運用二階段分類技術挖掘潛在中小企業借貸戶之研究

李御璽 顏秀珍
銘傳大學資訊工程學系
leeys@mail.mcu.edu.tw

丁明勇 郭家禎
銘傳大學資訊管理學系
chaichen29@gmail.com

趙家宏
銘傳大學電子工程學系
chiahong@mail.mcu.edu.tw

摘要

自政府實施金融自由化政策後，國內的銀行業面臨了激烈的競爭與挑戰。銀行的營收以貸款業務為主要來源之一。本研究目標鎖定中小企業(Small Medium Enterprise, 簡稱SME)，找出有資金需求的中小企業借貸戶之特徵，運用資料探勘(Data Mining)分類(Classification)技術，建立二階段的分類模型。由於資料集為不平衡(Unbalanced)的資料集，因此將使用抽樣(Sampling)的方式以平衡訓練資料集，並使用不同的分類技術來提高其模型預測少數類別的效能。

關鍵詞：中小企業借貸戶、分類模型、不平衡資料集

Abstract

After the government's financial liberalization policy, banks have confronted the severe competition and challenge. One of the bank's revenue is unsecured loan. Therefore, the aim of this study focuses on Small Medium Enterprise (SME) which has the demand feature of SME loan. The classification techniques of data mining are used to build the two-stage model. Because the SME dataset is unbalanced, a sampling technique is also proposed to make a balanced training dataset. Moreover, the study uses the different classification algorithms in each stage and enhances the predictive power on minority class.

Keywords: Small Medium Enterprise, Classification, Unbalanced

1. 前言

1990 年代起，政府實施一連串的金融自由化政策，開放民營銀行設立，公營銀行民營化，又通過信用合作社改制為商業銀行，因此台灣在短期間增加了很多銀行，帶動了二十多年來國內銀行之間的激烈競爭，使得金融機構面臨嚴酷的挑戰，各家銀行莫不絞盡腦汁提供多元化的服務，在市場中取得一席佔有之地，並與顧客達成良好的互動關係，進而提高顧客價值、提升產業競爭力。銀行內部擁有多個龐大資料量的資料庫，如何運用其資訊分析客戶消費行為，發掘客戶需求，尋找潛在的成交客戶進行行銷，成為決策者做決策時的重要參考依據，為目前各金融機構經營的首要目標。

1.1 銀行業概述

自政府積極採取一連串的金融改革措施之後，逐漸解除各項金融管制，開放民營銀行的設立，放寬銀行設立分行的限制，在眾多優勢的扶植之下，使得台灣金融體系邁向更自由化與國際化，截至 2011 年 9 月，根據中央銀行「金融統計月報」指出，如圖 1 台灣銀行總機構及其分行家數高達 3413 家[1]，由此可見，銀行業競爭是相當激烈。

其中，銀行的營收的主要來源為利息收入，以台灣銀行為例，其中利息收入所佔的營收比例為最大，約為 39.4%[2]，而其又是來自於各種不同型態之授信貸款業務，主要分為如下：抵押貸款(Secured Loan)，指向金融機構申請貸款時，將特定的資產抵押作為還款的保證，已獲得融通的資金，常見的抵押品如房屋地產等，銀行所承受的風險較為低，因此貸款的額度較高、利率較低；無抵押貸款(Unsecured Loan)，又稱為信用貸款(Fiduciary Loan)，即為

借款人不需提供擔保而發放的貸款，如現金卡(Cash card)、信用卡循環(Credit Card Revolving)、二胎房貸(2nd Mortgage)、二胎車貸(2nd Car Loan)等，因此無抵押貸款則以是借款人之信用來借貸，銀行所承受的風險相對來的高，因此貸款的額度較低，利率則提高。其中又可將貸款對象分為個人戶與法人戶(即為企業公司戶)，其對象若推行至中小企業(Small Medium Enterprise,簡稱 SEM)公司戶，則除了信貸本身利率高外，貸款金額也相較於個人戶提高許多，故可為銀行獲取更高的利潤。

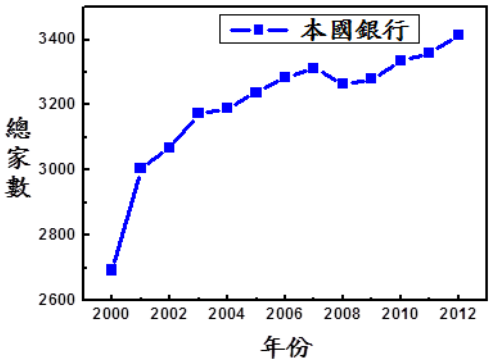

圖 1 本國銀行總家數
資料來源：中央銀行, 2011 年 9 月

1.2 中小企業借貸

台灣中小企業擁有技術創新、分工細緻、較小的金融風險等優點，以及它們所創造出的龐大國民財富和就業機會，是台灣經濟發展的基石。然而，其經營者在創業的過程，大多是靠技術起家，其經營基礎較為薄弱，不易凝聚資金，容易受到金融環境的影響，若非經營者本身有充裕的資金，在企業營運時，常會面臨到應收帳款與應付帳款之間等不同的資金需求，對於經營者是一大考驗。

所以說，中小企業在發展過程中，最需要的就是資金，若能夠找出這些有資金需求的中小企業，將可鞏固台灣經濟發展，也可增加銀行之收益。本研究目的為利用資料探勘技術，從銀行中小企業戶的資料庫中，建立一個中小企業借貸戶的分類預測模型，並提高其模型之類分類校能與發掘目標顧客的需求因素，藉此協助管理者制定相關的決策。

2. 相關研究

2.1 資料探勘定義

Fayyad and U.M. 指出資料探勘又稱為資料庫知識探索 (Knowledge discovery in database, KDD) [6]，將資料探勘是為整個知識發現過程中的一個必要的階段，也就是說資料探勘是知識發掘過程當中的一個環節，一是一個核心的步驟，其流程步驟為：

1.資料收集階段 (Selection)：先理解要應用的領域、熟悉相關知識，接著收集原始資料，而原始資料的來源很多，資料庫系統就是主要的資料收集工具，如日常進行的交易紀錄的資料庫。

2.前置處理階段(Preprocessing)：資料探勘牽涉了大量的準備工作與規劃過程，其中包含了去除資料錯誤與不一致的問題。

3.資料簡化與轉換階段(Transformation)：將此資料集轉換成為資料探勘程序中可以接受的資料型態，這是基於的資料探勘技術，會有不同之輸入資料型態。

4. 資料探勘階段(Data Mining)：此階段是知識發掘的一個核心，在資料集中尋找一些樣式與關聯。

5. 模型評估階段(Pattern Evaluation)：資料探勘萃取的知識不一定每個都是我們需要的，有些是有意義、有些是無意義，必須再經由一個樣式評估，目的為評估所找出來的是否有價值，過濾一些沒用的資訊，最後再將結果給與使用者去應用。

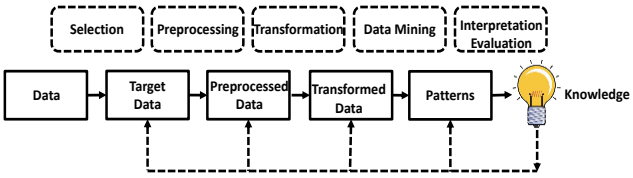

圖 2 KDD 流程圖
資料來源：Fayyad U.M. (1996)

這些程序是一個循環的關係，一直重複的步驟，最後才得到有用的知識，所以說資料庫知識探索是一連串的程序，資料探勘只是其中的一個步驟而已。

各學者專家因研究之角度不同，對其亦有不同之詮釋(Frawley et al. 1992, Grupe and Owrang 1995, Chen et al. 1996, Berry and Linoff

1997, Han and Kamber 2001, Turban et al. 2007)，綜觀諸位學者提出的定義，本研究將資料探勘定義為是一種支援企業策略的技術，從大量的資料中透過統計與機器學習等方式，發現出先前未知的知識過程，並轉換成有用的資訊與知識，輔助企業做出重要之決策。

2.2 資料探勘應用軟體之建構流程方法論

為了更有系統化及標準化進行探勘的流程，業者與軟體開發業者提出了一些資料探勘過程的參考模型或標準，來幫助其使用者能達到目標，如最具代表性之一的跨產業資料探勘標準作業程序(Cross Industry Standard Process for Data Mining, CRISP-DM)，其資料探勘作業程序主要是由 SPSS Inc.等多家公司在 1996 年聯合發展而成。另一為 SAS 公司提出了 SEMMA(Sample-Explore-Modify-Model-Assess)，強調結合其工具—Enterprise Miner 中的應用方法，代表資料探勘的五個步驟，如圖 3 說明如下：

1. 資料抽樣(Sample)：資料經過適當條件的篩選與過濾，可減少處理的資料量，進而節省系統資源與增進處理效率，而且經過資料篩選，可以突顯資料的規律性。

2. 資料探索(Explore)：資料探索就是一般進行深入調查的過程，當分析人員拿到了一個樣本資料集後，會進行以下特徵探索：資料是否達到原來設想的要求、其中有沒有什麼明顯的規律和趨勢、有沒有出現從未設想過的資料狀態、變數之間有什麼相關性它們可區分成怎樣一些類別，達到瞭解各個變數之間的複雜關係之目的。

3. 資料調整(Modify)：透過上述兩個步驟的操作，對資料的狀態和趨勢可能有了進一步的瞭解，對企業原來要解決的問題，可能也有更明確的想法。因針對問題的需要，可能要對資料進行增刪，也可能按照對整個資料探勘過程的新認識，要整合或者增加一些新的變數，以對現實狀態進行有效的描述。

4. 建立模型(Model)：是資料探勘工作的核心環節。當分析人員進行到這一步時，對應採用的技術已有了較明確的方向，資料結構和內容也有了充分的適應性。這時就可以運用統計方法，例如類神經網路、時間序列分析、決策樹模型和標準統計方法(集群分析、判別分析、羅吉斯迴歸、一般線性模型等)，來建立資料模型和發現知識。

5. 評價(Assess)：從上述過程中將會得出一系列的分析結果、模式或模型。若能得出一個直接的結論當然很好，但更多的時候，會得到對目標問題側面的描述，這時就要綜合它們的結果，提供合理的決策支援資訊。

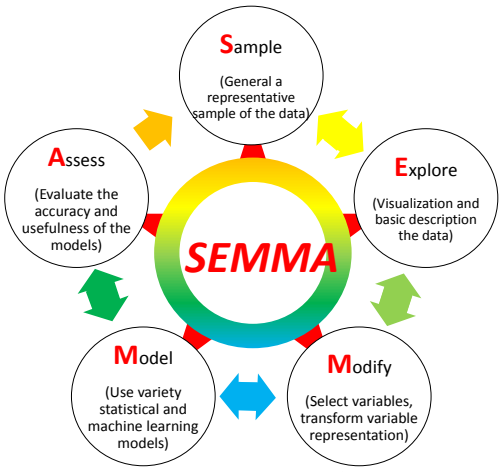

圖 3 SEMMA 示意圖

2.3 資料探勘的模式

資料探勘應其模式面可分為七種[7]：

1. 關聯規則(Association)：此技術是用來辨識歷史資料庫中找出哪些事物總是一併發生或者是哪些商品總是一起被購買，找出事物間隱藏性的關係。

2. 序列發現(Sequence discovery)：可進一步找出事物發生的時間先後順序。

3. 分群(Clustering)：分析前並不知道會以何種方式或根據來分類，分群技術可自動地依其相似性，將相似的事物分群。

4. 視覺化(Visualization)：利用視覺化的方式，簡單描述在複雜的資料庫中發生了什麼事，並將分析與淬取的結果呈現出來，以解釋複雜或繁瑣的內容。

5. 迴歸(Regression)：迴歸是使用一系列的現有數值，預測一個連續數值的可能值。

6. 預測(Forecasting)：預測分析主要是用來預測「連續變數」，利用過去的歷史數值來找出未來連續數值的變化狀況。

7. 分類(Classification)：針對欲處理且未分類的資料集合，根據已知類別(Class)的物件集合，將欲處理資料依據其屬性(Attributes)去完成分類的過程，並冀望能學習分類的規則，提供未來能自動分類之用。

2.4 資料探勘之分類技術

在資料探勘領域中，分類技術是受到相當廣泛的研究，並應用在各種領域中。對於分類模型而言，其分類能力的好壞對於知識發現是很重要的，可影響其分類結果效能之因素有三(如圖 4)：資料集、輸入屬性、分類演算法。

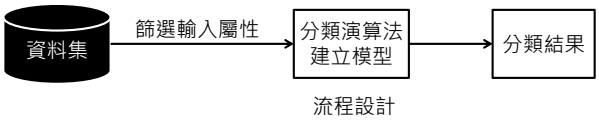

圖 4 資料探勘分類模型

2.4.1 資料集

在進行資料探勘前，通常可能面臨以下情形，資料集中的某幾筆資料帶有空值(Null Value)、錯誤值(Wrong value)、離群值(Outlier)等，會造成無法訓練並建立預測模型，即使將模型建立，也會是一個無效的模型，因此需先經過資料的前置處理(Pre-processing)，高品質的資料，才會有高品質的結果[8]。

除此之外，在處理分類問題時，經常假設訓練集資料是均勻或接近均勻地分布在不同目標類別(Target Class)。當資料集中的目標類別分布越均勻時，分類模型會有比較好的效能，但在許多實際應用的例子中可以發現經常為目標類別分佈不平衡的問題(Imbalanced Class Distribution Problem)，將會呈現多數類別(Majority Class)與極少數的資料為少數類別(Minority Class)，這時訓練的分類模型，為了提高整體資料的分類準確性，導致忽略了少數類別，會傾向於預測所有測試資料為多數類別，但是找這些少數類別才是決策者最重視的部分。先前已有許多研究提出解決不平衡資料集的問題，其中最常使用的方式為抽樣技術(Sample)，目的為試圖拉近多數類別與少數類別的比值，可分為兩種：減少多數類別抽樣法(Under Sampling)、增加少數類別抽樣法(Over Sampling)。一般而言，增加少數抽樣法的效能比減少多數抽樣法的效能還差[9]。

2.4.2 重要屬性

在建立分類模型前，必須計算屬性變數的重要程度，如此才能篩選重要且有鑑別能力的屬性變數作為分類模型的輸入，若將不重要的屬性輸入至分類模型中，將會導致模型增長分類的時間，更甚者可能影響分類模型的預測能力。

2.4.3 演算法

演算法的重點為選用的演算法是否適切，以及所使用的參數是否設定正確，常見的分類演算法有以下：

1. 決策樹(Decision Tree)：在資料探勘的領域中，決策樹被認為是一種樹狀結構的規則，樹的中間點(no-leaf nodes)代表測試的條件，樹的分支(branch)代表測試的結果，而樹的葉節點(leaf nodes)代表分類後所得到的分類標記，也就是表示分類的結果。
2. 類神經網路(Neural Network)：是模仿生物神經網路的運作方式，由一些高度連結的處理單元(節點或稱作神經元，neuron)，組成一動態的運算系統，類神經網路會透過不斷地自我調整，使得輸入的資訊經過神經元的運算後，能得到預設的輸出結果。
3. 邏輯斯迴歸(logistic regression)：欲利用 2 個或 2 個以上的自變項建立迴歸模式以估計每位受試者在某個事件所發生的機率。

2.4.4 相關文獻

最基本的模型建立方式為將經過處理後的資料集與屬性，直接投入分類演算法建立分類模型，已有學者發表相關研究如下：

黃怡華[3]：以某銀行消費性貸款授信戶為研究對象，應用分類法中的類神經網路與統計區別分析的方法，分別建置個人消費性貸款信用評等之分類模式，以區分銀行顧客之信用類別，實驗中發現，類神經網路在判斷不良顧客有較佳的準確率。接著再將此一分類結果，使用關聯規則去描述兩類別之顧客，使銀行能更詳盡的了解顧客特性：信用不良顧客，可提供銀行相關特徵及資訊，以便進行監控及防範之工作；信用良好的顧客之特徵及資訊則可以作為行銷規劃及相關金融商品設計之依據。

陳東和和黃謙順[4]：從現行龐大之基金交易資料集中，先使用 K-means 分群技術找出三種基金顧客的風險偏好類型(保險型、穩健型、積極型)，找出此三群顧客後，設為分類模型的目標變數，接著透過個人基本資料，以此建立

客戶之風險承受類別分類預測模型，將可應用於未知顧客，投入基本資料即便可預測出為哪一類型顧客，為顧客推薦適合其風險屬性之基金。

本研究認為此種分類模型，只訓練一次的模型，這當中必定還有資訊尚未被挖掘，因此，將會朝向建立二階段的分類模型的方式，並搭配使用不同演算法，希望藉由不同演算法的不同特點來提升模型之效能。

李御璽等人[5]使用資料探勘二階段分類技術模型，找出潛在有資金需求的中小借貸戶(已有貸款的為 G 公司戶，尚未來貸款的為 B 公司戶)。實驗的步驟為：先用整體資料去計算每個屬性的屬性重要程度，藉由門檻值選擇要進入分類預測模型的屬性，由於此資料集為不平衡資料集，也就是說資料集中的目標屬性已有貸款的G公司戶與尚未來貸款的B公司比例有偏差，所以使用減少多數類別的隨機抽樣方式(G 公司戶與 B 公司戶個數比為 1：1)，利用 SPSS Clementine 的 C5.0、C&RT、Logistic Regression 及 Neural Network 四種演算法，重複訓練五組模型後，將整體資料做為測試資料，選出訓練與測試模型正確率差距最小，且訓練模型正確率較高的作為第一階段的最佳模型。

接著進行第二階段，將第一階段的最佳模型中，預測為多數類別的資料當成訓練資料，重新訓練模型，以相同的方式計算重要屬性，隨機抽樣訓練五組模型，選出最佳模型。此研究鎖定的資金需求的中小企業即為第一階段實際為B公司戶但預測是G公司戶的客戶和第二階段實際為 B 公司戶但預測是 G 公司戶的客戶。

此篇研究提出了二階段的分類預測模型，但由於無法驗證結果是否真為資金需求的中小企業，即為無法鑑定此實驗流程所提升的效能。且不應以全體資料來當作測試資料，再加上訓練模型前，應預先將資料集分為訓練資料與測試資料，再經由訓練資料去計算重要性，選擇輸入模型的屬性，才可由測試資料得知其模型的效能。故本研究將針對此篇不足的方面進行修改。

3. 研究方法

本研究將依循 KDD 資料庫知識探索及 SEMMA，並採用 SAS Enterprise Miner 分析工具來完成研究目的，將建立兩階段分類模型。

研究流程為：步驟一，將取得的資料集隨機分為 80%訓練資料集，用來訓練分類模型；20%測試資料集，用來當作為之的新資料驗證模型。步驟二，計算並篩選訓練資料中的各變數重要程度。步驟三，使用分類演算法來建立第一階段分類模型。步驟四，擷取第一階段分類模型中，模型預測為未貸款顧客之資料作為第二階段模型訓練的資料。步驟五，計算並篩選訓練資料中的各變數重要程度。步驟六，使用與第一階段不同的演算法來建立第二階段的分類模型。步驟七，將兩階段的結果結合，即為二階段分類模型之實驗結果。步驟八，將測試資料投入二階段模型，可用來驗證本研究之方法是否可行。

3.1 資料來源

本研究資料來源為國內某銀行之中小企業公司戶的資料，資料集中總共有 32,681 筆中小企業公司戶的資料。其中，738 筆為已有來借貸之 SME 公司戶(簡稱 1)，31,943 筆為未來借貸之 SME 公司戶(簡稱 0)。0 與 1 的比例為 2.26%:97.74%。

3.2 欄位說明

1. 類別

表 1 類別屬性說明

屬性	欄位說明
ck	是否有支票存款(支存)
area	所在地區(地區代碼)
comp	是否為公司週邊法人戶

2. 數值

表 2 數值屬性說明

屬性	欄位說明
ck-saveall	數值支存存入總金額(一年)
ck-drawall	數值支存提領總金額(一年)
ck-savetime	數值支存存入總次數(一年)
ck-drawtime	數值支存提領總次數(一年)
ck-saveavg	數值支存平均每次存入金額(一年)
ck-drawavg	數值支存平均每次提領金額(一年)
ck-avg	數值支存平均餘額（半年）
dep-saveall	數值活期存入總金額(一年)
dep-drawall	數值活期提領總金額(一年)
dep-savetime	數值活期存入總次數(一年)

dep-drawtime	數值活期提領總次數(一年)
dep-saveavg	數值活期平均每次存入金額(一年)
dep-drawavg	數值活期平均每次提領金額(一年)
dep-avg	數值活期平均餘額（半年）
dep-9201	數值 92 年一月存入活期金額
fed-9201	數值 92 年一月外幣月底餘額
fed-avg	數值外幣平均餘額(一年)
ck-changame	數值支存交換票總金額(一年)（實際兌現的）
dep-changame	數值活存交換票總金額(一年)（實際兌現的）
ck-changtime	數值支存交換票總次數(一年)（實際兌現的）
dep-changtime	數值活存交換票總次數(一年)（實際兌現的）

3.3 資料前處理

由於中小企業借貸戶資料集中，類別屬性有空值的情形產生，因此本研究必須對有空值的屬性進行分析處理。根據本研究的分析結果，類別屬性地區(area)的欄位有 40 筆資料為空值。由於本研究沒有取得企業公司所在地與銀行間的距離屬性，因此在訓練及測試過程中，本研究排除地區(area)這個欄位。此外，另一個不納入考慮的欄位為是否為公司週邊法人戶(comp)，主要原因是有 88%的資料都不是公司。

3.4 計算重要度

資料集中的屬性分為類別屬性與數值屬性，我們將分開各別計算類別與數值的重要程度，計算方法如下：

1. 數值屬性

對於數值類別的輸入變數，本研究將採用下列的公式來計算前一節所提及的平均差與標準差以取得數值屬性欄位的重要度指數。

表 3 數值屬性之平均值與標準差表

	0	1
平均值	A	B
標準差	C	D

2. 類別屬性

對於類別屬性則是先行計算其各個屬性所出現的頻繁度（Frequency；F）、支持度（Support；S）以及信賴度（Confidence；C）後，再計算其屬性的重要程度，並將各類別屬性之子欄位的重要程度給予彙總。公式計算式子如下所示，其中公式中的 Cm 代表 Class m，An 代表屬性值 n。

Cm.An 的支持度：

$$S(Cm\_An) = \frac{F(Cm\_An)}{F(Cm)} \quad (1)$$

Cm.An 的信賴度：

$$C(Cm.An) = \frac{S(Cm.An)}{(S(C1.An)\,S(C2.An))} \quad (2)$$

屬性值 An 的重要性：

$$An = \frac{(|(C(C1.An) - C(C2.An))|)}{(F(C1) + F(C2))} * \frac{((F(C1.An) + (F(C1.An))}{(F(C1) + F(C2))} \quad (3)$$

3.5 模型效能評估

各模型的結果將依據下表混亂矩陣的各項公式，以計算取得個模型的實際產出的數據，以供比較。

表 4 混亂矩陣

	預測為 0	預測為 1
實際為 0	A	B
實際為 1	C	D

0 的預測準確率（Precision）：

$$P1 = \frac{B}{(B + D)} \quad (4)$$

0 的預測捕捉率（Recall）：

$$R1 = \frac{D}{(C + D)} \quad (5)$$

0 的 F-Measure 指標：

$$F1 = \frac{(2 * P1 * R1)}{(P1 + R1)} \quad (6)$$

3.6 實驗結果

本研究建立二階段分類模型，第一階段使用決策樹，結果混亂矩陣如下：

表 5 第一階段使用決策樹之混亂矩陣

	預測為 0	預測為 1
實際為 0	24640	913
實際為 1	267	323

模型預測準確率 Precision = 0.2613，捕捉率 Recall = 0.5475，F-Measure 指標 = 0.3538

第二階段使用類神經網路，結果混亂矩陣如下：

表 6 第二階段使用類神經網路之混亂矩陣

	預測為 0	預測為 1
實際為 0	24413	227
實際為 1	198	69

將兩階段實驗數據結合，即為二階段分類模型之結果，結果混亂矩陣如下：

表 7 兩階段實驗數據結合之混亂矩陣

	預測為 0	預測為 1
實際為 0	24413	1140
實際為 1	198	392

實驗之總結果為模型預測準確 Precision = 0.2559，捕捉率 Recall = 0.6644，F-Measure 指標 = 0.3695。

由此可得知，在進行一階段分類模型時，F-Measure 指標為 0.3538，而進行二階段分類模型後，F-Measure 指標提升至 0.3695。

4. 結論與未來工作

以資料探勘的分類技術，利用中小企業借貸戶之資料集，建立模型並提高其分類效能，藉此了解中小企業之借貸需求，銀行可更易於管理顧客，以達到增進銀行營收與降低成本的支出。

對於一般的資料探勘分類模式來說，建立一次的分類模型不夠精準，此研究將會使用兩階段的分類模型，試圖將第一次未分類完成的資料，以不同的分類技術進行第二次的分類模型，以達到提升整體的分類效能。從實驗的結果中，可發現進行二階段分類模型後，確實是可以提升 F-Measure 指標，未來工作將再嘗試以其他演算法來交互驗證之。

參考文獻

[1] 中央銀行，金融統計月報，民國 100 年 9 月。.

[2] 張明道，我國銀行業的機會與挑戰，民國 101 年。

[3] 黃怡華，”應用類神經網路與關聯法則於銀行消費性貸款”，國立成功大學資管所碩士論文，民國 93 年 6 月。

[4] 陳東和、黃謙順，”運用資料採礦技術於銀行基金客戶分群之研究”，**知識社群與系統發展研討會**，民國 97 年。

[5] 李御璽、顏秀珍、張韋豪、林基玄、鄭郁翰、楊乃樺、賴郁菁、廖晨涵，「運用分類技術發掘潛在中小企業借貸戶之研究」，**Journal of Information Technology and Applications**，第 1 卷，第 3 期，民國 96 年 12 月，頁 173~181。

[6] Fayyad, U. M. 1996, “Data Mining and Knowledge Discovery: Making Sense Out of Data,” ***IEEE Expert*** (11:5), pp.20-25.

[7] E.W.T. Ngai, Li Xiu, D.C.K. Chau 2009, “Application of data mining techniques in customer relationship management: A literature review and classification,” ***Expert Systems with Applications*** (36: 2), pp.2592-2602.

[8] Sven F. Crone, Stefan Lessmann, Robert Stahlbock 2006, “The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing,” ***European Journal of Operational Research***(173:3), pp. 781-800.

[9] Show-Jane Yen, Yue-Shi Lee 2009, “Cluster-based under-sampling approaches for imbalanced data distributions,” ***Expert Systems with Applications***(36:3), pp.5718–5727.