Kỷ yếu Hội nghị Khoa học công nghệ Quốc gia lần thứ XVII về Nghiên cứu cơ bản và ứng dụng Công Nghệ thông tin (FAIR); Hà Nội, ngày 08-09/8/2024
DOI: 10.15625/vap.2024.0242

MÔ HÌNH TÍCH HỢP TFT-TPE CHO DỰ BÁO DỮ LIỆU BẤT THƯỜNG CỦA NHU CẦU DU LỊCH VIỆT NAM TRONG GIAI ĐOẠN COVID-19

Ngô Văn Sơn[1, 3,*], Thái Thị Phương[2], Nguyễn Dương Thiện[3], Võ Viết Minh Nhật[4]

[1]Trường Đại học Khoa học, Đại học Huế
[2]Ban Đào tạo và Công tác Sinh viên - Đại học Huế
[3]Trường Du lịch - Đại học Huế
[4]Viện Khảo thí và Bảo đảm chất lượng giáo dục - Đại học Huế

ngovanson@hueuni.edu.vn, ttphuong@hueuni.edu.vn, ndthien@hueuni.edu.vn, vvmnhat@hueuni.edu.vn

**Corresponding author*

TÓM TẮT: *Dự báo đóng vai trò rất quan trọng trong sản xuất và kinh doanh. Tuy nhiên những sự kiện bất thường và không thể lường trước, như đại dịch Covid-19, đã khiến cho việc dự báo khó có thể đạt được chính xác. Trong những trường hợp này, việc xác định một phương pháp dự báo chính xác có thể thích ứng với những biến động bất thường là một yêu cầu cấp thiết. Bài báo đề xuất mô hình tích hợp Temporal Fusion Transformer (TFT) và Tree-Structured Parzen Estimator (TPE) nhằm có thể dự báo chính xác hơn đối với những sự kiện bất thường. TFT là một mô hình mạng nơ-ron học sâu được thiết kế đặc biệt cho việc xử lý dữ liệu chuỗi thời gian, mà nó có thể nắm bắt các xu hướng và mô hình hóa các biến đổi phức tạp của dữ liệu. Với TPE, đây là một phương pháp tối ưu hóa dựa trên cấu trúc dữ liệu giống cây để xác định tập siêu tham số tốt nhất. Tích hợp TFT và TPE có thể cung cấp một giải pháp hiệu quả cho bài toán dự báo dữ liệu bất thường. Mô hình TFT-TPE được đánh giá bởi một bộ dữ liệu nhu cầu du lịch Việt Nam trong giai đoạn Covid-19. Thực nghiệm cho thấy mô hình TFT-TPE có khả năng thích ứng và đạt được kết quả dự báo khá chính xác đối với dữ liệu bất thường khi có tích hợp với dữ liệu Covid-19.*

Từ khóa: *Transformer, TFT, TPE, chuỗi thời gian, dữ liệu bất thường, dự báo nhu cầu du lịch.*

I. GIỚI THIỆU

Dự báo chuỗi thời gian là chủ đề được quan tâm trong nhiều lĩnh vực nghiên cứu, từ dự báo thời tiết, biến đổi khí hậu, tăng trưởng kinh tế, biến động chứng khoán cho đến nhu cầu du lịch. Dự báo là việc đưa ra các tiên đoán trong ngắn hạn và dài hạn dựa trên những hiểu biết về các sự kiện trong quá khứ. Đối với dự báo dữ liệu chuỗi thời gian, các nhà nghiên cứu thường phân tích chuỗi dữ liệu được cho để tìm ra những điểm đặc trưng và sau đó chọn kỹ thuật thích hợp cho việc dự báo [1]. Tuy nhiên, việc dự báo chính xác trở nên khó khăn hơn khi xuất hiện những mẫu dữ liệu bất thường. Dữ liệu bất thường có thể là những biến động tăng giảm đột biến hay những thay đổi bất ngờ về xu hướng trong chuỗi dữ liệu khiến cho các kỹ thuật dự báo truyền thống trở nên kém hiệu quả [2]. Sự bất thường có thể do nhiều nguyên nhân khác nhau, như cú sốc thị trường, bùng phát dịch bệnh, lỗi thu thập dữ liệu hoặc hành vi kỳ lạ nào đó của con người. Khi đối mặt với dữ liệu bất thường, các phương pháp dự báo truyền thống không nắm bắt được các đặc điểm dị biệt, dẫn đến dự báo không chính xác. Những phương pháp dự báo truyền thống thường dựa trên các giả định về tính dừng và tính đều đặn, nên không đúng với những bất thường trong dữ liệu chuỗi thời gian. Dữ liệu bất thường gây ra những thách thức đáng kể cho các nhà dự báo do tính chất khó có thể đoán trước được của nó.

Đối với hướng tiếp cận dự báo chuỗi thời gian dựa trên học máy, trong quá trình huấn luyện, một bước không thể thiếu và cực kỳ quan trọng là khởi động giá trị ban đầu cho các tham số như số chu kỳ (epochs), số lớp,... Việc lựa chọn các tham số hợp lý thường dựa trên kinh nghiệm. Với mỗi bộ tham số, chúng ta phải huấn luyện mô hình, quan sát kết quả đạt được, đánh giá chúng, điều chỉnh lại tham số và lặp lại công việc này. Để tự động hóa quy trình trên, một số thuật toán như Grid Search hay Random Search [3] có thể được sử dụng. Tuy nhiên, các thuật toán này chỉ hoạt động hiệu quả với tập ít tham số vì không gian tìm kiếm sẽ tăng rất nhanh khi số lượng tham số lớn và tiêu tốn nhiều thời gian để tìm thấy giải pháp phù hợp. Thuật toán tối ưu hóa siêu tham số Tree-structured Parzen Estimator (TPE) [4] có thể giúp vượt qua trở ngại này bằng cách xây dựng một mô hình ước lượng phân phối có điều kiện đối với các siêu tham số dựa trên các mẫu đã thu thập.

Bài báo đề xuất một mô hình kết hợp nhằm nâng cao khả năng dự báo chính xác đối với các sự kiện bất thường và không lường trước, như nhu cầu du lịch trong bối cảnh đại dịch Covid-19. Chúng tôi sử dụng TFT (Temporal Fusion Transformer) [5], một công cụ mạnh mẽ cho dự báo chuỗi thời gian, nhằm nâng cao khả năng dự báo chính xác và thích ứng với biến động bất thường của dữ liệu. Tuy nhiên, việc xác định tập tham số ban đầu phù hợp cho TFT là một thách thức. TPE [4] là một phương pháp tối ưu hóa tham số có thể được sử dụng để cải thiện hiệu suất của TFT. Bằng cách tích hợp TPE vào TFT, mô hình TFT-TPE có thể cải thiện được hiệu suất dự báo. Chúng tôi đánh giá hiệu quả của mô hình TFT-TPE với một bộ dữ liệu nhu cầu du lịch trong giai đoạn dịch Covid-19. Kết quả thực nghiệm cho thấy rằng mô hình TFT-TPE đạt được độ chính xác dự báo cao hơn khi bổ sung dữ liệu bất thường về dịch Covid-19 làm đồng biến so với mô hình khi không được bổ sung dữ liệu về dịch Covid-19.

Các đóng góp chính của bài báo bao gồm:

• Đề xuất mô hình tích hợp TFT và TPE (mô hình TFT-TPE) và mô hình hóa bài toán dự báo nhu cầu du lịch thành bài toán tìm kiến trúc TFT tối ưu;

• Phân tích, chuẩn hóa và chuyển đổi dữ liệu chuỗi thời gian nhu cầu du lịch trong giai đoạn dịch Covid-19 thành bộ dữ liệu đầu vào cho mô hình TFT-TPE;

• Triển khai và đánh giá hiệu quả dự báo nhu cầu du lịch của mô hình TFT-TPE với bộ dữ liệu chuỗi thời gian về lượt khách quốc tế đến Việt Nam.

Các nội dung tiếp theo của bài báo được tổ chức như sau. Mục II tóm lược và đánh giá các nghiên cứu liên quan trong các năm trở lại đây, trong đó tập trung vào các mô hình dự báo dựa trên Transformer và các biến thể của nó. Trên cơ sở các phân tích, Mục III sẽ mô tả chi tiết mô hình TFT, phương pháp TPE và phân tích mô hình TFT-TPE được đề xuất và Mục IV mô tả việc cài đặt và đánh giá kết quả. Cuối cùng, kết luận được trình bày trong Mục V.

II. NGHIÊN CỨU LIÊN QUAN

Dự báo dữ liệu bất thường là một thách thức trong nhiều lĩnh vực khác nhau, như thương mại điện tử, vật tư y tế, tiêu thụ điện năng, sản xuất dược phẩm, và cả nhu cầu du lịch. Đã có một số phương pháp khác nhau được đề xuất nhằm giải quyết vấn đề này. Mục này đánh giá các nghiên cứu liên quan đến dự báo đối với dữ liệu chuỗi thời gian bất thường trong thời gian gần đây.

Trong nghiên cứu [6], nhóm tác giả đề xuất kết hợp thống kê dữ liệu lịch sử với hồi quy tuyến tính dùng trong dự báo nhu cầu thiết bị y tế với dữ liệu thưa thớt, tạm thời và bất thường. Để giảm bớt các ước tính thưa thớt, việc chuyển đổi dữ liệu lịch sử được thêm vào mô hình hồi quy tuyến tính. Tuy nhiên, đề xuất dựa trên thống kê và kỹ thuật hồi quy tuyến tính này gặp khó khăn trong việc nắm bắt các mối quan hệ phi tuyến phức tạp có trong dữ liệu dự báo nhu cầu thiết bị y tế. Không giống như các mô hình học sâu, có thể tự động học các mẫu và tính năng phức tạp từ dữ liệu, nghiên cứu trong [6] yêu cầu kỹ thuật trích rút tính năng thủ công và đặc tả mô hình, nên hạn chế khả năng thích ứng với đa biến trong dự báo.

Nghiên cứu [7] đã đưa ra giả thuyết rằng việc tận dụng kiến thức bên ngoài, như trong các sự kiện trên thế giới, có thể giúp cải thiện về dự báo trong điều kiện bất thường. Nghiên cứu này đã khai thác một kho lưu trữ các sự kiện thế giới trong 40 năm. Sau đó, D. Kalifa và cs. đã đề xuất một phương pháp mới dựa trên Transformer để xây dựng dữ liệu một ngày dựa trên mối quan hệ của các sự kiện trong ngày. Dữ liệu này sau đó được sử dụng để dự báo hành vi của người tiêu dùng trong tương lai. Một tập dữ liệu bán hàng thương mại điện tử từ eBay được thực nghiệm trên phương pháp được đề xuất. Kết quả thu được cho thấy phương pháp được đề xuất dự báo chính xác hơn so với các phương pháp thống kê như ARIMA hay phương pháp học sâu như Prophet.

Trong khi đó nghiên cứu [10] đề xuất một tiếp cận sử dụng Isolation Forest [11], Random Forest [12] và Decision Tree để dự báo hành vi tiêu thụ điện bất thường với độ chính xác cao. Phương pháp đề xuất sử dụng thuật toán Isolation Forest để gắn nhãn (bình thường hoặc bất thường) chỉ số tiêu thụ điện của đồng hồ điện thông minh. Từ việc gán nhãn này, dữ liệu chuỗi thời gian sẽ tạo thành nhiều chuỗi thời gian với độ dài khác nhau. Dựa trên dữ liệu này, hai thuật toán Random Forest và Decision Tree được áp dụng để dự báo khả năng xảy ra tình trạng tiêu thụ điện năng bất thường. Kết quả thực nghiệm cho thấy phương pháp đề xuất phát hiện và dự báo chính xác các trạng thái bất thường trước 30 phút. Không có sự khác biệt đáng kể về hiệu suất giữa Random Forest và Decision Tree với kích thước tập dữ liệu và độ dài chuỗi dữ liệu khác nhau.

Đối với trường hợp bệnh viện gặp khó khăn trong việc quản lý thuốc tồn kho (dư thừa hoặc thiếu hụt) khi xuất hiện nhu cầu bất thường đối với một số loại thuốc. Nghiên cứu [13] đã so sánh với một số phương pháp như Croston [14], TSB [15], SBA [16] và Kalaya [17] trên cùng dữ liệu nhu cầu thuốc bất thường. Kết quả cho thấy phương pháp Kalaya mang cho kết quả tốt nhất với dự báo bất thường trong khoảng thời gian 130 ngày. Nghiên cứu này chủ yếu so sánh các phương pháp dự báo chuỗi thời gian bất thường nhưng ở dạng đơn biến, mà chưa xét đến khả năng kết hợp dữ liệu ở dạng đa biến, cũng như khai thác mối quan hệ giữa các biến.

Nghiên cứu [18] đã chỉ ra rằng dự báo chuỗi thời gian gặp nhiều khó khăn với dữ liệu có tính không đều và bất thường. Các tác giả đã đề xuất sử dụng Support Vector Machines (SVM) và Ensemble Empirical Mode Decomposition (EEMD) [19] để dự báo dạng dữ liệu bất thường này. Dữ liệu chuỗi thời gian được phân tách thành các chuỗi con liên tục bằng kỹ thuật EEMD, SVM được áp dụng để lập mô hình cho từng chuỗi con. Cuối cùng, dự báo của các chuỗi con được tổng hợp lại để tạo ra dự báo tổng hợp. Kết quả thực nghiệm trên hai bộ dữ liệu cho thấy kết hợp SVM- EEMD vượt trội hơn so với SVM, ARIMA và Croston về các chỉ số RMSE, MAPE, MDRAE và MASE.

Trong dự báo nhu cầu mạng không dây, nghiên cứu trong [20] đề xuất mô hình dự báo nhu cầu sử dụng mạng không dây với các giá trị bất thường về lưu lượng do các yếu tố như nhiễu, tính di động, yêu cầu kết nối, nhằm nâng cao hiệu quả năng lượng và tránh mất kết nối mạng. Nghiên cứu này đã đề xuất một nhân nhúng đặc trưng (feature embedding (FE) kernel) cho Gaussian Process để dự báo nhu cầu lưu lượng với các giá trị cực trị. Kết quả thực nghiệm cho thấy kết hợp FE-GP giảm 32% so với S-ARIMA và giảm 17% so với Naive-GP. Đối với dự báo giá trị trung bình dài hạn, mô hình FE-GP có mức giảm 21% so với S-ARIMA và 12% so với Naive-GP.

Nghiên cứu [21] quan tâm đến nhu cầu sản phẩm hoặc dịch vụ mà những biến động do nhiều yếu tố mang tính không chắc chắn và bất thường sẽ gây khó khăn cho việc dự báo và quản lý chuỗi cung ứng. Trong nghiên cứu này, nhóm tác giả đã điều tra 843 chuỗi thời gian với các giá trị khác nhau của hệ số biến thiên (CoV - Coefficient of variations) và cho thấy khuyến mãi gây ra sự biến động trên toàn bộ chuỗi nhu cầu. Các tác giả đã đề xuất mô hình kết hợp giữa phân tách nhu cầu cơ bản và nhu cầu khuyến mại, từ đó tạo ra dự báo chính xác hơn trên các chuỗi biến động.

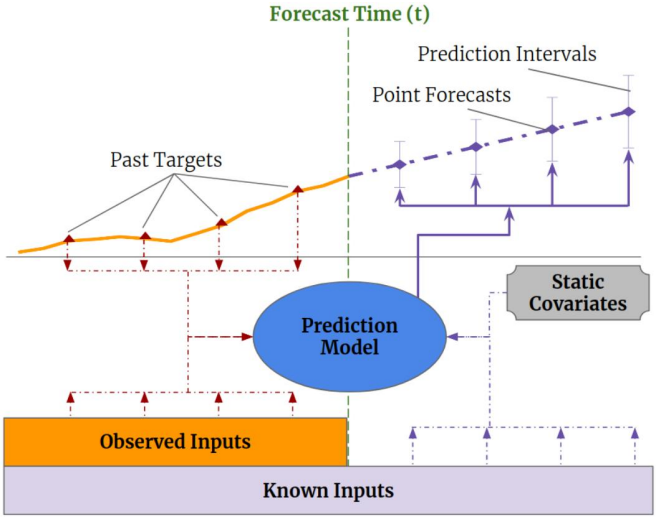
III. MÔ HÌNH TÍCH HỢP TFT-TPE

A. Temporal fusion transformer

Temporal Fusion Transformer (TFT) là một mô hình dựa trên Transformer được sử dụng để dự báo chuỗi thời gian đa chiều và có khả năng giải thích cao hơn so với các mô hình "hộp đen" thông thường. TFT kết hợp việc dự báo hiệu suất cao với khả năng giải thích, đặc biệt hữu ích trong việc dự báo các vấn đề phức tạp như chuỗi thời gian với sự biến đổi đồng nhất, dữ liệu từ tương lai đã được biết trước và các biến thời gian mà chỉ được quan sát trong quá khứ.

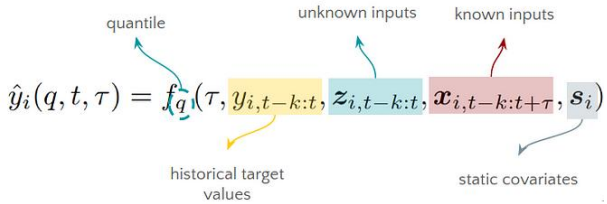
TFT tận dụng các cơ chế tự chú ý để nắm bắt các mối phụ thuộc phức tạp về thời gian trong dữ liệu. Để nắm bắt bản chất thời gian của dữ liệu, TFT đã sử dụng các kỹ thuật mã hóa thời gian khác nhau cho phép mô hình hóa sự kết hợp thông tin liên quan đến thời gian vào chuỗi đầu vào để nắm bắt tốt hơn các mô hình và xu hướng thời gian. TFT phân biệt bốn nhóm đặc trưng đầu vào (Hình 1):

- Nhóm đặc trưng 1 bao gồm các thông tin đã xảy ra trong quá khứ như doanh số bán hàng trước đó, số lượt khách du lịch,... Các thông tin này thay đổi theo thời gian và chỉ có thể biết chính xác vào thời điểm xem xét (t).
- Nhóm đặc trưng 2 là các thông tin khác có ảnh hưởng đến việc dự báo như thời tiết, chỉ số tiêu dùng,… Các thông tin này cũng thay đổi theo thời gian và chỉ có thể biết chính xác vào thời điểm xem xét (t).
- Nhóm đặc trưng 3 bao gồm các yếu tố cố định như ngày trong tuần, tuần trong tháng, ngày lễ,… Những thông tin này không thay đổi theo thời gian và có thể biết trước.
- Nhóm đặc trưng 4 là các thông tin không liên quan đến thời gian như định danh cửa hàng, khu vực địa lý,… Những thông tin này không thay đổi theo thời gian.



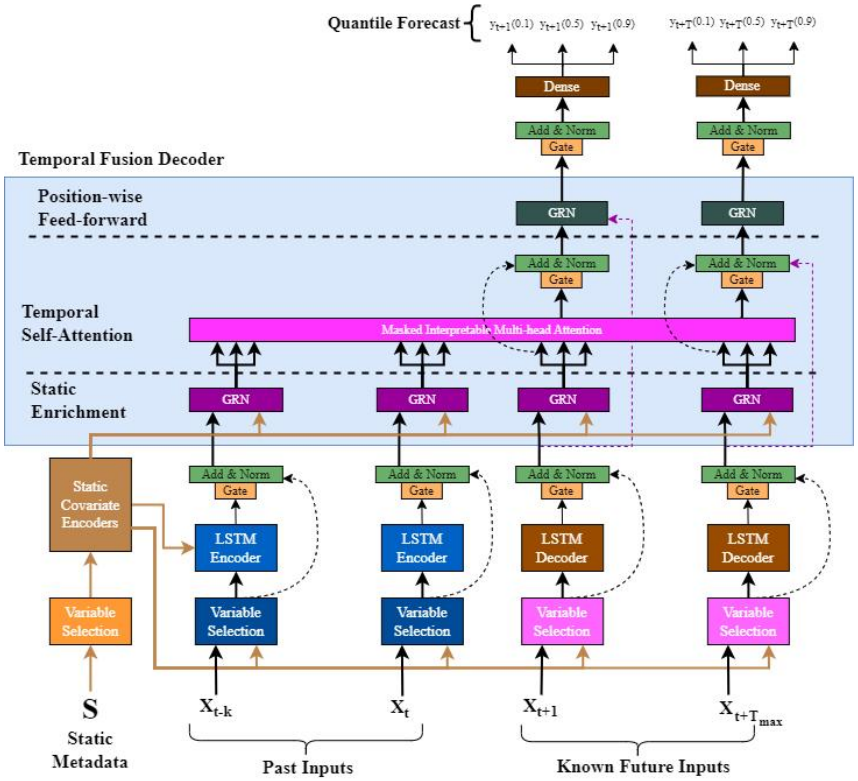
Hình 1. Minh họa dự báo đa biến (nguồn [5])

Khả năng giải thích mô hình là một đặc điểm nổi bật của TFT. Nó cung cấp thông tin chi tiết về cách đưa ra dự báo và cho thấy những bước thời gian nào trong quá khứ nào có ảnh hưởng nhất đến việc dự báo giá trị tương lai. TFT áp dụng các cơ chế chú ý, giúp làm nổi bật tầm quan trọng của các tính năng đầu vào và các yếu tố thời gian khác nhau. Cách tiếp cận này làm cho quá trình dự báo trở nên dễ hiểu và minh bạch hơn đối với người dùng. Thay vì chỉ đưa ra một giá trị dự báo duy nhất, TFT cung cấp các khoảng dự báo thông qua phân vị. Mỗi phân vị q dự báo cho τ-step-forward tại thời điểm t. τ-step-forward được hiểu là giá trị dự báo tại một thời điểm cụ thể trong tương lai, cách xa thời điểm hiện tại một khoảng thời gian τ. Đây là một thuật ngữ trong dự báo chuỗi thời gian, biểu thị việc dự báo giá trị tại bước thứ τ từ thời điểm hiện tại. Chẳng hạn, nếu τ = 3, thì 3-step-forward sẽ là giá trị dự báo sau ba bước thời gian từ hiện tại, như Hình 2.



Hình 2. Minh họa dự báo 3 bước (nguồn [5])

TFT sử dụng các lớp tuần tự để xử lý thông tin từ các đầu vào đã biết và đã được quan sát, cùng với các bộ mã hóa đồng nhất để biểu diễn các thông tin ngữ cảnh. Các lớp tuần tự này sau đó chọn lọc các biến phụ thuộc mẫu để loại bỏ thông tin không liên quan. TFT cũng có khả năng tự chú ý theo thời gian, giúp mô hình hiểu được mối quan hệ dài hạn trong dữ liệu. Hình 3 minh họa cấu trúc của mô hình TFT.



Hình 3. Mô hình cấu trúc TFT (nguồn [5])

TFT có khả năng tạo ra các biểu diễn đặc trưng cho từng loại đầu vào một cách hiệu quả bằng cách sử dụng các thành phần chuẩn, từ đó nâng cao hiệu suất dự báo cho nhiều nhiệm vụ khác nhau. TFT bao gồm năm thành phần cơ bản:

1) Các cổng (Gates): Được thiết kế để bỏ qua tất cả các yếu tố kiến trúc không được sử dụng, đồng thời cung cấp độ sâu và độ phức tạp có thể điều chỉnh của mạng, giúp thích ứng với các bộ dữ liệu và điều kiện khác nhau.

2) Mạng chọn biến (Variable Selection): Lựa chọn các biến đầu vào tương ứng tại mỗi bước thời gian.

3) Bộ mã hóa đồng biến tĩnh (Static Covariate Encoders): Tích hợp các đặc tính tĩnh vào mạng, giúp xử lý và mã hóa các đặc điểm không thay đổi theo thời gian.

4) Bộ giải mã hợp nhất thời gian (Temporal Fusion Decoder): Được thiết kế để giải mã thông tin đã được kết hợp và mã hóa từ các nguồn dữ liệu khác nhau nhằm tạo ra các dự báo chính xác theo thời gian. Các lớp tuần tự được sử dụng để xử lý cục bộ, trong khi các liên kết dài hạn dựa vào khả năng thu thập của cơ chế chú ý đa đầu mới có thể giải thích được.

5) Dự báo phân vị (Quantile Forecasts): Thay vì chỉ dự báo một giá trị trung bình duy nhất cho một thời điểm tương lai, dự báo phân vị cung cấp một phạm vi giá trị dự báo với các mức xác suất khác nhau. Điều này giúp thể hiện sự bất định trong dự báo và cung cấp một cái nhìn toàn diện hơn về các khả năng có thể xảy ra.

B. Tree-structured parzen estimator

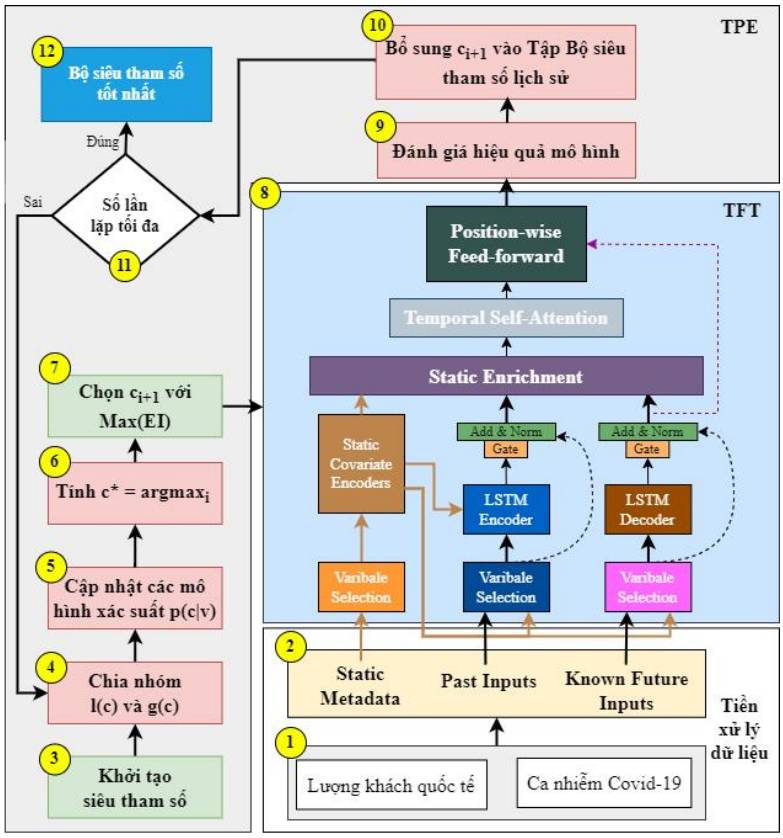
Với một số bài toán, như tìm kiếm kiến trúc TFT tốt nhất, hàm mục tiêu $f(x)$ là quá phức tạp để biểu diễn tường minh hoặc không thể phân tích được. Hàm $f(x)$ khi đó có thể được xem là hộp đen. Việc áp dụng tìm kiếm lưới hay tìm kiếm ngẫu nhiên [3] có thể gặp phải vấn đề về các bộ giá trị siêu tham số quá nhiều mất chi phí thời gian. Còn đối với phương pháp thủ công, việc chọn bộ giá trị siêu tham số phụ thuộc quá nhiều vào kinh nghiệm của người thực hiện huấn luyện. Vì vậy, TPE là phù hợp để tìm kiếm tự động các giá trị siêu tham số, giảm số lần đánh giá trên mỗi bộ siêu tham số và giảm tiêu tốn thời gian cho việc đánh giá này. Thực tế, cách thực hiện của TPE giúp nhảy ra khỏi các vị trí tối ưu hóa cục bộ.

TPE thường được sử dụng trong các bài toán có không gian siêu tham số lớn hoặc không gian đa chiều, nơi các phương pháp tìm kiếm thông thường có thể trở nên không hiệu quả. Bằng cách sử dụng một mô hình ước lượng phân phối để hướng dẫn quá trình tìm kiếm, TPE có thể tìm ra các bộ siêu tham số tốt một cách hiệu quả.

C. Tích hợp TFT-TPE

Tích hợp TPE vào TFT là nhằm tận dụng sức mạnh của TPE để tối ưu hóa bộ siêu tham số trong TFT. Cụ thể, quá trình này bắt đầu bằng việc sử dụng TFT để biểu diễn một phần không gian siêu tham số. TFT có khả năng biểu diễn các mối quan hệ phức tạp giữa các tham số, đặc biệt là trong các mô hình có dữ liệu dạng chuỗi hoặc không gian. Trong nghiên cứu

này, chúng tôi sử dụng TFT để tạo ra một mô hình cho hàm mục tiêu. Cập nhật tham số và kết quả của mỗi lượt là quá trình tinh chỉnh thông qua TPE để tạo ra một mô hình ổn định hơn.



Hình 4. Mô hình tích hợp TFT-TPE

Tiến trình thực hiện của mô hình TFT-TPE được thực hiện qua 12 bước (Hình 4).

Bước 1. Thu thập dữ liệu: Thu thập các chuỗi thời gian và những dữ liệu liên quan khác có thể ảnh hưởng đến kết quả dự báo.

Bước 2. Xử lý dữ liệu và tạo đặc trưng: Xử lý dữ liệu để chuẩn hóa và làm sạch dữ liệu. Xác định các đặc trưng quan trọng từ dữ liệu gốc bao gồm cả các đồng biến từ quá khứ và tương lai.

Bước 3. Khởi tạo không gian tìm kiếm siêu tham số: Thiết lập phạm vi và giá trị ban đầu cho các siêu tham số cần tối ưu. Điều này giúp định rõ giới hạn không gian tìm kiếm siêu tham số. Khởi tạo ngẫu nhiên một số bộ giá trị siêu tham số ban đầu và đánh giá hiệu suất của chúng.

Bước 4. Chia nhóm l(c) và g(c): Dựa trên kết quả quan sát, chia dữ liệu thành hai nhóm: nhóm hiệu suất tốt nhất $l(c)$ và nhóm còn lại $g(c)$.

Bước 5. Cập nhật các mô hình xác suất p(c|v): Xây dựng hoặc cập nhật các mô hình xác suất $l(c)$ và $g(c)$ dựa trên các quan sát.

Bước 6. Tính $c^* = argmax_i$: Tìm bộ giá trị siêu tham số c^* tối đa hóa tỷ lệ $\frac{l(c)}{g(c)}$.

Bước 7. Chọn c_{i+1} với Max(EI): Chọn bộ giá trị siêu tham số mới c_{i+1} tối đa hóa hàm Expectation Improvement (EI).

Bước 8. Huấn luyện mô hình với bộ giá trị siêu tham số mới c_{i+1} .

Bước 9. Đánh giá hiệu quả mô hình: Đánh giá hiệu suất của mô hình với c_{i+1} .

Bước 10. Bổ sung c_{i+1} vào tập bộ siêu tham số lịch sử.

Bước 11. Lặp lại Bước 2 đến khi số lần lặp tối đa.

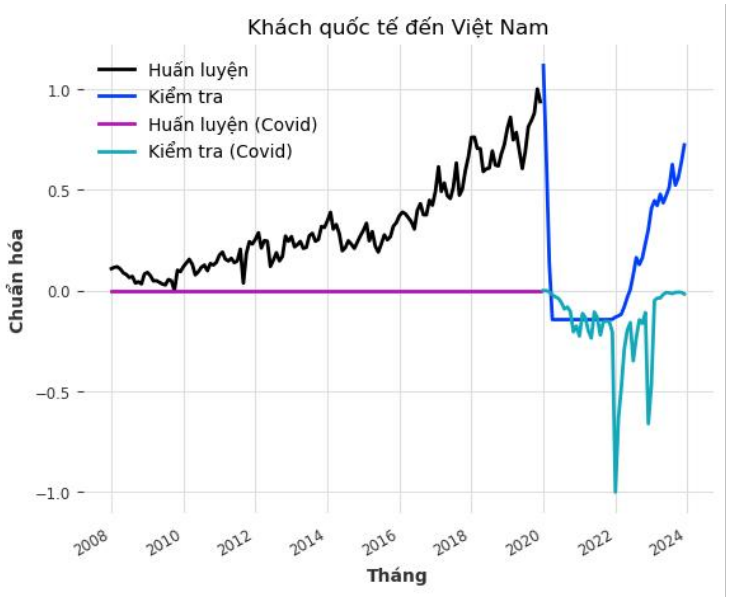
Bước 12. Chọn bộ giá trị siêu tham số có hiệu suất tốt nhất từ tập dữ liệu lịch sử.

Mô hình tích hợp TFT và TPE mang lại những lợi ích đáng kể từ cả hai phương pháp. Cụ thể, TFT cho phép biểu diễn các mối quan hệ phức tạp giữa các tham số, trong khi TPE cung cấp khả năng tìm kiếm hiệu quả trong không gian siêu tham số. Sự kết hợp này không chỉ tăng cường tính hiệu quả của quá trình tối ưu hóa siêu tham số mà còn giúp đạt được hiệu suất cao hơn trong thời gian ngắn hơn. Kết quả này cho thấy tiềm năng vượt trội của phương pháp tích hợp trong việc cải thiện hiệu quả và tốc độ của các mô hình dự báo.

IV. THỬ NGHIỆM VÀ PHÂN TÍCH KẾT QUẢ

A. Xử lý dữ liệu

Chúng tôi thu thập dữ liệu hàng tháng, từ 2008 đến 2023, về lượt khách quốc tế đến Việt Nam từ website của Cục du lịch quốc gia Việt Nam [22] và dữ liệu về các ca nhiễm mới Covid-19 theo ngày của Viet Nam [23]. Dữ liệu sau đó được tiền xử lý và chuẩn hóa để làm đầu vào cho mô hình TFT-TPE. Dữ liệu chuẩn hóa được chia thành hai phần chính: tập huấn luyện và tập kiểm tra, trong đó dữ liệu gần đây hơn được sử dụng để làm tập kiểm tra, trong khi dữ liệu cũ hơn được sử dụng để làm tập huấn luyện. Hình 4 mô tả dữ liệu hành tháng của lượt khách quốc tế đến Việt Nam và trường hợp nhiễm Covid-19 mới. Dữ liệu này được chuẩn hóa và dữ liệu ca nhiễm mới được đảo ngược giá trị thành giá trị âm để mô hình hiểu dự báo nhu cầu du lịch giảm khi biến về ca nhiễm Covid-19 mới giảm.



Hình 5. Dữ liệu hàng tháng về lượng khách du lịch và số ca nhiễm Covid-19 (nguồn [22, 23])

B. So sánh mô hình TFT-TPE trước và sau khi tích hợp dữ liệu Covid-19

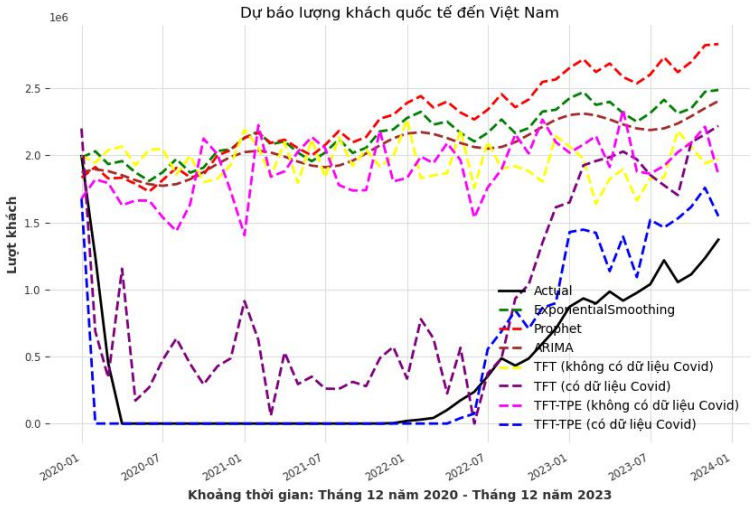
Việc phân tích và so sánh mô hình TFT-TPE trước và sau khi tích hợp dữ liệu Covid-19 là quá trình quan trọng để đánh giá tác động của dữ liệu này đến hiệu suất dự báo. Việc tích hợp dữ liệu Covid-19 giúp mô hình dự báo chính xác hơn, cho thấy các biến số Covid-19 có ảnh hưởng đáng kể và đã được mô hình nắm bắt tốt. Mô hình có thể dự báo chính xác hơn trong các khoảng thời gian ngắn sau khi tích hợp dữ liệu Covid-19, do mô hình đã học được các mẫu thay đổi do đại dịch.

Để đánh giá các mô hình huấn luyện, tiêu chí được sử dụng phổ biến là MAPE. Trong cài đặt này, chúng tôi so sánh mô hình TFT-TPE trước và sau khi tích hợp dữ liệu dịch Covid-19. Bộ giá trị siêu tham số được sử dụng trong cài đặt và không gian tìm kiếm được thiết lập như Bảng 2. Kết quả đánh giá được thể hiện trong Hình 6.

Bảng 1. Thông số mô phỏng

Tham số	Giá trị
input_chunk_length	24
output_chunk_length	6
hidden_size	77
lstm_layers	1
num_attention_heads	8
dropout	0.4202650976515916
batch_size	41
n_epochs	181
random_state	24

Việc tích hợp TPE vào TFT và sử dụng dữ liệu Covid-19 đã mang lại cải thiện đáng kể về độ chính xác trong dự báo nhu cầu du lịch, đặc biệt là trong giai đoạn xảy ra đại dịch. Khi so sánh biểu đồ dự báo nhu cầu du lịch không có và có dữ liệu Covid-19 cũng như mô hình TFT có tích hợp TPE và không có, chúng ta nhận thấy sự khác biệt rõ rệt. Theo Hình 5 không có dữ liệu Covid-19 cho thấy sự không chính xác rõ ràng, với khoảng cách lớn giữa đường dự báo và đường thực tế, đặc biệt trong các giai đoạn biến động mạnh do đại dịch. Mô hình TFT không có TPE cũng cho thấy dự báo không tốt bằng mô hình TFT-TPE.



Hình 6. So sánh các mô hình dự báo

Trong Hình 6, các mô hình khác TFT-TPE không phản ánh được thay đổi đột ngột trong nhu cầu du lịch khi Covid-19 bùng phát, dẫn đến dự báo sai lệch và không sát thực tế. Ngược lại, TFT-TPE với dữ liệu Covid-19 phản ánh chính xác hơn các biến động lớn, giảm mạnh trong nhu cầu du lịch ở các thời điểm dịch bùng phát và sự phục hồi sau đó. Khoảng cách giữa đường dự báo và thực tế với MAPE khoảng 36%, cho thấy mô hình có dữ liệu Covid-19 có độ chính xác cao hơn và phù hợp hơn với thực tế. Điều này chứng minh tầm quan trọng của việc cập nhật mô hình dự báo với các dữ liệu mới và quan trọng để duy trì độ chính xác và độ tin cậy trong bối cảnh thế giới thay đổi nhanh chóng.

Bảng 2. So sánh các mô hình theo chỉ số MAPE, MAE và RMSE

Mô hình	MAPE (%)	MAE	RMSE
ExponentialSmoothing	7.922592e+07	1.746187e+06	1.799656e+06
Prophet	7.886474e+07	1.860717e+06	1.902338e+06
ARIMA	7.580761e+07	1.654510e+06	1.708026e+06
TFT (không có dữ liệu Covid)	7.818912e+07	1.538437e+06	1.633344e+06
TFT (có dữ liệu Covid)	1.715005e+07	5.419773e+05	6.253479e+05
TFT-TPE (không có dữ liệu Covid)	7.136709e+07	1.502043e+06	1.573052e+06
TFT-TPE (có dữ liệu Covid)	**3.627463e+01**	**1.820286e+05**	**3.061875e+05**

Kết quả dự báo cho thấy TFT-TPE có độ chính xác cao hơn các mô hình khác và cả mô hình TFT. Mô hình TFT-TPE có sử dụng dữ liệu dịch Covid-19 dự báo chính xác hơn khi không có dữ liệu này. Như vậy, các mô hình dự báo phụ thuộc vào tính chất của dữ liệu đầu vào, khi dữ liệu không có quy luật và bất thường như lượng khách quốc tế trong giai đoạn Covid-19 càng khó dự báo và khó dự báo cho nhu cầu du lịch sau đại dịch. TPE-TFT đã thể hiện được khả năng vượt qua các trở ngại này để có thể dự báo chính xác đối với tập dữ liệu nhu cầu du lịch Việt Nam.

V. KẾT LUẬN

Bài báo đã đề xuất việc kết hợp TPE và TFT để thực hiện việc tìm kiếm tự động bộ giá trị siêu tham số tốt nhất. Bài báo đã triển khai thuật toán TFT-TPE với bộ dữ liệu thu thập được về lượt khách quốc tế đến Việt Nam từ năm 2004 đến 2023 và dữ liệu các ca mắc Covid-19 mới. Việc thu thập và chuẩn hóa dữ liệu cho quá trình huấn luyện cũng được mô tả. Kết quả thực nghiệm so sánh việc dự báo nhu cầu du lịch tại thời điểm bất thường cần có dữ liệu đồng biến liên quan đến chuỗi thời gian bất thường sẽ có kết quả tốt hơn.

Lời cảm ơn: *Bài báo được hoàn thành dưới sự hỗ trợ từ Đề tài cấp Đại học Huế: DHH2024-10-38 của Trường Du lịch - Đại học Huế.*

TÀI LIỆU THAM KHẢO

[1] V. Kotu and B. Deshpande, "Time Series Forecasting," *Data Science*, vol. 33, no. vi, Elsevier, pp. 395-445, 2019.

[2] S. Schmidl, P. Wenig, and T. Papenbrock, "Anomaly Detection in Time Series: A Comprehensive Evaluation," *Proc. VLDB Endow.*, vol. 15, no. 9, pp. 1779-1797, 2022, Doi: 10.14778/3538598.3538602.

[3] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, pp. 281-305, 2012.

[4] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," *Adv. Neural Inf. Process. Syst. 24 - 25th Annu. Conf. Neural Inf. Process. Syst. 2011, NIPS 2011*, pp. 1-9, 2011.

[5] B. Lim, S. Arık, N. Loeff, and T. Pfister, "Temporal Fusion Transformers for interpretable multi-horizon time series forecasting," *Int. J. Forecast.*, vol. 37, no. 4, pp. 1748-1764, 2021, Doi: 10.1016/j.ijforecast.2021.03.012.

[6] P. Liu, W. Ming, and C. Huang, "Intelligent modeling of abnormal demand forecasting for medical consumables in smart city," *Environ. Technol. Innov.*, vol. 20, p. 101069, 2020, Doi: 10.1016/j.eti.2020.101069.

[7] D. Kalifa, U. Singer, I. Guy, G. D. Rosin, and K. Radinsky, "Leveraging World Events to Predict E-Commerce Consumer Demand under Anomaly," in *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, Feb. 2022, vol. 1, no. 1, pp. 430-438, Doi: 10.1145/3488560.3498452.

[8] Z. Z. OO and S. PHYU, "Time Series Prediction Based on Facebook Prophet: A Case Study, Temperature Forecasting in Myintkyina," *Int. J. Appl. Math. Electron. Comput.*, vol. 8, no. 4, pp. 263-267, 2020, Doi: 10.18100/ijamec.816894.

[9] K. Sharma, R. Bhalla, and G. Ganesan, "Time Series Forecasting Using FB-Prophet," *CEUR Workshop Proc.*, vol. 3445, pp. 59-65, 2022.

[10] R. El-Hadad, Y. F. Tan, and W. N. Tan, "Anomaly Prediction in Electricity Consumption Using a Combination of Machine Learning Techniques," *Int. J. Technol.*, vol. 13, no. 6, pp. 1317-1325, 2022, Doi: 10.14716/ijtech.v13i6.5931.

[11] F. T. Liu, K. M. Ting, and Z. H. Zhou, "Isolation forest," *Proc. IEEE Int. Conf. Data Mining, ICDM*, pp. 413-422, 2008, Doi: 10.1109/ICDM.2008.17.

[12] A. Cutler, D. R. Cutler, and J. R. Stevens, "Random Forests," in *Ensemble Machine Learning*, C. Zhang and Y. Ma, Eds. New York, NY: Springer New York, 2012, pp. 157-175.

[13] X. Luochen and N. Hasachoo, "The Study of Irregular Demand Forecasting for Medicines: The Case Study of ABC Medical Center Hospital," *Proc. 2021 10th Int. Conf. Ind. Technol. Manag. ICITM 2021*, pp. 115-120, 2021, Doi: 10.1109/ICITM52822.2021.00028.

[14] J. D. Croston, "Forecasting and Stock Control for Intermittent Demands," *J. Oper. Res. Soc.*, vol. 23, no. 3, pp. 289-303, 1972, Doi: 10.1057/JORS.1972.50.

[15] R. H. Teunter, A. A. Syntetos, and M. Z. Babai, "Intermittent demand: Linking forecasting to inventory obsolescence," *Eur. J. Oper. Res.*, vol. 214, no. 3, pp. 606-615, 2011, Doi: 10.1016/j.ejor.2011.05.018.

[16] M. Z. Babai, A. Syntetos, and R. Teunter, "Intermittent demand forecasting: An empirical study on accuracy and the risk of obsolescence," *Int. J. Prod. Econ.*, vol. 157, no. 1, pp. 212-219, Nov. 2014, Doi: 10.1016/J.IJPE.2014.08.019.

[17] P. Kalaya, P. Termsuksawad, and T. Wasusri, "Forecasting Lumpy Demand for Planning Inventory: The Case of Community Hospitals in Thailand," *IEEE Int. Conf. Ind. Eng. Eng. Manag.*, pp. 686-690, Dec. 2019, Doi: 10.1109/IEEM44572.2019.8978541.

[18] Y. Bao, R. Zhang, T. Xiong, and Z. Hu, "Forecasting non-normal demand by support vector machines with ensemble empirical mode decomposition," *Adv. Inf. Sci. Serv. Sci.*, vol. 3, no. 3, pp. 81-91, 2011, Doi: 10.4156/aiss.vol3.issue3.11.

[19] Z. Wu and N. E. Huang, "Ensemble empirical mode decomposition: A noise-assisted data analysis method," *Adv. Adapt. Data Anal.*, vol. 01, no. 01, pp. 1-41, Jan. 2009, Doi: 10.1142/S1793536909000047.

[20] S. C. Sun and W. Guo, "Forecasting Wireless Demand with Extreme Values using Feature Embedding in Gaussian Processes," *IEEE Veh. Technol. Conf.*, vol. 2021-April, pp. 1-8, 2021, Doi: 10.1109/VTC2021-Spring51267.2021.9449040.

[21] M. Abolghasemi, E. Beh, G. Tarr, and R. Gerlach, "Demand forecasting in supply chain: The impact of demand volatility in the presence of promotion," *Comput. Ind. Eng.*, vol. 142, no. July 2019, p. 106380, 2020, Doi: 10.1016/j.cie.2020.106380.

[22] "Cục Du lịch Quốc gia Việt Nam." Accessed: May 18, 2024. [Online]. Available: https://vietnamtourism.gov.vn/statistic/international.

[23] Vietnam: Coronavirus Pandemic Country Profile - Our World in Data." Accessed: May 18, 2024. [Online]. Available: https://ourworldindata.org/coronavirus/country/vietnam?country=~VNM.

TFT-TPE INTEGRATED MODEL FOR FORECASTING ABNORMAL DATA OF VIETNAM TOURISM DEMAND DURING COVID-19

Ngo Van Son, Thai Thi Phuong, Nguyen Duong Thien, Vo Viet Minh Nhat

ABSTRACT: *Forecasting plays a vital role in production and business. However, abnormal and unpredictable events, such as the Covid-19 pandemic, have made accurate forecasting challenging. In these cases, identifying an accurate forecasting method that can adapt to unusual fluctuations is an urgent requirement. This article proposes a model integrating Temporal Fusion Transformer (TFT) and Tree-structured Parzen Estimator (TPE) to forecast abnormal events accurately. TFT is a deep learning neural network model designed explicitly for processing time series data, which can capture trends and model complex variations in data. With TPE, it is an optimization method based on a tree-like data structure to determine the best set of hyperparameters. Integrating TFT and TPE can effectively solve the problem of predicting abnormal data. The TFT-TPE model is evaluated by a Vietnamese tourism demand data set during Covid-19. Experiments show that the TFT-TPE model can adapt and achieve accurate forecasting results for abnormal data when integrated with Covid-19 data..*