以支撐向量法建構之自動化餘額代償信用評價模式

江坤林
國立台灣科技大學
資訊工程系
martystudio@gmail.com

簡立仁
國立台灣科技大學
資訊工程系
ljchien@niu.edu.tw

李育杰
國立台灣科技大學
資訊工程系
yuh-jye@mail.ntust.edu.tw

摘要

餘額代償是一種高收益但也高風險的金融商品，申請人過去都曾與其他銀行密切往來過一段期間，故已累積許多與銀行互動的紀錄於財團法人聯合徵信中心的資料庫中，一般銀行在進行代償徵審的工作上亦相當重視此紀錄。我們應用機器學習的技術來擷取正常戶與逾放戶之特徵資訊，佐助人工徵審以降低逾期放款的比率。目前在機器學習的範疇中，支撐向量法(Support Vector Machines)在分類問題實務上有優秀的表現。本研究旨在應用支撐向量法，以過去核貸客戶在聯徵中心所累積之歷史信用資料及其逾放與否作為訓練的樣本，透過機器學習的方法來建構自動化之餘額代償信用評價模式。經由北部某銀行提供之資料實證後，本模型確實可減銀行之損失。

關鍵詞：餘額代償、分類問題、機器學習、電子核貸、支撐向量法。

1 前言

近年來金融業的競爭愈形激烈，在高獲利因素吸引下，各家行庫紛紛推出各種形態的信用貸款，在電子媒體的促銷及廣告明星的代言下，高利率的信用貸款被包裝成圓夢的最佳工具，造成特殊需求人口負債上升。因之而生一種可降低利息支出的貸款商品-「餘額代償信用貸款」，逐漸在金融界受到重視，光以信用卡之循環信用餘額做為代償的目標市場，就約有 4 千多億[1]的潛在商機。餘額代償的商品特性為以中高利率(9~15%)之信用貸款來代償高利率(15~20%)之信用貸款，希望可以爭取到債信正常但想減輕貸款負擔之「好客戶」。若能建立一套好的分類機制，將風險控制在某一程度之下，餘額代償對銀行而言可以達到最高之邊際效益。因此，如何挑選「好客戶」就成為銀行徵審上的重要課題。

餘額代償在銀行的授信實務上，需要申請人提供相關之個人資料，並由銀行端在客戶授權下利用身分證字號，向「財團法人聯合徵信中心,JCIC」(以下簡稱「聯徵中心」)來查詢其過去在所有金融機構之繳款紀錄及其他信用相關資訊來作徵審。在申請人所提供之個人信用資料部份，均需提出相關證明文件，再由徵信人員以電話或面洽之方式來查證，過程相當耗時。而在聯徵中心所提供之資料方面，依政府規定從一般的貸款餘額到即時更新的現金卡繳款紀錄等，均需主動由借款銀行向聯徵中心通報，故聯徵中心所提供之資料可真實地反應申請人過去之信用狀況，且較具參考價值。根據北部某金融機構的統計，在成本的考量下，平均一份申請餘額代償的聯徵查詢報告至少會包含五種項目，約十頁左右之資料。雖然資料很豐富、數據很多，但在一般人工徵信實務上，卻僅會用到幾個經驗上認為較重點的數據，來評定申請人之信用，因此許多可能對徵審有幫助的資料就會被忽略。

目前已有很多銀行導入電子核貸系統(eLoan System)來輔助徵審人員辦理授信業務，除了可統一審核之標準，降低人員審核差異，更可提升核貸之品質。一般常見的核心技術不外乎使用決策樹(Decision Tree)、類神經網路(Neural Network)等機器學習演算法。近年來在機器學習及資料探勘的領域中，支撐向量法(Support Vector Machines，以下簡稱 SVM[13])對於分類問題(Classification Problem)提供最佳化的處理技術，且在目前的應用上都有極佳的表現。

本研究之目的希望結合具有最佳化效能的 SVM 與申請餘額代償客戶過去在聯徵中心所累積豐富的信用紀錄，在不使用申請人所提供之信用資料(可能需要進行徵信)下，建構一套餘額代償信用評價模式，提供貸放與否之建議給徵審人員參考，以降低銀行業者的授信風險、提高經營的績效。

本論文將在下一節提出一套餘額代償信用評價模式，並在第 3 節中對本研究之相關研究方法做介紹。在第 4 節則會以一系列之實驗來評量所建立的評價模式，最後在第 5 節作整篇論文的結論。

2 餘額代償信用評價模式

餘額代償為一種消費性貸款，由於消費性貸款之種類差異，並無統一之模式可用來分析借款人的信用，銀行徵授信人員除了依照該行的授信政策(lending policy)來對申請人進行條件之徵選外，更需透過以往累積之授信經驗來評估、預測借款者未來的履約能力，如一般常見的方式如 5P、5C。針對餘額代償之貸款形態，我們提出一套「餘額代償信用評價模式」，並以北部某金融機構之實際運作模

[1] 9406 之全台循環信用餘額為 473,539,271(千元)，資料來源：行政院金融監督管理委員會銀行局，http://www.boma.com.tw

式，設計一套電子化核貸系統。

餘額代償服務之申請人，一般擔負有數家銀行之信用貸款或信用卡卡債。通常代表其過去已經通過數家銀行之徵審且准予貸放，推論其轉貸時之個人狀況：

1. 由於本身之信用條件較差(如高風險行業或無固定收入者)，致貸款之利率偏高，但債信還正常，希望減輕貸款之利息者。
2. 同前項之條件，但可能因個人理財不慎、現金流量吃緊，到達快要無法正常履約之情況，轉貸後可能在六個月內就有無法履約之虞。

上述第 1 種客戶為銀行想爭取轉貸之好客戶，第 2 種則為銀行要在申請時就應拒絕之壞客戶。單從其個人提供之資料而言，在實務上對於邊際客戶之徵審難度較高。若能應用較詳實、完整的信用歷史紀錄(如聯徵中心之紀錄)與核貸銀行過去所累積之貸款經驗，建立一個申請人之信用行為預測模型(Model)，來描述可能會發生逾放客戶之特性，應可降低對第 2 種客戶核貸的比例。所以本論文提出之「餘額代償信用評價模式」將建構在聯徵中心之查詢結果上，以申請人在申貸時聯徵中心所提供之資料做為分類之依據，來進行分類。

由於聯徵中心提供之個人資料相當多，勢必要將銀行端與聯徵中心端之系統做連結，才能直接運用查詢結果。聯徵中心目前提供銀行會員機構之信用資訊產品包含八大類一百餘項，查詢的介面有四種：

1. 終端機型式(Terminal Type)：透過 IBM 終端機以撥接方式連線查詢，缺點是無法儲存資料，且只有分項查詢功能。
2. 財金資訊公司通道(FISC Gateway)：是透過主機與財金資訊公司之既有網路連線查詢，可儲存資料，但每筆查詢需另支付財金公司費用。
3. 網際網路(Internet)：是取得聯徵中心之帳號並使用 IC 晶片卡，直接於網頁介面來查詢，並將查詢結果透過加密之網頁回覆。一般以此方式查詢多以列印書面報告來處理，若要將結果轉換成電子媒體則需使用其他技術將結果轉換儲存才得以處理運用。
4. MQ：是透過 Message Queue 系統／MQ 直接與聯徵中心連線查詢，可將取得之資料加以整合分析，但前提需投入較大之建置成本。

而該金融機構目前是透過上述第三種方式查詢，故為配合資訊之取得，我們利用 PERL 開發一套聯徵查詢後端儲存系統 - JCIC Report Parser, RP，聯徵查詢的之純文字檔案轉存至內部之資料庫中，以利後續申請人樣本參數之建立，並可避免人工建立資料之錯誤。此功能運作類似 MQ，但省卻了部份系統建置成本。

本研究提出之系統架構圖如圖 1，其中 eGrade System 為自動化之信用評價系統，該系統包含以下幾個子系統：

1. JCIC Report Parser , RP：自動蒐集並拆解從網際網路查詢聯徵中心資料得來之純文字信用資料至銀行端之個人信用資料庫(Inst DB)中，由於聯徵中心提供之信用報告格式會不定期作修改，故本項子系統需不定期的視修改程度作維護。
2. Model Builder , MB：定期或不定期從 Inst DB 獲得之個人信用資料，並以一固定之貸放期間後是否發生逾期放款之結果做為資料之標籤(Label)，成為系統之訓練樣本，並進行機器學習作業，以建立分類之模型(Model)。
3. Grade Evaluator , GE：利用預先產生之模型，將查詢自聯徵中心之資料做測試，輸出信用評等等級，以提供審核人員核貸與否之參考。

樣本在經過一段時間之累積後，可以對模型重新調整、訓練，以符合當時的貸放環境。在訓練樣本之挑選上避免使用核貸過久且與目前金融環境差異過大之樣本，減少模型之分類偏差(bias)，以反應實際的市場現況。

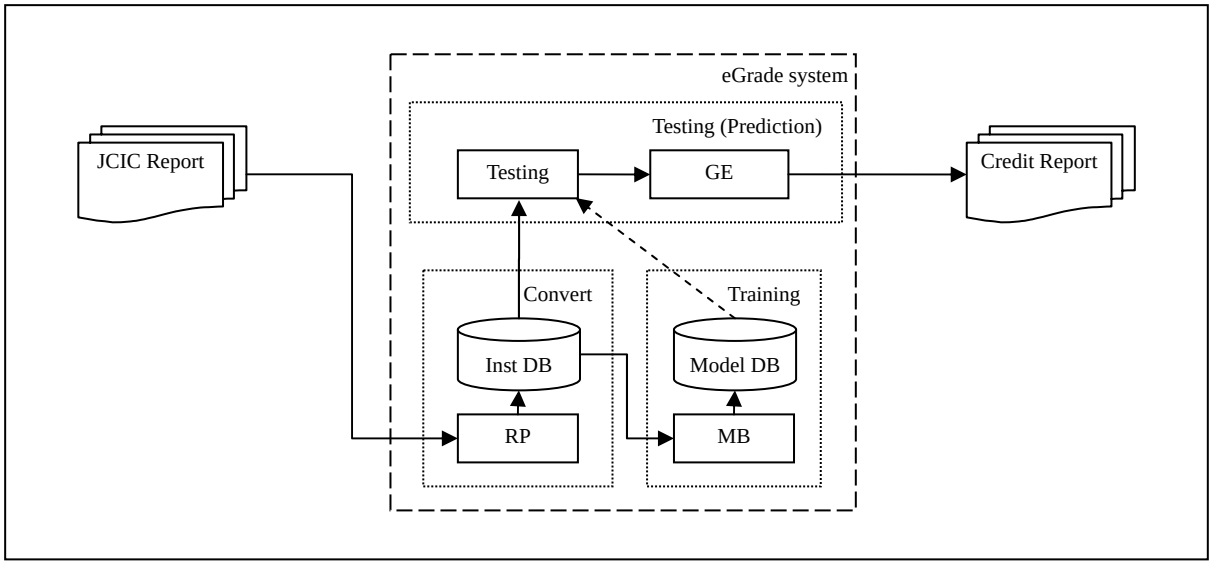


圖 1 餘額代償授信評價系統

故整體而言，本系統之特色為完全使用聯徵中心提供之歷史紀錄，在不需申請人提供資料的前提下，就可完成個人信用條件之初步審查。但此特色亦可能為本系統之限制，對於辦理非「餘額代償」之相關授信業務，可能無法直接套用，除了客戶群的差異之外，也可能因為申請人從未向金融機構借款，故在聯徵中心將查無任何還款紀錄。在銀行端的後續作業實務上，仍需參酌申請人提供之書面資料來進行徵審，這部份與銀行之貸放條件及授信政策有關。由於本論文旨在討論聯徵中心歷史紀錄對於客戶分類之影響，故對於是否可完全排除申請人提供之書面資料，則不在本論文討論範圍之內，但本系統亦可延伸增加對授信政策及條件之判斷。

3 研究方法

由於餘額代償之風險較一般消費性貸款商品高，所以在建立分類的模型上，除了要把整體分類的正確率提高外，對於實際將發生逾放客戶之鑑別率，亦要更加重視。我們這在此定義可能發生誤判的錯誤型別如下：

1. 錯誤型別一(Type I error)：將可能發生逾放之客戶分類為好客戶，銀行會損失本金。
2. 錯誤型別二(Type II error)：將好客戶分類為可能發生逾放之客戶，銀行僅減少利息收入。

我們進一步的將所有分類可能發生之情形，以混淆矩陣(confusion matrix)來說明，如表 1。其中 TP(True Positive)為是逾放戶且也被正確分類至逾放戶之人數， TN (True Negative)為是正常戶且也被正確分類至正常戶；而 FN(False Negative)及 FP(False Positive)即為先前定義之錯誤型別一及錯誤型別二。

表 1 分類可能之情況列表

分類為→	逾放戶	正常戶
逾放戶	TP	FN (Type I error)
正常戶	FP (Type II error)	TN

整體的正確率(Accuracy)[15]定義為：

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad \textbf{(1)}$$

在本論文之應用上，由於損失本金與減少利息收益兩者之成本(cost)相差比率很大，所以對於本金的損失，我們比較重視。故要降低本金的損失應該要降低 Type I error 之比率，這等同於提高 true positive 之正確率(TPR)，其定義如下：

$$TPR = \frac{TP}{TP + FN} \quad \textbf{(2)}$$

本研究其他核心應用技術的原理，則分別在以下各小節中說明。

3.1 特徵選取(Feature Selection)

在監督式學習(Supervised Learning)的系統中，特徵的選取扮演相當重要的角色，他會影響整體分類的效能、執行時間及對類別(Class)的描述。我們可以透過衡量各特徵的鑑別能力，來挑選較重要的特徵值，其中權重評分法(weight score approach)為一常用的衡量方式[6,14]。權重分數(Weight Score, WS)是針對每一資料特徵，計算其兩個類別(classes)平均值之差的絕對值與標準差之和的比率，計算方式如下：

$$w_j = \frac{\left| u_j^+ - u_j^- \right|}{\sigma_j^+ + \sigma_j^-} \quad \textbf{(3)}$$

其中 u_j^+ 及 u_j^- 為第 j 個特徵值的 positive 跟 negative 樣本之平均值，σ_j^+ 及 σ_j^- 則為標準差。所以若平均值的差越大或標準差的和越小，則此數值越大，代表此資料特徵對資料集之鑑別能力越好且越重要。

3.2 平滑式支撐向量法 (Smooth Support Vector Machine)

授信後是否發生逾放在本論文中將視作二元分類的問題，即發生逾期放款之案件(即壞客戶)視作 positive 的案例，以+1 表示；而未發生逾放的案件(即好客戶)視作 negative 的案例，以 -1 表示。為解決此二元分類的問題，我們將利用在資料探勘的領域中，被視為解決二元分類問題之優質演算機制 SVM 來建立模型。在本節中，我們對 SVM 及我們使用的 SVM 衍生之演算機制 SSVM[8]的原理作一個說明。

假設給予一訓練資料集 $\{(x^i, y_i)\}_{i=1}^m$，其中 $x^i \in R^n$ 代表輸入的資料點，而 $y_i \in \{-1, 1\}$ 則為該資料點對應的類別標記，所以整個集合代表在 n 維空間中，m 個帶有二元類別標記的資料點，我們可以將這些資料點用 $m \times n$ 的矩陣 A 來儲存，其中第 i 列(表為 A_i)即代表 x^i。另外定義一對角矩陣 D，矩陣中每一對角元素 D_{ii} 即代表 y_i。假設資料集為線性可分割(如圖 2)，則空間中存在一分割平面(虛線所示) $x^T w + b = 0$ (w 為分割平面的法向量，b 則為分割平面與原點的距離)，可以正確地區分兩類資料點，資料點與界限平面(實線所示)的的關係則如下式：

$$\begin{cases} x^T w + b \geq +1, & \text{for } x \in A_+ \\ x^T w + b \leq -1 & \text{for } x \in A_- \end{cases} \quad \textbf{(4)}$$

其中 A+及 A-分別為 positive 及 negative 樣本之集合，Margin 定義為兩界限平面的距離(即 $2/\|w\|_2$)，SVM 演算機制即假設在給定一訓練集的情況下，作最佳化運算來搜尋具有最大 Margin 的分割平面，換言之即在限制條件下，求 $\frac{1}{2}\|w\|_2^2$ 的最小解，因此可以

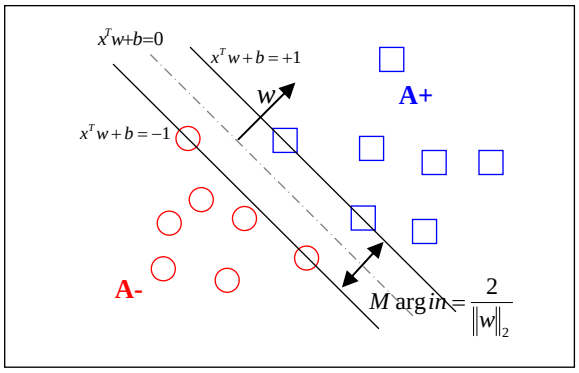


圖 2　SVM 在線性可分割之情況

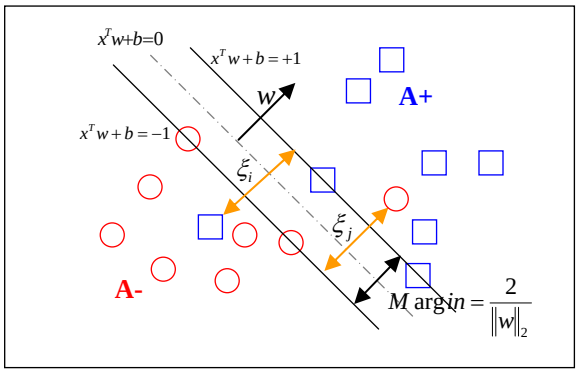


圖 3　SVM 在線性不可分割之情況

將 SVM 描述成為一個二次規劃問題(quadratic programming problem)：

$$\min_{(w,b)\in R^{n+1}} \frac{1}{2}(w^Tw)$$
$$subject\quad to\quad D(Aw+\mathbf{1}b)\geq \mathbf{1} \tag{5}$$

其中 $\mathbf{1}=[1,1,\ldots 1]^T\in R^m$。在此線性可分割的情形下，目標函數為二次凸面函數，因此存在一最佳解(w^*,b^*)，在界限平面(bounding plane) $x^Tw^*+b^*=\pm 1$上的資料點被稱為支撐向量(support vectors)，如果我們從訓練集中移去非支撐向量的資料點，重新學習的結果並不會因此改變，這是 SVM 極佳的特性及名稱的來由。亦即一旦獲致訓練的結果，我們可以只保留支撐向量即可維繫住完整的分類結果。

接下來說明如何處理非線性可分割資料集的情況，我們可以引進非負的鬆弛變數(slack variable), ξ。如圖 3 所示，若分割平面 $x^Tw+b=0$ 將第 i 個資料點分錯或資料點界於兩限制平面 $x^Tw+b=\pm 1$ 中，此時 ξ_i 的值將定為大於 0 的值，其他情形則為 0。如此對應說明圖形，我們可以列出傳統 SVM 對應的二次規劃問題：

$$\min_{(w,b,\xi)\in R^{n+1+m}} \frac{1}{2}(w^Tw)+C\mathbf{1}^T\xi$$
$$subject\quad to\quad D(Aw+\mathbf{1}b)+\xi\geq \mathbf{1},\quad \xi\geq \mathbf{0} \tag{6}$$

上式中 C 為一大於 0 的權重參數，如果我們著眼在減少訓練的誤差($\mathbf{1}^T\xi$)，可以選擇較大的 C 值；如果希望取得較寬 Margin，便可以選擇較小的 C 值。在實作上我們傾向在誤差小的情況下選擇較寬的 Margin 以避免發生 overfitting 的現象，所以如何選取一較佳表現的 C 值，就是 SVM 模式選擇(model selection)的一項課題。

接下來我們要說明的是本論文所採行的演算機制 SSVM。在不影響目標函數特性的前提下，我們將 SVM 目標函數中原本 1-norm 的 ξ 換成帶 $C/2$ 係數 2-norm 的 ξ 平方，對 Margin 的 n 維量度也改成 n+1 維量度，如此式(6)目標函數中的 $\frac{1}{2}\|w\|_2^2$ 被換成 $\frac{1}{2}\|(w,b)\|_2^2$，此時 SVM 最佳化的數學式可以改寫為：

$$\min_{(w,b,\xi)\in R^{n+1+m}} \frac{\mathbf{1}}{\mathbf{2}}(w^Tw+b^2)+\frac{C}{2}\xi^T\xi$$
$$subject\quad to\quad D(Aw+\mathbf{1}b)+\xi\geq \mathbf{1},\quad \xi\geq \mathbf{0} \tag{7}$$

定義 plus function $(\cdot)_+$ 為 $\max(\cdot,\mathbf{0})$，利用最佳解的充分必要條件(KKT conditions) [10]，可知產生最佳解時 $\xi=(\mathbf{1}-D(Aw+\mathbf{1}b))_+$，將此關係式代入原本的最佳化問題(7)中，便可將其改寫成無限制式的最佳化問題(unconstrained optimization problem)的模式求解：

$$\min_{(w,b)\in R^{n+1}} \frac{1}{2}(w^Tw+b^2)+\frac{C}{2}\|(1-D(Aw+1b))_+\|_2^2 \tag{8}$$

上式為無限制式的強凸形最小化問題(strongly convex minimization problem without any constraints)，因此會有唯一解。在不使用二次最佳化工具的前提上，擬套用牛頓法(Newton method)快速求解，但必須解決該式並非二次可微分(not twice differentiable)的問題。SSVM 利用平滑函數(smooth function)：

$$p(x,\alpha)=x+\frac{1}{\alpha}\log(1+e^{-\alpha x})\quad for\quad \alpha>0 \tag{9}$$

近似式(8)的 plus function 來解決這個問題，當平滑參數 α 愈大，函數(9)就愈接近 plus function，所以整個式子可以改寫如下：

$$\min_{(w,b)\in R^{n+1}} \frac{1}{2}(w^Tw+b^2)+\frac{C}{2}\|p(1-D(Aw+1b),\alpha)\|_2^2 \tag{10}$$

經過改寫的式子明顯為二次可微分，可以應用牛頓法求解。另外為了避免牛頓法求解過程中可能會發生振盪(oscillation phenomenon)的問題，SSVM 亦引用 Armijo Stepsize 的作法來解決這樣的問題。

利用核函數的技巧(kernel trick)[5]可輕易地將線性模式的 SVM 推廣至非線性模式，本文僅採用線性模式的作法。

3.3 非平衡資料集 (Imbalance Dataset)

在銀行授信實務上，餘額代償業務若在徵審人員善盡善良管理人之情況下，經驗上之逾放比率應維持在 10%以下(94 年 6 月金融機構之逾期放款比率平均為 2.46%)，故在預測可能發生逾放之分類問題上，資料集大多是非平衡(imbalance)的，這會降低模型對於未來可能發生逾放的鑑別力。目前已有些文獻探討[7,9]對於資料集使用重新取樣之技術，來增加特定類別的正確率。其原理為在原資料集中，調整特定類別在整個資料集中所佔之比率，一般而言有兩種方式：

1. Over-sampling：對特定類別重複取樣，增加其佔整體資料集之比率。
2. Down-sampling：對特定類別減少取樣，減少其佔整體資料集之比率。

這些方式對於 SVM 型式之演算法而言，有加重或減少對特定類別之重視或是懲罰(penalty)的作用，可以改變特定類別在分類上的正確率。在授信實務上，我們為了提高鑑別逾放戶之正確率，我們將在 SSVM 的模型訓練時，以倍數增加逾放戶(positive)的樣本份數，並以 Receiver Operating Curve，ROC[15] 曲線的覆蓋面積(Area Under the Curve，AUC)來評估整體分類的效能(較高之覆蓋面積代表具有較好的分類結果)。

4 實驗評估

本節將設計並實作第 2 節提出之評價模式，以北部某金融機構辦理餘額代償信用貸款之申請人為研究對象，並將該業務所做之聯徵查詢透過 RP 自動轉存至資料庫中。再依其他文獻及經驗上認有較有鑑別能力之特徵，整理成申請人之特徵資料集(訓練樣本)。所有之實驗將以線性的 SSVM 為主要演算法，以 10-fold Cross Validation ,CV [15]來評估。但因為 CV 為利用隨機抽樣之方式來進行訓練及評估結果，為求結果之客觀性，我們將對每個實驗重覆 25 次再取平均值，來增加結果之正確性，以避免樣本數過少，造成正確率之變動性過大。

4.1 資料集(Dataset)

本資料集來源以北部某金融機構在民國 93 年 7 月至 94 年 1 月間申請餘額代償服務之申請人為樣本，並以其過去被聯徵中心蒐集之歷史紀錄為特徵(Features)變數的來源，查詢聯徵的項目如表 2。

我們利用第 2 節中介紹之聯徵查詢後端儲存系統 - RP 來自動儲存自聯徵查詢之結果，包含未核貸之案件共計有 1,677 件樣本，並儲存在關聯式資料庫 - Inst DB 中。本研究在特徵之選定上，係參考相關理論及文獻之探討[2,3,4,11]，由於以往之文獻在使用聯徵查詢結果上著墨不多，故在與該金融機構授信相關資深人員討論後，我們將聯徵中心查詢之資料重新編整成經驗上較具鑑別性之特徵值

表 2 餘額代償之聯徵查詢項目

查詢項目	主要提供資訊說明	報送週期
B05/個人授信餘額變動資訊	本項查詢可了解申請人之過去 13 個月月底及前二、三年年底之各類放款之餘額變動情形，其中包含短期放款、中長期放款、擔保放款、遠期信用狀、押匯承兌、本票保證、保證款、總餘額(不含逾催保)、逾期、催收、呆帳等資料。	每月
B32/授信、還款紀錄與保證資訊－行庫別	本項查詢區分為主債務、共同債務、從債務、其他等四大項目，並均提供申請時前 12 個月份各類授信之還款情形紀錄。	每月
K21/信用卡主附卡資訊	本項查詢提供申請人過去申請自身之信用卡或其附卡之所有申辦歷史紀錄，包含發卡機構、卡別、啟用日、額度、停用日、停用原因等。	即時
K22/信用卡戶基本資訊彙總	本項查詢提供申請人過去申請信用卡時之個人基本資料，包含報送年月、報送機構、申請人姓名、出生日期、教育程度、戶籍地址、帳單地址、居住電話、辦公電話、任職機構、職稱、服務年資、年薪	即時
K23/信用卡戶繳款紀錄資訊	本項查詢提供申請人過去 12 個月份之所有有效之信用卡之繳款狀況，包含報送年月、發卡機構、卡名、額度、應繳金額、是否預借現金、繳款狀況(未消費、全額繳清無延遲、全額繳清有延遲、循環信用無延遲、循環信用有延遲、全額逾期未繳、未繳足最低金額)	即時
資料來源：財團法人聯合徵信中心		

共 33 個，如表 3。其中共有 27 個數值型(numerical)及 6 個非數值型(nominal)的特徵，各特徵變數如下：X1 (性別)是由身分證號第 2 碼取得，而 X2 (年齡)則是由出生年月日與申請日期做運算取得，特徵 X3(擁有自宅), X4(遺失身分證紀錄), X5(現金卡張數), X6(有效信用卡張數), X7(強停信用卡數), X8(學歷), X9(年收入), X18(擔負他人借款)等 8 個是由資料庫中直接取得。特徵 X11, X12, X13, X14, X15, X16 等 6 個是對過去 13 月之餘額變動值以一般數學統計之方式來描述其變化量，分別為取最大值、平均值、標準差等統計量而取得，但特徵 X15 則是以線性迴歸(linear regression)[12]之方式來描述過去 13 個月餘額之增減比率。我們對一連續數

表 3 特徵變數列表

特徵代號	特徵名稱	資料形態
X1	性別	Nominal
X2	年齡	Numerical
X3	擁有自宅	Nominal
X4	遺失身分證紀錄	Nominal
X5	現金卡張數	Numerical
X6	有效信用卡張數	Numerical
X7	強停信用卡數	Numerical
X8	教育程度	Nominal
X9	年收入	Numerical
X10	居住於大都會區	Nominal
X11	過去 13 個月債務之最大值	Numerical
X12	過去 13 個月債務之平均值	Numerical
X13	過去 13 個月債務之標準差	Numerical
X14	過去 13 個月債務之成長率	Numerical
X15	過去 13 個月擔放之最大值	Numerical
X16	過去 13 個月擔放之平均值	Numerical
X17	擔負他人債務	Nominal
X18	擔放額度	Numerical
X19	擔放餘額	Numerical
X20	擔放使用率	Numerical
X21	現金卡核准額	Numerical
X22	現金卡餘額	Numerical
X23	現金卡使用率	Numerical
X24	總授信之額度	Numerical
X25	總授信之餘額	Numerical
X26	總授信之使用率	Numerical
X27	信用卡總額度	Numerical
X28	信用卡平均額度	Numerical
X29	信用卡使用循環息次數	Numerical
X30	信用卡使用預借現金次數	Numerical
X31	信用卡全額逾期未繳之累計	Numerical
X32	被查詢次數	Numerical
X33	被同一家銀行查詢 2 次之次數	Numerical
資料來源：本研究整理		

值做線性迴歸可以得到其變動之斜率，圖 4 為一簡易的示意圖。特徵 X18, X19, X20, X 21, X 22, X 23, X 24, X 25, X 26 共 9 個，每 3 個一組，每一組之第

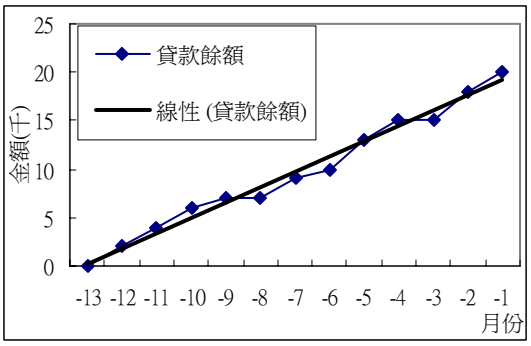


圖 4 餘額變動線性迴歸示意圖

3 個值為前 2 個值之比例，即為各類貸款之使用率，額度)對各家發卡銀行歸戶計算，以額度最高的卡片為準，一般個人在同一發卡機構發行之所有信用卡皆共用同一額度。特徵 X28(信用卡平均額度)為總額度除以有效發卡機構之家數。特徵 X29,X30, X31 為信用卡使用特徵描述之變數，分別為近 13 個月使用循環信用之次數、預借現金之次數、曾經發生全額逾期未繳之總金額。而特徵 X32, X33 分別為近三個月被金融機構查詢之次數及近三個月被同一金融機構查詢 2 次之家數，從這個數字可得知申請人近三個月申請貸款之密集程度，查詢次數高者可能代表最近曾向多家金融機構申請貸款。

4.2 資料前處理(Data preprocessing)

由於 SVM 只能對數值型特徵值直接進行運算，故所有的非數值型之特徵值將以 0、1 之數值取代成為數值型之值，而 X8(教育程度)則以 5 到 1 的數值分別代表博士、碩士、大專、高中、其他等五種教育程度。

我們從關聯式資料庫中，選取並組合先前訂定之特徵，將每一個樣本(Instance)轉換成一維的陣列，每個樣本為 $\{(x^i, y_i)\}_{i=1}^m$，其中 $x^i \in R^n$，共有 m 個樣本及 n 個特徵，每個樣本之標籤(Label) y_i 為是否發生逾期放款。逾期放款依目前法令[1]規定為：積欠本金或利息超過清償期三個月，或雖未超過三個月，但已向主、從債務人訴追或處分擔保品者。

在進行資料轉換時，有部份樣本核貸並未超過六個月(可能無法確認其是否為好客戶)或雖已發生逾期繳款但未超過逾期放款規定者均先行排除。再考量資料之完整性時，有部份案件在聯徵資料的查詢上並非如表 2 般完整(無查詢 B05)，或是有太多的特徵值為空(null)者，亦予以去除。最後考量逾期客戶與正常戶之核貸成本(cost)後，希望將更多之逾期放款樣本加入，故除原本由 RP 自動擷取之資料外，我們再將 RP 導入前之逾期案件樣本 90 件，以人工之方式建立。最後可供使用之有效樣本數為 391 筆，命名為 tw-33，其中逾期戶(Positive)為 124 筆(31.71%)，正常戶(Negative)為 267 筆(68.29%)。

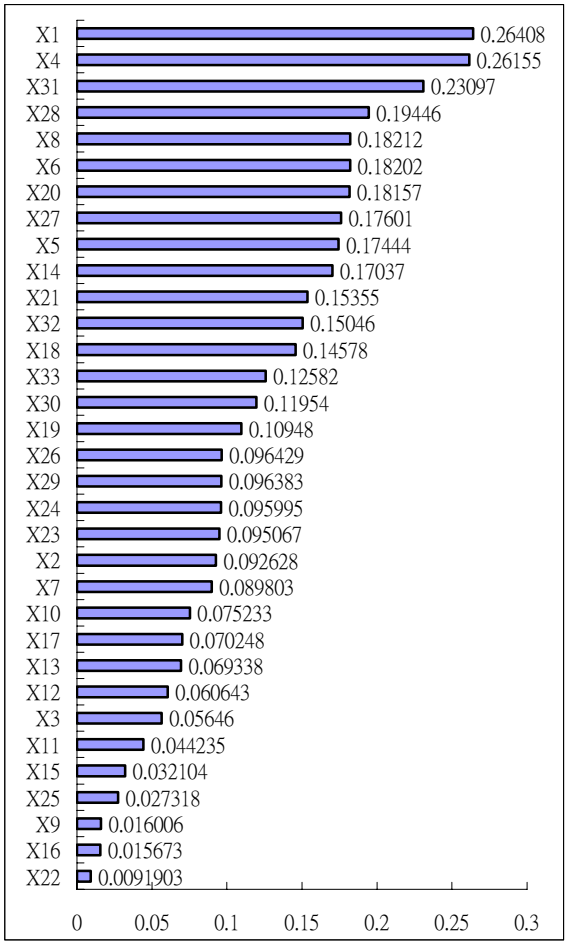


圖 5 各特徵之 WS

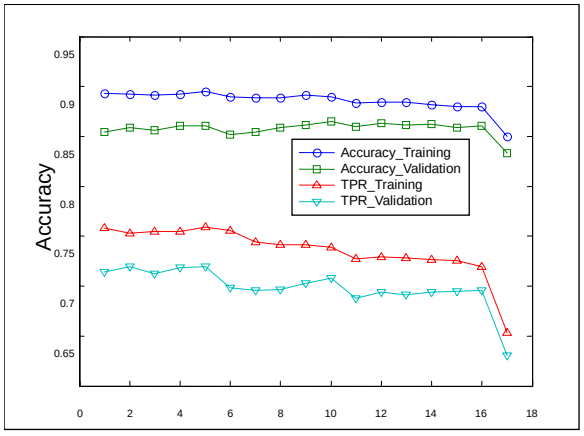


圖 6 依序移除 WS 較低特徵之影響

4.3 特徵選取(Feature Selection)

本實驗將以 WS 試算每個特徵的得分，作特徵選取後再使用 SSVM 來做實驗求其正確率，以挑選較重要之特徵，各特徵值之 WS 依得分之排序如圖 5。我們將後 50%(16 個)得分較小的特徵，從資料集中從最小的依序移除並使用 SSVM 來分類，結果如圖 6。圖中最左第一個點為未移除任何特徵之準確率。在移除到第 4 個得分較低的特徵(X25)之後，整體之 accuracy 及 true positive rate 開始有下滑的趨勢，所以最後我們將得分最少的 4 個特徵(X22, X16, X9, X25)移除，並更新本評價模式所之特徵集，重新將資料集命名為 tw-29。圖 7 為特徵選取前後之 ROC 曲線圖，tw-33 與 tw-29 兩曲線包含的面積(AUC)分別為 0.9075 與 0.9079，相差甚微，代表移除 WS 值最小的 4 個特徵後，SSVM 在此分類問題上的表現並不受影響，且我們可以觀察到 tw-29 在 FP-rate 前 10%時有較佳的表現。

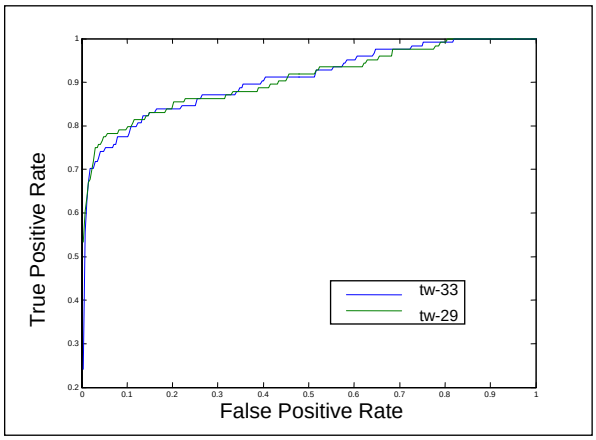
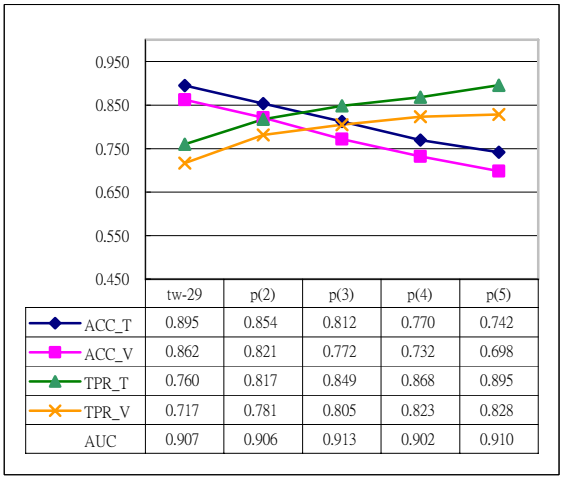


圖 7 特徵選取前後之 ROC 曲線

4.4 Over-Sampling 之影響評估

tw-29 為一個非平衡的資料集，目前比例為 124:267(1:2.15)，約 1:2 (positive vs. negative)。本實驗將使用超取樣(Over-Sampling)之方式，以倍數增加 positive 在訓練時的樣本數。藉由調整 positive 樣本之倍數來改變 positive 及 negative 之分佈比率，並找到合適之樣本比率。實驗結果如圖 8，其中 $p(i)$ 為複製 i 次 positive 樣本之資料集，部份(原始資料集、$p(3)$、$p(5)$)結果之 ROC 曲線比較如



	tw-29	p(2)	p(3)	p(4)	p(5)
ACC_T	0.895	0.854	0.812	0.770	0.742
ACC_V	0.862	0.821	0.772	0.732	0.698
TPR_T	0.760	0.817	0.849	0.868	0.895
TPR_V	0.717	0.781	0.805	0.823	0.828
AUC	0.907	0.906	0.913	0.902	0.910

圖 8 Over-Sampling 對正確率之影響

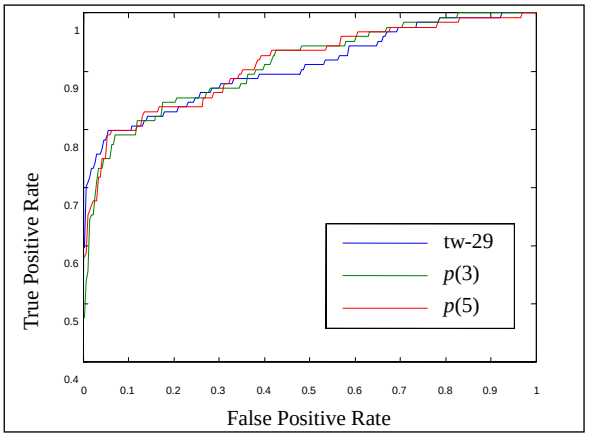


圖 9　使用倍增 positive 樣本之 ROC 曲線比較

圖 9。我們分別以訓練(Training)及評估(Validation)之正確率來觀察 Over-Sampling 對於正確率提升之影響。從圖 8 中可以得知，利用 Over-Sampling 的方式可有效提昇 TPR，然而整體之正確率會隨之下降(FN 增多)，至於應將 positive 樣本複製幾倍，可在實務上由利差與逾放比率來調整，求取授信銀行之最大利潤決定之。

5 結論

本論文提出之餘額代償信用評價模式，從以上的實驗結果中獲得實證，這也同時印證了 SVM 在分類問題上之絕佳性能。

在應用特徵選取(Feature Selection)後之資料集 tw-29 及 SSVM 作為核心技術後，整體 Accuracy 及 TPR 可達到 86.22%及 71.67%，若在授信總金額 10 億元逾放比率為 5%之情況下，則可減少 3,733.5 萬之逾放損失，對於資產品質的提升有相當之幫助。而在特徵的使用上可謂是創新，以往之文獻探討中大多會使用許多申請人提供之信用資料，再輔以聯徵中心之查詢結果做為信用評價之依據，其過程不但耗時，且在建檔之過程中很容易產生錯誤，在本論文中以低廉之成本自動取得聯徵中心之查詢結果，並獲得不錯之正確率，且可以減少人員在資料之審核及校對，大幅的降低人工之成本，相信可提供後繼在研究信用評等上一新的思維。

後續在於信用評等(Grade)的輸出上，我們認為可以透過 ROC 曲線之概念，將 SSVM 輸出之 y 值先行排序，以得分之高低與累計之 True Positive rate 為參考標的設定多個門檻值，將二元分類拓展成有機率概念之多元分類結果，以作為調整授信條件門檻之參考。

參考文獻

[1] 財政部，銀行資產評估損失準備提列及逾期放款催收款呆帳處理辦法，2004。

[2] 戴堅，個人消費性信用貸款授信評量模式之研究，國立中正大學國際經濟研究所，2004。

[3] 戴嘉甫，銀行現金卡客戶違約機率之衡量，義守大學管理科學研究所，2004。

[4] 鐘志明，現金卡二次授信風險實證分析，國立高雄第一科技大學風險管理與保險所，2004。

[5] C. J. C. Burges, “A tutorial on support vector machines for pattern recognition”, Data Mining and Knowledge Discovery, col.2, pp.121-167, 1998.

[6] Chia-Huang Chao, “Feature Selection for Microarray Gene Expression Data”, National Taiwan University Science and Technology, 2004.

[7] Nathalie Japkowicz, “Learning from Imbalanced Data Sets: A Comparison of Various Strategies.” , AAAI Press, Technical Report WS-00-05,pp. 10–15, 2000.

[8] Yuh-Jye Lee and O. L. Mangasarian, “SSVM: A Smooth Support Vector Machine for Classification”, Computational Optimization and Applications, 2001.

[9] Alexander Yun-chung Liu, ”The Effect of Oversampling and Undersampling on Classifying Imbalanced Text Datasets”, The University of Texas at Austin, 2004.

[10] O. L. Mangasarian, *Nonlinear Programming*, SIAM, 1994.

[11] Kasper Roszbach, ”Bank Lending Policy, Credit Scoring and the survival of Loans”, SVERIGES RIKSBANK WORKING PAPER SERIES,pp.7, 2003.

[12] Sen Ashish K., Srivastava M. S., *Regression analysis: theory methods and applications*, Springer-Verlag, 1990.

[13] Vladimir N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, 1995

[14] Weston, J., S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio and V. Vapnik, “Feature Selection for SVMs”, NIPS, 2000.

[15] Ian H. Witten, Eibe Frank, *Data Mining*, Morgan Kaufmann, 1999.