應用資料探勘技術於房貸信用風險預測之研究
Data Mining on Credit Risk Prediction of Mortgage

李御璽	顏秀珍	林基玄
銘傳大學資工系	銘傳大學資工系	台新銀行資訊服務處
leeys@mcu.edu.tw	sjyen@mcu.edu.tw	seanlin@taishinbank.com.tw

吳政瑋	粘嘉菖	鄭郁翰	張韋豪
銘傳大學資工所	銘傳大學資工所	銘傳大學資管所	銘傳大學資工所
silvemoonfox@hotmail.com	ruaddick@gmail.com	cutejolin@gmail.com	qoojohn7@hotmail.com

摘要

近年來，金融機構發放信用卡與現金卡浮濫。使得銀行因雙卡客戶逾期產生呆帳，而造成巨大的虧損。金融業者也注意到，同樣的情況也可能發生在房屋貸款(Mortgage)上。房屋貸款主要是借貸者因購屋需求，以房屋為擔保品向銀行進行借貸。如同信用卡借貸，一但房屋貸款出現大量違約的情況，對於銀行的虧損更是巨大。為了避免同樣的問題發生於房屋貸款上，金融機構也相當重視房屋貸款的信用風險(Credit Risk)評估。為了幫助金融機構評估房貸風險並審核房貸申請者，本研究運用資料探勘(Data Mining)中的分類(Classification)技術與 HMEQ 資料庫，利用過去銀行借貸者申請房屋貸款的資料，建立一套預測房貸信用風險的模型，幫助銀行或財務公司預測借貸者將來是否會有違約的情況發生。鑒於過去預測房屋貸款只建立一個分類模型導致過於主觀的缺點，本研究以五疊交互驗證法(Five-Fold Cross-Validation)為基礎，將資料集切割成五組訓練資料及其相對應的測試資料，分別進行資料前處理(Data Preprocessing)，並提出一套選取重要屬性(Significant Attribute Selection)的方法，最後建立模型進行探勘，並以五組分類模型的平均預測能力，作為最後整體模型的預測能力。實驗的結果顯示，經過資料前處理及挑選重要屬性後的模型，效能優於無處理過的分類模型。

關鍵詞：分類、房貸風險、資料探勘、屬性選擇

應用資料探勘技術於房貸信用風險預測之研究
Data Mining on Credit Risk Prediction of Mortgage

李御璽 顏秀珍 林基玄

銘傳大學資工系 銘傳大學資工系 台新銀行資訊服務處

leeys@mcu.edu.tw sjyen@mcu.edu.tw seanlin@taishinbank.com.tw

吳政瑋 粘嘉菖 鄭郁翰 張韋豪

銘傳大學資工所 銘傳大學資工所 銘傳大學資管所 銘傳大學資工所

silvemoonfox@hotmail.com ruaddick@gmail.com cutejolin@gmail.com qoojohn7@hotmail.com

摘要

近年來,金融機構發放信用卡與現金卡浮濫。使得銀行因雙卡客戶逾期產生呆帳,而造成巨大的虧損。金融業者也注意到,同樣的情況也可能發生在房屋貸款(Mortgage)上。房屋貸款主要是借貸者因購屋需求,以房屋為擔保品向銀行進行借貸。如同信用卡借貸,一但房屋貸款出現大量違約的情況,對於銀行的虧損更是巨大。為了避免同樣的問題發生於房屋貸款上,金融機構也相當重視房屋貸款的信用風險(Credit Risk)評估。為了幫助金融機構評估房貸風險並審核房貸申請者,本研究運用資料探勘(Data Mining)中的分類(Classification)技術與HMEQ 資料庫,利用過去銀行借貸者申請房屋貸款的資料,建立一套預測房貸信用風險的模型,幫助銀行或財務公司預測借貸者將來是否會有違約的情況發生。鑒於過去預測房屋貸款只建立一個分類模型導致過於主觀的缺點,本研究以五疊交互驗證法(Five-Fold Cross-Validation)為基礎,將資料集切割成五組訓練資料及其相對應的測試資料,分別進行資料前處理(Data Preprocessing),並提出一套選取重要屬性(Significant Attribute Selection)的方法,最後建立模型進行探勘,並以五組分類模型的平均預測能力,作為最後整體模型的預測能力。實驗的結果顯示,經過資料前處理及挑選重要屬性後的模型,效能優於無處理過的分類模型。

關鍵詞：分類、房貸風險、資料探勘、屬性選擇

1. 緒論

近年來，銀行與財務公司因為忽略信用卡及現金卡的風險管控，而爆發了卡債風暴，顧客違約呆帳的情況使銀行蒙受巨大的損失。而金融業者也注意到，同樣的情況可能發生在房屋貸款(Mortgage)上。房屋貸款主要是因購屋需求，以房屋作為擔保品向銀行進行借貸。為了避免同樣的情況發生在房屋貸款上,金融業者也相當重視房屋貸款的信用風險(Credit Risk)評估。

資料探勘(Data Mining)是指從大量的資料中找出隱含、未知且有用的資訊[1]。本研究使用資料探勘中的分類(Classification)技術 [2,3,4,5,6] 與 HMEQ 資料庫，利用過去銀行借貸者申請房屋貸款的資料，以 SPSS Clementine 8.5 探勘軟體為工具，建構一套分類模型，幫助銀行或財務公司預測房貸申請者將來是否會違約。

由過去研究得知，資料品質、屬性選擇、以及分類演算法的挑選將影響分類模型的正確率。若資料品質差,具有許多空值(Null Value)或錯誤值，將影響分析結果。此外，重要屬性是否收集完整，所使用的分類演算法是否適切,參數的設定等因素，都將影響分類模型的正確率。因此，本論文首先依照分層抽樣切割原始資料庫，以五疊交互驗證法(Five-Fold Cross-Validation)為基礎將資料分成五組訓練-測試資料(Training-Testing Data)，並對這五組訓練-測試資料，分別進行資料前處理(Data Preprocessing)的工作。處理的項目包括檢視原始資

料，填補空值、處理偏態屬性等。接著對每一組訓練資料分別選取重要的欄位。選出對於分類模型具有幫助的重要欄位後，對每組訓練資料建立分類模型，得到五組分類模型，最後以 F-measure 為評估方法，評估每一組分類模型的效能。詳細作法將在第三節中敘述。

2. HMEQ 資料庫描述

HMEQ 資料庫共有 5960 筆資料。每一筆紀錄代表一位顧客的行為，包括房屋貸款的資訊與顧客的資料。每筆紀錄共有 13 個屬性，包括 1 個二元的目標屬性與 12 個條件屬性。這 12 個條件屬性包括 10 個數值屬性與 2 個類別屬性。以下將介紹每個欄位的名稱、意義及屬性值。

屬性 01：Bad (二元目標屬性)：代表借貸者過去是有否違約的紀錄，共有 5960 筆紀錄。屬性值為 0 代表顧客沒有違約的紀錄，我們稱為好客戶，共有 4771 筆記錄。屬性值為 1 代表顧客有違約或嚴重滯延的記錄，我們稱為壞客戶，共有 1189 筆紀錄。

屬性 02：Loan (數值屬性)：為顧客的貸款金額。

屬性 03：Mortude (數值屬性)：目前抵押品之到期金額，指借貸者剩多少金額未償還。

屬性 04：Value (數值屬性)：借貸者的資產價值。

屬性 05：Yoj (數值屬性)：借貸者目前工作的年資，以年為單位。

屬性 06：Clage (數值屬性)：借貸者最長的授信期間。指借貸者貸款最久的時間，以月為單位。

屬性 07：Debtinc (數值屬性)：借貸者的負債與所得比例。

屬性 08：Derog (數值屬性)：借貸者的信用在聯徵中心被查詢的次數。

屬性 09：Delinq (數值屬性)：借貸者逾期還款的次數。

屬性 10：Ninq (數值屬性)：借貸者信用出現問題的次數。

屬性 11：Clno (數值屬性)：授信放款筆數。

屬性 12：Reason (類別屬性)：顧客貸款的原因，包括房屋修繕與負債整合。

HomeImp：房屋修繕。

DebtCon：負債整合。

屬性 13：Job (類別屬性): 顧客的職業，包括以下六個屬性值。

Office：白領階級中的一般職員。

Mgr：白領階級中的管理階級

ProfExe：藍領階級，有專門職業技術者

Self：自行創業者

Sales：行銷人員

Other：其他職業

表一為屬性空值分析表，可檢視 HMEQ 資料庫的資料品質。目標屬性 Bad 與條件屬性 Loan 皆無空值。條件屬性 Debtinc 則具有許多的空值。

表一：屬性空值分析表

屬性	無空值的資料筆數	各屬性空值所佔比例 (%)
Bad	5960	0.00
Loan	5960	0.00
Mortude	5442	8.69
Value	5848	1.88
Yoj	5445	8.64
Clage	5652	5.17
Debtinc	4693	21.26
Derog	5252	11.88
Delinq	5380	9.73
Ninq	5450	8.56
Clno	5738	3.72
Reason	5708	4.23
Job	5681	4.68

表二：各數值屬性值的統計資訊

屬性	最小值	最大值	平均值	標準差	中位數
Loan	1100	89900	18607.97	11207.480	16400
Mortude	2063	399550	73760.816	44457.610	63836
Value	8000	855909	101776.048	57385.776	87000
Yoj	0.000	41.0000	8.922	7.574	7.000
Clage	0.000	1168.23	85.810	85.810	174.390
Debtinc	0.524	203.312	33.78	8.602	34.962
Derog	0	10	0.255	0.846	0
Delinq	0	15	0.449	1.127	0
Ninq	0	17	1.186	1.729	1
Clno	0	71	10.139	10.139	20

表二為各數值屬性的統計資訊。包括各屬性的最大值、最小值、平均值、標準差、中位數等資訊。

3.研究方法

3.1 切割資料庫

傳統上只利用一組訓練資料來建立模型過於主觀，故本研究將資料庫切割成五組訓練-測試資料，分別建立五組分類模型。最後以這五組分類模型的平均效能作為整體模型的效能。

首先先分開 HMEQ 資料庫中，Bad 目標屬性中屬性值為 1 與目標屬性值為 0 的紀錄。將所有目標屬性值為 1 的紀錄所成的集合稱為 Bad_dataset，共有 1189 筆紀錄；所有目標屬性值為 0 的紀錄所成的集合稱為 Good_dataset，共有 4771 筆紀錄。接著以五疊交互驗證法為基礎，將 Bad_dataset 切成五份，分別為 Bdb_1、Bdb_2、Bdb_3、Bdb_4、Bdb_5。並將 Good_dataset 切成五份，分別為 Gdb_1、Gdb_2、Gdb_3、Gdb_4、Gdb_5。將 Bdb_i 與 Gdb_i 合併為 DB_i，$(1 \leqq i \leqq 5)$。例如將 Bdb_1 與 Gdb_1 合併成 DB_1，依此類推，可得到五個資料集。

然後以 DB_1 為測試資料(Testing Data)，以下稱為 $TestDB_1$，並合併剩餘的資料 DB_2~DB_5 為訓練資料(Training Data)，以下稱為 $TrainDB_1$。以 DB_2 為 $TestDB_2$，合併 DB_1、DB_3、DB_4 與 DB_5 為 $TrainDB_2$。依此類推，我們可以得到五個測試資料 $TestDB_i$ 與其相對應的訓練資料 $TrainDB_i$ (i=1~5)。

HMEQ 資料庫中共有 5960 筆記錄，每一組訓練資料共有 4768 筆，測試資料共有 1192 筆。訓練資料於 HMEQ 資料庫中所佔的比例約為 80%，測試資料約為 20%。每一組訓練-測試資料中，好壞顧客的比例與 HMEQ 資料庫中的比例相同，均為 80:20，如表三所示。

表三：訓練資料與測試資料好壞顧客分佈

資料集種類	訓練資料 4768 筆(80%)		測試資料 1192 筆(20%)	
目標屬性	好客戶	壞客戶	好客戶	壞客戶
資料筆數	3817	951	954	238

3.2 處理空值

由於每一組訓練-測試資料內的資料都含有空值，故需要加以處理。由於 Bad 和 Loan 屬性都沒有空值，故無須處理。而 Mortude、Value、Yoj、Clage、Debtinc 等數值屬性都有分佈偏態(Skewness Distribution)的情況發生，故填補空值的方式以填入中位數為主，雖然填補空值亦可採用平均數，但

較不適用於有偏態分布的屬性上。Derog、Delinq、Ninq、Clno 等屬性值的意義都是行為的次數，故填補空值的方式也以填補中位數為主。Reason 及 Job 等屬性為類別屬性，填補 Reason 及 Job 等屬性的空值，以填入眾數為主。Reason 屬性的空值填上負債整合(DebtCon)，而 Job 屬性的空值則填上其它(Other)。依此類推，我們可以得到五組已填補完空值的訓練-測試資料。接著將對每組訓練資料進行屬性重要程度分析。

3.3 重要屬性程度分析

訓練資料於建立分類模型前，會對所有屬性作分析，選擇重要且具有區分能力的屬性作為分類模型的輸入。並非所有的屬性對於分類模型皆有幫助，若將不重要的屬性輸入至分類模型中，不僅會拉長模型的學習時間，也會降低分類系統的正確率。在本研究中，挑選重要屬性的計算方式包括數值型屬性與類別型屬性的挑選。首先介紹挑選數值型重要屬性的計算方法。

3.3.1 數值型重要屬性計算方法

數值屬性的重要程度的計算方式如公式(1)所示。將訓練資料中的客戶依照目標屬性值，分成好顧客與壞顧客，並分別求其平均數與標準差。如公式(1)所示，好壞顧客的平均值差異越大，代表該屬性越能鑑別好壞顧客；若好壞顧客的標準差差異越大，代表該屬性的鑑別能力較不穩定。

$$\text{數值屬性重要程度} = \frac{|\text{壞顧客平均值} - \text{好顧客平均值}|}{(\text{壞顧客標準差} + \text{好顧客標準差})/2}$$

公式(1)

以 $TrainDB_1$ 中的 Delinq 屬性為例，表四為此屬性的平均值(Mean)及標準差(Standard Deviation)分別在目標屬性值 0 與 1 的分佈。根據公式(1)，Delinq 在 $TrainDB_1$ 中的重要屬性程度為 0.978 / 1.268 = 0.771。依此類推，我們可以得到所有數值屬性在各組訓練資料中的重要程度。

表四:屬性 Delinq 在 $TrainDB_1$ 中的平均值及標準差

	壞客戶(Bad=1)	好客戶(Bad=0)
平均值	0.219	1.197
標準差	0.634	1.902

3.3.2 類別型重要屬性計算方法

對於類別屬性，首先先計算各屬性所出現的頻率(Frequency；Freq)、支持度(Support；Sup)、與信賴度(Confidence；Conf)，再計算其屬性的重要程度。計算方法如下。公式中的 Cm 代表 Class m，An 代表屬性值 n。

Cm_An 的支持度: 公式(2)

$$\text{Sup}(\text{Cm\_An}) = \frac{\text{Freq}(\text{Cm\_An})}{\text{Freq}(\text{Cm})}$$

Cm_An 的信賴度: 公式(3)

$$\text{Conf}(\text{Cm\_An}) = \frac{\text{Sup}(\text{Cm\_An})}{(\text{Sup}(\text{C1\_An}) + \text{Sup}(\text{C2\_An}))}$$

屬性值 An 的重要性 = 公式(4)

$$\frac{(|(\text{Conf}(\text{C1\_An}) - \text{Conf}(\text{C2\_An}))| * ((\text{Freq}(\text{C1\_An}) + (\text{Freq}(\text{C1\_An})))}{(\text{Freq}(\text{C1}) + \text{Freq}(\text{C2}))}$$

類別屬性重要性= Sum(An 的屬性重要性) 公式(5)

以 $TrainDB_1$ 中的 Reason 屬性為例，表五為此屬性值在目標屬性值為 0 與 1 上所出現的頻率。

表五: Reason 屬性值於目標屬性值 0 與 1 之出現頻率

Reason 屬性值	目標屬性為 0 (好客戶)之出現頻率	目標屬性為 1(壞客戶)之出現頻率
DebtCon	2706	635
HomeImp	1111	316
總和	3817	951

表六：所有屬性的重要程度

	$TrainDB_1$	$TrainDB_2$	$TrainDB_3$	$TrainDB_4$	$TrainDB_5$	平均值	排名
Delinq	0.771	0.758	0.739	0.74	0.722	0.7460	1
Derog	0.547	0.573	0.582	0.563	0.562	0.5654	2
Clage	0.396	0.392	0.51	0.399	0.416	0.4226	3
Ninq	0.374	0.377	0.409	0.388	0.383	0.3862	4
Debtinc	0.4	0.355	0.344	0.307	0.343	0.3498	5
Loan	0.17	0.219	0.193	0.187	0.164	0.1866	6
Yoj	0.117	0.136	0.15	0.164	0.14	0.1414	7
Mortude	0.065	0.148	0.117	0.103	0.134	0.1134	8
Job	0.11	0.114	0.134	0.1	0.106	0.1128	9
Value	0.038*	0.084	0.12	0.084	0.094	0.0840	10
Reason	0.04*	0.043*	0.038*	0.044*	0.044*	0.0418*	11
Clno	0.012*	0.038*	0.001*	0.006*	0.005*	0.0124*	12

首先以公式(2)求出各屬性值的支持度。

$$\text{Sup(C0\_DebtCon)} = \frac{\text{Freq(C0\_DebtCon)}}{\text{Freq(C0)}} = \frac{2706}{3817} = 0.708$$

$$\text{Sup(C0\_HomeImp)} = \frac{\text{Freq(C0\_HomeImp)}}{\text{Freq(C0)}} = \frac{1111}{3817} = 0.091$$

$$\text{Sup(C1\_DebtCon)} = \frac{\text{Freq(C1\_DebtCon)}}{\text{Freq(C1)}} = \frac{635}{951} = 0.667$$

$$\text{Sup(C1\_HomeImp)} = \frac{\text{Freq(C1\_HomeImp)}}{\text{Freq(C1)}} = \frac{316}{951} = 0.332$$

再根據支持度並利用公式(3)，求出各屬性值的信賴度。

Conf(C0_DebtCon) =

$$\frac{\text{Sup(C0\_DebtCon)}}{(\text{Sup(C0\_DebtCon)} + \text{Sup(C1\_DebtCon)})} = \frac{0.708}{1.375} = 0.514$$

Conf(C1_DebtCon) =

$$\frac{\text{Sup(C1\_DebtCon)}}{(\text{Sup(C0\_DebtCon)} + \text{Sup(C1\_DebtCon)})} = \frac{0.667}{1.375} = 0.485$$

Conf(C0_HomeImp) =

$$\frac{\text{Sup(C0\_HomeImp)}}{(\text{Sup(C0\_HomeImp)} + \text{Sup(C1\_HomeImp)})} = \frac{0.291}{0.623} = 0.466$$

Conf(C1_HomeImp)=

$$\frac{\text{Sup(C1\_HomeImp)}}{(\text{Sup(C0\_HomeImp)} + \text{Sup(C1\_HomeImp)})} = \frac{0.332}{0.623} = 0.533$$

再根據所計算出來的支持度與信賴度，以公式(4)計算 DebtCon 與 HomeImp 的重要程度。

DebtCon 的屬性值重要程度 =

$$\frac{(|0.514 - 0.485| * (2706 + 635))}{(3817 + 951)} = 0.02$$

HomeImp 的屬性值重要程度 =

$$\frac{(|0.466 - 0.533| * (1111 + 316))}{(3817 + 951)} = 0.02$$

最後根據公式(5)，將 DebtCon 的屬性值重要程度與 HomeImp 的屬性重要程度加總，故 Reason 屬性於 $TrainDB_1$ 中的重要程度為 0.02 + 0.02 = 0.04。

表六為每一組訓練資料中，每種屬性的重要程度。表七中的「平均值」代表該欄位在五組訓練資料中的平均重要性。例如屬性 Loan 的平均重要性為 (0.17 + 0.219 + 0.193 + 0.187 + 0.164) / 5 =

表七: 對偏態欄位取對數後所有屬性的重要程度

	$TrainDB_1$	$TrainDB_2$	$TrainDB_3$	$TrainDB_4$	$TrainDB_5$	平均值	排名
Delinq	0.771	0.758	0.739	0.740	0.722	0.7460	1
Derog	0.547	0.573	0.582	0.563	0.562	0.5654	2
L_Clage	0.434	0.419	0.476	0.426	0.444	0.4398	3
Ninq	0.374	0.377	0.409	0.388	0.383	0.3862	4
L_Loan	0.309	0.340	0.316	0.308	0.287	0.3120	5
L_Debtinc	0.274	0.236	0.242	0.183	0.277	0.2424	6
L_Value	0.165	0.194	0.234	0.210	0.191	0.1988	7
L_Mortude	0.123	0.190	0.189	0.177	0.182	0.1722	8
L_Yoj	0.126	0.169	0.159	0.189	0.170	0.1626	9
Job	0.110	0.114	0.134	0.100	0.106	0.1128	10
Reason	0.04*	0.043*	0.038*	0.044*	0.044*	0.0418*	11
Clno	0.012*	0.038*	0.001*	0.006	0.005*	0.0124*	12

0.1866。依此類推，可得到其它屬性於五組訓練資料的平均重要程度。

在於本研究中，挑選重要屬性前需由使用者定義重要屬性門檻值，若屬性的重要程度超過該門檻值，我們才認為該屬性是一個重要的屬性，並可作為分類模型的輸入。於本研究中所設定的門檻值為 0.05，重要屬性值超過 0.05 以上的屬性，本研究才認為該屬性為重要的屬性。於表六中，未超過門檻值的屬性其右上角以 * 做為識別。例如屬性 Value 的重要程度為 0.038，在第一組訓練資料中並未大於 0.05，故數值右上角附注 * 表示之。此外，本研究挑選重要屬性的方式可分為兩種。第一種方法稱為個別選取法，第二種方式稱為投票選取法。以下將介紹這兩種方法。

3.3.3 以個別選取法挑選重要屬性

所謂的個別選取法是指若屬性 X，在第 Y 組訓練資料中的重要程度低於門檻值時，於建立第 Y 組訓練測試資料的模型時，屬性 X 不予列入模型的輸入。例如屬性 Value 在第一組訓練資料中的重要程度為 0.038，小於 0.05，故建立第一組模型時不考慮屬性 Value 為輸入。但屬性 Value 在其餘的訓練資料中皆有超過門檻值，故建立其餘模型時，考慮屬性 Value。因此，第一組訓練資料不考慮屬性 Value、Reason 及 Clno。其餘的訓練資料則不考慮屬性 Reason 及 Clno。

3.4 處理偏態分佈的屬性

由於 HMEQ 資料集中有許多屬性都有偏態分佈的情況，我們希望降低偏態分佈所造成的影響，故此在挑選重要屬性前，對於偏態屬性取 $\log_{10}$ 函數，目的在於降低偏態屬性中數值差異過大的問題。取 $\log_{10}$ 的偏態屬性包括 Loan、Mortude、Value、Yoj、Clage、Debtinc。而 Derog、Delinq、Ninq、Clno 等屬性的最大值和最小值差異不如 Loan、Mortude、Value、Yoj、Clage、Debtinc 等欄位大。例如: Clno 屬性中的最大值為 71，取 $\log_{10}$ 後為 1.85 (取到小數點後第二位)，且這些欄位的最小值為 0，也不適合取 $\log_{10}$。

表七為屬性 Loan、Mortude、Value、Yoj、Clage、Debtinc 取完 $\log_{10}$ 後的重要屬性程度分析。與表七比較後，發覺重要屬性除了 Debtinc 外，Loan、

表九:分類模型在四組實驗下之平均 F-measure

資料集		訓練資料				測試資料			
		C5.0		C&RT		C5.0		C&RT	
		AvgF0	AvgF1	AvgF0	AvgF1	AvgF0	AvgF1	AvgF0	AvgF1
未處理偏態屬性	個別選取法	94.21	71.694	91.664	53.854	92.62	63.302	91.514	52.78
	投票法	94.38	72.446	91.664	53.854	92.644	63.542	91.514	52.78
處理偏態屬性	個別選取法	95.00	78.596	91.664	53.854	93.062	69.772	91.514	52.78
	投票法	95.00	78.596	91.664	53.854	93.062	69.772	91.514	52.78

Mortude、Value、Yoj、Clage 等屬性的重要程度皆提升了，而其它屬性的重要程度則不變。需留意的是由於數值與類別屬性重要程度的計算方式不同，故無法比較數值屬性與類別屬性的重要程度。

3.5 分類模型的選擇與評估方法

在挑選完重要屬性後，接著針對每一組訓練資料建立其分類模型。本研究以 C5.0 決策樹與 C&RT 決策樹作為分類的演算法。分類的結果以混亂矩陣(Confusion Matrix)的方式呈現，如表八。我們依據混亂矩陣分別計算出準確率(Precision)與捕捉率(Recall)並求得 F-measure，並以 F-measure 為主要的評估方式。

表八: 混亂矩陣

	預測為好客戶	預測為壞客戶
實際為好客戶	A	B
實際為壞客戶	C	D

好客戶的預測準確率: $P0 = \frac{A}{(A+C)}$ 公式(6)

好客戶的預測捕捉率: $R0 = \frac{A}{(A+B)}$ 公式(7)

好客戶的 F-measure: $F0 = \frac{(2*P0*R0)}{(P0+R0)}$ 公式(8)

壞客戶的預測準確率: $P1 = \frac{D}{(B+D)}$ 公式(9)

壞客戶的預測捕捉率: $R1 = \frac{D}{(C+D)}$ 公式(10)

壞客戶的 F-measure: $F1 = \frac{2*P1*R1}{(P1+R1)}$ 公式(11)

針對五組訓練資料分別建立模型後，會得到五組好客戶的 F-measure，分別為 $F0_1$、$F0_2$、$F0_3$、$F0_4$、$F0_5$。我們加總這五組好客戶的 F-measure，並除以 5，以計算出整體模型中好客戶的平均 F-measure。同理，計算出壞客戶的平均 F-measure。

好客戶的平均 F-measure:

$$AvgF0 = \frac{(F01+F02+F03+F04+F05)}{5}$$ 公式(12)

壞客戶的平均 F-measure:

$$AvgF1 = \frac{(F11+F12+F13+F14+F15)}{5}$$ 公式(13)

4.實驗結果

本研究針對五組訓練-測試資料，填補完空值後，依據是否處理偏態屬性與不同的重要屬性選取法，而作了以下四組的實驗。第一組:採用個別選取法選擇重要屬性，且不處理偏態屬性。第二組:採用投票選取法選擇重要屬性，且不處理偏態屬性。

第三組:採用個別選取法選擇重要屬性，並處理偏態屬性。第四組:採用投票選取法，並處理偏態函數。最後分別以 C5 與 C&RT 為分類模型。實驗的結果如表九所示。

第一組實驗顯示，C5.0 在訓練資料的平均 F-measure 與測試資料的平均 F-measure 皆優於 C&RT。第二組實驗顯示，除了 C5.0 在訓練資料的平均 F-measure 與測試資料的平均 F-measure 皆優於 C&RT 外，可發覺採用投票選取法挑選重要屬性，提升了 C5.0 在訓練及測試資料的平均 F-measure，而 C&RT 則沒有改變預測能力。第三組實驗處理偏態屬性，並用個別選取法，發覺 C5.0 決策樹的效能優於第一組與第二組訓練資料所建的模型。第四組實驗處理偏態屬性，並使用投票法選取重要屬性，實驗結果與第三組實驗相同。理由是因為第四組實驗在採用投票選取法後，所得到的重要屬性與第三組實驗相同，除了 Clno 及 Reason 屬性外，其餘的屬性皆採用，故此建立出來的模型也相同。由以上四組實驗得知，採用投票選取法選取重要屬性使 C5.0 效能提升，而對偏態屬性取 $\log_{10}$ 具有使模型穩定的功能，並同時提升 C5.0 的分類效能。

5.結論與未來工作

本研究將資料探勘中資料前處理、欄位篩選等技術應用於預測房貸風險上。它幫助銀行利用分類模型預測好壞客戶，掌握客戶的房貸風險。本研究將原始資料庫，搭配分層抽樣並以五疊交互驗證法為基礎，將資料庫建立成五組模型，並於此架構下提出個別選取法與投票選取法挑選重要屬性。此架構相較於過去只採用一組模型評估風險，更為客觀。由於有違約的貸款申請者，與無違約的貸款申請者在資料庫中的比例往往呈現不平衡的分佈狀態。未來工作將朝向處理目標屬性值分佈不平衡的資料庫，並加以提升少數類別的預測能力，並以本研究的處理架構為基礎，延伸至其他貸款等問題，幫助銀行控管房貸、車貸及信用貸款等風險。

誌謝

這篇論文是國科會計劃(NSC 96-2221-E-130-017 & NSC 96-2221-E-130-005)研究成果的一部份。我們在此感謝國科會經費支持這項計劃的研究。

參考文獻

1. Han, J. and Kamber, M. (2000), Data Mining - Concepts and Techniques, Morgan Kaufmann Publishers, 2000.
2. Lee, Y. S. and Yen, S. J. (2002), “Neural-Based Approaches for Improving the Accuracy of Decision Trees,” Lecture Notes in Computer Science (LNCS): Data Warehousing and Knowledge Discovery, 2454, September 2002, pp. 114-123.
3. Lee, Y. S. and Yen, S. J. (2004), “Classification Based on Attribute Dependency,” Lecture Notes in Computer Science (LNCS): Data Warehousing and Knowledge Discovery, 3181, September 2004, pp. 259-268.
4. Lee, Y. S., et al. (2004), “A Data Mining Approach to Constructing Credit Risk Scoring Model,” Proceedings of 10th Conference on Information Management and Practice, 2004, pp. 1799-1813.
5. Quinlan, J. R. (1996), “Improved Use of Continuous Attributes in C4.5,” Journal of Artificial Intelligence Approach, 4, 1996, pp. 77-90.
6. Wang, K., Zhou, S. Q. and He, Y. (2000), “Growing Decision Trees on Support-Less Association Rules,” Proceedings of International Conference on Knowledge Discovery and Data Mining, 2000, pp .265-269.