運用地理資訊精進房貸風險預測

張皇裕
輔仁大學資訊工程系
henrychang001@gmail.com

徐嘉連
輔仁大學資訊工程系
alien@csie.fju.edu.tw

摘要

近幾年隨著房屋交易市場不斷升溫，因促進房屋貸款需求的提高，房貸也成為了各銀行重要的貸款產品，但前幾年因台灣的消費金融市場，在授信控管上過於寬鬆而導致了雙卡效應，使得銀行資產品質變壞，再加上美國次級房貸的影響，各銀行面對蓬勃發展的房貸業務，也開始擔心房貸資產品質的狀況，除嚴格的把關房貸業務的放款品質外，也開始利用資料探勘的技術，建置信用評分卡，期望能預測客戶未來可能產生逾期的狀況，或事先預測財務狀況即將變壞的客戶，即時予以處理進而保障銀行債權，如此將可大幅降低銀行風險，提昇銀行的獲利。

本研究運用房屋貸款擁有不動產擔保品的特性，使用 K-means 演算法分析擔保品的地理位置資訊，將客戶進行分群，再運用 C&RT 演算法針對各群客戶，進行房屋貸款的風險預測分析，有效的提高風險預測的正確率，並提供銀行在進行區域風險規劃時，可參考的一種更細緻的規劃方法。

關鍵詞：房貸、資料探勘、K-means、C&RT、地理資訊

Abstract

In recent years, along with the elevating mortgage loan market which promotes the mortgage loan demand increasing, mortgage loan has become an important business in all loan products among banks. Yet, several years ago in Taiwan's consumer financial market, the loose credit control causing the double card effect—credit card & cash card, also made the bank asset quality to go worse. In addition, influenced by US subprime mortgages, though facing the home loan's prosperous development, banks also worry about the asset quality of mortgage loan. Thus, besides the strict check on the quality of loan service, banks also start using the data mining technology to establish evaluation system, and expect this can forecast the customer who will possibly exceed the time limit to pay the mortgage loan in the future, or whose financial situation will soon go worse. So, banks could take action immediately to safeguard their creditor's rights, and consequently reduce the risk, promote the profit.

This research makes use of the characteristic that mortgage loan has the real estate collaterals, and uses K-means method to analyze the collaterals' geographical position information, and classify the customers. Then, using C&RT method to carry on the mortgage loan risk prediction analysis in each crowd of customers, it will effectively enhance the accuracy of risk prediction, and provide the banks one more careful method in regional risk planning.

Keywords : mortgage loan、data mining、K-means、C&RT、geographical.

1.導論

1.1 研究背景

目前市場上較具競爭力的銀行，可分為以企業金融或消費金融為主力的銀行，企業金融客戶數量少、金額大、單一客戶風險發生時，其損失金額也相對高，而消費金融客戶數量多、金額小，雖單一客戶發生風險損失的金額小，但相對上發生風險的客戶數量及機率也會比企業金融多很多，所以銀行在消費金融產品的風險的控管上，因客戶數量多且資料量大的特性,在風險控管上經常應用資料探勘的技術。

消費金融的產品可分為房貸、汽貸、信貸、現金卡及信用卡等五類。近幾年隨著房屋交易市場不斷升溫，因而也促進房屋貸款需求提高，房貸逐漸成為各銀行消費金融貸款業務的重頭戲，但在前幾年因整個消金市場，在授信控管上過於寬鬆而導致了雙卡效應，使得銀行資產品質變壞，因此各銀行面對蓬勃發展的房貸業務，也開始擔心房貸資產品質的狀況變化，除嚴格把關房貸業務的放款品質及數量外，也開始投入資料探勘的技術，建置信用評分卡，事先預測客戶未來可能產生逾期的機率，或事先預測財務狀況即將變壞的客戶，適時予以處理以保障銀行債權，如此將可大幅降低銀行風險。

1.2 研究動機及目的

有鑑於在房屋貸款的人工審查流程中，擔保品所在的地理位置，就放款相關專業人員而言，是影響放款案件的重要的因素之一，會直接影響到不動產的價值、流通性等，間接也會影響到客戶發生逾期風險時，銀行的債權會受到損失的程度，因此授信專業人員編製了區域風險參考準則，提供授信相關人員審查案件時參考，可見擔保品在地理上的位置，對於房貸案件有一定程度的影響。

本研究希望能在運用房屋貸款擁有不動產擔保品的特性，使用 K-means 演算法分析擔保品的地理位置資訊，將房貸客戶進行分群，再運用 C&RT 演算法針對各群客戶特性的不同，進行房屋貸款的風險預測分析，有效的提高風險預測的正確率，地理資訊分析的結果，期望提供銀行在進行區域風險規劃時，可做為規劃方式的參考,以設計出更細緻的區域風險劃分。

1.3 業務流程

借款人因為有住屋需求、投資房地產而購買房屋，或因有資金需求而將房屋設定抵押給金融機構，以取得貸款融資，再以分期攤還的方式歸還本金及利息，稱之為房屋貸款（簡稱房貸）。

在貸款產品中，房屋貸款屬於金額較大、貸款期間較長、貸款利率較低之貸款產品。如圖 1-1 中的房貸案件申請流程，逐項說明貸款步驟如下：

（一）申請房貸

借款人填寫貸款申請書及檢附借款人身份證明文件、土地及建物所有權狀、收入及財力證明等相關文件，至金融機構辦理房屋貸款的申請作業。

（二）洽談貸款內容

借款人與銀行業務人員接洽後，進行洽談貸款的相關內容及期望,包含貸款利率、額度、貸款期限及還款方式等內容。並交付貸款申請書及各項需檢附之文件，業務人員先進行初步審查，確認借款人符合公司規定的基本授信條件，再將房貸申請案件送件，進行案件審核的作業流程。

（三）不動產鑑價

鑑價人員接到指派的申請案件後，為了確實了解擔保品（申請貸款的房屋）目前的用途

及價值，進行現場的會堪，了解房屋使用狀況、房屋類型、座落區域、周邊生活機能…等因素，並進行不動產估價，依據不動產鑑價後的金額，可提供授信人員在計算可貸款成數及貸款金額時重要的參考依據。房屋所在的區域，會依專家經驗編製的台灣各地之區域風險參考準則，支援鑑價人員去判斷房屋所在區域之風險等級，並關連到貸款成數及利率的核定。

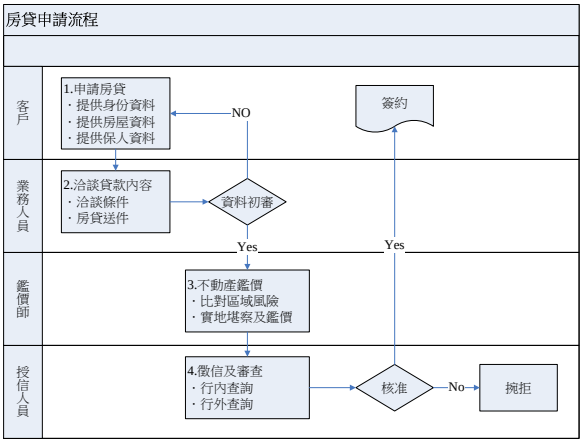

圖 1-1　房貸申請流程圖

（四）徵信及審查

授信人員針對借款人及保人，進行銀行內部的徵信查詢（意指查詢借款人及保人與銀行所有往來的狀況及信用）及銀行外部的聯合徵信查詢（意指查詢聯行徵信中心所記錄之個人及信用資料），依據借款人的基本條件、還款能力、不動產鑑價淨值、不動產所在之區域風險…等因素，決定貸款案件的婉拒或核給額度與利率。

2.研究方法

2.1 房貸相關文獻探討

（一）王仁宏與張雅君（2007），商業銀行房貸客戶違約因素之探討

以某商業銀行 2001 年 9 月至 2005 年 10 月間完成撥款的案件，因房貸客戶資料欄位的不足，增加使用房貸客戶在信用卡系統資料，以補足可供分析的欄位，使用 SPSS 的羅吉斯迴歸分析，整體的預測準確率為 80.39%，有顯著差異的欄位為利率、補貼率、年齡、性別、區域別。

（二）莊瑞珠與陳穆貞(2006)，金融機構住宅房屋貸款信用評分系統之建構研究

研究銀行自西元 1995 年 1 月起至 1996 年 12 月底之 2 年內所有顧客族群，使用 probit 分析、區別分析及羅吉斯迴歸分析三種分析方法，其中羅吉斯迴歸分析的預測力最好，對於正常戶的預測率為 94.9%，對於違約客戶的預測準確率為 63.3%。

（三）鄭歆蕊與吳宗正（2007），二階段預警模型之研究-以台南市房貸為例

以國內某金融機構針對台南市於 2004 年 6 月至 2006 年 4 月間，個人房屋貸款申請案件，使用羅吉斯迴歸分析後整體正確率為 88.6%，逾期戶正確率為 52.40%，有效的欄位變數為申請金額、年齡、性別、學歷、房屋類型、屋齡及鑑價金額等。

（四）李海麟與王瑜琳（2003），銀行消費者房屋貸款授信評量之實證分析

使用國內某金融機構，1996 年 10 月至 1999 年 10 月，使用羅吉斯迴歸分析原始申請件資料預測達 94.67%，研究發現在性別、年齡、年收入、貸款型態、婚姻情形、貸款區分及擔保品所有權者為申請表上影響授信品質的顯著變數。

（五）周俊賢（2003），房貸信用評量分析

以 1998 至 1999 二年期間銀行實際資料，以羅吉斯迴歸分析準確率為 93.2%，研究發現以夫婦職業及所得這二變數最為顯著，且特殊性在於欄位的選取是採用專業人員的多年經實務經驗，且分析資料而使用取樣資料而採全數資料進行分析。

2.2 資料探勘相關文獻探討

（一）張超（2008）分類樹中 C&RT 演算法與判別分析的比較及其醫學應用

探討決策樹中 C&RT 演算法和判別分析(Discriminant Analysis)在資料分析應用中的差異。首先介紹演算法的基本原理，再依各方法間的優缺點進行比較，最後以實例進行研究腮腺良惡性腫瘤在臨床體症及 CT 影像上的差異分析。證明 C&RT 演算法對於類別、數值型變數等不同屬性變數的運用非常靈活，且 C&RT 演算法和判別分析模型各有其優勢。

（二）闕昀澤（2006）新能源意見領袖生活型態與媒介使用行為之研究

此研究採用二種研究方法，一為利用網路調查問卷，再將資料使用因素及群集分析，以區分出不同生活型態的群集，另一為使用資料探勘中的 C&RT 進行決策樹模型建置，描述及分析對於能源宣導最有效的決策運用。結論為有三族群對於能源議題興趣較高，預測正確率為 71.1%

（三）張家鳳（2004）利用資料挖掘技術建構保險業差異化行銷模型

針對客戶的購買行為，採用 Apriori 演算法進行分析建置客戶購買保險產品險種之間的關聯性，得到三個最佳的搭配，醫療險搭配防癌險、醫療險搭配意外險及防癌險搭配意外險，並用 C&RT 及 C5.0 決策樹演算法分析保戶的特質，進行差異化行銷，得到 30 歲以下的保戶最有可能購買儲蓄險及投資險種。

（四）林育臣（2002），群聚技術之研究

依據以往的研究選出三個能表現出群聚特性的因素，分別為鑑別率、凝聚力、密度率，並將此三個因素量化為評估值，用於評估群聚結果的好壞，以 K-means 為例作者設計的系統會分析資料，提供最佳的分群參數提供給使用者參考，並提出選擇重要維度的演算法。

（五）李御璽與顏秀珍等（2003），資料探勘在銀行信貸風險評分模型上之研究

運用資料探勘中的分類技術，將客戶的個人基本資料以及信用貸款的相關欄位，建立一個信貸風險評分模型。並提出一個選擇重要欄位的方法，在資料探勘之前先進行欄位的篩選，結果為未經過事前欄位篩選對於預測正確率為 86.73%。但經過篩選欄位則預測的正確率可提昇至 95.74%。

2.3 研究流程

本研究的流程設計如圖 2-1 所示，對於研究流程中的各項作業說明如下：

（一）資料來源（Data Source）

使用國內某銀行房貸資料進行預測分析，資料內容分別包含客戶資料、帳務資料、及銀行外部資料（聯合徵信中心資料）等。

（二）資料預先處理（Data Preprocessing）

開發資料處理程式，將各資料源的資料，分別進行資料清理（Data Cleaning）、資料整合及轉換（Data Integration and Transformation），將資料整理為分析時可用之資料樣式。

（三）建置擔保品圖層

以地理座標定位工具軟體及 Mapinfo（地理資訊處理軟體）將擔保品的地址資料，進行地理座標定位及製作為地理圖層，提供分析運用。

（四）設計區域風險圖層

依授信時使用之區域風險參考準則，依其文字性的描述，繪製出區域風險的圖層，做為與客戶圖層比對、篩選時運用。

（五）以 Mapinfo 篩選出目標客戶

將客戶圖層與區域風險重疊後，再以 SQL 語法，篩選出此研究所要分析的目標客戶資料範圍。

（六）以 K-means 分析地理位置

使用 K-means 演算法，分析擔保品之地理位置，依所在之地理位置的進行分群。

（七）建立各群客戶之風險預測模型

經分群後在不同地理區域的客戶，依其特性上的不同，分別建置其風險預測之決策樹，以用於進行客戶可能逾期的風險預測。

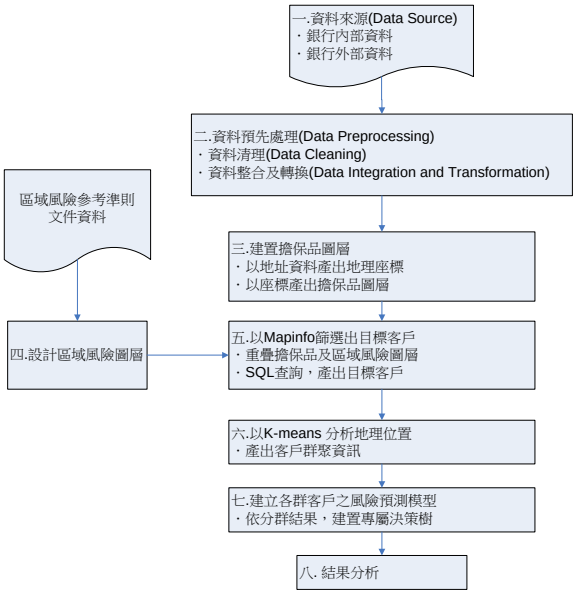

圖 2-1　研究流程圖

（八）結果分析

將無進行地理分群與有進行地理分群的二種資料之風險預測結果，進行比較分析二者間的差異與優劣，並研究地理分群對區域風險規劃的助益。

2.4 空間圖層處理工具介紹

本研究中以 Mapinfo V.7.5 為空間資料之處理工具，可將各式各種主題之圖層，合併於同一座標系上，進行資料分析或其他應用，此工具並提供 SQL 語法，用於處理及查詢空間資料。圖 2-2 為台灣鄉鎮行政區域圖層及河流圖層，重疊二個圖層後，可於系統上查看到重疊後的結果，並可以 SQL 或圖形工具選取或計算各式資訊，例如：計算距離、選取範圍內的物件等。

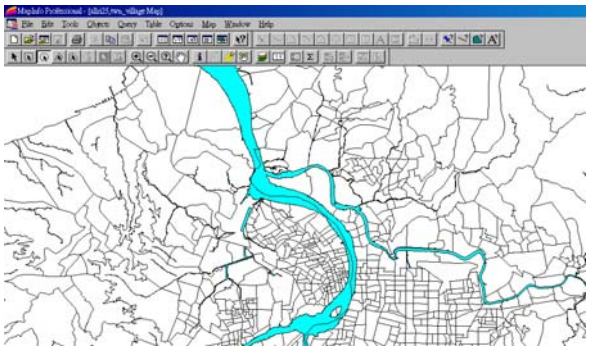

圖 2-2　圖層重疊範例圖

2.5 資料探勘工具介紹

本研究中使用的資料探勘工具為 Clementine V.10，此分析工具為 SPSS 公司所設計，提供 18 種統計及探勘的演算法，並提供各式資料處理與轉換的功能，對於設計整個資料探勘的流程有很大的助益，圖 2-3 為系統畫面範例。

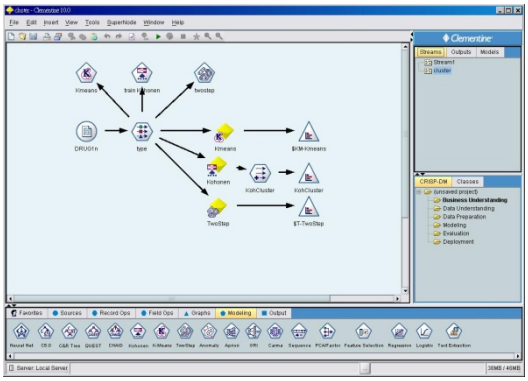

圖 2-3　Clementine 工具畫面

2.6 K-means

K-Means 為叢集分析演算法，將 *n* 個物件分為 *k* 個叢集，使每個叢集內的相似度高，而叢集間的相似度低，方法說明如下[12]：

（一）在高維度的資料中，選擇 K 個點為初始的中心點。

（二）依叢集內每個點的平均值，將每個點歸給取接近的中心點所在的叢集。

（三）重新計算每個叢集的中心，取代原來的

中心點。

（四）重複第 2 點直到不再變化，達到穩定為止。

K-means 演算法的優點為時間複雜度小計算快速，缺點為需自行決定分群的數量，為了達到較好的效果，可選擇不同的初始計算的中心點，再比較各個結果的好壞。

2.7C&RT

在資料探勘中決策樹經常用於資料分析與預測，在樹狀的資料結構中包含節點與樹葉，資料利用個節點的分類條件進行決策，C&RT（Classification and Regression Tree）是由 Breiman 等人在 1984 年所發展出來的一種決策樹演算法，對變數不同屬性的分析上可非常靈活的運用，當目標變數可為連續型數值和類別型的變數，如果目標變數是類別型資料，則會使用分類樹（classification trees），而當目標變數是連續型數值資料時，則可以採用迴歸樹（regression trees）。在決策樹模型建置時，是以單變數拆開進行遞迴運算，故能夠細分顯著影響的變數，產出的決策樹每個節點的分割都是二分法。C&RT 還有二個特殊的部份，一是在於處理遺漏資料，當有變數資料有遺漏時，會透過替代（surrogate）的方式處理，使用其他變數來進行模擬計算真正的分割，另一是對於處理數值性變數資料以迴歸運算，計算出節點上的數值分割點。

C&RT 演算法的運算過程包含三個步驟：

（一）樹狀結構的建立

使用遞迴二元分割的規則進行分割的動作，每一個節點向下分為二個資料子集合，在這個分割動作時是採用 Gini index 做為選擇依據，每一個子集合再尋找下一個變數，不停的將資料分為二個子集合，直到無法分割為止。

（二）樹狀結構的修剪

C&RT 初步計算出的樹狀結構會非常巨大，故需進行修剪的動作，依預測錯誤率及錯誤成本作為修剪判斷的依據基礎，期望以最少節點達到最有效的分類。

（三）選取最佳的樹狀結構

當所有的樹狀結構中的分支都被建置出來後，則可運用測試資料，進行交叉驗證法，將決策樹導出的結果定義成有用的規則。

3.資料處理

3.1 帳務系統資料庫說明

帳務系統的資料庫，主要存放客戶資料、帳務資料及擔保品資料及銀行外部資料等，其資料庫設計如圖 3-1，分別說明如下：

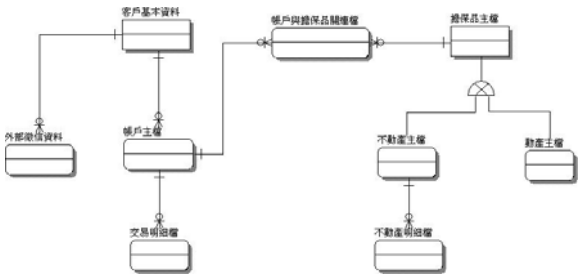

圖 3-1　帳務系統資料庫

（一）客戶基本資料

存放客戶基本資料，同一客戶申請多項產品，僅會存在一筆最新的資料，傳統帳務系統設計之客戶基本資料檔欄位較少、資料品質差、除地址資訊有寄送帳單需求外，其餘欄位資料因很少更新而資料新鮮度較舊。

（二）帳戶主檔

主要存放帳戶的資料，如餘額、利息、利率、期間、狀態及產品等，一個客戶會同時存在多筆帳戶資料。

（三）交易明細檔

客戶主動（如繳款、提款等）及客戶被動（如產生利息等）的交易明細，皆會被完整的保留下來，可提供帳務的查詢及帳單的列印等。

（四）擔保品主檔

因擔保品分為動產及不動產二類，故設計時進行正規化，將動產及不動產共同性的欄位，存放於主檔中，因一個擔保品可提供多個帳戶使用，一個帳戶可提供多擔保品，故帳戶與擔保品之間存在於多對多的關係，故需設計一關連檔記載其關連性。

（五）不動產主檔及明細檔

不動產主檔以一個擔保品的地址為一筆資料，記錄其坪數、鑑價金額、特殊事項記載等，因同一地址會存在多筆的地號建號，故設計明細檔存放資料。

（六）外部徵信檔（聯合徵信中心資料）

外部徵信檔在圖中為示意圖，實務上聯合徵信中心設計了非常多種類的查詢交易，共同點是都以客戶 ID 為查詢條件，故銀行在保留此類交易資料時，會以交易為單位，設計各式的資料表（Table）存放各式查詢的明細資料。

3.2 資料預先處理

開發資料處理程式，從帳務系統的資料庫中，選取 2008 年 4 月底，尚未結清的房貸客戶為範圍，再串連資料庫中各種性質的資料，加工運算為客戶階層的資料（Customer Level），例如將帳戶加工為一筆借款一筆資料，最後產出資料分析時方便使用之資料集。

3.3 地址資料清理處理

儲存於銀行系統上之擔保品地址資料，為一個字串型態的欄位，主要用途為提供相關人員查詢使用，故正確性較差，可能發生下列等品質不一的狀況：

缺乏完整行政區域名稱，例：內湖區內湖路一段１１號，未登打台北市

以簡稱登打，例：北市內湖路一段１１號，北市為簡稱

行政區未同步變更，例：台北市景美區景興街１號，未更新為文山區

錯別字，例：北縣盧洲市中山路１號，”盧”為錯別字

因分析的過程中，需使用地理座標定位程式處理地址資料，故需對地址資料進行資料品質的前處理，以人工進行清理作業，以提高地址資料品質，增加座標定位的準確度。

4.地理資訊處理與分析

4.1 地理座標定位作業

將所有地址資料，產出為文字檔案，欄位包含唯一鍵值（KEY）及地址資料共二個欄位，再使用定位工具，進行整批定位作業，作業完成後會產出各筆地址資料之地理座標（X，Y）及每筆資料定位時，決定座標之精準度，即完成定位作業，定位作業範例如圖 4-1 所示。

4.2 製作擔保品圖層

將定位系統產生之每筆地址的地理座標資料，以 Mapinfo 的轉入功能，匯入座標資料，產生如圖 4-2 之座標明細資料表，再使用其產生地理圖點的功能，指定使用橫麥卡托的台灣二分帶座標系，即可在圖層上產生座標資料的圖點，儲存後即完成擔保品圖層的製作。

利用 Mapinfo 將鄉鎮區域圖層、道路圖層、重要設施及剛產生之擔保品圖層，合併後，產生多重圖層顯示於同一畫面中，如圖 4-3 所示即可看到以台北市仁愛路附近為例的擔保品位置、道路資料及重要設施等資訊於同一圖面上。

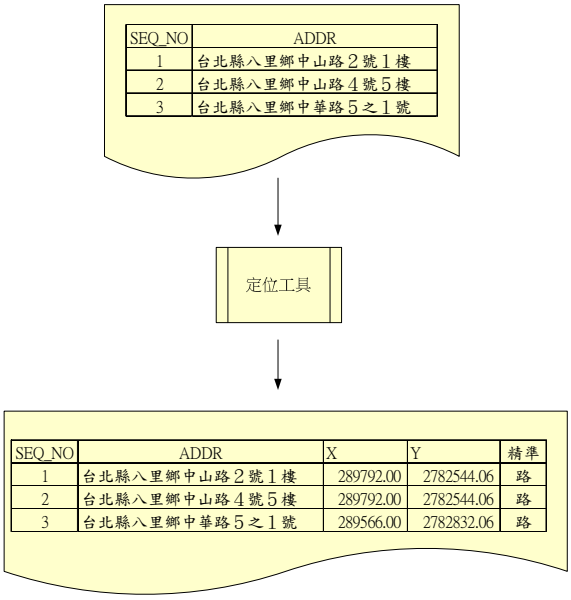

圖 4-1　定位作業範例圖

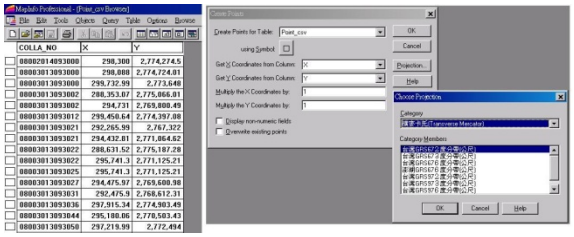

圖 4-2　Mapinfo 座標資料表

圖 4-3　擔保品圖層範例

4.3 繪製區域風險圖層

在區域風險參考準則上，以文字描述區域風險的劃分，提供予鑑價人員及授信人員參考，文字的描述範例如下：

低風險地區：

- 新店五重溪以東、北宜路以北。
- 新莊市全部。

高風險地區：

- 新店水源管制區。
- 泰山鄉全部。

本研究以輔大周圍的五個鄉鎮，分別為新莊、三重、泰山、五股、蘆洲等做為資料分析的範圍，因此將此五鄉鎮範圍內的區域風險，依文字敘述使用 Mapinfo 的繪圖功能，繪製這五個鄉鎮的區域風險區塊，並標上各區域為高風險（H）或低風險（L），再將擔保品圖層重疊後，如圖 4-4 即看到各擔保品所座落的區域風險，與圖 4-3 所呈現的地圖資訊有很大的不同。

接著進行篩選出落點於分析區域內的資料，方法是以 SQL 指令選取地理位置包含於繪製的區域內的資料，SQL 編寫介面如圖 4-5 右方所示，執行 SQL 指令後，可產出目標區域內的資料清單，如圖 4-5 左方資料明細清單。在選取範圍的同時並為每筆資料產出一個新欄位，存放標明每筆資料所在地理區域的風險。

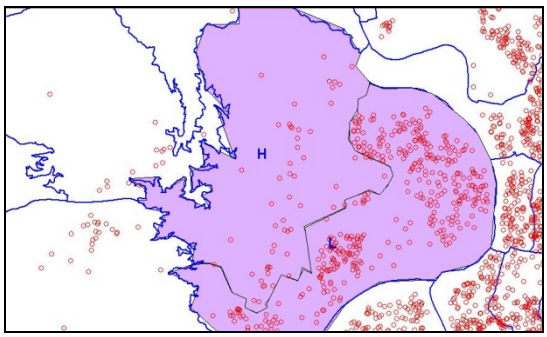

圖 4-4　區域風險圖層

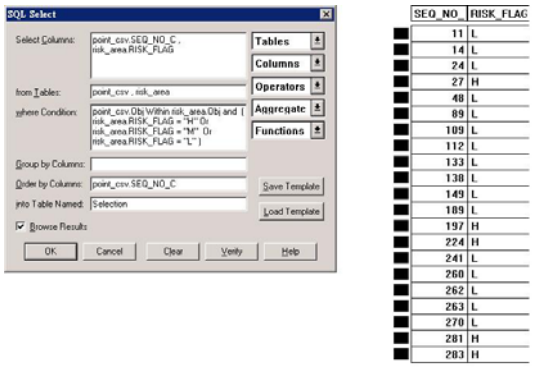

圖 4-5　以 SQL 選取目標客戶

4.4 分析資料的內容分布

經過地理區域的篩選後，已確認所有分析的目標資料範圍，並將資料的重要內容分佈，說明如下列各資料表。

如表 4-1 所示全部分析資料有共 4,851 筆，其中 5.3%的為逾期客戶，代表資料中的好壞比例為 17.8：1，此比例差異較大，因此會增加分析時困難度。

表 4-1 整體分析資料筆數

總筆數	逾期比例（%）
4,851	5.3

如表 4-2 所示，各鄉鎮中以新莊人數最多，五股人數最少；以五股的逾期比例最高，泰山的逾期比例最低。

表 4-2 各鄉鎮資料筆數

鄉鎮	總筆數	逾期筆數	逾期比例(%)
台北縣五股	296	19	6.42
台北縣泰山	449	11	2.45
台北縣蘆洲	1,036	65	6.27
台北縣三重	1,169	53	4.53
台北縣新莊	1,901	109	5.73

如表 4-3 所示，相關授信人員運用區域風險參考準則後，風險高低區域間的逾期比例，並非如一般的認知，風險高的區域其逾期比例較高，低的區域較低，反而呈現相反的狀況，而且逾期比例的差距高達 1.36%，這個差距換算為人數比後，為佔高風險地區的逾期人數的 1/3，代表其區域的劃分，可能有某一定程度上的不精準，後續將再針對此問題進行探討。

表 4-3 各區域風險之資料筆數

區域風險	總筆數	逾期筆數	逾期比例（%）
高	747	31	4.14
低	4104	226	5.50

4.5 以K-means進行地理位置分群

使用 SPSS Clementine V.10，將客戶依地理座標位置進行 K-means 分群，參考此次分析的資料範圍為 5 個鄉鎮，故將分群目標設定為分 5 群。

經 K-Means 分析後的結果，共取得 C1-C5 共 5 群資料，這 5 群的資料內容，統計如表 4-4 示，以其中幾個資料分佈的現象說明如下：

- C4 群人數最多，C3 群人數最少
- C1 群逾期比例最低，C3 群逾期人數比例最高

表 4-4 資料筆數分佈

分群	總筆數	逾期筆數	逾期比例
C1	898	35	3.90
C2	785	46	5.86
C3	287	20	6.97
C4	1573	74	4.70
C5	1308	82	6.27

再將分析完成後的 5 群資料，重新建置新的地理位置圖層，可於地圖上清晰的看到各群資料所在的地理位置分佈，以方便後續分析的進行及資料的判讀，圖形展現如圖 4-6 示，5 群資料分佈的地理位置，以文字說明如下：

（一）C1：為地圖上的綠色圓形點，主要分佈於三重市人口稠密地區。

- 平均房價為 4,453,790 元
- 平均年齡為 43.3 為 5 區中最高
- 單身人數為 24%為 5 區中最高
- 年齡 40 歲以下佔 44%為 5 區中最低

（二）C2：為地圖上的粉紅色菱形點，主要分佈於新莊市及泰山鄉南邊的地區。

- 平均房價為 4,491,098 元為 5 區中最高
- 年收入 100 萬以下的人數為 76%為 5 區中最高

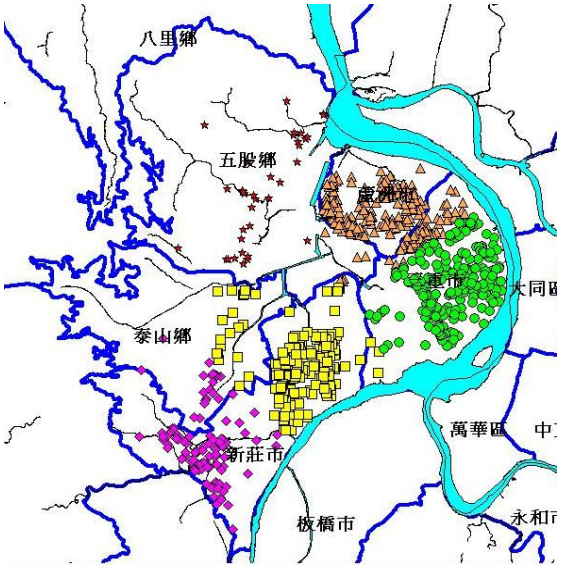

圖 4-6　各群客戶地理位置分佈

（三）C3：為地圖上的紅色星形點，主要分佈於五股鄉中間，該地區因位處於觀音山下，腹地較小且工業區較多，房屋流通較為不易且房價較低。

- 平均房價為 3,714,634 元為 5 區中最低
- 已婚人數為 36%為 5 區中最高
- 男生為 55%，女生為 45%為 5 區中差異最大
- 年齡 41 歲至 40 歲佔 37%為 5 區中最高

（四）C4：為地圖上的黃色方形點，主要分佈於新莊市人口稠密地區、泰山鄉北邊半個區域、及三重市南邊與新莊交界的地區。

- 平均房價為 4,440,697 元
- 已婚人數為 29%為 5 區中最低
- 年收入 100 萬以上為 39%為 5 區中最高
- 女生為 52%為 5 區中唯一女生人數大於男生的區域

（五）C5：為地圖上的淺棕色三角形點，主要分佈於蘆洲市人口稠密地區及與三重、新莊交界地區。

- 平均房價為 4,419,477 元
- 單身人數為 19%為 5 區中最低
- 高中以下教育程度為 56%，為 5 區中最高
- 大學以上教育程度為 7%，為 5 區中最低
- 年齡 40 歲以下佔 44%為 5 區中最低

5.風險預測分析

接續著進行風險預測分析，使用 SPSS Clementine 中的 C&RT 決策樹分析方法，設定預測錯誤的成本後，由探勘工具運算決策樹的節點條件，分析的方式分為二部份進行，首先以整體資料（簡稱 C0 群），建置決策樹，再以 C1-C5 分別依各群的特性，各別建置其風險預測的決策樹，分析比較二者間，地理上的分群是否會增加決策樹的預測準確度。

5.1 預測錯誤的成本計算

決策樹中的預測錯誤的成本，規劃以財務的角度進行成本試算，試算公式設計如下，在此尚無考慮相關的作業成本，僅使用毛利的概念進行計算：

公式 5-1　預測錯誤成本公式

預測錯誤成本 ＝逾期客戶造成的損失 ／正常客戶帶來的利潤＝（平均貸款金額－平均擔保品價值 x 0.7）／ （平均毛利率 x 平均貸款金額 x4）

說明如下：

逾期客戶造成的損失為，(平均貸款金額－平均擔保品價值 x0.7)，使用 0.7 的系數，係因擔保品送法拍後，會由 8 折開始進行一拍，通常需進法拍的案件無法一拍即順利拍出，故在此假設會於二拍中賣出，故賣出金額為 7 折即乘上 0.7。

平均毛利率指（客戶利率－資金成本率），意即概算的實質的獲利率，再乘上貸款金額，即得到客戶一年對銀行的獲利貢獻，經統計房貸客戶平均存在銀行 4 年，即會還清或轉貸到其他銀行，故最後將一年的獲利貢獻乘上 4，即得到一個正常客戶，從開始到結清對銀行的全部貢獻金額。

5.2 建置C0 群之風險預測決策樹

（一）進行取樣

因房貸這個產品，其擔保品的特性，故其逾期比例，相對其他產品為低，整體資料的好壞比例為 17.8：1，故分析時取樣的方式，需保留住逾期客戶的資料特性，方可建置準確性較高的預測模型。

全部資料共 4,851，其中 257 筆為逾期資料，以隨機的方式進行取樣。

訓練組：

- 逾期資料取樣 90%故為 257*0.9=231 筆
- 正常資料取樣逾期資料的 3 倍故為 231*3=693 筆

再將逾期資料重覆二份，以加強其逾期資料的特性，最後訓練組資料中，正常與逾期的資料比為 3：2

測試組：

將剩餘的資料歸入測試組中，統計筆數如表 5-1。

表 5-1　C0 群資料取樣筆數

	逾期筆數	逾期加強後筆數	正常筆數	原始筆數合計
訓練組	231	462	693	924
測試組	26	26	3,901	3,927
合計				4,851

（二）成本計算

以公式 5-1 計算成本如表 5-2，一個客戶逾期造成的損失金額，約相當於 4.2 位正常客戶的獲利，故將預測錯誤的成本設定為 4.2。

表 5-2　C0 群逾期客戶成本

平均 貸款金額	平均 房屋價值	平均逾期 客戶造成損失
3,557,820	4,402,600	476,000

平均 毛利率	平均客戶 貢獻金額	逾期損失 成本比例
0.79%	113,053	4.2

（三）訓練預測模型

C0 群的預測模型處理流程，設計如圖 5-1 所示，並將預測錯誤的成本設定為 4.2 如圖 5-2。

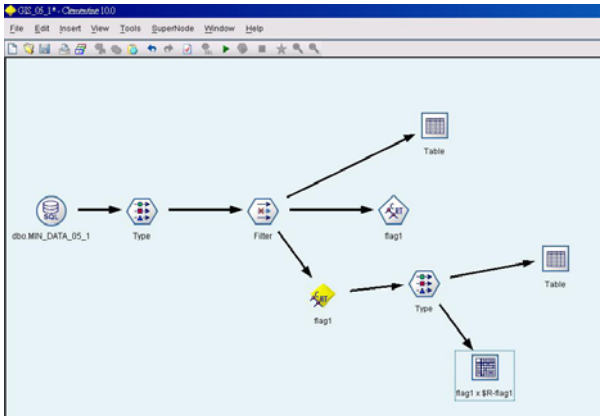

圖 5-1　C0 群 C&RT 分析設計流程

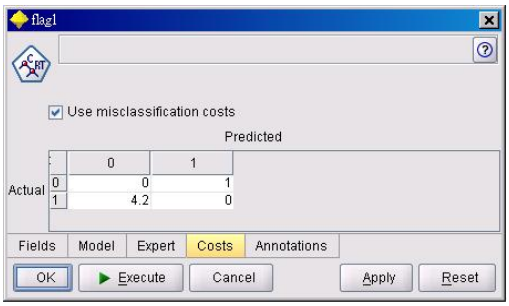

圖 5-2　C&RT 成本設定

決策樹建置後長像如圖 5-3，其中有 7 個路徑的預測結果為逾期，分別就其邏輯條件敘述如下：

- 沒有信用卡，且年薪<69 萬，且貸款

105 萬以上。

- 沒有信用卡，且年薪>69 萬，且貸款 507 萬以上。
- 信用卡 1 張以上，且近 1 年內信用卡曾經逾期 2 個月，且年薪<90 萬，且信用卡額度<7.4 萬，且教育程度為高中以上。
- 信用卡 1 張以上，且近 1 年內信用卡曾經逾期 2 個月，且年薪<90 萬，且信用卡額度>7.4 萬。
- 信用卡 1 張以上，且近 1 年內信用卡曾經逾期 2 個月，且年薪>90 萬，且信用卡張數<20.5 張，且星座為金牛座。
- 信用卡 1 張以上，且近 1 年內信用卡曾經逾期 2 個月，且年薪>90 萬，且信用卡張數>20.5 張。
- 信用卡 1 張以上，且近 1 年內信用卡未逾期，且星座為牡羊座、牡羊座、射手座、金牛座、處女座等。

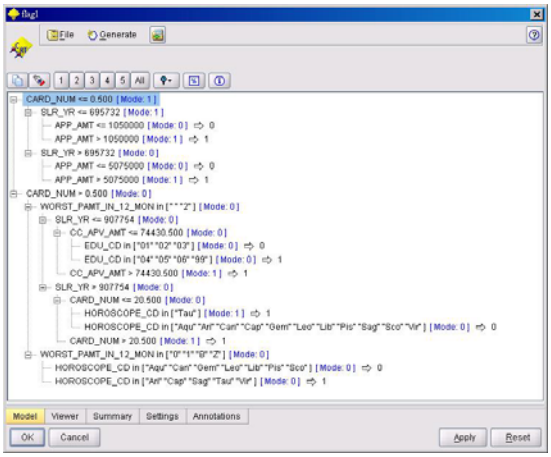

圖 5-3　C0 群決策樹內容

風險預測模型準確率及測試結果如表 5-3，訓練組正確率為 70.8%，測試組為 67.2%，歸納其原因在於逾期資料較少，導致無法有很高的準確度。

（四）測試結果比較

表 5-3　C0 群風險預測準確率

訓練組		預測	
		0	1
實際	0	376	317
	1	20	442
統計			
預測正確筆數		818	
預測錯誤筆數		337	
預測正確率		70.8%	

測試組		預測	
		0	1
實際	0	2619	1282
	1	6	20
統計			
預測正確筆數		2639	
預測錯誤筆數		1288	
預測正確率		67.2%	

6.分析結果之比較說明

6.1 風險預測決策樹之比較

風險預測的決策樹分別建置完成後，C0 群準確率為 70.8%，其餘各群由公式 6-1 以各區筆數加權計算後，準確率為 89.5%，大幅提高了 18.7%如表 6-1 所示，因此可得到一個結論，地理上的分群可將居住較為接近的人，分在同一群中，有助於提高建置風險預測模型之準確度。

公式 6-1　加權準確率計算公式

加權準確率＝（C1 準確率 x C1 筆數 ＋C2 準確率 x C2 筆數 ＋C3 準確率 x C3 筆數 ＋C4 準確率 x C4 筆數 ＋C5 準確率 x C5 筆數）／（C1 筆數＋C2 筆數＋C3 筆數＋C4 筆數＋C5

筆數）

表 6-1　各群加權合計準確率

分群	準確率	筆數	加權計算
C1	97%	898	869.58
C2	88%	785	690.65
C3	97%	287	277.43
C4	86%	1573	1,356.36
C5	88%	1308	1,148.49
合計		4,851	89.5%

6.2 區域風險比較

在研究成果產出的同時，我們再回過頭來檢視區域風險參考準則，與研究中所使用 K-means 分析完的結果，同為地理區域上的劃分，以新莊地區為例，進行二者間的差異比較。

以新莊市為例，如表 6-2 區域風險參考準則中，定義為新莊全區為 1 個低風險區域，目前逾期放款比例為 5.73%，而全部五個鄉鎮的平均的逾期比例為 5.3%，明顯於整體的平均值。

表 6-2　新莊市區域風險參考準則與分群結果比較

新莊市	區域風險參考準則	K-means 分群
區域劃分數量	1	3
區域風險	低	
逾放比例	5.73%	6.32%
		5.45%
		0.00%

經 K-Means 分群後位於新莊市之擔保品分佈如圖 6-1，以紅色圈圈出新莊市被分為 3 個區域，統計這 3 個區域的逾期比例如表 6-3，明顯的看到這 3 個區域的逾期比例有很大的差異。依據 K-means 分群後的區域以逾期比例的差異，重新修正區域風險的劃分，則其中 C2 區逾期比例高達 6.31%，故應將此區列為高風險區域，C4 區接近平均逾期比例，可列為低風險區域，因 C4 區逾期比例已接近平均值，反之若將 C4 區規劃為高風險地區，則可下降的逾期比例少，反而造成好客戶的人數降低，營運金額的下降，C5 區因筆數較少，故不列入比較。整理說明新的區域劃分如下：

- C2 區列為高風險區域。
- C4 區列為低風險區域。

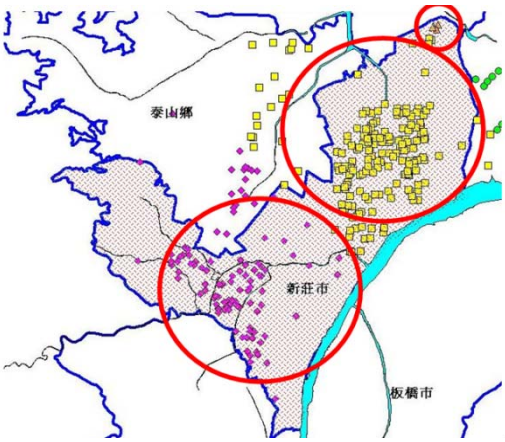

圖 6-1　新莊市擔保品座落分佈

如此一來有助於授信人員在審查時較嚴格的控管 C2 區，以降低新莊市的逾期比例，參考表 4-3　各區域風險之資料筆數，顯示以授信人員審查的嚴緊程度下，高風險地區的逾期比例比低風險地區低 24.7%，如將 C2 區控管至與 C4 區相同的逾期比例，則逾期比例可下降 2.21%。

- 在比例上即為逾期比例將從 6.32%下降為 4.11%。
- 在人數上即為逾期客戶數由 41 人下降為 26 人。
- 在損失金額計算上即為減少 15 人，C2 區的平均逾期損失金額為 423,844 元，故可減少 6,357,659 元的逾期金額損失。

以上說明可證明對於營運上，可減少較高損失金額而產生很大的貢獻。

表 6-3　新莊市各群逾期比例

分群	總筆數	逾期筆數	逾期比例
C2	649	41	6.32%
C4	1,247	68	5.45%
C5	5	0	0.00%

總結因地理區域範圍很大，經專家以人工規劃區域風險，仍會有不夠細緻的地方，因此利用資料探勘的技術協助區域規劃，則會有二個好處：一是在區域的劃分上可較為細緻化，另一是經建立區域劃分模型後，可在輸入地址後自動產生所在區域，達到自動化的作業及避免人工判斷上的誤差。

7.結論

此次研究發現依地理位置先進行分群，再分別進行建置其專屬的風險預測之決策樹，可比未分群直接建置風險預測決策樹，可大幅提高風險預測的準確度 18.7%，因居住在相同的地理區域上的人，生活條件上會較為接近接近，其許多屬性上的相似性，因此建置的預測風險的決策樹會較為準確。

比對新莊市在區域風險參考準則上的區域風險劃分，與 K-means 分析的結果，二者間的差異後，發現可建議將新莊市劃分為二個風險區域，一個為高風險另一為低風險，如此重新再劃分後，則預估可降低逾期比例 2.21%，減少 6,357,659 元的逾期金額損失，因此可提供銀行風險的控管人員，在區域風險參考準則上的區域風險可劃分的更加細緻，且有助於有效的控管逾期風險。

未來還可從三個方面進行更進一步的進行研究：

（一）本研究針對了五個鄉鎮的資料進行分析，未來可擴大分析範圍含蓋大台北地區或全台各鄉鎮，以建置其他區域風險預測模型。

（二）運用資料探勘技術分析地理資訊，可支援區域風險進行更細緻的劃分，未來可針對劃分風險區域的主題，更深入的劃分各個區域的風險，以達到更有效的風險控管。

（三）在本研究中運用了處理地理資訊的相關工具及資料探勘的工具，未來可將各個工具軟體串連，協助房貸進件流程建置一個自動化作業平台，以提高風險控管的準確度。

參考文獻

[1]王仁宏、張雅君（2007）。***商業銀行房貸客戶違約因素之探討***。世新大學管理學院財務金融學系碩士論文，台北市。

[2]李桐豪、呂美惠（2000）。***金融機構房貸客戶授信評量模式分析-Logistic 迴歸之應用***。台灣金融財務季刊，第 1 卷第 1 期，2000，頁 1-20。

[3]李海麟、王瑜琳（2003）。***銀行消費者房屋貸款授信評量之實證分析***。中正大學國際經濟研究所碩士論文。

[4]李壽田（2004）。***台灣金融業發行現金卡經營策略之研究***。中華大學科技管理研究所碩士論文。

[5]李御璽、顏秀珍（2004）。***資料探勘在銀行信貸風險評分模型上之研究***。2004 年第十屆資訊管理暨實務研討會。

[6]林育臣（2002）。***群聚技術之研究***。朝陽科技大學資訊管理研究所碩士論文。

[7]莊瑞珠、陳穆貞（2006）。***金融機構住宅房屋貸款信用評分系統之建構研究***。住宅學報，第十五卷第三期，頁 65-90。

[8]張超（2008）。***分類樹中C&RT演算法與判別分析的比較及其醫學應用***。數理醫藥學雜誌，2008 年 5 月 21 卷 2 期。

[9]張家鳳（2004）。***利用資料挖掘技術建構保險業差異化行銷模型***。世新大學資訊管理系碩士論文。

[10]鄭歆蕊、吳宗正（2007）。***二階段預警模型之研究-以台南市房貸為例***。國立成功大學統計學系碩博士班碩士論文。

[11]闕昀澤（2006）。***新能源意見領袖生活型態與媒介使用行為之研究***。私立世新大學廣播電視電影研究所碩士論文。

[12]Jiawei Han, Micheline Kamber（2003）。***資料探勘－概念與技術 Data Mining,Concepts and Techniques***（曾龍譯）。2003 年 9 月初版第 349 頁。台北：維科圖書有限公司