



THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

More Data Mining with Weka

Class 3 – Lesson 1

決策樹與規則

(Decision trees and rules)

Ian H. Witten

Department of Computer Science
University of Waikato
New Zealand

Lesson 3.1: 決策樹與規則

Class 1 探索Weka的介面；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 3.1 決策樹與規則

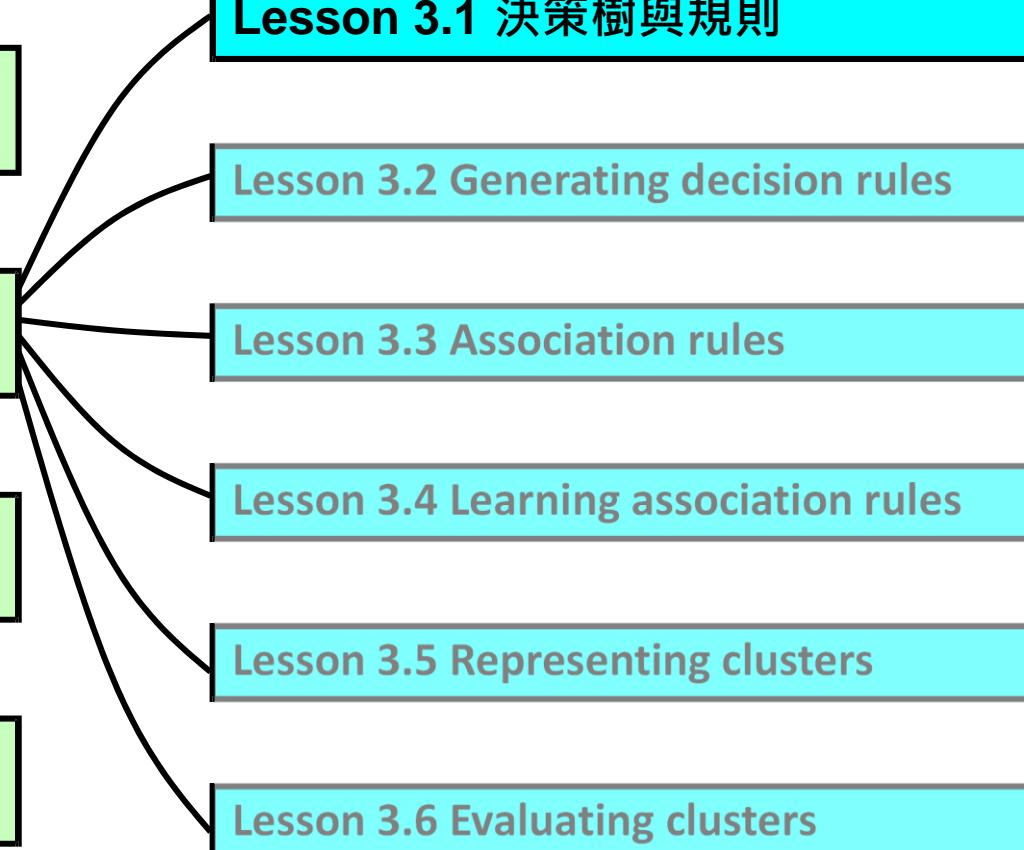
Lesson 3.2 Generating decision rules

Lesson 3.3 Association rules

Lesson 3.4 Learning association rules

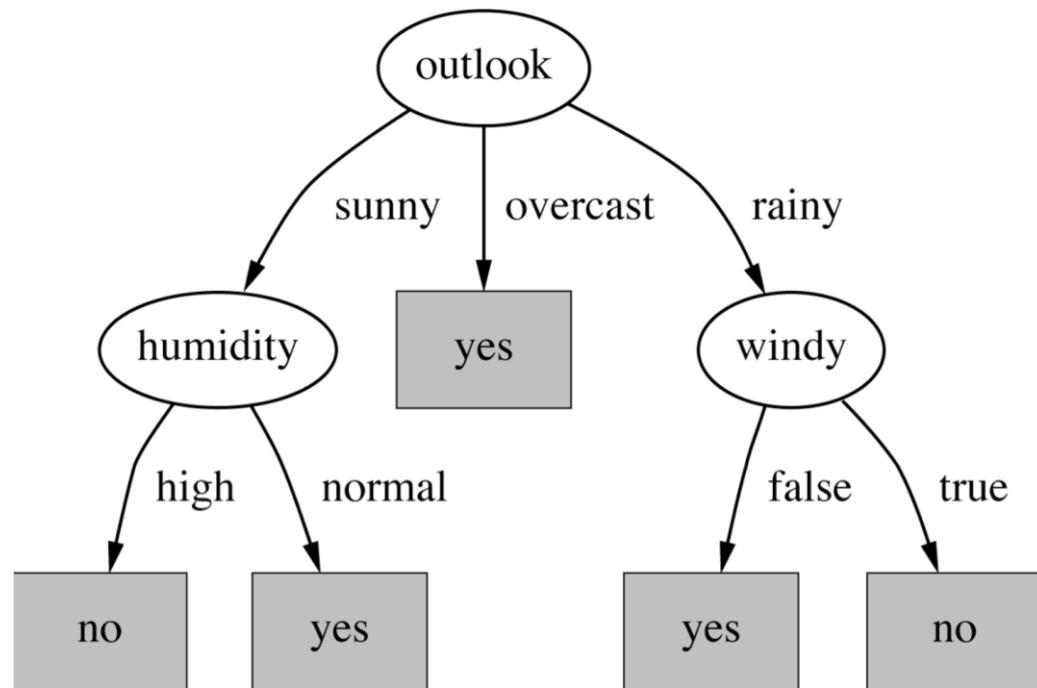
Lesson 3.5 Representing clusters

Lesson 3.6 Evaluating clusters



Lesson 3.1: 決策樹與規則

對於任何決策樹，我們都可以得到一組等效規則。



If $\text{outlook} = \text{sunny}$ 且 $\text{humidity} = \text{high}$ 則為 no
If $\text{outlook} = \text{sunny}$ 且 $\text{humidity} = \text{normal}$ 則為 yes

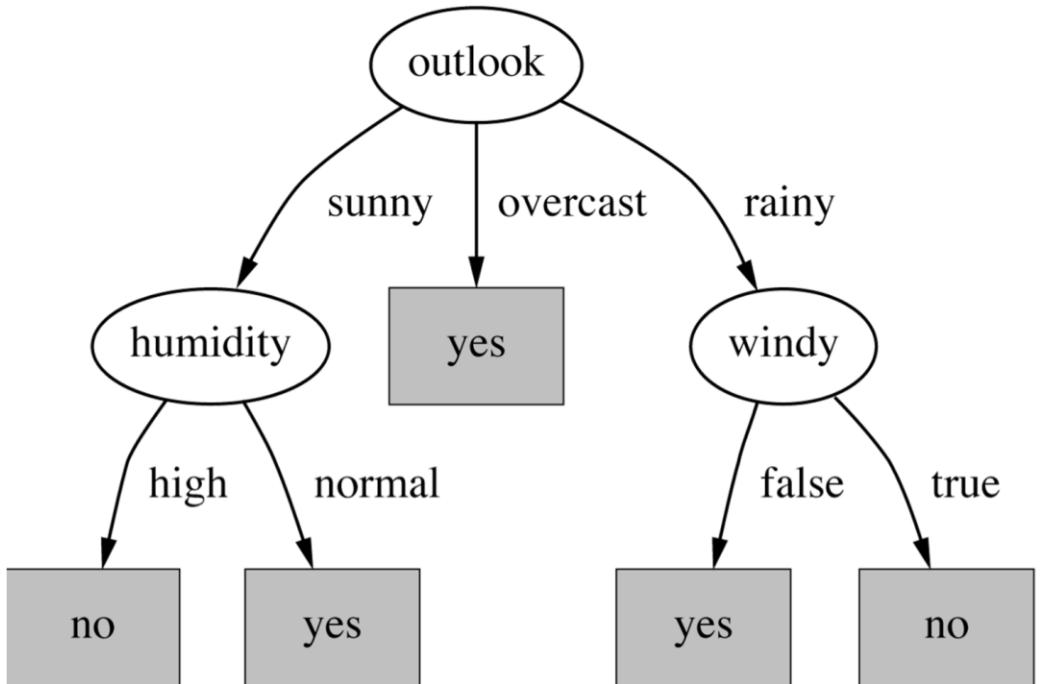
If $\text{outlook} = \text{overcast}$ 則為 yes

if $\text{outlook} = \text{rainy}$ 且 $\text{windy} = \text{false}$ 則為 yes

if $\text{outlook} = \text{rainy}$ 且 $\text{windy} = \text{true}$ 則為 no

Lesson 3.1: 決策樹與規則

對於任何決策樹，我們都可以得到一組等效依序規則(決策表)。



If ~~outlook = sunny 且 humidity = high 則為 no~~
If ~~outlook = sunny 且 humidity = normal 則為 yes~~ if ~~outlook = overcast 則為 yes~~
if ~~outlook = rainy 且 windy = false 則為 yes~~ if ~~outlook = rainy 且 windy = true 則為 no~~

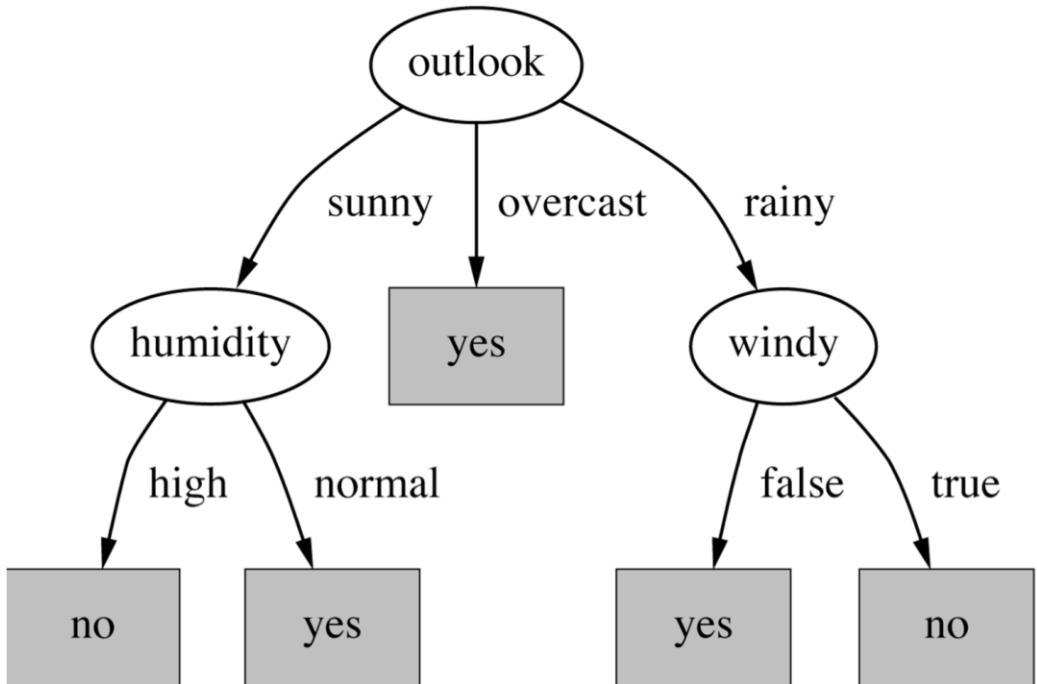
我們可以刪除第二條規則中的“濕度正常”，因為我們已經在第一條規則中處理了“濕度大”的情況，且這裡沒有其他的選擇。如果濕度不大，就會是正常。

同理，在第四條規則中，我們可以刪除“雨天”，因為我們已經處理了晴天和陰天，除去這些條件，天氣就只有雨天。

使用這樣的技巧，如果我們將規則做成決策表會更為簡單。

Lesson 3.1: 決策樹與規則

對於任何決策樹，我們都可以得到一組等效依序規則(決策表)。



If ~~outlook = sunny 且 humidity = high 則為 no~~
If ~~outlook = sunny 且 humidity = normal 則為 yes~~
if ~~outlook = overcast 則為 yes~~
if ~~outlook = rainy 且 windy = false 則為 yes~~
if ~~outlook = rainy 且 windy = true 則為 no~~

這裡是更簡單的規則集：

If ~~outlook = sunny 且 humidity = high 則為 no~~
if ~~outlook = rainy 且 windy = true 則為 no~~
否則為 yes

Lesson 3.1: 決策樹與規則

對於任何規則集可以得到一個同等的樹

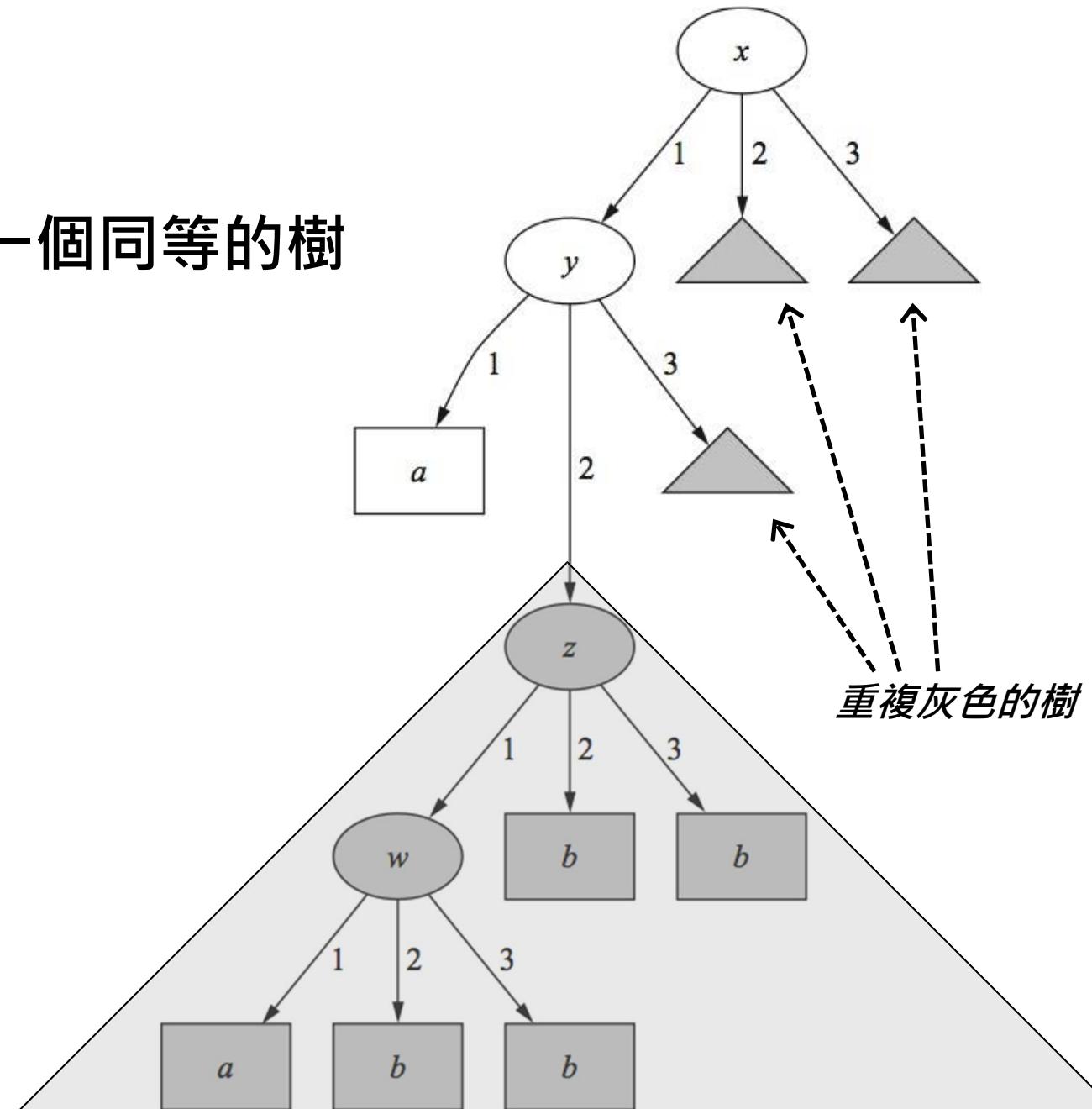
但是它可能非常複雜

```
if x = 1 且 y = 1 則 a  
if z = 1 且 w = 1 則 a  
否則為 b
```

假設x和y都有三個可能的數值，我們用x和y從左側分支，如果它們的值都是1，就是a。

如果y=2，我們就要看z和w的值。也就是下面這個灰色的樹，有點複雜。更麻煩的是，我們要重複這個決策樹。

如果y=3，我們也要重複這個灰色決策樹。



Lesson 3.1: 決策樹與規則

- ❖ 理論上，樹和規則具有同等的描述能力
- ❖ 但實際上他們非常不同
 - ... 因為規則通常用於表述一個決策表，若依序執行，會比決策樹小得多

- ❖ 人們喜歡規則：因為規則易於閱讀和理解
- ❖ 規則被視為獨立的「知識塊」
- ❖ ... 但那是誤導，規則並不真正相互獨立
 - 如果制成決策表，那麼對規則的解釋必須考慮之前的規則。

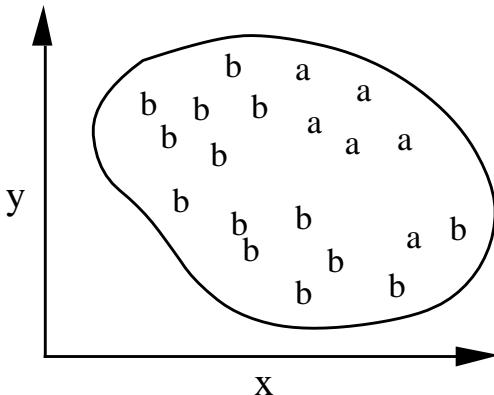
Lesson 3.1: 決策樹與規則

- ❖ 創建決策樹(由上而下，分而治之(**divide-and-conquer**)); 從決策樹讀取規則
 - 每個葉節點是一個規則
 - 非常直接，但是規則會包括重複的測試
 - 我們可以很容易刪除那些重複的測試，但是更為有效的變換卻不容易
- ❖ 替代方法: 覆蓋算法 (由下而上,割治算法(**separate-and-conquer**))
 - 針對每個類別，我們尋找能覆蓋所有實例的規則
 1. 找到一個有用的規則
 2. 用它劃分實例
 3. 繼續尋找能覆蓋同一類別中剩餘實例的規則

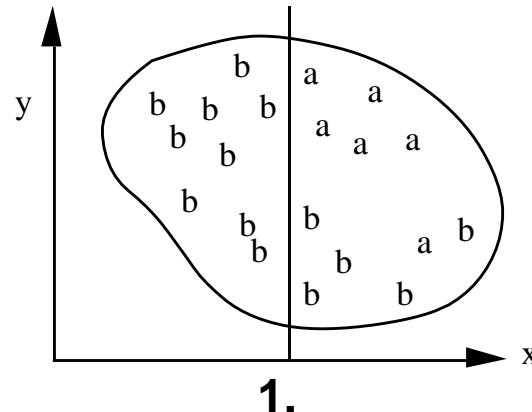
Lesson 3.1: 決策樹與規則

產生規則

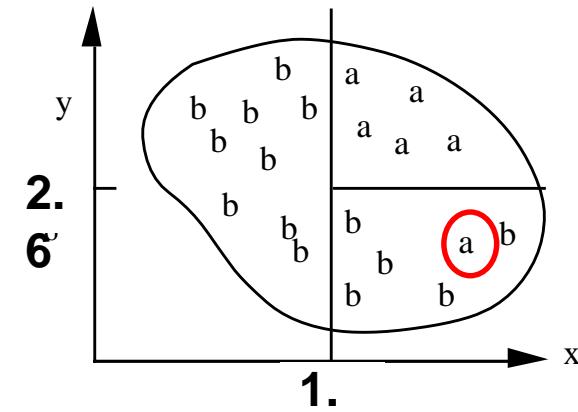
- ❖ 為類別a產生規則



所有實例都屬於類別a



1. 如果 $x > 1.2$, 類別為a



2. 如果 $x > 1.2$ 且 $y > 2.6$, 類別為a

- ❖ 為類別b可能產生的規則：

if $x \leq 1.2$ 則類別為b

if $x > 1.2$ 且 $y \leq 2.6$ 類別為b

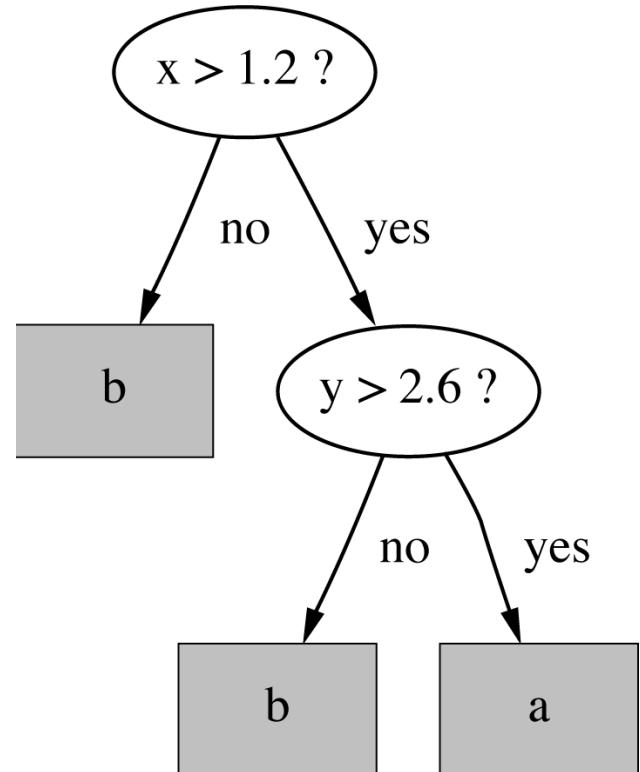
- ❖ 可以增加更多規則，找到最完美的規則集

我們可以為這個實例增加一條新的規則，或者我們可以選擇不增加，因為只漏掉了一個實例。

Lesson 3.1: 決策樹與規則

規則 vs. 決策樹

- ❖ 對應決策樹
 - 產生完全相同的預測
- ❖ 規則集可以更明確
 - E.g. 當決策樹包含重複的子樹的時候
- ❖ 還有：針對多類別的情況，
 - 覆蓋算法一次側重於一個類別
 - 決策樹算法做決策時考慮所有類別



Lesson 3.1: 決策樹與規則

簡單的由下而上的覆蓋算法制定規則的詳細步驟: PRISM

對於每個類別 C

 初始化 E 到 實例集

 當 E 包含 類別 C 中的 實例

 創建 一個 預測 類別 C 的 規則 R

 (左邊為空)

 直到 規則 R 達到 完美

 (或是 沒有 屬性 可以 使用 的 時候)

 對於 每個 未在 規則 R 中 提到 的 屬性 A 以及 數值 v

 考慮 增加 條件 $A = v$ 到 規則 R 的 左邊

 選擇 A 和 v 來 最大化 準確率

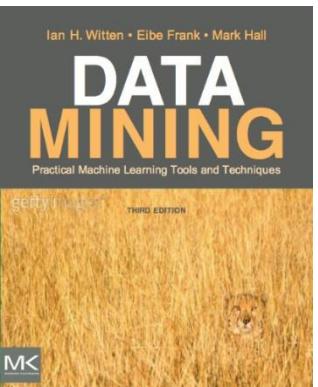
 (藉由 選擇 p 值 最大的 條件)

 增加 $A = v$ 到 規則 R

 從 E 中 刪除 R 覆蓋 的 實例

Lesson 3.1: 決策樹與規則

- ❖ 決策樹與規則具有相同表現能力
 - ...兩者中任何一個都可能更明確，這取決於數據集
- ❖ 規則可以通過自下而上的覆蓋過程來創建
- ❖ 規則通常作為決策表，依次執行
 - 如果多個規則給一個實例分配了不同的類別，依次執行時，第一條規則會起決定作用
 - 這些規則不是獨立的「知識塊」
- ❖ 和決策樹相比，人們更喜歡規則



Course text

- ❖ Section 4.4 *Covering algorithms: constructing rules*



THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

More Data Mining with Weka

Class 3 – Lesson 2

產生決策規則(*Generating decision rules*)

Ian H. Witten

Department of Computer Science
University of Waikato
New Zealand

weka.waikato.ac.nz

Lesson 3.2: 產生決策規則

Class 1 探索Weka的介面；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 3.1 決策樹與規則

Lesson 3.2 產生決策規則

Lesson 3.3 Association rules

Lesson 3.4 Learning association rules

Lesson 3.5 Representing clusters

Lesson 3.6 Evaluating clusters



Lesson 3.2: 產生決策規則

1. 通過部分決策樹創建規則: PART

- ❖ 創建規則
- ❖ 去除規則所覆蓋的實例
- ❖ 繼續為剩餘的實例創建規則

} 割治算法
(Separate and conquer)

為了創建規則而創造一棵樹!

- ❖ 為實例集創建和修建決策樹
- ❖ 從大的葉節點的讀取規則
- ❖ 不再參考決策樹(!)繼續執行覆蓋算法

(可以只創建部分決策樹，而不必創建完整的決策樹)

Lesson 3.2: 產生決策規則

2.遞增減少誤差修剪算法：

RIPPER 訓練集實例，以2:1的比例分成兩個數據集，成長集和修剪集

對於每個類別C

當成長集和修剪集都包含類別C中的實例

在成長集上使用PRISM來為C創建完美的規則

為修剪集上的規則計算價值 $w(R)$,

省略最後條件的規則 $w(R-)$

當 $w(R-) > w(R)$, 從規則中刪除最後一個條件，並重複前面的步驟

印出結果;從成長集和修剪集中刪除它覆蓋的實例

“價值(worth)”:

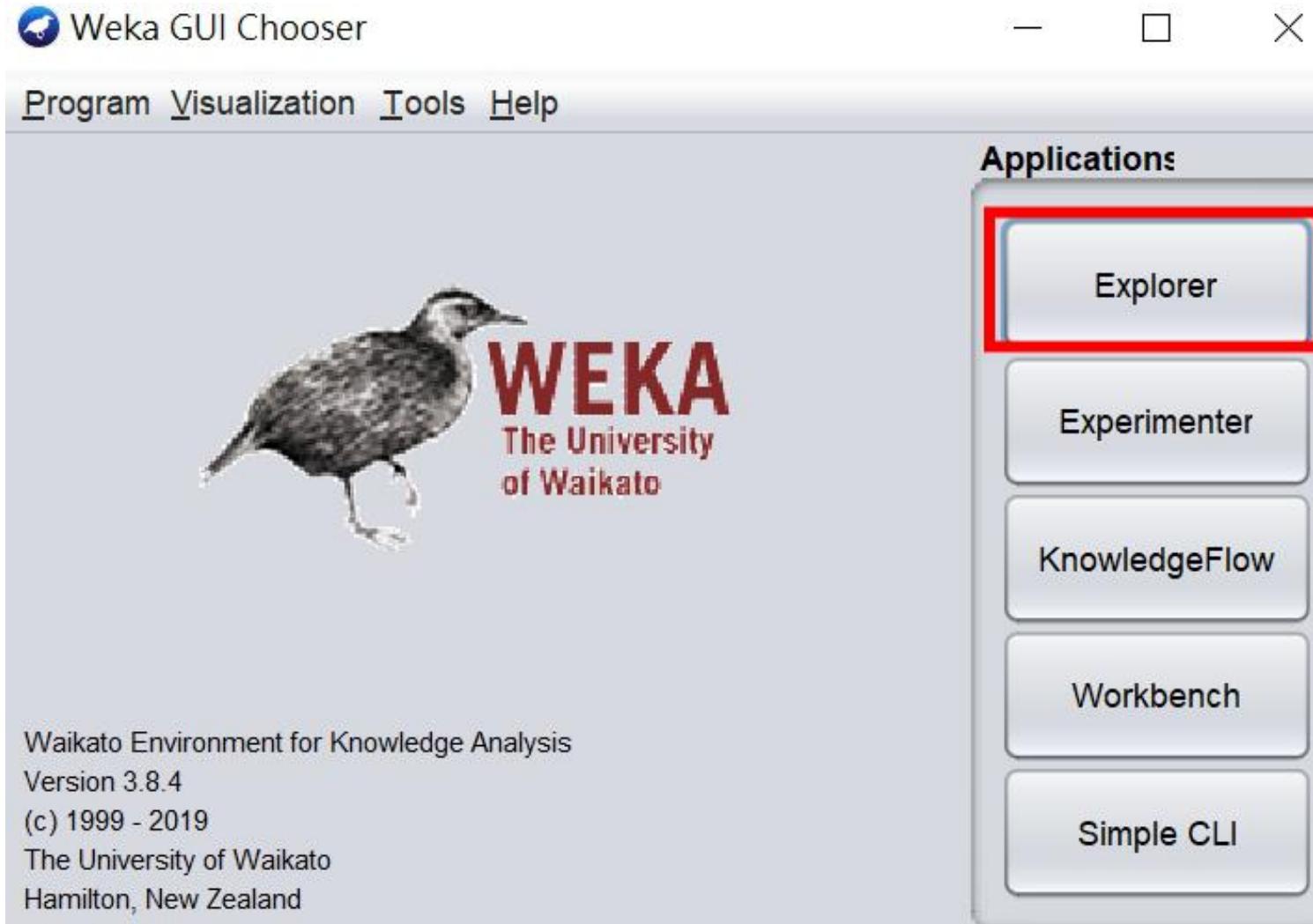
成功率?

某樣更複雜的東西?

...然後是極其複雜的全局優化步驟— RIPPER

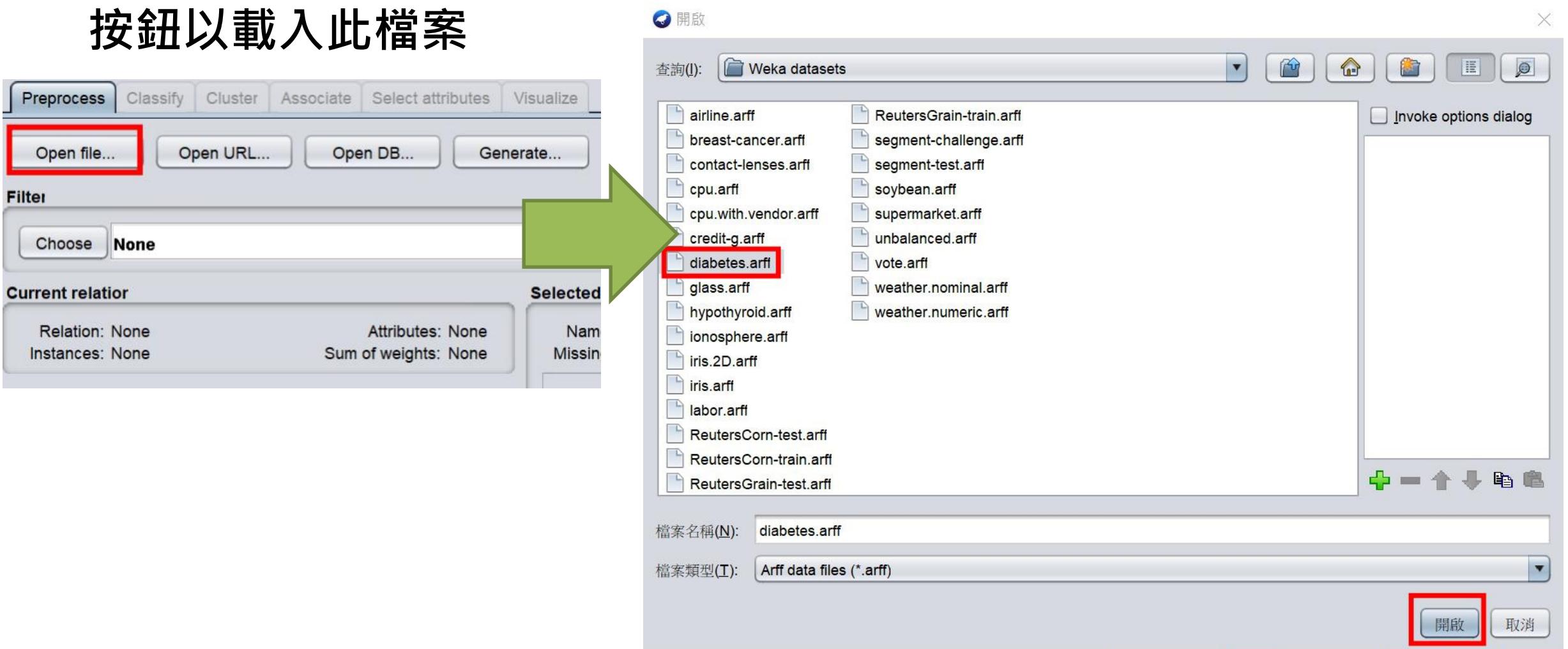
Lesson 3.2: 產生決策規則

1. 開啟Weka的Explorer



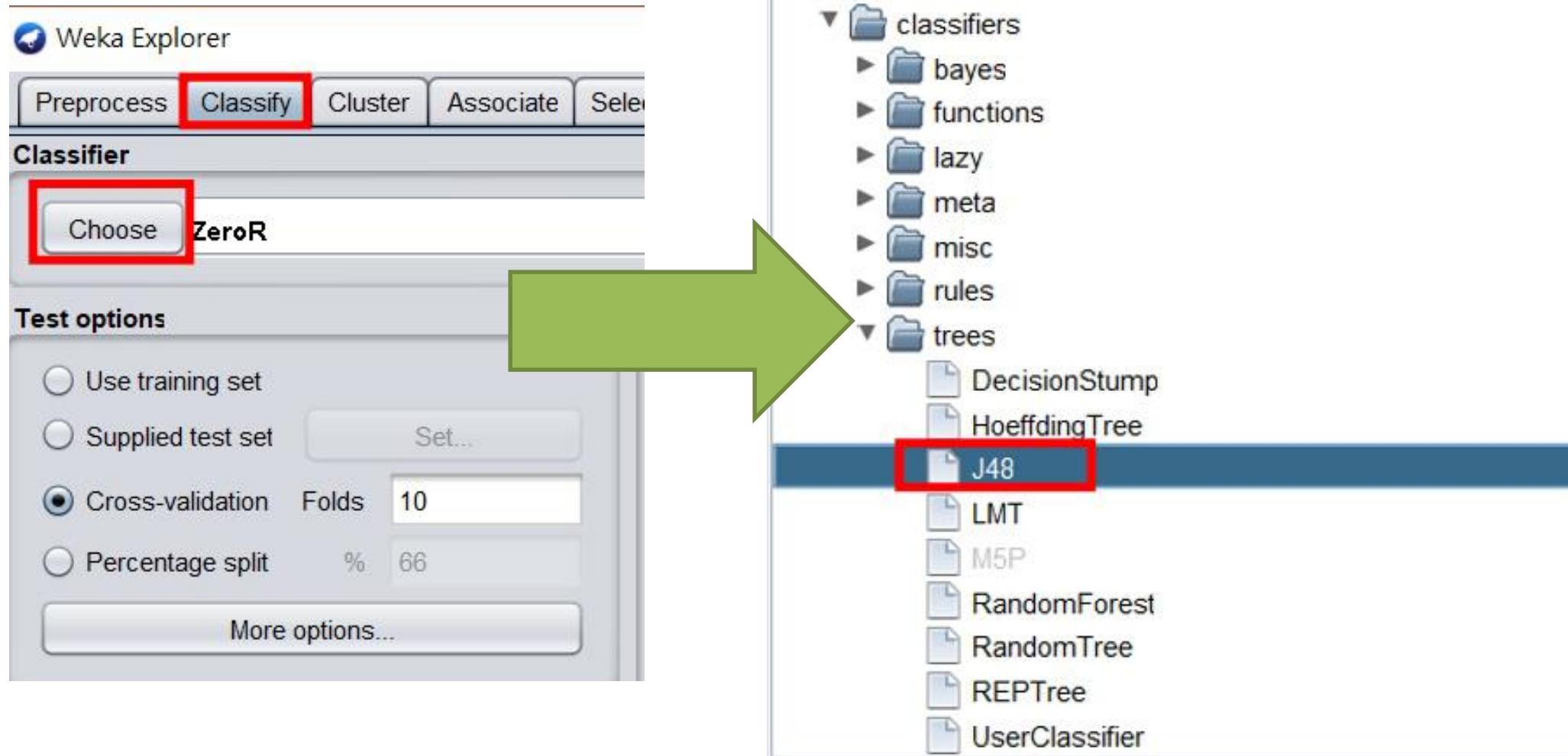
Lesson 3.2: 產生決策規則

2. 左鍵點擊Open file...開啟右圖視窗，進入自行複製的Weka datasets資料夾，左鍵單擊**diabetes.arff** 檔案後，再以左鍵單擊下方「開啟」按鈕以載入此檔案



Lesson 3.2: 產生決策規則

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 3.2: 產生決策規則

4. 回到Classify面板，左鍵單擊Start按鈕。

The image shows the Weka interface with the 'Classify' tab selected. On the left, the 'Test options' panel is visible, showing 'Cross-validation' selected with 'Folds' set to 10. A large green arrow points from this panel to the 'Start' button in the 'Classifier output' panel on the right. The 'Classifier output' panel displays the generated decision tree rules and performance metrics.

Classifier output:

```
| | | | | | | pedi > 0.396: tested_negative (3.0)
| | | | | | pres > 82: tested_negative (4.0)
| | | | age > 61: tested_negative (4.0)
| mass > 29.9
| | plas <= 157
| | | pres <= 61: tested_positive (15.0/1.0)
| | | pres > 61
| | | | age <= 30: tested_negative (40.0/13.0)
| | | | age > 30: tested_positive (60.0/17.0)
| | plas > 157: tested_positive (92.0/12.0)
```

Number of Leaves : 20
Size of the tree : 39

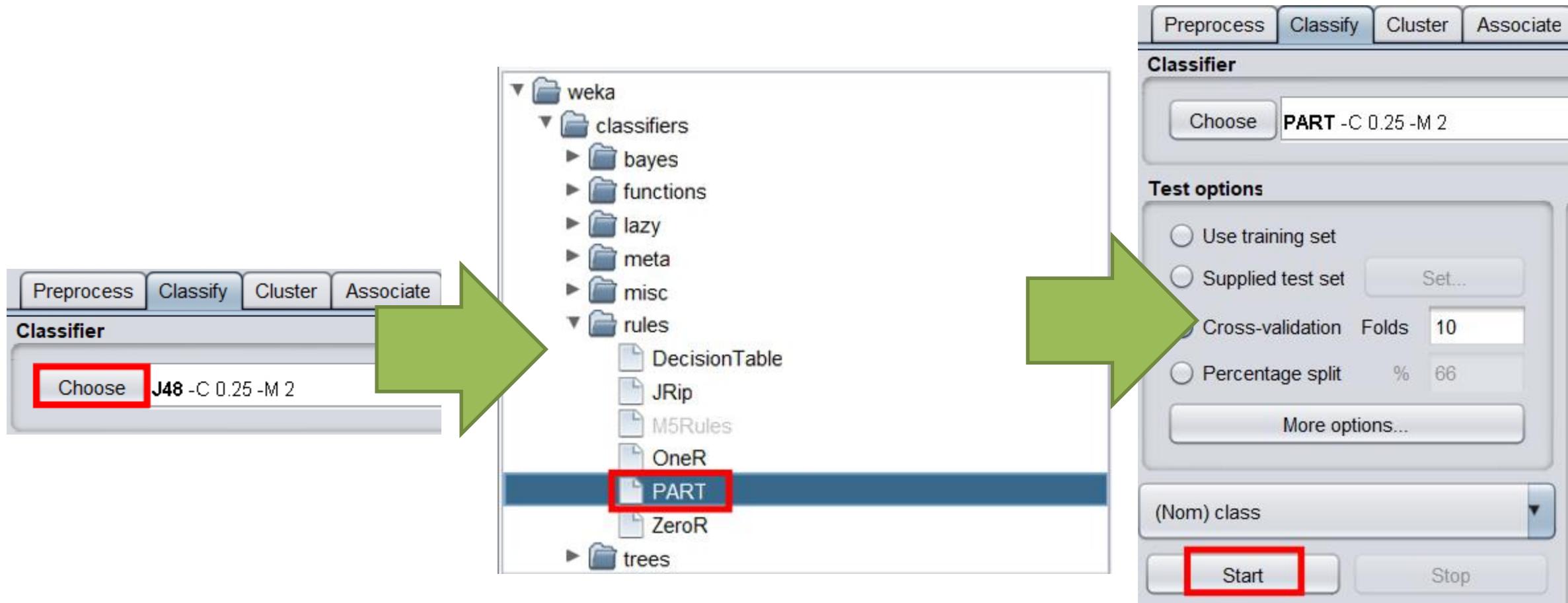
Time taken to build model: 0.31 seconds

==== Stratified cross-validation ====
==== Summary ====
Correctly Classified Instances 567
Incorrectly Classified Instances 201
Kappa statistic 0.4164
Mean absolute error 0.3158
Root mean squared error 0.4463
Relative absolute error 69.4841 %
Root relative squared error 93.6293 %
Total Number of Instances 768

有20個葉節點和39個節點。準確率是74%。

Lesson 3.2: 產生決策規則

5. 於Classify面板左鍵單擊Choose按鈕，左鍵單擊出現的選單中weka/classifiers/rules路徑下的PART分類器，再以左鍵單擊Start按鈕。



Lesson 3.2: 產生決策規則

▼ 執行結果

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose PART -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)

- 09:29:55 - trees.J48
- 09:38:13 - rules.PART

Classifier output

```
PART decision list
-----
plas <= 127 AND
mass <= 26.4 AND
preg <= 7: tested_negative (117.0/1.0)

plas > 154 AND
mass > 29.8: tested_positive (100.0/14.0)

plas <= 99 AND
age <= 25 AND
age <= 22: tested_negative (33.0)

age <= 28 AND
skin > 0 AND
skin <= 34 AND
age > 22 AND
preg <= 3 AND
plas <= 127: tested_negative (61.0/7.0)

plas <= 99 AND
insu <= 88 AND
insu <= 18 AND
skin <= 21: tested_negative (26.0/1.0)

age <= 24 AND
skin > 0 AND
mass <= 33.3: tested_negative (37.0)
```

Lesson 3.2: 產生決策規則

▼ 執行結果

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose PART -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)

09:29:55 - trees.J48
09:38:13 - rules.PART

Classifier output

```
age > 61 AND
preg > 4: tested_negative (11.0)

age <= 30 AND
pres > 72 AND
mass <= 42.8: tested_negative (41.0/7.0)

plas <= 89 AND
plas > 0: tested_negative (13.0/1.0)

: tested_positive (252.0/105.0)
```

Number of Rules : 13

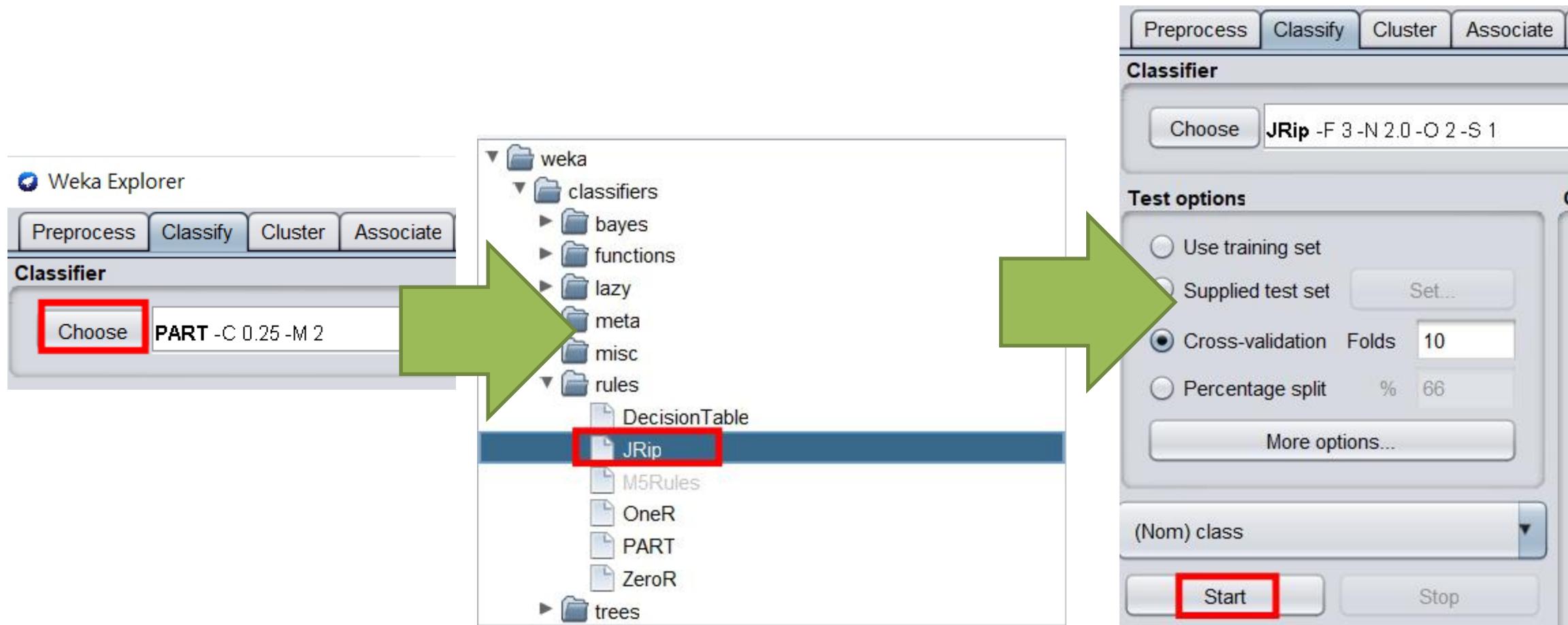
Time taken to build model: 0.07 seconds

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

Correctly Classified Instances 578 75.2604 %
Incorrectly Classified Instances 190 24.7396 %
Kappa statistic 0.439
Mean absolute error 0.3101
Root mean squared error 0.4149
Relative absolute error 68.224 %
Root relative squared error 87.0418 %
Total Number of Instances 768

Lesson 3.2: 產生決策規則

6. 於Classify面板左鍵單擊Choose按鈕，左鍵單擊出現的選單中weka/classifiers/rules路徑下的JRip分類器，再以左鍵單擊Start按鈕。



Lesson 3.2: 產生決策規則

▼ 執行結果

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose JRip -F 3-N 2.0-O 2-S 1

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)

- 09:29:55 - trees.J48
- 09:38:13 - rules.PART
- 09:40:48 - rules.JRip**

Classifier output

```
==== Classifier model (full training set) ====
JRIP rules:
=====
(plas >= 132) and (mass >= 30) => class=tested_positive (182.0/48.0)
(age >= 29) and (insu >= 125) and (preg <= 3) => class=tested_positive (19.0/4.0)
(age >= 31) and (pedi >= 0.529) and (preg >= 8) and (mass >= 25.9) => class=tested_positive (22.0/10.0)
=> class=tested_negative (545.0/102.0)

Number of Rules : 4

Time taken to build model: 0.14 seconds

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

Correctly Classified Instances      584          76.0417 %
Incorrectly Classified Instances   184          23.9583 %
Kappa statistic                      0.4538
Mean absolute error                  0.3419
Root mean squared error              0.4239
Relative absolute error              75.2322 %
Root relative squared error         88.933 %
Total Number of Instances           768

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

```

Lesson 3.2: 產生決策規則

Diabetes 資料集

- ❖ J48 74% 39個節點的樹
- ❖ PART 73% 13 個規則(25次測試)
- ❖ JRip 76% 4 個規則(9 次測試)

```
plas ≥ 132 and mass ≥ 30 -> tested_positive  
age ≥ 29 and insu ≥ 125 and preg ≤ 3 -> tested_positive  
age ≥ 31 and pedi ≥ 0.529 and preg ≥ 8 and mass ≥ 25.9 ->  
tested_positive  
-> tested_negative
```

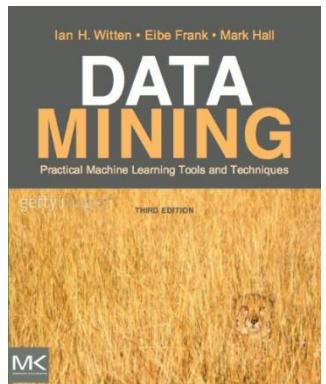
這是JRip的4條規則，RIPPER首先確定了大多數實例的類別(隱性結果)並把它放在一邊。RIPPER只為其他類別創建規則，而把大多數實例的類別作為默認條件。
測試結果為顯性是小類別，而測試結果為隱性是大類別。
僅有4條規則，9次測試，就得到最好的結果。

Lesson 3.2: 產生決策規則

- ❖ PART 快速且細緻
 - 重複創建決策樹，然後刪除，而且這並不像聽上去那麼浪費
- ❖ 遞增減少誤差修剪算法是標準算法
 - 使用成長集(Grow)和修剪集(Prune)
- ❖ RIPPER (JRip)使用遞增減少誤差修剪算法並全局優化
 - 制定規則來分類除了大多數類值之外的所有類值
 - 對大多數類別而言，上一規則為預設規則
 - 通常創建比PART還要少的規則

Course text

- ❖ Section 6.2 *Classification rules*





THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

More Data Mining with Weka

Class 3 – Lesson 3

關聯規則 (Association rules)

Ian H. Witten

Department of Computer
Science University of Waikato
New Zealand

weka.waikato.ac.nz

Lesson 3.3: 關聯規則

Class 1 探索Weka的介面；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 3.1 決策樹與規則

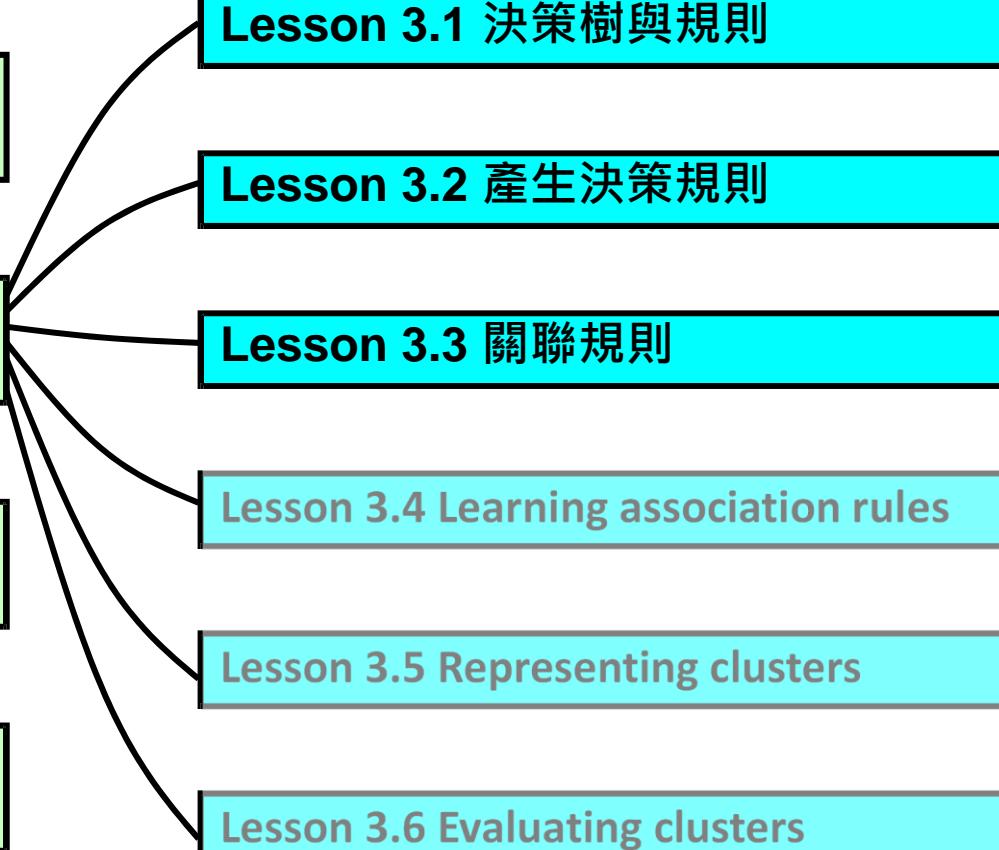
Lesson 3.2 產生決策規則

Lesson 3.3 關聯規則

Lesson 3.4 Learning association rules

Lesson 3.5 Representing clusters

Lesson 3.6 Evaluating clusters



Lesson 3.3: 關聯規則

- ❖ 關聯規則就是找出屬性之間的聯繫，關聯規則中沒有特殊的類別的屬性
- ❖ 規則可以預測任何屬性，或屬性的組合
- ❖ 需要一個不同種類的算法：“**Apriori**”

這裡有一些weather資料的關聯規則：

1. outlook = overcast ==> play = yes
2. temperature = cool ==> humidity = normal
3. humidity = normal & windy = false ==> play = yes
4. outlook = sunny & play = no ==> humidity = high
5. outlook = sunny & humidity = high ==> play = no
6. outlook = rainy & play = yes ==> windy = false
7. outlook = rainy & windy = false ==> play = yes
8. temperature = cool & play = yes ==> humidity = normal
9. outlook = sunny & temperature = hot ==> humidity = high
10. temperature = hot & play = no ==> outlook = sunny

Outlook	Temperature	Humidity	Windy	Play
k	p	y		
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes

Lesson 3.3: 關聯規則

- ❖ **支持度(Support):** 滿足規則的實例數量
- ❖ **置信度(Confidence):** 結論成立的實例的比例
- ❖ 設定最小置信度，尋找最小置信度下最大的支持度

			suppo	confiden
			rt	ce
1.	outlook = overcast	==> play = yes	4	100%
2.	temperature = cool	==> humidity = normal	4	100%
3.	humidity = normal & windy = false	==> play = yes	4	100%
4.	outlook = sunny & play = no	==> humidity = high	3	100%
5.	outlook = sunny & humidity = high	==> play = no	3	100%
6.	outlook = rainy & play = yes	==> windy = false	3	100%
7.	outlook = rainy & windy = false	==> play = yes	3	100%
8.	temperature = cool & play = yes	==> humidity = normal	3	100%
9.	outlook = sunny & temperature = hot	==> humidity = high	2	100%
10.	temperature = hot & play = no	==> outlook = sunny	2	100%

Lesson 3.3: 關聯規則

- ❖ 項目集(Itemset) 屬性和值配對的組合, e.g.

humidity = normal & windy = false & play =

支持度 = 4

- ❖ 這個項目集當中的7中可能的規則:

		支持度	置信度
If humidity = normal & windy = false	==>	4	4/4
play = yes	If humidity = normal & play = yes ==>	4	4/6
	windy = false		
If windy = false & play = yes	==> humidity = normal	4	4/6
If humidity = normal	==> windy = false & play = yes	4	4/7
If windy = false	==> humidity = normal & play = yes	4	4/8
If play = yes	==> humidity = normal & windy = false	4	4/9

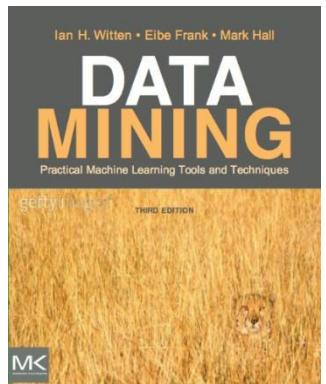
- ❖ 生成高支持度的項目集，從中獲取多個規則
- ❖ 策略:疊代地減少最小支持量，直到在給定的最小置信度下找到所需的規則數量為止

Lesson 3.3: 關聯規則

- ❖ 關聯規則比分類規則要多得多
 - 需要不同的技巧
- ❖ 支持度和置信度是兩個重要的指標
- ❖ Apriori是標準的關聯規則演算法
- ❖ 制定最小置信度，尋找支持度最大的規則
- ❖ 細節？ – 請見下一堂課

課程文本

- ❖ Section 4.5 *Mining association rules*





THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

More Data Mining with Weka

Class 3 – Lesson 4

學習關聯規則(*Learning association rules*)

Ian H. Witten

Department of Computer Science University
of Waikato
New Zealand

weka.waikato.ac.nz

Lesson 3.4: 學習關聯規則

Class 1 探索Weka的介面；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 3.1 決策樹與規則

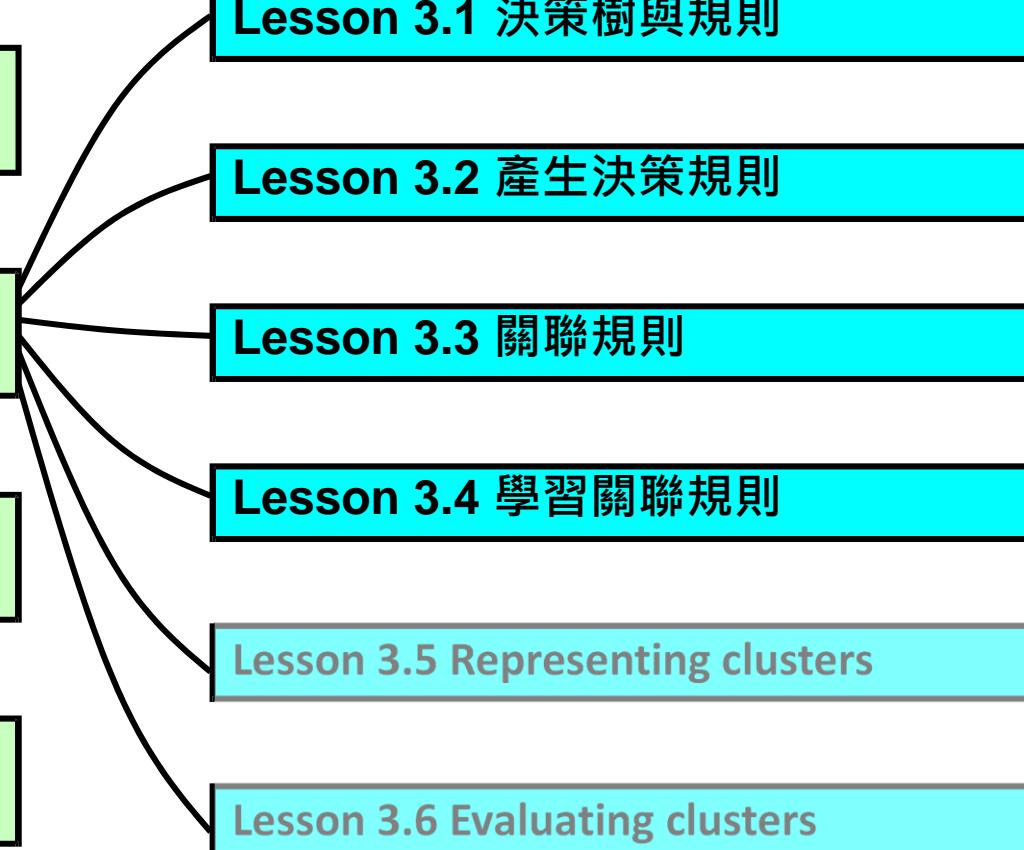
Lesson 3.2 產生決策規則

Lesson 3.3 關聯規則

Lesson 3.4 學習關聯規則

Lesson 3.5 Representing clusters

Lesson 3.6 Evaluating clusters



Lesson 3.4: 學習關聯規則

策略

- 指定最小的置信度
- 逐步減少覆蓋量，直到有足夠的規則符合最小置信率

單一項目集的7個可能的規則：

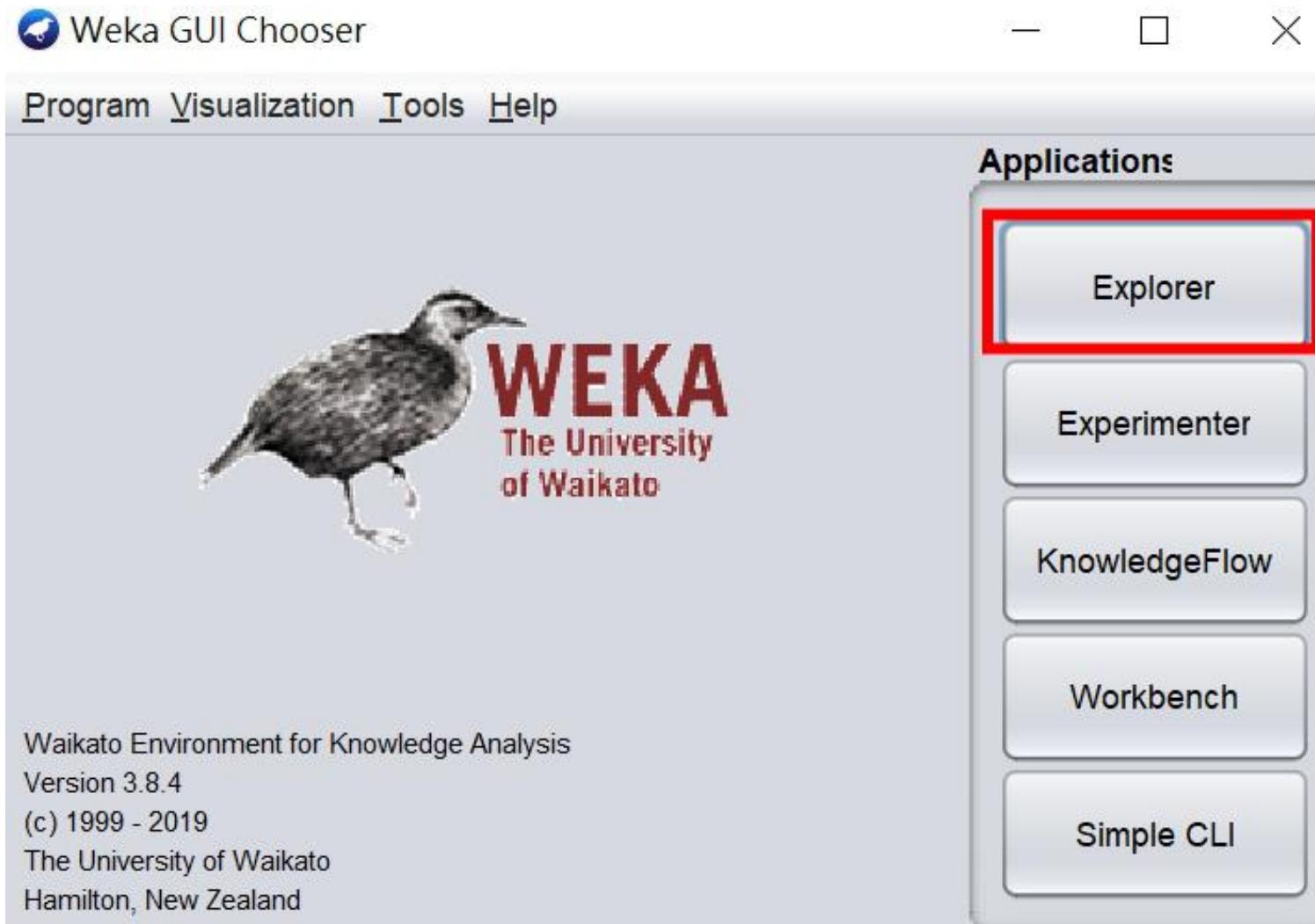
支持度 置信度

	If humidity = normal & windy = false play = yes If humidity = normal & play = yes ==> windy = false	4	4/4
	If windy = false & play = yes ==> humidity = normal	4	4/6
	If humidity = normal ==> windy = false & play = yes	4	4/7
1. 產生	If windy = false ==> humidity = normal & play = yes	4	4/8
2. 找到	If play = yes ==> humidity = normal & windy = false	4	4/9 度: 90%)
3. 繼續	==> humidity = normal & windy = false & play = yes	4	4/14 (集)

...不斷減少直到找到足夠的滿足條件的規則為止

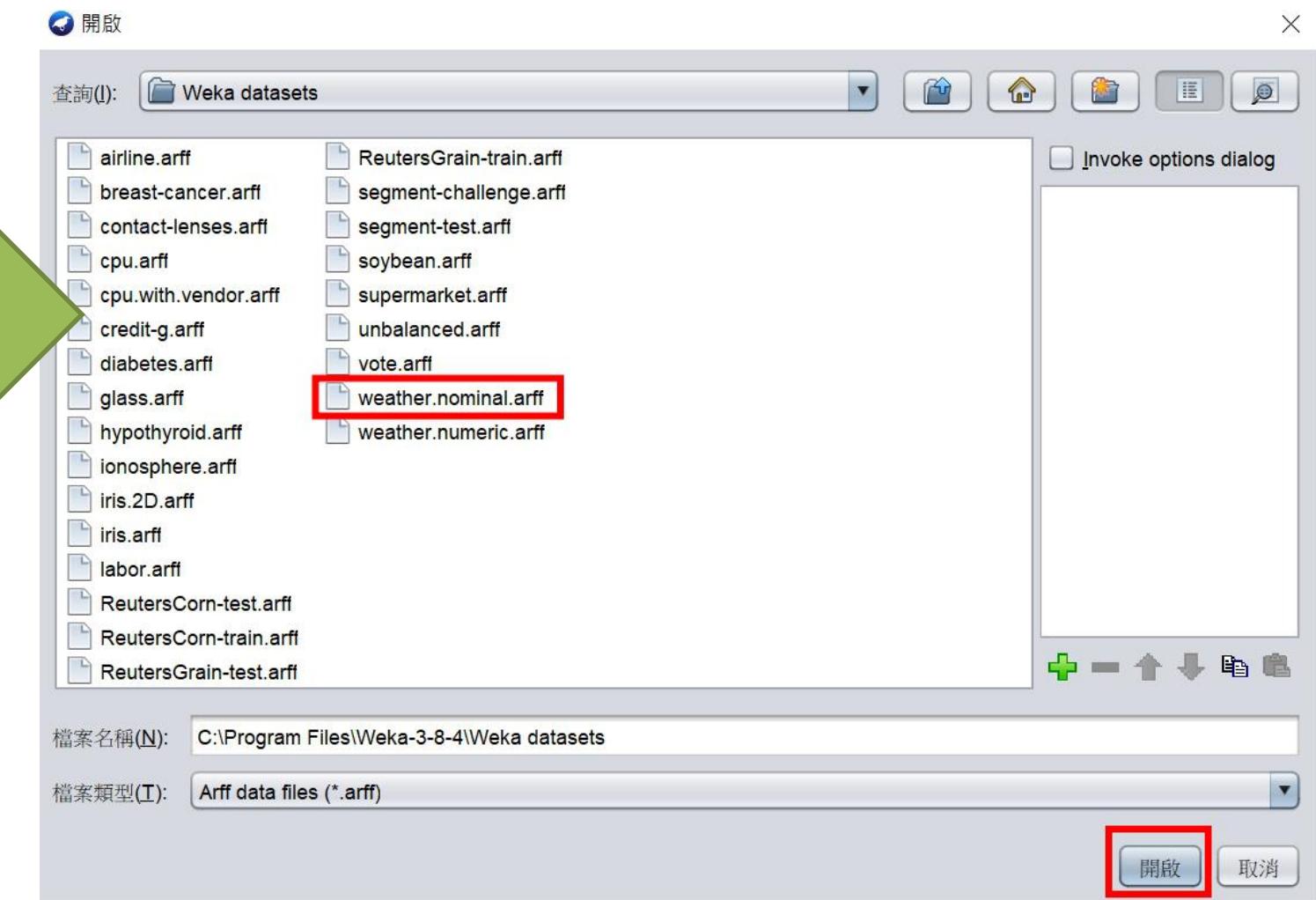
Lesson 3.4: 學習關聯規則

1. 開啟Weka的Explorer



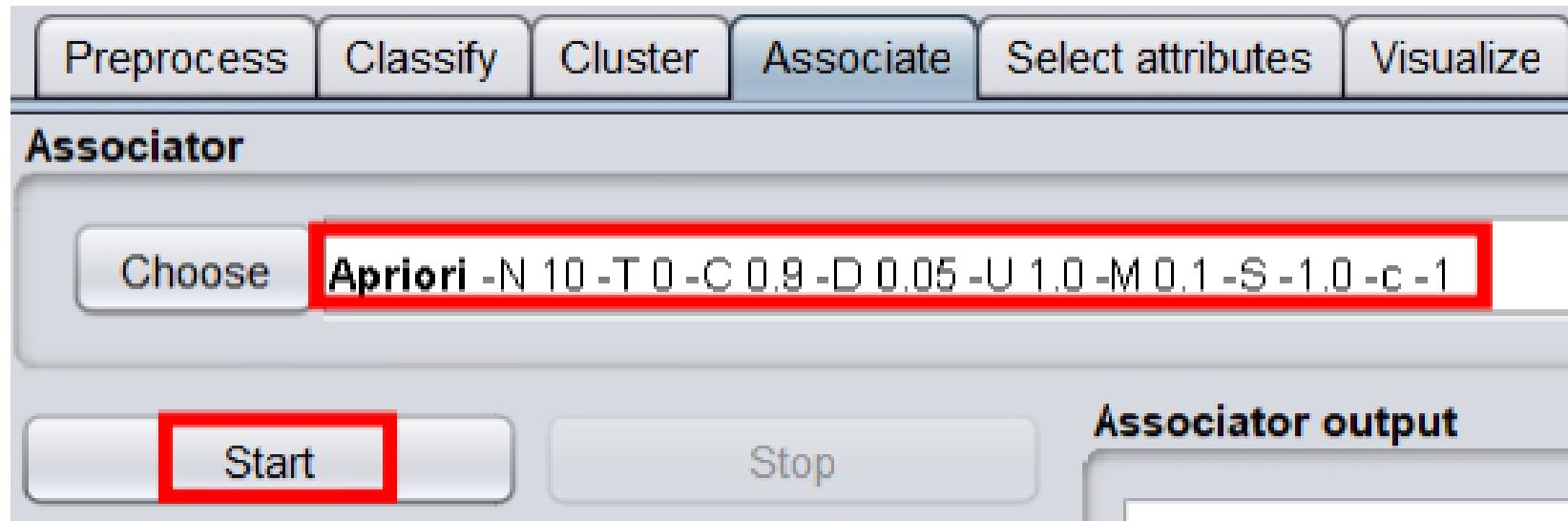
Lesson 3.4: 學習關聯規則

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**weather.nominal.arff**的檔案後，再以左鍵單擊下方”開啟”以載入此檔案



Lesson 3.4: 學習關聯規則

3. 確認選擇好Apriori分類器後，左鍵單擊Start按鈕。



Lesson 3.4: 學習關聯規則

▼ 執行結果

Preprocess Classify Cluster Associate Select attributes Visualize

Associator

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

Start Stop

Result list (right-click)

09:53:10 - Apriori

Associator output

```
=====
Minimum support: 0.15 (2 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 17

Generated sets of large itemsets:

Size of set of large itemsets L(1): 12
Size of set of large itemsets L(2): 47
Size of set of large itemsets L(3): 39
Size of set of large itemsets L(4): 6

Best rules found:

1. outlook=overcast 4 ==> play=yes 4    <conf:(1)> lift:(1.56) lev:(0.1) [1] conv:(1.43)
2. temperature=cool 4 ==> humidity=normal 4   <conf:(1)> lift:(2) lev:(0.14) [2] conv:(2)
3. humidity=normal windy=FALSE 4 ==> play=yes 4   <conf:(1)> lift:(1.56) lev:(0.1) [1] conv:(1.43)
4. outlook=sunny play=no 3 ==> humidity=high 3   <conf:(1)> lift:(2) lev:(0.11) [1] conv:(1.5)
5. outlook=sunny humidity=high 3 ==> play=no 3   <conf:(1)> lift:(2.8) lev:(0.14) [1] conv:(1.93)
6. outlook=rainy play=yes 3 ==> windy=FALSE 3   <conf:(1)> lift:(1.75) lev:(0.09) [1] conv:(1.29)
7. outlook=rainy windy=FALSE 3 ==> play=yes 3   <conf:(1)> lift:(1.56) lev:(0.08) [1] conv:(1.07)
8. temperature=cool play=yes 3 ==> humidity=normal 3   <conf:(1)> lift:(2) lev:(0.11) [1] conv:(1.5)
9. outlook=sunny temperature=hot 2 ==> humidity=high 2   <conf:(1)> lift:(2) lev:(0.07) [1] conv:(1)
10. temperature=hot play=no 2 ==> outlook=sunny 2   <conf:(1)> lift:(2.8) lev:(0.09) [1] conv:(1.29)
```

產生了10個規則，默認的數量是10。

我們可以看到這些規則的支持度，從4到3再到2。

數據集中只有兩個實例符合最後兩個規則。

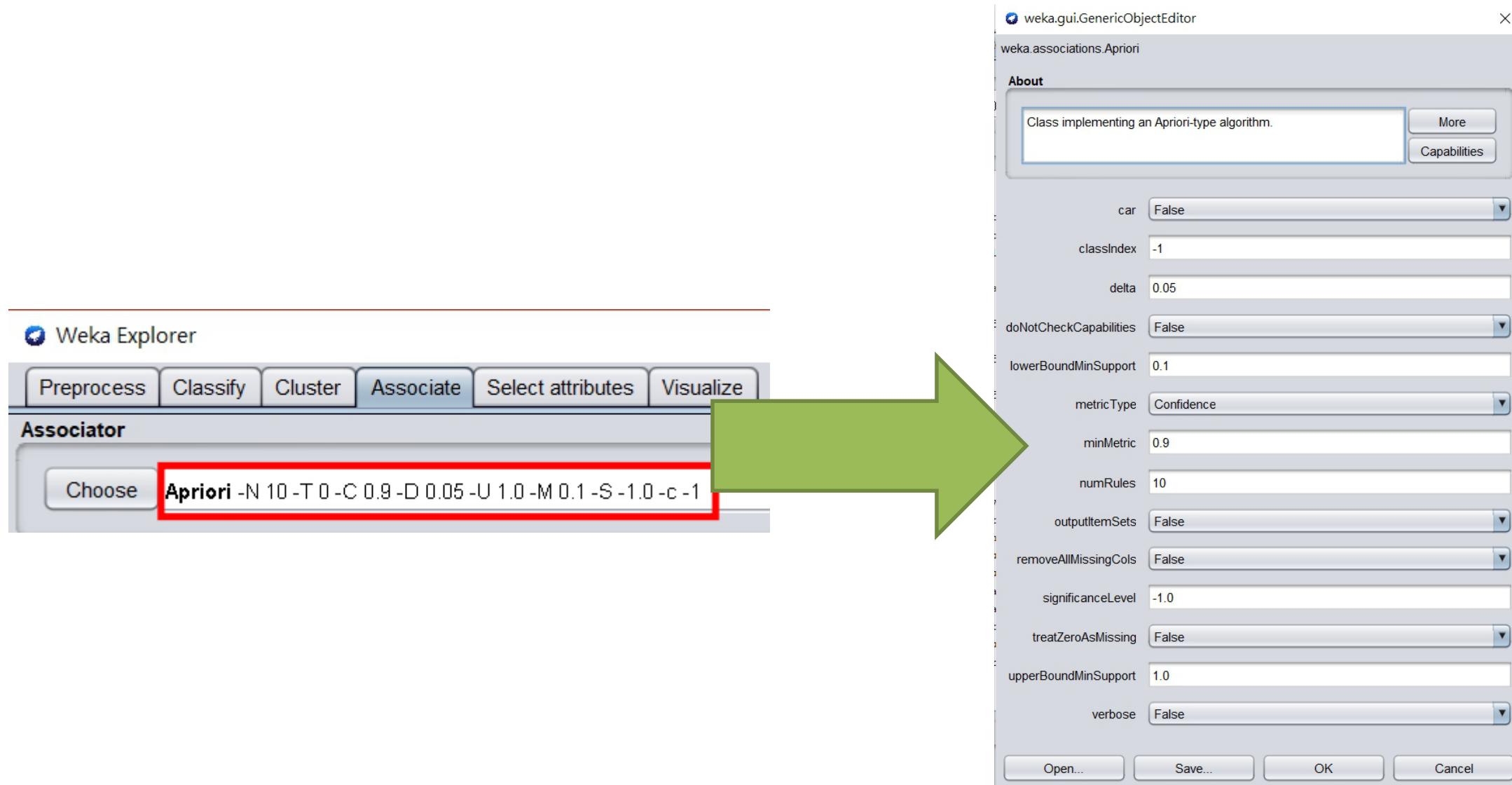
這些規則的置信率是百分百。

Lesson 3.4: 學習關聯規則

- ❖ Weather資料產生336個置信率為100%的規則!
 - 但只有8個規則支持度 ≥ 3 , 只有58個規則支持度 ≥ 2
- ❖ Weka : 我們指定最小置信度水平(**minMetric**, 預設為90%)；我們指定產生的規則的數量(**numRules**, 預設為10)
- ❖ 支持度是按實例數目的比例表示的
- ❖ Weka 執行Apriori演算法好幾次
 - 從**upperBoundMinSupport**開始 (通常從置信度100%開始)
 - 每次都由參數**delta**來減少置信度 (預設為5%)
 - 當得到高於最小置信率的規則時停止 (參數**numRules**)
 - ...或者達到另外一個參數**lowerBoundMinSupport**的限制時停止 (預設為10%)

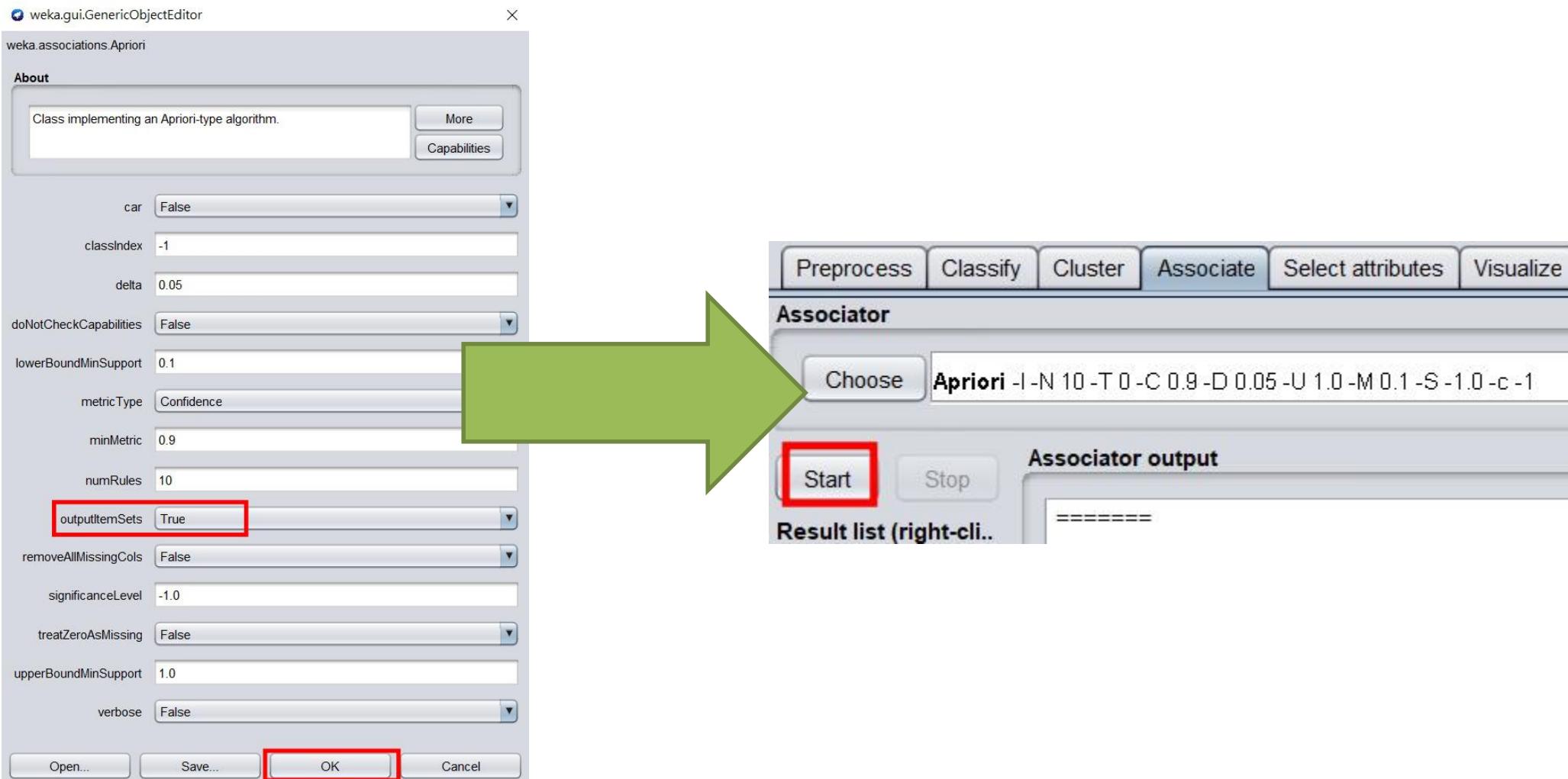
Lesson 3.4: 學習關聯規則

1. 左鍵單擊分類器名稱(左圖紅框處)，開啟配置視窗(右圖)。



Lesson 3.4: 學習關聯規則

2. 將參數outputItemSets設定為True，接著左鍵單擊OK按鈕回到Classify面板，左鍵單擊Start按鈕。



Lesson 3.4: 學習關聯規則

▼ 執行結果

Preprocess Classify Cluster Associate Select attributes Visualize

Associator

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

Start Stop

Associator output

Result list (right-click to copy)

09:53:10 - Apriori
09:58:23 - Apriori

```
humidity=high windy=TRUE play=no 2
humidity=high windy=FALSE play=yes 2
humidity=high windy=FALSE play=no 2
humidity=normal windy=TRUE play=yes 2
humidity=normal windy=FALSE play=yes 4

Size of set of large itemsets L(4): 6

Large Itemsets L(4):
outlook=sunny temperature=hot humidity=high play=no 2
outlook=sunny humidity=high windy=FALSE play=no 2
outlook=overcast temperature=hot windy=FALSE play=yes 2
outlook=rainy temperature=mild windy=FALSE play=yes 2
outlook=rainy humidity=normal windy=FALSE play=yes 2
temperature=cool humidity=normal windy=FALSE play=yes 2

Best rules found:

1. outlook=overcast 4 ==> play=yes 4    <conf:(1)> lift:(1.56) lev:(0.1) [1] conv:(1.43)
2. temperature=cool 4 ==> humidity=normal 4    <conf:(1)> lift:(2) lev:(0.14) [2] conv:(2)
3. humidity=normal windy=FALSE 4 ==> play=yes 4    <conf:(1)> lift:(1.56) lev:(0.1) [1] conv:(1.43)
4. outlook=sunny play=no 3 ==> humidity=high 3    <conf:(1)> lift:(2) lev:(0.11) [1] conv:(1.5)
5. outlook=sunny humidity=high 3 ==> play=no 3    <conf:(1)> lift:(2.8) lev:(0.14) [1] conv:(1.93)
6. outlook=rainy play=yes 3 ==> windy=FALSE 3    <conf:(1)> lift:(1.75) lev:(0.09) [1] conv:(1.29)
7. outlook=rainy windy=FALSE 3 ==> play=yes 3    <conf:(1)> lift:(1.56) lev:(0.08) [1] conv:(1.07)
8. temperature=cool play=yes 3 ==> humidity=normal 3    <conf:(1)> lift:(2) lev:(0.11) [1] conv:(1.5)
9. outlook=sunny temperature=hot 2 ==> humidity=high 2    <conf:(1)> lift:(2) lev:(0.07) [1] conv:(1)
10. temperature=hot play=no 2 ==> outlook=sunny 2    <conf:(1)> lift:(2.8) lev:(0.09) [1] conv:(1.29)
```

Lesson 3.4: 學習關聯規則

▼ 執行結果

The screenshot shows the Weka Explorer interface with the Associator tab selected. The 'Choose' dropdown is set to 'Apriori -I -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1'. The 'Associator output' panel displays the results of the association rule mining process.

Result list (right-click to copy)

- 09:53:10 - Apriori
- 09:58:23 - Apriori

Associator output

```
Size of set of large itemsets L(1): 12

Large Itemsets L(1):
outlook=sunny 5
outlook=overcast 4
outlook=rainy 5
temperature=hot 4
temperature=mild 6
temperature=cool 4
humidity=high 7
humidity=normal 7
windy=TRUE 6
windy=FALSE 8
play=yes 9
play=no 5

Size of set of large itemsets L(2): 47

Large Itemsets L(2):
outlook=sunny temperature=hot 2
outlook=sunny temperature=mild 2
outlook=sunny humidity=high 3
outlook=sunny humidity=normal 2
outlook=sunny windy=TRUE 2
outlook=sunny windy=FALSE 3
outlook=sunny play=yes 2
outlook=sunny play=no 3
outlook=overcast temperature=hot 2
outlook=overcast humidity=high 2
outlook=overcast humidity=normal 2
```

Lesson 3.4: 學習關聯規則

最小支持度: 0.15 (2個實例)

最小置信度: 0.9

運行循環次數: 17

產生大型項目集的大小 :

大型項目集的大小 $L(1)$: 12

大型項目集的大小 $L(2)$: 47

大型項目集的大小 $L(3)$: 39

大型項目集的大小 $L(4)$: 6

得出最好的規則為:

1. outlook = overcast 4 ==> play =

❖ 執行了17 次的Apriori演算法: 支持度 = 100%, 95%, 90%, ..., 20%, 15%

- 14, 13, 13, ..., 3, 2 個實例

❖ 設定參數 **outputItemSets** 可以看到項目集
度 > 3 他們基於最後的支持度 , 如2

支持度 ≥ 2 的 12 個單項集

outlook = sunny 5

outlook = overcast 4

...

play = no 5

支持度 ≥ 2 的 47 個雙項集

outlook = sunny & temperature = hot 2

outlook = sunny & humidity = high 3

...

支持度 ≥ 2 的 39 個三項集

outlook = sunny & temperature = hot & humidity = high

2

outlook = sunny & humidity = high & play =

no 3 outlook = sunny & windy = false &

支持度 ≥ 2 的 26 個四項集

outlook = sunny & humidity = high & windy =

false & play = no

Lesson 3.4: 學習關聯規則

Weka中的其他參數

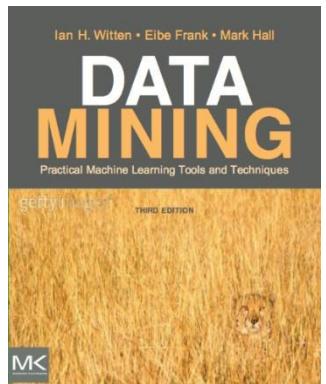
- ❖ **car**:始終產生預測類屬性的規則
 - 使用**classIndex**設置類屬性
- ❖ **significanceLevel**:根據統計測試篩選規則 (χ^2)
 - 不可靠，因為得經過許多測試，偶然才會發現重要的結果
 - 對於小的支持度，測試是不準確的
- ❖ **metricType**: 排行規則的不同措施
 - *Confidence*
 - *Lift*
 - *Leverage*
 - *Conviction*
- ❖ **removeAllMissingCols**:刪除所有值為 "缺失 "的屬性

Lesson 3.4: 學習關聯規則

- ❖ Apriori算法多次處理資料
 - 產生單項集、雙項集, ... 滿足最小支持度
 - 把每個項目集轉化成規則，查看它們的置信率
- ❖ 快速且有效率 (提供符合主記憶體的資料)
- ❖ Weka多次調用算法，逐步降低對覆蓋量的要求，直到得到足夠的高置信率的規則
 - 有很多參數控制這個循環過程

課程文本

- ❖ Section 11.7 *Association-rule learners*





THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

More Data Mining with Weka

Class 3 – Lesson 5

聚類的表達(*Representing clusters*)

Ian H. Witten

Department of Computer Science
University of Waikato
New Zealand

weka.waikato.ac.nz

Lesson 3.5: 聚類的表達

Class 1 探索Weka的介面；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 3.1 決策樹與規則

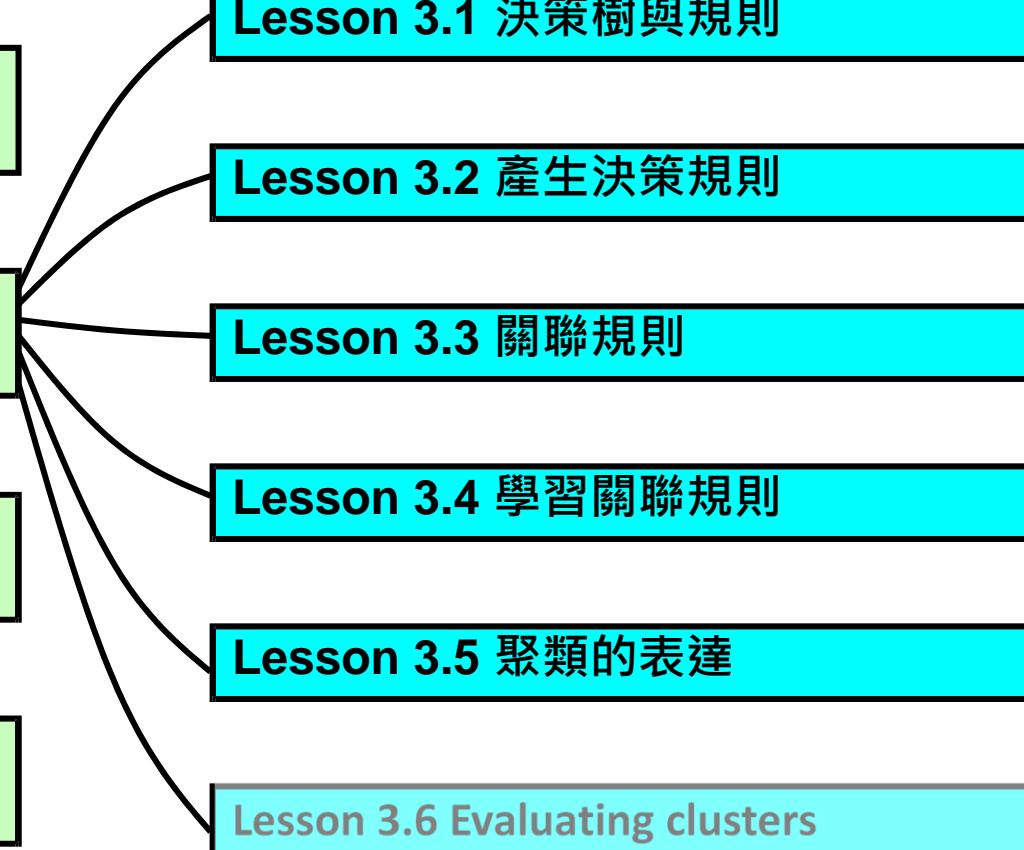
Lesson 3.2 產生決策規則

Lesson 3.3 關聯規則

Lesson 3.4 學習關聯規則

Lesson 3.5 聚類的表達

Lesson 3.6 Evaluating clusters



Lesson 3.5: 聚類的表達

- ❖ 在聚類中，沒有類屬性的概念
- ❖ 我們僅僅是試著把實例分成自然的組，或稱「聚類」

例子

- ❖ 在Explorer中檢驗 iris.arff檔案
- ❖ 想像你刪除了類別屬性
- ❖ 你可以通過聚類資料來恢復類別嗎？



鳶尾花

Setosa



鳶尾花 Versicolor

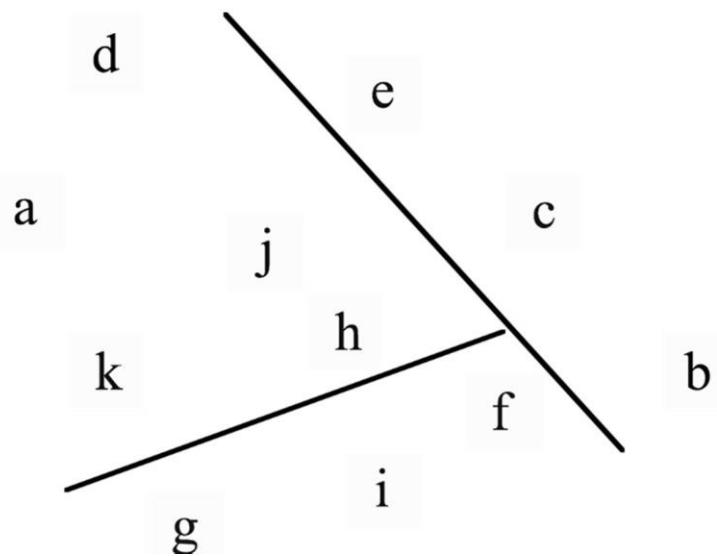


鳶尾花 Virginica

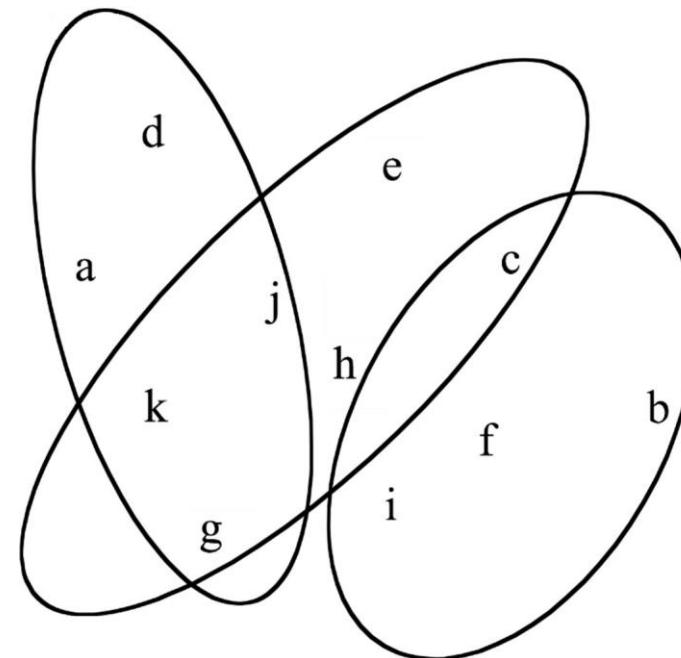
Lesson 3.5: 聚類的表達

聚類的種類

1. 不相交的集合(**Disjoint sets**)：
將實例空間分割成集合，這樣每個
實例空間只屬於一個聚類。



2. 重疊的集合(**Overlapping sets**)：
實例以一定的概率分屬於每個類。



Lesson 3.5: 聚類的表達

聚類的種類

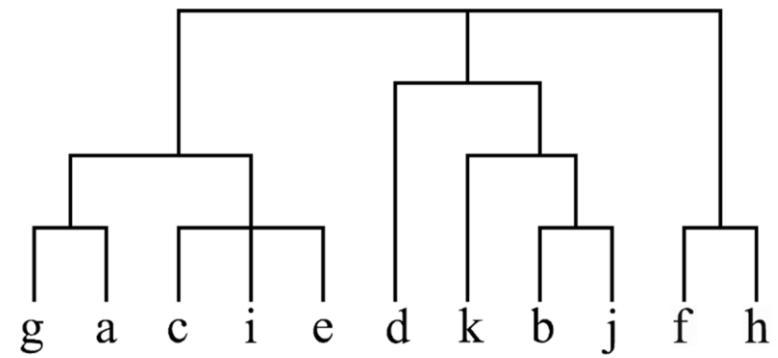
3. 概率聚類(Probabilistic clusters) :

以實例a說明：有40%的概率屬於聚類1，10%的概率屬於聚類2，50%的概率屬於聚類3。

a	0.4	0.1	0.5
b	0.1	0.8	0.1
c	0.3	0.3	0.4
d	0.1	0.1	0.8
e	0.4	0.2	0.4
f	0.1	0.4	0.5
g	0.7	0.2	0.1
h	0.5	0.4	0.1

...

4. 分層聚類(Hierarchical clusters)：
屬於底部聚類的實例(如a和g)，在底部被合並在一起，然後這些聚類再與上一層結合在一起，依此類推，直到頂部，整個資料集聚合成為一個大的聚類。這是樹狀圖，是樹的一種。



Lesson 3.5: 聚類的表達

KMeans:一個基於距離的疊代聚類(不相交的集合(disjoint sets))

1. 指定 k ,我們想要的聚類的數量
2. 算法隨機選擇 k 個點作為聚類的中心
3. 算法將數據集中的所有實例分配到最近的聚類中心
4. 對每個聚類，計算出所含實例的質心 (如平均值)
5. 這些質心將成為各個聚類新的中心
6. 回到開始重覆整個過程，直到聚類的中心不再變化

KMeans最小化每個實例至聚類中心的距離的平方的和
只是局部最小化，而不是全局的

Lesson 3.5: 聚類的表達

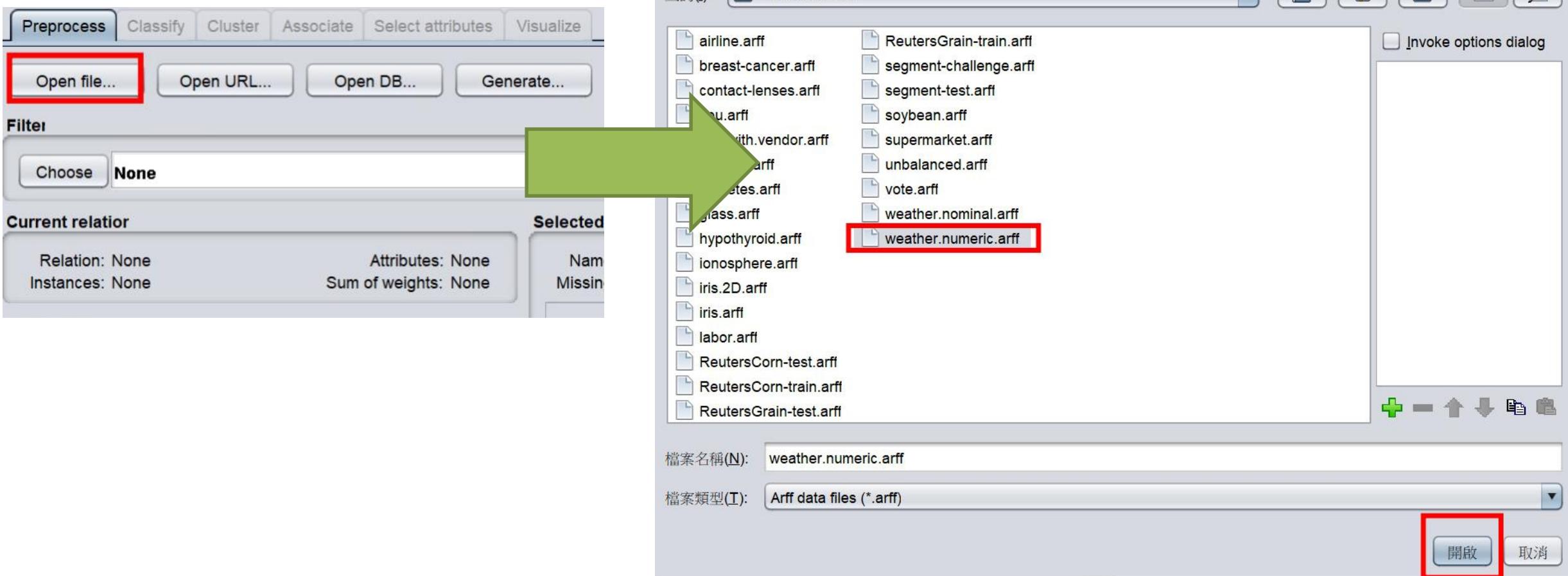
我們試著執行KMeans聚類法。

1. 開啟Weka的Explorer



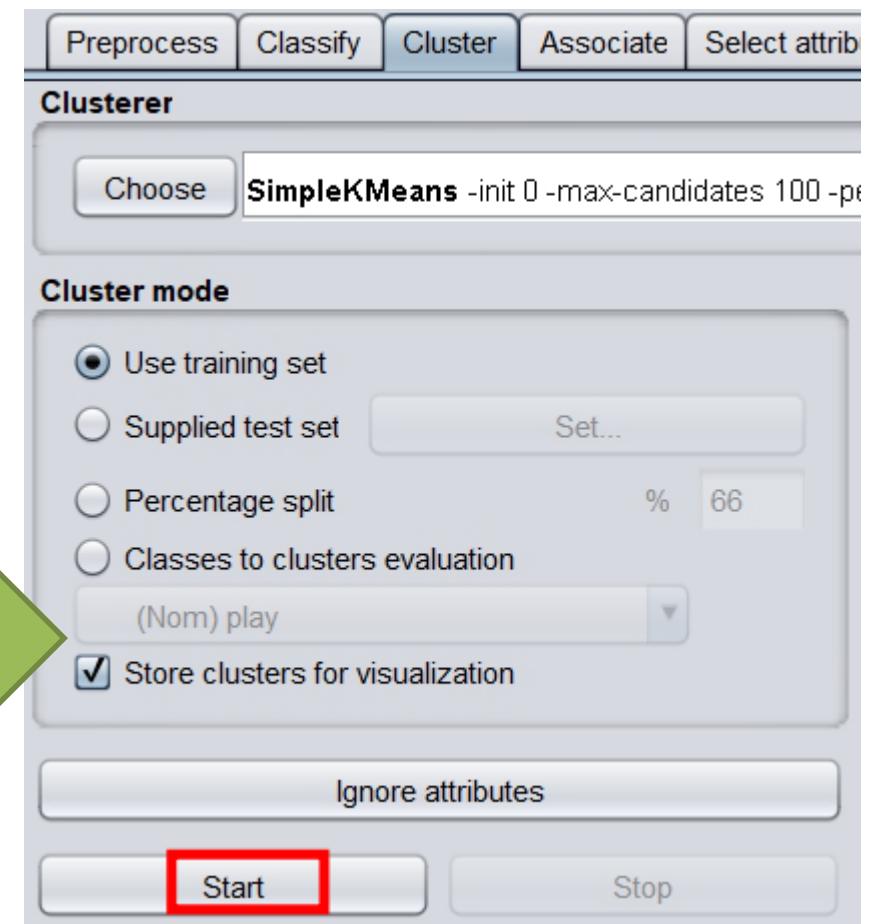
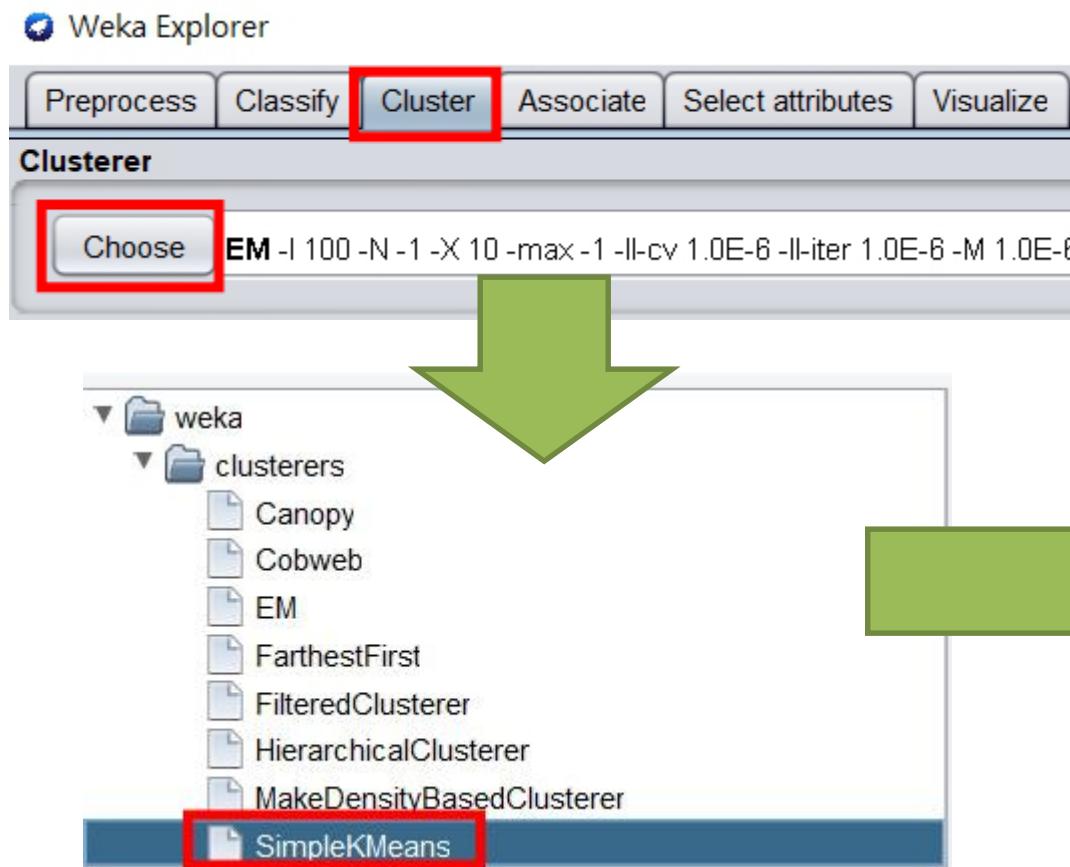
Lesson 3.5: 聚類的表達

2. 左鍵點擊Open file...開啟右圖視窗，進入自行複製的Weka datasets資料夾，左鍵單擊**weather.numeric.arff**的檔案後，再以左鍵單擊下方開啟按鈕以載入此檔案



Lesson 3.5: 聚類的表達

3. 於 Cluster面板，左鍵單擊Choose按鈕後，在出現的選單中左鍵單擊weka/cluster路徑下的SimpleKMeans聚類法。接著左鍵單擊Start按鈕。



Lesson 3.5: 聚類的表達

執行結果：

- 得到了兩個聚類。
- 其中一個包含9個實例，另一個有5個實例。
- 總平方誤差(sum of squared errors)是16.2——這是我們想最小化的。

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

- Preprocess, Classify, Cluster, Associate, Select attributes, Visualize** tabs are at the top.
- Clusterer** tab is selected.
- Choose** dropdown: SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance" -R first-last" -I 500 -num-slots
- Cluster mode**: Use training set (selected), Supplied test set, Percentage split, Classes to clusters evaluation, (Nom) play, Store clusters for visualization.
- Result list (right-click for options)**: 10:27:16 - SimpleKMeans
- Clusterer output** pane:
 - Number of iterations: 3
 - Within cluster sum of squared errors: 16.237456311387238 (highlighted by a red box)
 - Initial starting points (random):
 - Cluster 0: rainy,75,80, FALSE, yes
 - Cluster 1: overcast,64,65, TRUE, yes
 - Missing values globally replaced with mean/mode
 - Final cluster centroids:

Attribute	Full Data	Cluster# 0	Cluster# 1
(14.0)	(9.0)	(5.0)	

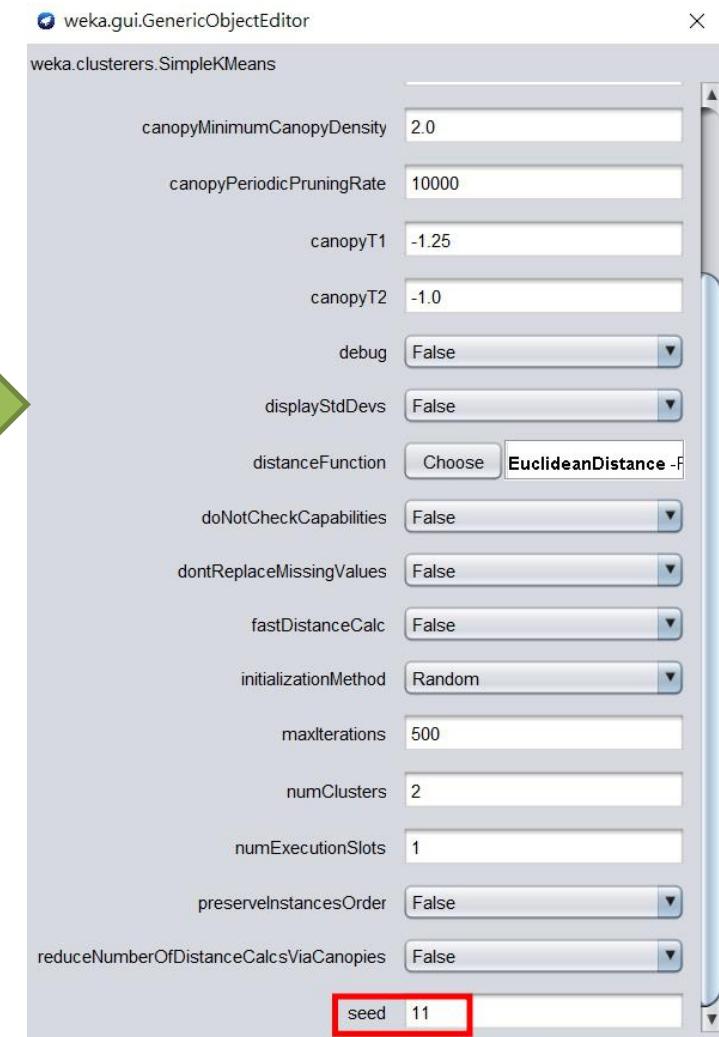
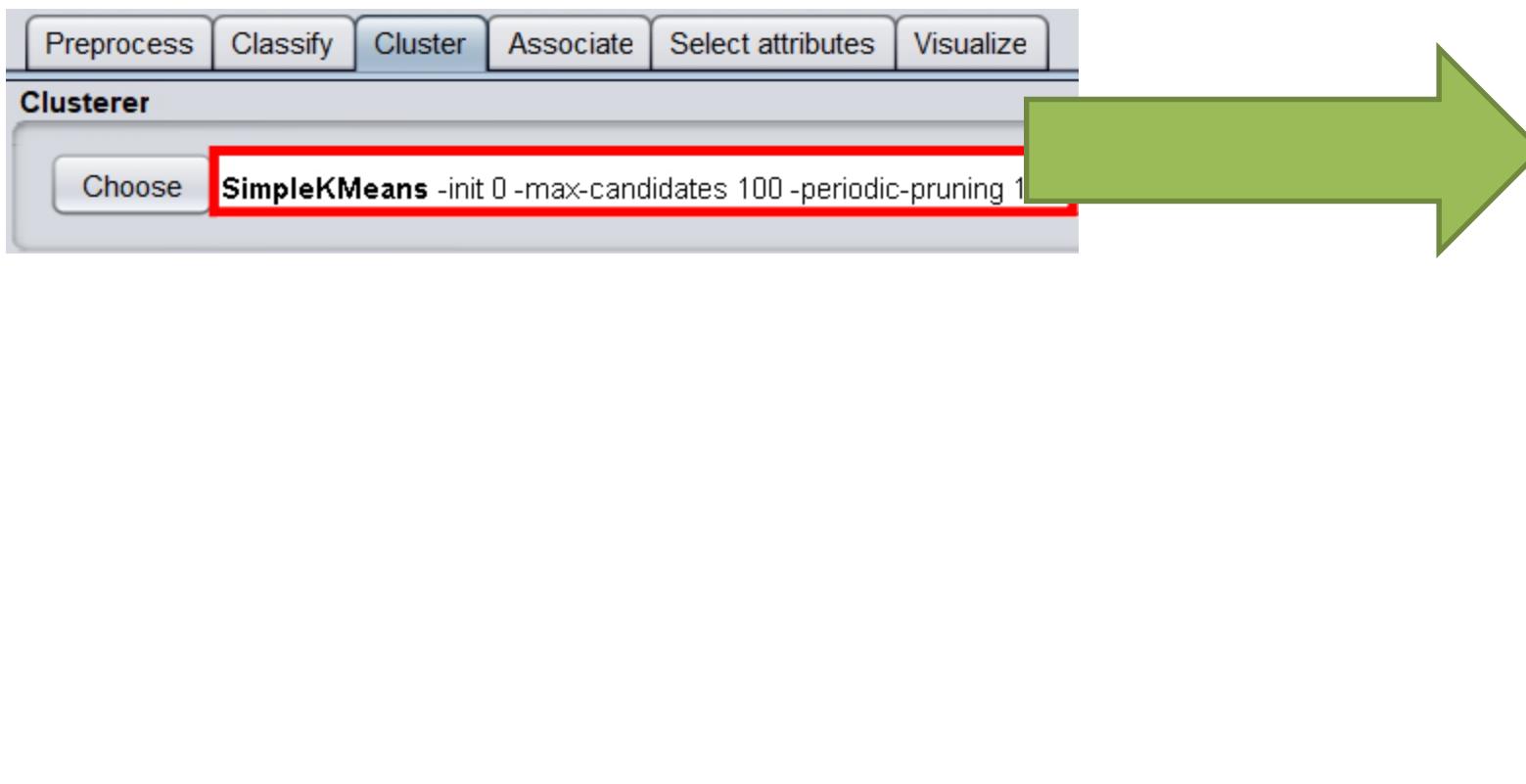
outlook	sunny	sunny	overcast
temperature	73.5714	75.8889	69.4
humidity	81.6429	84.1111	77.2
windy	FALSE	FALSE	TRUE
play	yes	yes	yes

 - Time taken to build model (full training data) : 0.01 seconds
 - ==== Model and evaluation on training set ===
 - Clustered Instances

0	9 (64%)
1	5 (36%)

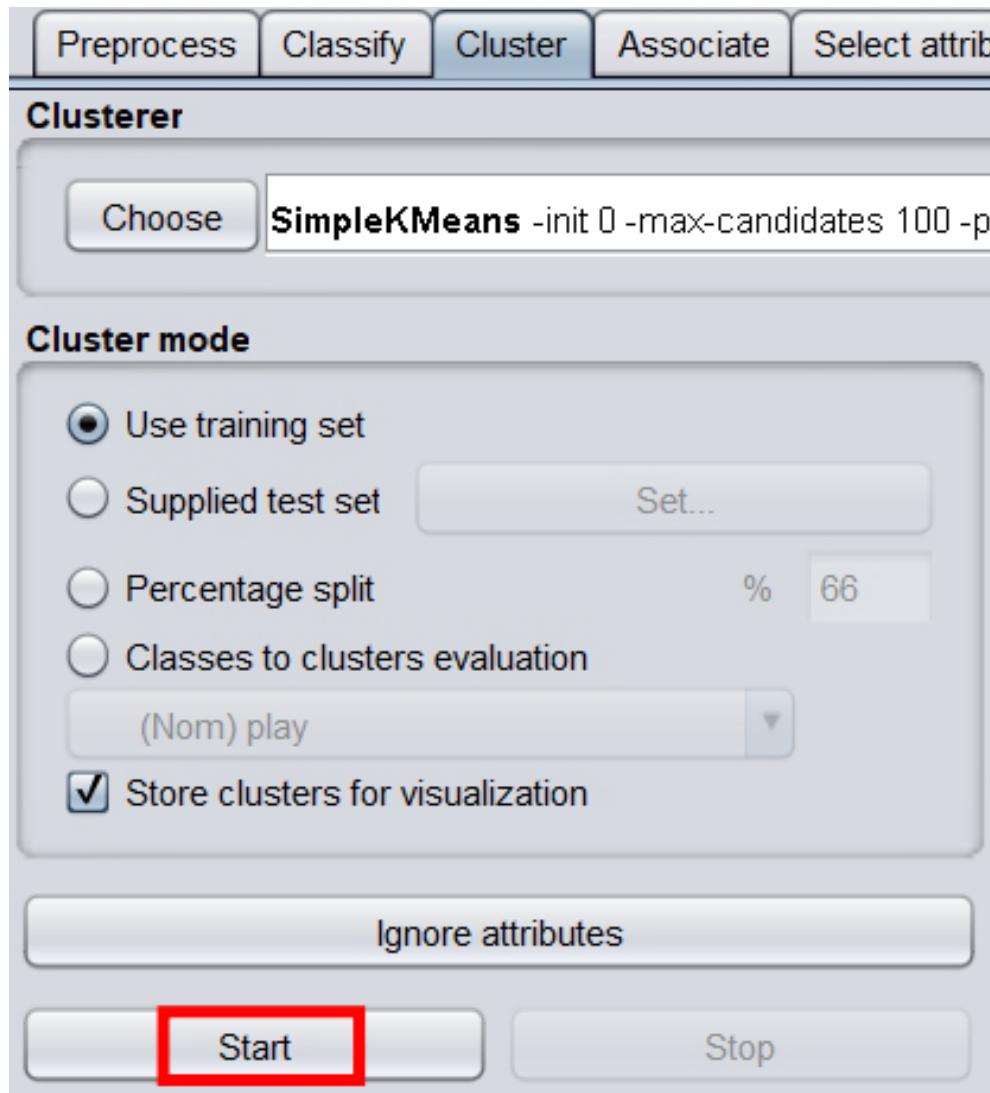
Lesson 3.5: 聚類的表達

4. 在Cluster面板，左鍵單擊聚類法名稱(左圖紅框處)，開啟配置視窗(右圖)。將參數seed的值設定為11，然後左鍵單擊OK按鈕。



Lesson 3.5: 聚類的表達

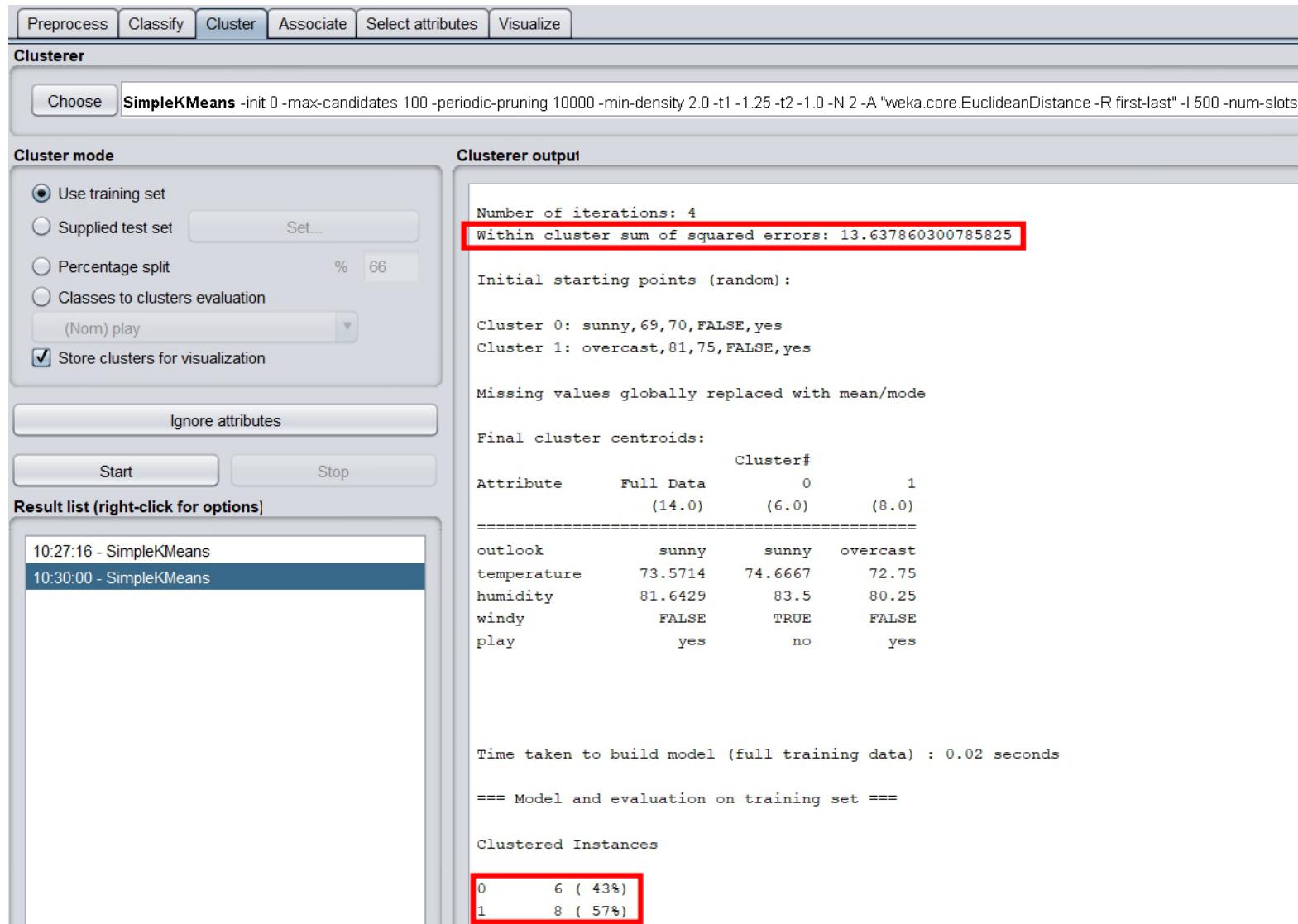
5. 回到Cluster面板，以左鍵單擊Start按鈕。



Lesson 3.5: 聚類的表達

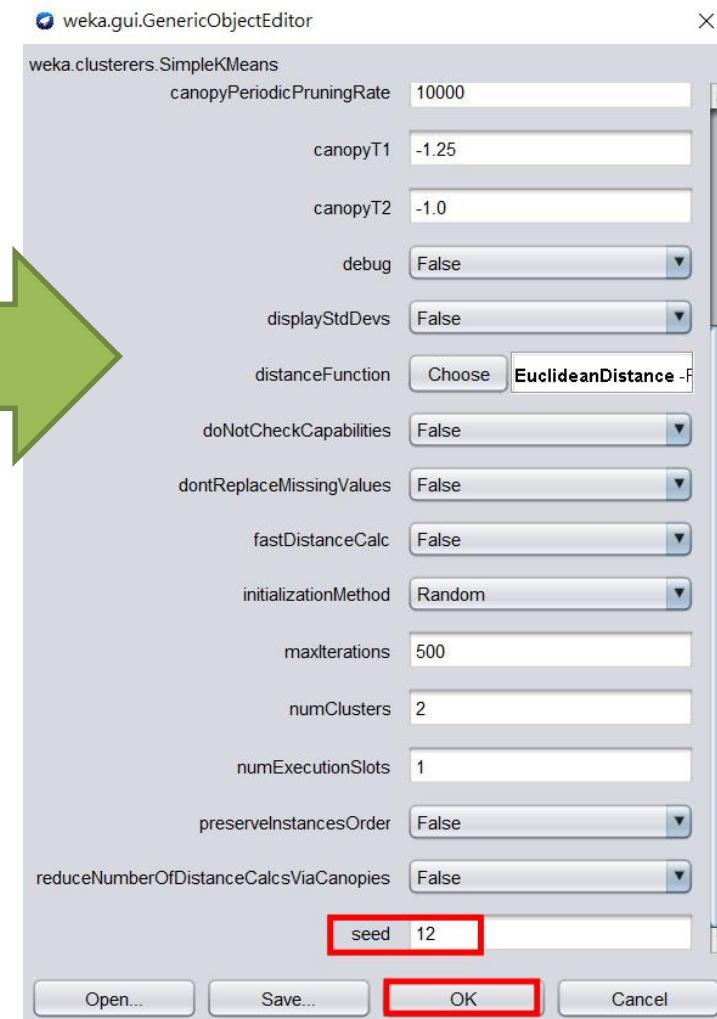
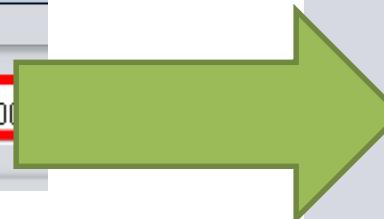
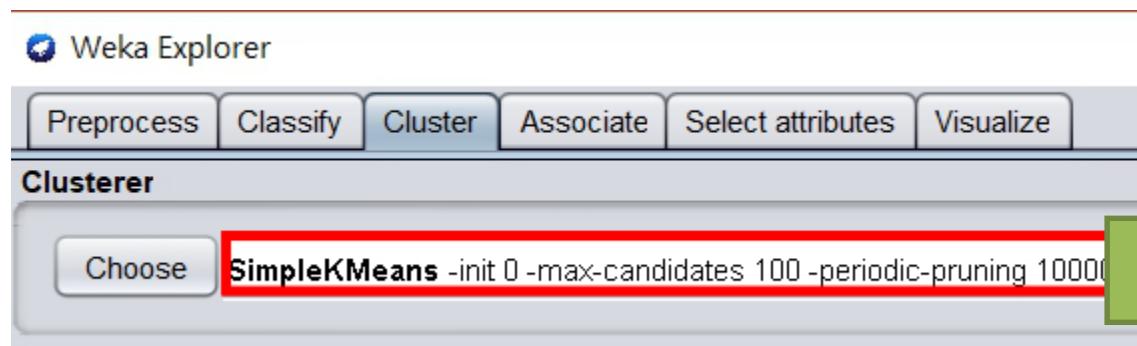
執行結果：

- 我們得到與剛才不同的聚類。
- 一樣是兩個聚類，其中一個包含6個實例，另外包含8個實例。
- 這次總平方誤差是 13.6。



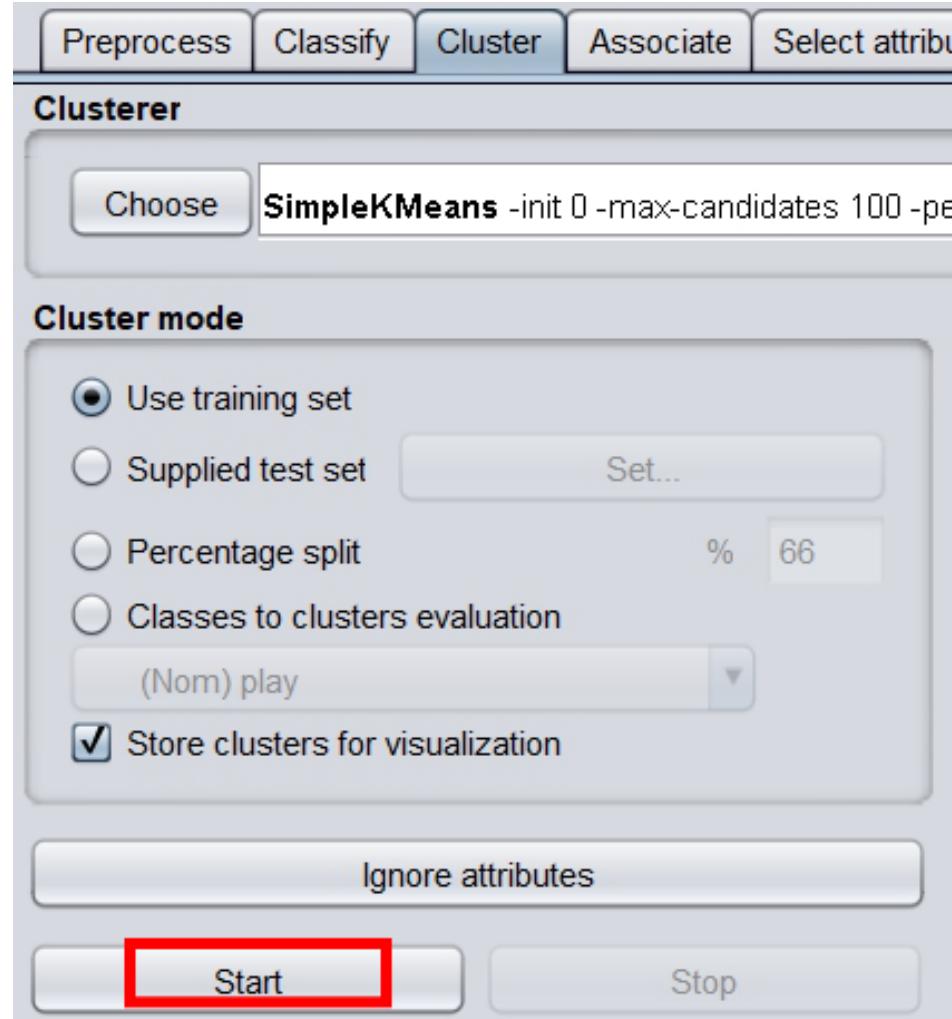
Lesson 3.5: 聚類的表達

6. 在Cluster面板，左鍵單擊聚類法名稱(左圖紅框處)，開啟配置視窗(右圖)。將參數seed的值設定為12，然後左鍵單擊OK按鈕。



Lesson 3.5: 聚類的表達

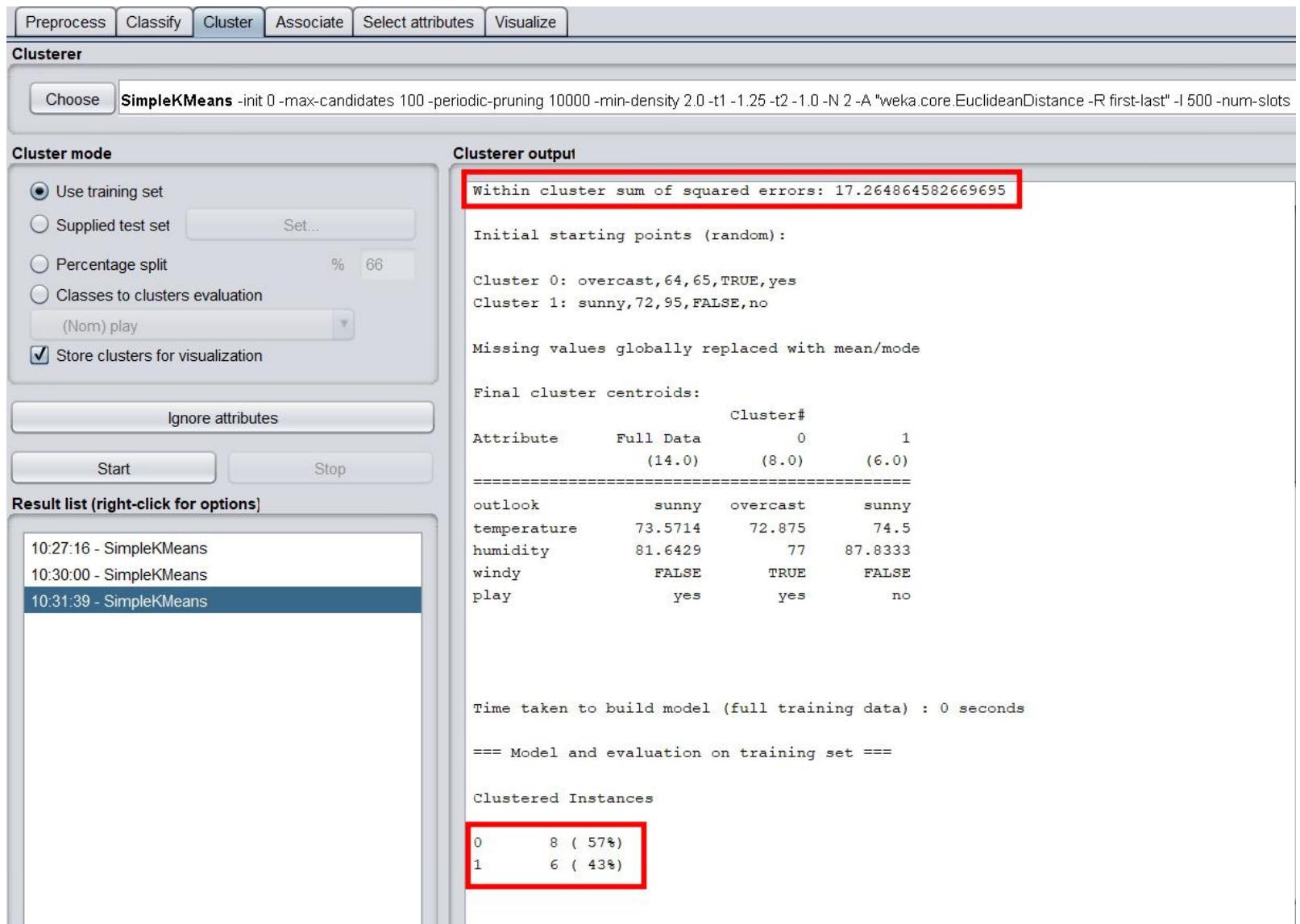
7. 回到Cluster面板，以左鍵單擊Start按鈕。



Lesson 3.5: 聚類的表達

執行結果：

- 再次得到不同的聚類。
- 一樣是兩個聚類，其中一個包含8個實例，另外包含6個實例。
- 這次總平方誤差是17.3。



Lesson 3.5: 聚類的表達

KMeans 聚類法

- ❖ 開啟 **weather.numeric.arff**
- ❖ Cluster面板; 選擇**SimpleKMeans**
- ❖ 注意參數: **numClusters, distanceFunction, seed** (預設值為10)

- ❖ 得到兩個聚類, 9 和 5 個實例, 總平方誤差(total squared error)為16.2
 {1/no, 2/no, 3/yes, 4/yes, 5/yes, 8/no, 9/yes, 10/yes, 13/yes} {6/no, 7/yes, 11/yes, 12/yes, 14/no}
- ❖ 將參數**seed**設為11
- ❖ 得到兩個聚類, 6 和 8 個實例, 總平方誤差為13.6
- ❖ 將參數**seed**設為12
- ❖ 總平方誤差為17.3

Lesson 3.5: 聚類的表達

XMeans: KMeans算法的延伸

- ❖ 自動選取聚類數量
- ❖ 你可以指定最小和最大的聚類數量
- ❖ 你可以指定4種不同的距離指標
- ❖ 使用kD-trees保證運行的效率

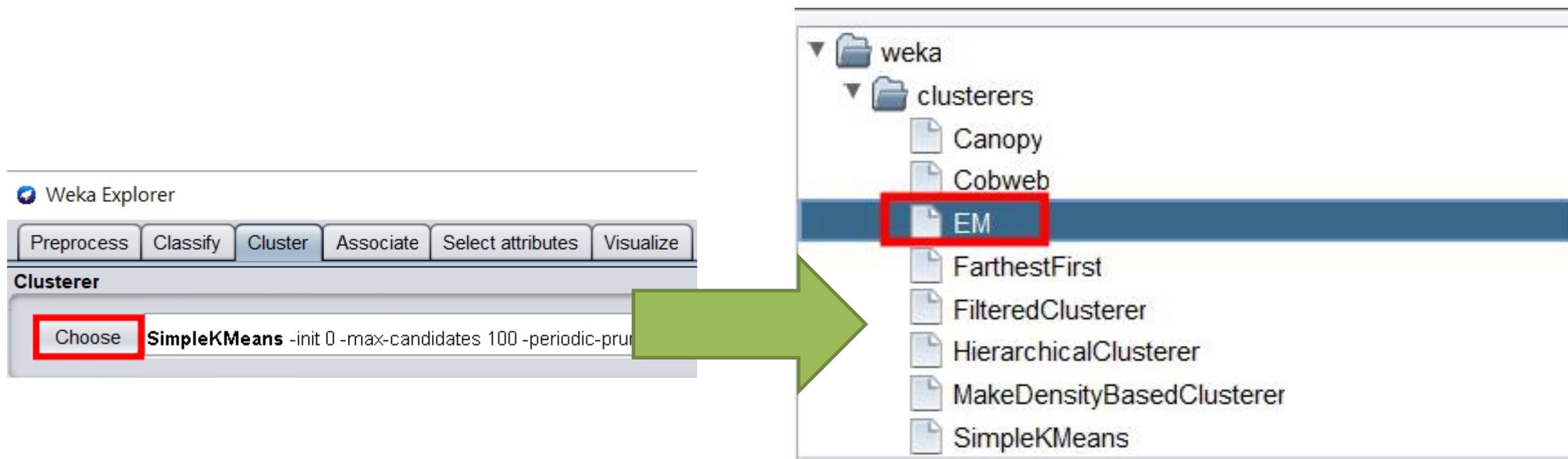
不能處理名詞性屬性

- ❖ 忽略weather資料中的名詞性屬性
outlook, windy, play

Lesson 3.5: 聚類的表達

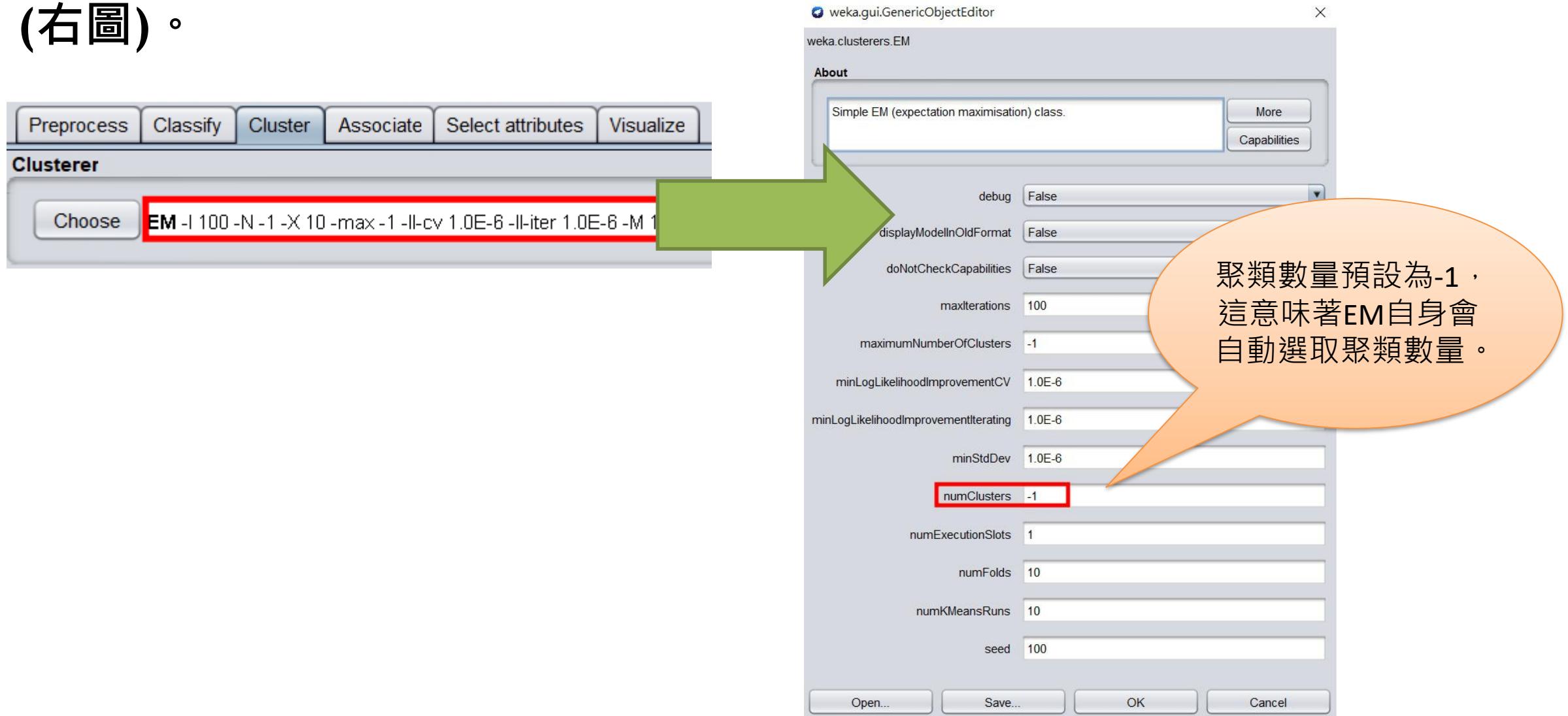
我們接著執行EM聚類法。

1. 於 Cluster面板，左鍵單擊Choose按鈕後，在出現的選單中左鍵單擊weka/cluster路徑下的EM聚類法。



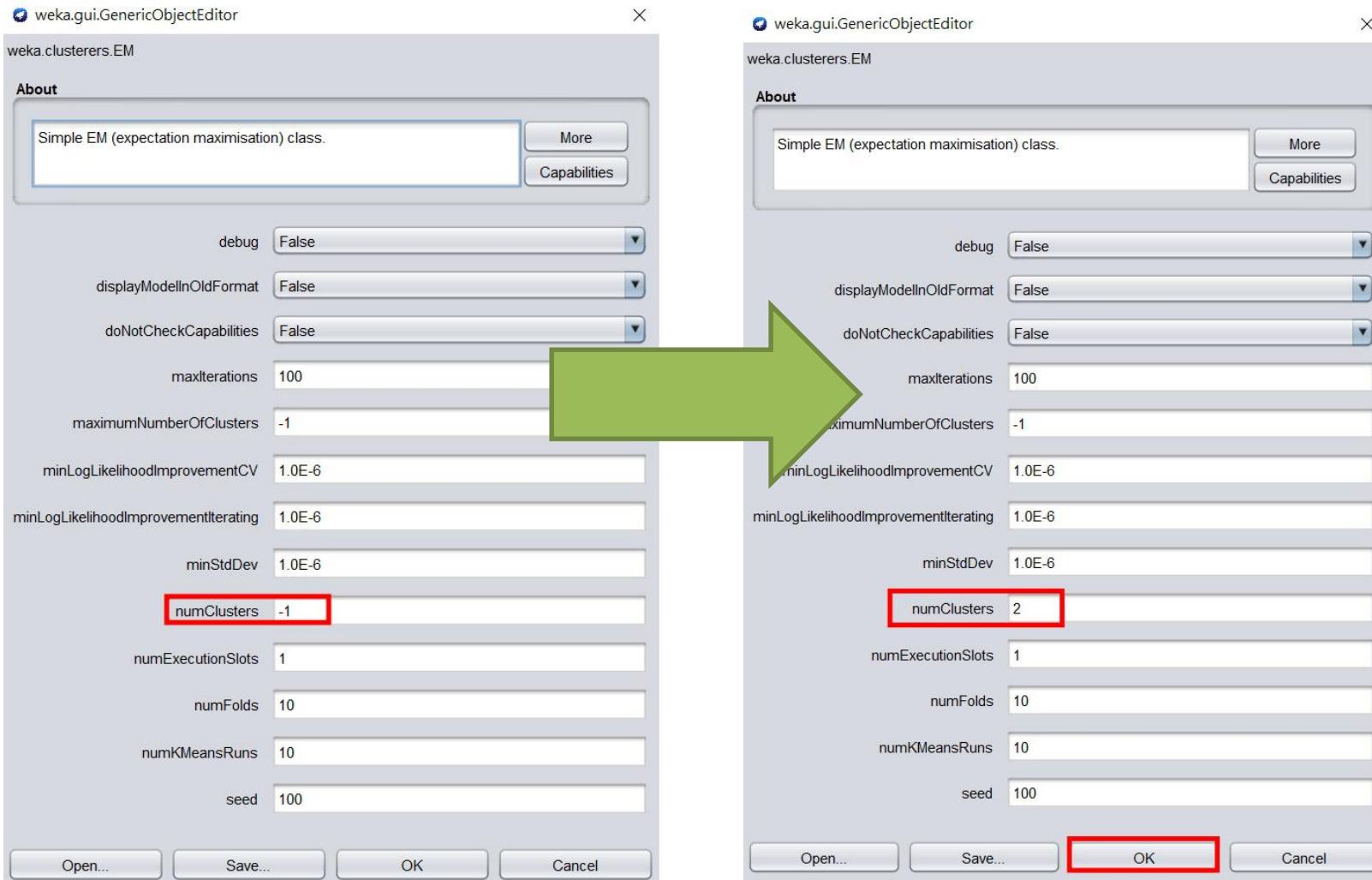
Lesson 3.5: 聚類的表達

2. 在Cluster面板，左鍵單擊聚類法名稱(左圖紅框處)，開啟配置視窗(右圖)。



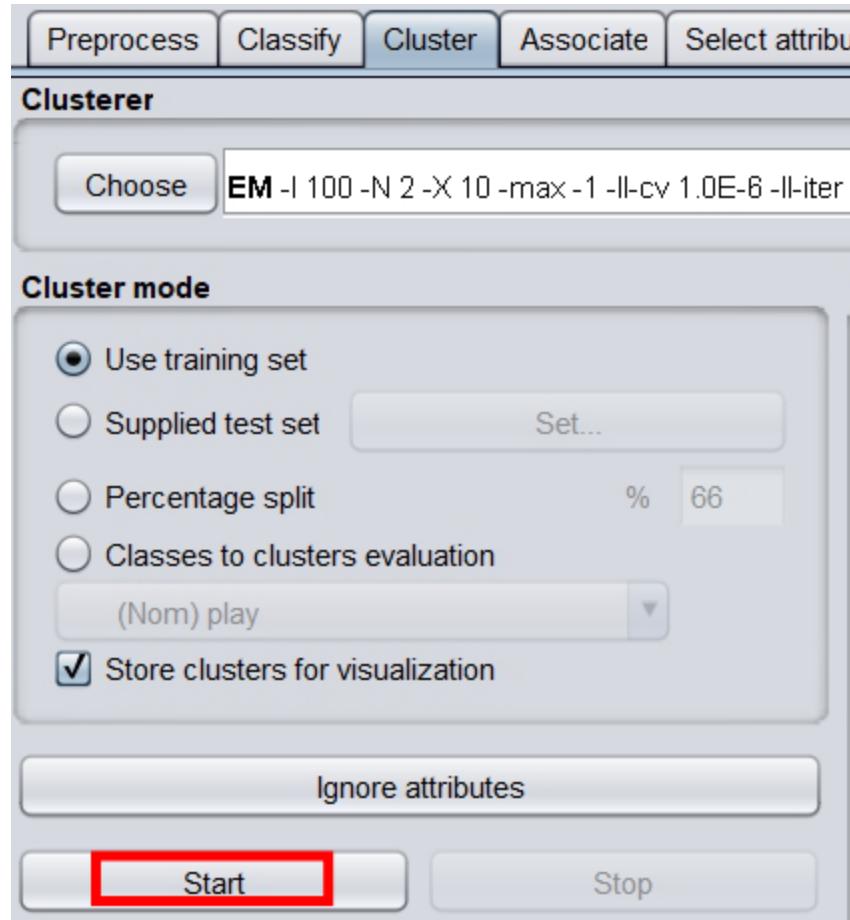
Lesson 3.5: 聚類的表達

3. 將參數numClusters的值由-1改為2，然後左鍵單擊OK按鈕。



Lesson 3.5: 聚類的表達

4. 回到Cluster面板，以左鍵單擊Start按鈕。



Lesson 3.5: 聚類的表達

執行結果：

- 得到兩個聚類。
- 聚類對應每個屬性得到概率。比如“outlook”屬性，在每個聚類中是“sunny”、“overcast”還是“rainy”的概率。

Attribute	Cluster	
	0 (0.35)	1 (0.65)
<hr/>		
outlook		
sunny	3.8732	3.1268
overcast	1.7746	4.2254
rainy	2.1889	4.8111
[total]	7.8368	12.1632
<hr/>		
temperature		
mean	76.9173	71.8054
std. dev.	5.8302	5.8566
<hr/>		
humidity		
mean	90.1132	77.1719
std. dev.	3.8066	9.1962
<hr/>		
windy		
TRUE	3.14	4.86
FALSE	3.6967	6.3033
[total]	6.8368	11.1632
<hr/>		
play		
yes	2.1227	8.8773
no	4.7141	2.2859
[total]	6.8368	11.1632
<hr/>		
Time taken to build model (full training data) : 0.1 seconds		
<hr/>		
==== Model and evaluation on training set ===		
<hr/>		
Clustered Instances		
<hr/>		
0	4 (29%)	
1	10 (71%)	

Lesson 3.5: 聚類的表達

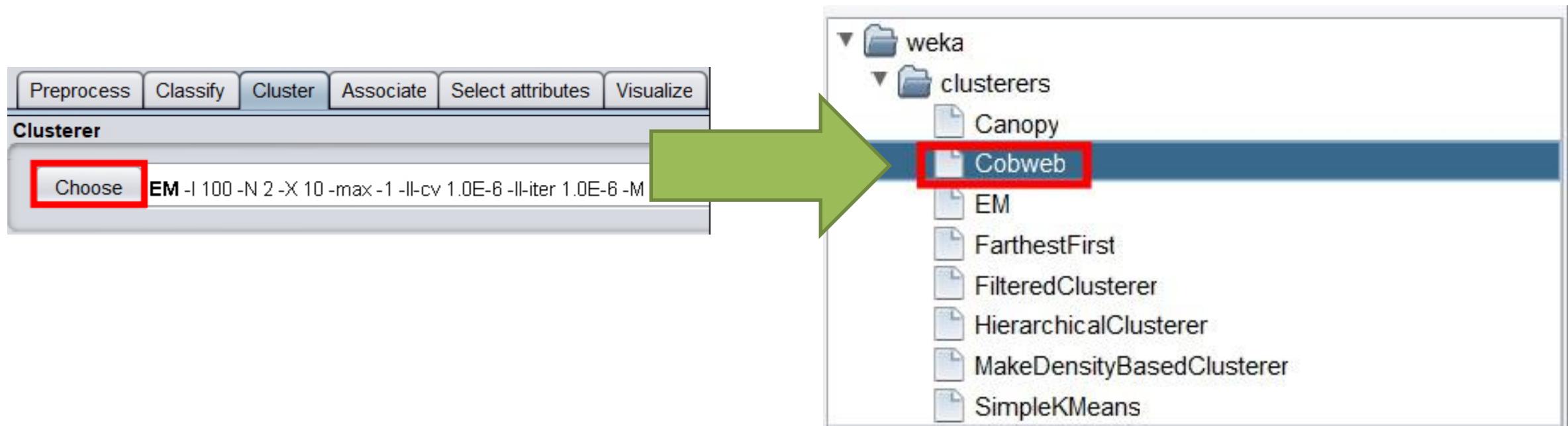
EM 聚類法(概率的(probabilistic)聚類,為期望最大化(Expectation Maximization)的縮寫

- ❖ Cluster 面板; 選擇 EM
- ❖ 將參數 numClusters 改為2 (-1為要求 EM決定聚類數)
- ❖ 注意parameters: maxIterations, minStdDev, seed (預設值為100)
 - 恢復名詞屬性
- ❖ 得到兩個聚類, 先前的概率為0.35以及0.65
- ❖ 在執行結果中:
 - 名詞屬性: 每個數值的概率
 - 數值屬性: 平均值和標準差
- ❖ 可以計算聚類成員任何實例的概率
- ❖ 整體質量測量:對數似然

Lesson 3.5: 聚類的表達

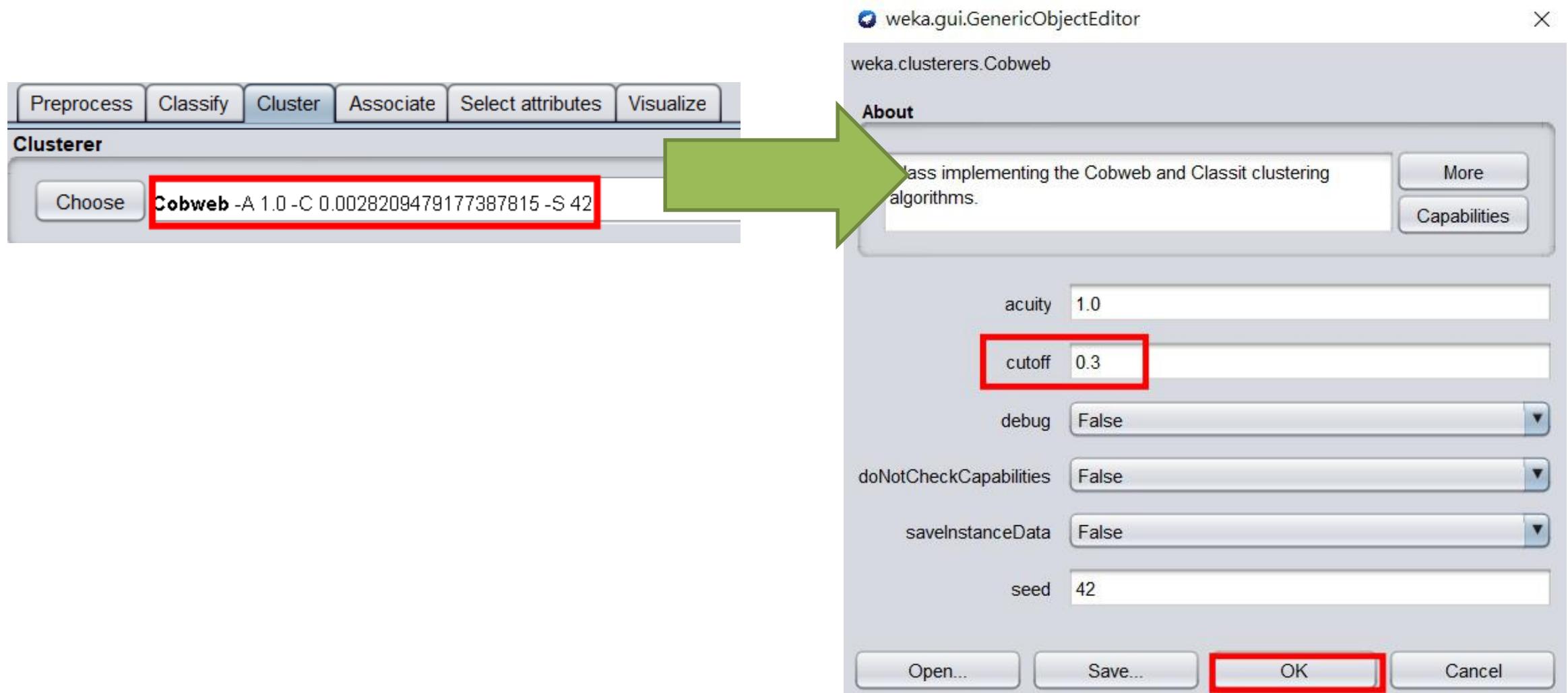
我們接著執行Cobweb聚類法。

1. 於 Cluster面板，左鍵單擊Choose按鈕後，在出現的選單中左鍵單擊weka/cluster路徑下的Cobweb聚類法。



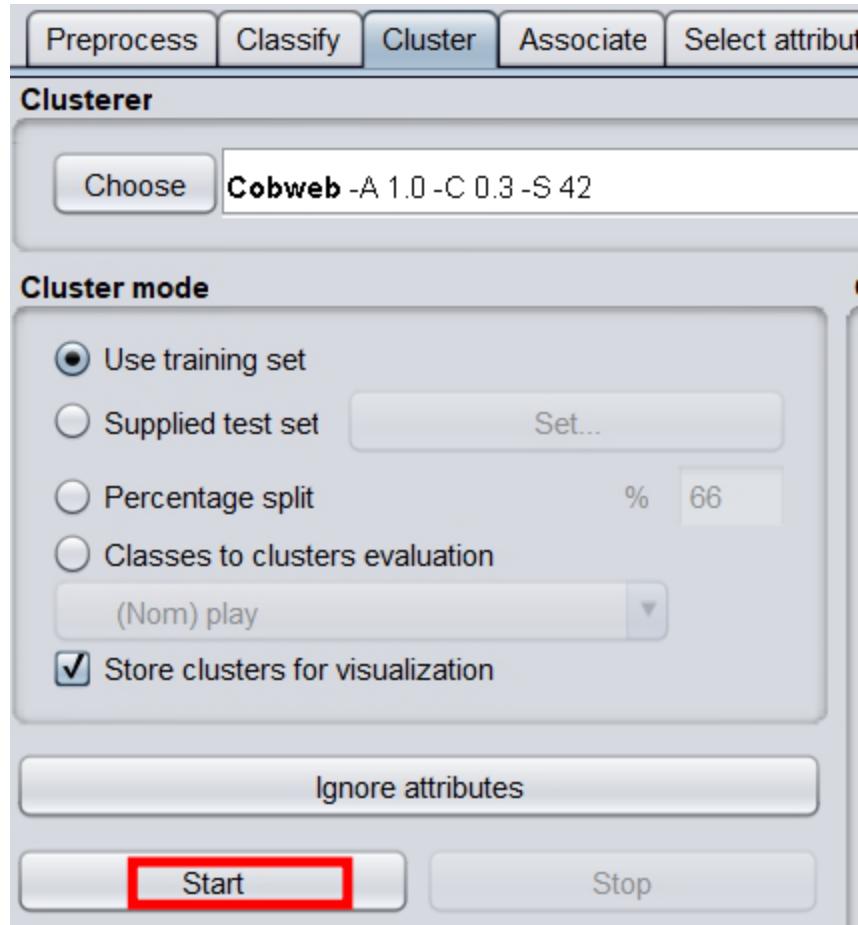
Lesson 3.5: 聚類的表達

2. 在Cluster面板，左鍵單擊聚類法名稱(左圖紅框處)，開啟配置視窗(右圖)。將參數cutoff的值設定為0.3，然後左鍵單擊OK按鈕。



Lesson 3.5: 聚類的表達

3. 回到Cluster面板，以左鍵單擊Start按鈕。



Lesson 3.5: 聚類的表達

▼ 執行結果：得到一棵樹。

The screenshot shows the Weka Clusterer interface with the Cobweb algorithm selected. The 'Cluster mode' section is set to 'Use training set'. The 'Clusterer output' window displays the hierarchical clustering structure:

```
| node 1 [5]
| | leaf 3 [1]
| node 1 [5]
| | leaf 4 [2]
| node 1 [5]
| | leaf 5 [1]
node 0 [14]
| leaf 6 [6]
node 0 [14]
| node 7 [3]
| | leaf 8 [1]
| node 7 [3]
| | leaf 9 [1]
| node 7 [3]
| | leaf 10 [1]
```

The 'Result list' window shows the following log entries:

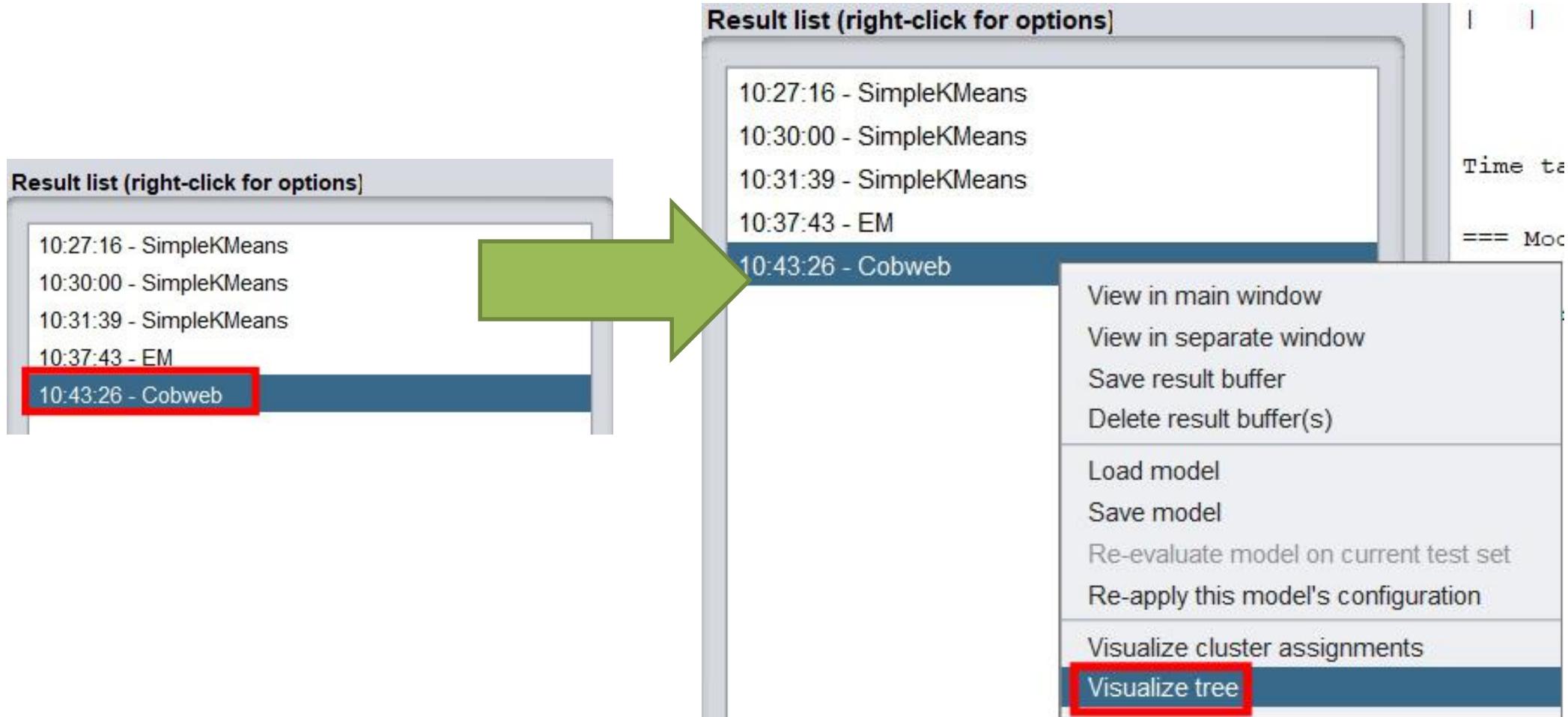
- 10:27:16 - SimpleKMeans
- 10:30:00 - SimpleKMeans
- 10:31:39 - SimpleKMeans
- 10:37:43 - EM
- 10:43:26 - Cobweb

The 'Clustered Instances' section lists the following data points:

Instance	Cluster	Percentage
2	1	7%
3	1	7%
4	2	14%
5	1	7%
6	6	43%
8	1	7%
9	1	7%
10	1	7%

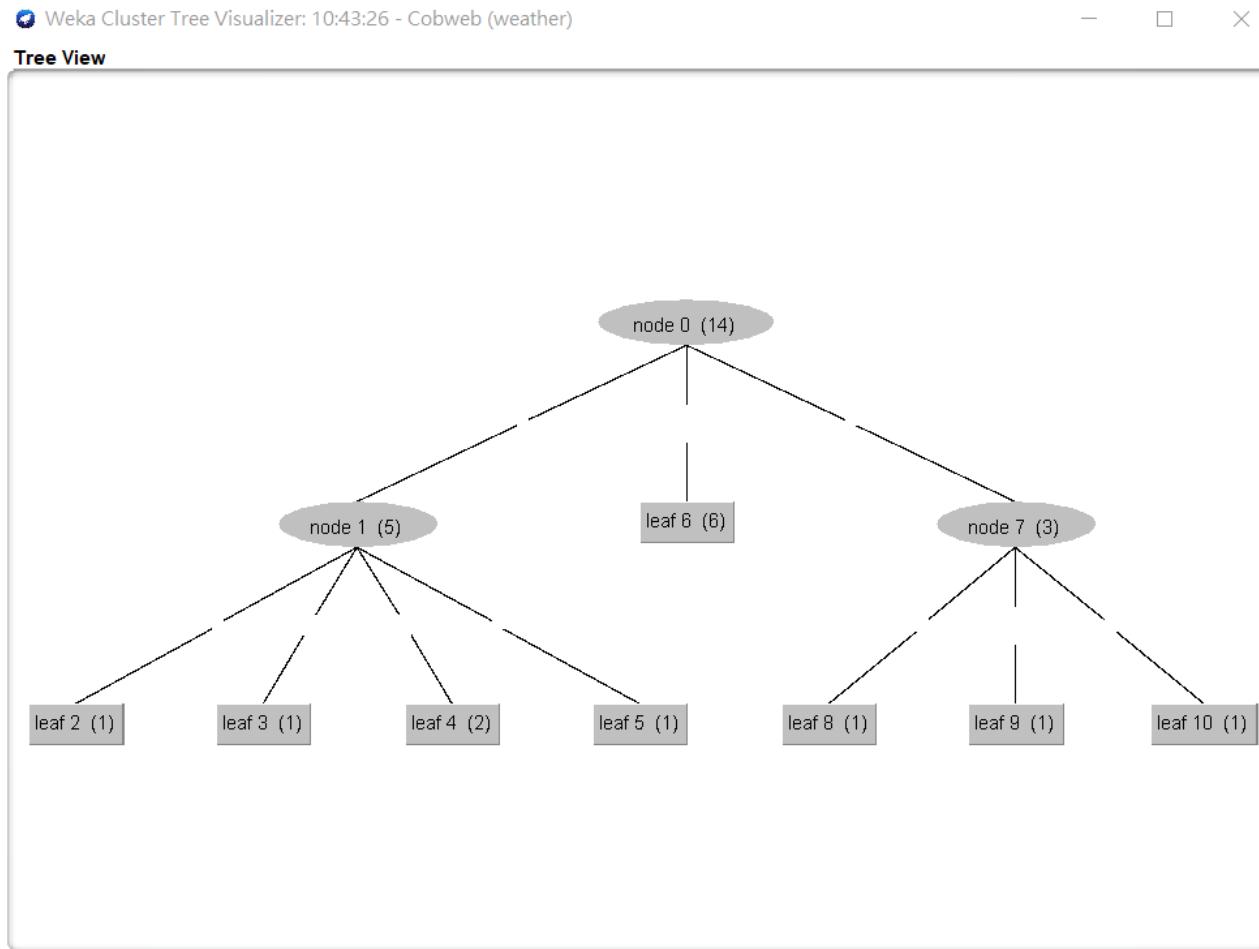
Lesson 3.5: 聚類的表達

4. 在執行結果列表中右鍵單擊剛才的Cobweb紀錄，於出現的選單中左鍵單擊Visualize tree選項。



Lesson 3.5: 聚類的表達

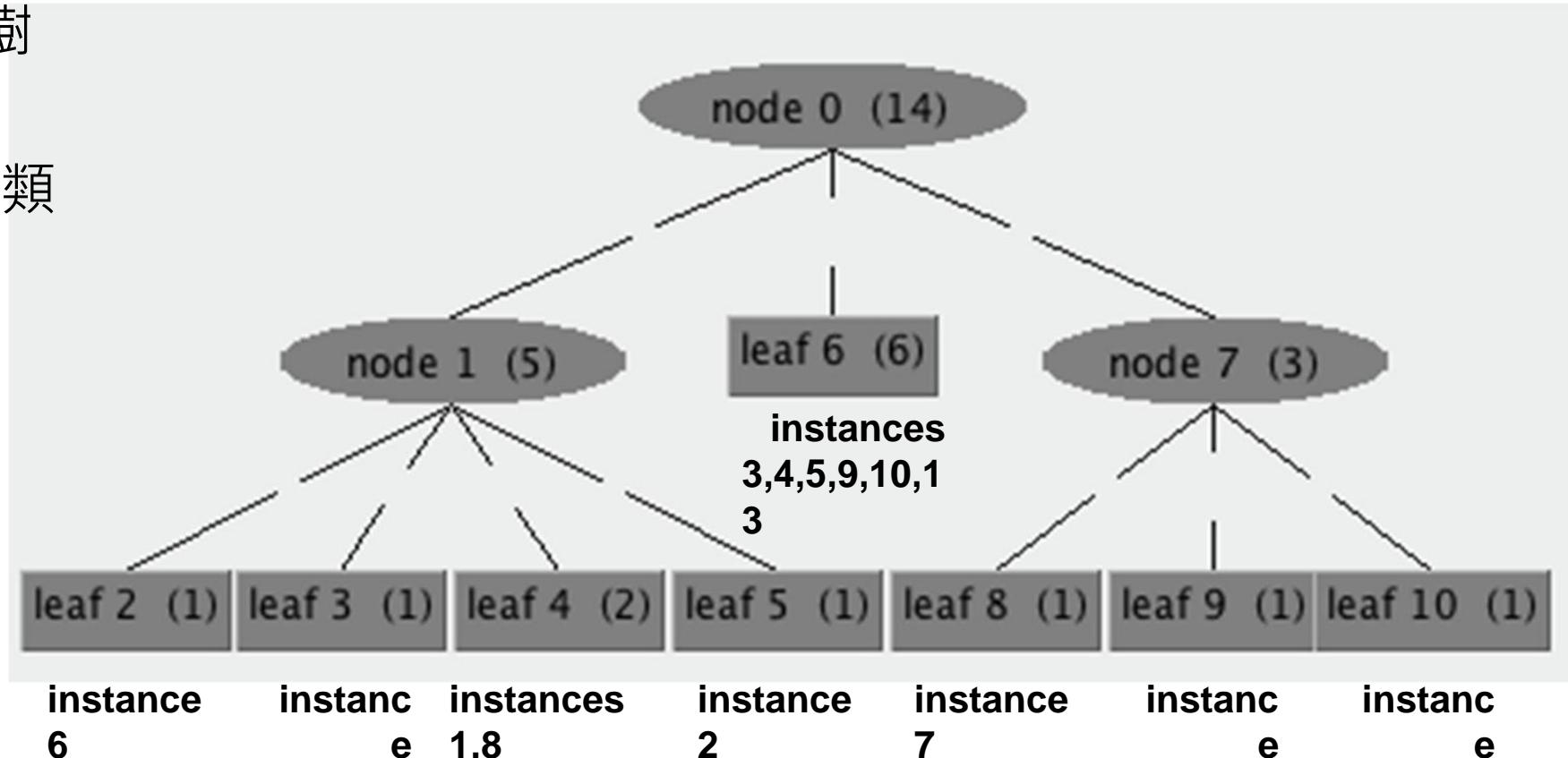
▼ 執行結果



Lesson 3.5: 聚類的表達

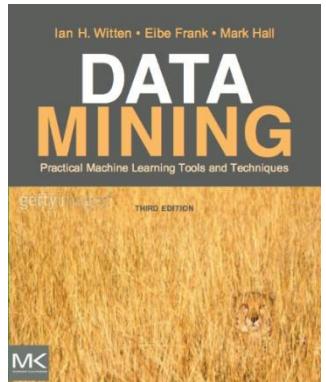
Cobweb 聚類法 (分層(hierarchical)聚類)

- ❖ Cluster 面板; 選擇 Cobweb
- ❖ 將參數Cutoff設定為0.3
- ❖ 視覺化這棵樹
- ❖ 得到10 個聚類



Lesson 3.5: 聚類的表達

- ❖ 聚類: 沒有類別屬性
 - ❖ 表達法:不相交的集合(disjoint sets),概率聚類(probabilistic),分層聚類(Hierarchical)
 - Weka中分別是 : SimpleKMeans (+XMeans), EM, Cobweb
 - ❖ Kmeans:基於距離的疊代聚類
 - ❖ 不同的距離指標
 - ❖ 評估聚類是困難的
- 課程文本
- ❖ Sections 4.8 and 6.8
Clustering





THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

More Data Mining with Weka

Class 3 – Lesson 6

聚類的評估

(*Evaluating clusters*)

Ian H. Witten

Department of Computer
Science University of Waikato
New Zealand

Lesson 3.6: 聚類的評估

Class 1 探索Weka的介面；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 3.1 決策樹與規則

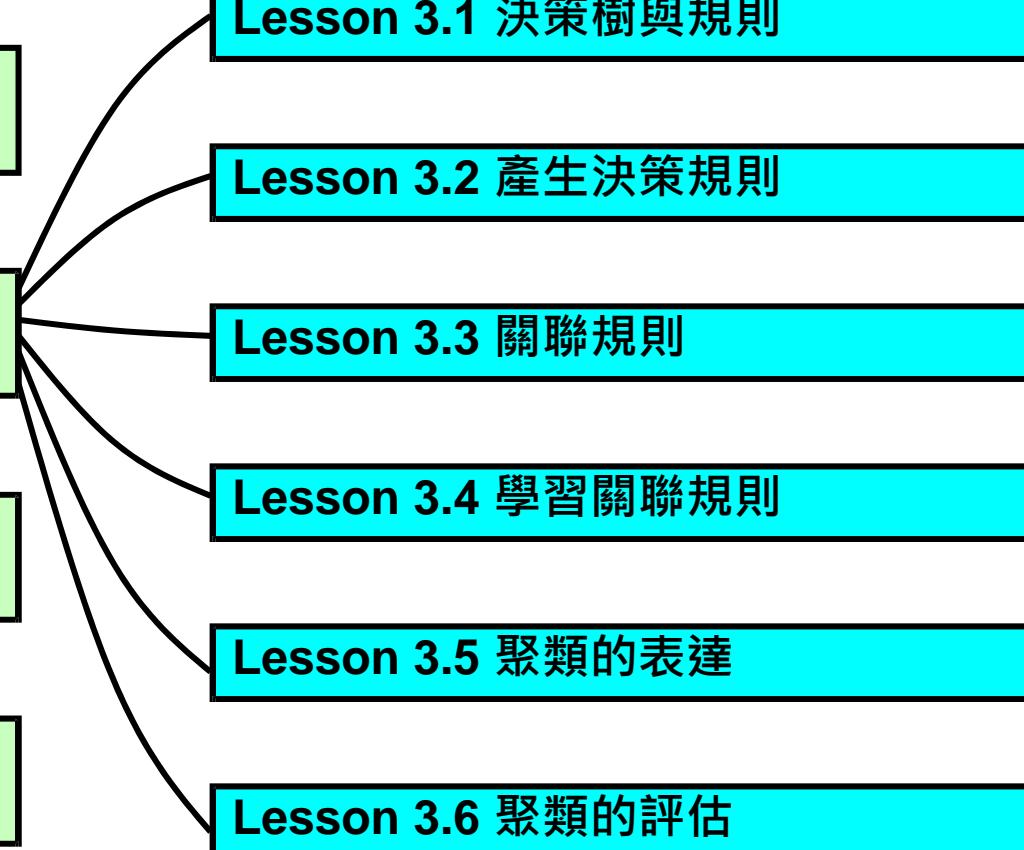
Lesson 3.2 產生決策規則

Lesson 3.3 關聯規則

Lesson 3.4 學習關聯規則

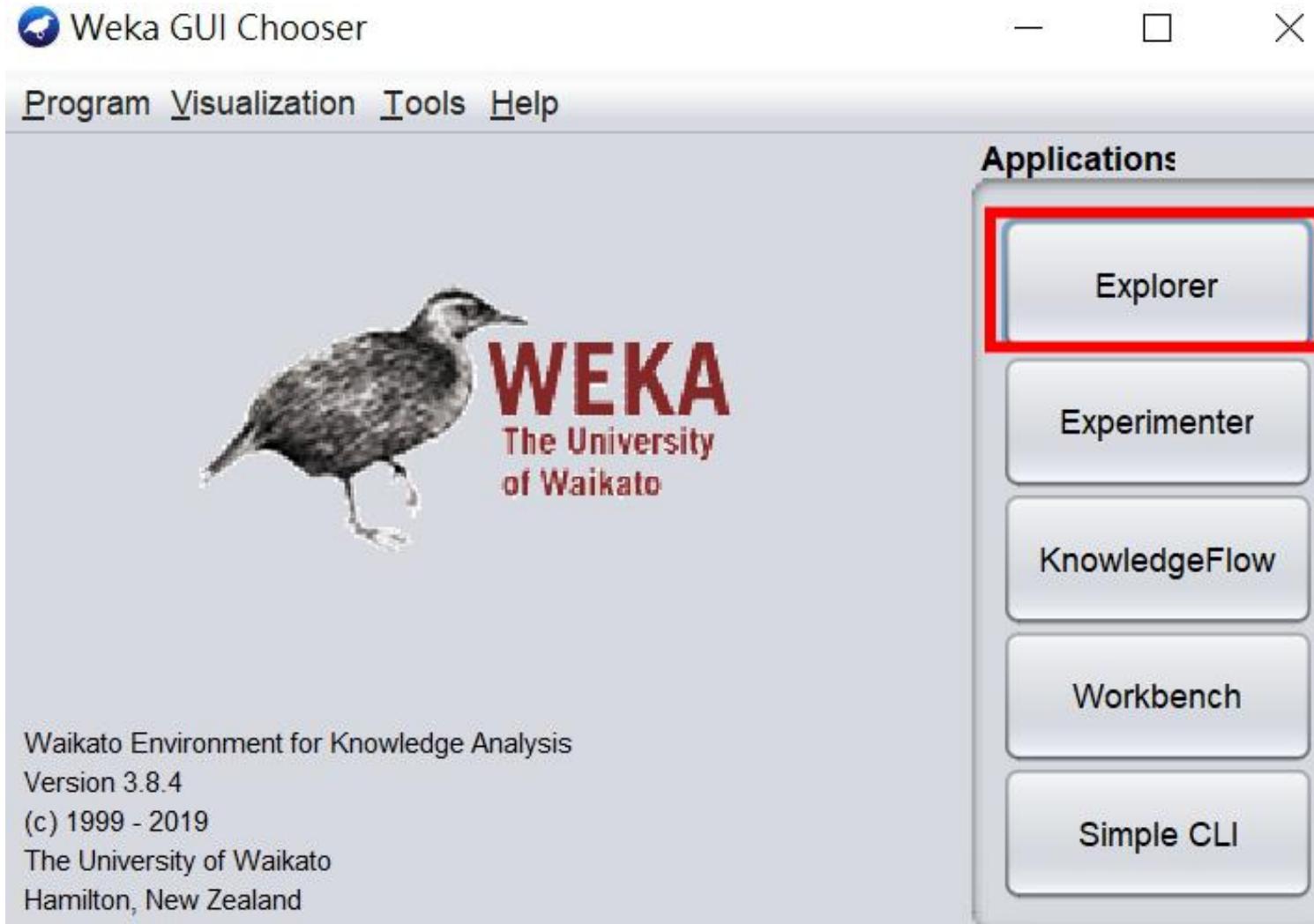
Lesson 3.5 聚類的表達

Lesson 3.6 聚類的評估



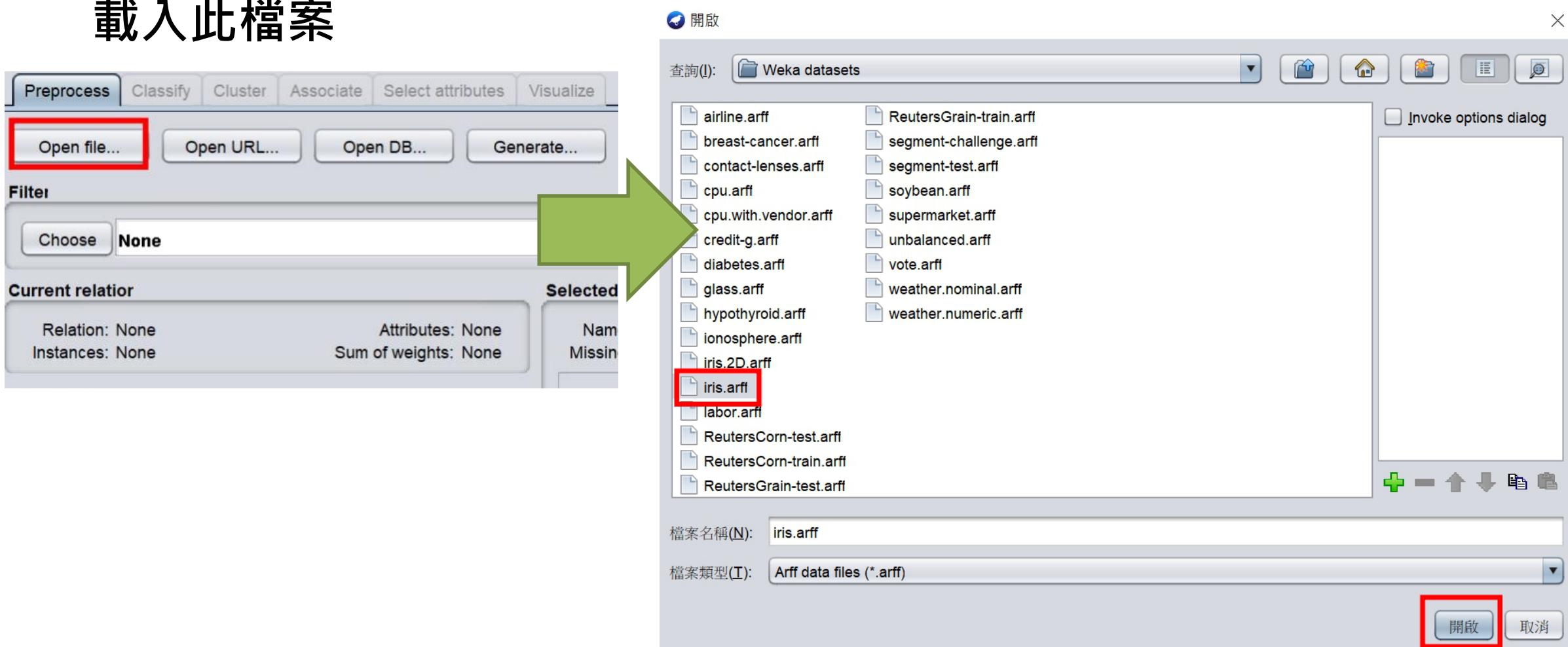
Lesson 3.6: 聚類的評估

1. 開啟Weka的Explorer



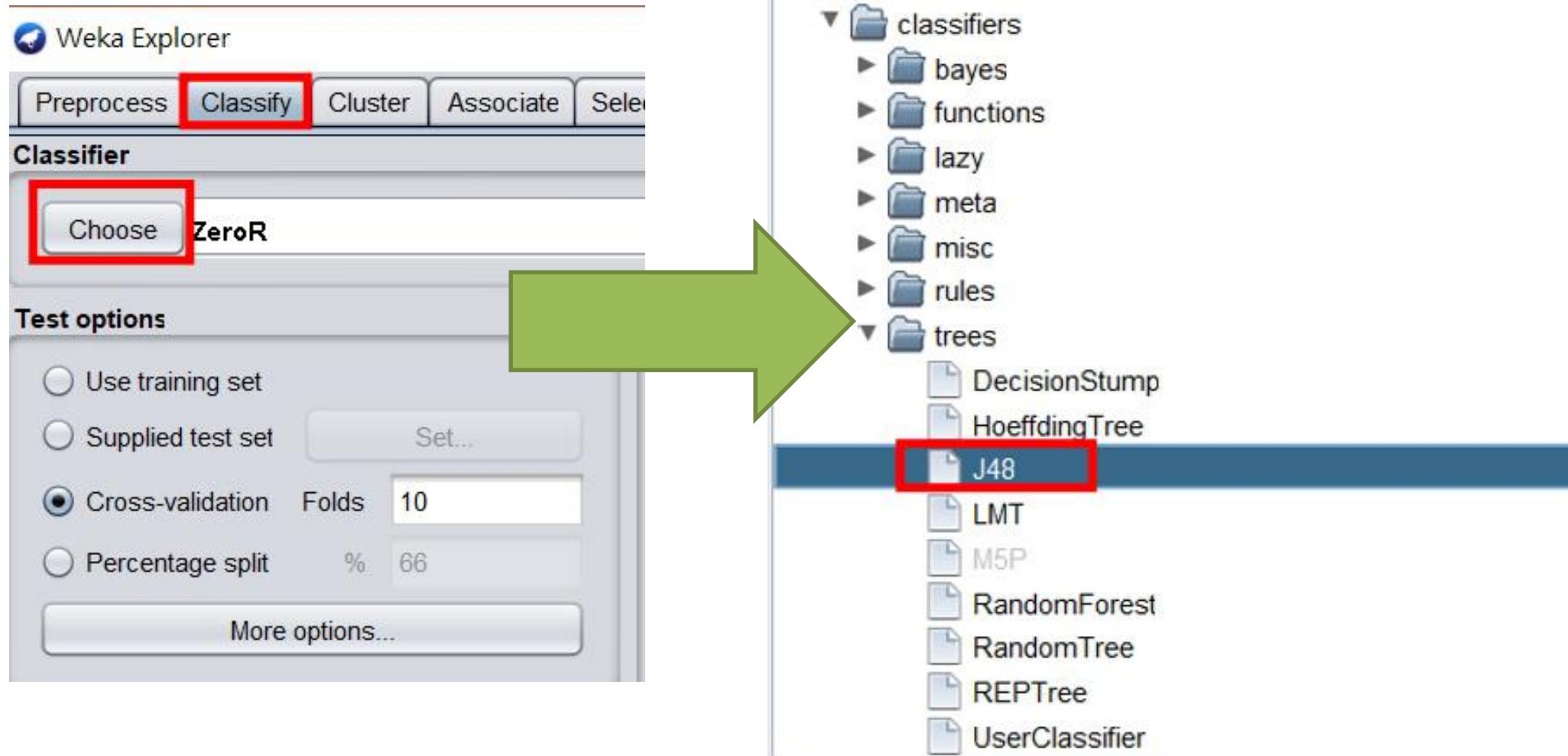
Lesson 3.6: 聚類的評估

2. 左鍵點擊Open file...開啟右圖視窗，進入自行複製的Weka datasets資料夾，左鍵單擊**iris.arff** 檔案後，再以左鍵單擊下方「開啟」按鈕以載入此檔案



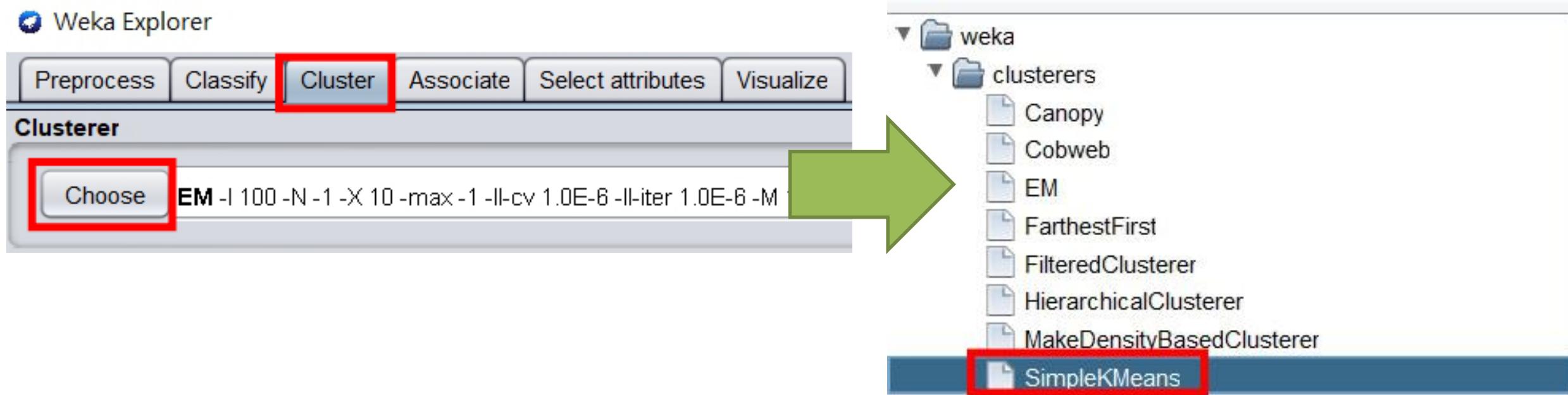
Lesson 3.6: 聚類的評估

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



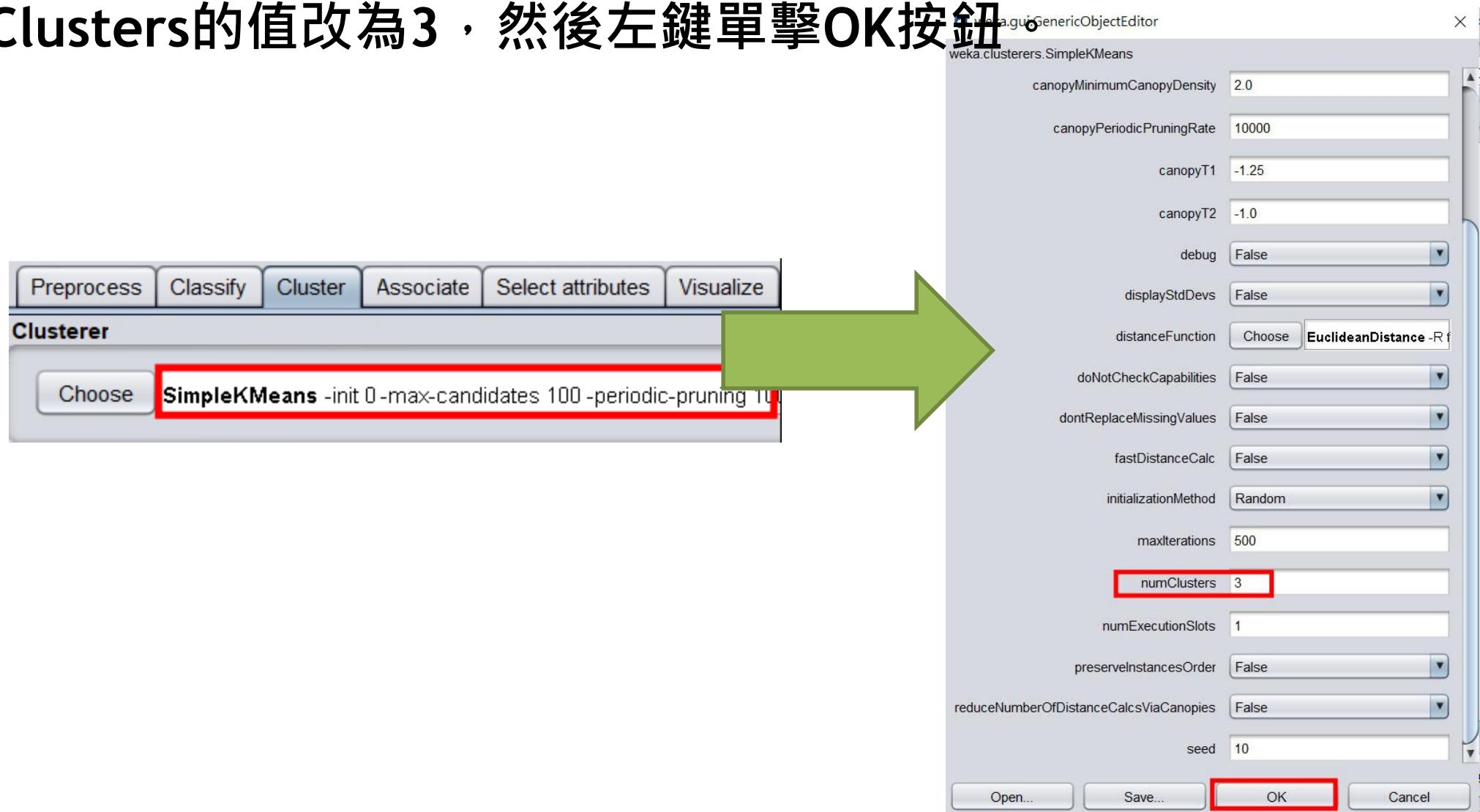
Lesson 3.6: 聚類的評估

4. 切換到Cluster面板，左鍵單擊Choose按鈕後，在出現的選單中左鍵單擊weka/cluster路徑下的SimpleKMeans聚類法。



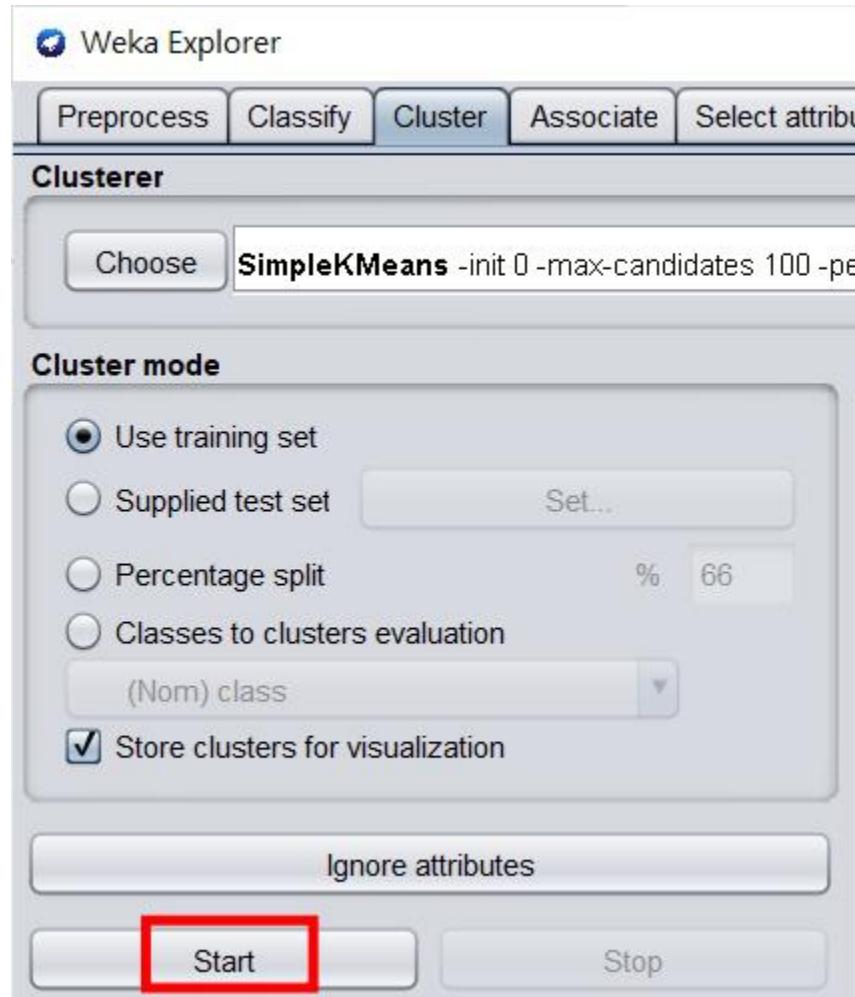
Lesson 3.6: 聚類的評估

5. 左鍵單擊分類器名稱(左圖紅框處)，開啟配置視窗(右圖)。接著將參數numClusters的值改為3，然後左鍵單擊OK按鈕。



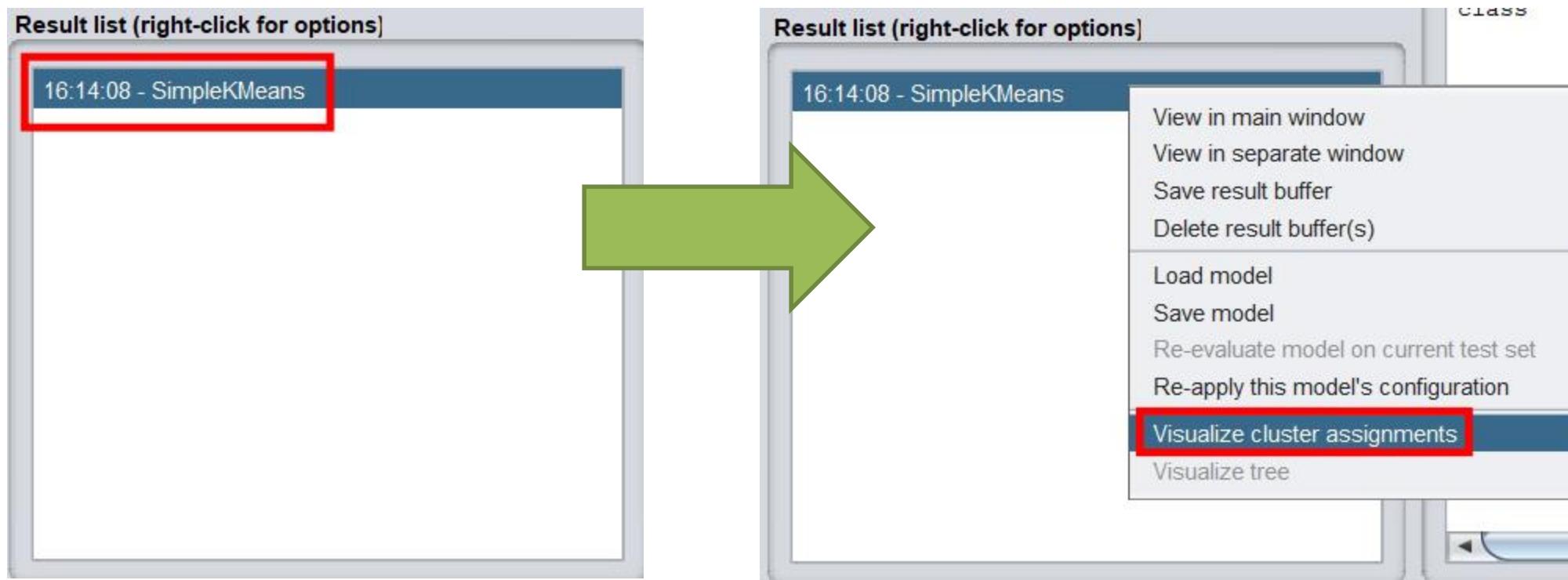
Lesson 3.6: 聚類的評估

6. 回到Cluster面板，左鍵單擊Start按鈕。



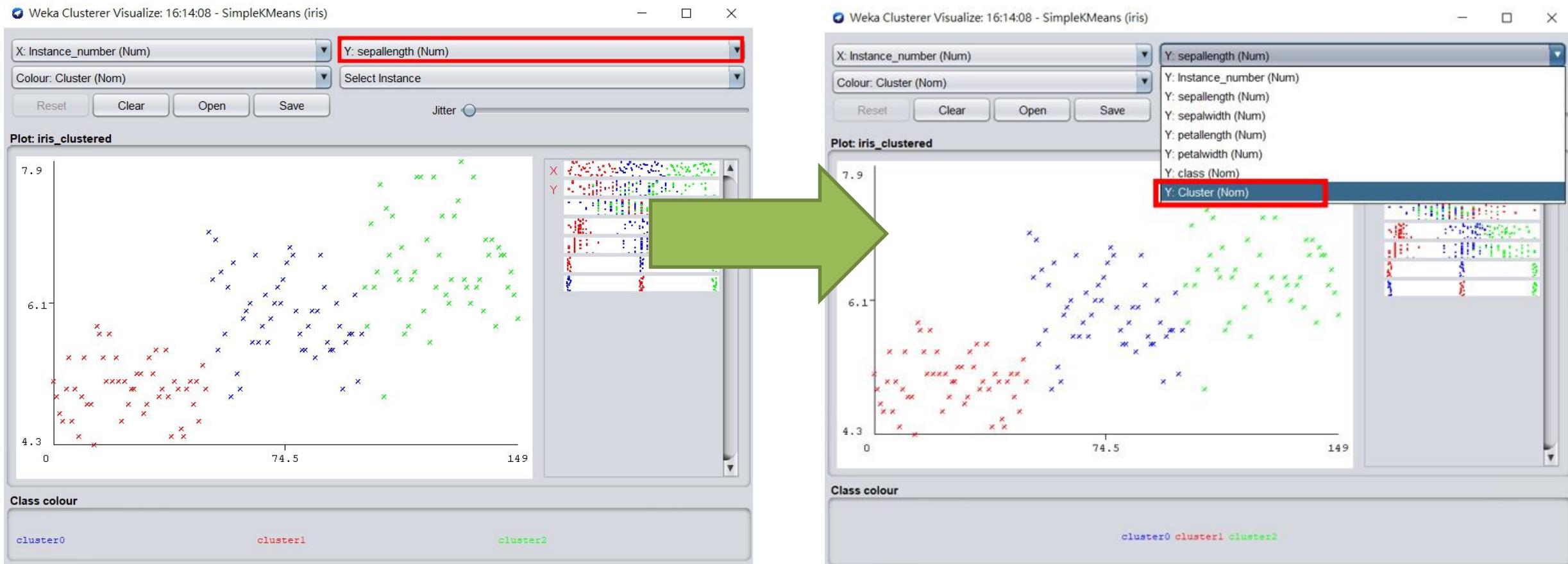
Lesson 3.6: 聚類的評估

7. 在執行結果列表中右鍵單擊剛才的執行紀錄，於出現的選單中左鍵單擊**Visualize cluster assignments**選項。



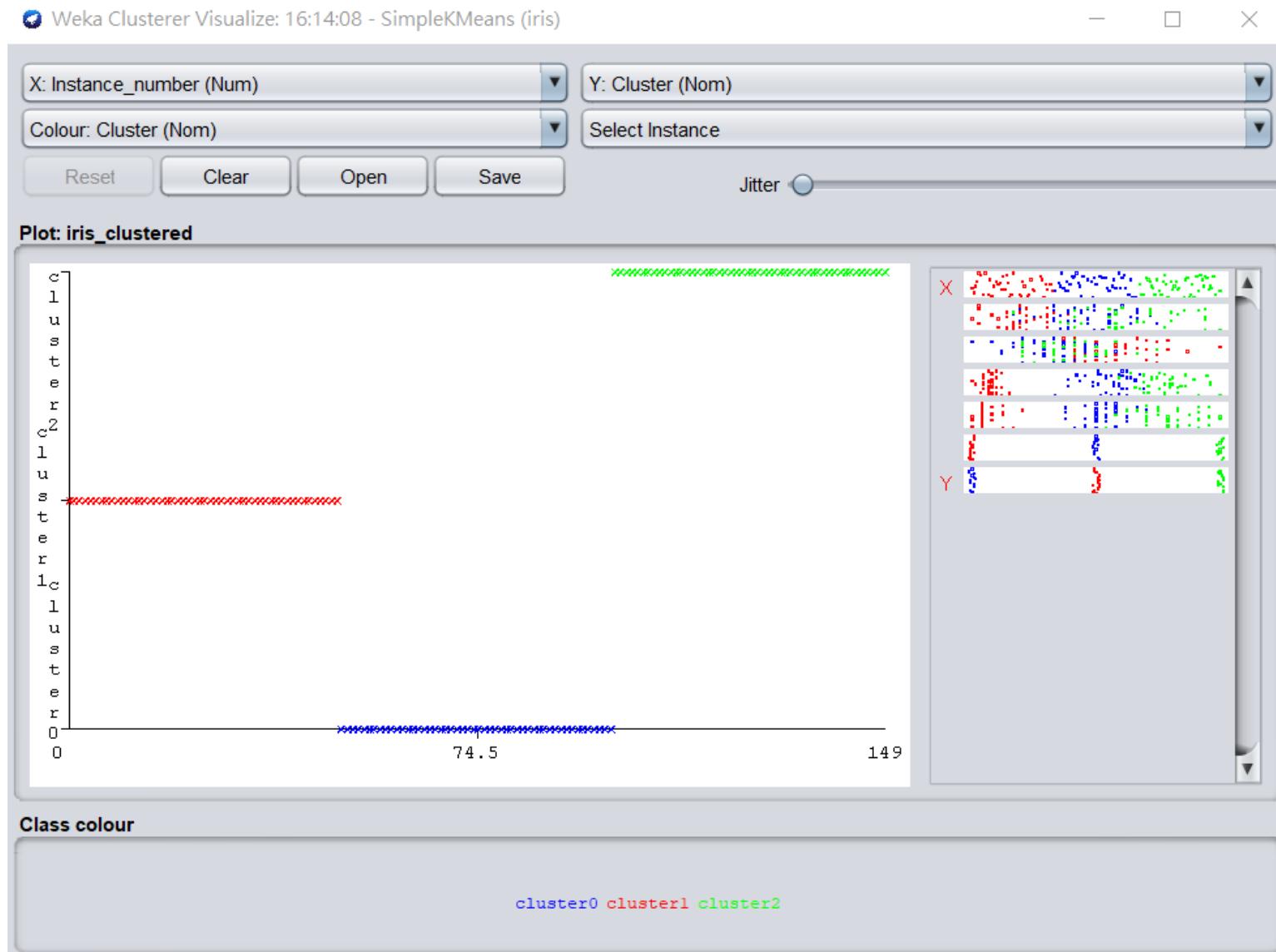
Lesson 3.6: 聚類的評估

8. 左鍵單擊Y軸代表項目(左圖紅框處)，接著在出現的選單中左鍵單擊 Y:Cluster (Nom) 選項。



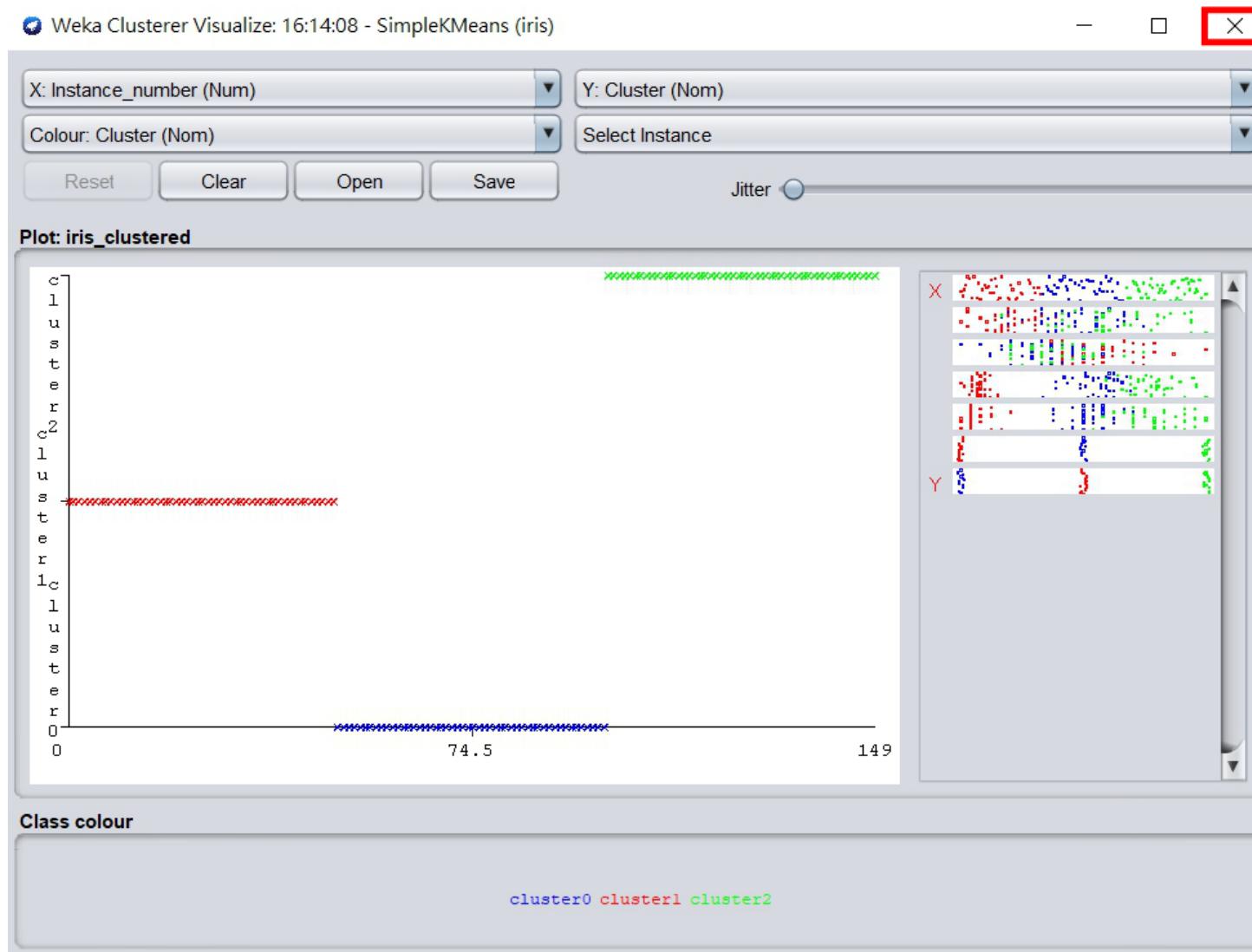
Lesson 3.6: 聚類的評估

- 結果的前50個實例是一類鳶尾花(圖中紅色部分)，中間50個是一類(圖中藍色部分)，後50個是另一類(圖中綠色部分)。
- 表現得太過於完美。如果一件東西在資料探勘中看上去過於完美，往往不可信。問題出在資料中有一個屬性叫做「類別」，進行聚類時不應該將類屬性包含進去。在Cluster面板中，我們可以做忽略某個屬性的操作。



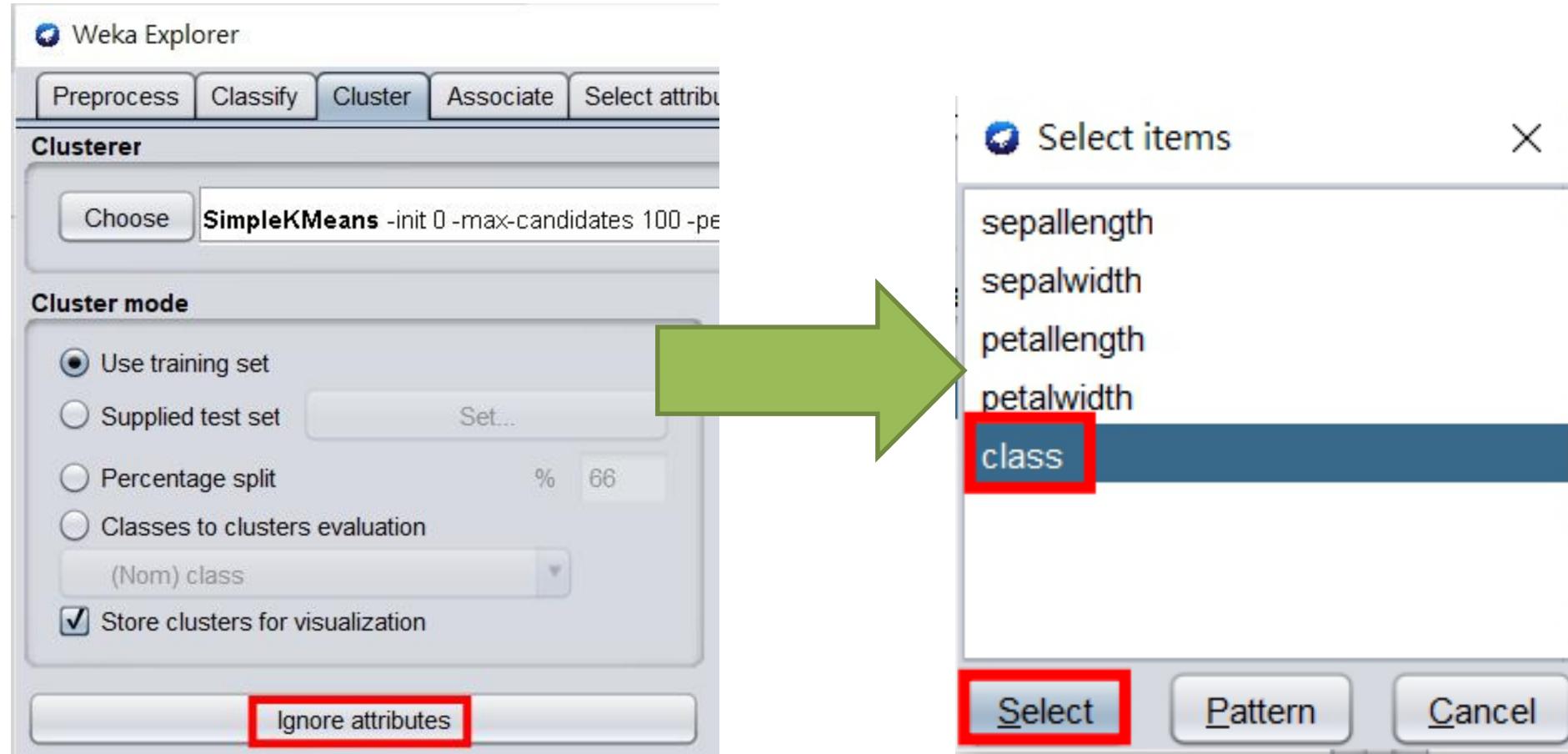
Lesson 3.6: 聚類的評估

9. 左鍵單擊視窗右上方的關閉按鈕回到Cluster面板。



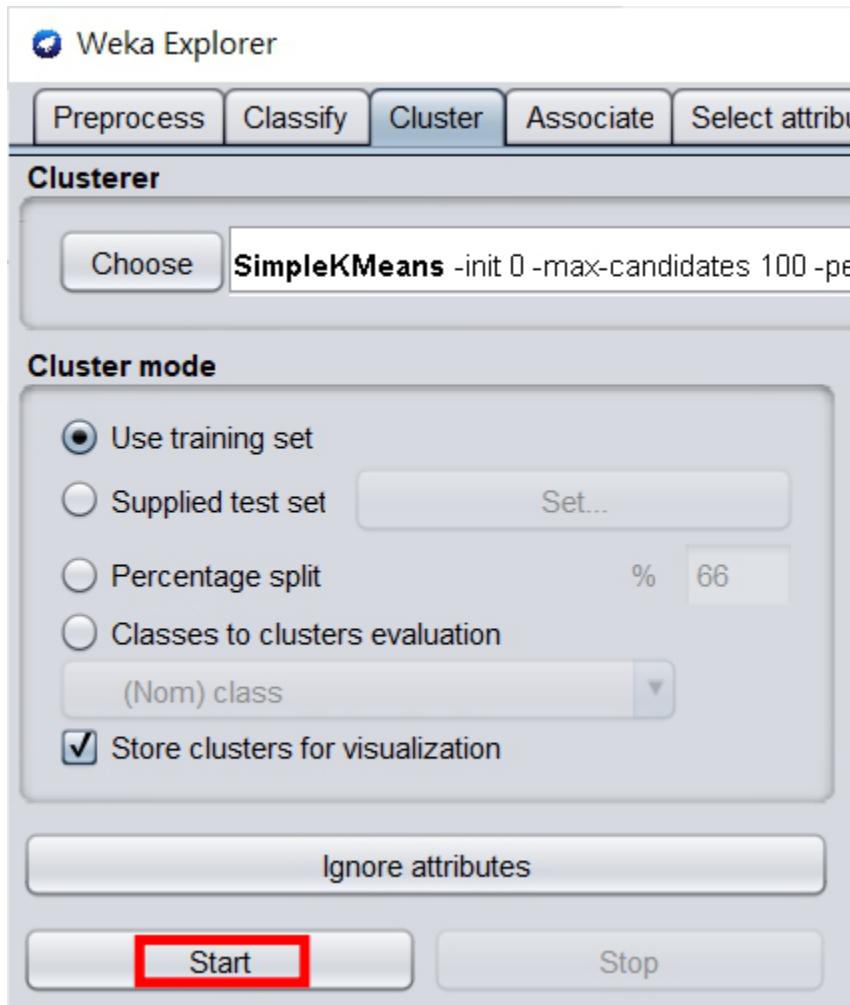
Lesson 3.6: 聚類的評估

10. 在Cluster面板中左鍵單擊Ignore attributes按鈕，並於出現的選單中，左鍵單擊class選項然後左鍵單擊Select按鈕。



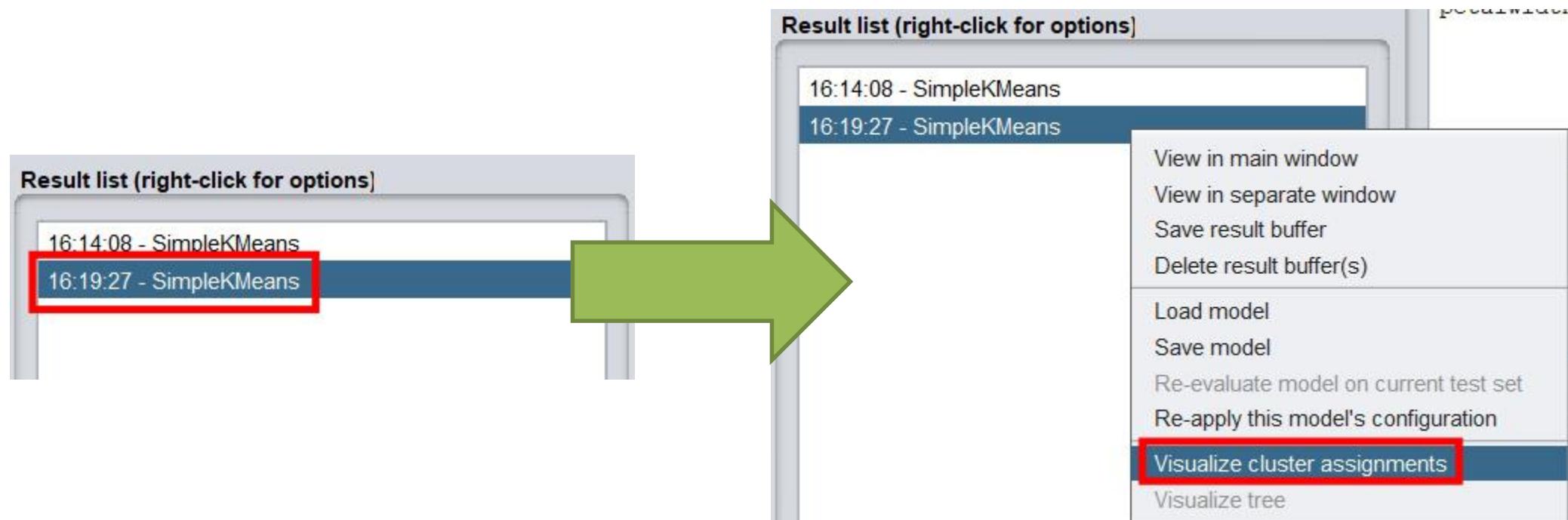
Lesson 3.6: 聚類的評估

11. 左鍵單擊Start按鈕再次運行。



Lesson 3.6: 聚類的評估

12. 在執行結果列表中右鍵單擊剛才的執行紀錄，於出現的選單中左鍵單擊**Visualize cluster assignments**選項。



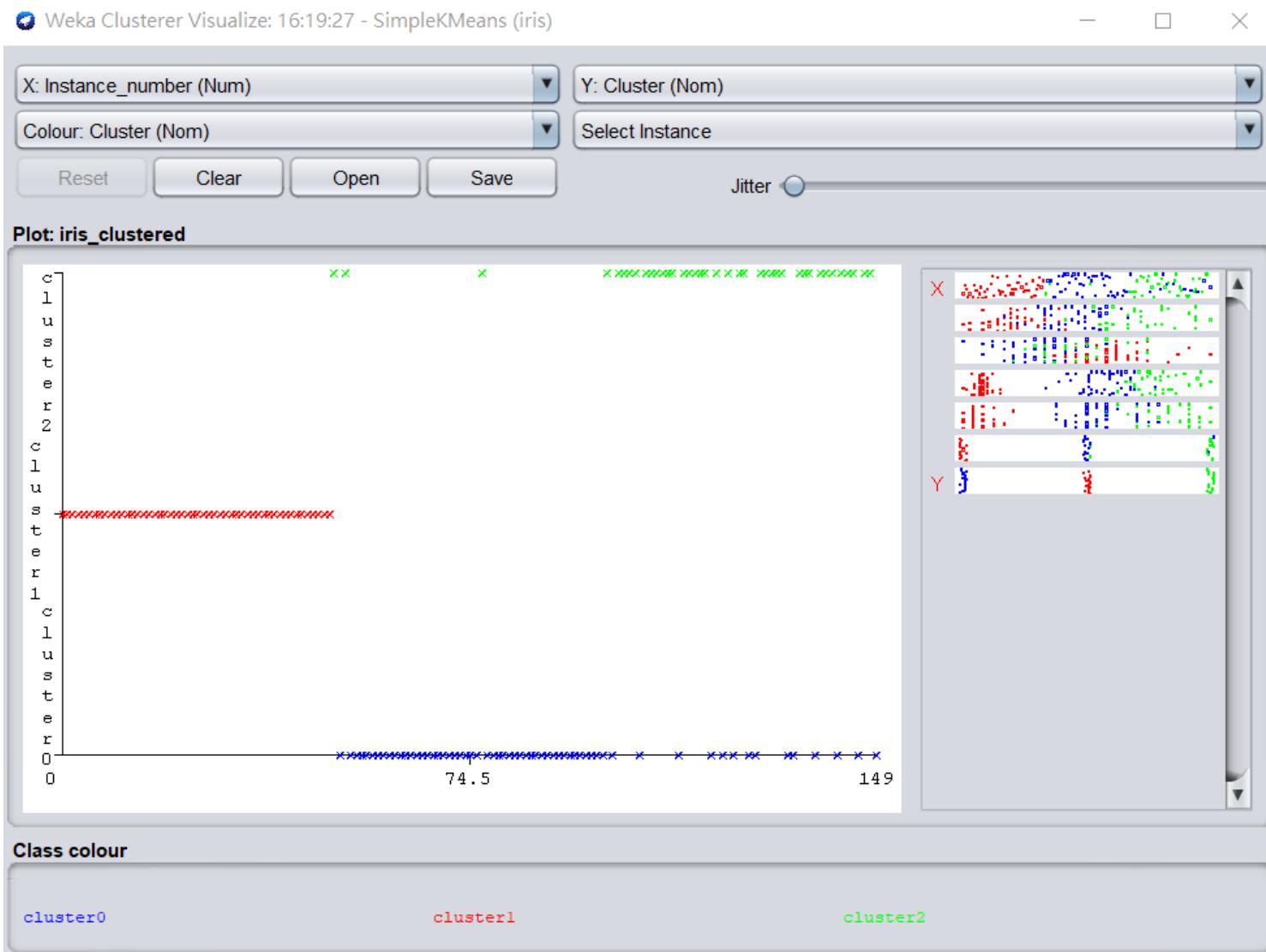
Lesson 3.6: 聚類的評估

13. 左鍵單擊Y軸代表項目(左圖紅框處)，接著在出現的選單中左鍵單擊Y:Cluster (Nom)選項。



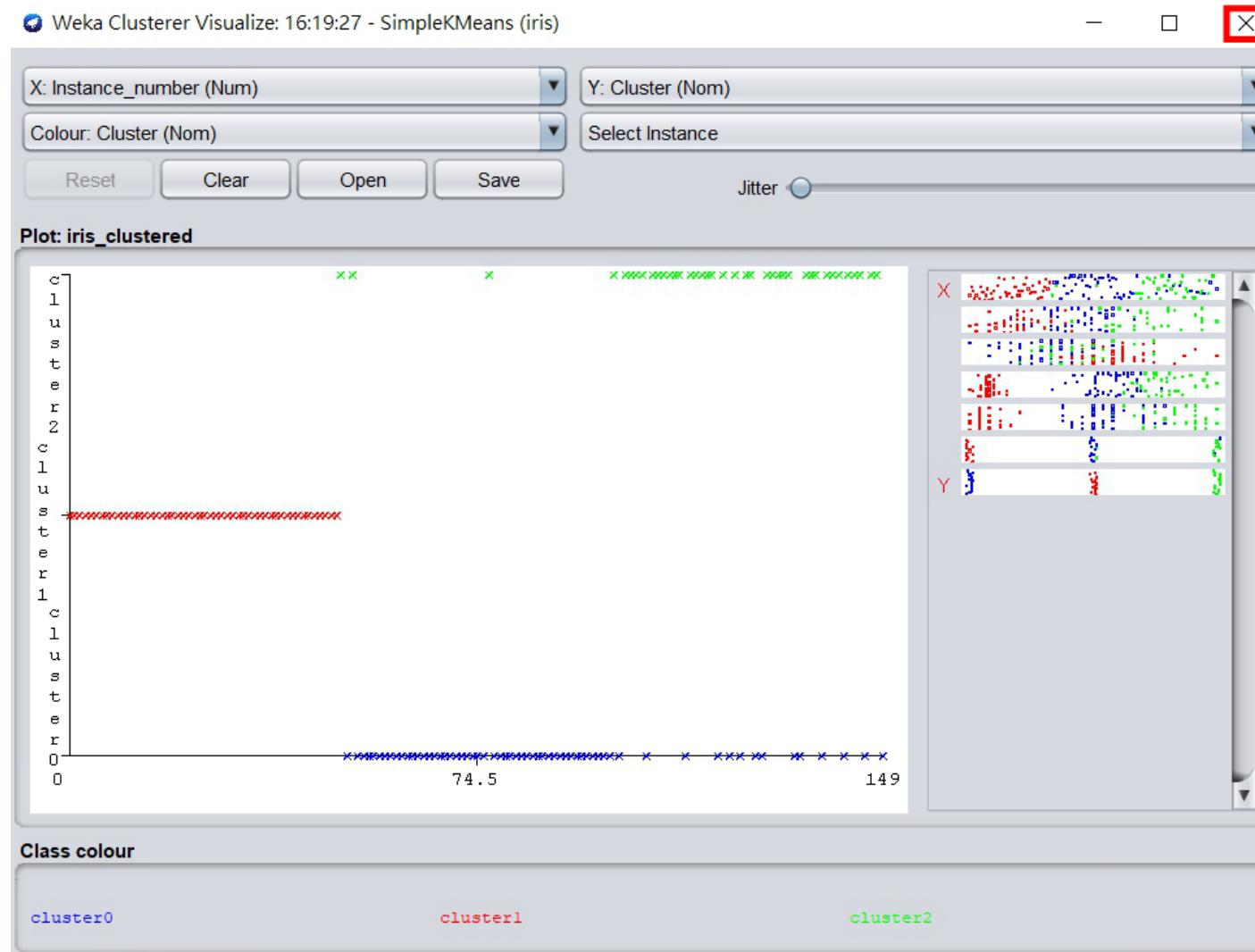
Lesson 3.6: 聚類的評估

- 我們可以看到紅色聚類看上去很不錯。
- 但是出現一些誤差：
 - 有幾個綠色的實例被分到了第二個聚類。
 - 有不少藍色的實例被分到了第三個聚類。



Lesson 3.6: 聚類的評估

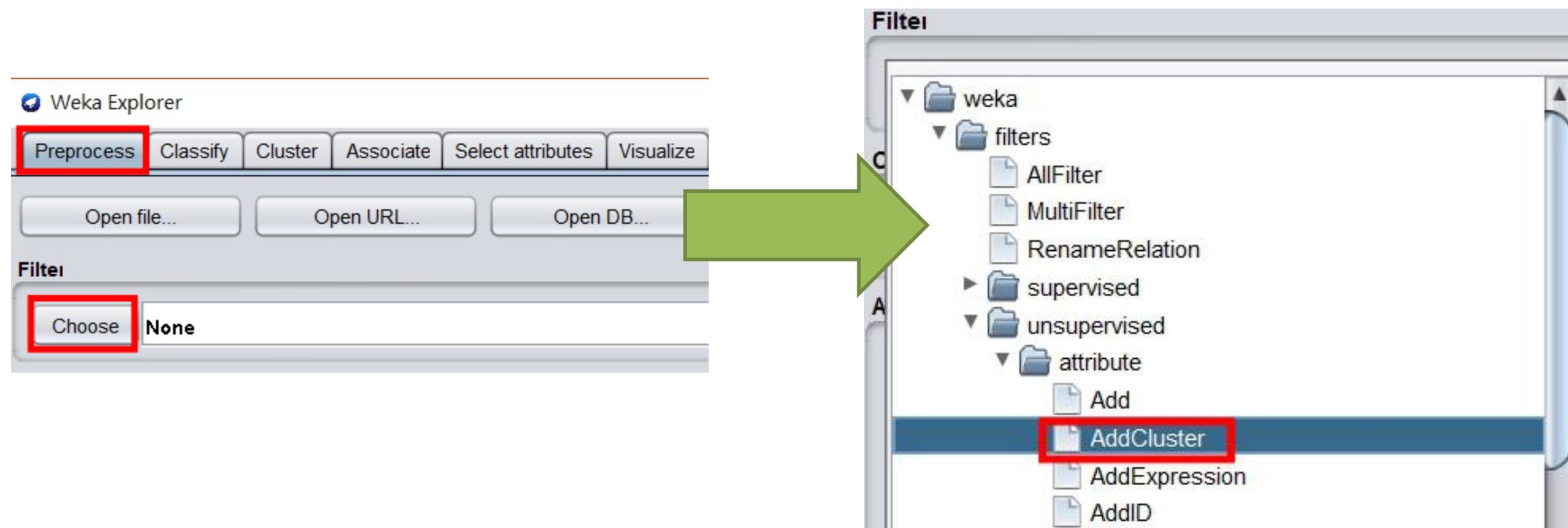
14. 左鍵單擊視窗右上方的關閉按鈕回到Cluster面板。



Lesson 3.6: 聚類的評估

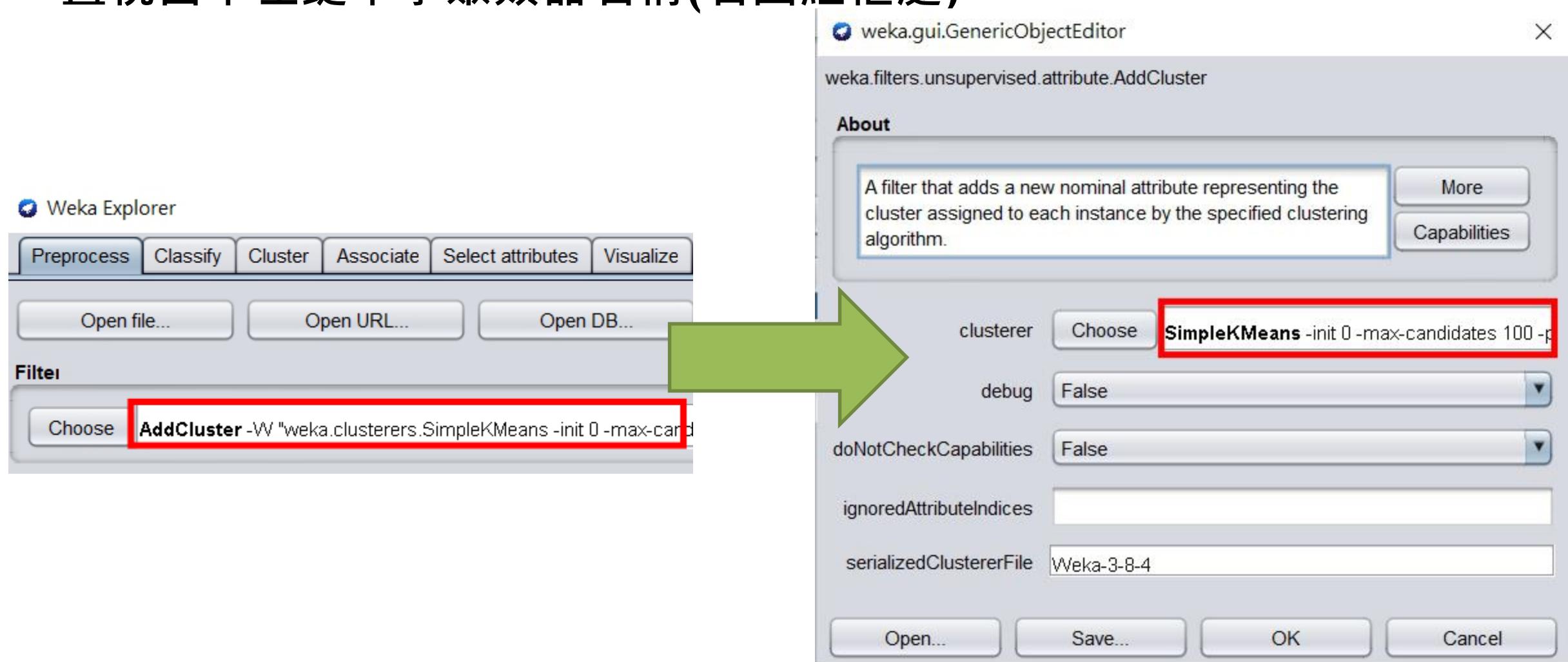
接著，我們使用AddCluster過濾器指定聚類器。

15. 切換到Preprocess面板，左鍵單擊Choose按鈕，並在出現的選單中左鍵單擊AddCluster過濾器。



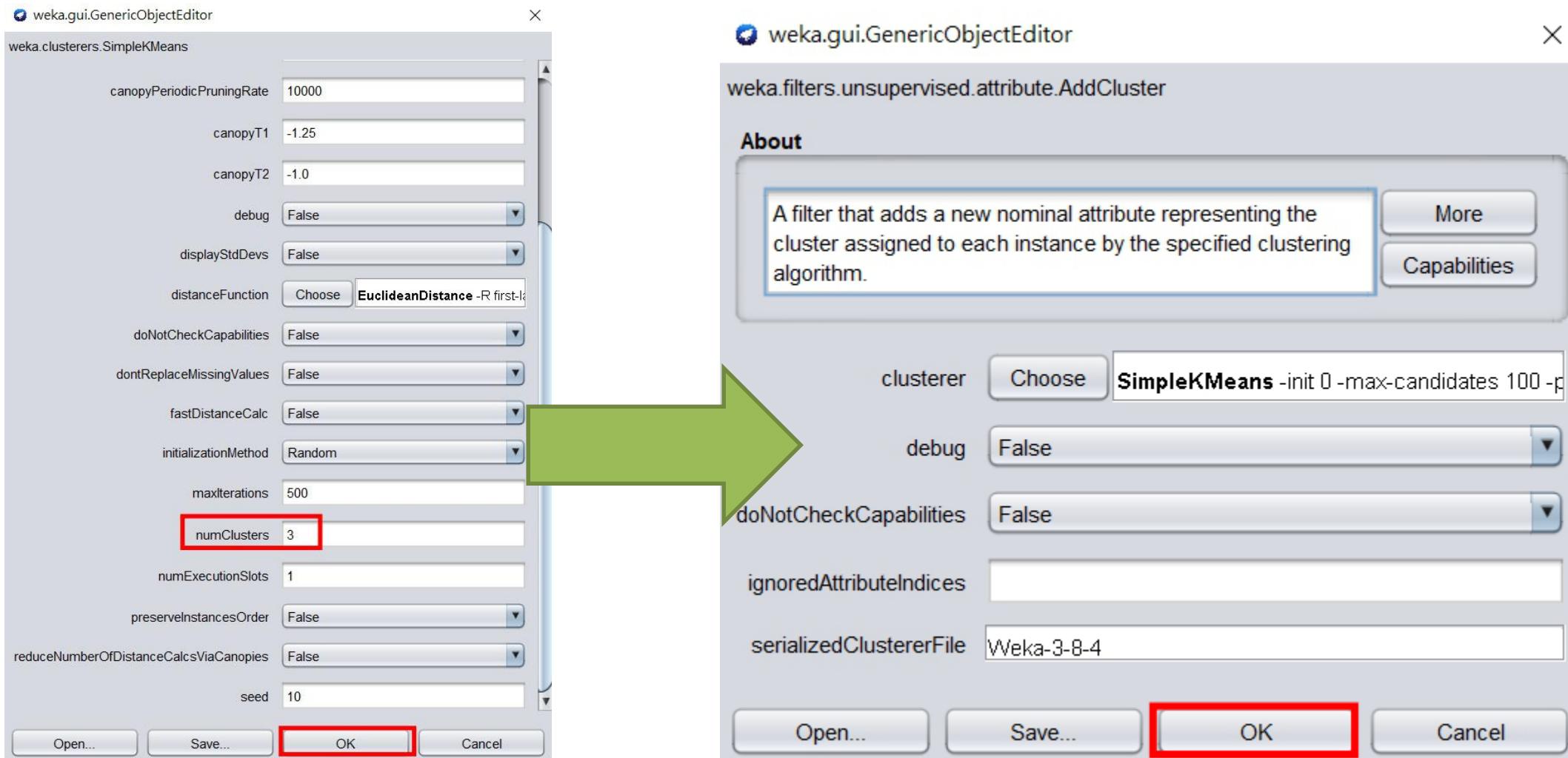
Lesson 3.6: 聚類的評估

16. 左鍵單擊過濾器名稱(左圖紅框處)，開啟配置視窗(右圖)。並在配置視窗中左鍵單擊聚類器名稱(右圖紅框處)。



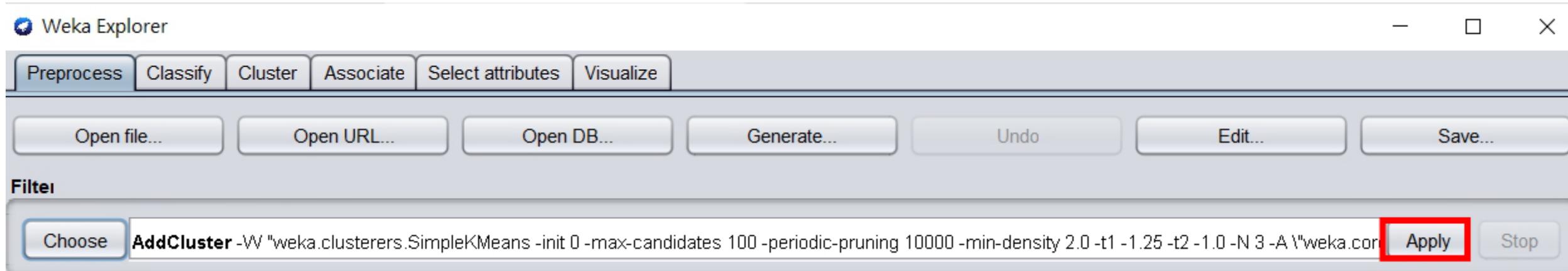
Lesson 3.6: 聚類的評估

17. 將參數numClusters設定為3，並以左鍵單擊OK按鈕回到過濾器配置視窗，再以左鍵單擊OK按鈕回到Preprocess面板。



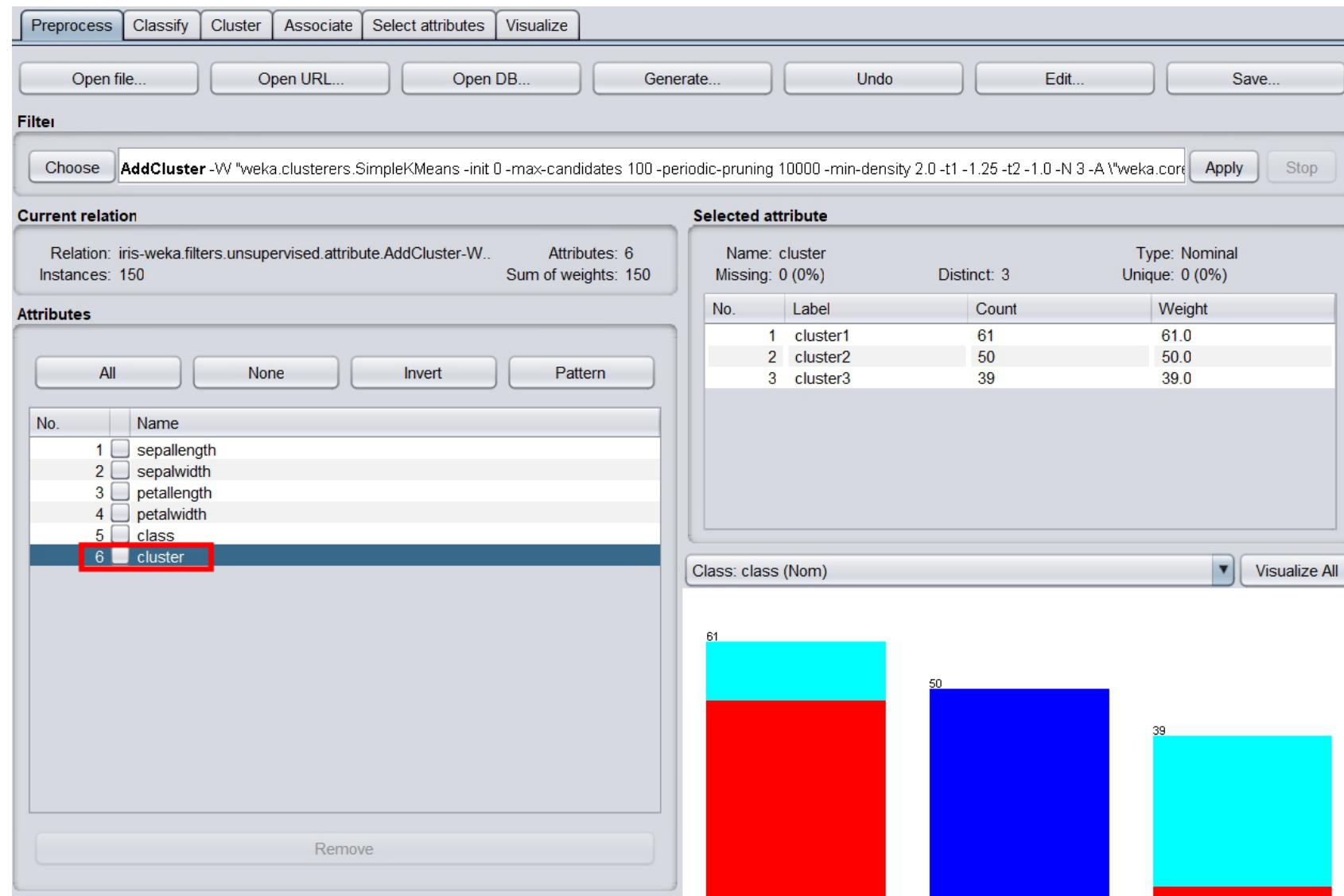
Lesson 3.6: 聚類的評估

18. 左鍵單擊Apply按鈕套用過濾器。



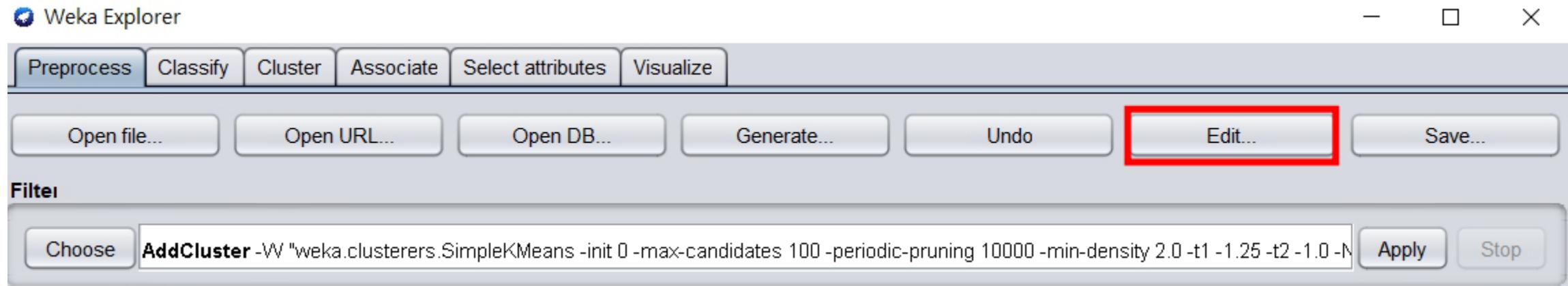
Lesson 3.6: 聚類的評估

▼我們得到了一個新的屬性，叫做“cluster”(第6個屬性)。



Lesson 3.6: 聚類的評估

19. 左鍵單擊Edit開啟編輯視窗查看。



Lesson 3.6: 聚類的評估

視覺化聚類

- ❖ Iris 資料, SimpleKMeans, 指定使用 3 個聚類
共3個聚類分別包含50個實例
- ❖ Visualize cluster assignments選項 (右鍵選單)
繪製聚類-實例數圖表以查看有哪些誤差
- ❖ 結果完美嗎? – 肯定不是!
忽略類別屬性; 3 個聚類, 分別有61, 50, 39個實例

一個聚類包含哪些實例？

- ❖ 使用了AddCluster非監督式屬性分類器
- ❖ 試著使用SimpleKMeans; Apply 以及點擊Edit

Lesson 3.6: 聚類的評估

接著我們操作類別-聚類評估(classes-to-clusters evaluation)。

1. 左鍵單擊視窗右上方的關閉按鈕回到Preprocess面板。

Viewer

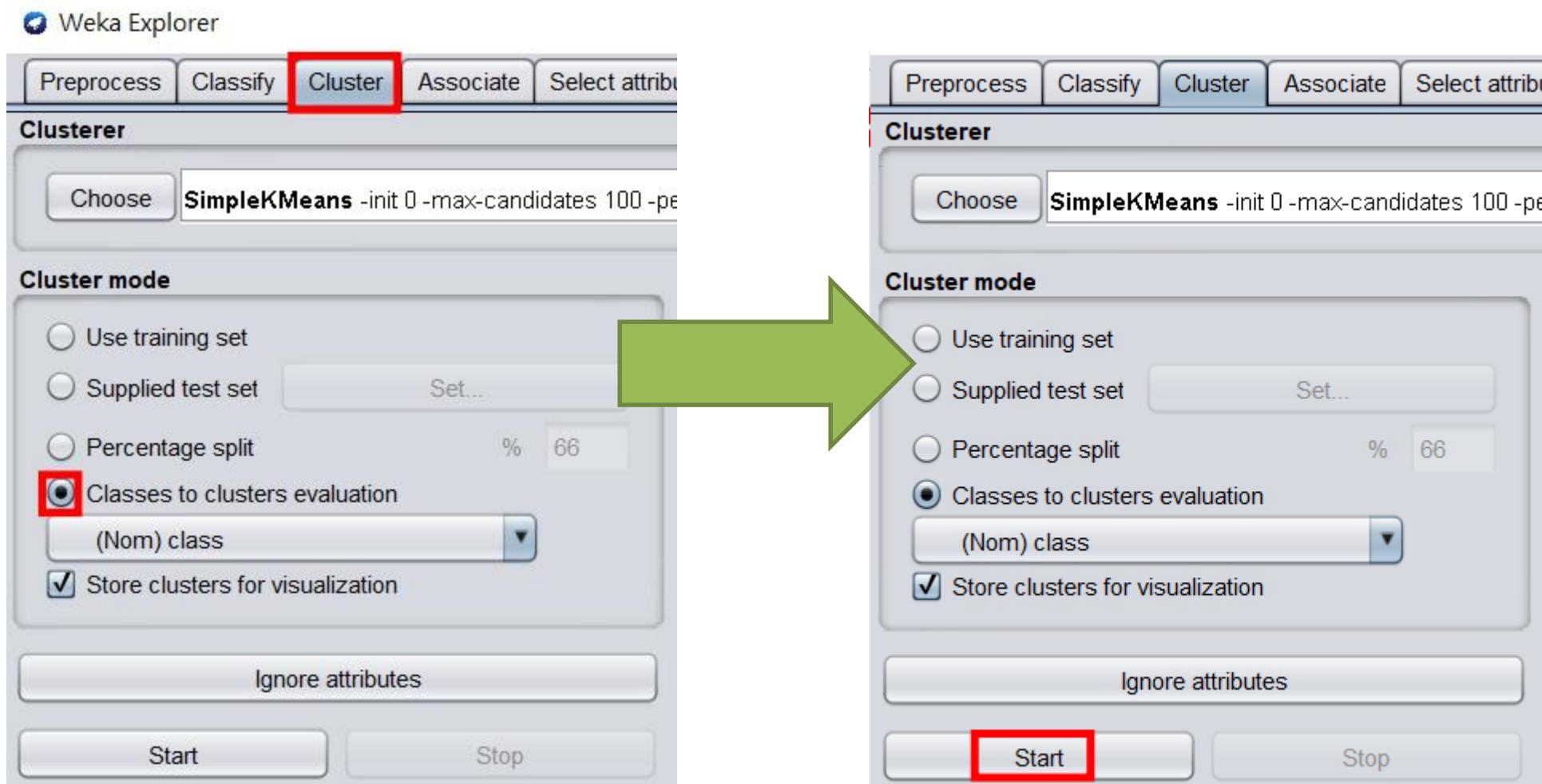
Relation: iris-weka.filters.unsupervised.attribute.AddCluster-Wweka.clusterers.SimpleKMeans -in

No.	1: sepallength Numeric	2: sepalwidth Numeric	3: petallength Numeric	4: petalwidth Numeric	5: class Nominal	6: cluster Nominal
...	6.3	3.3	6.0	2.5	Iris-virginica	cluster3
...	5.8	2.7	5.1	1.9	Iris-virginica	cluster1
...	7.1	3.0	5.9	2.1	Iris-virginica	cluster3
...	6.3	2.9	5.6	1.8	Iris-virginica	cluster3
...	6.5	3.0	5.8	2.2	Iris-virginica	cluster3
...	7.6	3.0	6.6	2.1	Iris-virginica	cluster3
...	4.9	2.5	4.5	1.7	Iris-virginica	cluster1
...	7.3	2.9	6.3	1.8	Iris-virginica	cluster3
...	6.7	2.5	5.8	1.8	Iris-virginica	cluster3
...	7.2	3.6	6.1	2.5	Iris-virginica	cluster3
...	6.5	3.2	5.1	2.0	Iris-virginica	cluster3
...	6.4	2.7	5.3	1.9	Iris-virginica	cluster3
...	6.8	3.0	5.5	2.1	Iris-virginica	cluster3
...	5.7	2.5	5.0	2.0	Iris-virginica	cluster1
...	5.8	2.8	5.1	2.4	Iris-virginica	cluster3
...	6.4	3.2	5.3	2.3	Iris-virginica	cluster3
...	6.5	3.0	5.5	1.8	Iris-virginica	cluster3
...	7.7	3.8	6.7	2.2	Iris-virginica	cluster3
...	7.7	2.6	6.9	2.3	Iris-virginica	cluster3
...	6.0	2.2	5.0	1.5	Iris-virginica	cluster1
...	6.9	3.2	5.7	2.3	Iris-virginica	cluster3
...	5.6	2.8	4.9	2.0	Iris-virginica	cluster1
...	7.7	2.8	6.7	2.0	Iris-virginica	cluster3
...	6.3	2.7	4.9	1.8	Iris-virginica	cluster1
...	6.7	3.3	5.7	2.1	Iris-virginica	cluster3
...	7.2	3.2	6.0	1.8	Iris-virginica	cluster3
...	6.2	2.8	4.8	1.8	Iris-virginica	cluster1
...	6.1	3.0	4.9	1.8	Iris-virginica	cluster1
...	6.4	2.8	5.6	2.1	Iris-virginica	cluster3
...	7.2	2.0	5.0	1.6	Iris-virginica	cluster2

Add instance Undo OK Cancel

Lesson 3.6: 聚類的評估

2. 切換到Cluster面板，左鍵單擊Cluster mode區域內的Classes to clusters evaluation選項的前方圓圈，並以左鍵單擊Start按鈕。



Lesson 3.6: 聚類的評估

得到了3個聚類，可以發現有17個被錯誤的劃分的實例。

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

Clustered Instances

0      61 ( 41%)
1      50 ( 33%)
2      39 ( 26%)

Class attribute: class
Classes to Clusters:

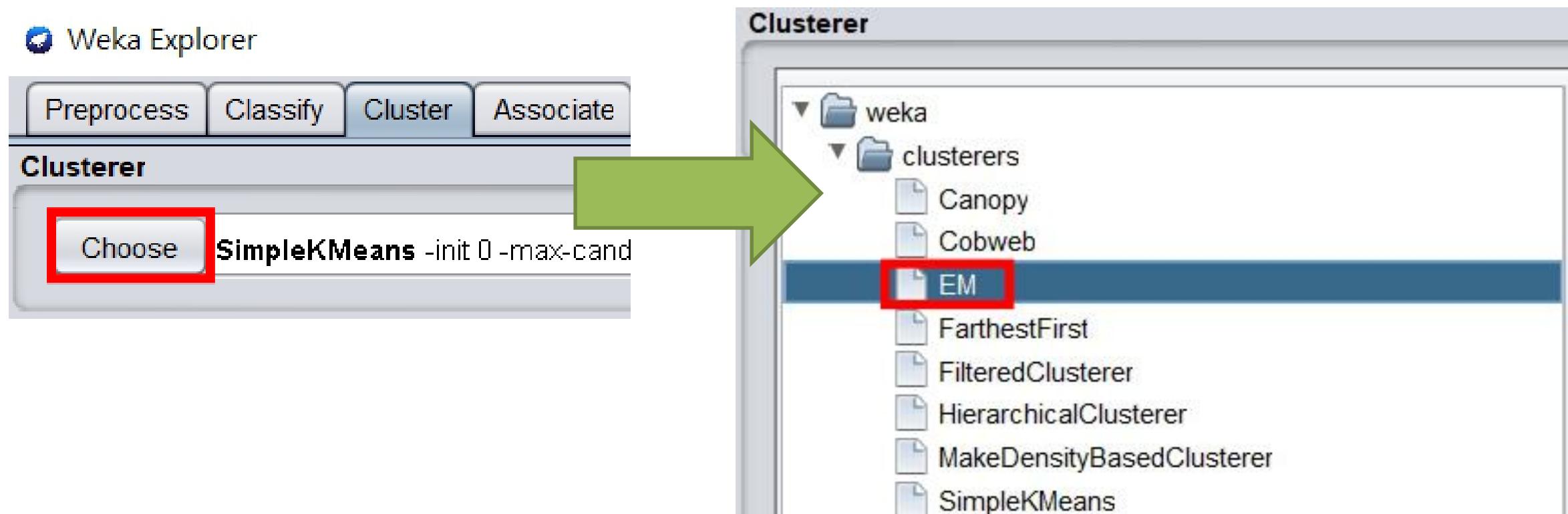
0 1 2 <-- assigned to cluster
0 50 0 | Iris-setosa
47 0 3 | Iris-versicolor
14 0 36 | Iris-virginica

Cluster 0 <-- Iris-versicolor
Cluster 1 <-- Iris-setosa
Cluster 2 <-- Iris-virginica

Incorrectly clustered instances : 17.0    11.3333 %
```

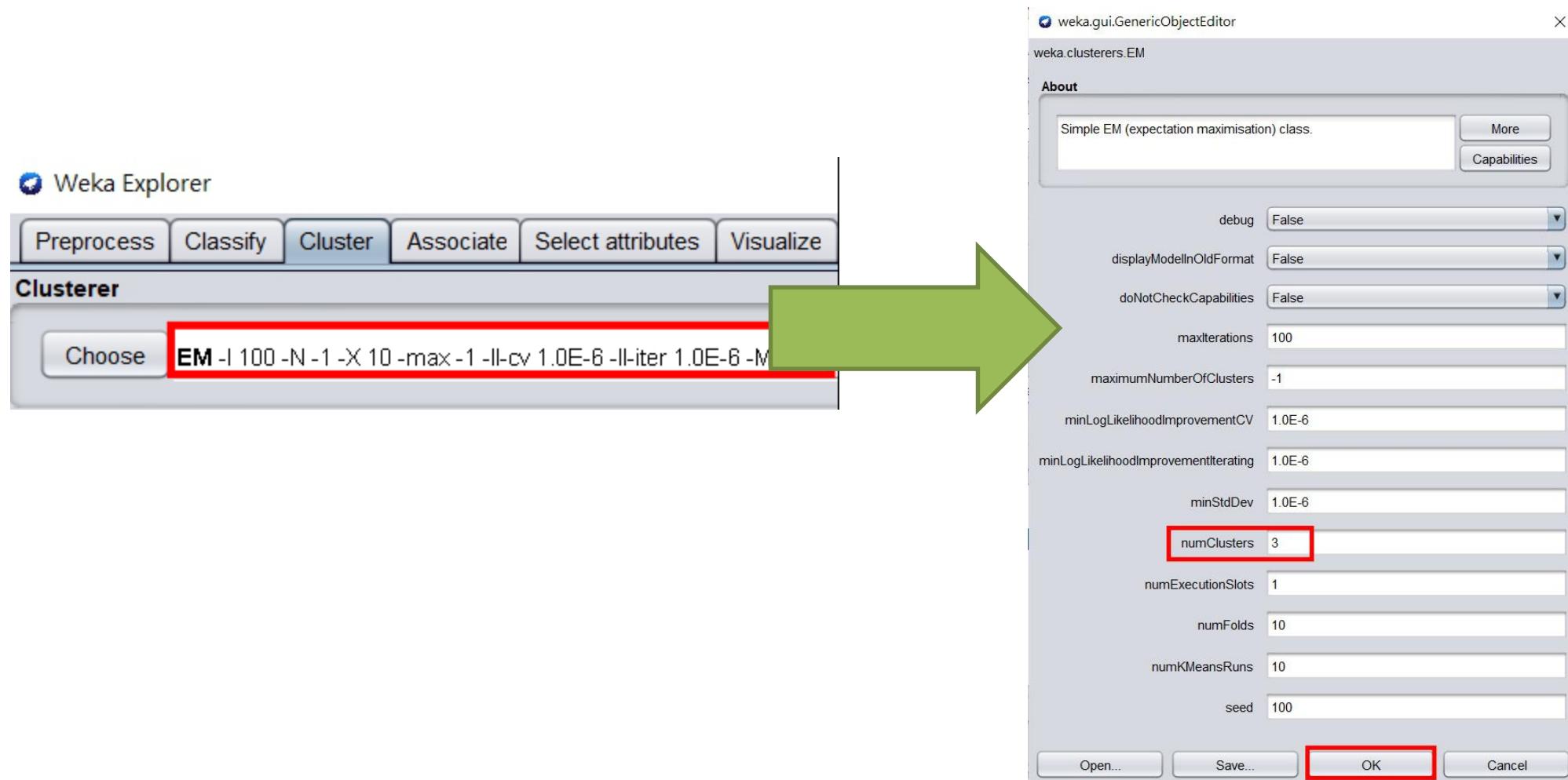
Lesson 3.6: 聚類的評估

3. 左鍵單擊Choose按鈕，於出現的選單中左鍵單擊EM聚類器。



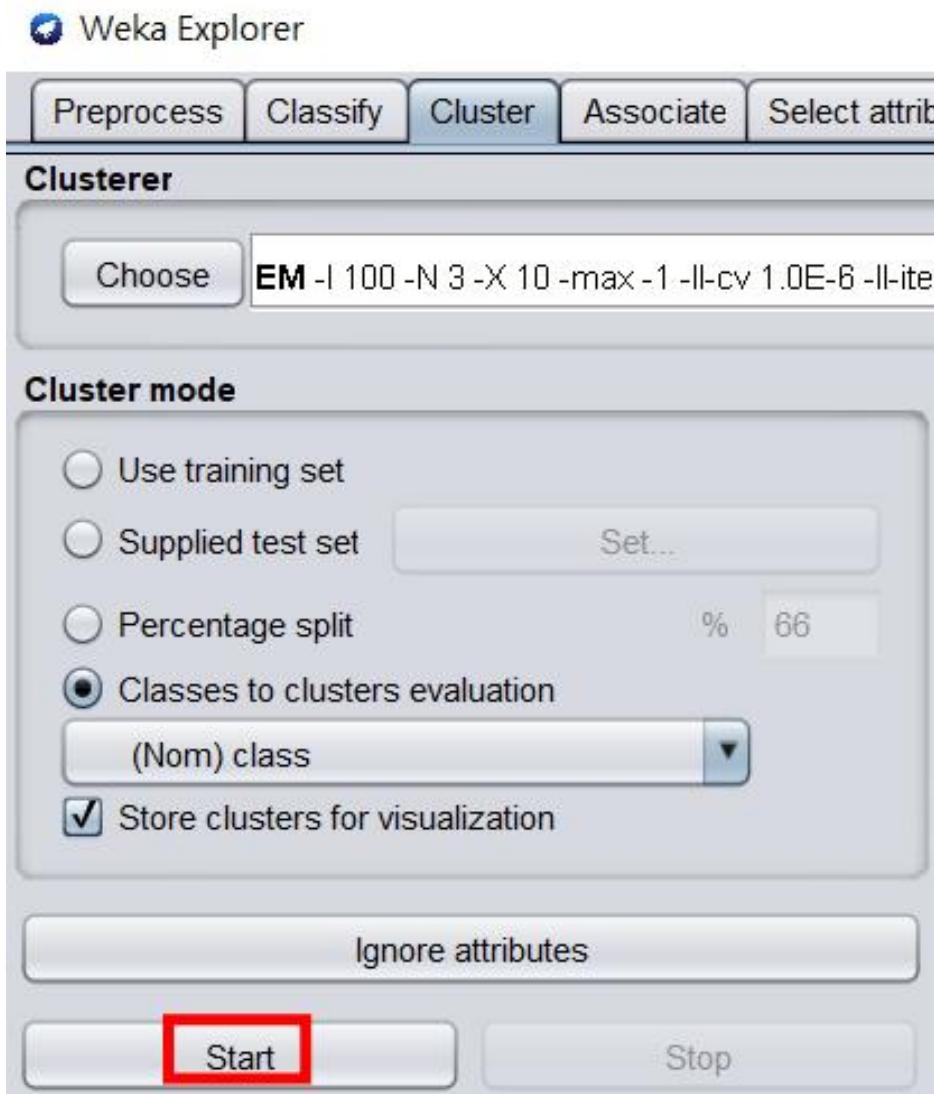
Lesson 3.6: 聚類的評估

4. 左鍵單擊聚類器名稱(左圖紅框處)，開啟配置視窗(右圖)。在配置視窗中將參數numClusters變更為3，再以左鍵單擊OK按鈕。



Lesson 3.6: 聚類的評估

5. 回到Cluster面板，左鍵單擊Start按鈕。



Lesson 3.6: 聚類的評估

類別-聚類評估(Classes-to-clusters evaluation)

- ❖ Iris 資料, SimpleKMeans, 指定3個聚類器
- ❖ 選擇Classes to clusters evaluation選項

SimpleKMeans運行結果 (3個聚類)

```
0 1 2 <- assigned to cluster  
0 50 0 | Iris-setosa  
47 0 3 | Iris-versicolor  
14 0 36 | Iris-virginica
```

```
Cluster 0 <- Iris-versicolor  
Cluster 1 <- Iris-setosa  
Cluster 2 <- Iris-virginica
```

Incorrectly clustered instances: 17 11%

可以看到聚類0主要由這47個實例構成，它們是versicolor。所以，我們認為聚類0代表versicolor。這裡有17個誤分的實例。

EM運行結果 (3個聚類)

```
0 1 2 <- assigned to cluster  
0 50 0 | Iris-setosa  
50 0 0 | Iris-versicolor  
14 0 36 | Iris-virginica
```

```
Cluster 0 <- Iris-  
versicolor Cluster 1 <-  
Iris-setosa Cluster 2 <-  
- Iris-virginica
```

Incorrectly clustered instances: 14.9%
EM的結果好得多，僅有14個誤分的實例，占數據集的9%。

Lesson 3.6: 聚類的評估

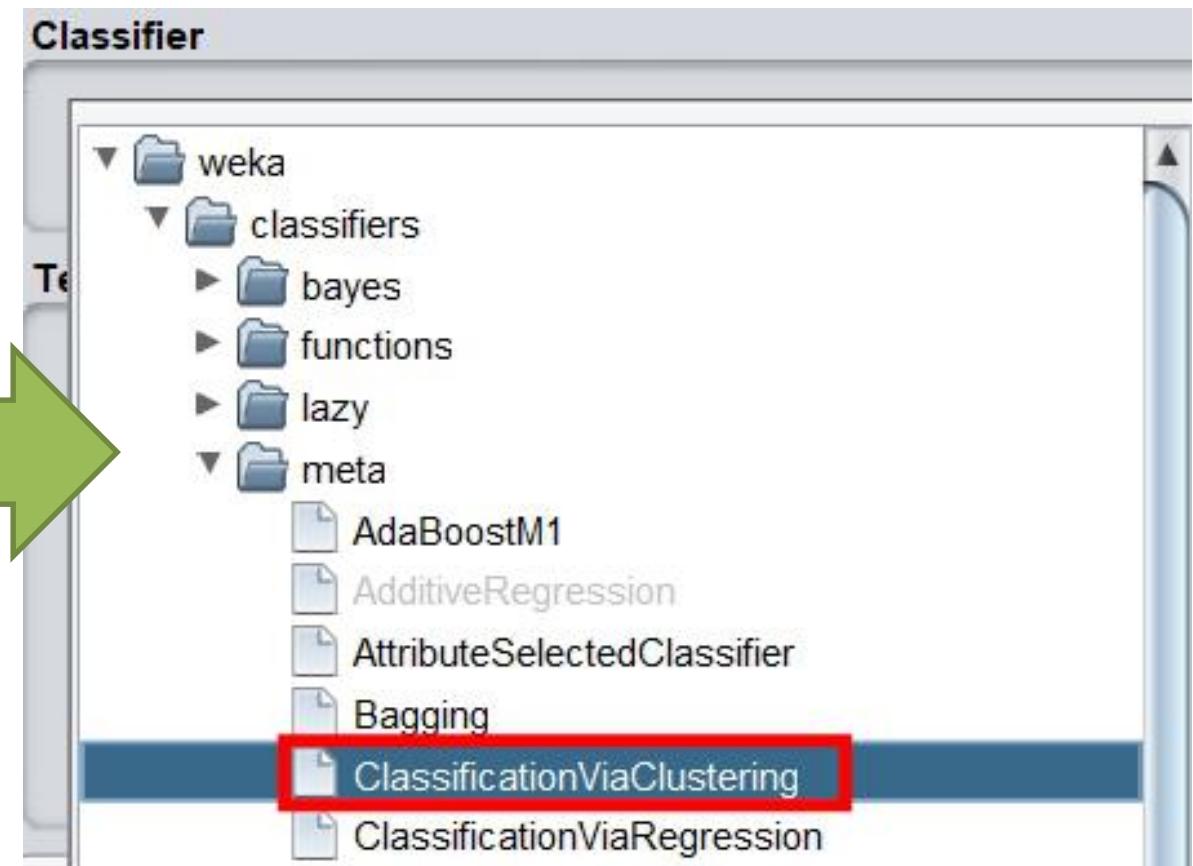
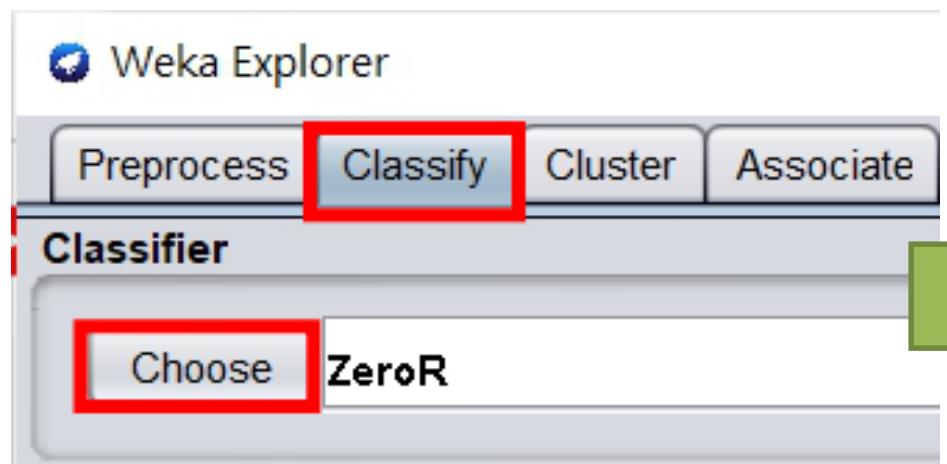
ClassificationViaClustering 元分類器

- ❖ 創建一個分類器：
 - 忽略類別
 - 聚類資料
 - 用最多的類別定義該聚類
- ❖ 顯然不能與其他分類技術競爭
- ❖ 比較聚類的好方法

Lesson 3.6: 聚類的評估

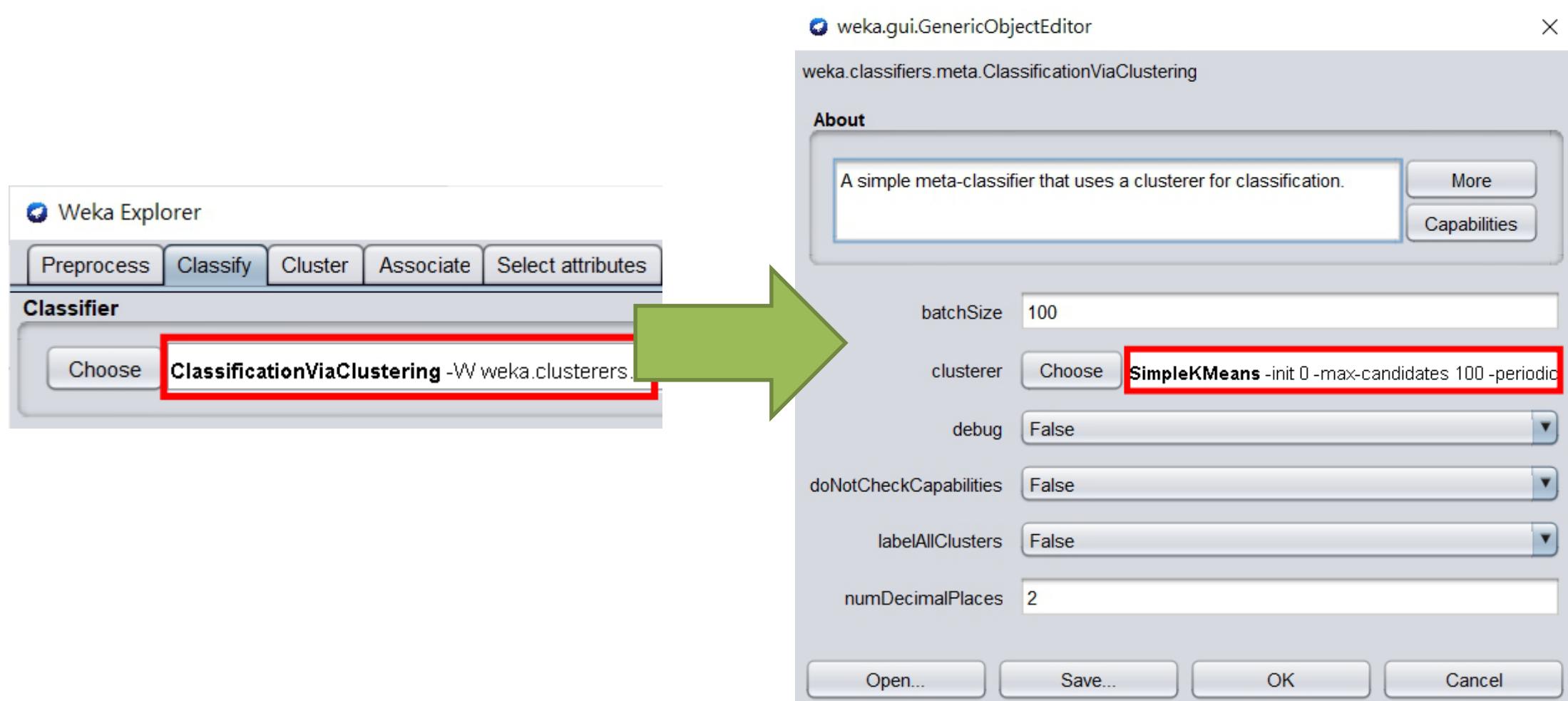
我們試著操作元分類器“ClassificationViaClustering”。

1. 切換至Classify面板，左鍵單擊Choose按鈕，於出現的選單中左鍵單擊ClassificationViaClustering。



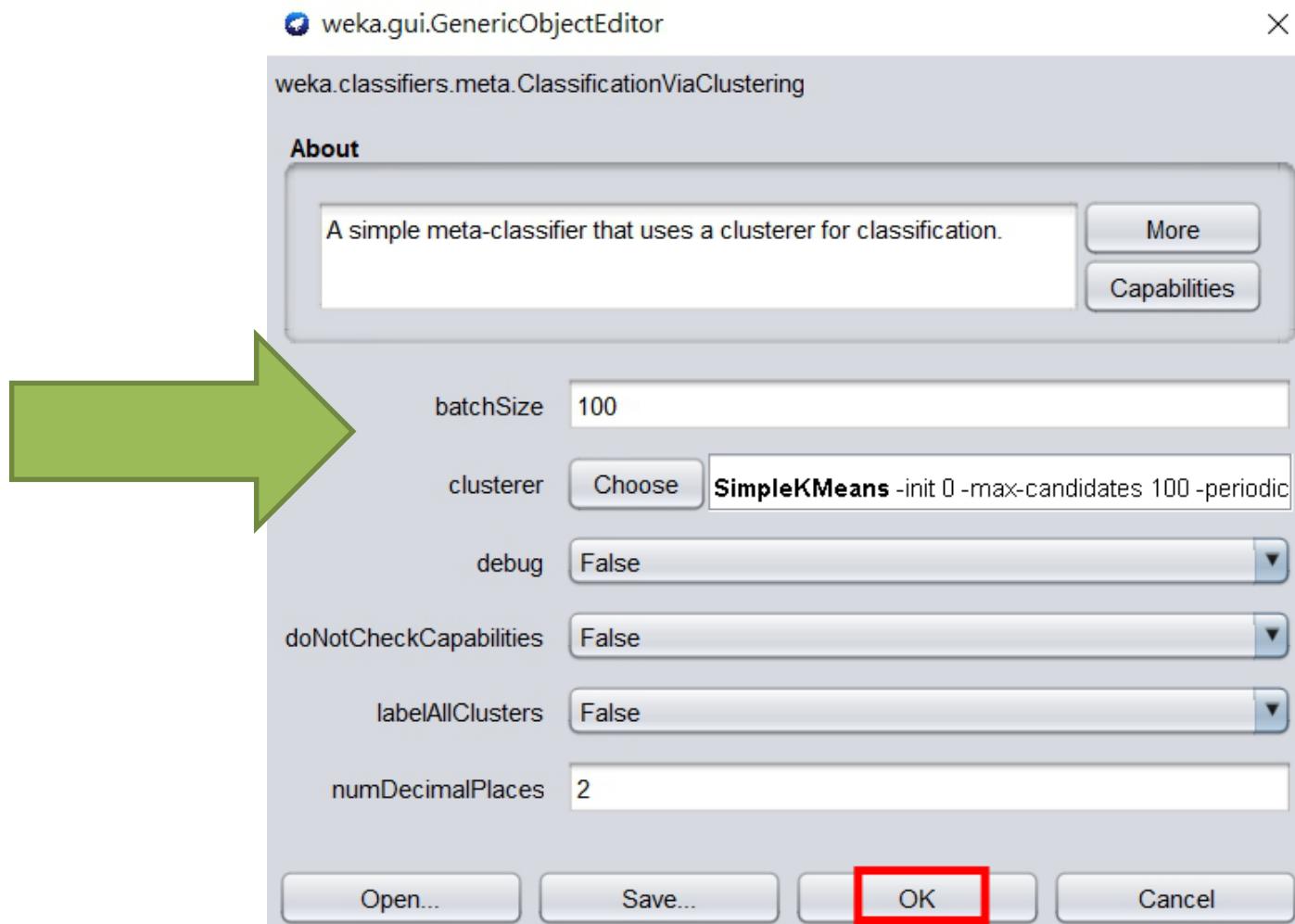
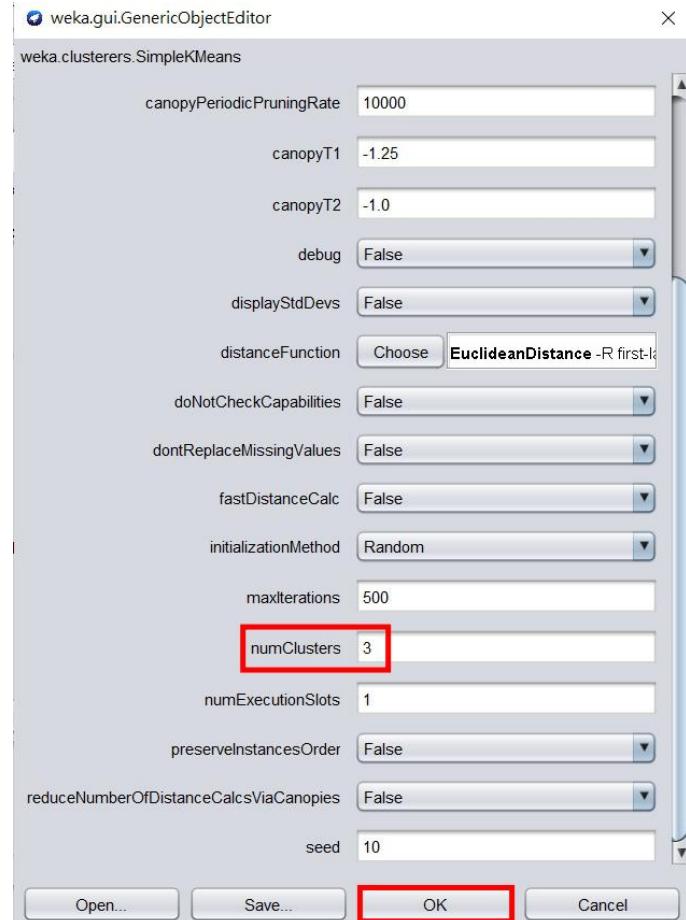
Lesson 3.6: 聚類的評估

2. 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聚類器名稱(右圖紅框處)。



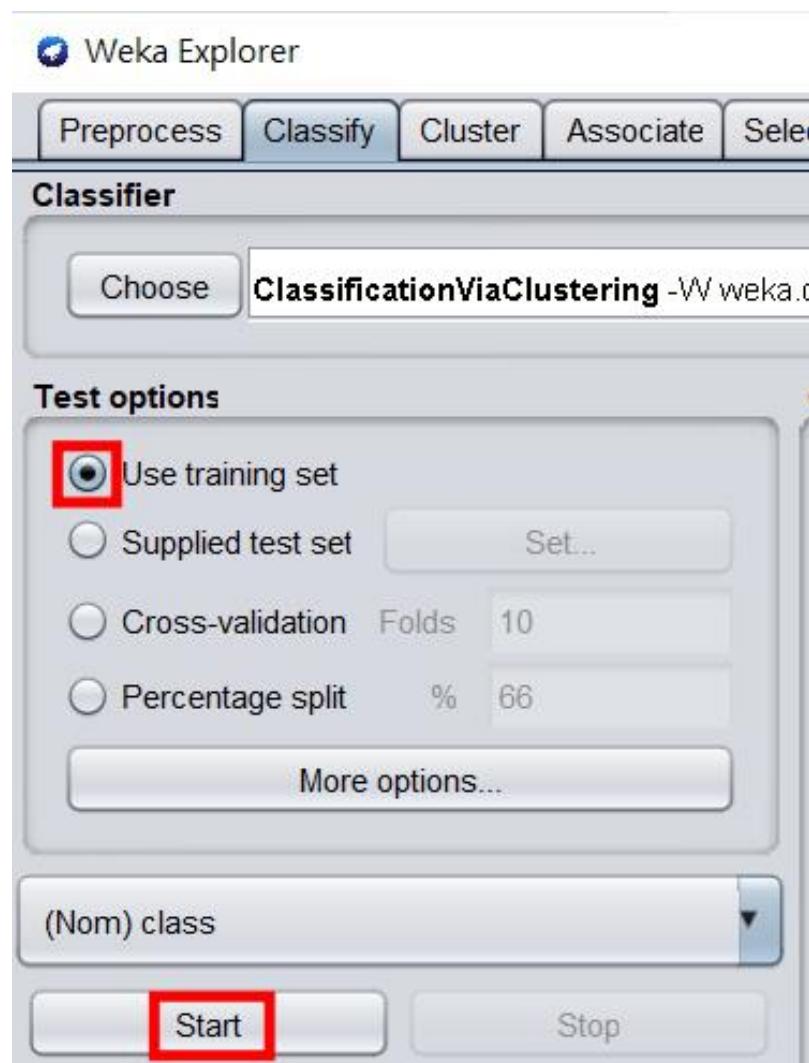
Lesson 3.6: 聚類的評估

3. 將參數numClusters設定為3後，左鍵單擊OK按鈕回到分類器配置視窗(右圖)，在此視窗左鍵單擊OK按鈕回到Classify面板。



Lesson 3.6: 聚類的評估

4. 在Test options區域內，左鍵單擊Use training set前方圓圈，接著再以左鍵單擊Start按鈕。



Lesson 3.6: 聚類的評估

- 我們看到了和之前一樣的矩陣。
- 一樣有17個誤差。這是用訓練數據評估的結果。

```
Classifier output

Time taken to test model on training data: 0.04 seconds

==== Summary ====

Correctly Classified Instances      133          88.6667 %
Incorrectly Classified Instances   17           11.3333 %
Kappa statistic                   0.83
Mean absolute error               0.0756
Root mean squared error          0.2749
Relative absolute error          17      %
Root relative squared error     58.3095 %
Total Number of Instances        150

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

          TP Rate  FP Rate  Precision  Recall   F-Measure  MCC    ROC Area  PRC Area
          1.000    0.000    1.000     1.000    1.000    1.000    1.000    1.000
          0.940    0.140    0.770     0.940    0.847    0.768    0.900    0.744
          0.720    0.030    0.923     0.720    0.809    0.742    0.845    0.758
Weighted Avg.      0.887    0.057    0.898     0.887    0.885    0.836    0.915    0.834

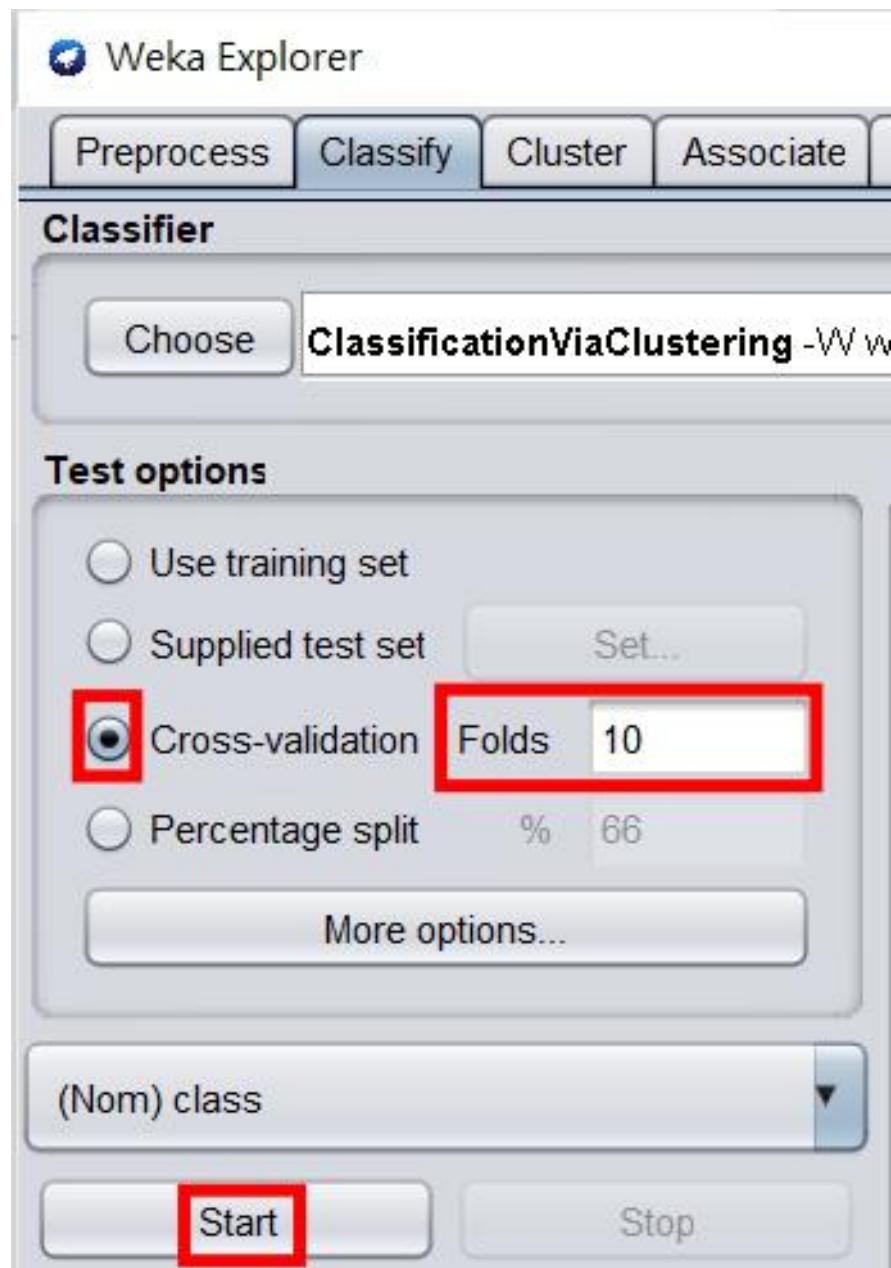
==== Confusion Matrix ====

  a  b  c  <-- classified as
50  0  0 |  a = Iris-setosa
 0 47  3 |  b = Iris-versicolor
 0 14 36 |  c = Iris-virginica
```

Lesson 3.6: 聚類的評估

我們改用交叉驗證。

5. 左鍵單擊Cross-validation前方圓圈，並確定後方輸入框內的值為10後，左鍵單擊Start按鈕。



Lesson 3.6: 聚類的評估

- 得到了19個誤差，
86%的正確率。
- 表現沒有之前那麼好。

```
Classifier output

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

    Correctly Classified Instances      129
    Incorrectly Classified Instances   20
                                         86 %
                                         13.3333 %

    Kappa statistic                   0.7987
    Mean absolute error              0.0895
    Root mean squared error          0.2991
    Relative absolute error          20.2694 %
    Root relative squared error     63.67 %
    UnClassified Instances           1
                                         0.6667 %
    Total Number of Instances        150

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

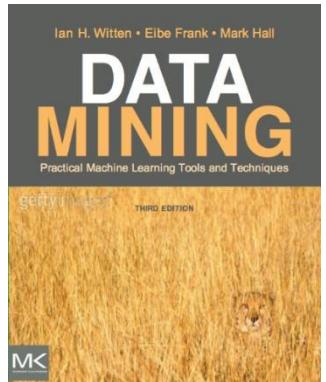
          TP Rate  FP Rate  Precision  Recall  F-Measure  MCC  ROC Area  PRC Area
    1.000    0.010    0.980    1.000    0.990    0.985    0.985    0.967
    0.920    0.162    0.742    0.920    0.821    0.727    0.880    0.709
    0.680    0.030    0.919    0.680    0.782    0.710    0.825    0.732
    Weighted Avg.    0.866    0.068    0.880    0.866    0.863    0.806    0.896    0.802

==== Confusion Matrix ====

    a  b  c  <-- classified as
49  0  0 |  a = Iris-setosa
 1 46  3 |  b = Iris-versicolor
 0 16 34 |  c = Iris-virginica
```

Lesson 3.6: 聚類的評估

- ❖ 評估聚類是困難的
 - SimpleKMeans: 使用聚類內部的誤差平方和
 - 但聚類算法的評估要在在特定的應用進行
- ❖ 視覺化
- ❖ **AddCluster** 過濾器顯示了每個聚類中的實例
- ❖ 類別-聚類評估(Classes to clusters evaluation)
- ❖ 透過聚類分類的分類器



課程文本

- ❖ Section 11.2, under *Clustering and association rules*
- ❖ Section 11.6 *Clustering algorithms*



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/