Clementine 中文版

實作與導入進階研討會

SPSS Taiwan Corp.

研討會議程

- 資料採礦導入
  - 爲什麼要導入資料採礦
  - 資料採礦的應用範圍
  - 要如何導入資料採礦
  - CRISP-DM
- 資料採礦導入過程可能遇到的困難
  - 商業理解階段
  - 資料理解階段
  - 資料準備階段
  - 塑模階段
  - 評估階段
  - 部署階段

研討會議程

- Clementine中文版
  - 功能綜覽
- 實機操作–資料採礦實作時面臨問題的解決之道
  - 商業理解
  - 資料理解
  - 資料準備
  - 塑模
  - 評估
  - 部署
- Q&A

資料採礦導入

資料採礦導入

- 企業/組織的現況
  - 企業組織運作及進行活動時產生大量的資料
  - 有足夠的IT技術把資料儲存起來
  - 面對劇烈的競爭，希望有更大的優勢
  - 如電信業所面臨的狀況：
    - 剛開始是要極力爭取手機市場搶客戶
    - 現在是人人都有手機的時代，競爭越來越激烈且複雜
    - 各種新規格的手機推出（如 3G）
    - Portable 的手機號碼

資料採礦導入

- 資料採礦的目的
  - 從大量的資料中找出有意義且意想不到的資訊
  - 如電信業可從資料中找出客戶的行爲模式，進而找出可能流失的客戶，採取行動挽留
- 將資料採礦導入企業/組織，大大提升競爭力
  - 配合 IT 技術，讓資料採礦融入企業組織的流程當中，如
    - 過去用 Call Center 的方式做 CRM
    - 現在可以做到針對每個客戶個人化的系統
    - 除了電腦之外，再加上手機、PDA 的應用，更貼心的個人化、自動化

資料採礦導入

- 資料採礦的應用
  - 降低成本
    - 找出目標客戶，不用亂槍打鳥
    - …
  - 增加營收
    - 向上銷售（Up-Selling）
    - 交叉銷售（Cross-Selling）
    - 拓展新客戶群
    - …
  - 提高價值
    - 客戶滿意度、忠誠度
    - 客戶區隔
    - …
  - 風險控管
    - 評估風險
    - …

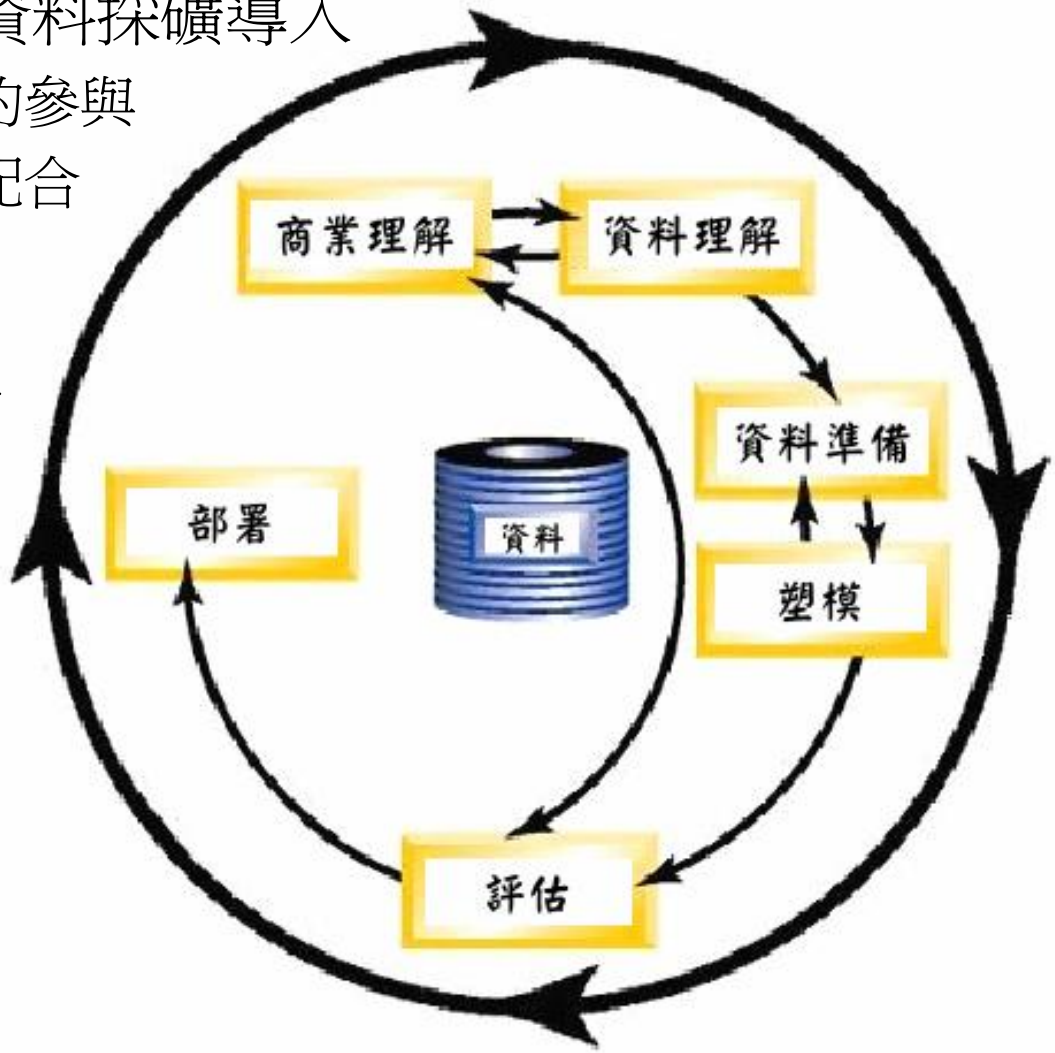
資料採礦導入

- 資料採礦的應用
  - 其他
    - 犯罪、詐欺偵測
    - 氣象預測
    - 交通流量控管
    - Text Mining
    - Image Mining
    - …

資料採礦導入

- 如何有效且成功地將資料採礦導入
  - 高層主管、相關人員的參與
  - 硬體設備及軟體工具配合
  - 資料採礦標準流程

CRISP-DM



資料採礦標準流程 CRISP-DM

- CRISP-DM
  - **Cr**oss-**I**ndustry **S**tandard **P**rocess for **D**ata **M**ining
  - SPSS 和 NCR 在 1996 年爲克萊斯勒做資料採礦時訂出的一套標準程序，集合專家意見修訂，目前版本爲1.0
  - 分爲六大步驟
    - 商業理解 (Business Understanding)
    - 資料理解 (Data Understanding)
    - 資料準備 (Data Preparation)
    - 塑模 (Modeling)
    - 評估 (Evaluation)
    - 部署(或佈署) (Deployment)

資料採礦導入的困難

資料採礦導入的困難

- 導入資料採礦難不難？
  - 爲什麼成功的案例不多？
  - 現實狀況到底和紙上談兵有多大的差別？
  - 這些問題是不能解決的嗎？

資料採礦導入過程
可能遇到的問題

1. 商業理解階段可能遇到的問題

- 沒有抓到真正企業/組織所面臨的問題
  - 最後做出的結果自然難以協助解決問題
  - 必須將商業性的問題具體化成爲資料採礦的目標
- 不了解目前企業組織運作的流程
  - 不了解資料從何而來
  - 做出來的結果難以和目前系統整合
  - …

2. 資料理解階段可能遇到的問題

- 不清楚資料的來源
  - 無法發現資料潛在的問題
  - 誤會資料的意義
- 不了解資料的儲存方式
  - 之後會在在整理或分析階段造成困難
- 沒有詳細檢視資料的狀況
  - 可能有不適合使用的變數或資料
  - 找出需要處理的問題（如遺漏值、錯誤的資料）

3. 資料準備階段可能遇到的問題

- 不同來源的資料合併
- 選擇資料和變數
  - 刪除不需要或不能用的資料和變數
- 把資料整理成能夠分析的格式
  - 不同的模型所需要的資料格式也不盡相同
  - 文字、數字資料的轉換等
- 產生新變數
  - 產生對分析會有幫助新變數
- 解決在資料理解階段發現的問題
  - 如遺漏值怎麼補？

4. 塑模階段可能遇到的問題

- 演算法要選用哪一種？
  - 若有兩種以上的演算法都適用，要選哪一個？
- 模型的參數怎麼設效果會比較好？
  - 如決策樹的層數要設幾層？子節點的資料個數要設多少？
- 是否能把不同模型的特性結合起來使用？
  - 混合模型

5. 評估階段可能遇到的問題

- 如何挑出最適合的模型？
  - 正確率最高不見得代表最能解決真正的問題
- 模型能帶來的效益有多少？
  - 使用此模型的成本、收益
  - 投資報酬率
- 做好的模型究竟值不值得導入現有流程？
  - 是否符合「商業理解」階段的衡量指標？
  - 是否需要重新修正模型？

6. 部署階段可能遇到的問題

- 把資料採礦結果部署出去的方式要如何選擇？
  - 給誰用？如何用？
  - 單機使用還是 Client-Server 或其他架構？
- 做好的成果要如何和其他軟、硬體結合？
  - 如何把採礦軟體做好的結果交給開發人員使用？
  - 安全性的考量
  - 效能的考量
  - 使用的方便性
  - …
- 部署後要如何監控、檢視、並反饋結果？

Clementine 中文版

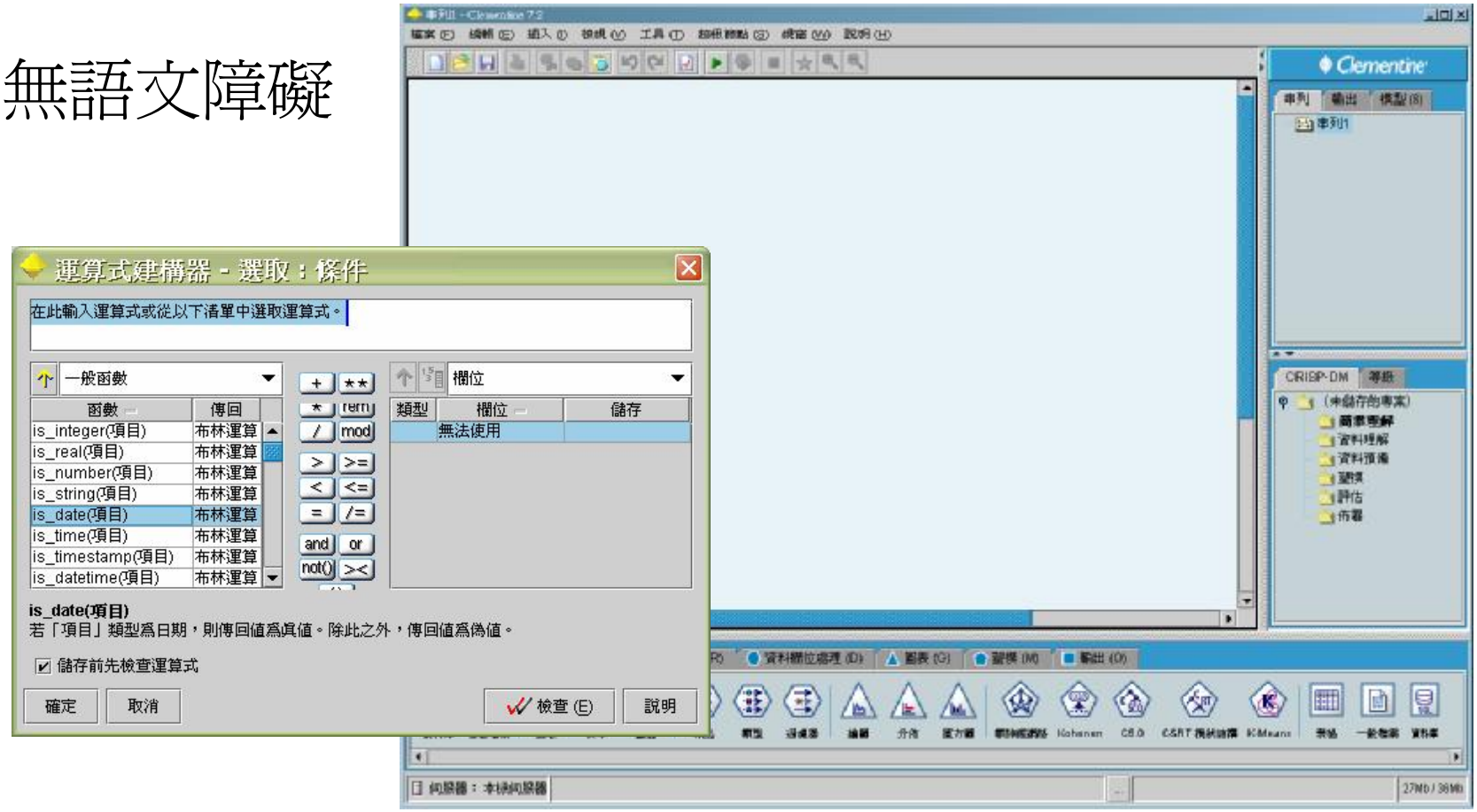
最佳資料採礦工具

SPSS Taiwan Corp.

Clementine 中文版功能綜覽

- 中文化介面

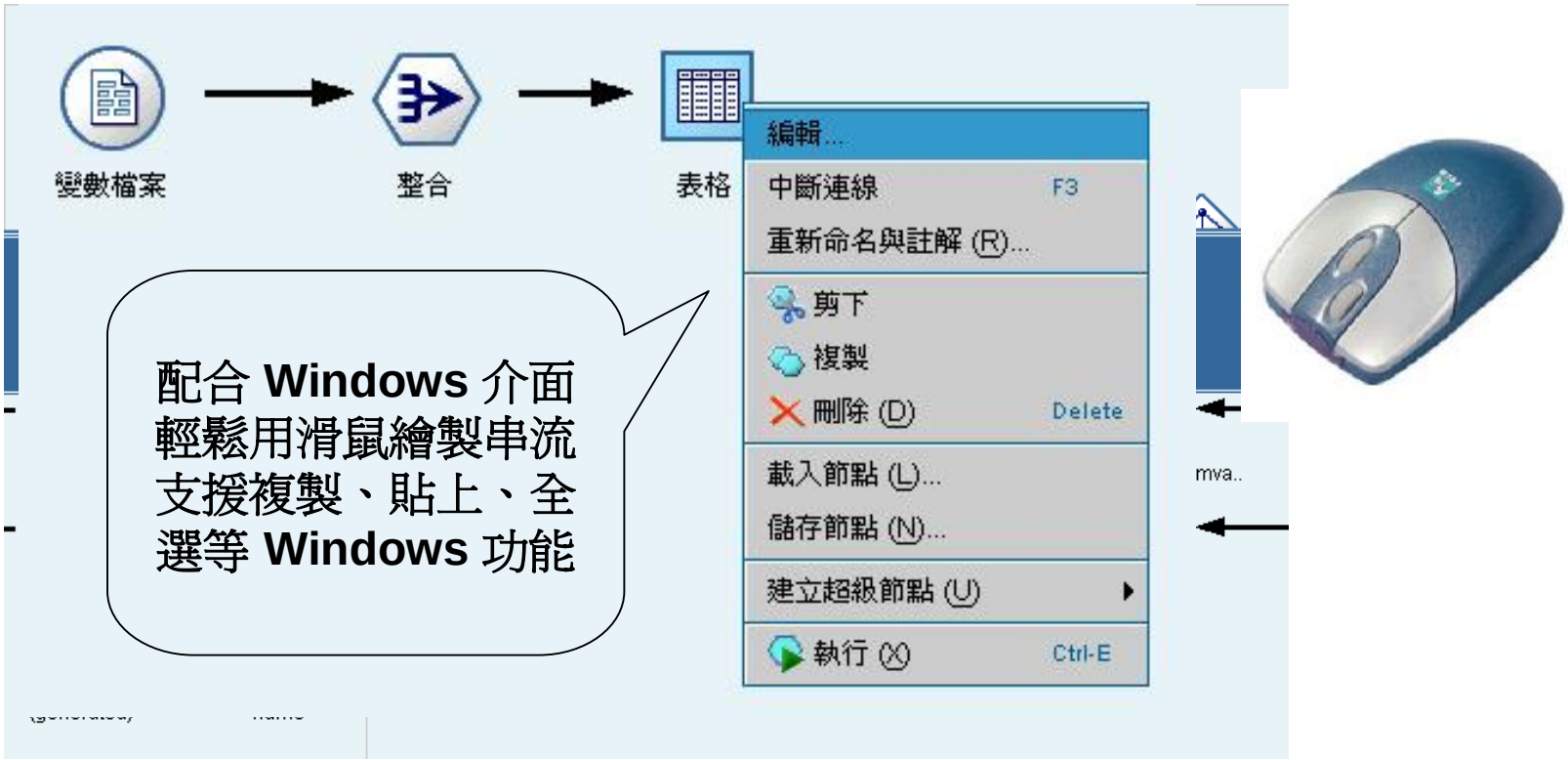
  無語文障礙



Clementine 中文版功能綜覽

§ 視覺化資料採礦

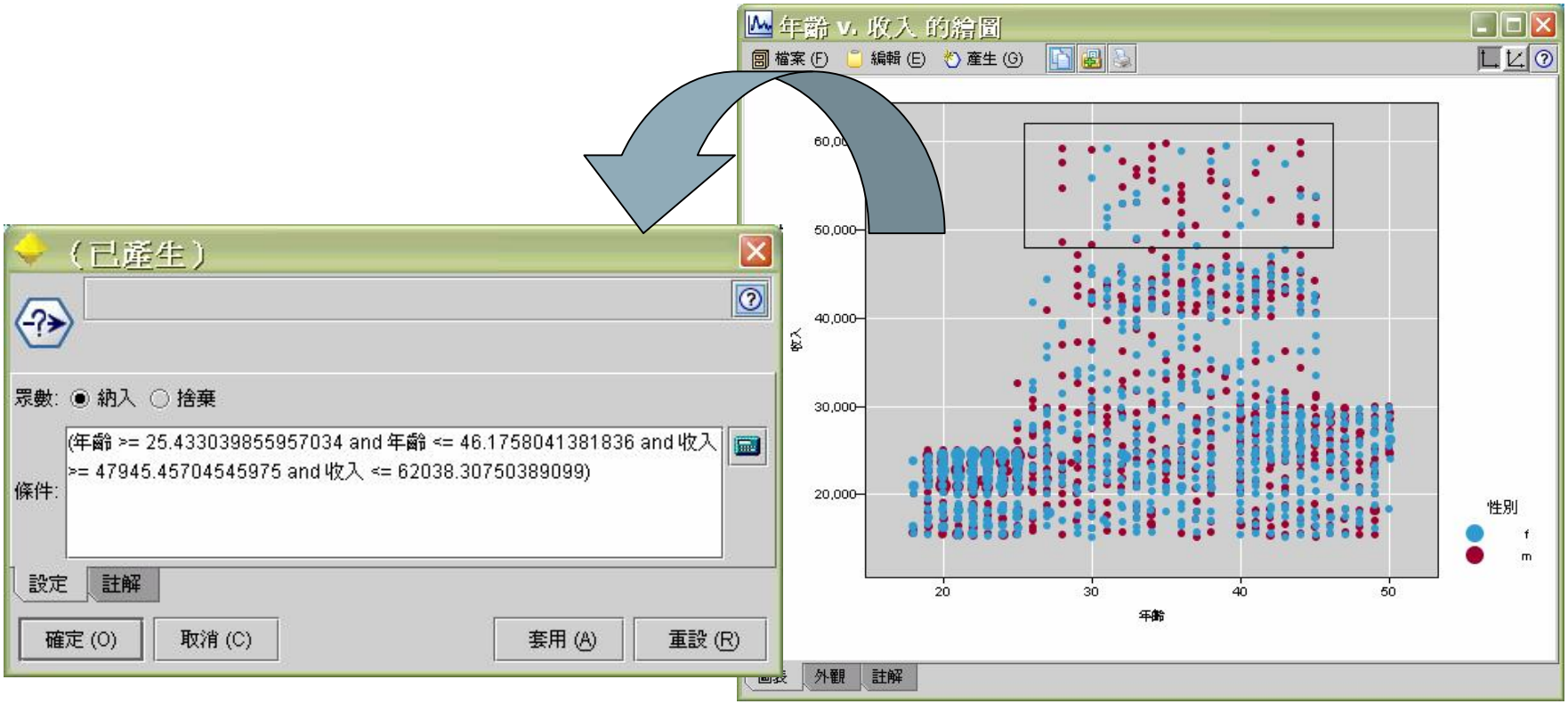
利用「串流」（Stream）進行資料採礦



Clementine 中文版功能綜覽

- 視覺化資料採礦

  用滑鼠選取有興趣的範圍，自動產生節點



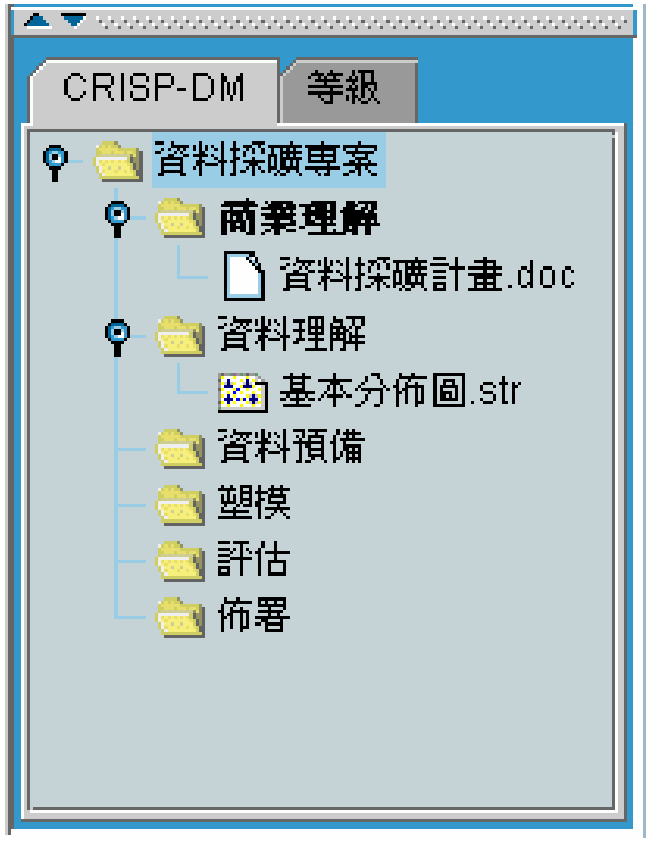
Clementine 中文版功能綜覽

- 軟體內建資料採礦標準流程CRISP-DM

  進行採礦時能跟著流程走

  可將外部文件也引入

  也可依自身需求增減資料夾



Clementine 中文版功能綜覽

- Clementine 提供採礦流程各階段會用到的工具

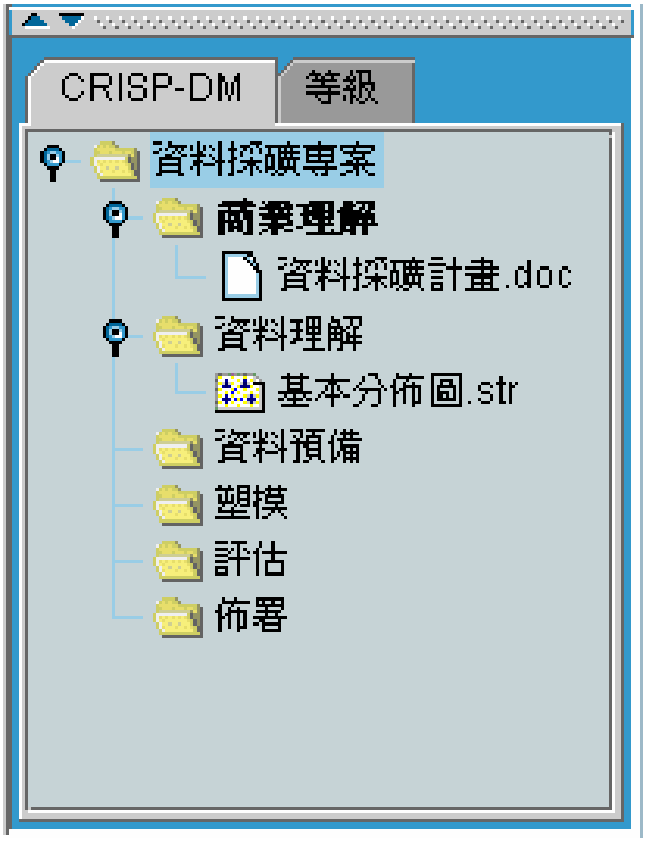
可讀取多種資料來源

多樣化圖表清晰呈現資料特性

完整的資料處理功能

12種統計和人工智慧的演算法

豐富的部署策略選擇



Clementine 中文版功能綜覽

§ 開放性

可和任何有支援ODBC的資料庫連結抓取資料，並將部份工作 Push-back 到資料庫中

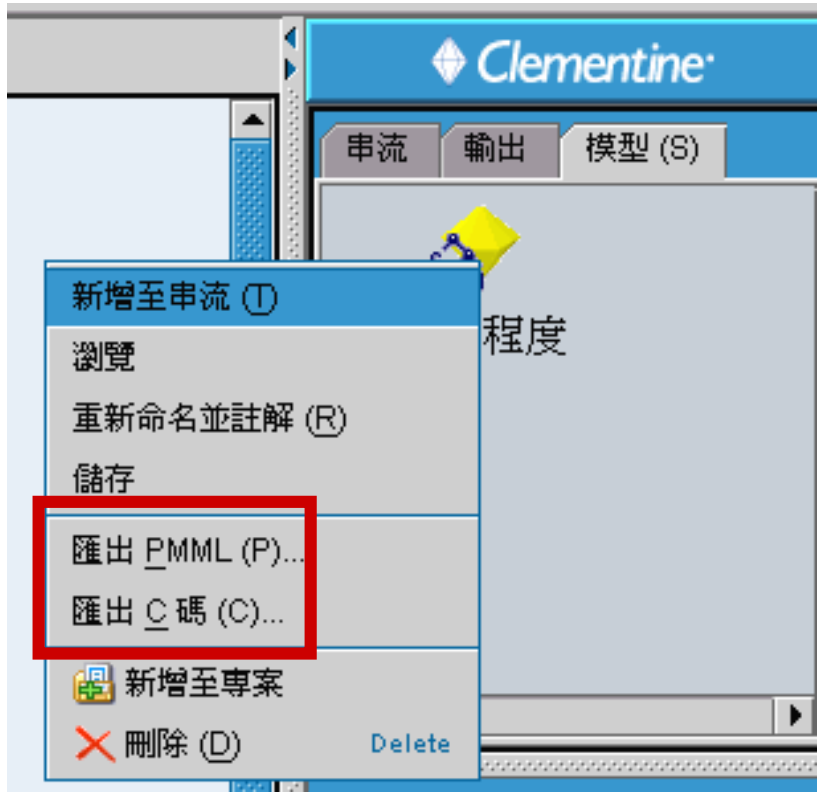
可自行將其他演算法或功能透過CEMI加入 Clementine 使用

Clementine 中文版功能綜覽

- 能將模型結果部署出去

  模型能輸出爲 PMML（Predictive Model Mark-up Language）

  模型能輸出爲 C 碼

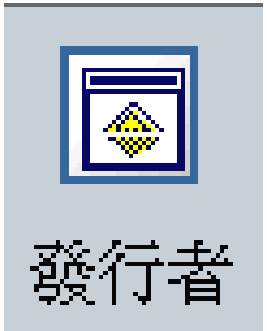


Clementine 中文版功能綜覽

- 能將整個採礦過程部署出去 – CSP

  透過 CSP（Clementine Solution Publisher ）可將「整個資料流程」包裝輸出爲 Image 檔案和 Parameter 檔案，日後無需使用 Clementine 即可進行該工作

  提供 API，程式撰寫人員可呼叫此 Image 檔，直接和系統整合



Clementine 中文版功能綜覽

- 效能
  - In-Database Mining（XML）
  - 快取
  - 設定使用記憶體
  - Batch排程執行
  - Client-Server 版本
  - …

實機展示

用Clementine中文版
解決採礦導入的問題

資料採礦導入過程可能遇到的問題與解決之道

1. 商業理解

資料採礦導入實作的問題與解決之道
1. 商業理解階段

- 如何把商業問題化爲資料採礦目標？
  - 比如商業問題是「A商品的業績不佳，要如何提升」，要如何用轉成採礦目標？
  - 要了解公司運作的方式，找出可行的方法，如：
    - 交叉銷售→關聯規則
    - 向上銷售→分類、…
    - 找出目標客戶群→分群、…
    - …

資料採礦導入過程可能遇到的問題與解決之道

2. 資料理解

資料採礦導入實作的問題與解決之道
2. 資料理解階段

- 讀取不同來源的資料
  - Clementine 提供多種資料來源的連結
    - ODBC（包括 Excel）
    - 各種文字檔案
    - SPSS 資料檔案
    - SAS 資料檔案
    - 使用者輸入
  - Clementine 可同時存取多種資料來源

資料採礦導入實作的問題與解決之道
2. 資料理解階段

- 讀入的資料形態是否正確？
  - 日期變數是否被當成是日期？
    - 數值型：20050801-20050701 = 100
    - 日期型：20050801-20050701 = 31天
  - 用數值表示的「類別」變數是否被視爲類別？
    - 如用 1 代表男性、2 代表女生
    - Clementine 本身有「防呆」機制，只有能用的資料型態能被選取

資料採礦導入實作的問題與解決之道
2. 資料理解階段

- 資料的特殊狀況
  - 最前面有幾行說明註解文字不是資料的內容
  - 資料兩側多了看不見的「空白」
  - 資料中有引號
  - …

資料採礦導入實作的問題與解決之道
2. 資料理解階段

- 如何開始了解資料？
  - 了解變數的意義
  - 了解單一變數的分佈情形
  - 了解各變數的品質
  - 了解各變數和目標變數的交互作用
  - 可用的方法：
    - 統計量
    - 圖形
    - 表格
    - …

資料採礦導入實作的問題與解決之道
2. 資料理解階段

- 利用統計量來了解資料
  - 了解資料的分佈是否異常
    - 平均數、最大最小值、全距、…
    - 使用時機：數值型變數
  - 了解資料的散度
    - 變異數、標準差、…
    - 使用時機：數值型變數
  - 兩變數之間的相關程度
    - 皮爾森相關係數
    - 使用時機：數值型變數

資料採礦導入實作的問題與解決之道
2. 資料理解階段

- 了解資料的遺漏狀況
  - 遺漏值的形式：
    - NULL
    - 空白
    - 用其他符號代表

資料採礦導入實作的問題與解決之道
2. 資料理解階段

- 用圖形來了解資料
  - 直方圖
    - 單一連續型變數
  - 分佈圖
    - 單一類別型變數
  - 收集圖
    - 兩個連續型變數
  - 多重繪圖
    - 多個連續變數對一個連續變數
  - 繪圖
    - 二~三個變數，不限類型
  - 關聯網
    - 類別變數各類項之間的關係

資料採礦導入過程可能遇到的問題與解決之道

3. 資料準備

資料採礦導入實作的問題與解決之道

3. 資料準備階段

- 不同來源資料的合併
  - 應分析需要，將多個資料表合併爲單一資料表
  - 合併的方式：
    - 增加資料筆數
    - 增加變數個數

資料採礦導入實作的問題與解決之道

3. 資料準備階段

- 遺漏值處理
  - 用其他資訊可以來填補
  - 填入固定的值
  - 填入平均數、眾數、其他函數等
  - 用模型來填補
  - …

資料採礦導入實作的問題與解決之道

3. 資料準備階段

- 資料格式的轉換
  - 資料原本儲存的樣子並不符合分析或其他工作的需要
  - 可用的轉換方式：
    - 設成旗標
    - 整合
    - 時間資料轉換
    - …

資料採礦導入實作的問題與解決之道

3. 資料準備階段

- 舊變數轉換、新變數產生
  - 生日轉成年齡
  - 身份證字號轉姓別
  - 電話轉居住區域
  - 由各行業本身的專業知識產生新的有用變數
  - …

資料採礦導入實作的問題與解決之道

3. 資料準備階段

- 資料切割
  - 爲何需要切割資料
    - 建立模型需將資料分成「訓練組」與「測試組」
    - 想建立不同區域或時間的模型
    - 想把特殊狀況特別分析
    - …
- 是否需要抽樣
  - 當資料量太大時，可以先抽一部份的資料，初步了解資料的狀況或是測試適合的模型

資料採礦導入實作的問題與解決之道

3. 資料準備階段

- 如何選擇有用的變數
  - 變數過多時不見得好
  - 不該使用無分析意義的變數
  - 資訊重覆（共線）的變數不宜同時使用
    - 用主成份分析
    - 用相關係數
    - 畫兩變數的散佈圖

資料採礦導入實作的問題與解決之道
3. 資料準備階段

- 平衡資料
  - 使用時機：稀有事件
  - 可用的方式
    - 誤差抽樣
    - 加權
    - 錯誤成本
  - 沒有平衡可能有什麼後果？
    - 導致模型只會預測我們不感興趣的目標
    - …

資料採礦導入實作的問題與解決之道

3. 資料準備階段

- 自訂節點
  - 可以利用超級節點做成客製化的節點
  - 不需要會寫程式
  - 且可設超級節點參數
  - 可儲存供日後使用

資料採礦導入過程可能遇到的問題與解決之道

4. 塑模

資料採礦導入實作的問題與解決之道

4. 塑模

- 如何選擇要用的模型？
- 模型的參數如何設定比較適合？
- 要建立單一的模型或是不同區域建不同的模型？
- 是否能把不同模型混合使用截長補短？

資料採礦導入實作的問題與解決之道
4. 塑模

- 分類模型（Classification）
  - 目標變數爲類別型時
    - 如：銀行將客戶分類成「會申請信用卡」及「不會申請」
  - 可選用的演算法
    - 類神經網路：無法產生能解釋的規則，可配適非常複雜的模型
    - 決策樹（C5 或 C&RT）：可產生規則
    - Logistic 迴歸：有統計上的前題假設

資料採礦導入實作的問題與解決之道
4. 塑模

- 估計模型（Estimation）
  - 目標變數爲連續型時
    - 如：零售商估計每個客戶的線上銷費額
  - 可選用的演算法
    - 類神經網路
    - 迴歸：有統計上的前題假設
    - C&RT：決策樹的方法

資料採礦導入實作的問題與解決之道
4. 塑模

- 分群模型（Clustering）
  - 沒有目標變數，非監督式
  - 如：電信業想把客戶做區隔
  - 可選用的演算法
    - K-Means：一維的分群，用「距離」的概念計算
    - Kohonen：利用類神經自我組織的演算法做二維度分群
    - 2-Step：可自動找出最適的群集數

資料採礦導入實作的問題與解決之道
4. 塑模

- 關聯規則模型（Association）
  - 找出各變數間的關聯性
  - 如：超市找出哪些商品可以組合販售
  - 可選用的演算法
    - Apriori：連續、類別變數都可用
    - GRI：只能用類別變數
    - 序列：有時間先後，只能處理類別型的，且需要有時間欄位

資料採礦導入實作的問題與解決之道
4. 塑模

- 混合模型
  - 爲什麼要用混合模型？
    - 模型各有其優缺點，混合模型可截長補短
  - 何時可使用混合模型？
    - 提高準確率
    - 協助解釋
    - 處理共線性問題
    - …

資料採礦導入過程可能遇到的問題與解決之道

5. 評估

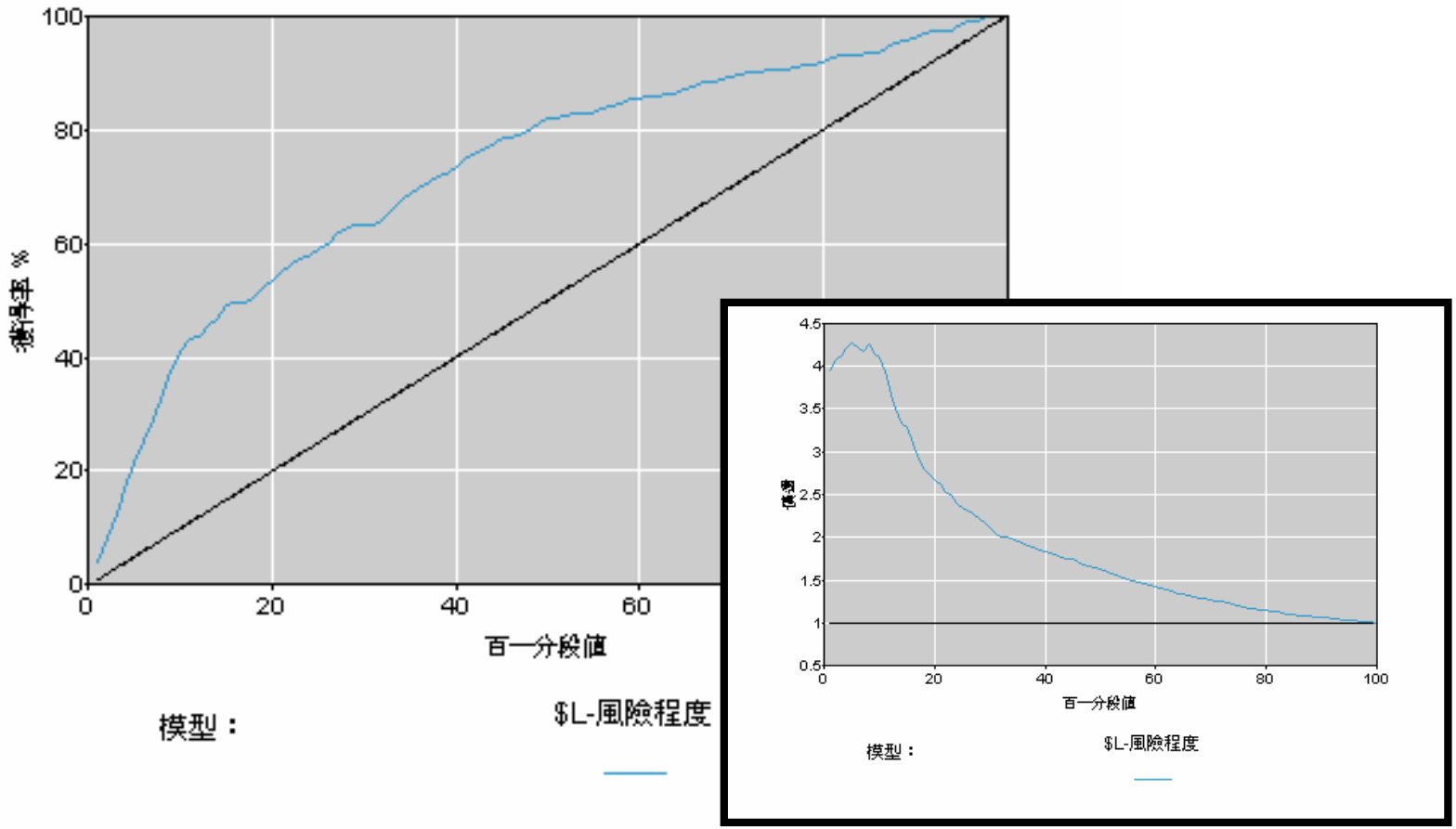
資料採礦導入實作的問題與解決之道
5. 評估

- 如何評估模型是否適合且達到商業目標？
  - 正確率
  - 增益圖
  - 投資報酬率圖（ROI）
  - 獲利圖

資料採礦導入實作的問題與解決之道

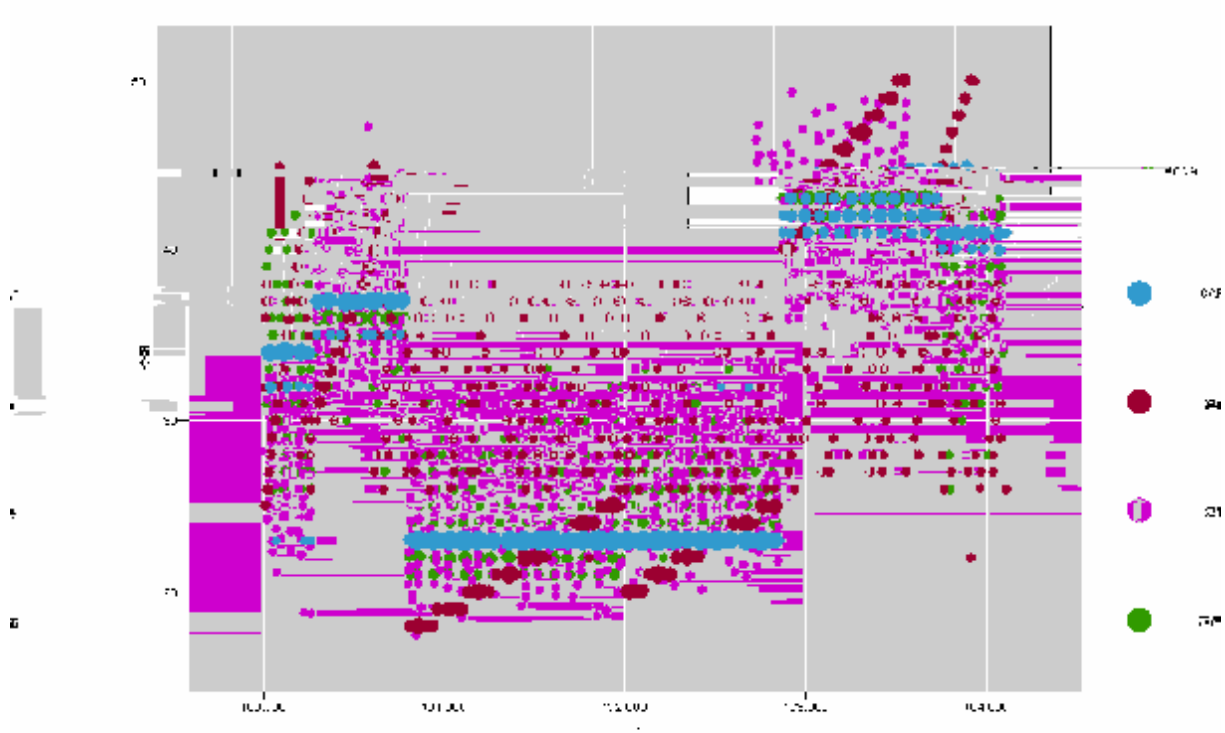
5. 評估

§ 增益圖要如何看？



資料採礦導入實作的問題與解決之道
5. 評估

- 還有什麼其他的評估方式？
  - 錯差矩陣（Coincidence Matrix）
  - 散佈圖畫預測結果
  - …



資料採礦導入過程可能遇到的問題與解決之道

6. 部署

資料採礦導入實作的問題與解決之道
6. 部署

- 將採礦結果部署
  - 如何自動化做固定的報表？
  - 要選批次的方式，還是線上即時自動化的應用？
  - 若建立自動化系統，要考慮什麼？
    - 選擇適合管道（網站、內部系統…等）
    - UI的設計
    - 如何把模型或整個資料採礦流程和現有系統結合
    - 效能
    - 安全性
    - …

採礦過程會遇到的問題
Clementine 通通幫您解決

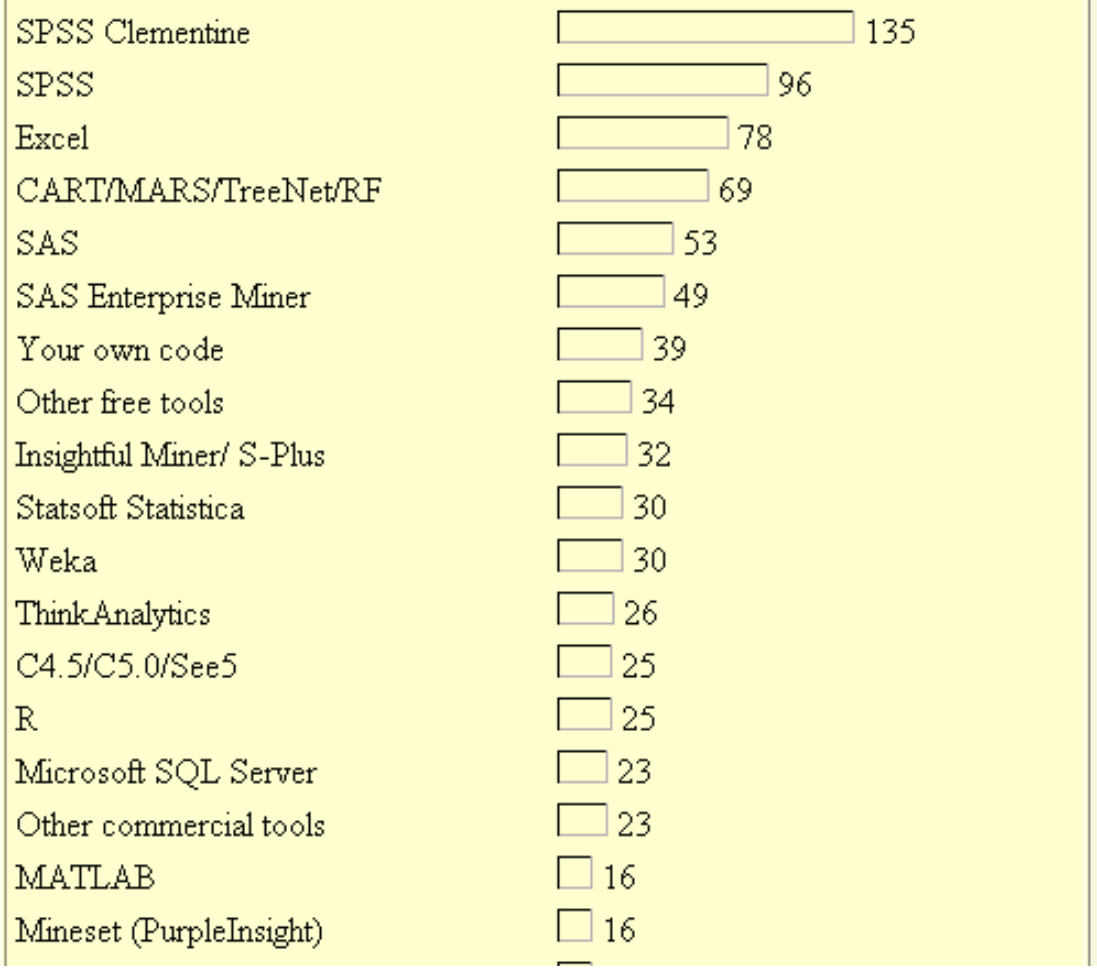
第一名資料採礦工具

第一名資料採礦工具

- KDnuggets 2005 年票選第一名，連續數年蟬連寶座
- www.kdnuggets.com

Poll

Data mining/Analytic tools you used in 2005 [376 voters, 860 votes total]



謝謝您的熱情參與

Lilian Chiu

spsstechsupport@mail.sinter.com.tw

SPSS Taiwan Corp. http://www.sinter.com.tw/SPSS

宏德國際軟體諮詢顧問股份有限公司

台北 TEL:02-25771100 新竹 TEL:03-5526100 台中 TEL:04-23283566 高雄 TEL:07-2251456