MineSet™ 3.0 企業版
使用者指南®

文件編號 007-4005-001CHT

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。

MineSet™3.0 企業版使用者指南®
文件編號 007-4005-001CHT

目錄

圖表清單

表格清單

關於本指南

*MineSet 3.0 企業版使用者指南*說明 MineSet 挖掘和可視化處理工具的特性和功能。關於 MineSet 產品的最新資訊，可以在 World Wide Web 上找到，網址爲 http://mineset.sgi.com。

本指南的適用物件

您不必是資料挖掘方面的專家就可以使用本指南或 MineSet，但是對資料及其代表含義的了解會有助於您更加容易解釋結果。如果您有資料挖掘技術的經驗，本指南仍然可以幫助您了解 MineSet 運算法則的工作方式，以及它們是如何應用和可視化的。

如果要使用「工具管理員」以便從資料庫中將資料匯入 MineSet 工具，您應該參考*《MineSet 3.0 企業版介面指南》*以執行相關任務。一旦資料載入各種可視化工具中，就不需要資料庫或程式背景了。

尋找 MineSet 資訊

本指南解決有關執行 MineSet 的任務。書中的大部分內容都是在講述如何執行各種工具。在「本文件的結構」中有逐章的摘要。

關於技術細節和更完整的說明，請參閱*《MineSet 3.0 企業版參考指南》*。關於指定的 MineSet 運算法則方面更專業的資訊，請參閱 http:www.sgi.com/software/mineset/mineset_data.html 中的白皮書清單。

關於下列的資訊，如：將 MineSet 匯出至其他位置或應用程式、用指令行操作或有關於系統管理，請參閱*《MineSet 3.0 企業版介面指南》*。

MineSet 也爲協力廠商的產品提供插入應用程式的方法，如 AcPro。

本文件的結構

本指南的前三章將介紹資料挖掘和 MineSet 產品。隨後的章節將集中介紹以下的特定工具和過程：

第 1 章，「資料挖掘和 MineSet 工具總覽」
本章簡短概述資料挖掘的原理、解釋術語，並介紹利用 MineSet 工具程式組進行分析和可視資料挖掘的過程。

第 2 章，「利用 MineSet 存取資料」
本章描述如何執行 MineSet 和如何使用一些基本的工具尋找資料。

第 3 章，「加工資料」
本章描述使用「工具管理員」轉換原始資料的需求和過程。說明研究資料的未知特徵的「欄重要性」和「聚類」工具。

第 4 章，「用分散和平板可視化工具檢查資料」
本章介紹「分散和平板可視化工具」的介面。無論是靜態的或是透過動畫演示，這些工具對於可視化多維度資料都很有價值。

第 5 章，「用樹狀可視化工具檢視資料」
本章描述「樹狀可視化工具」的介面。該工具對於可視化分層資料很有價值。

第 6 章，「用地圖可視化工具檢查資料」
本章描述「地圖可視化工具」的介面。該工具對於具有地理或空間背景的資料很有用。

第 7 章，「理解預測建模」
本章介紹和描述建模相反的預測建模，以及 MineSet 工具提供的各種分類器。

第 8 章，「用決策、選項和回歸樹建模和預測」
本章描述如何產生並使用「決策」、「選項」和「回歸」工具。這些工具對於按照屬性集合並根據這些屬性進行一系列的決策來分類資料很有價值。選項樹可以同時顯示拆分對多個屬性產生的影響。回歸樹對於預測根據連續值的屬性 (例如真實生活中的事件) 很有用。

第 9 章，「用決策表分類器和可視化工具建模和預測」
本章完整地描述「決策表」的介面。該工具對於在分類資料中作出的可視化決策和建立分類器很有價值。

第 10 章，「用証據分類器和可視化工具建模和預測」
本章描述如何產生並使用「証據分類器」。該工具對於檢查指定結果在指定屬性下出現的機率來分類資料很有價值。

第 11 章，「改進預測建模」
本章描述「混淆矩陣」、「損失矩陣」、「ROI 曲線」和「上升曲線」的用途及用法，將預測建模與真實生活聯繫起來。

第 12 章，「用聚類劃分資料」
本章描述「聚類可視化工具」的介面。該工具對於檢查資料和確定聚類模式很有價值。

第 13 章，「用關聯規則分析資料」
本章描述「關聯規則可視化工具」。該工具對於挖掘大的資料集和發現其中資料的相互關係很有價值。

MineSet 用戶詞彙表
詞彙表解釋本指南中頻繁使用的專門詞彙和詞組。

本指南中的插圖

本指南中的硬拷貝顯示螢幕畫面的黑白插圖。線上版本中的這些插圖都是彩色的。如果發現在印出的複本中看不清某個圖形或螢幕畫面，可以參考線上版本的頁面，以獲得更加清晰的插圖。

相關讀物

關於資料挖掘及相關技術的一般資訊，可以參閱以下讀物：

- *Data Mining Techniques for Marketing, Sales, and Customer Support*, Michael Berry 和 Gordon Linoff 合著，John Wiley & Sons 出版。
- *Data Mining Solution ： Methods and Tools for Solving Real World Problems*, Christopher Westphal 和 Teresa Braxton 合著，John Wiley & Sons 出版。

如果希望了解關於資料挖掘更加專業的資訊，*《MineSet 3.0 企業版介面指南》*中的附錄 A 列出了更多的讀物。

印刷慣例

本指南使用以下的類型慣例和符號：

斜體　　可執行名稱、檔名、程式變數、工具、公用程式、變數指令行參數、以及範例、代碼和語法語句中由使用者提供的變數

粗體　　關鍵字

固定寬度類型
　　螢幕上的指令行本文和提示

粗體固定寬度類型
　　使用者的輸入，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

資料挖掘和 MineSet 工具總覽

本章介紹資料挖掘的有關主題以及 MineSet 工具如何在此種環境下工作。如果對 MineSet 已經有所了解，可以跳過本章。有關主題包括：

MineSet 是一個進行資料分析的集合工具程式組。市場和營銷專家、金融分析家、保險商以及任何需要分析資料的人都可以使用 MineSet 的圖形使用者介面。對於希望爲商業和科學目的編寫決策支持應用程式的程式開發員，MineSet 也提供了介面。

資料挖掘技術

在這裡和本書最後的詞彙表中總覽了基本資料挖掘術語的用法。例如：在此種情況下，資料是進行一個商業或科學過程 (例如：客戶帳單、藥物測試或銷售點交易) 時最初收集的記錄集合。這些記錄組織爲帶有列和欄的表格，以便 MineSet 使用。可以賦予表格一個描述性名稱，如「用戶」，代表某個特定企業所有關於用戶的記錄。在資料挖掘環境中，通常將欄稱爲屬性，將列稱爲實例或記錄。

在 MineSet 使用的運算法則或數學公式中，有些稱爲導入工具，因爲它們可以導入要建立的模型。模型本身稱爲分類器，因爲它們對資料的列和欄進行分類。因此，您可以獲得欄和列的記錄內容，並以多種方式對它們進行排序或分類。

關於資料挖掘和資料挖掘方法

資料挖掘的目的是爲了發現資料中的規律，並將此種認識應用到問題解決中。資料挖掘可以解決的典型問題包括：

- 詐欺偵測
- 客戶波動分析
- 電話模式分析
- 市場目標
- 確定市場劃分
- 改善操作程序
- 改善醫療服務
- 市場供需分析

當使用查詢或線上分析處理 (OLAP) 收集資料庫中的資料資訊時，必須直接指定資料元件之間的所有關係。例如：要按區域查詢所有銷售額。此種方法預先假設您認爲銷售額隨區域變化，因此要用資料檢驗這個假設以証明或否認它的有效性。

這就是 OLAP 與基於發現的資料挖掘的結合點，在資料挖掘中可以發現未知的關係。此種方法允許資料本身向調查者提出結論。就是這種發現原來未知關係的能力將資料挖掘與 OLAP 及其他方法 (如查詢) 區分開來。

可視化資料挖掘以可視的形式展現資料，因此只要經由可視化處理就可以了解總體趨勢。此種可視化和研究複雜資料模式的能力在決策中是極其寶貴的。此種可視資料挖掘是描述性的，即：它對已測量或量化的現有資料進行描述。

分析資料挖掘使用運算法則自動從資料中建立模型。然後這些運算法則將根據選擇的算法以各種方法對資料進行處理。可以將資料分析結果以可視形式顯示，用未經試驗的資料檢驗它們，或是將模型適用到全新的資料集上。分析資料挖掘使用模型預測下一部分存在資料的特徵。

分析資料挖掘可以產生關於資料的模型。它用於工作的運算法則可分爲兩大類—監督和非監督建模方法。

在 MineSet 中，可視化處理與挖掘模型的結合與單獨應用任何一種技術相比，更能夠爲使用者提供深入的認識。

分析資料挖掘

在該論述中，資料出現在由欄和列組成的表格中。資料紀錄由列代表：資料屬性由表格的欄代表 (請參閱圖 1-1)。

	描述性屬性				標籤
	年齡	**婚姻狀況**	**性別**	**受僱年限**	**信用風險**
記錄 1	35	D	M	3	> 60
記錄 2	45	M	F	19	< 60
記錄 3	29	S	M	45	> 60

圖 1-1 資料表樣例

監督建模

監督建模任務的目標是根據記錄中其它欄 (屬性) 的值，預測同一個記錄中某一欄 (屬性) 的內容或值。要預測一個特殊屬性或特徵。該屬性稱爲*標籤*。經由對標籤和其他屬性之間關係的編碼，可以使用監督建模運算法則建立模型，對新的、未標示的資料進行預測。

然後，您就能夠使模型可視化，獲得對標籤和其他屬性之間關係的了解。例如：如果用戶已經離開公司 (通常被稱爲縮減或客戶波動)，就可以建立一個模型，既可以預測哪些客戶可能客戶波動，又有助於了解導致這種行爲的原因和模式。

兩種最普通的監督建模任務是*分類*和 *回歸*。如果標籤是離散的 (即：包含一個固定集合的值—例如：「是」或「否」)，則該任務稱爲分類；如果標籤是連續值 (即：可以取一個連續範圍中的任何值—例如：收入和股價)，則該任務稱爲回歸。

分類

分類任務是將分類標籤值分配給未標註的記錄。在此種情況下，記錄將劃分到預先定義中的群組中。例如：一個簡單的分組可能把客戶的帳單記錄歸爲以下特定的兩類：在 60 天內支付帳單的人，和超過 60 天支付帳單的人。另外的資料分類範例可能會按性別或收入來對客戶進行分組。分類過程也可預測標籤採取指定值的機率。例如：可以計算用戶在 60 天內支付帳單的機率。

在建立這些分類過程時，MineSet 利用運算法則，根據所提供的資料導入 (建立) 模型。MineSet 可以自動從訓練集中導入分類器。訓練集是原始資料的隨機子集。例如：可能決定只使用 40% 的資料來建立模型。模型一旦產生，就可用來對標籤值未知的記錄進行分類或預測分類機率。再如：當增加新記錄到表中時，標籤值可能是未知的。這時，模型就可以預測標籤欄的值。如果後來發現標籤值與預測的相同，模型即爲正確的。原始資料集中沒有用於訓練的部分可用來檢驗導入模型的準確性。

MineSet 可以導入四種分類模型，每一種都可以用可視化工具查看。根據運算法則在排序和分類資料時作出決策的方式，這些模型是不同的。例如：決策樹可能根據用戶交談是否超過 256 分鐘而將電話記錄分爲兩個組。而選項樹可能提供三種方式以搜尋可能波動的用戶。關於這些區別的詳細內容，請參閱第 7 頁的「MineSet 的資料挖掘工具」，有關範例將在稍後的章節中進行介紹。

回歸

除了標籤取值是連續而不是離散的以外，回歸是類似於分類的一種監督建模任務。例如：預測薪水或股價是一種回歸，儘管預測薪水是否在指定的範圍內 (例如：少於 $60K，$60K 到 $120K 之間，多於 $120K)，或股價將會上升還是下降是一種分類任務。

MineSet 還可以從訓練集或資料樣本中自動導入回歸器。回歸器一旦產生，就可用來預測問題中連續屬性的值，例如：薪水爲 $60,420。MineSet 目前只有一個回歸模型：回歸樹。和 MineSet 建立的所有其他模型一樣，回歸樹模型也可用 MineSet 可視化處理進行可視化。

評測模型的準確性

估計監督模型的準確性是資料挖掘過程中的重要部分。通常按照分類器的錯誤率或分類錯誤紀錄的比例對其進行評測。而回歸器的評估則是根據其預測誤差中標準偏差的絕對誤差而進行。在評測模型的準確性時，非建立模型所用的資料對其進行測試是非常重要的。

非監督建模

非監督建模發現行爲相似的資料段及其規則。既沒有正確答案的任何理念，也沒有明確公認的性能度量。非監督建模顯示相似的模式和區段，從整體的角度提供對資料的認識。模型無法直接用來進行預測；因此無需將部分資料保留下來作爲建立分類器的訓練集。MineSet 爲兩種最普通的非監督建模提供運算法則：關聯和聚類。

關聯

在產生關聯時，主要任務是描述 A 是否憑藉某種規則隱含著 B 。經典的關聯分組是市場供求分析，用以預測特定物品被同時購買的頻率。例如：購買嬰兒商品暗示著該客戶購買低焦油含量香煙的機率要比購買普通香煙的機率高，這個發現可能有助商店有差別地安排貨架。

聚類

聚類運算法則將資料分割爲具有相似特徵的記錄組或聚類組。例如：健康保險公司可能會發現下列特徵可以定義一個區段： 20 到 45 歲、技術員、孩子少於兩個、科幻電視迷、並且可支配收入爲每年 5,000 到 10,000 美元。然後，經由在新的科幻電視中插播電視廣告，就可以利用對這些人非常適合的健康保險節目更加有效地針對他們。

可視化資料挖掘

資料挖掘運算法則可由資料可視化技術進行補充，這些技術利用了人類大腦驚人的模式識別能力。以下的 MineSet 可視化工具爲工作範例：

- 地圖可視化工具—資料顯示在一張圖上，通常是一張地圖。
- 分散可視化工具—資料點的顯示是一維、二維或三維。額外的屬性可以對照爲顏色和大小。還有兩個附加的屬性可以對照到滑動桿，允許動畫和閃現，第八個屬性可以對照爲顏色代碼，透過可能有趣的動畫變數值組合來引導動畫。MineSet 中的欄重要性操作可以幫助識別指定任務要對照的重要屬性。
- 平板可視化工具—類似於分散可視化工具，其區別在於資料密度使用變化的不透明性顯示。得到的結果與單獨處理每個資料點所得的結果相似，並對於資料點太多以至於無法在分散可視化工具中顯示的資料集尤其有用。
- 樹狀可視化工具—資料對照到節點上以便觀察資料的分層分解。決策樹、選項樹和回歸樹都將資料顯示在各種分支的樹狀可視化處理中。

MineSet 的資料挖掘工具

如果有需要分類、回歸和聚類的資料集，可以使用下列這些 MineSet 工具：

- 決策樹導入和可視化工具—導入一個會產生分支格式的決策樹可視化處理之分類器。
- 選項樹導入和可視化工具—導入一個類似於決策樹導入和分類工具的分類器。但是，它還建立另外的樹並在分類過程中均衡它們，通常會提高準確性。
- 証據導入和可視化工具—建立自己的分類器並進行可視化處理，以根據所提供的資料顯示証據。
- 決策表導入和可視化工具—建立一個分層的可視化處理，顯示每一等級上的維度組。在保留上下文的同時，可以迅速地概化上尋獲得總覽，細化下尋查看細節。
- 回歸樹導入和可視化工具—導入一個預測真實值屬性的回歸器，即：結果帶有不同等級的值，而不是帶有預先確定的限制。可以使用樹狀可視化工具查看回歸器的結構。
- 關聯規則—對控制關聯 (其中 A 暗示 B) 的規則進行編碼，通常稱爲市場供求分析。使用分散可視化工具顯示規則，通常用長條和圓盤形。
- 聚類運算法則—根據特徵的相似性對資料進行分組，然後將其顯示爲一系列的方塊圖和柱狀圖，類似於統計可視化工具。預設情況下，聚類運算法則使用聚類可視化工具顯示結果，但是也可以使用其他的可視工具。
- 欄重要性分析—在區分兩個標籤值時，確定指定欄的重要性。用於觀察變數改變的各種影響，或提出對照到分散和平板可視化工具坐標軸的欄。

MineSet 中還包含額外的工具來輔助知識的發現過程：

- 統計可視化工具—以方塊圖和柱狀圖形式顯示資料，每一欄一個圖。連續欄顯示爲方塊圖；離散欄顯示爲柱狀圖。

- 柱狀圖可視化工具—以柱狀圖形式顯示資料，每欄資料一個柱狀圖，如果必要，可對連續欄進行分組。
- 記錄檢視器—將原始資料顯示爲試算表。

資料挖掘過程總覽

本部分介紹有關知識發現過程的特定任務。該過程是一個反複重複的過程，一旦發現新的模式並提高對資料的認識後，通常就再回到早期的階段，如 圖 1-2 所示。

圖 1-2 資料挖掘過程

該過程包括以下步驟：

1. 識別資料來源—請參閱第 9 頁的「識別資料」。
2. 準備資料—請參閱第 9 頁的「準備資料」。
3. 建立模型—請參閱第 10 頁的「建立模型」。
4. 評估模型—請參閱第 11 頁的「評估模型」。
5. 配置模型—請參閱第 11 頁的「配置模型」。

識別資料

識別資料任務將從決定需要哪些資料來解決問題開始。例如：關於客戶行爲的可預測性通常是一個必要目標。在重新定義問題時，調查者必須識別用於解決該問題的資料並探索其他可能的資料來源。

資料可能位於一個較爲偏僻的位置，或是以一種模糊的形式出現。有時存在多個互不相容的初始資料庫。而且，如果資料很少或不完整，就會需要更多的資料。新資料的形式將根據現有資料的形式收集起來。MineSet 支援多種商業資料庫 (Oracle、Informix、SQL) 的本地介面、ODBC 介面，並可讀取不同檔案格式的資料 (跳位符分隔的單一結構檔案、MineSet 二進位檔案、Excel、SPSS、MATLAB 等等。) 一些工具 (如：Conceptual Software, Inc. 的 DBMS/COPY) 可以將資料從 100 多種格式轉換爲 MineSet 可以使用的格式，並且 MineSet 的導入資料能力 (在「工具管理員檔案」功能表中) 也可進行類似的轉換 (關於詳細內容，請參閱*《MineSet 3.0 企業版參考指南》*和*《MineSet 3.0 企業版介面指南》*)。

準備資料

在載入至 MineSet 前，資料可能需要修改 (該步驟叫做清洗。) 特別地，常見的問題包括：

- 資料格式可能與 MineSet 的表示法不相容 (例如：來自舊式大型電腦的二進位、編碼、或 EBCDIC 字串)。
- 資料可能有拼寫錯誤或是錯誤，可能不完整，也可能取值有誤。
- 欄位說明可能不清楚，或者可能根據不同來源而有不同的含義。例如：訂購日期所指的可能是訂單已傳送、蓋戳、收到或輸入的日期。
- 資料可能過期；例如：客戶可能搬家、成員變更或改變消費模式。

即使是清楚易懂的資料也需要在應用於挖掘和可視化處理之前進行轉換。

轉換資料

轉換可以極大地提高模型的性能。例如：如果您正在分析電話公司的資料，可能會發現：相對單獨指定的任何一個元件，長途電話的費率 (按通話使用的分鐘總數計費的銷售額) 是一個更好的用戶行爲預測器。資料轉換是開發合理模型的核心部分。隨著項目的進展，您甚至可能會返回並以另外的方式轉換資料。可以透過下列方法來轉換資料：

- 增加欄，通常是對已有資料應用一個數學公式來建立新欄位。
- 移除無關、多餘、或包含明顯無用預測器的欄。
- 根據欄值的布林表達式篩選資料，以影響模型或可視化處理過程。例如：您可能只希望看到最重要的規則或最有利可圖的客戶區段。
- 對資料進行分組—將一個連續範圍內的資料分解爲幾個離散的區段 (例如： [1 - 10]，[11 - 20]，等等)。
- 組合資料—將記錄集合在一起，然後尋找總和、最大值、最小值和平均值。
- 對資料進行採樣以獲得資料的隨機子集 (透過百分比或計數)。
- 應用一個先前建立的分類器、回歸器或聚類模型，對帶有分類標籤的新記錄進行標示，或估計指定標籤值的機率。

在 MineSet 中，大部分的轉換是使用「工具管理員」中的「資料轉換」窗格來完成。插件模塊可用於建立轉換或特定種類的模型。

建立模型

知識發現過程的核心是模型的建立，該任務由分析型資料挖掘運算法則自動完成。可以使用所有的資料建立模型，也可以保留部分資料用來測試模型的準確性。建立模型可以提供多個導入模型的選項。這些選擇不僅會影響可視化處理顯示的方式，而且還會影響運算法則在建立分類器時的決策。在第 7 章「理解預測建模」中將進一步討論這些選項。

評估模型

評測模型的準確性可以進一步提煉對該模型及其用法的認識。一些模型，特別是「決策樹」分類器和「選項樹」分類器，可以直接通過可視化處理評估模型的不同部分並顯示它們。

MineSet 提供四種模型評測方法：誤差估計、混淆矩陣、上升曲線和 ROI (投資回報) 曲線。在第 11 章「改進預測建模」中將對這些方法進行說明。

配置模型

可以透過將模型應用到新資料上來配置它。新資料可能會產生新的問題，可能需要進一步的改進。

在《*MineSet 3.0 企業版教學課程*》的電信範例中，建立一個模型來確定哪些用戶可能波動 (即：離開他們的電話公司)。然後，客戶記錄將經由模型進行評估，以識別最可能波動的特定客戶。可以給這些客戶一定的刺激使他們留下。

可以繼續了解關於客戶波動資料集— (MineSet 軟體中準備好的電信用戶資料集) —的知識發現過程。當您完成該指南中的範例時，請思考一下您的商業或科學活動是如何向前發展與反複，進而建立起一個經過充分測試的分析資料挖掘模型。

在下一章，第 2 章「利用 MineSet 存取資料」您就可以利用 MineSet 及提供的樣本檔案直接開始工作。

利用 MineSet 存取資料

本章描述如何開始執行 MineSet，以及如何在「工具管理員」圖形使用者介面中使用程序查看資料。討論的主題包括：

另見*《MineSet 3.0 企業版教學課程》*以練習使用各種工具。當希望使用自己資料庫中的資料時，請參閱*《MineSet 3.0 企業版介面指南》*中的「設定 MineSet」。

在何處可以獲得 MineSet

可以從 CD 中或從 Internet 上安裝 MineSet。一旦安裝了 MineSet，(通常在您自己的系統上) 就可以立即開始使用它，並利用 MineSet 發行版提供的預先格式化資料檔案 (位於 MineSet 安裝目錄中的 *data* 目錄下) 進行練習。下一個部分將對 MineSet 操作進行簡短說明。如果希望立即開始，請參閱第 15 頁 的「執行 MineSet」。

MineSet 軟體是如何運作的

MineSet 以客戶端 / 伺服器的方式進行工作，伺服器過程可與客戶端執行在同一個系統中，也可以在另一個通常是更強大的系統中。伺服器必須安裝在客戶端可以找到的地方。「資料移動器」是一個在伺服器上執行的程序，用於存取資料庫檔案、執行資料轉換、執行挖掘操作並產生可視化處理檔案。這些可視化處理檔案隨後將傳送到客戶端。

「工具管理員」是一個客戶端過程，可以經由它提供的圖形使用者介面 (GUI) 完成與 MineSet 的大部分交互任務。這些交互任務是本章的主要部分。可以使用「工具管理員」指定資料來源、應用於資料的轉換集、使用的挖掘和可視化處理工具、以及如何儲存工作的結果。

一旦完成這些規範，就可指揮 MineSet 存取「工具管理員」中指定的所需資料，並執行其中指定的操作。然後，「工具管理員」將把該資訊傳送到伺服器「資料移動器」進程。如果要分析的資料存在於單一的結構檔案中，MineSet 將讀取一個配置 (*.schema*) 檔案，其中包含表格的交互資訊以及真正資料所在的 *.data* 檔案名稱。如果要分析的資料在 Oracle、Sybase 或 Informix 資料庫中，「資料移動器」將對資料庫執行查詢以取得需要的資料集。

.schema 檔案描述輸入資料檔案的格式。在 *《MineSet 3.0 企業版介面指南》* 中有這些檔案的詳細說明。

注意 : 在使用 MineSet 工具之前，請先按照 MineSet 發行版注意事項中的安裝和授權指示進行。系統管理員可能需要設定「資料移動器」配置檔案。*《MineSet 3.0 企業版介面指南》* 中介紹詳細的設定資訊。

MineSet 與應用程式

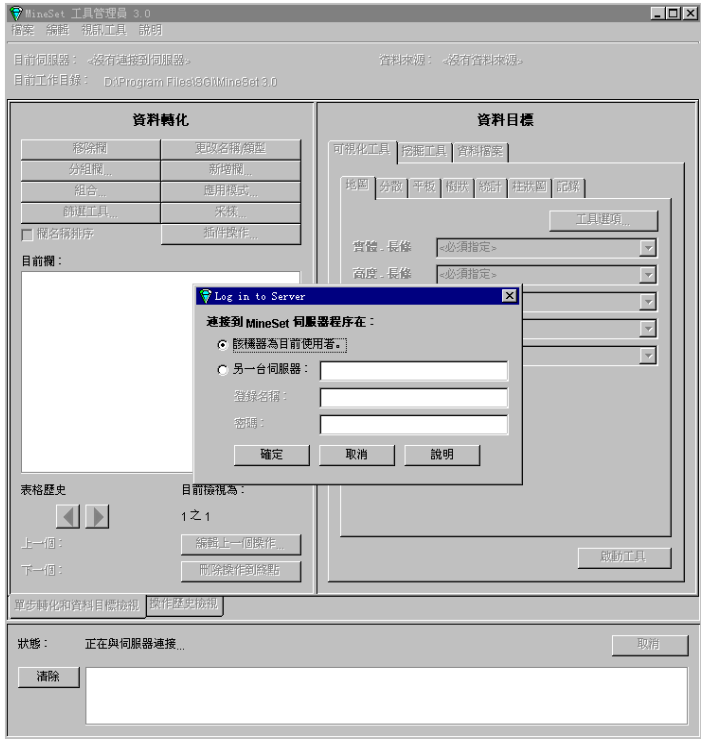
MineSet 具有一個應用程式介面 (API)，因此可將 MineSet 的可視化工具和資料挖掘引擎融合到應用程式中。關於詳細資訊，請參閱 *《MineSet 3.0 企業版介面指南》*，或 MineSet 網站。

執行 MineSet

執行 MineSet 的最簡單方法是在桌面上爲 *mineset.exe* 建立捷徑。要建立捷徑，可以從 MineSet 的安裝目錄中，進入 *\bin* 目錄，並按一下 *mineset.exe* 檔案。從「檔案」功能表中選擇 「建立捷徑」。

也可以從桌面上的「開始」>「程式集」>「MineSet 3.0 企業版」功能表進入 MineSet。可以用滑鼠左鍵選擇來執行它，或按住滑鼠右鍵將其拖曳到桌面上，然後再放開滑鼠右鍵。

1. 按一下 MineSet 捷徑圖示，將出現「MineSet 工具管理員」。
2. 如果預設情況下，「登錄到伺服器」對話方塊 (圖 2-1) 沒有出現，可以選擇「檔案」>「連接到伺服器」。如果希望將目前系統同時作爲客戶端和伺服器來使用，可以在產生的對話方塊中，按一下「將這台機器作爲目前使用者」。如果希望使用另一個系統作爲伺服器，請鍵入伺服器名稱、登錄名稱和密碼 (如果有的話)。



2-1 工具管理員登錄

3. 在「工具管理員」視窗中，選擇「檔案」>「開啓新的資料檔案」。在出現的「開啓資料檔案」視窗中，輸入要使用的資料路徑。如果希望用 MineSet 附帶的樣例文件進行練習，可以在 *MineSet 3.0\data* 目錄下找到它們，它位於 MineSet 的初始安裝位置。

   要使用樣本檔案工作，可在「開啓資料檔案」視窗中選擇 *churn.schema* 。該檔案是 MineSet 提供的一個資料樣表。右邊「預覽欄」窗格中的項目是資料欄，如圖 2-2 所示。按一下*確定*。

churn.schema 檔案讓您存取電信用戶的資料集。下一次執行 MineSet 時，系統將自動把您帶到最後一次結束 MineSet 時所處的狀態，並恢復您所做的一切選項選擇。

2-2 「開啓資料檔案」視窗

*《MineSet 3.0 企業版參考指南》*中包括一個可用來練習的樣例檔案清單；關於每個檔案的內容，請參閱「配置和資料檔案樣例」。

使用「工具管理員」視窗

「工具管理員」是執行所有操作的主要起點，因此對該部分進行快速瀏覽是很有幫助的。

2-3 「工具管理員」視窗

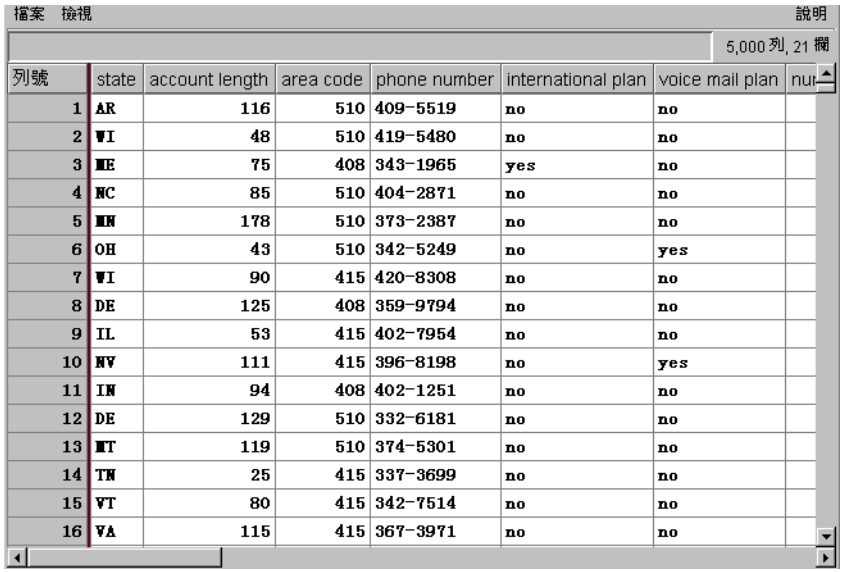
左邊的「資料轉換」窗格是轉換所有資料的地方。使用頂部按鈕進行的操作會影響以下面的「目前欄」。可以使用下部的按鈕在轉換的歷史操作中中前後瀏覽；底部的「狀態」視窗顯示目前資料集進行的操作。可以使用右邊的「資料目標」窗格在「可視化工具」和「挖掘工具」之間切換，或將目前狀態儲存在「資料檔案」中。以下將詳細敘述其中的每個功能。

用「記錄檢視器」查看原始記錄

有一個對選定的資料集熟悉的簡單方法，就是使用「記錄檢視器」查看資料庫記錄和欄中的資料值。記錄將以試算表的形式出現。

開啓「工具管理員」，如「執行 MineSet」所示，並按照以下步驟：

1. 在「工具管理員資料目標」窗格中，按一下「可視化工具」索引標籤。
2. 在下面一列的索引標籤中，按一下「記錄」以進入「記錄檢視器」。
3. 按一下窗格右下角的*調用工具*，將出現「記錄檢視器」(圖 2-4)。



列號	state	account length	area code	phone number	international plan	voice mail plan	nu
1	AR	116	510	409-5519	no	no	
2	WI	48	510	419-5480	no	no	
3	ME	75	408	343-1965	yes	no	
4	NC	85	510	404-2871	no	no	
5	MN	178	510	373-2387	no	no	
6	OH	43	510	342-5249	no	yes	
7	WI	90	415	420-8308	no	no	
8	DE	125	408	359-9794	no	no	
9	IL	53	415	402-7954	no	no	
10	NV	111	415	396-8198	no	yes	
11	IN	94	408	402-1251	no	no	
12	DE	129	510	332-6181	no	no	
13	MT	119	510	374-5301	no	no	
14	TN	25	415	337-3699	no	no	
15	VT	80	415	342-7514	no	no	
16	VA	115	415	367-3971	no	no	

圖 2-4 「記錄檢視器」畫面

如果希望以不同的方式查看結果，「記錄檢視器」可以進行各種排序、重新編號及調整大小 (請參閱下面的表 2-1。)

改變「記錄檢視器」欄

可以使用一些方法改變「記錄檢視器」的外觀，如表 2-1 所示：

表 2-1 處理記錄檢視器欄

目的：	**方法：**
調整欄大小	按一下右邊的欄分界線，並向需要的方向拖曳。
重新排列欄	按一下並將標題儲存格拖曳到需要的位置。
隱藏欄	在「檢視」功能表中選擇隱藏 / 顯示欄，取消對要隱藏的欄的選擇。
按欄值排列記錄順序	按一下欄的標題儲存格。
按多欄排列記錄順序	按住 Ctrl 鍵並按一下每個想要用於排序的欄標題，按下的順序就是希望它們排列的順序。
將排序反轉	再按一下欄的標題儲存格。
返回至原始順序	按一下列 # 欄標題。
對列進行重新編號	從「檢視」功能表中選擇「重新編號」(無法復原該功能)。
搜尋一個數值	從「檢視」下拉式功能表中選擇「搜尋」面板，在「尋找」欄位中輸入一個值，然後反白標示希望搜尋的欄，按一細化下尋*找下一個*或*尋找上一個*。要選擇多個欄，可以使用 Shift 按一下；對於非連續的欄，可以使用 Ctrl 按一下。

在「記錄檢視器」中篩選資料

可以對資料進行篩選，這樣就可以只看到特定範圍內的值。

1. 從「記錄檢視器」視窗中，選擇「檢視」>「篩選工具」面板。
2. 選擇一個欄或表達式，從中建立篩選表達式。例如：在汽車資料集中，可以建立表達式 `'cubicinches'>400` 來篩選那些容量大於 400 ci 的記錄。(請參閱圖 2-5)。

   可以選擇左邊的元件並按一下向右的箭頭來建立表達式；或是簡單地將表達式直接鍵入到本文欄位中。
3. 按一下*應用*，將立即開始篩選。面板將一直保留，直到關閉它。

要移除篩選，可以清除表達式視窗中的表達式，然後按一下*應用*，或從「記錄檢視器」視窗中選擇「檢視」>「移除篩選」。

圖 2-5 記錄檢視器篩選面板

可以開啓任意多個篩選面板。要一次應用多個篩選，應該先應用一個，對列進行重新編號，然後再應用下一個。

注意：無法復原對列的重新編號。要回到原始資料，必須重新開啓檔案。

在「記錄檢視器」中儲存資料

可以在「記錄檢視器」中儲存文件，使用「檔案」>「儲存」或「檔案」>「另存新檔」。

利用「另存新檔」，可以將資料儲存爲以下四種格式：二進位、ASCII、HTML 或純文字。當以二進位或 ASCII 格式儲存時，將儲存資料檔案 (表格) 和模式檔案 (向「工具管理員」和「資料移動器」描述資料檔案的內容)。HTML 格式將檔案儲存爲 HTML 表格。本文格式將檔案儲存爲跳位符分隔的形式，其中第一列爲欄標題。

用「統計可視化工具」查看記錄統計

用「統計可視化工具」可以發現資料集中記錄的詳細資訊。可以計算某些根據資料集中記錄數的統計。根據欄中的資料類型是*數值型*或*離散型*，分別把統計結果顯示爲方塊圖和柱狀圖。

認識方塊圖

方塊圖將顯示欄中數值的最小值、最大值、平均值、中值以及兩個四分位數 (25% 和 75%)，爲穿過垂直彩色長條的線。資料集總體標準偏差以 +/- 值表示。當不一致的值少於 50,000 個時，將顯示四分位數 (請參閱圖 2-6)。如果在欄中不同的值多於 50,000 個時，統計結果將顯示爲灰色垂直長條圖。關於術語*平均值*、*中值*和*標準偏差*，請參考詞彙表。

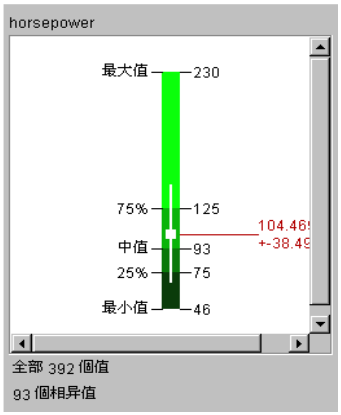


圖 2-6 「統計可視化工具」顯示的數值欄

認識柱狀圖

柱狀圖將顯示非數值資料 (如字串或分組) 欄的結果。(請參閱圖 2-7。) 這表示資料表中的欄可以包括字串 (如「yes」或「spore_color，」) 或分組的值 (如「10-70」)。可以有多達 100 個不同的項目。預設情況下，字串標稱所取的屬性值是按照計數降冪排列，但是可使用「檢視」下拉功能表來選擇另一種排序方法。字串的兩種排序方式分別爲：按照計數排列 (或在選取權重時按照權重排列)，或按照名稱的字母順序排列。如果不同的類別少於等於 100 個，欄面板還將包含不同值的計數。如果值已經分組，則分組的次序與排序方法無關。

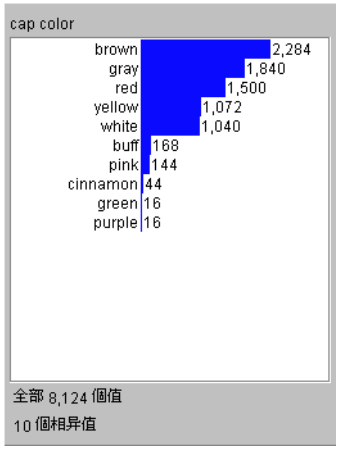


圖 2-7 「統計可視化工具」顯示的離散欄

建立資料的可視化處理後，當將滑鼠掠過一個欄位而沒有按下時，可在柱狀圖中看到已刪節的原文資訊。

執行「統計可視化工具」

在選擇了一個資料集後，執行「統計可視化工具」最簡單的方法就是：

1. 在「工具管理員資料目標」窗格中，按一下「可視化工具」索引標籤。
2. 在下面一列的索引標籤中，按一下「統計」索引標籤以進入「統計可視化工具」(圖 2-8)。
3. 按一下窗格右下角的*調用工具*。

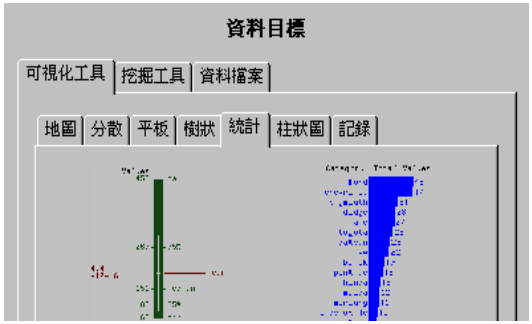


圖 2-8　　選定「統計可視化工具」的「資料目標」面板

關於執行工具的其他方法，請參閱 *《MineSet 3.0 企業版參考指南》*中的「統計可視化工具」。

使用樣本資料檔案工作

可以使用一些樣本資料檔案進行工作，以熟悉「統計可視化工具」的特性和功能。這些檔案在 *MineSet 3.0* 下的 *\examples* 目錄中，位於 MineSet 的初始安裝目錄。

從「工具管理員」功能表列上的「可視工具」功能表中啓動「統計可視化工具」，然後使用「檔案」>「開啓」下拉式功能表開啓所有 *.statviz* 檔案。

執行「柱狀圖可視化工具」

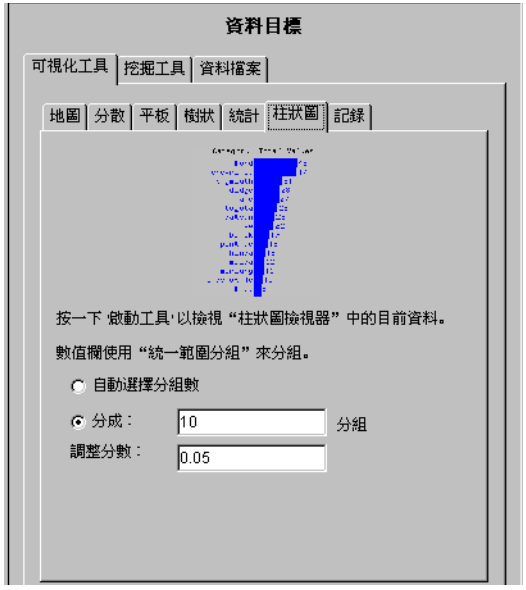
「柱狀圖可視化工具」自動將所有包括連續 (數值型) 值的欄分組，並將結果傳送到「統計可視化工具」進行顯示。關於*分組*，請參閱詞彙表。

「執行「柱狀圖可視化工具」的最簡單方法是執行下列步驟：

1. 在「工具管理員資料目標」窗格中，按一下「可視化工具」索引標籤。
2. 從工具選擇卡中，選擇「柱狀圖」(圖 2-9)。

3. 可以設定以下選項或直接到步驟 4 ：

   - 選取分組數，或允許 MineSet 進行選擇。
   - 設定修改分數，表示分組產生前所排除的極值分數。預設值爲 0.05。這將排除 5% 帶有極值的實例 (最低端 2.5%，最高端 2.5%)。修改會減小外層對臨界值產生的影響。外層是位於資料主體外部的單獨實例。



2-9 帶有「柱狀圖可視化工具」的「資料目標」面板

4. 按一下窗格右下角的*調用工具*。

   「工具管理員」視窗底部的「狀態」視窗將顯示運算法則處理資料的進程。然後目前的資料會以柱狀圖的形式顯示，從中可以對資料落在的位置有個快速的總體認識。請參閱第 23 頁 的「認識柱狀圖」。

在 MineSet 3D 可視化工具中瀏覽

本部分描述了各種 MineSet 可視化工具中的瀏覽控制項。樹狀可視化工具擁有的一組控制項將在下列的「在樹狀可視化工具中瀏覽」中進行介紹。非樹狀可視化工具擁有另一組控制項，在 第 29 頁 的「在「非樹狀可視化工具」中瀏覽」中作了介紹。

在樹狀可視化工具中瀏覽

「樹狀可視化工具」畫面就好像正透過一台照相機觀看場景。要變更檢視，可以改變照相機的位置 (觀察點)。本部分包括兩個表格，可作爲「樹」、「決策樹」、「選項樹」和「回歸樹」可視化工具控制項的快速參考。表 2-2 說明瀏覽按鈕。

表 2-2 「樹可視化工具」中的瀏覽圖示

圖示	操作
	將圖表返回爲首頁檢視設定的大小和位置。預設情況下，首頁檢視圖表的大小和位置是第一次調用可視化工具時所用的。可以使用下一個圖示改變首頁位置。
	爲圖設定新的首頁檢視。使用它來儲存特定的檢視或位置。
	將圖移到中央位置，讓視窗中能看到整個圖。
	復原上一次移動 (類似網頁瀏覽器上的「向後」按鈕)。
	重複已復原的移動 (類似網頁瀏覽器上的「向前」按鈕)。
	自一個節點移向根節點。
	自一個節點或長條移向左邊。
	自一個節點或長條移向右邊。
	自一個節點沿左邊路徑向樹的下方移動。
	自一個節點沿右邊路徑向樹的下方移動。
	跳現出一個功能表，其中包括來自目前節點的可能路徑。

表 2-3 列出的幾個操作可以用於樹狀可視化工具中的場景。大多數操作都可用可視化工具視窗上的一個控制項或滑鼠動作來完成。

表 2-3 操縱「樹狀可視化工具」場景

操作	滑動桿或滑輪	等同的滑鼠操作
在場景表面掠過	N/A	同時按住左右滑鼠按鈕 (或滑鼠中鍵) 並移動滑鼠
提高或降低長條高度以強調差別	高度滑動桿 (左上)	N/A
上下移動觀察點	H(水平) 滑輪	按住滑鼠右鍵，然後上下移動滑鼠。
左右移動觀察點	左右移動滑輪 (<-->)	同時按住左右滑鼠按鈕 (或滑鼠中鍵) 並左右移動滑鼠。
前後移動觀察點	Dolly(伸縮) 滑輪	同時按住左右滑鼠按鈕 (或滑鼠中鍵) 並上下移動滑鼠。
改變照相機的上下傾斜角度	Tilt(傾斜) 滑輪	N/A
沿指示的方向移動	N/A	同時按住 Alt 鍵和左右滑鼠按鈕 (或滑鼠中鍵) 並移動滑鼠。當向前移動時，觀察點也會按照目前的傾斜角度向下移動。類似地，當向後移動時，觀察點會按照目前的傾斜角度向上移動。
選擇一個節點的子節點	N/A	在父節點上按住 Ctrl 鍵並按一下滑鼠右鍵，然後在子節點上按一下，移動到子節點 (或使用展開瀏覽圖示)。

在「非樹狀可視化工具」中瀏覽

本部分包括兩個表格，可作爲「証據」、「地圖」、「分散」和「平板」可視化工具瀏覽控制項的快速參考。表 2-4 說明瀏覽按鈕。

表 2-4 「非樹狀可視化工具」中的瀏覽按鈕

按鈕	名稱	操作
	選取	改變程式至選擇模示 (箭頭)。在選取模示 (也稱爲選擇模示) 中，可以反白顯示 (刷過) 或選擇 (按一下) 圖中的元件。
	抓取	改變程式至抓取模式 (手形)。在抓取模式中，可以在視窗中移動圖表： — 要在視窗中移動圖表，可以按住滑鼠右鍵並移動滑鼠。 — 要旋轉圖表，可以按住滑鼠左鍵並移動滑鼠。 — 要對圖進行縮放操作，可以同時按住滑鼠的左右鍵 (或滑鼠中鍵) 並移動滑鼠。
	首頁	將圖表返回爲首頁檢視設計的大小和位置。預設情況下，首頁視圖表的大小和位置是第一次調用可視化工具時所用的。可以使用設定主檢視圖示來改變主檢視位置。
	設定首頁	爲圖設定新的首頁檢視。當您要儲存某個檢視或位置時，可以使用它。
	檢視全部	將圖移到中央位置，讓視窗中能看到整個圖。
	縮放	將所選取的點移動到窗格中間，並縮放它。當滑鼠游標成爲瞄準器形狀時，將它移動到希望看得更清楚的點，然後按一下滑鼠左鍵。
	3D	切換到 3D 視角。
	頂檢視	將圖改變爲頂檢視 (僅適用於「分散」和「平板」可視化工具)。
	前檢視	將圖改變爲前檢視 (僅適用於「分散」和「平板」可視化工具)。
	側檢視	將圖改變爲側檢視 (僅適用於「分散」和「平板」可視化工具)。

表 2-5 描述了非樹狀可視化工具的調整滑動桿和滑輪。

表 2-5 操縱「樹狀可視化工具」場景

操作	滑動桿或滑輪	滑鼠或鍵盤操作
在「選取」和「抓取」模式間切換	N/A	按下 Esc 鍵或瀏覽按鈕。
移動場景	N/A	在抓取模式中，按住滑鼠右鍵。將游標沿著希望圖移動的方向移動。
提高或降低塊、圓餅或長條圖的高度以強調差別	高度滑動桿 (左上)	N/A
繞坐標軸 X 旋轉場景	X 旋轉滑輪	在抓取模式中，按住滑鼠左鍵。將游標沿著希望圖旋轉的方向移動。
繞坐標軸 Y 旋轉場景	Y 旋轉滑輪	在抓取模式中，按住滑鼠左鍵。將游標沿著希望圖旋轉的方向移動。
將場景放大和縮小	Dolly(伸縮) 滑輪	在抓取模式中，按住左右滑鼠按鈕 (或滑鼠中鍵)。向下移動滑鼠放大場景，向上移動滑鼠縮小場景。
在細節級別中細化下尋 (僅適用於「決策表」和「地圖可視化工具」)	N/A	將滑鼠箭頭放在指定的圖表 (或所有圖表的背景) 上，然後按一下滑鼠右鍵。
在細節等級中概化上尋 (僅適用於「決策表」和「地圖可視化工具」)	N/A	將滑鼠箭頭放在指定的圖表 (或所有圖表的背景) 上，然後按住 Ctrl 並按一下滑鼠右鍵 (或按一下滑鼠中鍵)。

取得說明

除了在第 xv 頁 的「尋找 MineSet 資訊」中列出的資源以外，其他有關 MineSet 的資訊來源包括：

- 從 *《MineSet 3.0 企業版教學課程》*中可以迅速了解到各種工具的用法。
- 每個工具中都有上下文相關的說明：在感興趣的區域按 F1 鍵。
- http://mineset.engr.sgi.com/movies 上的線上展示影片總覽了一些工具的用法。

加工資料

本章介紹使用 MineSet 工具加工資料的概念。這些主題是建立分類器工作的預備知識，後者將在下章討論。本章包括以下幾個部分：

為何要加工資料？

面對一個陌生的資料集，可以先用「記錄檢視器」和「統計可視化工具」對其進行研究。(請參閱第 2 章，「利用 MineSet 存取資料」。) 本章中，將透過加工或轉換原始資料繼續該過程。例如：您可能會發現資訊太多以致於無法了解其意義。簡化可視化處理最簡單的方法就是對欄進行增加、移除或合併。該操作不會改變基本資料，改變的只是目前工作的會話檔案。應用的轉換順序稱爲該「工具管理員」會話的歷史。

「工具管理員」是一個圖形使用者介面 (GUI)，可以使用它完成與 MineSet 組件的大部分交互任務。可以用「工具管理員」選擇現有的資料來源，轉換或分析該資料，並使用任何一個 MineSet 工具可視化結果。

關於如何與 MineSet 可視化工具互動的詳細資訊，請參閱*《MineSet 3.0 企業版參考指南》*的指定工具項目。

注意：「工具管理員」可能需要一些手動操作，讓非 MineSet 建立的資料檔案能相容；請參閱 *《MineSet 3.0 企業版介面指南》*。

用「工具管理員」轉換資料

按鈕 (如圖 3-1 中所示) 的功能列於表 3-1 中。要轉換資料，可以從「目前欄」窗格中選擇欄，然後按一下適當的按鈕。

1. 執行 Mineset (請參閱第 2 章中的「執行 MineSet」)。
2. 使用「文件」 >「開啓新的資料文件」下拉功能表，向「工具管理員」視窗中選擇一個資料集。在「開啓資料檔案」對話方塊中按一下檔案，在「資料轉換」窗格的*目前欄* 視窗中將出現資料集的欄標題 (圖 3-1)。

圖 3-1 「資料轉換」窗格

表 3-1 「工具管理員資料轉換」窗格上的按鈕功能

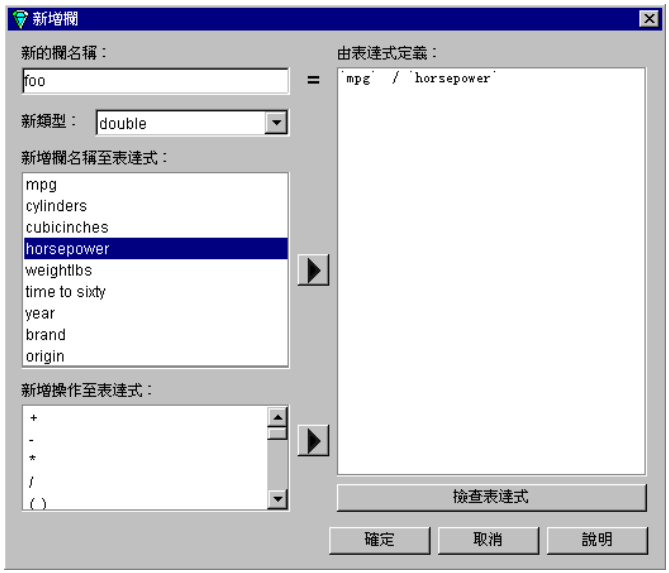
按鈕名稱：	**操作：**
移除欄	刪除目前所有選定的欄。
分組欄	將每個記錄分配到一定的範圍 (分組)。
組合	在分組的同時，執行一些欄的總和、平均值、最小值、最大值或計數組合。也可以排列索引組合的欄。
篩選	根據布林表達式建立資料的子集。只保留表達式判定爲真的記錄。
更改類型	可以更改欄的名稱和類型，如：從浮點改爲整數。(請參閱表 3-3 中的說明。)
增加欄	根據數學表達式增加新欄，表達式中可以包含其他欄。
應用模型	用現有的模型標示新記錄或修正資料。
樣本	在大資料集中選擇資料的隨機子集。
顯示已排序的欄	按字母順序排列資料集的欄。

移除和增加欄

要移除或增加欄，可以在「工具管理員資料轉換」窗格中進行以下操作：

移除欄—選擇與可視化處理或挖掘無關的欄，然後按一下「移除欄」按鈕。刪除無關的欄可以節省計算時間。要選擇多個非連續的欄以同時移除，可以在選擇新增欄時按住 Ctrl 鍵。對於連續欄使用 Shift 鍵。

增加欄—按一下*增加欄*按鈕打開對話方塊，在其中指定新欄的名稱和數學表達式 (圖 3-2)。例如：可以根據「age」欄增加一個稱爲「minor_age」的新欄，使用表達式：「如果年齡小於等於 18，則 minor_age 爲真；否則 minor_age 爲假。」這樣的表達式寫作：「if ('age'<= 18) then ('minor_age'= 1) else ('minor_age'= 0)。」



3-2 「增加欄」對話方塊

要完成該操作，按一下「增加欄」按鈕，將出現一個對話方塊 (圖 3-2)。

1. 在對話方塊中輸入新欄的名稱，並使用下面的跳現式功能表指定欄類型 (整數、字符串、浮點型等等)。欄名稱中不能包括空格。
2. 在大的本文項目區域中，填入表達式的定義：
   - 使用左下的捲軸清單從可用的欄和操作符中快速選擇。
   - 要在表達式中插入欄名稱和操作符，可以在捲軸清單中連按兩下，或者先選取它然後按一下清單的向右箭頭。
3. 按一下 *檢查表達式* 按鈕，檢查表達式語法是否正確。按一下「確定」，關閉狀態方塊。再次按一下「確定」，關閉「增加欄」對話方塊。

「增加欄」對話方塊將檢查類型的相容性：如果爲字串欄分配數值表達式 (反之亦然)，就會出現一個警告訊息，將自動更正新欄的類型。

為欄改變或建立新的分組

要組織模型或可視化處理，通常必須對記錄進行分組，尤其要建立動畫需要的滑動桿。可以更改分組的邊界，用不同的欄數目和範圍進行嘗試，以簡化可視化處理。

例如：如果希望將年齡範圍分爲： 0-18, 19-30, 31-50, 51-60, 61+ 。這些範圍表示爲：(...18], (18 ...30], (30 ... 50], (50 ...60], (60 ...]。數字旁邊的圓括號表示範圍中不包括該數字。數字旁邊的方括號表示範圍中包括該數字。請參閱詞彙表中的項目*分組*，以及*《MineSet 3.0 企業版參考指南》*中的「分組」。

在「工具管理員資料轉換」窗格中，按一下*欄分組* 以開啓一個對話方塊，指定分組選項 (圖 3-3)。

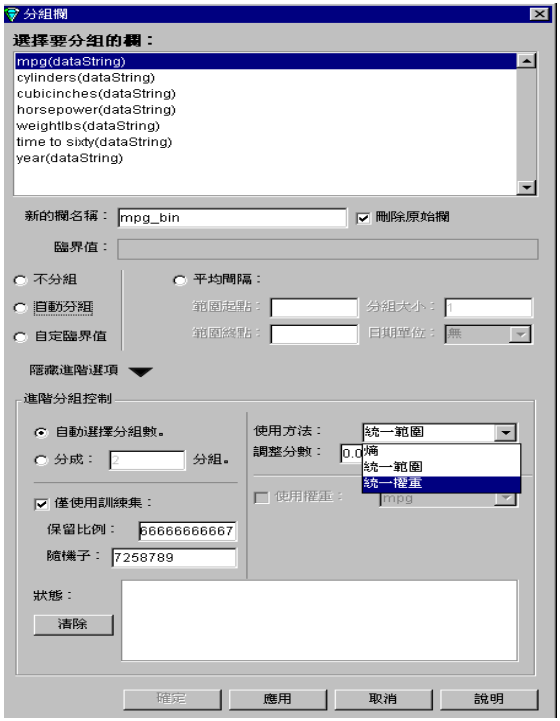
圖 3-3 「欄分組」對話方塊

1. 在對話方塊的頂部窗格中，按一下要分組的欄。預設值爲「未分組」。選擇分組方法，並在「新欄名稱」中選擇分組欄的名稱，或者直接鍵入新名稱。如果選擇多個欄進行分組，「新欄名稱」將處於非活動狀態。

   要刪除原來的欄，請確定已選取標示爲「刪除原來的欄」的方塊。

2. 選擇「自動分組」、「自定臨界值」或「平均分隔」(即平均分隔的分組)。
   - 自動分組讓 MineSet 經由機器學習建議分組臨界值。
   - 可以使用自定臨界值在「臨界值」本文欄位中指定分組臨界值。
   - 平均分隔可以讓您指定範圍的起點、終點及每個分組的大小。

   按一下「顯示進階選項」箭頭，將出現「進階分組控制」面板 (圖 3-4)。



3-4 「進階分組選項」面板

3. 在「進階分組選項」面板中，可以進行以下選擇：

表 3-2 進階分組選項

選擇	產生的操作
自動選擇分組數	MineSet 自動確定最佳的分組數。
分爲 ___ 組	將資料集劃分爲指定的組數
使用方法	在該功能表中可以選擇三種建立分組的方法。
熵	選取分組，使每個分組中的標籤值相似，讓熵或分組標籤的混亂度最小化。
統一範圍	將資料集在範圍內進行平均分布的分組。
統一權重	將資料集劃分爲具有相等權重 (實例數) 資料的組。
修剪因子	(僅適用於「統一範圍」和「統一權重」)。刪除分組臨界值以外的資料。外層是位於資料主體外部的單獨實例。
使用訓練集	將資料集劃分爲子集，並且僅使用佔總體「支持比」比例大小的子集建立分組。
支持比	指定用於建立分組的資料集的百分比。
隨機子	指定開始劃分資料集的起點。隨機子數值隨時間改變。如果希望下次執行時獲得完全相同的資料，必須指定相同的隨機子。

當選擇*自定臨界值*設定您自己的分組臨界值時，也可以輸入：

- *臨界值*

  在文字欄位中，輸入由逗號隔開的臨界值範圍，如果範圍中不包括該數字，則在其旁邊用圓括號；如果範圍中包括該數字，則在其旁邊用方括號。例如：所有 18 歲以上 (不包括 18 歲)，30 歲以下 (包括 30 歲) 的年齡表示爲： (18 ... 30].

  在「使用」的自定臨界值欄位中的日期項目需要使用雙引號。如果輸入的日期不帶有雙引號，系統將自動新增。

  "1/1/99", "2/1/99", "3/1/99", "4/1/99", "5/1/99", "6/1/99"

  也可以使用 *mm/dd/yyyyy* 格式 (例如：「5/22/1999」)。

- *平均分隔*

  當按一下該按鈕時，將會啓動下面的本文欄位。在適當的欄位中輸入分組範圍的起點、終點及分組大小。

對*平均分隔*臨界值使用的日期不要加雙引號。在該欄位中可以使用兩種日期格式 (1/1/96 或 01/01/1999)。

在*《MineSet 3.0 企業版參考指南》*中的「分組」中可以獲得詳細資訊。

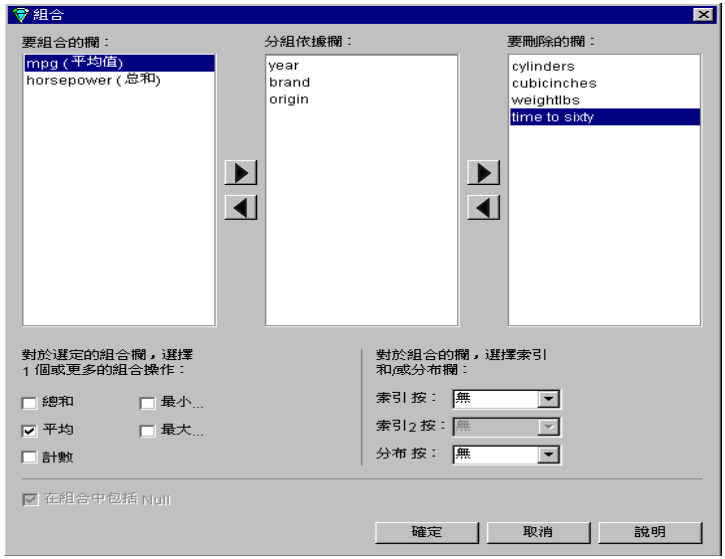
透過組合建立新欄

可以經由組合建立現有欄內容以外的新欄。例如：使用客戶波動資料集可以在欄中建立一個資料的子集，其中包含每個州的平均客戶波動數。也可以使用各種準則分布欄中的內容。因爲組合重新對資料進行加工，因此參考*《MineSet 3.0 企業版參考指南》*是很重要的，可以了解陣列和分布的背景，獲得對該特性更全面的認識。請參閱*組合*在詞彙表中的定義。

組合過程將原始資料集中的列集合分成新的組合列，以減少資料集中列的總數。

要建立組合，請按照以下步驟進行：

1. 在「工具管理員資料轉換」窗格中，按一下*組合*，開啓一個對話方塊以建立簡易組合、產生陣列或分布欄 (圖 3-5)。目前資料集的欄最初會出現在中間的「分組索引欄」文字欄位中。



3-5 「組合」對話方塊

2. 選擇要組合的欄，然後按一下左箭頭按鈕，將欄移到左邊。將「分組索引欄」項留在中間 (希望索引的欄)，然後將所有其餘的移到右欄中。(在選擇的同時按住 Ctrl 鍵以選擇多個欄。)

   要組合的欄—將欄移到此處以建立組合。組合欄中的值將求和、平均、指定最小值或最大值、或計數，取決於在面板底部選取的核取方塊。

   分組索引欄—保留在此處的欄在操作中不會改變。對於分組索引欄中每個具有相同組合值的記錄集合，產生表格中的輸出只有一個記錄。

   注意： 如果在分組清單中有許多欄和數值欄，在結果中得到的列可能要比期望的多。通常，在分組清單中使用的是具有很少值的幾欄。

   要移除的欄—移除的欄不用於組合。

3. 按一下欄下的核取方塊，指定操作中如何對值進行組合：總和、平均、最小值或最大值、或計數。很少使用最小值和最大值。

   對於數值欄可選擇任何組合。對於其他類型，只允許計數。每個選項都會提供獨立的結果；選擇平均和最大值將提供一個帶有平均值的欄，以及另一個帶有最大值的欄。預設欄名稱能夠識別用於建立它的過程 `count_state`、`avg_weightlbs` 等等。

4. 使用右下的跳現式功能表指定索引 (如果結果是陣列) 和欄分布 (如果可以分布結果)。

   陣列對於建立滑動桿很有用，而幾種可視化工具中的動畫都需要滑動桿。爲了顯示一個特徵是如何按照 (索引於) 另一個特徵而變化，必須事先將陣列中要索引的欄分組。關於陣列和分布的詳細說明，請參閱*《MineSet 3.0 企業版參考指南》*中的「組合」。

5. 按一下「確定」。

6. 請確定「資料轉換」視窗的「目前欄」本文方塊中顯示的新欄名稱是應用準則得到的。

在*《MineSet 3.0 企業版教學課程》*中詳細描述一個使用組合的範例，通過州來區分平均客戶波動。

透過篩選限制欄內容

可以用基於包含欄值的表達式選擇資料的子集，例如：可以只保留那些年齡小於 20 歲的記錄，或小於 25 英哩 / 加侖的記錄。一旦使用「工具管理員」載入資料集，就可以開始按以下步驟進行篩選：

1. 在「工具管理員資料目標」窗格中，按一下*篩選*，開啓一個對話方塊，使用數學表達式篩選資料 (圖 3-6)。

圖 3-6 「篩選」對話方塊

2. 從左邊選擇一個欄名稱和操作符，並在右邊建立一個表達式，或者直接在「由表達式定義」文字欄位中將其鍵入。產生的表格中只包括表達式判定爲真 (或者，如果是數值型，爲非零) 的記錄。請參閱*《MineSet 3.0 企業版參考指南》*中的「篩選面板」。關於表達式定義語言的完整介紹，請參閱*《MineSet 3.0 企業版介面指南》*。

更改欄類型或名稱

一些資料庫將數值型的值儲存爲字串。Oracle 資料庫將所有數字 (整數和實數) 以單一格式儲存，預設值爲「工具管理員」中的*雙精度*資料類型。使用*更改類型*按鈕以確保能正確處理這些值。使用同一個按鈕更改欄名稱。

更改欄類型

1. 在「工具管理員資料轉換」窗格中，按一下*更改名稱 / 類型*。將出現一個對話方塊，顯示欄標題及其類型 (圖 3-7)。
2. 選擇可以轉換的欄，然後按一下*新類型*功能表。清單中顯示可能的欄類型 (無效的類型顯示爲灰色)。對選擇的解釋請參閱表 3-3。

圖 3-7 「類型」跳現式清單

表 3-3 欄類型含義

類型	含義
整數	32 位元符號整數。
浮點	單精度浮點數，小數點爲可選的。
雙精度	雙精度浮點數，小數點爲可選的。
資料字串	*資料字串* 的值不是保存在一般字串表中，因此每個資料字串值都是獨立儲存的，即使多個值都是相同的。很少使用。
字串	儲存在一般字串表中的字元串，可以節省記憶體。如果無法確定要使用 *字串* 或 *資料字串*，請使用字串。
日期	指定爲資料庫中的日期類型。
分組	由分組操作建立的欄。

3. 選擇了新類型後，按一下*應用*，使更改生效。
4. 按一下*關閉*，結束對話方塊。

如果試圖將不合適的欄位 (如名稱) 轉換爲數字，產生的值將都爲零。

更改欄名稱

更改欄名稱可以在與更改欄類型相同的對話方塊中進行。

1. 從「更改類型」對話方塊中，選擇希望重新命名的欄，在「新名稱」文字欄位中鍵入新名稱，按一下*應用*。
2. 然後按一下*關閉*，退出對話方塊。

應用模型

如果事先建立了模型，可以使用*應用模型*按鈕標示目前表格中的新記錄、估計標籤值的機率、測試模型在目前表格上的性能、或在現有的模型上修正目前表格。關於相關的範例，請參閱第 7 章，「理解預測建模」。

資料採集

如果資料集太大，以至於不能有效工作，可能要使用一個隨機採樣的子集。載入資料檔案後，從「工具管理員資料轉換」窗格中：

1. 按一下*樣本*按鈕，開啓「採樣」對話方塊 (圖 3-8)。
2. 選擇一個選項：

   按一下「百分比」核取方塊，然後在文字欄位中鍵入希望採集的目前資料集的百分比。該百分比是近似值。

   按一下「計數」核取方塊，建立樣本中需要的記錄數。

圖 3-8 「採樣」對話方塊

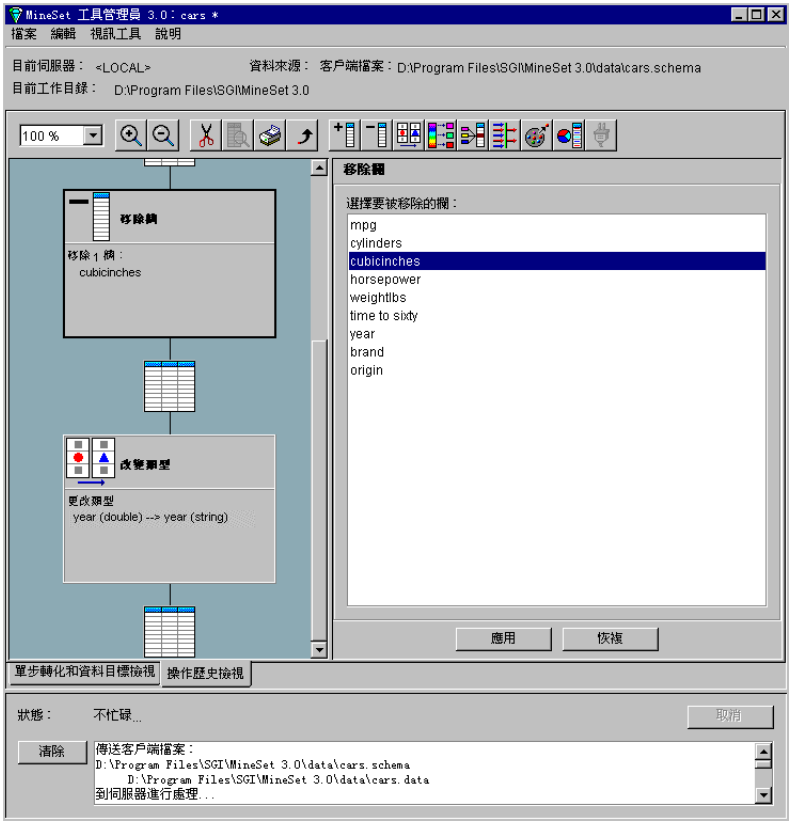
3. 按一下「*互*補樣本」核取方塊，以獲得除了在隨機樣本失敗的所有記錄。例如：如果在需要 10% 的樣本而沒有核取「*互*補樣本」方塊時，則當您核取該方塊時，就可以獲得其餘 90% 的資料。按一下「確定」。
4. (可選) 在「隨機子」欄位輸入數值，由此產生隨機樣本。如果沒有指定數值，根據目前時間產生的隨機數就作爲隨機子。如果希望不同的隨機樣本，必須指定不同的隨機子。對不同的資料集資料挖掘過程使用同樣的隨機子，可以使您每次都使用同樣的隨機樣本工作。當您需要測試新的在資料集中發現的支持程度時，請改變隨機子。請參閱詞彙表中的項目*隨機種子*。

用歷史表回溯歷史操作

要查看以前操作的歷史，以及在改變主意或發生錯誤而需要返回時，可以使用「資料轉換」窗格底部的兩個*歷史表*按鈕。按一下左箭頭按鈕，顯示先前步驟中的表格。按一下右箭頭按鈕，返回到目前狀態的表格。關於按鈕含義的說明，請參閱表 3-4。

表 3-4 歷史表含義

欄位	含義
目前檢視爲	計數將發生變化，指出正在查看的步驟。
上一操作：和下一操作：	請注意進行的操作，有助於跟蹤。
編輯上一個操作	開啓對話方塊，編輯顯示在*上一個：*的操作欄位。請注意以後操作中所用的刪除操作。
刪除操作到終點	移除目前步驟以後的所有操作。如果您還未到達操作歷史的結尾，「資料目標」面板將在按一下按鈕後再次啓動。
「操作歷史檢視」索引標籤	開啓「資料轉換」表格的一個完整歷史 (圖 3-9)。
「單獨轉換和資料目標檢視」索引標籤	將您帶到歷史中一個單獨的點，用「工具管理員」顯示事件的狀態。



3-9 「查看歷史」對話方塊

同*編輯上一個操作*，更改一個操作通常會影響 (有時會使其無效) 歷史中的後續操作。可以選擇一個指定的操作進行編輯、增加或檢視。當更改影響歷史時，操作歷史檢視會發出警告，並顯示更改後的新歷史。按一下圖表，返回到先前的操作，或按一下操作之間的表以到達操作之間的表格狀態。

要回到「工具管理員」中的上一個位置，可以按一下「單獨轉換和資料目標檢視」索引標籤。

加權記錄

在其他轉換資料的方法中，如果需要，加權記錄可以使一些記錄比其他記錄更重要，或不重要。例如：電話公司將所有詐欺電話儲存在資料集中，同時只儲存了非詐欺電話的一小部分。經過使用記錄加權，可以使每個記錄充份反映其對總體的真實作用。

一些資料集已經組合過，這些記錄具有一個與其自身相關的自然「計數」(例如：關於美國城市的統計通常具有一個相關的人口計數)。該計數屬性可以對照到權重，它相當於將每個記錄按照計數次數進行複製。

記錄加權的語義學即爲：記錄權重爲 2 相當於兩個記錄權重爲 1 的記錄。允許浮點型權重。

記錄的加權可以通過所有導入工具中的*進階選項*對話方塊完成 (請參閱*導入工具*在詞彙表中的定義)。載入資料集後，請按照以下步驟加權記錄；

1. 在「工具管理員資料目標」窗格中，選擇「挖掘工具」索引標籤，然後按一下「分類」索引標籤。
2. 從「模式」跳現式功能表中，選擇任何的導入工具。
3. 按一下*進階選項*按鈕，然後在出現的對話方塊中，按一下「使用權重」核取方塊。這表示在步驟 4 中選取的屬性將對記錄進行加權。
4. 爲權重選擇欄。
5. 如果步驟 4 中使用的屬性也作爲普通屬性出現，請按一下「權重屬性」核取方塊。如果沒有核取該方塊，權重屬性將不會影響導入過程。
6. 按一下*確定*，接受這些選項，然後按一下*開始*，執行選定的導入工具。

尋找重要的欄

欄重要性有助於發現在預測選定標籤欄的不同值的過程中最重要的欄。欄重要性和聚類的區別 (請參閱第 12 章，「用聚類劃分資料」) 在於：利用欄重要性可以決定使用哪個標籤來確定欄的重要性。而在聚類中，雖然資料本身就會顯示區分因素，卻不提供標籤。

例如：使用欄重要性尋找「分散可視化工具」中對照到坐標軸的最好的三欄。您只需選擇標籤並執行工具，便會出現一個視窗，其中的三欄就是最好的區分器。一個稱爲「純度」(0 到 100 的一個數) 的測量會告訴您用這些欄區分不同標籤的效果。增加更多的欄會提高純度。

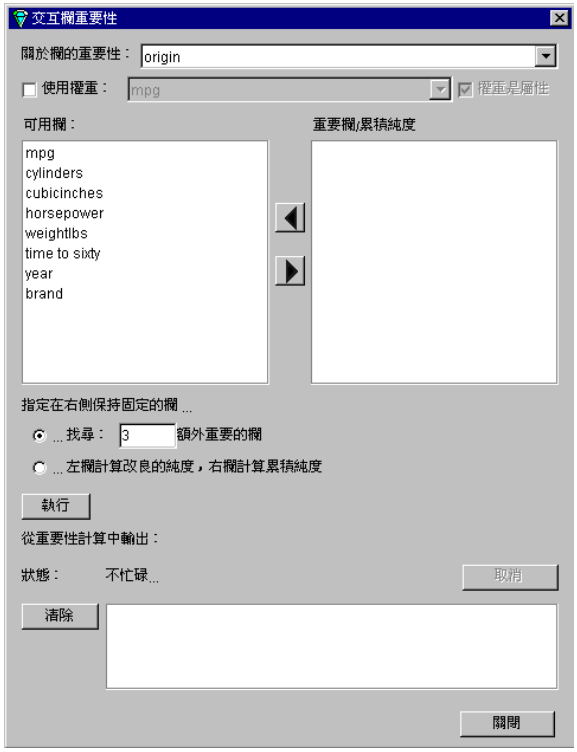
圖 3-10 「欄重要性」索引標籤

1. 在「工具管理員資料目標」窗格中，按一下「挖掘工具」索引標籤，然後再按一下「Col. Imp.」(欄重要性) 索引標籤，如圖 3-10 所示。
2. 若要以簡單模式執行，可以在跳現式功能表中選擇離散標籤，指定一數在「尋找」欄位中指定一個數字，然後按一下*開始*。底部的狀態視窗將出現結果。

   若要控制欄的選擇，可以按一下*進階模式*按鈕，這時會出現「進階欄重要性」對話方塊，(請參閱圖 3-10)。

3. 在對話方塊中，可以按一下「使用權重」核取方塊，並從「關於欄的重要性」跳現式功能表中選擇一欄。這將有助於決定對於確定重要性，欄是否相當於一個常規屬性。請參閱第 49 頁的「加權記錄」。關於工具操作的詳細內容，請參閱*《MineSet 3.0 企業版參考指南》*中的「欄重要性」。

4. 對話方塊中包含兩個欄名稱清單：左邊的清單中包含可用的屬性，右邊的清單中包含 (由您或「欄重要性」運算法則) 選爲重要的屬性。按一下向右箭頭，可以將選取的欄從左邊移到右邊。



3-11 「欄重要性」的進階模式

5. 在對話方塊的中間部分有兩個進階模式選項：

   - 尋找多個重要屬性

     按一下 *…尋找 [數字] 個其他的重要欄*上的選項鈕。如果直接按一下*開始*而沒有進一步的更改選項，則效果將與簡易模式一樣。找到指定數目的重要欄並移動至右邊的欄中，同時指定累積純度。

     但是，如果按一下向右箭頭，將欄名稱從左邊清單移到了右邊清單，就可以指定這些已包括的欄，並讓系統增加更多的欄。

     按一下*開始*，查看每個欄的累積純度，以及清單中先前的純度。純度爲 100 表示：使用指定的欄，可以完全區分資料集中不同的標籤值。

   - 對可用屬性排序

     可以檢驗將每一欄增加到右邊清單中時純度的增幅。如果使用汽車資料集，如圖 3-11 所示，可以將欄*汽缸*移到右邊的清單中，然後請求系統計算保留在左邊欄表中的每一欄能夠產生的純度增幅。爲右邊的欄 (已標記爲重要的) 計算累積純度。結果將顯示在右邊的欄中。

     按一下 *... 對左欄各欄計算改良純度，對右欄各欄計算累積純度*上的選項鈕。該子模式允許對過程進行精密控制。如果兩個欄排得非常近，您可能希望使一欄替換另一欄 (可能因爲它收集較便宜、更可靠、或者更易於認識)。

爲「分散和平板可視化工具」尋找最佳的三個坐標軸時，欄重要性很有用。爲「樹狀可視化工具」尋找良好的區分層 (區分開不同標籤值的分層，此時該標籤應被選擇爲「樹狀可視化工具」的關鍵字) 時，它也很有用。

關於使用「欄重要性」樣例檔案的詳細內容，請參閱*《MineSet 3.0 企業版參考指南》*中「欄重要性」。

用分散和平板可視化工具檢查資料

「分散」和「平板」可視化工具比較類似，其區別在於「分散可視化工具」將資料顯示爲一系列單獨的實體，而「平板可視化工具」顯示的資料點組合，看上去像是 3D 景觀中不同透明度和顏色的雲。要進一步了解它們各自的優點，請參閱*《MineSet 3.0 企業版參考指南》*。本章主要包括以下幾個主題：

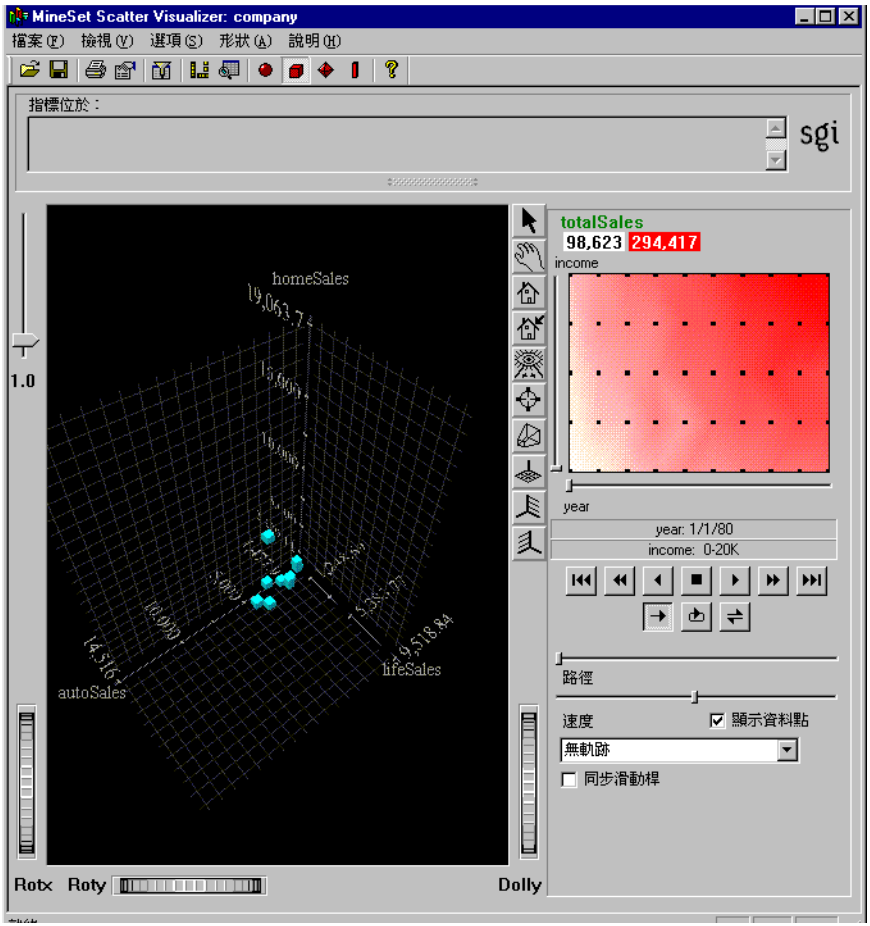
提供一些樣本配置和資料檔案，用於演示「分散」和「平板」可視化工具的特性和功能。這些檔案在 *MineSet 3.0* 下的 *\examples* 目錄中，位於 MineSet 的初始安裝目錄。

分散和平板可視化工具總覽

本部分解釋「分散」和「平板」可視化工具的用法，以及選擇使用時的一些原則。

分散可視化工具總覽

「分散可視化工具」的單個資料點對應於資料檔案中的列。當資料點少於 50,000，或對資料採取一些處理使其成爲小的組合集合時，此種可視化處理的效果很好。「分散可視化工具」產生的分散圖可以製成動畫以清楚地顯示關係 (圖 4-1)。



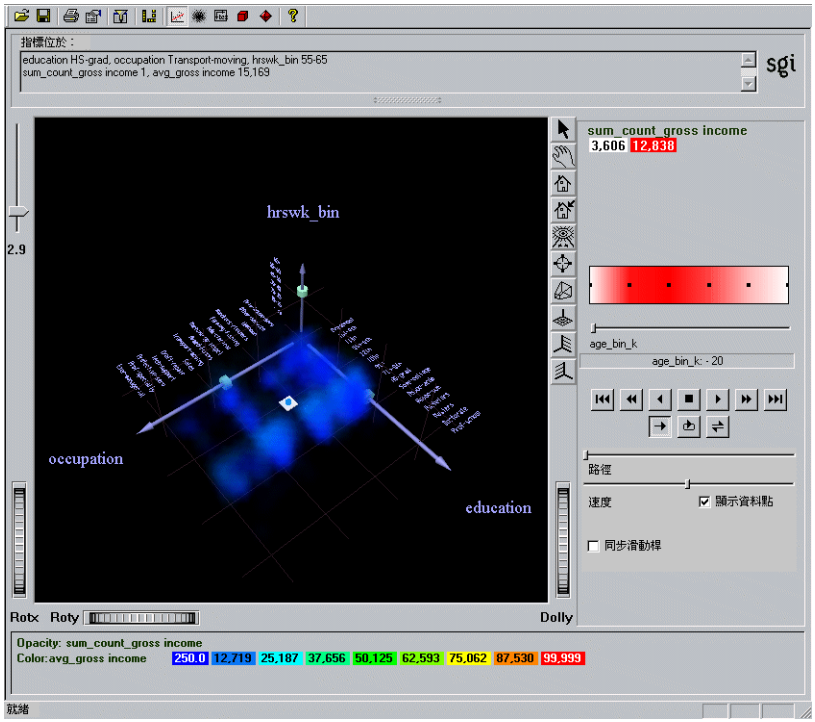
4-1 分散可視化工具螢幕樣本

「分散可視化工具」顯示 3D 景觀，其中的資料欄對照到實體和元件 (如坐標軸、大小和顏色)。如果將一個或兩個數值型變數對照到滑動桿，就可以移動實體的大小、顏色或位置。在圖 4-1 所示的範例中，資料代表數個公司隨時間變化的銷售額。如果將時間變數對照到一個滑動桿，將銷售變數對照到大小，則實體會隨時間滑動桿的移動而增長或縮短。該範例爲 *MineSet 3.0\examples\company.scatterviz*。

當播放回動畫路徑時，可以看到大小、顏色變化和資料點運動的趨勢或異常。可以在 3D 景觀中瀏覽以尋找感興趣的觀察點，或對變數值進行縮放以加強效果。要清除場景，可以對畫面進行過濾，僅顯示符合特定準則的實體。

平板可視化工具總覽

利用「平板可視化工具」，能夠可視化地分析多個變數之間的關係，當使用動畫特性時，可以更清楚地看到一些關係。「平板可視化工具」使用圖形物件，稱爲*平板*，代表資料點的組合。平板物件的顏色和不透明度在動畫過程中可以更改，而位置卻不能。



4-2 具有一維匯總滑動桿的平板可視化工具樣例

圖 4-2 顯示的是一個 3-D 景觀的「平板可視化工具」檢視，*adult94* 樣本資料集的欄對照到其中的坐標軸、滑動桿、顏色和不透明度。它類似於分散圖，只不過分散圖是單獨地畫出每個資料點，而「平板可視化工具」將彼此接近的資料點 (在同一個分組內) 組合起來並把它們畫爲一個單獨的平板物件。結果與你爲每個單獨的點畫一幅分散圖而得到的畫面很相似。所得影像可視爲 3D 的彩色柱狀圖。

從「平板可視化工具」的結果中，可以：

- 使用動畫面板可以了解資料中的整體移動及趨勢。變化的顏色和不透明度顯示實際的移動情況 (關於圖形的顏色版本，請參閱本書的線上版本)。
- 在 3D 景觀上掠過可以強調特定的維度或視點。
- 使用幅度滑動桿 (在主要視窗的左上方) 增強能見度，可以降低或提高平板不透明度。資料密集的區域顏色可能變化很小，因爲顏色是根據許多值的平均。
- 對畫面進行過濾，僅顯示符合特定準則的平板。可以對相應於坐標軸、滑動桿、權重、和顏色的欄進行過濾。
- 選取卷冊中單個平板的結構資訊。
- 用方塊選擇器定義一個選定的區域，以追溯原始資料或傳送到「工具管理員」。

例如：圖 4-2 中的左邊坐標軸顯示的是在軸上按平均收入排列的每個職業。行政管理職業列在坐標軸的末端，平均收入最高，並爲該值提供一個自然級數。另一方面，*教育* (圖 4-2 中右邊的坐標軸) 值的順序通常從低到高，但在少數情況下，順序上也會不規則。此種意外的順序可能很有意思，因爲它指出資料與期望不相符的地方。

為分散和平板可視化工具轉換資料

在轉換決策中要對資料有相當的計劃和了解。組成資料集的欄在對照到坐標軸、顏色或滑動桿之前，通常需要經過一定方式的處理或轉換。在查看一個結果後，可能需要經常返回並以另外的方式轉換資料。

在 MineSet 中，大部分的轉換是使用「工具管理員」中的「資料轉換」窗格來完成，(請參閱第 34 頁的「用「工具管理員」轉換資料」)。以下是一些可能需要考慮的轉換：

- 增加欄，通常是對已有資料應用一個數學公式。
- 移除無關、多餘、或包含明顯無用預測因子的欄。可以減小資料集的大小及加速處理。
- 篩選資料。例如：您可能只希望使用含有一定範圍內的值的特定屬性的記錄。
- 更改欄名稱或類型。
- 分組資料—將連續範圍的資料劃分到離散的區段中。
- 適用於「平板可視化工具」的組合資料—如果不用「工具管理員」手動組合資料，該工具將自動執行此任務。但是，最好在「工具管理員」中組合資料，以便在伺服器上完成任務，而不用將組合的資料集送到客戶端系統 (請參閱第 59 頁的「平板可視化工具的處理技術」)。
- 適用於「分散可視化工具」的組合資料—組合可以將大的資料集壓縮爲小的組合集合。組合後，表中的每一列都將對應於分散圖中的一個實體。在組合過程中可以爲動畫建立陣列欄。

「平板可視化工具」中的交互性與表現出來的資料點數目無關；它只取決於坐標軸維度中的分組數。如果資料集非常大，請在「工具管理員」中進行顯式組合。這可以是伺服器使用資料流操作執行進程，而非將整個資料集傳送到客戶端並在那裡進行組合要好得多 (請參閱第 59 頁的「平板可視化工具的處理技術」)。

表 4-1 所列爲可在「平板可視化工具」中對照到可視實體的欄類型。

表 4-1 平板可視化工具中允許對照的欄類型

實體	欄類型	要求
坐標軸	數值型	需要進行分組，手動或由「工具管理員」自動執行。
	分組型	可以直接對照。
	字串	可以直接對照。
顏色	數值型	可以直接對照。
不透明度	數值型	如果沒有對照，預設爲記錄計數。
滑動桿	數值型	必須進行分組，手動或由「工具管理員」自動執行。
匯總	數值型	可以直接對照。

表 4-2 所列爲可在「分散可視化工具」中對照到可視實體的欄類型。

表 4-2 平板可視化工具中允許對照的欄類型

實體	欄類型	要求
坐標軸	所有，包括陣列	可以直接對照。
實體 - 大小	數值型或陣列型	可以直接對照。
實體 - 顏色	所有，包括陣列	可以直接對照。
實體 - 標籤	所有，包括陣列	可以直接對照。
滑動桿	分組型	可以在組合過程中隱式對照。
匯總	數值型	可以在組合過程中隱式對照。

如果字串欄對照到坐標軸，則分組將定義爲該欄的標識值。這些值按順序排列，因此會與對照到顏色的屬性相關聯 (如果沒有屬性對照到顏色，則它們就與不透明度關聯)。沿著字串值的坐標軸查看顏色的變化可以了解該欄與對照到顏色的欄之關聯程度。

平板可視化工具的處理技術

在「工具管理員」中進行組合可以避免在客戶端上執行該任務。這意味著大部分的工作將由伺服器完成。通常，伺服器比客戶端更強大並執行資料流操作，因此不需將整個資料集載入記憶體中。而且，在「工具管理員」中組合可以避免將大的資料集傳送到客戶端。完成以上操作的步驟爲：

1. 使用「工具管理員」對用於坐標軸和滑動桿的數值欄進行分組。
2. 在按照坐標軸和滑動桿欄分組的同時，經由計數和平均組合對照到顏色的欄。
3. 將產生的計數組合對照到不透明度。
4. 將產生的平均組合對照到顏色。

以 *adult94* 資料集 (隨分布狀態提供) 爲例，以上所示的具體過程步驟列在下面。

1. 分組年齡和 `hours_per_week`。您將在需要分組值的滑動桿上使用前者，而將後者用於坐標軸上分組數值的範例。
2. 組合 `gross_income` 使用「計數」和「平均值」。保留「分組索引欄」窗格中的教育、職業、`age_bin` 和 `hours_per_week_bin`，並移除其他所有的欄。這將提供一個可視化處理，顯示 `gross_income` 按照其他三個準則變化的方式。
3. 將教育、職業和 `hours_per_week_bin` 對照到坐標軸。
4. 將 `avg_gross_income` 對照到顏色，`count_gross_income` 對照到不透明度，`age_bin` 對照到滑動桿。根據對照到坐標軸的準則，該對照使顏色和不透明度按年齡顯示收入 (平均和計數) 的變化。

當啓動工具時，請注意所有過程都是在伺服器上完成的，並且資料檔案 *adult94.splatviz.data* 中包含的列是原始資料中列的組合。產生的可視化處理如圖 4-2 所示。

執行分散和平板可視化工具

執行「分散」和「平板」可視化工具最簡單的方法是從「工具管理員」中：

1. 開啓「工具管理員」，選擇伺服器和資料來源。若需具體的介紹，請參閱第 15 頁的「執行 MineSet」。在第一個範例中，使用的是*客戶波動*資料集。
2. 在「工具管理員」的「資料目標」面板上，選擇「可視化工具」索引標籤；然後在下面一行索引標籤中，選擇「分散可視化工具」或「平板可視化工具」(圖 4-3)。

對照分散可視化工具可視元件

1. 在「工具管理員」的「資料轉換」窗格中，由捲動功能表中選擇，將「目當欄」對照到「可視元件」。例如：對於「分散可視化工具」，如果使用*客戶波動* 資料集，則爲坐標軸 1 選擇聲音郵件訊息的數目，爲坐標軸 2 選擇白天的費用總數，爲坐標軸 3 選擇晚上的費用總數。

圖 4-3 選定「 分散可視化工具」的「資料目標」面板

表 4-3 確認「分散可視化工具」可視元件的對照效果。

表 4-3 對照分散可視化工具中的可視元件

可視元件	操作：
坐標軸 1 坐標軸 2 坐標軸 3	只將資料分配給第一個坐標軸通常沒有用。將資料分配給全部的三個坐標軸以產生一個 3D 圖。
實體 - 大小 實體 - 顏色 實體 - 標籤	將大小、顏色和標籤分配給實體。用「分散可視化選項」對話方塊爲這些對照指定選項，可以按一下「分散可視化工具選項」按鈕開啓對話方塊。
匯總	確定匯總滑動桿背景的顏色
滑動桿 1 滑動桿 2	將欄直接對照到一個或兩個動畫滑動桿

2. 要復原對照，可以選擇元件，然後再從可用欄的清單中選擇 < 未指定 >。如果使用自己的資料集，可以考慮執行「欄重要性」工具以幫助確定可能的對照 (請參閱「尋找重要的欄」)。
3. 在「資料目標」窗格的右下角按一下*啓動工具*。

對照平板可視化工具元件

與「分散可視化工具」類似，「平板可視化工具」也要求在啓動前將欄連接到可視元件。在適合可視元件準則之前，一些欄可能需要處理。本範例使用的蘑菇資料集由系統提供，主要顯示如何對欄進行處理以滿足對照的要求。按第 60 頁的「執行分散和平板可視化工具」所述，從「工具管理員」開始，選擇蘑菇資料集，並選擇「平板可視化工具」。

1. 在「工具管理員」的「資料轉換」窗格中，由捲動功能表中，選擇將「目前欄」中的項目對照到「可視元件」中的項目。

圖 4-4 將欄對照到平板可視化工具的可視元件

在本範例，對於坐標軸 1 選擇產地，對於坐標軸 2 選擇氣味，對於坐標軸 3 選擇孢子印顏色。

表 4-4 確認「平板可視化工具」可視元件的對照效果。

表 4-4 對照平板可視化工具中的可視元件

可視元件	**操作：**
坐標軸 1 坐標軸 2 坐標軸 3	只將資料分配給第一個坐標軸通常沒有用。將資料分配給全部的三個坐標軸以產生一個 3D 圖。
顏色	需要一個數值欄以確定平板的顏色。如果沒有屬性對照到顏色，效果就會是單色的。
不透明度	預設情況下的記錄計數。如果在「工具管理員」中進行計數組合 (或加權欄的總和)，或者資料集中的欄已是根據計數或權重，就可以爲此元件使用該欄。

表 4-4 對照平板可視化工具中的可視元件

可視元件	操作：
滑動桿 1 滑動桿 2	將欄直接對照到一個或兩個滑動桿。欄必須是數值型或已分組。
匯總	確定匯總滑動桿背景的值。如果沒有對照匯總欄，預設情況下爲計數。如果對照了匯總欄，在匯總中會顯示該欄的加權平均值。

在對照欄中，如果需要數值欄，可以使用以下的表達式將兩值的字串欄變成新建立的數值欄： ('stringCol'=="value1")? 1:0。關於詳細內容，請參閱第 3 章中的「移除和增加欄」。

本範例中，增加一個整數欄，代表可食性的機率，以便將該特徵對照到顏色。蘑菇資料集只包含字串欄。非連續屬性 (如字元串) 不可能對照到顏色，因爲不可能平均這樣的屬性。

2. 在「工具管理員資料轉換」窗格中按一下*增加欄*按鈕。在產生的對話方塊中，建立一個新欄，名稱爲 `p_edible`，類型爲 `int`。表達式爲 `('edibility' == "edible")?100:0`。請注意引號的位置。

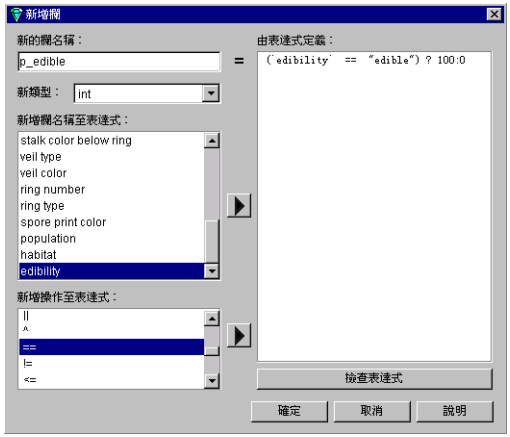


圖 4-5 爲對照增加一個整數類型的欄

3. 將 `p_edible` 對照到顏色，以根據顯示的顏色定義蘑菇是可食用或不可食用的。

   要復原對照，可以從下拉式功能表中選擇 < 未指定 >。

4. 在「工具管理員資料目標」窗格的右下角按一下*調用工具*。產生的可視化處理提供一些對蘑菇的有趣認識。

以下部分介紹 MineSet 產品提供的另一種存取樣本資料的方法。

檢視分散和平板可視化工具中的結果

與所有可視化工具一樣，如果在沒有指定配置檔案 (即：名稱結尾爲 *.scatterviz* 或 *.splatviz* 的檔案) 的情況下開啓「分散」或「平板」可視化工具，那么就只能使用「文件」和「幫助」下拉功能表。要顯示可視化工具主視窗中的所有功能表和控制項，必須開啓配置檔案。使用「檔案」>「開啓」以查看配置檔案的清單。這些檔案在 *MineSet 3.0\examples* 目錄中，位於 MineSet 的初始安裝目錄。

當開啓有效的配置檔案時，將會看見 3D 景觀。圖 4-1 所示爲 *company-total.scatterviz*，它顯示不同收入等級的人壽保險、汽車保險和家庭保險銷售額隨時間的變化。

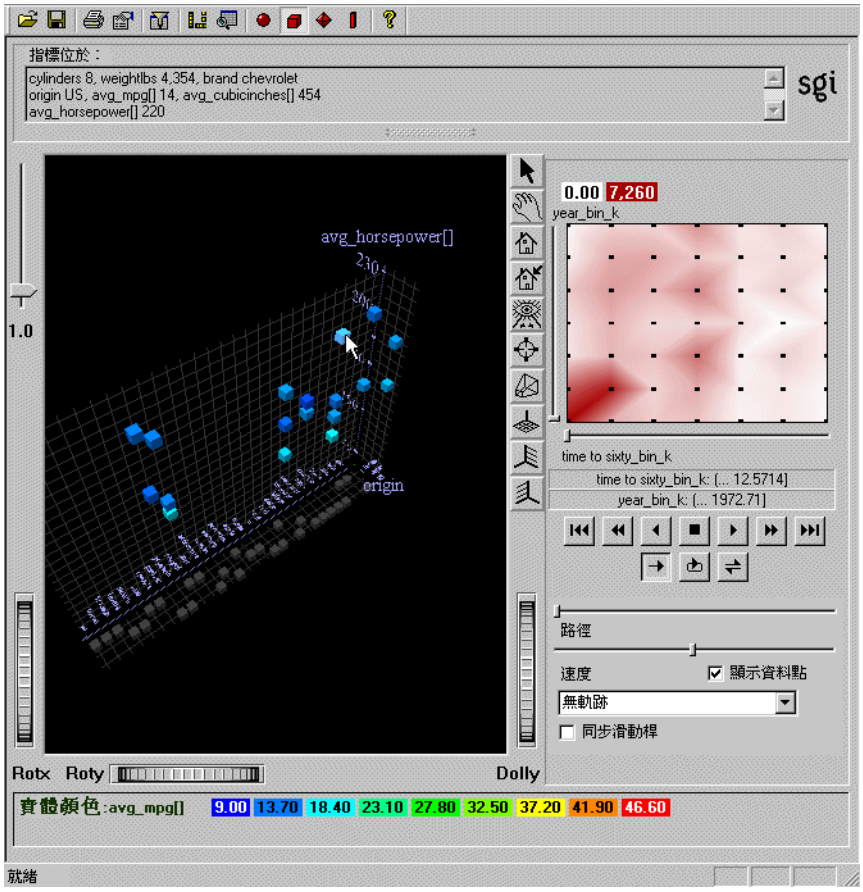
查看模式

「分散」和「平板」可視化工具的兩種檢視模式爲*抓取*和*選取*。要在這兩種模式之間切換，可以將游標移到主視窗中，然後按 Esc 鍵。按一下可視化工具視窗的主窗格邊界上的箭頭或手形，也可以從一種模式改變爲另一種模式。

分散可視化工具中的選擇模式

當處於「分散可視化工具」的選擇模式時：

- 要顯示「分散可視化工具」上部文字欄位中實體的有關資訊—將滑鼠移到物件上。
- 要選擇一個實體，在物件上按滑鼠左鍵。要選擇多個實體，按一下滑鼠左鍵的同時按住 Ctrl 鍵。
- 要清除選擇—按一下黑色背景。關於瀏覽的完整詳細內容，請參閱第 27 頁的「在 MineSet 3D 可視化工具中瀏覽」。



4-6 顯示選定實體的資訊

關於從指令行指定「分散」和「平板」可視化工具的資訊，請分別參閱它們在*《MineSet 3.0 企業版參考指南》*中的項目。

平板可視化工具中的選擇模式

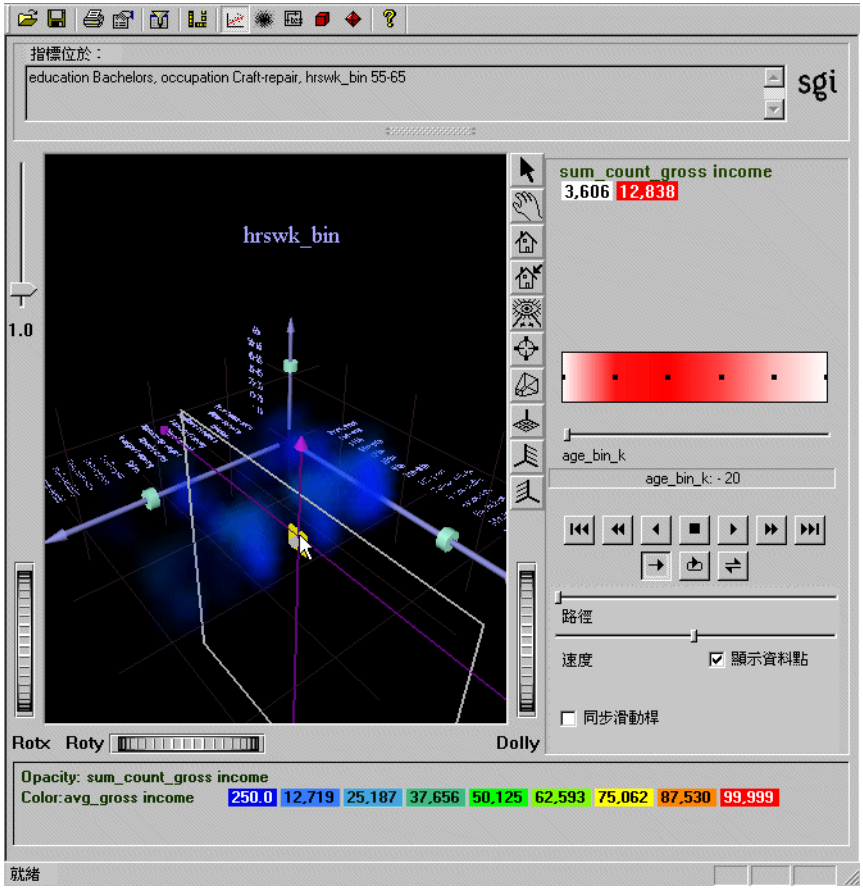
在「平板可視化工具」的*選取*模式中，可以移動 3-D 選取拖曳工具穿過資料點的密雲，尋找場景中區域的有關資訊。該選取拖曳工具由一個圓柱體和一個正方形組成。

- 要沿圓柱體的坐標軸平行移動，可以按一下圓柱體並將滑鼠指標沿希望的方向拖曳。
- 對於受正方形選定的平面限制的移動，可以按一下正方形並拖曳。使用 Shift 鍵將移動限制在平面內的一個坐標軸上。
- 要在平行與強制模式之間切換，可以在游標位於拖曳工具上時按下 Ctrl 鍵。(不需按滑鼠按鈕。)
- 另外，每個坐標軸都有一個圓盤，與選取拖曳工具的位置排成一行。在坐標軸上移動圓盤會使拖曳工具移動，反之亦然。

當選取拖曳工具位於資料上時，圓柱體的顏色改變爲資料下面雲團狀「平板」的顏色，而有關該區域的資訊會顯示在檢視區域頂部 (圖 4-7)。如果沒有資料，圓柱體將保持淡灰色，有關該位置的資訊會顯示在區域頂部以幫助瀏覽。

當完成拖曳並鬆開滑鼠按鈕時，目前所在「平板」圖形的資訊會顯示在頂部的選擇視窗中。如果移動了動畫滑動桿，該選取資訊就會更新。使用滑鼠可以將該選定資訊剪下並貼到其他的應用程式中，例如：報表或資料庫。

取消選取「選項」>「顯示選取拖曳工具」，可以將選取拖曳工具從場景中移除。



4-7 操作資料上方的選取拖曳工具

在第 27 頁的「在 MineSet 3D 可視化工具中瀏覽」中可以找到視窗圖示的說明。請參閱*《MineSet 3.0 企業版參考指南》*中的「分散可視化工具」和「平板可視化工具」以獲得可視化工具選項的補充說明，以及「顏色瀏覽器」以獲得改變顏色的詳細內容。

為分散可視化工具建立滑動桿

當您覺得一個值隨某一欄的值變化時，可以將該欄對照到滑動桿。如果欄是 (整數、浮點、資料點類型) 數值或分組的，就可以對照到滑動桿。如果欄已經分組，則其名稱後就會有 `_bin`。欄類型注釋在「目前欄」清單中的欄名稱之後，例如： total day call - double。在大多數的情況下，「工具管理員」中從欄到滑動桿的對照將自動建立滑動桿。

自動的滑動桿建立

如果在組合步驟中沒有明確地指定滑動桿，「工具管理員」將經由自動分組和組合建立滑動桿。這些自動操作會發生在按一下*調用工具*之後。沒有刪除或對照到可視實體的欄都將用來確定特定實體的數目 (即：它們將稱爲自動組合過程中被分組的欄)。按一下「工具管理員」中的*工具選項*，可以在「工具選項」對話方塊中找到目前滑動桿的索引。

要組合每一個對照到可視元件的數值欄。可以在「工具選項」面板中選擇組合的類型。如果希望對不同的實體進行不同的組合，必須手動建立滑動桿。

手動滑動桿建立

要保証預期的結果，最好用「工具管理員」明確地分組與組合。爲了完成該操作，首先要對計劃對照到可視實體的欄進行組合 (使用平均、計數或總和)，然後按其他的字串和分組欄進行分組，最後按希望成爲動畫滑動桿的分組欄來排列索引。對於所有按滑動桿變數索引的組合欄，結果將具有陣列欄。當滑動桿以此種方式指定時，它們將無法直接對照到滑動桿元件。關於組合的詳細內容，請參閱*《MineSet 3.0 企業版參考指南》*中的「組合」。

如果在目前資料表中有一個陣列欄，「分散可視化工具」中將有一個對應於該欄索引的滑動桿。對於索引，將有一個一維的 X 滑動桿；至於兩個索引，將同時建立 X 和 Y 滑動桿。目前表格中的所有陣列欄都必須具有相同的索引；否則，將不會建立滑動桿。關於建立陣列欄的詳細資訊，請參閱第 3 章中的「透過組合建立新欄」和*《MineSet 3.0 企業版參考指南》*。。

在「分散」和「平板」可視化工具中建立動畫

使用可視化工具主視窗右邊的動畫控制面板建立動畫。只有在對照到滑動桿時，才會出現動畫視窗。例如： *company.scatterviz* 就提供這樣的資料集。請參閱*《MineSet 3.0 企業版參考指南》*中的「動畫」。當您覺得一個值根據特定的準則變化 (如人口密度隨時間變化) 時，可以將該欄對照到滑動桿。*adultJobs.splatviz.* 是一個使用動畫滑動桿的「平板可視化工具」範例。

最有效的方式是爲滑動桿或坐標軸選擇獨立的屬性，爲坐標軸、大小和顏色選擇從屬的屬性。如果只有對照到一個滑動桿 (例如 *adultJobs.splatviz* 或 *adult94.scatterviz*)，則匯總視窗將被壓縮。該滑動桿的維度由它下面的標籤標示。如果資料集沒有對照到滑動桿的欄 (如 *brand.scatterviz*)，將不會出現滑動桿控制項。

使用可視化工具匯總視窗演示動畫

可以在「分散」或「平板」可視化工具動畫控制面板的匯總視窗部分建立動畫。一維或二維中的匯總視窗將根據主視窗中資料的組合顯示一個匯總值的陣列。在匯總滑動桿視窗中，每個黑點都有一個匯總值。匯總顏色指派爲高值，指定給白色的是低值 (內插在這兩種顏色之間的是中間值)。如果對於特定位置上的匯總變數沒有資料，或只有空值，它將顯示爲灰色，表示未知。使用該滑動桿左邊圖中的所有資料之指定組合計算匯總值。可以使用動畫面板上的「顯示資料點」核取方塊來關閉這些黑點。

該範例介紹如何使用準備好的 *company.scatterviz* 檔案建立動畫。

1. 在「工具管理員」功能表列中選擇「可視工具」，再選擇「3D 可視化工具」。
2. 在功能表中，選擇「檔案」>「開啓」功能表。
3. 開啓 *company.scatterviz* 檔案。該檔案在 *MineSet 3.0\examples* 目錄下，位於 MineSet 的初始安裝目錄。

4. 在匯總視窗中，選擇一個黑點作爲動畫路徑的起始點。在該點上按一下並按住滑鼠左鍵，然後在視窗上拖曳游標。鬆開滑鼠左鍵以結束路徑。
   - 也可以在一個點上按一下滑鼠左鍵定義起始點。然後拖曳一個獨立維度的滑動桿，並沿該維度畫一條直線。如果有兩個滑動桿，您也可以使用第二個滑動桿沿第二個坐標軸畫一條直線。
   - 定義路徑的另一個方法是：在一個點上按一下滑鼠左鍵，然後在另一個點上按一下滑鼠中鍵，作出連續的路徑段。該選項僅適用於三鍵滑鼠。
5. 使用匯總視窗下的 VCR 狀按鈕控制動畫。按一下第二列的路徑按鈕前，請先按一下第一列的播放按鈕。關於控制項的詳細內容，請參閱*《MineSet 3.0 企業版參考指南》*中的「動畫」。

在分散可視化工具中顯示動畫軌跡

可以用「分散可視化工具」顯示運動軌跡，以顯示實體不斷改變的動畫路徑。當建立動畫時，軌跡會以選定的形式顯示在每個選定的實體後面。可以從「分散可視化工具」的動畫控制面板中的運動選項功能表 (圖 4-8) 中選擇：

- 沒有軌跡—預設選項
- 線型軌跡—細的彩色線
- 漸弱軌跡—類似於線軌跡的透明彩色線，最近的位置也最不透明
- 管狀軌跡— 3D 管狀軌跡，其厚度隨實體在動畫路徑中運動時大小的變化而變化。管狀軌跡太多會明顯降低動畫速度。

所有軌跡都根據最初的實體進行顏色編碼。當實體改變顏色，如從紅色變爲藍色，則隨著匯總視窗下滑動桿的移動，對應的軌跡也將會在兩個位置之間逐漸改變顏色。軌跡建立在未對照屬性在整個路徑上保持不變的點之間。作爲範例，可以顯示收入隨年份的改變，並留下可視軌跡以顯示兩點之間的變化速度。

按照少數欄分組的組合資料可能非常適合用於顯示運動軌跡。最初，將顯示分散圖中受路徑影響的所有點的運動軌跡。按一下任何實體來選取，只讓選取的點顯示軌跡。這可以用來減少視覺混亂。具有空位置的實體在軌跡中表現爲中斷。

圖 4-8 所示爲具有管狀運動軌跡的「分散可視化工具」範例。

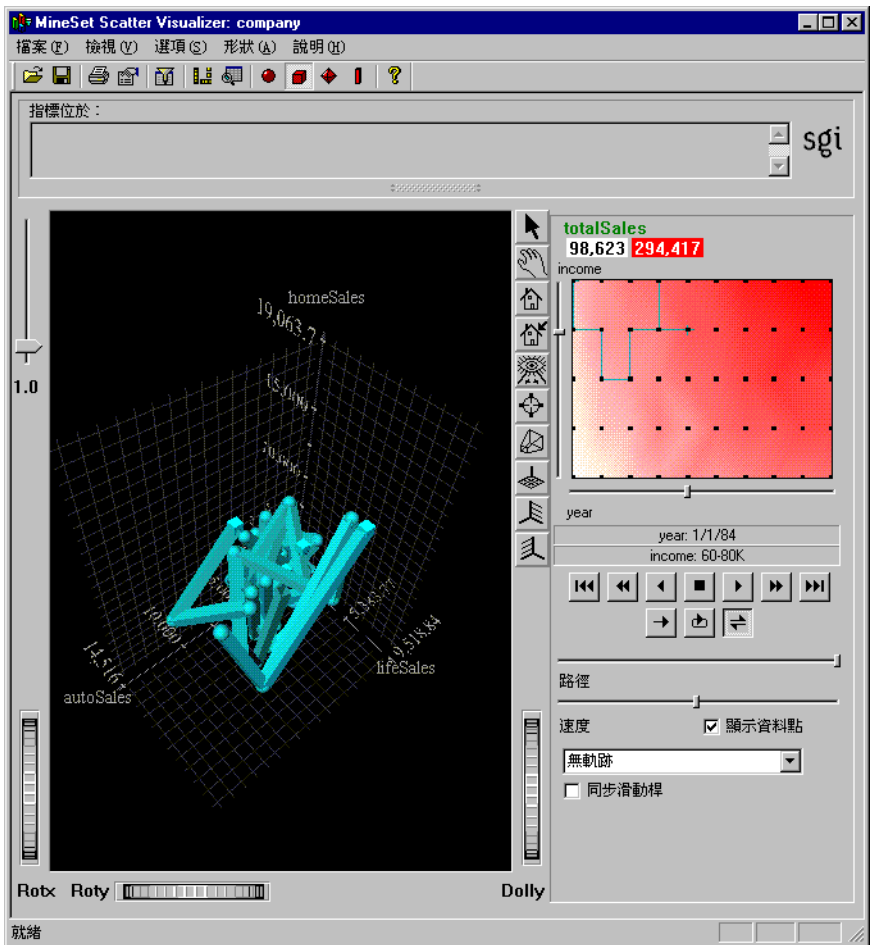


圖 4-8 分散可視化工具管狀運動軌跡範例

在平板可視化工具中演示動畫

從本質上看，在「平板可視化工具」中演示動畫與在「分散可視化工具」中是一樣的，除了「平板可視化工具」沒有管狀軌跡選項以外。關於 *adultJobs.splatviz* 的範例顯示在圖 4-9 中。

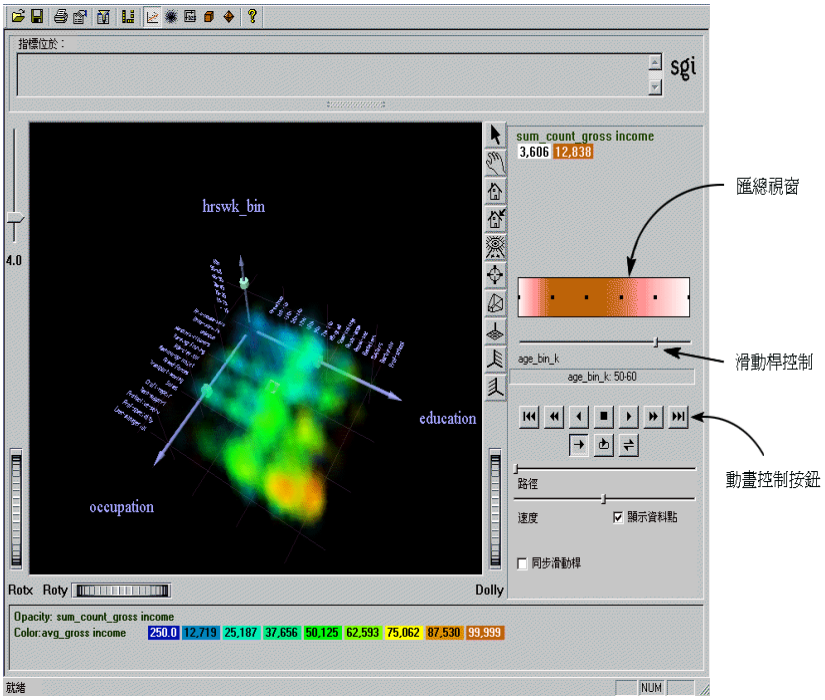


圖 4-9 具有匯總視窗和滑動桿控制的平板可視化工具動畫控制面板

如果配置檔案沒有指定滑動桿對照，將不會出現滑動桿控制項。

平板可視化工具中匯總視窗說明

匯總視窗 (請參閱 圖 4-9) 中顯示滑動桿維度中匯總屬性的組合。主視窗中的雲團狀平板代表匯總值的顏色密度。如果沒有欄顯式對照到匯總，則計數將用於顯示滑動桿上代表資料最多的位置。

在圖 4-9 所示的範例中，匯總滑動桿白色區域中的黑色圓點表示：當滑動桿處於該位置時，左邊的平板圖是由 3,606 個記錄組成的。滑動桿上最紅點處的圖擁有 12,838 個記錄。

平均間隔的黑色圓點指出離散資料點的準確位置。可以取消對動畫面板底部「顯示資料點」方塊的核取以隱藏這些黑點。這些位置之間的滑動桿位置使用基本資料的內插來產生影像。

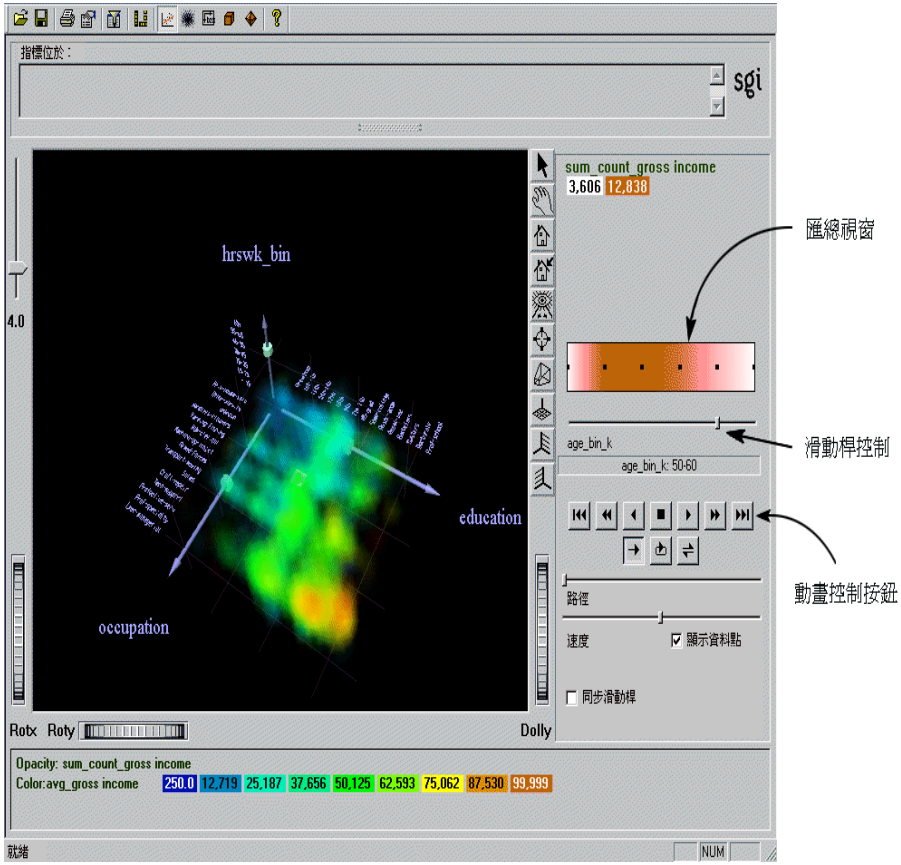
匯總視窗中的顏色密度

使用 *adultJobs.splatviz* 檔案可以解釋顏色密度對值的反映方式。匯總視窗顯示的顏色範圍從白色 (左) 到紅色 (中) 再到白色 (右)。紅色代表資料較多，而白色代表資料較少。在本範例中，滑動桿中間較高密度的紅色表明人口最密的範圍是 20-50 歲。

在匯總視窗中建立路徑

建立路徑的方式與在「分散可視化工具」中使用的完全一樣，並且控制項也相同。但是，當動畫演示時，雲團的大小和顏色平滑地變化是毫無意義的。狀態欄位中的資訊將顯示內插的資料值。當滑動桿停止移動時，滑動桿位置將繼續向前，直到最近的、沒有使用內插資料值的離散資料點。

匯總滑動桿上的每個分組位置都有一個表。這些表的每一列 (原始資料的組合) 定義場景中的一個雲團狀「平板」。匯總上對應於鄰近分組的表的列數不必相同，因爲位置之間的資料分布是不同的。例如：如果第 55 頁上圖 4-2 中的可視化處理將滑動桿缺口移到右邊 (請參閱圖 4-10) 而從顯示 40-50 歲變爲顯示 50-60 歲 ,則一些位置就會顯示以前未顯示的雲團，反之亦然。



4-10 滑動桿移動後改變的可視化處理 (請與圖 4-2 比較)

對於一維滑動桿上的內插，將會合併兩個鄰近的表，然後使用空間欄作爲唯一的關鍵字對其進行組合。權重是通過簡單內插得到的 (如果表格中缺少特定的列，將假定權重爲 0)。用於顏色的平均值也是內插得到的，但是由權重進行加權。

關於內插技術細節的範例，請參閱《*MineSet 3.0 企業版參考指南*》中的「平板可視化工具」。

處理分散和平板可視化結果

可以經由五個下拉式功能表得到「分散」和「平板」可視化工具的功能。它們是「檔案」、「檢視」、「選項」、「形狀」、和「說明」。如果在沒有指定配置檔案的情況下執行可視化工具，則只有「檔案」和「說明」功能表可以使用。一旦開啓檔案，就可使用其他功能表。「檔案」和「說明」功能表是標準的功能表，關於詳細的資訊，請參閱*《MineSet 3.0 企業版參考指南》*。

更改畫面

透過「檢視」功能表可以控制可視化工具視窗中顯示的特定部分。圖 4-11 所示爲「分散」和「平板」可視化工具的「檢視」功能表。

圖 4-11 分散和平板可視化工具的檢視功能表

下面表格將說明這些選項：

表 4-5 分散和平板可視化工具的檢視功能表選項

選項	描述
「篩選工具」面板	開啓「篩選」面板 (圖 4-12) 以處理畫面內容
設定背景顏色	開啓顏色選擇器以指定新的畫面背景顏色
視窗控件	顯示或隱藏主視窗周圍的外部控制項
動畫面板	顯示或隱藏所有的動畫控制面板
空位置	顯示或隱藏坐標軸上具有空位置或未知位置的實體

4-12 分散和平板可視化工具篩選面板

如果希望根據一定的準則篩選顯示的資料量，可以使用「篩選」對話方塊，如圖 4-12 所示。在「檢視」功能表中選擇「篩選」面板。可以使用上面的窗格根據字串欄進行篩選。要選擇一個值，可以按一下它。要取消對一個值的選擇，只需再次按一下它。要選擇欄中所有的值，可以按一下*全部設定*按鈕。要清除目前所有選擇，可以按一下*清除*按鈕。

可以使用底部的窗格同時根據字串欄和數值欄的值進行篩選。關於詳細資訊，請參閱*《MineSet 3.0 企業版參考指南》*中的「篩選面板」。

利用下方的「僅顯示篩選集」，可以指定主視窗中的景觀涵蓋整個資料集的範圍，還是只包括篩選後的資料。

按一下*應用*按鈕，開始篩選。如果在面板處於活動狀態時按下 Enter 鍵，則篩選將自動開始。

在分散可視化工具中選擇和追溯

可以使用「分散可視化工具選項」功能表追溯基本資料。

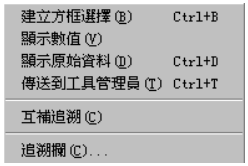


圖 4-13 分散可視化工具選項功能表

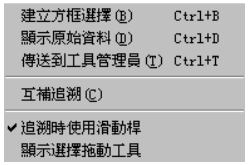
- *建立方塊選擇器作業*建立一個 3-D 的方塊選擇器，可以拉伸並移動它以選擇體積區域。當它處於活動狀態時，「記錄檢視器」表格將開啓，顯示內部實體組合資料的相關資訊。請確定游標處於「選擇」模式。要移動選擇方塊，可以用滑鼠左鍵按一下其中一面並沿需要的方向拖曳。拖曳時按住 Shift 鍵可以將運動限制在離拖曳最近的坐標軸上。要更改選擇方塊的大小，可以沿需要的方向拖曳灰度比例索引標籤。不能在體積界限之外調整大小或移動。灰度比例索引標籤將不斷調整大小以保持固定的螢幕大小。如果覺得它們顯得太大，可以將其縮小，它們將會減小相對於方塊的大小。
- *顯示值*可以顯示「記錄檢視器」表中選定實體的值。
- *顯示原始資料*可以檢索並顯示對應於已選定實體的記錄。產生的記錄顯示在「記錄檢視器」表中。如果沒有選擇任何實體，則該功能表項目爲灰色。
- *發送到工具管理員*將根據目前選擇，在「工具管理員」歷史表的開始處插入一個篩選操作。用於追溯的實際表達式由目前的選項決定。如果沒有選擇任何實體，則該功能表項目爲灰色。
- 使用*互補追溯*可以使*顯示原始資料*和*傳送到工具管理員*選項讀取所有*未*選取的資料。

- *追溯欄*將開啓一個面板，可以用來選擇用於追溯的欄。與其他可視工具不同，資料中沒有特定的欄指定爲資料的關鍵字「分散可視化工具」不可能確定使用者在追溯表達式中需要的欄。例如：資料可能是汽車的商標、型號和重量。可能需要追溯原始資料，指定應該考慮商標和型號，而不考慮重量。預設情況下，對照到圖形要求的所有欄在追溯中都視爲有意義的。其他的列則不是這樣。但是可以在「喜好設定」對話方塊中反白標示它們來達到此目的。

關於進一步的詳細內容，請參閱*《MineSet 3.0 企業版參考指南》*中的「追溯」。

在平板可視化工具中選擇和追溯

「平板可視化工具選項」功能表與「分散可視化工具」在用於追溯基本資料的選項方面稍有不同。



4-14 平板可視化工具選項功能表

- *建立方塊選擇作業*建立一個 3-D 的方塊選擇器，可以拉伸並移動它以選擇體積區域。當它處於活動狀態時，「記錄檢視器」表格將開啓，其中顯示內部實體代表的所有組合資料的相關資訊。關閉該視窗將清除目前所有選擇。所有選擇方塊內的實體或使用 shift 按一下選擇的實體將顯示在表格視窗中。要移動選擇方塊，可以用滑鼠左鍵按一下其中一面並沿需要的方向拖曳。拖曳時按住 Shift 鍵可以將運動限制在離拖曳最近的坐標軸上。要更改選擇方塊的大小，可以沿需要的方向拖曳灰度比例索引標籤。不允許在體積界限之外調整大小或移動。灰度比例索引標籤將不斷調整大小以保持固定的螢幕大小。如果覺得它們顯得太大，可以將其縮小，它們將會減小相對於方塊的大小。

- *顯示原始資料*可以檢索並顯示對應於「方塊選擇」選定的記錄。產生的記錄顯示在「記錄檢視器」表中。
- *傳送到工具管理員*將根據目前方塊選擇在「工具管理員」歷史的開始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確定。如果沒有選擇任何實體，將出現一個警告。
- 使用*互補追溯*可以使*顯示原始資料*和*傳送到工具管理員*選擇讀取所有*未*選取的資料。
- *在追溯中使用滑動桿*作用是在追溯時使用滑動桿定位。
- *顯示選取拖曳工具*可以切換選取拖曳工具的可見性 (預設情況下爲開啓)。當方塊選擇開始時，將移除選取拖曳工具，但是它可在方塊選擇處於活動狀態的同時開啓。

關於進一步的詳細內容，請參閱*《MineSet 3.0 企業版參考指南》*中的「追溯」。

用「形狀」功能表更改分散可視化工具畫面

可以使用「形狀」功能表更改「分散可視化工具」中繪製分散的方法。以下的不透明圖元可用來代表分散。

- *立方體*將繪製一個不透明的立方體，它的體積與對照到大小可視元件的屬性成正比。
- *菱形*將繪製一個線方塊三角形。它的方向隨其顏色變化，並且它的大小與對照到大小可視元件的屬性成正比。
- *球體*將繪製一個不透明的球體，它的體積與對照到大小可視元件的屬性值成正比。
- *長條*將繪製拉長的直條，它的高度由對照到大小可視元件的屬性值決定。

用「形狀」功能表更改平板可視化工具畫面

可以使用「平板可視化工具形狀」功能表更改繪製平板的方法。可以選擇交換交互性的準確度。在所有相似方法中，紋理是理想 Gaussian 密度分布最準確的表示法。因爲大部分的電腦都支援硬體輔助結構，因此紋理通常是最好的選擇。三種平板類型是：

- *線性*將繪製與高斯型平板線性近似的一小組三角形。
- *高斯型* 將繪製與高斯型平板近似的一大組三角形。
- *紋理*使用紋理對照三角形繪製最準確的表示法。在不支援硬體輔助結構對照的系統上，該過程可能會非常慢。

另外也可選擇以下的不透明圖形元件代表平板：

- *立方體*將繪製一個不透明的立方體，它的體積與對照到不透明度可視元件的屬性 (如果有的話) 或計數成正比。
- *菱形*將繪製一個線方塊三角形。它的方向隨其顏色變化，並且它的大小與對照到不透明可視元件的屬性 (如果有的話) 或計數成正比。
- *球體*將繪製一個不透明的球體，它的體積與對照到不透明度可視元件的屬性 (如果有的話) 或計數成正比。

用樹狀可視化工具檢視資料

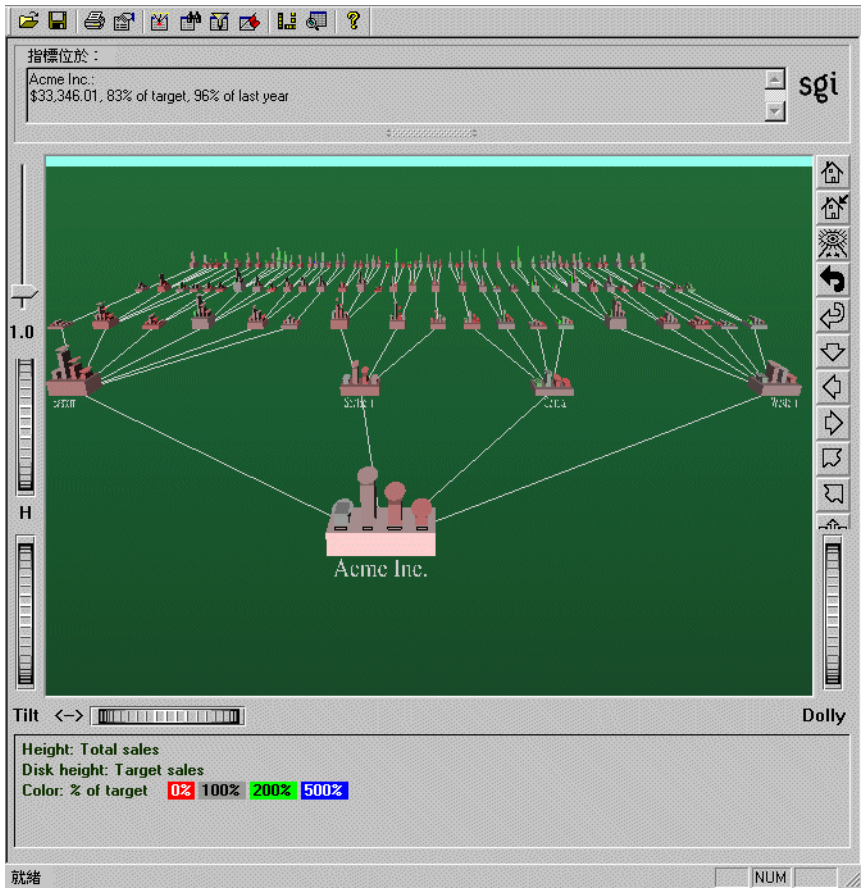
本章包括以下幾個部分：

提供一些樣本配置和資料檔案，用於演示「樹狀可視化工具」的特性和功能。這些檔案在 *MineSet 3.0* 下的 *examples* 目錄中，位於 MineSet 的初始安裝目錄。

樹狀可視化工具總覽

「樹狀可視化工具」是一個圖形介面，它將資料顯示爲 3D 景觀。它將資料以樹狀的形式分層顯示。樹的每一層根據不同屬性的值進行分支。每個樹節點都顯示一個代表它所有子樹資料的圖形。圖形表由基準塊組成，基準塊的高度和顏色取決於資料屬性。每個基準長條方塊和 / 或盤的編號、標籤、高度和顏色也由指定的屬性來確定。

如圖 5-1 所示，「樹狀可視化工具」將資料的定量特徵和相關特徵顯示爲分層連接的節點。每個節點由長條方塊和圓盤組成，其高度和顏色對應於資料值的組合 (通常爲總和、平均或計數)。連接節點的線稱爲邊，它顯示資料集與其子集的關係。



5-1 樹狀可視化工具主視窗的範例顯示

執行樹狀可視化工具

本部分說明如何使用「工具管理員」配置「樹狀可視化工具」。雖然「工具管理員」大大簡化了「樹狀可視化工具」的配置任務，但是您也可使用編輯器手動建立配置檔案。還有其他多種執行「樹狀可視化工具」的方法，關於詳細內容，請參閱*《MineSet 3.0 企業版參考指南》*。

要從「工具管理員」中執行「樹狀可視化工具」：

1. 選擇「檔案」>「連接到伺服器」，然後登錄到伺服器。
2. 選擇「檔案」>「開啓新的資料檔案」，並選擇或鍵入需要的檔名。
3. 在「資料目標」面板中，按一下「可視化 工具」索引標籤 (圖 5-2)。
4. 從工具索引標籤中，選擇「樹狀可視化工具」索引標籤。
5. 用「可視元件」旁的下拉式功能表將指定屬性從資料集對照到「資料目標」窗格中適當的「可視元件」。關於對照的詳細資訊，請參閱第 84 頁的「對照樹狀可視化工具的可視元件」。
6. 按一下「工具選項」按鈕，在「樹狀可視化工具配置選項」面板中指定「高度和顏色組合」。這一步對於獲得有意義的可視化處理極爲重要。

   預設情況下，父節點長條的高度 (或顏色) 代表所有子節點長條值的總和；但是，對於某些資料集，平均值、最大值、最小值或計數可能是更好的選擇。組合可以指定爲長條、基準塊和圓盤的高度及顏色。關於組合和其他內容的詳細資訊以及詳細的進階選項，請參閱*《MineSet 3.0 企業版參考指南》*。
7. 按一下*調用工具*，執行可視化工具。

「工具管理員」視窗底部的「狀態」方塊將顯示進程與統計結果。

對照樹狀可視化工具的可視元件

當選取「樹狀可視化工具」時，「資料目標」面板將顯示一些可視元件，可以將資料中的屬性對照到這些可視元件 (圖 5-2)。從每個可視元件旁的下拉功能表中選擇欄。表 5-1 列出可視元件以及其用法說明。通常，使用「樹狀可視化工具」的最好方法是用元件對照來進行試驗，直到找到屬性和可視元件的最好組合。

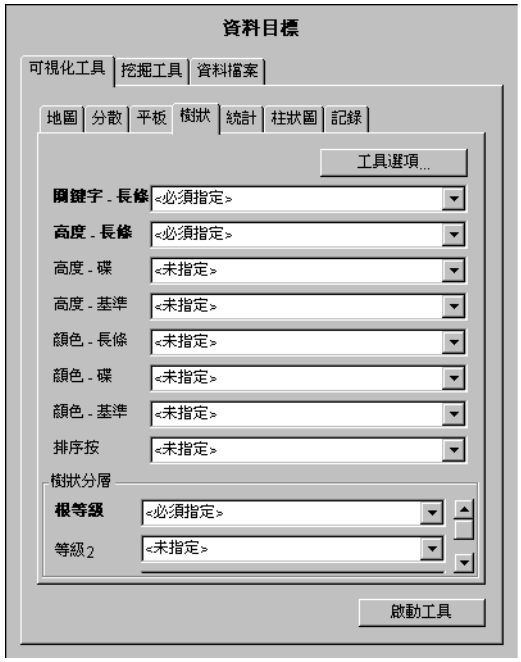


圖 5-2 選定樹可視化工具的工具管理員資料目標窗格

表 5-1 列出可視元件。

表 5-1 樹狀可視化工具可視元件

可視元件	註解
關鍵字 - 長條	指定長條方塊代表的值。如果欄是數值型，通常最好對欄進行分組。否則，MineSet 將爲每個屬性值建立一個長條方塊，讓可視化處理難以使用。
高度 - 長條	指定長條方塊高度代表的值。通常，長條方塊越高，代表的值就越大。查看「工具選項」面板中關於每級組合屬性的選項。
高度 - 圓盤	指定圓盤高度代表的值。可選的圓盤與長條方塊放在同一個位置。如果圓盤屬性的測量單位與長條方塊高度屬性的一樣 (例如：今年的銷售額和去年的銷售額)，則圓盤最有用。如果沒有指定對照，則不會顯示任何圓盤。
高度 - 基準	指定基準塊高度代表的值。如果沒有指定對照，缺少使用長條高度對照。查看「工具選項」面板中關於每級組合屬性的選項。
顏色 - 長條	指定長條方塊顏色代表的值。可以透過「工具選項」面板分配指定的顏色 (請參閱*《MineSet 3.0 企業版參考指南》*中的「顏色選擇」)，或允許 MineSet 自動分配顏色。查看工具選項面板中關於每級組合屬性的選項。
顏色 - 圓盤	指定圓盤顏色代表的值。可以透過「工具選項」面板分配指定的顏色 (請參閱*《MineSet 3.0 企業版參考指南》*中的「顏色選擇」)，或允許 MineSet 自動分配顏色。 只有在指定圓盤高度時，該選項才有效。

表 5-1 樹狀可視化工具可視元件

可視元件	註解
顏色 - 基準	指定基準塊顏色代表的值。可以透過「工具選項」面板分配指定的顏色 (請參閱 *《MineSet 3.0 企業版參考指南》* 中的「顏色選擇」) ,或允許 MineSet 自動分配顏色。查看工具選項面板中關於每級組合屬性的選項。 如果沒有指定對照，就會使用長條圖顏色對照。
按 ... 排序	按照選取屬性的值排列節點的配置。預設情況下，排列順序爲從左到右升序。
樹分層 根級別	指定如何將來自資料來源的表格轉換爲分層結構。樹的第二級具有的節點數與選定屬性值的數目一樣多。 如果屬性的可能值有很多，樹將會很大且難於分析。對欄進行分組可以減輕這種情況。
樹分層 等級 2, 3, ...	將資料劃分到更多的分層結構中。預設情況下，「可視元件」清單具有三個等級。如果指定了第三個等級，「樹狀可視化工具」會自動增加第四個等級。每指定一個額外等級的同時會增加另一個。可以根據需要指定任意多的等級。 如果屬性的可能值有很多，樹將會很大且難於分析。對欄進行分組可以減輕這種情況。

用樹狀可視化工具檢視結果

當「樹狀可視化工具」第一次啓動時，分層的根節點位於場景的前面，在主視窗底部附近。根節點的背後是它的子代。每個節點都由帶有長條方塊的基準組成。可以透過「工具管理員」更改長條方塊高度和顏色代表的值，或者手動更改 *.treeviz* 配置檔案 (請參閱 *《MineSet 3.0 企業版介面指南》* 中的「樹狀可視化工具」) ；通常基準塊代表所有長條方塊的組合。基準塊用邊連接起來，這些邊代表節點與其子代的連接。

子群組中的值可以自動在相鄰的高等級中求和與顯示。長條方塊下的基準塊可以提供有關所有長條方塊組合值的資訊。代表負值的長條方塊顯示在基準塊的頂部下面。可以經由取消基準高度面來更清晰地查看負值長條 (請參閱第 97 頁的「調整樹狀可視化工具畫面」，或*《MineSet 3.0 企業版參考指南》*中的「樹狀可視化工具」)。

樹狀可視化工具範例

圖 5-3 顯示的是開啓 stores.treeviz 檔案的場景。該檔案在 *MineSet 3.0\examples* 目錄中，位於 MineSet 的初始安裝目錄下。要想看到它，請從工具管理員「可視工具」功能表中，開啓「三維可視化工具」；然後在可視化工具「檔案」功能表中，選擇「開啓」。資料檔案中的每個記錄都包含以下欄：地區 (東部、南部、中部或西部)、州、城市、商店數目、產品類型 (用具、服裝、電器或家具)、今年的銷售額、去年的銷售額、以及目標銷售額。還有兩個新增加的欄：今年達到的銷售額占目標銷售額的百分比，和今年達到的銷售額占去年銷售額的百分比。表 5-2 顯示的是欄到「樹狀可視化工具」可視元件的對照。

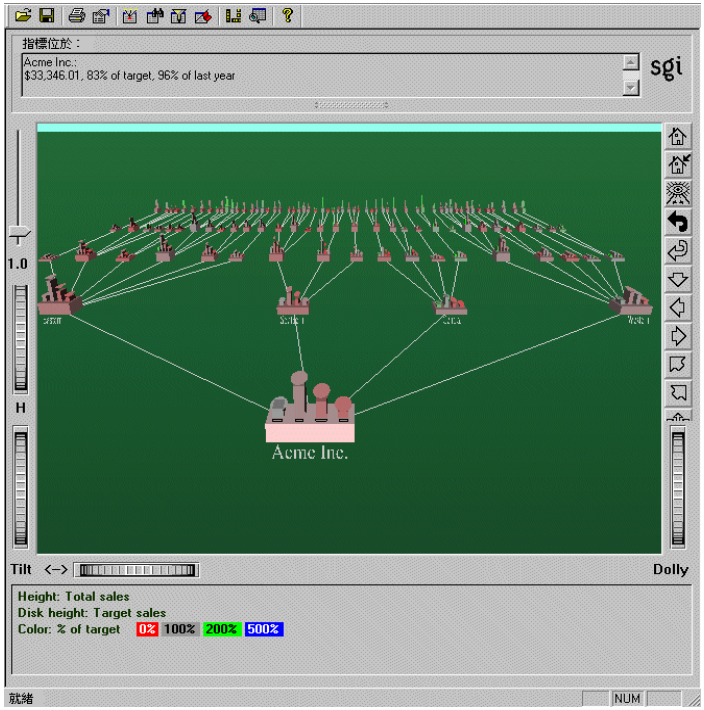


圖 5-3 「商店」資料集的樹狀可視化工具初始檢視

樹頂部節點中的長條方塊代表每種產品的總銷售額。圓盤代表目標銷售額。下一級將銷售額按地區分開，再下一級按州分開，等等。直到單個的商店。

在檢查「樹狀可視化工具」場景時，可以很容易地看到每個地區、州、城市或商店對於每種商品之目標銷售額的符合程度。也可以按照地區、州、產品類型等等來比較銷售額。

表 5-2 顯示該範例中對照到可視元件的欄。

表 5-2 「商店」資料的元件對照

欄	組合	對照
地區	無	等級 1
州	無	等級 2

表 5-2 「商店」資料的元件對照

欄	組合	對照
城市	無	等級 3
商店 ID	無	等級 4
產品類型	無	關鍵字 - 長條
今年的銷售額	總和	長條高度
去年的銷售額	總和	無
目標銷售額	總和	圓盤高度
目標百分比	無	長條方塊顏色
去年的銷售額百分比	無	無

對照是獲得有意義的可視化處理的關鍵。在熟悉該工具以前，實驗是尋找最佳對照模式的最好方法。

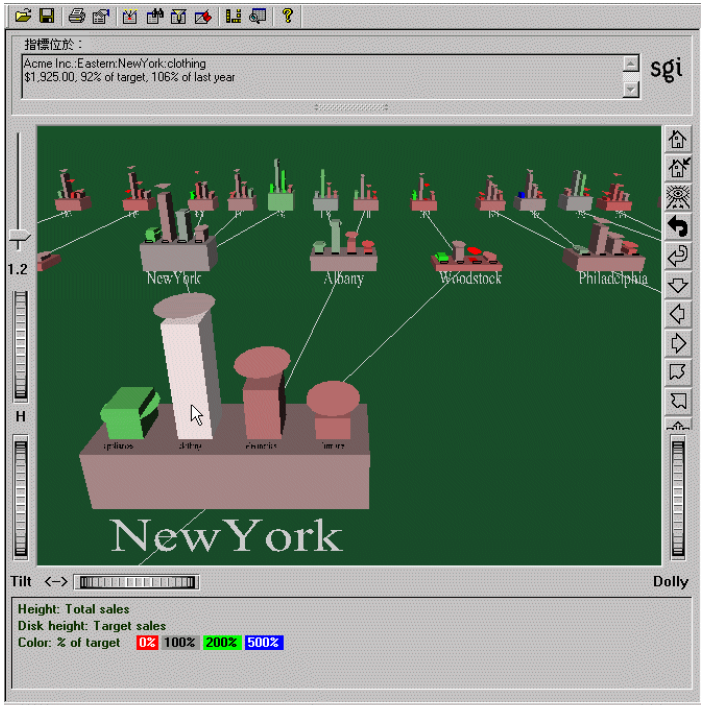
在樹狀可視化工具中更仔細地觀察資料

有兩種方法可以查看每個節點的細節：

- 反白標示節點或物件 (長條方塊或基準塊)，可以查看基本資料。(請參閱「反白標示物件或節點」)。
- 選擇節點或物件，將觀察點縮放到該位置，可以查看基本資料。(請參閱「選擇物件」)。可以一次選擇多個位置。選擇一個節點或物件也可以反白標示它。聚光燈可以持續追蹤該位置，即使它在場景的遠處。

反白標示物件或節點

要反白標示一個物件 (長條方塊或基準塊)，可以將滑鼠放在該物件上面。關於該物件的資訊將出現在檢視區域的左上部，「指針位於」標籤的下方。(圖 5-4)。要反白標示一個節點並取得關於該節點的資訊，可以將指標放在通向該節點的線上。



5-4 反白顯示物件及其基本資訊

選擇物件

要選擇一個物件並將鏡頭移動到它的位置，可以按一下該物件。按一下的同時按住 Shift 鍵則選擇物件而不直接縮放它。只要選定物件，就會顯示資訊。

如果按一下物件的同時也按住 Ctrl 鍵，就可以切換對該物件的選擇。如果目前沒有選定該物件，按一下選擇它，反之亦然。使用該技術可以同時選擇多個物件。當「指標位於」欄位中只顯示最後一個所選物件的資訊，可以選擇「選項」>「顯示數值」或追溯到所選物件後面的原始資料，進而看到所有被選的值。

當選擇物件時，在它的上面會出現一個白色的聚光燈 (圖 5-5)。在搜尋時，會出現一個黃色的聚光燈 (請參閱第 93 頁的「用搜尋面板尋找指定的物件」)。聚光燈是可見的，即使選定的物件是遠處背景中的一個子節點。

聚光燈的邊是物件的代用品：當移動指標到聚光燈的邊上時，就會反白標示相關物件，相關的資訊會出現在檢視的左上角。按一下聚光燈的邊，選擇相關的物件並 (如果沒有按下 Shift 鍵) 直接縮放它。聚光燈只有在邊上實線部分是活動的，而在中央透明部分是非活動的。在透明區域按一下可以選擇聚光燈後面的物件。

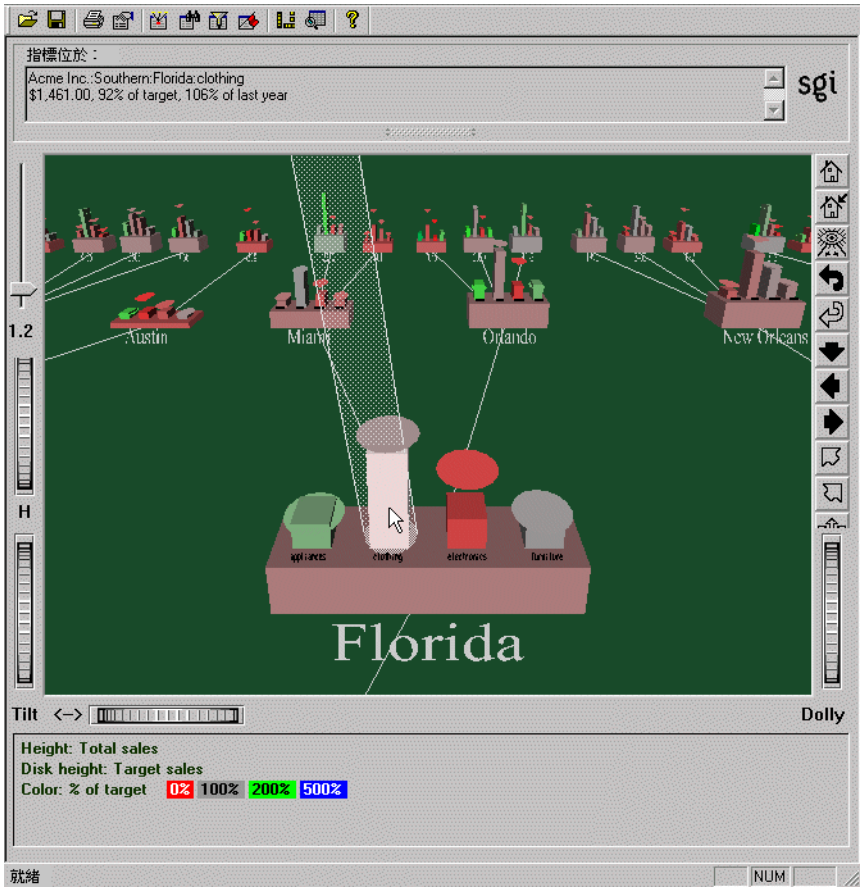


圖 5-5 選取 (反白標示的) 物件的範例

用全覽視窗觀察整個圖片

要開啓帶有整個分層的俯視圖視窗 (圖 5-6)，可以選擇「檢視」>「全覽檢視」。如果需要在每次檢視場景時顯示「全覽」，請設定配置檔案中的「全覽」選項 (請參閱 *《MineSet 3.0 企業版介面指南》* 中的「全覽」一節)。

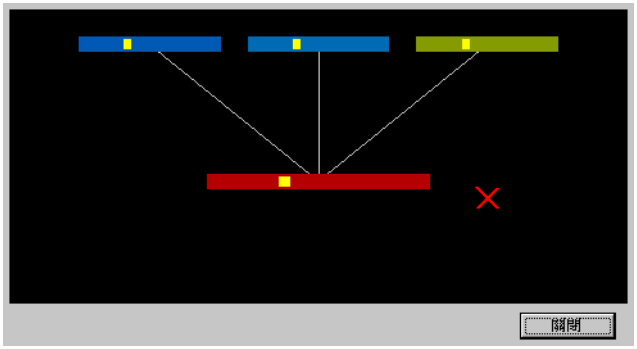


圖 5-6 樹狀可視化工具的全覽視窗

「全覽」視窗中的「X」表示目前的位置。全覽檢視可以幫助您追蹤在整個場景中的位置和觀察點。它還有助於迅速到達指定的位置：

- 要在全覽檢視中選擇節點、讓主檢視直接縮放它並反白標示它，可以按一下該節點。這類似在主視窗中按一下節點。
- 要將觀察直接縮放到新的位置，即使在該位置沒有節點存在，也可以同時在該位置按一下滑鼠的左右鍵 (或滑鼠中鍵)。

用搜尋面板尋找指定的物件

要在「樹狀可視化工具」視窗中搜尋指定的物件，可以選擇「檢視」>「搜尋」面板。填入搜尋規則，然後按一下「搜尋」。關於「搜尋」面板的詳細資訊，請參閱*《MineSet 3.0 企業版參考指南》*。

一旦搜尋完成，黃色聚光燈將反白顯示符合搜尋規則的物件 (請參閱圖 5-7)。要顯示黃色聚光燈下面物件的有關資訊，可以將指標移到該聚光燈上；資訊將出現在左上角，「指標位於」標籤下。要選擇並直接縮放到黃色聚光燈下的物件，可以按一下聚光燈；如果按一下的同時也按住 Shift 鍵，將不會進行縮放。

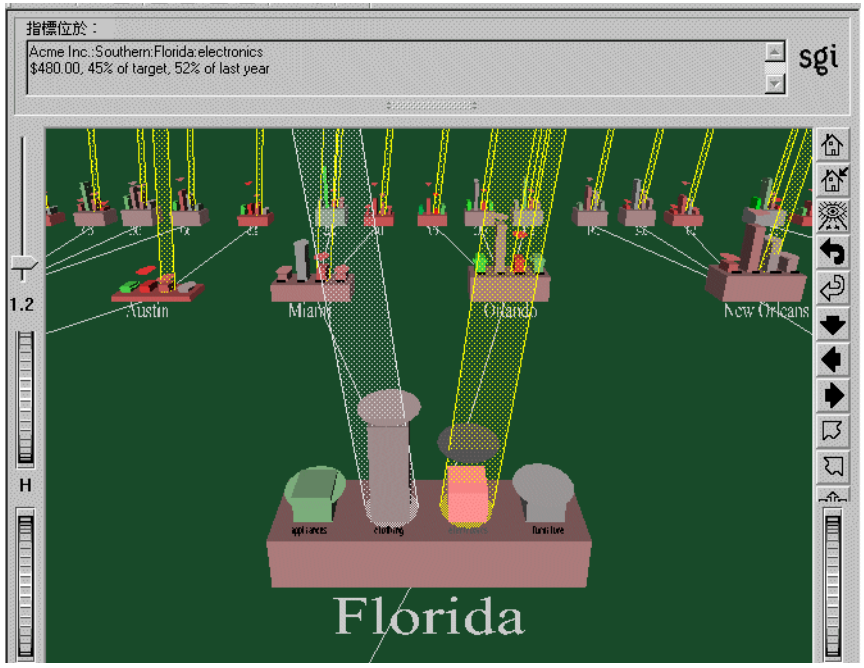


圖 5-7 樹狀可視化工具中的搜尋結果示例

用「標號」面板標記重要的位置

「標號」面板可以命名並儲存重要的位置 (或觀察點) 以備日後參考。所有標號都可以用主要檢視中的彩色旗幟顯示。如果標號代表選定的物件，就將旗幟放在該物件上。如果它代表觀察點的位置，就將旗幟放在該位置。要到標號，可以按一下旗幟。可以選擇「顯示」功能表中的「標號旗幟」開啓和關閉所有旗幟 (請參閱 *標號旗幟*，在第 97 頁的「調整樹狀可視化工具畫面」小節內)。

要儲存位置並用旗幟標示它 (可選) ：

1. 在樹狀可視化工具「檢視」的功能表中，選擇「標號面板」。將開啓一個視窗，如圖 5-8 所示。如果還沒有放置任何標號，清單將是空的。

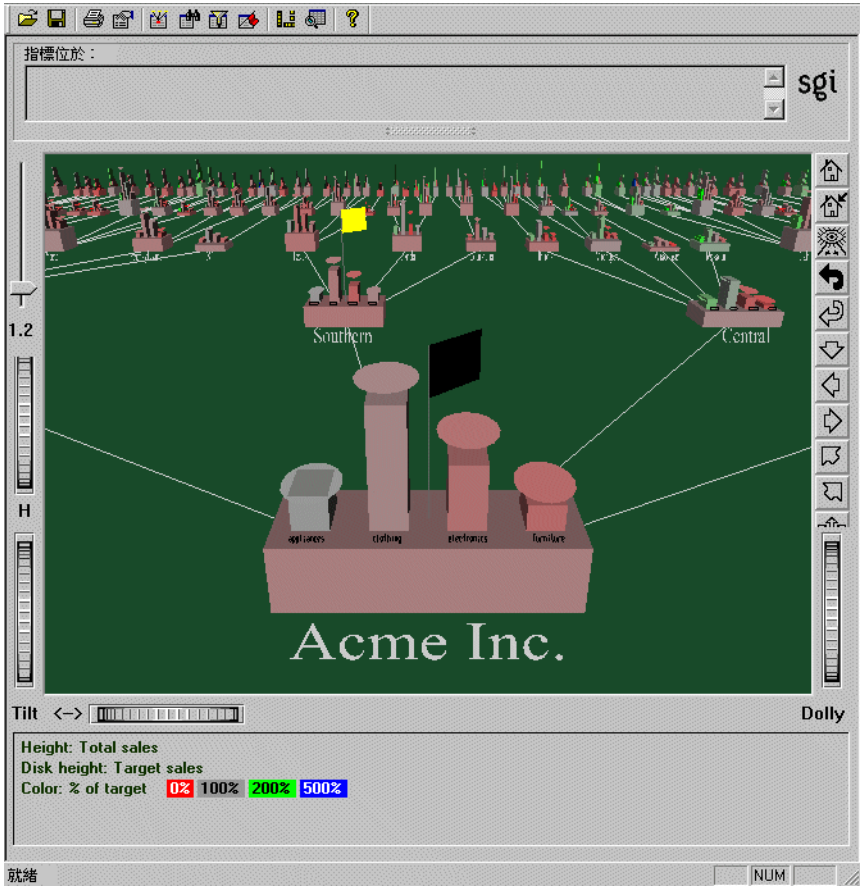
圖 5-8 樹狀可視化工具「標號面板」

2. 按一下*標號*按鈕，標示在主要視窗的目前位置。將顯示一個對話方塊，如圖 5-9 所示。

3. 輸入希望用於標號的名稱，並選擇一種顏色。預設名稱爲目前選定物件的名稱。選擇的顏色將控制出現在主視窗中的旗幟顏色，並且代表該標號。

圖 5-9 「選擇標號」對話方塊

圖 5-10 所示爲一個主視窗樣例，其中的旗幟代表建立的標號。



5-10 帶有標號旗幟的主視窗

「標號」視窗邊上的按鈕可以執行以下任務：

- 要轉到面板中選取標號代表的位置，可以按一下「標號」面板中的*跳到*按鈕。連按兩下面板中的標號可以產生同樣的效果。如果標號選定的物件已不存在 (因爲它已被篩選掉，或自標號建立以來資料發生過更改)，則顯示的位置將在原來物件位置的附近。
- 要刪除面板中選定的標號，可以按一下*刪除*按鈕。

- 要更改面板中選定標號的名稱或顏色，可以按一下*修改*按鈕。
- 要將面板中選定的標號按清單順序向上移動，可以按一下*向上*按鈕。
- 要將面板中選定的標號按清單順序向下移動，可以按一下*向下*按鈕。
- 要結束「標號」面板，可以按一下*關閉*按鈕。

儲存標號資訊的檔案與配置檔案具有相同的名稱，結尾爲 *.marks*。只要更改標號，就會儲存到該檔案中。如果刪除了所有標號，*.marks* 檔案也將移除。如果標號更改無法儲存 (例如：由於權限錯誤)，就會出現一個警告：此後再進行標號更改嘗試時，該警告將不再重複出現。

用「篩選」面板篩選資料

「篩選」面板將移除選取的資訊，因此可以調整顯示層次。您可以使用「篩選」面板強調指定的資訊，或壓縮資料量以獲得更好的性能。若要進入「篩選」面板，可從「樹狀可視化工具檢視」功能表中選擇。關於使用「篩選」面板的詳細資訊，請參閱*《MineSet 3.0 企業版參考指南》*中的「樹狀可視化工具」。

調整樹狀可視化工具畫面

表 5-3 描述用於調整「樹狀可視化工具」畫面的多個選項。這些樹狀可視化工具「顯示」功能表中的選項可以控制一些顯示參數。

表 5-3 樹狀可視化工具顯示參數

選項	描述
基準高度 (開關)	打開或關閉基準高度。要查看負數，或想讓方塊高度互相比較變得更容易，可以關閉該選項。打開該選項可以提供所有長條的匯總資訊。該切換開關的初始值可以用配置檔案中的語句「base height」來更改。
標號旗幟 (切換)	開啓和關閉代表標號的標記 (請參閱「用「標號」面板標記重要的位置」)。

表 5-3 樹狀可視化工具顯示參數

選項	描述
零 (子功能表)	*控制*零高度物件的顯示方式。預設情況下，它們的顯示與其他物件類似：高度爲零的實心立方體 (平面)。子功能表可以指定將它們顯示爲輪廓 (空心的正方形)，或完全隱藏起來 (不畫)。該切換開關的初始值可以用配置檔案中的選項「zero」來更改 (請參閱《*MineSet 3.0 企業版參考指南*》)。
空 (子功能表)	*控制*空高度物件的顯示方式。它的選項與零功能表中的一樣；但是空選項的預設值是將物件顯示爲一個輪廓。初始值可以用配置檔案中的選項「null」來更改 (請參閱《*MineSet 3.0 企業版參考指南*》)。

用地圖可視化工具檢查資料

「地圖可視化工具」在 3D 景觀上用長條圖或可識別的地理形狀來顯示資料。當資料具有地理背景或一些特定的地形配置時，該工具很有用。本章主要包括以下幾個主題：

提供一些配置檔案和資料檔案樣例，用於示範「地圖可視化工具」的特性和功能。這些檔案儲存在 *MineSet 3.0\examples* 目錄中，位於 MineSet 的初始安裝目錄。

地圖可視化工具總覽

「地圖可視化工具」可以用來觀察與空間相關的資料，您可以使用「地圖可視化工具」在三維景觀中瀏覽，可以對資料進行概化上尋或細部下尋，可以全覽資料或進一步檢視局部細節，還可以使用動畫來了解資料在一維或二維獨立變量空間中的變化方式。這裡所謂獨立的維度是指任意像年齡或年份這樣與其它欄無關的屬性。當資料集中含有對照到滑動桿的獨立變量時，動畫面板會出現在主視窗右側 (如 圖 6-1 所示。)

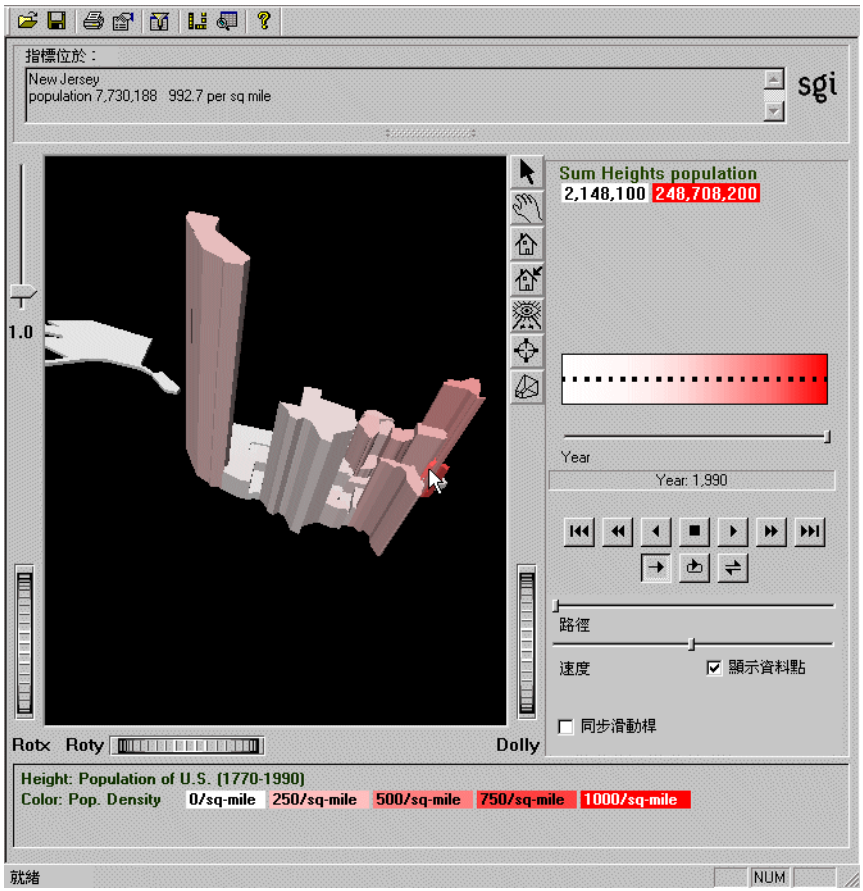


圖 6-1 使用地理形狀的地圖可視化工具樣例

景觀還可以由描繪成簡單輪廓的地理物件組成，在特定的位置放有「長條圖」圓柱體 (請參閱圖 6-2)。

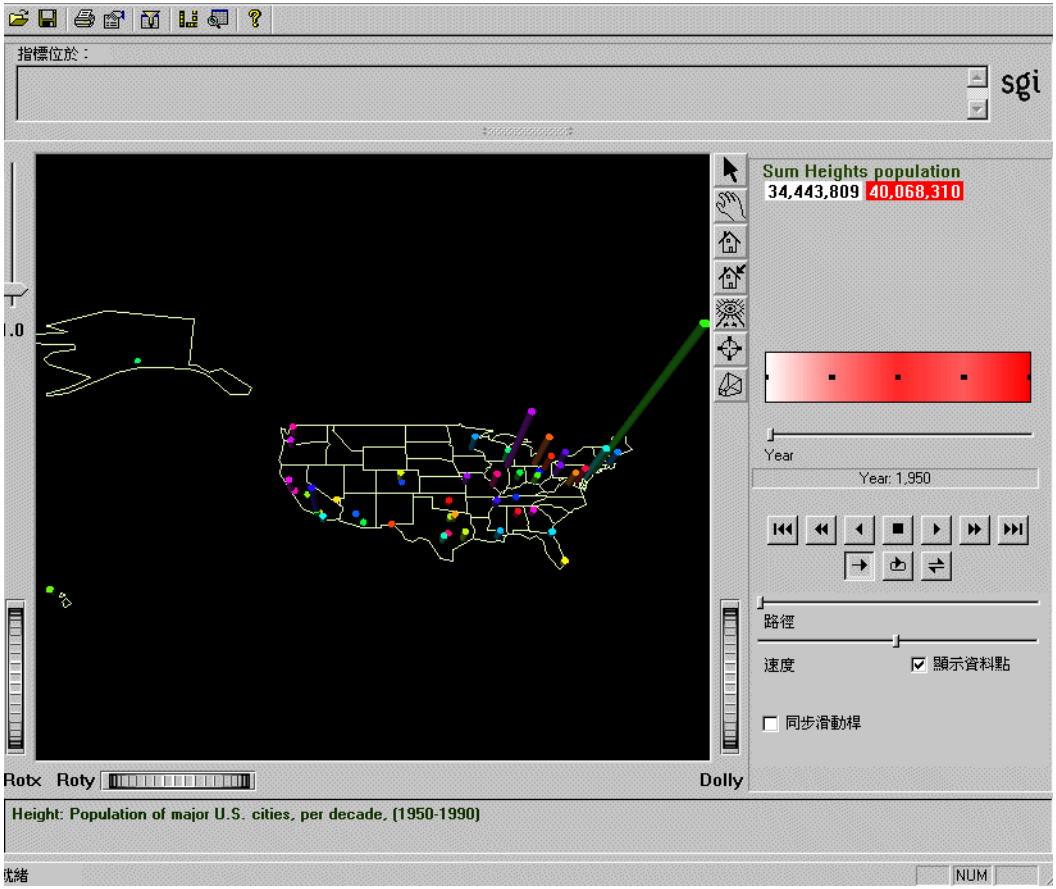


圖 6-2 帶有地理輪廓上相關群體長條圖的地圖可視化工具

另一個可能的景觀會在地圖上顯示連接著特定定位置的線，每條線具有各自的寬度和顏色 (請參閱圖 6-3)。這些線的特徵對應于普通物件和圓柱體的高度和顏色屬性。

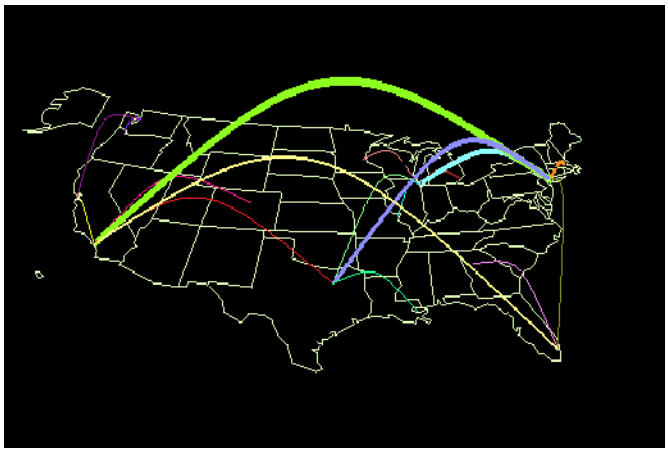


圖 6-3 顯示帶有指定端點的美國之地圖可視化工具樣本

執行地圖可視化工具

以下是執行「地圖可視化工具」最常用的方法：

1. 開啓「工具管理員」，選擇伺服器和資料來源。如果需要說明，請參閱第 2 章，「利用 MineSet 存取資料」。本範例使用「*客戶波動*」資料集。
2. 在「工具管理員資料目標」窗格中按一下「可視化工具」索引標籤；然後在下面一列的索引標籤中，按一下「地圖可視化工具」。

準備資料

本部分說明轉換地理資料的步驟，以便「地圖可視化工具」進行顯示。

組合資料

在地圖上描繪資料之前，必須先將資料轉換到較小的資料集。此種轉換稱爲組合。

1. 在「工具管理員資料轉換」窗格中按一下*組合*。

圖 6-4 在地圖可視化工具中組合

2. 將感興趣的欄移到「要組合的欄」清單中，然後使用底部的核取方塊標識需要進行的數學操作。將空間分組保留在中間的「分組索引欄」欄位中，然後將其他所有的欄從可視化處理中移除。關於操作的說明，請參閱 第 41 頁上的 圖 3-5。

例如：使用*客戶波動*資料集，在「組合」對話方塊中，按一下國際電話總數 (total intl calls) 欄，然後再按一下向左箭頭，將其移到左邊清單中。在底部核取方塊中核取 *平均*和*計數*，關閉*總和*。將「`state`」保留在中間的清單中，然後將其餘的所有項目移到右邊清單。(按住 Ctrl 鍵以選擇多個欄。) 按一下*確定*以應用您的選擇。

下一步是尋找適當的空間形狀。

選擇地圖形狀

1. 在「資料目標」窗格中，按一下「可視化工具」索引標籤上的*工具選項*按鈕，存取地理檔案。「工具管理員」將顯示「地圖可視化工具選項」對話方塊。
2. 按一下「實體」文字欄位右側的按鈕，選擇一個地理實體檔案。這些檔案在 MineSet 的初始安裝目錄中，位於 *MineSet 3.0\config\mapviz\gfx_files* 下。本範例使用 *usa.state.hierarchy*。
3. 按一下*確定*，找到選定的檔案，然後再按一下*確定*，關閉「地圖可視化工具選項」面板。關於 *.gfx 檔案*的詳細內容，請參閱*《MineSet 3.0 企業版參考指南》*中的「可視化工具」一節。

   下一步是將可視元件與欄相連接。

連接地圖可視化工具元件

將地圖元件連接到可視實體時，一些欄可能需要提前進行處理才能適合可視元件的要求：

- 如果欄擁有太多的離散值，請參閱第 3 章中的「爲欄改變或建立新的分組」。
- 如果欄的類型不正確，請參閱第 3 章中的「更改欄類型或名稱」。

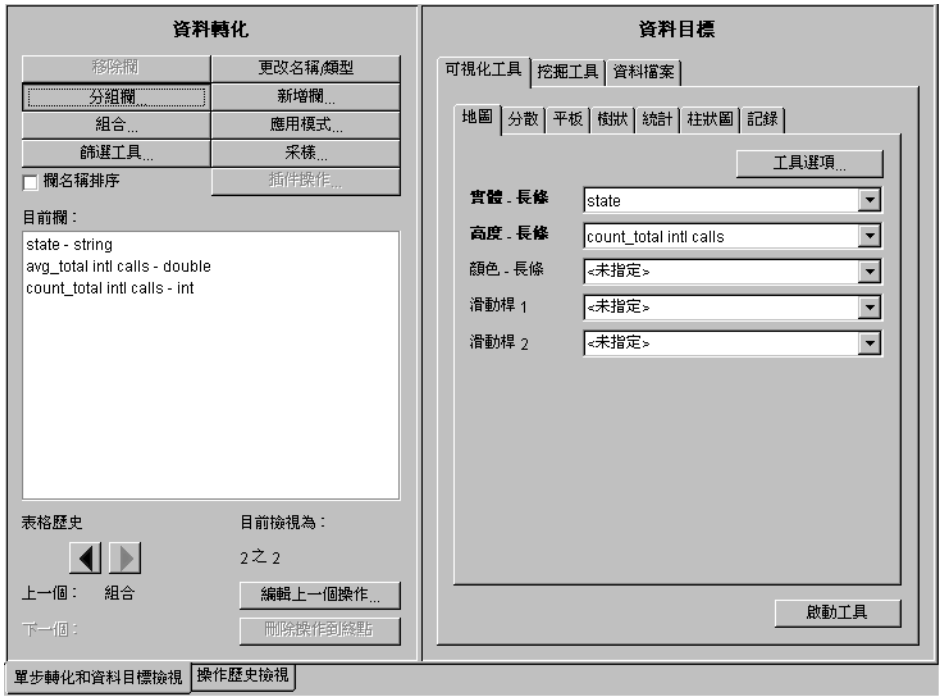


圖 6-5 將欄對照到地圖可視化工具的可視元件

1. 將在工具管理員「資料轉換」窗格中的欄對照到「資料目標」窗格中的實體。在每個實體旁的跳現式功能表中進行選擇以確定對照項目。

   對於該範例，爲實體 - 長條圖選擇「`state`」，爲高度 - 長條圖選擇「count_total intl calls」。

   關於對照可行性的進一步資訊，請參閱「對照到滑動桿」和「對照到地圖」。

   按一下*啓動工具*，查看資料集的地理分布。

對照到滑動桿

當您覺得一個值根據特定的準則變化 (如收入隨年齡變化) 時，可以將該欄對照到滑動桿，將其他欄對照到別的可視元件。例如：圖 6-1 中有個對照到「年」這一欄的滑動桿。這樣可以透過移動滑動桿觀察資料的逐年變化。

如果欄是 (整數、浮點數、雙精度類型) 的數值型或已分組的，就可以對照到滑動桿。如果欄已經分組，則其結尾就會帶有 `_bin` 字樣。欄類型注釋位於「目前欄」欄位中的欄名稱後，例
如：白天的電話總數 - 雙精度。

可以自動或手動建立滑動桿。當把欄對照到滑動桿時，「工具管理員」將自動產生滑動桿。滑動桿透過自動分組和組合來建立。請參閱詞彙表項目：*分組*和*組合*。這些自動操作會發生在按一下*啓動工具*之後。

如果在目前資料表格中有一個陣列欄，「工具管理員」可以建立對應於該陣列欄下標的滑動桿。「工具選項」對話方塊中指出目前的滑動桿下標。要顯示它們，可以在「工具管理員資料目標」窗格中按一下「可視化工具」索引標籤，然後再按一下*工具選項*按鈕。目前表格中的所有陣列欄都必須具有相同的下標；否則，將不會建立滑動桿。

關於建立陣列欄的詳細資訊，請參閱第 3 章中的「透過組合建立新欄」和*《MineSet 3.0 企業版參考指南》*中的「組合」。

關於自動和手動建立滑動桿的詳細資訊，請參閱*《MineSet 3.0 企業版參考指南》*中的「地圖、分散、平板可視化工具中的滑動桿建立」。

對照到地圖

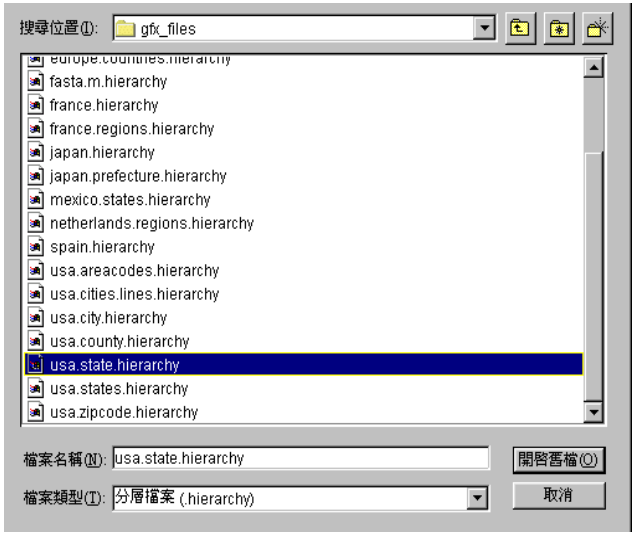
MineSet 提供多種可識別地形和其他空間實體組成的 *.gfx* 和 *.hierarchy* 檔案。除非將實體長條對照到一個圖形元件，否則無法得到可識別的地圖。如果希望顯示其他可視實體或地圖，可用以下步驟取代第 104 頁的「選擇地圖形狀」：

1. 在「工具管理員 可視化工具」索引標籤中，按一下*工具選項*按鈕。
2. 在對話方塊中選擇所需的選項。「地圖可視化選項」對話方塊，如圖 6-6 所示。
   - 對於地圖形狀的實體，可以按一下「實體檔案」文字欄位右邊的*尋找檔案*按鈕，然後選擇一個在「地圖可視化工具」主視窗中用於代表物件的 *.hierarchy* 地理實體檔案。可以在 */config/mapviz/gfx_files* 中找到這些範例。

- 對於矩形，可以按一下「輪廓檔案」文字欄位右邊的*尋找檔案*按鈕，指定要描繪的輪廓物件。將出現爲一個放置 3D 實體物件的平面。

「實體檔案」和「輪廓檔案」欄位是可選的。如果沒有提供「實體檔案」，可以看到實體物件由任意擺放在場景中的矩形組成。

關於其他選項的說明，請參閱*《MineSet 3.0 企業版參考指南》*中的「可視化工具」。



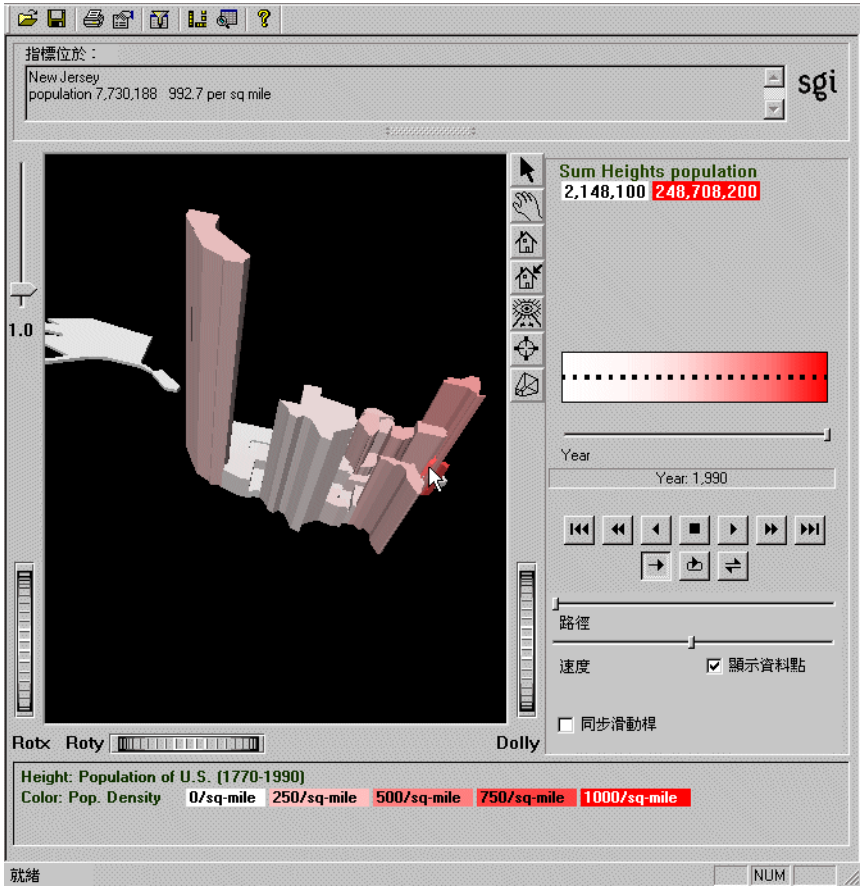
6-6 「地圖可視化工具選項」對話方塊

3. 完成了對「工具選項」對話方塊的更改後，請按一下*確定*，關閉「工具管理員」的螢幕。然後可以按一下*調用工具*。

檢視地圖可視化工具

與所有可視化工具一樣，如果在沒有指定配置文件的情況下執行了「地圖可視化工具」，那麼可視化工具主視窗是空的，並且只能使用「檔案」、「檢視」和「說明」下拉式功能表。要顯示所有功能表和控制項，請使用「檔案」>「開啓」，以查看配置檔案的清單。這些檔案在 *MineSet 3.0\examples* 目錄中，位於 MineSet 的初始安裝目錄。

如果已經指定有效的配置檔案，可視化工具將會顯示地理景觀。例如：圖 6-7 所示爲指定 *population.usa.mapviz* 並將*年份* 滑動桿移到極右處時得到的結果。



6-7 年份滑動桿位於 1990 年的 Population.usa.mapviz 範例

該畫面顯示美國各州的人口和人口密度。各州的人口由州圖形的高度代表。高度在主視窗右邊區域的整個動畫控制範圍中相互關聯著。

查看模式

兩種檢視模式爲*抓取*和*選取*。要在這兩種模式之間切換，可以將游標移到主視窗中，按一下視窗邊界上的箭頭或手形，然後按 Esc 鍵。想從一種模式改變爲另一種模式，關於詳細內容，請參閱第 64 頁。

概化上尋和細化下尋

只有當 *.data* 和 *.hierarchy* 檔案支援概化上尋和細化下尋時，才可以在選擇模式中細化下尋以獲得詳細資訊，或概化上尋以查看更全面的檢視。「地圖可視化工具」樣本檔案支援細化下尋和概化上尋。

表 6-1 追溯操作

目的：	**方法：**
細化下尋物件	在物件上按一下滑鼠右鍵。
細化下尋所有物件	在背景上按一下滑鼠右鍵。
概化上尋物件	在物件上按住 Ctrl 鍵並按一下滑鼠右鍵 (或滑鼠中鍵)。
概化上尋所有物件	在背景上按住 Ctrl 鍵並按一下滑鼠右鍵 (或滑鼠中鍵)。

可以重複細化下尋到資料支援的最細微級別的模組。物件的數目越大，「地圖可視化工具」在追溯操作後用於重建場景的時間越長，並且移動動畫控制項時的執行也越慢。

預設情況下，「地圖可視化工具」一開始將顯示最細微級別下的物件細節，因此，這時只有概化上尋 (更大的模組) 是正在活動狀態的。

可視化空資料

當在特定位置缺少資料或數據爲空時，將顯示一個問號。在「地圖可視化工具」中出現空值，可能由於資料庫或資料檔案包含空，也可能由於其他特定原因所引起 (請參閱 *《MineSet 3.0 企業版參考指南》* 中的「空值」一節)。

在地圖可視化工具中建立動畫

可以使用主視窗右邊的動畫控制面板建立動畫。只有當資料集對照到一個或兩個滑動桿時，才會出現動畫視窗。請參閱 *《MineSet 3.0 企業版參考指南》* 中的「動畫」。

處理地圖可視化工具結果

使用下拉式功能表「檢視」和「選項」，可以進入「地圖可視化工具」的擴充功能。

更改地圖可視化工具畫面

「檢視」功能表提供的選項匯總在表 6-2 中。關於進一步的詳細內容，請參閱 *《MineSet 3.0 企業版參考指南》* 中的「檢視」功能表。

表 6-2 「地圖可視化工具」的查看功能表選項

選項	描述
篩選面板	開啓「篩選」面板，根據選定準則對主視窗中顯示的實體數進行篩選。請參閱 *《MineSet 3.0 企業版參考指南》* 中的「篩選」一節。
顯示控制方塊	顯示或隱藏主視窗周圍的外部控制項
動畫面板	顯示或隱藏主視窗右邊的所有動畫控制面板
資料點	使 2D 匯總視窗中出現或隱藏由黑色圓點組成、用於指示精確資料值的柵格

表 6-2 「地圖可視化工具」的查看功能表選項

選項	描述
使用隨機顏色	使配置檔案中的顏色對照規定啓動或失效。
顯示 X-Y 坐標	用於形成和改進 *.gfx* 檔案，而非用於資料分析。關於詳細的內容，請參閱《*MineSet 3.0 企業版參考指南*》中的「地圖可視化工具」。

選項和追溯

「選項」功能表可以對「地圖可視化工具」中選取的資料執行表 6-3 中所列的操作。關於進一步的詳細內容，請參閱《*MineSet 3.0 企業版參考指南*》中的「地圖可視化工具」。

表 6-3 「地圖可視化工具」的選項功能表選擇

選項	描述
顯示值	顯示選定物件的值的「記錄檢視器」表格
顯示原始資料	檢索並將選取的記錄顯示爲「記錄檢視器」表
傳送到工具管理員	根據目前的選擇在「工具管理員」歷史的開始處插入一個篩選操作。
在細節追溯中使用滑動桿	確定在追溯表達式建立過程中是否使用了滑動桿定位。預設情況下，將追溯限制在由滑動桿位置定義的記錄中。
*互*補追溯	獲取*沒有*由*顯示原始資料*和*傳送到工具管理員*選定的所有資料。

關於追溯的詳細內容，請參閱《*MineSet 3.0 企業版參考指南*》中的「追溯」。

關於「地圖可視化工具」附帶樣例檔案的說明，請參閱《*MineSet 3.0 企業版參考指南*》中的「樣本檔案」。

理解預測建模

本章介紹預測建模並說明如何使用 MineSet 從資料集中產生預測模型。由 MineSet 產生的可視化處理有助於了解如何運作模型。可以使用追溯技術檢查基本資料。討論的主題包括：

隨後的章節中將深入介紹可視化處理和模型。關於進一步的技術資訊，請參閱*《MineSet 3.0 企業版參考指南》*。

預測建模總覽

MineSet 包含許多用於產生預測模型的資料分析挖掘運算法則。如果指定記錄中的其他幾個屬性值，預測模型就可以預測稱爲標籤的屬性值。如果在屬性中有足夠的資訊，模型就可以對標籤進行準確的預測。已預測的標籤指的是在指定記錄中的未知特徵。例如：在信用卡歷史資料集中，指定了屬性「年齡」的值、「性別」和「職業」等屬性的值，預測模型的任務可能就是預測「信用風險」屬性的值。

如果資料中含有未知的標籤，請使用預測建模，但是如果需要可視化含有已知標籤的資料，使用描述性建模更爲合適。MineSet 的兩類預測建模分別爲分類和回歸。分類預測離散值，回歸預測連續範圍內的值。隨後的章節將個別作詳細的介紹。本章主要討論所有預測模型的一般過程。

產生模型

用於產生預測模型的 MineSet 運算法則取決於將要預測的屬性類型。對於分類任務，MineSet 的運算法則將產生「証據」模型、「決策樹」、「選項樹」或「決策表」。對於回歸任務，MineSet 將產生「回歸樹」。

以下五個小節對 MineSet 支援的各種類型預測模型進行簡短的介紹，並且利用圖例說明它們如何產生。可以使用「工具管理員資料目標」窗格中的*進階選項*按鈕制定模型產生運算法則的運作方式。關於這些選項是如何影響模型產生器的詳細技術說明，請參閱*《MineSet 3.0 企業版參考指南》*中的「工具管理員中的導入工具模式」。

關於單獨的導入工具之相關讀物清單，請參閱*《MineSet 3.0 企業版介面指南》*中的附錄 A。

証據模型

本部分顯示如何產生稱爲「証據分類器」的預測模型，有關的詳細內容請參閱 第 10 章，「用証據分類器和可視化工具建模和預測」。

要建立預測模型，先使用「工具管理員」選擇資料集。關於詳細內容，請參閱第 2 章中的「執行 MineSet 」，該範例使用客戶波動資料集建立模型，用來預測哪些用戶可能改變他們的電信供應商。

1. 在「工具管理員資料目標」窗格中，按一下「挖掘工具」索引標籤。
2. 按一下「分類」索引標籤，然後從跳現式功能表中進行下列選擇：

   *模式：*僅用於分類器

   *導入工具：*証據

   離散標籤： churned (例如)

   即將導入証據分類器以幫助可能會波動的用戶特徵化。在該示例中，將使用所有的資料來建立模型，而不爲準確性的測試保留任何資料。

3. 按一下*開始*。

   「工具管理員」底部的「狀態」視窗將顯示導入過程的進度和匯總資訊。當導入步驟完成後，系統將自動啓動「証據可視化工具」，顯示模型的可視化處理，如 圖 7-1 所示。

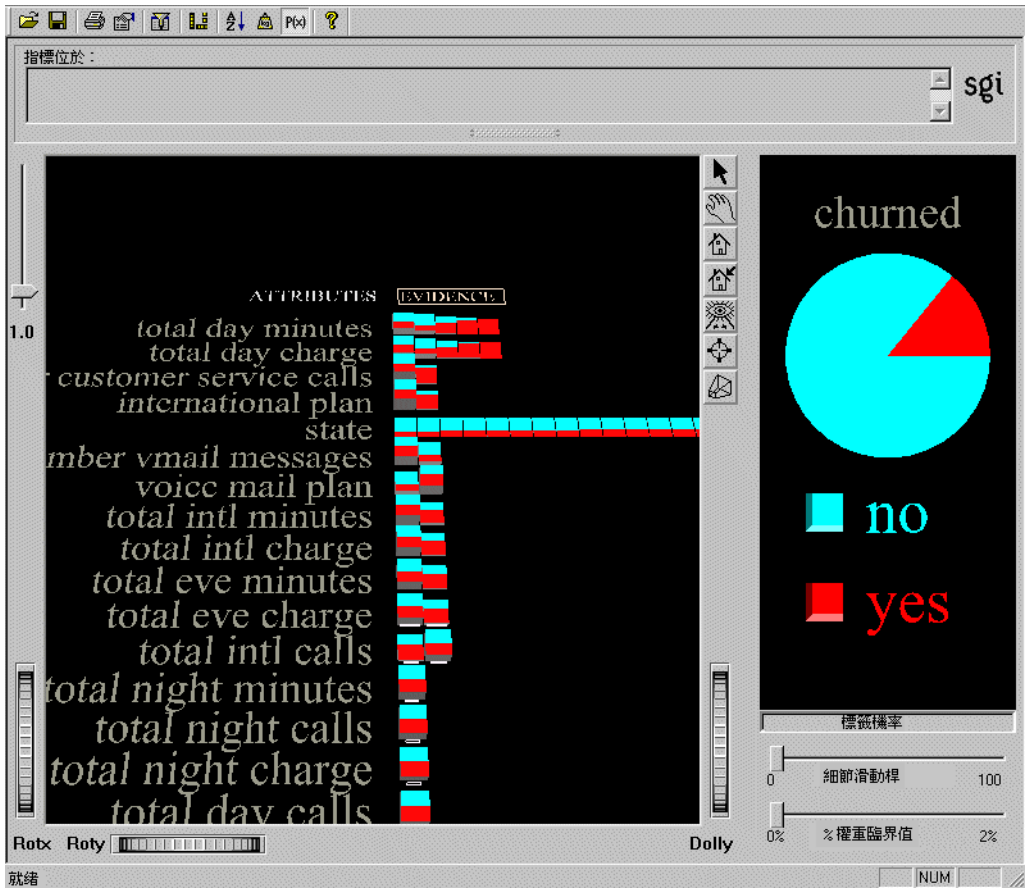


圖 7-1 「証據導入工具」爲客戶波動資料集產生的証據可視化處理

該模型顯示每個屬性預測客戶波動值的準確性。關於如何根據預測行爲使用模型，請參閱第 10 章，「用証據分類器和可視化工具建模和預測」。

決策樹模型

在上一個範例中，使用了「証據」方法導出並可視化分類器。如果選擇了「決策樹」方法，可以顯示各種屬性是如何相互作用的，即：屬性值的組合是如何影響已預測的標籤。有了「決策樹」，資料如何在以後的節點 (決策點) 上分布將取決於上一個節點作出的決策。在圖 7-2 所示的範例中，根決策點的分支爲白天的分鐘總數屬性是否大於或小於 264.45。這會影響後續分支中值的分布；顯然，較高的長條圖出現在可視化處理的左邊。

要產生「決策樹模型」，其步驟與產生「証據模型」的步驟一樣，除了在「分類」索引標籤上選擇「導入工具」時，需要選擇「決策樹」。在第 8 章，「用決策、選項和回歸樹建模和預測」中將全面說明該任務。該方法的結果如圖 7-2 所示。從根擴展出去的子樹中類的分布不同，這表明屬性之間的相互作用非常顯著。

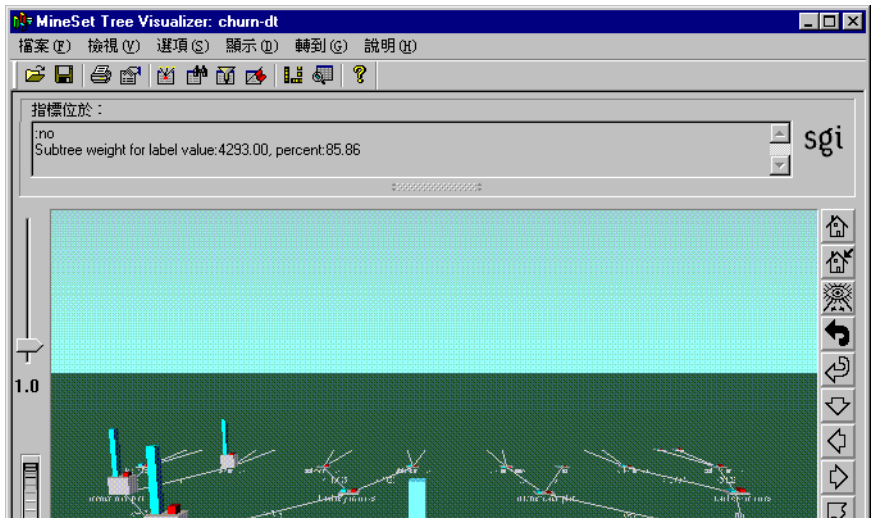


圖 7-2 由「決策樹導入工具」對客戶波動資料集產生的決策樹

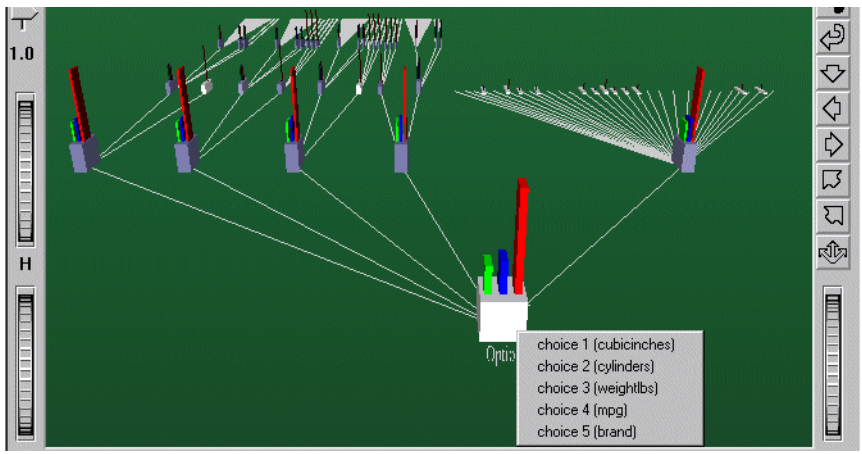
該決策樹每個節點上的長條圖代表標籤值 (分類) 的分布。當游標放在長條圖上時，畫面上的狀態區域會顯示該標籤值的記錄權重和百分比。每個節點的基準高度代表記錄的權重
(如果沒有使用權重，就用記錄計數)。

在這個範例中，代表決策樹根的小塊由標籤「白天的分鐘總數」所標示，表明這是預測客戶波動中一個最重要的因素—這些用戶的交談時間，劃分的臨界值爲 264.45 分鐘。在第 8 章，「用決策、選項和回歸樹建模和預測」中將對「決策樹」進行詳細說明。

選項樹模型

「決策樹導入工具」爲每個子樹選擇一個「最好」的屬性；然而可能有多個用以拆分的好屬性。此時，「選項樹」可以建立選項節點。「選項樹」與「決策樹」類似，但它是由「選項樹導入工具」產生的。

頂節點是一個「選項節點」，顯示有多個好屬性可以在根部選擇。可以設定每個節點產生的選項數，表示該點最好的屬性。然後將爲每個選項產生一個分支。圖 7-3 所示的範例使用汽車資料集。在第 8 章，「用決策、選項和回歸樹建模和預測」中將對「選項樹」進行詳細說明。



7-3 由選項樹導入工具對「汽車」資料集產生的選項樹

在範例資料集 (圖 7-3)，要預測汽車的產地是在歐洲、日本或美國。「決策樹導入工具」爲根選取了體積屬性。相反地，「選項樹導入工具」選擇多個選項：體積、汽缸、重量、英哩 / 加侖和商標都是根部的最佳選擇。在根上按一下 Ctrl 鍵和滑鼠右鍵可以看到這些選項。

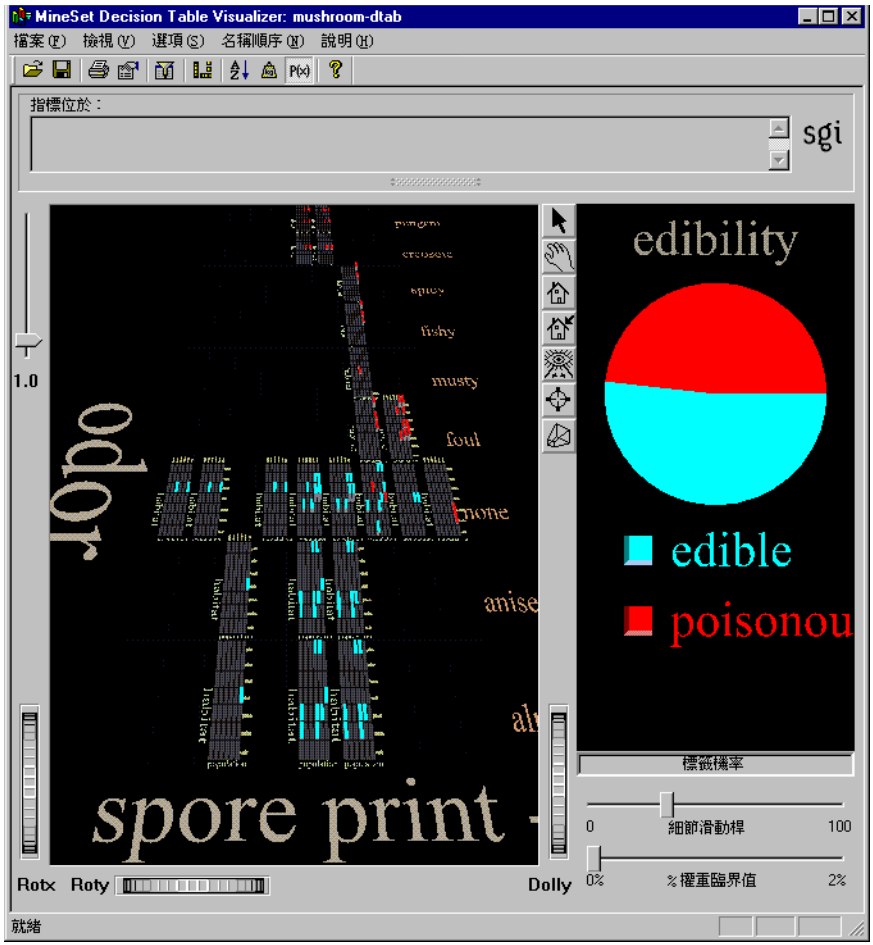
選項節點也可出現在根以外的其他地方。但在預設情況下，它們只出現在根或緊鄰根的下一級。

相對於「決策樹」，「選項樹」通常需要 10 到 15 倍的時間用來建立，並且規模要大得多，但是它們具有兩個顯著的優點：

- 彈性—可以使用選項節點看到多個合適的選項。可以使用選項節點從多個而不是單個屬性中進行選項設定。當掠過樹時，可以跟隨一個易於了解或適合用戶對該問題之背景知識的選項。
- 準確性—通常，「選項樹」的錯誤率低於「決策樹」。「選項樹」的分類是使每個選項給每個標籤值「投票」，然後將選票平均。這好比擁有一個專家小組，每個人都試圖根據不同的標準來預測標籤。選項節點將所有專家的選票平均，產生一個更加穩定、風險更小的分類器。

決策表模型

「決策表」導入的模型表示分層結構中屬性組之間的關聯。導入工具進行決策的方式與「決策樹」的一樣，但是屬性的評估遍及樹的所有等級，而非某個特定的子樹。然後，結果的表現形式是分層表而不是樹。在可視化處理中的小塊圖上按一下將顯示其詳細組成內容，如圖 7-4 所示。



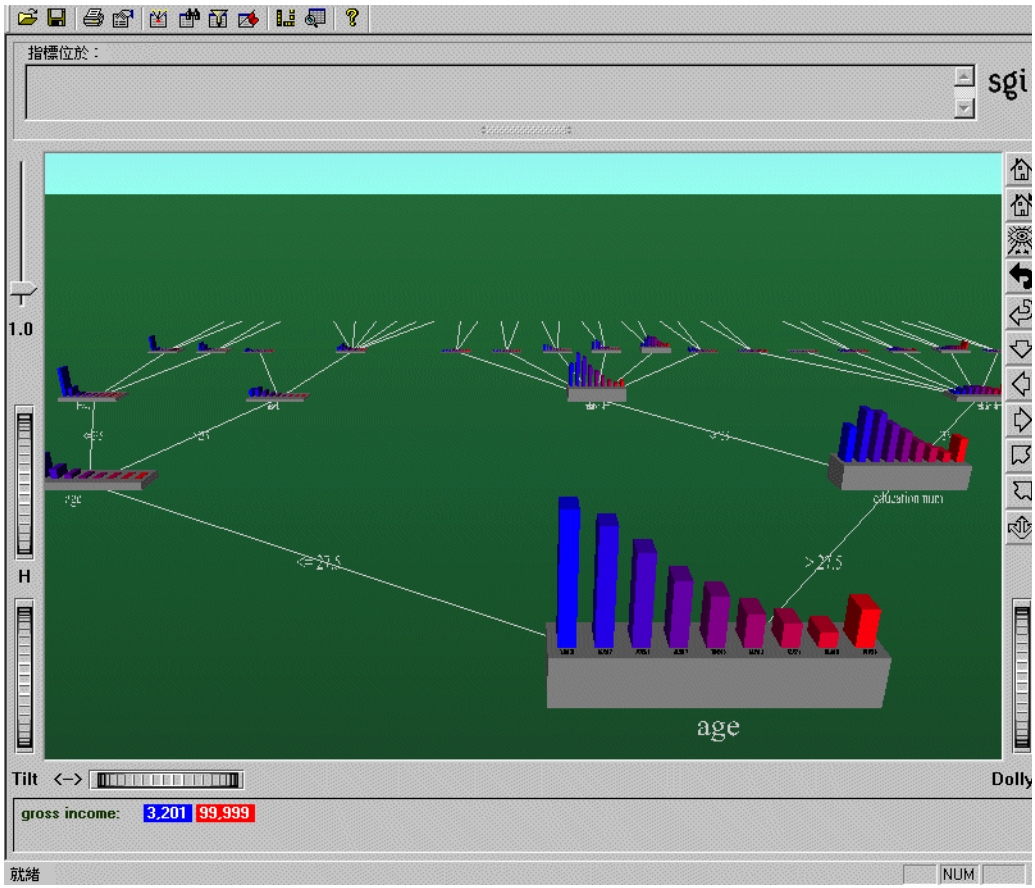
7-4 決策表導入工具對「蘑菇」資料集產生的決策表

該模型顯示預測蘑菇可食性的標籤機率。

關於如何使用「決策表」的說明，請參閱第 9 章，「用決策表分類器和可視化工具建模和預測」。

回歸樹模型

「回歸樹導入工具」產生一個回歸器，它是一個類似於分類器的預測模型。其區別在於回歸器預測的是具有連續值標籤，例如個人的年薪。而分類器只預測離散值，例如：「是」或「否」，或一個指定的範圍 (如 20-100)。



7-5 回歸器對「成人」資料集產生的回歸樹

下面的範例說明如何產生回歸器，在第 8 章，「用決策、選項和回歸樹建模和預測」中將介紹這個任務。

要建立預測回歸器，先使用「工具管理員」選擇一個資料集。關於詳細內容，請參閱第 2 章，「利用 MineSet 存取資料」。該範例使用成人資料集。

1. 在「工具管理員資料目標」窗格中，按一下「挖掘工具」索引標籤。
2. 按一下「回歸」索引標籤，然後從跳現式功能表中進行下列選擇：

   *模式：*僅回歸器

   *導入工具：*回歸樹

   *連續標籤：*總收入 (例如)

   即將導入一個回歸器，以幫助資料集中記錄代表的成人的總收入特徵化。在該示例中，將使用所有的資料來建立模型。
3. 按一下*開始*。

   「工具管理員」底部的「狀態」視窗將顯示導入過程的進度和匯總資訊，也可在對話方塊中觀看進程指示器。當導入步驟完成後，系統將自動調用「回歸樹可視化工具」，顯示一個模型的可視化處理，如圖 7-5 所示。某些資料集產生的時間較長。

可以爲擁有連續屬性 (即：屬性值在一個連續範圍內變化) 的資料集產生一個回歸器，如果沒有這樣的欄，可以從「工具管理員資料轉換」窗格中增加一個新的連續欄，請參閱第 3 章中的「移除和增加欄 」。

第 126 頁的「建立學習曲線」中將討論學習曲線的分類器模式。

評估預測模型

預測模型的目的是爲了進行預測，因此模型的有效性顯然取決於其預測的準確性。模型的準確性由錯誤率衡量。本部分討論 MineSet 建模模式的範圍：「僅用於分類器」使用所有的資料建模，不進行錯誤估計；估計錯誤率的方法通常有兩種—預留和交叉驗証。

在估計誤差的預留模式中，使用資料集的一部分 (通常爲三分之二) 來產生模型。導入工具使用該訓練資料集中的標籤構造模型，其餘的資料用於測試模型和估計錯誤率。

交叉驗証是估計分類器誤差的另一種方法，它將資料集拆分爲一定數量的資料夾或子集 (一般是 10 個)，然後建立相同數量的分類器。可以多次重複這個過程以提高估計的可靠性。

預留與交叉驗証兩種方法都是期望將來採集的記錄與訓練集的資料分布是相同的。*《MineSet 3.0 企業版參考指南》*的「誤差估計」和「交叉驗証」對此進行更加詳細的討論。

使用所有資料進行分類

建立模型的*僅用於分類器*模式將使用所有可用的資料來建立分類器。當您不考慮估計預測模型的錯誤率時 (例如：只是希望查看可視化處理)，這是很有用的。

假如您希望得到用*僅用於分類器*模式導入的蝴蝶花資料集的「決策樹」。這個範例講述建立並應用模型的整個程序。按照以下步驟，可以發現模型的誤差分類：

1. 使用「文件」 >「打開新的資料文件」下拉功能表，從「工具管理員」視窗中選擇一個資料集 . 選擇 *iris.schema* 資料集。
2. 在「工具管理員資料目標」窗格中，按一下「挖掘工具」索引標籤，然後按一下「分類」索引標籤並選擇
   - *模式：*僅用於分類器
   - *導入工具：*決策樹
   - *離散標籤：*蝴蝶花類型

   將使用所有的資料建立「決策樹」模型，確定用以區分不同類型蝴蝶花的屬性。
3. 按一下*開始*，執行「決策樹」。檢查，然後關閉可視化處理。建立好的模型使用了所有的資料。以下的步驟將顯示在此種情況下，如何評估正確和不正確的預測。

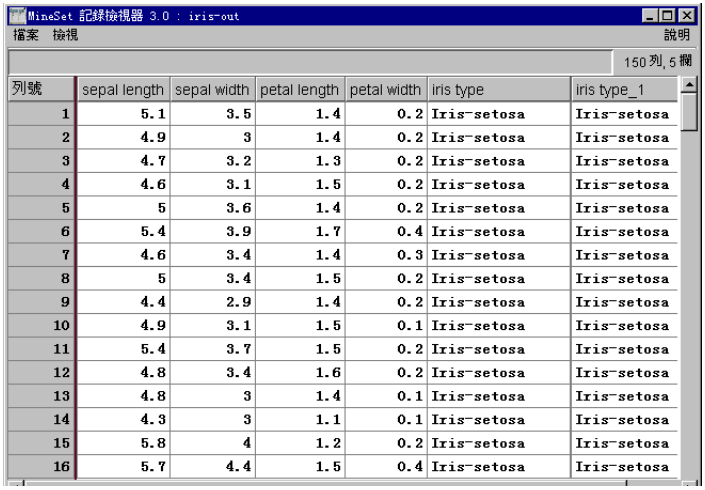
4. 在「工具管理員資料轉換」窗格中按一下*應用模型*。選擇 *iris-dt.class* (蝴蝶花決策表分類器)，然後按一下*確定*接受預設的新欄名稱 `iris type_1`，其中含有預測的標籤。

5. 按一下*增加欄*按鈕，建立另一個標識錯誤的新欄 (稱此欄爲 `iris_fault`)。然後，可以增加一欄，定義爲帶有表達式 (`'iris type'!= 'iris type_1'`) 的 **int** 類型。

   要建立表達式，可以從左邊的清單中選擇，然後按向右箭頭，或直接在本文欄位中鍵入。

6. 按一下*檢查表達式*，然後按一下「確定」。這將增加一個基於 `iris type_1` 的新欄 `iris_fault`。當分類器分類錯誤時，該欄中有一個 1；當分類正確時，該欄中有一個 0。

7. 驗證結果。在「工具管理員資料目標」窗格中，按一下「Viz 工具」，由「工具」跳現式功能表中選擇「記錄檢視器」，然後按一下*啓動工具*。圖 7-6 所示爲結果的一個範例。



列號	sepal length	sepal width	petal length	petal width	iris type	iris type_1
1	5.1	3.5	1.4	0.2	Iris-setosa	Iris-setosa
2	4.9	3	1.4	0.2	Iris-setosa	Iris-setosa
3	4.7	3.2	1.3	0.2	Iris-setosa	Iris-setosa
4	4.6	3.1	1.5	0.2	Iris-setosa	Iris-setosa
5	5	3.6	1.4	0.2	Iris-setosa	Iris-setosa
6	5.4	3.9	1.7	0.4	Iris-setosa	Iris-setosa
7	4.6	3.4	1.4	0.3	Iris-setosa	Iris-setosa
8	5	3.4	1.5	0.2	Iris-setosa	Iris-setosa
9	4.4	2.9	1.4	0.2	Iris-setosa	Iris-setosa
10	4.9	3.1	1.5	0.1	Iris-setosa	Iris-setosa
11	5.4	3.7	1.5	0.2	Iris-setosa	Iris-setosa
12	4.8	3.4	1.6	0.2	Iris-setosa	Iris-setosa
13	4.8	3	1.4	0.1	Iris-setosa	Iris-setosa
14	4.3	3	1.1	0.1	Iris-setosa	Iris-setosa
15	5.8	4	1.2	0.2	Iris-setosa	Iris-setosa
16	5.7	4.4	1.5	0.4	Iris-setosa	Iris-setosa

7-6 蝴蝶花錯誤分類範例

然後，可以將結果應用到一個分散圖，使新欄對照到顏色，顏色集合中綠色是 0 (沒有錯誤)，紅色是 1 (錯誤)。關於「分散可視化工具」的範例，請參閱第 4 章，「用分散和平板可視化工具檢查資料」。

預留誤差估計

不是使用所有的資料建立模型，而是僅使用部分資料作爲訓練集來導入分類器。「分類器和誤差」模式 (MineSet 的預設模式) 將自動將資料集劃分爲訓練和測試兩個子集。可以改變用作訓練集的資料比例以滿足要求。

要了解預留錯誤估計如何應用到選定的資料集，可以按照下列步驟：

1. 從 *iris.schema* 資料集開始 (請參閱第 122 頁)。
2. 在「工具管理員資料目標」窗格中，按一下「挖掘工具」索引標籤，然後按一下「分類」索引標籤並選擇：
   - *模式：* 分類器和錯誤估計
   - *導入工具：*決策樹
   - *離散標籤：*蝴蝶花類型

   將蝴蝶花資料集劃分爲兩個子集。訓練集中包含三分之二 (預設的支持記錄比例) 的記錄用於建立「決策樹」模型。剩餘的三分之一記錄用於估計模型的錯誤率。
3. 按一下*進階選項*按鈕，使用文字面板改變訓練集的支持比例，如圖 7-7 所示。也可以改變隨機子以獲得不同訓練集和測試集。請參閱詞彙表中的項目*隨機子*。按一下*確定*，關閉面板。

7-7 預留的錯誤估計選項

4. 按一下*開始*，執行「決策樹」導入工具。

主視窗底部的狀態區域中顯示使用測試集 (剩餘的三分之一資料) 估計的導入工具錯誤率。可以向下捲動，查看所有資訊。

交叉驗証誤差估計

爲建立最終的分類器或小資料集可以使用交叉驗証的級。使用的過程可以說明其原因。交叉驗証是對誤差取得更加精確的估計方法。在 *n-fold* 交叉驗証 (其中 *n* 代表賦予名稱一個數字，*fold* 是希望將資料劃分成的子集) 中，資料集被劃分爲 *n* 個獨立的 子集。每個子集將依次被留下，其餘的 *n*-1 個子集 (比最初指定的數目少 1) 組成爲一個訓練集。產生的模型使用留下的子集進行評測。然後，對這 *n* 個獨立的估計進行平均，並將得到的資料組合起來建立最終的模型。*N*-fold 交叉驗証使用的時間大約是「分類器和誤差」或「僅用於分類器」方法的 *n*+1 倍。

將交叉驗証應用到一個選定資料集並設定倍數：

1. 在「工具管理員資料目標」窗格中，選擇「挖掘工具」索引標籤，然後按一下「分類」。
2. 如果使用所有資料 (如在「僅用於分類器」模式中)，在「模式」跳現式功能表中選擇*誤差估計*以存取誤差。由於使用交叉驗証，執行的時間可能較長。當資料很少時，使用*誤差估計*。導入的分類器與*僅用於分類器*模式導入的分類器完全一樣。
3. 按一下*進階選項*按鈕，改變交叉驗証執行的方式。在「誤差估計選項」(請參閱 圖 7-8) 中，可以設定 Fold 和 Time 的值以重複誤差估計過程。按一下*確定*接受所有更改。

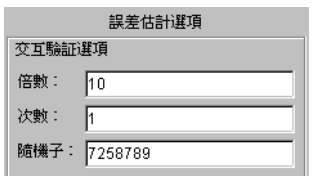


圖 7-8　交叉驗証的誤差估計選項

4. 按一下*開始*，執行運算法則。

關於「誤差估計」更加詳細的討論，請參閱 *《MineSet 3.0 企業版參考指南》*。

建立學習曲線

學習曲線顯示導入工具產生的分類器的誤差，與用於建立分類器的記錄數成反比。通常，用於產生分類器的記錄越多，誤差就越小。

經由指出希望調查訓練集的大小範圍，和對於每個訓練集大小希望建立導入工具的次數，可以建立學習曲線。要指出希望調查的訓練集之大小範圍，可以輸入記錄的起始數 (最小值) 和終止數 (最大值) 以及要檢查的點數。每個分類器都是使用記錄的隨機樣本產生的，並使用剩餘的記錄 (沒有用於訓練的記錄) 估計其誤差。

1. 從第 116 頁使用的 *churn.schema* 資料集開始。
2. 在「工具管理員資料目標」窗格中，按一下「挖掘工具」，然後按一下「分類」索引標籤並選擇：

   *模式：*學習曲線

   *導入工具：*決策樹

   離散標籤： churned

   系統將提示不計算電話號碼欄，因爲其中包含太多不同的值。可以刪除該欄，或是按一下「確定」。
3. 按一下*開始*。

   該過程可能需要較長的時間，因爲導入工具正在爲曲線上的每個點產生一個分類器。狀態視窗中會顯示訓練範圍、已經完成的訓練次數、和每次的平均誤差。

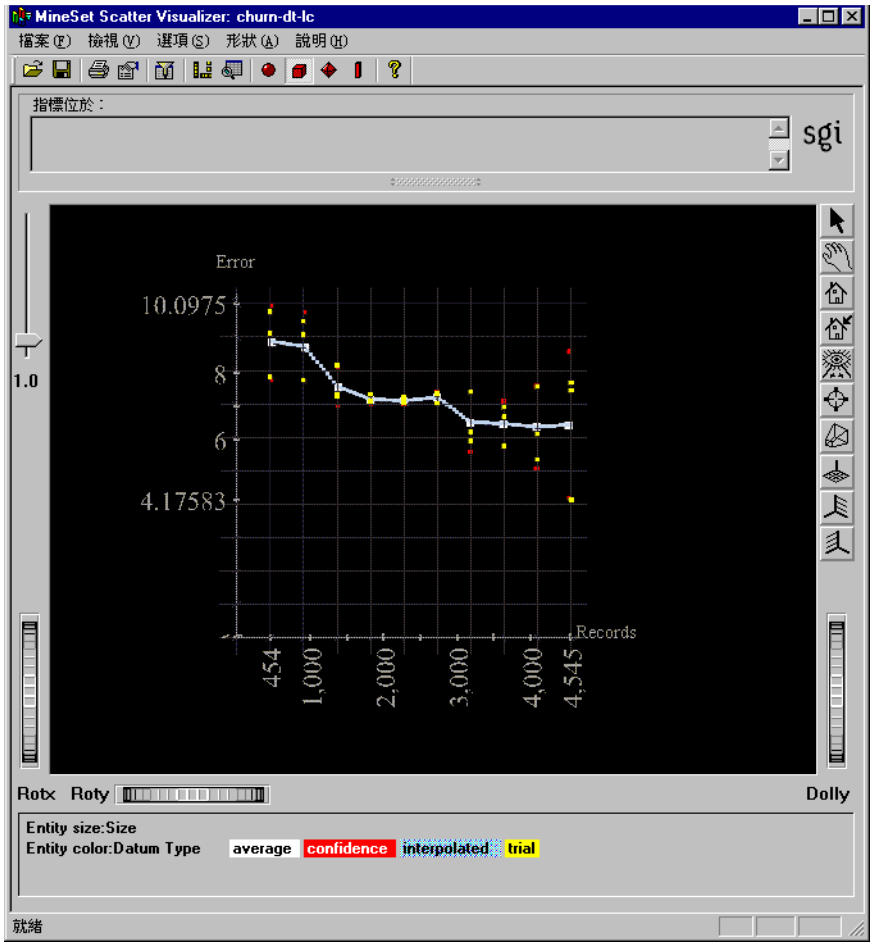


圖 7-9 帶有客戶波動標籤集的客戶波動資料集的學習曲線

圖 7-9 指出客戶波動資料集中指定記錄數的錯誤率。如果訓練了 454 個記錄，錯誤率爲 9.41%；如果幾乎訓練了所有的記錄，則錯誤率爲 5.27%。找出錯誤率不再降低的點，可以確定產生錯誤率的分類器所需的訓練記錄數。在這個例子中，對 3,000 個記錄進行訓練就已足夠。可以根據準確性適當的學習曲線上的點選取一個樣本大小。

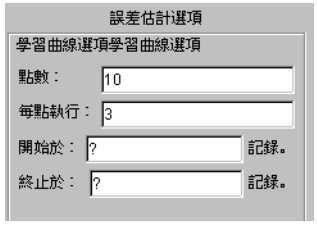
圖中顯示四種類型的點：

- 黃色的點是執行多次 (運算法則執行的次數) 後得到的實際誤差估計。
- 白色的點是平均值。
- 藍色的點 (有時像一條藍色的線) 內插在白色的點之間。
- 紅色的點顯示的是關於平均值的 95% 的置信度間隔，該平均值根據每次執行的實際誤差估計。

預設情況下，運算法則使用所有的記錄建立學習曲線，但是可以指定其他模式：

1. 使用前面的「學習曲線」、「決策樹」和「工具管理員資料目標」窗格中的客戶波動選項，按一下*進階選項*。
2. 可以在*進階選項*對話方塊中指定這些「學習曲線選項」(圖 7-10)
   - 學習曲線中的點數 — 必須大於 1
   - 每一點執行次數
   - 起始和終止點使用的記錄數。

   系統將自動計算中間每個點使用的記錄數。



7-10 學習曲線選項

如果範圍和點數都保留空白，系統將自動根據學習曲線中的點數和訓練集中的記錄總數計算。該預設值覆蓋訓練集的整個範圍。例如：假設有一個檔案含有 80,000 個記錄。如果在學習曲線中指定了 3 個點，運算法則會在 20,000、40,000 和 60,000 個記錄處產生點。通常，在較小的範圍上進行「放大」這是很有用的。例如：可以僅用 1,000 到 10,000 個記錄產生學習曲線。

增加每一點執行的次數會相應地增加執行時間，但是也會改善對誤差的估計。執行次數的預設值是 3。

可以爲所有的導入工具產生學習曲線。可以按照預設值執行，也可以確定曲線中的點數、執行的次數或記錄的範圍。表 7-1 所示爲特定選擇對學習曲線的影響。

表 7-1 對學習曲線結果的操作

要獲得該結果：	**方法：**
減小置信度間隔 (更高的準確性)	增加執行次數。 增加測試集的大小。
加速處理	使用小樣本 (請參閱第 46 頁的「資料採集」)。

關於導入工具操作的詳細資訊，請參閱*《MineSet 3.0 企業版參考指南》*. 中的「導入工具」。

應用預測模型

建立好預測模型後，可以將其應用到其他記錄，預測它們的標籤。例如：如果爲預測蝴蝶花 類型建立分類器，可以將該分類器應用到僅包含描述屬性的記錄，就會增加一個帶有預測蝴蝶花類型的新欄。

選擇模型

MineSet 產生的模型儲存到指定的伺服器目錄中。例如： *churn-dt.class* 是從客戶波動資料集中產生的決策樹模型。可以將這個分類器應用到指定到另一個資料表。使用「決策樹」爲該範例建立模型後，請繼續進行以下步驟：

1. 在「工具管理員資料轉換」窗格中按一下*應用模型*按鈕。
2. 在「測試和應用模型」對話方塊 (圖 7-11) 的「可用模型」窗格中選擇一個模型。接著，右邊的窗格會列出該分類器所需的欄名稱和類型。

如果目前記憶體載入的表格符合這些要求，則每個索引標籤底部的按鈕 (*應用模型*，*測試模型*，或*擬合資料到模型*) 將啓動。如果不符合要求，就會選取右邊清單中遺漏的欄，底部的按鈕將無效。

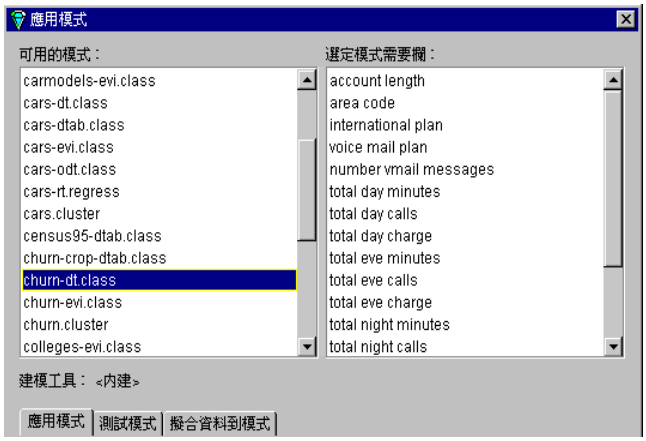


圖 7-11 「測試和應用模型」視窗：選擇分類器

應用模型

選定模型後，希望應用：

1. 在「測試和應用模型」對話方塊中，按一下「應用模型」索引標籤 (圖 7-12) 並進行下列選擇：

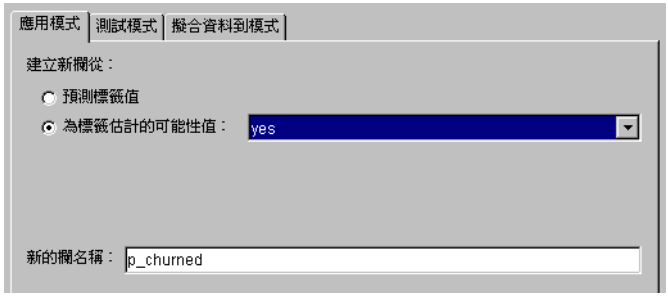
   預測的標籤值：增加要預測的欄，例如：是否某個用戶要波動。可以經由新欄看到模型使用新資料時的預測效果。

   標籤估計機率：估計每個記錄具有指定標籤值記錄的機率。從跳現式功能表中選擇可用的值。這將增加一個新欄，其中的數值表示估計的正確機率。

2. 在文字欄位中輸入新欄的名稱，例如： `p_churned`，代表用戶波動的機率。

3. 按一下「確定」。
4. 要將模型圖形化地顯示爲欄的表格，從「工具管理員資料目標」窗格中按一下「Viz 工具」，並在「工具」跳現式功能表中選擇「記錄檢視器」，然後按一下*調用工具*。

   顯示標示的欄 `churn` 和 `p_churned`，可以比較預測和結果。



7-12 應用模型面板

模型應用可能會在銷售活動中進行，其中模型是從某城市進行的商業活動中所產生，並根據該城市中的回應產生記錄 (標籤) 值的欄。可能會應用模型，然後商業活動郵件可以只傳送給另一個城市中的那些由分類器標示爲可能會作出回應的人，因此節省郵寄成本。

進行後續內容

以後的幾章將介紹查看資料的不同方式，並指出每種可視化工具特有的優勢。因此，您將發現自己從用「記錄檢視器」查看原始未轉換的資料開始，進而認識「統計」和「柱狀圖」可視化工具，然後又使用「欄重要性」工具搜尋重要的欄，以準備利用「分散」、「平板」或「地圖」可視化工具獲得已知資料的有用的可視化處理。

如果沒有想過從何開始，您可能希望從聚類開始，並讓這種方法引導知識的發現。

如果您已準備好要測試並應用一個模型，以及確定模型的準確性，可以在第 11 章，「改進預測建模」中找到幾種方法。

用決策、選項和回歸樹建模和預測

決策樹、選項樹和回歸樹看起來很相似，儘管它們的運算法則非常不同。本章介紹這三種根據樹的模型用法的類似和差別。要進一步了解導入工具使用的運算法則和可能的選項，請參閱*《MineSet 3.0 企業版參考指南》*。

本章包括以下幾個部分：

提供一些配置和資料檔案樣例，用於演示「樹狀可視化工具」的特性和功能。這些檔案在 *MineSet 3.0* 下的 *\examples* 目錄中，位於 MineSet 的安裝目錄。

決策、選項和回歸樹總覽

本部分解釋三種類型樹的用法，以及選擇使用時的一些原則。每種樹狀工具都由建立模型的導入工具和顯示結果樹的可視化工具組成。導入工具可以是分類器 (決策樹和選項樹) 或是回歸器 (回歸樹)。

決策樹

決策樹是預測模型。它使用相關的或已知的屬性值進行預測，以幫助確定標籤或未知屬性的值。對按標稱值 (通常是字元串，如「是」或「否」)，或只只在一個小範圍內取值屬性的值進行預測的任務稱爲分類。決策樹透過預測每個記錄的標籤值來對資料分類。用於分類的基本結構爲決策樹，如圖 8-1 所示。一旦「決策樹導入工具」建立資料分類器，「樹狀可視化工具」就可將它的結構顯示出來。

樹中的第一個元件是根節點，代表所有的資料。從此處，樹發出兩個或更多分支，每一個都代表具有不同指定屬性 (欄) 值的資料。例如：圖 8-1 是*汽車*資料集的決策樹可視化處理，其中的記錄是按體積分類的。第一個分支處是 169.5 立方英寸，比它小的在右邊的分支上，比它大的在左邊的分支上。樹可以在同一個屬性上拆分成多個節點。節點也可拆分成多個分支。例如：對於*蘑菇*資料集，記錄分爲可食用和不可食用的，而第一個分支又拆分成九個分支，對應於九種氣味類型。

物件將在分支末端到達葉節點，此處所有的記錄，或是說幾乎所有的記錄都具有相同的分類 (標籤)。在圖 8-1 中，分類是「決策樹」視窗底部列出的體積值的不同範圍。

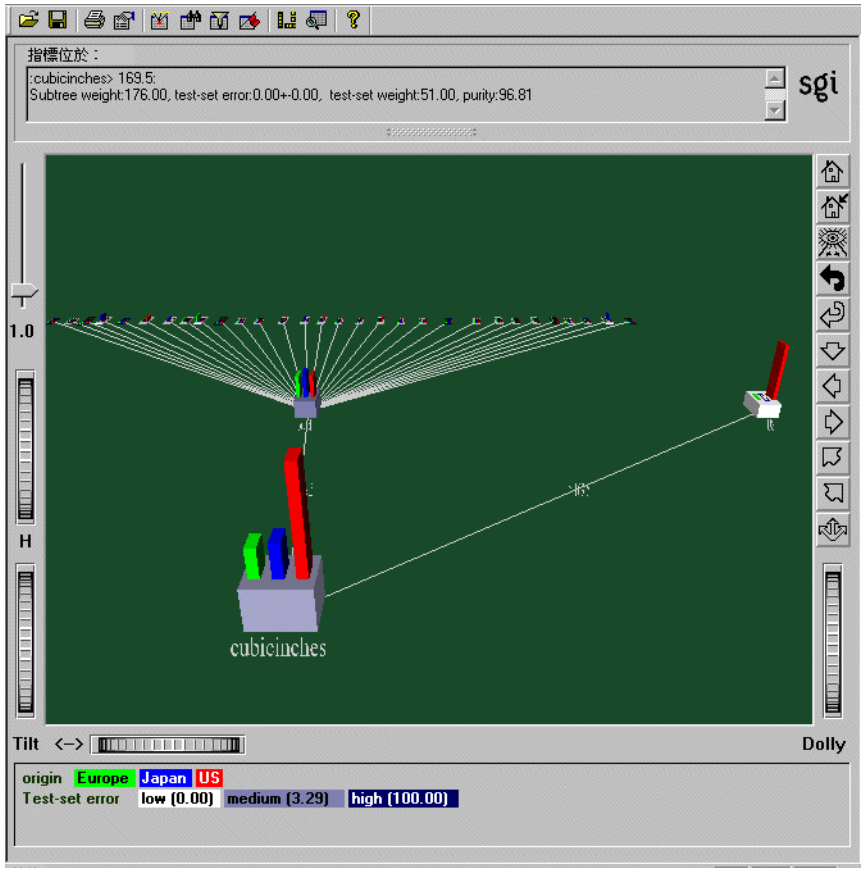


圖 8-1 汽車資料集的決策樹

選項樹

與「決策樹」分類器類似，「選項樹」分類器也爲每個記錄分配一個類別。用於分類的基本結構爲決策樹，如前面部分所述。一旦「選項樹導入工具」建立分類器，「樹狀可視化工具」就可將它的結構顯示出來。一個選項樹實際上由多個決策樹組成。它不是選取一個屬性來爲根節點拆分，而是選取多個，並爲每個屬性都產生一個決策樹。

圖 8-2 所示爲*汽車*資料集的選項樹，其中的目標是預測汽車的產地 (美國、日本或歐洲)。「選項樹」允許*選項節點*，因此擴展了「決策樹分類器」。「選項節點」提供多個選項，可以在樹中的決策節點進行選擇。例如：在圖 8-2 中，根就是具有五個選項的選項節點：體積、汽缸數、重量 (磅)、英哩 / 加侖和商標名稱。

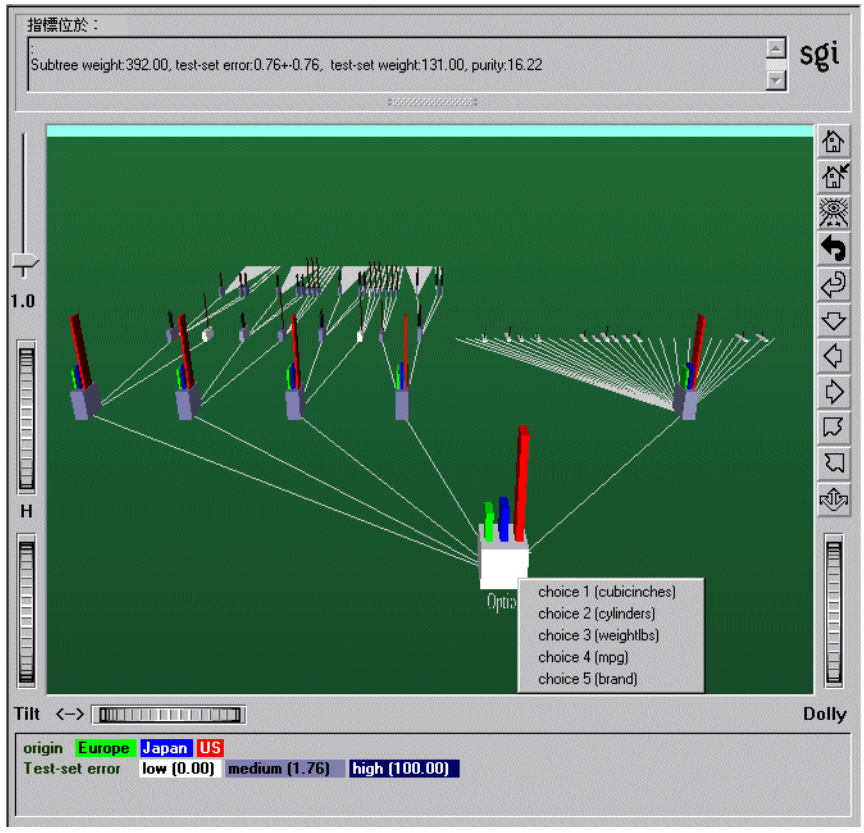
選項節點有兩個作用：

- 透過顯示可以選擇的多個選項，提高影響分類標籤確定因素的內容廣泛性。選項節點在節點處不是使用單獨一個屬性，而是提供多個選項。當檢視樹時，可以根據以前的經驗，選擇一個更易於了解或是對預測更好的選項，或是根據誤差估計選擇選項。

  在所示的*汽車*資料集中，可以選擇*汽缸*子樹，因爲它的值很少，或是選擇*重量（磅）*子樹，因爲它的估計誤差較小 (1.53)。誤差估計僅僅是估計；通常，如果兩個選項之間的誤差差別小於它們平均標準偏差的兩倍，則從統計上來看，誤差是不同的。

- 它們經由平均每個選項子樹的選票，降低建立不準確分類器的風險。每個選項產生的子樹都可以視爲一個「專家」。選項節點平均了這些專家的選票。這樣的平均可以產生錯誤率較低、更好的分類器。

  在汽車資料集中，如圖 8-2 所示，根節點的估計錯誤率爲 0.76%，比所有的子節點都低。有時，明顯的選項並不一定就是最好用的。例如：儘管商標似乎是此任務的一個明顯屬性，但是訓練集中可能不包含所有的商標 (實際上，它並沒有包含所有的商標)。對於未知的商標，「決策樹」猜測爲大多數分類 (美國)，犯了兩個錯誤。但是，當其他選項存在時，它們也被平均，並減少了錯誤。



8-2 汽車資料集的選項樹

但是，「選項樹」有兩個缺點：

- 預設設定下，建立選項樹需要的時間大約是建立決策樹所需時間的 10 到 15 倍。
- 建立的「樹狀可視化工具」檔案非常大，包含的節點是正常決策樹的 10 到 15 倍。

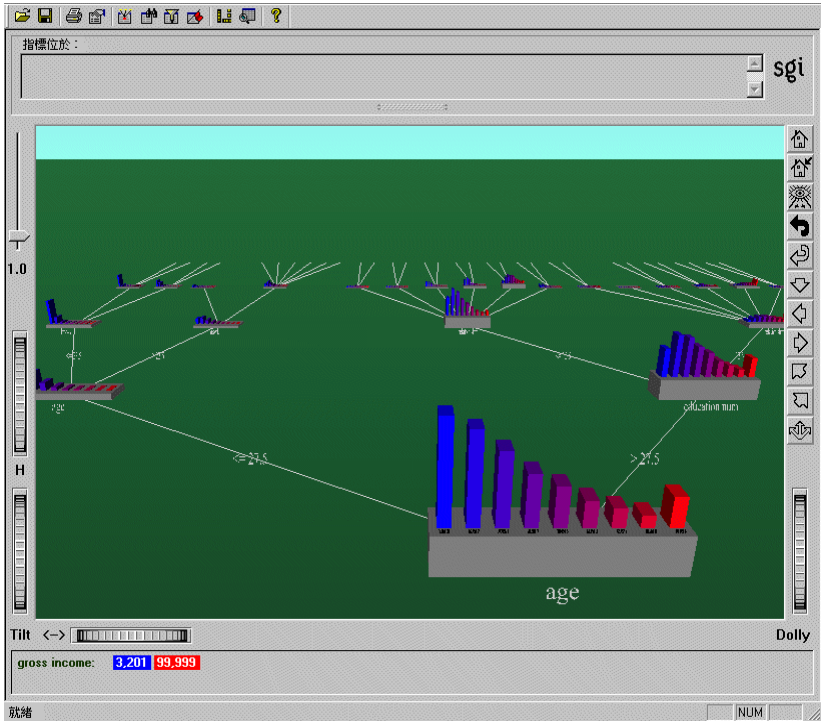
在資料集上執行「選項樹」導入工具，確定在綜合性和錯誤率方面的優點是否能夠平衡較長的導入時間的缺點，同時可以獲得額外的認識，例如：在建立決策樹時要移除或使用哪個屬性。

回歸樹

回歸樹的任務是在指定一組描述性屬性的情況下，預測連續標籤值。回歸和分類是類似的，除了在分類中，已預測標籤的取值只能是一小部分的離散值以外，例如：蝴蝶花可按類型分爲藍色、多色、純白色蝴蝶花。在回歸中，已預測的模型可以是連續範圍內的任意值，例如：個人的年薪可以是任何大於零的數。

當產生一個回歸器時，「選項樹可視化工具」會顯示相應的回歸樹。該可視化處理有助於認識回歸器及其如何進行預測。另外，它可提供對資料本身的寶貴認識。回歸器一旦產生，就可用於預測未標示記錄的標籤值。

用於回歸的基本結構爲回歸樹，如圖 8-3 所示。



8-3 成人資料集的回歸樹

執行決策、選項和回歸樹

執行樹狀工具最簡單的方法是從「工具管理員」中：

1. 在「工具管理員檔案」功能表中，連接並登錄到伺服器；再從同一個功能表中選擇「開啓新的資料檔案」，並選擇或鍵入需要的檔名。
2. 在「資料目標」窗格中，按一下「挖掘工具」(圖 8-4)，然後在下面一列的索引標籤中按一下「分類」或「回歸」：
   - 對於「決策樹」和「選項樹」，按一下「分類」索引標籤。
   - 對於「回歸樹」，按一下「回歸」索引標籤。
3. 從跳現式的「模式」功能表中選擇一個模式。關於四種可用模式的詳細資訊，請參閱《*MineSet 3.0 企業版參考指南*》中的「誤差估計」。
4. 從跳現式的「導入工具」功能表中，選擇一個導入工具。
5. 從「離散標籤」(對於「回歸樹」是「連續標籤」) 的跳現式功能表中，選擇希望用作標籤的屬性 (欄)。例如：對於*蘑菇*資料集，*可食性*是一個合理的選擇。對於*客戶波動*資料集，*已客戶波動*是一個合理的選擇。

圖 8-4　工具管理員資料目標窗格，分類索引標籤

6. 按一下*開始*，執行導入工具。「工具管理員」視窗底部的「狀態」窗格將顯示進程與統計結果。一旦 MineSet 完成計算，就會出現可視化工具視窗。

用決策樹可視化工具檢視結果

決策樹和選項樹可視化處理是類似的。它們都是由邊 (線) 連接的兩類節點組成：決策節點和葉節點。

決策節點上的標籤將指定在該節點上測試的屬性，用於測試屬性的值 (或值的範圍) 將顯示在邊上。每一個可能的屬性值都剛好符合一條邊。例如：圖 8-1 中決策樹的根測試屬性*重量磅*；從節點生出的兩條邊指定該屬性值的範圍 (小於等於 3018，和大於 3018)，所以每個可能值都能與右邊或左邊的分支相符合。

決策樹中的葉節點指定一個類別。在圖 8-1 中，如果跟隨右邊的分支從根到標示的葉 (... 16.5]，就會看到「決策樹」分類記錄的分界是：當達到 16.5 英哩 / 加侖或更少時，*重量磅*大於 3018 磅、*馬力*大於 141、以及*立方英吋*大於 311.5。

基準塊的顏色表示子樹的誤差估計：靛青顯示高誤差，灰色顯示中等誤差，白色顯示低誤差。如果沒有測試集記錄到達節點 (表示誤差估計不存在)，基準的顏色爲黑色。

每個節點頂上的垂直長條圖顯示節點處類別的分布。節點的基準具有高度和顏色。高度對應於到達該節點之訓練集記錄的權重 (如果沒有設定權重，就是記錄的數目)。通常，權重越高，每個節點處的類別分布就越可靠 (關於加權記錄的詳細資訊，請參閱第 3 章中的「加權記錄」)。

將滑鼠箭頭放在節點上面，會顯示以下資訊：

- *子樹權重*—指向節點下方的子樹中訓練集記錄的權重。該值對照到基準塊的高度。
- *測試集誤差 / 損失*—子樹誤差 (如果指定損失矩陣，即爲損失) 的估計。+/- 後面的數字是估計的標準偏差。標準偏差越高，誤差估計就越不準確。對於帶有很少記錄的葉，或者當測試集誤差接近於 0% 或 100% 時，誤差 / 損失估計和標準偏差較不可靠。

- *測試集權重*—到達節點的測試集記錄權重 (如果沒有設定權重，即爲記錄的數目)。
- *純度*—一個從 0 到 100 的數，表示節點處標籤值分布的傾斜度 (請參閱詞彙表中的項目*傾斜度*)。如果節點的記錄都來自單一的類別，則純度爲 100。如果標籤值擁有相同的權重，則純度爲 0。在修正之後計算純度 (請參閱詞彙表中的項目*修正*)。

只有「分類和錯誤估計」模式生成測試集誤差 / 損失和權重。可以使用「測試分類器」選項產生根據現存之分類器和測試集的可視化處理。

使用決策樹主視窗記錄分類

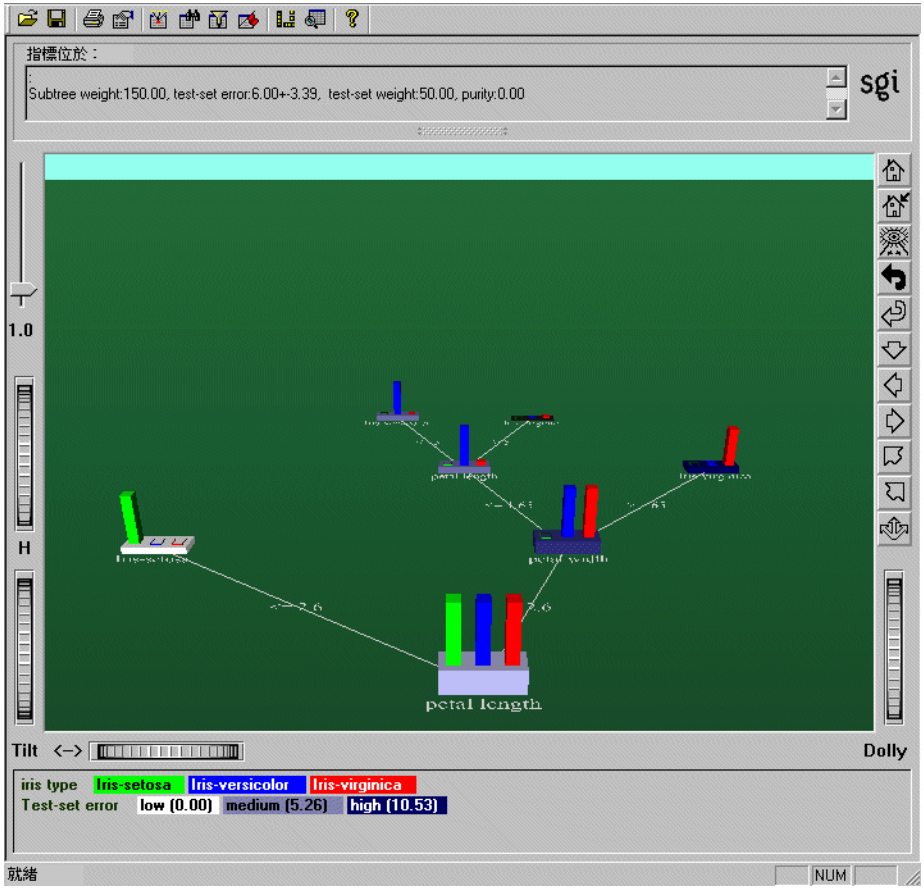
要分類具有未知標籤的記錄，可以從樹根開始，然後跟隨該記錄屬性值指示的分支。根據記錄屬性值著選擇適當的邊，就可以到達一個葉節點。與葉節點相關的標籤或分類就是預測的記錄分類。

有一些決策完成得很快，並且路徑較短。其他決策的路徑可能較長。通常，每個葉節點對應於一個規則，它是決策節點處的所有測試與從根通向它的沿線上所有值 (或值的範圍) 的結合。

爲*蝴蝶花*資料集 (圖 8-5) 產生的決策樹非常簡單，很適合演示。該檔案可從「工具管理員可視工具」功能表中開啓；選擇「 3D 可視化工具」，然後選擇 *examples**iris-dt.treeviz*。在這個樹的根部，誤差率是 6%，標準偏差是 3.39%。標準偏差高是因爲檔案小，測試集中只有 50 個記錄。純度爲 0.0，表示分布是同一的。

根的左邊子節點有 0 個測試集錯誤，且純度爲 100，因爲所有花瓣長度小於等於 2.6 英寸的記錄都是藍色蝴蝶花類；因此對於花瓣長度小於等於 2.6 英寸的所有記錄，藍色蝴蝶花的預測很可能非常準確。根的右邊子節點估計誤差爲 8.57%。在該子節點中，符合的記錄都是花瓣長度大於 2.6 英寸的，沒有任何記錄屬於藍色蝴蝶花類；因此這個類別很可能是多色蝴蝶花或純白色蝴蝶花。因爲在該節點處只有兩種可能性，因此純度比根部高 (36.91)。

決策樹葉節點將資料劃分爲共享同一個分類規則 (到達每個葉的路徑) 的聚類。經過查看葉節點，就可能看到共享相同屬性集的聚類。



8-5 蝴蝶花資料集的決策樹

其它有用的選項

「決策樹可視化工具」使用「樹狀可視化工具」進行顯示。「樹狀可視化工具」擁有多個有用的設備，例如：搜尋面板、篩選面板、全覽視窗，它可從「工具管理員」的「檢視」功能表中進入。以上所述詳見第 5 章，「用樹狀可視化工具檢視資料」和《*MineSet 3.0 企業版參考指南*》中的「樹狀可視化工具」。

決策樹也有各種選項，用於調整導入運算法則、修剪和評價測量。以上所述詳見《*MineSet 3.0 企業版參考指南*》中的「決策樹」。

用選項樹可視化工具檢視結果

決策樹和選項樹可視化處理基本上是類似的。但是，兩種類型的樹之間還是有些區別：

- 選項樹最左邊的選項是決策樹導入工具的唯一選項。當轉到右邊時，選項是按適合度得分來降冪排列的。適合度得分不一定會和顯示的測試集錯誤相符合。這是在預期中，因爲導入工具使用的是非理想的評分功能。測試集估計也有固有可變性：測試集越大，估計就越準確。
- 選項節點的錯誤率可以和它每個子節點的錯誤率不同。因爲選項節點平均了子節點的預測，因此它的錯誤率可能不同。在某些情況下，它的錯誤要比所有子節點的錯誤都要低，表現出平均的好處。
- 選項節點每個子節點上的範例 (長條中顯示的) 分布與選項節點本身的範例分布完全一樣。這是因爲選項節點沒有決策：選項顯示爲子節點。
- 「選項樹」可視化處理有一個很有用的在樹中瀏覽的特性，即：在選項節點上按一下滑鼠右鍵。將顯示子節點選項的清單。

在 *MineSet 3.0\examples\cars-odt.treeviz* 中可以發現簡單的「選項樹」價值檢驗。「選項樹可視化工具」使用「樹狀可視化工具」進行顯示。「樹狀可視化工具」具有多個有用的設備，例如：搜尋面板、篩選面板、全覽視窗。以上所述詳見第 5 章，「用樹狀可視化工具檢視資料」。

選項樹也有各種選項，用於調整導入運算法則、修剪和評估測量。以上所述詳見*《MineSet 3.0 企業版參考指南》*中的「決策樹」。

用回歸樹可視化工具檢視結果

回歸樹和決策樹可視化處理是類似的。它們都是由邊連接的兩類節點組成：決策節點和葉節點。

決策節點指定在該節點測試的屬性。用於測試屬性的值 (或值的範圍) 將顯示在邊上。每一個可能的屬性值都剛好符合一條邊。例如：圖 8-3 中回歸樹的根測試屬性年齡；從節點生出的兩條邊劃分該屬性的值 (小於 27.5 和大於等於 27.5)，以便每個可能值都能與右邊或左邊的分支相符合。如果值未知並且沒有用問號標注的邊，就會預測目前節點的平均值或中值標籤。

回歸樹中的葉節點預測一個值。在 *\examples\adult-rt.treeviz* 中可以找到此處討論的有用範例。在圖 8-3 中，如果跟隨最左邊的分支從根到標示 4002.300 的葉，就會發現回歸樹預測每周工作時間不足 35.5 小時的 19.5 歲以下的人平均總收入爲 $4002.30。

「回歸樹」中節點上的每個長條對應於連續標籤值的一個子區間。每個節點覆蓋的連續標籤值範圍可能不同。柱狀圖中每個節點上的長條表示記錄權重 (數) 在該範圍內的分布情況。長條數目由根處的記錄權重決定。最左邊的長條通常對應於最低的值。因此，每個節點的大小和中點可能不同。長條顏色表示長條覆蓋的子範圍的中間點。最大的範圍以左邊的藍色和右邊的紅色標示。僅包括有限標籤值範圍記錄的節點的柱狀圖範圍不是從藍色到紅色。

每個節點的基準塊都具有高度。高度對應於到達該節點之訓練集記錄的權重 (如果沒有設定權重，就是記錄的數目)。通常權重越高，節點處的分布就越可靠。

將滑鼠箭頭放在節點上面，會顯示以下資訊 (請參閱詞彙表中術語的定義) ：

- *子樹權重*—指向節點下方的子樹中訓練集記錄的權重。該值對照到基準的高度。
- *平均值*：連續標籤的平均值。
- *標準偏差*：連續標籤的標準偏差。標準偏差越高，模型就越不可靠。
- *中值*：連續標籤的中值。
- *絕對偏差*：連續標籤的絕對偏差。絕對偏差越高，模型就越不可靠。

使用回歸樹主視窗預測數值

要預測記錄的值，可以從樹根開始，然後跟隨該記錄屬性值指示的分支，按記錄屬性值跟隨著適當的邊，就可以到達一個葉節點。與葉節點相關的預測即爲預測的記錄值。

回歸樹的誤差估計

評估分類器時，固有測量就是誤差 (分類器預測標籤錯誤的範例數)。當提供損失矩陣時，不同類型的分類錯誤可能具有不同的相關代價。此時，損失就是固有測量。關於詳細的資訊，請參閱第 11 章中的「定義損失矩陣」。

對於預測真實值的回歸任務，沒有單獨的自然評價測量。頻繁使用的兩種測量是*標準偏差*和*絕對偏差*。標準偏差是均方誤差的平方根。絕對偏差是預測標籤值和實際標籤值差的絕對值的平均值。

其它有用的選項

「回歸樹可視化工具」使用「樹狀可視化工具」進行顯示。「樹狀可視化工具」具有多個有用的設備，例如：搜尋面板、篩選面板、全覽視窗。以上所述詳見第 5 章，「用樹狀可視化工具檢視資料」和*《MineSet 3.0 企業版參考指南》*中的「樹狀可視化工具」。

回歸樹也有各種選項，用於調整導入運算法則、修剪和評估度量。以上所述詳見*《MineSet 3.0 企業版參考指南》*中的「回歸樹」。

用決策、選項和回歸樹預測

到此爲止，本章已討論了利用已知的結果 (例如：客戶波動的或未波動的，和可食的或有毒的) 對資料進行分類，但是對於預測未知或將來的結果，決策、選項、和回歸樹可能最有用。可以根據其中已知分類的資料建立模型，然後使用該模型對其中未知分類的新資料進行分類。關於建立和應用模型的詳細資訊，請參閱第 7 章，「理解預測建模」和第 11 章，「改進預測建模」。

用決策表分類器和可視化工具建模和預測

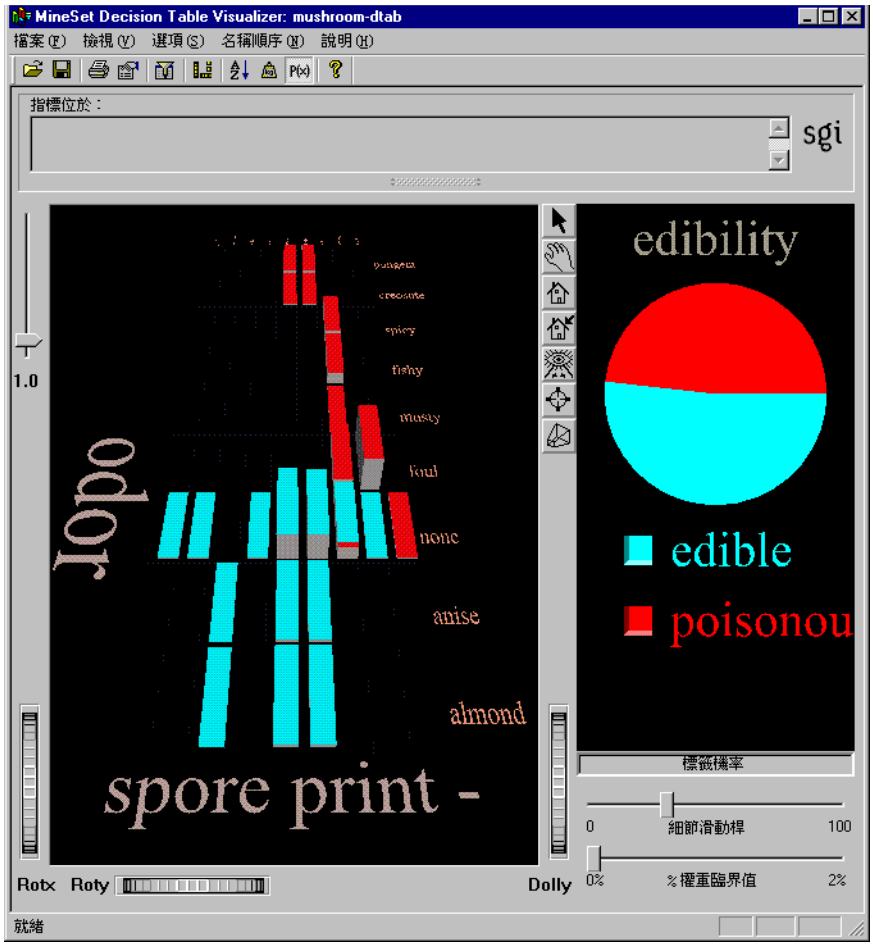
本章討論「決策表」的特性和功能，主要包括以下幾個部分：

我們提供一些配置和資料檔案樣例，用於演示「決策表」的特性和功能。這些檔案安裝在 *MineSet 3.0* 下的 *\examples* 目錄中，位於 MineSet 的初始安裝目錄。

決策表分類器總覽

決策表是執行分類的預測建模工具 (關於分類器和預測建模的詳細資訊，請參閱第 1 章，「資料挖掘和 MineSet 工具總覽」和第 7 章，「理解預測建模」)。它結合一個導入工具 (產生決策表模型的一個運算法則) 和一個可視化工具。與証據模型不同，「決策樹」模型不假設各屬性是相互獨立的。

決策表是資料的層次化分類，在每個層次上應用兩個屬性。「決策表」導入工具識別出對於分類資料最重要的屬性 (欄)，然後，可視化工具以一系列的塊狀圖圖形化顯示出產生的模型。可視化處理中的每個塊圖可以進一步劃分爲更小的塊，代表下一對最重要屬性。每個可視化處理可以包含多個等級，代表重要性依次遞減的屬性。圖 9-1 所示爲*蘑菇*資料集「決策表」可視化處理的頂級，其中確定可食性的兩個最重要屬性爲氣味和孢子印的顏色。

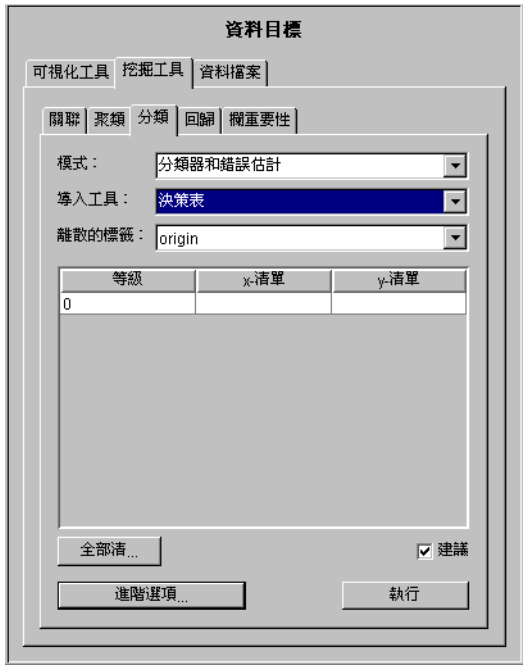


9-1 「蘑菇」資料集的決策表

執行決策表

建立「決策表」分類器最簡單的方法是：在「工具管理員」(關於更多選項，請參閱*《MineSet 3.0 企業版參考指南》*) 中選擇：

1. 從「工具管理員檔案」功能表中，連接並登錄到伺服器；
2. 從同一個功能表中選擇「開啓新的資料檔案」，並選擇或鍵入需要的檔名。
3. 在「資料目標」窗格中，按一下「挖掘工具」索引標籤 (圖 9-2)，然後從下面一列的索引標籤中選擇「分類」。



9-2 顯示分類器的工具管理員資料目標面板

4. 從跳現式的「模式」功能表中選擇一個模式。關於四種模式的詳細內容，請參閱第 7 章，「理解預測建模」。
5. 從跳現式的「導入工具」功能表中，選擇「決策表導入工具」。

6. 從跳現式的「離散標籤」功能表中，選擇想要用於標籤的屬性。
7. 在 x- 清單和 y- 清單下拉式功能表 (請參閱圖 9-2) 中，指定希望查看的屬性。可以用以下三種方法完成該操作：
   - 核取「建議」方塊，可以讓 MineSet 自動選擇屬性。MineSet 將識別預測分類最有用的屬性，並顯示它們。
   - 也可以從「資料目標」窗格的下拉式功能表中手動選擇屬性。如果需要一個起點，可以使用「欄重要性」尋找最重要的屬性 (請參閱第 3 章中的「尋找重要的欄」)。
   - 也可以選擇一部分屬性，然後讓 MineSet 選擇其餘的屬性。按照以上所述方法選擇屬性，然後選取「建議」方塊。MineSet 將對照其餘的有關屬性。
8. 按一下*開始*以執行「導入工具」。

   「工具管理員」視窗底部的「狀態」窗格將顯示進程與統計結果。也可以按一下進度對話方塊中的*取消*或*顯示可視化* 來中斷自動屬性建議過程。按一下*立即顯示 可視化* 將停止目前伺服器的計算，並使用中間的結果爲已對照的欄建構決策表。

注意： 關於其他「決策表」導入工具選項的詳細資訊，請參閱 *《MineSet 3.0 企業版參考指南》*中的「決策表」。

用決策表可視化工具檢視結果

「決策表」可視化工具有兩個窗格，左邊的「決策表」窗格和右邊的「標籤機率」窗格。圖 9-3 所示爲「*蘑菇*」資料集的「決策表」可視化處理，其中顯示最高兩層的細節。

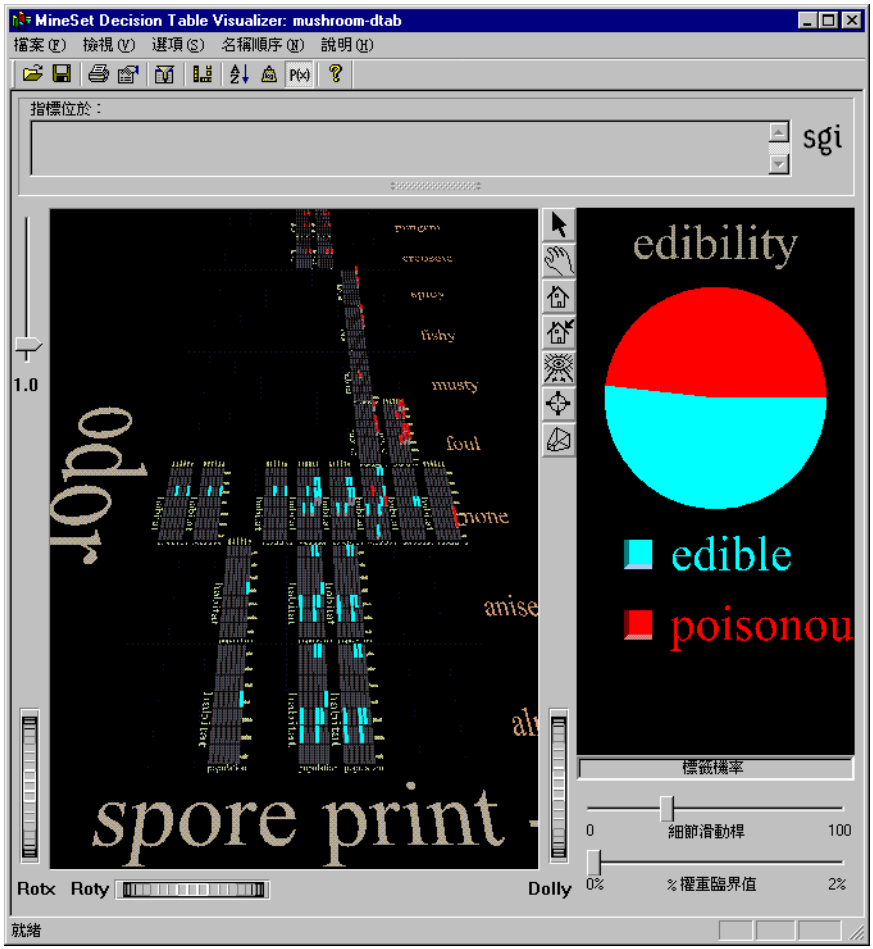


圖 9-3 蘑菇資料集的決策表可視化處理

檢視「決策表」窗格

左邊的「決策表」窗格由塊狀圖組成，即帶有彩色薄片的方塊圖，那些切片代表著具有特定屬性值記錄的標籤機率。標籤機率代表具有這些指定屬性值的記錄是某個特定類的機率。例
如：在圖 9-3 中，「決策表」執行在*蘑菇*資料集上。產生的圖表顯示具有白色孢子印和腥臭氣味的蘑菇有毒的機率是 100%。然而，如果具有白色孢子印卻沒有氣味，則有毒的機率只有 7.69%，可以食用的機率爲 92.31%。要了解這些百分比值，可以選擇左邊的一個塊，然後將滑鼠箭頭放在右邊標籤 (可食的或有毒的) 旁的彩色方塊上面。百分比值將顯示在功能表列和主視窗之間的區域內。

用滑鼠右鍵按一下，可以將「決策表」窗格中的元件進一步細分爲越來越小的塊狀圖，此過程叫做細化下尋。要更仔細地檢查塊狀圖：

- 要了解目前細節等級上兩個屬性的值，可以將滑鼠箭頭 (在選擇模式下) 放在所需的塊狀圖上。屬性值和記錄權重將顯示在功能表列和主視窗之間的區域內。塊狀圖的高度與權重成正比 (請參閱詞彙表中的項目*權重*)。
- 要細化下尋到下一個細節等級，可以將滑鼠箭頭放在需要的塊狀圖上，然後按一下滑鼠右鍵，或按一下背景在所有區塊上進行整體細化下尋。圖 9-5 所示爲對*蘑菇*資料集細化塊的特寫檢視。
- 要低層次中概化上尋回去，可以在按一下滑鼠右鍵的同時按住 Ctrl 鍵 (或使用滑鼠中
  鍵)。可以概化上尋一個單獨的區域，或是在背景上按住 Ctrl 鍵並同時按一下滑鼠右鍵
  (或使用滑鼠中鍵) 進行整體概化上尋。
- 要了解更高一層圖塊代表的屬性值和記錄權重，可以將滑鼠箭頭置於基準塊 (塊狀圖下面的灰色圖塊) 上。
- 要了解某圖塊的定義值，可以先選中最粗細節級別的基準塊，然後逐層選中相應層次的基準塊，直到目前層次的上一層爲止，相關的一對屬性值將出現在功能表列下面的選擇面板中。

在每個細節等級上，一個屬性的名稱顯示在塊狀圖陣列的左邊，它的值顯示在右邊；另一個屬性的名稱顯示在陣列的底部，它的值顯示在頂部。如果屬性的總數爲奇數，則最低的等級將只有一個屬性。

「標籤機率」窗格

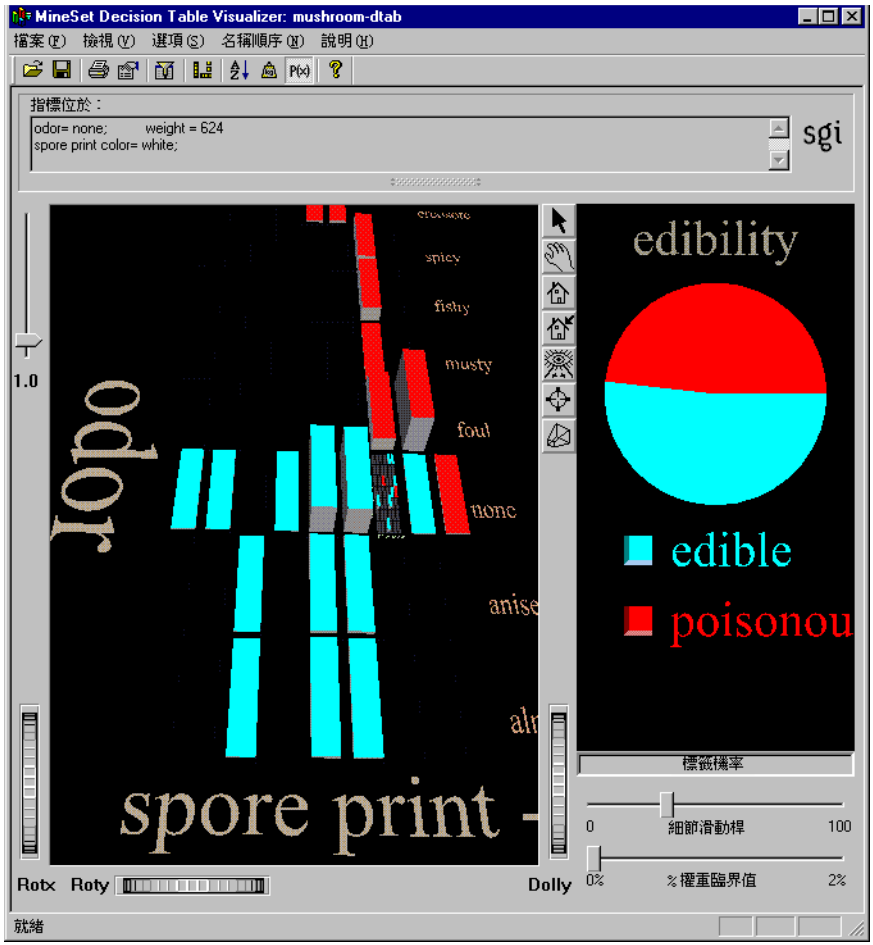
「標籤機率」窗格 (在「工具管理員」視窗的右側) 將顯示整個資料集標籤機率的圓餅圖。在圓餅圖下顯示有所有分類標籤的清單。

如果希望更仔細地檢查標籤機率，可以按照以下步驟進行：

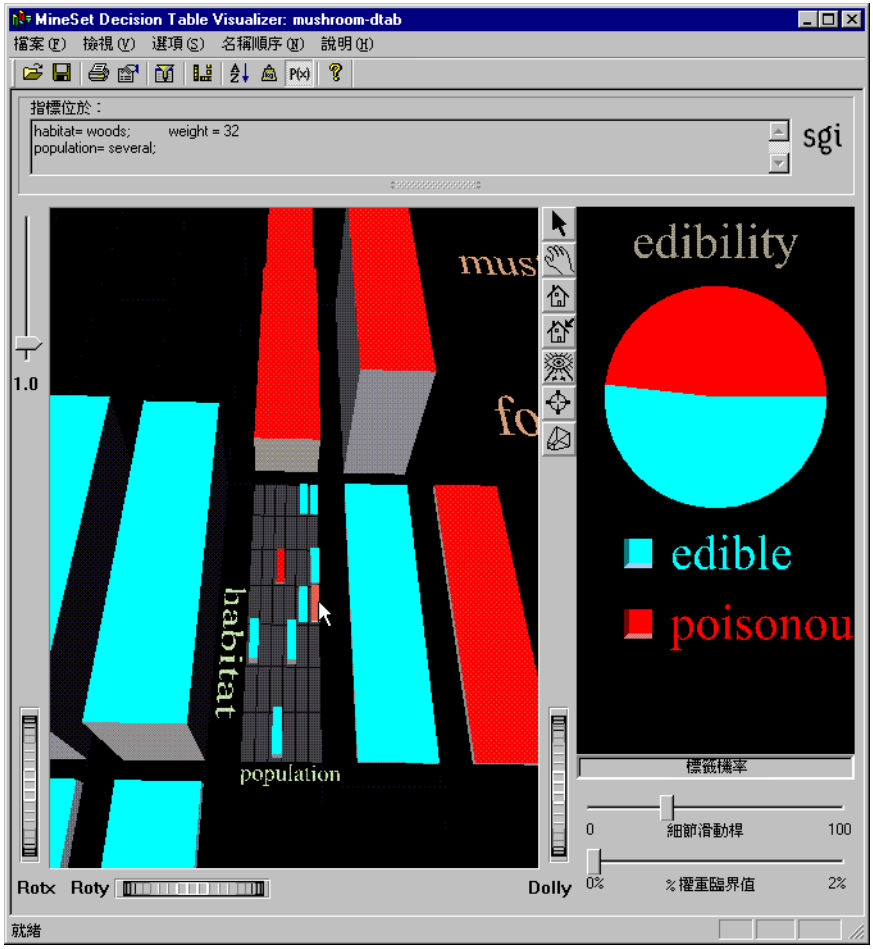
- 要了解指定屬性集合的標籤機率，可以按一下所需的塊狀圖。「標籤機率」窗格中的圓餅圖顯示屬性集合記錄的標籤機率，這個屬性集合由塊狀圖代表。
- 要了解屬性組合的標籤機率，可以使用 Ctrl 按一下所需的塊狀圖。按一下 Ctrl，可以選擇許多場景中不同細節等級上的區塊。可視化工具視窗右邊的「標籤機率」窗格將顯示選定記錄集的標籤機率。
- 要了解對每個標籤指定的百分比和置信度，可以將滑鼠箭頭放在所需標籤旁的彩色方塊上。數字將顯示在功能表列和「決策表」窗格之間的區域內。

「決策表」範例

在這個介紹如何使用「決策表」的範例中，假設指出哪些蘑菇可以安全地食用，哪些蘑菇是有毒的。開始時，需要從*蘑菇*資料集中建立一個決策表 (請參閱圖 9-1)。左邊的窗格顯示決策表中最高細節等級。最高等級選擇了氣味和孢子印，因爲它們是由「欄重要性」運算法則確定的兩個最重要屬性。只有一個頂級區塊顯示具有多個分類，其屬性值氣味等於無，孢子印顏色等於白色。當您細化下尋到下一個等級時 (在圖上按一下滑鼠右鍵)，將表示屬性生長地和總體 (請參閱圖 9-4)。該細化下尋區域的特寫如圖 9-5 所示。在這個等級上，混合的區塊不存在。因此，如果發現具有白色孢子印並且沒有氣味的蘑菇與其他一些蘑菇一起生長在樹林中，就可以確定它是有毒的。



9-4 在蘑菇資料集上細化下尋的決策表



9-5 細化下尋決策表的特寫

用決策表預測

到此爲止，本章已討論了利用已知的結果 (例如：可食的或有毒的) 對資料進行分類，但是對於預測未知的或將來發生的結果，「決策表」可能最有用。可以根據具有已知類資訊建立模型，然後使用該模型對未知類資訊新資料進行分類。關於建立和應用模型的詳細資訊，請參閱第 7 章，「理解預測建模」和第 11 章，「改進預測建模」。

關於功能表的詳細內容，請參閱*《MineSet 3.0 企業版參考指南》*中的「決策表」。

用証據分類器和可視化工具建模和預測

本章討論「証據分類器」和「証據可視化工具」的特性和功能，主要包括以下幾個部分：

提供一些配置和資料檔案樣例，用於演示「証據可視化工具」的特性和功能。這些檔案在 *MineSet 3.0* 下的 *\examples* 目錄中，位於 MineSet 的初始安裝目錄。

証據分類器和可視化工具總覽

決策表是執行分類的預測建模工具 (關於分類器和預測建模的詳細資訊，請參閱第 1 章，「資料挖掘和 MineSet 工具總覽」和第 7 章，「理解預測建模」)。它結合一個導入工具 (產生 Naive Bayes 模型的運算法則) 和一個可視化工具。與「決策樹」模型不同，「証據」模型假設屬性是獨立的，它甚至在屬性不是獨立時仍能產生合理的結果。

「証據可視化工具」顯示來自分類器的資訊，如圖 10-1 所示。可視化工具有助於認識分類的指定屬性 (欄) 之重要性。

「証據可視化工具」視窗

一旦建立分類器，結果就會顯示在「証據可視化工具」視窗中。最初，左邊窗格包含分類器使用的每個屬性的塊狀圖表列 (圖 10-1)。塊狀圖類似於其中顯示比例的圓餅圖，除了它是帶有矩形薄片的方塊。可以在塊狀圖 (代表証據)、圓餅圖 (代表機率) 和長條圖 (提供選取標籤的詳細資訊) 之間進行切換 (關於詳細的資訊，請參閱第 161 頁的「用証據可視化工具檢視結果」)。

圖 10-1 所示爲証據檢視，圖 10-2 所示爲機率檢視。

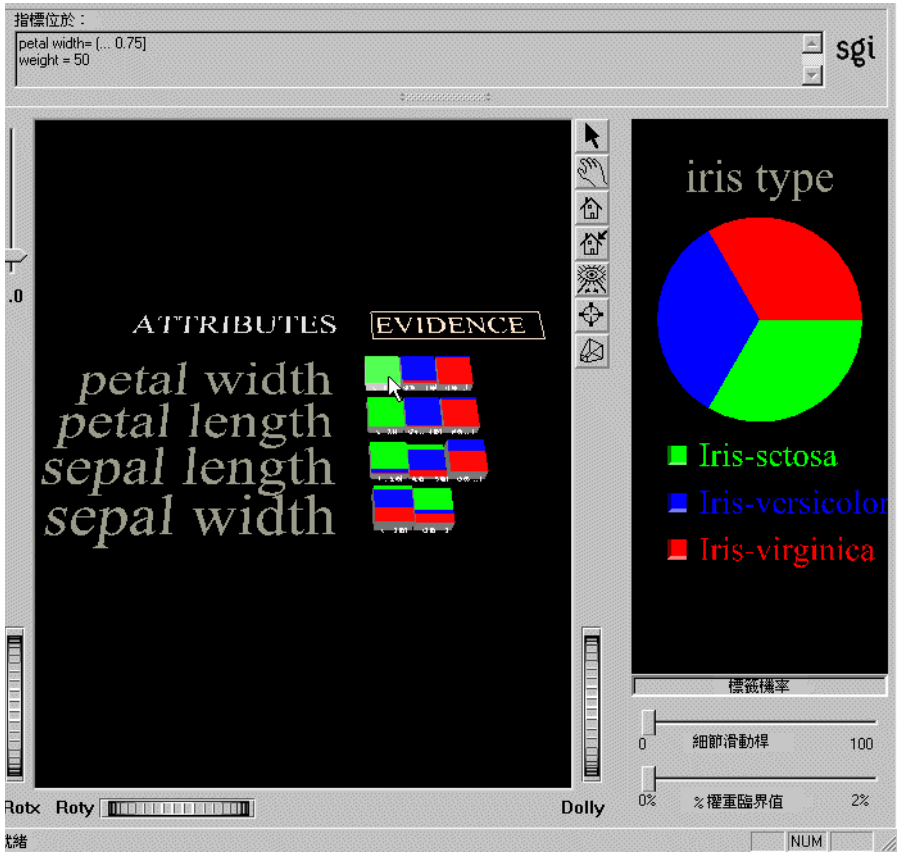


圖 10-1 應用到蝴蝶花資料集的証據可視化工具

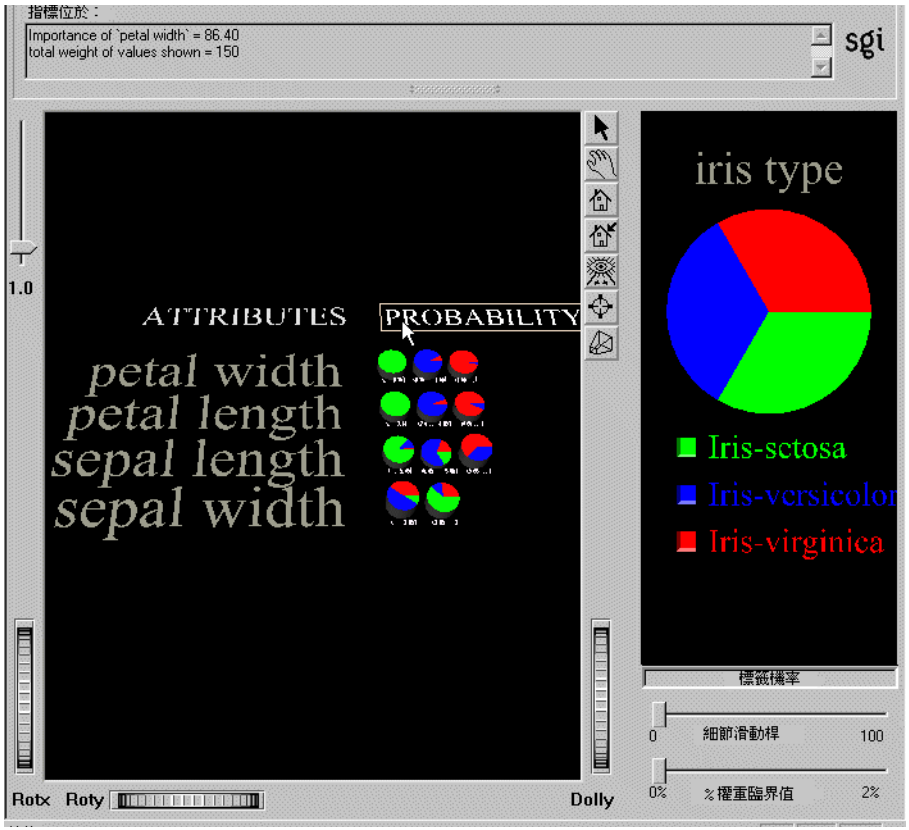


圖 10-2 顯示機率的証據可視化工具

執行証據工具

執行「証據」工具最簡單的方法是：在「工具管理員」(關於更多選項，請參閱 *《MineSet 3.0 企業版參考指南》*) 中：

1. 從「工具管理員檔案」功能表中，連接並登錄到伺服器。
2. 從同一個功能表中選擇「開啓新的資料檔案」，並選擇或鍵入需要的檔名。
3. 在「資料目標」窗格中，按一下「挖掘工具」索引標籤 (圖 10-3)，然後從下面一列的索引標籤中選擇「分類」。

4. 從跳現式的「模式」功能表中選擇一個模式。關於四種模式的詳細資訊，請參閱第 7 章，「理解預測建模」。
5. 從跳現式的「導入工具」功能表中，選擇「証據」。
6. 從跳現式的「離散標籤」功能表中，選擇想要用於標籤的屬性。可食性對於蘑菇資料集是一個很好的屬性。
7. 按一下*開始*以執行「導入工具」。

「工具管理員」視窗底部的「狀態」窗格將顯示進程與統計結果。

注意： 關於其他「証據」導入工具選項的詳細資訊，請參閱 *《MineSet 3.0 企業版參考指南》*中的「証據導入工具選項」。

圖 10-3 「工具管理員資料目標」窗格

用証據可視化工具檢視結果

一旦執行「証據可視化工具」，就可用多種方法檢視結果。在可視化工具視窗的左邊窗格中，可以在証據檢視、機率檢視和長條檢視之間切換：

- 証據檢視顯示塊狀圖，代表証據 (請參閱第 162 頁的 「証據檢視」)。
- 機率檢視顯示圓餅圖，代表機率 (請參閱第 165 頁的 「機率檢視」)。
- 長條檢視顯示長條圖，代表對特定標籤有利和不利的証據 (請參閱第 166 頁的 「長條檢視」)。

「標籤機率」窗格 (可視化工具視窗的右邊) 顯示一個圓餅圖，代表整個資料集中標籤值的分布 (請參閱「檢視証據檢視中的機率」)。該窗格是爲左邊窗格中的全部三種檢視 (証據、機率和長條) 而顯示。

証據檢視

當「証據可視化工具」第一次執行時，主窗格將顯示証據檢視 (圖 10-4)。

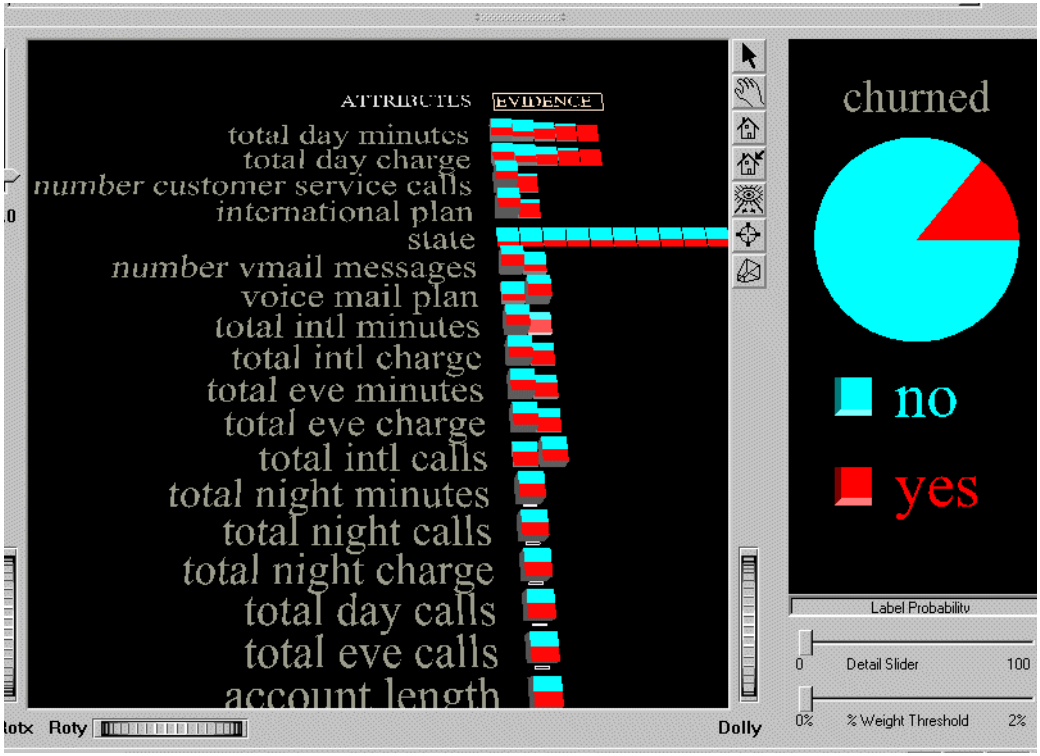


圖 10-4 用于客戶波動資料集的証據可視化工具塊狀圖

解釋証據檢視

「証據檢視」爲資料集中的每個屬性或欄顯示一列塊狀圖。每列所對應的每個離散屬性值都有一個區塊。如果屬性值不是離散的，而是連續的，MineSet 自動將值分爲幾個範圍或組 (例如： 0 到 5、5 到 10，等等)，劃分方式是使相鄰塊之間的差別最大化的方式。塊的高度與具有該屬性值的記錄權重成正比 (請參閱詞彙表中的項目*權重*)。如果沒有設定權重屬性，則高度就代表記錄的數目。

在塊狀圖中，薄片的尺寸代表條件機率分布 (証據)，而不是直接機率。例如：在*客戶波動*資料集中，如果一個區塊的一半爲「是」，一半爲「否」，就表示該塊代表的用戶特定屬性值與總體的波動機率相同。換言之，該屬性值沒有預測值。相反地，如果一個區塊的 90% 爲「是」，10% 爲「否」，即表示具有該特定屬性值的用戶比總體更可能客戶波動。但是，它並*不*表示：用戶有 90% 客戶波動的機率，因爲整個機率取決於多個變數，而不只是這一個。關於更詳細的說明，請參閱*《MineSet 3.0 企業版參考指南》*中的「証據導入工具」。

列的次序表示預測標籤時屬性的相對重要性。每個屬性都分配了一個介於 1 到 100 之間的重要性值。將滑鼠箭頭放在場景中屬性的名稱上，就可以看到這個值，該資訊顯示在功能表列和可視化工具窗格之間的區域內。例如：在圖 10-5 所示的*蘑菇*資料集中，蘑菇的氣味是可食性的標誌，因此氣味屬性排在清單的首位，重要性值爲 90.69。相反地，莖的形狀根本無法表示可食性，因此莖的形狀屬性排在清單的最後，重要性值爲 .75。

菌蓋的顏色屬性也是一個有趣的例子。它的重要性排在後面，因爲大部分蘑菇的菌幕是白色的。帶白色菌蓋的蘑菇幾乎被平均分成了可食的和有毒的，非常接近先驗機率 (即：資料集中任何蘑菇有毒的一般機率)。因此，從統計學的角度來看，該屬性沒有爲任何一類增加証據，因此重要性排在後面。但是，對於棕色、橙色或黃色的菌蓋顏色 (菌蓋顏色列中的第一、二、四塊)，圖片就完全不一樣了。棕色或橙色菌蓋的蘑菇 100% 可以食用。黃色菌蓋的蘑菇 100% 是有毒的。儘管聽起來很有說服力，但是菌蓋不是白色的蘑菇非常少，因此菌蓋顏色不能成爲好的區別特徵。

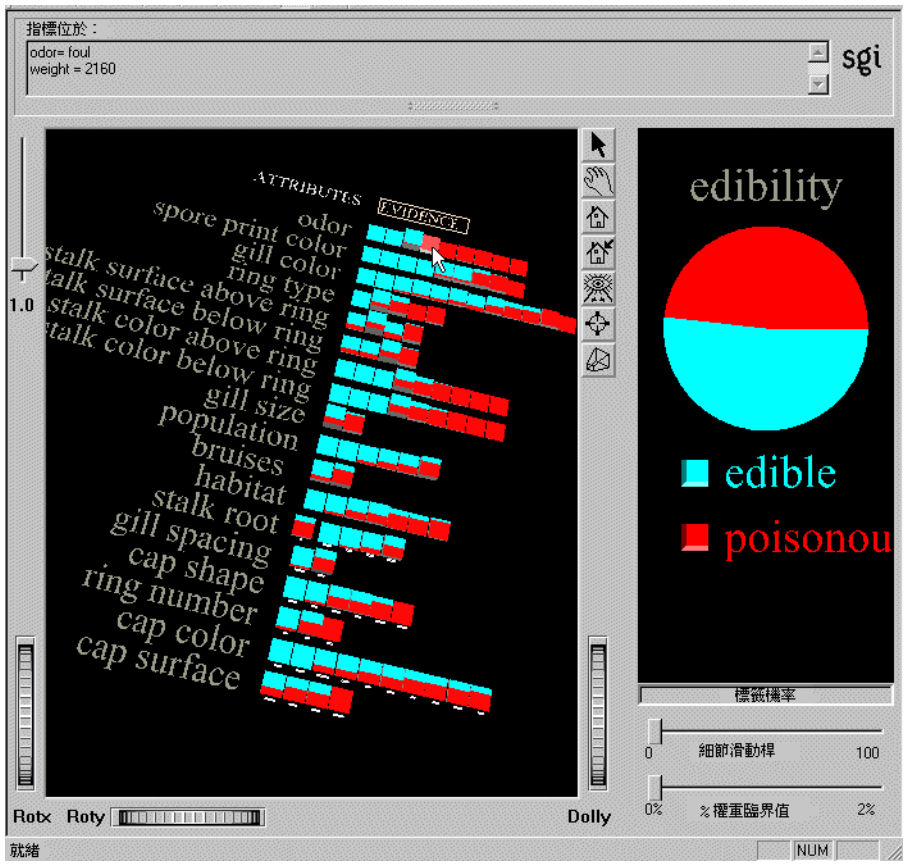


圖 10-5 証據可視化工具中的蘑菇資料集

檢視証據檢視中的機率

當「証據可視化工具」第一次開啓時，「標籤機率」窗格將顯示整個資料集的標籤機率。例如：在圖 10-4 中，使用*客戶波動*資料集，具有「是」的客戶波動標籤之隨機記錄的機率爲 14.14%。如果將滑鼠箭頭放在標籤旁的彩色方塊上，該標籤的機率和權重將顯示在視窗的左上方。

如果希望查看具有某個屬性特定值的記錄之機率圓餅圖，可以按一下代表該值的塊 (滑鼠必須處於選擇模式以選擇屬性)。「機率」窗格只會顯示符合選定值的記錄之機率圓餅圖。要取消對圖表的選定，可以按一下主窗格背景的任何地方。

要查看具有特定屬性組合的記錄之機率，可以用 Ctrl 按一下它們來選擇多個塊。「機率」窗格將顯示具有該種屬性組合的記錄之機率圓餅圖。因爲要計算每個屬性的機率，就好像它們之間相互獨立一樣，因此組合機率可能不完全準確。會出現一個警告提醒您這一點。將它們用作估計。只有在屬性真正獨立時，它們才是完全準確的，但這樣的情況很少。

有時在選擇屬性組合時，可能會選擇一些相互排斥的屬性。在這種情況下，「標籤機率」窗格將顯示一個灰色的圓餅。這表示具有該特定組合值的記錄很少或根本沒有。

要重新設定「標籤機率」窗格以顯示先驗機率，可以按一下主窗格背景上的任何位置。

機率檢視

機率檢視 (圖 10-6) 顯示代表機率的圓餅圖，而不是塊狀圖。要從証據檢視進入機率檢視，可以按一下圖表頂部白色輪廓方塊中的「証據」一詞，或在可視化工具「檢視」功能表中切換「顯示爲証據」核取方塊。要回到証據檢視，按一下「機率」一詞或再次切換核取方塊。

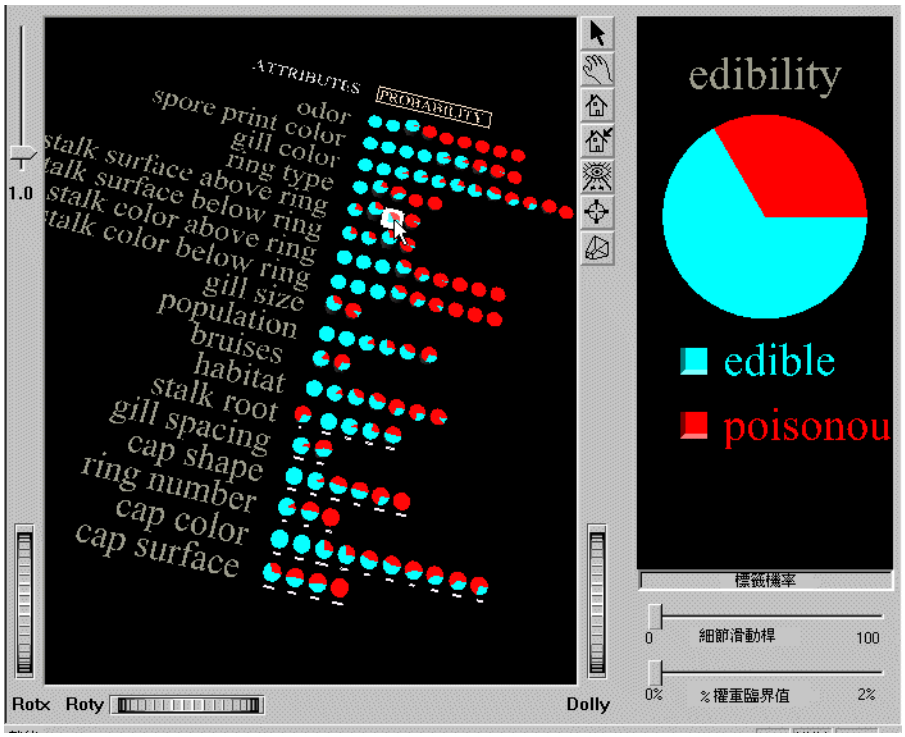
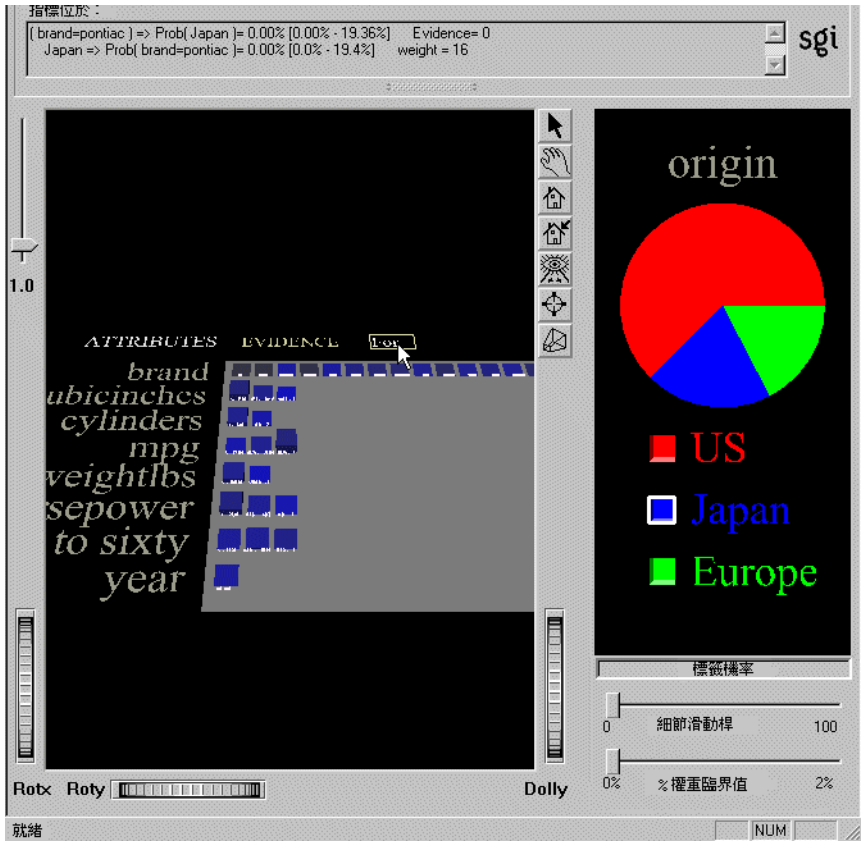


圖 10-6 証據可視化工具圓餅圖

長條檢視

長條檢視顯示有利於 (或不利於) 選定標籤值的証據。要進入長條檢視，可在「標籤機率」窗格中按一下所需標籤旁的彩色方塊。接著，主視窗將顯示長條圖 (圖 10-7)。每個長條的高度代表有利於選定標籤值的証據。要結束長條檢視，可以再按一下彩色方塊。

按一下長條檢視圖表頂部的白色輪廓方塊中的「有利」一詞，可以切換到「不利証據」模式。要回到「有利証據」模式，可以按一下「不利」一詞。



10-7 標籤值「日本」被選定的汽車資料集

當確定哪些值對預測特定標籤值最有幫助時，「有利証據」會很有用。相反情況下，即：當確定哪些值對標籤*不會*出現的預測最有幫助時，「不利証據」會很有用。關於這些值的計算方法之詳細技術說明，請參閱*《MineSet 3.0 企業版參考指南》*中的「証據導入工具」。

長條的顏色範圍可從完全飽和的顏色到完全灰色。長條的灰度以 95% 的置信度間隔爲準。而這又取決於該值的權重。因此，近似於灰色的長條具有較低的權重和大的置信度間隔。這些長條的高度不可能準確。相反地，完全飽和長條的高度和相應的証據值是可以信賴的，因爲它是根據較高的權重，代表許多記錄。當長條反白標示時，確實的記錄數 (*權重*) 反映在本文輸出行中。

改變証據可視化工具檢視

可以使用表 10-1 所示的各種操作加以檢查不同檢視中的結果。

表 10-1 操作「証據可視化工具檢視」

目的：	**操作：**
在塊狀圖 (証據) 和圓餅圖 (機率) 之間切換	按一下白色輪廓的「証據」或「機率」一詞，或切換「檢視」>「顯示爲証據」。
在長條圖和塊狀圖或圓餅圖之間切換	在「標籤機率」窗格中按一下所需標籤旁的彩色方塊。
在長條圖和塊狀圖或圓餅圖之間切換	在「標籤機率」窗格中按一下所需標籤旁的彩色按鈕。
在長條檢視中的「有利」和「不利」之間切換	按一下「有利」或「不利」旁的白色輪廓方塊。
顯示特定區塊或圓餅的屬性值和紀錄權重	將滑鼠箭頭放在塊或圓餅上。
顯示標籤機率	將滑鼠箭頭放在「標籤機率」窗格中標籤旁的方塊上。
爲具有特定屬性值的新記錄 (在右邊面板) 顯示標籤的預期分布	按一下代表該值的塊。
顯示屬性重要性和總權重	將滑鼠箭頭放在左邊的屬性名稱上。

表 10-1 操作「証據可視化工具檢視」

目的：	**操作：**
爲帶有特定屬性值組合的記錄取得估計的標籤機率分布	經由 Ctrl- 按一下選擇兩個或多個塊，並檢查「標籤機率」窗格中的圓餅圖。
清除選擇	按一下左邊窗格背景區域的任何地方。
查看具有選定屬性值的所有記錄	按一下代表需要的屬性值組合的區塊，然後從「選擇」功能表中選擇「顯示原始資料」。
查看*除*具有選定屬性值記錄以外的所有記錄	在「選擇」功能表中核取「*互補*追溯」，然後從同一個功能表中選擇「顯示原始資料」。
在圖表之間放大和縮小高度差別	使用主視窗左側的高度滑動桿。
減少 (或增加) 顯示的屬性列數，從最少 (或最多) 的預測屬性開始	使用細節滑動桿。移除重要性小於滑動桿值的屬性。
篩選出低計數或低權重的屬性值	使用權重臨界值滑動桿。移除權重小於滑動桿指示的百分比的值。
爲一些屬性值已知的新記錄預測分類標籤	按一下對應於這些屬性值的圖表。右側圓餅圖中最大的部份對應於預測的分類。

用証據分類器預測

到目前爲止，討論主要集中於使用「証據分類器」和「証據可視化工具」檢查標籤已知的資料。實際上，這個工具在預測未知資料中的標籤時最有用。在這個例子中，使用已知的資料建立模型 (請參閱 *《MineSet 3.0 企業版參考指南》*中的「訓練集」) 並儲存它，然後將其應用到標籤未知的資料上。請參閱 *《MineSet 3.0 企業版參考指南》*中的「應用模型」。

關於可視化工具功能表的詳細內容，請參閱 *《MineSet 3.0 企業版參考指南》*中的「証據可視化工具功能表」項目。

改進預測建模

本章建立在前面章節所發展出的概念上，使用「工具管理員」檢查分類中出現錯誤的代價，並將資料挖掘操作按實際需要進行調整。MineSet 導入工具的進階選項可以提高準確性，補償錯誤的代價，並且確定從何時起，追加投資將不會再帶來任何收益。本章包括以下部分：

在本指南的線上版本 (可從「說明」功能表中進入) 中可以看到彩色的螢幕插圖。

確保模型的準確性

可以用不同的方法來檢查已建立的模型的準確性。在第 121 頁的「評估預測模型」已經討論幾種方法。本部分繼續討論使用不同資料來源驗証模型的方法。如果還沒有模型，請回到第 7 章，「理解預測建模」，取得建立模型的說明。

測試模型

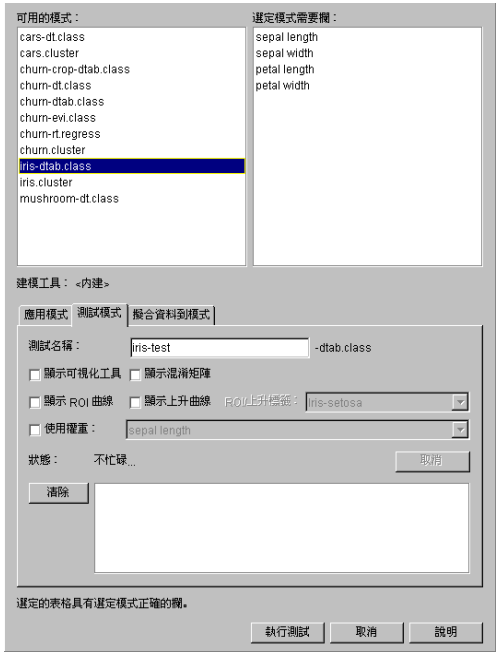
如果先前已建立了模型，MineSet 允許使用「測試模型」工具在目前使用的資料上測試它。資料集中的欄名稱和類型必須與建立模型時使用的相同。

與「應用模型」(詳見第 7 章中的「應用預測模型」) 不同，「測試模型」要求表的名稱和類型也要與建立模型時使用的相同。原因是測試必須從已具有正確答案的資料開始工作。將分配測試檔案一個預設名稱，結尾爲 *-test*，如 *iris-test*。

下面的範例從*蝴蝶花*資料集中建一個測試檔案。一旦登錄到伺服器上，就可以選擇資料集 (範例檔案爲 *MineSet 3.0\data\iris.schema*，位於 MineSet 最初安裝的目錄中)。

要在目前的資料集上測試模型，請按以下步驟進行：

1. 在「工具管理員資料轉換」窗格中按一下*應用模型*按鈕。
2. 在「測試和應用模型」對話方塊中，按一下「測試模型」索引標籤。
3. 從對話方塊頂部出現的可用模型清單中選擇模型的名稱 (如： *iris-dt.class*)，然後會出現預設的測試檔名 (圖 11-1)。



11-1 測試模型面板

如果目前表格中遺漏了某些欄，則它們的名稱會顯示在右邊「選定模型需要欄」清單中。請選擇其他模型或恢復被遺漏的欄。「測試模型」面板將提供以下選項：

表 11-1 測試模型面板選項

選擇	操作
測試名稱	顯示測試檔案的名稱，可以更改。
顯示 可視化 工具	顯示具有測試集表的分類器可視化處理。
顯示混合矩陣	顯示根據表格記錄的模型混合矩陣。請參閱第 178 頁的「顯示混合矩陣」。
顯示 ROI 曲線	顯示使用指定標籤值的模型 ROI 曲線。請參閱第 185 頁的「使用投資回報曲線尋找銷售利潤」。
顯示上升曲線	顯示使用指定標籤值的模型上升曲線。請參閱第 183 頁的「用上升曲線和 ROI 曲線評測模型」。
ROI/ 上升標籤	顯示可用標籤值的跳現式功能表。
使用權重	從加權記錄的跳現式功能表中選擇可用的屬性。請參閱第 3 章中的「加權記錄」。

4. 按一下*執行測試*。「測試模型」面板底部的本文欄位中將顯示結果，同時也顯示選定的可視化工具選項。

在這個例子中，整個 *iris.schema* 資料集執行在事先建立好的模型 (*iris-dt.class*) 上，在 150 個測試實例中發現了 4 個錯誤。

將資料擬合到模型

當有新資料時，可以使用「將資料擬合到模型」面板 (圖 11-2) 將目前表格中的資料擬合到事先建立好的模型中。雖然，這樣做會產生一個與原來模型結構相同的新模型；但是，新模型使用新表格資料更新機率估計 (請參閱「在誤差估計中修正」)。因爲表中的所有資料都擬合到模型的結構中，因此誤差估計不存在。「將資料擬合到模型」不能用於使用推進建立的模型。請參閱下面的「用推進提高準確性」。如果需要評測建立在獨立測試集 (與適合資料無關) 上的新模型的性能，可以使用「測試模型」工具。

要將資料擬合到現有的模型，請按以下步驟進行：

1. 在「工具管理員資料轉換」窗格中按一下*應用模型*按鈕。
2. 在「測試和應用模型」對話方塊中按一下「將資料擬合到模型」索引標籤。(圖 11-2) 在表 11-2 中會出現擬合到模型選項。

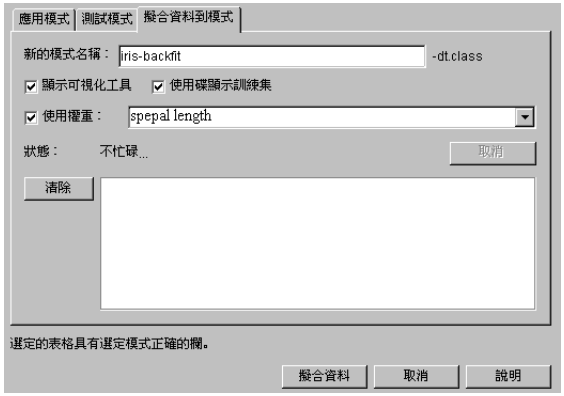


圖 11-2 「將資料擬合到模型」面板

「將資料擬合到模型」面板 (圖 11-2) 中有以下選項：

表 11-2 將資料擬合到模型選項

選擇	操作
顯示 可視化 工具	顯示新模型的可視化處理 (僅適用於根據樹的模型)。
用磁碟顯示訓練集	設定磁碟用以顯示可視化處理中訓練集資料所占的比例。
使用權重	從跳現式功能表中選擇作爲記錄權重的屬性。

在誤差估計中修正

當使用一個記錄集對模型修正時，就會更新了機率估計，但不會改變模型的結構。修正類似於將資料擬合到模型，預設情況下的模式是使用修正。修正在下列情況中很有用：

- 從小訓練集中建立一個模型，然後用大資料集對其修正。通常這比導入更大的模型結構要快。
- 需要更高的準確性。當計數、權重和機率顯示在模型的結構中時，它們反映所有的資料，而不僅是訓練集內的部分資料。

開啓修正：

1. 在「工具管理員」視窗的「資料目標」窗格中，按一下「挖掘工具」索引標籤。
2. 按一下「分類」索引標籤。
3. 選擇「分類器和錯誤估計」模式，然後按一下*進階選項按鈕*，開啓「進階選項」對話方塊。按一下「誤差估計選項」下的「修正」測試集核取方塊。當「推進」開啓時，無法使用核取標記。

用推進提高準確性

在一些情況下，錯誤率在選擇模型時是最重要的準則，因此需要開啓推進。推進是一種算法，它建立多個不同的模型並使用加權投票方式將模型的預測結合起來。推進將導入過程集中於相對於模型更爲困難的資料範例上，進而提高模型的準確性。但是，被「推進」的模型無法可視化。

要在選取的模型上啓動推進：

1. 在「工具管理員資料目標」窗格上，按一下「挖掘工具」。
2. 按一下「分類」索引標籤。
3. 在模式中按一下*進階選項按鈕*，開啓「進階選項」對話方塊。在「導入工具」選項部分按一下「推進」(無 可視化)。

推進經常但不總是會提高準確性。推進的模型無法可視化，但是仍然可以查看混淆矩陣、上升曲線、學習曲線和 ROI 曲線。推進是計算性很強的過程，通常需要的執行時間是不帶有推進的相對應之導入工具執行時間的 25 倍。推進的模型不能使用修正。

注意： 雖然推進使用的理論不是因爲超出雙類問題而產生，MineSet 還是可以讓您對具有任意多個值的標籤進行推進。此時，推進不太可能改進錯誤率。

推進反複地爲訓練集指定新的權重，並爲重新設定權重的訓練集導入模型。請參閱*《MineSet 3.0 企業版參考指南》*。以獲得關於推進的進一步資訊。

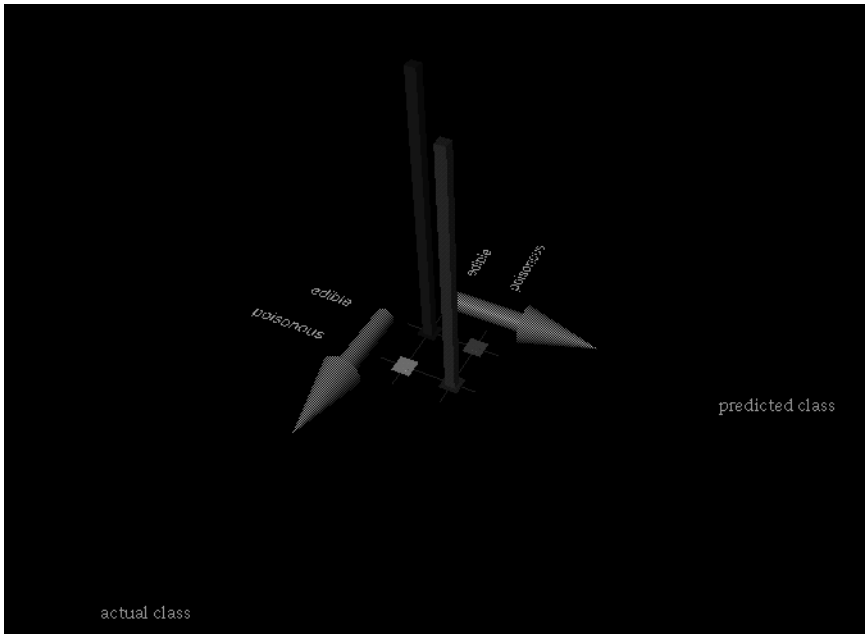
用混合矩陣和損失矩陣調整模型

當在分類中發現錯誤時，了解分類器發生混亂的方式以及這種混亂伴隨的耗損是很有幫助的。MineSet 提供兩種互補的矩陣來解決這個問題：混合矩陣和損失矩陣。

使用混合矩陣來調查錯誤

假如您要將蘑菇分類成有毒或可食用的。將蘑菇分類爲有毒的可能需要花費 2 美元 (蘑菇的成本)。但是，將一個有毒的蘑菇分類爲可食用的，可能會使您住在醫院中並花費 10,000 美元 (樂觀的估計)。MineSet 的混淆矩陣和損失矩陣功能可以處理此種情況，並且可以提供錯誤和不正確預測的詳細圖片。

圖 11-3 所示爲*蘑菇*資料集的混合矩陣，其決策樹僅使用了 10% 的資料進行訓練，遠少於建議的訓練集。關於詳細內容，請參閱第 7 章，「理解預測建模」。



11-3 「蘑菇」資料集的混淆矩陣

代表有毒蘑菇的 8 個記錄分類爲可食用的 (0.1%)；代表可食用蘑菇的 15 個記錄分類爲有毒的 (0.2%)。其餘的 3793 個可食用蘑菇和 3496 個有毒蘑菇的分類正確。儘管模型錯誤率僅爲 0.31% (小於 1%) ，但是按照上一段範例中的估計損失爲 $10000*8 + $2*15 = $80,030。

混淆矩陣用一個對錯誤類型的可視化處理，提供關於模型出錯的詳細圖片。兩個坐標軸分別代表測試集 (支持集) 中*預測*分類值和*實際*分類值。混合矩陣將先於修正計算。下面介紹在上述情景中如何使用混合矩陣。

顯示混合矩陣

在本範例，使用「工具管理員」視窗開啓資料檔案 *mushroom.schema*，然後研究在分類蘑菇爲可食用或有毒時的混亂情況。該檔案在 *MineSet 3.0\data* 目錄中，位於 MineSet 的初始安裝目錄。

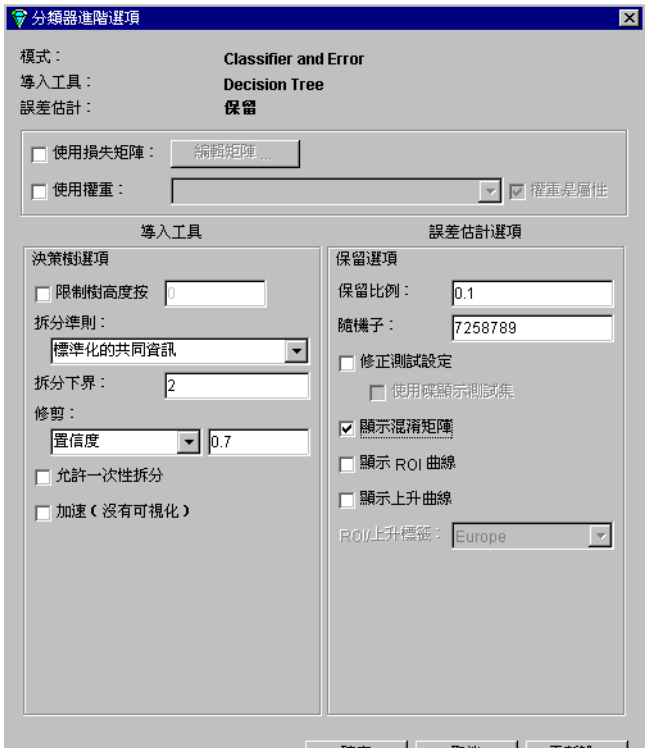
1. 在「工具管理員資料目標」窗格中按一下「挖掘工具」索引標籤；然後按一下「分類」索引標籤，並從以下的跳現式功能表中進行選擇：

   模式： 分類器和誤差估計

   導入工具： 決策樹

   離散標籤： 可食用性

2. 按一下*進階選項*，將出現「分類器進階選項」窗格，如圖 11-4 所示。



11-4 「分類器選項」窗格顯示的混淆矩陣

3. 在「誤差估計選項」窗格中，將支持比更改爲 0.1，關閉「修正」測試集，按一下「顯示混合矩陣」核取方塊，然後按一下「確定」。

   可以減小訓練集的大小，使分類器在展示時產生錯誤。典型的訓練集要大得多。

4. 在「工具管理員資料目標」窗格中按一下*開始*。

   當決策樹可視化處理出現時，請關閉它並檢查圖 11-5 所示的混淆矩陣。混淆矩陣顯示在分類器出現分類錯誤的地方。從這裡，可以根據對資料的了解建立一個「損失矩陣」，使特定錯誤的容錯性降低。

   圖 11-5 所示的畫面表示使用這個減少的訓練集，有 8 個有毒蘑菇錯誤分類爲可食用的，占總數的 0.1%。按一下可視化處理中的選擇物件，然後從「選項」功能表中選擇「顯示」，查看選擇清單

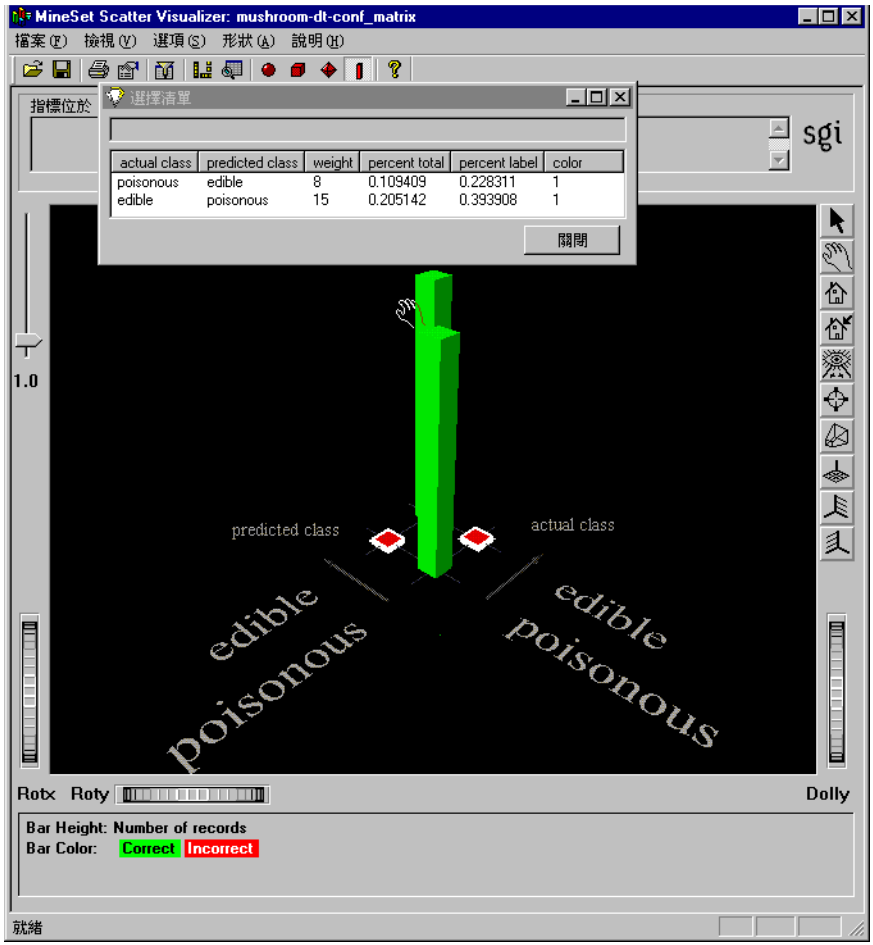


圖 11-5 顯示蘑菇資料集的混淆矩陣錯誤分類

定義損失矩陣

本部分解釋如何使用損失矩陣以減少成本高的分類錯誤類型。範例將繼續上一部分中場景，使用蘑菇資料集、決策樹導入工具、對可食性標籤分類。

1. 用「檔案」>「結束」來關閉「混淆矩陣」畫面，返回至「工具管理員」視窗。
2. 在「資料目標」窗格的「挖掘工具」中，按一下*進階選項*，打開「分類器選項」窗格。
3. 在窗格的第二部分，按一下*使用損失矩陣；*然後稍等片刻。
4. 按一下*編輯矩陣*，加權發生錯誤的成本。「損失矩陣」視窗類似於圖 11-6 所示。

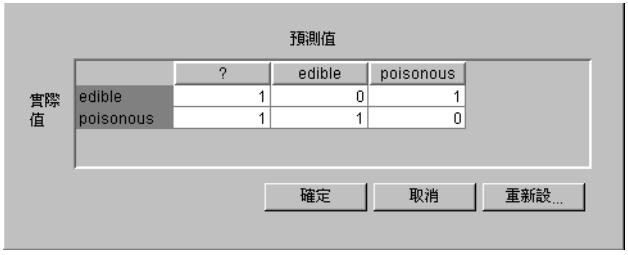


圖 11-6 「損失矩陣編輯」窗格

5. 在「損失矩陣」的「實際值」列中設定這些值，如圖 11-6 所示從左到右讀取：

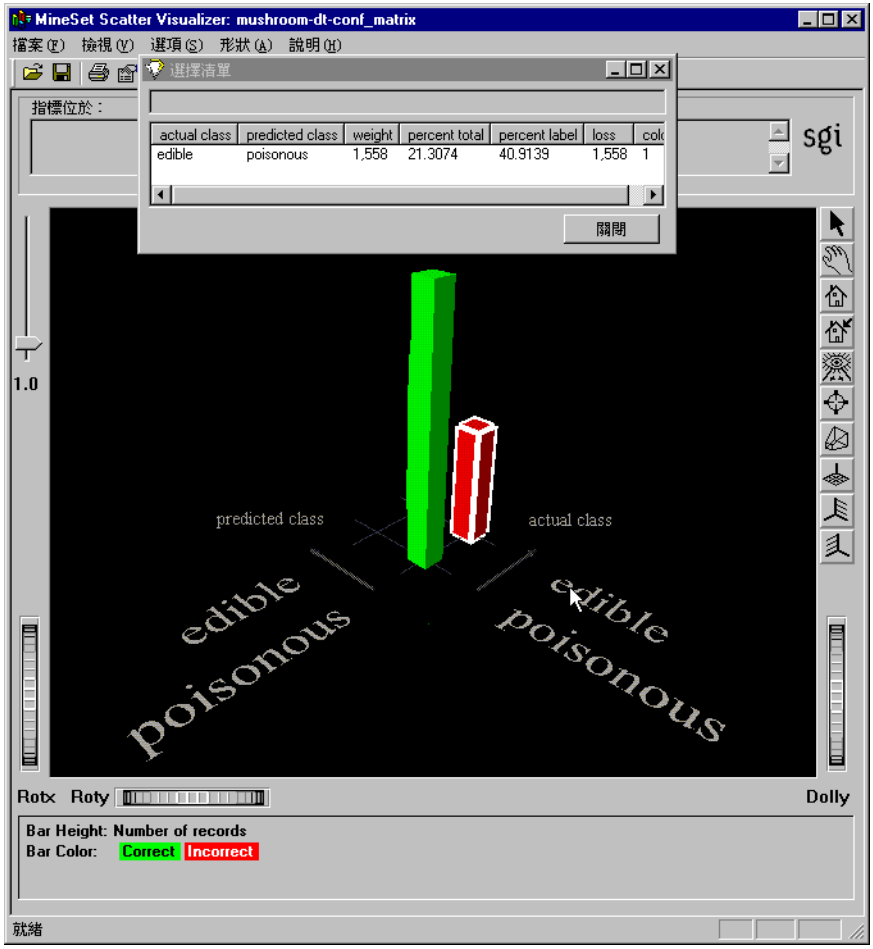
   可食用的： 1—0—1

   有毒的： 1—50—0

   這將把有毒蘑菇分類爲可食用蘑菇的成本被加權爲 50 倍於反向錯誤分類的成本。

6. 按一下*確定*以設定「損失矩陣」值，然後按一下*開始*執行分類器。

   新的分類器非常保守；不會發生將有毒蘑菇分類爲可食用蘑菇的錯誤。相反地，將可食用蘑菇分類爲有毒蘑菇的錯誤卻有 1558 個。使用蘑菇成本 ($2/ 個) 對住院成本 ($10,000) 的圖，新的估計損失爲 $10,000*0 + $2*1,558 = $3,116，僅爲沒有考慮損失的分類器成本的 3%。圖 11-7 所示爲產生的混合矩陣。



11-7 帶有損失矩陣的蘑菇資料集的混淆矩陣

關於誤差估計的詳細資訊，請參閱 *《MineSet 3.0 企業版參考指南》*。

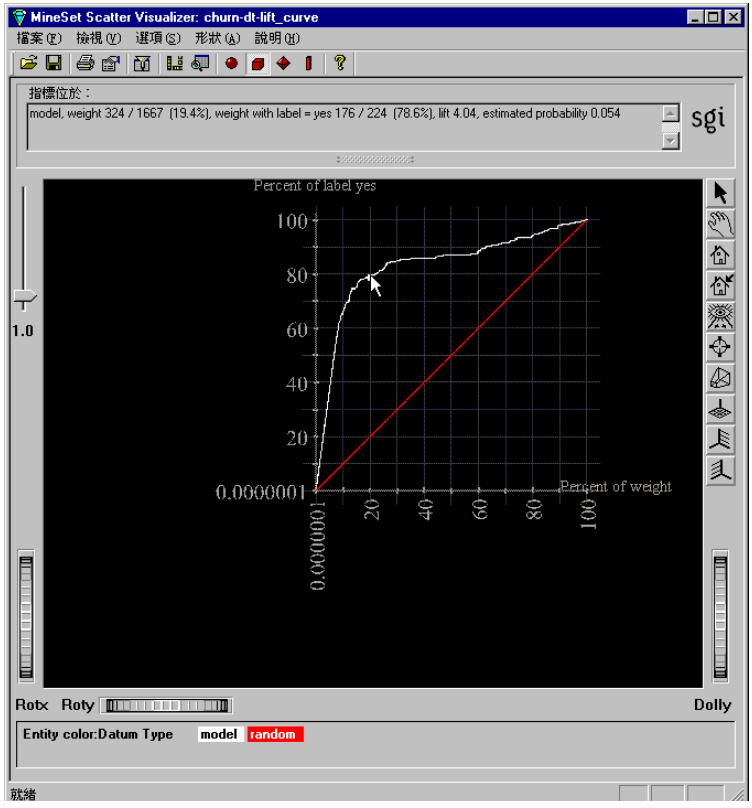
用上升曲線和 ROI 曲線評測模型

當商家從事一項長期項目時，能在下一階段更改計劃或放棄項目之前預測收益或損失。MineSet 可以判斷問題的結果，並在事件產生顯著影響之前進行修正。可以利用上升曲線和投資回報曲線 (ROI) 兩種方法來進行這項工作。

用上升曲線檢查預測

上升曲線顯示預測特定標籤值中隨機記錄順序與模型產生的順序之間的差別。例如：當建立一個預測哪些用戶可能波動的模型時，您可能希望在可能波動的用戶波動*之前*瞄準它們。上升曲線有助於實現此目標。本範例使用客戶波動資料集。

在上升曲線圖中，坐標軸 X 顯示從 0 到 100% 的記錄數，坐標軸 Y 顯示的記錄數對應於具有指定標籤值的用戶 (在此處，`Churn=yes`)。



11-8 上升曲線

圖中顯示兩條曲線，如 圖 11-8 所示。下面的曲線 (紅色) 顯示在指定隨機順序記錄下預期客戶波動的用戶數曲線。上面的曲線 (白色) 顯示按照模型對每個記錄的評價 (機率估計) 順序放置時客戶波動用戶數的百分比。將首先出現由模型識別爲最可能波動的用戶記錄；那些機率較小的記錄將出現在後面。模型排序的優點在於可以發現模型曲線與隨機曲線之間的差別。

建立該上升曲線時，在測試集中應用一個選定的模型。在下面的範例中，用一個指定的資料區段來進行訓練。然後將導入的模型在資料集的剩餘部分上執行。

要在選定模型上產生上升曲線：

1. 在「工具管理員資料目標」窗格中，按一下「挖掘工具」索引標籤。
2. 按一下「分類」索引標籤。
3. 選擇「分類器和錯誤估計」模式以及「決策樹導入工具」。
4. 按一下*進階選項*，然後在「分類器選項」跳現式功能表中，按一下「誤差估計選項」下的*顯示上升曲線*核取方塊。在 ROI/ 上升標籤跳現式功能表中選擇標籤值 (是)。
5. 按一下*確定*接受導入工具選項，然後關閉對話方塊。
6. 按一下*開始*。

該過程可能需要花費一些時間。關閉決策樹。產生的上升曲線將顯示任務列下之選擇視窗中所有選定點的詳細內容。將指標沿白色 (模型) 線移動，然後按一下各點，查看 `churn=yes` 的用戶百分比和升幅。查找曲線的拐點，在本範例中，拐點在分類器的估計機率爲 0.056 的地方。在這一點，向可能波動的用戶追加刺激能帶來的投資回報將迅速減少。下一步是將分類器應用到整個資料集 (請參閱第 7 章中的「使用所有資料進行分類」)。

使用投資回報曲線尋找銷售利潤

投資回報曲線 (ROI) 類似於上升曲線，但是它顯示損失而非錯誤方面的準確性，它取決於如何對損失矩陣加權 (請參閱第 180 頁的「定義損失矩陣」)。如果模型對其預測結果很肯定，則期望損失將較小，記錄會出現在 ROI 曲線的左邊。

ROI 曲線希望您對資料集中的每個單獨記錄都進行操作。該操作與選定的標籤值相關聯。例如：在*客戶波動*資料集中，與標籤「Yes」相關聯的操作可能是向某人傳送銷售材料。這可能會阻止某用戶波動，但是如果不加區別地傳送材料，成本就會非常昂貴。如果模型是用來預測是否向某人傳送郵件，則 ROI 曲線最高點顯示的就是在測試集中大約能夠節省多少錢。

可以從投資回報曲線中看到出現某種錯誤的成本，並指出繼續採取行動會無效的那一點。要使用「分類器和錯誤」模式在*客戶波動*資料集上產生 ROI 曲線，可以按照以下步驟進行：

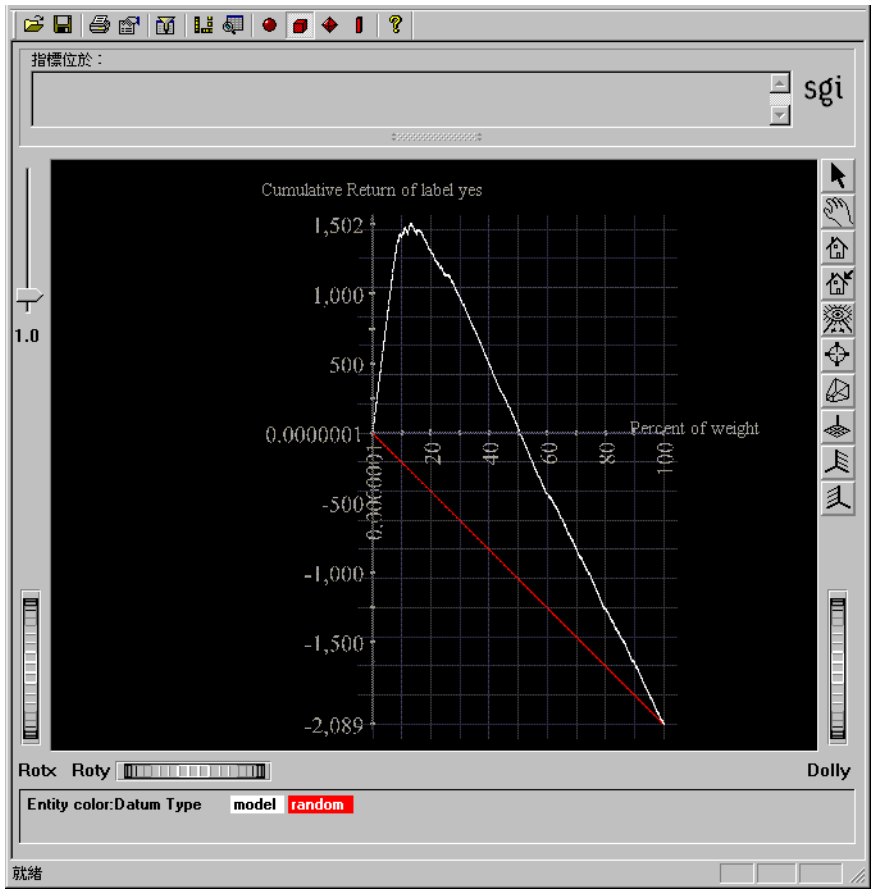
1. 在「工具管理員資料目標」窗格中，按一下「挖掘工具」索引標籤。
2. 按一下「分類」索引標籤，然後按一下*進階選項*按鈕。
3. 核取「修正」測試集、「顯示混淆矩陣」，然後在「分類器選項」窗格中按一下「使用損失矩陣」。
4. 在上面窗格中按一下*編輯矩陣*以加權發生錯誤的成本，如第 180 頁的「定義損失矩陣」。在本例中，用以下方法填充列：

   *實際值：*否： 1 0 2

   實際值： 是： 1 0 -10

   按一下*確定*以設定「損失矩陣」。
5. 將*顯示 ROI 曲線*設定爲開啓。
6. 在 ROI/ 上升標籤下拉式功能表中選擇「是」(對於客戶波動範例)。
7. 按一下*確定*。
8. 在「工具管理員」視窗中按一下*開始*。

將出現以下三個畫面視窗：「決策樹」、「混淆矩陣」和「ROI 曲線」。「混淆矩陣」顯示分類器在進行 churn=no 預測時更加保守，因此加權爲負數。在考慮損失的情況下，「混淆矩陣」顯示權重爲 4%。一邊的錯誤增加，但另一邊的錯誤卻減少了。使用「檔案」>「結束」以關閉「混淆矩陣」和「決策樹」畫面，然後查看「ROI 曲線」視窗圖 11-9。



11-9 投資回報曲線

ROI 曲線類似上升曲線。中間的水平線代表零收益和零損失。紅色線代表採用隨機樣本並向他們傳送郵件時的期望效果。如果向每個人都發送郵件，將會有所損失，因爲郵寄成本太高。但是，有一個最佳的投資回報點，即曲線的頂點 1448，或總體的 15.2%。

在使用 ROI 曲線之前填充「損失矩陣」時，請務必小心。特定預測標籤下的欄將決定爲該標籤值產生 的 ROI 曲線。如果要爲使用標籤客戶波動的客戶波動資料集而填充損失矩陣，則該欄中的項目需要表示出：在各類資料之上，與該標籤值相關的操作能夠帶來的期望收益或損失。例如：客戶波動中的欄「預測是」與列「實際值否」下的條目可能包含值 2，以表明向某個不會客戶波動的人郵寄小冊子的成本 (與「是」相關的操作) 是 2 塊美元。另一方面，「是」欄、「是」列下的條目的值是 -10，表示阻止了用戶的波動，除郵寄成本外將爲公司節省了 10 美元。這些成本難以估計，很小的變化就會顯著影響產生的 ROI 曲線之頂點位置。

關於進一步的詳細內容，請參閱 *《MineSet 3.0 企業版參考指南》*，以及查詢 MineSet 的網站 http://mineset.sgi.com，獲得最近的更新。。

用聚類劃分資料

聚類的目的是爲了確定資料集中哪些元件或特徵是類似的。這對於研究新的資料集是很有用的。例如：蝴蝶花資料集顯示三個明顯不同的蝴蝶花聚類—*藍色蝴蝶花*、*多色蝴蝶花*和 *純白色蝴蝶花*。本章包括以下主題：

聚類總覽

聚類是描述性的資料挖掘任務，所以不用將欄指定爲標籤。因此，將聚類歸爲非監督的學習運算法則。聚類模型通常是建立於整個資料集之上並在整個資料集上進行評估。

MineSet 將類似的紀錄分在一起組成聚類，目的是爲了使每個組中的相似性最大化。MineSet 提供兩種不同的聚類模式：單步的 k 平均值和重複的 k- 平均值。在*《MineSet 3.0 企業版參考指南》*的「聚類」中對兩種模式進行深入的說明。

一旦執行了聚類，就可直接使用「聚類可視化工具」查看聚類中心，它將顯示資料集中獨立的屬性。可以檢查最顯著的屬性，並了解它們之間的區別。但是，若要了解聚類間屬性的相互關係，「分散可視化工具」和「決策表」能提供更清楚的檢視。

要使用「分散可視化工具」查看聚類模型，需要確定應該使用「欄重要性」工具將哪些欄對照到不同的坐標軸。首先使用「應用模型」工具 (請參閱第 129 頁的「應用預測模型」)，然後從可用模型清單中選擇 < *資料集名稱* >.*cluster*。這將提供新的「聚類」欄，在「分散可視化工具」中將該欄對照到顏色。

用「工具管理員」啟動聚類

聚類沒有先決條件，也沒有必需的選項。預設情況下，MineSet 使用單步 k- 平均值聚類方法挖掘資料中的三個聚類。聚類一旦完成，就可看到聚類的評估，然後會出現「聚類可視化工具」。

1. 在「工具管理員資料目標」窗格中，按一下「挖掘工具」，然後再按一下「聚類」索引標籤，如圖 12-1 所示。

圖 12-1 「聚類」索引標籤

2. 要控制聚類方法的選擇，可以從「方法」跳現式功能表中選擇單步或重複 k- 平均值，如表 12-1 所示。

表 12-1 聚類方法選項

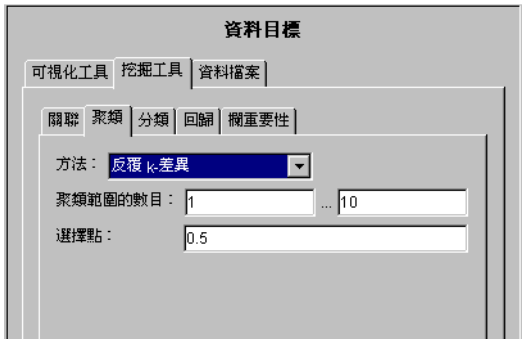
方法	描述
單步 k- 平均值	預設方法。指定要尋找的聚類數 (預設值爲 3)。
重複 k- 平均值	需要上界和下界以及選擇點 (0 到 1 之間的一個數—預設值爲 0.5)。

關於*選擇點*，**請參閱詞彙表定義。**

3. 按一下*開始*。

注意： 聚類是計算密集的操作，在處理較大的資料集時可能需要一定的時間，尤其是以重複 k- 平均值模式執行時。如果資料集中的記錄數多於 10,000，請嘗試聚類一個資料抽樣。

聚類分配了從 1 開始的連續數值名稱。雖然聚類名稱用數字表示，但是對聚類來說，次序不存在。



12-2 使用重複 K- 平均值算法聚類

在重複 k- 平均值中，根據聚類在拆分過程中的出處爲它們命名。使用連續的數字命名最初的聚類 (根據聚類的最小數)，就像在單步 k- 平均值中一樣。當拆分一個聚類時，兩個新的聚類將按已拆分的聚類命名，附加一個「A」或「B」。例如，命名爲「2-B-A」的聚類是在最初聚類中從聚類 2 中衍生而來，並已拆分了兩次。

關於聚類運算法則操作的詳細內容，請參閱*《MineSet 3.0 企業版參考指南》*中的「聚類」。

使用聚類樣例檔案工作

MineSet 提供的汽車樣本資料集比較簡單，處理的是一些熟悉的概念，如：馬力、車輛重量和達到 60 英哩 / 小時所需的時間。

使用「工具管理員檔案」>「開啓新的資料檔案」下拉式功能表，選擇 *cars.schema* 檔案。在 MinceSet 安裝目錄的 *data 目錄細化下尋找*。

1. 在「工具管理員資料目標」窗格中，按一下「挖掘工具」索引標籤。
2. 按一下「聚類」索引標籤，將「方法」設定爲單步 k- 平均值，將聚類數設定爲 3，然後按一下右下方的*開始*。

12-3 聚類可視化工具主視窗

3. 在產生的「聚類可視化工具」視窗 (圖 12-3) 中，按一下顯示聚類 1 的綠色長條，它控制長條圖和柱狀圖所代表屬性的優先次序。當在聚類 1 上按一下時，屬性重要性 (即：柱狀圖出現的次序，從上往下讀) 爲汽缸、體積、馬力、重量磅，然後是英哩 / 加侖。這個重要性次序僅由可視化處理的這個聚類控制。
4. 在其他的聚類中比較同一列，以了解聚類之間屬性的區別。
5. 選擇聚類 2，並注意屬性次序的不同。在本例中，產地較爲重要，其次是商標、體積、汽缸、重量磅，再其次是馬力，最後是英哩 / 加侖。

認識「聚類可視化工具」主視窗

「聚類可視化工具」顯示按列和欄排列的方塊圖。每一列顯示一個資料中的屬性，每一欄代表頂部窗格中指定的一個聚類。總體欄顯示對於整個資料集的統計。

按照第 190 頁的「用「工具管理員」啓動聚類」中的步驟，可以得到一個類似於圖 12-3 所示的畫面，其中的聚類按次序編號。運算法則無法偵測聚類代表的內容，因此需要人工翻譯解釋。屬性在「聚類可視化工具」視窗中顯示的次序是很重要的；預設情況下，屬性是按照用於聚類間彼此區分的重要性次序顯示的。

改變欄的次序：

1. 按一下含有聚類名稱的窗格，按區分該聚類與其他所有聚類的重要性次序重新顯示屬性。
2. 按一下含有「總體」的窗格以恢複預設的排序。

聚類的其它可視化處理

「聚類可視化工具」提供一種易用的聚類基本可視化處理。但是，這不是可視化聚類的唯一方式。*《MineSet 3.0 企業版教學課程》*中介紹一個使用「分散可視化工具」將聚類結果轉換爲畫面的範例。主要想法是應用剛建立的聚類模型建立一個新欄，然後使用選擇的可視化工具。對於查看聚類結果，「分散」、「証據」和「決策表可視化工具」特別有用。

用關聯規則分析資料

「關聯規則」挖掘資料的一致性，即：哪些資料實例可能與其他指定實例同時發生。這就是市場供求分析，它來自於一個隱喻：哪些產品可能在同一個購物籃中一起購買。本章主要包括以下幾個主題：

提供一些配置和資料檔案樣例，用於演示「關聯規則」的特性和功能。這些檔案在 *MineSet 3.0* 下的 *\examples* 目錄中，位於 MineSet 的初始安裝目錄。

關聯規則產生和可視化處理總覽

如果希望知道在超級市場的冷凍食品展示架上，爲何將雞蛋替代品與威化餅放在一起，則您可能已看到在其中發揮作用的關聯規則。關聯規則可以發現資料中的關係和隱含模式—古典資料驅動式的知識發現過程中不是很明顯的相互關係在此變得非常清晰，並且可用可視化形式顯示。關聯規則對於發現大資料集中的隱含模式尤其有用，可以用關聯規則指出資料集中的某些項目同時出現的頻率。這是一個非監督的學習運算法則，因爲它不是集中於任何一個特定的屬性上。而是同等對待所有的屬性，對於值得注意的規則進行全面搜尋。

關聯規則的形成有以下兩個階段：首先關聯規則產生器將對資料檔案進行處理，建立一個可用於可視化工具的檔案。然後可視化工具顯示該檔案，在本例中爲「分散可視化工具」。

關聯規則產生

關聯規則產生器可以產生簡易 (一對一) 和多路的關聯規則。本部分說明簡單關聯規則。關於多路規則的描述，請參閱第 204 頁的「說明多路關聯規則」。

一個簡單關聯規則規定：如果指定的 X 爲真 ,則在一定的機率下 Y 也爲真。MineSet 稱 X 爲規則的左側 (LHS)，稱 Y 爲規則的右側 (RHS)。

考慮顧客在一次單獨進商店時所購買的一組物品。在這種情況下，規則可能是：「80% 買尿布的人同時會買嬰兒奶粉。」整個百分比被稱作規則的*置信度*。

在本範例中,「尿布」是規則左側 (LHS) 的項目,「嬰兒奶粉」是規則右側 (RHS) 的項目。在此種情況下：

- 如果嬰兒奶粉出現在 RHS，則 LHS 就有助於確定何種措施可以增加該項目的銷售。
- 如果尿布出現在 LHS，RHS 就有助於確定如果商店中停止出售尿布，何種商品將會受到影響。

關聯規則產生器首先對輸入檔案進行處理，然後產生一個由規則組成的輸出檔案。如果 X 和 Y 爲記錄中的項目，則如下的規則：

X=>*Y*

表示只要 *X* 出現在記錄中，則預計 *Y* 將會以一定的頻率出現。

關聯的強度有四個數字來量化，匯總於 表 13-1

表 13-1 關聯規則組件

測量	描述	統計說明
支援度	LHS 和 RHS 同時發生的頻率	P(LHS∩RHS)
置信度	在所有 LHS 發生的情況中，RHS 也發生的分數，或者支援度除以 LHS 項目發生的頻率。	P(RHS | LHS)
期望置信度	RHS 項目發生的頻率	P(RHS)
上升度	置信度與期望置信度的比值	P(RHS | LHS):P(RHS)

請參閱以下詞彙在詞彙表中的定義：*支援度*、*置信度*、*期望置信度* 和 *上升度*。關於詳細的技術內容，請參閱*《MineSet 3.0 企業版參考指南》*中的「關聯規則」。

注意： 如果只知道 Y 和型式如 X=>Y 的規則，則對於 X 將一無所知。規則指定暗示只能由 LHS 到 RHS。

規則可視化處理

當關聯規則顯示在柵格圖上時，可以看到關聯規則的圖形化形式。左側 (LHS) 項目在一個坐標軸上，右側 (RHS) 項目在另一個坐標軸上。如圖 13-1 所示，規則的屬性顯示在其 LHS 和 RHS 項目的交叉點上。該畫面可以包括長條圖、圓盤和標籤。

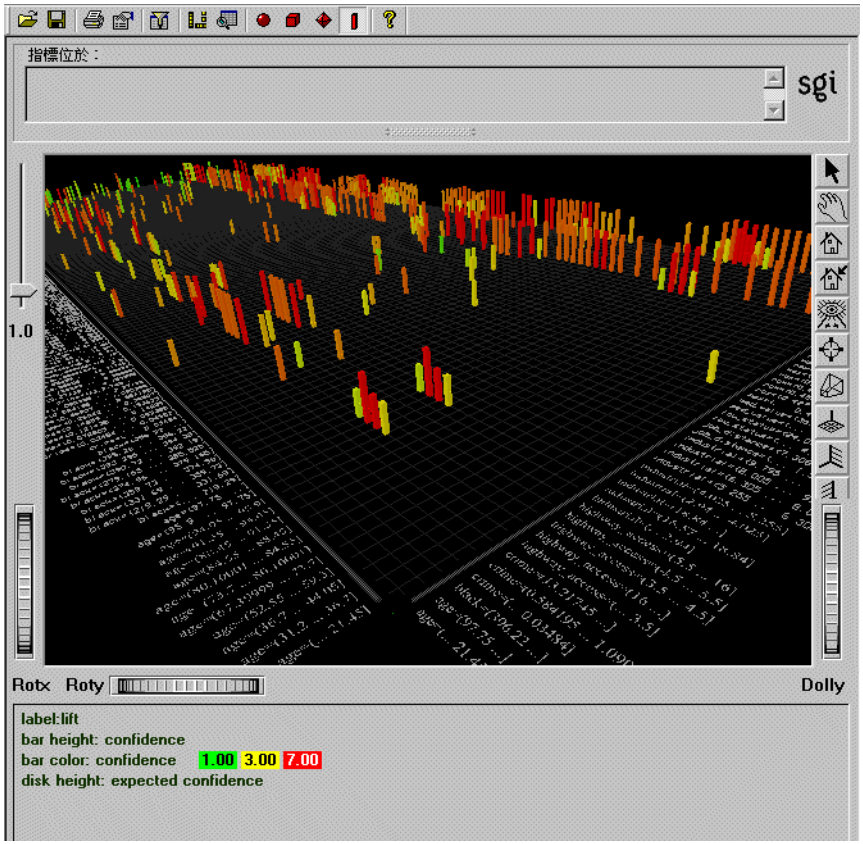


圖 13-1 關聯規則可視化工具主視窗的詳細檢視

主視窗的底部將顯示一個圖例，指示顯示屬性 (例如長條高度和顏色) 與基本規則關聯值 (例如置信度和支援度) 之間的對照。

執行關聯規則

本部分描述如何使用「工具管理員」簡化配置「關聯規則」工具的任務。

建立關聯

本範例中使用汽車資料集建立簡單關聯。您可能會發現以下屬性之間的關聯：英哩 / 加侖、馬力、加速度、重量、引擎大小、產地、品牌以及汽車的生產日期。例如：英哩數是否隨事件提高？引擎功率是否減小？

以下步驟是建立關聯最簡單的方法：

1. 開啓「工具管理員」，選擇伺服器，並選擇 *cars.schema* 作爲資料來源 (請參閱第 15 頁的「說明執行 MineSet」。
2. 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；在下面一列的索引標籤中，選擇關聯 (對應於關聯規則)。將顯示圖 13-2。

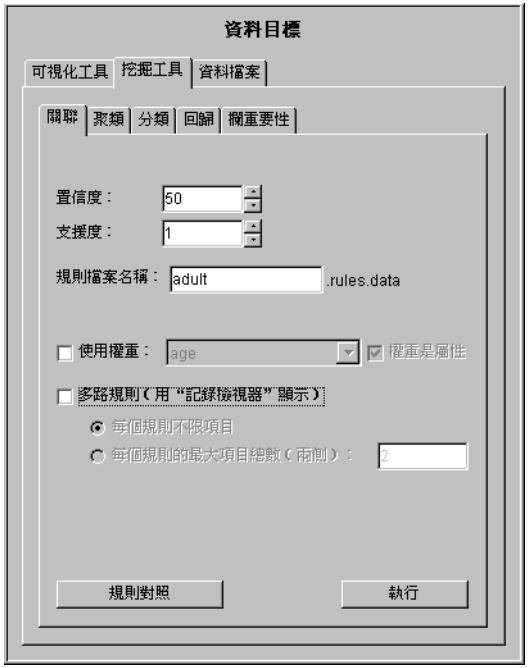


圖 13-2 關聯產生的初始工具管理員視窗

3. (可選步驟)「關聯規則」的運算法則只適用於離散值；因此連續類型的欄會自動分組。如果希望其它的分組，可以使用「資料轉換」窗格上的*欄分組*按鈕。(關於詳細內容，請參閱第 37 頁的「說明爲欄改變或建立新的分組」)。
4. 按一下*開始*，*執行「關聯規則產生器」*。或者選擇以下設定：

   置信度—指定規則的最小置信度臨界值。將不會產生置信度小於該值的規則。預設值爲 50%。可能的值爲 1-100。

   支持度—指定作爲記錄總數百分比的最小支持度臨界值。將不會產生支援度小於該值的關聯規則。預設值爲 1%。可能的值爲 1-100。
5. (可選性) 一旦完成關聯規則選項的選擇，可以按一下*規則對照*按鈕，將規則屬性 (如支援度、置信度、期望置信度以及上升度) 對照到可視元件。請參閱下面的「將規則屬性對照到可視元件」。

記錄加權

當希望將某些記錄指定爲更重要的記錄，或補償不均衡的採樣時，關聯規則允許記錄加權。如果沒有核取*按欄加權*方塊，則每個記錄的權重都是 1。只有當資料集中包含可以加權的欄時，*按欄加權*方塊才會處於活動狀態。

要開啓記錄加權，可以按一下*按欄加權*核取方塊。當核取該方塊時，會出現一個跳現式功能表讓您選擇包含每個記錄權重的欄。如果選取*權重是屬性*方塊，其中將包括關聯規則產生器發現的規則中的權重欄。如果沒有核取該方塊，則指定所有的記錄爲同樣的權重，以確定產生器發現的規則。關於記錄加權的進一步說明，請參閱第 3 章中的「加權記錄」。

將規則屬性對照到可視元件

「關聯規則」工具可以將規則屬性對照到畫面的可視元件。按一下「規則對照」按鈕，打開「將規則對照到 Rule 可視化 元件」面板，如圖 13-3 所示。

圖 13-3 關聯規則對照

可以對照的可視元件將在下面列出；前面有星號的項目是可選的：

表 13-2 將關聯規則對照到可視元件

可視元件	對照
高度 - 長條圖	指定長條高度代表的屬性。預設值爲「支持度」
* 高度 - 圓盤	指定圓盤高度代表的屬性。
* 顏色 - 長條圖	指定長條顏色代表的屬性。預設值爲「上升度」
* 顏色 - 圓盤	指定圓盤顏色代表的屬性。
* 標籤 - 長條圖	指定長條標籤代表的屬性。

說明「分散可視化工具」中解釋關聯規則

關聯規則使用有效配置檔案的說明顯示「分散可視化工具」中規則檔案的資料。要查看樣例檔案，可以在「MineSet 3D 可視化工具」中使用「檔案」>「開啓」，查看配置檔案的清單。這些檔案在 *MineSet 3.0\examples* 目錄中，位於 MineSet 的初始安裝目錄。例如：指定圖片中 *brand.rules.scatterviz* 的結果，如圖 13-4 所示。關聯規則檔案可以經由其名稱結尾的 *.rules.scatterviz* 來識別。

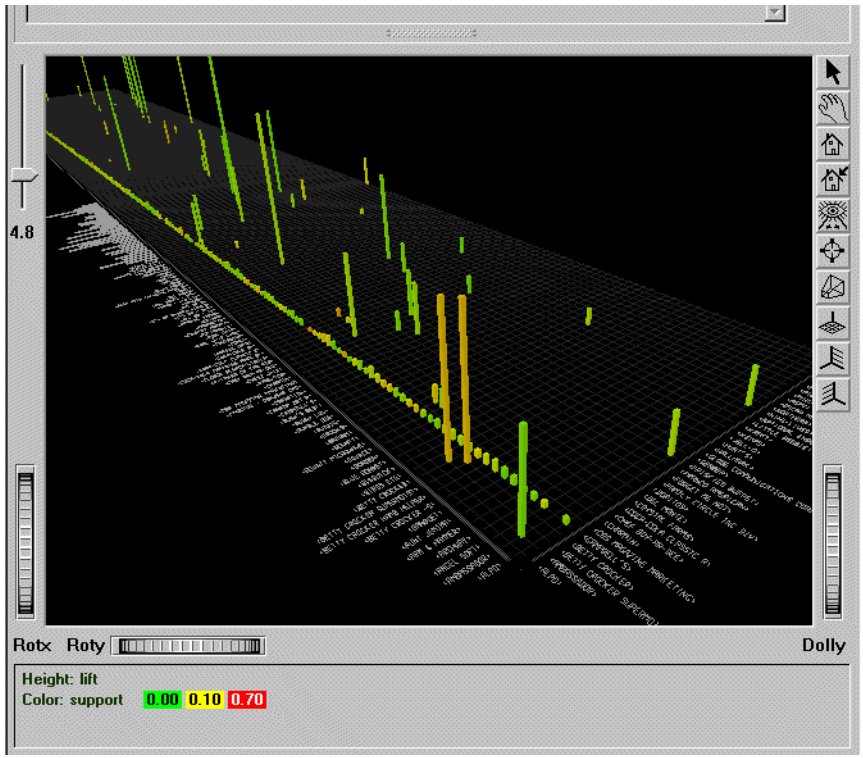


圖 13-4 在指定 brand.rules.scatterviz 樣例時的初始關聯規則檢視

規則顯示在「分散可視化工具」中的柵格圖上。左側 (LHS) 項目在一個坐標軸上，右側 (RHS) 項目在另一個坐標軸上。如圖 13-4 所示，規則的屬性顯示在其 LHS 和 RHS 項目的交叉點上。例如：長條高度對應於支援度，長條顏色對應於上升度。如果顯示的檢視太小，項目標籤不會出現在坐標軸的兩側。可以使用視窗右下的伸縮滑輪來放大檢視，直到項目標籤出現。當滑鼠爲選擇模示時，也可以將滑鼠指標放在單獨的長條上查看特定規則的標籤 (請參閱圖 13-5)。關於該特定規則的所有詳細內容都將顯示在檢視區域的左上角。

例如：在圖 13-4 中，長條高度對應於支援度，長條顏色對應於上升度。

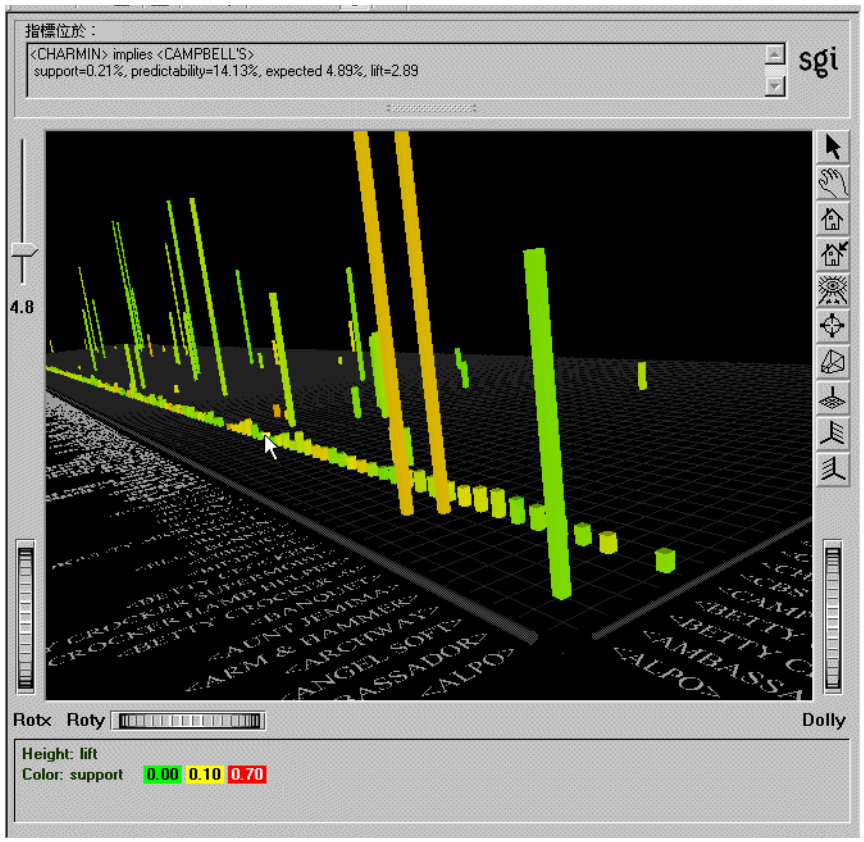


圖 13-5 代表規則的長條圖上的游標

如圖 13-5 所示，將滑鼠指標放在關聯規則物件上，會顯示出該物件的資訊。只要指標停留在物件上面，就會顯示資訊。但是有一定的顯示時間，如果將游標放在物件上面並按一下滑鼠左鍵，那麼在主視窗上方也會顯示同樣的訊息，該訊息將一直保留到游標離開該物件。另外，可以按住 CTRL 鍵並按一下選擇多個規則。

使用滑鼠可以將選擇視窗中的文本剪下並貼到其他的應用程式中，例如：報表或資料庫。

追溯

通過對選定的規則執行邏輯「與」來確定追溯表達式。因爲原始表格中的欄與 *.rules.data* 檔案中的欄不符合，規則產生器將產生一個專門的欄以幫助構建執行追溯時的篩選表達式。這意味著對追溯設置設定面板的更改將不會有效，因爲已經將一個專門的字串值欄對照到 *.rules.scatterviz* 檔案中用於追溯。

追溯規則時，MineSet 將顯示滿足規則的所有記錄。

多路關聯規則

在一些情況下，更加複雜的規則可能會很有用，這些規則在 LHS 和 / 或 RHS 上有多個項目。這些就是多路關聯規則。圖 13-6 所示即爲多路規則進行配置的「工具管理員關聯」面板。

如果選取「多路規則」核取方塊，關聯規則產生器將會產生所有滿足最小支援度和置信度臨界值的規則，包括那些在 LHS 和 RHS 中有多個項目的規則。這樣的規則可能是：「啤酒和空心粉暗示洋芋片、甜麵醬和葡萄酒。」

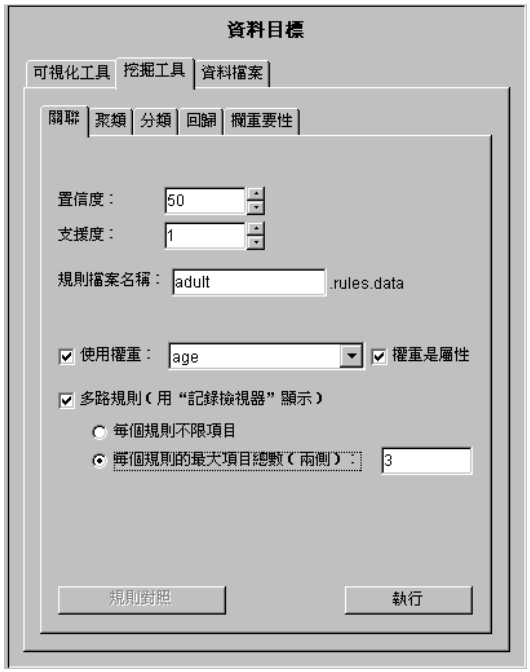


圖 13-6 建立多路關聯規則產生的初始工具管理員視窗

使用「記錄檢視器」可以顯示多路規則，因爲在「分散可視化工具」中沒有明顯的方法來顯示更複雜的規則。它們顯示爲一列一個規則。表格中前兩欄包含 LHS 和 RHS 中的項目數。以下的四欄包含支援度、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，項目由「和」分隔開。上面的範例規則中，LHS 包含兩個項目「啤酒和空心粉」，RHS 包含三個項目「洋芋 片、甜麵醬和葡萄酒。」

可以在「每個規則的最大項目總數」欄位中輸入一個數字，限制產生規則的大小。此數目指出規則中允許的項目數最大值。規則中的項目數是 LHS 和 RHS 中項目數的總和。上面的規則範例中有三個項目；簡易規則有兩個項目。

注意： 產生多路規則需要的時間較長。查看「工具管理員」狀態視窗，指出每次重複產生的規則數。如果產生的規則太多，可以取消操作，並增加最小支援度或置信度臨界值，或減少每個規則允許的項目數。

挖掘 MineSet 用戶詞彙表

組合
在組合 (如下) 操作中資料以各種方式組合起來。

組合
在一些指定的欄集合上，使用一個或多個操作匯總多列資料的過程。組合操作可以產生匯總統計、轉換表格，甚至可以產生根據原始資料的陣列。

運算法則
完成某些任務的正式過程。例如：導入工具運算法則是一組機器指令，它告訴電腦如何建立分類器或回歸器。

陣列
能夠儲存多個項目的資料類型。陣列類型由組合產生，可視化工具用它來儲存隨時間而活動的動態資料。MineSet 支援一維陣列 (也稱爲向量) 和二維陣列 (也稱爲矩陣)。

屬性
預測模型進行輸入的欄。在執行預測建模時，資料中的欄劃分爲屬性 (提供輸入) 和標籤 (代表模型的輸出)。

自動分組
這個運算法則可以選擇分組臨界值 (請參閱分組)，使不同分組中的標籤分布儘可能地不同。形式上，這個運算法則將使每個分組內的熵 (請參閱熵) 最小化。

修正
修正將整個資料集應用到由較小樣本建立的模型。產生的模型保留了原始結構。修正的目的是爲了在保持模型誤差估計的同時，使模型的內部分布準確地反映所有資料。

分組

分組將連續的資料劃分到離散的組中，將實際資料轉換爲分類資料。例如：一個連續的年齡範圍可以分組爲： 0-18、19-25、26-35 等等。這些群組或分組定義爲不會重疊的連續區域。

塊狀圖

在顯示資料分布的可視化處理中展示的分段區塊。類似於圓餅圖，但是爲方塊。

選擇點

選擇點是介於 0 和 1 之間的值，它起著引導的作用，例如：引導聚類數的選擇；較高的選擇點意味著較大的數，而較低的選擇點意味著較小的數。選擇點爲 1.0 通常選取上界。在聚類中，如果界限是一個和五個聚類，則值爲 0.4 的選擇點將選取兩個聚類，而值爲 0.8 的選擇點將選取四個聚類。值爲 1.0 的選擇點將選取五個。

分類器

試圖描述關於其他欄 (屬性) 的欄 (標籤) 的預測模型。從標籤已知的資料中建構的分類器，將用於預測標籤未知的新資料的標籤值。從內部來看，分類器是爲每個輸入列預測離散值的運算法則或數學公式。例如：從蝴蝶花資料集中建立的分類器可以預測指定花瓣和雄蕊的長度與寬度的蝴蝶花類型。分類器還會估計每個標籤值的機率。例如：從汽車資料集中建立的分類器可以預測特定汽車在美國製造的機率。分類器僅適用於離散標籤。回歸器類似於分類器，但用於連續標籤。

聚類

聚類確定資料集中的哪些元件是相似的。它按照一個運算法則或數學公式對記錄進行分組，這個運算法則或數學公式試圖找到相似記錄同時傾向的質心或中心。該過程將資料集劃分爲相互排斥的子集，不依賴於任何預先定義的分類。

條件機率

某個事件 (A) 在其他某個事件 (B) 發生的情況下發生的機率。寫作 P(A|B)，讀作「指定 B 時 A 的機率」。條件機率以矩形圖顯示在証據可視化工具左邊的視窗中，這些矩形圖顯示了在指定 (有條件的) 每個標籤值的情況下每個屬性值的相對機率。條件機率可以視爲指定標籤值的証據。

具有連續值的屬性

可以在整個連續值範圍內取值的屬性。在 MineSet 中，整數、雙倍和浮點型可以視爲連續的。

置信度

關聯規則 (X --> Y) 的置信度，將 *X* 和 *Y* *同時* 出現的頻率量化爲 *X* 出現的記錄數之分數。例如：如果置信度爲 50%，在 *X* 出現的記錄中，*X* 和 *Y* 同時出現的機率爲 50%。因此，若知道 *X* 會在記錄中出現，則 *Y* 也在該記錄中出現的機率爲 50%。

配置檔案

識別資料如何從 *.data* 檔案對照到滿足可視化工具要求的檔案。MineSet 配置檔案有下列幾種可能結尾： .treeviz、.scatterviz、.splatviz、.mapviz、.eviviz、.dtableviz、.statviz 和 .clusterviz

交叉驗証

估計預測錯誤的方法。交叉驗証將資料集拆分爲 k (一般爲 10) 個同樣大小、稱爲資料夾上的部分，建立 k 個預測模型，然後在經過其餘資料夾的訓練後，在不同的資料夾測試每個模型。可以多次重複這個過程以提高估計的可靠性。

資料清洗

資料清洗是經由修改資料的形式或內容以改進資料品質的過程，例如：移除或改正錯誤的資料值。

資料檔案

MineSet 的資料檔案是用索引標籤分隔的原始資料單一結構檔案，根據它來建立模型或可視化處理。MineSet 的資料檔案使用的結尾是 *.data* 。它們可能是二進位 (更小並且更快) 或 ASCII (可以人工讀取) 格式。

資料集

資料集是具有資料列和欄的表。其中的列有時稱爲記錄或實例。MineSet CD 中包括多種樣本資料集，例如：有關人們電信習慣的*客戶波動*資料集，大眾感興趣的*汽車*資料集，根據美國人口調查局篩選後的資訊的*成人*資料集。也可以將資料集描述爲一個模式及符合該模式的實例集合。通常假設這些實例沒有次序。

決策樹

根據分支系列測試的預測模型。每個測試都要檢查資料集中的單獨一欄，並用它確定下一個應用的測試。所有測試的結果將決定要預測的標籤。

離散屬性

具有有限數目的不同值的欄。在 MineSet 中，字串、分組值和日期被視爲離散的。如果整數用於分類器的標籤，也會將其視爲離散。

追溯

檢索可視化處理根據之部分原始資料的操作。

熵

「資訊論」中的一種度量，指的是資料的無序程度。資料集的熵越高，其中的值就越不同，越混雜。MineSet 中的許多挖掘運算法則都經由劃分資料來使熵最小化。

錯誤率

模型對資料集所做預測的正確率或錯誤率。錯誤率通常由獨立的測試集來決定，在學習過程沒有使用過該測試集。有時會使用更複雜的準確性估計技術，例如交叉驗証，尤其是對於只包含少量實例的資料集。

証據

機率估計的準確程度將影響最終的預測。証據分類器根據每個屬性的証據總和預測其標籤值。在形式上，証據是正態化條件機率的負對數。

期望置信度

期望置信度是資料集中 RHS 項目出現的頻率。因此，期望置信度和置信度之間的區別在於：因爲 LHS 項目的存在，預測能量中改變的度量。期望置信度指出項目之間不存在任何關係時的置信度。

篩選工具

在 MineSet 中，篩選工具通常是一個布林表達式，用來選擇或操作欄內容。例如：「m.p.g. < 25」表示只選擇顯示小於 25 英哩 / 加侖的項目 (即：對於該表達式爲真的記錄)。篩選工具是儲存起來、經過選擇的準則集合，這些準則指定資料集中的記錄子集。

預留

用於訓練模型的資料集的比例 (通常爲三分之二)，其餘部分將用於測試。預留方法是 MineSet 支援的最簡單的誤差估計方法。

假設檢驗

假設是可以檢驗有效性的假定解釋。假設檢驗是一種由上至下的方法，試圖証實或証僞事先的想法。在建立假設前，要爲注意到的行爲提出可能的解釋。經過延伸，假設將規定選取用來進行分析的資料。

導入工具

從訓練集中建立模型的運算法則。MineSet 支援四種分類導入工具 (決策樹、選項樹、証據、決策表) 以及一種回歸導入工具 (回歸樹)。

實例

用於記錄或列的術語。

知識發現

知識發現是尋找資料中新的、值得注意並且有用之模式的過程。資料挖掘是知識發現的子集。它允許根據資料提出可用於檢驗的新假設。

標籤

預測模型試圖預測的單獨欄。使用者必須預先選擇標籤。例如：「蝴蝶花類型」是蝴蝶花資料集中一個很好的候選標籤。

上升度

置信度與關聯規則產生器所建立規則的期望置信度的比值。通常，這個數字越大，規則就越值得注意。

上升曲線

評價模型預測準確性的方法。上升曲線可視地顯示分類器比隨機確定標籤的的優越處。

平均值

平均值可視爲樣本中資料值的總和除以資料分組數所得到的數。平均值通常由 x- 列表示，一般稱爲「平均數。」(例如： 1、2、3 的平均值 (1 +2 +3)/ 3 爲 2。)

中值

當數字以大小順序排列時中間的數。

模型

根據觀察到的現象描述，通常省略了某些細節。模型可能隱含預測。例如：如果形成直接郵件模型，就意味著「這就是我們認爲的直接郵件用戶的樣子。」學習者可以更改值並觀察更改對系統操作的影響。請參閱「分類器」。

可預測性

「置信度」的另一個術語。

先驗機率

分類標籤的先驗機率是在忽略所有屬性值的情況下，在隨機選擇記錄的資料中發現該標籤的機率。從數學的角度來說，這個紀錄數是帶有分類標籤的記錄數除以記錄總數所得到的。(請參閱「條件機率」。)

純度

「欄重要性」相關內容中的純度是標籤值分布的正確性度量。累積純度是資料關於標籤值劃分資料的好壞程度之度量。使用欄劃分的資料在「決策樹」中以同樣方式進行劃分時被認爲是重要的資料。分組中的每個集合都有自己的純度測量，並且分組內的純度測量是這些單獨測量的組合。對於分組中指定的集合，如果每個分類都有相同的表示法，則純度爲 0，如果每個記錄都是同一類，則純度爲 100。類似地，如果分組中每個集合的分類表示法相同，則累積純度爲 0；如果分組中的每個集合包含的記錄都有相同的分類，則累積純度爲 100。在 MineSet 中，純度取決於「熵」。

隨機子

隨機種子是爲選擇隨機樣本資料而選擇特定起始點的方法。當需要不同的隨機樣本時，可以指定另外的隨機子。對不同的資料集資料挖掘過程使用同樣的隨機子，可以使您每次都使用同樣的隨機樣本工作。當您需要測試研究中所發現的特定模式的穩定性時，可以改變隨機子。

範圍

欄中最大和最小可能值的差。

回歸器

一種預測模型，其中標籤的取值爲連續值。回歸器與分類器很相似。

關係型資料庫

關係型資料庫是資料倉庫的核心。資料及其之間的關係組織到表格中—每個項目中都包含相同欄位的記錄集合。一些欄位被指定爲關鍵字，以致按照關鍵欄位的指定值索引的搜尋可以快速檢索資料。如果在某個欄位中具有相同的值，不同表格中的記錄可能會連接在一起。具體範例如： Sybase、Informix、OLEDB、SQL server 和 Oracle。

規則

規定規則爲用於處理資料的行爲模式，例如：關聯規則或分類規則。它也是決策樹中從根到葉分類資料的唯一路徑。基於規則的系統按照指定的程序，將一組「if-then」規則應用到一組事實上以進行推斷。

投資回報 (ROI)
一個金融術語，它經由測量投資的利潤 (回報) 增值來衡量項目的價值。也稱爲 ROI。

列
關係表中的記錄。

模式
資料集中所有欄的描述。模式包括每一欄的名稱和類型資訊。模式可以儲存在結尾爲 .schema 的檔案中。

偏度
頻率分布的不對稱程度。

標準偏差
資料的分散測量。定義爲方差的平方根。

支援度
如果指定關聯規則 *X* --> *Y* (X 暗示 Y)，支援將把 X 與 Y 同時在檔案中出現的頻率量化爲總記錄數的分數。例如，如果支援度爲 1%，則 *X* 和 *Y* 將在總記錄數的 1% 中同時出現。

表
正態化的關係資料集。

測試集
測試集由資料集中保留用於測試導入的分類器錯誤率的記錄組成。請參閱「訓練集」。

訓練集
MineSet 中的訓練集是資料的子集，在主要資料操作之前劃分出來用於建立分類器或模型。它由資料庫中確定標籤的記錄所組成，根據描述性屬性。導入工具用它來學習如何構造分類器。請參閱「測試集」。

刪改因子

刪改因子在分組前指定要排除在值範圍外的資料集極值分數。預設的刪改因子爲 0.05。它排除 5% 具有極值的實例 (2.5% 的最小值和 2.5% 的最大值)。刪改會減小外層對臨界值產生的影響。

統一範圍

統一範圍是資料自動分組中使用的選擇，在這些資料中取值範圍分爲統一大小的子區間。

統一權重

統一權重是資料自動分組中使用的選擇，在這些資料中取值範圍分爲指定的分組數，以便每個分組中的記錄數目相同。如果開啓記錄加權，將對範圍進行劃分讓每個分組包含的總權重相同。

權重

MineSet 支援兩種權重類型：

記錄權重是應用到資料中每一列的數值。對權重爲 k 的列，挖掘工具將當作似乎存在 k 個這樣的列來處理。也支援非整數權重。權重值由資料中使用者選擇的欄提供。

屬性權重是使用者指定的欄重要性測量。聚類運算法則使用屬性權重影響距離測量。

索引

三畫

四畫

七畫

八畫

九畫

十畫

十一畫

十二畫

十三畫

十四畫

十五畫

十六畫

十七畫

十八畫

十九畫

二十畫

二十一畫

二十三畫