**Ứng dụng của CNN+LSTM trong dự báo PM2.5:**

**Dự đoán nồng độ PM2.5:** Mô hình CNN+LSTM có khả năng dự đoán nồng độ PM2.5 trong không khí dựa trên dữ liệu khí hậu và các yếu tố ảnh hưởng như nhiệt độ, độ ẩm và lượng mưa. Sự kết hợp giữa khả năng phân tích không gian của CNN và khả năng xử lý chuỗi thời gian của LSTM cho phép mô hình nắm bắt được các biến động theo thời gian và không gian của nồng độ PM2.5.

**Phân tích dữ liệu thời gian:** Mô hình CNN+LSTM giúp phân tích các chuỗi thời gian từ các trạm quan trắc, từ đó phát hiện các xu hướng và mẫu trong dữ liệu PM2.5. Điều này có thể hỗ trợ việc dự đoán các đợt ô nhiễm không khí trong tương lai.

**Tích hợp vào các mô hình dự báo truyền thống:** CNN+LSTM có thể được tích hợp vào các mô hình dự báo ô nhiễm không khí hiện có, nhằm cải thiện độ chính xác và giảm thiểu thời gian tính toán. Bằng cách kết hợp thông tin từ các mô hình truyền thống với khả năng học sâu của CNN+LSTM, có thể nâng cao hiệu quả trong việc dự báo nồng độ PM2.5.

**Ứng dụng của Machine Learning trong dự báo PM2.5:**

Dự đoán nồng độ PM2.5: Các mô hình Machine Learning, bao gồm CNN, LSTM và CNN+LSTM, có khả năng dự đoán nồng độ PM2.5 trong không khí dựa trên các yếu tố khí hậu như nhiệt độ, độ ẩm, áp suất và lượng mưa. Sự kết hợp giữa các phương pháp này giúp nắm bắt các biến động theo thời gian và không gian, cung cấp dự báo chính xác hơn về nồng độ ô nhiễm.

Phân tích dữ liệu lớn: Machine Learning có thể xử lý và phân tích các tập dữ liệu lớn từ nhiều nguồn khác nhau như trạm quan trắc ô nhiễm không khí, vệ tinh và các cảm biến môi trường. Điều này giúp nhận diện các xu hướng và mẫu trong dữ liệu PM2.5, từ đó hỗ trợ ra quyết định kịp thời trong việc bảo vệ sức khỏe cộng đồng.

Dự báo xu hướng ô nhiễm: Các mô hình Machine Learning có thể dự đoán xu hướng biến đổi nồng độ PM2.5 trong tương lai, giúp các cơ quan chức năng và cộng đồng chuẩn bị ứng phó với các đợt ô nhiễm không khí. Thông qua việc phân tích các yếu tố ảnh hưởng, mô hình có thể cung cấp thông tin chi tiết về nguyên nhân và thời điểm xảy ra ô nhiễm.

Tích hợp với mô hình khí hậu: Machine Learning có thể được tích hợp vào các mô hình khí hậu truyền thống để cải thiện độ chính xác trong việc mô phỏng và dự đoán nồng độ PM2.5. Sự kết hợp này giúp giảm thiểu độ phức tạp và thời gian tính toán, đồng thời nâng cao khả năng dự báo cho các tình huống ô nhiễm cụ thể.

Hệ thống cảnh báo sớm: Machine Learning có thể được sử dụng để phát triển hệ thống cảnh báo sớm cho nồng độ PM2.5, giúp cộng đồng nhận biết tình hình ô nhiễm không khí kịp thời và có các biện pháp phòng ngừa hiệu quả.

**Bối cảnh**

**Nguyên cứu từ CNN+LSTM**

**Phương pháp:**

Trong nghiên cứu này, chúng tôi áp dụng các phương pháp học sâu, cụ thể là Mạng Nơ-ron Tích chập (CNN) và Mạng Nơ-ron Bộ nhớ Dài Ngắn hạn (LSTM), để dự báo nồng độ PM2.5 trong không khí tại Thành phố Hồ Chí Minh. Mô hình CNN được sử dụng để khai thác các đặc điểm không gian từ dữ liệu thời gian thực, trong khi LSTM được triển khai để xử lý các mối quan hệ thời gian và xu hướng trong dữ liệu.

Dữ liệu đầu vào bao gồm các yếu tố thời tiết quan trọng như nhiệt độ, độ ẩm, và tốc độ gió, được thu thập từ năm 2018 đến 2023. Để nâng cao tính chính xác của các dự báo, dữ liệu đã được xử lý và làm mịn thông qua các phương pháp như chuẩn hóa và loại bỏ nhiễu. Các mô hình được phát triển và tinh chỉnh thông qua các kỹ thuật tối ưu hóa để cải thiện độ chính xác và khả năng tổng quát.

Kết quả từ mô hình sẽ được đánh giá thông qua các chỉ số như RMSE (Root Mean Square Error), MAE (Mean Absolute Error), và R² (Coefficient of Determination) để so sánh hiệu suất giữa các mô hình và xác định mô hình tối ưu nhất cho dự báo nồng độ PM2.5.

**Kết quả:**

**RMSE (Root Mean Square Error): 3.5072**

Giá trị RMSE thấp cho thấy mô hình dự đoán gần gũi hơn với dữ liệu thực tế. Với RMSE đạt 3.5072, mô hình cho thấy độ chính xác tốt trong việc dự đoán nồng độ PM2.5.

**MAE (Mean Absolute Error): 2.5244**

Với MAE là 2.5244, mô hình đã giảm thiểu được độ lệch trung bình giữa các giá trị dự đoán và thực tế, cho thấy nó có khả năng dự đoán nồng độ PM2.5 tương đối chính xác.

**R² (Coefficient of Determination): 0.9203**

Với R² là 0.9203, mô hình đã giải thích khoảng 92% biến động của nồng độ PM2.5, cho thấy mô hình có độ chính xác và hiệu quả cao trong việc dự đoán.

**Tóm lại**, các chỉ số RMSE, MAE và R² cho thấy mô hình CNN và LSTM đã hoạt động tốt trong việc dự đoán nồng độ PM2.5 tại Thành phố Hồ Chí Minh. Mô hình không chỉ có độ chính xác cao mà còn có khả năng giải thích tốt các yếu tố ảnh hưởng đến nồng độ PM2.5, từ đó hỗ trợ việc đưa ra các biện pháp cảnh báo và bảo vệ sức khỏe cộng đồng.

**Hạn Chế:**

Mặc dù mô hình CNN và LSTM cung cấp khả năng dự đoán nồng độ PM2.5 hiệu quả, vẫn tồn tại một số hạn chế đáng lưu ý:

***Dự đoán ngắn hạn*:**Các mô hình này có thể gặp khó khăn trong việc dự đoán chính xác các biến động ngắn hạn do độ phức tạp và tính biến đổi của dữ liệu khí hậu và môi trường. Sự biến đổi này có thể làm cho mô hình khó khăn trong việc nắm bắt các xu hướng tạm thời, dẫn đến dự đoán không chính xác trong những khoảng thời gian ngắn.

***Phụ thuộc vào dữ liệu*:**Mô hình học sâu thường phụ thuộc vào các giả định về dữ liệu đầu vào và có thể không phản ánh đầy đủ các điều kiện thực tế, đặc biệt khi dữ liệu đầu vào có nhiều yếu tố không chắc chắn hoặc bất định. Sự không chắc chắn trong dữ liệu đầu vào, chẳng hạn như giá trị sai lệch hoặc thiếu hụt, có thể dẫn đến những sai lệch trong dự đoán, làm giảm độ tin cậy của mô hình.

***Yêu cầu dữ liệu và tài nguyên***:Việc đào tạo các mô hình CNN và LSTM thường yêu cầu một lượng lớn dữ liệu và tài nguyên tính toán. Nếu dữ liệu không đủ phong phú hoặc đa dạng, khả năng tổng quát của mô hình có thể bị ảnh hưởng, dẫn đến hiệu suất không như mong đợi trong các tình huống thực tế.

**Nghiên cứu từ phương pháp học máy(Machine Learning)**

**Phương pháp**

Trong nghiên cứu này, chúng tôi áp dụng các phương pháp học máy (Machine Learning), cụ thể là các mô hình hồi quy như Hồi quy Rừng ngẫu nhiên (Random Forest), Hồi quy Vector Hỗ trợ (Support Vector Regression - SVR), và Hồi quy Tuyến tính (Linear Regression), để dự báo nồng độ PM2.5 trong không khí tại Thành phố Hồ Chí Minh. Các mô hình hồi quy được phát triển dựa trên dữ liệu thời gian thực của các yếu tố thời tiết từ năm 2018 đến 2023, bao gồm nhiệt độ, độ ẩm, gió, và thời gian. Dữ liệu đã được xử lý và làm mịn để tăng tính chính xác trong dự báo, và các mô hình được tinh chỉnh nhằm tối ưu hóa kết quả.

* **Rừng Ngẫu nhiên (Random Forest)**:

Là một mô hình học máy mạnh mẽ, Rừng ngẫu nhiên sử dụng một tập hợp các cây quyết định để cải thiện độ chính xác dự đoán và giảm thiểu vấn đề quá khớp (overfitting). Mô hình này có khả năng xử lý tốt các dữ liệu không đồng nhất và cung cấp thông tin về tầm quan trọng của từng yếu tố đầu vào trong dự đoán nồng độ PM2.5.

* **Hồi quy Tuyến tính (Linear Regression)**:

Là một trong những phương pháp cơ bản nhất trong học máy, hồi quy tuyến tính tìm mối quan hệ tuyến tính giữa các biến độc lập (nhiệt độ, độ ẩm, gió, thời gian) và biến phụ thuộc (nồng độ PM2.5). Mặc dù đơn giản, phương pháp này có thể cung cấp một cơ sở tốt để so sánh với các mô hình phức tạp hơn.

* **Hỗ trợ Vector Hồi quy (SVR - Support Vector Regression)**:

SVR là một phiên bản của phương pháp Hỗ trợ Vector cho hồi quy. Mô hình này tìm kiếm các hyperplane tối ưu để tối đa hóa khoảng cách giữa các điểm dữ liệu và hyperplane. SVR có thể hoạt động tốt trong các trường hợp dữ liệu không tuyến tính và có thể được điều chỉnh để phù hợp với nhiều loại dữ liệu khác nhau.

**Kết quả**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Time steps | SVR(linear) | SVR(poly) | Linear | Random Forest |
| 6 | Test RMSE: 5.8527  Test MAE: 5.0630  Test R2: 0.7781 | Test RMSE: 7.1252  Test MAE: 5.8979  Test R2: 0.6711 | Test RMSE: 3.1501  Test MAE: 2.2171  Test R2: 0.9357 | Test RMSE:3.3285  Test MAE: 2.3216  Test R2: 0.9282 |
| 12 | Test RMSE: 6.0264  Test MAE: 5.1875  Test R2: 0.7647 | Test RMSE: 7.0151  Test MAE: 5.7734  Test R2: 0.6812 | Test RMSE: 2.9453  Test MAE: 2.0559  Test R2: 0.9438 | Test RMSE:3.3442  Test MAE: 2.3338  Test R2: 0.9275 |
| 24 | Test RMSE: 6.3873  Test MAE: 5.5776  Test R2: 0.7357 | Test RMSE: 7.2302  Test MAE: 5.9796  Test R2: 0.6613 | Test RMSE: 2.8112  Test MAE: 1.9506  Test R2: 0.9488 | Test RMSE:3.3431  Test MAE: 2.3245  Test R2: 0.9276 |

**Với Time Steps = 12h**

Với Time Steps là 12h,Các mô hình hồi quy (Hồi quy Tuyến tính, SVR, Rừng Ngẫu nhiên) đã được đánh giá dựa trên các chỉ số RMSE, MAE và R². Mô hình Hồi quy Tuyến tính cho thấy hiệu suất tốt nhất với RMSE thấp (2.9453) và R² cao (0.9438), cho thấy khả năng dự đoán nồng độ PM2.5 rất chính xác. Mô hình Rừng Ngẫu nhiên cũng đạt được kết quả khả quan nhưng kém hơn một chút (RMSE: 3.3442, R²: 0.9275). Trong khi đó, các mô hình SVR (cả tuyến tính và polynomial) có độ chính xác thấp hơn, cho thấy không phù hợp cho bài toán dự đoán nồng độ PM2.5. Từ các kết quả này, các mô hình CNN và LSTM có thể tiếp tục được tối ưu hóa để đưa ra các biện pháp cảnh báo và bảo vệ sức khỏe cộng đồng hiệu quả hơn.

**Hạn Chế**

Mặc dù các mô hình hồi quy đã cho thấy hiệu suất tốt trong việc dự đoán nồng độ PM2.5, vẫn tồn tại một số hạn chế cần được xem xét:

**Khó khăn trong việc dự đoán biến động ngắn hạn**: Các mô hình này có thể gặp khó khăn trong việc dự đoán chính xác các biến động ngắn hạn do độ phức tạp và tính biến đổi của dữ liệu. Sự biến đổi này có thể làm cho mô hình khó khăn trong việc nắm bắt các xu hướng tạm thời, dẫn đến dự đoán không chính xác trong những khoảng thời gian ngắn.

**Phụ thuộc vào giả định về dữ liệu đầu vào**:Mô hình học sâu thường dựa vào các giả định về dữ liệu đầu vào và có thể không phản ánh đầy đủ các điều kiện thực tế. Điều này đặc biệt đúng khi dữ liệu đầu vào có nhiều yếu tố không chắc chắn hoặc bất định, dẫn đến những sai lệch trong dự đoán và giảm độ tin cậy của mô hình.

**Yêu cầu lớn về dữ liệu và tài nguyên tính toán**:Việc đào tạo các mô hình như CNN và LSTM thường cần một lượng lớn dữ liệu và tài nguyên tính toán. Nếu dữ liệu không đủ phong phú hoặc đa dạng, khả năng tổng quát của mô hình sẽ bị ảnh hưởng, dẫn đến hiệu suất không như mong đợi trong các tình huống thực tế.

**Nguyên cứu từ mô hình CNN**

**Phương pháp**

Mạng Nơ-ron Tích chập (CNN) là một loại mạng nơ-ron học sâu, được thiết kế đặc biệt để xử lý các dữ liệu có cấu trúc dạng lưới như hình ảnh, nhưng cũng có thể áp dụng hiệu quả trong các bài toán khác có đặc điểm không gian. Trong bối cảnh dự báo nồng độ PM2.5, CNN được sử dụng để phát hiện các mẫu không gian trong dữ liệu môi trường, giúp mô hình hiểu được mối liên hệ giữa các yếu tố như nhiệt độ, độ ẩm, tốc độ gió và nồng độ PM2.5.

**Kết quả**

|  |  |
| --- | --- |
| Time steps | CNN |
| 6 | Test RMSE: 3.0425  Test MAE: 2.1177  Test R2: 0.9400 |
| 12 | Test RMSE: 3.1271  Test MAE: 2.1709  Test R2: 0.9367 |
| 24 | Test RMSE: 3.0956  Test MAE: 2.1500  Test R2: 0.9379 |

**Với Time Steps = 12h**

Mô hình CNN cho kết quả khá ấn tượng với RMSE và MAE ở mức chấp nhận được và R² đạt gần mức tối ưu. Dựa trên các chỉ số này, mô hình CNN có thể coi là một lựa chọn khá cạnh tranh trong việc dự đoán nồng độ PM2.5, và có thể được tối ưu hóa thêm để cải thiện độ chính xác. So với các mô hình khác, CNN cho thấy khả năng học và dự đoán có triển vọng.