

MineSet™ 3.0 企業版 參考指南

文件號碼： 007-3558-001CHT

製作群

撰稿：Sandra Motroni 和 Helen Vanderberg

插圖：Dany Galgiani

製作：Linda Rae Sande

工程：Barry Becker、Amit Bleiweiss、Jeff Brainerd、Cliff Brunk、Eben Haber、Ara Jerahian、Andy Kar、Ed Karrels、Eser Kandogan、Alex Kozlov、Alan Norton、Peter Rathmann、Mario Schkolnick、Dan Sommerfield、Peter Welch 和 Brett Zane-Ulman。

(c)2000 Silicon Graphics, Inc.，版權所有

未經 Silicon Graphics, Inc. 書面許可，不得以任何方式複製或抄襲本文件的全部或部分內容。

有限和限制性權利圖例

政府部門對於該產品的使用、複製或公開受到以下條款的限制：FAR 52.227-14 中的「資料權利」條款和 / 或類似條款，或 FAR、DOD、DOE 或 NASA FAR 附錄中的後續條款。依據美國的版權法保留未出版發行的權利。合約商 / 製造商是 Silicon Graphics, Inc., 1600 Amphitheatre Pkwy., Mountain View, CA 94043-1351。

Silicon Graphics 是註冊商標，SGI、MineSet 和 Silicon Graphics 標誌都是 Silicon Graphics, Inc. 的商標。Oracle 是 Oracle 公司的註冊商標。Excel、Windows 和 Windows NT 都是 MicroSoft 公司的註冊商標。MATLAB 是 The Matchworks, Inc. 的商標。SPSS 是 SPSS, Inc. 的註冊商標。DBMS/COPY 是 Conceptual Software, Inc. 的商標。

「樹狀可視化工具」的美國專利號碼為 No. 5,528,735；5,555,354；5,671,381；和 5,861,885。「平板可視化工具」的美國專利號為 No. 5,861,891。「地圖可視化工具」、「分散可視化工具」和「平板可視化工具」中的 2D 滑動桿的專利正在申請中。「証據可視化工具」的專利正在申請中。「決策表」和「分散動畫」。

MineSet 3.0 企業版 參考指南

文件號碼：007-3558-001CHT

目錄

圖表清單	xiii
表格清單	xv
關於本指南	xvii
關於讀者	xvii
尋找資訊	xvii
文件的結構	xviii
本指南中的插圖	xviii
印刷慣例	xviii
1. 概念和說明	1
增加欄	1
增加欄按鈕	1
組合	4
陣列	6
分布欄	7
動畫	10
動畫控制面板	10
獨立維度控制滑動桿	10
動畫匯總視窗	12
動畫按鈕和滑動桿	13
動畫按鈕	13
動畫滑動桿	14
資料點和內插	14
運動軌跡	15

- 應用模型 15
 - 應用模型面板 16
 - 測試模式測試模式面板 17
 - 將資料擬合到模型面板 18
- 應用模型 18
- 關聯規則 20
- 關聯索引標籤 23
 - 檔案需求 23
 - 可視化關聯規則 23
- 關聯規則配置 24
 - 關聯規則選項 24
- 關聯規則對照按鈕 27
- 關聯規則可視化 28
 - 關聯規則的樣例檔案 29
 - 將檔案從 .ruleviz 轉換到 .scatterviz 29
- 自動分組 32
- 修正 33
- 分組 34
 - 分組選項 34
- 推進 (boosting) 37
- 更改欄類型或名稱 37
- 選擇點 38
- 分類器 38
- 分類器名稱 39
- 分類索引標籤 40
- 聚類 40
 - 用單步 k- 平均值方法聚類。 41
 - 重複 k- 平均值方法聚類 42
 - 聚類選項 44
 - 屬性權重 44
- 聚類選項對話方塊 45

聚類可視化工具	46
檔案需求	47
啓動聚類可視化工具	47
顏色選擇	48
用顏色選擇器來選擇顏色 (Windows)	48
用「顏色瀏覽器」來選擇顏色 (IRIX)	50
欄重要性索引標籤	52
欄重要性	52
尋找重要欄	53
分類器中欄重要性的不同	55
欄	57
指令行操作	57
配置檔案	58
混淆矩陣	58
成本複雜性	60
交叉驗證	61
資料整理	61
資料目標窗格	61
資料檔案索引標籤	62
資料導入	62
資料轉換面板	62
決策表	64
導入決策表	64
啓動決策表可視化工具	65
離散標籤	65
藉由「將欄對照到坐標軸」來「檢查資料」	66
解釋決策表	67
決策表選項	68
下拉式功能表	70
檢視功能表	70
名稱排序功能表	71

- 決策樹 71
 - 建立決策樹 72
 - IRIX 中的並行過程 73
 - 更多導入工具選項 73
 - 決策樹選項 73
 - 搜尋和篩選面板 76
- 離散標籤功能表 78
- 追溯 78
- 概化上尋和細化下尋 80
- 錯誤估計 80
- 証據模型 84
- 証據導入工具 84
 - 導入証據分類器 85
 - 啓動証據可視化工具 87
 - 証據導入工具選項 88
 - 証據可視化工具功能表 91
- 檔案功能表 92
 - Windows 系統 92
 - IRIX 系統 93
- 檔案需求 94
- 篩選按鈕 95
- 篩選面板 95
- 增量比 96
- 說明 (IRIX) 97
- 說明 (Windows) 97
- 柱狀圖可視化工具 98
- 歷史 98
- 預留 98
- 導入工具 99
- 工具管理員中的導入工具模式 100
 - 導入工具誤差選項 101
 - 進階導入工具選項 101
- 導入工具狀態視窗 102

國際化	104
在 IRIX 系統中設定地區	105
擴展到其它的語言和編碼 (僅適用 IRIX) 。擴展到其它的語言和編碼 (僅適用 IRIX) 。	105
重複 k- 平均值	107
拉普拉斯校正	108
學習曲線	108
上升曲線	110
損失矩陣	111
地圖可視化工具	115
地圖可視化工具的檔案需求	117
啟動地圖可視化工具	118
利用工具管理員來配置地圖可視化工具	119
產生 .gfx 和 .hierarchy 檔案	119
建立滑動桿和動畫	120
地圖可視化工具選項	121
地圖可視化工具檔案設定	123
挖掘工具索引標籤	123
多重選擇	124
公共資訊	124
簡單 -Bayes 方法	125
用視窗邊界工具在非樹狀可視化工具中瀏覽。	125
用視窗控件在樹狀可視化工具中瀏覽。	128
名稱排序功能表	129
標準化公共資訊	129
空	130
選項樹	131
導入選項樹	131
檔案需求	131
建立選項樹導入工具	132
IRIX 中的並行過程	132
選項樹選項	132
在 IRIX 系統中的「並行計算」	134

- 預測度 135
- 普遍度 135
- 刪改 136
- 隨機子 136
- 記錄檢視器 136
 - 啓動記錄檢視器 137
 - 對列重新編號 137
 - 在「記錄檢視器」中尋找 137
 - 儲存資料 137
- 記錄加權 138
- 回歸索引標籤 138
- 回歸樹 139
 - 導入「回歸樹」 139
 - 連續標籤 140
 - 回歸樹選項 140
 - 回歸程序中的錯誤估計 142
- 回歸器名稱 143
- 移除欄 143
- 投資回報曲線 143
- 儲存檔案 145
- 樣例檔案目錄 145
- 分散可視化工具 145
 - 檔案需求 146
 - 啓動分散可視化工具 147
 - 配置「分散可視化工具」 147
 - 爲「分散可視化工具」建立滑動桿 148
 - 分散可視化工具選項 148
 - 動畫控制面板 152
 - 在分散可視化工具中的空值處理 152
 - 樣例配置和資料檔案 152
 - 「選項」按鈕 153
- 爲地圖可視化、分散可視化、平板可視化建立的滑動桿。 154
- 排序欄名稱 154

平板可視化工具	155
平板可視化工具透明度	156
平板可視化工具檔案需求	158
啟動平板可視化工具	159
平板可視化工具形狀選項	160
儲存平板可視化工具設定	161
在平板可視化工具中的空值處理	162
為平板可視化工具建立的滑動桿	162
動畫控制面板	163
在平板可視化工具中的下拉式功能表	165
形狀功能表	165
配置和資料檔案樣例	166
拆分下限	166
拆分標準	167
統計可視化工具	168
如何讀取統計可視化工具	168
統計可視化工具下拉式功能表	170
統計可視化工具的檢視功能表	170
歷史表按鈕	171
「目前操作檢視為」欄位	171
「上一個」和「下一個」按鈕	171
工具管理員	175
「工具管理員」特性	176
訓練集	176
樹狀可視化工具	177
檔案需求	177
啟動樹狀可視化工具	178
樹狀可視化工具選項	178
儲存樹狀可視化工具設定	185

樹狀可視化工具下拉式功能表	185
檢視功能表	185
樹狀可視化工具選項功能表	191
樹狀可視化工具顯示功能表	192
跳至功能表 (樹狀可視化工具)	193
說明功能表	194
在樹狀可視化工具中的空值處理	194
樹狀可視化工具限制	195
樣例配置和資料檔案	196
修改因子	196
統一範圍	196
統一權重	196
檢視功能表	197
可視化過程工具	197
警告選項	199
Web 發布	199
加權	200
適應 Y2K 問題	200
A. 配置和資料檔案樣例	201
關聯規則樣例檔案	202
聚類樣例檔案	202
欄重要性樣例檔案	203

決策樹樣例檔案	204
客戶波動	205
轎車產地	205
預測性別	206
薪資因素	208
蝴蝶花分類	209
蘑菇分類	210
黨派隸屬	210
乳癌診斷	211
甲狀腺機能衰退診斷	211
Pima 糖尿病診斷	212
DNA 邊界	213
決策表樣例檔案	213
客戶波動	214
轎車產地	216
預測性別	217
薪資因素	218
蝴蝶花分類	221
蘑菇分類	222
黨派隸屬	223
乳癌診斷	224
甲狀腺機能衰退診斷	225
Pima 糖尿病診斷	226
DNA 邊界	227

証據可視化工具樣例檔案	227
客戶波動	228
轎車產地	229
預測性別	230
薪資因素	231
蝴蝶花分類	233
蘑菇分類	233
黨派隸屬	234
乳癌診斷	235
甲狀腺機能衰退診斷	236
Pima 糖尿病診斷	236
DNA 邊界	237
地圖可視化工具樣例檔案	238
選項樹樣例檔案	240
客戶波動	241
轎車產地	241
蝴蝶花分類	241
蘑菇分類	242
黨派隸屬	242
乳癌診斷	243
甲狀腺機能衰退診斷	243
DNA 邊界	243
回歸樹樣例檔案	243
客戶波動	244
轎車耗油量	244
薪資因素	245
蝴蝶花	246
糖尿病診斷	247
分散可視化工具樣例檔案	247
平板可視化工具樣例檔案	250
樹狀可視化工具樣例檔案	252
索引	255

圖表清單

■ 1-1	增加欄對話方塊	2
■ 1-2	組合對話方塊	5
■ 1-3	帶有匯總視窗和兩個滑動桿控制的動畫控制面板	11
■ 1-4	應用模型對話方塊：選擇分類器	16
■ 1-5	應用模型面板	17
■ 1-6	蝴蝶花資料集的錯誤劃分，範例 1	19
■ 1-7	Iris 資料集錯誤分類，範例 2	20
■ 1-8	初始的工具管理員視窗顯示多路關聯的產生	26
■ 1-9	關聯規則對照面板	27
■ 1-10	顏色開關列表	48
■ 1-11	顏色選擇對話方塊	48
■ 1-12	「顏色選擇器」對話方塊的 HSB 窗格	49
■ 1-13	「顏色選擇」對話方塊中的 RGB 窗格	50
■ 1-14	顏色開關列表	50
■ 1-15	顏色瀏覽器 (IRIX)	51
■ 1-16	Iris 資料集的混淆矩陣	59
■ 1-17	估計分類器的準確性	82
■ 1-18	分類器的交叉驗證 (k=3)	83
■ 1-19	模型的工具執行順序	99
■ 1-20	學習曲線	109
■ 1-21	上升曲線	111
■ 1-22	使用預設的設定值為蘑菇資料集建立的混淆矩陣	112
■ 1-23	帶有損失矩陣的蘑菇資料集之混淆矩陣	113
■ 1-24	帶有損失矩陣的蘑菇資料集混淆矩陣允許未知的預計	114

■ 1-25	地圖可視化工具樣例顯示 1990 年美國的人口分布	116
■ 1-26	對照到高度 (中上部的物件) 和對照到顏色 (右下部的物件) 的空值之方式	131
■ 1-27	投資回報曲線	144
■ 1-28	當 u 為高值或低值時透明度函數的形狀	156
■ 1-29	當 $u = 5.3$, 和 $u = 30$ 時的影像	157
■ 1-30	由統計可視化工具顯示的數字型欄	169
■ 1-31	統計可視化工具顯示的離散欄	170
■ 1-32	「歷史表按鈕」的「目前操作檢視為」欄位	171
■ 1-33	查看歷史對話方塊 (Windows)	173
■ 1-34	檢視歷史對話方塊 (IRIX)	174
■ 1-35	訓練集中的記錄樣例	176
■ 1-36	樹狀可視化工具配置選項對話方塊 (Windows)	179
■ 1-37	樹狀可視化工具配置選項對話方塊 (IRIX)	180
■ 1-38	樹狀可視化工具搜尋對話方塊 (Windows)	186
■ 1-39	樹狀可視化工具的搜尋對話方塊 (IRIX)	187
■ 1-40	在「樹狀可視化工具」中搜尋的「結果樣例」	189
■ 1-41	代表對照到高度、顏色、圓盤及標籤的「空值」	195
■ A-1	在客戶波動資料集中細化下尋	215
■ A-2	使用成人資料集的決策表可視化工具	219
■ A-3	成人資料集的進一步調查	220

表格清單

表 1-1	按照年齡組和州組合的消費模式	7
表 1-2	陣列以「年齡組」為索引	7
表 1-3	按照「年齡組」分布的欄	8
表 1-4	按照年齡組和薪資組分配的欄	8
表 1-5	總支出 (\$) 陣列中包括薪資組時的結果	9
表 1-6	利用年齡組和薪資組產生陣列的結果	9
表 1-7	利用 <i>薪資組</i> 進行分配以及利用 <i>年齡組</i> 進行索引的結果	9
表 1-8	關聯規則對照	28
表 1-9	自動分組選項	32
表 1-10	預設樣例檔案副檔名	94
表 1-11	「韓語」字型資源	106
表 1-12	在非樹狀可視化工具中的瀏覽按鈕	125
表 1-13	在非樹狀可視化工具中調節滑動桿和滑動輪	126
表 1-14	控制非樹狀可視化工具場景	127
表 1-15	樹狀可視化工具中的瀏覽圖示	128
表 1-16	在樹狀可視化工具中調節滑動桿和滑動輪	129
表 1-17	systune 參數	134
表 1-18	年齡在 40 到 50 之間	163
表 1-19	年齡在 50 到 60 之間	163
表 1-20	在表 1 和表 2 之間插值的中間過程	164

關於本指南

《*MineSet 3.0 企業版參考指南*》說明 MineSet 系統資料挖掘和可視化工具的技術特性和進階功能。MineSet 產品的目前資訊也可在下列網址中找到：
<http://www.sgi.com/software/mineset>。

關於讀者

該指南要求讀者已經從「工具管理員」中熟悉 MineSet 的操作。該指南的目的是向讀者闡述背景細節並說明 MineSet 如何運作。Windows 使用者將會發現指定的方法和路徑名是熟悉的。IRIX 的用戶應熟悉 UNIX 指令。

尋找資訊

該指南的大部分內容是探討從指令行和配置檔案中安裝和執行 MineSet 的有關問題。在一些樣例中有對應用程式介面的說明。每章的摘要可在「[文件的結構](#)」中找到。

如需使用 MineSet 工具的資訊，請參考《*MineSet for Windows 3.0 企業版使用者指南*》。

文件的結構

- 第 1 章，「[概念和說明](#)」
本章包括 MineSet 工具和概念的描述和解釋。
- 附錄 A，「[配置和資料檔案樣例](#)」
本章描述隨 MineSet 所附的資料和配置檔案樣例。

本指南中的插圖

該指南的大多數插圖來自 MineSet 3.0 for Windows。在一些樣例中，IRIX 和 Windows 的操作有本質的不同，本指南對兩種版本都進行說明。

印刷慣例

本指南將使用下列類型的慣例和符號：

- 斜體* 可執行檔名、檔名、程式變數、工具、公用程式、可變指令行參數、以及範例、代碼和語法語句中由使用者所提供的變數。
- 粗體** 關鍵字
- 固定寬度類型 螢幕上的指令行文字和提示
- 粗體固定寬度類型** 使用者輸入，包括鍵盤按鍵 (列印和非列印)；以及範例、代碼和語法語句中由使用者提供的文字。

概念和說明

增加欄

如果在傳送至分類器或可視化工具前，您想對資料集中的資料欄進行的新增、排除或排序操作可在「工具管理員」的「資料轉換」面板中執行。下面的「增加欄」按鈕項目描述「增加欄」的過程。「移除欄」以及「按欄名排序」會在各自的項目中討論。

增加欄按鈕

您可以用**增加欄**按鈕建立一個新欄，它的值來自於數學表達式的計算結果。例如，您可以新增一欄，其值為兩個現有欄的比值。按一下**增加欄**可得到一個對話方塊，讓您指定新欄的名稱和表達式 ([圖 1-1](#))。

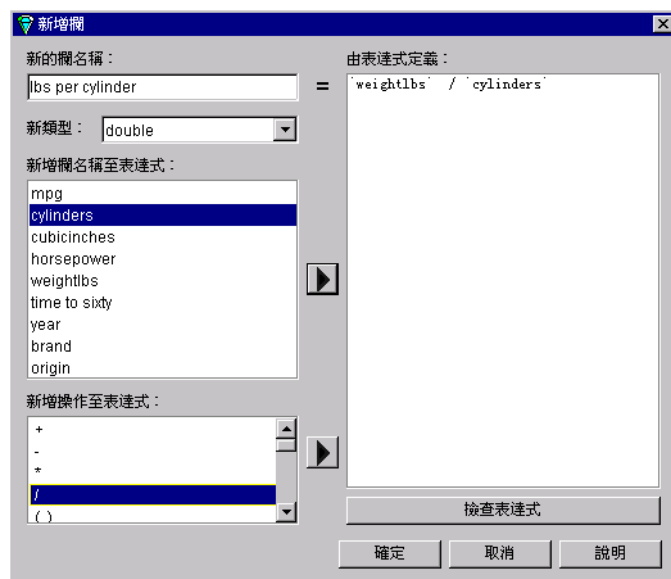


圖 1-1 增加欄對話方塊

對話方塊左邊有一個欄位可輸入新欄的名稱，跳現式功能表可指定欄的類型（整數型、字串型、浮點型等）。

對話方塊的右邊是大的文字輸入區域，可用來輸入定義了表達式。作為輸入列名和操作符的快捷方式，對話方塊左下方的滾動清單方塊中列出了目前表中的所有列和所有可能的操作符。要在表達式中插入一個欄名或操作符，可在捲軸清單方塊中連按兩下它，或先選取它然後按一下捲軸清單方塊右邊的箭頭按鈕。

在「篩選」和「增加欄」面板中使用的表達式語言與 C、C++ 和 Java 中使用的類似。基本操作符是一樣的：

+	加
-	減
*	相乘
/	除
()	括號
%	求餘數 (除法操作後所剩的餘數)
!	邏輯 NOT
~	邏輯 NOT
&&	邏輯 AND
	邏輯 OR
^	邏輯排除 OR
==	等於
!=	不等於
<=	小於或等於
<	小於
>=	大於或等於
>	大於
&	按位 AND
	按位 OR

表達式語言同樣提供下列操作：

isNull()	判斷括號中的值是否為空。
if()then()else()	如果第一對括號中的值為真，則該表達式傳回值就為第二對括號中的值。否則，傳回值就是最後一對括號中的值。
(x)?(y):(z)	C 語法 if/then/else
divide(x, y, z)	傳回 x 除以 y 的商，如果 y 等於 0，則傳回 0。

strlen(x)

該表達式傳回值等於字串中字元的個數。

子字串(x, y z)

傳回值是字串 x 的子串，該子串的主檢視位置從 y 算起，長度為 z。遵從 C、C++ 和 Java 中的慣例，字串第一個字元的索引號為 0。

要檢查您所建立的表達式，按一下 *檢查表達式按鈕*。如果有錯誤存在，就會出現一個對話方塊，指出是何錯誤及其位置。按一下 *確定*，就會自動檢查該表達式，並且對話方塊直到表達式修改正確後才會消失。

「增加欄」對話方塊可以檢查類型的相容性：如果您指定數字型表達式給一個字串欄 (反之亦然)，就會出現警告訊息，並會自動修正新欄的類型。

您還可以加入自己的使用者定義函數。請參閱 *《MineSet 3.0 企業版介面指南》* 中的「內置函數」。

組合

使用「工具管理員」中的 *組合* 按鈕需要對陣列及分布有基本的了解，請參閱「陣列」項目，該項目提供在組合功能中所使用概念的基本介紹。

您可利用「工具管理員」中的 *組合* 按鈕建立簡單的組合過程，產生陣列或分布欄。按一下該按鈕可產生「組合」對話方塊 (圖 1-2)。出現三個清單，在中間的清單中顯示目前表格中的欄。

要進行組合操作，選擇欄的名稱，然後按一下左邊和中間清單之間的左向箭頭按鈕。您可利用下面的跳現出式功能表指定索引值 (如果結果是陣列) 或一個分布欄 (如果結果已分布)。

要進行組合操作時，您可以利用底部的五個選項：求和、平均、最小最大值或計數。如果要對數值型欄進行組合，您可以選擇這些選項的任意搭配。至於其它類型，只允許計數。如果您選擇不止一個選項，就會得到多個結果。例如，選擇了平均和最大值運算，您將會得到一個平均值結果和另一個最大值結果。核取方塊「在組合中包括空值」允許您在計算中忽略或加入空值。

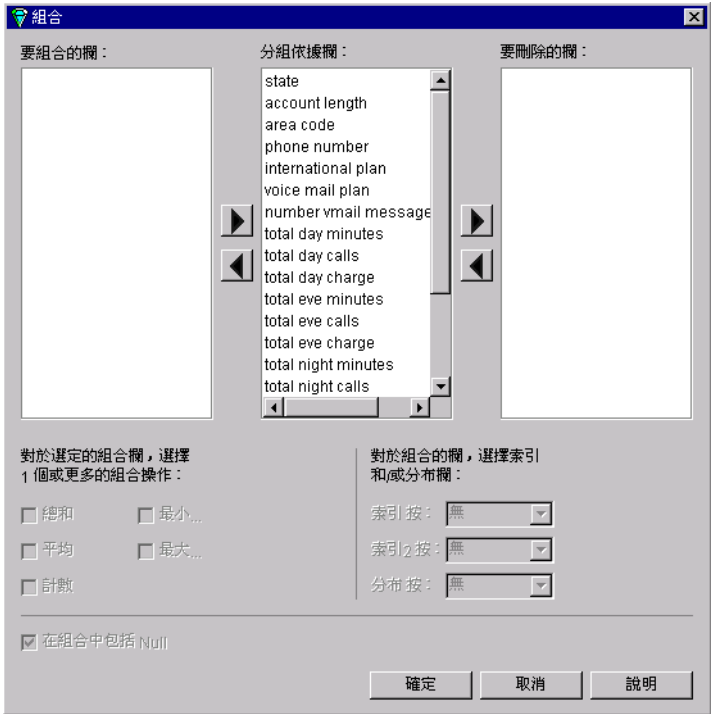


圖 1-2 組合對話方塊

下面提供欄名的三個清單：

- *要組合的欄。*
- *分組索引欄* (預設)；這樣可以保證在整個操作過程中該列表中的欄不發生變化。對於分組索引欄中具有相同組合值的每套記錄，在結果表中只能有一個記錄是輸出結果，其值出現在求和、最小值、最大值或計數組合欄中 (取決於面板底部的核選框)。
- *要移除的欄。*

在您用附加的組合標準對話方塊完成了後，「表格處理」視窗的目前欄本文方塊會顯示應用這些標準後所產生的新欄之名稱。

對話方塊右下方的功能表用於建立分布欄或陣列欄。您可利用這些索引和分配操作功能表指定一個索引陣列或分布欄。請參閱第 6 頁的「陣列」和第 7 頁的「分布欄」。

陣列

陣列是某一類型變數的集合，例如：浮點型、整數型、字元串等 (請參閱第 37 頁的「更改欄類型或名稱」取得可能類型的清單)。陣列都有一個索引，必須為已分組的欄。陣列可以是一維、二維、三維等等。一維陣列可視為清單。可以視 2D 陣列為表表格。陣列維度越高就越難將其可視化。

陣列對於「樹狀可視化工具」很有用處；如果您想自定在「分散可視化工具」、「平板可視化工具」及「地圖可視化工具」中所用的滑動桿，陣列是必要的。

例如，假設您的資料集是由根據年齡組和州所劃分的消費額來表示的。爲了限制儲存格數目，您可以將參與者的年齡分爲三組： 0-20, 21-40, 41-60. 結果表顯示在表 1-1。

表 1-1 按照年齡組和州組合的消費模式

州。	年齡 _ 組	總支出 (\$)
CA	0-20	\$50
CA	21-40	\$454
CA	41-60	\$693
NY	0-20	\$35
NY	21-40	\$541
NY	41-60	\$628

表 1-1 顯示六列，按照州分類。相同的值也可以用一維陣列欄表示，正如表 1-2 所示。在此種表示過程中，總支出 (\$) 成爲一個陣列，按照欄分組索引化 年齡 _ 組。

表 1-2 陣列以「年齡組」爲索引

州。	總支出 (\$) [年齡組]
CA	[\$50, \$454, \$693]
NY	[\$35, \$541, \$628]

分布欄

分布欄與陣列相似，但在幾個重要方面卻不同。分配操作爲每一個索引值產生新欄，取代產生一個新欄有許多值的方法。例如，如果在表 1-2 中的資料沒有成爲一個陣列，而是由年齡組進行分配，結果將顯示表 1-3。

表 1-3 按照「年齡組」分布的欄

州。	總支出 (\$) 0-20	總支出 (\$) 21-40	總支出 (\$) 41-60
CA	\$50	\$454	\$693
NY	\$35	\$541	\$628

分配樣例的範圍擴展 [表 1-1](#)，增加欄數而減少列數。

如果您有超過一個的欄分組 (例如，*年齡組*和*薪資組*)，您可以產生一個按照*年齡組*和*薪資組*組合索引的二維陣列。[表 1-4](#) 精修上面的樣例通過加入有區別特徵的*薪資組*欄 (只有代表 **California** 的行被顯示)。透過這些您可以檢視特定區域以及特定收入範圍中的消費總數。利用這些方法可以在資料集中尋找更精確的資訊。

表 1-4 按照年齡組和薪資組分配的欄

州。	年齡 _ 組	薪資組	總支出 (\$)
CA	0-20	\$0-\$25,000	\$30
CA	0-20	\$25,001-\$50,000	\$15
CA	0-20	超過 \$50,000	\$5
CA	21-40	\$0-\$25,000	\$120
CA	21-40	\$25,001-\$50,000	\$234
CA	21-40	超過 \$50,000	\$100
CA	41-60	\$0-\$25,000	\$101
CA	41-60	\$25,001-\$50,000	\$290
CA	41-60	超過 \$50,000	\$302

如果您將總支出 (\$) 變成陣列，在表 1-5 中顯示相同資料產生的結果：

表 1-5 總支出 (\$) 陣列中包括薪資組時的結果

州。	薪資組	總支出 (\$) [年齡組]
CA	\$0-\$25,000	[\$30, \$120, \$101]
CA	\$25,001-\$50,000	[\$15, \$234, \$290]
CA	超過 \$50,000	[\$5, \$100, \$302]

如果您想利用 年齡組 和 薪資組 產生陣列，結果將顯示在表 1-6：

表 1-6 利用年齡組和薪資組產生陣列的結果

州	總支出 (\$) [年齡組] [薪資組]
CA	[\$30, \$120, \$101, \$15, \$234, \$290, \$5, \$100, \$302]

最後，如果您利用 薪資組 來進行分配，以及利用 年齡組 來進行索引，表 1-7 將顯示出其結果：

表 1-7 利用 薪資組 進行分配以及利用 年齡組 進行索引的結果

州	總支出 (\$) [年齡組]，薪資 \$0-25,000	總支出 (\$) [年齡組]，薪資 \$25,001-\$50,000	總支出 (\$) [年齡組]，薪資超過 \$50,000
CA	[\$30, \$120, \$101]	[\$15, \$234, \$290]	[\$5, \$100, \$302]

對於每個陣列元件上面的樣例只有一個相關值，並且分配操作僅對存在的資料進行重組。MineSet 為那些輸出陣列元件含有多個值的資料集提供幾個組合選項。最普遍的選項就是求和。例如，當累計消費進行預算時這是很有用的。您也可以計算整個資料的最小、最大和平均值以及對其計數。

當對一個資料集進行上述操作 / 求和時，對於一個特殊的分組很有可能沒有合適的取值。在此種情況下，對於 *min*，*max*，*avg*，與 *sum* 組合操作，資料移動器都以空值處理。至於 *count* 操作，資料移動器為 0 值所填滿。

動畫

只要使用的資料集中至少有一個欄被對照到滑動桿元件時，「分散可視化工具」、「平板可視化工具」及「地圖可視化工具」會提供動畫功能。這就意味著您可以利用動畫顯示 1D 以上的資料集的變化，例如：時間。一般情況下，獨立的屬性，例如，時間或年齡是最好的滑動桿搭配，但是任何分組的欄都可以使用。

動畫控制面板

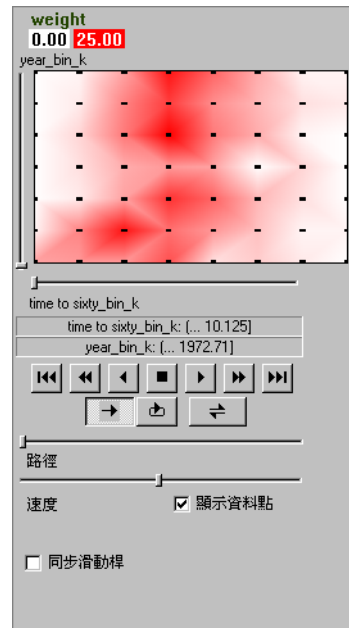
主可視化工具視窗右邊的動畫控制面板包含一個最多帶有兩個滑動桿的匯總視窗、一個資訊欄位、動畫按鈕、路徑滑動桿、速度滑動桿、資料點切換按鈕、同步滑動桿按鈕和「分散可視化工具」的管道軌跡功能表。

獨立維度控制滑動桿

在匯總視窗附近出現的滑動桿數目根據在可視化工具主視窗中所顯示的資料集。資料集可以有一、二維滑動桿或無滑動桿。

具有兩個獨立維度的資料集

如果資料集根據資料的不同具有獨立的 2D (例如 *company.scatterviz*)，則主視窗右邊的控制項將變為可見的 (請參閱圖 1-3)。



■ 1-3 帶有匯總視窗和兩個滑動桿控制的動畫控制面板

在主視窗的右邊是匯總視窗和滑動桿控制項。匯總視窗下面的水平滑動桿用於選擇第一獨立維度的資料點，而在其左面的垂直滑動桿用來選擇第二獨立維度的資料點。水平滑動桿所代表的維由其下面的標籤指明，垂直滑動桿所代表的維由其上面的標籤指明。

滑動桿在工具管理員中由以下一種方法指定：

- 對於「對照」、「分散」或「平板可視化」工具，將欄對照到「資料目標」窗格中的「滑動桿 1」和「滑動桿 2」元件，或者
- 利用「組合面板」建立一或兩個陣列欄。這些陣列欄將自動用於可視化工具的動畫滑動桿。

動畫匯總視窗

匯總視窗顯示全部的欄值，這些欄被對照為動畫滑動桿所有可能設定下的匯總可視化元件。也就是說，它顯示匯總屬性如何隨滑動桿變化。匯總視窗中越白的部分表示主視窗中實體所代表的匯總值越低。顏色越深，代表的值越高。

匯總視窗同樣顯示黑色的點，按照 1D 或 2D 範圍上間隔分布。這些點指示離散資料點的精確位置。您可利用「顯示資料點」核取方塊關閉這些黑色的點。

例如，當 *company.scatterviz* 檔案第一次開啓時，二維匯總視窗顯示從白（在左邊）到紅（在右邊）的顏色範圍。白色對應於低銷售量；紅色代表高組合銷售量。在該範例中，紅色越深則代表日常生活、汽車和家庭保險的銷售總值越高。

*動畫路徑*是指匯總視窗中代表資料移動到的黑點之間的路徑。在匯總視窗中有三種方法建立動畫路徑。

開啓檔案，可採用其中一種方法建立動畫路徑：

- 在匯總視窗中，選取一個黑點並定義起始點。按一下並按住滑鼠左鍵並將游標拖曳至視窗。鬆開滑鼠左鍵可結束路徑。
- 在匯總視窗中，按一下滑鼠左鍵來定義起始點。然後，移動游標至視窗的其它位置後按一下滑鼠中鍵來定義終止點。在兩點的中間出現一條直線。要加入更多的線段，繼續重複按一下滑鼠中鍵。三鍵滑鼠才能完成這一操作。
- 在匯總視窗中，按一下滑鼠左鍵來定義起始點。然後拖曳其中一個獨立維度滑動桿，沿著此維度畫一條直線。如果有兩個滑動桿，利用第二個滑動桿並在其控制下沿著坐標軸畫一條直線。

動畫按鈕和滑動桿

您可以利用在 2D 匯總視窗下面類似 VCR 的按鈕和滑動桿（路徑和速度）來控制動畫。

動畫按鈕

一旦在匯總視窗中繪製路徑（請參閱第 12 頁的「[動畫匯總視窗](#)」），您就可利用類似 VCR 的按鈕沿著路徑控制動畫。中間的**停止**按鈕以藍色反白標示，指示初始狀態。使用相鄰的**播放**按鈕（在**停止**的右邊）或**反向播放**（在左邊）所畫路徑中按向前或相反方向開始簡單運動。（**向前**和**反向**由路徑繪製的順序而定義，並不是從左向右或從右向左的運動。）

要停止和再啓動動畫，按一下**停止**按鈕，然後再按**播放**或**向後**按鈕。請注意：當您停止時，動畫繼續到最近的離散資料點。

在**播放**按鈕旁邊是**單步**按鈕，以及**單步向前**和**單步反向**按鈕。按一下其中一個按鈕將改變目前路徑位置到下一個離散資料點。

最外面的按鈕是**快速前進**和**快速反轉**按鈕。在**停止**狀態時按一下其中一個按鈕將使路徑位置改變到路徑的終止點（對於**向前**）或路徑起點（對於**反向**）。在**播放**狀態下按一下**快速**按鈕將加快動畫播放的速度。

動畫流程

在動畫按鈕的下面是三個動畫流程按鈕。

單次播放（預設）— 動畫向前或反向移運直到達到路徑的終點，然後停止。

循環— 當動畫到達路徑的終點時，會自動回到起點然後重新開始。

轉向— 當動畫到達路徑的終點時，它會反向沿著原路徑向另一端點運動；到達端點後動畫再次反向，循環也再次開始。

動畫滑動桿

經由分組和個種組合，可以自動或手動建立滑動桿，滑動桿可以自動或交互式的建立。該操作並沒有在「工具管理員」的目前歷史中顯示，但它們確實出現在工具的配置檔案中。

對照到「滑動桿 1」和「滑動桿 2」的欄最終形成在動畫中用到的欄之索引值（例如，顏色和大小）。這些欄必須是數字型（整數型、浮點型、雙精度型）或已分組。如果對照為滑動桿的欄已分組，該欄就不需自動分組，並且該欄會使用為滑動桿的索引。然而，如果該欄未分組，使用自動分組選項就可建立一個分組的欄。（請參閱第 34 頁的「分組」，以及個別工具項目可找到更多的資訊。）

當動畫停止時，您可以移動「路徑滑動桿」沿著路徑重新設定位置點。請注意：當您使用「路徑滑動桿」時，匯總視窗中的游標沿著畫好的路徑移動，而且繪製區域下面的和左邊的滑動桿將會隨游標的位置相應地移動。然後利用**播放**或**反向播放**按鈕從新指定的點重新開始動畫過程。您可以將「路徑滑動桿」拖曳到離散資料點之間的任意位置，然而當您放開滑動桿時，路徑位置就會移動到最近的離散資料點上。

使用「速度」滑動桿調節動畫沿著路徑播放的速度。

資料點和內插

當動畫播放時，在「分散」，「平板」或「地圖」可視化工具中對照成大小，顏色或坐標軸（位置）的組合變數在平滑的變化。然而，在「選擇」訊息方塊和游標位於欄位中所顯示的資訊只是位置最近的離散資料點的資料值，沒有顯示內插的資料值。

動畫按照下列的方式產生：如果您有了 1991 和 1992 年的資料，且它們對應於「分散可視化」工具中一個實體的大小。進一步假定 1991 年的大小為 20，1992 年的大小為 40，當您將年滑動桿從 1991 移向 1992 時，實體大小隨著 20 到 40 之間內插的結果而均勻變化。例如：1991 和 1992 的中點，其大小為 30，當您靠近 1992 時，大小就接近 40。然而，您無法終止離散資料點之間的動畫，並且您不能將路徑滑動桿拖曳到離散資料點之間的固定位置。

匯總視窗的資料點代表與資料檔案中實際資料相對應的位置。例如，20 和 40 可代表實際資料，但 30 則不能。在該範例中，在與每年相對應的滑動桿位置上匯總視窗中會有資料點。

請注意：並非所有的變數必須隨滑動桿的變化而變化。如果存在兩個滑動桿，一些變數會隨著其中一個滑動桿的變化而變化，而另外的變數將隨兩個滑動桿的變化而變化。

運動軌跡

在「分散可視化」工具中，軌跡功能表可以顯示動畫過程中運動點的軌跡。您可以選擇「線性形軌跡」、「淡出軌跡」、「管道軌跡」或「無軌跡」。請參閱 *《MineSet 3.0 for Windows 企業版使用者指南》* 中「在分散可視化工具和平板可視化工具中建立動畫」取得更多資訊。

應用模型

在「工具管理員」的「資料轉換面板」中的*應用模型*按鈕可以讓您：

- 採用以前建立的分類器並將它應用於新資料。
- 在目前資料表上測試先前建立的分類器。
- 將目前資料擬合到以前建立的分類工具的結構。

在對話方塊的左上部 (圖 1-4) 是目前伺服器上可用的所有分類器的清單。如果您選擇一個分類器，右邊會列出欄名稱和該分類器所需要的類型。如果這些必要條件與目前表格相符，底部的訊息將會提供提示。如果目前表沒有所選分類器所需的所有欄，底部的消息也會有相應提示：所遺失的欄會在右邊的清單中選取，並且底部的按鈕變為無效。



 1-4 應用模型對話方塊：選擇分類器

應用模型面板

「應用模型」面板是用來將以前建立的分類器應用於目前表格，如圖 1-5 所示。對於分類器有兩種應用模型：

- 為目前表格中的記錄*預測標籤值*。例如，如果您建立了一個分類器來測定客戶波動，您可以利用該選項增加一欄來標記每個客戶是否會產生客戶波動。
- *用於標籤值的估計機率值*。使用分類器來估計每個記錄含有指定標籤值的機率（例如，客戶波動 = yes），而不是用它來預測每筆記錄的標籤值。假設建立的分類器是用來測定客戶波動，您可以利用此選項來新增一欄，來表示每個客戶可能客戶波動的機率。

新欄名本文欄位可讓您指定新欄的名稱。

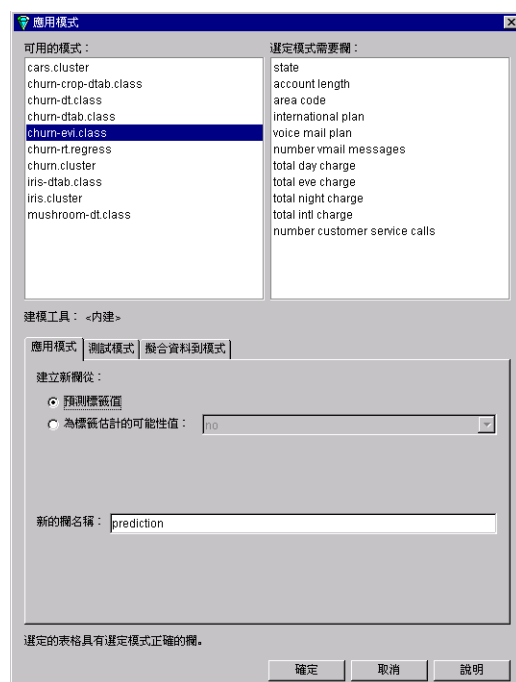


圖 1-5 應用模型面板

測試模式測試模式面板

使用「測試模型」面板在目前資料表中測試先前建立的分類器。該表必須包括選取的之分類器所需的一定名稱和類型的欄。不像「應用模型」，測試模型也要求表格包含一個標籤欄，該欄與建立分類器時所用的標籤欄擁有相同的名稱和類型。

「測試模型」面板所擁有的選項可以使您：

- 顯示表格記錄上應用分類器的混淆矩陣。
- 顯示指定標籤值應用分類器的上升曲線。
- 顯示指定標籤值應用分類器的 ROI 曲線。

- 使用作為測試模集的資料表，顯示分類器的可視化過程（這僅與「決策樹」和「選項樹」分類器有關）。
- 選擇用於記錄權重的屬性

「測試模型」面板底部的文字欄位顯示此結果。

將資料擬合到模型面板

使用「將資料擬合到模型面板」面板按以前建立的分類器調整目前表格中的資料。這將會產生一個與以前擁有相同結構的的新分類器；然而，新分類器利用表中的資料來更新機率估計（請參閱第 33 頁的「修正」）。因為表格中的所有資料都用來配合分類器的結構，所以沒有錯誤估計。「將資料擬合到模型面板」不能在使用推進建立的分類器上使用。當您想要在分離的測試集中評估新分類器的表現時（完全從適合資料中分離），可以使用測試模型。

「將資料擬合到模型面板」面板所具有的選項可以使您

- 顯示新分類器的可視化過程
- 為新分類器指定名稱
- 選擇用於記錄權重的屬性

應用模型

在建立預測模型後，您可將它應用於記錄以預測其標籤。例如，如果您建立了分類器（離散標籤的一種預測模型），對於預計蝴蝶花類型，您可以將分類器應用到只包含說明性屬性的記錄，就可以新增具有預測蝴蝶花類型的新欄。

為了確保資料質量，在建立分類器以後，為了找出分類器誤標的記錄，您可將它應用於訓練集。這些記錄可能會確保更進一步的調查。也許他們是「噪音」，或者會產生特別的直觀結果。

例如，假設導入蝴蝶花資料集的「決策樹」是使用*僅用於分類器*模式。如果您將分類器應用到資料集，將得到一個包含預測標籤的新欄 (`iris type_1`)。然後您可以增加新欄，以表達式 `(iris type != iris type_1)` 定義為 `int` 類型。無論分類器如何分類錯誤，新欄都為 1，在正確分類時為 0。圖 1-6 顯示「分散可視化工具」圖中的新欄以顏色設定對照為顏色時的資料，其中綠色為 0 (正確)，紅色為 1 (錯誤)。觀察這張圖，可以確定哪裡出錯。

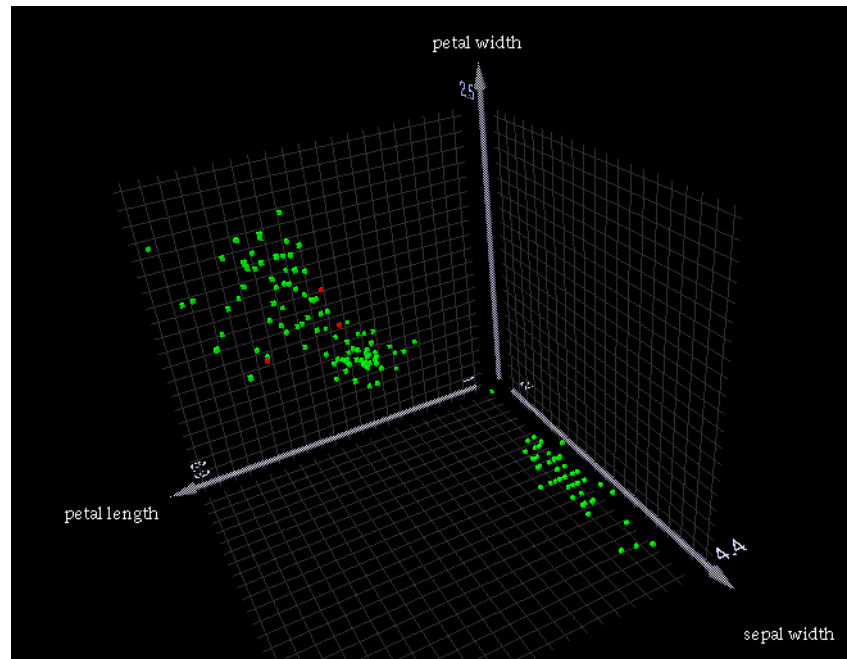


圖 1-6 蝴蝶花資料集的錯誤劃分，範例 1

另一個替代的方法是將新欄用表達式 `(iris type != iris type_1) + 0.01` 定義為浮點型。在原始標籤被照為顏色的條件下，可以使用「分散可視化工具」，並且該新欄對照到尺寸。不正確的預測顯示為大立方體；正確的預測顯示為小立方體 (請參閱圖 1-7)。

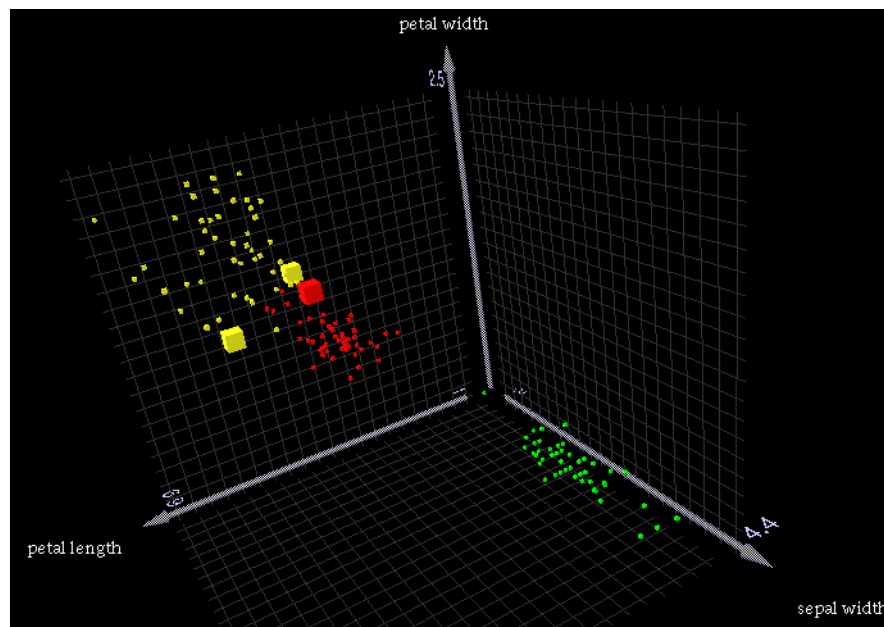


圖 1-7 Iris 資料集錯誤分類，範例 2

如需更多誤差和錯誤劃分的內容，請參考第 80 頁的「錯誤估計」。

關聯規則

大型資料集中，可以透過用關聯規則來建構、驗證及圖式化表達模式模型來進行資料挖掘。這些模式被關聯規則表達，這些規則指出在資料集中欄值一起出現的頻率。關聯規則的典型應用是市場供求分析，它將辨認哪些商品可能一起出現在消費者的「購物籃」中。

發現並圖式化關聯規則與許多行業有關。一些關聯規則可能產生的有用關聯的例子是超級市場存貨計劃、銷售計劃以及直銷中的附加郵寄。

在使用關聯規則時有兩個步驟：

1. 產生規則：資料檔案由「關聯規則產生器」處理，它可以建立由可視化工具使用的檔案。
2. 規則可視化：該操作顯示產生的關聯規則。

「關聯規則」可產生單一（一對一）和多路關聯規則。這部分描述單一關聯規則。關於多多路關聯規則規則的說明，請參閱第 25 頁的「多路規則」。

簡易的關聯規則可以這樣表達：如果 X 為真，則 Y 也肯定為真。MineSet 將 X 視為規則的左項 (LHS) 而 Y 為規則的右項 (RHS)。

應用關聯規則的範例就是得到消費者購買模式的市場供求資料。在此，市場供求就是消費者一次到商店所購買的商品。在此情形下的範例規則可能是：既買尿布又買嬰兒粉的人占百分之多少？這個百分數視為規則的置信度。

在該範例中，尿布是規則左項的項目 (LHS)，而嬰兒粉是規則右項的項目 (RHS)。

這些規則的一些應用有：

- 如果商品 A 出現在 RHS，而 LHS 可以幫助我們確定商店應該怎樣做才能促進其銷售。
- 如果商品 B 出現在 LHS，而 RHS 可以幫助我們確定如果商店不再經營商品 B，哪種商品將會受到影響。

「關聯規則產生器」處理一個輸入檔案，然後產生一個由規則組成的輸出檔案。如果 X 和 Y 是記錄中的屬性，則規則例如：

$X \Rightarrow Y$

表示無論 X 出現在哪個記錄中，則 Y 也很有可能一起出現。

關聯的強度由四個數字量化：

- 規則 *支援度* 量化了整個資料集中規則出現的頻繁度，或是 X 和 Y 作為整個記錄的一部分在資料集中一起出現的頻率。例如，如果支持為 1%，則 X 和 Y 在整個記錄的 1% 中出現。

您可以為產生的規則指定最小的支援臨界值。預設的最小支援度臨界值為 1%。最小支援度越小，產生的規則越多，工具的表現可能越慢。

滿足 *最小支援度臨界值* 的規則很重要，因為兩個原因：

- 只有當記錄中有相當的部分支援規則時，規則才會有商業價值。例如，如果每個買魚子醬的人同時也買伏特加酒，那麼規則魚子醬 $\Rightarrow$ 伏特加酒就有 100% 的置信度。但是，如果只有少數的人購買魚子醬（亦即支援度很低），則對於零售商來說規則就有侷限性。
 - 如果只有一小部分的記錄支援規則，則該規則就不具有統計顯著性。規則可能是因為偶然性，因此不要輕易根據規則做出判斷。
- 除以 LHS 規則出現的記錄數目後，規則的 *置信度* 量化了規則兩邊出現的記錄數目。例如，如果置信度為 50%，那麼 Y 在 X 所出現的記錄中有 50% 的出現機率。因此，如果知道 X 在一個記錄中出現，則 Y 也在同一記錄中出現的機率為 50%。您也可以為產生的規則指定最小置信度臨界值。最小置信度臨界值的預設值為 50%。
 - *期望置信度* 可以衡量那些左項和右項看似沒有關係的規則之置信度。它的計算是根據資料集中出現 RHS 項目的記錄數目。所以期望置信度與置信度之間的差別衡量了根據 LHS 項目表現而定的預測力的變化。
 - *上升* 代表置信度和期望置信度之間的比率。值越大代表規則越不可預測。這就告訴您，當試圖確定 RHS 是否在某一記錄中出現時，規則的 LHS 可以提供多少附加資訊。

關聯規則產生器不報告置信度小於期望置信度的規則。換句話說，如果 A 和 B 一起出現的頻率小於 B 單獨出現的頻率時，如 $A \Rightarrow B$ 的規則就不會列出。

注意：如果只給定 Y 和 $X \Rightarrow Y$ 類型的規則，至於 X 則一無所知。規則只指定從 LHS 到 RHS 之間的關係。

關聯索引標籤

「關聯」索引標籤位於「工具管理員」的「資料目標」窗格中的挖掘工具選項下，從中您可以產生關聯規則。這些規則，通常稱為市場需求分析，可以幫助您找出客觀現象的一般分類情況。

檔案需求

「關聯規則」需要下列檔案，這些檔案由「工具管理員」建立來產生規則可視化進程：

- 執行「關聯規則產生器」產生的規則檔案，命名為 *.rules.data* 檔案
- 帶有字尾 *.rules.schema* 的是結構檔案，因此您可以在「記錄檢視器」中檢視規則
- *.rules.scatterviz* 檔案

.rules 中綴並非必須，但是當這些檔案由「工具管理員」產生時便需要。

可視化關聯規則

有幾種方法可以啟動規則可視化工具：

- 利用「工具管理員」來配置和啟動「關聯規則」工具（請參閱第 24 頁的「[關聯規則配置](#)」）。一旦規則產生，「工具管理員」就自動啟動「分散可視化工具」。

- 如果您知道要用的配置檔案，可連按兩下檔案的圖示 (在檔案管理視窗中)。這樣就啓動「分散可視化工具」並自動載入您所指定的配置檔案。這只有當配置檔名以 `.scatterviz` 結尾時才起作用 (對於用「工具管理員」、「分散可視化工具」配置檔案都是這樣)。
- 輸入這個指令：
 - 在 DOS 指令行提示下：
`CD 檔案目錄`
`Viz [filename.scatterviz]`
 - 在 UNIX 介面指令行提示符下：
`scatterviz [filename.scatterviz]`

當啓動「分散可視化工具」時，您必須指定配置檔案，而不是資料檔案。

關聯規則配置

您可以利用「工具管理員」配置關聯規則的組成部分，這樣便可以大大地簡化任務。要在「工具管理員」的「資料目標」窗格中配置「關聯規則」檔案，選擇「挖掘工具」，然後按一下「關聯」。(關聯) 索引標籤。根據資料來源的不同，您可以分組或移除一些欄以簡化您的可視化過程。

關聯規則選項

關聯規則對話可讓您指定幾個選項。

置信度

除以 LHS 規則出現的記錄數後，該選項可定量描述規則兩邊出現的記錄數。例如，如果置信度為 50%，那麼 Y 在 X 所出現的記錄中有 50% 的出現機率。最小置信度臨界值的預設值為 50%。

支援度

該選項定量描述整個資料集中規則的普遍程度，或者說 X 和 Y 作為整個記錄數目的一部分在資料集中一起出現頻率。例如，如果支持為 1%，則 X 和 Y 在整個記錄的 1% 中出現。您可以為產生的規則指定最小支援度臨界值。預設的最小支援度臨界值為 1%。

加權

在下列情況下關聯規則可以允許記錄加權，即指定某些記錄比其它記錄更重要，或者因不均勻採樣而進行彌補時。如果沒有選取「使用權重」核取方塊，則每個記錄的權重皆為 1。當選取該核取方塊時，下拉式功能表可讓您選擇包含每個記錄權重的欄。*權重保留為屬性*核取方塊，如果選取的話，由「關聯規則產生器」所建立規則中將包含權重欄。如果位選取核取方塊，權重欄將從「產生器」建立的任何規則中排除。請參閱第 138 頁的「記錄加權」可找到進一步的說明。

多路規則

如果您選擇「多路規則」按鈕，「關聯規則產生器」產生所有滿足最小支援度和置信度臨界值的關聯規則，包括那些在 LHS 和 RHS 中擁有多個項目的規則。此種規則的範例可能是啤酒加空心粉，暗示著洋芋片、沙拉和白酒。

多路關聯規則產生帶有 LHS 和 / 或 RHS 上許多項目的規則。圖 1-8 說明為多路規則的產生而配置的工具管理員面板。顯示的版本來自 MineSet 3.0 for Windows。IRIX 版本是一樣的。

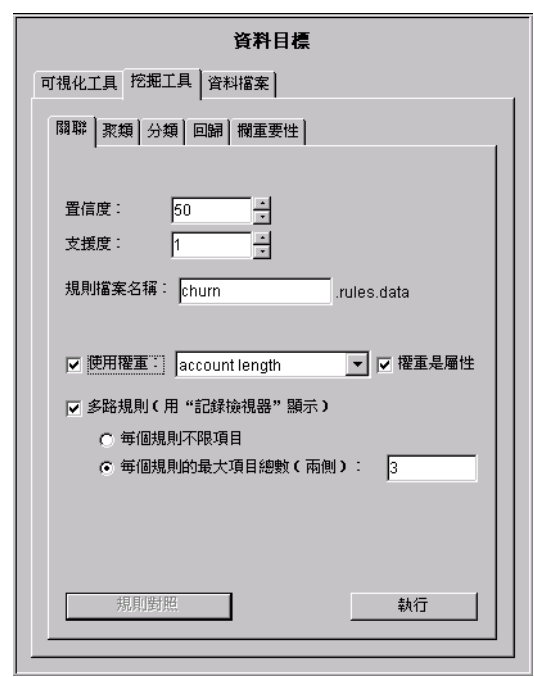


圖 1-8 初始的工具管理員視窗顯示多路關聯的產生。

使用「記錄檢視器」顯示多路規則而非「分散可視化工具」。每行顯示一條規則。表格的前兩欄包含 LHS 和 RHS 中項目的數目。接下來的四欄包含支援度、置信度、期望置信度和上升度值。最後兩欄包含 LHS 和 RHS 項目。在 LHS 和 RHS 欄中，項目之間由「and」分隔。在上面的樣例規則中，LHS 包含兩個代表啤酒和空心粉的項目。RHS 包含三個分別代表洋芋片、沙拉和白酒的項目。

您可以在每個規則的「最大項目數」欄位中輸入資料來限制所產生規則的大小。這個數字指出在任何規則中允許擁有的最大項目數。規則中的項目數是 LHS 和 RHS 中項數的總和。上述例子規則中有五個項目，簡易規則有兩項。

注意：產生多路規則要花費較多的時間。狀態視窗中指出每次重複產生的規則數目。如果產生太多的規則，刪除操作並增加最小支援度或置信度臨界值，或者降低在每個規則中項目的最大可允許項數。

關聯規則對照按鈕

對照規則可讓您將規則屬性對照為顯示內容中的可視元件。特殊的對照可以讓您檢驗假設以更進一步理解可視化過程。在「工具管理員」中的「資料目標」窗格中，按一下 *規則對照* 按鈕打開「關聯規則對照」面板在圖 1-9 中顯示。每個元件只顯示在跳現式功能表中的適當選擇。

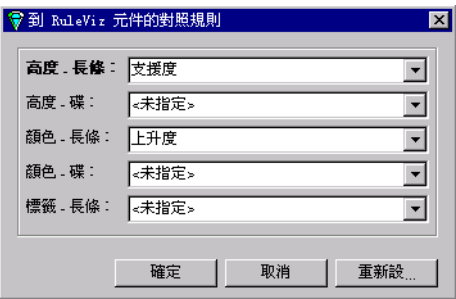


圖 1-9 關聯規則對照面板

您可以將從 *.rules.data* 檔案中自動產生的欄 (支援度、置信度、期望置信度和上升度) 對照為該面板中的可視元件；預設值顯示在表 1-8。請參閱第 24 頁的「關聯規則選項」可得到這些值的解釋：

表 1-8 關聯規則對照

可視元件	說明
高度 - 長條圖圖	可以指定長條圖高度所代表的意義 (預設情況是對照到支援度) 。
高度 - 盤	可指定盤高度所代表的意義。
彩色 - 長條圖	可指定長條圖顏色代表的意義 (預設情況下，對照至上升度) 。
顏色 - 盤	可以指定圓盤顏色所代表的含義。
標籤 - 長條圖	標籤 - 長條圖可讓您指定長條圖標籤所代表的意義。

關聯規則可視化

圖形化顯示關聯規則允許您開發和比較產生的規則。利用「分散可視化」工具在柵格場景中顯示規則。左項 (LHS) 項目在一個坐標軸上，而右項 (RHS) 項目在另一個上。規則的屬性和特徵在它的 LHS 和 RHS 項的結合處被顯示出來。顯示的內容包括長條圖、盤和標籤。

如果顯示的檢視太小，項目標籤不會出現在坐標軸上。可以用「伸縮」輪放大檢視直到項目標籤出現。您可以查看特殊規則的標籤，方法是當滑鼠處在選擇模式上時，將滑鼠指標放在單獨的列上。特殊規則的所有細節都會在檢視左上方的區域內顯示出來。

表示所示屬性 (例如，長條圖高度和顏色) 和與基本規則 (例如，置信度和支援度) 相關的值之間對照關係的圖例顯示在主視窗的底部。

長條圖高度對應於支持，而長條圖顏色對應於上升度值。當這一畫面放得足夠大時，LHS 和 RHS 軸會以項目名稱標示，除非這一項目已在配置檔案中關閉。

使用「工具管理員」(請參閱[表 1-8](#)) 或使用編輯器改變配置檔案，您可以修改標記以及長條圖和盤的高度和顏色所代表的內容。當一個變數對照到圓盤或長條圖顏色時，顏色對照便會自動產生。如果您希望改變這些預設的顏色對照，您可以編輯配置檔案。

如果將游標放在一個代表「關聯規則」物件的長條圖上，按一下滑鼠左鍵，資訊就會出現在「選擇視窗」中。按一下的同時按下 **Shift** 鍵就可以選擇多條規則。

追溯是一個術語，用來表示在可視化過程中觀察基本資料的行為。追溯表達式由對選定規則進行邏輯「與」的操作來決定。如果原始表中的欄與 `.rules.data` 檔案中的欄不相符，在執行追溯操作時，規則產生器產生一個特殊欄來幫助建立篩選表達式。這就意味著對追溯特性面板的修改無效，因為一個特殊的字串欄已經對照為對 `.rules.scatterviz` 檔案的追溯。

當您在一個或多個規則中追溯時，MineSet 顯示所有滿足規則的記錄。從可視化工具選項下拉式功能表中，選擇「顯示原始資料」。請參閱[第 78 頁](#)的「追溯」項目可得到關於追溯的更多資訊。

關聯規則的樣例檔案

MineSet 軟體包含樣例檔案以展示關聯規則的特徵和能力。請參閱[附錄 A](#)，「[配置和資料檔案樣例](#)」

將檔案從 `.ruleviz` 轉換到 `.scatterviz`

MineSet 2.6 版以前，關聯規則在自己的可視化工具中顯示，且配置檔案在格式上略有不同。

如果您有現存的 `.ruleviz` 檔案，便可將它們轉換成 `.scatterviz` 格式，方法是編輯現存的 `.ruleviz` 檔案並另存為 `.scatterviz` 檔案。[樣例 1-1](#) 和 [樣例 1-2](#) 說明了 `.ruleviz` 和 `.scatterviz` 格式之間的不同之處。[樣例 1-2](#) 有說明使用這樣變化的注釋。兩種配置檔案都使用相同的資料檔案。

注意：在舊的 `ruleviz` 檔案格式中，大小稱為高度，置信度稱為可預測度，支援度稱為普遍度。

樣例 1-1 group.ruleviz

MineSet 2,5

輸入

```
{
  檔案「Group.rules」
}
```

表達式

```
{
  double `pred/expected` = predictability/expected;
}
```

檢視

```
{
  高度可預測性；
  高度最大為 10；
  高度圖例開啓；

  期望的圓盤高度；
  圓盤高度圖例標籤「盤高度」：期望可預測性？

  顏色普遍度；
  顏色，多種顏色「白」，「紫」
  顏色比例 0 10；
  顏色圖例 ???0%?

  訊息暗示 %s\n 可預測性：%.2f 可預測性 / 期望：
    %.2f 普遍度：%.2f?LHS、RHS、可預測度、`pred/expected`、
    普遍度；

  選項柵格大小 3；
  選項隱藏圓盤距離 600；
  選項隱藏項距離 600；
}
```

樣例 1-2 group.rules.scatterviz

MineSet 3.0

輸入

```
{
  # 重新命名 group.rules 為 group.rules.data：
  檔案「Group.rules.data」
```

rules.data 檔案的結構為不變的

接下來。增加這些行：

```
int nlhs；
```

```

    int nrhs;
    浮點支援度;
    浮點置信度;
    float `expected confidence`;
    字串 LHS;
    字串 RHS;
}
expressions {
    float lift = confidence / `expected confidence`;
}

檢視
{
# 取代高度可預測性：
    規模置信度，比例為 1.；
    規模圖例標籤的長條圖高度：置信度？

# 取代期望的圓盤高度：
    圓盤高度「期望置信度」，比例為 1.；
    圓盤高度圖例標籤「盤高度」：期望置信度？

# 取代顏色流行度：
    顏色支援；
    顏色 多種顏色「白」、「紫」圖例標籤顏色：支援？
    顏色比例 0 9；
    顏色圖例 "0%" "9%";

# 增加雙軸對照（在舊檔案中不存在）：
    坐標軸 RHS，最大值為 100，按照字母順序排列；
    坐標軸 LHS，最大值為 100，按照字母順序排列；

# 確保形狀類型為長條圖：
    選項實體形狀長條圖；
    選項軸標籤大小 20；

    訊息暗示 %s\n support=%2.2f%%，confidence=%2.2f%%，
        期望置信度 %2.2f%%，lift=%2.2f"L、HS、RHS、支援度、置信度、
        「期望置信度」，上升；

    選項柵格顏色 "#202020";

    選項隱藏盤距離 600；
    選項隱藏實體標籤距離 600；
}

```

自動分組

「工具管理員」的「欄分組」面板給您選擇自動分組的選項。如果選擇您想要分組的欄，在 Windows 中選擇自動分組、或者在 IRIX 中從自動臨界值索引標籤中自動選擇分組的數目，則 MineSet 將會為您分組。當您想程式的機器學習功能為分組提供參考時，這是很有用的。請參閱第 34 頁的「分組」中分組選項的完整說明。表 1-9 列出可用的自動分組選項：

表 1-9 自動分組選項。

選項	Windows 位置	IRIX 位置	說明
自動分組	基本	N/A	自動分組所選的欄。
自動 x 選擇分組數	進階	自動臨界值	請求 MineSet 選擇分組的數目。
分為 _ 組	進階	自動臨界值	可讓您選擇分組數目。
使用方法	進階	自動臨界值	可在自動、統一範圍和統一權重之間選擇。
離散標籤	進階	N/A	可為分類選擇標籤。如果選擇自動方法，您必須從功能表中選擇標籤。
每組最小權重	進階	N/A	為了減少分組數，您可輸入每組的最小權重 (或在未設定權重的情況下輸入計數)。

在自動分組的過程中，選擇臨界值是為了使不同組中標籤的分配能儘可能地不同。這種方法持續產生臨界值來拆分範圍直到沒有明顯的多餘間隔。值的差別越大，選擇組數越多 (為對數關係)。

按一下「自動」核取方塊可通知 MineSet 根據實例中的總權重來自動確定最小權重：總權重越大，每組的最小權重就越高 (它們之間成對數關係)。

修正

修正可使您從一個訓練集中建立結構，然後對大資料集進行修正以改進機率估計。修正的任務就是使模型機率估計與基本資料一致。修正比導入大模型結構的速度快。當利用預留錯誤估計時，您省略對一部分資料的檢驗。當您經由模型結構修正所有資料時，最終模型的誤差便減小，因為計數、權重和機率反映整個資料集。您可以在任何導入工具「進階選項」面板中找到修正。

導入工具建立的模型擁有兩個部分：

- 結構「贊成決策樹」和「選項樹」，結構就是樹的形狀。對於證據，結構是每個屬性分組的數目，如果屬性為數值，其數目即為臨界值
- 機率估計，結構的每個特殊部分可用來估計每個分類出現的機率。這些估計的根據是在結構中不同點上訓練記錄的計數。對於「決策樹」，機率是由葉結點上面所記錄的權重來決定的。至於證據分類器，機率是由每個屬性值或區間的條件機率來決定的。條件機率是經由「證據可視化工具」左邊視窗中的矩形圖來顯示，該視窗顯示在指定（在 ... 條件之上）每個標籤值的情況下每個屬性間的相關機率。

用一套記錄來修正模型並不改變模型的結構，但會改變基於新資料的機率估計值。修正很有用，原因在於：

1. 可以從一個小的訓練集中建立結構，然後用大的資料集進行修正以改進結構中的機率估計。相對導入模型的結構，修正是一個更快的過程。
2. 當使用預留錯誤估計，則抽出一部分資料進行檢驗。一旦導入模型結構和估計誤差，則可以透過該結構對所有的資料進行修正，這樣可以減小最後模型的誤差。當計數、權重和機率出現在模型的結構時，它們反映所有資料的特徵而不僅是訓練集部分。

當使用可視化工具中的追溯功能時，可以看到對應於權重的資料顯示，它反映整個資料集的特徵。如果沒有使用修正功能，所顯示的權重僅代表訓練集。

對於所有的導入工具，您可以經由「資料目標」面板的「進階選項」按鈕來進入「修正實驗集」(IRIX 的「更多選項」)。在「資料目標」窗格中按一下「挖掘工具」並選擇「分類」或「回歸」索引標籤，然後為任何導入工具選擇「分類器」(或「回歸工具」)和誤差模式。當「加速」開啓時，修正核取標記就停用。

分組

分組可以把一個或多個欄中的資訊分組，處理過程中建立一個新欄(例如，關於年齡範圍的一欄 0-18、19-25、26-35 等等)。「工具管理員」中分組可節省計算時間並簡化可視化過程。如需分欄的細節，可請參閱《*MineSet 3.0 for Windows 企業版使用者指南*》中「為欄改變或建立新組」。

分組選項

對任一選取的欄，Windows 分組選項對話方塊初始會提供三種選擇：「不分組」、「自動分組」和「自定臨界值」。IRIX 版本提供兩種選擇，「自動臨界值」和「使用者指定臨界值」。

分組記號

MineSet 對組名使用修改的間隔記號：

(較低的臨界值 ... 較高的臨界值]

在較低臨界值旁邊的括號「)」表示該臨界值不包含在範圍之內。方括號，「]」指示臨界值包含在範圍之內。例如，(10.5 ...12.6] 表示值的範圍大於 10.5 並且包括 12.6。如果省略較低的臨界值，範圍就為小於並包含較高邊界的所有值。例如，(... 10.5] 即表示值的範圍小於或等於 10.5。如果省略較高的臨界值，範圍就為大於等於較低邊界的所有值。例如，(12.6 ...] 即表示值的範圍大於 12.6。

分組方法

將資料按類分組有三種方法：

- 必須為每個選擇離散標籤。臨界值的選擇是爲了使不同組中的標籤分布儘可能的不同。這種方法持續產生臨界值來拆分範圍直到沒有明顯的多餘間隔。

在每組最小權重文字欄位中，可以指定任何組中的最小權重；這就防止用比指定資料更小的權重來建立分組。如果兩個結果子區間中沒有包含指定的最小權重，則不會對區間進行拆分。預設情況下，每個記錄只有一個權重。在這種情況下，指定每組中「最小權重」與指定每組實例的最小數目一樣。

利用算法自動的設定值較爲每組指定最小權重的做法更好。根據記錄權重的總和，使用權重核取方塊可利用該算法爲每組計算最小權重：總權重越大，每組的最小權重就越高（它們之間成對數關係）。

- **統一範圍**算法將值範圍分解成指定的相同大小子區間的數目。極限範圍的較低和較高邊界包括資料觀察範圍之外的所有值。例如，如果屬性的值在 3-8 之間，並且您指定了四個組，那麼所找出的臨界值爲 4.25、5.5 和 6.75，對應於範圍：

- 小於 4.25
- 大於 5.5 小於等於 6.75
- 大於 5.5 小於等於 6.75
- 大於 6.75

極限範圍的較低和較高邊界包括資料觀察範圍之外的所有值。MineSet 這些組名的記號爲：

- (... 4.25]
- (4.25 ... 5.5]
- (5.5 ... 6.75],
- (6.75...]

- **統一權重**算法將值範圍分為指定數量的等權重組。與「統一範圍」在值域中臨界值將值劃分為等大小間隔的區間不一樣，而「統一權重」決定了將記錄分成等權重子集組的組臨界值。預設情況下，每個記錄只有一個權重。在此這種情況下，統一權重方法產生指定的組數，每組包含大約相等的實例數。

統一範圍和統一權重都可讓您指定修改因子，修改因子表示在分組之前排除在值範圍外的極值部分。預設的修改因子為 0.05。這就排除了 5% 的極值記錄 (其中 2.5% 為範圍中的最低值，2.5% 為範圍中最高值)。修改的目的就是減小在臨界值產生過程中異常值的影響。

所有的方法都可以讓您決定是自行指定組數或是讓算法自動選擇數目。至於「統一範圍」和「統一權重」進展，自動分組過程是根據相異值的數目來進行的：值的差別越大，選擇組越多 (為對數關係)。

一般情況下，在確定臨界值時所有可用的實例都將使用。當備分組的屬性後來用於導入模型時，錯誤估計指標將為最優。這是因為實驗集中的分布資訊用於確定臨界值。

*僅使用訓練集*阻止分組方法在識別臨界值時看到檢驗集中的記錄。這就提供分類器錯誤率的更實際的估計。*只使用訓練集*需要使用者指定相同的「預留率」和「隨機子」(請參閱第 101 頁的「導入工具誤差選項」) 用於為估計分類器誤差而建立的預留集。

「使用權重」功能表可以讓您用任何數字型屬性對實例進行加權。改變記錄加權會影響「自動」或「統一權重」方法，但不會影響「統一範圍」方法。

如果您按一下*應用*，「工具管理員」會挑選分組臨界值並將其顯示在用於選擇欄臨界值的本文欄位中。在「欄分組」視窗底部的文字欄位顯示分組算法的進度和任何出現的誤差。

推進 (boosting)

推進是一種算法，該算法建立幾個不同模型並利用加權投票法結合它們的預測過程。推進模型通過在很難應用於模型的資料中將歸納過程集中在樣例上的方式來經常改進準確性。在所有使用「推進」（無可視化）核取方塊的導入工具模式中，可用「更多導入工具」選項對話方塊中來開啓「推進」。

在一些情況下，在建立模型過程中最重要的問題是錯誤率。例如，假設您已經分析了資料集的客戶波動預計，到達了較為滿意程度，並準備模型，用來預計用戶中最有建立可能的波動者。在這點上，您不再對模型的可視化感興趣，因為您對相關的因素有了相當好的理解。您也想得到最佳的分類準確度。在此種情況下，您也許想要開啓推進。

推進並不一定會增加準確度，但是通常會這樣。無法可視化推進模型，雖然仍可看到其混淆矩陣、上升曲線、學習曲線和 ROI 曲線。推進是密集的計算過程，通常要比沒有推進的對應導入工具的計算時間長 25 倍。修正並不與推進分類器一起工作，這是因為特殊方法推進權重重複模型和用於訓練它們的記錄。

您可對帶有可取值數目的標籤使用推進。推進並不一定會減少導入模型的錯誤率；對於有多值標籤帶來的問題要特別重視此點。

更改欄類型或名稱

該「工具管理員」選項可使您改變欄的類型和名稱。請參閱 *《MineSet 3.0 for Windows 企業版使用者指南》* 中的「更改欄類型和名稱」得到更多資訊。

選擇點

當您從「工具管理員挖掘工具」的「聚類」索引標籤中按一下重複 **k**- 平均值時，選擇點是一個進階顯示選項。選擇點是聚類參考數目和 0 之間的一個點，較高的值建議了較大的類別數而較低值則建議了較小的類。選擇點為 1.0 時都會選取上邊界。在聚類過程中，如果您的界限是 1 到 5 個聚類，則選擇點為 0.4 時將選出兩個聚類，而當選擇點為 0.8 時將選出四個聚類。選擇點為 1.0 時都會選取五個。

分類器

分類器在指定數個其他屬性的情況下，預測一套資料的某一個屬性。分類器是模型的一種。屬性是資料集的本質特徵。例如，如果您有通訊公司的用戶資料，而且假設您已知用戶是否有聲音信件、出國打算以及使用電話的時間，就可利用分類器模型來預測用戶是否波動。已預測的屬性稱為標籤，用於預測的屬性則稱為描述性屬性。

MineSet 可以從訓練集中自動建立分類器。訓練集由提供標籤的資料中之記錄所組成 (也可參閱 [第 176 頁的「訓練集」](#))。例如，您為每個說明性屬性提供了只含有一欄的資料表 (例如，聲音信件計劃的現況，每天打電話的平均分鐘數)，該欄的內容為標籤 (客戶波動或不波動)。從訓練集中自動建立分類器的算法稱為導入工具。

當產生分類器時，MineSet 也產生可視化過程來幫助了解分類器如何操作。可視化過程也可以為資料本身提供有價值的直觀結果。一旦產生分類器，它就可可在標籤值未知的情況下對記錄分類。該值由分類器來預測。

分類器有兩個部分：結構和機率估計

- 結構「贊成決策樹」和「選項樹」，結構就是樹的形狀。對於證據，結構是每個屬性分組的數目，如果屬性為數值，其數目即為臨界值「決策表」的數學結構與「決策表」的相同，但其顯示與「證據」可視化過程相似。
- 機率估計結構的每一部分都估計了每個分類的機率。這些估計的根據是結構中不同點上訓練的記錄數目。對於「決策樹」，機率是由葉結點上面所記錄的權重來決定的。至於證據分類器，機率是由每個屬性值或區間的條件機率來決定的。當忽略其它所有的屬性時，先驗機率意味著在訓練資料集裡的隨機選擇記錄中發現標以例如糖尿病的特殊分類的機率，換句話說，即為帶有分類標籤的記錄總數除以資料集中記錄的總數。條件機率是，當您選擇糖尿病標籤時，落在例如 *age_60+* 分類中特殊記錄的機率（即建立在選擇標籤的條件上）。

注意：請參閱 *《MineSet 3.0 企業版介面指南》* 附錄 A 中關於分類工具的進一步讀物以及確認在 MineSet 樣例檔案裡所使用的資料集清單。

分類器名稱

正如「工具管理員」中所確定的，產生的分類器以階段檔名作為字首，至於合適的字尾，例如，*-dtable.class* 是指「決策表」的分類器，*-dt.class* 是指「決策樹」分類器。預設情況下，所有的分類器都儲存在伺服器的 *file_cache* 目錄，預設是 *mineset_files*。這些分類器可用於日後分類未標記的記錄；也就是說，它們可以用來對未標記的資料集預計標籤（請參閱第 15 頁的「應用模型」和 第 33 頁的「修正」）。

分類索引標籤

要進入 MineSet 分類器，在「工具管理員」的「資料目標」窗格中按一下「挖掘工具」索引標籤可找到「分類」索引標籤。您可在四個模式中選擇一個：「分類器和誤差估計」、「僅用於分類器」、「估計誤差」和「學習曲線」。這些模式都可與各種導入工具進行組合：「決策樹」、「選項樹」、「證據」和「決策表」。如需每個導入工具的詳細資訊，請參閱指南中對應的項目。如需使用這些分類器時的指示，請參閱 *《MineSet 3.0 for Windows 企業版使用者指南》*。

聚類

聚類索引標籤可使您利用 MineSet 的聚類功能，對於開發不熟悉資料集來說是非常有用的挖掘工具。因為聚類是一種描述性的挖掘工作與發現關聯規則相似，您不需將特殊的欄指定為標籤。除此之外，資料集從不分為訓練和檢驗集；分類模型總是從整個資料集中建立並進行評估。

聚類模型儲存為一套原形記錄，每類一個。這些代表聚類中所有資料的加權平均並認為是聚類中心或重心。不像標準資料庫記錄，聚類中心對於數字型欄以總結統計的形式，對於類目欄以柱狀圖的形式為每一欄保持分布。

從「工具管理員」中「資料目標」窗格的「挖掘工具」索引標籤中可執行聚類。聚類的目的是確定資料集中有哪些相似的特徵。然後，您可以看到聚類結果並以不同的參數進行實驗。

MineSet 中的所有聚類都在 k- 平均值方法的基礎上使用合併算法；算法將相似的記錄分組形成分類，在每個分組中其目標是使整體的相似性最大。

要使用聚類工具，在「工具管理員」主視窗中的「資料目標」窗格中選擇「挖掘工具」索引標籤。從索引標籤中選擇聚類。主要分類器面板就會顯露出來。

您可以從主「聚類」面板中按一下 *開始* 來啟動聚類過程；並沒有其它必需的選項。預設情況下，您將利用單一 **k- 平均值** 聚類方法來發現資料中的三個聚類。一旦聚類完成後，您將看到聚類的評估結果，然後「聚類可視化工具」將出現。

聚類面板提供下列選項：

- 允許在單步 **k- 平均值** 和重複 **k- 平均值** 方法之間進行選擇。預設情況下是單步 **k- 平均值**。兩種方法都在下面進行詳細的說明。
- 對於單步 **k- 平均值** 法的聚類數目，您必須指定要尋找的聚類數目。預設值為 3。
- 對於重複 **k- 平均值** 方法的聚類數目範圍，您必須指定聚類可能數目的上限和下限。預設為 1 ... 10.
- 對於重複 **k- 平均值** 方法的選擇點，您必須指定 *一個選擇點*。該值在 0 和 1 之間，並有助於選擇聚類的最終數目。請參閱第 42 頁的「**重複 k- 平均值方法聚類**」。預設值為 0.5。

注意：聚類是一個計算密集過程，要在較大的資料集上完成需花費一些時間，尤其在執行重複 **k- 平均值** 的模式下。如果資料集超過 10,000 筆記錄，最好對資料的樣本進行聚類。

用單步 **k- 平均值** 方法聚類。

術語 **k- 平均值** 是指一種方法，該函數根據記錄相似性來確定儘可能不錯的聚類方案。這種方法是 **MineSet** 中聚類的最簡單的形式。您指定所需的聚類數目，並且該算法將資料中的記錄分組使每個聚類中的整體離散度最小。離散度是指聚類的凝聚性；離散度越高，每個記錄離聚類中心越遠。從技術上，離散度可以經由記錄到聚類中心所指定的距離平方根的大小來度量。

算法本身是重複性的，其處理過程如下：

1. 例如，可選擇尋找五個聚類。
2. 五個聚類的中心在記錄空間中的隨機位置初始化。隨機子參數的選擇會導致不同的開始位置。

3. 資料中的每筆記錄都分配離自己中心最近的聚類。然後在每類新資料的基礎上重新計算聚類的中心位置。
4. 如果記錄與其它中心間的距離比其與所屬聚類中心間的距離還要近，則記錄將移至最近的聚類中。然後在每類新資料的基礎上重新計算聚類的中心位置。這一步驟將一直重複直到沒有新的改進。

這一算法確保在有限的重複次數中完成計算。

步驟 4 決定聚類的執行時間。因此，過程視窗在每執行一次步驟 4 時出現一個進度列，並伴有記錄指出到目前為止已進行多少次重複。您可以設定在計算停止前所允許的最小重複次數限制 (執行步驟 4)。預設值為 20。

聚類的名稱從 1 開始編號，雖然聚類名稱由數字代表，但對於聚類來說並不代表順序。

重複 k- 平均值方法聚類

重複 k- 平均值聚類是單步 k- 平均值聚類方法的一個更複雜的延伸，為「工具管理員」的「聚類」索引標籤上的一種方法。不像單步 k- 平均值方法，它並不需要指定所需建立的聚類之確切數目，但是需要下限和上限以及一個選擇點。該算法將在上下限之間挑選適合於資料集的一些聚類。選擇點是聚類參考數目和 0 之間的一個點，較高的值代表較大量的聚類而較低值則代表較小量的聚類。選擇點為 1.0 時都會選取上邊界。例如，如果界限為一和五個聚類，當選擇點為 0.4 時將選出兩個聚類，而當選擇點為 0.8 時將選出四個聚類。選擇點為 1.0 時都會選取五個。

要執行重複 k- 平均值方法，您要選擇三個參數：聚類數目的下限 (預設為 1)，聚類數目的上限 (預設為 10)，以及一個選擇點 (預設為 0.5)。有了這些參數，算法進行如下：

1. 利用聚類最小數目來執行單步 k - 平均值算法。這將產生初始聚類結果。
2. 找出離散度最大的聚類，並將它一分為二，建立兩個新聚類。初始聚類的一半記錄會分配到新聚類中，而另一半則分配給另一新聚類中。在所含資料的基礎上重新計算新聚類的中心。
3. 如果記錄與其它中心間的距離比其與現屬聚類中心間的距離還要近，則記錄將移至最近的聚類中。然後在每類新資料的基礎上重新計算聚類的中心位置。這與單步 k - 平均值方法的步驟 4 一樣。在單一 k - 均值算法，該步驟一直被重複直到沒有記錄移動或執行了所允許的最大的次數，該數值由最大重複次數參數指定 (預設為 20)。

步驟 2 和 3 一直重複著直到產生最大聚類數目為止：

該過程可在最大和最小聚類數目之間產生一系列聚類結果。

最終的呈現出聚類結果由選擇點參數來確定 (預設為 0.5)，過程如下：

經由測量每個聚類的平均離散度對每個聚類結果進行評估。隨著聚類數的增加，平均離散度總是下降。然而，它並不是均勻的降少。選擇點選擇所需的離散程度，用最小聚類數離散度到最大聚類數離散度的比例衡量。最終聚類結果就是所計算的聚類離散度與該值最相近的聚類結果。如果選擇點為 1.0，則總是挑選最大數目的聚類，如果選擇點為 0.0 則總是挑選最小的。

聚類的命名根據拆分過程中的不斷分離結合。最初的聚類 (在聚類數最小的基礎上進行) 用序數來命名，就好像單步 k - 平均值一樣。每次拆分某個聚類時，新的聚類名稱將與該聚類相同，但其名稱後面會個別加上「e」或「c」。例如，聚類 ?-B-A? 就是由初始聚類中導出而來的，而且拆分了兩次。

當建立聚類模型時，您可以在狀態視窗中顯示統計結果： iris 資料集中的樣例：

聚類結果：

```
-----  
記錄到重心的整體距離平方根： 0.216 +- 0.0928 RMS 從記錄到每個重心的距離：  
    聚類 1： 0.2306 +- 0.09484  
    聚類 2： 0.1921 +- 0.09932  
    聚類 3： 0.2247 +- 0.08058  
模型另存為 iris.cluster
```

資料是每個聚類離散度以及整個聚類最佳適合度的測量。根據不同資料集或不同的聚類數所聚類的離散度是不可比較的。

聚類選項

k- 平均值聚類算法根據記錄到聚類中心之間距離的計算。資料中的每一欄視為所有記錄的多維度空間中的獨立維度。預設情況下，資料中的每一欄對最後的距離有著同樣的影響。然而，您可以改變每一欄或屬性的影響力，方法是在「聚類進階選項」對話方塊中指定 *屬性權重*。

屬性權重

欄的屬性權重是一個大於或等於 0 的值，可指定在聚類算法的距離計算中確定欄的影響力。屬性權重設為 1 代表平均的影響力。設為 2 代表影響力為相同欄的兩倍。設為 0 會使聚類中的該欄不產生作用（就好像在進行聚類前將該欄從資料中移除一樣）。屬性權重不必為整數，而且可能小於 1。

按一下主聚類面板中 *進階選項* 或 *更多選項 (IRIX)* 按鈕可設定屬性權重。「進階選項」對話方塊的上部顯示目前屬性權重，所有值都預設為 1。

要改變一個或多個權重，選擇您期望改變權重的欄 (按住 **Shift** 並按一下可選擇或取消選擇範圍，按住 **Control** 並按一下可同時選擇或取消選擇多個值)。現在，在已選定的權重欄位中輸入新權重。

最後，按一下**設定**按鈕，和選定的權重就可以改變成新值。您可以利用**全部選定**按鈕同時選擇和設定所有的權重。

您需要不斷地調節屬性權重以發現更好的聚類結果。加權準則為：

- 含有大量值的字串或枚舉 (枚舉陣列) 類型欄應設低權重 (通常為 0)。如果使用該算法這些欄將導致聚類結果的嚴重偏差，他們以後將有助於解釋聚類結果。
- 如果發現有幾欄具有較強的相關性，調節權重使這些欄的「權重總和」為 1，否則，這些欄會過度地驅動聚類結果。
- 既然類目型 (字串或枚舉) 欄之間的距離趨向於更大，通常會將其權重設得比實值欄要低。

聚類選項對話方塊

您可以從「工具管理員」上的「挖掘工具」索引標籤中進入「聚類選項」對話方塊。選擇「聚類」索引標籤，然後選擇**進階選項**按鈕來得到「聚類選項」對話方塊。所列選項取決於您利用的是單一還是重複 k- 平均值方法。

- *屬性權重*

聚類提供對資料集中的每個屬性值進行不同方式的加權機會。請參閱第 200 頁的「**加權**」可得到關於該選項的更多資訊。

- *距離度量*

這種度量決定記錄與聚類中心之間的距離應該測量。預設的選擇是歐氏距離，該距離為多維度空間中沿直線的長度測量，在資料的每欄中為一維。跳現式功能表提供可替換的選項：曼哈頓距離。將每維度的坐標軸的距離相加以計算曼哈頓距離。這種度量方法因曼哈頓中橫向的街道街區而得名：不可能從點 A 直接到點 B，因為只能沿著平行於坐標軸 (街道) 的路來行進。

- *最大# 重複次數*

這個選項設定經過聚類算法所使用資料集數目的限制。更特別的是，它在單步 k- 平均值算法中限制步驟 4 執行的次數。請參閱第 41 頁的「用單步 k- 平均值方法聚類。」。

- *隨機子*

不同的隨機子將導致初始聚類中心的不同起始點。請參閱第 41 頁的「用單步 k- 平均值方法聚類。」。

- *使用權重*

與 MineSet 中大多數的挖掘工具一樣，聚類支援記錄加權。這一選項允許您指定一欄 (必須為數字型)，該欄將指定資料中每個記錄的權重。

- *權重保留為屬性*

如果核取該方塊，則指定的加權欄也會由聚類算法用為正常的屬性。如果未選擇該方塊，則指定的加權欄將不會用於屬性；它會從屬性加權部分中消失 (固有值為 0)。

聚類可視化工具

「聚類可視化工具」展示一系列方塊圖和柱狀圖。一旦執行聚類操作，便可以利用「聚類可視化工具」直接查看分類中心。

選擇性地，您可以在訓練資料上直接利用應用模型特徵 (請參閱第 15 頁的「應用模型」) 為每個記錄分配類。您可以利用 MineSet 中的許多其它工具來探索聚類結果。

檔案需求

「聚類可視化工具」需要以下檔案：

- 資料檔案由索引標籤分隔的欄位的列所組成。利用「工具管理員」可以很容易的建立這個檔案。您可以通過從資料來源（例如資料庫）中抽取資料來產生資料檔案，然後將其格式化以便於聚類可視化工具的使用。資料檔案有使用者定義的附檔名（由聚類可視化工具提供的樣例檔案具有 `.clusterviz.data` 附檔名）。
- 配置檔案描述輸入檔案的格式及如何進行顯示。工具管理員可以建立這個檔案，或者您自己可以利用任何編輯器（例如 `jot`，`vi`，`Emacs`，或您喜歡的文字編輯器）來產生這個檔案。

配置檔案必須有 `.clusterviz` 附檔名。當啟動「聚類可視化工具」或開啓檔案時，您必須指定配置檔案而不是資料檔案。

啟動聚類可視化工具

有幾種方法啟動「聚類可視化工具」：

- 利用「工具管理員」來配置並啟動「聚類可視化工具」。（請參閱第 175 頁的「工具管理員」瀏覽「工具管理員」功能，對所有 MineSet 工具都通用；請參閱《*MineSet 3.0 for Windows 企業版使用者指南*》，取得在「工具管理員」中利用聚類可視化工具的範例。）
- 從「工具管理員」中的「可視化工具」下拉式功能表中選擇「聚類可視化工具」。選擇「檔案」>「開啓」來開啓配置檔案。
- 如果您知道要使用的配置檔案，連按兩下該配置檔案的圖示。這樣就啟動「聚類可視化工具」並自動載入指定的配置檔案。只有當配置檔案以 `.clusterviz` 結尾時才起作用（對於利用「工具管理員」為「聚類可視化工具」建立的配置檔案來說都是這樣）。
- 從 UNIX 指令行中輸入：

```
clusterviz [configFile]
```

`configFile` 是可選的，指定所使用的配置檔案的名稱。如果未指定配置檔案，您必須用「檔案」>「開啓」來指定一個。

顏色選擇

選項對話方塊中的許多設定都有選擇顏色的選項。當使用「工具管理員」時，按一下 *進階選項* 按鈕 (Windows) 或 *更多選項* (IRIX) 按鈕來得到工具選項對話方塊。

用顏色選擇器來選擇顏色 (Windows)

要得到「顏色選擇器」，按一下顏色開關或「工具選項」中顏色清單裡的 + 符號或者「進階選項」面板。如果選取顏色開關清單，該選項清單會出現在 [圖 1-10](#) (最初是空的)。



圖 1-10 顏色開關列表

當按一下新顏色開關時，您將看到顏色反映在「顏色選擇器」的近期和預覽區域內 ([圖 1-11](#))。如果您改變主意，回到「最近」柵格中先前的顏色開關。一旦您選擇好顏色，按一下 *確定*，然後該顏色就會加到您的顏色清單中。不用取消「顏色選擇器」，您就可以加入幾種顏色。

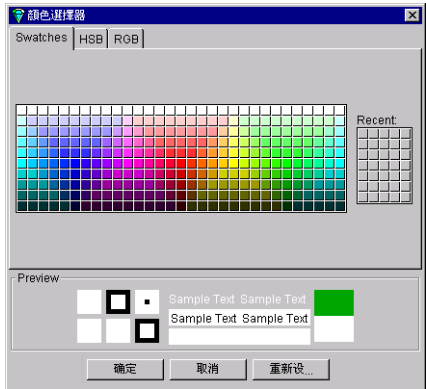


圖 1-11 顏色選擇對話方塊

您可以在顏色選擇器窗格中用 **HSB** (色度、飽和度和亮度)，或以 **RGB** (紅、綠和藍) 方式來選擇顏色。在 **HSB** 窗格中，按一下大的顏色方塊。白色的圓圈將會出現。

當選中 **H** 單選鈕時，您可以拖曳該方形區域中的圓圈來調節飽和度和亮度。要調節色度，移動彩色列旁的滑動桿。

當選取 **S** 單選鈕時，您可以在方形區域中拖曳圓圈來調節色度和明亮度。要調節飽和度，移動彩色列旁的滑動桿。

當選擇 **B** 單選鈕時，您可以在方形區域中拖曳圓圈來調節色度和飽和度。要調節明亮度，移動彩色列旁的滑動桿。

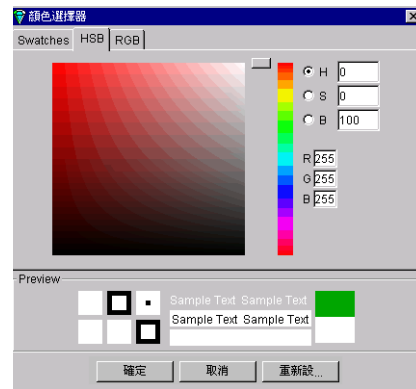


圖 1-12 「顏色選擇器」對話方塊的 HSB 窗格

在 **RGB** 窗格中，您可以利用滑動桿來指定紅、綠和藍色值的量以選擇顏色。

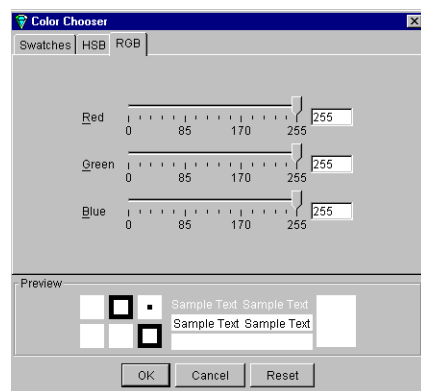


圖 1-13 「顏色選擇」對話方塊中的 RGB 窗格

用「顏色瀏覽器」來選擇顏色 (IRIX)

如果「顏色瀏覽器」是可用的，它將會出現在工具選項面板上。在 MineSet 中使用顏色開關，可以得到一個顏色清單選擇器。該部分描述如何為 MineSet 可視化工具選擇、應用和改變顏色選項。

如果只選擇一個顏色 (例如一個柵格顏色)，則會出現單一顏色開關。

按一下開關可產生顏色瀏覽器，讓您改變開關的顏色 (圖 1-15)。

如果選擇顏色開關清單，就會出現一欄開關 (它們最初是空的)，顯示如圖 1-14。



圖 1-14 顏色開關列表

要編輯顏色，用滑鼠左鍵按一下開關。也同樣選擇了用按鈕改變顏色的開關。如果用滑鼠中鍵按一下開關，便選取該開關，但是不會出現顏色選擇器。

在開關清單旁是四個按鈕。第一個是標有加號 (+) 的按鈕，它可以在清單的尾端加入新的顏色。增加開關後，顏色選擇器便出現，您可以選擇該開關的顏色。如果清單中已有最大數目的顏色，則「增加」按鈕停用。

「下一個」按鈕標有減號 (-)。該按鈕刪除選取的顏色。如果沒有選擇開關，或者清單中已有最小數目的顏色，則該按鈕無法使用。

在該「刪除」按鈕旁是另外兩個按鈕（向左和向右切換所選的顏色）。如果沒有選擇開關或這些開關已經在清單的底部，這些按鈕將停用。

如果顏色數量超過顯示面積所能容納的範圍，則捲軸箭頭就會增加至每個清單的底部。如果硬體用完了顏色，顏色開關會以本文標籤取代，將顏色以十六進位計數來顯示。

在可視化「配置選項」面板中的「顏色」面板上，當您按一下顏色開關或增加按鈕時，「顏色瀏覽器」（圖 1-15）就會出現。

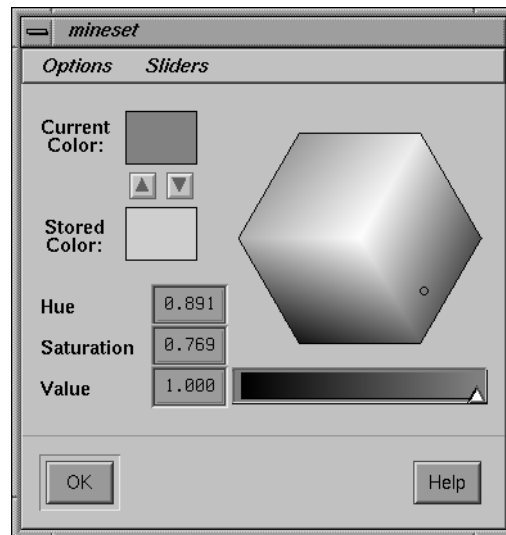


圖 1-15 顏色瀏覽器 (IRIX)

用「顏色瀏覽器」選擇顏色：

1. 有色六邊形中，您可以在小的白色圓圈上面移動滑鼠游標。
2. 按滑鼠左鍵按鈕，然後在六邊形的周圍移動滑鼠。小圓圈下面的顏色就會出現在 *目前* 顏色標籤旁的矩形中。當您選擇顏色時，這個矩形可作為調色板。
3. 當小圓圈在您想要顏色的上面時，鬆開滑鼠按鈕。所選開關立即換上所選的顏色。

您可在不取消「顏色瀏覽器」的情況下編輯幾個顏色；按一下選項面板中的任意顏色開關就可在已經出現的「顏色瀏覽器」中編輯該顏色。

當您對顏色作出選擇後，按一下 *確定* 按鈕。「顏色瀏覽器」視窗關閉。

欄重要性索引標籤

透過工具管理員的欄重要性索引標籤可以使用「欄重要性」工具。請參閱下面的「欄重要性」。

欄重要性

在「工具管理員」的「挖掘工具」面板上 *欄重要性* 可從標有「欄重要性」的索引標籤上執行。它有助於發現資料集中最重要的欄，並在此基礎上對所選的標籤欄可以區別不同值的差異。

欄重要性和其它資料挖掘工具，例如，聚類（在 [第 40 頁的「聚類」](#) 中討論）之間的差異是：利用欄重要性可以決定使用哪個標籤來確定欄重要性。在聚類中，資料本身會顯示識別因素。因為分類模型表示方法間的差異，不同屬性對於不同模型可能就顯得更為重要。MineSet 提供的樣例檔案，說明某種情況在欄重要性中非常有用，請參閱 [附錄 A，「配置和資料檔案樣例」](#)。

尋找重要欄

可以使用 *欄重要性*，假設您想要尋找能夠偵測 *適當信用風險* 的標籤，便能選擇它們並對照到「分散可視化工具」的坐標軸上。當您選擇標籤並按一下 *開始*，就會出現一個跳現式視窗，視窗中有三欄，這三欄就是最好的鑑別工具。一種稱為「純度」的測量（從 0 到 100）會告訴您欄區分不同標籤的優劣程度。增加更多的欄只能夠增加純度。

純度可作為標籤值分布斜率的一種度量。累計純度測量是拆分資料的純度的度量。利用欄來拆分資料與決策樹中的資料拆分同等重要。拆分中的每一設定都有其本身的純度測量，且拆分過程中的純度測量是這些單獨測量的結合。對於已知的拆分設定，如果每個分類具有同等代表性，其純度為 0；如果每個記錄為同一分類，則純度為 100。同樣地，如果拆分中的每個設定具有分類同等代表性，則累計純度為 0；如果每個拆分設定中包含的記錄都是相同的分類，則累計純度為 100。

「欄重要性」有兩種模式：

- 簡易模式

要啟動簡易模式，從跳現式功能表中選擇離散標籤，然後指定您要看到的欄號，然後按一下 *開始*。

- 進階模式

利用進階模式可以控制欄的選擇。要進入進階模式，按一下「欄重要性」面板中的 *進階模式*。出現一個對話方塊，您可以選擇加權屬性並決定在確定重要性時它是否表現為一般的屬性。對話方塊包含兩個欄名清單：左邊的清單包括可用的屬性，而右邊的清單包含選為重要的屬性（可經由人動或欄重要性算法來決定）。

進階模式可以以兩種不同的方式運作：尋找幾個新的重要屬性或者對可用屬性進行分級。

— 尋找幾個重要屬性

要進入這個子模式，按一下對話方塊中的前兩個單選鈕 (... *尋找 [數字] 個附加重要屬性*)。如果在沒有進一步變更的情況下按一下*開始*，其效果與「簡易」模式裡一樣，尋找重要欄的指定號碼並自動將其移到右邊的欄中。每個欄的旁邊，都指定了累計純度 (也就是說，向上所有欄的累計純度)。

選擇性地，將欄名從左清單移到右清單，可以事先指定想要包含以及讓系統增加更多的欄。例如，要選擇*柱面*欄並讓系統找到超過三個的欄，按一下*柱面*欄名，然後按一下清單之間的向右箭頭。

按一下*開始*可以看到每個欄的累計純度，還有清單先前的內容。純度為 100 意味著，使用指定的欄，您可以在資料集中正確地區別不同的標籤值。

— 對可用的屬性分級

利用進階模式您可以計算每個欄純度的變化，而這些欄要加入到已標為重要的欄中。例如，您可以將*柱面*移到右邊的清單中，然後請求系統計算左清單中剩下的每欄所產生的遞增改進純度。對於右邊（已經標為重要）的欄計算累計純度。

要進入這種子模式，按一下對話方塊底部兩個單選鈕的第二個 (... *對左欄各欄計算改良純度，對右欄各欄計算累計純度*)。該子模式允許對過程進行優化控制。如果兩個級別很接近，但您可能比較喜歡其中一個 (例如，因為它收集成本低、更可靠或更容易理解)。

欄的重要性取決於該欄先前是否標為重要。例如，當單獨的看*網*收入可能是一條好欄，但它可能沒有*薪資*重要，因為它們可能是非常有關係的。三個最好的欄並不一定由那些單獨級別很高的欄組成。如果兩欄分別以美元或另外一種貨幣形式來表示收入，它們的級別是一樣的；但是，一旦其中之一被選中，另外一個不能再提高區分能力。

對於為「分散可視化工具」和「平板可視化工具」尋找三個最佳坐標軸的過程來說，欄選擇是很有用的。將標籤選為在「樹狀可視化工具」中使用的關鍵字時，在為「樹狀可視化工具」尋找好的區別等級（即分開不同標籤值的等級）時它是很有用的。

利用自動離散化過程將所有浮點型的值（**雙精度** 或 **浮點型**）預先離散化。如果左邊清單中的欄沒有指定的值，算法將不予考慮；這是因為它具有單一值（例如，當它在離散為單一區間時），或因為它拆分的記錄數並不具有統計意義。

分類器中欄重要性的不同

這一部分描述「欄重要性」的差異，由「證據」和「決策表導入工具」所進行的重要性分級，以及利用「決策樹」來進行拆分。因為「欄重要性」使用全部資料，因此以下過程假定您在「僅用於分類器」模式中執行了導入工具，從而導入工具也就使用了全部資料。

離散過程

「欄重要性」算法和「證據導入工具」利用自動離散化算法（同樣的算法已應用於「工具管理員」中的自動分組）將所有連續屬性（欄）離散化。「決策樹算法」事先並不提供屬性離散化，而是在建成樹時尋找臨界值。

自動離散的優點是它將連續的範圍同時離散成幾個間隔，而「決策樹」只能進行二元拆分。

「決策樹」算法的主要優點在於將子集提煉離散為資料塊（當進行檢驗時，符合特殊的節點所形成）。因此，與「全局」離散化相對，該離散化過程對於那些記錄來說是「局部」的過程。

重要性函數

與排斥多重拆分的「決策樹」、「選項樹」和「決策表導入工具」形成對照，「證據導入工具」和「欄重要性」算法根據作為純度度量的「公共資訊」來對屬性進行分級。因此如果您允許「決策樹」找出重要欄，則結果將傾向於具有較少值的屬性而不是具有較多值的屬性。「決策樹」在預設情況下將變為「公共資訊」。

依照其它屬性

證據導入工具獨立為每個屬性分級。如果幾個屬性高度相關，則它們有相同的級別。如果在「欄重要性」中使用「進階選項」，並且在沒有任何屬性選為重要的情況下「... 計算改良純度」（也就是說，移到右邊的清單中），所顯示的屬性分級與由證據導入工具選擇的分類次序相符。

「欄重要性」算法、「決策樹導入工具」及「決策表導入工具」都提供比「證據導入工具」更有力的重要性功能。選擇重要級別要考慮到其它的欄。

在欄重要性中，將欄視為與右邊清單中的該組欄有重大關係。如果兩個欄高度相關，並且其中一欄已被選取，另外一欄可能永遠不會被選取；欄之所以被選取是因為可以提供關於標籤的更多資訊。

在「決策表導入工具」中，「建議」按鈕提供與「欄重要性」算法一樣的重要性工具。欄決定是否重要與對照方塊中已經存在的欄有關。但是，在「決策表導入工具」建議模式和「欄重要性」算法之間主要有三個不同。第一，「決策表導入工具」排斥多重拆分。第二，「決策表導入工具」能夠進行徹底的搜尋以找到更好的欄順序。最後，「決策表導入工具」並不會列出任何純度，而是僅用來將欄分級。

「決策樹導入工具」提供更靈活的重要性分級方法，因為不同的欄可以在不同的子樹上選取。例如，您可以選擇一欄作為根的左子節點，另一欄作為根的右子節點。當這些適用於「決策樹」時，對於選擇小部分欄來顯示「分散可視化工具」或「平板可視化工具」的情況則是不合適的。對於這些情況，「欄重要性」算法是比較好的，因為它建立了易忘記的「決策樹」，在其中任何層次的樹在每個節點都檢驗相同的列。使用「欄重要性」，必須為先前所選的欄組合選擇一個欄。

欄

請參閱[第 1 頁的「增加欄」](#)，[第 143 頁的「移除欄」](#)和 [第 154 頁的「排序欄名稱」](#)。

指令行操作

每個可視化工具以及 MineSet 本身可以由指令行中啟動，方法是在提示符下輸入工具名稱，例如：

Windows 系統類型：

```
Viz [configFile]
```

Irix 系統類型 (例如)：

```
scatterviz [configFile]
```

識別配置檔案 (*configFile*) 是可選的，但是如果不指定，工具啟動時只有「檔案」和「說明」功能表可操作。您必須利用「檔案」> 開啓來指定配置檔案。

在 IRIX 中，多種工具要使用指令：clusterviz，eviviz，scatterviz，splatviz，statviz，treeviz 和 mapviz. 關聯規則使用[第 20 頁的「關聯規則」](#)中的兩部份指令。

配置檔案

對於每個工具，MineSet 需要兩個檔案：

- 資料檔案由索引標籤分隔的欄位列組成，並具有 *.data* 附檔名。
- 配置檔案描述輸入資料的格式以及如何顯示它們。檔案將工具名稱的縮寫作為附檔名，例如，*eviviz*、*scatterviz*。

在利用「工具管理員」來指定那些您計劃使用的分類器或可視化工具的選項或參數時，您可以自動建立該配置檔案。選擇性地，您可以使用 `ascii` 文字編輯器為每個工具建立配置檔案。例如，開啓 Word Pad 或類似的文字編輯器並選擇「檔案類型：所有文件 (*.*)」。您可以用此種方式開啓任何 *.schema* 和 *.data* 檔案。在 *《MineSet 3.0 企業版介面指南》* 可找到更多的細節和範例。

混淆矩陣

混淆矩陣可以更加仔細地以圖描述由分類器產生的誤差。混淆矩陣顯示所產生的誤差類型，而不是只簡單的分析正確或錯誤預測的數目。在「工具管理員」的「挖掘工具」索引標籤環境下，對於所有的分類器，「顯示混淆矩陣」是一個「進階」選項（選擇「分類」索引標籤）。圖 1-16 顯示在 *iris* 資料集上導入的「決策樹」的混淆矩陣。

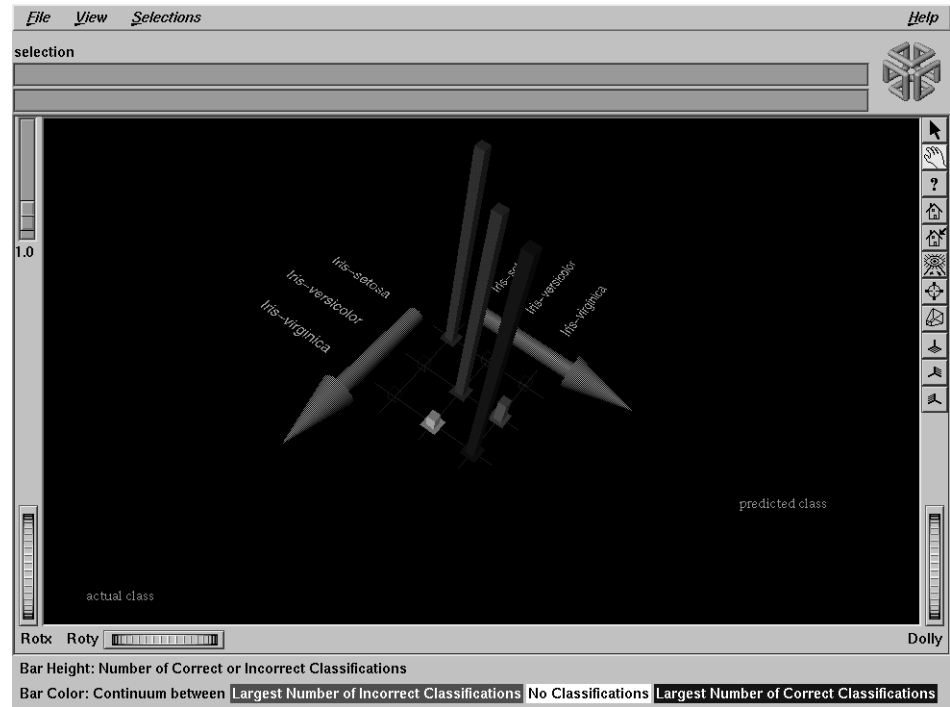


圖 1-16 Iris 資料集的混淆矩陣

兩個坐標軸代表：

- 由分類器預計的分類值，以及
- 在檢驗集中指定的實際分類值（預留設定）。

對角線上的項目是正確的預計。非對角線項目指示不正確的預計。這種代表方式顯示 多色蝴蝶花和純白色蝴蝶花經常會混淆，然而藍色蝴蝶花的預計總是正確的。

當產生不同錯誤類型的成本不平均時，使用損失矩陣而產生的錯誤通常是很有用的（請參閱第 111 頁的「損失矩陣」）。

注意：混淆矩陣顯示測試集產生的誤差；因此，如果資料的基本分布並沒有明顯地改變，就表示在實際情況下誤差期望的真實分布。MineSet 中的混淆矩陣在修正之前就進行計算，並且在是否應用修正的情況下都是一樣的。（也可參閱 [第 33 頁](#) 的「修正」。）

成本複雜性

成本複雜性是在 CART (分類和回歸樹) 中發展的進階修改方法選項。要尋找修改方法，從工具管理員中選擇分類挖掘工具索引標籤並選擇決策樹。按一下 [進階選項](#) 按鈕可得到顯示「修改選項」的對話方塊。（您也可以選擇「回歸挖掘工具」索引標籤並按照同樣的方法操作。）

成本複雜性的修改經由互換樹的誤差率（其成本）和樹上葉子的數目（其複雜性）試圖產生形狀最佳化的樹。在成本複雜度修改過程中，訓練集拆分為學習集和修改集。學習集用來成長一個修改樹。該樹修改後產生一系列具有遞減複雜性的樹。然後使用修改集在序列中找出成本最小的樹。標出最小成本樹的大小。學習和修改集合併並長成一棵樹。然後該樹修改為最小成本樹的大小。

成本複雜度修改參數允許您選擇比最小成本樹更小的樹。該參數指出標準差的值比可以接受的最小成本樹所花費的成本更大。將該參數設為 0 以選擇最小成本樹；設定參數為 0.5 以選擇尺寸最小的樹，其誤差率不大於 0.5，標準差比最小成本樹差。越高數目表示越多的修改。如果資料包含噪音（誤差和異常），降低值來建立更小的樹。如果樹修改成為一個節點，降低值以降低修改量並顯示更多樹的結構。

修改過程比限制樹高度或增加拆分下限過程的速度要慢，這是因為要建立整個樹然後再進行修改。然而，修改過程是有選擇地進行著，所以具有較低的誤差率。

交叉驗證

交叉驗證是估計分類器誤差的一種方法，其過程為將資料集拆分成一定數量 (k) 的子集 (通常為 10)，然後建立同樣數量的分類器。該過程可以重複多次來增加估計的可信度。導入工具訓練和檢驗了 k 次；每次都使用所有的資料減去一個不同的子集進行訓練，然後在預留的子集上進行檢驗。

「交叉驗證」是任何導入工具都可使用「估計誤差選項」。對話方塊允許您在交叉驗證中設定子集數以及過程的重複次數。您也可以將隨機子設定為不同的數或保持相同的數，以確保可以從相同點切入資料。

資料整理

資料很可能表現為含糊的形式，或與其它存在的資料不相容。欄位很可能遺漏或已無效。從這樣的複雜性中清除為一種有序狀況的過程稱為資料整理，並且在很多情況下是資料挖掘操作的預先處理過程。協力廠商提供工具可以支援該過程。

將資料從其它格式轉化為 MineSet 格式，然後做一些基本的資料整理操作，從「工具管理員」功能表列中選擇「檔案」>「匯入資料」。請參閱《MineSet 3.0 企業版介面指南》中「匯入資料檔案」部分。

資料目標窗格

「工具管理員」中的「資料目標」窗格提供這些主要選項：「可視化工具」、「挖掘工具」和「資料檔案」都位於索引標籤的頂部。「可視化工具」提供現有資料的直接可視化以顯示可提供直觀的相關性。「挖掘工具」建立資料模型 (伴隨可視化) 可用於預測資料特徵。「資料檔案」用於將操作資料儲存到檔案。

資料檔案索引標籤

在「工具管理員」的「資料目標」窗格中利用「資料檔案」索引標籤可以將操作資料儲存為將來可用於客戶機和伺服器上的資料檔案。如果您按一下「資料檔案」索引標籤，就會出現帶有兩個切換按鈕的面板，用它可以指定檔案儲存在伺服器上或客戶端上。選擇的客戶端檔名出現在「客戶端」核取方塊旁。如果選擇「客戶端」，*選擇新客戶端檔案*按鈕將帶出一個對話方塊，選定要儲存的檔案名稱。當您選擇「伺服器」，您可以將伺服器的檔名直接輸入鄰近「伺服器」核取方塊的文字欄位。

注意：對於伺服器檔案，不允許有路徑名；所有的伺服器檔案儲存在「資料移動」快取目錄中。

資料導入

MineSet 提供資料匯入功能，可從「工具管理員檔案」功能表中找到（「匯入資料」）。「導入資料」面板允許您選擇想要導入的檔案以及一些導入選項。請參閱 *《MineSet 3.0 企業版介面指南》* 來了解這些選項的說明及所支援的資料格式清單。

資料轉換面板

利用主要的「工具管理員」視窗左邊的「資料轉換」面板，您可以對要運作的資料表格進行操作。在您利用「檔案」功能表打開一個表格後，表格中的欄標題會出現在「資料轉換」面板的*目前欄*視窗中。顯示選項的功能為：

- *移除欄*可允許您刪除與目前可視化處理或挖掘過程無關的一個或多個欄。
- *欄分組*可讓您將每個記錄分配給一定欄值範圍內的群組。例如，年齡欄可以分為不同的範圍：(0...18]，(19...25]，(26...35]，等等「(」表示下限沒有包含在範圍內。「]」指出上限包含在區間範圍內（也可見「分組」）。

- **組合**—建立新欄，它代表所選欄的求和、平均、計數、最大或最小值。組合也能夠從以其它欄分組為索引的欄中產生陣列。應用組合操作產生的結果表格通常有很多欄比原始表格少很多行，因為新表中的每一列是原始表中幾個列的組合（請參閱 [第 4 頁的「組合」](#)）。
- **篩選**可讓您根據包含欄值的表達式來選擇資料的子集，例如，留下那些年齡小於 20 的記錄（請參閱「篩選」）。
- **更改名稱 / 類型**可讓您改變欄的名稱及類型（請參閱「更改名稱 / 類型」）。
- **增加欄**可讓您根據數學表達式來增加新欄（請參閱「增加欄」）。例如，使用表達式在「年齡」欄上增加「欄輔助值」：如果年齡小於或等於 18，則輔助值為 1；否則輔助值為 0。
- **應用模型**可讓您利用以前建立的分類器來標示新記錄、來估計標籤值的機率、在新資料上測試分類器或利用存在的分類器修正資料（請參閱「應用模型」）。
- **採樣**允許您選擇資料的隨機子集。對於很大的資料集這是有用的（請參閱採樣）。
- **按順序顯示各欄**允許您按照字母順序對欄進行排序。在 Windows 系統中，該功能顯示為核取方塊。
- **插件操作**只有當插件 API 可用時才出現。這些功能讓用戶可以像執行 MineSet 操作一樣來執行插件操作。請參閱 MineSet 首頁 (<http://mineset.sgi.com>) 及《MineSet 3.0 企業版介面指南》中的「MineSet 插件功能」，可以找到更多的資訊。

決策表

與決策樹相似，決策表是結構具有等級之分的表，但卻是使用一對屬性在每一層將資料分解而不僅是使用單一屬性。「決策表導入工具」為資料分類指定最重要的屬性（欄），然後可視化工具將其顯示為塊狀的套合圖。可視化結果中的每個塊狀圖依次分為更小的塊狀圖以代表下一個最重要的屬性。每個可視化過程都可以包含幾個代表重要程度遞減的屬性的層。請參閱 *《MineSet 3.0 for Windows 企業版使用者指南》* 獲得有關使用決策表可視化工具的資訊。

決策表使用的分類方法與決策樹的一樣。選取樣例所在區域的大多數分類來執行分類操作。如果已分類的記錄落入沒有訓練資料出現的區域，分類過程可以經由選擇表格等級中上一層的優勢分類來執行。

導入決策表

「決策表」分類器可以從資料中導入或自動產生。資料由記錄以及與每個記錄都相關的標籤所組成。（請參閱第 99 頁的「導入工具」。）

「決策表」分類器的自動導入是一個過程，使用其記錄計數（或更平常的情況下，記錄權重）計算表格等級中每個節點上的機率。在可視化過程中，每個節點上記錄的分布用一個塊狀圖來表示。

所有連續屬性被分為離散的組，分組原則是讓範圍內的類別分布儘可能的不同。範圍的數目可以自動確定。利用「工具管理員」進行明確的分組，您可以更改任何屬性的自動分組。

沿著坐標軸（屬性）的任意一列中塊狀圖的數目顯示由導入工具產生的離散範圍數目。如果只有一個範圍，就意味著該屬性本身在預計標籤時不起作用。初始時，標籤的先驗機率在「標籤機率」窗格中顯示。先驗機率是訓練集中每個分類的一部分。

在決策表等級中有三種方法可以將屬性對照到每一層的 X 和 Y 軸：手動、自動以及用功能搜尋的方式產生。

啟動決策表可視化工具

有幾種方法可以啟動「決策表可視化」工具：

1. 在「工具管理員」中的「分類」索引標籤執行「決策表導入工具」。在導入工具建立分類器後，會自動啟動「決策表可視化工具」。
2. 利用「工具管理員」從「可視化工具」按鈕中啟動 3D 可視化工具，然後開啓 `.dtbleviz` 檔案 (相關資訊請參閱第 175 頁的「工具管理員」中的「工具管理員」功能，在 MineSet 工具中是通用的)。
3. 如果您知道要利用的配置檔案，按一下該配置檔案的圖示。這樣就啟動了 3D 可視化工具並自動載入指定的配置檔案。只有當配置檔名以 `.dtbleviz` 結尾時才有效。(所有由「工具管理員」為「決策表可視化工具」建立的配置檔案都帶有 `.dtbleviz` 結尾。)
4. 在提示符下，輸入下列指令可以從 IRIX 中啟動「決策表可視化工具」：

```
dtbleviz [filename.dtbleviz]
```

其中檔案名是可選用的。如果您指定配置檔案，就必須使用「檔案」>「開啓」來選擇一個檔案。

離散標籤

「離散標籤」功能表提供欄的清單，該欄包含資料集中的離散值 (按一下 **離散標籤**：旁的箭頭)。離散屬性 (分組值、字元字串值或整數) 的值數有限。您應該以幾個值的欄來選擇標籤屬性，最好是兩到三個 (請參閱訓練集)。如果沒有離散屬性，功能表顯示沒有離散標籤，並且 **繼續按鈕** 也停用。然後您必須，利用「工具管理員」的「資料轉換」面板並經由分組或增加欄來建立一個離散屬性。

藉由「將欄對照到坐標軸」來「檢查資料」

您可以將欄對照到 X 和 Y 軸上來發現資料中屬性間的關係。在「工具管理員」視窗中「資料轉換」窗格的「目前欄」面板，選擇一個欄名，然後：

- 在 Windows 系統中，按一下 X 清單或 Y 清單中的儲存格，下拉式功能表會產生一個適當選擇的清單。
- 在 IRIX 系統中，在 X 坐標軸或 Y 坐標軸視窗中選擇一項。對照的頭兩欄顯示在最高層，接下來的對照按層次順序依次下降。

如果您對照奇數個屬性，最後一層會顯示與奇數屬性值相對應的單獨的一欄塊狀圖。如果兩個屬性之間的作用並不確定，則將一個屬性對照到 X 軸，而另外一個屬性對照到 Y 軸。

可視化過程可能會交換 X 和 Y 的對照，以求更好地保持整個塊圖方陣的長寬比，但是它不會將屬性移到到不同的層上除非屬性已經對照。

如果您並不知道將哪個屬性對照到哪個軸上，則按一下 **建議**和**繼續**按鈕。

「建議」有兩種不同方法：帶有或不帶有功能子集搜尋。在更多導入工具選項面板中：

- 如果建議使用特徵搜尋，而沒有選取核取方塊，則 MineSet 執行「欄重要性」來搜尋特徵。欄重要性在[第 52 頁的「欄重要性」](#)中會有進一步的說明。
- 如果建議使用特徵搜尋，而沒有選取核取方塊，MineSet 從「欄重要性」開始，然後從「欄重要性」建議的順序開始執行所有可能的順序。在每一階段都會執行錯誤估計，並且會檢查每個選項。可以預測地，這將是個冗長的過程，如果您不耐煩時可隨時按一下**停止**。功能搜尋持續的時間越長，分類器的結果越準確。

解釋決策表

每個分類標籤的先驗機率在「標籤機率窗格」中的餅圖都有說明，並出現在螢幕的右邊。分類標籤的先驗機率是在忽略所有屬性值的情況下，隨機抽取記錄時見到資料中標籤的機率。從數學上來說，該數為具有分類標籤的記錄除以記錄總數的比值。

左邊「主視窗」中每個矩形塊或塊狀圖的機率分布顯示在每個分類中，屬性值具有特別組合之記錄的比例。對於指定的資料這些機率是精確的。

預設情況下，沿坐標軸的標稱屬性值按照它們預測分類的優劣程度進行排序。這有助於找出重要屬性值。如果標籤是分組的屬性，就會使用代表最高值的類。如果標籤是標稱性的，則使用先驗機率餅圖中占有較大份額的分類以確定順序。當您選擇特殊的分類，並請求按照標籤機率進行排序（選擇「名稱排序」>「標籤機率」），該分類確定標稱值的順序。通常，標稱屬性值可以按照字母順序或權重順序來排序，分組屬性值總按它們的自然順序顯示。

當您細化下尋到一個矩形塊中（利用滑鼠右鍵），塊狀圖所代表的資料會重新顯示在新的塊狀圖方陣中，其中分配到明細層的屬性會用為坐標軸。塊狀圖在灰色基準塊的頂部。如果基準塊完全由塊狀圖覆蓋，這一層中的每個值組合由資料代表。如果基準塊完全由塊狀圖覆蓋，這一層中的每個值組合由資料代表。通常基準塊不會被覆蓋。顯示沒有資料存在的區域。在稀疏的資料集中，有大量的區域是空的。選取或選擇灰色基準塊與選擇塊狀圖的效果一樣，該塊狀圖在明細層的上一層。細化下尋到基準塊上使每個安置在其上的塊狀圖擴展到流程的下一層上。

如果屬性具有未知（空）值，則由問號（?）來注解未知值。如果存在空值並且沒有參與其他屬性值的排序，則空值顯示為第一個值。利用檢視 > 顯示空值可以切換「空值」的顯示。

「瀏覽」根據所執行的平台是 Windows 或 IRIX、以及所用的是兩鍵滑鼠或三鍵滑鼠而有所不同。在這裡說明了 Windows 平台下的兩鍵類型。您可以從桌面的「特性」面板中改變按鈕模式。

如果在左邊窗格中選取 (按一下滑鼠左鍵) 一塊圖，則右邊窗格中的機率餅圖會顯示機率分布，該分布正好與該塊圖顯示的相符。當選擇左邊的多個塊狀圖 (同時按下 **Ctrl** 與滑鼠左鍵)，則右邊的餅圖將顯示記錄集的機率分布 (與標籤有關)，該記錄集由所選餅圖定義。也可以選擇一整組餅圖，方法是選擇其下的灰色基準塊。這與在等級中的上一層選擇塊狀圖的效果相同。

分類會在右邊餅圖的下面列出，並按照每份的大小排序。擁有機率最大的分類在最上層。當選擇了左邊的值，這一順序改變為反映變化的機率餅圖。那些估計提供目前選擇的分類顯示在頂部。如果標籤是已分組的屬性，在份量大小方面分類的順序並沒有變化。除此之外，如果標籤是已分組的屬性，顏色就會按照連續譜系來分配：最高的組是紅色的；否則，就用隨機顏色來代表。

要得到更多資訊，將「標籤機率」窗格下面的「細節滑動桿」移到右邊。

決策表選項

選擇「進階選項」(Windows) 或「更多導入工具選項」(IRIX) 會出現「進階導入工具選項」對話方塊。該對話方塊由四個面板組成：

- 最上層的面板表示在「工具管理員」的「資料目標」面板中所做的選擇。
- 從上面數來第二個面板可讓您設定損失矩陣和加權屬性。請參閱第 111 頁的「損失矩陣」和第 200 頁的「加權」。
- 左下角的面板可讓您指定進一步的「導入工具」選項 (說明如下)。
- 右下角的面板可讓您指定「錯誤估計」選項 (除非在「資料目標」面板中選擇的是「僅用於分類器」模式，在這種情況下是空的)。面板中顯示的選項取決於您所選擇的「錯誤估計」類型 (請參閱第 18 頁的「應用模型」和 第 80 頁的「錯誤估計」)。

要微調「決策表」導入算法，您可以改變下列「決策表」導入工具選項。

- 最大尺寸

當擁有大型資料集時，這特別有用。預設情況下，限度是 10,000 個節點 (這些節點正好與可視化過程中的塊圖對應)。按一下核取方塊並輸入上限值以限制大小。限制數目可以加快導入過程並節省儲存空間。雖然限制規模會減少執行時間，但也會增加錯誤率。如果最大值小於顯示所有數值組合的必需值，可視化過程顯示的屬性數目可能比已對照的屬性數目少。

- 最大屬性數

該選項確定在搜尋時，可以允許多少不同的欄加入到「決策表」中。限制屬性會簡化結果並加快過程。這一限制僅影響由建議模式增加的欄。手動加入的欄並不受此限制的影響。

- 每組最小權重

「決策表」導入工具將所有的連續屬性離散化。該選項可使您定義每組中實例的最小數目。自動設定根據資料集的大小自動計算此一數目：資料集越大，組也就越大。如果資料集非常大，您可以得到更多想要的離散範圍。要減少組數，便增加該值。

細化下尋和概化上尋

在可視化工具中，您可以在「決策表」塊狀圖中細化下尋或概化上尋以檢視更多或更少的細節

(請參閱第 80 頁的「概化上尋和細化下尋」)。在其中一個塊狀圖上按一下滑鼠右鍵就可進行細化下尋。這樣將以更細的塊狀圖組來顯示原來的資料，這些資料按下一層的一組屬性進行分類。如果游標放在灰色基準塊上，按一下滑鼠右鍵細化下尋基準塊上所有的塊圖。如果游標放在背景中，按一下將細化下尋或概化上尋全部所有的塊圖。要進行概化上尋，在塊圖、基準塊或背景上按住 **Ctrl** 鍵並按一下滑鼠右鍵 (或使用滑鼠中鍵)。

下拉式功能表

您可使用五個下拉式功能表來存取「決策表可視化工具」的附加功能：「檔案」、「檢視」、「名稱排序」、「選項」以及「說明」。如果您沒有指定配置檔案就啟動「決策表可視化工具」，則只有「檔案」和「說明」功能表可用。請參閱「檔案」、「說明」及「選項」項目可以得到有關功能表的細節。

檢視功能表

「檢視」功能表可使您控制「主視窗」的顯示設定：

- 「篩選面板」所開啓的面板，可讓您根據所選的標準來篩選資料 (請參閱[第 95 頁的「篩選面板」](#))。
- 「設定背景顏色」可開啓「顏色選擇器」，並讓您選擇一個新的背景顏色 (請參閱[第 48 頁的「顏色選擇」](#))。
- 「視窗控件」可以隱藏或顯示主視窗周圍的外部控制。
- 「空位置」可切換空值所顯示的內容。如果存在空值，則顯示為第一個值，並與其它非空值略有偏移。
- 「使用場景檢視器」(或「利用檢查員檢視器」)可轉換成 3D 瀏覽的可選模式。既然滑鼠中鍵已用在「場景」檢視器中的瀏覽功能上，要使用細化下尋功能就必須在按一下滑鼠中鍵的同時按下 **Ctrl** 鍵。該選項不支援 **Windows**。
- 顯示為證據 (IRIX 上的「證據模式」) 顯示每個塊圖的條件機率，而不是最初根據記錄權重所顯示的分布。如果有一個或多個分類很小時，這是很有用的。
- 「工具列」和「狀態列」允許您顯示或隱藏「工具列」和「狀態列」。在 IRIX 上該選項不可用。

名稱排序功能表

「名稱排序」功能表可控制標稱屬性值如何排序，並提供這些選項：

- 按字母排序使具有標稱值的屬性從左到右（或從上到下）按字母順序進行排序。
- 「依據權重大小」使值從左到右進行排序，那些具有最大權重的記錄在左邊。
- 「依據標籤機率」（預設情況下）使標稱屬性值按照份量的大小進行排序，該份量對應其中一個分類。如果標籤是分組的屬性，則預設情況下會使用最高的組。如果標籤是標稱的，則使用先驗機率餅圖中占有最大份額的分類做為預設。如果選擇特殊分類，並且隨後就必須按照標籤機率進行排序，則選取的分類用於確定次序。在所有情況下，如果存在「空值」，它就為第一個值。

決策樹

「決策樹」是預測模型。使用獨立或已知的屬性值來進行預測，以幫助確定標籤值或未知屬性值。預計標稱值（通常為字串，例如「Yes」和「No」）或可僅可取較少值的任務是指分類過程。通過預計每個記錄的標籤，決策樹可以對資料進行分類。用於分類的結構是決策樹。一旦「決策樹導入工具」已對資料進行了分類，「3D 可視化工具」就會顯示它的結構。

「決策樹」方法顯示各種屬性間的相互作用，亦即屬性值的組合是如何影響預測標籤。預測標籤是指已知記錄中未知的特徵。在「決策樹」結構已知的條件下，資料在後序節點中的分布取決於前一節點所做的決策。

在完整的可視化過程中，決策樹中每個節點上的長條圖代表每個標籤值。您可以將游標放在長條圖上來顯示記錄數目（或權重）以及標籤值的百分比。在每個節點的基準塊處顯示了符合記錄的數目（或權重）。

建立決策樹

從資料中自動導入（或產生）「決策樹」分類器。登錄到伺服器並以平常的方式選取一個資料集後便可開始。

從「工具管理員」中選擇分類索引標籤，並且從導入工具彈出式功能表中選擇「決策樹」。除非您希望，否則您不必做更多的限定。只需按一下繼續按鈕。就可利用「樹狀可視化工具」來進行顯示生成的決策樹。在可視化過程結果中：

- 決策節點上的標籤指定在該節點上檢驗的屬性。
- 決策樹中的葉子節點指定一個分類。
- 基準的顏色表示子樹的錯誤估計。
- 每個節點上的垂直條顯示節點上分類的分布。

指到節點會顯示以下資訊：

- *決策樹權重*所指節點下之子樹中訓練集記錄的權重。該值對照為基準塊的高度。
- *檢驗集誤差 / 損失*子樹誤差的估計（或如果指定損失矩陣，則為損失）。+/- 後面的數是估計的標準差。標準差越高，錯誤估計越不準確。誤差 / 損失估計和標準差在葉節點只包含記錄或當檢驗集誤差接近 0% 或 100% 時可信度較低。
- *檢驗集權重*符合該節點的試驗集中的記錄權重（如果沒有設定權重則為記錄的數目）。
- *純度*是介於 0 到 100 之間的數，指出分配在節點上的標籤值的不均衡度。如果節點具有單一分類中的記錄，則純度為 100。如果標籤值具有相同的權重，純度為 0。純度是在修正之後進行計算的。

IRIX 中的並行過程

如果您已經安裝 MineSet 的多處理器版本，當分支包含超過 1000 條以上的記錄時，在「決策樹」中可以並行處理基於樹的算法。(請參閱第 134 頁的「在 IRIX 系統中的「並行計算」」。)您可以改變「工具管理員」中「特性」面板裡並行線程的數目來控制並行模式(請參閱第 92 頁的「檔案功能表」)。只有在 IRIX 系統中該選項才可用。

更多導入工具選項

當修正開啓並且將訓練集「顯示」為圓盤的選項選取時，垂直長條圖顯示以圓盤形式出現的訓練集的分布。圓盤的高度與長條圖的高度處在同一比例範圍內。您應該期望圓盤在長條圖上表現出與更多導入工具中用於「支持」比率的值一樣的高度。

決策樹選項

從「工具管理員決策樹」窗格中，選擇**進階選項**(或**更多選項**IRIX)彈出「分類器進階選項」對話方塊。該對話方塊由四個面板組成：

- 最上層的面板表示在「工具管理員」的「資料目標」面板中所做的選擇。
- 從上面數來第二個面板可讓您設定損失矩陣和加權屬性。請參閱第 111 頁的「損失矩陣」和第 200 頁的「加權」。
- 用左下邊的面板可以指定進一步的「導入工具」選項。
- 右下角的面板可讓您指定「錯誤估計」選項(除非在「資料目標」面板中選擇的是「僅用於分類器」模式，在這種情況下區域內是空的)。該面板中顯示的選項取決於您所選擇的「錯誤估計」類型(請參閱第 18 頁的「應用模型」)。

要微調「決策樹」導入算法，您可以改變一些「決策樹」導入工具選項。「分類器」對話方塊的「導入工具選項」部分提供以下功能：

- 經由 ... 限制樹高度

預設情況下，在「決策樹」中高度（層的數目）並無限制。按一下核取方塊並輸入上限值以限制高度。限制層的數目可以加快導入過程，改善並行過程的負載平衡度，並且不需要分解太多的節點可以研究「決策樹」。限制大小會增加誤差率，設定該選項並不影響最大層之前各層屬性的選擇和拆分過程。

- 拆分標準

這個選項提供五個拆分標準選擇。下面是技術性定義。對於一個指定的問題，很難說哪種標準是最好的。全部進行試驗，並選擇產生最低錯誤估計或形成最容易理解的「決策樹」。

公共資訊是指父節點和子節點純度加權平均之間純度的變化（即為 Δ ）。加權平均是根據每個子節點上記錄的數目來進行的。

標準化公共資訊（預設情況下）是公共資訊除以子節點數目的對數（基準塊 2）。

增量比是指在忽略標籤值的條件下，公共資訊除以拆分的 Δ 。「標準化公共資訊」和「增量比」將優先權給帶有少數值的屬性。

十番應用十番統計獨立性檢驗所有後選拆分。然後選擇導致標籤值最小獨立分類的拆分。

Gini 是 CART（「分類和回歸樹」）中使用的拆分標準。與「公共資訊」相似，Gini 測量父節點和子節點純度加權平均之間的純度變化。與「公共資訊」不同的是，Gini 計算節點純度，該純度減去節點處標籤機率平方根的和。

- 拆分下限

這是一個權重的下限 (或者，如果權重沒有設定，就為記錄的數目)，且至少必須存在於兩個子節點中。該選項的預設值為 2。例如，如果節點中有一個三向拆分，三個子節點中至少有兩個其權重必須為 2 或更大 (如果沒有設定權重則為兩筆記錄或更多)。這就提供另外一種限制「決策樹」大小以及提高執行速度的方法。

提高拆分下限是為增加機率估計的可靠度，因為每個葉子的記錄數較大。如果您期望資料包含大量的噪音 (誤差或異常)，或假如您利用樹來估計機率 (請參閱 **Heading2: 應用模型**)，將拆分下限增至 5 或更多。如果資料集很小 (< 100 筆記錄)，您可以減小該數到 1。

- 修改

「決策樹」是根據「樹高度」限制和「拆分」標準的限制而建立。統計檢驗是為確定在何種情況下某些子樹還沒有比一個葉節點更有意義，在此種情況下那些子樹會被修改掉。有三種修改選項可以讓您影響決策樹的修改操作：「置信度」、「成本複雜度」和「無」。

修改過程比限制樹高度或增加拆分下限過程的速度要慢，這是因為要建立整個樹然後再進行修改。然而，有選擇的進行修改將形成更準確的分類器。

*置信度*是預設的刪改方法。越高的值表示越多的修改；而越低的值修改越少。預設的修改置信度 0.7 指出應用於「決策樹」的推薦修改數量。如果資料包含噪音 (誤差和異常)，降低值來建立更小的樹。最低的可能性為 0 (不修改)；沒有上限。修改參數為 0 的情況下，只有當單一節點的誤差率至少與子樹的一樣低時，才會修改子樹。

成本複雜度交換樹的誤差率 (其成本) 和樹的葉節點數 (其複雜度) 並允許您選擇比最小成本樹更小的樹。這就反映了標準差的值比可以接受的最小成本樹所花費的成本更大。將該參數設為 0 得到最小開銷樹；設定參數為 0.5 得到大小最小的樹，其誤差率不大於 0.5，標準差比最小開銷樹差。預設的設定為 0，選擇成本最小且尺寸最小的樹。越高數目表示越多的修改。如果資料包含許多噪音 (誤差和異常)，增加該值來建立更小的樹。如果樹修改成為一個節點，降低值以降低修改量並顯示更多樹的結構。請參閱第 60 頁的「成本複雜性」可得到更多的資訊。

無不執行刪改。雖然這會產生對於訓練資料來說較為冗餘的樹，導致試驗資料較高的分類率，但它卻允許您調查決策樹的完整結構。

- 允許一次拆分

按一下核取方塊可使導入工具在擁有超過兩個值的標稱屬性上產生兩向拆分。正常情況下，標稱屬性上的拆分產生的分支與其擁有的值一樣多 (請參閱第 72 頁的「建立決策樹」)。例如，屬性「顏色」上的拆分會產生紅色、綠色、黃色以及藍色的分支。如果啟動一次拆分功能，導入工具也可以產生只有兩條線的拆分，例如紅色和非紅色。一次拆分正好隔離欄中的其中一個可能值。當資料包含有很多可能值的屬性，而其中一些值可以很好地區分標籤時，這種類型的拆分就會有效。

- 推進

「推進」用於提高分類準確性，雖然它是一個耗時過程。在推進過程中不執行可視化過程。

當執行過程繼續時，狀態視窗顯示估計誤差。估計誤差是該算法重複將新加權分布賦給訓練集以及在重設加權集上導入分類器的結果。請參閱第 37 頁的「推進 (boosting)」可得到更多的資訊。

搜尋和篩選面板

在「檢視」功能表中選擇「搜尋面板」和「篩選面板」(IRIX 中的「顯示」功能表) 會產生一個對話方塊，您可指定搜尋 / 篩選物件的標準。決策樹的項目選擇總是一樣的。這些將在下面描述。

可限制搜尋或篩選只能在特殊標籤上進行，或者選擇分類清單中的值（在最上層的窗格中）或使用分類項目（在底部的窗格中），這些都允許執行更強大的比較操作（例如「符合」）。往下捲動可以得到其它的標準。下面描述其它的項目：

- *子樹權重*可以限制對那些包含指定子樹權重（如果沒有設權重則為記錄數）的長條圖或基準塊的搜尋或篩選操作（取決於選擇的是長條圖或基準塊單選鈕）。例如，您就可以限制對權重不低於 50 的長條圖進行查詢操作。
- *檢驗屬性*可以將搜尋或篩選操作限制在正在檢驗並具有指定標籤值的節點上。請注意：決策節點標籤代表檢驗屬性，而葉節點標籤表示預計標籤。例如，如果選擇了*測試包含年齡的屬性*，就只考慮那些測試了年齡值的節點。
- *檢驗值*將搜尋或篩選操作限制在帶有指定收入線的節點上。
- *百分比*可以將搜尋或篩選操作限制在權重佔整個節點權重達到指定百分比的長條圖上。例如，您想找到所有對於一個指定分類所佔權重大於 80% 的節點。那麼按一下該分類標籤並選擇*百分比 > 80*。如果您選擇基準塊而不是長條圖（基準塊的值為 0）則設定這一項是毫無意義的。
- *純度*可以將搜尋或篩選操作限制在指定純度範圍的節點上。例如，如果想要看純度節點（有一個分類占優勢），您可以選擇*純度 > 90*。
- *檢驗集的子樹權重*可以透過指定檢驗集權重限制對子樹的查詢或篩選操作（如果沒有設定權重則為檢驗集記錄的數目）。
- *檢驗集誤差 / 損失*可以透過指定估計誤差或損失範圍限制對節點進行的查詢或篩選操作。
- *平均值誤差 / 損失標準偏差*可以透過限定標準差範圍限制對節點進行的查詢或篩選操作。
- *層*可以限制對特殊層或層區間進行的搜尋或篩選操作。例如，您可以只搜尋前五個層級。

對於決策樹，以下的項目或選項很少有用。

- 等級找出所有的節點和線，而這些節點和線與從根生出的路徑上的值相符合。然後標示這些節點的子節點。
- 設空值為 0 在決策樹導入工具中不使用，因為不會對決策樹產生空值。

一旦完成搜尋，黃聚光燈就會反白標示與搜尋標準相符的物件。要顯示黃色聚光燈下的物件資訊，將箭頭移到聚光燈上，則資訊會顯示在左上角，在標籤「游標位於」之下；在黃色聚光燈下選擇或縮放物件，用滑鼠左鍵按一下聚光燈即可；如果您在按下的同時也按了 Shift 鍵，就不會進行縮放操作。

一旦完成篩選操作，則畫面只會顯示與篩選標準相符的節點。

離散標籤功能表

「工具管理員」中「分類」索引標籤上的「離散標籤」功能表提供可能存在的離散標籤的清單。離散屬性（分組值、字元字串值或幾個整數）的值數有限。選擇只帶有極少值（例如，兩個或三個）的標籤是很明智的。如果沒有離散屬性，功能表顯示沒有離散標籤，並且開始繼續按鈕也停用。然後您必須，利用「工具管理員」的「資料轉換」面板並經由分組或增加欄來建立一個離散屬性。

追溯

在目前的可視化過程中能夠看到與選擇相對應的原始資料通常非常有用。這是指追溯。在可視化過程中選擇物件，並將下面的資料傳送至「工具管理員」，您便可以查看原始資料或利用其它工具來進行分析。即使某種工具擁有項目結尾所說明的附加功能，MineSet 中的所有工具都具有這樣的功能，並且某些工具還具有該項目結尾所說明的附加特徵。

在可視化過程「選項」下拉式功能表中有兩項可以執行 追溯：

- 傳送到工具管理員

當選擇該選項，產生可視化過程的操作歷史會置入「工具管理員」中。將與可視化工具中用戶選擇所對應的篩選操作加入至歷史中。將該篩選操作儘早的放入歷史中，下面將說明此限制。

- 顯示原始資料

利用「送入工具管理員」過程中，該選項該選項從產生可視化過程的操作歷程開始，並將篩選操作儘早增加到歷史中。移除在歷史的篩選之後進行的所有非篩選操作。新歷史用於產生顯示在「記錄檢視器」中的表。如果篩選操作能夠放在歷史的開始位置，則顯示的資料是原始記錄；否則，就會產生警告，指出不完全是原始資料。工具管理員的狀態不會改變 (除非當前沒有執行)。

在另一種情況下，「工具管理員」執行該操作。如果沒有執行「工具管理員」，它將自動啟動。

只有用「工具管理員」產生的可視化過程才可用於追溯操作。這些可視化過程的每個 *.schema* 檔案包含 *歷史* 部分，用來通知「工具管理員」關於檔案如何產生。幾個特殊的挖掘可視化過程，例如「學習曲線」、「混淆矩陣」以及「上升曲線」並不包括在歷史中，且不支援追溯。

當選擇用於追溯的物件，您就隱約地指定基於可視化表格的篩選語句。如果在可視化之前已經轉換表格，「工具管理員」可能必須改變篩選操作將它儘早放入歷史中。例如，如果該篩選操作是根據已欄分組而建立的，那麼「工具管理員」必須將它改變來引用分組之前的欄。

因為「工具管理員」無法為所有的操作調節篩選過程，因此篩選過程就不能放到歷史的起點處。特殊情況下，如果篩選過程引用由 *增加欄*，*組合*，或 *應用模型* 建立的欄，則篩選過程在歷史中的位置不會比建立欄的操作更早。此外，「工具管理員」無法將篩選操作過程移到現有的採樣操作之前，因為它會大大地改變採樣結果。

「顯示原始資料」模式試圖削減歷史以顯示儘早階段的記錄。然而，要防止該模式顯示多於所選的記錄，削減操作就不能移除任何其它的篩選過程（或者是由使用者建立或者是以前追溯產生的結果）。

概化上尋和細化下尋

概化上尋或細化下尋是指按一下實體來觀察更多或更少細節的操作。只有在「地圖」和「決策表可視化工具」中才可用。

細化下尋是用滑鼠右鍵按一下物件。概化上尋操作是，在按下 **Ctrl** 鍵的同時用滑鼠右鍵（或用滑鼠中鍵）按一下對象。在背景中進行的概化上尋或細化下尋操作會在整體範圍內執行。

錯誤估計

當建立分類器時，如果能夠知道將來的執行情況是非常有用的（也就是分類器誤差率代表的意義）。影響分類誤差率的因素包括：

- 訓練集中的可用記錄數目

因為導入工具必須從訓練集中學習，所以訓練集越大，分類器的可靠程度越高；但是訓練集越大，導入工具建立分類器所花的時間越長。隨著訓練集的規模降低，誤差率的改進也隨之下降（這就是一種減少回報效應）。

- 屬性數目

更多的屬性意味著導入工具要計算更多的組合、產生更多的問題以及花費更多的時間。請注意：有時隨機相關性會誤導導入工具，建立準確性更低的分類器（技術上稱為冗餘）。如果屬性與任務並不相關，則將它從訓練集中移除（可以用「工具管理員」執行）。

- 屬性中的資訊

有時屬性中並沒有足夠的資訊來預測標籤並保證較低的誤差率 (例如，根據眼睛的顏色試圖確定某人的薪資)。增加其它屬性 (例如，職業、每星期的小時數及年齡) 也許能夠降低誤差率。

- 未來未標示記錄的分布

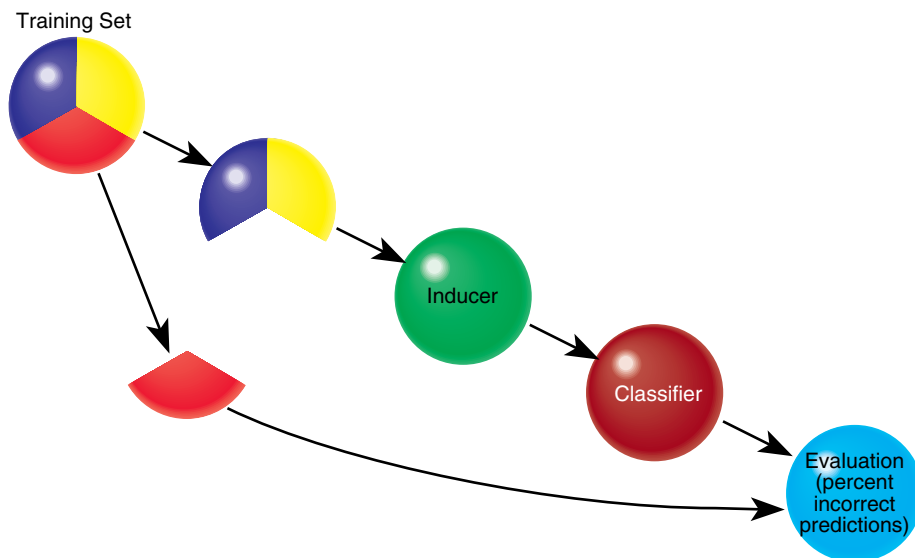
如果產生於分布中的記錄與訓練集中的不同，誤差率可能會很高。例如，如果您從包含家庭轎車的訓練集中建立了分類器，在嘗試對包含許多賽車的記錄進行分類時，該分類器也許並不適用，因為屬性值的分佈是很不相同的。

下面描述分類器中用於估計誤差率的兩種普通方法。兩種方法都假設未來的記錄將從和訓練集相同的分布中抽樣得來。

- 預留

一部分記錄 (通常為三分之二) 用於訓練集，而剩下的記錄為 檢驗集。導入工具只顯示資料的三分之二並建立分類器。然後，導入的分類器對檢驗集進行分類，並且檢驗集中的誤差率或損失就是估計誤差率或估計損失。圖 1-17 顯示這種錯誤估計方法。

當修正開啓並且「用盤顯示訓練集」的選項選取時，垂直長條圖顯示以圓盤形式出現的訓練集的分布。圓盤的高度與長條圖的高度比例一致。您應該期望圓盤在長條圖上的高度與更多導入工具中用於「預留」模式的比率值的高度一致。



■ 1-17 估計分類器的準確性

這種方法雖然很快，但是因為它僅僅使用資料的三分之二來建立分類器，因此就沒有充分地利用資料來進行學習。如果利用所有的資料，就有可能建立一個更為準確的分類器。

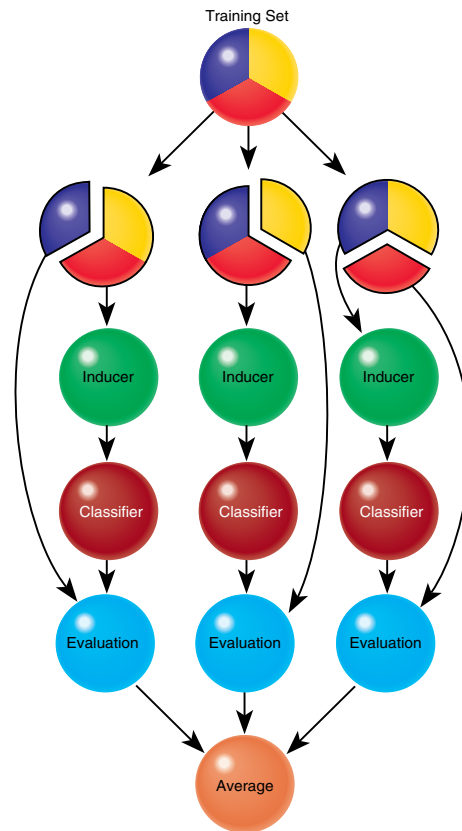
- 交叉驗證

資料分為 k 個互不重疊、大小相同的子集。導入工具訓練和檢驗了 k 次；每次都把所有的資料減去一個不同子集進行訓練，然後在預留子集上進行檢驗。估計誤差率是所得誤差的平均。圖 1-18 顯示 $k=3$ 的交叉驗證（注意預設值為 $k=10$ ）。

交叉驗證可以重複很多次 (t)。如果進行了 t 次，那麼 k -子集交互驗證就會有 $k*t$ 個分類器產生並被評估。這也就意味著交叉驗證的次數是 $k*t$ 次。預設情況下， $k=10$ 且 $t=1$ ，所以交叉驗證需要花費建立單獨分類器的 10 倍時間。

增加重複的次數 (t) 也就增加執行的時間並改善錯誤估計和相應的置信度間隔。

您可以增加或減少 k 。減少到 3 或 5 就會縮短執行時間；但是估計就會因為訓練集規模的變小而產生偏差。可以增加 k ，但這只有在對較小的資料集時才有用。



■ 1-18 分類器的交叉驗證 ($k=3$)

通常，預留估計應該在開發階段使用，並且資料集應有超過 5000 以上的記錄。交叉驗證應該用於分類器的最後建立階段，且資料集的規模較小。

修正是將整個資料集應用於利用訓練集中的資料所建立起來的分類器或模型。當利用預留錯誤估計時，您留下一部分的資料用於檢驗。當您經由分類器結構修正所有資料時，您可以削減最終分類器的誤差，因為計數、權重和機率反映的是整個資料集。如需詳細的討論，請參閱第 33 頁的「修正」。

證據模型

「證據分類器」模型將資料集中的每筆記錄指定一個分類。「證據可視化工具」顯示證據模型的結構。可視化工具有助於理解分類過程中特殊屬性值的重要性。同樣地，它也可用於了解分類過程，以及回答「What if」問題。請參閱《*MineSet 3.0 for Windows 企業版使用者指南*》可以得到如何使用和解釋「證據可視化工具」的更多資訊。

證據導入工具

證據導入工具建立一個模型，該模型假定對分類而言各屬性值的機率是獨立的。例如，對於每個蝴蝶花類別（藍色蝴蝶花、多色蝴蝶花和純白色蝴蝶花）來說，*iris* 資料集假設四個屬性（花萼長度、花萼寬度、花瓣長度以及花瓣寬度）都是獨立的。當然這個簡化的模型並不真實，但這個模型用於最初的資料開發卻非常有效並且它的分類預計表現在實際應用過程中也很好。

每個屬性值或值的範圍（對於分組的連續屬性）都剛好定義了一張圖表，依次為每個分類標籤指定條件機率。對已知的記錄進行分類，一種做法是計算每個分類的機率，其方法是將其先驗機率乘以矩陣中每一列的相應條件機率。最終的乘積為每個分類提供相對機率，並且得出最高的值就是預計分類。如果屬性的值未知，則會將其忽略。這些「空」值由位置略有位移的圖所代表，並且當您從 MineSet 提供的樣例中帶出 *iris.schema* 檔案時這些值會出現在左邊。請參閱附錄 A, “配置和資料檔案樣例,” 可得到路徑名，或遵循《*MineSet 3.0 for windows 企業版使用者指南*》中所提供的指示。

當您用「證據」導入工具顯示 iris 資料集時，每個分類標籤的先驗機率會用螢幕右側「標籤機率」窗格中的餅圖來說明。分類標籤的先驗機率是在忽略所有屬性值的情況下，隨機抽取記錄時見到資料中標籤的機率。從數學上來說，該數為具有分類標籤的記錄除以記錄總數的比值。要在「主視窗」中的「證據」(塊圖)和「機率」(餅圖)之間切換可以按一下「證據」標籤或在「檢視」功能表中切換「證據模式」。

條件機率，由螢幕左側「主視窗」中的塊狀圖來表達，顯示每個標籤值條件上每個屬性值間的相對機率。塊狀圖中份額的大小指出，在考慮記錄中的已知屬性值後分類器增加到先驗機率中的證據數量。如果份額的大小一樣，值就不相關，並且分類器向所有的分類增加相同數量的證據。

從技術上，圖中的份額代表了在分類標籤 L 的條件上，屬性值 A 的標準化條件機率。條件機率 $P(A|L)$ ，是指從帶有 A 值的 L 標籤中抽取的隨機記錄機率。在預設的設定下，根據記錄權重可計算機率。例如， $P(0.75 < \text{花瓣 寬度} < 1.65 | \text{iris-versicolor})$ 為 91.6，因為 iris-versicolor 標籤的記錄有 36 個，而其中有 33 個的花瓣寬度在這個區間內。

長條圖 / 證據模式中灰基準塊的高度顯示所選標籤的證據。整個證據值的支持由 $-\log[1 - P(L) / \sum(P(L_i))]$ 計算得出，而反對證據值由 $-\log[P(L) / \sum(P(L_i))]$ 計算得到。減去最小證據切換影響基準塊並同樣影響較小的長條圖。

重要性是關於標籤、對預測能力的度量。「主視窗」提供有價值的直觀結果，這不僅在於每個屬性的重要性影響著類值，而且還在於特殊屬性值的重要性。

導入證據分類器

證據分類器的自動用計數 (或權重) 來計算機率。證據分類器自動從資料中導入 (產生)。由記錄和與每個記錄相關的標籤所組成的資料稱為 *訓練集*。

用以下方法可以得到機率：

1. 所有的連續屬性都被離散化 (間隔取值)，因此這些組內的類分布盡可能的不同。如果系統中安裝多處理器版本的 **MineSet**，則範圍的數目將自動確定，這樣就可以啓動平行過程，使屬性能並行地分組。任何屬性的自動分組過程可以經由正確地執行「工具管理員」中的分組操作重新進行。
2. 先驗機率是訓練集中每個類比例。
3. 條件機率是指，訓練集中每個分類標籤條件上每個屬性值的機率。(該圖顯示屬性值已進行了標準化。)

每列中圖的數目就是由導入工具產生的離散範圍的數目。如果只有一個範圍，就意味著該屬性本身在預計標籤時不起作用。初始時，標籤的先驗機率在「標籤機率」窗格中顯示。

每個屬性值或值的範圍 (對於分組的連續屬性) 都剛好定義了一張圖表，依次為每個分類標籤指定條件機率。

- 拉普拉斯校正

這將會使機率偏向平均值，避了極端數 (例如 0 和 1)。如果在「證據導入工具」對話方塊的*進階選項*中 (或 **IRIX** 系統，*更多導入工具選項*) 選擇拉普拉斯校正選項，並且將因素設為空或 0 後，就可以應用因子為 $1/t$ 訓練集權重 " 的啓發式過程自動進行拉普拉斯校正。也請參閱第 108 頁的「拉普拉斯校正」。

- 每組最小權重

「證據導入工具」將所有的連續屬性離散化。該選項可使您定義每組中實例的最小數目。自動設定有一個可根據資料集規模來設定資料的啓發式過程：資料集越大，從總體上說，每個分組中所允許的最小記錄數目就越大，並且分組的寬度就越小。如果資料集非常大，您可以得到更多想要的離散範圍。要減少組數，便增加該值。

- 自動選擇欄

該選項提供一個過程，只選擇那些最有助於預測的欄。因為額外的欄可以降低證據分類器預測準確度的估計，所以該過程可自動搜尋好的欄子集。只有使用那些有用的欄。這一過程要花費很長的時間，尤其是在有很多欄的情況下。刪除那些高度相關並會降低準確性的欄很有用。自動欄選擇和「推進」不能同時使用，否則將會消耗很長的時間來完成。

自動選擇欄執行了一個搜尋過程，該過程用於搜尋可降低分類器誤差的欄的最優子集。這些欄的選擇過程使用包裝器方法來估計不同屬性集的誤差。利用交叉驗證估計分類器的誤差來進行每個屬性子集的評估。增加或移除欄的根據是利用「最好 - 第一」搜尋機制得到的錯誤估計。預設的模式下，搜尋從空的屬性集開始。選擇^{向後}選項，便開始搜尋整組的選項，速度會比較慢，因為較大的模型得從頭建立（請參閱《*MineSet 3.0 企業版介面指南*》中的證據可視化工具的資料檔案格式）。

啟動證據可視化工具

有幾種方法可以啟動「證據可視化工具」：

- 從「工具管理員」中的「分類」索引標籤上執行「證據導入工具」。在導入工具建立分類器後，會自動啟動「證據可視化工具」。請參閱以下的詳細內容，關於使用與「證據可視化工具」結合在一起的「工具管理員」。
- 用「工具管理員」從「可視化工具」功能表中啟動「3D 可視化工具」。
- 如果您知道要利用的配置檔案，連按兩下該配置檔案的菱形圖示。這樣就啟動了 3D 可視化工具並自動載入指定的配置檔案。只有當配置檔案以 *.eviviz* 結尾時才有效（對於用「工具管理員」為「證據可視化工具」建立的配置檔案都是這樣）。
- 要從 IRIX 指令行中啟動「證據可視化工具」，可以輸入：

eviviz [*configFile*]

configFile 是可選的，指定所使用的配置檔案的名稱。如果未指定配置檔案，您必須用「檔案」>「開啓」來指定一個。

啟動證據可視化工具的 IRIX 選項

`-quiet` 選項刪除顯示過程的跳現式對話方塊。在您的 `.Xdefaults` 檔案中加入一系列代碼可以使該選項永久可用

```
*minesetQuiet:TRUE
```

也請參閱第 199 頁的「警告選項」。

在 Windows 上，該選項在 3D 可視化工具的「特性」面板中可找到。

證據導入工具選項

選擇「進階選項」(IRIX 上為「更多導入工具選項」)可產生「導入工具選項」對話方塊。該對話方塊由三個面板組成：

- 最上層的面板顯示在「工具管理員」的「資料目標」面板中所做的選擇。「錯誤估計」的類型由該模型限定。
- 左下角的面板可讓您指定進一步的「導入工具」選項(說明如下)。
- 右下角的面板可讓您指定「錯誤估計」選項(除非在「資料目標」面板中選擇的是「僅用於分類器」模式，在這種情況下區域內是空的)。該面板中顯示的選項取決於您所選擇的「錯誤估計」類型(請參閱第 80 頁的「錯誤估計」)。

如果在「進階導入工具」選項中指定「損失矩陣」，標示為*使用損失矩陣*的按鈕就會出現在機率餅圖的右下角。當選取時(預設情況下)，「損失矩陣」會用於調節機率的顯示。最大的份額所顯示的為考慮「損失矩陣」的分類。要看到不使用「損失矩陣」時的機率狀態，取消選擇*使用損失矩陣*按鈕。灰份額表示一個經由編輯用來預計空值的欄存在著。如果灰份額是最大的份額，則分類器預測為空值。如需進一步說明，請參閱第 111 頁的「損失矩陣」。

在「證據可視化工具主視窗」中「選擇項目」

在選擇模式中，游標表現為箭頭。您可以反白標示物件（餅圖或長條圖），方法是將游標移動至該物件上。關於該物件的資訊就會出現在「主視窗」的上面。只要游標停留在物件上，就一直顯示資訊。

- 如果物件是一個餅圖，消息將採用這樣的格式：

< 屬性名稱 > : < 值或範圍 >
 權重 = <weight>

在此，*權重*是位於區間或具有該屬性值的所有資料點的權重和。圖高與該數成比例。除非使用記錄加權，否則權重就代表記錄計數。

- 如果物件是條，那么消息將採用這樣的格式：

(< 屬性 > = < 值 >) ==> Prob(< 選取標籤 >) = x% [low%-high%]
 Evidence=z
 < 選取標籤 > ==> Prob(< 屬性 > = < 值 >) = y% [low%-high%]
 權重 = <weight>

在此，*x* 是假設在記錄具有反白標示屬性值，且同時具有所選標籤值的機率。方括號的範圍，[low%-high%] 提供 95% 的置信度間隔。同樣地，*y*% 是在記錄具有反白標示屬性值的條件上並具有所選標籤值的機率。長條圖的高度代表證據而非機率。證據的數量 *z* 與長條圖的高度直接相關。如要指定要估計的類別，可以對證據進行求和計算（與機率不同，它必須相乘）。*權重* 是具有該值的資料點之權重。

從技術上，*支持證據*定義為

$$-\log \left[1 - \frac{P(A|L)}{\sum_{i=1}^N P(A|L_i)} \right]$$

而*反對證據*定義為

$$-\log \left[\frac{P(A|L)}{\sum_{i=1}^N P(A|L_i)} \right]$$

A 是屬性值，L 是所選的標籤值，而 N 是標籤值的數目。當計算長條圖高度時，上述括號中的表達式上加上一個很小的數以阻止長條圖無限制的增高。「主視窗」中的「支持」或「反對」的周圍有一個方塊說明可以進行按一下操作。在方塊上按一下可以切換表示法。

灰矩形基準塊的高度（上面有直立著的長條圖）代表由先驗機率提供的證據的數目。例如，如果標籤是轎車的汽缸，很少有三汽缸的轎車，所以當支持證據顯示時，其基準塊很低，而當反對證據顯示時，基準塊很高。您可以將頂部的單獨長條圖的高度加入此高度。

支持證據在確定預測特殊標籤值的過程中最有效的值時可能會有些用處。

證據的數量（長條圖高度）不是直接從反白標示顯示的機率中導出。相反地，證據取決於相對於其它（根據上述等式所得到的）標籤值機率的條件機率。

您也可用以下的方式從屬性列中選擇任意數量的值，即當游標位於屬性值之一的上方時按一下滑鼠左鍵。右邊標籤機率窗格中的大餅圖改變為反映您所選的項目；它現在顯示的是在剛選取之屬性值上的後驗機率。分類仍然按順序排列著，所以與最大份額對應的一個出現在右邊清單中的頂部。經由一起乘以每個屬性的條件機率，「證據可視化工具」形成新的後驗機率分布，然後將該結果乘以先驗機率並標準化為 1。

這種乘法操作與條件獨立性假設對應。當違反了該假設，並且出現幾個屬性實例時，預測類的機率就很可能是錯誤的（即使最終的分類是正確的）。當您執行導入工具時，顯示在「狀態」視窗中的估計誤差有助您判定這一假設的合理性。如果誤差率 / 損失為低時，則領域內的假設表現適當的健全性。

當在「標籤機率」窗格中選擇特殊標籤時，對於屬性的每個值「主視窗」顯示長條圖而不是塊圖或餅圖。列中的標題是「支持證據」，「支持」周圍的方塊指示它可以被選擇。

按一下「支持證據」中的「支持」，標題會切換顯示為「反對」。結果，長條圖高度改變成顯示反對標籤的證據。

因選擇長條圖而對右邊「標籤機率窗格」中的大機率餅圖所產生的效果與選擇塊圖或餅圖所產生的一樣。長條圖的高度表示支持或反對選定標籤的證據數目。因為機率的對數用於代表證據，所以長條圖高度就增加到累計證據中（反之，就必須乘以機率）。

證據可視化工具功能表

您可使用五個下拉式功能表來存取「證據可視化工具」的附加功能：「檔案」、「檢視」、「選擇」、「名稱排序」以及「說明」。如果沒有指定配置檔案就啟動「證據可視化工具」，則只有「檔案」和「說明」功能表可用。請參閱「檔案」和「說明」項目可以得到有關功能表的細節。

檢視功能表

利用「檢視」功能表可以控制「證據可視化」窗格中的顯示設定：關於一般功能表項目的資訊請參閱第 185 頁的「檢視功能表」。根據系統平台的的不同，功能表也包含一些附加選項：

- *顯示為證據* 切換「證據」的顯示。如果選取，畫面左側會顯示證據。如果沒有核取，畫面會顯示機率。
- *按重要性排序* 顯示屬性的排序，此種排序是根據分類所選標籤時屬性的用處來進行。如果關閉該選項，則屬性將顯示與「工具管理員」中「目前欄」下的相同順序。
- *減去最小證據* 當選取標籤並顯示長條圖的時候才應用。該選項開啓時（預設情況下），便減去整個標籤值中最小最小值對應的長條圖高度。每個屬性中每個值不同，該數量也不同，但是對於已知的屬性值，所減去的數量對應標籤值是固定的。啓動該選項會因為對所有標籤值都減去相同的量而放大一些小的差別。
- *使用拉普拉斯校正* 切換到使用「拉普拉斯校正」。如果在更多導入工具選項對話方塊中指定了拉普拉斯校正，就會利用該值，否則便提供預設值。
- *使用場景檢視器*（或 *使用檢查員檢視器*）可轉換成 3D 瀏覽的可選模式。

名稱排序功能表

「名稱排序」功能表可控制標稱屬性值如何排序。選擇的有關資訊請參閱第 129 頁的「名稱排序功能表」。

選擇功能表

「選擇」功能表可以對基本資料進行追溯。要進行追溯，首先選擇值和 / 或類的某種組合，然後在兩種追溯方法中選擇一個對基本記錄進行追溯。對於幾個工具該功能表是相似的，請參閱第 153 頁的「[「選項」按鈕](#)」可得到功能表項的解釋。

檔案功能表

絕大多數可視化工具的「檔案」功能表是相似的，並包含一些選項：

Windows 系統

- *開啓*載入並開啓一個配置檔案，並在主視窗中顯示。之前顯示的資料被清除。使用 *開啓* 來查看一個新的資料集，或在改變配置之後查看相同的資料集。
- *重新開啓* 重新開啓目前已開啓的檔案。在配置檔案或資料檔案更新後可以使用該功能。
- *儲存影像* 將目前可視化工具視窗中的狀態儲存到影像檔案中，並可以選擇要儲存整個視窗（包括所有的符號），或者只儲存主要的圖形物件畫面（預設情況下是整個視窗）。
- *列印影像* 將目前視窗中的狀態輸出至印表機。您可以利用「*列印*」對話面板（預設為系統的預設印表機）來指定輸出的印表機，並且可以像 *另存新檔* 對話方塊一樣，選擇是列印全部視窗或只列印主要畫面視窗。
- *預覽列印* 顯示即將列印的結果。
- *列印設定* 開啓一個對話面板來設定列印輸出（預設為系統的預設印表機）。
- *網上發布* 將目前檔案列印到網上相容的檔案中。

- *特性*開啓「特性」對話方塊來指定滑鼠對照，並在指令執行過程中切換警告訊息。在此還可指定聲音效果、動畫播放以及預設可視化工具的字體大小等。
- *啓動工具管理員*啓動「工具管理員」（如果還未執行），並恢復其在啓動「樹狀可視化工具」時的狀態。
- *最近檔案*開啓任意可視化工具近期使用過的檔案。功能表列出最近開啓的四個檔案。
- *結束*關閉所有的視窗並退出應用程式。

IRIX 系統

- *開啓*開啓並載入一個配置檔案，並在主視窗中顯示。之前顯示的資料被清除。使用*開啓*來查看一個新的資料集，或在改變配置之後查看同一資料集。
- *重新開啓*重新開啓目前已開啓的檔案。在配置檔案或資料檔案更新後可以使用該功能。
- *另存新檔*將目前可視化工具視窗中的狀態儲存到影像檔案中。使用者指定檔案名稱（對於「樹狀可視化工具」，預設為 *treeviz.rgb*），格式（預設為 *rgb*），並可選擇是儲存整個視窗，包括所有的圖例，或者只是儲存主要的圖形物件畫面（預設情況下是整個視窗）。
- *列印影像*將目前「可視化工具」視窗中的狀態輸出至印表機。您可以利用「列印」對話面板（預設為系統的預設印表機）來指定輸出的印表機，並且可以像*另存新檔*對話方塊一樣，選擇是列印全部視窗或只列印主要畫面視窗。
- *網上發布*將可視化工具儲存為 *.mtr* 檔案，該檔案適合在網上發布。
- *啓動工具管理員*啓動「工具管理員」（如果還未執行），並恢復其在啓動「樹狀可視化工具」時的狀態。
- *結束*關閉所有的視窗並退出應用程式。

檔案需求

大多數 MineSet 可視化工具除了配置檔案外至少需要兩個檔案：*.data* 檔案和 *.schema* 檔案，當您利用「工具管理員」調用工具時，檔案會自動產生。

.data 檔案由索引標籤分隔的欄位列所組成。您可以從資料來源（例如，Oracle、INFORMIX 或 Sybase 資料庫）中抽取資料來產生資料檔案並用您熟悉的文字編輯器（例如，WordPad、jot、vi 或 Emacs）對其進行格式轉換。請參閱 *《MineSet 3.0 企業版介面指南》* 可以找到所需的檔案格式。[表 1-10](#) 提供使用「工具管理員」為可視化工具建立的檔案附檔名類型。

表 1-10 預設樣例檔案副檔名。

工具	資料檔案附檔名	方案檔案附檔名	配置檔案附檔名
關聯規則	<i>.rules.data</i>	<i>.rules.schema</i>	<i>.rules.scatterviz</i>
聚類可視化工具	無	無	<i>.clusterviz</i>
決策表可視化工具	<i>.dtableviz.data</i>	無	<i>.dtableviz</i>
証據導入工具	無	無	<i>.Eviviz</i>
地圖可視化工具	<i>.mapviz.data</i>	<i>.mapviz.schema</i>	<i>.mapviz</i>
記錄檢視器	<i>.data</i>	<i>.schema</i>	無
分散可視化工具	<i>.scatterviz.data</i>	<i>.scatterviz.schema</i>	<i>.scatterviz</i>
平板可視化工具	<i>.splatviz.data</i>	<i>.splatviz.schema</i>	<i>.splatviz</i>
統計可視化工具	無	無	<i>.Statviz</i>
樹狀可視化工具	<i>.treeviz.data</i>	<i>.treeviz.schema</i>	<i>.treeviz</i>

當啓動可視化工具或開啓檔案時，您必須指定配置檔案而不是 *.data* 檔案或 *.schema* 檔案。「記錄檢視器」可以開啓任何 *.schema* 檔案。

資料檔案可以擁有使用者定義的附檔名（所提供的樣例有 *.data* 附檔名）。

篩選按鈕

「工具管理員」中「資料轉換」窗格上的按鈕可以使您利用數學表達式來篩選資料。結果表只包括表達式為真時的記錄 (或者，如果記錄為數字型的，則為非 0 值)。當按一下 *篩選*，就會產生「篩選」對話方塊。

該對話方塊可以讓您在左邊選擇欄名和操作符，在右邊建立表達式。如需表達式定義語言的完整說明，請參閱第 1 頁的「[增加欄](#)」。

篩選面板

從「檢視」下拉式功能表中可以進入「篩選面板」，以下部分描述這個選項。「分散」、「平板」、「對照」和「決策表可視化工具」使用該面板。「樹狀可視化工具」擁有一個類似的面板。「證據可視化工具」有其自己的相應功能部分。

篩選面板 開啓一個面板，您可以根據一個或多個標準減少在主要檢視區域中顯示的實體數目。您可以使用篩選面板來微調顯示設定，強調特別資訊或縮減顯示資訊的數量。

篩選範圍 (僅存於「分散可視化工具」) 可讓您指定圖形物件的高度是在整個資料集的範圍內，或者在篩選過的資料中進行比例變化。

篩選面板有兩個部分。最上層的部分可讓您根據字串變數進行篩選操作。選擇變數的所有值，按一下 *設定全部*。要清除目前選擇，按一下 *清除*。要選擇一個值就按一下它。要取消選擇一個值，只需再按一下。

下面的部分可讓您根據字串和數字變數的值進行篩選操作。當您操縱滑動桿時，只有那些值保持不變的變數才能用於篩選。捲動窗格可顯示不同等級的內容 (「包含」、「相等」、「符合」、「為空」)。

要篩選數字型值，就輸入一個值並選擇相關操作 (=, !=, >, <, >=, <=)。要篩選字母數字值，就輸入字串。您可以使用三種字串比較類型中的一種：

- 包含指出含有適當的字串。例如，**California** 包含了字串 **Cal** 和 **form**。
- 「相等」需要正好符合的字串。
- 符合允許通用字元：
 - 星號 (*) 代表任何數量的字元。
 - 問號 (?) 代表一個字元。
 - 方括號 ([]) 包含要配對的字元清單。

例如，**California** 與 **Cal***，**Cal?ifornia** 和 **Cal[a-z]ifornia** 相符。

在一些情況下 (通常與「工具管理員」中的分組有關)，出現值的選項功能表，而不是本文欄位。要忽略該變數，選擇「選項」功能表中的**忽略**。您可以使用帶有這些選項的相關操作符 (例如 >=)。這就意味著選擇了包括後續值的指定值。

除了數字型和字串型的比較操作以外，您還可指定 **is null** 操作，當值為空時則為真。

每個欄位的右邊是附加選項功能表，用該功能表可以指定「**and**」或「**or**」操作選項。例如，您可以指定銷售 >20 和 <40。已知變數可以有多個「**And**」或「**Or**」子句，但不能在單一變數中同時使用「**And**」和「**Or**」。

按一下**應用**按鈕來啟動篩選操作。如果在面板活動時，您按下 **Enter** 鍵，就會自動啟動篩選操作。

增量比

「增量比」是由除以熵 (純度的變化) 的「公共資訊」組成的「決策樹」拆分標準。該熵是忽略了標籤值的拆分所產生的熵。請參閱第 71 頁的「**決策樹**」中關於「決策樹」拆分標準的討論。

說明 (IRIX)

除了 F1 或 Shift-F1 鍵組合以外，每個工具中的「說明」功能表都可存取五個說明功能：

- *按一下得到說明*，將游標變為問號。將游標放在可視化工具主視窗中的物件上並按一下滑鼠會使說明螢幕出現；該螢幕包含該物件的訊息。關閉說明視窗後，游標復原到原來的形狀並取消說明功能的選擇。該功能的鍵盤快速鍵為 **Shift+F1**。(注意，在有些工具中也可以將箭頭狀的游標放在物件上，並按下 **F1** 功能鍵來存取該物件的說明螢幕。)
- *總覽*提供工具主要功能的簡單總結，包括如何開啓檔案以及交互操作結果檢視。
- *索引*提供完整說明系統的索引。該選項目前停用。
- *按鍵與捷徑*提供所有可使用加速鍵的可視化工具功能的快捷方式。
- *產品資訊*產生帶有可視化工具的版本號碼和版權注意事項之螢幕。

說明 (Windows)

除了 F1 鍵，每個工具中的「說明」功能表都提供對所有「MineSet 3.0 企業版」手冊的存取。*產品資訊*產生帶有可視化工具的版本號碼和版權注意事項之螢幕。

柱狀圖可視化工具

「柱狀圖可視化工具」將資料中的連續欄自動分組並將結果傳送至「統計可視化工具」。請參閱第 168 頁的「統計可視化工具」。您可以設定下列「柱狀圖可視化工具」選項：

- 您可以挑選分組數目或讓 MineSet 自動為您執行此作業。
- 您可以設定修改因子。修改因子說明部分極值將被排除在產生分組之前的值範圍以外。預設的修改因子為 0.05。這就排除了 5% 具極值的實例 (2.5% 具有範圍內的最低值而 2.5% 具有最高值)。修改的目的就是減小在臨界值產生過程中數值的影響。

歷史

您可以使用工具管理員中的歷史表按鈕來得到目前階段的之前操作歷史。請參閱第 171 頁的「歷史表按鈕」。您也可以利用檔案下拉功能表來儲存您的工作。您也可以在完成轉換操作後按一下「工具管理員」的「資料目標」窗格中「資料檔案」索引標籤來儲存資料檔案。當您恢復工作時，載入您想要執行的檔案或合適的階段，而操作歷史也在其中。

預留

預留是指在資料集的部分樣本上訓練模型的過程，使該模型使用剩餘的部分進行檢驗。該過程可在沒有用於訓練模型的資料上估算模型誤差率。用於訓練的資料的比例在 MineSet 中稱為「預留比率」。

導入工具

導入工具是一種從訓練集中建立預測模型的算法，需要帶有標籤的記錄。訓練集是導入工具用來「學習」如何建造模型的資料集中的一部分。一旦建立模型，就能可視化或使用其結構以分類未標記的記錄。執行導入工具可以是 CPU 和 I/O 密集過程。因為這個原因，MineSet 導入工具在 MineSet 伺服器上執行，而不在 MineSet 客戶機上執行（請參閱圖 1-19）。

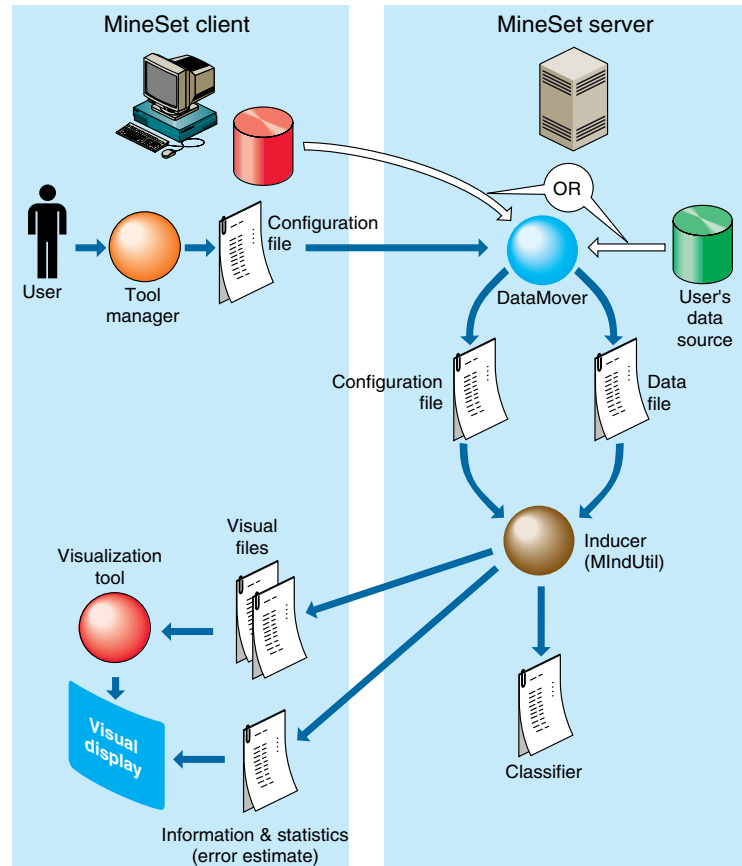


圖 1-19 模型的工具執行順序

導入工具需要訓練集，該訓練集是一個帶有屬性或特徵的表格，屬性之一指定為標籤。一旦建立模型，就可以對新記錄預測標籤。這些新記錄必須存在於擁有模型所用的全部屬性之表格中，並具有它們在訓練集中的名稱和類型。該表不需要包含標籤屬性。如果存在，在預計的過程中會將其忽略。

工具管理員中的導入工具模式

有四種模式可執行導入工具。

- 分類器和誤差估計 (或回歸器和誤差估計)
- 僅用於分類器 (或僅用於回歸器)
- 估計誤差
- 學習曲線

分類器和誤差估計模式 (或回歸器和誤差估計模式) 使用預留方法來建立模型：隨機抽取部分資料來進行訓練 (通常為三分之二) 而剩下的用於檢驗。預留比例可以在 **Windows** 系統的**進階選項**中設定，或者在 **IRIX** 系統中的**更多導入工具選項**中設定 (請參閱第 80 頁的「**錯誤估計**」)。該方法是預設方法並被推薦為初始開發方法。此種方法的速度較快並提供錯誤估計。

僅用於分類器 (或僅用於回歸器) 選項使用所有的資料建立模型。該方法沒有錯誤估計。當只有很少的資料或當您建立最終的模型時，可以利用該模式。

*估計誤差模式*估算使用全部資料建立的模型的誤差 (使用**僅用於分類器**或**僅用於回歸器**模式)。「估計誤差」使用了交互驗證，造成執行時間較長。(請參閱第 61 頁的「**交叉驗證**」。) 當只有少量資料時使用該方法。該導入模型與由**僅用於分類器**或**僅用於回歸器**模式導入的模型一樣。

*學習曲線模式*估算訓練集規模對分類器誤差率的影響 (請參閱第 108 頁的「**學習曲線**」)。

導入工具誤差選項

下列選項可用於微調導入工具的錯誤估計。該「誤差選項」的可用性取決於您選擇的導入工具模式。

在分類器和誤差估計(或回歸器和誤差估計)以及錯誤估計選項中，您可以設定隨機子用以確定資料如何分為訓練和檢驗集。改變隨機子將造成資料劃分為訓練集和檢驗集。如果錯誤估計略有變化，導入過程就會不穩定。

在分類器和誤差估計(或回歸器和誤差估計)選項中您可以設定記錄的「預留比率」以保留為訓練集。預設為 0.666667 (三分之二)。剩下的記錄用於估算誤差。

在估計誤差選項中，您可以設定交叉驗證倍數以及重複該過程的次數。請參閱第 61 頁的「交叉驗證」。

進階導入工具選項

MineSet 支援所有導入工具的進階選項。您可以考慮產生某種錯誤的成本，或對非均勻採樣過程的補償(真實總量中某些部分比其它部分的採樣密度高很多)。另一個選項可使您建立更複雜的模型，該模型在增加計算時間的代價下可能具有更好的準確性。

- **修正** 修正檢驗集設定選項是一個核取方塊，在使用分類器和誤差估計模式時適用於所有導入工具的 Windows 系統中的進階選項，或 IRIX 系統中的更多導入工具選項中可以找到。當「加速」開啓時，修正核取方塊就停用。請參閱第 33 頁的「修正」。
- **混淆矩陣** 顯示混淆矩陣選項是一個核取方塊，在使用「分類器和誤差估計」模式時適用於所有導入工具的 Windows 系統中的進階選項，或 IRIX 系統中的更多導入工具選項。請參閱第 58 頁的「混淆矩陣」。
- **投資回報曲線** 投資回報 (ROI) 曲線與「上升曲線」相似，但它以損失的形式而不是以誤差的形式顯示準確度；並考慮所使用的損失矩陣。顯示 ROI 曲線選項是適用於所有分類過程導入工具的一個核取方塊，可以在 Windows 系統的進階選項，或 IRIX 系統的更多導入工具選項中找到。ROI 曲線需要選擇的標籤值。然後為該標籤值產生並顯示 ROI 曲線。請參閱第 143 頁的「投資回報曲線」。

- **上升曲線**上升曲線為一圖形，該圖形顯示記錄的隨機排序與說明特殊標籤值時由分類器所建立的排序之間的差異。請參閱第 110 頁的「上升曲線」。
- **損失矩陣**「損失矩陣」可讓您改變產生某種誤差的成本或對其重新加權。損失矩陣與混淆矩陣一起可幫助減少誤差的成本。*使用損失矩陣*選項是一個適用於所有導入工具的核取方塊，它出現在 Windows 系統中*進階選項*，或 IRIX 系統中*更多導入工具選項*下的對話方塊的上方。請參閱第 111 頁的「損失矩陣」。
- **加權設定**記錄加權可讓您為每個記錄指定權重；採樣兩次的記錄得到的權重通常為 0.5，而其餘的部分其權重為 1。*使用加權*選項是一個適用於所有導入工具的核取方塊，出現在 Windows 系統中*進階選項*，或 IRIX 系統中*更多導入工具選項*下的對話方塊的上方。為權重選擇欄。*權重保留為屬性*選項用來確定導入工具是否可以使用該屬性完成模型化的目的。在某種情況下，即當權重是實驗設計一部分的分層樣例結果時，該模型就不提供對權重欄的存取，因為它不是真實世界中實體的性質。
- **學習曲線**學習曲線作為由一定量記錄所建立的模型功能，是一個顯示由導入工具所產生模型的誤差之圖形。一般情況下，產生模型所用的記錄越多則誤差越低。請參閱第 108 頁的「學習曲線」。

設定特殊選項

- 總是忽略類型陣列的屬性。
- 資料被視為字串。除非有少量的資料，否則因為離散屬性的限制，這樣的屬性通常會忽略。在執行導入工具之前，您應該將日期分組。

導入工具狀態視窗

在您按下*繼續*之後在「資料目標」面板中，「工具管理員」主視窗底部的「狀態視窗」顯示算法進度並展示導入模型的特別資訊。例如，它顯示了「決策樹」的節點數目，葉子數目以及「決策樹」的深度。該資訊自動儲存在您的工作站的階段檔案中，檔案名稱中帶有 *.out* 附檔名。

- 對於「分類器和誤差估計」(或「回歸程式和誤差」)，第一系列點代表了讀入檔案，然後顯示有關分類器建立進程的訊息，然後再顯示測試設定的分類進程。
- 對於「僅用於分類器」(或「僅回歸程式」)模式，沒有測試設定分類階段。
- 對於「估計誤差」，顯示了次數和倍數。
- 對於「學習曲線」，x 軸上的平均點將在線上說明出來，並且每次執行的平均點都將由一個點代表。

分類器和「誤差模式狀態視窗細節」

當您選擇「分類器」和「誤差模式」(或「僅用於回歸器」)，「狀態視窗」顯示：

- 用於將資料分為訓練和檢驗集的隨機子。
- 用於產生模型的記錄數目。
- 用於評估模型的記錄數目。
- 所預測的正確和不正確數目。
- 對於分類器，平均標準化均方差代表了機率估計的準確性。對於每個測試記錄，均方差是指 1 減去正確標籤值機率估計的平方，再加上其它標籤值 (不正確) 機率估計平方和的結果。標準化均方差是平均方根，是一個介於 0 和 1 之間的數。平均標準化平方根是在實驗集中利用適當的權重進行平均的標準化平方差 (加權平均)。
- 對分類器而言，分類誤差為不正確預計的百分比。
- 對於分類器，平均均方差和分類誤差表示均值的標準差以及均值的置信區間。如果資料來自於相同的分布，這就是一個可以從分類器得到的範圍。對於錯誤估計 (不是損失)，就會使用比通常的雙標準差規則更準確的公式。

估計誤差模式狀態視窗細節

當您已選擇「估計誤差」模式時，該「狀態」視窗包含有關分類器的下列資訊：

- 交叉驗證的次數和倍數。
- 隨機子。
- 標準差的估計準確性。
- 估計準確度的 95% 置信度間隔。

以下顯示有關回歸程式的資訊：

- 估計平方差誤差和估計平均絕對誤差。
- 有關上述準確度矩陣的 95% 置信度間隔。

設定特殊選項

- 總是忽略類型陣列的屬性。
- 日期被視為字串。除非有少量的日期，否則因為離散屬性的限制，這樣的屬性通常被忽略。在執行導入工具之前，您應該將日期分組。

國際化

在 Windows 系統中，可以經由桌面「控制面板」的「區域設定」元件來設定地區。該部可探討在 IRIX 系統中設定地區的問題。

MineSet 支援國際化資料集。圖形介面中的本文標籤是英文，但是您可以檢視用對應於資料編碼的語言表示的多位元欄名和資料值。只要安裝相對應的語言產品，MineSet 就會自動支援日語、中文和韓語的 EUC 編碼。有關其它語言和編碼的資訊，請參閱[第 105 頁的「擴展到其它的語言和編碼 \(僅適用 IRIX\)。擴展到其它的語言和編碼 \(僅適用 IRIX\)。」](#)。

在 IRIX 系統中設定地區

客戶端和伺服器系統中必須有您正在使用語言的地區和字型，以及任何用於遠程顯示的系統。要看到系統中安裝的地區清單，在 shell 提示符下輸入下列指令：

地區 -a

要設定地區，將環境變數 LANG 設定為合適的地區，而地區是從上述指令所產生的清單中選擇出來。例如，要將地區設定為 Japanese，EUC 編碼，請使用 csh，輸入下列指令：

```
setenv LANG ja_JP.EUC
```

然後從同樣的 shell 中啟動 MineSet。要為所有的應用程式永久地設定地區，請參閱 IRIX 文件。

擴展到其它的語言和編碼 (僅適用 IRIX)。擴展到其它的語言和編碼 (僅適用 IRIX)。

要讓 MineSet 在安裝過程所包含的地區外執行，需將資源檔案複製到合適的目錄中去並進行修改。MineSet 可視化工具以 2D 和 3D 字體使用「Open Inventor」。為使文字正常顯示，您必須安裝類型 III (一般稱為 CID 輪廓) 字體。

對於下列地區安裝過程中需包含資源檔案：

- ja_JP.EUC
- ko_KR.euc
- zh_CN.ugb
- zh_TW.ucns

在地區 *locale_name* 中執行 MineSet (請參閱第 105 頁的「在 IRIX 系統中設定地區」中如何列出安裝的地區)：

1. 照常安裝 MineSet。
2. 以 root 登錄。

3. 從 `/usr/lib/X11/app-defaults` 複製下列資源檔案至 `/usr/lib/X11/locale_name/app-defaults` 中：
- `clusterviz`

• `dtableviz`

• `Eviviz`

• `.mapviz`

• `MineSet`

• `.scatterviz`

• `.splatviz`

• `Statviz`

• `.treeviz`
4. 在 `/usr/lib/X11/locale_name/app-defaults` 中編輯資源檔案。您需要知道資源名稱以及您想用字體的特別要求（請參閱表 1-11 中的樣例）。
5. 在 `locale_name` 設定地區並啟動 `MineSet`。

樣例 1-1 為「韓語」的資源檔案變化

為「韓語」所需做的變化在表 1-11 中。列出的字體來自下列檔案的清單中：

- `/usr/lib/X11/fonts/ps2xlf_d_map.korean`

• `/usr/lib/X11/fonts/ps2xlf_d_map.korean.outline`。

表 1-11 「韓語」字型資源

檔案	「英語」資源（一些行被包圍起來）	「韓語」資源（一些行被包圍起來）
clusterviz Statviz	titleFont:screen12	titleFont:screen12， -ksg-mj-medium-r-normal--14-130-75-75-c-140-ksc5601.1987-0
clusterviz Statviz	gradationsFont:screen11	gradationsFont:screen11， -ksg-mj-medium-r-normal--12-110-75-75-c-120-ksc5601.1987-0
clusterviz Statviz	balloonFont:screen11	balloonFont:screen11， -ksg-mj-medium-r-normal--12-110-75-75-c-120-ksc5601.1987-0
clusterviz Statviz	xFontEncoding:ISO8859-1	xFontEncoding:ksc5601.1987-0

表 1-11 「韓語」字型資源

檔案	「英語」資源 (一些行被包圍起來)	「韓語」資源 (一些行被包圍起來)
dtableviz Eviviz , Mapviz , Scatterviz, Splatviz, Treeviz	myDefaultFont:Helvetica-Narrow	myDefaultFont:Helvetica-Narrow ; Gungso-Regular--KSC-H
MineSet	zoom2*fontList:--*-medium-r--*-6--* _****_*	zoom2*fontList:--*-medium-r--*-6--*-*-*-*-*-* ; -ksg--*-medium--*-12-: _****_*
	zoom3*fontList:--*-medium-r--*-8--* _****_*	zoom3*fontList:--*-medium-r--*-8--*-*-*-*-*-* ; -ksg--*-medium--*-12-: _****_*
	zoom4*fontList:--*-medium-r--*-10--* _****_*	zoom4*fontList:--*-medium-r--*-10--*-*-*-*-*-* ; -ksg--*-medium--*-14-:
	zoom5*fontList:--*-medium-r--*-12--* _****_*	zoom5*fontList:--*-medium-r--*-12--*-*-*-*-*-* ; -ksg--*-medium--*-14-:
	zoom6*fontList:--*-medium-r--*-14--* _****_*	zoom6*fontList:--*-medium-r--*-14--*-*-*-*-*-* ; -ksg--*-medium--*-18-:
	zoom7*fontList:--*-medium-r--*-16--* _****_*	zoom7*fontList:--*-medium-r--*-16--*-*-*-*-*-* ; -ksg--*-medium--*-24-:
	zoom8*fontList:--*-medium-r--*-24--* _****_*	zoom8*fontList:--*-medium-r--*-24--*-*-*-*-*-* ; -ksg--*-medium--*-24-:

重複 k- 平均值

重複 k- 平均值是一種由 MineSet 聚類工具所使用的方法。該方法將在用戶選擇的範圍內自動選擇聚類的數目。該算法產生幾個侯選聚類結果並利用離散度矩陣組合和用戶可選擇點選擇其中最好的一個方式來選擇聚類的數目。

拉普拉斯校正

「證據可視化工具」利用一個導入過程，計數（或權重）由此被用於計算機率。拉普拉斯校正可避免極端機率（0 或 1）。

要尋找拉普拉斯校正功能，從「工具管理員」的「挖掘工具」中按一下「分類」索引標籤並選擇「導入工具：」>「證據」。當您按一下「進階選項」按鈕，在產生的對話方塊中會顯示「證據選項」下的拉普拉斯核取方塊。

在一個樣例中使用拉普拉斯校正，在該例中我們並不將機率 1 分配給下述的事件，即「測出病人的 AIDS 為陽性就得了絕症」，而是機率接近 1，允許存在誤差或未被代表的樣例。拉普拉斯校正會使「證據可視化工具」機率偏向平均值，這樣就可避免極端值 0 和 1。

這就意味著「證據主視窗」中的每個圖對於每個類都有非 0 的切片。組中的記錄越少，傾向平均值的變化就越大。如果選擇拉普拉斯校正，並且因素為空或設為 0 時，則應用因素「1/ 訓練集權重」的啟發式過程，就可以進行自動拉普拉斯校正。

學習曲線

作為由大量記錄所建立的模型的一個功能，學習曲線是用於顯示導入工具產生的模型誤差的一個圖形。一般情況下，產生模型所用的記錄越多則誤差越低。

學習曲線是經由為曲線上的每個點產生指定數目的模型來建立。每個模型用隨機採樣的記錄產生，而其餘的記錄（那些沒有用於訓練的紀錄）用來估計誤差。

產生學習曲線將會耗費大量的計算時間。如果 t_i 是在訓練集 i 上訓練導入工具的次數（ i 的範圍從 1 到所有點的數目），並且對每個點執行 k 次，總次數為

$$k * \sum_i t_i$$

增加每點執行次數會成比例地增加執行時間，但會改善對均值的估計。執行次數的預設值為 3。

「分散可視化工具」篩選面板可用於篩選一些所顯示的資料類型 (平均點、置信度間隔、插值點或實際軌跡)。例如，您可能想要移除軌跡和置信區的資料點，只保留平均和插值的點。

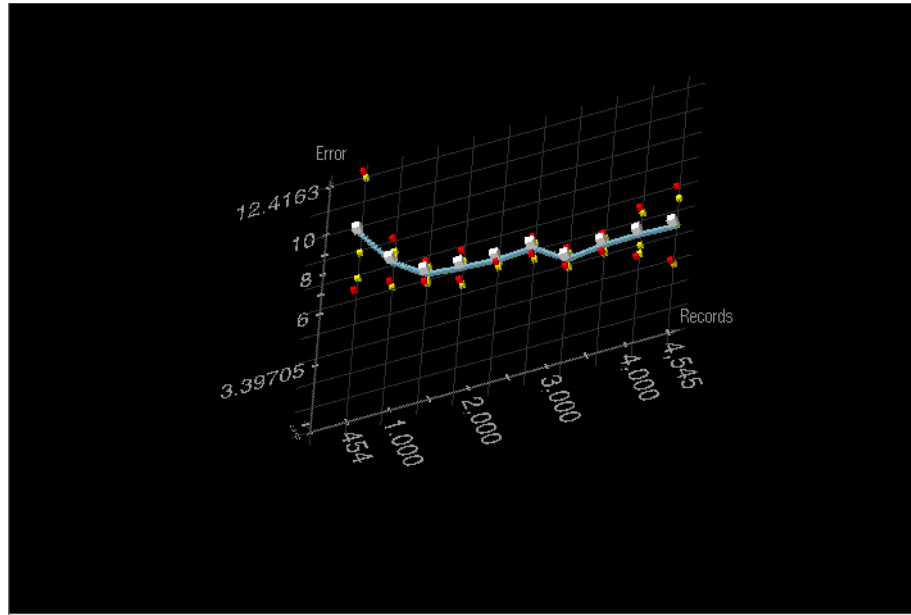


圖 1-20 學習曲線

「學習曲線」是「挖掘工具」索引標籤的「分類」功能表 (或「回歸」功能表) 中的一種模式。任何導入工具都可以使用它。當選取「學習曲線」模式時，*進階選項*對話方塊 (或*更多選項*對話方塊) 允許您在對話方塊的右側指定*學習曲線*選項，包括：

- 學習曲線中點數，
- 每個點的執行次數，
- 在起始點和結束點所用記錄的數量。

每個中間點所用的記錄數目將自動計算。

必須指定學習曲線中點的數目；它也必須大於或等於 1。可以指定起始和結束點的記錄數目以允許為指定範圍的訓練集產生學習曲線。如果這些選項為空，它們將根據學習曲線上的點數目和訓練集中的記錄數目自動的計算。該預設將覆蓋整個範圍的訓練集。例如，假設檔案包含了 80,000 筆記錄。如果您在學習曲線上指定 3 點，算法會在 20,000，40,000 和 60,000 個記錄上產生點。通常在較小的範圍內可用於放大。例如，可以在 1000 到 10,000 個記錄範圍內產生學習曲線。

上升曲線

上升曲線畫出記錄的隨機排序與在描述特殊標籤值時由分類器所建立的順序之間的差異。例如，您建立了一個模型來預計甚麼樣的客戶將要「客戶波動」，而現在想要在其波動之前找出這些客戶。上升曲線有助於完成此目標。

上升曲線是這樣的圖形，X 軸顯示記錄的數量從 0 到 100%，而 Y 軸顯示與那些擁有指定標籤值的用戶記錄的數目 (Churn=yes 在本例中)。在圖形 [圖 1-21](#) 中顯示兩條曲線。

較低的曲線 (紅色) 顯示在記錄隨機排序的情況下，波動用戶數量的期望。上面的曲線 (白色) 顯示根據每個記錄的分類器評價 (機率估計) 順序排列時，波動用戶的數量。那些代表分類器所識別最有可能波動的用戶之記錄最早出現；而較不可能波動的最後出現。通過分類器曲線和隨機曲線之間的差異可以看出分類器排序所表現出的優勢。

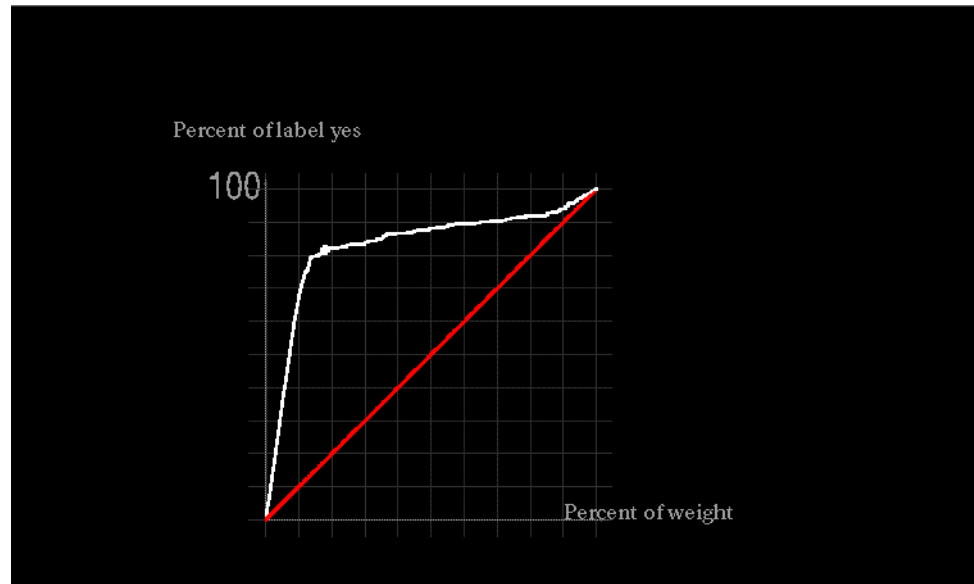


圖 1-21 上升曲線

在建立該上升曲線的過程中，將選取的分類器應用於該檢驗集。資料集的指定部分用於訓練，然後在資料集剩下的部分執行導入分類器。當使用「分類器和誤差估計」模式時，顯示上升曲線選項是進階選項(或更多選項)下的一個核取方塊，並適用於所有導入工具。「上升曲線」需要一個選定的標籤值。產生上升曲線並顯示該標籤值。

損失矩陣

損失矩陣的目的就是控制分類器可能產生的錯誤類型。在許多情況下，一些錯誤比其他的更致命。一個好例子來自於蘑菇資料集，其中蘑菇分為可食用或有毒的。將可食用蘑菇錯誤分類為有毒會損失一些錢(未吃蘑菇的價格)。將有毒的蘑菇錯誤分類為可食用將會損失上千倍的費用(住院的費用)。分類器利用損失矩陣來避免這些成本昂貴的錯誤。

損失矩陣和混淆矩陣結合使用最有效果。混淆矩陣辨識由分類器所產生錯誤的類型和數量。如果混淆矩陣指出分類器產生大量成本昂貴的錯誤，則在損失矩陣中賦予昂貴錯誤較高的權重也許是改進分類器表現的最佳途徑。可用「進階選項」(IRIX 系統中的更多選項)下的「使用損失矩陣」核取方塊來啟動損失矩陣。對於所有的分類導入工具都可用。

在下面例子中展示的過程：圖 1-22 顯示在只有 0.1 (10%) 的比例用於訓練集時，使用「決策樹導入工具」建立的蘑菇資料集的一個混淆矩陣。

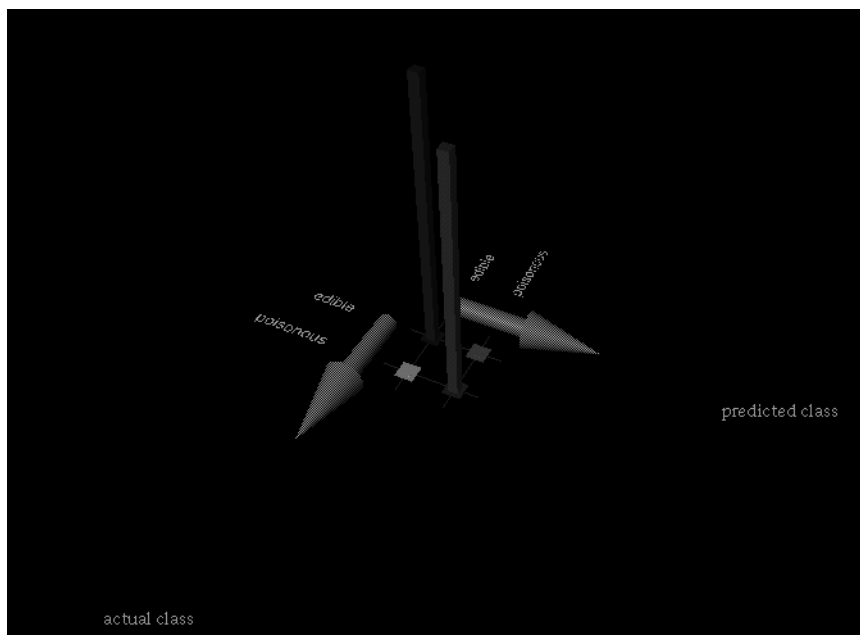


圖 1-22 使用預設的設定值為蘑菇資料集建立的混淆矩陣

代表有毒蘑菇的 8 筆記錄劃分為可食用 (0.1%)；代表可食用蘑菇的 15 筆記錄劃分為有毒 (0.2%)；3793 個可食用的蘑菇和 3496 個有毒蘑菇都已正確的分類。當分類器的錯誤率僅為 0.31% (小於百分之一) 時，我們的估計損失將為 $\$10000 * 8 + \$2 * 15 = \$80,030$ 或更糟。

圖 1-23 顯示同樣資料集的混淆矩陣，但這是在使用損失矩陣代表上述費用的條件下執行「決策樹導入工具」之後生成的。新分類器非常保守，沒有將有毒蘑菇劃分為可食用蘑菇；但是它卻將 1558 (1543+8) 個可食用蘑菇劃分為有毒蘑菇。估計的損失總和是這樣： $\$10000 * 0 + \$2 * 1558 = \$3116$ ，只有前述費用的 3%。

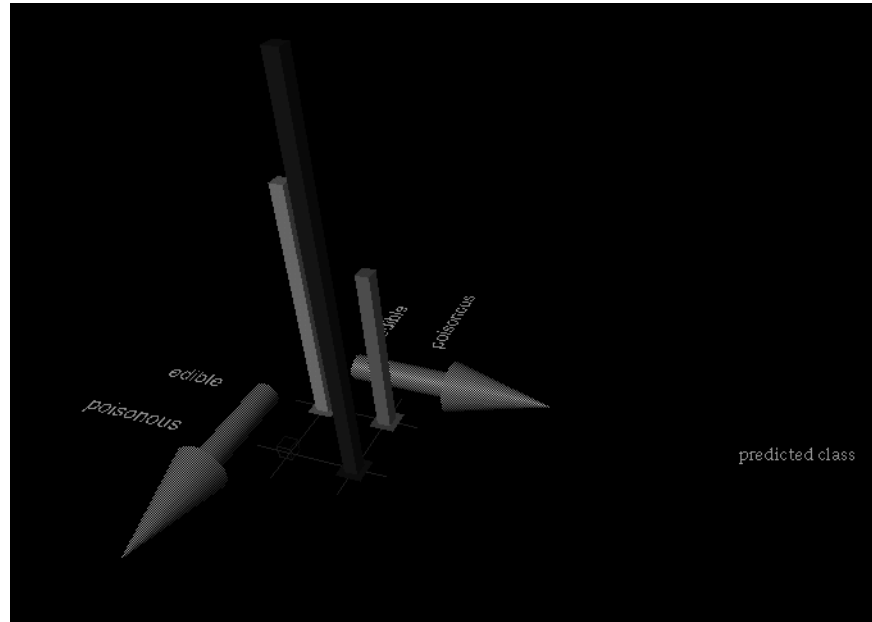
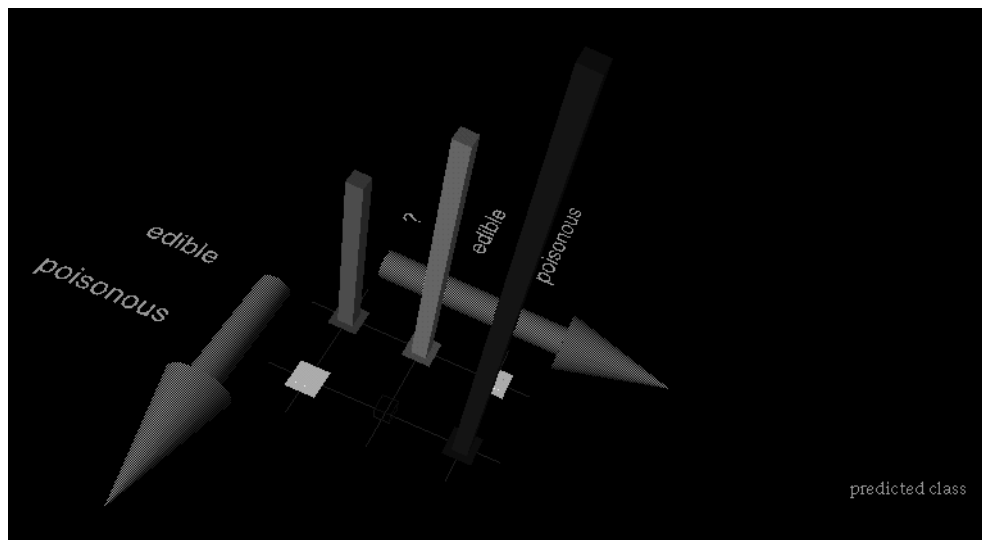


圖 1-23 帶有損失矩陣的蘑菇資料集之混淆矩陣

損失矩陣也有助於預計未知（空值）值，該值以問號顯示（?）。例如，假設必須花費 \$1，請教外面的專家來判斷蘑菇是否有毒還是可食用的。那樣的話，一些分類過程將導致未知預計。執行「決策樹導入工具」產生的混淆矩陣顯示在圖 1-24 中，其中有 1551 個為未知，只有 15 個可食用的蘑菇分類為有毒。整個的費用為 $\$10000 * 0 + \$1 * 1551 + \$2 * 15 = \1581 。



■ 1-24 帶有損失矩陣的蘑菇資料集混淆矩陣允許未知的預計

請注意：對於決策樹，損失矩陣的基礎是在樹葉處所做的機率估計。至於可靠估計：

1. 利用「決策樹」和「選項樹」中的「更多選項」將「拆分下限」從預設值增加到更高的值 (例如：5)。一般來說，訓練集越大噪音越多，該值就應該越高。
2. 使用大量訓練集。您也許需要大訓練集而得到更可靠的估計。
3. 使用選項樹。它們通常提供更好的機率估計並試圖降低損失。例如，執行上述蘑菇範例，將 \$10000 改變為 \$100 並且不允許出現未知值，則決策樹產生的估計損失為 \$1464，而選項樹產生的估計損失為 \$662。

使用損失矩陣選項是進階選項 (或更多選項) 下的核取方塊，它適用於所有分類過程的導入工具。編輯矩陣按鈕可用於定義損失矩陣。為避免預計產生未知值，在未知預計欄中填入矩陣中的最大值。

如果在「證據可視化工具」中指定「損失矩陣」，就會在**進階選項**(或 IRIX 系統中的**更多導入工具選項**)中的機率圓餅圖的右下方出現一個標示著**使用損失矩陣**的按鈕。當選取時(預設情況下)，「損失矩陣」會用於調節機率的顯示。最大的切片所顯示的為考慮「損失矩陣」的分類。要看到不使用「損失矩陣」時的機率狀態，取消選擇**使用損失矩陣**按鈕。當使用「損失矩陣」，就會出現一條灰色的切片，這是因為當您編輯損失矩陣時就會產生一個用於預計空值的欄。如果灰切片是最大的切片，則分類器預計為空值。

地圖可視化工具

「地圖可視化工具」向您顯示在 3D 場景中以長條圖表現的資料。該工具在建立與空間相關的可視化過程中有用，而在資料具有地理意義時尤其有用。

資料項目與可視場景中的長條圖物件相關。然而，這些物件具有可識別的形狀和位置。場景由一些這樣的物件組成，每個都具有獨立的高度和顏色(請參閱圖 1-25)。除了可以在場景中動態的進行瀏覽外，您還可以利用概化上尋和細化下尋來進行概括或觀察更高的精細度，以及使用動畫來觀察資料沿一維或二維參變量的變化。

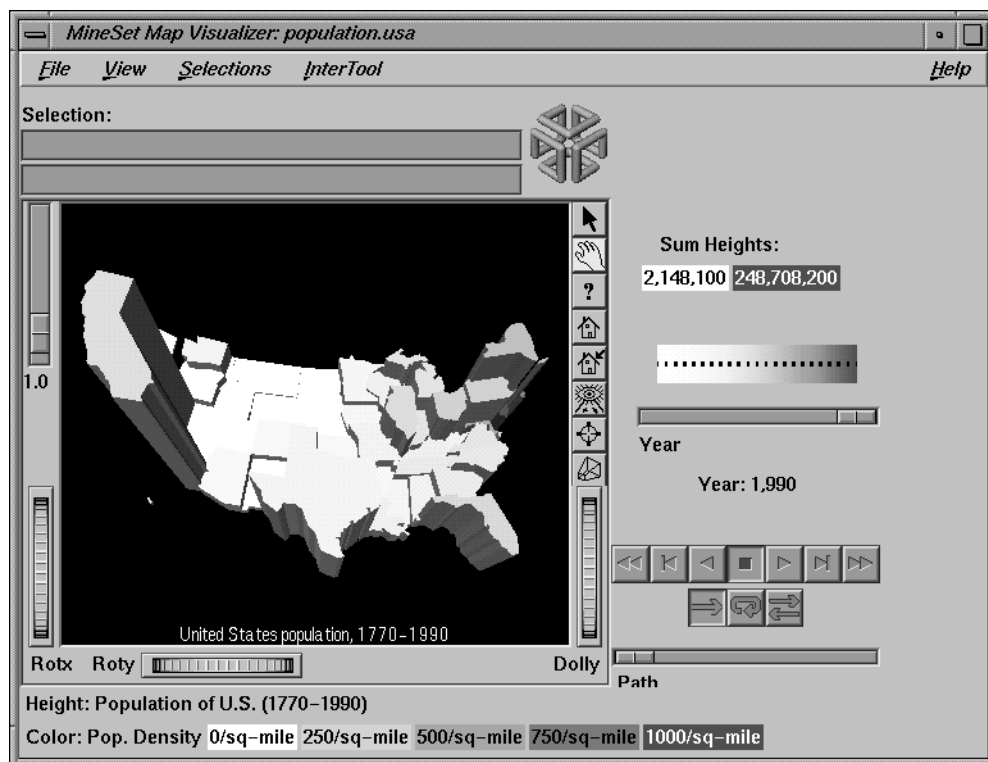


圖 1-25 地圖可視化工具樣例顯示 1990 年美國的人口分布。

場景也可以由這些地理物件輪廓的光滑平面組成，並在指定的位置產生「長條圖」。

另一種場景圖是由一些端點位於指定位置的線組成，這些線都具有獨立的寬度和顏色。線條是具有寬度和顏色屬性，而不具有高度和顏色屬性的任意形狀的物件和柱形實體。

地圖可視化工具的檔案需求

「地圖可視化工具」需要資料、gfx、等級和配置檔案：

- 資料檔案由索引標籤分隔的欄位列組成。一般情況下，「工具管理員」建立該檔案。您可以不用「工具管理員」產生該檔案（對於所需的檔案格式，請參閱 *《MineSet 3.0 企業版介面指南》* 中「為地圖可視化工具建立資料、配置、等級和 GFX 檔案」）。

資料檔案是從來源（例如，Oracle、INFORMIX 或 Sybase 資料庫）中提取原始資料的結果，然後將其轉化為「地圖可視化工具」能夠使用的格式。資料檔案有使用者定義的附檔名（「地圖可視化工具」所提供的樣例檔案有 .data 附檔名）。

- gfx 檔案由形狀描述和將要顯示的 1、2 或 3 維物件的位置所組成。

Gfx 必須有一個 .gfx 附檔名。MineSet 包含不同的 .gfx 檔案，包括按照州、電話區域代碼和郵政代碼等不同標準來劃分的美國地圖以及按省來劃分的加拿大地圖。您也可以手動產生 .gfx 檔案（請參閱 *《MineSet 3.0 企業版介面指南》* 中「為地圖可視化工具建立所需格式的資料、配置、等級和 GFX 檔案」）。

- 等級檔案由描述組成
 - 所顯示的不同圖形物件的欄名
 - 描述圖形物件形狀和位置的 .gfx 檔案的檔名
 - 一種對圖形物件等級關係的可選性說明，可讓您看到更多或更少的細節，稱為細化下尋和概化上尋。

等級檔案可以定義州和包含多個州的區域間之關係，並按照州或區域的層次來顯示人口總數。例如，`gfx_files/usa.state.gfx` 檔案描述美國 50 州的形狀；

`gfx_files/usa.state.hierarchy` 檔案描述將州劃分為區域，區域劃分為東西部以及東西部合成完整的美國這種等級關係。

如需進一步資訊，請參閱 *《MineSet 3.0 企業版介面指南》*，為「地圖可視化工具」建立「資料」、「配置」、「等級」以及「GFX」檔案。

- 配置檔案描述輸入資料的格式以及如何顯示它們。一般情況下，利用「工具管理員」建立該檔案。您也可以利用自己喜歡的文字編輯器（例如，WordPad、jot、vi 或 Emacs）而不是「工具管理員」來產生該檔案（請參閱《*MineSet 3.0 企業版介面指南*》中「為地圖可視化工具建立資料、配置、等級及 GFX 檔案」。

為了在「檔案」下拉功能表中選擇「開啓」時能被列出，配置檔案應具有 *.mapviz* 附檔名。當啓動「地圖可視化工具」或開啓一個檔案時，需指定配置檔案而不是資料檔案。

啟動地圖可視化工具

有幾種方法可以啓動「地圖可視化工具」：

- 利用「工具管理員」來配置和啓動「地圖可視化工具」。請先參閱《*MineSet 3.0 for Windows 企業版使用者指南*》中有關「工具管理員」大部分的功能之詳細內容，這些資訊適用於所有的 MineSet 工具。
- 如果您知道要利用的配置檔案，按一下該配置檔案的圖示。這樣就啓動「地圖可視化工具」並自動載入指定的配置檔案。這只有當配置檔案以 *.mapviz* 結尾時才有效（對於利用「工具管理員」為「地圖可視化工具」建立的配置檔案來說都是這樣）。
- 要從 IRIX 指令行中啓動「地圖可視化工具」，可以輸入：

```
mapviz [configFile]
```

configFile 是可選的，指定所使用的配置檔案的名稱。如果未指定配置檔案，您必須用「檔案」>「開啓」來指定一個。

啟動地圖可視化工具的選項

在 IRIX 系統中，您可以利用 **-quiet** 選項來刪除表示進度的跳現式對話方塊。加入一行可以使該選項永久可用

```
*minesetQuiet:TRUE
```

到您的 *.Xdefaults* 檔案。

您也可以設定一個警告執行語句，請參閱第 199 頁的「警告選項」項目。Windows 使用者可以從「檔案」>「特性」功能表中設定這些選項。

利用工具管理員來配置地圖可視化工具

該部分描述如何利用「工具管理員」來配置「地圖可視化工具」。雖然「工具管理員」大大地簡化配置「地圖可視化工具」的任務，但是您仍可使用文字編輯器為該工具手動建立一個配置檔案（請參閱《*MineSet 3.0 企業版介面指南*》中「為地圖可視化工具建立資料、配置、等級及 GFX 檔案」）。

產生 .gfx 和 .hierarchy 檔案

要使用「地圖可視化工具」，您必須為應用提供定義顯示圖形物件的兩個檔案：

- 一個或多個 .gfx 檔案，該檔案定義所要顯示的圖形物件的形狀。
- .hierarchy 檔案，描述多個互有關聯的地圖 (.gfx) 檔案之間的關係。

這些檔案不是由「工具管理員」所建立；它們必須已經是 MineSet 的一部分或由使用者建立。Windows 使用者可以在 MineSet 所安裝的 `\mapviz.gfx_files` 檔案目錄下找到它們。IRIX 使用者可以在 `/usr/lib/MineSet/mapviz/gfx_files` 目錄找到它們。如果您決定自己建立，可以在《*MineSet 3.0 企業版介面指南*》中的「為地圖可視化工具建立資料、配置、等級和 GFX 檔案」找到說明。

.gfx 和 .hierarchy 檔案是 MineSet 套裝軟體的一部分，它們包括

- 美國個州
- 美國個縣
- 美國 5 位郵政代碼所代表的區域
- 美國電話區域碼所代表的區域
- 加拿大個省及領土
- 墨西哥個州
- 澳大利亞的個州及領土
- 歐洲中西部個國家
- 法國和荷蘭的地區

「地圖可視化工具」需要的資料檔案具有

- 表示地理物件的一欄 (例如：州)。欄中的每一列必須指定唯一的地理物件 (在該例中，意味每個州對應一列)。
- 至少有一個數字型值的欄對照為 (利用數學表達式) 每個地理長條圖的高度和 / 或顏色。這些欄可以是標量、1D 陣列或 2D 陣列。如果該欄為陣列，必須使用滑動桿指定資料點以對照到高度和顏色。

如果高度和顏色對照為 1D 或 2D 陣列，這些陣列必須具有相同的索引 (請參閱 *《MineSet 3.0 企業版介面指南》* 中的「為地圖可視化工具建立資料、配置、等級及 GFX 檔案」)。

建立滑動桿和動畫

請參閱第 10 頁的「動畫」。

地圖可視化工具選項

按一下 **工具選項** 按鈕將顯示一個對話方塊，該對話方塊可允許您對「地圖可視化工具」中的一些選項預設值進行改變。

下面描述「地圖可視化工具」選項對話方塊中的按鈕和欄位。

地理位置

實體檔案 指定在「地圖可視化工具」主視窗中用於代表地理「實體」物件的 *.hierarchy* 檔案。

輪廓檔案 指定所要繪製物件的輪廓，該輪廓顯示為光滑平面，其上將放置 3-D 實體物件。

尋找檔案 按鈕可瀏覽檔案，並尋找所要使用的 *.hierarchy* 檔案。

實體檔案 和 **輪廓檔案** 欄位是可選的。如果不提供「實體檔案」，「地圖可視化工具」所建立的圖形實體物件將由畫面中任意形狀和隨意放置的簡單矩形組成。

高度

該部分指定最初高度 *比例值* (預設為 1.0) 以及是否在「地圖可視化工具」視窗的底部顯示高度圖例。

顏色

要使用這些「顏色」選項，您必須將欄對照為「資料目標」面板所需的 * *顏色 - 長條圖*。要得到如何選擇和改變顏色的詳細說明，請參閱第 48 頁的「顏色選擇」。

顏色清單—您可以利用顏色清單標籤旁的 + 按鈕來指定顏色清單。這會產生一個顏色編輯器，您可以利用它指定要加入清單的顏色。

對照—您可以指定在圖形顯示中所表現的顏色變化是 *連續* 或 *離散的*。如果您選擇 *連續*，作為在對照欄位中對照到顏色的值的一個功能，顏色值在顏色清單中所輸入的顏色間逐漸變化。

跳現式按鈕右邊的欄位可讓您輸入顏色所對照的指定值。在該欄位中值的數目必須與從「使用的顏色清單」欄位中輸入的顏色數目相同。

您可以按照需要在該欄位中輸入多種顏色用於顯示。如果欄中對照為 * 顏色 - 長條圖的顏色數目超過您所選擇的顏色的數目，「地圖可視化工具」將在執行期間加入適量的隨機抽取的顏色。如需選擇顏色的進一步資訊，請參閱第 48 頁的「顏色選擇」。

圖例開關—可讓您決定要顯示或隱藏顏色圖例。

標準化開關—可讓您確定「地圖可視化工具」是否在顏色欄的最大和最小值間自動按比例變化（稱為顏色標準化），正好與您手動指定臨界值相反。當**標準化開啓**可用時，臨界值必須在 0 到 100 之間，代表顏色欄中最大值到最小值之間範圍的百分數。

滑動桿

請參閱第 154 頁的「為地圖可視化、分散可視化、平板可視化建立的滑動桿」。

訊息欄位

該選項允許您指定當選取實體時所顯示的訊息。如需這個欄位中，可以使用的格式類型清單和說明，請參閱《*MineSet 3.0 企業版介面指南*》裡的「為地圖可視化工具建立資料、配置、等級和 GFX 檔案」中消息語句部分。

可讓您指定顯示在地圖可視化工具主視窗底部的字串。該字串必須包含在雙引號內。

執行欄位

該選項可讓您輸入在連接兩下一個實體時所執行的指令。格式與訊息語句一樣。如果不出現執行語句，連接兩下也不會起作用。

如需「執行」欄位的詳細說明，請參閱《*MineSet 3.0 企業版介面指南*》中的為「地圖可視化工具建立資料、配置、等級和 GFX 檔案」。

重設工具選項

在對「工具選項」對話方塊作出修改後，如果您想要將所有的設定重設為預設值，按一下 **重設** 按鈕。

接受工具選項

一旦您完成對「工具選項」對話方塊的修改，按一下 **確定** 將返回「工具管理員」主視窗。

地圖可視化工具檔案設定

「工具管理員」將「地圖可視化工具」的資訊儲存在幾個檔案中，並具有相同的字首：

- `<prefix>.mapviz.data` 包含資料。
- `<prefix>.mapviz.schema` 描述資料檔案。
- `<prefix>.mapviz` 包含地圖可視化工具所需的資訊。
- `<prefix>.mineset` 包含建立其它檔案所需的所有資訊。

要指定字首，使用「工具管理員」主視窗中「檔案」功能表的功能表選項 **儲存...**。如果您不指定字首，則它將根據資料來源的類型自行確定。

當您使用 **啟動工具** 按鈕，如果需要時將更新 `.data`、`.schema` 和 `.mapviz` 檔案。

挖掘工具索引標籤

從「工具管理員」中的「資料目標」窗格中按一下「發掘工具」索引標籤可得到下列索引標籤：

- 關聯，關聯規則
- 聚類，聚類
- 按照以下模式分類：僅用於分類器、分類器和誤差估計、估計誤差以及學習曲線。這些索引標籤中可用的導入工具是：「決策樹」、「選項樹」、「証據」和「決策表」。

- 按照回歸模式進行回歸：回歸器和誤差估計、僅用於回歸器、估計誤差以及學習曲線
- 欄重要性，欄重要性
- 任何插件，例如 ACpro

多重選擇

在大多數的工具中，利用 **Shift** 按一下滑鼠左鍵就可以完成選擇。如果在物件上按一下滑鼠左鍵的同時沒有按下 **Shift** 鍵，則在選擇游標下面物件的同時也取消了所有先前選擇的物件。按一下滑鼠左鍵時按下 **Shift** 鍵可以切換對該物件的選擇而不影響其它選擇。（「平板可視化」工具有不同的介面，可在「平板可視化工具」項目中找到其說明。）

當您選擇一項時，一個描述該項的訊息會出現在工具的主視窗中，在預設情況下，可視化工具只顯示最後選擇物件的訊息。分離的記錄查看視窗顯示一個表格，該表顯示所有選擇的值。在「樹」和「地圖可視化工具」中，選擇「選項」>「顯示值」來觀察該表。

如果已經為特殊的工具設定訊息，該消息也會出現在表中。拖曳欄之間的分離符來重新調整表格中欄的規模。您也可以按下一個值並在表格的頂部顯示該值的完整文字。

公共資訊

「公共資訊」是一個「決策樹」拆分標準，該標準控制著分類器進行決策的路線。「公共資訊」是指父節點和子節點純度加權平均之間純度的變化（也就是 Δ ）。加權平均是根據每個子節點上記錄的數目來進行的。請參閱第 71 頁的「決策樹」可得到其它拆分標準的討論。

簡單 -Bayes 方法

「證據導入工具」有時稱為 Naive-Bayes 或簡易 Bayes。它建立一個模型，該模型假定在已知分類的條件下，屬性值機率獨立。例如，對於每個蝴蝶花分類（藍色蝴蝶花、多色蝴蝶花和純白色蝴蝶花）來說，擁有四個屬性（花萼長度、花萼寬度、花瓣長度以及花瓣寬度）的 *iris* 資料集是獨立的。當然這個簡化的模型並不真實，但這個模型用於最初的資料開發卻非常有效並且它的分類預計表現在實際應用過程中也很好。

用視窗邊界工具在非樹狀可視化工具中瀏覽。

該部分由三個表組成，並可用於「證據」、「決策表」、「地圖」、「分散」以及「平板可視化工具」瀏覽控制的快速搜尋。[表 1-12](#) 描述瀏覽按鈕。

表 1-12 在非樹狀可視化工具中的瀏覽按鈕

按鈕	名稱	動作
	選取	改變程式至選取模式（箭頭）。在選取模式中，您可以反白標示（掠過）或選擇（按一下）圖中的元件。
	抓取	改變程式至抓取模式（手形）。抓取模式中，您可以在視窗中移動圖象： 若要旋轉圖表，按下滑鼠左鍵並移動滑鼠。 若要在視窗中移動圖表，按下滑鼠的左鍵和右鍵（如果系統配置的是三鍵滑鼠，可以使用滑鼠中鍵）並移動滑鼠。
	首頁	將圖返回至設計為首頁檢視的位置。當第一次啓動可視化工具時，這是起始位置。您可以設定主檢視按鈕來改變主檢視位置。
	設定主檢視	為圖設定新的主檢視。當您想儲存某個檢視或位置時使用。
	全部檢視	將圖移至視窗的中心位置，在視窗中便能看到全部的圖。
	搜尋對象	將您選擇的點移至窗格的中央並對其縮放。當滑鼠的游標變為目標形狀時，將它移到您想要更清楚檢視的一點，然後按下滑鼠左鍵不放（或利用滑鼠中鍵，如果您的系統配置三鍵滑鼠）。
	3D 透視模式	切換 3D 遠景。
	最上層檢視	將圖表變為最上層檢視（僅適用「分散」和「平板可視化工具」）。
	前檢視	改變圖表為前檢視（僅適用「分散」和「平板可視化工具」）。
	側檢視	改變圖表為側檢視（僅適用「分散」和「平板可視化工具」）。

表 1-13 描述在非樹狀可視化工具中的調節滑動桿和滑動輪。

表 1-13 在非樹狀可視化工具中調節滑動桿和滑動輪。

滑動桿或滑動輪	動作
高度滑動桿 (左上)	升高或降低塊、餅或長條圖的高度來強調差異。
細節滑動桿 (只適用於証據和決策表可視化工具)	篩選出不重要的屬性。
% 權重臨界值滑動桿 (只適用於「証據」和「決策表可視化工具」)	篩選出記錄權重小於資料集中，記錄權重總和指定百分數的屬性，一共為 2%。
Rotx 輪	繞 X 軸旋轉。
Roty 輪	繞 Y 軸旋轉。
縮放輪	將圖放大或縮小。

表 1-14 列出在非樹狀可視化工具中，可對圖表進行的幾項操作。

表 1-14 控制非樹狀可視化工具場景		
動作	滑動桿或滑動輪	滑鼠或鍵盤動作
在「選擇」和「抓取」模式中切換	N/A	按下 Esc 鍵或瀏覽按鈕。
移動畫面	N/A	抓取模式中，按一下並按住滑鼠右鍵。向您想要移動圖的方向移動游標。
升高或降低塊、餅或長條圖的高度來強調差異。	高度滑動桿（左上）	N/A
將畫面繞 X 軸旋轉。	Rotx 輪	抓取模式中，按一下並按住滑鼠左鍵。向您想要移動圖的方向移動游標。
將畫面繞 Y 軸旋轉。	Roty 輪	抓取模式中，按一下並按住滑鼠左鍵。向您想要旋轉圖的方向移動游標。
將畫面放大或縮小。	調節輪	抓取模式中，按一下並按住滑鼠左鍵（或滑鼠中鍵）。滑鼠向下移用於放大，向上移用於縮小。
在細節層中細化下尋 （只適用於決策表和地圖可視化工具）	N/A	將滑鼠箭頭放在指定的圖上（或所有圖的背景中）並按一下滑鼠右鍵。
在細節層中概化上尋 （只適用於決策表和地圖可視化工具）	N/A	將滑鼠箭頭放在指定的圖上（或所有圖的背景中）並在按下 Ctrl 時按一下滑鼠右鍵（或按一下滑鼠中鍵）。

用視窗控件在樹狀可視化工具中瀏覽。

「樹狀可視化工具」顯示是很好方法，它好像是經由照相機來檢視畫面。要改變檢視，您須改變照相機的位置（視點）。該部分由兩個表組成，並可用於「樹」、「決策樹」、「選項樹」以及「回歸樹狀可視化工具」控制的快速搜尋。[表 1-15](#) 描述瀏覽按鈕。

表 1-15 樹狀可視化工具中的瀏覽圖示

圖示	動作
	將圖返回至設計為首頁檢視的位置。當第一次啟動可視化工具時，這是起使位置。您可以使用下一個圖示來改變主檢視位置。
	為圖設定新的主檢視。用該功能來儲存檢視或位置。
	將圖移至視窗的中心位置，在視窗中便能看到全部的圖。
	恢復以前的移動（象「網上瀏覽器」中「後退」按鈕一樣）。
	重做已復原的移動（像 Web 瀏覽器上的「前進」按鈕一樣）。
	把節點移向樹根。
	從一個節點或長條圖移向左邊。
	從一個節點或長條圖移向右邊。
	沿著節點左邊路徑向下移動。
	沿著節點右邊路徑向下移動。
	從目前節點中跳現出一個可選路徑的功能表。

表 1-16 描述在樹狀可視化工具中的調節滑動桿和滑動輪。

表 1-16 在樹狀可視化工具中調節滑動桿和滑動輪。

滑動桿或滑動輪	動作
高度滑動桿 (左上)	升高或降低長條圖高度來強調差異。
H 輪	上下移動視點
傾斜輪	改變照相機的上下傾角
左右移動輪 (<-->)	將視點左右移動
調節輪	將照相機前後移動

名稱排序功能表

「證據可視化工具」和「決策表」上的「名稱排序」功能表可控制標稱 (命名的) 屬性值如何排序，並提供這些選項：

- 按字母排序使具有標稱值的屬性從左到右 (或從上到下) 按字母順序進行排序。
- 加權值從左到右進行排序，那些具有最大權重的記錄在左邊。
- 「標籤機率」 (預設情況下) 使標稱屬性值按照份量的大小進行排序，該份量對應其中一個分類。如果標籤是分組的屬性，則預設情況下會使用最高的組。如果標籤是標稱的，則使用先驗機率餅圖中占有最大份額的分類做為預設。如果選擇特殊分類，並且隨後就必須按照標籤機率進行排序，則選取的分類用於確定次序。在所有情況下，如果存在「空值」，它就為第一個值。

標準化公共資訊

「決策樹」工具使用「標準化公共資訊」作為預設拆分標準。從技術上，這是「公共資訊」除以子節點數目的以 2 為底的對數。請參閱第 71 頁的「決策樹」中關於「決策樹」拆分標準的討論。

空

一些可視化工具，例如，「平板可視化工具」以及「地圖可視化工具」，當含有未知資料值或空值的欄位對照為可視屬性時，便使用特殊的代表方式。(關於空值的討論，請參閱《*MineSet 3.0 企業版介面指南*》的「**MineSet 中的空值**」一章。當分組中對照至顏色的欄中每個記錄都有空值時，該平板過程的結果顏色為灰色。如果組合中的一個或多個記錄為對照到顏色的欄並具有非空值時，該 (或那些) 值則用於計算顏色。一個值與空值的和為空值，一個值和空值的平均就為該值 (也就是說， $\text{value} + \text{Null} = \text{Null}$ ； $\text{avg}(\text{val}, \text{Null}) = \text{val}$)。

當空值顯示在不同的可視化工具顯示視窗區域時：「選取視窗」、「選擇視窗」或「游標位於」區域，它顯示為問號 (?)。

對於含有空值 (對照於座標軸) 的數字欄，在座標軸定義的範圍之下有一個特殊的空值位置。這有助於顯示該空值與其它值不連續。利用「檢視」功能表中的「顯示空值位置」選項，可以關閉數字坐標軸的空值位置。至於對照到坐標軸的字串值欄，空值 (由 ? 代表) 被視為另一個值。

在「地圖可視化工具」中，當以下為真時會出現空值：

- 資料庫或資料檔案包含空值。
- 根據分組，「工具管理員」用於產生陣列並且沒有資料落入特殊的分組中。例如，30-40 歲的人口沒有資料，該組為空。
- 「工具管理員」用於產生陣列並指定空值枚舉選項。在此種情況下，對於組值為空的情況，建立多餘的陣列元件代表所有值的組合。「工具管理員」將問號 (?) 字元指派至特殊的分組。要查看該組的值，移動相應的滑動桿至最左邊位置。如果空組中沒有資料，與它相關的值也為空，並且「地圖可視化工具」也將相應的圖形物件表示為「空」物件。
- 空值的表達式和組合產生空值 (請參閱 **MineSet 中的空值**)。
- 當空值對照為可視屬性時，「地圖可視化工具」使用特殊的代表方式。空高度導致 0 高度的深灰物件；空顏色則導致具有適當高度 (由對照為高度的值來定義) 的物件，但顏色為深灰。

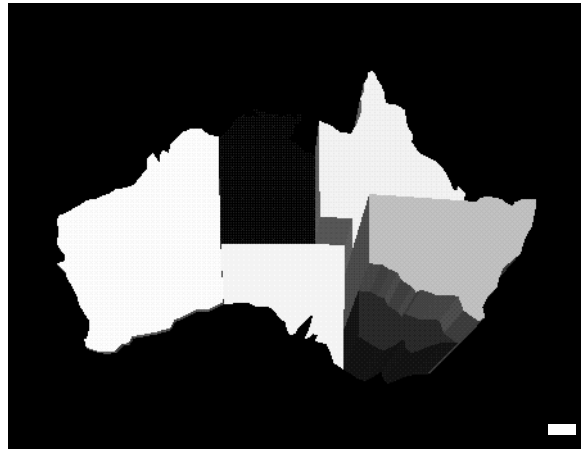


圖 1-26 對照到高度 (中上部的物件) 和對照到顏色 (右下部的物件) 的空值之方式

當選擇具有空值的物件時，問號 (?) 會顯示在選擇欄位中。在「地圖可視化工具」中對照的空值顯示在圖 1-26 中，其中空值對照為「高度」和「顏色」。

選項樹

「選項樹」是預測模型。使用獨立或已知的屬性值來進行預測，以幫助確定標籤值或未知屬性。「選項樹」為每個記錄指派一個類別，以開始資料的分類過程。用於分類的結構是決策樹，在第 71 頁的「決策樹」中有描述。「選項樹」以選項節點擴展一般的「決策樹」。「選項節點」顯示樹中決策節點處可供選擇的幾個選項。「選項樹」一般比普通的決策樹更大更複雜。

導入選項樹

從資料中自動導入 (產生) 「選項樹」模型。由記錄和一個與每個記錄相關的標籤所組成的資料稱為訓練集。

檔案需求

「選項樹」導入工具需要訓練集。從資料來源 (例如，Oracle、INFORMIX 或 Sybase 資料庫中的 MineSet ASCII 檔案、二元檔案或表格) 中抽取資料可以產生檔案。要應用產生

的分類器，您應該擁有一個具有分類器所需屬性的記錄資料集，即除了標籤屬性外具有所有其它訓練集中屬性的記錄資料集。

建立選項樹導入工具

從資料中自動導入（產生）「選項樹」分類工具。這個過程從登錄到伺服器並以平常的方式選擇一個資料集開始。

從「工具管理員」中選擇 *分類索引* 標籤，並且從 *導入工具* 彈出式功能表中選擇「選項樹」。除非您希望，否則您不必做更多的限定。只需按一下 *繼續按鈕*。「選項樹」使用「樹狀可視化工具」為其顯示。

IRIX 中的並行過程

如果您已經安裝 MineSet 的多處理器版本，當分支包含超過 1000 條以上的記錄時，在「決策樹」中可以並行處理基於樹的算法。（請參閱第 134 頁的「在 IRIX 系統中的「並行計算」」。）您可以改變「工具管理員」中「特性」面板的並行模式來控制線程的數目（請參閱「檔案功能表」）。只有在 IRIX 系統中該選項才可用。

選項樹選項

「選擇進階選項」（IRIX 上為「更多導入工具選項」）可產生「導入工具選項」對話方塊。該對話方塊由四個面板組成：

- 最上層的面板表示在「工具管理員」的「資料目標」面板中所做的選擇。
- 從上面數來第二個面板可讓您設定損失矩陣和加權屬性。請參閱第 111 頁的「損失矩陣」和第 138 頁的「記錄加權」。

- 左下角的面板可讓您指定「進階導入工具」選項(說明如下)。
- 右下角的面板可讓您指定「錯誤估計選項」(除非在「資料目標面板」中選擇的是僅用於分類器模式,在這種情況下此區域內是空的)。該面板中顯示的選項取決於您所選擇的「錯誤估計」類型(請參閱第 80 頁的「錯誤估計」)。

要微調「選項樹」導入工具算法,您可以選擇在「決策樹」中的任何選項,這些選項在第 71 頁的「決策樹」中描述。除此之外,還提供下面的選項。

- **最大根選項數**

該整數預設為 5,限制在根上建立的最大選項數目。該導入工具不允許出現最大數目的選項,因為其他的屬性可能級別更低。

- **減少值**

該整數預設為 2,定義在每一層中最大選項數目降低的量。最大根選項數的預設值為 5,這就說明決策樹節點的第二層至多有三個選項 ($5-2=3$)。而決策樹節點的第三層將限制為只有一個選項 ($3-2=1$)。再向下的層次都一樣限制為單個選項。

- **最小適合比率**

該比率確定何時將屬性作為選項排除。當導入工具給每個屬性一個適合得分時,則導入工具選擇最好的屬性以及其它良好並可作為選項的屬性。適合比率確定哪些選項應達到何種程度才能被選取。比率值 f 意味著侯選選項,屬性評分必須至少達到 $(1-f)*b$,其中 b 為最佳屬性的比數。適合比率為 1 則選擇所有的屬性(所以,如果有在其上拆分的屬性,就說明該屬性滿足了上述的限制選項)。適合比率為 0,就會建立「決策樹」(無選項節點)。預設值為 0.9。

導入「選項樹」所需的時間與所建立的選項節點數目有密切關係。因為選項節點一般都在頂部建立(對於可理解性和限制誤差來說最有效果),對導入「選項樹」所需時間的準確估計是所建立沒有子選項的選項數目乘以建立「決策樹」的時間。在預設設定下,根節點可以擁有超過五個的選項,並且每個子節點可以擁有超過三個的選項。則選項總數可達 15 (3 乘以 5)。如果最大根選項數增至 6,則選項數目的限制為 48 ($6*4*2$);如果增至 7,則選項數目的限制為 105 ($7*5*3$)。保持最大根選項數為 5,但將減少量改為 1,則選項的限制為 120 ($5*4*3*2$)。因此上例中所期望的導入時間將比普通「決策樹」所需的時間長兩個數量單位。降低最小擬合比選項通常導致比限制因素少的選項,因此減少了導入時間。

在 IRIX 系統中的「並行計算」

IRIX 系統的 MineSet 多處理器版本提供了並行處理。它允許將計算密集的任务以並行的方式執行。要在 64 位元版本上執行，需安裝並行伺服器並從「工具管理員」的「特性」功能表中選擇並行版本。

「資料移動」是在伺服器上執行的一個過程，提供對儲存在檔案中的資料庫和資料進行存取的功能，並為挖掘和可視化工具轉換資料。

IRIX 6.4 及以上的版本支援大型記憶體 (64 位元) 模式。如果您有 IRIX 6.2，您仍然可使用 32 位元的資料挖掘工具，但是您必須更新至 IRIX 6.5 以得到 64 位元的支援和 p 線程功能。要完全利用 64 位元定址，您也需要改變 **systune** 資源參數，該參數取決於您的系統配置。

systune 參數確定可用系統資源的預設限制。[表 1-17](#) 列出 IRIX 所推薦的 **systune** 參數值 (如需詳細內容，請參閱 **systune(1M)** 參考頁)。

表 1-17 **systune** 參數

參數	定義	推薦值
rlimit_pthread_cur	線程數的目前限制	1024
rlimit_rss_cur	記憶體使用的目前限制	機器上的實體記憶體數量
rlimit_vmem_cur	虛擬記憶體使用的目前限制	邏輯交換空間的大小或實體記憶體的兩倍
rlimit_nofile_cur	開啟檔案數的目前限制	1024 或線程數目的限制

注意：在設定新參數後，您必須重新啟動您的機器。

如果您已經安裝 MineSet 的多處理器版本，當分支包含超過 1000 筆以上的記錄時，可以並行處理基於樹的算法。樹上的每個節點可以估計在相應層上最有可能的拆分，這些任務可以並行處理。預設情況下，一個程式所能容納線程的最大數量將自動確定。您可以改變「工具管理員」中「特性」面板的並行模式來控制線程的數目。並行過程會造成記憶體碎裂化，會使可用於並行計算的最大資料集的規模小於在單處理器中計算的最大資料集的規模。

預測度

當用關聯規則進行工作時，如果您用左側 (LHS) 和右側 (RHS) 作為兩個坐標軸將柵格可視化，則預測度描述了在所有滿足 LHS 條件的記錄中同時滿足 RHS 條件的記錄所佔的比率，或者是普遍度除以 LHS 項目出現的頻率 (請參閱第 28 頁的「關聯規則可視化」可得到 RHS 和 LHS 的說明)。它指出 X 和 Y 一起出現的頻率，是所有出現 X 的記錄的一部分。例如，如果預測度為 50%，那麼在所有 X 出現的記錄中，X 和 Y 將一起出現的機率為 50%。因此，如果您知道 X 出現在一個記錄中，您可以預計在該記錄中 Y 有 50% 的機會出現。例如，如果記錄顯示了在人們買嬰兒尿布和煙之間相關性中的可預測度為 50%，那麼在買嬰兒尿布的人中，將會有 50% 的人也買煙。

普遍度

當用關聯規則進行工作時，如果您用左側 (LHS) 和右側 (RHS) 作為兩個坐標軸將柵格可視化，則普遍度說明了在 RHS 和 LHS 中一起出現的項目之頻率 (請參閱第 28 頁的「關聯規則可視化」可得到 RHS 和 LHS 的說明)。普遍度定量說明了全部記錄當中，檔案中 X 和 Y 一起出現的部分。例如，如果常用 (預置) 為 1%，X 和 Y 一起出現的記錄占總數的 1%。

刪改

修改是「決策樹」導入選項，其改變將影響樹的高度和「拆分下限」。預設的修改因素為 0.7。越高數目表示越多的修改。請參閱第 71 頁的「決策樹」可得到完整的討論。

隨機子

在 MineSet 中有幾個選項對話方塊可以指定隨機子。它是指產生算法的隨機數之起始點。然後，產生的隨機數用於資料集的採樣中。例如，每次採樣都使用相同的隨機子，就可確保每次在改善模型時都使用相同的資料選擇結果。如果您改變隨機子，將從基本資料集中得到不同的選擇結果。

記錄檢視器

「記錄檢視器」允許您直接查看您的資料。提供您一個熟悉它們資料值和欄的機會。「記錄檢視器」也允許您：

- 通過重置大小、重新排列或隱藏它們來進行操作
- 在任意的指定欄中根據值對資料進行排序或篩選
- 經由多欄進行排序 (只適用於 Windows)
- 在排序和篩選後對列進行重新編號
- 搜尋指定的值
- 將您的操作檔案以一定的格式儲存

啟動記錄檢視器

有幾種方法可以啟動記錄檢視器：

- 從「工具管理員」中：
 - 在「資料目標」面板中按一下「可視化工具」索引標籤。
 - 按一下「記錄」索引標籤 (Windows) 或從「工具」功能表 (IRIX) 中選擇「記錄檢視器」。
- 從「工具管理員可視化工具」功能表中，選擇「記錄檢視器」。然後，使用「記錄檢視器檔案」功能表來開啓檔案。
- 連按兩下 .schema 檔案圖示 (僅適用 Windows)。
- IRIX 使用者可以在介面的指令行提示下輸入該指令：

```
recordview [ filename ]
```

對列重新編號

「記錄檢視器」允許您在任意一點對列進行重新編號。如果您在排序和篩選之後執行該操作，就不能復原重新排序的操作。要回到原始資料，您必須重新開啓檔案。

要進行重新編號，選擇「檢視」>「重新對列編號」。

在「記錄檢視器」中尋找

「記錄檢視器」允許您在資料中尋找一個值。要開啓「搜尋」面板，選擇「檢視」>「搜尋」面板。要進行搜尋，請輸入值，反白標示您想要進行搜尋的欄，然後按一細化下尋找下一個或尋找上一個。

儲存資料

「記錄檢視器」允許您儲存資料，包括您對資料所做的任何更改。您可以利用「檔案」>「儲存」或「檔案」>「另存新檔」來儲存您的檔案。

如果您使用「儲存」，檔案將在原始的名稱和格式下進行儲存。如果這是您第一次儲存檔案，則它以 **MineSet** 二元格式進行儲存。「另存新檔」將帶出「儲存資料」視窗，您可以在其中輸入所需的檔案名稱及所希望的格式類型。

利用「另存新檔」功能，您可以將檔案按照四種格式進行儲存：二元、ASCII、HTML 或純文字格式。當您以二元或 ASCII 格式進行儲存時，同時也儲存了資料檔案和方案檔案。HTML 格式將檔案儲存為 HTML 表。本文格式將檔案以索引標籤分隔的形式進行儲存，欄的名稱為第一列的內容。

記錄加權

在一定的實驗設計中，真實人口部分比其它部分的採樣更頻繁。有時，您可能想要得到人口總數的 1% 樣本，那麼採樣總數 0.1% 的這一小部分只會產生 0.001% 的樣本，這就顯得太少（例如，您可能只得到了兩個人）。記錄加權讓您為每個記錄指定一個權重；因此採樣了兩次的人得到的權重為 0.5，而剩下人口的權重指定為 1。

另一個例子，電話公司將欺詐電話儲存在一個資料集中，而只存儲了非欺詐電話的一小部分。利用記錄加權，每個記錄都可指定為總量的真實部分。

最後，有些資料集已經組合，而且記錄中有相關的自然記數（例如，美國城市的統計資料通常擁有人口的相關記數）。該計數屬性可對照為權重，並複製按照計數量的記錄價值。

記錄加權的意義是權重為 2 的記錄與兩個權重為 1 的記錄相等。並且允許浮點型的權重。也請參閱第 200 頁的「加權」。

回歸索引標籤

回歸索引標籤允許您進入 **MineSet** 回歸工具。在「工具管理員」中的「資料目標」面板中，按一下「發掘工具」索引標籤來顯示「回歸」索引標籤。

回歸樹

回歸工具是一個執行回歸操作的可預測模型。在指定一套描述性屬性的條件下，回歸過程的任務是預測連續標籤值。回歸過程和分類過程相似，其區別是在分類過程中，預計的標籤只具有很少量的離散值。

產生回歸工具時，MineSet 也產生可視化過程。該可視化過程有助於您理解回歸工具以及它是如何進行預測的。除此之外，它還能夠為資料本身提供有價值的直觀結果。一旦產生回歸工具，就可為未標記的記錄預測標籤值。

導入「回歸樹」

使用稱為導入工具的算法後，就可以從資料自動導入（產生）「回歸樹」回歸工具。如果使用除了「僅用於回歸器」外的任何選擇，資料都會拆分而形成訓練集。訓練集由資料庫中的記錄組成，而且資料庫中的連續標籤也已知。例如，您可以提供帶有欄（例如，年齡、教育、職業、每周工作時數等等）的表格，其中有一欄包含說明性的屬性（收入總量）。登錄到伺服器並以平常的方式選擇一個資料集後便可開始。

Windows 使用者：

從「工具管理員」中，在「發掘工具」下選擇回歸索引標籤。

IRIX 使用者：

從工具管理員中選擇回歸索引標籤，導入工具功能表自動顯示了回歸樹。

除非您希望，否則您不必做更多的限定。除了標籤值為連續的而不是離散的以外，該選擇與「決策樹」的完全一樣。只需按一下啟動工具或繼續按鈕。「回歸樹」使用「樹狀可視化工具」為其顯示。

連續標籤

「連續標籤」功能表提供可能連續標籤的清單。該清單包括帶有數字型值的屬性。選擇您想要進行模型化操作的標籤屬性。例如，如要產生預測總收入的回歸程式，請選擇「總收入」。如果沒有連續屬性，功能表顯示沒有連續標籤，並且繼續按鈕也停用。只能為連續屬性產生回歸工具。如果資料集中沒有包含連續屬性，您可以利用「工具管理員資料轉換」面板增加一個新的連續欄。

回歸樹選項

選擇進階選項 (IRIX 中的更多導入工具選項) 可產生「回歸選項」對話方塊。該對話方塊由四個面板組成：

- 最上層的面板表示在「工具管理員」的「資料目標」面板中所做的選擇。
- 從上面數來第二個面板可指定損失矩陣，請參閱第 111 頁的「損失矩陣」，並設定加權屬性，請參閱第 200 頁的「加權」。
- 左下角的面板可讓您指定其它導入工具選項 (說明如下)。
- 右下角的面板可讓您指定「錯誤估計選項」(除非在「資料目標面板」中選擇的是僅用「僅用於回歸器」模式，在這種情況下此區域內是空的)。該面板中顯示的選項取決於您所選擇的錯誤估計類型 (也可參閱第 80 頁的「錯誤估計」)。

要微調「回歸樹」導入算法，您可以改變下列「回歸樹」導入工具選項。

- 經由 ... 限制樹高度

預設情況下，在回歸樹中高度 (層的數目) 並無限制。按一下核取方塊並輸入上限值以限制高度。限制層的數目可以加快導入過程，並且不需要分解太多的節點就可以研究「回歸樹」。雖然限制規模會降低執行時間，但會增加錯誤率。設定該選項並不影響在最大層之前所選的屬性。

- 拆分標準

在樹導入過程中，該選項允許您指定在競爭屬性拆分中用於選擇的標準。至於回歸樹，MineSet 支持四種拆分標準：

- 方差

這會造成「回歸樹導入工具」選擇那些能使樹中每點處的節點間方差最小的拆分。當產生葉子時，在葉子處的預測是符合葉子的記錄標籤值的平均。方差是資料集中每個值之間的平方差，除以成員或值的總數。該模式是統計中最常用的。

- 絕對偏差

這會造成導入工具選擇那些能使樹中每點處的節點間絕對偏差最小的拆分。當產生葉子時，在葉子處的預測是符合葉子的記錄標籤值的中值。

- 標準化方差

就像方差，這是一個拆分標準，但對於多重拆分會產生偏差。

- 標準化絕對偏差

就像絕對偏差，這是一個拆分標準，但對於多重拆分會產生偏差。

對於已知的問題，很難說哪種標準是最好的。對每個都進行試驗，並產生最低錯誤估計或形成最容易理解的「回歸樹」。

- 拆分下限

這是一個權重的下限（或者，如果權重沒有設定，就為記錄的數目），且至少必須存在於兩個子節點中。該選項的預設值為 5。例如，如果節點中有一個三向拆分，三個子節點中至少有兩個其權重必須為 5 或更多（如果沒有設定權重則為兩筆記錄或更多）。這就提供限制「回歸樹」規模的另一種方法。

提高拆分下限是為增加機率估計的可靠度，因為每個葉子的記錄數較大。這也會建立更小的樹並減少導入時間。如果您期望資料中包含噪音（錯誤或異常），則提高拆分下限。如果資料集很小（< 100 筆記錄），您可以降低拆分下限。

- 成本複雜性修改

成本複雜性的修改經由互換樹的誤差率（其成本）和樹上葉子的數目（其複雜性）試圖產生形狀最佳化的樹。在成本複雜度修改過程中，訓練集拆分為學習集和修改集。學習集用來成長一個修改樹。該樹修改後產生一系列具有遞減複雜性的樹。然後使用修改集在序列中找出成本最小的樹。標出最小成本樹的大小。學習和修改集合併並長成一棵樹。然後該樹修改為最小成本樹的大小。

成本複雜度修改參數允許您選擇比最小成本樹更小的樹。該參數指出標準差的值比可以接受的最小成本樹所花費的成本更大。將該參數設為 0 以選擇最小成本樹；設定參數為 0.5 以選擇尺寸最小的樹，其誤差率不大於 0.5，標準差比最小成本樹差。預設的設定為 0，在誤差率不大於標準誤差率但又不差於最小成本樹的條件下，選擇最小規模的樹。越高數目表示越多的修改。如果資料可能包含噪音（誤差和異常），增加該值來建立更小的樹。如果樹修改成為一個節點，降低值以降低修改量並顯示更多樹的結構。

修改過程比限制樹高度或增加拆分下限過程的速度要慢，這是因為要建立整個樹然後再進行修改。然而，有選擇的進行修改將形成更準確的回歸器。

回歸程序中的錯誤估計

當評估分類器時，自然矩陣有誤差（分類器預測錯誤標籤樣例的數目）。當提供損失矩陣時，不同類型的錯誤分類率會有不同的相關成本。在這種情況下，損失是一個自然度量。

至於回歸過程，其任務是預測一個實數值，該值沒有自然評估矩陣。經常使用的兩種度量是均值和誤差以及平均絕對誤差。在「平均差」中，使用預測標籤值和實際標籤值之間的平均方差。在平均絕對誤差中，使用預測標籤值和實際標籤值之間差的絕對值的平均。

回歸器名稱

產生的回歸器以階段檔案 (在「工具管理員」中確定) 的字首命名，而結尾為 *-rt.regress*，例如 *churn-rt.regress*。預設情況下，所有的回歸工具儲存在伺服器的 *file_cache* 目錄中，在 Windows 中預設為 *MineSet Files*，在 IRIX 中預設為 *mineset_files*。Windows 系統在目前工作目錄中將可視化過程儲存為 3D 長條圖圖的圖示。這些檔案有 *-rt.treeviz* 結尾。

這些回歸器可為未標記的資料集預測標籤。要應用已儲存的回歸器，您需要根據回歸器所使用的屬性選擇記錄資料集。該回歸器將為每個記錄預測新的連續標籤值。請參閱第 15 頁的「應用模型」和 第 33 頁的「修正」。

移除欄

當您需要簡化可視化過程或模型時，您可以從目前工作中移除欄，不用將它們從資料集中移除。在「工具管理員」的「資料轉換窗格」中，從「目前」欄文字窗格中選擇一欄並按一下 *刪除欄* 按鈕。這樣就從您的計算中，但並未從資料集中移除該欄。

投資回報曲線

投資回報 (ROI) 曲線與上升曲線相似，但它以損失的形式而不是以誤差的形式顯示準確度；並將所使用的損失矩陣考慮在內。

可以從任何導入工具的「進階選項」面板中的「工具管理員」對 ROI 曲線進行配置，只需使用「錯誤估計選項」部分下的核取方塊。在回歸工具中此功能無法使用。

ROI 曲線中按照每個記錄的期望損失進行排序的每個點，也就是那些由所選標籤值進行標記的點。同樣地，曲線中每個點的高度指出累計利潤 (與損失相反)，而不是該點以上所有記錄的累計準確度 (與誤差相反)。

在所選的標籤欄條件下（請參閱第 111 頁的「損失矩陣」）「損失矩陣」中的項目乘以分配給由分類器產生的分類機率可以計算期望損失。因此，如果分類器對於其預測很確定，那麼預期損失就較低，而且記錄將出現在 ROI 曲線的左邊。

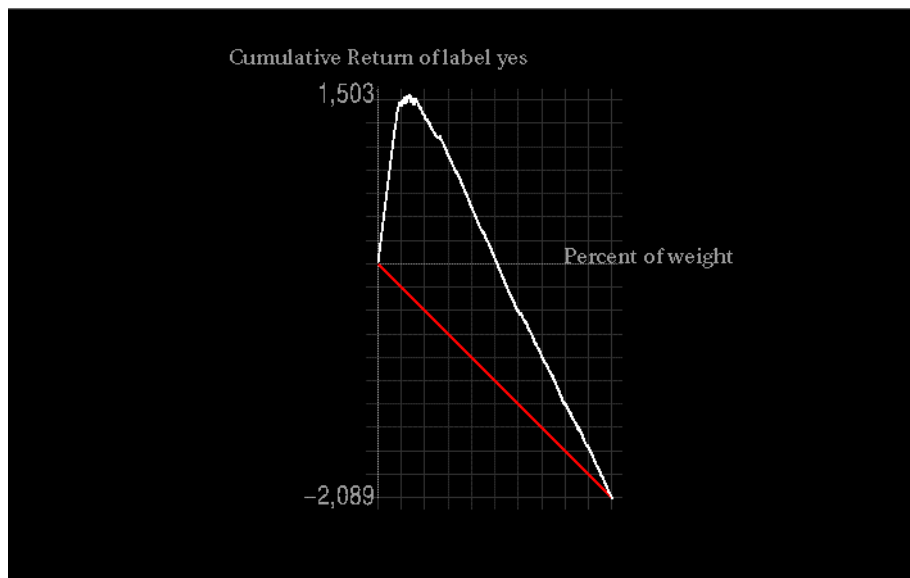


圖 1-27 投資回報曲線

ROI 曲線意義在於，用戶對於資料集中的每個記錄都應採取措施。該措施與所選擇的標籤值相關。例如，在客戶波動資料集中，與標籤是相關的措施，也許會向人們送去一些市場資料。這也許會阻止人們波動；但是如果不加區別的進行該行動代價是昂貴的。如果分類器用於預測是否向特殊的人傳送信函，則 ROI 曲線的頂點就顯示在檢驗集上必須保留大約多少錢。

當填入損失矩陣並與 ROI 曲線一起使用時，應特別小心。在一定預測標籤下的欄確定標籤值的結果 ROI 曲線。在所有可能的分類上，該欄中的項目代表採取與該標籤值相關的行動所得到的預期收益和損失。例如，客戶波動過程中，在「預計是」欄以及在「實際值為不」列下的成員可能包含值 2，以表示郵寄給將要客戶波動的人小冊子（該措施與「是」有關）的費用。另一方面，在「是」欄和「是」列下的成員可能以 -10 值來表示，如果能避免客戶客戶波動，公司將能省下郵寄費用 \$10。

儲存檔案

您可以儲存您正在工作的階段部分，方法是使用「工具管理員檔案」下拉式功能表，或按一下「工具管理員」中「資料目標」窗格的「資料檔案」索引標籤來儲存特殊轉換。當您想恢復工作時，從「工具管理員」中「檔案」下拉式功能表中載入該階段或檔案，則檔案和歷史表都在該處恢復。

樣例檔案目錄

MineSet 提供一系列樣例檔案供使用者使用。在前面的部分有進行詳細的說明。平台不同目錄也不同。

Windows 使用者可以在安裝 MineSet 的目錄下找到樣例檔案，在 `\examples` 下。

IRIX 使用者可以在 `/usr/lib/MineSet/examples` 中找到檔案。

分散可視化工具

利用「分散可視化工具」可直觀的分析幾個變數的關係，或從統計上或經由動畫過程。在沒有大量記錄 (少於 50,000) 的條件下，對於觀察單個資料點或將大量的記錄合併為不同組合結果的小集合時，它是很有用的。如果資料集有大量的記錄，則可考慮使用「平板可視化工具」。在「分散可視化工具」中進行分析可使用：

- 3D 場景
- 動畫控制面板包含一個 2D 滑動桿
- 圖形物件，稱為實體，在 3D 場景中可以進行動畫過程。在動畫過程中，實體的位置、顏色和大小都可以改變。

「分散可視化工具」將資料集中的每個記錄或列對照為 3D 場景中實體的方式來對資料進行可視化。資料中的變數可以對照為實體的大小、顏色以及位置。同樣地，您也可以將一個或兩個數字型變數對照到動畫控制面板中的滑動桿。如果對照為實體的大小、顏色和位置的變數取決於對照為滑動桿的變數，則用該滑動桿可執行動畫過程。例如，資料可代表幾個公司一段時間的銷售。如果時間對照為滑動桿而銷售變數對照為大小，就可以實現實體隨時間滑動桿增大或縮小的動畫。

在建立資料的可視化過程後，「分散可視化工具」可用不同的方法分析資料。您可利用動畫控制面板對動畫路徑進行 1D 或 2D 的追蹤。重新播放回所建立的動畫路徑，您可以觀察到實體大小、顏色和運動的趨勢或異常變化。在 3D 場景中，您可以對顯示結果定向來強調特殊的方向或角度。「分散可視化工具」可以對變數值進行放大以特別強調。同樣地，您也可以篩選顯示結果，只顯示那些符合標準的實體。

檔案需求

「分散可視化工具」需要下列檔案：

- 資料檔案由索引標籤分隔的欄位的列所組成。利用「工具管理員」可以很容易的建立這個檔案。如果您正在產生該檔案，請參閱 *《MineSet 3.0 企業版介面指南》* 中「為分散可視化工具建立的資料和配置檔案」尋找所需的檔案格式。

您可從資料來源（例如資料庫）中抽取資料來產生資料檔案，並將它按照「分散可視化工具」的需要進行格式轉化。資料檔案擁有使用者定義的附檔名（「分散可視化工具」工具中提供的檔案有 .data 附檔名）。

- 配置檔案描述輸入檔案的格式及如何進行顯示。「工具管理員」可以建立該檔案，或者您可利用自己喜歡的文字編輯器（例如，WordPad，jot，vi 或 Emacs）而不是工具管理員來產生該檔案（請參閱 *《MineSet 3.0 企業版介面指南》* 中的「為分散可視化工具建立資料和配置檔案」）。

配置檔案必須有 .scatterviz 附檔名。當啟動「分散可視化工具」或開啓檔案時，您必須指定配置檔案而不是資料檔案。

啟動分散可視化工具

有幾種方法可啟動「分散可視化工具」：

- 在「工具管理員」中的「可視化工具」索引標籤上執行「分散可視化工具」。請參閱第 147 頁的「配置「分散可視化工具」」可得到與「分散可視化工具」一起使用「工具管理員」的更多資訊。
- 用「工具管理員」從「可視化工具」功能表中啟動「3D 可視化工具」。(請參閱本書的「工具管理員」項目以及《*MineSet 3.0 for Windows 企業版使用者指南*》中關於「工具管理員」功能的部分，該部分對所有的 MineSet 工具都通用。)
- 如果您知道要使用的配置檔案，連按兩下該配置檔案的圖示。這樣就啟動「分散可視化工具」並自動載入您所指定的配置檔案。只有當配置檔名以 *.scatterviz* 結尾時才起作用 (對於用工具管理員、分散可視化工具為規則建立的配置檔案都是這樣)。
- 要從 IRIX 指令行中啟動「分散可視化工具」，可以輸入：

```
scatterviz [configFile]
```

configFile 是可選的，指定所使用的配置檔案的名稱。如果未指定配置檔案，您必須用「檔案」>「開啓」來指定一個。

配置「分散可視化工具」

- 在 Windows 系統中使用「工具管理員」：

在「工具管理員」的「資料目標」窗格中，按一下「可視化工具」索引標籤並選擇「分散」。「分散可視化工具」面板顯示您可以對照為欄的各種元件。在每個本文欄位右邊的跳現式功能表中選擇一欄。限制每個跳現式功能表中的可用選擇為那些合適類型的欄。

- 在 IRIX 系統中使用「工具管理員」

在「工具管理員」的「資料目標」窗格中，按一下「可視化工具」索引標籤；在「工具」跳現式功能表中選擇「分散可視化工具」。顯示星號的可視成員需要對照。從「目前欄」中選擇一欄並對照為右邊窗格中的元件。

- 用「文字編輯器」建立一個「配置檔案」

雖然「工具管理員」大大地簡化配置「分散可視化工具」的任務，但是您也可使用文字編輯器為該工具手動建立配置檔案。請參閱 *《MineSet 3.0 企業版介面指南》* 中「為分散可視化工具建立資料並配置檔案」。

為「分散可視化工具」建立滑動桿

當資料集中的欄單獨變化時，例如，當資料在一段期間內產生變化時，為「分散可視化工具」建立滑動桿是有用的。滑動桿是建立動畫的先決條件且可以手動或自動建立。請參閱 *《MineSet 3.0 for Windows 企業版使用者指南》* 中的「滑動桿建立」部分可得到更多的資訊。

分散可視化工具選項

要在「分散可視化工具」中改變不同的選項，返回「工具管理員」的「資料目標面板」，並選擇「可視化工具」索引標籤，然後選擇「分散可視化工具」。按一下 *工具選項* 按鈕可以產生用於顯示的新對話方塊。在這裡您可以改變「分散可視化工具」選項的預設值。

「分散可視化工具」對話方塊有四個基本選項塊：

- 實體
- 滑動桿
- 坐標軸
- 其他

實體選項

該選項組可讓您指定控制「分散可視化工具」圖形顯示外觀的實體的特徵：

- *實體圖例開關*可讓您決定要顯示或隱藏實體圖例。
- *實體尺寸*可讓您將實體按比例放大到最大尺寸、一定比例尺寸或預設尺寸（不調節）。您可以指定實體大小的圖例是顯示還是隱藏。
- *實體顏色* 可讓您控制實體所顯示的顏色。您可以：
 - 指定所用顏色清單
 - 指定對照類型
 - 將顏色清單對照為值清單
 - 指定顏色圖例是顯示還是隱藏
 - 對照顏色到實體
- *實體形狀*可讓您為實體選擇一個可視代表物：立方體、條形、球體或菱形。
- *實體標籤顏色*讓您在上面按一下來修改標籤的顏色。會讓「顏色選擇」對話方塊出現，利用它可以完成顏色的改變。
- *實體標籤尺寸*控制實體標籤的大小。較小的數字會減小規模，而較大的資料會增加規模。

要使用「顏色」選項，您必須將一欄對照為「資料目標」面板中的實體顏色。要得到如何選擇和改變顏色的詳細說明，請參閱第 48 頁的「顏色選擇」。

*顏色清單*可讓您利用顏色清單標籤旁的 + 按鈕來指定顏色清單。這會產生一個顏色編輯器，您可以利用它指定要加入清單的顏色。

*顏色對照*可讓您指定在圖形顯示中顏色的變化是*連續*或*離散*的。如果您選擇*連續*，作為在*顏色對照*欄位中對照到顏色的值的一個功能，顏色值在*顏色清單*欄位中輸入的顏色間逐漸地變化。

跳現式按鈕右邊的欄位可讓您指定對照該顏色的值。如果您不指定任何對照值，就會使用顏色變數值的範圍。請參閱第 48 頁的「顏色選擇」可得到關於選擇顏色的更多資訊。

匯總選項

匯總滑動桿允許在一或兩個附加變數上的進行動畫過程。滑動桿上的每個位置都有一個顏色，該顏色與對照為總和的變數組合值相對應。匯總選項可讓您指定在匯總視窗中顯示的變數應使用何種顏色。您也可以指定要顯示或隱藏匯總圖例，該匯總圖例是用於指示這些值。如需動畫的進一步說明，請參閱第 10 頁的「動畫」。

如果您有一個值的陣列，您可指定 X 或 Y 滑動桿。這些選項旁邊的跳現式按鈕提供可用關鍵字的清單，您可指定將哪個用做滑動桿。

滑動桿選項

「滑動桿」選項控制滑動桿對照的解釋過程。如需進一步資訊，請參閱第 148 頁的「為「分散可視化工具」建立滑動桿」。

坐標軸選項

坐標軸選項可讓您為每個坐標軸指定：

- 標籤 (如果該方塊為空，「分散可視化工具」就預設地使用每個坐標軸的欄名)。
- 顏色
- 每個坐標軸的規模尺寸 (可以為*最大尺寸*，*比例尺寸*，或*不調節*)。
 - *最大尺寸*可指定對於一個指定的尺寸坐標軸可以獨立地進行縮放。如果一個坐標軸具有最大尺寸，比另一個坐標軸大兩倍，則不管資料值的大小它都是另一個軸的兩倍長。當以不同的單位的坐標軸進行比較時，該選項是最有用的 (例如，收入對年齡)。對於非數字型的資料該選項不起作用。

- *比例尺寸*可根據其最大值的基礎對坐標軸進行縮放。如果兩個坐標軸具有相同的「比例規模」，但是其中一個具有最大值並且是另一個的兩倍，則前一個是後一個的兩倍長。當以相同的單位坐標軸比較時，該選項是非常有用的（例如，收入對支出）。該選項對於非資料型坐標軸不起作用。
- *不調節與比例尺寸*為 1.0 相同。
- 尺寸值
- 該坐標軸是否應延伸以包括 0 值。

其他選項

在對話方塊底部的「其他選項」，包含下列欄位：

- *訊息*允許您指定當選取實體時所要顯示的訊息。要列出和說明用於輸入欄位的格式型式，請參閱《*MineSet 3.0 企業版介面指南*》裡的為「分散可視化工具建立資料和配置檔案」中的「消息語句」部分。
- *執行*該選項可讓您輸入 UNIX 指令，該指令是在連按兩下一個實體時執行。格式與訊息語句一樣。如果不出現執行語句，連按兩下也不會起作用。如需執行欄位的詳細說明，請參閱《*MineSet 3.0 企業版介面指南*》中，為「分散可視化工具建立資料和配置檔案」中的「消息語句」部分。
- *隱藏標籤距離*控制實體標籤變為不可見的距離。較小的距離可以改善顯示的性能，但標籤很快就消失。資料越高，標籤隱藏的距離越大。
- *坐標軸標籤尺寸*控制坐標軸標籤的大小。較小的數字會減小尺寸，而較大的資料會增加尺寸。
- *柵格 (X, Y, Z) 尺寸*可指定各自坐標軸柵格線之間的時間隔。較小的數字會減小尺寸，而較大的資料會增加尺寸。如果指定 0，就不畫出柵格線。
- *柵格顏色*經由在顏色塊上面按一下可改變柵格的顏色。會讓「顏色選擇」對話方塊出現，利用它可以完成顏色的改變。

重設工具選項

如果想要將所有的選項恢復成預設值，按一下**重設**按鈕。

儲存新工具選項

一旦您完成對工具選項對話方塊的修改，按一下 **確定** 將返回「工具管理員」主視窗。

動畫控制面板

動畫控制面板，出現在主視窗的右邊，由最多帶有兩個滑動桿的匯總視窗、資訊欄位、動畫按鈕和動畫滑動桿組成。請參閱第 10 頁的「[動畫](#)」。

在分散可視化工具中的空值處理

當具有未知或空值的欄位對照為可視化屬性時，「分散可視化工具」會使用特殊的代表方式。(關於空值的討論，請參閱《*MineSet 3.0 企業版介面指南*》的「**MineSet 中的空值**」一章。) 當空值對照為實體尺寸時，該實體顯示成立方體的輪廓。當空值對照為實體的顏色時，實體顯示為深灰色。當空值顯示在「選擇視窗」或「游標位於」的區域時，它顯示為問號 (?)。

如果空值對照為實體的位置 x ， y ，或 z 時，則結果取決於「查看」功能表中的「顯示具有空位置」(請參閱第 197 頁的「[檢視功能表](#)」)。如果設定該選項，則坐標軸上帶有空位置的實體顯示在對應坐標軸範圍內。如果沒有設定該選項，就不顯示帶有空位置的實體。

樣例配置和資料檔案

我們提供樣例資料和配置檔案來展示「分散可視化工具」的特性和功能。這些檔案的進一步說明可參閱 [附錄 A](#)，「[配置和資料檔案樣例](#)」。

「選項」按鈕

選項按鈕對所有工具來說是相似的；不適合某種特殊工具的功能沒有顯示在功能表中。「選項」功能表可以對基本資料進行追溯。要執行追溯，首先選擇一或多個實體或平板，然後在兩種追溯方法中選擇一種對選定的記錄進行追溯。

- 建立選擇方塊建立 3-D 選擇器方塊中，可以伸展或轉換為選擇區域體積範圍。當選擇方塊啟動時，將會開啓一個「記錄檢視器」格式的表格，用以顯示所有選取實體所代表的組合資料的資訊。關閉視窗會移除選擇器方塊，而不取消它的選擇內容。選擇方塊中的任何實體或按下 **Shift** 來選擇的實體會顯示在表格視窗中。要改變選擇方塊，用滑鼠左鍵按一下它的一面，並將它按指定的方向拖曳。拖曳時按下 **Shift** 鍵會限制離移動方向最近的坐標軸運動。要改變選擇方塊的範圍，可在預定的方向內拖曳灰色標籤。在體積界限以外調整大小或移動是不允許的。灰色標籤自動調整大小來保持固定的螢幕大小。如果有時它們太大，您可以靠近選擇方塊，它們將根據方塊的大小減小它們的大小。
- 顯示值跳現出一個窗口顯示選取的實體。
- 顯示原始資料檢索並顯示與已選取實體對應的記錄。結果記錄顯示在表格檢視器中。如果沒有選取任何實體，將停用該選項（變灰）。
- 傳送至工具管理員根據目前（方塊）的選擇，在「工具管理員」歷史的起點處插入一個篩選操作。用於追溯的精確表達式由目前在主視窗中的選擇來確定。如果沒有選取任何實體，將停用該項目（變灰）。
- 互補追溯形成顯示原始資料和傳送到工具管理員選擇，當使用時，取得所有未選取的資料。

- *追溯欄*開啓一個面板，讓您選擇用於追溯的欄。不像其它的可視化工具，在資料中沒有特殊欄被指定為資料的關鍵字。可視化工具不可能確定在追溯表達式中使用者想要的欄。例如，您可能擁有品牌，模型以及重量等轎車資料。您可能想追溯這些原始資料，並確定需要考慮品牌和模型，而不考慮重量。預設情況下，所有已對照為圖形必備成分的欄在追溯中視為有意義的。其它的欄則不是如此，但可以在「追溯欄」對話方塊中將它們反白標示起來以達到此目標。

如需追溯的進一步說明，請參閱第 78 頁的「追溯」。

為地圖可視化、分散可視化、平板可視化建立的滑動桿

您可以將一個欄對照為滑動桿來顯示根據指定的標準值如何變化。如果欄是數字型的（整數、浮點型和雙精度型）或分組的就可以對照為滑動桿。如果欄已經分組，它在名稱後將帶有_bin。在「目前欄」欄位中欄的名稱之後標示欄的類型，例如：白天通話的總數 - 雙精度型。

在從「工具管理員」中「資料轉換」窗格的*目前欄*中，將數字型欄對照為一個或兩個滑動桿，您可以為「地圖」、「分散」以及「平板可視化工具」自動地建立滑動桿。請參閱第 13 頁的「動畫按鈕和滑動桿」。

排序欄名稱

您可以不改變資料集，按照字母順序對欄名進行排序以便於參照。從「工具管理員」中的「資料轉換」窗格中，按一下*按順序顯示各欄名稱*複選方塊 (Windows)，或 *排序欄名稱*按鈕 (IRIX)。

平板可視化工具

利用「平板可視化工具」可從統計角度或者透過動畫過程直觀的分析幾個變數的關係。對於具有大量記錄的資料集，它特別合適。對於量不太大的記錄，如果您想觀察個別資料點，可以選擇「分散可視化工具」。要進行資料分析，可使用

- 3D 場景
- 動畫控制面板包含一個 2D 滑動桿
- 稱為平板的圖形物件，代表資料點的組合。在動畫過程中可以改變平板的顏色和透明度（而不是位置或大小）。

「平板可視化工具」可以將欄對照為坐標軸、滑動桿、顏色和透明度來將資料可視化。所得到的 3D 場景結果可以視為與每個資料點都由分散畫出的分散圖相似。它並非真正的分散圖，因為聚在一起的資料點（落入相同的組內）已合併，並以單一的云團狀平板畫出。

每個對照到坐標軸或滑動桿的數字欄必須先進行分組。如果跳過該分組步驟，「工具管理員」會用自動統一分組執行該步驟（請參閱第 34 頁的「分組」）。字串欄可直接對照到坐標軸上。任何數字型欄都可對照到顏色。由於所有資料點都落入同一組使得該欄對照為顏色，而其值的平均可導出平板的顏色。平板的透明度根據同一組資料點數目的加權平均而定。如果沒有資料點對照為透明度，則記錄計數就用於確定透明度。結果可視化過程的交互性獨立於資料點數目所代表的意義；它只取決於在軸維度上分組的數目。

如果您的資料集很大，請在「工具管理員」中正確地進行組合。伺服器會處理該過程，而不用將所有的資料送至客戶端處進行組合。請參閱第 4 頁的「組合」。

在動畫控制面板中，最多能有兩個數字型欄可以對照為滑動桿。因為動畫面板中的滑動桿沿著滑動桿路徑從一點移向另一點，因此平板在動畫過程中改變其顏色和透明度。不像「分散可視化工具」，平板的位置和大小都不發生變化；它們在固定的、統一間隔的位置上。只有它們的顏色和透明度在變化，造成實際運動的幻覺。

如果字串欄對照到坐標軸上，則分組定義為該欄的相異值。經由平均對照為顏色的欄的值進行平均，沿字串坐標軸的值的順序可由對這些平均值的排序來自動決定。沿字串值坐標軸觀察顏色的變化，可看出該欄與對照為顏色的欄之間有相關性。如果沒有顏色，則透明度變數將用於確定順序。

平板可視化工具透明度

對照為透明度的欄應是記錄計數或用於對記錄進行加權的欄。平板透明度， α 根據下列關係建立：

$$\alpha = 1 - e^{-u \cdot weight}$$

權重是對照為透明度的欄（或者如果沒有對照為透明度的欄，則為記錄計數）。當權重變大時，該函數的形狀為透明度漸進於 1（完全透明）。當您調節主視窗左邊的透明度比例滑動桿時，變數 u 用於比例變化。圖 1-28 顯示該函數當 u 為高值或低值時的形狀。圖 1-29 顯示 u 為低值或高值時相同的可視化過程。

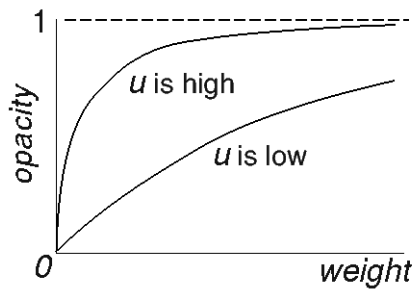


圖 1-28 當 u 為高值或低值時透明度函數的形狀

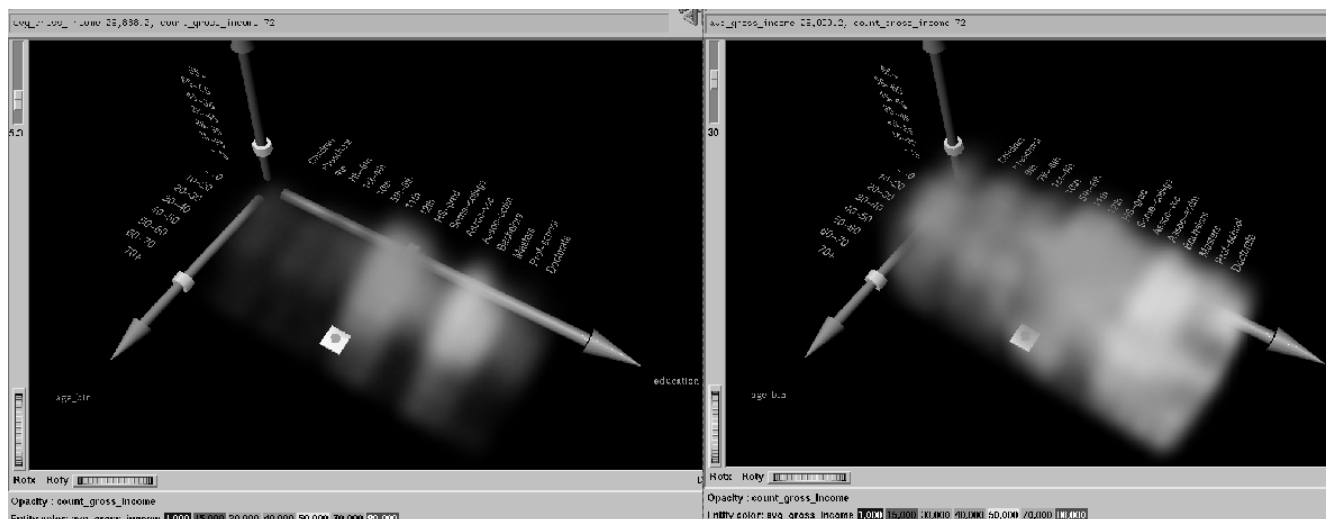


圖 1-29 當 $u = 5.3$ ，和 $u = 30$ 時的影像

如果沒有記錄對照為透明度，「分散可視化工具」會產生在組合時產生記錄計數的欄。這就意味著所有的記錄都進行相等的加權。在該欄中進行加總過程，當坐標軸和滑動桿欄進行分組時，在對照為顏色的欄上進行平均加總過程。所有其它的欄都是不必要且可移除的。您不需要將任何東西對照為透明度，除非您想對每個記錄進行加權，而且其權重不為 1。

關於在「工具管理員」而不是在客戶端進行處理的詳細內容，請參閱 [《MineSet 3.0 for Windows 企業版使用者指南》](#) 中「平板可視化工具」的「處理技術」部分。

當您啟動該工具，所有的過程將在伺服器上進行，而資料檔案 *adult94.splatviz.data* 包含原始資料中進行了列組合的列。

在一些情況下，您可能想要一個對記錄進行加權的欄。例如，如果您有一個資料集其中一欄為人口，另一欄為平均薪資（您想對照為顏色），您可以將人口對照為透明度，平均薪資對照為顏色；然後用「平板可視化工具」進行組合。該組合過程透過坐標軸和滑動桿欄進行分組，因此，它將求出透明度欄（在這種情況下為人口）的總和。新欄稱為 `sum_population`。`average_salary` 欄已修定，它仍然是平均薪資，但根據每列中人口值來加權。如此一來，平均薪資欄仍然顯示它所代表的所有人的平均薪資。

選擇性地，如果您因為資料集的容量而想避免客戶端的處理和儲存，可以按照下面所示在「工具管理員」中執行相同的組合過程：

1. 建立一個新欄，定義 `temp = population * avg_income`。
2. 執行組合過程：按照坐標軸和滑動桿欄進行分組，求出人口的總和與 `temp` 的總和。
3. 建立一個新欄，定義為
`avg_salary = sum_temp / sum_population`
這就會產生加權平均。
4. 現在您可以將 `sum_population` 對照為透明度，將 `avg_salary` 對照為顏色。

請注意：如果您沒有在「工具管理員」中正確地執行這些工作，「平板可視化工具」就會自動執行這些步驟。然而，如果您在「工具管理員」中執行了它們，會比在伺服器上進行要更有效，因為這樣會使客戶端能夠檢索較小的檔案。

平板可視化工具檔案需求

「平板可視化工具」需要下列檔案：

- 資料檔案由索引標籤分隔的欄位的列所組成。利用「工具管理員」可以很容易的建立這個檔案。如果您正在產生該檔案，請參閱 *《MineSet 3.0 企業版介面指南》* 中的為「平板可視化工具建立資料和配置檔案」尋找所需的檔案格式。

您可從資料來源（例如資料庫）中抽取資料來產生資料檔案，並將它按照「平板可視化工具」的需要進行格式轉化。資料檔案擁有使用者定義的附檔名（「分散可視化工具」工具中提供的檔案有 `.data` 附檔名）。

- 配置檔案描述輸入檔案的格式及如何進行顯示。「工具管理員」可以建立該檔案，或可用您所喜歡的文字編輯器來產生該檔案 (請參閱 [《MineSet 3.0 企業版介面指南》](#) 中的「為平板可視化工具建立資料和配置檔案」一章)。

配置檔案必須有 *.splatviz* 附檔名。當啟動「平板可視化工具」或開啓檔案時，您必須指定配置檔案而不是資料檔案。

啟動平板可視化工具

有幾種方法可以啟動「平板可視化工具」：

- 利用「工具管理員」來配置和啟動 3D 可視化工具。在 [《MineSet 3.0 for Windows 企業版使用者指南》](#) 中有詳細的描述。
- 從「工具管理員」的「可視化工具」下拉式功能表中選擇「平板可視化工具」。選擇「檔案」>「開啓」來開啓配置檔案。
- 如果您知道要利用的配置檔案，按一下該配置檔案的圖示。這樣就啟動了 3D 可視化工具並自動載入指定的配置檔案。只有當配置檔案以 *.splatviz* 結尾時才有效 (對於用「工具管理員」為「平板可視化工具」建立的配置檔案都是這樣)。
- 在 UNIX 指令行提示下，敲入：

```
splatviz [configFile]
```

configFile 是可選的，指定所使用的配置檔案的名稱。如果未指定配置檔案，您必須用「檔案」>「開啓」來指定一個。

啟動「平板可視化工具」的 IRIX 選項

-quiet 選項刪除顯示過程的跳現式對話方塊。加入一行可以使該選項永久可用：

```
*minesetQuiet:TRUE
```

到您的 *.Xdefaults* 檔案。

Windows 使用者可以透過檔案 > 特性功能表達到同樣的效果。

平板可視化工具形狀選項

平板選項可以為平板指定一定量的特徵，然後「平板可視化工具」將進行圖形顯示。

- **平板顏色**可控制用於平板的顏色。您可以
 - 指定所用顏色清單
 - 指定對照類型
 - 將顏色清單對照為值清單
- **平板形狀**可選擇用於繪製平板的一種方法：線性形、高斯形、紋理形、球形、立方體或菱形。請參閱第 165 頁的「[形狀功能表](#)」可得到更深入的說明。

要使用「顏色」選項，您必須將一欄對照為「資料目標」面板中的顏色。如果沒有在顏色清單中輸入，則使用預設的顏色圖。預設的顏色圖是一個從藍（最低值）到紅（最高值）的連續譜。要得到如何選擇和改變顏色的詳細說明，請參閱第 48 頁的「[顏色選擇](#)」。

顏色清單您可以利用顏色清單標籤旁的 + 按鈕來指定顏色清單。這會產生一個顏色編輯器，您可以利用它指定要加入清單的顏色。

顏色對照您可以指定在圖形顯示中所表現的顏色變化是**連續**或**離散**的。如果您選擇「連續」，作為在**顏色對照**欄位中對照到顏色的值的一個功能，顏色值在**顏色清單**欄位中所輸入的顏色間逐漸地變化。

跳現式按鈕右邊的欄位可讓您輸入對照該顏色的指定值。如果您不指定任何對照值，就會使用顏色欄中值的範圍。請參閱第 48 頁的「[顏色選擇](#)」可得到關於選擇顏色的更多資訊。

匯總選項

匯總選項可讓您指定在匯總視窗中使用何種顏色。只有當您將欄對照為匯總時才可用。

其他選項

在對話方塊底部的「其他選項」，包含下列欄位：

- *隱藏標籤距離* 控制坐標標籤（對於字串值坐標軸）變為不可見時的距離。增加該數值可以讓標籤在更遠的距離處出現。資料越高，標籤隱藏的距離越大。
- *坐標軸標籤尺寸* 控制坐標軸標籤的大小。較小的數字會減小尺寸，而較大的資料會增加尺寸。
- *柵格顏色* 經由在上面按一下可改變柵格的顏色。會讓「顏色選擇」對話方塊出現，利用它可以完成顏色的改變。
- *柵格(X, Y, Z) 尺寸* 可指定各自坐標軸柵格線之間的間隔。較小的數字會減小尺寸，而較大的資料會增加尺寸。如果尺寸為 0，則在該維度就沒有柵格線。

重設工具選項

按一下 *重設* 按鈕將所有選項的值設為預設值。

儲存平板可視化工具設定

當您按下 *啟動工具* 時，「工具管理員」在幾個檔案中儲存「平板可視化工具」的資訊，且所有檔案都具有相同的字首：

- *<prefix>.splatviz.data* 包含資料。
- *<prefix>.splatviz.schema* 描述資料檔案。
- *<prefix>.splatviz* 包含「平板可視化工具」所需的資訊。

按照目前工具選項儲存整個階段部分，從「工具管理員」的「檔案」功能表中使用這些功能表選項：

- *另存目前階段為 ...* 預設字首根據資料來源
- *另存目前階段為 ...* 指定您自己的字首

已儲存檔案的開頭為 *<prefix>.mineset*，並且包含 *MineSet* 返回目前狀態所需的所有資訊。

當您使用 *啟動工具* 後，如果需要將會更新 *.data*、*.schema* 和 *.splatviz* 檔案。

在平板可視化工具中的空值處理

當具有未知或空值的欄位對照為可視化屬性時，「平板可視化工具」會使用特殊的表示方式。(關於空值的討論，請參閱 [《MineSet 3.0 企業版介面指南》](#) 的「MineSet 中的空值」一章。) 當分組中對照至顏色的欄中每個記錄都有空值時，該平板過程的結果為灰色。如果組合中的一個或多個記錄對照到顏色的欄具有非空值時，該 (或那些) 值則用於計算顏色。一個值與空值的和為空值，一個值和空值的平均就為該值 (也就是說， $\text{value} + \text{Null} = \text{Null}$ ； $\text{avg}(\text{val}, \text{Null}) = \text{val}$)。

當空值在正文處出現時，它顯示為問號 (?)。(「選擇視窗」與「游標位於」區域在它們自己的部分中討論。)

對於含有空值 (對照於座標軸) 的數字欄，在座標軸定義的範圍之下有一個特殊的空值位置。這有助於顯示該空值與其它值不連續。利用「檢視」功能表中的「空位置」選項，可以關閉數字坐標軸的空值位置。至於對照到坐標軸的字串值欄，空值 (由 ? 代表) 被視為一個值。

為平板可視化工具建立的滑動桿

在主視窗旁的匯總視窗附近出現的滑動桿數目根據在配置檔案中指定的滑動桿對照。任何帶著識別標籤的滑動桿的出現，取決於資料集是有二維度、一維度或滑動桿對照。

對照到滑動桿 1 和滑動桿 2 上的欄最終會形成滑動桿的索引值。這些欄必須是數字型 (整數型、浮點型、雙精度型) 或已分組。如果對照為滑動桿的欄已分組，該欄就不需自動分組，並且該欄會使用為滑動桿的索引。然而，如果該欄未分組，使用自動統一分組就可建立一個分組的欄。(請參閱第 34 頁的「分組」可得到更多的資訊。) 在形成自動分組中用到的欄從目前的表格中刪去。

動畫控制面板

動畫控制面板，出現在主視窗的右邊，由最多帶有兩個滑動桿的匯總視窗、資訊欄位、動畫按鈕和動畫滑動桿組成。請參閱第 10 頁的「動畫」。

在每個單獨的滑動桿位置上（由匯總視窗的黑點指示）畫面會對應記憶體的資料表。要在一維滑動桿上進行插值，將兩個相鄰表合併，然後將空間欄作為特定鍵進行組合。當滑動桿從一個分組位置移向另一個時，則對每個平板的權重（對照為透明度）進行插值（如果表缺少特定列，則假定權重為 0）。用於平板的平均值也進行了插值，但只是按照權重進行加權。

樣例 1-1 插值過程

該樣例描述了插值過程的技術細節。假設我們想要顯示一個影像，該影像代表外部滑動桿之上在 40-50 歲表格和 50-60 歲表格之間的插值。讓表 1-18 和表 1-19 分別為年齡 =40-50 和年齡 =50-60 的表格，即兩個滑動桿的位置。

表 1-18 年齡在 40 到 50 之間

教育程度	職業	工作時間	收入	加權
HS-grad	Exec-Man.	15-25	25000	2
HS-grad	Mach-op	15-25	30000	1
碩士	技術人員	25-35	35000	3

表 1-19 年齡在 50 到 60 之間

教育程度	職業	工作時間	收入	加權
HS-grad	Exec-Man.	15-25	70000	1
職業的	Mach-op	35-45	40000	2

下面就是「平板可視化工具」進行插值的過程。對於表 1-18，增加了一個等於 $(1-t)weight$ 權重新欄以及一個等於 $(1-t)(weight)(value)$ 的權重新欄。對於表 2，增加了一個等於 $t(weight)$ 的權重新欄，以及一個等於 $t(weight)(value)$ 的權重新欄。合併兩個表。

以空間坐標軸為關鍵，並匯總兩個新欄的組合結果，將兩個表合併。這就確保對於所有的空間坐標軸，不存在具有相同分組值的兩列。最後，將總和值除以總和權重來得到內插值。在此種情況下，該內差值是對收入而言。如果 $t=.5$ ，最後的表格將為表 1-20。

表 1-20 在表 1 和表 2 之間插值的中間過程

教育程度	職業	工作時間	收入	加權
HS-grad	Exec-Man.	15-25	40000	1.5
HS-grad	Mach-op	15-25	30000	.5
碩士	技術人員	25-35	35000	1.5
職業的	Mach-op	35-45	40000	1

如果外部搜尋滑動桿是 2D，則使用雙線性插值。

該普查資料集包含將近 150,000 列。外部滑動桿的目的是在其中瀏覽，並顯示資料中附加維度的匯總資訊。紅色的區域代表總和值高的地區；白色區域顯示值較低的地區。當滑動桿停在黑點上時，則影像顯示未插值的資料。可以追蹤滑動桿的軌跡，並且利用滑動桿下的 VCR 控制面板顯示動畫。

要顯示動畫是如何形成，假定您具有 8 年的資料，1990-1997 (也就是說，在匯總視窗中有 8 個資料點)。可以由檢查平板如何隨滑動桿從一年移向另一年而進行變化的方式開始。假定 1990 中，在給定的位置平板的值為 20(被對照為顏色) 並且權重為 2，這就意味著它代表了 2 筆記錄。進一步假定 1991 中，同樣的平板的值為 40 權重為 200。

1991 中的平板比 1990 中的更加透明，這是因為它代表更多記錄的組合結果 (或者具有更大權重的記錄)。當您將年滑動桿從 1990 移到 1991 時，將會線性插入到 2 和 200 之間來改變權重。

計算經記錄權重加權所得兩個值的平均可以得到該值。例如，在 1990 和 1991 之間，權重為 101，而且該值為 $((1-.5)*2*20+.5*200*40)/((1-.5)*2+.5*200) = 39.8$ 。當您接近 1992，其大小也接近 40。

您不可能在離散資料點之間停止動畫過程，並且您不能將路徑滑動桿拖曳到離散資料點之間的固定位置。匯總視窗中的資料點代表與資料檔案中實際資料相對應的滑動桿位置。例如，值 20 和 40 代表了實際值的組合，而 39.8 則不是。

在平板可視化工具中的下拉式功能表

您可使用五個下拉式功能表來存取「平板可視化工具」的附加功能：它們為「檔案」、「檢視」、「選項」、「形狀」以及「說明」。它們的描述可參閱第 92 頁的「檔案功能表」、第 197 頁的「檢視功能表」、第 153 頁的「選項」按鈕以及第 97 頁的「說明 (IRIX)」中的項目。

形狀功能表

在該工具中使用的平板是用來對小雲團進行模型化 (請參閱 Lee Westover 所著的「*Proceedings of SIGGRAPH '90*」Vol.24, No.4, 第 367-376 頁「Footprint Evaluation for Volume Rendering」)。

「形狀」功能表可改變繪製平板的方法。您可以選擇交互式地交換準確度。在每個估計方法中，紋理平板是理想高斯型密度的最準確表達。因為大多數的電腦較支援硬體輔助結構，因此結構平板通常是最好的選擇。在 SGI 平台中，只有 Indy 或更早的系統受到較慢的軟體執行之限制。三種平板類型是：

- **線性形** 繪製少量的三角型來對高斯型平板進行線性估計。
- **高斯形** 繪製大量的三角型來估計高斯型平板。
- **結構** 利用結構對照矩形提供最準確的表示法。在不支援硬體輔助結構對照的機器上該過程是很慢的。

選擇性地，允許下列透明原語。

- *球形*繪製透明球體，其直徑隨…的立方根而變動
權重 (或 weight)。
- *立方體*繪製一個立方體，其寬度隨計數 (權重) 的立方根而變化。
- *菱形*繪製一個三角方塊架，其邊長隨計數 (權重) 的立方根而變化。

配置和資料檔案樣例

有一些由 MineSet 提供的樣例資料和配置檔案，它們用來展示「平板可視化工具」的特性和功能。每個檔案的詳細說明請參閱[附錄 A](#)，「[配置和資料檔案樣例](#)」

拆分下限

拆分下限是一個用於改進「決策樹導入工具」和「回歸樹導入工具」的選項。增加它會產生更小的樹，但卻會降低準確度。

拆分下限是權重的下限 (通常情況下，如果權重沒有設定，就為記錄的數目)，必須至少存在於兩個節點的子節點中。該選項的預設值為 2。例如，如果節點中有一個三向拆分，三個子節點中至少有兩個其權重必須為 2 或更多 (如果沒有設定權重則為兩筆記錄或更多)。這就提供限制「決策樹」規模的一種方法。

提高拆分下限是為增加機率估計的可靠度，因為每個葉子的記錄數較大。這也會建立更小的樹並減少導入時間。如果您認為資料包含噪音 (錯誤或異常)，或者在估計機率時使用樹 (請參閱[第 15 頁的「應用模型」](#))，則將拆分下限增至 5 或更多。如果您的資料集很小 (<100 筆記錄)，您可以減小該數到 1。請參閱[第 71 頁的「決策樹」](#)。

拆分標準

該選項提供「決策樹」中的三個拆分標準。下面是技術性定義。對於已知的問題，很難說哪種標準是最好的。可以全部進行試驗，然後選擇產生最低錯誤估計或最容易理解的「決策樹」。

*公共資訊*是指父節點和子節點純度加權平均之間純度的變化（即為*熵*）。加權平均是根據每個子節點上記錄的數目來進行的。

標準化公共資訊（預設情況下）是公共資訊除以子節點數目的以 2 為底的對數。

*增量比*是指在忽略標籤值的條件下，公共資訊除以拆分的*熵*。

*標準化公共資訊*和*增量比*將優先權指定給少量值的屬性。

提供給「回歸工具」的選項決定在回歸樹中欄的選擇。

*方差*確定在回歸樹中如何選擇欄。方差選擇的欄可以產生使節點方差最小化的拆分。「選擇方差」將產生回歸樹，其葉結點處有平均預測器。

*絕對偏差*選擇的欄可以產生使節點絕對偏差最小化的拆分。「選擇絕對偏差」將產生回歸樹，其葉結點處有中值預測。

標準化變數（預設情況下）為方差除以子節點數目的以 2 為底的對數。

*標準化絕對偏差*為絕對偏差除以子節點數目的以 2 為底的對數。

統計可視化工具

「統計可視化工具」，可以從「工具管理員」中「資料目標」窗格的「可視化工具索引標籤」中找到，它顯示帶有一系列小面板的視窗，每個在「工具管理員」窗格的*目前欄*中列出的欄分配一個。只有有限個欄面板可以同時顯示，使用邊捲軸長條圖，或者平行或垂直伸展「統計可視化工具」視窗，便可以看到更多欄面板。

在「工具管理員」資料集一定數量記錄的基礎上，利用「統計可視化工具」您可以看到某些統計結果。欄面板的格式根據欄類型以及該欄所存在的相異值的數目而變化。欄通常分為兩種類型：*數字型*和*離散型*，分別顯示為方塊圖及柱狀圖。

如何讀取統計可視化工具

方塊圖按照*數字型*欄繪製，每個數字欄由整數、浮點型、雙精度型或日期型值所組成。每個方塊圖面板顯示單一欄資料的統計結果，包括這些數字型值的最小值、最大值、平均值、中值以及兩個四分之一值 (25% 和 75%)。這些值顯示為橫穿具有綠色底紋垂直棒形的線，並且數量的標準差顯示為 $\pm$ 值。只有少於 50,000 個值的時候，才會顯示四分之一的值 (請參閱圖 1-30)。如果欄中的相異值超過 50,000 個，則統計結果顯示為灰色條。

平均值為欄資料的和除以記錄的總數。中間值為在指定欄中按照大小順序排列的數之中間數。標準差是欄中資料離散度的一種度量。

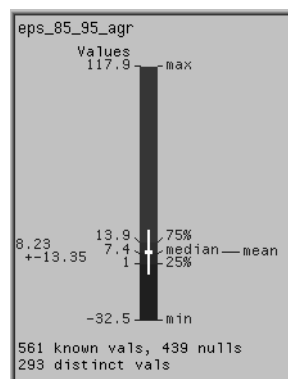


圖 1-30 由統計可視化工具顯示的數字型欄

柱狀圖用於繪製**離散型**(或**標稱型**)欄，具有非數字型(字串、分組或枚舉型)的值(請參閱圖 1-31)。離散型欄面板最多顯示 100 個相異值，以及各個相異值對應的實例數目的柱狀圖。離散列的預設順序是降冪排列，但是您可以利用「檢視」下拉式功能表來選擇另一種排序方式。如果只有 100 個或更少的列目，欄面板可以包含相異值的數量。

當顯示離散值時就可以使用柱狀圖，例如，是 / 否值或州名。不論方塊圖或柱狀圖，每個方塊都顯示資料集中記錄的數目，以「總數」表示，而該特殊方塊中的所代表的相異記錄數目以「相異值」表示。

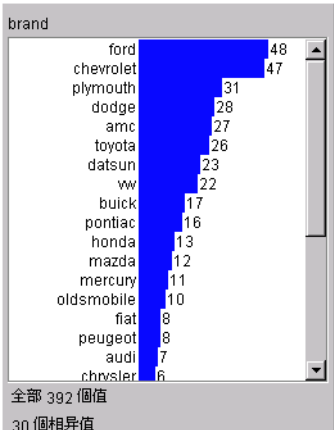


圖 1-31 統計可視化工具顯示的離散欄

如果您從圖示中開啓「統計可視化工具」，則在主視窗中只有「檔案」和「說明」功能表能夠使用。如要顯示主視窗的所有功能表和控制項，請開啓一個 *.statviz* 檔案。利用「檔案」>「開啓」可以看到可用配置檔案的清單。

統計可視化工具下拉式功能表

三個下拉式功能表可以讓您進入「統計可視化工具」的附加功能：它們是「標記檔案」、「檢視」和「說明」。如果沒有指定配置檔案就啓動「統計可視化工具」，則只有「檔案」和「說明」功能表可用。請參閱「檔案」和「說明」項目。

統計可視化工具的檢視功能表

「統計可視化工具」中的「檢視」下拉式功能表對柱狀圖和方塊圖進行分類：

- 按計數值排列名稱指定，名稱（離散）欄顯示為依照每值實例計數進行降冪排列的柱狀圖。
- 按字母順序排列名稱，指定必須按照字母順序來排列值。

歷史表按鈕

MineSet 允許您使用一系列操作來轉換資料表。一系列轉換記錄下來，並且透過「工具管理員」「資料轉換」面板下部的**歷史表按鈕**來回溯特定的轉換。利用這些按鈕您可以看到轉換過程中的每一步驟，並且當您犯錯誤時可以返回。當您按一下向左的箭頭按鈕，欄視窗會顯示在前一步驟出現的表。按一下向右箭頭按鈕將表格返回到目前狀態。

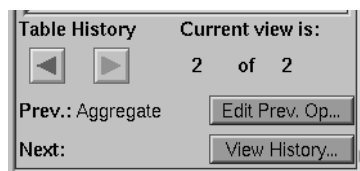


圖 1-32 「歷史表按鈕」的「目前操作檢視為」欄位

「目前操作檢視為」欄位

在歷史表按鈕的右邊是資訊欄位**目前檢視為**，它記錄您所做的轉換次數並且指出您正在查看的步驟。該欄位中的兩個整數表示您正在觀察轉換序列中的哪一步以及現存總步驟數)。例如，如果您做了兩次轉換，您可以檢視原始表 (總數 3 中的第 1 個)，在第一次變化之後的表 (總數 3 中的第 2 個)，或第二次變化之後的表 (總數 3 中的第 3 個)。

「上一個」和「下一個」按鈕

當您利用**歷史表按鈕**向前或向後查看以前的轉換時，**前一個**和**下一個**欄位 (在箭頭按鈕之下) 有助於您定位在歷史表中的位置。對於您所檢視的任何一個表，**前一個**：欄位會告訴您前一個轉換，與**下一個**：欄位會告訴您下一個轉換。

「編輯前一個操作」按鈕

「編輯前一個」操作按鈕允許您編輯在*前一個*欄位中顯示的操作。(該按鈕在下列情況下變灰，當*目前檢視*為：*某數的1*，因為這是指原始表格，以前沒有發生轉換。)當您按一下「編輯前一個操作」按鈕，則跳現出前面操作的對話方塊，您可以對該轉換進行改變。例如，如果之前的轉換是爲了要將欄分組，當您按一下「編輯前一個」操作時，出現欄分組對話方塊。

經由改變先前的轉換，您可以影響目前表格建立的後續轉換操作。例如，如果您刪除了在後續分組操作中用到的列，則分組操作就變爲無效。*編輯歷史*按鈕可讓您避免這樣的問題。

刪除操作至結束按鈕 (僅適用 Windows)

當您在歷史中做備份時，「工具管理員」的「資料目標」面板 (在主視窗的右邊) 便禁用 (變灰)，這是因爲您不在歷史的結尾。爲使「資料目標」面板再次啓用，您必須前往歷史的末端，或者按一下*刪除操作. 至結束*按鈕來刪除您在目前檢視後的所有操作。

操作歷史索引標籤

當您按一下*操作歷史索引標籤* (在 *IRIX* 中的*檢視歷史*按鈕)，顯示目前欄和資料目標的面板由一個顯示*資料轉換*表格全部歷史的面板所取代 (圖 1-33。)表現爲小矩形表的每個表格都包含欄的清單，並經由一個一個大的矩形圖表 (表示在表格上執行的操作) 與下個表格連結。

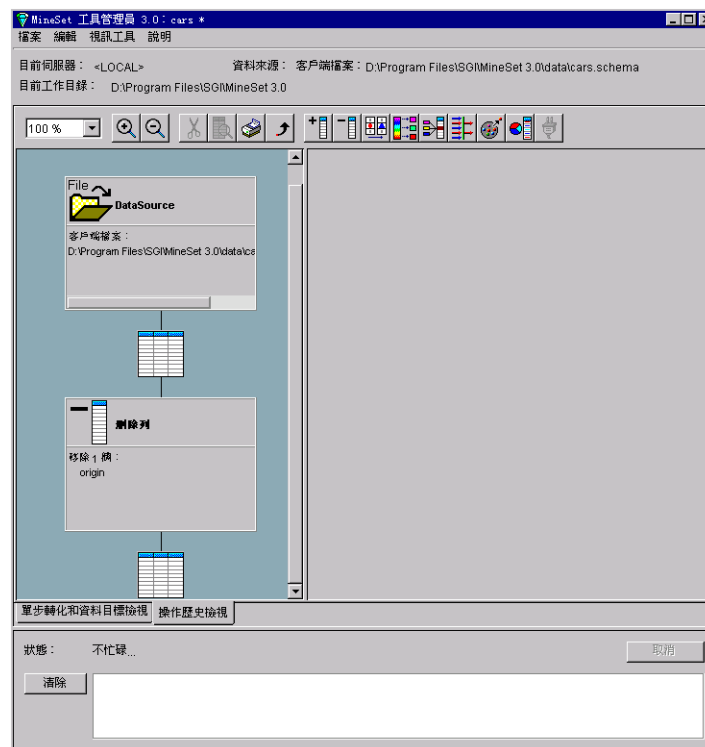
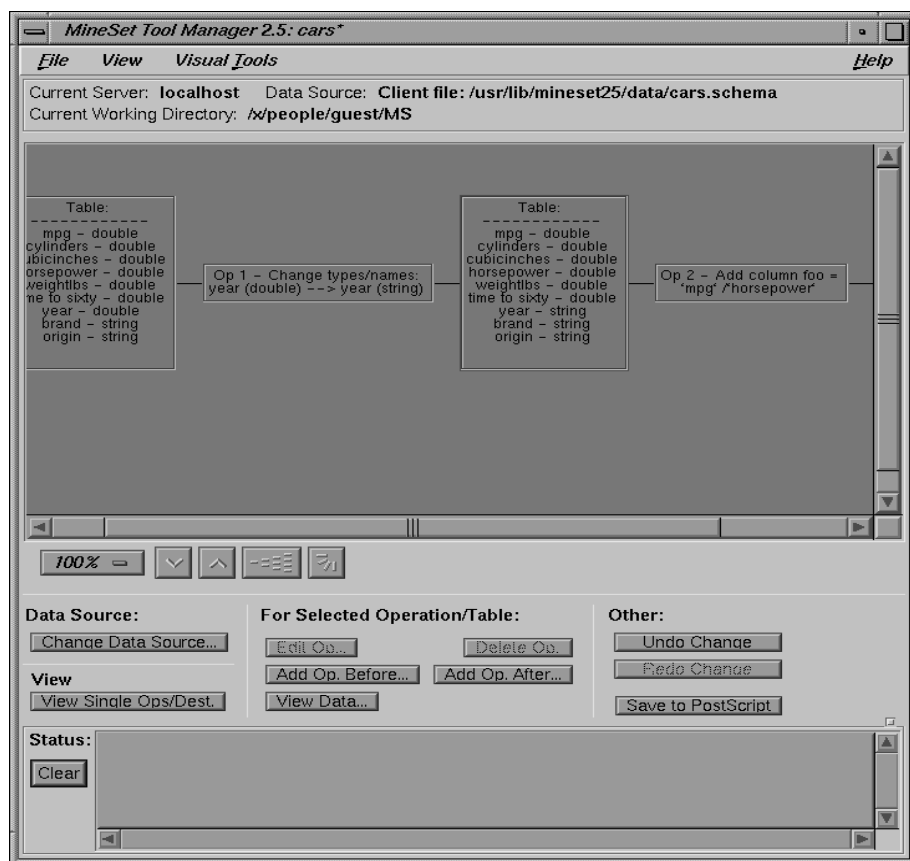


圖 1-33 查看歷史對話方塊 (Windows)

Windows 版本中工具列上的圖示可幫助您用「工具管理員」對資料集進行檢查和執行不同的操作。按一下左邊窗格的一個操作，然後按一下單步轉換和資料目標檢視索引標籤 (IRIX 上的檢視單步按鈕)，它呈現處於選定操作階段的「工具管理員」(請參閱圖 1-33 和圖 1-34。)按一下操作檢視歷史索引標籤 (在 IRIX 上的檢視歷史按鈕) 將返回先前的檢視。



■ 1-34 檢視歷史對話方塊 (IRIX)

與「編輯前一個」操作，改變一個操作通常會影響（有時會使其無影響）歷史中的後續操作。您可以選擇一個特定的操作來編輯、增加或查看。當編輯操作影響了歷史，在「檢視歷史」對話方塊將警告您並顯示新的歷史。

在 IRIX 版本中「檢視歷史」面板中的圖表視窗下面的一排按鈕允許您改變圖表的方向和大小。

工具管理員

「工具管理員」是用戶用於指定配置檔案、資料檔案和使用工具的圖形界面。該部分討論「工具管理員」的一般操作。關於如何具體使用，請參閱 *《MineSet 3.0 for Windows 企業版使用者指南》*。

「工具管理員」在 MineSet 客戶端執行。該過程一般都會遵循下列路徑：

1. 「工具管理員」開啓一個與 DataMover 的連接，該連接執行在 MineSet 伺服器上。在一些情況下，該伺服器可以是您客戶端的工作站，而在其它情況下則是不同的機器。
2. 「工具管理員」可以指定
 - 資料庫和表格，或在客戶端或伺服器上包含資料的二元或 ASCII 檔案。
 - 應用挖掘或可視化工具。
 - 資料如何經由工具選項顯示
 - 階段檔案儲存您工作的歷史

經「資料移動器」取得的資訊用於指導交互作用。結果，「工具管理員」產生配置檔案。該檔案包含用於執行後序步驟的使用者定義參數。

3. 「工具管理員」將步驟 2 配置檔案的複本傳送給「資料移動器」。「資料移動器」透過以下途徑處理檔案
 - 存取資料庫或檔案
 - 執行指定的資料轉換
 - 在需要時執行挖掘工具
 - 在需要時生成可視化處理檔案

這些可視化處理檔案包含 MineSet 工具可讀取並以指定格式儲存的資料。然後，這些可視化過程檔案的一份複本就轉給 MineSet 客戶端。

4. 「工具管理員」啓動適當的 MineSet 可視化工具。
5. 該工具存取可視化處理檔案並顯示資料。
6. 如果您產生一個模型，該模型可應用於新的資料。

「工具管理員」特性

「工具管理員」的「特性」對話方塊允許您設定下列選項：

- 啓動時自動恢復當您登錄 MineSet 時允許您返回上次進行使用的最後階段。MineSet 儲存歷史，並以您上次離開時的相同狀態開啓檔案。
- 使用二元資料檔案通知 MineSet 使用二元檔案，這樣可以減少處理時間。
- 最大屬性值數可讓您設定資料集中屬性值的截斷數目。在計算中將不使用具有更多單一值的任何欄。
- 並行過程可讓您為 IRIX 的並行過程設定選項。

訓練集

訓練集是一個包含屬性的表，其中之一指定為分類標籤。該標籤是您將用於預測的屬性。

該樣例的目標是在描述性屬性花萼長度、花萼寬度、花瓣長度及花瓣寬度的條件下預測蝴蝶花 (藍色蝴蝶花、多色蝴蝶花或純白色蝴蝶花) 的類型。圖 1-35 顯示樣例訓練集中的幾個記錄。

描述性屬性					標籤
	花萼長度	花萼寬度	花瓣長度	花瓣寬度	iris 類型
記錄 1	5.1	3.5	1.4	0.2	藍色蝴蝶花
記錄 2	5.9	3	5.1	1.8	純白色蝴蝶花
記錄 3	6.5	2.8	4.6	1.5	多色蝴蝶花
⋮	6.3	2.9	5.6	1.8	純白色蝴蝶花
⋮	6.5	3	5.8	2.2	純白色蝴蝶花

圖 1-35 訓練集中的記錄樣例

一旦建立模型，就可以為新記錄預測標籤。這些新記錄必須存在於擁有模型所用的全部屬性之表格中，並具有這些屬性在訓練集中相同的名稱和類型。該表不需要包含標籤屬性。如果存在，在預測的過程中就會將其忽略。

樹狀可視化工具

「樹狀可視化工具」是用來在 3D 場景中顯示資料的圖形界面，它將您的資料表現為等級塊（節點）和伴隨圓盤的長條圖，您可以動態的瀏覽、查看資料集的部分或全體。如何使用「樹狀可視化工具」在《*MineSet 3.0 for Windows 企業版使用者指南*》中有詳細的討論。

「樹狀可視化工具」將資料表現為以等級區別又相互連接的節點形式來顯示資料定量和相關的特徵。每個節點都包含其高度、顏色和圓盤位置與資料值組合相對應的長條圖。與節點相連的邊（顯示為線條）顯示各部分資料與其子集間的關係。

在次一級組中的值可以進行匯總並在相鄰的更高的層中自動地顯示。長條圖下面的基準塊可以提供所有長條圖組合的資訊。代表負值的長條圖顯示在基準塊頂部之下。使基準塊高度無效，您可以更清楚的觀察負值基準塊（請參閱第 192 頁的「樹狀可視化工具顯示功能表」，或《*MineSet 3.0 企業版介面指南*》中「為樹狀可視化工具建立資料和配置檔案」中的「基準塊高度」說明部分。

檔案需求

「樹狀可視化工具」需要下列檔案：

- 資料檔案由索引標籤分隔的欄位列所組成。利用「工具管理員」可輕易地建立該檔案，請參閱《*MineSet 3.0 for Windows 企業版使用者指南*》中的描述。如果您要自己生成該檔案，請參閱《*MineSet 3.0 企業版介面指南*》的為「樹狀可視化工具建立資料和配置檔案」找尋所需的檔案格式。

資料檔案有使用者定義的附檔名（「樹狀可視化工具」工具中提供的樣例檔案有 *data* 附檔名）。

- 配置檔案描述輸入資料的格式以及它們如何轉換等級結構。利用「工具管理員」可輕易地建立該檔案，請參閱《*MineSet 3.0 for Windows 企業版使用者指南*》中的描述。您也可利用自己喜歡的文字編輯器（例如，WordPad、jot、vi 或 Emacs）來產生該檔案，請參閱《*MineSet 3.0 企業版介面指南*》中「為樹狀可視化工具建立資料和配置檔案」。

配置檔案必須有 *.treeviz* 附檔名。當啟動「樹狀可視化工具」或開啓一個檔案時，指定的是配置檔案而不是資料檔案。

啟動樹狀可視化工具

有幾種方法啟動「樹狀可視化工具」：

- 如何使用工具管理員，請參閱 *《MineSet 3.0 for Windows 企業版使用者指南》*。
- 從「工具管理員」的「可視化工具」下拉式功能表中選擇「樹狀可視化工具」。選擇「檔案」>「開啓」來開啓配置檔案。
- 如果您知道要利用的配置檔案，按兩下該配置檔案的圖示。這樣就啟動了 3D 可視化工具並自動載入指定的檔案。只有當檔名以 *.treeviz* 結尾才有效（對於用「工具管理員」為「樹狀可視化工具」建立的配置檔案都是這樣）。
- 從 IRIX 指令行中輸入：

```
treeviz [configFile]
```

configFile 是可選的，指定所使用的配置檔案的名稱。如果未指定配置檔案，您必須用「檔案」>「開啓」來指定一個。

您可以開啓一個警告系統並在啟動工具時禁用對話方塊，請參閱第 199 頁的「警告選項」。

樹狀可視化工具選項

按一下 **工具選項** 按鈕可以產生用於顯示的新對話方塊。（圖 1-36 和 圖 1-37）。這就允許您改變「樹狀可視化工具」選項中的預設值。



圖 1-36 樹狀可視化工具配置選項對話方塊 (Windows)

在 Windows 版本中，對話方塊的頂部顯示三個索引標籤：長條圖，節點基準塊及圓盤圖。每個索引標籤都展示一系列選項，您可以對可視化過程進行詳細的配置。要得到關於選擇顏色的更多資訊，請參閱第 48 頁的「顏色選擇」。要指定每個對話方塊的「高度」部分，請參閱下一個第 180 頁的「標準化高度」。

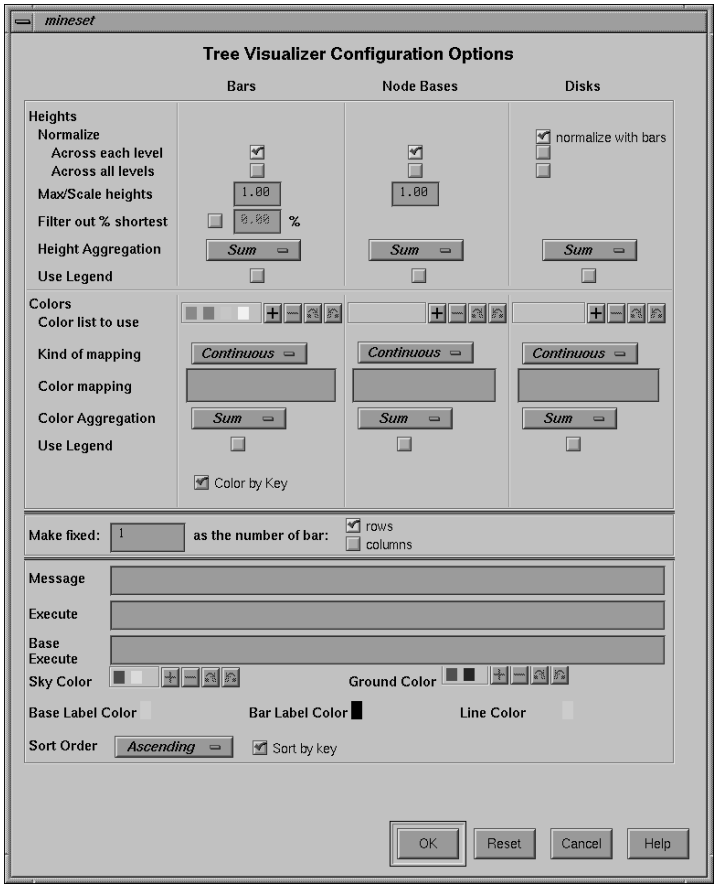


圖 1-37 樹狀可視化工具配置選項對話方塊 (IRIX)

在 IRIX 版本中，對話方塊的頂部顯示三欄：長條圖、節點基準塊和圓盤圖。在這個對話方塊中的所有選項都可進入。

標準化高度

該選項可使您對所有層次中的每個層次上長條圖、節點基準塊以及圓盤圖的高度進行標準化。將高度標準化確定高度變數的最大值；它把與該高度相關的所有值標準化。因此，如果最大值為 30.0，並且最大長條圖高度設為 1.0 (以任意單位)，則大小為 15.0 的值將對照為 0.5。

層內標準化獨立地標準化了等級的每一層。如果資料是按等級匯總的，該選項最為有用，並且防止等級的最高層將底層項目過度地壓縮。在所有層上進行標準化會將所有的層一起標準化，不管在等級中的位置。如果沒有為長條圖核取任何方塊，就不會有標準化過程。

節點基準塊的標準化獨立於「長條圖」。如果未核取方塊，用於長條圖的標準化方法也用於節點基準塊，雖然其值的標準化獨立進行。

如果圓盤存在並且選擇了與長條圖一起標準化：圓盤和長條圖代表相同的值具有相同的高度。如果在「圓盤」欄中選取其它標準化選項，則圓盤的標準化獨立於長條圖：不管最高的圓盤和最高的長條圖所代表的實際值，它們具有相同的高度。

最大 / 比例高度

該選項可讓您指定最高長條圖和節點基準塊的高度。預設值為 1.0 (任意單位)。如果在看到檢視以後，您覺得高度太低或太高，使用該欄位來進行調節。例如，在欄位中輸入 2 可以使所有長條圖的高度變為原來高度的兩倍；輸入 0.5 使所有長條圖變為原來的一半。

如果指定標準化過程，該值代表最高長條圖或基準塊的高度。如果沒有指定標準化過程，則所有的值都要用該值進行比例縮放。當對兩個不同的資料集進行比較時，後者可能會有用。

篩選掉 % 最短的

該選項可讓您篩選出只包含短條的節點。首先，計算畫面中最高的長條圖 (如果高度是按照層級來標準化，即每一層級中最高的長條圖)。只顯示那些至少包含一個長條圖的節點，而該長條圖高度相當於最高長條圖高度的適當百分比。例如，如果您在欄位中輸入 5%，那麼就會顯示至少包含一個高度不低於最高高度 5% 的長條圖的節點。(這些長條圖的先驅節點也有顯示)。這一選項可以粗略的篩選掉小的，不感興趣的節點。這並不是識別帶有一定值的特殊節點的精確原理。使用該選項可以加快緩慢、複雜畫面的表達或限制由許多在 0 值附近的長條圖所帶來的雜亂。

雖然篩選出小的節點，在等級上的任何計算中它們仍然要進行計數。

高度組合

預設情況下，父節點長條圖的高度是所有子節點長條圖高度的匯總；但是這些高度可以為平均值、最大值、最小值、計數或任何出現的值。這些值可用於長條圖高度、基準塊高度以及圓盤高度值。

顏色

利用這一套選項可以

- 指定所用顏色清單
- 指定對照類型
- 將顏色對照為長條圖、節點基準塊及圓盤

要使用這些「顏色」選項，您必須已經將一個屬性欄對照到「資料目標」面板所需的顏色 - 長條圖、顏色 - 圓盤或顏色 - 基準塊。要得到如何選擇和改變顏色的詳細說明，請參閱[第 48 頁的「顏色選擇」](#)。

*使用的顏色清單*可讓您使用顏色清單標籤旁的 + 按鈕來指定顏色清單。這會產生一個顏色編輯器，您可以利用它指定要加入清單的顏色。

*對照類型*可讓您指定在圖形顯示中顏色的變化是*連續*或*離散*的。如果您選擇「連續」，作為在*顏色對照*欄位中對照到顏色的值的一個功能，顏色值（長條圖、節點基準塊或圓盤）在「顏色清單」中所輸入的顏色間逐漸地變化。如果您選擇「離散」，顏色只在指定界限內變化。

顏色對照 讓您指定顏色所對照的值。

顏色組合

預設情況下，父節點長條圖的顏色是所有子節點長條圖值的匯總；但是這些顏色可以為平均值、最大值、最小值或任何出現的值。組合可用於長條圖顏色、基準塊節點顏色以及圓盤顏色值。

根據關鍵值來確定顏色

該選項可讓您透過它們的關鍵值來自動地對長條圖上色。如果指定另一個上色方法，則忽略該選項。如果您未指定顏色清單，或指定的顏色不充足，則隨機抽取一些附加顏色。如果指定額外的顏色，則它們會被忽略。

固定列欄

預設情況下，在一列上放置所有的長條圖。該選項允許改變欄或列的數目。如果既未選取列又沒選取欄，或者該數設為 0，則無論列或欄都不固定，並且顯示最為接近正方形的近似的列欄數。

訊息

該選項可讓您輸入您想要的任何訊息。訊息語句指定當指標在一個物件上移動或當選取一個物件時所顯示的訊息。預設情況下，對於基準塊使用與長條圖相同的訊息。如果未指定任何訊息，則使用包含所有欄名稱和值的訊息。

訊息的格式必須與正在使用的資料類型相符：

- 字串必須使用 %s。
- Ints 必須使用整數格式 (如 %d)。
- Floats 和 doubles 必須使用浮點格式 (如 %f)。

如需消息欄位的詳細說明，請參閱 [《MineSet 3.0 企業版介面指南》](#) 中為「樹狀可視化工具」建立「資料」和「配置檔案」一章裡的「消息語句」。

連按兩下執行和基準塊連按兩下執行

這些選項可讓您輸入一個當連按兩下長條圖或基準塊時所執行的指令。如果只填入執行欄位，則它既應用於長條圖也應用於基準塊。如果欄位都已填入，則「執行」應用與基準塊，而「基準塊執行」應用於基準塊。格式與訊息語句一樣。如果不出現執行語句，連按兩下也不會起作用。

如需「執行」欄位的詳細說明，請參閱《*MineSet 3.0 企業版介面指南*》中為「樹狀可視化工具」建立「資料」和「配置檔案」一章裡的「執行語句」。

天空顏色

您可以指定一種或兩種顏色。如果只指定一種顏色，則天空為純色。如果指定兩種顏色，則天空會在兩種顏色間漸變。當指定兩種顏色時，則第一種為天空頂部的顏色，而第二個為底部的顏色。

地面顏色

您可以指定一種或兩種顏色。如果只指定一種顏色，則地面為純色。如果指定兩種顏色，則地面會在兩種顏色間變化。至於地面，第一個顏色為遠處水平線的顏色，而第二個為近處地面的顏色。

基準塊標籤顏色

您可以指定基準塊前面的標籤顏色。

長條圖標籤顏色

您可以指定長條圖前面的標籤顏色。

線條顏色

您可以指定與基準塊相連的線的顏色。

排序順序

如果您選擇了按照關鍵值次序核取方塊，所顯示的節點按照既定順序排列。在核取方塊旁的功能表可指定按照升冪排列或按照降冪排列。

重設工具選項

在對「工具選項」對話方塊作出修改後，如果您想要將所有的設定重設為預設值，按一下**重設選項**按鈕。

儲存新工具選項

一旦您完成對工具選項對話方塊的修改，按一下 *確定* 將返回「工具管理員」主視窗。

儲存樹狀可視化工具設定

「工具管理員」將「樹狀可視化工具」的資訊儲存在幾個檔案中，並具有相同的字首：

- `<prefix>.treeviz.data` 包含資料。
- `<prefix>.treeviz.schema` 描述資料檔案。
- `<prefix>.treeviz` 包含「樹狀可視化工具」所需的資訊。
- `<prefix>.mineset` 包含建立其它檔案所需的所有資訊。

要指定字首，使用 *另存目前階段為...* 「工具管理員」主視窗中「檔案」功能表的功能表選項。如果您不指定字首，則它將根據資料來源的類型而定。

當您使用 *啟動工具* 按鈕，如果需要時，將更新 `.data`、`.schema` 和 `.treeviz` 檔案。

樹狀可視化工具下拉式功能表

利用六個標示為「檔案」、「檢視」、「選項」、「顯示」、「跳至」和「說明」的下拉式功能表，您可以進入「樹狀可視化工具」的所有功能。該「檔案」功能表對大多數 MineSet 工具是一樣的，請參閱第 92 頁的「[檔案功能表](#)」。

檢視功能表

查看功能表（在 IRIX 系統中為「顯示」功能表）包含四個選項：「全覽」、「搜尋面板」、「篩選面板」以及「標記面板」。為了與資料進行互動，每個選項都出現一個對話方塊。

搜尋面板

在「檢視」(或「顯示」)功能表中選擇搜尋會產生一個對話方塊，可讓您指定尋找物件的標準 (圖 1-38 和圖 1-39)。

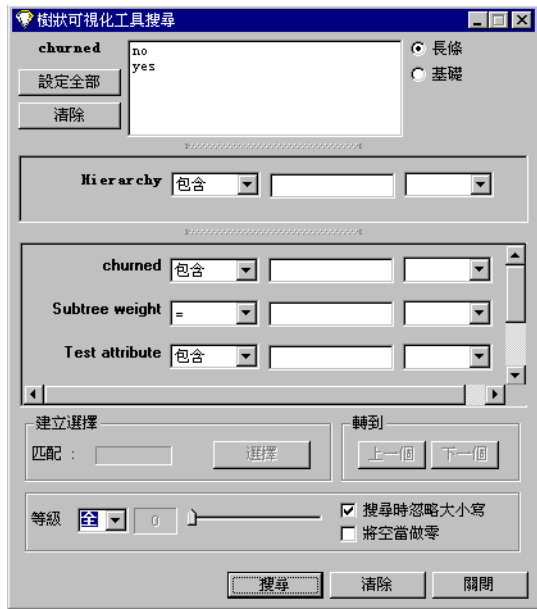


圖 1-38 樹狀可視化工具搜尋對話方塊 (Windows)

您可以指定要搜尋的等級。預設情況下，會對整個等級進行搜尋。要限制所搜尋的層，從選項功能表中選擇相關操作符 (例如 <=)，再指定層的操作數，使用「分層」滑動桿來選擇要進行搜尋的層。第 0 層是根的等級，第 1 層是其下面的一層，以此類推。例如，要選擇根與其下面的兩層，可選擇 <= 2。

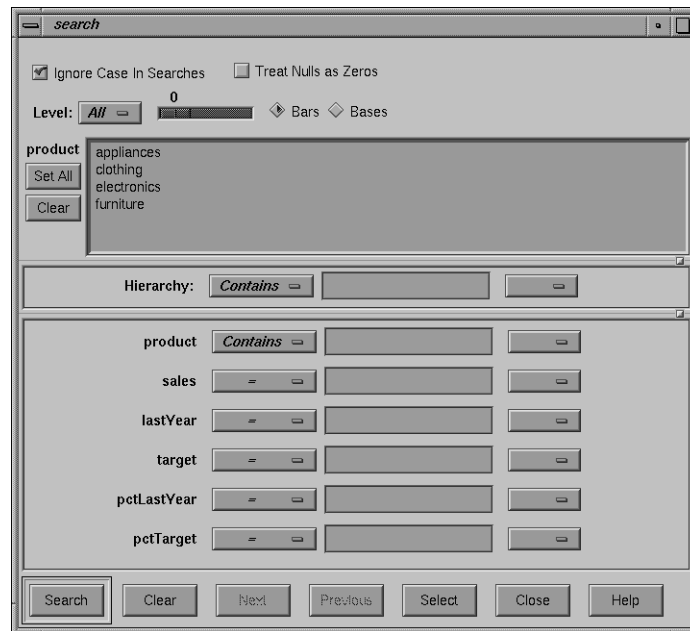
您也可以選擇是搜尋長條圖還是基準塊。

「等級」欄位可讓您指定要搜尋的節點。「等級」欄位下面是讓您為單獨欄指定搜尋標準的欄位 (在「目前欄」中定義：「工具管理員處理」窗格中的視窗。

要指定搜尋是否區分大小寫，按一下「搜尋」面板中的*在搜尋中忽略大小寫*核取方塊。例如，如果切換開關是開啓的（在該按鈕上有一個核選標記），則字串「Hello」與「hello」是相同的。

標有*將空值看作0*的核取方塊預設為關閉，在這種情況下，在搜尋中帶有空值的比較不能返回 TRUE。如果它是開啓的，則空被視為 0。

當對長條圖進行搜尋時，預設為在所有長條圖中進行搜尋。要搜尋指定的長條圖清單，您必須從列表中選擇它們。*設定全部*按鈕選中所有的長條圖；如果要對大多數長條圖進行搜尋，這是很有用的，您只需再取消選擇少量的長條圖即可。*清除*按鈕清除對所有的長條圖的選擇。如果沒有選擇長條圖，則忽略長條圖清單，並對所有的長條圖進行搜尋。



■ 1-39 樹狀可視化工具的搜尋對話方塊 (IRIX)

要搜尋數字型值，輸入該值，並選擇相關操作 (=, !=, >, <, >=, <=)。要搜尋字母型值，輸入您要搜尋的字串。您可以使用三種字串比較類型中的一種：

- 「包含」表示包含適當的字串。例如，**California** 包含了字串 **Cal** 和 **forn**。
- 「相等」則需要正好符合的字串。
- 「符合」允許萬用字元：
 - 星號 (*) 代表任何數量的字元。
 - 問號 (?) 代表一個字元。
 - 方括號 ([]) 包住要配對的字元清單。

例如，**California** 與 **Cal***，**Cal?fornia** 和 **Cal[a-z]fornia** 相符。

在一些情況下 (通常與「工具管理員」中的分組有關)，出現值的選項功能表，而不是本文欄位。要忽略一個變數，選擇「選項」功能表中的「忽略」。您可以使用選項中的相關操作符 (例如 >=)。這意味著指定值以及其後序值都已被選取。

除了數字型和字串型的比較操作以外，您還可指定為空操作，當值為空時則為真。

每個欄位的右邊是附加選項功能表，用該功能表可以指定「和」或「或」操作選項。例如，您可以指定銷售 >20 和 < 40。已知欄可以有多个「And」或「Or」子句，但是不能在單一欄中同時使用「And」和「Or」。

如果等級的不同層次有不同資料類型的關鍵字例如，頂層通過字串進行選擇，而第二層通過整數型進行選擇)，則「等級」搜尋欄位會視為字串並提供字串操作而不是數字型操作。

如果選擇搜尋中忽略大小寫核取方塊，則所有的字串搜尋比較都會不區分大小寫。

「搜尋」面板底部的按鈕包括一些或全部的下述功能：

- **搜尋**讓搜尋過程啟動。當面板在激活狀態並在按下 **Enter** 鍵後，該按鈕自動會啟動。
- **清除**關閉所有的搜尋聚光燈，並清除搜尋欄位中的值。
- **下一個**按照從左到右的順序選擇並縮放聚焦到下一個符合的物件。在選取最後一個符合的物件後，按一下 **下一個**將檢視返回至「首頁」的位置。只有在找到有符合的物件之後，**下一個**才有效。
- **上一個**使用與 **下一個**按鈕的相反順序進行選擇和縮放。
- **選擇**選擇所有與搜尋標準相符的物件。然後，可通過「選擇」功能表與這些物件進行交互。
- **關閉**關閉搜尋視窗並關掉聚光燈。如果「搜尋」面板重新開啓，則它所處的狀態與最近一次關閉之前的相同；再次按一下 **搜尋**則重複上次的搜尋。

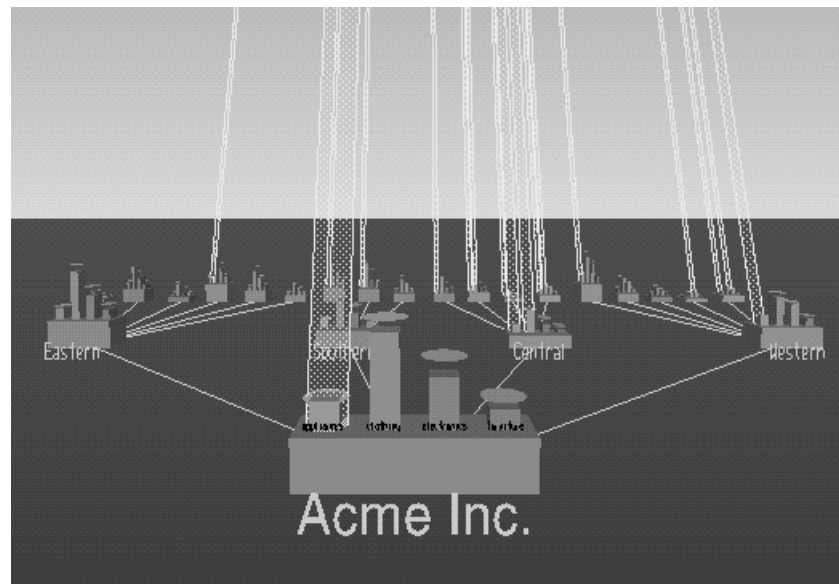


圖 1-40 在「樹狀可視化工具」中搜尋的「結果樣例」

一旦完成搜尋，黃聚光燈就會反白標示與搜尋標準相符的物件。(請參閱圖 1-40)。要顯示黃色聚光燈下的物件資訊，將箭頭移到聚光燈上，則資訊會顯示在左上角，在標籤游標位於：之下。要選擇並縮放黃色聚光燈下的物件，用滑鼠左鍵按一下聚光燈；如果您在按下時也按了 **Control** 鍵，則不進行縮放。

篩選面板

「篩選」面板篩選所選的資訊，可以微調顯示的等級。您可以使用「篩選」面板來強調指定資訊，或者減少資料量以求更好的顯示效果。對於大多數的 MineSet 工具，「篩選」面板是相似的，請參閱第 95 頁的「篩選面板」。

「樹狀可視化工具篩選」對話方塊中的欄位遵循在「搜尋」對話方塊中使用的相同約定。如果選擇在篩選中忽略大小寫核取方塊，則所有的字串篩選比較都區分大小寫。

如果節點不滿足篩選標準，沒有長條圖滿足標準，或沒有子節點滿足標準，則該節點不顯示。然而，會出現特殊的物件滿足篩選標準和情況，但是它在樹結構上的前輩不是這樣。同樣地，在相同節點中的其它長條圖也許不滿足該標準。因為這個位置在解釋上下關係時尤其重要，因此最好不要刪除這些長條圖。因此，您可以在控制這些物件應該如何繪製的三個單選鈕中，選取一個：**填色**、**輪廓及隱藏**。然而，請注意：如果物件由於「顯示 0」或「顯示空」功能表而以不填色的形式繪製，則它們遵循原有設定。例如，如果「空」是隱藏的，則不管篩選標準為何，它們永遠都是隱藏的。

對指定長條圖進行篩選時是一個例外。在這種情況下，會刪除其它長條圖且不占用空間，不管單選鈕如何設定。

該「高度滑動桿」可讓您篩選出只包含短長條圖的節點。該值的大小表示為最大值的百分比。首先，計算畫面中最高的長條圖（如果高度是按照層級來標準化，即每一層級中最高的長條圖）。那麼只顯示那些至少包含一個高度不小於最大高度的指定百分比的長條圖的節點。

例如，如果您在欄位中輸入 5%，那麼就會顯示包含至少一個長條圖的節點，而該條狀高度不低於最高條狀高度的 5%。(也顯示這些長條圖的祖先)。這一選項可以粗略的篩選出小的，不感興趣的節點。這並不是識別帶有一定值的特殊節點的精確方法；應該使用搜尋面板。使用該選項可以加快複雜畫面的顯示或限制 0 值附近的長條圖所帶來的雜亂。您也可以使用「高度篩選」指令來在配置檔案中設定篩選選項。

雖然篩選出小的節點，在層級上的任何計算中仍然要對它們進行計數。

在「高度篩選」滑動桿下面的「深度」滑動桿可讓您顯示層級，因此在任何時候只能顯示一定數量的層。當您在根等級時，只能看見用滑動桿指定的層數。列中的節點進行調整來最佳化其可視性。當在等級中移動到較低的節點時，附加列自動變為可見的。其上的節點自動調節它們的位置來容納新加入的節點；這樣，一些節點也許就要移動。請注意：總覽顯示等級中的所有節點；並不僅是頂部節點；這樣，總覽的配置也許就與主檢視的配置不相符。總覽中的 X 近似於主檢視中的對應位置；在兩個配置中不存在精確的對照。

要啟動篩選面板，按一下 **篩選** 按鈕。如果當面板活動時，您按下 **Enter** 鍵，就會自動啟動篩選操作。要關閉面板，按一下 **關閉** 按鈕。

樹狀可視化工具選項功能表

「選項」功能表可以對基本資料進行追溯。該功能表有五個項目。

- **顯示值**顯示所有選取物件值的表（記錄檢視器）格式。
- **顯示原始資料**檢索並顯示與已選取實體對應的記錄。結果記錄顯示在表格檢視器中。
- **傳送至工具管理員**根據目前的選擇方塊，在「工具管理員」歷史的起點處插入一個篩選操作。用於追溯的實際表達式由目前方塊選擇的範圍來確定。如果沒有選取任何實體，將出現一個警告訊息。

- *互補追溯*影響原始資料和傳送到工具管理員選擇，當使用時，取得所有未選取的資料。
- *標準化子樹*確定子樹中成員的最大高度，並且對與高度相關的所有值進行標準化。

如需追溯的進一步說明，請參閱第 78 頁的「追溯」。

樹狀可視化工具顯示功能表

「樹狀可視化工具顯示」功能表可讓您控制幾個顯示參數。

- *基準塊高度*是一個核取方塊，可以打開和關閉基準塊高度。要觀察負值或使之易於與長條圖高度比較，可將該選項關閉。將它開啓會提供所有長條圖的匯總資訊。用配置檔案中的「基準塊高度」語句可改變該切換的初始值。
- *標號標幟*是一個切換選項，可讓您開啓或關閉代表標記的標號（請參閱《*MineSet 3.0 for Windows 企業版使用者指南*》中的「標號面板」部分）。
- *零值*是一個子功能表，可以控制高度為 0 的物件如何顯示。預設情況下，它們與其它物件一樣地顯示：高度為 0 的立方體（平面）子功能表可讓它們顯示為輪廓（顯示為中空的正方形）或完全隱藏（不畫出）。利用配置檔案中的 0 選項可以改變初始值。
- *空值*是一個子功能表，可以控制高度為空的物件如何顯示。它與 0 功能表有相同的選項，但是空選項是預設值是將物件顯示為輪廓。利用配置檔案中的空選項可以改變初始值。

跳至功能表 (樹狀可視化工具)

「跳至」功能表重複主視窗右項上面的按鈕功能。它也指定一些功能的鍵盤快速鍵。

- *首頁*帶您到一個指定的位置。預設情況下，該位置為畫面的初始檢視點。初始時，該位置是在啟動「樹狀可視化工具」和指定配置檔案之後所顯示的第一個視點。如果您利用「樹狀可視化工具」進行工作並按一下*設為首頁*功能表項目，然後按一下*首頁*，則將返回您最後一次按下*設為首頁*時的視點位置。該功能的鍵盤快速鍵為 **Control+H**。
- *設定首頁*將「首頁」位置改變至目前位置。按一下*首頁*功能表項目，然後返回當您最後一次按下*設為首頁*時的視點位置。
- *全部檢視*顯示整個結構，保持畫面的傾斜度。要得到該畫面的俯檢視，傾斜畫面直向下，然後按一下*檢視全部*功能表項目。
- *向後*可讓您返回至前一個位置。如果您剛啟動「樹狀可視化工具」且未從初始位置移動，則該功能表項目變成灰色。該功能的鍵盤快速鍵為 **Control+B**。
- *向前*可讓您向前到您按一下*向後*功能表項目時的位置。如果您沒有按一下*向後*功能表項目，則*向前*功能表項目變成灰色。該功能的鍵盤快速鍵為 **Control+R**。
- *父節點*只有在選取物件時才活動。如果選擇長條圖，按一下該功能表項目便會選擇包含長條圖的基準塊。如果選擇基準塊，按一下該功能表項目便是將等級移到父節點上。一旦到達根節點 (等級的最高層)，*父節點*功能表就變灰。該功能的鍵盤快速鍵為 **Control+U**。
- *左移*可讓您向左選擇相鄰兄弟。如果選擇長條圖，便是選擇了其左邊的長條圖。如果選擇基準塊，如果父節點在左邊還有一個子節點，則也會選取該子節點。如果沒有選擇或者目前選擇在左邊沒有兄弟節點，則該按鈕變灰。
- *右移*可讓您向右選擇下一個兄弟節點。如果選擇長條圖，便是選擇了其右邊的長條圖。如果選擇基準塊，如果父節點在右邊還有一個子節點，則也會選取該子節點。如果沒有選擇或者目前選擇在右邊沒有兄弟節點，則該按鈕變灰。

- *第一個子節點*可讓您選擇目前節點的第一個子節點。如果沒有選擇，或選擇的長條圖或目前的選擇沒有子節點，則該按鈕變灰。
- *最後的子節點*可讓您選擇目前節點的最後子節點。如果沒有選擇，或選擇的長條圖或目前的選擇沒有子節點，則該按鈕變灰。

說明功能表

大多數 MineSet 工具的「說明」功能表是一樣的，請參閱[第 97 頁的「說明 \(IRIX\)」](#)。

在樹狀可視化工具中的空值處理

空代表未知值 (請參閱 [《MineSet 3.0 企業版介面指南》](#) 的「MineSet 中的空」)。

在「樹狀可視化工具」中，在下列情況下會出現空：

- 資料庫或資料檔案包含空值。
- 在配置檔案中沒有「跳過遺漏」選項 (請參閱 [《MineSet 3.0 企業版介面指南》](#) 的「為樹狀可視化工具建立資料和配置檔案」中的「跳過遺漏」) 以及資料表示等級中一個節點提供關鍵值，而對另一個卻沒有提供。例如，在代表州財政的過程中，如果沒有關於 Texas 州收入稅的記錄，則 Texas 的收入稅為空。這與一個記錄顯示 Texas 的收入稅為 0 不同，因為在這種情況下它顯示的是稅為 0。
- 當使用「工具管理員」在分組的基礎上來產生一個陣列並且沒有資料落入指定分組中，此時該組對應的值為空。例如，30-40 歲的年齡段沒有資料，則分該組為空。
- 當在「工具管理員」中產生陣列並指定空枚舉選項時，則建立與每個長條圖圖中的第一個長條圖對應的特殊陣列項目，用此代表空組中所有值的組合。(該長條圖標以問號(?)，代表空。如果空組中沒有資料，則與其相關的值也為空。

注意：如果整個資料中與空組相關的所有值為空，則「樹狀可視化工具」會忽略空組並不會顯示它。

- 空值的表達式和組合可以產生空值。

當空值對照為可視化屬性時，在「樹狀可視化工具」中使用特殊代表物。如果空對照為高度，則該物件以輪廓模式進行繪製（雖然可經由「顯示」功能表或配置檔案對其進行配置（請參閱第 192 頁的「樹狀可視化工具顯示功能表」）（《MineSet 3.0 企業版介面指南》的「為樹狀可視化工具建立資料和配置檔案」對長條圖或基準塊而言，看起來像一個空的正方形。（它並不像立方體，因為它沒有高度。）對圓盤而言，它看起來像一個圓圈。如果空值對照為顏色，則它被繪製為深灰色（請參閱圖 1-41）。

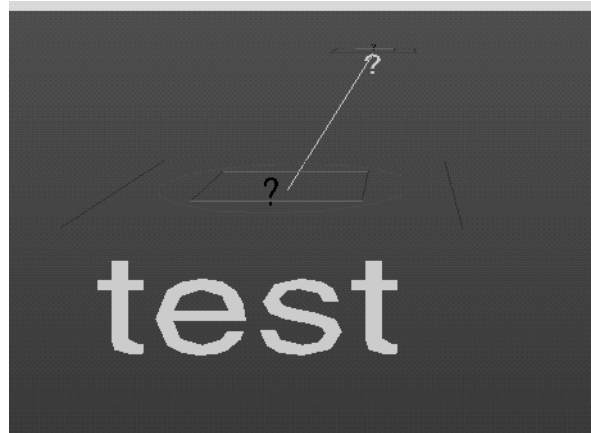


圖 1-41 代表對照到高度、顏色、圓盤及標籤的「空值」

當選擇具有空值的物件時，問號 (?) 會顯示在選擇欄位中。

樹狀可視化工具限制

至於「樹狀可視化工具」，「工具管理員」不支持下列項目：

- 資料未經組合就直接顯示的非組合等級結構
- 即時監控。
- 一些不常用的選項（跳過遺漏、總覽、收縮、根標籤、速度、爬升速度、葉空白、根葉空白、葉邊緣空白、初始位置、初始角度、長條圖標籤大小、基準塊標籤大小以及 lod）。

- 變數長度陣列。
- 在建立等級之後計算的表達式。例如，如果您想計算一個百分比，該百分比必須在等級組合過程之後才能計算，因為不可能對百分比進行組合。

樣例配置和資料檔案

所提供的配置檔案和資料檔案展示「樹狀可視化工具」的特性和功能。請參閱[附錄 A](#)，「[配置和資料檔案樣例](#)」，其中有使用「樹狀可視化工具」將「決策」，「選項」或「回歸樹」可視化的範例。

修改因子

修改因子是「進階分組」選項，它允許您在啟動分組操作之前將極端值（稱作離群值）從資料集中排除。預設的修改因子為 0.05。這就排除了 5% 具極值的實例（2.5% 具有範圍內的最低值而 2.5% 具有最高值）。修改的目的是減小在臨界值產生過程中離群值的影響。

統一範圍

統一範圍是在自動資料分組中使用的一個選擇，在資料分組過程中值範圍分為統一大小的子區間。請參閱[第 34 頁](#)的「[分組](#)」可得到使用「工具管理員」的欄分組按鈕進行應用的完整討論。

統一權重

統一加權是在自動資料分組中使用的一個選擇，在資料分組過程中的值範圍分為一定數量的等權重組，因此每個組包含相同數量的實例，如果每個實例的權重為 1，則每組的權重就相等。請參閱[第 34 頁](#)的「[分組](#)」可得到使用「工具管理員」的欄分組按鈕進行應用的完整討論。

檢視功能表

「檢視」功能表可控制不同的顯示選項，對於大多數可視化工具它是相似的。根據不同的平台，功能表包含下列一些或所有選項：

- **篩選面板**根據一個或多個標準，您可以減少在主要檢視區域中顯示的實體數目。您可以使用篩選面板來微調顯示設定，強調特別資訊或縮減顯示資訊的數量。在篩選面板右下方的「調整比例」核取方塊可以指定主視窗中的場景是覆蓋了整個資料集還是只有篩選的資料。請參閱第 95 頁的「**篩選面板**」填入該對話方塊，或將滑鼠在欄位上移動並按下 Shift 和 F1 鍵獲得幫助。
- **設定背景色**產生一個顏色選擇器，讓您指定一個新的背景顏色。
- **顯示視窗控件**可讓您隱藏或顯示在畫面視窗周圍的外部控制。
- **空位置**切換空值的顯示。
- **動畫面板**可讓您顯示或隱藏動畫控制面板。對於不具有獨立維度的資料集，該功能表項目停用。

可視化過程工具

該部分提供經 MineSet 「工具管理員」可找到的可視化過程工具的總覽，您可以使用不同的可視化範例來查看您的資料：

- **聚類可視化工具**顯示聚類或群組的統計結果，並將統計結果與整個資料集的聚類結果放在一起，這樣您便可以了解是何種功能將各類區分開來。
- **決策樹狀可視化工具**可讓您從分級的欄中觀察資料。例如，您可以根據產品類別、地理位置、業務推銷以及業務代表津貼計劃來檢視獲利率。資料每次分配兩個屬性，所以您可以細化下尋深入到每層的屬性組中。

- **證據可視化工具**可讓您了解特定的屬性值是如何有助於預測指定標籤的機率。例如，如果在蝴蝶花資料集中，您看到屬性花萼長度為 5.45...5.85，那麼您可以看到標籤為 iris-setosa 的機率為 86.54%。當「證據導入工具」建立模型時，「證據可視化工具」顯示不同屬性是如何貢獻於所產生的決策，並且允許「What-If」分析。
- **柱狀圖可視化工具**將資料中的連續類型的欄自動分組，並將結果傳送至統計可視化工具，以柱狀圖的形式顯示。
- **地圖可視化工具**可讓您可視化存在地理關聯性的資料。例如，您可以將不同國家區域可視化，顯示市場計劃的相關效應。「地圖可視化工具」的細化下尋功能可讓您將焦點集中在指定區域，並且在更小的地理單元中進行更細致的分析。
- **記錄檢視器**可讓您在類似試算表的列和欄中查看資料。
- **分散可視化工具**在 1、2 或 3 維空間中顯示資料點。附加屬性可對照為尺寸、顏色及形狀。最後，其他兩個屬性可以對照到滑動桿，在總共 8 維空間裡，允許動畫和閃現。當您移動滑動桿時，顯示的內容也隨之變化來反映獨立變數中的變化。「分散可視化工具」也用於顯示「關聯規則」。
- **平板可視化工具**與「分散可視化工具」在很多方面有相似之處，區別只是在顯示資料密度時使用了透明度。當需要表達大量的資料卻不需要畫出每個點時，該工具非常有用。
- **統計可視化工具**計算並顯示目前資料集的匯總資訊（最大值、最小值、中值、標準差、相異值以及四分之一值）。
- **樹狀可視化工具**在用等級關係分析資料的過程中很有用。交互式的「閃現」方法允許您在不同的等級層上檢查資料關係。例如，「樹狀可視化工具」可用來檢查公司的生產線，圖形顯示每個產品對公司收入的貢獻。等級的每個分支在子層上顯示越詳細的資訊，根據生產線與（最終）個體產品顯示收入狀況。

如果每個決策由樹中分離的節點來代表，則「樹狀可視化工具」也用於檢視「決策樹」、「選項樹分類器」及「回歸樹回歸工具」的結果模型。

警告選項

在 IRIX 中，當在指令行上進行操作時，有兩個 MineSet 選項會影響所有工具如何啓動：

- **-warnexecute** 指出如果您試圖執行一個在執行語句中指定的指令，則會出現一個警告，而您可以選擇執行或不執行指令。這是著眼於系統的安全而設計的，例如，從 Web 得到的檔案，並且當利用 mtr 檔案來執行指令時自動啓用。(請參閱 *《MineSet 3.0 企業版介面指南》* 得到關於 mtr 檔案的解釋。)

您可使該選項永久可用

— 在 IRIX 系統上經由增加：

```
*minesetWarnExecute:TRUE
```

到您的 .Xdefaults 檔案或

— 在 Windows 系統中經由改變值：

```
MINESET_WARN_EXECUTE
```

在註冊項中。您可以使用檔案 特性對話方塊來完成這些。

- **-quiet** 取消跳現出來顯示過程的對話方塊。加入一行可以使該選項永久可用

```
*minesetQuiet:TRUE
```

到您的 .Xdefaults 檔案。

在 Windows 中，這些選項在「3D 可視化工具」的「特性」對話方塊中可找到(在「檔案」功能表中可找到)。

Web 發布

MineSet Web 延伸允許在網上將 MineSet 軟體產生的檔案和資料可視化。MineSet 的 .mtr 附檔名可讓您將 MineSet 的配置、方案及資料檔案放入壓縮檔中，並作為 html 標籤插入網頁中。當該頁載入至 Internet Explorer 時，在瀏覽器視窗中可以啓動 MineSet 可視化工具。執行瀏覽器的機器必須安裝 MineSet 客戶端軟體。

如需範例和安裝說明，請參考 [《MineSet 3.0 企業版介面指南》](#)。

加權

通常權重代表實際記錄的計數，但是當資料集進行不均勻採樣時，一些記錄經由加權，也就是增加其重要性，可能會賦予更多的重要性。可以在「工具管理員」的「資料目標」窗格中選擇「挖掘工具」來加權記錄。如需任何導入工具，請按一下 [更多選項](#) (或 [進階選項](#)) 按鈕，以取得具有「使用加權」核選框的對話方塊。記錄加權的意義是權重為 2 的記錄與兩個權重為 1 的記錄相等。並且允許浮點型的權重。也請參閱 [第 138 頁的「記錄加權」](#)。

適應 Y2K 問題

MineSet 支援 Y2K 日期。對於美國地區，日期的輸入形式可能為 MM/DD/YY 或 MM/DD/YYYY。MineSet 對於 2 數位年份遵循 X/Open 標準：大於 68 的 2 數位年份可定為 1969 年到 1999 年，而小於或等於 68 的 2 數位年份定為 2000 年到 2068 年。

對於歐洲地區，日期的輸入形式可以為 DD/MM/YY 或 DD/MM/YYYY，對於 2 位數年的處理方法同上。

在其它地區，如果要輸入 2 數位年份，則在顯示中會自動擴展為 4 數位年份。

配置和資料檔案樣例

MineSet 中的多個配置和資料檔案樣例可以展示多個工具的功能。在這個部分將詳細描述和解釋每個工具的樣例檔案，具體項目按工具的字母順序排列。

- Windows 使用者可在 MineSet 安裝目錄中 *\examples* 下找到樣例檔案。
- IRIX 使用者可以在 */usr/lib/MineSet/examples* 中找到檔案

檔案描述如下：

- 第 202 頁的「關聯規則樣例檔案」
- 第 202 頁的「聚類樣例檔案」
- 第 203 頁的「欄重要性樣例檔案」
- 第 204 頁的「決策樹樣例檔案」
- 第 213 頁的「決策表樣例檔案」
- 第 227 頁的「證據可視化工具樣例檔案」
- 第 238 頁的「地圖可視化工具樣例檔案」
- 第 240 頁的「選項樹樣例檔案」
- 第 243 頁的「回歸樹樣例檔案」
- 第 247 頁的「分散可視化工具樣例檔案」
- 第 250 頁的「平板可視化工具樣例檔案」
- 第 252 頁的「樹狀可視化工具樣例檔案」

關聯規則樣例檔案

下面提供的資料和配置檔案樣例將在已預備資料集的基礎上對「關聯規則」進行可視化處理。其中一些檔案與等級資料集相對應。規則檔案包含經由執行「關聯規則產生器」得到的規則。按照慣例，包含規則的檔案應該具有 *.rules.data* 附檔名。每個配置檔案說明對應規則檔案是如何顯示。配置檔案必須有 *.scatterviz* 附檔名。在下面提到的檔案在安裝 MineSet 的目錄 *\examples* 中，對於 IRIX 用戶，則在 */usr/lib/MineSet/scatterviz/examples* 目錄下。

- *group.rules.data* 和 *group.rules.scatterviz*
這些檔案為每個產品組，例如，麵包和薰烤食品，牛奶以及碳酸飲料等提供產生的規則和配置規範。
- *category.rules.data* 和 *category.rules.scatterviz*
這些檔案為產品群組中的產品類別（例如冷藏或未冷藏牛奶）提供產生的規則和配置規範。
- *adult94.rules.data* 和 *adult94.rules.scatterviz*
這些檔案為調查資料庫提供產生的規則和配置規範，顯示婚姻狀況、受教育水平、年齡收入以及其它變數之間的關聯。
- *germanCredit.rules.data* 和 *germanCredit.rules.scatterviz*
這些檔案為德國的信用資料集提供產生的規則和配置規範，顯示與信用歷史、職業、存款及其它變數之間的關聯。
- *cars.rules.data* 和 *cars.rules.scatterviz*
這些檔案為轎車資料集提供產生的規則和配置規範，顯示不同屬性之間的關聯。

聚類樣例檔案

下面的樣例是聚類發揮作用的範例。該樣例與 MineSet 所提供的樣例資料集相關。它顯示如何利用「聚類」挖掘工具，並解釋不同的結果和選項。

轎車資料集比較簡單，主要探討一些熟悉的概念，如馬力、車重以及達到 60 mph 所需的時間。

當「聚類可視化工具」出現時，屬性根據它們在區分類的過程中作用的大小從上到下進行排列。

如果您選擇聚類 1，則該聚類控制由長條圖和柱狀圖所代表的屬性的排序優先權。在其它聚類中的屬性順序仍然以聚類 1 為基礎。例如，如果您按一下聚類 1，屬性序列為汽缸、重量，然後是每加侖英哩數。排序中的變化將僅出現在可視化過程中，而不是在基礎資料集中。您可以比較其它聚類的相同列，觀察屬性在不同聚類之間的差異。當您選擇聚類 2 時，您可以在更低的層上觀察到一個不同的屬性次序。在此種情況下，產地最重要，然後是汽缸、馬力、每加侖英哩數。

欄重要性樣例檔案

下面是「欄重要性」發揮作用的範例。該樣例與 MineSet 所提供的樣例資料集相關。它顯示如何使用「欄重要性」挖掘工具，並解釋不同的結果和選項。

當用戶更改他們的電信業務供應商時，這個術語稱為「客戶波動」，這是電訊產業中共同的問題。檔案 *churn.schema* 和 *churn.data* 即為這個範例。Windows 使用者可以在 MineSet 安裝目錄下的 *\data* 目錄中找到該範例。IRIX 使用者可以在 */usr/lib/MineSet/data* 中找到該範例。

簡易執行「欄重要性」模式產生下列三個屬性：

- 白天總費用。
- 客服電話數目。
- 州。

從進階模式中執行「計算改良純度」，您可以看到「白天總費用」和「白天總分鐘數」具有相同的純度級別 (48.67)。將其中之一移到右邊欄中 (例如，白天總分鐘數) 並重新執行「計算改良純度」，您會發現另外一個 (白天總費用) 沒有資料。有兩個屬性高度相關。

當「白天總分鐘數」在右邊一欄時觀察該屬性值，我們可以看到下列：

- 國際計劃 (4.1)
- 客服電話數目 (8.1)。
- 州 (4.7)

您可以選擇將「國際計劃」移向右邊，因為該資訊可迅速使用並容易測量。

另外兩個屬性 (客服電話數和州) 仍然非常重要 (實際上，它們的重要性在增加)，所以它們表現出與「國際計劃」不相關。

透過此種方式觀察屬性的重要性，您會決定何種屬性可以由同樣好 (或接近同樣好) 的來取代，但是更易於測量或理解。透過觀察純度，您可以確定附加屬性能起多大作用。例如，在上述的方案中，州明顯地提高了純度。在 *iris* 資料集中，所選擇的第三個屬性 (花瓣長度) 只是稍微提高了純度，簡單的，2D 分散圖將提供與 3D 分散圖幾乎一樣多的資訊。

決策樹樣例檔案

以下是「決策樹」導入工具發揮作用的範例。每一個範例都與 MineSet 所提供的樣例資料檔案相關。執行導入工具，您可以產生如下所述的 *-dt.treeviz* 檔案。

可以開啓相對應的 *.schema* 檔案而將資料檔案載入 MineSet，而檔案儲存在 *data* 目錄中 (例如，*churn.schema*)。可以從 *examples* 目錄中開啓具有 *-dt.treeviz* 附檔名的分類工具可視化檔案。

- Windows 使用者可以在 MineSet 安裝下的資料和樣例目錄下找到這些檔案。
- IRIX 使用者可以在 *data* 和 *examples* 中找到這些檔案，而上述兩個目錄在 */usr/lib/MineSet/treeviz/examples* 中。

客戶波動

當用戶將他們的電話電信業務從一個電訊公司轉向另一個，這個術語稱為「客戶波動」，這是電訊產業中共同的問題。在 *examples* 目錄中的檔案，*churn-dt.treeviz*，顯示一個為解決該問題而導入的「決策樹」分類器。通過在 *data* 目錄中檔案 *churn.schema* 上執行導入工具可以產生此檔案，注意其標籤設為客戶波動 (是、否)。該檔案是虛擬的，但卻是根據實際資料的模式而建立。

在該樹中，根的分割是在白天用戶交談的時間量 (白天總分鐘數) 為基礎進行的。每天交談超過 264 分鐘的用戶之客戶波動率比沒有超過的用戶明顯地高很多 (60% 比 11%)。這些很可能是最有利潤的用戶。

左邊的子樹代表每天交談少於 264 分鐘的用戶，他們的客戶波動率為 11%；但是如果他們打出超過三個的客服電話，則客戶波動率增至 49%。

右邊的子樹代表每天交談多於 264 分鐘的用戶。他們的客戶波動率為 59%；但是如果他們有聲音郵件計劃，則客戶波動率降為 9.3%。如果他們沒有聲音郵件計劃，則客戶波動率幾乎為 75%。

轎車產地

轎車資料集包含從 70 年代到 80 年代不同型號轎車的資訊。屬性包括重量、加速能力以及每加侖英里數 (mpg)。在 *examples* 目錄中的檔案，*cars-dt.treeviz*，顯示一個為解決該問題而導入的「決策樹」分類器。通過在 *data* 目錄中的 *cars.schema* 檔案上執行導入工具可以產生該檔案，注意其標籤設為產地 (日本、美國、歐洲)。如果您有一個關於轎車屬性的資料集，您會想知道不同產地轎車的特徵。

- Windows 使用者可以在安裝 MineSet 的目錄中 *examples\cars-dt.treeviz* 和 *\data\cars.schema* 下找到這些檔案。
- IRIX 使用者可以在 */usr/lib/MineSet/treeviz/examples/cars-dt.treeviz* 和 */usr/lib/MineSet/data/cars.schema* 中找到這些檔案

請注意：樹中左邊的分割是關於品牌，而根部的分割不是關於品牌的，這是因為「決策樹」導入工具較不支持多重分割，而視在引擎體積上的分割為更好的區分。您可以使用「工具管理員移除欄」功能來隱藏品牌，這樣使得問題更加有趣。

在「決策樹」中，您可以看到立方英吋對美國產轎車引擎體積是一個非常好的區分工具。有大引擎 (>169.5 立方英寸) 的轎車全部都是美國製造的，但是較小的轎車卻產自各地。選擇「選項」>「顯示原始資料」，您可以觀察到不是美國產的大引擎轎車是 Mercedes。請注意：在樹中的根節點 (也就是，整個訓練資料集) 有更多的美國轎車 (62.50%)，然而在體積屬性上進行一次分割後，很難預測具有小引擎轎車的產地。根的純度為 16.2 表明只有一個分類占多數 (在這種情況下為美國)。右節點 (> 169.5 立方英寸) 的純度為 96.81，表示我們已經找出非常純的子集 (幾乎所有具有大引擎的轎車都產自美國)。實際上，右邊子樹的誤差率估計為 0% (綠色基準)。根左邊節點的純度為 0.23 並且具有更高的誤差率 31.25% (橙色基準)。該子問題比最初的一個更難解決：每個分類的記錄數目近似一樣。

預測性別

成人資料集包含職業成人的資訊。該資料集是從美國調查局中所摘錄，它包含年齡超過 16 歲、每周工作至少一小時、每年收入超過 \$100 的人之相關資訊。您也許想知道怎樣將男性和女性特徵化，檔案 *adult-sex-dt.treeviz* 顯示為解決該問題而導入的「決策樹」分類器，該檔案在 *adult.schema* 上執行導入工具而產生，並且其標籤設為 *性別*。該資料集包含幾乎 50,000 筆記錄；所以當您在工作站上執行「決策樹導入工具」時將要花費幾分鐘的時間。

- Windows 使用者可以在 MineSet 安裝的目錄 `\examples\adult-sex-dt.treeviz` 和 `\data\adult.schema` 下找到這些檔案。
- IRIX 使用者可以在 `/usr/lib/MineSet/treeviz/examples/adult-sex-dt.treeviz` 和 `/usr/lib/MineSet/data/adult.schema` 中找到這些檔案

這些可視化過程結果提供了以下直觀結果：

- 婚姻關係是最重要的屬性。丈夫通常是男性。(有趣的是，有一位丈夫是女性，這顯示調查局的資料品質問題，並沒有意識到這是同性之間的婚姻。)同樣地，如果是妻子，她通常是女性，除了三個記錄例外。

為了使問題更加有趣，移除相關屬性並產生一個新「決策樹」。在此種情況下：

- 最重要的屬性就是婚姻狀況。
- 從基準塊的高度來看，大多數人爲已離婚、與一般公民結婚或從未結婚。只有少數是喪偶、分居、與軍人結婚、或離異。
- 在資料集中根節點的分布顯示集中男性居多。(該資料集包含就業成人的資訊而非代表整個人口。)
- 最左邊的節點包含離婚職業成人。我們可以看到這裡的分布比根部更加平衡 (60% 爲女性，40% 爲男性)。第二個節點包含已婚職業成人，我們可看到 89% 是男性。第三個節點包含從未結婚的職業成人，他們的數量大致與離婚組的相同，男性略多一點。最右邊的節點包含喪偶的職業成人，81% 爲女性 (也許是因爲她們較高的期望值)。「喪偶」一詞表示失去配偶的任何一方。

如果您想將新產品的目標集中在職業女性上，可以使用查詢面板來找出含有大量女性人口的部分。您可以選擇

- 性別與女性相符 (按一下視窗頂部的女性)
- 子樹權重 >1000
- 百分比爲 > 80

三個黃色聚光燈顯示相符節點。因爲兩個節點在同一條路徑上，所以觀察離根最近的節點 (在右邊)。該路徑翻譯爲規則

婚姻狀況 = 「喪偶」顯示其中 81.23% 爲女性

婚姻狀況 = 離婚和職業 =
管理文書暗示 87.67% 爲女性

在該訓練集中，在根部的 16,192 中，有 1233 (離異) 和 1045 (離婚和職業) 個女性滿足這些規則。該單一部分包含資料集中超過 14% 的職業女性的資訊。

薪資因素

如果您有一個職業成人的資料集，您可能想知道何種因素會影響薪資水準。您可以將記錄分為兩類：每年的收入在 \$50,000 以下或以上的。每筆記錄就有了一個屬性，該屬性具有兩個值的其中一個。" 50,000-" 和 "50,000+" 您可以執行 MineSet 分類器來確定是何種因素影響薪資。樣例檔案 *adult-salary-dt.treeviz* 顯示為了解決該問題而導入的「決策樹」分類器。在資料檔案 *adult.schema* 上執行導入工具，以及在使用者定義臨界值 50000 的基礎上對 *gross_income* 進行分組可以得到該檔案，注意標籤設定為 *gross_income_bin*。

- Windows 使用者可以在 MineSet 安裝的目錄 *\examples\adult-salary-dt.treeviz* 和 *\data\adult.schema* 下找到這些檔案。
- IRIX 使用者可以在 */usr/lib/MineSet/treeviz/examples/adult-salary-dt.treeviz* 和 */usr/lib/MineSet/data/adult.schema* 中找到這些檔案

這些可視化過程提供了以下直觀結果：

- 代表整個資料集的根節點顯示 76.07% 的職業成人年收入在 \$50,000 以下。
- 年齡是最重要的因素。27 歲以下的人中只有 3.07% 其收入超過 \$50,000。基準顏色為綠色，表示規則很準確的 (大約 3% 的誤差率)。
- 對於預測年齡超過 27 歲者的薪資，受教育水平是很重要的因素。調查局對每個人都指定了教育層次指標。「決策樹」在 12.5 處發生分割；層次 13 正好相對於學士學位。具有學士或更高學位者中約有 55% 其收入在 \$50,000 以上。
- 對於年齡大於 27 並且受到良好的教育的部分，婚姻關係是重要的薪資預測因素。對於那些已經結婚的人，賺得 \$50,000 的機會，對丈夫來說有 73%，對於妻子則有 75%。(但是，包含妻子的節點只有一個較小的基準塊，表示幾乎沒有女性符合該規則。) 如果在這組中的人沒有結婚，收入到達 \$50,000 以上的機會，男性會降為 27%，女性則為 25%。

蝴蝶花分類

在該資料集中，每個記錄都描述蝴蝶花朵的四個特徵：花萼寬度、花萼長度、花瓣寬度和花瓣長度。每個蝴蝶花進一步劃分為藍色蝴蝶花、多色蝴蝶花或純白色蝴蝶花類型。該目標就是理解如何按不同特徵來區分類型。

在執行分類器之前，按一下「工具管理員分類器」索引標籤中的「欄重要性」索引標籤；然後按一下繼續。您得到了特性重要性的排序，`petal_width`，`petal_length` 和 `sepal_length`。在「分散可視化工具」中，您可以將這些屬性對照到坐標軸上，`iris_type` 對照為顏色，並觀察所分的聚類。

檔案 `iris-dt.treeviz` 顯示為解決該問題而導入的「決策樹」分類器。在 `iris.schema` 上執行導入工具可產生該檔案。

- Windows 使用者可以在 MineSet 安裝的目錄中 `\examples\iris-dt.treeviz` 和 `\data\iris.schema` 下面找到這些檔案。
- IRIX 使用者可以在 `/usr/lib/MineSet/treeviz/examples/iris-dt.treeviz` 和 `/usr/lib/MineSet/data/iris.schema` 中找到這些檔案

執行「樹狀可視化工具」，您可以看到根節點有 6% 的誤差率，即使純度很低 (0)。純度測量分布的斜度，在根部的分布非常統一：每個標籤值有 50 個記錄。左邊的分支 (`petal-length <= 2.6` 英寸) 到達了只包含 `iris-setosas` 類型的綠色節點 (0 誤差)。另外一個分支利用 `petal_width` 上的檢驗，同樣也能夠很快地劃分類別。路徑 `petal-length > 2.6` 以及 `petal-width <= 1.65` 和 `petal-length > 5` 到達包含 4 個記錄的不純葉節點記錄。`iris-virginica` 類型有三個記錄，而 `iris-versicolor` 只有一個。「決策樹」並沒有分割該節點，因為被視為沒有意義 (預設情況下，每個分割必須包含兩個子節點其權重至少為 2)。節點顏色也為黑，表示沒有實例符合該節點，所以也沒有關於它的估計誤差率。

總結如下：花萼長度 ≤ 2.6 英寸的花預測為藍色蝴蝶花，那些花萼長度為 > 2.6 英寸或 ≤ 5 英寸以及花萼寬度 ≤ 1.65 英寸的預測為多色蝴蝶花，那些花萼長度為 > 2.6 英寸並且花萼寬度為 > 1.65 或花萼長度為 > 5 英寸並且花萼寬度為 ≤ 1.65 的預測為純白色蝴蝶花。

當使用「欄重要性」來判別資料時，因為「決策樹」在連續屬性上產生兩元分割，因此樹的根分割與欄重要性中的第一個屬性不同 (請參閱第 52 頁的「欄重要性」可得到更多的細節)。

蘑菇分類

檔案 *mushroom-dt.treeviz* 顯示為蘑菇分類過程而導入的「決策樹」分類器。在 *mushroom.schema* 上執行導入工具可以產生該檔案。

- Windows 使用者可以在 MineSet 安裝目錄中的 *\examples\mushroom-dt.treeviz* 和 *\data\mushroom.schema* 下找到這些檔案。
- IRIX 使用者可以在 */usr/lib/MineSet/treeviz/examples/mushroom-dt.treeviz* 和 */usr/lib/MineSet/data/mushroom.schema* 中找到這些檔案

分類目的是在指定的資料集上分辨哪些蘑菇可以食用而哪些有毒。在該資料集中有 8000 多筆記錄，執行該導入工具也許會花幾秒種。

每個蘑菇有很多特徵，包括菇帽顏色、褶紋和氣味。如果您建立「決策樹」分類器，您可以看到只用氣味屬性就可以對區分蘑菇是否有毒情況中的 50% 作出判斷。如果蘑菇沒有氣味，則有毒的機率為 3.4%。下一個屬性是莖的形狀。如果呈錐形，則蘑菇可以食用，但是如果越來越粗，則有 11.6% 的可能性是有毒的。有 1032 個蘑菇記錄到達該節點。您可以沿著樹往下到更深的節點來觀察其它屬性的作用。

黨派隸屬

該資料集包含投票記錄。目的是從關鍵投票資料的資訊中找出議員所屬黨派。資料集包括由 *議會季度年鑑 (CQA)* 所識別的美國眾議院每個議員在 16 次關鍵表決中的投票。CQA 列出九個類型的投票：投票贊成、配合贊成以及宣布贊成 (這三個為贊成)；投票反對、配合反對以及宣布反對 (這三個為反對)；交出選票、交出選票避免利益衝突以及不投票或持其它觀點 (這三個為未知處理)。

在執行分類器之前，觀察 16 個投票結果來檢視您是否能查覺最重要的特徵。然後執行「決策樹」分類器。

檔案 *vote-dt.treeviz* 顯示為解決該問題而導入的「決策樹」分類器。在 *vote.schema* 上執行導入工具可產生該檔案。

- Windows 使用者可以在 MineSet 安裝目錄中的 *\examples\vote-dt.treeviz* 和 *\data\vote.schema* 下找到該檔案。
- IRIX 使用者可以在 */usr/lib/MineSet/treeviz/examples/vote-dt.treeviz* 和 */usr/lib/MineSet/data/vote.schema* 中找到這些檔案

乳癌診斷

乳癌資料集包含有關婦女接受乳癌診斷的資訊。每筆記錄都包含病人細胞大小、凝塊厚度及邊緣附著力等屬性的記錄。最終的屬性是診斷為惡性或良性。檔案 *breast-dt.treeviz* 顯示為解決該問題而導入的「決策樹」分類器。在 *breast.schema* 上執行導入工具可產生該檔案。

- Windows 使用者可以在 MineSet 安裝目錄中 *\examples\breast-dt.treeviz* 和 *\data\breast.schema* 下找到這些檔案。
- IRIX 使用者可以在 */usr/lib/MineSet/treeviz/examples/breast-dt.treeviz* 和 */usr/lib/MineSet/data/breast.schema* 中找到這些檔案

該「決策樹」顯示儲存格大小的一致性是非常有力的區分屬性。根分布為 65% 對 35% (純度為 7.07)，根的兩個子節點更加偏移，其中左節點的誤差率僅為 1.29%。根本身是極好的區分器：如果您將樹高度限制為單一層，則誤差率為 7.3%。

甲狀腺機能衰退診斷

甲狀腺機能衰退疾病資料集與乳癌的相似，只是我們試圖預測地是甲狀腺機能衰退而不是癌症。檔案 *hypothyroid-dt.treeviz* 顯示為解決該問題而導入的「決策樹」分類器，在 *hypothyroid.schema* 上執行導入工具可產生該檔案。

- Windows 使用者可以在 MineSet 安裝目錄中 *\examples\hypothyroid-dt.treeviz* 和 *\data\hypothyroid.schema* 下找到這些檔案。
- IRIX 使用者可以在 */usr/lib/MineSet/treeviz/examples/hypothyroid-dt.treeviz* 和 */usr/lib/MineSet/data/hypothyroid.schema* 中找到這些檔案

在資料集中有 3163 筆記錄，其中的決大多數沒有甲狀腺機能衰退症狀 (95.23%)。這就意味著預測否定的判斷在多數時間裡是對的。但是，我們所擔心的正是那些模型預測為健康而實際上卻有甲狀腺機能衰退的人們。這種假否定非常嚴重的。從「更多導入工具」選項中選擇混淆矩陣，您會看到有五個錯誤分類的甲狀腺機能衰退的病人。

觀察「決策樹」，您可以看到根節點為綠色 (高準確度)。根上的 `fti` 上單一屬性顯示要給出許多否定的診斷是比較容易的。具有高 `fti` 的人 99.7% 是否定的，而那些未知值的人也是否定的 (也許醫生不測量此屬性是因為一些其它的屬性很明顯)，但是剩下的 (218 個人) 卻是很難診斷的病例。我們從 3163 筆記錄入手，但是真正「有趣」並值得挖掘的只有 218 個，這是因為要確定大多數病例的分類很容易。在該範例中，大多數資料並不令人感興趣，您便將注意力很快集中在一小部分人上。對於這 218 個人，您可以看到有 66% 是肯定而 34% 是否定的。

當您沿著樹向下，可以增加高度比例值 (滑動桿在可視化工具的左上部) 檢視不同的高度。在其中大多數人具有甲狀腺機能衰退的節點滿足條件「`fti<=64.5` 並且 `tsh>5.95`」。它包含 151 筆甲狀腺機能衰退記錄中的 140 筆。

Pima 糖尿病診斷

該資料集是關於糖尿病的診斷問題，它是利用統計學原理從 Phoenix Arizona 中的印地安部落收集而來。該目的是根據一些醫學特徵 (例如，血壓、體重、葡萄糖水平以及年齡) 來確定一個病人是否患有糖尿病。

檔案 `pima-dt.treeviz` 顯示為解決該問題而導入的「決策樹」分類器。在 `pima.schema` 上執行導入工具可產生該檔案。

- Windows 使用者可以在 MineSet 安裝目錄中 `\examples\pima-dt.treeviz` 和 `\data\pima.schema` 下找到這些檔案。
- IRIX 使用者可以在 `/usr/lib/MineSet/treeviz/examples/pima-dt.treeviz` 和 `/usr/lib/MineSet/data/pima.schema` 中找到這些檔案

DNA 邊界

在 DNA 資料集中有 3,186 筆記錄。DNA 邊界是從分子生物學領域的角度來界定的。結合連接處是 DNA 鏈上的點，在其上「多餘的」DNA 在蛋白質產生過程中就已移除。任務是識別 exon/intron 之間的界限，稱為 EI 位置；以及 intron/exon 之間的界限，稱為 IE 位置；或兩者都不是。IE 邊界稱為「接受者」而 EI 邊界為「捐獻者」。這些記錄最初從 GenBank 64.1 (genbank.bio.net) 中得到，其屬性提供一個有 60 個核甘的觀察視窗。分類點在視窗的中點，每個聯結處的一邊有為 30 個核甘。

在該範例中，「決策樹」的根節點顯示三個分類的分布。透過指向長條，您會發現其組成約有 24% 的 exon/intron，24% 為 intron/exon，而 52% 則為「無」。By pointing to the bars, you can see that the composition is about 24% exon/intron, 24% intron/exon, and 52% none. 根節點前面的「left_01」顯示這是一個需要首先檢視的重要屬性。「left_01」記號指示我們正在討論、處於連結結合處左邊的第一個核甘。第一個核甘（以及所有核甘）的屬性值選擇為「A」，「G」「T」「T」以及「C」核甘。如果「left_01」核甘是一個「G」，就沿著「G」分支到下一個節點，其分布顯示這樣的核甘比在根上的更像「exon/intron」或「intron/exon」：該分布對於「exon/intron」為 32%，而對於「intron/exon」為 43%，並且對於「沒有」為 24%。如果「left_01」核甘為「A」「T」「T」或「C」，則相應的「A」，「T」採用「T」或「C」分支並在三種情況下，「沒有」的機率急劇增加（分別為 87%、87% 和 95%）。該檢驗和分支過程就一直重複直到帶有預測分類（「exon/intron」）的最後節點「intron/exon」或「沒有」的最後節點。

對於該資料集，由於該領域的機率自然屬性，「證據分類器」可能比「決策樹分類器」更適合。可以由比較估計的誤差率來得到証實。

決策表樣例檔案

下面的樣例顯示「決策表」如何起作用，這些樣例都與 MineSet 所提供的樣例資料檔案相關。執行「決策表」導入工具（在「更多導入工具選項」中開啓「建議使用功能查詢」），您可以得到如下所述的 `-dtab.dtableviz` 和 `-dtab.dtableviz.data` 檔案。

注意：資料 (.data) 以及附加的配置 (.schema) 檔案在客戶端工作站的 *data* 目錄中。聚類可視化工具檔案，帶有 *-dtab.dtableviz* 附檔名，在客戶端工作站的 *examples* 目錄中。要將資料檔案載入 MineSet，開啓 .schema 檔案。

- Windows 使用者可以在 MineSet 安裝目錄的 \examples 和 \data 下找到這些檔案。
- IRIX 使用者可以在 /usr/lib/MineSet/treeviz/examples 和 /usr/lib/MineSet/data 中找到這些檔案

客戶波動

客戶波動是指電話用戶離開一家公司而轉投另一家公司。該樣例顯示造成用戶重選電話公司的原因。檔案 *churn.schema* 和 *churn.data* 用於產生這個樣例。Windows 使用者可以在 *Program Files\SGI\MineSet 3.0\data* 目錄下找到它們。IRIX 使用者可以在 */usr/lib/MineSet/data* 中找到這些檔案。

檔案 *churn-dtab.dtableviz* 顯示將屬性客戶波動為標籤而導入的分類器結構，該分類器的誤差率為 5.5%。紀錄中的 14.3% 代表波動的用戶。為第一細節層而選擇的兩個屬性是客服電話次數和白天總費用。通過觀察在這兩個屬性上的分布，您會發現客戶波動隨著白天總費用的增加而增加，只有在白天總費用少於 29.75 的情況下例外。如果服務電話的次數超過 3，則客戶波動率就高。大約 3/4 的記錄白天總費用小於 38 並且客服電話等於或小於 3。



A-1 在客戶波動資料集中細化下尋

當檢視用戶客戶波動時，從不清楚的區域開始進行細化下尋。圖 A-1 顯示對所有塊狀圖進行的細節下尋，其中沒有一個明顯佔多數的類別。下一個要考慮的屬性是國際計劃和聲音郵件的數目。在那些白天費用很高的用戶中，具有國際計劃及少數聲音郵件很明顯地與用戶波動率間有極大的相關性。選擇每個細化下尋區域中右下角的區塊，然後用滑鼠刷過右邊機率窗格中「churned=yes」旁的方塊，顯示在所選的區域中只有 3.4% 的用戶客戶波動。

現在，在每個先前的細化下尋區域中左上部的塊狀圖進一步細化下尋。這樣做時可以得到非常類似的分布。一致的模式顯示：較高的夜晚總費用及國際總費用與客戶波動之間呈較高的相關性。您可以細化下尋到另一層，來觀察國際通話總數的效果，但是該層有助於產生結論的記錄太少，您便無法太信任根據該樣例所得到的預測結果。如果您對國際通話如何影響客戶波動以及它與其它變數如何相關感興趣，可以返回「工具管理員」並正確地將國際通話總數對照為等級中更高的層次，再重新執行決策表。

雖然「州」與客戶波動有很好的相關性，但並不選擇它，因為算法對於取值較少的變數有內建的喜好設定。這就防止算法選擇像社會保障數目一樣的屬性，該數目唯一地辨識每一筆記錄，這樣，產生的訓練集高度準確，但對於劃分將來未標記的資料沒有任何用處。

轎車產地

轎車資料集包含從 70 年代到 80 年代轎車的不同模型資訊。屬性包括重量、加速能力以及每加侖英里數 (mpg)。檔案 *cars-dtab.dtableviz* 顯示為該資料集而導入的「決策表分類器」的結構。將標籤設為「產地」(日本、美國、歐洲)，並在 *cars.schema* 上執行導入工具可以產生該檔案。

- Windows 使用者在 MineSet 安裝目錄中 *\examples\cars-dtab.dtableviz* 和 *\data\cars.schema* 下可以找到這些檔案。
- IRIX 使用者可以在 */usr/lib/MineSet/examples/dtableviz/cars-dtab.dtableviz* 和 */usr/lib/MineSet/data/cars.schema* 中找到這些檔案

因為品牌屬性唯一地決定了產地，所以分類器的結構就極為簡單。所顯示的兩個屬性是品牌和汽缸型號。檢視各個品牌對汽缸數目的趨勢比較有趣。例如，Mazda 共有 21 種不同模型，Honda 有 18 種模型，但是它們都只有 5 個或更少的汽缸。相反地，Cadillac 只生產帶有 6 個或更多汽缸的轎車。

如果先移除品牌屬性，該樣例會變得更加有趣。另外一個有用的轉換可能是將汽缸屬性轉化為字串型，這樣可以顯示每個汽缸值，而非以一個分組形式出現。選擇性地，經由明確的對照，可以在品牌和汽缸之外建立額外的分支層。

預測性別

成人資料集包含從業成人的相關資訊。該資料集是從美國調查局中所摘錄。它包含年齡超過 16 歲、每周工作至少一小時、每年收入超過 \$100 的人之相關資訊。您也許想知道怎樣將男性和女性特徵化。

檔案 *adult-sex-dtab.dtableviz* 顯示為解決該問題而導入的「決策表分類器」的結構。在移除婚姻關係欄（使得分類器更加瑣細）後，將標籤設為性別，並在 *adult.schema* 上執行導入工具可以產生該檔案。對於每個值的組合，來方便地檢視記錄分布，您可以使用左邊比例滑動桿來縮放塊狀圖的高度。

- Windows 使用者可以在 MineSet 安裝目錄 *\examples\adult-sex-dtab.dtableviz* 和 *\data\adult-sex.schema* 下找到這些檔案。
- IRIX 使用者可以在 */usr/lib/MineSet/examples/dtableviz/adult-sex-dtab.dtableviz* 和 */usr/lib/MineSet/data/adult-sex.schema* 中找到這些檔案

在「決策表可視化工具」中，「標籤機率窗格」顯示從業者為男性的先驗機率要高於職業女性。「証據可視化工具」顯示婚姻狀況和職業對於確定性別是非常重要的屬性，但是，它並沒有顯示這兩個屬性之間的依賴性。最上層顯示幾個相互關係。例如，大部分從事手工業維修的人大多數與一般公民結婚（其中更特別的是，他們之中 98.6% 是丈夫），而職業為「Other-service」的人中大多數卻是「Never-married」（48% 為男性）。

起初您也許會覺得奇怪，「Marital_status = Married-civilian-spouse」的大多數人為男性，但是一旦您考慮到該資料是從納稅申報表中收集而來，您就會覺得這些男性的妻子們不工作似乎是蠻合理的，因為她們只是和丈夫一起被收入了檔案材料。

在「Occupation = Admin-clerical」中的離婚率最高。那些職業為「Other-service」的人其離婚率也很高，但是他們更傾向於分居，因為他們之中「分居」的數目甚至要比「Admin-clerical」的還高。

假設您想在「寡居」並且「職業 = 主管辦事員」的條件下找出是女性的機率。在証據可視化工具中，在這兩個屬性值上按一下就可以得到接近的答案（女性為 94.7%）。在此我們在這兩個屬性的交點處的塊上，按一下滑鼠左鍵即可得到正確的結果（95.2% 為女性）。

在「已婚城市夫婦」和「職業 = 未知」的對應塊上細化下尋至最下面一層，有一個比其它任何職業和婚姻狀態的組合還要顯著的模式。至於該塊，年輕的成員傾向為女性，更年長的成員傾向為男性。

薪資因素

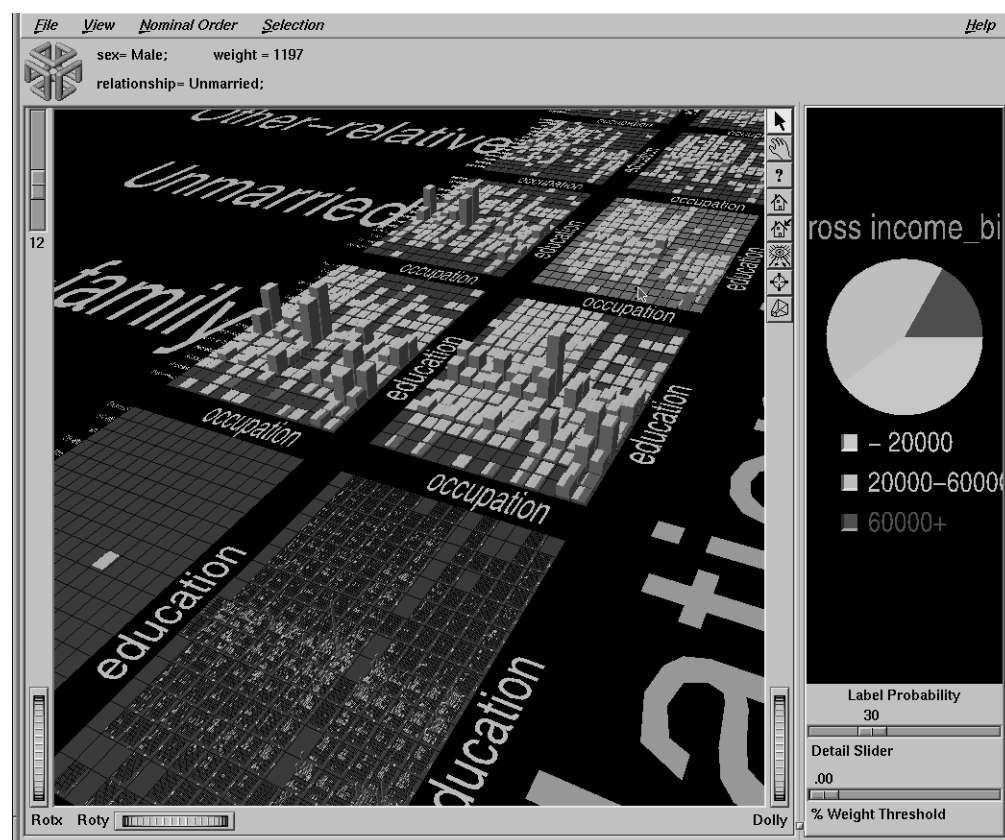
對於一個從業成人的資料集，您也許想找出影響薪資的因素。首先，將總收入分為兩組，少於 50,000 的為一組，超過 50,000 的為一組。然後，您可以執行 **MineSet** 分類器來確定影響薪資的因素。檔案 *adult-salary-dtab.dtableviz* 顯示為解決該問題而導入的「決策表」分類器。在 *adult.schema* 上執行導入工具可以產生該檔案，並利用使用者定義的臨界值將 *gross income* 分為五組。

- Windows 使用者可以在 **MineSet** 安裝目錄的 *\examples\adult-salary-dtab.dtableviz* 和 *\data\adult-salary.schema* 下找到這些檔案。
- IRIX 使用者可以在 */usr/lib/MineSet/examples/dtableviz/adult-salary-dtab.dtableviz* 和 */usr/lib/MineSet/data/adult-salary.schema* 中找到這些檔案

因為標籤是數字型，所以可用連續的色譜將顏色指定給每個分類。同樣在右邊機率窗格中的分類並不是經由份量大小來排序，這是因為它們具有數字順序。紅色分配給最高的組 (50,000+)。

在頂曾所選擇的兩個屬性是婚姻關係和「**Education_num**」。因為屬性「**education num**」僅是可能的不同教育程度的列舉，並不是您所想像的是教育的年數，所以並無特殊作用。但是，卻存在著一個近似的相關性。如果您比較想看到實際的字串值，可用教育程度來取代該欄。如果您只是移除 **education_num** 欄，並使用特徵查詢重新執行，則該算法可能不會在等級的頂部選取教育，因為它的記錄太多。

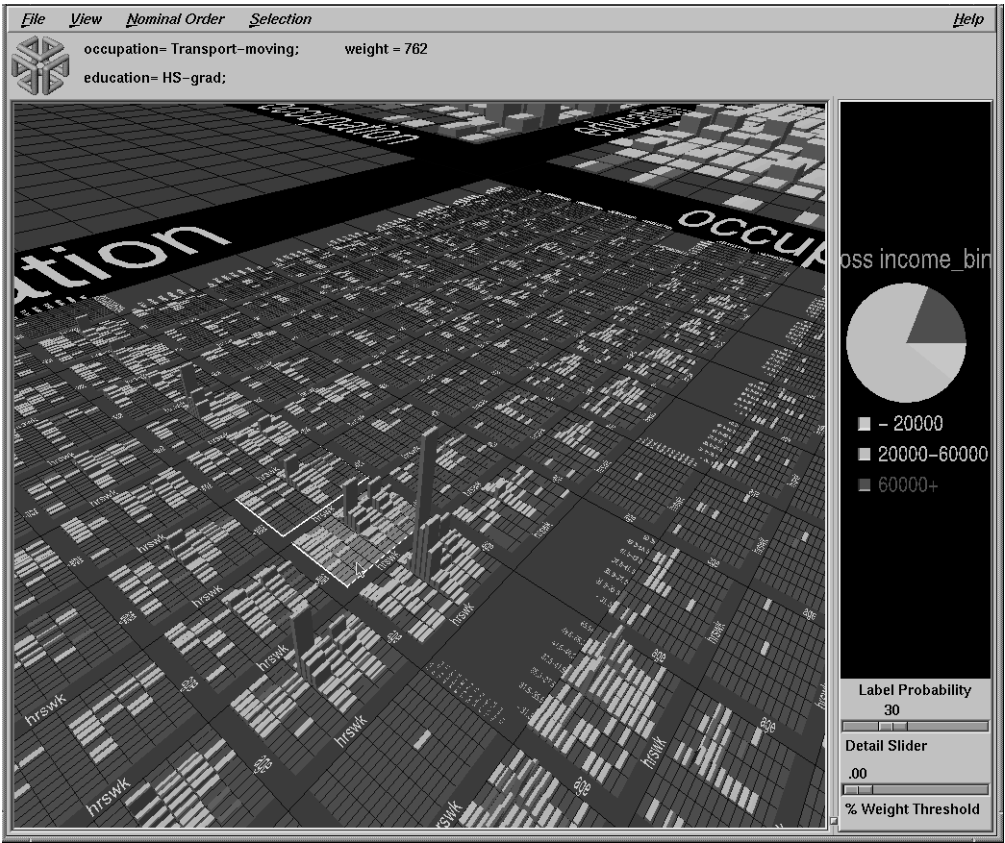
該模型中屬性的順序是自動選取來增加準確性的。應用領域知識來進行對照的模型通常可以提供更有用的可視化過程。*adult-salary3-dtab.dtableviz* 提供這樣的模型，並顯示在圖 A-2 中。在此薪資首先被分為 3 個區間 (20,000 和 60,000 是臨界值)。在第一層對照的屬性是：關係和性別；第二層為教育程度和職業；第三層為 **hrswk**(每周工作小時數) 和年齡。



A-2 使用成人資料集的決策表可視化工具

我們知道，在頂層「婚姻關係」和「性別」有很強的相關性。我們當然期望所有的丈夫都是男性，而所有的妻子都是女性，但是我們可以立即看到情況並非如此。透過按一下對應男性妻子的塊狀圖，我們可以看到其中有三個而他們的薪資都落在 20,000-60,000 的範圍內。可以在這些塊狀圖中細化下尋顯示關於這些異常記錄的更多資訊。您也許希望選擇這些塊（使用滑鼠左鍵）追溯其基本資料，由此您可以發現其它所有屬性的值。

在背景上按一下滑鼠右鍵，這樣會整體地細化下尋到下一層。現在，每個區塊都由教育程度和職業組成的一個矩陣所取代。對於每個方陣其排序是一樣的，並且整個的排序都是按照與收入之間相關性來進行。如果您選擇下拉式功能表中的「名稱順序」>「按權重排序」，則整個排序將按照記錄權重的大小來進行。最普遍的職業和教育程度層將出現在每個方陣的左下角。「高中畢業」是比例最多的教育程度，並且「專業技能型」是最普通的職業，但是其職業為專業技能型的高中畢業生卻不多。



A-3 成人資料集的進一步調查

返回按照收入的排序，您可以看到每個性別和婚姻關係組合的不同分配。在不屬於同一家庭關係的男性和女性之間並無很大的不同，但是在未婚男性和女性之間的差異卻十分明顯。(請參閱圖 A-3)。在男性方陣左下角中有很特別的紅色區塊聚類，而在女性方陣中卻沒有。稍微放大高度比例，您將會注意到女性方陣中在「職業 = 牧師」和「其他服務」處有個明顯的尖峰，而在對應的男性方陣中卻沒有這樣的尖峰。

在人口最多的區塊 (男性丈夫) 上按一下滑鼠右鍵。該操作可能會花掉一分鐘，這是因為可視化工具需要為該塊的下一層建造全部幾何形狀。該幾何形狀根據需要而建立，因為最初將其全部建立所需的時間過長，且過於浪費，而用戶很少對許多非常詳細的區域進行檢視。如果您在背景上誤擊了右鍵，而下一層分支量特別大，就必須等待較長的時間，因為計算量大的原因。如果細化下尋將花費很長的時間，就會出現一個帶有取消按鈕的進度長條。

考慮一下經由為職業和教育程度的每個組合顯示 hrswk 分布而分的許多年齡組。一個令人驚奇的事實是，不管顯示多少塊圖，在占有所有丈夫 2.5% 的地方有一個尖峰。(請參閱圖 A-3) 如果您必須為典型的丈夫挑選一個特徵，則下列說法應該合理：他中學畢業，做手藝維修，年齡在 41 到 59 之間，每周工作時數在 38 到 41 小時之間。

比較從事銷售且高中畢業的丈夫們與職業為執行經理且高中畢業的丈夫們之間的薪資分布，雖然在年齡分布和工時方面是相似的，但是對於經理們的這一組，屬於收入高於 60,000 類的機率為 34%，相較之下營業員的機率為 27%。要看到在頂部顯示的這些機率，首先按一下標籤機率窗格右邊 60000+ 收入類旁邊的按鈕，然後挑選左邊的塊狀圖。

蝴蝶花分類

在該資料集中，每個記錄都描述蝴蝶花朵的四個特徵：花萼寬度、花萼長度、花瓣寬度和花瓣長度。每個蝴蝶花進一步劃分為藍色蝴蝶花、多色蝴蝶花或純白色蝴蝶花類型。該目標就是把握每個蝴蝶花類型的特徵。

在執行分類器之前，按一下「工具管理員分類器」索引標籤中的「欄重要性」索引標籤；然後按一下「建議」，再按一下繼續。您得到了特性重要性的排序。花萼寬度、花萼長度和花瓣長度。在「分散可視化工具」中，您可以將這些對照到坐標軸上，將蝴蝶花類型對照為顏色，並觀察所分的聚類。

檔案 *iris-dtab.dtableviz* 顯示為解決該問題而導入的「決策表分類器」結構。在 *iris.schema* 上執行導入工具可產生該檔案。

- Windows 使用者可以在 MineSet 安裝的目錄 `\examples\iris-dtab.dtableviz` 和 `\data\iris.schema` 下找到這些檔案。
- IRIX 使用者可以在 `/usr/lib/MineSet/examples/dtableviz/iris-dtab.dtableviz` 和 `/usr/lib/MineSet/data/iris.schema` 中找到這些檔案

在「決策表可視化工具」中，我們可以看到「花萼寬度」是非常好的區別屬性。當增加花瓣寬度時，您可以看到那些花萼寬度在 0.75 到 1.65 之間的記錄，所有多色蝴蝶花可視顏色的實例都出現在「花瓣寬度 < 3.05」的分組中。

在純度不到 100% 的三個區塊中細化下尋。頂部的兩個塊中都包含唯一的 *iris-versicolor* 實例，使它們無法純淨。至於「花瓣寬度 < 3.05」的塊，要分離異常的 *iris-versicolor* 將有困難。但是，至於那個特殊的塊，可使用花萼寬度來分離多色蝴蝶花。

蘑菇分類

檔案 *mushroom-dtab.dtableviz* 顯示為解決該問題而導入的「決策表分類器」結構。在 *mushroom.schema* 上執行導入工具可以產生該檔案。

- Windows 使用者可以在 MineSet 安裝目錄中的 `\examples\mushroom-dtab.dtableviz` 的 `\data\mushroom.schema` 下找到這些檔案。
- IRIX 使用者可以在 `/usr/lib/MineSet/examples/dtableviz/mushroom-dtab.dtableviz` 和 `/usr/lib/MineSet/data/mushroom.schema` 中找到這些檔案

目的是在指定的資料集上理解哪些蘑菇可以食用而哪些有毒。在該資料集中有 8000 多筆記錄，執行該導入工具也許會花幾分鐘。請注意：在支持三分之一進行準確性估計的預設模式下，將保留三分之一的記錄用於測試模式。

每個蘑菇有很多特徵，包括菇帽顏色、褶紋和氣味。在「決策表可視化工具」中，氣味和莖的形狀出現在頂層。請注意：氣味可以很好地區分蘑菇是否可以食用。只有出現沒有氣味並且莖形狀為「向上擴大」時，才會出現不明確的情形。所以我們自然地在這個單獨的區塊中細化下尋。現在，我們只看到具有這兩個值的記錄，而這兩個值是根據菌紋和菌褶大小的值來分類，注意菌紋和菌褶大小之間的交互作用，該作用很難用其它分類器辨認。

既然在該資料集中的所有記錄都是標稱的，則這些值會根據在預測可食性過程中所產生的效果進行排序。您也許想將這些值按照字母或根據權重進行排序（普遍度）。要達到這一點，從名稱順序功能表中選取合適的方法。如果您單獨考慮菌紋或菌褶大小，將不能預測完全可食用或有毒蘑菇的大分類，但是若將它們一起考慮，會發現如果沒有菌紋且菌褶較大，則該類型中的 814 個蘑菇可以食用。相反地，如果有擦傷且菌褶較窄，則該類型中所有的 11 個蘑菇都有毒。要區分另外兩種情況，我們必須更進一步細化下尋。

在「決策表可視化工具」中，向右移動「% 權重臨界值」滑動桿。最後，那些帶有霉味的蘑菇對應塊狀圖將從畫面中刪除。原因是只有不到 1% 的記錄標記為「氣味 = 霉味」。

黨派隸屬

該資料集包含投票記錄。目的是在指定關鍵投票資料的條件下找出議員所屬黨派。資料集包括由議會季度年鑑 (CQA) 所識別的美國眾議院每個議員在 16 次關鍵表決中的投票。CQA 列出九個類型的投票：投票贊成、配合贊成以及宣布贊成（這三個為贊成）；投票反對、配合反對以及宣布反對（這三個為反對）；交出選票、交出選票避免利益衝突以及不投票或持其它觀點（這三個為未知處理）。

在執行分類器之前，觀察 16 個投票結果來檢視您是否能查覺最重要的特徵。然後執行「決策表可視化工具」。對於該資料集，您也許想要按照字母順序排列，所以「是」類型的投票都出現右上角，而「不」類型投票則在左上角。

在頂層，我們可以看到「合成燃料公司縮減」和「醫生費用凍結」。在這些變數中有個令人驚奇的關係，用任何其它模型基本上都不可能得出。除了三個以外其它所有投票反對醫生的都是民主黨。幾乎每個民主黨成員既投票反對費用凍結同時又投票反對合成燃料縮減 (206 個中只有 3 個不符合該模式)。令人驚訝地，除 5 個以外所有投票贊成醫生費用凍結的共和黨成員，都投票反對合成燃料縮減。在這些完全不相關的問題中存在的奇怪關係暗示著這些結果以某種方式結合著，而這需要更深入的調查。

頂層中大多數的塊狀圖除了中間的一個 (只包含 6 個記錄) 以及在兩個問題上都投贊成票的一個以外幾乎都是純淨的。在此我們可以在另一個等級進行，以區分該組中代表的政治傾向。利用「反衛星實驗禁令」和「採納財政議案」對該組的 55 個代表進一步區分。

檔案 *vote-dtab.dtableviz* 顯示為解決該問題而導入的「決策表分類器」。在 *vote.schema* 上執行導入工具可產生該檔案。

- Windows 使用者可以在 MineSet 安裝目錄的 *\examples\vote-dtab.dtableviz* 和 *\data\vote.schema* 下找到這些檔案。
- IRIX 使用者可以在 */usr/lib/MineSet/examples/dtableviz/vote-dtab.dtableviz* 和 */usr/lib/MineSet/data/vote.schema* 中找到這些檔案

乳癌診斷

乳癌資料集包含有關婦女接受乳癌診斷的資訊。每筆記錄都包含病人細胞大小、凝塊厚度及邊緣附著力等屬性的記錄。最終的屬性是診斷為惡性或良性。該檔案 *breast-dtab.dtableviz* 顯示為解決該問題而導入的「決策表分類器」，在 *breast.schema* 上執行導入工具可以產生該檔案。

- Windows 使用者可以在 MineSet 安裝目錄的 *\examples\breast-dtab.dtableviz* 和 *\data\breast.schema* 下找到這些檔案。
- IRIX 使用者可以在 */usr/lib/MineSet/examples/dtableviz/breast-dtab.dtableviz* 和 */usr/lib/MineSet/data/breast.schema* 中找到這些檔案

在「決策表可視化工具」中，在等級的頂部顯示細胞形狀的分割和均勻性。如果兩個屬性同時都為低值，則指定的樣例有 99.2% 為良性；相反，兩個值都比較高的訓練記錄 100% 為惡性。

在純度不是很高的四個塊狀圖中細化下尋。現在邊緣附著力和單細胞核也用於區分過程。在該層每個塊狀圖中的記錄越少；其結果是噪音更多，並且更難掌握趨勢。邊緣附著力和單細胞核都高的值似乎更傾向於惡性腫瘤，但是並不確定。請注意：單細胞核的第一個值為空。這些空值塊狀圖的分布比其它的塊狀圖更值得懷疑，所以您也許希望取消選擇檢視 > 顯示空值來將它們隱藏。

如果您整體細化下尋兩層以上時，您會注意到一些有趣的特徵。這些塊狀圖非常小，而大量的多維度空間區域是空的。只有一些小區域，很多記錄都進行聚類。當所有的值都低時，有一個巨大尖峰 (100% 良性)。尖峰可占資料的 20%。

甲狀腺機能衰退診斷

甲狀腺機能衰退資料集與乳癌的一樣。檔案 *hypothyroid-dtab.dtableviz* 顯示為解決該問題而導入的「決策表分類器」。在 *hypothyroid.schema* 上執行導入工具可產生該檔案。

- Windows 使用者可以在 MineSet 安裝目錄的 `\examples\hypothyroid-dtab.dtableviz` 和 `\data\hypothyroid.schema` 下找到這些檔案。
- IRIX 使用者可以在 `/usr/lib/MineSet/examples/dtableviz/hypothyroid-dtab.dtableviz` 和 `/usr/lib/MineSet/data/hypothyroid.schema` 中找到這些檔案

在資料集中有 3163 筆記錄，其中的決大多數沒有甲狀腺機能衰退症狀 (95.45%)。這就意味著預測否定的判斷在多數時間裡是對的。但是，我們所擔心的正是那些模型預測為健康而實際上卻有甲狀腺機能衰退的人們。這種假否定非常嚴重的。

爲了避免假否定，在這種情況下，您也許想調節損失矩陣使後驗機率向預測甲狀腺機能衰退方面傾斜。將本來患有疾病的人診斷爲健康就要付出很高的代價，而將實際健康的人診斷爲有病只意味著不得不進行更準確的檢查或對他們進行不必要的治療。

在該資料集中使用「決策表導入工具」，我們注意到：

- 如果 **fti** 爲空，則 **tbg** 就不爲空且 **t3** 幾乎都爲空（再細化下尋一層）。
- 只有當 **fti** 爲空時，**t4u** 才會是空的，但有兩個情形例外。
- 當 **fti** 爲較低值時甲狀腺機能衰退更容易發生。

Pima 糖尿病診斷

該資料集是關於糖尿病的診斷問題，利用統計學從 Phoenix Arizona 中的印地安部落收集而來。該目的是根據一些醫學特徵（例如，血壓、體重、葡萄糖水平以及年齡）來確定一個病人是否患有糖尿病。

檔案 *pima-dtab.dtableviz* 顯示爲解決該問題而導入的「決策表分類器」結構。在 *pima.schema* 上執行導入工具可產生該檔案。

- Windows 使用者可以在 MineSet 安裝目錄的 *\examples\pima-dtab.dtableviz* 和 *\data\pima.schema* 下找到這些檔案。
- IRIX 使用者可以在 */usr/lib/MineSet/examples/dtableviz/pima-dtab.dtableviz* 和 */usr/lib/MineSet/data/pima.schema* 中找到這些檔案

使用「決策表可視化工具」後我們注意到：

- 在細節的第二層上您可以看到小於 28 歲的女性幾乎沒有懷過六次孕的（只有 4 人）。您希望在這些記錄中細化下尋以得到這四人的相關資訊。
- 血糖濃度高、體重過大就表示發生糖尿病的可能性越大。

DNA 邊界

檔案 *dna-dtab.dtableviz* 顯示為解決該問題而導入的「決策表分類器」結構。在 *dna.schema* 上執行導入工具可產生該檔案。

- Windows 使用者可以在 MineSet 安裝目錄中 *\examples\dna-dtab.dtableviz* 和 *\data\dna.schema* 下找到這些檔案。
- IRIX 使用者可以在 */usr/lib/MineSet/examples/dtableviz/dna-dtab.dtableviz* 和 */usr/lib/MineSet/data/dna.schema* 中找到這些檔案

在 DNA 資料集中有 3,186 筆記錄。DNA 邊界是從分子生物學領域的角度來界定。結合連接處是 DNA 鏈上的點，在其上「多餘的」DNA 在蛋白質產生過程中就已移除。任務是識別 exon/intron 之間的界限，稱為 EI 位置；以及 intron/exon 之間的界限，稱為 IE 位置；或兩者都不是。IE 邊界稱為「接受者」而 EI 邊界為「捐獻者」。這些記錄最初從 GenBank 64.1 (*genbank.bio.net*) 中得到。該屬性提供一個有 60 個核甘的觀察視窗。分類點在視窗的中點，每個聯結處的一邊有為 30 個核甘。

從「決策表可視化工具」中，您可以看到一個驚人的模式，而在任何其它分類器模型中都不很明顯。在頂層在 left_01 和 right_02 之間存在著很明顯的相互作用。Exon/intron 只在 right_02 = T 時才存在。而 Intron/extron 只在 left_01=G 時才存在。對於 left_01 和 right_01 的其它值只有很少的份量結合。

證據可視化工具樣例檔案

下面的樣例顯示分類器如何發揮作用。各樣例都與 MineSet 所提供的樣例資料集相關。執行導入工具，您可以產生如下所述的 *.eviviz* 檔案。

資料檔案可以開啓相對應的 *.schema* 檔案而載入 MineSet，而檔案儲存在 *data* 目錄中（例如，*churn.schema*）。分類器可視化檔案，具有 *.eviviz* 副檔名，可以從 *examples* 目錄中找到並開啓。

客戶波動

客戶波動是指一個電話用戶離開一家公司到另一家公司。該樣例顯示造成用戶變更電話公司的原因。

檔案 *churn.schema* 和 *churn.data* 用於產生這個樣例。要將資料檔案載入 MineSet，開啓 *.schema* 檔案。

- Windows 使用者可以在 MineSet 安裝目錄的 *\examples\churn.data* 和 *\data\churn.schema* 下找到這些檔案。
- IRIX 使用者可以在 */usr/lib/MineSet/examples/eviviz/churn.data* 和 */usr/lib/MineSet/data/churn.schema* 中找到這些檔案

檔案 *churn.eviviz* 顯示將屬性客戶波動作爲標籤而導入的分類器結構。該分類器的誤差率爲 12%。對於記錄，14.1% 代表波動的用戶。兩個最重要的屬性，白天總分鐘數和白天總費用明顯地相關。如果從「更多導入工具」選項中選擇「自動功能選擇」之後執行導入工具，則只利用四個屬性（白天總費用、服務電話數目、聲音郵件業務以及聲音郵件訊息數目）就可將誤差率降到 10.5%。白天總費用超過 53.78 的所有 29 個用戶都客戶波動了。

大量的客服電話也是客戶波動的預測工具。大量客服電話可能表示在使用複雜設備或接收不可靠服務時出現障礙。具有國際業務的用戶也很有可能客戶波動。有些州的用戶比其它州的更有可能客戶波動，California 和 New Jersey 有最多的客戶波動，而 Virginia 爲最少。要觀察擁有記錄數超過記錄總數的 2% 的州，則將「% 權重臨界值」滑動桿滑到右邊，這樣將從顯示中刪除絕大部分州的值。如果您也選擇「名稱順序 > 加權」，具有最多記錄的州 West Virginia (WV) 會出現在最左面。在區分客戶波動的過程中，許多屬性（清單的低部）是沒有用的。請注意：白天費用是很好的預測工具，而夜間費用則不是。

轎車產地

轎車資料集包含從 70 年代到 80 年代轎車的不同模型資訊。屬性包括重量，加速能力，以及每加侖英里數 (*mpg*)。檔案 *cars.eviviz* 顯示為解決該問題而導入的「証據分類器」結構。在 *cars.schema* 執行導入工具可以產生該檔案，其標籤設為產地 (日本、美國、歐洲) 並將汽缸欄轉變為字串類型。將此汽缸欄轉換為字串類型來觀察到所有的值並避免自動離散化。

- Windows 使用者可以在 MineSet 安裝目錄 `\examples\cars-eviviz` 和 `\data\cars.schema` 下找到這些檔案。
- IRIX 使用者可以在 `/usr/lib/MineSet/examples/eviviz/cars.eviviz` 和 `/usr/lib/MineSet/data/cars.schema` 中找到這些檔案

如果您有一個關於轎車屬性的資料集，您可能想知道不同產地轎車的特徵。從右邊餅圖中標籤值的分布可以看出，資料集中的多數轎車都是美國製造的 (62.5%)，而一小部分產自日本 (20.2%) 和歐洲 (17.3%)。因為每個品牌都只與產地國家相關，因此很明顯地品牌是產地最好的預測工具。因此，它具有最高的重要性並且在清單的頂部。透過觀察圓餅圖的高度，可以看出許多轎車有四個汽缸，大多數重量小於 3000 lbs 並且大多數可以在 13 至 20 秒的範圍內將速度提高到每小時 60 英哩。

觀察個別屬性值份量的分布。如果轎車的引擎大小為 >169 立方英寸，就幾乎可以肯定它產自美國；並且可以肯定不是產自日本。另一個餅圖顯示美國轎車通常有 6 或 8 個汽缸，每加侖的英哩數較低，大馬力 (超過 134)，大重量 (超過 2981 lbs) 並具有較強的加速能力。日本汽車具有較好的汽油英哩數，3 或 4 個汽缸 (一些為 6 汽缸)，並且具有較小的引擎。如果您在「標籤機率窗格」中按一下了「歐洲」，您可以看到代表歐洲轎車的証據的長條圖。例如，如果轎車有五個汽缸，則肯定是產自歐洲。但是，相應餅圖的高度顯示在資料中只有三個轎車有五個汽缸。如果轎車有較好的英哩數，則很明顯地它產自歐洲。如果轎車的英哩數小於 41，它產自歐洲的機會就為 83%。如果一個轎車為歐洲產，其英哩數多於 41 mpg 的機會是 10.4%。但是只有 2% 的日本轎車 - 無美國轎車有這個範圍內的 mpg，所以這是產地為歐洲有最強的証據。

假如您只知道一個轎車有 40mpg 且重量為 3000lbs，而又想預測轎車的產地。選擇合適的餅圖 (或長條圖)：mpg=30.95-41.15 並且 weightlbs=2981.5+。顯示在右邊的機率結果為美國 84%，歐洲 16%。不可能是日本轎車，這是因為在訓練集中沒有日本轎車的重量為 lbs>2981.5。如果您選中拉普拉斯校正選項 (值為 0.5) 並再次執行導入工具，您將得到不同的結果：歐洲的機率為 16%，美國的機率為 82%，而日本的機率為 2%。這是因為拉普拉斯校正可以防止塊狀圖中的任何份量完全變為 0。無疑地，對於日本為何無法製造重量超過 2981 lbs 的轎車並沒有基本的原因，因此當機率 (餅圖) 一起相乘時，並沒有刪除預測為日本轎車的機率。

預測性別

成人資料集包含從業成人的相關資訊。該資料集是從美國調查局中所摘錄。它包含年齡超過 16 歲、每周工作至少一小時、每年收入超過 \$100 的人之相關資訊。您也許想知道怎樣將男性和女性特徵化。檔案 *adult-sex.eviviz* 顯示為解決該問題而導入的「證據分類器」結構。在移除婚姻關係欄 (使得分類器更加瑣細) 後，將標籤設為性別，並在 *adult.schema* 上執行導入工具可以產生該檔案。

- Windows 使用者可以在 MineSet 安裝目錄 *\examples\adult-sex.eviviz* 和 *\data\adult-sex.schema* 下找到這些檔案。
- IRIX 使用者可以在 */usr/lib/MineSet/examples/eviviz/adult-sex.eviviz* 和 */usr/lib/MineSet/data/adult.schema* 中找到這些檔案

在「證據可視化工具」中，「標籤機率窗格」顯示從業者為男性的先驗機率要高於職業女性。

- 婚姻狀況對於性別來說是最重要的預測工具。如果一個工人是已婚市民，有很大的機率為男性。但是，如果一個工人寡居並且正在工作，則很有可能為女性。
- 所列出的第二個屬性顯示職業。研究這些可以知道在特定的性別中普遍是何種職業。從左到右按照男性所占比例的降冪列出不同的職業：軍隊 (100%)，手動維修 (95%)，交通運輸 (95%) 以及種植漁業 (94%)。而女性則代以房屋內務 (94%) 以及主管辦事員 (67%)。按一下「標籤機率窗格」中「女性」旁的按鈕，然後將滑鼠移至「職業 = 主管辦事員」上，可以看到 23% 的女性其工作為主管辦事員。相反地，假設一個人的工作為主管辦事員，則其性別為女性的機率為 67%。

假設您要在此種條件下，即「寡居」並且「職業 = 主管辦事員」中找出為女性的機率。(處於選擇模式時) 當在「女性」旁的方塊上移動滑鼠時，按一下該屬性值並閱讀頂部的文字可以完成上面所說的步驟。

- 如果一個人的工作種類是「私人企業」或「非私人企業」，則其為男性的機率就比較高。相反地，如果一個人的工作種類為「政府職員」，則其為女性的條件機率就比較高，但是後驗機率（在考慮先驗機率之後）並不高（按一下它並觀察右邊的後驗機率）。在為州政府工作的條件上，女性所占的份額增加，但是並未達到由此預測其為女性的程度。
旋轉該檢視，您可以從圖的高度中看出大多數人在私人企業中工作。
- 透過觀察總收入屬性，您會發現收入範圍越高，為男性的機率越大。
- 教育程度並沒有對性別提供多少資訊，除了博士學位，在其中男性較多一些。
- 不同的職業會因男性或女性而有不同的分布。
- 種族屬性顯示，在條件機率中非裔美國女性工作的百分比要高於其他種族的百分率。按一下該值可以看到：在男性和女性之間後驗機率大致相等。
- 在該資料集中男性每周工作的時數要多於女性。

薪資因素

如果您有一個從業成人的資料集，您可能想知道何種因素會影響薪資。首先，將 `gross_income` 分為五組，臨界值為 10,000、20,000、30,000 以及 60,000。如此一來，每個記錄的屬性為這五組值中的一個。您可以執行 `MineSet` 分類器來確定是何種因素影響薪資。檔案 `adult-salary.eviviz` 顯示為解決該問題而導入的「証據分類器」結構。在 `adult.schema` 上執行導入工具可以產生該檔案，並利用使用者定義的臨界值將 `gross_income` 分為五組。

- Windows 使用者可以在 MineSet 安裝目錄中的 `\examples\adult-salary.eviviz` 和 `\data\adult.schema` 下找到這些檔案。
- IRIX 使用者可以在 `/usr/lib/MineSet/examples/eviviz/adult-salary.eviviz` 和 `/usr/lib/MineSet/data/adult.schema` 中找到這些檔案

「證據導入工具」中的屬性按照重要性劃分等級；這樣，可以視婚姻關係、婚姻狀況、年齡、職業、教育程度、每周時數及性別為重要。因為標籤是數字型，所以可用連續的色譜將顏色指定給每個分類。紅色分配給最高的組 (60,000+)。根據份量的大小，在「標籤機率窗格」中列出分類標籤。當您按一下「主視窗」中的值，類標籤的次序會改變，以保持該標籤在頂部為最大預測類。

- *婚姻關係*顯示夫妻所賺的錢可能比未婚工人或無家庭的工人還要多。妻子的收入比丈夫的收入略高。
- *婚姻狀況*顯示大多數人結過婚 (從左邊數來的第二個圖很高)。已婚工人比未婚工人所賺的薪資高。
- *年齡*顯示年齡是一個關鍵的因素。直到 61 歲時，許多人都已退休，賺 \$50,000 以上的機率隨年齡的增長而增加。
- 不同的職業對應不同的機率。管理型和專業型的工作可保證年收入超過 \$60,000。
- *教育程度*是重要的因素。當只考慮教育程度時，對於收入超過 \$60,000 的證據大都來自最高證據給予教育程度為碩士、博士或從專業學校進入大學的那些工人。
- 每周時數顯示工作的時數越多，賺取更多錢的證據越高。
- *性別*顯示女性會每年收入少於 \$60,000 提供更多的證據。
- 調節 *權重百分數*滑動桿來移除 `native_country` 的值，具有權重較低的教育程度以及職業值也會移除。

蝴蝶花分類

在該資料集中，每個記錄都描述蝴蝶花朵的四個特徵：花萼寬度、花萼長度、花瓣寬度和花瓣長度。每個 蝴蝶花進一步劃分為藍色蝴蝶花、多色蝴蝶花或純白色蝴蝶花類型。該目標就是把握每個蝴蝶花類型的特徵。

在執行分類器之前，按一下「工具管理員分類器」索引標籤中的「欄重要性」索引標籤；然後按一下 *繼續*，您會得到這些特徵重要性的等級順序：花萼寬度、花萼長度和花瓣長度。在「分散可視化工具」中，您可以將這些對照到坐標軸上，*iris_type* 對照為顏色，並觀察自然聚類結果。

檔案 *iris.eviviz* 顯示為解決該問題而導入的「証據分類器」結構。在 *iris.schema* 上執行導入工具可產生該檔案。

- Windows 使用者可以在 MineSet 安裝目錄的 `\examples\airis.eviviz` 和 `\data\iris.schema` 下找到這些檔案。
- IRIX 使用者可以在 `/usr/lib/MineSet/examples/eviviz/iris.eviviz` 和 `/usr/lib/MineSet/data/iris.schema` 中找到這些檔案

在「証據可視化工具」中，我們可以看到花萼長度和花萼寬度是極好的區分屬性，而花瓣長度和花瓣寬度略差。向右移動重要性臨界值滑動桿可以看到以花瓣為基礎的屬性首先消失。

蘑菇分類

檔案 *mushroom.eviviz* 顯示為解決該問題而導入的「証據分類器」結構。在 *mushroom.schema* 上執行導入工具可以產生該檔案。

- Windows 使用者可以在 MineSet 安裝目錄中的 `\examples\mushroom.eviviz` 和 `\data\mushroom.schema` 下找到這些檔案。
- IRIX 使用者可以在 `/usr/lib/MineSet/examples/eviviz/mushroom.eviviz` 和 `/usr/lib/MineSet/data/mushroom.schema` 中找到這些檔案

目的是在指定的資料集上預測哪些蘑菇可以食用而哪些有毒。在該資料集中有 8000 多筆記錄，執行該導入工具也許會花幾秒鐘。請注意：支持三分之一做準確性估計的預設模式下模式下，將保留三分之一的記錄用於測試。

每個蘑菇有很多特徵，包括菇帽顏色、褶紋和氣味。預設情況下，「証據導入工具」按照重要性（也就是在預測標籤過程中的有效性）來排序屬性。氣味和孢子顏色出現在清單的頂部，因為這些屬性，在塊圖中的分布隨著值的不同而有很大不同。既然在該資料集中的所有記錄都是標稱的，則這些值會根據在預測可食性過程中所產生的效果由左至右進行排序。您也許想將這些值按照字母或根據權重進行排序。要達到這一點，從名稱順序功能表中選擇合適的方法。改變指標為箭頭（按一下主畫面右上角的箭頭圖示或按下 Esc 鍵），然後按一下右邊窗格中有關該分類的按鈕，您就可以看到有毒蘑菇的特徵。較高的長條圖與表示蘑菇有毒的值相關。

在「証據可視化工具」中，向右移動「細節」滑動桿，可以從畫面中移除具有最低重要性的屬性。目前所知，最重要的屬性是氣味，它的重要性為 92；而所有其它屬性的重要性都小於 48。所有的值都是很好的區別器，但是如果沒有氣味（無），就會出現兩類的混淆。「証據可視化工具」可讓您觀察或許是關鍵的特定值，即使屬性本身不一定重要。例如，`stalk_color_below_ring` 不是理想的區分屬性，這是因為在大部分時間它的值為白色。因為顏色為白色的可食性蘑菇和有毒蘑菇其數量相等，因此白色沒有預測力。當 `stalk_color_below_ring` 為灰色或淺黃色時，它可以進行極準確的區分，但是很少有蘑菇帶有這種顏色。

黨派隸屬

該資料集包含投票記錄。目的是在指定關鍵投票資料的條件下找出議員所屬黨派。資料集包括由 *議會季度年鑑* (CQA) 所識別的美國眾議院每個議員在 16 次關鍵表決中的投票。CQA 列出九個類型的投票：投票贊成、配合贊成以及宣布贊成（這三個為贊成）；投票反對、配合反對以及宣布反對（這三個為反對）；交出選票、交出選票避免利益衝突以及不投票或持其它觀點（這三個為未知處理）。

在執行分類器之前，觀察 16 個投票結果來檢視您是否能查覺最重要的特徵。然後執行「証據可視化工具」。對於該資料集，您也許想按照字母順序將值進行排列，以使沒有選票的在最左邊，未決定的在中間，而代表是的在右邊。

一些結果清楚地指示一個人的黨派傾向。民主黨傾向於投票贊成醫生費用凍結以及對 El Salvador 的援助，而共和黨則投票贊成採納財政議案以及對 Nicaragua 中的 Contras 進行援助。

在黨派中移民問題並沒有引起分歧，然而政客們在這上面有明確的立場，因為 235 人中只有 7 人對該問題無所適從。

檔案 *vote.eviviz* 顯示為解決該問題而導入的「證據分類器」結構。在 *vote.schema* 上執行導入工具可產生該檔案。

- Windows 使用者可以在 MineSet 安裝目錄下 *\examples\vote.eviviz* 和 *\data\vote.schema* 中找到這些檔案。
- IRIX 使用者可以在 */usr/lib/MineSet/examples/eviviz/vote.eviviz* 和 */usr/lib/MineSet/data/vote.schema* 中找到這些檔案

乳癌診斷

乳癌資料集包含有關婦女接受乳癌診斷的資訊。每筆記錄都包括病人細胞大小、凝塊厚度及邊緣附著力等屬性，最終的屬性是診斷為惡性或良性。檔案 *breast.eviviz* 顯示為解決該問題而導入的「證據分類器」結構。在 *breast.schema* 上執行導入工具可以產生該檔案。

- Windows 使用者可以在 MineSet 安裝目錄中 *\examples\breast.eviviz* 和 *\data\breast.schema* 下找到這些檔案。
- IRIX 使用者可以在 */usr/lib/MineSet/examples/eviviz/breast.eviviz* 和 */usr/lib/MineSet/data/breast.schema* 中找到這些檔案

在「證據可視化工具」中，您可以看到 *sample_code_number* 分布處在一個平分為兩部分的範圍裡，也就意味著它並不能指出乳癌是良性或惡性的。

甲狀腺機能衰退診斷

甲狀腺機能衰退資料集與乳癌的一樣。檔案 *hypothyroid.eviviz* 顯示為解決該問題而導入的「證據分類器」結構。在 *hypothyroid.schema* 上執行導入工具可產生該檔案。

- Windows 使用者可以在 MineSet 安裝的目錄中 `\examples\hypothyroid.eviviz` 和 `\data\hypothyroid.schema` 下找到這些檔案。
- IRIX 使用者可以在 `/usr/lib/MineSet/examples/eviviz/hypothyroid.eviviz` 和 `/usr/lib/MineSet/data/hypothyroid.schema` 中找到這些檔案

在資料集中有 3163 筆記錄，其中的決大多數沒有甲狀腺機能衰退症狀 (95.45%)。這就意味著預測否定的判斷在多數時間裡是對的。但是，我們所擔心的是那些模型預測為健康而實際上卻有甲狀腺機能衰退的人們，這種假否定非常嚴重。

觀察一下 *tsh* 介於 6.35 和 27.5 之間的塊狀圖。它顯示許多關於甲狀腺機能衰退的證據。然而，按一下它時，在右邊的後驗機率餅仍然預測為「否定」，這是因為「否定」的先驗機率太大。

為了避免假否定，在這種情況下，您也許想調節損失矩陣使後驗機率向預測甲狀腺機能衰退方面傾斜。將本來患有疾病的人診斷為健康就要付出很高的代價，而將實際健康的人診斷為有病只意味著進行更準確的檢查或對他們進行不必要的治療。

在「證據可視化工具」中，您可以看到 *fti* 很重要。對於甲狀腺機能衰退，頭兩個範圍提供很多證據。

Pima 糖尿病診斷

該資料集是關於糖尿病的診斷問題，利用統計學從 Phoenix Arizona 中的印地安部落收集而來。該目的是根據一些醫學特徵（例如，血壓、體重、葡萄糖水平以及年齡）來確定一個病人是否患有糖尿病。

檔案 *pima.eviviz* 顯示為解決該問題而導入的「證據分類器」結構。在 *pima.schema* 上執行導入工具可產生該檔案。

- Windows 使用者可以在 MineSet 安裝目錄中 *\examples\pima.eviviz* 和 *\data\pima.schema* 下找到這些檔案。
- IRIX 使用者可以在 */usr/lib/MineSet/examples/eviviz/pima.eviviz* 和 */usr/lib/MineSet/data/pima.schema* 中找到這些檔案

在「證據可視化工具」中，您可以看到許多無關的屬性。當 *plasma_glucose* 增加時，得糖尿病的機率也隨著增加。當年齡大於 27 歲時，如果懷孕的次數較高（超過 6），也將成為很好的指標。。

DNA 邊界

檔案 *dna.eviviz* 顯示為解決該問題而導入的「證據分類器」結構。在 *dna.schema* 上執行導入工具可產生該檔案。

- Windows 使用者可以在 MineSet 安裝目錄下 *\examples\dna.eviviz* 和 *\data\dna.schema* 中找到這些檔案。
- IRIX 使用者可以在 */usr/lib/MineSet/examples/eviviz/dna.eviviz* 和 */usr/lib/MineSet/data/dna.schema* 中找到這些檔案

在 DNA 資料集中有 3,186 筆記錄。DNA 邊界是從分子生物學領域的角度來界定。結合連接處是 DNA 鏈上的點，在其上「多餘的」DNA 在蛋白質產生過程中就已移除。任務是識別 *exon/intron* 之間的界限，稱為 EI 位置；以及 *intron/exon* 之間的界限，稱為 IE 位置；或兩者都不是。IE 邊界稱為「接受者」而 EI 邊界為「捐獻者」。這些記錄最初從 GenBank 64.1 (*genbank.bio.net*) 中得到，其屬性提供一個有 60 個核甘的觀察視窗。分類點在視窗的中點，每個聯結處的一邊有為 30 個核甘。

從「證據可視化工具」中，您可以看到位於中心附近的屬性都被選為很重要，而遠離結合處的屬性就比較不重要。

如果您按一下並選擇對應於「left_01」的左邊窗格中的圖表：G 和「left_02:A」，然後，右邊標籤機率窗格中的餅圖將改變以顯示由「證據分類器」所預測的每個分類之機率分布。在這兩個值的條件下，圓餅圖顯示了所建立的證據模型將最高的機率分配給了「intron/exon」，其後是「exon/intron」和「none」。

如果您啟動自動功能選擇，雖然執行時間會增加很多 (有時以小時記)，但準確性卻只略有改進。在這種情況下，執行自動特徵選擇時應有所考慮。

地圖可視化工具樣例檔案

所提供的配置和資料檔案樣例展示「地圖可視化工具」的特性和功能。

- Windows 使用者可以在 MineSet 安裝目錄中的 `\examples\mapviz` 下找到這些檔案。`.gfx` 和 `.hierarchy` 檔案可以在 `\config\mapviz` 中找到。
- IRIX 使用者可以在 `/usr/lib/MineSet/examples/mapviz` 中找到這些檔案。`.hierarchy` 和 `.gfx` 檔案可以在 `/usr/lib/MineSet/mapviz/gfx_files` 中找到。

- `blocks.mapviz`、`blocks.data`、`blocks.gfx` 和 `blocks.hierarchy`
該簡單的樣例顯示四個相鄰的方塊，每個方塊的高度和顏色隨著 `blocks.data` 中基本資料的不同而不同。您可以使用滑鼠中鍵 您可以使用滑鼠中鍵來概化上尋以檢視上部或下部區塊組的組合值，接著繼續概化上尋檢視概化成一個方塊圖的組合值。您可以利用滑鼠右鍵細部下尋來觀察重新出現的更詳細物件。

- `population.australia.mapviz`、`population.australia.data`、`australia.states.gfx` 和 `australia.states.hierarchy`
該資料檔案中所為澳大利亞的每個州及地區包含一條記錄。每一列包含三個以索引標籤分隔的項目：有關州或地區、人口值以及地區面積。

該樣例圖形顯示 1991 年澳大利亞各州和地區的人口和人口密度。圖形物件的高度代表了相關的人口數量；顏色代表了相關的人口密度。在顯示結果下部的圖例描述顏色範圍及其相關值。

- `population.canada.mapviz`、`population.canada.data`、`canada.provinces.gfx` 和 `canada.provinces.hierarchy`
該資料檔案所為加拿大的每個省和地區包含一條記錄。在該範例中，每列包含 13 個以空格分隔的值 (每個值代表 1871 至 1991 年之間的每個十年)。

該樣例圖形顯示從 1871 至 1991 年間，以十年為間隔加拿大各省和地區的人口和人口密度。動畫控制面板可讓您動態查看一段時間後的資料集的變化。動畫操作在[第 10 頁](#)的「[動畫控制面板](#)」中有說明。

- *population.europe.mapviz*、*population.europe.data*、*europe.countries.hierarchy* 和 *europe.countries.gfx*
當圖形顯示出現時，所顯示的為 1992 年中歐和西歐國家的人口與人口密度。
- *population.usa.mapviz*、*population.usa.data*、*usa.state.gfx* 和 *usa.state.hierarchy*
當圖形顯示出現時，會顯示從 1770 至 1990 年間美國的人口與人口密度。動畫控制可讓您根據時間動態查看人口和密度的變化。
- *population.usa.city.mapviz*、*population.usa.city.data*、*usa.state.gfx*、*usa.state.hierarchy* 和 *usa.city.gfx* 和 *usa.city.hierarchy*
usa.state.gfx 檔案指定美國，顯示為背景。*usa.city.gfx* 檔案確定在該背景中的城市位置。*.data* 檔案說明每個城市的人口。

該樣例圖形顯示 1950 至 1990 年間美國最大的 48 個城市的人口。沒有資料對照為顏色。動畫控制可允許您根據時間動態查看人口變化。

- *perhouse.perage.mapviz*、*perhouse.perage.data*、*usa.state.gfx* 和 *usa.state.hierarchy*
該樣例以圖形顯示從 1988 年 7-8 月到 1991 年 5-6 月間用戶的家庭消費資料。顏色對照到消費者家庭成員的性別；高度代表指定時期和年齡群組中每個家庭平均花費的美元數量。該資料有兩個獨立維度：時間和年齡主視窗中顏色濃度最高的區域表示最高的消費，也就是「1989 年 5-6 月 (年齡：30-39)」以及「1990 年 5-6 月 (年齡：30-39)」。
- *telecom.mapviz*、*telecom.data*、*usa.city.lines.gfx*、*usa.city.lines.hierarchy*、*usa.state.gfx* 和 *usa.state.hierarchy*
該樣例以圖形顯示帶有弧線的平面地圖。這些線連接兩個端點。這些線具有不同的數值、寬度和顏色。在該範例中，寬度和顏色為隨機；但是它們與數量和端點之間連接的持久度相關。

- *fasta.m.data*、*fasta.m.mapviz*、*fasta.m.gfx* 和 *fasta.m.hierarchy*

該樣例資料檔案包含兩個完整染色體組 (承蒙 Tom Flores 博士之允許發表歐洲生物資訊研究所) 間序列比較的部分結果。當進行圖形顯示時，科學家可以很快找出並確定兩組基因之間相似的區域。如上所示的在可視化資料探索中顯示大量資訊的能力可以被擴展以用來加入更多有關個體基因的資訊。科學家可以更容易地探索資料，因此可以更好理解相似基因序列的目的和功能。

在該樣例中，「圖」是生物有機體的圓形基因，稱為**生殖道菌原體 (MG)**。MG 基因分為 500 個相等區段，每個代表基因中 1000- 核苷序列。滑動桿選擇第二個基因的一段，稱作**流感嗜血菌 (HI)**，用於兩個基因之間的交叉比較。「動畫控制面板」中的「匯總視窗」指出哪一區段顯示出了最大的相似性，且您可以移動滑動桿來仔細研究那些有趣的特殊區段。因此「圖」上長條的高度和顏色指出每個 MG 段和每個 HI 段的相關相似性，而較高的長條對應於相似性的較大度量。相似性透過「倒數值」來衡量，其範圍從 0.0 到 1.0。

選項樹樣例檔案

下面的樣例顯示選項樹可能有用時的情況。這些樣例都與 MineSet 所提供的樣例資料檔案相關。執行導入工具，您可以產生如下所述的 *-odt.treeviz* 檔案。本文所述的每個任務，其方案和目標在「樹狀可視化工具」的「樣例檔案」中都進行了描述。在此我們說明了幾個範例資料集選項樹的優缺點。

注意： 擁有 *.schema* 附檔名的分類器可視化資料檔案，位於客戶端工作站的 *data* 目錄中。分類器可視化檔案，帶有 *-odt.treeviz* 附檔名的可視化檔案，在客戶端工作站的 *examples* 目錄中。要將資料檔案載入 MineSet，請開啓 *.schema* 檔案。

- Windows 使用者可以在 MineSet 安裝目錄的 `\examples` 和 `\data` 子目錄下找到這些檔案。
- IRIX 使用者可以在 `/usr/lib/MineSet/examples/treeviz` 和 `/usr/lib/MineSet/data` 中找到這些檔案。

客戶波動

該資料集的「選項樹」顯示出白天總費用、白天總分鐘數及客服電話對於根來說是好屬性：它們幾乎具有相同的估計誤差率。根據對資料的理解及其喜好設定，您可以選擇下降到某個子樹。請注意：當右子樹開始於客服電話時，第二個檢驗在日總費用或日總分鐘數（作為根的左選項）上進行。但是，由於已經在一個屬性上出現了拆分，所以臨界值將不會相同。

轎車產地

該資料集的「選項樹」顯示幾個對於根的好屬性，它們包括：體積、汽缸、重量 (lbs)，速度 (mpg) 及品牌。請注意：根的估計誤差其任何一個子節點都要低。

蝴蝶花分類

這是一個「選項樹」應用效果比「決策樹」差的樣例。「決策樹」的根擁有 6% 的錯誤，而「選項樹」的根具有 8% 的錯誤，「選項樹」的效果似乎差一些。但是：

- 誤差估計的標準偏差相當高：3.88% 和 3.39%。在統計中的經驗估計法認為如果偏差小於兩倍標準差，在 95% 的置信水平下，此偏差在統計上並非顯著的。此範例中，兩種分類模式間 2% 的偏差甚至不大於單倍標準差；因此，在 95% 的置信度水平下，分類誤差率很可能不具統計差異。

- 至於小檔案 (蝴蝶花有 150 筆記錄)，不同隨機子會有不同結果。例如，在不改變「決策樹分類器」誤差率的情況下將隨機子變為 3，可以將「選項樹分類器」的誤差從 8% 降至 4%。這並不意味著已產生更準確的分類器，因為其誤差估計並不穩定。因為只有 50 筆記錄用於檢驗，每一個錯誤就要占 2% 的誤差率。4% 和 8% 之間的差別只是多犯了兩個錯誤。
- 至於小檔案 (蝴蝶花只有 150 筆記錄)，使用 MineSet 中的「誤差估計」選項。它在更窄的置信度間隔內會產生較好的估計。當您執行該模式，狀態視窗顯示「決策樹」分類器的誤差率為 4.67% +/- 1.73%，而「選項樹」分類器的誤差率為 4.00% +/- 1.61%。在該樣例中，差異並不明顯，但是「選項樹」效果略優。
- 即使「選項樹」的誤差率較高，但在分配機率估計方面，它們可能 (通常也是) 比較好的。對於該資料集，「決策樹」的估計均方差為 3.94；「選項樹」為 3.67 (雖然 95% 的置信等級的差異並不明顯)。

蘑菇分類

該資料集的「選項樹」顯示在根節點處選擇的五個選項而其總估計誤差率為 0。觀察該結果，您也許更傾向於使用左選項 (菌斑)，因為雖然準確性相似，但它比氣味更容易檢驗 (所導入「決策樹」的根檢驗)。您也許想移除氣味和菌褶大小兩個屬性，然後為了準確性 (使估計誤差率為 0%) 再建立正規的「決策樹」。

然而請注意，移除根選項而由「決策樹」所選取的兄弟選項來取代並不一定會產生「選項樹」所顯示的具有相同準確率的分類器。被刪除的屬性也許是樹的下會要用到的屬性。例如，從轎車資料集中刪除品牌屬性將明顯地增加誤差率，即使是在根部，五個選項中有四個根本沒有用到它。

黨派隸屬

該資料集與 蝴蝶花資料集的表現十分相似。「選項樹」與「決策樹」有同樣的誤差率。在「估計誤差」模式下，交叉檢驗估計顯示在誤差率和平均的均方根誤差方面，「選項樹」比「決策樹」略好一點 (但在 95% 的置信水平下並不顯著)。

乳癌診斷

對於分類器和誤差估計以及估計誤差模式，「選項樹」的誤差率比「決策樹」略低，但是差異並不明顯（在 5% 置信水平下）。

甲狀腺機能衰退診斷

該資料集的誤差率非常低（小於 1%），因為被檢查是否患有甲狀腺機能衰退的大部分（95%）皆未患有此病。如果我們利用損失矩陣來避免錯誤否定（對實際患有甲狀腺機能衰退但診斷為否定的情況給予 100 分的懲罰），我們可以看到「選項樹」的損失將明顯低於「決策樹」：182 比 523（全部），或 0.17 比 0.5（每筆記錄）。在 95% 的置信水平下，差異很明顯。

DNA 邊界

至於該資料集，「選項樹」比「決策樹」略微準確；但是觀察根選項後，您會注意到它會選擇左 1、2，和右 1、2、5。如果更接近已知背景知識屬性的邊界的屬性更重要，您可能會想要在右 5 上排除選項分割。在您將根選項的最大數目更新為 4（自 5 向下），錯誤率從 5.65% 漲至 6.59%。如果根不再將右 5 作為選項來使用，這也許很奇怪；而將根選項數目從 5 改為 4 的另一個結果是減少在樹的深層結構中所出現的選項數目（因為降低了參數）。這樣使得其它 4 個子樹的個別誤差率增加。但是，選項樹的誤差率（在 95% 置信水平下）仍然明顯低於「決策樹」的誤差率 7.06% +/- 0.79%。

回歸樹樣例檔案

下述樣例描述了可以使用回歸樹的情況，並且反白顯示「回歸樹導入工具」的一些功能。這些樣例都與 MineSet 所提供的樣例資料檔案相關。執行導入工具，您可以產生將在下述描述的 *-rt.regress files* 檔案。

注意： 擁有 *.schema* 附檔名的資料檔案，位於客戶端工作站的 *data* 目錄中；而回歸可視化檔案，帶有 *-rt.treeviz* 附檔名，位於客戶端工作站的 *examples* 目錄中。要將資料檔案載入 MineSet，請開啓 *.schema* 檔案。

- Windows 使用者可以在 MineSet 安裝目錄的 *\examples* 和 *\data* 子目錄下找到這些檔案。
- IRIX 使用者可以在 */usr/lib/MineSet/examples/treeviz* 和 */usr/lib/MineSet/data* 中找到這些檔案。

客戶波動

客戶波動資料集包含在電信公司用戶的通話模式上產生的資訊。在分類範例中，該資料集用於確定是何種因素導致用戶波動，或離開該公司轉到其他的競爭對手處。在回歸範例中，我們將試圖確定是何種因素影響每天公司從用戶手中索取的費用。

檔案 *churn-rt.treeviz* 顯示在該資料集上產生的並用於預測白天總費用的「回歸樹」。有趣的是，整個樹只在一個屬性（白天總分鐘數）上產生分支；它連續不斷深入地分隔該屬性成更小的範圍。這是因為白天總分鐘數正比於白天總費用 - 用戶為使用該系統的時間支付費用。「回歸樹」能夠反映這個事實。

轎車耗油量

轎車資料集包含從 70 年代到 80 年代初轎車的不同型號資訊。該資料集中的屬性包括重量、提速能力以及每加侖英里數 (mpg)。檔案 *cars-rt.treeviz* 顯示在該資料集上導入的「回歸樹」回歸工具，並將每加侖英里數作為連續標籤。

按一下頂部的節點，我們可以看到資料集中轎車的平均 mpg 在 23.5 左右。在資料集的「回歸樹」中第一個分割顯示對轎車英里數最大的因素是轎車本身的重量。「決策樹」發布了眾所皆知的事實，就是越重的轎車其英里數越低。透過觀察基準節點的兩個子節點，我們注意到右邊的子節點比左邊的更藍一些，這就是說它的每加侖英里數較少。反白標示該節點，我們可以看到當轎車的重量小於 3018lbs. 時，其 mpg 大約為 28.3，而當重量大於 3018lbs 時大約為 16.6mpg。

至於馬力更強的轎車，我們可以看到下一個分割是在轎車的馬力，而且馬力越大的轎車其應哩數越低。下一層的分割出現在轎車製造的年份上，可以看出越新的轎車英哩數越大。現在，讓我們搜尋一個不尋常的轎車使用篩選面板，我們試圖發現一個節點其平均 mpg <24 而最大值 >30。在篩選過程中，我們很快將樹減少至一個節點，而轎車重量小於 3018 lbs，功率大於 77 馬力，生產於 1980 以前。在這個範圍裡只存在一輛不尋常的轎車；選擇節點上最右邊的條並選項 ">" 顯示原始資訊" 菜單項，我們看到這輛是 1978 年 Dodge 型，重量大約 2000lbs，功率 83 馬力，它每一加侖汽油可以多跑 33.5 英哩。

薪資因素

成人資料集是從美國統計局處摘錄的，其中包含職業成人的相關資訊。它包含年齡超過 16 歲、每周工作至少一小時、每年收入超過 \$100 的人之相關資訊。我們可以利用「回歸樹導入工具」來確定是何種因素在影響一個人的薪資；並在已知其它資訊的條件下，對其應得的薪資水平進行粗略的估計。

檔案 *adult-rt.treeviz* 顯示在成人資料集上導入的「回歸樹回歸工具」，並將總收入作為連續標籤。請注意：該資料集非常大（大約有 50000 筆記錄），因此在該資料集上導入回歸工具的過程要花費幾分鐘的時間。

頂部節點的長條圖是表示統計局資料中薪資值的柱狀圖。請注意：可用資料量隨薪資層的降低而逐漸減少。我們有許多關於每年收入為 \$3,000 的資料，但是隨著圖的深入它們在減少。在柱狀圖的最後部分該趨勢發生逆轉，指出存在相當數量的每年收入大約為 \$100,000 的資料。這個差別也許是資料本身存在的趨勢，也許是因為採樣存在偏差。

「回歸樹」中第一個分割產生在年齡屬性上。正如所預期的，年輕人所賺的錢比年長者少。從頂節點及其兩個子節點，我們可以看到有關這三組人群的一些匯總統計資料。我們注意到在該項研究中，每個人的平均薪資在 \$33,500 左右，而年齡小於 27 歲的人其平均薪資為 \$14,300 左右；年齡大於 27 歲的人其平均薪資在 \$40,000 左右。

在年齡小於 27 歲的下兩次分割中，我們看到回歸樹先將他們分成兩個種類：23 歲及以下的人群和 23 歲以上的人群。有意思的是，這兩個分支上的後續分割都產生在每週工作屬性的小時上，這顯示出對於這兩個年齡範圍來說，人們工作的時間越多，總收入也就越高。

現在將注意力放在年齡超過 27 歲的人群，我們發現在所受教育程度上迅速產生三個分割。受過 13 年以上教育的人（與學士學位相對應）會賺更多的錢。透過觀察被教育程度分割的兩個子節點，我們會發覺每年收入的 \$90,000 左右的人都至少接受某種進階的教育。

我們可以使用篩選面板，迅速找出每年平均收入超過 \$50,000 的人群的位置。在篩選面板中，選擇平均值 >50000。頂層節點在該篩選過程中消失，這是因為要賺如此多的錢並不是常見的事。具有學士學位且年齡超過 27 歲者落入該範圍內。沿著第一個分割的左分支一直到末端，我們在該範圍內發現另一組人群：年齡超過 36 歲的已婚男性，每周工作超過 35 小時，並且受過良好的教育（10 年或更多）。

如果我們再次使用篩選工具，尋找絕對偏差大於 \$25,000 的節點，我們可以發現，經濟條件表現出的變化範圍最大的人群。在該篩選過程中第一個被保留的節點是那些年齡超過 27 歲且具有學士學位的人群。該節點上的柱狀圖顯示以平均值為中心而分布的屬性，但只具有年收入為 \$100,000 左右人群的數目。

蝴蝶花

在該資料集中的每筆記錄都描述了蝴蝶花的五個特徵：花萼寬度、花萼長度、花瓣寬度、花瓣長度以及蝴蝶花的類型。該回歸過程的目的是根據其它特徵來預測花萼寬度。檔案 *iris-rt.treviz* 顯示為預測花萼寬度而在該資料集上執行「回歸樹導入工具」的結果。

觀察頂部節點，我們可以看到在花萼寬度值中存在一個間隙，其中沒有花存在。在該資料集上分割的「回歸樹導入工具」首先使用花萼長度變數。花萼長度小於 2.6 時，可能的花萼寬度子集很有限。另一方面，花萼長度值大於 2.6 指出了一個更大花萼長度的均勻分布。那些花萼長度小於 2.6 的蝴蝶花其平均花萼寬度為 0.24，而對應那些花萼長度大於 2.6 的蝴蝶花其花萼平均寬度為 1.68。對於具有大花萼長度的蝴蝶花，我們可以看到在花萼長度上再次有三個分割，這一次的臨界值為 4.85。相同變數上兩個連續的分割說明在這兩個變數中可能存在著某種受限制的功能性關係。

回到那些花萼寬度小於 2.6 的蝴蝶花，我們可以看到下面的分割是在花瓣寬度屬性上進行的。有趣地，樹中這部分的值似乎已分隔開了，那些花瓣寬度小於 3.25 的蝴蝶花呈現出的值位於三個狹窄而且分開的範圍內。

糖尿病診斷

該資料集是關於糖尿病的診斷問題，它是利用統計學亞利桑那州鳳凰城的中印地安部落收集而來。檔案 *pima-rt.treeviz* 顯示在該資料集上執行「回歸樹導入工具」的結果，並將血糖水準作為待預測的連續變數。

該樹的第一個分割出現在糖尿病指標上，顯示出患有糖尿病的人其血糖水平比正常人高 (141 比 110)。下一個分割是依 2 小時血清胰島素屬性進行的，當其值高於 125 時將導致高血糖。

「回歸樹」從頂部節點開始遵循每個節點處的決策進行預測，對新記錄進行檢查。例如，如果某個糖尿病病人的 2 小時血清胰島素為 110，那麼就可以預測其血糖水平為 105。

分散可視化工具樣例檔案

這裡提供的樣例資料檔案和配置檔案展示「分散可視化工具」的特性和功能。下面的 *.data* 和 *.scatterviz* 檔案位於 *examples* 目錄中。要將資料檔案載入 MineSet，請開啓 *.schema* 檔案。

- Windows 使用者可以在 MineSet 安裝目錄的 *\examples* 子目錄下找到這些檔案。
- IRIX 使用者可以在 */usr/lib/MineSet/examples/scatterviz* 中找到這些檔案。

「分散可視化工具」樣例檔案如下所示：

- *company.data*
該檔案包含數個保險公司對三個產品種類的假想銷售資料：人壽保險、汽車保險和家庭保險。該資料跨越 10 年（增量為一年），包括了五個收入範圍（用戶的年收入）。
- *company.scatterviz*
該檔案指定年份到一個滑動桿維度，收入範圍到另一個滑動桿。人壽保險、汽車保險和家庭保險的銷售額成為「分散可視化工具」場景觀中的三個維度。滑動桿匯總視窗中的顏色濃度代表所有公司在全部保險業務的總銷售值。
- *company-total.scatterviz*
該檔案包含與 *company.scatterviz* 中相同的內容，不同的是由公司在所有保險業務範圍內的總銷售額確定每個公司的規模。
- *company-life.scatterviz*
該檔案包含與 *company.scatterviz* 中相同的內容，不同的是每個物件的顏色指出作為整個銷售額一部分的人壽保險銷售額。
- *store-type.data* 和 *store-type.scatterviz*
這兩個檔案顯示不同類型的商店在不同產品上的三年期間銷售額。由滑動桿所代表的獨立變數是時間。每個實體代表一種商店類型（例如，食品店、藥局、服務站等等）。資料檔案包含每種商店類型的產品組銷售總額，例如，酒精飲料、麥片等等。資料的時間跨越 36 個月，按月增長。

該配置檔案使用月份作為滑動桿維度。一個坐標軸為酒精飲料的銷售，另一個為煙草製品的銷售額。沒有使用第三個坐標軸。
- *brand.data* 和 *brand.scatterviz*
這些檔案顯示不同類型商店中數種軟性飲料品牌的銷售額。在該資料集中，實體的品牌和商店的類型都對照到坐標軸。銷售總額被對照為每個品牌的大小。顏色對照是隨機的。由於沒有獨立的變數，便沒有滑動桿存在。

- *cars.data* 和 *cars.scatterviz*
這些檔案顯示數個轎車模型的重量、馬力、型號年代與加速能力。坐標軸為立方英寸、mpg 及加速到 60 哩的時間。重量對照為實體大小。
- *people.data* 和 *people.scatterviz*
這些檔案顯示了一個假想的人口抽樣中的身高、體重、人口密度以及膽固醇水平。
- *nl.births.data* 和 *nl.births.scatterviz*
這些檔案顯示荷蘭的人口出生模式。顯示每個區域的人口密度、出生率以及人口總數。動畫滑動桿對照至母親的年齡和年份。
- *adult94.data* 和 *adult94.scatterviz*
這些檔案顯示使用分散可視化工具應用於 *adult.data* 的複雜範例。可視化過程中的三個坐標軸為 *avg_hrswk* (即每周平均的工作時數)、*avg_gross_income* 和 *avg_education_num*。不幸地是「教育數」與受教育的年數並不完全對應，但很接近。右邊的滑動桿可在不同的年齡範圍內進行動畫過程。按照職業、種族和性別的分組來建立每個組合結果。這就意味著對於這三個屬性存在一個對應每個取值組合的實體。顏色正如圖例所示，代表不同的職業。每個實體的大小對應於記錄計數。匯總滑動桿同樣按照資料密度進行著色。要想知道該可視化過程如何建立，您可以從「檔案」功能表中選擇「啟動工具管理員」。這將開啓「工具管理員」，並在這期間建立該樣例。

初始時，畫面顯示年齡小於 20 歲者的有關資訊。請注意：平均工作時數 (約為 14) 和平均收入 (約為 \$4000) 較低。如果您利用滑動桿在年齡屬性上進行動畫過程，並且從三個垂直的視角觀察畫面 (使用主視窗右下的三個按鈕)，您將注意到有不同的趨勢出現。例如，如果您定向畫面只觀察每周小時數與收入之間關係，您會發現人們工作的時間隨著年齡的增長而延長一直到 25 歲，然後每周工作的時間很少超過 49 小時，直到退休。然而，收入將一直增加直到 50 歲，便處於平穩狀態，然後再次降低。實際的趨勢多少與職業的選擇和其它的因素有關。

假設您對手藝維修和特殊行業兩種職業間的比較感興趣。可以開啓「篩選」面板 (檢視 > 顯示篩選面板) 並從職業清單中選擇「手藝維修」和「特殊行業」。現在，當您移動時，可以看到從事「特殊行業」的人在實際中養往開始於低收入，但是隨著年齡的增長很快趕上從事「手藝維修」的人。「特殊行業」在教育坐標軸上比「手藝維修」高很多。您也許希望只顯示女性或者特定的種族來進一步限制篩選過程。同樣地，或者在動畫過程中試圖選擇一些不同的運動軌跡。

- *census.data* 和 *census.scatterviz*

這些檔案也顯示組合統計資料的圖形。原始資料集包含大約 150,000 列。在組合過程後，對於教育、性別、企業 1 及職業 (它們作為分組依據欄) 的每個組合會出現一個立方體 (一個組合結果)。

平板可視化工具樣例檔案

這裡提供的配置檔案和資料檔案展示「平板可視化工具」的特性和功能。這些檔案在 *examples* 目錄中。

Windows 使用者可以在 MineSet 安裝目錄的 *\examples* 子目錄下找到這些檔案。

IRIX 使用者可以在 */usr/lib/MineSet/examples/splatviz* 中找到這些檔案。

- 蘑菇

mushroom.data 檔案包含超過 5,000 個蘑菇的預先組合資料。分組依據欄為：**odor**、**gill_color** 和 **cap_color**。對於原始資料中這三個欄的每個組合，都有一個計數和一個平均可食性指標，當其為 0 時表示可以食用，為 1 時則有毒。因為蘑菇不可能部分有毒，因此在 0 到 1 之間的可食性指標意味著在組合結果中，有些蘑菇可以食用，而另一些有毒。

該可視化過程機將每個欄的值都根據平均可食性沿著坐標軸進行排序。很明顯地，氣味是可食性的最好判定因素。請注意：大多數的平板不是 0 就是 1，這意味著這三個欄在對兩類蘑菇進行劃分時都有用處。實際上，欄重要性功能選擇了對照到坐標軸上的欄。向下移動透明度滑動桿可以確定擁有最高計數的平板。最透明的平板代表 288 只具有共同氣味、**gill_color** 以及 **cap_color** 值的蘑菇。要進行確認，可以用 **sum_count_poison>280** 進行篩選並選中剩下的平板以觀察它們的計數。請注意：所有 **gill_color=buff** 的蘑菇都是有毒的。

- adultJobs*

adultJobs.data 檔案從 *adult94* 中導出，*adult94* 是個發行版中提供的資料集。它使用依據教育、職業、每周工作時數 (已分組) 及年齡 (已分組) 分組的組合結果來進行建立。*gross_income* 欄按照計數和平均進行組合。對於使用「平板可視化工具」來得到的顯示結果，*age_bin* 對照為滑動桿，將其它分組欄對照為坐標軸。*count_gross_income* 欄對照為透明度，而 *avg_gross_income* 對照為顏色。

當滑動桿處在最左邊的位置時，圖的顏色幾乎完全是藍色的。這就意味著不管職業、教育程度或工作的小時數，大多數小於 20 歲的年輕人收入較低。向右移動滑動桿，可以注意收入隨著更高的教育程度和不同的職業如何朝坐標軸的頂端快速升高。經由透明度的變化，您可以看到最為普遍的教育程度類型是 HS，即大專或學士學位。

移動「匯總」滑動桿可以顯示收入的分布是如何按照年齡坐標軸欄變化的。
- adultJobs2*

adultJobs2 也是根據 *adult94* 資料集。在此，坐標軸表示工作類別、教育程度以及職業。兩個對照為滑動桿的欄為年齡 (已分組) 和每周的工時 (已分組)。收入再次被計數和平均組合併，分別對照到顏色和透明度。因為在 2D 滑動桿上有更多的位置，因此代表每個位置的紀錄就較少。這會造成顏色和透明度的更大變化。「匯總」滑動桿 *hrs_per_week* 維度中間的紅色區域指出幾乎所有的人每周的工時都在 35 和 45 之間。請注意：一些職業與特定的工作類別有固定的關聯。例如，軍隊中每個人的工作類別都為 Fed-Government。
- censusIncome*

這個範例根據一個與 *adult94* 相似的資料集，不過由於其規模較小，這個資料集沒有包括資料的分布資訊。要試圖理解毛收入 (*gross income*) 和總收入 (*total income*) 間的差別，必須將 *gross_income*、*total_income* 和 *hrs_per_week* 對照到坐標軸。顏色表示年齡。經由研究影像，我們可以知道有很多記錄為 *total_income=gross_income*，但是也有不少比例的記錄具有較高的 *total_income*，而 *gross_income* 為 0。令人驚奇的是，在很多情況下 *gross_income* 大於 *total_income*。

注意不同年齡的人聚群所類集中的地方。許多老人 (黃色) 在 `hrs_per_wk=0` 的平面內。他們應該是退休人員。許多子節點和年輕人 (藍色) 在 `gross_income=total_income=0` 的線上。注意，非常透明的平板位於該範圍的外邊緣的附近。這些位置包括所有落入坐標軸上的最大組中的點。例如，`total_income` 最高的組為 70,300+。高於 70,300 的任何點都進入了該組。

要更好地觀察變化的密度，請調節透明度滑動桿。在低的不透明性等級中，對角線表示出大多數人 `gross_income=total_income`，或他們只有 `total_income` 而沒有 `gross_income`。當您提高等級時，您可以看到幾乎所有的卷冊都包含資料。該資料集包含 150,000 筆記錄。

- *客戶波動*

客戶波動是指用戶離開一家公司轉向另一家公司。該樣例顯示電話公司用戶的波動。用於產生該樣例的資料是 `churn.schema`。

使用欄重要性，我們可以發現 `total_day_charge`、`number_customer_service_calls` 和 `international_plan` 是重要的區別因素。這些欄對照為坐標軸。然後我們建立一個新數字型欄，`churn`，相同於 `churned==Yes`，將它對照為顏色。

在可視化結果中，紅色區域表示較高的波動可能。該區域該區域表示撥過三通以上客服電話，但 `total_day_charge` 很低的用戶對應於高 `churn` 值。您也許想對高消費用戶進行加權，賦予比其它用戶更高的權重。要達到此目標，建立一個新欄，`total_charge` 等於

```
`total_day_charge`+`total_eve_charge`+`total_night_charge`
```

或該總和的冪。然後將 `total_charge` 欄對照為顏色。這就意味著每個記錄按照 `total_charge` 進行了加權。現在，可視化過程顯示 `total_day_charge` 坐標軸高端附近令人感興趣的區域。

樹狀可視化工具樣例檔案

這裡提供的配置檔案和資料檔案展示「樹狀可視化工具」的特性和功能。這些檔案在 *examples* 目錄中。

- Windows 使用者可以在 MineSet 安裝目錄的 `\examples` 子目錄下找到這些檔案。
- IRIX 使用者可以在 `/usr/lib/MineSet/examples/treeviz` 和 `/usr/lib/MineSet/data` 中找到這些檔案。

「樹狀可視化工具」樣例檔案如下所示：

- *store.data* 和 *store.treeviz*
當圖形顯示出現時，這些檔案會顯示有關連鎖店的假想銷售資料。樹中的等級包括所有商店、區域、州、城市以及個別商店。分層結構中的每一個等級皆顯示四個產品。在該配置中，高度代表銷售額；顏色代表完成目標美元量的百分比。
- *stateRevenue.data* 和 *stateRevenue.treeviz*
當圖形顯示出現時，這些檔案會顯示 1992 年度每個州的財政預算成分，該資料可從美國統計局中得到。(從 <http://www.census.gov/govs/state/stfin92.dat>)。高度代表了稅收額。背景中下降的節點顯示各種稅收對顯示在根節點中的整體收入的貢獻。
- *beer.data* 和 *beer2.data*，和 *beer.treeviz* 和 *beer2.treeviz*
當圖形顯示出現時，這些檔案會顯示根據啤酒用戶調查假想資料。樹包含三個層：
 1. 第一層是關於種類的 (例如，啤酒或淡啤酒)。
 2. 第二層是品牌代碼 (隨機分配)。
 3. 第三層是個別產品代碼；例如：12 包裝與 6 包裝 (隨機分配)。

每個圖包含七個長條，代表七個年齡組。長條高度代表了該年齡組的消費額。顏色代表被男性和女性消費的百分比。品牌、產品以及在這些檔案中使用的資料只是樣例。

beer.treeviz 和 *beer2.treeviz* 產生相同的圖形結果，但是它們是按照不同的方式建構而成。在 *beer.treeviz* 中，每種類型的啤酒由一個記錄代表，每個記錄代表男性和女性的消費量；這些值儲存在一個枚舉陣列中。

在 *beer2.treeviz* 中，每種啤酒有七筆記錄，每筆記錄代表一個年齡組。請注意：在 *beer* 檔案中，年齡組包含在配置檔案中；在 *beer2* 檔案中，它們包含在資料檔案中。

beer 檔案比 *beer2* 檔案所需的儲存空間少；但是配置檔案略微複雜一些。在一些情況下，以 *beer2* 檔案所使用的格式來產生資料可能更容易一些。

索引

Symbols

% 最短選項, 181
* 萬用字元, 96, 188
? 萬用字元, 96, 188
[] 萬用字元, 96, 188

Numbers

0 指令, 192
0 值
物件與, 192
2D 組合, 12
3D 景觀, 155, 177
64 位元支援, 134
systune 參數, 134

A

adultJobs.data, 251
adult-salary.dtableviz, 218
adult-salary.eviviz, 231, 232, 233, 235, 236, 237
adult-salary.schema, 231
adult.schema, 206, 217, 230
adult-sex.dtableviz, 217
adult-sex-dt.treeviz, 206
adult-sex.eviviz, 230
anl.births.scatterviznl.births.scatterviz, 249

australia.states.gfx, 238
australia.states.hierarchy, 238
avg 關鍵字
顏色值與, 182

B

beer2.data, 253
beer2.treeviz, 253
beer.data, 253
beer.treeviz, 253
blocks.data, 238
blocks.gfx, 238
blocks.hierarchy, 238
blocks.mapviz, 238
brand.data, 248
brand.scatterviz, 248
breast.dtableviz, 224
breast-dt.treeviz, 211
breast.eviviz, 235
breast.schema, 224, 235

C

canada.provinces.gfx, 238
canada.provinces.hierarchy, 238
cars.data, 249
cars-dt.treeviz, 205

cars.eviviz, 229
cars.scatterviz, 249
cars.schema, 205, 229
censusIncome 資料檔案, 251
churn-dt.treeviz, 205
churn.schema, 252
.clusterviz.data 檔案, 47
company.data, 248
company.scatterviz, 248
company-total.scatterviz, 248

D

dna.dtableviz, 227
dna.eviviz, 237
DNA 資料集, 213, 227, 237
DNA 邊界資料集, 243

E

europe.countries.gfx, 239
europe.countries.hierarchy, 239

F

fasta.m.data, 240
fasta.m.gfx, 240
fasta.m.hierarchy, 240
fasta.m.mapviz, 240

G

gfx 檔案, 117
產生, 119-120
樣例, 238, 239

Gini, 74

H

hypothyroid.dtableviz, 225, 226, 227, 228
hypothyroid.eviviz, 236
hypothyroid.schema, 211, 225, 236

I

Internet 檔案, 199
iris.dtableviz, 222, 224
iris-dt.treeviz, 209
iris.eviviz, 233
iris.schema, 209, 233
Iris 分類資料集, 209, 221, 233, 241

L

LANG, 在國際化中, 105

M

.mapviz 副檔名, 118
max 關鍵字
顏色值與, 182
MineSet
工具
總覽, 175
MINESET_WARN_EXECUTE 變數, 199
MineSet 的 mtr 附檔名, 199
mtr 檔案, 199
mushroom.data, 250
mushroom.dtableviz, 222
mushroom-dt.treeviz, 210

mushroom.eviviz, 233, 235, 236, 237
mushroom.schema, 210, 222, 233

N

Naive-Bayes, 125
nl.births.data, 249

P

people.data, 249
people.scatterviz, 249
perhouse.perage.data, 239
perhouse.perage.mapviz, 239
pima.dtableviz, 226
pima-dt.treeviz, 212
pima.schema, 212, 226, 237
Pima 糖尿病診斷資料集, 212, 226, 236
population.australia.data, 238
population.australia.mapviz, 238
population.canada.data, 238
population.canada.mapviz, 238
population.europe.data, 239
population.europe.mapviz, 239
population.usa.cities.data, 239
population.usa.cities.mapviz, 239
population.usa.data, 239
population.usa.mapviz, 239

R

ROI 曲線, 101, 143
ROI 曲線選項, 101

S

.scatterviz 檔案名, 146
ScatterViz 選項對話方塊, 148-151
.splatviz.data 檔案, 161
.splatviz.schema 檔案, 161
stateRevenue.data, 253
stateRevenue.treeviz, 253
store.data, 253
store.treeviz, 253
store-type.data, 248
store-type.scatterviz, 248
systune 參數
 64 位元支援與, 134
 rlimit__nofile_cur, 134
 rlimit__rss_cur, 134
 rlimit__vmem_cur, 134
 rlimit_pthread_cur, 134

T

telecom.data, 239
telecom.mapviz, 239
.treeviz 副檔名, 178

U

UNIX 指令, 122, 151, 183
UNIX 啟動指令
 關聯規則, 24
usa.cities.gfx, 239
usa.cities.hierarchy, 239
usa.cities.lines.gfx, 239
usa.cities.lines.hierarchy, 239
usa.states.gfx, 239
usa.states.hierarchy, 239

V

vote.dtableviz, 224
vote-dt.treeviz, 211
vote.eviviz, 235
vote.schema, 235

W

-warnexecute 選項, 199
Web 檔案, 199
「What if」問題, 84

X

.Xdefaults 檔案, 199
X 滑動桿, 150

Y

Y2K 相容性, 200
Y2K 應變能力, 200
Y 滑動桿, 150

二畫

人口採樣, 138
人口樣例檔案, 238, 239
十番, 74

三畫

下載 Internet 檔案, 199
上升, 22

上升曲線, 17, 111

子節點
選擇, 193

子樹權重選項, 77

小值, 篩選出, 190

工具
多重選擇與, 124
總覽, 175

工具管理員

工具選項
平板可視化工具, 161
地圖可視化工具, 121-123
分散可視化工具, 148-152
樹狀可視化工具, 178-185

多重規則
圖, 26

配置選項
平板可視化工具, 161
地圖可視化工具, 119-123
決策樹導入工具, 76
証據可視化工具, 87
樹狀可視化工具, 185
關聯規則, 24

已計算的欄, 1

四畫

不區分大小寫查詢, 188
不區分大小寫篩選工具, 190
互補追溯指令, 153, 192
分析
關係, 145, 155, 177
分析模式和趨勢, 20
分組, 34
分組欄按鈕, 34
分組欄選項, 62
分割下限選項, 75

分割標準選項, 74, 167

分類

分配記錄, 71, 84, 131

分類順序

指定, 184

分類標籤, 100

查詢, 77

分類器, 38-39

上升曲線, 111

已定義, 38

投資回報曲線, 101

修正, 33, 101

記錄加權, 102, 138

產生, 38, 72, 85, 131

損失矩陣, 102, 114

準確性, 80

檢驗, 100

預測未知值, 113

混淆矩陣, 58-60, 101

學習曲線, 102, 108

選項, 109-110

檢視輸出結果, 103

應用於記錄, 33

檢視輸出結果, 102

欄重要性與, 55

分類器和誤差模式, 34, 100, 101, 111

檢視輸出結果, 103

分類器選項對話方塊, 73

分類類型, 81, 82

升冪排列順序, 184

反白標示的物件

分散可視化工具, 153

反向播放按鈕, 13

天空顏色, 184

支持分類, 81, 101

支持比率, 101

支援, 22

最小臨界值, 22

支援大量記憶體, 134

systune 參數, 134

文件

印刷慣例, xviii

日期

導入工具與, 102, 104

日期, Y2K 相容性, 200

比較

字串

篩選類型與, 96

資料集, 181

父按鈕, 193

五畫

主視窗

地圖可視化工具

標記, 122

決策表, 69, 80

統計可視化工具, 170

證據可視化工具, 91

加拿大地圖, 120, 238

加權指令, 71, 129

功能表

平板可視化工具, 165

決策表可視化工具, 70

統計可視化工具, 170

證據可視化工具, 91-92, 131, 144

樹狀可視化工具, 185

包含查詢選項, 96, 188

另存新檔

記錄檢視器, 138

市場供求分析, 21

平方差, 142

平板

已定義, 155

標記, 161

顏色選項, 160

- 繪圖選項, 160, 165
- 顯示, 160
- 平板可視化工具, 252
 - 功能表, 165
 - 組合資料點, 157, 158
 - 空值與, 162
 - 重設, 預設, 161
 - 配置
 - 「工具管理員」與, 161
 - 動畫控制面板, 163
 - 匯總視窗, 160
 - 啟動選項, 159
 - 開始, 159
 - 重設預設值與, 161
 - 資料檔案, 158
 - 樣例檔案, 166, 250-252
 - 選項, 161
 - 重設, 161
 - 檔案需求, 158
 - 顏色對照, 160
 - 顯示資料, 155, 165
- 平板可視化工具選項對話方塊, 161
- 平板形狀選項, 160
- 平板選項, 160
- 平板顏色選項, 160
- 平板類型功能表, 165
- 平均誤差 / 損失標準偏差選項, 77
- 平面地圖, 239
- 必需的檔案
 - 平板可視化工具, 158
 - 地圖可視化工具, 117
 - 分散可視化工具, 146
 - 樹狀可視化工具, 177
 - 選項樹導入工具, 132
- 本文欄位
 - 用於選擇欄的臨界值是, 36
- 未標記的記錄, 81
- 正規化高度, 180-181
- 用戶消費樣例檔案, 239

- 用戶調查樣例檔案, 253
- 甲狀腺機能衰退診斷資料集, 211, 225, 236, 243
- 目前檢視, 171
- 目前欄本文方塊, 6
- 目前欄視窗, 62

六畫

- 交叉檢驗分類, 61, 82, 101
- 交互資訊選項, 74, 167
- 任意關鍵詞
 - 顏色值與, 182
- 光滑平面, 116
- 先驗機率, 85, 86
- 列印影像指令, 92, 93
- 印刷慣例, xviii
- 向上移動指令, 193
- 向右移動指令, 193
- 向左移動指令, 193
- 向前執行指令, 193
- 向前播放按鈕, 13
- 向後執行指令, 193
- 組合, 4-6, 177
 - 2D, 12
 - 長條高度, 182
 - 資料點, 157, 158
 - 選項, 5, 6
 - 顏色值, 182
- 組合按鈕, 4
- 組合對話方塊, 4, 5
- 組合選項, 63
- 組合欄選項, 6
- 回歸工具, 139
- 回歸樹導入工具
 - 誤差估計, 142
 - 選項, 140

- 分割下限, 141
- 分割標準, 141
- 成本複雜性修改, 142
- 經由 ... 限制樹高度, 140
- 總覽, 139
- 地區
 - 資源檔案, 105
 - 範例, 106
- 地區, 國際化中, 105
- 地理物件, 117, 121
- 地理區域, 120
 - 訊息, 122
 - 圖例, 122
 - 縮放比例, 121
- 地理檔案選項, 121
- 地圖可視化工具, 240
 - 主視窗
 - 標記, 122
 - 地理位置, 120
 - 空值與, 130
 - 重設, 預設, 121
 - 配置
 - 「工具管理員」與, 119-123
 - 啟動選項, 118
 - 開始, 118
 - 資料檔案, 117, 120
 - 樣例檔案, 240
 - 篩選面板, 95-96
 - 選項, 121-123
 - 重設, 123
 - 啟動, 118
 - 儲存, 123
 - 儲存預設值, 123
 - 檔案需求, 117
 - 顏色對照, 121-122
 - 顯示資料, 116
- 地圖可視化工具選項對話方塊, 121-123
- 在聚類中使用加權, 46
- 在聚類中權重就是屬性, 46

- 在檢視中穿過, 171
- 在「樹狀可視化工具」中查詢的「樣例結果」, 189
- 多重規則, 25
 - 顯示, 26
 - 「工具管理員」與, 26
- 多個值, 124
- 多處理器版本, 73, 132, 134, 135
- 多維資料點, 11
- 字母數字值
 - 查詢, 188
 - 篩選, 96
- 字串
 - 比較, 96
 - 查詢, 188
 - 對照, 156
 - 篩選, 96
- 收入樣例檔案, 253
- 百分比選項, 77
- 自動臨界值
 - 統一範圍, 36
 - 統一權重, 36
- 自動離散算法, 55
- 自動欄選擇選項, 87

七畫

- 估計機率值模式, 16
- 伺服器
 - 連接至, 175
- 坐標軸
 - 標記, 151, 161
 - 隱藏標籤, 161
 - 顯示選項, 150
- 坐標軸標籤大小選項, 151, 161
- 坐標軸選項, 150
 - 不可調節, 151
 - 比例尺寸, 151

- 最大規模, 150
- 快退按鈕, 13
- 快進按鈕, 13
- 投票記錄範例, 210, 223, 234, 242
- 投資回報曲線, 101, 143
- 改變顏色, 50
- 改變類型選項, 63
- 更多分類器選項指令, 73, 88, 132
- 決策表可視化工具, 227
 - 功能表, 70
 - 樣例檔案, 213-227
- 決策樹
 - 分割, 74, 167
 - 空值與, 78
 - 查詢, 76-78
 - 修改, 75
 - 純度測量, 72, 77
 - 設定選項, 73-76, 132-133
 - 節點
 - 檢視資訊, 72
 - 誤差 / 損失估計, 72, 77
 - 篩選, 76
 - 顯示, 102
- 決策樹分類器
 - 查詢物件, 76-78
 - 產生, 72
 - 總覽, 71
- 決策樹導入工具, 76, 213
 - Gini, 74
 - 十番, 74
 - 修改方法
 - 成本複雜性, 76
 - 置信度, 75
 - 配置, 76
 - 樣例檔案, 204-213
 - 調整歸納算法, 74
 - 檢視節點資訊, 72
 - 總覽, 71
 - 欄重要性工具和, 56
- 客戶波動資料集, 203, 205, 214, 228, 241, 252

八畫

- 並行化, 134
- 並行計算, 73, 132, 135
- 乳癌診斷資料集, 211, 224, 235, 243
- 依照模型的資料調整模式, 18
- 依照模型調整資料, 18
- 使用方法功能表
 - 自動, 32
 - 每組最小權重, 35
- 使用加權選項, 102
- 使用損失矩陣選項, 102, 114
- 使用權重功能表, 36
- 使固定選項, 183
- 命名
 - 欄, 2
- 固定選項, 190
- 底色, 184
- 性別屬性資料集, 206, 217, 230
- 拉普拉斯校正選項, 86
- 物件
 - 0 高度與, 192
 - 地理的, 117, 121
 - 空高度與, 192
 - 查詢, 76-78, 186-189
 - 選擇
 - 空值與, 131, 195
 - 樹狀可視化工具, 190
 - 檢視所選取, 124
 - 顯示訊息
 - 地圖可視化工具, 122
 - 分散可視化工具, 151
 - 樹狀可視化工具, 183
- 空指令, 192
- 空值, 152, 194
 - 平板與, 162
 - 決策樹與, 78
 - 物件與, 192
 - 預測, 113

- 對照, 152, 195
- 表格
 - 分類器與, 100
 - 處理選項, 62
 - 儲存資料, 62
- 表格處理視窗, 6
- 表格歷史按鈕, 171
- 表達式, 2
- 長條, 177
 - 固定, 183
 - 查詢, 187
 - 負值與, 177
 - 高度, 180
 - 組合, 182
 - 標記
 - 顏色, 184
 - 縮放比例, 181
 - 顏色選項, 182
 - 根據按鍵, 183
 - 對照到, 182
 - 標籤, 184
- 長條標籤顏色選項, 184

九畫

- 信用資料庫樣例檔案, 202
- 保險樣例檔案, 248
- 按比例篩選指令, 95
- 按字母排序的指令, 71, 129
- 按鈕
 - 樹狀可視化工具
 - 查詢對話方塊, 189
- 按照重要性排序指令, 91
- 按欄分組選項, 6
- 指令行選項, 啟動
 - 關聯規則, 24
- 柱狀圖可視化工具, 98
 - 修改因素, 98

- 柵格
 - 線間隔, 151, 161
 - 顏色選項, 151, 161
- 柵格 (X, Y, Z) 規模選項, 151, 161
- 柵格顏色選項, 151, 161
- 查詢, 76-78, 186-189
 - 指定查詢標準, 188
 - 記錄檢視器, 137
 - 萬用字元, 96, 188
- 查詢面板指令, 76
- 查詢對話方塊, 189
- 查詢聚光燈, 190
 - 關閉, 189
- 爲空操作符, 96, 188
- 相關操作符, 188
 - 篩選資料, 96
- 美國地圖, 120, 239
- 負值, 177
- 重要性 (已定義), 85
- 重要性分級, 56
- 重設工具選項
 - 平板可視化工具, 161
 - 地圖可視化工具, 123
 - 樹狀可視化工具, 184
- 重設選項按鈕, 123, 151, 161, 184
- 重設, 預設
 - 平板可視化工具, 161
 - 地圖可視化工具, 121
 - 分散可視化工具, 148
 - 樹狀可視化工具, 178
- 重新開啓指令, 92, 93
- 重新編號列
 - 記錄檢視器, 137
- 重置標籤大小, 149
- 重複 k- 平均值聚類, 42
- 限制樹高度於選項, 74
- 降冪排序次序, 184

首頁位置, 193
 設定, 193
首頁指令, 193

十畫

修正分類器, 33, 101
修正測試設定選項, 34, 101
修正模型, 34
修改方法
 決策樹導入工具
 成本複雜性, 76
 置信度, 75
修改因素, 98
修改因素選項, 75
修改顏色, 50
書面文件
 印刷慣例, xviii
根據按鍵來分類選項, 184
根據鍵來確定顏色選項, 183
格式
 訊息, 183
浮點數, 55
純度, 53, 72, 77
 檢驗, 54
純度測量, 56, 72, 77
純度選項, 77
記錄
 分配給類, 71, 84, 131
 分類器與, 33
 可用性, 80
 未標記的, 81
 模型與, 18-20
記錄加權, 33, 102, 138
 關聯規則與, 25
記錄檢視器, 136
 另存新檔, 138

查詢, 137
重新編號列, 137
開始, 137
儲存資料, 137
總覽, 136
訊息, 124
 地圖可視化工具, 122
 分散可視化工具, 151
 樹狀可視化工具, 183
訊息選項, 122, 183
訓練集, 19, 100
退出
 樹狀可視化工具, 93
退出指令, 93
迴旋按鈕, 14
追溯
 關聯規則與, 29
追溯資料集
 限制, 79
配置
 平板可視化工具
 「工具管理員」與, 161
 地圖可視化工具
 「工具管理員」與, 119-123
 決策樹導入工具, 76
 証據可視化工具, 87
 樹狀可視化工具
 「工具管理員」與, 185
配置檔案, 175
 平板可視化工具, 159
 載入, 159
 地圖可視化工具, 118
 載入, 118
 樣例, 238, 239
 將 .ruleviz 轉換為 .scatterviz 檔案, 29
 分散可視化工具, 146
 載入, 147
 樣例, 248
 証據可視化工具

- 載入, 87
- 樹狀可視化工具, 177
 - 載入, 92, 93, 178
 - 樣例, 253
- 關聯規則樣例, 202
- 配置關聯規則
 - 「工具管理員」與, 24
- 陣列
 - 地理位置, 120
 - 滑動桿與, 150
 - 導入工具與, 102, 104
- 高度組合選項, 182
- 高度篩選滑動桿, 190
- 高斯形指令, 165, 166

十一畫

- 停止按鈕, 13
- 停用過程對話方塊, 88, 118, 199
- 動畫, 115, 155
- 動畫流程按鈕, 13
- 動畫控制面板, 10-15
 - 按鈕, 13
 - 停止動畫, 13
 - 開始動畫, 13
 - 匯總視窗, 12
- 動畫控制面板 (平板可視化工具), 163
 - 匯總視窗, 160
- 動畫控制面板 (分散可視化工具)
 - 匯總視窗, 150
 - 顯示, 197
- 區別大小寫查詢, 187
- 參數
 - 顯示選項, 192
- 國際化, 107
 - LANG, 105
 - 地區, 105
 - 資源檔案, 105

- 資源檔案樣例, 106
- 設定地區, 105
- 資源檔案, 105
 - 範例, 106
- 擴展到其它的語言和編碼, 105
- 基準, 177
 - 標記, 184
 - 選擇, 193
 - 縮放比例, 181
 - 顏色選項, 182
 - 對照到, 182
 - 標籤, 184
 - 線條, 184
- 基準高度指令, 192
- 基準執行選項, 183
- 基準標籤顏色選項, 184
- 執行 UNIX 指令, 122, 151, 183
- 執行功能表 (樹狀可視化工具), 193
- 執行執行語句
 - 分散可視化工具, 199
- 執行聚類分析, 42
- 執行語句
 - 分散可視化工具
 - 執行, 199
 - 開啓警告, 199
- 執行選項, 122, 151, 183
- 將 .ruleviz 轉換為 .scatterviz 檔案, 29
- 將空值設為 0 選項, 78
- 常數指令, 165, 166
- 推進, 37
- 排序, 34, 184
- 啓動
 - 平板可視化工具, 159, 159, 159
 - 重設預設值與, 161
 - 地圖可視化工具, 118, 118
 - 決策表導入工具, 64
 - 分散可視化工具, 118, 118, 146, 147, 147, 147, 159, 159

- 證據可視化工具, 87-88, 118
- 樹狀可視化工具, 178, 178
- 啓動工具管理員指令
 - 樹狀可視化工具, 93
- 啓動可視化規則, 23
- 啓動記錄檢視器, 137
- 啓動關聯規則, 23
- 條件機率, 85, 86
- 清除按鈕
 - 查詢對話方塊, 187, 189
- 深度滑動桿, 191
- 球體指令, 166
- 產生關聯規則, 21
- 產品組樣例檔案, 202
- 產品類別樣例檔案, 202
- 移除欄選項, 6, 62
- 第一個子節點指令, 194
- 符合查詢選項, 96, 188
- 組, 34
- 規則檔案, 23
 - 樣例, 202
- 設定地區, 105
- 設定首頁指令, 193
- 連接, 175
- 連接至伺服器, 175
- 連續顏色設定, 121, 149, 160, 182
- 速度滑動桿, 14

十二畫

- 單一 k- 平均值聚類, 41
- 單步按鈕, 13
- 尋找特定值, 186-189
 - 決策樹, 76-78
- 循環按鈕, 13
- 分散可視化工具

- 空值與, 152
- 重設, 預設, 148
- 動畫控制面板
 - 匯總視窗, 150
 - 顯示, 197
- 開始, 118, 146, 147, 159
- 資料檔案, 146
- 載入檔案, 146
- 樣例檔案, 247
- 選項, 148-152
 - 儲存, 152
- 選擇物件, 153
- 選擇欄, 55
- 檔案需求, 146
- 顏色對照, 149
- 關聯規則與, 23
- 景觀, 155, 177
- 最大 # 根選項, 133
- 最大值 / 比例高度選項, 181
- 最小支援度臨界值, 22
- 最小適合比率, 133
- 最小關鍵字
 - 顏色值與, 182
- 最後的子節點指令, 194
- 期望置信度, 22
- 減去最小證據指令, 91
- 等級, 177
 - 穿過, 193
- 等級資料
 - 樣例檔案, 202
- 等級選項, 78
- 等級檔案, 117
 - 指定, 121
 - 產生, 119-120
 - 樣例, 238, 239
- 等級欄位, 186
- 結構指令, 165, 166
- 證據分類器, 84

產生, 85
 證據可視化工具
 主視窗, 91
 功能表, 91-92, 131, 144
 配置, 87
 啟動選項, 88
 開始, 87-88, 118
 預測, 85
 樣例檔案, 227
 機率, 85, 86
 校正, 86
 選擇項目, 89
 總覽, 84
 證據窗格
 選擇項目, 89
 證據導入工具
 欄重要性工具和, 55, 56
 進階模式, 53-54
 進階模式按鈕, 53
 開始
 平板可視化工具
 重設預設值與, 161
 決策表導入工具, 64
 分散可視化工具, 146
 證據可視化工具, 87-88, 118
 開啓檔案
 分散可視化工具, 146
 樹狀可視化工具, 92, 93, 178

十三畫

僅分類器模式, 100, 103
 匯總值, 177
 匯總視窗 (平板可視化工具), 160
 匯總視窗 (「分散可視化工具」), 12, 150
 匯總圖例, 150
 匯總選項, 150, 160
 圓盤

高度, 180
 顏色選項, 182
 對照到, 182
 損失矩陣, 102, 114
 新欄名本文欄位, 16
 滑動桿
 在平板可視化工具中建立, 162
 建立, 148, 154
 對照選項, 150
 滑動桿控制, 10, 14
 滑動桿選項, 150
 準確性
 推進, 37
 準確性 (分類器的), 80
 檢驗, 100
 節點, 177
 決策樹
 檢視資訊, 72
 基準高度, 181
 尋找特定值, 186
 圓盤高度, 180
 篩選, 190
 選擇子節點, 193
 置信度, 21, 22
 期望, 22
 萬用字元
 地圖可視化工具, 96
 樹狀可視化工具, 188
 資料庫伺服器
 連接至, 175
 資料移動工具
 連接至, 175
 資料集
 分類, 71, 84, 131
 比較, 181
 追溯
 限制, 79
 採樣, 100
 尋找特定值, 76-78, 186-189

- 預測, 38
 - 混淆矩陣與, 58
- 篩選, 181, 190-191
- 選擇多個值, 124
- 儲存, 62
- 顯示資料, 116
 - 3D 景觀, 155, 177
- 動畫控制面板, 10-15
- 資料集, 分等級的
 - 樣例檔案, 202
- .data 檔案副檔名, 94, 117, 146, 158, 177
- 資料檔案
 - 平板可視化工具, 158
 - 樣例, 250
 - 地圖可視化工具, 117, 120
 - 樣例, 238, 239
 - 分散可視化工具, 146
 - 樣例, 248
 - 樹狀可視化工具, 177
 - 樣例, 253
 - 選項樹導入工具, 132
- 資料檔案面板, 62
- 資料檔案索引標籤, 62
- 資料點, 14
 - 組合, 157, 158
 - 多維的, 11
- 資料轉換面板, 62
- 路徑滑動桿, 14
- 載入檔案
 - 分散可視化工具, 146
 - 聚類可視化工具, 47
 - 樹狀可視化工具, 92, 93, 178
- 預設, 重設
 - 平板可視化工具, 161
 - 地圖可視化工具, 121
 - 分散可視化工具, 148
 - 樹狀可視化工具, 178
- 預測, 38, 85
 - 混淆矩陣與, 58

- 預測未知值, 113
- 預測離散標籤值模式, 16
- 預算, 253

十四畫

- 圖例
 - 地理區域, 122
 - 匯總, 150
 - 實體, 149
 - 關聯規則, 28
- 圖例開啓選項, 122
- 實體
 - 大小, 149
 - 空值與, 152
 - 標記, 149
 - 選擇, 151
 - 顏色選項, 149
 - 顯示, 149
- 實體圖例開啓選項, 149
- 實體標籤大小選項, 149
- 實體標籤顏色選項, 149
- 實體檔案欄位, 121
- 實體顏色選項, 149
- 對照
 - 地理位置, 120
 - 字串, 156
 - 空值與, 152, 195
 - 關聯規則與, 27
- 對照選項, 121
- 對照顏色種類選項, 182
- 端點, 116
 - 樣例檔案, 239
- 算法
 - 調節, 74, 133
- 網上發布指令, 93
- 網絡連接, 175

- 聚光燈, 190
 - 關閉, 189
- 聚類中的距離度量, 45
- 聚類中屬性權重, 45
- 聚類算法
 - 重複 k- 平均值, 42
 - 單一 k- 平均值, 41
- 與操作, 96, 188
- 誤差估計
 - 回歸樹導入工具, 142
- 誤差估計模式, 61, 100, 101
 - 檢視輸出結果, 103, 104
- 誤差估計選項
 - 平方差, 142
 - 平均絕對誤差, 142
- 誤差率
 - 增加準確度, 37
- 誤差 / 損失估計, 72, 77
- 誤差選項 (導入工具), 101-102
- 說明功能表
 - 樹狀可視化工具, 194
- 遠處水平線, 184
- 遞減選項, 133

十五畫

- 增加欄按鈕, 1
- 增加欄選項, 63
- 增益比率選項, 74, 167
- 層次選項, 77
- 播放一次按鈕, 13
- 數字
 - 查詢, 188
 - 篩選, 96
- 數量, 177
- 數學表達式, 1

- 樣例選項, 63
- 樣例檔案
 - 平板可視化工具, 166, 250-252
 - 回歸樹可視化工具, 243
 - 地圖可視化工具, 240
 - 決策表可視化工具, 213-227
 - 決策樹導入工具, 204-213
 - 信用資料庫, 202
 - 產品組, 202
 - 產品類別, 202
 - 分散可視化工具, 247
 - 証據可視化工具, 227
 - 聚類可視化工具, 202
 - 調查資料庫, 202
 - 樹狀可視化工具, 196
 - 選項樹導入工具, 240
 - 關聯規則產生器, 202
 - 欄重要性, 203
- 標記旗幟指令, 192
- 標準化子樹指令, 192
- 標準化交互資訊選項, 74, 167
- 標準化高度選項, 180
- 標準化開啓選項, 122
- 標準差, 77
- 標稱次序功能表, 71, 92, 129
- 標籤
 - 主視窗, 122
 - 平板, 161
 - 坐標軸, 151, 161
 - 兩者間的距離, 151, 161
 - 長條
 - 顏色, 184
 - 重置大小, 149
 - 基準, 184
 - 實體, 149
 - 導入工具與, 78, 100
 - 顏色選項
 - 長條, 184
 - 基準, 184

標籤機率指令, 71, 129

模式

修正, 34

記錄加權, 33

載入事先存在的, 15

選擇, 15

應用於記錄, 18-20

混淆矩陣, 17, 58-60, 101

歐洲地圖, 120, 239

編輯, 172

顏色, 50

編輯矩陣按鈕, 114

線性指令, 165, 166

線條顏色選項, 184

線間隔 (柵格), 151, 161

線顏色, 184

調查資料庫樣例檔案, 202

輪廓, 192

輪廓選項, 190

輪廓檔案欄位, 121

銷售樣例檔案, 248, 253

十六畫

學習曲線, 102, 108-110

選項, 109-110

檢視輸出結果, 103

學習曲線模式, 109

導入工具, 99

分類標籤, 100

追蹤過程, 102

執行, 100

執行模式, 100

設定選項, 101

誤差選項, 101-102

導入工具選項對話方塊, 88

操作符

相關的, 188

篩選資料, 96

樹狀可視化工具

分類器與, 102

功能表, 185

列印, 92, 93

空值與, 194

查詢物件, 186-189

穿過, 193

重設, 預設, 178

退出, 93

配置

「工具管理員」與, 185

得到資訊, 190

開始, 178

資料檔案, 177

載入檔案, 92, 93, 178

聚光燈資訊, 189, 190

樣例檔案, 196

篩選資料, 181, 190-191

選項, 178-185

重設, 184

儲存, 185

選擇子節點, 193

選擇物件, 190

空值與, 131, 195

選擇欄, 55

儲存預設值, 185

檔案需求, 177

顏色對照, 182

樹狀可視化工具的查詢對話方塊 (IRIX), 187

樹狀可視化工具選項, 178

樹狀可視化工具選項對話方塊, 179

樹狀可視化工具選擇功能表, 191

機率, 85

校正, 86

產生, 86

機率估計, 33, 39

歷史視窗, 172

澳大利亞地圖, 120, 238

篩選

- .data, 181, 190-191
- 地圖, 95
- 決策樹, 76
- 篩選出 % 最短的選項, 181
- 篩選按鈕, 95
- 篩選面板
 - 地圖可視化工具, 95-96
 - 樹狀可視化工具, 190-191
- 篩選面板指令, 76, 95, 190
- 篩選選項, 63
- 選項節點, 133
 - 已定義, 131
- 選項樹分類器, 131
 - 產生, 131
- 選項樹導入工具
 - 必需的檔案, 132
 - 樣例檔案, 240
 - 調整歸納算法, 133
 - 總覽, 131
- 選擇功能表
 - 証據可視化工具, 92
 - 樹狀可視化工具, 191
- 選擇多個值, 124
- 選擇物件
 - 空值與, 131, 195
 - 分散可視化工具, 153
 - 樹狀可視化工具, 190
- 選擇基準, 193
- 選擇實體, 151
- 選擇模式
 - 証據可視化工具, 89
- 選擇模型, 15
- 選擇顏色, 50, 52
- 隨機種子, 101
- 隨機樣例, 100
- 靜態選項, 88, 118, 199

十七畫

- 儲存工具選項
 - 地圖可視化工具, 123
 - 分散可視化工具, 152
 - 樹狀可視化工具, 185
- 儲存資料, 62
 - 記錄檢視器, 137
- 應用分類器選項, 63
- 應用按鈕, 96
- 應用模式面板, 16
- 應用模型
 - 估計機率值模式, 16
 - 預測離散標籤值模式, 16
- 應用模型按鈕, 15
- 檔名
 - 平板可視化工具, 158, 159
 - 分散可視化工具, 146
 - 樹狀可視化工具, 94, 117, 177
- 檔案功能表
 - 網上發布, 93
- 檔案需求
 - 平板可視化工具, 158
 - 地圖可視化工具, 117
 - 分散可視化工具, 146
 - 樹狀可視化工具, 177
 - 選項樹導入工具, 132
- 檢查表達式按鈕, 4
- 檢視
 - 平板可視化工具
 - 顯示選項, 161
 - 目前, 171
 - 地圖可視化工具, 115, 116
 - 顯示選項, 121-123
 - 穿過, 171
 - 分散可視化工具
 - 顯示選項, 148-152
 - 樹狀可視化工具
 - 穿過, 193

- 聚光燈資訊, 189, 190
- 顯示選項, 178-185
- 檢視功能表, 185
 - 決策表可視化工具, 70
 - 分散可視化工具, 197
 - 証據可視化工具, 91
- 檢視全部指令, 193
- 檢視歷史按鈕, 172
- 檢驗分類器準確性, 100
- 檢驗值選項, 77
- 檢驗設定誤差 / 損失選項, 77
- 檢驗模式面板, 17
- 檢驗模型
 - 上升曲線, 17
 - 混淆矩陣, 17
- 檢驗屬性選項, 77
- 縮放比例
 - 地理區域, 121
 - 長條, 181
 - 基準, 181
 - 實體, 149
- 薪資因素資料集, 208, 218, 231
- 隱藏
 - 標籤, 151, 161
- 隱藏標籤, 151, 161
- 隱藏標籤距離選項, 151, 161
- 隱藏選項, 190

十八畫

- 歸納算法
 - 調節, 74, 133
- 簡易 Bayes, 125
- 簡易模式, 53
- 離散化算法, 55
- 離散標籤, 78
- 離散標籤功能表, 78

- 離散顏色設定, 121, 149, 160, 182
- 離散屬性, 55, 78
- 顏色, 48, 182
 - 天空, 184
 - 平板, 160
 - 地面, 184
 - 改變, 50
 - 長條, 182
 - 根據按鍵, 183
 - 標籤, 184
 - 柵格, 151, 161
 - 根據關鍵字填入, 183
 - 基準, 182
 - 標籤, 184
 - 圓盤, 182
 - 實體, 149
 - 標籤
 - 長條, 184
 - 基準, 184
 - 線條, 184
- 顏色組合選項, 182
- 顏色清單, 121, 149, 160, 182
- 顏色開關, 48, 50
- 顏色對照
 - 平板可視化工具, 160
 - 地圖可視化工具, 121-122
 - 分散可視化工具, 149
 - 樹狀可視化工具, 182-184
 - 空值與, 195

- 顏色對照選項, 149, 160, 182
- 顏色編輯器, 121, 182
- 顏色選項, 182
- 顏色瀏覽器, 51
 - 開啓, 50

十九畫

- 轎車產地資料集, 205, 216, 229, 241

關係，分析，145, 155, 177

關閉按鈕

查詢對話方塊，189

關閉指令，93

關聯規則，20-29

上升，22

支援，22

最小臨界值，22

市場供求，21

多重，25

顯示，26

記錄加權，25

追溯，29

配置

「工具管理員」與，24-28

產生，21

分散可視化工具與，23

期望置信度，22

開始，23-24

置信度，21, 22

期望，22

對照資料於，27-28

樣例檔案，202

顯示，28

關聯規則產生器，20-24

組件，21, 23

輸出，21

檔案需求，23

總覽，20

顯示圖例，28

關聯規則對照面板，27

關鍵字

為列著色與，183

二十畫

蘑菇分類資料集，210, 222, 233, 242, 250

用於 ... 的混淆矩陣，112, 113, 114

警告，199

黨派傾向資料集，210, 223, 234, 242

二十一畫

屬性，38, 78, 100

可用性，80

檢驗，77

離散化算法，55

欄

已計算，1

組合選項，5, 6

命名，2

選擇，52, 53, 55

檢視，62

欄位名，2

欄重要性工具

重要性分級，56

純度測量，56

從屬性，56

樣例檔案，203

模式，53-54

離散屬性與，55

欄重要性算法，55

二十二畫

權重是屬性選項，102

二十三畫

變數

作為篩選工具，95

顯示

.data, 116, 155, 177

動畫控制面板，10-15

分類器輸出，102

平板，160

- 決策樹, 102
- 決策樹節點, 72
- 訊息, 124
 - 地圖可視化工具, 122
 - 分散可視化工具, 151
 - 樹狀可視化工具, 183
- 動畫控制面板, 197
- 實體, 149
- 標籤, 151, 161
- 選擇物件, 124
- 顯示上升曲線選項, 111
- 顯示功能表 (樹狀可視化工具), 185, 192
- 顯示值指令, 191
- 顯示原始資料指令, 153, 191, 226
 - 用法討論, 79
- 顯示動畫面板指令, 197
- 顯示參數, 192
- 顯示視窗裝飾指令, 197
- 顯示資料, 12
- 顯示混淆矩陣選項, 101
- 顯示選項
 - 平板可視化工具, 161
 - 地圖可視化工具, 121-123
 - 分散可視化工具, 148-152
 - 樹狀可視化工具, 178-185
- 顯示關聯規則, 28
- 「下一個」按鈕, 189
- 「下一個」欄位, 171
- 「上一個」按鈕, 189
- 「上一個」欄位, 171
- 「另存目前歷史為 ...」指令, 185
- 「另存新檔」指令, 93
- 「其他選項」選項, 151, 161
- 「或」運算, 96, 188
- 「查詢中忽略大小寫」選項, 187, 188
- 「查詢」按鈕, 189
- 「查詢」指令, 186
- 「相等」查詢選項, 96, 188
- 「設定全部」按鈕
 - 查詢對話方塊, 187
- 「設定每組最小權重」選項, 86
- 「尋找檔案」按鈕, 121
- 「開啓」指令, 92, 93
- 「傳送到工具管理員」指令, 153, 191, 226
 - 用法討論, 79
- 「實體大小」選項, 149
- 「實體形狀」選項, 149
- 「實體選項」選項, 149
- 「增加欄」對話方塊, 4
- 「編輯前一個操作」按鈕, 172
- 「篩選中忽略大小寫」選項, 190
- 「選擇」按鈕, 189
- 「顏色選擇」對話方塊, 151, 161