ỨNG DỤNG MẠNG SVM TRONG MÔ HÌNH HỖN HỢP CHO BÀI TOÁN DỰ BÁO THÔNG SỐ THỜI TIẾT

APPLICATION OF SVM NETWORK IN A HYBRID MODEL FOR WEATHER FORECASTING

Đỗ Văn Đỉnh

TÓM TẮT

Dự báo thời tiết là bài toán có tính thực tiễn và có ý nghĩa quan trọng đối với ngành nông nghiệp, công nghiệp và dịch vụ. Đã có nhiều phương pháp để xuất để dự báo thông số thời tiết này [3, 7, 8, 10], tuy nhiên các thông số của mô hình dự báo phụ thuộc vào điều kiện địa lý và sự phát triển kinh tế của khu vực cần dự báo. Do đó, đối với các khu vực dự báo khác nhau cần phải xác định lại các thông số của mô hình hoặc đề xuất mô hình mới phù hợp hơn. Bài báo đề xuất sử dụng mạng SVM *(Support Vector Machine)* trong mô hình hỗn hợp [2] để dự báo thời tiết (nhiệt độ lớn nhất và nhỏ nhất) trong ngày. Các số liệu đầu vào là giá trị lớn nhất, nhỏ nhất của nhiệt độ, độ ẩm, tốc độ gió và giá trị trung bình của lượng mưa, số giờ nắng ngày trước đó. Đầu vào mô hình được đánh giá và lựa chọn sử dụng thuật toán khai triển theo giá trị kỳ dị SVD (*Singular Value Decomposition*). Chất lượng của giải pháp đề xuất được kiểm nghiệm trên số liệu quan trắc thực tế (2191 ngày từ 01/01/2010 đến 31/12/2015) ở tỉnh Hải Dương.

Từ khóa: *Mô hình hỗn hợp, máy véc-tơ đỡ, dự báo thông số thời tiết.*

ABSTRACT

Weather forecast is a practical problem and have important implications for agriculture, industry and other services. There have been different proposed methods to forecast the weather parameters [3, 7, 8, 10], but the parameters of the prediction model depends on the geographical conditions and the economic development of the given area. Therefore, for every new location, we need to find the parameters of the model or to propose a more suitable model. This paper proposes to use the SVM network (*Support Vector Machine*) in a hybrid model [2] to forecast the daily weather parameters (maximum temperature and minimum temperature). The input data is the historical values of maximum and minimum temperatures, humidity, wind speed and average values of rainfall, sun hours for past days. Model inputs are evaluated and selected using linear decomposition coefficients estimated using SVD (*Singular Value Decomposition*). The quality of the proposed solution is tested on real environment data (taken from 01/01/2010 to 31/12/2015, 2191 days) of Hai Duong province.

Keywords: *Hybrid model, support vector machines, environment parameters estimation.*

Trường Đại học Sao Đỏ
Email: dodinh75@gmail.com
Ngày nhận bài: 10/10/2018
Ngày nhận bài sửa sau phản biện: 18/10/2019
Ngày chấp nhận đăng: 20/02/2020

1. ĐẶT VẤN ĐỀ

Dự báo nhiệt độ không khí là một trong những nội dung chính của dự báo thời tiết, nó có ý nghĩa quan trọng đối với ngành nông nghiệp, công nghiệp và dịch vụ, nhằm phòng chống và hạn chế thiên tai, thiết lập kế hoạch sản xuất, khai thác tiềm năng khí hậu.

Diễn biến của nhiệt độ không khí rất phức tạp, nó chịu ảnh hưởng của rất nhiều các yếu tố khác như độ ẩm, áp suất khí quyển, lượng mưa, tốc độ gió, bức xạ nhiệt, sự phát triển các thành phần kinh tế,… Hiện nay, các mô hình dự báo nhiệt độ sử dụng phổ biến nhất được chia thành hai dạng là mô hình dự báo tất định *(Deterministic Model)* và mô hình dự báo thống kê *(Statistical Model)* [2]. Trong đó, mô hình dự báo tất định được xây dựng dựa trên quá trình diễn biến thời tiết, nó đòi hỏi một hệ thống cơ sở hạ tầng đủ mạnh và người vận hành có trình độ về công nghệ thông tin. Ngược lại, các mô hình dự báo thống kê đơn giản hơn, nó không đòi hỏi quá cao về mặt cơ sở hạ tầng hay quá chi tiết về các thông số ảnh hưởng đến thông số thời tiết cần dự báo vì mô hình này có khả năng tự động xây dựng mối quan hệ tuyến tính cũng như phi tuyến giữa các thông số cần dự báo và các thông số khác.

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 nơ-rôn nhân tạo *(ANN - Artificial Neural Network)* [6-10], trong số đó, các mô hình ứng dụng mạng nơ-rôn nhân tạo đã đạt được những tiến bộ đáng kể và nghiên cứu ứng dụng rộng rãi trong thời gian qua [1, 6, 7, 9]. Thuật toán máy véc-tơ đỡ SVM được Vapnik giới thiệu năm 1995 [4], đã được nghiên cứu thử nghiệm trong lĩnh vực dự báo thời tiết và thu được những kết quả khả quan, trong hầu hết các nghiên cứu đã được công bố, mô hình dự báo nhiệt độ không khí dùng mạng SVM đều cho kết quả tốt hơn so với các mô hình ANN kiểm chứng [8-11]. Mặt khác, trong bài báo này nhóm tác giả ứng dụng mạng nơ-rôn SVM trong mô hình hỗn hợp [2] để dự báo nhiệt độ không khí, kết quả nghiên cứu thực nghiệm cho thấy ứng dụng mạng SVM trong mô hình hỗn hợp dự báo nhiệt độ không khí cho kết quả khả quan hơn so với các mô hình mạng ANN khác (như mạng RBF, MLP, MLR, Elman, BRtree,…).

2. ỨNG DỤNG PHỐI HỢP SVD VÀ SVM TRONG MÔ HÌNH HỖN HỢP ĐỂ DỰ BÁO

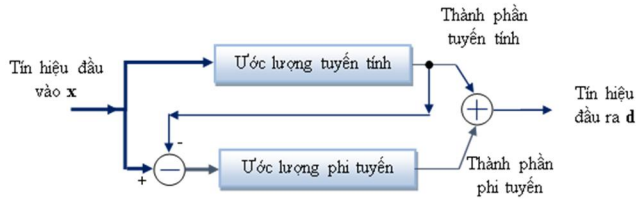
2.1. Mô hình hỗn hợp

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 [1, 2]. Theo [2], mô hình hỗn hợp đã được 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 tác giả đề xuất ứng dụng phối hợp SVD và SVM trong mô hình hỗn hợp để dự báo nhiệt độ thấp nhất (T_{min}) và nhiệt độ cao nhất (T_{max}) trong ngày.

2.1.1. Cấu trúc của mô hình hỗn hợp

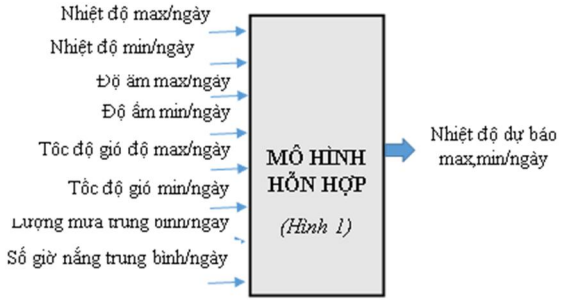
Sơ đồ cấu trúc của mô hình hỗn hợp được trình bày như hình 1, tín hiệu đầu vào (**x**) là véc-tơ chứa các số liệu quá khứ; tín hiệu đầu ra (**d**) là tổng của hai thành phần ước lượng: ước lượng tuyến tính và ước lượng phi tuyến.



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 nhiệt độ cao nhất, thấp nhất trong ngày như hình 2 .



Hình 2. Cấu trúc mô hình dự báo nhiệt độ 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:

$$\mathbf{d} = f(x) \approx \text{Linear}(x) + \text{NonLinear}(x) \quad (1)$$

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 $\{\mathbf{x}_i, \mathbf{d}_i\}$, 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:

$$\forall i: \text{Linear}(\mathbf{x}_i) \approx \mathbf{d}_i$$
$$\text{hay } e = \frac{1}{2}\sum_{i=1}^{p}\left\|\text{Linear}(\mathbf{x}_i) - \mathbf{d}_i\right\|^2 \to \min \quad (2)$$

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:

$$\forall i: \text{NonLinear}(\mathbf{x}_i) \approx \mathbf{d}_i - \text{Linear}(\mathbf{x}_i) \text{ hay}$$
$$e = \frac{1}{2}\sum_{i=1}^{p}\left\|\text{NonLinear}(\mathbf{x}_i) - (\mathbf{d}_i - \text{Linear}(\mathbf{x}_i))\right\|^2 \to \min \quad (3)$$

Giả thiết rằng giá trị T_{max} được ước lượng theo (5) (Giá trị T_{min} làm tương tự):

$$\begin{aligned} T_{max}(d) \approx f_{1,2,...,K}(&T_{max}(d-i), T_{min}(d-i), \\ &RH_{max}(d-i), RH_{min}(d-i)), Win_{max}(d-i), \\ &Win_{min}(d-i), ShAll(d-i), RainAll(d-i) + \\ &+\sum_{i=1}^{K}\begin{bmatrix} a_{i1}\cdot T_{max}(d-i) + a_{i2}\cdot T_{min}(d-i) \\ +a_{i3}\cdot RH_{max}(d-i) + a_{i4}\cdot RH_{min}(d-i) \\ +a_{i5}\cdot Win_{max} + a_{i6}\cdot Win_{min}(d-i) \\ +a_{i7}\cdot ShAll(d-i) + a_{i8}\cdot RainAll(d-i) \end{bmatrix} \end{aligned} \quad (4)$$

Trong đó, f() là hàm phi tuyến, a_{ij} là các hệ số của mô hình tuyến tính, RH_{max}: độ ẩm cao nhất trong ngày; RH_{min}: độ ẩm thấp nhất trong ngày; Win_{max}: tốc độ gió lớn nhất trong ngày; Win_{min}: tốc độ gió nhỏ nhất trong ngày; ShAll: số giờ nắng trong ngày; RainAll: lượng mưa trùng bình trong ngày. Mô hình phi tuyến được xấp xỉ bằng mạng SVM.

2.2. Các thuật toán xây dựng mô hình tuyến tính và phi tuyến

2.2.1. Ứng dụng thuật toán SVD để tối ưu hóa mô hình tuyến tính [1, 2]

Bài toán xây dựng mô hình tuyến tính có thể đưa về giải tìm nghiệm **x** của hệ phương trình: **A.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:

$$\min\|A.x - b\| = \min\|r\| = ? \quad (6)$$

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

$$\mathbf{A}^+ = \mathbf{U.S}^+\mathbf{.V}^T \quad (7)$$

với **U**, **V** là các ma trận trực giao

$$\mathbf{S}^+ = \text{diag}\left(\frac{1}{\sigma_1}, \frac{1}{\sigma_2}, ..., \frac{1}{\sigma_r}\right) \in \mathbb{R}^{nxm} \text{ - ma trận đường chéo.}$$

Khi đó nghiệm tối ưu của phương trình (5) được xác định bởi:

$$\mathbf{x} = \mathbf{A}^+\mathbf{.b} \quad (8)$$

2.2.2. Mạng SVM và ứng dụng ước lượng thành phần phi tuyến

Cho tập dữ liệu gồm N mẫu huấn luyện $\{(x_1, y_1), \ldots, (x_N, y_N)\}$ trong đó xi $\in R^D$ là các véc-tơ đầu vào (D chiều) và $y_i \in \{\pm 1\}$ là

mã lớp của véc-tơ 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.x + b = 0 để tách tập dữ liệu trên thành 2 lớp, trong đó w là véc-tơ 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 [12], thuật toán biệt thức tuyến tính của Fisher [13]. Tuy nhiên, trong thuật toán SVM, siêu phẳng tối ưu được cho là siêu phẳng có tổng khoảng cách tới các véc-tơ 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:

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

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

trong đó, C > 0 là tham số chuẩn tắc (*Regularization Parameter*), ξ_i là biến lỏng. Bài toán (9) 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 (10), ta sẽ thu được các giá trị tối ưu w* và b* của siêu phẳng. Chỉ có các mẫu có $\alpha_i \geq 0$ mới được gọi là các véc-tơ đỡ. Cuối cùng, hàm đầu ra 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. Mô hình hỗn hợp ước lượng T_{max}, T_{min} trong ngày

2.2.3.1. Ước lượng thành phần tuyến tính

Từ phương trình (4), hàm quan hệ tuyến tính giữa T_{max} của ngày d với T_{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 véc-tơ $\mathbf{a} = [a_1, a_2, ..., a_k]^T$ 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?.

$$\begin{cases} a_1 \cdot T_{max}(K) + a_2 \cdot T_{max}(K-1) + ... + a_k \cdot T_{max}(d-K) \approx T_{max}(K+1) \\ a_1 \cdot T_{max}(K-1) + a_2 \cdot T_{max}(K-2) + ... + a_k \cdot T_{max}(d-K+1) \approx T_{max}(K+2) \\ \quad ... \qquad\qquad ... \qquad\qquad ... \\ a_1 \cdot T_{max}(N_{max}-1) + a_2 \cdot T_{max}(N_{max}-2) + ... + a_k \cdot T_{max}(N_{max}-K) \approx T_{max}(N_{max}) \end{cases} \quad (12)$$

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

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 ta sử dụng K = 60 - tương đương 2 tháng số liệu trước đó - là đủ lớn để dự báo ngày tiếp theo).

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

- Xác định thành phần có giá trị tuyệt đối nhỏ nhất trong véc-tơ **a**. Thành phần này sẽ tương ứng với ngày trong quá khứ ít ảnh hưởng tới ngày dự báo. 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 như vậy ta xây dựng hàm quan hệ tuyến tính giữa T_{max} của ngày d với T_{min}, RH_{max}, RH_{min}, Win_{max}, Win_{min}, RainAll và RhAll của các quá khứ ta được phương trình (14).

$$T_{max}(d) \approx \sum_{i=1}^{K} \begin{Bmatrix} a_{i1} \cdot T_{max}(d-i) + a_{i2} \cdot T_{min}(d-i) + a_{i3} \cdot RH_{max}(d-i) \\ + a_{i4} \cdot RH_{min}(d-i) + a_{i5} \cdot Win_{max}(d-i) \\ + a_{i6} \cdot Win_{min}(d-i) + a_{i7} \cdot RainAll(d-i) + a_{i8} \cdot RhAll(d-i) \end{Bmatrix} \quad (14)$$

Khi xác định được mối quan hệ tuyến tính giữa T_{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).

$$NL(d) = T_{max}(d) - \sum_{i=1}^{K} \begin{Bmatrix} a_{i1} T_{max}(d-i) + a_{i2} \cdot T_{min}(d-i) + a_{i3} \cdot RH_{max}(d-i) \\ + a_{i4} \cdot RH_{min}(d-i) + a_{i5} \cdot Win_{max}(d-i) \\ + a_{i6} \cdot Win_{min}(d-i) \\ + a_{i7} \cdot RainAll(d-i) + a_{i8} \cdot RhAll(d-i) \end{Bmatrix} \quad (15)$$

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

2.2.3.2. Mô hình ước lượng phi tuyến

Khi xác định được các thông số mô hình tuyến tính, ta tiến hành xây dựng mạng nơ-rôn 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. Để kiểm nghiệm chất lượng các mô hình mạng nơ-rôn ước lượng thành phần phi tuyến, trong bài báo tác giả sử dụng các mô hình mạng MLP, MLR, Elman, BRtree và SVM. Các mô hình này có cấu trúc được lựa chọn bằng phương pháp thử nghiệm để chọn ra mô hình có sai số kiểm tra nhỏ nhất. Cụ thể, mạng MLP và MLR được lựa chọn có 30 nơ-rôn ẩn (1 lớp ẩn), mạng Elman có 15 nơ-rôn ẩn, mô hình BRTree được lựa chọn với 221 nút [4].

3. KẾT QUẢ VÀ THẢO LUẬN

Mô hình nghiên cứu được xây dựng trên nền phần mềm Matlab®, với SVM sử dụng LSSVMlabv1.8_R2009b_R2011a và được thiết kế theo các bước sau: chuẩn bị dữ liệu, lựa chọn đặc tính cho mô hình dự báo, xây dựng kiến trúc mạng, lựa chọn phương pháp và huấn luyện mạng, đánh giá độ tin cậy.

3.1. Kết quả ước lượng thành phần tuyến tính

3.1.1. Kết quả ước lượng T_{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 giá trị nhiệt độ cao nhất (T_{max}) cần dự báo:

- Ảnh hưởng của T_{max} trong quá khứ đến T_{max} dự báo, ta xác định được 5 ngày có hệ số phụ thuộc lớn là: d-1, d-7, d-11 và d-18. Tiếp tục khảo sát sự phụ thuộc của T_{max} vào các số liệu T_{min}, RH_{max},RH_{min}, Win_{max}, Win_{min}, RainAll, ShAll trong quá khứ bằng cách làm hoàn toàn tương tự ta được:

- Ảnh hưởng của T_{min} trong quá khứ đến T_{max} dự báo là các ngày d-1, d-7, d-12 và d-22; Ngày d-22 xa ngày dự báo nên ta có thể loại.

- Giá trị RH_{max} trong quá khứ ảnh hưởng đến T_{max} dự báo là d-1, d-2, d-4 và d-7.

- Các giá trị RH_{min} trong quá khứ ảnh hưởng đến T_{max} dự báo d-1, d-2, d-5 và d-57; Do ngày d-57 xa ngày dự báo nên loại.

- Ảnh hưởng của tốc độ gió max (Win_{max}) đến T_{max} là d-1, d-2, d-30 và d-59; Các ngày d-30 và d-59 xa ngày dự báo nên loại.

- Ảnh hưởng của tốc độ gió min (Win_{min}) đến T_{max} là d-1, d-7, d-11 và d-52; Ngày d-52 loại do xa ngày dự báo.

- Sự phụ thuộc của T_{max} vào lượng mưa trung bình là các ngày d-51, d-55, d-57 và d-60. Các ngày này xa ngày dự báo nên loại.

- Ảnh hưởng của số giờ nắng ngày tới T_{max} là d-24, d-50, d-56 và d-60. Loại do xa ngày dự báo.

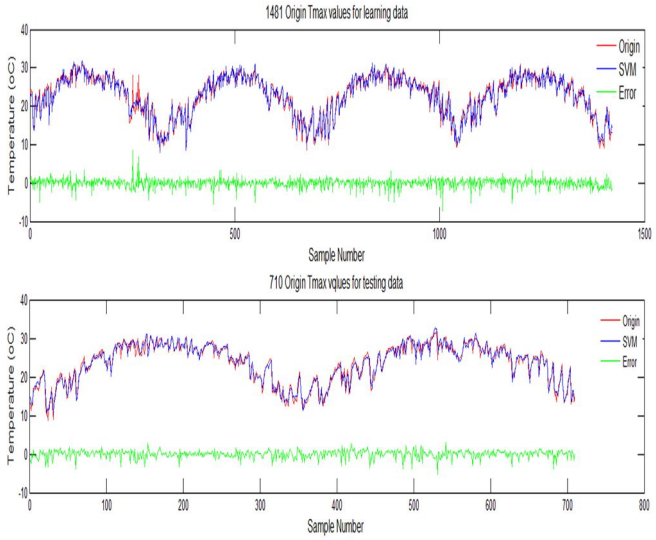
Tổng hợp lại ta có mô hình được lựa chọn để dự báo giá trị T_{max} của ngày thứ d sẽ gồm 19 số liệu quá khứ:

$$\begin{aligned} T_{max}(d) = {} & 0,808 \cdot T_{max}(d-1) + 0,084 \cdot T_{max}(d-7) \\ & + 0,062 \cdot T_{max}(d-11) + 0,07 \cdot T_{max}(d-18) \\ & + 0,828 \cdot T_{min}(d-1) + 0,077 \cdot T_{min}(d-7) \\ & + 0,067 \cdot T_{min}(d-12) + 0,571 \cdot RH_{max}(d-1) \\ & + 0,101 \cdot RH_{max}(d-2) + 0,059 \cdot RH_{max}(d-5) \\ & + 0,081 \cdot Win_{max}(d-1) + 0,044 \cdot Win_{max}(d-2) \\ & + 0,071 \cdot Win_{min}(d-1) + 0,054 \cdot Win_{min}(d-7) \\ & + 0,05 \cdot Win_{min}(d-11) \end{aligned}$$

Kiểm tra chất lượng của mô hình sử dụng 710 ngày số liệu cuối trong tập số liệu 2191 ngày. Các kết quả tính toán được thể hiện trong bảng 1.

Bảng 1. Kết quả sai số khi sử dụng mô hình tuyến tính để ước lượng T_{max}, T_{min}

	Sai số học			Sai số kiểm tra		
	MAE	MRE (%)	MaxMAE	MAE	MRE (%)	MaxMAE
T_{max}	0,78	3,83	8,47	0,75	3,53	5,30
T_{min}	1,04	5,34	8,78	1,01	4,95	7,42



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

3.1.2. Kết quả ước lượng cho T_{min}

Thực hiện ước lượng nhiệt độ thấp nhất (T_{min}) tương tự T_{max} ta xác định các yếu tố ảnh hưởng lớn nhất đến giá trị nhiệt độ thấp nhất (T_{min}) cần dự báo:

- Ảnh hưởng của T_{min} trong quá khứ đến T_{min} dự báo, ta xác định được 5 ngày có hệ số phụ thuộc lớn là: d-1, d-2, d-3 và d-7. Tiếp tục khảo sát sự phụ thuộc của T_{min} vào các số liệu T_{max}, RH_{max}, RH_{min}, Win_{max}, Win_{min}, RainAll và ShAll trong quá khứ:

- Ảnh hưởng của T_{max} trong quá khứ đến T_{min} dự báo là d-1, d-7, d-11 và d-60. Loại ngày d-60.

- Giá trị RH_{max} trong quá khứ ảnh hưởng đến T_{min} dự báo gồm d-1, d-4, d-7 và d-12.

- Các giá trị RH_{min} trong quá khứ ảnh hưởng đến T_{min} dự báo d-1, d-2, d-6 và d-55. Ngày d-55 ở xa ngày dự báo nên bỏ qua.

- Ảnh hưởng của tốc độ gió Win_{max} đến T_{min} là d-1, d-2, d-28 và d-59. Loại ngày d-28, d-59 do xa ngày dự báo.

- Ảnh hưởng của tốc độ gió Win_{min} đến T_{min} là d-1, d-2, d-30 và d-60. Loại ngày d-30, d-60 do ở xa ngày dự báo.

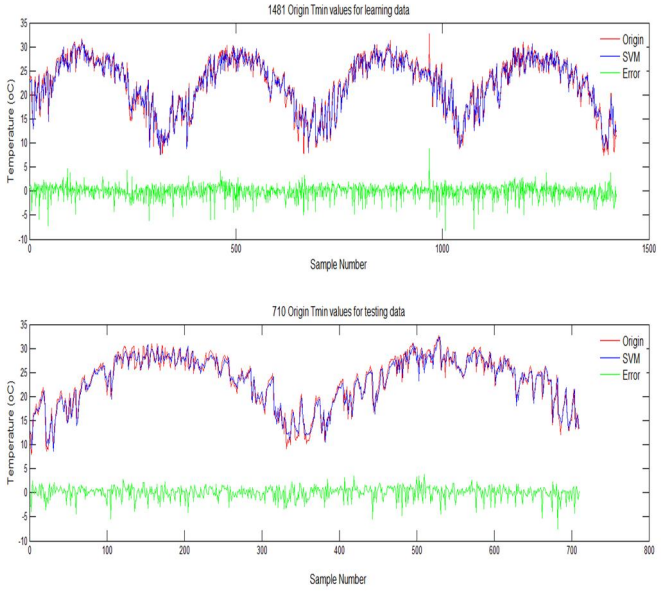
- Lượng mưa trung bình ngày không ảnh hưởng đến T_{min} dự báo do ở xa ngày dự báo.

- Số giờ nắng các ngày ảnh hưởng đến T_{min} là d-1, d-55, d-56 và d-60. Loại các ngày d-50, d-56 và d-60 do ở xa ngày dự báo.

Tổng hợp lại ta có mô hình được lựa chọn để dự báo giá trị T_{min} của ngày thứ d sẽ gồm 18 số liệu quá khứ:

$$T_{min}(d) = 0,89 \cdot T_{min}(d-1) - 0,135 \cdot T_{min}(d-2) + 0,073 \cdot T_{min}(d-3) + 0,101 \cdot T_{min}(d-7) + 0,807 \cdot T_{max}(d-a) + 0,113 \cdot T_{max}(d-7) + 0,075 \cdot T_{max}(d-7) + 0,63 \cdot RH_{max}(d-1) + 0,127 \cdot RH_{max}(d-4) + 0,094 \cdot RH_{max}(d-12) + 0,791 \cdot RH_{min}(d-1) + 0,092 \cdot RH_{min}(d-2) + 0,077 \cdot RH_{min}(d-6) + 0,09 \cdot Win_{max}(d-1) + 0,037 \cdot Win_{max}(d-2) + 0,079 \cdot Win_{min}(d-1) + 0,058 \cdot Win_{min}(d-2) + 0,067 ShAll(d-1)$$

Kiểm tra chất lượng của mô hình sử dụng 710 ngày số liệu cuối trong tập số liệu 2191 ngày. Các kết quả tính toán được thể hiện trong bảng 1.



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

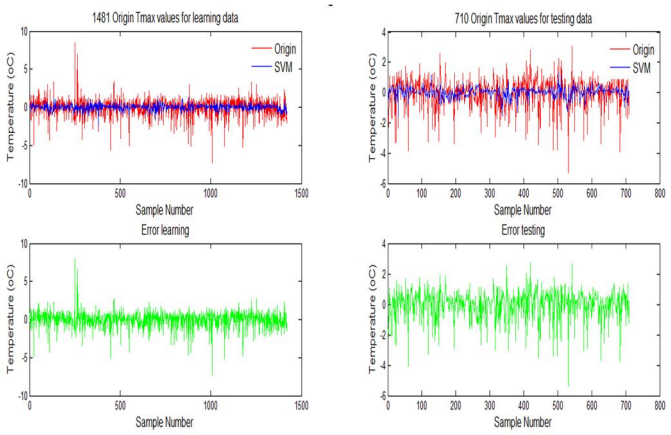
3.2. Kết quả ước lượng thành phần phi tuyến

3.2.1. Kết quả ước lượng T_{max}

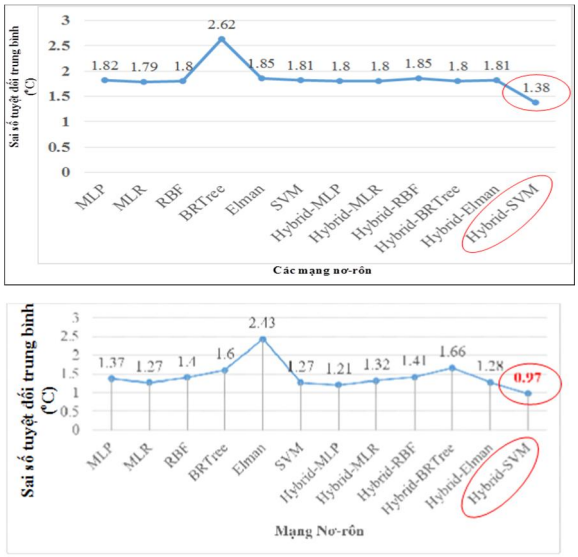
Sau khi đã xác định các thông số của mô hình tuyến tính, ta tiến hành xây dựng mạng Nơ-rôn ứng với 19 đầu vào, 1 đầu ra (ứng với giá trị nhiệt độ cao nhất cần dự báo); Kết quả các thành phần sai số khi ước lượng phi tuyến như bảng 2.

Bảng 2. Tổng hợp sai số khi sử dụng các mô hình mạng nơ-rôn khác nhau ước lượng T_{max}, T_{min}

Mạng nơ-rôn	Sai số học						Sai số kiểm tra					
	MAE		MRE (%)		MaxMAE		MAE		MRE (%)		MaxMAE	
	T_{max}	T_{min}	T_{max}	T_{min}	T_{max}	T_{min}	T_{max}	T_{min}	T_{max}	T_{min}	T_{max}	T_{min}
MLP	1,08	1,34	5,11	6,58	1,08	1,34	1,02	1,37	4,62	6,36	1,02	1,39
MLR	0,78	1,04	3,83	5,35	8,47	8,79	0,75	1,02	3,52	4,98	5,30	7,48
SVM	**0,71**	**0,93**	**3,43**	**4,65**	**8,13**	**8,33**	**0,70**	**0,97**	**3,28**	**4,67**	**5,31**	**7,43**
Elman	0,95	1,30	4,49	6,38	0,95	1,30	1,05	1,40	4,74	6,50	1,05	1,40
BRTree	0,38	0,52	1,79	2,56	4,50	5,36	0,97	1,41	4,51	6,61	7,75	7,43

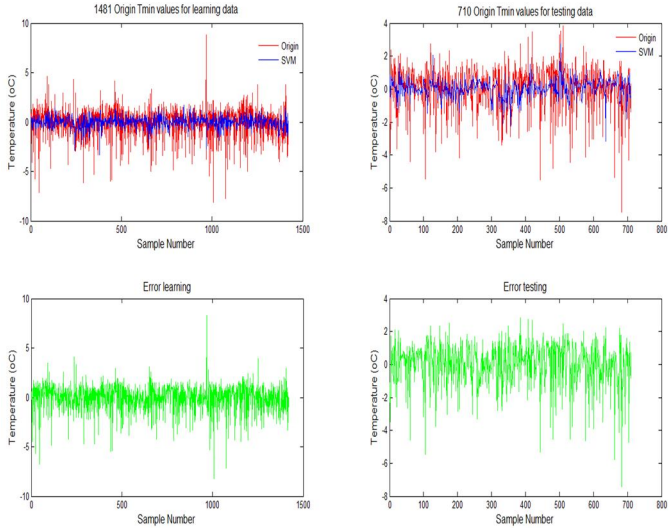


Hình 5. Sai số học và sai số kiểm tra kết quả ước lượng thành phần phi tuyến T_{max}



Hình 6. Đồ thị biểu diễn sai số tuyệt đối trung bình của các mô hình đã thử nghiệm khi dự báo T_{max} (trái) và T_{min} (phải)

3.2.2. Kết quả ước lượng T_{min}



Hình 7. Kết quả ước lượng thành phần phi tuyến T_{min} cho bộ số liệu học và kiểm tra

Sau khi đã xác định các thông số của mô hình tuyến tính, ta tiến hành xây dựng mạng nơ-rôn ứng với 18 đầu vào, 1 đầu ra (ứng với giá trị nhiệt độ thấp nhất cần dự báo); Kết quả các thành phần sai số khi ước lượng phi tuyến như bảng 2.

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 giữa nhiệt độ 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 mô hình mạng nơ-rôn khác nhau; Qua thực nghiệm cho thấy sai số học và sai số kiểm tra khi dự báo ngắn hạn nhiệt độ cao nhất (T_{max}) và thấp nhất (T_{min}), kết quả thu được tốt nhất khi sử dụng mạng SVM. Vì vậy, ta thấy rằng ứng dụng mạng SVM trong mô hình hỗn hợp cho bài toán dự báo một số thông số thời tiết là phù hợp, sai số học và sai số kiểm tra ở mức trung bình, đặc biệt là sai số kiểm tra sẽ có giá trị tương đối ổn định. Kết quả sai số trung bình tuyệt đối dưới 1%.

TÀI LIỆU THAM KHẢO

[1]. Trần Hoài Linh, 2009. *Mạng nơ-rôn và ứng dụng trong xử lý tín hiệu*. NXB Bách Khoa.

[2]. Nguyễn Quân Nhu, 2009. *Nghiên cứu và ứng dụng mạng nơ-rôn và lô-gic mờ cho bài toán dự báo phụ tải điện ngăn hạn*. Luận án Tiến sĩ, Đại học Bách khoa Hà Nội.

[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), quyển 2, trang 35-39.

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

[5]. Đỗ Văn Đỉnh, 2018. *Xây dựng mô hình dự báo một số thông số khí tượng cho địa bàn tỉnh Hải Dương,* Luận án Tiến sĩ, Đại học Bách khoa Hà Nội.

[6]. Parag P Kadu et al. *Temperature Prediction System Using Back propagation Neural Network An Approch*. International Journal of Computer Science & Communication Networks,Vol 2(1), pp. 61-64.

[7]. Mohsen Hayati and Zahra Mohebi, 2007. *Temperature forecating based on neural network approach*. World applied sciences journal 2(6), pp. 613-620.

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

[9]. Y.Radhika and M.Shashi, 2009. *Atmospheric Temperature Prediction using Support Vector Machines*. International Journal of Computer Theory and Engineering, Vol. 1, No. 1, pp. 55-58.

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

[11]. 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.

[12]. 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.

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

AUTHOR INFORMATION

Do Van Dinh

Sao Do University