Dự báo thời tiết

Huy Chan Huynh, Trang Huyen Thi Nguyen
University of Information Technology, Ho Chi Minh City, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam
{19521609, 21520488}@gm.uit.edu.vn

Abstract

"Dự báo thời tiết" nhằm nghiên cứu và phát triển một hệ thống dự báo thời tiết chính xác và hiệu quả cho khu vực Thành phố Hồ Chí Minh. Được xây dựng dựa trên các dữ liệu thời tiết lịch sử, hệ thống này sử dụng các thuật toán và mô hình học máy để phân tích và dự đoán thời tiết trong vòng 3 giờ tiếp theo tại thành phố Hồ Chí Minh. Dự án tập trung vào việc thu thập dữ liệu từ các cảm biến thời tiết, bao gồm thời gian, nhiệt độ, lượng mưa, độ ẩm, độ che phủ của mây, áp suất không khí, tốc độ gió, và thời tiết cụ thể tổng quan trên Thành phố Hồ Chí Minh. Các dữ liệu này được xử lý và phân tích để tạo ra mô hình dự báo thời tiết. Để đạt được độ chính xác cao, đồ án sử dụng một số phương pháp dự báo thời tiết như mô hình hồi quy Logistic, XGBClassifier và Mô hình SVC (Support Vector Classifier). Các thuật toán này sẽ phân tích và học từ các dữ liệu thời tiết lịch sử để tìm ra mẫu và xu hướng, từ đó dự đoán thời tiết trong 3 giờ tiếp theo. Kết quả dự báo sẽ được hiển thị dưới 2 loại nhãn là mưa, không mưa và thông báo trực tuyến cho người dùng thông qua ứng dụng di động hoặc trang web. Sử dụng các tiêu chuẩn đánh giá để chọn ra mô hình phù hợp nhất. Từ đó, nhóm tiếp tục hoàn thiện, phát triển bộ dữ liệu và mô hình cũng như các ứng dụng thực tiễn liên quan. Điều này giúp người dùng cập nhật thông tin thời tiết nhanh chóng và chuẩn xác, giúp họ đưa ra quyết định và chuẩn bị phù hợp cho các hoạt động trong vòng 3 giờ tiếp theo ở Thành phố Hồ Chí Minh.

1 Giới Thiệu

Dự báo thời tiết là vấn đề được quan tâm bởi vì kết quả dự báo sẽ tác động đến đời sống hàng ngày của mỗi chúng ta. Trong dự báo thời tiết, nhiều yếu tố biến đổi khó lường của thiên nhiên nên có độ phức tạp lớn do đó độ chính xác hạn chế và cần các phương pháp để giải quyết nó rất được chú trọng. Bài toán dự báo thời tiết là bài toán có một vai trò quan trọng trong đời sống kinh tế, xã hội. Dự báo đúng sẽ giúp con người đưa ra những quyết định đúng đắn. Trong đề tài này, để xây dựng một hệ thống dự báo chính xác và đáng tin cậy về tình hình thời tiết trong khoảng thời gian ngắn, cụ thể là trong 3 giờ tới, nhóm đã sử dụng một số mô hình để dự đoán như mô hình hồi quy Logistic, XGBClassifier và Mô hình SVC (Support Vector Classifier), nhóm sẽ nói rõ hơn về các mô hình này ở phần sau. Nhóm mong đợi độ chính xác của mô hình vượt 80

2 Dữ Liệu

2.1 Thu Thập Dữ Liệu

Để dự đoán tình hình thời tiết tại Thành phố Hồ Chí Minh, trong bài báo này, nhóm sử dụng dữ liệu được thu thập từ 01/01/2020 đến 30/04/2023, theo từng thời điểm cách nhau 3 giờ trong ngày ,số liệu này được lấy ở WorldWeatherOnline ngày tháng cụ thể của từng điểm dữ liệu thời tiết được ghi lại rõ ràng, theo thứ tự. Nhóm dùng thư viện selenium của python để thu thập dữ liệu. Source code trong file crawler.ipynb. Dữ liệu thu thập bao gồm: Hour, Temperature, Forecast, Rain, Rain-rate, Cloud-rate, Pressure, Wind, Gust, Day, Month, Year, Weather.

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

Table 1: Bảng mô tả các thuộc tính

2.2 Tiền xử Lí Dữ Liệu

Sau khi thu thập dữ liệu từ WorldWeatherOnline, 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. Chúng tôi chuẩn hóa giá trị của các cột *Temperature, Forecast, Rain-rate, Cloud-rate, Pressure, Wind, Gust* sang kiểu dữ liệu số (float) để phù hợp với các phân tích; chuẩn hóa giá trị của các cột *Rain-rate, Cloud-rate* 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.

- Rút trích đặc trưng: Ở bước này nhóm lấy giá trị của thuộc tính *weather* (lưu vào 1 list) chuyển giá trị về tập 0,1 (không mưa và mưa). Việc chuyển đổi này giúp cho dữ liệu thành đặc trưng phù hợp để mô hình có thể hiểu, học từ đó và dễ dàng cho việc phân tích

Figure 1 và *Figure 2* là ví dụ về bộ dữ liệu trước và sau khi được xử lý.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
0	00:00	26	°c	28	°c	0.0	mm	0%	10%	1015	mb	8	km/h	11	km/h	1	1	2020	Clear
1	03:00	25	°c	26	°c	0.0	mm	0%	19%	1014	mb	8	km/h	13	km/h	1	1	2020	Clear
2	06:00	24	°c	26	°c	0.0	mm	0%	10%	1015	mb	8	km/h	11	km/h	1	1	2020	Clear
3	09:00	29	°c	30	°c	0.0	mm	0%	5%	1016	mb	7	km/h	15	km/h	1	1	2020	Clear
4	12:00	34	°c	34	°c	0.0	mm	0%	8%	1015	mb	10	km/h	17	km/h	1	1	2020	Clear

Figure 1: Dữ liệu thu thập được

	Hour	Temperature	Forecast	Rain	Rain_rate	Cloud_rate	Pressure	Wind	Gust	Day	Month	Year	Weather
0	0	26.0	28.0	0.0	0.0	0.10	1015.0	8.0	11.0	1	1	2020	0
1	3	25.0	26.0	0.0	0.0	0.19	1014.0	8.0	13.0	1	1	2020	0
2	6	24.0	26.0	0.0	0.0	0.10	1015.0	8.0	11.0	1	1	2020	0
3	9	29.0	30.0	0.0	0.0	0.05	1016.0	7.0	15.0	1	1	2020	0
4	12	34.0	34.0	0.0	0.0	0.08	1015.0	10.0	17.0	1	1	2020	0

Figure 2: Dữ liệu sau khi xử lý

2.3 Phân tích DL

Thống kê trên tập huấn luyện (Train), tập kiểm thử (Test) và toàn bộ dữ liệu (All) nhóm thu thập được thể hiện trong Table 2. Bảng thống kê thể hiện tổng số điểm dữ liệu thời tiết, trung bình Nhiệt độ đo được, trung bình Nhiệt độ cảm nhận được, trung bình Lượng mưa, trung bình Độ ẩm, trung bình độ che phủ của mây, trung bình Áp suất, trung bình Tốc độ gió trung bình, trung bình Tốc độ gió cao nhất

	Train	Test	All
Số điểm dữ liệu thời tiết	13222	3306	16528
Trung bình Nhiệt độ đo được			27.827
Trung bình Nhiệt độ cảm nhận được			30.985
Trung bình Lượng mưa			0.22
Trung bình Độ che phủ của mây			0.5
Trung bình Áp suất			3.15
Trung bình Tốc độ gió trung bình			9.49
Trung bình Tốc độ gió giật			15.336

Table 2: Tổng quan bộ dữ liệu đã thu thập

Figure 3 thể hiện sự phân bố các thuộc tính của bộ dữ liệu, từ đó ta thấy được sự phức tạp của dữ liệu.

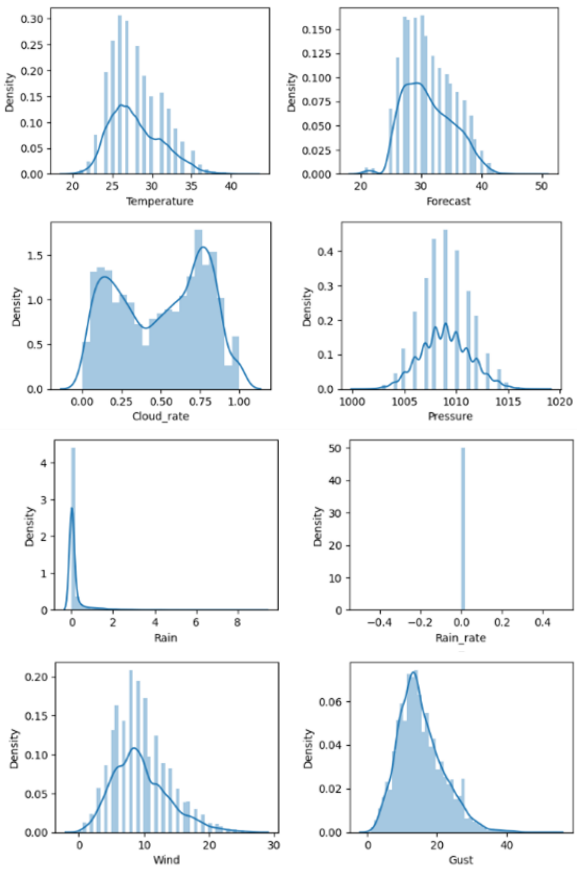

Figure 3: Biểu đồ phân bố các thuộc tính trong bộ dữ liệu

Sự phân bố thời tiết mưa hay không mưa được thể hiện qua biểu đồ tròn ở Figure 4.

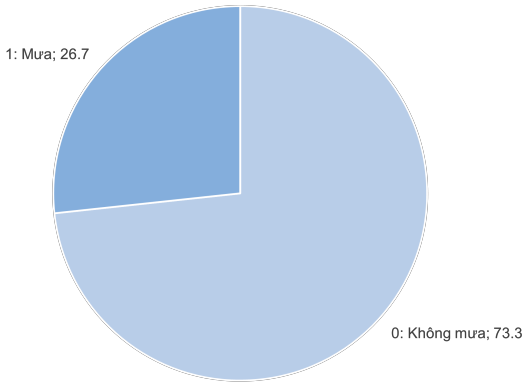

Figure 4: Biểu đồ phân bố tỉ lệ thời tiết trong bộ dữ liệu

Một số từ khoá của từng nhãn thời tiết được thể hiện trong *Table 3*.

Nhãn	Từ khóa
0 (No rain)	Clear, Cloudy, Overcast, Partly cloudy, Thundery outbreaks possible.
1 (Rain)	Heavy rain at times , Light rain shower, Moderate or heavy rain shower, Moderate rain at times, Patchy light rain with thunder, Patchy rain possible

Table 3: Một số từ khoá của từng nhãn thời tiết

3 Phương pháp máy học

Trước khi tiến hành huấn luyện, tập dữ liệu được chuẩn hóa bằng hàm StandardScaler() của thư viện sklearn giúp thuật toán máy học chạy nhanh hơn và chính xác hơn. Tăng dữ liệu với các nhãn có mưa, bằng cách lặp lại các điểm dữ liệu mưa để tăng độ cân bằng cho mô hình bằng việc sử dụng hàm RandomOverSampler() với các tham số sampling-strategy'='minority', random-state=22 trong thư viện imblearn. Sử dụng dữ liệu thu thập được, chúng tôi dự đoán thời tiết của thành phố Hồ Chí Minh bằng ba phương pháp: mô hình hồi quy Logistic, XGBClassifier và Mô hình SVC.

3.1 Mô hình Hồi quy Logistic

Một thuật toán rất nổi tiếng trong thống kê được sử dụng để dự đoán một số giá trị (Y) cho một tập hợp các tính năng (X). Thuật toán Hồi quy Logistic thuộc học máy có giám sát để phân loại dữ liệu. Mô hình hồi quy Logistic áp dụng cho biến phụ thuộc là biến định tính hoặc định lượng chỉ có hai giá trị (mưa hoặc không mưa) hay nhị phân là 0 hoặc 1. Điều này phù hợp với bài toán dự báo tình hình thời tiết cụ thể là phân tích thời tiết. Đầu ra của bài toán đó là xác định thời tiết trong 3 giờ tiếp theo mưa hay không mưa. Mô hình Logistic Regression sử dụng mô hình có sẵn trong thư viện sklearn và các giá trị tham số là mặc định.

Trong hồi qui tuyến tính đơn, biến độc lập x và phụ thuộc y là biến số liên tục liên hệ qua phương trình:

$$P(y = 1|x) = \frac{1}{1 + exp(-z)}$$

Trong đó:

- P(y=1|x) là xác suất biến phụ thuộc (y) có giá trị bằng 1 dựa trên các biến độc lập (x).
- exp(-z) là hàm mũ cơ số e với z là một hàm tuyến tính của các biến độc lập.

Hàm tuyến tính z thường được tính bằng cách kết hợp các biến độc lập với các trọng số tương ứng:

$$z = \beta_0 + \beta_1.x_1 + \beta_2.x_2 + ... + \beta_p.x_p$$

Trong đó:

- $$\beta_0 + \beta_1 + \beta_2 + ... + \beta_p$$

  là các tham số hồi quy (coefficients) cần được xác định thông qua quá trình huấn luyện mô hình.

- x, x, ..., x là các biến độc lập được sử dụng để dự đoán xác suất.

3.2 XGBClassifier

Mô hình XGBClassifier là một ứng dụng của thuật toán Gradient Boosting Decision Trees (GBDT). Mô hình XGBClassifier là sự kết hợp của các cây quyết định và kỹ thuật tăng cường gradient. GBDT là một phương pháp ensemble learning, trong đó một loạt các cây quyết định (decision trees) được xây dựng tuần tự và mỗi cây được điều chỉnh dựa trên các lỗi của cây trước đó. Quá trình này tạo ra một mô hình mạnh mẽ bằng cách kết hợp các cây yếu thành một mô hình tổng hợp. XGBClassifier: sử dụng mô hình có sẵn trong thư viện xgboost và các giá trị tham số là mặc định. XGBClassifier có một số tham số quan trọng để điều chỉnh quá trình huấn luyện và tránh overfitting, bao gồm learning rate (tỷ lệ học), số lượng cây, độ sâu của cây và hệ số điều chuẩn.

Mô hình XGBClassifier cung cấp một phương pháp mạnh mẽ để xây dựng mô hình phân loại dựa

trên cây quyết định và kỹ thuật tăng cường gradient, và thường mang lại hiệu suất tốt trong nhiều bài toán phân loại. Vì vậy chúng tôi nghĩ nó sẽ hoạt động tốt trong việc phân loại thời tiết trong bài toán này.

3.3 Mô hình SVC (Support Vector Classifier)

Mô hình Support Vector Classifier (SVC) là một bộ phân loại dựa trên Support Vector Machines (SVM) trong thư viện scikit-learn để tạo ra mô hình phân loại tối ưu giữa các lớp dữ liệu. Nó được sử dụng cho bài toán phân loại, trong đó mục tiêu là dự đoán và phân loại các điểm dữ liệu vào các lớp khác nhau. Mô hình SVC có nhiều ưu điểm, bao gồm khả năng tách biệt các lớp dữ liệu phức tạp, khả năng xử lý các tập dữ liệu lớn, và tính tổng quát cao. Tuy nhiên, nó cũng có một số hạn chế, như đòi hỏi tính toán phức tạp và nhạy cảm với các nhiễu dữ liệu. Mô hình SVC (Support Vector Classifier): sử dụng mô hình có sẵn trong thư viện sklearn và các giá trị tham số là mặc định.

3.4 Độ đo

Nhóm sử dụng độ đo là độ chính xác 'Compute Area Under the Receiver Operating Characteristic Curve' và các độ đo khác trong ma trận nhầm lẫn confusion matrix. AUC-ROC là một độ đo phổ biến trong bài toán phân loại, đo lường khả năng của mô hình phân loại để phân biệt giữa các lớp dữ liệu. Đường cong ROC biểu thị mức độ phân biệt của mô hình, và AUC-ROC tính diện tích dưới đường cong ROC. Một AUC-ROC gần 1.0 cho thấy mô hình có khả năng phân loại tốt, trong khi AUC-ROC gần 0.5 cho thấy mô hình không có khả năng phân loại tốt hơn ngẫu nhiên.

Confusion matrix là một bảng tổng hợp kết quả phân loại của mô hình. Nó phân chia các dự đoán thành bốn phần: true positive (TP), true negative (TN), false positive (FP) và false negative (FN). Các độ đo chính trong confusion matrix bao gồm:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 - score = \frac{2 * (Precision * Recall)}{Precision + Recall} \quad (4)$$

4 Phân tích mô hình và nhận định về kết quả

Với label tại thời điểm hiện tại (không chỉnh sửa)

	Logistic Regression	XGBClassifier	SVC
Accuracy	0.99990	1.0	0.99951
Precision	0.99923	0.99944	0.98871
Recall	0.99923	0.99979	0.98661
F1-score	0.99923	0.99979	0.98766
Time	0.032	0.13	18.9

Table 4: Đánh giá các phương pháp máy học tại thời điểm hiện tại

Với label là 3 tiếng sau (để có thể dự đoán)

	Logistic Regression	XGBClassifier	SVC
Accuracy	0.83441	0.95152	0.88126
Precision	0.70383	0.85749	0.75523
Recall	0.74960	0.88262	0.79508
F1-score	0.70976	0.86853	0.76789
Time	0.021	0.273	42.912

Table 5: Đánh giá các phương pháp máy học trong 3 giờ tiếp theo

Với label là 3 tiếng sau và bỏ các thuộc tính "day, month, year"

	Logistic Regression	XGBClassifier	SVC
Accuracy	0.82908	0.93362	0.86657
Precision	0.70115	0.81027	0.74480
Recall	0.74772	0.84556	0.78348
F1-score	0.70589	0.82399	0.75682
Time	0.019	0.254	41.173

Table 6: Đánh giá các phương pháp máy học trong 3 giờ tiếp theo và loại bỏ các thuộc tính "day, month, year"

Với label là 3 tiếng sau và bỏ thuộc tính "day, year", còn "month" sẽ được chia theo mùa nắng (1,2,3,4,12) và mưa (5,6,7,8,9,10,11) ở Việt Nam

	Logistic Regression	XGBClassifier	SVC
Accuracy	0.83692	0.93534	0.87293
Precision	0.70716	0.81339	0.74988
Recall	0.75551	0.84920	0.79657
F1-score	0.71206	0.82731	0.76207
Time	0.018	0.280	40.370

Table 7: Đánh giá các phương pháp máy học trong 3 giờ tiếp theo và thời gian được chia theo 2 mùa mưa và nắng

5 Phân tích lỗi, hướng phát triển

Cách 1: với độ chính xác tuyệt đối ở cả 3 mô hình cho thấy tính hoàn thiện của bộ dữ liệu. Cách 2: việc thay đổi label thành 3 tiếng sau gây thiếu tính đúng đắng của bộ dữ liệu, dẫn đến việc độ chính xác của cả 3 mô hình bị giảm. Cách 3: việc giảm

thuộc tính của điểm dữ liệu làm cho bộ dữ liệu thiếu mất tính đầy đủ và thiếu tính đúng đắng của cách 2 làm cho độ chính xác của mô hình giảm mạnh. Cách 4: kết quả tốt hơn 1 chút so với cách 3 thể hiện hướng đi chính xác đối với bộ dữ liệu, phản ánh đúng thời tiết thực tế ở Việt Nam.

Về nhận xét chung: mô hình XGBClassifier là độ chính xác cao nhất và thời gian thực thi nhanh, phù hợp cho mô hình dự báo thời tiết. Logictis regresstion tuy thời gian tốt nhưng không đảm bảo độ chính xác cao như XGBClassifier. Đối với SVC thì lại kém ở mọi mặt so với 2 mô hình trước, thời gian thực thi lâu gấp nhiều lần.

6 Kết luận

Trong bài báo này, nhóm đã thử nghiệm bài toán trên nhiều mô hình khác nhau như: mô hình Logistic Regression, XGBClassifier và Mô hình SVC (Support Vector Classifier). Nhìn chung, các mô hình đều cho kết quả khá khả quan với các độ đo trên. Sau khi so sánh, chúng tôi nhận thấy mô hình học sâu đưa ra kết quả khá tốt trên bộ dữ liệu với độ chính xác từ 80-95 đạt được mục tiêu ban đầu đề ra. Tuy nhiên, mô hình vẫn còn một số điểm cần cải thiện để hiệu suất nâng cao hơn nữa để có thể áp dụng vào thực tế. Về khía cạnh bộ dữ liệu, chúng tôi cần thêm một số thuộc tính có liên quan mật thiết đến thời tiết để kết quả dự đoán có kết quả chính xác cao hơn.

7 Bảng phân công

Bảng phân công công việc trong nhóm được thể hiện trong *Figure 5*

Công việc	Thời gian	Phân công (%)	
		Huy	Trang
Thu thập dữ liệu	10/05 - 20/05	70	30
Tiền xử lý dữ liệu	20/05 - 01/06	40	60
Xây dựng mô hình	01/06 - 08/06	70	30
Đánh giá mô hình	09/06 - 11/6	50	50
Phân tích và hướng phát triển	12/6 - 16/6	50	50
Làm slide thuyết trình	17/6 - 20/6	40	60
Viết Báo cáo	15/6 - 24/6	40	60

Figure 5: Bảng phân công công việc trong nhóm