資料探勘在汽車保險預測模型上之研究

李御璽[1]　顏秀珍[2]　羅隆晉[3]　吳政瑋[4]　粘嘉菖[5]

[1]銘傳大學資訊工程學系 {leeys@mail.mcu.edu.tw}
[2]銘傳大學資訊工程學系 {sjyen@mail.mcu.edu.tw}
[3]銘傳大學資訊工程研究所 {masatotaiki@gmail.com}
[4]銘傳大學資訊工程研究所 {silvemoonfox@hotmail.com}
[5]銘傳大學資訊工程研究所 {ruaddick@gmail.com}

摘要

汽車為今日社會不可或缺的交通運輸工具。隨著國民所得的提昇，汽車的數量因而大量的增加。和其它的保險一樣，汽車保險乃是累積社會大眾的保險資源，將嚴重損失合理分攤，使意外事故所造成的影響降至最低，間接減少社會問題與社會成本的一項工具。汽車保險的保費高低，更是與其保險後是否會來理賠有著密切的關係。有鑑於此，本研究運用資料探勘(Data Mining)中的分類(Classification)技術與汽車保險資料集(Car Insurance Dataset)，建構一個客戶在保險後是否會來保險理賠的分類預測模型。本研究首先處理資料中有離群值(Outlier)的屬性，並提出一個屬性選擇(Attribute Selection)的方法，先將對建構分類模型不重要的屬性刪除，再進行資料探勘。實驗的結果顯示，本研究所提出的屬性選擇方法不需借助任何統計軟體，在實務應用上略勝一籌。

關鍵詞：屬性選擇、汽車保險、分類、資料探勘

1. 緒論

資料探勘(Data Mining)意指從大量資料中去尋找有用或潛在的資訊或知識(Han and Kamber, 2006)、(Witten and Frank, 2005)、(Berry and Linoff, 2004)、(Roiger and Geatz, 2003)。近年來資料庫技術日益成熟，並廣泛的運用在各個領域之中。本研究運用資料探勘中的分類(Classification)技術(Anthony and Bartlett, 2002)、(Quinlan, 1996)、(Chauvin and Rumelhart, 1995)、(Quinlan, 1993)、(Lee, Yen and Wu, 2006)、(Lee and Yen, 2004)與汽車保險資料集(Car Insurance Dataset)來建構一個分類預測模型，以預測一個客戶在保險後是否會來理賠。

從過去的研究中得知，分類模型的效能會受到資料 (Data)、屬性 (Attribute)、以及分類演算法 (Classification Algorithm)的影響。若資料品質差(有空值(Null Values)、錯誤值(Wrong Value)或離群值(Outlier))，容易造成分析結果不正確(Pyle, 1999)。此外，重要屬性是否收集完整，以及參數設定是否正確，都是影響分類模型正確率的關鍵(Lee, et al., 2008)、(Lee, et al., 2007)。因本研究所取得的汽車保險資料集，其資料品質相當優良，沒有空值，且相關的屬性也收集地相當完整，故在此資料集上，本研究的重點就放在篩選重要屬性進入分類模型中，以提高模型的預測能力，使分類的結果最佳化。

本論文的章節安排如下：下一節將介紹本研究所使用的汽車保險資料集與資料集中各欄位的意義。第三節將說明本研究對於離群值的分析及處理方式，並詳述本研究所提出的屬性重要程度分析法，最後更與以統計為基礎的屬性選擇方法做比較。做結論之前，本研究將以實驗來驗證屬性選擇的重要性。

2. 汽車保險資料集

本研究利用汽車保險資料集作為分析的對象。整個資料集中有 1,200 位客戶資料。每筆客戶資料包含 1 個目標屬性(Target Attribute)，7 個類別型條件屬性(Categorical Input Attribute)以及 4 個數值型條件屬性(Numerical Input Attribute)。本研究將客戶分為兩類：一為好客戶-無理賠紀錄的客戶，有 319 筆資料；另一為壞客戶-有理賠紀錄的客戶，有 881 筆資料。以下將針對此 11 個條件屬性及 1 個目標屬性所代表的意義作詳細的描述。

屬性 1：sex (類別屬性)：性別
屬性 2：age (數值屬性)：年齡
屬性 3：marrage (類別屬性)：婚姻狀態
屬性 4：job(類別屬性)：職業
屬性 5：income (數值屬性)：月收入
屬性 6：child_cnt (類別屬性)：家中小孩數級距
屬性 7：region(類別屬性)：居住區域
屬性 8：distance (數值屬性)：通勤距離
屬性 9：car_type (類別屬性)：車型
屬性 10 : car_age(數值屬性) : 車齡
屬性 11：parking_type (類別屬性)：停車型態
屬性 12 : claim_cnt(目標屬性) : 保險後是否有理賠

為了訓練及測試模型的成效，本研究將這 1,200 位客戶的資料，以分層抽樣的方式(從 319 位沒有理賠紀錄及 881 位有理賠紀錄的客戶資料中，各取近 80%作為訓練資料，剩下的 20%作為測試資料)將原始資料分成 931 筆訓練資料及 269 筆測試資料。整個訓練及測試資料的分佈如下所示：

訓練資料：931 筆(240 筆(25%)無理賠紀錄，691 筆(75%)有理賠紀錄)
測試資料：269 筆(079 筆(63%)無理賠紀錄，190 筆(37%)有理賠紀錄)

3. 研究方法

此節將說明本研究對於離群值的處理方式，並詳述本研究所提出的屬性重要程度分析法，以及以統計為基礎的屬性重要程度分析法。

3.1 屬性離群值分析

本研究所使用的汽車保險資料集中，沒有屬性有空值及失誤值的情況產生，但某些數值屬性中卻存在著離群值。本研究採用統計學中「信賴區間」的概念(Moore, et al., 2003)。以平均值 ± 3 × 標準差當做正常值的方式，來計算出離群值。表一為分析原始資料的結果。

表一: 屬性離群值分析表

屬性	最小值	最大值	平均值	標準差	離群值
age	9	70	39.04	10.538	小於 7.426 或大於 70.654
income	15862	193000	68966.78	35714.571	小於-38176.9 或大於 176110.5
distance	0	42.849	14.044	7.061	小於-7.139 或大於 35.227
car_age	0.009	111.44	3.412	6.603	小於-16.397 或大於 23.221

由表一可以知道只有 age 屬性沒有離群值，而 income、distance 和 car_age 屬性均有離群值存在。為了不讓離群值對往後分析造成影響，我們採用天花板/地板的方法，將大於/小於平均值加/減 3 倍標準差的離群資料，修改為平均值加/減 3 倍標準差。

3.2 屬性重要程度分析

本研究在進行分類前，會對所有的屬性進行分析，並選擇重要且具有區分能力的屬性作為分類系統之輸入。若將不重要的屬性輸入至分類模型中，除了會拉長模型訓練的時間，也會降低分類模型的預測能力。傳統挑選重要屬性的方式是利用人的經驗及直覺來篩選屬性，但是這種方式往往缺乏科學數據支持。因此本研究提出一個有系統的屬性分析方法，藉由系統化方式給定每個屬性一個重要性分數，最後再利用人工來篩選重要的屬性，以使日後分析結果更加客觀。本研究在下兩個小節中，將詳述如何分別計算類別屬性及數值屬性的重要程度。

3.2.1 類別屬性分析方法

針對類別屬性分析方法，以下我們用 marriage 類別屬性為例，其結果列於表二。

表二: 類別屬性 marriage 分析結果

屬性名稱	屬性值	無理賠紀錄			有理賠紀錄			屬性值重要程度
		頻率	支持度	信賴度	頻率	支持度	信賴度	
marriage	單身	50	0.156	0.509	133	0.15	0.49	0.002
	已婚	231	0.724	0.526	574	0.651	0.473	0.035
	離婚	38	0.119	0.376	174	0.197	0.623	0.043
0.082	加總	319			881			0.082

在表二中，marriage 屬性具有 3 個屬性值，分別為單身、已婚、離婚。本研究分別分析其在有理賠紀錄及無理賠紀錄兩類別群組中的資料分佈。頻率代表出現次數，支持度則表示頻率在有理賠或無理賠紀錄類別群組中出現之機率。例如，單身屬性值在無理賠紀錄類別群組中出現 50 次，因此它在該類別群組中所佔之機率(支持度)為 50/319= 0.156。而單身屬性值在有理賠紀錄類別群組中出現 133 次，因此它在該類別群組中所佔之機率(支持度)為 133/881=0.15。此外本研究透過信賴度的計算來判斷屬性值是偏向何

種類別屬性。計算方式為無理賠紀錄和有理賠紀錄個別支持度除以兩支持度總和。例如，單身在有理賠紀錄中的支持度為 0.15，在無理賠紀錄中的支持度為 0.156，所以單身在有理賠紀錄中的信賴度為 0.15/(0.15+0.156)=0.509；在無理賠紀錄中的信賴度為 0.156/(0.15+0.156)=0.49，因此單身屬性值是偏向無理賠紀錄類別(因為 0.509> 0.49)。

基於信賴度的計算，最右邊欄位的數值代表屬性值重要程度，其計算公式為：ABS(無理賠信賴度-有理賠信賴度)*((無理賠頻率+有理賠頻率)/總人數)，其中 ABS 為取絕對值的函數。例如，單身屬性值重要程度為 0.002=ABS(0.509-0.49)*((50+133)/1200)。此公式含意為信賴度差異越大，屬性值所佔的總比例越高，其重要性越高。最後，本研究將所有屬性值重要程度加總來代表此屬性重要程度。因此，marriage 屬性重要程度為 0.082= 0.002+0.035+0.043。本研究運用相同的計算方式處理其他類別屬性。表三列出所有類別屬性的重要程度與排名。

表三: 所有類別屬性的重要程度

屬性值	重要程度	排名
car_type	0.177	1
sex	0.112	2
job	0.106	3
region	0.099	4
marriage	0.082	5
child_cnt	0.079	6
parking_type	0.073	7

由表三觀察得知，region、marriage、child_cnt 和 parking_type 4 個屬性鑑別力較低，所以本研究將此四個屬性進行過濾，只保留 car_type、sex 和 job 這三個鑑別力較高的屬性來訓練分類模型，以訓練出正確率較高的分類模型。

3.2.2 數值屬性分析方法

這一小節中本研究將詳述數值屬性的分析方法。表四為針對 income 數值屬性進行分析之結果。

表四: 數值屬性 income 的分析結果

屬性名稱	無理賠紀錄		有理賠紀錄		屬性重要程度
	平均值	標準差	平均值	標準差	
income	58799.25	21346.894	72581.861	38817.45	0.458

在表四中，本研究分別統計無理賠紀錄及有理賠紀錄兩類別群組的平均值(Mean)及標準差(Standard Deviation)。屬性重要程度為判斷屬性是否顯著依據，其計算公式為：

ABS(無理賠平均值-有理賠平均值)/((無理賠標準差+有理賠標準差)/2)，其中 ABS 為取絕對值的函數。例如，INCOME 的重要程度為 0.458=ABS(58799.25-72581.861)/((21346.894+38817.45)/2)。因此本研究可從每個屬性顯著程度，得知哪些屬性對理賠紀錄會有顯著影響，並根據此數值予以排名。排名結果如表五所示。

表五: 所有數值屬性的重要程度

屬性值	重要程度	排名
income	0.458	1
distance	0.146	2
car_age	0.104	3
age	0.044	4

由表五得知 age 屬性的鑑別力較低，所以本研究將其過濾，以訓練出正確率較高的分類模型。

3.3 以統計為基礎的屬性重要程度分析

在屬性重要程度分析法中，針對類別屬性的重要程度，傳統上可用統計分析中的卡方檢定進行相關性檢定。而數值屬性資料可用 T 檢定與 ANOVA 檢定進行重要屬性分析。以下就這些方式來比較傳統以統計為基礎的方法與我們所提方法之差異。表六顯示我們的方法與卡方檢定在類別屬性重要程度計算上之差異。

表六: 我們的方法與卡方檢定在類別屬性重要程度計算上之差異

屬性值	我們方法 重要程度	我們 排名	卡方檢定 卡方值	卡方檢定 P-value	卡方 排名
car_type	0.177	1	47.726	0.000	1
sex	0.112	2	11.755	0.001	4
job	0.106	3	17.052	0.002	2
region	0.099	4	13.570	0.001	3
marriage	0.082	5	10.036	0.007	5
child_cnt	0.079	6	8.709	0.069	7
parking_type	0.073	7	6.609	0.037	6

要利用卡方檢定去計算各個類別型的條件屬性與目標屬性間的相關程度，首先必需先計算其卡方值，然後根據卡方值查表算出 P-value。P-value 越小代表此屬性與目標屬性間的相關程度越大，也越重要。在 95%的信心水準之下，P-value<0.05 則為顯著(重要)的屬性。由表六可以看出，卡方檢定只會過濾掉 child_cnt(小孩個數)這個屬性。我們的

方法則選擇前 3 個屬性作為重要的屬性。

由於樣本變異數是否與母體相同會影響 T-Value 之計算，因此要利用 T 檢定去計算各個數值型的條件屬性與目標屬性間的相關程度，首先，我們必需利用 F 檢定去檢驗每個屬性的樣本變異數是否與母體相同。表七顯示 F 檢定後的結果。

表七：F 檢定後的結果

屬性值	F檢定F-value	F檢定P-value
income	103.811	0.000
distance	10.657	0.001
car_age	1.160	0.282
age	13.413	0.000

在表七中，P-value 越小代表屬性的樣本變異數與母體變異數越有差距。由表七可以看出，在 95%的信心水準下，只有 car_age 的樣本變異數與母體變異數無差異。其他皆有差異。根據表七的結果，表八的 T-value 即可計算出來。

表八: 我們的方法與 T 檢定在數值屬性重要程度計算上之差異

屬性值	我們方法 重要程度	我們 排名	T 檢定 T-value	T 檢定 P-value	T 檢定 排名
income	0.458	1	-7.771	0.000	1
distance	0.146	2	2.318	0.021	2
car_age	0.104	3	-1.580	0.114	3
age	0.044	4	-7.000	0.484	4

接著根據 T-value 查表算出 P-value。P-value 越小代表此屬性與目標屬性間的相關程度越大，也越重要。在 95%的信心水準之下，P-value<0.05 則為顯著(重要)的屬性。由表八可以看出，T 檢定會保留 income 和 distance 的屬性，比我們的方法少一個。表九顯示我們的方法與 ANOVA 檢定在數值屬性重要程度計算上之差異。

表九: 我們的方法與 ANOVA 檢定在數值屬性重要程度計算上之差異

屬性值	我們方法 重要程度	我們 排名	ANOVA F-value	ANOVA P-value	ANOVA 排名
income	0.458	1	36.186	0.000	1
distance	0.146	2	4.655	0.031	2
car_age	0.104	3	2.495	0.114	3
age	0.044	4	0.446	0.504	4

要利用 ANOVA 檢定去計算各個數值型的條件屬性與目標屬性間的相關程度，首先必需先計算其 F-value，然後根據 F-value 查表算出 P-value。P-value 越小代表此屬性與目標屬性間的相關程度越大，也越重要。在 95%的信心水準之下，P-value<0.05 則為顯著(重要)的屬性。由表九可以看出，ANOVA 和 T 檢定所保留的屬性相同，仍然比我們的方法少一個。若不管信心水準，只管排名，本研究所提的方法與 T 檢定的順序大致相同。因此，本研究所提的方法與傳統以統計為基礎的方法差異不大。然而，本研究所提的方法簡單、易懂，而且不需借助任何統計的軟體，在執行的效能上略勝一籌。

4. 實驗結果

在實驗上，本研究將資料分成訓練資料(約 80%)及測試資料(約 20%)，其資料的分佈情形如第二節所示。根據第三節的屬性分析方法，本研究最後選擇 3 個類別屬性和 3 個數值屬性來訓練分類模型。這 6 個屬性分別為:

屬性 1：sex (類別屬性)：性別
屬性 2：job(類別屬性)：職業
屬性 3：income (數值屬性)：月收入
屬性 4：distance (數值屬性)：通勤距離
屬性 5：car_type (類別屬性)：車型
屬性 6：car_age(數值屬性)：車齡

在分類效能的評估上，對於分兩類之分類系統時，通常會將分類的結果建立如表十的混亂矩陣(Confusion Matrix)，並利用四種指標(Van Rijsbergen, 1979)、(Ricardo and Berthier, 1999)、(Lee, et al., 2005)：正確率(Accuracy)、精確率(Precision)、捕捉率(Recall)及 F-Measure 來評估一個分類系統的好壞。其中，AA 代表系統預測為無理賠紀錄的客戶且實際也為無理賠紀錄的客戶的人數；AP 代表系統預測為有理賠紀錄的客戶但實際為無理賠紀錄的客戶的人數；PA 代表為系統預測為無理賠紀錄的客戶但實際為有理賠紀錄的客戶的人數；PP 代表為系統預測為有理賠紀錄的客戶且實際為有理賠紀錄的客戶的人數。

表十: 混亂矩陣及其預測值

	預測為無理賠紀錄的客戶	預測為有理賠紀錄的客戶
實際為無理賠紀錄的客戶	AA	AP
實際為有理賠紀錄的客戶	PA	PP

這四種指標的公式如下所示：

1. 分類的正確率 CA =(AA+PP)/(AA+AP+PA+PP)
2. 無理賠紀錄客戶的精確率 P = AA /(AA + PA)

3. 無理賠紀錄客戶的捕捉率 R = AA/(AA+AP)
4. 無理賠紀錄客戶的 F-Measure =(2*P*R)/(P+R)

由於本研究主要是對預測無理賠紀錄客戶的結果有興趣(因為無理賠的客戶，對公司而言獲利較大)，因此在第 2, 3, 4 項指標中，本研究僅針對無理賠客戶的情形去評估。在實驗上，本研究使用 SPSS Clementine 中的分類演算法來建構分類模型，並比較我們所提出的方法與傳統以統計為基礎的方法所挑選的屬性，在分類效能上的差異。表十一為模型在測試資料上的分類結果。

表十一: 模型效能比較

分類演算法	評估指標	我們的方法挑選的欄位	統計的方法挑選的欄位
決策樹(C&RT)	分類正確率	75.40%	74.97%
	無理賠紀錄客戶的 F-Measure	0.1965	0.1402
類神經網路	分類正確率	73.58%	74.22%
	無理賠紀錄客戶的 F-Measure	0.23	N/A
羅吉斯迴歸	分類正確率	73.90%	74.22%
	無理賠紀錄客戶的 F-Measure	N/A	0.244

本研究共使用 3 種分類演算法來建立分類預測模型。這 3 種方法分別為決策樹(Decision Tree)-C&RT、類神經網路(Neural Network)以及羅吉斯回歸(Logistic Regression)。

C&RT 是一種建構二元(Binary)分岔的分類迴歸樹演算法。它在 1984 年由 Breiman, Friedman, Olshen 與 Stone 四人所提出(Breiman, Friedman, Olshen and Stone, 1984)。決策樹是採用樹狀分岔的架構來產生分類規則。其建置流程是以全部的訓練資料作為根節點，再根據最佳變數產生分岔並產生子節點。決策樹按照此方式持續生長，並利用修剪技術來移除不必要的分支。最後，再根據每個子節點案例的分佈狀況，指派其分類結果及信賴度。決策樹的缺點為當訓練資料太多而需要取樣時，其分類預測的機率(信賴度)會因只取部分樣本的原故而無法準確的預測。

類神經網路(Artificial Neural Network)是一種計算系統，它使用大量簡單相連的人工神經元來模擬生物神經網路的能力。人工神經元從外界環境或其它人工神經元取得資訊，之後利用簡單的運算，並輸出結果到外界環境或其它人工神經元。若依網路架構分類，則可區分為前向式(Forward)及回饋式(Feedback)二種架構。在本論文中，我們採用前向式架構的倒傳遞類神經網路(Backpropagation Neural Network, BPN)(Anthony and Bartlett, 2002)(Chauvin and Rumelhart, 1995)作為我們分類的演算法。類神經網路的缺點為其分類預測的過程就像是黑盒子(Black Box)，使得模型無法以一般所能理解的方式說明其決策過程。

羅吉斯迴歸(Logistic Regression)是離散選擇法模型之一，屬於多重變數分析範疇，

是社會學、生物統計學、臨床、數量心理學、市場營銷等統計實證分析的常用方法。

由表十一的實驗結果可觀察出，決策樹(C&RT)中之分類正確率和無理賠客戶的F-Measure 用我們的方法挑選屬性時最好。而類神經網路中之分類正確率用統計方法挑選屬性時最好，無理賠客戶的 F-Measure 用我們的方法挑選屬性時最好。另外，在羅吉斯迴歸中之分類正確率和無理賠客戶的 F-Measure 則用統計方法挑選屬性時最好。總結以上實驗結果，本研究所提挑選屬性的方法與傳統以統計為基礎的方法，其分類效能差異不大。但從效益上來說，本研究所提的方法簡單、易懂，且不需借助任何的統計軟體，在實務應用上略勝一籌。

5. 結論與未來工作

本研究以汽車保險資料集為實驗對象，找出無理賠紀錄之重要因素。透過屬性重要性分析和分類模型效能評估的協助，使本研究可以客觀地觀察實驗的過程及結果。本研究以汽車保險資料集為例，未來希望能將本研究所提出的方法運用至更廣的領域中，可為本研究帶來更高的附加價值。此外，在挑選重要屬性方面，未來也將與現有已發展的方法進行相互比較，以強化研究結果的說服力。

誌謝

這篇論文是國科會計劃(NSC 97-2221-E-130-012 & NSC 97-2221-E-130-013)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. Anthony, M. and Bartlett, P. L. Neural Network Learning: Theoretical Foundations, Cambridge University Press, 2002.
2. Berry, M. J. A. and Linoff, G. S. Data Mining Techniques for Marketing, Sales and Customer Relationship Management, Wiley Publishers, Second Edition, 2004.
3. Breiman, L., Friedman, J., Olshen, R. and Stone, C. “Classification and Regression Trees”, Wadsworth International Group, 1984.
4. Chauvin, Y. and Rumelhart, D. Backpropagation: Theory, Architectures and Applications, Lawrence Erlbaum Associates, 1995.
5. Han, J. and Kamber, M. Data Mining - Concepts and Techniques, Morgan Kaufmann Publishers, Second Edition, 2006.
6. Lee, Y. S., et al. “Data Mining on Credit Risk Prediction of Mortgage”, Proceedings of International Conference on Information Management, 2008.
7. Lee, Y. S., et al. “The Study of Customer-Oriented Knowledge Discovery Technology - Loan of Small Medium Enterprise”, Proceedings of Conference on Electronic Commerce and Digital Life, 2007, pp. 228-237.
8. Lee, Y. S., et al. “Performance Evaluation on a Classification System”, Proceedings of Conference on Artificial Intelligence and Applications, 2005.

9. Lee, Y. S. and Yen, S. J. “Classification Based on Attribute Dependency”, Proceedings of International Conference on Data Warehousing and Knowledge Discovery, 2004, pp. 259-268.
10. Lee, Y. S., Yen, S. J. and Wu, Y. C. “Using Neural Network Model to Discover Attribute Dependency for Improving the Performance of Classification”, Journal of Informatics and Electronics (1:1), 2006, pp. 9-19.
11. Moore, D. S., et al. The Practice of Business Statistics: Using Data for Decisions, W. H. Freeman & Co, 2003.
12. Pyle, D. Data Preparation for Data Mining, Morgan Kaufmann Publishers, 1999.
13. Quinlan, J. R. “Improved Use of Continuous Attributes in C4.5”, Journal of Artificial Intelligence Approach (4), 1996, pp. 77-90.
14. Quinlan, J. R. C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, 1993.
15. Ricardo, B. Y. and Berthier, R. N. Modern Information Retrieval, Addison-Wesley, 1999.
16. Roiger, R. J. and Geatz, M. W. Data Mining: A Tutorial-Based Primer, Addison- Wesley, 2003.
17. Van Rijsbergen, C. J. Information Retrieval, Butterworths, 1979.
18. Witten, I. H. and Frank, E. Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann Publishers, Second Edition, 2005.

Data Mining on Predicting the Claims in Car Insurance

Yue-Shi Lee[1] Show-Jane Yen[2] Lung-Jin Ro[3] Cheng-Wei Wu[4] Chia-Chang Nien[5]

[1] Department of CSIE, Ming Chuan University {leeys@mail.mcu.edu.tw}
[2] Department of CSIE, Ming Chuan University {sjyen@mail.mcu.edu.tw}
[3] Department of CSIE, Ming Chuan University {masatotaiki@gmail.com}
[4] Department of CSIE, Ming Chuan University {silvemoonfox@hotmail.com}
[5] Department of CSIE, Ming Chuan University {ruaddick@gmail.com}

Abstract

Cars are transports which are indispensable to the modern society. With the increase of G.D.P., the numbers of cars are on the increase, too. Like the other coverage, the car insurance accumulates the insurance resources from the public, which share the risk reasonably and lower the influence on the cost caused by an accident. Simultaneously, this tool indirectly reduces some social problems and social cost. The premium of car insurance is closely related to how to pay the compensation afterwards. In view of it, this research uses the classification technology of data mining and a car insurance dataset to build a classification model for predicting whether the customer will or not to pay the compensation afterwards. Our research firstly deals with the attributes which has the outlier value and then proposes an attribute selection method to filter out the unimportant attributes before building the mining model. The experimental results show that the proposed attribute selection method can be easily applied to the practical applications without using any statistical software.

Keyword: Attribute Selection, Car Insurance, Classification, Data Mining