Journal of educational equipment: Applied research, Volume 2, Issue 283 (February 2023)
ISSN 1859 - 0810

Lựa chọn thuật toán cây quyết định tốt nhất để dự đoán và phân loại hành động của sinh viên

Nguyễn Văn Quang*

**ThS. Công nghệ thông tin, Trường Đại học Hải Phòng*
Received: 27/01/2023; Accepted: 30/01/2023; Published: 06/02/2023

Abstract: *Since the student's success rate reflects the success of educational organizations, so the trend of increasing student's success became the goal of all educational organizations. Besides that, the student's willingness of studying higher education after complete secondary school is one of the most important goals to the educational Organizations. Many reasons affect this willingness and revealing these reasons may enhance the student's will. Data mining tools (especially Decision Tree Algorithms) can be considered as the best choice to find the hidden patterns in order to achieve these goals. The experimental dataset used in this work is data set about Portuguese student on two courses (Mathematics (395 instances) and Portuguese (Portuguese language course which holds 659 instances)) which was collected and analyzed by Paulo Cortez and Alice Silva, University of Minho, Portugal. Three Decisions Tree algorithms (J48, RepTree and Hoeffding Tree (VFDT)) are applied and experimented in this work. The results showed that J48 algorithm mostly proper to classify and predict both students' willingness to complete higher education and success in courses.*

Keywords: *Educational Data Mining, Decision Tree Algorithms, J48 Algorithm, RepTree Algorithm, VFDT Algorithm.*

1. Đặt vấn đề

Khai thác dữ liệu, còn được gọi là Khám phá tri thức trong cơ sở dữ liệu (KDD), là lĩnh vực khám phá thông tin mới và có khả năng hữu ích từ lượng lớn dữ liệu. Trong những năm gần đây, người ta ngày càng quan tâm đến việc sử dụng khai thác dữ liệu để điều tra các câu hỏi khoa học trong nghiên cứu giáo dục (GD), một lĩnh vực điều tra được gọi là khai thác dữ liệu GD. Khai thác dữ liệu GD (còn được gọi là "EDM"). Các nhà nghiên cứu EDM nghiên cứu nhiều lĩnh vực khác nhau, bao gồm học tập cá nhân từ phần mềm GD, học tập hợp tác được máy tính hỗ trợ, kiểm tra thích ứng với máy tính (và kiểm tra rộng hơn) và các yếu tố liên quan đến việc học sinh (HS) không đạt hoặc không theo học các khóa học.

Một trong những lĩnh vực chính của ứng dụng EDM là cải tiến các mô hình HS dự đoán các đặc điểm hoặc thành tích học tập của HS trong các trường phổ thông, cao đẳng và các cơ sở GD khác. Dự đoán kết quả học tập của HS với độ chính xác cao rất hữu ích trong nhiều bối cảnh ở tất cả các cơ sở GD để xác định HS chậm và phân biệt HS có thành tích học tập thấp hoặc HS yếu có khả năng có thành tích học tập thấp. Sản phẩm cuối cùng của các mô hình sẽ có lợi cho giáo viên, phụ huynh và các nhà lập kế hoạch GD không chỉ để thông báo cho HS trong quá trình học tập, liệu hành vi hiện tại của họ có thể liên quan đến kết quả tích cực và tiêu cực trong quá khứ hay không, mà còn đưa ra lời khuyên để khắc phục vấn đề.

Ba thuật toán Cây quyết định (C4.5 (J48), RepTree và Hoeffding Tree) được áp dụng. Tập dữ liệu chính bao gồm hai tệp Giá trị được phân tách bằng dấu phẩy (CSV) được lấy từ Kho lưu trữ học máy của UCI cho mức tiêu thụ rượu của sinh viên (SV) trong hai khóa học (Ngôn ngữ Bồ Đào Nha và Toán học). Các tệp tập dữ liệu nguồn chứa (1044 trường hợp trong các trường hợp) với 32 thuộc tính. Một số thao tác tiền xử lý như (làm sạch dữ liệu, tạo cột và xóa cột) được triển khai để hợp nhất hai tệp nguồn này trong một tập dữ liệu.

2. Nội dung nghiên cứu

Trong [1] các nhà nghiên cứu đã so sánh các phương pháp và kỹ thuật khai thác dữ liệu khác nhau để phân loại SV dựa trên dữ liệu sử dụng Moodle của họ và điểm cuối kỳ đạt được trong các khóa học tương ứng của họ. Họ đã phát triển một công cụ khai thác cụ thể để tạo cấu hình và thực hiện các kỹ thuật khai thác dữ liệu dễ dàng hơn cho người hướng dẫn. Họ đã sử dụng dữ liệu thực tế từ bảy khóa học Moodle với SV Đại học Cordoba. Họ tuyên bố rằng một mô hình phân loại thích hợp cho việc sử dụng GD phải chính xác và dễ hiểu đối với người hướng dẫn để được sử dụng cho việc ra quyết định.

Trong [2] các phương pháp và kỹ thuật khai thác dữ liệu khác nhau đã được so sánh trong quá trình

dự đoán thành công của SV, áp dụng dữ liệu thu thập được từ các cuộc khảo sát được thực hiện trong học kỳ mùa hè tại Đại học Tuzla, Khoa Kinh tế, năm học 2010-2011, giữa các SV năm thứ nhất và dữ liệu được lấy trong quá trình tuyển sinh. Sự thành công được đánh giá bằng điểm đạt tại kỳ thi. Tác động của các biến nhân khẩu học - xã hội của HS, kết quả đạt được ở trường trung học và từ kỳ thi tuyển sinh, và thái độ đối với việc học có thể ảnh hưởng đến sự thành công, tất cả đều được điều tra.

Trong [3] các kỹ thuật khai thác dữ liệu nhằm tiếp cận thành tích của HS ở trường trung học sử dụng dữ liệu trong thế giới thực. Hai lớp cốt lõi (Toán học và Tiếng Bồ Đào Nha) được mô hình hóa theo các nhiệm vụ hồi quy và phân loại nhị phân/năm cấp độ. Bốn mô hình DM (tức là Cây quyết định, Rừng ngẫu nhiên, Mạng thần kinh và Máy véc tơ hỗ trợ) và ba lựa chọn đầu vào (ví dụ: có và không có điểm trước) đã được thử nghiệm. Kết quả cho thấy có thể đạt được độ chính xác dự đoán tốt, với điều kiện là có sẵn các điểm của học kỳ đầu tiên và/hoặc thứ hai. Mặc dù thành tích của HS bị ảnh hưởng nhiều bởi các đánh giá trong quá khứ, một phân tích giải thích đã chỉ ra rằng cũng có những yếu tố khác có liên quan.

2.1. Khai thác dữ liệu GD

Các phương pháp EDM thường khác với các phương pháp từ tài liệu khai thác dữ liệu rộng hơn, trong việc khai thác rõ ràng nhiều cấp độ phân cấp có ý nghĩa trong dữ liệu GD. Các phương pháp từ tài liệu tâm lý học thường được tích hợp với các phương pháp từ tài liệu khai thác dữ liệu và máy học để đạt được mục tiêu này. Ví dụ: khi khai thác dữ liệu về cách HS chọn sử dụng phần mềm GD, có thể đáng để xem xét đồng thời dữ liệu ở cấp độ gõ phím, cấp độ câu trả lời, cấp độ phiên, cấp độ HS, cấp độ lớp học và cấp độ trường học. Các vấn đề về thời gian, trình tự và bối cảnh cũng đóng vai trò quan trọng trong nghiên cứu dữ liệu GD.

Sau khi một cấu trúc quan tâm đến GD (chẳng hạn như hành vi ngoài nhiệm vụ hoặc liệu một kỹ năng có được biết hay không) đã được xác định theo kinh nghiệm trong dữ liệu, nó có thể được chuyển sang các tập dữ liệu mới. Việc chuyển giao các cấu trúc không phải là chuyện nhỏ – thông thường, cùng một cấu trúc có thể khác nhau một cách tinh tế ở cấp độ dữ liệu, trong dữ liệu từ một ngữ cảnh hoặc hệ thống khác – nhưng phương pháp học chuyển giao và ghi nhãn nhanh đã thành công trong việc tăng tốc quá trình phát triển hoặc xác thực một mô hình cho một bối cảnh mới. Điều này đã dẫn đến nhiều phân tích khai thác dữ liệu GD được sao chép trên dữ liệu từ một số hệ thống hoặc bối cảnh học tập.

2.2. Thuật toán cây quyết định

Cây là đồ thị có hướng bắt đầu bằng một nút và phân nhánh thành nhiều nút. Chúng là nền tảng cho khoa học máy tính (cấu trúc dữ liệu), sinh học (phân loại, tâm lý học (lý thuyết quyết định) và nhiều lĩnh vực khác. Cây phân loại và hồi quy được sử dụng để dự đoán. Trong hai thập kỷ qua, chúng đã trở nên phổ biến như là những lựa chọn thay thế cho hồi quy, phương pháp sắp xếp cây đã trở nên phổ biến đến mức một số chương trình thương mại hiện đang cạnh tranh để thu hút sự chú ý của các nhà nghiên cứu thị trường và những người khác đang tìm kiếm phần mềm.

4.1 J48: J48graft là phiên bản mở rộng của J48 xem xét việc ghép các nhánh bổ sung vào cây trong giai đoạn xử lý hậu kỳ (Webb, 1999). Quá trình ghép cố gắng đạt được một số sức mạnh của các phương pháp tập hợp chẳng hạn như cây được đóng gói và tăng cường trong khi duy trì một cấu trúc có thể hiểu được. Nó xác định các vùng của không gian phiên bản trống hoặc chỉ chứa các ví dụ bị phân loại sai và khám phá các phân loại thay thế bằng cách xem xét các thử nghiệm khác nhau có thể đã được chọn tại các nút phía trên lá chứa vùng được đề cập.

2.3. Cây đại diện

RepTree xây dựng một cây quyết định hoặc cây hồi quy bằng cách sử dụng giảm mức tăng/giảm phương sai thông tin và cắt tỉa nó bằng cách cắt tỉa giảm lỗi. Được tối ưu hóa về tốc độ, nó chỉ sắp xếp các giá trị cho các thuộc tính số một lần. Nó xử lý các giá trị bị thiếu bằng cách chia các phiên bản thành nhiều phần, giống như C4.5. Bạn có thể đặt số lượng phiên bản tối thiểu trên mỗi lá, độ sâu tối đa của cây (hữu ích khi tăng cường cây), tỷ lệ phương sai tập huấn luyện tối thiểu cho một phần tách (chỉ các lớp số) và số lần gấp để cắt tỉa.

2.4. Độ sâu của cây cuốc đất

Độ sâu tối đa của cây (hữu ích khi tăng cường cây), tỷ lệ phương sai tập huấn luyện tối thiểu cho một phần tách (chỉ các lớp số) và số lần gấp để cắt tỉa.

Cây Hoeffding (VFDT) là một thuật toán cảm ứng cây quyết định gia tăng, bất cứ lúc nào có khả năng học hỏi từ các luồng dữ liệu lớn, giả sử các ví dụ tạo phân phối không thay đổi theo thời gian. Cây Hoeffding khai thác thực tế là một mẫu nhỏ thường có thể đủ để chọn thuộc tính phân tách tối ưu. Ý tưởng này được hỗ trợ về mặt toán học bởi giới hạn Hoeffding, định lượng số lượng quan sát (trong trường hợp của chúng tôi là các ví dụ) cần thiết để ước tính một số thống kê trong một độ chính xác quy định (trong trường hợp của chúng tôi là mức độ tốt của một thuộc tính). Một tính năng hấp dẫn về mặt lý thuyết của Cây Hoeffding không được chia sẻ bởi những người học cây quyết

định gia tăng khác là nó có sự đảm bảo chắc chắn về hiệu suất. Sử dụng giới hạn Hoeffding, người ta có thể chỉ ra rằng đầu ra của nó gần giống với đầu ra của một người học không gia tăng bằng cách sử dụng vô số ví dụ.

2.6. Mô hình cây quyết định

2.6.1. Tiền xử lý dữ liệu

Bộ dữ liệu (Bộ dữ liệu về mức tiêu thụ rượu của SV) phụ thuộc vào mô hình này. Dữ liệu bao gồm hai bộ dữ liệu student-mat.csv (Khóa học toán có 395 phiên bản) và student-por.csv (Khóa học tiếng Bồ Đào Nha có 659 phiên bản). độ sâu tối đa của cây (hữu ích khi tăng cường cây), tỷ lệ phương sai tập huấn luyện tối thiểu cho một phần tách (chỉ các lớp số) và số lần gấp để cắt tỉa.

2.6.2. Cây Hoeffding

Cây Hoeffding (VFDT) là một thuật toán cảm ứng cây quyết định gia tăng, bất cứ lúc nào có khả năng học hỏi từ các luồng dữ liệu lớn, giả sử rằng các ví dụ tạo phân phối không thay đổi theo thời gian. Cây Hoeffding khai thác thực tế là một mẫu nhỏ thường có thể đủ để chọn thuộc tính phân tách tối ưu. Ý tưởng này được hỗ trợ về mặt toán học bởi giới hạn Hoeffding, định lượng số lượng quan sát (trong trường hợp của chúng tôi là các ví dụ) cần thiết để ước tính một số thống kê trong một độ chính xác quy định (trong trường hợp của chúng tôi là mức độ tốt của một thuộc tính). Một tính năng hấp dẫn về mặt lý thuyết của Cây Hoeffding không được chia sẻ bởi những người học cây quyết định gia tăng khác là nó có sự đảm bảo chắc chắn về hiệu suất. Sử dụng giới hạn Hoeffding, người ta có thể chỉ ra rằng đầu ra của nó gần giống với đầu ra của một người học không gia tăng bằng cách sử dụng vô số ví dụ.

2.7. Kiểm tra thuật toán cây quyết định

Trong thử nghiệm đầu tiên để xây dựng nhóm cây quyết định, tất cả các thuộc tính được chọn trừ khi (G1, G2, G3, G1Grade, G2Grade) và G3Grade được giữ nguyên để sử dụng trong việc xây dựng cây. Đã xóa thuộc tính vắng mặt và thay vào đó sử dụng AbsRate. Chế độ kiểm tra là 10 xác thực chéo với lớp mục tiêu cao hơn.

Thuật toán đại diện cho tên của thuật toán được sử dụng trong quá trình thử nghiệm.

- CCI (Phiên bản được phân loại chính xác) biểu thị số lượng phiên bản được phân loại chính xác chia cho tổng số phiên bản và nhân với 100.

- ICI (Phiên bản được phân loại không chính xác) biểu thị số lượng phiên bản được phân loại không chính xác chia cho tổng số phiên bản và nhân với 100.

- Độ chính xác: của thuật toán biểu thị tỷ lệ phần trăm các trường hợp được phân loại chính xác từ tất cả các trường hợp được phân loại thực sự.

- Thu hồi phản ánh số phân chia của các phiên bản được phân loại chính xác theo tổng số của tất cả các phiên bản (giá trị thu hồi gần như giống với CCI).

- F-Measure: được đo từ các giá trị thu hồi và độ chính xác (gấp đôi giá trị của độ chính xác nhân với thu hồi chia cho giá trị tổng của thu hồi và độ chính xác).

- Vùng ROC là viết tắt của Vùng đặc tính hoạt động của máy thu mô tả hiệu suất của bộ phân loại mà không tính đến chi phí phân phối hoặc lỗi của lớp.

- Thời gian: lấy phần thứ hai để dựng cây.

Nhóm cây quyết định thứ hai được xây dựng dựa trên (Khóa học, trường học, giới tính, tuổi, địa chỉ, famsize, Pstatus, Medu, Fedu, Mjob Fjob, Reason, người giám hộ, thời gian đi lại, thời gian học, thất bại, trường học, famsup, trả tiền, hoạt động, nhà trẻ, internet, lãng mạn, gia đình, thời gian rảnh rỗi, bệnh gút, sức khỏe, vắng mặt, AbsRate và G3Grade). Chế độ Kiểm tra là xác thực chéo 10 lần để xây dựng cây với tỷ lệ chính xác cao và lá cuối cùng (lớp được đặt cao hơn (HS có muốn học đại học hay không) để xây dựng quyết định

- Cây phân loại và dự đoán HS dựa trên Thuộc tính đầu vào.

3. Kết luận

Bài báo này đã liệt kê và so sánh kết quả của việc thực hiện ba giải thuật cây quyết định khác nhau. Đồ thị cây quyết định bị ảnh hưởng bởi số thuộc tính đầu vào và thuộc tính lớp kết thúc. Hai lớp chính (Thành công của SV (G3Grade) và Sẵn sàng học đại học (cao hơn)) được chọn để xây dựng biểu đồ cây. Kết quả cho thấy J48 là thuật toán cây quyết định tốt nhất có thể được sử dụng làm sơ đồ dự đoán và phân loại hành động của SV. Việc lựa chọn J48 xuất phát từ các kết quả được so sánh bên cạnh số lượng nút trong biểu đồ ảnh hưởng đến khả năng hiển thị của cây.

Tài liệu tham khảo

1. Romero, Cristóbal, et al. “Data mining algorithms to classify students.” Educational Data Mining 2008. 2008.

2. Osmanbegović, Edin, and Mirza Suljić. “Data mining approach for predicting student performance.” Economic Review 10.1 (2012).

3. Cortez, Paulo, and Alice Maria Gonçalves Silva. “Using data mining to predict secondary school student performance.” (2008).

4. Kabakchieva, Dorina. “Predicting student performance by using data mining methods for classification.” Cybernetics and information technologies 13.1 (2013): 61-72.