第五章應用案例

應用案例 5.1

1-800-Flowers的企業分析和資料探勘應用

1-800-Flowers為禮品零售業中最知名與最成功的品牌之一。30多年來，這家總部位於紐約的公司提供給全世界顧客各個場合需要的最新鮮花卉以及最佳植物、禮物籃、美食、糕點以及可愛填充玩具。由Jim McCann在1976年成立，14年前在架設自己的網站後，1-800-Flowers很快成為直接訂購電子商務的領導。

問題

與許多其他電子商務公司一樣，在成功後，1-800-Flowers必須即時決定增加留客率、降低成本以及讓最佳顧客能夠一再回流。當企業由單一家花店成長為服務3,000多萬名顧客的線上禮品零售商時，必須要盡可能的成為佼佼者，才能在競爭中保持領先。

解決方案

堅信親密顧客關係的價值，1-800-Flowers想要透過分析手中每一筆資料，更了解顧客的需求與需要。1-800-Flowers決定使用SAS資料探勘工具更深入挖掘他們的資料資產，以找出其顧客的新模式以及將知識轉變為企業交易。

結果

根據McCann，SAS的企業分析與資料探勘工具讓1-800-Flowers即使在經濟規模變大時，企業還是能成長。在其他零售商為存活而努力時，1-800-Flowers的營收持續成長，並且在過去五年成長了近乎兩倍。

分析的特定助益如下：

- **更有效的行銷活動**(more efficient marketing campaigns)。1-800-Flowers區隔顧客以大幅減少寄發傳單的時間。顧客知識管理部門的副董事長，Aaron Cano表示：「過去必須要花2至3星期的時間，現在只要2至3天。這讓我們有時間作更多分析以及確實傳達相關訊息。」
- **減少郵件、增加回應率**(reduced mailings, increased response rates)。公司能夠顯著減少行銷郵件，同時增加回應率以及更仔細選擇電視與廣播廣告。
- **更佳的顧客經驗**(better customer experience)。回流顧客再次登入1-800-Flowers.com時，該網站馬上顯示顧客感興趣的相關選項。Cano表示：「如果顧客通常都是購買鬱金香送給老婆，我們會顯示出最新以及最佳的鬱金香商品」。
- **增加回流銷售**(increased repeat sales)。公司的最佳顧客會多次回流通常是因為1-800-Flowers了解顧客身分以及喜好。公司讓購物經驗更簡單以及有效，並且在聯絡點就對顧客行銷。因為使用企業分析與資料探勘，1-800-Flowers減少其營業費用，增加80%最佳顧客群保留率，吸引2,000萬名新顧客以及將整體回流企業從40%增加為50%以上（10%的所有品牌回流銷售等於是4,000萬額外的企業營收）。

資料來源：“SAS Helps 1-800-Flowers.com Grow Deep Roots with Customers,” sas.com/success/1800flowers.html (accessed on May 26, 2009) “Data Mining at 1-800-Flowers,” kdnuggets.com/news/2009/n10/3i.html (accessed on May 23, 2009).

應用案例 5.2

執法機關使用資料探勘更進一步打擊犯罪

在這不佳的經濟條件中，全世界各地的警察局都正面對著困難時刻，除了更少的線索、更多案件以及越來越複雜的犯罪外，必須以持續縮減的資源來打擊犯罪。在英國的一處警察局中，調查人員發現這些挑戰限制了他們能處理的案件。許多案件沒有絕對線索，例如沒有明顯證據的闖空門以及汽車失竊，除非有新的證據發現，否則通常都會擱置一旁。所以，警察部門決定使用可以輕易快速在未解決的犯罪案例中找出模式與趨勢的方法。

在警察部門中的每個電子檔都包含竊盜以及作案手法 (modus operandi, MO) 的實體說明。儘管許多缺乏證據的案件之前都先另外存檔，現在該警局正在以更快的速度重新審查。在 PASW Modeler（之前為 Clementine）中，資料塑模器使用兩個 Kohonen 類神經模式以及 MO，然後結合分群檢視說明哪些實體群組與 MO 群組相同。如果找到相符者，發現加害人有一個或一個以上罪行，就可能可以偵結由同一個人所犯下之懸而未決的案件。

分析團隊更進一步使用統計方法確認相同點的重要性，以調查分群。如果分群顯示出有同樣的犯罪案件，警方就會重新開始並且調查其他罪行。或者，如果罪犯未知，但是大型分群顯示為同一人所犯，就會結合這些案件，進行重新排序。警方也調查檔案上屢犯的行為，目標為確認符合其行為模式的罪行。警方希望 PASW Modeler 可以啟動舊案件並且連結已知加害人。

另一個在美國的警察部門也面對類似的挑戰：沒有足夠的資源，再加上數目漸增的案件。為了產生大範圍罪行與社區失序的永續性解決方法，警方採用社區為主的警察理念，即需要市民與社區機構合作夥伴的參與，以及仔細分析刑案資訊的全面方法。基本的流程目標為藉由確定根本原因、教育社區問題的嚴重性以及與社區擬定協調解決方法，有效解決該原因，找出長期解決方法。主要的挑戰為說服社區其之參與對有效執行的重要性。

警方使用 PASW 統計分析以及資料探勘軟體工具，執行更廣泛的資料分析，藉以發現與案件有強力相關的變數以及評估市民對於社區警力的滿意度。此分析的結果，強烈證明社區參與加上智慧資料分析是在經濟困難時，擬訂有效長期解決方法的必要成份。

全球警方使用創新的二十一世紀方法，應用資料探勘技術避免犯罪活動，提升打擊犯罪技術。應用成功的故事在主要資料探勘工具、解決方案提供者以及主要顧問公司網站上都有記載（例如， SPSS、 SAS、 StatSoft、 Salford Systems）。

資料來源：Based on C. McCue “Connecting the Dots: Data Mining and Predictive Analytics in Law Enforcement and Intelligence Analysis,” *Police Chief Magazine*, Vol. 70, No. 10, October 2003; “Police Department Fights Crime with SPSS Inc. Technology,” spss.com/success/pdf/WMPCS-1208.pdf (accessed July 25, 2009); “North Carolina Law Enforcement Agency Identifies Crime Areas and Secures Community Involvement,” spssshowcase.co.uk/success/pdf/CMPDCS-0109.pdf (accessed on September 14, 2009).

應用案例 5.3

車禍與駕駛分心

駕駛人分心是高速公路安全的重要議題。由美國國家高速公路交通安全管理局 (National Highway Traffic Safety Administration, NHTSA) 在 1996 年發佈的一項研究指出，約 25%～30%車禍造成的傷害是因為駕駛人分心。在 1999 年，根據美國國家統計與分析中心 (National Center for Statistics and Analysis, NCSA) 的事故報告分析系統 (Fatality Analysis Reporting System, FARS)， 11%的死亡車禍（即 4,462 人死亡）都是因為在駕駛時分心。

一份研究遂抽取交通意外的分心因素模式，以資料探勘擷取 FARS 車輛資料的相關性與關聯性，將三種資料探勘技術（ Kohonen 型類神經網路、決策樹以及多層感知類神經網路）都用來找出不同的相關與潛在可能解釋高意外發生率的分心因素組合。 Kohonen 型類神經網路確認類神經分群並且顯示資料收集中輸入變數模式，決策樹探討並且將每件意外對連續事件的影響分類以及指出分心的駕駛與體能／精神狀況之間的關係，最後，多層感知類神經網路模式經訓練與測試後發現在交通意外中分心與其他駕駛人相關因素之間關係。也使用 SPSS Clementine 探勘 FARS 資料集中三種模式種類。

預測與探討模式發現發生車禍的 1,255 位駕駛人主要都是因為分心。在其他數個輸出變數中，車尾、面對面以及其他角度撞擊對車禍發生以及嚴重性有顯著影響。

資料來源：W. S. Tseng, H. Nguyen, J. Liebowitz, and W. Agresti, “Distractions and Motor Vehicle Accidents: Data Mining Application on Fatality Analysis Reporting System (FARS) Data Files,” *Industrial Management & Data Systems*, Vol. 105, No. 9, January 2005, pp. 1188 – 1205; and J. Liebowitz, “New Trends in Intelligent Systems,” Presentation made at University of Granada, docto-si.ugr.es/seminario2006/presentaciones/jay.ppt (accessed May 2009).

應用案例 5.4

恐怖份子金援來源探勘

2001年9月11日世界貿易中心遭受恐怖攻擊，強調出開放來源情報的重要性。美國愛國者法案以及美國國土安全部門 (Department of Homeland Security, DHS) 的成立意味著資訊科技與資料探勘技術的潛在應用，以偵測洗錢與其他形式的恐怖活動金援。執法機構已經透過銀行與其他金融服務機構之間的正常交易，鎖定洗錢活動。

執法機構現在鎖定以國際貿易價格做為恐怖活動金援工具的部分。涉及洗錢者利用國際貿易，偷偷地在不引起政府注意的情況下，將錢運往其他國家。利用高報進口價值與低報出口價值來達成轉移。例如，國內進口商與國外出口商可以形成夥伴關係並且低報進口貨物價值，以將錢轉入國內，產生與關稅詐欺、逃稅以及洗錢等相關罪行。因此，國外出口商可能是恐怖組織的一員。

資料探勘技術鎖定來自美國商務部以及商業相關機構的進口與出口交易資料分析。追蹤超過最上層 25%的進口價格以及低於最底層 25%的出口價格。重點在於公司之間可能會利用美國可課稅項目與稅額，進行不正常價格轉移。觀察到的價格差異可以與避／逃漏稅、洗錢或恐怖金援有關。但所觀察到的價格差異可能也會是美國貿易資料錯誤所產生。

反之，資料探勘也會產生有效的資料評估，可協助打擊恐怖主義。資訊科技以及資料探勘技術在財務交易上應用可以對情報資訊有更多貢獻。

資料來源：J.S. Zdanowic, "Detecting Money Laundering and Terrorist Financing via Data Mining," *Communications of the ACM,* Vol. 47, No. 5, May 2004, p. 53; and R. J. Bolton, "Statistical Fraud Detection: A Review," *Statistical Science,* Vol. 17, No. 3, January 2002, p. 235.

應用案例 5.5

癌症研究的資料探勘

根據美國癌症協會研究，在 2009 年約有 150 萬個新診斷出的癌症病例。癌症是美國以及全世界第二大的常見死因，僅次於心臟血管疾病。今年，預計約有 56 萬 2,340 位美國人會死於癌症，每天有 1,500 人，也就是四個死亡病例中就有一個。

癌症的特色為不正常細胞無法控制成長與擴散。如果成長或擴散不能控制，就會造成死亡。即使確切的成因未知，一般認為癌症是由外在因素（例如，香菸、受感染器官、化學物質以及放射物質）以及內在因素（例如，遺傳變異、荷爾蒙、免疫情況與代謝突變）所引發。這些成因可能會一起或依序發生以引發或助長致癌物質。癌症的治療包括手術、放射線、化療、荷爾蒙治療、生物治療以及標靶治療。存活統計依癌症的種類以及診斷發現的階段有非常不同的結果。

在 1996～2004 年所診斷出的所有癌症，其 5 年相關存活率從 1975～1977 年的 50%上升為 66%。存活率的改善反映出特定癌症早期診斷與治療改善的進展。癌症的預防與治療需要更進一步改善。

即使傳統上癌症研究本質為臨床與生物，近年來資料導向分析研究是常見的互補方法。在醫學領域中，資料與分析導向研究已經有成功應用，也確定能夠增進臨床與生物研究的新方向。使用各種類型資料，包括分子、臨床、文獻為主以及臨床實驗資料與適合的資料探勘工具與技術，研究人員已經能夠確認新模式，為無癌社會奠定良好基礎。

在一份研究中， Delen (2009) 使用三個常用的資料探勘技術（決策樹、人工類神經網路以及支援向量機）與邏輯迴歸開發前列腺癌存活率的預測模式。資料集包含約 12 萬筆記錄以及 77 個變數。使用 k 摺交叉驗證法進行塑模、評估以及比較。結果顯示在此領域，支援向量機有最準確的預測（測試正確率達 92.85%），接著分別為人工類神經網路與決策樹。再者，使用敏感度分析為主的評估方法，研究也顯示出與前列腺癌存活因素有關的新模式。

在一份相關研究中， Delen 等人 (2006) 使用兩個資料探勘模式演算法（人工類神經網路與決策樹）以及邏輯迴歸，以大量資料集（20 多萬個病例）開發乳癌存活率的預測模式。基於績效比較目的，使用 10 摺交叉驗證法來測量預測模式，結果顯示決策樹（C5 演算法）為最佳預測方法，其保留樣本的正確率為 93.6%（為文獻報告中最佳正確率）；接著為 91.2%正確率的人工類神經網路以及 89.2%正確率的邏輯迴歸。進一步分析預測模式發現，存活因素的重要性排序可以做為未來臨床與生物研究的基礎。

這些範例（與其他在醫學文獻中範例）顯示進階資料探勘技術可以用來開發具高程度預測力與說明力的模式。雖然資料探勘方法能夠抽取深入隱藏在複雜醫學資料庫中的模式與關係，但若沒有醫學專家的合作與反饋意見，則結果不會有太大作用。透過資料探勘方法發掘的模式，可以由在問題領域擁有多年經驗的醫學專家來評估，以決定是否合乎邏輯、可以採取行動以及可以保證是新的研究方向。簡而言之，資料探勘不會取代醫學專家與研究人員，而是相輔相成，提供資料導向的新研究方向與達成最終拯救人類性命目標。

資料來源：D. Delen, "Analysis of Cancer Data: A Data Mining Approach," *Expert Systems,* Vol. 26, No.1, 2009, pp. 100-112; J. Thongkam, G. Xu, Y. Zhang, and F. Huang, "Toward Breast Cancer Survivability Prediction Models Through Improving Training Space," *Expert Systems with Application, 2009, in press;* D. Delen, G. Walker, and A. Kadam, "Predicting Breast Cancer Survivability: A Comparison of Three Data Mining Methods," *Artificial Intelligence in Medicine,* Vol. 34, No. 2, 2005, pp. 113-127.

應用案例 5.6

Highmark. Inc., 使用資料探勘管理保險成本

Highmark, Inc. 位於賓州匹茲堡，長久以來為其會員與社區提供可負擔、高品質健康照護服務。 Highmark, Inc. 在 1996 年創立，由賓州的兩家特許業者 Blue Cross 與 Blue Shield Association 所合併而成： Pennsylvania Blue Shield（現在為 Highmark Blue Shield）與賓州西部的 Blue Cross 計畫（現在為 Highmark Blue Cross Blue Shield）。 Highmark 目前是美國最大的健康保險公司之一。

管理照護機構的資料

在管理照護機構，像 Highmark，傳輸的資料量非常大。通常被視為佔據儲藏空間與處理上很麻煩的這些資料最近成為新知識來源。資料探勘工具與技術提供實際應用，可以分析病患資料與揭開以更低成本進行更佳管理的迷思。這是大部分管理照護公司嘗試達成的目標。

每一天，管理照護公司收到顧客數百萬筆資料項目，每一筆資訊都會更新會員的歷史案例。公司在處理時注意到資料的效用，並且使用分析工具抽取治療成本高過一般費用的病患分群。早期在使用電腦技術抽取病患相關可行動資訊時，皆受限於必須建立兩個不同疾病的連結。例如，軟體工具能夠掃描資料與報告糖尿病患或冠狀血管心臟病是治療成本最高者。然而，這些報告型軟體工具在發現這些病患罹病原因或為什麼有些病患容易受到某些疾病負面影響部分並沒有效率。藉由分析不同疾病與病患檔案的多維資訊以及產生簡單關係與相關性資料，探勘工具可以解決部分上述問題。

管理照護組織因為資料而不勝負荷，並且有些公司不希望新增資料探勘應用來增加複雜性。基於各種理由，他們可能希望掃描資料，但是不能決定為什麼或如何分析他們的資料。此種狀況對病患以及公司而言，有了良好的轉機，因為健康保險法規明定了有效資料與結構分析。

資料探勘需求

市場壓力驅使管理照護組織變得更有效率，因此可以認真考慮資料探勘。顧客要求更多與更佳服務，競爭者持續激烈競爭，這些都造成必須及時推出在設計與提供上更客製化的產品。

此客製化讓問題回到原點，即主要醫療費用產生的原因與區塊。許多組織開始使用資料探勘軟體來預測比較容易生病以及治療成本比較高的對象。規劃未來讓組織可以利用預防措施來過濾昂貴的病患並且降低醫療照護費用。另一個預測研究的應用為保費管理。擁有大批員工屬於高費用級數的資方團體，保費會增加。

根據歷史性資料，預測塑模可以預知哪些病患可能會成為公司的財務負擔。例如，預測塑模應用會將糖尿病患列為增加醫療費用的高風險，這本身不是個值得注意的資訊。但是，在 Highmark 的資料探勘裡可以找出糖尿病患以及其他病患與環境相關參數之間的關係；即，有特定心臟情況的病患可能在罹患糖尿病方面屬於高風險。此關係的界定是因為服用心臟病藥之後容易引發糖尿病。 Highmark 主管表示並證實如果沒有監測心臟病

患的用藥，就不會發現這兩者的關係。醫學研究也成功地證實許多與病患情況有關的併發症。資料探勘為更佳偵測以及適當介入計畫奠立了基礎。

資料來源：Condensed from G. Gillespie, "Data Mining: Solving Care, Cost Capers," *Health Data Management*, November 2004, findarticles.com/p/articles/mi_km2925/is_200411/ai_n8622737 (accessed May 2009); and "Highmark Enhances Patient Care, Keeps Medical Costs Down with SAS," sas.com/success/highmark.html (accessed April 2006).

應用案例 **5.7**

預測顧客流失——不同工具的競爭

在 2003 年，杜克大學 /NCR Teradata 中心尋求辨識最佳預測塑模技術，以協助無線電信通訊提供者管理棘手問題：顧客流失。雖然在零售階層，其他產業也面對顧客流失給競爭者問題，每年無線顧客轉換服務提供者的比率約為 25%或是每個月 25%。在 1990 年代早期，新訂戶的成長率為 50%，電信公司把焦點放在爭取新顧客而非留住顧客。然而，在成長率比較緩慢的年代，低到 10%，留住顧客很明顯地對整體的獲利力來說相形重要。

留住顧客的關鍵為預測哪個顧客最可能流失給競爭者並且提供他最有價值的誘因挽留。為了有效執行此策略，必須能夠開發高度準確預測——顧客流失計分卡，由此可以得知留住顧客的努力可以鎖定在相關顧客身上。

資料

資料由主要無線電信公司提供，含括 2001 年前半年的顧客資料。帳號資料為過去六個月接受該公司服務的 10 萬位顧客資料。為了協助塑模流程，針對已轉換服務顧客（在結束後 60 天離開公司）使用超取樣，所以樣本中有一半為已轉換顧客，另一半為 60 天後，還留在公司之顧客。有 170 個潛在預測因子，含括各類典型服務提供者可能有的種類。預測因子資料包括：

- 人口因子 (demographics)：年齡、地點、子女數目與年齡等。
- 財務因子 (financials)：信用額度、有無信用卡。
- 產品詳細資料 (product details)：手機價格、手機功能等。
- 電話使用 (phone usage)：號碼與通話費類型期間等。

評比標準

提供資料支援預測塑模開發。參與者（資料探勘軟體公司、大學研究中心、其他非營利與顧問公司等）必須使用最佳的模式來預測兩組不同顧客的流失機率：2001 年下半資料取得的 5 萬 1,306 個「目前」樣本以及 2002 年第 1 季 10 萬 462 位顧客資料取得的「未來」樣本。一般認為預測「未來」資料比較困難，因為外部因素以及行為模式可能隨著時間而改變。在真實情況下，預測模式都應用在未來模式中，比賽主辦單位都想要複製類似的脈絡。

每個競賽中的參賽者都必須以由大而小的顧客流失機率順序排列目前與未來的分數樣本。比賽主辦單位使用真正的顧客流失狀態計算兩個預測模式的績效指標：整體的吉

尼係數以及頂層的增益。計算兩個目前與未來樣本的指標，所以每個參賽者都有四個績效分數。評比標準在許多位置皆有詳細說明，包括比賽網站。頂層增益最容易解釋：它測量模式中最可能流失的顧客之間真正流失的數目。

結果

如果想要嘗試將模式以時間或評比標準進行最佳化，那麼競賽者可以開發每個指標的不同模式，其是所有類別的優勝者，其使用 TreeNet 軟體建立模式。 TreeNet 是廣為人知建立精確分類模式用來提升決策樹分析的創新形式。在所有參賽者中，雖然並不是所有方法都可以在比賽中適當呈現，但評審發現決策樹與邏輯迴歸通常最能精準預測顧客流失率。

Salford 的 TreeNet 模式在 171 個可能變數中最能夠找出最多的流失顧客因子，所以對預測顧客的流失而言很重要。在前 10%的顧客中， TreeNet 找出比其他競爭模式多出 35%至 45%的顧客流失，可以在隨機樣本中多找出三倍。對有大型顧客群的公司，這可以用來辨識數千位每個月可能潛在流失的顧客。以適當挽留活動留住這些顧客的努力可以為公司每年省下數百萬元。

資料來源：Salford Systems, “The Duke/NCR Teradata Churn Modeling Tournament,” salford-systems.com/churn.php (accessed April 20, 2009); and W. Yu, D. N. Jutla, and S. C. Sivakumar, “A Churn-Strategy Alignment Model for Managers in Mobile Telecom,” *Proceedings of the Communication Networks and Services Research Conference*, IEEE Publications, 2005, pp. 48-53.