Journal of educational equipment: Applied research, Volume 2, Issue 297 (September 2023)
ISSN 1859 - 0810

Khai thác dữ liệu giáo dục

Hoàng Văn Lâm*

**ThS. Công nghệ thông tin, Trường Đại học Hải phòng*
Received: 30/7/2023; Accepted: 7/8/2023; Published: 14/8/2023

Abstract. *The adoption of learning management systems in education has been increasing in the last few years. Various data mining techniques like prediction, clustering and relationship mining can be applied on educational data to study the behavior and performance of the students. This paper explores the different data mining approaches and techniques which can be applied on Educational data to build up a new environment give new predictions on the data. This study also looks into the recent applications of Big Data technologies in education and presents a literature review on Educational Data Mining and Learning Analytics.*

Keywords**:** *EDM, Prediction, Clustering, Relationship Mining*

1. Đặt vấn đề

Hiện nay có rất nhiều nghiên cứu về lĩnh vực khai thác dữ liệu. Khai thác dữ liệu giáo dục là một lĩnh vực nghiên cứu chính còn được gọi là EDM. Nó nhằm mục đích nghĩ ra và sử dụng các thuật toán để cải thiện kết quả giáo dục và giải thích các chiến lược giáo dục để đưa ra quyết định tiếp theo. Bài viết này thảo luận về một số thuật toán khai thác dữ liệu được áp dụng trong các lĩnh vực liên quan đến giáo dục. Các thuật toán này được áp dụng để trích xuất kiến thức từ dữ liệu giáo dục và nghiên cứu các thuộc tính có thể góp phần tối đa hóa hiệu suất. Trên thực tế, việc học ban đầu bắt đầu trong lớp học và dựa trên các mô hình hành vi, nhận thức và kiến tạo [1],[2]. Các mô hình hành vi dựa vào những thay đổi có thể quan sát được trong hành vi của học sinh để đánh giá kết quả học tập. Các mô hình nhận thức dựa trên sự tham gia tích cực của giáo viên vào quá trình học tập. Trong các mô hình kiến tạo, học sinh phải tự học từ những kiến thức sẵn có. Một thuật ngữ mới "Chủ nghĩa kết nối" được mô tả là "sự khuếch đại việc học tập, kiến thức và hiểu biết thông qua việc mở rộng mạng lưới cá nhân" đã xuất hiện trong những năm gần đây. Theo Siemens, việc học không còn là nút thắt giúp cải thiện trải nghiệm học tập của sinh viên và giảm nhu cầu có sự tham gia trực tiếp của Giáo sư. Trên thực tế, môi trường học tập truyền thống đã dần chuyển sang môi trường học tập dựa vào cộng đồng [4].

2. Nội dung nghiên cứu

2.1 Khai thác dữ liệu, khái niệm và thách thức

Khai thác dữ liệu giáo dục có thể được định nghĩa là "Một môn học mới nổi liên quan đến việc phát triển các phương pháp khám phá các loại dữ liệu duy nhất đến từ môi trường giáo dục và sử dụng các phương pháp đó để hiểu rõ hơn về sinh viên cũng như môi trường mà họ học tập"[5]. EDM là quá trình chuyển đổi dữ liệu thô do hệ thống giáo dục biên soạn thành thông tin hữu ích có thể được sử dụng để đưa ra quyết định sáng suốt và trả lời các câu hỏi nghiên cứu. Nhưng sự phát triển của khai thác và phân tích dữ liệu trong lĩnh vực Giáo dục còn khá muộn so với các lĩnh vực khác. Tuy nhiên, thách thức đối với việc khai thác dữ liệu giáo dục của học tập trực tuyến là do các tính năng cụ thể của nó về dữ liệu. Trong khi nhiều loại dữ liệu có các khía cạnh tuần tự, việc phân bổ dữ liệu giáo dục theo thời gian có những đặc điểm riêng biệt; ví dụ, một kỹ năng có thể gặp nhiều lần trong một năm học, nhưng được tách ra theo thời gian và trong bối cảnh các hoạt động hoàn toàn khác nhau [6]. Ngoài ra, các phương pháp khai thác dữ liệu giáo dục đã thành công trong việc mô hình hóa một loạt hiện tượng liên quan đến việc học tập của học sinh trong các hệ thống và mô hình thông minh trực tuyến đang đạt được độ chính xác cao hơn hàng năm và đang được xác nhận để có thể khái quát hóa hơn theo thời gian. Có những khía cạnh quan trọng cần được thảo luận để biện minh cho sự phát triển độc đáo của dữ liệu giáo dục, đó là nhận thức ngày càng tăng rằng không phải tất cả thông tin quan trọng đều được lưu trữ trong một luồng dữ liệu; sự cải thiện về chất lượng mô hình, được thúc đẩy bởi những cải tiến liên tục về phương pháp luận và tầm quan trọng của việc tồn tại rằng có nhiều ví dụ về máy dò được công bố hơn là các máy dò được sử dụng để thúc đẩy sự can

thiệp, như Ellucian [6] [7] đã cung cấp cho các Giáo sư các báo cáo về liệu sinh viên có nguy cơ bỏ học hoặc trượt một khóa học hay không và hướng dẫn Giáo sư cách can thiệp, mang lại kết quả tốt hơn cho người học. Nghiên cứu về giáo dục [8] đã dẫn đến một số cải tiến sư phạm mới. Công nghệ dựa trên máy tính đã thay đổi cách chúng ta sống và học tập. Ngày nay, việc sử dụng dữ liệu được thu thập thông qua [6] các công nghệ này đang hỗ trợ vòng chuyển đổi thứ hai trong mọi lĩnh vực và học tập với những thành tựu khác nhau.

Khai thác dữ liệu là một công nghệ mới mạnh mẽ có tiềm năng lớn giúp các Trường học và Đại học tập trung vào những thông tin quan trọng nhất trong dữ liệu họ đã thu thập về hành vi của sinh viên và người học tiềm năng [9]. Khai thác dữ liệu liên quan đến việc sử dụng các công cụ phân tích dữ liệu để khám phá các mẫu và mối quan hệ chưa biết trước đây trong các tập dữ liệu lớn. Những công cụ này có thể bao gồm các mô hình thống kê, thuật toán toán học và phương pháp học máy.

2.2. Đánh giá tài liệu

Nhiều cuộc điều tra đã được thực hiện để chứng minh tầm quan trọng của kỹ thuật "Khai phá dữ liệu" trong giáo dục, chứng tỏ đây là một khái niệm mới đối với mục đích trích xuất thông tin hợp lệ và chính xác về hành vi và hiệu quả trong quá trình học tập [10] .

Trong lĩnh vực kỹ thuật giáo dục, "Khai thác dữ liệu" cũng được sử dụng để phân tích chương trình giảng dạy và chủ đề của các chủ đề nghiên cứu hiện tại, cũng như phân tích kết quả học tập của sinh viên . Đã có một số cuộc điều tra được thực hiện theo đối tượng nghiên cứu được đề xuất này. Ví dụ, Bhardwaj đã sử dụng thuật toán Naïve Bayes để dự đoán kết quả học tập của học sinh dựa trên 13 biến số . Các kết quả được sử dụng để xây dựng một mô hình nhằm xác định trước những học sinh có nguy cơ thất bại và từ đó kích hoạt chương trình hướng dẫn và tư vấn. Varghese, Tommy và Jacob trong nghiên cứu của họ đã sử dụng thuật toán "K Mean" để phân cụm 8000 sinh viên dựa trên 5 biến (trung bình đầu vào trong điểm trung bình của các bài kiểm tra/bài kiểm tra ở trường Đại học, điểm trung bình của các bài báo, ghi chú và ghi chú của hội thảo). làm việc theo tần số). Kết quả cho thấy có mối quan hệ chặt chẽ giữa việc đi học đầy đủ và kết quả học tập của học sinh. Gulati và Sharma cho rằng kiến thức thông qua phân tích bằng "Khai thác dữ liệu" có thể cải thiện hệ thống giáo dục về định hướng, hiệu quả hoạt động của sinh viên và quản lý tổ chức. Ayesha Mustafa đã chỉ đạo một nghiên cứu về đánh giá, có tính đến sự phát triển của việc học và phân tích các bài kiểm tra vào đầu và cuối khóa học. Bresfelean đã thực hiện một nghiên cứu dựa trên kết quả của sinh viên và mức độ dễ dàng thực hiện những kết quả này. Cortez và Silva đã tiến hành một nghiên cứu về hệ thống giáo dục ở Bồ Đào Nha và kết quả cho thấy có thể đạt được dự đoán tốt và chính xác. Điều này được thiết lập bởi các công cụ phát triển giúp cải thiện việc quản lý giáo dục trong trường học và hiệu quả học tập, đây là một lợi ích rất quan trọng. Theo Sun [9], kết quả của mối quan hệ giữa đánh giá và học tập là một công cụ quan trọng để giám sát và định hướng một nền giáo dục có chất lượng. Noaman và Al-Twijri đã công bố một nghiên cứu gần đây áp dụng cho yêu cầu đầu vào của Đại học Ả Rập Saudi. Họ đã sử dụng các thuật toán và kỹ thuật mà họ đã phát triển cũng như một mô hình phù hợp với công chúng và các biến mô tả nó. Họ đã tính đến việc nhập học đầu vào với tần suất ghi chú trong quá trình giáo dục trước đây, ghi chú nhập học và thậm chí cả các đặc điểm mô tả nhu cầu của trường Đại học. Một số nghiên cứu cho thấy tác động của việc sử dụng Moodle bằng cách áp dụng Data Mining . Sun [9] mô tả các kỹ thuật khai thác dữ liệu khác nhau có thể được áp dụng để thúc đẩy việc học tập của sinh viên trên nền tảng kỹ thuật số. Aslam và Ashraf đã sử dụng thuật toán phân cụm để cung cấp mô hình học tập của học sinh. Một số cuộc điều tra đã thảo luận về cách dữ liệu hoạt động để Dữ liệu cải thiện hệ thống giáo dục và tiếp thu kiến thức trong lớp học. Vince Kellen trong nghiên cứu điển hình của mình đã mô tả việc triển khai công cụ phân tích có cấu trúc cho Khai thác dữ liệu - HANA của SAP tại Đại học Kentucky, công cụ này ước tính giá trị "k-score" cho mỗi sinh viên. Giá trị này sẽ quyết định sự tham gia và hướng dẫn tiếp theo để học sinh đạt thành tích tốt. Grafsgaard, Wiggins, Boyer, Wiebe và Lester đã phát triển một hệ thống nhận dạng nét mặt dựa trên sự thất vọng hoặc hiểu biết của học sinh trong lớp. Họ cũng sử dụng các thuật toán để phát hiện những hành vi không được nói ra và liên kết chúng với kiến thức thu được. Seong Jae Lee cũng mô tả kỷ lục về việc sử dụng các mô hình dự đoán hành vi con người.

2.3 Các phương pháp khai phá dữ liệu trong dữ liệu giáo dục

Khai thác dữ liệu là lĩnh vực khoa học máy tính nhKhai thác dữ liệu là lĩnh vực khoa học máy tính nhằm tìm ra các yếu tố và mô hình tiềm năng khác nhau để giúp đưa ra quyết định.

Khai thác dữ liệu có thể tạo điều kiện thuận lợi cho Bộ nhớ tổ chức. Khai thác dữ liệu , còn được gọi phổ biến là Khám phá tri thức trong cơ sở dữ liệu, đề cập đến việc trích xuất hoặc "khai thác" kiến thức từ một lượng lớn dữ liệu. Một hệ thống giáo dục thường có một số lượng lớn dữ liệu giáo dục. Dữ liệu này có thể là của sinh viên dữ liệu, dữ liệu giáo viên, dữ liệu cựu sinh viên, dữ liệu tài nguyên, v.v. EDM tập trung vào phát triển các phương pháp khám phá các loại dữ liệu độc đáo đến từ bối cảnh giáo dục. Những dữ liệu này đến từ nhiều nguồn, bao gồm cả dữ liệu từ truyền thống trực tiếp đối mặt với môi trường lớp học, phần mềm giáo dục, chương trình học trực tuyến, v.v.

Kỹ thuật khai thác dữ liệu được sử dụng để hoạt động trên khối lượng lớn dữ liệu nhằm khám phá các mẫu và mối quan hệ ẩn hữu ích cho việc ra quyết định. Các thuật toán và kỹ thuật khác nhau như Phân loại, Phân cụm, Hồi quy, Trí tuệ nhân tạo, Mạng thần kinh, Quy tắc kết hợp, Cây quyết định, Thuật toán di truyền, Phương pháp lân cận gần nhất, v.v., được sử dụng để khám phá kiến thức từ cơ sở dữ liệu.

2.4. Dự đoán

Các kỹ thuật hồi quy (hình 4) có thể được điều chỉnh cho phù hợp với vị từ [25]. Phân tích hồi quy có thể được sử dụng để mô hình hóa mối quan hệ giữa một hoặc nhiều biến độc lập và biến phụ thuộc. Trong khai thác dữ liệu, các biến độc lập là các thuộc tính đã được biết đến và các biến phản hồi là những gì chúng ta muốn dự đoán. Thật không may, nhiều vấn đề trong thế giới thực không chỉ đơn giản là dự đoán. Do đó, các kỹ thuật phức tạp hơn (ví dụ: hồi quy logistic, cây quyết định hoặc mạng lưới thần kinh) có thể cần thiết để dự báo các giá trị trong tương lai.

3. Kết luận

Ngày càng có nhiều mối quan tâm nghiên cứu về việc sử dụng khai thác dữ liệu trong giáo dục. Lĩnh vực mới nổi này, được gọi là Khai thác dữ liệu giáo dục, liên quan đến việc phát triển các phương pháp khám phá kiến thức từ dữ liệu có nguồn gốc từ môi trường giáo dục. Khai thác dữ liệu là một lĩnh vực cực kỳ rộng lớn bao gồm việc sử dụng các kỹ thuật và thuật toán khác nhau để tìm kiếm mẫu. Bài viết này chỉ là một đánh giá đơn giản về lĩnh vực mới nổi này và nhằm mục đích làm nổi bật tầm quan trọng của nghiên cứu của nó. Có sự cải thiện đáng kể về mức độ phân loại chính xác hơn sau khi thực hiện thao tác trích chọn dữ liệu so với dữ liệu thô. Hơn thế nữa, với tất cả các trường hợp, thời gian thực hiện mô hình phân lớp cũng nhanh hơn đáng kể sau khi trích chọn thuộc tính quan trọng.Ví dụ, xét với giải thuật phân loại tốt nhất thuộc nhóm Bayes, trước khi trích chọn thuộc tính, BayesNet cho kết quả phân lớp chính xác là 89.33% với thời gian xây dựng mô hình phân lớp là 0.19s, trong khi độ chính xác của phân lớp được tăng lên 90% và thời gian xử lý chỉ còn 0.08s sau khi dữ liệu đã được trích chọn. Thêm vào đó, có thể thấy với giải thuật tốt nhất thuộc nhóm Cây quyết định, mô hình phân lớp dựa trên RandomForest cho kết quả độ chính xác phân lớp sau khi trích chọn dữ liệu cao hơn trước khi trích chọn 0.66%, trong khi thời gian xử lý giảm 0.25s.

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.

http://www.cs.waikato.ac.nz/~ml/weka/

5. Kumar, S. Anupama, and M. N. Vijayalakshmi. “Efficiency of decision trees in predicting student’s academic performance.” First International Conference on Computer Science, Engineering and Applications, CS and IT. Vol. 2. 2011.

6. Witten, Ian H., and Eibe Frank. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, Second Edition, 2005.

7. Witten, Ian H., and Eibe Frank. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, Third Edition, 2011.

8. Wilkinson, Leland. “Classification and regression trees.” Systat 11 (2004): 35-56.

9. Baker, R. S. J. D. “Data mining for education.” International encyclopedia of education 7 (2010): 112-118.