應用資料挖掘於交通事故資料分析

[1]吳冠宏 [2]吳信宏 [3]郭廣洋
[1]臺中健康暨管理學院資訊科學與應用學系碩士研究生
[2]國立彰化師範大學企業管理學系副教授
[3]中央警察大學交通學系副教授

摘要

台灣的交通雖然越來越便利，然而交通事故仍是層出不窮，時常可見車禍的發生，至今如何改善交通安全的問題仍然是個十分重要的議題。隨著時代的進步與需要，資料庫中的資料量日趨龐大，如何從資料庫中找出有用的知識以解決問題更是受到重視。本篇論文將利用資料挖掘(Data mining)技術中的Two-Step分群法與K平均值演算法(K-Means method)來分析大量的車禍資料以進行分群工作，並利用關聯規則(Association rules)從其中找出有意義的資訊，提供建議或決策以減少車禍的發生。

一、研究動機與目的

近年來汽、機車的數量不斷上昇，隨之帶來的交通事故也日趨嚴重，根據行政院衛生署[1]統計資料指出，事故傷害在台灣地區仍位於第五大死亡原因，光是民國91年因為事故傷害而傷亡的人數就高達8千多人。依據高速公路年報[2]分析顯示，民國91年高速公路交通量成長約4.56%，全線交通事故之肇事率為0.003件／百萬車公里，死亡率為0.003人／百萬車公里，受傷率為0.003人／百萬車公里，另就肇事原因分析，以「駕駛不當」、「未保持行車安全間距」為首，共佔全年事故之36.21%；就肇事車種分析，大貨車與聯結車之事故比例相對於所佔交通組成仍屬偏高。

一般而言，道路類別與行駛其上之車輛數間具有某種正相關的關係，即當道路面積越大時行駛其上之車輛數越多，這會使發生道路交通事故的可能性更高，由於道路面積的變化會影響道路交通事故資料的統計基礎，因此進行道路交通事故分析時，需將此納入考量。而且歷年以來發生於省道之交通事故所肇致的死亡人數，居各類道路之冠，顯見在省道上發生的交通事故雖少於市區道路，但其所肇致之事故嚴重性卻高於市區道路[3]。

本研究將利用資料挖掘(Data mining)的技術，使用Two-Step分群法與K平均值演算法(K-Means method)來分析大量的車禍資料以進行分群工作，並利用關聯規則(Association rules)從其中找出有意義的資訊，從交通事故的資料中找出一些特徵，提供建議或決策以減少車禍的發生。

二、文獻探討

本研究文獻探討針對資料挖掘之定義、應用資料挖掘的文獻方面、應用交通事故分析的文獻方面與資料挖掘相關技術作探討。

2.1 資料挖掘之定義

Frawley等[4]對資料挖掘的定義是：「從資料庫中挖掘出不明確、前所未知以及潛在有用的資訊過程」。Grupe與Owrang [5]認為資料

探勘乃是從現存資料中剖析出新事實及發現專家們尚未知曉的新關係。而Cabena等[6]則解釋資料挖掘是將未知且有效的資訊從大型資料庫抽出的過程、並且將萃取出的有用資訊提供給主管做決策。Kleissner [7]則解釋為：「資料挖掘是一種新的且不斷循環的決策支援分析過程，它能夠從組合在一起的資料中，發現隱藏價值的知識，提供給企業專業人員參考」。綜合以上學者的解釋得知，資料挖掘是從龐大的資料庫中，利用資訊科技、統計方法等，找出潛在有價值的資訊並支援主管決策活動。

隨著時代不斷的演進，許多企業組織已經收集並儲存相當豐富的客戶、供應商、和商業伙伴的資料。但遺憾的是，卻沒有能力找出隱藏在資料中有價值的資訊，造成人工分析上的困難。因此這些企業組織無法將資料化為知識。在這種情況下，近年來已有愈來愈多的資料探勘方面的應用來支援企業決策活動[8]。

在應用於交通事故分析的文獻方面，陳敬明[9]透過肇事資料之蒐集分析依所建構之評定模式確認易肇事地點，分析其肇事原因，研擬改善對策，以減低肇事之發生機率及嚴重性。吳偉碩[10]利用逐步迴歸分析方法建立一適當之肇事率預測分析模式，針對易肇事地點找出主要的肇事原因並提出有效改善方案供決策者參考。魏開元[11]利用類神經網路，針對路口肇事與工程因素兩者間之相互關係進行探討，經訓練後之鑑別模式對各類型肇事之誤判率均低於15%，而平均誤判率則為6.07%；而數量預估模式之誤判率亦低於25%，平均值為6.67%。程銘鎮[12]以複迴歸分析來進行國道中山高速公路交通事故發生原因之分析探討，以瞭解肇事之基本原因並找出那些高危險群的路段與肇事因子。張敏亮、吳信宏與郭廣洋[13]針對交通事故的現場環境，利用關聯規則來挖掘當中隱含的資訊，以提供給相關部門，對於交通工程、道路環境設計或建造時一個參考，使得肇事發生率能獲得有效的改善，降低交通事故對於人民生命財產的威脅。

2.2資料挖掘相關技術

資料挖掘可運用的範圍相當的廣，然而不同的技術所針對的問題也不相同，本研究將利用以下的技術，包括Two-Step分群法、K平均值演算法與關聯規則，來進行集群分析與關聯規則之分析。

2.2.1 Two-Step 分群法

Two-Step 分群法是使用凝集聚群法的一種演算法，可以幫助我們處理大量數據資料的分群工作，並且能處理連續和某一範圍的變數或屬性。透過 Two-Step 分群分析，可以把數據資料歸類，因此在同一組內的記錄則會有較高相似度。Two-Step 分群法的演算步驟如下[14]：

步驟一：在預先分群階段根據群集距離接近的順序來分群，可以一筆一筆的瀏覽資料記錄並且決定目前的資料記錄是否應該和以前形成的群集合併，或是根據距離標準自己形成一個新的群集。

步驟二：這個分群階段採取來自預先分群階段的輸入和群集所產生子群集，將其加入所屬的群集。

2.2.2 K 平均值演算法

K 平均值演算法是在非階層式集群分析中最常被使用的方法。K 平均值演算法需事先決定集群的群數，由於選擇群數不當會造成其群與群的差異不明顯，故在選擇分群時，最好選擇不同的群數多做幾次演算，以得到合理的解釋，其分群演算步驟如下[15,16]：

步驟一：先決定要分成幾群，選擇種子來當群

集的中心點。

步驟二：將各別的每筆資料與每個中心點計算距離，看比較接近那一點，則視為該群組。

步驟三：將每一群的每筆資料平均，計算出新的中心點。

步驟四：再次利用算出的中心點將每筆資料重新分群。

步驟五：直到每筆資料的群組皆已固定，不會再分到其他的群組，則結束。

2.3 關聯規則

由於電腦廣泛應用和自動化的資料收集工具，大量的交易數據已經被收集並且儲存在資料庫裡。關聯規則可以讓我們從其中挖掘出有趣的規則來幫助行銷、決策，以及商業管理[17]。關聯規則主要是從大量的資料中，找出項目或屬性之間共同的關係，此法則的描述是「A→B」，即在交易集合之中，當 A 資料項出現時 B 資料項也會出現。關聯規則的評估指標，一般以支持度(Support)和信賴度(Confidence)為主。支持度即為，在規則中與項目一起出現的紀錄佔全部筆數的百分比，信賴度為此條規則的預測強度，為了避免過多無意義的規則，當支持度與信賴度高於所定義的門檻值時，該規則才成立。其計算方式如下：

$$\text{Support}(A\rightarrow B)=\frac{\#\text{records containing both A \& B}}{\text{total \# of records}} \quad (1)$$

與

$$\text{Confidence}(A\rightarrow B)=\frac{\#\text{records containing both A \& B}}{\#\text{records containing A}} \quad (2)$$

三、案例實證

3.1 案例資料說明

本研究資料來自於行政院內政部警政署所使用的道路交通事故調查表，此調查表為車禍發生時由警方人員所填寫的調查表格，具有民國 88 年度 A1、A2 類(A1 類係指造成人員當場或二十四小時內死亡之交通事故；A2 類係指造成人員受傷之交通事故)的車禍現場與發生事故人員的詳細資料，資料筆數為 22491 筆，資料型態為文字和數字，表格欄位包含天候、道路類別、路面狀況、損失金額等。

取得研究資料後，先進行資料的預先處理，由於資料之中可能存在者不完整或不一致的值，且為了避免資料遭極端值扭曲，故把異常樣本及極端值予以剔除並進行標準化，台灣快車道路數目多半為偶數，快車道路數為 2 表示左右各為一線道，快車道路數為 0 表示該路段無快車道，如產業道路、市區小巷道路，由於奇數的快車道路如 1、3、5、7 在台灣仍屬少見，故在本研究裡予以刪除，在資料之中仍發現極少數的交通事故發生在速限 200 以上的地點，此為極端值且明顯的不合常理，在這邊也予以刪除。

為了避免無效因子或干擾因子影響結果，故先進行列聯相關分析，分析資料是否具有顯著性。由於交通部推動「路權優先、安全第一」專案以確保民眾生命財產安全[18]，這表示民眾的生命和財產是十分重要的，除了寶貴的人命損失外，交通事故每年亦造成台灣嚴重的金錢損失，遠遠超過外界側目的一八八億罰單收入，嚴重侵蝕我國的經濟與生產力[19]，民眾在車禍事故當中的損失金額與車禍事故的嚴重性一般而言是相關的，通常損失金額的高低，會隨者車禍事故的嚴重性而改變，故將所有的資料欄位與損失金額做列聯相關分析，分析其

與損失金額是否有顯著相關。

一般上顯著水準會定在 0.05，而比較嚴格的會定在 0.01，本研究則採用較嚴格 0.01，在顯著水準達 0.01 時，表示該欄位與損失金額具有顯著的相關性，因此利用來進行分析的欄位有快車道數目、速限、道路類別、損失金額，如表一所示。

表一 選擇輸入的欄位

欄位名稱	分析欄位	卡方檢定值
快車道數目	快車道數目*損失金額	0.052 **
速限	速限*損失金額	0.329 **
道路類別	道路類別*損失金額	0.380 **
損失金額	損失金額*損失金額	1.000 **

**：在顯著水準為 0.01 時(雙尾)，相關顯著。

由於道路類別的資料屬於定性的資料，而多數的分群研究均需為定量的資料，故我們將利用 Ganti 等[20]提出的研究，對文字型態的資料做轉換。轉換主要是利用類別間其相關程度加以量化計算成為數值型態。因為損失金額與道路類別為顯著相關，且依常理判斷，在不同的道路類別的情況之下，其可能損失金額也不相同，故我們將利用損失金額來計算不同道路類別之間的相關性。

在道路類別中，其資料欄位有國道、省道、縣道、鄉道、市區道路、村里道路、專用道路共七種類別，以各道路類別的平均損失金額做為其之間的距離，並將其距離正規化，使其值轉換後介於零到一之間，其轉換公式為：道路類別轉換值 =(類別的平均損失金額-全部的平均損失金額)/(全部平均損失金額的最大值-全部平均損失金額的最小值)。道路類別的數值轉換後整理於表二。

表二 數值轉換表

	國道	省道	縣道	鄉道	市區道路	村里道路	專用道路
轉換前	1	2	3	4	5	6	7
轉換後	1	0.5068	0.4896	0.3486	0	0.2177	0.2015

3.2 分群與比較

本研究是利用 Clementine 分析軟體，來進行車禍事故資料的分群，以下使用 Two-Step 和 K-Means 分群法來進行分析，在非階層式集群分析中 K 平均值演算法則是最常被使用的方法，因為 Clementine 分析軟體提供 Two-Step 分群法，所以將該方法加入，利用這二種方法來進行分群的工作，並比較何種方法在車禍事故分析中較為合適。

3.3 Two-Step 模式分群結果

首先從 Two-Step 分群結果來看，在 Clementine 分析軟體中的 Two-Step 分群法，可以幫助我們選擇最佳的分群數並且進行分群，由表三中我們可以發現 Two-Step 分群法將資料分成二群，大部份的資料都分佈在群組一，只有少數的車禍事故分佈在群組二，其數字意義為該群組於該欄位所有資料加總後的平均數，如群組一的平均快車道數目為 1.367，表示群組一的平均快車道數目介於快車道數目 0 與 2 之間，道路類別 0.003 表示道路類別趨近於市區道路，而群組一的平均快車道數目為 1.367，趨近資料庫中所有事故件數的平均快車道數目 1.54；速限 42.096 低於資料庫中所有事故件數的平均速限 44.58；主要道路類別是市區道路；平均損失金額為 $ 13211.5 元低於資料庫中所

有事故件數的平均損失金額＄20686.2 元。群組一的快車道數目、速限、與損失金額均明顯低於群組二且趨近於所有資料的平均。

表三 Two-Step 分群法的分群結果

群組	件數	平均快車道數目	速限	道路類別	平均損失金額
1	18700	1.367	42.096	0.003	＄13211.5
2	3791	2.343	55.948	0.473	＄54926.8
總數	22491	1.54	44.58	0.09	＄20686.2

在群組二方面，平均快車道數目為 2.343，高於資料庫中所有事故件數的平均快車道數目 1.54；速限 55.948 高於資料庫中所有事故件數的平均速限 44.58；主要道路類別是省道或縣道；平均損失金額為＄54926.8 元，高於資料庫中所有事故件數的平均損失金額＄20686.2 元。群組二的快車道數目、速限、與損失金額均明顯高於群組一且高於所有資料的平均。

針對表三的結果，對每個群組做更詳細的分析可得到下面之結果。由表四到表七與圖一可以得知，群組一在平均快車道數目方面，0 個快車道數佔了群組一的 58.2%；在速限方面，速限低於 40 的佔了 86.2%，而在速限 40 情況下發生了 15912 次車禍，佔了群組二的 85.1%；在道路類別方面，僅市區道路就佔了 96.5%；在平均損失金額方面，損失低於＄10,000 元的件數佔了群組一的 79.9%。由平均快車道數目、速限、道路類別、與平均損失金額可以得知，在群組一車禍事故當中，具有快車道數目少、速限低、主要發生在市區道路且損失金額低的特性，故本研究將此群組定義為輕微事故。

由表四到表七與圖二可以得知，群組二在平均快車道數目方面，高於 0 個的快車道數佔了群組二的 70.5%；在速限方面，速限高於 50 的佔了 61.3%；在道路類別方面，國道與省道就佔了群組二的 59.3%；在平均損失金額方面，損失高於＄10,001 元的件數佔了群組二的 59.3%。由平均快車道數目、速限、道路類別與平均損失金額可以得知，在群組二車禍事故當中，具有快車道數目多、速限高、主要發生在國道與省道且損失金額高的特性，故本研究將此群組定義為嚴重事故。

表四 Two-Step 分群法平均快車道數目次數分配表

平均快車道數	群組一			群組二		
	件數	百分比	累積百分比	件數	百分比	累積百分比
0 以下	10879	58.2	58.2	1118	29.5	29.5
2	4236	22.7	80.8	958	25.3	54.8
4	2750	14.7	95.5	1339	35.3	90.1
6	608	3.3	98.8	271	7.1	97.2
8	227	1.2	100.0	105	2.8	100.0

表五 Two-Step 分群法的速限次數分配表

速限	群組一			群組二		
	件數	百分比	累積百分比	件數	百分比	累積百分比
30	204	1.1	1.1	47	1.2	1.2
40	15912	85.1	86.2	1420	37.5	38.7
50	1441	7.7	93.9	325	8.6	47.3
60	1097	5.9	99.8	1061	28.0	75.3
70	46	0.2	100.0	318	8.4	83.6
80	-	-	-	49	1.3	84.9
90	-	-	-	248	6.5	91.5

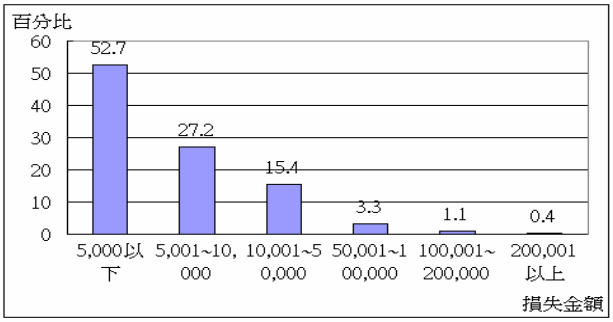
100	–	–	–	323	8.5	100.0

表六 Two-Step 分群法的道路類別數目次數分配表

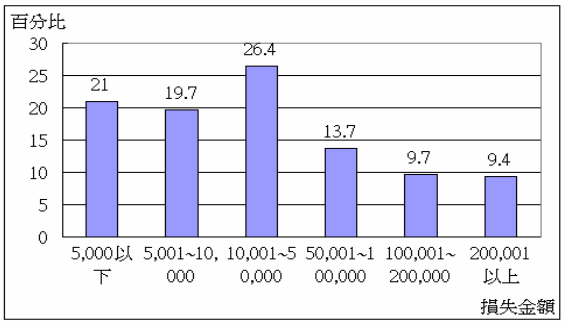
道路類別	群組一			群組二		
	件數	百分比	累積百分比	件數	百分比	累積百分比
國道	–	–	0.0	685	18.1	18.1
省道	–	–	0.0	1562	41.2	59.3
縣道	–	–	0.0	471	12.4	71.7
鄉道	81	0.4	0.4	103	2.7	74.4
市區道路	18039	96.5	96.9	659	17.4	91.8
村里道路	548	2.9	99.8	299	7.9	99.7
專用道路	32	0.2	100.0	13	0.3	100.0

表七 Two-Step 分群法的損失金額次數分配表

損失金額	群組一			群組二		
	件數	百分比	累積百分比	件數	百分比	累積百分比
5,000 以下	9863	52.7	52.7	797	21.0	21.0
5,001~10,000	5078	27.2	79.9	747	19.7	40.7
10,001~50,000	2871	15.4	95.3	1002	26.4	67.2
50,001~100,000	609	3.3	98.5	518	13.7	80.8
100,001~200,000	204	1.1	99.6	369	9.7	90.6
200,000 以上	75	0.4	100	358	9.4	100



圖一 Two-Step 分群法的群組一損失金額長條圖



圖二 Two-Step 分群法的群組二損失金額長條圖

在 Two-Step 分群法的分群結果中，我們可以發現群組一與群組二的區分並不是非常的明顯，有些項目的重疊性頗高。在平均快車道數目方面，群組一主要分佈於 0 與 2 個車道，而群組二主要分佈於 0、2 與 4 個車道，在二個群組中主要分佈有重疊的部份是 0 與 2 個車道；在速限方面，群組一的主要分佈於速限 40，而群組二主要分佈於速限 40 與 60，在二個群組中主要分佈有重疊的部份是速限 40；在道路類別方面，群組一主要分佈於市區道路，而群組二主要分佈於國道、省道與市區道路，在二個群組中主要分佈有重疊的部份是市區道路；在平均損失金額方面，群組一的主要分佈於 5,000 以下、5,001~10,000 與 10,001~50,000，而群組二主要分佈於 5,000 以下、5,001~10,000、10,001 ~50,000 與 50,001~100,000，在二個群

組中主要分佈有重疊的部份是 5,000 以下、5,001~10,000 與 10,001~50,000。

3.4 K-Means 模式分群結果

使用 K-Means 分群法時，需要事先訂定集群數目，我們為了與 Two-Step 分群法進行分群比較，故利用 Two-Step 分群法所決定的分群數來進行分群，而 Two-Step 分群法的分群數為二群集，故 K-Means 分群法的分群數選擇分成二群集，從 K-Means 分群法的分群結果來看，由表八中我們可以發現，只有少數的車禍事故分佈在群組一，而大部份的資料都分佈在群組二，而群組一的快車道數目為 4.20，高於資料庫中所有事故件數的平均快車道數目 1.54；速限 85.39 高於資料庫中所有事故件數的平均速限 44.58；主要道路類別是國道與省道；平均損失金額為＄73920.2 元，高於資料庫中所有事故件數的平均損失金額＄20686.2 元，群組一的快車道數目、速限、與損失金額明顯高於群組二且高於所有資料的平均。

在群組二方面，群組二的快車道數目為 1.42，趨近資料庫中所有事故件數的平均快車道數目 1.54；速限 42.71 低於資料庫中所有事故件數的平均速限 44.58；主要道路類別是市區道路；平均損失金額為＄18250.7 元，低於資料庫中所有事故件數的平均損失金額＄20686.2 元，群組二的快車道數目、速限、與損失金額明顯低於群組一且趨近於所有資料的平均。

表八 K-Means 分群法的分群結果

群組	件數	平均快車道數目	速限	道路類別	平均損失金額
1	984	4.20	85.39	0.74	$73920.2
2	21507	1.42	42.71	0.06	$18250.7
總數	22491	1.54	44.58	0.09	$20686.2

針對表八的結果，對每個群組做更詳細的分析可得到下面之結果。由表九到表十二與圖三可以得知，群組一在平均快車道數目方面，高於 4 個快車道數佔了群組一的 82.9%，而在車道數目 4 個的情況下發生了 509 次車禍，佔了群組一的 51.7%；在速限方面，速限至少在 70 以上的佔了 100%；在道路類別方面，僅國道就佔了 62.8%；在平均損失金額方面，損失高於＄10,001 元以上的件數佔了群組一的 75.7%。由平均快車道數目、速限、道路類別、與平均損失金額可以得知，在群組一車禍事故當中，具有快車道數目多、速限高、主要發生在國道與省道且損失金額高的特性，故本研究將此群組定義為嚴重事故。

由表九到表十二與圖四可以得知，群組二在平均快車道數目方面，2 個以下的快車道數佔了群組二的 79.2%，而在車道數目零個的情況下發生了 11864 次車禍，佔了群組一的 55.2%；在速限方面，速限在 60 以下的佔了 100%；在道路類別方面，僅市區道路佔了 86.4%；在平均損失金額方面，損失低於＄10,000 元的件數佔了群組二的 75.5%；由平均快車道數目、速限、道路類別與平均損失金額可以得知，在群組二車禍事故當中，具有快車道數目少、速限低、主要發生在市區道路且損失金額低的特性，故本研究將此群組定義為輕微事故。

表九 K-Means 分群法平均快車道數目次數分配表

平均快車道數目	群組一			群組二		
	件數	百分比	累積百分比	件數	百分比	累積百分比
0 以下	133	13.5	13.5	11864	55.2	55.2
2	35	3.6	17.1	5159	24.0	79.2
4	509	51.7	68.8	3580	16.6	95.8
6	216	22.0	90.8	663	3.1	98.9

表十 K-Means 分群法的速限次數分配表

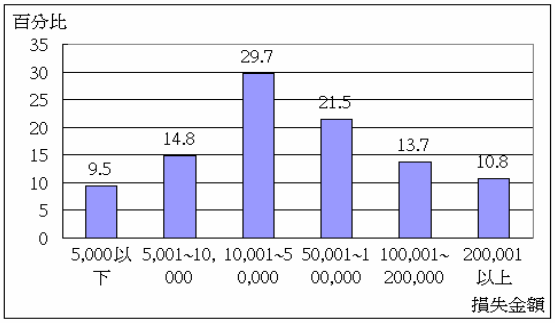
速限	群組一			群組二		
	件數	百分比	累積百分比	件數	百分比	累積百分比
30	-	-	-	251	1.2	1.2
40	-	-	-	17332	80.6	81.8
50	-	-	-	1766	8.2	90.0
60	-	-	-	2158	10.0	100.0
70	364	37.0	37.0	-	-	-
80	49	5.0	42.0	-	-	-
90	248	25.2	67.2	-	-	-
100	323	32.8	100.0	-	-	-

表十一 K-Means 分群法道路類別數目次數分配表

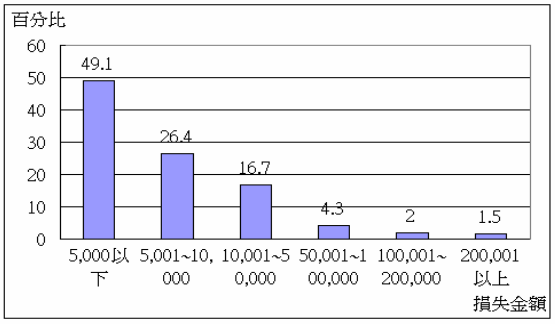
道路類別	群組一			群組二		
	件數	百分比	累積百分比	件數	百分比	累積百分比
國道	618	62.8	62.8	67	0.3	0.3
省道	200	20.4	83.2	1362	6.3	6.6
縣道	17	1.7	84.9	454	2.1	8.7
鄉道	2	0.2	85.1	182	0.8	9.5
市區道	138	14.0	99.1	18559	86.4	95.9
村里道	7	0.7	99.9	840	3.9	99.8

表十二 K-Means 損失金額次數分配表

損失金額	群組一			群組二		
	件數	百分比	累積百分比	件數	百分比	累積百分比
5,000 以下	93	9.5	9.5	10567	49.1	49.1
5,001~10,000	146	14.8	24.3	5679	26.4	75.5
10,001~50,000	292	29.7	54.0	3581	16.7	92.2
50,001~100,000	212	21.5	75.5	915	4.3	96.4
100,001~200,000	135	13.7	89.2	438	2.0	98.5
200,000 以上	106	10.8	100	311	1.5	100



圖三 K-Means 分群法的群組一損失金額長條圖



圖四 K-Means 分群法的群組二損失金額長條圖

由表九到表十二與圖三可以發現，群組一的車禍事故發生件數很少，但是損失金額卻是一般車禍的三倍以上，若針對減少金錢上的損

失，可以從此群組著手。表九到表十二與圖四可以發現，群組二是二群組之中車禍發生次數最多，損失金額較低，若是要降低車禍發生之次數，可針對此群組加以分析研究。

在 K-Means 分群法的分群結果中，我們可以發現群組一與群組二的區分很明顯，有些項目的重疊性較低甚至完全沒有重疊。在平均快車道數目方面，群組一的主要分佈於 4 與 6 個車道，而群組二主要分佈於 0 與 2 個車道，在二個群組中主要分佈沒有重疊的情形；在速限方面，群組一的主要分佈於速限 70、90 與 100，而群組二主要分佈於速限 40，在二個群組中主要分佈沒有重疊的情形；在道路類別方面，群組一的主要分佈於國道與省道，而群組二主要分佈於市區道路，在二個群組中主要分佈也沒有重疊的情形；在平均損失金額方面，群組一的主要分佈於 5,001~10,000、10,001~50,000、50,001~100,000 與 100,001~200,000，而群組二主要分佈於 5,000 以下、5,001~10,000 與 10,001~ 50,000，在二個群組中主要分佈重疊的部份是 5,001~10,000 與 10,001~50,000。

3.5 Two-Step 與 K-Means 分群法之比較

在 Two-Step 分群法與 K-Means 分群法的結果之中，我們可以發現 Two-Step 分群法與 K-Means 分群法都能有效的將資料區分成二群，但從表四到表七、圖一到圖二與表九至表十二、圖三到圖四的相對比較之下，可以明顯的看出 Two-Step 分群法的群集分佈與 K-Means 分群法的群集分佈有著顯著不同的差異。Two-Step 分群法的群集分佈像是呈現著平均分配；而 K-Means 分群法的群集分佈像是呈現著峰態。

一般而言，群集中的觀察值具有高度相似性或以幾何學說法而言，群集內部差異越小越好，群集間的差異越大越好[21]。從 Two-Step 分群法與 K-Means 分群法的次數分配表互相比較之下，K-Means 分群法的二個群組在不同的值所佔的比重有明顯的差異；而 Two-Step 分群法在群與群之間的相異性較沒有 K-Means 分群法的那麼明顯。

首先我們從平均快車道數目次數來看，Two-Step 分群法所區分出來的二個群組具有較低的相異性，二個群組的主要分佈大部份都在 0、2 與 4 個車道，而 K-Means 分群法所區分出來的二個群組就具有較高的相異性，二個群組的主要分佈大部份都不同，而且群組一主要分佈於快車道數目較多的車道，群組二主要分佈於快車道數目較少的車道；在平均速限中可以明顯的看出，Two-Step 分群法的平均速限次數分配表中，雖然速限 80 至 100 的件數全部分佈於群組二之中，但速限 40 的件數在群組二仍佔了 37.5%的比例，而 K-Means 分群法的平均速限次數分配表中，損失金額較低的群組完全分佈在速限 60 以下，而損失金額較高的群組分佈完全在速限 70 以上，群與群之間的相異性明顯比 Two-Step 分群法高，明顯區分出二群組的不同；在道路類別中，Two-Step 分群法與 K-Means 分群法所區分的群組則沒有太大的差異；損失金額方面，Two-Step 分群法與 K-Means 分群法所區分的群組都很相似，但以項目重疊性來說，則是以 K-Means 分群法的項目重疊較少。

雖然 Two-Step 分群法也能有效的將資料分群，但從分群的效果上來看，K-Means 分群法可以更明顯的區分出二個群組，故我們認為 K-Means 分群法在車禍事故中具有比較好的分群能力。

3.6 關聯規則

在 K-Means 分群法的分群結果中，分出了嚴重事故群與輕微事故群，接下來我們再利用關聯規則對嚴重事故群進行分析。

在關聯規則分析中，Support 表示事件發生的機率、Confidence 表示事件發生後其為真的機率，我們對於事件發生後其為真的機率較高者有興趣，也就是對 Confidence 較高者有興趣，因為某些車禍事故雖然不常發生，但一發生可能會相當的嚴重，故本研究將關聯規則的 Support 設為 20、Confidence 設為 80，結果在嚴重事故群中發現了二條關聯規則其信賴度高達 1 之情形，請參考表十三。有一條是發生在無裝設號誌情況中，其支持度有 61.1%而信賴度高達 1，其可能發生的情況是在晴天且路面狀況良好，因為該路段沒有裝設號誌，所以駕駛者開車容易車速過快，比較不會小心謹慎，一但有緊急情況，很有可能會來不急反應而造成較嚴重的事故。而另一條規則的支持度只有 26.1%，雖然發生的機會約為四分之一，但若發生了那一定是在國道上速限 90 以上的地方發生追撞，若能針對國道增設警告標誌，減少追撞的發生將能減少嚴重事故的發生。

表十三 嚴重事故群的關聯規則

Association Rule	Support	Confidence
車道線附標記 & 無慢車道 & 無裝設號誌 & 標誌適當 → 道路類別(國道)	60.6%	0.987
天候(晴) & 號誌無動作 & 路面狀況(乾燥) →無裝設號誌	61.1%	1
分道設施(無劃分或無慢車道) & 無裝設號誌& 快車道鋪設(柏油) → 事故位置(快車道)	64.7%	0.841
事故類型(追撞) & 速限 90 以上 →道路類別(國道)	26.1%	1

3.7 總結

在研究當中，我們使用了 Two-Step 分群法、K-Means 分群法與關聯規則三種分析方法，其中以 Two-Step 分群法與 K-Means 分群法進行分群，結果發現這二種分群方法，都能區分出嚴重事故群與輕微事故群，但以 K-Means 分群法所區分出來的群組，分群效果明顯優於 Two-Step 分群法，我們認為 K-Means 分群法在車禍事故中具有比較好的分群能力。

關聯規則中發現，在國道上速限 90 以上的地方發生追撞，雖然發生機率只有 26.1%，若一發生必定相當嚴重，故建議交通部可針對該點加以改進，若能針對速限高的路段增設警告標誌，或加強宣導開快車其發生事故的嚴重性。

四、結論與建議

道路交通事故的發生在台灣仍然是時常可見，如何在眾多的車禍事故中找出其共同特性，進一步的加以改善以減少傷亡，使民眾的生命安全受到保障是非常重要的。本研究對道路交通事故調查表所記錄的車禍資料，使用 Two-Step 分群法與 K-Means 分群法，二種方法來進行分群、比較，結果發現在道路交通事故分析上，K-Means 分群法的分群效果明顯比 Two-Step 分群法的分群效果好，可以更明顯的區分出事故的嚴重、輕微，更從關聯規則中得知，當道路的快車道數目越多或是在速限高的路段，交通事故的嚴重度也就越高。

一般而言，在快車道數目多時，車輛行駛的速度也隨之變快，然而當車禍發生時，情況會比一般車禍更為嚴重，損失金額也隨之提高，而這些情形多半發生在道路類型是國道或省道的情況下。故交通部應加以改善國道與省道的道路交通規則與道路環境，並持續交通安

全的宣導，讓民眾有正確的行車觀念，並告誡駕駛人車速勿過快，使駕駛人了解發生車禍事故的危險性或是在快車道數目多的地段設置警告標語，提醒民眾應保持安全距離、注意行車安全，以減少交通事故的發生。

由於時間上的因素，本研究只取了民國 88 年度的車禍事故資料，並針對道路交通事故的道路環境進行分析，未來仍可以加入多個年度、當事人方面的資料來進行分析，以找出事故的其他趨勢，得到更詳細的結論。

參考文獻

[1] 行政院衛生署(2003)，臺灣地區主要死亡原因，Online，http://www.doh.gov.tw/statistic/data/死因摘要/91 年/表一.xls。

[2] 國道高速公路局(2002)，高速公路年報，Online，http://www.freeway.gov.tw/11_91_05_02. asp。

[3] 國立教育資料館(2003)，交通安全教育資訊年刊，Online，http://www.nioerar.edu.tw:82/basis1/ 691/newpage2.htm。

[4] Frawley, W.J., Paitesky-Shapiro, G., Matheus, C.J. (1991), Knowledge Discovery in Databases: An Overview, AAAI/MIT Press.

[5] Grupe, F.H., Owrang, M.M. (1995), "Data mining discovering new knowledge and cooperative advantage," Information Systems Management, Vol. 12, No. 4, pp. 26-31.

[6] Cabena, P., Hadjinian, P., Stadler, R. (1997), Discovering Data Mining from Concept to Implementation, Prentice-Hall.

[7] Kleissner, C. (1998), "Data mining for the enterprise," IEEE Proc. 31st Annual Hawaii International Conference on System Sciences, 7, pp. 295-304.

[8] 葉涼川，2000，CRM Data Mining 應用系統建置，美商麥格羅．希爾國際公司。

[9] 陳敬明，1999，臺十五線易肇事地點評定與改善對策之研究，國立交通大學交通運輸研究所碩士論文。

[10] 吳偉碩，2000，台南環線高快速公路肇事特性分析與安全改善之研究，國立交通大學交通運輸研究所碩士論文。

[11] 魏開元，1998，由肇事碰撞構圖及類神經網路推導肇事工程因素研究，國立成功大學交通管理學系碩士論文。

[12] 程銘鎮，2002，國道高速公路交通事故發生原因之分析與探討，國立雲林科技大學工業工程與管理研究所碩士論文。

[13] 張敏亮、吳信宏、郭廣洋 2004，「應用資料探勘於車禍肇事環境之關聯規則」，2004 產業管理創新研討會論文集，pp. 526-531。

[14] Universitat Hamburg (2004), Two-Step Cluster, Online, http://www.rrz.uni-hamburg.de/RRZ/

[15] Software/SPSS/Algorith.120/twostep_cluster.pdf.

[16] Buttrey, S.E., Karo, C. (2002), "Using K-nearest-neighbor classification in the leaves of a

tree," Computational Statistics and Data Analysis, Vol. 40, pp. 27-37.

[17] avidson, I. (2002), "Understanding K-means non-hierarchical clustering," SUNY Albany Technical Report 02-2.

[18] Han, J., Fu, Y. (1995), "Discovery of multiple-level association rules from large databases," Proceedings of the 21st International Conference on Very Large Data Bases, pp. 420-431.

[19] 交通安全入口網，2003，目前執行專案，http://168.motc.gov.tw/gip/ct?xItem=17 23&ctNode=709。

[20] 交通部運輸研究所，2003，交通事故傷亡比罰單更驚人， http://www.iot.gov.tw/ct.asp?xItem=103778&ctNode=1066。

[21] Ganti, V., Ramarkrishnan, R., Gehrke, J., Powell, A., French, J. (1999) "Cluster large datasets in arbitrary metric spaces," Proceedings of the 15th International Conference on Data Engineering, Sydney, Australia, pp. 502-511.

[22] 彭文正，2001，資料挖礦-顧客關係管理暨電子行銷之應用，數博網資訊股份有限公司，台北。