鄭麗珍、李麗美（2014），『探勘不平衡資料集中之突顯樣式-以國道事故資料為實證研究』，*資訊管理學報*，第二十一卷，第二期，頁 161-184。

探勘不平衡資料集中之突顯樣式－以國道事故資料為實證研究

鄭麗珍*
東吳大學資訊管理學系

李麗美
交通部臺灣區國道高速公路局

摘要

在資料探勘的分類問題中，大多數演算法都是設計在資料類別分布平均的情況下去訓練分類模型。然而，在實務應用上，資料類別分布不平衡是常見的狀況，在這樣的資料集設計的分類方法是很重要的研究議題。此外，透過分類模型所找到的規則常瑣碎複雜，透過突顯樣式探勘可以整理篩選出具有區分找出兩個類別之間的顯著差異與獨特識別的規則。然而，過去沒有相關研究在不平衡資料集上作突顯樣式探勘。本研究提出一個新的研究架構，基於關聯規則分類的方法，調整資料的權重於計算支持度，以探勘出不平衡資料集之突顯樣式，並加入不同年份間的突顯樣式變化探勘。本研究以真實之國道交通事故資料集為實證基礎，此資料為一個嚴重不平衡的資料集，死亡事故僅佔全部事故資料的百分之一比例都不到。然而，主管機關一直努力探求了解死亡事故發生原因，希望可以透過各項因應措施，增進行車安全減低死亡事故發生。本研究將透過提出之研究架構，找出一般及稀有死亡事故的肇事因子間關聯，並分析不同年度間肇事因子，找出一些重要的樣式，提供交通管理單位參考。

關鍵詞：關聯規則分類、突顯樣式、不平衡資料集、高速公路事故、權重支持度

* 本文通訊作者。電子郵件信箱：lijen.cheng@gmail.com
2013/7/27 投稿；2013/10/1 修訂；2013/12/17 接受

Cheng, L.C. and Lee, L.M. (2014), 'Mining Emerging Patterns from Imbalance Dataset – A Case Study on Freeway Accident Database', *Journal of Information Management*, Vol. 21, No. 2, pp. 161-184

Mining Emerging Patterns from Imbalance Dataset – A Case Study on Freeway Accident Database

Li-Chen Cheng*
Department of Computer Science and Information Management, Soochow University

Li-Mei Lee
Taiwan Area National Freeway Bureau, MOTC

Abstract

Traditional associative classification is used to search frequent patterns at the balance datasets. However, most real life datasets are imbalance. To discover special rare patterns from imbalance dataset is an important job. Currently, the freeway becomes the main transportation route at Taiwan. Because of the high speed and heavy traffic, accidents at highway would cause more serious injuries than other roads. The serious injury accidents are very small part among the accident data. The impact factors of these special cases are the most important issue. This study proposes a framework to explore the most significant reasons for serious accidents. The framework combines the associative classification method with the emerging patterns mining to discover rare and serious incidents. The weight of each accident is adjusted by the severity of accident. Since the rare items can be discovered by the proposed formula of calculation support. The results of an experiment that was conducted on a real accidents data demonstrated the efficacy of the proposed approach. After analysing these accidents, we provide some suggestions.

Keywords: Associative Classification, Emerging Patterns, Imbalance Dataset, Freeway Accident, Weight Support

* Corresponding author. Email: lijen.cheng@gmail.com
2013/7/27 received; 2013/10/1 revised; 2013/12/17 accepted

壹、導論

資料探勘技術可以在大量資料中，不需先行了解資料分佈下，透過相關的方法找出有趣且重要的樣式，這些隱含資訊可以協助決策者分析。目前已經成功應用在各種領域，其中關聯規則用來分析預測消費者行為，最著名的例子，就是在威瑪超市的購物資料中，找出啤酒跟尿布常被一起購買的特殊消費行為（Agrawal et. al. 1993; Agrawal & Srikant 1994）；分類技術也早已普遍應用在醫療或是金融產業，例如建立分類模型識別信用卡客戶是否流失或是預測病患是否得到某種疾病。

在分類技術中，核心技術是先透過已分類好的資料庫，建立一個識別分類的模型以預測未來趨勢。通常模型中會透過適當的規則把不同類別的資料準確的區分出來。目前分類技術所建立的模型，有兩大重要的議題需要克服：首先大多數分類模型都是建構在資料分布均勻的資料集，然而實務上的資料集卻多半屬於類別分布不平衡，像客戶流失的資料集中，屬於流失的客戶類別是少數。此時訓練資料集中大多數資料是某一類別，這樣所建購之模型會偏向僅能精準預測多數類別。然而，行銷人員在乎的是如何有效識別流失客戶，這卻是少數類別。因此，如何克服不平衡資料集以建立分類模型，是一個很重要的議題。此外，多數的分類模型，會透過產生許多分類規則來建立模型以預測。甚至可能為了有效精準的分類，而產生過多的推理規則，讓決策者無所適從。本研究將針對以上這兩議題，提出一個創新的探勘架構來解決這兩個問題。

在本研究的架構中，首先諮詢領域專家意見，依據該資料集的特質以調整資料集中每筆交易的重要性，解決不平衡資料集的問題。本研究採用關聯規則分類的技術以建立模型，透過支持度及信賴度門檻限制來篩選出重要的分類規則。另一方面，更結合突顯樣式（Emerging Patterns: EPS）探勘的技術，以解決模型規則太多的困境。突顯樣式意指在兩個不同類別的推理規則中，找出不同類別間獨特的差異。篩選其對應類別特別具有代表性的專屬規則，提供管理者豐富的資訊。目前也應用上在不同的領域，例如：在眾多導致病人得癌症生化因素中，聚腺苷酸化（polyadenylation site prediction）是很關鍵的物質，利用突顯樣式探勘，可以提供醫學研究上識別出得癌症與正常人這兩不同類別間聚腺苷酸化的訊號差異（Dong & Li 1999; Li & Wong 2002）。

本研究提出之創新突顯樣式探勘架構，可以解決不平衡的分類資料中突顯樣式的擷取，除此之外還可處理不同年度的樣式變化分析，在交通事故研究上此架構據我們所知是目前尚未有的。將透過國道 1 號的事故資料集，來做實證分析與研究。事故資料是一個極不平衡的資料集，嚴重死亡事故與一般事故這兩類資料比率將近 1 比 180。且國道 1 號為歷史最久、交通量最大且使用人次最多，發生事

故比率相對其它國道為高，故含較完整的事故樣態。透過本研究創新的探勘架構，可以探勘出哪些肇事原因常在嚴重的死亡車禍事件中出現，卻極少或是沒有出現在一般事故中或是反之亦然。藉此找出嚴重事故與一般事故之差異性規則。提供交通管理單位參考，藉以規劃因應或改善措施，制定決策對症下藥，避免類似問題發生。接著整理出不同類別不同年度間的突顯樣式規則，以提供不同的面向的資訊給交通決策者專業決策者。

本文共分為五章，第一章為緒論，第二章為文獻探討，第三章為研究架構，說明本研究的步驟及方法，第四章為實證研究，第五章為結論。

貳、文獻探討

一、關聯規則分類

資料探勘領域中，關聯規則為資料探勘領域中最常被應用的模式之一，由 Agrawal 等學者提出，用來探勘頻繁項目集及強關聯規則，後來經廣泛研究成為熱門的資料探勘技術。購物籃分析（market basket analysis）即為著名的商業應用，藉由日積月累的龐大交易資料庫中，發掘顧客購買不同商品的關聯來分析客戶的購買習慣，作為未來行銷參考（Han & Kamber 2006）。

關聯規則探勘以支持度（support）與信心水準（confidence）兩個參數，用來判斷找出的規則（X→Y）是否有意義，支持度越高代表項目 X 出現的頻率相對越高，信心水準（Confidence）越高，代表同時包含 X 與 Y 的交易筆數佔整個資料庫包含 X 的交易筆數的百分比越高。支持度與信心水準可以用來判斷規則的有效性，大幅度縮減規則數量，一個有效的關聯規則其支持度必須大於或等於事先設定的兩個參數值，最小支持度（mini support）與最小信心水準（mini confidence）。

關聯規則分類根據已知資料及其分類屬性值，建立以規則為基礎之分類模型，接著利用此分類模型預測新資料的類別，假設 Z 代表關聯規則的項目集，C_i 代表分類的類別屬性，如公式(1)所示。

$$Z \rightarrow C_i \tag{1}$$

關聯規則分類兼具關聯規則探勘及分類的特性，除可顯示屬性間的有效關聯，亦可用描述與類別之間的關係，在某些領域上之分類結果比傳統決策樹優，因此最近被使用於分類。關聯規則分類法較傳統決策樹的優點有，第一，傳統分類器如歸納法（induction）或決策樹，當規則建立後，會刪除訓練資料集內與其相關之屬性，因此，所產生的往往是所有規則的子集，可能錯失該屬性在某些案例上扮演重要角色的詳細的規則（Thabtah et. al. 2006）。而關聯規則分類，可以探勘出資料中所有屬性間的關聯，並且很成功的用於的分類上（Ali et. al. 1997; Liu et. al.

1998）。第二，決策樹在產生的決策規則，當資料量很大會產生很多的規則，且很難理解讓使用者無所適從哪些是重要的規則。反之，關聯規則分類比傳統分類器更能精確的分類以及發現更多規則（Liu et. al. 1998; Antonie et. al. 2003）。在複雜的分類中，也比決策樹、C4.5、naive bayes 及支持向量機可以提供更有效的分類及更準確的預測」（Dong et. al. 1999 ; Yin & Han 2003; Li et. al. 2001; Wang & Karypis 2005; Veloso et. al. 2006）。近年來，關聯規則分類異軍突起，乃因其可以有效篩選出精確重要的規則給決策者。

關聯規則分類雖有以上優點，但對於稀有項目的探勘並不擅長，因為大部份關聯規則所提出之演算法，著眼於找出高頻繁項目關聯規則之效率，對低頻繁項目且不平衡資料探勘的討論相對較少。然而，現實生活中發生的問題，如金融詐騙、電腦網路犯罪及恐怖行動，這些問題都有共同特色，就是這些問題都是眾多資料中的稀少案例（Cao et. al. 2008）。但卻可能是關鍵項目，為決策者最有興趣了解的，因此近年來逐漸引起學者關注，陸續提出低支持度項目關聯規則探勘並發展出各種演算法，一般稱為稀有樣式探勘。

對於低支持度的稀有樣式進行關聯探勘時，因有最小支持度及最小信賴度門檻限制，若調高支持度，則低支持度但有趣樣式會被忽視，若將支持度設低雖可能探勘出，但產生太多無趣規則且執行績效不彰等問題。近年來，各種解決方法被陸續提出，有的學者以發展不同演算法找出稀有樣式（Koh & Rountree 2005; Szathmaryet. al. 2007; Troiano et. al. 2009; Weng 2011）。部分學者以改變支持度的計算方式或定義有意義函數，再配合演算法找出稀有但有趣樣式（Liu et. al. 1999; Yun et. al. 2003; Zhou & Yau 2007; Romero et. al. 2010）。另一派學者以多重支持度的觀念，針對不同商品給定不同的支持度，來過濾出稀有樣式（Liu 1999; Hu & Chen 2006; Chen & Huang 2013; Huang 2013; Hu et. al. 2013）。

交通事故資料也有同樣的情形，在所有的事故中，嚴重事故佔的比例很低但影響很大，非常適合用稀有類別關聯規則探勘，本研究參考之前學者之研究並配合交通事故嚴重程度特性來改變支持度，尋找特殊的樣態做觀察。

二、突顯樣式

突顯樣式可以在兩個不同集合中篩選出支持度有明顯差異的項目集，在有時間欄位之資料庫中可以顯示趨勢變化、比對出不同分類資料之差異和變化。這些差異和變化可以提供管理者豐富的資訊，而成長率即扮演此重要區分變化的角色，其表示該類別對另一類別之支持度比例。突顯樣式探勘的技術，是一個很重要且很有價值的分類工具應用在很多實務上的問題（García-Borroto et. al. 2012）。

突顯樣式在許多應用上已被證明有很大效用，尤其對疾病了解及偵測相關的

基因譜分析上（Li & Wong 2002; Dong et. al. 2005）。如在醫學領域應用上，聚腺苷酸化（polyadenylation site prediction）的預測是一個具有挑戰性的問題，因為該物質是生物演進和疾病如了解癌症的關鍵元素，是提供治愈癌症研究的答案，利用突顯樣式探勘類別間獨特的差異，提供醫學研究上識別出不具明顯訊號的聚腺苷酸化（George et. al. 2011）。近年來也應用於串流資料的分類（Alhammady 2007）與網路異常偵測（Ceci et. al. 2008）等實務領域。

突顯樣式是透過資料探勘的方法，找出在事先定義好的兩類資料中，找出足以代表這兩類資料的獨特樣式，最早是由 Dong and Li（1999）提出突顯樣式探勘之觀念。其原理為 D1 與 D2 為兩個擁有相同欄位、不同交易紀錄之資料集，項目集 X 於 D1 及 D2 的支持度分別以以 supp1(X)及 supp2(X)表示，則項目集 X 從 D1 到 D2 的成長率 GrowthRate(X)如公式(2)所示。

$$\text{GrowthRate}(X) = \begin{cases} 0 & \text{if } supp_1(X) = 0 \text{ and } supp_2(X) = 0 \\ \infty & \text{if } supp_1(X) = 0 \text{ and } supp_2(X) \neq 0 \\ \dfrac{supp_2(X)}{supp_1(X)} & \text{otherwise.} \end{cases} \quad (2)$$

當項目集 X 於 D_1 及 D_2 支持度為零時，則 *GrowthRate*(X)=0；當項目集 X 於 D_1=0 但 $D_2 \neq 0$ 時，則 GrowthRate(X)=∞，此項目 X 被稱為跳躍式突顯樣式（jumping emerging patterns: JEPs）；一般的情況下，*GrowthRate*(X)=*supp2*(x)/*supp1*(X)，當 *GrowthRate*(X)大於一門檻值 σ 時，則稱項目集 X 為突顯樣式，代表此樣式在不同資料集中，其支持度有明顯變大，突顯樣式與跳躍式突顯樣式均為突顯樣式探勘之目標。

三、資料探勘用於交通事故之研究

過去交通事故的研究，大致上可以分為兩大類，一是以統計為主，另一則是資料探勘技術。統計方面的研究，有學者採用卜瓦松迴歸及負二項迴歸進行肇事因果分析（戚培芳 1997; 林郁志 1998）；也有學者使用依序羅機模式（陳志和 1999）及二元羅吉特模式（楊思瑜 2003）評估事故傷害程度。由於傳統的統計模型需要先行假設，如相依變數和解釋變數間的線性函數形式，所以統計模型雖已被廣泛用於交通事故受傷嚴重程度之分析，但仍受到統計先天上之限制，當這些交通事故數據與統計假設不相符時，產生之統計模型將受到挑戰，甚至可以產生錯誤估計和推論（Mussone et. al. 1999）。

資料探勘的技術適合用在大量資料分析，其特性在於不需先行了解資料分佈

下，可以透過不同演算法找出有用的樣式。近年來也陸續應用於交通事故分析上。最常被應用的研究技術是分類法，透過不同的分類演算法，建立模型以預測未來趨勢。

近年探勘事故的研究主要可分為兩大方向：分類與分群。分類方法主要是先利用測試資料建立模型或進而用測試資料預測未來趨勢。常用於交通事故受傷嚴重程度與事故因子的預測，透過駕駛、車輛等因素與受傷嚴重性等級建立模型，進而找出規則預知那些是高受傷風險車輛為的重要因素。目前，各種不同的分類法，像分類及回歸樹（Classification and Regression Tree: CART）（陳文杰 2004; Chang & Wang 2006）、類神經網路（Artificial Neural Network Models: ANN）（周雍傑 2000; 黃昶斌 2004; Abdelwahab & Abdel-Aty 2001; Nefti & Oussalah. 2004; Solomon et. al. 2006; Xie et. al. 2007）、決策樹（Chong et. al. 2004; Chong et. al. 2005; Solomon et. al. 2006）及支持向量機（Chong et. al. 2005）等多種模型，都成功應用在事故方面的研究也證明其重要性。例如學者建立駕駛、車輛、道路狀況及環境因素與受傷嚴重性關係模型，找出高受傷風險車輛為最重要因素（Sohn et. al. 2003）。也有學者利用類神經網路建立分類模型（Abdelwahab et. al. 2001; Nefti & Oussalah. 2004; Delen et.al. 2006; Solomon et. al. 2006; Xie et. al. 2007），例如 Abdelwahab and Abdel-Aty 學者用在駕駛人、車、道路及環境特性探討司機受傷嚴重程度（Abdelwahab et. al. 2001）。此外，亦有學者比較決策樹、類神經網路、支持向量機等多種模型之研究（Chong et. al. 2004; Chong et. al. 2005; Solomon et. al. 2006）。例如 Chong 等學者將傷害程度分為「沒有受傷」、「可能的傷害」、「非失能傷害」、「失能傷害」及「致命傷」5 種分類，比較決策樹及類神經網路，找出三個最重要的因素是安全帶使用，光線及酒精使用。

分群是將一群個體依據相似程度群組在一起的過程，所形成的群集有群內相似群間相異的特性；分群有時單獨使用，如進行對個體分佈的了解及群集特性的觀查及分析，有時也作為其它方法的前處理步驟（Han & Kamber 2006）。在交通事故分群上，有學者用 Two-Step、K-Means 對車禍資料分群效果，結果顯示 K-Means 分群效果較好（吳冠宏等 2006）；更多學者以分群結合其它方法，如 GIS、關聯規則、多變量分析…等進行交通事故研究（Sohn et. al. 2003; Depaire et. al. 2008; Anderson 2009; 吳冠宏等 2004; 黃湄清 2005）。

參、研究架構

在此節我們將介紹本研究所提出之不平衡資料集的突顯樣式探勘架構。此架構以關聯規則分類法為基礎，結合突顯樣式探勘與稀有項目探勘的概念，解決不平衡資料集中探勘重要資訊的問題。本研究架構之模組如下：

模組一：調整資料庫中每筆交易的權重：此步驟將根據實證案例的資料意義，參酌領域專家建議，修正資料庫中每筆交易的權重，以加重稀有資料的權重。

模組二：關聯規則分類：此分類法兼具關聯規則探勘及分類的特性，將找出識別各分類屬性間的關聯規則，可以利用支持度與信賴度篩選出重要的規則。

模組三：突顯樣式探勘：在兩個不同類別集合中，透過本研究所定義之成長率篩選規則篩選出支持度有明顯差異的突顯樣式。

模組四：不同年度突顯樣式分析：透過本研究所提出的步驟，將整理與歸納出觀察不同年度突顯樣式消失及新增之變化趨勢。

本研究將以國道 1 號的事故資料，來做實證分析與研究。後面章節將以此資料庫為例來介紹本研究提出之不平衡資料集之突顯樣式探勘架構，針對各模組一一做詳細說明。

一、調整資料庫中每筆交易的權重

此模組主要考量每個事故產生之嚴重性不同，每個事故不應該視為相同的比重。為展現資料集的實質意涵來調整各筆交易的權重，本研究參考過去學者調整交易比重之作法（Tao et. al. 2003）。經與領域專家討論，決定以國內交通部運輸研究事故當量來調整各筆交易的權重（林大煜 1982）。

在交通事故相關研究中，依照事故嚴重性，會給予不同的比重。生命財產損失越大者，事故嚴重性越大；又國道事故資料僅詳細記載之人員傷亡資料，並無法知道財產損失之金額。因此，在衡量事故嚴重性時，皆以死傷人數來量化。經與領域專家討論，依國道事故資料特性並參考交通事故相關肇事嚴重性法之研究，採用運輸研究所事故當量中所定義之肇事嚴重性當量，來調整交易權重。

肇事嚴重性法是肇事路段判斷指標之一，其考量路段發生事故所累積之死亡、受傷及財務損失之價值觀不同。在計算事故嚴重程度時，合併考慮而採取「當量」之方式計算，其對死亡人數加重最大，其次為受傷人數。運輸研究所事故當量為該領域最常衡量肇事嚴重性法之一（蘇志哲 2003）。公式(3)計算的數值代表該路段事故狀況，依照路段累積事故計算出每一路段的肇事嚴重性值 （林大煜 1982）。本研究是以每個肇事事件為一筆交易，故將每一路段肇事嚴重性值回算至每一事故之肇事嚴重程度。因此修改公式(3)為公式(4)以計算出每一事故嚴重程度。

每一路段肇事嚴重性值之計算：$9.5\times F+3.5\times J+TAN$ (3)

每一事故肇事嚴重性值之計算：$9.5\times F+3.5\times J+1$ (4)

F：肇事死亡人數

J：肇事受傷人數

TAN：總肇事次數

例如：某一事故，其死亡人數為 1 人，受傷人數為 3 人，依照公式(4)則該事故之肇事嚴重程度為 21。

本研究要找出不同嚴重類型事故的相關因子，利用公式(4)計算出每一事故之肇事嚴重程度，做為每筆交通事故之交易權重。數值越大代表死傷人數越多，事故嚴重程度越高，對用路人生命威脅也越大，應越受重視並加以防範。希望透過加重的方式，在極為不平衡的交通事故資料中，以發掘出並突顯出稀有之嚴重事故，有利於找出嚴重事故之因子關聯性提供管理單位參考。

二、關聯規則分類

此模組將進行關聯規則分類探勘，根據專家建議將事故依嚴重程度分為嚴重及一般兩種類別，本研究將透過此步驟找出一般事故及嚴重事故之影響因子關聯性。

其中嚴重事故是最為重要，卻可能會因為資料量稀少，無法用傳統支持度設定有效探勘出來。傳統關聯規要找出稀有樣式，必須支持度設定很低，將導致找出很多沒有意義的資訊。Romero et. al.（2010）為避免這種情形，提出條件支持度的計算。本研究參考其精神，修改關聯規則分類支持度為公式(5)(6)，分別為一般及嚴重事故權重支持度計算公式。這樣調整後，可以有效解決稀有事件之傳統支持度太小無法探勘出來之遺憾。

$$\text{一般事故權重支持度} = \text{Gsup}(\text{Y} \rightarrow \text{一般事故}) = \frac{\text{support}(\text{Y} \cup \text{一般事故})}{\text{support}(\text{一般事故})} \quad (5)$$

$$\text{嚴重事故權重支持度} = \text{Ssup}(\text{X} \rightarrow \text{嚴重事故}) = \frac{\text{support}(\text{X} \cup \text{嚴重事故})}{\text{support}(\text{嚴重事故})} \quad (6)$$

三、突顯樣式探勘

在前一模組，透過關聯規則分類後，已經篩選出分屬嚴重及一般事故事故之事故關聯規則。本階段將利用成長率以探勘出兩類別各自的突顯樣式，詳細步驟描述如下：

步驟 1： 將所有透過關聯規則分類的結果，依右式分為嚴重及一般事故兩類別。所有探勘出來之規則分成嚴重事件集合 SR 及一般事故事件集合

GR 兩集合。

步驟 2： 分別找出 SR 及 GR 兩集合之規則於一般及嚴重事故權重支持度。例如:規則 r_i: X→嚴重事故， r_i∈SR，其嚴重事故權重支持度為 Ssup(r_i)，找出其項目集 X 於一般事故之權重支持度 Gsup(r_i)；規則 r_j: Y→一般事故，r_j∈GR，其一般事故權重支持度為 Gsup(r_j)，找出其項目集 Y 於嚴重事故之權重支持度 Ssup(r_j)。

步驟 3： 計算 SR 及 GR 兩集合內，其項目集的成長率。

規則（r_i: X→嚴重事故），項目集 X 的成長率如公式(7)

$$\text{GrowthRate}(X)=\begin{cases}0 & ,\text{if } Gsup(r_i)=0 \text{ and } Ssup(r_i)=0\\ \infty & ,\text{if } Gsup(r_i)=0 \text{ and } Ssup(r_i)\neq 0\\ \dfrac{Ssup(r_i)}{Gsup(r_i)} & ,\text{otherwise.}\end{cases} \tag{7}$$

規則（r_j: Y→一般事故） ，項目集 Y 的成長率如公式(8)

$$\text{GrowthRate}(Y)=\begin{cases}0 & ,\text{if } Ssup(r_j)=0 \text{ and } Gsup(r_j)=0\\ \infty & ,\text{if } Ssup(r_j)=0 \text{ and } Gsup(r_j)\neq 0\\ \dfrac{Gsup(r_j)}{Ssup(r_j)} & ,\text{otherwise}\end{cases} \tag{8}$$

步驟 4： 以成長率篩選規則，成長率超過門檻之規則為突顯樣式。成長率的門檻值篩選是一件很重要的工作，通常要在實驗時觀察找出一個適當的門檻值點，以確保探勘結果。

四、不同年度突顯樣式分析

本模組分析突顯樣式在不同年度是否新增、消失及共同存在之趨勢變化， 詳細步驟說明如下。

假設 Y 年度中，SR(Y)及 GR(Y)分別為嚴重及一般事故突顯樣式集合，SR(Y+1)及 CR(Y+1)分別為 Y+1 年度嚴重及一般是故突顯樣式集合。

步驟 1： 規則 r_i 屬於 SR(Y)，如果 r_i 亦屬於 SR(Y+1)，則 r_i 為 Y 及 Y+1 年共同都有之嚴重事故突顯樣式，否則 r_i 為 Y+1 年消失的樣式。

步驟 2： 規則 r_i 屬於 SR(Y+1)，如果 r_i 亦屬於 SR(Y)，則 r_i 為 Y 及 Y+1 年共同都有之嚴重事故突顯樣式，否則 r_i 為 Y+1 年新增的樣式。

步驟 3： 規則 r_i 屬於 GR(Y)，如果 r_i 亦屬於 GR(Y+1)，則 r_i 為 Y 及 Y+1 年共同都有之一般事故突顯樣式，否則 r_i 為 Y+1 年消失的樣式。

步驟 4： 規則 r_i 屬於 GR(Y+1)，如果 r_i 亦屬於 GR(Y)，則 r_i 為 Y 及 Y+1 年共同都有之一般事故突顯樣式，否則 r_i 為 Y+1 年新增的樣式。

肆、實證研究

一、實驗資料

交通事故依嚴重性分為 A1、A2 及 A3 三種等級，A1 為造成人員當場或二十四小時內死亡之事故，A2 為造成人員受傷或超過二十四小時死亡之交通事故，A3 為財物損失之交通事故，為簡化分類，本研究將事故種類為兩類，一類嚴重事故意指造成人員死亡事故，包含所有 A1 事故和超過 24 小時死亡之 A2 事故；另一類為一般事故，是指 A3 事故及未有人員死亡之 A2 事故。

本研究事故資料係採用高速公路 98-99 年國道 1 號事故資料，因其歷史最久交通量最大及使用人次最多，事故樣態較完整，故選擇以國道 1 號為研究對象。

98 與 99 年的事故資料，在不同類別與不同年度的統計資料，如表 1 所示，可以看出死亡嚴重事故數量相當稀少。

表 1：98-99 年交通事故統計

年度	國道名稱	嚴重事故		一般事故		合計	A1 事故與全部國道 1 號事故比率
		A1	A2 死亡	A2 受傷	A3		
98	國道 1 號	34	1	429	5889	6353	1:187
99	國道 1 號	32	4	509	8426	8971	1:280

本研究受限無法於取得當事人及道路幾何特性等資料。故僅就取得交通事故資料庫中，委請有 10 年以上經驗之資深交通管理專家，協助挑選有意義欄位，最後整理出 27 個重要事故因子欄位，如表 2 所示。

表 2：事故嚴重程度因子欄位

影響因子類別	影響事故因子欄位
時間	月、時、時段、星期、假期
天候環境	天候、光線、視距
道路環境	方向、里程、事故位置、路面鋪裝、路面狀態、路面缺陷、道路型態、障礙物
交通工程	號誌種類、號誌動作、車速、分向設施、快車道或一般車道間、快慢車道間
車輛機械	車種
事故類型及原因	事故類型、事故型態、主要肇因類別、主要肇因

在進行分析前，依照相關專家的建議處理空值與衝突資料部分。在這些欄位中，為了進一步呈現事故的原因，將一些特殊欄位產生衍生欄位。如：原始事故資料欄位是「日期」本研究將之轉換成「星期」欄位與「假期」欄位，期望能深入找出事故的背景因素。「假期」欄位依人事行政局所頒布之政府機關辦公日曆表轉換。「車速」欄位原本為連續性數值轉，本研究依照專家見建議換為 < 70、70-90 及 90-110 區間值。雖然，利用統計資料可以看出嚴重事故發生的比率極低，並觀察出各類事件的單一肇事因子。然而，透過本研究所提出的架構更可以出探勘出那些肇事原因常在嚴重的死亡車禍事件中出現，卻極少或是沒有出現在一般事故中或是反之亦然。藉此找出嚴重事故與一般事故之差異性規則，以對交通管理者提出建議。

事故資料經整理後，依本研究流程計算每一事故之肇事嚴重程度值，依權重支持度進行關聯規則分類，再以成長率找出突顯樣式，最後分析不同年度突顯樣式消長變化趨勢。

二、實驗結果

本研究將實驗分為兩階段來進行，第一階段透過本研究所提的研究架構，分別篩選找出國道 1 號 98 年及 99 年這兩年度交通事故突顯樣式，第二階段為分析突顯樣式在不同年度之變化趨勢前一階段交通事故突顯樣式變化，如圖 1。下面章節將分開作說明。

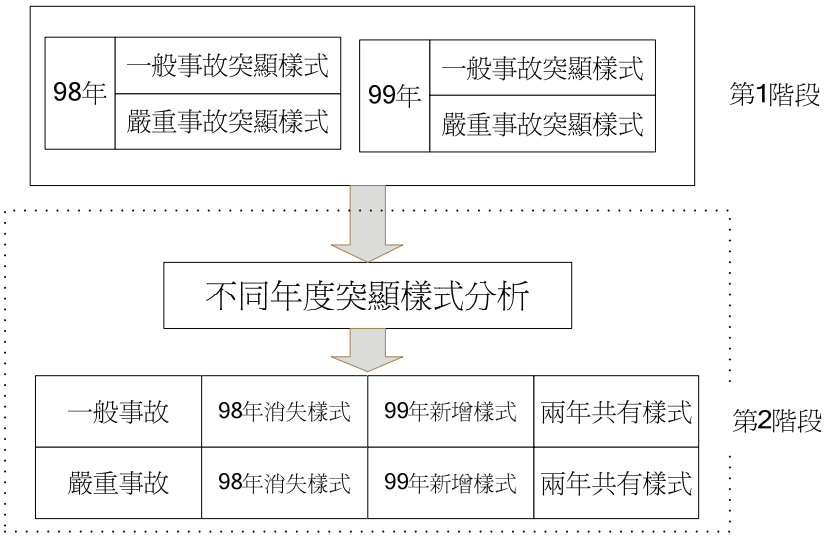

圖 1：實驗步驟圖

本研究透過多次實驗，摘錄其中最小權重支持度分別為 10%與 15%的結果，用圖 2(a)(b)呈現。觀察圖中一般及嚴重事故規則兩集合的規則成長率，成長率 3 時，規則數量銳減，成長率 3 亦能區分出資料集的差異。又實驗顯示在最小權重支持度設為 10%且成長率達 3 以上時，進行關聯規則分類，較能在一般及嚴重事故類別之關聯規則中，找出有趣與重要的樣式。因此擇定其為門檻值，超過門檻值之規則為突顯樣式。

根據事故統計資料，99 年的事故總數多於 98 年的事故。由圖 2(a)(b)的結果中發現 99 年不管在嚴重事故或是一般事故的規則數目都明顯多於 98 年的結果。當事故事件較多時的資料庫，當權重支持度相同時，規則數自然較多，此現象不依成長率門檻選擇而有不同變化。

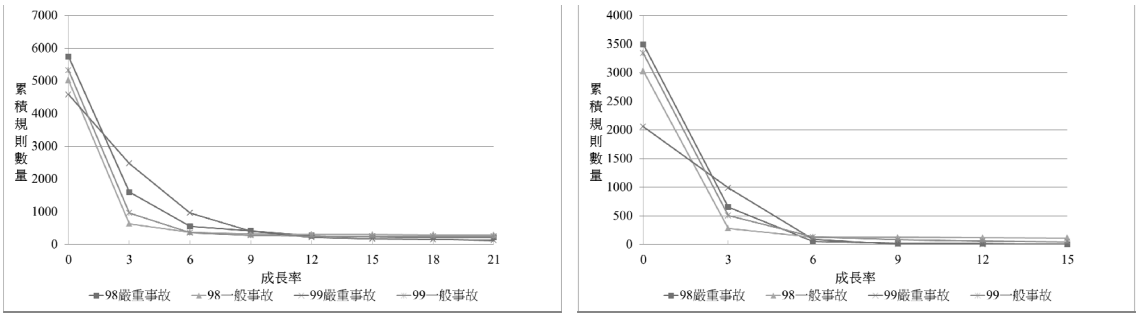

圖 2：最小權重支持度設為(a)10%之實驗結果(b)15%之實驗結果

第 1 階段實驗結果找出 98 年及 99 年突顯樣式，計有 98 年一般事故 633 筆突顯樣式，嚴重事故樣式 1605 筆；99 年一般事故 966 筆突顯樣式，嚴重事故樣式 2477 筆。在跨年度分析中，未發現一般事故規則，在另一年度變為嚴重事故規則者，或嚴重事故規則變為一般事故規則之現象。有了這些規則可以提供給管理單位，制定規則或是加強取締，以減少這方面的事故發生。

本研究請一位資深專家協助摘錄較具代表性之一般事故及嚴重事故突顯樣式探勘結果。由於所挑選的規則事故次數大於等於 2，所以 98 年之傳統支持度（Sup）需大於等於 0.03%，99 年之 Sup 需大於等於 0.02%，其中成長率為∞之規則為跳躍式突顯樣式。接者，邀請三位平均年資 13 年以上的交通流量與事故數據分析專家，驗證本研究挑選出的規則有用性。在表 3，4 呈現的各類事故之突顯樣式規則，都增加了專家的認同度，每條整理出來的規則其平均認同度都超過 4 以上（李克特 5 尺度）。

除此之外，增加探勘 100 年的事故資料，來驗證 98-99 找出規則之真確性。結果分別摘錄於後面章節，本研究發現一般事故的共同突顯樣式，非常穩定出現每一年的事故中。然而，嚴重事故數量極少，易因交通管理措施改變或隨機特殊之個案而影響支持度，以致改變找出之樣式，此為本研究之研究限制。

（一）一般事故突顯樣式

表 3：一般事故突顯樣式變化

編號	一般事故突顯樣式	98 年	99 年	變化	Sup(%) (98→99)	Gsup(%) (98→99)	Ssup(%) (98→99)	成長率 (98→99)	100 年	平均同意值
1	光線 = 夜間有照明 and 車速 = 90-110 and 障礙物 = 無障礙物	∨		消失	20.7→18.6	20.9→18.5	1.8→19.2	11.5→1.0		4.33
2	主要肇因 = 未保持行車安全距離 and 方向 = 北 and 車速=90-110		∨	新增	23.3→22.5	19.6→19.6	15.8→2.0	1.2→9.8	∨	4.33
3	車速=<70 and 視距 = 良好 and 路面鋪裝 = 柏油		∨	新增	18.8→19.3	16.2→17.0	5.5→2.0	2.9→8.5		4.00
4	車種 = 小客車 and 事故類型及型態 = 車與車-追撞 and 假期 = 平日		∨	新增	31.6→28.1	26.7→24.0	10.1→2.7	2.6→8.9	∨	4.67
5	車種 = 小客車 and 事故類型及型態 = 車與車-追撞 and 時段 = 平常		∨	新增	26.6→26.2	22.8→23.2	7.9→0	2.9→∞	∨	5.00
6	時段 = 下午尖峰	∨	∨	共同	21.4→20.4	18.8→17.2	0→5.4	∞→3.2	∨	4.67
7	光線 = 夜間有照明 and 車速=90-110 and 事故類型 = 車與車	∨	∨	共同	19.6→17.1	19.0→16.3	4.1→2.0	4.6→8.2		4.33
8	光線=夜間有照明 and 主要肇因 = 未保持行車安全距離	∨	∨	共同	17.3→15.2	14.9→12.6	0→0	∞→∞	∨	4.67

編號	一般事故突顯樣式	98 年	99 年	變化	Sup(%) (98→99)	Gsup(%) (98→99)	Ssup(%) (98→99)	成長率 (98→99)	100 年	平均同意值
9	車種 = 小客車 and 事故類型及型態 = 車與車-追撞 and 障礙物 = 無障礙物	∨	∨	共同	46.8→45.4	40.5→39.5	6.1→0	6.6→∞	∨	5.00
10	車種 = 小客車 and 事故類型及型態 = 車與車-追撞 and 主要肇因類別 = 駕駛人	∨	∨	共同	46.1→44.5	39.3→38.2	9.5→0	4.1→∞	∨	4.33
11	車種 = 小客車 and 事故類型 = 車與車 and 方向 = 北	∨	∨	共同	30.6→30.2	27.7→27.2	5.5→0	5.0→∞	∨	4.67
12	車種 = 小客車 and 事故類型 = 車與車 and 假期 = 假日	∨	∨	共同	19.8→22.2	18.4→19.3	1.8→4.5	10.1→4.3	∨	4.33
13	主要肇因 = 未保持行車安全距離 and 車種 = 小客車 and 方向 = 南	∨	∨	共同	19.7→19.0	16.2→15.3	1.8→0	8.9→∞	∨	4.33
14	主要肇因 = 未保持行車安全距離 and 車種 = 小客車 and 天候 = 晴	∨	∨	共同	32→27.5	26→22.2	0→0	∞→∞	∨	4.67
15	主要肇因 = 未保持行車安全距離 and 車種 = 小客車 and 假期 = 平日	∨	∨	共同	27.9→24.5	22.5→19.5	0→0	∞→∞	∨	4.67

表 3 為摘錄具代表性之一般事故突顯樣式探勘結果，說明如下：

1. 編號 1 為消失樣式，原因為 Ssup 由 98 年 1.8%大幅增加至 99 年 19.2%，表示 99 年度此樣式之嚴重事故大幅增加，以致 99 年成長率大幅降低變成消失樣式，值得管理單位持續觀察。
2. 編號 2~5 為新增樣式，新增原因均為 99 年度之 Ssup 大幅下降以致成長率高於 3。
   (1) 樣式 3 為車速小於 70、視距良好且為柏油路面之樣式，兩年度 Gsup 達 16.2%以上，但 Ssup 低於 5.5%，可見低速行駛並非絕對安全，但較少發生嚴重事故。
   (2) 樣式 5 為小客車在平常時段發生車與車-追撞之樣式，此樣式在 99 年度之 Ssup 為 0%，代表 99 年度未發生此種嚴重事故，所以是成長率為∞的跳躍式突顯樣式。
3. 編號 6~15 為 98~99 年均共同出現之突顯樣式，其中編號 6、8~11、13~15 為跳躍式突顯樣式。
   (1) 樣式 6 中，98 年為跳躍式突顯樣式，在下午 17~19 時之尖峰時段，Ssup 為 0%，表示發生嚴重事故件次數為 0，但 99 年 Ssup 提高為 5.4%，雖均為兩年度之共同樣式，但值得觀察。

(2) 樣式 7~8 均為夜間有照明之樣式，表示在此條件下，發生事故時，以一般事故為多，尤其樣式 8 之 Ssup 為 0%，代表兩年度均未發生此種嚴重事故，顯示夜間有照明之重要性。

(3) 樣式 9~15 均為小客車之樣式，顯示小客車多為一般事故主角，事故類型及型態多為車與車-追撞，顯示小型車犯這個毛病的比例偏高。

（二）嚴重事故突顯樣式

嚴重事故突顯樣式規則數較多，本研究僅列部分於表 4。

表 4：嚴重事故突顯樣式變化

編號	嚴重事故突顯樣式	98 年	99 年	變化	Sup(%) (98→99)	Ssup(%) (98→99)	Gsup(%) (98→99)	成長率 (98→99)	100 年	平均同意值
1	主要肇因=車輪脫落或輪胎爆裂 and 車種=小貨車	∨		消失	0.05→0.02	13.6→6.7	3.5→3.3	3.9→2.0		4.0
2	車種=聯結車 and 方向=北 and 車速=70-90	∨		消失	0.06→0.02	13.4→4.68	1.9→2.63	9.2→1.78		4.33
3	事故類型及型態=車本身-撞護欄（樁） and 車種=小客車 and 時段=深夜		∨	新增	0.02→0.06	1.8→12.7	1.7→1.6	1.0→7.9		4.67
4	事故類型及型態=車本身-撞護欄（樁） and 車種=小客車 and 主要肇因類別=駕駛人		∨	新增	0.03→0.04	4.2→13.9	2.8→2.1	1.5→6.5	∨	4.0
5	事故類型及型態=車本身-撞護欄（樁） and 車種=小客車 and 車速=90-110		∨	新增	0.06→0.08	8.4→19.8	4.5→5.1	1.9→3.9	∨	3.67
6	主要肇因=酒醉（後）駕駛失控 and 車種=小客車 and 車速=90-110		∨	新增	0.03→0.03	5.9→10.5	3.4→2.3	1.7→4.5		4.0
7	主要肇因=酒醉（後）駕駛失控 and 車種=小客車 and 假期=假日		∨	新增	0.02→0.03	1.8→10.5	1.6→1.4	1.1→7.6		4.33
8	車種=小貨車 and 事故類型=車本身 and 天候=晴	∨	∨	共同	0.06→0.04	16→13.9	4.7→4.4	3.4→3.2	∨	3.67

表格 4 之突顯樣式均與車種相關，消失樣式有 2 筆，新增樣式有 5 筆，共同樣式有 1 筆，說明如下：

1. 編號 1 及 2 為 99 年度消失之樣式，消失原因均為 99 年之 Ssup 降低，以致成長率低於 3 門檻值之故。
2. 編號 3~7 為 99 年度新增之樣式，新增原因均為 99 年之 Ssup 大幅提高至 10.5%以上，所以為 99 年度之的新增規則。

(1) 樣式 3~5 樣式為小客車車本身-撞護欄（樁）之規則，尤其車速為 90-110

km/hr 高速下，99 年嚴重程度高達 19.8%。

(2) 樣式 6~7 樣式為小客車酒醉（後）駕駛失控之規則，其在 90-110 km/hr 高速下或假日條件下，成長率均大於 4.5。

3. 編號 8 跨年度之共同嚴重事故突顯樣式，編號 8 為小貨車在天候為晴時發生之車本身之事故規則，值得交通管理單位觀察。

在本次實驗中找出 98~99 年一般及嚴重事故突顯樣式，用意於找出兩者差異性及樣式變化趨勢，由以上探勘規則顯示，在一般事故主角仍為小客車追撞事故，顯示小客車犯這個毛病的比例偏高，雖非嚴重事故，但是對於財產損失或造成交通壅塞等社會成本仍不小，仍值得有關單位參酌加強防治。在嚴重事故中其它如高速行駛、天候晴於嚴重事故探勘出，並影響事故嚴重性；嚴重事故突顯樣式如小客車酒醉（後）駕駛失控及小客車撞護欄（樁）樣式於 99 年度新增樣式及小貨車及車本身為兩年度均有之共同樣式。對於新增嚴重事故樣式及持續發生之突顯樣式，建議有關單位若欲降低肇事嚴重程度，應將資源集中投入此樣式之肇事防治。

伍、結論

在分類技術中，大多數研究都在建構在均衡分布的資料集中，提出新的模型會方法以提升分類的精準度。然而，這類研究所建構的分類模型，會在不平衡資料集中失去優勢，造成分類不準。但是在很多實際應用時，不平衡資料集卻占多數，像客戶流失問題、信用卡詐欺問題、疾病識別等，甚至本研究所採用之高速公路交通事故資料庫更是一極不平衡的資料集。因此，在此類不平衡資料集中的分類模型建構是很重要的課題。除此之外，分類模型為了增加準確度會產生很多分類規則，這也會讓決策者較難採用，這也是分類方法的一個困境。本研究提出一創新的分類架構可以在不平衡資料集中探勘出突顯樣式，這架構可以解決上述的兩大問題。

本研究架構結合關聯規則分類與突顯樣式探勘的概念，提出一個創新的研究架構可以有效區分找出兩個類別之間的差異性，篩選出分別隸屬這兩類資料集的獨特識別規則。然而過去沒有相關研究提出方法在不平衡資料集上作突顯樣式探勘。本研究透過高速公路交通事故資料庫當作實證研究對象，將事故分為一般事故與嚴重死亡事故兩類，其中死亡事故佔整體事故很低的比例。因此本研究以權重支持度進行關聯規則分類，並配合突顯樣式分析交通事故，探勘一般及嚴重事件規則及差異，最後分析事故因子規則之趨勢變化。

本研究在高速公路交通事故資料庫探勘篩選出重要樣式，呈現與整理給專家協助解讀，並整理出有用樣式在本研究的實驗結果中。透過本研究的結果發現，

在大部份高速公路事故中，道路及交通環境並非嚴重事故與一般事故差異主因，排除天候、時間等不可抗因素後，其它如機械及人為因素可能才是主因，管理單位若能採取因應措施，加強對用路人宣導、加強對車輛機械管理及配合執法單位查緝，相信將可降低或避免嚴重事故發生。

本研究的結果可以發現一般事故的突顯樣式，不但專家認同度高且在 100 年探勘的資料驗證中，也證明這些肇因依然存在。然而，嚴重事故數量極少，易因交通管理措施改變或隨機特殊之個案而影響支持度，以致改變找出之樣式。這部分為本研究之研究限制。除此之外，本研究實證資料受限於資料取得為高速公路交通事故資料庫的限制，對於交通事故資料當事人受傷等級、財務損失、駕駛人性別、年齡、起迄旅程、保護裝備及國道 1 號道路幾何特性數化資料無法取得，以及未進行車種、時段、路段、假期等欄位與通行量正規化處理，這些都為本研究之限制。在有限條件下，本研究僅能以現有之事故資料比較不同年度之事故嚴重程度影響因子，找出因子間關聯性，並針對不同年度影響差異作更詳細分析，並探討其變化情形，作為管理單位在制定交通管理改善措施之參考。未來，若基於這樣的模型，加入更多元化的資訊進來探勘，相信可以達到更豐富的資訊。

致謝

感謝審查委員無私的付出，提供許多的寶貴建議使本論文之內容更臻完美；本研究承蒙國科會專案部分經費贊助（計畫編號：NSC 100-2410-H-031 -010 -MY2；NSC 102-2410-H-031 -058 -MY3），謹致謝忱。

參考文獻

陳文杰（民 93），『應用資料挖掘技術於高速公路交通肇事次數之研究』，未出版碩士論文，國立嘉義大學運輸與物流工程研究所，嘉義市。

周雍傑（民 89），『以類神經網路探討都市地區肇事嚴重程度之研究』，未出版碩士論文，國立成功大學交通管理研究所，台南市。

黃昶斌（民 93），『以類神經網路探討都市地區肇事嚴重程度』，未出版碩士論文，國立交通大學交通運輸研究所，新竹市。

黃湄清（民 94），『利用資料探勘技術於台灣地區肇事危險判別之研究』，未出版碩士論文，國立中央大學土木工程研究所，桃園縣。

戚培芳（民 86），『中山高速公路肇事分析模式之研究』，未出版碩士論文，國立交通大學交通運輸研究所，新竹市。

林大煜（民 71），『臺灣地區道路交通事故分析及建立電腦資訊系統之研究』，交通部運輸研究所，台北市。

林郁志（民 87），『都市地區肇事嚴重程度之分析研究-以臺南市為例』，未出版碩士論文，國立成功大學交通管理研究所，台南市。

陳志和（民 88），『都市地區肇事嚴重程度預測模式之研究』，未出版碩士論文，國立成功大學交通管理研究所，台南市。

楊思瑜（民 92），『小型車事故特性分析及嚴重程度預測模式之研究—以桃竹苗地區為例』，未出版碩士論文，逢甲大學交通工程與管理研究所，台中市。

吳冠宏、吳信宏、郭廣洋（2006），『應用分群技術於交通事故資料分析』，品質學報，13 卷，第三期，頁 305-312。

吳冠宏、吳信宏、郭廣洋（2004），『應用資料挖掘於交通事故資料分析』，中華民國品質學會第 40 屆年會高雄市分會第 30 屆年會暨第 10 屆全國品質管理研討會論文集，高雄，台灣，頁 72-81。

蘇志哲（民 92），『易肇事地點改善作業手冊之研訂』，交通部運輸研究所，台北市。

Abdelwahab, H.T. and Abdel-Aty, M.A. (2001), 'Development of artificial neural network models to predict driver injury severity in traffic accidents at signalizes intersection', *Transportation Research Record*, Vol. 1746, pp. 6-13.

Agrawal, R., Imilienski, T. and Swami, A. (1993), 'Mining association rules between sets of items in large databases', *Proceedings of ACM SIGMOD International Conference on Management of Data*, Washington, USA, May 26-28, pp. 207-216.

Agrawal, R. and Srikant, R. (1994), 'Fast algorithms for mining association rules', *Proceedings of the Twentieth International Conference on Very Large Data Bases*, Santiago, Chile, September 12-15, pp. 487-499.

Ali, K., Manganaris, S. and Srikant, R. (1997), 'Partial classification using association rules', *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, Newport Beach, California, USA, pp. 115-118.

Alhammady, H. (2007), 'Mining streaming emerging patterns from streaming data', *Proceedings of the IEEE/ACS International Conference on Computer Systems and Applications*, Amman, Jordan, May 13-16, pp. 432-436.

Anderson, T.K. (2009), 'Kernel density estimation and K-means clustering to profile road accident hotspots', *Accident Analysis and Prevention*, Vol. 41, pp. 359-364.

Antonie, M.L., Zaiane, O.R. and Coman, A. (2003), *Associative Classifiers for Medical Images, Mining Multimedia and Complex Data*, Springer Berlin Heidelberg, New York, US.

Cao, L., Zhao, Y. and Zhang, C. (2008), 'Mining impact-targeted activity patterns in imbalanced data', *IEEE Transaction on Knowledge and Data Engineering*, Vol. 20, No. 8, pp. 1053-1066.

Ceci, M., Appice, A., Caruso, C. and Malerba. D. (2008), 'Discovering emerging patterns for anomaly detection in network connection data', *Proceedings of the Seventeenth International Symposium*, Toronto, Canada, May 20-23, pp. 179-188.

Chang, L.Y. and Wang, H.Y. (2006), 'Analysis of traffic injury severity: an application of non-parametric classification tree techniques', *Accident Analysis and Prevention*, Vol. 38, pp. 1019-1027.

Chen, S.S. and Huang, C.K. (2013), 'An efficient model for mining precise quantitative association rules with multiple minimum supports', *International Journal of Innovative Computing, Information and Control*, Vol. 9, No. 1, pp. 207-222.

Chong, M.M., Abraham, A. and Paprzycki, M. (2004), 'Traffic accident analysis using decision trees and neural networks', in Isaias, P. et al. (Eds.), *IADIS International Conference on Applied Computing*, IADIS Press, Portugal, Vol. 2, pp. 39-42.

Chong, M.M., Abraham, A. and Paprzycki, M. (2005), 'Traffic accident analysis using machine learning paradigms', *Informatica*, Vol. 29, pp. 89-98.

Delen, D., Sharda, R. and Bessonov, M. (2006), 'Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks', *Accident Analysis and Prevention*, Vol. 38, No. 3, pp. 434-444.

Depaire, B., Wets, G. and Vanhoof, K. (2008), 'Traffic accident segmentation by means of latent class clustering', *Accident Analysis and Prevention*, Vol. 40, pp. 1257-1266.

Dong, G. and Li, J. (1999), 'Effcient mining of emerging patterns: discovering trends and differences', *Proceedings of the Fifth International Conference Knowledge Discovery and Data Mining*, San Diego, CA, USA, August 15-18, pp.43-52.

Dong, G., Li, D. and Wong, L. (2005), 'The use of emerging patterns in the analysis of gene expression profiles for the diagnosis and understanding of diseases', *New Generation of Data Mining Applications*, IEEE Press/Wiley, New Jersey, US, pp. 331-354.

George, T., Ioannis, K. and Ioannis, V. (2011), 'PolyA-iEP: a data mining method for the effective prediction of polyadenylation sites', *Expert Systems with Applications*, Vol. 38, No. 10, pp. 12398-12408.

García-Borroto, M., Martínez-Trinidad, J. and Carrasco-Ochoa, J. (2012), 'A survey of emerging patterns for supervised classification', *Artificial Intelligence Review*, October, No. 6, pp. 1-17.

Han, J. and Kamber, M. (2006), *Data Mining:Concepts and Techniques*, 2nd Ed., Elsevier, Boston, US.

Hu, Y.H. and Chen, Y.L. (2006), 'Mining association rules with multiple minimum supports: a new mining algorithm and a support tuning mechanism'. *Decision Support Systems*, Vol. 42, No. 1, pp. 1-24.

Hu, Y.H., Wu, F. and Liao, Y.J. (2013), 'An efficient tree-based algorithm for mining sequential patterns with multiple minimum supports', *Journal of Systems and Software*, Vol. 86, No. 5, pp. 1224-1238.

Huang, C.K. (2013), 'Discovery of fuzzy quantitative sequential patterns with multiple minimum supports and adjustable membership functions', *Information Sciences*, Vol. 222, pp. 126-146.

Koh, Y.S. and Rountree, N. (2005), 'Finding sporadic rules using apriori-inverse', *Proceedings of the Ninth Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Hanoi, Vietnam, May 18-20, pp. 97-106.

Li, W., Han, J. and Pei, J. (2001), 'Accurate and efficient classification based on multiple class-association rules', *Proceedings of the IEEE International Conference on Data Mining*, San Jose, CA, USA, November 29-December 2, pp. 369-376.

Li, J. and Wong, L. (2002), 'Identifying good diagnostic gene proups from gene expression profiles using the concept of emerging patterns', *Bioinformatics*, Vol. 18, pp. 725-734.

Liu, B., Hsu, W. and Ma, Y. (1998), 'Integrating classification and association rule mining', *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, New York, USA, August 27-31, pp. 80-86.

Liu, B., Hsu, W. and Ma, Y. (1999), 'Mining association rules with multiple minimum supports', *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Disconvery and Data Mining*, San Diego, CA, USA, August 15-18, pp. 337-341.

Mussone, L., Ferrari, A. and Oneta, M. (1999), 'An analysis of urban collisions using an artificial intelligence model', *Accident Analysis and Prevention*, Vol. 31, No. 6, pp. 705-718.

Nefti, S. and Oussalah, M. (2004), 'A neural network approach for railway safety prediction', *Proceedings of IEEE International Conference on Systems, Man and Cybernetic*s, Hague, Netherlands, Oct 10-13, Vol. 4, pp. 3915-3920.

Romero,C., Romero, J.R., Luna, J.M. and Ventura, S. (2010), 'Mining rare association rules from e-learning data', *Proceedings of the 3rd International Conference on Educational Data Mining*, Pittsburgh, PA, USA, June 11-13, pp.171-180.

Solomon, S., Nguyen, H., Liebowitz, J. and Agresti, W. (2006), 'Using data mining to improve traffic safety programs', *Industrial Management & Data Systems*, Vol. 106, No. 5, pp.621-643.

Sohn, S.Y. and Lee, S.H. (2003), 'Data fusion, ensemble and clustering to improve the classification accuracy for the severity of road traffic accidents in Korea', *Safety Science*, Vol. 41, pp. 1-14.

Szathmary, L., Napoli, A. and Valtchev, P. (2007), 'Towards rare itmeset mining', *Proceedings of the Ninteenth IEEE International Conference on Tools with Artificial Intelligence*, 2007, Patras, Greece, October 29-31, pp. 305-312.

Tao, F., Murtagh, F. and Farid, M. (2003), 'Weighted association rule mining using weighted support and significance framework', *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Washington, DC, USA, August 24-27, pp. 661-666.

Thabtah, F., Cowling, P. and Hammoud, S. (2006), 'Improving rule sorting, predictive accuracy and training time in associative classification', *Expert Systems with Applications*, Vol. 31, No. 2, pp. 414-426.

Troiano, L., Scibelli, G. and Birtolo, C. (2009), 'A fast algorithm for mining rare itemsets', *Proceedings of the Ninth International Conference on Intelligent Systems Design and Applications*, Pisa, Italy, November 30-December 2, pp. 1149-1155.

Veloso, A., Meira, W.Jr. and Zaki, M.J. (2006), 'Lazy associative classification', *Proceedings of the Sixth IEEE International Conference on Data Mining*, Hong Kong, China, December 18-22, pp. 645-654.

Wang, J. and Karypis, G. (2005), 'HARMONY: efficiently mining the best rules for classification', *Proceedings of the Fifth SIAM International Conference on Data Mining*, Newport Beach, California, USA, April 21-23, pp. 205-216.

Weng, C.H. (2011), 'Mining fuzzy specific rare itemsets for education data', *Knowledge-Based Systems*, Vol. 24, pp. 697-708.

Xie, Y., Lord, D. and Zhang, Y. (2007), 'Predicting motor vehicle collisions using Bayesian neural network models: an empirical analysis', *Accident Analysis and Prevention*, Vol. 39, pp. 922-933.

Yin, X. and Han, J. (2003), 'CPAR: classification based on predictive association rules', *Proceedings of the 3rd SIAM International Conference on Data Mining*, San Francisco, CA, USA, May 1-3, pp. 331-335.

Yun, H., Ha, D., Hwang, B. and Ryu, K.H. (2003), 'Mining association rules on significant rare data using relative support', *The Journal of Systems and Software*,

Vol. 67, pp. 181-191.

Zhou, L. and Yau, S. (2007), 'Efficient association rule mining among both frequent and infrequent items', *Computers & Mathematics with Applications*, Vol. 54, pp. 737-749.