

Nghiên cứu và ứng dụng thuật toán học máy: Tài liệu tham khảo cho ngành Trí tuệ nhân tạo

Nguyễn Thành Long*

*ThS. Trường Đại học Tài nguyên và Môi trường Hà Nội

Received: 2/11/2023; Accepted: 13/11/2023; Published: 20/11/2023

Abstract: This research article presents the application of several machine learning algorithms to estimate dust concentration in Bac Ninh city. This study focuses on using machine learning methods to predict airborne dust concentrations in Bac Ninh city and evaluate the performance of the applied algorithms. The results of the study can provide important information about air quality in this area, and open up the potential for widespread application of machine learning in environmental forecasting and monitoring.

Keywords: Research, application, machine learning algorithms, Reference documents, artificial intelligence industry

1. Đặt vấn đề

Trong thời đại hiện đại, ngành trí tuệ nhân tạo (AI) đang trở thành trung tâm của sự chú ý toàn cầu, với sức mạnh của thuật toán học máy đóng vai trò không thể phủ nhận. Bài báo này nghiên cứu và tìm hiểu ứng dụng của các thuật toán, từ đó không chỉ

trường tại thành phố Bắc Ninh. Dữ liệu này bao gồm thông tin chi tiết về các chỉ số nồng độ bụi trong không khí, được ghi lại theo khoảng thời gian cụ thể và điều kiện môi trường. Tương ứng với các mẫu là mức độ ô nhiễm và không ô nhiễm theo bảng 2.1.

Bảng 2.1. Các thang đo chất lượng không khí

AQI	Mức độ ô nhiễm không khí	Ý nghĩa sức khỏe	Tuyên bố cảnh báo (đối với PM2.5)
0 - 50	Tốt	Chất lượng không khí được coi là đạt yêu cầu và ô nhiễm không khí gây ra ít hoặc không có rủi ro	Không có
51 -100	Vừa phải	Chất lượng không khí ở mức chấp nhận được; tuy nhiên, đối với một số chất gây ô nhiễm, có thể có mối lo ngại về sức khỏe ở mức độ vừa phải đối với một số rất ít người nhạy cảm bất thường với ô nhiễm không khí.	Trẻ em và người lớn năng động và những người mắc bệnh về đường hô hấp, chẳng hạn như hen suyễn, nên hạn chế hoạt động ngoài trời kéo dài.
101-150	Không lành mạnh cho các nhóm nhạy cảm	Thành viên của các nhóm nhạy cảm có thể bị ảnh hưởng sức khỏe. Công chúng nói chung không có khả năng bị ảnh hưởng.	Trẻ em và người lớn năng động và những người mắc bệnh về đường hô hấp, chẳng hạn như hen suyễn, nên hạn chế hoạt động ngoài trời kéo dài.
151-200	Không khỏe mạnh	Mọi người đều có thể bắt đầu bị ảnh hưởng về sức khỏe; thành viên của các nhóm nhạy cảm có thể bị ảnh hưởng sức khỏe nghiêm trọng hơn	Trẻ em và người lớn năng động và những người mắc bệnh về đường hô hấp, chẳng hạn như hen suyễn, nên tránh hoạt động ngoài trời kéo dài; mọi người khác, đặc biệt là trẻ em, nên hạn chế hoạt động ngoài trời kéo dài
201-300	Rất không tốt cho sức khỏe	Cảnh báo sức khỏe về tình trạng khẩn cấp. Toàn bộ dân số có nhiều khả năng bị ảnh hưởng hơn.	Trẻ em và người lớn năng động và những người mắc bệnh về đường hô hấp, chẳng hạn như hen suyễn, nên tránh mọi hoạt động gắng sức ngoài trời; mọi người khác, đặc biệt là trẻ em, nên hạn chế hoạt động ngoài trời.
300+	Nguy hiểm	Cảnh báo về sức khỏe: mọi người có thể bị ảnh hưởng sức khỏe nghiêm trọng hơn	Mọi người nên tránh mọi hoạt động gắng sức ngoài trời

đưa ra cái nhìn tổng quan về tầm quan trọng của chúng trong ngành AI mà còn để cung cấp tài liệu tham khảo chất lượng.

2. Nội dung nghiên cứu

2.1. Thu thập dữ liệu

Để đảm bảo tính đáng tin cậy của nghiên cứu, chúng tôi đã thu thập dữ liệu về nồng độ bụi từ các trạm quan trắc môi

2.2. Tiền xử lý dữ liệu

Trước khi đưa vào ứng dụng các thuật toán học máy, chúng tôi đã thực hiện quá trình tiền xử lý dữ

Bảng 2.2: Bảng số liệu các trạm quan trắc môi trường tại thành phố Bắc Ninh

	Temp	NOX	SO2	O3	PM-10	PM-2.5	NO	NO2	CO
00:00 26/02/2020	17.4815	18.323	36.9389	47.5431	71.9046	32.5675	0.0024	18.3205	484.2535
01:00 26/02/2020	17.2813	18.47	37.0124	43.3199	62.2364	32.5149	-0.0006	18.4706	407.9796
02:00 26/02/2020	17.2327	48.8289	37.0761	14.2536	59.7413	39.6827	4.5104	44.3185	395.191
03:00 26/02/2020	16.5238	29.7647	37.4386	30.8589	59.5568	57.973	1.042	28.7226	363.9062
.....									

liệu. Điều này bao gồm việc xử lý dữ liệu nhiễu, điều chỉnh giá trị còn thiếu, và chuẩn hóa dữ liệu để đảm bảo tính nhất quán và hiệu suất tốt của mô hình.

Sau khi đã chuẩn hóa dữ liệu theo yêu cầu mẫu dữ liệu có cấu trúc như sau.

2.3. Lựa chọn thuật toán

Tôi đã tìm hiểu ba thuật toán học máy phổ biến để dự đoán nồng độ bụi đó là: Random Forest (RF), Support Vector Machines (SVM), và Thuật toán Artificial Neural Network (ANN). Sự kết hợp giữa các thuật toán này được thiết kế để đảm bảo sự đa dạng trong mô hình hóa và dự đoán.

Ưu nhược điểm của ba loại thuật toán này như sau.

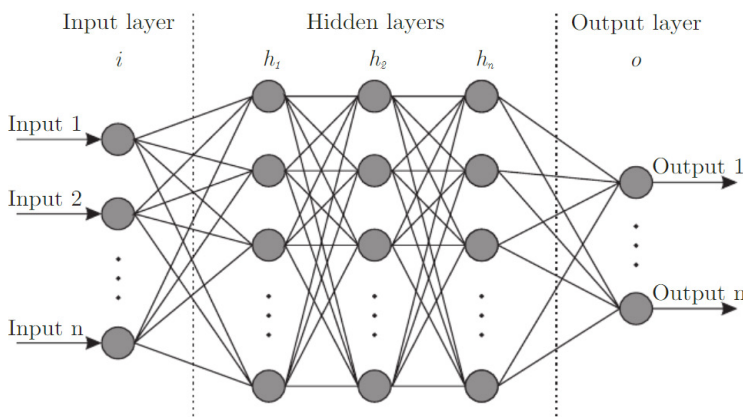
- Thuật toán Random Forest (RF) có khả năng xử lý nhiễu và dữ liệu nhiễu, thích hợp cho dữ liệu lớn và nhiều chiều. Dễ triển khai và linh hoạt trong việc áp dụng vào nhiều loại dữ liệu. Hạn chế là không hiệu quả khi xử lý dữ liệu rời rạc.

- Hệ thống mạng Support Vector Machines (SVM) làm việc hiệu quả với dữ liệu có nhiều chiều. Phù hợp cho cả hai loại dữ liệu. Có thể sử dụng với nhiều loại dữ liệu và tác vụ khác nhau.

- Đối với thuật toán Artificial Neural Network (ANN) thì điểm mạnh là có khả năng học tự nhiên và mô phỏng khả năng học của não người. Phù hợp với nhiều loại dữ liệu và tác vụ. Có khả năng tự động điều chỉnh trọng số trong quá trình học. Nhược điểm là yêu cầu một lượng lớn dữ liệu để đạt được hiệu suất tốt và kết quả chính xác. Cần nhiều tài nguyên tính toán, đặc biệt là với mô hình lớn và phức tạp.

Từ sự phân tích trên tôi đã lựa chọn mô hình Artificial Neural Network (ANN): Nó là mô hình thích hợp cho ước lượng nồng độ bụi với dữ liệu lớn và có tính nhiễu, đặc biệt là dữ liệu không gian chiều cao.

2.4. Cấu trúc mô hình ANN



Hình 2.1 – Cấu trúc mạng ANN

*Lớp đầu vào (Input Layer)

- Số lượng nút: Số lượng nút trong lớp này phụ thuộc vào số lượng đặc trưng (features) của dữ liệu đầu vào.

- Chức năng: Nút ở đây biểu diễn các đặc trưng của dữ liệu và là nơi tiếp nhận thông tin đầu vào.

*Lớp ẩn (Hidden layers)

- Số lượng lớp ẩn và nút trong mỗi lớp: Số lượng và kích thước của các lớp ẩn được thiết kế tùy thuộc vào độ phức tạp của vấn đề. Mỗi nút trong lớp này thực hiện các phép tính và truyền giá trị đến lớp kế tiếp.

- Chức năng: Lớp ẩn giúp mô hình học được các biểu diễn phức tạp từ dữ liệu đầu vào.

*Lớp đầu ra (Output Layer)

- Số lượng nút: Tùy thuộc vào loại vấn đề (ví dụ: một nút cho dự đoán giá trị liên tục, nhiều nút cho phân loại).

- Chức năng: Nút ở đây biểu diễn kết quả dự đoán hoặc phân loại của mô hình.

*Trọng số và hệ số Bias

- Trọng số (Weights): Là các tham số được điều chỉnh trong quá trình huấn luyện, ảnh hưởng đến đầu ra của mỗi nút.

- Hệ số Bias: Là giá trị cố định được thêm vào trước khi đưa qua hàm kích hoạt, giúp điều chỉnh đầu ra của mỗi nút.

- *Hàm kích hoạt: Hàm kích hoạt là hàm xác định cách mỗi nút tính tổng đầu vào của nó và trả ra đầu ra. Các hàm kích hoạt phổ biến bao gồm ReLU, Sigmoid, và Tanh.

2.5. Huấn luyện và kiểm thử mô hình

- Dữ liệu đã được chia thành hai phần: một phần được sử dụng để huấn luyện mô hình và phần còn lại

dùng để kiểm thử hiệu suất. Quá trình huấn luyện được thực hiện với một loạt các tham số khác nhau để tối ưu hóa hiệu suất của mỗi thuật toán.

- Dữ liệu của tôi chứa 10 đặc trưng, nên lớp đầu vào của mô hình Neural Network sẽ có 10 nút. Dưới đây là một số thông số có thể áp dụng khi xây dựng mô hình.

- *Chọn cấu trúc mạng Neural Network

- Lớp đầu vào (Input Layer):

10 nút (tương ứng với số lượng đặc trưng).

- Lớp ẩn (Hidden Layer): Sử dụng một hoặc nhiều lớp ẩn với số lượng nút tùy chọn. Trong nghiên cứu này tôi dùng lớp ẩn với 64 nút.

- Lớp đầu ra (Output Layer): 1 nút (dự đoán PM-2.5).

**Chọn hàm mất mát (Loss Function)*: Với bài toán dự đoán giá trị liên tục, có thể sử dụng Mean Squared Error (MSE):

**Chọn thuật toán tối ưu hóa và tỉ lệ học (Optimizer and Learning Rate)*: Adam là một lựa chọn phổ biến và tỉ lệ học thường được chọn là 0.001:

**Thiết lập tập dữ liệu và huấn luyện mô hình*

X_train, y_train là dữ liệu huấn luyện

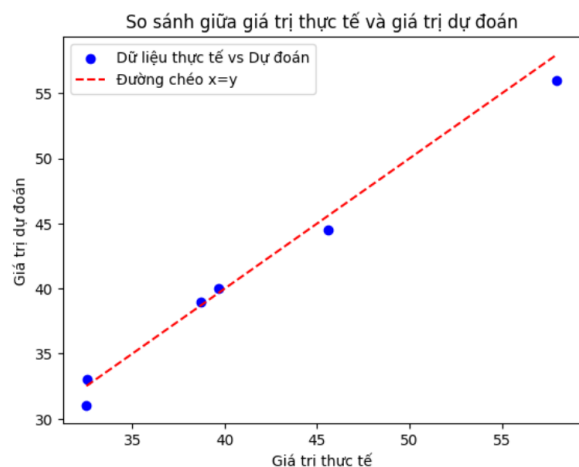
X_val, y_val là dữ liệu validation

```
model.fit(X_train, y_train, epochs=50, batch_size=32, validation_data=(X_val, y_val))
```

**Đánh giá và kiểm thử mô hình*

X_test, y_test là dữ liệu kiểm thử

```
loss = model.evaluate(X_test, y_test)
```



Đồ thị 2.1. So sánh giữa giá trị thực tế và giá trị dự đoán

- Các tham số đánh giá mô hình

+ Mean Squared Error (MSE): 1.2958977583333313

+ Mean Absolute Error (MAE): 0.9396166666666659

+ Root Mean Squared Error (RMSE): 1.1383750517001554

- Dựa vào kết quả mô hình có các giá trị tham số đánh giá, chúng ta có thể đưa ra một số nhận xét về hiệu suất của mô hình:

+ Mean Squared Error (MSE): 1.2959. Giá trị MSE mô tả mức độ bình phương trung bình của sai số giữa giá trị dự đoán và giá trị thực tế. Trong

trường hợp này, MSE có giá trị khá cao (1.2959), có thể cho thấy mô hình đang mắc phải mức độ sai số lớn khi dự đoán nồng độ bụi.

+ Mean Absolute Error (MAE): 0.9396. Giá trị MAE mô tả trung bình giá trị tuyệt đối của sai số. MAE có giá trị 0.9396, cho thấy mức độ chênh lệch trung bình giữa giá trị dự đoán và giá trị thực tế. Giá trị này cũng là một chỉ số quan trọng và thấp hơn so với MSE.

- RMSE: RMSE là căn bậc hai của MSE và giữ lại đơn vị của dữ liệu. Trong trường hợp này, giá trị RMSE là 1.1384, chỉ ra mức độ sai số trung bình giữa dự đoán và thực tế, nhưng vẫn có độ lớn cao.

Nhận xét chung, các giá trị MSE, MAE, và RMSE đều có giá trị khá cao, có thể làm nghi ngờ về độ chính xác của mô hình. Đối với bài toán ước lượng nồng độ bụi, có thể cần xem xét lại mô hình hoặc thử nghiệm các phương pháp tinh chỉnh để cải thiện hiệu suất

2.6. Ý nghĩa và ứng dụng

- *Ý nghĩa nghiên cứu*: Nghiên cứu này mang lại cái nhìn sâu rộng về khả năng ứng dụng của học máy trong dự đoán nồng độ bụi, đồng thời đưa ra những phát hiện quan trọng về mức độ ô nhiễm không khí tại Bắc Ninh.

- *Ứng dụng thực tế*: Kết quả của nghiên cứu có thể được áp dụng trong việc phát triển các hệ thống dự báo nhanh chóng và chính xác về chất lượng không khí. Điều này có thể hỗ trợ cơ quan quản lý môi trường và cộng đồng trong việc đưa ra quyết định và thực hiện các biện pháp giảm ô nhiễm.

3. Kết Luận

Nghiên cứu này đặt ra những bước quan trọng để hiểu và ứng dụng các thuật toán học máy trong dự báo chất lượng không khí. Kết quả đều cung cấp cơ sở cho những nghiên cứu tiếp theo và hướng phát triển, chú trọng vào mối quan hệ giữa ô nhiễm không khí và sức khỏe cộng đồng tại Bắc Ninh và có thể mở rộng ra các địa phương khác.

Tài liệu tham khảo

[1]. Gánh nặng từ vòng và kinh tế của bụi mịn pm2.5 tác động đến bệnh tim mạch tại thành phố Hồ Chí Minh năm 2018 - Tạp chí Y học dự phòng, Tập 31, số 6 - 2021

[2]. <https://aqicn.org/scale/>

[3]. <https://tainguyenvamoitruong.vn/so-tai-nguyen-va-moi-truong-bac-ninh-cid11051.html>

[4]. Prediction of wind pressure coefficients on building surfaces using Artificial Neural Networks