SỬ DỤNG MÔ HÌNH CHUỖI THỜI GIAN VÀ HỌC SÂU ĐỂ DỰ ĐOÁN NHIỆT ĐỘ

Nguyễn Thị Huyền Trang[1,2] and Nguyễn Thị Mai Trinh[1,2]

[1] Trường đại học Công nghệ Thông Tin Đại học quốc gia
[2] Đường Hàn Thuyên, Thủ Đức, Thành phố Hồ Chí Minh, Việt Nam {21520488, 21522718}@gm.uit.edu.vn

Tóm tắt nội dung Trước sự biến đổi không ngừng của khí hậu và tình trạng nóng lên toàn cầu, nhiệt độ trở thành một yếu tố quyết định đối với cuộc sống và công việc hàng ngày. Việc theo dõi mẫu nhiệt độ theo ngày càng cần chính xác hơn, đây là một thách thức quan trọng trong việc dự báo nhiệt độ. Báo cáo này hướng tới phát triển hệ thống dự báo nhiệt độ sử dụng mô hình Chuỗi thời gian và mô hình Học sâu trên dữ liệu về thời tiết của Thành phố Hồ Chí Minh - Việt Nam. Nghiên cứu tập trung vào việc dự đoán nhiệt độ theo mốc 3h 1 lần và so sánh bốn mô hình: ARIMA, SARIMA, LSTM, CNN-LSTM và LSTMs . Nhóm đã tiến hành các thực nghiệm để tìm ra mô hình và phương pháp phù hợp cho nhiệm vụ dự báo nhiệt độ trên bộ dữ liệu đã thu thập. Kết quả thu được cho thấy hiệu quả và hạn chế của từng mô hình, phương pháp trong bài toán dự đoán nhiệt độ.

Keywords: Chuỗi thời gian · Học sâu · Dự báo nhiệt độ · ARIMA · SARIMA · LSTM · CNN-LSTM · LSTMs

1 GIỚI THIỆU

Trong ngữ cảnh của biến đổi khí hậu lan rộng toàn cầu, việc dự báo nhiệt độ trở nên vô cùng quan trọng để hiểu và ứng phó với sự biến đổi không ngừng của thời tiết. Bằng cách sử dụng các phương pháp Học máy, chúng ta có thể phân tích các mẫu dữ liệu lịch sử dạng chuỗi thời gian và dự báo nhiệt độ trong tương lai với độ chính xác cao.

Nghiên cứu này tập trung vào việc thực hiện các thử nghiệm sử dụng mô hình Chuỗi thời gian và mô hình Học sâu trên dữ liệu về nhiệt độ của Thành phố Hồ Chí Minh. Mục tiêu của chúng tôi là dự đoán nhiệt độ và so sánh năm mô hình khác nhau: ARIMA (AutoRegressive Integrated Moving Average), SARIMA (Seasonal AutoRegressive Integrated Moving Average), LSTM (Long Short-Term Memory Networks), CNN-LSTM (mô hình kết hợp CNN và LSTM) và LSTMs (Multi-variable Long Short-Term Memory Networks) trong việc dự đoán.

Thực nghiệm của chúng tôi nhằm tìm ra các bước, yếu tố quan trọng trong mỗi mô hình cho bài toán dự báo nhiệt độ dựa trên thời gian, độ chính xác, khả

năng xử lý trên bộ dữ liệu đã thu thập. Kết quả nghiên cứu sẽ giúp chúng ta hiểu rõ hiệu quả và hạn chế của mỗi mô hình, ý nghĩa một số phương pháp, yếu tố trong ứng dụng dự đoán nhiệt độ. Nhóm hy vọng báo cáo này có thể cung cấp các phương pháp dự báo nhiệt độ chính xác và tin cậy, hỗ trợ trong việc ra quyết định và lựa chọn phù hợp dựa trên các dự báo, dự đoán nhiệt độ.

Cấu trúc bài báo cáo như sau. Chi tiết về bộ dữ liệu và phương pháp xây dựng được trình bày trong phần 2. Phần 3 là quy trình thực hiện, cho biết các công việc mà chúng em thực hiện trong đề tài này. Nhóm phân tích kết quả thực nghiệm và nhận xét về hiệu xuất các mô hình trong phần 4. Tại phần 5 của báo cáo, chúng em đề cập đến lỗi và hướng phát triển của bài toán. Cuối cùng, kết luận được trình bày trong phần 6 của báo cáo.

2 BỘ DỮ LIỆU

2.1 Nguồn và quá trình thu thập

Phục vụ cho bài toán, nhóm thu thập dữ liệu từ trang web WorldWeatherOnline - một trang web cung cấp dữ liệu thời tiết đáng tin cậy.

Nhóm dùng thư viện selenium của python để thu thập dữ liệu. Bộ dữ liệu được thu thập là thông tin thời tiết trong khoảng thời gian từ 01/01/2020 đến 30/04/2023, theo từng thời điểm cách nhau 3 giờ trong ngày ,số liệu này được lấy ở ngày tháng cụ thể của từng điểm dữ liệu thời tiết được ghi lại rõ ràng, theo thứ tự. Dữ liệu thô gồm các thuộc tính: Hour, Temperature, Forecast, Rain, Rain-rate, Cloud-rate, Pressure, Wind, Gust, Day, Month, Year, Weather.

Các thuộc tính được mô tả trong Bảng 1

2.2 Tiền xử lý dữ liệu

Sau khi thu thập dữ liệu, nhóm đã kiểm tra tổng quan dữ liệu và tiến hành tiền xử lý.

- Loại bỏ dữ liệu nhiễu: Dữ liệu thực tế thu thập được chứa những thông tin không cần thiết như đơn vị đo, v.v... Nhóm xử lý chúng bằng hàm *drop()* để xóa bỏ các cột không cần thiết. Điều này giúp làm sạch dữ liệu và làm giảm khả năng mô hình bị ảnh hưởng bởi dữ liệu không chính xác. Ngoài ra dữ liệu bị thiếu mất tên các thuộc tính, sử dụng hàm *rename()* để bổ sung tên thuộc tính, kiểm tra các ô dữ liệu trống và lấp đầy.
- Chuẩn hóa dữ liệu: Dữ liệu thời tiết xuất hiện trong nhiều định dạng và phân phối khác nhau. Nhóm chuẩn hóa giá trị của các cột sang kiểu dữ liệu số (float) để phù hợp với các phân tích, chuẩn hóa giá trị về đoạn [0,1]. Điều này đảm bảo tính nhất quán và đồng nhất của dữ liệu, giúp cho mô hình có thể hiểu và học từ dữ liệu một cách hiệu quả hơn.

Tên thuộc tính	Mô tả thuộc tính
Hour	Giờ
Temperature	Nhiệt độ (độ C)
Forecast	Nhiệt độ cảm nhận được (độ C)
Rain	Lượng mưa (mm)
Rain-rate	Độ ẩm (phần trăm)
Cloud-rate	Độ che phủ của mây (phần trăm)
Pressure	Áp suất (mb)
Wind	Tốc độ gió trung bình (km/h)
Gust	Tốc độ gió cao nhất (km/h)
Day	Ngày
Month	Tháng
Year	Năm
Weather	Thời tiết

Bảng 1. Bảng mô tả các thuộc tính

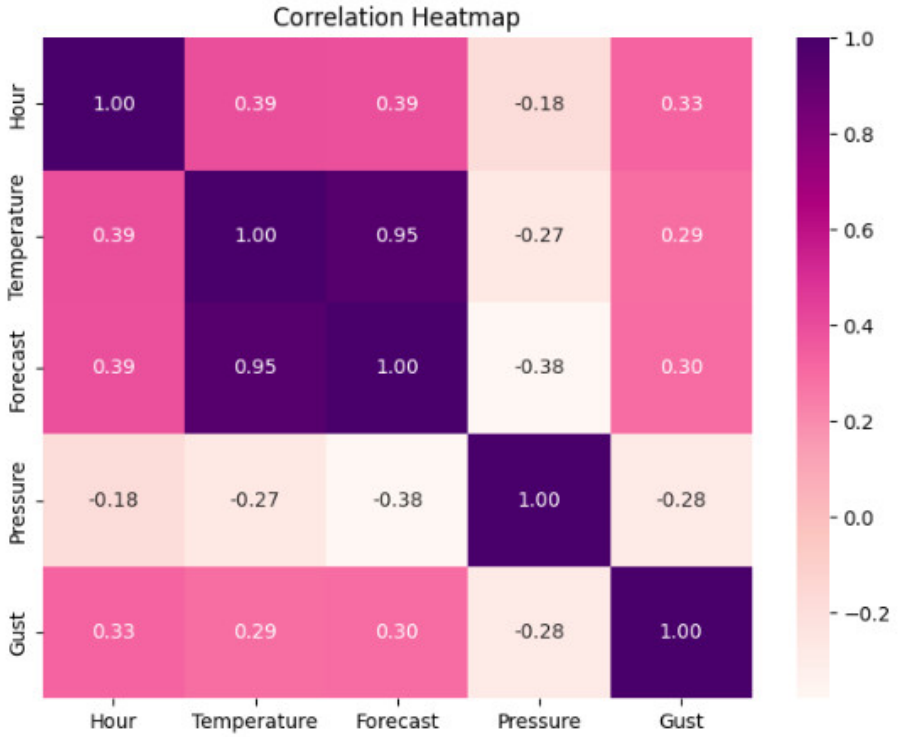
	feature	Test statistics	P-value	99% confidence	95% confidence	90% confidence
0	Hour	-2.894939e+14	0.000000e+00	-3.431023 -> stationary	-2.861837 -> stationary	-2.566928 -> stationary
1	Temperature	-5.329000e+00	4.764035e-06	-3.431025 -> stationary	-2.861838 -> stationary	-2.566929 -> stationary
2	Forecast	-5.579000e+00	1.407882e-06	-3.431025 -> stationary	-2.861838 -> stationary	-2.566929 -> stationary
3	Pressure	-7.014000e+00	6.819626e-10	-3.431025 -> stationary	-2.861838 -> stationary	-2.566929 -> stationary
4	Gust	-9.436000e+00	5.032719e-16	-3.431025 -> stationary	-2.861838 -> stationary	-2.566929 -> stationary

Bảng 2. Bảng ADF của bộ dữ liệu

2.3 Phân tích dữ liệu

Dựa vào Bảng 2 - bảng ADF của bộ dữ liệu; Hình 1 - ma trận tương quan; và để phù hợp cho bài toán, nhóm chỉ giữ lại 5 thuộc tính có ảnh hưởng nhất đến biến mục tiêu là: Hour, Temperature, Forecast, Pressure, Gust.

Dữ liệu mục tiêu là nhiệt độ (Temperature) có tính chất chu kỳ và biến động theo thời gian nên nhóm lựa chọn sử dụng mô hình chuỗi thời gian và mô hình



Hình 1. Ma trận tương quan

học sâu để dự đoán. Nhóm thực hiện 5 mô hình Chuỗi thời gian và Học sâu để dự đoán nhiệt độ, bao gồm:

- Mô hình chuỗi thời gian: ARIMA (AutoRegressive Integrated Moving Average), SARIMA (Seasonal AutoRegressive Integrated Moving Average)
- Mô hình học sâu: LSTM (Long Short-Term Memory Networks), CNN-LSTM (mô hình kết hợp CNN và LSTM), và LSTMs (Multi-variable Long Short-Term Memory Networks)

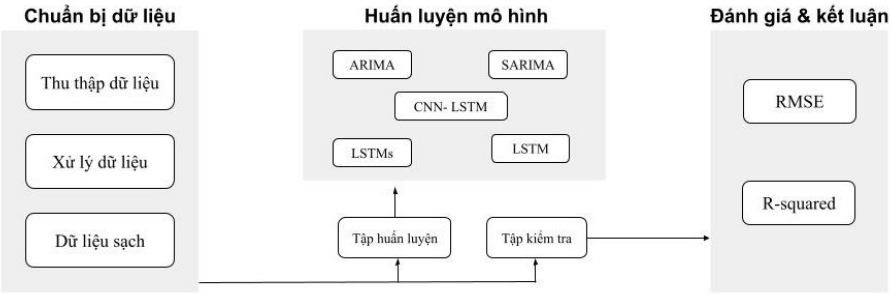
3 THỰC NGHIỆM

3.1 Quy trình thực hiện

Sau khi thực hiện tiền xử lý và làm sạch, nhóm thu được một bộ dữ liệu sạch. Sau đó, dữ liệu được chia thành tập huấn luyện (với tỷ lệ là 80%) và tập kiểm tra (với tỷ lệ là 20%). Tập dữ liệu huấn luyện được sử dụng để huấn luyện hệ thống dự báo nhiệt độ với các mô hình dự báo chuỗi thời gian và mô hình học sâu bao gồm: ARIMA, SARIMA, CNN-LSTM, LSTM và LSTMs. Tất cả các mô

hình trên được áp dụng trên tập dữ liệu thử nghiệm và sử dụng hai độ đo là RMSE và R-squared để so sánh và đánh giá.

Quy trình xây dựng mô hình được thể hiện trong Hình 2



Hình 2. Quy trình xây dựng mô hình

3.2 Mô hình CNN-LSTM

Mô hình CNN-LSTM là sự kết hợp một lớp Convd cùng với 1 lớp LSTM. Nhóm xây dựng với hi vọng rằng: Các bộ lọc của CNN có thể phát hiện các mẫu thời gian quan trọng, như chuỗi nhiệt độ đột ngột tăng cao hoặc giảm đột ngột. Sau đó, lớp LSTM để mô hình hóa mối quan hệ giữa các đặc trưng thời gian và dự báo các giá trị nhiệt độ tiếp theo trong chuỗi.

Bằng cách này, mô hình có thể học được cả các mẫu thời gian quan trọng và mối quan hệ giữa chúng để dự báo nhiệt độ trong tương lai. Nếu thành công có thể tối ưu hóa hiệu suất dự báo trên bộ dữ liệu chuỗi thời gian.

3.3 Độ đo đánh giá

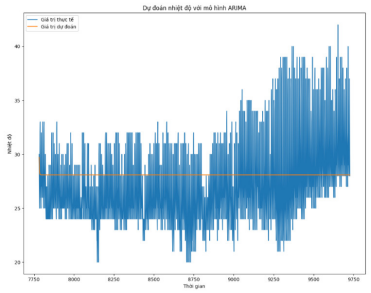
Nhóm sử dụng hai độ đo đánh giá là **RMSE** (Root Mean Squared Error)[2] và **R - Squared** (còn gọi là hệ số xác định)[3].

3.4 So sánh hiệu suất các mô hình

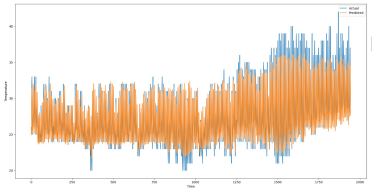
Bảng 3 là Bảng kết quả độ đo đánh giá các mô hình

4 KẾT QUẢ THỰC NGHIỆM

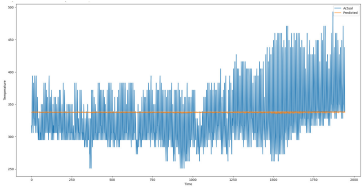
Báo cáo này tập trung thử nghiệm phương pháp học máy trong bài toán dự báo nhiệt độ. Cụ thể là 5 mô hình (ARIMA, SARIMA, LSTM, CNN-LSTM và LSTMs) được triển khai và so sánh về hiệu suất.



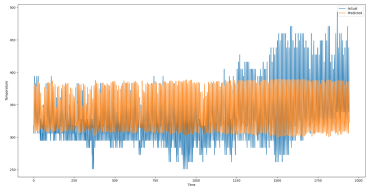
Hình 3. Biểu độ dự đoán nhiệt độ của mô hình ARIMA



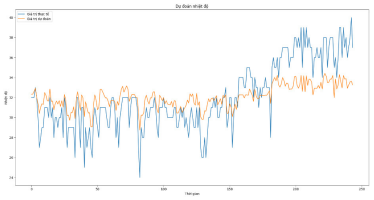
Hình 4. Biểu độ dự đoán nhiệt độ của mô hình LSTM có dùng Min-Max Scaler



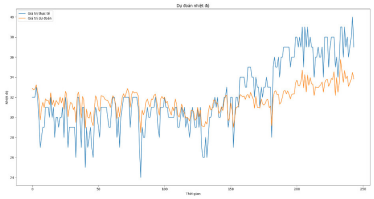
Hình 5. Biểu độ dự đoán nhiệt độ của mô hình CNN-LSTM



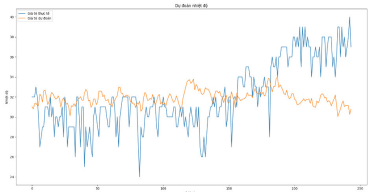
Hình 6. Biểu độ dự đoán nhiệt độ của mô hình CNN-LSTM có dùng Min-Max Scaler



Hình 7. Biểu độ dự đoán nhiệt độ của mô hình LSTMs (Tổ hợp đầu vào Forecast, Pressure)



Hình 8. Biểu độ dự đoán nhiệt độ của mô hình LSTMs (Tổ hợp đầu vào Forecast, Gust)



Hình 9. Biểu độ dự đoán nhiệt độ của mô hình LSTMs (Tổ hợp đầu vào Pressure, Gust)

Model		RMSE	R2
ARIMA		3.90195	-0.039313
SARIMA		1.602933	0.91974
LSTM không dùng Min-Max Scaler		1.268084	0.890316
LSTM dùng Min-Max Scaler		14.318538	0.884426
CNN-LSTM không dùng Min-Max Scaler		42.07519	0.00204
CNN-LSTM dùng Min-Max Scaler		25.611774	0.630222
LSTMs không dùng Min-Max Scaler đầu vào là cả 5 biến		1.428089	0.860782
LSTMs dùng Min-Max Scaler	Cả 5 biến	0.026087	0.999953
	Temperature, Forecast	1.263757	0.856596
	Temperature, Pressure	1.670558	0.749414
	Temperature, Gust	1.501059	0.797684
	Forecast, Pressure	2.372906	0.494413
	Forecast, Gust	2.316165	0.518303
	Pressure, Gust	3.520659	-0.112968
	Temperature, Forecast, Pressure	1.587377	0.773747
	Temperature, Forecast, Gust	1.358844	0.834204
	Temperature, Pressure, Gust	2.117579	0.597363
	Forecast, Pressure, Gust	2.197493	0.5664
	Temperature, Forecast, Pressure, Gust	1.363009	0.833186

Bảng 3. Bảng kết quả độ đo đánh giá các mô hình

4.1 Mô hình ARIMA

Mô hình ARIMA cần được cải thiện: Với giá trị âm của R-squared (-0.039313) và RMSE (3.90195) cũng khá cao, mô hình ARIMA không thể cho ra dự báo tốt. Nhận định này có thể được thấy rõ qua [Hình 4], đường nhiệt độ dự đoán chỉ là 1 đường thẳng, tức nhiệt độ gần như không đổi, điều này là vô lý. Mô hình ARIMA cần được cải thiện hoặc điều chỉnh để cải thiện hiệu suất dự báo của nó, đặc biệt nó cần giải quyết được tính mùa vụ của dữ liệu.

4.2 Mô hình SARIMA

Mô hình SARIMA dù không sử dụng phương pháp Scale dữ liệu vẫn cho kết quả dự đoán rất tốt (1.6029 RMSE, 0.9197 R-squared), dự đoán tương đối chính xác xu hướng diễn biến nhiệt độ. Kết quả dự đoán tốt cũng khẳng định SARIMA có khả năng xử lý tốt yếu tố mùa vụ trong dữ liệu mà ARIMA không làm được. Tuy nhiên, mô hình không thích ứng tốt với những biến đổi của dữ liệu theo thời gian.

4.3 Mô hình LSTM

Mô hình LSTM dù sử dụng phương pháp Scale dữ liệu hay không cũng cho kết quả dự đoán tương đối khả quan, độ đo R-squared đều trên 0.88. Mô hình LSTM không dùng Min-Max Scaler có RMSE thấp (1.27) và R-squared cao (0.89). Mô hình LSTM dùng Min-Max Scaler có RMSE cao (14.318538) và R-squared thấp hơn (0.88), nhiệt độ dự đoán và nhiệt độ thực tế chênh lệch khá lớn như [Hình 4]. Điều này cho thấy việc sử dụng Min-Max Scaler trong trường hợp này gây phản tác dụng, ảnh hưởng xấu đến năng suất mô hình.

4.4 Mô hình CNN-LSTM

Mô hình CNN-LSTM khi không dùng Min-Max Scaler cho kết quả dự đoán rất tệ, R-squared chỉ 0.002. Trong khi đó, đô đo mô hình CNN-LSTM sử dụng MinMaxScaler đã cải thiện rất nhiều: RMSE từ 42.08 giảm xuống 25.61, R-squared tăng từ 0.002 lên 0.63. Kết quả cho thấy việc sử dụng Min-Max Scaler có thể làm tăng RMSE và giảm R-squared của mô hình CNN-LSTM. Tuy nhiên các độ đo vẫn không quá khả quan, mô hình này không dự báo tốt, nhiệt độ được dự đoán không đáng tin cậy, thấy rõ tại [Hình 6].

4.5 Mô hình LSTMs

- Mô hình LSTMs dự báo nhiệt độ khá tốt kể cả khi không thực hiện phương pháp Scale dữ liệu. Điều này được thể hiện bởi con số độ đo RMSE 1.428089, và R-squared 0.860782.
- Mô hình LSTMs khi sử dụng phương pháp MinMaxScaler cho hiệu suất mô hình khá tốt, và phân hoá theo từng tổ hợp biến đầu vào.
- Việc lựa chọn biến đầu vào cho mô hình dự đoán LSTMs là vô cùng quan trọng. Cụ thể quan sát [Bảng 3] cùng là LSTMs nhưng với các tổ hợp biến đầu vào khác nhau cho ra kết quả dự đoán chênh lệch rất lớn.
- Việc sử dụng nhiều biến đầu vào có thể cung cấp thông tin đa dạng và cải thiện khả năng dự đoán so với việc chỉ sử dụng một biến đơn lẻ. (Mô hình LSTMs với 5 biến đầu vào cho độ đo R-squared 0.999953 cao hơn đáng kể so với R-squared 0.884426 của mô hình LSTM đơn biến)
- Các biến có thể có mức độ ảnh hưởng khác nhau. Điều này được thấy qua việc cùng là bộ 2 đầu vào và có sự góp mặt của biến 'Forecast' thì khi kết hợp với biến 'Temperature', độ đo RMSE đạt 1.2638, R-squared đạt 0.8566, kết hợp cùng biến 'Gust' mô hình đạt được RMSE 2.3162 và R-squared 0.5183, khi kết hợp cùng biến 'Pressure' lại thu được RMSE 2.3729 và R-squared 0.5183.
- Đầu vào 'Temperature' xuất hiện trong 7 tập hợp đầu vào và cho kết quả dự đoán tương đối tốt (độ đo R-squared đạt trong khoảng 0.5974 đến 0.99). Điều này cho thấy 'Temperature' là một biến quan trọng trong việc dự báo nhiệt độ (Temperature). Điều này hoàn toàn hợp lý.
- Kết quả dự đoán cùng với độ chính xác có thể cung cấp thông tin về sự ảnh hưởng của các biến đầu vào lên biến mục tiêu (Temperature).

– Không nên lựa chọn các biến đầu vào có mối tương quan thấp hoặc không có quan hệ rõ ràng với biến mục tiêu. Việc này có thể làm giảm đáng kể hiệu suất mô hình dự đoán. Cụ thể, 3 tổ hợp biến đầu vào là 'meanpressure' - 'Forecast' và 'Pressure' - 'Forecast' và 'Gust' - 'Pressure' và 'Gust' (các tổ hợp chứa 'Pressure' và 'Gust' - các thuộc tính gần như không tương quan với biến mục tiêu, độ tương quan < 0.3) đều cho hiệu quả mô hình không cao, đạt R-squared lần lượt 0.4944, 0.5183 và -0.1130, độ đo RMSE cũng khá lớn [Bảng 3]. Quan sát Hình[Hình 7], Hình 8], [Hình 9] cho thấy với ba tổ hợp biến đầu vào trên mô hình LSTMs không thể dự đoán nhiệt độ.

4.6 Nhận xét chung

Thực nghiệm mà nhóm thực hiện đã làm rõ hơn về ưu nhược điểm của từng mô hình [Bảng 4]. Báo cáo giúp nhóm hiểu hơn về tác dụng, ý nghĩa của một vài yếu tố, phương pháp như: yếu tố mùa vụ trong dữ liệu, kĩ thuật Scale dữ liệu cụ thể là phương pháp MinMaxScaler, ảnh hưởng của độ tương quan giữa biến đầu vào và biến mục tiêu,..

5 PHÂN TÍCH LỖI VÀ HƯỚNG PHÁT TRIỂN

Nguyên nhân mô hình ARIMA được xây dựng nhưng không thể dự đoán chính xác là do ARIMA không xử lý được các yếu tố mùa vụ có trong dữ liệu mục tiêu (Temperature).

Mô hình ARIMA đã được nâng cấp và thay thế bởi mô hình tiên tiến hơn đó là SARIMA. Mô hình kết hợp CNN-LSTM cho khả năng dự đoán rất thấp khi không sử dụng MinMaxScaler, R-squared chỉ đạt 0.00204, mô hình này không thể dự đoán nhiệt độ.

Ba mô hình còn lại là SARIMA, LSTM và LSTMs dù vẫn còn tồn tại khuyết điểm như việc SARIMA không thích ứng tốt với sự biến đổi của dữ liệu, LSTM đòi hỏi lượng lớn dữ liệu, LSTMs cần xác định biến đầu vào phù hợp, ... nhưng về cơ bản các mô hình này đều có khả năng dự đoán nhiệt độ trên bộ dữ liệu đã thu thập.

Mô hình dự báo nhiệt độ vẫn còn nhiều tiềm năng để phát triển như việc sử dụng các phương pháp tinh chỉnh tham số, tăng cường đặc trưng khác nhau vào mô hình hay xây dựng mô hình kết hợp.

6 KẾT LUẬN

Nhóm đã xây dựng thành công hệ thống dự báo nhiệt độ trên bộ dữ liệu được thu thập. Qua thực nghiệm, nhóm đã tìm ra được mô hình dự đoán cho kết quả tốt nhất trên bộ dữ liệu: mô hình LSTMs với cả 5 biến đầu vào có sử dụng MinMaxScaler, với độ đo RMSE:0.026, R-squared: 0.99, giá trị dự đoán rất sát với thực tế.

Mô hình	Ưu điểm	Nhược điểm
ARIMA	- Dễ hiểu và triển khai - Có thể xử lý các chuỗi thời gian không mô phỏng được bằng các phương pháp đồng thời - Tính toán tương đối nhanh với dữ liệu có kích thước nhỏ	- Không xử lý được các mối quan hệ phi tuyến và phức tạp trong dữ liệu, các yếu tố mùa vụ mạnh - Yêu cầu các giả định về tuyến tính, không nhiễu và phân phối chuẩn
SARIMA	- Xử lý tốt dữ liệu có yếu tố mùa vụ - Xây dựng trên lý thuyết chặt chẽ của ARIMA, có khả năng ước lượng các tham số một cách hiệu quả - Cung cấp thông tin đáng tin cậy	- Yêu cầu dữ liệu phải: ổn định, tính tuyến tính, tính chu kỳ của chuỗi thời gian - Không thích ứng tốt với sự biến đổi của dữ liệu
LSTM	- Xử lý tốt dữ liệu có mối quan hệ phụ thuộc dài hạn - Có khả năng mô hình hóa mối quan hệ phi tuyến và xử lý dữ liệu nhiễu tốt - Khả năng tự học và ghi nhớ các mẫu dữ liệu dài hạn và phức tạp - Có thể thích ứng với sự biến đổi của dữ liệu theo thời gian và dự báo các xu hướng và mẫu phức tạp	- Đòi hỏi lượng lớn dữ liệu huấn luyện để được kết quả tốt - Cần điều chỉnh kỹ thuật huấn luyện như kích thước batch, số lớp ẩn, số đơn vị ẩn để đạt hiệu suất tốt - Tính toán phức tạp, thời gian huấn luyện lâu
CNN-LSTM	- Có khả năng cải thiện kết quả - Mô hình thể hiện khả năng xử lý dữ liệu chuỗi thời gian phức tạp. - Tích hợp ưu điểm của CNN và LSTM	- Đòi hỏi phải lựa chọn phương pháp xử lý dữ liệu một cách cẩn thận - Hiệu suất dự báo của mô hình vẫn chưa đạt mức khả quan so với các mô hình khác.
LSTMs	- Cải thiện khả năng dự đoán nhờ việc kết hợp nhiều biến đầu vào để hiểu rõ hơn các mối quan hệ giữa chúng - Xử lý tốt các dữ liệu đa biến - Cung cấp khả năng dự đoán chính xác hơn	- Đòi hỏi có đủ dữ liệu và hiểu rõ mối quan hệ giữa các biến đầu vào - Cần đảm bảo tính nhất quán và chuẩn hóa các biến đầu vào - Khó khăn trong việc xác định, tối ưu hóa số lượng và cách kết hợp các biến đầu vào

Bảng 4. Bảng so sánh ưu, nhược điểm các mô hình

Mô hình ARIMA và CNN-LSTM không sử dụng MinMaxScaler không có khả năng dự đoán nhiệt độ trên bộ dữ liệu. Mô hình CNN-LSTM khi sử dụng MInMaxScaler đã cho kết quả dự đoán cải thiện rất nhiều: R-squared từ 0.002 lên 0.63. Qua đây, nhóm thấy rõ hơn về tác dụng của việc sử dụng MinMaxScaler trong quá trình huấn luyện mô hình.

Các mô hình SARIMA, LSTM cho kết quả dự đoán tốt, độ đo R-squared đều trên 0.88. Điều này khẳng định tính tối ưu của SARIMA so với ARIMA bởi việc giải quyết được tính mùa vụ có trong dữ liệu.

Tuy nhiên, lưu ý ở sự thay đổi độ đo của mô hình CNN-LSTM và LSTM khi sử dụng MinMaxScaler, ta cần sử dụng các phương pháp xử lý dữ liệu như MinMaxScaler hợp lý tránh tác dụng ngược.

Hiệu suất dự đoán của mô hình LSTMs phụ thuộc vào tổ hợp biến đầu. Vì vậy việc việc lựa chọn biến đầu vào phù hợp cho mô hình là vô cùng quan trọng.

Từ đây ta thấy được việc lựa chọn mô hình phù hợp, nhóm thuộc tính phù hợp và phương pháp xử lý phù hợp với bộ dữ liệu cụ thể giúp nâng cao hiệu suất cao, độ chính xác của mô hình trong giải quyết bài toán bất kỳ.

Tài liệu

[1] "World Weather API and Weather Forecast", 2024. [Trực tuyến]. Địa chỉ: https://www.worldweatheronline.com/

[2] “Root Mean Square Error (RMSE)”, 2023. [Trực tuyến]. Địa chỉ: https://c3.ai/glossary/data-science/root-mean-square-error-rmse/

[3] Jason Fernando, "R-Squared: Definition, Calculation Formula, Uses, and Limitations", 2023. [Trực tuyến]. Địa chỉ: https://www.investopedia.com/terms/r/r-squared.asp

[4] Anqi Xie, Hao Yang, Jing Chen, Li Sheng and Qian Zhang, "A Short-Term Wind Speed Forecasting Model Based on a Multi-Variable Long Short-Term Memory Network", 2021. [Trực tuyến]. Địa chỉ: https://www.mdpi.com/2073-4433/12/5/651