以 XpertRule Miner 預測銀行客戶信用評等之研究

周彥宇、蔡汶叡、許曉芬、莊惟至、王櫻蒨
南華大學資訊管理系
李翔詣助理教授
hylee@mail.nhu.edu.tw
南華大學資訊管理系

摘　要

現今的「微利時代」，消費者荷包緊縮，盲目的行銷活動將嚴重影響企業獲利的能力。銀行為爭奪有限的客源，除投入大量的行銷成本以刺激消費外，亦不斷的以各種方案，讓此大眾化的行銷方式不僅造成行銷資源的浪費，回應率也不高，更衍生逾繳帳款等信用風險問題。金融之成敗最重視的是信用風險控管能力，能精準掌握客戶風險的變化，利用資料探勘的自我學習能力，可及時反應出變化，減少可能的呆帳損失。

壹、緒論

因經濟的不景氣，促使民眾辦理申請借貸有增加的趨勢；以銀行來說，各種相關的業務資料十分龐雜，若要審核資料是否達到一定門檻標準，在作業上需耗費相當多時間，使銀行不易快速的篩選資料。例如客戶申請借貸時，銀行在前置作業的處理上必須先評估此客戶的信用程度好壞作爲第一優先的篩選，才能進一步的決定是否可放款借貸給予某顧客。但因缺乏適當的分析工具，使銀行在作業上必須仰賴經驗累積的判斷或歷史報表來找出相關資訊，這不僅耗時也不符經濟成本。另外，因銀行呆帳的比例在民國 89 年至 91 年及 92 年至 95 年間有高升的情況，由此可見，呆帳對於銀行是個值得深入思考的問題，若想要減少呆帳發生的風險，第一步必須先嚴格把關過濾信用程度低的客戶，才能幫助降低呆帳不斷高升的機率。

貳、相關文獻探討

一、資料採礦(Data Mining)

資料採礦可解釋爲資料庫之知識發掘(Knowledge Databases，簡稱 KDD)，爲資料倉儲(Data warehouse)應用方式中重要的一環；其主要的目的是用來預測未來的趨勢及找出未知的樣式。資料採礦是利用分類、群集分析、關連性、序列分析、機器自我學習及其它統計方法自龐大的歷史資料中將所隱藏的資訊挖掘出來，它使用了許多統計分析與 Modeling 的方法，到資料中找尋有用的特徵(Patterns)及關連性(Relationships)。以下幾點爲資料採礦的主要功能：

(一)分類(classification)

按照分析對象的屬性分門別類加以定義，建立群組(class)。例如，將信用申請者的風險屬性，區分爲高度、中度及低度風險申請者。

(二)推理(estimation)

根據既有的連續性數值之相關屬性資

料，以獲得某一屬性未知的值。例如，按照信用申請者之教育程度、行爲來推估其信用卡消費量。

(三)預測(prediction)

根據對象屬性之過去觀察值來推估該屬性未來的值。例如，由顧客過去之刷卡消費量預測其未來之刷卡消費的額度。

(四)關聯分組(affinity grouping)

從所有物件決定哪些相關物件應該放在一起。例如，超市中相關之盥洗用品(牙刷、牙膏)可放在同一個貨架上。在客戶行銷系統上，可透過此功能來確認交叉銷售(cross selling)的機會，以設計出吸引人的產品群組。

(五)同質分組(clustering)

將異質母體中區隔爲較具同質性的群組。它類似於行銷術語中的區隔化，但同質分組爲先假定事先未對於區隔加以定義，而是由資料中自然產生區隔的結果。

二、決策樹

決策樹是建立分類模式(classification models)的方法之一，使用此演算法，可針對給定的資料運用歸納的方式產生樹狀結構的模式，以離散或連續的屬性做預測分析。爲了將輸入的資料做分類，決策樹的每一個節點(node)都是一個判斷式，其內部的每一個節點(internal node)會對映到某項屬性的測試，每一個分支代表被測試的屬性當中一個可能的值，而每一個葉節點(leaf node)則會對映到一個布林函數值。

決策樹適用於各種分類的問題，應用的層面相當廣泛，例如：銀行的信用卡授信、建構分類型專家系統、直效行銷回應、顧客流失預測等，都可利用決策樹來產生容易理解的規則，從中歸納出一些規律性，以幫助使用者可利用此預測結果進行後續動作

三、關聯法則

關聯法則此模型是用來探討資料項目間的關係，找出在某一事件或資料中會同時出現的項目。例如：某顧客買了筆記型電腦，則此顧客同時購買隨身碟的機率是80%、而最著名的例子爲美國的超級市場Wal-Mart ，經日月時間的累積銷售資料中發現，每到星期四尿布與啤酒經常會被一起購買，於是將兩樣物品放在一起，使顧客易於同時拿取購買。

關連法則的優點在於試圖找出多條規則，且每一條規則都可以得到一個相對應的結論；而其缺點則爲需花費較多的時間，且所產生出來的規則不一定是適用的。

參、軟體功能簡介

本研究主要以 Attar Software 公司所開發的資料採礦軟體 XpertRuleR Miner 作爲研究。

它在資料探勘的應用主要分爲三大部分：(1)決策樹(2)關聯法則(3)交易關聯。

在銀行的資料上，我們主要是採用決策樹來作爲運用(1)決策樹的技術來產生界定影響判定信用程度好或壞的關鍵屬性其因素與優先次序，藉由決策樹將一連串的問題和規則將資料分類，由相似的型態來推測相同的結果，以階層樹狀的方式呈現分類法則，可用來預測新客戶的信用好壞之判別。這些法則需經過評估篩選，去除低信賴、重複與不合理的法則，保留有合理解釋與推論價值的部分於法則文字檔中。

肆、軟體特色

XpertRule Miner 爲一套圖形化環境的資料採礦軟體，介面爲英文，主要的功能是從歷史資料中發掘隱藏的資訊，其特色有：

(一)、軟體爲圖形化介面操作

(二)、提供高效能挖掘元件的資料採礦系統

(三)、資料採礦的結果可以嵌入 ActiveX 元件，以做爲其他的應用

(四)、資料採礦的結果可輸出爲檔案

伍、研究方法（軟體研究工具與技術）

本研究進行資料分析前，會先進行資料整理的動作。首先將原始資料之遺漏值、重複值與錯誤值予以刪除，以方便日後使用各種軟體做不同的分析，避免因錯誤的資料而影響結果。爲了驗證決策樹的正確率，因此我們根據 1000 筆經過整理的資料，我們隨機挑選分爲訓練組 800 筆與測試組 200 筆兩組資料；以測試組 200 筆來進行訓練組 800 筆所產生的決策樹之驗證。

利用 XpertRule Miner 決策樹功能，將屬性和資料匯入進行分類，並根據探勘的目的找出一個最終的決策變數(outcome)及影響決策變數的相關欄位(attribute)和排除(excluded)沒有相關的欄位，讓決策樹的準確度提高。決策變數爲離散型態，會出現設定最小分支個數、分支最大顯著水準、分割準繩的視窗。本研究將根據系統建議值爲預設值；第一個參數爲最小分支個數：預設爲 24，表示樹中節點的樣本數低於 24 的設定門檻值就會停止往下分支；第二個參數爲分支最大顯著水準，也是終止分支預設爲 0.5%，最後一個參數則爲分割準繩：entropy 值。最後而產生決策樹，若結果顯示葉節點的Good與Bad百分比值過於接近(ex：54%與 46%)，就採用人工操作使葉節點在進行個別分支，讓葉節點比例值更爲顯著。

由於XpertRuleMiner的決策樹(Decision tree)屬於二元樹(Binary Tree)，而 Polyanalyst 則是多元樹。而因爲本研究的資料多爲類別屬性以二元樹分類法較不明確，故想透過 Polyanalyst 所產生的決策樹來比較二者間的差別。Polyanalyst 從一開始的資料匯入、排除不相關的屬性以及決策變數的設定等，都與 XpertRule Miner 屬性的設定相同。透過 Polyanalyst 所產生的決策樹可以手動調整決策的分割方式，利用數值的調配讓錯誤率降低；此外 Polyanalyst 的系統也內建混亂矩陣的模式，類似下表 1-1 與表 1-2 表示方式，可根據顯示將得知資料筆數是否正確或資料無法定義。

透過訓練組 800 筆的資料分別匯入 XpertRule Miner 與 Polyanalyst 中所得到的決策樹，我們將使用人工的方式利用測試組 200 筆的資料來驗證其決策樹的結果是否有可看性。

關聯法則在本研究中，會先將原先的資料屬性匯入模式中，以支持度及信賴度兩個指標來評估所產生的法則是否成立。我們先將支持度與信賴度的篩選條件分別調整爲達 0.6、0.7、0.8、0.9 以上的數值，因設定值過低會產生許多雜亂的規則，且得到的法則會不具意義及代表性。但因設定條件爲 0.7 以上的數值，所產生的規則爲零；若設 0.6 的條件數值，法則結果則過少；因此我們採用系統的預設值 0.5 做爲依據準則產生關聯法則，從中得到了六條關聯法則。我們希望透過其結果的運用，可以用來幫助銀行發展適當的的行銷手法以拉攏回饋顧客；如在此六條法則中，其中有一資料集爲沒有其他分期付款計畫及信用程度爲好的情況下，二者同時發生的支持度機率爲 0.63，此資料集的情況又劃分爲兩種狀況：一、沒有其他分期付款計畫的情形下，信用程度爲好的機率爲 0.78；二、情況與一相反，爲信用程度是好的情形下，沒有其他分期付款計畫的機率爲 0.87。由此我們可以推測此類型的客戶可能爲保守型的民眾，因此銀行可以推薦此客戶群投資一些低風險的方案，如定存、基金等。

預測 實際	Good	Bad
Good	125/147*100% =85%	22/147*100% =15%
Bad	31/53*100% =58%	22/53*100% =42%

表 1-1XpertRule 測試組 200 筆之驗証

預測 實際	Good	Bad
Good	125/147*100% =85%	22/147*100% =15%
Bad	31/53*100% =58%	22/53*100% =42%

表 1-2 Polyanalyst 測試組 200 筆之驗証

陸、系統使用環境

一、硬體

作業系統	Window XP
記憶體	256MB 以上
硬碟空間	80GB 以上
操作方式	鍵盤、滑鼠

二、軟體

探勘軟體	XpertRule Miner、 Polyanalyst
資料處理輔助	Excel、Access、Word
資料庫	UCI 所提供的某德國 銀行客戶歷史資料

柒、研究結論

根據研究結果顯示 XpertRule 與 Polyanalyst 不同在於前者爲二元樹，後者爲多元樹，然透過 XpertRule 與 Polyanalyst 的研究分析探討德國銀行中的客戶資料(訓練組)，得知兩者決策樹都以「帳戶狀況」當成根節點，「信用記錄」、「貸款金額」、「其他分期付款計畫」屬性爲次分類；由以上推知在德國銀行客戶信用中，這些條件是此德國銀行分類客戶信用的重要依據。透過以上的法則，可以使銀行在給予顧客申請貸款時的一個依據，藉由此的分類法則的結果來進行借款策略。

雖然此研究的資料庫來源是藉由德國銀行的歷史資料來進行資料採礦的分析測試；但在採礦結果中，這樣的步驟流程及結果，對於台灣的銀行或者是需要靠報表或者是經驗法則來經營的公司企業是有實質上的幫助；因給予一定的判斷結果或者是預測的法則，能在運作上帶給更快速方便的工作效率。雖然資料採礦的過程及結果，不一定一次即能讓使用者得到滿意的結果，而是需要經過不斷的嘗試及驗證；雖然在研究過程中，需耗費相當大的心力與遭受打擊及挫折，但若得到結果時則會對此研究領域有所幫助，知識就有一定的重要性。

而在關聯法則的運用上，支持度及信賴度這兩個指標通常被用來作爲評估規則是否成立的標準，若支持度和信賴度的門檻設定過高會不容易產生規則，進而遺漏可能的重要關聯規則；但是門檻設定值過低又會產生許多雜亂不可靠的規則。所以支持度和信賴度的設定值還需要靠分析者的經驗才能得到有用的資訊。

捌、使用心得

我們是第一次研究資料探勘這個領域，從一開始的模型的構思到軟體實作的階段之中都曾遭遇挫折及失敗，如使用的資料庫變動、可請教軟體知識的人員不足、探勘項目及屬性的選擇等。以資料庫更動的選擇來說，在大三下的研究分析時，先使用醫院的資料庫來做探討，但由於醫院站在保護病人隱私權的立場下能給予的資料有限，因此才轉向銀行客戶歷史資料來探討挖掘，雖然過程是一波三折，但之間的變動讓我們對軟體的熟悉度更爲增加。

而本研究的工具使用上，耗費相當大且多的時間與心思。從一開始系上安排的教育訓練課程至自我摸索研究過程中，遇到很多問題及很多不明白的地方，如決策樹的設定值、過濾刪除等，然可以詢問軟體操作的對象認知範圍有限，故透過閱讀論文的方式及組員間互相討論摸索嘗試的經歷來進行本研究的操作流程。

根據本研究發現資料探勘的範圍是廣大的，光是演算法的部分就很多及複雜而決策樹又有不同的劃分法。另外，將資料丟進探勘軟體中所得到的結果就不一定是有用的，而是要經過多次的揣測及驗證的結果；經過本研究之後，讓我們了解到原來只有碰觸到資料探勘的皮毛而已，還有很大的空間等著我們去探索。

參考文獻

【1】曾憲雄、蔡秀滿、蘇東興、曾秋蓉、王慶堯著，「資料探勘」，旗標出版股份有限公司，97 年 1 月出版

【2】李秀琴，「論文-應用人工智慧於人類慢性疾病管理」，92 年 6 月

【3】賴信良，「論文-資料挖掘在教育上的應用-以國小學童「體適能測驗」爲例」，91 年 5 月

【4】資料來源：永豐金控研究總處整理，2008 年 12 月

【5】資料庫來源網站，「UCI(German Credit Data)DataSet-http://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)」

【6】皮托科技股份有限公司，「http://www.pitotech.com.tw/」

附件

使用手冊