運用決策樹於中學生飲酒預測之研究

Using Decision Tree to Predict Drinking of Middle School Students

莊奕婕[1] 林昀葆[2] 黃謙順[3]

CHUANG,YI CHIEH[1] LIN,YUN PAO[2] HUANG,CHIEN SHUN[3]

[1] 中國文化大學 資訊管理學系研究所 研究生

[1] Private Chinese Culture University of Information Management Student

[2] 中國文化大學 資訊工程學系研究所 研究生

[2] Private Chinese Culture University of Computer Science and Information Engineering Student

[3] 中國文化大學 資訊管理學系研究所 教授

[3] Private Chinese Culture University of Information Management Professor

摘要

根據國內外新聞及調查皆可發現，中學生飲酒的情形越趨普遍，引發的社會案件也漸趨頻繁，相關調查更指出飲酒年齡有逐漸下降的趨勢，此外對於身理及心理皆尚在發育階段的中學生而言，飲酒也有礙身心發展；本研究為了瞭解中學生飲酒與學習表現之相關因素。透過葡萄牙學生問卷調查收集的資料集，並使用資料探勘技術來分析學生飲酒與家庭、朋友間的因素，建立決策樹模型預測學生飲酒情況。研究結果發現影響學生飲酒的相關屬性，包括性別、和朋友出去頻率、上課缺席次數等。本研究的預測模型的準確率高達 84.0%。希望本研究預測模型能協助相關單位及早發覺有飲酒情形的中學生加以介入並關懷。

關鍵字：學生飲酒預測，學生飲酒因素、決策樹

Abstract

According to domestic and foreign research, the alcohol-drinking condition of middle school students is becoming more and more common and causing a number of social incidents. Many relevant researches even pointed out that alcohol-drinking age has become younger. Besides, for students who are still in their development stage, alcohol is harmful to physical and psychological health. The purpose of this study is to understand and find the factors about the middle school students drinking problem and learning situation.Through the questionnaire and data collection from Portugal students. We use data mining technology to analyze the factor of students drinking problem between family and friends, and establish a decision tree to predict students drinking problem. The accuracy of the prediction model reaches as high as 84.0%. Hoping the proposed model can be used to help the concerned department discover the student with drinking problem earlier.

Keywords：Student alcohol consumption prediction, Students drinking factors,

decision tree

壹、緒論

現今含有酒精類飲品對許多人來說已被視為生活休閒必需品，然而青少年飲酒比例也逐漸升高。根據林美嫻(2007)研究指出，青少年階段是健康行為養成的關鍵期，影響著其未來的健康與生活品質。吸菸、飲酒、嚼檳榔、過去兩個星期沒有運動、沒有體重控制及沒有每天吃水果六項行為男生的發生率顯著高於女生。

全世界 4%至 30% 的癌症死亡都可歸因於飲酒。即便是適量的飲酒也會增加其風險：每天飲用一杯就會增加 4%的乳腺癌患病風險，而過量飲酒則會增加 40%至 50%的風險。重度飲酒更是削弱了免疫系統，飲酒還會導致危險性行為，增加性傳播感染的機會(Nogrady B., 2015)世界衛生組織指出在西元 2012 年全球因使用酒精致死約 330 萬例。飲酒不僅產生酒精依賴，也會增加罹患超過 200 種疾病的風險，包括肝硬化與某些癌症。(陳筱蕾，2014）

貳、文獻探討

一、飲酒行為影響因素

許多研究顯示影響青少年行為發展有很大的原因在於同儕之間的相互影響及家庭環境因素。根據洪兆嘉(2009)研究結果顯示，開始飲酒的年齡、同儕飲酒、和酒品可取得性會直接與飲酒行為類型有關。另外，開始飲酒的時間、同儕飲酒及同儕勸酒的壓力會透過認知因素間接與飲酒行為類型有關。在劉美媛、周碧瑟(2001)研究中指出，家庭因素中的單親家庭結構型態、父母對你的評價等，學校因素上課業表現、出席情形、打工狀況等以及年齡、原住民、父母、兄弟姊妹飲酒等都是與飲酒有關的因素。學生飲酒對生活影響方面，在 Statistic Brain 網站，統計了美國大學生狂飲至爛醉數據為 44%，而飲酒的大學生之中為了喝醉而飲酒的佔 48%，因飲酒而錯過課程的大學生佔 21%。

二、資料探勘技術－決策樹

資料探勘，意指利用一個龐大數據庫建立模型，並從中找出隱藏的特殊關聯及特徵，發掘各種的隱藏資訊，提供決策支援之用。透過資料探勘技術，可從原本雜亂無章的資料中找出有用的資訊，像是藉由分析特定顧客的消費行為，找出顧客特質、消費特徵等，助於經營者做決策或行銷策略時參考使用。

決策樹(decision tree)，依照各項條件自動分割，再運用歸納方法找出資料的規則，可用來做決策或預測的模型。決策樹為提供分類及預測的常用方法，是一種以樹狀結構來展現資料個別變數。決策樹以每個內部節點表示一個屬性評估欄

位，以每個分支代表一個可能的欄位輸出結果，以每個樹葉節點來代表不同分類的類別標記。ID3、C4.5 及 CHAID 等皆是決策樹常用的演算法。

參、研究方法

為了瞭解青少年飲酒對其學習表現之影響。本研究透過 UCI data set 網站內「Student Alcohol Consumption Data Set」所提供的資料集並使用軟體工具 Weka 建立預測模型。其資料來源為葡萄牙兩所大學以問卷的方式所收集而成。其中統計資料分為修習數學課的學生及修習葡萄牙語的學生(共 1,044 筆資料)。

本研究將修習葡萄牙語及數學課的兩筆資料集合併為一個資料集並將當中原始資料屬性的 Dalc(工作日星期一至星期五的飲酒情形)與 Walc(週末星期六、日的飲酒情形)轉換成新的屬性 alc 代表一週的飲酒情形，計算公式如下：

$$\text{alc} = \frac{Dalc * 5 + Walc * 2}{7}$$

並將 alc 屬性進行等寬方法將資料的取值範圍按照等距離劃分公式如下：

$$W = (B - A)/N$$

W 是寬度，B 是取值範圍的最大值，A 是取值範圍的最小值，N 是取值範圍的個數，以 alc 屬性為例，值範圍從 0 到 5，套用等寬公式 W 寬度等於 2.5，區間落在[0-2.5],(2.5-5)，並以前者標記成 low 為低度飲酒者，後者標記成 high 為高度飲酒者。

本研究使用的資料集和家庭有相關的包括學生家庭型態、父母是否同居、家庭對教育的支持度、家庭關係等，另外也有學生的性別、年齡、每週讀書時間、上課缺席次數、和朋友出去、在校學習成績等共 30 個資料屬性。

本研究模型採用資料探勘技術決策樹 C4.5，並以十折交叉驗證來測試演算法。十折交叉驗證是將數據集分成十份，輪流將其中九份做訓練一份做測試，進而產生預測模型。並以混淆矩陣進行驗證，混淆矩陣會呈現真確定(TP)、誤確定(FP)、真否定(TN)和誤否定(FN)。再透過決策樹交叉驗證其準確率(Accuracy)、精確率(Precision)及召回率(Recall)。準確率為預測實際有飲酒學生占全部資料的百分比，精確率為預測有飲酒的學生中，實際上真正有飲酒的學生資料，召回率為實際值中有多少比率是模型所預測出來的，公式如下。

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

肆、結果與討論

探討學生飲酒與學生表現之分析結果，本研究使用 Weka 軟體進行分析，而 Weka 中以建構決策樹的 C4.5 演算法，名稱為 J48，建立預測模型，以十折交叉驗證進行訓練與測試。

混淆矩陣行代表實際類別，列代表預測類別，實際類別為 low 且預測類別為 low 相符合有 796 個，預測類別為 low 且實際類別為 high 的有 103 個，實際類別為 high 且預測類別為 high 相符合有 85 個，預測類別為 high 且實際類別為 low 的有 60 個。經過 J48 演算法預測出來的學生飲酒類別中，評估準確率(Accuracy)、精確率(Precision)、召回率(Recall)結果。如表 1。

表 1　J48 預測結果

Accuracy	Precision	Recall
84.387%	0.885	0.930

根據研究結果之決策樹來看，會影響學生飲酒情形主要有以下幾個因素，包括性別、和朋友出去頻率、上課缺席次數、每週讀書時間以及家庭對教育的支持度等皆有一定的影響。其中以男性占為多數，另外自由時間愈多，經常和朋友出去的學生飲酒情形愈普遍。

伍、未來展望

本研究之預測模型評估準確率達到 84.387%，顯然所得到的結果有相當的可靠性。希望能夠藉由此模型來協助相關單位提早發現中學生飲酒的情形加以介入關懷並深入探討中學生飲酒的原因，再針對其原因制定各項輔導方案以及制定其他相關政策來加以輔導，期望能達到減少青少年飲酒的現象。

參考文獻

中文部份

劉美媛、周碧瑟(2001)。臺灣在校青少年飲酒盛行率與相關因素的探討。臺灣公共衛生雜誌，*20*(2)，143-152。

陳筱蕾編(2014)。世界衛生組織呼籲各國政府致力於防止與酒精有關的死亡與疾病。國家衛生研究院電子報，*554*。

洪兆嘉(2009)。從社會認知理論探討青少年的飲酒行為及相關因素。國立臺灣大學衛生政策與管理研究所，臺北市。

林美嫻(2007)。臺灣地區青少年危害健康行為及其相關因素研究。國立臺

灣師範大學衛生教育學系，臺北市。

英文部份

College Student Alcohol Drinking Statistics. Retrieved August 14, 2016, from http://www.statisticbrain.com/college-student-alcohol-drinking-statistics/

Nogrady,B. (2015).*Is alcohol actually bad for you?* Retrieved September 2, 2015, from http://www.bbc.com/future/story/20150901-is-alcohol-really-bad-for-you

STUDENT ALCOHOL CONSUMPTION Data Set. Retrieved March 3, 2016, from https://archive.ics.uci.edu/ml/datasets/STUDENT+ALCOHOL+CONSUMPTION