資料採礦在學生流失偵測上之應用

Application of Data Mining Techniques for Detection of Student Drop out

邱宏彬 許依宸
Chiu, Hung-Pin Hsu, Yi-Chen
南華大學資訊管理研究所
Department of Information Management, Nan-hun University

摘要

依據教育部資料顯示，本國大專校院數量不斷增加，大學日間部學生招生人數亦隨之增加，然而，根據內政部統計資料表示，本國出生人口總數相對持續減少中，以致大專院校招生情況日益嚴峻，亦影響自國小、國中至高中職未來招生困難之處。對上述問題，本研究運用資料採礦技術，以某大學 94-96 學年度大學日間部學生資料為例，分析學生歷史學籍資料，以建立學生流失預測模型並進行預測評估，同時探索流失學生之表徵。本研究藉由測試驗證預測模型效益之可行性，經研究顯示，成績為學生流失的主要影響因素。最後，本研究針對研究結果，提出相關建議，以供學校參考以改善學生流失之情況。

關鍵字：資料採礦、學生流失、決策樹、類神經網路

Abstract

According to the data from Ministry of Education, the number of college increase constantly. The enrollment of students in the day school also increases. However, the statistics from Ministry of Internal Affair showed that total birth rate of Taiwan is decreasing. Therefore, the enrollment status is getting tougher. This research is based on the day school student enrollment data between 2005~2007 in a University. By using the data mining techniques to analyze the historical records of students, this research tried to find the potential reasons of why students drop out, and to create a prediction model for student drop out. The analyzing and mining results from this research will provide relevant suggestions to help the university to improve the situation of student drop out.

Keyword: data mining, student drop out, decision trees, neural networks

1. 緒論

依據教育部高教司統計處資料顯示，本國大學校院數量由民國 75 學年度的 28 所至 97 學年度已增設到 147 所(如圖 1 所示)，而大學日間部的預計總招生人數亦由 91 學年度的 96,847 至 96 學年度增加到 110,099 人。然而，根據內政部統計表示，民國 85 年之前，每年出生人口皆達到 30 萬人口的水準，自 86 年後出生人口趨於減緩，至 95 年出生人口僅 204,000 多人，經上述可知少子化趨勢並非短期現象。

根據教育部高教司統計處的統計，一般校院於 96 學年度第 2 學期因學業成績不及格遭退學學生共計 3,955 人，相較 96 學年度第 1 學期退學人數 3,519 人，增加 436 人。此外， 於 97 學年度第 2 學期截至，各公私立大學院校的休學人數統計高達 5,138 人，人數相當可觀。面臨招生人數逐年下降，休、退學人數亦逐年增加的雙面夾殺之下，大專校院如何輔導學生在校順利完成學業，成為急需探討的議題。

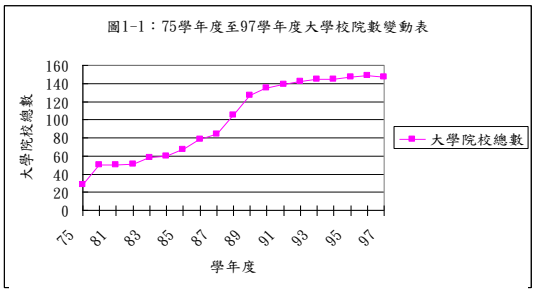

圖 1、75 至 97 學年度大學校院數變動

另一方面，某大學自民國 85 年成立至今，學生規模由大學日間部 77 人增加為 5,000 多人，大學日間部學系自兩個系擴增為 22 個系。然而，大學日間部學生報到人數由 95 學年度的 1,059 人下降到 97 學年度的 916 人，而大學日間部學生流失人數亦由 95 學年度第 1 學期的 45 人， 至 97 學年度第 1 學期增加為 97 人，故每年學生流失人數持續攀升令人堪憂。

以服務業而言，維繫舊客戶相較擴張企業競爭優勢，如擴大經營規模、市場占有率等方法具有更大的獲利價值(Colgate et al., 1996)；以利潤角度，維持既有客戶成本遠低於開發新客戶群的成本(Heskett et al., 1989)。由此可知，與客戶建立緊密且長久的關係，是企業創造利潤的重要關鍵，如何預防客戶流失是相當重要的(Keaveney, 1995)。雖學校經營不同於企業，隨著各大專院校轉型後，招生競爭日趨激烈，顧客關係管理(Customer Relationship Management, CRM)亦逐漸被運用在學校招生方面。以顧客關係管理觀點，中途離校生即客戶流失，若能運用顧客關係管理的方式經營與學生的關係，從整體來看對學校經營應有所助益 (Jaishankar et al., 2000)。

因學校人員與教師日常行政工作與教學活動已相當繁忙，如何協助學校及早發現潛在可能流失學生，以減少其流失率，是一大重要關鍵。由於學生人數眾多，察覺潛在可能流失學生相當不易，本研究運

用資料採礦技術分析學生歷史學籍資料，探索在何種情況下容易發生流失學生的可能，建立流失預測模型並提供相關建議，以降低學生流失的情形發生。

本研究給予學生流失的定義為：「已註冊繳費，擁有本校學籍資料之學生，因故無法完成學業(即休學、退學與轉學)，中途退出而未取得畢業證書的學生」。

2. 文獻探討

2.1.資料採礦

資料採礦(data mining)由 Fayyad (1991)首先提出，其目的為從龐大的原始資料中找出規則。Frawley 等人(1991)認為資料採礦是一大量自動化的過程，運用統計分析從大量的資料集中，挖掘出有用的、潛在與先前未知的特徵或資料趨勢。Kleissner (1998)表示，資料採礦是發現企業資料中所隱含的知識，以供組織決策者進行支援決策的分析過程。因此，資料採礦不外乎是發現有用型樣的過程，其過程為自動或半自動的方法，並為組織帶來經濟優勢 (Ian, H. W. and Eibe, F., 2005)。

藉由資料採礦技術協助，可增進組織企業對顧客需求與行為的瞭解，有助於企業提供客製化的服務，強化與顧客之間的連結、溝通與互動，亦即可發掘大量關於顧客特徵和購買模式有益於行銷的知識(Shaw et al., 2001)。大多數公司，運用資料採礦作為策略的基礎，協助其打敗競爭者、確認新顧客並降低成本 (Davis, 1999)。

一般常用的資料採礦技術主要可分為：分類(classification)、推估(estimation)、群集化(cluster)、關聯法則(association rule)、序列(sequential)與描述(description)等共六種，各類技術可依其特性適用於不同的領域，如金融保險、零售製造、醫療生技與教育等各行各業之中 (尹相志，2007)。

2.1.1.決策樹

決策樹為資料採礦技術建立分類模式最常見的方法之一，其針對給定資料利用歸納的方式產生樹狀結構的模式。決策樹的每一個節點為一個判斷式，判斷式針對一個變數去判斷輸入的資料大於或等於或小於某個數值，每一個節點因可將輸入的資料分成若干類。

此外，可透過修剪機制抑制決策樹的增長，除去無用的規則。其中，可分為事前(prepruning)與事後修剪(postpruning)，前者即訓練模型的同時，將該節點設為葉節點，使該節點停止生長；後者則將已產生的決策樹多餘的規則修剪去。決策樹主要優點在於利用樹狀結構表示具有規則與解釋力，可用文字來表達且易於瞭解，在轉化為資料庫語言，亦可讓落在特定類別的資料紀錄得以被搜尋。

2.1.2. 類神經網路

類神經網路(neural network)以模擬人類大腦神經細胞的的運作方式，由高度連結的處理單元(神經元)構成一組運算系統，透過不斷自我學習加以調整，使得輸入的

資訊經過神經元運算後可得預設的輸出結果。

如圖 2 為例，其中 X 是輸入變數值，W 為輸入變數的權重，輸入變數值(X)乘上相對應的權重植(W)等於外部輸入的神經脈衝(Y1)，神經脈衝必須大於門檻值，才能傳遞至神經元。當脈衝進入神經元後，神經元透過加總函數把所有的神經脈衝累加，透過轉換函數(activation Function)，產生新的神經脈衝向外傳遞。將神經元彼此連結，建立類神經網路架構，亦即一個神經元的輸出可成為下一個類神經網路的輸出脈衝。類神經網路優點不僅預測能力的準確度高，具有部分容錯的功能，且可應用於連續與類別變數型態的預測問題上。

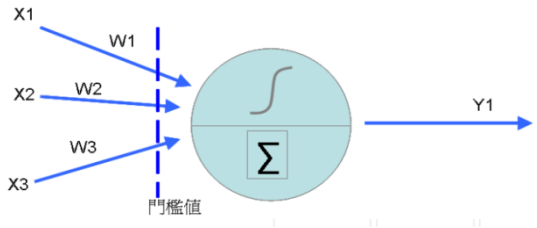

圖 2、神經元架構

2.2.相關研究

運用資料採礦技術在學生流失相關文獻如下：

王建華（2004）運用決策樹分析缺曠課情形及學業成績，針對技職五專學生行為特徵並比較人格測驗之分析結果，以探討中途離校生之出缺席狀況及學業成績關聯，其研究發現中途離校生缺課情形嚴重且成績較差。

趙瑞麟(2007)利用決策樹 C5.0、CART 與類神經網路等技術建立學生流失預測模型，針對進修院校二技二專學生，探討科系、性別、年紀、原畢業學校分類、居住地區、學業成績等級與操行成績等級等變項對學生流失的影響，研究結果發現決策樹演算法表現較佳、學生流失主因與主要分類變項等關鍵。

3. 研究方法

3.1 研究對象

本研究以某大學 94 至 96 學年度入學之大學日間部學生學籍歷史資料為研究對象，依據校務行政系統的學生資料，蓋括學生性別、科系、入學方式、學期成績、歷年學業平均成績、歷年修課學分數、歷年實得學分數、居住地區等資料欄位為主要的研究變數。因資料獲取問題，僅限於資訊管理、資訊工程以及建築與景觀等三個學系的學生資料，而取樣時間為 97 學年度第 2 學期，故其就學狀態為 97 學年度第 2 學期的狀態。

3.2.增益圖

為瞭解所建立預測模型的效益，本研究採用一種普遍使用的資料採礦模型評估圖－增益圖(lift chart)，以協助評估預測模型的優劣。由於，研究取樣可能面臨屬性分佈比例不勻的問題，若使用百本分比表示方式，將無法展現某一屬性於各自子群中所占的比例。如本研究樣本性別比

例分佈不均而言，假設採用其流失之百分比方式表示，則原本所占比例較高之性別，其流失人數換算比例會相對較高。為此改用增益值計算的方式，其值則表示傾向流失的機率，比較基準為相對的增益。

增益值即「事件在子群的發生機率除以該事件在原始母體之發生機率」，增益值大於 1 表示事件在子群的發生機率大於該事件在原始母體之發生機率。因此，該條件下被視為傾向發生。如表 1 之男學生退學為例。

增益值＝事件在子群之發生機率/該事件在原始母體之發生機率

＝（男生退學人數/總退學人數）/（男生總人數/總體人數）

＝ (64/77)/(522/704) =1.12

表 1、就學狀態一覽表

預測模型	實際已流失人數	實際未流失人數
已流失人數	43	74
未流失人數	13	80
正確率	58.87%	
錯誤率	41.42%	

3.3.分類矩陣與 3R 指標

資料採礦模型評估指標上，可利用分類矩陣(confusion matrix)進行效益檢測。分類矩陣主要為檢視錯誤的分佈狀態，即建立模型後，以測試組資料進行測試，驗證模型預測結果的分佈狀態，計算方式如表 2 為例，如下所示：

表 2、預測模型之分類矩陣

就學狀態 / 性別	在學	校內	休學	退學	轉學	總人數
女	132	1	5	13	31	182
男	381	2	14	64	61	522
總人數	513	3	19	77	92	704

預測正確率＝(預測已流失且實際已流失人數＋預測未流失且實際未流失)/總體人數
＝(43+80)/(43+13+74+80)＝58.57%
預測錯誤率＝(預測已流失且實際未流失人數＋預測未流失且實際已流失)/總體人數
＝(74+13)/(43+13+74+80)＝41.42%

我們可藉由分類矩陣獲得三種資訊判斷模型之成效，稱為 3R 指標。分別代表回應率（re sponse rate），表示預測名單中找出多少稀有事件；其次，反查（recall）為預測稀有事件佔總體稀有事件的比例；再次，間距縮減（range reduce）透過資料採礦模型搜尋稀有事件，顯示計算名單縮小的範圍。計算方式，如表 2 為例：

回應率＝預測會流失實際已流失/所有預測已流失的學生＝43/（43+74）＝36.75%
總體回應率＝所有實際已流失/全體學生＝（43+13）/（43+13+74+80）＝26.67%
反查＝預測會流失實際已流失/所有實際已流失的學生＝43/（43+13）＝76.79%
間距縮減＝所有預測已流失的學生/全體學生＝（43+74）/（43+13+74+80）＝55.71%

經上述計算，回應率 36.75%與總體回應率相較下，可發現原始總體回應率為 26.67%，運用預測模型提升 1.37 倍為 36.75%。一般而言，反查數值越高越好，但是反查與回應率兩者會產生互斥的情況，故期望回應率愈高，則必須把回應率的門檻提高，以排除回應率低的名單，相對的會造成反查的降低。針對間距縮減而言，運用此預測模型可使學生名單縮減至原本的 55.71%，涵蓋總體 76.79%會流失的學生（反查），使得回應率提升原先的 1.37 倍。

4. 資料分析

4.1 分析工具

本研究採用 SQL Server 2005 為資料採礦工具，其擁有決策樹、群集演算法、類神經網路、線性迴歸與關聯規則等多種演算法，並具有豐富的視覺化圖形，不僅使分析者易於瞭解模型規則與內容，更可透過互動的機制，觀察潛在模型的趨勢走向。此外，還提供分類矩陣、增益圖、利潤圖與散布圖等資料模型評估工具，且利用整合式商業智慧 (business intelligence, BI) 工具，提供資料管理功能，可視為一個功能完備的資料庫平台。

4.2.樣本特性分析

本研究樣本共 704 筆，就學狀態為休學、退學與轉學者，合計高達 26.7%（表 3）。在研究樣本性別分佈上，男女比例明顯不均，以男生占大多數（表 4）。以流失比例而言，女生以轉學最多，其次為退學；而男生則為轉學與退學最多（表 5）。為深入瞭解不同性別流失趨勢，採用增益值計算其傾向發生機率，可知女生轉學增益值較高，而退學則為男生增益值較高（表 6）。另外，針對系別就學狀態依據增益值進行分析（表 7），就學狀態為休學以建築與景觀系增益值最高，其次，退學與轉學皆以資訊工程系最高。

表 3、學生就學狀態表

就學狀態	人次數	百分比
在學	513	72.9
校內轉系	3	0.4
休學	19	2.7
退學	77	10.9
轉學	92	13.1
總和	704	100.0

表 4、學生性別比例表

	人數	百分比
女生	182	26
男生	522	74
總和	704	100.0

表 5、就學狀態性別分佈一覽表

就學狀態	女生	男生
在學	72	73
校內轉系	1	0
休學	3	3
退學	7	12
轉學	17	12

單位：百分比

表 6、不同性別之就學狀態增益值表

就學狀態 / 性別	在學	在校轉系	休學	退學	轉學	總和
女生人數	132	1	5	13	31	182
增益值	1.00	1.29	1.02	0.65	**1.3**	
男生人數	381	2	14	64	61	522
增益值	1.00	0.90	0.99	**1.12**	0.89	
總和	513	3	19	77	92	704

表 7、學系就學狀態一覽表

就學狀態 / 學系	在學	在校轉系	休學	退學	轉學	總和
建築與景觀系	96	1	5	9	15	126
增益值	**1.05**	1.86	1.47	0.65	0.91	17.9%
資訊工程系	168	0	3	31	33	235
增益值	0.98	0	0.47	**1.21**	**1.07**	33.4%
資訊管理系	249	2	11	37	44	343
增益值	**1.00**	1.37	1.19	**0.99**	**0.98**	48.7%
總和	513	3	19	77	92	704

表 8、居住縣市增益值

狀態 / 縣市	已流失	未流失	已流失增益值	未流失增益值
蒙古	0	1	0.00%	136.43%
[illegible]	5	15	93.62%	102.33%
嘉義縣	8	[illegible]	[illegible]99%	97.45%
彰化縣	7	22	90.39%	[illegible]
台北市	19	38	***124.82%***	90.96%
台北縣	18	41	***114.24%***	94.81%
[illegible]	4	22	57.61%	115.44%
花蓮市	[illegible]	26	**111.33%**	95.87%
花蓮縣	2	[illegible]	[illegible]	83.38%
台南市	9	17	129.62%	[illegible]
台南縣	14	32	***113.97%***	94.91%
[illegible]	[illegible]	1	0.00%	136.43%
澎湖縣	2	3	[illegible]	[illegible]
總計	188	516	100.00%	100.00%

由學生居住縣市進行分析可知，台北市、台北縣、雲林縣、高雄市與屏東縣為學生流失最多的縣市，經計算增益值可知台北縣市與台南縣已流失增益值顯著（表 8），顯示其傾向已流失。

學生的入學年度學期以第一學期為主，另行針對異動學年學期進行統計，則可顯示學生資料產生異動筆數，如休學、退學與轉學等。以表 9 之 94 學年第 1 學期為例，其資料異動於 95 學年度第一學期筆數為 33，明顯較其他學年度學期異動筆數居多，且同樣的情形皆發生於 95、96 學年度。此外，由圖 3 可發現資料異動的時間在每學年第一學期達到高峰，94 學年第 1 學期入學的學生，資料異動在 95 學年第 1 學期達到高峰，往後 95、96 學年度亦皆發生。經上述，可顯示新生入學一年後，升二年級之際產生資料異動較為頻繁。

表 9、入學學年學期與異動學年學期交叉表

在學人數 \ 異動學年學期		學年度	94		95		96		97		總次數
		學期	1	2	1	2	1	2	1	2	
入學學年	94學年第1學期	152人	4	5	**33**	5	7	0	1	0	207
	95學年第1學期	157人	0	0	3	11	**32**	9	17	2	231
	96學年第1學期	188人	0	0	0	0	4	12	**46**	13	263
總和		497人	4	5	36	16	43	21	65	15	701

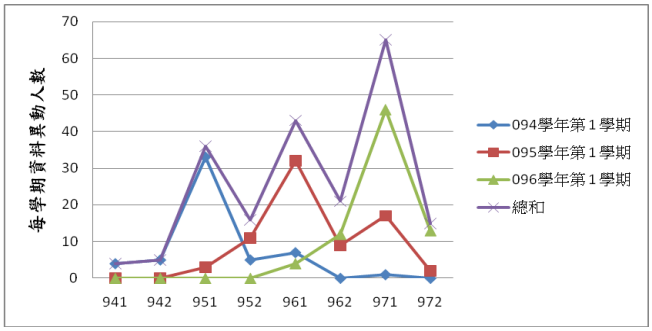

圖 3、資料異動時間趨勢圖

透過學生入學方式分析顯示，以考試分發比例為首，其次為個人申請，再次為轉學考。由表 10 可知，就學狀態為休學者，其入學方式為轉學考的增益值為大於考試分發者，顯示轉學生傾向休學程度的極高。此外，就學狀態為退學者，其入學方式以考試分發與轉學考的增益值較高，顯示兩者都有退學的傾向，而就學狀態為轉學者，以考試分發及個人申請較高，表示兩者皆有轉學的傾向。

表 10、就學狀態與入學方式交叉表

就學狀態 / 入學方式	休學	在學	校內	退學	轉學	總和
四技申請	0	1	0	0	0	1
外國學生	0	1	0	0	0	1
申請入學	0	2	0	0	2	4
考試分發	8	338	3	50	65	464
增益值	0.64	**1.00**		**0.99**	1.07	
身心障礙	0	3	0	1	0	4
個人申請	2	94	0	5	17	118
增益值		**1.09**			1.10	
進學班甄	2	4	0	0	0	6
學校推薦	0	16	0	2	1	19
增益值		1.16				
轉學考	7	54	0	19	7	87
增益值	**2.98**	0.85		**2.00**	0.62	
總和	19	513	3	77	92	704

4.3.決策樹分析

依據本研究對學生流失的定義，將學生就學狀態為在學與校內轉系歸為「未流失」，共計 514 筆；休學、退學與轉學列為「已流失」，共計 187 筆。「已流失」與「未流失」之比例約為 1：2.75。透過資料前處理，將原始資料進行整合、清除與轉換後，挑選出進行決策樹分析所需的輸入欄位和預測欄位(表 11)。之後，將資料分為訓練組及測試組，使用訓練組資料進行模型的建立，待模型建立後使用測試組資料進行驗證與評估。

表 11、決策樹分析輸入欄位

輸入欄位	學號、性別、科系、入學方式、居住地區、歷年學業平均成績、平均每學期不及格學分數
預測欄位	流失與否

在資料採礦過程中，因預測欄位值分佈不均，故產生稀有事件的問題，即本研究所預測學生流失與否欄位的「已流失」。為此，本研究採用誤差抽樣(error sampling)方法，為處理稀有事件最常採用的一種技

巧，不按照原先值的分配等比例抽樣，而將稀有事件透過抽樣的方式將其比重提高。

進行資料採礦前，本研究事先將資料分為訓練組及測試組，其中測試組占總資料的 30%，即 210 筆資料(已流失為 56 筆，未流失為 154 筆，其已流失與未流失之比例為 1：2.75)。此後，將訓練組資料中所含已流失資料全部抽出後，剩餘未流失資料分別抽出 131 筆、262 筆及 360 筆資料，分別構成含已流失與未流失比例為 1：1、1：2 與 1：2.75（1：2.75 即扣除測試組資料後的全部資料）等三組訓練組資料。

分別依 1：1、1：2 及 1：2.75 訓練組資料建立決策樹模型，於各個模型之中，分別調整其演算法參數並測試篩選後，刪除模型無差異之組合，以 COMPLEXITY _PENALTY 與 MINIMUM_SUPPORT 三組不同參數分別建立模型，如表 12 所示。其中，COMPLEXITY _PENALTY 表示複雜性懲處，其值愈接近 1，決策樹的成長會受到較多的抑制，而產生分岔較少樹狀規則；MINIMUM_SUPPORT 則為每個規則節點所需最小案例數。

表 12、篩選之三組參數

參數項目	第 1 組	第 2 組	第 3 組
COMPLEXITY_PENALTY	0.5	0.1	0.05
MINIMUM_SUPPORT	10	5	5

4.3.1 評估預測模型

模型建立後，分別代入測試組資料分別進行測試，並利用分類矩陣、3R 指標與增益圖評估模型最佳模型。由於，分類矩陣與3R指標分析結果難以判定模型優劣，故本研究以增益圖作為衡量預測模型的依據。以增意圖之圖 4 為例，橫軸與縱軸皆由百分比所構成，橫軸百分比代表資料採礦模型根據機率從高至低排序後的名單占總體百分比；縱軸則是此名單中稀有事件的人數占總體稀有事件人數的百分比。

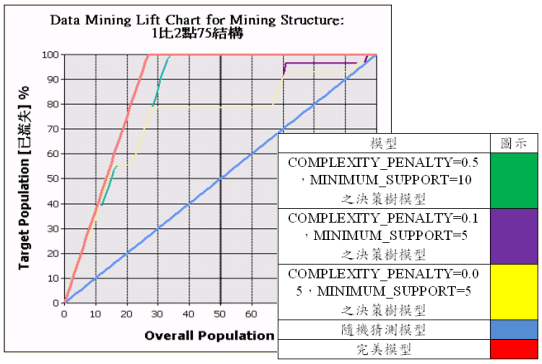

圖 4、訓練組資料 1：2.75 之決策樹模型增益圖

圖中有一藍色呈現 45 度的斜線，代表隨機的狀態。以增益圖的定義，代表當篩選一半的名單去檢視學生流失狀況時，將會包含全體名單一半的學生流失數量(即隨機亂猜)，而完美模型的紅線則顯示極佳的預測結果。正常的模型增益圖必定要比 45 度線向第二象限彎曲，增益圖愈向上彎曲，顯示模型效果愈好，且所有模型的曲線必須介於隨

機猜測與完美模型之間，其愈接近完美模型愈好(尹相志，2007)。

圖中粗線之橫軸表示整體母體擴展，即總人數百分比。以圖 4 為例，黃線為訓練組資料 1：2.75，COMPLEXITY_PENALTY=0.05，MINIMUM_SUPPORT=5 之決策樹模型，對應在橫軸 50%對應到縱軸的值為約 79%，代表資料採礦根據流失機率由最高至最低排序的前 50%名單中抓到會流失的學生，占總體會流失的學生約 79%。相對的，剩下的 50%學生中，僅占總體會流失學生約 21%。如下表 13 不同比例訓練組資料模型增益圖比較中，分數為模型曲線下面積與完美曲線下面積的比值，分數愈接近 1，則表示模型預測力愈高。

表 13、不同比例訓練組資料模型增益圖比較（以母體擴展 50%為比較基準）

序列、模型		分數	目標母體%	預測機率%
隨機猜測模型			50.00	
理想模型			100.00	
訓練組 1：1	COMPLEXITY_PENALTY=0.5，MINIMUM_SUPPORT=10	0.90	76.79	49.85
	COMPLEXITY_PENALTY=0.1，MINIMUM_SUPPORT=5	0.83	73.21	48.21
	COMPLEXITY_PENALTY=0.05，MINIMUM_SUPPORT=5	0.81	71.43	34.37
訓練組 1：2	COMPLEXITY_PENALTY=0.5，MINIMUM_SUPPORT=10	0.89	82.14	29.19
	COMPLEXITY_PENALTY=0.1，MINIMUM_SUPPORT=5	0.85	78.57	27.13
	COMPLEXITY_PENALTY=0.05，MINIMUM_SUPPORT=5	085	78.57	27.13
訓練組 1：2.75	COMPLEXITY_PENALTY=0.5，MINIMUM_SUPPORT=10	**0.96**	100	16.29
	COMPLEXITY_PENALTY=0.1，MINIMUM_SUPPORT=5	0.86	78.57	20.71
	COMPLEXITY_PENALTY=0.05，MINIMUM_SUPPORT=5	0.85	78.57	20.71

因分數指的是整體的面積，而目標母體值則僅部分母體擴展，故以分數高低做為模型優劣的衡量，本研究以母體擴展 50%作為各模型比較優劣的基準。如表 13 顯示訓練組資料 1：2.75，參數為 COMPLEXITY_ PENALTY =0.5，MINIMUM_SUPPORT=10 之決策樹模型之分數 0.96 為最高，故判斷此為最佳預測模型。

4.3.2 驗證預測模型

由訓練組資料 1：2.75，COMPLEXITY_PENALTY=0.5，MINIMUM_SUPPORT=10 之決策樹模型，所建立的決策樹模型如圖 5 所示。其中，Total AV#為歷年學業平均成績，可顯示學生學習成績為學生流失與否之重要預測因素。

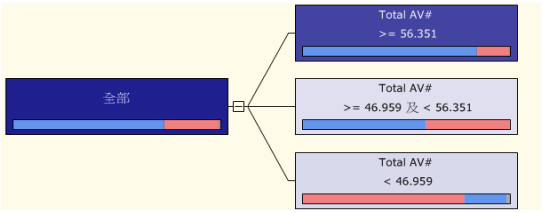

圖 5、訓練組資料 1：2.75，COMPLEXITY_PENALTY=0.5，MINIMUM_SUPPORT=10 之決策樹模型

由決策樹模型獲得規則與其所代表意涵如下表 14 至 16 所示。以表 14 決策樹規則一為例，若以歷年學業平均成績≧56.351 去預測學生流失與否，則有 16.29%機率為已流失，83.44%的機率為未流失。

表 14、決策樹規則一

編號	節點路徑	值	案例	機率%
規則一	歷年學業平均成≧56.351	已流失	62	16.29
		未流失	322	83.44
		遺漏	0	0.28

編號	節點路徑	值	案例	機率%
規則二	46.959≦歷年學業平均成績＜56.351	已流失	17	40.34
		未流失	25	59.03
		遺漏	0	0.62

表 15、決策樹規則二

表 16、決策樹規則三

編號	節點路徑	值	案例	機率%
規則三	歷年學業平均成績 46.959	已流失	52	77.29
		未流失	13	20.77
		遺漏	0	1.93

經上述三條規則之後，進行原始資料檢視。首先，歷年學業平均成績≧56.351 的學生就學狀態，已流失之就學狀態以轉學居多（圖 6），顯示規則一歷年學業平均成績≧56.351，其學生多為自願性流失。其次，規則二，由於資料筆數較少，檢視其原始資料並未發現其他特殊狀況。再次，規則三之歷年學業平均成績＜46.959 原始資料中則發現，其已流失之就學狀態以退學為主，為非自願性流失，如圖 7 所示。

而訓練組資料 1：2.75，COMPLEXITY_PENALTY=0.5，MINIMUM_SUPPORT=10 所得之決策樹模型經測試組資料測試後所得分類矩陣如表 17 所示，其正確率可高達 81.90%，錯誤率為 18.09%。經上述，可顯示決策樹預測模型的預測能力與適用性皆有相當不錯的成效。

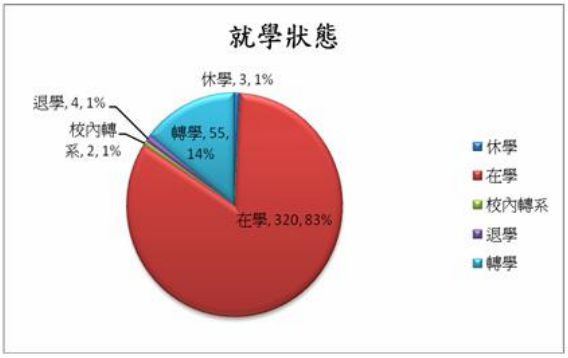

圖 6、規則一之就學狀態圖

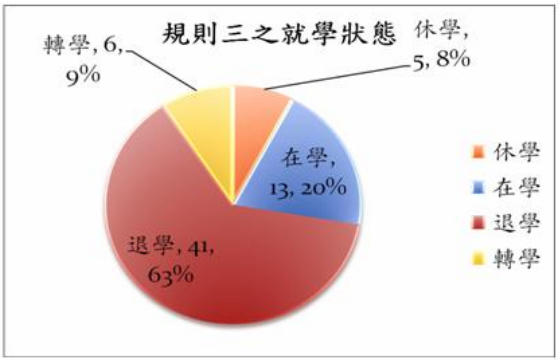

圖 7、規則三之就學狀態

表 17、決策樹預測模型之分類矩陣

模型訓練組資料 1：2.75	預測	實際已流失	實際未流失
COMPLEXITY_PENALTY=0.5，MINIMUM_SUPPORT=10	已流失	22	4
	未流失	34	150
	正確總數（百分比）	172（81.90%）	
	錯誤總數（百分比）	37（18.09%）	

4.4 類神經網路分析

神經網路模型僅能處理連續數值，若輸入變數為類別變數時，會先將輸入變數轉換為虛擬變數（選項轉為 1 或 0 的編碼方式）。類神經網路透過修正權重的模式來產生學習成果，演算法本身不具有變數篩選功能，即所有的變數皆計算出屬於自己的權重。因此，在 SQL Server 2005 類神經網路演算法的檢視器中，主要呈現資料內部的機率分佈架構。

4.4.1 評估與驗證模型

根據前一節所設計的訓練組及測試組資料，分別進行類神經網路模型之製作。在調整各個模型之內建參數之後，發現調整參數後差異變化不大，故皆不予調整，分別代入測試組資料。透過分類矩陣與 3R 指標評估模型各有優劣，故本研究則用增益圖作為比較基準。由表 18 可知，類神經網路 1：2.75 模型之分數為 0.78 與其他兩

者之差距甚微，因而採用決策樹獲得較佳模型之訓練組資料 1：2. 75，進行類神經網路分析，其模型如圖 8 所示。

表 18、類神經網路模型增益圖比較表

模型	分數	目標母體%	預測機率%
隨機猜測模型		50.00	
理想模型		100.00	
1：1 模型	0.77	73.21	45.59
1：2 模型	0.75	69.64	23.03
1：2.75 模型	0.78	73.21	16.67

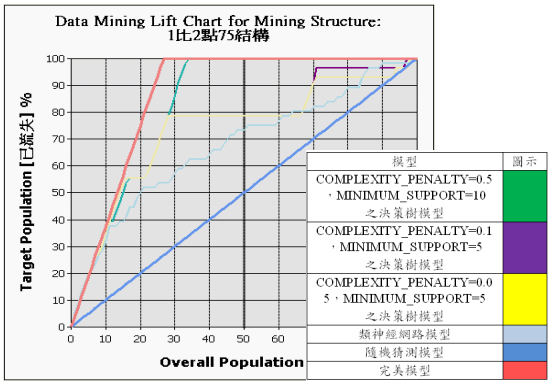

圖 8、訓練資料 1：2.75 類神經網路模型增益圖

下圖 9 為訓練組全部資料的所有變數選項組合，依照顯著性由高至低排列，即 Attribute 欄位依照其顯著性由高至低排列，顯著性為系統自動計算出之分數，而 Favors 則依據自動計算增益值，決定為已流失或未流失。其中，Total AV#為歷年學業平均成績；Lose AV#為平均每學期不及格學分數，正常應為正數，若為負數者，其因於學生進行學分抵免，抵免學分列入每學期的實得學分數，故造成當學期的實得學分數大於其修課學分數的情形發生。經圖 9 顯示，以歷年學業平均成績 5.305-50.526 之顯著性為最高，其傾向已流失；其次，為居住地區為離島，傾向為已流失；再次，是歷年學業平均成績 76.759-93.918，其傾向為未流失，依此列舉之。


Variables:

Attribute	Value	Favors 已流失	Favors 未流失
Total AV#	5.305 - 50.526		
Area	離島		
Total AV#	76.759 - 93.918		
Application	其他		
Lose AV#	7.036 - 20.981		
Area	外國		
Lose AV#	-14.999 - -1.054		
Area	東部		
Application	學校推薦		
Application	個人申請		
Total AV#	50.526 - 63.643		
Total AV#	63.643 - 76.759		
Sex	女		
DEPART#	資訊工程學系		
DEPART#	資訊管理學系		
Lose AV#	2.991 - 7.036		
Area	北部		
Lose AV#	-1.054 - 2.991		
Application	轉學考		
DEPART#	建築與景觀學系		
Application	考試分發		
Area	南部		
Sex	男		
Area	中部		

圖 9、類神經網路輸出檢視

由於檢視表乃顯示全部變數之顯著性及已流失與未流失傾向，細部資料還需進一步進行相關之選取。表 19 為擷取顯著性較高的前 3 項變數。其中，可知第一列歷年學業平均成績為 5.305-50.526 的變數，其 Score 最高，即顯著性最高。由系統依發生機率協助我們自動算出增益值決定其傾向為已流失或未流失，依此類推，可看出各變數值之顯著性高低及其已流失與未流失之傾向。

表 19、神經網路輸出變數喜好值

Attribute	Value	Favors 傾向已流失	Favors 傾向未流失
Total AV#	5.305~50.526	Score: 100 Probability of Value1: 72.06% Probability of Value2: 27.65% Lift for Value1: 3.97 Lift for Value2: 0.34	
Area	離島	Score: 64.11 Probability of Value1: 51.73% Probability of Value2: 47.98% Lift for Value1: 2.85 Lift for Value2: 0.59	
Total AV#	76.759~93.918		Score: 59.9 Probability of Value1:4.85% Probability of Value2:94.87% Lift for Value1: 0.27 Lift for Value2: 1.16

在類神經網路建立預測模型時，可由使用者自行輸入過濾條件來限縮母體的範圍，使用者可以透過輸入區域的屬性(Attribute)選擇所要的變數。由於，輸入的過濾條件僅限於察看原始模型之細部資料，與原模型的差異不大，僅於顯著性高低的不同，仍可提供我們作參考。以圖 10 為例，選擇學生入學方式為輸入條件，可發現每一種入學身份的流失與否，成績為最重要的因素，藉此可觀察變數選項中可預測變數選項的分佈機率。

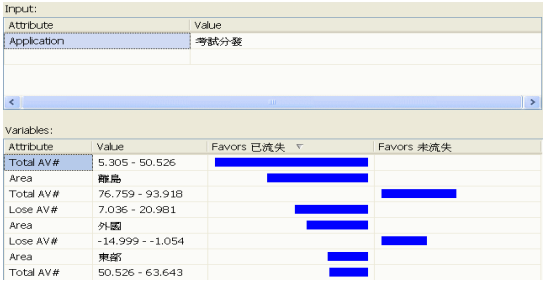

圖 10、類神經網路檢視輸入

我們輸入過濾條件為歷年學業平均成績在 76.759-93.918 來進行細部檢視，如圖 11 所示。歷年學業平均成績在 76.759-93.918 者，傾向於未流失，以 Lose AV#平均每學期不及格學分數-14.999 --1.054 為顯著性最高，其次為入學身份為個人申請、科系為資訊工程系，依此類推。

此外，細部資料顯示居住地區為離島或外國，以及入學身份為其他者，傾向已流失。檢視原始資料後發現，居住地區為外國者僅 1 人，且為未流失。由於居住於離島或東部的資料亦算少數，故發現資料筆數較少。此外，藉由增益值計算之後，容易造成已流失或未流失的偏好顯示異常，且屬性顯著性亦提高許多。由此，可知透過類神經網路檢視器僅能協助我們瞭解變數之間的交互作用，並非權重高低，故我們僅能使用此檢視器瞭解變數重要性的程度。

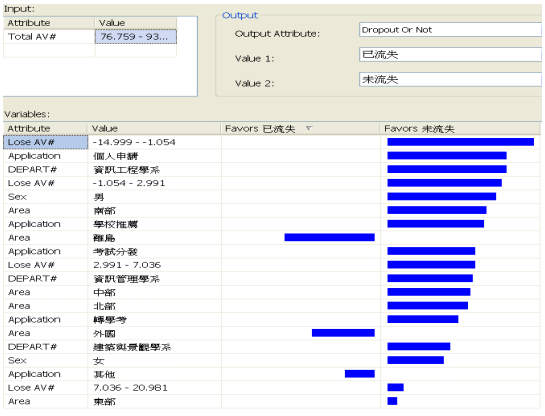

圖 11、類神經網路模型檢視

5. 結論與未來工作

本研究運用學生歷史學籍資料進行分析，試圖發掘流失學生之表徵，以預測可能流失之學生，建立預測模型並提供可行之建議作為參考，以協助教師輔導學生有所依據並找出重點輔導對象，以減少學生流失的情形。

5.1.結論

本研究實驗結果發現，如下述五項：

(1)學生一年級升二年級時的流失率較其他年級高。因學習困難或適應不佳等因素，為此新生入學時，應加強對系所、課程與校園等介紹，使新生能短時間熟悉環境並融入團體生活，並配合舉辦課外活動以增進其認同及歸屬感，進而減少其流失。

(2) 性別比例而言，以退學方面，男生增益值較高，而轉學方面則是女生增益值較高。對潛在流失學生進行輔導時，可引導學生至本系不同專業領域的興趣方向，或提供轉系方案，以學校觀點認為校內轉系並未流失學生。

(3) 以學系而言，就學狀態為休學時，建築與景觀系增益值高，其次為資訊管理系，再次為資訊工程系；為退學時，以資訊工程系增益值較高,其次為資訊管理系，再次是建築與景觀系；為轉學時，以資訊工程系增益值較高，其次為資訊管理系，再次是建築與景觀系。

(4) 入學方式為轉學考者，其休學及退學的增益值較高。起因於轉學生無法融入新群體與課業學習不佳，易於成為孤立的個體以致選擇流失。因此，可藉由建立學長姐制與相關輔導措施，克服人際與課業問題，或可減少其流失。

(5) 研究樣本以居住縣市為台北縣市、雲林縣、高雄市與屏東縣等縣市為最多，建議可針對未流失增益值較高的縣市加強招生及宣傳。

由上述可知，可針對自願性流失之學生，校方與教師可加強其生活輔導，增進學生對系上的歸屬感，而非自願性流失之學生，則需同時進行生活與課業輔導。本研究實驗發現決策樹模型，以成績為主要分類依據，還可參考平均不及格學分數及入學方式是否為轉學考等兩項因素；則類神經網路分析以成績為主要因素，而居住地區也可能是影響為流失與否的潛在因素之一。

此外，針對學生的已流失與未流失分類仍以歷年學業平均成績為主要分類依據，顯示學生的學習情況是決定學生流失與否

之重要原因。整體而言，決策樹較類神經網路模型的預測能力高（圖 4、8 所示），透過決策樹與類神經網路兩種方法皆發現學生學習成績為分類的重要因素，可藉由成績因素觀察學生流失與否的可能發生。

5.2. 未來工作

本研究以某大學 94 至 96 學年度入學之大學日間部學生學籍歷史資料為研究對象，因資料獲取問題，僅限於資訊管理、資訊工程以及建築與景觀等三個學系的學生資料。未來研究可增加分析樣本的數量，並可納入學生流失其他相關的因素，如學校名聲與環境、系所課程、家庭、經濟與個人等影響因素，將可發現更多影響學生流失之潛在因素。另外，本研究取得資料欄位時，未將休學、退學與轉學先後順序欄位分開，以致無法了解其個別的流失因素是否具有異同性質，可進一步進行資料庫修正，將其納入實驗範疇之一。此外，利用資料採礦技術建立學生流失之預測模型，亦可適用於顧客流失等其他相關議題上，其預測模型可廣泛加以使用。

參考文獻

[1] 尹相志(2007),《SQL Server 2005 Data Mining 資料採礦與 Office 2007 資料採礦增益集》，台北：悅知文化。

[2] 王建華（2004），《資料挖掘技術再技職院校中途離校生輔導之應用-以醒吾技術學院為例》，碩士論文，國防管理學院國防資訊研究所。

[3] 趙瑞麟（2007），《以資料探勘技術降低學生流失之研究 -以某技術學院附設進修院校為例》，碩士論文，雲林科技大學資訊管理系碩士班。

[4] 內政部戶政司人口統計資料＜http://sowf.moi.gov.tw/stat/month/m1-02.xls/＞

[5] 教育部高教司統計處＜http://www.edu.tw/files/site_content/B0013/overview03.xls/＞

[6] Colgate, M.(1996), “The use of personal bankers in New Zealand: an exploratory study”, New Zealand Journal of Business, 18(2), 103-122.

[7] Davids, M.(1999), “How to avoid the 10 Biggest Mistake in CRM”, Journal of Business Strategy, Vol. 4, 22-26.

[8] Fayyad, U. M. and Irani, K. B.(1991), “Machine learning algorithm (GID3*) for automated knowledge acquisition: Improvements and extensions,” General Motors Research Report CS-634, Warren, MI:CM research labs.

[9] Frawley, W. J., G. Piatetsky-Shapiro, and Matheus. C. J.(1992), “Knowledge Discovery in Databases: An Overview”, AI Magazine, 213-228.

[10] Heskett, J. L., Sasser Jr., E., and Hart, C. W.(1989), *Service breakthrough*, New York: The Free Press.

[11] Jaishankar, G., Mark, J. A. and Kristy, E. R.(2000), “Understanding of Customer Base of Service Providers:An Examination of Differences between Switchers and Stayers”, Journal of Marketing, Vol. 64, 65-87.

[12] Ian, H. W. and Eibe, F.(2005), *Data Mining: Practical machine learning tools and techniques*, San Francisco：Morgan Kaufmann.

[13] Keaveney, S. M. (1995), “Customer Switching in Service Industries: An Exploratory Study”,Journal of Marketing, Vol. 59, 71-82.

[14] Kleissner, C. (1998), “Data mining for the enterprise”, Proceedings of the Thirty-First Hawaii International Conference, pp.295-304.

[15] Shaw, M. J., Subramaniam, C., Tan, G. W., and Welge, M. E.(2001), “Knowledge management and data mining for marketing”, Decision Support Systems, 127-137.