

南 華 大 學

資訊管理學系

碩士論文

資料採礦在學生學業成績表現上之應用  
A Study on Student Course Performance using Data  
Mining Techniques



研究生：劉妃娜

指導教授：邱宏彬

中華民國 101 年 07 月

# 南 華 大 學

資訊管理研究所

碩 士 學 位 論 文

資料採礦在學生學業成績表現上之應用

A Study on Student Course Performance using Data Mining

Techniques

研究生：劉妃娜

經考試合格特此證明

口試委員：陳宗榮

李朋諳

邱宏林

指導教授：邱宏林

系主任(所長)：



口試日期：中華民國 101 年 6 月 30 日

# 資料採礦在學生學業成績表現上之應用

學生：劉妃娜

指導教授：邱宏彬

南 華 大 學 資 訊 管 理 學 系 碩 士 班

## 摘 要

據資料顯示，國內大專校院數量近年逐年不斷增加，學生能選擇的大專科目同時也隨之增加。學生在選擇其就讀科系或選擇研習課程時，所選擇的科目也不一定是選擇適合學生本身。本研究以某私立大學資訊管理系日間部之學生成績資料為例，運用 SQL Server 2008 所提供之資料採礦工具，分析學生成績資料之變化，針對結果提供相關建議，並設法改善學生學習狀況，以期能夠提高學生學習興趣，將阻力化為助力。

**關鍵詞：**資料採礦、類神經網路、學生學業表現

# A Study on Student Course Performance using Data Mining Techniques

Student : Fei Na, Liu

Advisor : Dr. Hung Pin, Chiu.

Department of Information Management  
The Graduate Program  
Nan-Hua University

## ABSTRACT

In view of the increasing number of college courses, students might not select the courses that suit them best. In order to provide suggestions for students when they encounter difficulty of choosing their courses, this study utilizes the data mining tool of SQL Server 2008 to analyze the data collected from students at a department of information management in certain university. Based on the suggestions, the students are expected to improve their course performance.

Key words: Data Mining, Neural Network, Student Course Performance

# 目 錄

論文口試合格證明 .....	ii
中文摘要 .....	iii
英文摘要 .....	iv
目錄 .....	v
圖目錄 .....	vii
表目錄 .....	viii
第一章 緒論 .....	1
第一節 研究背景與動機 .....	1
第二節 研究目的 .....	3
第三節 研究範圍與限制 .....	3
第四節 研究步驟 .....	3
第五節 論文架構 .....	5
第二章 文獻探討 .....	7
第一節 資料採礦定義 .....	7
第二節 資料採礦運用 .....	9
第三節 類神經網路 .....	10
第四節 資料前置處理 .....	12
第五節 相關研究之回顧與探討 .....	15
第三章 研究方法 .....	17
第一節 研究對象 .....	17
第二節 Microsoft SQL Server 資料採礦工具 .....	17
第三節 資料處理方式 .....	17
第四章 類神經網路分析 .....	19

第一節 模型之建立 .....	19
第二節 類神經網路分析 .....	22
第五章 結論與建議 .....	41
第一節 研究結論 .....	41
第二節 研究建議 .....	42
參考文獻 .....	43

# 圖 目 錄

圖 1- 1：75 學年度至 99 年度大學校院數變動表.....	2
圖 1- 2：研究流程圖.....	6
圖 2- 1：Knowledge Discovery in Databases, KDD 流程圖(黃南傑，2004).....	8
圖 4- 1：成績變數分析圖.....	20
圖 4- 2：成績變數分析圖.....	21
圖 4- 3：系統與分析類成績散佈圖.....	22
圖 4- 4：管理與決策類成績散佈圖.....	23
圖 4- 5：類神經網路檢視器-輸出項為系統與服務類.....	23
圖 4- 6：類神經網路檢視器-輸出項為系統與服務類.....	24
圖 4- 7：變數視窗-程式類過濾條件為 36.400-59.857，輸出屬性為系統與服務類....	25
圖 4- 8：變數視窗-程式類過濾條件為 36.400-59.857，輸出屬性為管理與決策類....	26
圖 4- 9：變數視窗-程式類過濾條件為 74.625-92.200，輸出屬性為系統與服務類....	27
圖 4- 10：變數視窗-程式類過濾條件為 74.625-92.200，輸出屬性為管理與決策類..	28
圖 4- 11：變數視窗-管理類過濾條件為 46.316-68.916，輸出屬性為系統與服務類..	29
圖 4- 12：變數視窗-管理類過濾條件為 46.316-68.916，輸出屬性為管理與決策類..	30
圖 4- 13：變數視窗-管理類過濾條件為 82.097-98.000，輸出屬性為系統與服務類..	31
圖 4- 14：變數視窗-管理類過濾條件為 82.097-98.000，輸出屬性為管理與決策類..	32
圖 4- 15：變數視窗-語文類過濾條件為 52.476-68.769，輸出屬性為系統與服務類..	33
圖 4- 16：變數視窗-語文類過濾條件為 52.476-68.769，輸出屬性為管理與決策類..	34
圖 4- 17：變數視窗-語文類過濾條件為 78.220-89.167，輸出屬性為系統與服務類..	35
圖 4- 18：變數視窗-語文類過濾條件為 78.220-89.167，輸出屬性為管理與決策類..	36
圖 4- 19：變數視窗-數學量化類過濾條件為 25.561-56.450，輸出屬性為系統與服務類.....	37
圖 4- 20：變數視窗-數學量化類過濾條件為 25.561-56.450，輸出屬性為管理與決策類.....	38
圖 4- 21：變數視窗-數學量化類過濾條件為 74.368-98.000，輸出屬性為系統與服務類.....	39
圖 4- 22：變數視窗-數學量化類過濾條件為 74.368-98.000，輸出屬性為管理與決策類.....	40

## 表 目 錄

表 1-1：75 學年度至 98 學年度大學校院數變動表.....	1
表 2-1：類神經網路之優缺點比較表。.....	11
表 5-1：細部分析結果，由本研究整理。.....	42



# 第一章緒論

本研究之研究背景與動機、研究目的、以及研究範圍與限制將於本章說明。第一節為研究背景與動機，說明目前資訊管理學系學生成績之現況；第二節為本研究之目的；第三節為本研究之研究範圍與限制。

## 第一節 研究背景與動機

在過去 60、70 年代時的台灣，大專院校的存在並不似現今一樣普遍。而當時的業界所需要的人才類型大多偏向勞力。而現今，由於社會工作型態的轉變以及業界所需之人才類型較之以往有很大的不同；大學校院設立也日益增多，科系方面較之以往，則更加多元化。

依據我國教育部統計，國內目前大學校院數量由民國 75 學年度的 28 所增設至 99 學年度的 148 所（表 1-1：75 學年度至 99 學年度大學校院數變動表）。而系（所）方面，為因應職場需求，也從傳統上的文、史、法、醫等系（所）陸續增加了各種新式系（所），如：資訊管理學系、應用數位媒體學系、航空太空工程學系……等。

表 1-1：75 學年度至 98 學年度大學校院數變動表

學年度	大 學				學 院				總計
	國立	市立	私立	小計	國立	市立	私立	小計	
75	9	0	7	16	6	0	6	12	28
...	...	...	...	...	...	...	...	...	...
80	13	0	8	21	14	1	14	29	50
81	13	0	8	21	14	1	14	29	50
82	13	0	8	21	14	1	15	30	51
83	15	0	8	23	16	1	18	35	58
84	16	0	8	24	17	1	18	36	60
85	16	0	8	24	19	2	22	43	67
86	20	0	18	38	19	2	19	40	78
87	21	0	18	39	20	2	23	45	84

學年度	大 學				學 院				總計
	國立	市立	私立	小計	國立	市立	私立	小計	
88	21	0	23	44	23	2	36	61	105
89	25	0	28	53	22	2	50	74	127
90	27	0	30	57	21	2	55	78	135
91	27	0	34	61	21	2	55	78	139
92	30	0	37	67	19	2	54	75	142
93	34	0	41	75	15	2	53	70	145
94	40	1	48	89	9	1	46	56	145
95	40	1	53	94	10	1	42	53	147
96	41	1	58	100	9	1	39	49	149
97	41	1	60	102	7	1	37	45	147
98	41	1	63	105	8	1	35	44	149
99	44	1	67	112	5	1	30	36	148

說明：1.92、95、96、98、99 學年度之學校數為第一學期資料。

2.此表當中所陳示之資料不包含專科學校（15 所）、軍事院校、警大及空大。

3.資料來源：教育部高教司統計處。

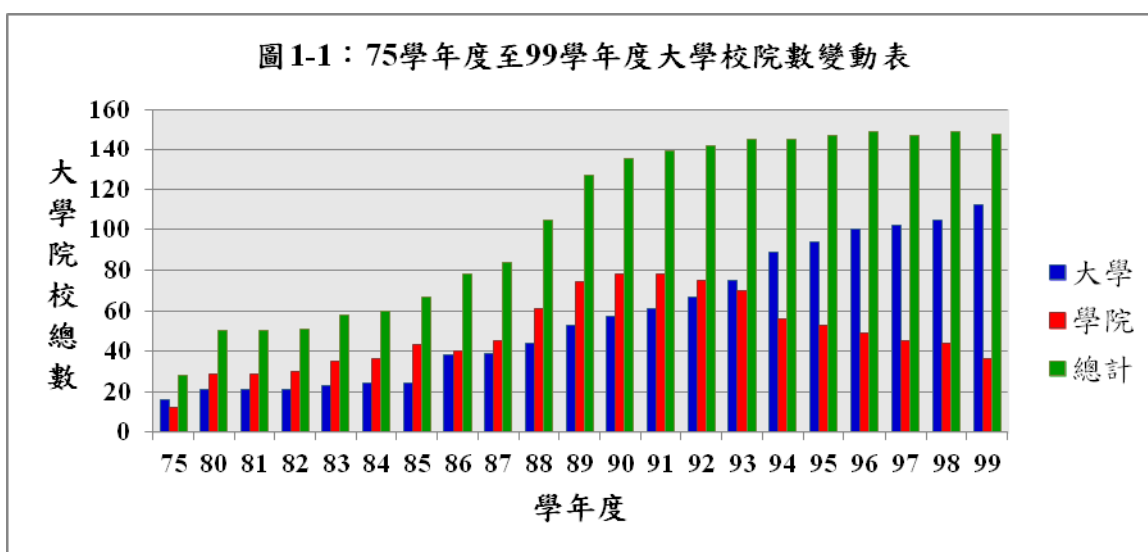


圖 1-1：75 學年度至 99 年度大學校院數變動表

隨著大學校院逐年的增設，現今國人教育程度普遍提高。而由於科技技術的進步以及資訊的快速傳遞，業界所需人力從以往的高人力演變成為高科技人才。然而，隨著時空背景的轉變，學校方面與業界之間所存在已久的斷層問題則日趨嚴重。

如何讓學生在他們在校就學期間，除了了解現今業界型態以及所需之人力類型之外，從大學中眾多課程裡，選擇適合自己的特質以及適合發展的課程研習，且將所選擇出來之課程獲得所需之知識，累積其就業競爭力，已成為各學系的眾多重要課題之一。

## 第二節 研究目的

由於現今國人的學歷普遍逐年往上提昇，為求可滿足業界之高科技人才之需求，學校除了針對學生於在學期間之課業學習情形、生活態度等多方面情形多加留意之外，更應盡可能協助學生於在校期間多元化學習，使其可學以致用，以期滿足業界需求之人力。

本研究擬運用資料採礦之技術，分析學生成績資料之變化，並且從中發現學生在學習哪些課程方面有學習上之困難；此外，針對資料採礦所得之結果，提供相關之建議，並設法改善學習不佳的學生學習狀況，以期能夠提高學生學習興趣，將阻力化為助力。

## 第三節 研究範圍與限制

本研究之研究範圍係以某私立大學 90 學年度至 93 學年度入學之大學日間部資訊管理學系之學生為主。然而本研究之資料分析來源僅為單一學校單一科系之資料庫資料，恐難以含概至他校各系（所），但研究結果仍可供他校各系（所）作為改進學習狀況之參考。

## 第四節 研究步驟

本研究將探討學生成績變化之相關因素，以「某私立大學 90~93 學年度大學日間部資訊管理學系學生之成績資料」，做為研究使用之資料庫。以 Microsoft SQL Server 2008 分析工具，來深入探討學生成績之分佈

之重要因素，並依照下列步驟完成本研究。

研究步驟如下所示：

- 一、研究動機與目的之確認
- 二、文獻探討
- 三、研究架構及使用方法之擬定
- 四、資料採礦與分析
- 五、結果與建議

上述各步驟之說明如下：

一、研究動機與目的之確認：

確認本研究之研究動機與目的後，以確定之方向蒐集資料及進行研究與操作。

二、文獻探討：

蒐集與本研究相關之文獻並從中瞭解相關有用知識。

三、研究架構及使用方法之擬定：

針對本研究之欲研究之問題，擬定適合之研究架構以及尋找適合之研究方法。

四、資料採礦與分析：

本研究以「某私立大學 90~93 學年度大學日間部資訊管理學系學生之成績資料」，做為本研究使用之資料庫。從此資料庫的學生歷史成績資料中，分析學生成績從入學起至畢業時之變化。為求所獲得之分析結果之正確性，在進行資料分析之前，應先充份檢視並了解資料庫中之各個資料欄位之意義，並瞭解各項資料之使用特性及限制條件。

五、結果與建議：

從分析結果中，找出學生成績相對較高之課程類別，並且針對資料採礦之結果提供可行之建議學生選課方向，以期可提高學生之成績及未來就業出路。

## 第五節 論文架構

本研究共分為五章，其各章節架構之詳細如下：

第一章 緒論，說明本研究之研究動機與背景、研究目的、研究步驟、研究範圍與限制以及本研究之論文架構。

第二章 文獻探討，針對相關文獻分別探討資料採礦之定義，並且運用之資料採礦方法加以深入分析探討。

第三章 研究方法，針對研究資料來源作概略簡介，其中包含研究對象、研究架構、資料處理方法及資料採礦工作之介紹。

第四章 資料分析，針對研究資料先做基本分析，並且對於所需要更進一步研究之部分加以整理並進行資料採礦分析，分析後獲得之結果進行比較。

第五章 結論與建議，針對第四章資料分析後所獲得之結果進行統整，並提出相關結論以及相關建議。

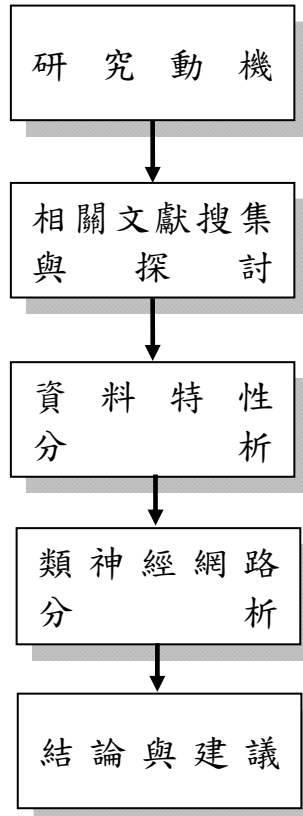


圖 1-2：研究流程圖

## 第二章文獻探討

### 第一節 資料採礦定義

所謂「資料採礦」的定義，可由兩種角度去定義：一、學者定義。  
二、廣義定義。

#### 一、學者定義：

Frawley：認為所謂的資料採礦就是指從資料庫中挖掘出潛在、明確、且非常有用資訊的過程。

Grupe & Owrang：認為資料採礦是指從已經存在的資料庫當中挖掘出專家仍未知的新事實。

Fayyad：定義知識發掘（knowledge discovery）為從大量資料中選取合商的資料，進行資料處理、轉換等工作，再進行資料採礦與結果評估的一系列過程，也就是說資料採礦只是知識發掘過程當中的一個步驟。

Berry & Linoff：定義資料採礦為使用自動或半自動的方法，對大量資料作分析，找出有意義的關係或法則。

#### 二、廣義定義：

可解釋為資料庫之知識發掘（Knowledge Discovery in Databases, 簡稱 KDD）。也就是說可以從一個大型資料庫裡頭所儲存的大量資料當中萃取出有趣知識，這個大型資料庫有可能是線上作業的資料庫（on-line database），也有可能是資料倉儲（data warehouse）。

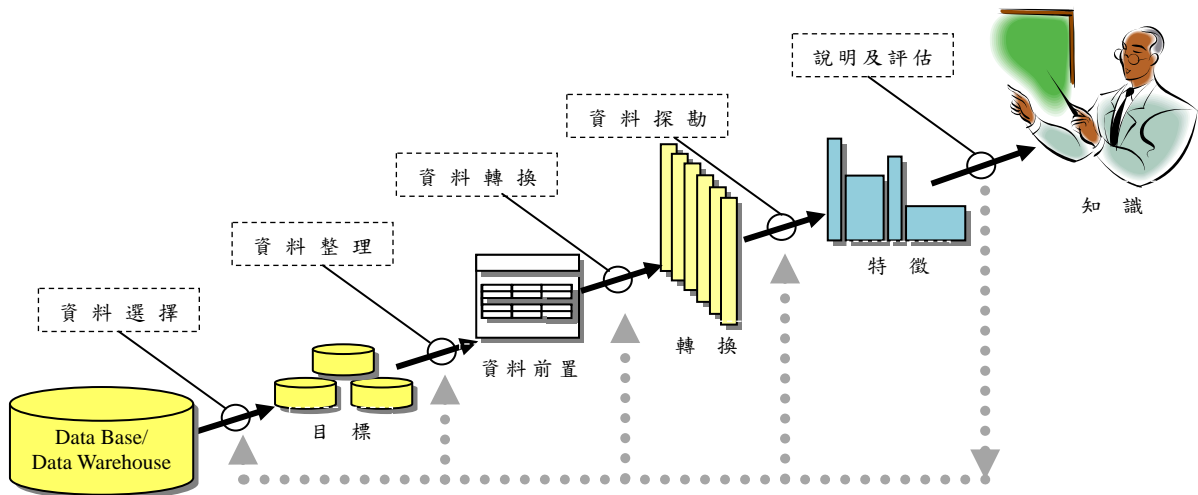


圖 2-1：Knowledge Discovery in Databases, KDD 流程圖(黃南傑，2004)

一般來說，大致上常用的資料採礦技術主要有下述六類：一、分類 (Classification)。二、推估 (Estimation)。三、群集化 (Cluster)。四、關聯法則 (Association Rule)。五、序列 (Sequential)。六、描述 (Description)。

一、分類 (Classification)：預測類別變數的過程，我們稱之為「分類」。即依據我們已知的資料及其分類的屬性，建立出資料的分類模型；接著，利用此分類模型來預測新資料的類別。例如：顧客的購物習性分類模型……等。

二、推估 (Estimation)：如何運用處理過後連續性數值的結果，給定一些輸入資料以推估未知的連續性變數的值。例如：金融商品價格之預測……等。

三、群集化 (Cluster)：群集意為物以類聚。即依資料本身的自我相似性 (self-similarity) 而群集在一起。群集 (Clusters) 的意義則要經由事後之闡釋方能得知。例如：文件分類、產品製程瑕疵分布。

四、關聯法則 (Association Rule)：指的是從歷史資料中，找出哪些事件總是相伴發生。例如：產品自動化推薦……等。



五、序列 (Sequential): 即在同質分組中透過序列來找出事物「先後」發生的順序，有時，這樣的規則被稱為時序規則 (Sequential Pattern)。例如：網頁瀏覽序列分析……等。

六、描述 (Description): 指的是在資料採礦的過程中，除了分析的預測模型以外，更重要的是在分析與處理資料的過程中，透過資料視覺化以及觀察來找出許多有意義的規則。

## 第二節 資料採礦運用

由於現今資訊傳遞快速，資料量的產生與累積的速度更是驚人。資料採礦之技術已被廣泛運用在許多行業之中。然而，各行各業依其工作特質的不同，用來支援之資料採礦的技術亦有所不同。例如：金融業、教育業、電信業、醫療業、製造業、保險業……等所採之資料採礦運用方式皆不盡相同。

一、金融業：利用資料採礦之技術替客戶做信用評等、資產管理、股匯市行情預測……等服務。

二、教育業：運用資料採礦之技術來分析學生來源、替學生做課程規劃、未來生涯規劃……等。

三、電信業：電信業者已採用或嘗試應用資料採礦之技術，以期可有效地支援目標行銷、顧客流失預測、新產品或是服務上市、或進行詐欺之偵測。

四、醫療業：運用資料採礦之技術來做醫學分析、臨床病狀分析。此外，這項技術也被運用到基因方面之分析。

五、製造業：用來做銷售分析、庫存管理或是庫存分析上之運用。

六、保險業：就保險業者而言，可運用資料採礦之技術來分析保戶要求理賠之模式，並可加強稽核以防止詐領之情形。(許依宸，2009)

隨著進入電子商務與知識經濟時代的來臨，除了競爭者與日俱增之外，客戶同時也變得更具有判斷力。企業為求可提供更佳的服務，以期爭取更多的客戶或是企業整體營業收入，業者經營方向也改變成為以顧客為中心之經營方式。因此，資料採礦之運用，除了可協助企業能夠掌握更多更精準之客戶需求及其偏好之外，同時更可制定因應措施來應用。

### 第三節 類神經網路

在早期，如果有人認為可以讓電腦學會思考，一定會被有所質疑，覺得那人只是在異想天開，認為此情景只能是在電影情節中才會出現的橋段。

然而，這對於科學家來說，並非是不可能發生的事。自從 IBM 在 1997 年推出的超級電腦「深藍 (Deep Blue)」，打敗了當時的世界棋王卡斯帕洛夫 (Garry Kasparov) 之後，電腦運用它最大的武器，可以在瞬間算出百萬個可能性，而且永不疲倦。此後，電腦除了無法在圍棋上贏過人腦之外，幾乎是已經處於不敗的地位。在早期，研究人員為求能夠在語音及影像辨認上，可以獲得與人腦相似的功能，從 1940 年代起，就已經著手開始這方面的研究。他們仿造最簡單的神經元模式，開始建立最原始的類神經網路 (Artificail Neural Network, ANN) (謝邦昌、鄭宇庭、蘇志雄，2011)。

類神經網路的原始想法、基本構造跟神經生物學中的神經元構造相似。根據 Freeman (1992) 的定義，類神經網路是模仿生物神經網路的資訊處理系統，透過使用大量簡單連接的人工神經元來模仿生物神經網路的能力。類神經網路的優點在於不需要瞭解系統的數學模型為何，直接以神經網路取代系統的模式，一樣可以得到輸入與輸出之間的關係 (謝邦昌、鄭宇庭、蘇志雄，2011)。

類神經網路可以協助解決在特徵識別、預測、最佳化問題、結合式記憶和控制的領域 (Jain, Mao, and Mohiuddin, 1996)，類神經網路的優點包含：

一、平行處理能力：可處理大量資料。

二、容錯能力：若輸入雜訊資料，仍不會影響其運作。

三、結合式記憶 (associative memory)：其又稱為內容定址記憶 (content addressable memory)，它可以記憶曾經訓練過的輸入樣式以及對應的理想輸出值。

四、解決最佳化 (optimization)：可用於處理非演算法表示的問題，或是以演算法處理很費時者。

五、能處理多對多關聯問題：即不受傳統迴歸必須用多對一函數的限制，為多輸入和多輸出的系統。

當然，任何事皆有一體兩面，類神經網路有其優點，亦有其缺點所在，表 2-1 即為其優缺點之整理以供參考 (謝邦昌、鄭宇庭、蘇志雄，2011)：

表 2-1：類神經網路之優缺點比較表。

編號	優點	缺點
1	可建構非線性的模型，且準確度高。	因其中變數(即隱藏層)可為一層或二層，數目也可設為任意數目，且有學習速率等參數需設定，工作相當費時。
2	有良好的推廣性，對於未知的輸入亦可得到正確的輸出。	使用迭代方式更新鍵結值與閾值，計算量大，相當耗費電腦資源。

編號	優點	缺點
3	可接受不同種類的變數作為輸入，適應性強。	其解有無限多組，無法得知哪一組的解為最佳解。
4	可應用的領域相當廣泛，模型建構能力強。	訓練過程中無法得知需要多少神經元個數，太多或太少的神經元均會影響系統的準確性；因此，需以試誤之方式得到適當的神經元個數。
5	具模糊推論能力，允許輸出入變數具模糊性，歸納學習較難具備此能力。	以建立數值結構(含加權值的網路)來學習，其知識結構為隱性，缺乏解釋能力。而歸納學習以建立符號結構(如：決策樹)來學習，其知識結構為顯性，具解釋能力。

#### 第四節 資料前置處理

當一筆很大量且未經處理過之資料時，大多會有下列幾項之問題：

- 一、資料不完整 (data incomplete)：在資料中可能某些屬性值會有遺缺，或是缺少某些分析時需要用到的屬性。
- 二、資料有雜訊 (noise)：資料有錯誤或是特例 (outlier) 造成。
- 三、資料不一致 (data inconsistency)：由不同來源整合而得所產生。例如：商品在台灣是以台幣，在歐洲則是歐元計價。
- 四、有高品質的資料，才有高品質的採礦結果。

所謂的資料倉儲 (Data Warehouse) 亦即是資料前置處理。資料前置處理在整個知識發掘過程當中，是最為重要且最為耗時的一項程序。資

料前置處理的主要工作有：一、資料整合。二、資料清理。三、資料轉換。

一、資料整合：資料來源多有不同，且眾多資料中總會有重覆之處，是以透過資料整合這一步驟來解決多重資料來源的整合問題。資料整合的主要工作有：

(一)消除資料不一致：1.消除資料重複性。2.綱目不一致 (schema conflict)。

(二)消除資料重複性：1.數值重複。2.綱目重複。

二、資料清理：資料清理之目的在於確認資料之正確性以及資料之完整性，使得資料採礦之工作可以順利進行。在資料清理當中，常見的資料正確性問題約有下列幾項：

(一)屬性的有效值或有效範圍：在設定資料時，理應留意資料屬性之設定是否有錯誤。例如：性別屬性的值不是男性就是女性；生日月份應介於 1 和 12 之間。

(二)數值的唯一性：有時資料上之數值應具備其絕對唯一性，以防有混淆之情形發生。例如：代表個人身分之身分證字號或是代表顧客之顧客編號應不可有重複。

(三)參考完整性 (referential integrity)：資料上之記載應具有參考完整性之特質，例如：存在於訂單資料表中的會員編號必須同時存在於會員資料表中，這樣才具有其參考完整之特性。

(四)資料的合理性驗證：資料之填寫應具有合理性之驗證，借此功能以期資料之填寫具有合理性。例如：從會員的生日計算出該會員的年齡只有 3 歲，但該會員所填寫的學歷卻是碩士，顯然該會員所填之資料不合理。

此外，在資料清理當中，常見的資料完整性問題則約有下列幾項：

(一)是否缺少採礦所需的屬性：在設定資料時，應當留意在所設資料值當中，是否有缺少即將進行採礦之屬性值。如：想要在資料當中，採礦顧客性別與購買商品種類的關係時，卻發現資料庫中並未包含性別此一屬性。

(二)是否只包含統計整合過的資訊，而缺少詳細的單筆資料：例如：當老闆想要分析在財務報表中，每月份之營業額以了解在哪幾個月當中，營業額為最高，卻發現財務報表中僅記錄著每季之營業額記錄，而缺少每月份之營業額記錄。

其他應做的清理工作包含：

(一)遺缺填補：人工填補或自動填補。

(二)雜訊消除：雜訊使採礦結果有相當大的偏差，必須將雜訊移除或將資料做平緩化處理 (smoothing)。

三、資料轉換：資料轉換，其主要目的是將資料內容轉換成更容易採礦或是採礦結果可信度更高的狀態。資料轉換工作分為下列幾項：

(一)資料統整 (data aggregation)：資料統整，指的是將所有資料之加總、統計或是建立資料方塊 (data cube)，或是將資料做初步整理，使資料更適合採礦。

(二)資料一般化：資料一般化指的是將資料的概念階層 (concept hierarchy) 向上提升，例如：會員的居住地址用城市或是南北東西中等五區取代原始詳細地址，此為資料精簡的方法之一。

(三)建立新屬性：利用資料中舊有之屬性建立即將做採礦時所需之新屬性。

## 第五節 相關研究之回顧與探討

本節介紹幾篇運用資料採礦技術在學生學業成績表現上之應用之相關文章。

傅說道(2000)探討以資料採礦中之群集分析針對南部某高中之高三畢業生，以其入學國中基測成績與高中在校成績對於大學學測之相關性做分析。其研究顯示學生在經過高中教育洗禮後，其學習所呈現之高中在校成績前百分比與參加大學學測各科之平均成績，除國文、社會達前標外，其餘皆高於全國頂標，可見其智育素質均達優等之程度；學測總分與高中在校成績總分中呈現高度之正相關；基測英文未達滿分之學生其學測英文的成績相較之下也較低。在高中在校成績方面，社會及英文的成績相對於學測的成績較為吻合；其餘學科之關聯性均有相當程度之正相關，在培育高等人才方面及十二年國民基本教育的政策下，明星高中仍有其必要性的存在。

溫侑柯(2005)探討大學入學聯招成績與南華大學學生在學成績之相關性，使用資料採礦中之關聯法則對入學聯招成績與南華在校成績做分析。其研究發現英文入學成績好的同學，在學校各科的整體表現上也較好；而歷史入學成績好的同學幾乎所有在學科目都很差；歷史入學成績差的同學在學成績幾乎都好。

簡伊汝(2009)研究目的在於建構學生基本能力預測學習優勢之預測模式，運用不同的方法來建構其預測模式，並加以比較分析其預測能力之較適切性，將較為適切之預測模式提供資訊管理系學生在修課選擇上之參考。

李佳玲(2001)與陸炳杉(2003)以高中在校學業成績來預測學測成績，其方法是以簡單迴歸來做預測模型；彭重恩(2005)以高中學生在校之數學、生物、地球科學……等科目之學業成績做為解釋變數，分別預測學測數學和自然兩考科成績；所用之方法是利用複迴歸建立其預測模型。

任眉眉等(2005)則應用 92 和 93 學年度各校系最低錄取加權總分、各單一考科考生成績累計表和各校系之採記指考科目與加權方法，及前一學年度各校系最低錄取加權總分等資料，研究「預測指考成績」建立一落點預測之模式。

陳超(2007)探討國中基測，以及高三學生的模擬考與大學學測成績之相關性。以南投縣立旭光高中為例，利用相關分析及線性迴歸模型等統計方法進行資料分析。



## 第三章研究方法

本研究主要針對前述之研究背景與動機、研究目的以及相關文獻探討等方面，提出本研究之架構並加以說明資料處理之方式，以作為後續研究之基礎。首先，在此章第一節說明研究對象之概況；第二節提出本研究之架構；第三節說明資料處理方式。

### 第一節 研究對象

本研究以某私立大學資訊管理系 90 學年度至 93 學年度入學之大學日間部學生之在校成績作為資料來源。其中，因各學年度學生之入學方式不盡相同，且最後實際完全修畢其所有學業學分之畢業學生人數皆不一致；是故，將欲進行分析之資料中，刪除所有於修業期間，途中休、退或轉學之學生資料。亦即，將未完整在該校修讀所有畢業課程規劃學分之學生資料不予以列入分析。

### 第二節 Microsoft SQL Server 資料採礦工具

在現行實務操作中，有各式各樣的資料庫管理系統，如：Microsoft 的 Access，還有 Informix、MySQL、Microsoft SQL Server、Oracle……等。本研究採用 Microsoft SQL Server 2008 作為資料採礦工具。

所謂的 Microsoft SQL Server 是一個完整的商業智慧（Business Intelligence, BI）平臺。此項平臺提供了可用於建構典型和創新的分析應用程式所需的各種特性、工具和功能給使用者使用，並於程式中引進了大量新的資料採礦功能，讓企業可以給出這些問題和其他問題的答案。

### 第三節 資料處理方式

資料的前置處理為資料採礦中最重要之關鍵步驟，其主要目的在於

取得最高預測力及讓模型結果符合使用者需求，主要包含：資料淨化、資料轉換及資料整合等。

一、資料淨化：所謂的資料淨化主要是在確認資料的完整性及正確性。將所得資料中所用不到的資料欄位予以刪除，即為資料淨化。

二、資料轉換：為使所得資料之內容可更容易進行資料採礦，通常會將部分資料進行轉換。

三、資料整合：將所得的不同相關資料進行整合，以俾進行相關後續分析。

本研究以某私立大學之資訊管理學系為例，以 90 學年~93 學年入學學生之課程學習成績資料做為資料來源。該系之畢業生所在校之成績顯示，學生在各學年皆有不同之入學方式，並且在最後實際完成學業順利畢業之學生數也不盡相同；故在研究前先將資料做整理，刪除所有不具分析價值之價值之欄位予以刪除，其中 90 學年度~93 學年度入學後若是經由轉學考入學，或是途中休、退學之學生，亦即未完整在該校修讀所有畢業課程規劃之學生的資料較不為完整，故不將其列入有效資料之中。

## 第四章類神經網路分析

### 第一節 模型之建立

首先，先將學生成績資料稍做整理，將大一、大二的成績分別為程式類、管理類、語文類、數學量化類等四大類別；將大三、大四的成績分別為系統與服務類、管理與決策類。

同時，將此四類的平均成績做為輸入變數  $X_1, X_2, X_3, X_4$  為連續值(如圖 4-1)。

$X_1$ =語文類：學生大一國文、英文、英語聽講平均成績；

$X_2$ =數學量化類：學生微積分、統計學、管理數學平均成績；

$X_3$ =程式類：計算機概論、程式設計、資料結構、企業資料通訊平均成績；

$X_4$ =管理類：管理學、資訊管理導論平均成績。

在輸出變數方面，即為該系所規劃之三大教育目標主軸。即大三、大四「管理與決策」、「系統與服務」之平均成績做為輸出變數  $Y_1, Y_2$  為連續值(如圖 4-1)。

$Y_1$ =管理與決策類：管理資訊系統、決策支援系統、專案管理、人力資源管理、生產與作業管理、統計學(二)、供應鏈管理、行銷管理、財務管理、電子商務概論、網路行銷、知識管理、科支管理、醫療資訊管理、ERP 企業資源規劃、CRM 顧客關係管理、研究方法、事業行銷、企業資訊系統、資料採礦概論…等「管理與決策類」相關必選修課程平均成績；

$Y_2$ =系統與服務類：資料庫系統與設計、資料庫系統管理、XML 可

延伸標記語言、多媒體製作、CISCO 網路實務(一)、網路服務技術、網路及資料庫應用系統、作業系統、系統分析與設計、資訊安全、演算法、網頁程式設計、JAVA 實務程式設計、物件導向技術…等「系統與服務類」相關必、選修課程平均成績。

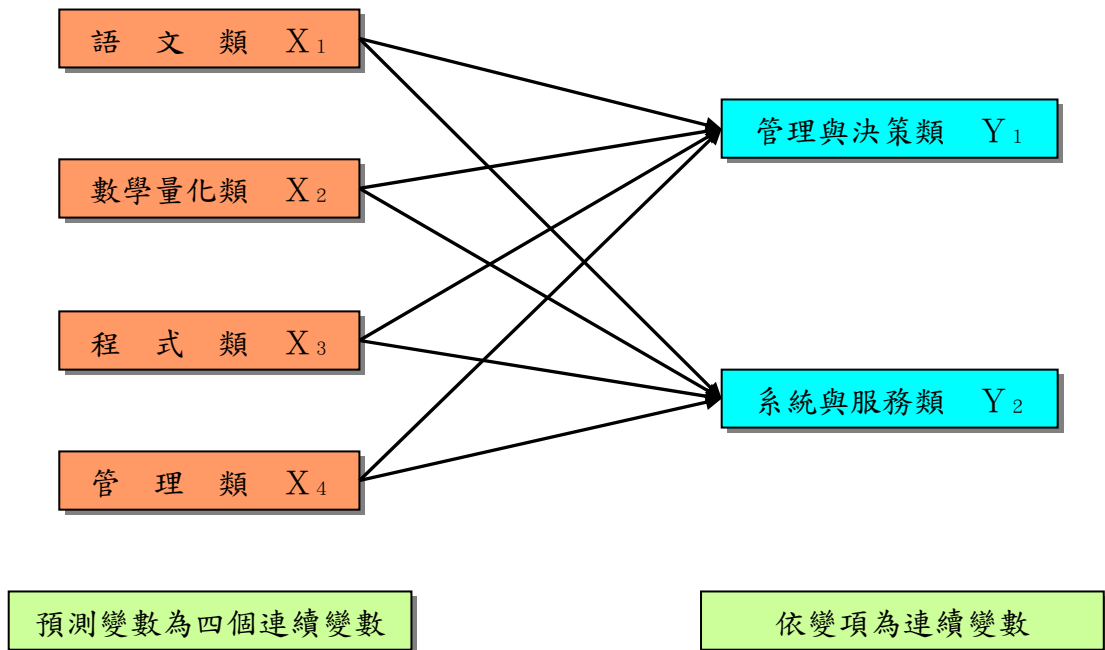


圖 4-1：成績變數分析圖

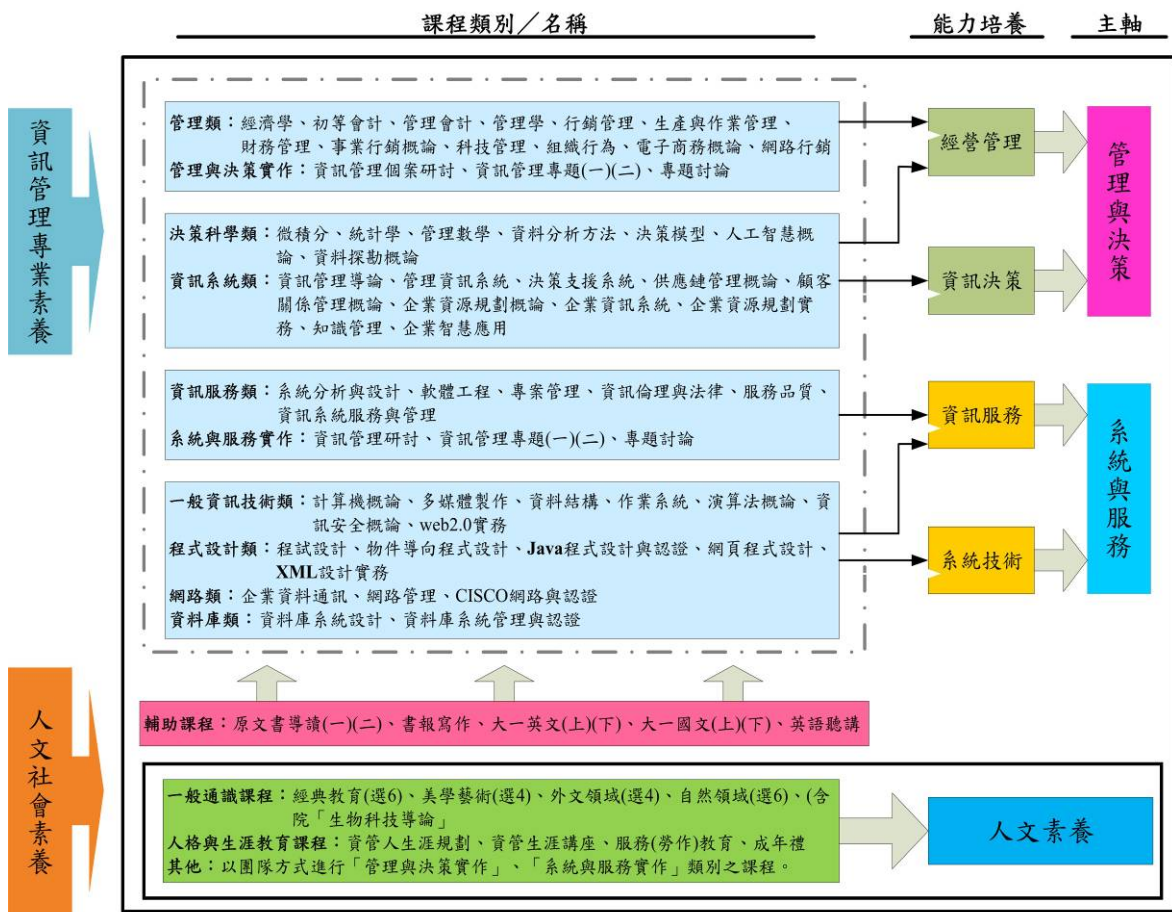


圖 4-2：成績變數分析圖

將學生成績資料匯入 SQL Server 2008 資料採礦工具中，利用 Microsoft SQL Server 2008 內建之類神經網路採礦模型建立資料模型並且進行資料分析，詳如下列：

資料準備好後，接著進行資料採礦作業，其步驟如下：

一、在 Microsoft SQL Server Management Studio 中新增資料庫，選取新增之資料庫，按右鍵選擇 Tasks，選 Import Data，之後選取存在電腦中之資料庫。

二、在 SQL Server Business Intelligence Development Studio 中新增一個 Analysis Service 專案，將資料來源設定為剛新增之資料庫；之後設定資料來源檢視，採礦結構設為類神經網路。

三、依匯入之資料建立決策樹模型，並調整其參數與類別後，分別建立模型。

四、取得容易與文字搭配視覺化圖型之方式呈現所分析之資料型態，利於最後結果之判讀。

## 第二節 類神經網路分析

將資料匯入分析工具後，取得系統與分析類、管理與決策類成績散佈圖(如圖 4-3，4-4)

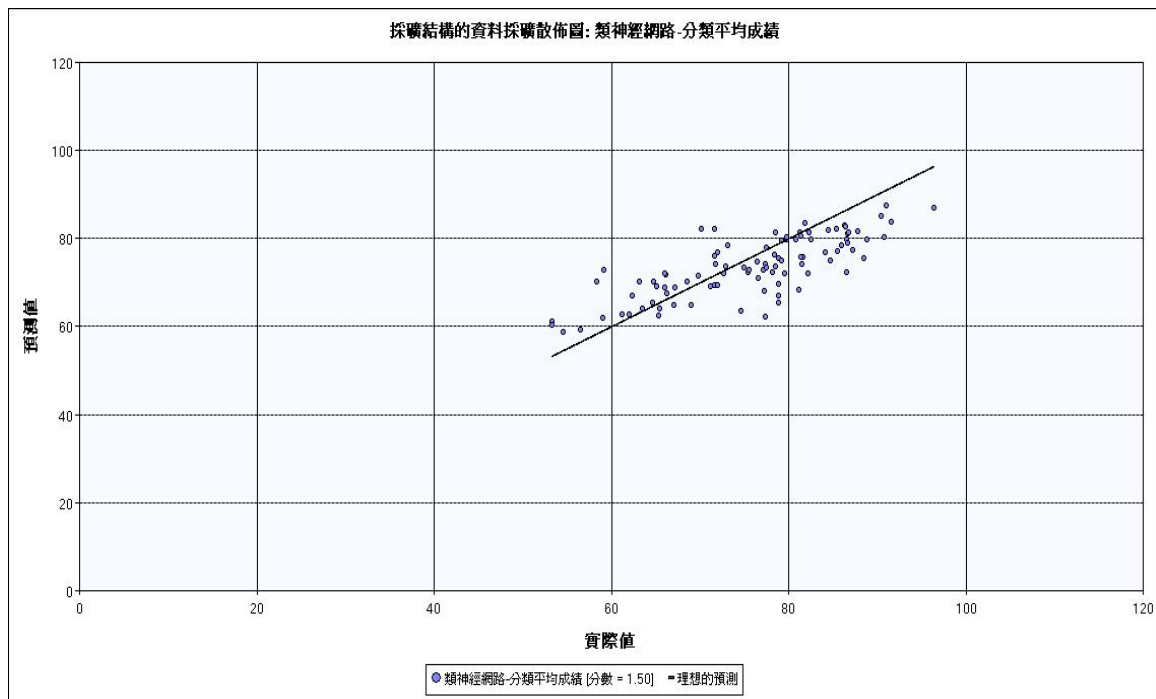


圖 4-3：系統與分析類成績散佈圖

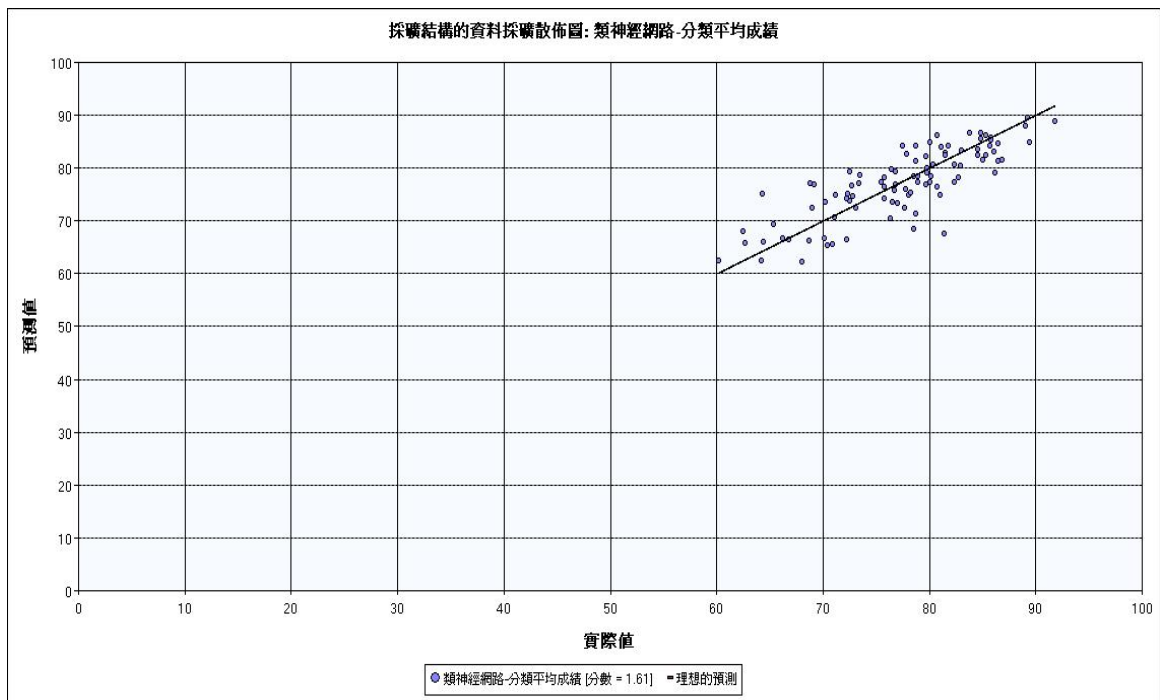


圖 4-4：管理與決策類成績散佈圖

在類神經網路檢示器中，當輸出項目為系統與服務類時，程式類與管理類之成績在 74.625-92.200 與 82.097-98.000 中，皆落點在高分 (80.065-95.000) 區塊中(如圖 4-5)。

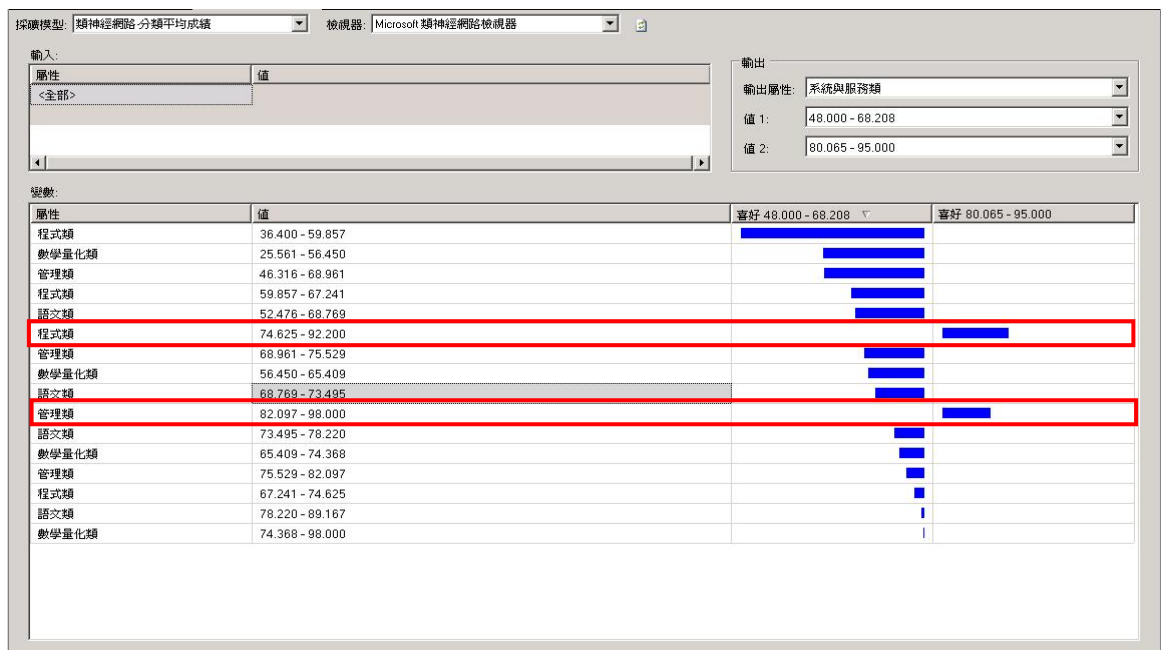


圖 4-5：類神經網路檢視器-輸出項為系統與服務類

在類神經網路檢示器中，當輸出項目為管理與決策類時，管理類、程式類、語文類如為 70-90 分之區塊，則落點在高分(80.065-95.000)區塊中(如圖 4-6)。

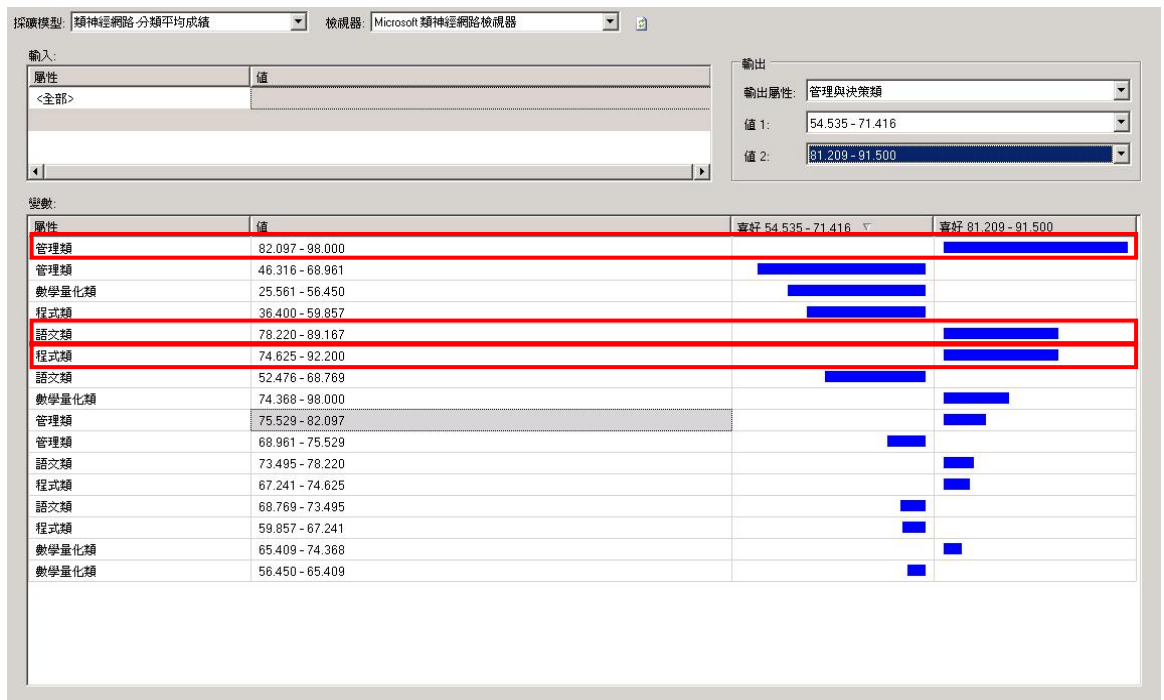


圖 4-6：類神經網路檢視器-輸出項為系統與服務類



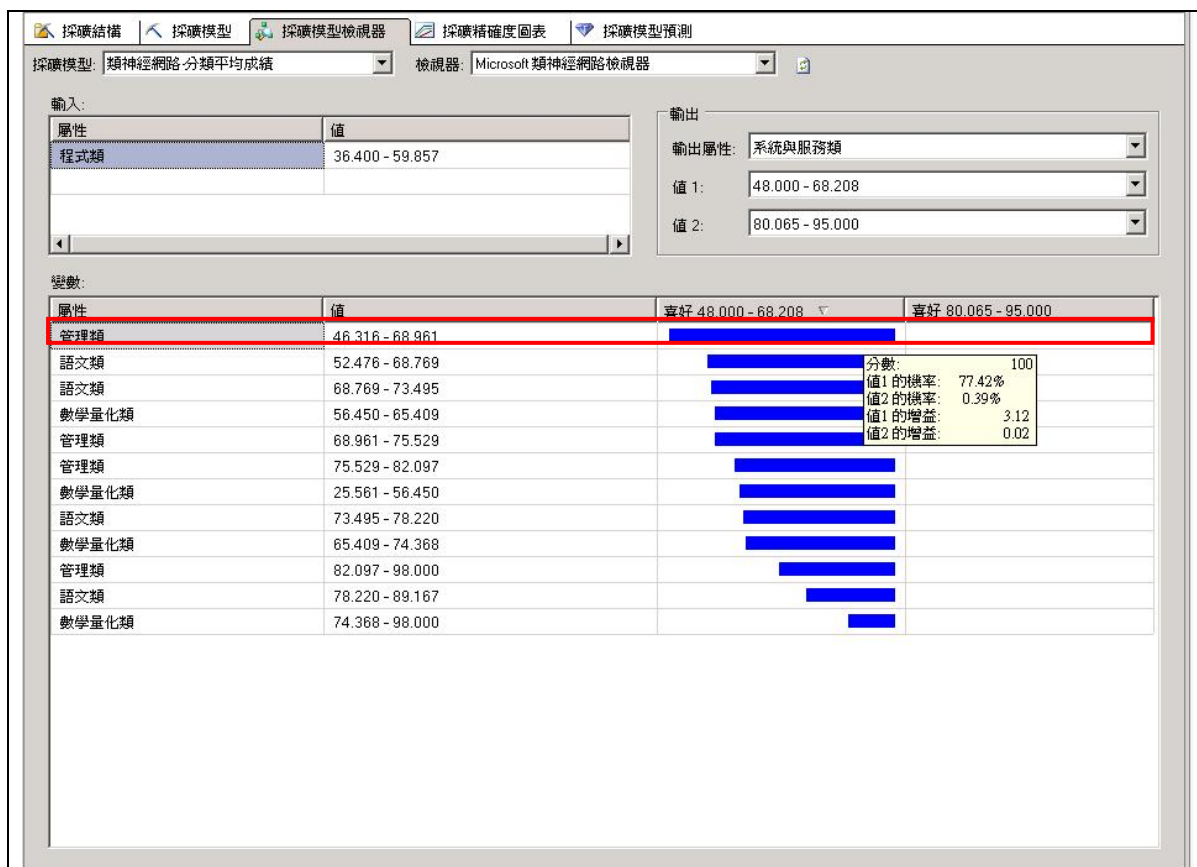


圖 4-7：變數視窗-程式類過濾條件為 36.400-59.857，輸出屬性為系統與服務類

此外，當輸入項目的屬性分別再細分時，其輸出之結果亦有所不同。當輸入屬性為程式類，輸入值為 36.400-59.857 分時，輸出屬性為系統與服務類，輸出值 1 為 48.000-68.208 分、輸出值 2 為 80.065-95.000 分時，發生值 1 的機率為 77.42%，增益值為 3.12；發生值 2 的機率為 0.39%，增益值為 0.02，即在此條件之下，管理類分數為 46.316-68.961 分時，發生值 1 的機率會高於母體 3.12 倍，傾向發生值 1；其變數視窗如圖 4-7。

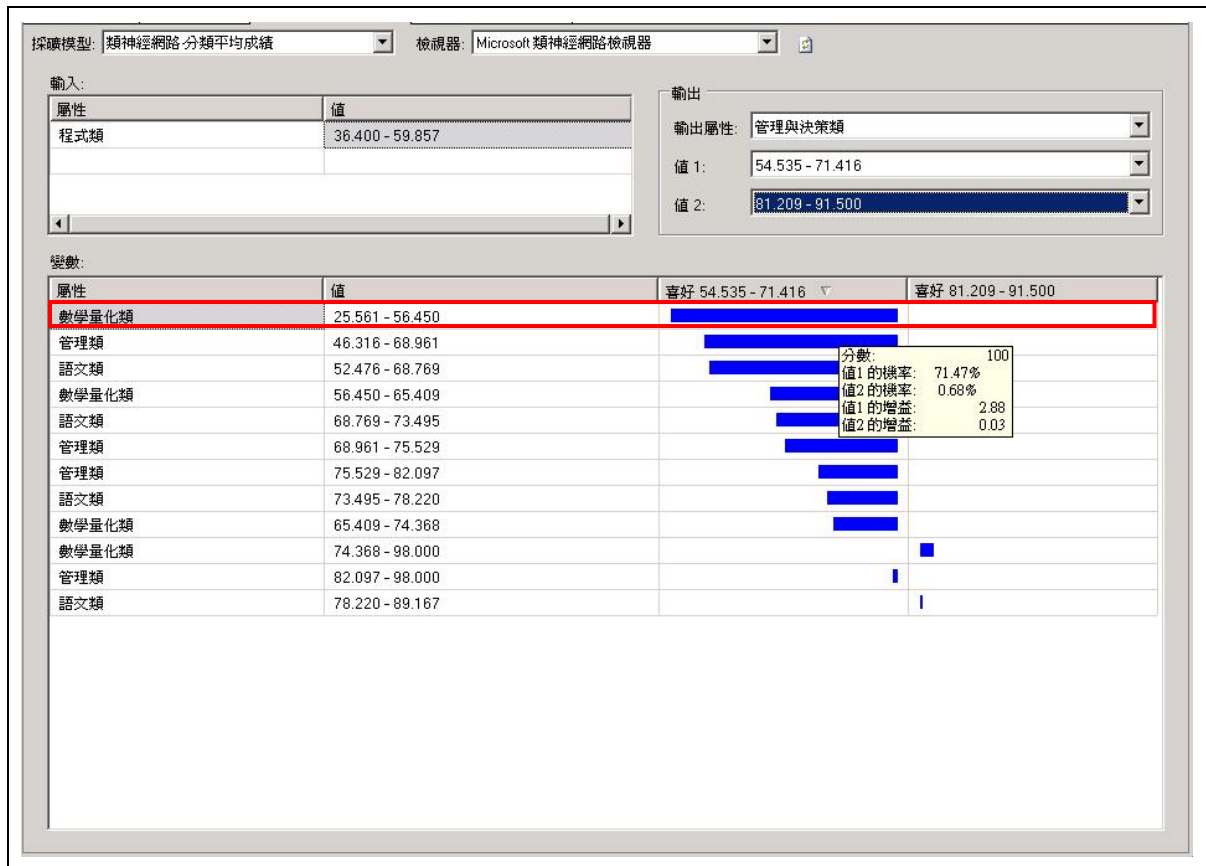


圖 4-8：變數視窗-程式類過濾條件為 36.400-59.857，輸出屬性為管理與決策類

當輸入屬性為程式類，輸入值為 36.400-59.857 分時，輸出屬性為管理與決策類，輸出值 1 為 54.535-71.416 分、輸出值 2 為 81.209-91.500 分時，發生值 1 的機率為 71.47%，增益值為 2.88；發生值 2 的機率為 0.68%，增益值為 0.03，即在此條件之下，數學量化類分數為 25.561-56.450 分時，發生值 1 的機率會高於母體 2.88 倍，傾向發生值 1；其變數視窗如圖 4-8。

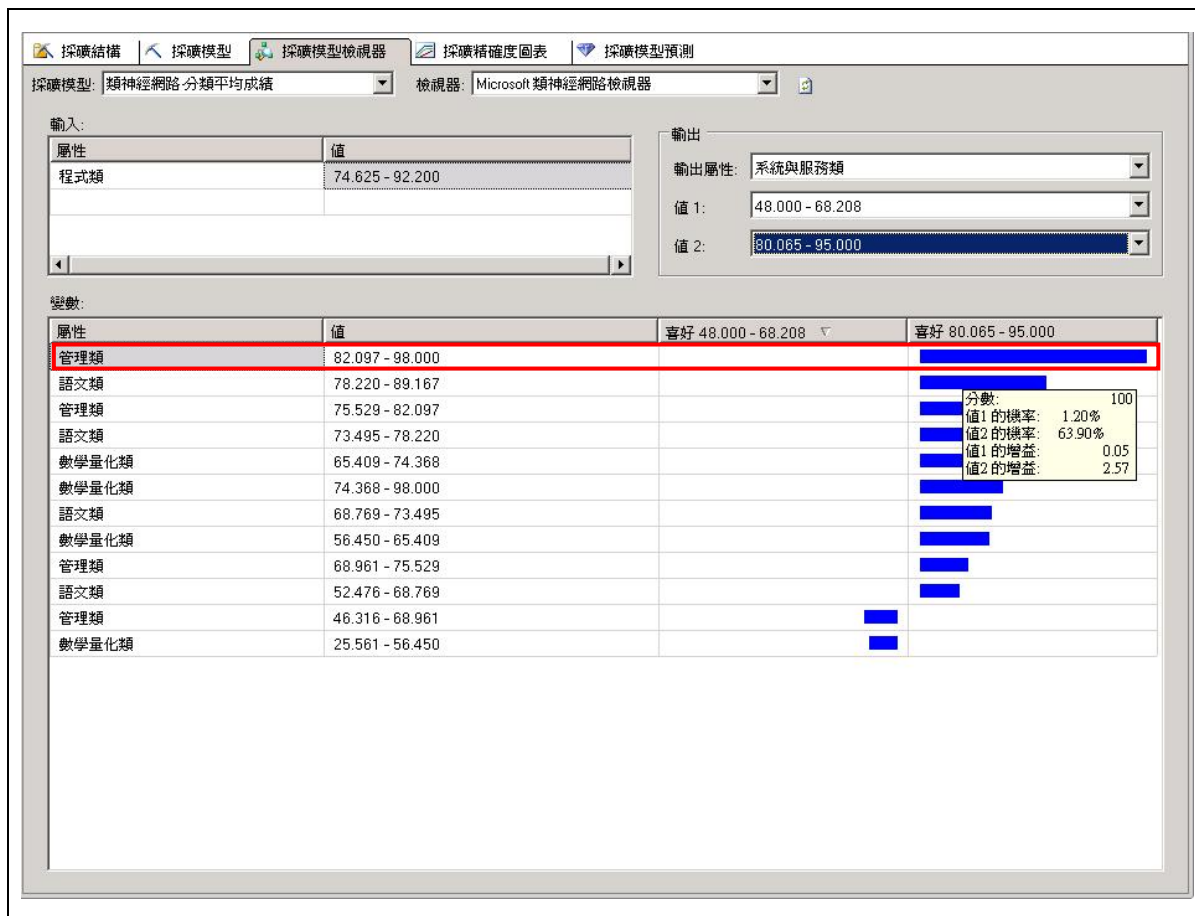


圖 4-9：變數視窗-程式類過濾條件為 74.625-92.200，輸出屬性為系統與服務類

當輸入屬性為程式類，輸入值為 74.625-92.200 分時，輸出屬性為系統與服務類，輸出值 1 為 48.000-68.208 分、輸出值 2 為 80.065-95.000 分時，發生值 1 的機率為 1.20%，增益值為 0.05；發生值 2 的機率為 63.90%，增益值為 2.57，即在此條件之下，管理類分數為 82.097-98.000 分時，發生值 2 的機率會高於母體 2.57 倍，傾向發生值 2；其變數視窗如圖 4-9。

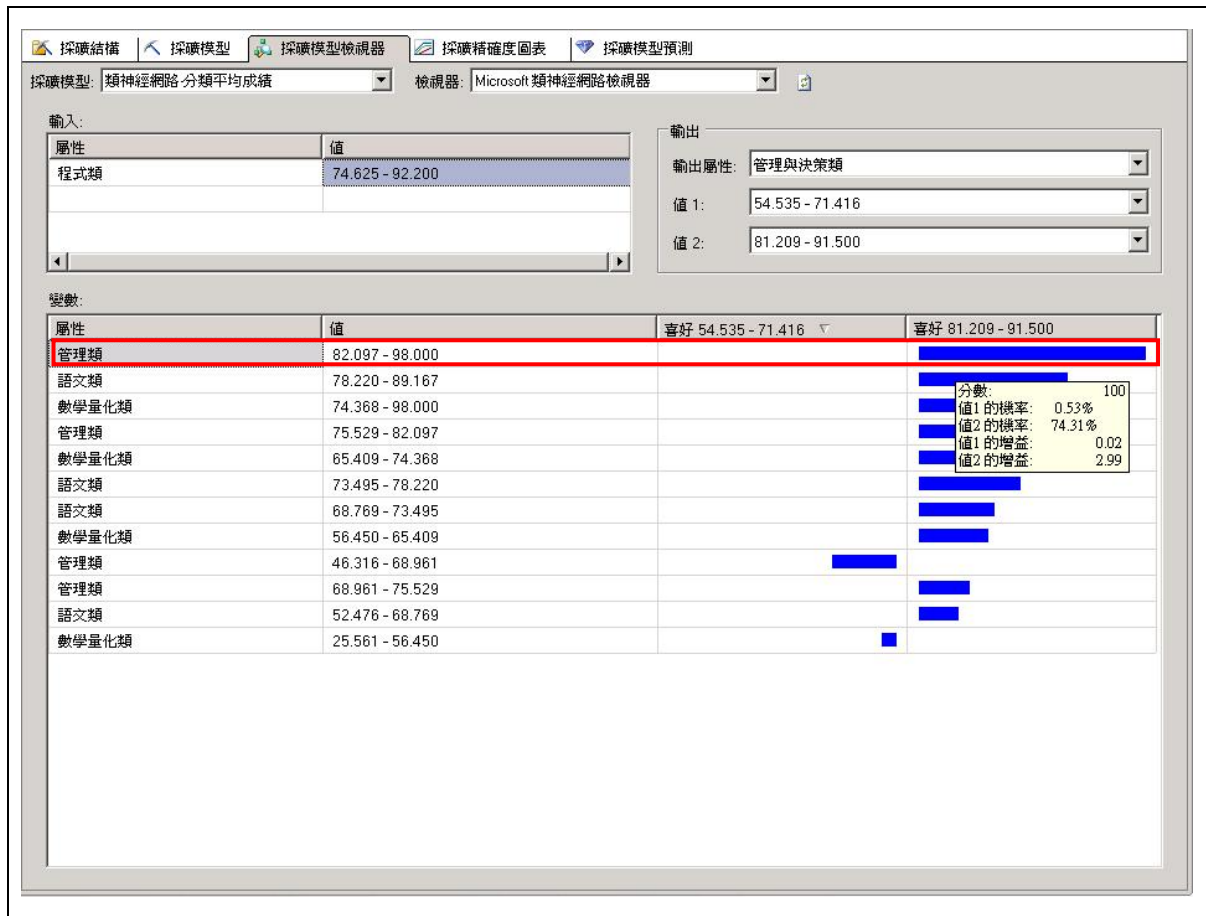


圖 4-10：變數視窗-程式類過濾條件為 74.625-92.200，輸出屬性為管理與決策類

當輸入屬性為程式類，輸入值為 74.625-92.200 分時，輸出屬性為管理與決策類，輸出值 1 為 54.535-71.416 分、輸出值 2 為 81.209-91.500 分時，發生值 1 的機率為 0.53%，增益值為 0.02；發生值 2 的機率為 74.31%，增益值為 2.99，即在此條件之下，管理類分數為 82.097-98.000 分時，發生值 2 的機率會高於母體 2.99 倍，傾向發生值 2；其變數視窗如圖 4-10。

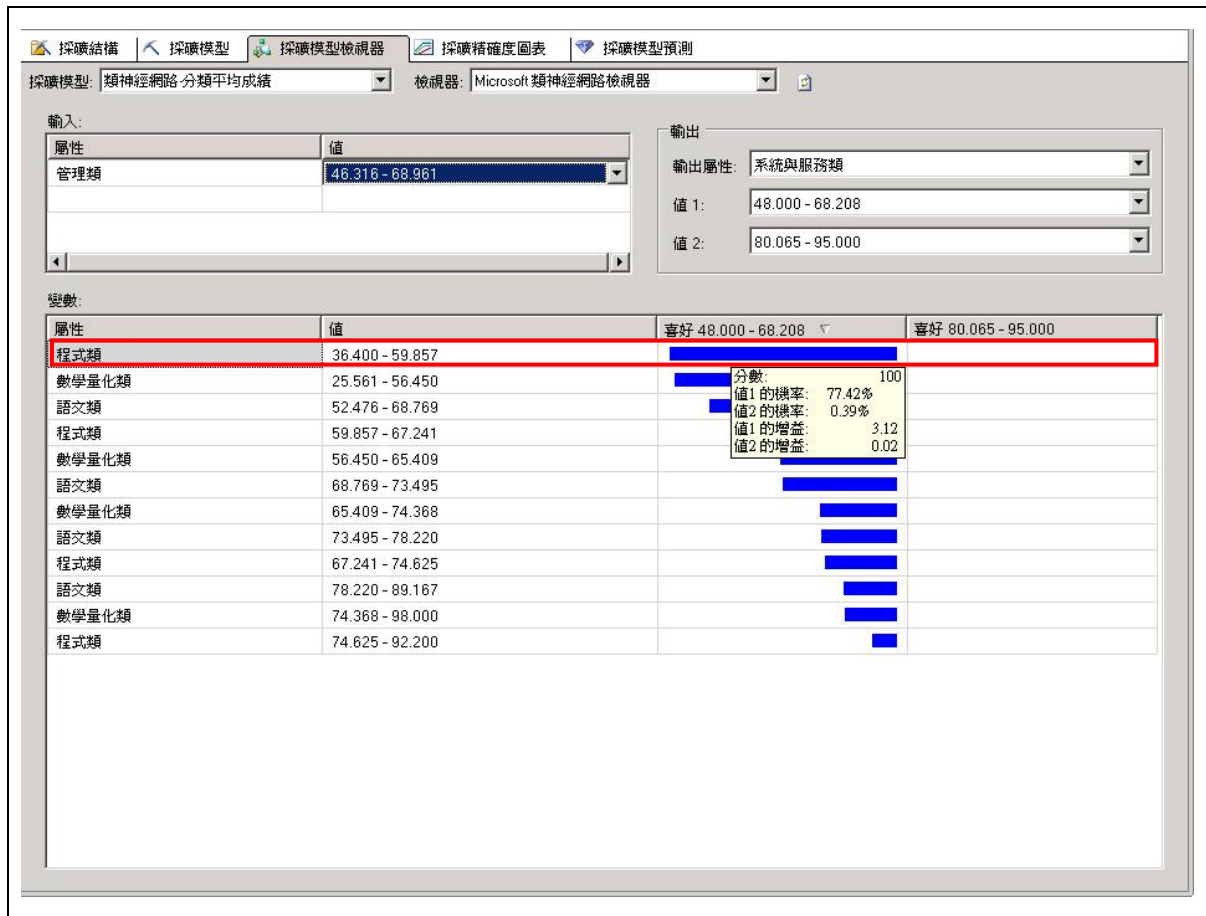


圖 4-11：變數視窗-管理類過濾條件為 46.316-68.916，輸出屬性為系統與服務類

當輸入屬性為管理類，輸入值為 46.316-68.916 分時，輸出屬性為系統與服務類，輸出值 1 為 48.000-68.208 分、輸出值 2 為 80.065-95.000 分時，發生值 1 的機率為 77.42%，增益值為 3.12；發生值 2 的機率為 0.39%，增益值為 0.02，即在此條件之下，程式類分數為 36.400-59.857 分時，發生值 1 的機率會高於母體 3.12 倍，傾向發生值 1；其變數視窗如圖 4-11。

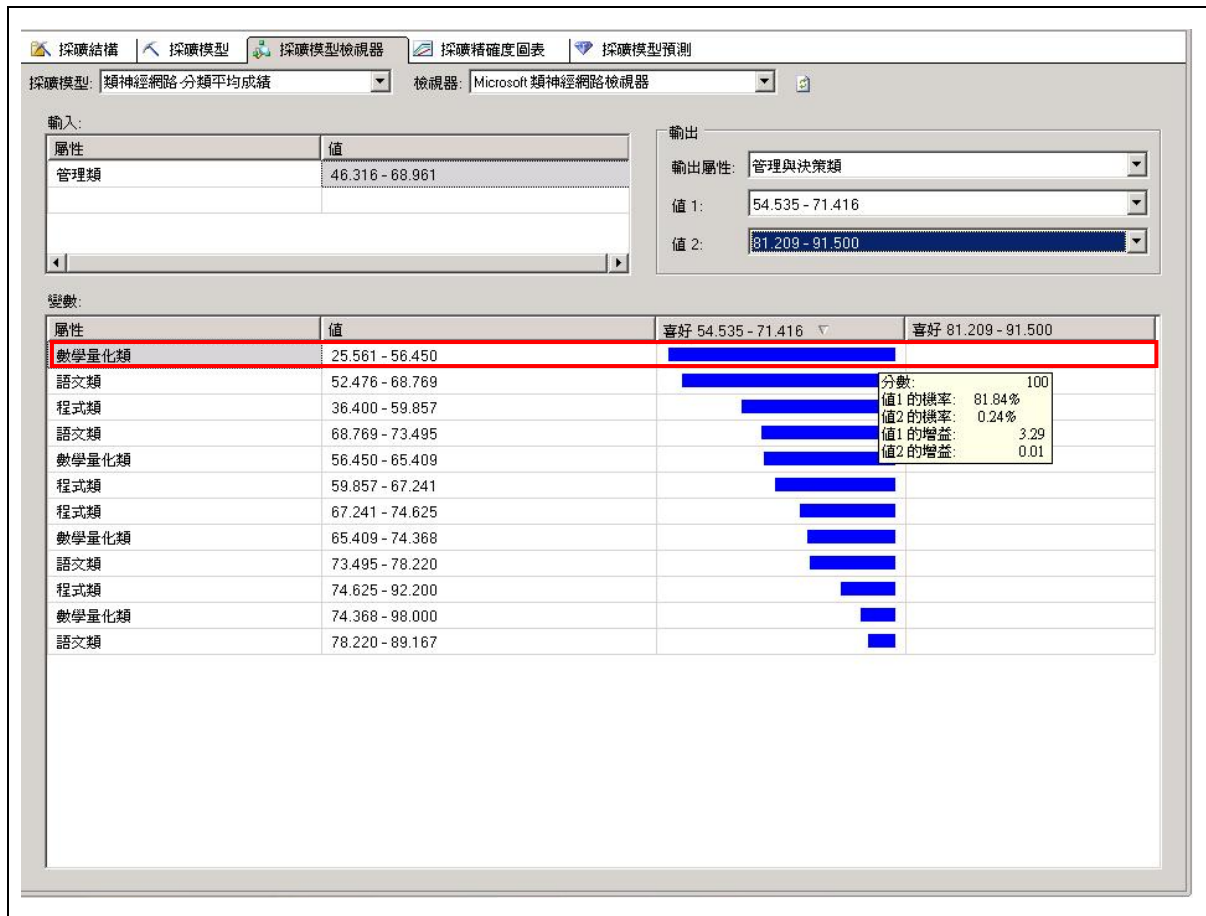


圖 4-12：變數視窗-管理類過濾條件為 46.316-68.916，輸出屬性為管理與決策類

當輸入屬性為管理類，輸入值為 46.316-68.916 分時，輸出屬性為管理與決策類，輸出值 1 為 54.535-71.416 分、輸出值 2 為 81.209-91.500 分時，發生值 1 的機率為 81.84%，增益值為 3.29；發生值 2 的機率為 0.24%，增益值為 0.01，即在此條件之下，數學量化類分數為 25.561-56.450 分時，發生值 1 的機率會高於母體 3.29 倍，傾向發生值 1；其變數視窗如圖 4-12。

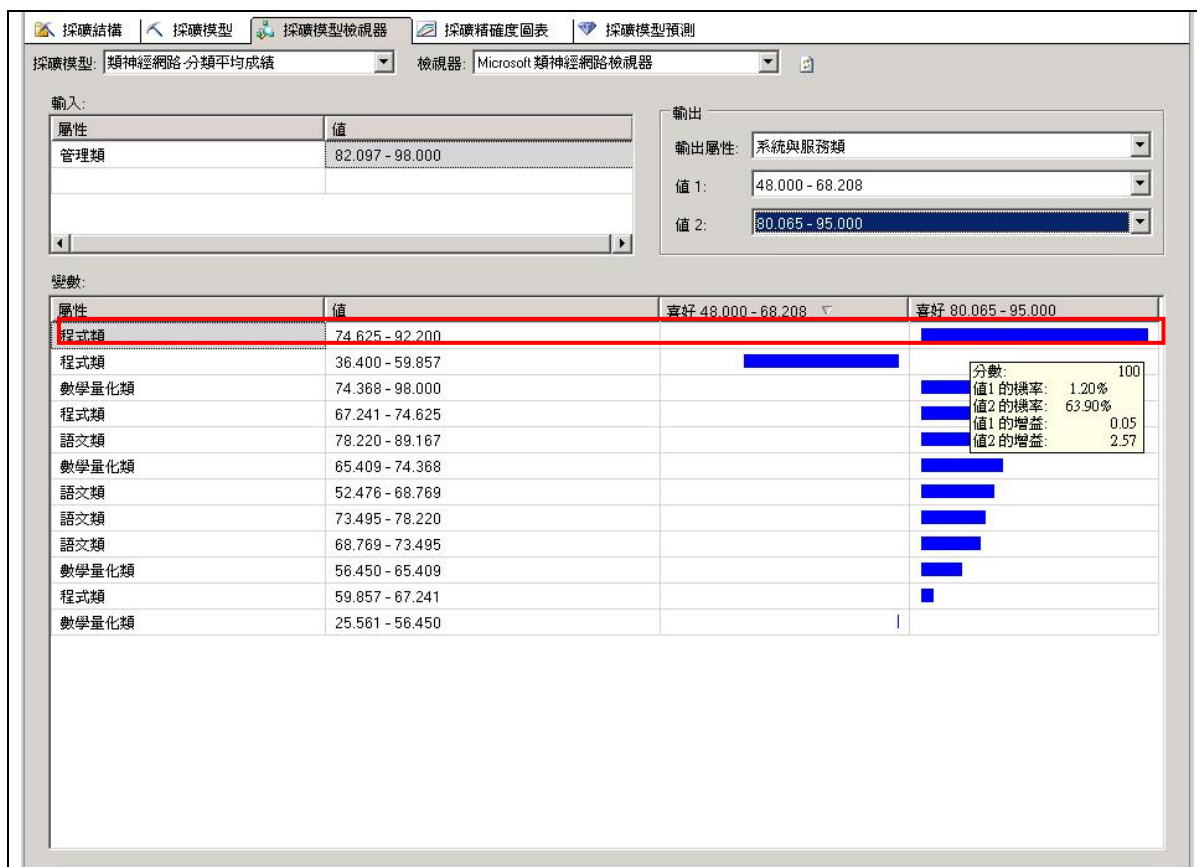


圖 4-13：變數視窗-管理類過濾條件為 82.097-98.000，輸出屬性為系統與服務類

當輸入屬性為管理類，輸入值為 82.097-98.000 分時，輸出屬性為系統與服務類，輸出值 1 為 48.000-68.208 分、輸出值 2 為 80.065-95.000 分時，發生值 1 的機率為 1.20%，增益值為 0.05；發生值 2 的機率為 63.90%，增益值為 2.57，即在此條件之下，程式類分數為 74.625-92.200 分時，發生值 2 的機率會高於母體 2.57 倍，傾向發生值 2；其變數視窗如圖 4-13。

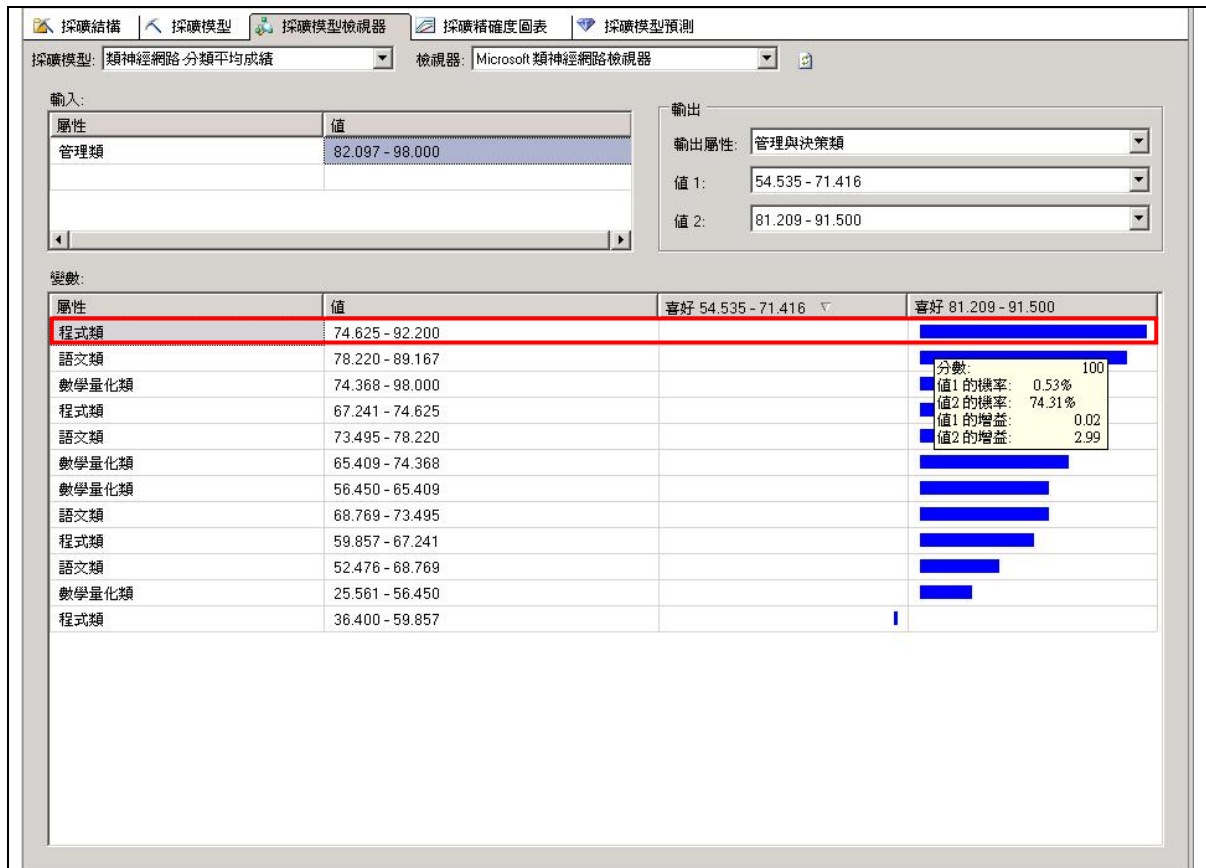


圖 4-14：變數視窗-管理類過濾條件為 82.097-98.000，輸出屬性為管理與決策類

當輸入屬性為管理類，輸入值為 82.097-98.000 分時，輸出屬性為管理與決策類，輸出值 1 為 54.535-71.416 分、輸出值 2 為 81.209-91.500 分時，發生值 1 的機率為 0.53%，增益值為 0.02；發生值 2 的機率為 74.31%，增益值為 2.99，即在此條件之下，程式類分數為 74.625-95.200 分時，發生值 2 的機率會高於母體 2.99 倍，傾向發生值 2；其變數視窗如圖 4-14。



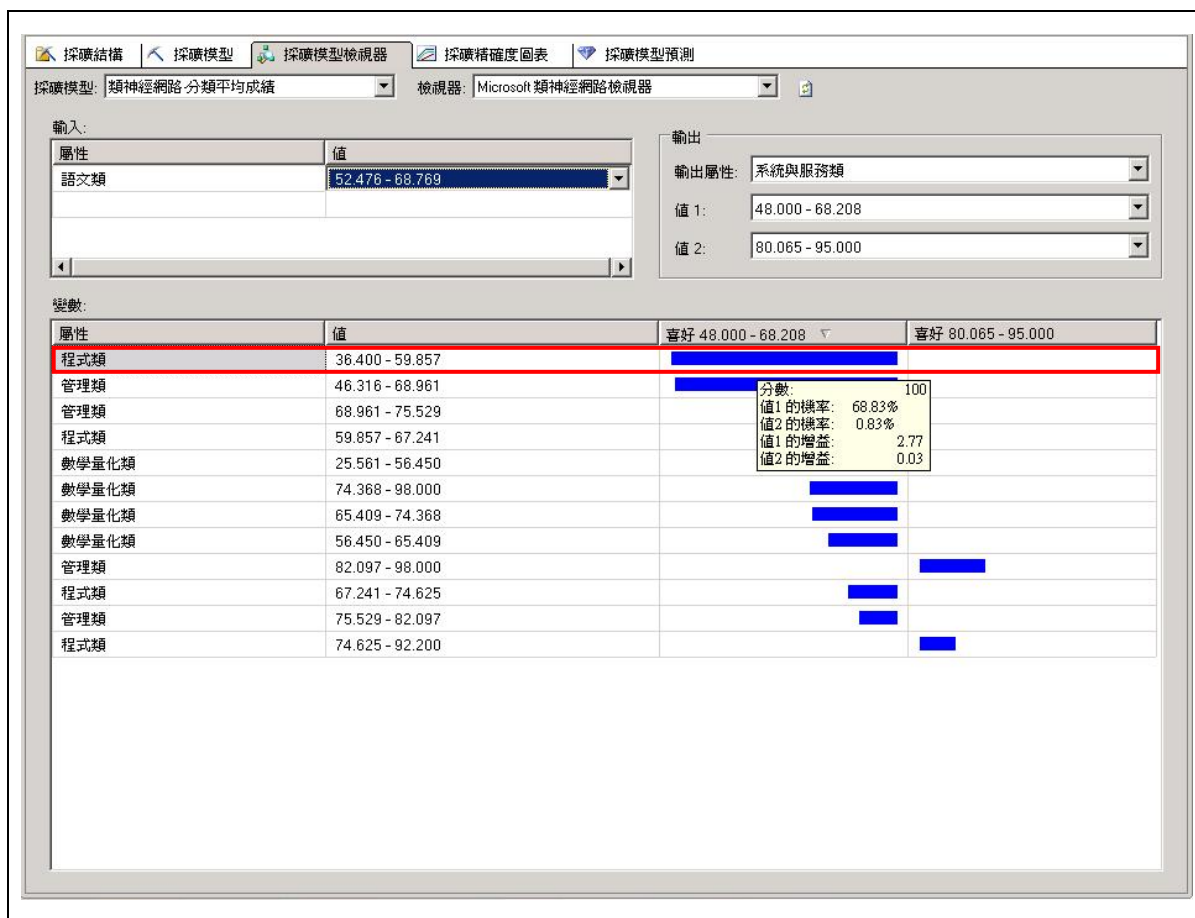


圖 4-15：變數視窗-語文類過濾條件為 52.476-68.769，輸出屬性為系統與服務類

當輸入屬性為語文類，輸入值為 52.476-68.769 分時，輸出屬性為系統與服務類，輸出值 1 為 48.000-68.208 分、輸出值 2 為 80.065-95.000 分時，發生值 1 的機率為 68.83%，增益值為 2.77；發生值 2 的機率為 0.83%，增益值為 0.03，即在此條件之下，程式類分數為 36.400-59.857 分時，發生值 1 的機率會高於母體 2.77 倍，傾向發生值 1；其變數視窗如圖 4-15。

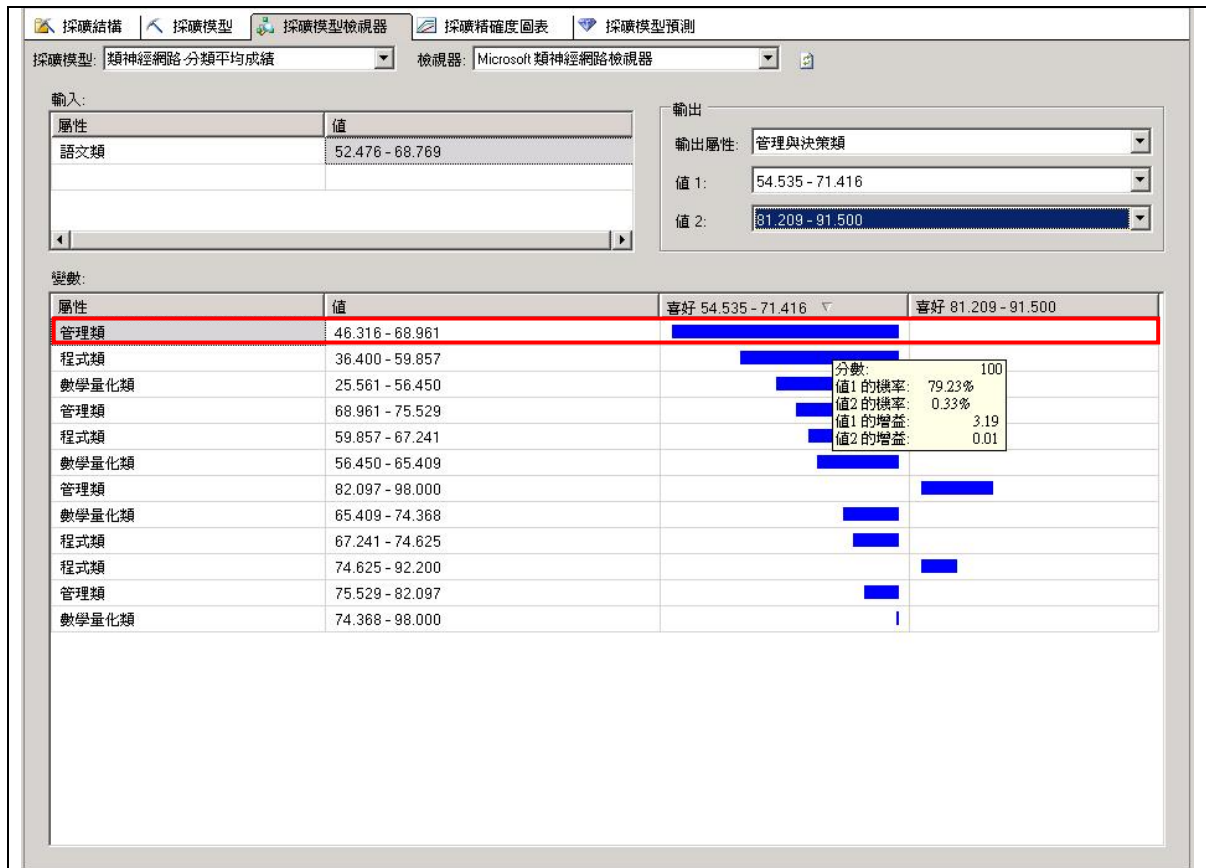


圖 4-16：變數視窗-語文類過濾條件為 52.476-68.769，輸出屬性為管理與決策類

當輸入屬性為語文類，輸入值為 52.476-68.769 分時，輸出屬性為管理與決策類，輸出值 1 為 54.535-71.416 分、輸出值 2 為 81.209-91.500 分時，發生值 1 的機率為 79.23%，增益值為 3.19；發生值 2 的機率為 0.33%，增益值為 0.01，即在此條件之下，管理類分數為 46.316-68.961 分時，發生值 1 的機率會高於母體 3.19 倍，傾向發生值 1；其變數視窗如圖 4-16。

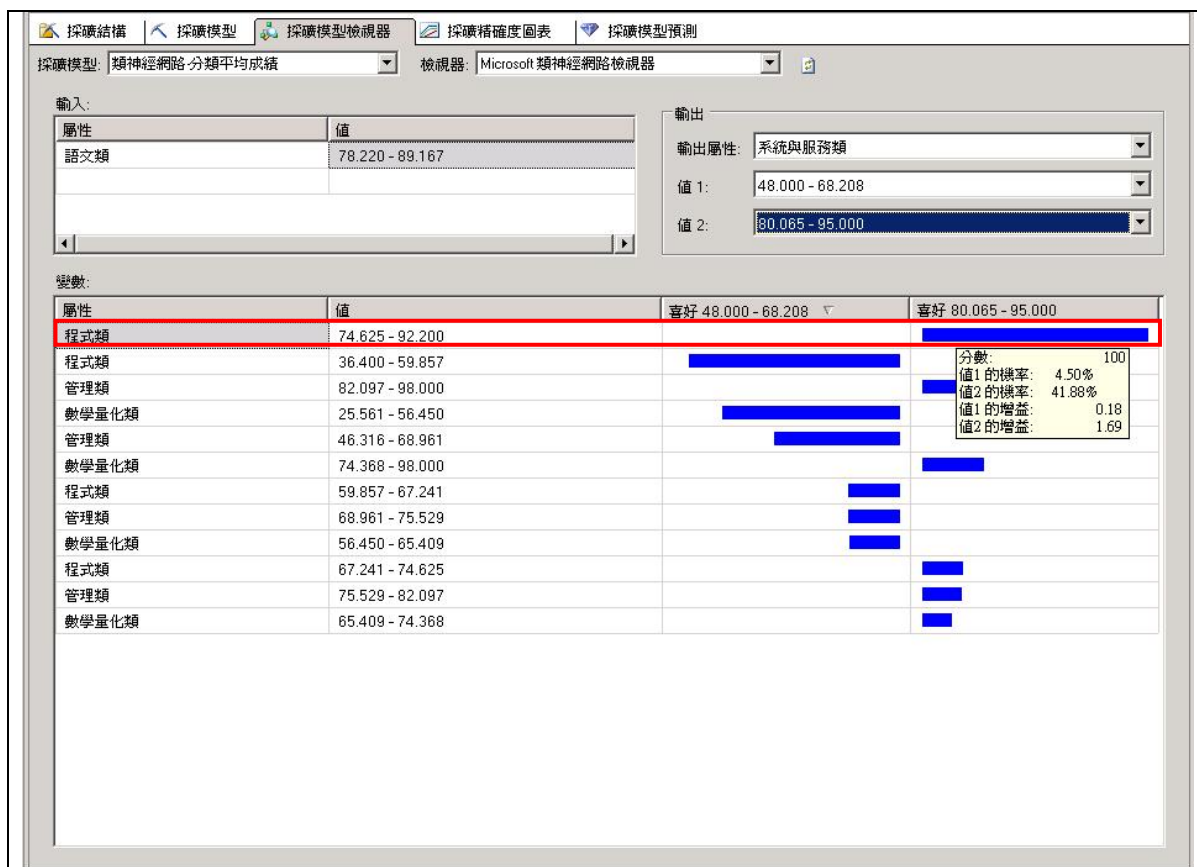


圖 4-17：變數視窗-語文類過濾條件為 78.220-89.167，輸出屬性為系統與服務類

當輸入屬性為語文類，輸入值為 78.220-89.167 分時，輸出屬性為系統與服務類，輸出值 1 為 48.000-68.208 分、輸出值 2 為 80.065-95.000 分時，發生值 1 的機率為 4.50%，增益值為 0.18；發生值 2 的機率為 41.88%，增益值為 1.69，即在此條件之下，程式類分數為 74.625-92.200 分時，發生值 2 的機率會高於母體 1.69 倍，傾向發生值 2；其變數視窗如圖 4-17。

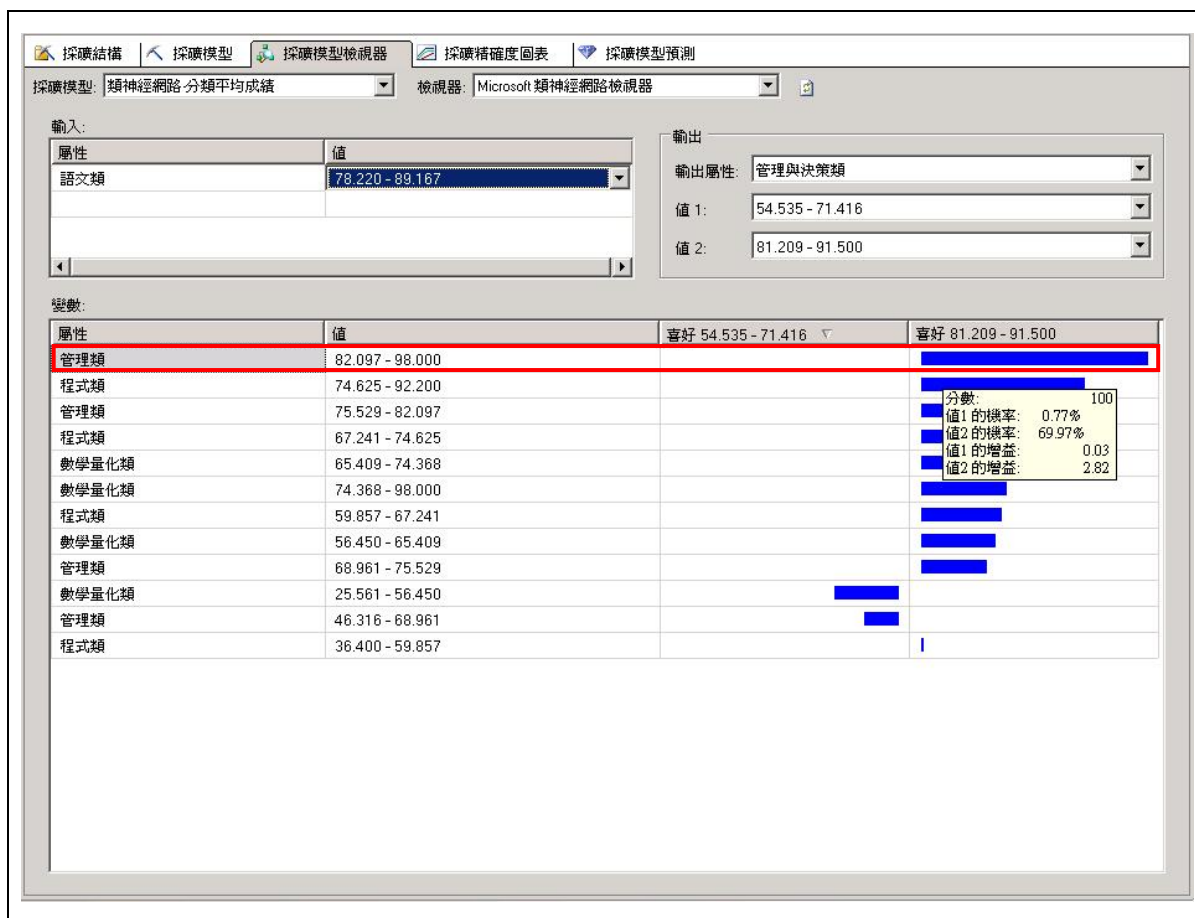


圖 4-18：變數視窗-語文類過濾條件為 78.220-89.167，輸出屬性為管理與決策類

當輸入屬性為語文類，輸入值為 78.220-89.167 分時，輸出屬性為管理與決策類，輸出值 1 為 54.535-71.416 分、輸出值 2 為 81.209-91.500 分時，發生值 1 的機率為 0.77%，增益值為 0.03；發生值 2 的機率為 69.97%，增益值為 2.82，即在此條件之下，管理類分數為 82.097-98.000 分時，發生值 2 的機率會高於母體 2.82 倍，傾向發生值 2；其變數視窗如圖 4-18。

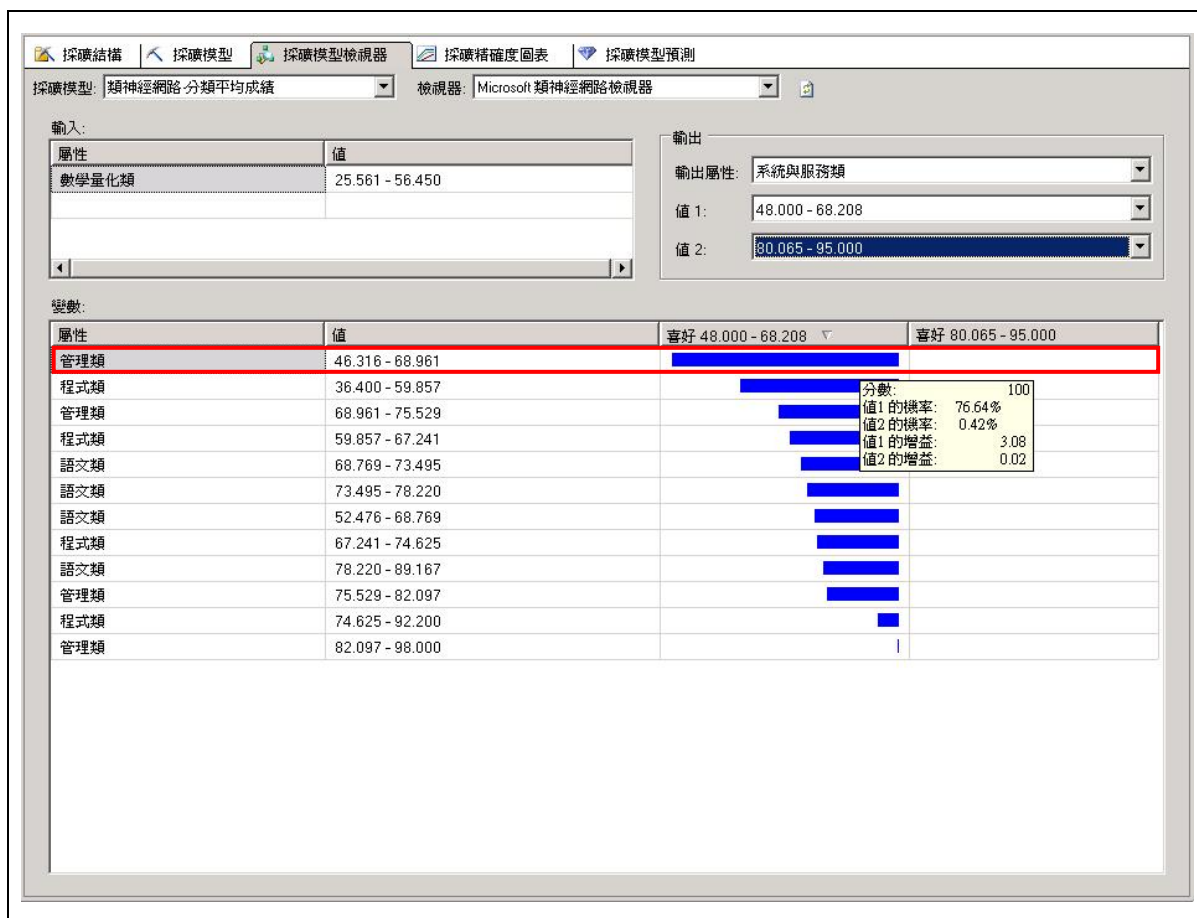


圖 4-19：變數視窗-數學量化類過濾條件為 25.561-56.450，輸出屬性為系統與服務類

當輸入屬性為數學量化類，輸入值為 25.561-58.450 分時，輸出屬性為系統與服務類，輸出值 1 為 48.000-68.208 分、輸出值 2 為 80.065-95.000 分時，發生值 1 的機率為 76.64%，增益值為 3.08；發生值 2 的機率為 0.42%，增益值為 0.02，即在此條件之下，管理類分數為 46.316-68.961 分時，發生值 1 的機率會高於母體 3.08 倍，傾向發生值 1；其變數視窗如圖 4-19。

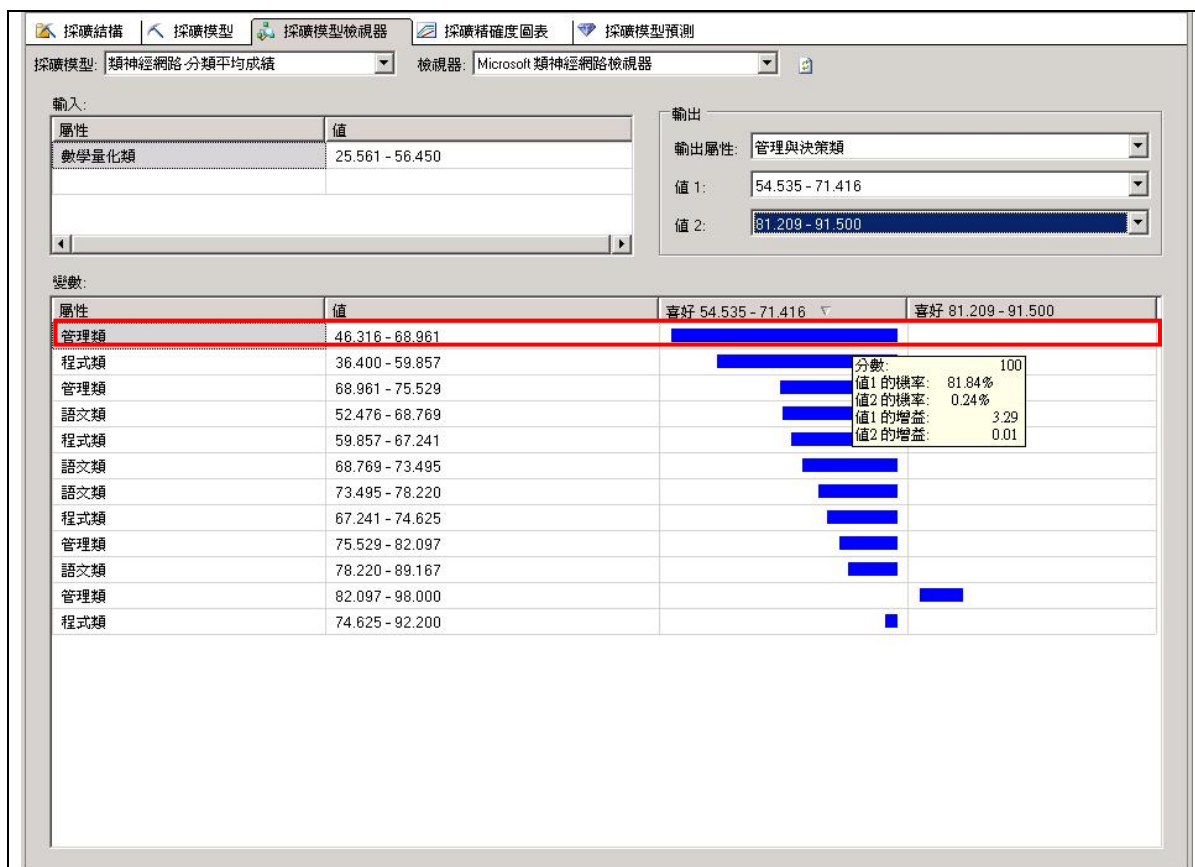


圖 4-20：變數視窗-數學量化類過濾條件為 25.561-56.450，輸出屬性為管理與決策類

當輸入屬性為數學量化類，輸入值為 25.561-58.450 分時，輸出屬性為管理與決策類，輸出值 1 為 54.535-71.416 分、輸出值 2 為 81.209-91.500 分時，發生值 1 的機率為 81.84%，增益值為 3.29；發生值 2 的機率為 0.24%，增益值為 0.01，即在此條件之下，管理類分數為 46.316-68.961 分時，發生值 1 的機率會高於母體 3.29 倍，傾向發生值 1；其變數視窗如圖 4-20。

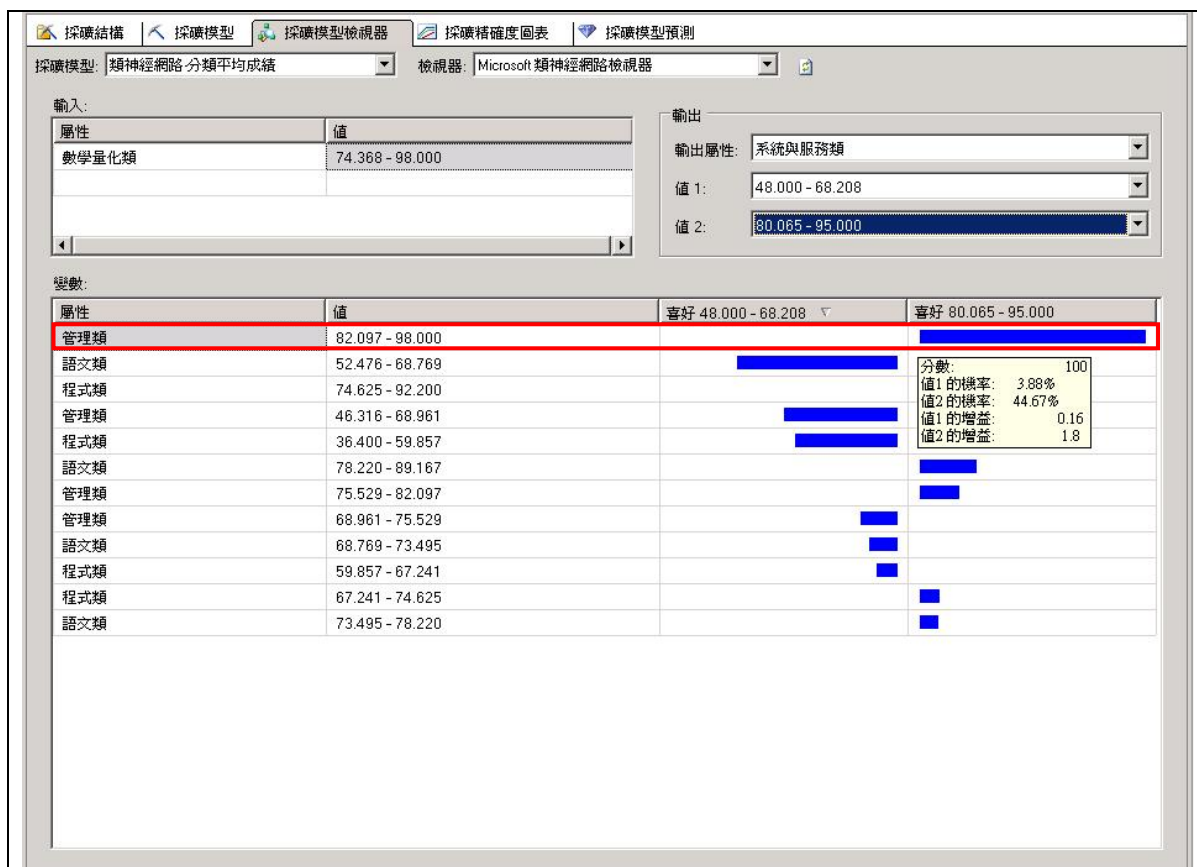


圖 4- 21：變數視窗-數學量化類過濾條件為 74.368-98.000，輸出屬性為系統與服務類

當輸入屬性為數學量化類，輸入值為 74.368-98.000 分時，輸出屬性為系統與服務類，輸出值 1 為 48.000-68.208 分、輸出值 2 為 80.065-95.000 分時，發生值 1 的機率為 3.88%，增益值為 0.16；發生值 2 的機率為 44.67%，增益值為 1.8，即在此條件之下，管理類分數為 82.097-98.000 分時，發生值 2 的機率會高於母體 1.8 倍，傾向發生值 2；其變數視窗如圖 4-21。

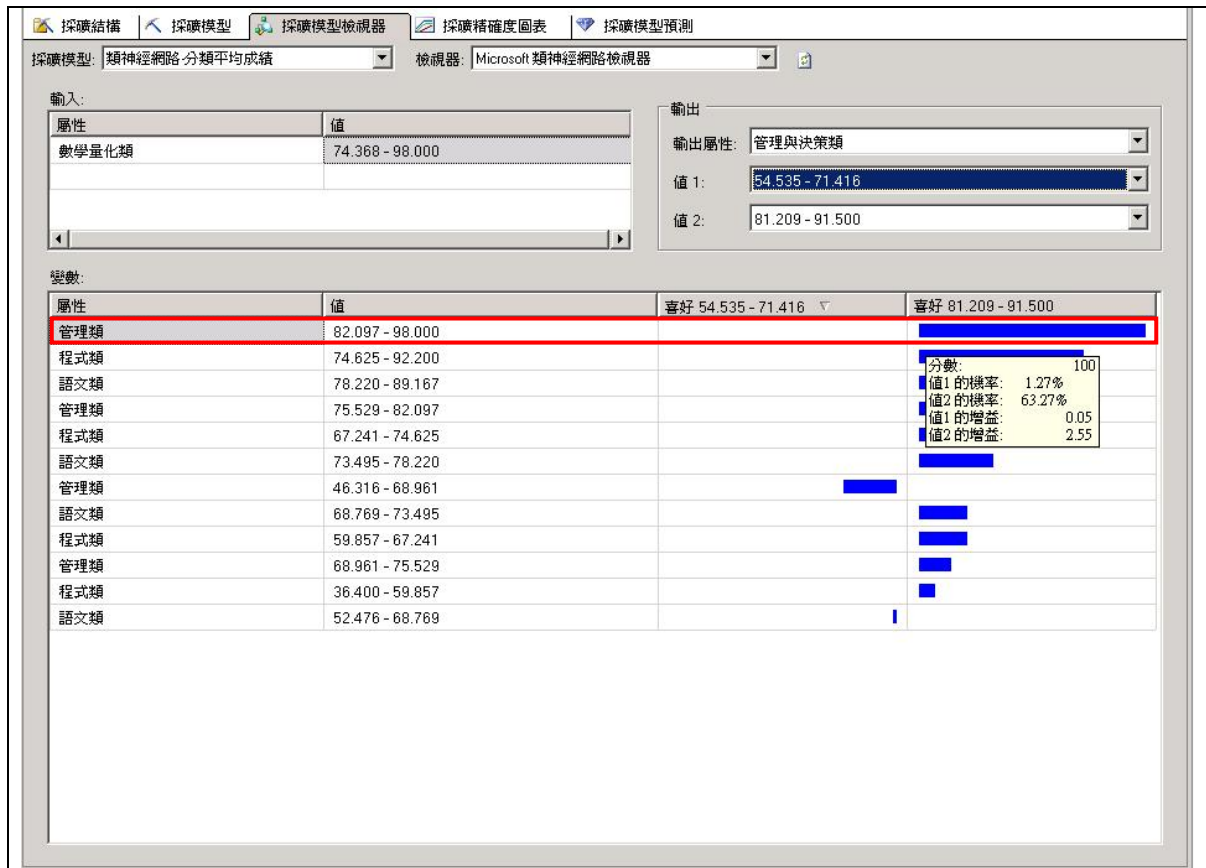


圖 4-22：變數視窗-數學量化類過濾條件為 74.368-98.000，輸出屬性為管理與決策類

當輸入屬性為數學量化類，輸入值為 74.368-98.000 分時，輸出屬性為管理與決策類，輸出值 1 為 54.535-71.416 分、輸出值 2 為 81.209-91.500 分時，發生值 1 的機率為 1.27%，增益值為 0.05；發生值 2 的機率為 63.27%，增益值為 2.55，即在此條件之下，管理類分數為 82.097-98.000 分時，發生值 2 的機率會高於母體 2.55 倍，傾向發生值 2；其變數視窗如圖 4-22。



## 第五章結論與建議

### 第一節 研究結論

由於學生在學習之過程中，最忌諱的是不清楚自己未來想發展的方向而胡亂選擇研讀課程，造成未來在畢業時對自己信心不足及無專業專長。而學生在大學入學時，所面臨到的上課方式是必須由自己從系所中所開出之課程，去選擇所想要研讀以及必修的課程科目；與往常在國高中時的填鴨式方法完全不同。

如何讓學生在他們在校就學期間，除了解現今業界型態以及所需之人力類型之外，從大學中眾多課程裡，選擇適合自己的特質以及適合發展的課程研習，且將所選擇出來之課程獲得所需之知識，累積其就業競爭力，已成為各學系的眾多重要課題之一。

從分析結果「系統與服務類」、「管理與決策類」的成績散佈圖結果顯示，系統與服務類相對分數偏低分，落點有位於 60 分之位置；管理與決策類相對分數偏低分，落點皆位於 60 分以上。並且可從中發現管理與決策類分數高於系統與服務類；此外，由於預測出來之效果較為集中分佈在理想線附近，所以，預測效果頗佳。

從細部分析後，所得之結果，詳如下表 5-1 所示。在編號 1，輸入屬性為程式類，輸入值為 36.400-59.857 分時，輸出屬性為系統與服務類，輸出值 1 為 48.000-68.208 分、輸出值 2 為 80.065-95.000 分時，發生值 1 的機率為 77.42%，增益值為 3.12；發生值 2 的機率為 0.39%，增益值為 0.02，即在此條件之下，管理類分數為 46.316-68.961 分時，發生值 1 的機率會高於母體 3.12 倍，傾向發生值 1。

表 5- 1：細部分析結果，由本研究整理。

編號	輸入屬性	輸入分數值	輸出屬性	發生值1(分數)	發生值2(分數)	高變數屬性	分數值	增益值1	增益值2	建議選課方向
1	程式類	36.400-59.857	系統與服務類	48.000-68.208	80.065-95.000	管理類	46.316-68.961	3.12	0.02	管理與決策類
2	程式類	36.400-59.857	管理與決策類	54.535-71.416	81.209-91.500	數學量化類	25.561-56.450	2.88	0.03	系統與服務類
3	程式類	74.625-92.200	系統與服務類	48.000-68.208	80.065-95.000	管理類	82.097-98.000	0.05	2.57	系統與服務類
4	程式類	74.625-92.200	管理與決策類	54.535-71.416	81.209-91.500	管理類	82.097-98.000	0.02	2.99	系統與服務類
5	管理類	46.316-68.916	系統與服務類	48.000-68.208	80.065-95.000	程式類	36.400-59.857	3.12	0.02	管理與決策類
6	管理類	46.316-68.916	管理與決策類	54.535-71.416	81.209-91.500	數學量化類	25.561-56.450	3.29	0.01	系統與服務類
7	管理類	82.097-98.000	系統與服務類	48.000-68.208	80.065-95.000	程式類	74.625-92.200	0.05	2.57	系統與服務類
8	管理類	82.097-98.000	管理與決策類	54.535-71.416	81.209-91.500	程式類	74.625-95.200	0.02	2.99	管理與決策類
9	管理類	52.476-68.769	系統與服務類	48.000-68.208	80.065-95.000	程式類	36.400-59.857	2.77	0.03	管理與決策類
10	語文類	52.476-68.769	管理與決策類	54.535-71.416	81.209-91.500	管理類	46.316-68.961	3.19	0.01	系統與服務類
11	語文類	78.220-89.167	系統與服務類	48.000-68.208	80.065-95.000	程式類	74.625-92.200	0.18	1.69	系統與服務類
12	語文類	78.220-89.167	管理與決策類	54.535-71.416	81.209-91.500	管理類	82.097-98.000	0.03	2.82	管理與決策類
13	數學量化類	25.561-58.450	系統與服務類	48.000-68.208	80.065-95.000	管理類	46.316-68.961	3.08	0.02	管理與決策類
14	數學量化類	25.561-58.450	管理與決策類	54.535-71.416	81.209-91.500	管理類	46.316-68.961	3.29	0.01	管理與決策類
15	數學量化類	74.368-98.000	系統與服務類	48.000-68.208	80.065-95.000	管理類	82.097-98.000	0.16	1.8	系統與服務類
16	數學量化類	74.368-98.000	管理與決策類	54.535-71.416	81.209-91.500	管理類	82.097-98.000	0.05	2.55	管理與決策類

## 第二節 研究建議

由於類神經網路中的解有很多組，本研究從各類別中分別挑選輸入屬性之輸入值與輸出屬性之發生值 1、2 的最低分與最高分的級別做研究。建議可由其他參數進行分析，建立更加完整之建議選課方向參考以期使之更加完善，以供學生做為大學二年級升三年級時選課之參考。

## 參考文獻

### 一、中文部分

- 1.黃南傑，「高效率拆解之關聯規則探勘」，南台科技大學資訊管理研究所碩士論文，民93年
- 2.溫侑柯，「應用資料探勘之關聯法則探討大學入學成績對在學成績的影響-以資管系為例」，南華大學資訊管理研究所碩士論文，民95年
- 3.許依宸，「資料採礦在學生流失偵測上之應用」，南華大學資訊管理研究所碩士論文，民98年
- 4.簡伊汝，「學生基本能力預測學習優勢之研究以某私立大學資訊管理學系為例」，國立雲林科技大學工業工程與管理系研究所碩士論文，民98年
- 5.傅說道，「國中基測與高中在校成績對於大學學測成績關係之研究-以南部某高中為例」，南台科技大學資訊管理研究所碩士論文，民99年
- 6.李佳玲，「大學入學測驗與高中在校成績關係之研究」，國立臺北師範學院國民教育研究所碩士論文，民90年
- 7.陸炳杉，「多元入學學生學業成就之研究-以高雄市立中正高級中學為例」，國立高雄師範大學工業科技教育學系碩士論文，民92年
- 8.彭重恩，「高中學生生活科技及數理成績與大學學科能力測驗數理科成

- 績之相關研究」，國立臺灣師範大學工業科技教育研究所碩士論文，民  
94 年
- 9.任眉眉、陳日昇、詹嘉豪，「統計與落點分析：大學指考選填志願的輔助利器」，中國統計學報，第 43 卷，2005，頁 165-181。
- 10.謝邦昌、鄭宇庭、蘇志雄，SQL Server 2008 R2 資料採礦與商業智慧，  
台北市，基峯資訊，2011
- 11.尹相志，SQL Server 2008 Data Mining 資料採礦，台北市，悅知文化、  
精誠資訊，2009

## 二、網頁部分

1. 教育部高教司統計處
2. 中華資料採礦協會
3. [http://vega.cs.tku.edu.tw/~cyh/data\\_mining/DM\\_Exercise\\_Chapter1.html](http://vega.cs.tku.edu.tw/~cyh/data_mining/DM_Exercise_Chapter1.html)