資料探勘的發展與挑戰

■翁慈宗

如果把資料看成是地表的泥土和石頭，
而資訊是隱藏在裡面的礦藏，
資料探勘就是把埋藏在地表下的
礦藏挖掘出來。

資料探勘的定義

現今社會的各行各業在處理日常事務時，幾乎都離不開資料。例如到便利商店購物，或到圖書館借還書，即使是日常使用的手機裡，也存放著各式各樣的資料。主要是因爲資料的儲存裝置已非常便宜，不需再一個位元一個位元地斤斤計較。

但伴隨而來的是，資料的產生和儲存的速度遠超過人們所能分析和消化的速度。在這樣一個周遭都是資料的環境中，要擔心的已不是手邊沒有資料可以分析，而是煩惱如何有效率地把手邊的資料轉化成有用的資訊。因爲有很多的資訊隱藏在資料的背後，如果能從資料中挖掘出有用的資訊加以運用，可以增加個人或組織的競爭力。

以台灣較具全球競爭力的半導體和面板產業爲例，工廠在製造產品時，爲了生

資料探勘就好像是把埋藏在地表下的礦藏挖掘出來，因此資料探勘也稱為資料採礦。

產的效率和提高產品的品質，都會從生產線上不斷地蒐集資料，除了確認生產線是否正常運作外，也可用來做生產線的診斷和改善。但由於這些資料的量相當大，而企業內能正確分析資料的人才又有限，因此這些資料都是經年擺放在電腦的硬碟中，從未分析過。

企業空有這些資產卻無法有效地利用，資料探勘的目的就是希望在堆得像山一樣高的資料中，使用自動或半自動的方式把隱藏在資料中的有用資訊發掘出來。如果把資料看成是地表的泥土和石頭，而資訊是隱藏在裡面的礦藏，資料探勘就是把埋藏在地表下的礦藏挖掘出來，因此資料探勘也稱爲資料採礦。

資料探勘和一般用統計方法來進行資料分析有個很大的不同。統計分析希望能用單一規則或單一模式來顯示資料的性質，但大多數的資料是由多條規則或多個模型混合產生的。資料探勘則是除了運用統計的概念來判讀資料的性質外，更善用電腦的高運算效能，來歸納哪一條規則或

在目前周遭都是資料的環境中，要擔心的已不是手邊沒有資料可以分析，而是煩惱如何有效率地把手邊的資料轉化成有用的資訊。

有很多的資訊隱藏在資料的背後，
如果能從資料中挖掘出有用的資訊加以運用，可以增加個人或組織的競爭力。

探勘的目的是希望在像山一樣高的資料中，使用全自動、半自動的方式發掘資訊。

哪一個模型適合用來解讀哪一部分的資料，如此可使資料的詮釋更具彈性，也較能發掘出眞正隱藏在資料背後的資訊。因此簡單來說，資料探勘就是以統計學和電腦科學爲基礎，所發展出來能快速分析資料的方法。

爲了釐清資料探勘和一般資料檢索的不同，以某一銀行的信用卡發放爲例。如果只是把現有的信用卡使用者做一些資料的整理，以了解目前的使用狀況，那只是一般的資料檢索而已，資料庫管理系統大都會提供相關的資料檢索功能來彙整出這些資訊。

但資料探勘要做的是如何從信用卡使用者的個人、刷卡和付款資料，研判哪些新的申請者應核發信用卡，哪些應拒絕核發。其餘的新申請者則是無法用電腦來自動判斷，因而人工只需處理這一部分的新申請者。如此不但可節省審核的人力，也可進一步了解適合和不適合核發信用卡的主要原因，而一般的資料檢索並無法提供這一類的資訊。

另外以異常值的偵測爲例，從統計學的觀點來看，如果資料服從常態分配，則可以設定離平均值正負三個標準差以上的是異常值。由於可以從一組資料計算出平均數和標準差，因此判定的標準非常明確，很多製造程序或產品都是採用這樣的標準來判斷生產線的運作是否正常。

但如果用這樣的標準來進行網路異常入侵的偵測，由於異常入侵的種類非常多，其異常的行爲必須有一般電腦的使用

行爲來做對照，才能猜測某一作業是否是入侵行爲。更棘手的是，很多異常入侵的方式是新發展出來的，以前從未發生或被發現過，因此無法用一般的資料檢索方式來進行網路的異常入侵偵測。

圖片來源：李男設計

以銀行的信用卡發放為例，資料探勘是從信用卡使用者的個人、刷卡和付款資料，研判哪些新的申請應該核發信用卡。

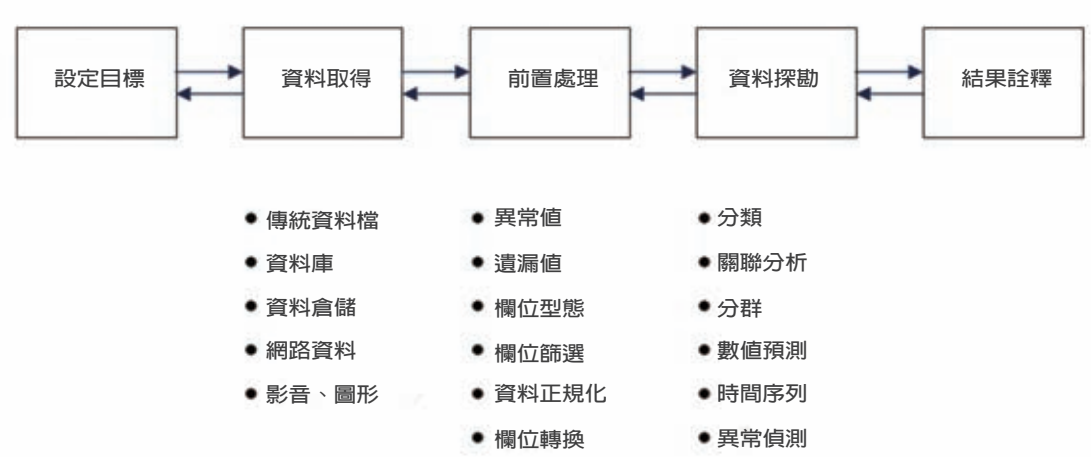
爲了能從資料中挖掘出有用且非顯而易見的資訊，資料探勘的進行主要分5個步驟。首先是設定目標，了解進行資料探勘的目的。因爲在目的不明的狀況下，無法知道應蒐集哪些資料來做分析，或該使用哪一類型的工具來進行探勘。而且資料探勘的成本比使用統計方法或資料檢索的成本高，因此要確認統計方法和資料檢索都無法產生所需要的資訊後，才適合使用資料探勘。

在目標明確之後，接著便是蒐集或整理出適合探勘的資料。由於現在資料的來源有很多種方式，例如是電腦中一筆筆循序擺放的檔案，或是資料庫中彼此有關聯性的檔案，或是放在資料倉儲中的大量歷史資料，或是網頁資料，這些資料可能是集中存放或是散布在多部電腦主機中。資料型態除了傳統的文（字）數字資料外，還有網頁格式、語音和圖像資料。資料探勘的第2個步驟，就是把這些不同來源和格式的資料蒐集起來，並用適當的格式來存放。

蒐集好的資料並不見得就可以進行資料探勘，因爲資料中可能有些有異常，或是有些欄位的值無法取得，這些都會直接影響到資料探勘的結果。因此第3個步驟是做資料的前置處理。這個步驟除了處理異常值和遺漏值外，還要考量欄位是擺放數字還是文字，以及每個欄位的必要性。另外爲了避免欄位的度量對結果造成影響，還要考慮是否進行資料的正規化。例如衡量長度時使用公尺或公分，會使該欄位有一百倍的差距，而這一百倍的差距有可能嚴重影響到該欄位在資料探勘時的重要性。

欄位轉換則是考量到有時候必須結合多個欄位才能產生有用的資訊。例如單純只用年齡、身高或體重單一個欄位，很難判斷一個人是否過重，但如果把這三個欄位轉換成體脂肪一個欄位，這個新的欄位會更適合用來判斷一個人是否過重。因此若能做出適當的欄位轉換，會更容易取得有意義的探勘結果。

資料探勘除了分類、關聯分析及分群這3種公認的類型外，其餘的類型到目前爲止並沒有一致的見解，因此後續會針對這3



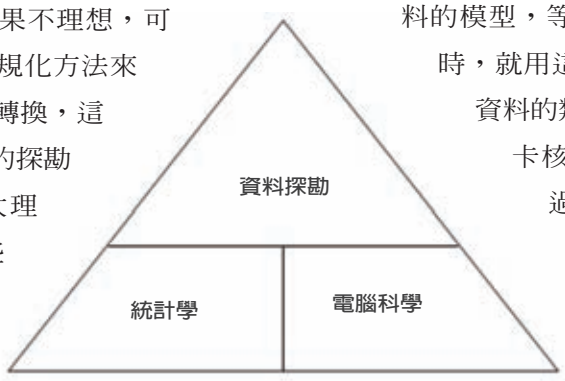
資料探勘的程序

種公認的類型做進一步的說明。對於一開始設定的目標，要了解適用的資料探勘類型是哪一種，否則無法得到有用的資訊。甚至往往爲了能得到有意義的探勘結果，必須結合多種類型的資料探勘工具。因此了解各類型資料探勘的定義和資料分析方法後，才能選擇適當的工具進行資料探勘。

最後對於探勘結果的詮釋，則要倚賴線上工作人員具備的背景知識來解讀探勘的結果。因此這一部分最好由資料探勘的人員和產生資料的人員合作，除了能適切地詮釋探勘結果外，也才能知道該如何把結果應用在實務上。

若發覺探勘中的某一步驟有問題，則應考慮回到上一步驟去檢討和修正。例如認爲資料探勘的結果不理想，可以考慮使用不同的正規化方法來處理資料或進行欄位轉換，這樣或許可以產生較佳的探勘結果。如果還是不太理想，可能是因爲有些有用的欄位沒有被選取，如果能回到資料取得的步驟重新檢視，或許所得的結果就能讓人耳目一新。

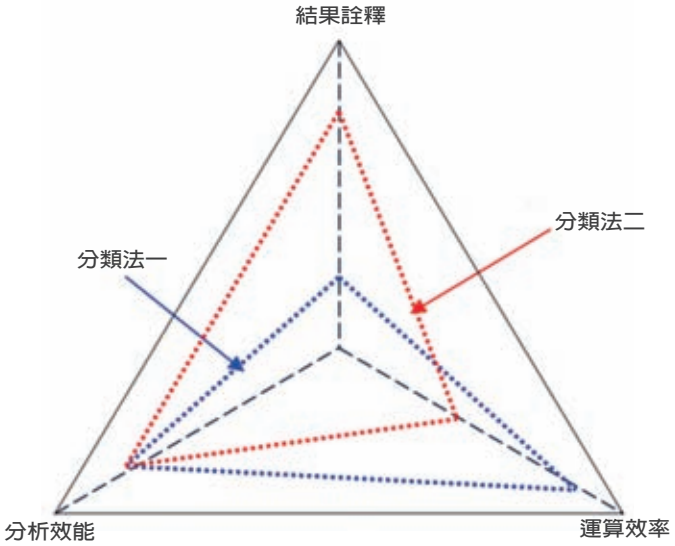
幾乎在所有的資料探勘案例中，都要不斷地回到上一步驟去檢視，並進一步從探勘結果中思考應如何修正才能獲得較佳的結果。在如此不斷來回的測試中，累積對資料探勘的經驗，也才能得到滿足當初所設定目標的結果。



資料探勘、統計學和電腦科學間的關係。

資料探勘的類型

雖然資料探勘的類型該怎麼劃分，到現在還沒有一致的見解，但公認分類、關聯分析和分群3種類型的資料探勘方式，是其中最主要的。分類的方式主要是從現有的資料中，歸納出一個較能解釋這些資料的模型，等將來有新的資料產生時，就用這個模型來預測這筆新資料的類別值。例如前述信用卡核發的例子，就是先透過現有的信用卡申請者和使用者的資料，歸納出一個含有發放、拒絕及待判定3個類



兩個分類法的優缺點比較。這二個分類法的分析效能（即分類正確率）差不多，分類法一的優勢在於運算的效率較佳，可以在較短的時間內學習出分類的模型，而分類法二雖然運算上較費時，但學習出來的分類模型較有助於結果的詮釋，也就是較能了解一筆資料被預測為哪一個類別值的原因。當資料量很大時，處理效率變得很重要，因此選擇分類法一會較合適。但若是資料量並非十分龐大，而且著重的是預測為哪一個類別值的原因說明，就應選擇分類法二來產生分類模型。

別值的模型，然後對於任一新的信用卡申請者，就把他的申請資料放入這模型中進行類別值的推論。

這一類型的分類工具主要包含歸納和推論二個步驟，目的不外乎是希望分類的正確率能提高。以上述的例子來看，就是希望能儘量把信用卡發放給具有消費能力的使用者，而且能儘量拒絕沒有付款能力的申請者，如此可以使信用卡公司有較佳的獲利。

但並非所有分類問題的焦點都是擺在如何提高分類正確率。例如疾病診斷的分類預測，只分成有病和沒病二個類別值。假使一味地追求分類正確率，由於在資料中大多是沒有病的，因此歸納出的分類模型會把大多數的人預測成沒有病。然而一個有病的人若因爲被推論爲沒病而延誤治療的時效性，這種情況所要付出的代價，遠比一個沒病的人被誤診爲有病，到最後證明只是虛驚一場的代價要高得多。因此有些分類的工具會把分類錯誤的成本列爲優先考量，不會只是追求提高分類正確率。

關聯分析最早是由美國的量販店發展出來的，用來了解顧客的購買行爲中，是否會有一些物品存在一起被購買的關聯性。最有名的例子就是，該量販店發覺有一部分的顧客會同時買尿布和啤酒。由於尿布是嬰兒用品而啤酒是大人喝的，因此一開始對這二種物品經常被一起購買感到有些驚訝。但深入了解後才知道，原來大多數有嬰兒的家庭，其男主人假日會一邊帶小孩一邊看電視轉播的運動節目，因此會準備啤酒來喝。

探勘出這樣的購買行爲後，量販店便改變物品擺放的位置，把尿布和啤酒擺在相鄰的貨架或走道，讓顧客較容易找到所需要的物品，因而提高了顧客對該量販店的服務滿意度。因此這種關聯分析一開始又稱爲「購物籃分析」，用來探索各類銷售物品之間的關聯性。

關聯分析主要是處理一個變數只有二個可能值的情況，例如購物籃分析中一筆顧客的交易資料會記錄買了那些物品，有買的物品的值是 1，而沒有購買的物品的值是 0。由於量販店的交易資料往往數以萬計，因此這一類的資料探勘工具，最重要的就是必須很有效率地處理大量資料。

除了購物籃分析外，這種關聯分析也可以用來處理其他同樣是每個變數只有二個可能值的資料。適用的例子如在醫院中，可蒐集病人曾罹患哪些疾病的資料，有患過的病的值是 1，沒有患過的病的值是 0，這樣的資料可用來探勘哪些疾病的發生具有關聯性。

雖然資料探勘工具就是希望能自動分析或過濾大量的資料，但當資料筆數相當龐大時，如何在有限的時間內整理出有用的結果，對任何一個資料探勘工具都是莫大的挑戰。

圖片來源：李易設計

以銀行的信用卡發放為例，透過資料探勘不但可節省審核的人力，也可進一步了解適合和不適合核發信用卡的主要原因，而一般的資料檢索並無法提供這一類的資訊。

曾有一個國外的研究探討當學生某些學科表現不佳時，適合使用哪些輔導措施來提升這些學科的成績，需加強的學科的值是1，表現已經不錯的學科的值是0；有被採用的輔導措施的值是1，沒有被採用的輔導措施的值是0。其探勘的結果可用來了解哪些輔導措施對於加強特定的學科能力可以產生效果，將來才能選擇適合的措施輔導學生。

另一個類似的例子，是把關聯分析用來處理犯人出獄後應採用哪種輔導措施。由於每個犯人犯的罪都不盡相同，適合的輔導措施也不一樣，這樣的探勘結果可以知道什麼樣的輔導措施，較能有效防止哪一類的犯人出獄後再度犯罪，對於犯人和社會來講都是有利的。

資料的分群是希望儘量把相似的資料歸在同一群，並把不相似的資料儘量分在不同群。當針對一整組的資料進行其他的探勘工作時，有時候結果會過於複雜而無法予以適當地詮釋，或是無法得到較精簡且有用的結果。遇到這種情況時，若能先把資料進行分群，然後再對每一群的資料個別進行探勘，往往能得到較有用且清晰的結果。因此分群大都是和其他的探勘工具一起使用。

分群的主要工作就是衡量二筆或二群資料的相似程度，才有辦法判定它們是否適合放在同一群內。但光是衡量相似度的計算方式就有非常多種，採用不同的方式來計算相似度時，得到的結果很有可能大不相同，這是進行資料分群的困難之一。分群的另一個難處是，事先並不知道應把資料分成多少群才是最適當的，而且沒有一個客觀又通用的指標，可以用來衡量目前的分群結果是否是最佳的，只能在參考一些指標後主觀地判定最佳的分群數目。

不管是分類、關聯分析、分群或其他的資料探勘工具，衡量一個資料探勘方法的優劣主要是從3方面：分析效能、運算效率和結果詮釋。現今的資料探勘方法很少能在這3方面都具有絕對的優勢，因此應從這3方面來認識一個資料探勘工具，然後根據自身

只要有資料，就會有探勘的結果。
但這結果到底是泥土還是重要的礦藏，取決於整個資料探勘的過程是否有步步為營，
千萬不要以為把資料輸入資料探勘工具取得結果就完成了。

的需求選擇適合的工具以進行資料的探勘。

資料探勘的未來挑戰

當資料的產生和儲存越來越容易之際，對於資料探勘工具的需求越來越多，要求也越來越高。因此爲了因應資料處理的新需求，資料探勘方法也不斷地演進。

對於一般文數字的資料，現行的資料探勘方法要面對的挑戰是如何處理大量的資料。例如本文一開始提到的半導體或面板業的製程資料筆數便非常多，或現在地球上空各式各樣的衛星，每天不斷地蒐集地表或太空的資料。

雖然資料探勘工具是希望能自動分析或過濾大量的資料，但當資料筆數相當龐大時，如何在有限的時間內整理出有用的結果，對任何一個資料探勘工具都是莫大的挑戰。

除了資料筆數很多之外，另外一個會讓資料量很大的原因是一筆資料內的欄位數非常多。一筆資料所包含的欄位個數稱爲該筆資料的維度，在大多數的資料檔中，一筆資料頂多就是由數十個欄位組成，因此在處理上不會造成太大的問題。但當資料的維度是幾千或幾萬個時，就會變成一個很大的困擾。

高維度的資料首先要面對的就是所謂「維度的詛咒」，當維度越來越高時，任二筆資料會越來越不相似，而且若用統計的方式來檢驗，每一筆高維度的資料都會像是一筆異常的資料。這種「維度的詛咒」，會讓使用統計原理的資料探勘工具變得非常不適合處理高維度的資料。況且維度很高時，一筆資料包含雜訊或記錄不精確的機會便提高，因此要處理高維度的資料探勘工具，必須克服這一方面的問題。

高維度的資料如基因資料，一個人類染色體的基因數在二萬個以上，表示一筆人類基因資料的維度就是二萬多個。又或者在做英文的文件分類時，一般書寫常用的英文字在一萬個以上，也就是一筆資料或一份文件的維度會在一萬個以上。面對這樣高維度的資料，如何快速又正確地處理，是一件非常具挑戰性的任務。

除了龐大的資料量外，現在的資料型態除了文字和數字外，也可以是聲音和影像這類在儲存和處理上都需特別費心的資料型態。尤其是現在的網路非常發達，網頁資料包含文數字、圖形、聲音、影像等各式各樣的資料，而且是提供資料的重要來源。

但目前大多數的資料探勘方法，主要還是用來處理文數字的資料，要把這些方法擴展成可以處理非文數字的資料並不容易。因爲聲音和影像的儲存格式有很多種，要萃取出聲音或影像的欄位或維度以便後續的資料探勘，到現在還沒有一定的標準做法，表示這種非文數字的資料探勘還有很長的一段路要走。

使用資料探勘要有的一個概念是，只要有資料，就會有探勘的結果。但這結果到底是泥土還是重要的礦藏，取決於整個資料探勘的過程是否有步步爲營，千萬不要以爲把資料輸入資料探勘工具取得結果就完成了。如果資料的品質不佳，或對資料探勘工具沒有充分的認識，不僅會讓投注的心力都白費了，還會挖掘到一些產生誤導的資訊，如此就得不償失了。 □

翁慈宗
成功大學資訊管理研究所