HƯỚNG TIẾP CẬN HỒI QUY MỚI CHO DỰ BÁO TỐC ĐỘ GIÓ

Nguyễn Hoàng Huy*, Hoàng Thị Thanh Giang

Khoa Công nghệ thông tin, Học viện Nông nghiệp Việt Nam

**Tác giả liên hệ: nhhuy@vnua.edu.vn*

Ngày nhận bài: 20.07.2020 Ngày chấp nhận đăng: 08.09.2020

TÓM TẮT

Trong bài báo này, chúng tôi giới thiệu một hướng tiếp cận sử dụng hồi quy tuyến tính (Linear Regression - LR) trong hai bước, được gọi là two-step LR, để dự báo cho dữ liệu có cấu trúc không - thời gian (spatio - temporal data). Ở bước đầu tiên tất cả các đặc trưng được chia thành các nhóm con và sử dụng hồi quy tuyến tính cho mỗi nhóm con đặc trưng để có các giá trị hồi quy tương ứng với mỗi nhóm. Bước hai áp dụng hồi quy tuyến tính một lần nữa cho các giá trị hồi quy thu được ở bước một để tạo ra giá trị hồi quy cuối cùng. Cách tiếp cận sử dụng two-step LR có hiệu năng tốt nhất khi dự báo tốc độ gió. Dự báo tốc độ gió hữu ích cho tích hợp năng lượng gió vào lưới điện bởi vì năng lượng gió được sinh bởi tuabin gió, có mối quan hệ mật thiết với tốc độ gió. Sự khó dự đoán trước và thay đổi liên tục của tốc độ gió là một trong những khó khăn căn bản nhất của việc tích hợp này.

Từ khóa: Dữ liệu không - thời gian, dữ liệu số chiều cao, dự báo tốc độ gió.

A Novel Regression Approach for wind Speed Forecasting

ABSTRACT

The paper presents a spatio-temporal data forecasting approach using Linear Regression (LR) in two steps called two-step LR. In the first step, all features were divided into subgroups and Linear Regressions was utilized to obtain a regression value for each feature subgroup. In the second step, Linear Regressions was applied again to these regression values to generate the final regression value. The approach using two-step LR had state-of-the-art performance for a wind speed forecasting problem. Wind speed forecasting would be useful for the integration of wind energy into the power grid because wind power generated by wind turbines has an intimate relationship with wind speed and unpredictability and variability of wind speed is one of the fundamental difficulties of this integration system.

Keyworks: spatio-temporal data, high dimensional data, wind speed forecasting.

1. ĐẶT VẤN ĐỀ

Các hệ thống thu thập dữ liệu hiện đại có khả năng sản sinh lượng lớn dữ liệu, trong đa số trường hợp sẽ cho số lượng lớn đặc trưng ứng với mỗi mẫu dữ liệu. Trong một số trường hợp, các mẫu dữ liệu được thu thập trong thời gian dài có thể dẫn đến phân bố không ổn định, hay thậm chí là dữ liệu không liên quan, ví dụ như EEG (Nguyen Hoang Huy & cs., 2014), hoặc dữ liệu vận tốc gió (Lei & cs., 2009). Trong những trường hợp này, chúng ta có thể phân tích dữ liệu trong khoảng thời gian ngắn hơn, với số lượng mẫu dữ liệu ít hơn, để làm phân bố dữ liệu ổn định hơn (Nguyen Hoang Huy & cs., 2014). Tuy nhiên trong các bài toán hồi quy thực tế, vấn đề này sẽ dẫn đến tình trạng là số lượng mẫu dữ liệu n không đủ lớn so với số lượng đặc trưng d (vấn đề dữ liệu số chiều cao). Không may, khi n không đủ lớn so với d, vấn đề hồi quy thống kê trong cả lý thuyết và thực tế sẽ khó giải quyết hơn (Bai & cs., 2019; Bickel & Levina, 2008; Cai & Zhang, 2019; Hastie & cs., 2009; Lei & cs., 2018).

Một số hướng tiếp cận đã được đưa ra để giải quyết vấn đề hồi quy dữ liệu số chiều cao nói trên (nghĩa là khi n lớn hơn so với d). Hầu hết các phương pháp này sử dụng các mô hình

đơn giản với số tham số ít hơn, như "naive Bayes", hay hồi quy thưa (sparse regression) (Bickel & Levina, 2004; Hastie & cs., 2009; Hastie & cs., 2015), để tránh việc phải ước lượng quá nhiều tham số trong các mô hình hồi quy. Tuy nhiên, trong thực tế dữ liệu không phải lúc nào cũng thỏa mãn các giả thiết của phương pháp này. Ví dụ như trong nhiều tình huống dữ liệu không thỏa mãn giả thiết thưa, thậm chí ngay cả khi giả thiết này được thỏa mãn thì phương pháp hồi quy dựa trên giả thiết thưa, cũng không đảm bảo sẽ hoạt động tốt do vấn đề tương tác giữa các đặc trưng (Cai & Liu, 2011).

Khi dữ liệu không thỏa mãn giả thiết thưa, một tính chất quan trọng khác của dữ liệu số chiều cao thường thỏa mãn trong thực tế và được khai thác đó là tính khả tách (trong dữ liệu không - thời gian) (Bai & cs., 2019; Genton, 2007). Loại dữ liệu này có ma trận hiệp phương sai phân tách được, nghĩa là có thể viết thành tích tensor của ma trận hiệp phương sai không gian và ma trận hiệp phương sai thời gian. Cho đến nay, chỉ có một vài phương pháp sử dụng tính chất này để giải quyết các bài toán phân loại hoặc hồi quy đối với dữ liệu số chiều cao, tuy nhiên những phương pháp này yêu cầu thêm các giả thiết như mô hình trung bình cộng tính (Huizenga & cs., 2002; Leiva & Roy, 2014).

Hoang & cs. (2014) đã đề xuất phương pháp two-step LDA để tránh việc phải ước lượng đồng thời nhiều tham số khi áp dụng mô hình phân tích khác biệt tuyến tính (LDA). Two-step LDA áp dụng LDA trong hai bước thay vì một lần duy nhất cho tất cả các thuộc tính. Đầu tiên, LDA được áp dụng cho các tập con đặc trưng. Sau đó LDA được áp dụng vào các giá trị kết quả thu được từ bước thứ nhất. Two-step LDA yêu cầu tính toán ít hơn bởi vì nó không cần trải qua các quy trình tối ưu các tham số như tham số chỉnh hóa trong phân tích khác biệt tuyến tính chỉnh hóa (regularized LDA), và có hiệu năng tốt nhất trong phân loại EGG. Đối với dữ liệu có tính chất khả tách (dữ liệu thỏa mãn giả thiết ma trận hiệp phương sai khả tách), chúng tôi đã chứng minh được tỷ lệ lỗi lý thuyết của two-step LDA tương đương với phương pháp Bayes với tỉ lệ lỗi tối ưu nhất, đồng thời đưa ra hướng dẫn cách nhóm các đặc trưng trong bước đầu tiên của two-step LDA.

Trong khi two-step LDA được thiết kế để giải quyết bài toán phân loại dữ liệu số chiều cao có tính khả tách, như trong dữ liệu không - thời gian EEG và vẫn là câu hỏi mở, nếu phương pháp này có thể được mở rộng cho bài toán hồi quy. Trong bài này, chúng tôi mở rộng two-step LDA thành two-step LR để xử lý dữ liệu không - thời gian khả tách số chiều cao. Giống như two-step LDA khi phân loại dữ liệu EEG, chúng tôi chỉ ra bằng thực nghiệm rằng two-step LR hiệu quả với bài toán dự báo tốc độ gió (dữ liệu không - thời gian), cho kết quả tốt hơn phương pháp mới nhất dựa vào hướng tiếp cận học sâu (Deep Learning).

Cần lưu ý rằng, dự báo tốc độ gió là một trong những bài toán quan trọng trong khoa học khí tượng (Lei & cs., 2009). Gần đây đã có nhiều hướng tiếp cận dựa vào dữ liệu để giải quyết bài toán này từ các phương pháp phân tích thống kê theo chuỗi thời gian như Persistence Forecasting, Autoregressive Model cho đến sử dụng mạng thần kinh nhân tạo như Wavelet Transform-Based Artificial Neural Networks (WT-ANN), ANN-based ST và LS-based ST (Bali & cs., 2019; Sanandaji & cs., 2015; Tascikaraoglu & cs., 2016). Trong khi ANN-based ST, LS-based ST là những phương pháp khai thác cấu trúc không - thời gian (ST) của dữ liệu tốc độ gió, sử dụng mạng thần kinh nhân tạo, bình phương tối thiểu (LS). Gần đây thì các tác giả trong bài báo Ghaderi & cs. (2017) đã đưa mô hình học sâu LTSM cho toàn bộ dữ liệu để dự báo tốc độ gió, mà bỏ qua việc xem xét cấu trúc không - thời gian. Nó được coi là phương pháp tốt nhất hiện nay để giải quyết bài toán dự báo tốc độ gió (Bali & cs., 2019; Ghaderi & cs., 2017).

2. PHƯƠNG PHÁP NGHIÊN CỨU

Trong nghiên cứu này, chúng tôi phân tích, tổng hợp lại cơ sở lý thuyết của hồi quy tuyến tính, rồi trên cơ sở đó chúng tôi đề xuất phương pháp hồi quy mới two-step LR. Trong mô hình hồi quy tuyến tính, giả sử có các mẫu huấn luyện độc lập $\{(x_s, y_s) \in R^d \times R, s = 1, \ldots, n\}$ từ một đám

đông chưa xác định có phân bố P(x, y) nào đó. Cho một mẫu mới x của đám đông trên, chúng ta cần tìm hàm hồi quy $y = \hat{\ }$ cho vector đặc trưng x, để có thể dự đoán giá trị y chưa biết ứng với quan sát mới x càng chính xác càng tốt. Trong nghiên cứu này, chúng tôi mở rộng two-step LDA (Hoang & cs., 2014) thành two-step LR để xác định hàm hồi quy tuyến tính $\hat{\ }$. Tương tự như two-step LDA, two-step LR áp dụng hồi quy tuyến tính trong hai bước.

2.1. Hồi quy tuyến tính

Hồi quy tuyến tính đã chứng minh hiệu quả cao cho nhiều tập dữ liệu khác nhau nếu đủ nhiều mẫu huấn luyện, sao cho $\frac{d\sqrt{\log d}}{\sqrt{n}} \to 0$, xem Bickel & Levina (2008) và Hastie & cs. (2009). Tuy nhiên nếu n không đủ lớn so với d thì phương pháp này có hiệu năng không tốt, thậm chí ngay cả khi phân bố dữ liệu xấp xỉ hay là phân bố chuẩn. Chính xác hơn, khi $n < d + 1$ ma trận hiệp phương sai mẫu Σ là ma trận kỳ dị, và hồi quy tuyến tính mẫu là không xác định. Một số phương pháp đã được đưa ra để giải quyết vấn đề này như Hastie & cs. (2009) và Lei & cs. (2018). Các phương pháp phổ biến thường dựa vào kỹ thuật chỉnh hóa, như hồi quy Ridge và hồi quy tuyến tính Lasso. Các phương pháp Lasso dựa trên giải thiết thưa. Tuy nhiên, có những thuộc tính có thể làm giảm tỉ lệ lỗi của hồi quy tuyến tính Lasso hoặc phân tích khác biệt Lasso thông qua mối tương quan với những đặc trưng khác mặc dù mỗi thuộc tính đó không có ảnh hưởng gì lên hàm phân biệt hoặc hồi quy.

Trọng tâm của nghiên cứu này là đưa ra hướng tiếp cận mới cho xây dựng hàm hồi quy cho các mô hình tuyến tính trong không gian số chiều trung bình:

$$y = X\beta + \epsilon$$

trong đó $\mathbf{y} = (y_1, \ldots, y_n)^T$, X là ma trận thiết kế Gaussian kích thước $n \times d$, với mỗi hàng độc lập sinh từ cùng một phân bố $x_i \sim N(0, \Sigma)$, β là vector tham số thực sự với kích thước $d \times 1$, và ϵ là vector lỗi ngẫu nhiên kích thước $n \times 1$ với các phần tử $\epsilon_1, \ldots, \epsilon_n$ là các biến ngẫu nhiên độc lập có cùng phân bố và $E[\epsilon_i] = 0$, $Var(\epsilon_i) = \sigma^2 < \infty$, d có thể lớn hơn n ($d = 0(n)$). Để đơn giản hóa và không mất tính tổng quát, chúng ta có thể giả sử hệ số tự do của hàm hồi quy và giá trị trung bình của tất cả các biến đều bằng 0. Giả thiết này có thể đạt được bằng cách trung tâm hóa bởi trung bình mẫu. Hệ số của mô hình hồi quy tuyến tính (LR) có thể được xác định bằng phương pháp bình phương tối tiểu, nghĩa là tìm β làm tối tiểu lỗi

$$\beta = \arg\min_{\beta \in R^d} \sum_{i=1}^{n} \left\| y_i - x_i^T \beta \right\|^2$$

Để tối tiểu lỗi này, lấy đạo hàm tương ứng với β để được hệ phương trình gồm n phương trình, d ẩn. Nếu $d \le n$, giải hệ phương trình này ta được:

$$\beta = \left(X^T X\right) - 1 X^T y \left(1\right) \; n \ll$$

Khi $d + 2 > n$, $X^T X$ là một ma trận kích thước $d \times d$, nhưng hạng của nó thấp hơn n. Nếu $n \ll$, ma trận $X^T X$ không khả nghịch, thậm chí điều kiện xấu (ill-conditioned) với hầu hết các giá trị riêng bằng 0. Xây dựng mô hình hồi quy tuyến tính sử dụng phương pháp bình phương tối tiểu trong trường hợp này hoàn toàn thất bại. Cách đơn giản nhất để xử lý trường hợp này là thay thế ma trận nghịch đảo bằng ma trận giả nghịch đảo Moore-Penrose. Một vài hướng tiếp cận khác là dựa trên kỹ thuật chỉnh hóa như hồi quy Ridge, hồi quy Lasso cũng đã được đưa ra. Chúng tôi đề xuất hướng tiếp cận mới, được gọi là two-step LR.

2.2. Phương pháp two-step LR

Tương tự như two-step LDA (Nguyen Hoang Huy & cs., 2014), two-step LR cũng xử lý trong hai bước. Ở bước đầu tiên two-step LR phân chia tất cả các đặc trưng thành *q* các tập con rời nhau $x_g, x_{sg} \in R^{pg}$, $g = 1, \ldots, q$, $s = 1, \ldots, n$, $x = \left[x_1^T, \ldots, x_q^T\right]^T$, $x_s = \left[x_{s1}^T, \ldots, x_{sq}^T\right]^T$, $p_1 + \ldots + p_q = d$. Cách xác định các tập con đặc trưng là rất quan trọng và chúng tôi kế thừa từ two-step LDA. Để đơn giản hóa, trong bài báo này chúng

tôi thiết lập $p_1 = \ldots = p_q$ và $d = pq$. Sau đó hồi quy tuyến tính được áp dụng cho mỗi tập con đặc trưng $\mathbf{x}_g$ để được hàm hồi quy tuyến tính $\hat{}$

$\hat{}$ $^{T}\beta_g$

trong đó, β_g được xác định bằng cách áp dụng phương pháp bình phương tối thiểu trên các mẫu huấn luyện:

$\{x_{sg} \in R^p, g = 1,\ldots, q; s = 1,\ldots, n\}$

Trong trường hợp $p + 2 > n$, ma trận nghịch đảo ở công thức (1) được thay thế bởi ma trận giả nghịch đảo Moore-Penrose để xác định β_g. Trong bước hai, hồi quy tuyến tính được áp dụng một lần nữa với điểm kết quả tính ở bước một:

$\hat{}$ $\hat{}$ $\hat{}$ $\hat{}$ $\hat{}$

với $s = 1,\ldots, n$ để được hàm hồi quy two-step $f^*(x)$ cuối cùng. Điều đó có nghĩa $f^*(x)$ xác định như sau:

$\hat{}$ $\hat{}$ $\hat{}$

trong đó $\hat{}$ là hàm hồi quy tuyến tính. Hình 2 mô tả lược đồ của quá trình thực hiện two-step LR.

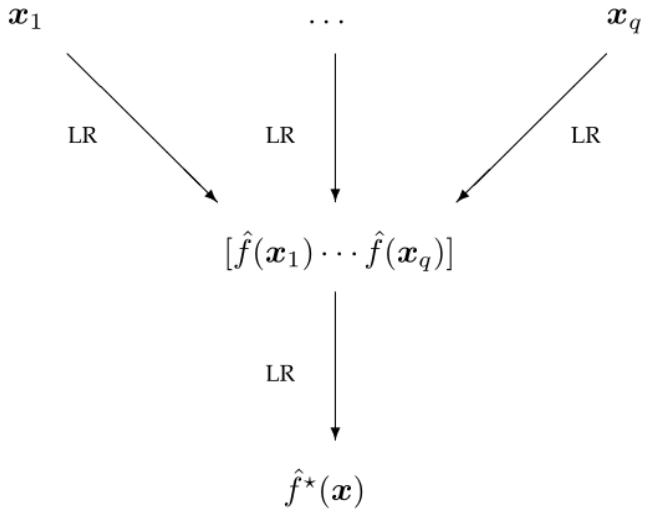
2.3. Dự báo dữ liệu không - thời gian dựa trên two-step LR

Tương tự two-step LDA (Nguyen Hoang Huy & cs., 2014), khi áp dụng two-step LR cho dữ liệu không - thời gian, chúng tôi sử dụng dữ liệu từ tất cả các địa điểm tại từng thời điểm để dự báo tại một thời điểm cụ thể ở bước đầu tiên. Sau đó tất cả kết quả dự đoán ở bước đầu tiên được kết hợp để tạo ra kết quả dự báo cuối cùng tại một thời điểm xác định. Trong phần này chúng tôi đưa ra quy trình áp dụng two-step LR để dự báo dữ liệu tại các địa điểm khác nhau, tại h thời điểm tiếp theo sử dụng dữ liệu ở l thời điểm trước đó. Chúng tôi thực hiện điều đó bằng h bước sau:

Bước 1: Dự báo dữ liệu ở mỗi địa điểm, tại thời điểm t + 1 bằng cách áp dụng two-step LR cho khối dữ liệu tại l thời điểm, từ thời điểm t – l + 1 đến t, ở tất cả các địa điểm.

Bước 2: Dự báo dữ liệu ở mỗi địa điểm, tại thời điểm t + 2 bằng cách áp dụng two-step LR cho khối dữ liệu bao gồm dữ liệu thực tại l – 1 thời điểm từ t – l + 2 đến t, ở tất cả các địa điểm, gộp với dữ liệu được dự báo tại thời điểm t + 1, đây là kết quả từ bước 1.

Bước 3: Dự báo dữ liệu ở mỗi địa điểm, tại thời điểm t + 3 bằng cách áp dụng two-step LR cho khối dữ liệu bao gồm dữ liệu thực tại l – 2 thời điểm từ thời điểm t – l + 3 đến t, ở tất cả các địa điểm, gộp với dữ liệu dự báo tại 2 thời điểm từ t + 1 đến t + 2, đây là kết quả từ bước 1, 2.



Hình 2. Lược đồ của two-step LR

Cứ tiếp tục lặp lại như vậy cho đến bước h

Bước h: Dự đoán dữ liệu ở mỗi địa điểm, tại thời điểm t + h bằng cách áp dụng two-step LR cho khối dữ liệu bao gồm dữ thực tại l – h + 1 thời điểm từ thời điểm t – l + h đến t, ở tất cả các trạm, gộp với dữ liệu dự báo tại h – 1 thời điểm từ t + 1 đến t + h - 1, đây là kết quả từ các bước 1, 2,..., h – 1. Quá trình dự báo trong h bước cho h thời điểm sau thời điểm t, sử dụng dữ liệu tại l thời điểm trước đó được mô tả như sau:

$$\left(sd_{t-l+i},...,sd_{t},sd_{t+1},...,sd_{t+i-1}\right)\xrightarrow{\text{Two-Step LD}}sd_{t+i}$$

i = 1, 2,..., h

trong đó h, l là những tham số cho trước, $sd_{t,}$ sd_{t} lần lượt là khối dữ liệu thực và dự đoán tại tất cả các địa điểm (spatial data) vào thời điểm t.

3. KẾT QUẢ VÀ THẢO LUẬN

3.1. Bài toán dự báo tốc độ gió

Năng lượng gió đã được phát triển nhanh chóng và ngày càng trở thành năng lượng tái tạo quan trọng ở nhiều vùng trên thế giới, đặc biệt ở những nước châu Âu (Lei & cs., 2009). Tích hợp năng lượng gió vào lưới điện trên diện rộng là thiết yếu và nhiều thách thức do bản chất ngẫu nhiên của gió. Sự tích hợp sẽ thuận tiện hơn nếu dự báo chính xác được năng lượng gió trong ngắn hạn (Ghaderi, 2017). Có nhiều hướng tiếp cận để dự đoán năng lượng gió, tuy nhiên hướng tiếp cận dựa vào dự đoán tốc độ gió vẫn được xem là hướng tiếp cận nổi bật nhất (Tascikaraoglu & cs., 2016).

Nhiều phương pháp dự báo tốc độ gió được đưa ra, có thể chia làm 2 loại: phương pháp vật lý, và phương pháp thống kê (Lei & cs., 2009). Phương pháp vật lý khai thác nhiều thuộc tính vật lý như địa hình, áp suất, nhiệt độ, có lợi thế trong dự báo tốc độ gió dài hạn. Phương pháp thống kê thường dựa vào giá trị lịch sử, như mô hình ARMA, và thường cho kết quả tốt trong dự báo tốc độ gió ngắn hạn. Bài báo này tập trung vào dự báo tốc độ gió ngắn hạn dựa vào dữ liệu tốc độ gió lịch sử.

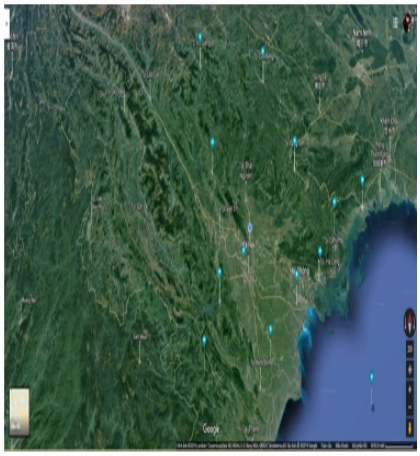
Mô hình ARMA và một số trường hợp đặc biệt như mô hình AR, mô hình Persistence rất thông dụng trong dự báo tốc độ gió (Lei & cs., 2009). Chúng được xem là các mô hình chuỗi thời gian đơn giản nhất nhưng có thể vượt trội nhiều mô hình phức tạp khác trong dự báo tốc độ gió ngắn hạn (Sanandaji & cs., 2015; Tascikaraoglu & cs., 2016). Để cải tiến độ chính xác dự đoán, nhiều mô hình tương quan không gian được đưa ra để khai thác mối quan hệ tốc độ gió ở những vị trí khác nhau. Tuy nhiên chúng chỉ gần như áp dụng các phương pháp học máy như mạng nơron nhân tạo (ANN-based ST), phương pháp bình phương tối tiểu (LS-based ST) đối với tất cả dữ liệu không - thời gian hoặc với dữ liệu đã được biến đổi thông qua biến đổi Wavelet (WT-ANN),... (Lei & cs., 2009; Sanandaji & cs., 2015; Tascikaraoglu & Uzunoglu, 2014; Tascikaraoglu & cs., 2016).

Gần đây các thuật toán học sâu như Deep Learning-based Spatio-Temporal Forecasting (DL-STF) được sử dụng để dự báo tốc độ gió (Ghaderi & cs., 2017; Yu & cs., 2019; Wu & cs., 2019). Tương tự những phương pháp trên, nó khai thác toàn bộ dữ liệu không - thời gian như dữ liệu đầu vào cho thuật toán dự đoán, sử dụng Recurrent Neural Networks (RNN) và Long Short Term Memory (LSTM) (Ghaderi & cs., 2017). Phương pháp này vượt trội các kết quả dự báo tốc độ gió gần đây. Tuy nhiên, tất cả các phương pháp trên đều không dựa vào cấu trúc không - thời gian bên trong của dữ liệu tốc độ gió, như tính khả tách của ma trận hiệp phương sai.

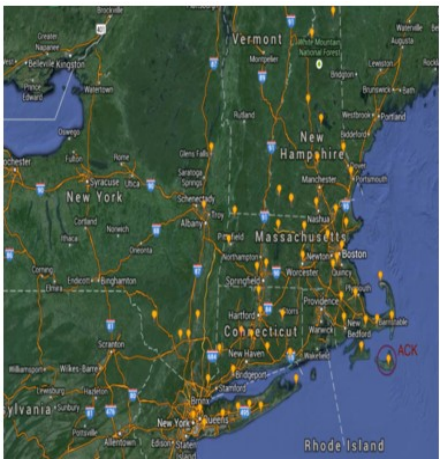
3.2. Mô tả dữ liệu

3.2.1. Tập dữ liệu NCHMF

Tập dữ liệu gió NCHMF từ trung tâm dự báo khí tượng thủy văn quốc gia Việt Nam được đo ở 13 trạm thời tiết Hà Giang, Cao Bằng, Tuyên Quang, Hòa Bình, Nam Định, Hà Đông, Phú Liễn, Lạng Sơn, Bãi Cháy, Tiên Yên, Móng Cái, Bạch Long Vĩ, Hội Xuân. Những trạm này ở miền bắc Việt Nam với kinh độ từ 104.044220 đến 107.848208, vĩ độ từ 20.020846 đến 22.401052, như trong hình 3. Tốc độ gió ở trạm Bạch Long Vĩ thay đổi nhanh và không ổn định như các trạm khác khác. Dữ liệu quan sát từ ngày 01/10/2016 đến 01/01/2019. Tốc độ gió được đo ba giờ một lần.



Hình 3. Vị trí trạm khí tượng đo tốc độ gió của Trung tâm Dự báo khí tượng thủy văn quốc gia Việt Nam



Hình 4. Vị trí trạm đo dữ liệu tốc độ gió METAR

3.2.2. Tập dữ liệu METAR

Tập dữ liệu tốc độ gió hàng giờ METAR được thu thập từ các báo cáo thời tiết tại 57 cảng sân bay ở bờ biển phía đông Hoa Kỳ, bao gồm Massachusetts, Connecticut, New York, New Hampshire. Hình 4 cho thấy vị trí của các cảng sân bay này. Dấu đỏ là sân bay ACK, nằm trên một hòn đảo. Tốc độ gió ở đảo đó thay đổi rất nhanh, tương tự như những gì ở trạm thời tiết Bạch Long Vĩ của Việt Nam, một trong 13 trạm thời tiết ở hình 3. Tốc độ gió từ 06/01/2014 đến 20/02/2014 được sử dụng để kiểm tra hiệu năng của các phương pháp học máy được nghiên cứu. Đây là thời điểm, tốc độ gió không ổn định hơn tất cả các khoảng thời gian khác.

3.3. Kết quả ứng dụng two-step LR

Tương tự two-step LDA (Nguyen Hoang Huy & cs., 2014), two-step LR xác định các nhóm con đặc trưng gồm tất cả các đặc trưng tại mỗi thời điểm. Chúng tôi cũng không áp dụng bất kỳ kỹ thuật học máy nào như chỉnh hóa để nâng cao hiệu năng của hồi quy tuyến tính được thực hiện ở mỗi bước. Do đó không có sự thiết lập siêu tham số nào khác của two-step LR. Hơn nữa độ phức tạp tính toán của two-step LR giảm đi do chỉ áp dụng hồi quy tuyến tính trên mỗi nhóm con đặc trưng.

Bảng 1 so sánh hiệu suất của phương pháp được chúng tôi đưa ra với các phương pháp khác trên tập dữ liệu METAR. Để so sánh sai số của

các phương pháp, chúng tôi sử dụng ba độ đo thông dụng là MAE, RMSE và NRMSE. Trong thử nghiệm này chúng tôi chọn l = 12, h = 6 theo Ghaderi (2017), đây là tham số cho hiệu năng tốt nhất của DL-STF trên tập dữ liệu METAR. Lựa chọn l = 12, h = 6 có nghĩa là two-step LR và các phương pháp khác sử dụng d = 684 = 57 × 12 giá trị quan trắc (đặc trưng) từ 57 trạm và 12 thời điểm (giờ) trước đó để dự đoán giá trị tốc độ gió trong 6 giờ tiếp theo. Chúng tôi sử dụng dữ liệu tốc độ gió từ 6.012 giờ liên tiếp (250,5 ngày), hình thành 6.000 mẫu huấn luyện để học mô hình dự báo và các mẫu kiểm tra là tốc độ gió trong giai đoạn không ổn định nhất từ 06/01/2014 đến 20/02/2014 như đã nêu trong bài báo của Ghaderi & cs. (2017). Cụ thể hơn về các phương pháp khác được trình bày cụ thể trong các bài báo của Sanandaji & cs. (2015) và Tascikaraoglu & cs. (2016).

Bảng 2 trình bày 3 sai số trung bình của tất cả các trạm trên tập dữ liệu METAR. Chúng ta có thể thấy hiệu năng dự đoán trên ACK hoặc tất cả các trạm của two-step LR trội hơn DL-STF, phương pháp tốt nhất hiện nay.

Hình 5 biểu diễn dữ liệu tốc độ gió thực tế (đường màu xanh) và tốc độ gió dự đoán (đường màu đỏ) từ dữ liệu kiểm tra trên 16 trạm quan sát. Đồ thị đầu tiên trong hình ứng với trạm quan sát ở cảng sân bay ACK.

Trong tập dữ liệu NCHMF, có 3 giá trị tốc độ gió bị thiếu và chúng tôi đã thay thế chúng bằng giá trị tốc độ gió đo tại thời điểm trước đó (3 tiếng trước), ở cùng trạm. Bảng 3 biểu diễn hiệu năng của DL-STF và two-step LR khi sử dụng tất cả dữ liệu từ 13 trạm thời tiết với l = 12, h = 6. Bằng cách này thì DL-STF và two-step LR có thể khai thác tất cả các thông tin tương tác ẩn giữa các trạm. Qua bảng 3, chúng ta có thể thấy two-step LR có hiệu năng dự báo tốc độ gió tốt hơn hoặc bằng phương pháp đang cho kết quả tốt nhất hiện nay là DL-STF, xem bài báo Ghaderi & cs. (2017).

Bảng 1. Sai số của các phương pháp khác nhau trên trạm ACK

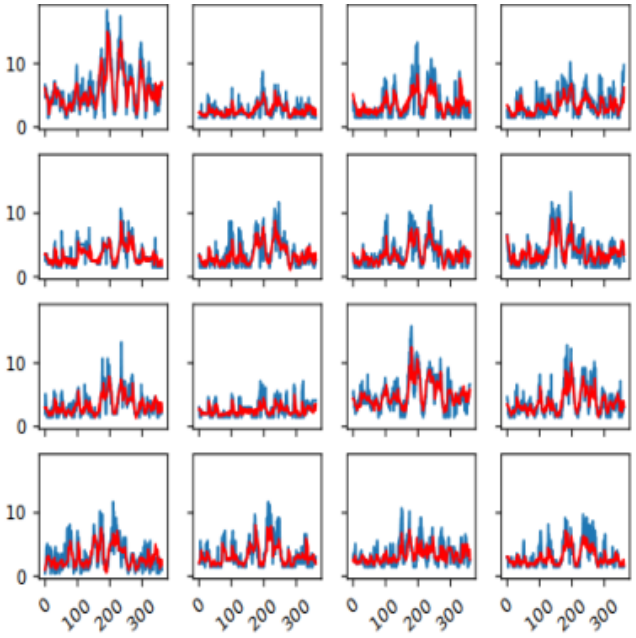
Method	MAE (m/s)	RMSE (m/s)	NRMSE (%)
Persistence Forecasting	2,14	2,83	16,86
AR of order 1	2,07	2,76	16,44
AR of order 3	2,07	2,76	16,40
WT-ANN	1,82	2,47	14,68
ANN-based ST	1,80	2,30	13,69
LS-based ST	1,72	2,20	13,08
DL-STF	1,63	2,19	13,08
Two-Step LR	1,40	1,93	11,48

Bảng 2. Sai số trung bình trên tất cả các trạm sử dụng DL-STF, two-step LR

Method	MAE (m/s)	RMSE (m/s)	NRMSE (%)
DL-STF	1,18	1,62	16,28
Two-Step LR	1,09	1,44	14,32

Bảng 3. Sai số trung bình của DL-STF và Two-Step LR trên trạm Bạch Long Vĩ và cả 13 trạm

Method	Locations	MAE (m/s)	RMSE (m/s)	NRMSE (%)
DL-STF	Bach Long Vi	1,70	2,36	13,86
	All Stations	0,82	1,16	19,09
Two-Step LR	Bach Long Vi	1,67	2,27	13,34
	All Stations	0,82	1,07	18,04



Hình 5. So sánh giữa tốc độ gió thực tế và dự báo trên dữ liệu kiểm tra

4. KẾT LUẬN

Hiệu suất của hồi quy tuyến tính bị ảnh hưởng bởi số chiều. Để giải quyết vấn đề này, chúng tôi giới thiệu phương pháp áp dụng hồi quy tuyến tính trong hai bước, được gọi là two-step LR. Hướng tiếp cận này được gợi ý từ two-step LDA và tính khả tách của ma trận hiệp phương sai của dữ liệu tốc độ gió. Với dữ liệu tốc độ gió có số chiều cao trung bình, hiệu năng của cách tiếp cận này tốt hơn các phương pháp mới nhất. Ngày nay, có nhiều phương pháp điều chỉnh hồi quy tuyến tính cho dữ liệu có số chiều cao như là hồi quy Lasso và các cải tiến của nó. Tuy nhiên, với hiểu biết của tôi, các thuật toán đó chưa được thử nghiệm cho dự báo tốc độ gió. Trong tương lai, hướng tiếp cận hai bước sử dụng những thuật toán này nên được khảo sát tỉ mỉ.

TÀI LIỆU THAM KHẢO

Bali V., Kumar A. & Gangwar S. (2019). Deep Learning based Wind Speed Forecasting-A Review. 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence). India. pp. 426-431.

Bai Z., Li H. & Pan G. (2019). Central limit theorem for linear spectral statistics of large dimensional separable sample covariance matrices. Bernoulli. 25(3): 1838-1869.

Bickel P.J. & Levina E. (2004). Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. Bernoulli. 10(6): 989-1010.

Bickel P.J. & Levina E. (2008). Covariance regularization by thresholding. The Annals of Statistics. 36: 2577-2604.

Cai T. & Liu W. (2011). A direct estimation approach to sparse linear discriminant analysis. Journal of the American Statistical Association. 106(496): 1566-1577.

Cai T. & Zhang L. (2019). High dimensional linear discriminant analysis: optimality, adaptive algorithm and missing data. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 81(4): 675-705.

Genton M.G. (2007). Separable approximation of space-time covariance matrices. Environmetrics. 18: 681-695.

Ghaderi A., Sanandaji B. M. & Ghaderi F. (2017). Deep forecast: Deep learning-based spatio-temporal forecasting. 34th ICML Time Series Workshop. Sydney, Australia.

Hastie T., Tibshirani R. & Friedman J. (2009). The Elements of Statistical Learning: DataMining, Inference, and Prediction. Springer Series in Statistics, Springer-Verlag.

Hastie T., Tibshirani R. & Wainwright M. (2015). Statistical Learning with Sparsity: The Lasso and Generalizations. Chapman & Hall/CRC Press.

Huizenga H.M., De Munck J.C., Waldorp L.J. & Grasman R.P.P.P. (2002). Spatiotemporal EEG/MEG source analysis based on a parametric noise covariance model. IEEE Transactions on Biomedical Engineering. 49: 533-539.

Huy N.H., Frenzel S. & Bandt C. (2014). Two-step linear discriminant analysis for classification of eeg data. In M. Spiliopoulou, L. Schmidt-Thieme and R. Janning, editors, Data Analysis, Machine Learning and Knowledge Discovery. Springer, Cham. pp. 41-50.

Lei M., Shiyan L., Chuanwen J., Hongling L. & Yan Z. (2009). A review on the forecasting of wind speed and generated power. Renewable and Sustainable Energy Reviews. 13: 915-920.

Lei L., Bickel P.J., Karoui N.E. (2018). Asymptotics for high dimensional regression M-estimates: fixed design results. Probability Theory and Related Fields. 172 (3-4): 983-1079.

Leiva R. & Roy A. (2014). Classification of Higher-order Data with Separable Covariance and Structured Multiplicative or Additive Mean Models. Communications in Statistics - Theory and Methods. 43(5): 989-1012.

Sanandaji B.M., Tascikaraoglu A., Poolla K. & Varaiya P. (2015). Low dimensional models in spatio-temporal wind speed forecasting. American Control Conference. Chicago, USA. pp. 4485-4490.

Tascikaraoglu A. & Uzunoglu M. (2014). A review of combined approaches for prediction of short-term wind speed and power. Renewable and Sustainable Energy Reviews. 34: 243-254.

Tascikaraoglu A., Sanandaji B. M., Poolla K. & Varaiya P. (2016). Exploiting sparsity of inter-connections in spatio-temporal wind speed forecasting using wavelet transform. Applied Energy. 165 (1): 735-747.

Yu R., Gao J., Yu M., Lu W., Xu T., Zhao M., Zhang J., Zhang R. & Zhang Z. (2019). LSTM-EFG for wind power forecasting based on sequential correlation features. Future Generation Computer Systems. 93: 33-42.

Wu Y.X., Wu Q. B. & Zhu J.Q. (2019). Data-driven wind speed forecasting using deep feature extraction and LSTM. IET Renewable Power Generation. 13(12): 2062-2069.