Tạp chí Khoa học Đại học Cần Thơ
Phần A: Khoa học tự nhiên, Công nghệ và Môi trường
ISSN 1859-2333 | e-ISSN 2815-5599

DOI:10.22144/ctujos.2025.135

ĐÁNH GIÁ ẢNH HƯỞNG CỦA CÁC THAM SỐ GIẢI THUẬT BO-GP CHO MÔ HÌNH LIGHTGBM ĐỐI VỚI DỰ BÁO PHỤ TẢI NGẮN HẠN

Trần Thanh Ngọc*
Khoa Công nghệ Điện, Trường Đại học Công nghiệp Thành phố Hồ Chí Minh, Việt Nam
**Tác giả liên hệ (Corresponding author): tranthanhngoc@iuh.edu.vn*

Thông tin chung (Article Information)

Nhận bài (Received): 18/02/2025
Sửa bài (Revised): 12/03/2025
Duyệt đăng (Accepted): 17/07/2025

Title: *Evaluation of the Impact of BO-GP Algorithm Parameters on the LightGBM Model for Load Forecasting*

Author: *Tran Thanh Ngoc**

Affiliation(s): *Faculty of Electrical Engineering Technology, Industrial University of HCM City, Viet Nam*

TÓM TẮT

Lưu đồ giải thuật được đề xuất trong bài báo nhằm đánh giá ảnh hưởng của các tham số trong giải thuật Bayesian Optimization với Gaussian Process (BO-GP) đối với mô hình LightGBM trong bài toán dự báo phụ tải ngắn hạn. Các siêu tham số được khảo sát bao gồm: learning_rate, n_estimators, max_depth, subsample và colsample_bytree. Hiệu suất của giải thuật BO-GP được đánh giá dựa trên hai tham số chính: số vòng lặp tối ưu hóa (n_iter) và số lần chia tập dữ liệu huấn luyện (k-fold). Để kiểm tra độ ổn định của mô hình, hệ số phân tán CD% được tính toán bằng cách lặp lại BO-GP n = 30 lần. Tập dữ liệu phụ tải đỉnh hàng ngày từ bang Victoria, Úc được sử dụng trong nghiên cứu. Kết quả thực nghiệm cho thấy, giá trị tối ưu của n_iter là 80, trong khi giá trị mặc định là 50. Tương tự, k-fold đạt hiệu suất tốt nhất tại 4, 6 và 7, trong khi giá trị mặc định là 3. Tầm quan trọng của việc lựa chọn tham số phù hợp được nhấn mạnh trong nghiên cứu nhằm tối ưu hóa mô hình LightGBM khi áp dụng vào dự báo phụ tải.

Từ khóa: *Dự báo phụ, mô hình LightGBM, giải thuật BO-GP*

ABSTRACT

This paper proposes an algorithm flowchart to evaluate the impact of parameters in the Bayesian Optimization using the Gaussian Process (BO-GP) algorithm on the LightGBM model for short-term load forecasting. The investigated hyperparameters include learning_rate, n_estimators, max_depth, subsample, and colsample_bytree. The performance of the BO-GP algorithm is assessed based on two primary parameters: the number of optimization iterations (n_iter) and the number of folds in cross-validation (k-fold). To assess the model's stability, the coefficient of dispersion (CD%) is calculated by repeating the BO-GP process n = 30 times. The study utilizes daily peak load data from Victoria, Australia. The experiments revealed that the optimal n_iter value was 80, compared to the default of 50. Similarly, optimal performance for k-fold was observed at values of 4, 6, and 7, while the default was 3. The study highlights the importance of selecting appropriate parameters of the BO-GP algorithm to optimize the LightGBM model for load forecasting.

Keywords: *Load forecasting, LightGBM model, BO-GP algorythm*

1. GIỚI THIỆU

Dự báo nhu cầu phụ tải có vai trò quan trọng trong vận hành, truyền tải, phân phối và bán lẻ hệ thống điện. Các mô hình dự báo phụ tải chính xác có thể hỗ trợ các nhà quản lý hệ thống điện trong việc tối ưu hóa sản xuất, phân phối và tiêu thụ năng lượng (Fan & Hyndman, 2010). Các phương pháp dự báo phụ tải đã phát triển từ các mô hình kinh điển như: hồi quy tuyến tính (Vu et al., 2017), làm mịn hàm mũ (Kahraman & Akay, 2023), ARIMA (Kien et al., 2023), đến các mô hình tiên tiến như mạng nơ-ron nhân tạo (ANN) (Le et al., 2019), học máy SVM (Tran et al., 2024), học sâu như CNN, LSTM (Wang et al., 2024), gần đây là các mô hình học kết hợp như: XGBoost, Catboost và LightGBM (Tran & Nguyen, 2024). Trong đó, LightGBM là một thuật toán mạnh mẽ, được phát triển dựa trên Gradient Boosting Decision Trees (GBDT), có ưu điểm vượt trội về hiệu suất tính toán, khả năng xử lý dữ liệu lớn và độ chính xác cao.

Giống như nhiều mô hình học máy khác, hiệu suất của LightGBM phụ thuộc vào việc lựa chọn các siêu tham số của nó. Một số siêu tham số quan trọng của mô hình LightGBM bao gồm: learning_rate (tốc độ học), n_estimators (số lượng cây quyết định), max_depth (độ sâu tối đa của cây quyết định), subsample (tỷ lệ mẫu con), colsample_bytree (tỷ lệ đặc trưng được chọn cho mỗi cây) (Liang et al., 2021). Việc lựa chọn các siêu tham số tối ưu giúp mô hình đạt được độ chính xác cao nhất, do đó, việc tìm ra các giá trị siêu tham số tối ưu là rất quan trọng trong việc áp dụng mô hình LightGBM. Nhiều phương pháp tối ưu hóa khác nhau, chẳng hạn như Grid Search, Random Search, Genetic Algorithms và Bayesian Optimization đã được sử dụng để giải quyết thách thức này (Yang & Shami, 2020). Trong các thuật toán này, giải thuật Bayesian Optimization with Gaussian Process (BO-GP) là một trong những kỹ thuật tiên tiến, giúp giảm số lần thử nghiệm cần thiết để tìm ra cấu hình siêu tham số tốt nhất. BO-GP hoạt động bằng cách sử dụng Gaussian Process (GP) để mô hình hóa hàm mục tiêu và hướng dẫn quá trình tìm kiếm siêu tham số một cách hiệu quả.

Giải thuật BO-GP có một số tham số ảnh hưởng đến kết quả tìm kiếm, chẳng hạn như n_iter (số vòng lặp tối ưu hóa), không gian siêu tham số và hàm mục tiêu. Trong đó số lần lặp n_iter là tham số quan trọng nhất trong BO-GP vì đó là số lần lặp lại quá trình cập nhật mô hình GP để tìm kiếm siêu tham số tốt nhất (Tran & Nguyen, 2024). Ngoài ra, thuật toán BO-GP thường được kết hợp với quy trình xác thực chéo để tránh tình trạng quá khớp. Dữ liệu được chia thành các phần k-fold trong xác thực chéo, được gọi là xác thực chéo k-fold (Papadopoulos et al., 2023). Do đó, hai tham số có ảnh hưởng nhất trong BO-GP là số lần lặp và giá trị k-fold. Thuật toán BO-GP đã được sử dụng trong nhiều nghiên cứu để xác định siêu tham số tối ưu cho các mô hình học máy. Tuy nhiên, hầu hết các tác giả sử dụng các giá trị mặc định cho hai tham số này: số lần lặp n_iter = 50 và k-fold = 3 (Yang & Shami, 2020; Zwolle, 2022; Starukhin & Diukarev, 2024; Dwi et al. , 2025). Hiện tại, rất ít kết quả nghiên cứu đề cập đến tác động của các tham số này đối với thuật toán Bo-GP trong các mô hình LightGBM cho các vấn đề hồi quy, đặc biệt là trong bối cảnh dự báo phụ tải.

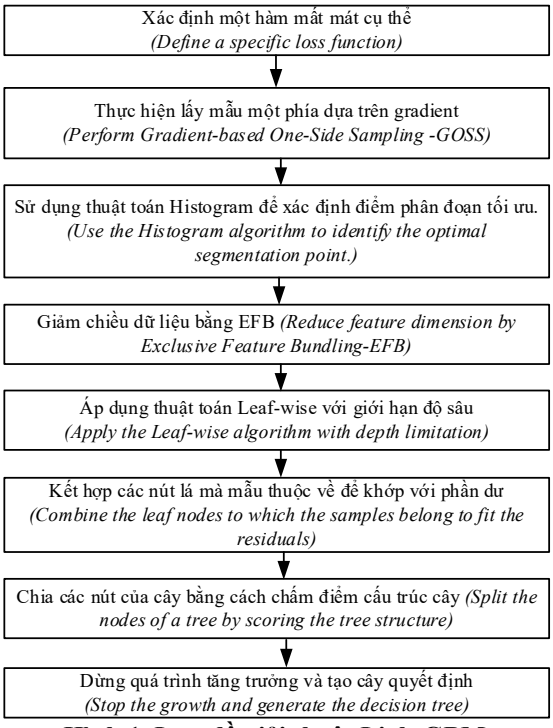
Để giải quyết những thiếu sót của nghiên cứu hiện tại, một lưu đồ giải thuật được đề xuất nhằm đánh giá ảnh hưởng của các tham số trong giải thuật BO-GP đối với mô hình LightGBM trong bài toán dự báo phụ tải. Các siêu tham số của mô hình LightGBM được khảo sát bao gồm: learning_rate, n_estimators, max_depth, subsample và colsample_bytree. Hiệu suất của giải thuật BO-GP được đánh giá dựa trên hai tham số chính: n_iter và k-fold. Để kiểm tra độ ổn định của giải thuật, hệ số phân tán CD% được tính toán bằng cách lặp lại quá trình tối ưu hóa 30 lần. Tập dữ liệu phụ tải đỉnh hàng ngày từ bang Victoria, Úc được sử dụng trong nghiên cứu. Kết quả thực nghiệm làm sáng tỏ tầm quan trọng của việc lựa chọn tham số phù hợp cho giải thuật BO-GP khi áp dụng trong mô hình LightGBM, góp phần nâng cao độ chính xác trong dự báo phụ tải.

2. MÔ HÌNH LGB VÀ GIẢI THUẬT BO-GP

2.1. Mô hình LGB

LightGBM là một thuật toán do Microsoft Research Asia phát triển, dựa trên khung GBDT, nhằm cải thiện hiệu quả tính toán để xử lý các bài toán dự đoán trên dữ liệu lớn một cách tối ưu hơn. Quy trình của thuật toán LightGBM được chia thành 8 bước cụ thể, như mô tả trong Hình 1 (Zhang & Gong, 2020). LightGBM cải tiến mô hình GBDT so với các phiên bản trước bằng cách tích hợp hai kỹ thuật quan trọng: Gradient-based One-Side Sampling (GOSS) và Exclusive Feature Bundling (EFB). GOSS hoạt động dựa trên nguyên tắc rằng các gradient lớn chứa nhiều thông tin về sự thay đổi của hàm mất mát, do đó mang lại thông tin quan trọng hơn. Vì vậy, nó ưu tiên lựa chọn các mẫu có gradient lớn, đồng thời lấy ngẫu nhiên một tập hợp con các mẫu có gradient nhỏ hơn. Phương pháp này tối ưu hóa việc sử dụng dữ liệu mẫu trong quá trình

huấn luyện, qua đó cải thiện hiệu suất mô hình. Trong khi đó, EFB giúp giảm số lượng đặc trưng (features) bằng cách kết hợp các đặc trưng không hoạt động đồng thời (mutually exclusive), qua đó tối ưu hóa kích thước và hiệu quả xử lý dữ liệu.



Hình 1. Lưu đồ giải thuật LightGBM

Giống như các mô hình học máy khác, độ chính xác của mô hình LightGBM phụ thuộc một cách đáng kể vào các siêu tham số của nó. Các siêu tham số chính của mô hình LightGBM là: (Liang et al., 2021; Vuong & Pham, 2023):

Learning rate (tốc độ học): Xác định kích thước bước tại mỗi lần lặp dựa trên hàm gradient mất mát. Tốc độ học thấp hơn giúp mô hình huấn luyện chậm hơn nhưng ổn định và có khả năng đạt được tối ưu toàn cục. Trong khi đó, tốc độ học cao giúp hội tụ nhanh nhưng có nguy cơ dừng lại ở các cực trị địa phương.

Maximum depth (độ sâu tối đa): Một số nguyên dùng để kiểm soát khoảng cách tối đa từ nút gốc đến nút lá trong cây quyết định. Độ sâu tối đa lớn hơn giúp mô hình học được nhiều đặc điểm phức tạp của dữ liệu, nhưng cũng làm tăng nguy cơ quá khớp (overfitting), khiến mô hình kém hiệu quả khi áp dụng cho dữ liệu mới.

Number of estimators (số lượng cây quyết định): Đề cập đến số lượng cây quyết định được sử dụng trong mô hình tổng hợp. Số lượng cây lớn hơn thường cải thiện hiệu suất dự báo nhưng cũng làm tăng thời gian huấn luyện và yêu cầu tài nguyên tính toán nhiều hơn.

Subsample (tỷ lệ mẫu con): Kiểm soát tỷ lệ dữ liệu huấn luyện được lấy ngẫu nhiên để phát triển mỗi cây. Giá trị nhỏ hơn 1 giúp giảm sự tương quan giữa các cây, qua đó cải thiện tính tổng quát của mô hình.

Min_child_weight (trọng số nhỏ nhất của nút con): Xác định tổng trọng số nhỏ nhất (hoặc số lượng điểm dữ liệu tối thiểu) cần thiết trong một nút lá. Giá trị cao hơn giúp mô hình đơn giản hơn, giảm nguy cơ quá khớp.

Colsample_bytree (tỷ lệ đặc trưng được chọn cho mỗi cây): Xác định tỷ lệ các đặc trưng (features) được chọn ngẫu nhiên cho mỗi cây. Điều này giúp giảm sự phụ thuộc vào một số đặc trưng cụ thể và cải thiện tính tổng quát.

Min_child_samples (số lượng mẫu tối thiểu trong một nút lá): Quy định số lượng mẫu tối thiểu cần có trong một nút lá. Tham số này giúp kiểm soát kích thước của nút lá, từ đó tránh việc chia nhỏ quá mức và giảm nguy cơ quá khớp.

Feature_fraction (tỷ lệ đặc trưng): Xác định tỷ lệ các đặc trưng được chọn ngẫu nhiên trước khi xây dựng mỗi cây quyết định. Tham số này giúp cải thiện tính đa dạng của mô hình bằng cách giảm sự phụ thuộc vào các đặc trưng cụ thể.

Bagging_fraction (tỷ lệ mẫu con cho bagging): Kiểm soát tỷ lệ dữ liệu huấn luyện được sử dụng ngẫu nhiên để xây dựng các cây (dùng cho bagging).

Max_bin (số lượng bin tối đa cho histogram): Xác định số lượng bin tối đa được sử dụng trong thuật toán histogram. Nhiều bin hơn có thể tăng độ phân giải của dữ liệu, cải thiện độ chính xác, nhưng cũng làm tăng thời gian tính toán và yêu cầu bộ nhớ.

2.2. Giải thuật BO-GP

Bayesian Optimization (BO) là một thuật toán lặp phổ biến dùng để giải quyết các bài toán tối ưu hóa siêu tham số. BO dự đoán điểm cần đánh giá tiếp theo dựa trên kết quả từ các bước trước đó. BO hoạt động dựa trên hai thành phần chính:

– Mô hình đại diện (Surrogate Model): Xấp xỉ hàm mục tiêu bằng cách khớp các điểm dữ liệu đã quan sát, từ đó tạo ra một phân phối xác suất cho các giá trị mục tiêu.

– Hàm chọn lựa (Acquisition Function): Quyết định cách cân bằng giữa khám phá (exploration) và

khai thác (exploitation) khi chọn các điểm dữ liệu mới để thử nghiệm.

Quy trình cơ bản của Bayesian Optimization bao gồm các bước sau:

1. Xây dựng mô hình đại diện xác suất để xấp xỉ hàm mục tiêu.

2. Tìm kiếm giá trị siêu tham số tối ưu bằng cách sử dụng hàm chọn lựa.

3. Áp dụng các giá trị siêu tham số tối ưu vào hàm mục tiêu và đánh giá hiệu suất.

4. Cập nhật mô hình đại diện bằng cách tích hợp các quan sát mới nhất.

5. Lặp lại quá trình trên cho đến khi đạt số vòng lặp tối đa hoặc thỏa mãn tiêu chí dừng.

Gaussian Process (GP) là một mô hình đại diện tiêu chuẩn được sử dụng để mô hình hóa hàm mục tiêu trong Bayesian Optimization. Khi có tập dữ liệu huấn luyện $D = \{(xi, yi)\}_{i=1}^{N}$, GP có thể ước lượng phân phối xác suất có điều kiện của giá trị đầu ra y dựa trên giá trị đầu vào x và dữ liệu đã quan sát.

$$p(y \mid x, D) = N(y \mid \hat{\mu}, \sigma^2) \qquad (1)$$

Trong đó: p (y| x, D) là phân phối xác suất có điều kiện của biến ngẫu nhiên y với điều kiện dữ liệu đã quan sát D và biến đầu vào x, N (y|$\hat{\mu}$, σ^2) là phân phối chuẩn Gaussian có $\hat{\mu}$ là kỳ vọng toán học và σ^2 là phương sai.

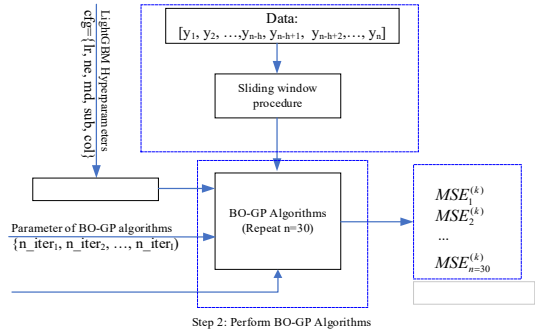
Sau khi có dự đoán từ mô hình GP, các điểm cần đánh giá tiếp theo được chọn dựa trên khoảng tin cậy của phân phối dự đoán. Mỗi điểm dữ liệu mới được kiểm tra thì được thêm vào tập dữ liệu huấn luyện và mô hình BO-GP được cập nhật lại với dữ liệu quan sát mới để cải thiện dự đoán. Quá trình này lặp lại cho đến khi đạt tiêu chí dừng.

3. MÔ HÌNH ĐỀ XUẤT

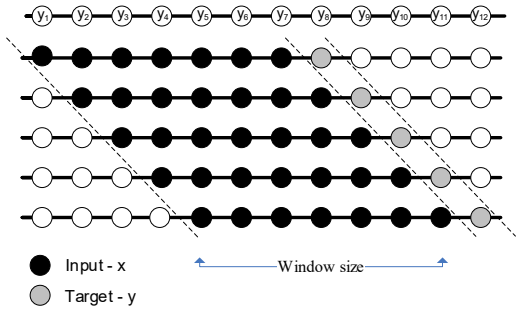
Hình 2 trình bày lưu đồ của giải thuật nhằm đánh giá mức độ ảnh hưởng của các tham số trong giải thuật BO-GP đối với mô hình LightGBM. Lưu đồ bao gồm ba bước chính: xử lý dữ liệu, thực hiện giải thuật BO-GP và phân tích kết quả.

Bước 1: Xử lý dữ liệu

Dữ liệu phụ tải đỉnh đầu vào được áp dụng chu trình cửa sổ trượt (Sliding window procedure) để tạo thành các tập dữ liệu input-target (x, y). Chu trình cửa sổ trượt được minh hoạ như Hình 3 bên dưới, trong đó kích thước cửa sổ (window size) là một tham số quan trọng của chu trình, xác định kích thước của tập ngõ vào x (Davtyan et al., 2020; Papadopoulos et al., 2023).



Hình 2. Lưu đồ giải thuật BO-GP đề xuất



Hình 3. Chu trình cửa sổ trượt

Bước 2: Thực hiện lặp lại n=30 lần giải thuật BO-GP.

Mô hình LightGBM được thiết lập dựa vào các siêu tham số cần xác định tối ưu. Trong nghiên cứu này, các siêu tham số đại diện được chọn là: learning_rate (lr), n_estimators (ne), max_depth (md), subsample (sub) và colsample_bytree (col). Mô hình LightGBM sau khi được thiết lập đóng vai trò là hàm mục tiêu đầu vào cho thuật toán BO-GP, nhằm xác định cấu hình siêu tham số tối ưu.

Các tổ hợp con của tham số n_iter và k-fold được xác định và thiết lập trong giải thuật BO-GP. Tham số n_iter có I phần tử, tham số k-fold có K phần tử. Như vậy có I x K tổ hợp của 2 tham số này. Điều này cho phép đánh giá mức độ ảnh hưởng của giải thuật BO-GP đối với các tham số n_iter và k-fold.

Đồng thời, thuật toán BO-GP được thực hiện lặp lại 30 lần để đánh giá độ ổn định và tính nhất quán của các kết quả tối ưu hóa siêu tham số. Quá trình lặp lại này giúp giảm thiểu ảnh hưởng của các yếu tố ngẫu nhiên và đảm bảo tính tin cậy của mô hình.

Ngõ ra của giải thuật BO-GP là giá trị sai số giữa giá trị dự đoán và giá trị thực. Trong nghiên cứu này,

giá trị sai số là MSE với phương trình được xác định như biểu thức (2) dưới đây được sử dụng, trong đó y là giá trị thực tế và $\hat{y}$ là giá trị dự báo:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (2)$$

Chú ý rằng trong lưu đồ của Hình 2, ngõ ra MSE có I x K đại lượng, tương ứng với số tổ hợp của I thành phần n_iter và K thành phần k-fold. Mỗi đại lượng lại có 30 giá trị, tương ứng với 30 lần lặp.

Bước 3: Phân tích kết quả.

Để đánh giá mức độ ngẫu nhiên của sai số MSE với 30 lần lặp, hệ số mức độ phân tán của dữ liệu CD% được tính toán theo công thức sau (Brown, 1998):

$$CD\% = \frac{100\sigma}{\mu} \qquad (3)$$

Trong đó, σ là độ lệch chuẩn (standard deviation) và μ là giá trị trung bình (mean) của n=30 lần lặp. Nếu giá trị CD% vượt quá 30%, thì kết quả không ổn định.

Đồng thời, các biểu đồ thể hiện mối tương quan giữa giá trị sai số MSE với các tham số n_iter và k-fold cũng được thiết lập để đánh giá mức độ ảnh hưởng của chúng.

4. KẾT QUẢ VÀ THẢO LUẬN

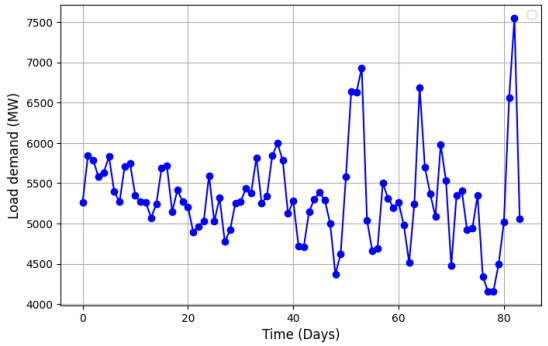
4.1. Dữ liệu và thiết lập thí nghiệm

Trong nghiên cứu này, dữ liệu phụ tải đỉnh hàng ngày (peak load) từ bang Victoria (VIC), Úc (https://aemo.com.au) được sử dụng. Dữ liệu được chia thành các tập đầu vào x và đầu ra y bằng chu trình cửa sổ trượt với kích thước cửa sổ là 7. Bảng 1 trình bày các đặc điểm thống kê của tập dữ liệu đầu ra y, bao gồm: số lượng quan sát, giá trị trung bình, độ lệch chuẩn, giá trị nhỏ nhất và giá trị lớn nhất. Hình 4 hiển thị biểu đồ dạng sóng của phụ tải y theo thời gian từ ngày 10 tháng 10 năm 2021 đến ngày 1 tháng 1 năm 2022.

Bảng 1. Đặc tính tập dữ liệu y

Tham số	Count	Mean	Std	Min	Max
Giá trị	84	5332	584	4156	7547

Bảng 2 trình bày phạm vi giá trị của các siêu tham số được khảo sát trong mô hình LightGBM. Các tham số này bao gồm: learning_rate, n_estimators, max_depth, subsample, và colsample_bytree với phạm vi cụ thể được nêu trong cột "Phạm vi khảo sát" và kiểu dữ liệu tương ứng (kiểu float hoặc integer) trong cột "Kiểu". Những phạm vi và kiểu dữ liệu này cung cấp cái nhìn tổng quan về các siêu tham số được xem xét trong quá trình tối ưu hóa.



Hình 4. Biểu đồ tập dữ liệu y

Bảng 2. Giá trị các siêu tham số khảo sát cho mô hình Lightgbm

Siêu tham số	Phạm vi khảo sát	Kiểu
learning_rate	[0,1-0,3]	float
n_estimators	[50-1000]	integer
max_depth	[1-16]	integer
subsample	[0,1-1,0]	float
colsample_bytree	[0,1-1,0]	float

Kết quả thể hiện ở Bảng 3 trình bày phạm vi các giá trị của các tham số được khảo sát cho giải thuật BO-GP, bao gồm n_iter và k-fold, cùng với số lượng tổ hợp tương ứng. Cột "Phạm vi khảo sát" liệt kê các giá trị của từng tham số, trong đó các giá trị mặc định được làm nổi bật bằng chữ in đậm và gạch chân để nhấn mạnh tầm quan trọng của chúng trong các thiết lập tiêu chuẩn. Cột "Số tổ hợp" thể hiện tổng số tổ hợp tham số được tạo ra từ các giá trị khảo sát, với tổng cộng 49 tổ hợp (tương ứng với 7 giá trị của n_iter và 7 giá trị của k-fold).

Bảng 3. Giá trị các tham số khảo sát cho giải thuật BO-GP

Tham số	Phạm vi khảo sát	Số tổ hợp
n_iter	[20, 30, 40, **50**, 60, 70, 80]	49
k-fold	[2, **3**, 4, 5, 6, 7, 8]	

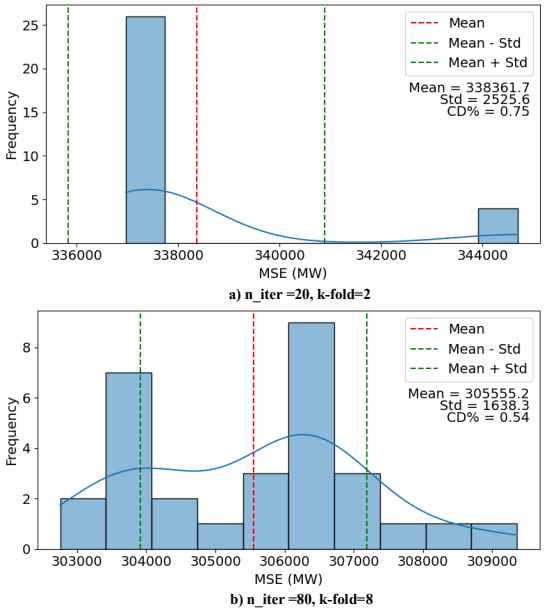
4.2. Kết quả và phân tích

4.2.1. Phân tích sự phân tán trong kết quả

Hình 5 trình bày biểu đồ phân bố (histogram) của sai số MSE cho tổ hợp đầu tiên và tổ hợp cuối cùng, tương ứng với 49 tổ hợp như được trình bày trong Bảng 3, với mỗi tổ hợp được lặp lại 30 lần để đảm bảo tính ổn định của kết quả. Cụ thể, Hình 5a thể hiện tổ hợp với tham số n_iter = 20 và k-fold = 2,

trong khi Hình 5b thể hiện tổ hợp với tham số n_iter = 80 và k-fold = 8.

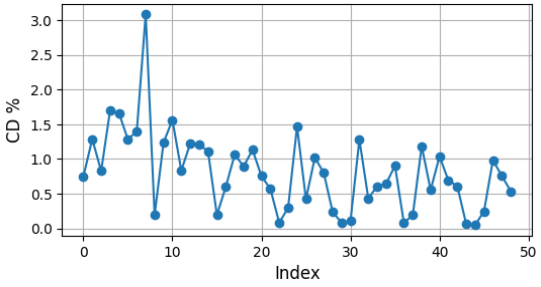
Kết quả trong Hình 5 phản ánh rõ tính chất ngẫu nhiên của các thuật toán tối ưu hóa qua nhiều lần thực thi. Cụ thể, trong Hình 5a, giá trị MSE dao động trong khoảng từ 336,000 đến 344,000 MW với tần suất xuất hiện từ 1 đến 25 lần. Trong khi đó, ở Hình 5b, giá trị MSE nằm trong khoảng từ 303,000 đến 309,000 MW với tần suất từ 1 đến khoảng 9 lần. Ngoài ra, các tổ hợp tham số khác được liệt kê trong Bảng 3 cũng cho thấy xu hướng tương tự. Do tính ngẫu nhiên này, việc chỉ dựa vào một lần chạy duy nhất có thể dẫn đến kết quả không đáng tin cậy, làm nổi bật sự cần thiết của việc thực hiện nhiều lần lặp để đảm bảo tính chính xác và độ ổn định của kết quả. Đây là lý do vì sao mỗi tổ hợp trong nghiên cứu đã được thực hiện 30 lần lặp để giảm thiểu ảnh hưởng của các yếu tố ngẫu nhiên và đảm bảo kết luận có độ tin cậy cao.



Hình 5. Biểu đồ phân bố sai số MSE

Kết quả ở Hình 6 thể hiện biểu đồ hệ số phân tán (CD%) cho các trường hợp khảo sát được liệt kê trong Bảng 3. Mỗi điểm trên đồ thị đại diện cho một tổ hợp tham số cụ thể. Trong Hình 6, điểm đầu tiên biểu thị giá trị CD% cho tổ hợp với tham số n_iter = 20 và k-fold = 2, trong khi điểm cuối cùng tương ứng với tổ hợp n_iter = 80 và k-fold = 8.

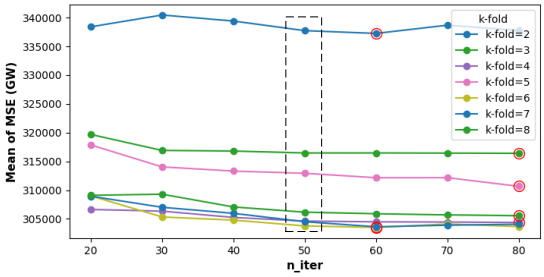
Kết quả phân tích từ Hình 6 cho thấy các hệ số phân tán CD% có giá trị rất thấp, phản ánh mức độ ổn định cao của các kết quả thu được. Giá trị CD% lớn nhất xấp xỉ 3,1%, thấp hơn nhiều so với ngưỡng tham chiếu 30%, vốn thường được xem là giới hạn chấp nhận trong các nghiên cứu thống kê. Điều này cho thấy rằng các giá trị thống kê của sai số MSE có mức độ tập trung cao xung quanh giá trị trung bình, đồng thời khẳng định tính nhất quán và độ tin cậy của kết quả thu được khi áp dụng các tổ hợp tham số khác nhau.



Hình 6. Biểu đồ hệ số phân tán CD%

4.2.2. Đánh giá tác động của các tham số đến độ chính xác của thuật toán BP-GP

Kết quả thể hiện ở Hình 7 cho thấy tỷ lệ lỗi trung bình MSE của thuật toán BO-GP theo tham số n_iter. Mỗi đường trên biểu đồ tương ứng với một giá trị tham số k-fold cụ thể cho từng thuật toán. Trong Hình 7, các điểm dữ liệu đại diện cho giá trị mặc định của tham số n_iter được làm nổi bật bằng các khung đường nét đứt, trong khi các giá trị nhỏ nhất của MSE được khoanh tròn để nhấn mạnh hiệu suất tối ưu.

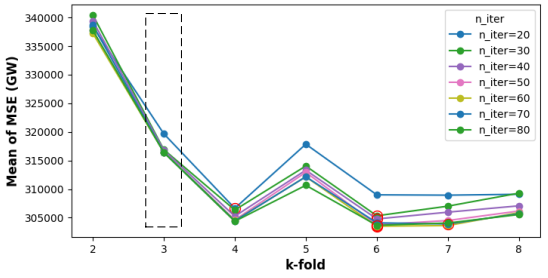


Hình 7. Ảnh hưởng của tham số n_iter

Hiệu suất của mô hình với các giá trị tham số mặc định được đánh giá, kết quả từ Hình 7 cho thấy rằng các giá trị mặc định tạo ra lỗi MSE cao hơn so với các giá trị tối ưu được khảo sát. Cụ thể, đối với tham số n_iter trong thuật toán BO-GP, giá trị trung bình của sai số MSE có xu hướng giảm dần khi n_iter tăng, với mức giảm rõ rệt và ổn định hơn sau khi vượt qua giá trị mặc định n_iter = 50 và đạt trạng thái bão hòa ở các mức n_iter = 60, 70, và 80. Trong đó, giá trị tham số n_iter tối ưu chủ yếu là 80, giúp

đạt được hiệu suất tốt nhất và vượt xa hiệu quả so với giá trị mặc định là 50.

Hình 8 trình bày các biểu đồ tương tự như trong Hình 7 nhưng được biểu diễn dựa trên các giá trị của tham số k-fold để kiểm tra ảnh hưởng của nó đến hiệu suất của các thuật toán. Mỗi đường trong biểu đồ tương ứng với một giá trị cụ thể của tham số n_iter. Các giá trị mặc định của tham số k-fold được đánh dấu bằng các khung nét đứt, trong khi các giá trị nhỏ nhất của sai số trung bình MSE được làm nổi bật bằng các vòng tròn để nhấn mạnh hiệu suất tối ưu.

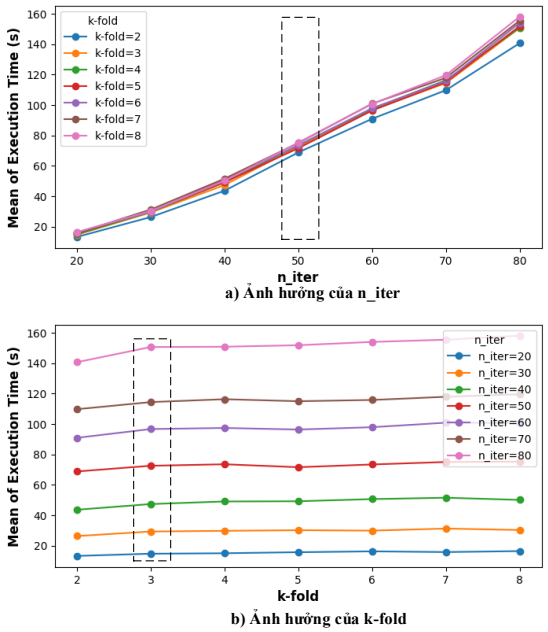


Hình 8. Ảnh hưởng của tham số k-fold

Kết quả từ Hình 8 được phân tích cho thấy rằng các giá trị mặc định (k-fold = 3) không phải là tối ưu, khi so sánh với các giá trị tham số khác. Cụ thể, các giá trị k-fold thay thế (không phải giá trị mặc định) như 4, 6 và 7 mang lại giá trị trung bình sai số MSE thấp hơn đáng kể so với giá trị mặc định. Điều này cho thấy rằng hiệu suất của mô hình có thể được cải thiện rõ rệt thông qua việc tối ưu hóa tham số k-fold, thay vì phụ thuộc vào các giá trị mặc định. Những phát hiện này nhấn mạnh tầm quan trọng của việc hiệu chỉnh tham số để đạt được hiệu quả tối ưu trong quá trình huấn luyện mô hình.

4.2.3. Đánh giá ảnh hưởng của các tham số đối với thời gian chạy chương trình.

Hình 9 trình bày biểu đồ minh họa các đường đặc trưng về thời gian thực thi trung bình của các thuật toán theo tham số n_iter (Hình 9a) và tham số k-fold (Hình 9b). Kết quả được phân tích cho thấy thời gian chạy của chương trình tăng gần như tuyến tính theo giá trị n_iter, phản ánh mối quan hệ trực tiếp giữa số vòng lặp và thời gian xử lý do mỗi vòng lặp bổ sung thêm khối lượng tính toán cho quá trình tối ưu hóa. Điều này cho thấy rằng khi tăng số vòng lặp, thuật toán cần nhiều thời gian hơn để hoàn thành quá trình huấn luyện. Trong khi đó, thời gian chạy của chương trình hầu như không thay đổi đáng kể khi tham số k-fold thay đổi. Sự khác biệt nhỏ giữa các giá trị k-fold cho thấy rằng yếu tố này có ảnh hưởng không đáng kể đến thời gian thực thi tổng thể. Điều này có thể do quy trình chia tách dữ liệu trong phương pháp cross-validation không làm tăng đáng kể khối lượng tính toán so với quá trình tối ưu hóa siêu tham số. Kết quả này nhấn mạnh rằng, để cải thiện hiệu suất tính toán, việc tối ưu số vòng lặp n_iter có tác động lớn hơn so với việc điều chỉnh tham số k-fold.



Hình 9. Ảnh hưởng các tham số đối với thời gian chạy chương trình.

5. KẾT LUẬN

Ảnh hưởng của các tham số n_iter và k-fold trong giải thuật BO-GP đối với hiệu suất mô hình LightGBM trong bài toán dự báo phụ tải ngắn hạn được đánh giá trong nghiên cứu. Kết quả thực nghiệm cho thấy, giá trị tối ưu của tham số n_iter chủ yếu là 80 thay vì 50 như mặc định, trong khi k-fold đạt hiệu suất tốt nhất tại 4, 6 và 7 thay vì giá trị mặc định là 3. Điều này chứng minh rằng việc lựa chọn tham số phù hợp cho các giải thuật tối ưu như BO-GP có tác động đáng kể đến độ chính xác của mô hình dự báo LightGBM, giúp giảm sai số MSE đáng kể so với việc sử dụng giá trị mặc định. Ngoài ra, hệ số phân tán CD% thấp chứng tỏ tính ổn định cao của các kết quả tối ưu hóa, khẳng định tính tin cậy của phương pháp đề xuất. Hướng nghiên cứu tiếp theo có thể tập trung vào việc khám phá các mô hình tiên tiến hơn như N-Beats hoặc Transformer, cũng như khảo sát hiệu quả của các giải thuật tối ưu hóa khác như Bayesian Optimization with Tree-structured Parzen Estimator và Genetic Algorithm nhằm nâng cao độ chính xác và tính linh hoạt của mô hình trong các bài toán dự báo phức tạp hơn.

TÀI LIỆU THAM KHẢO (REFERENCES)

Brown, C. E. (1998). *Applied multivariate statistics in geohydrology and related sciences*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-80328-4

Davtyan, A., Rodin, A., Muchnik, I., & Romashkin, A. (2020). Oil production forecast models based on sliding window regression. *Journal of Petroleum Science and Engineering, 195*, 107916. https://doi.org/10.1016/j.petrol.2020.107916

Dwi, U., Fathoni, D., & Sivi, N. A. (2025). Analisis Pengaruh Bayesian Optimization Terhadap Kinerja SVM Dalam Prediksi Penyakit Diabetes. *Infotek: Jurnal Informatika Dan Teknologi, 8*(1), 140–150. https://doi.org/10.29408/jit.v8i1.28468

Fan, S., & Hyndman, R. J. (2010). Short-term load forecasting based on a semi-parametric additive model. *IEEE Transactions on Power Systems, 25*(3), 1475–1483. https://doi.org/10.1109/TPWRS.2009.2036017

Kahraman, E., & Akay, O. (2023). Comparison of exponential smoothing methods in forecasting global prices of main metals. *Mineral Economics, 36*, 427–435. https://doi.org/10.1007/s13563-022-00354-y

Kien, D. T., Huong, P. D., & Minh, N. D. (2023). Application of SARIMA model in load forecasting in Hanoi City. *International Journal of Energy Economics and Policy, 13*(3), 164–170. https://doi.org/10.32479/ijeep.14121

Liang, S., Peng, J., Xu, Y., & Ye, H. (2021). Passive fetal movement recognition approaches using hyperparameter tuned LightGBM model and Bayesian optimization. *Computational Intelligence and Neuroscience, 2021*, 6252362. https://doi.org/10.1155/2021/6252362

Le, T. N., Pham, V. H., & Nguyen, M. H. (2019). Application of artificial neural networks for power output forecasting of thermal power plants. *Journal of Science and Technology – The University of Danang, 17*(3), 29–33. https://jst-ud.vn/jst-ud/article/view/1906

Papadopoulos, D. N., Javan, F. D., Najafi, B., Mamaghani, A. H., & Rinaldi, F. (2023). Handling complete short-term data logging failure in smart buildings: Machine learning-based forecasting pipelines with sliding-window training scheme. *Energy and Buildings, 301*, 113694. https://doi.org/10.1016/j.enbuild.2023.113694

Tran, T. N., & Nguyen, Q. D. (2024). Research on the influence of genetic algorithm parameters on XGBoost in load forecasting. *Engineering, Technology & Applied Science Research, 14*(6), 18849–18854. https://doi.org/10.48084/etasr.8863

Tran, T. N., Nguyen, Q. D., & Lam, B. M. (2024). Impact of kernel functions on support vector machine models in classification and regression problems. In *Proceedings of the International Conference on Sustainable Energy Technologies (ICSET)* (pp. 813–820). Springer. https://doi.org/10.1007/978-981-97-1868-9_80

Vu, N. H. M., Khanh, N. T. P., Cuong, V. V., & Binh, P. T. T. (2017). Forecast on Vietnam electricity consumption to 2030. In *Proceedings of the 2017 International Conference on Electrical Engineering and Informatics (ICELTICs 2017),* 72–77). https://doi.org/10.1109/ICELTICS.2017.8253238

Vuong, T. C., & Pham, H. H. (2023). Application of the LightGBM algorithm in land cover classification of Ly Son Island District, Vietnam. *Journal of Geodesy and Cartography Science, 56*, 51–57. https://doi.org/10.54491/jgac.2023.56.685

Wang, C., Li, X., Shi, Y., Jiang, W., Song, Q., & Li, X. (2024). Load forecasting method based on CNN and extended LSTM. *Energy Reports, 12*, 2452–2461. https://doi.org/10.1016/j.egyr.2024.07.030

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing, 415*, 295–316. https://doi.org/10.1016/j.neucom.2020.07.061

Starukhin, Y., & Diukarev, V. (2024). AUTOMATION OF TEXT DATA PROCESSING USING NLP. *The American Journal of Engineering and Technology, 6*(07), 24–39. https://doi.org/10.37547/tajet/Volume06Issue07-04

Zhang, D., & Gong, Y. (2020). The comparison of LightGBM and XGBoost coupling factor analysis and prediagnosis of acute liver failure. *IEEE Access, 8*, 220990–221003. https://doi.org/10.1109/ACCESS.2020.3042848

Zwolle, M. G. B. (2022). Analysing environmental ratings' ability to model firms' emission behaviour. *Econometrie*. http://hdl.handle.net/2105/62071