HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

KHOA SAU ĐẠI HỌC

TIỂU LUẬN KHAI PHÁ DỮ LIỆU

Đề tài: NGHIÊN CỨU VỀ DỰ BÁO NĂNG LƯỢNG MẶT TRỜI DỰA TRÊN MÔ HÌNH HỌC MÁY

Giảng viên : TS. Đặng Hoàng Long

Lớp : D24CQKH01-B

Nhóm : 13

Thành viên : Lương Đức Thuận

Đinh Thị Hương Thảo

Hồ Thức Huy

Đỗ Minh Tuấn

Hà Nội – 2024

MỤC LỤC

NỘI DUNG

THUẬT NGỮ VIẾT TẮT

Thuật ngữ	Viết đầy đủ	Giải thích nghĩa
IoT	Internet of Things	Internet kết nối vạn vật
AI	Artificial Intelligence	Trí tuệ nhân tạo
DC	Direct Current	Dòng điện một chiều
AC	Alternating Current	Dòng điện xoay chiều
Inverter	Inverter	Bộ nghịch lưu
ARMA	AutoRegressive Moving Average	Mô hình dự báo ARMA
ARIMA	AutoRegressive Integrated Moving Average	Mô hình dự báo ARIMA
RNN	Recurrent Neural Network	Mô hình dự báo RNN
LSTM	Long Short Term Memory	Mô hình dự báo LSTM
GRU	Gate Recurrent Unit Network	Mô hình GRU
TCN	Temporal Convolution Network	Mô hình dự báo TCN
Trend	Trend	Xu hướng
Seasonality	Seasonality	Mùa vụ
Cycles	Cycles	Chu kì
Stationality	Stationality	Tính dừng
Univariate	Univariate	Đơn biến
Multivariate	Multivariate	Đa biến
Receptive Field	Receptive Field	Trường tiếp nhận
Kernel	Kernel	Bộ lọc
1D Convolution	1D Convolution	Tích chập một chiều
Causal Convolution	Causal Convolution	Tích chập một chiều
Dilated Convolution	Dilated Convolution	Tích chập giãn nở
Forget gate	Forget gate	Cổng quên
Dilation rate	Dilation rate	Độ giãn nở
Window_size	Window_size	Độ dài dữ liệu đầu vào
MSE	Mean Square Error	Hàm lỗi trung bình bình phương

MAE	Mean Absolute Error	Hàm lỗi trung bình tuyệt đối
Coefficient of Determination	Coefficient of Determination	Hệ số xác định

LỜI MỞ ĐẦU

Trong bối cảnh nhu cầu năng lượng toàn cầu ngày càng gia tăng và các nguồn tài nguyên hóa thạch dần cạn kiệt, việc tìm kiếm và phát triển các nguồn năng lượng tái tạo đã trở thành một ưu tiên chiến lược của nhiều quốc gia. Trong số các nguồn năng lượng tái tạo hiện nay, năng lượng mặt trời nổi bật như một giải pháp thân thiện với môi trường, có tiềm năng lớn và phù hợp với xu hướng phát triển bền vững. Tuy nhiên, tính chất không ổn định và phụ thuộc nhiều vào các yếu tố tự nhiên của năng lượng mặt trời đã đặt ra nhiều thách thức lớn trong việc khai thác và tối ưu hóa nguồn năng lượng này. Một trong những vấn đề quan trọng cần giải quyết là làm thế nào để dự đoán chính xác lượng năng lượng mặt trời trong tương lai nhằm hỗ trợ quá trình lập kế hoạch, phân phối và sử dụng hiệu quả. Việc dự báo năng lượng mặt trời không chỉ giúp giảm thiểu lãng phí tài nguyên mà còn tăng cường khả năng hòa lưới điện và đảm bảo tính ổn định của hệ thống điện. Đây là bài toán không đơn giản, bởi dữ liệu liên quan đến năng lượng mặt trời thường mang tính phi tuyến, không đồng nhất và bị ảnh hưởng bởi nhiều yếu tố phức tạp như điều kiện thời tiết, vị trí địa lý và thời gian trong ngày. Dự báo chuỗi thời gian là một lĩnh vực nghiên cứu có vai trò then chốt trong việc giải quyết các bài toán như trên. Với sự phát triển mạnh mẽ của công nghệ khoa học dữ liệu và trí tuệ nhân tạo trong những năm gần đây, các phương pháp hiện đại như học sâu (Deep Learning) đã chứng minh hiệu quả vượt trội trong việc dự báo chuỗi thời gian. Do vậy, trong tiểu luận lần này nhóm em lựa chọn nghiên cứu một kiến trúc mô hình cho dự báo năng lượng mặt trời dựa trên bài toán dự báo chuỗi thời gian là sự kết hợp của hai mô hình Temporal Convolutional Network và Gate Recurrent Unit Network. Mục đích của nghiên cứu là để đánh giá khả năng dự báo của sự kết hợp giữa hai mô hình này. Bố cục của tiểu luận bao gồm 3 chương:

Chương 1: Giới thiệu chung

Chương 2: Bài toán dự báo chuỗi thời gian và các giải pháp

Chương 3: Áp dụng và triển khai mô hình dựa trên kĩ thuật học máy

Bằng sự cố gắng và nỗ lực nhóm em đã hoàn thành xong tiểu luận. có sự hạn chế về mặt thời gian và mức độ hiểu biết của nhóm nên không thể tránh khỏi những thiếu sót trong quá trình nghiên cứu. Vì thế, nhóm em rất mong nhận được những lời góp ý và xi sự chỉ bảo thêm của thầy và các bạn để nhóm em có thêm những kiến thức phục vụ cho học tập cũng như công việc sau này.

CHƯƠNG 1: GIỚI THIỆU CHUNG

1.1. Tổng quan về năng lượng mặt trời và hệ thống pin năng lượng mặt trời

Năng lượng là một khái niệm trong vật lý, thể hiện khả năng thực hiện công, tạo ra biến đổi hoặc duy trì hoạt động trong tự nhiên. Năng lượng tồn tại dưới nhiều dạng khác nhau như thế năng, động năng, nhiệt năng, điện năng, hóa năng và năng lượng hạt nhân. Một trong những đặc điểm quan trọng nhất của năng lượng là nó không thể tự nhiên sinh ra hay mất đi mà chỉ có thể chuyển hóa từ dạng này sang dạng khác. Quy luật này được thể hiện rõ qua định luật bảo toàn năng lượng. Ví dụ, nhiệt năng từ than đốt có thể được chuyển hóa thành điện năng trong các nhà máy nhiệt điện hoặc năng lượng mặt trời được chuyển thành điện năng thông qua các tấm mặt trời.

Các nguồn năng lượng tự nhiên bao gồm năng lượng hóa thạch như dầu mỏ, khí đốt tự nhiên, than đá và các nguồn năng lượng sạch như gió, nước, bức xạ mặt trời và các nguyên tố phóng xạ. Trong nhiều thập kỷ qua, con người đã dựa chủ yếu vào các nguồn năng lượng hóa thạch để phục vụ nhu cầu phát triển kinh tế và xã hội. Tuy nhiên, các nguồn năng lượng này có hạn và không thể tái tạo. Việc khai thác quá mức và thiếu bền vững đã dẫn đến tình trạng cạn kiệt tài nguyên thiên nhiên. Hơn nữa, quá trình sử dụng nhiên liệu hóa thạch phát sinh một lượng lớn khí CO2 và các chất ô nhiễm khác vào bầu khí quyển, làm gia tăng hiệu ứng nhà kính và là nguyên nhân chính gây biến đổi khí hậu, nóng lên toàn cầu cũng như suy thoái môi trường tự nhiên.

Bên cạnh đó, năng lượng hạt nhân đã được chứng minh là có thể tạo ra nguồn năng lượng khổng lồ với hiệu suất cao. Tuy nhiên, công nghệ này đi kèm với nhiều thách thức lớn như nguy cơ rò rỉ phóng xạ trong các sự cố thiên tai, thảm họa kỹ thuật, cũng như vấn đề xử lý chất thải phóng xạ an toàn và hiệu quả. Những hệ lụy từ các sự cố hạt nhân trong lịch sử đã cho thấy đây không phải là giải pháp hoàn toàn tối ưu nếu không được kiểm soát chặt chẽ và an toàn.

Trong bối cảnh đó, việc chuyển đổi từ năng lượng truyền thống sang các nguồn năng lượng tái tạo bền vững là giải pháp cấp bách và cần thiết hơn bao giờ hết. Năng lượng tái tạo bao gồm năng lượng gió, năng lượng mặt trời… Những nguồn năng lượng này không chỉ dồi dào, có thể tái tạo vô hạn mà còn thân thiện với môi trường, hạn chế phát thải khí nhà kính và ô nhiễm không khí. Đặc biệt, năng lượng mặt trời và năng lượng gió đang ngày càng được đầu tư phát triển nhờ sự tiến bộ của công nghệ, giúp giảm đáng kể chi phí sản xuất và tăng hiệu quả khai thác.

Hiện nay, biến đổi khí hậu đang trở thành mối đe dọa toàn cầu với những hiện tượng thời tiết cực đoan như bão lũ, hạn hán và mực nước biển dâng cao. Việc mở rộng sử dụng các nguồn năng lượng tái tạo sẽ đóng vai trò quan trọng trong việc giảm thiểu phát thải khí CO2, góp phần hạn chế sự gia tăng nhiệt độ Trái Đất và bảo vệ môi trường sống. Để đạt được mục tiêu này, các quốc gia trên thế giới cần chung tay đẩy mạnh nghiên cứu, đầu tư phát triển công nghệ năng lượng sạch, đồng thời xây dựng các chính sách

khuyến khích doanh nghiệp và cộng đồng tham gia vào quá trình chuyển đổi năng lượng bền vững.

1.1.1. Năng lượng mặt trời và tiềm năng phát triển

1.1.2. Tác động của năng lượng mặt trời

1.1.3. Xu hướng phát triển năng lượng mặt trời

1.1.4. Hệ thống pin năng lượng mặt trời

1.2. Vai trò của dự báo năng lượng mặt trời

1.2.1. Tối ưu hóa việc sử dụng năng lượng tái tạo trong hệ thống điện lưới

1.2.2. Giảm phụ thuộc vào năng lượng không tái tạo

1.2.3. Hiệu quả trong phát triển hệ thống lưu trữ năng lượng

1.2.4. Hỗ trợ hoạch định chính sách năng lượng

1.3. Dự báo năng lượng mặt trời dựa trên bài toán dự báo chuỗi thời gian

1.3.1. Nhiệm vụ và thách thức của bài toán dự báo chuỗi thời gian

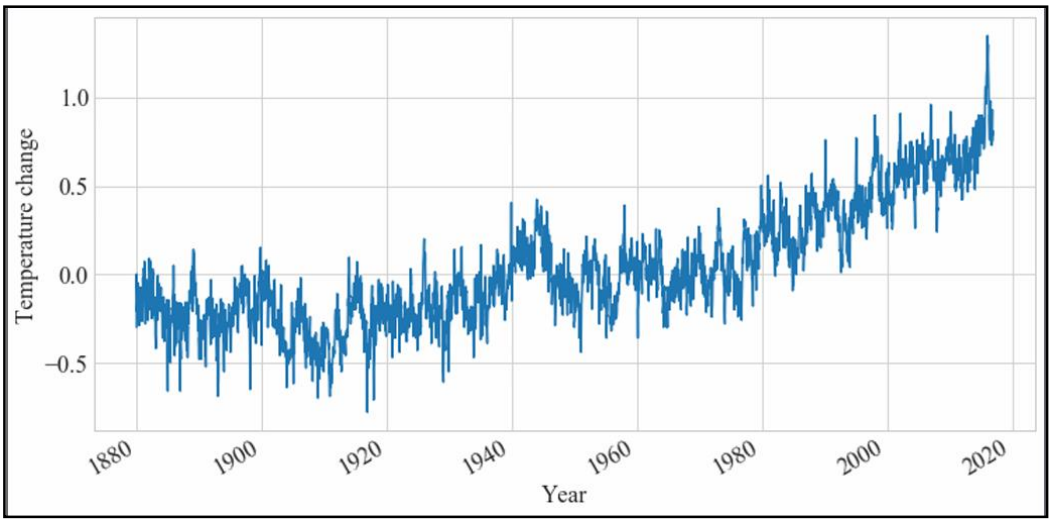
1.3.2. Vai trò của dữ liệu trong dự báo năng lượng mặt trời

CHƯƠNG 2: BÀI TOÁN DỰ BÁO CHUỖI THỜI GIAN VÀ CÁC GIẢI PHÁP

2.1. Các đặc điểm cơ bản của bài toán dự báo chuỗi thời gian

Trong bài toán dự báo chuỗi thời gian, đặc điểm cơ bản nổi bật nhất đó chính là đặc điểm về mặt dữ liệu của bài toán. Ngoài ra, một đặc điểm cơ bản nữa của bài toán dự báo chuỗi thời gian còn được thể hiện về độ dài của thời gian dự báo. Trong phần này trình bày những đặc điểm cơ bản nhất trong bài toán dự báo chuỗi thời gian.

2.1.1. Tính tuần tự và sự phụ thuộc thời gian của dữ liệu



Hình 2.1. Dữ liệu nhiệt độ từ năm 1880 đến năm 2019

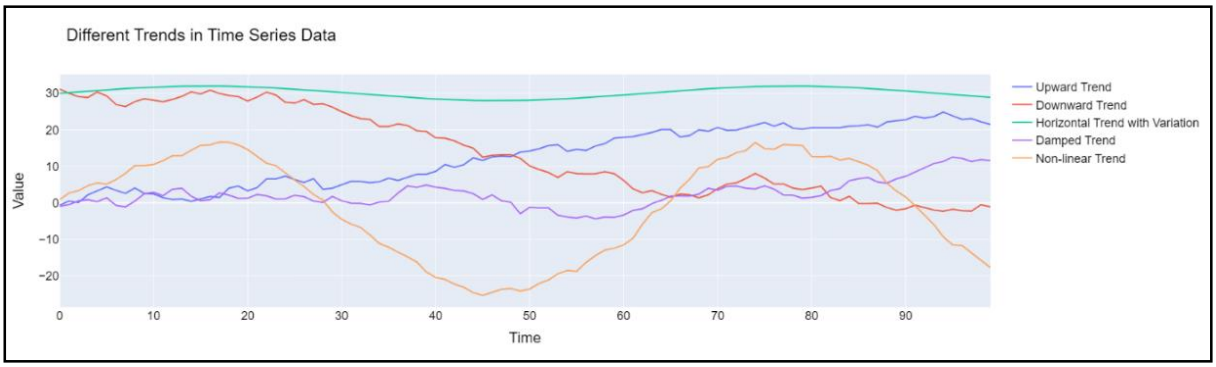
Trong bài toán dự báo chuỗi thời gian, các giá trị dữ liệu được thu thập liên tục theo thời gian và có sự phụ thuộc lẫn nhau. Điều này có nghĩa là giá trị tại một thời điểm cụ thể thường chịu ảnh hưởng bởi các giá trị ở các thời điểm trước đó trong chuỗi dữ liệu. Mối quan hệ này đóng vai trò quan trọng trong việc dự đoán các giá trị tương lai do có thể quan sát được những đặc điểm của dữ liệu theo thời gian chẳng hạn như: xu hướng, mùa vụ, chu kì của dữ liệu.

Ví dụ trong dự báo thời tiết nhiệt độ ngày hôm nay thường có liên quan đến nhiệt độ của những ngày trước đó. Tương tự trong dự báo giá cổ phiếu, giá cổ phiếu hiện tại có thể phụ thuộc vào xu hướng giá đóng cửa trong các ngày trước. Sự phụ thuộc này giúp xây dựng các mô hình dự báo hiệu quả, tận dụng thông tin lịch sử để dự đoán các giá trị trong tương lai.

2.1.2. Tính biến động của dữ liệu chuỗi thời gian

a. Xu hướng của dữ liệu chuỗi thời gian

Xu hướng (Trend) trong dữ liệu chuỗi thời gian đề cập đến hướng di chuyển của dữ liệu qua thời gian, có thể là tăng lên hoặc giảm đi. *Hình 2.2* mô phỏng về xu hướng của dữ liệu theo thời gian. Dựa vào *Hình 2.1* có thể thấy xu hướng của nhiệt độ tăng dần theo thời gian từ năm 1880 đến năm 2019.



Hình 2.2. Mô phỏng xu hướng của dữ liệu chuỗi thời gian

Một số xu hướng phổ biến của dữ liệu chuỗi thời gian:

- Upward Trend: Dữ liệu có xu hướng tăng theo thời gian.
- Downward Trend: Dữ liệu có xu hướng giảm theo thời gian.
- Horizontal Trend: Dữ liệu duy trì ổn định hoặc biến đổi rất ít theo thời gian.

b. Mùa vụ của dữ liệu chuỗi thời gian

Tính mùa vụ (Sesonality) trong dữ liệu chuỗi thời gian đề cập đến các biến động có tính chất lặp lại đều đặn trong một khoảng thời gian nhất định thường có tính ngắn hạn chẳng hạn như hàng ngày, hàng tuần, hàng tháng hoặc hàng năm và có thể dự đoán được. Việc nhận diện và phân tích tính mùa vụ đóng vai trò quan trọng trong bài toán dự báo chuỗi thời gian. Một số tính mùa vụ phổ biến chẳng hạn như: Holiday Sesonality, Weekly Seasonality, Monthly Seasonality, Annuality Seasonality.

Holiday Seasonality: Thể hiện sự thay đổi thường được gây ra bởi các sự kiện đặc biệt như ngày lễ, sự kiện đặc biệt nào đó.

Weekly Seasonality: Thể hiện sự thay đổi lặp lại trong khoảng thời gian 7 ngày.

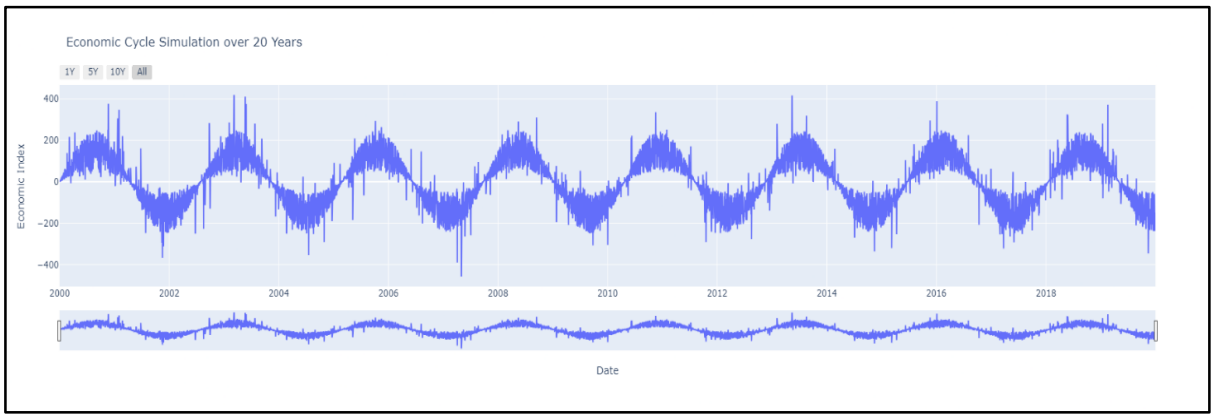
Monthly Seasonality: Thể hiện sự thay đổi lặp lại theo từng tháng trong một năm. Ví dụ, hóa đơn tiền điện thường tăng cao trong các tháng mùa hè khi nhu cầu sử dụng điều hòa tăng, hoặc doanh thu bán lẻ tăng vào tháng cuối năm do các sự kiện mua sắm cuối năm.

Annual Seasonality: Thể hiện tính mùa vụ xảy ra theo từng.

c. Chu kì trong dữ liệu chuỗi thời gian

Chu kì (Cycles) trong dữ liệu chuỗi thời gian là những biến đổi lặp lại theo một khoảng thời gian dài, thường không có tính cố định. Chu kỳ trong dữ liệu chuỗi thời gian phản ánh sự thay đổi liên tục qua các giai đoạn. Ví dụ trong lĩnh vực kinh tế, một chu kì bao gồm tăng trưởng, suy thoái, khủng hoảng và phục hồi. Thời gian của một chu

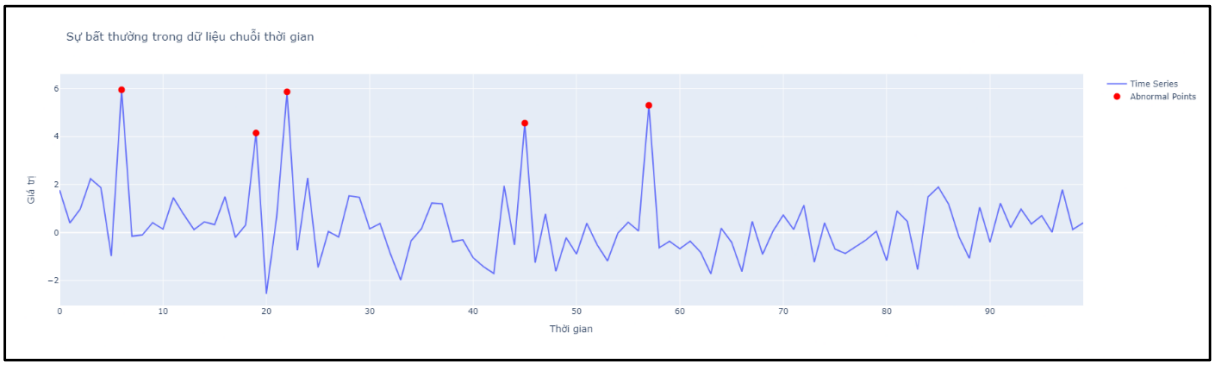
kì không cố định mà có thể thay đổi tùy thuộc vào các yếu tố ngoại cảnh tác động. Khác với tính mùa vụ thường xảy ra trong một khoảng thời gian cố định, chu kì trong dữ liệu chuỗi thời gian không đều đặn và phức tạp hơn.



Hình 2.7. Mô phỏng chu kì của dữ liệu

d. Nhiễu trong dữ liệu chuỗi thời gian

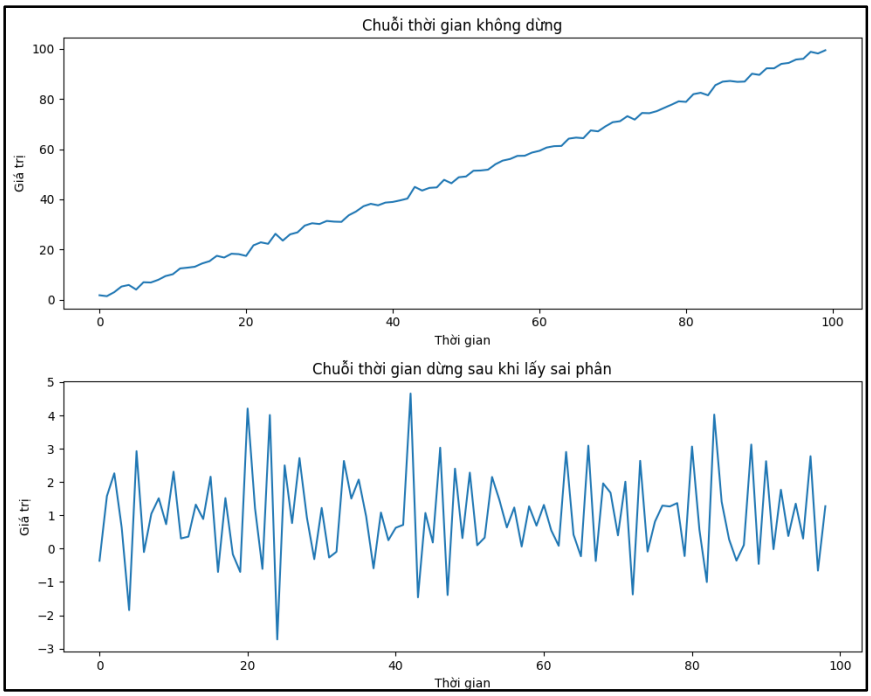
Sự bất thường trong dữ liệu hay còn gọi là nhiễu trong dữ liệu chuỗi thời gian đề cập đến những thay đổi đột ngột, khác thường so với các xu hướng hoặc quy luật đã xuất hiện trước đó trong dữ liệu. Những thay đổi này thường xảy ra một cách ngẫu nhiên, khó giải thích rõ ràng và không thể dự đoán trước.



Hình 2.8. Mô phỏng nhiễu (bất thường) trong dữ liệu chuỗi thời gian

2.1.3. Tính dừng của dữ liệu chuỗi thời gian

Tính dừng (stationarity) là một thuộc tính quan trọng trong phân tích và dự báo chuỗi thời gian. Một chuỗi thời gian được gọi là dừng khi các đặc điểm thống kê của nó, chẳng hạn như kỳ vọng, phương sai và tự tương quan, không thay đổi theo thời gian. Điều này có nghĩa là chuỗi không có xu hướng (trend) hay mùa vụ (seasonality) biến đổi theo thời gian.



Hình 2.9. Mô phỏng tính dừng của dữ liệu chuỗi thời gian

Trong thực tế, chuỗi thời gian thường là chuỗi thời gian có tính không dừng do ảnh hưởng của xu hướng, mùa vụ hoặc biến động lớn trong dữ liệu.

2.1.4. Phân loại dữ liệu chuỗi thời gian

Dữ liệu chuỗi thời gian là tập hợp các quan sát được ghi nhận theo một trật tự thời gian, thường ở các khoảng thời gian cách đều nhau. Trong phân tích chuỗi thời gian, việc xác định loại dữ liệu là đơn biến (univariate) hay đa biến (multivariate) đóng vai trò quan trọng trong việc lựa chọn phương pháp phân tích và xây dựng mô hình dự đoán.

a. Dữ liệu chuỗi thời gian đơn biến

Dữ liệu chuỗi thời gian đơn biến (Univariate) chỉ chứa một đặc trưng được quan sát theo thời gian. Mục tiêu chính thường là phân tích hoặc dự đoán giá trị trong giá trị trong tương lại dựa trên các giá trị quá khứ của chính nó chẳng hạn như dự báo nhiệt độ của một ngày dựa trên dữ liệu nhiệt độ của các ngày trước đó hay dự báo giá cổ phiếu dựa trên giá đóng của của ngày trước đó.

b. Dữ liệu chuỗi thời gian đa biến

Dữ liệu chuỗi thời gian đa biến (Multivariate) chứa nhiều đặc trưng được quan sát đồng thời theo thời gian. Các đặc trưng này có thể có mối quan hệ với nhau và việc khai thác mối quan hệ này phản ánh những mối quan hệ giữa các đặc điểm của các đặc trưng dữ liệu trên thực tế. Việc khai thác mối quan hệ này giúp mô hình dự báo khách quan hơn, chính xác hơn chẳng hạn như dự báo năng lượng pin mặt trời dựa trên các đặc trưng thời tiết như nhiệt độ, độ ẩm, độ che phủ mây …

2.1.5. Các kiểu dự báo trong bài toán chuỗi thời gian

Trong bài toán dự báo chuỗi thời gian, các vấn đề dự báo thường được quan tâm là dự báo ngắn hạn, dự báo trung hạn và dự báo dài hạn.

Dự báo ngắn hạn thường liên quan đến việc dự báo các giá trị trong tương lai gần, thường là trong khoảng thời gian từ vài ngày đến vài tuần. Mục tiêu của dự báo ngắn hạn là cung cấp thông tin chi tiết và chính xác về các thay đổi ngắn hạn trong chuỗi thời gian, chẳng hạn như sự thay đổi đột ngột của các yếu tố chẳng hạn như nhiệt độ, giá trị cổ phiếu, hay lưu lượng giao thông trong tương lai gần. Các mô hình dự báo ngắn hạn chủ yếu tập trung vào việc nắm bắt các xu hướng ngắn hạn và sự biến động nhanh chóng trong dữ liệu.

Dự báo trung hạn thường dự báo các giá trị trong tương lai trong khoảng thời gian từ vài tuần đến vài tháng. Mục tiêu của dự báo trung hạn là nắm bắt các xu hướng tổng thể hoặc sự biến động theo chu kỳ của chuỗi thời gian. Ví dụ, trong dự báo nhu cầu năng lượng, dự báo trung hạn có thể giúp dự đoán mức tiêu thụ điện trong vài tuần tới dựa trên các yếu tố như mùa vụ hoặc sự thay đổi của nhu cầu.

Dự báo dài hạn tập trung vào việc dự báo các giá trị trong tương lai xa, thường kéo dài từ vài tháng đến vài năm. Mục tiêu chính của dự báo dài hạn là nhận diện và nắm bắt các xu hướng lớn, cũng như các yếu tố tác động có thể thay đổi hoặc ảnh hưởng trong một khoảng thời gian dài. Các mô hình dự báo dài hạn thường được sử dụng để dự đoán những biến động hoặc sự phát triển quan trọng trong các lĩnh vực như kinh tế, thị trường tài chính, hay các yếu tố môi trường. Các mô hình này cần phải xử lý và phân tích các dữ liệu dài hạn, giúp đưa ra các quyết định chiến lược và chuẩn bị cho các thay đổi quan trọng trong tương lai.

2.2. Các phương pháp dự báo trong bài toán dự báo chuỗi thời gian

Để giải quyết vấn đề dự báo trong chuỗi thời gian, đã có nhiều phương pháp được xây dựng nhằm phân tích, dự báo các giá trị trong tương lai. Về cơ bản, các phương pháp được chia thành 2 nhóm là các phương pháp truyền thống và các phương pháp hiện đại dựa trên kĩ thuật học máy. Ngoài ra, các phương pháp kết hợp giữa 2 phương pháp cơ bản trên chẳng hạn như sự kết hợp giữa các mô hình cổ điển và các mô hình học máy hoặc sự kết hợp giữa các mô hình học máy nhằm tăng cường khả năng trích xuất thông tin và dự báo trong chuỗi thời gian. Trong phần này trình bày khảo sát một số những phương pháp vừa nêu.

2.2.1. Phương pháp dự báo dựa trên các mô hình cổ điển

a. Dự báo cho dữ liệu chuỗi thời gian có tính dừng

Mô hình ARMA là mô hình phổ biến nhất để dự báo chuỗi dữ liệu thời gian có tính dừng. Mô hình ARMA giả định dữ liệu đầu vào đã có tính dừng. Mô hình ARMA là mô

hình kết hợp của 2 thành phần cơ bản: AR (AutoRegressive – Tự hồi quy) và MA (Moving Average – Trung bình trượt).

AR (AutoRegressive) sử dụng mối quan hệ giữa một quan sát và một số quan sát trước đó để dự đoán các giá trị trong tương lai chẳng hạn như dự báo nhiệt độ cho ngày mai bằng cách sử dụng dữ liệu từ vài ngày trước.

Công thức tổng quát của AR(p): $X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \epsilon_t$ (2.1)

Trong đó : X_t là giá trị hiện tại của chuỗi thời gian

c là hằng số

ϕ_i là tham số của mô hình AR

p là bậc của mô hình AR

ϵ_t là nhiễu trắng

MA (Moving Average) sử dụng nhiễu trắng trong quá khứ và hiện tại để dự đoán giá trị hiện tại. Ví dụ như MA giả định là nhiệt độ ngày hôm nay cũng bị ảnh hưởng bởi nhiễu được tạo ra khi dự đoán nhiệt độ của những ngày trước.

Công thức tổng quát của MA(q): $X_t = c + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}$ (2.2)

Trong đó : X_t là giá trị hiện tại của chuỗi thời gian

c là hằng số

ϕ_i là tham số của mô hình MA

q là bậc của mô hình MA

ϵ_t là nhiễu trắng

Mô hình ARMA kết hợp cả hai thành phần AR và MA để dự đoán giá trị của chuỗi thời gian. Công thức tổng quát của mô hình ARMA (p, q) là:

$$X_t = c + \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} \quad (2.3)$$

Trong đó : X_t là giá trị hiện tại của chuỗi thời gian

c là hằng số

ϕ_i là tham số của mô hình MA

p là bậc của mô hình MA

θ_i là tham số của mô hình MA

q là bậc của mô hình MA

ϵ_t là nhiễu trắng

Mô hình ARMA là mô hình đơn giản, dễ triển khai và đã được ứng dụng rộng rãi.

b. Dự báo cho chuỗi dữ liệu có tính không dừng

Mô hình ARIMA là mô hình được áp dụng cho dự báo chuỗi thời gian đối với đặc điểm dữ liệu chuỗi thời gian có tính không dừng. Dữ liệu không dừng chứa các đặc điểm như xu hướng của dữ liệu hoặc mùa vụ của dữ liệu. Mô hình ARIMA bao gồm 3 thành phần: AR (Auto Regression), MA (Moving Average), I (Intergrated).

Auto Regression: Đây là thành phần tự hồi quy bao gồm tập hợp các độ trễ của biến hiện tại. Độ trễ bậc p chính là giá trị lùi về quá khứ p bước thời gian của chuỗi. Độ trễ dài hoặc ngắn trong quá trình AR phụ thuộc vào tham số trễ p. Cụ thể, quá trình AR(p) của chuỗi x(t) được biểu diễn như sau:

$$AR(p) = \phi_0 + \phi_1 x_{t-1} + \phi_2 x_{t-2} + ... + \phi_p x_{t-p} \quad (2.4)$$

Moving Average: Thành phần này hồi quy trên các sai số (residuals) của mô hình tại các thời điểm trước đó. Nó giúp giảm ảnh hưởng của các nhiễu ngẫu nhiên trong dữ liệu. Độ trễ bậc q là số lượng giá trị sai số trong quá khứ được sử dụng. Quá trình trung bình trượt MA(q) được biểu diễn như sau:

$$MA(q) = \mu + \epsilon(t) + \theta_1 \epsilon(t-1) + \theta_2 \epsilon(t-2) + \cdots + \theta_q \epsilon(t-q) \quad (2.5)$$

Intergrated: Là quá trình lấy sai phân do yêu cầu chuỗi phải đảm bảo tính dừng. Hầu hết các chuỗi đều tăng hoặc giảm theo thời gian. Khi biến đổi sang chuỗi dừng, các nhân tố ảnh hưởng thời gian được loại bỏ và chuỗi sẽ dễ dự báo hơn. Để biến đổi từ chuỗi không dừng sang chuỗi dừng, phương pháp đơn giản nhất là lấy sai phân. Quá trình sai phân được thực hiện như sau:

Sai phân bậc một: $I(1) = \triangle(x_t) = x_t - x_{t-1}$ (2.6)

Sai phân bậc d: $I(d) = \Delta^d(x_t) = \Delta(\Delta(\ldots\Delta(x_t)))$ d lần lặp (2.7)

Tổng hợp ba thành phần AR, I, MA của mô hình ARIMA, phương trình hồi quy ARIMA (p, d, q) được biểu diễn như sau:

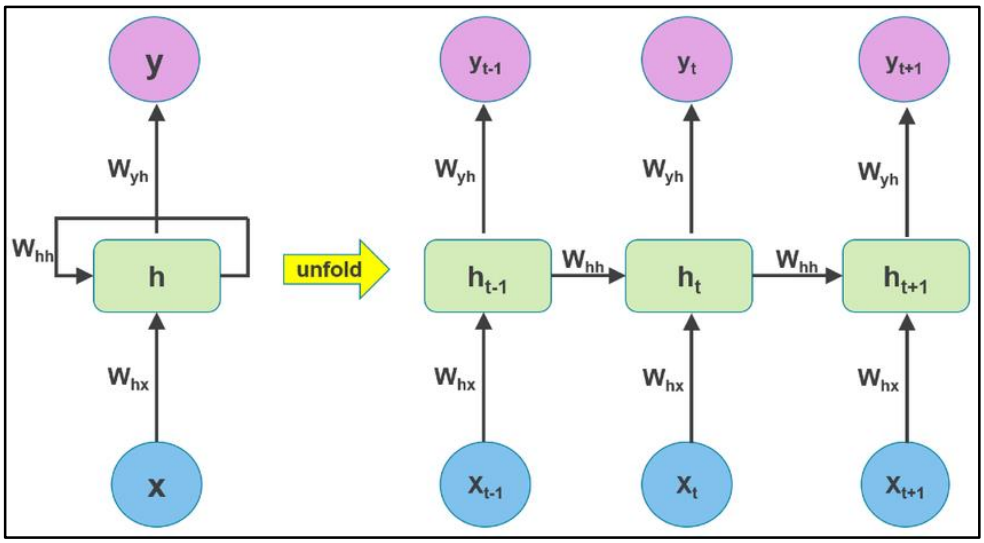
$$\Delta x_t = \phi_1 \Delta x_{t-1} + \cdots + \phi_p \Delta x_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \quad (2.8)$$

Trong đó : Δx_t là giá trị sai phân bậc d và ε_t là chuỗi nhiễu trắng

Mô hình ARIMA là sự cải tiến của mô hình ARMA khi bổ sung thành phần lấy sai phân để biến chuỗi dữ liệu không dừng thành chuỗi dữ liệu dừng.

2.2.2. Phương pháp dự báo dựa trên các mô hình hiện đại

a. Mô hình Recurrent Neural Network



Hình 2.10. Cấu trúc mô hình RNN

Mô hình RNN (Recurrent Neural Network) là một mô hình được thiết kế cho các bài toán nhận dữ liệu đầu vào có tính tuần tự. Điều này rất phù hợp cho các bài toán dự báo chuỗi thời gian, vì dữ liệu trong chuỗi thời gian có tính tuần tự về mặt thời gian và các giá trị trong chuỗi có tính phụ thuộc lẫn nhau giữa các thời điểm.

Một tế bào RNN (RNN cell) gồm các thành phần là dữ liệu đầu vào (x_t) tại thời điểm t, trạng thái ẩn lưu trữ thông tin từ các bước trước đó (h_t) và kết quả đầu ra tại thời điểm t (y_t)

RNN cập nhật trạng thái ẩn theo công thức:

$$h_t = f(W_h h_{t-1} + W_x x_t + b_h) \quad (2.9)$$

Trong đó: h_t là trạng thái ẩn hiện tại

h_{t-1} là trạng thái ẩn trước đó

x_t là dữ liệu đầu vào hiện tại

W_h là ma trận trọng số kết nối trạng thái ẩn giữa các bước thời gian

W_x là ma trận trọng số kết nối đầu vào và trạng thái ẩn

b_h là bias

f là hàm kích hoạt phi tuyến

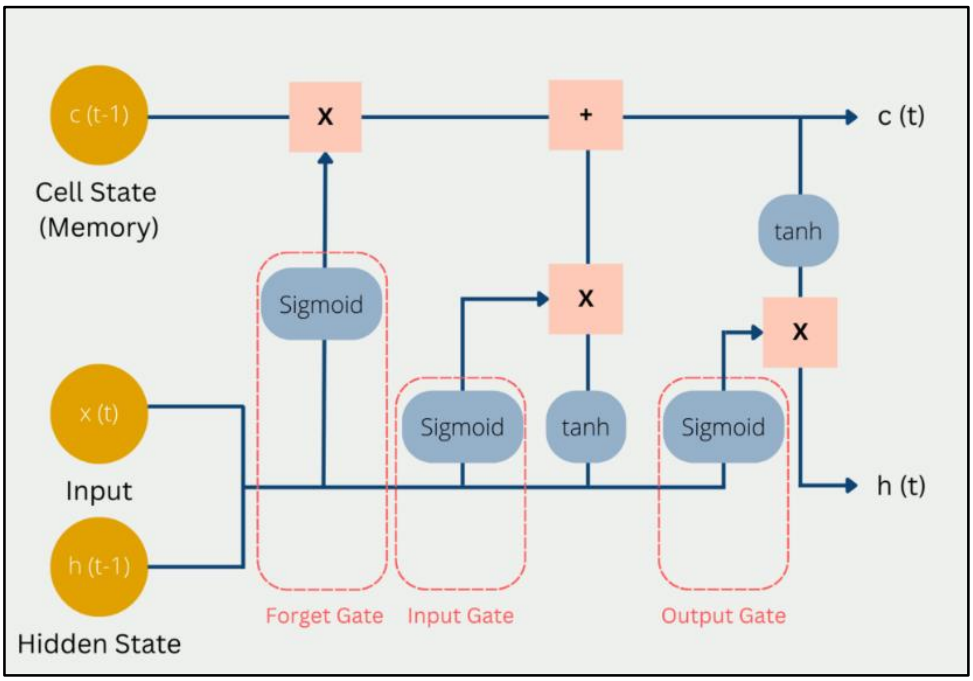
RNN dự đoán đầu ra theo công thức: $y_t = g(W_y h_t + b_y)$ (2.10)

Trong đó: W_y là ma trận trọng số kết nối giữa trạng thái ẩn và đầu ra

b_y là bias

g là hàm kích hoạt phi tuyến

b. Mô hình Long Short Term Memory



Hình 2.11. Cấu trúc mô hình LSTM

Mô hình RNN giải quyết các bài toán tuần tự , tuy nhiên mô hình RNN gặp vấn đề bị mất thông tin khi dữ liệu đầu vào quá dài do hiện tượng mất đạo hàm trong quá trình huấn luyện. Mô hình LSTM là một biến thể của mô hình RNN, được thiết kế đặc biệt để có thể ghi nhớ được thông tin dài hạn dựa trên các cổng: forget gate, input gate, output gate.

Cổng forget gate (cổng quên) quyết định những thông tin nào từ trạng thái trước đó sẽ bị loại bỏ theo công thức:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{2.11}$$

Trong đó: f_t là đầu ra của cổng quên tại thời điểm t

σ là hàm sigmoid

W_f là ma trận trọng số của cổng quên

b_f là bias

h_{t-1} là trạng thái ẩn từ các bước đó

x_t là dữ liệu đầu vào tại thời điểm t

Cổng input gate quyết định thông tin nào sẽ được lưu vào trạng thái bộ nhớ hiện tại theo công thức:

$$\begin{aligned} i_t &= \sigma(Wi \cdot [ht-1, x_t] + b_i) \\ \tilde{C}_t &= tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \\ C_t &= f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \end{aligned} \tag{2.12}$$

Trong đó: i_t là đầu ra của cổng nhập

$\tilde{C}_t$ là trạng thái bộ nhớ sau khi cập nhật

C_t là trạng thái bộ nhớ mới

Cổng output gate quyết đinh lấy thông tin nào làm đầu ra theo công thức:

$$\begin{aligned} o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\ h_t &= o_t \odot tanh(C_t) \end{aligned} \tag{2.13}$$

Trong đó: h_t là trạng thái ẩn đầu ra của mô hình

2.2.3. Phương pháp dự báo dựa trên các mô hình kết hợp

Việc kết hợp các mô hình dự báo đơn lẻ thành một kiến trúc mô hình mới cho dự báo chuỗi thời gian có khả năng mang lại hiệu suất tốt hơn so với sử dụng từng mô hình riêng lẻ. Chẳng hạn, các phương pháp kết hợp như hoặc CNN-LSTM đã được áp dụng rộng rãi trong các bài toán chuỗi thời gian.

Ví dụ trong mô hình kết hợp CNN-LSTM, mạng CNN được sử dụng để trích xuất các đặc trưng cục bộ từ dữ liệu chuỗi thời gian, chẳng hạn như các mẫu hoặc xu hướng ngắn hạn. Sau đó, các đặc trưng này được chuyển đến mạng LSTM, một mạng có khả năng ghi nhớ thông tin dài hạn, để phân tích các mối quan hệ và xu hướng dài hạn trong chuỗi. Sự kết hợp này tạo cơ sở cho việc dự đoán các giá trị tương lai với độ chính xác cao hơn.

Trên thực tế, dữ liệu chuỗi thời gian thường chứa nhiều biến động bất thường và không thể dự đoán trước, điều này đòi hỏi các mô hình phải đủ linh hoạt và mạnh mẽ để xử lý hiệu quả. Việc sử dụng các kiến trúc kết hợp giúp tận dụng ưu điểm của từng mô hình, từ đó giảm thiểu các sai số và tăng cường độ tin cậy của các dự báo.

2.3. Lựa chọn phương pháp

2.3.1. Lựa chọn phương pháp dự báo chuỗi thời gian

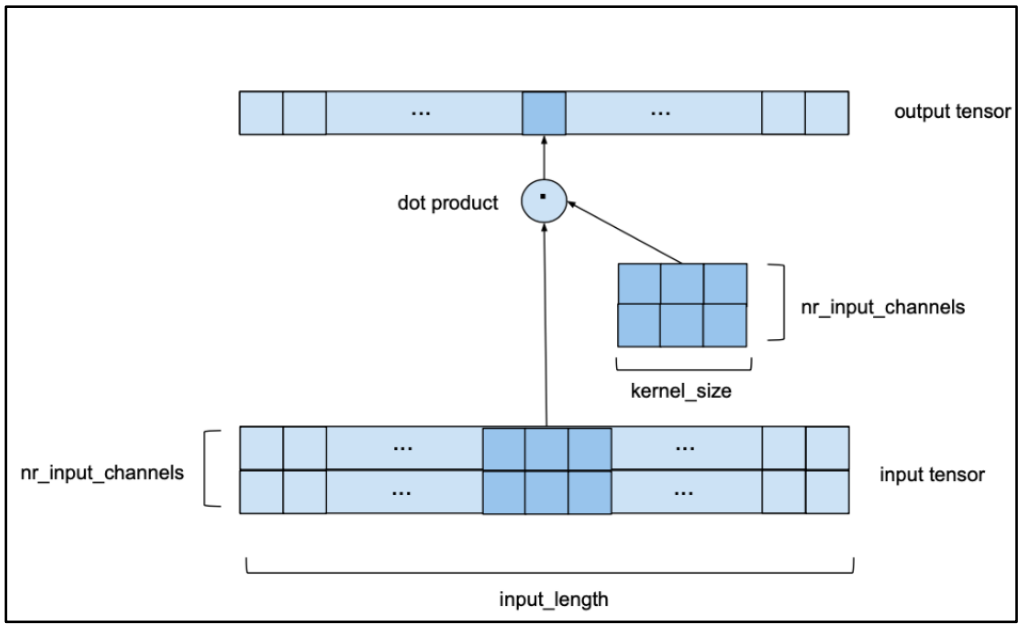
Với những khảo sát về các phương pháp dự báo trong bài toán chuỗi thời gian nhóm em đã trình bày tại các phần phía trên và dữ liệu chuỗi thời gian được các hệ thống IoT giám sát và thu thập, trong phần này đề xuất một mô hình dự báo cho bài toán chuỗi thời gian dựa trên sự kết hợp của các mô hình học máy. Cụ thể, mô hình đề xuất là sự kết hợp của 2 mô hình Temporal Convolutional Networks (TCN) và Gated Recurrent Unit Networks (GRU). Hai mô hình này cũng như sự kết hợp của hai mô hình sẽ được trình bày ngay sau đây.

a. Temporal Convolutional Networks

Mạng Temporal Convolutional Networks (TCN) là một mô hình mạng nơ ron sử dụng các lớp tích chập 1 chiều để xử lý tính toán giá trị tại thời điểm t trong bài toán dự báo chuỗi dữ liệu theo thời gian. Các thành phần của mạng TCN bao gồm: 1D Convolution, Causal Convolution, Dilated Convolution

1D Convolution sẽ khởi tạo ma trận bộ lọc có kích thước số lượng đặc trưng đầu vào nhân kích thước của kernel. Kernel của 1D sẽ trượt tuần tự trên chuỗi dữ liệu đầu vào, điều này phù hợp với tính toán trong chuỗi dữ liệu theo thời gian và tính toán đầu ra thông qua phép tích chập. *Hình 2.13* minh họa cách tính giá trị đầu ra với 1D Convolution.

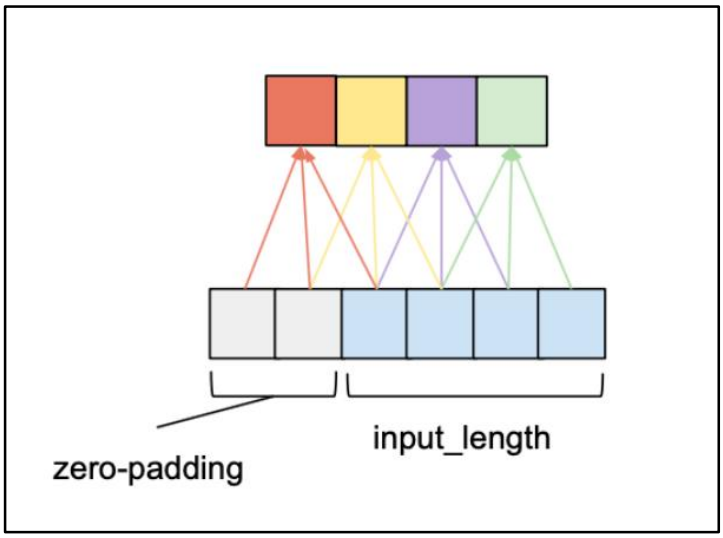
Với dự báo đa biến trong dữ liệu chuỗi thời gian, các đặc trưng được xếp chồng tạo thành các kênh đặc trưng, một bộ lọc có kích thước bằng kích thước bộ lọc nhân số biến trong dữ liệu chuỗi thời gian sẽ trượt tuần tự trên dữ liệu đầu vào và dùng phép tính tích chập để tính toán giá trị đầu ra.



Hình 2.13. Tích chập 1D trong mạng TCN

Causal Convolution sử dụng 1D Convolution để tính toán giá trị đầu ra với ràng buộc một giá trị được tính toán phụ thuộc vào các giá trị đứng trước nó. Điều này phù hợp với bài toán dự báo chuỗi thời gian khi tại thời điểm t, một giá trị được dự báo dựa trên các giá trị lịch sử. Causal Convolution sử dụng cơ chế "causal padding" để thêm các giá trị 0 vào đầu chuỗi dữ liệu. Số lượng giá trị 0 được thêm vào đầu chuỗi dữ liệu bằng kích thước bộ lọc trừ 1.

Hình 2.14 mô tả phép Causal Convolution. Giả sử chuỗi đầu vào có chiều dài là 4, kích thước bộ lọc là 3, tại thời điểm thứ 2 giá trị được tính toán đầu ra phụ thuộc vào giá trị thứ nhất và giá trị thứ 2 mà không sử dụng các giá trị đằng sau nó để tính toán.



Hình 2.14. Causal Convolution

Dilated Convolution giúp mạng TCN mở rộng được trường tiếp nhận thông tin hơn trong chuỗi đồng thời quan sát được các giá trị xa hơn trong chuỗi dữ liệu mà không làm tăng số lượng tham số tính toán *Hình 2.15* mô tả phép tính Dilated Convolution với các độ giãn (dilation rate) khác nhau.

Công thức tính giá trị đầu ra tại thời điểm t :

$$y[t] = \sum_{k=0}^{K-1} w_k . x[t - d.k] \tag{2.14}$$

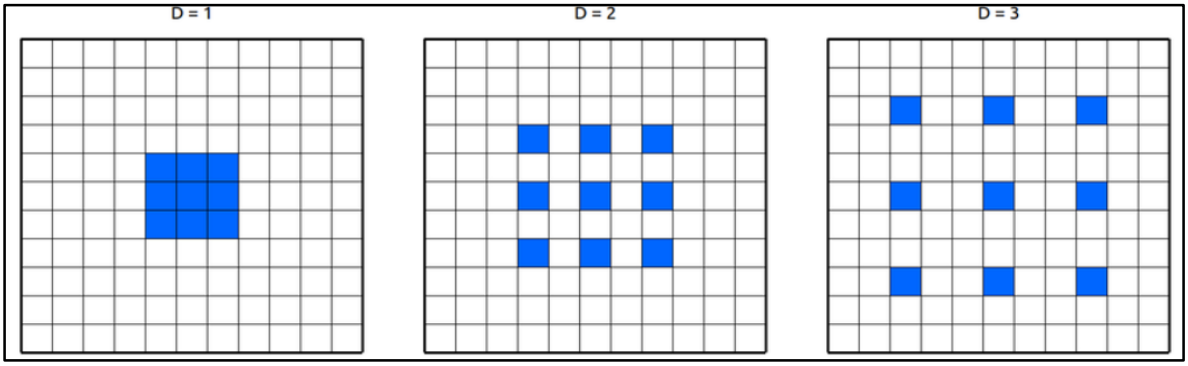
Trong đó: $y[t]$ là giá trị đầu ra tại thời điểm t

K là kích thước của bộ lọc (kernel size)

w_k là trọng số tại vị trí k của bộ lọc

$x[t - d.k]$ là giá trị đầu vào tại thời điểm t – k*d

d là độ giãn (dilation rate)



Hình 2.15. Dilated Convolution trong mạng TCN

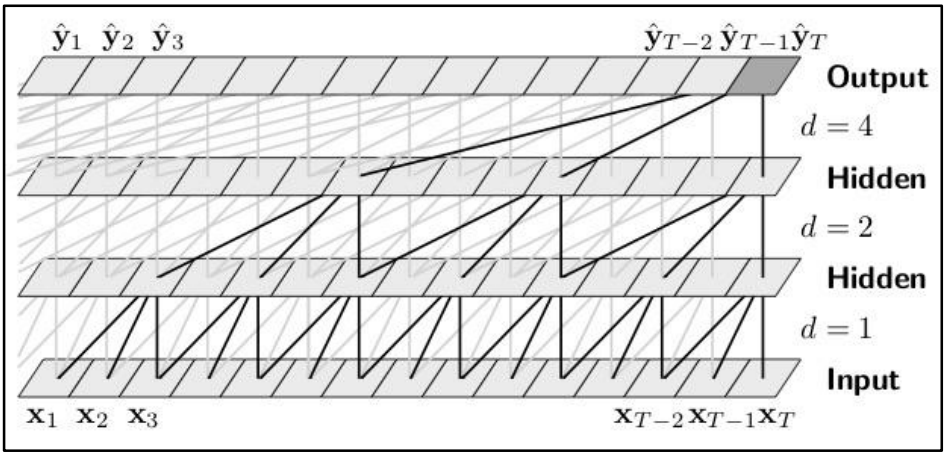
Trường tiếp nhận (Receptive Field) là độ dài chuỗi thông tin mà mạng có thể tiếp cận thông qua các lớp 1D Convolution và độ dãn d. Nếu mạng TCN có L lớp và mỗi lớp có kích thước bộ lọc K, trường tiếp nhận được tính theo công thức

$$ReceptiveField = 1 + \sum_{l=0}^{L-1} (K-1) * d_l \tag{2.15}$$

Trong đó: L là số lớp của mạng TCN

d_l là độ giãn (dilation rate) tại lớp thứ l

Để tăng kích thước trường tiếp nhận có thể thiết kế mạng TCN với nhiều lớp có độ giãn (dialtion rate) tăng theo cấp số nhân của cơ số 2. Điều này có thể giúp mạng TCN tính toán một giá trị trong tương lai dựa trên toàn bộ giá trị lịch sử trong chuỗi. *Hình 2.16* mô tả một mạng TCN gồm 3 tầng với độ giãn (dilation rate) tăng theo cấp số nhân tính toán một giá trị đầu ra tại thời điểm t phụ thuộc vào toàn bộ những giá trị trước đó.



Hình 2.16. Mô hình TCN với độ giãn khác nhau

Với các đặc điểm như trên, mô hình TCN là một phương pháp mới trong dự báo chuỗi thời gian với các phép tính tích chập thường xuất hiện trong các bài toán xử lý ảnh. Mô hình TCN đơn giản hơn và nhẹ hơn so với các mô hình tuần tự phức tạp trong xử lý chuỗi thời gian mà vẫn có thể cho kết quả tốt.

b. Gated Recurrent Unit Networks

Mạng GRU là một biến thể đơn giản hơn của mạng LSTM, được thiết kế để giảm độ phức tạp trong tính toán mà vẫn duy trì hiệu suất tốt khi xử lý các chuỗi dữ liệu dài. Mạng GRU bao gồm sử dụng hai cổng chính là cổng cập nhật (Update gate) và cổng đặt lại (Reset gate).

Cổng cập nhật quyết định bao nhiêu thông tin từ trạng thái trước đó cần được giữ lại và bao nhiêu thông tin mới từ đầu vào cần được thêm vào trạng thái hiện tại theo công thức:

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t] + b_z) \tag{2.16}$$

Trong đó: z_t là đầu ra của cổng cập nhật tại thời điểm t

W_z là ma trận trọng số của cổng cập nhật

h_{t-1} là trạng thái ẩn từ bước trước đó

x_t là dữ liệu đầu vào tại thời điểm t

σ là hàm sigmoid

Cổng đặt lại (Reset gate) quyết định bao nhiêu thông tin từ trang thái trước đó bị bỏ đi theo công thức:

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t] + b_r) \tag{2.17}$$

Trong đó: r_t là đầu ra của cổng đặt lại tại thời điểm t.

W_r là ma trận trọng số của cổng đặt lại

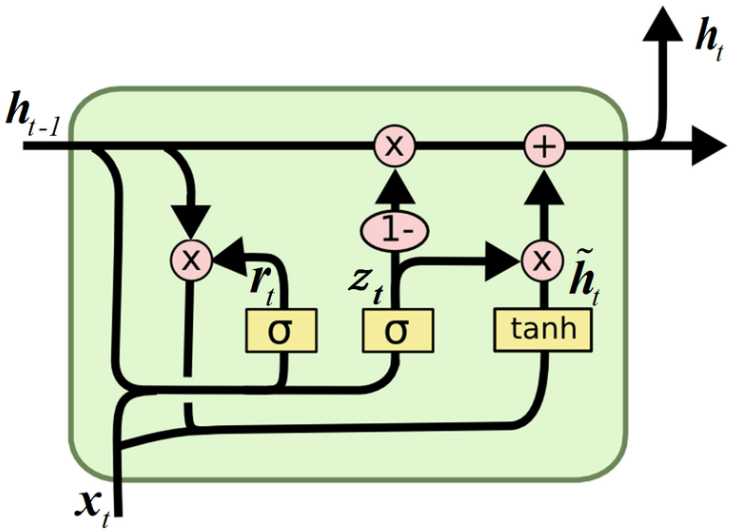
b_r bias

Trạng thái ẩn hiện tại được tính bằng cách kết hợp thông tin từ trạng thái trước đó sau khi đi qua cổng đặt lại và thông tin mới từ đầu vào theo công thức:

$$\begin{aligned} \tilde{h}_t &= tanh(W_h \cdot [r_t \odot h_{t-1}, x_t] + b_h) \\ h_t &= (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \end{aligned} \tag{2.18}$$

Trong đó: $\tilde{h}_t$ là trạng thái ẩn sau khi qua cổng đặt lại

h_t là trạng thái ẩn mới đầu ra



Hình 2.17. Cấu trúc mạng GRU

c. Hợp nhất mô hình TCN – GRU

Kết hợp mô hình TCN và GRU nhằm tận dụng ưu điểm của cả hai kiến trúc, giúp tăng cường khả năng dự báo chính xác trong các bài toán chuỗi thời gian. Mạng TCN (Temporal Convolutional Network) được sử dụng để trích xuất thông tin cục bộ từ chuỗi thời gian. Bằng cách áp dụng Causal Convolution và Dilated Convolution, TCN có thể mở rộng phạm vi tiếp nhận dữ liệu mà không làm mất đi tính chất tuần tự, đảm bảo rằng thông tin tương lai không bị sử dụng trong quá trình dự đoán. Mạng GRU (Gated Recurrent Unit) nhận các đặc trưng từ TCN, GRU sẽ xử lý thông tin thông qua khả năng lưu trữ và duy trì các thông tin dài hạn. Điều này cho phép GRU học được các mối quan hệ dài hạn trong chuỗi dữ liệu, từ đó dự đoán chính xác các giá trị tương lai dựa trên toàn bộ lịch sử giá trị của dữ liệu đầu vào.

CHƯƠNG 3: ÁP DỤNG VÀ TRIỂN KHAI MÔ HÌNH DỰA TRÊN KĨ THUẬT HỌC MÁY

3.1. Phân tích và tiền xử lý dữ liệu

Trong phần này trình bày quá trình phân tích và tiền xử lý dữ liệu cho tập dữ liệu là tập dữ liệu GEFCOM2014. Tập dữ liệu này thu thập dữ liệu về năng lượng mặt trời theo thời gian kèm theo các đặc trưng liên quan và đã được sử dụng trong các nghiên cứu khoa học gần đây.

3.1.1. Phân tích dữ liệu

a. Tập dữ liệu GEFCOM2014

TIMESTAMP	VAR78	VAR79	VAR134	VAR157	VAR164	VAR165	VAR166	VAR167	VAR169	VAR175	VAR178	VAR228	POWER
2012-04-01 06:00:00	0.036996	0.099045	94676.9375	72.374039	0.641353	1.333368	-1.728431	292.077148	11815767.0	7558415.0	14198503.0	0.003960	0.057244
2012-04-01 07:00:00	0.080911	0.121323	94708.0625	81.798737	0.753142	1.457923	-1.034620	291.069336	12274591.0	8798617.0	14925342.0	0.004970	0.088718
2012-04-01 08:00:00	0.036159	0.139069	94748.8125	87.854065	0.788338	2.374826	-1.089040	289.073486	12351290.0	10041167.0	15112951.0	0.006477	0.030064
2012-04-01 09:00:00	0.036372	0.072609	94785.8125	88.793488	0.502275	1.985531	-0.963010	288.031250	12351290.0	11257316.0	15112951.0	0.006725	0.000128
2012-04-01 10:00:00	0.014353	0.035797	94817.7500	90.450668	0.501918	1.999518	-0.930320	287.405762	12351290.0	12460132.0	15112951.0	0.006745	0.000000

Hình 3.1. Tập dữ liệu GEFCOM2014

Tập dữ liệu GEFCOM2014 được lấy từ bộ dữ liệu về năng lượng mặt trời tại cuộc thi Global Energy Forecasting Competition (GEFCOM2014). Bộ dữ liệu chứa dữ liệu năng lượng mặt trời được thu thập theo giờ tại ba khu vực cụ thể nằm trên nước Australia. Tập dữ liệu GEFCOM2014 cung cấp 12 đặc trưng về thời tiết ảnh hưởng đến sản lượng năng lượng mặt trời và cung cấp giá trị năng lượng mặt trời tương ứng cho mỗi khu vực khai thác năng lượng mặt trời. Dữ liệu được thu thập từ ngày 1 tháng 4 năm 2012 cho đến ngày 1 tháng 7 năm 2014, tương ứng với 821 ngày.

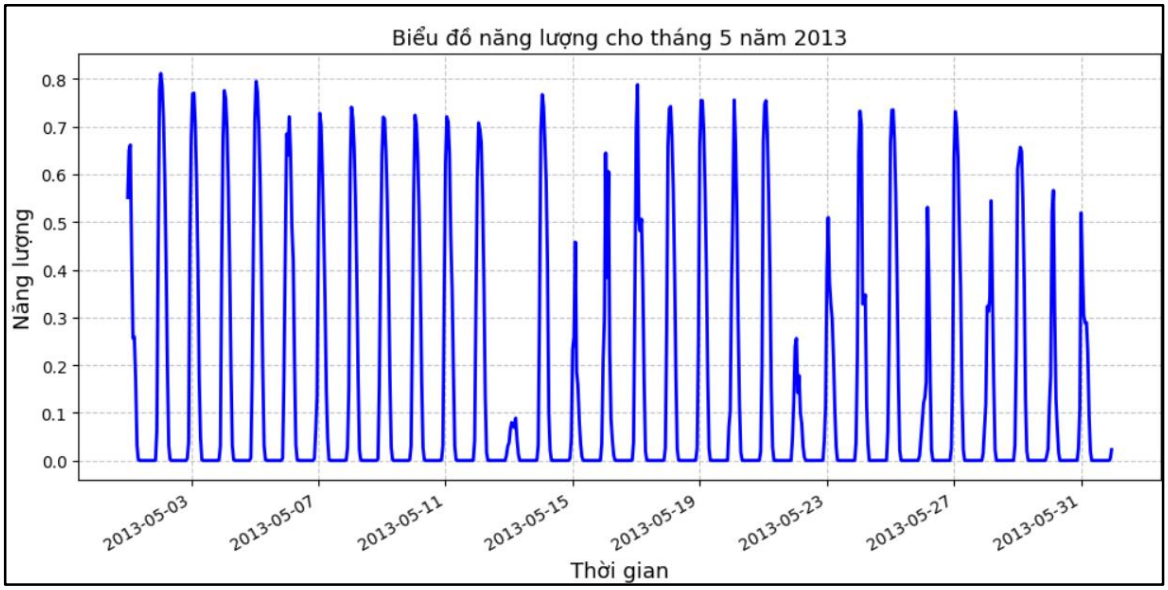
Do tập dữ liệu GEFCOM2014 cung cấp các đặc trưng về thời tiết giống nhau cho ba khu vực khai thác năng lượng mặt trời nên nhóm em sử dụng dữ liệu của khu vực 1 để tiến hành phân tích và tiền xử lý dữ liệu. *Bảng 1* mô tả chi tiết các đặc trưng trong tập dữ liệu GEFCOM2014.

Bảng 1. Mô tả các đặc trưng tập dữ liệu GEFCOM2014

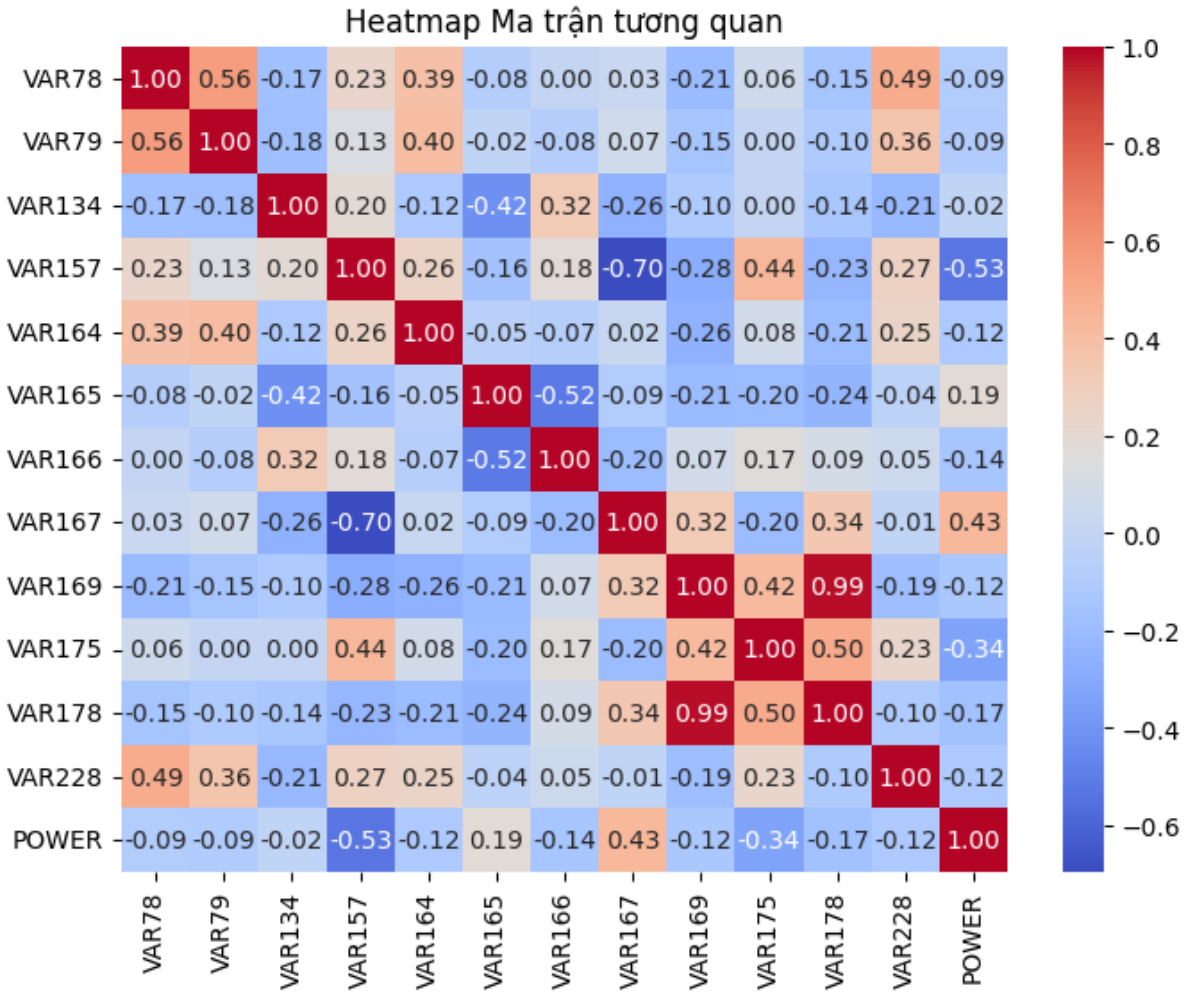
Mã đặc trưng	Mô tả
VAR78	Tổng lượng nước dạng lỏng trong mây
VAR79	Tổng lượng nước dạng đá trong mây
VAR134	Áp suất bề mặt
VAR157	Độ ẩm tương đối ở mực 1000mbar
VAR164	Tổng độ che phủ của mây

VAR165	Tốc độ gió hướng đông-tây ở độ cao 10 mét
VAR166	Tốc độ gió hướng bắc-nam ở độ cao 10 mét
VAR167	Nhiệt độ ở độ cao 2 mét
VAR169	Bức xạ mặt trời chiếu xuống bề mặt
VAR175	Bức xạ nhiệt chiếu xuống bề mặt
VAR178	Bức xạ ròng tầng cao nhất
VAR228	Tổng lượng mưa
POWER	Năng lượng mặt trời

Để phân tích các đặc điểm của dữ liệu chuỗi thời gian trong tập dữ liệu GEFCOM2014, cụ thể là biến năng lượng mặt trời theo thời gian, *Hình 3.2* minh họa sự biến đổi của năng lượng mặt trời trong suốt tháng 5 năm 2013. Từ *Hình 3.2*, có thể nhận thấy rõ tính mùa vụ lặp lại hàng ngày của năng lượng mặt trời. Các giá trị đỉnh tương đối đồng đều, phản ánh rằng trong giai đoạn này năng lượng mặt trời được khai thác ổn định theo từng ngày. Ngoài ra, *Hình 3.2* cũng ghi nhận một số giá trị bất thường, có thể giải thích là do ảnh hưởng của điều kiện thời tiết bất lợi trong những ngày đó.



Hình 3.2. Trực quan hóa năng lượng mặt trời tháng 5 năm 2013 trên tập dữ liệu GEFCOM2014



Hình 3.3. Ma trận tương quan biểu diễn mối quan hệ tuyến tính của các biến

3.1.2. Tiền xử lý dữ liệu

Tập dữ liệu GEFCOM2014 cung cấp dữ liệu theo dạng chuỗi thời gian. Quy trình tiền xử lý dữ liệu được diễn ra tuần tự bao gồm các bước xử lý dữ liệu bị thiếu, lựa chọn đặc trưng, chuẩn hóa dữ liệu, chia dữ liệu và tạo dữ liệu tuần tự để đưa vào mô hình huấn luyện.

Đầu tiên, sau khi kiểm tra, tập dữ liệu GEFCOM2014. Để giải quyết vấn đề này, các giá trị ở thời gian gần nhất trước đó sẽ được điền bổ sung tại thời điểm bị thiếu dữ liệu. Sau khi dữ liệu đã được điền bổ sung, dữ liệu đã đảm bảo được tính đầy đủ và liên tục về mặt thời gian để có thể xử lý các bước tiếp theo.

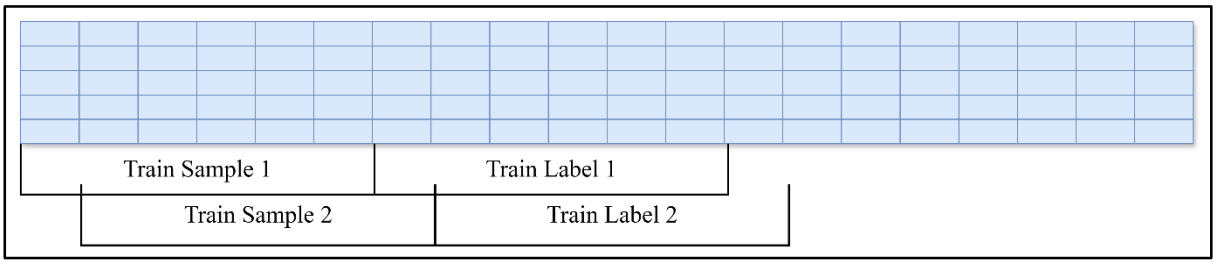
Tiếp theo sau khi xử lý dữ liệu bị thiếu sẽ lựa chọn các đặc trưng để đưa vào mô hình huấn luyện. Đối với bài toán dự báo năng lượng mặt trời, các đặc trưng ảnh hưởng đến năng lượng mặt trời cần được đưa vào để đảm bảo tính khách quan của dự báo. Tuy nhiên feature VAR169 và VAR178 có sự tương quan mạnh lẫn nhau có thể gây ra vấn đề đa cộng tuyến nên trong trường hợp này ta loại bỏ feature VAR178. Ngoài ra, do tính chất bài toán dự báo chuỗi thời gian là dự báo các giá trị trong tương lai dựa trên các giá

trị trong quá khứ, do đó bản thân đặc trưng về năng lượng mặt trời cũng được đưa vào mô hình huấn luyện nhằm mục đích tăng cường độ chính xác dự báo. Do vậy trong nghiên cứu này sẽ thiết lập các cột về thời gian làm chỉ số và không đưa vào mô hình huấn luyện, các đặc trưng bao gồm đặc trưng ảnh hưởng đến năng lượng mặt trời và năng lượng mặt trời trong 2 tập dữ liệu sẽ được sử dụng làm dữ liệu đầu vào của mô hình dự báo.

Do tính chất của dữ liệu chuỗi thời gian là tuần tự và liên tục, việc chuẩn hóa dữ liệu là rất quan trọng để mô hình dự báo có thể hội tụ nhanh chóng trong quá trình huấn luyện. Tại bước chuẩn hóa dữ liệu này, sử dụng kĩ thuật chuẩn hóa Standard Scaler.

Tiếp theo sẽ chia dữ liệu thành ba tập train, validation, test trên 2 tập dữ liệu trên theo tỉ lệ 80%, 10%, 10% và đảm bảo tính tuần tự về mặt thời gian. Tập train để đưa vào huấn luyện mô hình dự báo, tập validation để kiểm thử trong quá trình huấn luyện mô hình dự báo, tập test để kiểm tra dự đoán sau khi quá trình huấn luyện mô hình dự báo kết thúc.

Bước cuối cùng trong quá trình chuẩn bị dữ liệu cho mô hình dự báo chuỗi thời gian là tạo dữ liệu tuần tự (hay còn gọi là tạo dữ liệu dạng cửa sổ). Mục đích của bước này là chuyển đổi dữ liệu lịch sử thành một dạng mà mô hình có thể sử dụng để dự báo các giá trị trong tương lai. Đây là bước rất quan trọng vì trong bài toán dự báo chuỗi thời gian, mô hình sẽ học từ các giá trị quá khứ để dự đoán các giá trị tương lai. Cấu trúc tạo dữ liệu tuần tự thường được mô tả bằng cách chia dữ liệu thành các cửa sổ (windows) có kích thước cố định. Mỗi cửa sổ chứa một dãy các giá trị liên tiếp của chuỗi thời gian tại thời điểm trước đó (Train Sample) và mô hình sẽ sử dụng dãy này để dự báo giá trị tiếp theo hoặc một số giá trị tiếp theo trong tương lai (Train Label). *Hình 3.6* mô tả cách tạo dữ liệu dạng chuỗi. Cấu trúc này giúp đảm bảo mô hình học được các mối quan hệ tuần tự và phụ thuộc thời gian giữa các giá trị dữ liệu. Trong nghiên cứu này, các giá trị lịch sử được đưa vào dự đoán bao gồm tất cả các đặc trưng đã được lựa chọn tại bước lựa chọn đặc trưng để dự báo các giá trị năng lượng mặt trời trong tương lai. Giá trị về năng lượng mặt trời chính là nhãn của mô hình dự báo.



Hình 3.6. Mô tả tạo dữ liệu tuần tự

3.2. Triển khai mô hình dự báo

3.2.1. Thiết lập mô hình dự báo cho quá trình thử nghiệm

Đầu tiên, thiết lập một lớp đầu vào để tiếp nhận dữ liệu với kích thước bao gồm độ dài của chuỗi dữ liệu lịch sử và số lượng đặc trưng**.** Tiếp theo, xây dựng 5 khối mạng TCN với tham số dilation rate được tăng dần theo cấp số nhân và số lượng bộ lọc theo thứ tự 2 – 4 – 8 – 16 - 32. Việc này giúp mở rộng trường tiếp nhận (receptive field) và cho phép mô hình học được các mối quan hệ dài hạn trong dữ liệu lịch sử. Tham số padding thiết lập là 'causal' để đảm bảo tính toán các giá trị tương lai sẽ phụ thuộc vào các giá trị trong quá khứ. Các lớp BatchNomalize nhằm mục đích chuẩn hóa các đầu ra để giúp mô hình hội tụ nhanh chóng trong quá trình huấn luyện.

Sau đó, sử dụng một lớp GRU với số đơn vị 64 lớp để học thêm các phụ thuộc tuần tự trong chuỗi dữ liệu một cách dài hạn.

Cuối cùng, lớp đầu ra được thiết lập với kích thước tương ứng với số lượng giá trị cần dự báo trong tương lai. Tại lớp này, sử dụng hàm kích hoạt 'relu' đảm bảo mô hình dự báo giá trị luôn lớn hơn hoặc bằng 0 do giá trị của năng lượng mặt trời không thể nhỏ hơn 0. *Bảng 3* thể hiện chi tiết kiến trúc mô hình dự báo.

Sau khi đã thiết lập kiến trúc mô hình, thiết lập hàm mất mát là hàm Mean Square Error (MSE). Để mô hình có thể học và cập nhật trọng số, thiết lập hàm tối ưu hóa là hàm Adam với tốc độ học ban đầu là 0.001. Tuy vào quá trình học mà tốc độ học có thể được thay đổi để mô hình có thể tổng quát hóa dữ liệu một cách tốt nhất.

Bảng 2. Thiết lập kiến trúc mô hình dự báo

Lớp	**Các tham số**	**Kích thước đầu ra**
Input Layer	(window_size,num_features)	(window_size,num_features)
Conv1D BatchNomalize (Block 1)	filters=num_features, kernel_size=3, dilation_rate=1, activation='relu', padding='causal'	(window_size, num_features)
Conv1D BatchNomalize (Block 2)	filters=num_features, kernel_size=3, dilation_rate=2, activation='relu', padding='causal'	(window_size, num_features)
Conv1D BatchNomalize (Block 3)	filters=num_features, kernel_size=3, dilation_rate=4, activation='relu', padding='causal'	(window_size, num_features)

Conv1D BatchNomalize (Block 4)	filters=num_features, kernel_size=3, dilation_rate=8, activation='relu', padding='causal'	(window_size, num_features)
Conv1D BatchNomalize (Block 5)	filters=num_features, kernel_size=3, dilation_rate=16, activation='relu', padding='causal'	(window_size, num_features)
Dense	units=gru_unit	(window_size, gru_unit)
GRU	units=gru_unit	(window_size, gru_unit)
Output Layer (Linear)	units = length output prediction	(window_size, length output prediction)

3.2.2. Triển khai huấn luyện mô hình

Sau khi thiết lập mô hình dự báo, tiến hành huấn luyện trên hai tập dữ liệu với các kịch bản khác nhau tương ứng với độ dài chuỗi dữ liệu lịch sử đầu vào và thời gian dự báo khác nhau.

Kịch bản 1: Dữ liệu lịch sử đầu vào và thời gian dự báo là 24 bước thời gian, tương đương với 1 ngày.

Kịch bản 2: Dữ liệu lịch sử đầu vào và thời gian dự báo là 48 bước thời gian, tương đương với 3 ngày.

Kịch bản 3: Dữ liệu lịch sử đầu vào và thời gian dự báo là 144 bước thời gian, tương đương với 6 ngày.

Mô hình được huấn luyện với 50 lần lặp và sử dụng GPU T4x2 nhằm tăng tốc quá trình đào tạo và suy luận, đảm bảo hiệu quả và tiết kiệm thời gian xử lý.

3.3. Phân tích và đánh giá kết quả

3.3.1. Các phương pháp đánh giá

a. Mean Square Error

Mean Square Error (MSE) là một chỉ số đo lường sự sai lệch trung bình giữa giá trị dự đoán và giá trị thực tế trong bài toán hồi quy. Cụ thể, MSE được tính bằng cách lấy bình phương của sự sai lệch giữa giá trị dự đoán và giá trị thực tế, sau đó tính trung bình của tất cả các sai lệch này. Công thức tính MSE:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (3.2)$$

MSE là một chỉ số quan trọng trong các bài toán hồi quy vì nó đo lường mức độ chính xác của mô hình dự đoán bằng cách đánh giá sự sai lệch giữa các giá trị thực tế và dự đoán. MSE càng nhỏ, mô hình càng chính xác trong việc dự đoán giá trị thực tế.

b. Mean Absolute Error

Mean Absolute Error (MAE) là một chỉ số đánh giá sự sai lệch tuyệt đối trung bình giữa giá trị dự đoán và giá trị thực tế. Công thức tính MAE:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (3.3)$$

MAE đánh giá sự sai lệch tuyệt đối trung bình, có nghĩa là MAE cho biết mức độ trung bình của sự sai lệch giữa các giá trị thực tế và dự đoán mà không tính đến dấu hiệu của sai lệch (dương hay âm). MAE dễ hiểu hơn so với MSE và không có sự chênh lệch quá mức đối với các sai số lớn, điều này khiến cho nó trở thành một chỉ số phù hợp hơn trong các bài toán mà các ngoại lai có thể xảy ra.

c. Coefficient of Determination

Coefficient of Determination (R2 scores) là một chỉ số đánh giá mức độ phù hợp của mô hình hồi quy với dữ liệu thực tế. R^2 cho biết phần trăm biến thiên trong dữ liệu mà mô hình có thể giải thích. Giá trị R^2 có thể nằm trong khoảng từ 0 đến 1 (hoặc âm trong một số trường hợp đặc biệt). Công thức tính R^2:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (3.4)$$

R^2 là một thước đo quan trọng trong các mô hình hồi quy vì nó cho biết tỷ lệ phần trăm sự biến động của giá trị thực tế mà mô hình có thể học được. Nếu giá trị R^2 gần bằng 1, điều này có nghĩa là mô hình học được hầu hết sự biến động trong dữ liệu và có thể dự đoán chính xác. Ngược lại, nếu R^2 gần bằng 0, mô hình không học được biến động trong dữ liệu và dự báo kém hiệu quả.

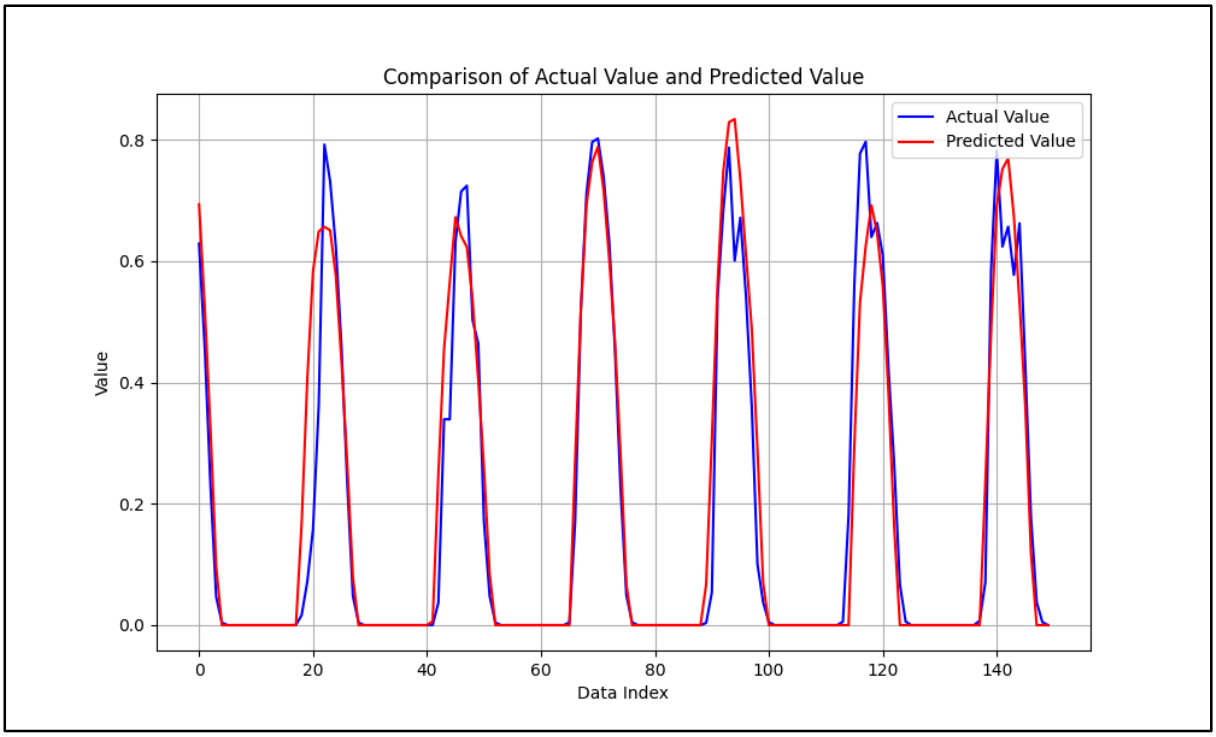
3.3.2. Đánh giá kết quả thử nghiệm

Sau khi huấn luyện mô hình kết thúc, mô hình sẽ được tải và dự báo các giá trị năng lượng trên tập dữ liệu test đã chia trước đó từ tập dữ liệu sau đó tiến hành đánh giá kết quả dựa trên các kịch bản thử nghiệm đã nêu.

a. Đánh giá kết quả thử nghiệm trên tập dữ liệu GEFCOM2014

Bảng 3. Kết quả thử nghiệm trên tập dữ liệu GEFCOM2014 với độ dài dữ liệu lịch sử

Độ dài dữ liệu lịch sử	MSE	MAE	R2 scores
24	0.007757	0.054802	80.25%
48	0.009267	0.055472	85.75%
144	0.005461	0.043656	89.55%



Hình 3.7. Kết quả so sánh giá trị dự đoán và thực tế với 144 bước thời gian trên tập dữ liệu GEEFCOM2014

Dựa vào *Bảng 4* cho thấy mô hình dự báo kết quả năng lượng mặt trời rất tốt. Các tham số đánh giá như MSE và MAE phản ảnh độ sai lệch giữa giá trị dự đoán và giá trị thực tế là rất nhỏ. Mô hình dự báo đã học được tính biến động của chuỗi dữ liệu thời gian trong tập dữ liệu với chỉ số R2 đạt gần 90% khi nhận đầu vào là 144 bước thời gian. Điều này có thể khẳng định, với 144 bước thời gian đầu vào mô hình dự báo đã học được các đặc điểm của chuỗi dữ liệu thời gian từ đó đã dự báo một cách gần chính xác.

Hình 3.7 trực quan hóa giá trị dự đoán và giá trị trên thực tế với 144 bước thời gian chuỗi dữ liệu lịch sử cho thấy mô hình dự báo chính xác cao.

3.4. Kết luận chương

Trong chương này trình bày áp dụng và triển khai mô hình dự báo TCN-GRU đã được đề xuất trước đó. Mô hình TCN-GRU được huấn luyện trên hai tập dữ liệu đã được sử dụng trong các nghiên cứu gần đây. Kết quả dự báo cho thấy mô hình TCN-GRU đã dự báo rất tốt trên cả hai tập dữ liệu. Trong thời gian tiếp theo cần tối ưu hóa mô hình để có thể triển khai trên các thiết bị biên nhằm đáp ứng khả năng mở rộng và quản lý năng lượng mặt trời thông minh.

KẾT LUẬN CHUNG

Năng lượng tái tạo là xu hướng về mặt năng lượng ở thời điểm hiện tại cũng như trong tương lai. Tiềm năng phát triển của năng lượng tái tạo là rất lớn và một trong số đó là năng lượng mặt trời. Năng lượng mặt trời đã và đang được đầu tư và khai thác trên một quy mô lớn. Việc quản lý thông minh sẽ giúp tối ưu hóa khả năng vận hành cũng như khai thác nguồn năng lượng này. Dự báo chính xác năng lượng mặt trời là điều cần thiết và rất hữu ích cho nhiệm vụ trên. Đã có rất nhiều nghiên cứu cho bài toán dự báo chuỗi thời gian, trong tiểu luận này nghiên cứu và áp dụng bài toán chuỗi thời gian cho dự báo năng lượng mặt trời dựa trên kĩ thuật học máy và cho kết quả tốt trên tập dữ liệu có sẵn. Trong tương lai, cần tối ưu mô hình dự báo để có thể triển khai trên thực tế và đảm bảo độ chính xác cao của dự báo.

Đây là những kết luận và định hướng cho tương lai của tiểu luận. Mặc dù đã cố gắng hết sức, nhóm em tự nhận thức rằng vẫn còn những thiếu sót cần khắc phục. Rất mong nhận được những ý kiến đóng góp từ thầy/cô và các bạn để nhóm em có thể hoàn thiện hơn trong các nghiên cứu tiếp theo.