TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU GIÁO DỤC

Nguyễn Ngọc Đan Thanh¹, Hà Thị Thúy Vi²

AN OVERVIEW OF EDUCATIONAL DATA MINING

Nguyen Ngoc Dan Thanh¹, Ha Thi Thuy Vi²

Tóm tắt – Dữ liệu giáo dục đại học ngày càng lớn, nhưng việc khai phá và trích xuất những tri thức có ý nghĩa phục vụ công tác quản lý chưa được quan tâm nhiều. Công cụ khai phá hiện tại chưa mang lại hiệu quả. Nghiên cứu này trình bày ba kỹ thuật khai phá dữ liệu giáo dục: (1) Kỹ thuật phân loại, (2) Mô hình dự đoán, (3) Kỹ thuật gom cụm. Đồng thời, nghiên cứu đề xuất một số giải pháp phân tích và trực quan hóa dữ liệu, dự đoán năng lực học tập của sinh viên và gom cụm người học. Trên cơ sở đó, nhà quản lý giáo dục có thể lựa chọn giải pháp phù hợp để khai phá dữ liệu phục vụ công tác quản lý và đào tạo một cách hiệu quả.

Từ khóa: dữ liệu giáo dục đại học, khám phá tri thức, khai phá dữ liệu, khai phá dữ liệu giáo dục, kho dữ liệu.

Abstract – Higher education data is growing, but the exploitation and extraction of meaningful knowledge for management have not been paid much attention. The existing mining tools are not effective. This study aims to introduce three techniques for educational data mining: (1) Classification techniques, (2) Predictive models, (3) Clustering techniques. Simultaneously, the study also proposes some solutions to analyze and visualize data, predict students' learning capacity and assemble learners. Thereby, education managers could choose appropriate data mining solutions for effective management and training.

Keywords: Higher education data, Knowledge discovery in database, Data mining, Educational data mining, data warehousing.

I. GIỚI THIỆU

Trong giáo dục đại học, các nhà quản lý đặc biệt chú trọng đến chất lượng của chương trình đào tạo và việc làm sau khi tốt nghiệp của sinh viên, đó là hai yếu tố quyết định sự tồn tại của một cơ sở giáo dục. Trong điều kiện gia tăng mạnh của dữ liệu giáo dục đại học hiện nay, việc ứng dụng các kỹ thuật khai phá dữ liệu để trích xuất những tri thức có ý nghĩa từ các bộ dữ liệu lớn nhằm mục đích cải tiến các quyết định quản lý đang là mối quan tâm hàng đầu. Việc khai phá dữ liệu mang lại rất nhiều lợi ích thiết thực. Các cơ sở giáo dục có thể phân công nguồn nhân lực hiệu quả hơn và chủ động trong quản lý kết quả của sinh viên. Cơ sở giáo dục có thể cải thiện chính sách, nâng cao chiến lược và cải thiện chất lượng hệ thống quản lý. Ví dụ khi khai phá dữ liệu về sinh viên như: sở thích, điểm số, thời gian học tập, môi trường học tập, chúng ta có thể dự đoán được một sinh viên có khả năng vượt qua một khóa học hay không hay tạm dừng khóa học. Mặt khác, ở khía cạnh quản lý và xây dựng thông tin chiến lược, việc khai phá dữ liệu tuyển sinh sẽ giúp người quản lý xây dựng kế hoạch đào tạo phù hợp với đối tượng người học. Trong nghiên cứu này, nhóm tác giả sẽ làm rõ thế nào là khai phá dữ liệu giáo dục với các kỹ thuật khai phá dữ liệu và những ứng dụng có thể triển khai trong lĩnh vực giáo dục đại học.

^{1,2}Khoa Kỹ thuật và Công nghệ, Trường Đại học Trà Vinh

Email: ngocdanthanhd@tvu.edu.vn

^{1,2}School of Engineering and Technology, Tra Vinh University

II. TỔNG QUAN NGHIÊN CỨU

A. Tình hình nghiên cứu trong nước

Khai phá dữ liệu giáo dục là một lĩnh vực đang được ứng dụng mạnh mẽ nhằm mục đích nâng cao chất lượng đào tạo và năng lực học tập của người học. Trong báo cáo Hội thảo quốc gia về Công nghệ Thông tin năm 2013 tại Trường Đại học Cần Thơ, nhóm tác giả Huỳnh Lí Thanh Nhân và Nguyễn Thái Nghe đã đề cập đến hệ thống dự đoán kết quả học tập của sinh viên sử dụng thư viện mã nguồn mở MyMediaLite [1]. Trong báo cáo, nhóm tác giả đề cập đến nhiều nghiên cứu liên quan đến việc sử dụng khai phá dữ liệu trong việc dự đoán năng lực sinh viên như Cây quyết định, KNN, Bayes, Luật kết hợp, ... Tuy nhiên, kết quả của nghiên cứu này chỉ thể hiện dưới dạng quy luật chung, chỉ gom nhóm sinh viên mà không dự đoán riêng cho từng cá nhân. Từ đó, nhóm tác giả đã đề xuất phương pháp Biased Matrix Factorization trong xây dựng hệ thống dự đoán.

B. Tình hình nghiên cứu ngoài nước

Trên thế giới, nhiều ứng dụng được triển khai trong lĩnh vực khai phá dữ liệu giáo dục. Ứng dụng phương pháp học máy để dự đoán khả năng tốt nghiệp của sinh viên Đại học California, Los Angeles dựa trên các thông tin về môn học và tiến độ học tập của sinh viên trong hơn ba năm học [2]. Trước khi được khai phá, dữ liệu giáo dục cần phải được trải qua giai đoạn tiền xử lý nhằm mục đích phân cụm dữ liệu. Có nhiều thuật toán để làm điều này và một nhóm tác giả đã tổng hợp các thuật toán khai phá dữ liệu được đề cập trong [3]. Ngoài ra, mô hình dự đoán khả năng bỏ học của sinh viên sử dụng các kỹ thuật khai phá dữ liệu và trên cơ sở đó lập kế hoạch ngăn cản tình trạng này đã được nhóm tác giả nghiên cứu trong [4]. Một cách tiếp cận dự đoán đa lớp để đo lường kỹ năng nhận thức của sinh viên được đề cập trong [5]. Kết quả nghiên cứu cho thấy phương pháp tiếp cận này tối ưu hơn so với các kỹ thuật đo lường kỹ năng nhận thức của sinh viên hiện có với độ chính xác đạt 0.979.

Khai phá dữ liệu giáo dục và phân tích dữ liệu học tập ngày càng được quan tâm bởi sự hữu ích

của nó trong việc dự đoán hành vi và tri thức của sinh viên. Nhiều tác giả đã trình bày tổng quan về vấn đề khai phá dữ liệu trong giáo dục [6]. Đồng thời, nhóm tác giả cũng đề cập đến một ứng dụng phổ biến và hữu ích trong giáo dục là dự đoán về hiệu suất học tập trong tương lai của sinh viên dựa trên điểm số hiện tại của họ. Một số kỹ thuật khai phá dữ liệu được sử dụng để dự đoán sớm nguy cơ trong học tập của sinh viên như kỹ thuật phân lớp Naive Bayes, kỹ thuật KNN, kỹ thuật SVM, mạng Noron được nghiên cứu và đề cập trong [7], [8]. Một nghiên cứu về các đặc trưng của sinh viên xét trên hai khía cạnh: tổng thời gian dành cho bài tập về nhà và mức độ nghiêm túc khi làm bài tập về nhà của sinh viên để dự đoán xác định các loại bài tập về nhà cho sinh viên [9].

III. KHAI PHÁ DỮ LIỆU GIÁO DỤC

A. Khái niệm khai phá dữ liệu giáo dục

Khai phá dữ liệu giáo dục (EDM – Educational Data Mining) là một miền con của khai phá dữ liệu và có liên quan đến việc phân tích, nghiên cứu dữ liệu từ cơ sở dữ liệu học thuật. Thông qua việc khai phá các bộ dữ liệu lớn này, các phương pháp khai phá dữ liệu khác nhau, chúng ta có thể xác định các mẫu đặc trưng sẽ giúp nghiên cứu, dự đoán và cải thiện kết quả học tập của người học [10].

Một nhóm tác giả nghiên cứu dữ liệu giáo dục từ năm 1995 đến năm 2010 và đề xuất mô hình xác định quy trình hoạt động của ứng dụng khai phá dữ liệu trong hệ thống giáo dục [11], được minh họa như Hình 1.

B. Các kỹ thuật khai phá dữ liệu giáo dục và ứng dụng

Kỹ thuật khai phá dữ liệu được sử dụng để thao tác trên một tập dữ liệu lớn nhằm phát hiện các thông tin hữu ích hỗ trợ việc ra quyết định. Hai khái niệm khai phá dữ liệu và khám phá tri thức trong cơ sở dữ liệu thường được xem là đồng nghĩa nhưng thực chất khai phá dữ liệu là một giai đoạn của quá trình khám phá tri thức. Cụ thể, các giai đoạn trong quá trình khám phá tri thức được minh họa như Hình 2.

Hình 1: Quy trình hoạt động của hệ thống khai phá dữ liệu giáo dục

Có nhiều thuật toán và kỹ thuật được sử dụng trong quá trình khám phá tri thức, cụ thể: Sự phân loại (Classification), Phân cụm (Clustering), Hồi quy (Regression), Trí tuệ nhân tạo (Artificial Intelligence), mạng Nơ-ron (Neural Networks), Luật kết hợp (Association Rules), Cây quyết định (Decision Trees), Thuật toán di truyền (Genetic Algorithm), Phương pháp người láng giềng gần nhất (Nearest Neighbor method).

Khai phá dữ liệu là một trong các tiến trình của khám phá tri thức nên có sự liên quan về các kỹ thuật khai phá dữ liệu giữa chúng. Trong khai phá dữ liệu giáo dục, các kỹ thuật này cũng thường được sử dụng. Chúng được liệt kê trong Bảng 1.

Mỗi ứng dụng đều mang lại những lợi ích khác nhau nhưng tùy theo thực trạng hiện tại và nhu cầu cần cải tiến tại cơ sở giáo dục, nhà quản lý sẽ chọn lựa, triển khai những ứng dụng phù hợp.

C. Phân tích và trực quan hóa dữ liệu

Phương pháp phân tích và trực quan hóa dữ liệu giúp làm nổi bật các thông tin hữu ích và hỗ trợ ra quyết định. Ví dụ, những ứng dụng này có

thể giúp các nhà giáo dục và quản lý khóa học phân tích các hoạt động khóa học và thông tin sử dụng của sinh viên để có cái nhìn tổng quát về việc học tập của họ. Kỹ thuật thống kê và trực quan hóa dữ liệu là hai kỹ thuật được áp dụng rộng rãi và liên quan đến việc thu thập, phân tích, giải thích và trình bày dữ liệu. Kỹ thuật này tương đối dễ dàng để có được số liệu thống kê cơ bản mô tả từ các phần mềm thống kê như SPSS, R, ... Trực quan hóa dữ liệu là việc sử dụng các kỹ thuật đồ họa để giúp mọi người hiểu và phân tích dữ liệu. Có nhiều nghiên cứu hướng đến trực quan hóa dữ liệu giáo dục khác nhau như mô hình phản ánh hành vi người dùng theo giờ trên các diễn đàn học tập trực tuyến, các đồ thị thống kê thông tin hoạt động của người học, điểm số các bài đánh giá, sự tham gia lớp học nhằm làm cơ sở để dự đoán trạng thái có thể hoàn thành khóa học trong tương lai hay không.

D. Dự đoán kết quả học tập của sinh viên

Dự đoán kết quả học tập của sinh viên là một ứng dụng phổ biến nhất của lĩnh vực khai phá dữ liệu giáo dục. Có nhiều mô hình và kỹ thuật

An overview of the steps that compose the knowledge discovery in databases [Fayyad et al. 1996]

Hình 2: Quy trình khám phá tri thức [11]

Bảng 1: Các kĩ thuật khai phá dữ liệu trong giáo dục

STT	Kĩ thuật khai phá dữ liệu	Ứng dụng khai phá dữ liệu trong giáo dục
1	Phân loại (Classification)	Phân loại và mô tả sơ lược sinh viên, xác định phong cách học tập và sở thích của họ
2	Mô hình dự đoán (Predictive Modeling)	Tạo ra các mô hình có thể dự đoán sinh viên đậu hay rớt một khóa học hoặc thậm chí là tốt nghiệp hay không.
3	Gom cụm (Clustering)	Gom nhóm những sinh viên giống nhau (dựa trên sự tương đồng về hành vi hay năng lực học tập) hoặc nhóm các khóa học, các bài đánh giá tương tự, tìm kiếm các mô hình học tập cộng tác.
4	Mô hình khai phá mới (Emerging Pattern Mining)	Tìm ra các mô hình xác định sự khác biệt đáng kể trong hành vi của những sinh viên tốt nghiệp so với những sinh viên không tốt nghiệp hoặc những mô hình có thể giải thích những thay đổi trong hành vi của các thể hệ sinh viên trong những năm khác nhau.
5	Lựa chọn cộng tác và khuyến nghị (Collaborative filtering and recommendations)	Đề xuất các mục tiêu học tập phù hợp dựa trên phân tích kết quả học tập và hành vi của người học, khuyến nghị các lớp bổ túc kiến thức cho sinh viên.
6	Phân tích trực quan (Visual Analytics)	Tạo điều kiện thuận lợi cho tiến trình giáo dục hoặc kết quả học tập thông qua trực quan hóa dữ liệu, ví dụ trực quan hóa sự cộng tác của các sinh viên.

khác nhau được áp dụng như mạng nơ-ron, hồi quy, phân tích tương quan,... để phân tích dữ liệu giáo dục. Kết quả phân tích này sẽ giúp dự đoán được sự thành công của sinh viên trong một khóa học hoặc thậm chí là có thể tốt nghiệp hay không.

E. Phân cụm người học

Kĩ thuật phân cụm được sử dụng để phân chia tập dữ liệu lớn thành nhiều tập dữ liệu con. Nhiều kĩ thuật phân cụm khác nhau được sử dụng để nhóm các sinh viên như kĩ thuật phân phân cụm theo thứ bậc, kĩ thuật K-means và phân cụm dựa

trên mô hình. Các nhóm người học sẽ được phân chia dựa trên các đặc điểm và khả năng của họ. Những nhóm học này có thể được người hướng dẫn định hướng và xây dựng mô hình học tập riêng cho nhóm nhằm phát huy những lợi thế của họ trong quá trình hoạt động học tập theo nhóm.

IV. KẾT LUẬN VÀ KHUYẾN NGHỊ

Bài viết trình bày tổng quan về các ứng dụng và kĩ thuật được sử dụng trong lĩnh vực khai phá dữ liệu giáo dục. Những thông tin này góp phần giúp các nhà quản lí giáo dục hình dung được

các ứng dụng có thể triển khai nhằm cải tiến các quy trình hay các nhân tố cần thiết để nâng cao chất lượng giáo dục.

Trường Đại học Trà Vinh có nhiều loại hình đào tạo (*Tập trung, không tập trung, vừa làm vừa học, trực tuyến*) với nguồn dữ liệu về chương trình đào tạo, người học, sự tương tác của người học với khóa học, dữ liệu đánh giá chương trình đào tạo... Đây là nguồn dữ liệu phong phú mà chúng ta có thể khai thác để giúp nhà quản lý định hướng ra quyết định nhằm nâng cao chất lượng đào tạo của nhà trường. Từ thực tế đó, Trường Đại học Trà Vinh hoàn toàn có thể triển khai các ứng dụng:

- (1) Xây dựng kho dữ liệu chuyên biệt (về sinh viên, chương trình đào tạo, nhân sự);
- (2) Dự đoán năng lực học tập của sinh viên;
- (3) Phân tích, trực quan hóa dữ liệu.

Những ứng dụng này sẽ mang lại nhiều lợi ích thiết thực: giúp nhà quản lý có thể phân tích, đánh giá lại chương trình đào tạo; giúp dự đoán kết quả học tập hỗ trợ sinh viên lập kế hoạch học tập hay giảng viên có thể điều chỉnh phương pháp giảng dạy phù hợp với đối tượng người học.

TÀI LIỆU THAM KHẢO

- [1] Huynh-Ly, Thanh-Nhan, Thai-Nghe, Nguyen. (2013). *A system for predicting students's course result using a free recommender system library - MyMediaLite*.
- [2] Xu, J., Moon, K. H., Van Der Schaar, M. (2017). A machine learning approach for tracking and predicting student performance in degree programs. *IEEE Journal of Selected Topics in Signal Processing*, 11(5), 742-753.
- [3] Dutt, A., Ismail, M. A., Herawan, T. (2017). A systematic review on educational data mining. *IEEE Access*, 5, 15991-16005.
- [4] Burgos, C., Campanario, M. L., de la Pena, D., Lara, J. A., Lizcano, D., Martínez, M. A. (2018). Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. *Computers & Electrical Engineering*, 66, 541-556.
- [5] Ahmad, S., Li, K., Amin, A., Anwar, M. S., Khan, W. (2018). A Multilayer Prediction Approach for the Student Cognitive Skills Measurement. *IEEE Access*, 6, 57470-57484.
- [6] Romero, C., Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1), 12-27.

- [7] Marbouti, F., Diefes-Dux, H. A., Madhavan, K. (2016). Models for early prediction of at-risk students in a course using standards-based grading. *Computers & Education*, 103, 1-15.
- [8] Shahiri, A. M., Husain, W. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, 72, 414-422.
- [9] Flunger, B., Trautwein, U., Nagengast, B., Lüdtke, O., Niggli, A., Schnyder, I. (2017). A person-centered approach to homework behavior: Students' characteristics predict their homework learning type. *Contemporary Educational Psychology*, 48, 1-15.
- [10] K. Shyamala and S.P. Rajagopalan, 2006. Data Mining Model for a Better Higher Educational System. *Information Technology Journal*, 5: 560-564.
- [11] Aher, S. B., Lobo, L. M. R. J. (2011, March). Data mining in educational system using Weka. In *International Conference on Emerging Technology Trends (ICETT)* (Vol. 3, pp. 20-25).