Tạp chí Khoa học Đại học Cần Thơ
Số chuyên đề: Khoa học tự nhiên
ISSN 1859-2333 | e-ISSN 2815-5599

DOI:10.22144/ctujos.2024.341

DỰ BÁO CHUỖI THỜI GIAN VỚI MỘT SỐ MÔ HÌNH HỌC MÁY VÀ ỨNG DỤNG

Lê Trung Can và Trần Phước Lộc*

Khoa Khoa học Tự nhiên, Trường Đại học Cần Thơ
**Tác giả liên hệ (Corresponding author): tploc@ctu.edu.vn*

Thông tin chung (Article Information)

Nhận bài (Received): 01/05/2024
Sửa bài (Revised): 25/06/2024
Duyệt đăng (Accepted): 01/08/2024

Title: *Forecasting time series with some machine learning models and applications*

Author(s): *Le Trung Can and Tran Phuoc Loc**

Affiliation(s): *Can Tho University*

TÓM TẮT

Nghiên cứu này trình bày việc phân tích và dự báo dữ liệu chuỗi thời gian bằng cách sử dụng các mô hình học máy khác nhau. Các phương pháp được sử dụng bao gồm Holt-Winters, ARIMA, hồi quy tuyến tính (LR), rừng ngẫu nhiên (RF), máy tăng cường độ dốc (GBM) và học máy tự động (AutoML). Các phương pháp tìm kiếm lưới nâng cao cũng được áp dụng cho ARIMA, RF và GBM để tối ưu hóa mô hình. Dữ liệu lưu lượng nước hàng tháng tại trạm đo trên Sông Tiền ở Tân Châu từ năm 1992 đến 2021 được sử dụng để huấn luyện và kiểm tra các mô hình. Kết quả cho thấy mô hình GBM với tìm kiếm lưới nâng cao cho độ chính xác vượt trội so với các mô hình khác.

Từ khoá: *Chuỗi thời gian, dự báo, học máy, lưu lượng nước*

ABSTRACT

This study presents the analyses and forecasts of time series data using different machine learning models. Methods used include Holt-Winters, ARIMA, random forest (RF), gradient boosting machine (GBM), and automatic machine learning (AutoML). Advanced grid search methods are also applied to ARIMA, RF, and GBM for model optimization. Monthly water flow data at the Tien River measuring station in Tan Chau, Vietnam, from 1992 to 2021 are used to train and test the models. The results show that the GBM model with advanced grid search has superior accuracy compared to other models.

Keywords: *Time series, forecast, machine learning, river discharge*

1. GIỚI THIỆU

Dự báo là một trong những kênh thông tin quan trọng giúp chúng ta chuẩn bị tốt hơn cho tương lai và đưa ra quyết định dựa trên những gì được dự đoán sẽ xảy ra. Trong thời đại công nghệ thông tin hiện nay, việc sử dụng các mô hình toán học và học máy (ML) để dự báo dữ liệu đã trở nên phổ biến.

Dữ liệu chuỗi thời gian là một loại dữ liệu quan trọng trong nhiều lĩnh vực như kinh tế, y tế, sinh học, giáo dục và xã hội học. Dữ liệu này bao gồm các quan sát được thu thập tại các thời điểm cách đều nhau. Những ví dụ phổ biến của dữ liệu chuỗi thời gian bao gồm giá chứng khoán, nhiệt độ trung bình hàng ngày, lưu lượng nước hàng tháng, mức tiêu thụ năng lượng và sản lượng nông nghiệp. Việc phân tích và dự báo chuỗi thời gian giúp chúng ta hiểu rõ hơn về các quy luật trong quá khứ và đưa ra những dự đoán chính xác về tương lai, từ đó hỗ trợ quá trình ra quyết định trong nhiều lĩnh vực khác nhau. Đối với dữ liệu này, một số mô hình chuỗi thời gian như mô hình hồi quy, Holt-Winters, ARIMA,... được sử

dụng phổ biến để phân tích và dự báo. Những năm gần đây, với sự phát triển nhanh chóng của máy tính có cấu hình mạnh và các thuật toán ML, học sâu và trí tuệ nhân tạo cho phép chúng ta thực hiện dễ dàng các phép tính phức tạp hơn trước đây để phân tích các bộ dữ liệu.

ML là một nhánh của trí tuệ nhân tạo, tập trung vào việc phát triển các thuật toán cho phép máy tính học từ dữ liệu và cải thiện hiệu suất theo thời gian mà không cần sự can thiệp nhiều của con người. ML gồm nhiều loại khác nhau, thường bao gồm học có giám sát (như hồi quy và phân loại), học không giám sát (như phân cụm và giảm số chiều dữ liệu), học bán giám sát, học tăng cường, học truyền dẫn, và học theo nhóm. Mục tiêu của học tập theo nhóm là cải thiện hiệu suất dự đoán hoặc phân loại bằng cách kết hợp các dự đoán từ nhiều mô hình ML khác nhau. Nó có thể được coi như một cách bù đắp cho các thuật toán đơn lẻ có hiệu suất không cao. Các thuật toán này thường được sử dụng để giảm độ chệch, phương sai, hoặc cải thiện dự báo.

Với dữ liệu chuỗi thời gian, ML đã chứng minh được khả năng vượt trội trong việc xử lý các tập dữ liệu lớn và phức tạp, cho phép chúng ta dự đoán chính xác hơn so với các phương pháp truyền thống. Nghiên cứu này giới thiệu và sử dụng một số mô hình ML theo nhóm để phân tích dữ liệu chuỗi thời gian. Rừng ngẫu nhiên (random forest – RF), một phương pháp ML dựa trên tập hợp các cây quyết định, được giới thiệu bởi Breiman (2001). Máy tăng cường độ dốc (gradient boosting machines - GBM), một phương pháp ML mạnh mẽ khác, phát triển dựa trên kỹ thuật tăng cường, khi nhiều mô hình yếu được kết hợp để tạo ra một mô hình mạnh hơn (Friedman, 2001). ML tự động (automated machine learning - AutoML), một kỹ thuật giúp tự động hóa quá trình xây dựng các mô hình ML dựa trên các kỹ thuật khác nhau. Trong bài viết này, hàm từ thư viện h2o trong ngôn ngữ R (LeDell & Poirier, 2020) được sử dụng để huấn luyện mô hình.

Trong những năm gần đây, vấn đề biến đổi khí hậu như sự gia tăng nhiệt độ toàn cầu, hạn hán, lũ lụt, thiếu hụt nguồn nước ngọt, xâm nhập mặn,... đã gây ra những ảnh hưởng tiêu cực đến đời sống sinh hoạt và sản xuất của con người. Do đó, công tác dự báo và cảnh báo lũ, xâm nhập mặn đóng một vai trò hết sức quan trọng và thu hút sự quan tâm của nhiều cơ quan, trung tâm dự báo và nhà khoa học. Trung tâm Dự báo Khí tượng Thủy văn Quốc gia đã ứng dụng mô hình hoá (MIKE 11) để tính toán dòng chảy lũ hạ lưu và dự báo mặn cho đồng bằng sông Cửu Long (ĐBSCL). Kết quả mô phỏng cho thấy mô hình này có khả năng dự báo tốt về xu thế và có độ chính xác cao (Hải và ctv., 2020). Khi sử dụng mô hình MIKE 11 để mô phỏng xâm nhập mặn tại ĐBSCL, kết quả kiểm định cho thấy sự tương quan tốt giữa mực nước và độ mặn tại các trạm thủy văn (Toàn và ctv., 2020). Tại tỉnh Bến Tre, ngoài phương pháp dự báo thống kê và kinh nghiệm, các chuyên gia đã áp dụng công nghệ mới và sử dụng mô hình MIKE 11 để hiệu chỉnh và kiểm định số liệu mực nước, đạt kết quả tốt cho công tác dự báo chi tiết mực nước và độ mặn (Lam và ctv., 2022). Ngoài ra, các thuật toán ML đã được ứng dụng để dự báo xâm nhập mặn ở ĐBSCL nhằm quản lý nguồn nước ngọt và giảm thiểu tác động của xâm nhập mặn. Phương pháp k-hàng xóm gần nhất (KNN) đã được sử dụng để dự báo độ mặn trên sông Hàm Luông, Bến Tre, cho kết quả khá chính xác (Hoài và ctv., 2022). Các thuật toán như LR, RF, và mạng nơ-ron nhân tạo (ANN) cũng được áp dụng để dự đoán xâm nhập mặn hàng tuần từ năm 2012 đến 2020, với mô hình ANN đạt hiệu suất cao (Pham et al., 2022). Các mô hình học sâu như LSTM, MLP, CNN, và Transformer đã được dùng để kiểm tra mực nước hàng giờ tại các cửa sông Mê Kông, trong đó LSTM có độ chính xác và tin cậy cao (Tran et al., 2022).

Mặc dù hiệu quả của việc sử dụng các mô hình ML trong việc phân tích và dự báo đã được chứng minh là khá hiệu quả, nhưng các áp dụng của nó trong lĩnh vực khí tượng thủy văn vẫn còn ít, đặc biệt là trong dự báo lưu lượng dòng chảy của nước. Do đó, động lực chính và đóng góp của nghiên cứu này là khái quát lý thuyết và ứng dụng các mô hình ML vào việc phân tích và dự báo dữ liệu thời gian, đặc biệt là dữ liệu về lưu lượng dòng chảy của nước ở trạm đo sông Tiền tại Tân Châu, An Giang. Việc áp dụng các mô hình trong nghiên cứu này sẽ góp phần làm đa dạng hơn các mô hình dự báo chính xác và kịp thời, từ đó hỗ trợ thêm thông tin cho các cơ quan chức năng và cộng đồng trong việc chuẩn bị và ứng phó với những thách thức về thời tiết và biến đổi khí hậu, đồng thời bổ sung thêm tài liệu tham khảo cho các nhà nghiên cứu, giảng viên và sinh viên trong vấn đề dự báo dữ liệu chuỗi thời gian.

2. KIẾN THỨC LIÊN QUAN

2.1. Chuỗi thời gian

Chuỗi thời gian là một dãy các giá trị quan sát được sắp thứ tự theo diễn biến thời gian cách đều nhau theo năm, quý, tháng, ngày, hoặc giờ,... Ví dụ về chuỗi thời gian là độ cao của thủy triều, lưu lượng nước hàng tháng tại một trạm đo, và số lượng hàng bán được của một công ty trong một quý. Chuỗi thời

gian thông thường bao gồm bốn thành phần cơ bản là xu hướng, mùa, chu kì và ngẫu nhiên.

2.2. Học máy

ML là lĩnh vực trong trí tuệ nhân tạo, cho phép máy tính tự học từ dữ liệu và cải thiện hiệu suất theo thời gian. Nó liên quan đến việc xây dựng các mô hình dự đoán và phân loại dựa trên dữ liệu huấn luyện. Các thuật toán ML giúp máy tính nhận biết mẫu, tìm ra quy luật ẩn sau dữ liệu và thực hiện dự đoán cho dữ liệu mới. ML đang được áp dụng rộng rãi trong nhiều lĩnh vực như xử lý ngôn ngữ tự nhiên, thị giác máy tính, tài chính, y học và đặc biệt trong dự báo dữ liệu chuỗi thời gian.

2.3. Quy trình dự báo chuỗi thời gian

Quy trình dự báo chuỗi thời gian, tương tự như các lĩnh vực phân tích dự báo khác, thường bao gồm các bước chính sau đây:

Bước 1. Chuẩn bị dữ liệu: bao gồm các công việc có thể như làm sạch dữ liệu sau khi thu thập, mã hóa lại các biến, chuẩn hóa, biến đổi hoặc tạo những biến mới, và chia chuỗi dữ liệu thành tập huấn luyện và kiểm tra. Thông thường, tập kiểm tra chiếm tối đa 30% dữ liệu để đảm bảo hiệu suất tính toán.

Bước 2. Huấn luyện mô hình: tập huấn luyện được sử dụng để thiết lập các mô hình khác nhau từ các mô hình được lựa chọn.

Bước 3. Kiểm tra và so sánh: Độ chính xác hoặc các tiêu chí chấm điểm sai số trên cả tập huấn luyện và kiểm tra của từng mô hình được so sánh để lựa chọn mô hình tiềm năng.

Bước 4. Cải thiện hiệu suất: Các mô hình tiềm năng sẽ được điều chỉnh các tham số và lặp lại quá trình huấn luyện để cải thiện hiệu suất.

Bước 5. Dự báo: Mô hình tối ưu nhất được sử dụng cho toàn bộ dữ liệu ban đầu và đưa ra dự báo cho các bước thời gian tiếp theo.

2.4. Chấm điểm dự báo

Gọi $\{Y_t,\ t = 1,2,\dots,T\}$ là một chuỗi thời gian và $\hat{Y}_t$ là giá trị dự báo tương ứng. Gọi $n\ (n \le T)$ là cỡ mẫu của tập dữ liệu xem xét, chẳng hạn như tập huấn luyện hoặc tập kiểm tra được tách ra từ chuỗi $\{Y_t\}$. Để so sánh hiệu quả giữa các mô hình, một số phép đo lường hoặc chấm điểm sai số sau thường được sử dụng:

– *Trung bình bình phương sai số (MSE):* Đây là trung bình của bình phương sai số giữa giá trị thực tế và giá trị dự đoán.

$$\mathrm{MSE} = \frac{1}{n}\sum_{t=1}^{n}\left(Y_t - \hat{Y}_t\right)^2 . \qquad (1)$$

– *Căn bậc hai của trung bình bình phương sai số (RMSE):* Đây là căn bậc hai của MSE, giúp đưa sai số về cùng đơn vị với giá trị dự đoán và thực tế.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(Y_t - \hat{Y}_t\right)^2}. \qquad (2)$$

– *Trung bình tuyệt đối sai số (MAE):* Đây là trung bình của giá trị tuyệt đối sai số giữa giá trị thực tế và giá trị dự đoán.

$$\mathrm{MAE} = \frac{1}{n}\sum_{t=1}^{n}\left|Y_t - \hat{Y}_t\right| . \qquad (3)$$

– *Trung bình tuyệt đối phần trăm của sai số (MAPE):* Đây là trung bình của giá trị tuyệt đối phần trăm sai số giữa giá trị thực tế và giá trị dự đoán.

$$\mathrm{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{Y_t - \hat{Y}_t}{Y_t}\right| . \qquad (4)$$

Ngoài ra, còn các chỉ số như hệ số xác định (Coefficient of Determination - R^2), R^2 hiệu chỉnh, Akaike Information Criterion (AIC) và Bayesian Information Criterion (BIC),... Trong nghiên cứu này, R^2 hiệu chỉnh được sử dụng để xác định trong mô hình LR, AIC trong việc tìm kiếm nâng cao các mô hình ARIMA, RMSE trong việc tìm kiếm nâng cao các mô hình ML, và MAPE trên tập huấn luyện và tập kiểm tra để so sánh hiệu quả giữa các mô hình.

Phần tiếp theo trình bày tóm tắt một số phương pháp dự báo phổ biến và ML cho dữ liệu thời gian.

3. MỘT SỐ MÔ HÌNH DỰ BÁO

3.1. Holt-Winters

Holt (2004) và Winters (1960) đã mở rộng mô hình Holt cho các dữ liệu chuỗi thời gian với cả thành phần xu hướng và mùa, gọi là mô hình HW. Hai dạng mô hình HW cơ bản bao gồm một phương trình dự báo và ba phương trình làm trơn được cho sau đây:

– Mô hình HW$^+$ được biểu diễn:

$$\begin{aligned}\hat{Y}_{t+h|t} &= (l_t + hb_t)s_{t+h-m(k+1)},\\ l_t &= \alpha(Y_t - s_{t-m}) + (1-\alpha)(l_{t-1} + b_{t-1}),\\ b_t &= \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1},\\ s_t &= \gamma(Y_t - l_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m}, \qquad (5)\end{aligned}$$

trong đó, $0 \le \alpha \le 1, 0 \le \beta \le 1,\ 0 \le \gamma \le 1,\ m$ là tần suất mùa, k là phần nguyên của $(h-1)/m$, h là

số bước dự báo tiếp theo tính từ thời điểm t; l_t, b_t, và s_t lần lượt là các thành phần mức độ, xu hướng và mùa của chuỗi.

– Mô hình HW* được biểu diễn:

$$\hat{Y}_{t+h|t} = l_t + hb_t + s_{t+h-m(k+1)},$$
$$l_t = \alpha\left(\frac{Y_t}{s_{t-m}}\right) + (1-\alpha)(l_{t-1} + b_{t-1}),$$
$$b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1},$$
$$s_t = \gamma\left(\frac{Y_t}{l_{t-1} + b_{t-1}}\right) + (1-\gamma)s_{t-m}. \quad (6)$$

3.2. ARIMA

Mô hình trung bình trượt tích hợp tự động hồi quy (ARIMA), kí hiệu là ARIMA(p,d,q), là một mô hình thống kê được sử dụng để phân tích và dự đoán dữ liệu chuỗi thời gian. Mô hình này kết hợp các thành phần tự hồi quy AR(p), tích hợp I(d) với d là số lần sai phân chuỗi đến khi đạt trạng thái dừng, và trung bình trượt MA(q) để loại bỏ tính không ổn định và mô hình hóa cấu trúc của chuỗi thời gian. Mô hình ARIMA(p,d,q) có thể được biểu diễn sau đây:

$$Y'_t = c + \phi_1 Y'_{t-1} + \phi_2 Y'_{t-2} + \cdots + \phi_p Y'_{t-p} + \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + \cdots + \theta_q\epsilon_{t-q}, \quad (7)$$

trong đó Y'_t là sai phân của chuỗi Y_t (sai phân đến d lần), ϕ_i và θ_j $(i = 1, \ldots, p, j = 1, \ldots, q)$ lần lượt là các tham số của mô hình AR(p) và MA(q), và ϵ_t là nhiễu trắng.

Ngoài ra, mô hình SARIMA, kí hiệu ARIMA$(p,d,q)(P,D,Q)_m$, là một dạng mở rộng của ARIMA cho chuỗi thời gian có thành phần theo mùa, với việc sử dụng ba thành phần mùa SAR(P), SI(D), và SAM(Q) trong đó D là sai phân theo mùa bậc D, và m là tần suất mùa của chuỗi.

3.3. Hồi quy tuyến tính

LR là một mô hình toán học mô tả mối quan hệ tuyến tính giữa một biến phụ thuộc và một hoặc nhiều biến độc lập. Mô hình này được biểu diễn qua phương trình:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_r X_{ri} + \epsilon_i,$$

trong đó,

– Y là biến phụ thuộc,

– X_j, $j = 1, 2, \ldots, r$, là các biến giải thích hay các biến độc lập,

– $\beta_0, \beta_1, \ldots, \beta_r$ là các hệ số của mô hình,

– ϵ là phần dư,

– $i = 1, 2, \ldots, n$ là ký hiệu quan sát thứ i trong mẫu có n quan sát.

Các hệ số của mô hình hồi quy tuyến tính thường được ước lượng bằng phương pháp bình phương tối thiểu. Hồi quy tuyến tính là một kỹ thuật thống kê đã được sử dụng lâu đời, hiệu quả và dễ dàng áp dụng cho phần mềm và tính toán, cung cấp một công thức toán học dễ giải thích để đưa ra các dự đoán.

3.4. Rừng ngẫu nhiên

Rừng ngẫu nhiên (RF) là một phương pháp học tập kết hợp rất phổ biến trong các bài toán toán hồi quy và phân loại. Thuật toán này kết hợp đầu ra của nhiều cây quyết định để tạo ra một dự đoán duy nhất, do đó nó thường linh hoạt, dễ sử dụng và có độ chính xác cao. Cụ thể, RF tạo ra một bộ phân loại từ nhiều cây quyết định, mỗi cây được xây dựng trên một tập con ngẫu nhiên của dữ liệu và một tập con ngẫu nhiên của các đặc trưng hoặc biến. Thay vì dựa vào một cây quyết định duy nhất, thuật toán này tận dụng sức mạnh của "đám đông" để tạo ra dự đoán chính xác hơn. Số lượng cây lớn hơn trong rừng giúp tăng độ chính xác và ngăn chặn vấn đề quá khớp.

Các bước thực hiện của phương pháp này tương tự như các bước chung của ML. Tuy nhiên, bước huấn luyện mô hình có thể mô tả tóm tắt thành 3 bước con như sau:

Bước 1: Lấy mẫu tái lập (bootstrap) ngẫu nhiên từ tập huấn luyện để tạo thành một tập dữ liệu con.

Bước 2: Chọn ngẫu nhiên s biến $(s \leq r)$, với r là tổng số các đặc trưng (biến) đầu vào của mô hình, và xây dựng mô hình cây quyết định dựa trên các biến này và tập dữ liệu con ở bước 1. Lặp lại bước 1 và 2 nhiều lần để được nhiều cây quyết định.

Bước 3: Thực hiện lấy trung bình giữa các cây quyết định để đưa ra dự báo.

3.5. Máy tăng cường độ dốc

GBM là một thuật toán ML dựa trên cơ sở học tập theo nhóm và được sử dụng rộng rãi vì tính linh hoạt và hiệu quả. GBM hoạt động bằng cách xây dựng tuần tự nhiều mô hình yếu, thường là cây quyết định, nhưng khi kết hợp lại, chúng tạo thành một mô hình mạnh mẽ hơn. GBM được sử dụng rộng rãi trong cả phân loại và hồi quy. Các biến thể phổ biến của GBM bao gồm XGBoost, LightGBM (được Microsoft phát triển) và CatBoost.

Các bước thực hiện của phương pháp này tương tự như các bước chung của ML. Tuy nhiên, trong bước huấn luyện mô hình, GBM hoạt động bằng cách tối ưu hóa một hàm mất mát, thông qua việc sử dụng kỹ thuật tăng cường độ dốc. Trong mỗi vòng lặp, GBM cố gắng tối ưu hóa hàm mất mát bằng cách di chuyển theo hướng dốc âm của hàm mất mát.

Trong quá trình này, mỗi bộ phân loại yếu mới được thêm vào mô hình dự đoán mạnh để giảm thiểu hàm mất mát. Cuối cùng, các dự đoán của các tham số yếu được kết hợp lại với nhau dựa trên trung bình trọng số của chúng để đưa ra dự đoán cuối cùng.

3.6. Học máy tự động

ML tự động (AutoML) là tập hợp các công cụ và quy trình tự động hóa các bước trong việc phát triển mô hình ML. AutoML được xây dựng trong các thư viện của R hay Python. Tùy vào loại thư viện, nó thường được tích hợp sẵn các phương pháp ML phổ biến khác nhau như RF, GBM, XGBoost, CatBoost, và LightGBM,... Khi được gán các biến đầu vào và cài đặt các tham số phù hợp, AutoML sẽ tự động huấn luyện dữ liệu, điều chỉnh tham số và lựa chọn mô hình tối ưu nhất. Vì lẽ đó, AutoML rất thuận tiện cho người sử dụng, nhưng nhược điểm là tốn rất nhiều thời gian và tài nguyên tính toán vì phải quét qua nhiều phương pháp cùng lúc.

Trên đây là tóm tắt một số phương pháp phổ biến trong dự báo chuỗi thời gian. Ngoài ra, còn một số phương pháp ML khác và người đọc có thể tìm hiểu thêm trong các tài liệu của Hyndman and Athanasopoulos (2018) và Krispin (2019).

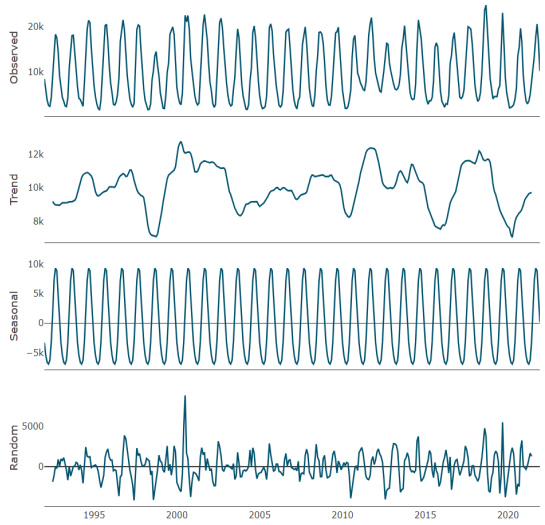
4. ỨNG DỤNG

4.1. Giới thiệu

Phần này trình bày các phân tích cơ bản và dự báo cho dữ liệu về lưu lượng nước hàng tháng tại trạm đo đạt trên sông Tiền, khu vực Tân Châu, An Giang, Việt Nam, từ tháng 1 năm 1992 đến tháng 12 năm 2021. Tân Châu là một thị xã nằm ở tỉnh An Giang, thuộc vùng đầu nguồn của Sông Tiền, một nhánh quan trọng của hệ thống sông Mekong. Với vị trí địa lý chiến lược, Tân Châu giữ vai trò quan trọng trong việc điều tiết và quản lý nguồn nước từ thượng nguồn đổ về. Do đó, phân tích và dự báo lưu lượng nước tại Tân Châu là việc làm thiết yếu để bảo vệ đời sống và nông nghiệp của người dân vùng hạ lưu và ĐBSCL. Sự biến đổi của dòng nước ảnh hưởng trực tiếp đến canh tác, thu hoạch và nguồn nước sinh hoạt của hàng triệu người ở hạ nguồn sông. Bằng cách dự báo chính xác, các nhà quản lý có thể phòng chống lũ lụt, hạn hán, và xâm nhập mặn hiệu quả, từ đó giảm thiểu thiệt hại và đảm bảo nguồn cung cấp nước ổn định. Điều này không chỉ giúp phát triển kinh tế bền vững mà còn nâng cao chất lượng cuộc sống của người dân trong khu vực.

4.2. Các thông tin cơ bản về dữ liệu

Hình 1 cung cấp tính trực quan hóa về các thành phần của chuỗi. Trong Hình 1, biểu đồ trên cùng chính là dữ liệu thời gian thực tế của chuỗi, qua đó ta có thể nhận xét rằng lưu lượng nước ở Tân Châu từ tháng 1 năm 1992 đến tháng 12 năm 2021 có tính mùa hàng năm rõ ràng. Dữ liệu tương đối ổn định trong giai đoạn đầu từ 1992 đến 2010, nhưng biến động nhiều hơn trong giai đoạn từ sau năm 2010. Sự thay đổi cấu trúc dữ liệu này có thể là một trở ngại không nhỏ khi áp dụng các mô hình dự báo bởi vì nó có thể làm tăng giá trị sai số của mô hình. Biểu đồ thứ 2 cung cấp tính xu hướng của dữ liệu thông qua phương pháp trung bình trượt, qua đó ta thấy rằng dữ liệu có xu hướng không rõ ràng và thay đổi liên tục. Biểu đồ ở hàng thứ 3 cho thấy tính mùa vụ hàng năm của dữ liệu là tương đối ổn định. Cuối cùng, biểu đồ Random cho biết yếu tố ngẫu nhiên hoặc bất thường của chuỗi. Các yếu tố này cũng có những biến đổi lớn ở năm 2000 và một số năm khác, điều này cho thấy có những yếu tố không thể dự đoán được đã ảnh hưởng đến lưu lượng nước.



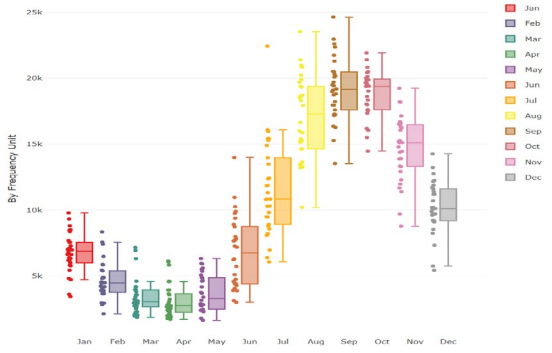
Hình 1. Biểu đồ các thành phần dữ liệu lưu lượng nước ở Tân Châu từ 1/1992 đến 12/2021

Bảng 1 và Hình 2 cung cấp các giá trị thống kê mô tả và biểu đồ hộp về lượng mưa theo từng tháng từ 1992 đến 2021. Trong Bảng 1, các giá trị được liệt kê ra bao gồm giá trị nhỏ nhất (Min), trung bình (Mean), trung vị (Median), độ lệch chuẩn (SD) và giá trị lớn nhất (Max) theo tháng của dữ liệu gồm 360 quan sát. Có thể thấy rằng lưu lượng nước trung bình ở Tân Châu thấp nhất vào tháng 4 (3.093 m^3/s) và cao nhất vào tháng 9 (19.189 m^3/s). Bên cạnh đó, Hình 2 cung cấp các thông tin trực quan hơn như ở Bảng 1 rằng lưu lượng nước ở Tân Châu phân bố theo mùa 12 tháng rõ rệt. Khi ở các tháng đầu năm, lưu lượng nước giảm xuống đến thấp nhất vào tháng 4, sau đó tăng liên tục đến đỉnh điểm vào tháng 9, và giảm dần ở các tháng cuối năm.

Bảng 1. Bảng thống kê mô tả dữ liệu theo tháng

Tháng	Min	Mean	Median	SD	Max
1	3.427	6.768	6.873	1.503	9.775
2	2.111	4.644	4.454	1.429	8.346
3	1.844	3.412	3.049	1.353	7.149
4	1.698	3.093	2.723	1.251	6.136
5	1.620	3.712	3.272	1.393	6.319
6	3.000	6.773	6.728	2.686	13.989
7	6.061	11.554	10.835	3.554	22.432
8	10.208	17.079	17.298	3.108	23.547
9	13.529	19.189	19.151	2.392	24.648
10	14.468	18.883	19.395	1.757	21.930
11	8.771	14.680	15.107	2.459	19.246
12	5.420	10.179	10.092	1.984	14.264

Ghi chú: Min: giá trị nhỏ nhất, Mean: trung bình, Median: trung vị, SD: độ lệch chuẩn, Max: giá trị lớn nhất.



Hình 2. Biểu đồ hộp lưu lượng nước theo tháng từ 1992 đến 2021

Từ các thông tin trên, ta có thể hiểu được quy luật về tính mùa của lưu lượng nước các tháng và đưa ra những dự báo cho các tháng tiếp theo.

4.3. Xây dựng mô hình dự báo

Phần này áp dụng các mô hình đã đề cập ở trên để huấn luyện mô hình và dự báo lưu lượng nước ở Tân Châu cho 12 tháng tiếp theo, gồm các bước sau đây:

Chuẩn bị dữ liệu: Để chuẩn bị dữ liệu đầu vào cho mô hình LR và các mô hình ML, biến Y được đặt là các giá trị về lưu lượng nước. Các biến độc lập $\boldsymbol{X} = \{X_1, X_2, X_3\}$, trong đó X_1 là biến bao gồm các tháng từ 1 đến 12, X_2 là lag_12 (độ trễ bậc 12), và X_3 là trung bình trượt bậc 3 về 1 phía. Lưu ý rằng, khi dữ liệu có tính mùa mạnh mẽ như các phân tích ở trên, việc sử dụng các biến X_1 và X_2 giúp các mô hình LR và ML nắm bắt được tốt hơn tính mùa vụ của dữ liệu, trong khi biến X_3 giúp mô hình nhận diện xu hướng, từ đó giúp nâng cao hiệu quả của các mô hình dự báo. Sau đó, dữ liệu ban đầu được tách thành tập huấn luyện và tập kiểm tra, trong đó tập kiểm tra bao gồm 72 quan sát cuối cùng (20% dữ liệu) và tập huấn luyện là phần còn lại của chuỗi.

Huấn luyện mô hình: Sử dụng các mô hình HW^+, HW^*, ARIMA, LR, RF, GBM, và AutoML để huấn luyện mô hình trên tập huấn luyện.

Kiểm tra và so sánh: Giá trị MAPE trên tập huấn luyện và kiểm tra được dùng để so sánh hiệu quả giữa các mô hình.

Cải thiện hiệu suất: Một số mô hình tiềm năng sau đó được áp dụng phương pháp tìm kiếm nâng cao, tức là huấn luyện mô hình với các tham số khác nhau, để tìm ra mô hình tối ưu nhất.

Dự báo: Mô hình tối ưu nhất sau đó được sử dụng để dự báo cho 12 tháng tiếp theo cho toàn bộ dữ liệu ban đầu.

Bảng 2 cung cấp giá trị MAPE cho tất cả mô hình trên hai tập huấn luyện và kiểm tra. Trước tiên, ta thấy rằng giá trị MAPE của 3 mô hình HW^+, HW^*, và ARIMA (mô hình SARIMA$(1,0,1)(1,1,0)_{12}$ thu được từ hàm *auto.arima()* trong R) là khá tương đồng trên tập huấn luyện, nhưng khác biệt đáng kể trên tập kiểm tra, trong đó ARIMA có giá trị MAPE thấp nhất trên cả 2 tập.

Tiếp theo, giá trị MAPE của mô hình LR trên 2 tập lần lượt là 10,12% và 15,77%, nhỏ hơn của mô hình ARIMA ở trên. Chú ý thêm rằng kết quả phân tích mô hình LR với biến phụ thuộc Y và biến độc lập X cho thấy các biến X_1, X_2 và X_3 đều có ý nghĩa thống kê trong mô hình với mức ý nghĩa 5%, hệ số R^2 hiệu chỉnh là 0,97% và thống kê F có $p_{value} < 0{,}001$, cho thấy mô hình hiệu quả cao trong việc xấp xỉ dữ liệu trên tập huấn luyện.

Sử dụng cùng dữ liệu đầu vào như mô hình LR, các mô hình RF, GBM và AutoML cho kết quả MAPE trên tập huấn luyện là 7,26%, 6,76% và 9,2%, và trên tập kiểm tra lần lượt là 17,02%, 14,55%, và 14,84%. Từ kết quả này ta thấy rằng GBM cho kết quả tốt nhất trong các phương pháp xem xét.

Tiếp theo, để cải thiện thêm hiệu quả tính toán, phương pháp tìm kiếm nâng cao cho 3 mô hình ARIMA, RF và GBM, gọi là ARIMA_grid, RF_grid, và GBM_grid đã được áp dụng. Với ARIMA_grid, mô hình ARIMA theo mùa được thiếp lập, SARIMA$(p,d,q)(P,D,Q)_{12}$, trong đó gán các giá trị hệ số $d = 0, D = 1$, và p, q, P, Q nhận các giá trị từ 0 đến 2, và cho máy tính quét qua hết tất cả các tổ hợp tham số để tìm ra mô hình tối ưu nhất.

Với mô hình RF_grid, một vài tham số được thiết lập, ví dụ như tỷ lệ mẫu lựa chọn cho mô hình rừng ngẫu nhiên là 0,95, 0,9 và 0,7, độ sâu tối đa là 1, 3 và 30, và số lượng cây là 5.000,... Cuối cùng, với mô hình GBM_grid, số lượng câu là 100, 200 và 500, độ sâu tối đa là 5, 10, 15, và tỷ lệ học lần lượt là 0,0001, 0,001, và 0,01, ...

Kết quả từ việc huấn luyện mô hình theo tìm kiếm nâng cao cho thấy sự cải thiện về hiệu suất, trong đó RF_grid có MAPE thấp nhất trên tập huấn luyện (4,76%) và GBM_grid có MAPE thấp nhất trên tập kiểm tra (13,05%).

Từ các kết quả ở Bảng 2 cho thấy, GBM_grid cung cấp kết quả có thể chưa phải là hoàn hảo nhưng nó là tốt nhất trong các mô hình xem xét, do đó nó được chọn để mô hình hóa cho toàn bộ dữ liệu và dự báo cho tương lai.

Bảng 2. Bảng giá trị MAPE (đơn vị: %) của các mô hình trên tập huấn luyện và kiểm tra

Mô hình	Tập huấn luyện	Tập kiểm tra
Huấn luyện mô hình theo truyền thống		
Holt-winters^{+}	13,08	61,95
Holt-winters*	14,45	29,80
ARIMA	12,93	25,98
LR	10,12	15,77
RF	7,26	17,02
GBM	6,76	14,55
AutoML	9,2	14,84
Huấn luyện mô hình với tìm kiếm nâng cao		
ARIMA_grid	10,89	23,84
RF_grid	4,76	16,86
GBM_grid	**7,29**	**13,05**

Bảng 3 và Hình 3 cung cấp dự báo lưu lượng nước ở Tân Châu 12 tháng tiếp theo. Qua đó có thể thấy rằng lưu lượng nước năm 2022 vẫn diễn ra theo như quy luật của những năm trước đó, thấp vào các tháng 2, 3 và 4, và cao vào các tháng 8, 9, và 10. Ngoài ra, Hình 3 cho thấy rằng kết quả dự báo (màu cam) khá phù hợp với dữ liệu thực tế (màu xanh).

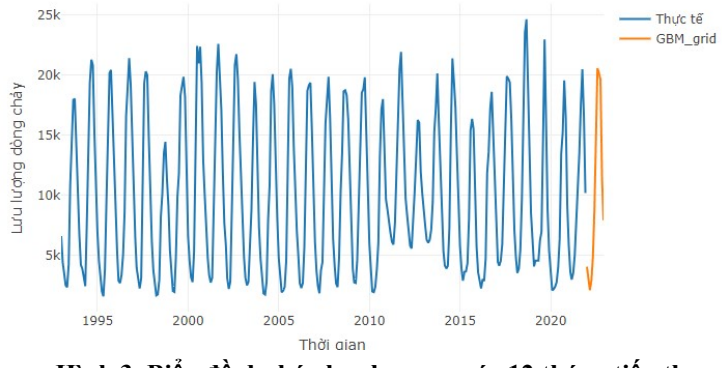
Bảng 3. Dự báo lưu lượng nước 12 tháng tiếp theo

Tháng	Giá trị dự báo
1	4.083,198
2	2.876,603
3	2.126,913
4	2.915,217
5	4.860,142
6	8.795,440
7	16.603,804
8	20.580,583
9	20.161,382
10	19.633,424
11	11.461,749
12	7.932,453

5. KẾT LUẬN

Nghiên cứu đã trình bày tóm tắt kiến thức liên quan đến một số phương pháp ML trong việc phân tích và dự báo chuỗi thời gian. Các phương pháp này đã được áp dụng vào bộ dữ liệu thực tế về lưu lượng nước ở Tân Châu và nhận được kết quả tốt nhất từ thuật toán GBM_grid với kỹ thuật tìm kiếm nâng cao.

Dựa vào những kết quả của nghiên cứu, có thể thấy rằng các phương pháp ML không chỉ cải thiện độ chính xác của dự báo mà còn có thể mở ra nhiều cơ hội ứng dụng trong các lĩnh vực khác nhau, đặc biệt là các vấn đề liên quan đến dự báo sự thay đổi của các yếu tố thời tiết và thiên tai. Ngoài ra, các phương pháp học sâu, một dạng phức tạp hơn của ML, rất tiềm năng trong dự báo dữ liệu chuỗi thời gian nên có thể được xem xét để áp dụng trong thời gian tới.



Hình 3. Biểu đồ dự báo lưu lượng nước 12 tháng tiếp theo

TÀI LIỆU THAM KHẢO

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, *29*(5), 1189-1232. https://www.jstor.org/stable/2699986

Hải, Đ. V., Huệ, L. T., & Trí, Đ. Q. (2020). Nghiên cứu ứng dụng mô hình hóa xây dựng phần mềm dự báo lũ, xâm nhập mặn sông Cửu Long hiển thị kết quả dự báo mặn lên Google Earth. *Tạp chí Khí tượng Thủy văn*, *710*, 33-42. DOI: 10.36335/VNJHM.2020(710).33-42

Hoài, N. P., Huyền, T. T. P., Nguyễn, L., Hiền, T. N., Thái, T. T., & Lâm, L. L. (2022). Đánh giá khả năng dự báo mặn trên sông Hàm Luông của thuật toán k-nearest neighbors. *Tạp chí Khoa học và Công nghệ Thủy lợi*, *74*, 1-9.

Holt, C. C. (2004). Forecasting seasonals and trends by exponentially weighted moving averages. *International journal of forecasting*, *20*(1), 5-10. https://doi.org/10.1016/j.ijforecast.2003.09.015

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts.

Krispin, R. (2019). *Hands-On Time Series Analysis with R: Perform time series analysis and forecasting using R*. Packt Publishing Ltd.

Lam, Đ. H., Phương, N. H., Đạt, N. Đ., & Giang, N. T. (2022). Xây dựng mô hình MIKE 11 phục vụ công tác dự báo thủy văn và xâm nhập mặn tỉnh Bến Tre. *Tạp chí Khí tượng Thủy văn*, *740*(1), 38-49. DOI:10.36335/VNJHM.2022 (740(1)).38-49

LeDell, E., & Poirier, S. (2020). H2o automl: Scalable automatic machine learning. In *Proceedings of the AutoML Workshop at ICML (Vol. 2020)*. San Diego, CA, USA: ICML.

Pham, N. H., Pham, B. Q., & Tran, T. T. (2022). Apply Machine Learning to Predict Saltwater Intrusion in the Ham Luong River, Ben Tre Province. *VNU Journal of Science: Earth and Environmental Sciences, 38*(3), 79-92. https://doi.org/10.25073/2588-1094/vnuees.4852

Toàn, C. H., Đông, P. N., Hoàng, T. H., Hải, T. C., & Hồng, V. N. (2020). Nghiên cứu dự báo xâm nhập mặn cho khu vực đồng bằng sông Cửu Long. *Hội nghị Khoa học Trường Đại học Khoa học Tự nhiên, ĐHQG-HCM lần 12.*

Tran, T. T., Nguyen, L. D., Hoai, P. N., Pham, Q. B., Huyen, P. T. T., Dong, N. P., ... & Hien, N. T. (2022). Long short-term memory (LSTM) neural networks for short-term water level prediction in Mekong river estuaries. *Songklanakarin Journal of Science & Technology*, *44*(4), 1057-1066. DOI: 10.14456/sjst-psu.2022.138

Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Management Science*, *6*(3), 324-342. https://www.jstor.org/stable/2627346