DỮ LIỆU LỚN VÀ TƯƠNG LAI CỦA THỐNG KÊ VỚI SỰ HIỆN DIỆN CỦA TRÍ TUỆ NHÂN TẠO

Đoàn Dũng[*]

Thống kê đang trải qua những thay đổi sâu rộng do sự phát triển mạnh mẽ của dữ liệu lớn (Big data) và tiến bộ trong trí tuệ nhân tạo (AI). Thống kê truyền thống tập trung vào việc phân tích mẫu nhỏ và thiết kế khảo sát, giờ đây các nhà thống kê phải tận dụng Big data và các kỹ thuật học máy để rút trích thông tin từ các bộ dữ liệu khổng lồ và ngày càng không cấu trúc (Cox & Ellsworth, 1997). Sự chuyển đổi này mang lại cả cơ hội và thách thức cho tương lai của thống kê. Mặc dù thống kê vẫn sẽ cần thiết để phân tích các khối lượng và đa dạng dữ liệu mới, các nhà thống kê cần phải thích ứng kỹ năng và vai trò hoặc rủi ro mất đi tính liên quan. Bài viết này tìm hiểu cách mà Big data và AI đang làm thay đổi thực hành và nghiên cứu về thống kê, đồng thời phác thảo những hàm ý quan trọng cho các nhà thống kê.

Sự trỗi dậy của Big data: Big data đề cập tới các bộ dữ liệu cực lớn và tích lũy nhanh chóng vượt quá khả năng xử lý của hệ thống cơ sở dữ liệu truyền thống (Diebold, 2012). Do số hóa và Internet vạn vật, dự kiến lượng dữ liệu toàn cầu sinh ra hàng năm sẽ đạt 175 zettabytes vào năm 2025 (Statista, 2022). Nguồn gốc của chúng bao gồm các bài viết trên mạng xã hội và giao dịch trực tuyến đến đọc số từ cảm biến và thí nghiệm khoa học. Nơi mà các nhà thống kê từng tập trung vào khảo sát đại diện, Big data giờ đây cho phép quan sát trực tiếp và theo dõi cả quần thể ở quy mô chưa từng có (Fan et al., 2014). Phân tích Big data đòi hỏi các kỹ thuật tính toán thuật tính toán mới. Việc tổng hợp các bộ dữ liệu đa dạng từ nhiều nguồn và rút trích thông tin trong thời gian thực trở nên cần thiết cho năng lực thống kê (Dhar, 2013).

Vai trò ngày càng tăng của AI và học máy: AI đề cập đến các hệ thống có thể thực hiện các nhiệm vụ thường yêu cầu trí tuệ của con người như nhận diện hình ảnh, giọng nói và ra quyết định (Russell & Norvig, 2016). Trong AI, học máy cung cấp các kỹ thuật cho hệ thống tự động phát hiện mô hình và học từ dữ liệu mà không cần lập trình rõ ràng (Mitchell, 1997). Các mô hình học sâu sử dụng mạng nơ-ron đã gặt hái được thành công đáng kể trong các ứng dụng như nhận diện hình ảnh và xử lý ngôn ngữ tự nhiên bằng cách phân tích các bộ dữ liệu không gắn nhãn lớn (LeCun et al, 2015). Các nhà thống kê giờ đây đang áp dụng các cách tiếp cận AI/ML để khám phá Big data không cấu trúc và phát hiện các mô hình ẩn với tốc độ mà qua các thủ tục thống kê truyền thống một mình là không thể (Alpaydin, 2020).

Cơ hội cho các nhà thống kê trong cảnh quan dữ liệu mới: Khi cả khu vực công cộng và tư nhân đều đối mặt với các thách thức quản lý dữ liệu mới phong phú và phức tạp, nhu cầu đối với các nhà thống kê được trang bị kiến thức bổ sung trong cả phương pháp cổ điển và phương pháp mới đang tăng lên (Hardin et al., 2015). Các dự án AI/ML ngày càng yêu cầu các nhà thống kê không chỉ dành cho phân tích mà còn đảm bảo quản trị đúng đắn của các công cụ

[*] *Hội Thống kê Việt Nam*

phân tích và xử lý công bằng các chủ đề nhạy cảm (Yu et al., 2021). Các lĩnh vực nghiên cứu thống kê hoàn toàn mới cũng đang phát triển tập trung vào mô hình dự đoán, thuật toán học sâu và đánh giá hệ thống AI (Giles et al., 2019). Hơn nữa, vẫn còn các thách thức nghiên cứu mở xung quanh khả năng giải thích, minh bạch, công bằng và trách nhiệm từ các hệ thống tự động mà cung cấp các lĩnh vực hấp dẫn cho sự đóng góp thống kê (Rudin, 2019).

Thách thức: Mặc dù hứa hẹn mang lại sự hiểu biết sâu rộng, sự trỗi dậy của Big data và AI tiên tiến đặt ra những mối quan ngại xã hội chính đáng. Khả năng ra quyết định không công bằng hoặc có định kiến được nhấn mạnh ở quy mô lớn yêu cầu cải thiện phương pháp luận cho việc đánh giá công bằng mô hình và tính chung chung (Mehrabi et al., 2019). Vai trò công việc cũng có thể bị gián đoạn khi một số nhiệm vụ thông thường trở nên tự động hóa, yêu cầu đào tạo thống kê phát triển để duy trì tính liên quan, mặc dù các vị trí chuyên môn mới cũng đang được tạo ra đồng thời (Katz & Evans, 2018). Giáo dục thống kê và con đường sự nghiệp do đó phải nhanh chóng thích nghi để trang bị cho các chuyên gia với bộ kỹ năng kết hợp cả lý thuyết đã thiết lập và ngành học tính toán mới nổi (Horton, 2015).

Kết Luận: Cuối cùng, sự tăng trưởng nhanh chóng của dữ liệu và tiến bộ công nghệ đã thay đổi đáng kể cơ hội và yêu cầu trong lĩnh vực thống kê. Khi lượng và loại hình dữ liệu số ngày càng tăng lên, thống kê sẽ vẫn giữ vai trò quan trọng trong việc hiểu biết và xử lý những kho thông tin rộng lớn này. Tuy nhiên, các nhà thống kê cần phải tiếp nhận các công cụ từ trí tuệ nhân tạo và học máy để phân tích và rút trích thông tin từ những bộ Big data hơn và phức tạp hơn nhiều so với trước đây. Mặc dù vẫn còn những thách thức trong việc đảm bảo tính khách quan trong phân tích, minh bạch và chuẩn bị cho nghề nghiệp, sự kết hợp giữa thống kê với Big data và AI đã mở ra những lĩnh vực nghiên cứu mới đầy hứa hẹn và các vai trò chuyên môn sáng tạo. Những nhà thống kê đứng ở tuyến đầu của những tiến bộ này sẽ có vị trí thuận lợi để tạo ra hình thức công nghệ dựa trên dữ liệu cuối cùng có sức ảnh hưởng và mang lại lợi ích cho xã hội.

Tài liệu tham khảo:

1. Alpaydin, E. (2020). Giới thiệu về học máy. MIT Press.

2. Cox, L. A., & Ellsworth, P. C. (1997). Áp dụng phương pháp kiểm soát quy trình thống kê vào dữ liệu khoa học xã hội. Quality & Quantity, 31(4), 329-347.

3. Dhar, V. (2013). Khoa học dữ liệu và dự đoán. Communications of the ACM, 56(12), 64-73.

4. Diebold, F. X. (2012). Quan điểm cá nhân về nguồn gốc và sự phát triển của "Big data": Hiện tượng, thuật ngữ và lĩnh vực.

5. Fan, J., Han, F., & Liu, H. (2014). Thách thức của việc phân tích Big data. National Science Review, 1(2), 293-314.

6. Giles, C. L., Wurzer, D., Ohrndorf, L., & Becker, D. (2019). Nắm bắt AI và thống kê. Bản in trước arXiv arXiv:1909.06642.

7. Hardin, J., Hoerl, R., Horton, N. J., Nolan, D., Baader, G., Buck, S., ... & Westfall, P. H. (2015). Khoa học dữ liệu trong giáo trình thống kê: Sinh viên chuẩn bị để "suy nghĩ với dữ liệu". The American Statistician, 69(4), 343-353.

8. Horton, N. J. (2015). Thách thức và cơ hội cho thống kê và giáo dục thống kê: Nhìn lại, nhìn về phía trước. The American Statistician, 69(2), 138-145.

9. Katz, R. L., & Evans, P. (2018). Ảnh hưởng kinh tế của công nghệ số đối với thị trường lao động. Trong Ảnh hưởng của công nghệ số lên thị trường lao động và nền kinh tế rộng lớn hơn (tr. 65-81). IGI Global.

(Xem tiếp trang 17)

(Tiếp theo trang 20)

10. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Học sâu. Nature, 521(7553), 436-444.

11. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019, Tháng 5). Tổng quan về thành kiến và công bằng trong học máy. Trong Bản Ghi Hội Nghị của IEEE (Vol. 16, No. 5, tr. 120-135). IEEE.

12. Mitchell, T. M. (1997). Học máy. 1997. Burr Ridge, IL: McGraw Hill, 45, 37.

13. Rudin, C. (2019). Dừng giải thích mô hình học máy dạng "hộp đen" cho các quyết định quan trọng và sử dụng mô hình có thể giải thích thay thế. Nature Machine Intelligence, 1(5), 206-215.

14. Russell, S. J., & Norvig, P. (2016). Trí tuệ nhân tạo: một cách tiếp cận hiện đại. Malaysia; Pearson Education Limited.

15. Statista. (2022). Lượng dữ liệu/thông tin được tạo ra toàn cầu từ năm 2010 đến 2025. https://www.statista.com/statistics/871513/worldwide-data-created/

16. Yu, K. H., Beam, A. L., & Kohane, I. S. (2018). Trí tuệ nhân tạo trong y tế. Nature Biomedical Engineering, 2(10), 719-731.