地理學報　第三十八期：93-109 (2004)
JOURNAL OF GEOGRAPHICAL SCIENCE (38): 93-109 (2004)


# 預測型模式在空間資料探勘之比較與整合研究

## 以集集大地震引致山崩之空間資料庫為例

# A Comparitive and Integrated Study of a Predictive Model in Spatial Data Mining

## The Case of Chi-Chi Earthquake-induced Landslide Spatial Database

鄒明城*　　孫志鴻**

Ming-Cheng Tsou　　Chin-Hong Sun

**Abstract**

Using a single model to forecast spatial phenomena will not produce good estimation in the prediction of individual pixels, even with good overall accuracy. A new strategy, which combines several models based on different philosophies, not only reduces the uncertainty of predictive modeling but also improves its accuracy. This study integrates a Decision Tree algorithm , the Artificial Neural Network, the Bayes Classfier, and Exemplar-based Concept Learning, with each individually applied to a spatial data warehouse. The results of each model and two kinds of modeling-integration methods, including horizontal integration and vertical integration, were then evaluated. In a case study, we chose Chi-Chi earthquake-induced landslide to test the prediction accuracy and obtained good results.

* 工業技術研究院能源與資源研究所副研究員
Associate Researcher, Energy and Resources Laboratories, Industrial Technology Research Institute.
** 國立臺灣大學地理環境資源學系教授
Professor, Department of Geography, National Taiwan University.

We believe that the same methodology can also been used in other cases of environment issues for which there is plentiful GIS digital data.

**Keywords: data mining, Geographic Information System, earthquake-induced landslide, predictive model.**

## 摘　要

使用單一預測模型在空間現象的預測上，即便有不錯的整體預測率，但不保證能在個別的像素或網格上產生良好的預測結果。本研究提出一個新的策略，將不同設計哲學的模式，包含決策樹演算法、類神經網路、貝氏分類器以及案例式概念學習等四個模式加以整合，並且以資料倉儲為資料探勘的基礎平臺，各個模式以及水平、垂直整合之預測結果分別被加以評估及比較。透過這樣的整合方式不僅可以減少模式的不確定性，更提升了預測上的正確性。再以集集大地震引致山崩之空間資料庫為案例，來進行資料探勘預測模式效能的評估。獲致良好的結果，相信類似的方法論可以應用在其他具有豐富空間資料的環境議題研究上。

**關鍵字：資料探勘、地理資訊系統、地震山崩、預測型模式**

## 前　言

由於過去數十年來，有關地球之研究文獻、數位資料、衛星影像等資訊成長極快，但大眾的應用卻極少。因此美國前任副總統高爾先生，於 1998 年 1 月 31 日，在加州科學中心的演講中，即以「數位地球：在二十一世紀瞭解我們的星球 (The Digital Earth: Understanding our planet in the 21st Century)」(The Second Interagency Digital Earth Workshop, 1999) 為題，呼籲其政府及民間各界共同研發並整合現有資源，將許多數據型資料，包括環境、社會、經濟、人口等，轉化為易理解的、可快速查詢、並能提供地理空間參考資料的「數位地球」資訊，以協助人們了解週遭環境所面對的問題 (Gore, 1999)。

迎接數位地球的來臨，並不是只要把所有空間資訊機械式的拼湊在一起，而是要把原始的數據轉化為可以理解的知識，這也是目前 GIS 所面臨的問題。人們對於空間資料庫的應用已不滿足於僅對資料庫進行查詢與檢索。僅用查詢檢索不能幫助用戶從資料中找出帶有結論性的有用資訊。這樣，資料庫中蘊藏的豐富知識，就得不到充分的發掘與利用，形成「資料豐富而知識貧乏」的現象。另外，從人工智慧的角度來看，專家系統雖已應用於部份空間問題的解決上，但是，知識獲取仍是專家系統研究上的瓶頸。傳統上，知識的獲取是透過知識工程師對於領域專家的訪談，再由其中擷取出專家的經驗與知識並轉換成可以輸入電腦的格式，這樣的過程不但冗長繁複，而且很容易錯失掉部份資訊。因此，有必要考慮從資料庫中發現新的知識，這就是資料庫知識發現 (KDD, Knowledge Discovery in Database) 觀念，也叫做資料探勘 (Data Mining) (Chen *et al.*, 1996)，這是一個迅速發展的新領域，綜合了機器學習、資料庫、專家系統、模式識別、統計、基於知識的系統 (knowledge-based system) 以及視覺化等領域的相關技術 (Koperski *et al.*, 1996; Miller and Han, 2001)，已廣泛的應用在當前許多商業

問題的解決上。在空間問題的研究上也有同樣的需求與研究，所謂的空間資料探勘與知識發現 (Spatial Data Mining and Knowledge Discovery, SDMKD) 也就因應而生，它們的處理方式不同於一般的屬性資料，特別著重於空間的意涵，企圖由空間資料庫中找出不明顯記錄且有趣的樣式與特徵，目前正逐漸受到重視，研究的方向從基礎空間資料探勘演算法的研究 (Adbelmoty *et al*.,1993; Lu *et al*., 1993; Ester *et al*. 1997; Koperski *et al*., 1998; Miller and Han, 2001)，到各種空間相關的應用研究，如地圖資訊的擷取 (Malerba *et al*., 2001)、遙測影像資訊擷取、環境特徵的製圖 (Eklund *et al*., 1998)、時空樣式的擷取 (Openshaw, 1994)、空間互動模式分析 (Arentze *et al*., 2000; Smyth, 2001) 等，領域十分廣泛。

Han 與 Kamber (2000) 將資料探勘模式依其產生結果的應用分為二大類，一為預測型模式 (predictive)，它是根據目前的資料狀況，建立能預測未來現象、樣式或趨勢改變的模式。另一則為描述型模式 (descriptive)，它用來描述資料與現象，並找出其間的關係以及樣式 (pattern)。Weiss 與 Indurkhya (1998) 另將資料探勘預測模式分為三類，分別是 1. 數學式，2. 距離式，3. 邏輯式。數學方法與邏輯方法均係建立在新案例度量上的直接運算，數學式模式主要以加法或乘法的數學運算為主要運算方式，邏輯式則建立在屬性值的布林或邏輯值比較與運算上，而距離式方法則建立在新案例與儲存案例相似度的度量上，此三種方法各有優缺點。本研究嘗試以預測型模式為研究的對象，並以集集大地震引致的山崩為案例，利用地理資訊系統的空間資料處理能力，將山崩資料與相關背景空間資料予以轉化整合，建立成提供資料探勘使用的資料倉儲。再分別使用三種資料探勘技術中各具代表性的預測型模式，有屬於數學式的類神經網路 (Neural Network)、距離式的案例式概念學習 (Exemplar-Based Concept Learning)、與邏輯式的決策樹 (Decision Tree) 等技術，另外再加上統計技術上的貝式分類器 (Bayes Classfier) 等四種預測模式，分別探討他們在空間資料探勘上的效能，最後再將各模式加以結合成一整合模式，以達到提昇預測正確率的效果。

## 研究架構

本研究之架構如圖 1，在進行研究之前，先儘可能蒐集相關學者在空間資料探勘與地震引致山崩之研究文獻，了解相關技術應用的限制與可行性及引致山崩機制的環境因子探討。透過這些資料的分析，建立接下來研究上相關背景知識的基礎。有了這些背景知識後，便進行相關影響環境因子資料的蒐集，資料來源多樣化，包括地理圖層資料以及屬性資料，先進行必要的前處理，再將這些資料分別建置進入地理資訊系統與關聯式資料庫中，做為資料探勘的基礎資料。

相關的基礎資料為了做進一步的處理，以建立適合資料探勘技術使用的乾淨資料，須先將各個向量型地理圖層資料轉化為網格資料、清理空的資料，將不必要的屬性資料剔除，同時也利用現有的資料再衍生出新的資料欄位。這些經過清理乾淨後的龐大資料，必須加以整合後再輸入資料倉儲中，透過適當的採樣技術後，可做為各資料探勘模式使用的資料來源。接下來根據使用目的的不同，分別依格式輸入所需要的資料。

預測型模式包含有，類神經網路、決策樹演算、案例式概念學習以及貝氏分類器等四種。首先以決策樹演算法模式找出具有預測力的環境因子，再以這些因子提供其他模式建立上輸入環境因子選擇

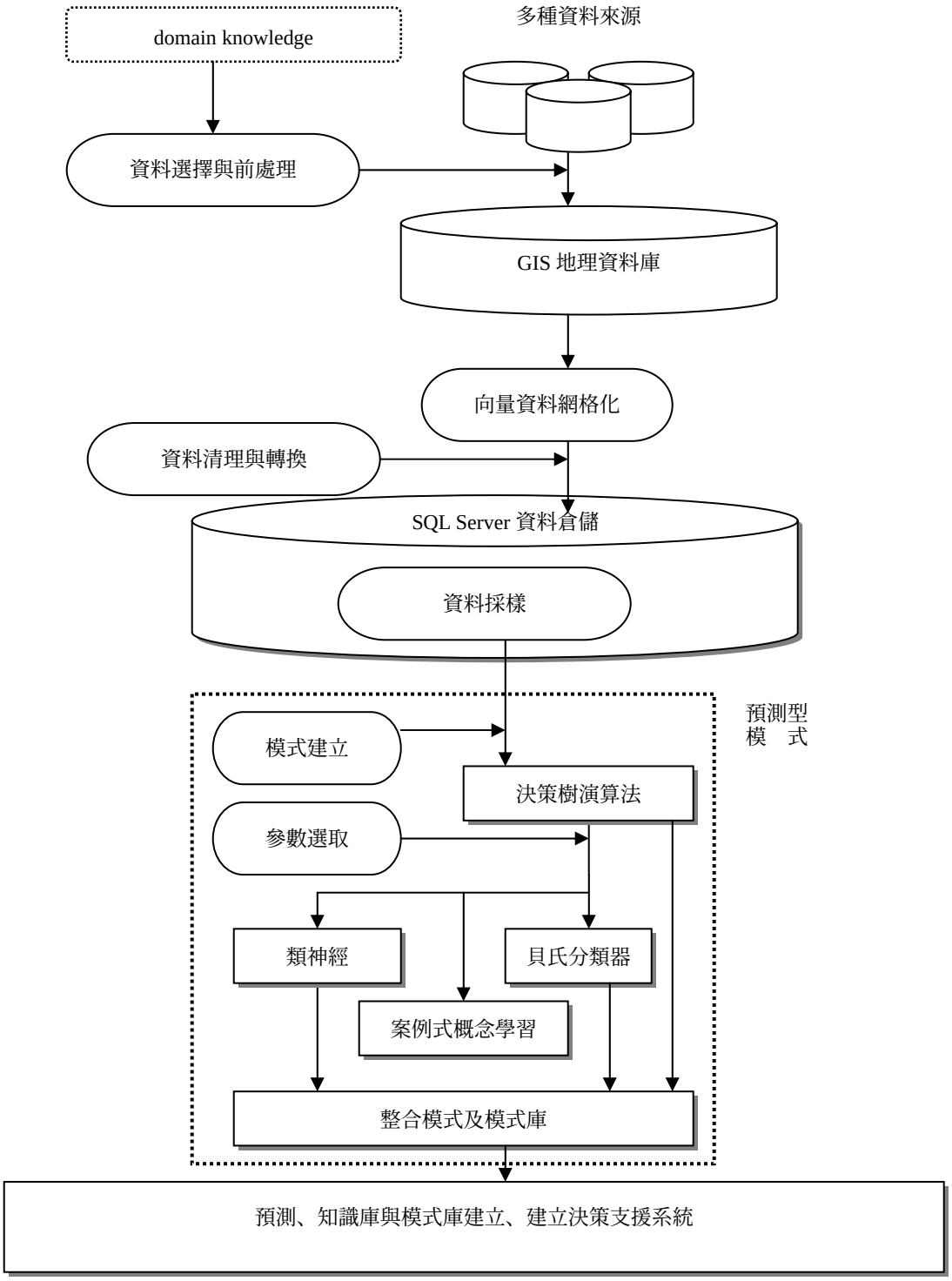
domain knowledge
多種資料來源
資料選擇與前處理
GIS 地理資料庫
向量資料網格化
資料清理與轉換
SQL Server 資料倉儲
資料採樣
預測型
模　式
模式建立
決策樹演算法
參數選取
類神經
貝氏分類器
案例式概念學習
整合模式及模式庫
預測、知識庫與模式庫建立、建立決策支援系統

**圖 1　研究架構**

的參考，適度的降低資料維度，以提升探勘效率。各模式分別參考這些因子，輸入必要的資料以建立模式並且進行測試，接著再將這些模式予以整合成單一運算結果，並且進行測試，以便獲得最佳的預測效果，並建立成空間決策支援系統所需的模式庫與知識庫。

# 模式建立

## (一) 決策樹演算法

決策樹演算法是一種以遞迴方式把資料集合切分成越來越具同質性次集合的分類演算法，其模式結構以樹狀展開，子節點為切割的資料集，終端節點又稱葉節點，並以能取得最高增益比值的節點來做為樹狀結構成長分支的依據 (Quilan, 1993)，所產生的樹狀結構相較於其他演算法 (不論是迴歸類的公式或是類神經網路完全封閉式的黑箱結構)，比較容易讓人理解 (Openshaw and Openshaw, 1997)，這一點其實是相當重要的，因為在進行研究時，有時候後不只是希望有準確的模型就好，對決策者來說，更重要的是能否從規則的內容中獲得啟發，此時，規則是否能夠以人類所能理解的形式呈現就顯得相當重要了。決策樹分岔的準則是根據增益比值 (gain ratio) 來計算，它首先根據每一節點選擇一個具有最大資訊增益值 (information gain) 屬性，對於每一節點下純度尚未淨化的資料集 $S$ 之熵值計算如下：

$$Entropy(S) = \sum_{i=1}^{c} - P_i \log_2 (P_i)$$

$P_i$ 是指預測因子選項 i 對於分類資料的切分率，即類別 i 在資料集 S 中所佔的比率，c 為類別的數目，增益值計算如下：

$$Gain(S, A) = Entropy(S) - \sum_{v} \frac{|S_v|}{|S|} Entropy(S_v)$$

其中，|S| 為資料集 $S$ 樣本數，$|S_v|$ 為資料集中屬性 A 等於 v 的樣本數。

Entropy 表示亂度，這個觀念來自熱力學，用來表示物體分佈的分散狀況，亂度越高，則越無規則，因此決策樹演算法的目標就是希望能夠降低資料分類結果的亂度。比較各個屬性值的增益比值，挑選具有最高增益比值的因子來當做分岔依據。

一旦決策樹建立完成後，它可以被轉換成一組組如以下形式的法則：

*if (坡度 < a) and (距離車籠埔斷層距離 < b) and (地質分佈 = c) then*

*山崩預測 = 是*

這裏 a、b 為數值資料，c 為類別型資料

由於決策樹演算法需要計算增益比值，有挑選具增益因子的能力，能從一大堆因子中篩選出重要的因子，故可與其它模式進行整合 (Berry and Linoff, 2000)。本研究所使用的建置方法則是以 Microsoft SQL Server 的 Analysis Service 來建置 Decision Tree 模式，它屬於微軟 SQL Server 2000 資料庫管理系統的套件之一，是微軟進軍資料探勘商業智慧的解決方案之一，除了決策樹演算法外，還包括聚類分析

模組。可以透過微軟 DSO (Decision Support Object) 之 COM 元件，以 VB Net 程式來呼叫使用，其呼叫操作語法與 SQL 類似，可方便的進行建模、預測與評估。

### (二) 類神經網路

在商業、科學和學術理論中，類神經網路的研究持續受到關注，這是因為類神經網路已被證實用於數值的預測和連續性的演進均有良好的成效，對於非線性問題的處理也有良好的表現，本研究選擇以倒傳遞網路 (Back-Propagation Network; BPN) 做為數學式模式的代表，建構其中之一個預測模式。

類神經網路提供一個數學模式來嘗試去小型化人腦。知識通常被表示成互相連接所成的多層處理器，這些處理器常被視做神經元，以表示人腦神經中的關聯性。每個節點各有加權值連接至鄰接層中的其他點，個別的節點接收從所連接的節點輸入，並且使用一個簡單且包含加權的函數來運算輸出值。

倒傳遞網路 (Back-Propagation Network; BPN) 屬於監督式的學習網路，是目前類神經網路學習模式中應用最廣的模式，由於它的「隱藏層」，使得網路可以表現輸入處理單元間的相互影響。通常給定的隱藏層數夠多，理論上可以逼近任何的函式 (Hagan *et al*., 1995)，但是事實上，處理的單元數越多，越慢達到收斂的狀態，隱藏層的單元數目一般可用 (輸入＋輸出) /2、或√ (輸入 ㄨ輸出) (葉怡成，1993)，此外一般問題取一層隱藏層已經足夠。

本研究關於類神經網路的部份，是以 C++程式自行建立模式及操作介面，再透過之前決策樹演算法增益比值的計算後，以決策樹演算法所找出的因子設定為類神經網路模式的輸入因子的值，在經過正規劃化成 0-1 之間的實數，透過類神經網路內部權重矩陣的運算後，得到一個介於 0-1 之間輸出值 (1：表山崩，0：表未山崩)，它代表了發生山崩與不會發生崩的可能值，越接近 1 者可能性越高。經過嘗試與錯誤多次調整後，以一層隱藏層、30 個內部節點所構成的網路結構獲得最佳的效果，以此做為模式評估與整合的依據。

### (三) 案例式概念學習

案例式概念學習演算法屬於距離式資料探勘技術，具有處理數值與類別型資料而無需做轉換的能力，較一般案例式推理 (Case Based Reasoning) 型的資料探勘技術執行上較有效率，非常適用於類似本研究案例之使用。

案例式概念學習會根據訓練資料集 (Training Data Set) 中標示為分類欄位的值來建立分類別，產生一個三層狀的樹狀結構，圖 2 為這個樹狀結構的大略型式。在資料層裏的每個節點，代表資料庫裏的每一筆資料範例，用以定義概念層的概念類別。在概念層裏的節點，則摘要統計其在資料層的子節點所包含的屬性值。樹的根節點則儲存了整個資料的摘要資訊。

案例式概念學習所使用的評估方式為類別相似分數 (以距離的觀念來評估類別內案例間彼此的接近程度)，它被儲存在根節點和概念層的節點裏。類別相似分數可以評估全部概念類別的相似度。運作方式大致如下；

1. 假設經輸入訓練集資料產生樹狀結構，有：
   a. 一群概念層節點 C1, C2…Cn。
   b. 將 C1, C2…Cn 的類別相似分數相加，然後除以 n 得到平均類別相似分數 S。

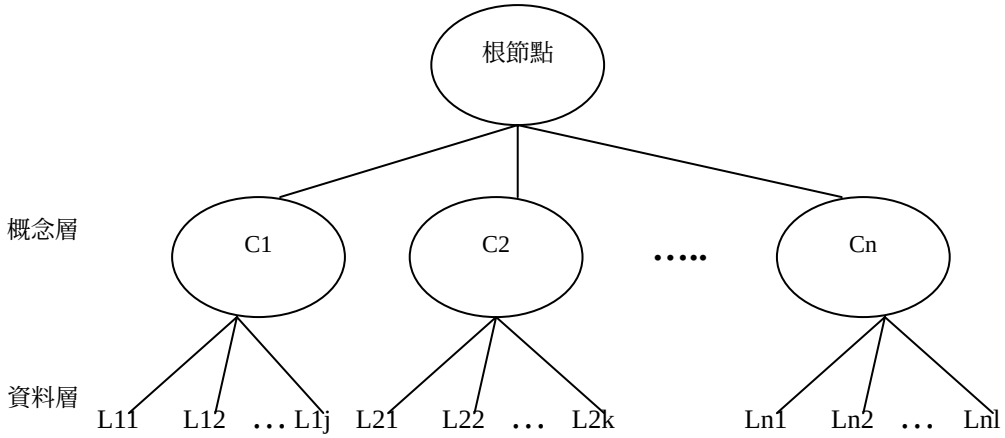


**圖 2　案例式概念學習法架構**

c. 一筆要被分類的新資料案例 I。

2. 將 I 指定為某個概念節點的資料範例，並且使得平均類別相似分數 S 最小，則該案例 I 就可以被歸類至此分類至此概念節點，完成分類的任務。

類別相似度計算公式如下：

$$FR(C) = 2/(N \times (N-1)) \times \Sigma\ Sim(a,b)$$

FR (C)：為某一類別 C 的類別相似分數

N: 為類別 C 內所有的案例數

Σ Sim (a, b)：為類別 C 內所案例間彼此相似度分數的總合

$$Sim(E1, E2) = E1\ INT\ E2 / ((E1\ INT\ E2) + (E1\ DIF\ E2))$$

Sim (E1, E2)：為兩個案例間的相似度

E1, E2: 為二個個別的案例

INT: 代表 E1, E2 二案例之間的交集，即，屬性資料相同的個數

DIF: 代表 E1, E2 二個案例中，屬性資料不同的個數

關於交集與差集個數的計算，對於類別型的屬性資料，只需判斷彼此是否相等即可，對於數值型的屬性資料，則以彼此數值差的絕對值除以儲存在母節點所紀錄該類別的標準差之值，並且超過設定的閥值來求得，因此可以同時針對數值型、類別型、混合型以及資料不完全的紀錄 (missing data item) 進行建模與分類。

本研究關於案例式概念學習的部份，是以 Roiger 與 Geatz (2002) 所著之 *Data Mining – A Tutorial-Based Primer* 一書中所附的 ESX 程式為工具，以 Excel 軟體做為運作的平臺，透過 VBA 程式的撰寫，做為介面的設計與建模操做的工具。並以之前決策樹演算法增益比值的計算後所列因子的值做為輸入值，以山崩與否做為計算的輸出值 (為零或一的輸出)。

### (四) 貝式分類器 (Bayes Classfier)

貝式分類器係統計方法的一種，提供一個簡單但強而有力的監督式分類技術。此模式假設所有的輸入屬性彼此間是獨立，而且具有同等的重要性，雖然這樣的假設並不是很正確，但在實際運用上仍可提供可接受的結果，是一個相當被廣泛應用的監督式分類技術。其方法是建立在貝氏理論的基礎 (Bayes Theorem) 上，說明如下：

$$P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)}$$

其中 H 是要被檢定的假設，E 是有關假設的證據。

由分類觀點來看，被檢定的假設是應變數，代表被預測的類別，證據是由輸入的屬性值所決定，$P(H|E)$ 是 $H$ 為真的情況下得到證據 E 之條件機率。$P(H)$ 是先期機率 (a priori probability)，表示任何證據出現之前的假設機率。對於數值資料的處理，貝式分類器是以機率密度函數來計算條件機率，假設某屬性值為常態分佈時，條件機率計算如下：

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

其中，e 為指數函數；μ 為給定屬性的平均值；σ 為屬性類別的標準差；x 為屬性值

本研究關於貝氏分類器的部份，為自行撰寫的 VB.Net 程式，除了計算分類的類別外，並可輸出隸屬類別的機率值，以供模式整合之用。

### (五) 整合模式

在資料探勘的專案中不宜拘泥單一演算法的改進，而喪失了其他演算法可能帶來不一樣的看法與知識增益。Weiss 與 Indurkhya (1998)、Berry 與 Linoff (2000) 及 Roiger 與 Geatz (2002) 均認為如果預測能力是最高追求目標的話，則採用結合數個模式來運作或結合其運算結果，可以獲得令人滿意的答案。

整合的方式可以有兩種，一種是垂直整合，將數個具有互補功能的模式整合成單一個較強的模式，所獲得的答案只有一個。另一種做法為水平式的整合，分別使用數個方法來做預測，各自獲得答案，最後的答案是透過整合數個模式之答案而獲得。

水平整合上常見的做法，係針對單一演算模式，利用靴帶式引導建模法 (Bagging) 將訓練資料切割成大小相等的數個次訓練資料集，再以各個次訓練資料集來建立多個模式，最後再將各個模式的預測結果以投票的方式來決定，當被預測為某分類獲得多數票時，則認定為該分類。當預測的結果為數值資料時，則使用各個模式的預測值加總平均。這樣的做法雖然簡單，但因各模式的效能可能參差不齊，原本較佳的模式可能因而被較差的模式所降低，而且僅針對單一演算模式。本研究採垂直與水平且多種演算模式整合並用的做法，在垂直整合上先以決策樹演算法做為前導，計算各因子的增益值，挑選具影響力的因子，作為其它模式輸入因子的參考，以降低資料的維度與複雜度，提升計算上的效率。而水平整合方式，則係分別求得各個演算模式預測所得到分佈於 0 與 1 之間的機率值乘以加權值

後，再將各組答案加總起來獲得最後的結果。另外，本研究亦分別計算投票式與算術平均的整合方式以供做為效能分析上的比較。對於各個演算法的處理如下：

**1. 決策樹演算法：**

對於決策樹演算法來說，是以 Microsft SQL Server 的 Analysis Service 來計算預測類別的機率值，故若預測結果是山崩，該值為發生山崩的機率值，若預測結果是非山崩，則該值為不會發生山崩的機率值。為求統一計算方便，我們將它們全部轉為山崩的機率值，即將原來非山崩的機率值改為山崩機率值，例如，若預測結果是非山崩且機率為 0.9，則經過反算後山崩機率只有 0.1，即 (1-0.9) = 0.1。

**2. 類神經網路：**

類神經網路因其輸出值原本就是位於 0 與 1 之間的值，越接近 1 者，山崩機率越高，故無需經過反算，該輸出值可以直接拿來與其他模式結果進行加權運算。

**3. 案例式概念學習**

案例式概念學習演算法會為每個輸入欲預測的資料，計算該筆紀錄在被計算出來的預測分類中之代表性分數，表示此一紀錄對其他在同一類別裏的所有資料範例的平均相似度，它的值也世介於 0-1 之間，但並不代表為機率值，為了取得其機率值，我們以每一類別中獲有最高代表性分數的值當作該類別的代表分數，或稱做類別原型 (class prototype)。將每一個資料範例的代表性分數除以類別原型的代表性分數，即可獲得每一資料範例的可信度或機率值，接著再按照之前的做法，將機率值轉換為山崩機率值，即可與其他模式的結果整合運算。

**4. 貝式分類器**

貝氏分類器的演算原就是從機率的觀點來計算的，故可以計算預測類別的機率值，然後再按照先前的轉換方式，轉換為山崩的機率。

**5. 結果值的加權加總**

結果值的加權加總是一個嘗試錯誤的過程，必須經過多次的調整與測試才能獲得最佳的結果，這是一個最佳化的求取，由於本研究只有使用四個模式，經過幾次的排列組合後即已獲得最佳的結果。最後發現，若以決策樹演算法的結果權重佔 50%，類神經演算法的結果權重佔 30%，貝式分類器分類值權重佔 15%，案例式概念學習演算法的結果權重佔 5%，加總之後的分數可以獲得高於任一個別模式所得的預測結果 (詳細評比參見稍後說明)。儘管沒有明確的規律定義權重的配比，但是可以發現，原則上對於較高預測率的模式，給以較高的權重，通常可以獲得更理想的正確率。

# 案例研究—集集大地震引致山崩地理資料庫

## (一) 研究區域資料概述

由於集集大地震所誘發的山崩大多集中在臺灣中部區域，故以臺灣中部山區為研究區域。所涵蓋的範圍為東經 120°36’ 到 120°01’，北緯 23°33’ 到 24°20’ 之間。(如圖 3)。

本研究參考之前文獻 (童啟哲，2001；許煜煌，2002；鄒明城與孫志鴻，2004)，一共選擇了十七個因子圖層進行空間資料庫的建立 (如表 1)，做為模式的輸入因子，其中較特殊的是除了部份原生圖

層外，還包括了由原圖層衍生出來的新圖層，如九格平均坡度、坡向與九格最大最小坡度、坡向差。而崩塌地的部份，則是以工業技術研究院能源與資源研究所 (2000) 受農委會水土保持局之委託，辦理集集大地震崩塌地調查與治理規劃，在經過 1999 年 4 月 9 日至 7 月 24 日間之災前衛星影像 (SPOT) 與 9 月 27 日災後之衛星影像及航空照片判釋後，所提出的 21,969 筆變異點網格化之後的資料。在完成以上各圖層之後，再以 ArcView 分別將以上十七個因子之向量式主題圖層予以網格化，為了配合 DTM 的大小，將每一個圖層的解析度均設為 40m ×40m。之後，再以程式將各個圖層網格結合成一筆一筆的紀錄，每一筆紀錄內之每一個欄位均對應到某一個主題圖的格網值，其資料量非常的龐大，共計有約 122 萬餘個網格，以這做為研究的母體，其中屬於崩塌的網格大約只有約 6 萬個。然後再透過 SQL Server 之 Data Transformation Service 轉換之後，可以確保資料的適用性，這些紀錄資料在轉進 SQL Server 資料庫管理系統中，可成為資料倉儲，透過資料庫管理的功能，可以很方便的處理這些資料以提供資料探勘時的資料來源。

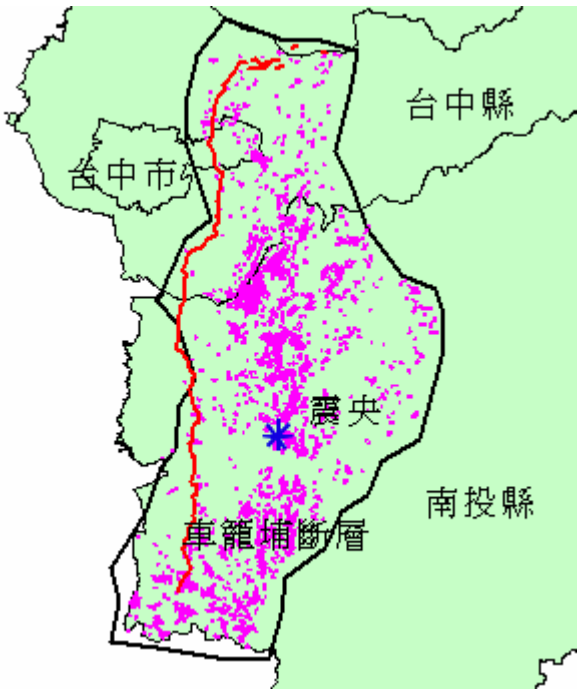


**圖 3　研究區域及地震山崩圖**

表 1　地震山崩之輸入圖層

| 原生因子圖層 | 衍生因子圖層 |
|---|---|
| 高程 | 九格點之平均坡度 |
| 坡向 | 九格點之最大最小坡度差 |
| 坡度 | 九格點之平均坡向 |
| 距離車籠埔斷層距離 | 九格點之最大最小坡向差 |
| 距離斷層破碎帶距離 | 地震強度 (Arias Intensity) |
| 距離道路距離 | |
| 距水系距離 | |
| 距震央距離 | |
| 地質分佈狀況 | |
| 垂直向地表加速度 | |
| 東西向地表加速度 | |
| 南北向地表加速度 | |

## (二) 資料化簡與採樣

雖然，資料探勘是想從大量資料中找尋知識，也就是說資料量越大越有潛力獲得較佳的結果，但是，許多學者 (Weiss and Indurkhya, 1998; Berry and Linoff, 2000; Groth, 2000) 均認為，大量的資料並不能保證獲得比小量資料更佳的預測結果，大多建議以較小量的資料來建立，至少一開始應該如此。除了計算上及時間上的成本太高之外，Weiss 與 Indurkhya (1998) 認為，預測型模式會試著去迎合 (fit) 大量的資料，即便是隨機取樣的資料，也含有許多例外，故資料量越大所含的例外也就越多，為了迎合這樣的資料型態，所得的模式也就越容易出錯。

本研究所面對的資料母體數高達 122 萬餘筆，其中屬於山崩的部份不到 6 萬筆，為了處理這樣大量的資料，即便是目前的電腦軟硬體技術，處理上亦力有未逮。因此需要採取採樣的技術來簡化資料，但是，山崩資料相對於母體而言過於稀少，若採一般隨機抽樣，可能造成對於稀少事件的不易掌握。Berry 與 Linoff (2000) 建議可以採用超採樣 (Oversampling) 的技術，增加樣本中稀有事件的比率，他認為維持在 10% 至 40% 的比率，通常可以獲得不錯的結果。故本研究將資料分成山崩與未山崩二部份，為了方便於 Excel 中做分析，因此分別從屬於山崩的六萬筆資料中隨機抽樣 2 萬筆、另從未山崩之 116 萬筆資料中抽樣 4 萬筆混合成 6 萬筆樣本。由於本研究所使用的四個預測型模式皆為監督式分類模式，必須有訓練資料作為模式建立的學習樣本以及測試資料作為模式驗證的基準，故由六萬筆採樣資料集中的隨機選取五萬筆為訓練資料集，做為模式建構的基礎，待模式建置完成後，再以其它另外一萬筆資料做為模式驗證與評估依據。

# 模式效能評估與驗證

資料探勘技術之所以有價值，可以說它的本質就是在協助利用過去的經驗來找出「稀有事件」，而且，所要尋找的稀有事件一定能夠為決策者帶來高度價值或是嚴重損失，但是因為太稀少，所以更需要資料探勘技術來找出它，發生於企業界的有顧客流失、呆帳風險、電話銷售等例子，而本研究也有這樣的特性，山崩與未山崩的資料比率達 3:97，山崩本身相對於未山崩資料就是一個稀有事件，而且對於國家社會會造成嚴重的損失，因此一些在應用在資料探勘技術上的評估標準，也適用於本研究中。驗證的方式如下；

## (一) 各模式效能評估

資料探勘上有所謂的 commission error 及 omission error, commission errorr 就是把未山崩的資料預測為山崩，而將原為山崩資料當作未山崩則是 omission error。對於本研究來說，通常 omission error 的成本比 commission error 來的高，因此評估本研究模式的第一步，就必須先從錯誤狀態的分類著手，而使用的工具為錯差矩陣 (表 2)。其中包含了二個最重要的指標，即回應率 (response rate)、反查 (recall)

**表 2　錯差矩陣**

| | 實際**山崩**數 | 實際未**山崩**數 |
|---|---|---|
| 預測為**山崩**數 | A | B |
| 預測為**非山崩**數 | C | D |

**回應率：**即在預測為山崩的集合中，多少是屬於真正的山崩，回應率越高代表模式效果越好。公式如下：

response_rate = A/ (A+B)

commission error = 1 – response_rate

**反查：**為在所有實際山崩集合中，有多少比率被模式找出來，越高越好。這個指標對本研究特別重要，回應率高或正確率高不見得模式就好，更重要的是反查要高，如何把稀有事件找出來更重要，本研究中山崩資料只佔全部資料不到 3%，若有一個模式把所有結果都預測為不會山崩，那麼這個模式就會有高達 97% 的正確率，但是這樣的模式對於決策者毫無意義。對於山崩的預測分類不同於一般的分類問題，它的精神在於對於山崩的真正掌握率。公式如下：

recall = A/ (A+C)

omission error = 1 - recall

表 3、4、5、6、7 分別為案例式概念學習、類神經網路、貝氏分類器、決策樹與整合型模式的錯差矩陣，表 8 為各個模式之評估指標。可以發現整合型模式有最高的正確率，達到 89%，其次為決策樹演算法，有 86% 的正確率，再其次為類神經網路，正確率為 84%，而貝氏分類器雖然有空間獨立與常態分佈這樣的統計假設限制，但仍然有 78% 的正確率，最差的為案例式概念學習演算法，正確率只有 70%。正如先前所述，本研究透過對於各模式計算結果值加權平均而獲得最高的正確率，較次高

的決策樹演算法提高了約 3%。基本上各模式的反查與回應率表現排名大致與正確率排名相符，但特別值得注意的是，案例式概念學習演算法雖然只有 70% 的正確率，但卻擁有 87% 的反查率，也就是說該模式可以掌握高達 87% 的山崩資料，而貝氏分類器雖然有 78.1% 的正確率，高於案例式概念學習演算法，但是貝氏分類器的反查率確只有 81.6%，也就是說貝氏分類器的 omission error 較大，對於山崩的掌握，案例式概念學習反而較貝式分類器好甚至高於類神經網路。

**表 3　案例式概念學習錯差矩陣**

| | 實際山崩 | 實際未崩 |
|---|---|---|
| 預測山崩 | 4362 | 2263 |
| 預測未崩 | 654 | 2722 |

**表 4　類神經網路錯差矩陣**

| | 實際山崩 | 實際未崩 |
|---|---|---|
| 預測山崩 | 4241 | 835 |
| 預測未崩 | 775 | 4150 |

**表 5　決策樹錯差矩陣**

| | 實際山崩 | 實際未崩 |
|---|---|---|
| 預測山崩 | 4479 | 618 |
| 預測未崩 | 537 | 4367 |

**表 6　貝氏分類器錯差矩陣**

| | 實際山崩 | 實際未崩 |
|---|---|---|
| 預測山崩 | 4093 | 1268 |
| 預測未崩 | 923 | 3717 |

**表 7　整合模式錯差矩陣**

| | 實際山崩 | 實際未崩 |
|---|---|---|
| 預測山崩 | 4536 | 628 |
| 預測未崩 | 480 | 4357 |

**表 8　各模式對於地震山崩預測之評估指標比較**

| | 正確率 | 反　查 | 回應率 |
|---|---|---|---|
| 案例式概念學習 | 70% | 86.9% | 65.8% |
| 貝式分類器 | 78.1% | 81.6% | 76.3% |
| 類神經網路 | 83.9% | 84.5% | 83.6% |
| 決策樹 | 86.2% | 88.8% | 84% |
| 整合模式 | 89% | 90.4% | 87.8% |

### (二) 水平整合效能評估

本研究分別進行了加權平均、算術平均與投票法等三種整合方式，發現加權平均具有最高的正確率，較個別模式中最高的決策樹演算法提高了將近 3% 的正確率，算術平均整合只稍微提高了一點正確率，投票法甚至還不如決策樹演算法，反而將正確率拉低了，這是由於各模式輸出的結果必須先以類別的型態來輸出，再進行投票表決，部份模式之輸出必須先由數值型態轉換為類別型態，方可進行整合，在轉換的過程中捨去掉部份具有機率概念的資料。由此可知，使用機率值並以加權平均的整合方式，具有最佳的效能，可以將正確率提升至 89%。至於不同整合方式之反查與回應率的排名基本上與正確率相同，均較正確率為高。

**表 9　不同水平整合方法之比較**

| | 加權平均整合 | 算數平均整合 | 投票法整合 |
|---|---|---|---|
| 正確率 | 89.0% | 86.6% | 85.8% |
| 反查 | 90.4% | 87.5% | 86.5% |
| 回應率 | 87.8% | 85.4% | 82.6% |

### (三) 垂直整合效能評估

由於決策樹演算法需要計算各因子的增益比值，有挑選具增益因子的能力，故能從一大堆因子中篩選出重要的因子，透過決策樹演算法分析顯示，除了與坡向有關的三個因子外 (即坡向、九格平均坡向、九格最大最小坡向差)，其餘的輸入因子均對地震山崩有所影響，且影響程度各不相同，其程度由大至小排列，分別為地質分佈、九格平均坡度、距車籠埔斷層距離、Ia 值、南北向地表加速度、東西向地表加速度、垂直地表加速度、距震央距離、高程、九格最大最小坡度差、距道路距離、距水系距離、距斷層破碎帶距離、坡度。本研究以決策樹演算法分析出來的因子提供其他模式輸入的參考，做為垂直整合的方法，把不具重要性的因子排除。並以最具影響力的地質分佈因子與不具影響力的坡向因子做對比說明，由表 10 可發現，雖然去除掉 3 項與坡向有關的因子，但對於各模式之正確率影響卻非常小，均在 0.05% 以內，因為此三項因子透過決策樹演算法的增益分析後，並不具影響力，已被排除在決策樹建構的節點外。相反的，如果將決策樹演算法所計算出最具影響力的單一因子 (地質分佈) 去除掉的話，則降了將近 2% 的正確率 (表 11)。故可知決策樹演算法對於重要因子的挑選與垂直整合上確有貢獻，可降低模式建立上資料的維度，提升演算效率，並提供對於各輸入因子於輸出結果影響程度大小之了解。

表 10 去除 3 項與坡向相關因子 (不具影響力) 後之各模式評估

| | 案例式學習 | 類神經 | 貝氏分類器 |
|---|---|---|---|
| 去除 3 項與坡向相關因子後之正確率 | 69.7% | 77.6% | 83.6% |
| 使用全部因子之正確率 | 70% | 78.1% | 83.9% |

表 11 去除地質分佈 (最具影響力) 單一因子後之各模式評估

| | 案例式學習 | 類神經 | 貝氏分類器 |
|---|---|---|---|
| 去除地質分佈單一因子後之正確率 | 68.2% | 76.4% | 82.5% |
| 使用全部因子之正確率 | 70.0% | 78.1% | 83.9% |

# 結論與建議

1. 面對如研究區這樣資料量的大的研究範圍時，適當的抽樣是必須的。本研究對於相對稀少的山崩資料以超採樣 (Oversampling) 的方式進行。
2. 透過預測型資料探勘技術的分析，在地震山崩的預測上可以獲得令人滿意的結果。四種預測模式決策樹演算法、類神經網路、貝氏分類器、案例式概念學習分別有 86%、84%、78%、70% 的正確率。
3. 將模式計算結果以機率值輸出，並搭配各機率值的加權平均水平整合後，可以將正確率提升至 89%。經以嘗試錯誤方式多次權重配比實驗後發現，對於正確率較高者給予較高的權重，採決策樹演算法、類神經網路、貝氏分類器、案例式概念學習四模式的權重分別為 0.5、0.30、0.15、0.05 時，可獲得最佳的效果。未來，當模式數目增加時，關於權重配比最佳化的計算可考慮以遺傳演算法來自動求得。
4. 以決策樹演算法優先建模，做為模式垂直整合的方法，透過決策樹演算法的增益比計算後發現，除了與坡向有關的因子外，其餘的輸入因子均對地震山崩有所影響，且影響程度有所差別，這些因子提供了其他模式在建模時因子選擇上的參考，降低資料維度提升演算效率。
5. 本研究係以預測型資料探勘模式為比較研究的對象，並以集集大地震區域相關地理資料庫做為測試案例，故提出模式之適用範圍應為研究區域附近，未來遭受類似斷層錯動引起地震影響下，所引致山崩危險的預測，礙於大範圍資料收集上的困難，所使用的因子仍屬有限，未來若能加入如植被、覆土厚度、覆土性質、地下水狀況等因素，當能提供更具說服力的結果說明，但本研究在方法論以及評估結果比較上，仍可做為其他地區預測模式建立上的參考。

# 引用文獻


工業技術研究院能源與資源研究所 (2000) **九二一震災系列調查（一）崩塌地調查治理規劃**，行政院農業委員會水土保持局。

許煜煌 (2002) **以不安定指數法進行地震引致坡地破壞模式分析**, 國立臺灣大學土木工程研究所碩士論文。

葉怡成 (1993) **類神經網路模式應用與實作**，臺北：儒林圖書。

童啟哲 (2001) **應用地理資訊系統於地震引致坡地破壞多變量模式分析**, 國立臺灣大學土木工程研究所碩士論文。

鄒明城、孫志鴻 (2004) 資料探勘技術在集集大地震引致山崩之研究，**地理學報**，36: 117-131。

Adbelmoty, A. I., Williams, M. H. and Paton, N.W. (1993) Deduction and deductive databases for geographic data handling, *Advance in Spatial Databases. Lecture Notes in Computer Science*, No. 692, 441-464.

Arentze, T. A., Hofman, F., van Mourik, H., Timmermans, H. J. P. and Wets, G. (2000) Using decision tree induction systems for modeling space-time behavior, *Geographical Analysis*, 32 (4): 52-72.

Berry, M. J. and Linoff, G. S. (2000) Mastering Data Mining: The Art and Science of Customer Relationship Management, New York: Wiley.

Chen, M.-S., Han, J. and Yu, P. S. (1996) Data mining: an overview from a database perspective, *IEEE Transactions on Knowledge and Data Engineering*, 8 (6): 20-35.

Eklund, P. W., Kirkby, S. D. and Salim, A. (1998) Data mining and soil salinity analysis, *International Journal Of Geographical Information Science*, 12: 247-268.

Ester, M., Kriegel, H.-P. and Sander, J. (1997) Spatial data mining: a database approach, *Proceedings of the 5th International Symp. On Spatial Database (SSD'97)*, 47-66.

Gore, A. (1999) *The Digital Earth Vision, The Second Interagency Digital Earth Workshop*, http://www.digitalearth.gov. [ 1 March 2000]

Groth, R. (2000) *Data Mining – Building Competitive Advantage*, New York: Prentice Hall.

Hagan, M. T., Demuth, H. B. and Beale, M. (1995) *Neural Network Design*, New York: PWS Publishing Company.

Han, J. and Kamber, M. (2000) *Data Mining: Concepts and Techniques*, New York: Morgan Kaufmann Publisher.

Koperski, K., Adhihary, J. and Han, J. (1996) Spatial data mining: progress and challenges survey paper, *SIGMOD'96 Workshop on Ressearch Issues on Data Mining and Knowledge Discovery*, 35-50

Koperski, K., Adhihary, J. and Han, J. (1998) Mining knowledge in geographical data, *Communication of ACM*, http://db.cs.sfu.ca/sections/publication/kdd/kdd.html [ 1 April 2000]

Lu, W., Han, J. and Ooi, B.C. (1993) Discovery of general knowledge in large spatial database, *Proceedings of the Far East Workshop on Geographic Information Systems*, 275-289.

Marlerba, D., Esposito, E., Lanza, A. and Lisi, F. (2001) Geographic data mining and knowledge: an overview. In: Miller, H. J. and Han, J. (eds.) *Geographic Data Mining and Knowledge Discovery*, New York: Taylor and Francis, 291-314

Miller, H. J. and Han, J. (eds.) (2001) Geographic data mining and knowledge: an overview, *Geographic Data Mining and Knowledge Discovery*, New York: Taylor and Francis, 3-33.

Openshaw, S. (1994) Two exploratory space-time attribute pattern analyzers relevant to GIS, *Spatial Analysis and GIS*, London: Taylor and Francis, 83-104.

Openshaw, S. and Openshaw, C. (1997) *Artificial Intelligence in Geography*, New York: John Wiley and Sons.

Quinlan, J. R. (1993) *C4.5: Programs for Machine Learning*, New York: Morgan Kaufman.

Roiger, R. J. and Geatz, M. W. (2002) *Data Mining-A Tutorial-Based Primer*, New York: Addison Wesley.

Smyth, C. S. (2001) Mining mobile trajectories. In: Miller, H. J. and Han, J. (eds.) *Geographic Data Mining and Knowledge Discovery*, New York: Taylor and Francis, 337-361.

Weiss, S. M. and Indurkhya, N. (1998) *Predictive Data Mining-A Practical Guide*, New York: Morgan Kaufmann.

93 年　4 月 27 日　收稿

93 年　8 月 20 日　修正

93 年 11 月 25 日　接受