

Tổng quan về lợi ích và hạn chế của khai thác dữ liệu trong nghiên cứu giáo dục

Trần Quang Huy*

*ThS. Khoa Công nghệ thông tin, Trường Đại học Hải Phòng

Received: 28/7/2023; Accepted: 02/8/2023; Published: 15/8/2023

Abstract: Education researchers are increasingly interested in applying data mining approaches, but to date, there has been no overarching exposition of their methodological advantages and disadvantages to the field. This is partly because the use of data mining in education research is relatively new, so its value and consequences are not yet well understood. Yet statisticians, sociologists and those who study computer-based education have discussed the methodological merits of data mining in education research. This article brings together their perspectives, providing an interdisciplinary overview of potential benefits and drawbacks. The benefits, regardless of scholar background, largely emphasize the speed and ease with which data mining approaches can help explore very large datasets. Perceived drawbacks, however, differ based on disciplinary expertise. For example, statisticians question data mining's exploratory nature and non-reliance on sampling theory, while sociologists raise concerns about an excessive reliance on data in research designs and in understandings of education.

Keyword: Data mining, Education research

1. Đặt vấn đề

Khai thác dữ liệu (KTDL) là một quá trình khám phá các mẫu trong dữ liệu một cách có hệ thống và tự động hoặc bán tự động (Witten, Frank và Hall, 2011). Nó thường được tiến hành trên các tập dữ liệu rất lớn, khó có thể kiểm tra đầy đủ thông qua các phương pháp mô tả và suy luận truyền thống. KTDL đã trở thành một chủ đề nóng trong nghiên cứu giáo dục, điều này đặt ra câu hỏi: Các học giả đã xác định những lợi ích và mối quan tâm nào khi sử dụng KTDL trong nghiên cứu giáo dục? Phần tổng quan này giải quyết câu hỏi bằng cách (1) xác định mẫu đại diện cho các quan điểm học thuật về giá trị của việc KTDL đối với nghiên cứu giáo dục, (2) kiểm tra chặt chẽ diễn ngôn để hiểu bối cảnh, động cơ và sự khác biệt của các quan điểm khác nhau và (3) tổng hợp các quan điểm này một cách toàn diện và cô đọng. Các bài viết có liên quan đã được xác định thông qua việc tìm kiếm các tác phẩm được bình duyệt liên quan đến “KTDL” trong cơ sở dữ liệu ERIC vào tháng 8 năm 2015. ERIC, do Bộ Giáo dục Hoa Kỳ tài trợ, đã được chọn vì nó được coi là “cơ sở dữ liệu thư mục quốc gia hàng đầu về tài liệu giáo dục” (Hệ thống Thư viện Đại học Pittsburgh, 2015), và nó chỉ bao gồm các tài liệu tham khảo liên quan đến giáo dục. Việc tìm kiếm được giới hạn ở các tác phẩm được xuất bản giữa năm 2005 and 2015, xác định 137 bài báo tạp chí học thuật và 1 tài liệu ERIC. Trong số đó, có 13 cuộc thảo luận mang tính khái niệm hoặc lý thuyết thực chất về giá trị của việc KTDL như một phương pháp nghiên cứu giáo dục. Các tài liệu mang

tính khái niệm chính được trích dẫn bởi các bài báo này, một số tài liệu giáo dục bên ngoài, cũng đã được kiểm tra khi thích hợp.

Sau khi xác định được các bài báo liên quan, mỗi bài báo đều được đọc cẩn thận để hiểu những tuyên bố chính về tiện ích của việc KTDL trong nghiên cứu giáo dục và những biện minh cho từng tuyên bố. Trong lần đọc chi tiết ban đầu, chúng tôi đã cẩn thận giữ lại những lập luận có vẻ giống nhau và hiểu càng nhiều càng tốt về các giả định cũng như hàm ý của chúng. Trong khi các học giả nhìn chung đồng ý về lợi ích tiềm tàng của việc KTDL, họ nêu ra một loạt mối lo ngại dựa trên chuyên môn của họ. Ví dụ, các học giả có nền tảng thống kê đã xác định các vấn đề khác với những vấn đề được đào tạo về xã hội học khoa học hoặc phân tích học tập.

Các bài báo nhìn chung lạc quan về cách KTDL có thể đóng góp về mặt phương pháp cho giáo dục và nghiên cứu giáo dục; một số có vẻ quá lạc quan (AlShammari, Aldhafiri, & Al-Shammari, 2013; ElAtia, Ipperciel, & Hammad, 2012), và một số tỏ ra chỉ trích (Gasevic, Dawson, & Siemens, 2015; Reimann, Markauskaite, & Bannert, 2014). Đã có sự đồng thuận chung về việc KTDL là gì và tại sao nó được sử dụng, đồng thời có chung cảm giác không thể tránh khỏi về việc sử dụng rộng rãi nó trong giáo dục. Một số so sánh và đối chiếu việc KTDL với thống kê truyền thống (Grover & Mehra, 2008; Zhao & Luan, 2006), hóa ra lại là một khung lý thuyết quan trọng để hiểu được những lợi ích và hạn chế của việc KTDL.

Lợi ích tiềm năng của việc sử dụng KTDL trong nghiên cứu giáo dục

Hầu hết các học giả đều lạc quan về lợi ích mà việc KTDL có thể mang lại cho lĩnh vực này. Một lý do quan trọng cho sự nhiệt tình này là, về mặt lý thuyết, việc KTDL có thể dẫn đến sự hiểu biết sâu sắc hơn về từng người học, từ đó có thể cải thiện trải nghiệm học tập của họ (Berland, Baker, & Blikstein, 2014; Papamitsiou & Economides, 2014). Vì việc học liên quan đến nhiều con đường phức tạp nên các phương pháp có thể giúp phát hiện những khuôn mẫu như vậy có thể đặc biệt có giá trị (Berland và cộng sự, 2014; Martin & Sherin, 2013). KTDL thậm chí có thể cần thiết khi các bộ dữ liệu giáo dục trở nên lớn hơn và phức tạp hơn. Một số người đã chỉ ra rằng với quy mô ngày càng tăng của các bộ dữ liệu giáo dục có sẵn, chúng ta không thể không KTDL (Grover & Mehra, 2008).

KTDL cũng có thể mang lại sự đóng góp độc đáo khác với các phương pháp thống kê truyền thống. Ngược lại với các phương pháp thống kê truyền thống được thiết kế để phân tích các mẫu nhỏ, KTDL được thiết kế để phân tích hiệu quả các tập dữ liệu rất lớn (Grover & Mehra, 2008). Điều này cho phép KTDL để cung cấp thông tin khi nào và như thế nào khi cần (Berland và cộng sự, 2014; Luan & Zhao, 2006) và phát hiện những thông tin bất ngờ (và hữu ích) (Elatia et al., 2012; Thuneberg & Hotulainen, 2006). Một lợi ích độc đáo khác của KTDL là nó có thể giúp phân tích các dạng dữ liệu phi truyền thống theo những cách hiệu quả và hiệu quả. KTDL có thể được áp dụng cho dữ liệu về văn bản, vị trí, âm thanh, hình ảnh, tương tác và quan hệ xã hội (Grover & Mehra, 2008; Papamitsiou & Economides, 2014). Điều này có thể giúp mở rộng phạm vi phân tích của các chuyên ngành giáo dục định tính truyền thống. Lang và Baehr (2012) đã sử dụng việc khai thác văn bản để hiểu rõ hơn về mối quan hệ giữa hướng dẫn soạn thảo bài viết và kết quả học tập của học sinh. Thông qua KTDL, họ có thể phân tích số lượng dữ liệu văn bản lớn hơn so với thông thường trong nghiên cứu giáo dục viết văn và tự tin hơn về kết quả của mình.

Tuy nhiên, các học giả thận trọng chống lại việc sử dụng mù quáng việc KTDL trong nghiên cứu giáo dục. Mối quan tâm nảy sinh từ việc xem xét các nguyên tắc thống kê truyền thống, xã hội học khoa học và từ việc kiểm tra các hoạt động gần đây trong phân tích học tập và KTDL giáo dục. Những mối quan ngại từ quan điểm thống kê truyền thống sẽ được thảo luận trước tiên, vì đây là những vấn đề cơ bản nhưng phức tạp và có thể được chia sẻ rộng rãi bởi nhiều nhà nghiên cứu giáo dục đã xem xét việc KTDL.

2. Nội dung nghiên cứu

2.1. Mối quan tâm từ quan điểm của thống kê truyền thống

Mặc dù có mối liên hệ rõ ràng với số liệu thống kê, nhưng dữ liệu mznzng, thường sử dụng các thuật toán “kỳ lạ”) và dường như đang vận hành hầu hết (y trong hộp đen), đã tạo ra aJair(y mức độ khó chịu cao trong cộng đồng thống kê). việc khai thác tập trung vào việc thiếu lý thuyết trong việc tìm kiếm những dự đoán tốt nhất và do đó, máy tính sẽ nhận được quá nhiều sức mạnh. Điều này trực tiếp(y trái ngược với cách hiểu truyền thống về dữ liệu).

Sự khác biệt về tầm quan trọng của lý thuyết này nhấn mạnh mối quan tâm của các nhà thống kê truyền thống về KTDL (Grover & Mehra, 2008; Zhao & Luan, 2006). Các hoạt động KTDL thường không có nền tảng vững chắc trong nghiên cứu trước đó và do đó ít đóng góp về mặt xác nhận hoặc giải thích lý thuyết. Họ thường không giả định một lý thuyết lấy mẫu, vì vậy họ không thể đưa ra những khái quát thống kê thuyết phục về một dân số lớn hơn. Nếu không dựa vào nền tảng và lý thuyết lấy mẫu, sẽ không có thử nghiệm giả thuyết hoặc giá trị ý nghĩa nào (thường được hiểu là “sự chặt chẽ về mặt thống kê”) gắn liền với kết quả. Cuối cùng, việc KTDL có thể làm tăng khả năng kết luận sai rằng một phát hiện là quan trọng hoặc quan trọng (sự lạm phát của lỗi Loại I). Lỗi như vậy có thể xảy ra do người KTDL có rất ít nền tảng lý thuyết và không biết điều gì là quan trọng hoặc không quan trọng đối với những gì đã biết. Nó cũng có thể xảy ra nếu người KTDL liên tục khám phá cùng một dữ liệu, sử dụng các phương pháp hoặc điều kiện khác nhau, làm tăng khả năng giải thích mối quan hệ giả mạo là hợp lệ hoặc quan trọng.

Tuy nhiên, như Zhao và Luan (2006) giải thích, những hạn chế của KTDL không nhất thiết là có tính tàn phá và thống kê truyền thống cũng giúp làm sáng tỏ lý do tại sao. Đầu tiên, trong khi lý thuyết có thể hướng dẫn các quan sát và mang lại mức độ an tâm rằng những phát hiện quan trọng thực sự tồn tại, nó cũng có thể khiến các nhà nghiên cứu mù quáng không nhìn thấy điều gì là quan trọng, hoặc thậm chí dẫn họ đi sai hướng. John Tukey đã đưa ra ví dụ tương tự về một nhà phân tích dữ liệu như một nhà phân tích dữ liệu. thâm từ “mở ra nhiều ý tưởng, khả năng và đặc điểm riêng” và một nhà thống kê (truyền thống) với tư cách là thẩm phán “kiểm tra và kiểm tra các giả thuyết được xác định rõ ràng” (Tukey, 1962, được Zhao & Luan tóm tắt,

2006, trang 11). Để xây dựng dựa trên sự tương tự của Tukey, các thám tử có định kiến mạnh mẽ về cách tội phạm suy nghĩ và hành động có thể bỏ lỡ những manh mối quan trọng không phù hợp với định kiến của họ hoặc cân nhắc quá nhiều bằng chứng ủng hộ mạnh mẽ quan điểm cụ thể của họ, không giải quyết được vụ án. Điều này cũng áp dụng trong nghiên cứu khoa học xã hội, nơi không phải lúc nào cũng thận trọng khi có quá nhiều giả định về những gì tồn tại và cách mọi thứ hoạt động. Khi cần hiểu một hiện tượng chưa có đầy đủ lý thuyết, bản chất lý thuyết của việc KTDL có thể là một điểm mạnh chứ không phải là điểm yếu.

2.2. Việc KTDL không dựa trên lý thuyết lấy mẫu

Việc KTDL không dựa trên lý thuyết lấy mẫu cũng không đặc biệt đáng lo ngại nếu kỹ thuật này được sử dụng chủ yếu để xây dựng các mô hình cụ thể phản ánh các điều kiện địa phương, thay vì xây dựng sự hiểu biết toàn cầu. Khi các công ty và tổ chức KTDL, mục đích của họ thường là dự đoán thông tin về khách hàng của chính họ và hướng dẫn việc ra quyết định trong tương lai gần. Các tổ chức như vậy thường không quan tâm liệu thông tin đó có đúng hay không, trong toàn bộ ngành và do đó không cần phải lấy mẫu ngẫu nhiên của các công ty. Zhao và Luan (2006) nói thêm rằng việc tạo ra các mô hình toàn cầu, thay vì cụ thể, là "một nhiệm vụ đầy tham vọng và thậm chí không thực tế" (trang 12). Họ nhận xét:

Mô hình là sự đơn giản hóa thực tế và mô hình toàn cầu loại trừ các chi tiết cấp thấp, chỉ tập trung vào mức độ trừu tượng cao tóm tắt cấu trúc dữ liệu vì nó giả định tính đồng nhất trong tổng thể. Một mô hình tổng quát hóa toàn cầu thường chứa ít thông tin chi tiết hơn một mô hình cụ thể. Nhưng thực tế cực kỳ phức tạp, đặc biệt đối với các ngành khoa học xã hội, đầy khó khăn và mơ hồ xuất phát từ những thiếu sót trong đo lường, thiết kế và phân tích.

3. Kết luận

Bản chất chung của các mô hình thống kê truyền thống giải thích ngưỡng chấp nhận thấp của các mô hình thống kê và tại sao các nhà khoa học xã hội không hiếm khi đưa ra các kết quả giải thích ít hơn 20% phương sai trong một biến phụ thuộc. Sự tương phản giữa KTDL và thống kê truyền thống không chỉ đơn giản là thống kê truyền thống đạt được kiến thức tổng quát hơn. Đứng hơn, đó là một sự cân bằng trong đó, "mô hình hồi quy thống kê điển hình sử dụng một vài biến để khái quát hóa cho toàn bộ dân số, [trong khi] KTDL mang lại tiềm năng tận dụng thông tin ở mức độ chi tiết và cụ thể hơn" (Zhao & Luan, 2006, trang 12).

Có một cách khác để suy nghĩ về vai trò của việc lấy mẫu trong bối cảnh KTDL: Miễn là có thông tin và khả năng tính toán cần thiết để phân tích toàn bộ tổng thể thì đơn giản là không cần phải lấy mẫu. Thống kê truyền thống phần lớn được phát triển như một phương tiện thực tế và kinh tế để hiểu một hiện tượng - nó đưa ra lý do biện minh cho việc đưa ra những tuyên bố ngay cả khi người ta chỉ nhìn vào một phần rất nhỏ của nó. Cho đến tận những năm 1970, hầu hết các phân tích thống kê đều được thực hiện bằng tay (Zhao & Luan, 2006), điều đó có nghĩa là có một hạn chế đáng kể về lượng thông tin mà một người có thể đưa vào phân tích một cách hợp lý. Việc thu thập và lưu trữ dữ liệu rất tốn kém, đặc biệt là trước khi việc sử dụng cơ sở dữ liệu điện tử và liên lạc trực tuyến trở thành thói quen, ngăn cản việc phân tích dữ liệu dân số phong phú. Trong nhiều thập kỷ qua, thông tin về dân số ngày càng sẵn có cũng như khả năng tính toán. Các khu học chánh, viện giáo dục đại học, giáo dục tiểu bang và địa phương, sở y tế và dịch vụ xã hội cũng như hệ thống tư pháp hình sự hiện nay thường có hồ sơ điện tử của mọi người từng là một phần của hệ thống của họ. Nhiều câu hỏi nghiên cứu từng yêu cầu lấy mẫu không còn cần đến việc lấy mẫu nữa vì dữ liệu đã có sẵn cho toàn bộ dân số. Mặc dù việc KTDL không yêu cầu người dùng phải tuân theo lý thuyết lấy mẫu nhưng đó không phải là vấn đề nghiêm trọng miễn là dữ liệu được khai thác từ tất cả hoặc hầu hết dân số mà người ta hy vọng có thể hiểu được.

Mối lo ngại về việc thiếu ý nghĩa thống kê gắn liền với kết quả KTDL là một biến thể của mối quan tâm lấy mẫu. Ý nghĩa thống kê là thước đo độ không đảm bảo liên quan đến lỗi lấy mẫu. Trong một số trường hợp, không cần đánh giá khả năng kết quả là do lỗi lấy mẫu.

Tài liệu tham khảo

1. Asterisks (*) denotes articles that were analyzed in the literature review.
2. AlShammari, I. A., Aldhafiri, M. D., & Al-Shammari, Z. (2013). A meta-analysis of educational data mining on improvements in learning outcomes. *College Student Journal* 47(2), 326-333. *
3. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. Washington, D.C: National Research Council and National Academy of Education.
4. Berland, M., Baker, R. S., & Blikstein, P. (2014). Educational data mining and learning analytics: Applications to constructionist research. *Technology, Knowledge and Learning*, 19(1-2), 205-220. *