



DOI:10.22144/ctujos.2026.113

XÂY DỰNG MÔ HÌNH CHUỖI THỜI GIAN DỰA VÀO SỰ CẢI TIẾN CÁC MỐI QUAN HỆ MỜ VÀ MÔ HÌNH HỒI QUY VECTOR HỖ TRỢ

Đoàn Hoài Phong¹, Phạm Hoàng Trúc Linh¹, Châu Ngọc Thơ¹, Lê Đại Nghiệp², Trần Bạch Mai¹ và Võ Văn Tài^{1*}

¹Trường Khoa học Tự nhiên, Đại học Cần Thơ, Việt Nam

²Khoa Cơ bản, Trường Đại học Nam Cần Thơ, Việt Nam

*Tác giả liên hệ (Corresponding author): vvtai@ctu.edu.vn

Thông tin chung (Article Information)

Nhận bài (Received): 24/08/2025

Sửa bài (Revised): 18/09/2025

Duyệt đăng (Accepted): 23/05/2026

Title: Building time series model based on improved fuzzy relationships and support vector regression model

Author(s): Doan Hoai Phong¹, Pham Hoang Truc Linh¹, Chau Ngoc Tho, Le Dai Nghiep², Tran Bach Mai¹ and Vo Van Tai^{1*}

Affiliation(s): ¹College of Natural Science, Can Tho University, Viet Nam;

²Basis College, Nam Can Tho University, Viet Nam

TÓM TẮT

Mô hình chuỗi thời gian được xây dựng trong nghiên cứu dựa trên sự cải tiến dữ liệu đầu vào cho một mô hình học máy. Đầu tiên, chuỗi nguồn được chia thành các cụm với một số lượng thích hợp và những phần tử trong mỗi cụm được xác định một cách tự động căn cứ vào mức độ biến động của chuỗi. Kết quả này sau đó được sử dụng để xây dựng các mối quan hệ mờ giữa mỗi phần tử và các cụm để từ đó xây dựng nguyên tắc mờ hóa chuỗi. Chuỗi mờ hóa này sau đó được sử dụng làm dữ liệu đầu vào để áp dụng mô hình hồi quy vector hỗ trợ. Kết hợp những cải tiến này, một mô hình dự báo mới đã được xây dựng. Áp dụng trên cổ phiếu Công ty Cổ phần Dược Hậu Giang, mô hình đề nghị đã cho kết quả tốt, vượt qua nhiều mô hình khác, bao gồm các mô hình học máy và thống kê.

Từ khóa: Chuỗi thời gian, cổ phiếu, dữ liệu được mờ hóa, mô hình dự báo, phân tích cụm

ABSTRACT

This study develops a time series model based on improvements in the input data for a machine learning model. First, the original series is partitioned into clusters with an appropriate number, and the elements within each cluster are automatically determined according to the variability of the series. The results are then used to establish fuzzy relationships between each element and the clusters, from which fuzzy rules for the series are constructed. The fuzzified series is subsequently employed as input data for applying the support vector regression model. By integrating these enhancements, a new forecasting model has been built. Applied to the stock data of Hau Giang Pharmaceutical Joint Stock Company, the proposed model produced superior results, outperforming several other models, including both machine learning and statistical approaches.

Keywords: Cluster analysis, forecasting model, interpolated data, stock, time series

1. GIỚI THIỆU

Chuỗi thời gian (CTG) là kiểu dữ liệu được lưu trữ theo diễn thời gian như giờ, ngày, tháng, năm.... Trong sự phát triển của xã hội và nhiều lĩnh vực khác nhau, dữ liệu CTG thường được lưu trữ nên kiểu dữ liệu này rất phổ biến và đồ sộ trong thực tế (De Lima Silva et al., 2019; Nguyen & Vo, 2023). Trong phân tích dữ liệu CTG, có thể nói dự báo là hướng được các nhà nghiên cứu quan tâm nhiều nhất bởi vì nó là thông tin đầu vào không thể thiếu cho mọi kế hoạch và chiến lược hợp lý. Một dự báo đúng có thể mang lại nhiều lợi ích và một dự báo không tốt có thể gây ra nhiều thiệt hại (Alyousifi et al., 2021; Pham et al., 2022).

Trong xây dựng mô hình dự báo cho CTG, hai hướng phát triển thường được đề cập là thống kê và học máy. Dựa vào thống kê, mô hình hồi quy được xem là trường hợp đơn giản, được sử dụng phổ biến hiện nay (Sulandari et al., 2020; Pant & Kumar, 2022). Tuy nhiên, để xây dựng mô hình này, phải biết được dạng đồ thị của dữ liệu (Neto & De Carvalho, 2008; Cai et al., 2015; Muti & Yildiz, 2023). Đây là hạn chế rất lớn khi áp dụng mô hình này vào thực tế. Một hướng phát triển quan trọng theo hướng thống kê khi xây dựng mô hình dự báo cho CTG là những mô hình đặc trưng cho kiểu dữ liệu này, được gọi là mô hình chuỗi thời gian (MHCTG). MHCTG được xây dựng theo hai hướng mờ và không mờ. Theo hướng không mờ, mô hình tự hồi quy tích hợp trung bình trượt (ARIMA) được xem là mô hình phổ biến, được áp dụng rộng rãi trong thực tế hiện nay. Tuy nhiên, để áp dụng hiệu quả mô hình này, dữ liệu phải dừng và sai số là ồn trắng. Vì thực tế, dữ liệu không đáp ứng được điều này nên ARIMA cũng có nhiều hạn chế (Ospina et al., 2023; Vo & Le, 2022). Theo hướng mờ, nhiều mô hình khác nhau đã được đề xuất, bao gồm những mô hình mờ hóa dữ liệu nguồn (Phamtoan & Vovan, 2023) và dự báo trực tiếp (Abbasov & Mamedova, 2003; Vovan, 2019). Các mô hình mờ hóa tạo ra một chuỗi thời gian mới mà các phần tử có sự liên kết với nhau theo một nguyên tắc nào đó để làm dữ liệu đầu vào cho các mô hình dự báo trực tiếp. So với các mô hình không mờ, mô hình mờ được đánh giá có ưu điểm hơn vì nó không đòi hỏi điều kiện dữ liệu, có sự linh động và thích ứng với nhiều chuỗi thời gian hơn. Tuy nhiên, việc xác định các tham số cũng như mối quan hệ mờ ứng với từng tập dữ liệu cụ thể luôn là một thách thức đối với các mô hình này (Tinh, 2020; Nguyen & Vo, 2023).

Các MHCTG dựa vào học máy hiện đang là một hướng nghiên cứu và ứng dụng quan trọng, nhằm khắc phục những hạn chế của các phương pháp thống kê truyền thống. Nguyên tắc chung của các mô hình này là khai thác khả năng học từ dữ liệu để nhận diện mẫu, khám phá quan hệ tiềm ẩn và dự báo giá trị tương lai mà không cần quá nhiều giả định chặt chẽ về phân phối hay tính dừng của chuỗi. Trong số các mô hình học máy được áp dụng phổ biến cho CTG, có thể kể đến hồi quy vector hỗ trợ (SVR) (Rodriguez-Perez & Bajorath, 2022), mô hình K-Nearest Neighbors (KNN) (Henderi et al., 2021), Decision Tree (DT) (Seto et al., 2022) và Random Forest (RF) (Couronne et al., 2018). Các mô hình này có ưu điểm nổi bật là khả năng xử lý dữ liệu phi tuyến, dữ liệu lớn và phát hiện cấu trúc phức tạp, từ đó cải thiện độ chính xác dự báo. Tuy nhiên, chúng cũng tồn tại hạn chế nhất định như yêu cầu khối lượng dữ liệu huấn luyện lớn, tốn kém về tính toán, khó diễn giải kết quả và đôi khi dễ bị quá khớp.

Các phân tích trên cho thấy các MHCTG phát triển theo hướng thống kê và học máy đều có những ưu điểm và khuyết điểm chưa thể khắc phục cho đến hiện tại. Do đó, nghiên cứu này tập trung tận dụng những ưu điểm của chúng để kết hợp với nhau nhằm tạo ra một mô hình mới vừa tận dụng được sức mạnh dự báo của mô hình học máy, vừa nâng cao tính ổn định và khả năng giải thích của mô hình thống kê. Cụ thể, nghiên cứu cải tiến mô hình mờ hóa dữ liệu đầu vào để áp dụng hiệu quả cho mô hình học máy thích hợp.

2. THUẬT TOÁN ĐỀ NGHỊ

Cho CTG $X = \{x_1, x_2, \dots, x_n\}$. Mô hình dự báo đề nghị cho X bao gồm 6 bước như sau:

Bước 1. Chuẩn hoá X về thang đo 10 theo quy tắc:

$$y_i = 10 \cdot \frac{x_i}{\max\{x_i\}}, i = 1, 2, \dots, n. \quad (1)$$

Sau bước này, ta có chuỗi $Y = \{y_1, y_2, \dots, y_n\}$.

Bước 2. Chia Y thành các cụm với số lượng thích hợp bởi các bước:

Bước 2.1. Gọi $Z^{(t)} = \{z_1^{(t)}, z_2^{(t)}, \dots, z_n^{(t)}\}$ là tập các phần tử đại diện của các cụm tại vòng lặp t . Tại vòng lặp đầu tiên ($t = 0$), xem mỗi phần tử của Y là một cụm riêng biệt, ta có n cụm với các trọng tâm là chính các phần tử đó.

$$Z^{(0)} = \{z_1^{(0)}, z_2^{(0)}, \dots, z_n^{(0)}\} = Y = \{y_1, \dots, y_n\}.$$

Bước 2.2. Cập nhật trọng tâm của chùm tại vòng lặp $t + 1$ theo công thức

$$z_i^{(t+1)} = \frac{\sum_{i'=2}^n f(z_i^{(t)}, z_{i'}^{(t)}) \cdot z_{i'}^{(t)}}{\sum_{i'=2}^n f(z_i^{(t)}, z_{i'}^{(t)})}, \quad (2)$$

trong đó

$$f(z_i^{(t)}, z_{i'}^{(t)}) = \begin{cases} \exp\left(\frac{-|z_i^{(t)} - z_{i'}^{(t)}|}{\lambda}\right) & \text{khi } |z_i^{(t)} - z_{i'}^{(t)}| \leq d_s, \\ 0 & \text{khi } |z_i^{(t)} - z_{i'}^{(t)}| > d_s, \end{cases} \quad (3)$$

với

λ là hằng số,

d_s là trung bình khoảng cách giữa từng đôi các phần tử:

$$d_s = \frac{2}{n(n-1)} \sum_{i < l} |z_i^{(t)} - z_l^{(t)}|, \quad (4)$$

với $i, l \in \{0, 1, 2, \dots, n\}$.

Trong nghiên cứu này, chúng tôi khảo sát tham số λ để chọn giá trị phù hợp nhất dựa vào các tham số đánh giá mô hình.

Bước 2.3. Lặp lại Bước 2.2 cho đến khi tồn tại $t = t_0$ sao cho điều kiện sau được thỏa mãn:

$$\max_i \{|z_i^{(t_0)} - z_i^{(t_0+1)}|\} < \varepsilon, i = 1, 2, \dots, n, \quad (5)$$

trong đó: ε là một số dương nhỏ tùy ý mà nó được chọn là 0,001 trong nghiên cứu này. Sau bước này, nếu $Z^{(t_0)}$ có c phần tử thì tập Y được chia thành c chùm.

Bước 3. Xác định các phần tử cụ thể trong mỗi chùm $w_j, j = 1, 2, \dots, c$ theo nguyên tắc:

Những phần tử trong Z hội tụ về cùng một phần tử đại diện trong $Z^{(t_0)}$ được xếp vào cùng một chùm. Tính trọng tâm c_j của chùm $w_j, j = 1, 2, \dots, c$.

Bước 4. Xác định mối quan hệ mờ từ mỗi phần tử y_i tới chùm w_j theo các bước sau:

Bước 4.1. Khởi tạo ma trận phân vùng ban đầu $V^{(0)}$ với c hàng và n cột $V^{(0)} = [\mu_{ji}^{(0)}]_{c \times n}$, trong đó $\mu_{ij}^{(0)}$ được chọn ngẫu nhiên sao cho thỏa được điều kiện sau:

$$\mu_{ji}^{(0)} \in [0; 1], \sum_{j=1}^c \mu_{ji}^{(0)} = 1, 0 < \sum_{i=1}^n \mu_{ji}^{(0)} < n.$$

Bước 4.2. Tính các phần tử đại diện cho các chùm tại lần lặp $t, t = 0, 1, 2, \dots$, sử dụng công thức sau:

$$c_j^{(t)} = \frac{\sum_i \mu_{ji}^{(t)} z_i}{\sum_i \mu_{ji}^{(t)}}, 1 \leq j \leq c. \quad (6)$$

Bước 4.3. Cập nhật ma trận phân vùng mới $V^{(t+1)}, t = 0, 1, \dots$, với các phần tử được xác định bởi công thức

$$\mu_{ji}^{(t+1)} = \frac{|c_j^{(t)} - z_i|}{\sum_{h=1}^c |c_h^{(t)} - z_i|}, 1 \leq j \leq c, 1 \leq i \leq n, \quad (7)$$

Bước 4.4. Lặp lại Bước 4.3 cho đến khi điều kiện sau thỏa mãn:

$$\max_{ji} (|\mu_{ji}^{(t+1)} - \mu_{ji}^{(t)}|) < \varepsilon. \quad (8)$$

Khi Bước 4 dừng, chúng ta có được ma trận phân vùng $V^{(t+1)} = [\mu_{ji}^{(t+1)}]_{c \times n}$.

Bước 5. Dự báo

Bước 5.1. Dự báo y_i theo quy tắc sau:

$$\hat{y}_i = \sum_{j=1}^c \mu_{ji} \cdot c_j, i = 1, 2, \dots, n, \quad (9)$$

trong đó c_j là trọng tâm của chùm w_j được xác định từ Bước 3.

Bước 5.2. Từ giá trị của \hat{y}_i , dự báo x_i theo nguyên tắc:

$$\hat{x}_i = \frac{\hat{y}_i \cdot \max\{x_i\}}{10}. \quad (10)$$

Như vậy, sau Bước 5, ta thu được chuỗi mới là một phiên bản làm mờ của chuỗi ban đầu:

$$\hat{X} = \{\hat{x}_1, \hat{x}_2, \dots, \hat{x}_n\}.$$

Bước 6. Sử dụng mô hình hồi quy vector hỗ trợ (SVR) với dữ liệu đầu vào \hat{X} được thiết lập từ Bước 5 để thực hiện dự báo cho tương lai.

SVR là một trong những mô hình chuỗi thời gian dựa vào học máy, được sử dụng phổ biến ngày nay. So với các phương pháp học máy khác, SVR có những ưu điểm nổi bật như xử lý tốt các quan hệ phi tuyến thường gặp trong chuỗi thời gian tài chính, có tính thích ứng cao với các tập dữ liệu khác nhau và ít xảy ra hiện tượng quá khớp ngay cả khi dữ liệu ít (Rodriguez-Perez & Bajorath, 2022).

Thuật toán đề nghị có thể áp dụng cho dữ liệu thực tế bởi một chương trình được thiết lập trên phần mềm Python.

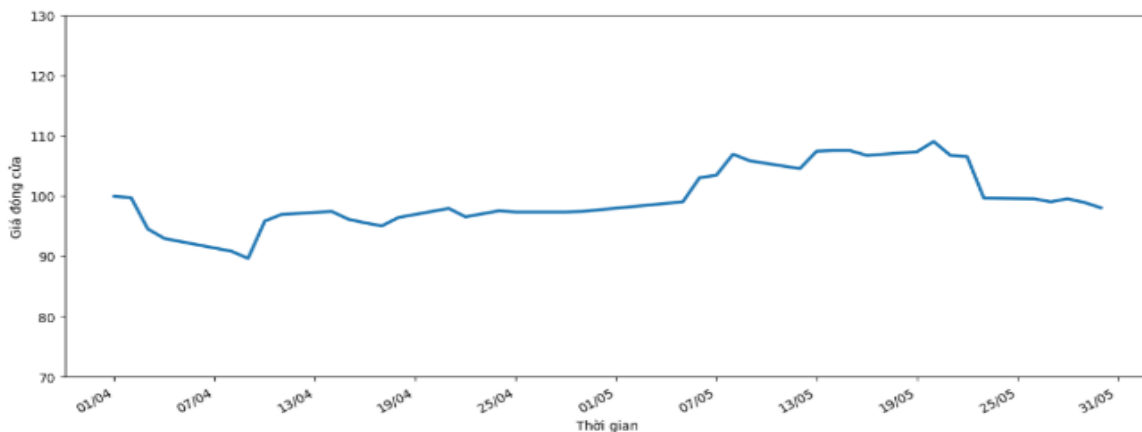
3. ỨNG DỤNG TRONG DỰ BÁO CỔ PHIẾU DHG

3.1. Dữ liệu và phương pháp thực hiện

Tập dữ liệu sử dụng trong nghiên cứu này được thu thập từ thị trường chứng khoán Việt Nam. Cụ thể là giá đóng cửa cổ phiếu của Công ty Cổ phần Dược Hậu Giang, có mã chứng khoán DHG. DHG là một trong những doanh nghiệp dược phẩm hàng

đầu tại Việt Nam, có lịch sử niêm yết lâu năm và được nhiều nhà đầu tư quan tâm.

Khoảng thời gian thu thập dữ liệu kéo dài từ ngày 01/04/2025 đến ngày 30/05/2025, bao gồm 40 phiên giao dịch. Bộ dữ liệu bao gồm thông tin về ngày giao dịch và giá đóng cửa tương ứng. Giá đóng cửa là mức giá được khớp tại thời điểm kết thúc phiên giao dịch và được sử dụng rộng rãi trong các phân tích tài chính. Bộ số liệu được lấy miễn phí từ trang website <https://s.cafef.vn/>, được trình bày chi tiết trong Bảng 1 và minh họa bởi Hình 1.



Hình 1. Đồ thị của cổ phiếu DHG từ ngày 01/4/2025 đến ngày 30/5/2025

Bảng 1. Số liệu cổ phiếu DHG từ ngày 01/4/2025 đến ngày 30/5/2025

Thời gian	X_i	Thời gian	X_i
01/04/2025	99,90	05/05/2025	99,00
02/04/2025	99,70	06/05/2025	103,00
03/04/2025	94,50	07/05/2025	103,40
04/04/2025	92,90	08/05/2025	106,90
08/04/2025	90,80	09/05/2025	105,80
09/04/2025	89,60	12/05/2025	104,50
10/04/2025	95,80	13/05/2025	107,40
11/04/2025	96,90	14/05/2025	107,50
14/04/2025	97,40	15/05/2025	107,50
15/04/2025	96,10	16/05/2025	106,70
16/04/2025	95,50	19/05/2025	107,30
17/04/2025	95,00	20/05/2025	109,00
18/04/2025	96,40	21/05/2025	106,70
21/04/2025	97,90	22/05/2025	106,50
22/04/2025	96,50	23/05/2025	99,60
23/04/2025	97,00	26/05/2025	99,50
24/04/2025	97,50	27/05/2025	99,00
25/04/2025	97,30	28/05/2025	99,50
28/04/2025	97,30	29/05/2025	98,90
29/04/2025	97,40	30/05/2025	98,00

Hình 1 cho thấy cổ phiếu DHG có xu hướng giảm từ ngày 01/4/2025 đến ngày 09/4/2025, sau đó nó có xu hướng tăng nhẹ đến ngày 20/5/2025, sau thời gian này nó lại có xu hướng giảm dần đều.

Từ dữ liệu được xem xét, các bước được thực hiện như sau:

Bước 1. Chia mỗi tập dữ liệu thành 2 phần, tập huấn luyện và tập kiểm tra với tỷ lệ lần lượt là 80% và 20%.

Bước 2. Xây dựng mô hình đề nghị trên tập huấn luyện và sử dụng tập huấn luyện để dự báo cho tập kiểm tra.

Bước 3. So sánh mô hình đề nghị và các mô hình quan trọng khác để lựa chọn mô hình tốt nhất căn cứ vào các tham số sau:

- Sai số tuyệt đối trung bình:

$$MAE = \frac{1}{n} \sum_{i=1}^n (|\hat{X}_i - X_i|). \quad (11)$$

- Sai số phần trăm tuyệt đối trung bình:

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left(\frac{|\hat{X}_i - X_i|}{X_i} \cdot 100 \right). \quad (12)$$

- Sai số bình phương trung bình:

$$MSE = \frac{1}{n} \sum_{i=1}^n (\hat{X}_i - X_i)^2. \quad (13)$$

Trong các mô hình dự báo, các tham số này càng nhỏ thì độ chính xác của mô hình càng cao và ngược lại.

Bảng 2. Tập huấn luyện và giá trị được chuẩn hoá

Chỉ số chùm	x_i	y_i	Chỉ số chùm	x_i	y_i
3	99,90	9,17	3	97,50	8,94
3	99,70	9,15	3	97,30	8,93
3	94,50	8,67	3	97,30	8,93
2	92,90	8,52	3	97,40	8,94
1	90,80	8,33	3	99,00	9,08
1	89,60	8,22	4	103,00	9,45
3	95,80	8,79	4	103,40	9,49
3	96,90	8,89	5	106,90	9,81
3	97,40	8,94	5	105,80	9,71
3	96,10	8,82	4	104,50	9,59
3	95,50	8,76	5	107,40	9,85
3	95,00	8,72	5	107,50	9,86
3	96,40	8,84	5	107,50	9,86
3	97,90	8,98	5	106,70	9,79
3	96,50	8,85	5	107,30	9,84
3	97,00	8,90	5	109,00	10,00

3.2. Kết quả thực hiện

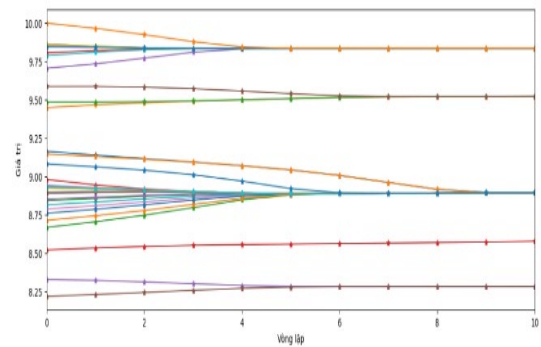
Mô hình đề nghị thực hiện cho tập dữ liệu DHG gồm các bước sau:

Bước 1. Tính các giá trị chuẩn hoá theo công thức (1), ta được giá trị đã chuẩn hoá y_i trong Bảng 2.

Bước 2. Với chuỗi y_i , công thức (2), (3), (4) được sử dụng để thỏa điều kiện (5), sau 10 lần lặp, thuật toán dừng lại. Khi đó, 32 phần tử của y_i hội tụ về 5 phần tử:

$$\{8,28 \ 8,52 \ 8,91 \ 9,51 \ 9,84\}.$$

Quá trình này được minh họa bởi Hình 2.



Hình 2. Sự hội tụ của Bước 2 qua 10 vòng lặp

Hình 2 cho thấy 32 phần tử ban đầu đã hội tụ về 5 phần tử. Do đó, ta chia tập Y thành 5 chùm.

Bước 3. Xác định các phần tử cụ thể trong mỗi chòm, ta có

$$w_1 = \{8,33; 8,22\}, w_2 = \{8,52\},$$

$$w_3 = \left\{ \begin{array}{l} 9,17; 9,15; 8,67; 8,79; 8,89; 8,94; 8,82; \\ 8,76; 8,72; 8,84; 8,98; 8,85; \\ 8,90; 8,94; 8,93; 8,94; 8,93; 9,08 \end{array} \right\},$$

$$w_4 = \{9,45; 9,49; 9,59\},$$

$$w_5 = \{9,81; 9,71; 9,85; 9,86; 9,86; 9,79; 9,84; 10,00\}.$$

Với 5 chòm đã được xác định, tính trọng tâm của mỗi chòm, ta nhận được cụ thể kết quả

$$c_1 = \{8,28\}, c_2 = \{8,52\}, c_3 = \{8,90\},$$

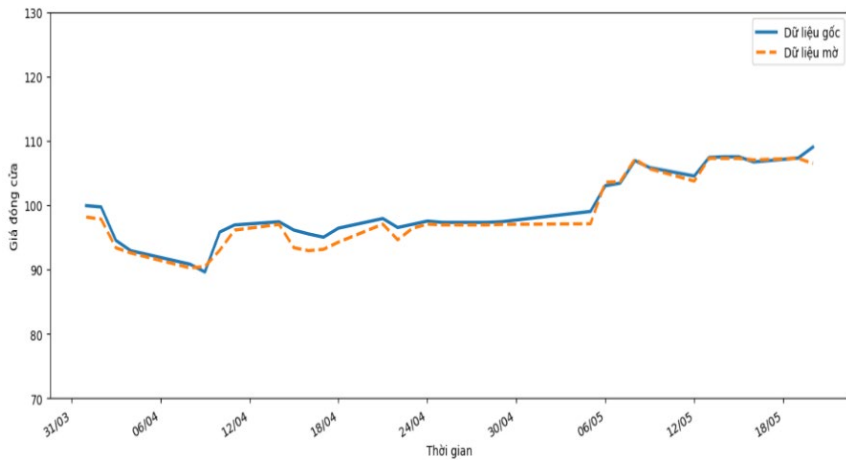
$$c_4 = \{9,51\}, c_5 = \{9,84\}.$$

Bước 4. Sau 27 vòng lặp của Bước 4, thuật toán dừng lại. Khi đó, ta xác định được mỗi quan hệ mờ $y_i, i = 1, 2, \dots, 32$ thuộc vào từng chòm $w_j,$

$j = 1, 2, \dots, 5$ bởi ma trận U như sau:

$$U = \begin{bmatrix} 0,032 & 0,030 & 0,061\dots & 0,001 & 0,007 & 0,004 \\ 0,150 & 0,150 & 0,825\dots & 0,001 & 0,015 & 0,010 \\ 0,556 & 0,605 & 0,096\dots & 0,002 & 0,023 & 0,017 \\ 0,211 & 0,171 & 0,012\dots & 0,016 & 0,300 & 0,870 \\ 0,051 & 0,044 & 0,006\dots & 0,980 & 0,656 & 0,099 \end{bmatrix}$$

Bước 5. Thực hiện mờ hóa cho Y và cho X lần lượt theo công thức (9) và (10), ta nhận được Bảng 3 và được minh họa bởi Hình 3.



Hình 3. Chuỗi dữ liệu gốc và dữ liệu mờ

Bảng 3. Kết quả mờ hoá chuỗi thời gian DHG

\hat{y}_i	\hat{x}_i	\hat{y}_i	\hat{x}_i
9,00	98,12	8,90	97,04
8,97	97,80	8,89	96,91
8,56	93,35	8,89	96,91
8,49	92,55	8,90	96,99
8,28	90,23	8,91	97,08
8,30	90,45	9,50	103,54
8,53	92,97	9,51	103,63
8,82	96,09	9,83	107,15
8,90	96,99	9,69	105,62
8,57	93,37	9,52	103,72
8,52	92,91	9,84	107,26
8,54	93,12	9,84	107,25
8,65	94,23	9,84	107,25
8,90	97,03	9,82	107,01
8,68	94,60	9,84	107,26
8,84	96,38	9,77	106,45

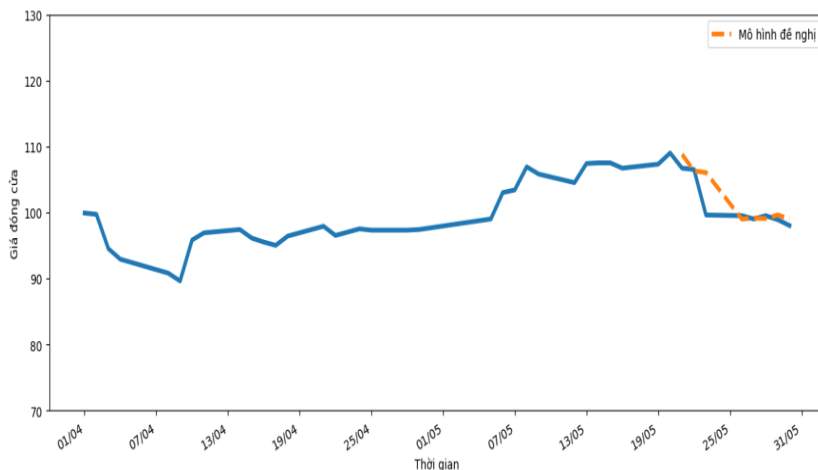
Kết quả được thể hiện ở Hình 3 cho thấy dữ liệu mờ hóa thích ứng với khuynh hướng của dữ liệu gốc nhưng có tính trơn hơn.

Bước 6. Áp dụng mô hình SVR để dự báo cho tập kiểm tra gồm 8 ngày, ta nhận được Bảng 4 và được minh họa như Hình 4.

Để đánh giá hiệu quả của mô hình đề nghị, nghiên cứu so sánh nó với các mô hình khác, bao gồm các mô hình dựa vào học máy và thống kê. Các mô hình này đã được nhiều nhà nghiên cứu quan tâm và đã được sử dụng phổ biến. Đó là các mô hình KNN, DT, RF, SVR, LR và ARIMA(0,1,0). Kết quả so sánh được trình bày trong Bảng 4.

Bảng 4. Kết quả dự báo của mô hình đề nghị

Ngày	21/05	22/05	23/05	26/05	27/05	28/05	29/05	30/05
Dữ liệu gốc	106,7	106,5	99,6	99,5	99,0	99,5	98,9	98,0
Dự báo	108,8	106,3	106,0	99,0	99,1	99,1	99,6	99,0



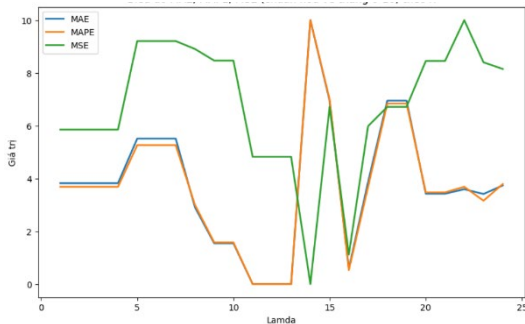
Hình 4. Kết quả dự báo DHG của mô hình đề nghị

Bảng 5. Các tham số của mô hình đề nghị và các mô hình được so sánh

Thực tế	ARIMA	RF	KNN	DT	LR	SVR	MHDN
106,70	109,00	107,37	107,60	107,50	109,25	109,02	108,78
106,50	109,00	107,37	107,58	106,70	106,88	106,33	106,28
99,60	109,00	107,36	107,58	109,00	106,72	106,36	106,04
99,50	109,00	105,02	106,98	107,40	99,82	98,90	98,99
99,00	109,00	103,76	106,12	107,40	99,94	99,56	99,14
99,50	109,00	102,72	98,84	106,90	99,55	99,08	99,06
98,90	109,00	102,72	98,84	106,90	100,07	99,68	99,61
98,00	109,00	102,72	98,84	106,90	99,45	98,98	99,00
MAE	8,038	3,918	3,264	6,375	1,747	1,572	1,441
MAPE	0,081	0,039	0,033	0,064	0,017	0,016	0,014
MSE	75,426	20,222	21,681	52,498	7,729	6,685	5,980

Kết quả thể hiện ở Bảng 3 cho thấy giá trị MAE của các mô hình từ 1,412 đến 6,357, trong đó mô hình có giá trị MAE cao nhất là DT và thấp nhất chính là mô hình đề nghị. Chúng ta cũng có kết quả hoàn toàn tương tự cho tham số MAPE và MSE. Như vậy, với tập dữ liệu này, mô hình đề nghị đã cho kết quả dự báo tốt nhất.

Kết quả trên được thực hiện với tham số $\lambda = 11$. Khảo sát các giá trị λ khác nhau và tính các tham số đánh giá MAE, MAPE và MSE tương ứng, sau đó chuẩn hóa cùng về thang đo 10, ta nhận được Hình 5.



Hình 5. Đồ thị của MAE, MAPE và MSE theo sự thay đổi của tham số λ

Kết quả được thể hiện tại Hình 5 cho thấy khi λ thay đổi từ 2 đến 11 cả 3 tham số MAE, MAPE và MSE đều có xu hướng giảm khá nhanh, sau đó chúng có xu hướng tăng. Như vậy, với $\lambda = 11$ thì giá trị MAE, MAPE và MSE có giá trị nhỏ nhất. Do đó, các kết quả trình bày ở trên đã thực hiện với tham số $\lambda = 11$ được xem là tối ưu.

4. KẾT LUẬN

Nghiên cứu này đã xây dựng được một mô hình dự báo cho chuỗi thời gian dựa trên sự cải tiến dữ

TÀI LIỆU THAM KHẢO

- Abbasov, A. M., & Mamedova, M. H. (2003). Application of fuzzy time series to population forecasting. *Vienna University of Technology*, 12, 545-552. <https://doi.org/10.1080/18756891.2013.808426>
- Alyousifi, Y., Othman, M., Husin, A., & Rathnayake, U. (2021). A new hybrid fuzzy time series model with an application to predict PM10 concentration. *Ecotoxicology and Environmental Safety*, 227, 112875. <https://doi.org/10.1016/j.ecoenv.2021.112875>
- Cai, Q., Zhang, D., Zheng, W., & Leung, S. C. (2015). A new fuzzy time series forecasting model combined with ant colony optimization and auto-regression. *Knowledge-Based Systems*, 74, 61-68. <https://doi.org/10.1016/j.knsys.2014.11.003>
- Couronne, R., Probst, P., & Boulesteix, A. L. (2018). Random forest versus logistic regression: a large-scale benchmark experiment. *BMC bioinformatics*, 19(1), 270. <https://doi.org/10.1186/s12859-018-2264-5>
- De Lima Silva, P. C., Sadaei, H. J., Ballini, R., & Guimarães, F. G. (2019). Probabilistic forecasting with fuzzy time series. *IEEE Transactions on Fuzzy Systems*, 28(8), 1771-1784. <https://doi.org/10.1109/TFUZZ.2019.2922152>
- Henderi, H., Wahyuningsih, T., & Rahwanto, E. (2021). Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer. *International Journal of Informatics and Information Systems*, 4(1), 13-20. <https://doi.org/10.47738/ijis.v4i1.73>
- Muti, S., & Yıldız, K. (2023). Using linear regression for used car price prediction. *International Journal of Computational and Experimental Science and Engineering*, 9(1), 11-16. <https://doi.org/10.22399/ijcesen.1070505>
- Neto, E. D. A. L., & De Carvalho, F. D. A. (2008). Centre and range method for fitting a linear regression model to symbolic interval data. *Computational Statistics & Data Analysis*, 52(3), 1500-1515. <https://doi.org/10.1016/j.csda.2007.04.014>
- Nguyen, L. H., & Vo, T. V. (2023). A new fuzzy time series forecasting model based on clustering technique and normal fuzzy function. *Knowledge and Information Systems*, 65(8), 3489-3509. <https://doi.org/10.1007/s10115-023-01875-w>
- Ospina, R., Gondim, J. A., Leiva, V., & Castro, C. (2023). An overview of forecast analysis with ARIMA models during the COVID-19 pandemic: Methodology and case study in

liệu đầu vào cho mô hình hồi quy vector hỗ trợ. Việc cải tiến dựa trên sự kết hợp của kỹ thuật phân tích chùm mờ tự động để xác định mối quan hệ mờ giữa các phần tử và phân tích chùm mờ để từ đó đề xuất nguyên tắc mờ hóa dữ liệu nguồn. Dữ liệu sau khi được mờ hóa có sự liên kết tốt hơn so với dữ liệu nguồn nên đã làm tăng hiệu quả dự báo cho mô hình học máy. Một chương trình trên phần mềm Python cho mô hình đề nghị để áp dụng cho số liệu thực đã được thiết lập. Mô hình đề nghị cho cổ phiếu DHG được áp dụng đã cho kết quả tốt hơn một số mô hình học máy và thống kê khác qua các tham số đánh giá.

Trong tương lai, nghiên cứu sẽ tập trung cải tiến giai đoạn mờ hóa dữ liệu nguồn theo các hướng khác nhau để có sự tối ưu và tương thích tốt hơn cho những chuỗi thời gian vốn rất đa dạng. Nghiên cứu cũng thử nghiệm nhiều mô học sâu khác nhau cho bước cuối cùng để cải thiện hiệu quả dự báo trong thực tế.

LỜI CẢM TẠ

Đề tài được tài trợ bởi Đại học Cần Thơ, mã số TSV2025-08.

- Brazil. *Mathematics*, 11(14), 3069.
<https://doi.org/10.3390/math11143069>
- Pant, M., & Kumar, S. (2022). Particle swarm optimization and intuitionistic fuzzy set-based novel method for fuzzy time series forecasting. *Granular Computing*, 7(2), 285-303.
<https://doi.org/10.1007/s41066-021-00265-3>
- Pham, D. T., Nguyen, D. T. H., & Vo, T. V. (2022). Improving the ANFIS forecasting model for time series based on the fuzzy cluster analysis algorithm. *International Journal of Fuzzy System Applications (IJFSA)*, 11(1), 1-20.
<https://doi.org/10.4018/IJFSA.313602>
- Pham, D. T., & Vo, T. V. (2023). The fuzzy cluster analysis for interval value using genetic algorithm and its application in image recognition. *Computational Statistics*, 38(1), 25-51.
<https://doi.org/10.1007/s00180-022-01215-6>
- Rodriguez-Perez, R., & Bajorath, J. (2022). Evolution of support vector machine and regression modeling in chemoinformatics and drug discovery. *Journal of Computer-Aided Molecular Design*, 36(5), 355-362.
<https://doi.org/10.1007/s10822-022-00442-9>
- Seto, H., Oyama, A., Kitora, S., Toki, H., Yamamoto, R., Kotoku, J. I., & Moriyama, T. (2022). Gradient boosting decision tree becomes more reliable than logistic regression in predicting probability for diabetes with big data. *Scientific reports*, 12(1), 15889.
<https://doi.org/10.1038/s41598-022-20149-z>
- Sulandari, W., Subanar, S., Lee, M. H., & Rodrigues, P. C. (2020). Time series forecasting using singular spectrum analysis, fuzzy systems and neural networks. *MethodsX*, 7, 101015.
<https://doi.org/10.1016/j.mex.2020.101015>
- Tinh, N. V. (2020). Enhanced forecasting accuracy of fuzzy time series model based on combined fuzzy C-mean clustering with particle swarm optimization. *International Journal of Computational Intelligence and Applications*, 19(2), 2050017.
<https://doi.org/10.1142/S1469026820500170>
- Vo, V., T. (2019). An improved fuzzy time series forecasting model using variations of data. *Fuzzy Optimization and Decision Making*, 18, 151-173.
<https://doi.org/10.1007/s10700-018-9290-7>
- Vo, V. T., & Le, T. T. T. (2022). A fuzzy time series model based on improved fuzzy function and cluster analysis problem. *Communications in Mathematics and Statistics*, 10(1), 51-66.
<https://doi.org/10.1007/s40304-019-0>