以資料探勘技術建置適性化的課業成績表現關聯模組之研究

A Study on the Adaptive Association Rules of the Learning Performance of Courses by Using Data Mining Technology

張元勳[1]，陳瑞堂[2]，李建億[3*]
[1]國立台南大學數位學習科技學系
[2]國立台南大學數位學習科技學系
[3]國立台南大學數位學習科技學系
[*] leeci@mail.nutn.edu.tw

【摘要】現今大學的成績預測大多以學生的背景或是學習歷程作為判斷標準，背景與課程表現之間較無關聯性，而學習歷程則在修課過程中才有辦法取得。因此，如何在選課前得知未來的成績表現與課程間的關聯性以作為學生選修此門課程的決策參考，將是值得探究的議題。本研究使用某綜合大學十個學年度的學生為研究對象，以其所有必選修課程之學生修課資料做為研究資料。本究結果顯示 (1)學生以往課程成績與未來之課程成績具關聯性；(2)群集分析可提升關聯規則之預測準確率；(3)群集數量會影響關聯規則之預測準確度。
【關鍵字】 學習歷程；教育資料探勘；學習成效；

Abstract: *Most of the current university performance prediction systems use students' background information or learning history as the criteria for judgment. Background information has little correlation with course performance, and learning history can only be obtained during the course. Known before. Therefore, how to know the correlation between future academic performance and courses before selecting courses so as to serve as a reference for students to make decisions on taking this course will be a topic worth exploring. This study uses the students who entered a comprehensive university in ten academic years as the research subjects, and the student course records of all required courses as the research data. The results of this study show that (1) students' past course grades are correlated with their future course grades; (2) cluster analysis can improve the prediction accuracy of association rules; and (3) the number of clusters affects the prediction accuracy of association rules.*
Keywords: Learning Portfolios, Educational Data Mining, learning effectiveness

1. 前言

近年來許多研究學者與大學機構合作，在教育資料探勘領域針對學生學習成效方面取得了豐碩的成果，諸如探討針對學生在網路上的發言以預測學習成效(Smirnov, 2020)、學習困難學生的學習特徵(Botelho et al., 2019)、根據 MOOC 影片觀看行為預測學習表現(曾敬翔, 2020) 、使用行為序列分析學習行為(余禎祥, 2019)或學生學習策略、通過向學生回饋學習報告掌握班級整體學習情況進行個性化干預學習(吳靜薇, 2014)等等。過去的研究聚焦在：預測學生是否會通過課程(胡詠翔, 2019)、從學生背景找出學習規則(陳嘉鴻, 2014)、從作業提交時間檢測風險學生、找出學習困難學生的學習行為之特徵(Botelho et al., 2019)等等。雖然以上研究針對存在學習風險的學生根據預測結果進行學習預警，但因學習不主動、學習能力相較其他學生存在困難的學生外，努力學習但不見成效或是不常主動學習最後只是低分通過的學生，或是排名在中段希望想再往上進步的學生不僅是授課老師所關注，也應該給予這些學生適時的鼓勵或提醒。此外，如何將學生依照其不同特徵，區分為較符合真實狀態的若干群集，再加搭配其建置合適的課業成績表現關聯模組，也將是本研究探討的重要議題。

2. 文獻探討

推薦系統乃透過蒐集使用者的使用數據，將使用者所需之資訊從巨量且雜亂的資訊中過濾出來，提升使用者的決策效能。隨著資訊暴漲，推薦系統也被運用於各個領域，許多學者也針對推薦系統運用於不同領域進行研究，從推薦系統擁有如此多元化的運用領域，可得知推薦系統在這個資訊量不斷增長的時代裡的重要性(Roy & Dutta, 2022)。目前學生在選擇課程時，僅能憑藉自身的興趣與前人的經驗做為選擇，缺少其他具有科學性、數據化的資訊做為參考，因此本研究欲對教務資料進行分析，再將產生之模型做為推薦系統之推薦資料，期望能透過此推薦系統學生使學生選擇適合自身的課程，進而提升學生的課程滿意度與學習成效。

關聯規則(Association Rule)的定義為：假設 I={i1,i2,...,im}是所有交易項目的集合。資料庫 D={t1,t2,…,tn}是所有交易紀錄 t 的集合，產生規則形式為 X→Y，其中 X ⊆ I, Y ⊆ I 且 X∩Y = ø。每條關聯法則皆需有支持度(support)與信賴度(confidence)此兩個參數是評斷一條法則具備價值的重要依據。一條具有意義的關聯法，需檢查其提升值(Lift)是否大於 1 (Han et al., 2022)。此外，本研究將使用群集分析(Cluster analysis)，它是一個多變量分析程序，根據樣本的資料特徵，將所有樣本以分組的方式區分，使其形成相似的若干群集，以進行樣本間的分群過程，由於方法是透過計算資料間的距離，若資料特徵為類別性的維度，則需要透過其他方法來量測樣本的相似度。群集分析技術有許多類型，本研究由於樣本數較少，且需要控管群集數量並解釋分群規則，因此本研究選擇最為普遍的 K-means 做為本研究的群集分析演算法(Aljarbouh et al.,2022).。

3. 研究方法

3.1. 研究對象與資料來源

本研究使用之樣本資料為某大學 10 個學年度某系的所有必選修課程之學生修課資料，所使用到的樣本資料欄位分別為學生學號（經過校方加密處理）、學年度、學期、課程名稱、修別、開課年級、成績，此資料為 411 名學生之資料，總計有 17352 筆資料。本研究樣本的總課程數為 140 門，其中 59 門為必修(42%)，81 門為選修(58%)。本研究樣本的總資料筆數為 17352 筆，其中未通過課程門檻 60 分的佔 2039 筆(12%)，通過課程門檻佔 15313 筆(88%)，本研究以前 8 個學年度入學的學生資料做為訓練資料，最後 2 個學年度入學的學生資料做為測試資料，以進行預測模型訓練。原始的教務資料欄位中，因其資料欄位的型態與後續資料探勘所需的格式不同，因此需要透過資料整合與轉換的方式，產生衍生欄位，以符合後續探勘之需求，例如，將成績進行轉換，因不同課程之間授課老師評定成績的標準有不同的標準，直接進行後續動作，將因其佔比關係，不易找出有效的模組。因此，本研究將不及格的獨立出來，自成一個區間 F，其餘通過的成績則按其成績之排名，最終得到之排名區間欄位為「F、1、2、3、4」。

3.2. 研究工具

本研究使用 SAS 9.4 中的 SAS Enterprise Guide 8.1 與 SAS Enterprise Miner 15.1 做為研究的主要工具，SAS 是統計分析系統 (https://www.sas.com/zh_tw/software/enterprise-miner.html)，此跨國軟體研發公司所開發用於商業統計分析的套裝軟體，能提供建造出最合適的探勘模型。

4. 研究結果

4.1. 利用關聯法則建立學生課業成績表現關聯模組

生成之關聯規則模型根據其支持度的高低，取前 200 條規則做為後續之運用，此 200 條規則之支持度平均值為 11.55%，信賴度平均值為 43.58%，提升值平均值為 1.75，第一類預測準確率平均值為 33%，第二類預測準確率平均值為 53.43%。由於部分的規則之預測準確率較低，若規則之預測準確率只有 50%甚至低於 50%，此規則將無使用價值。因此，本研究將除去第一類預測準確率低於 50%的規則，並將剩餘之規則，依照第一類預測準確率，每間隔

2%做一次分類，藉此找出表現較佳之分類規則。透過比較不同分類規則，規則數在第一類預測準確率 65%以後低於 10 筆，規則數、準確率由於與分類規則相依，呈現持續下降與上升的趨勢。支持度先降後升，信賴度先升後降，而提升值方面持續上升。由於支持度、信賴度、提升值間的差異不大，而第一類預測準確率 60%以上的平均準確度與規則數量較符合本研究之需求，因此本研究使用第一類預測準確率 60%以上的作為關聯規則模組，如表 1 所示。

表 1

預測準確度 60%以上的關聯規則

	TRUE1	FALSE1	總人數1	成功率1		TRUE2	FALSE2	總人數2	成功率2	Support(%)	Confidence	PseudoLift	Rule
2	8	3	11	72.73%		10	1	11	90.91%	13.51	50.56	2.13	資工一上微積分(一)_4 ⟹ 資工一下微積分(二)_4
3	11	1	12	91.67%		11	1	12	91.67%	12.91	46.24	1.69	資工二下計算機網路_F ⟹ 資工三下畢業專題實作(一)_1
4	7	4	11	63.64%		8	3	11	72.73%	12.31	46.07	1.78	資工一上微積分(一)_4 ⟹ 資工一下離散數學_4
5	9	3	12	75.00%		11	1	12	91.67%	12.31	47.13	2.09	資工二上線性代數_4 ⟹ 資工二下機率與統計_4
6	5	3	8	62.50%		6	2	8	75.00%	12.01	51.95	2.40	資工二上工程數學_4 ⟹ 資工三上資料庫系統_4
7	10	2	12	83.33%		10	2	12	83.33%	11.71	44.83	2.37	資工二上線性代數_4 ⟹ 資工二下組合語言與系統程式_4
8	8	4	12	66.67%		10	2	12	83.33%	11.71	44.83	2.20	資工二上線性代數_4 ⟹ 資工三上演算法_4
9	6	2	8	75.00%		6	2	8	75.00%	11.41	49.35	2.61	資工二上工程數學_4 ⟹ 資工二下組合語言與系統程式_4
10	9	6	15	60.00%		11	4	15	73.33%	11.41	48.72	2.39	資工一下程式設計實習(二)_4 ⟹ 資工三上計算機組織_4
11	10	5	15	66.67%		11	4	15	73.33%	11.11	47.44	2.51	資工一下程式設計實習(二)_4 ⟹ 資工二下組合語言與系統程式_4
12	8	4	12	66.67%		11	1	12	91.67%	10.81	38.71	1.74	資工二下計算機網路_F ⟹ 資工四上畢業專題實作(二)_1
13	9	3	12	75.00%		11	1	12	91.67%	10.51	40.23	2.13	資工二上線性代數_4 ⟹ 資工二下計算機網路_4
14	9	6	15	60.00%		13	2	15	86.67%	10.21	43.59	2.30	資工一下程式設計實習(二)_4 ⟹ 資工二下計算機網路_4
15	5	3	8	62.50%		6	2	8	75.00%	10.21	44.16	2.33	資工二上工程數學_4 ⟹ 資工二下計算機網路_4
16	8	4	12	66.67%		10	2	12	83.33%	10.21	39.08	1.78	資工二上線性代數_4 ⟹ 資工三下畢業專題實作(一)_4
17	4	1	5	80.00%		4	1	5	80.00%	10.21	40.00	2.61	資工三上作業系統_4 ⟹ 資工三下網路規劃與管理_4
18	8	4	12	66.67%		10	2	12	83.33%	10.21	39.08	1.91	資工二上線性代數_4 ⟹ 資工三上計算機組織_4
19	16	6	22	72.73%		19	3	22	86.36%	10.21	53.97	2.64	資工二下組合語言與系統程式_4 ⟹ 資工三上演算法_4
20	9	3	12	75.00%		10	2	12	83.33%	9.91	37.93	1.91	資工二上線性代數_4 ⟹ 資工二下數位系統實驗_4
21													
22	159	67	226	70.35%		188	38	226	83.19%	11.21	44.94	2.19	平均

4.2. 結合群集分析技術建立學生課業成績表現關聯模組

群集分析結果，其中群集 1 最多，所有項目表現均為最差；群集 2，20 人，計算機概論、離散數學表現第三，其他項目表現第二；群集 3 人數 13 人，所有項目表現均為最佳；群集 4 共 17 人，在程式設計（一）、程式設計（二）表現第二，其他項目表現第三，如表 2 所示。

表 2.

分群結果

	群集平均值					
群集人數	群集	程式設計(一)	程式設計(二)	計算機概論	離散數學	rank_Score_Mean
28	1	24.57%	20.91%	22.97%	19.59%	36.98%
20	2	32.59%	36.98%	62.02%	50.63%	58.49%
13	3	76.58%	81.88%	81.09%	74.81%	78.00%
17	4	59.42%	61.95%	44.96%	36.24%	57.73%

在預測準確率方面，預測準確率一表現最差的為群集 4，然而群集 4 的預測準確率二的排名卻提升到第二高，由此可知運用於不同類型的群集資料，預測準確率會產生不同程度的影響。在分群規則方面，第一群為所有項目表現都最差的一群，而第 3 群為所有項目表現都最佳的一群，此兩個群集為資料中最極端的兩個群集。剩餘的群與 2 與群 4 的各項表現都居中，第二群在計算機概論與離散數學表現較佳，第四群在程式設計（一）與程式設計（二）表現較佳，此二群集為完全相反的兩個群集，如果不經過分群便直接進行分析，所得到的關聯模型價值將會降低。而所有提升數據中，最重要為預測準確率與規則涵蓋率，如表 3 所示。

表 3.

分群後預測效能分析比較表

		TRUE1	FALSE1	總人數1	成功率1		TRUE2	FALSE2	總人數2	成功率2	Support(%)	nfidence(%	PseudoLift		規則數	涵蓋率
4-1		486	174	660	73.64%		581	79	660	88.03%	26.39%	55.57%	1.32		71/200	35.50%
4-2		55	21	76	72.37%		59	17	76	77.63%	15.85%	40.92%	1.40		19/200	9.50%
4-3		255	110	365	69.86%		265	100	365	72.60%	22.02%	50.30%	1.46		54/200	27.00%
4-4		41	18	59	69.49%		50	9	59	84.75%	15.43%	43.31%	1.28		15/200	7.50%
4-總和平均		837	323	1160	72.16%		955	205	1160	82.33%	22.61%	50.87%	1.37		159/800	19.88%
無分群		159	67	226	70.35%		188	38	226	83.19%	11.21%	44.94%	2.19		19/200	9.50%

5. 結論

由於使用的是成績排名分區，並不是直接使用成績做為規則依據，成績不達及格標準的自成一區，其餘成績則依據排名做為分區，不僅能讓成績較差的學生選擇通過率較高之選課組合，提升其課程通過率。更能使得成績較優秀的學生選擇出成績表現較佳之選課組合，提升其總成績排行。由於學生的組成相當多元，部分學生擅長實作，部分學生擅長思考，每名學生都有自身的特色，若將所有學生不經分類就直接進行分析，所產生的關聯模組準確度勢必會降低，因此本研究先將學生進行群集分析，再對各群集進行關聯規則分析，以降低群集內資料的特徵差異性。因此結合群集分析來建立關聯規則模組，依據研究結論是可提升學習預警預測準度。

參考文獻

吳靜薇(2014).應用決策樹與關聯規則於學生成績之分析－以台南市某職業學校為例.南台科技大學資訊管理研究所碩士論文.

余禎祥. (2019). 磨課師學習分析的軟體框架開發. 逢甲大學資訊工程學系碩士論文. https://hdl.handle.net/11296/g5b8f9

胡詠翔. (2019). 大規模開放線上課程學習分析促進科技學科教學知識之研究. 教學實踐與創新, 2(1), 77-114. https://doi.org/10.3966/261654492019030201004

陳嘉鴻. (2014). 結合藍海策略與資料探勘技術分析 Moodle 網路教學平台之教學成效-以國小資訊課程為例.義守大學. https://hdl.handle.net/11296/c642v6

曾敬翔. (2020). 在 MOOCs 上基於影片學習活躍度預測課程表現系統. 清華大學資訊工程學系碩士論文 https://hdl.handle.net/11296/zefyh4

Aljarbouh, A., Tsarev, R., Robles, A. S., Elkin, S., Gogoleva, I., Nikolaeva, I., & Varyan, I. (2022). Application of the K-medians Clustering Algorithm for Test Analysis in E-learning. In Proceedings of the Computational Methods in Systems and Software (pp. 249-256). Cham: Springer International Publishing.

Botelho, A. F., Varatharaj, A., Patikorn, T., Doherty, D., Adjei, S. A., & Beck, J. E. (2019). Developing Early Detectors of Student Attrition and Wheel Spinning Using Deep Learning. IEEE Transactions on Learning Technologies, 12(2), 158-170.

Han, J., Pei, J., & Tong, H. (2022). Data mining: concepts and techniques. Morgan Kaufmann.

Roy, D., & Dutta, M. (2022). A systematic review and research perspective on recommender systems. Journal of Big Data, 9(1), 59.

Smirnov, I. (2020). Estimating educational outcomes from students' short texts on social media. EPJ Data Science, 9(1), 27. https://doi.org/10.1140/epjds/s13688-020-00245-8