XỬ LÝ DỮ LIỆU THIẾU BẰNG BIỂU ĐỒ CHUẨN HÓA ĐƠN VỊ (SLP) VÀ SUPPORT VECTOR REGRESSION (SVR)

HANDLING MISSING DATA USING STANDARDIZED LOAD PROFILE (SLP) AND SUPPORT VECTOR REGRESSION (SVR)

Nguyễn Tuấn Dũng[1,*], Nguyễn Thanh Phương[2]

TÓM TẮT

Trong những năm gần đây, việc nghiên cứu và ứng dụng các kỹ thuật khai thác dữ liệu gặp phải nhiều khó khăn, thách thức lớn, trong đó có vấn đề thiếu những giá trị thuộc tính của dữ liệu. Có nhiều nguyên nhân khác nhau dẫn tới vấn đề này: thiết bị thu thập bị hỏng, có sự từ chối cung cấp dữ liệu nhằm bảo vệ tính riêng tư, có sai sót khi nhập dữ liệu hoặc có các sự cố xảy ra trong quá trình truyền dữ liệu,... Trong đó, việc thiếu dữ liệu phục vụ công tác nghiên cứu, dự báo phụ tải điện là một trong những vấn đề nan giải đối với ngành điện. Hiện các Công ty điện lực đang thực hiện việc này bằng cách nội suy từ các giá trị đo đếm của các ngày trước, giờ trước một cách thủ công, không chuẩn xác làm ảnh hưởng không nhỏ đến kết quả phân tích, xử lý dữ liệu trong quá trình nghiên cứu, dự báo phụ tải. Bài báo đề xuất một phương pháp xử lý dữ liệu thiếu bằng cách xây dựng Biểu đồ chuẩn hóa đơn vị (SLP) trên cơ sở bộ dữ liệu phụ tải điện quá khứ (chu kỳ 60 phút), kết hợp các giải thuật học máy SVR (NN/RD) để xây dựng lại đường đặc tuyến phụ tải từ đó ước lượng các dữ liệu đã mất hoặc không ghi nhận được trong quá trình đo đếm.

Từ khóa: *Thiếu dữ liệu; ước lượng; số liệu đo đếm; phụ tải điện; Biểu đồ chuẩn hóa đơn vị; SVR.*

ABSTRACT

In recent years, the research and application of data mining techniques encountered many difficulties and major challenges, including the lack of attribute values of data. There are many different reasons for this problem: the device is broken, the data is refused to protect the privacy, data entry mistakes or incidents occur during data transmission. In particular, the lack of data for electricity load research and forecasting is one of the problems for the electricity industry. Currently, the power companies are doing this by interpolating from the measured values of previous days and hours manually, which significantly affects the results of data analysis during the load forecasting process. The paper proposes a method of processing missing data by building a Standardized Chart (SLP) based on past load data (60-minute cycle), combining machine learning algorithms SVR (NN / RD) to rebuild the load curve, thereby we can estimate the data missed or not recorded during the measurement.

Keywords*: Missing data; estimation; measured data; electrical load; Standardized load profile; SVR.*

[1]Tổng Công ty Điện lực TP.HCM
[2]Trường Đại học Công nghệ TP.HCM
[*]Email: dungnt@hcmpc.com.vn
Ngày nhận bài: 20/10/2018
Ngày nhận bài sửa sau phản biện: 20/01/2019
Ngày chấp nhận đăng: 25/02/2019

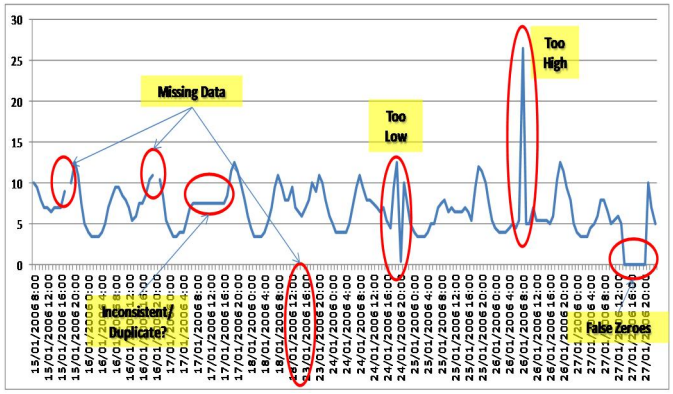
1. ĐẶT VẤN ĐỀ

1.1. Dữ liệu thiếu trong quá trình khai thác cơ sở dữ liệu (CSDL)

Cho đến nay, có nhiều phương pháp xử lý giá trị thiếu đã được đề xuất và áp dụng [1, 2]. Các phương pháp này cho phép xử lý trực tiếp các giá trị thiếu, tuy nhiên chúng cũng có thể mang những thông tin nhiễu vào tập dữ liệu đang xét. Việc xử lý các giá trị thiếu cần phải được cân nhắc và thực hiện một cách thận trọng, nếu các nhà nghiên cứu sử dụng phương pháp xử lý dữ liệu bị mất mà không cẩn trọng xem xét các giả định cần thiết của phương pháp đó thì họ có nguy cơ có kết quả sai lệch và gây hiểu nhầm [2]. Cho đến nay, việc xử lý giá trị thiếu trong các CSDL vẫn là đề tài thu hút sự quan tâm của nhiều nhà nghiên cứu và ứng dụng.

Một nhiệm vụ vô cùng quan trọng khi xây dựng một phương pháp xử lý giá trị thiếu là phải hiểu được cơ chế sinh ra các giá trị thiếu trong CSDL cần xử lý. Nắm bắt được cơ chế sinh ra giá trị thiếu trong một tình huống cụ thể sẽ giúp xây dựng được một phương pháp xử lý thích hợp và hiệu quả.

1.2. Dữ liệu thiếu trong nghiên cứu phụ tải điện



Hình 1. Các lỗi thường gặp trong ghi nhận dữ liệu

Trong quá trình vận hành, thu thập dữ liệu đã xuất hiện nhiều sự cố làm gián đoạn việc ghi nhận các dữ liệu đo đếm như: sự cố truyền dẫn tín hiệu từ công tơ đo đếm về Kho dữ liệu làm mất gói dữ liệu truyền về; lỗi tại thiết bị đo đếm; lỗi do mất nguồn điện; lỗi do cài đặt thiết bị đo đếm không đúng; lỗi do xử lý dữ liệu bằng phương pháp thủ

công; hoặc do việc thu thập dữ liệu bằng thủ công,... dẫn đến dữ liệu ghi nhận được không phù hợp như: dữ liệu có có giá trị bằng 0 (Fasse Zero); trùng lặp dữ liệu (Inconsistent/Duplicate); thiếu chuỗi dữ liệu (Missing Data); dữ liệu thiếu chính xác, quá cao hoặc thấp bất thường (Too High/Too Low).

2. CÁC PHƯƠNG PHÁP NGHIÊN CỨU

Cho đến nay vẫn chưa có một phương pháp nào được khuyên sử dụng riêng cho việc xử lý dữ liệu thiếu trong các ứng dụng khai thác dữ liệu. Đặc biệt, là làm thế nào để có thể xử lý giá trị thiếu trong một CSDL dữ liệu khổng lồ.

2.1. Một số phương pháp xử lý dữ liệu thiếu đã được nghiên cứu [3, 4, 5]

2.1.1. Phương pháp loại bỏ: Nếu xảy ra trường hợp thiếu dữ liệu cho một biến bất kỳ nào đó, giải pháp đơn giản là loại bỏ thuộc tính bị thiếu của dữ liệu ra khỏi qua quá trình phân tích đánh giá của chuỗi dữ liệu.

Phương pháp này có ưu điểm là đơn giản, ít tốn thời gian hơn bất kỳ phương pháp nào khác. Nhưng nó lại có hai điểm hạn chế quan trọng: *i)* thứ nhất là nếu chúng ta áp dụng vào trong thực tế có thể gây mất mát nhiều đặc tính của dữ liệu; *ii)* thứ hai là nếu phân bố dữ liệu thiếu trong tập dữ liệu không thuộc trường hợp (MCAR) thì việc loại bỏ tất cả các bộ dữ liệu có giá trị thiếu sẽ làm sai lệch nghiêm trọng kết quả.

2.1.2. Phương pháp gán ghép: Phương pháp này thay thế các giá trị bị thiếu bằng một giá trị dự đoán được xem là hợp lý và sau đó thực hiện các phân tích cho chuỗi dữ liệu đã được bổ sung. Gán ghép trung bình: Tính giá trị trung bình dữ liệu của X bằng cách sử dụng các giá trị không bị mất và sử dụng nó để gán ghép cho giá trị thiếu.

2.1.3. Phương pháp hồi quy tuyến tính

Khi hai thuộc tính định lượng nào đó có mối quan hệ tuyến tính với nhau, chúng ta có thể xây dựng một phương trình hồi quy tuyến tính, trong đó thuộc tính có giá trị thiếu là biến phụ thuộc, biến còn lại là biến độc lập và sử dụng phương trình hồi quy cho việc dự đoán các giá trị thiếu của biến phụ thuộc thông qua các giá trị đã biết của biến độc lập.

Phương pháp hồi quy tuyến tính thường gặp phải hai vấn đề: *i)* thứ nhất, mô hình quan hệ giữa các thuộc tính có phải tuyến tính không. Nếu mối quan hệ này là không tuyến tính, các giá trị thiếu ước lượng được có thể bị sai lệch lớn so với các giá trị thực; *ii)* thứ hai, thường thì trong cùng một bộ dữ liệu, các thuộc tính có quan hệ chặt với thuộc tính có giá trị thiếu cũng có giá trị thiếu.

2.2. Phương pháp xử lý dữ liệu thiếu trong nghiên cứu phụ tải điện

Một số phương pháp ước lượng số liệu đo đếm của các phụ tải điện bị lỗi trong quá trình thu thập dữ liệu của các Công ty điện lực thường được sử dụng như [12]:

- Nội suy tuyến tính: nội suy từ đường đặc tính xu thế tiêu thụ điện;

- Ngày tương đồng: sử dụng dữ liệu ngày tương đồng của tuần hiện tại hoặc tuần trước;

- Tự động ước lượng: sử dụng trong trường hợp dữ liệu bị thiếu không quá bảy (07) ngày;

- Kiểm tra trực quan đồ thị: để biết được dữ liệu bị sai và quyết định về dữ liệu được ước lượng;

- Hiệu chỉnh ước lượng số liệu thủ công: được sử dụng khi dữ liệu bị thiếu nhiều hơn bảy (07) ngày;

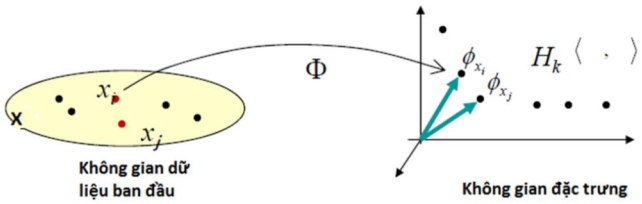
- Hiệu chỉnh ước lượng giá trị trung bình các tuần của ngày tham chiếu: dựa vào dữ liệu của bốn (04) tuần gần nhất.

Tuy nhiên, các cách làm này được thực hiện một cách thủ công và phụ thuộc rất nhiều vào năng lực kinh nghiệm của chuyên gia thực hiện việc ước lượng.

2.3. Bộ hồi quy dựa theo vector hỗ trợ - Support vector regression (SVR)

Ý tưởng cơ bản của SVR là ánh xạ không gian đầu vào sang một không gian đặc trưng nhiều chiều mà ở đó, ta có thể áp dụng được hồi qui tuyến tính (mà nếu ta áp dụng trực tiếp hồi qui tuyến tính thì không hiệu quả).

Đặc điểm của SVR là cho ta một giải pháp thưa (sparse solution); nghĩa là để xây dựng được hàm hồi qui, ta không cần phải sử dụng hết tất cả các điểm dữ liệu trong bộ huấn luyện. Những điểm có đóng góp vào việc xây dựng hàm hồi qui được gọi là những Support Vector. Việc phân lớp cho một điểm dữ liệu mới sẽ chỉ phụ thuộc vào các support vector.



Hình 2. Biến đổi không gian dữ liệu sang không gian đặc trưng (thủ thuật Kernel)

Hàm hồi qui cần tìm có dạng:

$y = f(x) = w^T\Phi(x) + b$

Trong đó: $w \in R^m$ là vector trọng số; T là kí hiệu chuyển vị; $b \in R$ là hằng số; $x \in R^n$ là vector đầu vào; $\Phi(x) \in R^m$ là vector đặc trưng; Φ làm hàm ánh xạ từ không gian đầu vào sang không gian đặc trưng [6, 7, 8].

Như vậy, mục tiêu của việc huấn luyện SVR là tìm ra được w và b.

Cho tập huấn luyện $\{(x1, t1), (x2, t2), \ldots, (xN, tN)\} \subset R^n \times R$. Với bài toán hồi qui đơn giản, để tìm w và b ta phải tối thiểu hóa hàm lỗi chuẩn hóa:

$$\frac{1}{2}\sum_{n=1}^{N}\{y_n - t_n\}^2 + \frac{\lambda}{2}\|w\|^2$$ với λ là hằng số chuẩn hóa

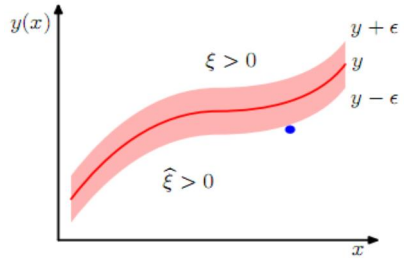
Để có được một giải pháp thưa, ta sẽ thay hàm lỗi trên bằng hàm lỗi *ε-insensitive*. Đặc điểm của hàm lỗi này là nếu trị tuyệt đối của sự sai khác giữa giá trị dự đoán y(x) và giá trị đích nhỏ hơn ε (với $\varepsilon > 0$) thì nó coi như độ lỗi bằng 0.

Như vậy bây giờ, ta phải tối thiểu hóa hàm lỗi chuẩn hóa sau:

$$C\sum_{n=1}^{N} E_{\varepsilon}(y(x_n)-t_n)^2+\frac{1}{2}\|w\|^2$$

Với $y_n = w^T\Phi(x_n)+b$, C là hằng số chuẩn hóa giống như λ nhưng được nhân với hàm lỗi thay vì $\|w\|^2$.

Để cho phép một số điểm nằm ngoài ống ε, ta sẽ đưa thêm các biến lỏng (slack variable) vào. Đối với mỗi điểm dữ liệu x_n, ta cần hai biến lỏng $\xi_n \geq 0$ và $\hat{\xi}_n \geq 0$, trong đó $\xi_n > 0$ ứng với điểm mà $t_n > y(x_n) + \varepsilon$ (nằm ngoài và phía trên ống) và $\hat{\xi}_n \geq 0$ ứng với điểm mà $t_n < y(x_n) - \varepsilon$ (nằm ngoài và phía dưới ống).



Hình 3. Minh họa cho các biến lỏng ξ_n

Điều kiện để một điểm đích nằm trong ống là: $y_n - \varepsilon \leq t_n \leq y_n + \varepsilon$ với $y_n = y(x_n)$. Với việc sử dụng các biến lỏng, ta cho phép các các điểm đích nằm ngoài ống (ứng với các biến lỏng > 0) và như thế thì điều kiện bây giờ sẽ là:

$$t_n \leq y_n + \varepsilon + \xi_n$$

$$t_n \geq y_n - \varepsilon - \hat{\xi}_n$$

Như vậy, ta có hàm lỗi cho SVR:

$$C\sum_{n=1}^{N}(\xi_n+\hat{\xi}_n+\frac{1}{2}\|w\|)^2$$

Mục tiêu của ta là tối thiểu hóa hàm lỗi này với các ràng buộc:

$$\xi_n \geq 0; \hat{\xi}_n \geq 0$$

$$t_n \leq y_n + \varepsilon + \xi_n$$

$$t_n \geq y_n - \varepsilon - \hat{\xi}_n$$

Dùng hàm Lagrange và điều kiện Karush-Kuhn-Tucker, ta có bài toán tối ưu hóa tương đương:

$$-\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}(a_n-\hat{a}_n)(a_m-\hat{a}_m)k(x_n,x_m)$$

$$-\varepsilon\sum_{n=1}^{N}(a_n-\hat{a}_n)+\sum_{n=1}^{N}(a_n-\hat{a}_n)t_n$$

Với k là hàm nhân: $k(x, x') = \Phi(x)^T\Phi(x')$. Bất kỳ một hàm nào thỏa điều kiện Mercer thì đều có thể được dùng làm hàm nhân. Hàm nhân được sử dụng phổ biến nhất là hàm Gaussian: $k(x_i,x_j) = \exp(-\gamma\|x_j - x_j\|^2)$

Cực đại hóa với các ràng buộc:

$$0 \leq a_n \leq C$$

$$0 \leq \hat{a}_n \leq C$$

$$\sum_{n=1}^{N}(a_n-\hat{a}_n)=0$$

Từ đây, ta có hàm hồi qui của SVR:

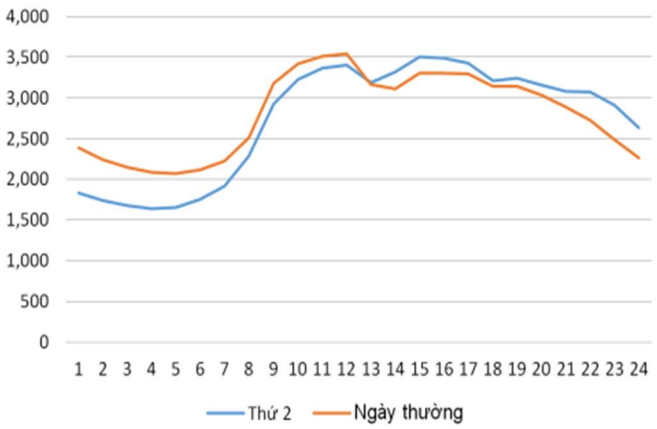
$$y(x)=\sum_{n=1}^{N}(a_n-\hat{a}_n)k(x_n,x_m)+b$$

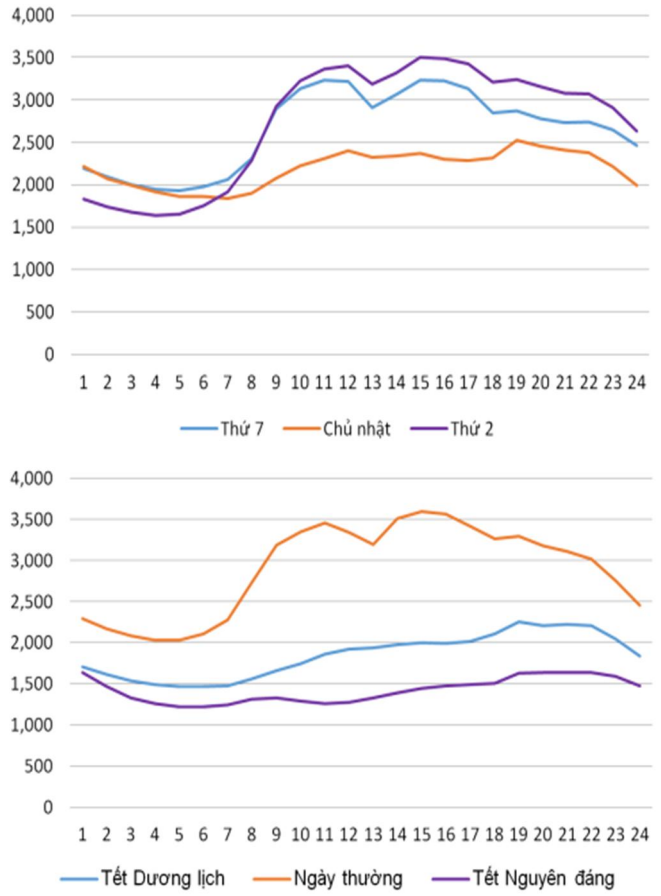
Như vậy, với SVR sử dụng hàm lỗi ε-insensitive và hàm nhân Gaussian ta có ba tham số cần tìm: hệ số chuẩn hóa C, tham số γ của hàm nhân Gaussian và độ rộng của ống ε [9]. Cả ba tham số này đều ảnh hưởng đến độ chính xác dự đoán của mô hình và cần phải chọn lựa kỹ càng. Nếu C quá lớn thì sẽ ưu tiên vào phần độ lỗi huấn luyện, dẫn đến mô hình phức tạp, dễ bị quá khớp. Còn nếu C quá nhỏ thì lại ưu tiên vào phần độ phức tạp mô hình, dẫn đến mô hình quá đơn giản, giảm độ chính xác dự đoán. Ý nghĩa của ε cũng tương tự C. Nếu ε quá lớn thì có ít vectơ hỗ trợ, làm cho mô hình quá đơn giản. Ngược lại, nếu ε quá nhỏ thì có nhiều vectơ hỗ trợ, dẫn đến mô hình phức tạp, dễ bị quá khớp. Tham số γ phản ánh mối tương quan giữa các vectơ hỗ trợ nên cũng ảnh hưởng đến độ chính xác dự đoán của mô hình.

2.4. Biểu đồ chuẩn hóa đơn vị (SLP)

Quan sát đồ thị phụ tải các ngày trong một tuần và một số ngày lễ đặc biệt trong năm của khu vực thành phố Hồ Chí Minh (hình 4) ta thấy: sự biến đổi giữa các ngày thường (từ thứ 3 đến thứ 6) không có nhiều biến động và có cùng một kiểu biểu đồ phụ tải. Đối với đồ thị phụ tải ngày thứ 2 thì có sự biến đổi khác biệt với ngày thường tại khoảng thời từ 0h00 đến 9h00, do có sự chuyển tiếp nhu cầu từ ngày chủ nhật.

Đối với đồ thị phụ tải ngày thứ 7 thì có sự biến đổi nhưng không nhiều so với ngày thường, chủ yếu nhu cầu phụ tải suy giảm vào buổi chiều tối, do bắt đầu cho ngày nghỉ cuối tuần. Riêng đối với đồ thị phụ tải ngày Chủ nhật thì hoàn toàn khác với các ngày thường (nhu cầu sử dụng điện xuống thấp).

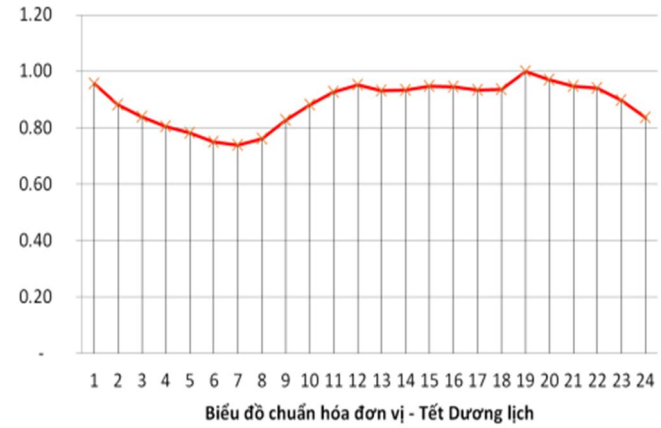


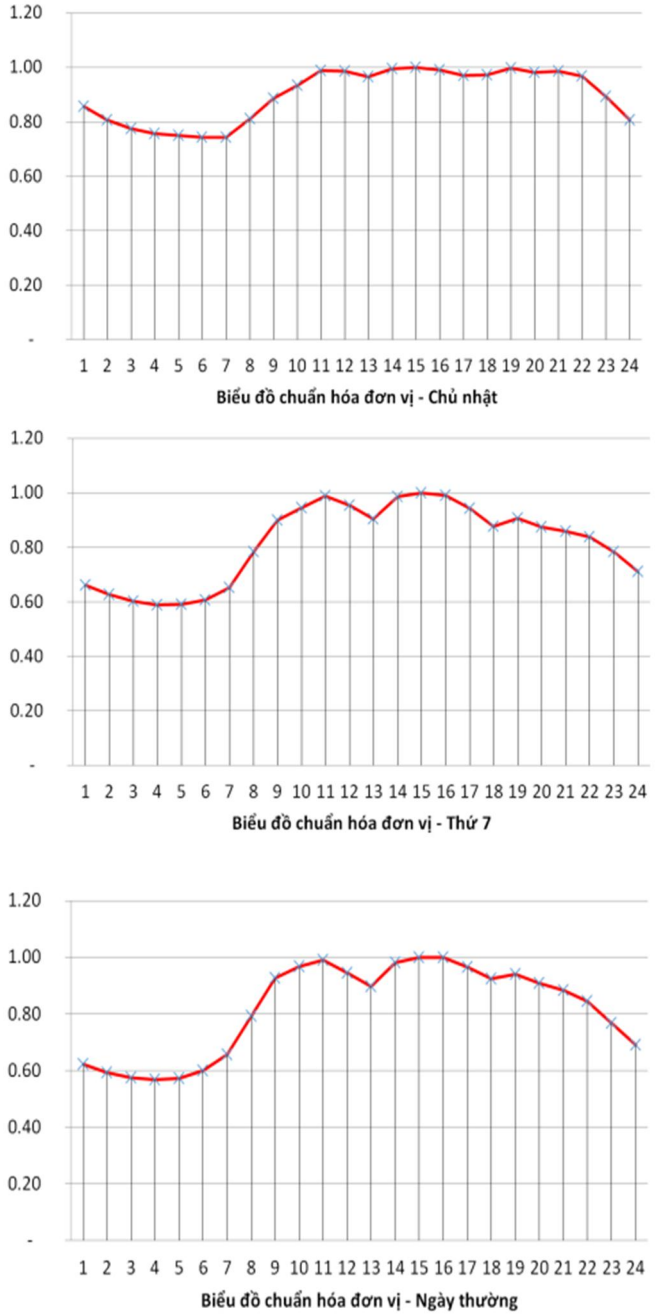


Hình 4. Đồ phụ tải một số ngày trong năm

Khi quan sát biểu đồ phụ tải các ngày Tết Dương lịch và Tết Âm lịch thì chúng ta thấy sự khác biệt hoàn toàn, đồ thị gần như bằng phẳng và nhu cầu phụ tải xuống khá thấp do đây là các ngày nghỉ. Riêng ngày Tết Âm lịch thì nhu cầu phụ tải xuống thấp nhất, do đây là kỳ nghỉ kéo dài nhất trong năm (có thể từ 6 - 9 ngày).

Biểu đồ phụ tải chuẩn hóa đơn vị (Standardized Load Profiles - SLP) được xây dựng bằng cách lấy giá trị công suất thu thập theo chu kỳ 60 phút chia cho công suất cực đại của nó. Cần phải xây dựng SLP cho 365 ngày/ năm. Một số SLP điển hình:





Hình 5. SLP một số ngày trong năm

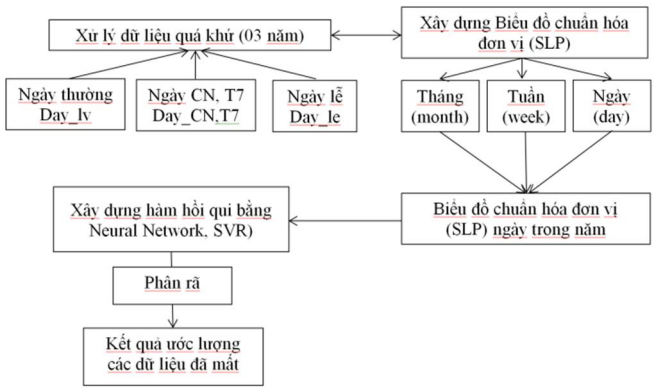
Qua quan sát, biểu đồ phụ tải chuẩn hóa đơn vị thể hiện được hết tất cả các đặc tính tải theo từng thời điểm, mùa vụ và các ngày lễ, Tết (Dương lịch, Nguyên Đán),… chúng ta thấy mức độ tương đồng của SLP về mặt hình dáng, độ lớn từng chu kỳ. Do đó, Biểu đồ phụ tải chuẩn hóa đơn vị (SLP) chính là một điểm đặc biệt và cũng là bộ thông số đầu vào quan trọng của quá trình huấn luyện của các thuật toán học máy SVR (NN) để xây dựng lại đường đặc tuyến phụ tải từ đó ước lượng các dữ liệu đã mất hoặc không ghi nhận được trong quá trình đo đếm.

- **Lưu đồ giải thuật**:

Bài báo đề xuất một phương pháp xử lý dữ liệu thiếu bằng cách xây dựng Biểu đồ chuẩn hóa đơn vị (SLP) trên cơ

sở bộ dữ liệu phụ tải điện quá khứ chu kỳ 60 phút/lần của 03 năm trước đó. Đồng thời, kết hợp các giải thuật SVR (NN) để xây dựng lại hàm hồi qui (đường đặc tuyến phụ tải) từ đó ước lượng các dữ liệu đã mất hoặc không ghi nhận được trong quá trình đo đếm.

Trên cơ sở SLP của từng chu kỳ của bộ dữ liệu trong quá khứ, chúng ta có thể xây dựng bộ dữ liệu SLP cho các chu kỳ cần dự báo trong tương lai và cần chuẩn xác đến từng chu kỳ, từng loại ngày (ngày lễ, ngày thường, ngày làm việc, ngày nghỉ,…), từng tuần, từng tháng.



Hình 6. Lưu đồ giải thuật xử lý dữ liệu thiếu

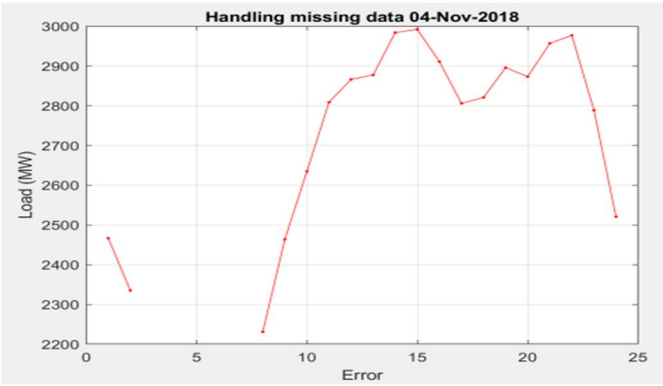
Biểu đồ phụ tải chuẩn hóa đơn vị (SLP) sẽ được đưa vào các modules xây dựng hàm hồi qui theo giải thuật SVR (Support Vector Regression), NN (Neural Network) để xây dựng các hàm hồi qui. Sau đó sử dụng bộ dữ liệu nêu trên để kiểm tra, đánh giá sai số của các hàm hồi qui, từ đó lựa chọn ra được hàm hồi qui có sai số thấp nhất để làm hàm hồi qui ước lượng dữ liệu thiếu.

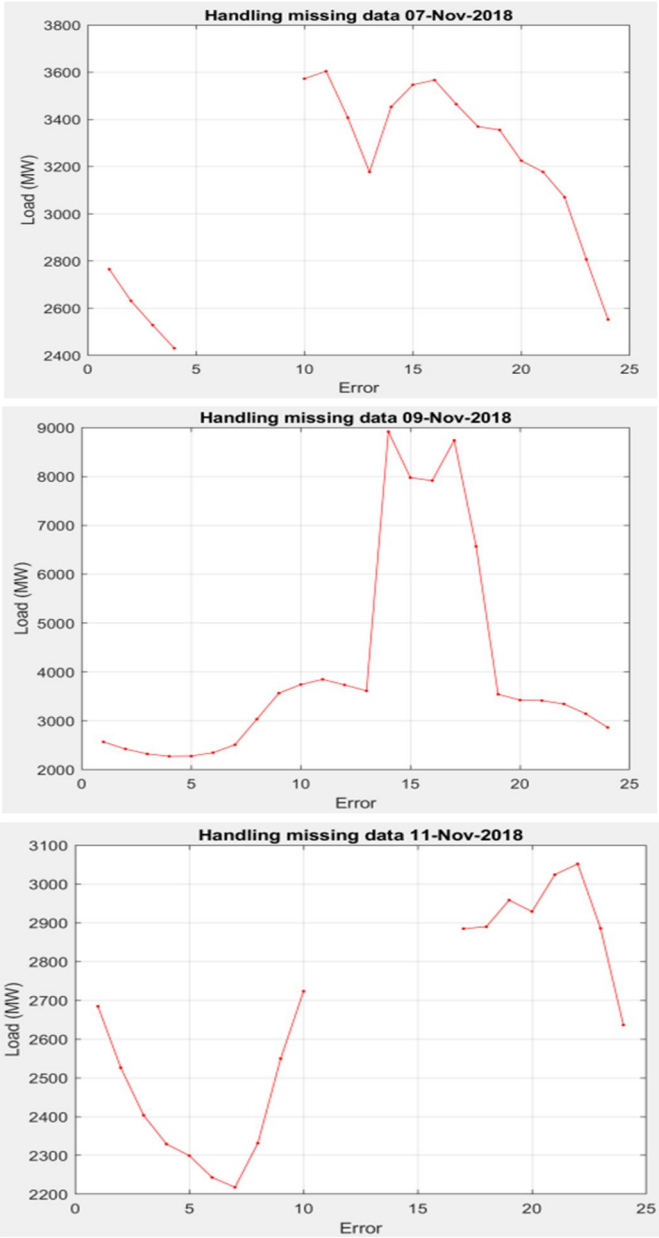
3. KẾT QUẢ NGHIÊN CỨU

3.1. Dữ liệu đầu vào

Dữ liệu đo đếm của phụ tải sử dụng trong việc xây dựng thuật toán gồm: số liệu công suất (P_{max}), điện năng tiêu thụ ($A_{tổng}$) và nhiệt độ (t^0) theo từng giờ, từng ngày trong tháng của các phụ tải tại Tổng công ty Điện lực TP.HCM. Xét một chuỗi dữ liệu đo đếm trong khoảng thời gian từ ngày 01/01/2014 đến 17/12/2018.

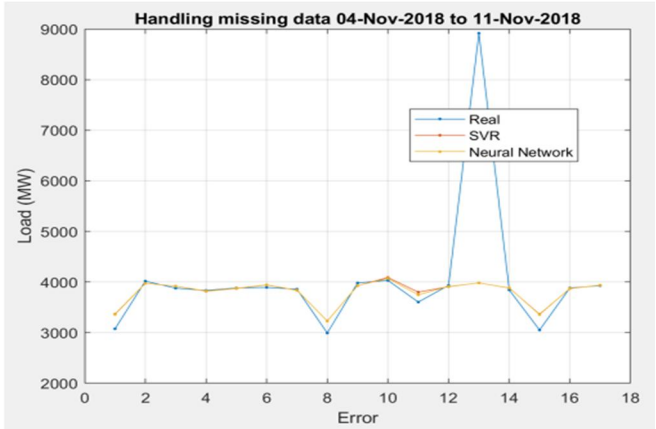
Trong đó có một số chu kỳ dữ liệu điện năng tiêu thụ ($A_{tổng}$) bị thiếu do gián đoạn đo đếm (lỗi giá trị = 0) và lỗi ghi nhận vượt quá (lớn bất thường), để phục vụ nghiên cứu thì cần phải hiệu chỉnh.





Hình 7. Một số ngày dữ liệu bị lỗi một vài chu kỳ

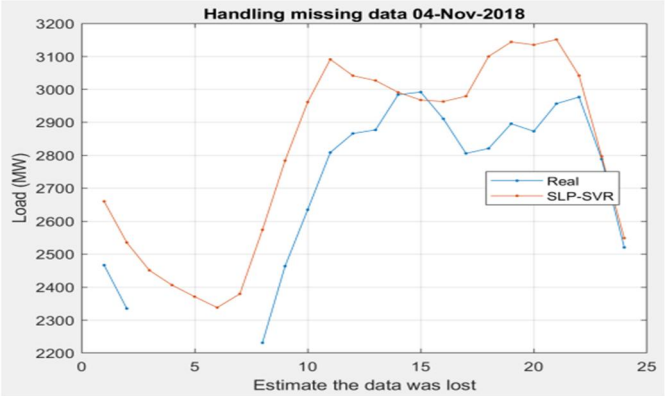
3.2. Kết quả xử lý dữ liệu thiếu



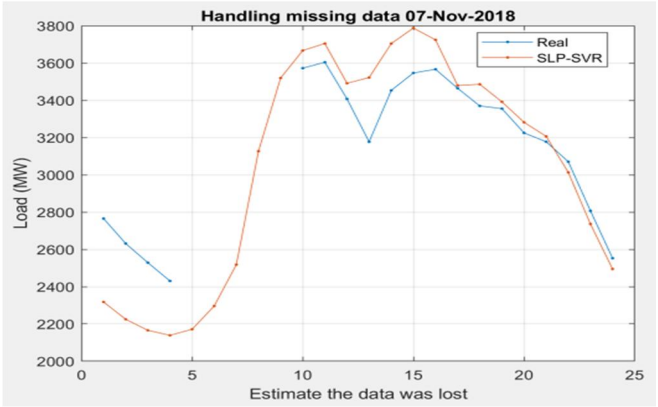
Hình 8. Đường cong phụ tải được xây dựng lại

Đặc điểm của SVR là cho ta một giải pháp thưa (sparse solution); nghĩa là để xây dựng được hàm hồi qui, ta không cần phải sử dụng hết tất cả các điểm dữ liệu trong bộ huấn luyện, những điểm có đóng góp vào việc xây dựng hàm hồi qui được gọi là những Support Vector (việc phân lớp cho một điểm dữ liệu mới sẽ chỉ phụ thuộc vào các support vector). Dựa trên mối quan hệ tuyến tính của ba thành phần số liệu công suất (P_{max}), điện năng tiêu thụ ($A_{tổng}$) và nhiệt độ (t^0), cùng với bộ SLP – SVR (NN) bài báo đã xây dựng lại dường cong phụ tải các ngày bị lỗi

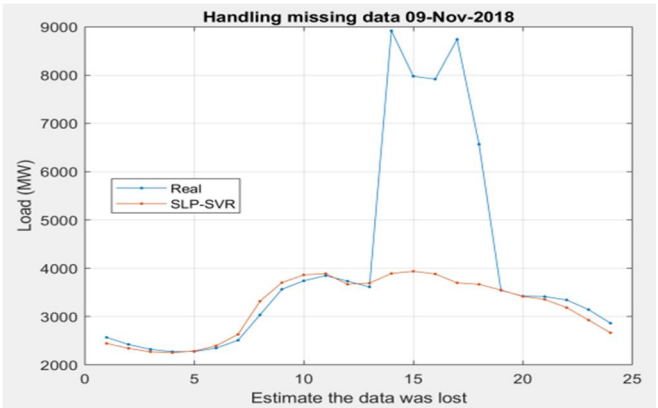
Trên cơ sở SLP của từng chu kỳ của năm 2018 đã xây dựng, chương trình sẽ xây dựng lại biểu đồ phụ tải theo từng chu kỳ của các ngày bị lỗi để xuất ra kết quả ước lượng dữ liệu.



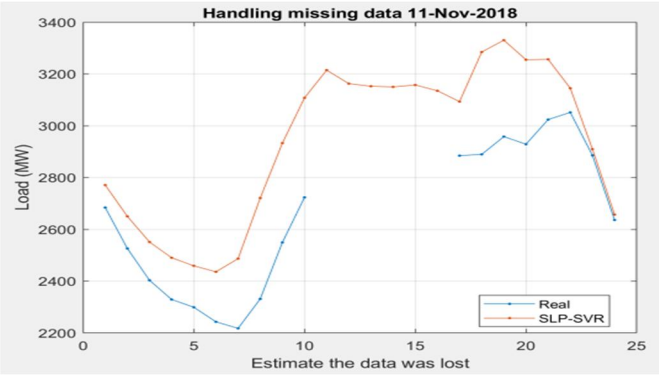
Hình 9. Dữ liệu được xây dựng lại ngày 04/11/2018



Hình 10. Dữ liệu được xây dựng lại ngày 07/11/2018



Hình 11. Dữ liệu được xây dựng lại ngày 09/11/2018



Hình 12. Dữ liệu được xây dựng lại ngày 11/11/2018

4. KẾT LUẬN

Dựa trên mối quan hệ tuyến tính của ba thành phần số liệu công suất (P_{max}), điện năng tiêu thụ ($A_{tổng}$) và nhiệt độ (t^0), cùng với bộ SLP - SVR (NN), bài báo đã xây dựng được công cụ tự động ước lượng các dữ liệu bị lỗi mà trước đây phải thực hiện thực một cách thủ công. Biểu đồ chuẩn hóa đơn vị (SLP) đã góp một phần không nhỏ trong kỹ thuật ước lượng lại dữ liệu bị lỗi. Tuy dữ liệu ước lượng chưa hoàn toàn trùng khớp nhưng phần nào góp phần tạo công cụ nhằm nâng cao độ tin cậy trong việc phân tích, xử lý dữ liệu trong quá trình nghiên cứu phụ tải điện.

TÀI LIỆU THAM KHẢO

[1]. J. W. Grzymala-Busse and M. Hu, 2000. *A comparison of several approaches to missing attribute values in data mining.* Proceedings of the Second International Conference on Rough Sets and Current Trends in Computing RSCTC'2000, October 16-19, 2000, Canada, 340-347.

[2]. Jochen Hardt, Max Herke, Tamara Brian, Wilfried Laubach, 2013. *Multiple Imputation of Missing Data: A Simulation Study on a Binary Response.* Open Journal of Statistics, 3, 370-378

[3]. SAS Institute, 2005. *Multiple Imputation for Missing Data: Concepts and New Approaches.*

[4]. Yuan Yang C., 2011. *Multiple imputation for Missing Data: Concepts and New Development* (SAS Version 9.0). SAS Institute Inc., Rockville, MA)

[5]. Nakai M and Weiming Ke., 2011. *Review of Methods for Handling Missing Data in Longitudinal Data Analysis*. Int. Journal of Math. Analysis. Vol. 5, no.1, 1 -13.

[6]. V.Vapnik, 1995. *"The nature of statistical learning theory"*. Springer, NY.

[7]. S.R. Gunn, 1998: *Support Vector Machines for Classification and Regression*, Technical Report, Image Speech and Intelligent Systems Research Group, University of Southampton.

[8]. V. Cherkassky, Y. Ma, 2002. *Selection of Meta-parameters for Support Vector Regression*. International Conference on Artificial Neural Networks, Madrid, Spain, Aug. pp. 687 - 693.

[9]. D. Basak, S. Pal, D.C. Patranabis, Oct. 2007: *Support Vector* Regression, Neural Information Processing – Letters and Reviews, Vol. 11, No. 10, pp. 203 – 224.

[10]. A.J. Smola, B. Schölkopf, Aug. 2004: *A Tutorial on Support Vector Regression*, Statistics and Computing, Vol. 14, No. 3, pp. 199 – 222.

[11]. Understanding Support Vector Machine Regression and Support Vector Machine Regression, http://www.mathworks.com.

[12]. Thông tư số 33/2011/TT-BCT ngày 06/09/2011 của Bộ Công Thương về Quy định nội dung, phương pháp, trình tự và thủ tục nghiên cứu phụ tải điện