

WEKA 3-5-5 Explorer 使用者指南

原文版本 **3.5.5**

翻譯 王娜

校對 **C6H5NO2**

Pentaho 中文討論群

組 QQ 群：**12635055** 論壇：

<http://www.bipub.org/bipub/index.asp>

<http://bbs.wekacn.org/>

目 錄

1	啓動WEKA.....	3
2	WEKA Explorer.....	5
2.1	標籤頁	5
2.2	狀態列	5
2.3	Log 按鈕.....	5
2.4	WEKA 狀態圖示	5
3	預處理.....	6
3.1	載入數據	6
3.2	當前關係	6
3.3	處理屬性	7
3.4	使用篩選器	7
4	分類.....	10
4.1	選擇分類器	10
4.2	測試選項	10
4.3	Class屬性	11
4.4	訓練分類器	11
4.5	分類器輸出文本	11
4.6	結果列表	12
5	聚類.....	13
5.1	選擇聚類器（Clusterer）	13
5.2	聚類模式	13
5.3	忽略屬性	13
5.4	學習聚類	14
6	關聯規則.....	15
6.1	設定	15
6.2	學習關聯規則	15
7	屬性選擇.....	16
7.1	搜索與評估	16
7.2	選項	16
7.3	執行選擇	16
8	視覺化.....	18

8.1	散點圖矩陣	18
8.2	選擇單獨的二維散點圖	18
8.3	選擇實例	19
參考文獻	20

啓動 WEKA

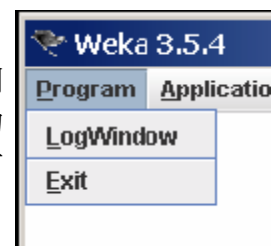
WEKA中新的功能表驅動的 GUI 繼承了老的 GUI 選擇器（類 `weka.gui.GUIChooser`）的功能。它的MDI（“多重文件介面”）外觀，讓所有打開的視窗更加明瞭。



這個功能表包括六個部分。

1. Program

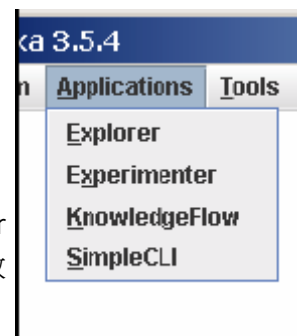
- z **LogWindow** 打開一個日誌視窗，記錄輸出到 stdout 或 stderr 的內容。在 MS Windows 那樣的环境中，WEKA 不是從一個終端啓動，這個就比較有用。
- z **Exit** 關閉WEKA。



2. Applications

列出 WEKA 中主要的應用程式。

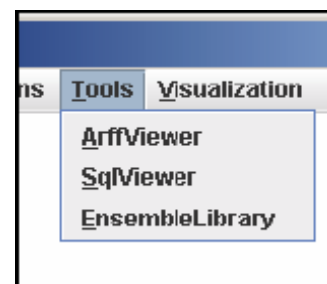
- z **Explorer** 使用 WEKA 探索資料的環境。（本文檔的其它部分將詳細介紹這個環境）
- z **Experimenter** 運行演算法試驗、管理演算法方案之間的統計檢驗的環境。
- z **KnowledgeFlow** 這個環境本質上和 Explorer 所支持的功能是一樣的，但是它有一個可以拖放的介面。它有一個優勢，就是支持增量學習（incremental learning）。
- z **SimpleCLI** 提供了一個簡單的命令列介面，從而可以在沒有自帶命令列的作業系統中直接執行 WEKA 命令。



3. Tools

其他有用的應用程式。

- z **ArffViewer** 一個 MDI 應用程式，使用電子表格的形式來查看 ARFF 檔。
- z **SqlViewer** 一個 SQL 工作表，用來通過 JDBC 查詢資料庫。
- z **EnsembleLibrary** 生成集成式選擇（Ensemble Selection）[5] 所需設置的介面。

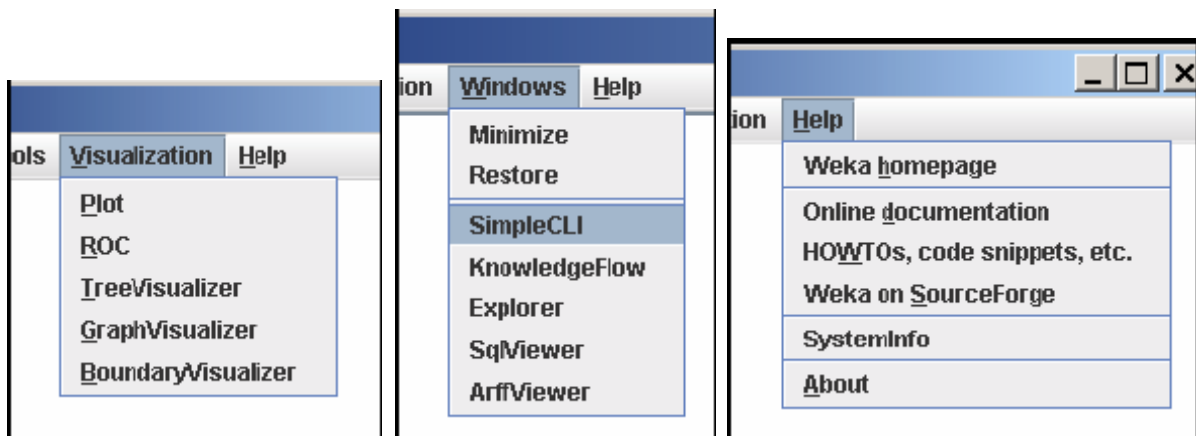


4. Visualization

WEKA 中資料視覺化的方法。

- z **Plot** 作出資料集的二維散點圖。
- z **ROC** 顯示預先保存的 ROC 曲線。

- z **TreeVisualizer** 顯示一個有向圖，例如一個決策樹。
 - z **GraphVisualizer** 顯示 XML、BIF 或 DOT 格式的图片，例如貝葉斯網路（Bayesian network）。
 - z **BoundaryVisualizer** 允許在二維空間中對分類器的決策邊界進行視覺化。
5. **Windows** 所有已打開的視窗都列在這裡。
- z **Minimize** 最小化所有當前的視窗。
 - z **Restore** 還原所有最小化過的視窗。
6. **Help** WEKA 的線上資源可以從這裡找到。
- z **Weka homepage** 打開一個流覽器視窗，顯示 WEKA 的主頁。
 - z **Online documentation** 連結到 WekaDoc 維琪文檔 [4]。
 - z **HOWTOs, code snippets, etc.** 通用的 WekaWiki [3]，包括大量的例子，以及開發和使用 WEKA 的基本知識（HOWTO）。
 - z **Weka on Sourceforge** WEKA 項目在 Sourceforge.net 的主頁。
 - z **SystemInfo** 列出一些關於 Java/WEKA 環境的資訊，例如 CLASSPATH。
 - z **About** 不光彩的“About”窗口。



如果從終端啟動 WEKA，會有一些文字在終端視窗中出現。這些文字是可以忽略的，除非某些東西出錯了——這時它可以說明找到錯誤的原因。（LogWindow 也可以顯示那些資訊。）

這份文檔也可以從線上的 WekaDoc Wiki [4] 中找到，它將集中闡述如何使用 Explorer，而不會逐個解釋 WEKA 中的資料預處理工具和學習演算法。要獲得關於各種篩選器（filter）和學習演算法的更多資訊，可參考 Data Mining [2] 一書。

1 WEKA Explorer

1.1 標籤頁

在窗口的頂部，標題列下是一排標籤。當 Explorer 首次啟動時，只有第一個標籤頁是活動的；其他均是灰色的。這是因為在探索資料之前，必須先打開一個資料集(可能還要對它進行預處理)。

所有的標籤頁如下所示：

1. **Preprocess.** 選擇和修改要處理的資料。
2. **Classify.** 訓練和測試關於分類或回歸的學習方案。
3. **Cluster.** 從數據中學習聚類。
4. **Associate.** 從資料中學習關聯規則。
5. **Select attributes.** 選擇資料中最相關的屬性。
6. **Visualize.** 查看資料的互動式二維圖像。這些標籤被啟動後，點擊它們可以在不同的標籤頁面上進行切換，而每一個頁面上可以

執行對應的操作。不管位於哪個頁面，視窗的底部區域(包括狀態列、log 按鈕和 Weka 鳥)仍然可見。

1.2 狀態列

狀態 (Status) 欄出現在視窗的最底部。它顯示一些資訊讓你知道正在做什麼。例如，如果 Explorer 正忙於裝載一個檔，狀態列就會有通知。

提示 — 在狀態列中的任意位置右擊滑鼠將會出現一個小功能表。這個功能表給了你兩個選項：

1. **Memory Information.** 在 log 欄中顯示 WEKA 可用的記憶體量。
2. **Run garbage collector.** 強制運行 Java 垃圾回收器，搜索不再需要的記憶體空間並將之釋放，從而可為新任務分配更多的記憶體。注意即使不強制運行，垃圾回收也是一直作為幕後工作在運行的。

1.3 Log 按鈕

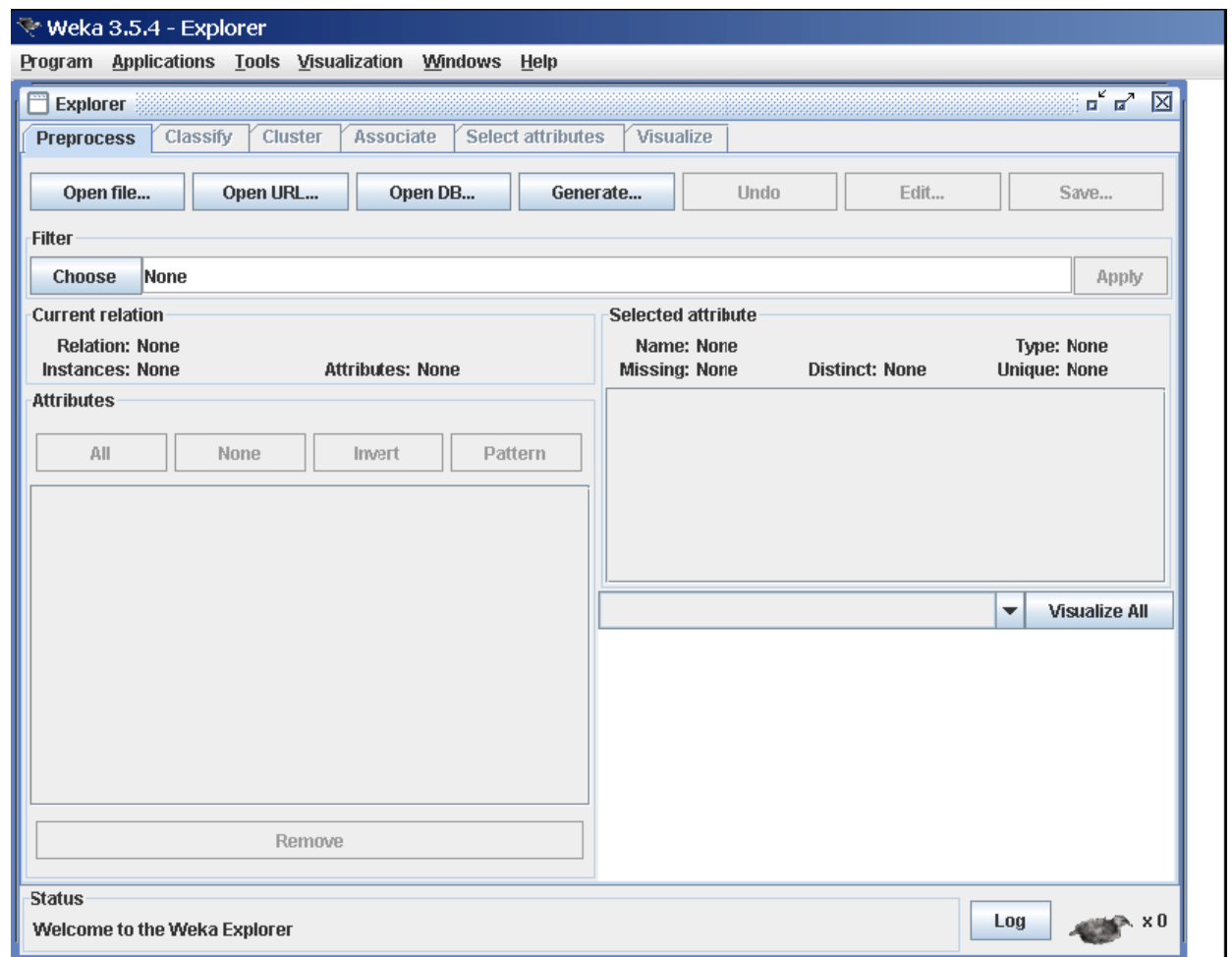
點擊這個按鈕，會出現一個單獨的視窗，包含一個可拖動的文本區域。文本的每一行被加了一個時間戳記，顯示了它進入日誌 (log) 的時間，一旦在 WEKA 中執行某種操作時，該日誌就會記錄發生了什麼。對於使用命令列或者 SimpleCLI 的人，日誌也將完整地記錄分類，聚類，特徵提取等任務的設置字元，使得它們可被複製/粘貼到其它地方。但關於資料集和 class 屬性¹的選項仍然要由使用者給出（例如，分類器 (classifier) 的 -t，或者篩選器的 -i 和 -o）

1.4 WEKA 狀態圖示

狀態列的右邊是 WEKA 狀態圖示。當不運行任何進程時，WEKA 鳥會坐下並打一個小盹。×符號旁的數字顯示了正運行的併發進程的數量。當系統空閒時，它是零，而當進程的數量增長時，它也會增長。任意進程啟動後，小鳥會站起來並到處活動。如果它仍然是站著的，但是很長時間內不動，那麼它生病了：某個地方出錯了！在這種情況下，應該重新啟動 WEKA Explorer。

¹ 在分類或回歸任務中，class 屬性是預設的目標變數。注意這與下文中的分類型屬性不是一個概念——譯注。

2 預處理



2.1 載入數據

預處理頁頂部的前4個按鈕用來把資料載入WEKA：

1. **Open file....** 打開一個對話方塊，允許你流覽本地檔案系統上的資料檔案。
2. **Open URL....** 請求一個存有資料的 URL 位址。
3. **Open DB....** 從資料庫中讀取資料（注意，要使之可用，可能需要編輯 weka/experiment/ DatabaseUtils.props 中的文件）
4. **Generate....** 從一些資料生成器（DataGenerators）中生成人造資料。

使用 **Open file...** 按鈕可以讀取各種格式的檔：WEKA 的 ARFF 格式，CSV 格式，C4.5 格式，或者序列化的實例²格式。ARFF 檔通常副檔名是.arff，CSV 文件副檔名是.csv，C4.5 文件副檔名是.data 和 .names，序列化的實例物件副檔名為.bsi。

2.2 當前關係

載入資料後 預處理面板就會顯示各種資訊 **Current relation** 一欄（“current relation”指目前裝載的資料，可理解為資料庫術語中單獨的關係表）有3個條目：

1. **Relation.** 關係的名稱，在它裝載自的檔中給出。使用篩選器（下文將詳述）將修改關係的名稱。

² 只有本段文字中的“實例”是 JAVA 語言中實例的概念；而後文中的“實例”都將指資料集中的記錄——譯注。

2. **Instances.** 資料中的實例(或稱資料點/記錄) 的個數。
3. **Attributes.** 資料中的屬性(或稱特徵) 的個數。

2.3 處理屬性

在 **Current relation** 一欄下是 **Attributes** (屬性) 欄。有四個按鈕，其下是當前關係中的屬性清單。該列表有3列：

1. **No..** 一個數位，用來標識資料檔案中指定的各屬性的順序。
2. 選擇框。允許勾選關係中呈現的各屬性。
3. **Name.** 資料檔案中聲明的各屬性的名稱。

當點擊屬性清單中的不同行時，右邊 **Selected attribute** 一欄的內容隨之改變。這一欄給出了清單中當前高亮顯示的屬性的一些描述：

1. **Name.** 屬性的名稱，和屬性清單中給出的相同。
2. **Type.** 屬性的類型，最常見的是分類型 (Nominal) 和數值型 (Numeric)。
3. **Missing.** 資料中該屬性缺失(或者未指定)的實例的數量(及百分比)。
4. **Distinct.** 資料中該屬性包含的不同值的數目。
5. **Unique.** 唯一地擁有某值的實例的數目 (及百分比)，這些實例每個的取值都和別的不一樣。

在這些統計量的下面是一個清單，根據屬性的不同類型，它顯示了關於這個屬性中儲存的值的更多資訊。如果屬性是分類型的，清單將包含該屬性的每個可能值以及取那個值的實例的數目。如果屬性是數值型的，列表將給出四個統計量來描述資料取值的分佈—最小值、最大值、平均值和標準差。在這些統計量的下方，有一個彩色的長條圖，根據長條圖上方一欄所選擇的 **class** 屬性來著色。(在點擊時，該欄將顯示一個可供選擇的下拉清單。) 注意僅有分類型的 **class** 屬性才會讓長條圖出現彩色。最後，若點擊 **Visualize All** 按鈕，將在一個單獨的視窗中顯示資料集中所有屬性的長條圖。

回到屬性清單，開始時所有的選擇框都是沒有被勾選的。可通過逐個點擊來勾選/取消。以上的4個按鈕也可用於改變選擇：

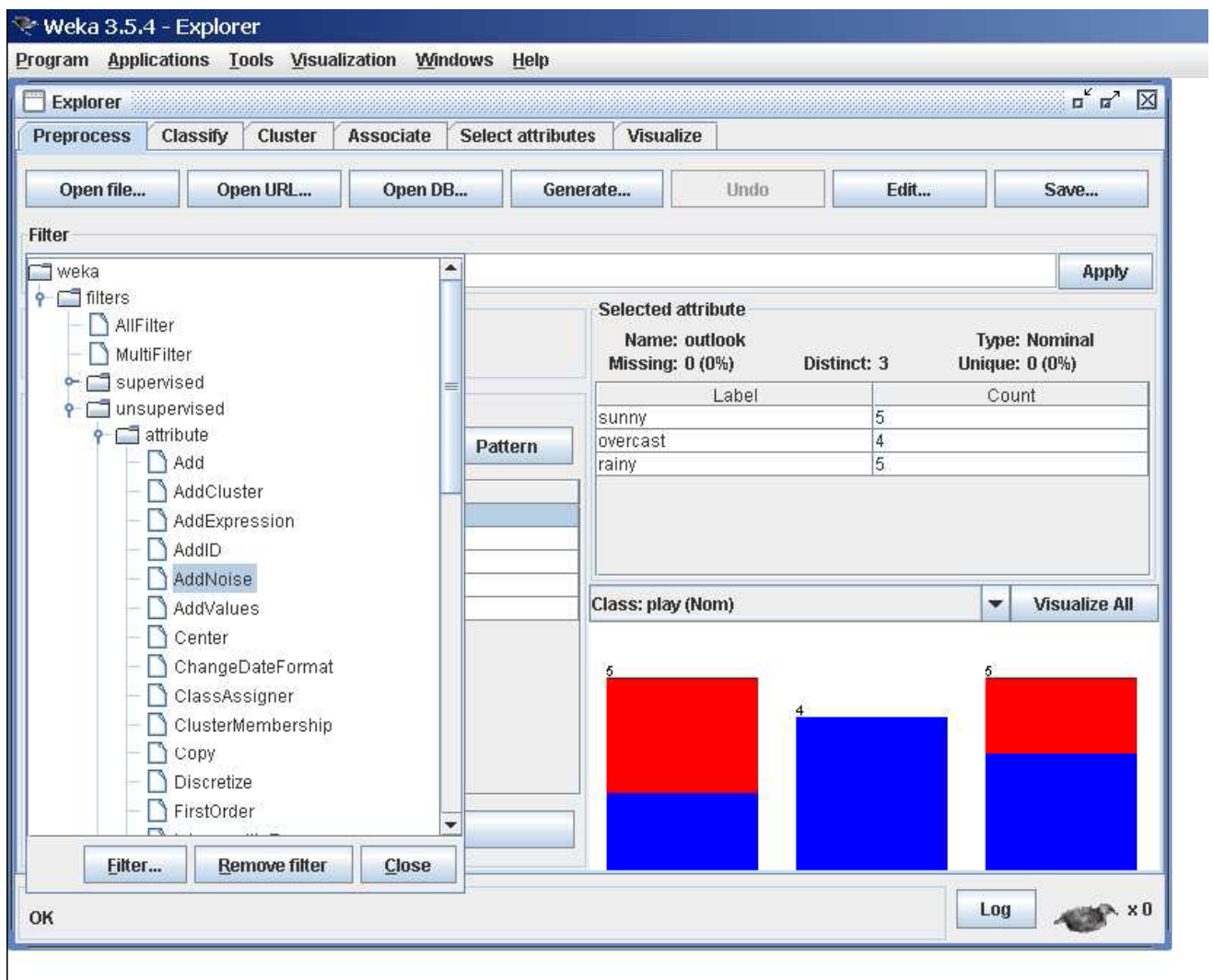
1. **All.** 所有選擇框都被勾選。
2. **None.** 所有選擇框被取消 (沒有勾選)。
3. **Invert.** 已勾選的選擇框都被取消，反之亦然。
4. **Pattern.** 讓使用者基於 Perl 5 規則運算式來選擇屬性。例如，用 `*_id` 選擇所有名稱以 `_id` 結束的屬性。

選中了想要的屬性後，可通過點擊屬性清單下的 **Remove** 按鈕刪除他們。注意可通過點擊位於 **Preprocess** 面板的右上角的 **Edit** 按鈕旁的 **Undo** 按鈕來取消操作。

2.4 使用篩選器³

在預處理階段，可以定義篩選器來以各種方式對資料進行變換。**Filter** 一欄用於對各種篩選器進行必要的設置。**Filter** 一欄的左邊是一個 **Choose** 按鈕。點擊這個按鈕就可選擇 WEKA 中的某個篩選器。選定一個篩選器後，它的名字和選項會顯示在 **Choose** 按鈕旁邊的文字方塊中。用滑鼠左鍵點擊這個框，將出現一個 **GenericObjectEditor** (通用物件 編輯器) 對話方塊。用滑鼠右鍵 (或 **Alt+Shift+左鍵**) 點擊將出現一個功能表，你可從中選擇，要麼在 **GenericObjectEditor** 對話方塊中顯示相關屬性，要麼將當前的設置字元複製到剪貼 板。

³ 篩選器的英文原文是 **filter**，與資料庫術語中的篩選有關。但是 WEKA 中的 **filter** 不僅能提供篩選功能，還涵蓋了其他各種資料變換。——譯注。



GenericObjectEditor 對話方塊

GenericObjectEditor 對話方塊可以用來配置一個篩選器。同樣的對話方塊也用於配置其他物件，例如分類器（classifier）和聚類器（clusterers）（見下文）。視窗中的欄位反映了可用的選項。點擊它們中間的一個便可改變 filter 的設置。例如，某項設置可能是一串文本字元，這時將字串輸入相應的文字方塊中即可。或者它可能會給出一個下拉清單，列出可供選擇的幾個狀態。也可能是其他一些操作，根據所需的資訊而有所區別。如果把將滑鼠指針停留在某個欄位上，會出現一個小提示來給出相應選項的資訊。而有關該篩選器和它的選項的更多資訊可通過點擊 GenericObjectEditor 視窗頂部 **About** 面板中的 **More** 按鈕來獲得。

除了 **More** 按鈕，某些物件也會在關於欄中顯示一些有關其功能的簡短描述。點擊 **More** 按鈕，會出現一個視窗來描述了不同的選項分別起什麼作用。有的還另外一個 **Capabilities** 按鈕，它能列出該物件可處理的屬性和 class 屬性的類型。

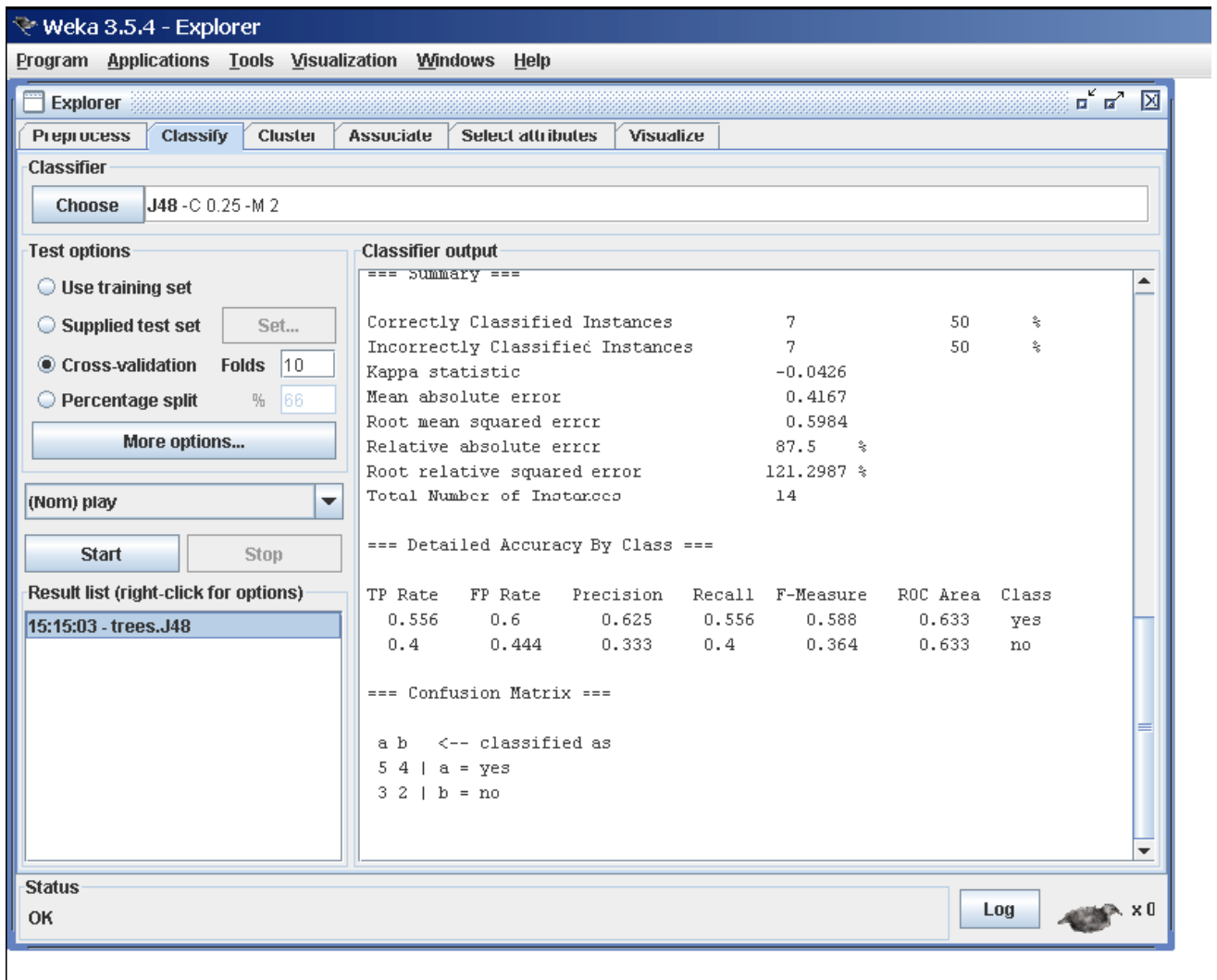
GenericObjectEditor 對話方塊的底部有4個按鈕。前兩個 **Open...** 和 **Save...** 允許存儲對該物件的配置，以備將來之用。**Cancel** 按鈕用於直接退出，任何已作出的改變都將被忽略。當前選擇的物件和設置令人滿意後，點擊 **OK** 返回到主 Explorer 視窗。

應用篩選器

選擇並配置好一個篩選器後，就可通過點擊 Preprocess 面板的 **Filter** 欄右邊的 **Apply** 按鈕將之應用於資料集上。然後 Preprocess 面板將顯示轉換過的資料。可點擊 **Undo** 按鈕取消改變。你也可使用 **Edit...** 按鈕在一個資料集編輯器中手動修改你的數據。最後，點擊 Preprocess 面板右上角的 **Save...** 按鈕將用同樣的格式保存當前的關係，以備將來使用。

注意：一些篩選器會依據是否設置了 `class` 屬性來做出不同的動作。（點擊長條圖上方那一欄時，會出現一個可供選擇的下拉清單。）特別的，“supervised filters”（監督式篩選器）需要設置一個 `class` 屬性，而某些“unsupervised attribute filters”（非監督式屬性篩選器）將忽略 `class` 屬性。注意也可以將 `Class` 設成 `None`，這時沒有設置 `class` 屬性。

3 分類⁴



3.1 選擇分類器

在 **classify** 頁面的頂部是 **Classifier** 欄。這一欄中有一個文字方塊，給出了分類器的名稱和它的選項。左鍵點擊文字方塊會打開一個 **GenericObjectEditor**，可以像配置篩選器那樣配置當前的分類器。右鍵（或Alt+Shift+左鍵）點擊也可以複製設置字元到剪貼板，或者在 **GenericObjectEditor** 中顯示相關屬性。**Choose** 按鈕用來選擇 WEKA 中可用的分類器。

3.2 測試選項

應用選定的分類器後得到的結果會根據 **Test Option** 一欄中的選擇來進行測試。共有四種測試模式：

1. **Using training set.** 根據分類器在用來訓練的實例上的預測效果來評價它。
2. **Supplied test set.** 從檔載入的一組實例，根據分類器在這組實例上的預測效果來評價它。點擊 **Set...** 按鈕將打開一個對話方塊來選擇用來測試的檔。
3. **Cross-validation.** 使用交叉驗證來評價分類器，所用的折數填在 **Folds** 文字方塊中。

⁴ WEKA 中的分類和回歸都放入了 **classify** 頁面中，相應的工具都叫做分類器（**classifier**）。參考4.3節。

4. **Percentage split.** 從資料集中按一定百分比取出部分資料放在一邊作測試用，根據分類器這些實例上預測效果來評價它。取出的資料量由 **%** 一欄中的值決定。注意：不管使用哪種測試方法，得到的模型總是從所有訓練資料中構建的。點擊 **More options** 按鈕可以設置更多的測試選項：

1. **Output model.** 輸出基於整個訓練集的分類模型，從而模型可以被查看，視覺化等。該選項預設為選中的。
2. **Output per-class stats.** 輸出每個 class 的準確度/回饋率（precision/recall）和正確/錯誤（true/false）的統計量。該選項也是預設選中的。
3. **Output evaluation measures.** 輸出熵估計度量。該選項預設沒有選中。
4. **Output confusion matrix.** 輸出分類器預測結果的混淆矩陣。該選項預設選中。
5. **Store predictions for visualization.** 記錄分類器的預測結果使得它們能被可視化表示。
6. **Output predictions.** 輸出測試資料的預測結果。注意在交叉驗證時，實例的編號不代表它在資料集中的位置。
7. **Cost-sensitive evaluation.** 誤差將根據一個價值矩陣來估計。**Set...** 按鈕用來指定價值矩陣。
8. **Random seed for xval / % Split.** 指定一個隨即種子，當出於評價的目的需要分割資料時，它用來隨機化資料。

3.3 Class 屬性

WEKA 中的分類器被設計成經過訓練後可以預測一個 class 屬性，也就是預測的目標。有的分類器只可用來學習分類的 class 屬性；有的則只可用來學習數值型的 class 屬性（回歸問題）；還有的兩者都可以學習。

預設的，資料集中的最後一個屬性被看作 class 屬性。如果想訓練一個分類器，讓它預測一個不同的屬性，點擊 **Test options** 欄下方的那一欄，會出現一個屬性的下拉清單以供選擇。

3.4 訓練分類器

分類器，測試選項和 class 屬性都設置好後，點擊 **Start** 按鈕就可以開始學習過程。分類器忙於訓練時，下方的小鳥會動來動去。可以通過點擊 **Stop** 按鈕，在任意時刻停止訓練過程。

訓練完成後，會發生幾件事。右邊的 **Classifier output** 區域會被填充一些文本，描述訓練和測試的結果。在 **Result list** 欄中會出現一個新的條目。接下來我們會觀察這個結果列表，但我們先來研究輸出的文本。

3.5 分類器輸出文本

Classifier output 區域的文本有一個捲軸以便流覽結果。按住 **Alt** 和 **Shift** 鍵，在這個區域點擊滑鼠左鍵，會出現一個對話方塊，讓你用各種格式（目前可用 **JPEG** 和 **EPS**）保存輸出的結果。當然，可以通過放大 **Explorer** 視窗來獲得更大的顯示區域。輸出結果 可分為幾個部分：

1. **Run information.** 給出了學習演算法各選項的一個清單。包括了學習過程中涉及到的關係名稱，屬性，實例和測試模式。
2. **Classifier model (full training set).** 用文本表示的基於整個訓練集的分類模型。

所選測試模式的結果可以分解為以下幾個部分：

3. **Summary.** 一系列統計量，描述了在指定測試模式下，分類器預測 class 屬性的準確程度。

4. **Detailed Accuracy By Class.** 更詳細地給出了關於每一類的預測準確度的描述。
5. **Confusion Matrix.** 給出了預測結果中每個類的實例數。其中矩陣的行是實際的類，矩陣的列是預測得到的類，矩陣元素就是相應測試樣本的個數。

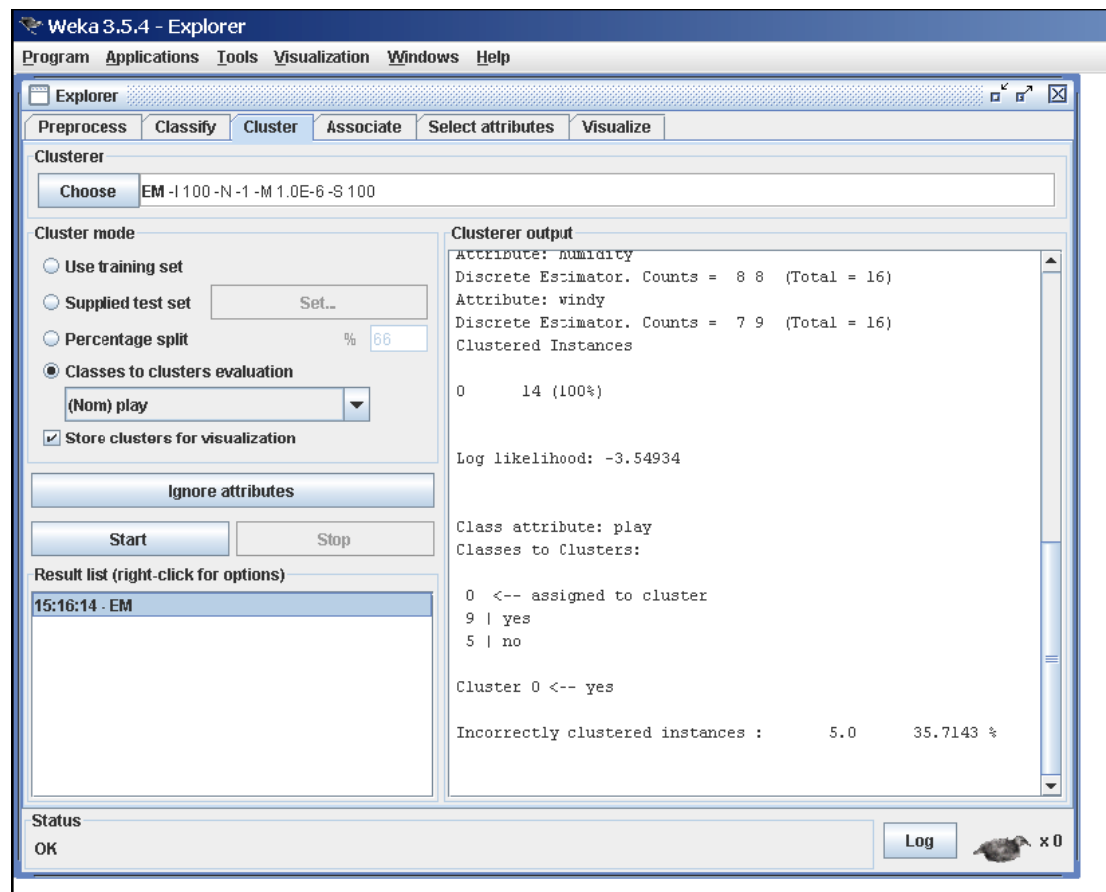
3.6 結果列表

在訓練了若干分類器之後，結果列表中也包含了若干個條目。左鍵點擊這些條目可以在生成的結果之間進行切換流覽。右鍵點擊某個條目則會彈出一個功能表，包括如下的選項：

1. **View in main window.** 在主視窗中顯示輸出該結果（就象左擊該條目一樣）。
2. **View in separate window.** 打開一個獨立的新視窗來顯示結果。
3. **Save result buffer.** 彈出一個對話方塊，使得輸出結果的文本可以保存成一個文本檔。
4. **Load model.** 從一個二進位檔案中載入以前訓練得到的模型物件。
5. **Save model.** 把模型物件保存到一個二進位檔案中。物件是以 Java “序列化”的形式保存的。
6. **Re-evaluate model on current test set.** 通過 Supplied test set 選項下的 Set 按鈕指定一個資料集，已建立的分類模型將在這個資料集上測試它的表現。
7. **Visualize classifier errors.** 彈出一個視覺化視窗，把分類結果做成一個散點圖。其中正確分類的結果用叉表示，分錯的結果用方框表示。
8. **Visualize tree or Visualize graph.** 如果可能的話，把分類模型的結構用圖形來表示（例如決策樹（decision tree）和貝葉斯網路（Bayesian network）模型）。圖形視覺化選項只有在貝葉斯網路模型建好之後才會出現。在流覽決策樹圖形時，可以在空白處右擊滑鼠彈出一個功能表，也可以拖動滑鼠來拖動決策樹，還可以在節點上按一下滑鼠查看它對應的訓練實例。Ctrl鍵+左鍵點擊會縮小圖形，Shift鍵+拖曳會得到一個方框並放大其中的圖形。這個圖形視覺化工具本身應該能夠解釋它的作用。
9. **Visualize margin curve.** 創建一個散點圖來顯示預測邊際值。這個邊際值的定義為：預測為真實值的概率與預測為真實值之外其它某類的最高概率之差。例如，提升式（boosting）演算法可以通過增加訓練資料上的邊際值來使得它在測試資料上表現得更好。
10. **Visualize threshold curve.** 生成一個散點圖，以演示預測時改變各類之間的閾值後取得的平衡。例如說，默認的閾值是0.5，那麼一個實例要預測成為“positive”，它是“positive”的預測概率必須大於0.5。這個曲線可以用來在 ROC 曲線分析中演示準確度/回饋率之間的平衡（正確的 positive 率對錯誤的 positive 率），也可用於其它類型的曲線。
11. **Visualize cost curve.** 生成一個散點圖，如 [1] 中所描述的那樣，給出期望價值（expected cost）的一個顯式表達。

在特定的情況下某些選項不適用時，它們會變成灰色。

4 聚類



4.1 選擇聚類器 (Clusterer)

現在我們應該熟悉選擇和配置物件的過程了。點擊列在視窗頂部的 **Clusterer** 欄中的聚類演算法，將彈出一個用來選擇新聚類演算法的 `GenericObjectEditor` 對話方塊。

4.2 聚類模式

Cluster Mode 一欄用來決定依據什麼來聚類以及如何評價聚類的結果。前三個選項和分類的情形是一樣的：**Use training set**、**Supplied test set** 和 **Percentage split**（見4.1節）——區別在於現在的資料是要聚集到某個類中，而不是預測為某個指定的類別。第四個模式，**Classes to clusters evaluation**，是要比較所得到的聚類與在資料中預先給出的類別吻合得怎樣。和 **Classify** 面板一樣，下方的下拉清單是用來選擇作為類別的屬性的。

在 **Cluster mode** 之外，有一個 **Store clusters for visualization** 的勾選框，該框決定了在訓練完演算法後可否對資料進行視覺化。對於非常大的資料集，記憶體可能成為瓶頸時，不勾選這一欄應該會有所幫助。

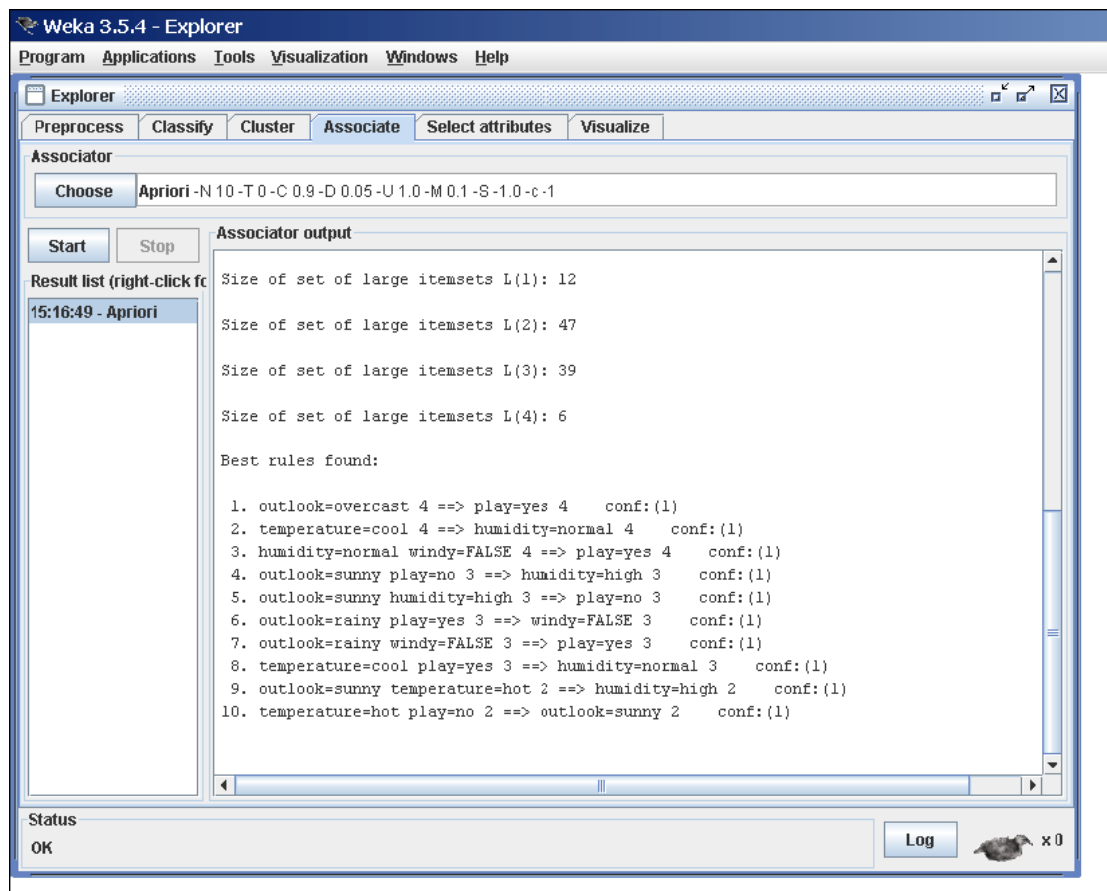
4.3 忽略屬性

在對一個資料集聚類時，經常遇到某些屬性應該被忽略的情況。**Ignore attributes** 可以彈出一個小視窗，選擇哪些是需要忽略的屬性。點擊視窗中單個屬性將使它高亮顯示，按住 **SHIFT** 鍵可以連續的選擇一串屬性，按住 **CTRL** 鍵可以決定各個屬性被選與否。點擊 **Cancel** 按鈕取消所作的選擇。點擊 **Select** 按鈕決定接受所作的選擇。下一次聚類演算法運行時，被選的屬性將被忽略。

4.4 學習聚類

Cluster 面板就像**Classify**面板那樣，有一個 **Start/Stop** 按鈕，一個結果文本的區域和一個結果列表。它們的用法都和分類時的一樣。右鍵點擊結果列表中的一個條目將彈出一個相似的功能表，只是它僅顯示兩個視覺化選項：**Visualize cluster assignments** 和 **Visualize tree**。後者在它不可用時會變灰。

5 關聯規則



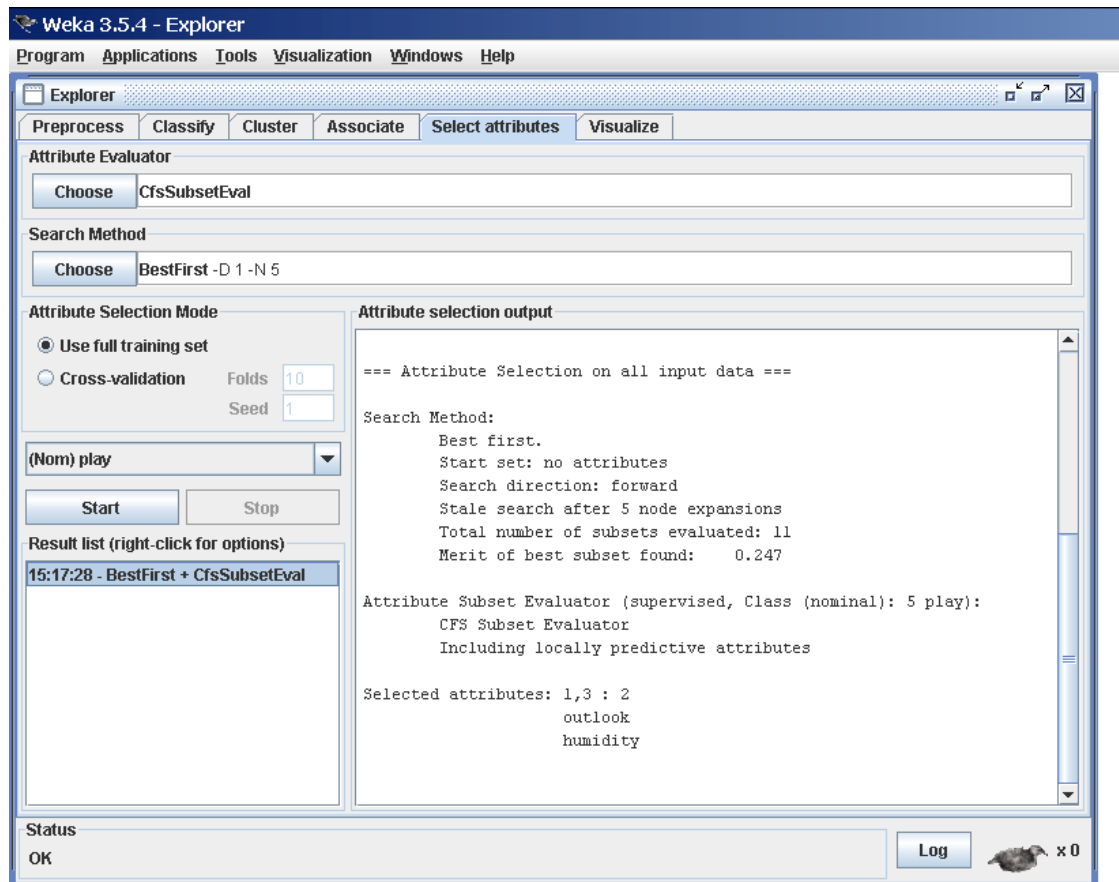
5.1 設定

這個面板包含了學習關聯規則的方案。這裡的學習器也可以跟其它面板的聚類器，篩選器和分類器一樣選擇和配置。

5.2 學習關聯規則

為關聯規則學習器設置好合適的參數後，點擊 **Start** 按鈕。完成後右鍵點擊結果列表中的條目可以查看或保存結果。

6 屬性選擇



6.1 搜索與評估

屬性選擇是說搜索資料集中全部屬性的所有可能組合，找出預測效果最好的那一組屬性。為實現這一目標，必須設定兩個東西：屬性評估器（evaluator）和搜索策略。評估器決定了怎樣給一組屬性安排一個表示它們好壞的值。搜索策略決定了要怎樣進行搜索。

6.2 選項

Attribute Selection Mode 一欄有兩個選項。

1. **Use full training set.** 使用訓練資料的全體好決定一組屬性的好壞。
 2. **Cross-validation.** 一組屬性的好壞通過一個交叉驗證過程來決定。**Fold** 和 **Seed** 分別給出了交叉驗證的折數和打亂資料時的隨機種子。
- 和 **Classify** 部分（4.1節）一樣，有一個下拉清單來指定 class 屬性。

6.3 執行選擇

點擊 **Start** 按鈕開始執行屬性選擇過程。它完成後，結果會輸出到結果區域中，同時結果列表中會增加一個條目。在結果列表上右擊，會給出若干選項。其中前面三個（**View in main window**, **View in separate window** 和 **Save result buffer**）和分類面板中是一樣的。還可以視覺化精簡過的資料集（**Visualize reduced data**），或者，如果使用過主成分分析那樣的屬性變換工具，則能視覺化變換過的資料集（**Visualize transformed data**）。精簡過/變換過的資料能夠通過 **Save reduced data...** 或 **Save transformed data...** 選項來保存。

如果想在精簡/變換訓練集的同時進行測試，而又不使用在分類器面板中的

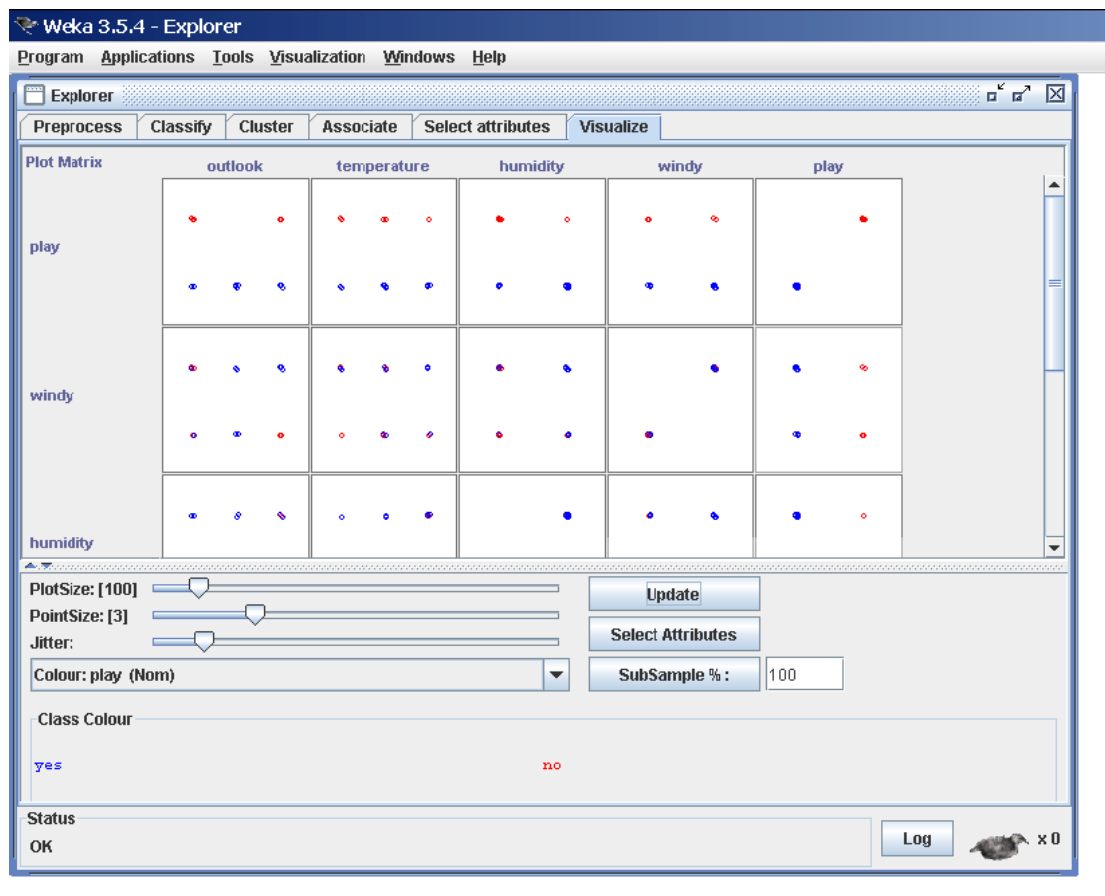
AttributeSelectedClassifier，那麼最好在命令列或者 SimpleCLI 中使用批量模式（“-b”）的 AttributeSelection 篩選器（這是一個 supervised attribute filter）。這一批量模式允許指定額外的輸入和輸出檔對（選項 -r 和 -s），處理它們的篩選器的設置是由訓練檔（由 -i 和 -o 選項給出）決定的。下面
是 Unix/Linux bash 中的一個例子：

```
java weka.filters.supervised.attribute.AttributeSelection \  
-E "weka.attributeSelection.CfsSubsetEval " \  
-S "weka.attributeSelection.BestFirst -D 1 -N 5" \  
-b \  
-i <input1.arff> \  
-o <output1.arff> \  
-r <input2.arff> \  
-s <output2.arff>
```

注意：

- z 每一樣末尾的反斜線告訴 bash 命令還沒有結束。使用 SimpleCLI 時必須把命令寫在同一行而不能使用反斜線。
- z 這裡假設 WEKA 已經在 CLASSPATH 中了，否則還要加上 -classpath 選項。
- z 整個篩選器的設置會在日誌中輸出，就像運行正規的屬性選擇時的設置一樣。

7 視覺化



WEKA 的視覺化頁面可以對當前的關係作二維散點圖式的視覺化流覽。

7.1 散點圖矩陣

選擇了 **Visualize** 面板後，會為所有的屬性給出一個散點圖矩陣，它們會根據所選的 class 屬性來著色。在這裡可以改變每個二維散點圖的大小，改變各點的大小，以及隨機地抖動（jitter）資料（使得被隱藏的點顯示出來）。也可以改變用來著色的屬性，可以只選擇一組屬性的子集放在散點圖矩陣中，還可以取出資料的一個子樣本。注意這些改變只有在點擊了 **Update** 了按鈕之後才會生效。

7.2 選擇單獨的二維散點圖

在散點圖矩陣的一個元素上點擊後，會彈出一個單獨的視窗對所選的散點圖進行可視化。（前面我們描述了如何在單獨的視窗中對某個特定的結果進行視覺化——例如分類誤差——這裡用了相同的視覺化控制項。）

資料點散佈在視窗的主要區域裡。上方是兩個下拉清單選擇用來選擇打點的坐標軸。左邊是用作 x 軸的屬性；右邊是用作 y 軸的屬性。

在 x 軸選擇器旁邊是一個下拉清單用來選擇著色的方案。它可以根據所選的屬性給點著色。在打點區域的下方，有圖例來說明每種顏色代表的是什麼值。如果這些值是離散的，可以通過點擊它們所彈出的新視窗來修改顏色。

打點區域的右邊有一些水準橫條。每一條代表著一個屬性，其中的點代表了屬性值的分佈。這些點隨機的在豎直方向散開，使得點的密集程度能被看出來。在這些橫條上點擊可以改變主圖所用的坐標軸。左鍵點擊改變 x 軸的屬性；右鍵點擊改變 y 軸的。橫條旁

邊的“X”和“Y”代表了當前的軸用的那個屬性（“B”則說明 x 軸和 y 軸都是它）。

屬性橫條的上方是一個標著 **Jitter** 的游標。它能隨機地使得散點圖中各點的位置發生偏移，也就是抖動。把它拖動到右邊可以增加抖動的幅度，這對識別點的密集程度很有用。如果不使用這樣的抖動，幾萬個點放在一起和單獨的一個點看起來會沒有區別。

7.3 選擇實例

很多時候利用視覺化工具選出一個資料的子集是有說明的。（例如在 **classify** 面板的 **UserClassifier**（自訂分類器），可以通過互動式的選取實例來構建一個分類器。）

在 y 軸選擇按鈕的下方是一個下拉按鈕，它決定選取實例的方法。可以通過以下四種方式選取資料點：

1. **Select Instance.** 點擊各資料點會打開一個視窗列出它的屬性值，如果點擊處的點超過一個，則更多組的屬性值也會列出來。
2. **Rectangle.** 通過拖動創建一個矩形，選取其中的點。
3. **Polygon.** 創建一個形式自由的多邊形並選取其中的點。左鍵點擊添加多邊形的頂點，右鍵點擊完成頂點設置。起始點和最終點會自動連接起來因此多邊形總是閉合的。
4. **Polyline.** 可以創建一條折線把它兩邊的點區分開。左鍵添加折線頂點，右鍵結束設置。折線總是打開的（與閉合的多邊形相反）。

使用 **Rectangle**，**Polygon** 或 **Polyline** 選取了散點圖的一個區域後，該區域會變成灰色。這時點擊 **Submit** 按鈕會移除落在灰色區域之外的所有實例。點擊 **Clear** 按鈕會清除所選區域而不對圖形產生任何影響。

如果所有的點都被從圖中移除，則 **Submit** 按鈕會變成 **Reset** 按鈕。這個按鈕能使前面所做的移除都被取消，圖形回到所有點都在的初始狀態。最後，點擊 **Save** 按鈕可把當前能看到的實例保存到一個新的 ARFF 檔中。

參考文獻

- [1] Drummond, C. and Holte, R. (2000) Explicitly representing expected cost: An alternative to ROC representation. *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Publishers, San Mateo, CA.
- [2] Witten, I.H. and Frank, E. (2005) *Data Mining: Practical machine learning tools and techniques*. 2nd edition Morgan Kaufmann, San Francisco.
- [3] *WekaWiki* – <http://weka.sourceforge.net/wiki/>
- [4] *WekaDoc* – <http://weka.sourceforge.net/wekadoc/>
- [5] Ensemble Selection on *WekaDoc* – http://weka.sourceforge.net/wekadoc/index.php/en:Ensemble_Selection