運用資料探勘技術於預測國中生成績與分析其學習狀況

Using Data Mining Technologies to Predict the Scores of Junior High Students and the Analysis of Their Studies

李仁鐘
華梵大學資訊管理系教授
johnlee@hfu.edu.tw

陳郁榮
華梵大學資訊管理系研究生
m9725007@cat.hfu.edu.tw

摘要

本研究以某國中某班學生的在校平時成績、段考成績與復習考成績為研究主題，期能從實際數據中，探討其關係，並提供訊息給老師及學生。本研究目的在於探討各學生之國文、英文、數學、自然、社會平時成績與復習考的各科成績之相關情形，以期能提供訊息給老師，讓老師給予學生督促及輔導。

本研究以資料探勘(Data Mining)技術中之支援向量迴歸（Support Vector Regression, SVR）、倒傳遞類神經網路（Back-Propagation Network, BPN）及迴歸樹（Regression Tree）軟體進行預測，且分析各科成績所預測之結果，再將結果之訊息及建議提供給學校老師，做為輔導學生之參考。

關鍵詞：預測成績、資料探勘、倒傳遞類神經網路、支援向量迴歸、迴歸樹

Abstract

The objective of this study is based on the scores from junior high students according to their scores of quiz, exam, and review test. The relationship among the data is expected as a reference for either teachers or students. The main purpose is to analyze students' test and exam scores in English, Mandarin, math, nature science, and social science. Thus teacher can supervise and give students a proper guidance.

The study uses the Data Mining software like SVR (Support Vector Regression), BPN (Back-Propagation Network), and Regression Tree to survey, predict, and analyze every test result in order to provide a series of data as a reference for teachers.

Keywords: Score Predicting, Data Mining, Support Vector Regression, Back-Propagation Network, Regression Tree.

1. 研究動機與目的

1.1 研究動機

在學習生涯當中，國民中學的階段是最備受重視及關注，也是最為重要的。許多家長因為工作關係，無法兼顧孩子的學業，因此，學校老師便扮演起這重要的角色，而學生的平時成績及段考成績，即是老師參考的指標。然而，國中階段正值大多中學生的青春期，學生可能因為感情問題或生活方面的因素而影響到學習，導致成績下滑，老師可能無法察覺到每一位學生的學習狀況及成績的細微變化。因此，如何有效率且有方法的從這些平時及段考

成績中判斷並統計出學生成績的變化，是一項很重要的課題。

1.2 研究目的

本研究使用某國中之某班的七年級上、下學期所有的平時考及段考成績去預測其上、下學期的複習考成績，使用資料探勘技術，包括倒傳遞類神經網路(Back-Propagation Network, BPN)、支援向量迴歸（Support Vector Regression, SVR）及迴歸樹（Regression Tree）軟體三種技術去預測，將實際數值與預測數值互相比較，再將結果提供給老師做為參考。以預測結果與實際複習考成績做比對，觀察學生的複習考成績與預測成績，瞭解學生平時小考到複習考期間的學習是否穩定、正常，進而知道學生上課時的學習和回家後的用功程度是否影響到複習考之成績。

2. 文獻探討

本研究目前所使用到的資料探勘技術，包括：支援向量迴歸、倒傳遞類神經網路、及迴歸樹共三種的資料探勘技術。

2.1 支援向量機（Support Vector Machine, SVM）

支援向量機(Support Vector Machine, SVM)是由 Vapnik 在1995年和AT&T實驗室團隊所發展出的學習演算法(Vapnik, 1995)，它是一種統計學習理論（Statistical Learning Theory）。

支援向量機初始的學習架構是以小樣本的學習來得到最佳的學習與歸納能力，其主要是利用分隔超平面(Separating Hyperplane)的方法，尋找最大的邊界(Margin)，進而將資料分隔成兩類或多類的類別(Class)(Burges, 1998)。1997年，Drucker 和Burges等人提出了以支援向量機為基礎所發展出的一個新的迴歸技術，稱之為支援向量迴歸(Support Vector Regression, SVR)，其主要是利用已知的資料去對未知的變數進行預測(林智仁，LibSVM)，而在支援向量迴歸利用訓練資料產生迴歸方程式的過程中，並非所有的訓練資料對於建立迴歸方程式都是有幫助的，這些資料中，也會存在著雜訊(Noise)或離群值(Outlier)，其都會影響最後所預測結果的準確率，為解決此問題，則可利用損失函數(Loss Function)和懲罰係數(Penalty Parameter)來解決(蔡承益，2007)。

2.2 倒傳遞類神經網路(Back-Propagation Network, BPN)

倒傳遞類神經網路模式是目前類神經網路學習模式中最具代表性，應用最普遍的模式。P. Werbos於1974年在其論文中提出了加入隱藏層的學習演算法，使得網路可表現輸入處理單元間的交互影響，他突破了在1957年所提出的感知器模式，因此模式缺乏隱藏層的學習演算法的因素，其學習能力受到很大的限制，導致無法解決XOR的問題，因而發明了倒傳遞類神經網路(Back-Propagation Network, BPN) (Chang, 2008)。

倒傳遞類神經網路是一種前饋式網路，具有監督式學習的過程，其基本原理是利用最陡坡降法(The Gradient Steepest Descent Method)的觀念，將誤差函數予以最小化推導出誤差法則(Delta Rule)，其構想是透過連續性修正值來降低實際輸出與期望輸出的差距。而學習過程則是藉由訊息正向傳播(Forward-Pass)與誤差負向傳播(Backward-Pass)兩階段所組成。正向傳播在運作時，其權重是固定的；而負向傳播運作時，其權重先是不變，待取到誤差參數值後，再配合所選取的學習法則以調整其權重(陳贊仁，2009)。

2.3 決策樹(Decision Tree)

決策樹是利用樹的觀念，依事物的特徵，將事物區分為不同的種類，每一種類再對應不同的決策模式。在現有的決策樹分類器中，最被廣為使用的就是 ID3 (Quinlan, 1986)、C4.5 (Quinlan, 1993)、CART (Breiman et al., 1984)，此三種方法適用的資料型態是文字符號或是離散狀態，其處理的問題以分類(Classification)問題為主。1996 年，Quinlan 改良傳統的 C4.5，使其能有效的處理連續性數值的資料。而本研究所採用的就是此種決策樹方法，稱為迴歸樹(Regression Tree)，此方法為 C4.5 的改良，可處理連續性的資料型態(Jeffrey T. W., 2008)。

迴歸樹是將估計的模型以決策樹的方式呈現，依據資料的屬性，將資料分割成若干區塊，再各自於區塊中選取重要的自變數，建立個別的迴歸模型。其概念與決策樹相同，都是依據資料屬性進行分類，但不同的是，迴歸樹的終端節點是一條迴歸方程式；而決策樹的終端節點則是該筆資料所屬的類別，因此，迴歸樹所處理的是屬於連續數值的資料型態，而決策樹則是處理分類型態的資料(陳樹衡，2007)。

當資料中遇有遺漏值時，可用「？」表示，系統會給予一個平均的值帶入，並運算之。其正確率以平均誤差值(Average Error)表示，其方程式如下：

$$\text{Average Error} = \frac{\sum_{i=1}^{n}\left|(d_i - y_i)\right|}{n} \qquad (1)$$

d_i：實際值
y_i：預測值
n ：資料總筆數

3. 實驗方法與分析結果

3.1 研究資料

本研究之資料取自於臺北縣某國中之某班 34 位學生七年級上、下學期各科所有的平時成績、段考成績及複習考成績來實驗，透過資料探勘的技術，以複習考成績為目標值預測之，期能準確預測出正確值或近似值，並提供結果及訊息給老師，以助輔導學生學習之參考用。

3.2 分析結果

本研究目前所試驗過的資料探勘技術，包含了支援向量機、倒傳遞類神經網路及迴歸樹軟體。

在使用支援向量機、倒傳遞類神經網路及迴歸樹軟體預測出結果後，因此三種資料探勘技術所使用之正確率的表示方法皆不同，為公平起見及方便比較，在此皆以均方根誤差(Root Mean Squared Error, RMSE)表示, 其方程式如下：

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(d_i - y_i)^2}{n}} \qquad (2)$$

d_i： 原始成績
y_i： 預測成績
n： 學生人數

當所計算出之值越小，則代表其預測越準確，結果如下表所示：

表 1 三種資料探勘技術之上學期各科 RMSE 之值

科目	支援向量機	倒傳遞類神經網路	迴歸樹軟體
國文	0.1	10.325	7.859
英文	0.099	9.331	4.69
數學	0.1	11.712	10.305
自然	0.099	6.772	6.647
社會	0.099	6.418	6.016

表 2 三種資料探勘技術之下學期各科 RMSE 之值

科目	支援向量機	倒傳遞類神經網路	迴歸樹軟體
國文	0.1	4.751	7.979
英文	0.1	4.276	4.701
數學	0.1	11.615	14.45
自然	0.1	6.378	8.281
社會	0.1	5.641	7.285

由上面兩張表格得知，支援向量機是此三種資料探勘技術中，預測之結果最為準確的，而迴歸樹軟體在上學期的預測能力比倒傳遞類神經網路佳；但下學期則反之。

為驗證此三種方法是否適用於預測成績，本研究進行 T 檢定求取其 *P* 值，觀察三者的顯著差益與否。將上學期和下學期的實際成績及預測成績匯進 Excel 中，並設定其參數，即可取得 *P* 值，其結果如下表所示：

表 3 三種資料探勘技術之上學期各成績之 *P* 值

科目	支援向量機	倒傳遞類神經網路	迴歸樹軟體
國文	0.499	0.213	0.498
英文	0.498	0.205	0.499
數學	0.499	0.376	0.499
自然	0.498	0.451	0.497
社會	0.498	0.455	0.498

表 4 三種資料探勘技術之下學期各成績之 *P* 值

科目	支援向量機	倒傳遞類神經網路	迴歸樹軟體
國文	0.499	0.489	0.497
英文	0.499	0.499	0.499
數學	0.499	0.479	0.498
自然	0.498	0.486	0.498
社會	0.498	0.484	0.499

當 *P* 值＞0.05 時，代表實際值與預測值之間無顯著差異；而 *P* 值＜0.05 時，代表實際值與預測值之間有顯著差異。由表三及表四得知，此三種資料探勘之技術不管在哪一科目，其 *P* 值皆＞0.05，故本資料之實際值與預測值之間無顯著差異，此三種方法皆適用於預測成績之研究上。

為提供有效的訊息及結果給老師，本研究以實際成績比預測成績差三科以上之學生，視為老師必需注意之名單，其統計結果分為上、下兩學期，一方面可看出哪些學生的成績有進步；另一方面也可看出哪些學生的成績有下滑或依舊處於退步狀態，其統計結果如下表所示：

表 5 上學期之實際成績比預測好之統計結果（以座號表示）

科目總數	支援向量機	倒傳遞類神經網路	迴歸樹軟體
五科成績皆比預測好	4、10、11、16、21、22、25、26、27、28、29、32	26	

四科成績比預測好	6、15、19、20、30、31、33	11、19、22、28	1、7、16、17、19、27、28、29、31、32
三科成績比預測好		1、3、4、5、13、14、17、21、27、32、34	4、10、11、13、14、18、20、22、26

表 6 下學期之實際成績比預測好之統計結果（以座號表示）

科目總數	支援向量機	倒傳遞類神經網路	迴歸樹軟體
五科成績皆比預測好	4、6、10、11、16、20、25、26、27、28、29、30、32	10、12、13、16、27	10、13、27、29
四科成績比預測好	15、19、21、23、31	4、5、21	1、4、16、17、21、28、31
三科成績比預測好	2、13、33	1、8、14、15、17、20、23、26、28、29、31	12、14、15、18、19、20、23

由表 5 上學期之模擬實驗結果得知此三種方法共同預測之實際成績比預測好的學生共有八位，其分別為座號 4、11、19、22、26、27、28、32 號，其中以座號 19 號的同學預測最為準確，不論是在支援向量機、倒傳遞類神經網路或迴歸樹軟體，皆坐落於四科實際成績比預測成績好；而表 6 為下學期之模擬實驗結果，由此表得知，三種方法共同預測到下學期之實際成績比預測成績好的學生共有十二位，其座號分別為 4、10、13、15、16、20、21、23、27、28、29、31 號，其中座號 10 號及 27 號同學在此三種方法所預測之五科實際成績皆比預測成績好，而 21 號同學之成績預測結果，在此三種方法中，皆坐落於四科實際成績比預測好。

表 7 上學期之實際成績比預測差之統計結果（以座號表示）

科目總數	支援向量機	倒傳遞類神經網路	回歸樹軟體
五科成績皆比預測差	1、5、8、9、12、17、18、34	9、23、24	9
四科成績比預測差	7、24	6、7、15、18、25	6、15、24、30
三科成績比預測差	2、3、13、14、23	16、31、34	2、12、21、25、33、34

表 8 下學期之實際成績比預測差之統計結果（以座號表示）

科目總數	支援向量機	倒傳遞類神經網路	迴歸樹軟體
五科成績皆比預測差	1、5、7、8、12、17、18、34	11、25、34	34
四科成績比預測差	3、9、14、22、24	7、18、32、33	2、8、9、11、22、30、33
三科成績比預測差	13	2、3、6、9、19、22、24、	3、5、6、7、24、25、26、

		30	32

由表 7 之模擬結果顯示，不論是在支援向量機、倒傳遞類神經網路或迴歸樹軟體，皆顯示上學期五科實際成績皆比預測結果差的學生為 9 號同學，而 24 號和 34 號同學的預測結果雖然都坐落於不同的地方，但都有出現在此三種資料探勘技術的模擬結果之中，因此，也被列為老師所該關注之名單。再觀察表 8 之下學期成績預測之模擬結果得知，座號 34 號同學在此三種資料探勘技術中，其實際成績皆比預測成績低；而 9 號及 24 號同學在下學期之預測，依然處於被關注名單中，因此建議老師也要多注意此兩位同學之學習狀況及態度。

四、結論

由模擬實驗結果得知，不論是支援向量機、倒傳遞類神經網路或是迴歸樹軟體，三者皆適用於本研究之預測成績之上，其中尤以支援向量機最為準確。由表五及表六之模擬結果得知，其 4 號及 27 號同學，在上、下學期之成績皆比平時成績有進步之趨勢，老師可予以獎勵，也可請同學向其兩位學生請教讀書之方法且讓兩位同學扮演小老師的角色，輔導較跟不上進度之同學。由表七及表八之模擬結果得知，34 號同學在上學期被列為觀察名單之中，而下學期所預測之結果，其實際成績皆比預測成績低，由此得知，該名學生之成績有明顯下滑的現象，老師應予以輔導或是特別督導其學習狀況；而 9 號及 24 號同學之成績，依然屬於被關注名單之中，因此老師應該也要特別注意此兩位學生之學習狀況。由表八也可發現，成績退步之學生的比率有明顯增加，其結果可能與課程的難易度有關，建議老師可依學生之能力，調整其授課方式亦或是讓學生多做練習。

五、建議與未來研究方向

綜合前述，由數據歸納出在實際成績與預測成績的差異方面，老師可依學生之學習能力及狀況方面，制定學習時程表，並把時程表一同交給家長與學生，如此一來，在老師、家長以及學生三方面的配合，可以更容易去幫助學生改善其學習的效率及狀況。而在學校方面，某些成績較差的學生，可能礙於許多因素，不敢向老師發問，因此，老師可以藉同儕之間的力量，請成績較優異的學生扮演小老師的角色，在課後之餘輔導成績較差的學生，或是讓學生組成一個類似讀書會的團體，彼此間可以互相發問，找尋共同的問題或疑慮，並利用下課時間一同向老師請教；而在平時測驗時，也可獎勵成績進步的學生，並予以鼓勵，讓學生們在學習遇到挫折時，有額外的動力使他們成長。

在未來之研究方向，本研究將使用其他的資料探勘技術去實驗，例如自適應性神經模糊推理系統來做預測，期能找到更省時、更準確之方法，並繼續蒐集更多的成績資料，以提供更快速且更精確之結果給老師作為參考。

參考文獻

[1] 林智仁，LibSVM, URL: http://www.csie.ntu.edu.tw/~cjlin

[2]陳贊仁，”以倒傳遞網路設計籃球運動彩券推薦模式”，***大同大學資訊工程研究所碩士論文***，2009。

[3]陳樹衡、郭子文、棗厥庸，”以決策樹之迴歸樹建構住宅價格模型—台灣地區之實證分析”，***住宅學報第十六卷第一期***，pp. 2-7，2001。

[4]蔡承益，”使用 SOM-SVR 混合型系統搭配屬性篩選模式應用於臺灣股票指數期貨預測”，***國立高雄第一科技大學資訊管理系碩士論文***，2007。

[5]Burges, C.,“A tutorial on support vector

machines for pattern recognition," ***Data Mining and Knowledge Discovery***, Vol. *2,* No. 2, pp. 121-167, 1998.

[6]Chang, T. H., "Using Back-Propagation Network to Predict Proper Cyclosporine Dosage in Patients After Kidney Transplantation," ***Thesis for Master of Science, Department of Computer Science and Engineering***, Tatung University, 2008.

[7] Jeffrey T. W., "Subpixel Urban Land Cover Estimation: Comparing Cubist, Random Forests, and Support Vector Regression," ***Photogrammetric Engineering & Remote Sensing***, pp. 1213–1222, 2008.

[8] Quinlan, J. R., "Introduction of Decision Tree," ***Machine Learning***. Vol. 1, pp. 81-106, 1986.

[9] Quinlan, J. R., "***C4.5: Programs for Machine Learning***," CA: Morgan Kaufmann, 1993.

[10] Quinlan, J. R., "Improved use of continuous attributes in C4.5," ***Journal of Artificial Intelligence Research***, Vol. 4, pp. 77–90, 1996.

[11] Vapnik, B. N., ***The Natural of Statistical Learning Theory***, 2nd Edition, Springer-Verlag, N.Y., USA, 1995.