Báo Cáo Kỹ Thuật: Mô Hình Dự Đoán Khí Thải Tàu Biển

Báo cáo

Ngày 30 tháng 12 năm 2025

1 Kiến trúc Mô hình (Model Architecture)

Mô hình được xây dựng là một hệ thống hồi quy đa biến (Multi-output Regression) sử dụng thuật toán Gradient Boosting. Cụ thể, kiến trúc bao gồm hai thành phần chính:

1.1 Thuật toán cốt lõi: LightGBM

Mô hình cơ sở (Base Estimator) là **LightGBM (Light Gradient Boosting Machine)**. Đây là một khung (framework) gradient boosting sử dụng thuật toán học dựa trên cây quyết định (tree-based learning algorithms).

Các đặc điểm kỹ thuật chi tiết của kiến trúc này bao gồm:

- **Cơ chế phát triển cây (Tree Growth Strategy):** LightGBM sử dụng cơ chế phát triển theo lá (leaf-wise growth) thay vì theo mức (level-wise growth) như các thuật toán truyền thống. Điều này cho phép giảm loss nhanh hơn và độ chính xác cao hơn trên dữ liệu lớn.
- **Xử lý biến phân loại (Categorical Features):** Sử dụng phương pháp Fisher (histogram-based) để tìm điểm cắt tối ưu cho các biến phân loại như `valve_type` và `is_man`.

1.2 Chiến lược Đa đầu ra (Multi-Output Strategy)

Do bài toán yêu cầu dự đoán cùng lúc nhiều loại khí thải ($NO_x, SO_x, CO_2, \dots$) cho các giai đoạn hoạt động khác nhau (E1, E2, ...), mô hình sử dụng lớp bao `MultiOutputRegressor` từ thư viện Scikit-Learn.

- **Cơ chế:** Với N biến mục tiêu (target variables), hệ thống sẽ khởi tạo và huấn luyện N mô hình LightGBM độc lập.
- **Đầu vào (X):** 10 đặc trưng kỹ thuật của tàu.
- **Đầu ra (Y):** Tập hợp các giá trị khí thải liên tục.

1.3 Không gian Đặc trưng (Feature Space)

Mô hình sử dụng 10 biến đầu vào sau khi tiền xử lý:

STT	Tên đặc trưng	Mô tả / Loại dữ liệu
1	`year_ship`	Năm sản xuất tàu (Numerical)
2	`rpm`	Vòng tua máy (Numerical)
3	`is_man`	Loại động cơ (Categorical - Encoded)
4	`valve_type`	Loại van (Categorical - Label Encoded)
5	`P_main`	Công suất chính (Numerical)
6	`P_aux`	Công suất phụ trợ (Numerical)
7	`v_trip`	Vận tốc hành trình (Numerical)
8	`v_maneuver`	Vận tốc điều động (Numerical)
9	`v_max`	Vận tốc tối đa (Numerical)
10	`time_anchor`	Thời gian neo đậu (Numerical)

Bảng 1: Danh sách các đặc trưng đầu vào của mô hình

2 Tối ưu hóa Mô hình (Model Tuning)

Quá trình tinh chỉnh siêu tham số (Hyperparameter Tuning) được thực hiện để tối đa hóa hiệu suất dự đoán. Quá trình này được thực hiện như sau:

2.1 Phương pháp Tuning

Sử dụng phương pháp tìm kiếm ngẫu nhiên (**RandomizedSearchCV**) hoặc tìm kiếm lưới (Grid Search) trên không gian tham số của LightGBM. Mô hình được đánh giá và xếp hạng dựa trên hệ số xác định R^2 (Rank_R2).

2.2 Không gian tham số (Search Space)

Các tham số chính đã được thử nghiệm bao gồm:

- **learning_rate:** Tốc độ học (e.g., 0.05).
- **n_estimators:** Số lượng cây quyết định (e.g., 500).
- **num_leaves:** Số lượng lá tối đa trên một cây, quyết định độ phức tạp của mô hình.
- **max_depth:** Độ sâu tối đa của cây (e.g., 10, 15, hoặc -1 cho không giới hạn).
- **min_child_samples:** Số lượng mẫu tối thiểu cần thiết để tạo một lá (giúp chống overfitting).
- **subsample:** Tỷ lệ mẫu dữ liệu được sử dụng để huấn luyện mỗi cây.

2.3 Kết quả Thử nghiệm Chi tiết

Bảng dữ liệu dưới đây liệt kê các chỉ số đánh giá (R^2, MAE, MAPE) và bộ tham số tương ứng cho 10 cấu hình hàng đầu.

Rank	R2	MAE	MAPE (%)	Time (s)	Subsample	Leaves	Estimators	Min Child	Max Depth	LR
1	0.9750	570216.11	5.51	25.51	0.8	31	500	20	10	0.05
2	0.9749	572706.97	5.52	25.67	0.8	31	500	20	-1	0.05
3	0.9749	572632.39	5.52	25.86	0.8	31	500	20	15	0.05
4	0.9746	577249.65	5.48	46.13	1.0	63	1000	20	-1	0.01
5	0.9742	581995.29	5.64	69.58	0.8	127	1500	20	15	0.05
6	0.9740	600039.18	5.74	26.29	1.0	63	500	20	-1	0.10
7	0.9737	603326.16	5.76	67.03	0.8	63	1500	20	10	0.10
8	0.9645	813740.39	7.98	13.82	0.8	31	500	50	-1	0.05
9	0.9641	811301.96	8.08	22.42	1.0	127	1000	50	-1	0.05
10	0.9641	811422.38	8.08	22.05	0.8	31	1000	50	-1	0.05

Bảng 2: Top 10 kết quả tinh chỉnh siêu tham số (Hyperparameter Tuning Results)

Chú thích:

- **LR:** Learning Rate (Tốc độ học).
- **Leaves:** num_leaves (Số lá tối đa).
- **Min Child:** min_child_samples (Số mẫu tối thiểu ở một lá).
- **Time:** mean_fit_time (Thời gian huấn luyện trung bình).

Kết quả Tối ưu (Best Configuration) Theo bảng xếp hạng trong `tuning_model.csv`, cấu hình mô hình tốt nhất (Rank 1) đạt được kết quả như sau:

- **Độ chính xác (R^2):** ≈ 0.975 (Rất cao)
- **Sai số tuyệt đối trung bình (MAE):** $\approx 570,216$
- **Sai số phần trăm (MAPE):** $\approx 5.51\%$
- **Thời gian huấn luyện trung bình:** ≈ 25.5 giây

Bộ siêu tham số tối ưu (Optimal Hyperparameters):

Tham số	Giá trị tối ưu
Learning Rate	0.05
N Estimators	500
Num Leaves	31
Max Depth	10
Min Child Samples	20
Subsample	0.8

Bảng 3: Cấu hình siêu tham số của mô hình tốt nhất

2.4 Phân tích Kết quả

Cấu hình tốt nhất (Rank 1) đạt $R^2 \approx 0.975$ với thời gian huấn luyện rất nhanh (khoảng 25.5 giây). Điều thú vị là các mô hình có độ sâu cây giới hạn (`max_depth=10`) hoạt động tốt hơn một chút so với không giới hạn (`max_depth=-1`), cho thấy việc kiểm soát độ phức tạp của cây giúp mô hình tổng quát hóa tốt hơn.