Poly Analyst 6.0 實作以本系畢業生成績及心臟病資料爲例

沈永濠、涂韡瀚、李子建、孫一寧
南華大學資訊管理系
李翔詣教授
hylee@mail.nhu.edu.tw
南華大學資訊管理系

摘 要

隨著網際網路以及資料庫技術的發達，人們搜尋資料變得越來越容易，透過網路我們可以快速的獲得大量的資料，於是在資料充斥的情況之下，現代的資訊技術已不再是如何管理大量的資料，而是如何從大量的資料中獲得真正有用的資訊，資料庫系統的成熟導致企業對於資料探勘技術的人才需求越來越殷切，為了提供足夠的資訊給管理人員作為決策的參考開創新的商業契機，資料探勘技術和人才正在快速發展和渴求。應用 Poly 這套軟體從資料庫裡找出有用的資訊來熟練瞭解 Poly 這套軟體，比如經由改變 Poly 軟體上的參數，來分隔找出實驗組和對照組並且互相印證，讓在分析重要屬性的過程中，提昇最後分析出來的結果使正確率能更加提昇，這個步驟讓我們對於 Poly 這套軟體各個功能和參數的改變會對成果有何影響能夠有更深入的了解。

本研究主要希望利用各個資料集來使用 Poly 這個軟體來執行整個資料探勘的流程並且熟悉這套軟體，從學生的資料庫中來執行資料探勘中的資料前處理的部份，再利用心臟病病人資料範本。從心臟病病因分佈情形來執行分析的部份，並藉由分析結果提供病患和醫生做為診斷的參考依據，以這個目的來使用 Poly 軟體分析這筆資料當作實際應用的部份。主要為什麼會使用這兩個資料庫的原因為，心臟病是一個很完整的資料，是由 UCI 網站下載而來，曾經受過多項研究使用，因此是一個已經經過整理，而不需要前處理的步驟便可直接使用的資料庫，主要分析為使用到各種不同專業技術，包含一般的統計分析、叢集分析、決策樹等方法來探討心臟病所引發的症狀分部特性，利用心臟病資料集來分析心臟病屬性對於結果有何影響的過程中，熟析每個參數對於結果的影響、功能的使用方法和注意事項，另外利用學生資料庫的原始檔案來熟息資料探勘的資料前處理的部份，希望可以找出南華大學資管系學生的一些成績上的相關性，並將原始不完全完整的資料做新增、修該、刪除含資料格式轉換的動作。期望從探勘結果和過程中來學習資料探勘的應用並且熟練使用 Poly 這套軟體，在藉著這份研究能讓之後使用 Poly 這套軟體的人能夠作為參考依據。

壹、研究動機

爲了提供幫助學習的資料、增進學習的效率、提供完善的資訊，只有充分的瞭解軟體的的每個功能和參數所影響分析結果的關係，藉此學校可以提供比較完善的教學範例、和使用軟體的說明等，並且針對學生資料和心臟病資料等兩個資料集來直接使用軟體執行整個資料探勘的流程。然而，爲了因應現在時代科技的進步資料探勘流程中以往傳統的統計分析工具已經不敷使用，統計分析只能幫助管理者針對已發生的現象擬定研究假設，並檢定是否達到統計上得顯著意義，進而以樣本之結果回推母群體之現象。所以利用更強的分析工具來挖掘出潛在更能幫助管理人員決策的資訊。

隨著資料庫系統的強大和網際網路的發達，獲得大量的資料越來越容易，但是如何從龐大的資料量中獲得有用的資訊變成現在企業所關注的問題。針對企業中龐大又複雜的資料光是使用傳統統計分析軟體已經無法滿足企業對於資料探勘結果的需求，只有將各種的專業技術統合而成的資料探勘軟體才能符合現代龐大又複雜資料庫的需求，只有將管理與科技的結合之下產生的軟體才能符合現代企業對於資料探勘軟體的需求。

管理、科技、統計是目前影響企業能否成功重要的主軸，更是未來企業是否能夠掌握市場先機的重要關鍵。(台大醫療探勘論文 2004)因此本研究使用了 Poly Anlysist 6.0 這個軟體來作爲學習與應用資料探勘的媒介，並且研究結果能夠提供後來學習資料探勘和使用 Poly 這套軟體的人能夠作爲參考。

貳、相關文獻探討

一、 資料探勘定義

資料探勘(Data Mining)利用資料來建立一些模擬真實世界的模式，利用這些模式來描述資料中的特徵以及關係。從龐大的資料庫中選取合適的資料，進行資料的處理、轉換等工作，再進行資料探勘與結果評估的一系列過程。

有許多學者曾對資料探勘做過定義，微軟的 Data Mining 小組 Fayyad 等人(1996) 對 Data mining 的定義是：它是知識發掘的一個過程，應用某些電腦計算技術，在可接受的運算效率限制下，找出建立於資料之特性樣式。因此 Data mining 是利用自動或半自動的方式分析大量的資料，以找出有意義的資料關係或法則（湯玲郎、林信忠，2000）。

一般而言，資料探勘(Data Mining)可以解釋爲資料庫之知識發覺(Knowledge Discovery in Databases，簡稱 KDD)。我們可以從一個大型資料庫中所儲存的大量資料從中萃取出一些有用的知識。

（一）資料探勘目的

1. 瞭解資料的特徵與關係可以提供作決策所需要的資訊。
2. 資料的特徵可幫助做預測。

（二）資料探勘過程

1. 選取（設定目標）

分析現有的模型，以確認資料探勘應用的領域，進一步設定此模型的目標及評估準則，並逐一考量影響此模型潛藏的因素。這在企劃過程中有絕對的意義，亦關係著資料探勘是否成功的關鍵。

2. 資料前處理

資料前處理主要包含資料整合(data integration)、資料清理(data cleaning)以及資料轉換(data transformation)等三項工作。

(1) 資料整合:最主要的目的便是解

決多資料來源的整合問題。

(2) 資料清理:主要目的是確認資料的正確性以及完整性。

(3) 資料轉換:主要目的是將資料內容轉換成更容易探勘或是探勘結果可信度更高的狀態

資料探勘前必須來確定資料是否完整一致並確認資料無誤，因爲資料本身如果是有缺陷，當錯誤的資料在進行資料探勘時勢必會影響到探勘的結果，所以當資料越龐大時，資料除錯的工作就會越顯得重要!

3. 資料探勘工作

此階段爲資料探勘真正核心。藉由資料探勘演算法找出隱藏在資料背後的規則、特性、型樣。

4. 結果分析

解釋並評估前階段所產生的結果。一般而言，會將結果以圖表的方式呈現出來，讓使用者對於分析結果能深刻瞭解。(曾詠淑，1999)

整個資料探勘的過程中，可以發現第一及第二階都是對於資料的事前準備工作，直至第三階段才正式進入資料探勘的主軸；由此可知，在從事探勘工作前需做足許多的準備工作。資料探勘過程中花費在準備工作以及規劃過程所佔的時間相當多；在實際執行上，大部份的時間以及精力是花費在前置作業上。

（三）資料探勘功能

1. 分類

按照分析對象的屬性分門別類加以定義，建立類組。

2. 推估

根據既有的連續性數值之相關屬性資料，已獲得某一屬性未知之值。

3. 預測

根據對象屬性之過去觀察值來推估該屬性未來之值。

4. 關聯分組

從所有物件決定哪些相關物件應該放在一起。

5. 同質分組

將異質母體中區隔爲較具同值性之群組。

參、PolyAnalyst 軟體功能簡介

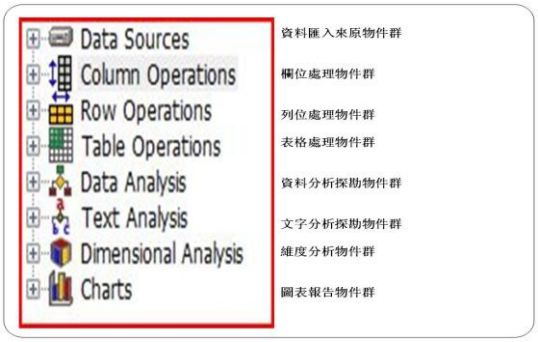
PolyAnalyst 是使研究變得容易的一個多目的的數據採集系統，以及決定支援。用戶能在交互式分析過程中發現隱藏的規則。PolyAnalyst 可以協助使用者充分地利用現有的資料庫，透過 PolyAnalyst 工具，能讓這大量的資料作更有效率的運用，淬煉出有用資訊和獲得其潛在的情報及知識，如此才能提供更充足的決策情報。

由於 PolyAnalyst 本身提供進 17 種演算法，但是如果以解決問題來說的話，使用者不一定會使用到全部的工具，所以在工具的選擇上，我們會依照目的不同而使用不同的演算法，以我們在操作 PolyAnalyst 這套軟體的操作步驟大致上可以分成以下幾個步驟

一、匯入資料

由於我們的研究資料是所提供的資料格式是.data 檔，你可以利用微軟的 Excel 來打開資料，或是一般的文字編輯器也都可以，只是以操作方便上來說，通常我們把.data 的檔案先儲存爲純文字檔，因爲資料的本身並沒有顯示屬性，都是數值資料的呈現，屬性的說明是額外放在說明檔中，所以我們可以在 Excel 那邊就先把屬性和資料先整合完成，在利用 poly 匯入已經有屬性和數值資料的檔案，上述這個動作在 PolyAnalyst 的工具中，也可以使用欄位處理中的 Modify columns 來做欄位掩蓋、次

序變更、重新命名、變更資料型式，但必須注意到的是當使用者在做欄位的重新命名，Poly 會把最上面的欄位視爲你要修改的欄位，所以當匯入的資料如果本身沒有屬性的話，必須先預留一列空白欄位來給 Modify columns 做修改的動作，否則資料可能會發生減少一筆的情況!這邊也是資料前處理必須先準備的階段，確保資料是否完整一致並確認資料無誤，藉由上述資料前處理的步驟來逐步確認資料的正確性!



上圖是 PolyAnalyst 所提供的工具

二、 尋找資料特徵和衍生欄位

當資料已經匯入 poly 後，可以先藉由一般統計的表單，尋找是否有明顯的特徵值，或是將資料做排序的動作，藉由 Poly 的 Row Operation 利用 Sort Row 對資料中表格欄位各列進行排列順序，會更容易的觀察資料， 經由初步的分析影響問題的可能因素，但通常會發現原始資料並不足以我們解決問題,所以可能要借助於 PolyAnalyst 的 Derive 來做衍生欄位的動作，將原始資料經由運算後衍生出所需的欄位，期望能夠藉由充足來資料來做後續的資料探勘動作! PolyAnalyst 也提供許多圖形化的功能，例如:之前我們在做汽車耗油量分析的時候，利用 poly 裡面的 plot-散佈圖來觀察影響汽車耗油量的重要因素。

三、 資料探勘/結果分析

主要是利用 PolyAnalyst 裡面的工具來幫助我們做資料探勘，例如我們可以利用 Poly 的 Clutering 來做叢集分析，我們可以設定要劃分的群數，當執行 Clutering 之後，系統會自動開始作群集，其結果將列出 Setting 值，以及 Clustering 結果，分群之後，系統會將資料個別摘要出屬性平均值。或是利用 Poly 裡面的決策樹，藉由決策樹 (Decision Tree)來產生規則，從決策樹的分類技術來分析資料，由於決策樹能夠產生易於了解的規則。藉由樹的節點(node)與分支(branches)可以簡單呈現資料分類之規則，圖表的呈現方式，讓使用者對於分析結果能深刻瞭解。

肆、研究目的

在產業中，企業將資料探勘之技術應用在顧客關係管理及產品行銷已有許多先例，透過大量收集顧客資料，根據客戶的屬性(性別、年齡、職業、消費行爲等)加以分群，經過分群後之結果及呈現出具同樣特性及特徵之消費者族群，因此該結果也可提供企業做爲日後區隔行銷目標之決策參考。因此本研究藉由資料探勘工具 Poly Analysis 6.0 並使用學生資料庫來執行資料前處理的部份和使用心臟病資料集分析心臟病特徵與年齡層分佈及心臟病患者病因的相關程度。並且利用訓練組和對照組來實地演練資料探勘的流程和熟悉這套新的 Poly 資料探勘軟體。

此研究利用學校學生資料庫的資料前處理和分析探勘心臟病個案之病患病因屬性分佈情形，在學術研究上，本研究參考過去一貫使用之統計分析方法，利用統計方法找出顯著的屬性，更利用訓練組與對照組的方式來驗證探勘軟體之效度，並可從探勘結果中發現統計分析無法呈現之結果，另外，此研究主要希望能提供使用軟體的之後，從過程中熟悉每個功能和參數，最後所研究的結果能夠提供後來使用 Poly 這套軟體或學習資料探勘的人能夠

有幫助。

歸納本研究的目的如下

一、 學習熟悉資料探勘的整個流程。

二、 探討軟體中部份功能的功用。

三、 探討軟體中改變參數後影響分析結果時的情況。

四、 利用使用後的結果來幫助之後學習資料探勘或使用這套軟體的人。

本研究除了利用目前業界炙手可熱的資料探勘工具 PolyAnalysis、統計分析方法、關聯分析，更蒐集了各種相關資訊。企圖瞭解資料探勘整個的流程和 Poly 軟體中的給個功能和參數的功用，期望可提供學習資料探勘和使用 Poly 這套軟體的人一些參考依據。

伍、研究方法

將各種的屬性資料作除錯和整理之後，先用一般的統計分析，找出顯著意義的目標屬性，以便日後的資料探勘使用，然後藉由資料探勘工作來推估出資料中隱含的規則來，並藉由探勘出來的結果來進行規則推算和屬性分析，之後再配和一般統計分析的結果，希望能由這兩種的分析的結果來獲得最好的探勘模式。

首先將原始的網路資料檔下載，利用.csv 檔進行資料的匯入，然後藉由資料探勘軟體 PolyAanalyst 來進行資料前置處理，刪除不必要的欄位和延伸新的欄位，此外，各欄位的資料型態有些不是正確的資料形態，必須利用 Modify Columns 來更變欄位資料的型式。資料經過一般統計分析找出顯著的屬性，目的在作資料前處理以利日後的資料探勘分析，資料探勘分析這個階段牽涉到繁雜的過程，利用複雜的演算法分析大量的資料，以找出隱含於資料中的規則和模式。首先選擇資料，我們選擇具有代表性的屬性，然後將人爲錯誤、電腦系統錯誤、欄位遺漏等錯誤排除，對於資料不足的部份，我們以原先既存的變項中延伸出本研究要探討的變項，以彌補資料不足的部份，並且利用編碼與資料的轉換，整理原始資料的格式和屬性可能無法用於統計軟體或資料探勘軟體的格式需求，所以需將原始資料加以從新編碼和轉換以方便軟體使用，資料探勘過程中，利用適當的演算法和對應的參數設定，並經由軟體的訓練和學習之後，探勘出一些隱藏而且有幫助的訊息。這次研究使用的軟體是 PolyAanalyst 6.0，使用的探勘技術有叢集分類、關聯分析及決策樹等…，將探勘所得到的模式特徵及關聯性和一般統計方法所得到的結果作比較分析，驗證探勘結果在統計資料中是否達到顯著意義，藉由傳統的統計分析和新興的資料探勘技術作結合，將資料輸出成有用的資訊。

一、 心臟病患者資料

心臟病患者資料收集，包含病患年齡及診斷資料，來源於 UCI Machine Learning Repository 網站提供，網址爲 http://archive.ics.uci.edu/ml/。此資料集具有 1 個目標屬性，13 個條件屬性(7 個爲類別屬性，6 個爲數值屬性)，總共 270 筆病患資料。其中利用目標屬性將病患群分爲兩類(有心臟病；無心臟病)，以下對 13 個條件屬性作敘述。

(一) 數值屬性

1. 屬性 1:病患年齡 age(記錄病患年齡資料)。
2. 屬性 4:休息時血壓 resting_blood_pressure(血壓是心臟收縮，把血輸送到全身各部時，血液在血管內流動的壓力。血壓通常以 120/80mmHg 之收縮壓／舒張壓爲正常血壓，第一個數字是收縮壓，也就是心臟收縮或擠壓時的血壓。第二個

數字是舒張壓，也就是當心臟休息時二次心跳之間的血壓)。

3. 屬性 5:血液中膽固醇含量 serum_cholestoral(當血清中膽固醇含量過多，易引起動脈硬化症、高血壓；含量太低，則可能爲營養不良)。
4. 屬性 8:最大心跳速率 maximum_heart_rate_achieved(正常人心臟的跳動是有規律的，且會隨身體狀況(活動量、情緒)而加快或減慢，只要每分鐘規則跳動 60 到 80 次(不超過 100 次)。
5. 屬性 10:運動心電圖圖 ST 下降程度 oldpeak(S-T 是指心室收縮到心室擴張的時間正常的心電圖的每個波的長度)。
6. 屬性 12:冠狀動脈被阻塞的數量 number_of_major_vessels(冠狀動脈受到了阻塞，讓冠狀動脈能夠輸送的氧氣和養分的量減少了，使得心臟不能夠補充到正常狀況下所需要的能量)。

(二) 類別屬性

1. 屬性 2:病患性別 sex(記錄病患性別 0=女性；1=男性)。
2. 屬性 3:胸痛類型 chest(記錄病患胸痛類型 1=主體脈剝離引起主動脈瘤急性發作時，所引發的疼痛；2=心包膜炎引起的胸痛，深呼吸時會使疼痛加重； 3=官能性神經症所引發的疼痛；4=心絞痛，當供應心臟氧氣的冠狀動脈變厚、變硬、變窄使心臟缺氧，導致心絞痛)。
3. 屬性 6:空腹的血糖含量是否>120 fasting_blood_sugar(正常空腹血糖介於 80~120，0=否；1=是)。
4. 屬性 7:休息時的心電圖結果 resting_electrocardiographic_results(由心電圖的變化可以診斷心律不整的及各種心臟疾病所引起的心臟型態上的變化，0=正常；1=輕微易常；2=異常)。
5. 屬性 9:運動是否引發心絞痛 exercise_induced_angina(心絞痛是心肌缺氧的臨床表徵，大部分由於膽固醇積聚而使動脈血管逐漸變小，造成冠狀動脈硬化所致。0=否；1=是)。
6. 屬性 11:運動心電圖 ST 間最高的斜率 slope(1=完全正常；2=連續兩個以上電極 ST 段上升超過 1 毫米以上；3=連續兩個以上電極出現 ST 段下降超過 1 毫米以上)。
7. 屬性 13:地中海型貧血 thal(3=正常；6=地中海貧血甲型；7=地中海型貧血乙型)。

陸、研究流程

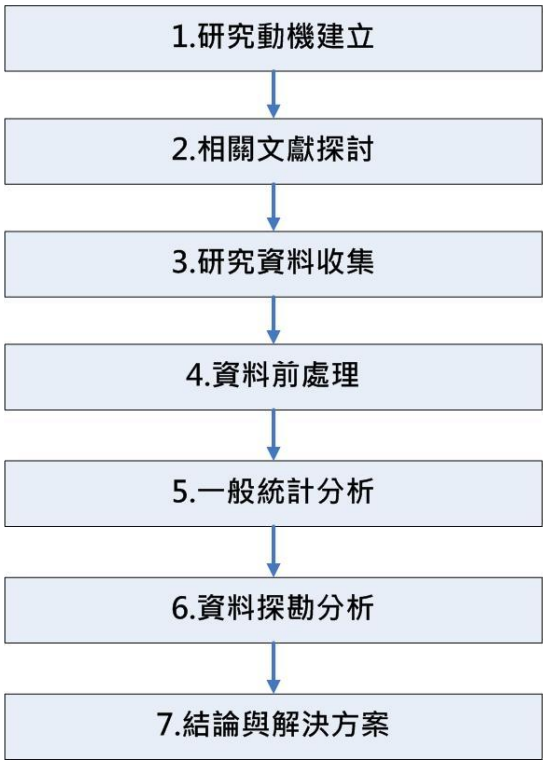
在觀察許多案例後，發現最終結果發生時，事前是有許多徵兆的，卻常被忽略，所以便從這方面來做探討與研究，剛開始先研讀其他相關案例研究，了解一些研究分析的方向，同時也學習資料探勘的技術與應用。

而相關研究研讀完後，便做資料收集的動作，我們選擇網路上的資料庫來做分析，接著我們便開始做資料探勘的前處理的步驟也就是將資料整合、資料清理和資料轉換的工作，在確認資料完整無誤之後，便可以做一般的統計分析了，先找出較顯著的變項，方便以後資料探勘的效率和運作。

最後則是資料探勘與分析的部份，利用不同的工具，將我們想要知道的問題一一明朗化，做出我們想要的分析，最後則是依據我們所做的分析來歸納和討論出解決方案，或是將研究結果呈現給醫院，希望可以給予一些幫忙，以達成本研究的最

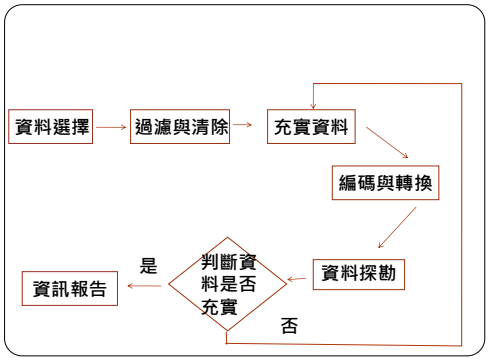
終目的。

一、 以下則是此研究的流程圖



二、 資料探勘分析

在這個階段進入較複雜的資料分析過程，其中運用到大量的分析方式，例如線性分析、叢及分析及決策樹等，然而，爲了要充分了解資料探勘的功用之前，有一些標準步驟是要遵循的，以下就是本研究的步驟



柒、結果呈現

練習分析各種案例後，漸漸了解分析的步驟與方法，以心臟病資料庫爲例：每種症狀數值屬性在心臟病中都是一種指標，也許某些屬性高於某個數值之後就會成爲心臟病高危險群，但是也許某些屬性徵狀都同時擁有的時候才會發生心臟病，找出這些關聯的屬性就可以幫助醫師在診斷病人的時候可以更準確的診斷出病患是否有沒有患有心臟病，並且根據數據來建議病患應注意飲食還是生活習慣多做注意，如此就可以達成提早預防的目的。

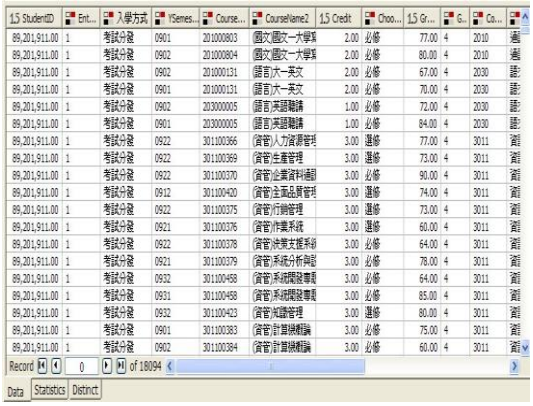
在作統計分析和資料探勘之後所演算出來的規則和模式，我們將由這些推演預測出來的訊息當作資訊，再由這些資訊中找出一些能當做心臟病診斷時候有參考價值的資訊，預期心臟病中某些屬性看似重要其實並不是重點屬性，而某些看似不是重要的資訊才是真正重要的診斷決策依據，讓醫師在作診斷時能更有效並且準確的診斷出病人是否患有心臟病。

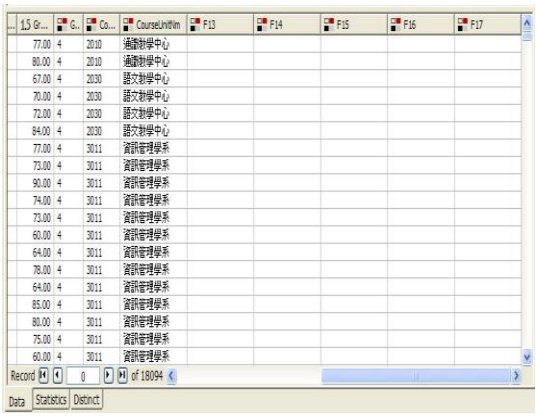
學生資料庫則是注重在資料前處理的部份，而我們的主要目的在於熟悉 poly 這套軟體並且在使用和學習這套軟體時，改變其中某些重要的參數能達到讓分析出來的結果能夠更明顯而且準確，以下則是我們的結果。

一、 資料前處理結果

資料前處理的部分，我們以本系畢業生的成績，做爲我們的資料來源。主要的目爲：資料整合、資料清理以及資料轉換等三項工作。下圖是匯入資料時，未經過前處理，各別欄位資料型態的呈現：

未經前處理的資料圖



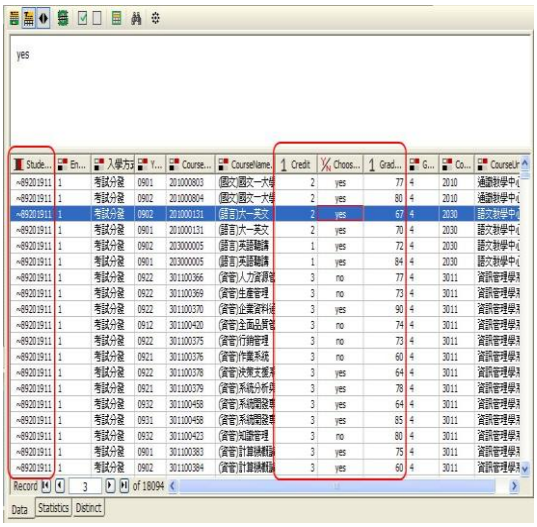


我們利用 PolyAnalyst 裡面的節點工具：Modify Columns，將資料的型態作更爲正確的呈現，並且勾除掉後面的空白資料欄位。

StudentID (學號) 改成 ID 字串型態 Credit (學分) 改成整數資料型態

ChooseNm (主修/選修) 改成布林資料型態 GradeResults (成績結果) 改成整數資料型態

執行前處理的結果

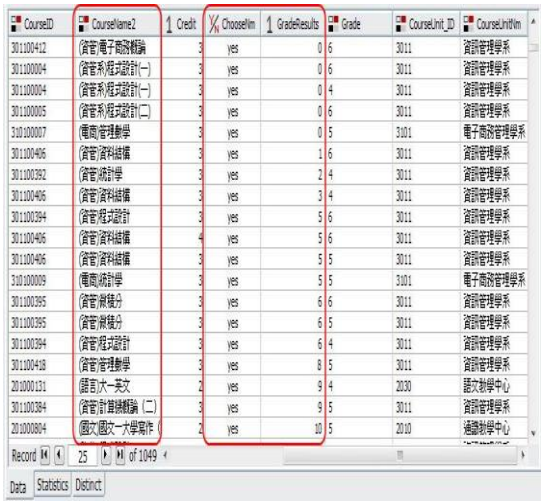


可以看到圈起來的欄位資料型態，經由 Modify Columns 的執行後，資料型態更爲正確，且更接近原始資料的型態，有利於爾後的分析工作。

當確認資料是正確無誤並且沒有缺陷的時候，接下來可以利用 PolyAnalyst 裡的節點工具：Filter Rows，將我們想要資料挑出來，也可以利用此方法過濾掉不正確的資料。

例如：當學生成績是不及格的情況下，主要是哪些主修的課程。

前處理的結果



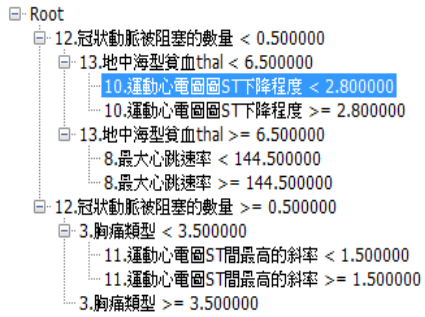
此圖是藉由 Filter Rows 和 Sort Rows 兩個節點而形成的結果。

Filter Rows 主要是挑出不及格(<=59 的資料)的資料和主修科目(yes 爲主修/no 爲選修)，Sort Rows 則是將成績欄位和課程名稱做排序(成績結果會由小排到大，課程名稱相同的會排列一起)，藉由這樣的呈現，我們可以看出不及格的資料，是屬於什麼主修科目。

二、 心臟病資料不分群結果

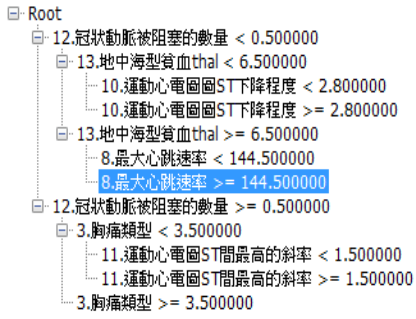
心臟病資料不分群的部份主要是用心臟病資料中分割的百分之七十的訓練組資料，主要過程有：更改型態、蛇行圖、決策樹、驗證等步驟，主要目的是用來對照分群之後所產生之決策樹的驗證結果，以期可以證實分群後之決策樹的驗證錯誤率是比未分群的驗證結果更好。

不分群資料決策樹模型



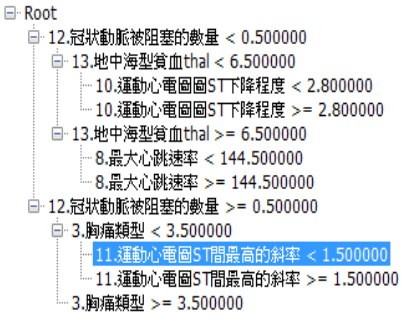
Decision	no
Classification errors	7.69231%
p-value	9.18565e-013
log(p-value)	-27.716
Number of records	78 (41.2698%)
no	72 (92.3077%) [1.66154]
yes	6 (7.69231%) [0.173077]

在冠狀動脈阻塞數量小於 1 並且地中海貧血類別不是 7 和運動心電圖下降程度小於 2.8 的病人只有 7.69%的機率會得到心臟病。



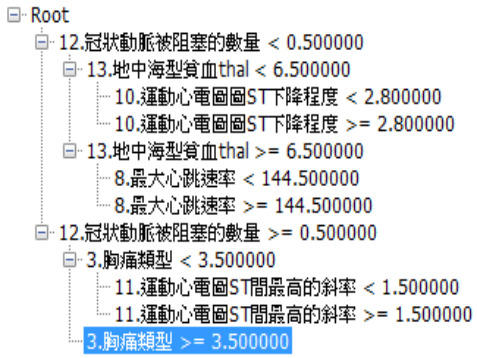
Decision	no
Classification errors	33.3333%
p-value	0.240922
log(p-value)	-1.42328
Number of records	18 (9.52381%)
no	12 (66.6667%) [1.20000]
yes	6 (33.3333%) [0.750000]

在冠狀動脈阻塞數量小於 1 並且地中海貧血類別是 7 和最大心跳速率大於等於 144.5 的病人有 33.3%的機率會得到心臟病。



Decision	no
Classification errors	11.1111%
p-value	0.00288173
log(p-value)	-5.84937
Number of records	18 (9.52381%)
no	16 (88.8889%) [1.60000]
yes	2 (11.1111%) [0.250000]

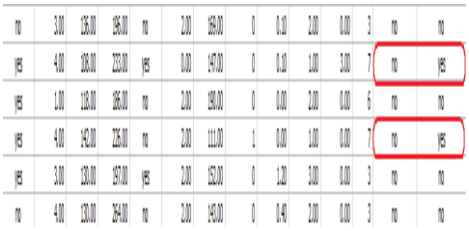
在冠狀動脈阻塞數量小於 1 並且胸痛類型類別不是第四型和運動心電圖 ST 最高斜率小於 1.5 的病人有 11.1%的機率會得到心臟病。



Decision	yes
Classification errors	4.16667%
p-value	0
log(p-value)	-31.4161
Number of records	48 (25.3968%)
no	2 (4.16667%) [0.07500...
yes	46 (95.8333%) [2.15625]

在冠狀動脈阻塞數量大於 1 並且胸痛類型類別是第四型的病人有 95.8%的機率會得到心臟病。

決策樹之驗證結果



紅線圈起來的地方就是驗證之結果與預期結果不符合的地方。

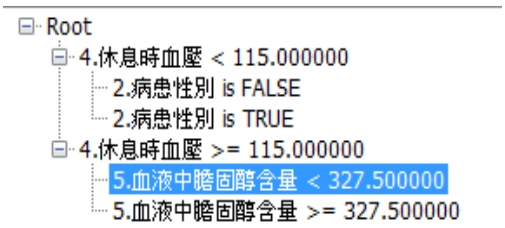
Actual/Predicted	NO	YES
NO	36(44.44%)	9(11.11%)
YES	11(13.58%)	25(30.86%)

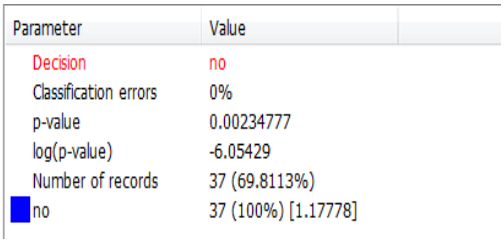
這是驗證的混亂矩陣，其中實際上爲 NO 又判別爲 NO 的有 36 筆，實際爲 NO 又判別爲 YES 的 9 筆，實際爲 YES 但判別爲 NO 的有 11 筆，實際爲 YES 且判別也是 YES 的有 25 筆，整個驗證組的資料筆數有 81 筆判別錯誤的有 20 筆，這個決策樹的正確率大約在 75.3%。

三、心臟病資料分群探勘結果

心臟病資料經由叢集分析後分別建置叢集一及叢集二決策模型，並且在建置決策模型的過程中將原布林型態資料轉換爲數值型態資料，以統一整體分析資料。

(一) 叢集一決策樹模型

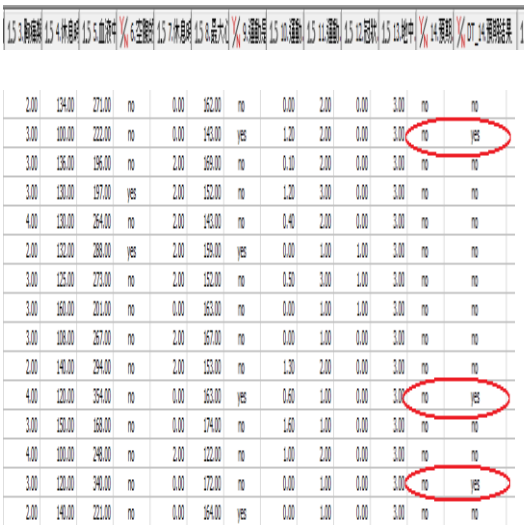




Parameter	Value
Decision	no
Classification errors	0%
p-value	0.00234777
log(p-value)	-6.05429
Number of records	37 (69.8113%)
no	37 (100%) [1.17778]

叢集一決策模型病患群大部分落在休息時血壓>=115 且血液中膽固醇含量<327.5 患者顯示無心臟病(其餘枝葉因資料筆數過少因此有待檢驗)。

(二) 叢集一決策模型驗證

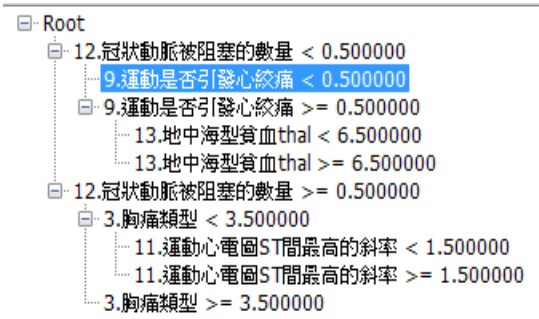


(三) 叢集一決策模型混亂矩陣

Actual/Predicted	NO	YES
NO	30(76.92%)	4(10.25%)
YES	5(12.82%)	0(0.00%)

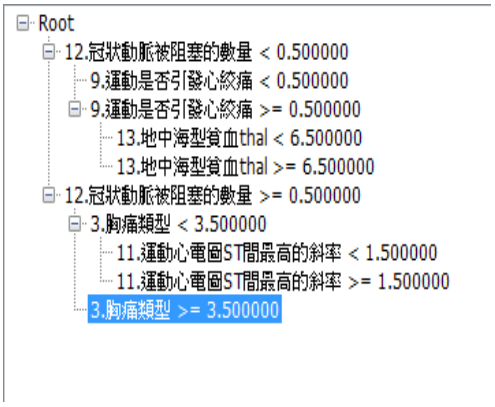
叢集一決策模型經由測試組資料驗證過後，無心臟病患者正確判斷爲無心臟病共有 30 筆；無心臟病患者錯誤判斷爲有心臟病共有 4 筆，而有心臟病患者錯誤判斷爲無心臟病共有 5 筆；有心臟病患者正確判斷爲有心臟病共 0 筆，整體決策模型一錯誤判斷爲 9 筆，決策模型一正確率爲 76.92%。

(四) 叢集二決策樹模型



Parameter	Value
Decision	no
Classification errors	16.3265%
p-value	1.35036e-008
log(p-value)	-18.1203
Number of records	49 (36.0294%)
no	41 (83.6735%) [1.89660]
yes	8 (16.3265%) [0.292159]

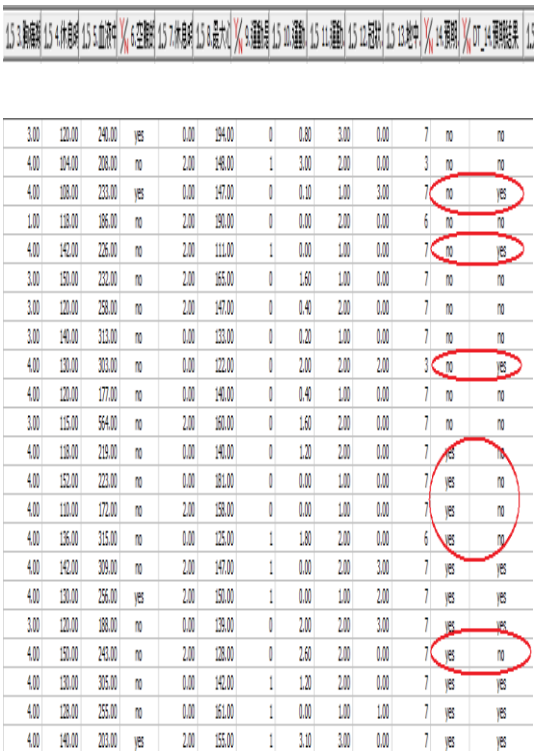
叢集二病患群如冠狀動脈無阻塞且運動無引發心絞痛患者大部分爲無心臟病。



Parameter	Value
Decision	yes
Classification errors	2.38095%
p-value	8.29944e-010
log(p-value)	-20.9097
Number of records	42 (30.8824%)
no	1 (2.38095%) [0.05396...
yes	41 (97.619%) [1.74687]

叢集二病患群如冠狀動脈有阻塞且胸痛類型爲心絞痛患者，大部分患有心臟病(其餘枝葉因資料筆數過少因此有待檢驗)。

(五) 叢集二決策模型驗證



(六) 叢集二決策模型混亂矩陣

Actual/Predicted	NO	YES
NO	8(19.04%)	3(7.14%)
YES	7(16.66%)	24(57.14%)

叢集二決策模型經由測試組資料驗證過後，無心臟病患者正確判斷爲無心臟病共有 8 筆；無心臟病患者錯誤判斷爲有心臟病共有 3 筆，而有心臟病患者錯誤判斷爲無心臟病共有 7 筆；有心臟病患者正確判斷爲有心臟病共 24 筆，整體決策模型一錯誤判斷爲 10 筆，決策模型一正確率爲 76.19%。

四、 結論

心臟病資料探勘主要嘗試有經過分群的資料與未經過分群的資料所建置出決策樹的差異，其中未經過分群的心臟病資料決策模型正確率爲 75.3%；經過分群的兩個叢集各別建置的決策模型正確率分別爲 76.92%及 76.19%。

雖然本研究經由分群所建置的決策模型與未分群所建置的決策模型正確率差異不大，但是經由研究過程的嘗試及與專家研討的結果，如果資料筆數非常大且複雜的話，有經過分群所建置的決策模型正確率較高且與未分群所建置的決策模型相較下也較精準，因此本研究亦驗證了此理論的正確性。

捌、使用心得

對於一套完全沒有使用過的軟體，從學校一開始安排的教育訓練，到我們熟悉整套軟體，確實花費不少功夫，一開始是資料難尋，嘗試過各種不一樣的資料庫做練習，汽車性能分析、病床使用率、乳癌、心臟病，還有學生資料庫，在試過這麼多不同的資料之後，我們發現，資料的收集有著一定的困難度，以病床使用率來說，我們也是因爲資料不完整所以停止分析的動作，而資料的前處理當然也幫韓這個部份，接下來當有資料了以後，要將資料轉成怎樣的檔案匯入較利於資料的完整呈現，也有不一樣的匯入方式，另外如何將不完整的原始資料作整理，也是一大學問，將不完整的資料刪除、型態的修改、資料的排序和資料的補充，如何將現有的資料做延伸以利於分析，這些都是要靠經過不斷的測試和累積的經驗依照不同的步驟做整理，雖然只是資料前處理，但其實卻是整個資料分析最重要的部份。

前處理完成後，接下來也就是資料分析的部份，如何將一堆的資料轉換成有用的資訊，也就是這個軟體的最大目的與功用，一開始利用不同的資料下去做功能熟析的動作，關聯分析、購物籃分析、線性迴歸分析、到最後因爲熟析個個功能節點，同時也慢慢知道資料分析的步驟和過程，在最後心臟病的分析裡面首先將資料分爲訓練組和測試組，接著利用蛇形圖找出屬性重要性，再投入分群，便同時和不分群的資料做比較，找出最佳決策樹，經過一次一次反覆的測試，找出最佳決策樹，在由測試組下去做驗證，將結果做一個呈現。

這次使用軟體最主要的收穫就是瞭解資料分析的過程和步驟，而不是向無頭蒼蠅一樣亂跳，因爲了解程序，對於使用 Pply 這套軟體更能方便使用，將有幫助的功能，快速的分析出我們想要的東西，重點其實是在整個流程，軟體只是輔助我們的一個工具，先知道如何規劃再來使用，才是最正確的步驟。

參考文獻

【1】周歆凱，利用「資料探勘技術」探討急診高資源耗用者之特性，國立台灣大學醫療機構管理研究所碩士論文，2004。

【2】柯睿明 (Rami Karjian) ，以製造業爲師改善台灣醫院效率，商業週刊第 972 期(Business Weekly)，2006 年 7 月 10 日。

【3】李御璽，「資料探勘在心臟並預測模型上之研究」，碩士論文，銘傳大學資訊管理研究所，2007。

【4】曾憲雄、蔡秀滿、蘇東興、曾秋蓉、王慶堯著，「資料探勘」旗標書局出版股份有限公司。

【5】心臟病資料來源
http://tw.babelfish.yahoo.com/translate_url?doit=done&tt=url&trurl=http%3A%2F%2Farchive.ics.uci.edu%2Fml%2F&lp=en_zt&.intl=tw&fr=yfp

附 件

實作成果完整投影片

Poly Analyst 6.0 使用手冊