DỰ BÁO MỨC ĐỘ Ô NHIỄM KHÔNG KHÍ TẠI BA TỈNH MIỀN BẮC VIỆT NAM: PHƯƠNG PHÁP HỌC MÁY VÀ HỌC SÂU

1st Cao Hoai Sang
Information System
name of organization (of Aff.)
Ho Chi Minh City, Viet Nam
21522541@gm.uit.edu.vn

2nd Nguyen Tran Gia Kiet
Information System
name of organization (of Aff.)
Ho Chi Minh City, Viet Nam
21522258@gm.uit.edu.vn

3rd Thi Thanh Cong
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

4th Nguyen Hoang Dang Khoa
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

5th Cu Ngoc Hoang
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

6th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

***Tóm tắt nội dung*—Mục tiêu chính của nghiên cứu này là dự đoán chất lượng không khí của ba tỉnh được chỉ định (Bắc Ninh, Hà Nội và Quảng Ninh) ở khu vực phía Bắc Việt Nam bằng cách kết hợp các thuật toán học máy và học sâu. Các mô hình bao gồm Gauss Newton Method Non-Linear, Residual Convolutional Neural Networks (ResCNN), Neural Hierarchical Interpolation for Time Series Forecasting (N-HiTS), Dynamic Linear Model (DLM), Simple Exponential Smoothing (SES), Linear Regression (LR), Autoregressive Integrated Moving Average (ARIMA), Recurrent Neural Network (RNN), Gated Recurrent Unit (GRU), Long Short Term Memory (LSTM). Hiệu quả của tất cả các mô hình được đề cập trên được đo lường bằng Mean Absolute Percentage Error (MAPE), Root Mean Squared Error, nhằm đạt được độ chính xác tối đa trong dự báo chuỗi thời gian chất lượng không khí chính xác.**

***Index Terms*—Nonlinear regression, Gauss-Newton, generalized least squares, iteratively reweighted least squares**

I. Giới Thiệu

Với sự gia tăng dân số nhanh chóng ở Việt Nam cùng với sự công nghiệp hóa ngày càng mạnh mẽ ở các khu vực trọng yếu, vấn đề chất lượng không khí nhanh chóng trở thành mối quan tâm chính, đặc biệt là ở các tỉnh đông dân cư ở khu vực phía Bắc Việt Nam. Sự suy giảm chất lượng không khí không chỉ gây ra những rủi ro lớn đối với sức khỏe của cư dân mà còn đe dọa sự cân bằng sinh thái mong manh của khu vực. Trong bối cảnh này, việc dự đoán chính xác và kịp thời các mức độ chất lượng không khí là rất cần thiết cho các chiến lược giảm thiểu hiệu quả và can thiệp y tế công cộng. Nghiên cứu của nhóm chúng tôi sử dụng sự kết hợp của các mô hình học máy và học sâu để dự đoán chất lượng không khí, bằng cách tận dụng sức mạnh của các thuật toán phức tạp như Gauss Newton Method Non-Linear, Resilient Convolutional Neural Networks (ResCNN), Neural Hierarchical Interpolation for Time Series Forecasting (N-HiTS), Dynamic Linear Model (DLM), Simple Exponential Smoothing (SES), Linear Regression (LR), Autoregressive Integrated Moving Average (ARIMA), Recurrent Neural Network (RNN), Gated Recurrent Unit (GRU), và Long Short Term Memory (LSTM).

Việc bao gồm một tập hợp đa dạng các mô hình cho phép chúng tôi khám phá các khía cạnh khác nhau của dữ liệu và so sánh hiệu suất của các phương pháp khác nhau. Chẳng hạn, các mô hình thống kê truyền thống như SES, LR, và ARIMA cung cấp tính giải thích và sự đơn giản, làm cho chúng trở thành các mô hình cơ sở có giá trị để so sánh. Mặt khác, các kiến trúc học sâu như RNN, GRU, và LSTM xuất sắc trong việc nắm bắt các phụ thuộc phi tuyến tính và phụ thuộc thời gian dài hạn, những đặc điểm thường có trong dữ liệu chuỗi thời gian chất lượng không khí. Ngoài ra, các phương pháp sáng tạo như ResCNN và N-HiTS mang lại những lợi thế độc đáo trong việc xử lý các cấu trúc không gian và thứ bậc trong dữ liệu, nâng cao độ chính xác của các dự đoán.

II. Nghiên cứu liên quan

A. Gauss-Newton nonlinear method

Vào năm 2015, ứng dụng dự báo phục hồi sau xuất viện của Phạm Thị Hương được sử dụng trong luận văn thạc sĩ khoa học [1]. Sử dụng phương pháp Gauss-Newton phi tuyến để ước lượng giá trị nhỏ nhất của bình phương sai số. [1]

III. Tài nguyên

A. Bộ dữ liệu

Dự báo chất lượng không khí là rất quan trọng để giảm thiểu các tác động tiêu cực của ô nhiễm lên sức khỏe con người và môi trường. Phân tích chuỗi thời gian nổi lên như một công cụ

Identify applicable funding agency here. If none, delete this.

mạnh mẽ trong lĩnh vực này, cho phép dự đoán các mức độ chất lượng không khí trong tương lai dựa trên dữ liệu lịch sử.

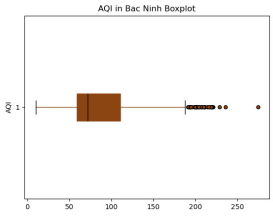
Do các vấn đề phổ biến liên quan đến bụi mịn ở khu vực phía Bắc, nhóm đã chọn một bộ dữ liệu chi tiết về chất lượng không khí của ba thành phố lớn ở miền Bắc Việt Nam: Bắc Ninh, Hà Nội và Quảng Ninh.

Bộ dữ liệu chủ yếu kéo dài từ năm 2021 đến năm hiện tại, 2024, bao gồm 6 cột tương ứng với các thành phần khác nhau trong không khí và chất lượng không khí được đánh giá thông qua Chỉ số Chất lượng Không khí (AQI): nồng độ PM2.5, nồng độ PM10, nồng độ O3, nồng độ NO2, nồng độ SO2 và nồng độ CO.

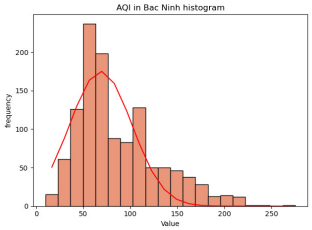
B. Thống kế mô tả

Bảng I
HANOI, BACNINH, QUANGNINH'S DESCRIPTIVE STATISTICS

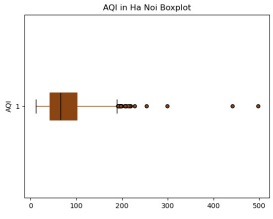
	HaNoi	BacNinh	QuangNinh
Count	2779	1190	929
Mean	75.283	86.87	32.117
Std	42.635	43.484	25.576
Min	11	10.0	5.0
25%	42.0	59.0	19.0
50%	66.0	72.0	25.0
75%	101.0	111.0	38.0
Max	498	275.0	500.0



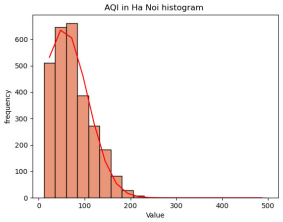
Hình 1. BacNinh AQI's boxplot



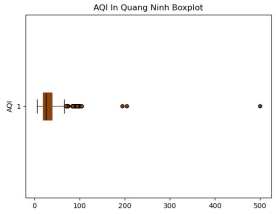
Hình 2. BacNinh AQI's hostogram



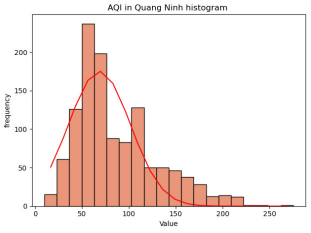
Hình 3. HaNoi AQI's boxplot



Hình 4. HaNoi AQI's hostogram



Hình 5. QuangNinh AQI's boxplot



Hình 6. QuangNinh AQI's hostogram

IV. PHƯƠNG PHÁP LUẬN

A. Gauss newton method nonlinear

1) Least Squares: Khoảng cách giữa một đường cong được khớp và một quan sát được gọi là sai số dư, hoặc lỗi.

$$Residuals = y_i - \widehat{y}_i$$

Tổng của các sai số bình phương được tính bằng phương trình sau:

$$SSE = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$$

Trong đó:
- y_i là các giá trị quan sát được
- $\widehat{y}_i$ là các giá trị khớp

2) Phương pháp Newton: Với hàm $y = y_0 e^{-kt}$, chúng ta tìm giá trị nhỏ nhất của SSE. Chúng ta tìm giá trị k bằng phương pháp Newton.

$$SSE = \sum_{i}^{n}(y_i - y_0 e^{-kt_i})^2$$

$$k_{\text{new}} = k_{\text{old}} - \frac{f'(k_{\text{old}})}{f''(k_{\text{old}})}$$

Hoặc chúng ta có thể giải thích bằng ma trận Hessian như sau:

$$\begin{pmatrix} k_{\text{new}} \\ y_{0,\text{new}} \end{pmatrix} = \begin{pmatrix} k_{\text{old}} \\ y_{0,\text{old}} \end{pmatrix} - H^{-1}G$$

Trong đó:
- $H = \begin{bmatrix} \frac{\partial^2 f}{\partial k^2} & \frac{\partial^2 f}{\partial k \partial y_0} \\ \frac{\partial^2 f}{\partial y_0 \partial k} & \frac{\partial^2 f}{\partial y_0^2} \end{bmatrix}$
- $G = \begin{bmatrix} \frac{\partial f}{\partial k} \\ \frac{\partial f}{\partial y_0} \end{bmatrix}$

Vấn đề với phương pháp Newton trong hồi quy phi tuyến là việc tính toán ma trận Hessian và nghịch đảo của nó gặp khó khăn. Để giải quyết vấn đề này, phương pháp Gauss-Newton thay thế bằng cách xấp xỉ ma trận Hessian.

Chúng ta có thể viết lại SSE như sau:

$$SSE = \sum_{i}^{n} r_i^2 = r^T r$$

Trong đó:
- r là một vector chứa các sai số dư

3) Phương pháp Gauss-Newton: Chúng ta lấy đạo hàm của SSE theo các tham số trong mô hình bằng quy tắc chuỗi, chúng ta có được phương trình sau:

$$\frac{\partial SSE}{\partial \beta_j} = 2\sum_i^n r_i \frac{\partial r_i}{\partial \beta_j}$$

Sau đó, bỏ số hai vì nó sẽ không ảnh hưởng đến việc ước tính các tham số. Tương ứng với ma trận Jacobian.

$$J_r = \begin{bmatrix} \frac{\partial r_1}{\partial \beta_1} & \frac{\partial r_1}{\partial \beta_2} \\ \frac{\partial r_2}{\partial \beta_1} & \frac{\partial r_2}{\partial \beta_2} \\ \vdots & \vdots \\ \frac{\partial r_n}{\partial \beta_1} & \frac{\partial r_n}{\partial \beta_2} \end{bmatrix}$$

Với phương trình sau cho tổng các sai số bình phương (SSE):

$$SSE = \sum_{i=1}^{n} (y_i - y_0 e^{-kt_i})^2$$

Chúng ta có:

$$\frac{\partial^2 SSE}{\partial \beta_j \partial \beta_k} = \sum_i^n \left(\frac{\partial r_i}{\partial \beta_j} \frac{\partial r_i}{\partial \beta_k} + r_i \frac{\partial^2 r_i}{\partial \beta_j \partial \beta_k} \right)$$

Sự khác biệt chính giữa phương pháp Newton và Gauss-Newton là phương pháp Gauss-Newton bỏ qua $r_i \frac{\partial^2 r_i}{\partial \beta_j \partial \beta_k}$. Do đó, đạo hàm bậc hai được xấp xỉ bằng hàm sau:

$$\frac{\partial^2 SSE}{\partial \beta_j \partial \beta_k} \approx \sum_i^n \left(\frac{\partial r_i}{\partial \beta_j} \frac{\partial r_i}{\partial \beta_k} \right) = J_r^T J_r$$

Sử dụng quy tắc cập nhật sau trong phương pháp Newton. Đối với Gauss-Newton, đơn giản chỉ cần cắm vào xấp xỉ cho ma trận Hessian và gradient.

$$\begin{pmatrix} k_{\text{new}} \\ y_{0,\text{new}} \end{pmatrix} = \begin{pmatrix} k_{\text{old}} \\ y_{0,\text{old}} \end{pmatrix} - (J_r^T J_r)^{-1} J_r^T r$$

Với β là một vector cột với các tham số được ước tính. Đối với ví dụ đơn giản chỉ ước tính hai tham số, phương trình trông như sau:

$$\beta_{\text{new}} = \beta_{\text{old}} - (J_r^T J_r)^{-1} J_r^T r(\beta_{\text{old}})$$

$$\begin{pmatrix} k_{\text{new}} \\ y_{0,\text{new}} \end{pmatrix} = \begin{pmatrix} k_{\text{old}} \\ y_{0,\text{old}} \end{pmatrix} - (J_r^T J_r)^{-1} J_r^T r \begin{pmatrix} k_{\text{old}} \\ y_{0,\text{old}} \end{pmatrix}$$

B. AUTOREGRESSIVE INTEGRATED MOVING AVERAGE (ARIMA)

ARIMA là viết tắt của "Autoregressive Integrated Moving Average". Mô hình ARIMA thường được sử dụng để dự báo dữ liệu chuỗi thời gian đơn biến. Mô hình ARIMA có thể xử lý một chuỗi thời gian nếu chuỗi đó là dừng và không có dữ liệu bị thiếu. Phương pháp này được sử dụng trong nhiều nghiên cứu để dự báo.

ARIMA là sự kết hợp của 3 thành phần, Auto-Regressive – AR, Integrated – I và Moving Average – MA tương ứng với các tham số p, d và q đại diện cho ba thành phần chính của mô hình, trong đó:

• $p - AR(p)$: Tham số p đại diện cho số lượng các quá trình tự hồi quy trong thành phần tự hồi quy (AutoRegressive) của mô hình ARIMA. Nó chỉ ra số lượng ngày quá khứ của chuỗi dữ liệu mà được sử dụng để dự đoán giá trị hiện tại. Mỗi giá trị quá khứ được sử dụng là một hệ số trong mô hình tự hồi quy. Giá trị của p phụ thuộc vào sự phụ thuộc tạm thời trong chuỗi dữ liệu và có thể được xác định bằng cách sử dụng các phương pháp như đồ thị tự tương quan (ACF - AutoCorrelation Function) hoặc hàm tương quan một lệnh (PACF - Partial AutoCorrelation Function). Phương trình tự hồi quy AR được tổng quát như sau:

$$Y_t = c + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \ldots + \varphi_p Y_{t-p} + \epsilon_t$$

Trong đó:
– Y_t đại diện cho giá trị dữ liệu tại thời điểm t
– c là hằng số chặn
– φ là hệ số AutoRegressive(AR)
– ϵ_t là sai số ngẫu nhiên
– p là số bậc

• $q - MA(q)$: Tham số q đại diện cho số lượng các thành phần trung bình động (Moving Average) trong mô hình ARIMA. Nó chỉ ra số lượng giá trị trung bình động được sử dụng để dự đoán giá trị hiện tại.

$$Y_t = c + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \ldots + \theta_p \epsilon_{t-p} + \epsilon_t$$

Trong đó:
– Y_t đại diện cho giá trị dữ liệu tại thời điểm t
– c là hằng số chặn
– θ là hệ số Moving Average(MA)
– ϵ_t là hệ số tương quan
– p là số bậc

• $d - I(d)$: Tham số d đại diện cho số lần lấy đạo hàm-sai phân (differencing) trên chuỗi dữ liệu ban đầu để loại bỏ xu hướng (trend) và/hoặc thành phần mùa vụ (seasonality) hay chuyển đổi dữ liệu thành chuỗi dừng. chuỗi dừng là chuỗi có trung bình, phương sai và tự tương quan không đổi theo thời gian. Một chuỗi thời gian được coi là chuỗi dừng nếu nó có trung bình không đổi, phương sai không đổi và tự tương quan không đổi. Chuỗi dừng là Công thức tính sai phân tại thời điểm t như sau

$$\Delta y_t = y_t - y_{t-1}$$

Sau khi kết hợp tất cả, ta có ARIMA(p, d, q) được biểu diễn như sau:

$$\Delta y_t = c + \varphi_1 \Delta y_{t-1} + \ldots + \varphi_p \Delta y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q}$$

Trong đó:
– Y_t đại diện cho giá trị dữ liệu tại thời điểm t
– c là hằng số chặn
– θ là hệ số Moving Average(MA)
– ϵ_t là hệ số tương quan
– φ là hệ số AutoRegressive(AR)

TÀI LIỆU

[1] P. T. Huong, "Linear regression, polynomial regression, and applications,"Master's thesis in science, 2015.

[2] H. Choi, C. Jung, T. Kang, H. J. Kim, and I. -Y. Kwak, "Explainable time-series prediction using a residual network and gradient-based methods,"in IEEE Access, vol. 10, pp. 108469-108482, 2022.

[3] Doan Vo Duy Thanh, Nguyen Van Cuong, Vo Tan Phat, Le Khac Hong Phuc, Nguyen Duy Du, Nguyen Van Quang, Pham Minh Duc, Pham Hong Vinh, and Nguyen Canh Thuong. *Gated Recurrent Unit (GRU)*. Accessed May 26, 2024. https://d2l.aivivn.com/chapter_recurrent-modern/gru_vn.html.

[4] Nejedly P, Ivora A, Viscor I, Koscova Z, Smisek R, Jurak P, Plesinger F. Classification of ECG using ensemble of residual CNNs with or without attention mechanism. Physiol Meas. 2022 Apr 28;43(4). doi: 10.1088/1361-6579/ac647c. PMID: 35381586.