NGHIÊN CỨU ỨNG DỤNG KỸ THUẬT TRÍ TUỆ NHÂN TẠO TRONG BÀI TOÁN DỰ BÁO GIÁ MỘT SỐ MẶT HÀNG

Nguyễn Thái Sơn

Khoa Công nghệ thông tin, Trường Đại học Đại Nam, Hà Đông, Hà Nội, Việt Nam

ARTICLE INFORMATION

Journal: *Vinh University Journal of Science*
ISSN: *1859-2228*

Volume: 52
Issue: 3A
Correspondence:
thaison.nguyenn@gmail.com

Received: *04 August 2023*
Accepted: *17 August 2023*
Published: *20 September 2023*

Citation:
Nguyễn Thái Sơn (2023). Nghiên cứu ứng dụng kỹ thuật trí tuệ nhân tạo trong bài toán dự báo giá một số mặt hàng. ***Vinh Uni. J. Sci.*** *Vol. 52 (3A), pp. 116-138 doi: 10.56824/vujs.2023a082*

TÓM TẮT

Sự biến động của giá hàng hóa chủ chốt có tác động đáng kể đến nền kinh tế toàn cầu. Do đó, việc dự báo giá các mặt hàng này nhận được nhiều sự quan tâm từ các nhà nghiên cứu. Nghiên cứu này được triển khai nhằm phát triển các mô hình dựa trên các kỹ thuật trí tuệ nhân tạo trong một ngày tới dự báo giá thị trường cho các mặt hàng chủ chốt gồm đồng, dầu thô, khí đốt và bạc. Dữ liệu về giao dịch hàng hóa được thu thập từ 1/2000 đến 10/2019. Các mô hình dựa trên kỹ thuật trí tuệ nhân tạo khác nhau là hệ suy diễn mờ nơron thích nghi (ANFIS), mạng nơron nhân tạo (ANN), phương pháp xử lý dữ liệu nhóm (GMDH), mạng học sâu (LSTM) được phát triển. Các chỉ số đánh giá RMSE, MAPE, MAE, R và Theil's U được sử dụng. Kết quả cho thấy mô hình đề xuất dựa trên kỹ thuật GMDH vượt trội trong dự báo giá cả hàng hóa.

Từ khóa: Mặt hàng chủ chốt; Dự báo giá; GMDH; ANFIS; ANN; LSTM.

1. Mở đầu

Dự báo giá cả của một số mặt hàng và chỉ số chứng khoán được quan tâm nghiên cứu vì tính thực tiễn cao. Kết quả của dự báo chính là thông tin trong tương lai của một hiện tượng dựa trên các thông tin sẵn có như dữ liệu trong quá khứ và các thông tin sự kiện có thể ảnh hưởng đến hiện tượng đó. Dự báo dựa trên chuỗi thời gian (time series) là một trong những kỹ thuật phổ biến. Có thể được chia các phương pháp dự báo làm 02 nhóm chính: (1) Các phương pháp dự báo cổ điển dựa trên các lý thuyết về thống kê; (2) Các phương pháp dự báo hiện đại sử dụng kỹ thuật trí tuệ nhân tạo.

Đã có nhiều công trình chứng minh tính hiệu quả của dự báo khi sử dụng các phương pháp dự báo sử dụng kỹ thuật trí tuệ nhân tạo trong dự báo giá cả và chỉ số chứng khoán như Bakir và cộng sự [1] dự báo chỉ số thương mại điện tử, Behmiri và Pires Manso [2] dự báo giá dầu thô, Jeenanunta và cộng sự [3] dự báo giá cổ phiếu, Jubinski và Lipton [4] dự báo giá vàng, bạc và dầu, Kristjanpoller và Minutolo [5] dự báo giá vàng. Các nghiên cứu này đã xem xét các thao tác lựa chọn số đầu vào, tiền xử lý dữ liệu, thuật toán huấn luyện nhằm tăng khả năng xử lý và độ chính xác dự báo của mô hình.

Trong lĩnh vực nghiên cứu dự báo giá ở Việt Nam, có một số công trình nổi bật như: Huỳnh Quyết Thắng và cộng sự [6] dự báo chỉ số VNINDEX, Ngô Văn Toàn và cộng sự [7] dự báo giá vàng. Tuy nhiên, trong lĩnh vực dự báo giá ở Việt Nam, các kỹ thuật chỉ dừng lại ở việc thử nghiệm một số cấu trúc, kỹ thuật đơn lẻ để dự báo mà chưa có nghiên cứu nào thực hiện đánh giá và so sánh nhiều kỹ thuật với nhau. Bài báo này sẽ thực hiện so sánh khả năng dự báo ngắn hạn (dự báo ngày tiếp theo) giá cả hoặc chỉ số của một số mặt hàng. Đồng thời xem xét ảnh hưởng của các bộ dữ liệu khác nhau lên kết quả dự báo. Từ đó đánh giá tổng quan về khả năng ứng dụng trí tuệ nhân tạo vào dự báo giá cả và chỉ số của một số mặt hàng, đồng thời đưa ra một số khuyến nghị về việc sử dụng thực tế.

2. Tổng quan

2.1. Bài toán dự báo giá

Trong những năm gần đây, giá cả các mặt hàng chủ chốt có những biến động mạnh theo sự lên xuống của đồng USD, biến động tình hình chính trị của thế giới và thị trường tài chính. Kim loại quý, dầu thô được coi là giải pháp an toàn đối với các nhà đầu tư. Để thành công trong quá trình đầu tư vào một mặt hàng cần có phương pháp dự báo giá, là công việc đòi hỏi nhiều kỹ năng, kiến thức, kinh nghiệm cũng như tư duy [4].

Khoa học dự báo đã hình thành từ đầu những năm 60 của thế kỉ 20. Là một phần quan trọng trong hoạch định, khi tiến hành dự báo phải căn cứ vào số liệu thu thập được, xác định xu hướng vận động của các hiện tượng trong tương lai nhờ vào các mô hình toán học. Ngày nay, dự báo giá là một nhu cầu thiết yếu của hoạt động kinh tế [8] để giảm bớt rủi ro. Đồng thời kết quả dự báo là căn cứ để đưa ra các quyết định kịp thời, tạo ra lợi thế cạnh tranh.

Phương pháp dự báo chuỗi thời gian được sử dụng phổ biến, được tiếp cận bởi hai nhóm phương pháp chính [2]:

(1) Các phương pháp cổ điển dựa vào các kỹ thuật thống kê như Autoregressive Integrated Moving Average (ARIMA), Generalized Autoregressive Conditional Heteroskedasticity (GARCH) và Seasonal Autoregressive Integrated Moving Average (SARIMA).

(2) Các phương pháp dự báo hiện đại dựa trên vào các kỹ thuật trí tuệ nhân tạo như mạng nơron (ANN), máy vectơ hỗ trợ (SVM), học sâu (deep learning), …

Việc xây dựng các mô hình dự báo bằng các phương pháp cổ điển gặp nhiều khó khăn vì các mô hình này dựa vào các giả thuyết chặt chẽ và các phân bố xác suất [9] và các phương pháp cổ điển không thể tìm ra những mối quan hệ phi tuyến trong dự báo [10]. Nhiều nghiên cứu đã chỉ ra rằng các mô hình dự báo dựa trên kỹ thuật trí tuệ nhân tạo cho kết quả tốt hơn so với các mô hình dựa trên kĩ thuật thống kê [5, 11]. Một số ví dụ như, mạng nơron nhân tạo có khả năng dự báo với giá trị lỗi bình phương trung bình bằng một nửa so với các mô hình kinh tế lượng [12]. Tapia Cortez và cộng sự [13] đã chứng minh rằng các kỹ thuật học máy hiệu quả hơn trong việc dự báo giá quặng thô. Do đó, ứng dụng các kỹ thuật trí tuệ nhân tạo sẽ cho kết quả chính xác, ổn định hơn trong bài toán dự báo giá [14, 15].

Nghiên cứu này tập trung vào bài toán dự báo giá của một số mặt hàng chủ chốt trên thị trường quốc tế như giá của một số kim loại quý, dầu thô và gas. Các mô hình dự

báo giá sẽ được phát triển dựa trên một số kỹ thuật trí tuệ nhân tạo, được đánh giá độ chính xác dựa trên các chỉ tiêu $RMSE, MAPE, MAE, R$ và $Theil's\ U$.

2.2. Dự báo chuỗi thời gian

Phương pháp được xây dựng dựa trên sự phụ thuộc của các giá trị quan sát trong dãy số. Các yếu tố gồm tính xu hướng, tính mùa, tính chu kì và tính ngẫu nhiên có thể được kể đến là nguồn gốc tạo ra đặc tính dao động [16], cụ thể:

- Tính xu hướng: Thể hiện chiều hướng biến động, tăng hoặc giảm của hiện tượng nghiên cứu trong một thời gian dài.

- Tính mùa: Biến động ở một số thời điểm (tháng/quý) nào đó được lặp đi lặp lại qua nhiều năm. Kích thước ảnh hưởng mùa đều đặn hàng năm thì được gọi là có tính cộng (additive), còn kích thước của ảnh hưởng mùa tỉ lệ với giá trị trung bình được gọi là có tính nhân (multiplicative).

- Tính chu kỳ: Biến động có tính chất lặp đi lặp lại trong khoảng thời gian nhất định.

- Tính ngẫu nhiên: Biến động thất thường, hầu như không thể dự đoán.

Bốn thành phần trên có thể kết hợp với nhau theo mô hình nhân:

$$y_t = T_t.S_t.C_t.I_t \tag{1}$$

với: y_t là giá trị quan sát và T_t, S_t, C_t , I_t lần lượt là các thành phần xu hướng, thời vụ, chu kỳ, ngẫu nhiên ở thời điểm t.

Khi xem xét sự biến động của các thành phần trong chuỗi thời gian, chuỗi thời gian chỉ gồm hai thành phần: một phần xu hướng và một phần mang tính chu kỳ, vì phần mùa vụ và các thành phần ngẫu nhiên đã bị loại bỏ.

$$SI = \frac{TCSI}{TC} = \frac{y_t}{\overline{y_t}} \tag{2}$$

trong đó: $\overline{y_t}$ là số trung bình di động ứng với y_t.

Sau đó, xu hướng của chuỗi thời gian có yếu tố mùa vụ thì cần loại bỏ yếu tố mùa vụ ra khỏi dãy.

$$CTI = \frac{TCSI}{S} = \frac{y_t}{I_s} \tag{3}$$

trong đó I_s là chỉ số thời vụ.

Biến động của yếu tố ngẫu nhiên có thể được xác định theo cách sau:

$$I_t = \frac{y_t}{TI_sI_c} \tag{4}$$

trong đó: I_c là chỉ số chu kì.

Ngoài ra, chuỗi thời gian còn có các yếu tố khác như: tuyến tính và tính ổn định. Các phép sai phân thường được sử dụng để chuyển chuỗi thời gian không dừng về chuỗi dừng [17]. Ngoài ra, còn tồn tại hệ số tương quan. Với một chuỗi thời gian x, với giả thiết giá trị trung bình không đổi, ta có thể xấp xỉ hệ số tương quan giữa x_t và x_{t+k} như sau:

$$r_k = \frac{\sum_{t=1}^{N-k}(x_t-\bar{x})(x_{t+k}-\bar{x})}{\sum_{t=1}^{N}(x_t-\bar{x})^2} \tag{5}$$

trong đó, r_k là hệ số tương quan tại độ trễ k và $\bar{x}$ là giá trị trung bình của x.

2.3. Các phương pháp dự báo cổ điển

Các mô hình tự hồi quy (AR): Mô hình chuỗi thời gian tự hồi quy hoàn toàn có cấu trúc như sau:

$$Y_t = \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \cdots + \alpha_p Y_{t-P} + u_t \quad (6)$$

Trong đó Y_t là quan sát thứ 1 đối với biến phụ thuộc sau khi trừ đi giá trị trung bình của chính nó, u_t là thành phần sai số có động thái tốt có trung bình bằng 0 và phương sai không đổi và không tương quan với u_s nếu $t \neq s$ (nhiễu trắng). Thành phần hằng số được bỏ qua vì Y_t được biểu diễn dạng độ thiên lệch khỏi giá trị trung bình. Y_t đã được mô hình hoá chỉ với quá khứ của nó và không có các biến độc lập khác. Đây là các mô hình tự hồi quy, AR và mô hình trong phương trình (2) được gọi là mô hình AR(p), với p là bậc tự hồi quy.

Các mô hình trung bình trượt (MA): Mô hình trung bình trượt - MA bậc q, ký hiệu là MA(q) có dạng như sau:

$$Y_t = v_t - \beta_1 v_{t-1} - \beta_2 v_{t-2} - \cdots - \beta_q v_{t-q} \quad (7)$$

Với v_t là chuỗi sai số nhiễu trắng; Y_t là tổ hợp tuyến tính của các biến ngẫu nhiên nhiễu trắng.

Mô hình ARMA, phối hợp giữa các công thức tự hồi quy và trung bình trượt, có dạng tổng quát như sau:

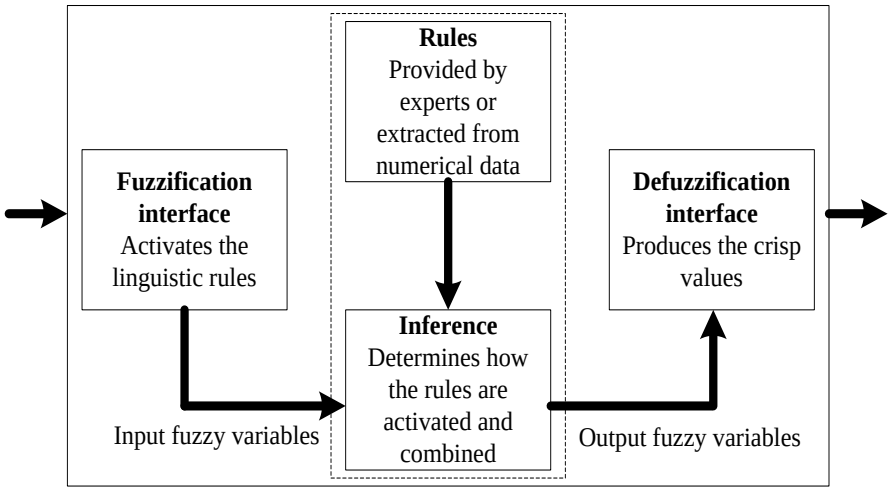
$$Y_t = \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \cdots + \alpha_p Y_{t-P} + u_t + v_t - \beta_1 v_{t-1} - \beta_2 v_{t-2} - \cdots - \beta_q v_{t-q} \quad (8)$$

3. Một số kỹ thuật trí tuệ nhân tạo trong dự báo

Một bài toán được xem là hồi quy nếu đầu ra (nhãn) không được chia thành các nhóm mà là một giá trị thực cụ thể (miền giá trị là liên tục). Các bài toán dự báo (giá cổ phiếu, giá nhà, …) được xếp vào bài toán hồi quy. Bài toán dự báo giá của một số mặt hàng là bài toán hồi quy với đầu ra là giá của ngày tiếp theo (là một giá trị liên tục). Trong bài báo này sử dụng ANFIS, ANN, GMDH và LSTM để dự báo giá của một số mặt hàng.

3.1. Hệ suy diễn mờ nơron thích nghi (ANFIS)

Mạng nơron và lý thuyết mờ là các công nghệ tính toán mềm được sử dụng để xây dựng các hệ thống thông minh. Hệ suy diễn mờ (FIS) sử dụng các luật mờ if-then khi thu thập tri thức từ các chuyên gia [18], được sử dụng rộng rãi trong nhiều lĩnh vực như tối ưu hóa, điều khiển và xác định hệ thống [19]. Một hệ suy diễn mờ FIS đơn giản được trình bày trong Hình 1. Tuy nhiên, hệ mờ không có khả năng tự học hỏi và điều chỉnh [20]. Trong khi đó, mạng nơron có khả năng học hỏi từ môi trường, tự tổ chức và thích nghi. Chính vì lý do đó, việc kết hợp hệ suy diễn mờ với mạng nơron trong một hệ nơron mờ được sử dụng để tạo ra các hệ cơ sở luật mờ. Trong khi nhiệm vụ của các quy tắc if-then mờ là mô hình hóa kiến thức chuyên gia, thì mạng nơron thần kinh tối ưu hóa các hàm thành viên để giảm thiểu tỷ lệ lỗi trong đầu ra. Trong các hệ nơron mờ, ANFIS, đề xuất bởi Jang, là phương pháp phổ biến nhất. Trong các hệ suy diễn mờ, các luật if-then mờ được xác định bởi chuyên gia thì trong ANFIS, nó được tự động tạo ra bởi dữ liệu đầu vào, đầu ra, và khả năng học của mạng nơron.



Hình 1*: Hệ suy diễn mờ [21]*

ANFIS là một mạng nơron truyền thẳng nhiều lớp. Một ANFIS bao gồm năm lớp và mỗi lớp được hình thành bởi một số nút và chức năng nút. Có hai loại nút: nút thích ứng và nút cố định. Các nút thích ứng được đánh dấu bằng các ô vuông đại diện cho các bộ tham số, có thể được điều chỉnh. Các nút cố định được đánh dấu bằng các vòng tròn và các bộ tham số của chúng được cố định trong hệ thống. Kiến trúc của ANFIS gồm năm lớp: lớp mờ hóa, lớp luật, lớp chuẩn hóa, lớp giải mờ và một nút tổng hợp. Giả thiết một ANFIS hai đầu vào, x và y, hai luật, và một đầu ra, f, như trong Hình 2. Mỗi nút trong cùng lớp thực hiện chức năng giống nhau. Một FIS có hai đầu vào và hai luật mờ if-then [22] được biểu diễn như sau:

Luật 1: $If\ x\ is\ A_1\ and\ y\ is\ B_1\ then\ f_1 = p_1x + q_1y + r_1$
Luật 2: $If\ x\ is\ A_2\ and\ y\ is\ B_2\ then\ f_2 = p_2x + q_2y + r_2$ (9)

Trong đó x và y là các đầu vào; A_1, A_2, B_1, B_2 là các biến ngôn ngữ; p_i, q_i, r_i, (i=1 or 2) là các tham số kết quả, được xác định trong quá trình huấn luyện; f_1 và f_2 là các giá trị đầu ra mờ. Công thức (2.2) thể hiện dạng 1 của luật mờ if-then trong đó đầu ra là một hàm tuyến tính. Giá trị đầu ra cũng có thể là một hằng số [23], như sau:

Luật 1: $If\ x\ is\ A_1\ and\ y\ is\ B_1\ then\ f_1 = C_1$
Luật 2: $If\ x\ is\ A_2\ and\ y\ is\ B_2\ then\ f_2 = C_2$ (10)

Trong đó C_i ($i = 1$ hoặc 2) là các giá trị hằng số. Công thức (10) chính là dạng 2 của luật mờ if-then.

Đối với các bài toán phức tạp thì dạng 2 được sử dụng để xác định mối quan hệ giữa đầu vào và đầu ra [24].

Chức năng của các lớp trong Hình 2 được giải thích như sau:

Lớp 1 - mờ hóa, chứa các nút thích nghi được đại diện bởi i. Các nút tạo các giá trị thành viên. Đầu ra của các nút được xác định qua công thức sau:

$$O_{1,i} = \mu A_i(x), i = 1,2 \quad \text{và} \quad O_{1,i} = \mu B_{i-2}(y), i = 3,4 \tag{11}$$

Với $O_{1,i}$ là đầu ra của nút i trong lớp 1, và $\mu A_i(x)$ và $\mu B_{i-2}(y)$ là các hàm thành viên mờ của A_i và B_{i-2}. Các hàm thành viên mờ có thể là dạng hình tam giác, hình thang, hàm Gaussian.

Lớp 2 - quy tắc, chứa các nút quy tắc hình tròn, có nhãn là Π. Một đầu ra đại diện cho một sản phẩm của các tín hiệu đầu vào. Nghĩa là, nút cố định nhận các đầu vào từ các nút thích nghi tương ứng, và mỗi giá trị đầu ra của nút biểu diễn cường độ của một quy tắc đã cho:

$$O_{2,i} = w_i = \mu A_i(x) \times \mu B_i(y), i = 1,2 \quad (12)$$

Với $O_{2,i}$ là đầu ra của nút i trong lớp 2 và w_i là cường độ một quy tắc.

Lớp 3 - chuẩn hóa, mỗi nút là một nút cố định hình tròn với số lượng các nút giống như trong lớp 2. Nút thứ i được tính là tỷ lệ của cường độ quy tắc của nút thứ i so với tổng tất cả các cường độ của các quy tắc:

$$O_{3,i} = \overline{w_i} = \frac{w_i}{w_1 + w_2}, i = 1,2 \quad (13)$$

Với $O_{3,i}$ là đầu ra của nút i trong lớp 3, và $\overline{w_i}$ là cường độ quy tắc được chuẩn hóa.

Lớp 4 - giải mờ, mỗi nút là một nút thích nghi hình vuông, có số lượng giống như lớp 3. Đầu ra từ mỗi nút là giá trị kết quả trọng số của một quy tắc nhất định:

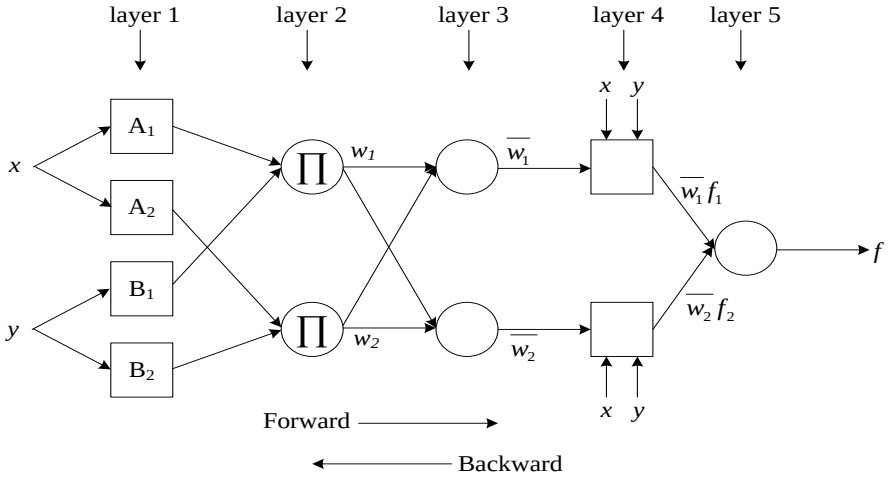
$$O_{4,i} = \overline{w_i} f_i, i = 1,2 \quad (14)$$

Trong đó $O_{4,i}$ là đầu ra của nút i trong lớp 4, $\overline{w_i}$ là đầu ra của lớp 3, với $\{p_i, q_i, r_i\}$ là tập các giá trị. Các giá trị trong lớp này được gọi là các giá trị kết quả của mô hình Sugeno mờ.

Lớp 5 - tổng kết, chỉ chứa một nút đầu ra, có hình tròn, được biểu thị là ∑, tính tổng sản lượng của ANFIS, là tổng của các đầu ra của tất cả các nút thích nghi trong lớp 4:

$$O_{5,i} = \sum_i \overline{w_i} f_i = \frac{\sum_i w_i f_i}{w_i}, i = 1,2 \quad (15)$$

Với $O_{5,i}$ là đầu ra của nút i trong lớp 5.



Hình 2: *Kiến trúc ANFIS với hai đầu vào và hai luật [25]*

ANFIS sử dụng một thuật toán học lai để hiệu chỉnh mạng. Việc kết hợp thuật toán hồi phục lại với thuật toán xấp xỉ hoặc thuật toán truyền lại được sử dụng trong thuật toán học lai ghép để tối ưu hóa các tham số trong các lớp 1 và 4. Thuật toán lai ghép là việc kết hợp hai phương pháp lan truyền ngược và phương pháp sai số bình phương nhỏ nhất.

Đối với mô hình mờ, mối quan hệ phi tuyến vào-ra phụ thuộc rất nhiều vào các phân vùng mờ của không gian vào-ra. Trong mạng nơron mờ việc điều chỉnh này có thể xem như là vấn đề tối ưu dùng giải thuật học để giải quyết. Đầu tiên giả định các hàm liên thuộc có một hình dạng nhất định. Sau đó ta thay đổi các thông số của hình dạng đó qua quá trình học (huấn luyện) bằng mạng nơron. Như vậy cần một tập dữ liệu ở dạng các cặp vào-ra mong muốn để cho mạng nơron học và cũng cần phải có một bảng các luật ban đầu dựa trên các hàm phụ thuộc đó [26].

3.2. Mạng nơron nhân tạo (ANN)

Là công cụ tính toán phổ biến trong lĩnh vực trí tuệ nhân tạo, có cấu trúc gồm một tập các đơn vị tính toán và được chia thành nhiều lớp như ví dụ Hình 3. Mức độ liên kết giữa các đơn vị được xác định bởi một tập giá trị trọng số. Tham số Bias (thiên vị) được sử dụng để tăng độ thích nghi của mạng với bài toán đặt ra. Số lớp và các đơn vị trong mỗi lớp phụ thuộc vào từng bài toán và được xác định bằng thử nghiệm. Số lượng đơn vị của lớp ra bằng số biến của vectơ lời giải.

Mạng nơron nhân tạo gồm có một nhóm các nơron nhân tạo (nút) nối với nhau, và xử lý thông tin bằng cách truyền theo các kết nối và tính giá trị mới tại các nút. Trong đó mạng perceptron nhiều lớp (Multilayer Perceptron -MLP), hay còn gọi là mạng truyền thẳng nhiều lớp, mở rộng của mô hình mạng perceptron, là mạng nơron nhân tạo được sử dụng phổ biến nhất, đặc biệt là mạng MLP có một lớp ẩn. Các nghiên cứu cho thấy rằng một mạng nơron truyền thẳng nhiều lớp với một lớp ẩn có thể xấp xỉ hóa tất cả các hàm số liên tục [27, 28], do đó được ứng dụng trong rất nhiều lĩnh vực [29]. Hình 3 là một mạng nơron truyền thẳng nhiều lớp gồm 3 lớp. Với R, N, và S là số lượng nút vào, nút ẩn và nút ra; iw và hw là các trọng số của nút vào và nút ẩn; hb và ob là các vectơ độ lệch bias của lớp ẩn và lớp ra; x là vectơ các đầu vào; ho là các vectơ đầu ra của lớp ẩn; và y là vectơ đầu ra. Mạng nơron trong Hình 3 được trình bày thông qua công thức sau:

$$h_{oi} = f\left(\sum_{j=1}^{R} iw_{i,j}.x_j + hb_i\right), \text{ với } j = 1,\dots,N \quad (16)$$

$$y_i = f\left(\sum_{k=1}^{N} hw_{i,k}.ho_k + ob_i\right), \text{ với } i = 1,\dots S \quad (17)$$

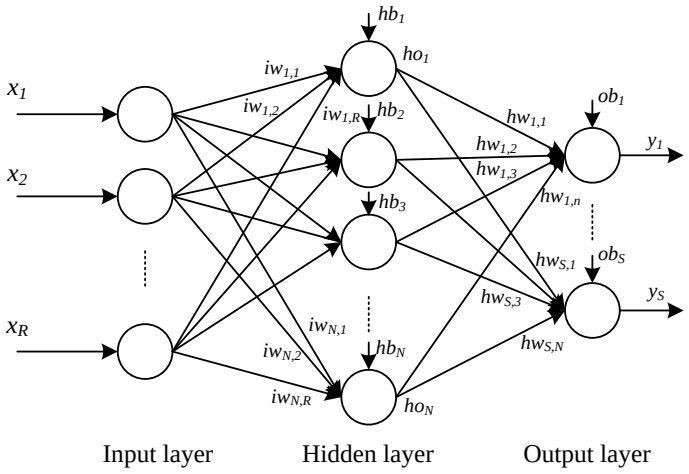
Trong đó, f là hàm kích hoạt (hàm chuyển)

Khi xây dựng một mô hình mạng nơron, cần phải xác định số lớp và số nút trong mỗi lớp. Một mạng có nhiều lớp và nút thì mạng sẽ phức tạp. Khi độ phức tạp của mô hình quá cao sẽ có hiện tượng quá khớp (overfiting), có thể dẫn đến việc dự đoán nhầm nhiễu, và chất lượng mô hình không còn tốt trên dữ liệu kiểm tra [30].

Chức năng của một mạng nơron được quyết định bởi cấu trúc mạng (số lớp, số nút trên mỗi lớp, liên kết giữa các lớp), các trọng số của các liên kết. Cấu trúc mạng thường cố định, và các trọng số được quyết định bởi các thuật toán huấn luyện. Tiến trình điều chỉnh các trọng số để mạng "nhận biết" được quan hệ giữa đầu vào và đích mong muốn được gọi là học (learning) hay huấn luyện (training). Nhiều thuật toán đã được áp dụng để tìm ra tập trọng số tối ưu làm giải pháp cho các bài toán, chia làm hai nhóm chính: học có giám sát (supervised learning) và học không có giám sát (unsupervised learning).

Học có giám sát là mạng được huấn luyện bằng cách cung cấp cho nó các cặp mẫu đầu vào và các đầu ra mong muốn (target values). Sự khác biệt giữa các đầu ra thực tế so với các đầu ra mong muốn được thuật toán sử dụng để thích ứng các trọng số trong mạng. Điều này thường được đưa ra như một bài toán xấp xỉ hàm số: cho dữ liệu huấn luyện bao

gồm các cặp mẫu đầu vào x, và một đích tương ứng t, mục đích là tìm ra hàm $f(x)$ thoả mãn tất cả các mẫu học đầu vào.



Hình 3: *Mạng nơron truyền thẳng ba lớp*

Để huấn luyện một mạng và xét xem nó thực hiện tốt đến đâu, ta cần xây dựng một hàm mục tiêu (cost function) để cung cấp cách thức đánh giá khả năng mô hình. Có một số hàm cơ bản được sử dụng như tổng bình phương lỗi (sum of squared error - SSE) và trung bình bình phương lỗi (mean squared error - MSE). Trong quá trình huấn luyện, sẽ đạt được phương án tối ưu hoặc gần tối ưu tương ứng với các vectơ trọng số và độ lệch. Giả thiết là có m cặp đầu vào và đầu ra mong muốn, x_k, t_k, với $k = 1,2, \dots m$. Trong quá trình huấn luyện, các giá trị iw, hw, hb, và ob sẽ được thay đổi để tối thiểu hóa hàm mục tiêu E, giả thiết E sử dụng hàm MSE sẽ được biểu diễn như sau:

$$MSE = \frac{1}{m}\sum_{k=1}^{m} e_k^2 = \frac{1}{m}\sum_{k=1}^{m}(t_k - y_k)^2 \quad (18)$$

Với y_k là đầu ra thực tế và t_k là đầu ra mong muốn

3.3. Phương pháp xử lý dữ liệu nhóm (GMDH)

GMDH, đề xuất bởi Ivakhnenko, là các mô hình toán học và thuật toán hồi quy phi tuyến [31], được ứng dụng chủ yếu trong khai phá dữ liệu, dự báo, tối ưu và nhận dạng [32-34]. GMDH có thể được biểu diễn dưới dạng một tập hợp các nơron trong đó liên kết giữa các cặp khác nhau trong mỗi lớp được thực hiện thông qua đa thức bậc hai, tạo ra các nơron mới ở lớp tiếp theo. Có thể biểu diễn mối quan hệ giữa các biến đầu ra và đầu vào dưới dạng rời rạc phức tạp bằng chuỗi Volterra.

$$y = a_0 + \sum_{i=1}^{n} a_i x_i + \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} x_i x_j + \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} a_{ijk} x_i x_j x_k + \cdots \quad (19)$$

Trong biểu thức trên, được biết đến là đa thức Kolmogorov- Gabor, $X = (x_1, x_2, \dots, x_r)$ là vectơ đầu vào, y là biến đầu ra, có thể được biểu diễn bằng một hệ thống đa thức bậc hai bán phần chỉ gồm hai biến (nơron) dưới dạng:

$$G(x_i, x_j) = a_0 + a_1 x_i + a_2 x_j + a_3 x_i x_j + a_4 x_i^2 + a_5 x_j^2 \quad (20)$$

Thuật toán GMDH gồm các bước chính sau:

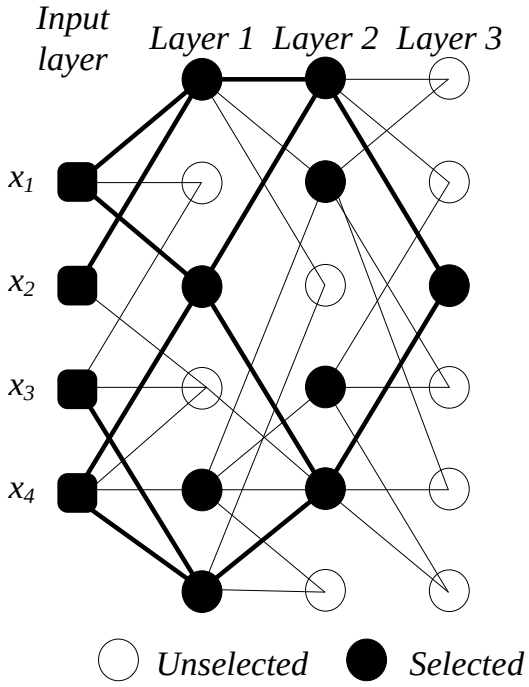
(1) Xác định tất cả nơron với đầu vào bao gồm tất cả các cặp giá trị biến đầu vào. Do đó, bao gồm $\frac{r(r-1)}{2}$ cặp (nơron).

(2) Dùng tập hợp số liệu xác nhận để chọn các nơron phù hợp nhất với tiêu chí lựa chọn.

(3) Nếu điều kiện dừng thỏa mãn (mạng phù hợp với dữ liệu với độ chính xác mong muốn hoặc việc giới thiệu các nơron mới không làm tăng đáng kể khả năng xấp xỉ của mạng nơron), sau đó sẽ cho dừng. Ngược lại, sử dụng đầu ra của các nơron phù hợp nhất để hình thành vectơ đầu vào cho lớp tiếp theo, rồi sau đó chuyển sang Bước 1.

Trong thuật toán GMDH, các lớp được xây dựng liên tiếp với các liên kết phức tạp là các mục của một đa thức. Lớp ban đầu chỉ đơn giản là lớp đầu vào, được thực hiện bằng cách hồi quy của các biến đầu vào và sau đó chọn các biến tốt nhất. Lớp thứ hai được tạo bằng cách tính hồi quy của các giá trị trong lớp đầu tiên cùng với các biến đầu vào, có nghĩa là GMDH về cơ bản xây dựng các đa thức của đa thức.

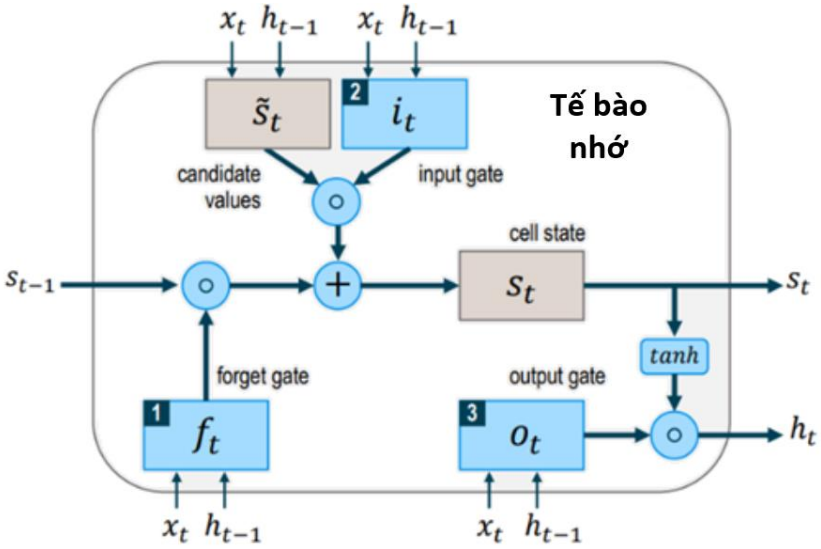
Số lượng mỗi nơron được xác định thông qua số đầu vào. Vì tất cả các kết hợp từng cặp của các đầu vào đều được xem xét nên số lượng nơron là $h = \binom{p}{2}$ với chuỗi thời gian được dịch chuyển thời gian về quá khứ p đơn vị để tạo ra các đầu vào. Một ví dụ về kiến trúc GMDH được trình bày trong Hình 4 gồm 03 lớp và 04 đầu vào. Vì có 4 đầu vào nên số lượng nơron ở mỗi lớp là 06. Tại mỗi nơron, giá trị đầu ra mong muốn được tính thông qua các hệ số và giá trị đầu vào. Dựa vào chỉ số lỗi bình phương trung bình (MSE), 4 nơron được chọn và hai nơron bị xóa khỏi mạng, những đầu ra từ các nơron được chọn sẽ thành đầu vào cho các nơron ở lớp kế tiếp. Các bước này được lặp đi lặp lại cho tới lớp cuối cùng. Tại lớp cuối cùng, chỉ một nơron duy nhất được chọn. Giá trị đầu ra của lớp cuối cùng là giá trị dự báo.



Hình 4: *Kiến trúc của GMDH*

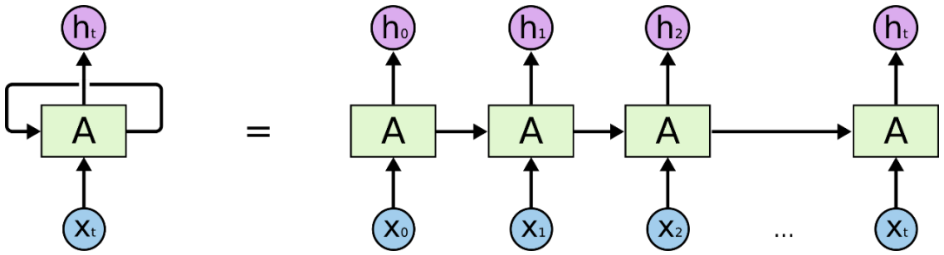
3.4. Mạng học sâu (Long Short Term Memory - LSTM)

Mạng học sâu có 2 mô hình lớn là Convolutional Neural Network (CNN) cho bài toán có input là ảnh và Recurrent neural network (RNN) cho bài toán dữ liệu dạng chuỗi, trong đó RNN chứa các vòng lặp bên trong cho phép lưu lại các thông tin.



Hình 5: *Cấu trúc RNN [35]*

Hình 5 mô tả cấu trúc RNN với đầu vào x_t và đầu ra h_t. Một vòng lặp cho phép thông tin có thể được truyền từ bước này qua bước khác của mạng nơron. Các vòng lặp khiến chúng tạo thành một chuỗi danh sách các mạng sao chép nhau. Mạng bộ nhớ dài-ngắn (Long short term memory networks), thường được gọi là LSTM - là một dạng đặc biệt của RNN, có giải quyết được các vấn đề phụ thuộc xa [36]. Mạng RNN cơ bản trong thực tế không có khả năng ghi nhớ thông tin từ các bước có khoảng cách xa và do đó những phần tử đầu tiên trong chuỗi đầu vào không có nhiều ảnh hưởng đến các kết quả tính toán dự đoán phần tử cho chuỗi đầu ra trong các bước sau.



Hình 6: *Kiến trúc bên trong của một tế bào LSTM [37]*

Mạng LSTM có thể bao gồm nhiều tế bào LSTM liên kết với nhau, có kiến trúc bên trong như Hình 6. Ý tưởng của LSTM là bổ sung thêm trạng thái bên trong của tế bào (cell internal state) s_t và ba cổng sàng lọc các thông tin đầu vào và đầu ra cho tế bào bao gồm forget gate f_t, input gate i_t, và output gate o_t. Tại thời điểm t, các cổng đều lần lượt nhận giá trị đầu vào x_t và giá trị h_{t-1} có được từ đầu ra của bước thời gian trước đó $t-1$. Thông tin được sàng lọc tại các cổng khác nhau, cụ thể:

- Forget gate: Loại bỏ những thông tin không cần thiết nhận được khỏi cell internal state.

- Input gate: Chọn lọc những thông tin cần thiết nào được thêm vào cell internal state.

- Output gate: Xác định những thông tin nào từ cell internal state được sử dụng như đầu ra.

Với:

$W_{f,x}, W_{f,h}, W_{\tilde{s},x}, W_{\tilde{s},h}, W_{i,x}, W_{i,h}, W_{o,x}, W_{o,h}$ là các ma trận trọng số trong mỗi tế bào LSTM.

$b_f, b_{\tilde{s}}, b_i, b_o$ là các vectơ bias.

$s_t, \tilde{s}$ lần lượt là các vectơ đại diện cho internal state và candidate value.

h_t là giá trị đầu ra của tế bào LSTM.

Trong quá trình lan truyền xuôi (forward pass), s_t và h_t được tính như sau:

Đầu tiên, tế bào LSTM quyết định những thông tin nào cần được loại bỏ từ cell internal state ở bước thời gian trước đó s_{t-1}. Giá trị f_t tại bước thời gian t được tính dựa trên x_t, giá trị h_{t-1} từ tế bào LSTM ở bước trước đó và bias b_f của forget gate. Hàm sigmoid function biến đổi tất cả các giá trị kích hoạt về miền có giá trị trong khoảng từ 0 (hoàn toàn quên) và 1 (hoàn toàn ghi nhớ):

$$f_t = \sigma\left(W_{f,x}\, x_t + W_{f,h}\, h_{t-1} + b_f\right) \tag{21}$$

Ở bước tiếp theo, tế bào LSTM quyết định những thông tin nào cần được thêm vào s_t. Bước này bao gồm hai quá trình tính toán đối với $\widetilde{s_t}$ và f_t. Giá trị $\widetilde{s_t}$ được tính như sau:

$$\widetilde{s_t} = tanh\left(W_{\tilde{s},x}\, x_t + W_{\tilde{s},h}\, h_{t-1} + b_{\tilde{s}}\right) \tag{22}$$

Giá trị kích hoạt i_t được tính như sau:

$$i_t = tanh\left(W_{i,x}\, x_t + W_{i,h}\, h_{t-1} + b_i\right) \tag{23}$$

Ở bước thứ ba, giá trị mới của cell internal state s_t được tính dựa trên kết quả tính toán thu được từ các bước trước với phép nhân Hadamard theo từng phần tử được ký hiệu bằng $\odot$.

$$s_t = f_t \odot s_{t-1} + i_t \odot \widetilde{s_t} \tag{24}$$

Giá trị đầu ra h_t được tính dựa trên hai phương trình sau:

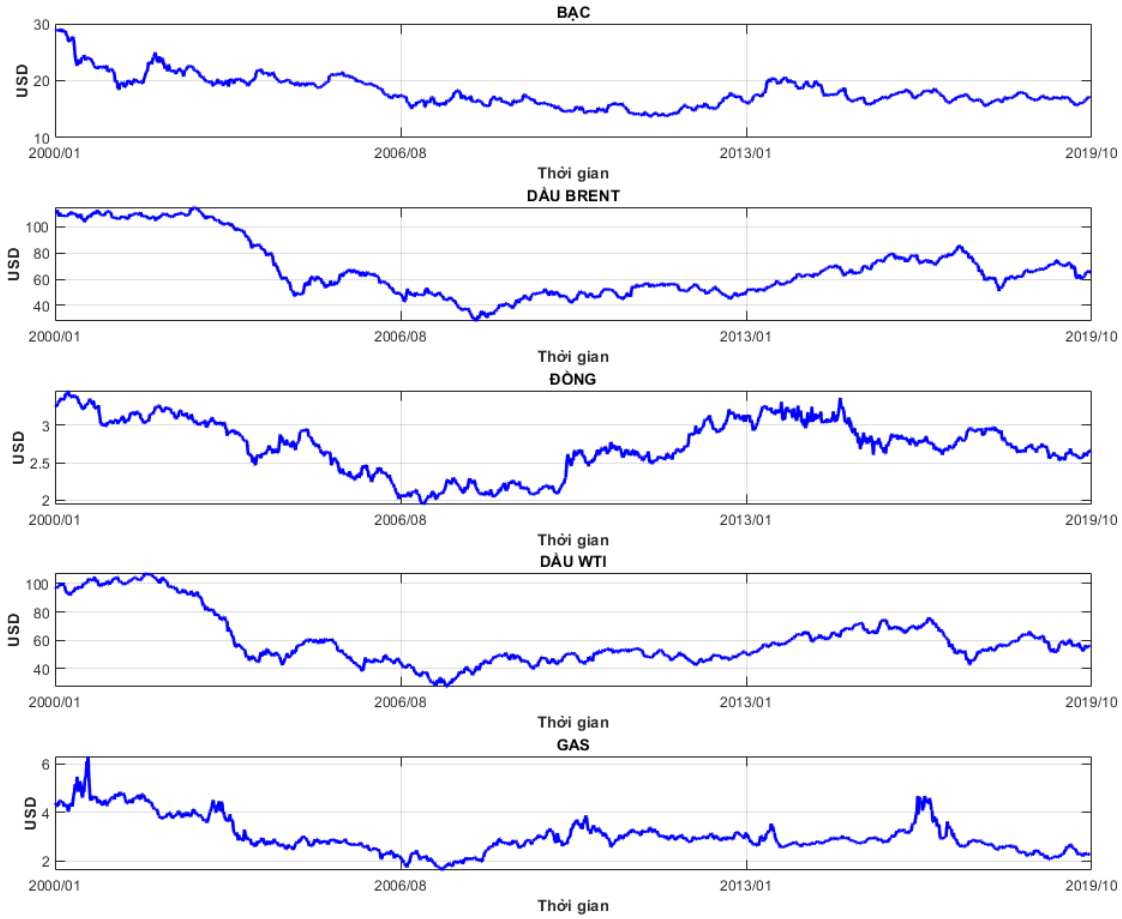
$$o_t = \sigma\left(W_{o,x}\, x_t + W_{o,h} h_{t-1} + b_o\right) \tag{25}$$

$$h_t = o_t \odot tanh\,(s_t) \tag{26}$$

4. Xây dựng mô hình dự báo

4.1. Thu thập dữ liệu

Dữ liệu sử dụng trong nghiên cứu là giá hàng ngày của một số mặt hàng chính trên thị trường bao gồm bạc, dầu thô (Brent and WTI), đồng và gas từ ngày 03/01/2001 đến ngày 30/09/2019, bao gồm 4995 mẫu. Dữ liệu được thu thập chủ yếu từ https://www.investing.com/, một nền tảng cung cấp dữ liệu, báo giá theo thời gian thực, biểu đồ, công cụ tài chính, tin nóng và các bài phân tích trên 250 sàn giao dịch trên khắp thế giới với 44 phiên bản ngôn ngữ. Sự biến động về giá trong giai đoạn dữ liệu nghiên cứu được thể hiện trong Hình 7.



Hình 7: *Giá hàng ngày của các mặt hàng từ ngày 03/01/2001 đến 30/09/2019*

Các dữ liệu lỗi hoặc bị mất (missing data) cần phải được xử lý, loại bỏ trước khi xây dựng mô hình để tránh ảnh hưởng đến độ chính xác của kết quả dự báo. Trong nghiên cứu này phương pháp được đề xuất bởi Grubbs [38] được sử dụng nhằm xác định và loại bỏ các số liệu quá lớn/quá khác so với phần còn lại. Đối với các dữ liệu bị mất, phương pháp nội suy từ các số liệu lân cận đã được sử dụng.

4.2. Xây dựng các mô hình dự báo

Các bước xây dựng mô hình được trình bày trong Hình 8, cụ thể:

Bước 1: Chuẩn bị dữ liệu

Bước này sẽ thực hiện thu thập dữ liệu và tiền xử lý dữ liệu. Dữ liệu sau khi được thu thập, cần phải được xử lý, làm sạch và biến đổi trước khi một kỹ thuật học máy có thể được huấn luyện trên những bộ dữ liệu này. Các kỹ thuật này bao gồm: xử lý dữ liệu bị khuyết, mã hóa các biến nhóm (encoding categorical variables), chuẩn hóa dữ liệu, …

Bước 2: Phân chia dữ liệu

Bước này chuẩn bị dữ liệu để xây dựng mô hình. Kiểm tra chéo (cross-validation) thường được sử dụng để chia tập dữ liệu thành hai phần, phục vụ cho huấn luyện (training

datasets) và kiểm tra mô hình (testing dataset). Trong nghiên cứu này, tác giả sử dụng 75% dữ liệu cho mục đích huấn luyện, phần còn lại 25% dữ liệu cho mục đích kiểm tra.

Bước 3: Xây dựng mô hình

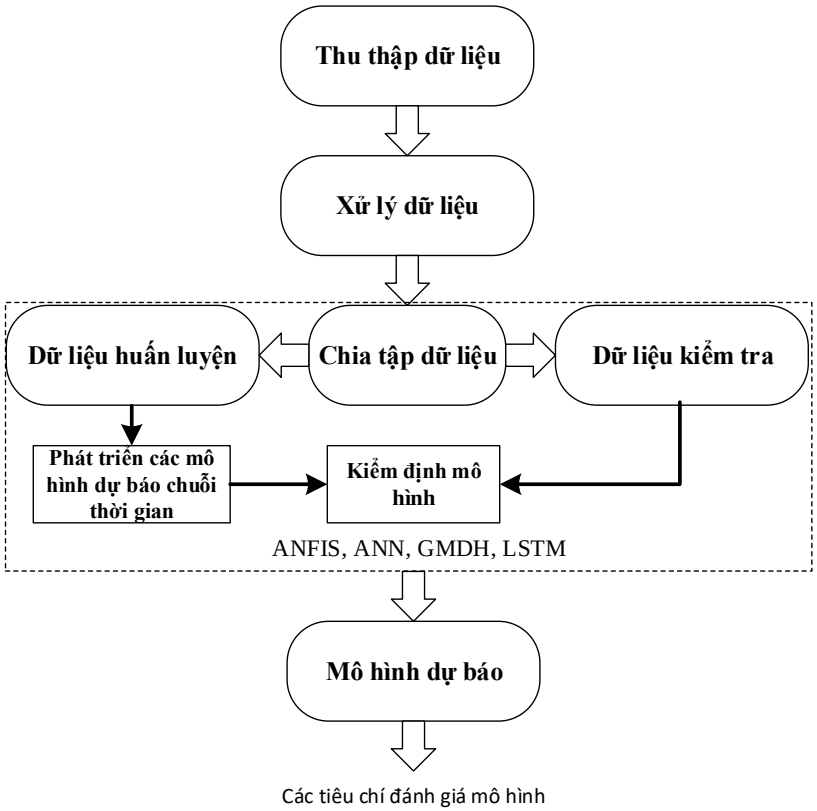
Mục đích của bước này là tìm ra hàm $f(x)$ và gán nhãn cho dữ liệu, thường được gọi là học hay training. Trong đó: x là các dữ liệu đầu vào, y là đầu ra của dự báo. Đối với bài toán dự báo chuỗi thời gian, nếu đầu ra $y = x[s]$, thì các đầu vào sẽ là $x[s-1]$, $x[s-2]$, …

Các thuật toán học giám sát (supervised learning) là ANFIS, ANN, GMDH, LSTM đã được sử dụng.

Bước 4: Kiểm tra: Các dữ liệu mới sẽ được đưa vào để kiểm tra, đánh giá.

Bước 5: Đánh giá và chọn ra mô hình tốt nhất

Việc đánh giá được thực hiện thông qua các chỉ số lỗi tìm được trên tập dữ liệu. Nếu không đạt được kết quả mong muốn thì các tham số (turning parameter) của các thuật toán phải được thay đổi để tìm ra các mô hình tốt hơn và thực hiện kiểm tra, đánh giá lại. Cuối cùng chọn ra mô hình dự báo tốt nhất.



Hình 8: *Các bước xây dựng mô hình*

Các tham số của mô hình

Đối với mô hình ANFIS, phương pháp phân chia lưới (grid partition) được sử dụng để tạo ra hệ suy diễn mờ FIS (Fuzzy Inference System). Hàm thành viên Gaussian được sử dụng trong mô hình ANFIS. Hàm thành viên đầu ra là tuyến tính. Số lượng luật fuzzy là 16, giá trị learning rate là 0.01.

Đối với mô hình ANN, mạng nơron truyền thẳng nhiều lớp với một lớp ẩn được sử dụng. Lớp ẩn có 12 units, hàm kích hoạt là ReLU (Rectifier Linear Unit) được sử dụng để tăng tốc độ tính toán. Hàm chi phí là MSE.

Đối với mô hình GMDH, số lớp tối đa là 03, số nơron tối đa trong mỗi lớp là 25. Trong mô hình GMDH, tham số giới hạn lựa chọn (Selection pressure) là giá trị ngưỡng để xác định số lượng nơron của mỗi lớp. Sau khi tính các giá trị chỉ số của các nơron, nơron nào có hiệu quả thấp nhất (giá trị MSE) sẽ bị loại bỏ khỏi lớp. Tham số giới hạn lựa chọn nằm trong khoảng từ 0 đến 1.

Đối với mô hình dựa trên LSTM, số lượng lớp là 03. Kiến trúc tuần tự (sequential structure) được sử dụng. Một số tham số khác như sau: số unit lớp ẩn là 50; hàm ReLU (Rectified Linear Unit) được sử dụng là hàm kích hoạt; hàm chi phí (cost function) là MSE.

Các chỉ số đánh giá mô hình

Sai số dự báo là chênh lệch giữa giá trị thực và giá trị dự báo nhằm đánh giá chất lượng hay sự phù hợp của mô hình dự báo tại cùng một thời điểm. Sai số dự báo cũng là căn cứ để thực hiện việc điều chỉnh mô hình dự báo.

Căn của sai số bình phương trung bình (Root Mean Squared Error- RMSE):

$$RMSE = \sqrt{\frac{1}{m}\sum_{k=1}^{m}(t_k - y_k)^2} \quad (27)$$

Với t_k là giá trị mong muốn, y_k là giá trị dự báo của mô hình, m là tổng số mẫu.

Sai số tương đối trung bình (Mean Absolute Percent Error- MAPE)

$$MAPE = \frac{1}{m}\sum_{k=1}^{m}\left|\frac{t_k - y_k}{t_k}\right| \quad (28)$$

Sai số tuyệt đối trung bình MAE (Mean Absolute Error)

$$MAE = \frac{1}{m}\sum_{k=1}^{m}|t_k - y_k| \quad (29)$$

Các chỉ số MAE và MSE và RMSE có đặc tính, công năng như nhau và thường cho cùng một kết quả khi đánh giá. Tuy nhiên, nếu giá trị sai số $\varepsilon_t = t_k - y_t$ đều nhau thì nên chọn MSE để đánh giá. Ngược lại, nếu giá trị sai số ε_t quá khác biệt, MAE nên được lựa chọn. Tiêu chí RMSE là căn bậc 2 của tiêu chí MSE nên hai tiêu chí về bản chất là một; điều khác biệt là giá trị của tiêu chí RMSE bé hơn.

Tiêu chí MAPE giúp đánh giá sai số một cách tương đối, do đó thường được áp dụng khi đánh giá sai số dự báo với các bộ số liệu khác nhau. Ngược lại, với cùng một bộ số liệu nhưng áp dụng nhiều phương pháp dự báo khác nhau thì không nên áp dụng tiêu chí MAPE vì tính phức tạp trong tính toán.

Hệ số tương quan R: Có giá trị từ -1 đến 1, được dùng để đo lường mức độ phụ thuộc tuyến tính giữa giá trị thực tế và giá trị dự báo. Hệ số tương quan bằng 0 (hay gần 0) có nghĩa là không có liên hệ giữa hai biến số; ngược lại nếu bằng -1 hay 1 có nghĩa là giữa giá trị thực tế và giá trị dự báo có một mối liên hệ tuyệt đối. Nếu $R < 0$ có nghĩa là khi t tăng cao thì y giảm và ngược lại; nếu R > 0 có nghĩa là khi t tăng cao thì y cũng tăng, và khi t giảm cao thì y cũng giảm theo.

$$R = \frac{\sum_{k=1}^{m}(t_k - \bar{t})(y_k - \bar{y})}{\sqrt{\sum_{k=1}^{m}(t_k - \bar{t})^2 . \sum_{k=1}^{m}(y_k - \bar{y})^2}} \quad (30)$$

Với $\bar{t} = \frac{1}{m}\sum_{k=1}^{m} t_k$ và $\bar{y} = \frac{1}{m}\sum_{k=1}^{m} y_k$.

Theil's U: Hệ số này được sử dụng để so sánh các mô hình dự báo, công thức như sau:

$$U = \frac{\sqrt{\sum_{k=1}^{m}(t_k - y_k)^2}}{\sum_{k=1}^{m} t_k^2 + \sum_{k=1}^{m} y_k^2} \quad (31)$$

Giá trị U nằm trong khoảng từ 0 đến 1, U càng tiến về 0 thì mô hình dự báo càng chính xác.

4.3. Kết quả và thảo luận

Các chỉ số đánh giá mô hình được trình bày trong Bảng 1. Các giá trị $RMSE$, $MAPE$, MAE, R và $Theil's\ U$ của mô hình GMDH tương ứng là 0,0368; 0,0098; 0,0265; 0,9949 và 0,0067 đối với giá đồng; 1,1430; 0,0143; 0,8644; 0,9987 và 0,0080 đối với giá dầu thô Brent; 0,0962; 0,0202; 0,0626; 0,9913 và 0,0154 đối với giá gas; 0,2674; 0,0101; 0,1843; 0,9955 và 0,0073 đối với giá bạc; 1,1226; 0,0154; 0,8392; 0,9982 và 0,0089 đối với giá dầu thô WTI. Mô hình được đánh giá là tốt khi các giá trị $RMSE$, $MAPE$, và MAE nhỏ, R gần giá trị 1 và $Theil's\ U$ gần giá trị 0. Trong Bảng 1, các giá trị tốt nhất đối với mỗi chỉ số được in đậm và nghiêng. Dễ dàng nhận thấy mô hình dự báo dựa trên GMDH là mô hình tốt nhất (04 tiêu chí tốt nhất trong 05 tiêu chí). Hầu hết các mô hình dựa trên kỹ thuật trí tuệ nhân tạo đều cho kết quả chấp nhận được.

Bảng 1: *Các chỉ số đánh giá các mô hình dự báo*

Loại hàng	Mô hình	RMSE	MAPE	MAE	R	Theil's U
Đồng	ANFIS	0,0405	0,0110	0,0297	***0,9942***	0,0073
	ANN	0,0376	0,0100	0,0271	0,9948	0,0068
	GMDH	***0,0368***	***0,0098***	***0,0265***	0,9949	***0,0067***
	LSTM	0,0378	0,0101	0,0274	0,9948	0,0069
Dầu thô Brent	ANFIS	1,1780	0,0145	0,8775	0,9986	0,0083
	ANN	1,1485	0,0144	0,8664	***0,9987***	0,0081
	GMDH	1,1430	***0,0143***	***0,8644***	***0,9987***	***0,0080***
	LSTM	***1,2169***	0,0156	0,9407	***0,9987***	0,0086
Gas	ANFIS	0,1725	0,0507	0,1482	***0,9908***	0,0271
	ANN	***0,0962***	0,0206	0,0636	0,9914	***0,0154***
	GMDH	***0,0962***	***0,0202***	***0,0626***	0,9913	***0,0154***
	LSTM	0,1725	0,0545	0,1482	0,9908	0,0271
Bạc	ANFIS	0,2798	0,0106	0,1941	***0,9951***	0,0076
	ANN	0,2729	0,0102	0,1866	***0,9951***	0,0075
	GMDH	***0,2674***	***0,0101***	***0,1843***	0,9955	***0,0073***
	LSTM	0,2833	0,0106	0,1950	0,9953	0,0077

Loại hàng	Mô hình	RMSE	MAPE	MAE	R	Theil's U
Dầu thô WTI	ANFIS	1,1609	0,0157	0,8601	***0,9981***	0,0092
	ANN	1,1236	0,0155	0,8428	0,9982	***0,0089***
	GMDH	***1,1226***	***0,0154***	***0,8392***	0,9982	***0,0089***
	LSTM	1,1231	***0,0154***	0,8421	0,9982	***0,0089***

Các Hình 9 đến Hình 13 thể hiện mức độ phù hợp giữa giá trị thực tế và giá trị dự báo bởi mô hình GMDH. Hình 9a - 13a là các giá trị thực tế và giá trị dự báo tương đối khớp nhau. Giá trị lỗi được thể hiện trong Hình 9b - 13b. Biểu đồ tần suất lỗi được thể hiện trong Hình 9c - 13c.

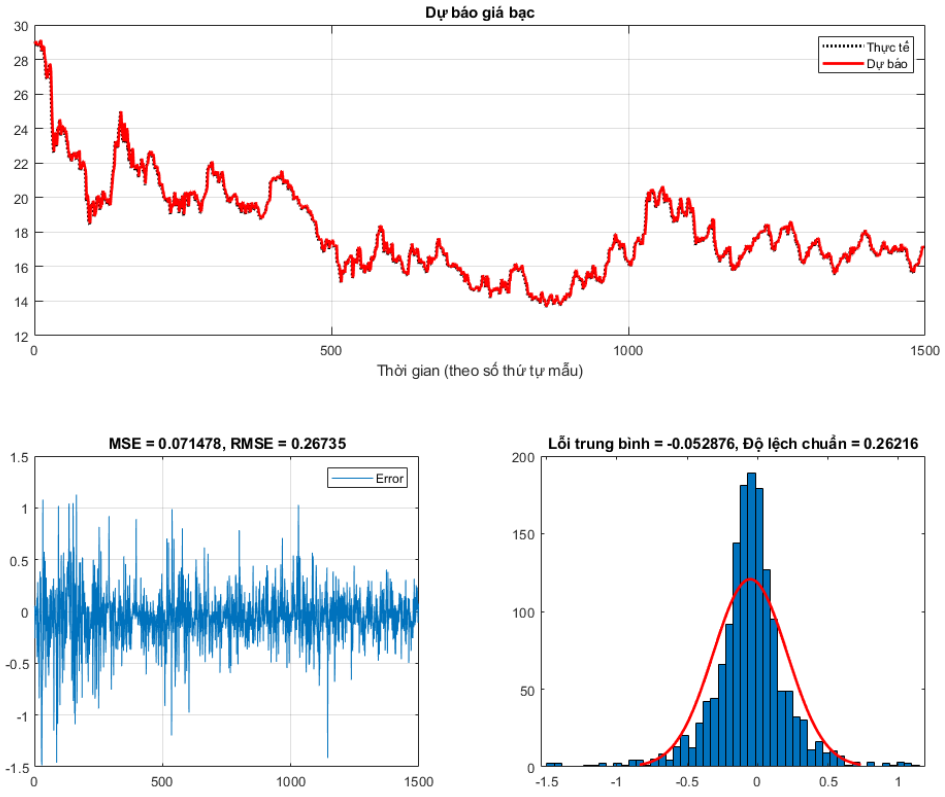
Đối với dự báo giá bạc, các giá trị lỗi như sau: $MSE = 0{,}071478$, $RMSE = 0{,}26735$, $Lỗi\ trung\ bình = -0{,}052876$ và Độ $lệch\ chuẩn = 0{,}26216$.

Đối với dự báo giá dầu thô Brent, các giá trị lỗi như sau: $MSE = 1{,}3065$, $RMSE = 1{,}143$, $Lỗi\ trung\ bình = -0{,}069199$ và Độ $lệch\ chuẩn = 1{,}1413$.

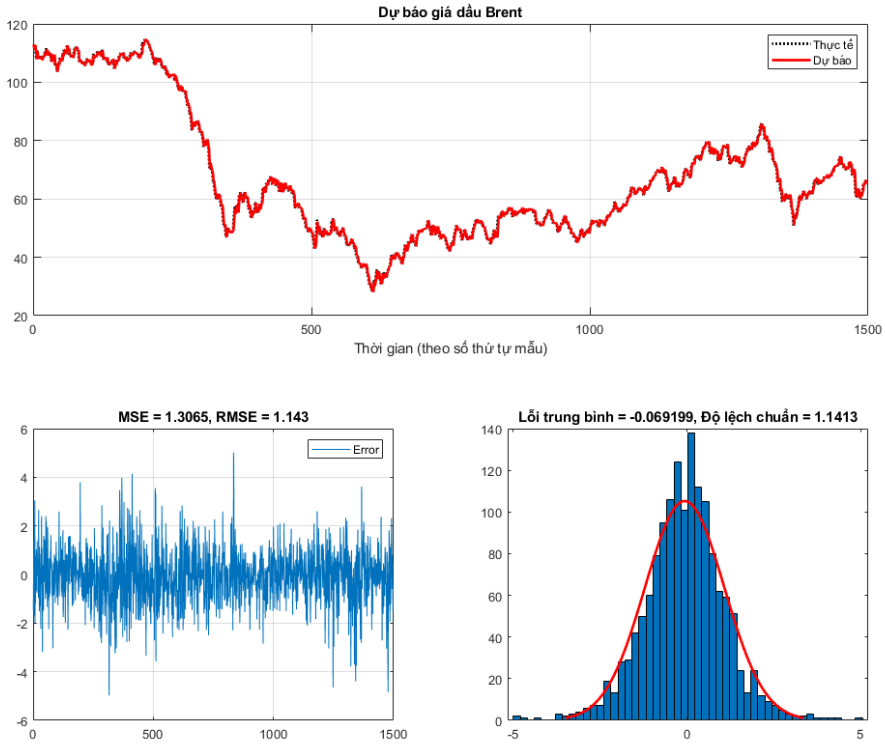
Đối với dự báo giá dầu thô WTI, các giá trị lỗi như sau: $MSE = 1{,}2602$, $RMSE = 1{,}1226$, $Lỗi\ trung\ bình = -0{,}1082$ và Độ $lệch\ chuẩn = 1{,}1178$.

Đối với dự báo giá gas, các giá trị lỗi như sau: $MSE = 0{,}0092529$, $RMSE = 0{,}096192$, $Lỗi\ trung\ bình = -0{,}0067529$ và Độ $lệch\ chuẩn = 0{,}095987$.

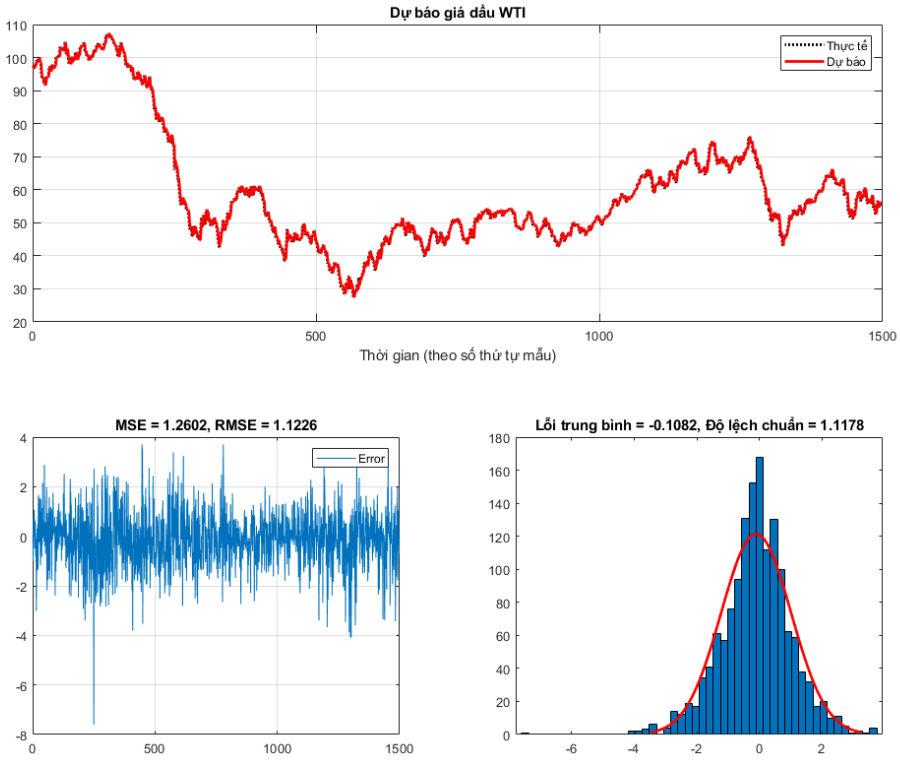
Đối với dự báo giá đồng, các giá trị lỗi như sau: $MSE = 0{,}0013519$, $RMSE = 0{,}036768$, $Lỗi\ trung\ bình = -0{,}0060239$ và Độ $lệch\ chuẩn = 0{,}036283$.



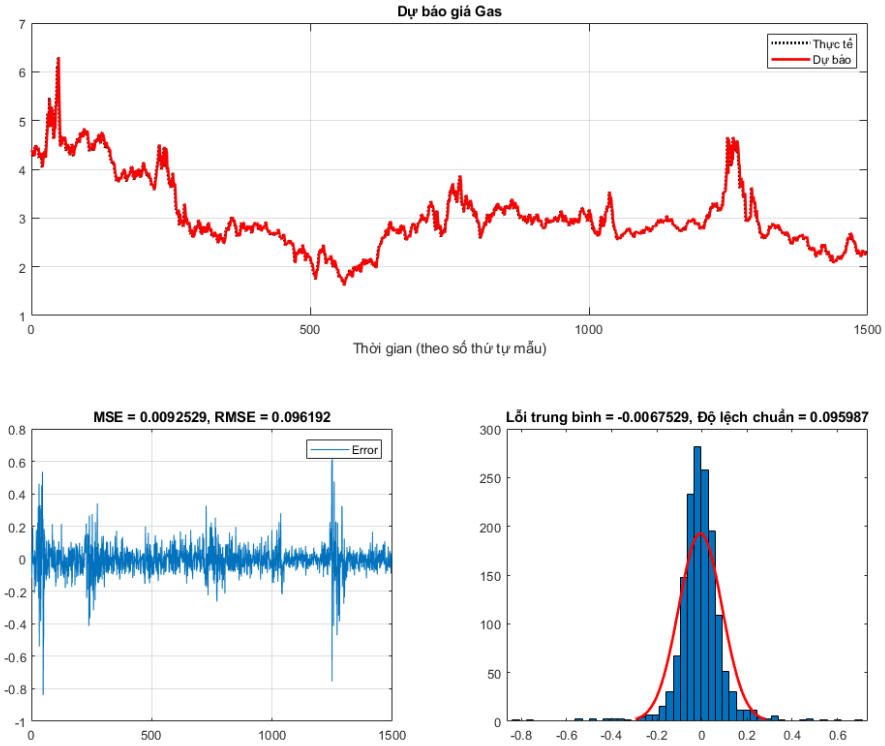
Hình 9: *Kết quả dự báo giá bạc (GMDH)*



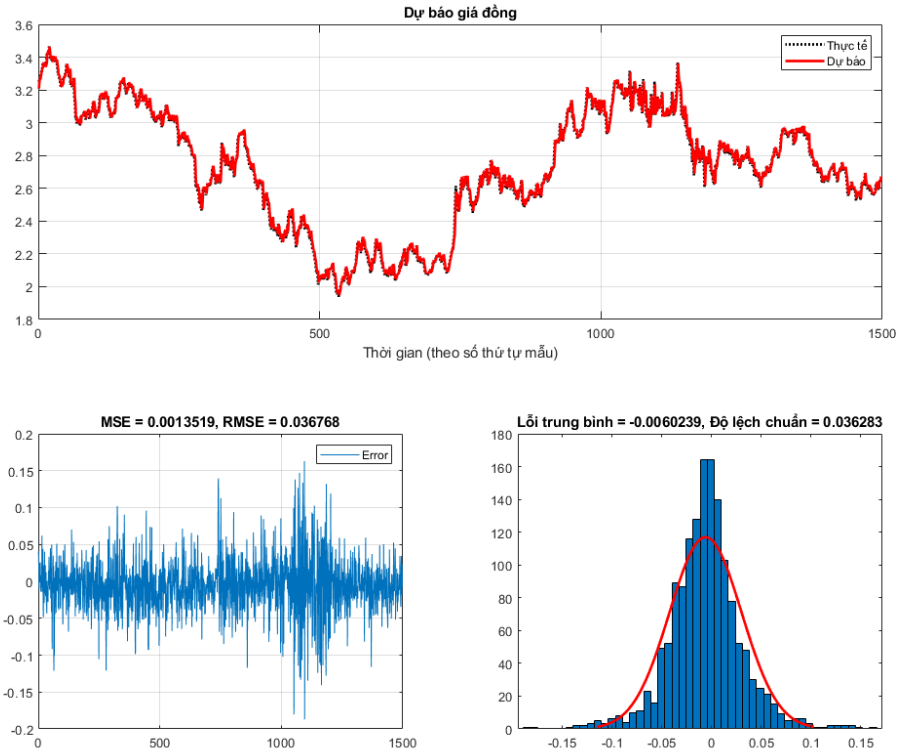
Hình 10: *Kết quả dự báo giá dầu thô Brent (GMDH)*



Hình 11: *Kết quả dự báo giá dầu thô WTI (GMDH)*

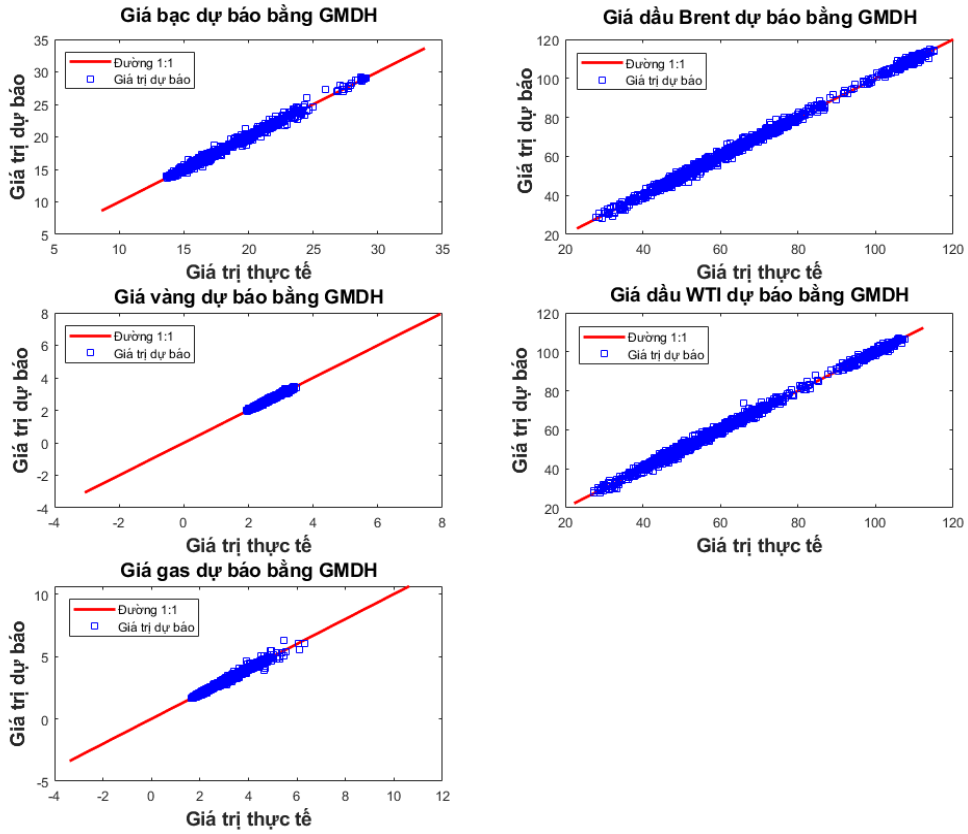


Hình 12: *Kết quả dự báo giá gas (GMDH)*



Hình 13: *Kết quả dự báo giá đồng (GMDH)*

Việc so sánh các giá trị thực tế và giá trị dự báo của mô hình GMDH cũng được thể hiện qua Hình 14. Đường thẳng 1:1 thể hiện giá trị thực tế và giá trị dự báo trùng nhau. Nếu tập các giá trị thực tế và giá trị dự báo tập trung quanh đường thẳng 1:1 thì giá trị dự báo gần với giá trị thực tế. Quan sát Hình 14 nhận thấy giá trị dự báo bởi mô hình GMDH và giá trị thực tế tương đối khớp nhau do các điểm tập trung gần nhau và có thể vẽ được một đường thẳng đi qua các điểm này. Hình 14 thể hiện mối tương quan là rất mạnh.



Hình 14: *Giá trị thực tế và giá trị dự báo của mô hình GMDH*

Dựa vào các kết quả thu được, có thể kết luận rằng mô hình dựa trên kỹ thuật GMDH cho kết quả tin cậy nhất. Do đó, với bộ dữ liệu đã thu thập được, GMDH có thể được ứng dụng trong việc dự báo giá.

5. Kết luận và hướng phát triển

Dự báo giá của các mặt hàng chính trên thị trường là một lĩnh vực nghiên cứu thu hút sự quan tâm của nhà kinh tế, nhà phân tích dữ liệu vì tính ứng dụng cao của kết quả dự báo. Trong nghiên cứu này, tác giả đã phát triển các mô hình dự báo giá của một số mặt hàng chính dựa trên các kỹ thuật trí tuệ nhân tạo bao gồm ANFIS, ANN, GMDH và LSTM. Kết quả thực nghiệm đã khẳng định rằng mô hình dự báo GMDH cho độ chính xác dự báo cao nhất và tin cậy nhất.

Do hạn chế về mặt số liệu nên các yếu tố khác ảnh hưởng đến giá các mặt hàng như chỉ số S&P 500, chỉ số ổn định chính trị chưa đưa được vào mô hình dự báo. Hướng

nghiên cứu tiếp theo sẽ là: nâng cấp các mô hình đã được xây dựng trong bài báo thành một hệ hỗ trợ ra quyết định hoàn chỉnh phục vụ cho dự báo giá của một số mặt hàng, bao gồm 05 thành phần: hệ thống máy tính, cơ sở dữ liệu, quản lý mô hình, quản lý cơ sở tri thức, giao tiếp với người dùng. Đồng thời cần tiếp tục bổ sung thêm các nhân tố khác có thể ảnh hưởng đến sự biến động của các mặt hàng. Để tăng độ chính xác dự báo của các mô hình, việc tinh chỉnh các tham số cũng cần được xem xét.

TÀI LIỆU THAM KHẢO

[1] Bakir H., Chniti G., Zaher H., "E-Commerce price forecasting using LSTM neural networks," *Int. J. Mach. Learn Comput,* 8:169-174, 2018. DOI: 10.18178/ijmlc.2018.8.2.682

[2] Bashiri Behmiri N., Pires Manso J. R., "Crude Oil Price Forecasting Techniques: A Comprehensive Review of Literature," *SSRN Electron J.,* 2013 DOI: 10.2139/ssrn.2275428

[3] Jeenanunta C., Chaysiri R., Thong L., "Stock Price Prediction With Long Short-Term Memory Recurrent Neural Network," *In: 2018 International Conference on Embedded Systems and Intelligent Technology & International Conference on Information and Communication Technology for Embedded Systems (ICESIT-ICICTES), IEEE*, pp. 1-7, 2018. DOI: 10.1109/ICESIT-ICICTES.2018.8442069

[4] Jubinski D., Lipton A., "VIX, Gold, Silver, and Oil: How do Commodities React to Financial Market Volatility," *J. Account Financ*, 2013.

[5] Kristjanpoller W., Minutolo M. C., "Gold price volatility: A forecasting approach using the Artificial Neural Network-GARCH model," *Expert Syst. Appl.,* 2015. DOI: 10.1016/j.eswa.2015.04.058

[6] Thắng H. Q., Vũ P. Đ., Vinh T. V., "Dự đoán xu thế chỉ số chứng khoán Việt Nam sử dụng phân tích hồi quy quá trình Gauss và mô hình tự hồi quy trung bình động," *Chuyên san Các công trình Nghiên cứu và Phát triển về Công nghệ thông tin và Truyền thông*, 2018. DOI: 10.32913/rd-ict.vol1.no39.571

[7] Toàn N. V., Quốc N. P., Thạch N. H., "Dự báo giá vàng Việt nam sử dụng mô hình Garch," *Tạp chí Trường Đại học An Giang,* 2016.

[8] Husain A. M., Bowman C., *Forecasting Commodity Prices: Futures Versus Judgment,* IMF Work Pap., 2004. DOI: 10.5089/9781451846133.001

[9] Yazdani-Chamzini A., Yakhchali S. H., Volungevičiene D., Zavadskas E. K., "Forecasting gold price changes by using adaptive network fuzzy inference system," *J. Bus. Econ. Manag.,* 2012 DOI: 10.3846/16111699.2012.683808

[10] Yu L., Wang S, Lai K. K., "Forecasting crude oil price with an EMD-based neural network ensemble learning paradigm," *Energy Econ.,* 2008. DOI: 10.1016/j.eneco.2008.05.003

[11] Haidar I., Kulkarni S., Pan H., "Forecasting model for crude oil prices based on artificial neural networks," *In: ISSNIP 2008 - Proceedings of the 2008 International*

Conference on Intelligent Sensors, Sensor Networks and Information Processing. 2008. DOI: 10.1109/ISSNIP.2008.4761970

[12] Panella M., Barcellona F., D'Ecclesia R. L., "Forecasting energy commodity prices using neural networks," *Adv. Decis Sci.,* 2012 DOI: 10.1155/2012/289810

[13] Tapia Cortez C. A., Saydam S, Coulton J., Sammut C., "Alternative techniques for forecasting mineral commodity prices," *Int J Min Sci Technol.,* 2018. DOI: 10.1016/j.ijmst.2017.09.001

[14] Ebtehaj I., Bonakdari H., Zaji A. H. et al., "GMDH-type neural network approach for modeling the discharge coefficient of rectangular sharp-crested side weirs," *Eng Sci Technol an Int J.,* 2015. DOI: 10.1016/j.jestch.2015.04.012

[15] Amanifard N., Nariman-Zadeh N., Farahani M. H., Khalkhali A., "Modelling of multiple short-length-scale stall cells in an axial compressor using evolved GMDH neural networks," *Energy Convers Manag*., 2008. DOI: 10.1016/j.enconman.2008.05.025

[16] Chatfield C., *The analysis of time series: theory and practice*, Springer, 2013.

[17] Stepnicka M., Donate J. P., Cortez P. et al, "Forecasting seasonal time series with computational intelligence: contribution of a combination of distinct methods," *In: Proceedings of the 7th conference of the European Society for Fuzzy Logic and Technology*, Atlantis Press, pp 464-471, 2013. DOI: 10.2991/eusflat.2011.7

[18] Yusof N., Bahiah N., Shahizan M., Chun Y., " A Concise Fuzzy Rule Base to Reason Student Performance Based on Rough-Fuzzy Approach," In: Fuzzy Inference System - Theory and Applications, 2013. DOI: 10.5772/37773

[19] Singh R., Kainthola A., Singh T. N., "Estimation of elastic constant of rocks using an ANFIS approach," *Appl Soft Comput J.,* 2012. DOI: 10.1016/j.asoc.2011.09.010

[20] Ata R., Kocyigit Y., "An adaptive neuro-fuzzy inference system approach for prediction of tip speed ratio in wind turbines," *Expert Syst Appl*., 2010. DOI: 10.1016/j.eswa.2010.02.068

[21] Ghenai C. et al., "Short-term building electrical load forecasting using adaptive neuro-fuzzy inference system (ANFIS)," *J. Build Eng*., 52:104323. 2022. DOI: 10.1016/j.jobe.2022.104323

[22] Takagi T., Sugeno M., "Derivation of fuzzy control rules from human operator's control actions," In: IFAC Proceedings Seriesm, 1984. DOI: 10.1016/S1474-6670(17)62005-6

[23] Sugeno M., "An introductory survey of fuzzy control," *Information Sciences*, 1985. DOI: 10.1016/0020-0255(85)90026-X

[24] Wei M., Bai B., Sung A. H., et al., "Predicting injection profiles using ANFIS," *Information Sciences*, 2007. DOI: 10.1016/j.ins.2007.03.021

[22] Foroughi B., Nhan P. V., Iranmanesh M. et al., "Determinants of intention to use autonomous vehicles: Findings from PLS-SEM and ANFIS," *J. Retail Consum Serv*. 70:103158. DOI: 10.1016/j.jretconser.2022.103158

[26] Jyh-Shing R Jang, Chuen-Tsai Sun, Eiji Mizutani, *Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence*, Book Review, IEEE Transactions On Automatic Control, 1997. DOI: 10.1109/TAC.1997.633847

[27] Cuomo S., Di Cola V. S., Giampaolo .F, et al., "Scientific machine learning through physics-informed neural networks: Where we are and what's next," *J. Sci. Comput.*, 92:88, 2002. DOI: 10.1007/s10915-022-01939-z

[28] Raviv L., Lupyan G., Green S. C., "How variability shapes learning and generalization," *Trends Cogn. Sci*., 26:462-483, 2022.. DOI: 10.1016/j.tics.2022.03.007

[29] Masini R. P., Medeiros M. C., Mendes E. F., "Machine learning advances for time series forecasting," *J. Econ. Surv*., 37:76-111, 2023. DOI: 10.1111/joes.12429

[30] Caruana R., Lawrence S., Giles L., "Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping," *In: Advances in Neural Information Processing Systems*, 2001. DOI: 10.1109/IJCNN.2000.857823

[31] Ivakhnenko A. G., "Polynomial Theory of Complex Systems," *IEEE Trans Syst Man Cybern*, 1971. DOI: 10.1109/TSMC.1971.4308320

[32] Teng G. E., He C. Z., Xiao J., Jiang X. Y., "Customer credit scoring based on HMM/GMDH hybrid model," *Knowl Inf. Syst.*, 36:731-747, 2013. DOI: 10.1007/s10115-012-0572-z

[33] Teng G., He C., Gu X., "Response model based on weighted bagging GMDH," *Soft Comput*., 2014. DOI: 10.1007/s00500-014-1225-9

[34] R. Y. M. Li, Simon Fong, Kyle Weng Sang Chong, "Forecasting the REITs and stock indices: group method of data handling neural network approach," *Pacific Rim. Prop. Res. J.*, 23:123-160, 2017. DOI: 10.1080/14445921.2016.1225149

[35] Ma M., Liu C., Wei R. et al., "Predicting machine's performance record using the stacked long short-term memory (LSTM) neural networks," *J. Appl. Clin. Med. Phys.*, 23:e13558, 2022. DOI: 10.1002/acm2.13558

[36] Schmidhuber J (2015) Deep Learning in neural networks: An overview. Neural Networks. DOI: 10.1016/j.neunet.2014.09.003

[37] Hunt K. M. R., Matthews G. R., Pappenberger F., Prudhomme C., "Using a long short-term memory (LSTM) neural network to boost river streamflow forecasts over the western United States," *Hydrol Earth Syst. Sci.*, 26:5449-5472, 2022. DOI: 10.5194/hess-26-5449-2022

[38] Grubbs F. E., "Procedures for detecting outlying observations in samples," *Technometrics*, 11:1-21, 1969. DOI: 10.1080/00401706.1969.10490657

[39] https://www.investing.com

ABSTRACT

RESEARCH ON THE APPLICATION OF ARTIFICIAL INTELLIGENCE TECHNIQUES IN PRICE FORECASTING OF SOME COMMODITIES

Nguyen Thai Son
Faculty of Information Technology, Dai Nam University, Ha Dong, Hanoi, Vietnam
Received on 04/8/2023, accepted for publication on 17/8/2023

The global economy is significantly impacted by changes in the price of primary commodities. As a result, both the academic and professional sectors have paid attention to price predictions for major commodities. The goal of this study is to build an artificial intelligence-based model for one-day market price predictions for important commodities like copper, crude oil, gas, and silver. The information on commodity trading was gathered between 01/2000 and 10/2019. Different models based on group method of data handling (GMDH), long short-term memory (LSTM), artificial neural network (ANN), and adaptive neuro fuzzy inference system (ANFIS) were developed. Theil's U, RMSE, MAPE, MAE, R, and other performance indices were used to compare the models. The findings demonstrated that, in terms of commodity price prediction, the suggested model based on GMDH technique performs better than alternative approaches. A viable alternative for price prediction is the GMDH-based model. For economists and professionals involved in commodity price forecasting, the GMDH can be a useful tool.

Keywords: Major commidities; price prediction; artificial intelligence techniques; GMDH; ANFIS; ANN; LSTM.