

PHÂN TÍCH ẢNH HƯỞNG CỦA DỮ LIỆU ĐA NGUỒN TRONG DỰ BÁO CHUỖI THỜI GIAN BẰNG MẠNG MLP

Ngô Văn Sơn^{1,2}, Võ Viết Minh Nhật³

¹Trường Đại học Khoa học, Đại học Huế

²Trường Du lịch, Đại học Huế

³Viện Khảo thí và Bảo đảm chất lượng giáo dục, Đại học Huế

Email: ngovanon@hueuni.edu.vn, vvmnhat@hueuni.edu.vn

Ngày nhận bài: 8/12/2024; ngày hoàn thành phản biện: 20/12/2024; ngày duyệt đăng: 20/12/2024

TÓM TẮT

Bài báo nghiên cứu hiệu quả của việc tích hợp dữ liệu đa nguồn trong dự báo chuỗi thời gian bằng mô hình mạng nơ-ron đa lớp (MLP), ứng dụng vào dự báo lượt khách quốc tế đến Việt Nam theo tháng. Bên cạnh dữ liệu quá khứ, mô hình còn sử dụng dữ liệu tìm kiếm từ Google Trends và thông tin ngày lễ quốc gia. Kết quả cho thấy dữ liệu ngoại sinh giúp cải thiện độ chính xác, đặc biệt trong dự báo ngắn hạn ($h=1$), với sMAPE giảm rõ rệt khi kết hợp đầy đủ các nguồn dữ liệu. Tuy nhiên, tác động của các nguồn dữ liệu giảm dần theo độ dài dự báo và mức độ hiệu quả giữa các nguồn không đồng đều; dữ liệu ngày lễ có tác động ổn định hơn so với tìm kiếm trực tuyến. Nghiên cứu khẳng định tiềm năng của dữ liệu đa nguồn trong nâng cao chất lượng dự báo và chỉ ra hạn chế của MLP khi độ dài dự báo tăng lên, từ đó đề xuất hướng tiếp cận các mô hình học sâu tiên tiến hơn.

Từ khóa: Du lịch Việt Nam, dự báo nhu cầu du lịch, dữ liệu công cụ tìm kiếm, Google Trends.

1. MỞ ĐẦU

Dự báo chuỗi thời gian là một chủ đề nhận được nhiều sự quan tâm trong lĩnh vực học máy và trí tuệ nhân tạo. Phần lớn các nghiên cứu hiện nay tập trung vào việc khai thác một chuỗi dữ liệu đơn lẻ. Tuy nhiên, trong môi trường thực tế, hệ thống thường thu nhận dữ liệu từ nhiều cảm biến, hệ thống quan trắc hoặc nguồn dữ liệu thứ cấp khác. Điều này đặt ra thách thức mới trong việc xử lý, tích hợp và học biểu diễn dữ liệu từ nhiều nguồn để phục vụ dự báo hiệu quả. Trong nghiên cứu này, chúng tôi khai thác lợi thế của mô hình mạng nơ-ron đa lớp (MLP - Multi-layer Perceptron) trong việc học mối quan hệ phi tuyến giữa các đặc trưng đầu vào đến giá trị đầu ra của chuỗi thời gian, từ

đó đề xuất một mô hình có khả năng tích hợp dữ liệu từ nhiều nguồn để cải thiện chất lượng dự báo.

Một trong các nguồn dữ liệu có thể khai thác để sử dụng trong dự báo nhu cầu du lịch là xu hướng tìm kiếm về du lịch trên công cụ tìm kiếm như Google. Ngày nay, nhiều hoạt động kinh doanh đã chuyển từ các giao dịch vật lý sang môi trường trực tuyến với nền tảng web [1]. Sự phát triển của công nghệ thông tin đã tạo ra một lượng lớn dữ liệu lớn từ người dùng, như các bình luận, bình chọn, các truy vấn tìm kiếm hay các bài phân tích. Đối với các doanh nghiệp, dữ liệu lớn có thể có ý nghĩa quan trọng trong việc hỗ trợ người quản lý ra quyết định. Các nhà quản lý đã bắt đầu nhận thức được tầm quan trọng và giá trị của "dữ liệu lớn" và một hướng nghiên cứu liên quan đến "dữ liệu công cụ tìm kiếm" cũng đã xuất hiện. Thực tế, đã có khá nhiều nghiên cứu về khai thác giá trị của dữ liệu công cụ tìm kiếm xuất hiện trong những năm trở lại đây [2][3][4]. Google và Baidu là hai công cụ tìm kiếm cung cấp dữ liệu tìm kiếm lớn nhất thế giới. Baidu chủ yếu phổ biến tại Trung Quốc, trong khi Google được sử dụng khắp toàn cầu. Với khả năng cung cấp dữ liệu có sẵn và dễ truy cập, Google Trends trở thành mục tiêu để nghiên cứu dữ liệu công cụ tìm kiếm. Trong bài báo này, chúng tôi thu thập dữ liệu nhu cầu du lịch (khách du lịch quốc tế đến) Việt Nam từ Google Trends [5].

Du lịch là một ngành công nghiệp không khói, đóng vai trò quan trọng trong nền kinh tế của nhiều địa phương và quốc gia cả trực tiếp và gián tiếp bằng cách cung cấp việc làm và nguồn thu ngoại tệ. Điều này giải thích tại sao rất nhiều quốc gia nhỏ phụ thuộc nhiều vào du lịch. Dự báo chính xác nhu cầu du lịch do đó là một nhu cầu cấp thiết đối với các quốc gia và địa phương nhằm có thể xây dựng chính sách, hoạch định kế hoạch, chuẩn bị nguồn lực nhằm cung cấp tốt nhất các sản phẩm và dịch vụ cho khách du lịch. Các mô hình dự báo truyền thống dựa trên dữ liệu lịch sử khách du lịch đến không còn đảm bảo độ chính xác và xu hướng kết hợp với những loại dữ liệu khác, chẳng hạn dữ liệu công cụ tìm kiếm, trở thành xu hướng mới cho dự báo du lịch trong kỷ nguyên dữ liệu lớn [6][7].

Nghiên cứu trong bài báo này có mục đích khám phá việc kết hợp dữ liệu công cụ tìm kiếm với dữ liệu lịch sử khách du lịch đến tại một điểm đến nhằm nâng cao hiệu quả dự báo. Google Trends được sử dụng để thu thập dữ liệu công cụ tìm kiếm, kết hợp với dữ liệu lịch sử khách du lịch quốc tế đến ở hai điểm đến Việt Nam. Kết quả dự báo sau đó so sánh với dữ liệu du lịch từ Cục Du lịch Quốc gia Việt Nam [8].

Trong bối cảnh có nhiều mô hình học sâu mạnh mẽ như LSTM, GRU hay Transformer đã được áp dụng hiệu quả trong dự báo chuỗi thời gian, nghiên cứu này lựa chọn mạng no-ron truyền thẳng nhiều lớp (MLP) với mục đích làm bước thử nghiệm ban đầu cho việc tích hợp dữ liệu đa nguồn trong một kiến trúc đơn giản, dễ huấn luyện và ít yêu cầu về tài nguyên tính toán. MLP không có cấu trúc tuần tự như các mô hình hồi tiếp (RNN), nhưng có thể học được mối quan hệ phi tuyến giữa các đặc trưng đầu

vào và đầu ra nếu đầu vào được xây dựng hợp lý. Lựa chọn này cho phép chúng tôi tập trung đánh giá tác động của từng nguồn dữ liệu (lượng khách đến, tìm kiếm, ngày lễ) đến chất lượng dự báo trong một môi trường kiểm soát, từ đó cung cấp nền tảng thực nghiệm và cơ sở so sánh cho các nghiên cứu mở rộng sau này với các kiến trúc mạng sâu hơn. Các đóng góp của bài báo gồm:

- Bài báo đề xuất một mô hình dự báo chuỗi thời gian với khả năng tích hợp linh hoạt dữ liệu từ nhiều nguồn khác nhau với mô hình MLP.
- Bài báo chỉ ra rằng không phải mọi loại nguồn dữ liệu đều mang lại hiệu quả dự báo như nhau; trong đó dữ liệu ngày lễ cho thấy đóng góp ổn định hơn so với dữ liệu tìm kiếm trực tuyến.
- Thực nghiệm với nhiều độ dài dự báo (1–6 tháng) giúp làm rõ hiệu quả của việc thêm dữ liệu đa nguồn giảm dần theo thời gian dự báo khi sử dụng mô hình MLP.
- Trên cơ sở các kết quả thực nghiệm, bài báo đề xuất việc mở rộng nghiên cứu với các mô hình sâu hơn như LSTM, GRU hoặc Transformer để khai thác tốt hơn mối quan hệ chuỗi dài hạn khi sử dụng dữ liệu đa nguồn.

Nội dung tiếp theo của bài báo gồm: phần 2 phân tích các nghiên cứu liên quan đến dự báo du lịch. Phần 3 mô tả việc thu thập dữ liệu, xử lý dữ liệu và đề xuất các mô hình dự báo đa nguồn dữ liệu. Phần 4 so sánh các kết quả dự báo thực nghiệm khác nhau của mô hình được đề xuất. Phần 5 kết luận bài báo này.

2. CÁC NGHIÊN CỨU LIÊN QUAN

Google, công cụ tìm kiếm lớn nhất thế giới (chiếm hơn 90% thị phần), bắt đầu công bố dữ liệu Google Trends từ tháng 1/2004. Dữ liệu này tổng hợp truy vấn tìm kiếm theo từ khóa, thời gian, danh mục và vị trí dựa trên địa chỉ IP. Do quy trình lấy mẫu của Google, dữ liệu tải xuống cho cùng một truy vấn vào các ngày khác nhau có thể hơi khác nhau. Để giảm thiểu sai số, nghiên cứu áp dụng phương pháp của Carrieré-Swallow và Labbé (2013) [9]: thu thập dữ liệu hàng ngày trong 30 ngày và lấy trung bình. Chuỗi thời gian hàng tháng được chuẩn hóa theo tổng số truy vấn tại từng địa phương hoặc quốc gia, với giá trị nằm trong khoảng [0, 100]. Chi tiết về việc dữ liệu Google Trends được mô tả tại www.google.com/trends.

Đã có nhiều nghiên cứu khám phá vấn đề nhu cầu du lịch ở các địa phương và quốc gia khác nhau, vì điều đó là rất quan trọng đối với nền kinh tế của địa phương và quốc gia này. Trước đây, các mô hình chuỗi thời gian được sử dụng để phân tích và dự báo nhu cầu du lịch bằng cách xem xét dữ liệu lịch sử trong quá khứ [10][11][12]. Gần đây, việc kết hợp các phương pháp khác nhau trong dự báo là một hướng tiếp cận đã được nghiên cứu và thử nghiệm nhằm cải thiện độ chính xác của dự báo [13][14][15].

Các bài báo này chủ yếu tập trung vào vào các mô hình dự báo du lịch dựa trên kết hợp các nguồn dữ liệu chuỗi thời gian khác nhau, trong đó đặc biệt tập trung vào dữ liệu truy vấn tìm kiếm.

Trong [16], T. Havranek và cộng sự nghiên cứu hiệu quả của dữ liệu Google Trends trong dự báo lượng khách du lịch và số đêm lưu trú tại Praha (2010–2016). Họ so sánh độ chính xác giữa các mô hình có và không sử dụng dữ liệu tìm kiếm, cũng như giữa biến tần suất cao (Google Trends hàng tuần) và thấp (dữ liệu hàng tháng). Kết quả cho thấy Google Trends có tiềm năng cải thiện dự báo, đặc biệt khi sử dụng dữ liệu tìm kiếm trước 2 tháng và 1 tuần. Mô hình dùng Google Trends hàng tuần cho kết quả vượt trội so với các mô hình khác.

Trong nghiên cứu của [17], các tác giả sử dụng mạng nơ-ron để dự báo nhu cầu du lịch của du khách Nhật Bản đến Hồng Kông, với điểm nổi bật là tích hợp ngày lễ như một biến ngoại sinh. Kết quả cho thấy ngày lễ không chỉ phản ánh yếu tố mùa vụ mà còn ảnh hưởng rõ rệt đến hành vi du lịch, góp phần nâng cao độ chính xác dự báo. Nghiên cứu này là cơ sở quan trọng cho việc tiếp tục khai thác dữ liệu ngày lễ trong các mô hình dự báo sau này, bao gồm cả nghiên cứu của chúng tôi.

Mặc dù các nghiên cứu trước đã bước đầu kết hợp một số nguồn dữ liệu như truy vấn tìm kiếm hoặc ngày lễ vào mô hình dự báo chuỗi thời gian, song hầu hết vẫn chỉ tập trung vào từng nguồn riêng lẻ, thiếu sự tích hợp toàn diện giữa các yếu tố có ảnh hưởng đến nhu cầu du lịch. Việc chỉ sử dụng một nguồn dữ liệu khiến mô hình dễ bị thiên lệch hoặc bỏ sót những tín hiệu quan trọng từ các nguồn thông tin khác. Nghiên cứu của chúng tôi hướng đến khắc phục các hạn chế trên thông qua việc tích hợp dữ liệu từ nhiều nguồn (bao gồm dữ liệu lịch sử, dữ liệu truy vấn tìm kiếm từ Google Trends và dữ liệu ngày lễ) vào một mô hình mạng MLP. Cách tiếp cận này không chỉ tận dụng được các tín hiệu ngắn hạn và yếu tố thời vụ từ dữ liệu truy vấn tìm kiếm và ngày lễ, mà còn kế thừa được tính ổn định và xu hướng dài hạn từ chuỗi thời gian lịch sử. Trong nghiên cứu này, chúng tôi lựa chọn mạng MLP làm khung mô hình dự báo nhằm thử nghiệm khả năng khai thác dữ liệu đa nguồn trong một kiến trúc học sâu đơn giản, ít phức tạp. Việc lựa chọn MLP không nhằm mục tiêu tối ưu hóa tuyệt đối độ chính xác, mà thay vào đó tập trung vào việc phân tích hiệu quả tổng hợp thông tin từ nhiều nguồn dữ liệu khác nhau trong một mô hình học sâu cơ bản. Qua đó, nghiên cứu góp phần đặt nền tảng cho việc phát triển các mô hình nâng cao hơn trong tương lai, khi khả năng tích hợp dữ liệu và năng lực tính toán được mở rộng.

Phần tiếp theo của bài báo sẽ trình bày chi tiết phương pháp khai thác dữ liệu đa nguồn trong mô hình MLP, bao gồm quy trình chuẩn bị dữ liệu, chiến lược kết hợp các nguồn dữ liệu và cấu hình huấn luyện mô hình. Qua đó, bài báo cũng sẽ phân tích rõ mức độ ảnh hưởng của từng loại dữ liệu (tìm kiếm, ngày lễ) đến hiệu quả dự báo trên bài toán dự báo lượng khách quốc tế đến Việt Nam theo tháng.

3. MÔ HÌNH HÓA MẠNG MLP DỰ BÁO ĐA NGUỒN DỮ LIỆU

3.1. Thu thập dữ liệu

Nghiên cứu sử dụng dữ liệu lượt du khách quốc tế đến Việt Nam theo tháng, thu thập từ Cục Du lịch Quốc gia Việt Nam [8], trong giai đoạn 01/2008–12/2023. Dữ liệu từ 01/2008 đến 09/2020 được dùng để huấn luyện mô hình, với 12 tháng cuối làm tập xác thực; dữ liệu từ 10/2020 đến 12/2023 được sử dụng để kiểm tra mô hình. Để đảm bảo dữ liệu phản ánh chính xác sự quan tâm của du khách quốc tế, việc lựa chọn từ khóa phù hợp là rất quan trọng. Quy trình lựa chọn được thực hiện qua nhiều bước: sử dụng công cụ Google Suggest, phỏng vấn du khách quốc tế và tham khảo ý kiến chuyên gia du lịch. Sau đó, hệ số tương quan Pearson giữa lượng tìm kiếm của từng từ khóa và số lượng khách quốc tế đến Việt Nam theo tháng được tính toán. Kết quả cho thấy từ khóa "visit to Vietnam" có tương quan cao nhất và được chọn làm chỉ báo tích hợp vào mô hình. Bảng 1 trình bày chỉ số Google Trends của từ khóa này trong giai đoạn 2008–2023.

Bảng 1: Mô tả dữ liệu thu thập

Tháng	Du khách đến	Lượng tìm kiếm	Ngày lễ
2008-01	399.556	20	0
2008-02	411.032	15	0
2008-03	414.332	13	1
...	...	...	
2023-12	1.371.135	85	1

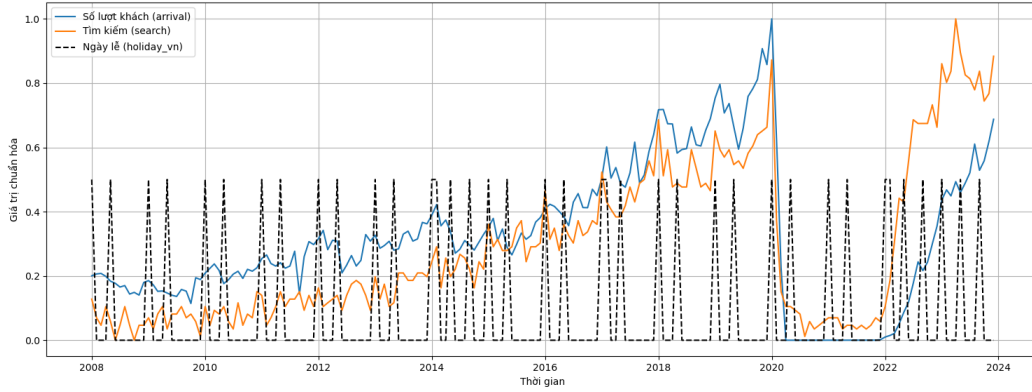
Cuối cùng nguồn dữ liệu được sử dụng cho dự báo nhu cầu du lịch là dữ liệu về các ngày lễ ở Việt Nam từ tháng 1/2008 đến tháng 12/2023. Dữ liệu này được sử dụng làm nguồn dữ liệu tương lai (nghĩa là dữ liệu được biết trước tại thời điểm dự báo).

3.2. Chuẩn hóa dữ liệu

Trước khi đưa vào mô hình học máy (mạng MLP), dữ liệu đầu vào bao gồm lượt du khách quốc tế đến (Arrival), lượng tìm kiếm du lịch (Search) và ngày lễ (Holiday) đã được chuẩn hóa về miền giá trị $[0, 1]$ thông qua phương pháp MinMaxScaler. Việc chuẩn hóa giúp mô hình huấn luyện tốt hơn, tránh hiện tượng một số nguồn dữ liệu có giá trị lớn lấn át ảnh hưởng các nguồn còn lại, đặc biệt quan trọng đối với các mô hình không dựa trên gradient chuẩn hóa như MLP.

Như minh họa ở Hình 1, sau khi chuẩn hóa, các nguồn dữ liệu vẫn giữ được xu hướng và đặc điểm biến động theo thời gian, giúp mô hình khai thác hiệu quả các mẫu thời gian. Chuẩn hóa giúp dễ so sánh và trực quan hóa dữ liệu từ các nguồn khác nhau. Cụ thể, dữ liệu arrival có biên độ lớn, phản ánh xu hướng tăng dài hạn và gián đoạn do

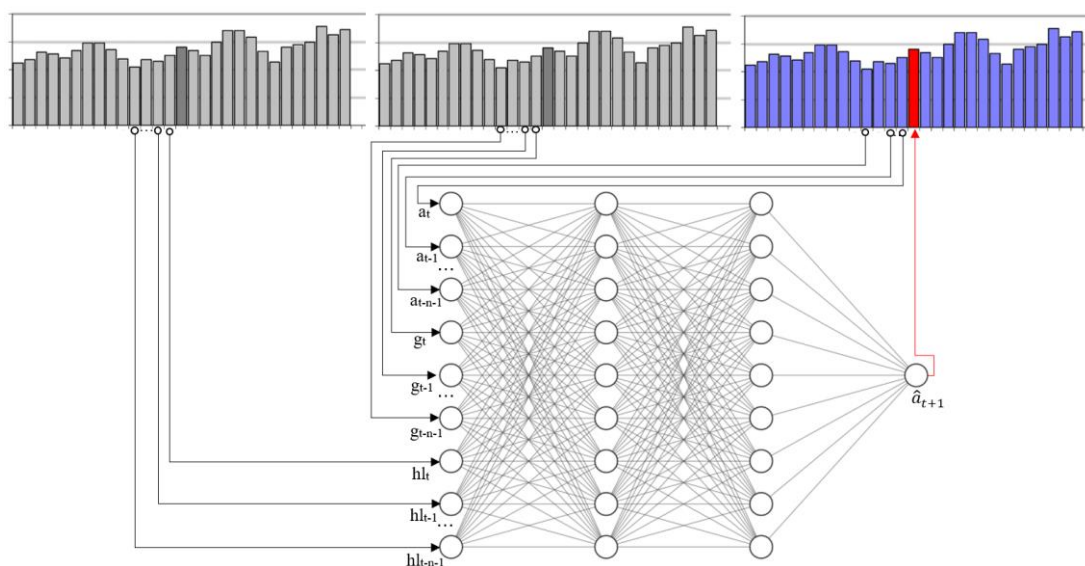
COVID-19; dữ liệu Search tăng mạnh từ 2022; Holiday được biểu diễn bằng đường đứt nét, là đặc trưng rời rạc không bị ảnh hưởng bởi chuẩn hóa. Việc áp dụng MinMaxScaler là bước tiền xử lý quan trọng khi tích hợp dữ liệu đa nguồn với thang đo khác nhau.



Hình 1: Mô tả dữ liệu từ 3 nguồn: lịch sử du khách đến, lượng tìm kiếm và ngày lễ

3.3. Xây dựng mạng MLP cho việc dự báo đa nguồn

Mạng MLP là một loại kiến trúc mạng nơ-ron nhân tạo phổ biến, bao gồm các nơ-ron được tổ chức theo từng lớp: một lớp đầu vào, một hoặc nhiều lớp ẩn và một lớp đầu ra. Thông tin trong mạng MLP được truyền một chiều từ lớp đầu vào đến lớp đầu ra qua các lớp ẩn và không có các kết nối hồi tiếp (feedback). Do đó, MLP còn được xếp vào nhóm các mạng truyền thẳng (feedforward neural networks). Hình 2 thể hiện quá trình lấy dữ liệu đầu vào cho một mô hình mạng MLP, để dự đoán lượt khách du lịch trong tương lai $\hat{a}_{t+1}$. Cụ thể: Lượt khách du lịch a là dữ liệu lịch sử về số lượng khách du lịch qua các thời điểm trước ($t, t-1, \dots, t-n-1$). Lượng tìm kiếm g là dữ liệu về số lượng tìm kiếm theo từ khóa thông tin du lịch trên mạng tại thời điểm $t, t-1, \dots, t-n-1$. Dữ liệu các kỳ nghỉ lễ hl là các đặc trưng nhị phân biểu thị kỳ nghỉ lễ tại thời điểm $t, t-1, \dots, t-n-1$ (có lễ thì 1, không lễ thì 0).



Hình 2: Mô hình MLP dự báo với đa nguồn dữ liệu

Quá trình học của mạng MLP diễn ra bằng cách điều chỉnh trọng số giữa các nơ-ron nhằm tối thiểu hóa sai số giữa đầu ra và giá trị mục tiêu. Thuật toán phổ biến là lan truyền ngược kết hợp với tối ưu hóa bằng gradient descent. Sai số được lan truyền ngược từ đầu ra về đầu vào để cập nhật trọng số theo hướng giảm của gradient hàm lỗi. Nhờ khả năng học phi tuyến, khái quát hóa tốt và tính linh hoạt cao. MLP được ứng dụng rộng rãi trong dự báo chuỗi thời gian, đặc biệt với dữ liệu phức tạp như nhu cầu du lịch, vốn chịu ảnh hưởng từ nhiều yếu tố ngoại sinh.

Tổng số đặc trưng đầu vào của mô hình là $3 \times n$, tương ứng với tổng số đặc trưng thời gian từ ba nguồn. Việc tích hợp các nguồn dữ liệu này nhằm tận dụng ưu thế riêng của từng loại:

- Dữ liệu du khách đến (arrival) phản ánh xu hướng dài hạn và các mẫu hành vi lặp lại trong quá khứ.
- Dữ liệu truy vấn tìm kiếm (search) đại diện cho mức độ quan tâm tức thời và kỳ vọng tương lai của du khách, có thể đóng vai trò như chỉ báo sớm.
- Dữ liệu ngày lễ (holiday) là yếu tố thời vụ ảnh hưởng mạnh đến hành vi du lịch trong ngắn hạn.

Bằng cách đưa cả ba nguồn dữ liệu này vào mô hình một cách đồng thời, mạng MLP có khả năng học các mối quan hệ phức tạp và phi tuyến giữa các yếu tố, từ đó nâng cao độ chính xác của dự báo so với mô hình chỉ sử dụng một nguồn dữ liệu đơn lẻ.

4. CÀI ĐẶT VÀ PHÂN TÍCH KẾT QUẢ

4.1. Phân chia tập dữ liệu và xử lý biến ngoại sinh

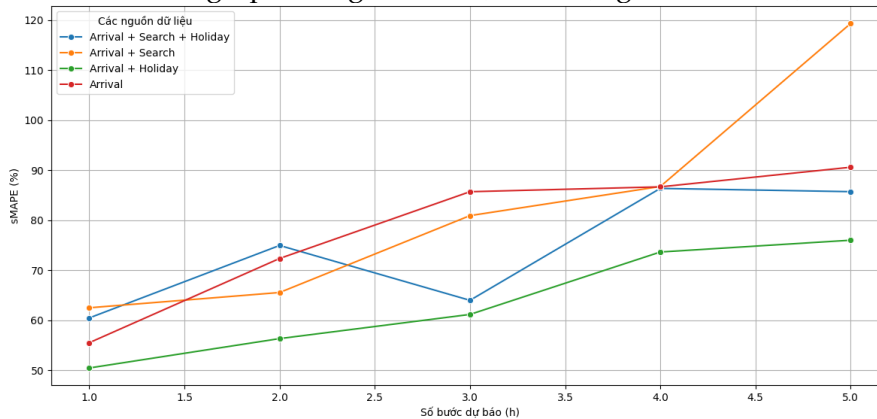
Với dữ liệu là số lượng du khách quốc tế đến Việt Nam theo tháng từ tháng 1/2008 đến tháng 12/2023 (Hình 1), tập dữ liệu được chia thành ba tập: Tập huấn luyện (Training set): gồm dữ liệu từ tháng 1/2008 đến tháng 8/2019, được sử dụng để học các đặc điểm của chuỗi thời gian; Tập đánh giá (Validation set): là 12 tháng kế tiếp, từ tháng 9/2019 đến tháng 8/2020, được dùng để giám sát quá trình huấn luyện, điều chỉnh siêu tham số và áp dụng cơ chế dừng sớm (early stopping); và Tập kiểm tra (Test set): bao gồm dữ liệu từ tháng 9/2021 đến tháng 12/2023, nhằm đánh giá hiệu quả dự báo của mô hình trong điều kiện thực tế. Cũng trong Hình 1, hai loại biến ngoại sinh (exogenous variables) được đưa vào để nâng cao hiệu quả dự báo. Biến ngoại sinh quá khứ *Search* (lượng tìm kiếm) được dùng làm đặc trưng đầu vào cùng với biến mục tiêu quá khứ và có mặt đầy đủ trong cả tập huấn luyện, đánh giá và kiểm tra. Biến *Holiday_vn* (ngày lễ tại Việt Nam) là biến ngoại sinh tương lai có thể biết trước, được sử dụng trong giai đoạn dự báo theo từng bước với chiều dài bằng số bước dự báo h (ví dụ: $h = 1, 3$ hoặc 6 tháng). Mỗi lần trượt cửa sổ, mô hình được cung cấp thông tin ngày lễ tương ứng với thời gian dự báo, giúp cải thiện độ chính xác trong các giai đoạn có yếu tố mùa vụ hoặc sự kiện đặc biệt.

4.2. Dự báo và phân tích kết quả

Hiệu quả dự báo được đánh giá trong nghiên cứu này dựa trên tiêu chí sMAPE

$$sMAPE = \frac{100\%}{n} \sum_{t=1}^n \frac{|F_t - A_t|}{\frac{|A_t| + |F_t|}{2}} \quad (1)$$

trong đó, F_t là giá trị thực tế tại thời điểm t ; A_t là giá trị dự báo tại thời điểm t ; n là tổng số mẫu (trong bài báo là số tháng). Ngoài sMAPE là tiêu chí chính, nghiên cứu còn sử dụng thêm hai chỉ số phổ biến là MAE (Mean Absolute Error) và RMSE (Root Mean Square Error). MAE đo mức sai lệch trung bình giữa dự báo và thực tế, còn RMSE nhấn mạnh các sai số lớn hơn, giúp đánh giá mức độ biến động của sai số dự báo.



Hình 3: So sánh hiệu quả (sMAPE) dự báo ngắn hạn của sự kết hợp các nguồn dữ liệu

Thực nghiệm đánh giá hiệu quả dự báo (dựa trên công thức 1) với các trường hợp sau: chỉ dựa trên lịch sử du khách đến (Arrival), kết hợp lịch sử du khách đến và lượng tìm kiếm (Arrival + Search), kết hợp lịch sử du khách và ngày lễ (Arrival + Holiday) và kết hợp cả ba nguồn dữ liệu (Arrival + Search + Holiday). Với dự báo ngắn hạn (độ dài dự báo $h = 1$), việc kết hợp cả ba nguồn dữ liệu (Arrival + Search + Holiday) cho kết quả tốt nhất với sMAPE đạt 44.52%. Điều này cho thấy, với dự báo ngắn hạn, thông tin từ các nguồn dữ liệu thực sự có ích, giúp mô hình nắm bắt các yếu tố tác động ngắn hạn lên nhu cầu du lịch, từ đó dự báo chính xác hơn so với chỉ dựa vào dữ liệu lượt khách quá khứ. Tuy nhiên, khi tăng độ dài dự báo lên (dự báo dài hạn), hiệu quả của việc thêm nguồn dữ liệu dần suy giảm (Bảng 2). Từ $h = 2$ trở đi, sMAPE đều tăng lên rõ rệt dù có hay không sử dụng nguồn dữ liệu và đến $h = 6$, ngay cả khi dùng cả dữ liệu Search và Holiday, sMAPE cũng vượt mức 100%, chứng tỏ mô hình bắt đầu gặp khó khăn trong việc nắm bắt xu hướng dài hạn. Việc này phản ánh hạn chế của mô hình MLP khi không có cơ chế ghi nhớ tuần tự, đồng thời cũng cho thấy rằng tác động của các yếu tố ngắn hạn như tìm kiếm và lễ hội sẽ khó duy trì được ảnh hưởng mạnh mẽ trong các dự báo dài hạn.

Có thể nhận thấy rằng, không phải nguồn dữ liệu nào cũng có hiệu quả như nhau. Khi so sánh riêng giữa kết hợp với Search hay Holiday, rõ ràng biến Holiday mang lại cải thiện tốt hơn cho mô hình so với Search ở hầu hết các độ dài dự báo. Điều này phần nào hợp lý vì các dịp lễ thường trực tiếp kéo theo sự gia tăng thực tế về lượng khách du lịch, trong khi dữ liệu tìm kiếm có thể chỉ phản ánh sự quan tâm, chưa chắc chuyển hóa thành hành vi du lịch thực tế, đặc biệt trong bối cảnh bất ổn như giai đoạn đại dịch COVID-19.

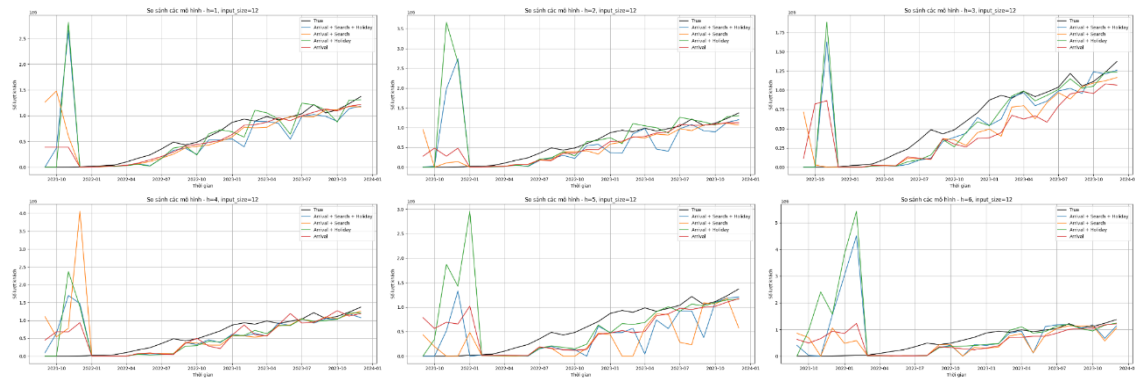
Bảng 2. Bảng đánh giá dự báo theo độ dài dự báo (h)

h	Nguồn dữ liệu	sMAPE	MAE	RMSE	h	Nguồn dữ liệu	sMAPE	MAE	RMSE
1	S1	44,52	265.887	742.594	4	S1	82,25	303.801	741.811
1	S2	67,29	215.419	572.386	4	S2	89,27	332.084	774.886
1	S3	50,46	204.484	552.181	4	S3	73,63	275.939	552.201
1	S4	55,51	118.720	161.877	4	S4	86,65	249.653	339.620
2	S1	56,61	406.425	1.153.903	5	S1	71,72	399.407	851.206
2	S2	70,53	351.094	921.113	5	S2	98,96	491.893	1.017.856
2	S3	56,33	319.021	863.835	5	S3	75,99	367.137	737.329
2	S4	72,34	165.849	207.997	5	S4	90,57	303.369	396.681
3	S1	73,93	222.241	553.328	6	S1	101,41	760.716	1.805.743
3	S2	78,96	223.440	611.719	6	S2	108,06	794.871	1.643.134

3	S3	61,16	155.450	379.303	6	S3	82,83	637.930	1.386.811
3	S4	85,68	249.877	328.381	6	S4	95,28	346.509	453.040

S1: Arrival + Search + Holiday; S2: Arrival + Search; S3: Arrival + Holiday; S4: Arrival

Ngoài ra, việc độ dài dự báo (h) càng tăng thì sMAPE cũng tăng đều với tất cả các trường hợp, khẳng định rằng việc dự báo càng xa càng làm giảm độ chính xác, đặc biệt đối với mô hình học máy đơn giản như MLP. Thực tế, MLP có thể khai thác tốt mối quan hệ ngắn hạn giữa các nguồn dữ liệu, nhưng khả năng suy luận về chuỗi dài hạn, xu hướng vĩ mô hay biến động phức tạp là rất hạn chế. Hình 3 so sánh kết quả dự báo giữa các kết hợp nguồn dữ liệu khi thay đổi độ dài dự báo h .



Hình 3: So sánh hiệu quả dự báo (sMAPE) với độ dài dự báo thay đổi

Từ kết quả thực nghiệm có thể rút ra kết luận rằng nguồn dữ liệu rất hữu ích cho dự báo ngắn hạn nhưng cần được cân nhắc kỹ lưỡng về chất lượng và tính ổn định của nguồn dữ liệu. Đồng thời, khi mục tiêu là dự báo xa hơn, cần kết hợp thêm các mô hình có khả năng ghi nhớ tốt hơn như RNN, LSTM hoặc Transformer để tận dụng tốt hơn các nguồn dữ liệu đa dạng.

5. KẾT LUẬN

Nghiên cứu đã đánh giá tác động của dữ liệu đa nguồn trong dự báo chuỗi thời gian bằng mô hình MLP, với ứng dụng dự báo khách quốc tế đến Việt Nam. Việc bổ sung dữ liệu tìm kiếm và ngày lễ giúp cải thiện đáng kể độ chính xác trong dự báo ngắn hạn ($h=1$), đặc biệt khi kết hợp đầy đủ các nguồn. Tuy nhiên, hiệu quả giảm dần theo độ dài dự báo và dữ liệu ngày lễ cho thấy tác động ổn định hơn so với tìm kiếm. Mô hình MLP tỏ ra kém hiệu quả trong dự báo dài hạn, dù có thêm dữ liệu ngoại sinh. Nghiên cứu cho thấy tiềm năng của dữ liệu đa nguồn, đồng thời khuyến nghị áp dụng các mô hình học sâu mạnh hơn như LSTM, GRU hoặc Transformer cho các bài toán dự báo độ dài lớn.

TÀI LIỆU THAM KHẢO

- [1] J. Winter and R. Ono (2015). Introduction to the Future Internet: Alternative Visions, *Public Adm Inf Technol*, vol. 17, pp. 1–16.
- [2] L. Tang, J. Li, H. Du, L. Li, J. Wu, and S. Wang (2021). Big Data in Forecasting Research: A Literature Review, *Big Data Res*, vol. 27, pp. 100289.
- [3] M. Fathi, M. Haghi Kashani, S. M. Jameii, and E. Mahdipour (2022). Big Data Analytics in Weather Forecasting: A Systematic Review, *Arch Computat Methods Eng* 29, vol. 29, pp. 1247–1275.
- [4] H. Li, M. Hu, and G. Li (2020). Forecasting tourism demand with multisource big data, *Annals of Tourism Research*, vol. 83.
- [5] S. P. Jun, H. S. Yoo, and S. Choi (2018). Ten years of research change using Google Trends: From the perspective of big data utilizations and applications, *Technological Forecasting and Social Change*, vol. 130, 2018, pp. 69–87.
- [6] Y. Li, Z. Lin, and S. Xiao (2022). Using social media big data for tourist demand forecasting: A new machine learning analytical approach, *J. Digit, Econ*, vol. 1, no. 1, pp.32–43, 2022
- [7] M. Á, R. Reina (2020). Big Data: Forecasting and Control for Tourism Demand, *Theory and Applications of Time Series Analysis. ITISE 2019. Contributions to Statistics. Springer*.
- [8] Cục Du lịch Quốc gia Việt Nam (2024). Số liệu thống kê Khách quốc tế, <https://vietnamtourism.gov.vn/statistic/international> (accessed Aug, 15, 2024).
- [9] Y, Carrière-Swallow and F, Labbé (2013). Nowcasting with Google trends in an emerging market, *J. Forecast*, vol. 32, pp. 289–298.
- [10] C. Lim and M. McAleer (2001). Forecasting tourist arrivals, *Ann. Tour. Res.*, vol. 28, no. 4, pp. 965–977.
- [11] V. Cho (2003). A comparison of three different approaches to tourist arrival forecasting, *Tour. Manag.*, vol. 24, no. 3, pp. 323–330.
- [12] H. Hassani, E. S. Silva, N. Antonakakis, G. Filis, and R. Gupta (2017). Forecasting accuracy evaluation of tourist arrivals, *Ann. Tour. Res.*, vol. 63, pp. 112–127.
- [13] M. E. Nor, A. I. M, Nurul, and M. S, Rusiman (2018). A Hybrid Approach on Tourism Demand Forecasting, *J. Phys, Conf. Ser.*, vol. 995, no. 1, pp. 1–11.
- [14] M. L, Shen, H. H. Liu, Y. H, Lien, C. F. Lee, and C. H, Yang (2019), Hybrid approach for forecasting tourist arrivals, *ACM Int. Conf. Proceeding Ser.*, vol. Part F1479, pp. 392–396.
- [15] Purwanto, Sunardi, F. T. Julfia, and A. Paramananda (2019). Hybrid model of ARIMA-linear trend model for tourist arrivals prediction model in Surakarta City, Indonesia, *AIP Conf. Proc.*, vol. 2114, no. June.
- [16] T. Havranek and A. Zeynalov (2021). Forecasting tourist arrivals: Google Trends meets mixed-frequency data, *Tour. Econ.*, vol. 27, no. 1, pp. 129–148.
- [17] Law, R., & Au, N. (1999). A neural network model to forecast Japanese demand for travel to Hong Kong. *Tourism Management*, 20(1), 89–97.

ANALYZING THE IMPACT OF MULTI-SOURCE DATA ON TIME SERIES FORECASTING USING MLP NETWORKS

Ngo Van Son^{1,2}, Vo Viet Minh Nhat³

¹ University of Sciences, Hue University

² School of Hospitality and Tourism, Hue University

³ Institute for Educational Testing and Quality Assurance, Hue University

Email: ngovanson@hueuni.edu.vn, vvmnhat@hueuni.edu.vn

ABSTRACT

This paper investigates the effectiveness of integrating multi-source data in time series forecasting using a multilayer perceptron (MLP) model with a specific application to monthly forecasts of international tourist arrivals to Vietnam. In addition to historical data, the model incorporates online search data from Google Trends and information on national holidays. The results show that exogenous data sources improve forecasting accuracy, particularly in short-term forecasts ($h=1$), significantly reducing sMAPE when all data sources are combined. However, the influence of exogenous variables diminishes as the forecasting horizon increases and the effectiveness of different sources varies; holiday data exhibits a more consistent positive impact than online search data. The study highlights the potential of multi-source data in enhancing forecast performance and points out the limitations of MLP models for long-term forecasting, suggesting future exploration of more advanced deep learning approaches.

Keywords: Vietnam tourism, tourism demand forecasting, search engine data, Google Trends.



Ngô Văn Sơn nhận bằng Thạc sĩ Khoa học máy tính tại Trường Đại học Khoa học, ĐH Huế năm 2010. Ông hiện công tác tại trường Du lịch, ĐH Huế.

Lĩnh vực nghiên cứu: dự báo chuỗi thời gian, ứng dụng di động, xử lý ngôn ngữ tự nhiên, chatbot và hệ thống tư vấn.



Võ Viết Minh Nhật nhận bằng Tiến sĩ về Công nghệ thông tin năm 2007 tại Đại học Quebec ở Montreal (Canada), đạt danh hiệu phó giáo sư năm 2016. Hiện ông công tác Viện Khảo thí và Bảo đảm chất lượng giáo dục - Đại học Huế.

Lĩnh vực nghiên cứu: mạng chuyển mạch quang, công nghệ RFID, đảm bảo đa dạng chất lượng dịch vụ, tối ưu hóa.