私立東海大學資訊工程與科學研究所

碩士論文

指導教授：許玟斌 博士

以分層抽樣之規則歸納法探勘信用卡族群共同特性

Mining the Commonality of Credit Card Holders Based on Stratified Sampling and Rule Induction

研究生：蔡明富

中 華 民 國 九 十 四 年 五 月 二 十 七 日

摘　要

塑膠貨幣的興起，改變了人們消費與理財的觀念，根據銀行局統計，民國87年至93年的信用卡循環信用餘額平均每年的成長率為20.85%，現金卡放款餘額民國93年5月至94年2月平均每月的成長率為3.43%。這些數據究竟隱藏了什麼樣的資訊？它要告訴我們什麼樣的訊息？循環利息是信用卡發卡行主要收入來源之一，但相對的其風險也是最高，本研究利用資料探勘的技術，尋找產生這個族群的共同規則，提供發卡銀行卡務推廣參考，例如客戶償債能力、消費能力與對發卡行獲利能力之預測，以及信用額度准駁的依據。

自資訊科技發展以來，不但資訊的產生速度以倍速增長，資訊的累積速度更隨著儲存科技的進步以人類未知的度量方式爆炸式的成長，面對如此龐大的資料，如何從其中萃取出可用的知識及還原資料本身所要表達的涵義，儼然成了人類近代科學中另一門熱門且渴望解決的話題。

在眾多的分類（Classification）演算法中，規則歸納法(Rule Induction)是找出同質資料之產生規則最常用的方法，但隨著資料屬性的增多以及訓練資料量的龐大，使得整個尋找規則的學習(Learning)時間倍增，本研究將運用分層抽樣的技術，隨機抽取樣本再尋找規則，以進行信用卡族群共同特性之分析。

關鍵字：知識探索，資料探勘，分層抽樣，規則歸納，循環利息。

Abstract

In the advent of plastic cash, eventually we changed the spending and saving habits. According to the statistics of the Department of Finance announced in October 2003, the cyclic credit amount has increased 32.33 percent compared with one year ago. The cumulated interests and annual revenues increased 31.4 percent compared with same period of last year, as well. What is the implication concealed on this report? What should credit card issuing institutions respond to this information? The cyclic credit from the cardholders is the main income of issuing bank; consequently the relative risk it carried is very high.

In our research, we are aimed to figure out the commonality of cardholders who apply cyclic credit, based on customer's ability in paying back the debt, expense pattern, credit allowance and profitability of issuing bank. With the popularity of information technology, we are surrounded by data generated in an unpredictable speed. Facing this gigantic dataset, to extract useful knowledge and its implication becomes a hot research topic waiting us to resolve.

Among various classification algorithms, rule induction is the most applied technique in searching the rules out from homogeneous dataset. But with more attributes of entities and large amount of training dataset studied, the learning time in inducing the rules also doubled. So that we tried to stratify the dataset first, followed by inducing the rules, hopefully it would save processing time significantly.

Keywords: knowledge discovery, data mining, stratified sampling, rule induction, cyclic credit

誌　謝

又是鳳凰花開梅雨紛紛的時節，鬱鬱的午後瀰漫著令人傷感的氣息，漫步在文理大道上，回首的不只是綿延的綠蔭，更有那兩年來提攜我的教授身影，以及陪我一路走來的夥伴笑容，沒有你們，我很難獨自走過這段研究生的生涯。

首先感謝 許玟彬教授的傾囊相授，對我這個遲鈍的學生從無怨言，除了灌輸我們正確的研究方法與知識，更教導我們豁達的人生觀與嚴謹的處世態度，讓我在人生的旅途上有了新的體認和看法。同時要感謝在百忙之中撥冗的口試委員鄭國揚教授、陳文雄教授、唐順明教授及姜文忠教授，有您鉅細靡遺的指教，讓學生獲益匪淺，讓論文的內容更臻完備，立論更趨周延，在此，謹致上學生最高的敬意與謝意。

其次要感謝同窗好友孝先、青雲、呈祥、榮信、姿秀，兩年來你們的相互提攜與幫忙，使的艱辛的研究生涯變的平順與充滿歡笑，此外也要感謝學長津華、永明與學姐怡靜、海燕，以及幾位學弟們，感謝這段期間你們所給予的協助。

最後要感謝的是在背後默默支持我的母親與妻子還有松俊、鎮羽兩位頑皮的小可愛，謝謝你們無怨無悔的支持，讓我能夠無後顧之憂的完成學業，尤其是妻子在這段期間獨力的肩負照顧小孩的責任，讓原本文靜柔弱的妳受盡了苦頭。

研究生的生活是辛苦的，但卻是值得的，經過這樣的歷練讓我的視野更加的寬廣，讓我的思路更加有組織更嚴密，這一切都要歸功於曾經幫助過我的人，感謝你們大家，謝謝。

蔡明富 謹誌於

私立東海大學資訊工程與科學研究所

中華民國九十四年六月

目　次

圖目錄

表目錄

第 1 章.　緒論

近十年來，國內金融界歷經政府為順應金融自由化而開放民營銀行之設立，為增加台灣之國際競爭力而加入世界貿易組織（World Trade Organization, WTO），以及 2000 年通過「金融機構合併法」與 2001 年通過「金融控股公司法」等之重大變革。金融環境之競爭愈趨激烈，除了縱橫之整合外，更因企業之西進壓縮企業金融之獲利空間[03]，而掀起一場消費金融大餅之爭奪戰，各式的信用卡、現金卡商品不斷推陳出新，令人目不暇給。在如此劇烈的爭奪下，客戶品質的良窳、促銷對象範圍的精確，便成了影響獲利及風險之重要因素。

現任 CA 公司執行副總裁的馬克．貝瑞尼契(Mark J. Barrenechea, 2001)在 Oracle 公司擔任高級副總裁時出書「e-Business or Out-of-Business」，強調「企業要不電子化否則就被淘汰」[14]，此句話道出企業在這資訊科技進步一日千里的洪流中，不得不接受的事實，而金融業更是個高度資訊化的產業，甚至可以大膽的說「金融業沒有電子化的早就不存在」。如何應用適當的探勘技術，從電子化後累積的龐大資料中找出正確的資訊，是企業在做正確決策與行動之前最重要的工作。

1.1.　研究動機

從銀行局公佈的金融機構損益概況表中可以得知（如表 1）[06]，國內銀行業的主要收入來源為「利息收入」，而利息收入主要是資金需求者支付銀行提供資金時所產生的利息而來。如前所述，銀行要在獲得此項收入前必須先提供一筆資金給予資金需求者，也因此資金需求者的還款品質成了此項業務風險的所在。另外，由於貨幣市場的逐漸開放，大型企業要從初級市場上籌集資金已相當容

易，如發行普通公司債、可轉換公司債（Euro-Convertible Bond, ECB）等，在僧多粥少的情況下，銀行逐漸重視以一般消費大眾為對象的消費金融。

表 1　金融機構損益概況表

單位：億元

年度	總營業收入	利息收入	利息收入占總營業收入百分比
83 年	8,562.6	7,217.1	84.29%
84 年	9,651.1	8,302.5	86.03%
85 年	10,322.1	8,532.3	82.66%
86 年	11,372.1	9,374.1	82.43%
87 年	12,994.9	10,961.4	84.35%
88 年	12,985.9	11,043.7	85.04%
89 年	13,297.7	11,644.1	87.56%
90 年	12,795.8	10,939.2	85.49%
91 年	9,896.8	8,414.7	85.02%
92 年	8,387.7	6,520.2	77.74%
93 年	8,972.1	6,570.2	73.23%

資料來源：財政行政院金融監督管理委員會　銀行局

銀行局統計國內信用卡業務動用循環信用逐年的上升（如表 2、圖 1）[07]、直到民國 93 年年增率不升反降，其原因是消費金融增加現金卡業務分食信用卡預借現金業務（如表 3、圖 2）[08]，此外我們可以從聯合信用卡中心（NCCC）所公佈的年報[01]中看出：至 93 年底經由 NCCC 所處理的信用卡簽帳金額為 5,652 億元，較 92 年成長 726 億元（如圖 3），其中經由自動櫃員機（Automated Teller Machine, ATM）做信用卡預借現金的交易金額將近 100 億，較前一年度增加將近 20 億（如圖 4）。從這些數字我們可以看出國人消費習慣與儲蓄觀念的改變，這些業務雖然給銀行帶來高的利潤，但相對的也帶來了高的風險，因此客戶品質的良窳以及促銷方案推出市場的速度則是消費金融獲利的關鍵所在。

由於金融業是個高度資訊化的產業，各金融機構莫不累積了大量的客戶個人資料以及交易資訊，如何從這些龐大的資料中，在最短的時間、最快速的方法，

找出特定族群的特徵，作為行銷策略及對象、額度訂定等之參考，以期比對手更早在市場上推出新產品佔領市場並吸收到品質較佳的客戶，已是各金融機構最迫切之需求。而在企業界與學術界對於如何區隔客戶品質的優劣，常利用不同的市場區隔變數，對客戶的消費態度或行為進行研究。當代行銷學大師，西北大學凱洛格管理學院（Kellogg School of Management）國際行銷學名譽教授，菲利普．科特勒（Philip Kotler, 1998）指出就消費者市場而言，主要的區隔變數有地理變數(區域、國家、城市大小等)、人口統計變數（Demographic Variables）(性別、年齡、婚姻、所得、教育等)、心理變數(社會階層、生活型態、人格等)及行為變數(購買動機、購買目的、使用率、忠誠度等) [02、11]，這些變數值中以「人口統計變數」為銀行在面對客戶服務時，在第一時間就能獲得的資料。

是以本研究的動機是嘗試在現行的規則歸納技術中，以人口統計變數為資料屬性，探討縮短特徵尋找時間的可行性，以加快尋找的速度，找出合宜的客戶屬性，提供未來在這一方面發展決策時的一項參考依據。

表 2　信用卡業務統計表

單位：百萬元

年度	簽帳金額	預借現金	循環信用餘額	循環信用年增率
87 年	491,097	39,638	124,908	0
88 年	597,786	51,389	152,768	22.30%
89 年	719,770	79,768	205,656	34.62%
90 年	771,862	103,779	259,875	26.36%
91 年	873,595	132,488	316,328	21.72%
92 年	998,885	178,398	399,847	26.40%
93 年	1,254,451	205,781	457,932	14.53%

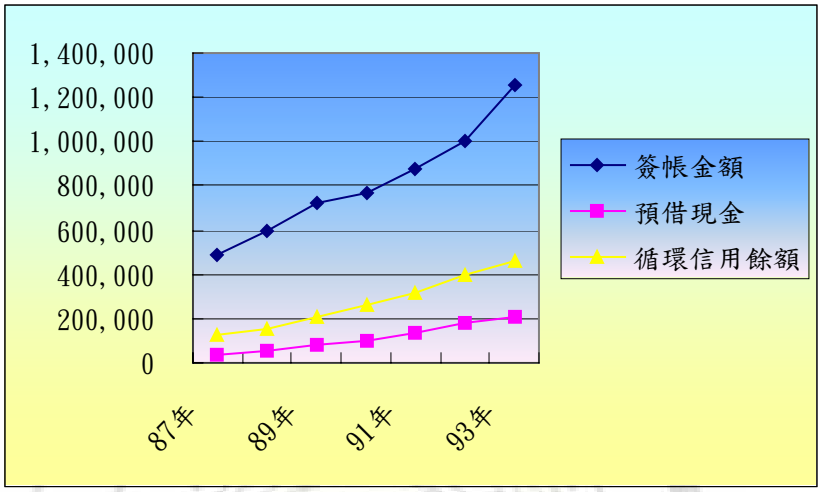

圖 1　信用卡業務統計圖

資料年月	放款餘額(百萬元)	月增率	逾放比率
93 年 5 月	180,990	0.00%	2.843%
93 年 6 月	193,456	6.89%	1.360%
93 年 7 月	199,745	3.25%	1.418%
93 年 8 月	207,603	3.93%	1.389%
93 年 9 月	214,370	3.26%	0.869%
93 年 10 月	222,114	3.61%	0.837%
93 年 11 月	233,447	5.10%	0.811%
93 年 12 月	241,504	3.45%	0.626%
94 年 1 月	249,035	3.12%	0.798%
94 年 2 月	253,293	1.71%	0.835%

表 3　現金卡業務統計表

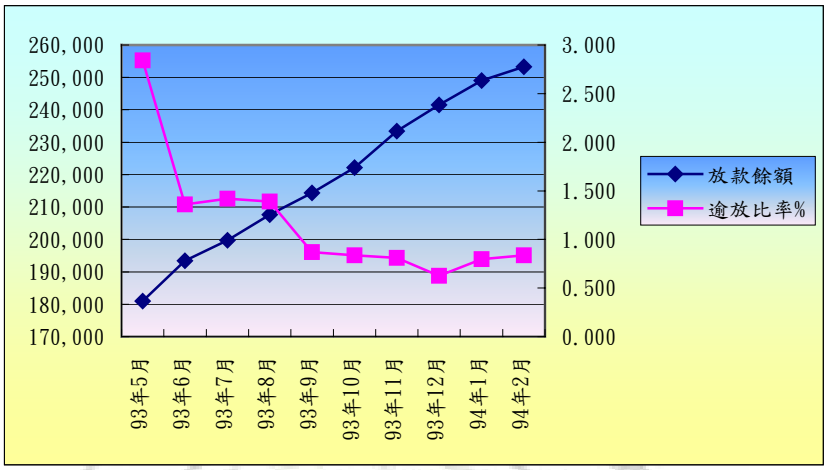

圖 2　現金卡業務統計圖

資料來源：行政院金融監督管理委員會　銀行局

單位：新台幣億元

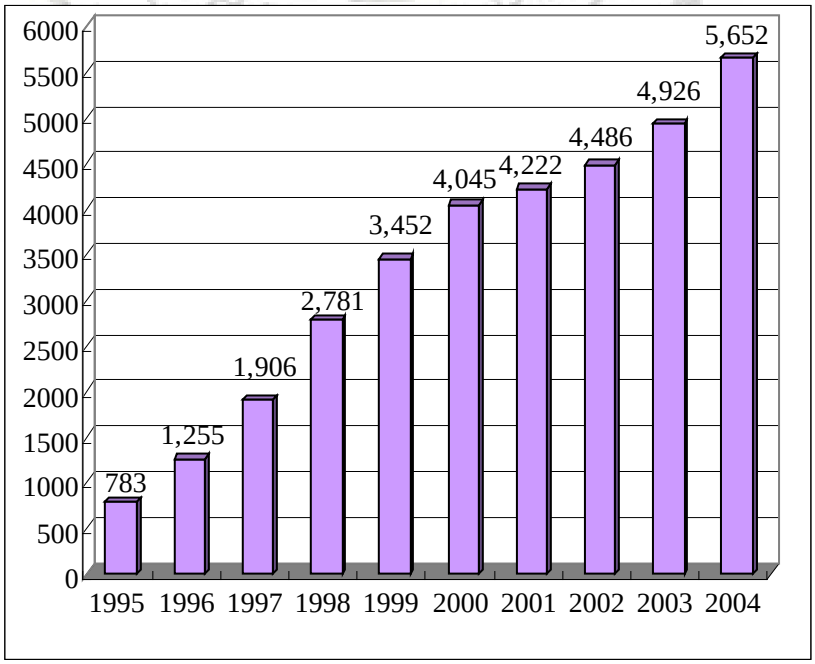

圖 3　聯合信用卡處理中心國內清算簽帳金額統計圖

單位：新台幣百萬元

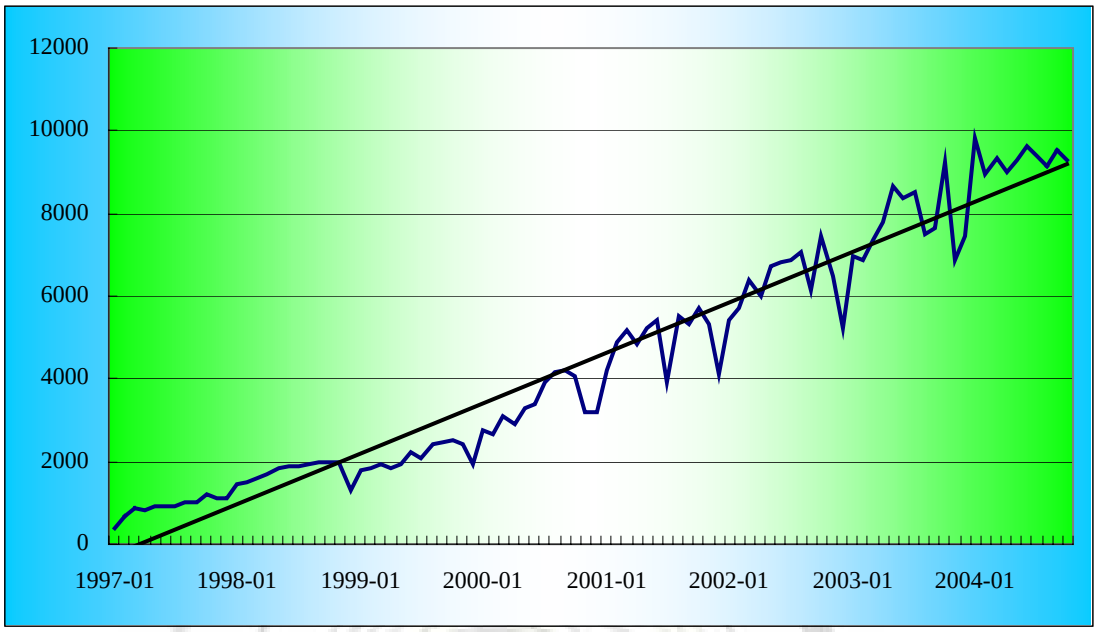

圖 4　聯合信用卡處理中心信用卡 ATM 預借現金金額成長趨勢圖

資料來源：聯合信用卡處理中心

1.2.　研究目的

目前資料探勘的技術繁多，其中以規則歸納法最適合應用在對已知分類的結果，進行該類別族群特徵之歸納，找出該類族群共同特徵之規則。然而因為企業在電子化後，所累積的資料量相當龐大，往往因為訓練資料之空間過大，使得規則尋找時間相對的加長，反而使尋找出的結果失去時效性，在這分秒必爭的極度競賽裡，越快搶先在市場上提出約略合適的產品及服務比落後在市場上推出精確的產品及服務來的重要。

抽樣在統計的領域中，早已是耳熟能詳的方法，在遇到母體過大，無法取得完整的母體時，抽樣技術便是最常應用的解決方法，而在母體中的每一個樣本被選中的機率相等的條件下，具有簡單且快速特性的「簡單隨機抽樣」是最常被

使用的，但在進行抽樣時，由於「簡單隨機抽樣」並未對抽樣方式做任何的限制，因而比較容易產生扭曲的結果，若扭曲的情形嚴重時，將造成樣本資料集和原始資料之間的關聯程度較不明顯而不具代表性。

是以本研究將探討使用隨機分層抽取樣本的方法，藉以縮小訓練資料空間，縮短規則尋找的時間，並與原來的所有訓練資料集以未進行隨機分層抽取樣本所做的結果做比較，探討在合理可容忍的分類正確度內，隨機分層抽取樣本的規則歸納法之可行性。

再則，為求這些規則的簡化與正確，我們通常會在前置處理的過程中，將具有連續性數值的資料屬性離散化，以減少規則尋找的時間與降低規則的複雜度，例如年齡屬性我們可能會分為每十歲一個群族，然而這樣的主觀分法，往往忽略了資料集原有的分佈特質，進而影響規則尋找的結果。

1.3. 論文章節概要

本論文之架構與各章節概述如下：

第一章 緒論

介紹本研究背景、動機與目的。

第二章 相關理論探討

探討與彙整國內外學者對於知識探索、資料探勘、機器學習、規則歸納法、抽樣理論等分析技術，並作概要的簡述。

第三章 研究方法

(1).說明本研究的構想、步驟與方法。

(2).詳述本研究所應用到的理論與方法。

第四章 模擬實驗

(1). 對於所要解決的問題作進一步的定義與描述，以及設計實驗的環境、變數與理論應用的做法。

(2). 以銀行信用卡中心轉換後的資料作為本研究的資料來源，依據資料各欄位屬性定義級距，對於非離散的欄位則以模糊分群法將其離散化，並將資料集依客戶貢獻狀態分成 Good、Bad、Lethargy 三類（層）。

(3). 說明應用分層抽樣技術在此三層分別抽出樣本的結果。

(4). 說明用資料探勘中的 AQ 分類演算法進行規則尋找結果。

(5). 對於實驗結果進一步分析與解釋。

第五章 結論與建議

本研究結論與未來研究方向。

第 2 章. 相關理論探討

2.1. 知識探索

人類的商業活動，隨著資訊化的深入與普及，所有與商業活動有關的交易資訊，包括參與交易者本身的資訊與交易行為產生的資訊等，都被留在日益發展的資料庫系統裡，面對這些數量龐大的資料，該如何從其中獲的有用的資訊，逐漸受到業界與學界的重視，在 1992 年由 Frawley（1992） 等首先定義「知識探索」(Knowledge Discovery, KD)的內涵，是指從資料之中嚴謹的粹取出事先未知的、潛在有用的知識[36]，並定義 KD 流程的基礎框架(Framework)（如圖 5），要完成知識探索的過程主要包含五個要素：

(1).存放資料本身的資料庫。

(2).被探索的領域相關的知識。

(3).處理過程參與的使用者。

(4).探索知識所使用的方法。

(5).對所探索出的知識的表達與應用。

1996 年 Fayyad 等更進一步的定義資料庫的知識探索 (Knowledge Discovery in Databases, KDD)一詞：

KDD 是一個指出資料中令人信服的、潛在有用的一個非細瑣流程，其最終目標是瞭解資料的樣式[26]。他將 KDD 的發展與進行分為五個流程（如圖 6）：

(1). 資料的選擇 (Data Selection)。

(2). 預先的處理 (Preprocessing)。

(3). 資料的轉換 (Data Transformations)。

(4). 資料的探勘 (Data Mining)。

(5). 結果的詮釋與評估 (Interpretation/ Evaluation)。

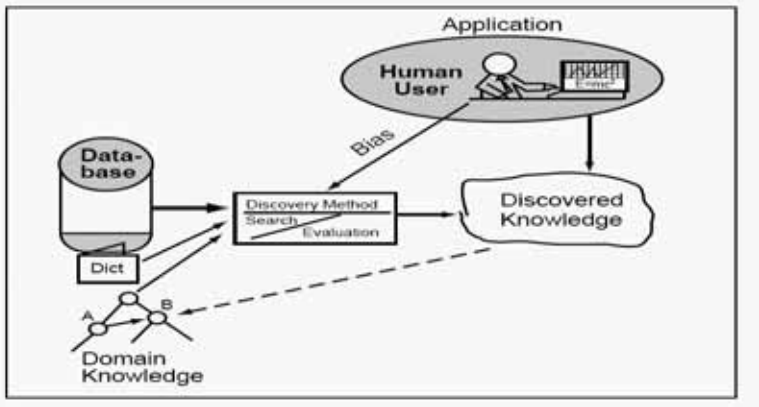

圖 5　在資料庫中進行知識探索的架構圖

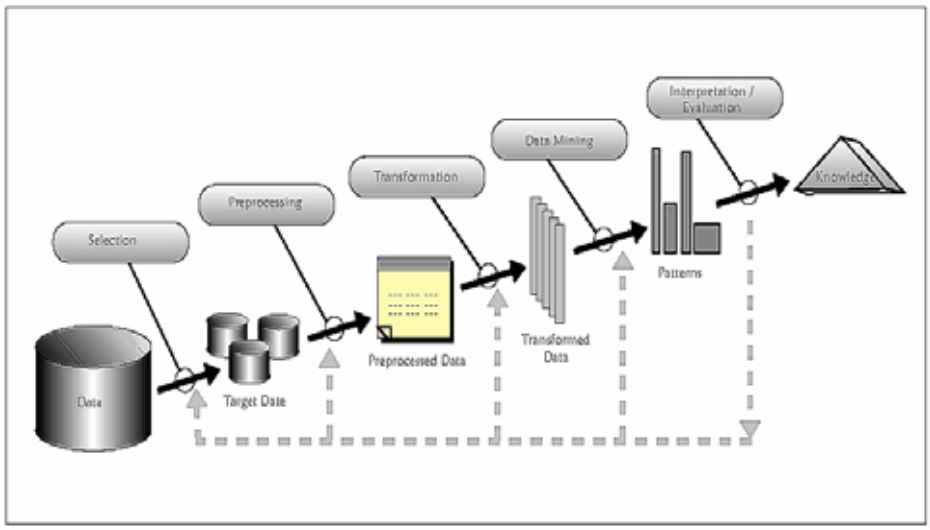

圖 6　KDD 的處理步驟示意圖

Brachman and Anand (1996) [21]更詳細的定義完成上述流程可以分為下列

九個步驟：

(1).加強對目的領域的應用與知識的了解，從使用者的角度清楚的定義進行知識探索的目標。

(2).針對希望探索的資料建立標的資料庫。

(3).對資料作前置處理，包括雜訊的去除與解釋、收集模組化所必須的資訊、掌握漏失資料的欄位以及定義與時間、順序有關的資訊。

(4).資料的歸納與規劃，包括尋找對完成目標有用的資料屬性欄位、應用維度精簡(Dimensionality Reduction) 或轉換的方法(Transformation Methods)簡化資料。

(5).依據第一步驟所定義的目標，選擇適合的資料探勘(Data Mining)方法，如摘要法(Summarization)、分類 (Classification)、分群 (Clustering)、迴歸分析法(Regression Analysis)等。

(6).進行探索的分析(Analysis)、模組化以及假設(Hypothesis)的選擇，包括資料探勘演算法以及尋找資料樣式(Data Patterns)方法的選擇，決定使用的模組化及使用的參數值。

(7).執行探勘尋找所希望的資料樣式，例如分類的規則(Rules)或決策樹(Decision Trees)、分群後的群組。

(8).根據探勘的結果，解譯資料的樣式與所包含的意義。

(9).完成報告與現行的知識做比較，進一步應用所得的知識，改善現行的作業，並擴及其他的系統。

2.2. 資料探勘

「資料探勘」一詞是泛指從巨大的資料庫中粹取，綜合出未知資訊為主軸的複雜活動之通俗的講法，它是資料庫知識探索所有處理程序中的一個步驟 ，另一方面也是指有關於為了現實生活問題所存在的大量資料所作的各種領域的研究或發展的演算法及開發的軟體環境[27、34]，與 KDD 相同的 DM 的處理通常亦可以分為下列五個步驟（如圖 7）：

(1). 資料選擇

由資料探勘處理的選擇目標和工具所組合的，辨識所要採掘的資料，然後選擇適合的輸入項屬性和輸出項的資訊來呈現給交付之工作。

(2). 資料轉換

包括下列這些操作：

a. 以想要的方法來組織資料。

b. 轉換資料的形式(如將符號轉為數字)。

c. 定義新的屬性 縮減資料的幅員。

d. 消除雜訊(去除不必要的部份)與主題無關的部份。

e. 標準化。

f. 如果適當的話，決定如何處理遺失的資料的策略。

(3). 資料的探勘

就步驟的本身而論，接下來是對這些轉換後的資料進行採掘，使用一種或多種的技術來粹取感興趣的樣式。

(4). 結果詮釋與驗證（Result Interpretation and Validation）。

對所探勘出的結果進一步的解釋建立模型，並應用已知的評估方法及未使用的資料庫中資料來測試它的正確度。

(5). 組織所探索到的知識（Incorporation of the Discovered Knowledge）。

將結果呈現給決策者，做選擇或決定與目前的認知所潛在的衝突，將被粹取的知識應用於新探勘到的模型。

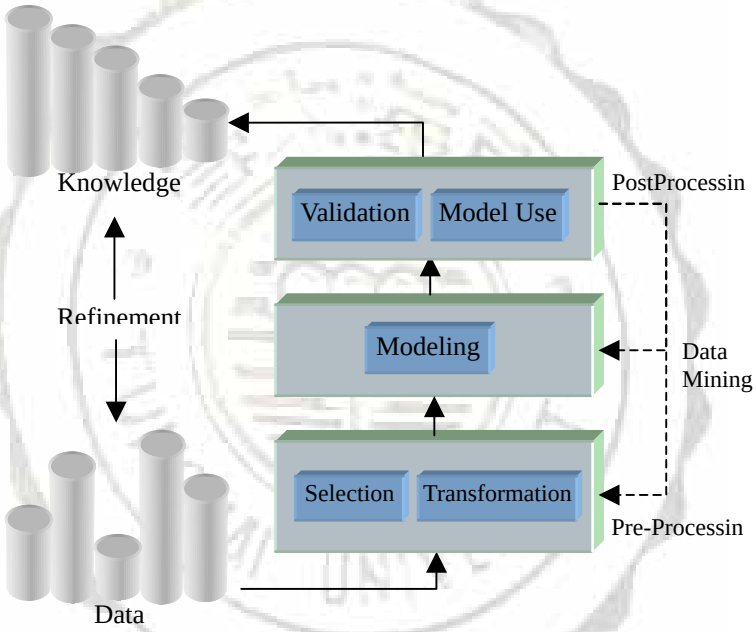

圖 7　資料探勘的處理步驟

在進行資料探勘時，依據所要探勘的目的及資料的性質，通常會進行下列幾種工作[23、34]：

(1). 摘要（Summarization）

主要是對給予的資料產生結構上或特質的概略性描述。它能採取多個形式： 數字(如簡單的統計描述語：平均值、標準差等)，圖表形式(如直方圖，散佈圖（Scatter Plots）)，或者是 ”if-then” 形式的規則。 它可能在整個資料庫裡或者在選擇的子集裡關於所推測的對象提供描述。

(2). 分群（Clustering）

主要是對給予未知歸類特性的資料中找出彼此特性相似的群組，其目的是要將組與組之間的差異找出來，同時也要將一個組之中的成員的相似性找出來。

(3). 分類（Classification）

是根據一些變數的數值做計算，再依照結果作分類。（計算的結果最後會被分類為幾個少數的離散數值，例如將一組資料分為 “可能會回應” 或是 “可能不會回應” 兩類）。分類會用一些已經分類的資料來研究它們的特徵，然後再根據這些特徵對其他未經分類或是新的資料做預測。這些我們用來尋找特徵的已分類資料可能是來自我們的現有的歷史性資料，或是將一個完整資料庫做部份取樣，再經由實際的運作來測試；譬如利用一個大的郵寄對象資料庫的部份取樣來建立一個 分類模型（Classification Model），以後再利用這個模型來對資料庫的其他資料或是新的資料作預測。

(4). 迴歸分析（Regression）。

迴歸分析（Regression）屬於建造一個些許透通的模型的監督式（Supervised）學習問題，其主要目的是做預測，目標是使用一系列的現有數值發展一種能以一個或多個預測變數的數值做為應變數，以預測一個連續數值的可能值的方法，可透過古典或更先進統計方法和以經常在分類任務過程中使用的符號式（Symbolic）方法來進行。

(5). 變動及偏移偵測(Change and Deviation Detection)。

這項工作主要是在發現資料的實際內容與被預期的內容（先前所預估的）或標準值間是否有顯著的變動、誤差或偏移，這些變動可以包含時

間上的偏差或群組間的差異。

(6). 依賴度（從屬性）模型 (Dependency Modeling)。

依賴度模型的問題在於發現一個模型以描述屬性間顯著的的依賴或從屬關係，這些依賴度通常被以“if antecedent is true then consequent is true”的“if-then”規則形式表示。

(7). 時序問題 (Temporal Problems)。

Time-Series Forcasting 與 Regression 很像，只是它是用現有的數值來預測未來的數值。Time-Series Forcasting 的不同點在於它所分析的數值都與時間有關。Time-Series Forcasting 的工具可以處理有關時間的一些特性，譬如時間的階層性（例如每個星期五個或六個工作天）、季節性、節日、以及其他的一些特別因素如過去與未來的關連性有多少。

(8). 因果關係 (Causation Modeling)。

這是一個在資料的屬性中發現因果關係的問題，使用一個 ”if-then” 形式的因果規則，表明條件（前項）和規則的當然結果（後項）之間有相互關係。

(9). 關聯規則 (Association Rule)

是要找出在某一事件或是資料中會同時出現的東西。Association 主要是要找出下面這樣的資訊：如果 項目 A 是某一事件的一部份，則項目 B 也出現在該事件中的機率有 n %。（例如：如果一個顧客買了低脂乳酪以及低脂優酪乳，那麼這個顧客同時也買低脂牛奶的機率是 85%。）

(10). 屬性導向歸納法 (Attribute Oriented Induction)

屬性導向歸納法是一種以歸納屬性為基礎的資料分析技術，其技術核心為線上資料歸納方法，將相關式表格(relational dataset)資料集合中的每

一個屬性，檢查其資料的分佈，判斷應歸納到那個相關的抽象層級[12]。

(11).樣式導向相似性搜尋 (Pattern-Based Similarity Search)

在時間或時間-空間資料庫搜索相似的樣式，經常會應用到兩種查詢類型：

a. 物件關聯相似度查詢（Object-relative similarity query），亦即相似度查詢（similarity query）或範圍查詢（range query）

   在所收集到的物件中，尋找使用者指定的範圍或距離中，符合的物件。

b. 完全關聯相似度查詢（All-pair similarity query），亦即空間聯合（spatial join）

   目標是找到彼此都是在一段使用者指定的範圍或距離內的全部相符的要素。

(12). 資料方塊法 (Data Cube)

資料方塊法一般概念為將經常被要求的高成本計算具體化，尤其是計數(count)、總計(sum)、求平均數(average)、取最大值(max)等的歸納函數，將歸納後的具體化景觀儲存在一個多重維度資料庫(資料方塊)，可供決策支援、知識發現及其他應用做參考[12]。

(13).序列樣式探勘 (Sequence Pattern Mining)。

在包含時序關係的資料庫中尋找一定數量所支持的序列樣式，主要是找出關聯順序進行行為模式上的預測，例如若 A 事件發生，則 B 事件可能接著會發生。

2.3. 機器學習

機器學習(Machine Learning, ML)通常與分類(Classification)畫上等號,是根據一些變數的數值做計算，再依照結果作分類，而這個計算與分類的工作是透過機器利用演算法經由自動學習的過程，由機器尋找出分類的結果。

機器學習是用來協助知識獲取的工作，要從龐大的資料中作知識粹取的工作,光靠人力是很難達到的，唯有借助機器可快速執行重複運算的能力，經過適當的演算法和理論,才有辦法協助我們來探索前所未知的知識，也因此成熟的機器學習理論研究，也成了發展人工智慧系統裡，重要的一環。Peter Clark（1990）闡述機器學習並不是為了解決外部的問題，而是改善對知識本身的陳述[24]，Langley（1996）定義機器學習的演算法改善本身的執行效率是根據演算法本身的經驗[30]。ML 常見的演算法如下[29]：

(1). 古典的統計方法（Classical Statistical Methods），例如線性區別分析（Linear Discrinant Analyses）、二元區別分析（Quadratic Discriminant Analyses）、邏輯區別分析（Logistic Discriminant Analyses）。

(2). 現代的統計技術（Modern Statistical Techniques），例如投影追蹤分類法（Projection Pursuit Classification, PPC）、密度推估法（Density Estimation）、k 個最近鄰居分類法（k-Nearest Neighbor），因果網路（Casual Networks）、貝氏理論（Bayes theorem）。

(3). 類神經網路（Neural Networks），例如倒傳遞網路（Back-Propagation Network）、Kohonen 網路模型、機率神經網路(Probabilistic Neural Network，PNN)、霍普菲爾網路(Hopfield Neural Network, HNN) 與適應共振理論網路(Adaptive Resonance Theory Network, ART)、雙向聯想記憶

網路(Bidirectional Associative Memory Network, BAM)、放射函數網路（Radial Function Networks）。

(4). 支持向量機（Support Vector Machine, SVM）。

(5). 決策樹方法（Decision Tree Method s），例如 ID3、CN2、C4.5、T2、Lazy decision trees、OODG、OC1、AC、BayTree、CAL5、CART、ID5R、IDL、TDIDT、PROSM 等。

(6). 決策規則演算法（Decision Rule Algorithms）例如 AQ 系列、LERS 等。

(7). 分類學習系統（Learning Classifier Systems）例如 GOFFER-1 、MonaLysa 、XCS 等。

(8). 關聯規則演算法（Association Rule Algorithms），例如 APPIORI。

2.4. 規則歸納法

在機器學習眾多的理論當中，規則歸納(Rule Induction)的學習理論是最被廣泛討論及應用之一，規則歸納的主要涵義是從群訓練案例中尋找出最佳的、正確的、可了解的分類方法的規則[24]。規則歸納依據分類結果的表達方式大致上可分為兩大類，一類是結果以樹狀(Tree)的表達方式，以 ID3 系列[22]演算法為最具代表性，另一類的結果是以 if…then… 的條列式規則表達方法，以 AQ 系列演算法為最具代表性。

ID3 系列演算法的概念最早在 1966 年由 Hunt、Marin、Stone等人正式發表提出，當時演算法的名稱稱為概念式學習系統（Concept Learning System, CLS）。到了 1970 年代人工智慧(Artificial Intelligence AI)研究者 J.R. Quinlan 應用這個

觀念發展出實際的應用程式，並命名為ID3(Itemized Dichotomizer 3) ，自此ID3廣為學術界研究和引用，並發展出更多改良的演算法，圖 8 為直至 80 年代ID3系列的發展史。

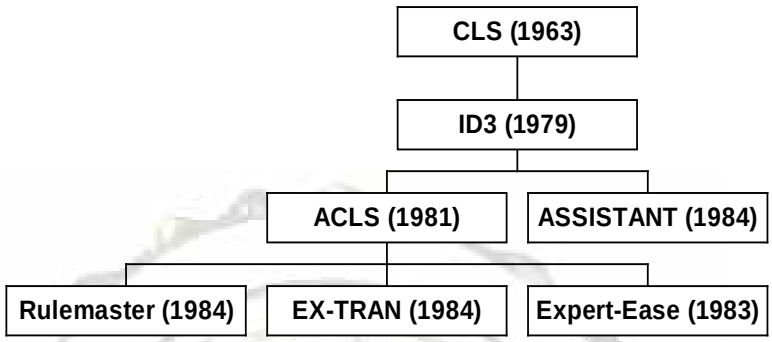

圖 8　ID3 家族早期的系統發展史

擬似最佳解演算法(Algorithm for Quasi-optimal solutions, AQ) [28]的概念最早是在 1969 年由Michalski（1969）所提出。AQ系列演算法與ID3 系列演算法不同的是其產生的結果以 if…then… 的條列式規則表達，而且不管輸入的資料的屬性是數值資料或文字資料，因此廣為決策支援系統(Decision Support System, DSS)及專家系統(Expert System, ES)所應用，在 1983 年Michalski 更引用了STAR演算法，改善了AQ演算法的正確性及效率，並命名為AQ15。在後續的專家學者的改進下，AQ系列演算法已研發了相當多的新理論，如下[20、22、28、33、37、38]：

(1). 1983 年 Larson 和 Michalski 的 AQ11。

(2). 1989 年 Clark 和 Niblett 的 CN2。

(3). 1986 年 Michalski, Mozetic, Hong, 和 Lavrac 的 AQ15。

(4). 1991 年 Bloedorn 和 Michalski 的 AQ17-DCI。

(5).1993 年 Bloedorn、Wnek、Michalski 和 Kaufman 的 AQ17-HCI。

(6).2000 年 Kaufman 和 Michalski 的 AQ18。

(7).2001 年 Michalski 和 Kaufman 的 AQ19。

(8).2004 年 Michalski 和 Kaufman 的 AQ21。

2.5. 分群演算法

分群演算法中最常用的是由 Makhoul 等（1985）所提出來的 K-平均法（K-means）[32]，然而為了解決 K-平均法對於分群的效能與邊界之不足，在 1969 年 Ruspini（1969）提出了模糊版本的 C-平均法演算法，後來由 J. C. Dunn(1973)[25] 和 J. C. Bezdek(1981) [18] 將模糊理論實際應用在集群分析上，提出了「模糊化 C - 平均法」(Fuzzy C-means Method)。

Fuzzy C-means（FCM）不但在分群能力上不輸傳統的 K-means 演算法，更可由其歸屬度矩陣(Relation Matrix)提供了分群的可靠度或邊界資訊。這在考量分群效果或衡量資料的集中程度時，就是有力的幫手[09]。

2.6. 抽樣理論

廣義的說「統計適用以蒐集和解釋關於某一特定研究（或調查）領域的數據，自不定性(Uncertainty)和變異(Variation)狀況中抽取結論的概念與方法。」乃是面對不確定的情況下，能夠幫助我們做出明智抉擇的一種科學方法 [13、19]，而統計在各行各業早已被廣泛的應用，凡是對於一些不確定或變異的現象，幾乎都可以用統計方法來處理。

「統計推論」則是利用樣本統計量去推論母體的特質，而樣本是否具有代表性會受抽樣方法的影響，因此抽樣方法非常重要。利用樣本而非母體的主要原因如下[04] ：

(1).有限的資源。

(2).毀壞性的試驗。

(3).概念性的母體無法全部觀察。

(4).樣本較母體小。

(5).在資料搜集與整理時較容易且精確。

抽樣基本目的乃在訊息之搜集作成結論，以供決策參考。有效抽樣應具有準則有下：

(1).有效原則：

抽樣應該符合調查目的之需要，所獲訊息價值應超過所支付成本。

(2).可測量原則：

抽樣的正確程度必須能夠測量，否則抽樣就失去意義。

(3).簡單原則：

抽樣必須保持簡單性要求。俾使抽樣順利進行，以避免不必要之節外生枝。

抽樣的類型可分為機率抽樣與非機率抽樣兩類[10] ，分述如下：

(1).機率抽樣 (Probability-Sampling)：

機率抽樣的原則：（隨機性原則）

母體中的每一個樣本被選中的機率相等。隨機抽樣具有健全之統計理論基礎，可用機率理論加以解釋，是一種客觀而科學的抽樣方法，在市場調查中通常都用隨機抽樣。

a. 簡單隨機抽樣 (Simple Random Sampling)：

母體中全部個體，完全委諸均勻機率分佈抽取樣本，使每一個體被抽出之機率均為已知且相等。簡單隨機抽樣為其它各種隨機抽樣方法之基礎。簡單隨機抽樣法樣本之取得，對母體編號後以利用隨機數表依機率抽取。

採用簡單隨機抽樣之時機：

- 母體小，母體名冊令人滿意且為母體訊息唯一來源。
- 單位訪問成本不受樣本單位所在地遠近之影向。

b. 系統抽樣 (Systematic Sampling)：

把母體的單位進行排序，再計算出抽樣距離，然後按照這一固定的抽樣距離抽取樣本。第一個樣本採用簡單隨機抽樣的辦法抽取。

K（抽樣距離）=N（母體規模）/n（樣本規模）

前提條件：母體中個體的排列對於研究的變量來說，應是隨機的，即不存在某種與研究變量相關的規則分佈。可以在調查允許的條件下，從不同的樣本開始抽樣，對比幾次樣本的特點。如果有明顯差別，說明樣本在母體中的分佈成某種循環性規律，且這種循環和抽樣距離重疊。此法優點，抽樣操作簡單，但有發生抽樣誤差的可能為其缺點。

c. 分層抽樣 (Stratified Sampling)：

先將母體中的所有單位按照某種特徵或標誌（性別、年齡等）劃分成若干類型或層次，然後再在各個類型或層次中採用簡單隨機抽樣或用系統抽樣的辦法抽取一個子樣本，最後，將這些子樣本合起來構成母體的樣本。

兩種方法：

- 先以分層變量將母體劃分為若干層，再按照各層在母體中的比例從各層中抽取。
- 先以分層變量將母體劃分為若干層，再將各層中的元素按分層的順序整齊排列，最後用系統抽樣的方法抽取樣本。

分層抽樣是把異質性較強的母體分成一個個同質性較強的子母體，再抽取不同的子母體中的樣本分別代表該子母體，所有的樣本進而代表母體。

分層標準：

- 以調查所要分析和研究的主要變量或相關的變量作為分層的標準。
- 以保證各層內部同質性強、各層之間異質性強、突出母體內在架構的變量作為分層變量。
- 以那些有明顯分層區分的變量作為分層變量。

分層的比例問題：

- 按比例分層抽樣：根據各種類型或層次中的單位數目占母體單位數目的比重來抽取子樣本的方法。

- 不按比例分層抽樣：有的層次在母體中的比重太小，其樣本量就會非常少，此時採用該方法，主要是便於對不同層次的子母體進行專門研究或進行相互比較。如果要用樣本資料推斷母體時，則需要先對各層的數據資料進行加權處理，調整樣本中各層的比例，使數據恢復到母體中各層實際的比例架構。

d. 群集抽樣 (Cluster Sampling)：

抽樣的單位不是單個的個體，而是成群的個體。它是從母體中隨機抽取一些小的群體，然後由所抽出的若干個小群體內的所有元素構成調查的樣本。對小群體的抽取可採用簡單隨機抽樣、系統抽樣和分層抽樣的方法。

優點：簡便易行、節省費用，特別是在母體抽樣框難以確定的情況下非常適合。

缺點：樣本分佈比較集中、代表性相對較差。

一般來說，類別相對較多、每一類中個體相對較少的做法效果較好。

分層抽樣與群集抽樣的區別：

分層抽樣要求各子群體之間的差異較大，而子群體內部差異較小；群集抽樣要求各子群體之間的差異較小，而子群體內部的差異性很大。換句話說，分層抽樣是用代表不同子群體的子樣本來代表母體中的群體分佈；群集抽樣是用子群體代表母體，再透過子群體內部

樣本的分佈來反映母體樣本的分佈。

e. 分段抽樣 (Subsampling)：

按照元素的隸屬關係後層次關係，把抽樣過程分為幾個階段進行。適用於母體規模特別大，或者母體分佈的範圍特別廣時。

類別與個體之間的平衡問題：

- 各個抽樣階段中的子母體同質性程度。
- 各層子母體的人數。
- 研究所能提供的人力和經費

缺點：每級抽樣時都會產生誤差

措施：增加開頭階段的樣本數，同時適當的減少最後階段的樣本數。

f. 複合抽樣 (Replicated Sampling)

將母體分為若干層，用系統抽樣法選取樣本。因此有分層抽樣及系統抽樣優點。

(2). 非機率抽樣：

不是按照等機率原則，而是根據人們的主觀經驗或其他條件來抽取樣本。常用于探索性研究。

a. 便利抽樣(Convenience Sampling)

在樣本之選擇只考慮到接近樣本或衡量便利。如訪問過路行人即為一例。

b. 判斷抽樣(Judgement Sampling)

在母體之構體極不相同且樣本數很小之時，根據抽樣設計者之判斷來選擇樣本個體，設計者必須對母體有關特徵具有相當了解。在編製物價指數時，有關產品項目選擇及樣本地區之決定，即採用判斷抽樣。

c. 配額抽樣 (Quota Sampling)：

選擇「控制特徵」，作為將母體細分類之標準，將母體細分為幾個子母體，按比較分發各子母體樣本數大小，訪查員有極大自由去選擇子母體中之樣本個體，只要完成配額調查，即告完成。

配額抽樣與分層抽樣之比較：

前者注重的是樣本與母體在架構比例上的表面一致性；后者一方面要提升各層間的異質性與同層的同質性，另一方面也是為了照顧到某些比例小的層次，使得所抽樣本的代表性進一步提升，誤差進一步減小。在機率上，前者是按照事先規定的條件，有到達站尋找；后者是客觀地、等機率地到各層中進行抽樣。

d. 滾雪球抽樣 (Snowball Sampling)

利用隨機方法或社會調查選出原始受訪者。再根據原始受訪者提供訊息去取得其它受訪者。本法之目的乃母體很難尋找或十分稀少。

第 3 章. 研究方法

本章主要敘述研究的構想、步驟以及所應用的相關理論與演算法，整個架構則是建基在第二章所描述的知識探索五個流程與九個步驟。主要的步驟可分為三大主軸，首先為資料蒐集與分層抽樣，其次為規則歸納與驗證，最後為實驗結果與比較。

3.1. 研究構想

如第一章緒論所述，由於消費金融已為各銀行所重視，國內外對於消費金融相關研究非常多，大都著眼在使用動機、風險控管、推廣促銷與客戶忠誠度方面，在推論結果上通常著重於準確度或正確率，對於依據客戶消費結果良窳分類，尋找各族群共通之特徵，以提供後續業務推展、決定授信額度之依據，往往由於資料量龐大、計算時間過長，所找出的規則無法經常更新，致使線上所使用的准駁規則與最近的消費習性有所差距。

因此經過精密、複雜的演算法所推論的結果，即使當下的準確度或正確率高，但當運用於實際資料時，反因時間的差距而有所誤差，因此若能夠在可容許的誤差範圍內，快速地找出當時資料的特徵規則，不但節省為了精密計算所投下的計算資源，更可快速得反應當時資料的屬性。

而客戶忠誠度深深影響企業獲利之結果[17]，在影響忠誠度諸多因素中[05、15] （如圖 9），以「人口統計變數」為銀行在提供消費者服務時第一時間就能直接獲取的資料，因而若能從歷史交易資料庫中，依據人口統計變數資料，歸納出各類族群在消費後繳款情形的共同特徵之規則，如此客戶提出申請時，便可依據客戶所填寫申請文件中的人口統計變數，於第一時間提出客戶未來消費習

性之預測值，提供額度准駁決策之依據。

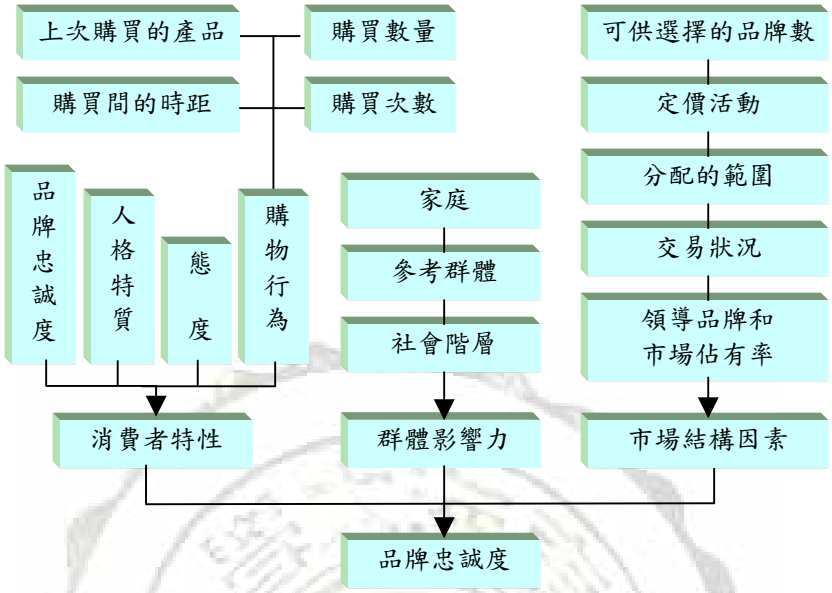

圖 9　影響忠誠度的因素

3.2.　研究方法與步驟

本研究過程主要的步驟可分為三大主軸，首先為資料蒐集與分層抽樣，其次為規則歸納與驗證，最後為實驗結果與比較，研究的步驟如下：

(1). 進行問題之定義與知識探索的目標：

問題之定義：探討在可容許的誤差範圍內，以較快速的方式，取得資料分類規則之可行性。

知識探索的目標：以對發卡行貢獻狀態分類，尋找信用卡客戶人口統計變數資料之規則，提供未來發卡或業務推廣之參考。

(2). 定義各項參數及抽樣方式。

(3). 資料收集、資料編組及分類，對於非離散性資料（年齡）應用模糊分群法離散化。

(4). 定義樣本及分層空間並進行資料分層與抽樣（如圖 10）。

(5). 以所有的資料組隨機抽取 80% 為訓練組資料，剩下的 20% 為驗證組資料。

(6). 依據訓練組資料由 iAQ 軟體進行規則尋找產生分類規則，再以驗證組資料驗證分類結果，並計算正確率（如圖 11）。

(7). 依據公式 2 抽取樣本，重複步驟五與六 30 次，取得分層抽樣平均正確率。

(8). 比較原始資料與分層抽樣兩者之正確率。

(9). 對於實驗結果進行解釋，並討論結果與分析。

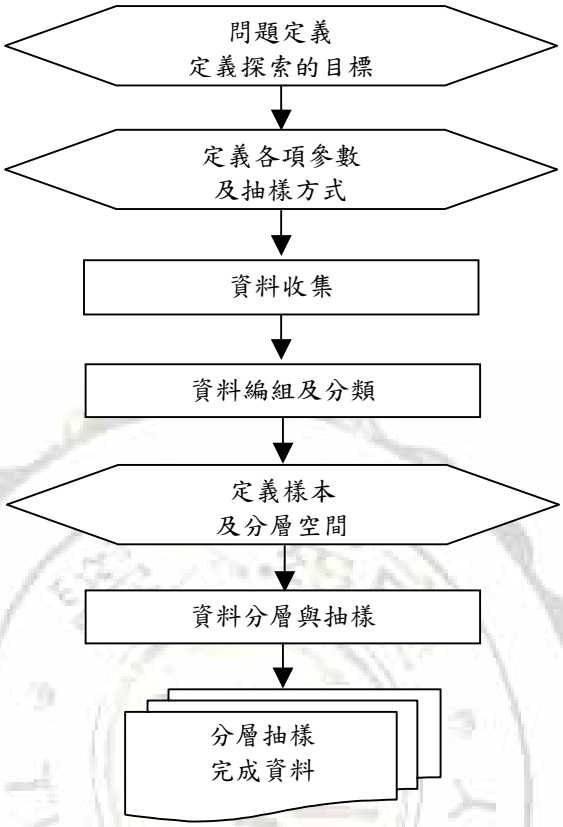

圖 10 資料收集與抽樣流程圖

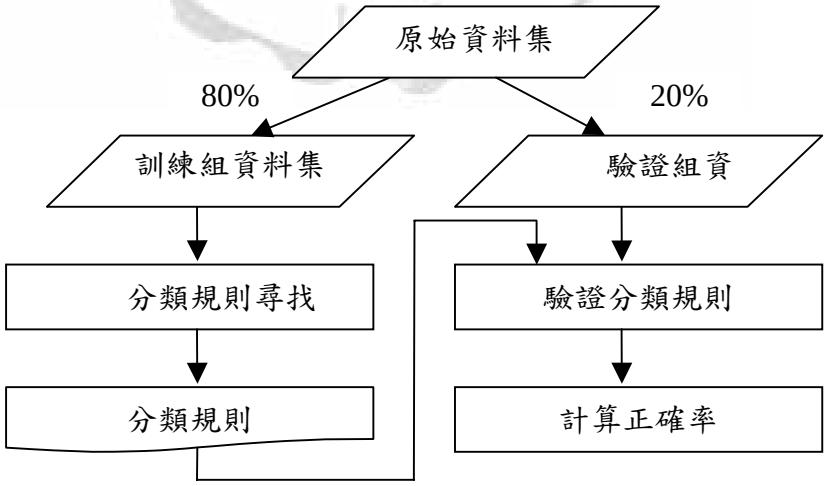

圖 11 分類規則的尋找與驗證

3.3. 研究理論架構

3.3.1. 模糊分群演算法

Fuzzy C-means（FCM）是一種允許一個數據屬於兩群或更多的群組的方法。有時也被稱為模糊ISODATA，經常被在圖樣識別過程中使用。與K-means最大的不同在於融入了模糊的概念，每一個資料點x不再絕對地屬於任何群組，取而代之的是將x隸屬於某個群組的程度用一個介於 0~1 之間的數字來表示。假設c（c_1, c_2, …, c_c）為預期的分群數目，欲分群的資料為n點（x_1, x_2, …, x_n），矩陣U以一個 $c \times n$的來表示每個資料點隸屬於每個群組的程度，任一點x_j隸屬於各個群組的程度總和應該正好等於 1。

於是
$$\sum_{i=1}^{c} u_{ij} = 1, \forall j = 1,2,..., n \quad \text{.....（3-1）}$$

根據矩陣 U，我們可以定義出我們的目標函數（objective function）J：

$$J_m = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}{}^{m} \left\| x_j - c_i \right\|^2 \quad \text{.....（3-2）}$$

其中，m為大於 1 的任何數實，u_{ij}是x_j在第j群的歸屬度，x_j是矩陣中第j層第i個資料點，c_i則是每一群的中心點。$\left\| x_j - c_j \right\|^2$是$c_i$與 x_j之間的距離函數。

模糊的劃分則是透過上面顯示的目標函數方程式（3-1）以疊代的方式，反覆的更新歸屬度u_{ij}與中心點c_i進行最佳化。

其中：

$$u_{ij} = \left[\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{kj}} \right)^{\frac{2}{m-1}} \right]^{-1} \quad \text{.....(3-3)}$$

$$c_i = \frac{\sum_{j=1}^{n} (u_{ij})^m x_j}{\sum_{j=1}^{n} (u_{ij})^m} \quad \text{.....(3-4)}$$

整個疊代的程序直到 $\max_{ij}\left\{\left|u_{ij}^{k+1} - u_{ij}^{k}\right|\right\} < \varepsilon$，$\varepsilon$ 為介於 0 和 1 之間的結束門檻值，k 為重複執行的步驟。

根據上面的定義，整個 Fuzzy C-means 分群法的步驟如下：

1. 隨機填寫矩陣 U 中各行列位置的數值(介於 0~1)，但須滿足方程式(3-1)。

2. 在第k次步驟，根據方程式（3-4），計算各群聚的群聚中心c_i。

3. 根據方程式（3-2），計算本步驟的目標函數並記錄為J_{now}。假如J_{now}已經小於某個標準，或者這次的分群改良效果($J_{pre} - J_{now}$)(J_{pre}為上一次的J_{now})已經過小，則結束本演算法。

4. 根據方程式（3-3）計算新的矩陣 U，並回到步驟二。

3.3.2. 分層抽樣理論

如 2.6 節所述，利用抽樣技術及機率統計理論，除了可節省調查人力，物力，時間及經費外，更可獲得既定精確估計值，以代表母群體特徵。本研究著重於依據客戶貢獻度分為已列呆帳、催收、逾期的 Bad、無消費的 Lethargy 以及使用循

環信用，繳息正常的 Good 三類（層）。由於此三層異質性很高，為了要確保抽出樣本的特質分配與母體相一致，因此採用分層抽樣法，先依據各層應佔樣本比例均計算妥當，後依照簡單隨機抽樣法抽出所需要的樣本數[10]。

本研究分層抽樣原則係依據 William Mendenhall（1995）等所提的分層抽樣方法[35]，如下所述：

總抽樣樣本數為 n 可以公式（3-5）計算

$$n = \frac{\sum_{i=1}^{L} \frac{N_i^2 p_i q_i}{w_i}}{N^2 D + \sum_{i=1}^{L} N_i p_i q_i} \quad \text{……..……(3-5)}$$

每一層抽樣樣本數為n_i可以公式（3.3.2-2）所示

$$n_i = n \frac{N_i \sqrt{p_i q_i / c_i}}{\sum_{i=1}^{L} N_i \sqrt{p_i q_i / c_i}} \quad \text{……..……(3-6)}$$

其中：

N 為母體(population)樣本總數，本實驗取樣本總數為 1999 戶。

N_i 為每一層之樣本總數，本實驗N_1(Bad)為 600 戶，N_2(Lethargy)為 400 戶，N_3(Good)為 999 戶。

w_i為第i層占樣本總數的比例。

D 為 $\frac{B^2}{4}$。

B 為容許估計誤差上限(Bound on the error of estimation)，本實驗定為 0.1，

故 D 為 0.0025。

L 為欲分層之層數，依據表一的客戶貢獻狀態分為 3 層。

p_i　　為第 i 層的正確分類比率，其中 $p_1 = 0.3$ ， $p_2 = 0.2$ ， $p_3 = 0.5$ 。

q_i　　為 $1 - p_i$ 。

c_i 為第 i 層抽樣成本,因為本實驗樣本均由資料庫所得,故所有的 c_i 均為 1。

i 為第 i 層之編號。Bad 為 i=1，Lethargy 為 i=2，Good 為 i=3。

3.3.3. 規則歸納法概述

規則歸納法是知識探勘的領域中最常用的方式，這是一種由一連串的「如果.../則...（If / Then）」之邏輯規則對資料進行細分的技術，在實際運用時如何界定規則為有效是最大的問題，通常需先將資料中發生數太少的項目先剔除，以避免產生無意義的邏輯規則。

簡單的規則歸納法演算步驟可用兩個迴圈完成，如下：

(1). 首先依據類別排序。

(2). 最外層的迴圈以類別為主，由第一個類別開始尋找符合此類別的規則，直到所有類別尋找完畢，完成規則的尋找，結束整個尋找工作。在開始每一層迴圈前，將所有此類別的資料放入候選區，作為規則測試資料。

(3). 內層迴圈主要針對上一層迴圈的類別作規則的尋找，並以上一層所準備的候選區測試資料尋找規則。

a. 假設第一個屬性內容足以歸類到此類別，以此規則驗證所有資料。

b. 若假設成立，則輸出此條規則，並將此筆資料自候選區測試資料剔除。

c. 若第一個屬性無法滿足條件，則假設第二個屬性，並重複步驟 b。

d. 如果單一屬性無法滿足條件，則增加一個屬性一併考慮。

e. 重複步驟 i ~ iv，直到候選區測試資料被清空。

(4). 重複步驟 1 ~ 3 直到所有類別均完成。

以表 4 為例，其中 NO 為資料記錄編號，F1 ~ F4 為資料屬性，Class 為已知的分類：

NO	F1	F2	F3	F4	class
1	0	1	0	2	Zero
2	1	1	0	2	Two
3	0	0	0	1	Zero
4	0	1	1	0	One
5	0	0	1	3	Zero

表 4 原始測試資料集

經過第一步驟依據 class 排序後，產生如表 5

NO	F1	F2	F3	F4	class
4	0	1	1	0	One
2	1	1	0	2	Two
1	0	1	0	2	Zero
3	0	0	0	1	Zero
5	0	0	1	3	Zero

表 5 排序後的原始測試資料集

經過第 1 迴圈以類別 One 找出第一條規則：

If F4 = 0 then D = One

由於沒有其他候選區測試資料，繼續尋找類別 Two，找出：

If F1 = 1 then D = Two

繼續第三個類別 Zero，找到類別 Zero 第一條規則

If F2 = 0 then D = Zero

此時已無法用單一個屬性作為判斷條件，來滿足所有的測試條件，所以必須增加一個屬性來判斷，於是找到的最後一條規則

If F1 = 0 And F3 = 0 then D = Zero

最後得到的規則關係圖（如圖 12）如下：

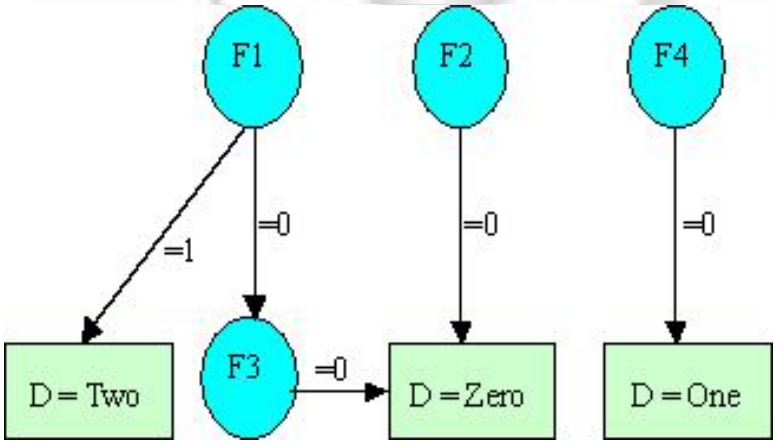

圖 12 規則歸納後的規則關係圖

3.3.4. AQ 演算法

AQ 演算法和 ID3 不同的是 ID3 是產生決策樹，而 AQ 是產生決策規則，ID3 是將每一個屬性排序，並計算每一個屬性的熵（Entropy），以 Entropy 最低的屬性作為分支頂點（Note）再往下分支，AQ 演算法則是先依據欲分類的類別排序，再依據每一個類別去尋找符合條件的規則。

AQ 演算法主要有兩層的循環，最外層是以每一個類別為主，每一個類別都必須測試'，最內層則是單一類別中測試每一個屬性找出可以符合類別的條件。

AQ 演算法如下所示（如圖 13）：

```
let examples = a set of training examples
let classes = the set of all classes

procedure aq(examples, classes);
let allrules = {}
for each class c in classes:
      sort examples into pos (members of c) and neg (the rest)
      generate rules by aqrules(pos, neg, c)
      add rules to allrules
endfor
return allrules.

procedure aqrules(pos, neg, c):
let rules = {}
for each member of pos (each 'seed') not covered by any rule on rules:
      call aqrule (seed, neg, c) to generate a rule covering seed
      add rule to rules
end for
return rules

procedure aqrule(seed, neg, c):
let mgc = the most general <condition> ('true')
let star initially contain only the mgc
for each negative example n in neg:
      for each <condition> c in star:
            if     c covers n
            then remove c from star
               & generate all specialisation of c with still cover seed
                  but no longer cover n by adding an extra attribute test to c
                  ('c' becomes 'c & test' for each test true of seed and false of n)
               & add them all to star.
            if     size of star > maxstar (a user-defined constant)
            then remove worst <condition>s in star until size of star = maxstar.
      endfor
endfor
select the best <conition> bestcond in star
return the rule 'if bestcond then predict c'
```

圖 13 AQ 演算法

第 4 章. 模擬實驗

4.1. 問題定義

如前所述，銀行在新增一個客戶時，「人口統計變數」在第一時間就能直接由客戶所填寫的資料中獲取。而一般信用卡客戶申請信用卡的動機大約有下列幾種情形：

(1).人情壓力，來自親友的業績壓力。

(2).辦卡可以兌換贈品。

(3).卡片提供足夠吸引的配套促銷方案或活動：如高額的旅遊意外險、免費拖吊、百貨公司週年慶可兌換贈品、搭配公益活動等。

(4).本身之需求：如經常搭飛機出差、旅遊、不喜歡帶太多現金。

(5).可以先消費後付款：本身經常透支生活費，或由父母配偶支付卡款。

(6).可以預借現金：沒有儲蓄觀念、月光族。

第(1)(2)類的人較容易形成辦卡後沒有消費的呆卡，第(3)(4)類的人，雖然會刷卡消費，但大部份收入固定，自主能力強，每月均會繳清消費款，銀行大多只能賺取手續費，第(5)(6)類是銀行又愛但又怕受傷害的族群，這一類的人通常會動用到循環息，是銀行信用卡主要的利息收入來源，但卻也常因收入不穩定造成逾期繳款甚至轉列催收、呆帳而造成銀行的損失。

因此本研究將客戶在辦卡一年後的消費情形分為已列呆帳、催收、逾期的 Bad，無消費的 Lethargy，以及使用循環信用,繳息正常的 Good 三群，希望從歷

史的交易資料庫中，能快速找出這三群個別的共同特徵，提供後續的應用。

4.2. 實驗設計

4.2.1. 資料欄位說明

根據 Kotler（1989）指出消費者的人口統計變數包括了：年齡、性別、教育、職業、宗教、種族、家庭人數、家庭生命週期、國籍、所得等[15、16]。本研究所探討的變數有以消費者在申請時所填寫的資料為主，包括年齡、性別、教育程度、婚姻狀況、居住地、住宅情形、行業別、年資等八個變數。

各變數對於信用卡的影響程度分述如下：

(1).年齡：影響信用卡客戶消費觀念、還款能力等。

(2).性別：影響消費能力。

(3).教育程度：影響消費、還款觀念。

(4).婚姻狀況：影響消費型態、還款穩定性。

(5).居住地：城鄉的差距影響消費習性。

(6).住宅情形：影響消費觀念、還款穩定性。

(7).行業別：影響消費、還款觀念。

(8).年資：影響消費型態、還款穩定性。

表 6 為各變數之級距：

表 6　各項參數定義與級距

屬性名稱	參數定義與級距
年齡(Age)	12~31：A1， 32~39：A2， 40~47：A3 48~58：A4， 59 以上：A5
性別(Sex)	男：Male，女：Female
教育程度 (Education)	國中小：Primary，高中、專科：Senior， 大學：College，碩士：Master，博士：Doctor
婚姻狀況 (Marital)	未婚：Single，已婚：Married
居住地 (Zip)	鄉鎮：Rural，縣轄市：Town，省轄市：City，直轄市：Capital
住宅情形 (House)	公司宿舍：E，自有無房貸：F，自有已房貸：M 親屬：P，租賃:R，其他：O
行業別 (Trade)	傳統產業：A，電子產業：B，金融業：C，服務業：D，其他： E
年資 (Seniority)	未滿 1 年：S1，1~3 年：S2，3~5 年：S3，5~10 年：S4， 10 年以上：S5
客戶分類 (Group)	已列呆帳、催收、逾期：Bad 無消費：Lethargy 使用循環信用，繳息正常：Good

4.2.2.　模糊分群說明

以發卡行信用卡採簡單隨機抽樣方式抽樣之 6 萬多戶資料為例：

若將年齡分為五群，分別為 0~20、21~30、31~40、41~50、51~100，統計數如表 7，其分佈圖如圖 14：

表 7　原始年齡層的個群統計數

族群	0~20	21~30	31~40	41~50	51~100
統計數	129	11944	23677	21267	11784

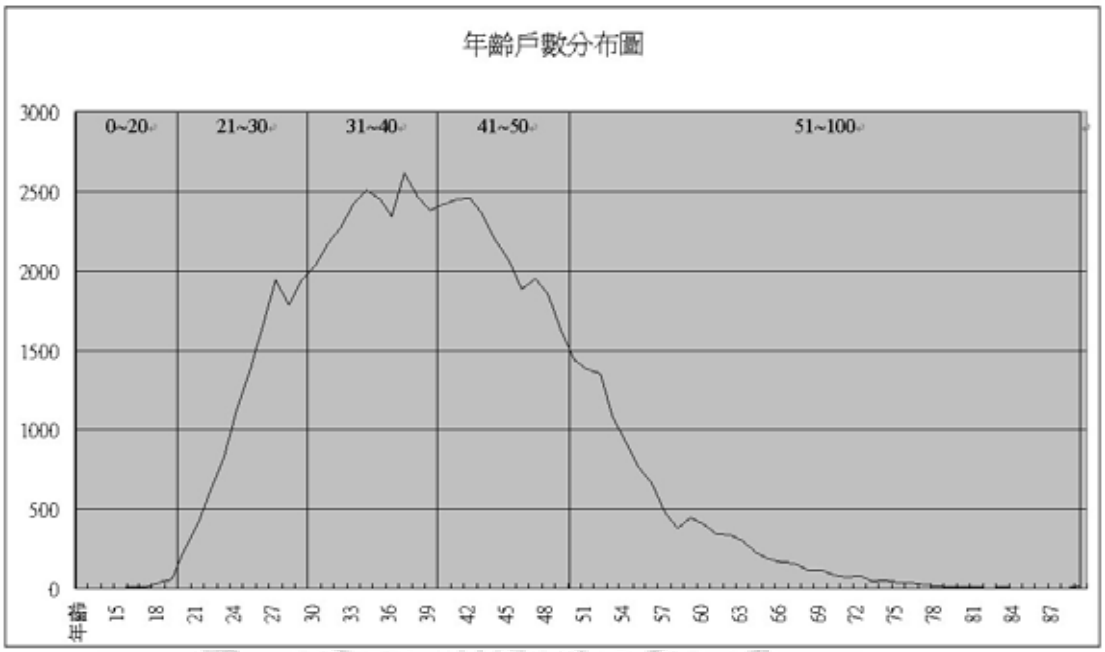

圖 14 年齡層分佈與原始級距分群圖

若改以模糊分群法進行族群之分類，表 8 為以模糊分群法後，各群所佔人數統計表：

表 8　模糊分群法各群所佔人數統計表

年齡族群	12~31	32~39	40~47	48~58	59~93
族群編號	A1	A2	A3	A4	A5
統計數	14110	19259	18202	13531	3699

所得之年齡層族群分類結果如圖 15：

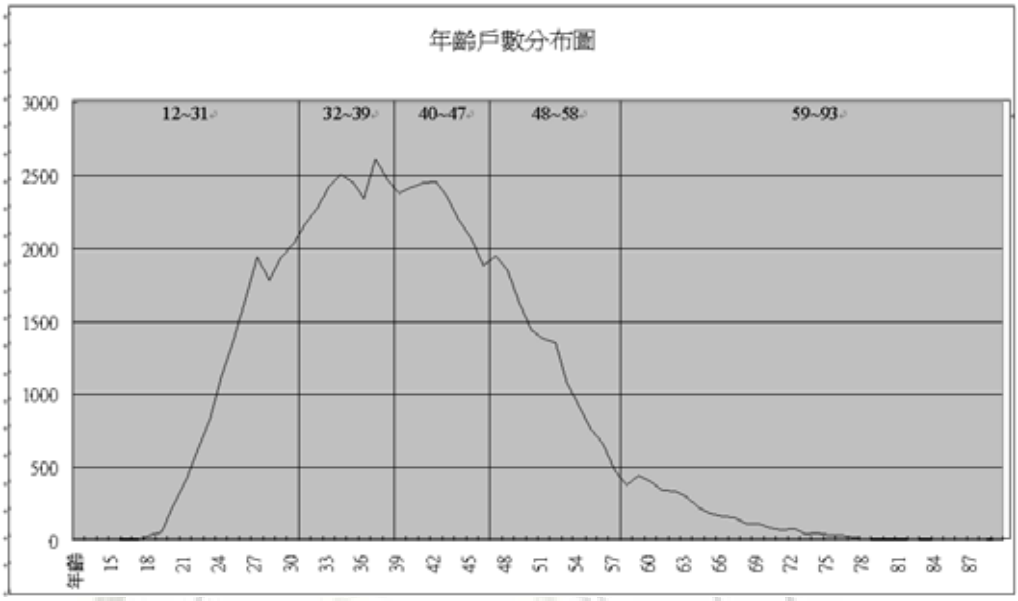

圖 15 年齡層分佈與模糊分群法進行族群之分群圖

由圖 14、表 7 與圖 15、表 8 之比較，我們可觀察到經過模糊分群法後所得到的分群級距較接近資料的原始分佈情形，而以此分群之結果進行規則尋找，亦較能減少錯誤之發生。

4.2.3. 分層抽樣原則

經過公式(3-5)、(3-6)計算 本研究將抽樣的參數如表 9：

表 9　各層抽樣結果

n_1	Bad　　層抽樣樣本數	25
n_2	Lethargy　層抽樣樣本數	16
n_3	Good　　層抽樣樣本數	42
n	抽樣結果總數	83

4.2.4.　　使用 AQ 軟體說明

本研究使用 MLI 實驗室(Machine Learning and Inference Laboratory)所提供的 iAQ 軟體[31]，AQ 軟體提供 AQ 規則歸納演算法，依據輸入的資料產生所尋找的規則，iAQ 軟體畫面如圖 16 所示。

圖 16　iAQ 軟體畫面

4.2.5. 實驗環境與過程說明

本實驗使用的硬體設備如下：

CPU ： INTEL Pentium IV 2.4 G

MEMERY： 512 MB

主機板 ： 華碩 P4P800

資料庫 ： Microsoft Access 2000

使用程式 ： Microsoft Excel 2000 VBA

為完成如圖 11 之流程，本實驗將以 Access 及 Excel 相互搭配，以 VBA 撰寫抽樣及驗證之相關程式，主要的步驟如下：

(1). 實驗先由 Access 中匯入 Excel，以簡單隨機方式分成訓練資料組 80%及驗證資料組 20%，並以訓練資料組產生 iAQ 所需要的輸入資料。

(2). 執行 iAQ 產生規則。

(3). 由 iAQ 產生的結果中抽取出規則資料匯入 Excel 中，並以驗證資料組驗證這些規則，紀錄正確率。

(4). 實驗的第一輪迴使用未分層抽樣之 1999 筆資料，作為我們希望獲得規則的主要來源，其餘 30 個輪迴使用分層抽樣後的資料，作為正確率的比較組，彼此在進行交互驗證，以比較未分層抽樣與分層抽樣正確率之差別。

4.3. 研究結果與分析

4.3.1. 實驗結果

經過實際資料執行之結果，彙整如表 10

表 10 各項驗證規則正確率彙整與 CPU 耗費時間表

規則產生來源	分層抽樣後80%訓練組	分層抽樣後80%訓練組	未分層抽樣80%訓練組	未分層抽樣80%訓練組	CPU 耗費時間(秒)
驗證資料來源	分層抽樣後20%驗證組	未分層抽樣20%驗證組	分層抽樣後20%驗證組	未分層抽樣20%訓練組	
未分層				81.25%	1271.938
分層第 1 次	87.50%	78.45%	93.75%		0.429
分層第 2 次	62.50%	76.94%	81.25%		6.453
分層第 3 次	87.50%	84.96%	100.00%		4.155
分層第 4 次	75.00%	84.46%	87.50%		9.047
分層第 5 次	81.25%	82.46%	93.75%		0.75
分層第 6 次	81.25%	86.97%	93.75%		3.594
分層第 7 次	81.25%	84.46%	93.75%		1.922
分層第 8 次	75.00%	83.71%	93.75%		1.922
分層第 9 次	87.50%	89.22%	100.00%		1.375
分層第 10 次	75.00%	80.45%	87.50%		0.625
分層第 11 次	81.25%	84.96%	87.50%		0.531
分層第 12 次	87.50%	79.20%	93.75%		1.781
分層第 13 次	93.75%	79.45%	100.00%		1.453
分層第 14 次	68.75%	88.22%	93.75%		4.453
分層第 15 次	87.50%	86.72%	100.00%		2.107
分層第 16 次	81.25%	85.96%	93.75%		1.531
分層第 17 次	87.50%	84.46%	93.75%		2.199
分層第 18 次	81.25%	84.71%	100.00%		0.235
分層第 19 次	93.75%	79.70%	100.00%		6.958
分層第 20 次	75.00%	86.22%	100.00%		12.467
分層第 21 次	75.00%	84.21%	81.25%		6.513
分層第 22 次	68.75%	86.97%	93.75%		4.77
分層第 23 次	87.50%	88.22%	100.00%		3.018
分層第 24 次	81.25%	88.22%	87.50%		2.734
分層第 25 次	62.50%	77.19%	100.00%		5.156
分層第 26 次	81.25%	80.95%	100.00%		4.547
分層第 27 次	62.50%	86.22%	93.75%		2.5
分層第 28 次	87.50%	81.45%	87.50%		0.375
分層第 29 次	75.00%	82.96%	87.50%		5.109
分層第 30 次	93.75%	82.46%	87.50%		1.859
分層之平均	80.21%	83.68%	93.54%		3.3522667

由未分層抽樣之 1999 筆資料所產生客戶分類 Bad 之規則共 81 條（前 40 條如表 11），Good 之規則共 56 條（前 40 條如表 12），Lethargy 之規則共 48 條（前 40 條如表 13）。

表 11 客戶分類 Bad 之規則前 40 條

AGE	SEX	Education	Marital	Zip	House	Trade	Seniority
=A3,A4	=Female		=Married		=P,R	=D	=S1
=A3,A4					=F,P,O	=A	=S5
=A2	=Male	=Senior,College	=Single		=P,R	=A,E	=S3,S4,S5
=A2,A5	=Male	=Senior,College		=Rural,Town	=P,R	=D	=S4,S5
=A3,A4	=Male			=Rural,Town	=F,P	=A	=S4,S5
=A4	=Male	=Primary,Senior,College			=E,F,M	=D,E	=S1,S5
=A4	=Male	=Primary,College		=Rural,Capital	=F,M,R	=E	=S1,S4,S5
=A2,A3,A4	=Female	=Primary,Senior	=Married	=Town,Capital	=P,R	=A,E	
=A4		=Primary,Senior	=Married	=Rural,Town,City	=P,R,O	=A,E	=S1,S4,S5
=A2,A4		=Senior,College		=Rural,Capital	=P,R	=A,E	=S2,S3
		=Senior		=Rural,Town	=R	=A	
=A2,A4,A5	=Male	=Senior,College		=Town	=R	=A,E	=S1,S5
=A4	=Female			=Capital	=F,R	=E	=S1,S2,S4
	=Female	=Primary,College	=Married		=M,R	=A,E	=S2,S4,S5
=A2,A4	=Female	=Primary,College	=Married	=Rural,City,Capital	=F,P	=D,E	=S1,S4
		=Primary		=Rural,Town	=F,O	=E	=S1
=A3,A5	=Female	=Primary,Senior			=P,O	=A,B,E	=S3,S5
=A3,A4,A5	=Male	=Primary,Senior		=Town,City,Capital	=E,R	=A,E	=S1
=A3,A4	=Female	=Senior,College		=Town,Capital		=A,E	=S2
=A2,A4				=Rural,Capital	=P	=A,E	=S2,S3
=A2		=Primary,College	=Married	=Rural,City,Capital	=P	=D,E	=S1,S4
=A2,A4		=Primary,College		=Rural	=F,M,P	=E	=S1
=A3,A4	=Male		=Single	=Town	=F,R	!C	=S2,S3,S4
=A3,A4	=Male			=Rural,Town	=O	=B,E	=S1
=A3,A4		=Senior	=Married	=Rural	=E,R,O	=A,C	
=A3,A4,A5	=Male	=College	=Married	=Rural,Town,Capital	=P,R	=B,D	=S1,S3
=A2,A5	=Female	=Primary,College		=Rural	=F,O	=C,E	=S1,S4,S5
=A4,A5		=Senior,College		=Town,Capital	=M,P	=C,E	=S2,S4,S5
	=Female	=Primary,College	=Married		=P	=D,E	=S5
=A2,A4	=Female	=Primary,Senior	=Married	=Rural	=P,R	=C,D,E	
=A3,A4	=Male	=College		=Town,Capital	=R	=E	=S1,S3
=A2,A5		=Senior	=Married	=Rural	=F	=A	=S1,S5
=A3	=Male		=Single	=Rural,Town	=P,R	=A,E	=S2
=A3,A4	=Male	=College	=Single	=Rural,Town	=P,R,O	=E	=S1,S2
=A3,A5		=Primary,College	=Married	=Rural,Capital	=F,P	=E	=S3
=A3			=Married		=P	=B	=S1,S4,S5
=A3,A5			=Married	=Rural,City,Capital	=P	=B	=S1,S4,S5
=A4,A5		=College		=Town	=F,P	=B,C,E	=S5
=A4,A5	=Male		=Married	=Capital	=P,R	=A,B	=S1
=A3,A5	=Male	=College		=Rural,Town	=P	=A	=S1,S5

表 12 客戶分類 Good 之規則前 40 條

AGE	SEX	Education	Marital	Zip	House	Trade	Seniority
=A2,A3	=Female	=Senior,College		=Rural,Capital	=M,P,O	=A,B,C	=S1
=A2,A3,A4		=College	=Single		=F,P	=A,C	=S1
=A2,A3,A5	=Female	=Senior,College		=Rural,Capital	=M,P,O	=A,C	=S1
=A3		=Senior,College	=Single	=Rural,City	=E,F,R	=A,B	=S1
=A3		=Senior,College		=Rural,Capital	=M,P	=A	=S1
=A2,A3	=Male	=Senior,College	=Single	=Rural,Capital	=F,P	=A,C	=S1
=A2		=Senior		=Rural,Town	=E,M,P	=A,C	=S1
=A3,A4,A5	=Female			=Town	=E,M,P	=A	=S1
=A4		=Senior		=Rural,Town	=F,M	=A,B	=S1
=A3		=Senior	=Married	=Rural,Capital	=M,P	=A,C,E	=S1
=A2,A4	=Female			=Town,Capital	=E,F	=A,B	=S1
	=Male	=Senior		=Town	=F,M,P	=B,D	
=A3,A4,A5	=Female	=Senior	=Married	=Town	=E,M,P	=A,C	=S1
=A2			=Single	=City,Capital	=P,R	=A,B	=S1
=A3		=Senior,College	=Single	=Rural,Capital	=F,M	=B,C,E	=S1
=A3		=Senior,College	=Single	=Rural,Town,Capital	=M,R	=B,C	=S1
=A3,A4	=Female	=Senior		=Town,Capital	=F,M,P	=C,D	=S1
=A3		=Senior,College		=Town	=E,O	=A,E	=S1,S4
=A3			=Married	=Town,City	=M,P,R	=A	=S1
=A2,A4	=Male	=Senior,College		=Town	=P	=C,D	=S1
=A2,A4,A5	=Female	=Senior		=Town	=M,P	=A,B	=S1,S4
=A3,A4	=Male	=College	=Single	=Town	=F,P,R,O	=A,B	=S1
=A2,A3		=Senior		=Town,Capital	=F	=A	=S1
=A2,A4,A5	=Female			=Town	=E,F	=A	=S1
=A2		=College		=Rural,Town	=P,R	=A,B	=S1
	=Male	=College		=Rural	=E,F,M	=A	=S1
=A3,A4	=Male	=College	=Married	=Rural,Town	=F,M	=A,C	
=A3,A5	=Female	=College		=Town	=F	=B,C,E	=S1
=A3,A5	=Male	=Primary,Senior		=Rural,Town	=F	=B,E	=S1,S5
=A2,A5	=Female			=Town,Capital	=F,P	=D	=S3,S4
=A3,A4	=Female	=Senior		=Rural,Town,Capital	=P	=B	=S1
=A4,A5	=Male	=Senior,College		=Town	=M	=B	=S1
=A3,A4	=Male		=Single	=Capital	=P,O	=A	=S1
=A3,A4		=College		=Rural	=F	=C	=S1
=A3,A5	=Male	=Senior		=Rural	=P	=A,B,E	=S1
=A4,A5		=College		=Rural	=F,P	=B,D	=S1
=A2	=Female	=College	=Single	=Rural	=P	=A,E	=S5
=A3,A4	=Male		=Married	=Town	=P	=A,D	=S1,S2
=A3		=College		=Town	=R	=B	=S1
=A3	=Male	=Senior,College		=Town	=P	=B,E	=S5

表 13 客戶分類 Lethargy 之規則前 40 條

AGE	SEX	Education	Marital	Zip	House	Trade	Seniority
		=Master		=Town,Capital	=P,R,O	=D,E	=S1,S5
=A3	=Female	=College,Master				=D	=S1,S4
=A2,A3,A5		=College,Master	=Married	=Rural,Town	!R	=D	=S1,S5
	=Female	=Senior,College,Master		=City,Capital	=P,O	=D,E	=S1
=A3,A4,A5				=Town,Capital	=E,F,O	=D	=S2,S5
		=College,Master	=Married	=Rural,Town	=E,M,P,O	=D	=S1,S5
	=Female			=Rural,Capital	=F,O	=D	
=A3	=Male	=College,Master	=Married	=Rural,Capital	=E,F,P	=D,E	=S1,S2,S5
=A2,A3		=Senior,College,Master		=Rural,Capital	=E,M,O	=E	=S1,S2,S3
=A3,A5		=College,Master		=Town,Capital	=E,M,P	=C,D	=S1
=A2,A3,A5		=College,Master	=Married	=Town	=F,P,O	=E	=S2,S4
				=City,Capital	=P,R,O	=D	=S2,S4
=A3,A4	=Female		=Single	=Rural,Capital	=P,R	=D,E	=S2,S4,S5
=A3,A5	=Female		=Married	=Capital	=F,O	=E	=S1,S4
	=Female	=College,Master		=Town	=E,P	=D,E	=S5
=A3,A4	=Male	=Senior,College,Master			=P,O	=D	=S5
		=College,Master	=Married	=Rural	=E,F	=D,E	=S1,S2
=A3		=Primary,Senior,Master	=Single	=Rural,Capital	=P,R	=D,E	=S2,S4,S5
=A3,A5		=College		=Town	=P	=D,E	=S3,S4
=A3	=Male	=College,Master	=Single		=P,O	=D	=S1,S5
=A3,A5		=Primary,College,Master	=Married	=Town,Capital	=F	=E	=S4,S5
=A2,A4		=College,Master		=Town,City	=F,P	=E	=S1
	=Female		=Single	=Rural,Capital	=R,O	=E	=S2,S3,S4
=A2,A4	=Female	=College,Master	=Married	=Town,City		=E	=S1
=A3,A4		=College	=Married	=Capital	=O	=E	=S1,S3
=A3,A4	=Male	=Primary,Senior,Master		=Town	=P,R	=E	=S1
=A3,A5		=College,Master	=Married		=F	=E	=S4,S5
=A5		=Primary,College		=Town	=F,P,O	=E	=S4
=A3	=Male		=Single	=Capital	=P	=D,E	
=A5			=Married	=Town	=F	=B,D	=S1
=A5	=Male	=Primary,Senior	=Married		=M,P	=C,E	=S1,S5
=A3,A4	=Male	=College,Master	=Single	=Town	=E,M,R	=A,C,E	=S1
=A4,A5		=Senior,Master		=Town	=F	=A,E	=S1,S4
=A5	=Male	=Primary,Senior	=Married	=Rural,Capital	=F,P	=C,E	=S1
=A4,A5			=Single	=Town		=C,D	=S1,S2
=A5	=Male	=Senior			=M	=A	=S1
	=Male		=Single	=Rural		=D	=S2,S5
=A3	=Male		=Single	=Capital	=R	=A,D	=S1
=A2	=Female	=College	=Single	=Rural,Capital		=E	=S4
	=Male		=Single	=Town	=E,F	=C,E	=S1

4.3.2. 結果分析

規則歸納法是從案例與結果中推導出案例的屬性與結果間的關係規則，是資料探勘方法中最直接也最為一般人所瞭解之分析方式，經常應用在對龐大的資料庫中獲得資料屬性之特性規律，分層抽樣法則有助於縮減龐大資料的規則推演時間。

經由實驗所彙整的表 10、11、12、13 我們得到以下的結果：

以母體 1999 筆之樣本未抽樣進行規則尋找 CPU 所耗費的時間(約 1272 秒)為分層抽樣後(平均 3.35 秒)之 379.7 倍，若隨樣本空間之加大，相信此差距會更大，分層抽樣後之平均分類正確率為 80.21%，未分層之正確率為 81.25%，兩者差距不大，若以未分層之規則驗證分層抽樣後之驗證組資料，其正確率高達 93.54%，有幾組甚至達 100%，其原因應是經隨機抽樣後，驗證組資料量變小之原因，但若以分層抽樣後之規則驗證未分層之驗證組資料，平均正確率為 83.68%，反而比未分層之規則的 81.25%高出 2.43%，經由實驗結果得，分層抽樣與未分層抽樣之結果差距相近，然其所耗費的時間差距甚大。

分層抽樣對於縮減龐大資料的規則推演時間的確有極大的改善，且對於未分層抽樣之原始資料所做的推論結果差異不大，將來將可應用於對於需要縮短時間爭取時效之應用上。

第 5 章. 結論與建議

5.1. 結論

由於金融環境的競爭白熱化，各行庫莫不使出渾身解數，從各個構面爭取業績、盈餘的成長率，以期在市場上佔有一席之地，為達到此一目的，導入自動化流程是各行庫最積極進行的首要動作，例如自動信用評分系統、自動貸款系統、進件系統等，然而雖然是自動化但仍需人為介入，依據經驗法則逐項定義每條評分標準與規則，而使得人才的選擇顯的極為重要，造成挖角的情形經常上演。

因此一套合宜的資料探勘工具，是企業產生企業智慧重要的依據，而規則歸納法是資料探勘的領域中最常使用的方法，經這種由一連串的“如果－則”（ If － Then）之邏輯規則對資料進行細分的技術所產生的規則，可以直接轉換成自動化流程裡的規則資料庫所需要的格式，而本研究就是希望能提供一套符合此需求的機制，根據實驗之結果證實，本研究所提出的構想將可快速的達到規則獲得的目的。

5.2. 建議及未來研究方向

本研究所提出的構想僅是雛型的架構，在實際運用時如何界定規則為有效是其中的問題，如何事先將資料中發生數太少的項目先行剔除，以避免無意義的邏輯規則產生，是筆者未來研究的方向，若能再結合金控公司的優勢，整合旗下各子公司的資料庫，建立更完整的消費者消費資料，將使銀行能夠在最適當的時機提供更符合消費者需求的產品與協助，達到與消費者共同成長的雙贏局面。

參考文獻

中文部分（依作者姓名筆劃排列）

01. 中心年報，”93 年度年報”，聯合信用卡中心，民國 94 年 5 月 5 日取自：http://www.nccc.com.tw/yearbook/cover.htm。

02. 方世榮譯，Philip Kotler 著，「行銷管理學—分析、計畫、執行與控制」，東華書局，一版，民國 87 年。

03. 李紀珠，「台灣開放民營銀行設立之經驗與展望」，國改研究報告，財金（研）091-063 號，財團法人國家政策研究基金會，民國 91 年。

04. 林惠玲與陳正倉，「應用統計學」，雙葉書廊有限公司，二版，民國 91 年。

05. 林欽榮，「消費者行為」，揚智文化事業公司，初版，民國 91 年。

06. 金融機構損益概況，“金融統計”，行政院金融監督管理委員會　銀行局，民國 94 年 5 月 5 日取自：http://www.boma.gov.tw/public/data/boma/stat/index/index-9.xls。

07. 信用卡業務統計，“金融統計”，行政院金融監督管理委員會　銀行局，民國 94 年 5 月 5 日取自：http://www.boma.gov.tw/public/data/boma/stat/cc/index-7.xls。

08. 現金卡重要業務及財務資訊揭露，“金融資訊揭露”，行政院金融監督管理委員會　銀行局，民國 94 年 5 月 5 日取自：http://www.boma.gov.tw/public/Attachment/4121174371.xls。

09. 張智星，”線上教材”，“資料群聚與樣式辨認”，”2-4：模糊 C-means 分群法″，民國 94 年 5 月 5 日取自：

民國 94 年 5 月 5 日取自：

http://neural.cs.nthu.edu.tw/jang/books/dcpr/模糊 C-means 分群法.pdf。

10. 張堯庭、董麓、謝邦昌，「抽樣調查之理論及其應用方法」，中國統計出版社，民國 87 年。

11. 曾干育，「溫泉旅館遊客利益區隔之研究－以苗栗泰安地區為例」，國立朝陽科技大學休閒事業管理系碩士論文，民國 93 年。

12. 賴志東，「資料挖掘文獻，資料歸納性的研究」，陳彥良教授網站，國立中央大學館理學院，民國 94 年 5 月 11 日取自：

http://www.mgt.ncu.edu.tw/~ylchen/database/induction.doc。

13. 戴久永，「統計概念與方法」，三民書局，民國 84 年八月。

14. 戴至中譯，Mark J. Barrenechea 著，“e-Business or Out of Business”，「甲骨文革命—主宰未來電子商業的 ORACLE 模式」，麥格羅-希爾，民國 90 年。

15. 簡芃榛，「統計方法應用於消費性業務銀行顧客關係管理系統之研究」，國立成功大學統計學系碩士論文，民國 92 年。

16. 魏啟林譯，Philip Kotler 著，「行銷學精論」，華泰書店，民國 78 年。

17. 顧淑馨譯，Frederick F. Riechheld 著，「忠誠度：企業獲得利潤的基石」，智庫出版社，民國 93 年。

西文部分（依作者姓氏字母排列）

18. Bezdek, J. C., “Pattern Recognition with Fuzzy Objective Function Algorithms,” Plenum Press, New York, 1981.

19. Bhattacharyya, G. K. and Richard A. J., “Statistical Concepts and Methods,” New York: John Wiley & Sons, 1997.

20. Bloedorn, E., Wnek, J.; Michalski, R.S., and Kaufman, K., “AQ17: A Multistrategy Learning System: The Method and User’s Guide,” Report of Machine Learning and Inference Labratory, MLI-93-12, Center for AI, George Mason University, 1993.

21. Brachman, R. and Anand, T., “The process of knowledge discovery in databases: A human-centered approach. In Advances in Knowledge Discovery and Data Mining,” U.M. Fayyad, G. Piatetsky -Shapiro, P. Smyth, and R. Uthurusamy, Eds. AAAI Press/The MIT Press, Cambridge, Mass, pp. 37-57, 1996.

22. Bramer, M.A., “Induction of Classification Rules from Examples: A Critical Review,” Proceedings of Data Mining 96, London, April 1996, Unicom Conferences, pp. 140-166, 1996.

23. Chen, M.S., Han, J., and Yu, P.S., “Data Mining: An Overview from Database Perspective,” IEEE Transactions on Knowledge and Data Engineering, Vol. 8, No.6 , pp. 866-883, 1996.

24. Clark, P., “Machine learning: Techniques and recent developments,” In A. R. Mirzai, editor, Artificial Intelligence: Concepts and Applications in Engineering, pp. 65–93, Chapman and Hall, 1990.

25. Dunn, J. C., “A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters,” Journal of Cybernetics, Vol. 3, pp. 32-57, 1973.

26. Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P., “From Data Mining to Knowledge Discovery: An Overview,” In Advances in Knowledge Discovery and

Data Mining, Chapter 1, pp. 1-34, AAAI Press and the MIT Press, 1996.

27. Hegland M., “Data Mining – Challenges, Models, Methods and Algorithms,” Publications of ANU Data Mining group, Draft, 2003.

28. Kaufman, K. A. and Michalski, R. S., “The AQ18 Machine Learning and Data Mining System: An Implementation and User’s Guide,” Reports of the Machine Learning Laboratory, George Mason University, 2000.

29. Kusiak, A.,“Feature Transformation Methods in Data Mining,” IEEE Transactions on Electronics Packaging Manufacturing, Vol. 24, No. 3, pp. 214-221, 2001.

30. Langley, P., “Elements of machine learning,” San Francisco: Morgan Kaufman, 1996.

31. Machine Learning and Inference Laboratory, iAQ, Retrieved May 27, 2005 from the World Wide Web: http://www.mli.gmu.edu/msoftware.html.

32. Makhoul, J., Rpucos, S., and Gish, H., “Vector quantization in speech coding,” Proceedings of IEEE, 73, pp.1551 - 1558., 1985.

33. Michalski R.S., Mozetic, I., Hong, J. and Lavrac, N., “The Multi-Purpose Incremental Learning System aq15 and Its Testing Application to Three Medical Domains,” Proc. Fifth Nat'l Conf. Artificial Intelligence, pp. 1041-1045, 1986.

34. Olaru, C. and Wehenkel, L., “Data Mining,” IEEE Computer Applications in Power, Vol. 12, no. 3, pp. 19-25, 1999.

35. Schaeffer, R.L., Mendenhall, W., Ott, L., “Elementary Survey Sampling,” Duxbury Press, A Division of Wadsworth Publishing Company, pp. 76-82, 1995.

36. William J. Frawley, Piatetsky-Shapiro, G. and Christopher J. Matheus, “Knowledge Discovery in Databases-An Overview,” AI Magazine, pp. 57-70, 1992.

37. Wnek, J. and Michalski, R.S., “Hypothesis-driven Constructive Induction in AQ17-HCI: A Method and Experiments,” Machine Learning, Springer Science +Business Media B.V., Vol. 14, No. 2, pp. 139-168, 1993.

38. Wojtusiak, J., "AQ21 User's Guide," Reports of the Machine Learning and Inference Laboratory, MLI 04-3, George Mason University, Fairfax, 2004.