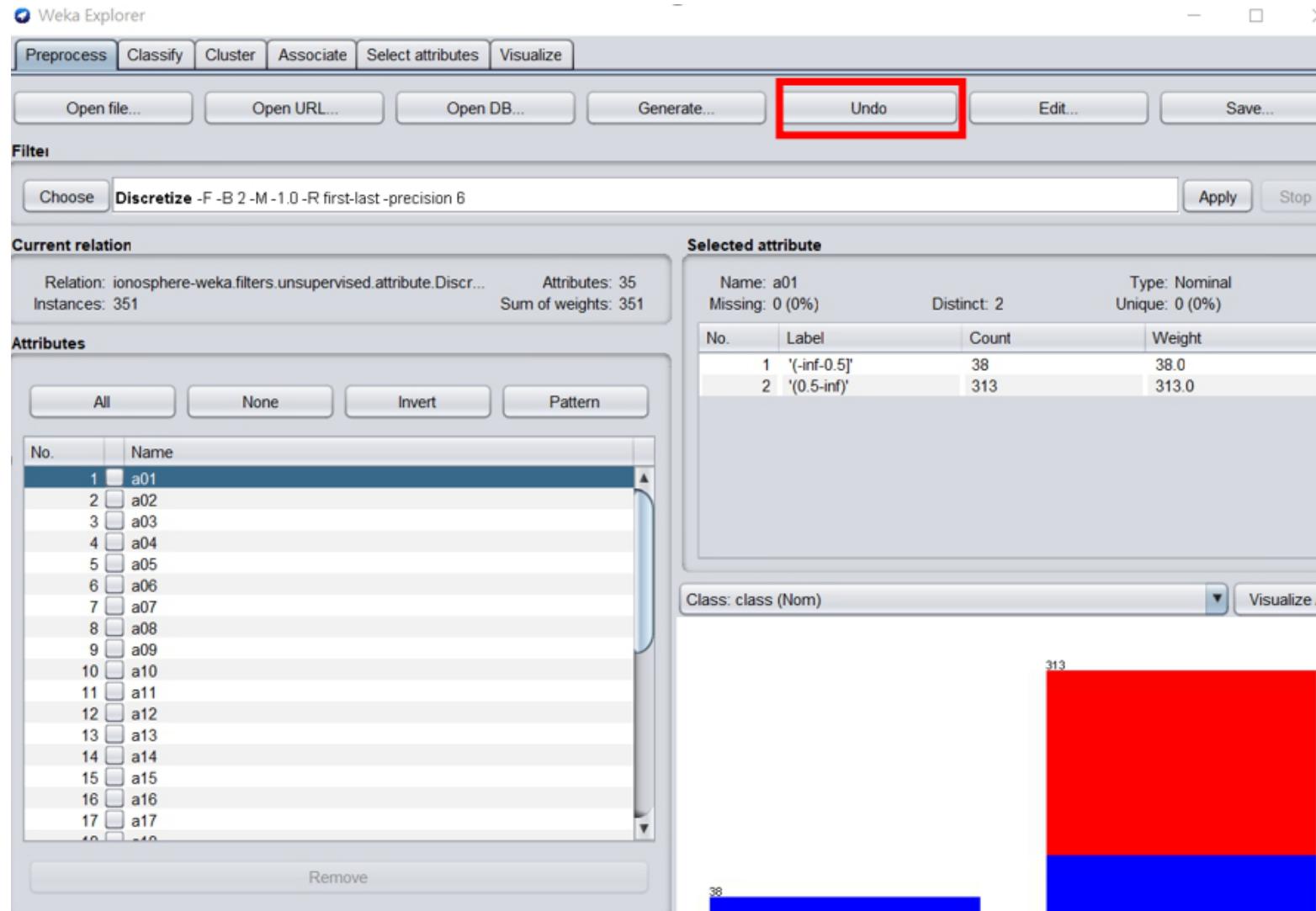
## Lesson 2.1: 離散化數字屬性

7. 在配置視窗中的bins參數後的輸入框輸入2，並確定useEqualFrequency參數設定為True，然後左鍵單擊視窗下方OK按鈕。



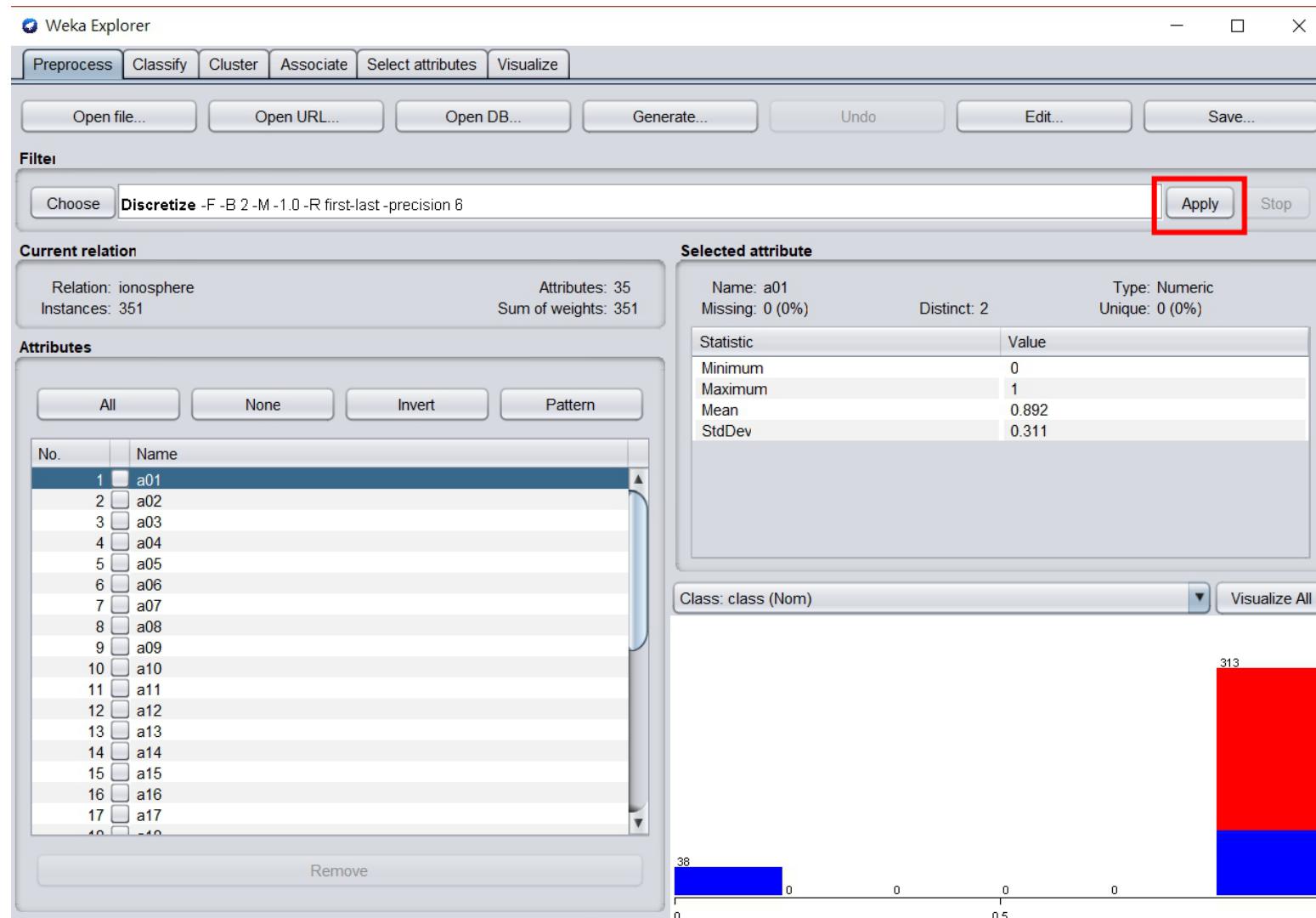
## Lesson 2.1: 離散化數字屬性

8. 左鍵單擊視窗上方Undo按鈕撤銷剛才的過濾器套用。



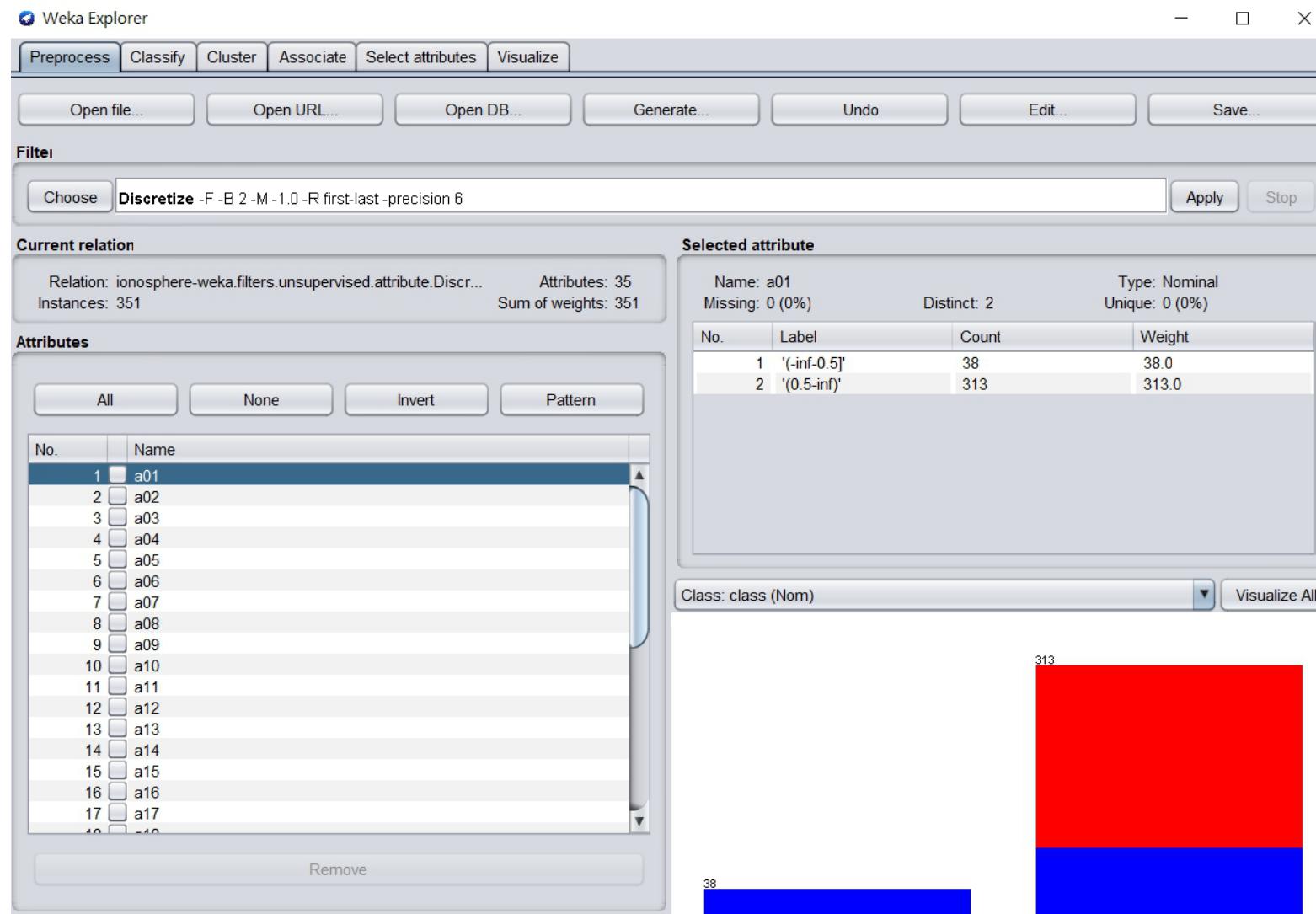
## Lesson 2.1: 離散化數字屬性

9. 左鍵單擊視窗右上方的Apply按鈕，套用剛配置的過濾器。



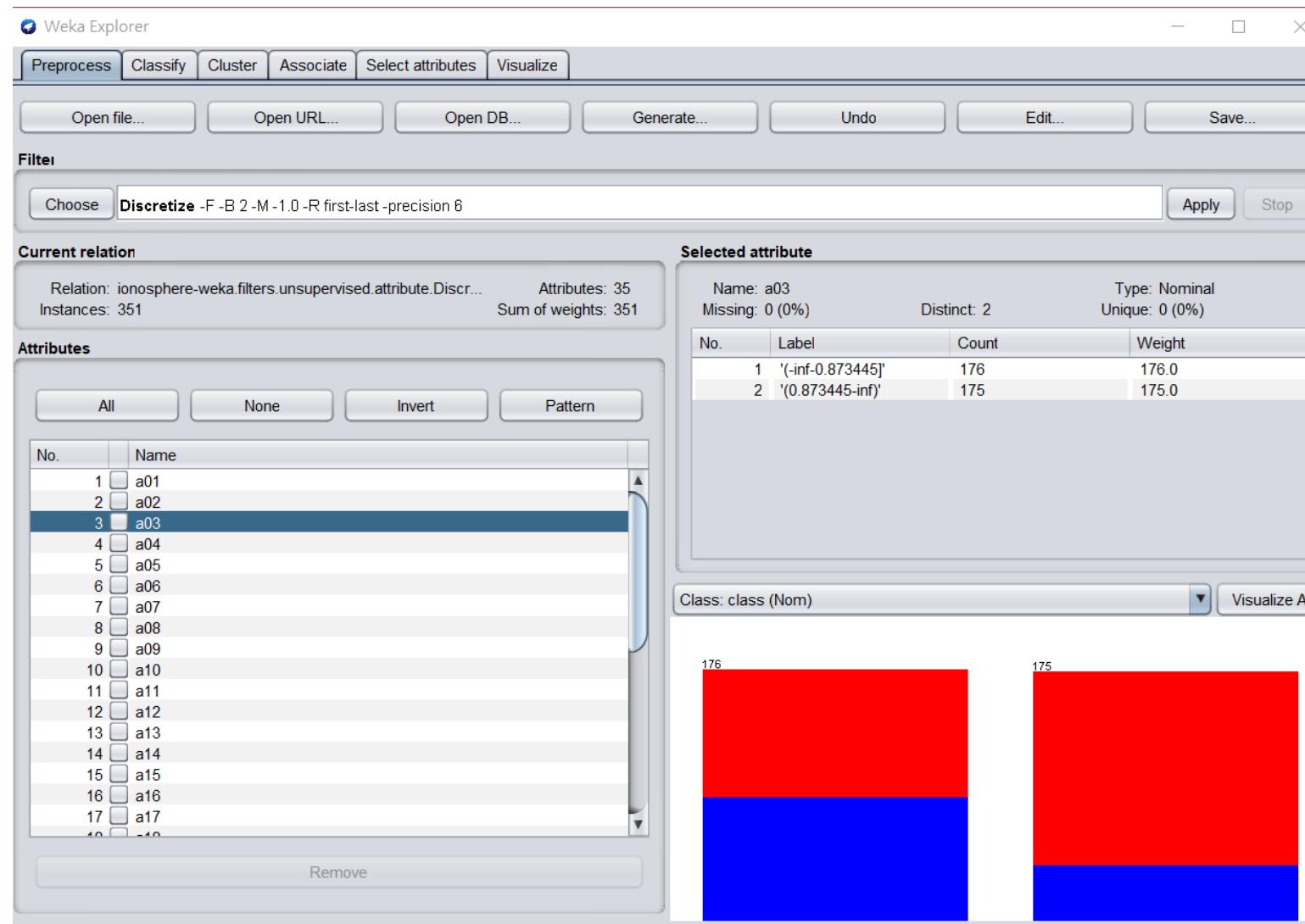
# Lesson 2.1: 離散化數字屬性

▼ 執行結果：a01。



# Lesson 2.1: 離散化數字屬性

▼執行結果：a03。



# Lesson 2.1: 離散化數字屬性

▼執行結果：a04。

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

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

Filter

Choose Discretize -F -B 2 -M 1.0 -R first-last -precision 6 Apply Stop

Current relation

Relation: ionosphere-weka.filters.unsupervised.attribute.Discr... Attributes: 35  
Instances: 351 Sum of weights: 351

Selected attribute

Name: a04 Missing: 0 (0%) Distinct: 2 Type: Nominal Unique: 0 (0%)

No.	Label	Count	Weight
1	'(-inf-0.0167]'	176	176.0
2	'(0.0167-inf)'	175	175.0

Attributes

All None Invert Pattern

No.	Name
1	a01
2	a02
3	a03
4	<b>a04</b>
5	a05
6	a06
7	a07
8	a08
9	a09
10	a10
11	a11
12	a12
13	a13
14	a14
15	a15
16	a16
17	a17
18	-18

Remove

Class: class (Nom)

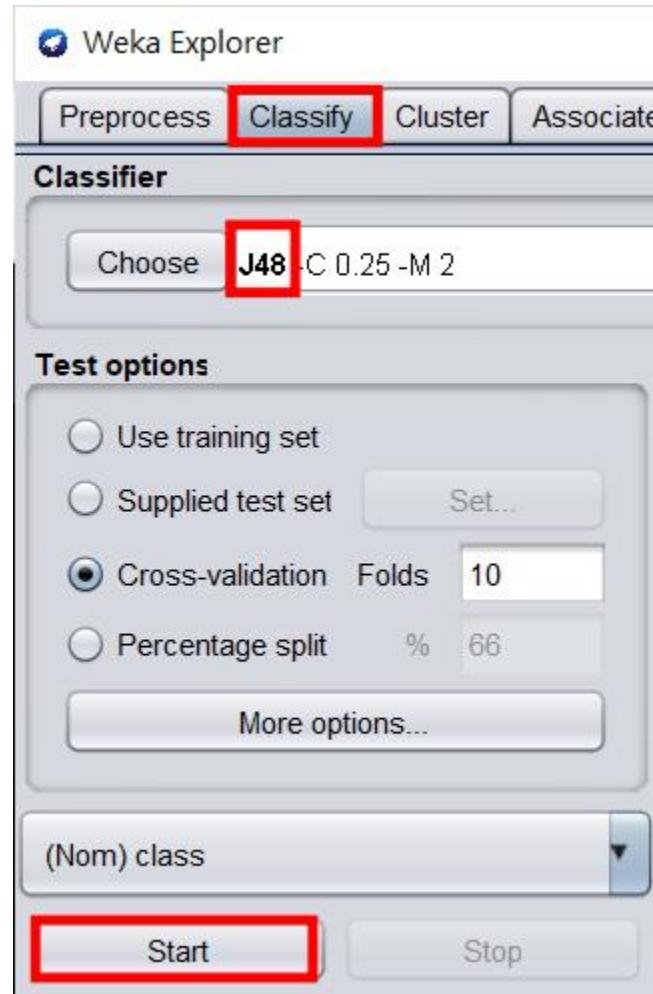
Visualize All

176

175

## Lesson 2.1: 離散化數字屬性

10. 切換到Classify界面，確定分類器為J48後，左鍵單擊下方Start按鈕。



# Lesson 2.1: 離散化數字屬性

▼執行結果：得到82.906%準確率。

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

- Weka Explorer** window title.
- Preprocess Classify Cluster Associate Select attributes Visualize** tabs.
- Classifier** panel:
  - Choose**: J48 -C 0.25 -M 2
  - Test options**:
    - Use training set
    - Supplied test set **Set...**
    - Cross-validation Folds 10
    - Percentage split % 66
  - (Nom) class** dropdown.
  - Start Stop** buttons.
- Classifier output** panel:
  - Time taken to build model: 0.02 seconds
  - ==== Stratified cross-validation ====
  - ==== Summary ====

Correctly Classified Instances	291	82.906 %
Incorrectly Classified Instances	60	17.094 %
  - Other metrics:
    - Kappa statistic 0.6206
    - Mean absolute error 0.2137
    - Root mean squared error 0.3914
    - Relative absolute error 46.4024 %
    - Root relative squared error 81.5757 %
    - Total Number of Instances 351
  - ==== Detailed Accuracy By Class ====

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Cl
a	0.714	0.107	0.789	0.714	0.750	0.622	0.811	0.740	b
b	0.893	0.286	0.848	0.893	0.870	0.622	0.811	0.816	g
Weighted Avg.	0.829	0.221	0.827	0.829	0.827	0.622	0.811	0.789	
  - ==== Confusion Matrix ====

a	b	<-- classified as
90	36	a = b
24	201	b = g

## Lesson 2.1: 離散化數字屬性

### 等頻裝箱(Equal-frequency binning)

- ❖ **ionosphere.arff**; 使用 **J48** **91.5%** (35 個節點)
- ❖ **equal-frequency**, **40** 個箱子 **87.2%** (61 個節點)
  - *a01*: 只有 2 個箱子
  - *a03*: 只有在 +1 的地方有高峰、在 -1 和 0 的地方有小高峰 (檢查 **Edit...** 視窗)
  - *a04*: 只有在 1, 0, 以及 +1 的地方有高峰
- ❖ **10** 個箱子 **89.5%** (48 個節點)
- ❖ **5** 個箱子 **90.6%** (28 個節點)
- ❖ **2** 個箱子 (請觀察屬性長條圖!) **82.6%** (47 個節點)
- ❖ 該用多少個箱子?

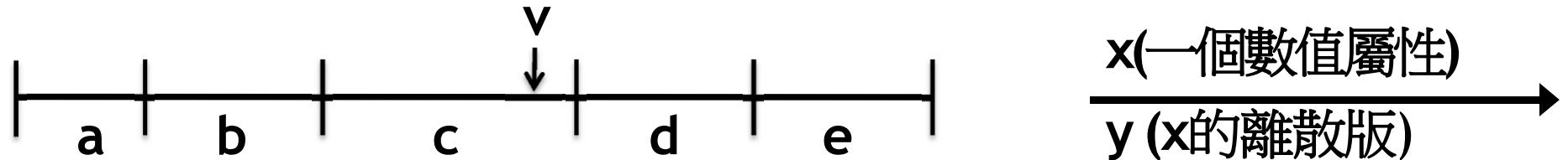
「K均衡區間離散法(proportional k-interval discretization)」理論： $\propto \sqrt{\text{實例的數量}} / \text{箱子數量}$  應該和數據數量的平方根成正比。

但這條理論並不能幫你決定最終的箱子數量，因為它沒有給出比例的常量。

從實驗結果可以看出等頻區間裝箱法的結果不如等份裝箱的結果。樹也沒有顯著變小。

## Lesson 2.1: 離散化數字屬性

如何使用排序信息？



屬性的數值版本中，在頂端的屬性值可看到一個屬性值  $v$ ，屬性的不同值之間是有順序的。

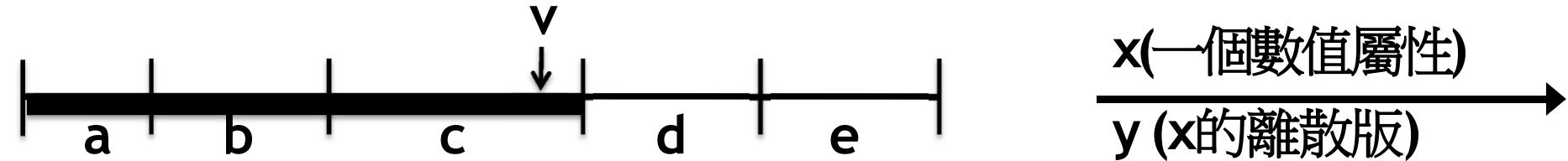
然而，當我們把數值離散到五個不同的箱子中，箱子間是沒有排序信息的。這會是一個問題，因為在離散前我們或許需要測試一個決策樹是否  $x < v$ 。

## Lesson 2.1: 離散化數字屬性

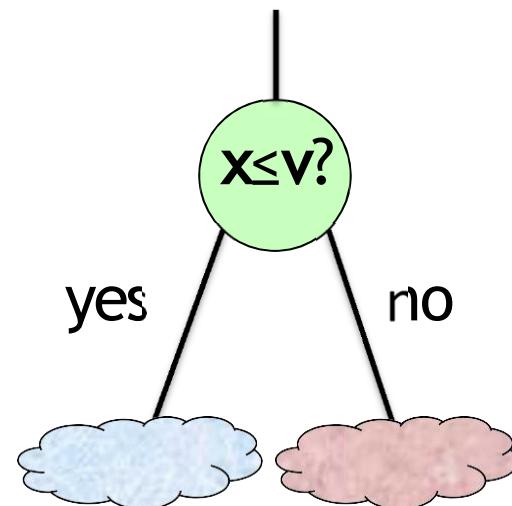
如何使用排序信息？

離散後，為了做同樣的測試，我們需要知道是否 $y=a?$ ， $y=b?$ ， $y=c?$ ，所以在每個節點下重覆測試決策樹。

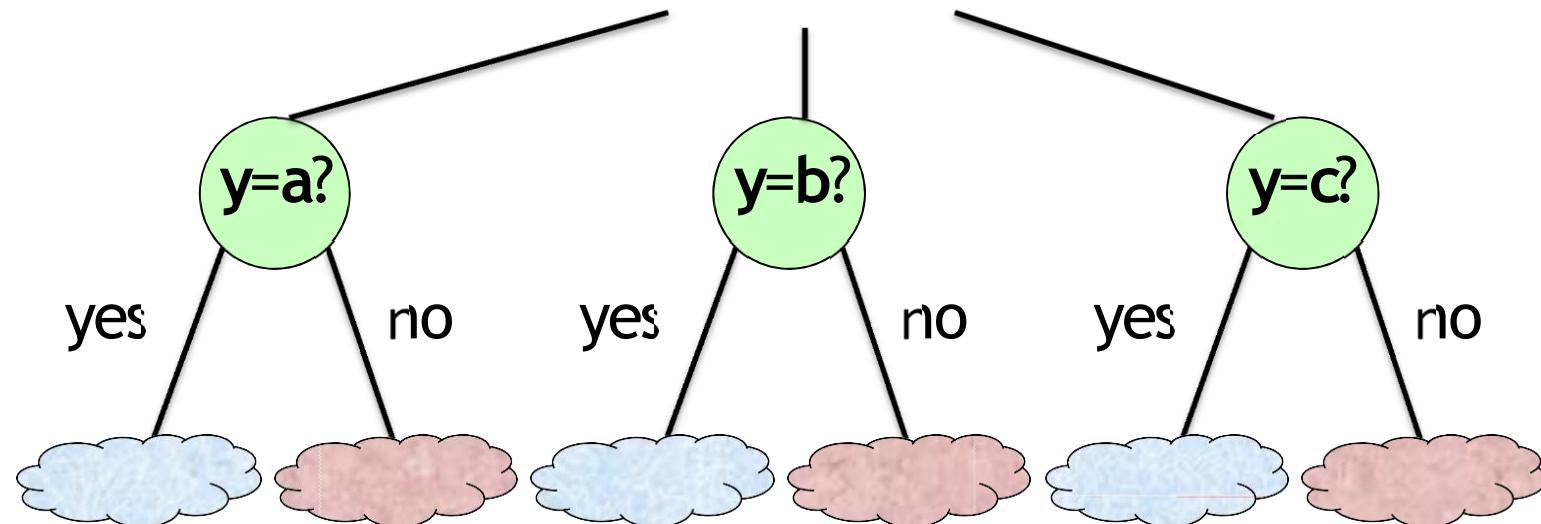
但很顯然這麼做效率很低，而且容易造成不理想的結果。



before



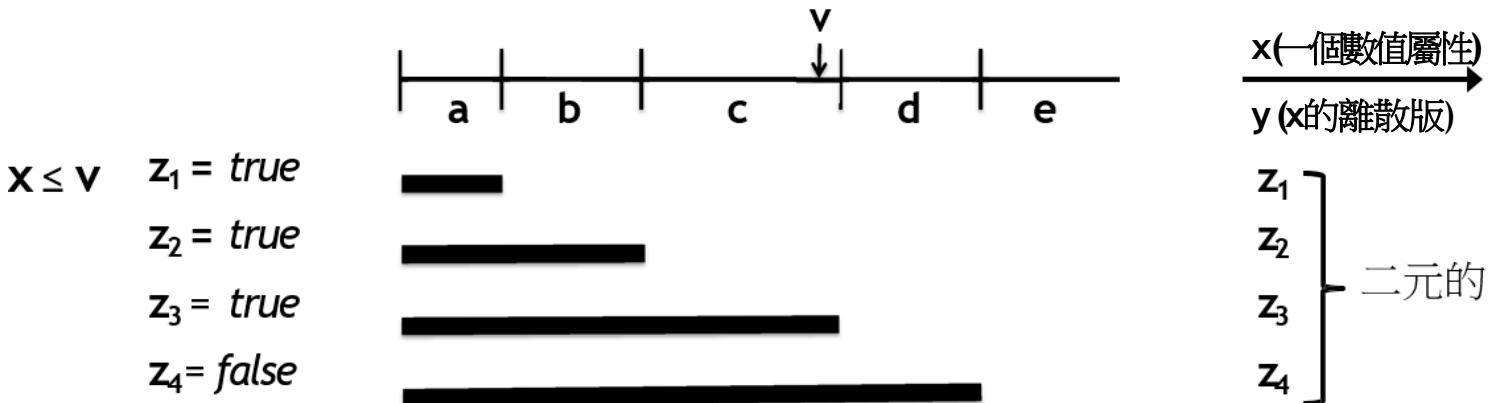
after



## Lesson 2.1: 離散化數字屬性

使用排序信息的解決辦法

- ❖ 將一個離散後的用有 $k$ 個值的屬性轉換為 $k-1$ 個二元屬性
- ❖ 如果一個特定的實例原本屬性的值是 $v$ , 將第一個 $i$ 的二元屬性設定為  
**true**，其餘的設定為**false**



我們不需要離散化數值屬性為從a到e五個中的一個值，而是離散為四個不同的二元屬性，共 $k-1$ 個二元屬性。

第一個屬性 $z_1$ 表示值 $v$ 是否落入這個範圍(a)；

第二個屬性 $z_2$ 表示 $v$ 是否落入這個範圍(a或b)；

第三個屬性 $z_3$ 表示是否落入這個範圍(a或b或c)；

第四個屬性表示它是否落入前面的四個範圍(a或b或c或d)。

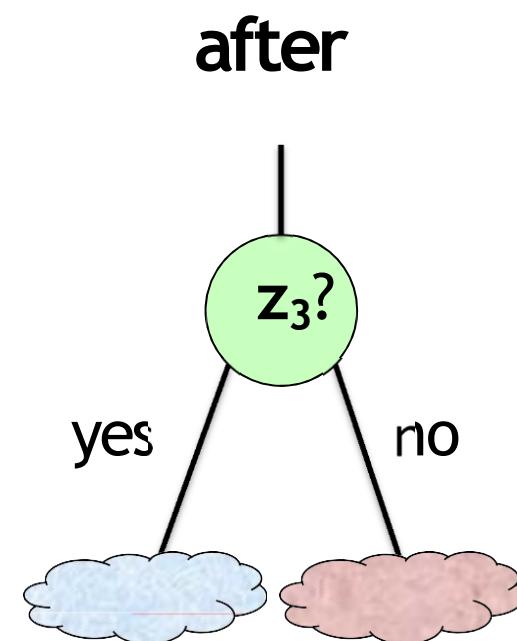
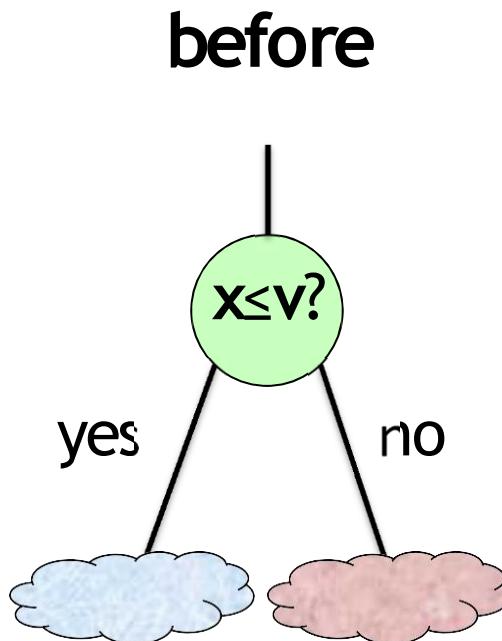
當我們測試 $x$ 是否小於 $v$ ，如是，則 $z_1$ 、 $z_2$ 、 $z_3$ 都是真值， $z_4$ 是假值。

因此，對二元屬性的對應測試就是 $z_3$ 屬性是否為真。

## Lesson 2.1: 離散化數字屬性

使用排序信息的解決辦法

- ❖ 將一個離散後的用有 $k$ 個值的屬性轉換為 $k-1$ 個二元屬性
- ❖ 如果一個特定的實例原本屬性的值是 $i$ ,將第一個 $i$ 的二元屬性設定為true,其餘的設定為false

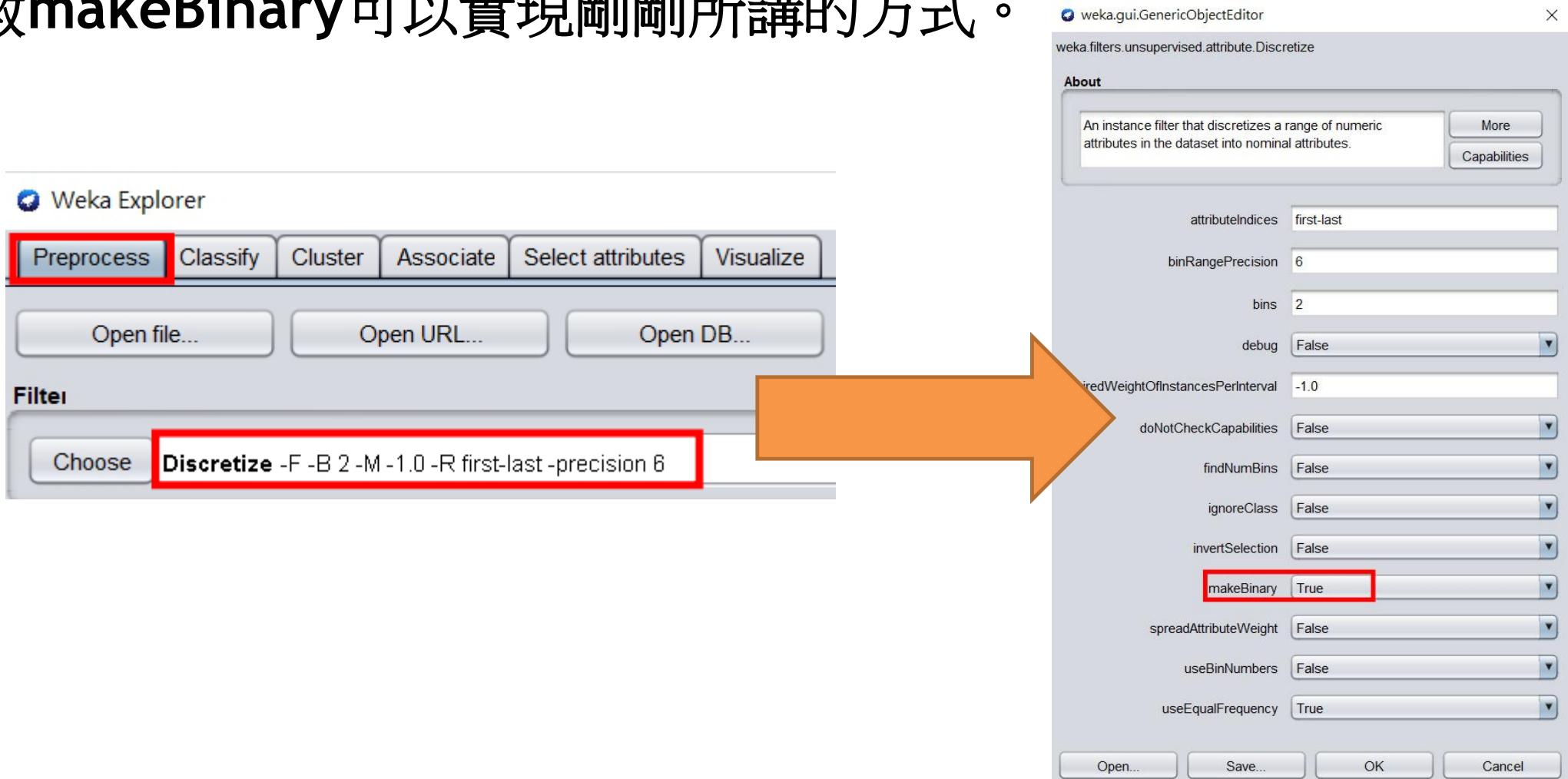


我們來看之前一個決策樹，測試是否 $x \leq v$ ，對應的測試是否就是 $z_3$ 屬性為真。

這樣我們就能提高樹結構的效率，而不需要重複測試不同的子樹。

## Lesson 2.1: 離散化數字屬性

切換到Preprocess面板，左鍵單擊左圖紅框處開啟右圖配置視窗。參數makeBinary可以實現剛剛所講的方式。

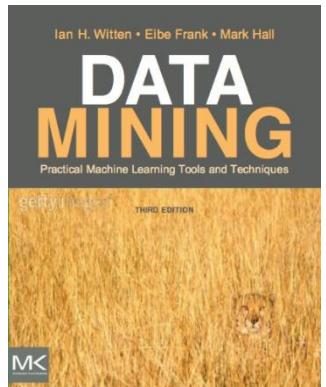


## Lesson 2.1: 離散化數字屬性

- ❖ 等份裝箱(Equal-width binning)
- ❖ 等頻裝箱(Equal-frequency binning或稱histogram equalization")
- ❖ 應該選擇幾個箱子？
- ❖ 利用數值的排序信息
  
- ❖ 下一節課：所有類別值都混在一起的「監督式(supervised)」離散化

### 課程文本

- ❖ *Section 7.2 Discretizing numeric attributes*





THE UNIVERSITY OF  
**WAIKATO**  
*Te Whare Wānanga o Waikato*

# *More Data Mining with Weka*

Class 2 - Lesson 2

監督式離散化以及 *FilteredClassifier*

(*Supervised discretization and the FilteredClassifier*)

Ian H. Witten

Department of Computer Science University of  
Waikato  
New Zealand

# Lesson 2.2: 監督式離散化以及*FilteredClassifier*

Class 1 探索Weka界面，處理大數據

Class 2 細散化以及文本分類

Class 3 分類規則、關聯規則、聚類

Class 4 選擇屬性以及計算成本

Class 5 神經網路，學習曲線和表現優化

Lesson 2.1 細散化

Lesson 2.2 監督式離散化

Lesson 2.3 Discretization in J48

Lesson 2.4 Document classification

Lesson 2.5 Evaluating 2-class classification

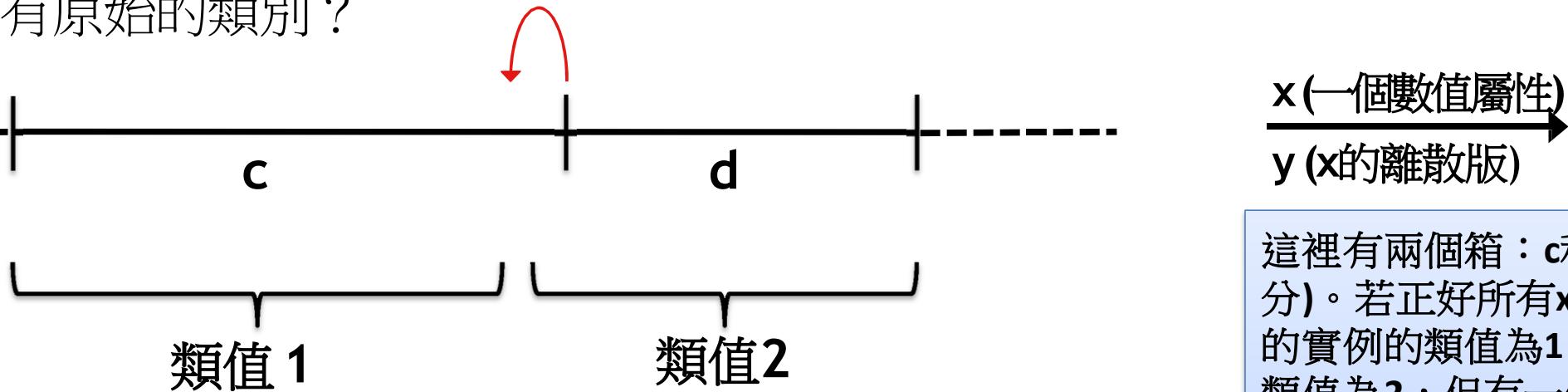
Lesson 2.6 Multinomial Naïve Bayes



## Lesson 2.2: 監督式離散化以及FilteredClassifier

### 將數字屬性轉換為名詞屬性

- ❖ 如果一個箱子裡所有的屬性都擁有一個類別，且所有屬性在下一個箱子裡擁有非第一種類別的類別，則哪個屬性擁有原始的類別？



- ❖ 所有類別值都混在一起 - 「監督式(supervised)」離散化

這裡有兩個箱：**c**和**d**(所有箱的一部分)。若正好所有x屬性值落在**c**箱內的實例的類值為1，**d**箱內的實例的類值為2，但有一個例外：有一x屬性值是在**c**箱內的實例的類值為2。

我們可以發現，這一特別的屬性及離散後的屬性，和它的實例的類別有一種精確的對應關係——解釋了當我們在做離散的決定時，可以將類值也考慮進來。

## Lesson 2.2: 監督式離散化以及FilteredClassifier

# 將數字屬性轉換為名詞屬性

- ❖ 使用熵啟發法(C4.5首創、位於Weka內的J48)
  - ❖ 如：weather.numeric.arff 資料集中的*temperature*屬性

- ❖ 選擇熵最小的分支點（信息增益最大）。
  - ❖ 驟迴地重複執行，直到達到某個停止的標準。

64	65	68	69	70	71	72	75	80	81	83	85
yes	no	yes	yes	yes	no	no	yes	no	yes	yes	no

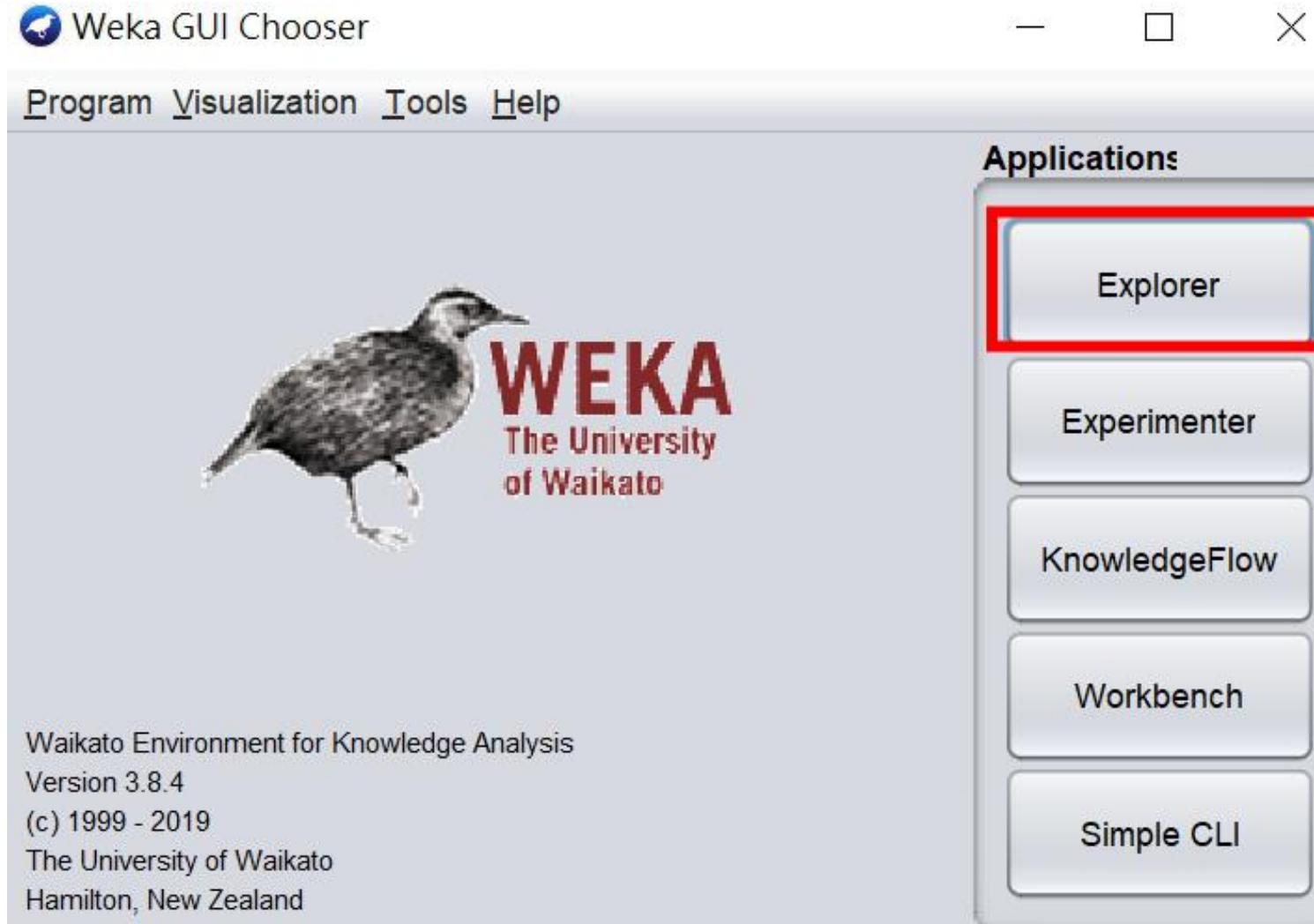
## Lesson 2.2: 監督式離散化以及*FilteredClassifier*

### 監督式離散化：以信息增益為基礎

- ❖ `ionosphere.arff`; 使用 **J48** 91.5% (35個節點)
- ❖ `Supervised`資料夾下的`attribute`資料夾下的`discretize`: 檢驗參數
- ❖ 套用 `filter`: 屬性的範圍從1-6個箱子(bins)
- ❖ 使用 **J48?** - 但是有一個關於**cross-validation**的問題！
  - 因為測試集已經被使用來幫助設定離散邊界了 - 所以這是作弊!!!

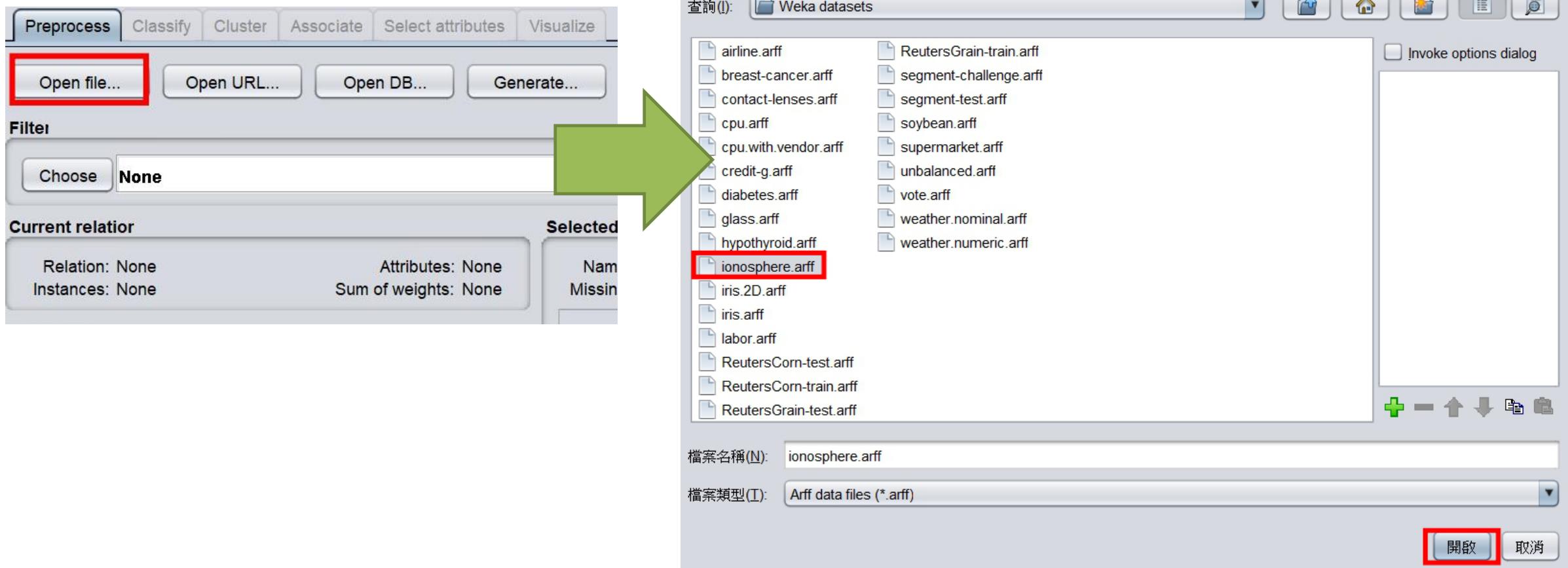
## Lesson 2.2: 監督式離散化以及FilteredClassifier

### 1. 開啟Weka的Explorer



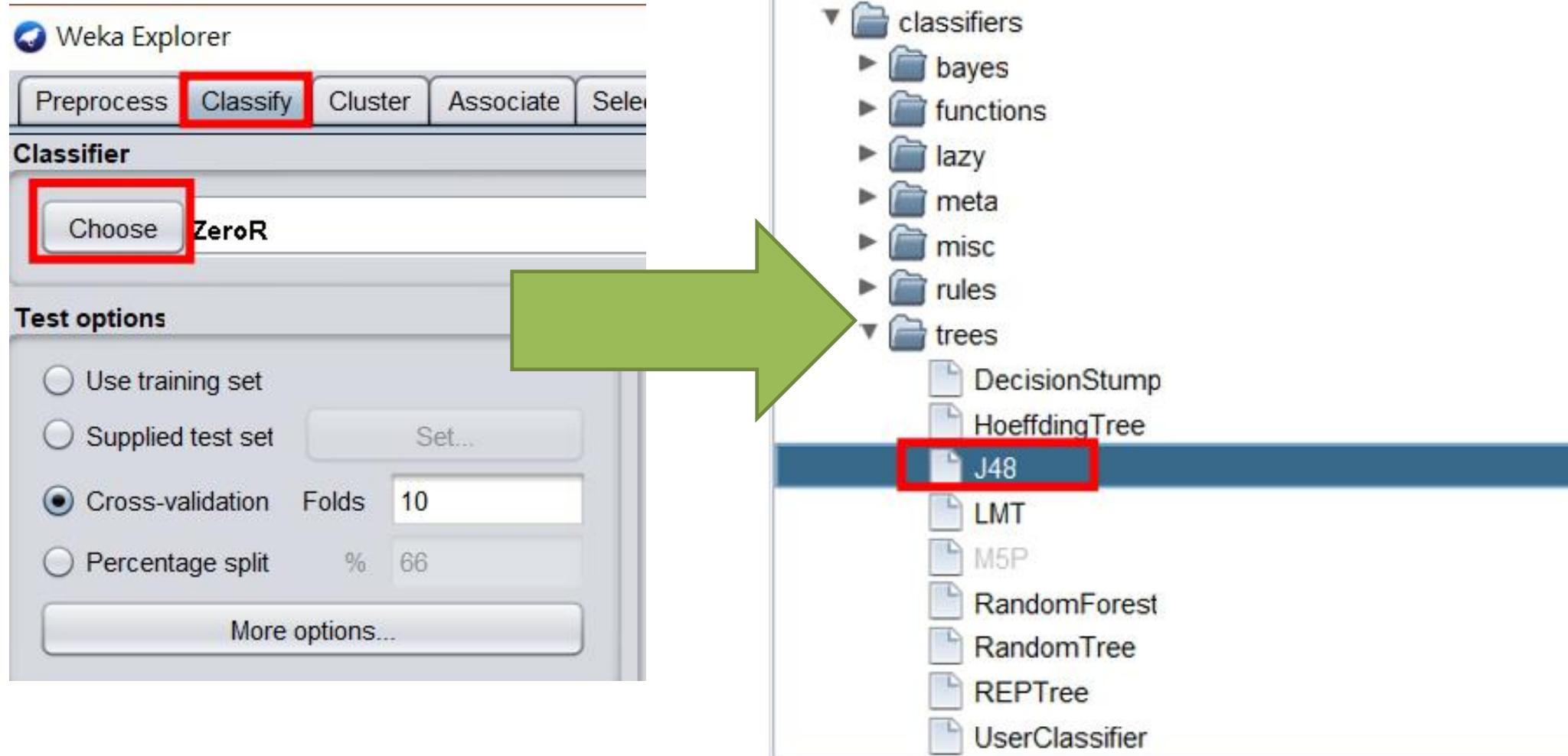
## Lesson 2.2: 監督式離散化以及FilteredClassifier

2. 左鍵點擊Open file...開啟右圖視窗，進入自行複製的Weka datasets，左鍵單擊**ionosphere.arff**的檔案後，再以左鍵單擊下方”開啟”以載入此檔案



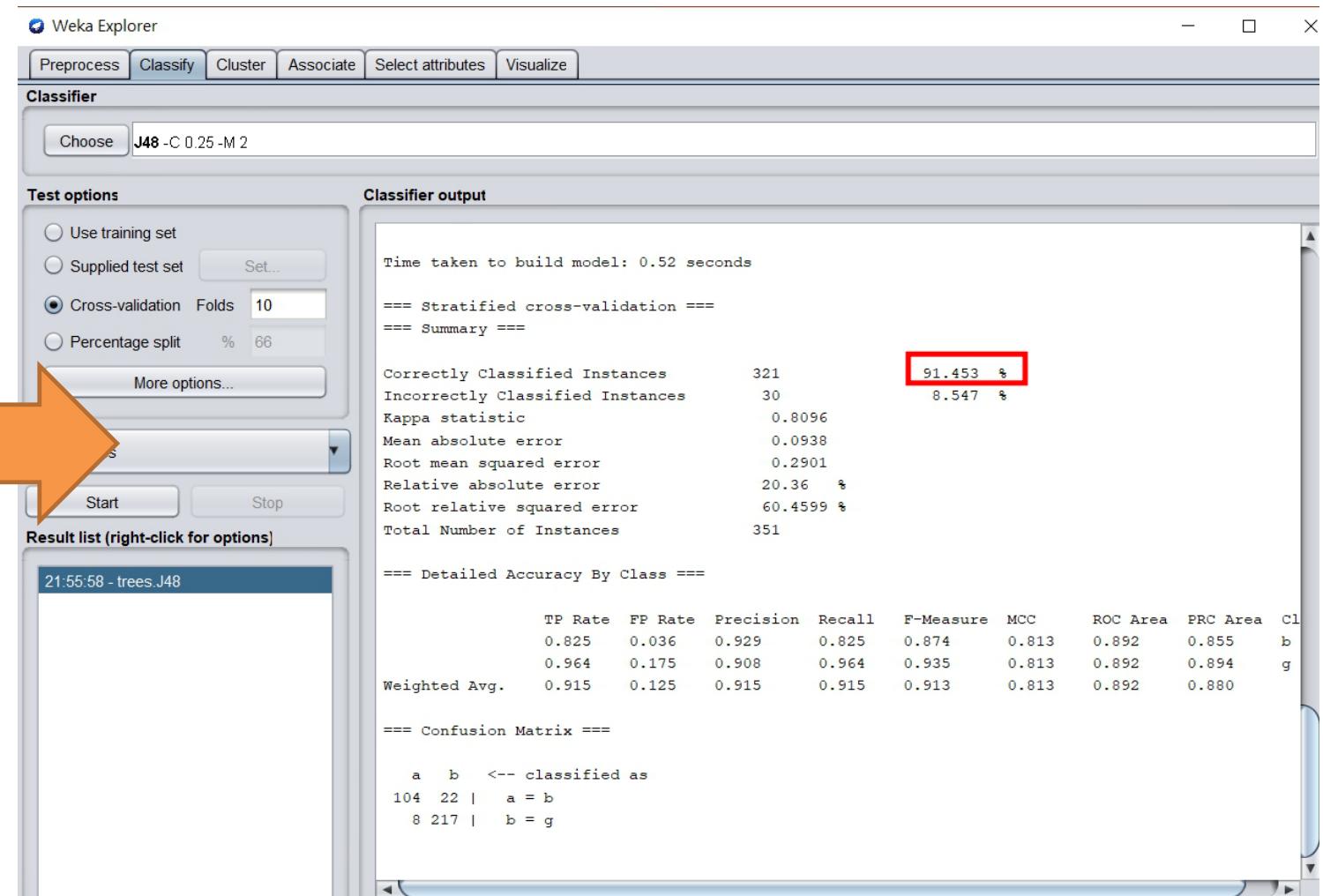
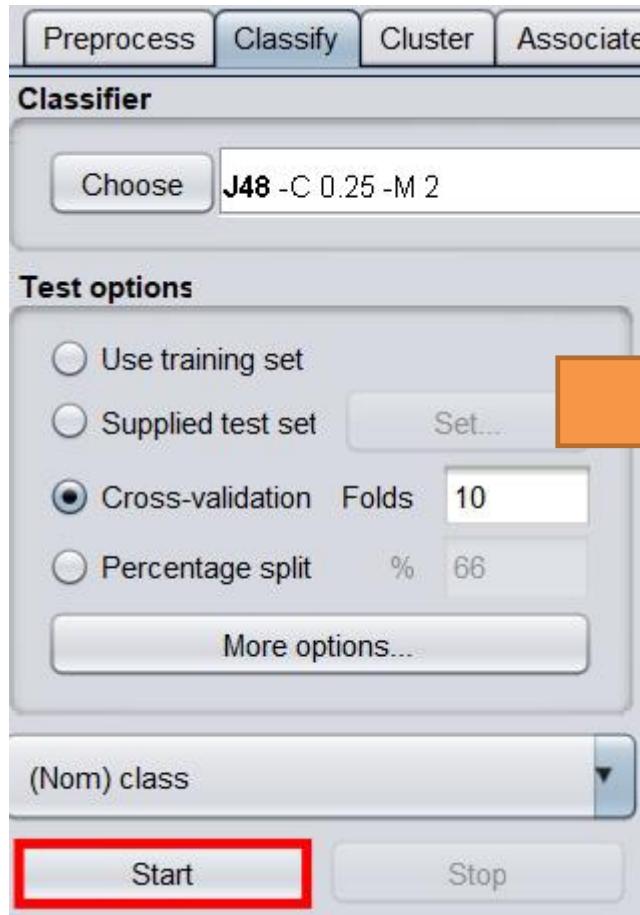
## Lesson 2.2: 監督式離散化以及FilteredClassifier

3. 切換到Classify介面點選Choose鈕，在出現的選單中左鍵單擊trees資料夾下的J48



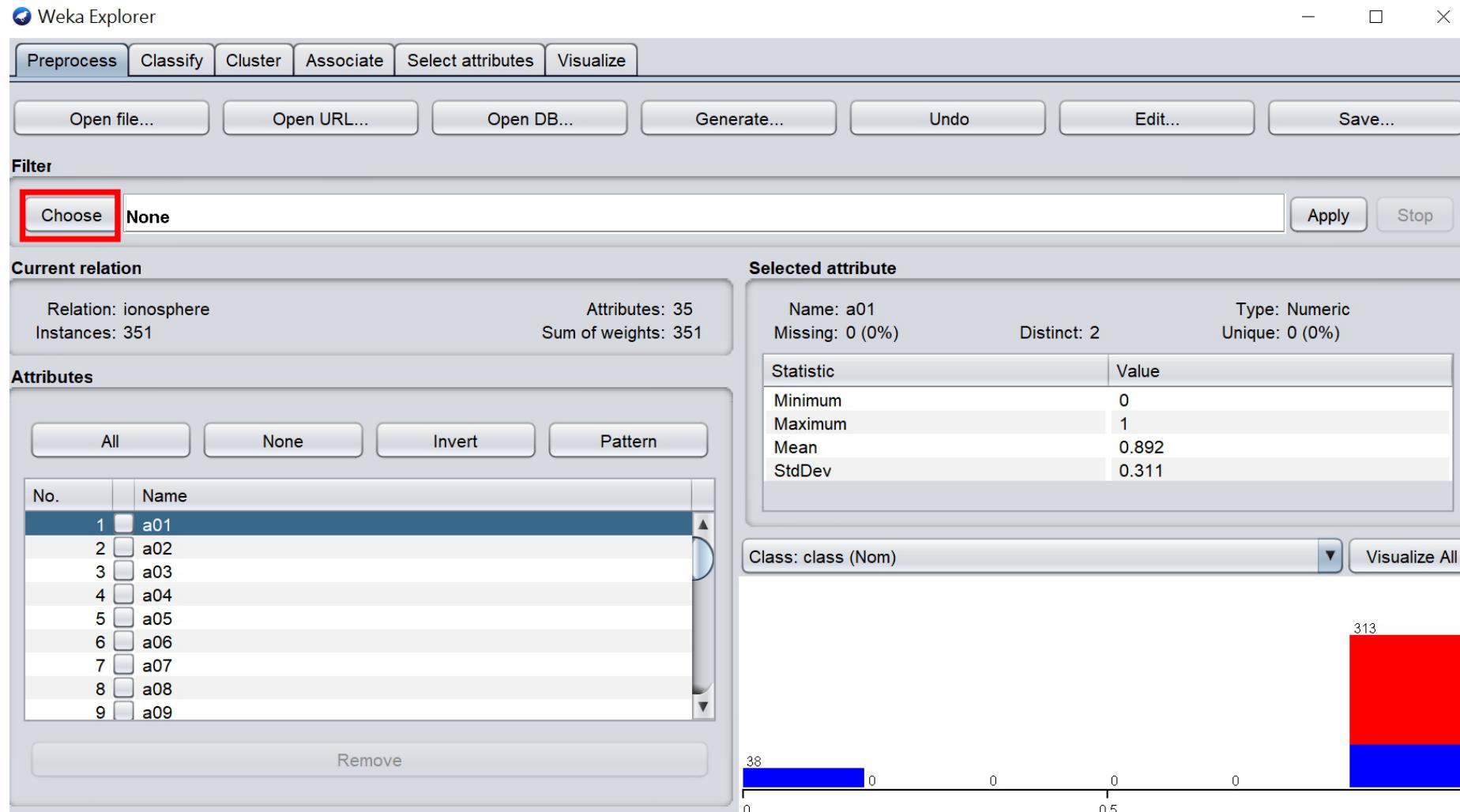
## Lesson 2.2: 監督式離散化以及FilteredClassifier

4. 左鍵單擊Start按鈕運行J48分類器得到91.453%準確率。



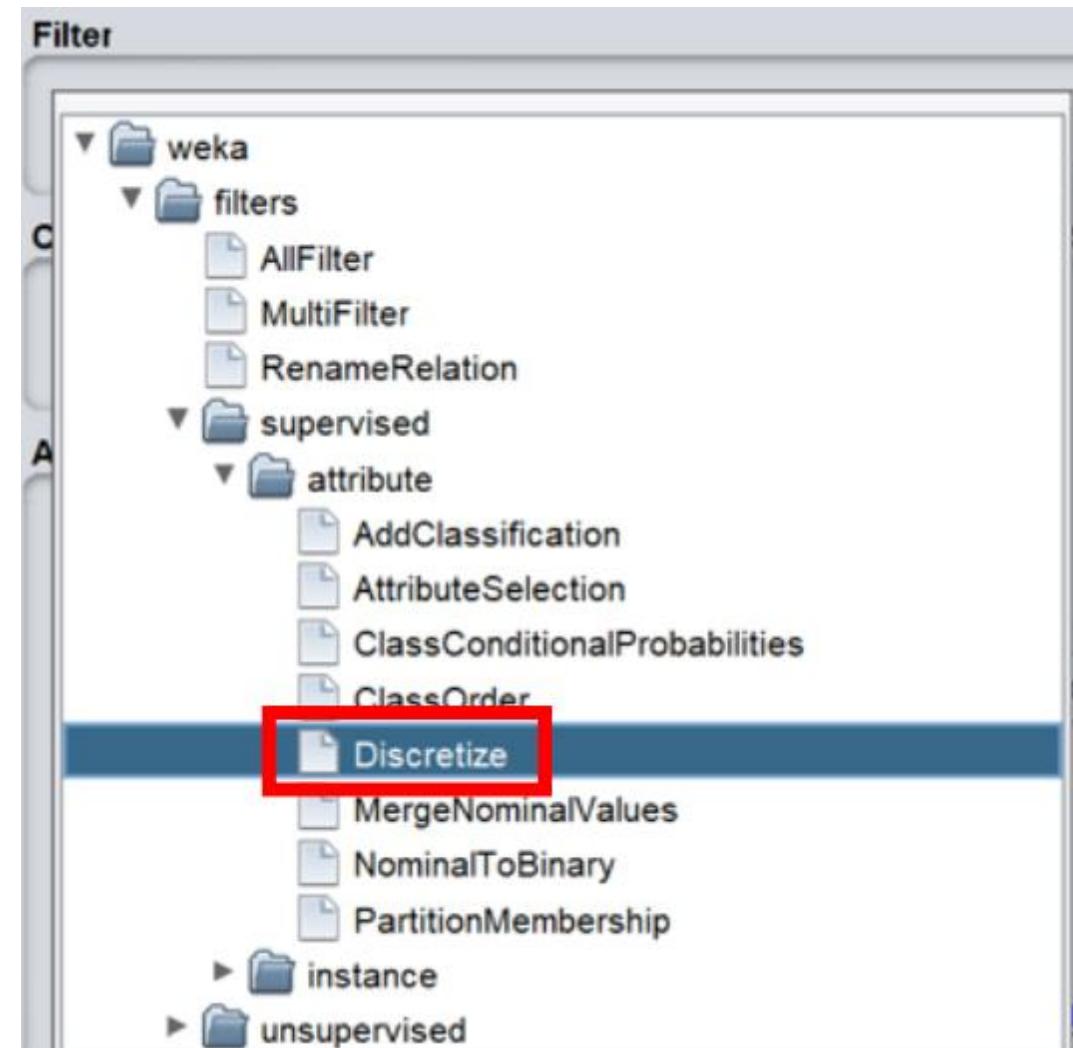
## Lesson 2.2: 監督式離散化以及FilteredClassifier

5. 切換到Preprocess面板，左鍵單擊Choose按鈕選擇過濾器。



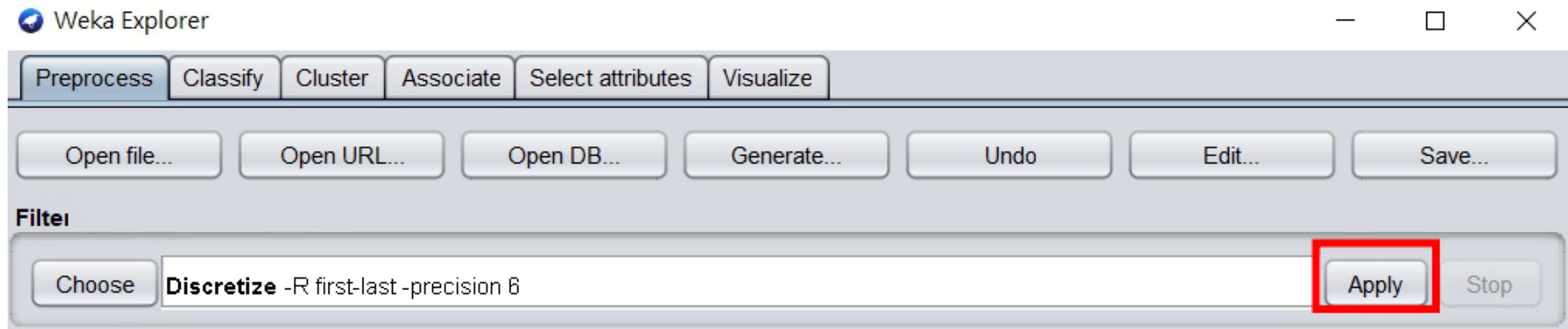
## Lesson 2.2: 監督式離散化以及FilteredClassifier

6. 左鍵單擊filters資料夾下的supervised資料夾下的attribute資料夾下的Discretize過濾器。



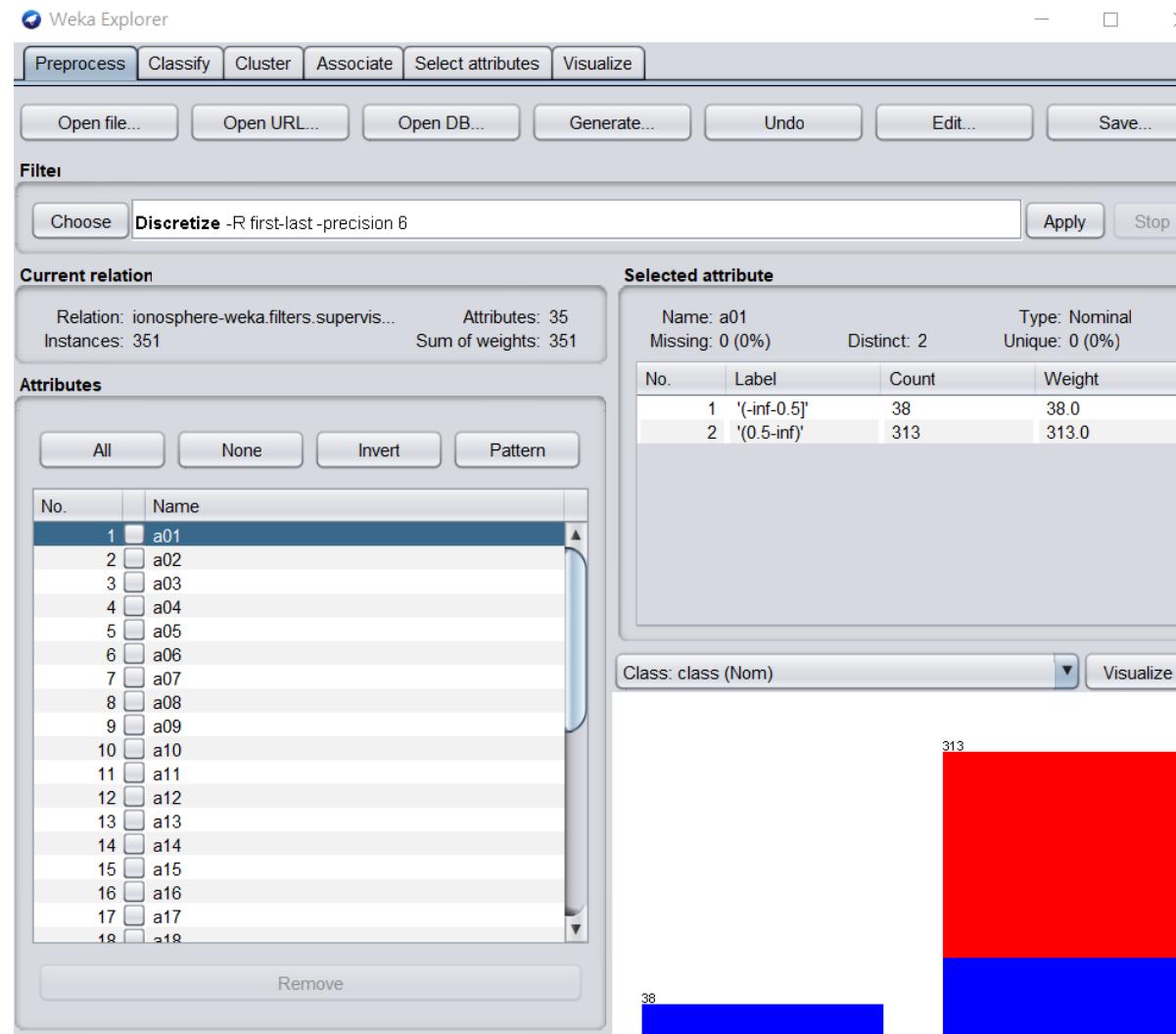
## Lesson 2.2: 監督式離散化以及FilteredClassifier

7. 回到Preprocess面板，左鍵單擊Apply按鈕套用過濾器。



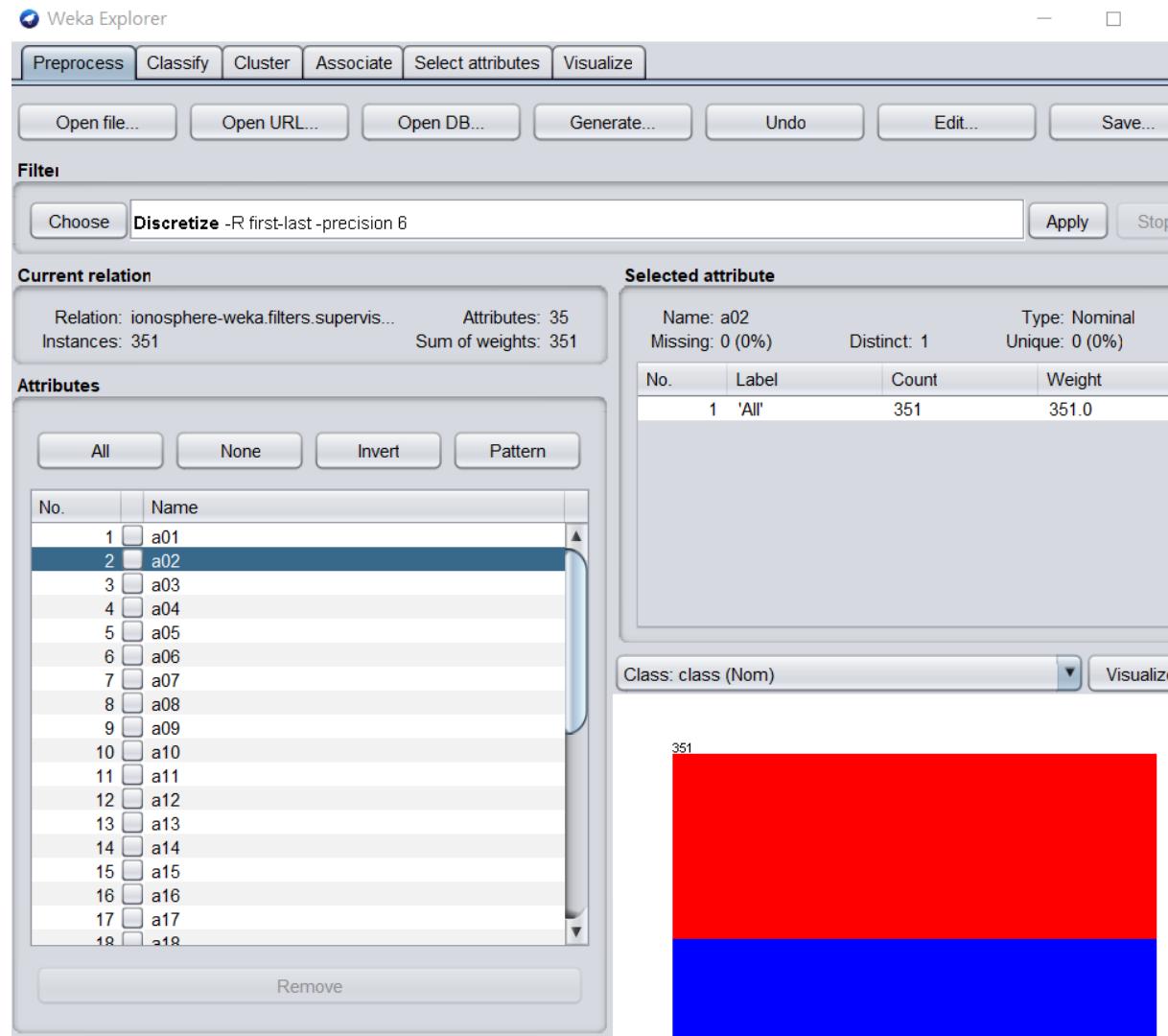
## Lesson 2.2: 監督式離散化以及FilteredClassifier

▼運行結果：a01。



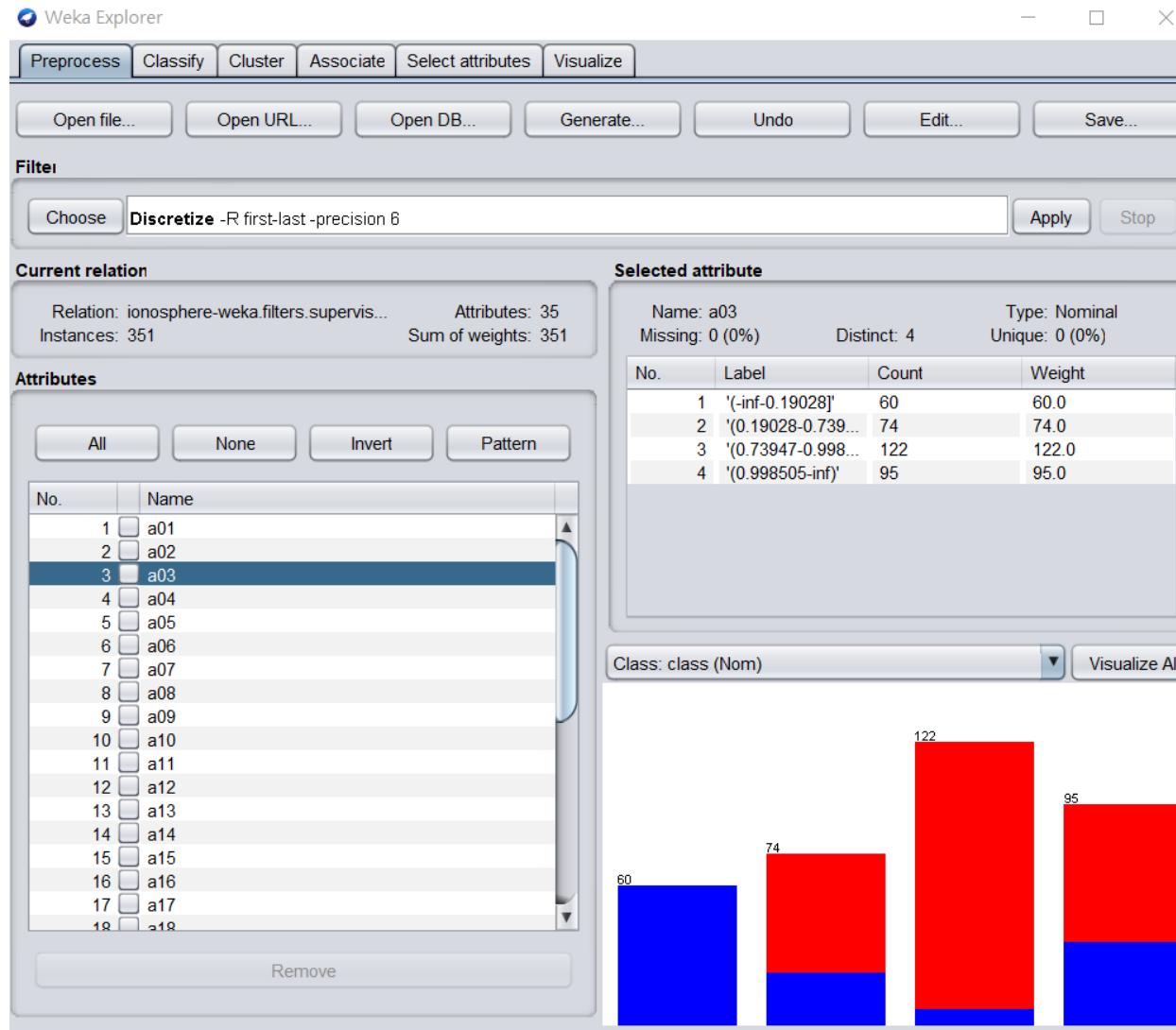
## Lesson 2.2: 監督式離散化以及FilteredClassifier

▼運行結果：a02。



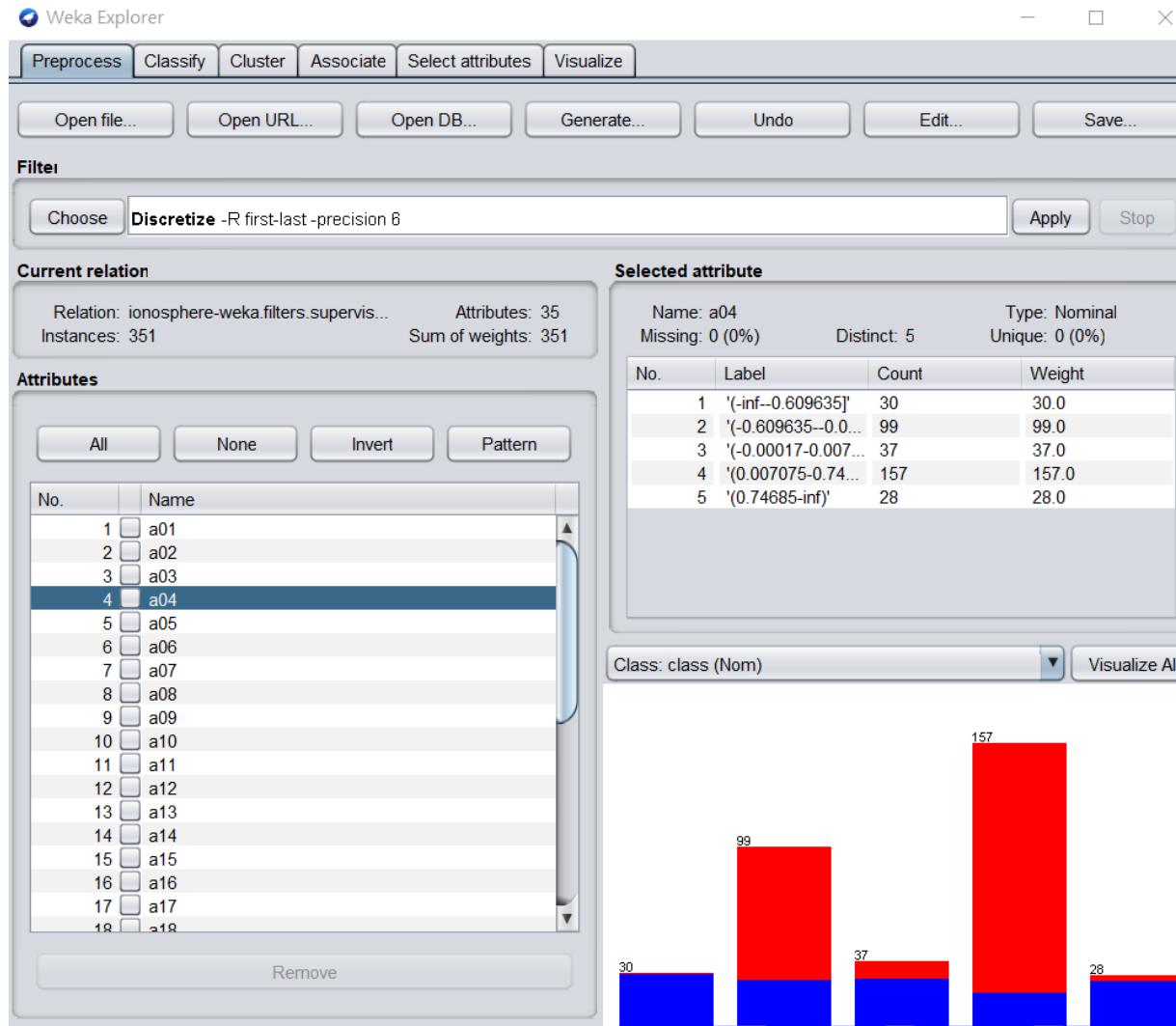
## Lesson 2.2: 監督式離散化以及FilteredClassifier

▼運行結果：a03。



## Lesson 2.2: 監督式離散化以及FilteredClassifier

▼運行結果：a04。



## Lesson 2.2: 監督式離散化以及*FilteredClassifier*

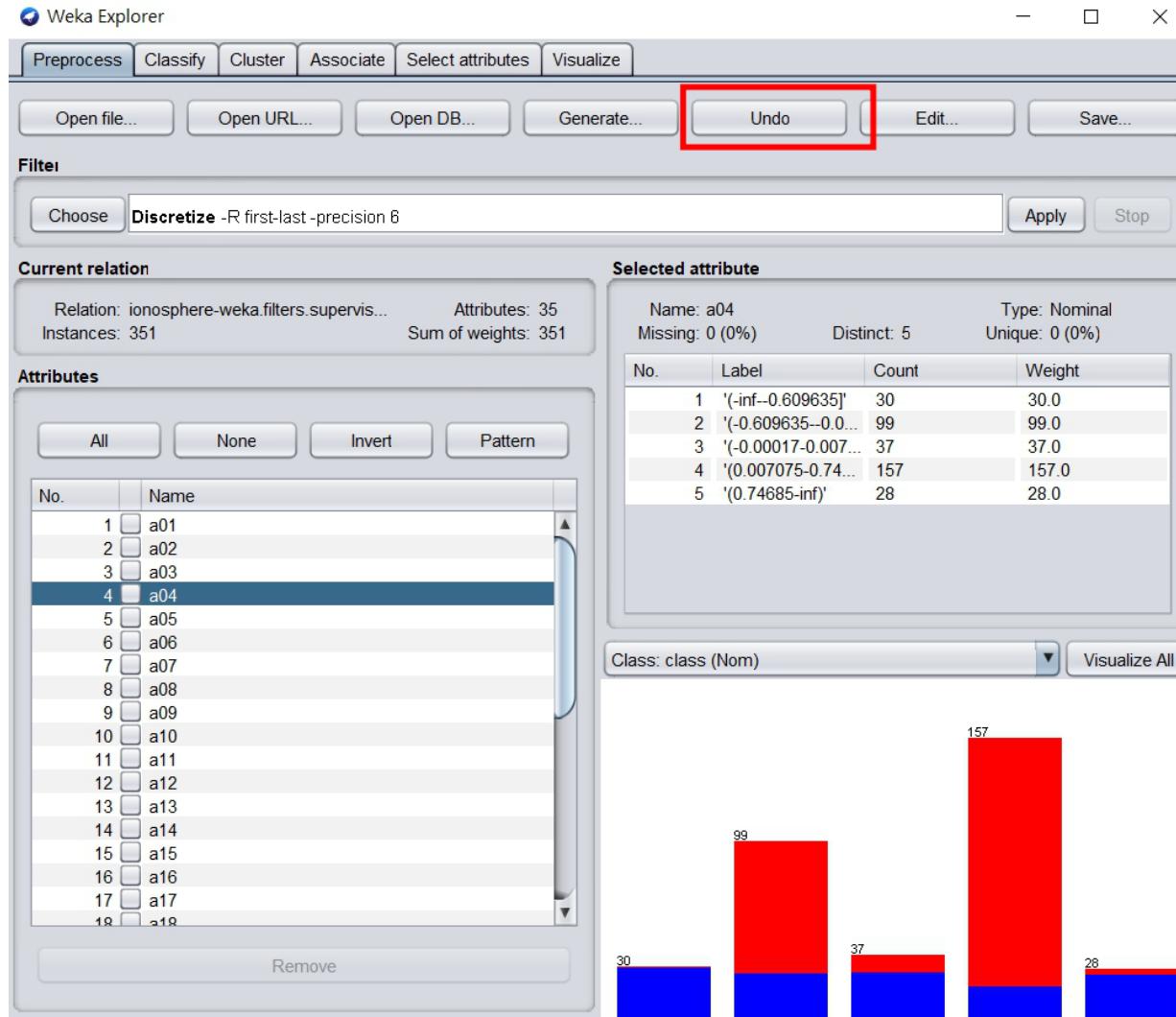
### 監督式離散化：以信息增益為基礎

- ❖ `ionosphere.arff`; 使用 **J48** 91.5% (35個節點)
- ❖ `Supervised`資料夾下的`attribute`資料夾下的`discretize`: 檢驗參數
- ❖ 套用 `filter`: 屬性的範圍從1-6個箱子(bins)
- ❖ 使用 **J48?** - 但是有一個關於**cross-validation**的問題！
  - 因為測試集已經被使用來幫助設定離散邊界了 - 所以這是作弊!!!
- ❖ (撤銷過濾器)
- ❖ `Meta`資料夾下的`FilteredClassifier`: 查看「More」資訊
- ❖ 設定過濾器及**J48**分類器; 運行: 91.2% (27個節點)
- ❖ 配置過濾器，設定參數`makeBinary` 92.6% (17個節點)
- ❖ 使用 `makeBinary`預先離散化進行作弊 94.0% (17個節點)

## Lesson 2.2: 監督式離散化以及FilteredClassifier

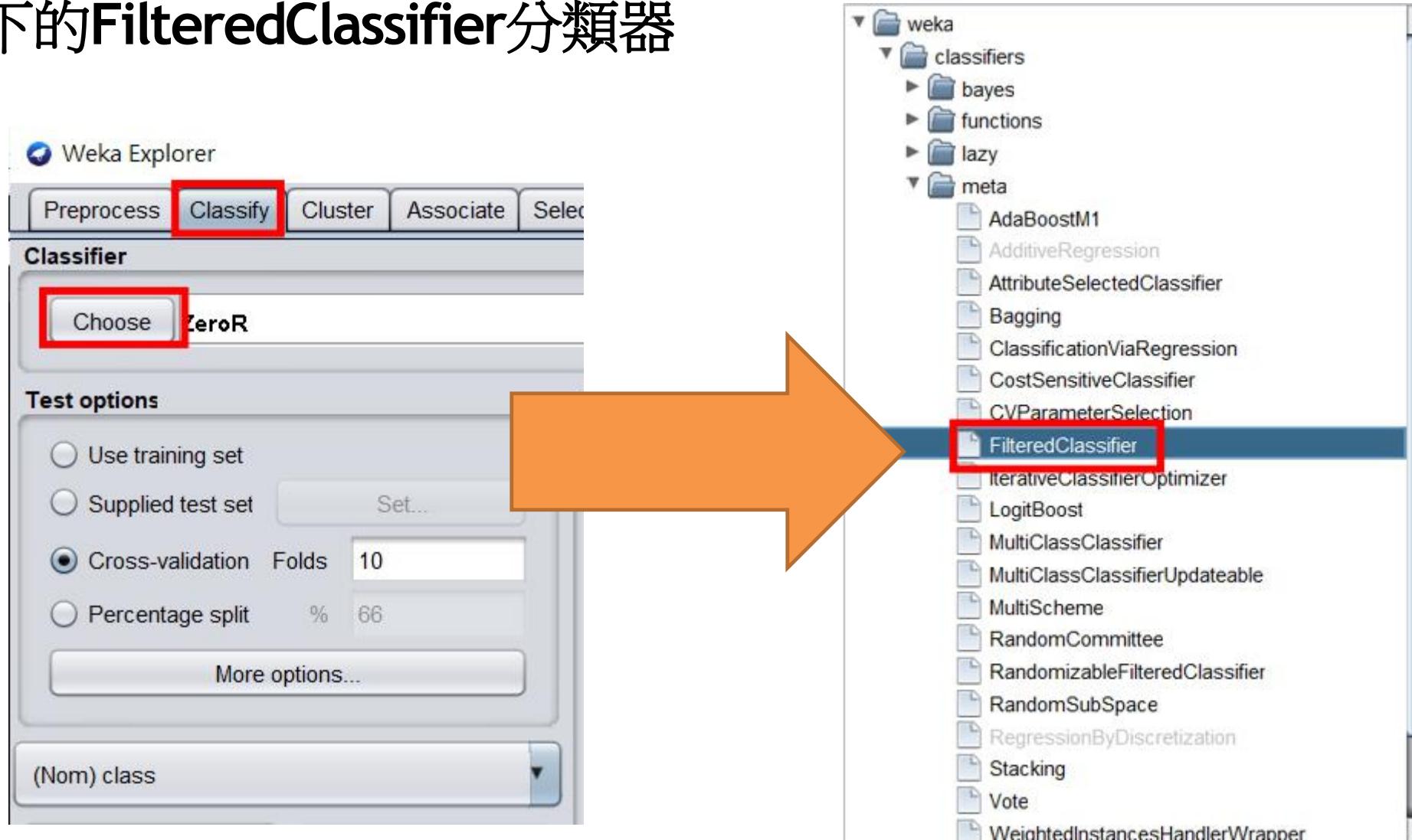
首先，設定過濾器及J48分類器並運行看看輸出結果。

1. 左鍵單擊Undo按鈕撤銷剛才的過濾器影響。



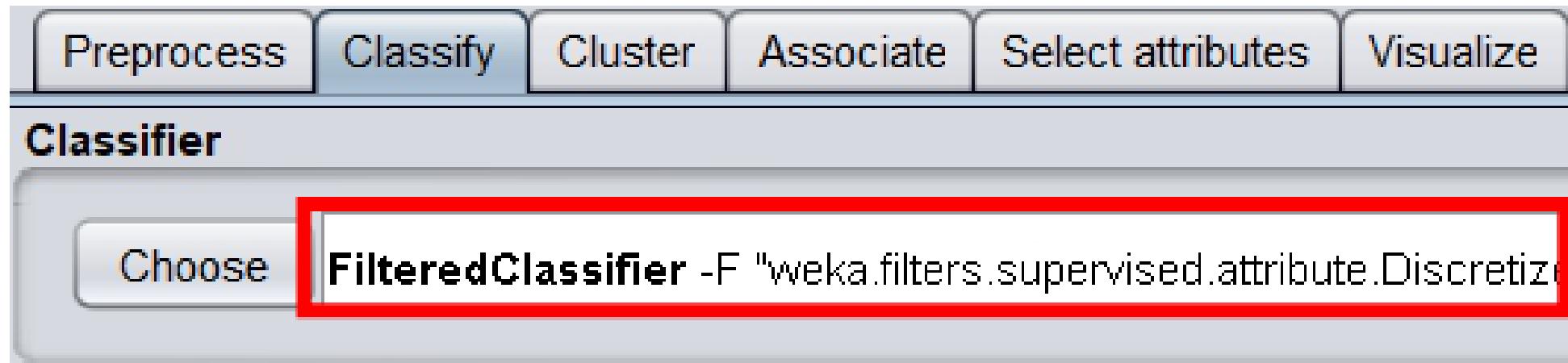
## Lesson 2.2: 監督式離散化以及FilteredClassifier

2. 切換到Classify介面點選Choose鈕，在出現的選單中左鍵單擊meta資料夾下的FilteredClassifier分類器



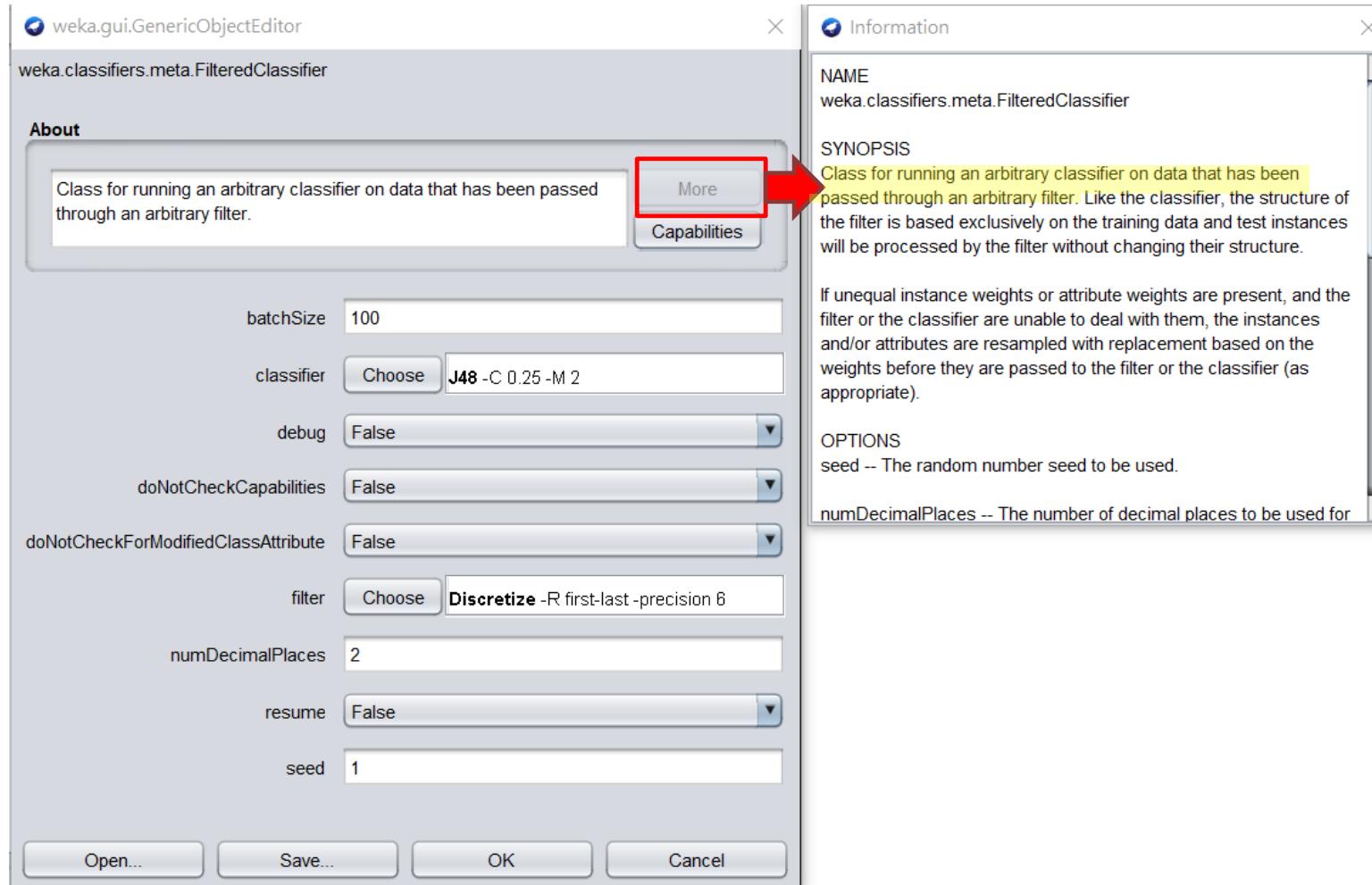
## Lesson 2.2: 監督式離散化以及*FilteredClassifier*

3. 左鍵單擊紅色方框處進行參數配置。



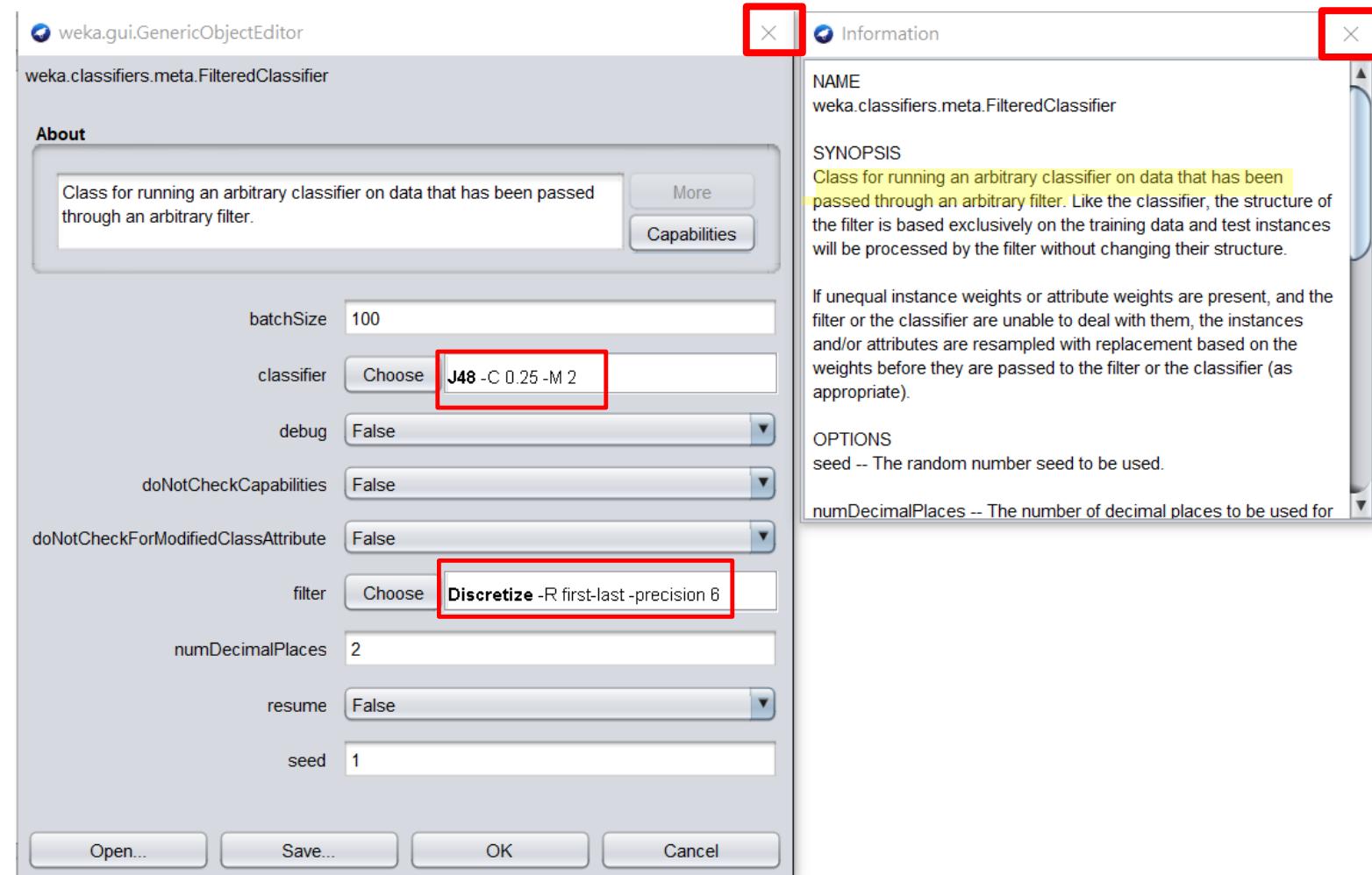
## Lesson 2.2: 監督式離散化以及FilteredClassifier

4. 左鍵單擊More按鈕查看Information，可以看到「處理被任意過濾器過濾過的資料的分類器」，這正是我們需要的。



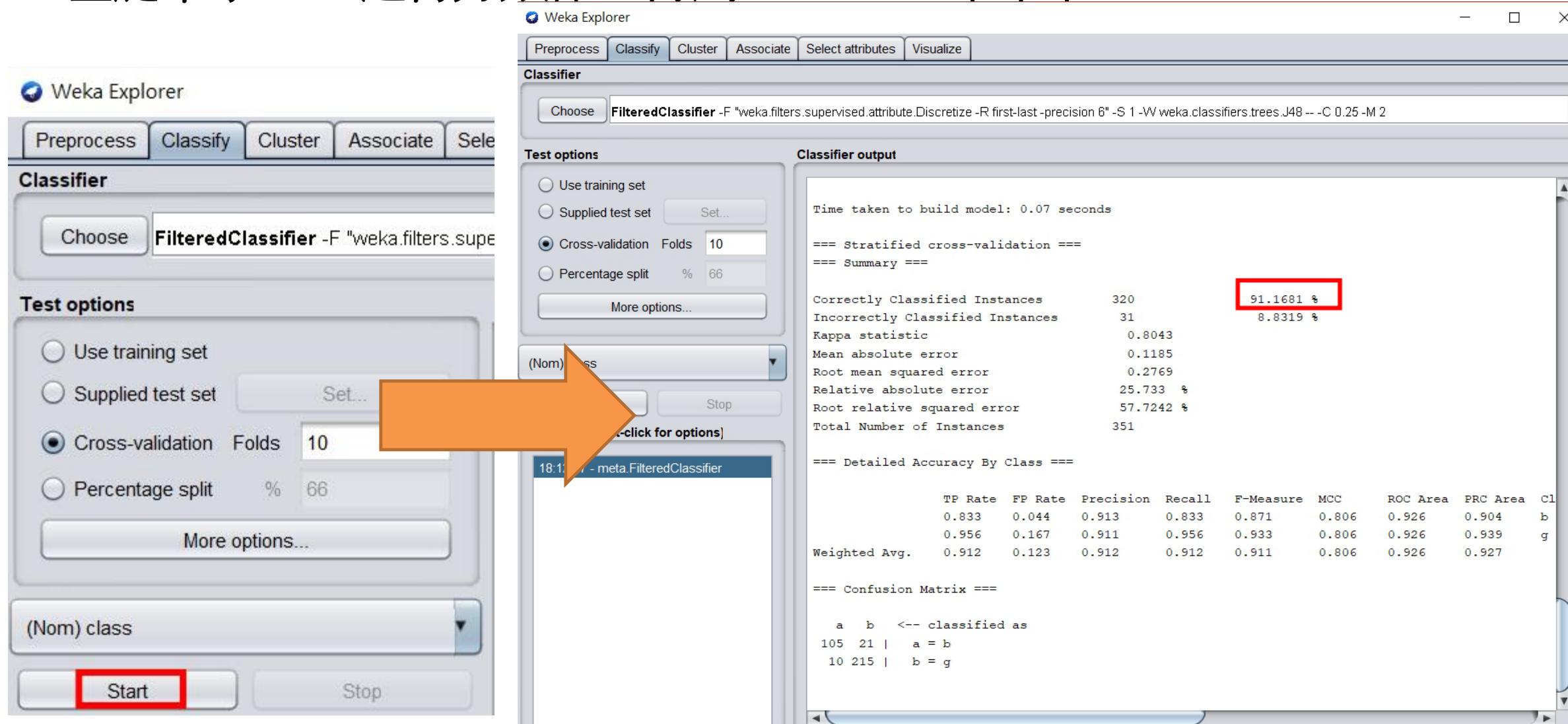
## Lesson 2.2: 監督式離散化以及FilteredClassifier

5. 參數classifier和filter的預設剛好都是我們所需要的。於是我們可以左鍵單擊information視窗右上的關閉按鈕，再以左鍵單擊配置視窗的關閉按鈕回到Classify面板。



## Lesson 2.2: 監督式離散化以及FilteredClassifier

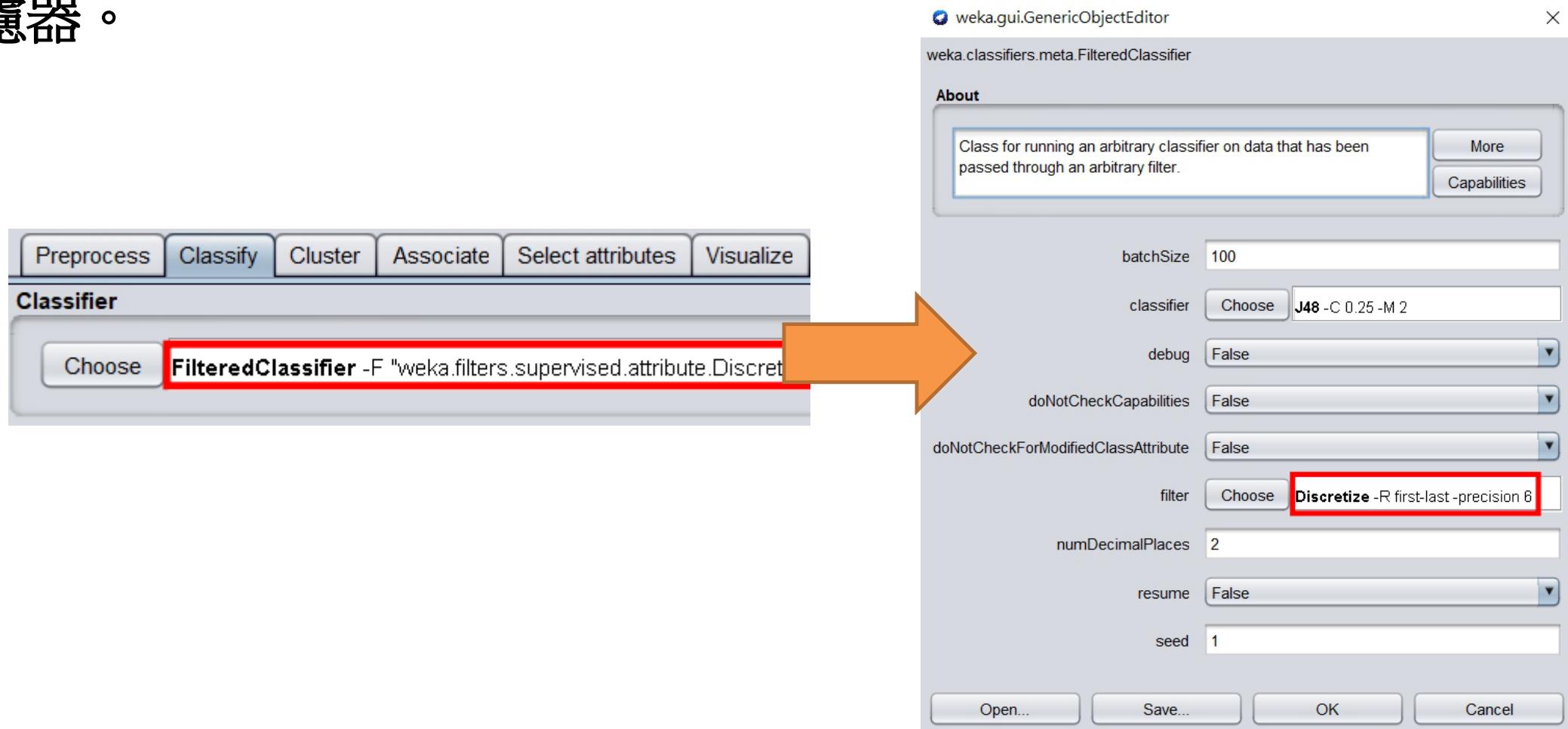
6. 左鍵單擊Start運行分類器，得到91.1681%準確率。



## Lesson 2.2: 監督式離散化以及FilteredClassifier

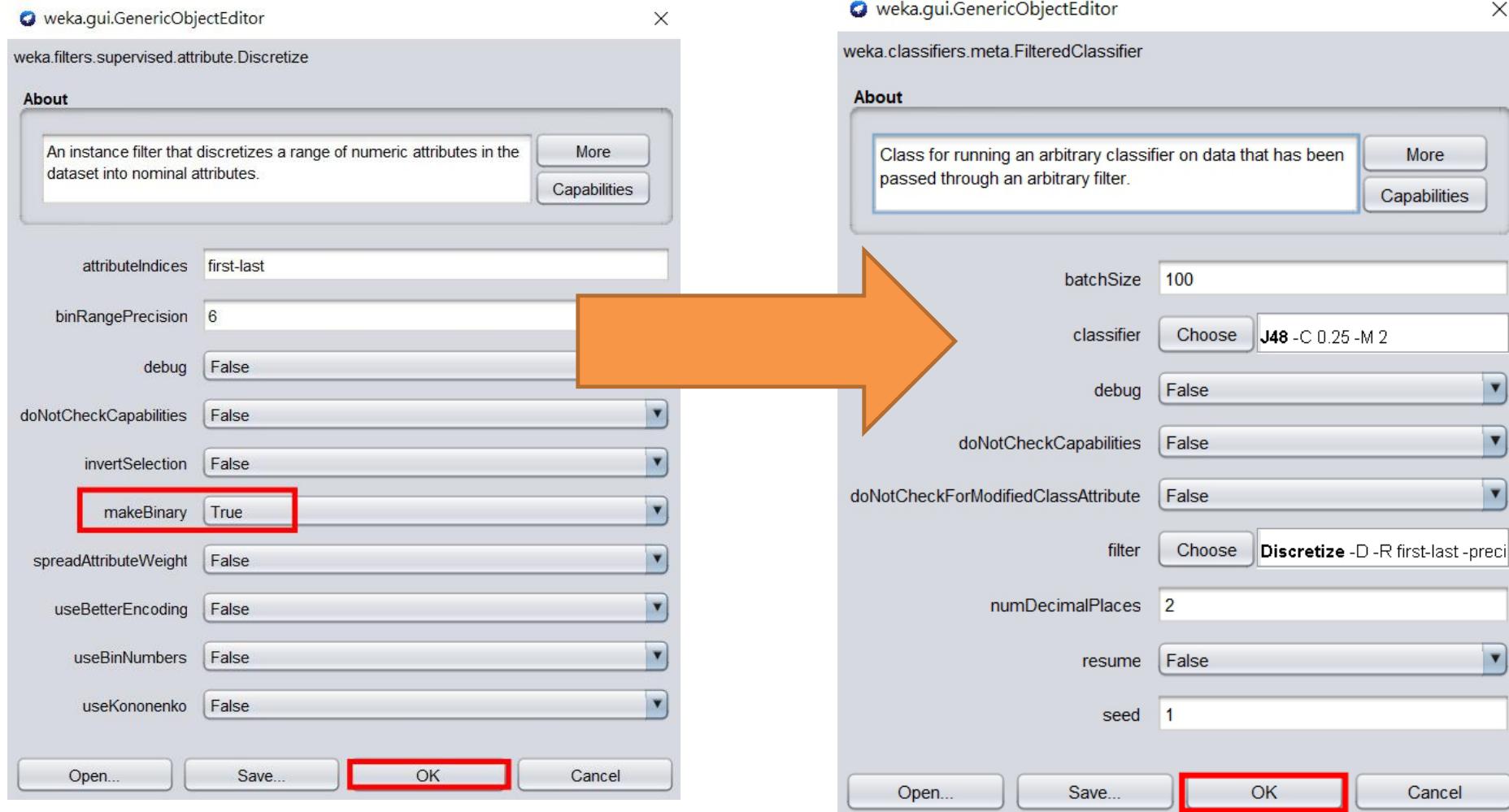
接著，配置過濾器，設定參數**makeBinary**。

1. 左鍵單擊左圖紅框處進入右圖配置視窗，以左鍵單擊右圖紅框處配置過濾器。



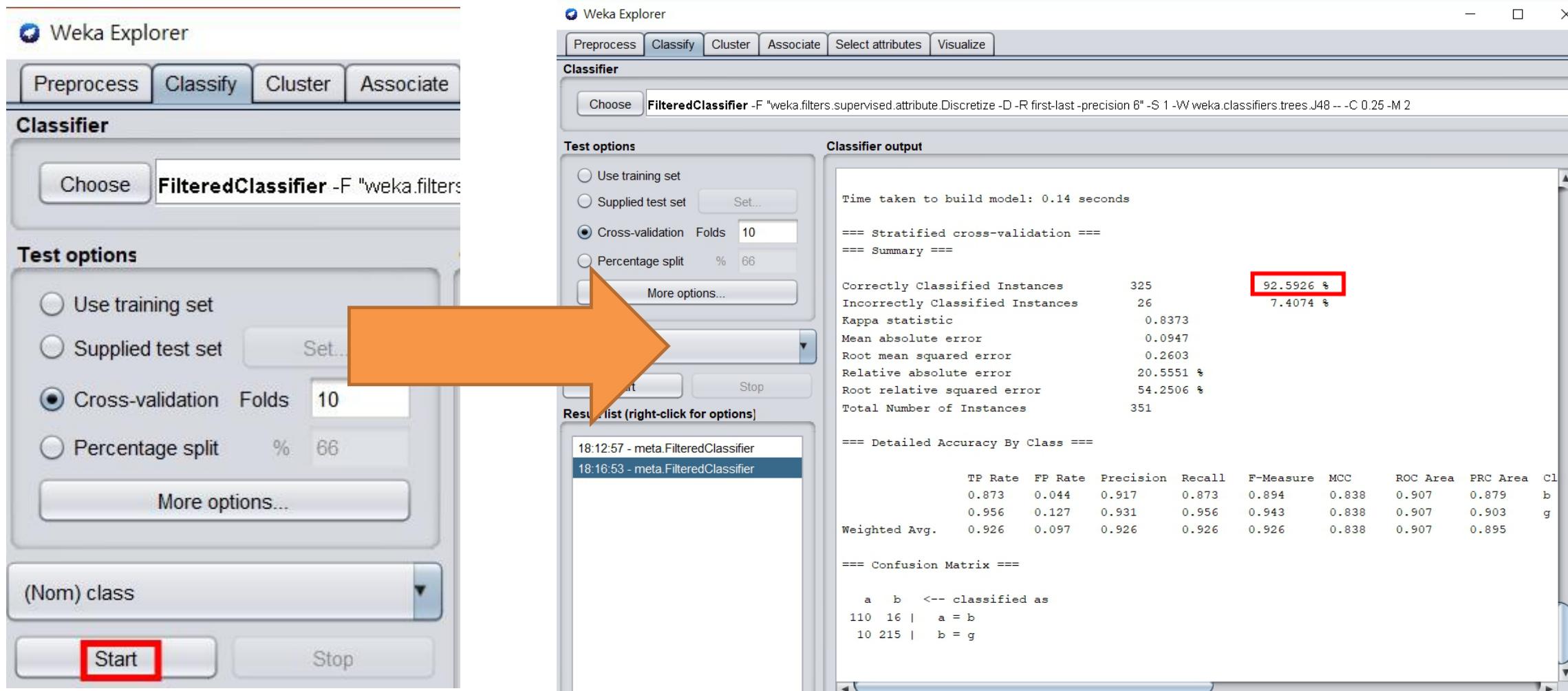
## Lesson 2.2: 監督式離散化以及FilteredClassifier

2. 將makeBinary參數設定為True並按下下方OK按鈕，回到  
FilteredClassifier配置視窗後按下下方OK按鈕。



## Lesson 2.2: 監督式離散化以及FilteredClassifier

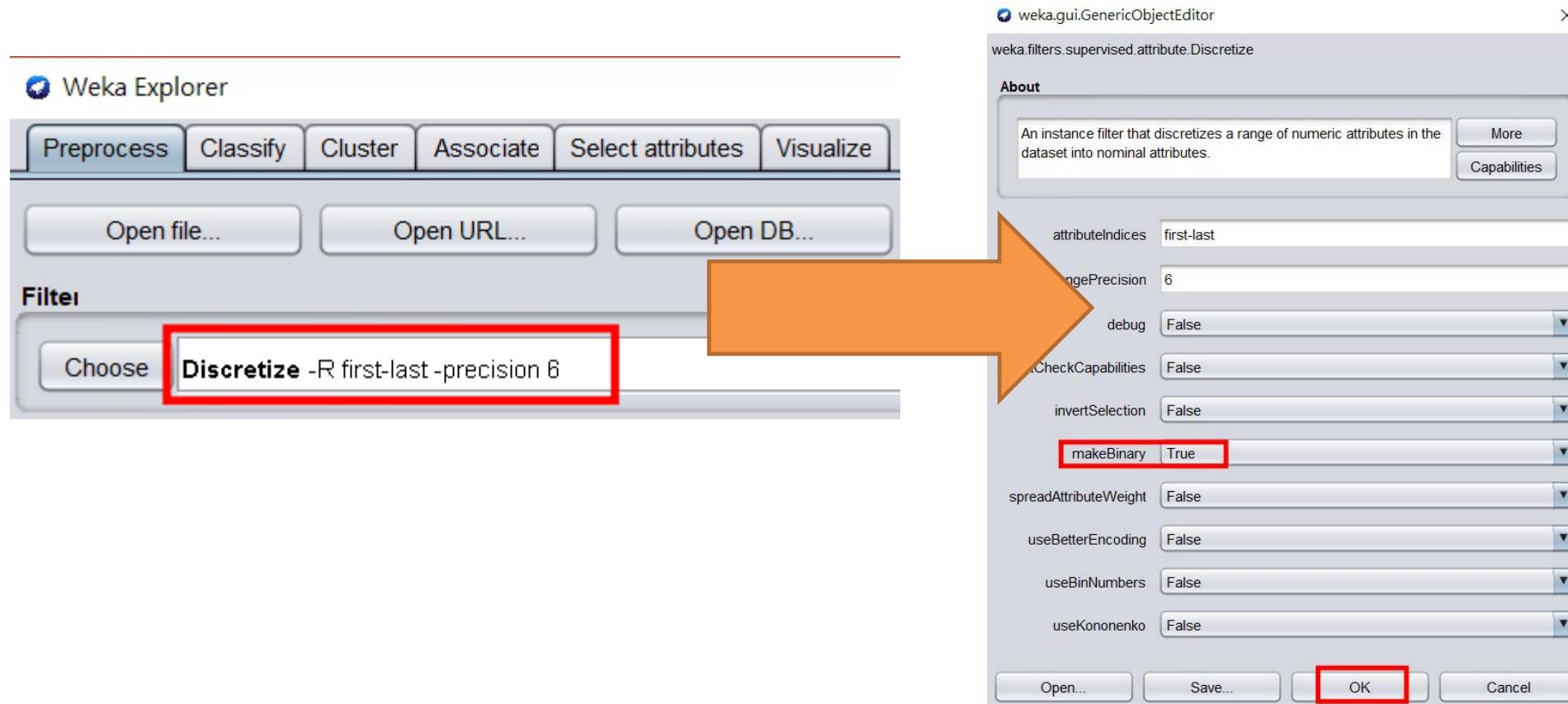
3. 回到Classify面板，左鍵單擊Start按鈕運行分類器，得到92.5926%準確率。



## Lesson 2.2: 監督式離散化以及*FilteredClassifier*

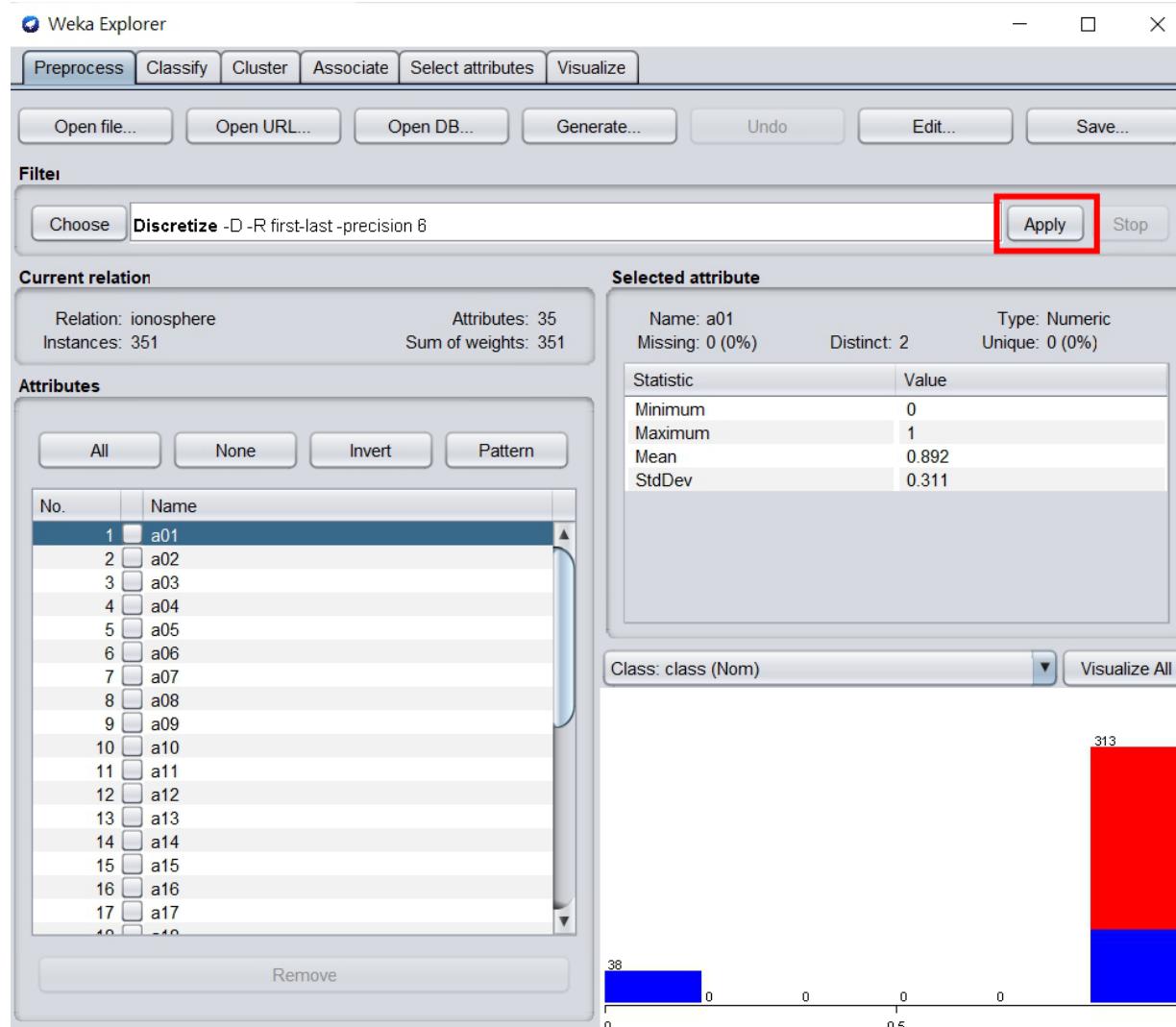
接著，使用 **makeBinary** 預先離散化試試看作弊。

1. 切換到 Preprocess 面板，左鍵單擊左圖紅框處，開啟右圖配置視窗。然後將 **makeBinary** 參數設定為 **True**，並按下下方 **OK** 按鈕。



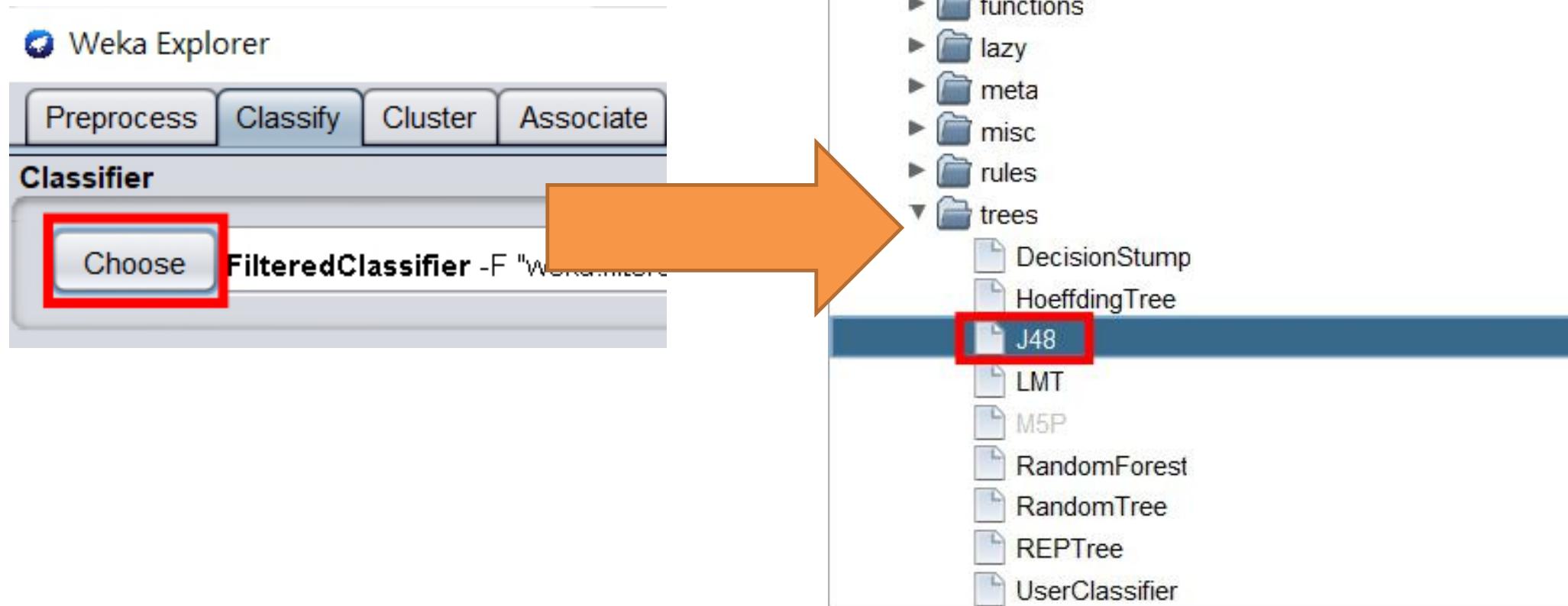
## Lesson 2.2: 監督式離散化以及FilteredClassifier

2. 回到Preprocess面板後，左鍵單擊Apply按鈕套用過濾器。



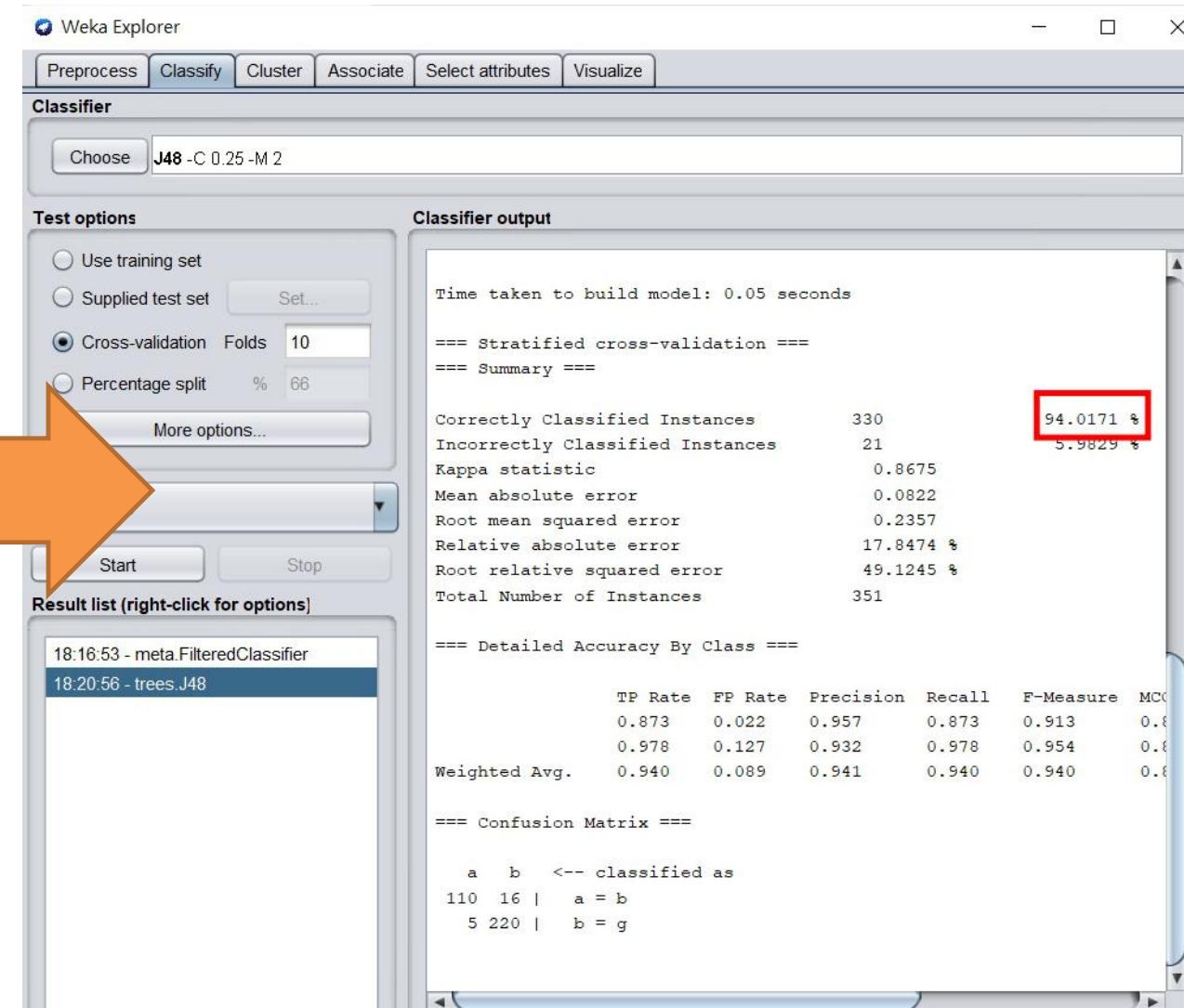
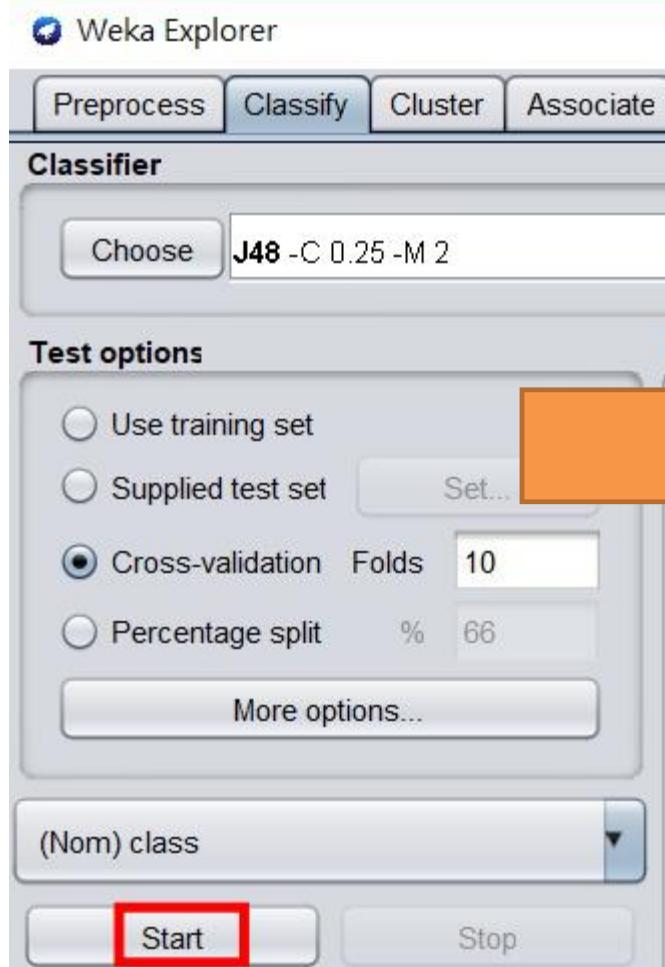
## Lesson 2.2: 監督式離散化以及*FilteredClassifier*

3. 切換到Classify介面點選Choose鈕，在出現的選單中左鍵單擊trees資料夾下的J48分類器



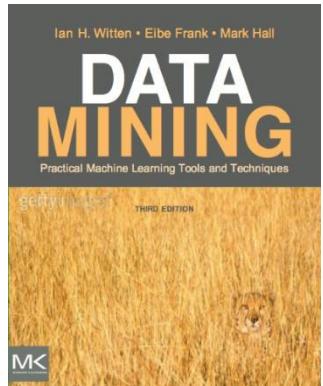
## Lesson 2.2: 監督式離散化以及FilteredClassifier

4. 左鍵單擊Start按鈕運行分類器，得到94.0171%準確率。



## Lesson 2.2: 監督式離散化以及**FilteredClassifier**

- ❖ 監督式離散化
  - 當製造離散邊界時，會把所有類別值都混在一起
- ❖ 對於測試集，必須使用訓練集決定的離散方式
- ❖ 當使用交叉驗證的時候你該怎麼做呢？
- ❖ **FilteredClassifier:** 就是為這種情況而設計的分類器
- ❖ 對其他監督式過濾器也很有用



### 課程文本

- ❖ Section 7.2 *Discretizing numeric attributes*
- ❖ Section 11.3 *Filtering algorithms*, subsection “Supervised filters”



THE UNIVERSITY OF  
**WAIKATO**  
*Te Whare Wānanga o Waikato*

# *More Data Mining with Weka*

Class 2 - Lesson 3

使用J48進行離散化

(*Discretization in J48*)

Ian H. Witten

Department of Computer Science  
University of Waikato  
New Zealand

# Lesson 2.3: 使用J48進行離散化

Class 1 探索Weka界面，處理大數據

Class 2 細散以及文本分類

Class 3 分類規則、關聯規則、聚類

Class 4 選擇屬性以及計算成本

Class 5 神經網路，學習曲線和表現優化

Lesson 2.1 細散化

Lesson 2.2 監督式離散化

Lesson 2.3 使用J48進行離散化

Lesson 2.4 Document classification

Lesson 2.5 Evaluating 2-class classification

Lesson 2.6 Multinomial Naïve Bayes

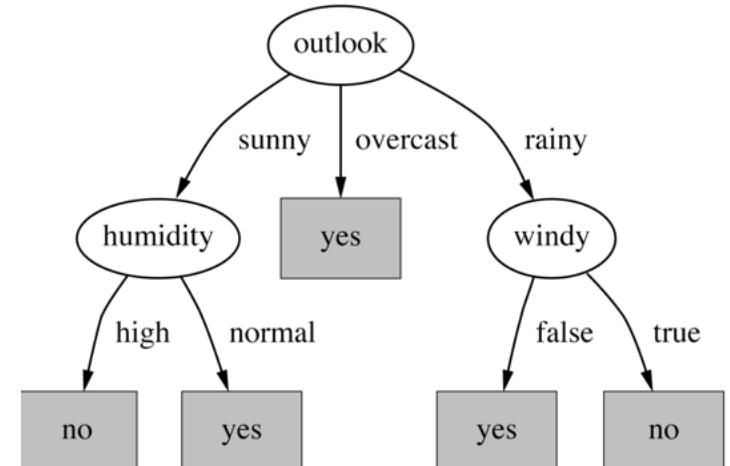


## Lesson 2.3: 使用J48進行離散化

J48如何處理數值性屬性？

由上而下遞迴地分而治之(*divide-and-conquer*) (複習)

- ❖ 選擇根節點屬性
  - 為每個可能的屬性值創造分支
- ❖ 將實例們分支成子集合
  - 從此節點延伸出來的每個分支都有一個子集合
- ❖ 對每個分支重複遞迴地執行這樣的操作
  - 只使用到達此分支的實例



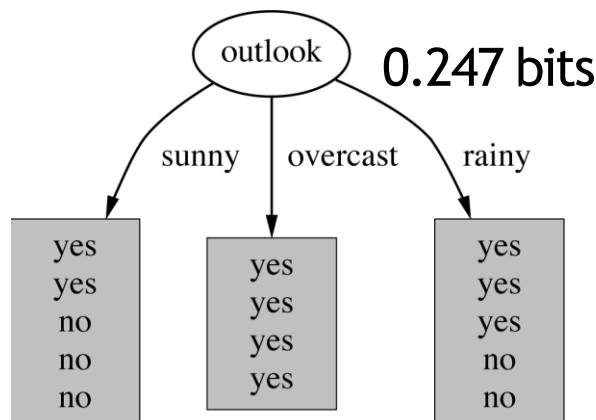
## Lesson 2.3: 使用J48進行離散化

Q: 用哪一個屬性分支最好呢？

A (J48): 那個擁有最大信息熵的屬性

### 信息增益

- 藉由知道屬性值來得到訊息量
- (分支前配給的信息熵)- (分支後配給的信息熵)
- $\text{entropy}( p_1, p_2, \dots, p_n ) = - p_1 \log p_1 - p_2 \log p_2 \dots - p_n \log p_n$



## Lesson 2.3: 使用J48進行離散化

*temperature*(溫度)屬性的信息增益

- ❖ 分支點是一個數字...而這裡有無限多個數字！
- ❖ 從訓練集中相鄰值之間的中間分割
- ❖  $n-1$  個可能 ( $n$  是訓練集的實例個數); 將全部都試一次！

9 yes, 5 no

分支前的熵= 0.940bits

64	65	68	69	70	71	72	75	80	81	83	85
yes	no	yes	yes	yes	no	no	yes	no	yes	yes	no

4 yes, 1 no      5 yes, 4 no

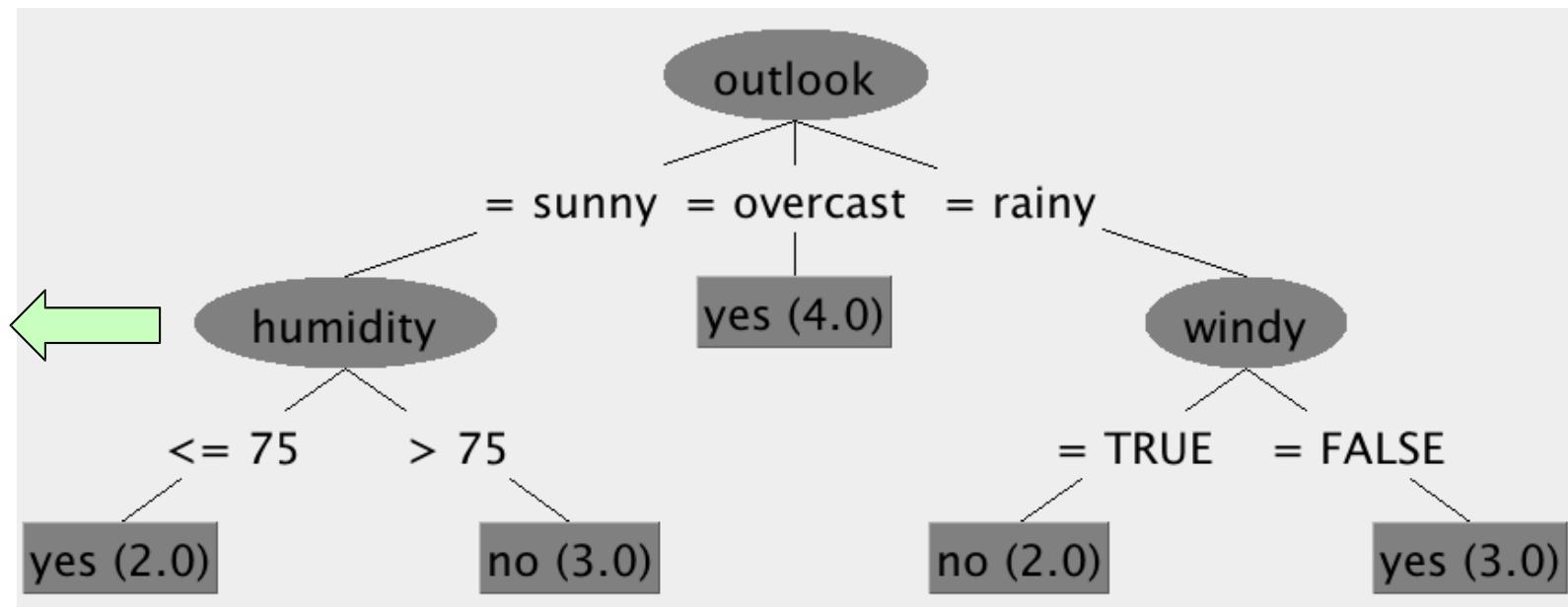
分支後的熵= 0.939bits

信息增益 = 0.001bits

## Lesson 2.3: 使用J48進行離散化

繼續往下看，我們於**humidity**屬性進行分支

Outlook	Temp	Humidity	Wind	Play
Sunny	85	85	False	No
Sunny	80	90	True	No
Sunny	72	95	False	No
Sunny	69	70	False	Yes
Sunny	75	70	True	Yes



*humidity* 把 no和yes 分開。

我們從{70,70} 和 {85}中間值進行分支,如75 (!)

## Lesson 2.3: 使用J48進行離散化

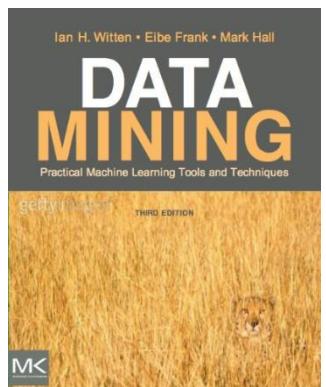
創造樹的時候進行離散化 vs. 預先離散化(pre-discretization)

- ❖ 更具體的情況有助於決定離散邊界
- ❖ 但是我們的決定僅基於整體訊息的小子集合
  - ... 特別是越向下，直到樹的底部，我們可以使用的實例也會越來越少
- ❖ 對於所有內部節點，到達它的實例必須對每個數字屬性進行單獨排序
  - ... 且排序會有複雜度 $O(n \log n)$
  - ... 但如果採用一個更好的資料結構可以避免重複排序

## Lesson 2.3: 使用J48進行離散化

- ❖ C4.5/J48很早就採用離散化
- ❖ 預先離散化(**pre-discretization**)是一種選擇，開發/完善得較晚
  - 監督式離散化本質上使用相同的熵啟發法
  - 可以保留數字屬性所蘊含的排序信息
- ❖ J48內部離散化是否會優於預先離散化？
  - 有人支持，有人反對
- ❖ 這是一個實驗性問題
  - 對於其他分類器也是如此

課程文本  
*Section 6.1 Decision trees*





THE UNIVERSITY OF  
**WAIKATO**  
*Te Whare Wānanga o Waikato*

# *More Data Mining with Weka*

Class 2 - Lesson 4

文本分類

*(Document classification)*

Ian H. Witten

Department of Computer Science University of  
Waikato  
New Zealand

# Lesson 2.4: 文本分類

Class 1 探索Weka介面，處理大數據

Class 2 離散以及文本分類

Class 3 分類規則、關聯規則、聚類

Class 4 選擇屬性以及計算成本

Class 5 神經網路，學習曲線和表現優化

Lesson 2.1 Discretization

Lesson 2.2 監督式離散化

Lesson 2.3 使用J48進行離散化

Lesson 2.4 文本分類

Lesson 2.5 Evaluating 2-class classification

Lesson 2.6 Multinomial Naïve Bayes



## Lesson 2.4: 文本分類

### 一些訓練資料

Document text	Classification
The price of crude oil has increased significantly	yes
Demand of crude oil outstrips supply	yes
Some people do not like the flavor of olive oil	no
The food was very oily	no
Crude oil is in short supply	yes
Use a bit of cooking oil in the frying pan	no

這裡有六篇文檔。每個文檔里面只有一句話，且被分為**yes**和**no**兩類。  
通過閱讀，你會發現這些文檔都是關於油的。

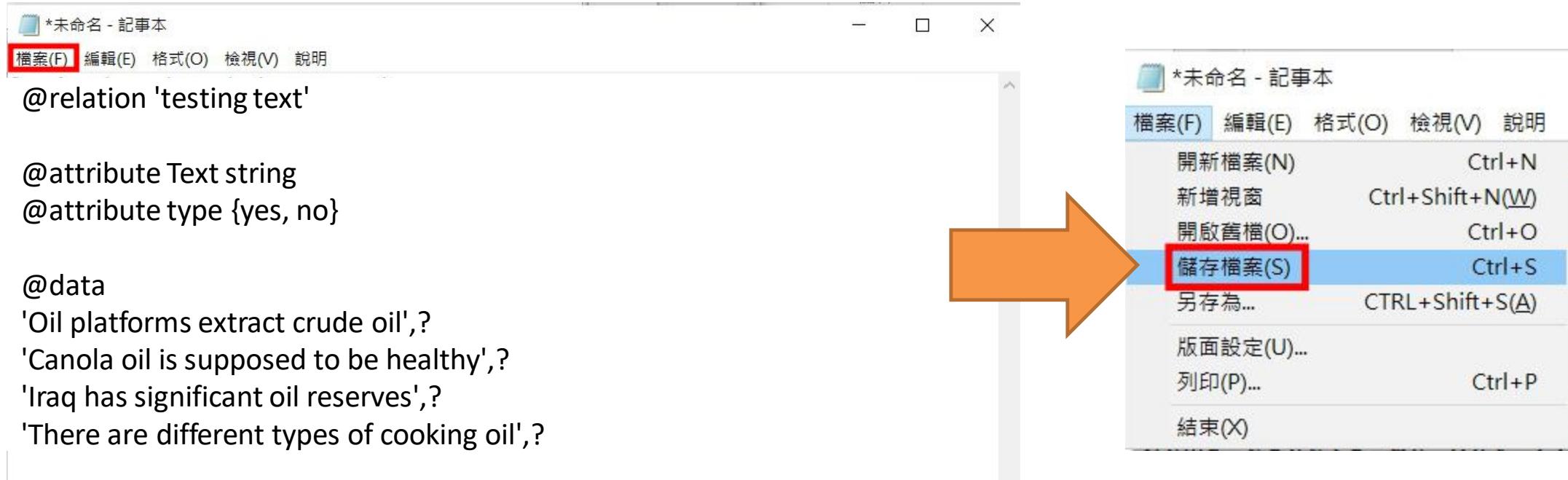
**yes**類的文檔是關於地下開採的油；**no**類的文檔是關於食用油，比如 “the food was very oily”。

## Lesson 2.4: 文本分類

- ❖ 將文本載入Weka; 注意屬性要標註“string”
- ❖ 套用 **StringToWordVector**過濾器 (非監督式屬性過濾器)
- ❖ 創造了 33 種新的屬性
  - *Crude, Demand, The, crude, has, in, increases, is, of, oil, ...*
- ❖ 二元的, 數值的屬性
- ❖ 使用 **J48** (必須設定類別屬性(class attribute))
- ❖ 使用訓練集評估
- ❖ 視覺化這顆樹

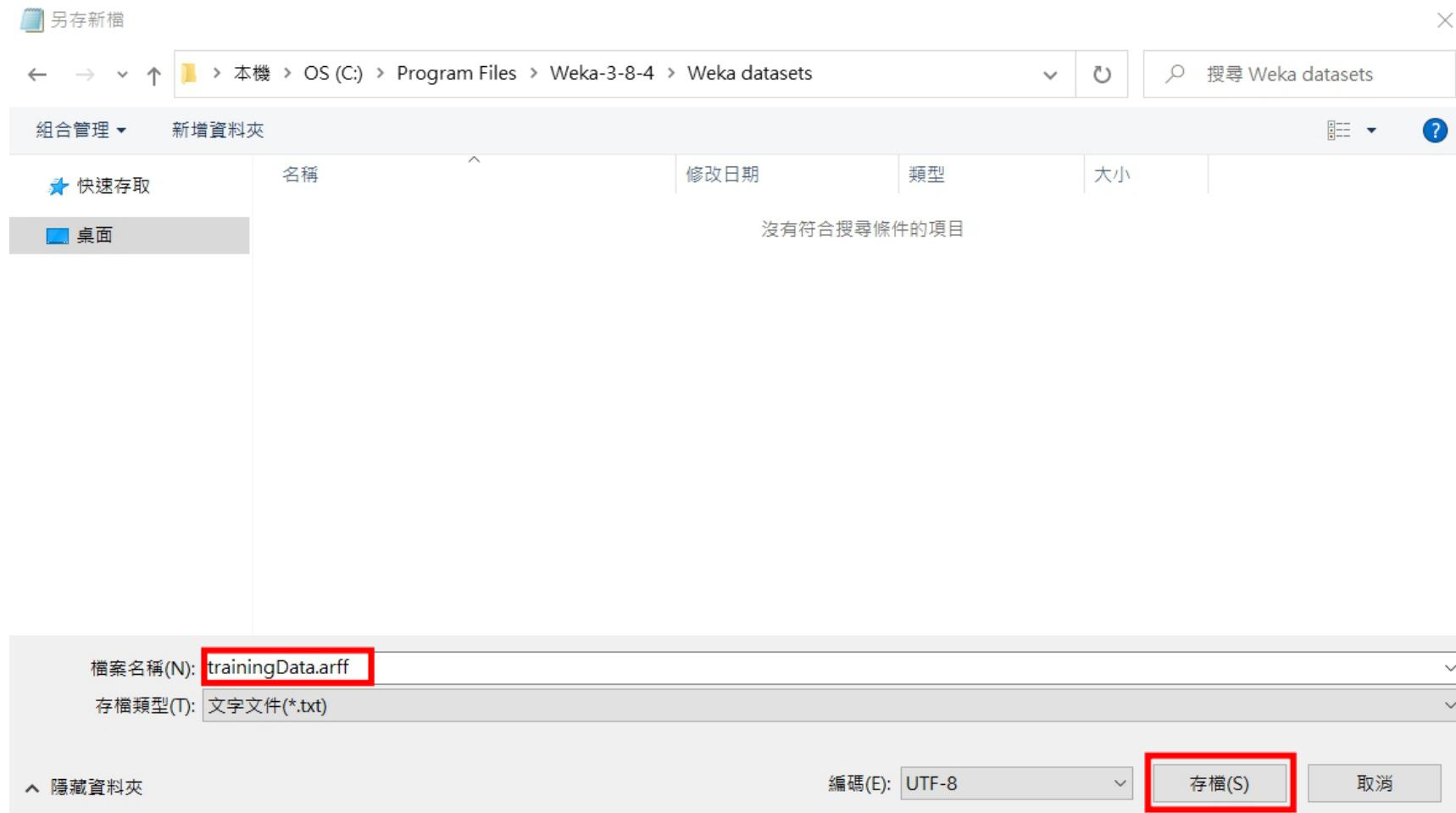
## Lesson 2.4: 文本分類

1. 首先將下列訓練文本複製並貼到記事本，然後左鍵單擊工具列中的檔案(F)按鈕，並在選單中左鍵單擊儲存檔案(S)。



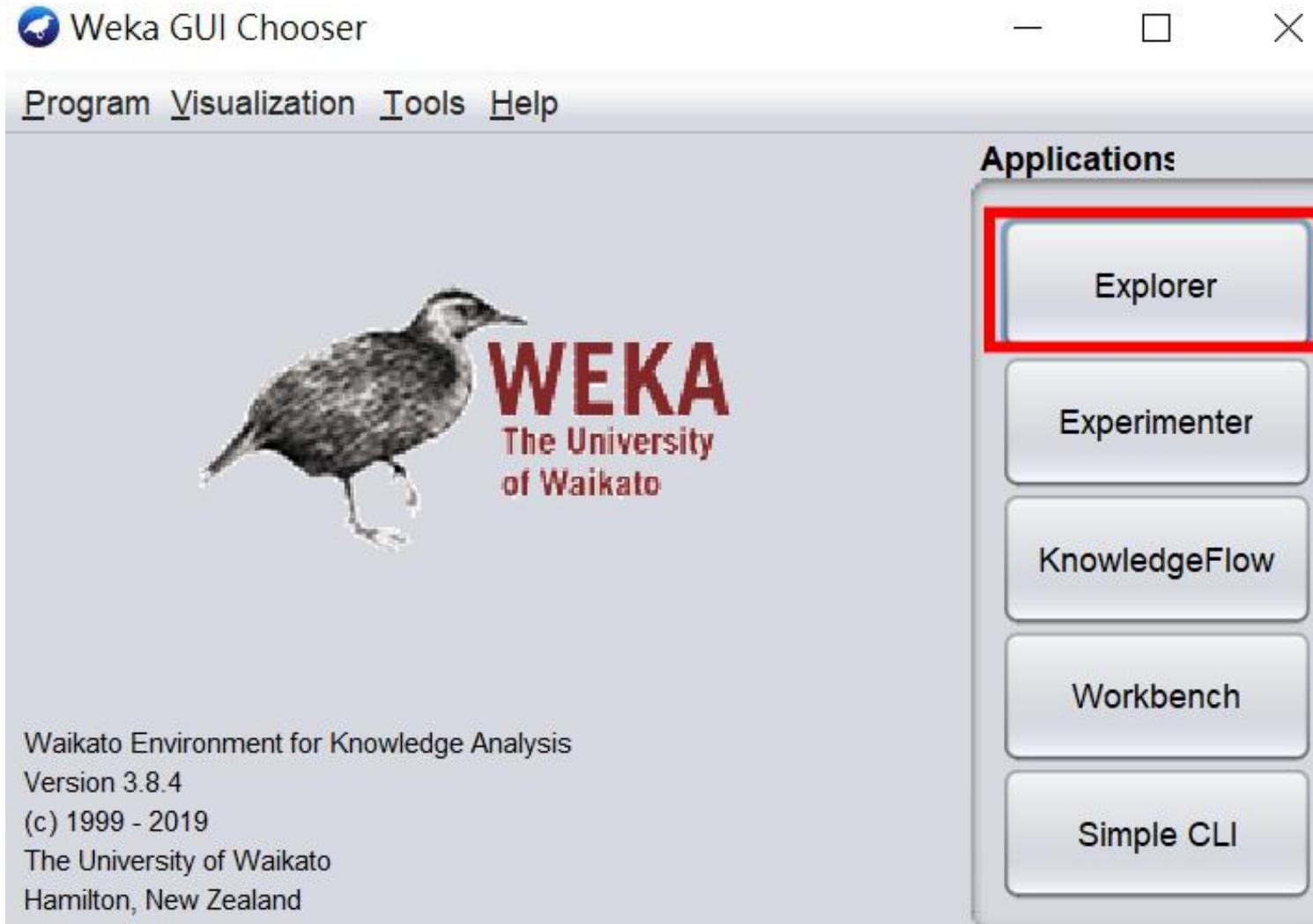
## Lesson 2.4: 文本分類

2. 將檔案儲存在自行複製的Weka datasets資料夾中，並命名為 **trainingData.arff**，接著左鍵單擊下方存檔(S)按鈕。



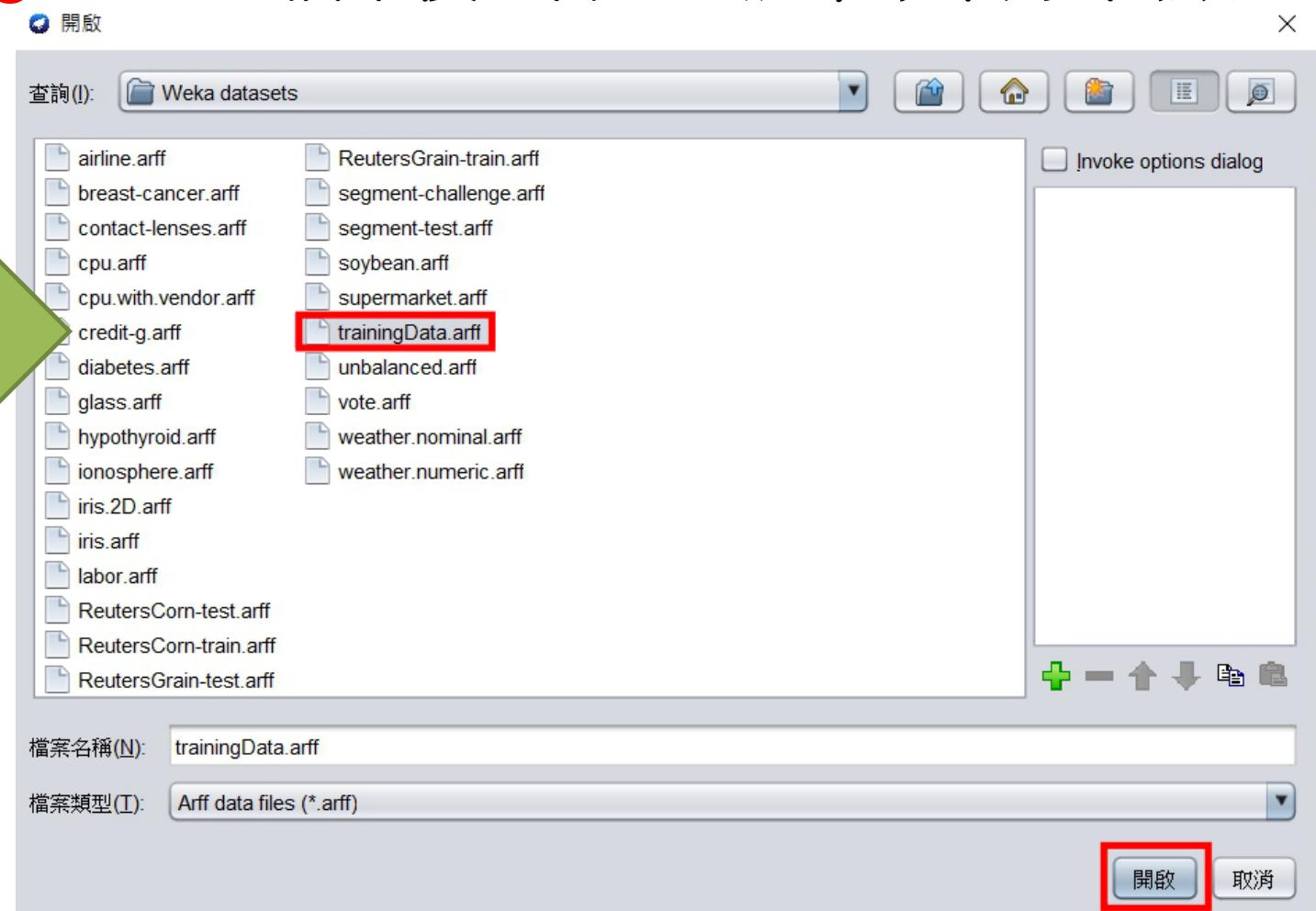
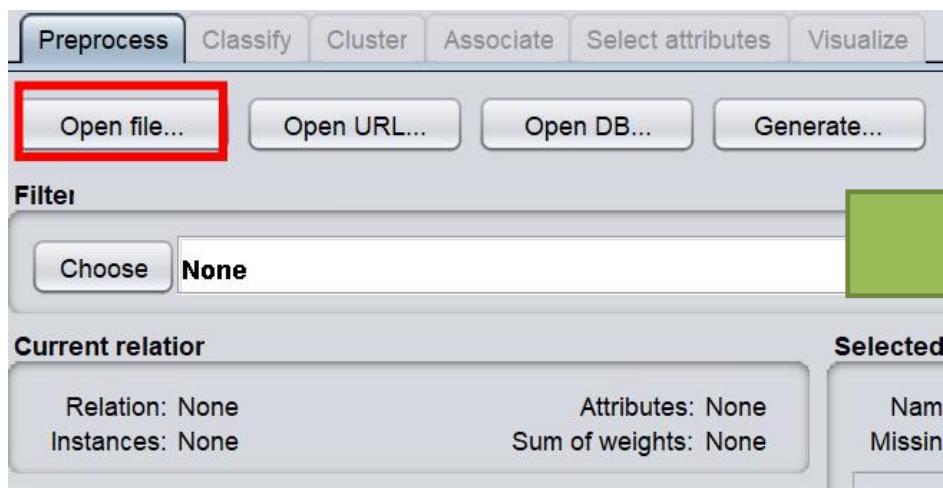
## Lesson 2.4: 文本分類

### 3. 開啟Weka的Explorer



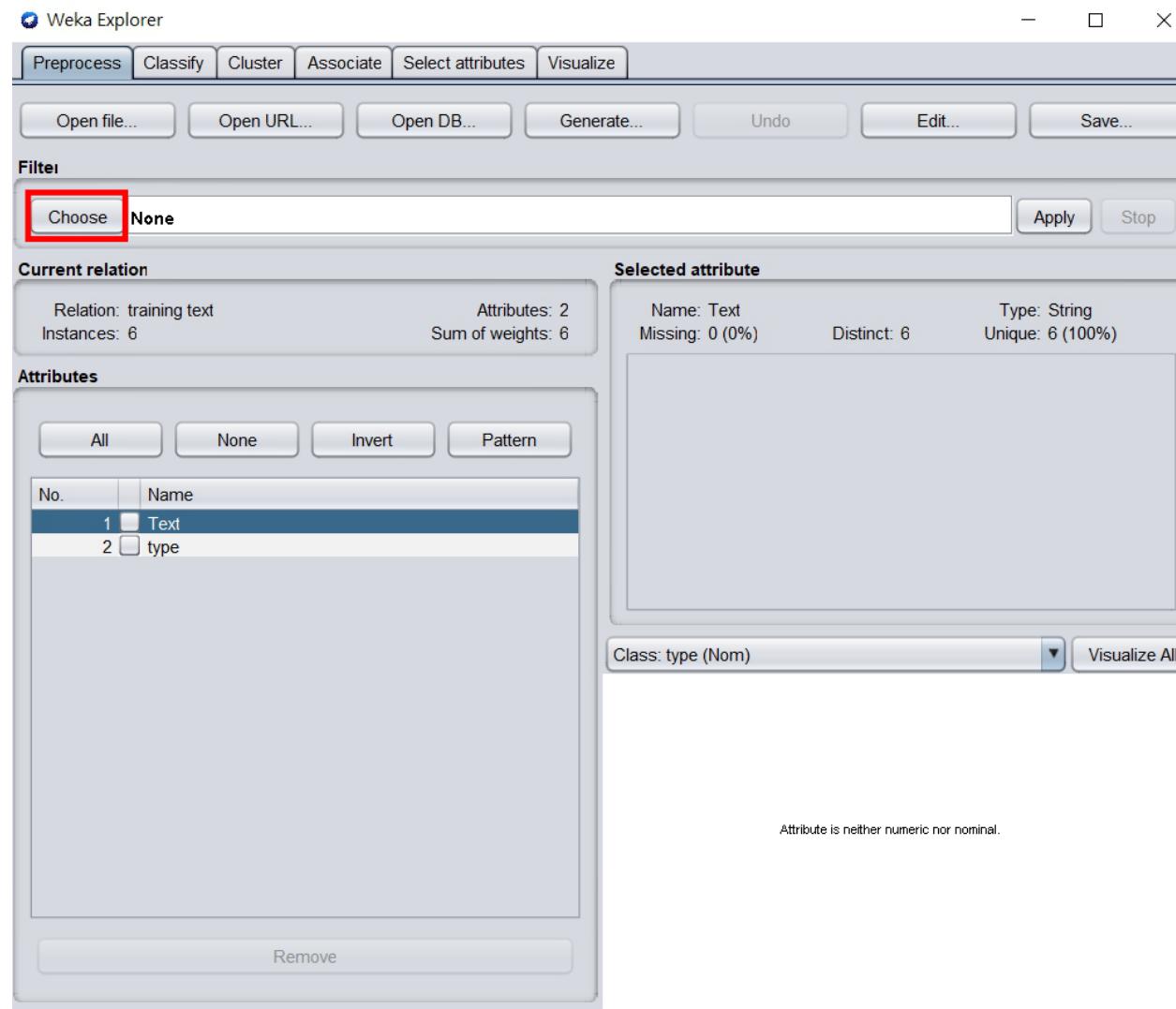
## Lesson 2.4: 文本分類

4. 左鍵點擊Open file...開啟右圖視窗，進入自行複製的Weka datasets，左鍵單擊剛才儲存的**trainingData.arff**檔案後，再以左鍵單擊下方”開啟”以載入此檔案



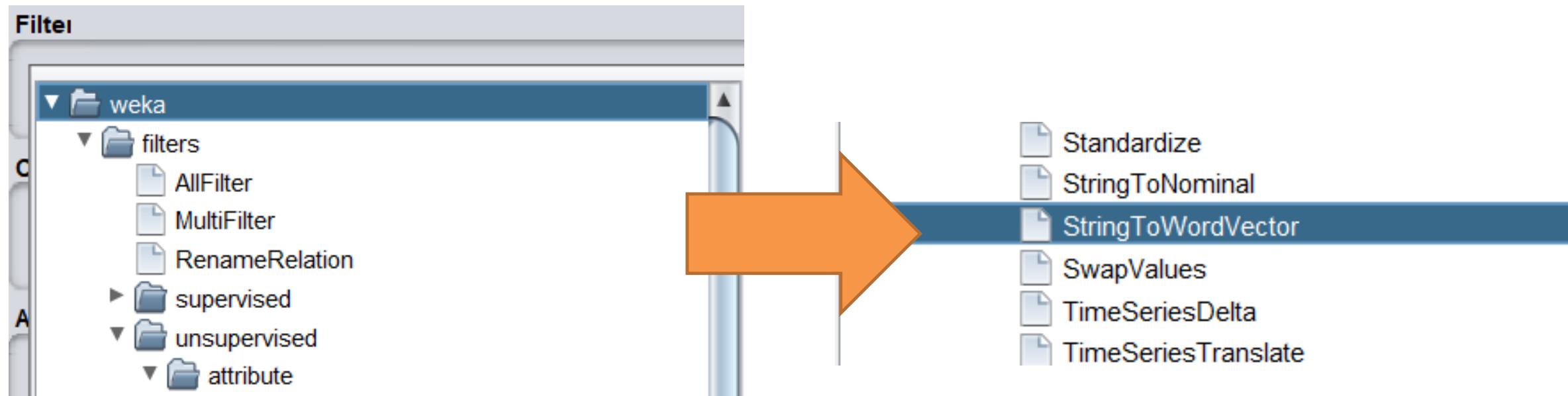
## Lesson 2.4: 文本分類

5. 在Preprocess面板中左鍵單擊Choose按鈕選擇過濾器。



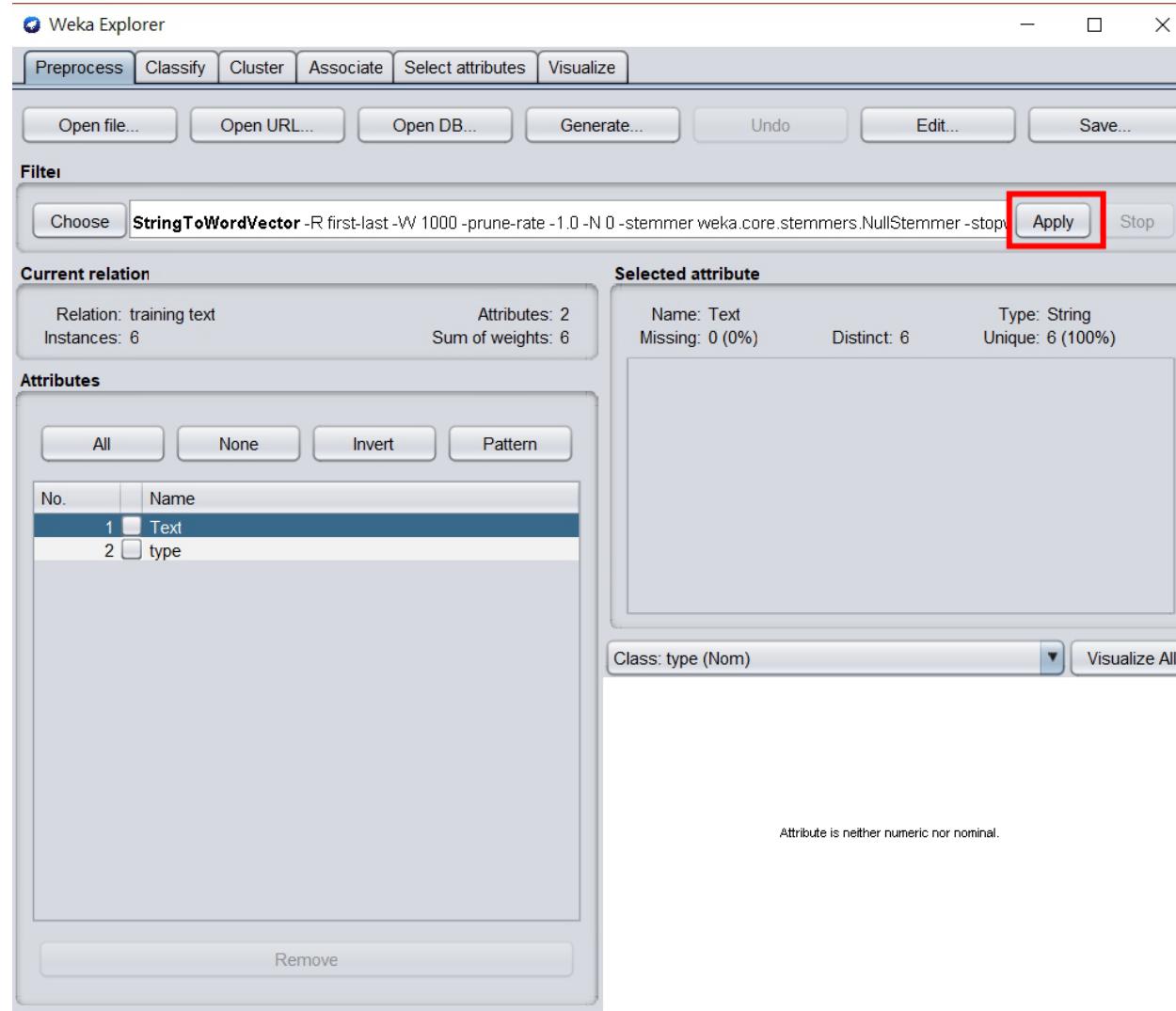
## Lesson 2.4: 文本分類

6. 左鍵單擊weka/unsupervised/attribute路徑下的StringToWordVector。



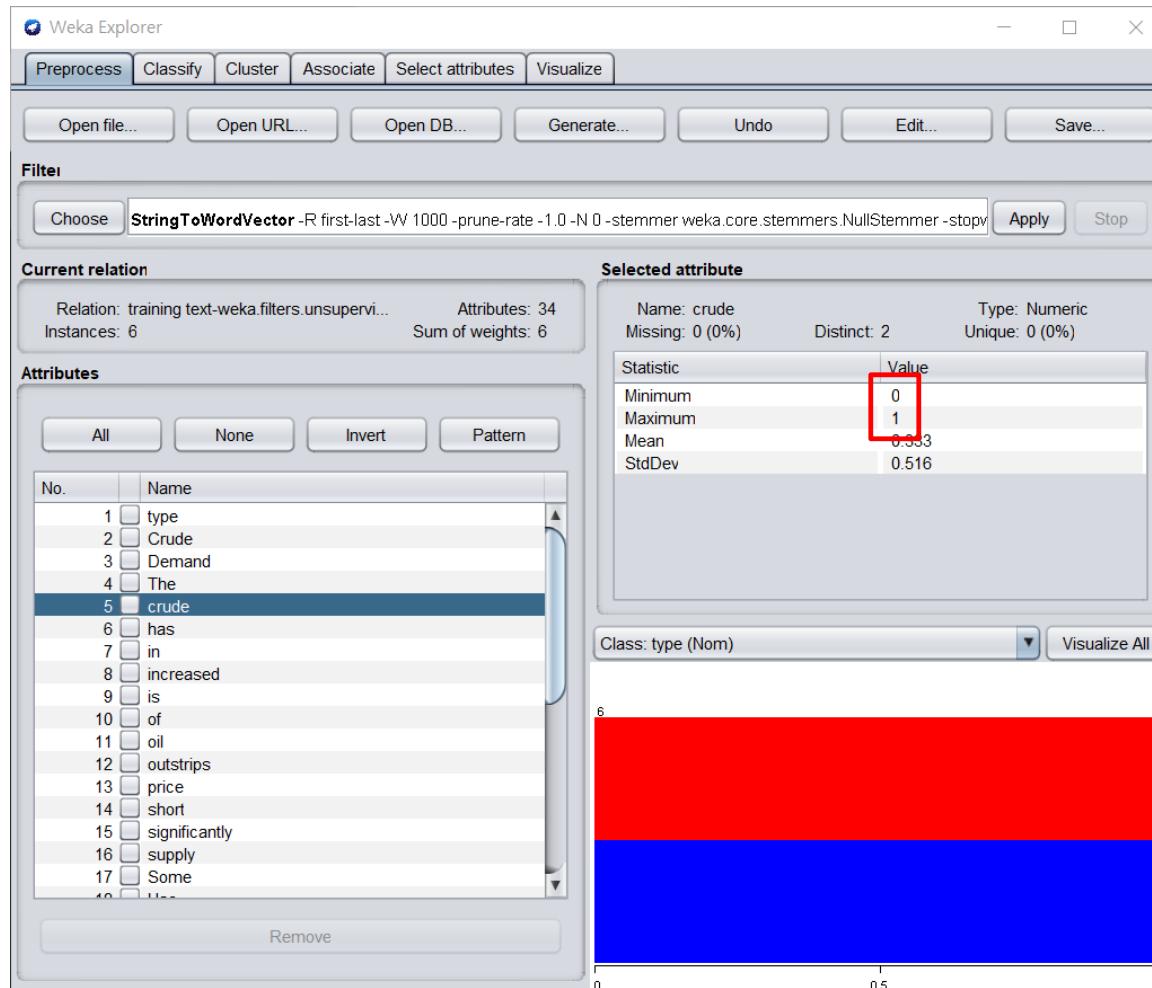
## Lesson 2.4: 文本分類

7.回到Preprocess面板，左鍵單擊Apply按鈕套用過濾器。



## Lesson 2.4: 文本分類

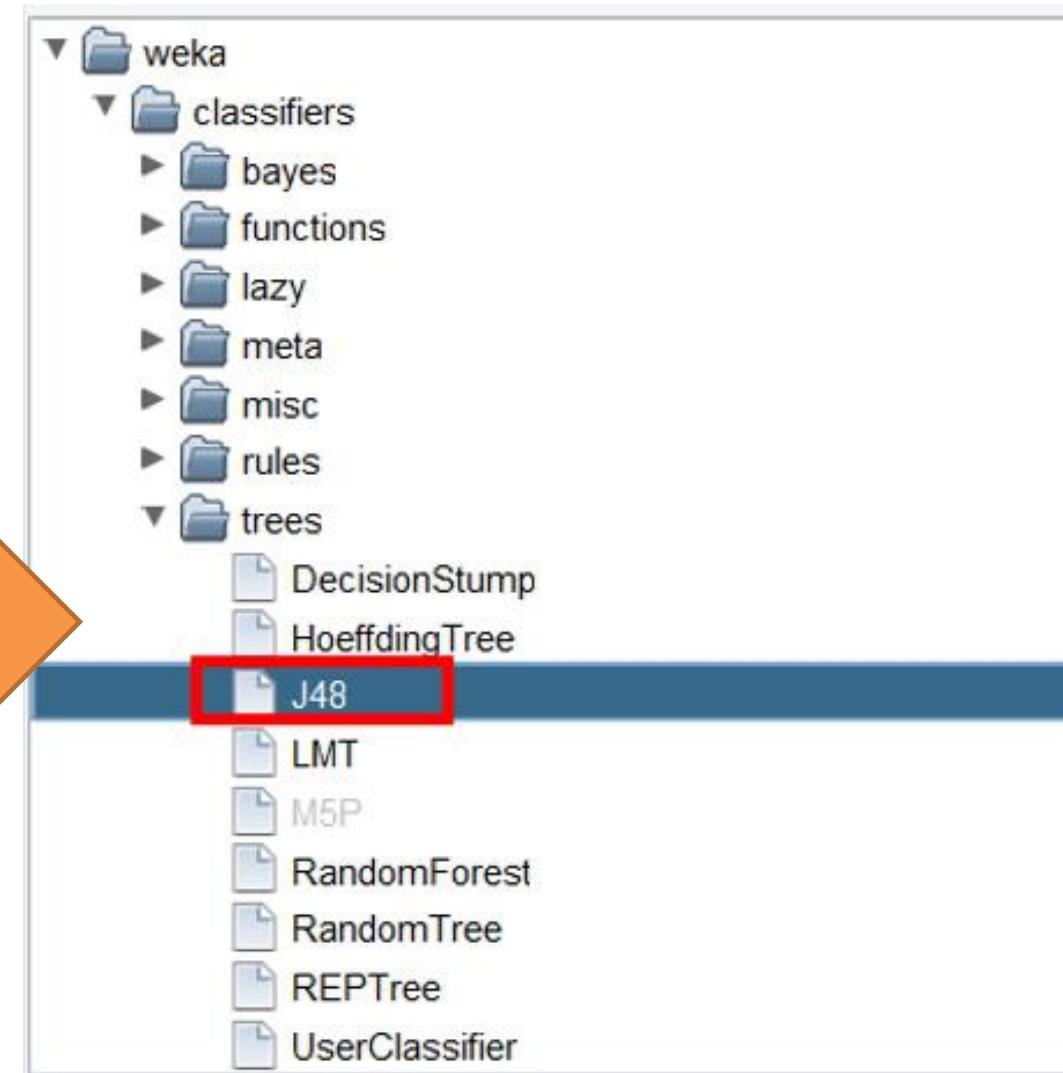
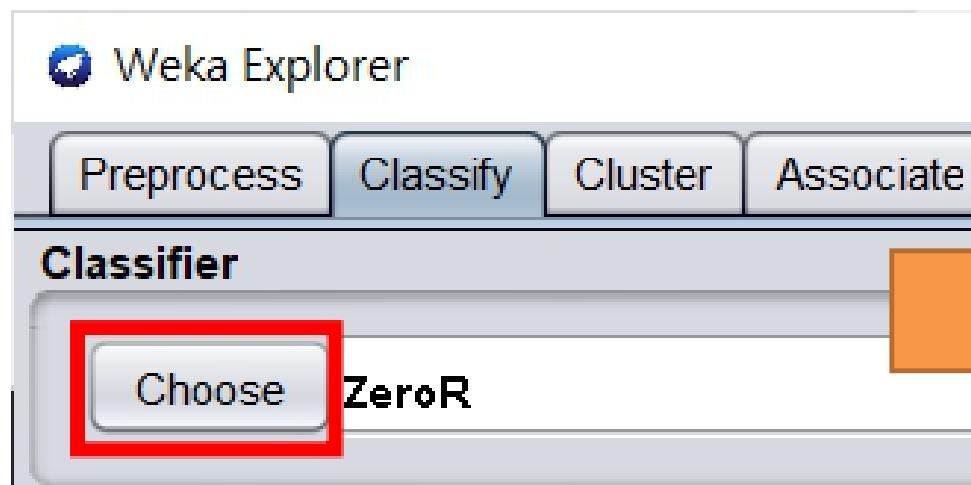
▼執行結果：以crude說明，它的屬性值是一個數字，有兩個值，0和1：如果沒有在文中出現，屬性值為0；如果在文中出現了，屬性值為1。



## Lesson 2.4: 文本分類

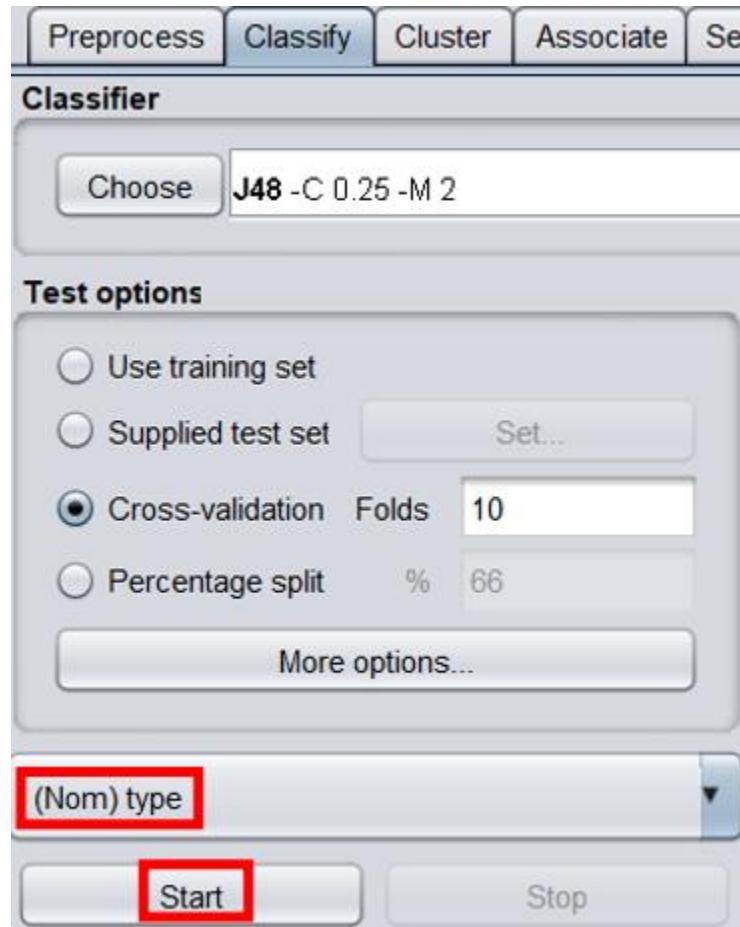
試著用分類器運行看看。

8. 切換到Classify介面點選Choose鈕，在出現的選單中左鍵單擊trees資料夾下的J48分類器。



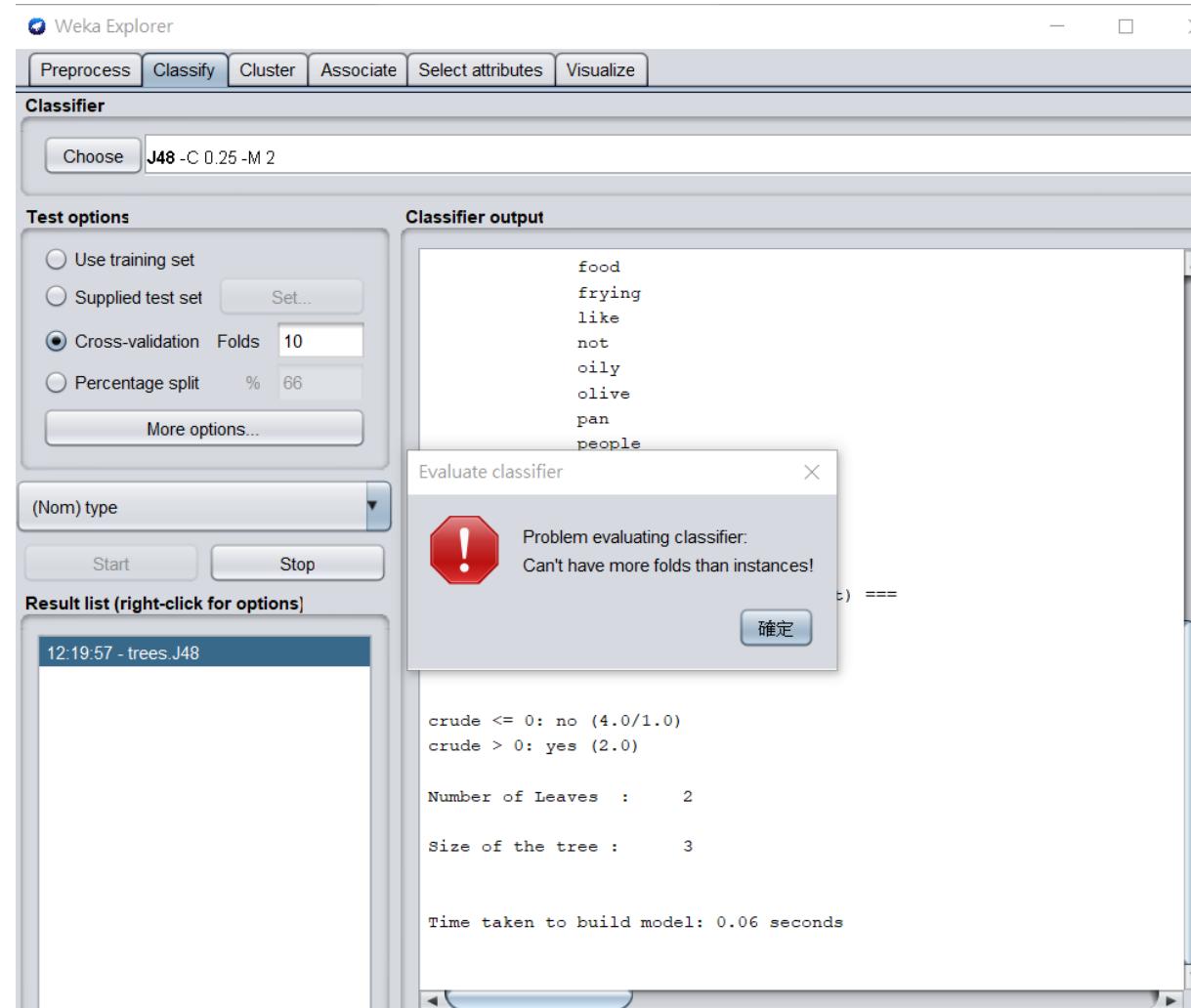
## Lesson 2.4: 文本分類

9.回到Classify面板，不同於教學影片的是J48並沒有反白，因為系統預設選好(Nom)type的選項了！我們接著左鍵單擊Start運行J48。



## Lesson 2.4: 文本分類

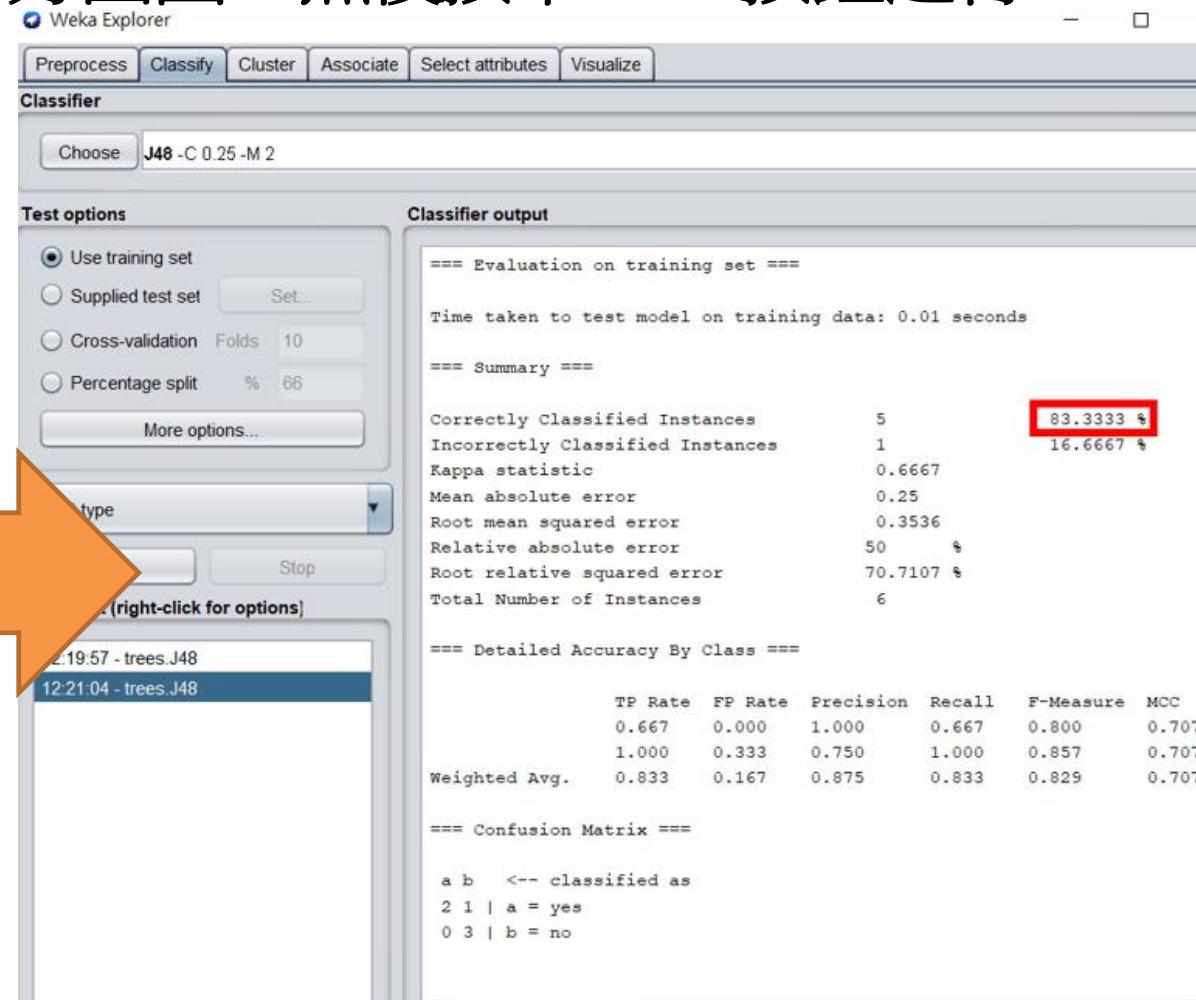
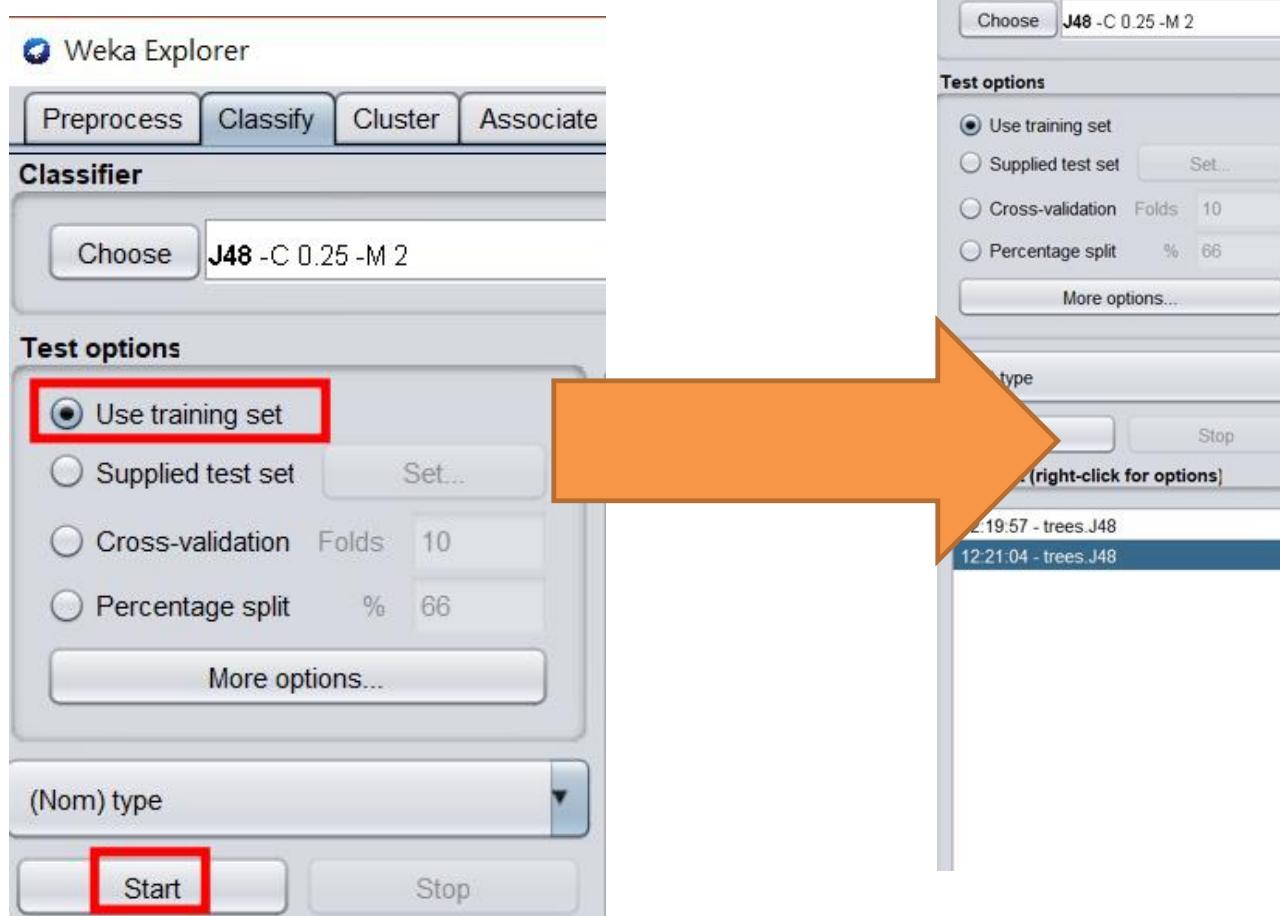
▼執行結果：發生錯誤。因為訓練集只有六個實例，我們不能使用十層交叉驗證。



## Lesson 2.4: 文本分類

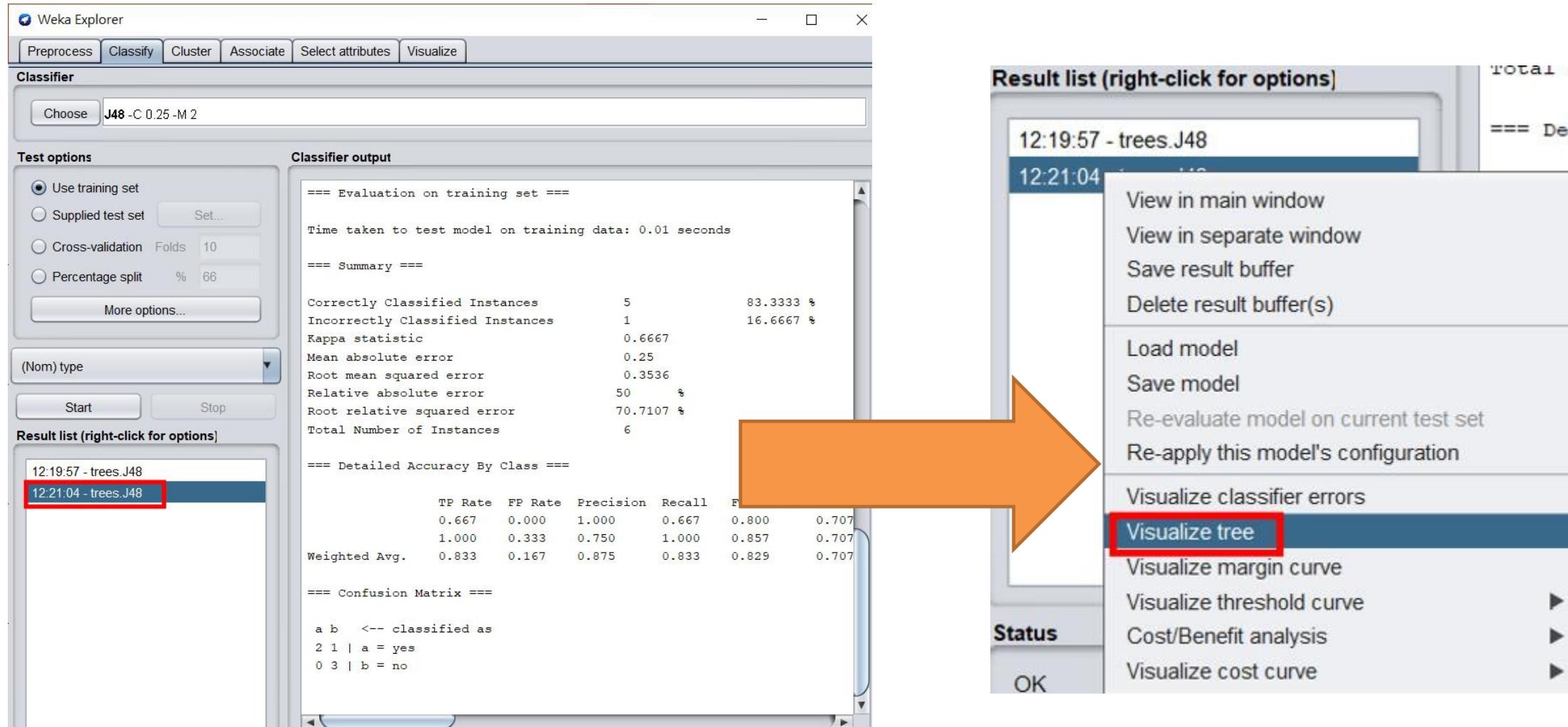
我們試著用訓練資料執行看看。

10. 左鍵點選Use training set前方圓圈，然後按下Start按鈕運行。  
得到83.333%準確率。



## Lesson 2.4: 文本分類

11. 右鍵單擊剛才的執行紀錄，在出現的選單中左鍵單擊Visualize tree。

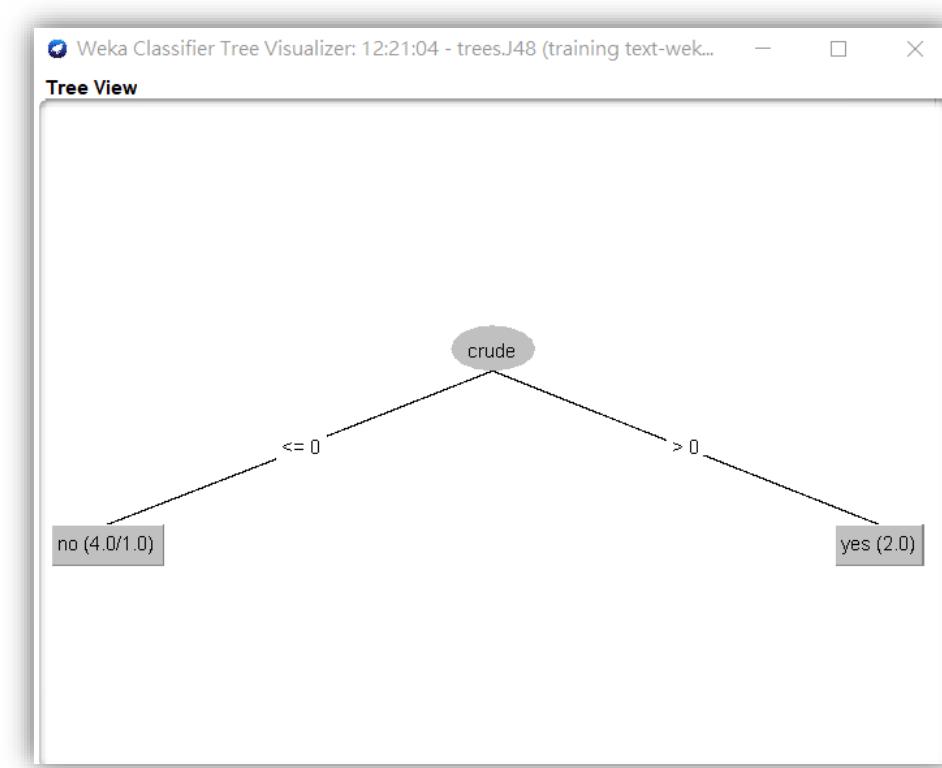


## Lesson 2.4: 文本分類

12. 可以看到決策樹只測試了一個單詞"crude"。  
如果"crude"沒有出現，那麼歸為"no"類文檔(關於食物)。  
如果“crude”出現了，說明是“yes”類文檔(關於地下開採的油)。

### ▼執行結果中的規則

```
==== Classifier model (full training set) ====  
  
J48 pruned tree  
-----  
  
crude <= 0: no (4.0/1.0)  
crude > 0: yes (2.0)  
  
Number of Leaves : 2  
  
Size of the tree : 3
```



## Lesson 2.4: 文本分類

- ❖ 使用提供的測試集
  - 設定 “*Output predictions*”
- ❖ 出現問題：Problem evaluating classifier
- ❖ 套用**StringToWordVector** 到測試檔案上?
  - 仍然出現問題：“Problem evaluating classifier”
- ❖ 解決辦法：**FilteredClassifier**
  - *StringToWordVector* 從訓練集創造屬性
  - *FilteredClassifier* 對測試集使用相同的屬性
- ❖ 結果：
  - document 1 得到“yes”; Documents 2, 3, 4 得到“no”
  - (雖然 document 3 應為“yes”)

### 一些測試資料

Document text	Classification
Oil platforms extract crude oil	Unknown
Canola oil is supposed to be healthy	Unknown
Iraq has significant oil reserves	Unknown
There are different types of cookingoil	Unknown

## Lesson 2.4: 文本分類

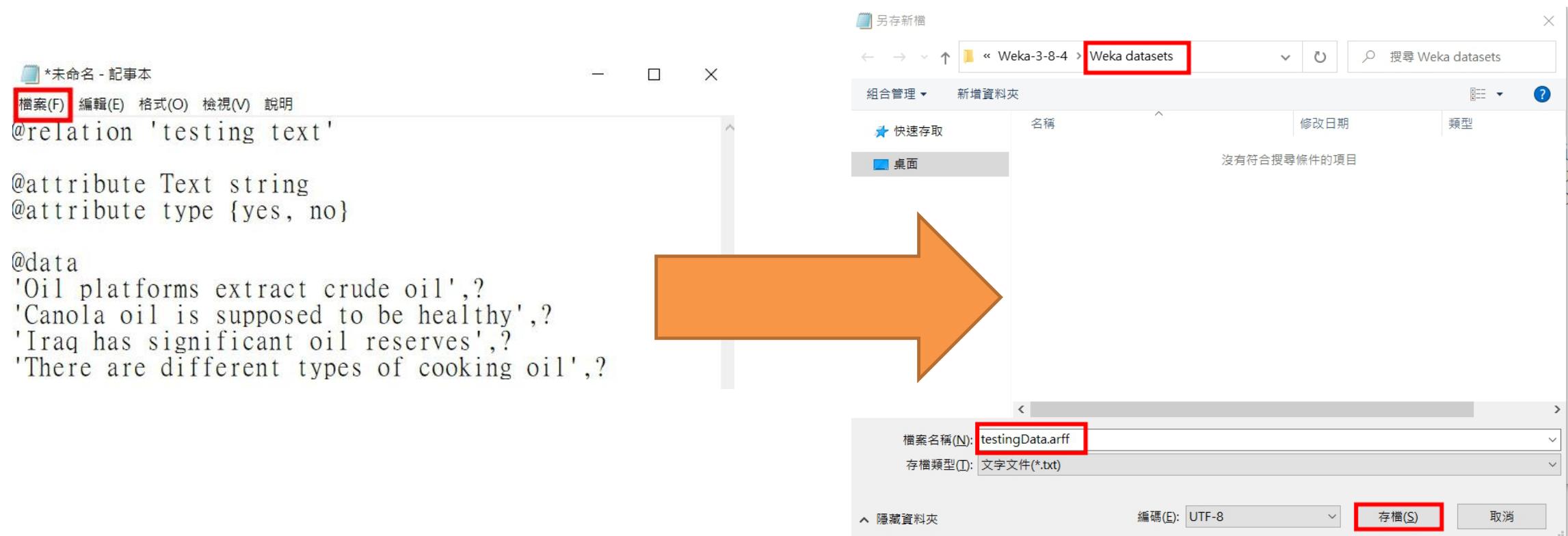
接下來我們製作一份測試資料。

1. 開啟一個新的記事本並將貼上下列測試文本的內容。

```
@relation 'testing text'  
  
@attribute Text string  
@attribute type {yes, no}  
  
@data  
'Oil platforms extract crude oil',?  
'Canola oil is supposed to be healthy',?  
'Iraq has significant oil reserves',?  
'There are different types of cooking oil',?
```

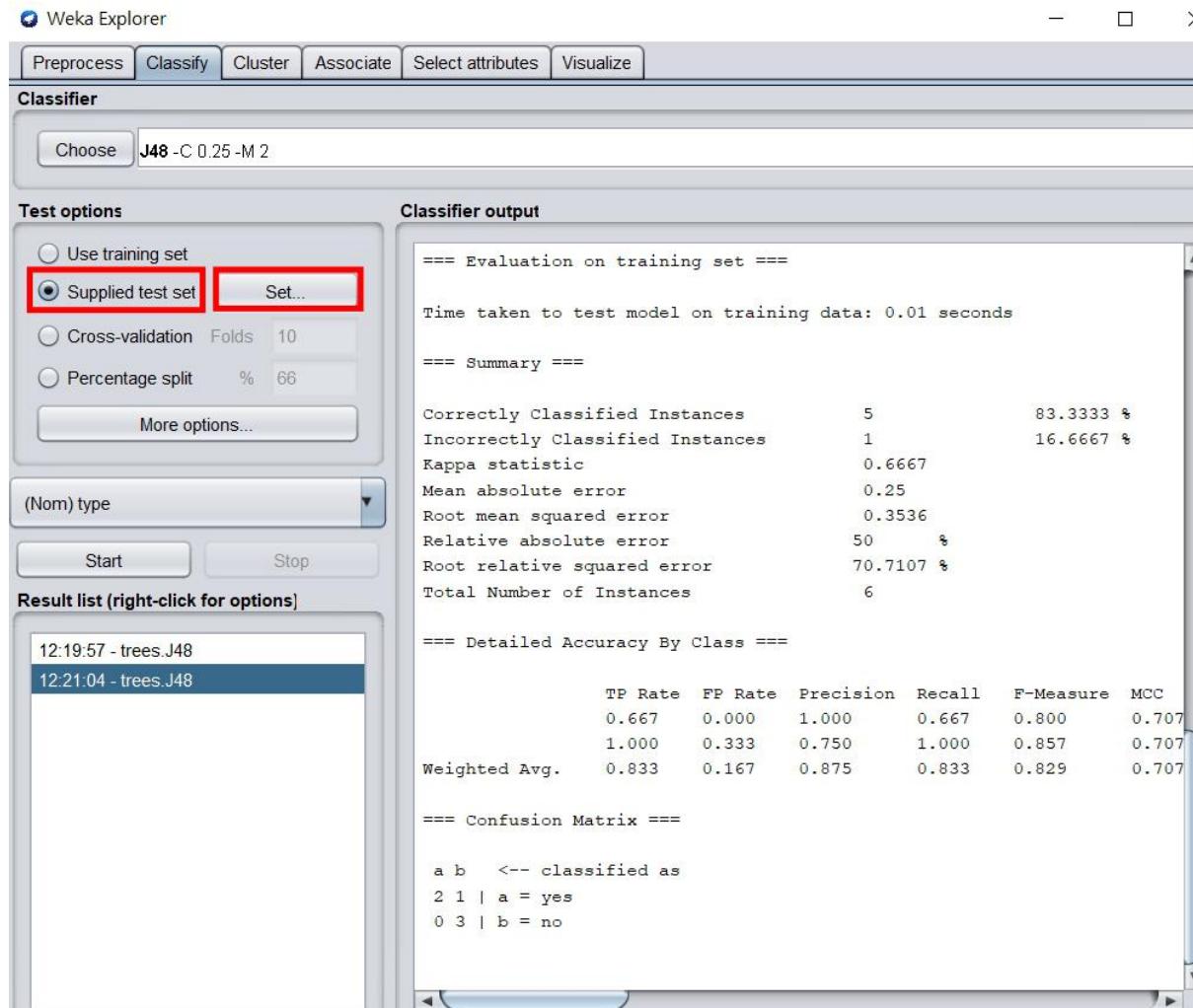
## Lesson 2.4: 文本分類

2. 左鍵單擊工具列中的檔案(F)按鈕，江文本命名為testingData.arff，並在選單中左鍵單擊儲存檔案(S)。



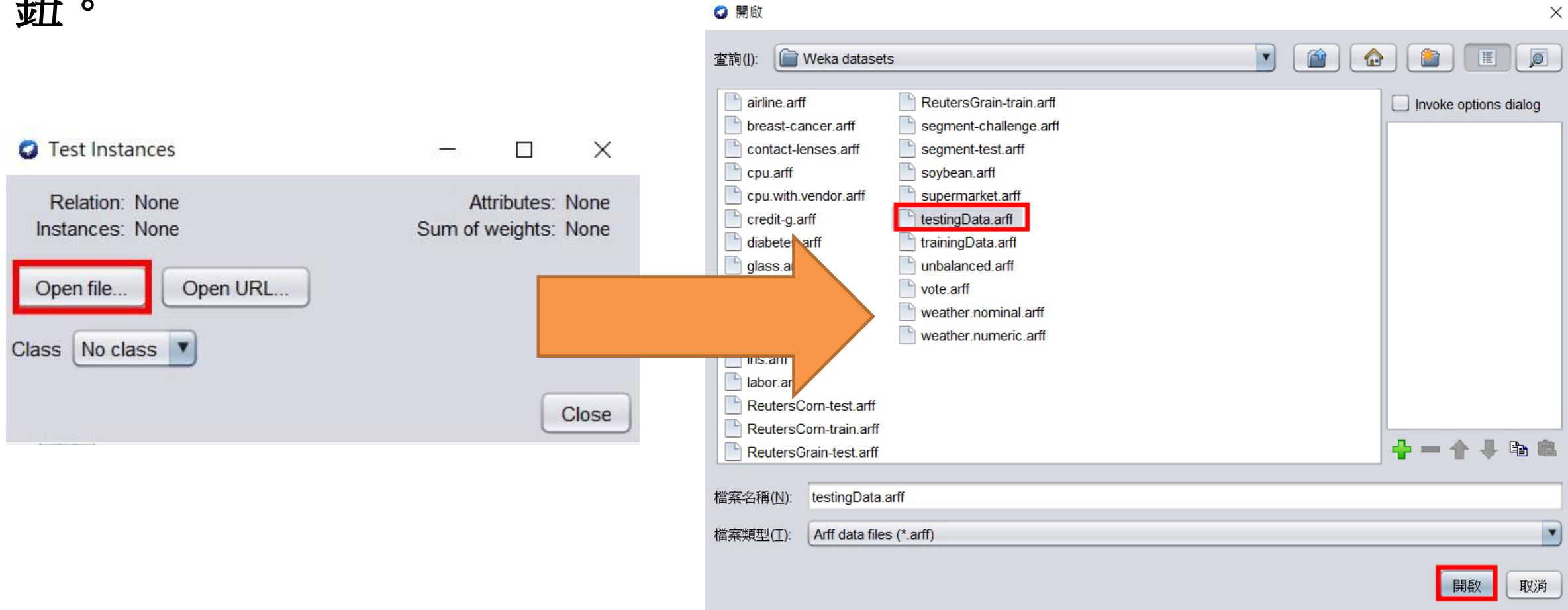
## Lesson 2.4: 文本分類

3.回到Explorer的Classify面板左鍵點選Supplied test set前方圓圈，並以左鍵單擊後方的Set...按鈕。



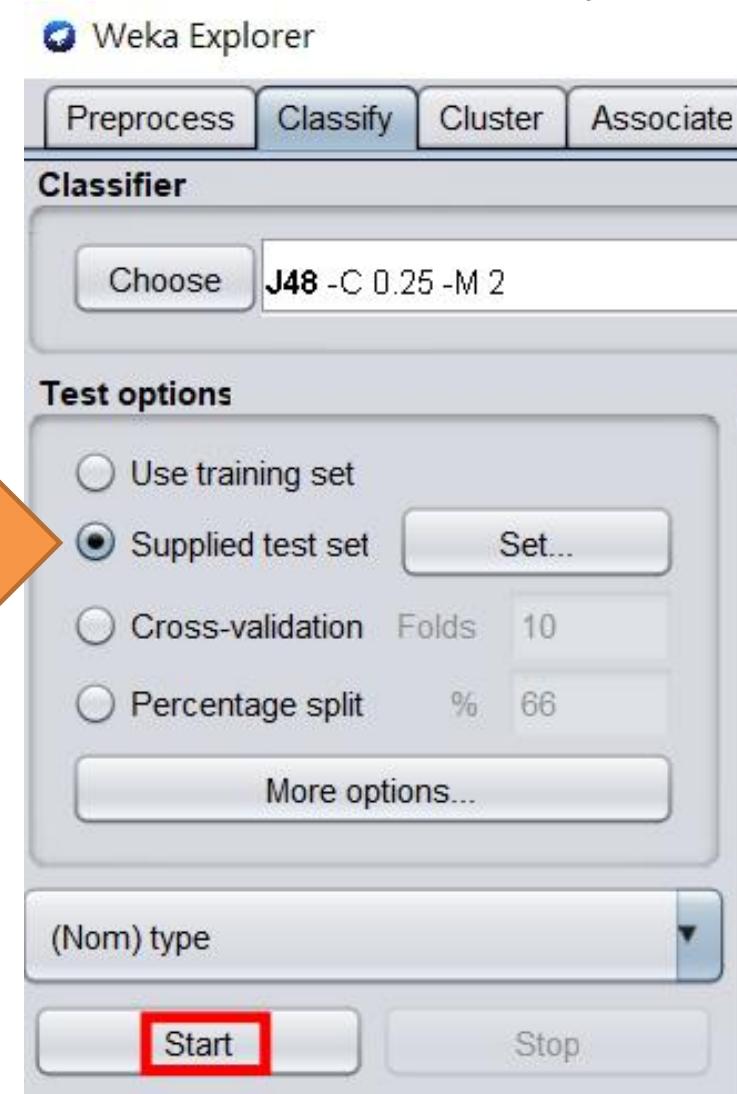
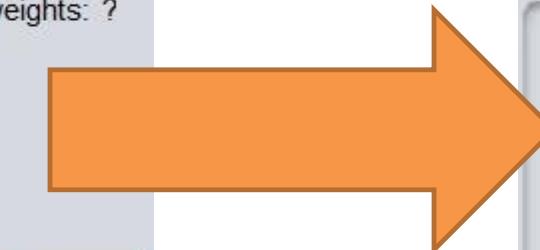
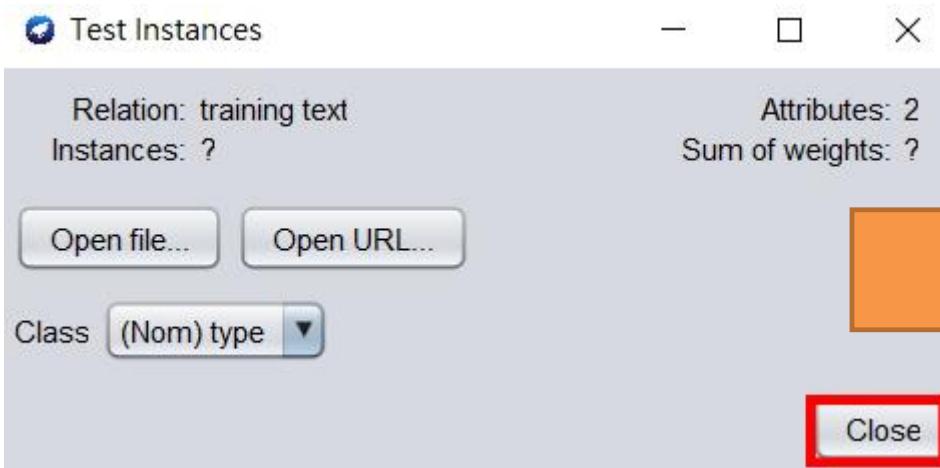
## Lesson 2.4: 文本分類

4. 在出現的**Test Instances**視窗中左鍵單擊**Open file...**按鈕，並在開啟檔案的視窗中選擇剛剛創建的**testingData.arff**檔案，然後左鍵單擊下方開啟按鈕。



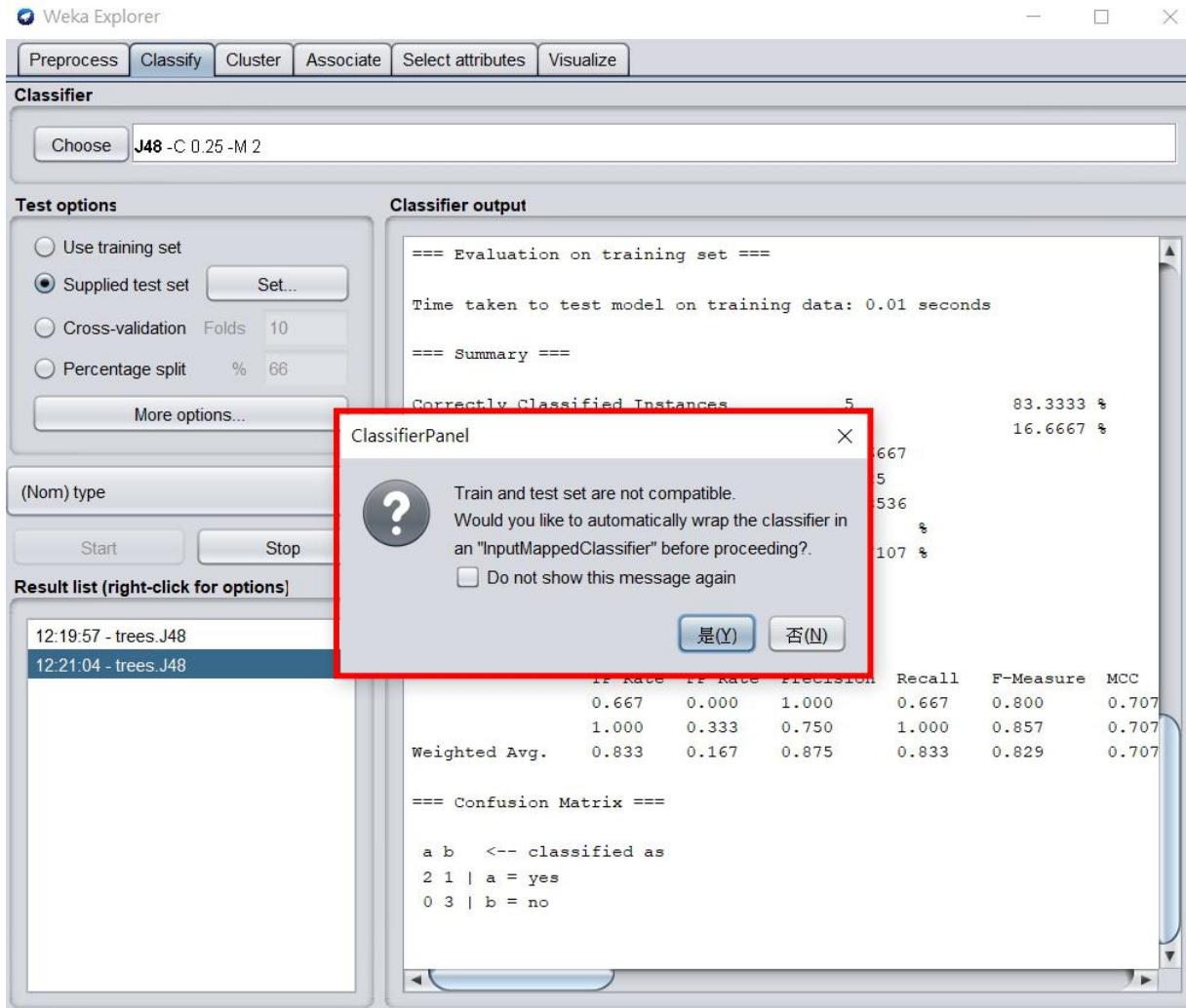
## Lesson 2.4: 文本分類

5. 回到Test Instances視窗，左鍵單擊Close按鈕回到Classify面板，並按下Start按鈕運行。



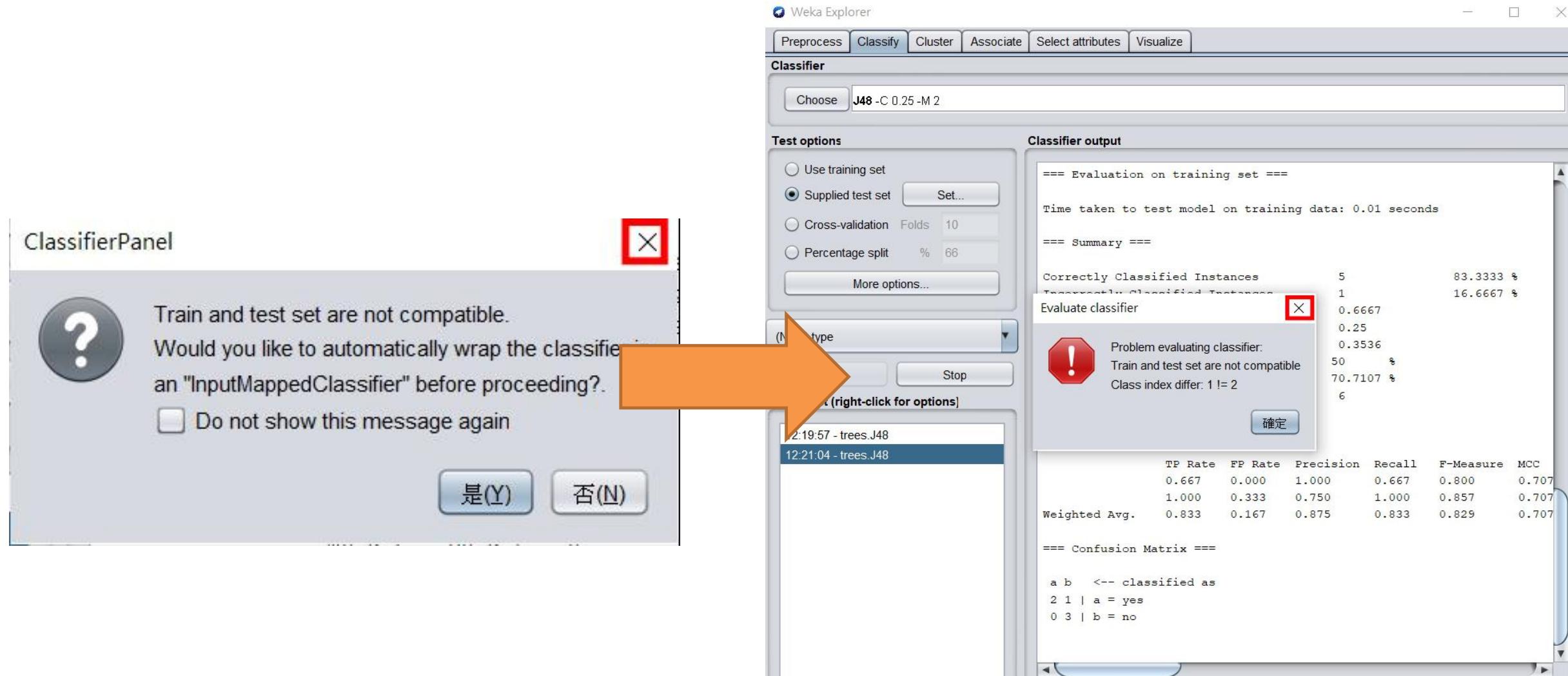
## Lesson 2.4: 文本分類

▼執行結果：出現一個如何評估分類器問題。因為測試集是一個包含字符串屬性的ARFF文件，但是訓練集是一個包含單詞屬性的ARFF文件。



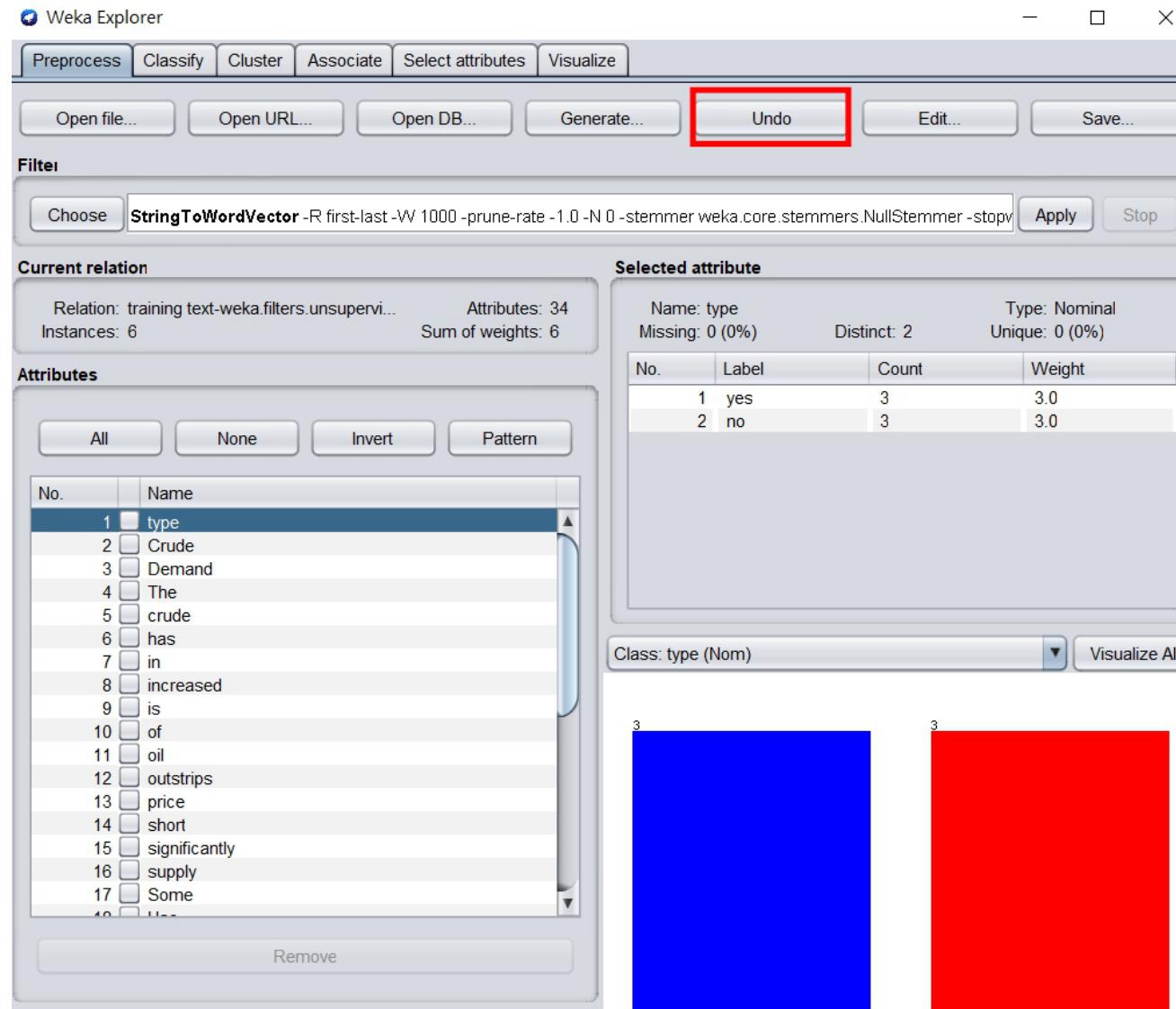
## Lesson 2.4: 文本分類

### 6. 左鍵單擊兩則錯誤訊息右上方的關閉按鈕。



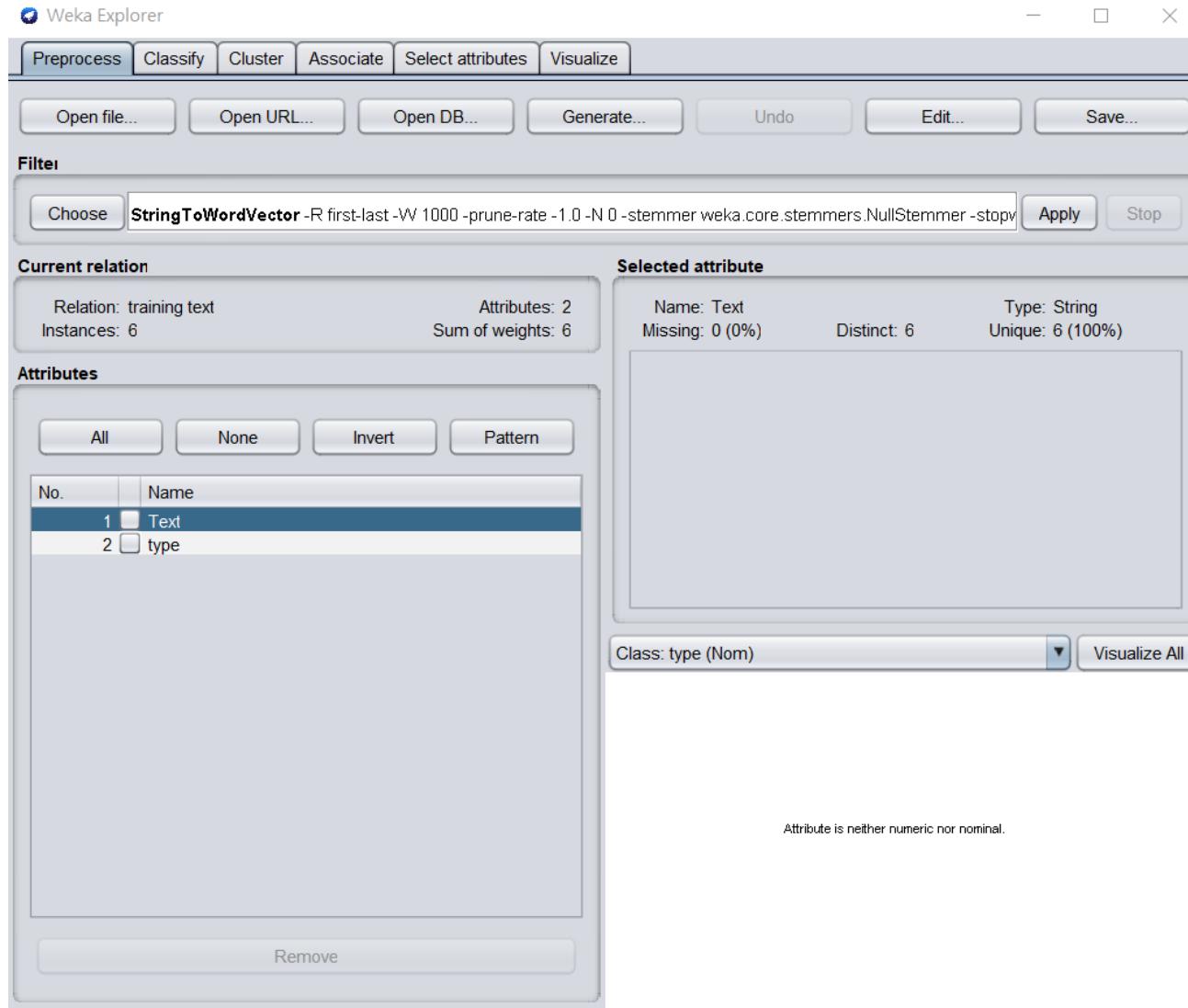
## Lesson 2.4: 文本分類

7. 切換到Preprocess面板，左鍵單擊Undo按鈕撤銷過濾器的影響。



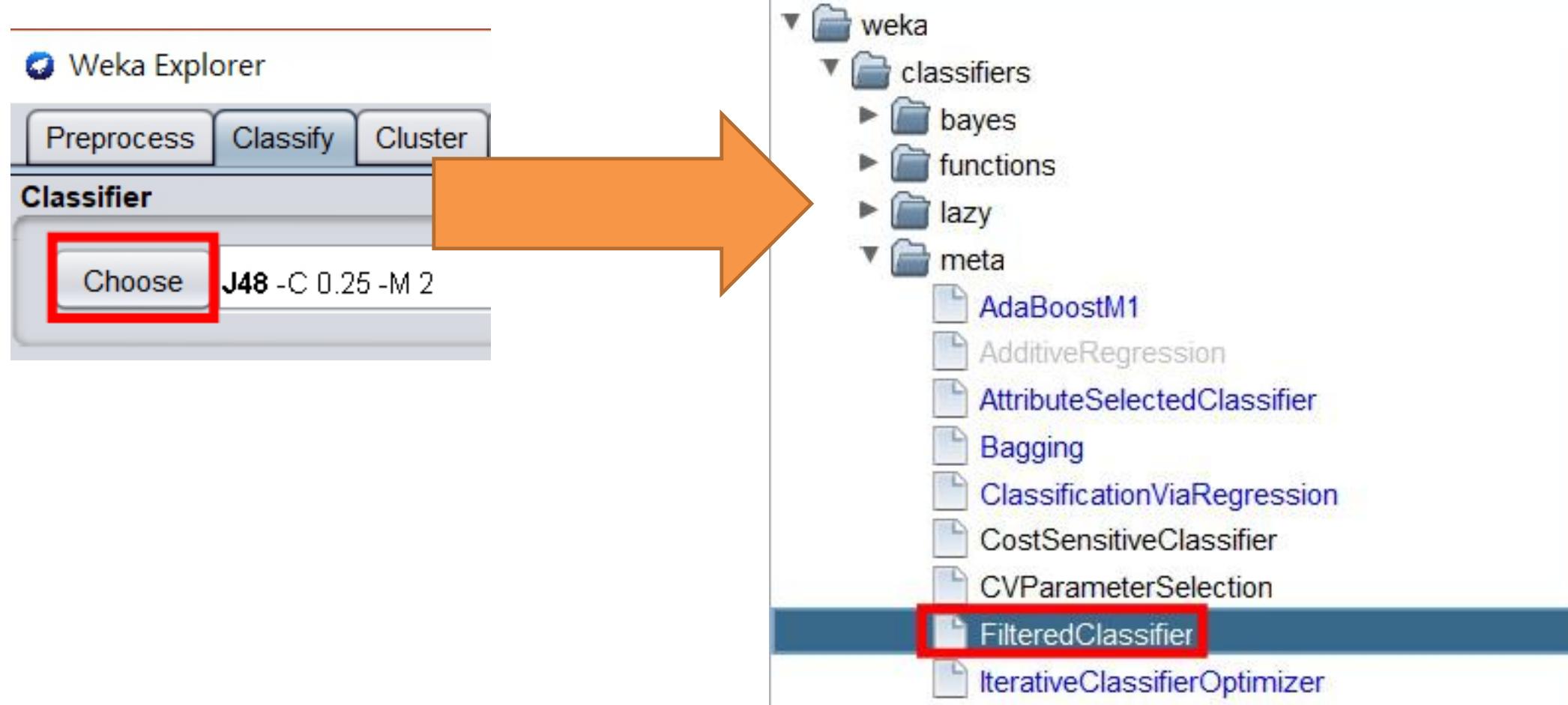
## Lesson 2.4: 文本分類

▼執行結果：回到一開始只有兩個屬性的樣子。



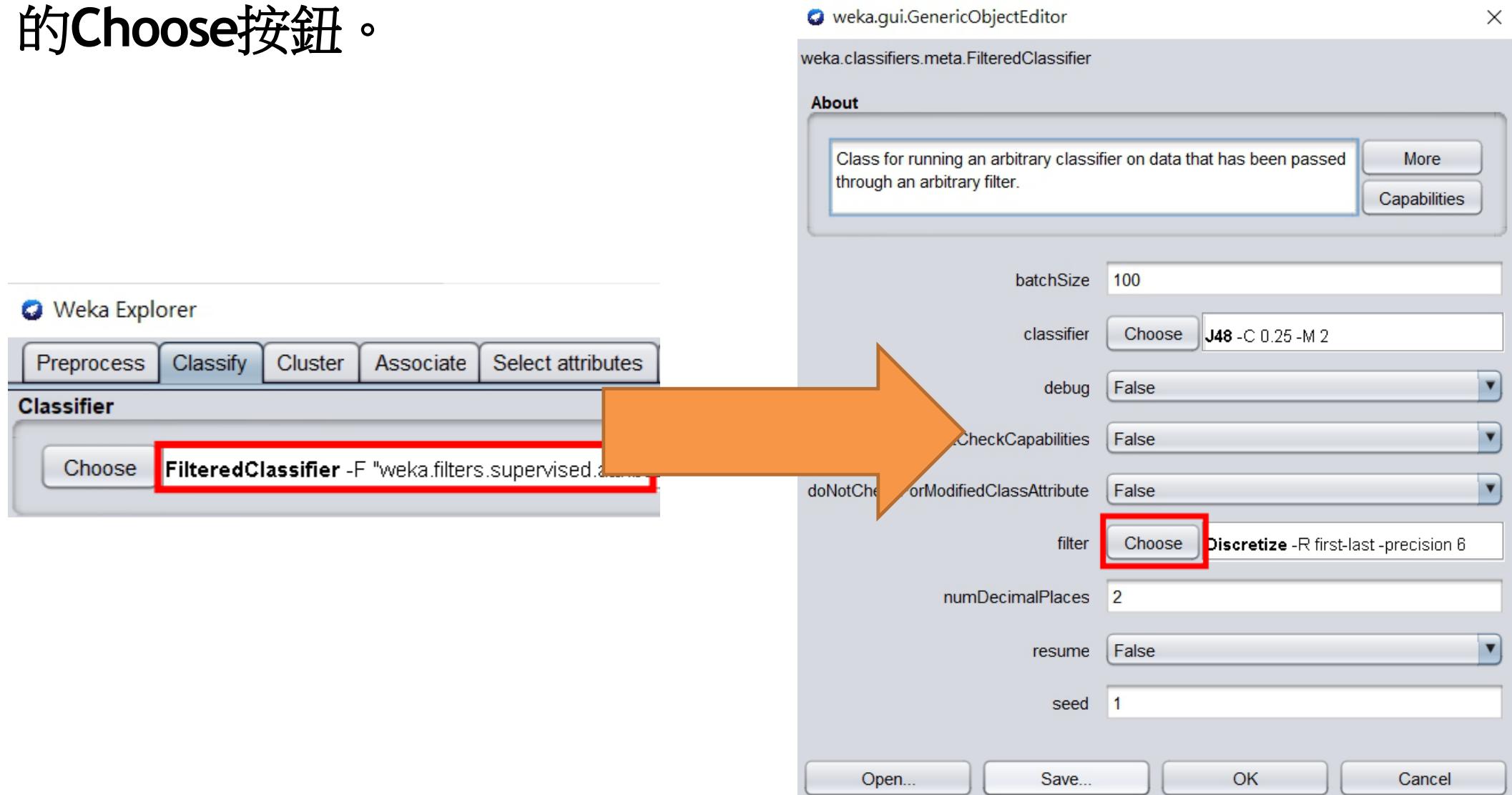
## Lesson 2.4: 文本分類

8. 切換到Classify介面左鍵單擊Choose鈕，並在出現的選單中以左鍵單擊meta資料夾下的FilteredClassifier分類器。



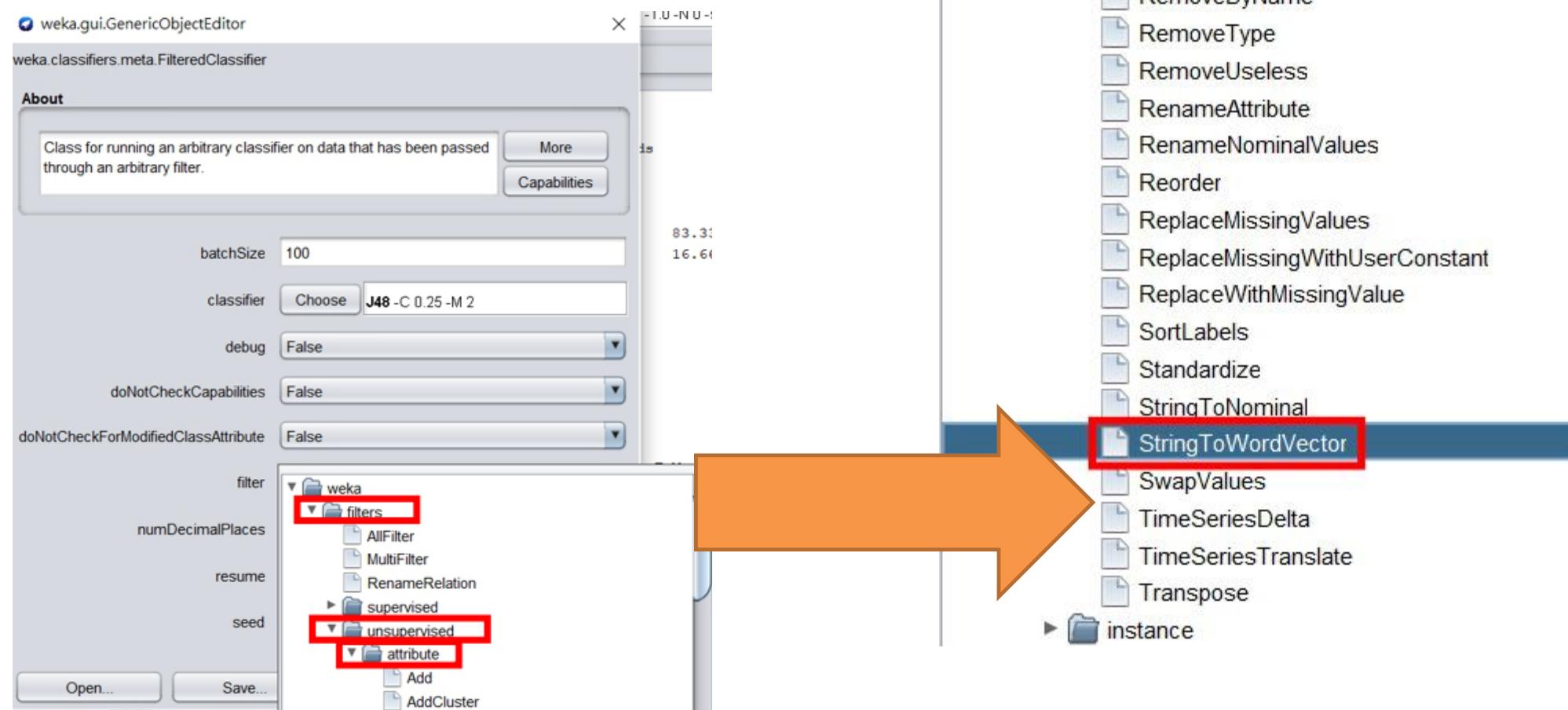
## Lesson 2.4: 文本分類

9. 左鍵單擊左圖紅色方框處開啟右圖配置視窗，左鍵單擊filter參數右方的Choose按鈕。



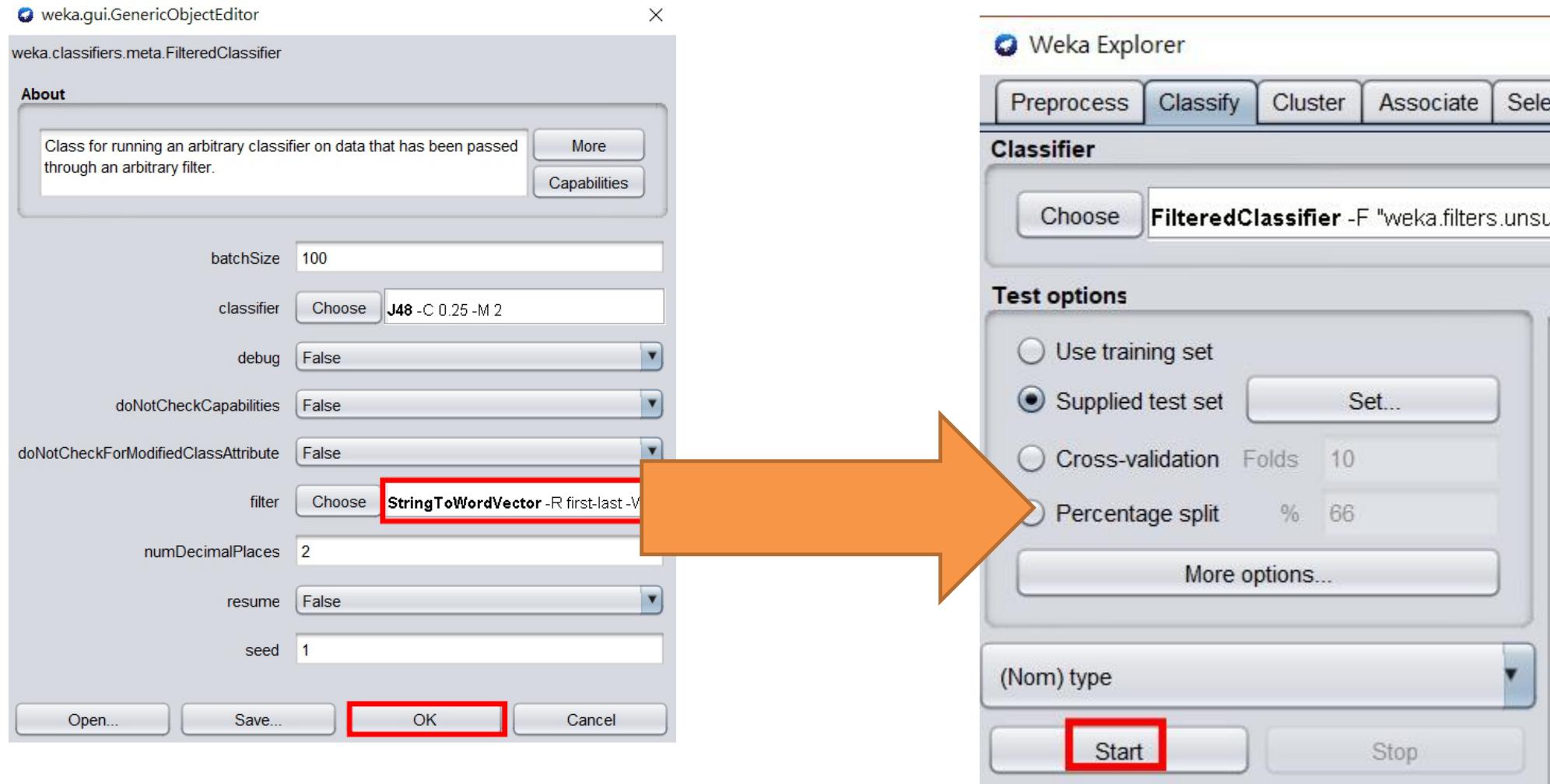
## Lesson 2.4: 文本分類

### 10. 左鍵單擊filters/unsupervised/attribute資料夾下的StringToWordVector。



## Lesson 2.4: 文本分類

11. 確認選擇好**StringToWordVector**之後，左鍵單擊下方OK按鈕回到**Classify**面板，再以左鍵單擊**Start**按鈕運行分類器。



# Lesson 2.4: 文本分類

## ▼運行結果

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

**Classifier**

Choose **FilteredClassifier** -F "weka.filters.unsupervised.attribute.StringToWordVector -R first-last -W 1000 -prune-rate -1.0 -N 0 -stemmer weka.core.stemmars.NullStemmer -stopword-patterns null"

**Test options**

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

(Nom) type

Start Stop

**Result list (right-click for options)**

20:24:03 - meta.FilteredClassifier  
20:24:31 - meta.FilteredClassifier

**Classifier output**

```
Size of the tree :      3

Time taken to build model: 0.01 seconds

==== Evaluation on test set ===

Time taken to test model on supplied test set: 0 seconds

==== Summary ===

Total Number of Instances          0
Ignored Class Unknown Instances   4

==== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall   F-Measure  MCC    ROC Area  PRC Area  Cl
          ?        ?        ?         ?        ?        ?        ?        ?        ?        yes
          ?        ?        ?         ?        ?        ?        ?        ?        ?        no
Weighted Avg.    ?        ?        ?         ?        ?        ?        ?        ?        ?

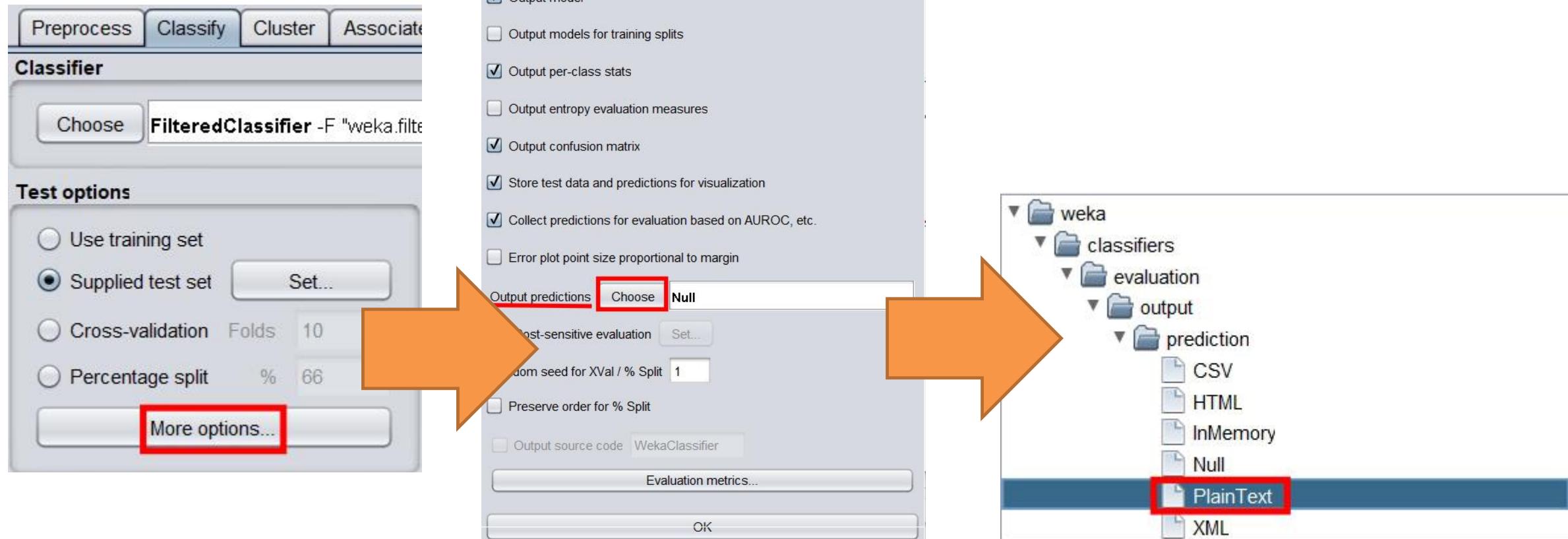
==== Confusion Matrix ===

a b    <-- classified as
0 0 | a = yes
0 0 | b = no
```

## Lesson 2.4: 文本分類

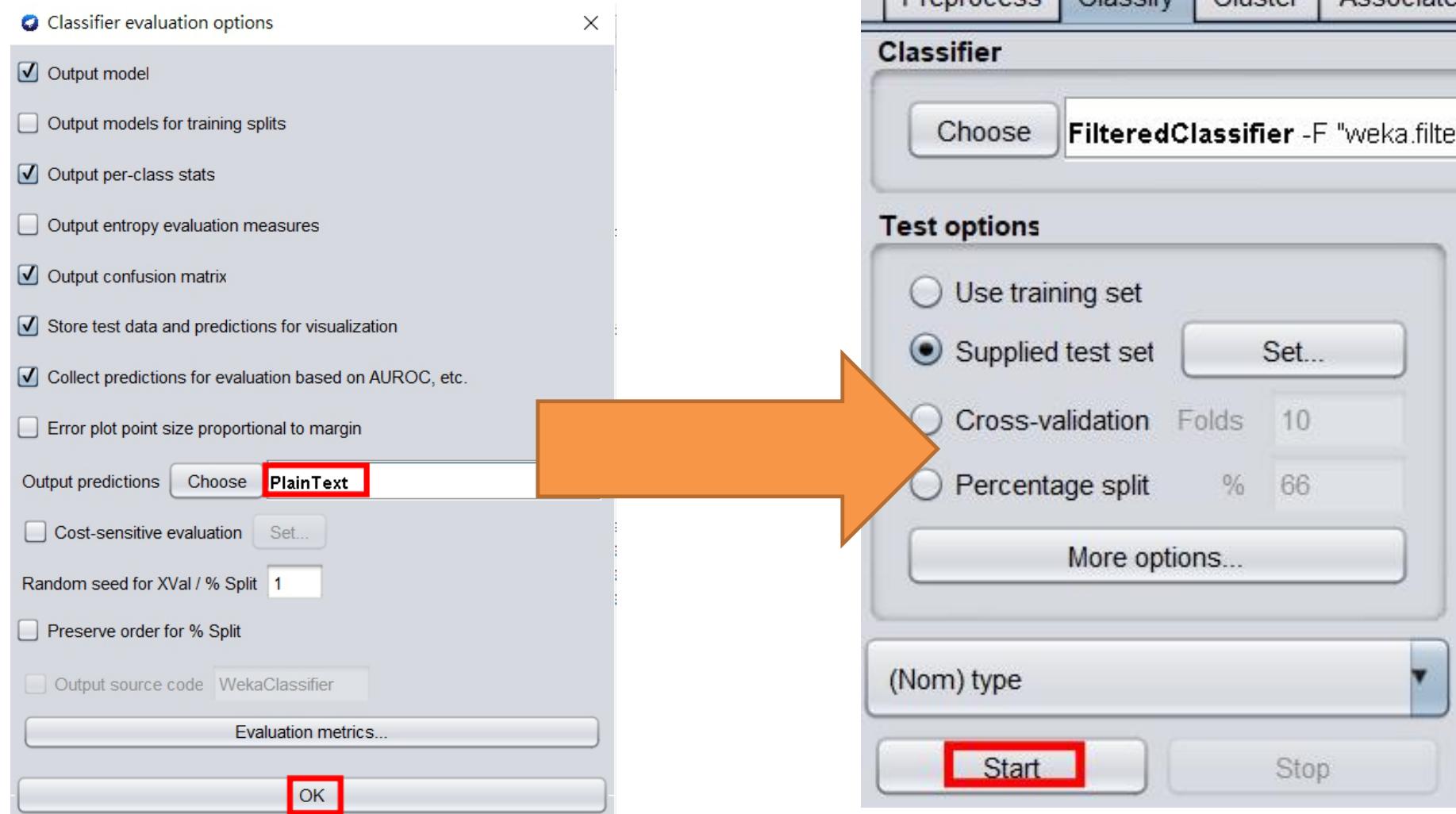
如何導出預測結果？

12. 左鍵單擊Classify面板中Test option區域內的More options...按鈕，再以左鍵單擊Output predictions右方的Choose按鈕，於出現的選單中左鍵單擊PlainText。



## Lesson 2.4: 文本分類

13.確定選擇好PlainText後，左鍵單擊OK按鈕回到Classify面板，左鍵單擊Start按鈕。



## Lesson 2.4: 文本分類

▼運行結果：預測中有一個“yes”和3個“no”。

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose FilteredClassifier -F "weka.filters.unsupervised.attribute.StringToWordVector -R first-last -W 1000 -prune-rate 1.0 -N 0 -stemmer weka.core.stemmers.NullStemmer -stop

Test options

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

(Nom) type

Start Stop

Result list (right-click for options)

20:24:03 - meta.FilteredClassifier  
20:24:31 - meta.FilteredClassifier

Classifier output

```
==== Predictions on test set ====
inst#    actual   predicted error prediction
      1       1:?
      2       1:?
      3       1:?
      4       1:?

==== Evaluation on test set ====
Time taken to test model on supplied test set: 0.04 seconds

==== Summary ====
Total Number of Instances          0
Ignored Class Unknown Instances    4

==== Detailed Accuracy By Class ====
TP Rate  FP Rate  Precision  Recall  F-Measure  MCC  ROC Area  PRC Area  Cl
?        ?        ?         ?        ?        ?        ?        ?        ?        yes
?        ?        ?         ?        ?        ?        ?        ?        ?        no
Weighted Avg.  ?        ?        ?         ?        ?        ?        ?        ?        ?

==== Confusion Matrix ====
a b    <-- classified as
0 0 | a = yes
```

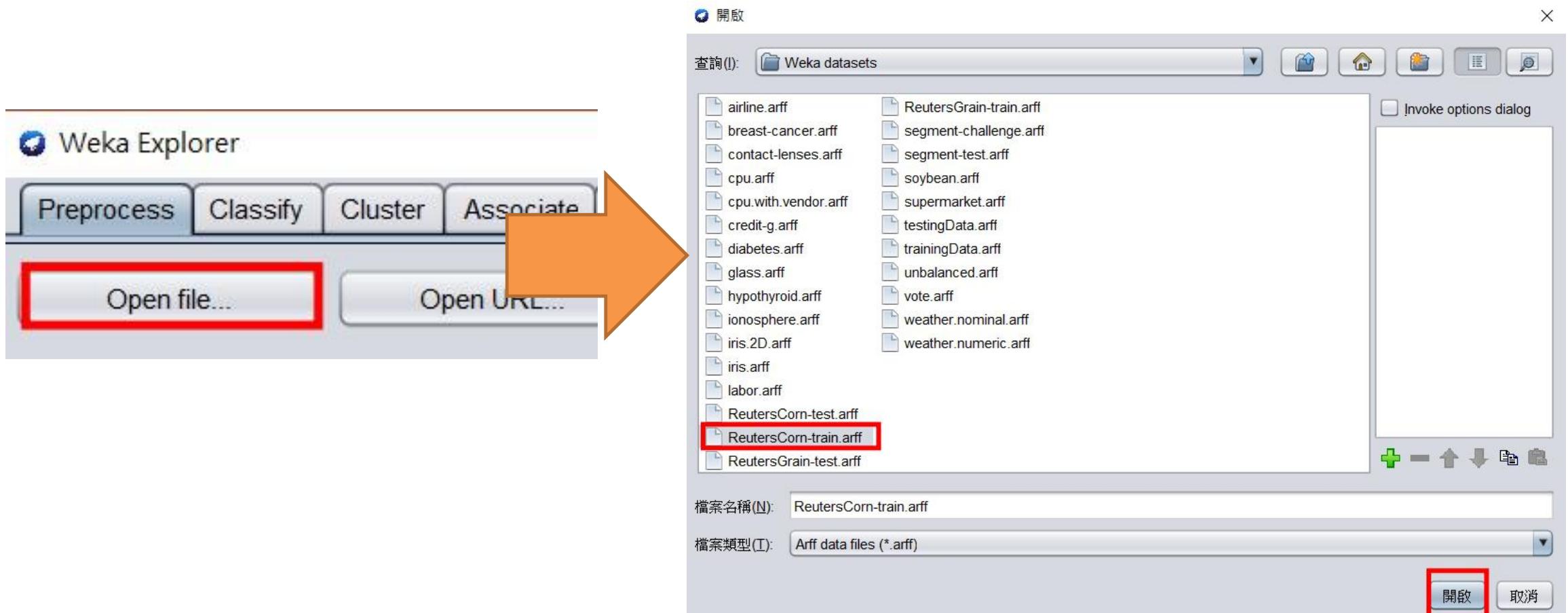
## Lesson 2.4: 文本分類

- ❖ 查看資料集: **ReutersCorn-train.arff**
  - 它很龐大**big**: 1554 個實例, 2 種屬性
- ❖ 套用 **StringToWordVector**
  - 它變的很巨大: 1554個實例, 2234種屬性(!)
- ❖ 測試集: **ReutersCorn-test.arff**
- ❖ 使用**StringToWordVector** 以及 **J48** 的**FilteredClassifier**
  - (運行需要稍等一下)
- ❖ 97% 分類器準確率
- ❖ 查看模型
- ❖ 查看混淆矩陣(**confusion matrix**):
  - 分類器在**24**穀物相關(*corn-related*)的文本上的準確率:  $15/24 = 62\%$
  - 在剩餘的**580**個文本上的準確率:  $573/580 = 99\%$
- ❖ 優化整體的分類器準確率是對的嗎？

## Lesson 2.4: 文本分類

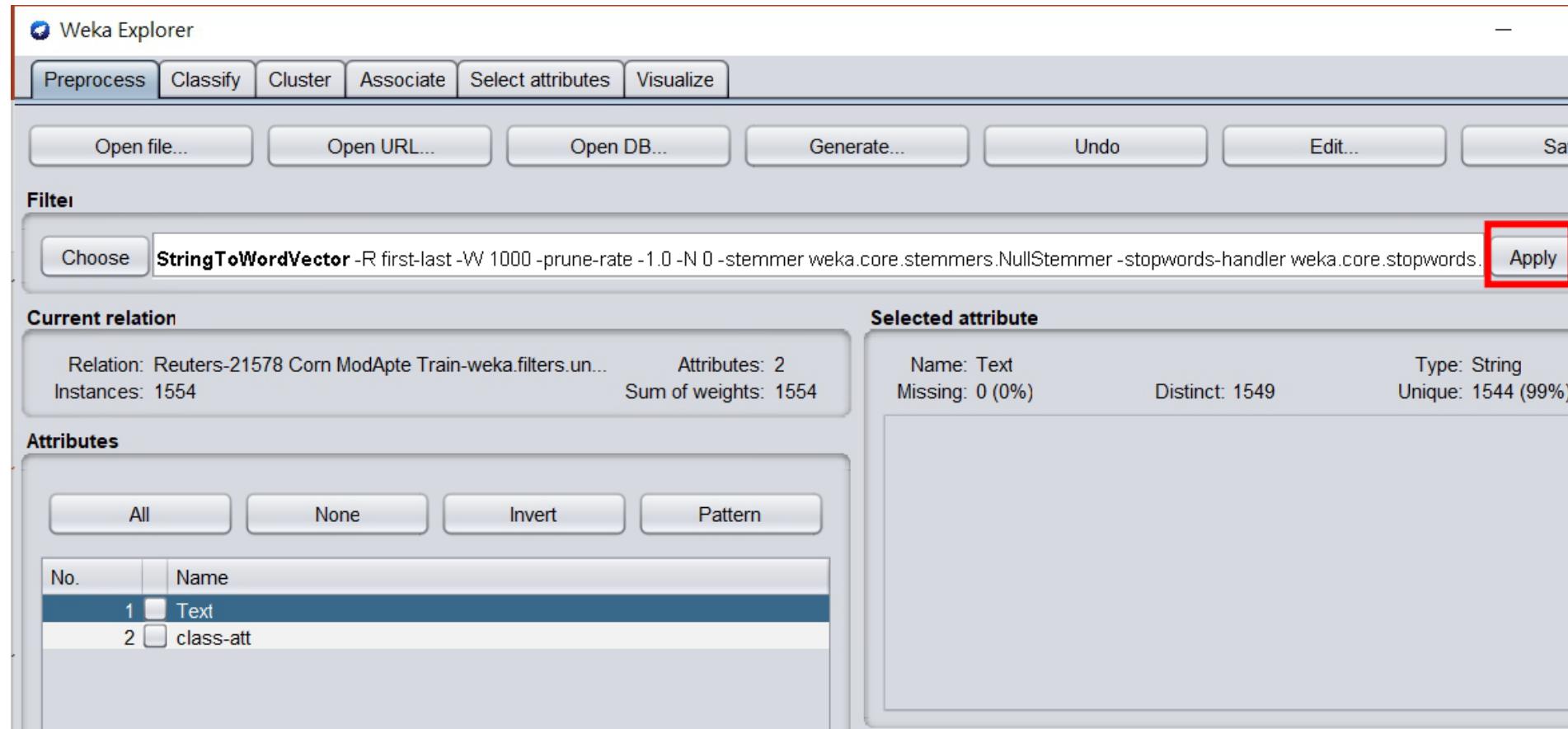
接下來我們試試另一個資料集。

1.回到Preprocess面板，左鍵單擊Open file...按鈕開啟右側視窗，再以左鍵單擊ReutersCorn-train.arff檔案，然後按下下方開啟按鈕。



## Lesson 2.4: 文本分類

### 2. 套用StringToWordVector過濾器。



# Lesson 2.4: 文本分類

## ▼運行結果

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

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

**Filter**

Choose `StringToWordVector -R first-last -V 1000 -prune-rate -1.0 -N 0 -stemmer weka.core.stemmers.NullStemmer -stopwords-handler weka.core.stopwords` Apply Stop

**Current relation**

Relation: Reuters-21578 Corn ModApte Train-weka.filters.un... Attributes: 2234  
Instances: 1554 Sum of weights: 1554

**Attributes**

All None Invert Pattern

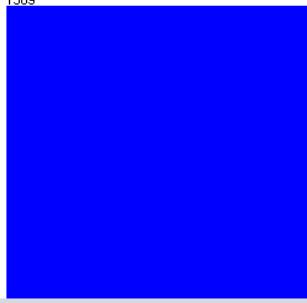
No.	Name
1	class-att
2	&#3
3	&lt
4	-
5	--
6	0
7	0/92
8	00
9	000
10	1
11	10
12	100
13	11
14	12
15	13
16	14
17	15
18	150

**Selected attribute**

Name: class-att Type: Nominal  
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count	Weight
1	0	1509	1509.0
2	1	45	45.0

Class: class-att (Nom) Visualize All

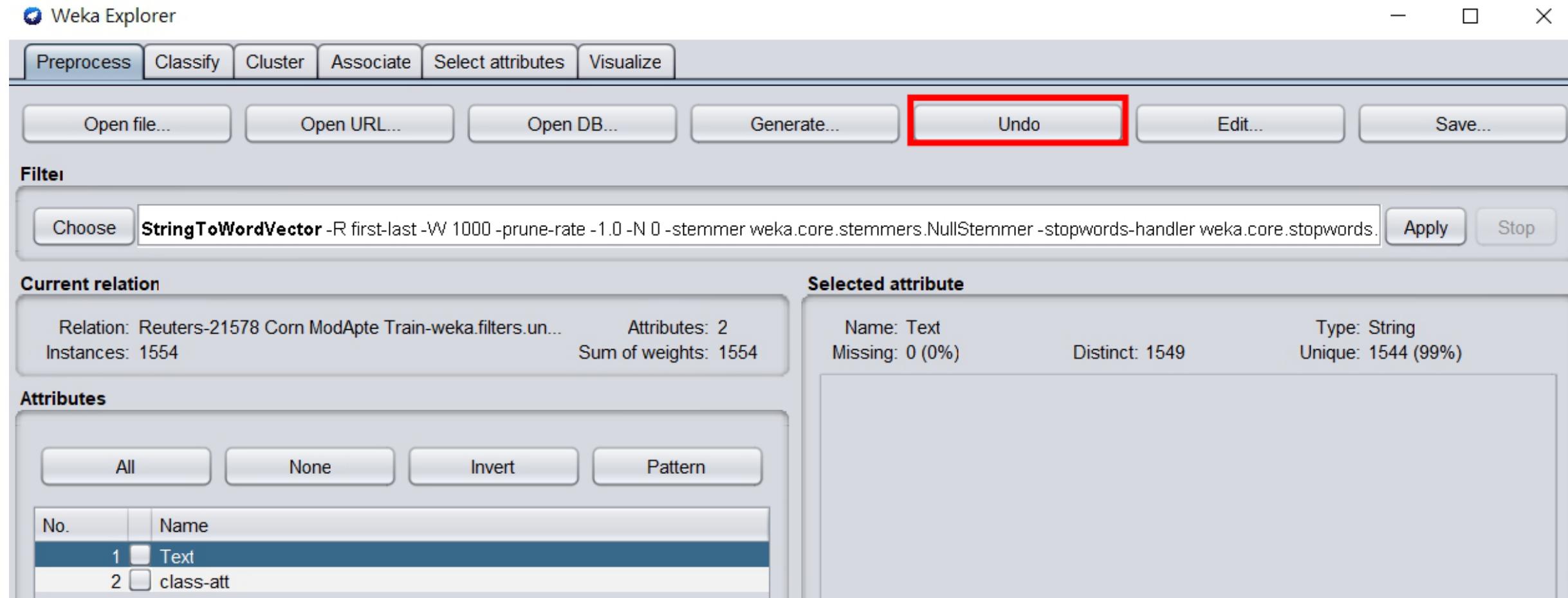


1509

45

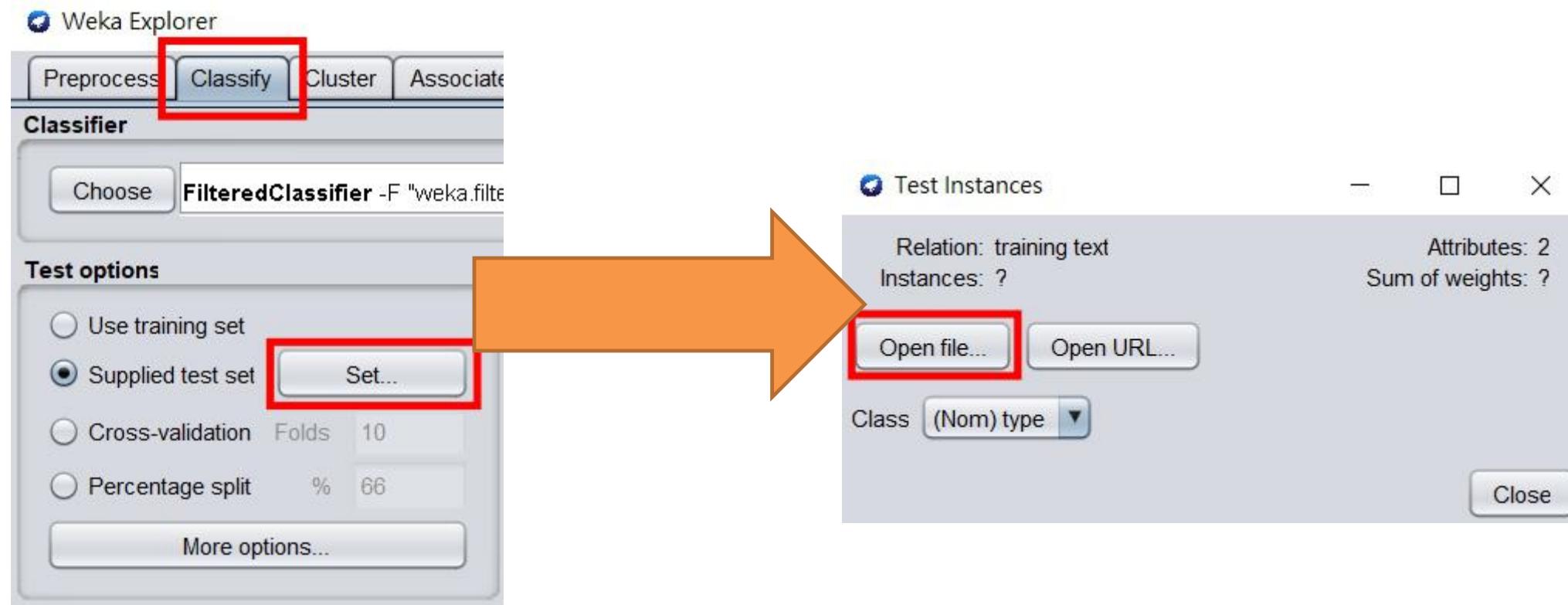
## Lesson 2.4: 文本分類

3. 左鍵單擊Undo按鈕撤銷過濾器影響。



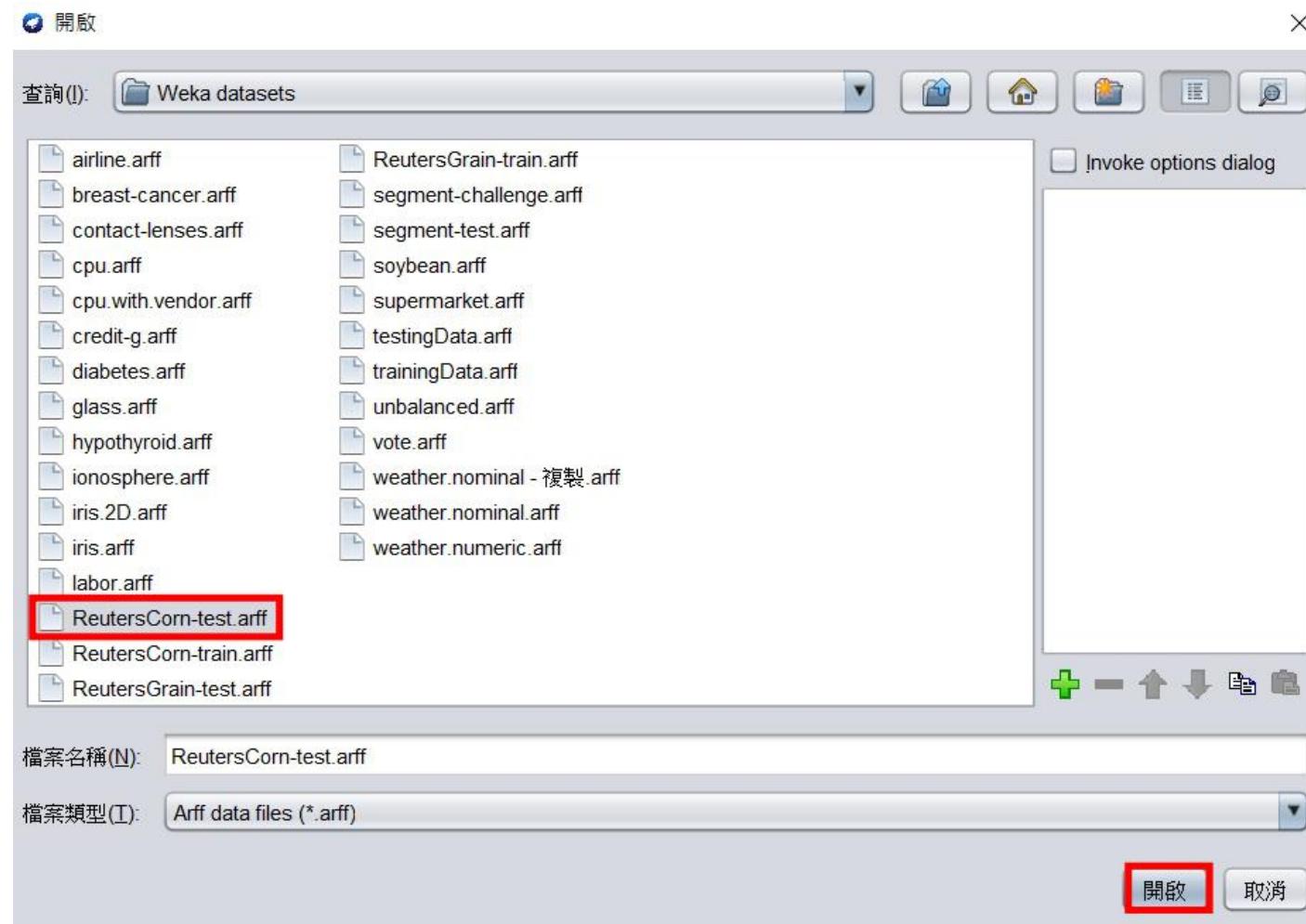
## Lesson 2.4: 文本分類

4. 切換到Classify面板，左鍵單擊Set...按鈕並在開啟的視窗左鍵單擊Open file...按鈕。



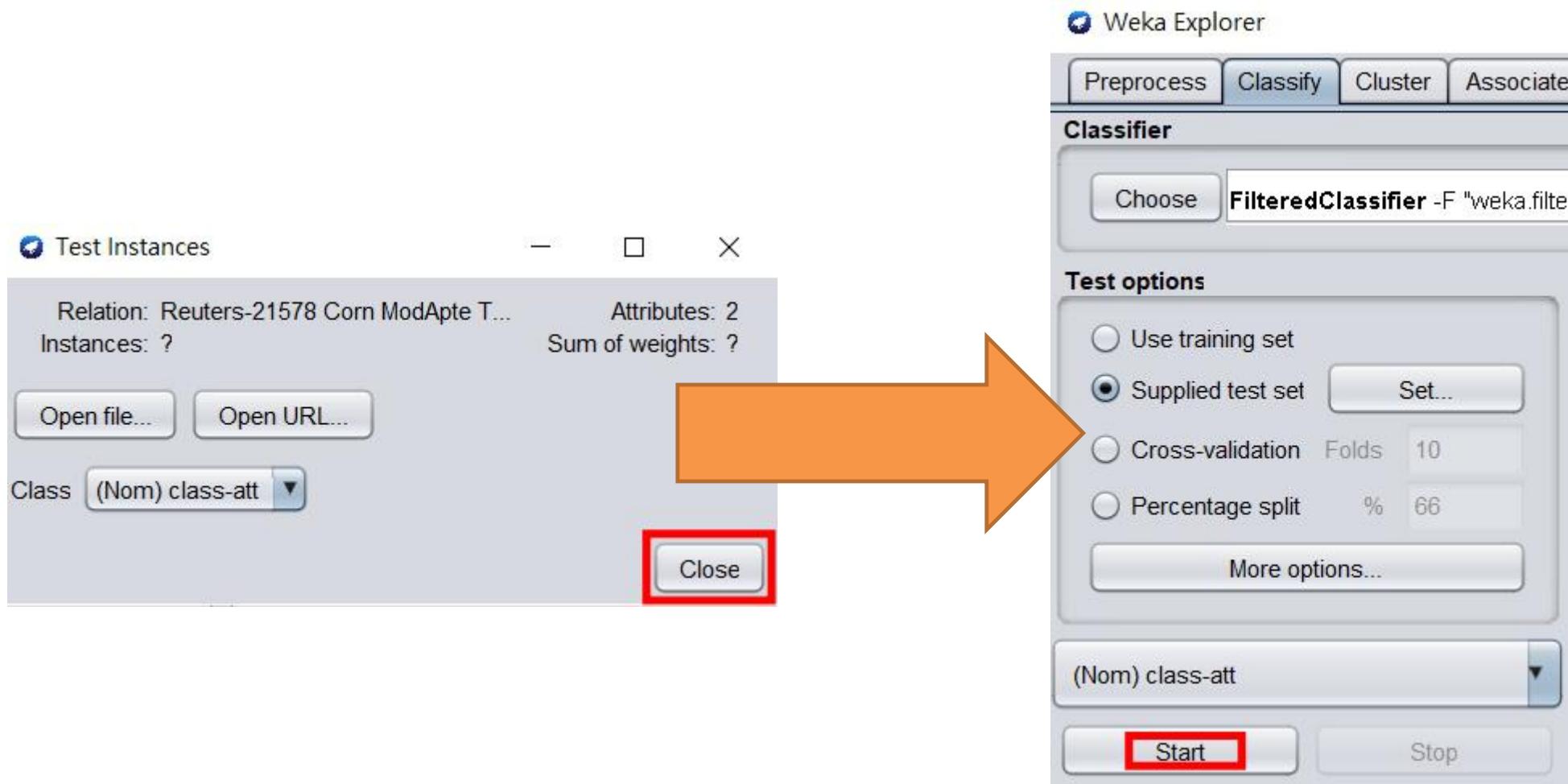
## Lesson 2.4: 文本分類

5. 左鍵單擊ReutersCorn-test.arff檔案，並按下下方開啟按鈕。



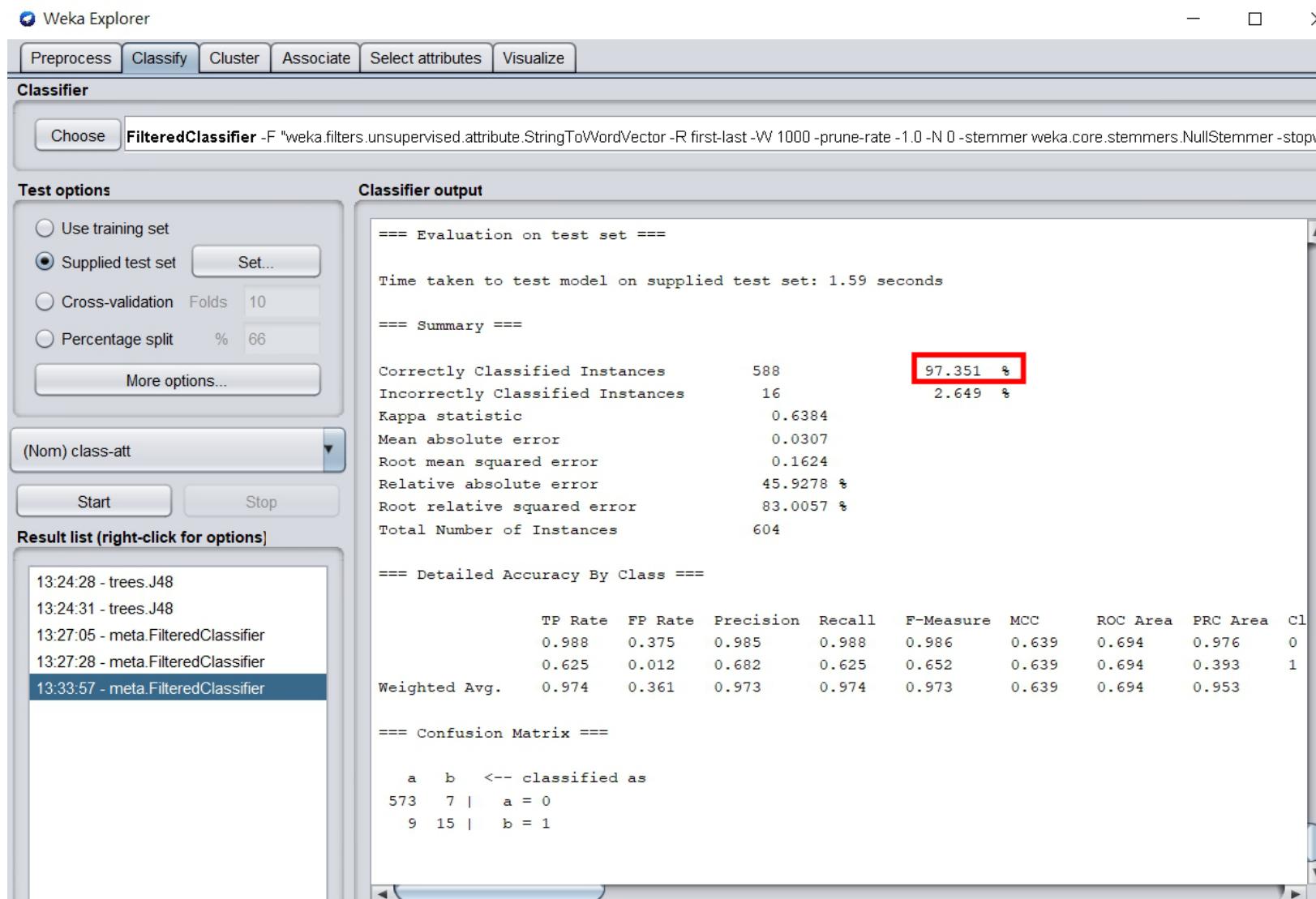
## Lesson 2.4: 文本分類

6. 回到Test Instances視窗，左鍵單擊Close按鈕回到Classify面板，在以左鍵單擊Start按鈕運行分類器。



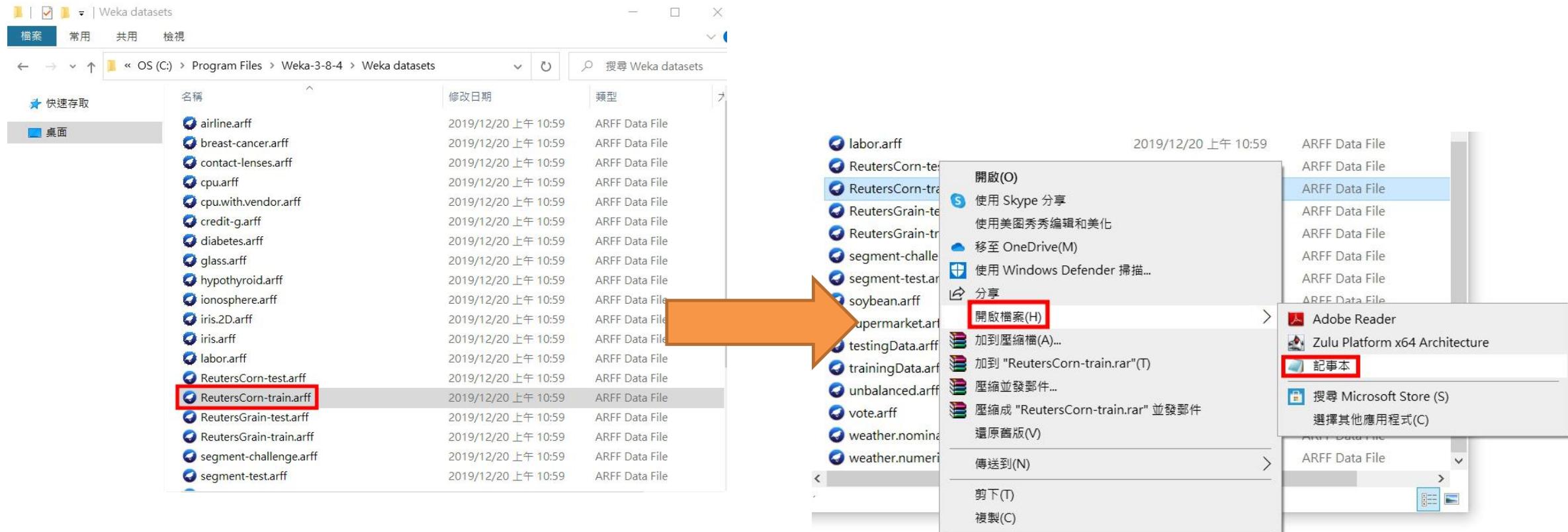
## Lesson 2.4: 文本分類

▼運行結果：得到97.351%準確率。



## Lesson 2.4: 文本分類

7. 進入自行複製的Weka datasets資料夾，以右鍵單擊ReutersCorn-train.arff檔案。於出現的選單中，將游標移至開啟檔案(H)的選項上方，並在右方出現的選單中選擇記事本。



## Lesson 2.4: 文本分類

▼觀察文本內容。我們用“\n”代表換行符因為Weka會認為把換行當作一個新的實例。

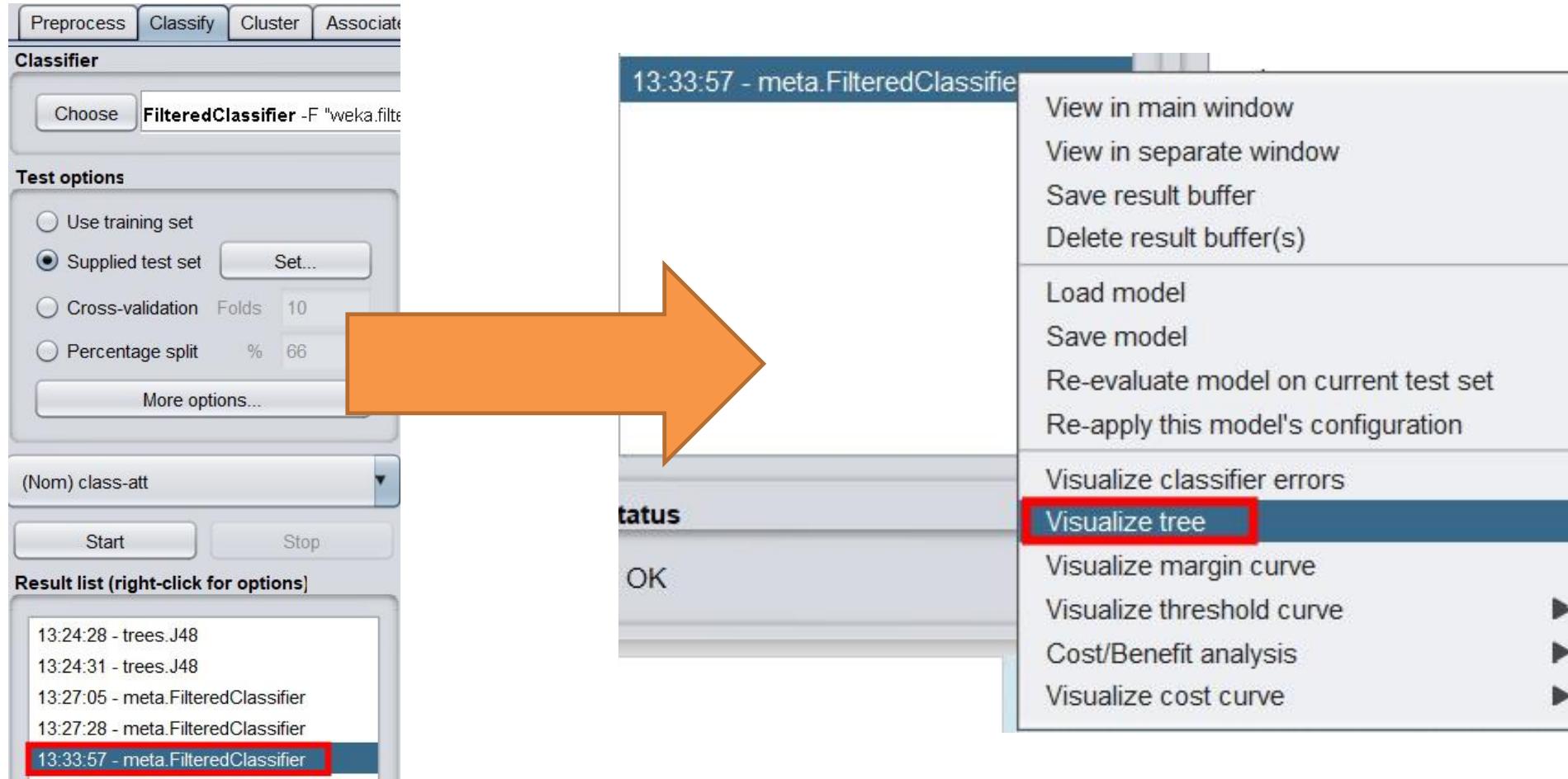
```
ReutersCorn-train.arff - 記事本
檔案(F) 檢查(E) 格式(O) 檢視(V) 說明
@relation 'Reuters-21578 Corn ModApte Train-weka.filters.unsupervised.attribute.NumericToBinary-
weka.filters.unsupervised.instance.RemoveFolds-S0-N5-F1'

@attribute Text string
@attribute class-att {0,1} 有兩個屬性：一個字符串屬性和值為0或1的類屬性。

@data 這是第一個字符串的開頭，它是一個很長的字符串。
'BAHIA COCOA REVIEW Showers continued throughout the week in\nthe Bahia cocoa zone, alleviating the drought since early\nJanuary and improving prospects for the coming temporao,\nalthough normal humidity levels have not been restored,\nComissaria Smith said in its weekly review.\n The dry period means the temporao will be late this year.\n Arrivals for the week ended February 22 were 155,221 bags\nof 60 kilos making a cumulative total for the season of 5.93\nmln against 5.81 at the same stage last year. Again it seems\nthat cocoa delivered earlier on consignment was included in the\narrivals figures.\n Comissaria Smith said there is still some doubt as to how\nmuch old crop cocoa is still available as harvesting has\npractically come to an end. With total Bahia crop estimates\naround 6.4 mln bags and sales standing at almost 6.2 mln there\nare a few hundred thousand bags still in the hands of farmers,\nmiddlemen, exporters and processors.\n There are doubts as to how much of this cocoa would be fit\nfor export as shippers are now experiencing difficulties in\nobtaining +Bahia superior+ certificates.\n In view of the lower quality over recent weeks farmers have\nsold a good part of their cocoa held on consignment.\n Comissaria Smith said spot bean prices rose to 340 to 350\nreais per arroba of 15 kilos.\n Bean shippers were reluctant to offer nearby shipment and\nonly limited sales were booked for March shipment at 1,750 to\n1,780 dls per tonne to ports to be named.\n New crop sales were also light and all to open ports with\nJune/July going at 1,850 and 1,880 dls and at 35 and 45 dls\nunder New York July, Aug/Sept at 1,870, 1,875 and 1,880 dls\nper tonne FOB.\n Routine sales of butter were made. March/April sold at\n4,340, 4,345 and 4,350 dls.\n April/May butter went at 2.27 times New York May,\nJune/July\nat 4,400 and 4,415 dls, Aug/Sept at 4,351 to 4,450 dls and at\n2.27 and 2.28 times New York Sept and Oct/Dec at 4,480 dls\nand\n2.27 times New York Dec, Comissaria Smith said.\n Destinations were the U.S., Covertible currency areas,\nUruguay and open ports.\n Cake sales were registered at 785 to 995 dls for\nMarch/April, 785 dls for May, 753 dls for Aug and 0.39 times\nNew York Dec for Oct/Dec.\n Buyers were the U.S., Argentina, Uruguay and convertible\ncurrency areas.\n Liquor sales were limited with March/April selling at 2,325\nand 2,380 dls, June/July at 2,375 dls and at 1.25 times New\nYork July, Aug/Sept at 2,400 dls and at 1.25 times New York\nSept and Oct/Dec at 1.25 times New York Dec, Comissaria Smith\nsaid.\n Total Bahia sales are currently estimated at 6.13 mln bags\nagainst the 1986/87 crop and 1.06 mln bags against the 1987/88\ncrop.\n Final figures for the period to February 28 are expected to\nbe published by the Brazilian Cocoa Trade Commission after\ncarnival which ends midday on February 27.\nReuter\n#0 → 意味著文檔類別為0。對於這個數據集而言，代表這不是一篇關於穀物的文檔。
'NATIONAL AVERAGE PRICES FOR FARMER-OWNED RESERVE The U.S. Agriculture Department\nreported the farmer-owned reserve national five-day average\nprice through February 25 as follows (Dlrs/Bu-Sorghum Cwt) -\nNatl Loan Release Call\nAvge
Rate-X Level Price Price\nWheat 2.55 2.40 IV 4.65 --\nVI 4.45 --\nCorn 1.35 1.92 IV 3.15 3.15\nNatl Loan Release Call\nAvge Rate-X Level Price Price\nOats 1.24 0.99
1986 Rates.\n\nX -
```

## Lesson 2.4: 文本分類

8. 回到Classify面板，右鍵單擊剛才的執行紀錄，於出現的選單中左鍵單擊Visualize tree選項。

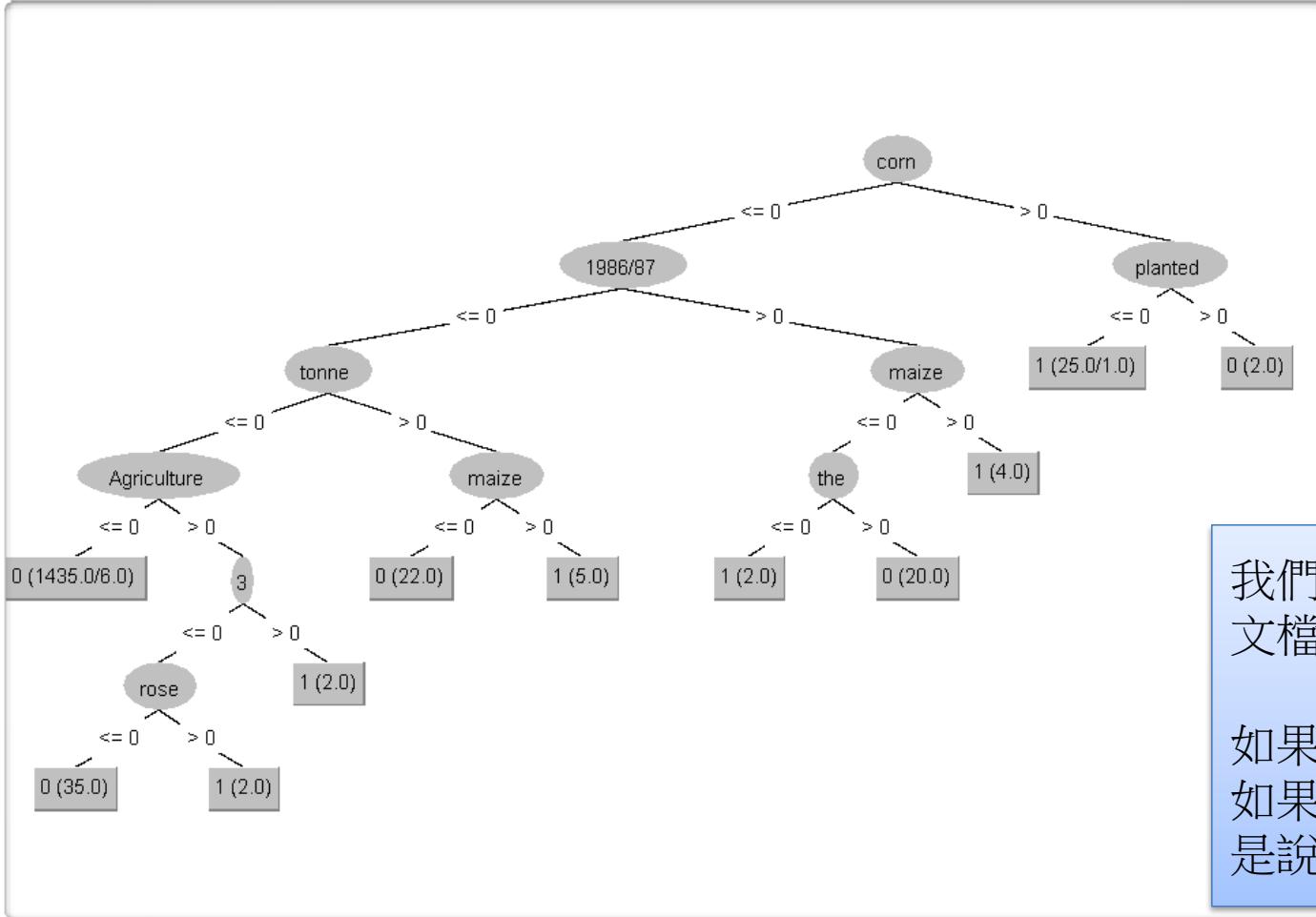


# Lesson 2.4: 文本分類

## ▼觀察樹狀結構。

Weka Classifier Tree Visualizer: 13:33:57 - meta.FilteredClassifier (Reuters-21578 Corn ModApte Train-weka.f... — □ ×

Tree View



我們得到了節點為 "corn"（穀物）的分支，如果文檔包含 "corn"，我們便尋找單詞 "planted"。

如果包含 "planted"，那麼預測分類為 0。

如果不包含 "planted"，那麼預測分類為 1，也就是說，與穀物有關。

## Lesson 2.4: 文本分類

- ❖ 字串(String)屬性
- ❖ **StringToWordVector** 過濾器: 創造許多屬性
- ❖ 查看**StringToWordVector**的選項
- ❖ J48對文本資料的模型
- ❖ 整體的分類器準確率
  - 可能並不是我們真正關切的



THE UNIVERSITY OF  
**WAIKATO**  
*Te Whare Wānanga o Waikato*

# *More Data Mining with Weka*

Class 2 - Lesson 5

評估二類分類器

(*Evaluating 2-class classification*)

Ian H. Witten

Department of Computer Science University of Waikato  
New Zealand

# Lesson 2.5: 評估二類分類器

Class 1 探索Weka界面，處理大數據

Class 2 離散以及文本分類

Class 3 分類規則、關聯規則、聚類

Class 4 選擇屬性以及計算成本

Class 5 神經網路，學習曲線和表現優化

Lesson 2.1 Discretization

Lesson 2.2 監督式離散化

Lesson 2.3 使用J48進行離散化

Lesson 2.4 文本分類

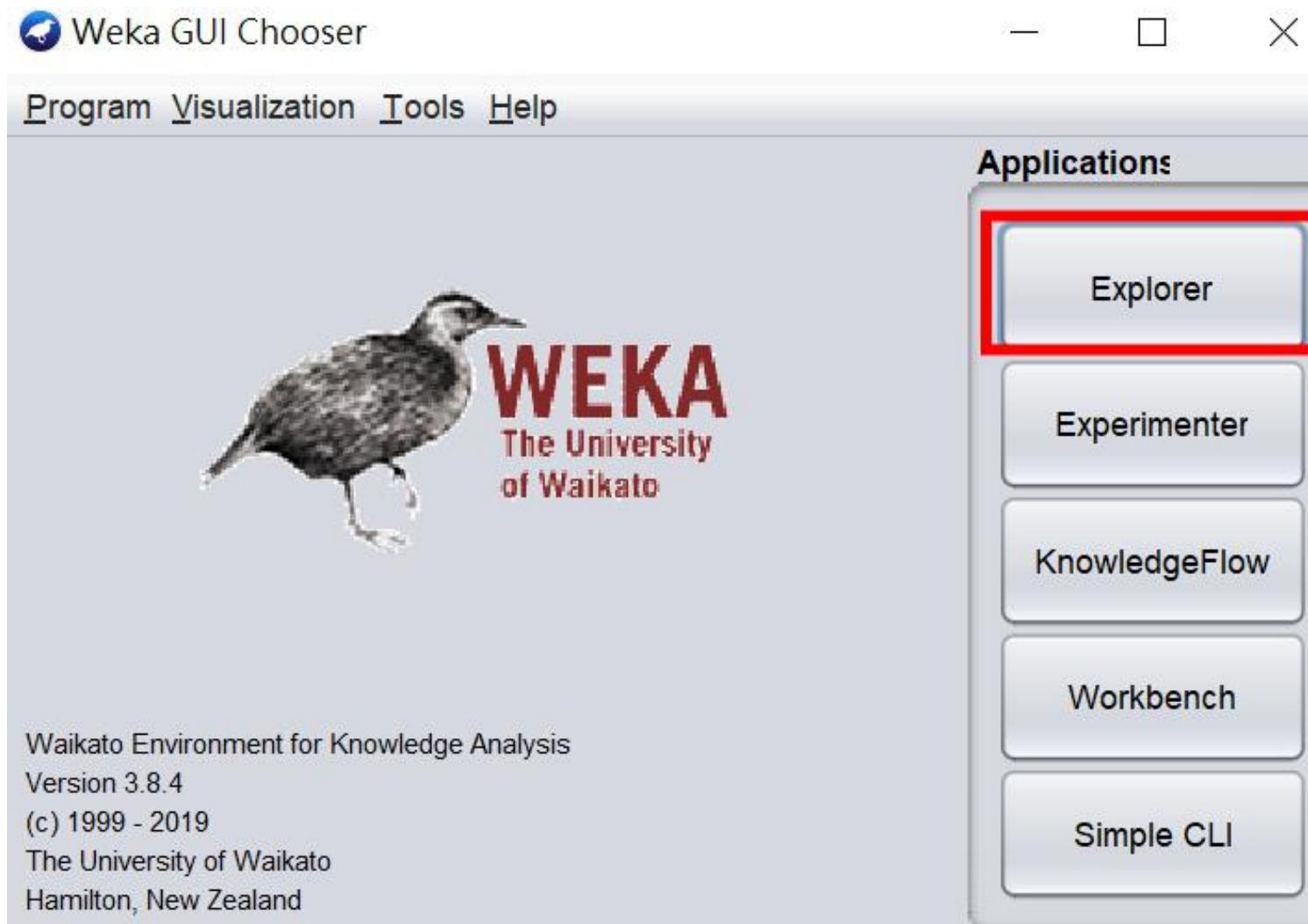
Lesson 2.5 評估二類分類器

Lesson 2.6 Multinomial Naïve Bayes



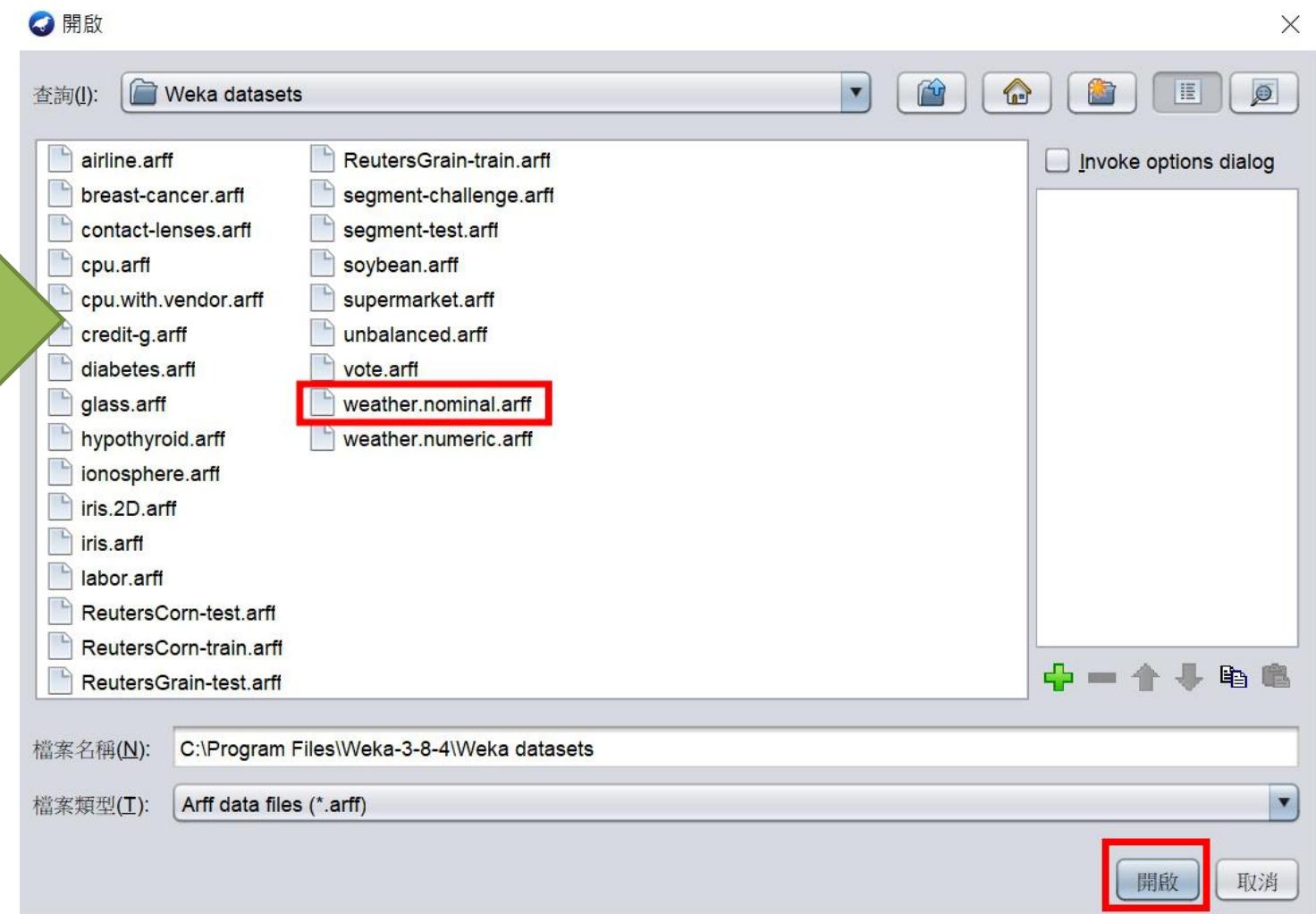
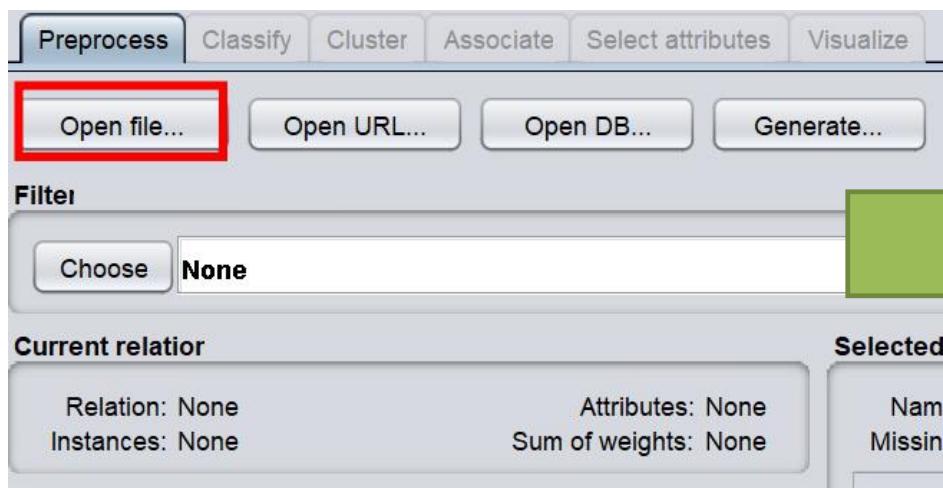
## Lesson 2.5: 評估二類分類器

### 1. 開啟Weka的Explorer



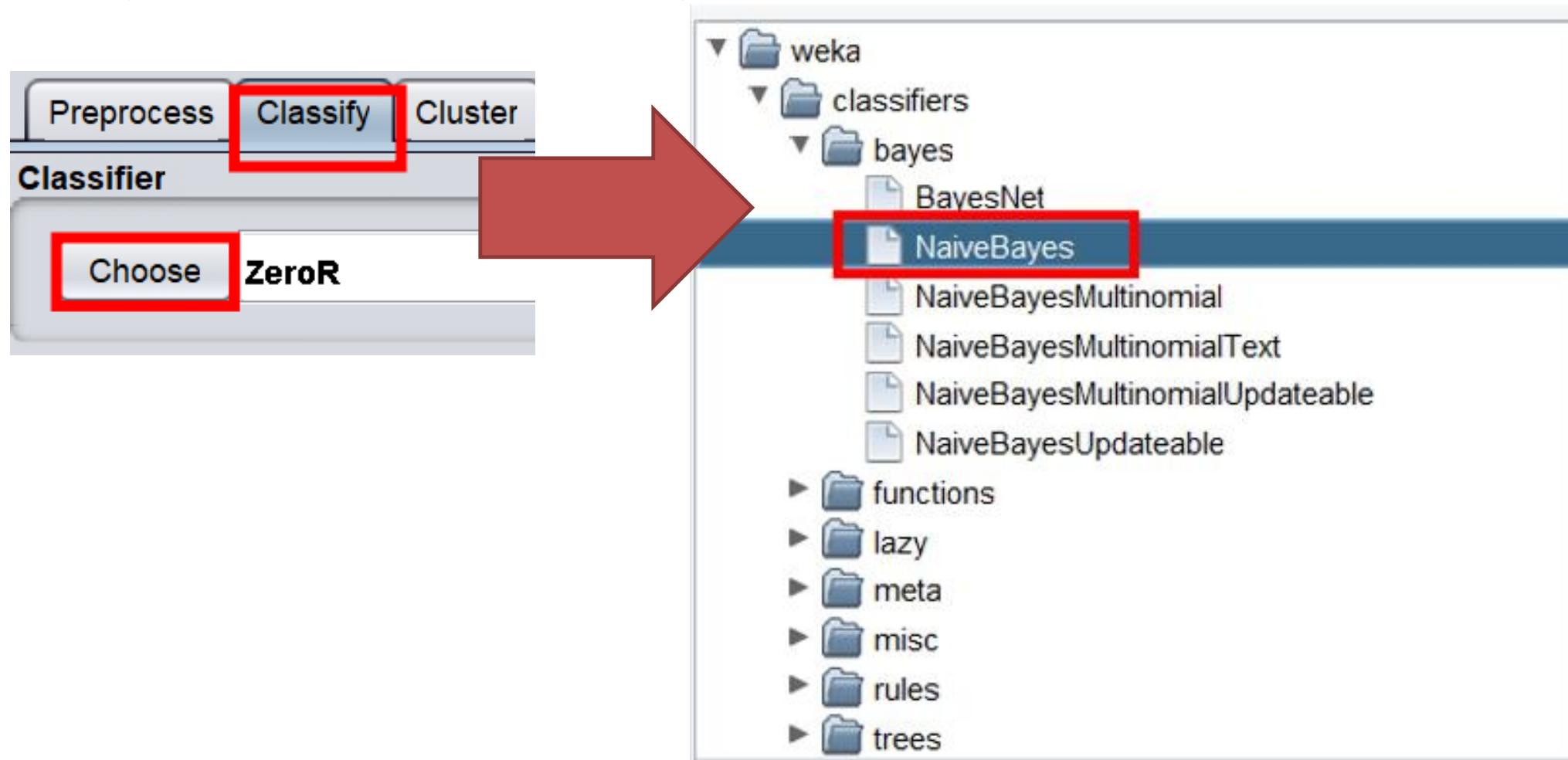
## Lesson 2.5: 評估二類分類器

2. 左鍵單擊Open file...按鈕開啟右圖視窗，進入自行複製的Weka datasets，左鍵單擊**weather.nominal.arff**的檔案後，再以左鍵單擊下方”開啟”按鈕以載入此檔案



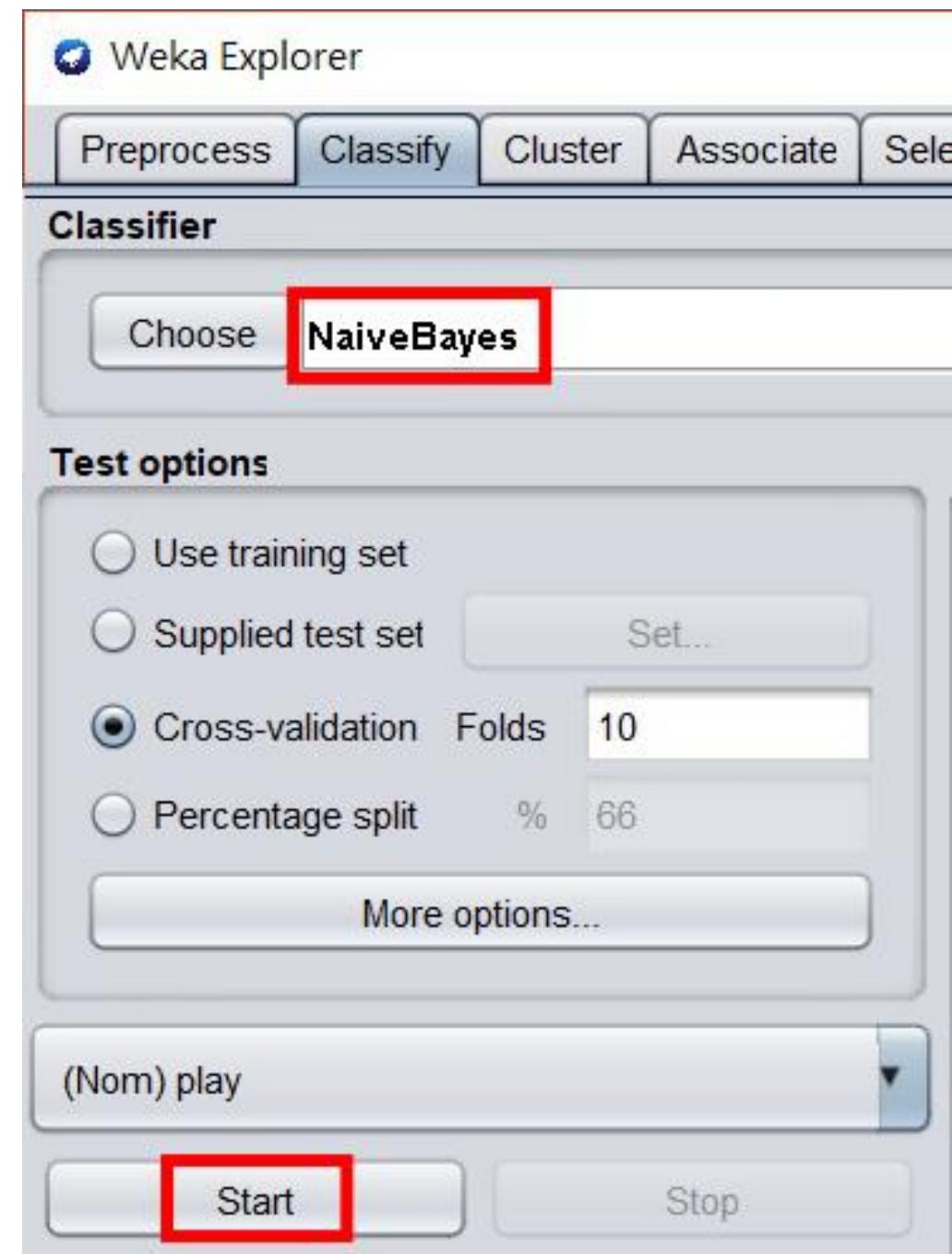
## Lesson 2.5: 評估二類分類器

3. 切換到Classify介面以滑鼠左鍵單擊Choose，在彈出的視窗以左鍵單擊bayes資料夾下的Naive Bayes



## Lesson 2.5: 評估二類分類器

4. 確認選擇了NaiveBayes分類器後，左鍵單擊下方Start運行分類器。



# Lesson 2.5: 評估二類分類器

## ▼執行結果

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

**Classifier**

Choose **NaiveBayes**

**Test options**

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

(Nom) play

Start Stop

**Result list (right-click for options)**

21:59:25 - bayes.NaiveBayes

**Classifier output**

```
Time taken to build model: 0.01 seconds

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

Correctly Classified Instances      8          57.1429 %
Incorrectly Classified Instances   6          42.8571 %
Kappa statistic                   -0.0244
Mean absolute error               0.4374
Root mean squared error           0.4916
Relative absolute error           91.8631 %
Root relative squared error      99.6492 %
Total Number of Instances         14

==== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall   F-Measure  MCC     ROC Area  PRC Area  Cl
      0.778    0.800    0.636     0.778    0.700     -0.026   0.578    0.697    ye
      0.200    0.222    0.333     0.200    0.250     -0.026   0.578    0.557    no
Weighted Avg.      0.571    0.594    0.528     0.571    0.539     -0.026   0.578    0.647

==== Confusion Matrix ===

 a b  <- classified as
7 2 | a = yes
4 1 | b = no
```

## Lesson 2.5: 評估二類分類器

Weather data; Naïve Bayes; 10-fold cross-validation

```
====Confusion Matrix ===
```

```
a b    <-- classified as
```

```
7 2 | a = yes
```

```
4 1 | b = no
```

這是輸出結果中的混淆矩陣。

有7個a被歸為a類，2個a被錯歸為b類。  
1個b被歸為b類，4個b被錯歸為a類。

## Lesson 2.5: 評估二類分類器

我們將 “yes” 當作肯定的類別

Weather data; Naïve Bayes; 10-fold cross-validation

====Confusion Matrix ===			
		a	b
a	classified as	<--	
	a = yes	7	2
b	b = no	4	1

正確肯定  
(true positives) → 7  
錯誤肯定  
(false positives) → 4  
  
(否定的實例被錯分到肯定的類。它們  
看上是肯定的，其實它們是錯誤的)

    → 錯誤否定  
(false negatives)      2  
    → 正確否定  
(true negatives)      1

正確肯定率(TP rate)，即正確肯定的數量7，除以肯定實例(也就是yes)的總數9，得到0.78。

$$\text{TP rate: } \text{TP} / (\text{TP} + \text{FN}) = 7/9 = 0.78$$

類別a(yes)的準確率

錯誤肯定率(FP rate)，即錯誤肯定的數量4，除以否定實例(也就是no)的總數5得到，0.80。

$$\text{FP rate: } \text{FP} / (\text{FP} + \text{TN}) = 4/5 = 0.80$$

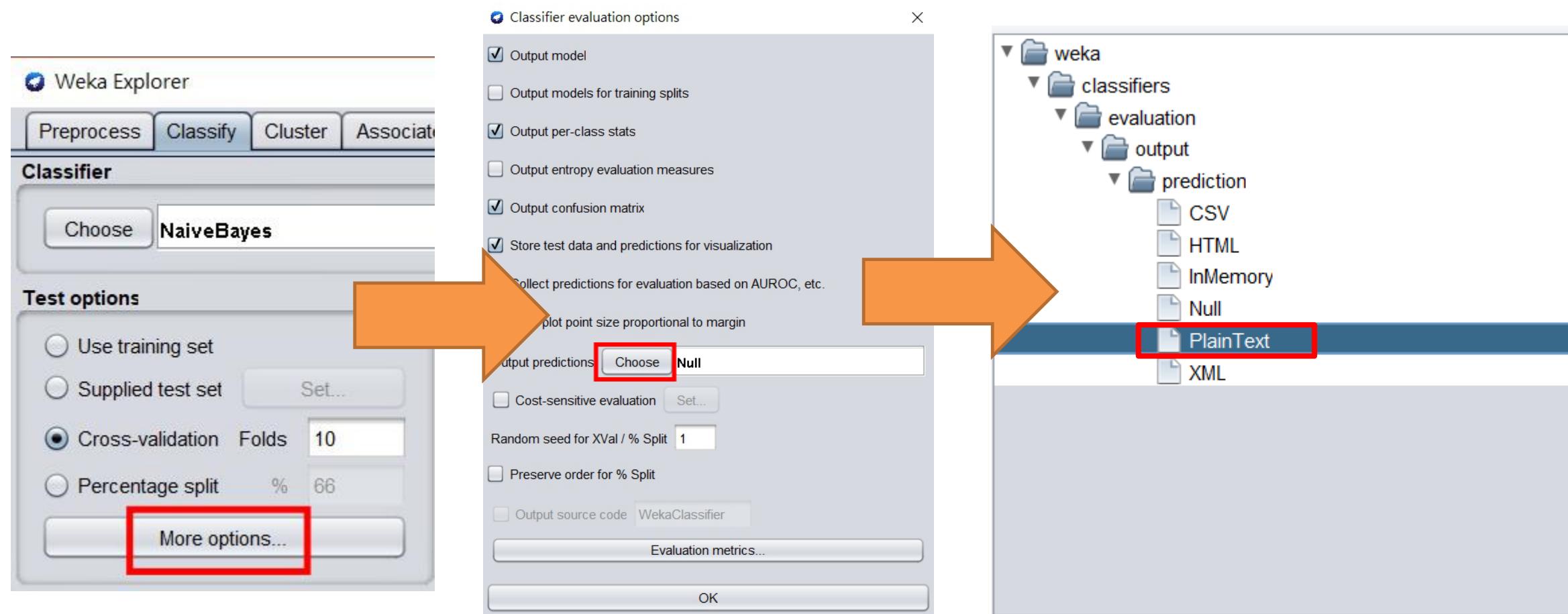
也就是1-類別b(no)的準確率

這節課的重點是這些數值都具有權衡關係——你可以犧牲類別a的準確率來換取類別b的準確率；反之亦然。達到提高準確率的目的。

## Lesson 2.5: 評估二類分類器

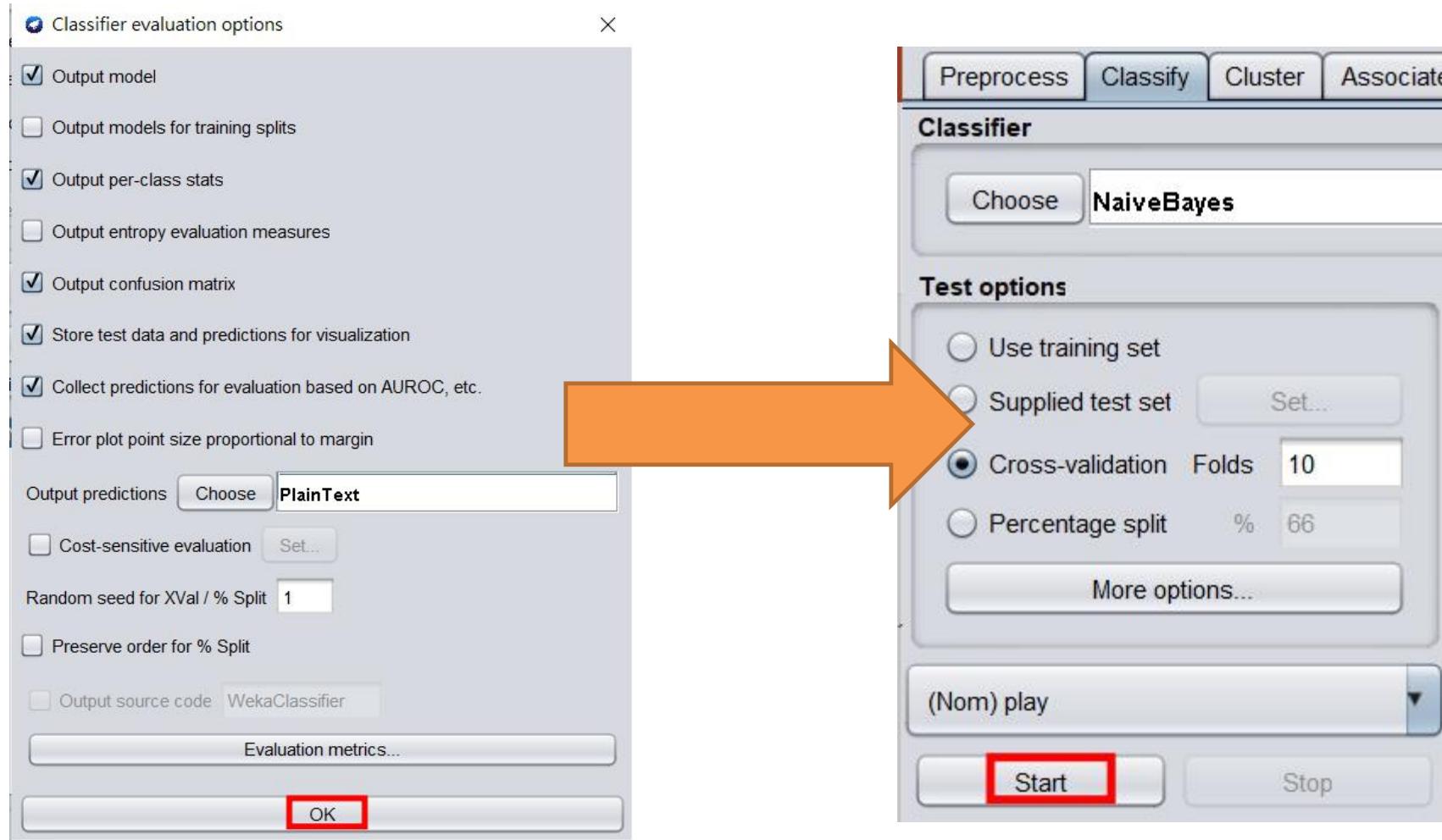
我們接著將預測結果輸出。

1. 左鍵單擊More options...按鈕，在出現的視窗中左鍵單擊Choose按鈕，並在出現的選單中左鍵單擊PlainText。



## Lesson 2.5: 評估二類分類器

2. 左鍵單擊OK按鈕回到Classify面板，接著左鍵單擊Start運行。



## Lesson 2.5: 評估二類分類器

### ▼執行結果

**Naive Bayes**做預測的方式：對比“yes”和“no”的可能性哪個大，預測可能性較大者。

如：**Naive Bayes**預測實例1是“yes”的可能性是92.6%，“no”的可能性是7.4%（兩者相加為1）。

然而**Naive Bayes**最終預測結果是錯誤的，實例1實際上是“no”——這就是錯誤欄中有加號的原因。

The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. In the 'Classifier' section, 'NaiveBayes' is chosen. The 'Test options' panel shows 'Cross-validation Folds 10' selected. The 'Classifier output' panel displays the following text:

```
Time taken to build model: 0 seconds

==== Predictions on test data ====

inst# actual predicted error prediction
1 2:no 1:yes + 0.926
2 1:yes 1:yes - 0.825
1 2:no 1:yes + 0.636
2 1:yes 1:yes - 0.808
1 2:no 2:no - 0.718
2 1:yes 2:no - 0.656
1 2:no 1:yes + 0.579
2 1:yes 1:yes - 0.541
1 2:no 1:yes + 0.515
1 1:yes 2:no - 0.632
1 1:yes 1:yes - 0.84
1 1:yes 1:yes - 0.554
1 1:yes 1:yes - 0.757
1 1:yes 1:yes - 0.778

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

Correctly Classified Instances 8 57.1429 %
Incorrectly Classified Instances 6 42.8571 %
Kappa statistic -0.0244
Mean absolute error 0.4374
Root mean squared error 0.4916
```

The 'Predictions on test data' section is highlighted with a red box.

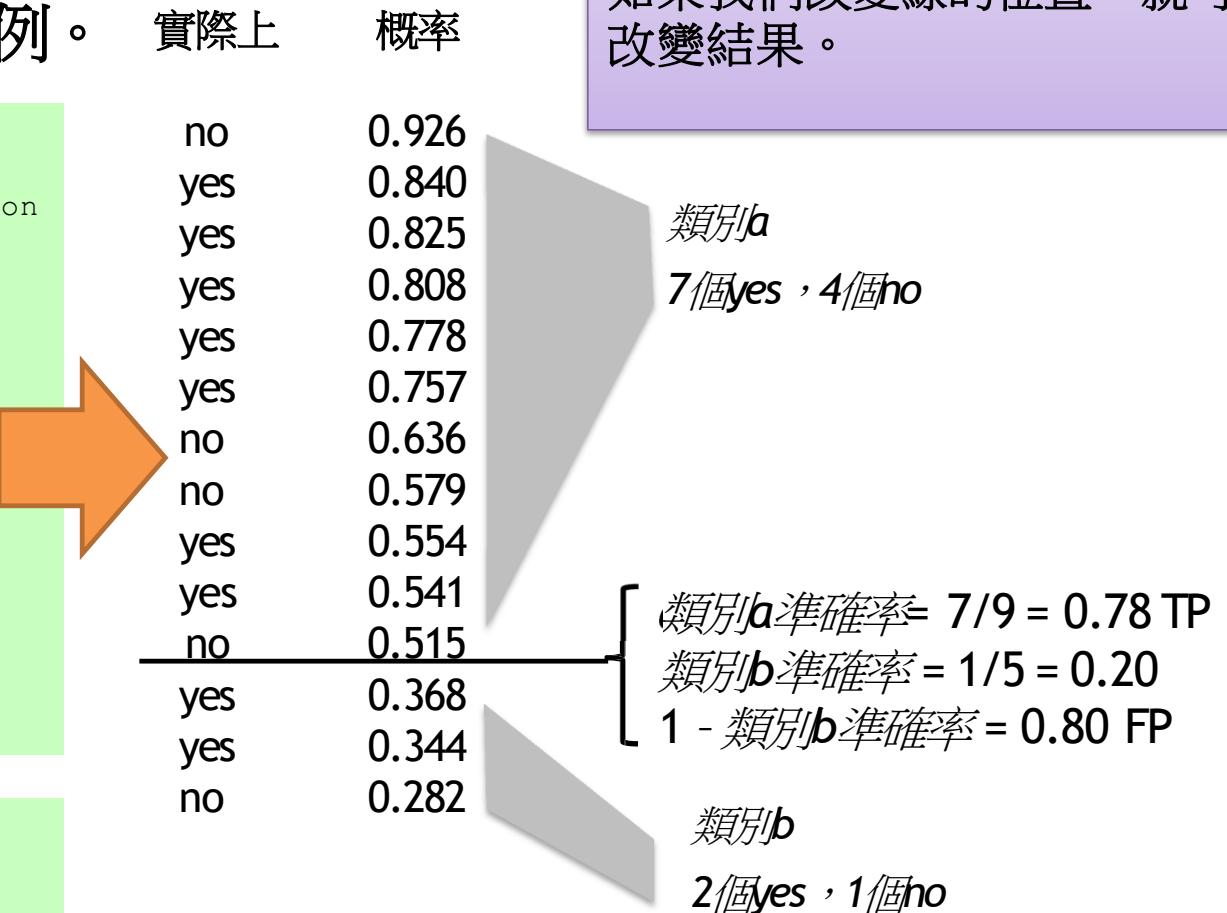
## Lesson 2.5: 評估二類分類器

不同概率的門檻：同樣的數據，簡化為右側的表，表中僅包含實際類別和預測“yes”類別的概率，並依照概率高低由高至低排序實例。 實際上

```
==== Predictions on test data ====
```

```
inst#, actual, predicted, error, probability distribution
 1      2:no      1:yes      + *0.926  0.074
 2      1:yes     1:yes      *0.825  0.175
 1      2:no      1:yes      + *0.636  0.364
 2      1:yes     1:yes      *0.808  0.192
 1      2:no      2:no       0.282 *0.718
 2      1:yes     2:no       + 0.344 *0.656
 1      2:no      1:yes      + *0.579  0.421
 2      1:yes     1:yes      *0.541  0.459
 1      2:no      1:yes      + *0.515  0.485
 1      1:yes     2:no       + 0.368 *0.632
 1      1:yes     1:yes      *0.84   0.16
 1      1:yes     1:yes      *0.554  0.446
 1      1:yes     1:yes      *0.757  0.243
 1      1:yes     1:yes      *0.778  0.222
```

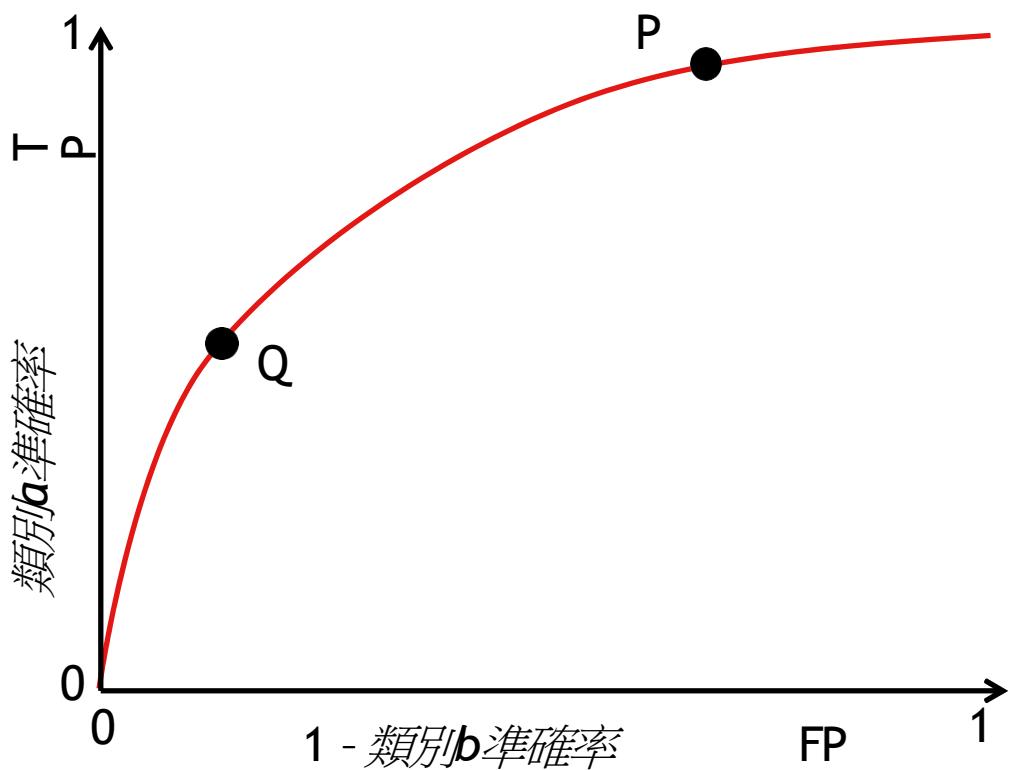
```
a b    <-- classified as
7 2 | a = yes
4 1 | b = no
```



這就像是Naive Bayes在0.5的可能性上劃了條線，線以上的所有實例預測為“yes”，線以下的所有實例預測為“no”。  
如果我們改變線的位置，就可以改變結果。

## Lesson 2.5: 評估二類分類器

### 不同概率的門檻



實際上	概率
no	0.926
yes	0.840
yes	0.825
yes	0.808
yes	0.778
yes	0.757
no	0.636
no	0.579
yes	0.554
yes	0.541
no	0.515
yes	0.368
yes	0.344
no	0.282

如果我們將表中水平線向下挪移，圖中的數值會沿紅線向上。  
假設我們把線劃在第一個實例下方，將所有實例分類為“no”以得到100%“no”分類的準確率，那麼會導致錯誤肯定率為0以及“yes”分類的準確率是0%，正確肯定率也為0，最後得到圖中的0點，。

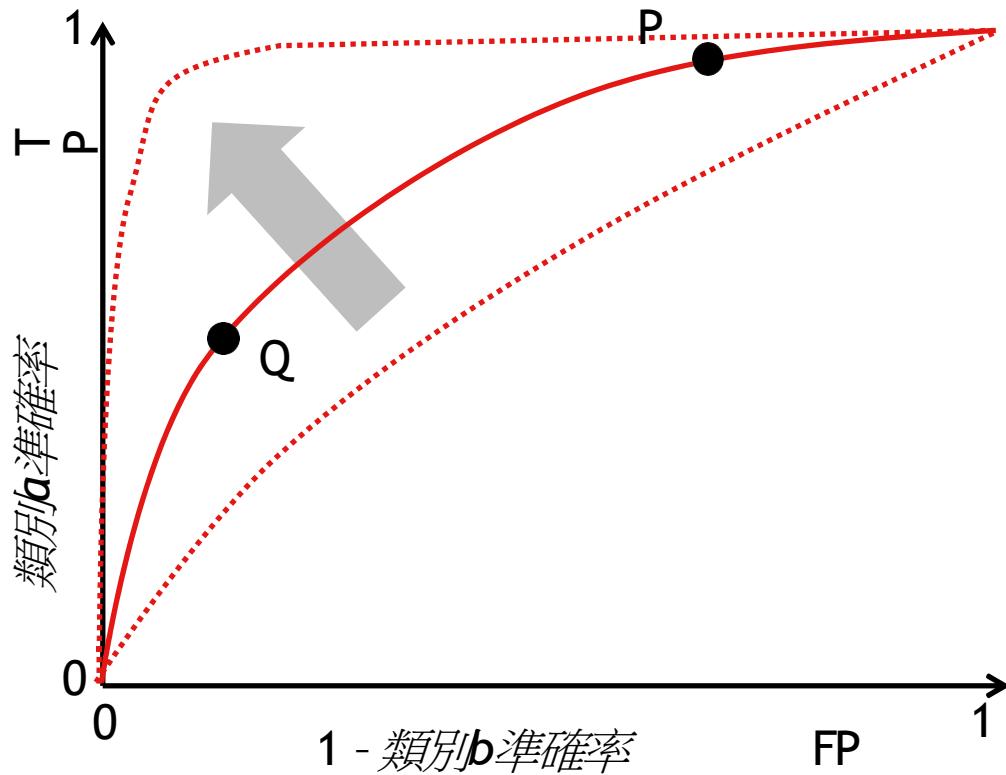
$$\left[ \begin{array}{l} \text{類別a準確率} = 5/9 = 0.56 \text{ TP} \\ \text{類別b準確率} = 4/5 = 0.80 \\ 1 - \text{類別b準確率} = 0.20 \text{ FP} \end{array} \right]$$

$$\left[ \begin{array}{l} \text{類別a準確率} = 7/9 = 0.78 \text{ TP} \\ \text{類別b準確率} = 1/5 = 0.20 \\ 1 - \text{類別b準確率} = 0.80 \text{ FP} \end{array} \right]$$

通過把線劃在不同點上，我們可以權衡a類別和b類別的準確率。

## Lesson 2.5: 評估二類分類器

### 不同概率的門檻



不同的機器學習方法會給出不同的紅色曲線。例如這條紅線下的虛線，但其準確率要低於有P、Q兩點的Naive Bayes線。我們的目標是左上角那條紅虛線，那表示a類和b類最高的準確率，是我們希望的結果。

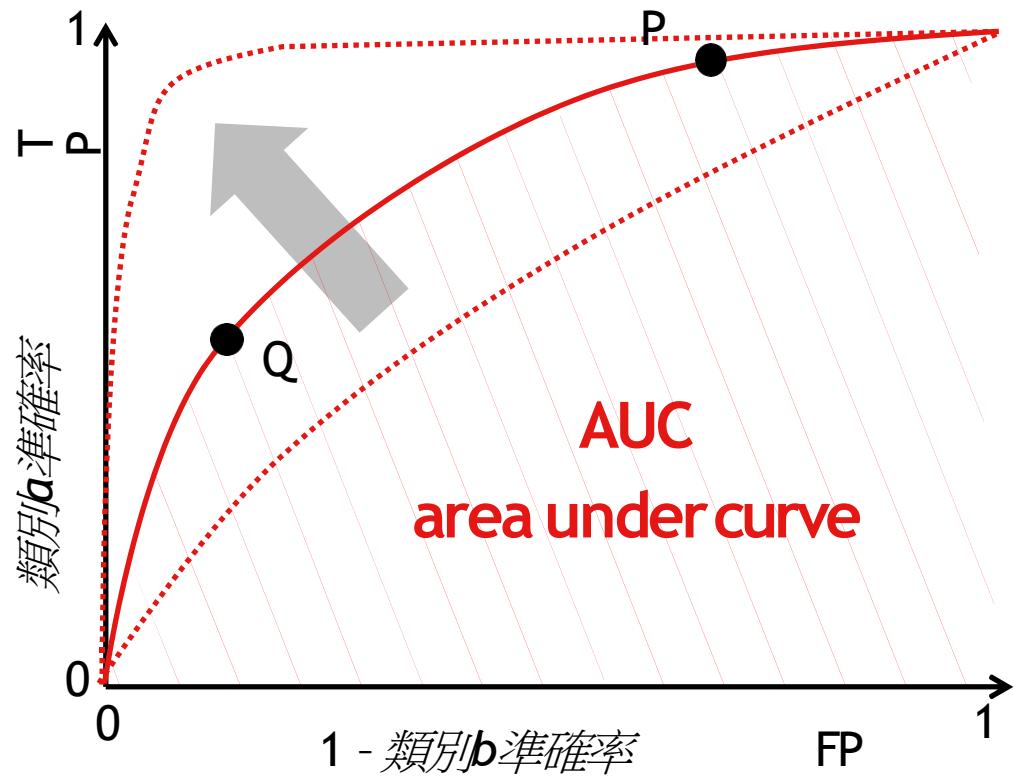
實際上	概率	
no	0.926	類別a準確率 = $5/9 = 0.56$ TP
yes	0.840	類別b準確率 = $4/5 = 0.80$
yes	0.825	1 - 類別b準確率 = 0.20 FP
yes	0.808	
yes	0.778	
yes	0.757	
no	0.636	
no	0.579	
yes	0.554	
yes	0.541	類別a準確率 = $7/9 = 0.78$ TP
no	0.515	類別b準確率 = $1/5 = 0.20$
yes	0.368	1 - 類別b準確率 = 0.80 FP
yes	0.344	
no	0.282	

通過把線劃在不同點上，我們可以權衡a類別和b類別的準確率。

## Lesson 2.5: 評估二類分類器

評估某一分類器整體性能的方法，以有P、Q兩線的Naive Bayes線為例，是根據曲線下方的區域：如果區域面積大，說明我們找到了已用不同權衡關係、不同臨界值評估了的最佳分類器，且此區域面積不會因選擇某一特定的權衡而被影響。

不同概率的門檻



實際上	概率
no	0.926
yes	0.840
yes	0.825
yes	0.808
yes	0.778
yes	0.757
no	0.636
no	0.579
yes	0.554
yes	0.541
no	0.515
yes	0.368
yes	0.344
no	0.282

類別a準確率 =  $5/9 = 0.56$  TP  
類別b準確率 =  $4/5 = 0.80$   
1 - 類別b準確率 = 0.20 FP

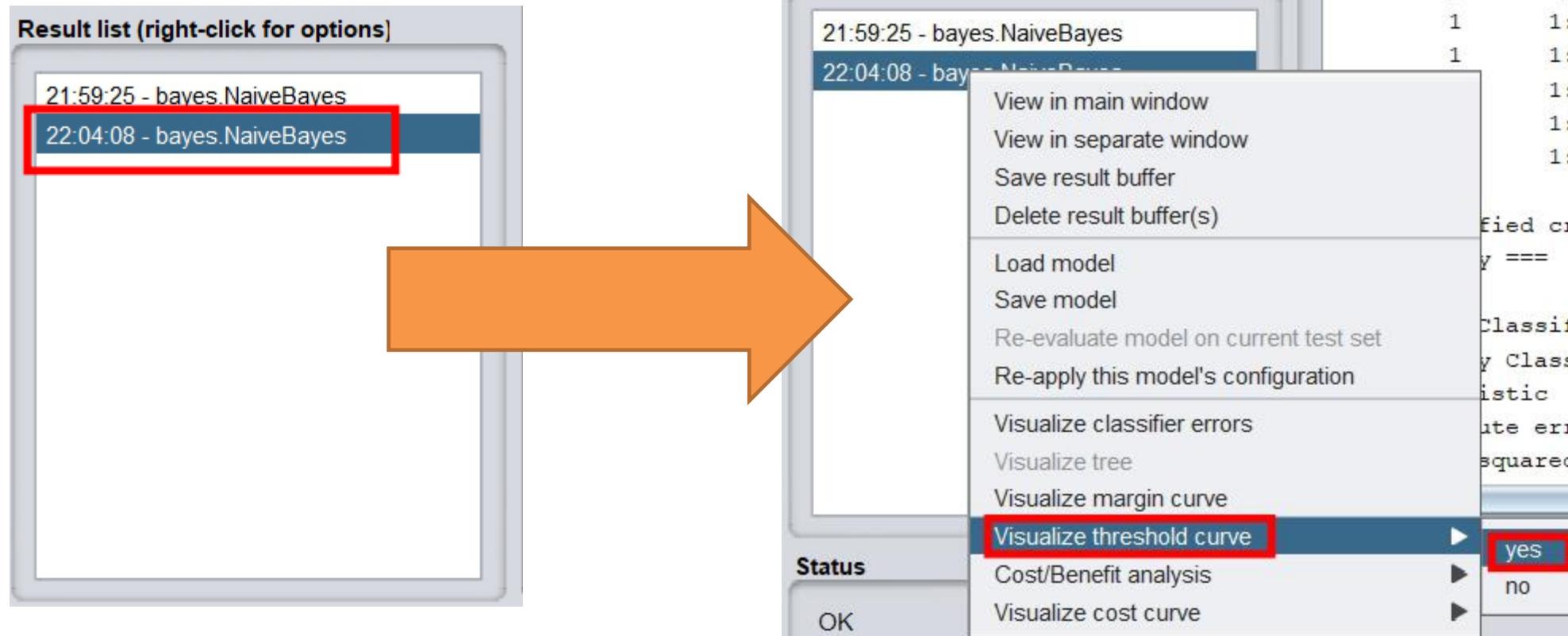
類別a準確率 =  $7/9 = 0.78$  TP  
類別b準確率 =  $1/5 = 0.20$   
1 - 類別b準確率 = 0.80 FP

通過把線劃在不同點上，我們可以權衡a類別和b類別的準確率。

## Lesson 2.5: 評估二類分類器

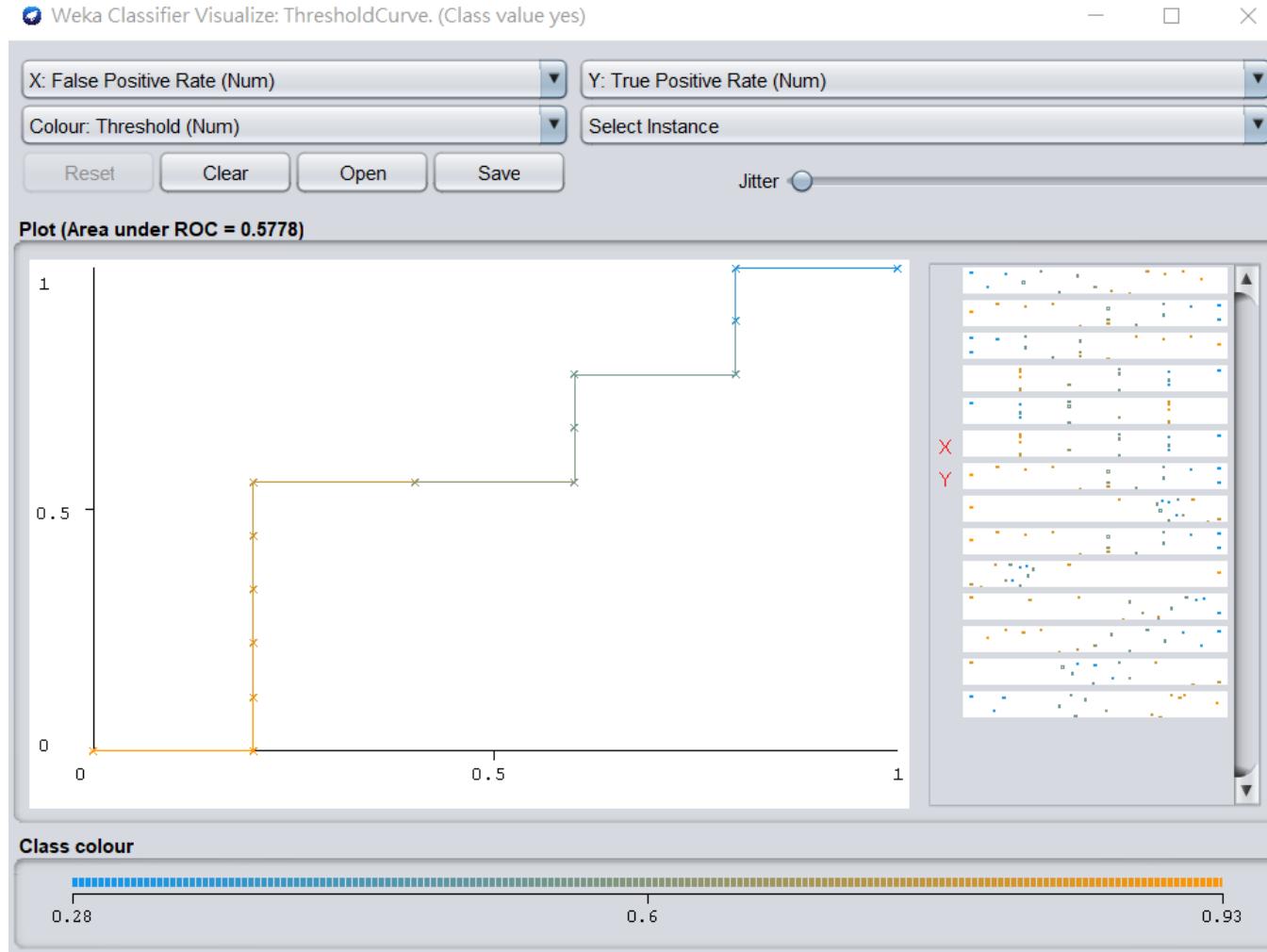
我們接著查看ROC曲線。

1.右鍵單擊剛才的執行記錄，在出現的選單中將滑鼠游標停在**Visualize threshold curve**選項上直到出現**yes/no**選單。最後，在**yes/no**選單中左鍵單擊**yes**選項。



# Lesson 2.5: 評估二類分類器

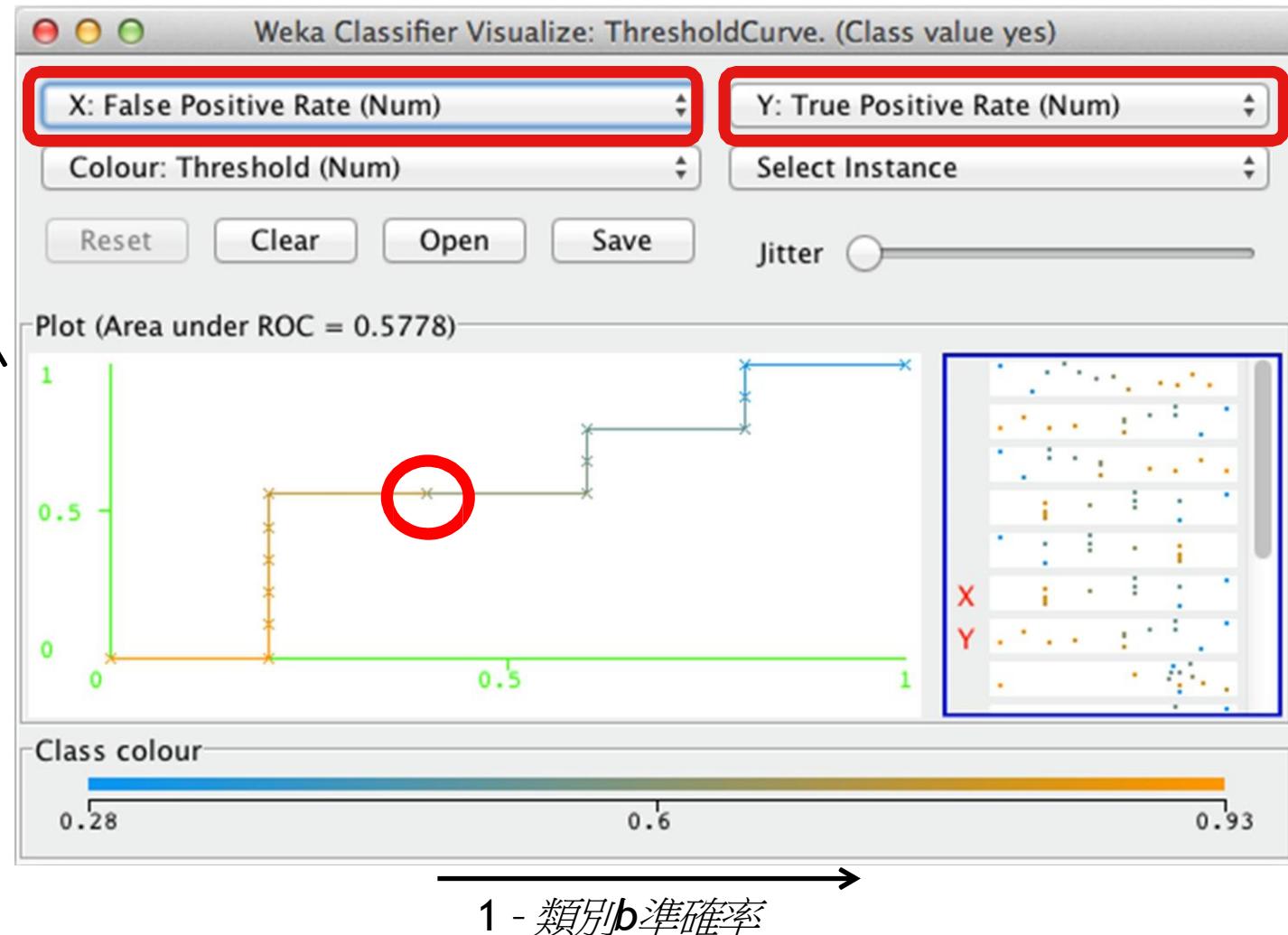
## ▼執行結果



# Lesson 2.5: 評估二類分類器

y軸和x軸分別代表正確肯定率和錯誤肯定率，圖上每點對應表中的每個數值。表中共有15點，圈出的這一點對應的錯誤肯定率是 $2/5$ ，正確肯定率是 $5/9$ 。

“ROC”曲線 (Receiver Operating Characteristic: 因相關歷史命名)

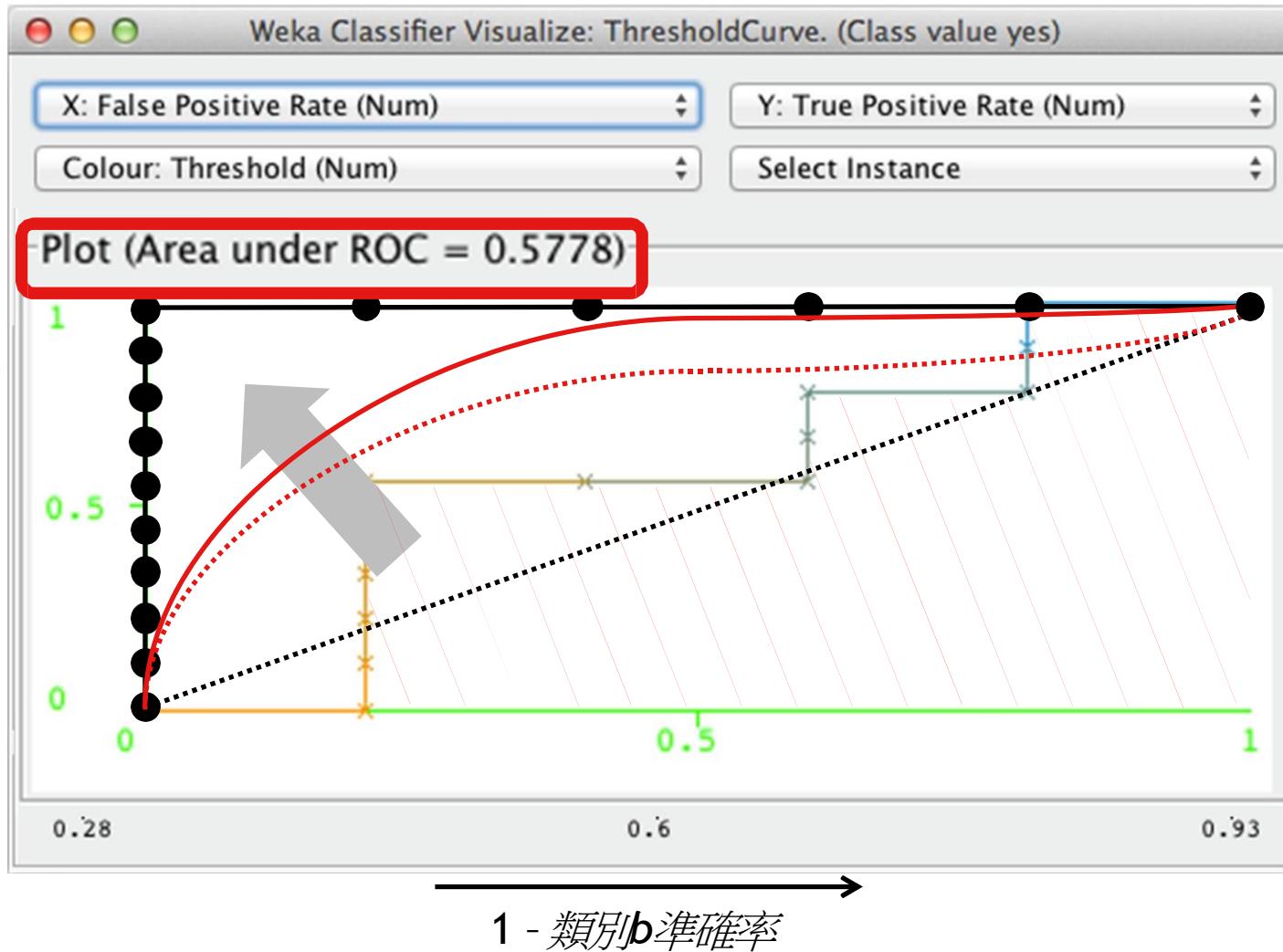


actual	probability	1 - accuracy on class b	accuracy on class a
		FPRate	TPRate
no	0.926	0/5	0/9
yes	0.840	1/5	0/9
yes	0.825	1/5	1/9
yes	0.808	1/5	2/9
yes	0.778	1/5	3/9
yes	0.757	1/5	4/9
no	0.636	1/5	5/9
no	0.579	2/5	5/9
yes	0.554	3/5	5/9
yes	0.541	3/5	6/9
no	0.515	3/5	7/9
yes	0.368	4/5	7/9
yes	0.344	4/5	8/9
no	0.282	4/5	9/9
		5/5	9/9

我們想測量的是曲線下面的區域，此曲線稱為ROC(Receiver Operating Characteristic)曲線。

## Lesson 2.5: 評估二類分類器

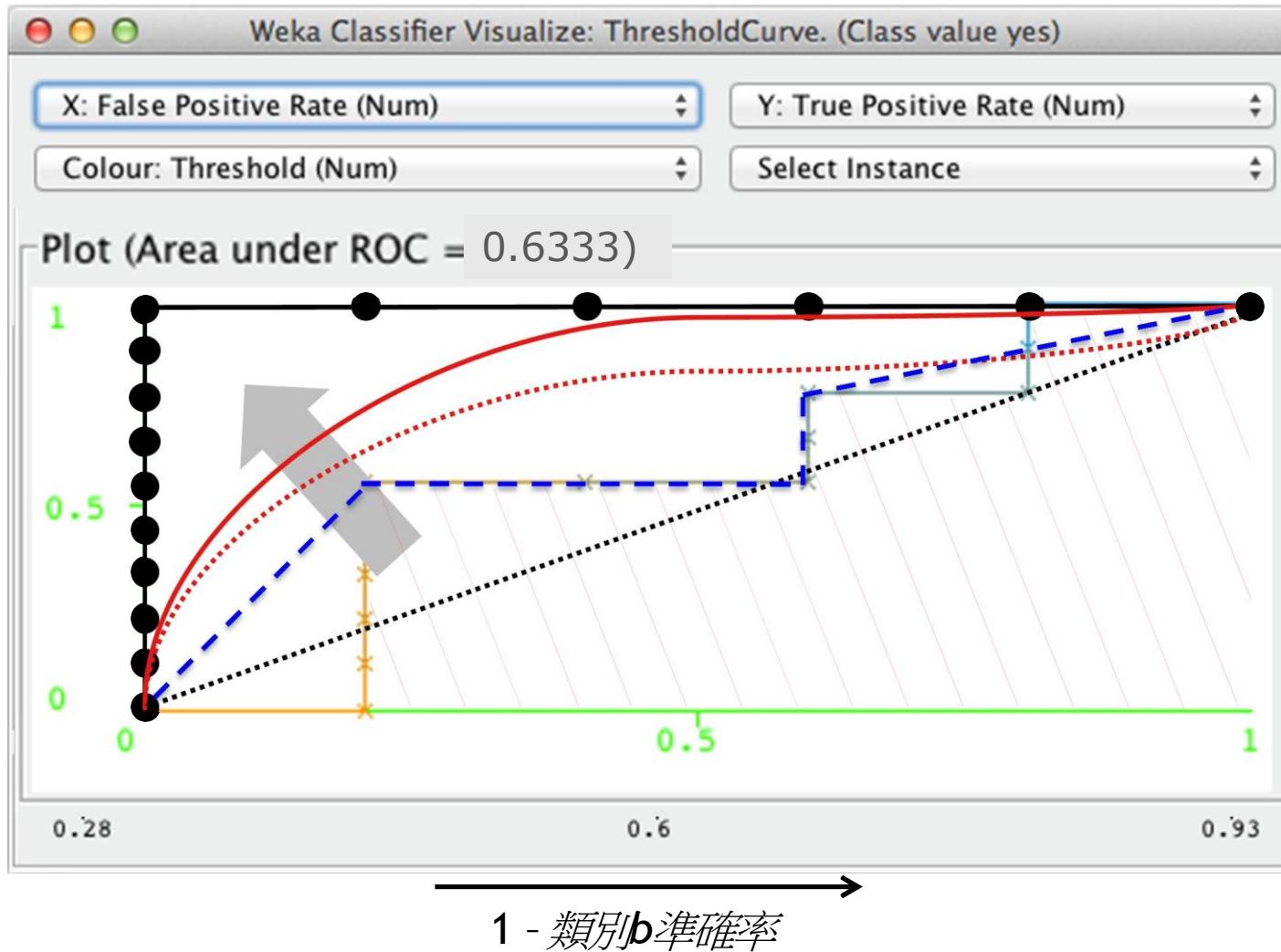
### 優化“ROC”曲線



Weka會顯示ROC曲線下的區域(圖中紅框)，面積為0.5778。如果我們能夠找到一個能向左上角擴展區域的分類器，結果會更好。

## Lesson 2.5: 評估二類分類器

----- 使用J48的ROC曲線: Area under ROC = 0.6333



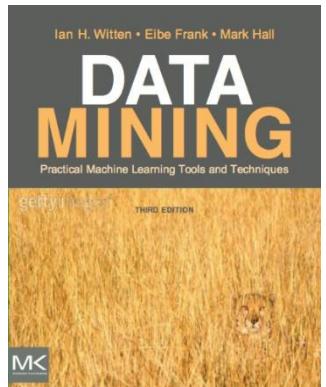
實際上，如果我們在同一資料集上使用J48，我們會得到一條曲線(圖中藍色虛線)。曲線下面積是0.6333，高於Naive Bayes，是一條更好的曲線。

## Lesson 2.5: 評估二類分類器

- ❖ “每個類別的準確率”臨界值曲線
  - 對比某一類的準確率與另一類的準確率，描繪出兩個類的權衡關係
- ❖ ROC 曲線: TP rate (y 座標) 對比 FP rate (x 座標)
  - 從左下方到右上方
  - 好的分類器會像左上角延伸
  - 對角線對應的是隨機選擇的結果
- ❖ AUC (在[ROC] 曲線下方的區域) - 衡量整體性能
  - 對應分類器判斷隨機選擇的肯定測試實例高於隨機選擇的否定測試實例的可能性

### 課程文本

- ❖ Section 5.2 *Counting the cost*, subsection “ROC curves”





THE UNIVERSITY OF  
**WAIKATO**  
*Te Whare Wānanga o Waikato*

# *More Data Mining with Weka*

Class 2 - Lesson 6

*Multinomial (多項式) Naïve Bayes*

Ian H. Witten

Department of Computer Science University of  
Waikato  
New Zealand

# Lesson 2.6: Multinomial Naïve Bayes

Class 1 探索Weka界面，處理大數據

Class 2 離散以及文本分類

Class 3 分類規則、關聯規則、聚類

Class 4 選擇屬性以及計算成本

Class 5 神經網路，學習曲線和表現優化

Lesson 2.1 Discretization

Lesson 2.2 監督式離散化

Lesson 2.3 使用J48進行離散化

Lesson 2.4 文本分類

Lesson 2.5 評估二類分類器

Lesson 2.6 Multinomial Naïve Bayes



## Lesson 2.6: Multinomial Naïve Bayes

記得Naïve Bayes嗎？

- ❖ 基於例證 $E$ 事件 $H$ 出現的概率
- ❖ 例證分成幾個獨立的部分

$$P_1[H \cap E] = \frac{Pr[E | H]Pr[H]}{Pr[E]}$$

↓                                    ↓  
後來的概率                          一開始的概率  
    ↑                                   ↑  
    類別                                  實例

$$Pr[E | H] = Pr[E_1 | H]Pr[E_2 | H]...Pr[E_n | H]$$

文本分類： $E_i$ 是單字*i*出現的次數

- ❖ 但是
  - *Naïve Bayes*認為沒出現的單字與有出現的單字同等重要
  - *Naïve Bayes*忽略一個單字的重複的次數多寡
  - *Naïve Bayes*在處理常見的、不常見的單字上，方法是一樣的

## Lesson 2.6: Multinomial Naïve Bayes

### Multinomial Naïve Bayes

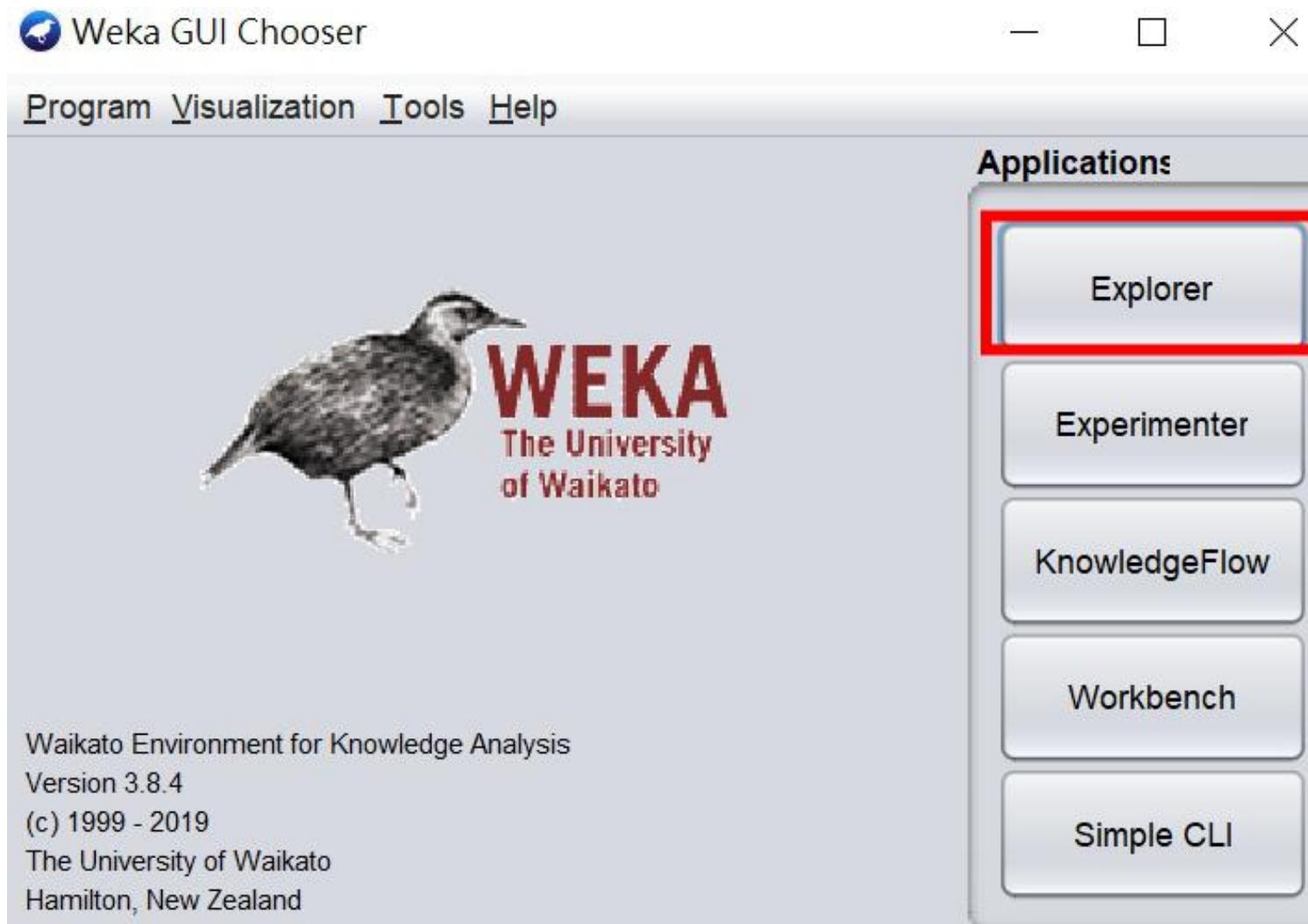
(給好奇它如何運作的人)

$$\Pr[E | H] = \frac{\Pr[E_1 | H] \Pr[E_2 | H] \dots \Pr[E_n | H]}{N! \times \prod_{i=1}^k \frac{p_i^{n_i}}{n_i!}}$$

- ❖  $p_i$ 是單字*i*對類別*H*所有文檔的概率
- ❖  $n_i$ 是它出現在這個文件中的次數
- ❖  $N = n_1 + n_2 + \dots + n_k$ 為本文檔的字數  
(階層“!”用來計算不同詞序的概率)

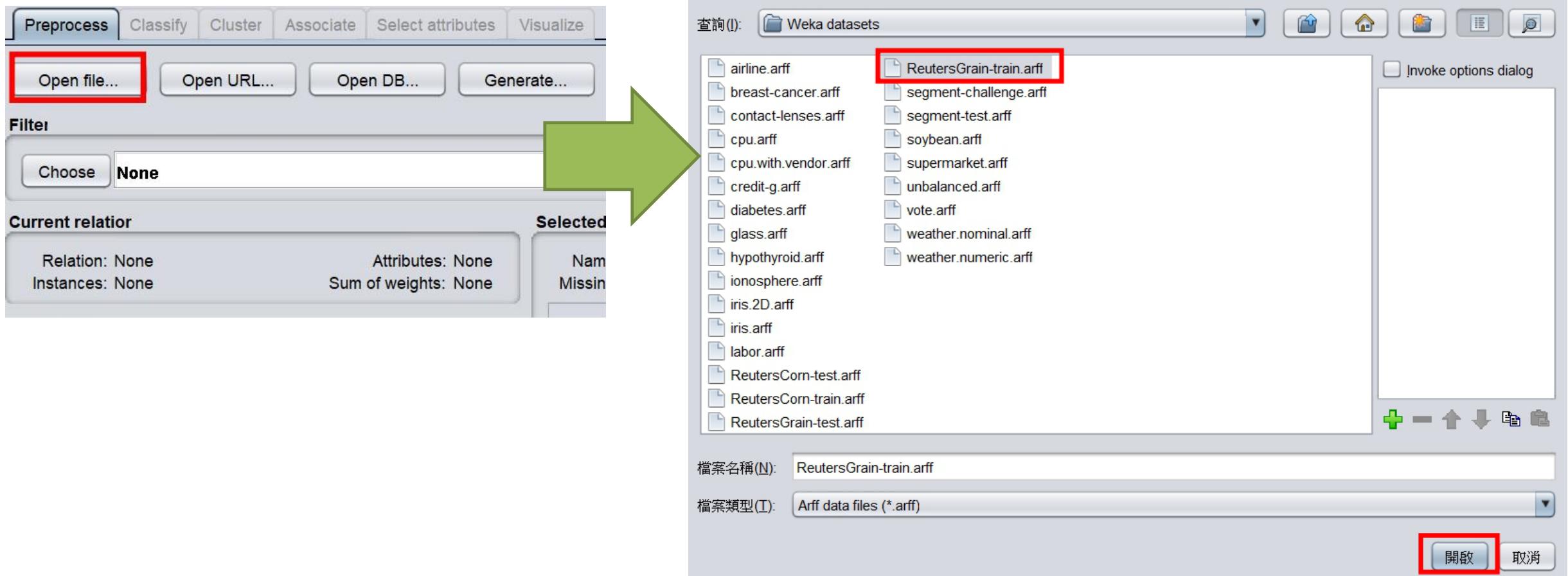
# Lesson 2.6: Multinomial Naïve Bayes

## 1. 開啟Weka的Explorer



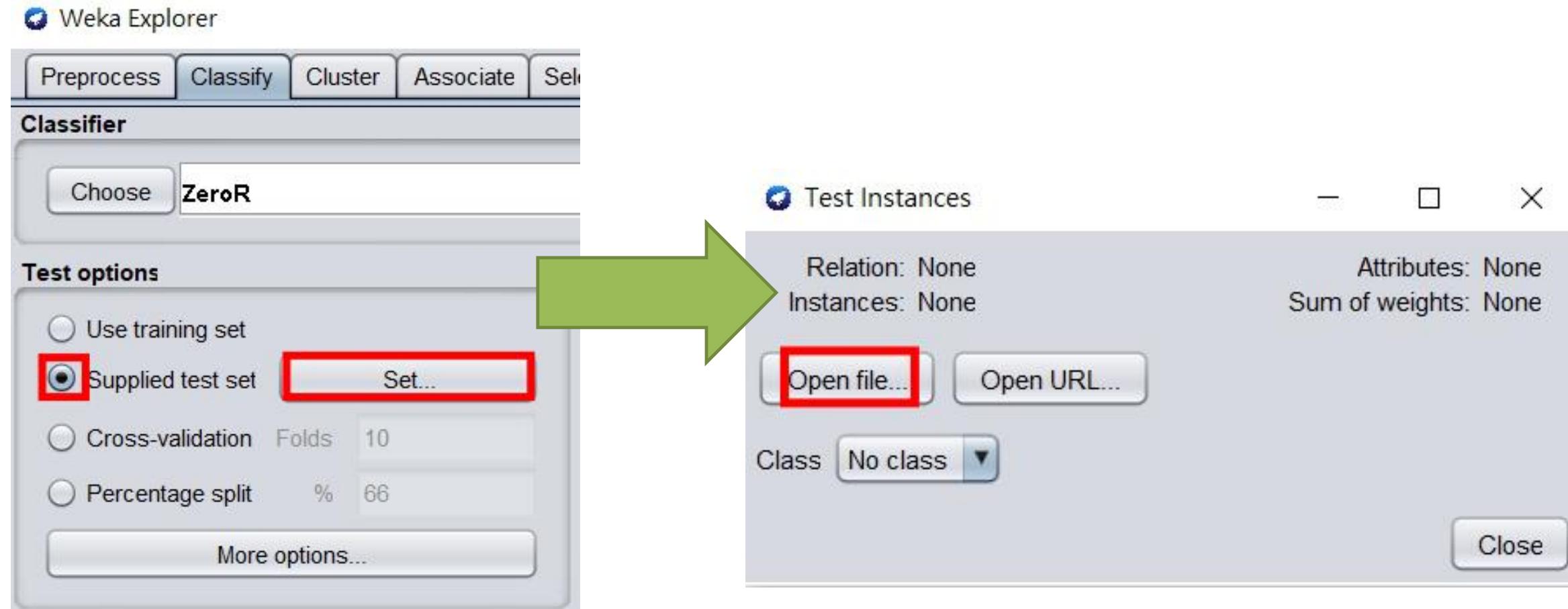
## Lesson 2.6: Multinomial Naïve Bayes

2. 左鍵單擊Open file...按鈕開啟右圖視窗，進入自行複製的Weka datasets，左鍵單擊**ReutersGrain-train.arff**的檔案後，再以左鍵單擊下方“開啟”按鈕以載入此檔案



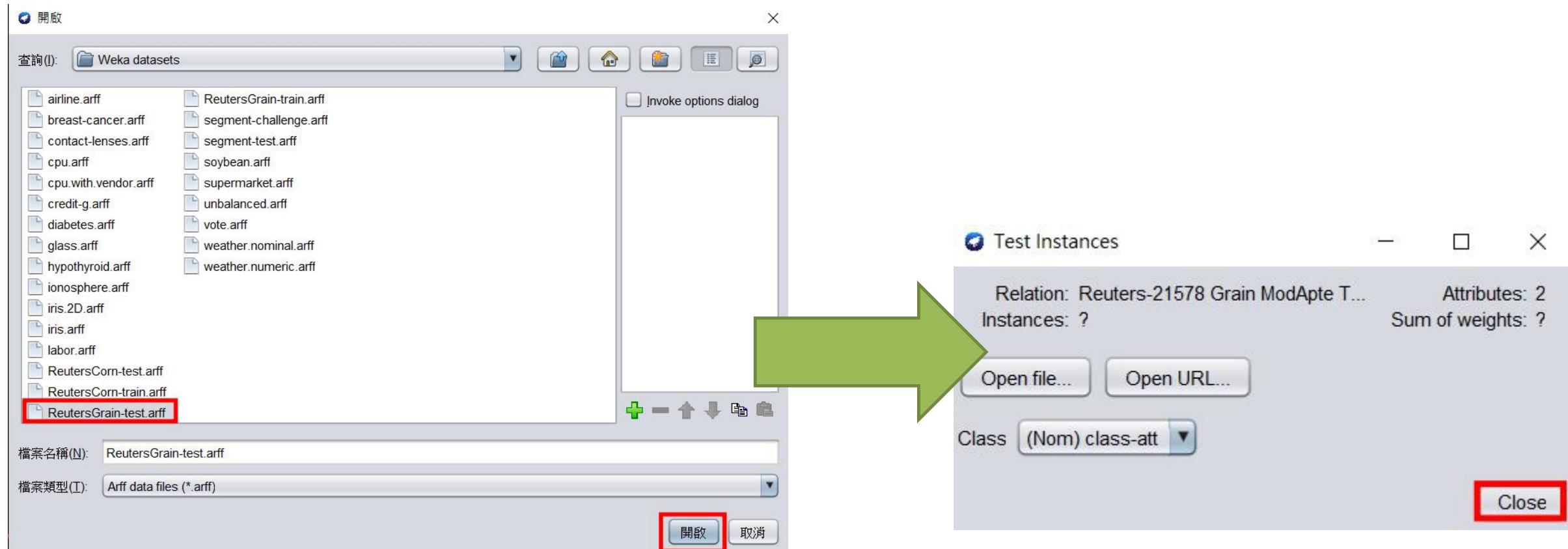
## Lesson 2.6: Multinomial Naïve Bayes

3. 切換到Classify面板，左鍵點選Supplied test set前方圓圈，並以左鍵單擊後方Set按鈕。在出現的視窗(右圖)中以左鍵單擊Open file...按鈕。



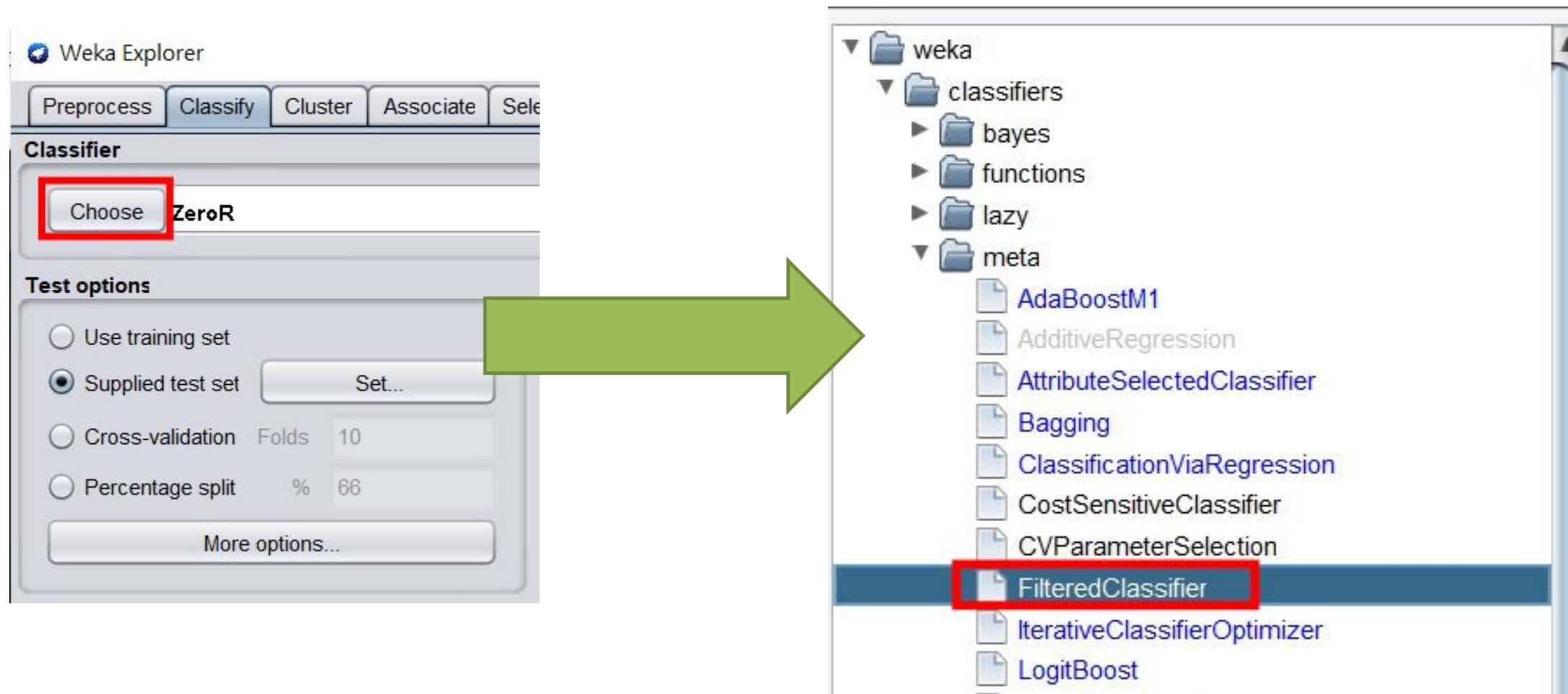
## Lesson 2.6: Multinomial Naïve Bayes

4. 進入自行複製的Weka datasets，左鍵單擊**ReutersGrain-test.arff**檔案，然後以左鍵單擊下方的開啟按鈕回到Test Instances視窗(右圖)。接著在Test Instances視窗左鍵單擊Close按鈕。



## Lesson 2.6: Multinomial Naïve Bayes

5. 在Classify面板中，左鍵單擊Choose按鈕，左鍵單擊weka/classifiers/meta路徑中的FilteredClassifier分類器。



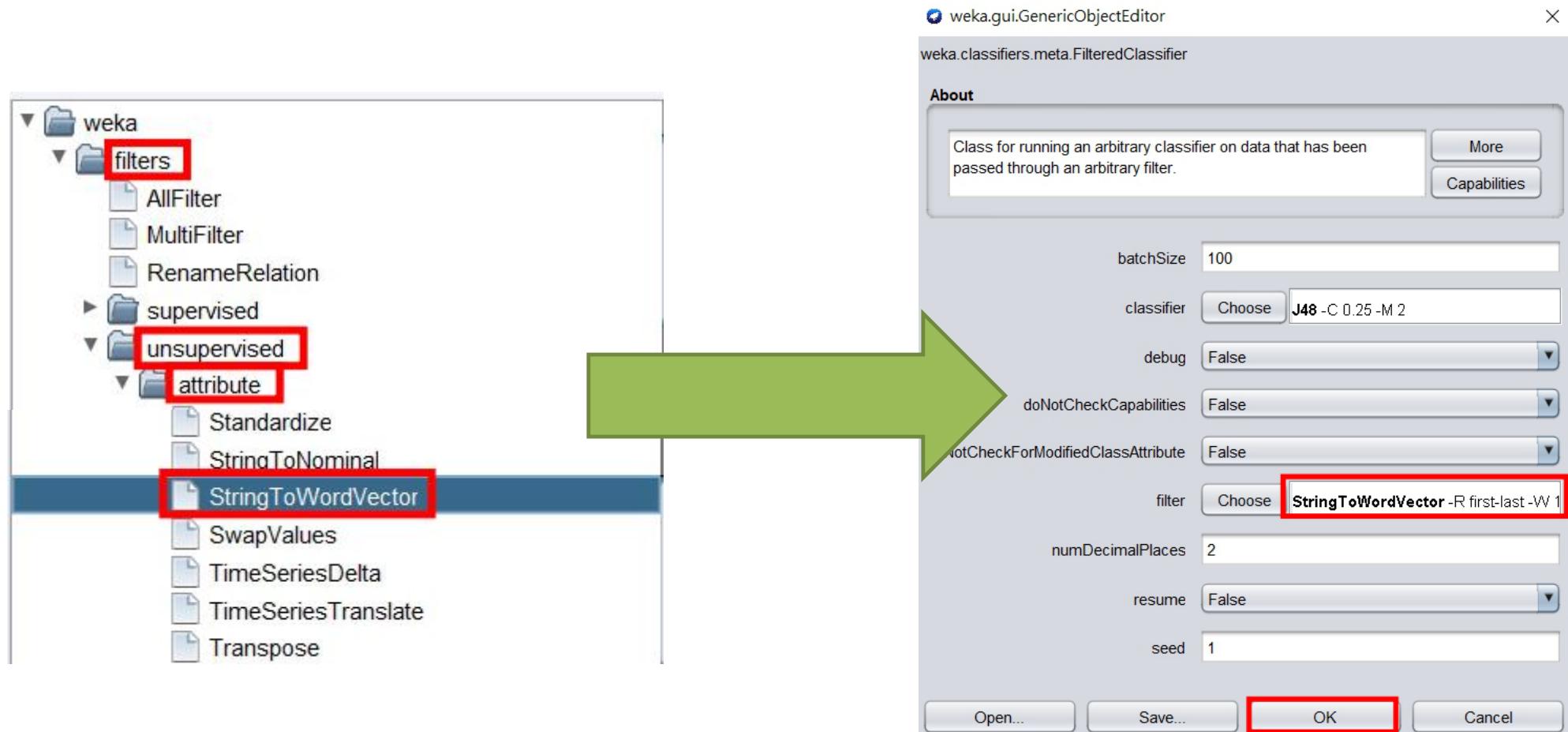
## Lesson 2.6: Multinomial Naïve Bayes

6. 左鍵單擊分類器名稱(左圖紅框處)，開啟配置視窗(右圖)。在配置視窗中左鍵單擊Choose按鈕。



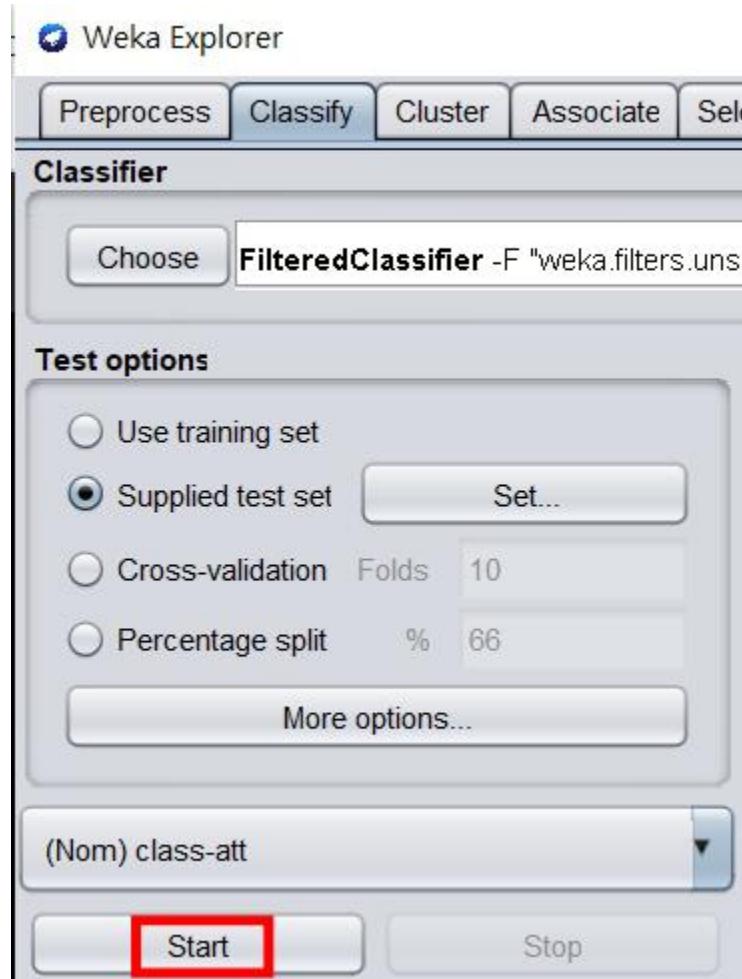
## Lesson 2.6: Multinomial Naïve Bayes

7. 左鍵單擊出現的選單中weka/filters/unsupervised/attribute路徑下的StringToWordVector過濾器。回到配置視窗(右圖)後，確認選擇好StringToWordVector過濾器，左鍵單擊OK按鈕。



## Lesson 2.6: Multinomial Naïve Bayes

8. 回到Classify面板，左鍵單擊Start按鈕。



# Lesson 2.6: Multinomial Naïve Bayes

## ▼執行結果

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose FilteredClassifier -F "weka.filters.unsupervised.attribute.StringToWordVector -R first-last -W 1000 -prune-rate 1.0 -N 0 -stemmer weka.core.stemmers.NullStemmer -"

Test options

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

(Nom) class-att

Start Stop

Result list (right-click for options)

22:44:28 - meta.FilteredClassifier

Classifier output

```
==== Evaluation on test set ====
Time taken to test model on supplied test set: 0.66 seconds
==== Summary ====
Correctly Classified Instances      582      96.3576 %
Incorrectly Classified Instances    22      3.6424 %
Kappa statistic                      0.7563
Mean absolute error                  0.043
Root mean squared error              0.1859
Relative absolute error              28.9093 %
Root relative squared error         63.3132 %
Total Number of Instances           604
```

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

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
a	0.995	0.333	0.966	0.995	0.980	0.768	0.906	0.981
b	0.667	0.005	0.927	0.667	0.776	0.768	0.906	0.767
Weighted Avg.	0.964	0.302	0.963	0.964	0.961	0.768	0.906	0.961

==== Confusion Matrix ====

	a	b	<-- classified as
a	544	3	a = 0
b	19	38	b = 1

得到96%的準確率，但是穀物類的準確率並不樂觀。57個實例中只有38個正確。另外，我們可以在結果中查看ROC區域。

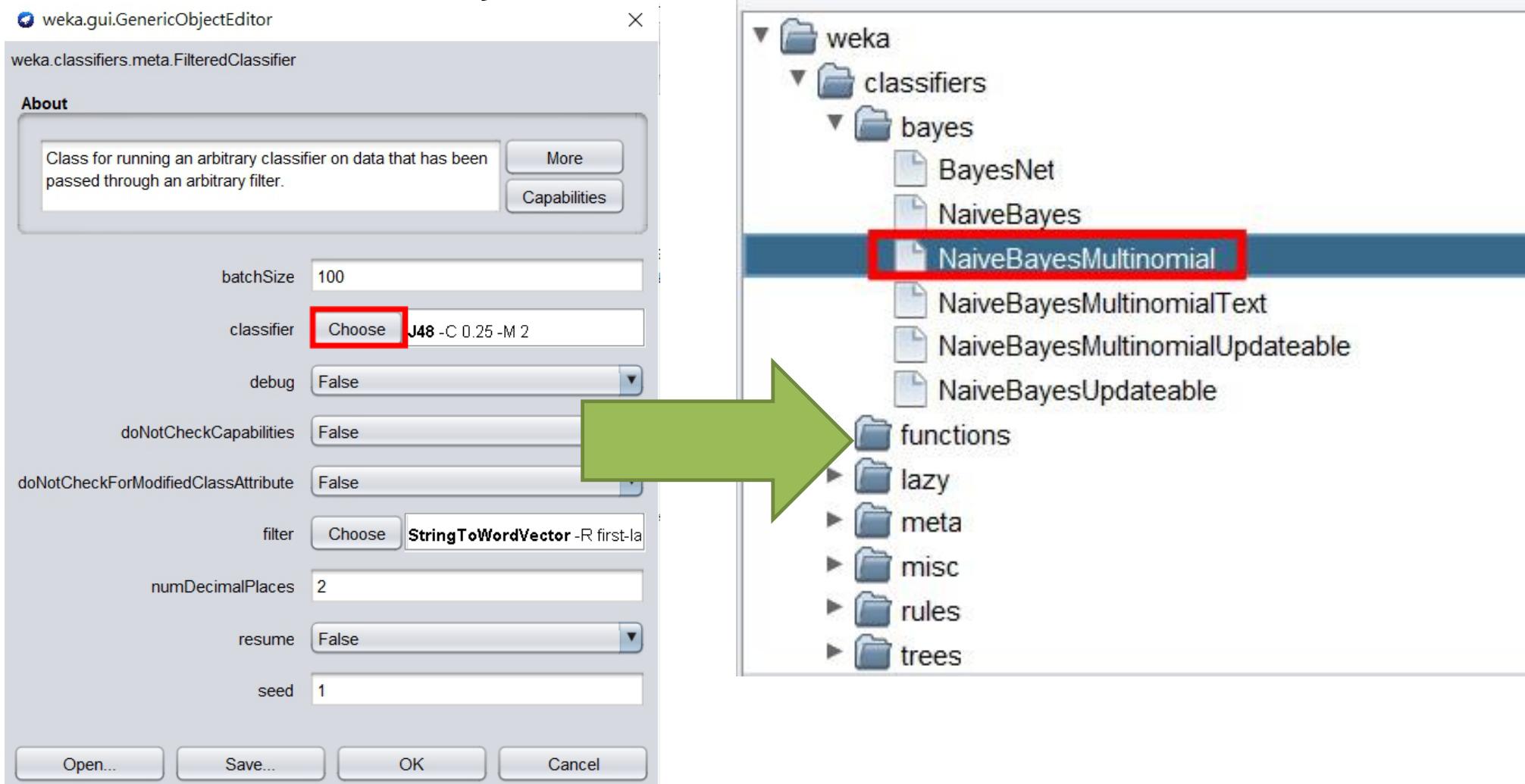
## Lesson 2.6: Multinomial Naïve Bayes

9. 左鍵單擊分類器名稱(圖中紅框處)，以開啟配置視窗。



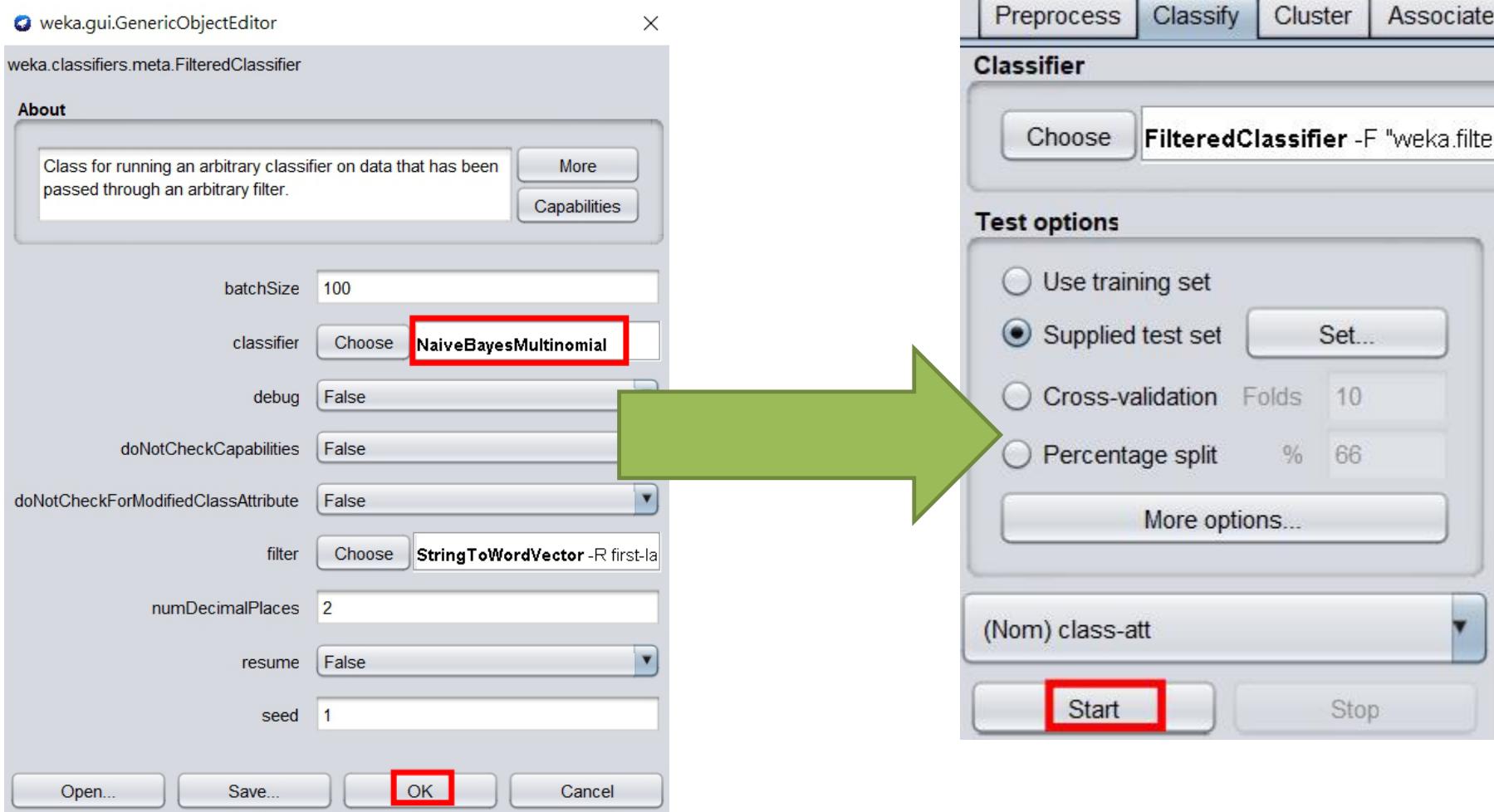
## Lesson 2.6: Multinomial Naïve Bayes

10. 在配置視窗(左圖)中左鍵單擊Choose按鈕。左鍵單擊出現的選單中 weka/classifiers/bayes 路徑下的 NaiveBayesMultinomial 分類器。



## Lesson 2.6: Multinomial Naïve Bayes

11. 確認選擇好NaiveBayesMultinomial分類器後，左鍵單擊OK按鈕。回到Classify面板，左鍵單擊Start按鈕。



# Lesson 2.6: Multinomial Naïve Bayes

## ▼執行結果：

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose FilteredClassifier -F "weka.filters.unsupervised.attribute.StringToWordVector -R first-last -W 1000 -prune-rate 1.0 -N 0 -stemmer weka.core.stemmers.NullStemmer -sto

Test options

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

(Nom) class-att

Start Stop

Result list (right-click for options)

22:44:28 - meta.FilteredClassifier  
22:49:57 - meta.FilteredClassifier

Classifier output

```
==== Evaluation on test set ====
Time taken to test model on supplied test set: 0.25 seconds
==== Summary ====
Correctly Classified Instances      548          90.7285 %
Incorrectly Classified Instances   56           9.2715 %
Kappa statistic                   0.6016
Mean absolute error               0.0946
Root mean squared error          0.2944
Relative absolute error           63.6592 %
Root relative squared error     100.2715 %
Total Number of Instances        604
```

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

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	C1
a	0.907	0.088	0.990	0.907	0.947	0.637	0.973	0.997	0
b	0.912	0.093	0.505	0.912	0.650	0.637	0.973	0.818	1
Weighted Avg.	0.907	0.088	0.944	0.907	0.919	0.637	0.973	0.980	

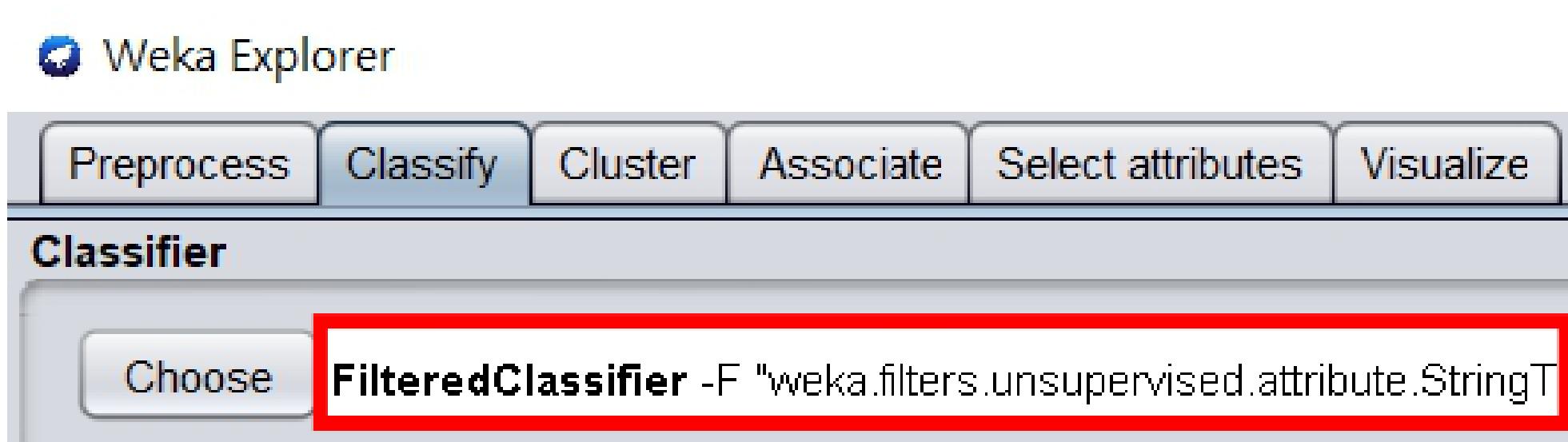
==== Confusion Matrix ====

		<-- classified as	
		a	b
a	496	51	a = 0
	5	52	b = 1

雖然沒有得到非常好的分類準確率，但是得到較好的ROC區域，穀物類準確率也較高：  
**52/57**。

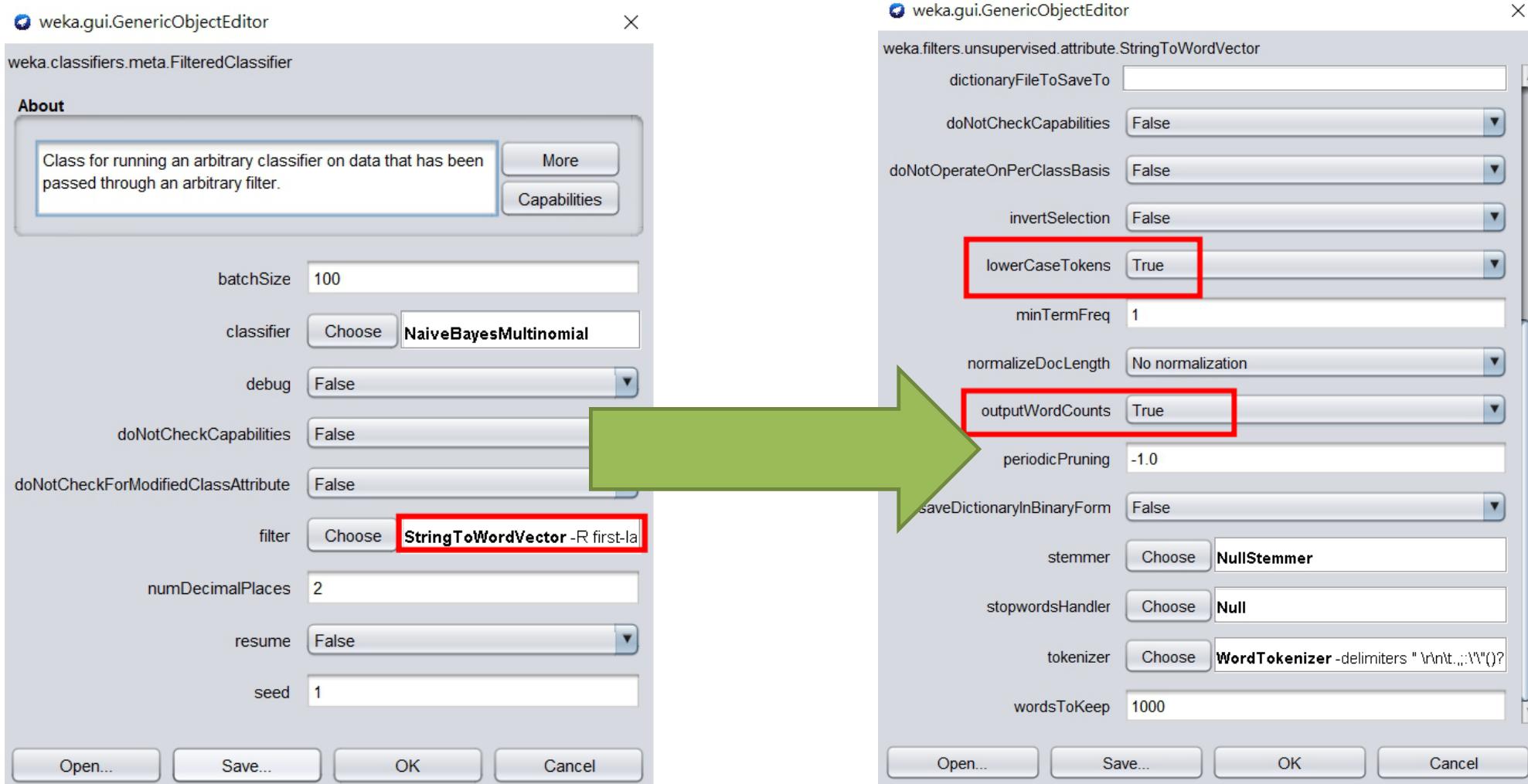
## Lesson 2.6: Multinomial Naïve Bayes

12. 左鍵單擊分類器名稱(圖中紅框處)，開啟配置視窗。



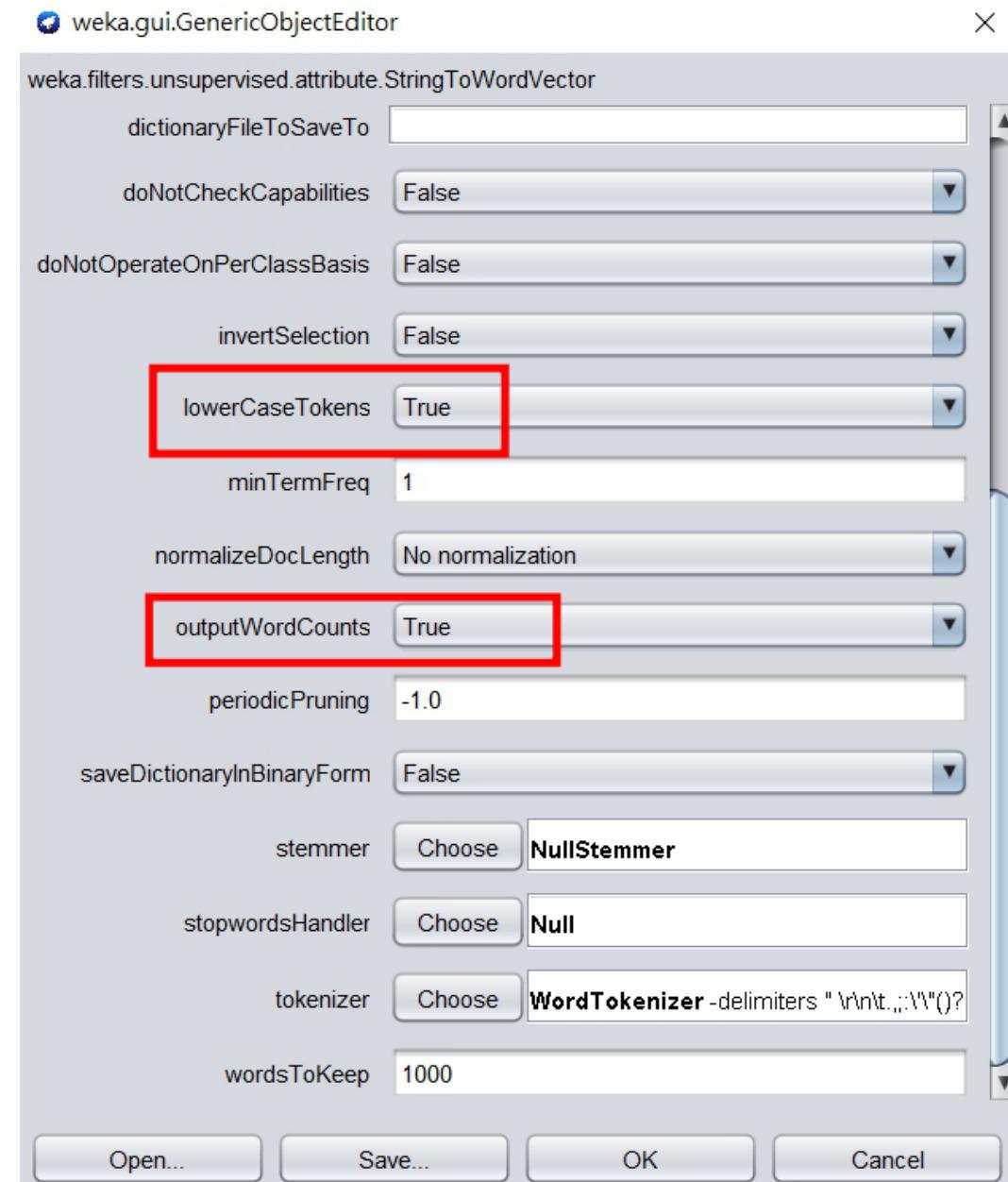
## Lesson 2.6: Multinomial Naïve Bayes

13. 在配置視窗(左圖)中左鍵單擊過濾器名稱(左圖紅框處)，開啟配置視窗(右圖)。將參數lowerCaseTokens和outputWordCounts設定為True。



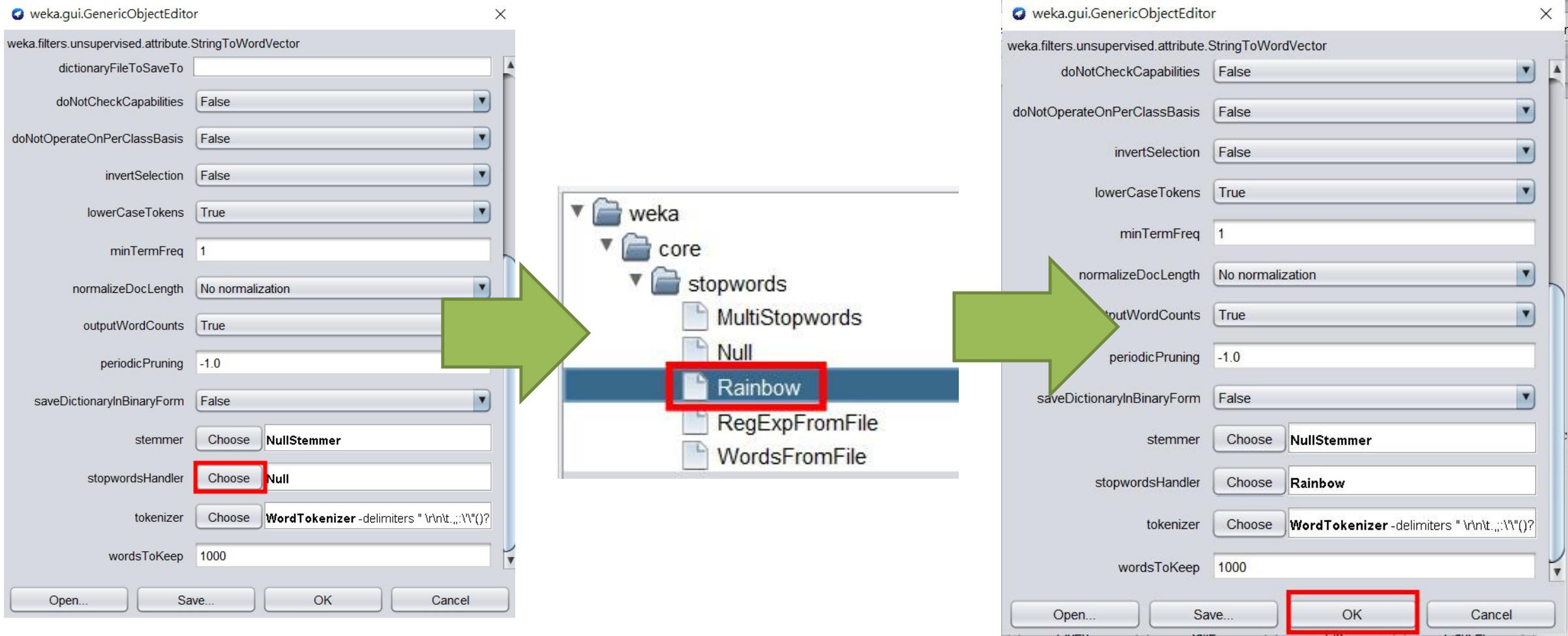
## Lesson 2.6: Multinomial Naïve Bayes

- 參數**lowerCaseTokens**可以將所有的字母都變為小寫，讓大小寫單字都會被視為同樣的詞。
- 參數**outputWordCounts**默認設置是如果文檔包含這個單詞，輸出為1，反之為0。我們將它設定為True，可以計算文檔中出現某個單詞的數量，適合Multinomial Naive Bayes。



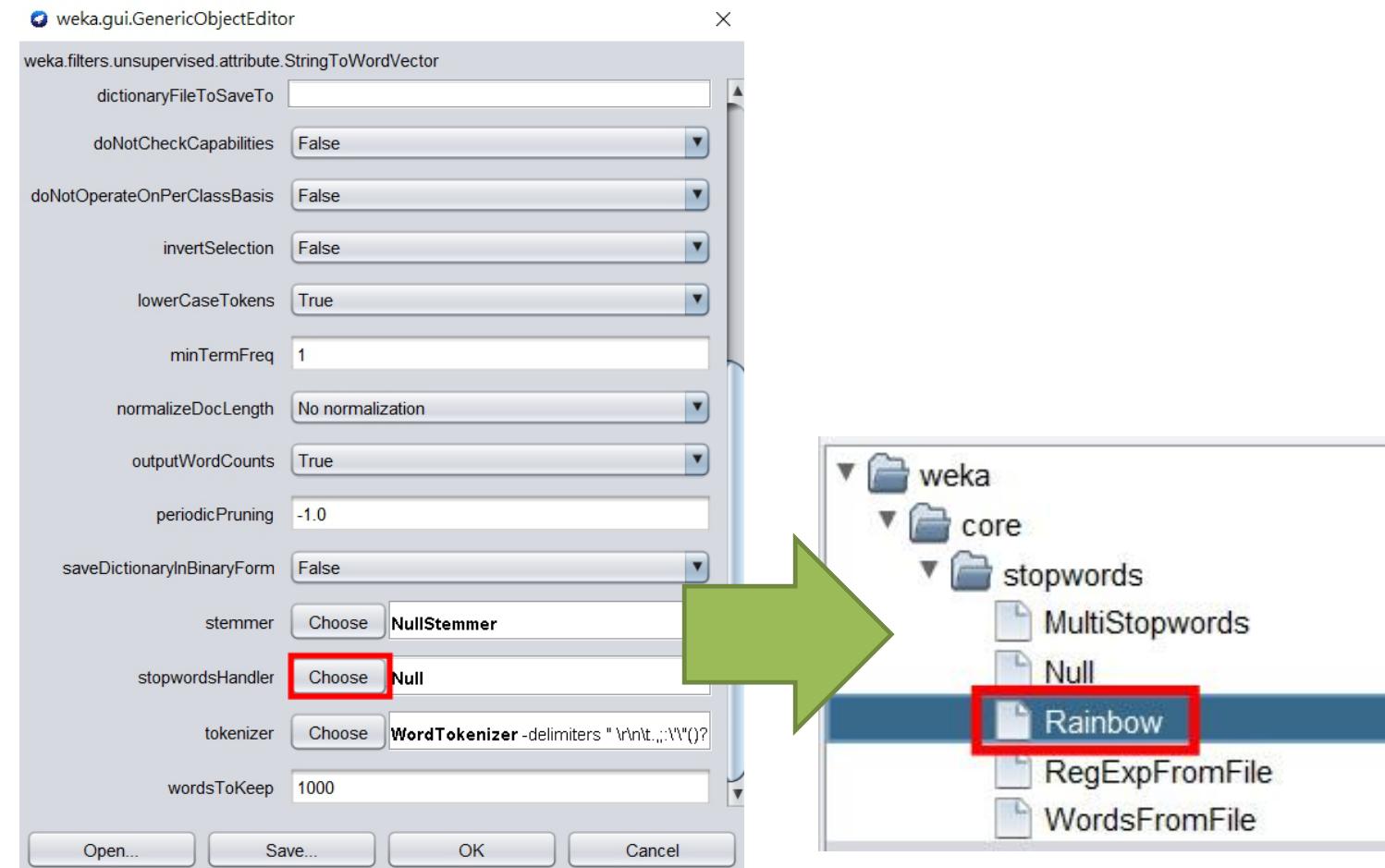
## Lesson 2.6: Multinomial Naïve Bayes

14. 接著左鍵單擊參數stopwordsHandler後方的Choose按鈕，在出現的選單中左鍵單擊Rainbow，接著左鍵單擊OK按鈕。



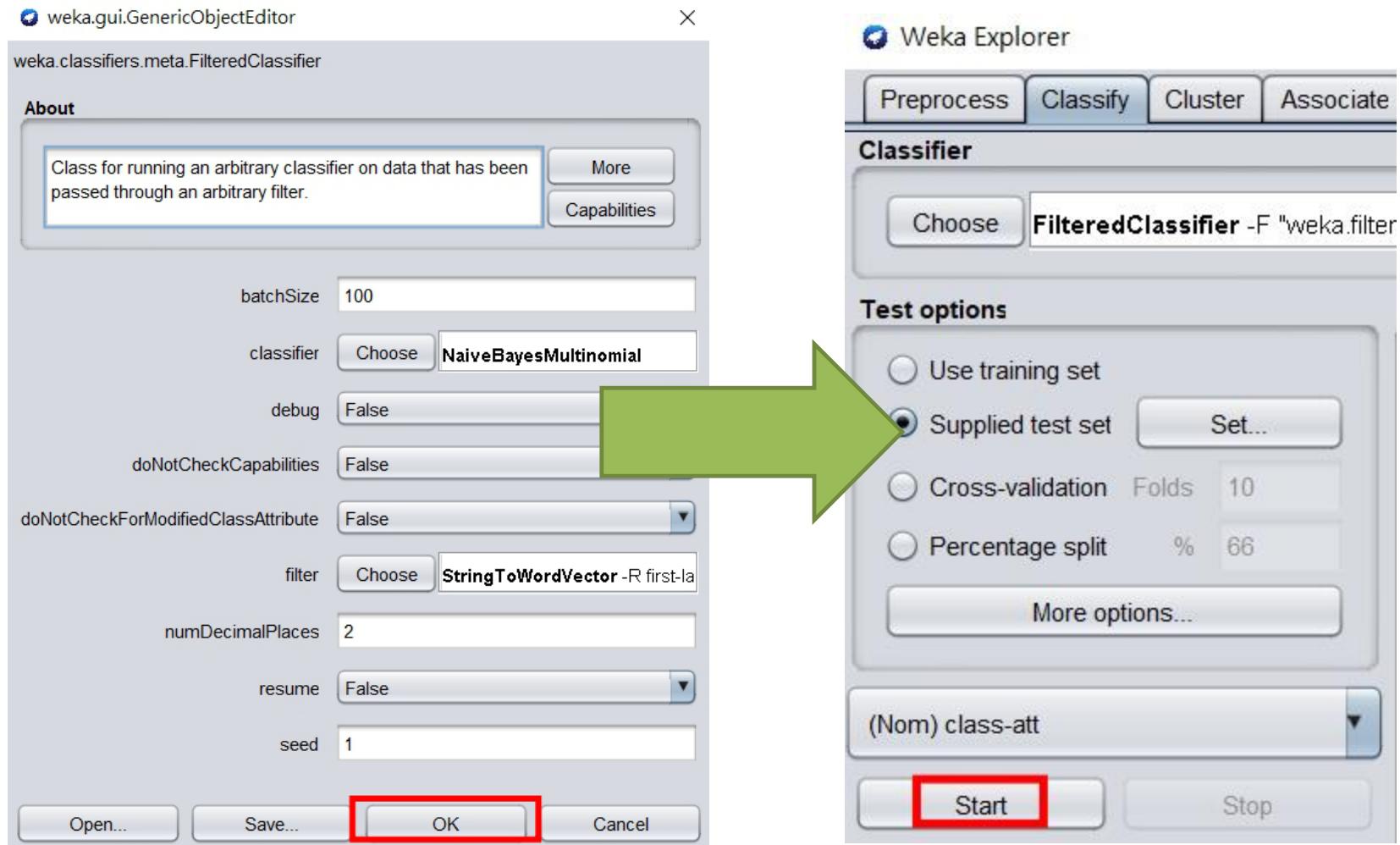
## Lesson 2.6: Multinomial Naïve Bayes

- “Stop words” 是英文中的常見詞彙，如"and"和 "the"。
- 影片中的參數"stoplist"如果設置為真，系統就會忽略Weka 的"stoplist" 中的單詞。
- 最新版本中，此功能為參數StopwordsHandler。



## Lesson 2.6: Multinomial Naïve Bayes

15. 於配置視窗中左鍵單擊OK按鈕回到Classify面板，左鍵單擊Start按鈕。



# Lesson 2.6: Multinomial Naïve Bayes

## ▼執行結果

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose FilteredClassifier -F "weka.filters.unsupervised.attribute.StringToWordVector -R first-last -W 1000 -prune-rate -1.0 -C -N 0 -L -stemmer weka.core.stemmers.NullStemmer"

Test options

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

(Nom) class-att

Start Stop

Result list (right-click for options)

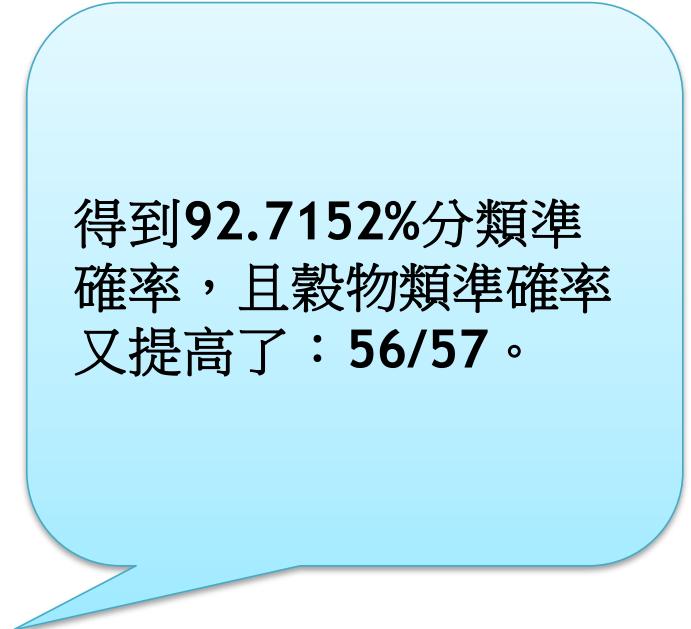
22:44:28 - meta.FilteredClassifier  
22:49:57 - meta.FilteredClassifier  
23:00:49 - meta.FilteredClassifier

Classifier output

==== Evaluation on test set ====  
Time taken to test model on supplied test set: 0.29 seconds  
==== Summary ====  
Correctly Classified Instances 560 92.7152 %  
Incorrectly Classified Instances 44 7.2848 %  
Kappa statistic 0.6796  
Mean absolute error 0.0745  
Root mean squared error 0.2638  
Relative absolute error 50.1639 %  
Root relative squared error 89.8286 %  
Total Number of Instances 604

==== Detailed Accuracy By Class ====  
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area C1  
0.921 0.018 0.998 0.921 0.958 0.714 0.978 0.998 0  
0.982 0.079 0.566 0.982 0.718 0.714 0.976 0.782 1  
Weighted Avg. 0.927 0.023 0.957 0.927 0.936 0.714 0.978 0.977

==== Confusion Matrix ====  
a b <- classified as  
504 43 | a = 0  
1 56 | b = 1



得到92.7152%分類準確率，且穀物類準確率又提高了：56/57。

## Lesson 2.6: Multinomial Naïve Bayes

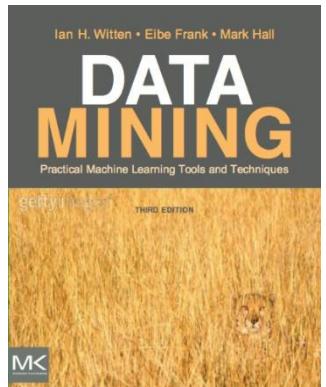
- ❖ 訓練集: `ReutersGrain-train.arff`; 測試集: `ReutersGrain-test.arff`
- ❖ 分類器: 搭配`StringToWordVector`的`FilteredClassifier`
- ❖ **J48** 得到 96% 分類器準確率
  - 穀物相關準確率: 38/57, 其他準確率: 544/547;  $ROC\ Area = 0.906$
- ❖ **NaiveBayes**: 80% 分類器準確率
  - 穀物相關準確率: 46/57, 其他準確率: 439/547;  $ROC\ Area = 0.885$
- ❖ **NaiveBayesMultinomial**: 91% 分類器準確率
  - 穀物相關準確率: 52/57, 其他準確率: 496/547;  $ROC\ Area = 0.973$
- ❖ 在`StringToWordVector`中設定`outputWordCounts`  
**NaiveBayesMultinomial**: 91% 分類器準確率
  - 穀物相關準確率: 54/57, 其他準確率: 496/547;  $ROC\ Area = 0.962$
- ❖ Set `lowerCaseTokens`, `useStoplist` in `StringToWordVector`  
**NaiveBayesMultinomial**: 93% 分類器準確率
  - 穀物相關準確率: 56/57, 其他準確率: 504/547;  $ROC\ Area = 0.978$

## Lesson 2.6: Multinomial Naïve Bayes

- ❖ Multinomial Naïve是專門為文檔設計的機器學習方法
  - 採用出現的單詞，而不是未出現的單詞
  - 可以檢索單詞在文檔中出現的頻率
  - 區別對待常見的和非常見的
- ❖ Multinomial Naïve比Naïve Bayes快上許多!
  - Multinomial Naïve忽略未出現在文檔中的詞
  - Weka在內部使用資料的稀疏表示
- ❖ StringToWordVector 過濾器有許多有趣的選項
  - 儘管他們不一定會給出你想要的結果!
  - 以「稀疏資料」格式輸出結果，MNB利用了這種格式

### 課程文本

- ❖ Section 4.2 *Statistical modeling*, under “Naïve Bayes for document classification”





THE UNIVERSITY OF  
**WAIKATO**  
*Te Whare Wānanga o Waikato*

# *More Data Mining with Weka*

Department of Computer Science  
University of Waikato  
New Zealand



Creative Commons Attribution 3.0 Unported License



[creativecommons.org/licenses/by/3.0/](http://creativecommons.org/licenses/by/3.0/)

[weka.waikato.ac.nz](http://weka.waikato.ac.nz)