Dự báo mực nước sông cao nhất, thấp nhất trong ngày sử dụng mô hình hỗn hợp

The daily highest and lowest river water levels are forecasted using a hybrid model

Đỗ Văn Đỉnh[1], Nguyễn Trọng Quỳnh[1], Vũ Văn Cảnh[2], Phạm Văn Nam[3]

Email: dodinh75@gmail.com

[1]Trường Đại học Sao Đỏ
[2]Trung tâm GDNN-GDTX Tp Chí Linh, Hải Dương
[3]Trường Đại học Công nghiệp Hà Nội

Ngày nhận bài: 08/02/2021
Ngày nhận bài sửa sau phản biện: 29/3/2021
Ngày chấp nhận đăng: 31/3/2021

Tóm tắt

Dự báo mực nước sông cao nhất và thấp nhất trong ngày là bài toán có tính thực tiễn cao. Đã có nhiều phương án đề xuất để dự báo hai đại lượng này, nhưng do các thông số của mô hình dự báo phụ thuộc vào điều kiện địa lý và phát triển kinh tế khu vực cần dự báo. Vì vậy, đối với từng khu vực dự báo cần phải xác định lại các thông số của mô hình hoặc đề xuất mô hình mới phù hợp để đạt được độ chính xác cao hơn. Bài báo đề xuất ứng dụng mô hình hỗn hợp để ước lượng mực nước sông cao nhất và thấp nhất trong ngày, trong đó thành phần tuyến tính được ước lượng bằng SVD (Singular Value Decomposition) và thành phần phi tuyến ước lượng bằng máy học véctơ hỗ trợ SVM (Support Vector Mechine). Số liệu đầu vào là giá trị mực nước sông cao nhất, thấp nhất và trung bình của các ngày trước đó. Chất lượng của giải pháp đề xuất được kiểm nghiệm trên bộ số liệu thực tế (1.460 ngày, từ 01/01/2017 đến 31/8/2020) ở khu vực hạ lưu sông Thái Bình, Trạm thủy văn Bá Nha, huyện Thanh Hà, tỉnh Hải Dương. Kết quả sai số trung bình tuyệt đối đạt 0.12.

Từ khóa: *SVD; SVM; mô hình tuyến tính; mô hình phi tuyến; mô hình hỗn hợp.*

Abstract

The daily forecast of the highest and lowest river water level is a problem with a high applicability. There have been many proposed methods to forecast these two quantities, however the parameters of the forecasting model depend on geographical conditions and regional economic development to forecast. Therefore, for each forecast area it is necessary to redefine the model parameters or propose a suitable new model to achieve higher accuracy. The paper proposes to apply a hybrid model to estimate the day's highest and lowest river water levels, in which the linear component is estimated by Singular Value Decomposition (SVD) and the nonlinear component estimated by Support Vector Mechine (SVM) vector. Input data is the value of the highest, lowest and average river water level of the previous days. The quality of the proposed solution was tested on the actual data set (1.460 days, from January 1, 2017 to August 31, 2020) in the downstream area of Thai Binh river, Ba Nha hydrological station, Thanh Ha district, Hai Duong province. The result of the absolute average error is 0.12.

Keywords: *SVD; SVM; linear models; nonlinear models; hybrid model.*

1. ĐẶT VẤN ĐỀ

Bài toán dự báo là một trường hợp đặc biệt của bài toán ước lượng và xây dựng mô hình ánh xạ giữa đầu vào và đầu ra. Dự báo mực nước sông có ý nghĩa thực tiễn trong việc phát triển kinh tế của khu vực, đặc biệt trong sản xuất nông nghiệp. Biết được mực nước sông hàng ngày sẽ giúp các trạm Thủy văn khu vực điều tiết lưu lượng nước để phục vụ sản xuất như tưới, tiêu, chống úng nội đồng...

Người phản biện: 1. GS.TSKH. Thân Ngọc Hoàn
2. PGS.TS. Nguyễn Văn Tiềm

Đã có nhiều mô hình dự báo thống kê được nghiên cứu và ứng dụng thành công trên thế giới như phương pháp hồi quy phi tuyến tính, phi tuyến; phương pháp giá trị cực trị (Extreme Value) và mạng nơron nhân tạo (ANN – Artificial Neural Network) [7-11], trong số đó, các mô hình ứng dụng mạng nơron nhân tạo đã đạt được những tiến bộ đáng kể và nghiên cứu ứng dụng rộng rãi trong thời gian qua [1, 3-5, 7-11]. Thuật toán máy học véctơ hỗ trợ (SVM) được Vapnik giới thiệu năm 1995 [6], đã được nghiên cứu thử nghiệm trong lĩnh vực dự báo các thông số khí tượng, môi trường và thu được những kết quả khả quan, trong hầu hết các nghiên cứu đã được công

bố, mô hình dự báo dùng kỹ thuật SVM đều cho kết quả tốt hơn so với các mô hình ANN kiểm chứng [12-14]. Trong bài báo này, nhóm tác giả đề xuất ứng dụng kỹ thuật SVM trong mô hình hỗn hợp [2] để dự báo mực nước sông ở khu vực hạ lưu sông Thái Bình. Kết quả nghiên cứu thực nghiệm cho thấy ứng dụng kỹ thuật SVM trong mô hình hỗn hợp dự báo mực nước sông cho kết quả khả quan.

2. PHỐI HỢP SVD VÀ SVM TRONG MÔ HÌNH HỖN HỢP DỰ BÁO MỰC NƯỚC SÔNG

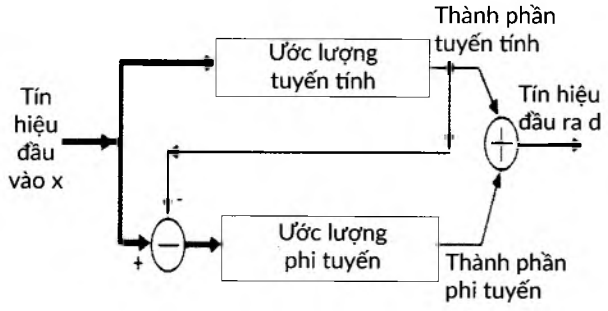
2.1. Mô hình hỗn hợp

Mô hình hỗn hợp [2] được tác giả đề xuất để dự báo ngắn hạn phụ tải điện và cho kết quả khả quan; để ước lượng thành phần tuyến tính tác giả sử dụng thuật toán khai triển theo các giá trị kỳ dị SVD, phần ước lượng phi tuyến sử dụng mạng MLP.

Trong bài báo này, nhóm tác giả đề xuất phối hợp SVD và SVM trong mô hình hỗn hợp để dự báo mực nước sông cao nhất (HL_{max}) và mực nước sông thấp nhất (HL_{min}) trong ngày ở khu vực hạ lưu sông Thái Bình.

2.1.1. Cấu trúc của mô hình hỗn hợp

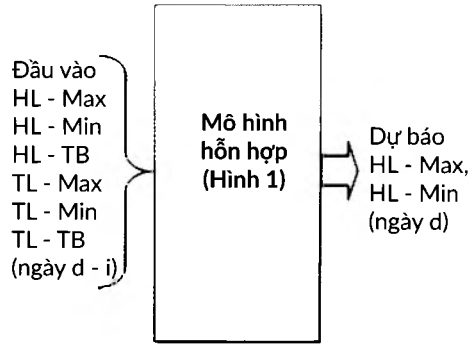
Sơ đồ cấu trúc của mô hình hỗn hợp được trình bày như Hình 1, tín hiệu đầu vào (x) là véctơ chứa các số liệu quá khứ; tín hiệu đầu ra (d) là tổng của hai thành phần ước lượng: ước lượng tuyến tính và ước lượng phi tuyến.



Hình 1. *Cấu trúc của mô hình hỗn hợp [2]*

Khi sử dụng mô hình hỗn hợp, để giảm bớt mức độ phức tạp của mô hình phi tuyến, trước hết cần ước lượng thành phần tuyến tính, sau đó ta loại thành phần tuyến tính khỏi các số liệu đầu vào để nhằm chỉ giữ lại thành phần phi tuyến trong tín hiệu của đối tượng. Tín hiệu còn lại này sẽ được dùng để huấn luyện khối phi tuyến hay nói cách khác: Sai số còn lại từ khối tuyến tính trở thành đầu vào của khối phi tuyến.

Cấu trúc của mô hình dự báo mực nước sông cao nhất (HL_{max}) và mực nước sông thấp nhất (HL_{min}) ở khu vực hạ lưu trong ngày như Hình 2; trong đó HL_{max} (mực nước cao nhất ở hạ lưu), HL_{min} (mực nước thấp nhất ở hạ lưu), HLTB (mực nước trung bình ở hạ lưu), TL_{max} (mực nước cao nhất ở thượng lưu), TL_{min} (mực nước thấp nhất ở thượng lưu), TLTB (mực nước trung bình ở thượng lưu), ngày d là ngày dự báo, ngày d-i là các ngày trước ngày dự báo.



Hình 2. *Cấu trúc mô hình dự báo mức nước cao nhất, thấp nhất trong ngày*

2.1.2. Mô tả toán học của mô hình hỗn hợp

Từ sơ đồ Hình 1 ta có:

$$d = f(x) \approx Linear(x) + NonLinear(x) \quad (1)$$

Mô hình tuyến tính (Linear(x)) được xác định trước sau đó sẽ xác định mô hình phi tuyến (NonLinear(x)). Với bộ số liệu gồm p mẫu {xi, di}, i = 1, 2,... p, mô hình tuyến tính được xác định trên cơ sở tối ưu hóa hàm sai số trên tập mẫu số liệu này:

$$\forall i : Linear(x_i) \approx d_i$$

$$\text{hay } e = \frac{1}{2} \sum_{i=1}^{p} \left\| Linear(x_i) - d_i \right\|^2 \to \min \quad (2)$$

Khi xác định được mô hình tuyến tính, phần sai số còn lại sẽ được xấp xỉ bởi mô hình phi tuyến bằng các thuật toán tối ưu hóa hàm sai số phi tuyến:

$$\forall i : NonLinear(x_i) \approx d_i - Linear(x_i)$$

hay

$$e = \frac{1}{2} \sum_{i=1}^{p} \left\| NonLinear(x_i) - (d_i - Linear(x_i)) \right\|^2 \to \min \quad (3)$$

Giả thiết rằng giá trị HL_{max} được ước lượng theo (4) (Ước lượng giá trị HL_{min} làm tương tự):

$$\begin{aligned} HL_{\max}(d) \approx\ & f_{1,2,\ldots,K}(HL_{\max}(d-i), \\ & HL_{\min}(d-i), HL_{TB}(d-i), TL_{\max}(d-i)), \\ & TL_{\min}(d-i), TL_{TB}(d-i) + \\ & + \sum_{i=1}^{K} \begin{bmatrix} a_{i1} \cdot HL_{\max}(d-i) + a_{i2} \cdot HL_{\min}(d-i) + \\ + a_{i3} \cdot HL_{TB}(d-i) + a_{i4} \cdot TL_{\max}(d-i) + \\ + a_{i5} \cdot TL_{\min}(d-i) + a_{i6} \cdot TL_{TB}(d-i) \end{bmatrix} \end{aligned} \quad (4)$$

Trong đó:

f(): hàm phi tuyến;

a_{ij}: các hệ số của mô hình tuyến tính;

Mô hình phi tuyến được xấp xỉ bằng SVM.

2.2. Xây dựng mô hình hỗn hợp

2.2.1. Lựa chọn đặc tính đầu vào của mô hình dự báo [1, 2]

Ứng dụng thuật toán khai triển theo giá trị kỳ dị để lựa chọn đặc tính đầu vào của mô hình dự báo. Bài toán xây dựng mô hình tuyến tính có thể đưa về giải tìm nghiệm x của hệ phương trình:

$A \cdot x = b$ (5)

Trường hợp số phương trình nhiều hơn số ẩn nên thường không có nghiệm duy nhất, khi đó nghiệm của hệ phương trình trên được xác định từ bài toán tối ưu hóa sai số (còn gọi là residue r) định nghĩa bởi:

$$\min\left\|A \cdot x - b\right\| = \min\left\|r\right\| = ? \quad (6)$$

Nghiệm của bài toán tối ưu (6) có thể được xác định dựa trên kết quả phân tích ma trận A $\epsilon^{-m\times n}$ theo các giá trị kỳ dị. Theo [1, 2], với ma trận không vuông, ta có thể xác định ma trận $A^+ \epsilon^{-m\times n}$ từ phân tích SVD của ma trận A. Với $A = U\cdot S\cdot V^T$ thì.

$A^+ = V\cdot S^+\cdot U^T$ (7)

Với U, V: các ma trận trực giao:

$$S^+ = diag(\frac{1}{\sigma_1},\frac{1}{\sigma_2},\ldots,\frac{1}{\sigma_r}) \in \mathbb{R}^{n\times m}$$

Ma trận đường chéo.

Khi đó nghiệm tối ưu của phương trình (6) được xác định bởi:

$x = A^+ \cdot b$ (8)

2.2.2. Ứng dụng SVM ước lượng thành phần phi tuyến

Cho tập dữ liệu gồm N mẫu huấn luyện {(x1, y1),..., (xN, yN)} trong đó là các véctơ đầu vào (D chiều) và yi ϵ {±1} là mã lớp của véctơ đầu vào. Bài toán nhị phân chỉ phân loại 2 lớp, được mã tương ứng là lớp +1 và lớp -1. Ta cần tìm một siêu phẳng $w \cdot x + b = 0$ để tách tập dữ liệu trên thành 2 lớp.

Trong đó:

w: véctơ pháp tuyến của siêu phẳng, có tác dụng điều chỉnh hướng của siêu phẳng;

Giá trị b có tác dụng di chuyển siêu phẳng song song với chính nó.

Có thể, có nhiều siêu phẳng để phân tách tập dữ liệu và cũng đã có nhiều thuật toán để giải bài toán này, chẳng hạn như thuật toán Perceptron của Rosenblatt [15], thuật toán biệt thức tuyến tính của Fisher [16]. Tuy nhiên, trong thuật toán SVM, siêu phẳng tối ưu được cho là siêu phẳng có tổng khoảng cách tới các véctơ gần nhất của hai lớp là lớn nhất. Bên cạnh đó, để đảm bảo tính tổng quát hóa cao, một biến lỏng (Slack Variable) được đưa vào để nới lỏng điều kiện phân lớp. Bài toán đưa đến việc giải quyết tối ưu có ràng buộc:

$$\min_{w,b,\xi} \frac{1}{2} w^T \cdot w + C\sum_{i=1}^{N} \xi_i$$

Sao cho:

$$y_i(w^T \cdot x_i + b) + \xi_i - 1 \geq 0; \xi_i \geq 0, \forall i \in [1, N] \quad (9)$$

Trong đó:

C >0: Tham số chuẩn tắc (Regularization Parameter);

ξ_i: Biến lỏng.

Bài toán (10) có thể được giải bằng phương pháp SMO (Sequential Minimal Optimization). Phương pháp này đưa đến giải bài toán đối ngẫu quy hoạch toàn phương (Quadratic Programming):

$$\max_{\alpha} L(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \cdot \alpha_j \cdot y_i \cdot y_j \cdot \Phi(x_i) \cdot \Phi(x_j) \quad (10)$$

Thỏa mãn: $0 \leq \alpha_i \leq C, \forall i \in [1, N]$ và $\sum_{i=1}^{N} \alpha_i \cdot y_i = 0$ với α_i là các nhân tử Lagrange. Sau khi có được các giá trị α_i từ bài toán (11), ta sẽ thu được các giá trị tối ưu w^* và b^* của siêu phẳng. Chỉ có các mẫu có $\alpha \geq 0$ mới được gọi là các véctơ hỗ trợ. Cuối cùng, hàm quyết định phân lớp có dạng:

$$f(x) = \text{sgn}\left(\alpha_i \cdot y_i \cdot \Phi(x_i) \cdot \Phi(x_j) + b^*\right) \quad (11)$$

Gọi $K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$ là hàm nhân của không gian đầu vào. Theo đó, tích vô hướng trong không gian đặc trưng tương đương với hàm nhân $K(x_i, x_j)$ ở không gian đầu vào. Như vậy, thay vì tính trực tiếp giá trị tích vô hướng, ta thực hiện gián tiếp thông $K(x_i, x_j)$ cho các tính toán tiếp theo.

2.2.3. Ước lượng HL_{max}, HL_{min} trong ngày

2.2.3.1. Ước lượng thành phần tuyến tính

Từ phương trình (4), hàm quan hệ tuyến tính giữa HL_{max} của ngày d với HL_{max} của các ngày quá khứ và được xác định từ hệ phương trình ước lượng xấp xỉ như trong công thức (12) và (13). Từ (13) ta cần xác định véctơ $a = \left[a_1, a_2, \ldots, a_K\right]^T$ để đạt cực tiểu của hàm sai số ước lượng. Trong thực tế áp dụng, ta còn cần trả lời hai câu hỏi: (1) Cần sử dụng bao nhiêu số liệu trong quá khứ?, (2) Đó là những số liệu nào?.

$$\begin{cases} a_1 \cdot HT_{\max}(K) + a_2 \cdot HL_{\max}(K-1) + \ldots + a_K \cdot HL_{\max}(d-K) \approx HL_{\max}(K+1) \\ a_1 \cdot THL(K-1) + a_2 \cdot HL_{\max}(K-2) + \ldots + a_K \cdot HL_{\max}(d-K+1) \approx HL_{\max}(K+2) \\ \ldots \qquad \ldots \qquad \ldots \\ a_1 \cdot HL_{\max}(N_{\max}-1) + a_2 \cdot HL_{\max}(N_{\max}-2) + \ldots + a_K \cdot HL_{\max}(N_{\max}-K) \approx HL_{\max}(N_{\max}) \end{cases} \quad (12)$$

$$\Leftrightarrow \begin{bmatrix} HL_{\max}(K) & HL_{\max}(K-1) & .... & HL_{\max}(1) \\ HL_{\max}(K-1) & HL_{\max}(K-2) & ... & HL_{\max}(2) \\ \vdots & \vdots & \vdots & \vdots \\ HL_{\max}(N_{\max}-1) & HL_{\max}(N_{\max}-2) & ... & HL_{\max}(N_{\max}-K) \end{bmatrix} \cdot \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_K \end{bmatrix} \approx \begin{bmatrix} HL_{\max}(K+1) \\ HL_{\max}(K+2) \\ \vdots \\ HL_{\max}(N_{\max}) \end{bmatrix} \quad (13)$$

Phương pháp xác định thích nghi được thực hiện như sau:

- Trước tiên ta sử dụng một số lượng lớn số liệu quá khứ (trong nghiên cứu sử dụng K = 60 - Tương đương 2 tháng số liệu trước đó - là đủ để dự báo ngày tiếp theo).

- Với K số liệu quá khứ, ta xác định véctơ $a=\left[a_1,a_2,....,a_K\right]^T$ của hàm ước lượng tuyến tính $HL_{\max}(d)\approx\sum_{i=1}^{K}\left[a_i\cdot HL_{\max}(d-i)\right]$ bằng SVD.

- Xác định thành phần có giá trị tuyệt đối nhỏ nhất trong véctơ a. Thành phần này sẽ tương ứng với ngày trong quá khứ ít ảnh hưởng tới ngày dự báo và ta loại bỏ khỏi bộ số liệu trong quá khứ, giảm K=K-1, và quay lại bước 2 nếu $K>K_{min}$ chọn trước. Quá trình lặp các bước 2-3 cho đến khi K giảm xuống một giá trị đủ nhỏ có thể chấp nhận được để mô hình không quá phức tạp. Cụ thể, trong bài báo ta chọn $K_{min}<5$.

Tương tự, tiếp tục xây dựng hàm quan hệ tuyến tính giữa T_{max} của ngày d với HL_{min}, HLTB, TL_{max}, TL_{min} và TL_{TB} của các quá khứ ta được phương trình (14).

$$HL\max(d)\approx\sum_{i=1}^{K}\begin{Bmatrix} a_{i1}\cdot HL\max(d-i)+a_{i2}\cdot HL\min(d-i)+ \\ a_{i3}\cdot HL_{TB}(d-i)+a_{i4}\cdot TL\max(d-i)+ \\ +a_{i5}\cdot TL\min(d-i)+a_{i6}\cdot TL_{TB}(d-i) \end{Bmatrix} \quad (14)$$

Khi xác định được mối quan hệ tuyến tính giữa HL_{max} của ngày d với các ngày trong quá khứ, ta tính sai số chênh lệch giữa số liệu thực tế và số liệu ước lượng như phương trình (15).

$$NL(d)=HL_{\max}(d)-\sum_{i=1}^{K}\begin{Bmatrix} a_{i1}\cdot HL_{\max}(d-i)+a_{i2}\cdot HL_{\min}(d-i)+ \\ a_{i3}\cdot HL_{TB}(d-i)+a_{i4}\cdot TL_{\max}(d-i)+ \\ +a_{i5}\cdot TL_{\min}(d-i)+a_{i6}\cdot TL_{TB}(d-i) \end{Bmatrix} \quad (15)$$

Đây sẽ là phần phụ thuộc phi tuyến còn lại giữa HL_{max} với các ngày trong quá khứ. Hoàn toàn thực hiện tương tự khi xây dựng các mô hình ước lượng cho HL_{min}.

2.2.3.2. Ước lượng thành phần phi tuyến

Khi xác định được các thông số mô hình tuyến tính, ta tiến hành xây dựng mạng nơron nhân tạo để ước lượng thành phần phi tuyến. Giá trị chênh lệnh (phương trình (15)) được sử dụng là đầu vào cho mô hình ước lượng thành phần phi tuyến dùng kỹ thuật SVM.

3. KẾT QUẢ VÀ THẢO LUẬN

Mô hình nghiên cứu được xây dựng trên nền phần mềm Matlab® 2010b, với SVM sử dụng LSSVMlabv1.8_R2009b_R2011a và được thiết kế theo các bước sau: (1) Chuẩn bị dữ liệu, (2) Lựa chọn đặc tính cho mô hình dự báo, (3) Xây dựng kiến trúc mạng, (4) Lựa chọn phương pháp và đào tạo mạng, (5) Đánh giá độ tin cậy. Chất lượng của mô hình được kiểm nghiệm trên số liệu thực tế tại Trạm thủy văn Bá Nha, huyện Thành Hà, tỉnh Hải Dương (1460 ngày từ 01/01/2017 đến 31/8/2020).

3.1. Kết quả ước lượng thành phần tuyến tính

3.1.1. Kết quả ước lượng HL_{max}

Bằng phương pháp phân tích SVD kết hợp với kinh nghiệm thực tế ta xác định các yếu tố ảnh hưởng lớn nhất đến mức nước cao nhất (HL_{max}) cần dự báo:

- Ảnh hưởng của HL_{max} trong quá khứ đến HL_{max} dự báo, ta xác định được 5 ngày có hệ số phụ thuộc lớn là: d-1, d-4, d-8, d-12 và d-15. Tiếp tục khảo sát sự phụ thuộc của HL_{max} vào các số liệu HL_{min}, HL_{TB}, TL_{max}, TL_{min}, TL_{TB} trong quá khứ bằng cách làm hoàn toàn tương tự ta được:

- Ảnh hưởng của HL_{min} trong quá khứ đến T_{max} dự báo là các ngày d-1, d-15 và d-18; Ngày d-27 xa ngày dự báo nên ta có thể loại.

- Giá trị HLTB trong quá khứ ảnh hưởng đến HL_{max} dự báo là d-1, d-2, d-4 và d-8. Ngày d-36 loại.

- Các giá trị TL_{max} trong quá khứ ảnh hưởng đến HL_{max} dự báo d-1; Ngày gần nhất d-47 xa ngày dự báo nên loại.

- Ảnh hưởng của TL_{min} đến HL_{max} là d-5, d-12, d-15 và d-16; ngày d-42 xa ngày dự báo nên loại.

- Ảnh hưởng của TL_{TB} đến HL_{max} là d-4, d-13 và d-14; ngày d-33 và d-51 xa ngày dự báo loại.

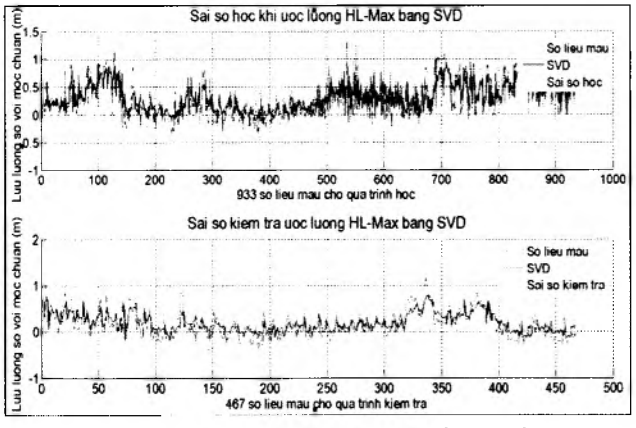
Tổng hợp lại ta có mô hình được lựa chọn để dự báo giá trị HL_{max} của ngày thứ d sẽ gồm 20 số liệu quá khứ có hệ số ảnh hưởng lớn nhất đến ngày dự báo d, như phương trình (16).

$$
\begin{aligned}
HL_{\max}(d) = {} & 0.522 \cdot HL_{\max}(d-1) + 0.142 \cdot HL_{\max}(d-4) + \\
& + 0.107 \cdot HL_{\max}(d-8) + 0.108 \cdot HL_{\max}(d-12) + \\
& + 0.081 \cdot HL_{\max}(d-1) + 0.117 \cdot HL_{\min}(d-1) - \\
& - 0.065 \cdot HL_{\min}(d-15) + 0.092 \cdot HL_{\min}(d-18) + \\
& + 0.539 \cdot HL_{TB}(d-1) - 0.211 \cdot HL_{TB}(d-2) + \\
& + 0.160 \cdot HL_{TB}(d-4) + 0.085 \cdot HL_{TB}(d-8) + \\
& + 0.084 \cdot TL_{\max}(d-1) + 0.198 \cdot TL_{\min}(d-5) + \\
& + 0.176 \cdot TL_{\min}(d-12) - 0.113 \cdot TL_{\min}(d-15) + \\
& + 0.167 \cdot TL_{\min}(d-16) + 0.138 \cdot TL_{TB}(d-4) + \\
& + 0.119 \cdot TL_{TB}(d-13) - 0.115 \cdot TL_{TB}(d-14)
\end{aligned} \quad (16)
$$

Kiểm tra chất lượng của mô hình, tác giả sử dụng 467 ngày số liệu cuối trong tập số liệu 1.460 ngày. Kết quả tính toán các thành phần sai số được thể hiện trong Bảng 1.

Bảng 1. *Kết quả sai số khi sử dụng mô hình tuyến tính (SVD) để ước lượng HL_{max}, HL_{min}*

	Sai số học			Sai số kiểm tra		
	MAE	**MRE**	**Max MAE (%)**	**MAE**	**MRE (%)**	**MaxMAE**
HL_{max}	0,19	3,45	0,95	0,17	3,65	0,69
HL_{min}	0,17	3,02	0,93	0,16	3,62	0,65



Hình 3. *Kết quả ước lượng thành phần tuyến tính cho HL_{max} của bộ số liệu học và bộ số liệu kiểm tra*

3.1.2. Kết quả ước lượng cho HL_{min}

Thực hiện ước lượng mực nước thấp nhất (HL_{min}) tương tự ước lượng HL_{max} ta xác định các yếu tố ảnh hưởng lớn nhất đến mức nước thấp nhất (HL_{min}) cần dự báo:

- Ảnh hưởng của HL_{min} trong quá khứ đến HL_{min} dự báo, ta xác định được 3 ngày có hệ số phụ thuộc lớn là: d-1, d-12 và d-24. Tiếp tục khảo sát sự phụ thuộc của HL_{min} vào các số liệu HL_{max}, HL_{TB}, TL_{max}, TL_{min} và TL_{TB} trong quá khứ:

- Ảnh hưởng của HL_{max} trong quá khứ đến HL_{min} dự báo là d-1 và d-22. Loại ngày d-55.

- Giá trị HL_{TB} trong quá khứ ảnh hưởng đến HL_{min} dự báo gồm d-1, d-12 và d-23.

- Các giá trị TL_{max} trong quá khứ ảnh hưởng đến HL_{min} dự báo d-1, d-8 và d-23. Ngày d-33 ở xa ngày dự báo nên bỏ qua.

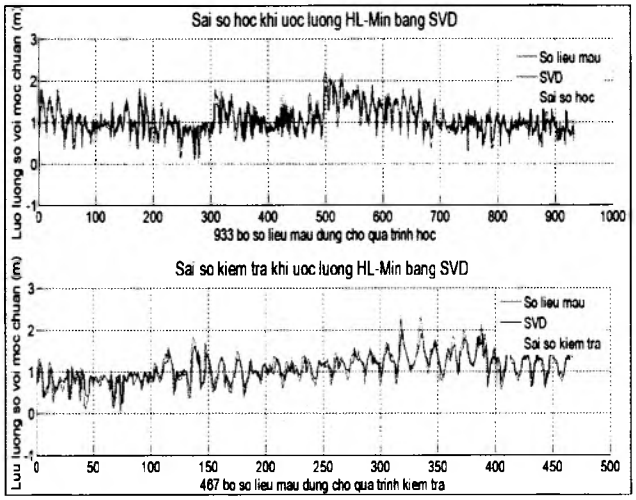
- Ảnh hưởng của TL_{min} đến HL_{min} là d-1 và d-12 Loại ngày d-43 do xa ngày dự báo.

- Ảnh hưởng của TL_{TB} đến HL_{min} là d-1, d-8 và d-24. Loại ngày d-33 do ở xa ngày dự báo.

Tổng hợp lại ta có mô hình được lựa chọn để dự báo giá trị HL_{min} của ngày thứ d sẽ gồm 16 số liệu quá khứ:

$$
\begin{aligned}
HL_{\min}(d) = {} & 0.699 \cdot HL_{\min}(d-1) + 0.160 \cdot T_{\min}(d-12) + \\
& + 0.093 \cdot HL_{\min}(d-24) + 0.101 \cdot T_{\min}(d-7) + \\
& + 0.807 \cdot T_{\max}(d-a) + 0.113 \cdot T_{\max}(d-7) + \\
& + 0.600 \cdot HL_{\max}(d-1) + 0.535 \cdot HL_{\max}(d-22) + \\
& + 0.808 \cdot HL_{TB}(d-1) + 0{,}228 \cdot HL_{TB}(d-12) + \\
& + 0.165 \cdot HL_{TB}(d-23) + 0.443 \cdot TL_{\max}(d-1) + \\
& + 0.444 \cdot TL_{\max}(d-8) + 0.361 \cdot TL_{\max}(d-23) + \\
& + 0.131 \cdot TL_{\min}(d-1) + 0.217 \cdot TL_{\min}(d-12) + \\
& + 0.377 \cdot TL_{TB}(d-1) + 0.285 \cdot TL_{TB}(d-8) + \\
& + 0.272 \cdot TL_{TB}(d-24)
\end{aligned} \quad (17)
$$

Kiểm tra chất lượng của mô hình, tác giả sử dụng 467 ngày số liệu cuối trong tập số liệu 1.460 ngày. Các kết quả tính toán sai số được thể hiện trong bảng 1. Khi ước lượng thành phần tuyến tính bằng phương pháp khai triển hệ số tuyến tính trong mô hình hỗn hợp sai số trung bình tuyệt đối kiểm tra với HL_{max} (0.17) và HL_{min} (0.16). Sau khi có mô hình tuyến tính, phần sai số còn lại sẽ được xấp xỉ bởi mô hình phi tuyến sử dụng kỹ thuật SVM và được trình bày ở mục 3.2 của bài báo.



Hình 4. *Kết quả ước lượng thành phần tuyến tính cho HL_{min} của bộ số liệu học và bộ số liệu kiểm tra*

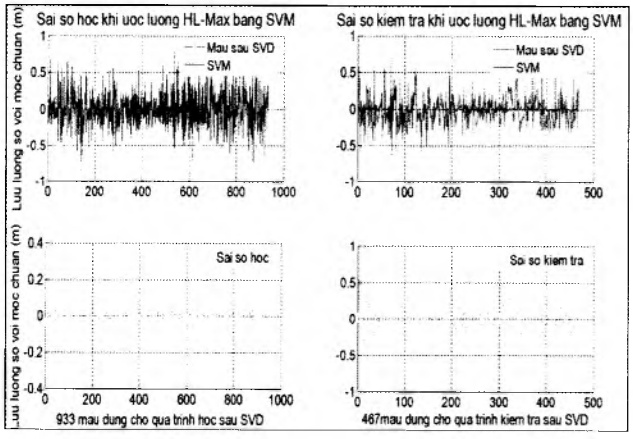
3.2. Kết quả ước lượng thành phần phi tuyến

3.2.1. Kết quả ước lượng HL_{max}

Sau khi đã xác định các thông số của mô hình tuyến tính, ta tiến hành xây dựng mạng nơron ứng với 20 đầu vào, 1 đầu ra (tương ứng với mức nước cao nhất cần dự báo); Kết quả các thành phần sai số khi ước lượng phi tuyến như Bảng 2.

Bảng 2. *Kết quả sai số khi sử dụng SVM ước lượng thành phần phi tuyến để dự báo HL_{max}, HL_{min}*

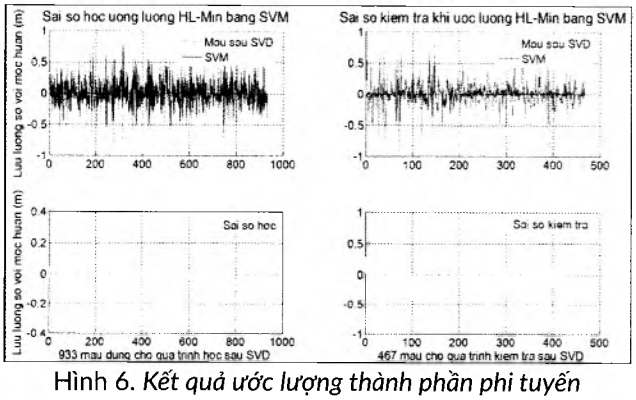
	Sai số học			Sai số kiểm tra		
	MAE	MRE (%)	MaxMAE	MAE	MRE (%)	MaxMAE
HL_{max}	0.017	1,52	0.22	0.12	2,35	0.61
HL_{min}	0.022	1,43	0.26	0.12	2.33	0.87



Hình 5. *Sai số học và sai số kiểm tra khi ước lượng HL_{max} bằng SVM trong mô hình phi tuyến*

3.2.2. Kết quả ước lượng HL_{min}

Sau khi đã xác định các thông số của mô hình tuyến tính, ta tiến hành xây dựng mô hình phi tuyến với 16 đầu vào, 1 đầu ra (ứng với mực nước thấp nhất cần dự báo); Kết quả các thành phần sai số khi ước lượng phi tuyến như Bảng 2. Sai số trung bình tuyệt đối khi ước lượng HL_{max} và HL_{min} bằng SVM trong mô hình phi tuyến tương đối nhỏ khoảng 0,12. Với sai số này phù hợp với bài toán dự báo mực nước sông.



Hình 6. *Kết quả ước lượng thành phần phi tuyến cho HL_{min} bằng SVM trong mô hình phi tuyến*

4. KẾT LUẬN

Khi ước lượng các bài toán phi tuyến, để giảm bớt mức độ phức tạp của giải pháp, mô hình hỗn hợp tách riêng thành phần tuyến tính và thành phần phi tuyến để xử lý.

Thành phần tuyến tính được xác định thông qua việc sử dụng khai triển theo cá giá trị kỳ dị (SVD). Thuật toán này cho phép xác định được hàm quan hệ tuyến tính mực nước sông cao nhất (hoặc thấp nhất) của một ngày và các ngày trước đó từ hệ các phương trình ước lượng xấp xỉ được viết dưới dạng ma trận có số hàng nhiều hơn số cột.

Thành phần phi tuyến được xác định thông qua việc sử dụng kỹ thuật máy học véctơ hỗ trợ SVM; Qua thực nghiệm cho thấy sai số học và sai số kiểm tra khi dự báo ngắn hạn mực nước sông khu vực hạ lưu cao nhất (HL_{max}) và thấp nhất (HL_{min}) đạt khoảng 0,12, với sai số này phù hợp cho bài toán dự báo. Vì vậy, ta thấy rằng phối hợp SVD và SVM trong mô hình hỗn hợp cho bài toán dự báo mực nước sông là phù hợp.

TÀI LIỆU THAM KHẢO

[1]. Trần Hoài Linh (2009), *Mạng nơron và ứng dụng trong xử lý tín hiệu*, NXB Bách Khoa.

[2]. Nguyễn Quân Nhu (2009), *Nghiên cứu và ứng dụng mạng nơron và logic mờ cho bài toán dự báo phụ tải điện ngắn hạn*, Luận án Tiến sĩ.

[3]. Đỗ Văn Đỉnh, Đinh Văn Nhượng và Trần Hoài Linh (2015), *Ứng dụng mô hình hỗn hợp trong ước lượng giá trị lớn nhất và nhỏ nhất của nhiệt độ môi trường ngày*, Tạp chí khoa học và công nghệ - Đại học Đà Nẵng, số 11(96).2015, quyển 2, trang 35-39.

[4]. Hồ Việt Hùng, Lê Xuân Hiền, Giha Lee (2018), *Ứng dụng mạng thần kinh nhân tạo dự báo lưu lượng dòng chảy sông Hồng tại Sơn Tây dựa trên dữ liệu thượng lưu*, Tuyển tập công trình Hội nghị khoa học Cơ học Thủy khí toàn quốc lần thứ 2018.

[5]. Hồ Việt Hùng, Hồ Việt Tấn (2020), *Ứng dụng mạng nơron LSTM (Long Short-Term Memory neural networks) dự báo mực nước tại cống-Âu Thuyền Cầu Cất trong hệ thống Bắc Hưng Hải*, Hội nghị.

[6]. V. Vapnil (1995), *Support-Vector Networks, Machine Learning*, 20, 273-297.

[7]. A. Filippo et al (2012), *Application of Artificial Neural Network (ANN) to improve forecasting of sea, Ocean & Coastal Management 55*, pp. 101-110.

[8]. Erik M.M.R et al (2014), *Comparison of artificial neural networks and harmonic analysis for sea level forecasting (Urias coastal lagoon, Mazatlan, Mexico)*, Ciencias Marinas, 40(4), pp. 251-261.

[9]. Lyubka Pashova and Silviya Popova (2011), *Daily sea level forecast at tide gauge Burgas, Bulgaria using artificial neural networks*, Journal of Sea Research 66, pp. 154-161.

[10]. Mohammed s. et al (2015), *Development of a sequential Artificial Neural Network for predicting river water levels based on Brahmaputra and Ganges water levels*, Neural Comput Applic, DOI 10.1007/s00521-015-1871-6.

[11]. Fazlina A.R. etal (2014), *Flood water level modeling and prediction using NARX neural network: Case Study at Kelang River*, IEEE 10th International Colloquium on Signal Processing & ít Applications (CSPA2014), 7-9 Mac. 2014, Kuala Lumpua, Malaysia, pp. 204-207.

[12]. H. Wang and D. Hu (2005), *Comparison of SVM and LS-SVM for regression, in Neural Networks and Brain*, 2005. ICNN&B'05. International Conference on, vol. 1. IEEE, 2005, pp. 279–283.

[13]. Ani Shabri (2015), *Least Square Support Vector Machines as an Alternative Method in Seasonal Time Series Forecasting*, Applied Mathematical Sciences, Vol. 9, 2015, no. 124, pp. 6207 – 6216.

[14]. T. Joachims (1998), *Making large-Scale Support Vector Machine Learning Practical, in Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press, Cambridge, MA, 1998.

[15]. D.E. Rumelhart; G.E. Hinton and R.J. Williams (1986), *Learning internal representations by error propagation*, Rumelhart, D.E. et al. (eds.): Parallel distributed processing: Explorations in the microstructure of cognition (Cambridge MA.: MIT Press), 318-362.

[16]. R.A. Fisher (1936), *The Use of Multiple Measurements in Taxonomic Problems*, in Annals of Eugenics, No 7, 1936, pp. 179-188.

THÔNG TIN TÁC GIẢ

Đỗ Văn Đỉnh

- Tóm tắt quá trình đào tạo, nghiên cứu (thời điểm tốt nghiệp và chương trình đào tạo, nghiên cứu):

+ Năm 1998: Tốt nghiệp Đại học chuyên ngành Tự động hóa, Trường Đại học Bách khoa Hà Nội;

+ Năm 2005: Tốt nghiệp Thạc sĩ chuyên ngành Kỹ thuật tự động hóa, Trường Đại học Bách khoa Hà Nội;

+ Năm 2018: Tốt nghiệp Tiến sĩ chuyên ngành Kỹ thuật điều khiển và Tự động hóa, Trường Đại học Bách khoa Hà Nội;

- Công việc hiện tại: Cán bộ quản lý, nghiên cứu và giảng viên Trường Đại học Sao Đỏ;

- Các nghiên cứu chính: Ứng dụng trí tuệ nhân tạo trong các giải pháp đo lường, điều khiển và tự động hóa, các thiết bị đo thông minh, hệ chuyên gia;

- Email: dinh.dv@saodo.edu.vn, dodinh75@gmail.com;

- Điện thoại: 0982 586 160

Nguyễn Trọng Quỳnh

- Tóm tắt quá trình đào tạo, nghiên cứu (thời điểm tốt nghiệp và chương trình đào tạo, nghiên cứu):

+ Năm 2005: Tốt nghiệp Đại học Kỹ thuật công nghiệp Thái Nguyên, chuyên ngành Tự động hóa;

+ Năm 2018: Tốt nghiệp Thạc sĩ ngành Kỹ thuật điện tử, Trường Đại học Sao Đỏ;

- Công việc hiện tại: Giảng viên khoa Điện, Trường Đại học Sao Đỏ;

- Lĩnh vực quan tâm: Tự động hóa;

- Email: trongquynhk36ib@gmail.com;

- Điện thoại: 0986 836 399.

Vũ Văn Cảnh

- Tóm tắt quá trình đào tạo, nghiên cứu (thời điểm tốt nghiệp và chương trình đào tạo, nghiên cứu):

+ Năm 2012: Tốt nghiệp Đại học chuyên ngành CNTT, Trường Đại học Sư phạm Hà Nội;

- Tóm tắt công việc hiện tại: Giáo viên Trung tâm GDNN&GDTX Thành phố Chí Linh, tỉnh Hải Dương;

- Lĩnh vực quan tâm: Xử lý ảnh, Multimedia, tăng cường thực tại ảo;

- Email: tuancanh482@gmail.com;

- Điện thoại: 0987 930 925.

Phạm Văn Nam

- Tóm tắt quá trình đào tạo, nghiên cứu (thời điểm tốt nghiệp và chương trình đào tạo, nghiên cứu):

+ Năm 2006: Tốt nghiệp Đại học chuyên ngành Đo lường, Trường Đại học Bách khoa Hà Nội;

+Năm 2008: Tốt nghiệp Thạc Sĩ chuyên ngành Đo lường, Trường Đại học Bách khoa Hà Nội;

+ Năm 2018: Tốt nghiệp tiến sĩ chuyên ngành Kỹ thuật điều khiển và Tự động hóa, Trường Đại học Bách khoa Hà Nội;

- Tóm tắt công việc hiện tại: Giảng viên, khoa Điện, Trường Đại học Công nghiệp Hà Nội;

- Lĩnh vực quan tâm: Thị giác máy và mạng nơron AI;

- Email: nampv@haui.edu.vn;

- Điện thoại: 0979 427 781.