Ứng Dụng Dữ Liệu Lớn và Quản Lý Tài Sản Hạ Tầng Trong Ngành Điện Lực

Lê Thanh Nam [1,*,2,3,4]

[1] *Tiến Sĩ, Kĩ Sư Xây Dựng: Viện Đô Thị Thông Minh và Quản Lý (ISCM), Trường Công Nghệ và Thiết Kế, Đại Học UEH, tp. Hồ Chí Minh - www.iscm.ueh.edu.vn*

[2] *Giám Đốc Kĩ Thuật, Chuyên Gia Soạn Hợp Đồng (FIDIC): Công Ty Tư Vấn ARCADIS, phụ trách thị trường Đông Nam Á và Ấn Độ - www.arcadis.com*

[3] *Giám Đốc: Công ty TNHH ASQ Việt Nam - www.asq.vn*

[4] *Cố Vấn Kĩ Thuật: Công ty EMAPTA - www.emapta.com*

Liên Hệ:
Di Động: +84-9-8378-0100 (Việt Nam) / +91-98-1044-4723 (Ấn Độ) Email: namlt@ueh.edu.vn

TÓM TẮT

Ngành điện lực là một trong những ngành xương sống cho sự phát triển kinh tế của quốc gia. Trong những năm gần đây, các công ty điện lực đã chứng kiến và trải nghiệm nhiều thử thách và sự đổi mới trong cả kĩ thuật và quản lý, và phần nào đã có những bước tiến đáng kể trong công tác chuyển đổi số (digitization) cùng với khái niệm về hệ thống truyền tải điện thông minh(Smart Grids). Các hệ thống điện thường luôn hoạt động trong tình trạng áp lực, chủ yếu là do sự gia tăng hàng năm trong nhu cầu sử dụng điện, sự thiếu hụt về nguồn cung về nhiên liệu/năng lượng để sản xuất ra điện, và các ràng buộc về môi trường áp đặt lên hệ thống sản xuất điện và sự mở rộng các đường dây tải điện.

Ở Việt Nam, tổng công ty điện lực Việt Nam (EVN) cùng các bộ ngành liên quan đã soạn thảo và đề xuất "Qui Hoạch Điện VIII" hướng tới sự phát triển ổn định và bền vững chiến lược đến năm 2050. Một trong những nội dung quan trọng của Qui Điện VIII có chú trọng đến việc phát triển và nâng cao khả năng ứng dụng các công nghệ tiên tiến và áp dụng dữ liệu lớn cũng như trí thông minh nhân tạo vào công tác xây dựng, điều hành và quản lý các hệ thống điện.

Bài viết này cung cấp cho người đọc một cái nhìn tổng quan về các ứng dụng phân tích dữ liệu lớn, công tác quản lý cơ sở hạ tầng trong ngành điện lực, và các vấn đề liên quan đến việc triển khai. Chú trọng của bài viết sẽ là các ứng dụng mới đã mang lại giá trị tích cực trong ngành, một số bài học rút ra từ việc triển khai ứng dụng tại một số tổ chức, và một số ý tưởng và chủ đề cần khám phá. Tổng quan các nghiên cứu khoa học trong ngành khoa học dữ liệu trong ngành điện sẽ được đề cập nhằm tạo ra được cái nhìn tổng quát cho việc ứng dụng và tiếp tục phát triển của ngành trong tương lai. Cuối cùng, bài viết nêu ra các cơ hội và thách thức cũng như các mục tiêu và chiến lược để đạt được các kết quả có ý nghĩa.

Từ Khóa: Dữ Liệu Lớn, Quản Lý Tài Sản, Cơ Sở Hạ Tầng, Độ Tin Cậy, Trí Thông Minh Nhân Tạo

1 GIỚI THIỆU

Có rất nhiều định nghĩa và cách hiểu khác nhau về phân tích dữ liệu lớn trong các ứng dụng ngành điện, đặc biệt là công tác sản xuất, truyền tải, và phân phối. Điều này chủ yếu là vì lý do có khá nhiều cách tiếp cận trong ngành khoa học dữ liệu và mục đích rộng liên quan (ví dụ như việc sử dụng các công cụ toán xác xác thống kê, trí thông minh nhân tạo (AI), học máy). Tính phức tạp và đa dạng của phân tích dữ liệu lớn trong

ngành còn đến từ nhiều cách thức ứng dụng trong các lĩnh vực như quản lý tài sản (Asset Management), điều hành, điều khiển hệ thống, an toàn và an ninh, các quyết định lập kế hoạch và thị trường. Trong đó, có một vấn đề nền tảng làm cho công tác phân tích dữ liệu lớn trong ngành điện lực đặc biệt khó khăn là liên quan đến dung lượng của dữ liệu. Dung lượng của dữ liệu trong ngành thường ở cấp độ Terabyte, lớn hơn rất nhiều cấp độ Petabyte là cấp độ đã được coi là lớn trong các lĩnh vực khác. Cuối cùng, khái niệm "phân tích dữ liệu" còn là khá mơ hồ và đôi khi bị hiểu sai bởi vì hầu hết các ứng dụng cơ bản hay ứng dụng truyền thống trong ngành điện là dựa trên việc xử lý các dữ liệu được đo đạc (measurement data). Công việc xử lý truyền thống này đôi khi không được coi là phân tích dữ liệu lớn trong ngành khoa học dữ liệu. Ví dụ như, các phương pháp truyền thống liên quan đến tính toán ước lượng trạng thái (state estimation) hay vị trí sự cố (outage location) thường không được coi là phân tích dữ liệu lớn nếu công tác phân tích này chỉ đơn thuần dựa vào các phương trình/công thức toán được phát triển trên nền tảng vật lý. Trong khi đó, ở một khía cạnh khác, việc phán đoán sự cố dựa trên các mô hình nền tảng dữ liệu sử dụng các dữ liệu về thời tiết và mất điện có độ phân giải cao thì được coi là phân tích dữ liệu lớn.

Bài viết này không chú trọng đến việc đưa ra một định nghĩa rành mạch về phân tích dữ liệu lớn trong ngành điện mà chỉ chú trọng đến việc tổng quan hóa và nêu ra một số nghiên cứu và ứng dụng gần nhất sử dụng các phương pháp tiên tiến trong ngành khoa học dữ liệu trong ngành điện, đặc biệt chú trọng đến công tác quản lý tài sản cơ sở hạ tầng ngành điện, quản lý sự cố, và tính tích hợp với năng lượng tái tạo.

Có thể nói, một trong những bài báo nghiên cứu đầu tiên liên quan đến việc ứng dụng dữ liệu lớn trong ngành điện là bài báo của tác giả Kezunovic et al. (2013). Tuy nhiên, nghiên cứu này chỉ dừng lại ở dạng mô hình và lý thuyết mà chưa đưa ra ứng dụng cụ thể vào thực tế. Việc ứng dụng vào thực tế mới chỉ được tiến hành trong vài năm gần đây.

Với mục đích hướng tới việc truyền tải kiến thức và đào tạo, ở bài viết này, một số khái niệm cơ bản cấu thành nên công tác phân tích dữ liệu lớn trong ngành điện và các khía cạnh đặc thù của ngành điện sẽ được đề cập.

2 TẦM QUAN TRỌNG VÀ TÍNH KHẢ THI CỦA PHÂN TÍCH DỮ LIỆU LỚN TRONG NGÀNH ĐIỆN

2.1 Sự Tác Động của Phân Tích Dữ Liệu Lớn

Sự thay đổi mạnh mẽ trong ngành điện dường như là chưa có tiền lệ, bao gồm sự chuyển dịch trong phân bổ tỷ lệ các dạng năng lượng (Energy Mix), yêu cầu và nhu cầu ngành càng cao của khách hàng, sự xuất hiện các kĩ thuật tiên tiến và thiết bị mới, và các mô hình kinh doanh mới. Chính những thay đổi này dẫn đến tính phức tạp (Complexity) và tính không chắc chắn (Uncertainty), cũng như mang tới các thách thức và cơ hội mới. Ở trung tâm của việc giải quyết các thách thức chính là việc đưa ra các quyết định tốt hơn và tối ưu hơn trong giai đoạn vận hành và lập kế hoạch, bao gồm cả công tác đầu tư dài hạn và phát triển chính sách.

Như chúng ta đã biết, lưới điện thông minh ngày nay đã và đang tiếp tục được trang bị các thiết bị tiên tiến có khả năng cảm biến (sensing) và thu thập dữ liệu với độ phân giải cao hơn nhiều so với trước đây. Dữ liệu mới kèm theo các phương pháp phân tích mới có thể hỗ trợ các mục tiêu trong quản lý hệ thống điện như việc đạt được độ bền thiết bị cao hơn, hiệu quả kinh tế tốt hơn và giảm khí thải.

Phân tích dữ liệu vì thế trở thành một phần quan trọng giúp doanh nghiệp nâng cao khả năng ứng dụng các công nghệ hiện có nhằm đưa ra các quyết định tối ưu hơn. Các ứng dụng của phân tích dữ liệu cho ngành điện là rất nhiều và có thể được xác định trong nhiều hoạt động của ngành. Do đó, phân tích dữ liệu lớn đã, đang và sẽ là một bước đổi mới quan trọng trong đối trong các tổ chức cơ quan ngành điện lực.

2.2 Nguồn Dữ Liệu

Việc áp dụng các giải pháp phân tích dữ liệu lớn trong các hệ thống phân phối và truyền tải điện khác nhau thường được tập trung vào việc khai thác các nguồn dữ liệu không có tính đồng nhất (heterogeneous), là các khối dữ liệu có chất lượng riêng biệt, độ phân giải không gian/và thời gian khách nhau, và có cách trình bày thông tin khác nhau. Sẽ là khả thi để nâng cao ứng dụng nguồn dữ liệu này thông qua các phương pháp như: (a) kết hợp các nguồn dữ liệu mới và nguồn dữ liệu truyền thống, ví dụ thông qua việc sử dụng lý thuyết hợp nhất dữ liệu (data fusion) (Simões Costa et al., 2013); (b) trích xuất và kết hợp thông tin từ các dữ liệu có định dạng khác nhau (ví dụ như dữ liệu ở hạng ảnh và ở dạng văn bản) thông qua các mô hình đa chiều (multimodal learning) (Srivastava and Salakhutdinov, 2012) hoặc qua một mạng lưới thông tin có tính không đồng nhất (heterogeneous) (Sun et al., 2016); và (c) kết hợp dữ liệu từ các nguồn dữ liệu phân tán về mặt địa lý, ví dụ như việc sử dụng các phương pháp Vector Autoregressive (VAR) truyền thống (Cavalcante et al., 2017) hay phương pháp học sâu (Deep Learning) (Zhu et al., 2020).

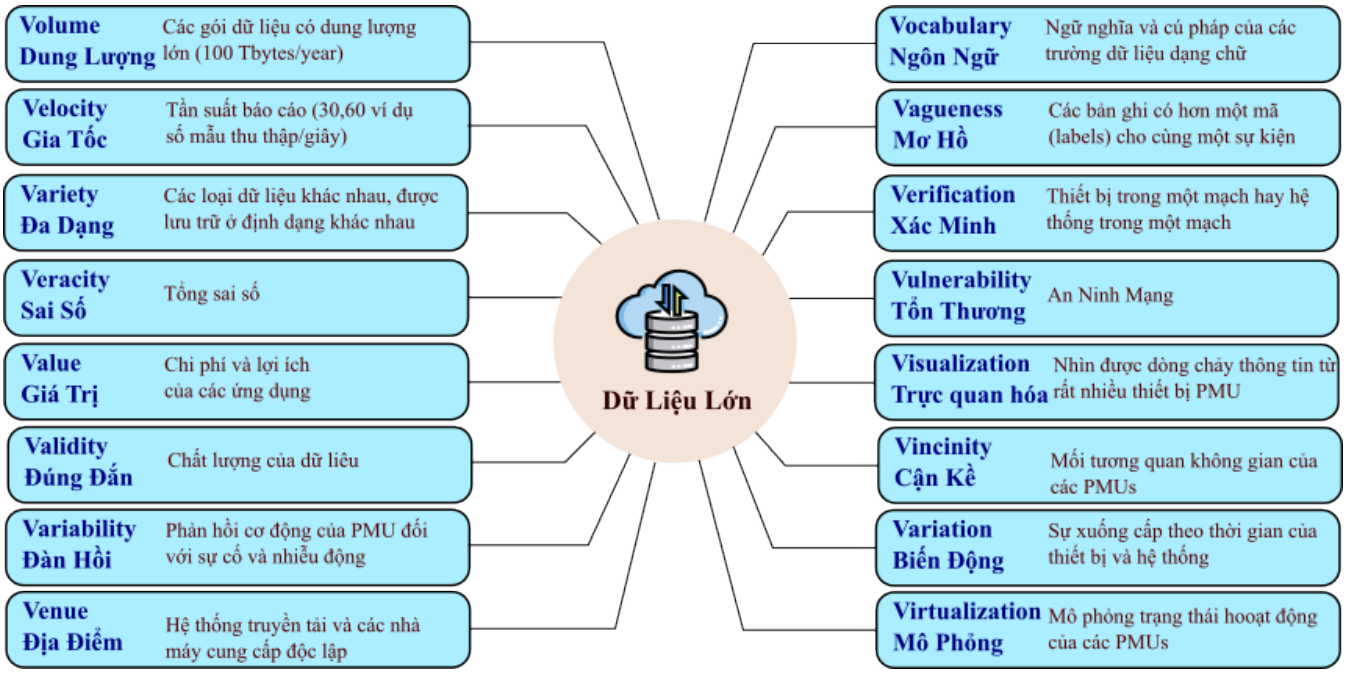
Nguồn dữ liệu đến từ các lĩnh vực khác nhau:

- **Hạ tầng mạng lưới điện**: Các doanh nghiệp quản lý và điều hành hệ thống đang cải thiện và nâng cao khả năng quan sát của hệ thống mạng bằng cách lắp đặt các đơn vị đo pha (PMU) có thể cung cấp dữ liệu với tốc độ báo cáo cao (ví dụ, 30 đơn vị đo lường trên mỗi giây của độ lớn điện áp/dòng điện, pha và tần số) và các thiết bị đầu cuối điểm từ xa (Remote Terminal Units - RTU) tại các trạm biến áp và các công tơ điện thông minh lắp đặt cho khách hàng. Các cảm biến được dùng cho việc giám sát từ xa các trạm biến áp cũng đang được thử nghiệm để theo dõi tình trạng làm việc của thiết bị và cải thiện chất lượng dịch vụ (Leitão et al., 2015).
- **Nhà máy điện tái tạo**: Các nhà máy sản xuất điện năng lượng tái tạo đang tiến hành lắp đặt và vận hành các cảm biến giám sát trong turbin và trong các cánh quạt (có những trường hợp lên đến hàng trăm cảm biến), các cảm biến này thu nhập hơn 10.000 điểm dữ liệu mỗi giây). Các dữ liệu này có thể được dùng cho công tác dự đoán sự cố và bảo trì thiết bị (giảm chi phí vận hành và bảo trì; dữ liệu từ mạng lưới dự báo thời tiết số, các cảm biến phân tán địa lý (ví dụ, cánh quạt điện gió, các cảm biến đo độ chiếu sáng), máy ảnh ghi lại bầu trời và hình ảnh vệ tinh có thể được kết hợp để cải thiện khả năng dự báo công suất (và thời tiết) trong nhiều khung thời gian khác nhau (Sweeney et al., 2020). Trong việc dự báo điện năng tái tạo, quy mô kết nối và sử dụng dữ liệu cũng đã tăng từ một địa điểm đến hàng trăm địa điểm (Messner and Pinson, 2019).
- **Người tiêu dùng và mạng xã hội**: trong khi đang ở giai đoạn đầu của triển khai, sự lan rộng của các thiết bị liên kết Internet trong các căn hộ và tòa nhà thông minh đang tạo điều kiện cho các dịch vụ về năng lượng và phi năng lượng dựa trên dữ liệu (Ahmed et al., 2016), ảnh hưởng của chúng phụ thuộc vào việc giải quyết các thách thức như bảo vệ/quyền riêng tư dữ liệu và tương tác với người tiêu dùng. Hơn nữa, sự gia tăng của mạng xã hội đã cho phép các công ty điện lực hiểu và tương tác với khách hàng tốt hơn trước rất nhiều(Moreno-Munoz et al., 2016). Các nhà nghiên cứu cũng đã thử kết hợp dữ liệu Twitter vào việc phát hiện cục bộ mất điện (Sun et al., 2016).
- **Thị trường điện**: Trong vài năm qua, đặc biệt tại châu Âu, tính minh bạch của thị trường điện đã được cải thiện đáng kể, và sau khi xuất bản Quy định (EU) số 543/2013 (EU, 2013), lượng dữ liệu công khai có sẵn đang tăng lên (Hirth et al., 2018), bao gồm cả truy cập vào các đề xuất riêng lẻ từ các công ty cung cấp điện trên thị trường (thông thường có sẵn với độ trễ vài tháng). Cùng xu hướng đó đang diễn ra ở Mỹ, với các nền tảng như bộ sưu tập dữ liệu Form EIA-930 cung cấp nguồn dữ liệu toàn diện và tập trung cho các dữ liệu hoạt động hàng giờ của lưới điện cao áp tại 48 tiểu bang. Dữ liệu mở này có thể được sử dụng cho các mục đích khác nhau: cải thiện kỹ năng dự báo giá điện bằng cách kết hợp

phân tích giá điện từ các khu vực khác nhau (Ziel and Weron, 2018) hoặc đánh giá ảnh hưởng quy mô lớn của việc phát điện từ nguồn năng lượng tái tạo được cấp bởi lưới điện liên kết với các quốc gia láng giềng (Zugno et al., 2013).

- **Môi trường và thời tiết**: Dữ liệu thời tiết đóng vai trò quan trọng trong việc dự đoán điều kiện hoạt động, bao gồm cả lỗi hỏng hóc thiết bị và đường dây. Dữ liệu từ các trạm thời tiết trên mặt đất, vệ tinh và tài nguyên radar có sẵn từ cơ sở dữ liệu của chính phủ. Mạng lưới cảm biến chuyên dụng, chẳng hạn như mạng lưới phát hiện sét quốc gia ở Mỹ, cũng là nguồn cung cấp dữ liệu thời tiết khá hữu ích. Nhiều dịch vụ dự báo thời tiết cũng có sẵn để cung cấp các tính năng được tính trước của các tập dữ liệu thời tiết. Ngoài ra, dữ liệu về rừng và nông nghiệp, đất đai, di cư động vật và các điều kiện xung quanh khác cũng có thể có sẵn từ nhiều nguồn khác nhau. Cách sử dụng dữ liệu chính xác cao bằng cách sử dụng các cơ sở dữ liệu chuyên dụng như hệ thống (LIDAR - Light Detection and Ranging) hoặc khảo sát bằng thiết bị bay drone cũng được nghiên cứu và đưa vào thực tiễn. Những dữ liệu này thường không được thu thập trong phạm vi chức năng của ngành điện nhưng chúng lại là một nguồn dữ liệu quan trọng cho ngành này. Chẳng hạn, trong dự báo tải, các nghiên cứu đã chuyển từ việc sử dụng nhiệt độ được thu thập tại một trạm đến nhiều biến thời tiết và nhiều trạm thời tiết (Hong et al., 2015). Trong dự báo điện mặt trời, dữ liệu hình ảnh bầu trời được sử dụng nhiều để phát hiện đám mây (Kleissl, 2013).

Các nhà nghiên cứu và chuyên gia hiện nay tập trung vào việc khai thác dữ liệu hiện có và khám phá các nguồn dữ liệu mới và dữ liệu với quy mô lớn để theo đuổi các cải tiến trong lập kế hoạch và vận hành lưới điện. Có nhiều khía cạnh quan trọng khác của phân tích dữ liệu lớn, chẳng hạn như xây dựng thuật toán có thể tận dụng môi trường tính toán hiệu suất cao và mở rộng kích thước mô hình để nắm bắt các đặc trưng chi tiết trong dữ liệu (Wang et al., 2016). Một ví dụ khác là sử dụng dữ liệu tải và thời tiết hàng giờ để dự báo tải dài hạn, thông thường dựa trên dữ liệu hàng tháng (Hyndman and Fan, 2009; Hong et al., 2014). Một bài đánh giá gần đây về phân tích dữ liệu lấy từ các công tơ thông minh liệt kê 10 bộ dữ liệu công khai về nhu cầu điện (Wang et al., 2018). Hình 1 là một ví dụ mô tả các thuộc tính cơ bản của dữ liệu lớn trong ngành điện.



Hình 1. Các thuộc tính cơ bản của dữ liệu lớn trong ngành điện

2.3 Một Số Lưu Ý Quan Trọng Khi Tạo Ra Các Tập Dữ Liệu

Cách tạo các tập dữ liệu như thế nào là khá quan trọng để sử dụng và ứng dụng các công cụ phân tích dữ liệu lớn. Một số cân nhắc sau cần được quan tâm khi thiết lập, tạo lập, và thu thập các gói dữ liệu:

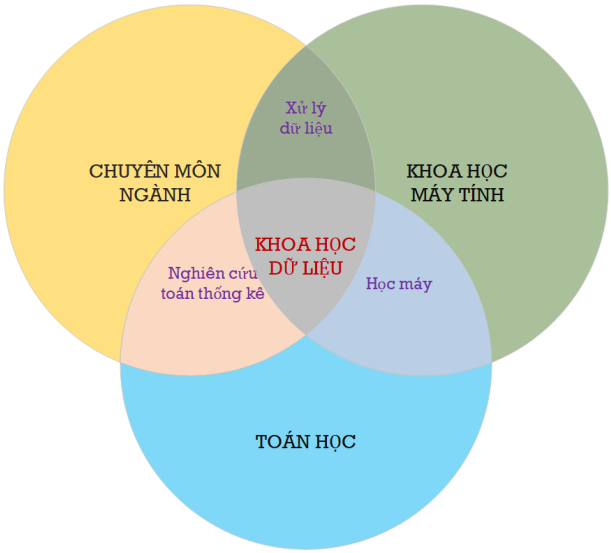
- Tính đồng bộ và mối tương quan theo không gian và thời gian
- Tính mở rộng
- Dữ liệu trống
- Sự đa dạng của dữ liệu xấu
- Các loại dữ liệu có độ tin cậy thấp/có tính biến thiên cao

Cách mà mỗi yếu tố này phản ánh lên các ứng dụng dữ liệu lớn trong lưới điện nằm ngoài phạm vi của bài viết này, nhưng chắc chắn đáng khám phá khi các tập dữ liệu mới được thêm vào và kết hợp với nhau trong quá trình xử lý và phân tích.

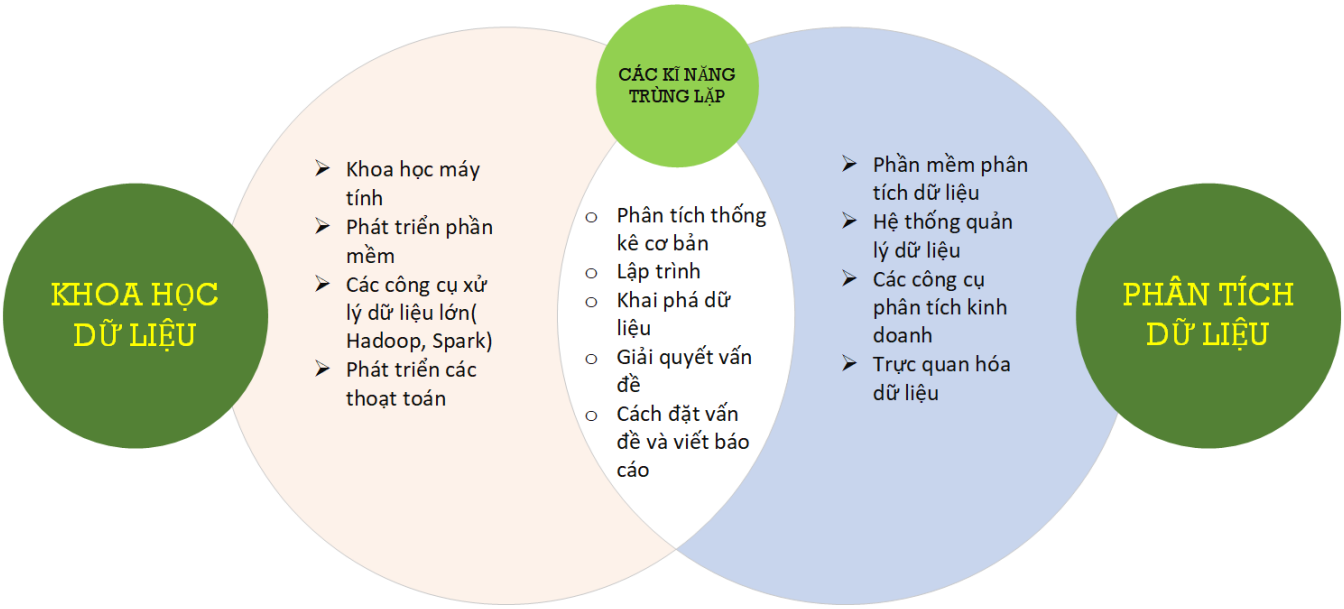
3 THÁCH THỨC CỦA CÔNG VIỆC PHÂN TÍCH DỮ LIỆU LỚN

3.1 Nền Tảng của Khoa Học Dữ Liệu

Mục tiêu của khoa học dữ liệu là trích xuất giá trị từ dữ liệu. Các bước của vòng đời quản lý dữ liệu bao gồm thu thập dữ liệu; tiền xử lý (khám phá, lấy mẫu, giảm chiều/thu gọn các đặc trưng và thuộc tính, tạo đặc trưng và thuộc tính, biến đổi, làm sạch và tích hợp); xử lý phân tích (mô hình hóa, thường bao gồm nhiều khối mô hình xây dựng liên kết với nhau); giải thích và báo cáo kết quả (Aggarwal et al., 2015). Những kỹ năng chính cần thiết trong lĩnh vực này thường được xem là đa ngành ở giao điểm của khoa học máy tính, toán học, thống kê và các lãnh vực khoa học kĩ thuật và xã hội khác (Hình 2). Trên mặt kỹ thuật, các thách thức chính thường liên quan đến dữ liệu lớn, trí thông minh nhân tạo và các phương pháp học máy, trong khi quá trình khoa học dữ liệu áp dụng cũng có thể yêu cầu các kỹ năng khoa học xã hội, giao tiếp và kinh doanh (Hình 3).



Hình 2. Giao thoa các mảng kiến thức - khoa học dữ liệu



Hình 3. Các kĩ năng giao thoa - khoa học dữ liệu

Một cái nhìn toàn diện về khoa học dữ liệu nhấn mạnh rằng "khoa học dữ liệu hơn là sự kết hợp của thống kê và khoa học máy tính" vì "nó yêu cầu đào tạo làm sao để tổng hợp và kết nối giữa các kĩ thuật thống kê và sức mạnh của máy tính trong một khung làm việc lớn hơn, giải quyết từng vấn đề một cũng như giải quyết các câu hỏi cụ thể trong từng lĩnh vực". Các nhà nghiên cứu cũng có cùng quan điểm cho rằng ngành khoa học dữ liệu đòi hỏi: (1) hiểu về bối cảnh của dữ liệu, (2) đánh giá các trách nhiệm liên quan đến việc sử dụng dữ liệu riêng tư và công khai; và (3) giao tiếp rõ ràng về những gì có thể và không thể được suy ra từ một bộ dữ liệu.

Các thành phần cốt lõi của khoa học dữ liệu là các phương pháp dựa trên học máy (Machine Learning - ML) để tìm kiếm các khuôn mẫu (pattern) trong dữ liệu và dựa vào đó để cung cấp thông tin về hiện tượng được mô tả bởi dữ liệu và các dự đoán liên quan đến các sự kiện trong tương lai. Trong học máy, mục tiêu là học/tạo ra một hàm (ánh xạ) mô tả và chứa đựng các dự liệu đầu vào (thường gọi là các biến giải thích - explanatory variable) với tham số đầu ra quan sát được (gọi là biến phản ứng - response variable). Một biểu diễn đơn giản hóa của thực tế được tạo ra cho mục đích này, gọi là mô hình, được sử dụng để ước tính phản ứng chưa biết cho các trường hợp mới dựa trên các biến giải thích quan sát được quan tâm, và quá trình này được gọi là suy luận hoặc đơn giản hơn là dự đoán.

Các mục tiêu của học máy thường được nhóm thành các nhiệm vụ mô tả và nhiệm vụ dự đoán. Các nhiệm vụ mô tả nhằm khám phá các mô hình có thể diễn giải được mô tả dữ liệu trong quá khứ, và các nhiệm vụ dự đoán là những nhiệm vụ mà mục tiêu là xác định các mô hình được quan sát trong dữ liệu huấn luyện để ước tính các dự đoán về các rủi ro và các kết quả khác trong tương lai. Các nhiệm vụ mô tả thường là không giám sát, có nghĩa là chỉ các biến giải thích được xem xét trong phân tích. Các mục tiêu mô tả phổ biến bao gồm phân cụm dữ liệu (Albarakati and Obradovic, 2019; Gligorijevic et al., 2016b), khám phá mối liên hệ (Gligorijevic et al., 2016a), và phát hiện sự khác biệt so với hành vi bình thường, bao gồm phân tích giá trị cực đại, phát hiện giá trị ngoại vi và xác định các mô hình mới (Tan et al., 2016). Các nhiệm vụ dự đoán là giám sát, có nghĩa là chúng không chỉ yêu cầu các biến giải thích mà còn giá trị của biến phụ thuộc đang được dự đoán. Các ví dụ thực tế bao gồm đánh giá rủi ro và chẩn đoán bệnh (Ghalwash et al., 2013).

Có các phương pháp bán giám sát và tự đào tạo, trong đó dữ liệu huấn luyện bao gồm một số dữ liệu có nhãn và nhiều dữ liệu không có nhãn.

Trong phân loại, các biến phản hồi được dự đoán là một lớp (ví dụ, một trong vài loại nhãn dữ liệu), hoặc trong trường hợp hồi quy, đó là một giá trị liên tục. Một trong những phương pháp thường được sử dụng cho phân loại là cây quyết định (decision tree). Thuật toán Hunt (Tan et al., 2016), một trong những phương pháp cây quyết định sớm nhất, đề xuất quy trình tổng quát của việc phân chia dữ liệu dựa trên giá trị của một thuộc tính duy nhất và tiếp tục đệ quy trên các tập con nếu lớp không đủ thuần khiết ở các tập con. Nhiều phương pháp đã được đề xuất để đo độ không thuần khiết của tập dữ liệu và xác định phân nhánh tiếp theo (ví dụ: entropy trong CART (Breiman, 2017) hoặc chỉ số Gini trong ID3 (Quinlan, 1986) và C4.5 (Quinlan, 2014)), cũng như để tỉa cây, từ đó cải thiện khả năng tổng quát của mô hình. Cây quyết định dễ hiểu và không tốn kém nguồn lực để xây dựng và rất nhanh trong việc phân loại các trường hợp không rõ ràng. Chúng cũng khá bền với nhiễu và có thể xử lý các thuộc tính trùng lặp hoặc không liên quan, nhưng không tính đến sự tương tác giữa các thuộc tính. Một trong những hạn chế của cây quyết định là chúng yêu cầu tỉa, nếu không chúng sẽ trở nên quá lớn dẫn đến các vấn đề kĩ thuật như quá khớp (overfit). Giới hạn này được giải quyết thành công bằng Random Forests, được xây dựng như một tập hợp các cây quyết định không tương quan (Breiman, 2001). Tập hợp này được lấy cảm hứng từ phương pháp Bagging, một phương pháp tổng hợp dựa trên lấy mẫu bootstrap, được phát triển để giảm phương sai mà không làm tăng sai số (Breiman, 1996). Trong Random Forests, ý tưởng này được mở rộng hơn nữa bằng cách giới hạn mỗi nút chỉ xem xét một tập con ngẫu nhiên nhỏ của các thuộc tính. Giải pháp kết quả được chứng minh là chính xác hơn so với thuật toán AdaBoost.

Một kỹ thuật thay thế phổ biến có thể xử lý các tương tác giữa các biến giải thích là phân loại một trường hợp mới bằng cách tính khoảng cách đến k láng giềng gần nhất trong tập huấn luyện và dự đoán lớp dựa trên đa số hoặc đa số có trọng số của các láng giềng được xác định kết quả của chúng. Đây là một phương pháp học ít tốn kém vì mô hình không được xây dựng rõ ràng và thời gian suy luận cần thiết để phân loại một trường hợp mới khá lớn. Nó cũng yêu cầu so sánh mỗi điểm dữ liệu mới với mỗi điểm dữ liệu trong tập huấn luyện. Ngoài ra, kỹ thuật này không dễ sử dụng khi nhiều giá trị thuộc tính bị thiếu, vì trong những trường hợp như vậy, phương pháp dựa trên khoảng cách để xác định láng giềng gần nhất có thể không đáng tin cậy. Một số trong số những giới hạn này có thể được vượt qua bằng cách sử dụng nhánh mạng gần nhất (proximity graphs), trong đó các nút mạng được kết nối nếu các điều kiện hình học nhất định được đáp ứng. Trong một công thức như vậy, các thuật toán đồ thị hiệu quả khác nhau (ví dụ: cây khung nhỏ nhất và tam giác hóa) có thể được sử dụng để xác định láng giềng gần nhất và tương quan hơn.

Một phương pháp phân loại thay thế toán học nghiêm ngặt hơn là ước tính xác suất hậu nghiệm của lớp mục tiêu bằng cách sử dụng định lý Bayes. Một phương pháp khá đơn giản nhưng trơn tru và mạnh mẽ, được gọi là phương pháp Naive Bayes. Phương pháp này giả định rằng các giá trị thuộc tính độc lập có điều kiện với nhau, cho trước nhãn lớp y. Trong trường hợp như vậy, xác suất có điều kiện của lớp của tất cả các thuộc tính có thể được phân tách thành tích của xác suất có điều kiện của mỗi thuộc tính. Phương pháp này ngăn được nhiễu, giá trị bị thiếu và các thuộc tính không liên quan. Tuy nhiên, sự độc lập có điều kiện giữa các biến giải thích là một giả định mạnh mẽ không hợp lệ trong nhiều ứng dụng. Đối với các kịch bản như vậy, một lớp mô hình mạng lưới xác suất được gọi là Mạng tin cậy Bayesian đã được phát triển bằng cách mô hình hóa các phụ thuộc có điều kiện thông qua các mạng vô hướng và có hướng. Sự suy luận chính xác trên các mạng như vậy gọ là NP-hard (Cooper, 1990), do đó, các ứng dụng của chúng bị giới hạn cho số lượng thuộc tính nhỏ hơn hoặc cho các loại cấu trúc mạng lưới đặc biệt.

Một phương pháp phân loại theo xác xuất cũng khá hiệu quả là phương pháp hồi quy logistic. Khá nhiều ứng dụng sử dụng phương pháp này để dự đoán xác xuất mất điện liên quan đến thời tiết đã được triển khai.

Điểm mạnh của phương pháp hồi quy logistic so với phương pháp láng giếng gần (k-nearest neighbour) là nó có thể giải quyết các vấn đề có nhiều chiều, bởi vì phương pháp này không dựa vào việc đo lường sự giống nhau giữa các điểm dữ liệu. Một lợi ích khác của phương pháp này là các tham số trọng lượng tương ứng với các thuộc tính riêng lẻ và do đó có tính giải thích khá dễ dàng. Tuy nhiên, sự hiện diện của một số lượng lớn các thuộc tính không liên quan là một thách thức đối với hồi quy logistic, và phương pháp này không áp dụng cho việc phân loại các trường hợp với giá trị rỗng, điều này có thể là một hạn chế lớn khi áp dụng trong thực tế với gói dữ liệu tồn tại số lượng lớn giá trị rỗng.

Mô hình hồi quy logistic có thể được xem như một trường hợp của mô hình tuyến tính tổng quát. Các mô hình mạnh mẽ khác trong danh mục này bao gồm Support Vector Machines (SVM) và Mạng neural đa tầng (Multilayer Neural Network - MNN). Trong phương pháp SVM, vấn đề tối ưu hóa được định hình như tìm kiếm ranh giới lớn nhất phân tách siêu mặt phẳng cho một khu vực lớn tồn tại ở mỗi bên của ranh giới quyết định (Cortes and Vapnik, 1995). Điều này được định hình như một vấn đề lập trình phi tuyến tính bị ràng buộc được biểu thị dưới dạng một hàm của các hệ số của siêu mặt phẳng phân tách, được giải quyết bằng phương pháp bội số Lagrange. Đối với phân loại phi tuyến tính, dữ liệu được chuyển đổi một cách ngầm định thành không gian đa chiều, nơi vấn đề có thể được phân tách theo dạng tuyến tính. Điều này được đạt được bằng cách sử dụng thủ thuật Kernel, để giảm bớt vấn đề thành tình huống phân loại tuyến tính. Sử dụng các Kernel được chọn cẩn thận (Gaussian, đa thức, hoặc sigmoid) cho phép xấp xỉ các ranh giới quyết định tùy ý. Các lợi ích chính của phương pháp SVM là nó chống lại nhiễu và giảm thiểu overfitting trong khi tìm kiếm giá trị nhỏ nhất toàn cục của hàm mục tiêu. Tuy nhiên, chi phí tính toán của SVM là cao, và vẫn là một thách thức để sử dụng mô hình này khi các biến mô tả bị thiếu một phần trong dữ liệu quan sát.

Mạng nơ-ron tiến thẳng đa tầng (Multiplayer Feed-Forward Neural Network - FNN) cũng được sử dụng thành công cho phân loại trong nhiều ứng dụng khác nhau (Zhang, 2000). Ví dụ như việc áp dụng thành công trong việc huấn luyện FNN để phân biệt giữa dòng vào từ biến áp và dòng lỗi (Perez et al., 1994). Mô hình này có ít nhất một lớp đơn vị ẩn, mỗi lớp tính toán một hàm phi tuyến mượt và khả vi của tổng đầu vào có trọng số (ví dụ, hàm sigmoid). Trong mô hình này, vấn đề cập nhật các tham số khi phát hiện lỗi ở đầu ra thường được giải quyết bằng cách truyền ngược lỗi từ đầu ra về các lớp trước đó. Trong quá trình này, lỗi của một nút trong lớp ẩn được ước tính dưới dạng một hàm của các ước tính lỗi và trọng số trong các nút trong lớp trước đó, và giá trị này được sử dụng để cập nhật trọng số của nút ẩn này bằng cách tính toán độ dốc lỗi đối với các trọng số trong nút (Werbos, 1994). FNN có thể xấp xỉ các hàm tùy ý, và do đó mạnh mẽ hơn SVM. Tuy nhiên, khi thiết kế một mạng, overfitting phải được giải quyết cẩn thận. Ngoài ra, nhiễu trong dữ liệu có thể gây ra vấn đề huấn luyện, vì mô hình có thể hội tụ đến một cực tiểu cục bộ, và quá trình huấn luyện có thể yêu cầu một thời gian dài, do đó giới hạn tới việc triển khai cho các ứng dụng thực tế. Vấn đề khác với phương pháp FNN truyền thống là học các mạng sâu rất khó, do tác động kết hợp của việc bão hòa hàm kích hoạt sigmoid khi truyền ngược các lỗi nhỏ, dẫn đến sự hội tụ rất chậm. Tiến bộ lớn đã được đạt trong việc giải quyết vấn đề này, được gọi là vấn đề gradient biến mất (Vanishing Gradient Problems), trong những năm gần đây. Điều này, cùng với tiến bộ trong cơ sở hạ tầng tính toán phân tán dựa trên GPU và sự có sẵn của các tập dữ liệu rất lớn, đã cho phép phát triển các mạng nơ-ron sâu hiệu quả, vượt trội hơn rất nhiều so với tất cả các phương pháp truyền thống, bao gồm thị giác máy tính (Computer Vision), xử lý ngôn ngữ tự nhiên, phát âm. trong những năm gần đây, nhiều kiến trúc học sâu đã được đề xuất và phát triển để xử lý dữ liệu với nhiều thuộc tính khác nhau, tiêu biểu trong số đó là phương pháp mạng nơ-ron tích chập (Convolutional Neural Network) áp dụng để xử lý dữ liệu toàn mạng lưới điện (ví dụ như các dữ liệu về ảnh) và mạng nơ-ron hồi quy cho chuỗi dữ liệu thời gian (Werbos, 1994; LeCun et al., 2015).

3.2 Khía Cạnh Kĩ Thuật

Có một thực tế đã được nhiều chuyên gia và nhà nghiên cứu tổng kết là, thực tế làm các công việc phân tích và mô hình liên quan đến xử lý dữ liệu lớn thường chỉ chiếm khoảng 10% tổng số công việc (số giờ, tài nguyên) liên quan đến cả quá trình xử lý dữ liệu. Có đến 90% công việc lại trực tiếp liên quan đến thiết lập quy trình làm việc, quản lý dữ liệu (và lưu trữ) cũng như các khía cạnh tính toán. Do đó, việc quan sát và tập trung nguồn lực vào một số khía cạnh liên quan đến kỹ thuật trong phân tích dữ liệu lớn là rất quan trọng. Chúng đã được xác định là các thách thức chính cho sự thành công của phân tích dữ liệu lớn (Wickham and Grolemund, 2023).

Ở trung tâm của khái niệm phân tích dữ liệu lớn cơ bản là việc chúng ta cho rằng dữ liệu cần được xử lý là "lớn". Để định nghĩa đúng dữ liệu lớn và các tính năng cần thiết của nó, người đọc được giới thiệu đến (De Mauro et al., 2016). Các ví dụ điển hình liên quan đến việc thu thập dữ liệu PMU, cũng như dữ liệu có độ phân giải cao ở mức tài sản, (ví dụ, từ các tuabin gió, bộ biến tần điện từ (PV), công tơ thông minh, v.v.). Tốc độ thu thập dữ liệu này ở các thang thời gian từ giây đến phút, và đồng thời ở nhiều vị trí địa lý khác nhau, trong khi cũng bao gồm nhiều loại biến số khác nhau. Nói chung, các dữ liệu như vậy trong lưới điện bao gồm các dữ liệu đo lường điểm, hình ảnh và có thể là văn bản. Một số khía cạnh của dữ liệu lớn này cho các hệ thống điện, từ thách thức đến ứng dụng, đã được Arghandeh and Zhou (2017) đề cập gần đây .

Việc đảm bảo tính của giao tiếp dữ liệu, đảm bảo tính toàn vẹn của dữ liệu, cũng như đảm bảo chất lượng dữ liệu, là các bước cần thiết và quan trọng trước khi tiến hành thiết kế và triển khai một giải pháp dựa trên dữ liệu (Cai and Zhu, 2015). Ngược lại, những khía cạnh liên quan đến tính sẵn có và chất lượng dữ liệu đã được xem xét trong một thời gian khá lâu bởi cộng đồng khoa học máy tính (Batini et al., 2009). Tiếp theo đó là việc làm sạch dữ liệu và sửa đổi các bộ dữ liệu có liên quan, mặc dù người ta nên nhận thức rằng những hành động này thực sự có thể ảnh hưởng đến thông tin ban đầu trong các bộ dữ liệu. Đối với các ứng dụng trong các thị trường điện liên quan đến năng lượng tái tạo, một vấn đề khá truyền thống liên quan đến ví dụ là điền vào các khoảng trống trong chuỗi thời gian, tức là có thể có các khoảng thời gian mà dữ liệu không có sẵn, do các lỗi trong việc đăng nhập, lưu trữ hoặc truyền dữ liệu. Để điền vào những khoảng trống này, đã có nhiều nhiều phương pháp được đưa ra như việc tận dụng dữ liệu xung quanh khoảng thời gian đó, hay dữ liệu với phụ thuộc không gian thời gian (đặc biệt liên quan đến việc tạo ra năng lượng tái tạo dựa trên thời tiết), tính sẵn có của dữ liệu ở các mức tổng hợp khác nhau, cũng như mối quan hệ vật lý giữa các biến số quan tâm.

Ngoài những khía cạnh liên quan đến dữ liệu, các phương pháp dựa trên dữ liệu được sử dụng trên các tập dữ liệu lớn đòi hỏi năng lực tính toán đáng kể để giải quyết các vấn đề mô phỏng và tối ưu hóa liên quan. Chẳng hạn như việc sử dụng phương pháp tính toán hiệu năng cao (High Performance Computing - HPC) cho việc tập trung hóa xử lý. Phương pháp HPC ngày càng trở nên phổ biến trong các ứng dụng phân tích lưới điện và thị trường điện (Khaitan and Gupta, 2012).

Phân tích dữ liệu lớn cần được đặt trong khung quản lý tổng thể để giải quyết một hay nhiều vấn đề cụ thể. Trong thực tế, đó chỉ là một công cụ bổ sung để hỗ trợ hoạt động và ra quyết định. Do đó, trước khi đầu tư vào các thiết lập dữ liệu lớn cụ thể và các công cụ phân tích, vấn đề và phương pháp giải quyết vấn đề liên quan cần được xác định rõ ràng. Ví dụ, nếu các dự báo được sử dụng để hỗ trợ ra quyết định, loại dự báo (ví dụ, xác định hoặc xác suất) và các sản phẩm dự báo (độ phân giải, chuẩn hóa, vv.) nên được quyết định dựa trên vấn đề quyết định cụ thể. Ngoài ra, khi chuyển từ dữ liệu sang dạng thứ cấp để phục vụ phân tích, thường nên tạo ra một lớp trích xuất/chú giải đúng loại và mức thông tin từ dữ liệu thô để về sau dễ liên hệ.

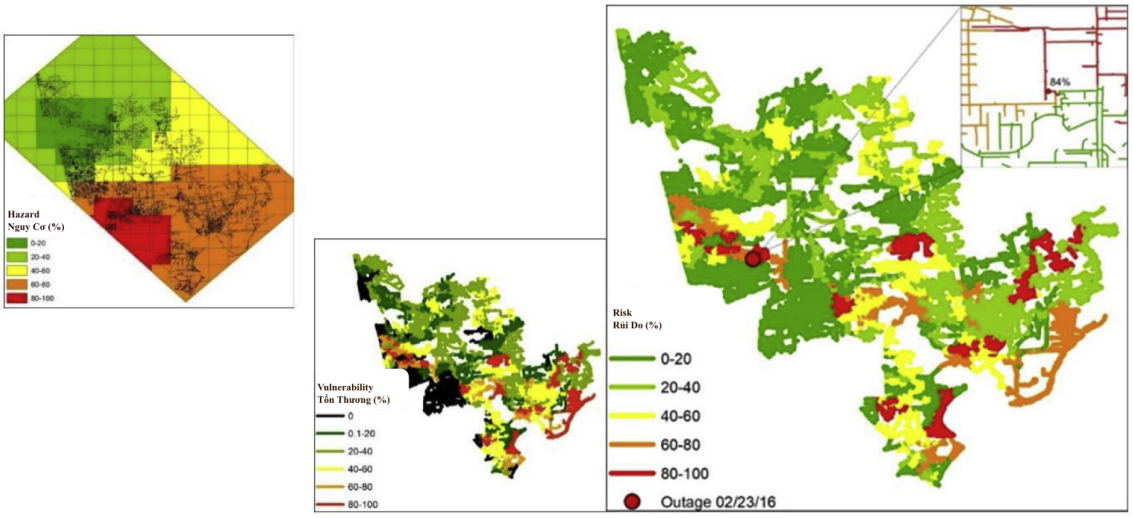
Các yêu cầu bổ sung có thể mang lại một cấp độ phức tạp khác cho phân tích dữ liệu lớn. Yêu cầu quan trọng liên quan đến dữ liệu chính nó (và các luồng dữ liệu liên quan) là cách xử lý an ninh mạng và quyền

riêng tư trong lưới điện. Hiện nay, an ninh mạng đại diện cho một thành phần quan trọng của hệ thống năng lượng phân phối trong tương lai, mà trên đó phân tích dữ liệu lớn có thể được thực hiện (Li et al., 2017). Do đó, công tác thiết lập cho việc phân tích dữ liệu lớn, cũng như các công cụ được sử dụng, cần phải mạnh mẽ để có thể chịu đựng được việc loại bỏ các dữ liệu quan trọng hoặc làm giả dữ liệu. Ngoài ra, quyền riêng tư dữ liệu đang trở thành mối quan tâm ngày càng tăng vì nếu dữ liệu đang được thu thập được chia sẻ, có thể suy ra thông tin các nhân về các tài sản hoặc người tiêu dùng cụ thể. Mối quan tâm về quyền riêng tư đặc biệt hiện nay khi các công tơ thông minh được triển khai rộng rãi, cho phép một số người có thể có thông tin về thói quen tiêu dùng của người khách hàng để tiến hành các hoạt động tiếp thị và tội phạm.

3.3 Khung Đưa ra Quyết Định

Việc sử dụng phân tích dữ liệu lớn tất yếu sẽ dẫn đến việc nâng cao khả năng ra quyết định. Do đó, việc trực quan hóa trong toàn bộ quá trình xử lý nên được đặc biệt quan tâm vì các nhà quản lý và điều hành thường đưa ra các quyết định dựa trên các con số và bảng biểu hay đồ thị được trực quan hóa.

Trong trong những vấn đề giành được sự quan tấm lớn hiện nay, không những chỉ trong ngành điện, mà còn trong công tác quản lý hạ tầng nói chung đó là việc quản lý rủi do. Trong công tác quản lý rủi do, ngoài việc đưa ra các mức rủi do mang tính chất định tính, các nhà quản lý cũng cần phải yêu cầu các kĩ sư và nhà phân tích đưa ra các con số định lượng và trực quan hóa các nguy cơ gây ra rủi do, mức độ tổn thương mạng lưới và rủi do cho từng bộ phân và khu vực trên toàn mạng lưới (Hình 4). Như ví dụ ở Hình 4, một bản đồ rủi do được thiết lập dựa trên việc tích hợp của hai bản đồ nền tảng là bản đồ nguy cơ và bản đồ tổn thương.



Hình 4. Bản đồ thể hiện nguy cơ, sự tổn thương, và rủi do của mạng lưới

Như chúng ta đã thấy, việc đánh giá rủi do là khá phức tạp, đặc biệt là trên phương diện mạng lưới. Do đó, để tính toán được một cách có hệ thống và đạt được một mức độ chính xác nhất định, chúng ta cần phải xây dựng một khung công việc (framewwork) hay một dòng chảy công việc (work flow) nền tảng. Các khung công việc hay dòng chảy công việc này cần được thiết lập cho từng cấp độ khác nhau. Ở phạm vi quốc tế, ví dụ chúng ta có thể thấy đó là khung công việc cho đánh giá rủi do được đưa ra bởi Văn phòng Cứu trợ Khẩn cấp của Liên Hợp Quốc (UNDRO), hay Văn phòng Liên Hợp Quốc về Giảm thiểu Rủi ro Thiên tai (UNISDR). Ở mức độ quốc gia, chúng ta có thể lấy ví dụ về các tiêu chuẩn được đưa ra bởi Cơ quan Quản lý Khẩn cấp Liên bang (FEMA) của Hoa Kỳ.

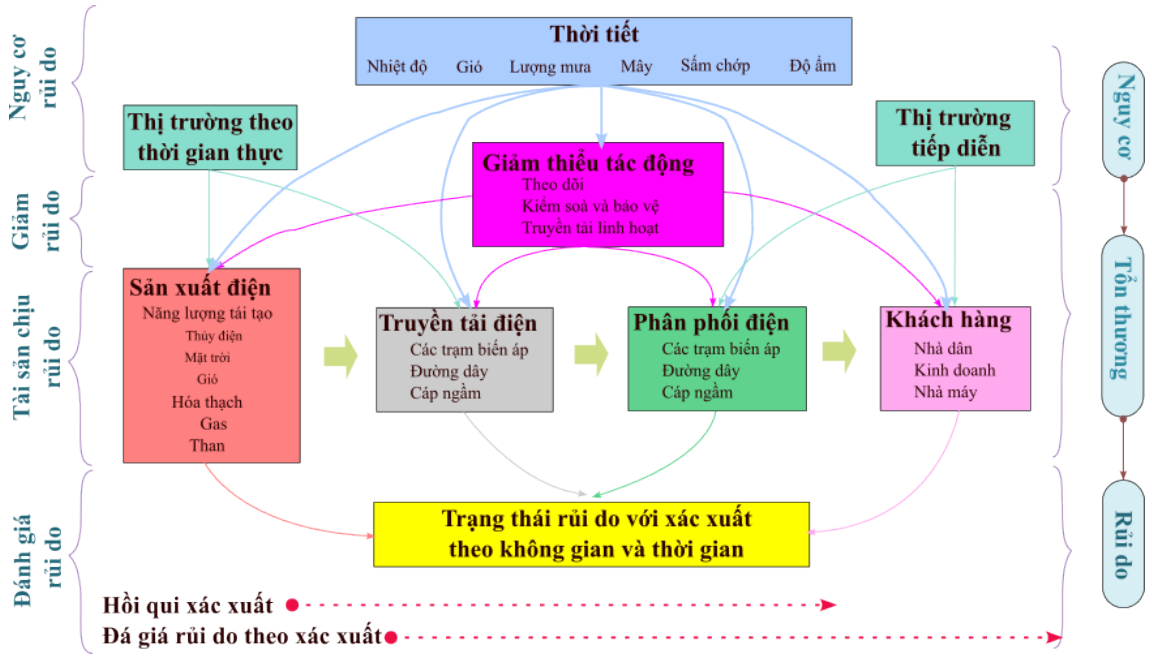
Rủi được thể hiện qua công thức tính toán đơn giản sau.

$$\mathbf{R}isk = \mathbf{H}azard \times \mathbf{V}ulnerability \times \mathbf{C}onsequences \tag{1}$$

Nguy cơ (H) và sự tổn thương (V) có thể được xác định bằng xác xuất. Nguy cơ có thể được tính toán dựa trên xác xuất của một sự kiện nào đó (chẳng hạn như động đất, lũ lụt, mưa lớn) xảy ra với mức độ hay độ mạnh khác nhau (thường gọi là Intensity). Tổn thương (V) liên quan trực tiếp tới yếu tố bản thân của thiết bị tài sản, như xự xuống cấp hay các thông số thiết kế và độ mạnh của nguy cơ. Ở nhiều góc độ khác, tổn thương còn liên quan đến các yêu tố quản lý và sự sẵn sàng của doanh nghiệp trước các sự cố và hiện tượng bất thường.

Thiệt hại (C) cần phải được tính toán cụ thể và thường được đo bởi đơn vị tiền tệ. Ở đây, thiệt hại nên được hiểu theo nghĩa rộng, tức là không chỉ dừng lại ở việc đo lường thiệt hại cho doanh nghiệp, mà còn phải đo lường thiệt hai cho cộng đồng và cho môi trường. Để tính toán và lượng hóa được thiệt hại cần phải xây dựng một hệ thống hay cấu trúc thông xuốt để loại trừ cách tính trùng lặp (thường được hiểu là thiết lập một Impact Hierachy). Một số phương pháp tiêu biểu có thể được áp dụng để tính toán H, V, và C là phương pháp Fault-Tree Analysis (FTA), hay còn gọi là phân tích lỗi theo dạng cây và phương pháp Event Tree Analysis (ETA) hay còn gọi là phân tích sự kiện theo dạng cây.

Việc tính toán rủi do là quan trọng, nhưng cũng chỉ là một bước nếu đặt trong một mô hình toán tối ưu (Optimization Model) khi mà các nhà quản lý cần quan tâm đến nguồn lực nội tại cũng như các khó khăn phải đối mặt (Constraints). Do đó, sau khi tính toán được rủi do, cần xây dựng mô hình tối ưu với một hàm mục tiêu cụ thể và các giàng buộc đi kèm. Hình 5 mô tả khái quan một khung làm việc cho ví dụ về dự báo mất điện mạng lưới gây ra bởi thời tiết.



Hình 5. Khung ra quyết định cho việc đánh giá rủi do

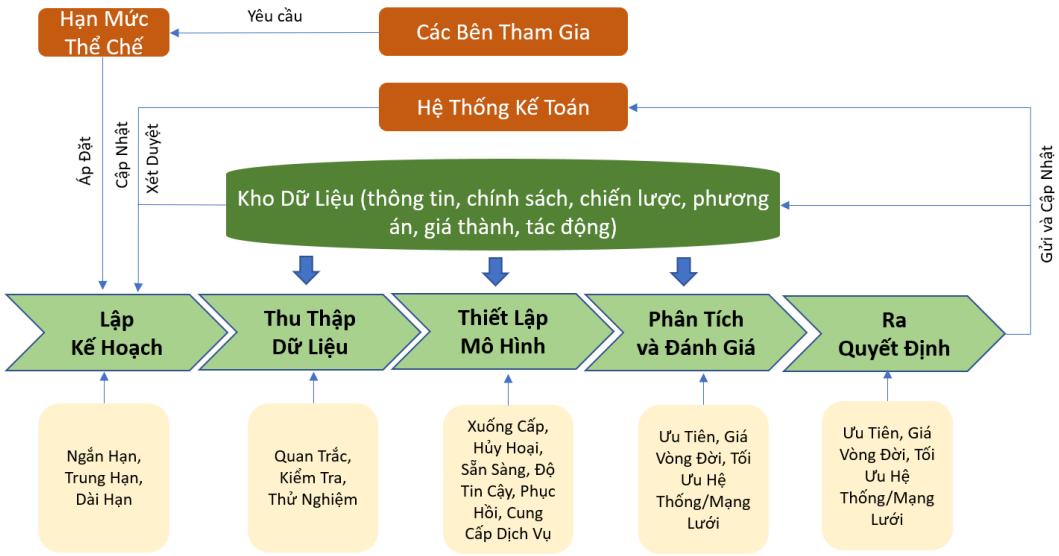
4 ỨNG DỤNG

4.1 Quản Lý tài Sản Cơ Sở Hạ Tầng và Sự Cố Mất Điện

Tại các thị trường có sự cạnh tranh về điện, tư nhân hóa và các yêu cầu kỹ thuật hoặc quy định bắt buộc các công ty điện phải tối ưu hóa công tác quản lý và vận hành. Trong thực tế hoạt động của công ty điện, quản lý tài sản và quản lý sự cố mất điện được xử lý bởi các nhóm khác nhau và thường được đề cập trong việc lập kế hoạch từ ngắn hạn đến dài hạn. Việc quản lý tài sản và quản lý sự cố mất điện có mối tương quan với nhau và chính vì thế việc sử dụng dữ liệu của cả hai nhóm này trên cùng một hệ thống và nền tảng phân tích sẽ là một trong những ưu tiên nên được đầu tư và quan tâm thích đáng.

Các công ty và doanh nghiệp điện lực thường chú trọng nhiều hơn đến việc quản lý sự cố mất điện chứ chưa có đầu tư đích đáng vào quản lý tài sản. Ở các nước phát triển, có một thực tế là việc quản lý tài sản nói chung và ngành điện nói riêng cũng chỉ được quan tâm trong vòng 15 năm trở lại đây. Đã có khá nhiều phần mềm quản lý tài sản, gọi chung là CMMS (Computerized Maintenance and Management Systems) đã được phát triển và đưa vào sử dụng. Nhưng ở các nước đang phát triển, trong đó có Việt Nam, quản lý tài sản là một khái niệm tương đối mới và chưa có một trường đại học nào đào tạo về ngành này.

Về cơ bản, theo tác giả, quản lý tài sản có thể được định nghĩa là "**Tối ưu hóa việc phân bổ nguồn tài nguyên có hạn cho công tác làm mới hay duy tu vả bảo dưỡng các thiết bị và hệ thống sẵn có. Việc tối ưu hóa phải được tính đến lợi ích và ảnh hưởng của doanh nghiệp cũng như của các bên liên quan và môi trường**". Hình 6 mô tả các qui trình và công đoạn cần thiết cấu thành lên một hệ thống quản lý tài sản hoàn chỉnh.



Hình 6. Qui trình và công đoạn trong quản lý tài sản hạ tầng

4.1.1 Phân Loại Tài Sản

Quản lý tài sản trong các hệ thống điện có thể được phân loại rộng rãi thành bốn loại chính dựa trên các thời gian thời gian thực, ngắn hạn, trung hạn và dài hạn.

Quản lý tài sản thời gian thực chủ yếu bao gồm các nguyên tắc bền vững của hệ thống điện chính và xử lý các sự cố gián đoạn không đáng kể của thiết bị hệ thống điện và sự cố gián đoạn lưới điện. Bằng cách cải

thiện nhận thức tình hình, các nhà điều hành lưới điện có thể giám sát và kiểm soát hệ thống một cách hiệu quả. Quản lý tài sản ngắn hạn cố gắng tối đa hóa tỷ lệ lợi nhuận liên quan đến đầu tư tài sản. Giá trị chủ yếu phụ thuộc vào giá thị trường không chắc chắn thông qua việc điều chỉnh và tác động vào thị trường. Đánh giá rủi ro thị trường là một yếu tố quan trọng và phân phối doanh thu/lợi nhuận được đạt thông qua phân tích tính lợi nhuận. Lập lịch bảo trì tối ưu thuộc quản lý tài sản trung hạn. Nó hướng dẫn kế hoạch bảo trì đến các mục tiêu mong muốn của toàn hệ thống một cách đầy đủ.

Các nỗ lực được tập trung vào tối ưu hóa phân bổ tài nguyên tài chính hạn chế ở những nơi và thời điểm cần thiết để quản lý gián đoạn tối ưu mà không đánh đổi tính đáng tin cậy của hệ thống. Sử dụng rộng rãi các công nghệ cảm biến thông minh và giám sát để đánh giá tình trạng làm việc và độ tin cậy của thiết bị và hệ thống theo thời gian và tối ưu hóa các kế hoạch bảo trì tương ứng. Đầu tư vào kế hoạch mở rộng hệ thống điện, cũng như triển khai rộng rãi các nguồn điện phân tán, thuộc phạm vi của quản lý tài sản dài hạn, trong đó các nhà đầu tư, các đối thủ cạnh tranh và các bên liên quan khác được mời tham gia vào các kế hoạch kinh tế trong tương lai.

4.1.2 Tác Động của Thời Tiết và Sự Cố Mất Điện

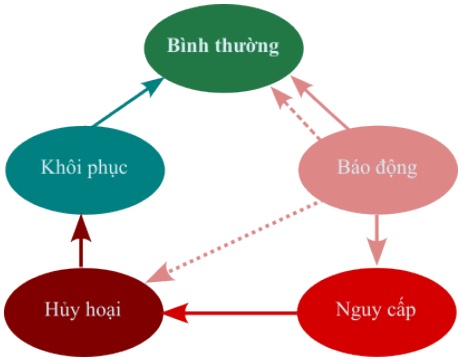
Tác động của thời tiết đối dẫn đến sự cố mất điện và gián đoạn trong hệ thống điện có thể được phân loại thành trực tiếp và gián tiếp:

- Tác động trực tiếp đối với tài sản hạ tầng điện: Loại tác động này bao gồm tất cả các tình huống khi điều kiện thời tiết nghiêm trọng trực tiếp gây ra sự cố với tài sản. Ví dụ như sét đánh vào thiết bị, tác động của gió khiến cho cây hoặc cành cây chạm vào đường dây, v.v. Những sự cố này được đánh dấu là sự cố do thời tiết gây ra.
- Tác động gián tiếp đối với tài sản hạ tầng điện: Loại tác động này xảy ra khi thời tiết tạo ra một tình huống trong mạng lưới mà gián tiếp gây ra sự cố của thiết bị. Các ví dụ bao gồm: điều kiện thời tiết nóng làm tăng nhu cầu, gây quá tải cho đường dây dẫn đến giảm độ cao của đường dây, tăng nguy cơ sự cố do tiếp xúc với cây, thiết bị tiếp xúc với tác động của thời tiết trong dài hạn dẫn đến xuống cấp và giảm khả năng cung ứng dịch vụ, v.v. Những loại gián đoạn này được đánh dấu là lỗi thiết bị.

Ngoài hai khái niệm trên, thực tế là chúng ta có thể chia sự xuống cấp của thiết bị theo hai xu hướng: (1) Xuống cấp theo thời gian và có thể quan sát được (Manifest Deterioration) và (2) xuống cấp bất ngờ hay đột ngột (Latent/Sudden) do các nguyên nhân như thiên tai.

4.1.3 Cơ Bản về Quản Lý Sự Cố Mất Điện

Khả năng theo dõi đồng thời nhiều mối đe dọa do thời tiết gây ra và đánh giá các tác động và thiệt hại đến tài sản ngành điện hay các sở hạ tầng (hạ tầng giao thông) liên quan đến việc hỗ trợ cho ngành điện là rất quan trọng. Đơn giản bởi vì, lưới điện được phân bổ trên các vùng rộng lớn với các nhà máy phát điện thường đặt ở những vùng xa xôi. Tiêu thụ lớn ở các khu đô thị có nghĩa là lưới truyền tải phải đưa điện từ các nhà máy phát điện ở xa đến các trung tâm tiêu thụ và hệ thống phân phối phải cung cấp các đường dẫn tiện ích cho từng khách hàng. Để đạt được điều này, lưới phải trải qua các trạng thái hoạt động khác nhau (Hình 7). Bằng cách kết hợp quản lý tài sản và quản lý sự cố mất điện, chúng ta có thể xử lý và giảm thiểu các tác động gây thiệt hại một cách hiệu quả nhất.



Hình 7. Các trạng thái của lưới điện

4.1.4 Tiên Đoán Mất Điện trên Mạng Lưới Truyền Tải

Kiến thức từ dữ liệu thu thập trong quá khứ có thể được sử dụng để dự đoán những sự cố mất điện trên lưới điện liên quan đến thời tiết trong vòng 1-3 giờ trước. Các biến liên quan đến không gian được thêm vào bộ dữ liệu để tính toán được mối liên hệ tương hỗ giữa các sự kiện và các nút trên mạng lưới. Chương trình Hệ thống Quan sát Bề Mặt Tự Động (ASOS) được sử dụng để thu thập các đo lường thời tiết lịch sử cho các tham số sau: Hướng gió [độ], Tốc độ gió [km/h], Gió giật [m/s], Nhiệt độ [C], Điểm sương [F], Độ ẩm tương đối [%], Áp suất [mb], Lượng mưa/giờ [mm]. Cơ sở Dữ liệu Dự báo Kỹ thuật số Quốc gia (NDFD) của Mỹ được sử dụng để trích xuất dữ liệu dự báo thời tiết trong quá khứ và được sử dụng để kiểm tra khả năng theo thời gian thực tính phản xạ của hệ thống tương ứng với các xác suất mất điện khác nhau theo thời gian.

Việc bố trí tối ưu các thiết bị bảo vệ sét cho dây điện và cột điện nhằm giảm thiểu tổng rủi ro của các sự cố và độ nhiễu liên quan đến sét khi vẫn duy trì trong giới hạn ngân sách yêu cầu. Mạng lưới và tác động của nó được mô hình hóa bằng một mạng có trọng số đa chế độ sử dụng dữ liệu từ nhiều nguồn khác nhau. Mô hình rủi ro được phát triển (sử dụng Gaussian Conditional Random Fields (GCRF) (Radosavljevic et al., 2013)) có xem xét đến tác động tích lũy của các sự cố sét trước đó để tạo ra một ước tính chính xác hơn về độ nhiễu của bộ cách điện và dự đoán hiệu suất của bộ cách điện khi có tình huống xảy ra quá áp do sét đánh trong tương lai.

4.1.5 Đánh giá tình trạng hoạt động và sự xuống cấp của trạm biến áp

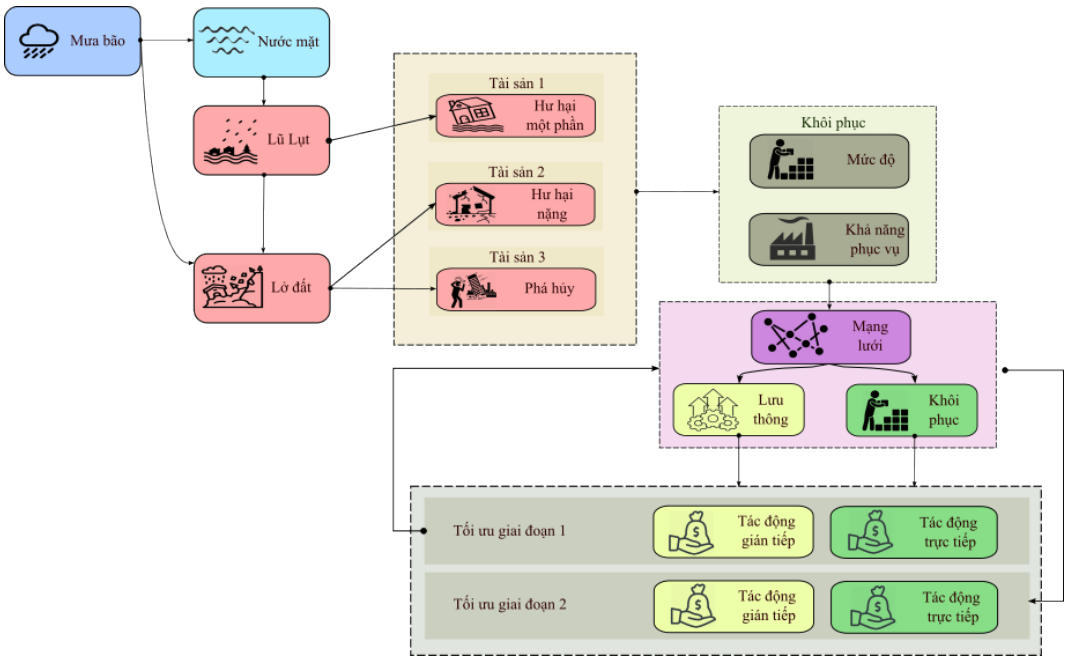
Các phương pháp truyền thống để đánh giá tình trạng hoạt động (sức khỏe) của trạm biến áp được phát triển bằng cách sử dụng kiến thức lĩnh vực về các quá trình vật lý và hóa học xảy ra bên trong bồn dầu của bộ biến áp, sau đó được xác nhận bằng các nghiên cứu kinh nghiệm. Một số ví dụ là việc sử dụng phương pháp tam giác Duval, tỷ số khí IEC hoặc phân tích khí chính. Việc thu thập dữ liệu ngày càng tăng (ví dụ: phân tích khí và dầu định kỳ, các cảm biến thu thập dữ liệu thời gian thực, v.v.) bởi các công ty điện đã thúc đẩy sự phát triển các phương pháp phân tích dữ liệu dựa trên học máy có giám sát để phân loại tình trạng bộ biến áp và loại ra các lỗi. Một số ví dụ là việc sử dụng mạng nơ-ron nhân tạo đa tầng, Support Vector Machine và mạng Bayesian .

Các thuật toán học máy có giám sát đối mặt với các thách thức sau: (i) đa số các thuật toán chỉ cung cấp khả năng giải thích thấp cho người ra quyết định; (ii) thiếu dữ liệu về sự cố có chất lượng cao; và (iii) dữ liệu được gán nhãn về tình trạng biến áp trong hầu hết các trường hợp không có sẵn hoặc được xác định bởi con người (tức là phân loại chủ quan của tình trạng). Sử dụng học máy không giám sát là một giải pháp thay thế và hấp dẫn, nhưng trong tài liệu vẫn còn hạn chế. Phương pháp không giám sát đầu tiên là chỉ số

sức khỏe được mô tả trong tài liệu tham khảo (Jahromi et al., 2009) tóm tắt sức khỏe tổng thể của tài sản bằng cách kết hợp kết quả quan sát vận hành, kiểm tra trường và kiểm tra phòng thí nghiệm thành một chỉ số duy nhất. Tuy nhiên, các hạn chế chính của chỉ số này là: (i) định nghĩa kinh nghiệm của trọng số cho mỗi tiêu chí, và (ii) thiếu thông tin về loại lỗi. Các phương pháp thay thế khác là phân cụm (Islam et al., 2016) và học bán giám sát với Low Dimensional Scaling (Mirowski and LeCun, 2012).

4.1.6 Dự đoán về thiệt hại cơ sở hạ tầng thảm khốc gây ra tình trạng mất điện

Một trong những vấn đề mà ít có nghiên cứu nào đề cập đó là vấn đề liên quan đến phục hồi (restoration) khi có thiên tai xảy ra, các thiên tai như mưa lớn lũ lụt gây ra sự cố mất điện trên diện rộng và để phục hồi việc cấp điện nhiều khi cần phải huy động nguồn lực không những trong ngành điện mà còn các ngành và cơ quan chính quyền như quân đội và người dân. Hơn nữa, công tác phục hồi thường diễn ra không phải trong một vài ngày mà có thể kéo dài và tốn nhiều nguồn lực hơn, cần sự điều phối cao và chặt chẽ hơn. Trong những tình huống như vậy, chúng ta có thể sử dụng các công cụ toán tối ưu để hỗ trợ. Các công cụ toán này được xây dựng với biến đầu vào cả bên trong và bên ngoài ngành điện, cả nguy cơ, tổn thương, và tác động của nguy cơ gây ra, cũng như nguồn lực trong một khoảng thời gian cụ thể. Hình 8 biểu diễn mối quan hệ tương hỗ lẫn nhau trong công tác phục hồi lưới điện khi có sự cố xảy ra do tác động của thiên tai. Phân tích dữ liệu lớn có thể được sử dụng để dự đoán các sự cố thảm khốc do các sự kiện thời tiết khắc nghiệt như bão, lốc xoáy, lũ lụt và sóng thần.



Hình 8. Phục hồi mạng lưới điện khi có sự cố gây ra bởi thiên tai

4.2 Phân tích dữ liệu thu thập từ các công tơ thông minh

Ứng dụng phân tích dữ liệu lớn cho dữ liệu thu thập từ các công tơ thông minh có thể là

- Các ứng dụng về phân tích dự trên một công tơ thông minh
- Các ứng dụng về phân tích dự trên nhiều công tơ thông minh
- Các công tơ thông minh được kết nối với các mô hình mạng lưới

Số lượng đồng hồ thông minh được lắp đặt ở Mỹ và Châu Âu đã vượt quá 50% trên tổng số lượng công tơ điện tính tới thời điểm năm 2020. Điều này hiển nhiên cung cấp cơ hội lớn cho việc phân tích dữ liệu để nâng cao quản lý khách hàng và hoạt động và kế hoạch quản lý lưới điện. Các công tơ thông minh cung cấp các chỉ số về năng lượng, công suất và điện áp ở mức độ phân giải cao, thường là một giờ hoặc 15 phút. Số liệu lấy từ công tơ điện thông minh được sử dụng cho phân tích như dự báo, phân loại tải cho từng đối tượng khách hàng, ước tính tải. Khi kết hợp với các nguồn dữ liệu khác và các hệ thống tiện ích, phân tích dựa trên dữ liệu từ công tơ thông minh có thể được mở rộng và tạo thêm lợi ích cho hoạt động của doanh nghiệp.

4.3 Dự báo và phân tích năng lượng tái tạo

Ngày nay tại các trang trại điện gió, đặc biệt là các trang trại điện gió ngoài khơi, việc thu thập dữ liệu ở mức độ động cơ gió được thực hiện với độ phân giải một giây. Tương tự, đối với các nhà máy năng lượng mặt trời, dữ liệu có thể được thu thập ở mức độ biến tần và với độ phân giải cùng cấp độ giây (Gilbert et al., 2020). Những dữ liệu ở mức độ rất tinh vi này được sử dụng để cải thiện phân tích và dự báo. Ngoài ra, vì khả năng sinh năng lượng từ các nguồn tái tạo bắt đầu trở nên đa dạng và phân tán địa lý một cách dày đặc, ta cũng có thể sử dụng tất cả dữ liệu được thu thập tại các trang trại để cải thiện dự báo.

Ngoài ra, còn có rất nhiều nguồn dữ liệu khác có liên quan, chủ yếu là liên quan đến quan sát và dự báo khí tượng, mô tả các quy trình phức tạp và mang lại khối lượng dữ liệu rất lớn.

- Sky imagers có tiềm năng lớn cho mô hình hóa và dự báo năng lượng mặt trời với độ phân giải cao vì chúng theo dõi các đám mây di chuyển và tác động của chúng đến các tấm pin mặt trời (Chow et al., 2011). Chúng có thể cung cấp hình ảnh của bầu trời trên các nhà máy điện mặt trời mỗi 30 giây;
- Radar thời tiết cũng đã chứng minh tầm ảnh hưởng trong đánh giá và mô hình hóa các chế độ làm việc của các trang trại điện gió. Tùy thuộc vào công nghệ, hình ảnh radar có thể có sẵn giữa mỗi phút và mỗi 10 phút, với bán kính hình ảnh từ 60 đến 250 km;
- Dữ liệu từ LIDAR ngày càng được coi là rất liên quan đến các đo lường gió ở phía trước của các tuabin gió và chúng được tích hợp vào các phương pháp dự báo, hoặc nói chung là các quan sát tiềm năng mới về gió được sử dụng trong dự báo thời tiết và năng lượng tái tạo. LIDAR cung cấp các đo lường gió cho hình nón mà chúng quét (theo chiều dọc hoặc chiều ngang, tùy thuộc vào cách chúng được thiết lập) mỗi vài giây.

Ngoài ra, có thể đề cập đến hình ảnh vệ tinh, có tiềm năng quan trọng cho năng lượng gió, mặt trời và sóng. Tuy nhiên, tần suất cập nhật thấp hơn khiến chúng ít quan trọng hơn ở thời điểm hiện tại. Thông tin từ các loại thiết bị này được gọi là thông tin được cảm biến từ xa.

Hiện nay, sự sẵn có của số lượng lớn dữ liệu này yêu cầu những thay đổi cơ bản trong phân tích và dự báo năng lượng tái tạo, không chỉ về phương pháp mà còn về các mô hình kinh doanh. Chúng ta mong đợi sẽ có nhiều công trình đổi mới xuất hiện trong tương lai, đề xuất các phương pháp dựa trên phương trình vi phân ngẫu nhiên, học sâu, học phân tán và liên kết, cũng như dữ liệu về thị trường.

4.4 Hiệu quả và tối ưu hóa năng lượng

Các công nghệ về lưới điện thông minh cung cấp tiềm năng lớn để tăng cường hiệu quả sử dụng năng lượng trong các lĩnh vực khác nhau. Tuy nhiên, hiện tại, các hoạt động tăng cường hiệu quả năng lượng chủ yếu bị hạn chế trong việc thực thi tiêu chuẩn ISO 50001 như: (i) lắp đặt thiết bị bổ sung (đồng hồ đo, cảm biến, v.v.) để đo lường tiêu thụ năng lượng; (ii) lắp đặt phần cứng mới và thay thế thiết bị; và (iii) trực quan hóa dữ liệu, tìm các mô hình bất thường; và xác định các quy trình tiêu thụ năng lượng cao. Thực tiễn tiêu

chuẩn này cung cấp việc giám sát và nhận thức về tiêu thụ năng lượng cho các quyết định của con người, nhưng nó không cho phép phân tích theo chỉ đạo và kiểm soát quy trình tự động.

Sự xuất hiện của công nghệ internet-of-things cung cấp điều kiện kỹ thuật cho mô hình hóa động lực dữ liệu trong tối ưu hóa năng lượng. Một số ví dụ là việc sử dụng học tăng cường dạng batch (fitted Q-iteration) để điều khiển một nhóm máy sưởi nước điện gia đình cho dịch vụ đáp ứng nhu cầu (Ruelens et al., 2014); fitted Q-iteration kết hợp với auto-encoders để tối ưu hóa năng lượng trong máy sưởi nước điện (Ruelens et al., 2018); học tăng cường sâu cho tối ưu hóa năng lượng dự đoán các trạm bơm nước thải (Filipe et al., 2019); và mô hình cây cho tiêu thụ năng lượng của tòa nhà để lên lịch hệ thống sưởi ấm tối ưu (Kouzelis et al., 2015). Phương pháp này không yêu cầu việc mô hình hoá đầy đủ các phương trình quá trình vì sự hiểu biết của nó được thực hiện theo thời gian thực thông qua dữ liệu. Tuy nhiên, sẵn có dữ liệu và thời gian (ví dụ: số lần tương tác với hệ thống vật lý) cần thiết để "huấn luyện" các phương pháp tối ưu dữ liệu vẫn là thách thức thực tế cho sự áp dụng trong ngành công nghiệp.

5 CƠ HỘI VÀ THÁCH THỨC TRONG TƯƠNG LAI

Các thách thức và cơ hội liên quan đến công tác phân tích dữ liệu lớn trong tương lai có thể được tóm tắt theo liệt kê dưới đây:

- Dự đoán các sự kiện trước thời gian và cho phép triển khai các chiến lược giảm thiểu rủi ro và tính toán rủi ro.
- Tối ưu hóa sử dụng dữ liệu đã thu thập cho lợi ích của doanh nghiệp và người sử dụng. Kết hợp mô hình vật lý và dữ liệu để cải thiện việc phân tích nguyên nhân gốc rễ của các vấn đề. Ngăn chặn các sự cố và hạn chế trong vận hành hệ thống điện hiện tại liên quan đến tác động và hậu quả kinh tế lớn.
- Thông báo cho người dùng về những hạn chế cung cấp điện sắp tới và sử dụng các nguồn lực phân tán hiệu quả hơn để giảm thiểu sự cố và những tình huống khẩn cấp khác.
- Thúc đẩy sự đổi mới bằng cách sử dụng kinh nghiệm đa ngành từ các lĩnh vực khác nhau nhưng có mối liên hệ mạnh về thuộc tính dữ liệu và yêu cầu phân tích.
- Sử dụng dữ liệu từ nhiều nguồn phổ biến khác nhau như máy tính bảng, điện thoại thông minh và các thiết bị điện tử cá nhân khác, khiến doanh nghiệp và các bên liên quan của doanh nghiệp phân tích dữ liệu trở thành các chủ thể cùng tham gia.
- Thay thế một số nhiệm vụ được thực hiện bởi các nhà điều hành và chuyên gia ngày nay bằng các thuật toán và quy trình tự động hóa.
- Xây dựng được các mô phỏng không gian và thời gian liên quan đến sự xuống cấp tài sản thiết bị và mạng lưới, đồng thời tích hợp định lượng nguy cơ rủi do, tổn thương, và rủi do vào mô phỏng để đưa ra các quyết định quản lý tối ưu, đặc biệt là trong các hoàn cảnh khó khăn và tình trạng khẩn cấp.

GHI CHÚ QUAN TRỌNG

Bài viết này được viết dựa trên việc tổng hợp kiến thức từ các bài báo được trích dẫn liệt kê trong phần Tài Liệu, đặc biệt là bài viết của các tác giả Kezunovic et al. (2020); Angadi, Ravi V et al. (2020). Tác giả đã sử dụng tối ưu công cụ ChatGPT được phát triển bởi công ty OpenAI để làm công tác biên soạn cho phù hợp với người đọc là các cán bộ và nhà quản lý của các Công Ty Điện Lực EVN ở Thành Phố Hồ Chí Minh, dựa vào gợi ý của ChatGPT, tác giả tiếp tục chỉnh sửa nội dung và câu chữ cho phù hợp hơn với ngôn ngữ Tiếng Việt. Ngoài việc sử dụng ChatGPT, tác giả cũng đối chiếu phần biên soạn nội dung với công cụ AI được tích hợp trên phần mềm Bing và Edge của Microsoft.

TÀI LIỆU

Aggarwal, C. C. et al. (2015). *Data mining: the textbook*, vol. 1 (Springer)

Ahmed, E., Yaqoob, I., Gani, A., Imran, M., and Guizani, M. (2016). Internet-of-things-based smart environments: state of the art, taxonomy, and open research challenges. *IEEE Wireless Communications* 23, 10–16. doi:10.1109/MWC.2016.7721736

Albarakati, N. and Obradovic, Z. (2019). Multi-domain and multi-view networks model for clustering hospital admissions from the emergency department. *International Journal of Data Science and Analytics* 8, 385–403. doi:10.1007/s41060-018-0147-5

Angadi, Ravi V, Venkataramu, P. S, and Daram, Suresh Babu (2020). Role of big data analytics in power system application. *E3S Web Conf.* 184, 01017. doi:10.1051/e3sconf/202018401017

Arghandeh, R. and Zhou, Y. (2017). *Big data application in power systems* (Elsevier)

Batini, C., Cappiello, C., Francalanci, C., and Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM computing surveys (CSUR)* 41, 1–52

Breiman, L. (1996). Bagging predictors. *Machine learning* 24, 123–140

Breiman, L. (2001). Random forests. *Machine learning* 45, 5–32

Breiman, L. (2017). *Classification and regression trees* (Routledge)

Cai, L. and Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. *Data Science Journal*

Cavalcante, L., Bessa, R. J., Reis, M., and Browell, J. (2017). Lasso vector autoregression structures for very short-term wind power forecasting. *Wind Energy* 20, 657–675. doi:https://doi.org/10.1002/we.2029

Chow, C. W., Urquhart, B., Lave, M., Dominguez, A., Kleissl, J., Shields, J., et al. (2011). Intra-hour forecasting with a total sky imager at the uc san diego solar energy testbed. *Solar Energy* 85, 2881–2893. doi:https://doi.org/10.1016/j.solener.2011.08.025

Cooper, G. F. (1990). The computational complexity of probabilistic inference using bayesian belief networks. *Artificial Intelligence* 42, 393–405. doi:https://doi.org/10.1016/0004-3702(90)90060-D

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning* 20, 273–297

De Mauro, A., Greco, M., and Grimaldi, M. (2016). A formal definition of big data based on its essential features. *Library review* 65, 122–135

EU (2013). *Commission Regulation (EU)*. Tech. Rep. 543, Eurepean Union

Filipe, J., Bessa, R. J., Reis, M., Alves, R., and Póvoa, P. (2019). Data-driven predictive energy optimization in a wastewater pumping station. *Applied Energy* 252, 113423. doi:https://doi.org/10.1016/j.apenergy.2019.113423

Ghalwash, M. F., Radosavljevic, V., and Obradovic, Z. (2013). Extraction of interpretable multivariate patterns for early diagnostics. In *2013 IEEE 13th International Conference on Data Mining*. 201–210. doi:10.1109/ICDM.2013.19

Gilbert, C., Messner, J. W., Pinson, P., Trombe, P.-J., Verzijlbergh, R., van Dorp, P., et al. (2020). Statistical post-processing of turbulence-resolving weather forecasts for offshore wind power forecasting. *Wind Energy* 23, 884–897. doi:https://doi.org/10.1002/we.2456

Gligorijevic, D., Stojanovic, J., Djuric, N., Radosavljevic, V., Grbovic, M., Kulathinal, R. J., et al. (2016a). Large-scale discovery of disease-disease and disease-gene associations. *Scientific reports* 6, 1–12

Gligorijevic, D., Stojanovic, J., and Obradovic, Z. (2016b). Disease types discovery from a large database of inpatient records: A sepsis study. *Methods* 111, 45–55. doi:https://doi.org/10.1016/j.ymeth.2016.07.021. Big Data Bioinformatics

Hirth, L., Mühlenpfordt, J., and Bulkeley, M. (2018). The entso-e transparency platform – a review of europe's most ambitious electricity data platform. *Applied Energy* 225, 1054–1067. doi:https://doi.org/10.1016/j.apenergy.2018.04.048

Hong, T., Wang, P., and White, L. (2015). Weather station selection for electric load forecasting. *International Journal of Forecasting* 31, 286–295. doi:https://doi.org/10.1016/j.ijforecast.2014.07.001

Hong, T., Wilson, J., and Xie, J. (2014). Long term probabilistic load forecasting and normalization with hourly information. *IEEE Transactions on Smart Grid* 5, 456–462. doi:10.1109/TSG.2013.2274373

Hyndman, R. J. and Fan, S. (2009). Density forecasting for long-term peak electricity demand. *IEEE Transactions on Power Systems* 25, 1142–1153

Islam, M. M., Lee, G., and Hettiwatte, S. N. (2016). Incipient fault diagnosis in power transformers by clustering and adapted knn. In *2016 Australasian Universities power engineering conference (AUPEC)* (IEEE), 1–5

Jahromi, A. N., Piercy, R. C. M., Cress, S., Service, J., and Fan, W. (2009). An approach to power transformer asset management using health index. *IEEE Electrical Insulation Magazine* 25, 20–34

Kezunovic, M., Pinson, P., Obradovic, Z., Grijalva, S., Hong, T., and Bessa, R. (2020). Big data analytics for future electricity grids. *Electric Power Systems Research* 189, 106788. doi:https://doi.org/10.1016/j.epsr.2020.106788

Kezunovic, M., Xie, L., and Grijalva, S. (2013). The role of big data in improving power system operation and protection. In *2013 IREP Symposium Bulk Power System Dynamics and Control - IX Optimization, Security and Control of the Emerging Power Grid*. 1–9. doi:10.1109/IREP.2013.6629368

Khaitan, S. K. and Gupta, A. (2012). *High performance computing in power and energy systems* (Springer)

Kleissl, J. (2013). *Solar energy forecasting and resource assessment* (Academic Press)

Kouzelis, K., Tan, Z., Bak-Jensen, B., Pillai, J. R., and Ritchie, E. (2015). Estimation of residential heat pump consumption for flexibility market applications. *IEEE Transactions on Smart Grid* 6, 1852–1864

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature* 521, 436–444

Leitão, A., Carreira, P., Alves, J., Gomes, F., Cordeiro, M., and Cardoso, F. (2015). Using smart sensors in the remote condition monitoring of secondary distribution substations. In *Proc. of the 23rd International Conference on Electricity Distribution (CIRED)*

Li, Z., Shahidehpour, M., and Aminifar, F. (2017). Cybersecurity in distributed power systems. *Proceedings of the IEEE* 105, 1367–1388

Messner, J. W. and Pinson, P. (2019). Online adaptive lasso estimation in vector autoregressive models for high dimensional wind power forecasting. *International Journal of Forecasting* 35, 1485–1498. doi:https://doi.org/10.1016/j.ijforecast.2018.02.001

Mirowski, P. and LeCun, Y. (2012). Statistical machine learning and dissolved gas analysis: A review. *IEEE Transactions on Power Delivery* 27, 1791–1799. doi:10.1109/TPWRD.2012.2197868

Moreno-Munoz, A., Bellido-Outeirino, F., Siano, P., and Gomez-Nieto, M. (2016). Mobile social media for smart grids customer engagement: Emerging trends and challenges. *Renewable and Sustainable Energy Reviews* 53, 1611–1616. doi:https://doi.org/10.1016/j.rser.2015.09.077

Perez, L., Flechsig, A., Meador, J., and Obradovic, Z. (1994). Training an artificial neural network to discriminate between magnetizing inrush and internal faults. *IEEE Transactions on Power Delivery* 9, 434–441. doi:10.1109/61.277715

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning* 1, 81–106

Quinlan, J. R. (2014). *C4. 5: programs for machine learning* (Elsevier)

Radosavljevic, V., Ristovski, K., and Obradovic, Z. (2013). Gaussian conditional random fields for modeling patients response to acute inflammation treatment. In *ICML 2013 workshop on Machine Learning for System Identification*

Ruelens, F., Claessens, B. J., Quaiyum, S., De Schutter, B., Babuška, R., and Belmans, R. (2018). Reinforcement learning applied to an electric water heater: From theory to practice. *IEEE Transactions on Smart Grid* 9, 3792–3800. doi:10.1109/TSG.2016.2640184

Ruelens, F., Claessens, B. J., Vandael, S., Iacovella, S., Vingerhoets, P., and Belmans, R. (2014). Demand response of a heterogeneous cluster of electric water heaters using batch reinforcement learning. In *2014 power systems computation conference* (IEEE), 1–7

Simões Costa, A., Albuquerque, A., and Bez, D. (2013). An estimation fusion method for including phasor measurements into power system real-time modeling. *IEEE Transactions on Power Systems* 28, 1910–1920. doi:10.1109/TPWRS.2012.2232315

Srivastava, N. and Salakhutdinov, R. R. (2012). Multimodal learning with deep boltzmann machines. In *Advances in Neural Information Processing Systems*, eds. F. Pereira, C. Burges, L. Bottou, and K. Weinberger (Curran Associates, Inc.), vol. 25

Sun, H., Wang, Z., Wang, J., Huang, Z., Carrington, N., and Liao, J. (2016). Data-driven power outage detection by social sensors. *IEEE Transactions on Smart Grid* 7, 2516–2524. doi:10.1109/TSG.2016.2546181

Sweeney, C., Bessa, R. J., Browell, J., and Pinson, P. (2020). The future of forecasting for renewable energy. *Wiley Interdisciplinary Reviews: Energy and Environment* 9, e365

Tan, P.-N., Steinbach, M., and Kumar, V. (2016). *Introduction to data mining* (Pearson Education India)

Wang, P., Liu, B., and Hong, T. (2016). Electric load forecasting with recency effect: A big data approach. *International Journal of Forecasting* 32, 585–597. doi:https://doi.org/10.1016/j.ijforecast.2015.09.006

Wang, Y., Chen, Q., Hong, T., and Kang, C. (2018). Review of smart meter data analytics: Applications, methodologies, and challenges. *IEEE Transactions on Smart Grid* 10, 3125–3148

Werbos, P. J. (1994). *The roots of backpropagation: from ordered derivatives to neural networks and political forecasting*, vol. 1 (John Wiley & Sons)

Wickham, H. and Grolemund, G. (2023). R for data science: Visualize, model, transform, tidy and import data (2nd edition). *O'REILLY*

Zhang, G. (2000). Neural networks for classification: a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 30, 451–462. doi:10.1109/5326.897072

Zhu, Q., Chen, J., Shi, D., Zhu, L., Bai, X., Duan, X., et al. (2020). Learning temporal and spatial correlations jointly: A unified framework for wind speed prediction. *IEEE Transactions on Sustainable Energy* 11, 509–523. doi:10.1109/TSTE.2019.2897136

Ziel, F. and Weron, R. (2018). Day-ahead electricity price forecasting with high-dimensional structures: Univariate vs. multivariate modeling frameworks. *Energy Economics* 70, 396–420. doi:https://doi.org/10.1016/j.eneco.2017.12.016

Zugno, M., Pinson, P., and Madsen, H. (2013). Impact of wind power generation on european cross-border power flows. *IEEE Transactions on Power Systems* 28, 3566–3575. doi:10.1109/TPWRS.2013.2259850