

TỐI ƯU HÓA ƯỚC TÍNH MỨC TIÊU THỤ NĂNG LƯỢNG TRONG CÁC TÒA NHÀ DỰA TRÊN CÁC THUẬT TOÁN TRÍ TUỆ NHÂN TẠO

Trần Đức Học^{a,*}, Lê Tấn Tài^a

^aKhoa Kỹ thuật Xây dựng, Trường Đại học Bách khoa Thành phố Hồ Chí Minh,
số 268 Lý Thường Kiệt, quận 10, thành phố Hồ Chí Minh, Việt Nam

Nhận ngày 11/10/2019, Sửa xong 03/02/2020, Chấp nhận đăng 03/02/2020

Tóm tắt

Mô phỏng và dự báo năng lượng tiêu thụ đóng vai trò quan trọng trong việc thiết lập chính sách năng lượng và đưa ra quyết định theo hướng phát triển bền vững. Nghiên cứu này sử dụng phương pháp kỹ thuật thống kê và công cụ trí tuệ nhân tạo bao gồm mạng nơ-ron thần kinh (ANNs – Artificial neural networks), máy hỗ trợ véc tơ (SVM – Support vector machine), cây phân loại và hồi quy (CART - Classification and regression trees), hồi quy tuyến tính (LR - Linear regression), hồi quy tuyến tính tổng quát (GENLIN - Generalized linear regression), tự động phát hiện tương tác Chi-squared (CHAID - Chi-square automatic interaction detector) và mô hình tổng hợp (Ensemble model) để dự đoán mức tiêu thụ năng lượng trong các căn hộ tòa nhà chung cư. Bộ dữ liệu để xây dựng mô hình gồm 200 mẫu được khảo sát ở nhiều chung cư tại TP. Hồ Chí Minh. Mô hình đơn có hiệu quả tốt nhất trong quá trình dự đoán là CART, trong khi đó mô hình được tổng hợp tốt nhất là CART + GENLIN.

Từ khoá: ước tính; tòa nhà; tiêu thụ năng lượng; khai phá dữ liệu, trí tuệ nhân tạo.

OPTIMIZING ESTIMATION ACCURACY OF ENERGY CONSUMPTION IN RESIDENTIAL BUILDINGS BASED ON A COMPARISON OF ARTIFICIAL INTELLIGENCE ALGORITHMS

Abstract

Energy simulation and forecasting play an important role in setting energy policy and making decisions in pursuit of sustainable development. This research uses statistical tools and artificial intelligence techniques including artificial neural networks, support vector machines, classification and regression trees, linear regression, generalized linear regression, chi-square automatic interaction detector, and ensemble model to predict the energy consumption in apartment buildings. A dataset of two hundred samples, which were obtained from residential buildings in Ho Chi Minh City, Vietnam, were used to evaluate the performance of the developed models. As a result, CART is the best single model, meanwhile the best ensemble model is CART + GENLIN.

Keywords: estimation; residential buildings; energy consumption; data mining.

[https://doi.org/10.31814/stce.nuce2020-14\(1V\)-04](https://doi.org/10.31814/stce.nuce2020-14(1V)-04) © 2020 Trường Đại học Xây dựng (NUCE)

1. Giới thiệu

Hiện nay, tiết kiệm năng lượng là một trong những vấn đề cấp thiết không chỉ trong phạm vi từng quốc gia mà đã trở thành mối quan tâm của toàn thế giới [1]. Biến đổi khí hậu và tăng giá các loại năng lượng cùng với chất lượng cuộc sống con người ngày càng cao [2] dẫn đến sự phụ thuộc nhiều vào các nguồn năng lượng đã tạo ra những thách thức cho đội ngũ kiến trúc sư, kỹ sư phải không

*Tác giả chính. Địa chỉ e-mail: tdhoc@hcmut.edu.vn (Học, T. Đ.)

ngừng sáng tạo và cải tiến các thiết kế trong ngành xây dựng. Thiết kế của các tòa nhà hiện nay phải không chỉ tạo sự tiện nghi, thoải mái cho người sử dụng mà còn cần tối ưu hóa sử dụng năng lượng và giảm thiểu tối đa tác động xấu đối với môi trường.

Trên thế giới, năng lượng sử dụng trong các tòa nhà chiếm đến 20% của tổng năng lượng cần sử dụng cho con người [3]. Ở Việt Nam tỷ lệ này là 27% [2, 4]. Riêng đối với Mỹ và các nước Châu Âu tỷ lệ này chiếm đến 40% [5]. Do đó, việc ước tính mức tiêu thụ năng lượng trong các tòa nhà có ý nghĩa rất quan trọng trong việc tối ưu hóa hiệu suất sử dụng năng lượng nhằm mục đích tiết kiệm năng lượng và giảm thiểu tác động đối với môi trường.

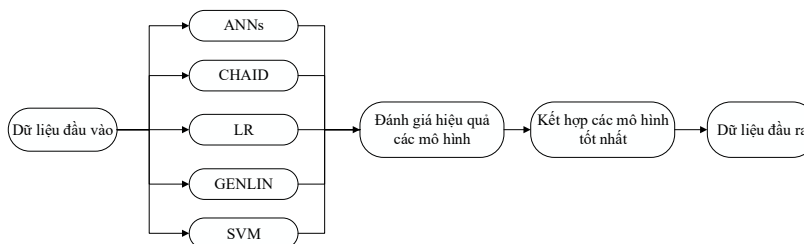
Tuy nhiên, hệ thống năng lượng trong tòa nhà tương đối phức tạp đối với từng công trình nhất định [6]. Các thiết bị tiêu thụ năng lượng chủ yếu trong tòa nhà như: hệ thống thiết bị điều hòa không khí, tủ lạnh, bếp điện...[7]. Kích thước phòng và đặc điểm cửa sổ cũng được thay đổi theo từng loại công trình. Ngoài ra, các yếu tố về điều kiện thời tiết cũng gây ảnh hưởng không nhỏ đối với nhu cầu sử dụng năng lượng của tòa nhà [1].

Do sự phức tạp trong thiết kế các thiết bị ảnh hưởng năng lượng tòa nhà nên việc ước tính chính xác mức tiêu thụ năng lượng gặp nhiều khó khăn. Trong những năm gần đây, rất nhiều nghiên cứu về phương pháp dự đoán đã được đề xuất và áp dụng rộng rãi cho nhiều ngành nghề. Những phương pháp này bao gồm các phương pháp kỹ thuật, thống kê và trí tuệ nhân tạo. Những phương pháp được sử dụng rộng rãi nhất hiện nay gồm: mạng nơ-ron thần kinh (ANN), máy hỗ trợ véc tơ (SVM), cây phân loại và hồi quy (CART), hồi quy tuyến tính (LR), hồi quy tuyến tính tổng quát (GENLIN), tự động phát hiện tương tác Chi-squared (CHAID) được sử dụng trong chương trình SPSS của IBM [8].

Năm 1965, Nilsson [9] giới thiệu về ý tưởng mô hình kết hợp cho các vấn đề phân loại dữ liệu. Kỹ thuật mô hình tổng hợp tổ hợp các điểm mạnh của các mô hình riêng lẻ nhằm mục đích tạo ra sự ước tính tốt hơn. Bởi vì kết hợp nhiều mô hình riêng lẻ sẽ giảm được lỗi giống nhau trong quá trình ước lượng. Do vậy, sử dụng phương pháp nhiều mô hình riêng lẻ kết hợp sẽ tạo ra một mô hình dự đoán mạnh với tính tổng quát hóa cao. Nghiên cứu này sử dụng các kết quả thu được từ các mô hình riêng lẻ tốt nhất sau đó được tổng hợp và xây dựng các mô hình kết hợp để ước lượng mức tiêu thụ năng lượng trong tòa nhà. Mô hình tổng hợp này sử dụng thuật toán logic chéo k -fold trên bộ dữ liệu đã được thu thập từ trước.

2. Các mô hình ước tính và phương pháp đánh giá

Mục tiêu chính của khai phá dữ liệu có được bằng cách kết hợp các phương pháp công nghệ của nhiều lĩnh vực khác nhau, bao gồm khoa học máy tính, thống kê, phân tích dữ liệu trực tuyến, máy học, và các hệ thống chuyên gia [10]. Công nghệ khai phá dữ liệu hiện được áp dụng trong quá trình dự đoán của nhiều lĩnh vực. Hình 1 thể hiện quy trình hoạt động của các mô hình dự đoán đơn bao gồm sáu kỹ thuật khai phá dữ liệu, ANNs, CART, CHAID, LR, GENLIN, và SVMs. Các mô hình đơn được sử dụng để tự động tạo và so sánh kết quả liên tục.



Hình 1. Sơ đồ khối mô hình ước tính mức tiêu thụ năng lượng trong tòa nhà chung cư

2.1. Mạng thần kinh nhân tạo (ANNs)

Mô hình mạng thần kinh nhân tạo là một công cụ mạnh mẽ để giải quyết các vấn đề phức tạp. Mạng thần kinh xử lý các nhân tố như các nơ-ron trong não người, các nhân tố đơn được sắp xếp theo tầng lớp. Mạng thần kinh nhân tạo đã được sử dụng để dự đoán mức tiêu thụ năng lượng ở nhiều nghiên cứu trước đây [11–13]. Trong mạng nơ-ron đa lớp, lớp đầu tiên là tập hợp các nút dữ liệu đầu vào về đặc điểm của căn phòng, sẽ có một hay nhiều lớp ẩn chứa các nút tính toán và một lớp đầu ra chứa một nút biểu thị mức tiêu thụ năng lượng.

Thuật toán máy học được sử dụng rộng rãi và hiệu quả nhất để huấn luyện mạng thần kinh đa lớp là thuật toán lan truyền ngược. Ngưỡng kích hoạt của từng nơ-ron trong lớp ẩn được tính như sau:

$$net_k = \sum w_{kj}o_j \quad \text{và} \quad y_k = f(net_k)$$

trong đó net_k là ngưỡng kích hoạt của nơ-ron thứ k ; j là tập các nơ-ron ở lớp trước; w_{kj} là trọng số kết nối giữa nơ-ron k và nơ-ron j ; o_j là đầu ra của nơ-ron j , và y_k là hàm truyền.

$$f(net_k) = \frac{1}{1 + e^{-net}}$$

Công thức huấn luyện và trọng số cập nhật w_{kj} trong từng chu kỳ t là

$$w_{kj}(t) = w_{kj}(t-1) + \Delta w_{kj}(t)$$

Giá trị thay đổi $\Delta w_{kj}(t)$ được tính như sau

$$\Delta w_{kj}(t) = \eta \delta_{pj} o_{pj} + \alpha w_{kj}(t-1)$$

trong đó η là tham số tốc độ huấn luyện; δ_{pj} là sai số lan truyền; o_{pj} là kết quả đầu ra của nơ-ron j cho lần thứ p ; α là tham số khuếch đại, và $w_{kj}(t-1)$ là giá trị thay đổi cho w_{kj} trong chu kỳ trước.

2.2. Cây phân loại và hồi quy (CART)

Cây phân loại và hồi quy là một phương pháp cây quyết định để xây dựng cây phân loại và cây hồi quy theo loại biến phụ thuộc của nó, vừa có thể theo kiểu phân loại hoặc kiểu số [14]. Với các trường dự đoán như nhau có thể sử dụng nhiều lần ở các cấp cây khác nhau. Phương pháp cây quyết định vượt trội hơn các mô hình kỹ thuật khác khi áp dụng các vấn đề có tính logic cao [8].

Tùy thuộc vào trường mục tiêu, ba biện pháp đo lường có thể được sử dụng để xác định vị trí phân chia cho các mô hình cây phân loại và hồi quy. Chẳng hạn như, biến Gini thường được dùng cho trường mục tiêu tượng trưng trong khi phương pháp độ lệch bình phương nhỏ nhất sẽ tự động chọn các mục tiêu liên tục mà không giải thích được chúng. Chỉ số Gini $g(t)$ tại một nút t trong mô hình cây quyết định, được xác định theo phương trình sau:

$$g(t) = \sum_{j \neq i} p(j|t)p(i|t)$$

trong đó i và j là các loại trường mục tiêu

$$p(j|t) = \frac{p(j, t)}{p(t)}; \quad p(jt) = \frac{\pi(j)N_j(t)}{N_j}; \quad \text{và} \quad p(t) = \sum_j p(j, t)$$

trong đó $p(j)$ là giá trị xác suất trước cho loại j ; $N_j(t)$ là số lượng mẫu trong loại j của nút t , và N_j là số lượng mẫu của loại j của nút gốc. Khi chỉ số Gini được sử dụng để cải thiện sau quá trình phân tách trong quá trình cây phát triển, chỉ các mẫu trong nút và nút gốc với giá trị hợp lệ cho bộ dự đoán phân tách được sử dụng để tính $N_j(t)$ và N_j .

2.3. Tự động phát hiện tương tác Chi-squared (CHAID)

Kỹ thuật tự động phát hiện tương tác Chi-squared để phân loại dữ liệu được phát triển bởi Kass [15]. Nó kiểm tra tính độc lập bằng cách sử dụng kiểm định Chi-square để đánh giá việc tách một nút có cải thiện độ sạch dữ liệu đáng kể hay không. Cụ thể, bộ dự đoán có liên kết mạng nhất (theo giá trị p -value) với biến trả lời tại mỗi nút được sử dụng làm nút chia. Nếu bộ dự đoán được kiểm định cho thấy không có sự cải thiện đáng kể về mặt thống kê thì không có sự phân tách nào được thực hiện và thuật toán dừng lại.

Tự động phát hiện tương tác Chi-squared toàn diện được phát triển để giải quyết các hạn chế của CHAID [16]. Tuy nhiên, kỹ thuật CHAID toàn diện có thể không tối ưu hóa phân tách cho các biến dự báo vì nó dừng việc hợp nhất các biến phân loại ngay khi xác định tất cả các biến phân loại còn lại khác nhau đáng kể. Kỹ thuật CHAID toàn diện tránh việc mô hình quá phù hợp với cây quyết định đã phát triển đầy đủ vào dữ liệu để huấn luyện bằng cách liên tục hợp nhất các bộ dự đoán phân loại cho đến khi chỉ còn hai biến phân loại tốt nhất. Sau đó, nó xác định bộ dự đoán trong mỗi chuỗi các phép hợp nhất và tính giá trị p -value được điều chỉnh cho bộ biến phân loại giúp mang lại sự liên kết tốt nhất với biến mục tiêu. Do đó, CHAID toàn diện tìm ra sự phân tách tốt nhất cho mỗi bộ dự đoán và chọn bộ dự đoán nào để phân tách dựa trên giá trị p -value đã điều chỉnh.

2.4. Hồi quy tuyến tính (LR)

Mô hình hồi quy tuyến tính đa biến (LR) là một phần mở rộng của hồi quy đơn giản, nó xác định mối quan hệ giữa hai hoặc nhiều biến [17]. Công thức chung của mô hình là:

$$Y = \beta_0 + \sum_{i=1}^n \beta_i X_i + \varepsilon$$

trong mô hình đề xuất, Y là mức tiêu thụ năng lượng của căn hộ chung cư; β_0 là hằng số; β_i là hệ số hồi quy ($i = 1, 2, \dots, n$); ε là sai số, và X_i đại diện cho các nhân tố cụ thể. Mô hình hồi quy tuyến tính áp dụng bốn phương pháp hồi quy tuyến tính đa biến bằng cách sử dụng bình phương cực tiểu: stepwise, forward và backward.

2.5. Mô hình tuyến tính tổng quát (GENLIN)

Mô hình tuyến tính tổng quát được phát triển bởi Nelder và Wedderburn [18]. Mô hình có thể phân tích các phân phối xác suất khác nhau (ví dụ như phân phối chuẩn, nhị thức, Poisson và gamma) cho một biến phụ thuộc sử dụng hàm liên kết làm mô hình tính toán để xác định mối quan hệ giữa các yếu tố dự báo tuyến tính và hàm phân phối trung bình.

Mô hình tuyến tính tổng quát linh hoạt hơn và có mối quan hệ thực tế hơn so với hồi quy đơn. Mẫu phân phối giả định của các điểm dữ liệu và mối quan hệ giữa X và Y được xác định theo phương trình sau:

$$\eta = g(E(Y)) = X_i \beta_i + O, Y \sim F$$

trong đó η là bộ dự đoán tuyến tính, O là biến bù, X_i là biến độc lập, β_i là hệ số độ dốc và F là phân phối của Y .

Ba thành phần của mô hình tuyến tính tổng quát bao gồm một biến kết quả Y với phân phối ngẫu nhiên cụ thể và giá trị kỳ vọng μ và phương sai $\sigma^2(E(Y) = \mu)$. Một hàm liên kết $g(\cdot)$ kết nối giá trị kỳ vọng (μ) của Y để biến đổi các giá trị dự đoán của $\eta[\eta = g(\mu)]$; và một mô hình cấu trúc tuyến tính.

2.6. Máy hỗ trợ véc tơ (SVM)

Vapnik [19] là người đầu tiên giới thiệu máy hỗ trợ véc tơ. Các máy hỗ trợ véc tơ được tạo bởi các hàm ánh xạ đầu vào – đầu ra từ một tập dữ liệu. Hàm này giải quyết cả vấn đề phân loại lẫn hồi quy. Thông thường, mô hình hồi quy sử dụng hồi quy hỗ trợ véc tơ epsilon [20] để tìm một hàm $f(x)$ có độ lệch ε tối đa được thu thập từ các dữ liệu đầu ra y_i . Trong máy hỗ trợ véc tơ hồi quy, đầu vào đầu tiên được ánh xạ vào một vùng đặc tính không gian n chiều bằng cách sử dụng hàm phi tuyến như sau:

$$f(x, \omega) = \langle \omega, x \rangle + b \quad \text{với} \quad \omega \in \chi, b \in \chi$$

Chất lượng của hàm $f(x)$ có thể được ước tính dựa trên tổn thất của hàm $L(x)$ như sau:

$$L_\varepsilon = [y, f(x, \omega)] = \begin{cases} 0 & \text{nếu } |y - f(x, \omega)| \leq \varepsilon \\ |y - f(x, \omega)| & \text{khác} \end{cases}$$

Đặc tính mới nhất của hỗ trợ véc tơ hồi quy là sử dụng tổn thất không nhạy cảm ε để tính toán hàm hồi quy tuyến tính cho không gian đặc tính cao hơn đồng thời giảm độ phức tạp của mô hình bằng cách giảm thiểu $\|\omega\|^2$. Hàm này được đưa ra bằng cách lấy tổng không âm của các hàm ξ_i và ξ_i^* , trong đó $i = 1, \dots, n$ được sử dụng để xác định các mẫu huấn luyện từ vùng không nhạy cảm ε . Do đó, hỗ trợ véc tơ hồi quy có thể được coi là một phiên bản thu nhỏ của hàm sau:

$$\min \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^n (\xi_i + \xi_i^*) \quad \text{với} \quad \begin{cases} y_i - f(x_i, \omega) \leq \varepsilon + \xi_i^* \\ f(x_i, \omega) - y_i \leq \varepsilon + \xi_i \\ \xi_i, \xi_i^* \geq 0, i = 1, \dots, n \end{cases}$$

trong đó hằng số $C \geq 0$ xác định sự cân bằng giữa độ phẳng của $f(x, \omega)$ và dung sai cho độ lệch lớn hơn so với ε .

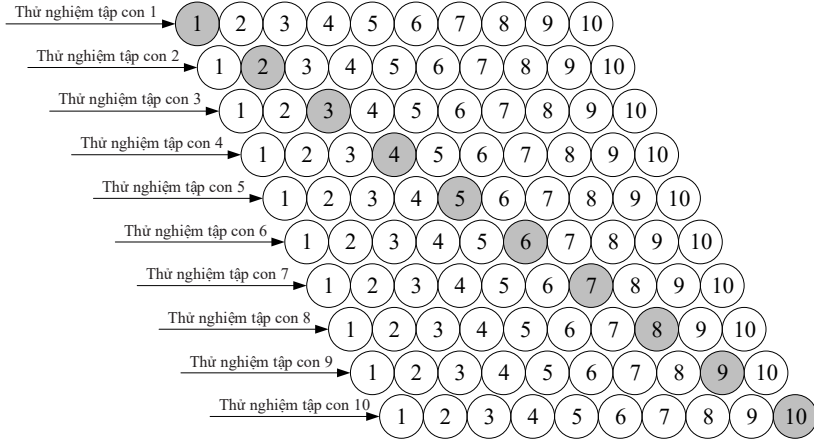
2.7. Các mô hình kết hợp

Các mô hình được xếp hạng dựa vào quá trình dự đoán và sau đó các mô hình có tỷ lệ dự đoán tốt nhất được kết hợp lại tạo thành mô hình kết hợp. Phương pháp kết hợp được thể hiện bằng phép toán là $g : \mathbb{R}^d \rightarrow \mathbb{R}$ với một biến dự đoán X và biến phản hồi Y . Mỗi phương pháp sử dụng một thuật toán xác định để đưa ra một hàm ước tính $g(\cdot)$. Ước tính bằng một hàm kết hợp $g_{en}(\cdot)$ tạo ra được bằng cách kết hợp tuyến tính của các hàm riêng lẻ như sau:

$$g_{en}(\cdot) = \sum_{j=1}^n c_j * g(\cdot)$$

trong đó c_j chứa các hệ số kết hợp tuyến tính, là giá trị trung bình của trọng số khác nhau.

Nhìn chung, phương pháp mô hình kết hợp ước tính chính xác hơn so với những mô hình riêng lẻ thông thường [21, 22]. Các nghiên cứu thường áp dụng thuật toán xác thực chéo k lần để giảm thiểu sai số liên quan đến lấy mẫu ngẫu nhiên của việc huấn luyện. Kohavi [23] đã xác nhận rằng thử nghiệm 10 lần đem lại thời gian tính toán và phương sai tối ưu. Phương pháp này phân chia tập mẫu dữ liệu thành 10 tập con, tiến hành xây dựng và xác thực mô hình 10 lần, chọn 1 tập dữ liệu khác để kiểm tra, huấn luyện mô hình bằng 9 tập dữ liệu và sử dụng tập còn lại để kiểm tra tính chính xác của mô hình được minh họa ở Hình 2. Độ chính xác của mô hình được tính bằng độ chính xác trung bình của 10 mô hình trong 10 lần xác thực.



Hình 2. Phương pháp xác thực chéo 10 lần

2.8. Phương pháp đánh giá hiệu suất

Để đánh giá độ chính xác của quá trình dự đoán của các mô hình đơn lẻ và mô hình kết hợp, các phương pháp sau đã được sử dụng:

- Phần trăm sai số trung bình tuyệt đối

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{y - y'}{y} \right|$$

- Sai số trung bình tuyệt đối

$$MAE = \frac{1}{n} \sum_{i=1}^n |y - y'|$$

- Sai số toàn phương trung bình

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y' - y)^2}$$

Ta sử dụng chỉ số tổng hợp (SI) thông qua ba phương pháp thống kê MAPE, MAE, RMSE với công thức như sau:

$$SI = \frac{1}{m} \sum_{i=1}^m \left(\frac{P_i - P_{\min,i}}{P_{\max,i} - P_{\min,i}} \right)$$

trong đó m là số phương pháp đánh giá; P_i là hiệu suất thứ i . Giá trị của SI là từ 0 đến 1; SI càng gần 0 độ chính xác mô hình càng cao.

3. Thiết lập thực nghiệm

3.1. Mô tả và chuẩn bị dữ liệu

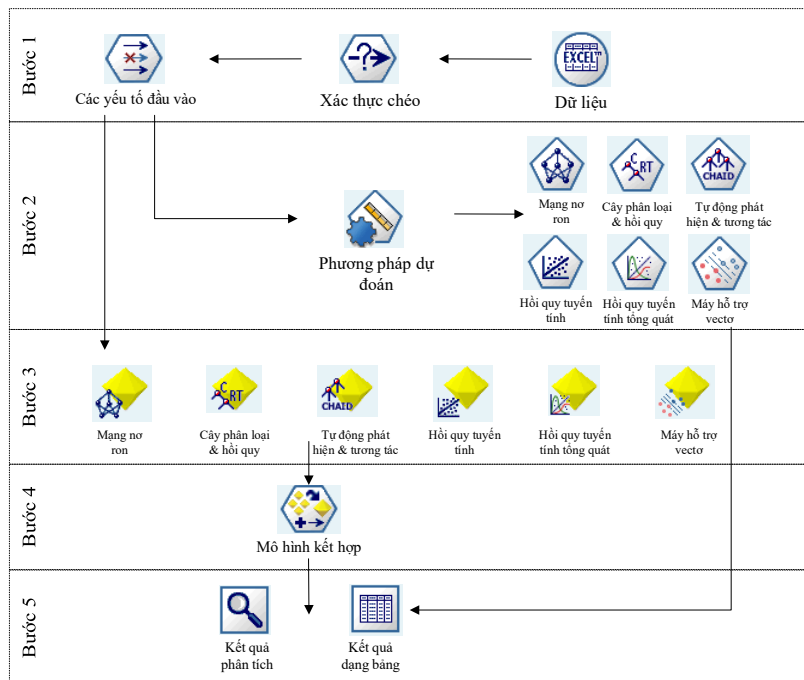
Dữ liệu được khảo sát bằng bảng câu hỏi và được đưa đến từng hộ gia đình. Bảng câu hỏi bao gồm 9 nhân tố ảnh hưởng nhất đến nhu cầu sử dụng điện của căn hộ chung cư cũng chính là 9 biến

đầu vào của bộ dữ liệu được thể hiện ở Bảng 1. Bộ dữ liệu gồm có 200 mẫu được khảo sát ở nhiều chung cư tại Thành phố Hồ Chí Minh. Trung bình mức tiêu thụ năng lượng theo tháng được thu thập từ tháng 10 năm 2018 đến tháng 4 năm 2019.

Các chung cư được khảo sát nằm ở các quận khác nhau trên địa bàn Thành phố Hồ Chí Minh để đảm bảo tính bao phủ về vị trí. Đồng thời, các chung cư có mức độ tiện nghi từ trung bình cho đến cao cấp. Các tòa nhà này hoàn thành từ năm 2015 với trên 90% cư dân sinh sống. Đối với căn hộ chung cư, dữ liệu được thu thập từ các hộ gia đình có người đang sinh sống. Trong mỗi tầng sẽ tiến hành khảo sát các căn hộ chung cư ở các vị trí và đặc điểm khác biệt (vị trí giữa/góc, hướng căn hộ, số phòng ngủ, ...) để đảm bảo tính đa dạng của đối tượng được khảo sát [24].

Bảng 1. Mô tả dữ liệu

Tên nhân tố	Tên biến	Đơn vị	Loại
Tổng diện tích	X_1	m^2	Range
Tổng số lượng cửa sổ	X_2	Cái	Range
Hướng cửa sổ	X_3		Set
Loại kính cửa sổ cách nhiệt	X_4	Có/không	Flag
Số lượng máy điều hòa	X_5	Cái	Range
Số lượng lò nướng	X_6	Cái	Range
Kích cỡ tủ lạnh	X_7	Lít	Range
Tổng số người sinh sống	X_8	Người	Range
Mật độ tòa nhà	X_9	[1,5]	Range
Mức tiêu thụ năng lượng	Y	kWh	Range



Hình 3. Cấu trúc mô hình ước tính mức tiêu thụ năng lượng trong tòa nhà chung cư

3.2. Cấu trúc mô hình

Các tham số của mô hình được đặt mặc định trong quá trình so sánh thử nghiệm của các kỹ thuật khai phá dữ liệu nhằm đảm bảo cho mô hình hoạt động một cách khách quan, dễ dàng và thỏa mãn về mức độ hoạt động và độ chính xác.

Hình 3 minh họa các bước dùng các mô hình để dự đoán mức độ tiêu thụ năng lượng thông qua phần mềm SPSS của IBM [8].

- Bước 1: Nhập dữ liệu đầu vào nút nguồn dựa trên thuật toán xác thực chéo.
- Bước 2: Sử dụng nút dự đoán số để đào tạo dữ liệu.
- Bước 3: Sử dụng mô hình đơn để kiểm tra dữ liệu.
- Bước 4: Kết hợp các mô hình thông qua nút kết hợp.
- Bước 5: Đánh giá kết quả phân tích thông qua bảng kết quả.

4. Kết quả thực nghiệm

Phần này đưa ra kết quả phân tích và so sánh hiệu suất mô hình kết hợp với sáu mô hình dự đoán đơn được đề xuất bằng hệ số tổng hợp SI. Từ kết quả xếp hạng hiệu suất ta kết hợp các mô hình đơn với nhau nhằm tăng hiệu quả dự đoán.

Bảng 2 thể hiện hiệu suất của các mô hình bao gồm ANNs, CART, CHAID, LR, GENLIN và SVM. Trong đó, mô hình có hiệu suất tốt nhất trong quá trình dự đoán là CART với chỉ số tổng hợp SI là 0,0. Tiếp sau đó là mô hình GENLIN cũng có hiệu suất tương đối tốt với chỉ số tổng hợp SI 0,131. Xếp vị trí thứ ba và thứ tư lần lượt là hai mô hình CHAID (SI = 0,243) và LR (SI = 0,353). Hai mô hình dự đoán ANNs (SI = 0,51) và SVM (SI = 1,00) có hiệu suất thấp nhất.

Bảng 2. Tổng hợp kết quả thử nghiệm của các mô hình đơn

Model	MAPE (%)	MAE (kWh)	RMSE (kWh)	SI
ANNs	28,822	55,917	68,414	0,510 (5)
CART	6,289	11,833	22,959	0,000 (1)
CHAID	14,106	30,306	45,592	0,243 (3)
LR	19,078	40,813	50,706	0,353 (4)
GENLIN	11,112	23,589	31,647	0,131 (2)
SVM	44,193	83,235	108,369	1,000 (6)

Tiếp theo ta tiến hành kết hợp các mô hình có hiệu suất tốt nhất với nhau nhằm tăng hiệu quả của quá trình dự đoán. Ta sử dụng 4 mô hình kết hợp như sau:

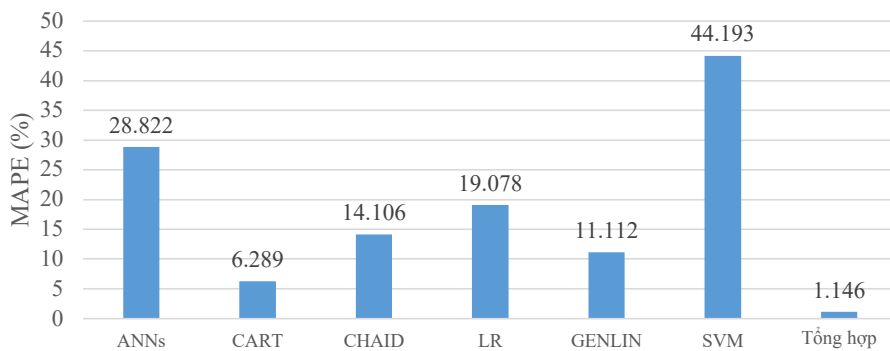
- Mô hình kết hợp 5 mô hình đơn tốt nhất: CART + GENLIN + CHAID + LR + ANNs
- Mô hình kết hợp 4 mô hình đơn tốt nhất: CART + GENLIN + CHAID + LR
- Mô hình kết hợp 3 mô hình đơn tốt nhất: CART + GENLIN + CHAID
- Mô hình kết hợp 2 mô hình đơn tốt nhất: CART + GENLIN

Sau khi kết hợp các mô hình với nhau ta có được hiệu suất các mô hình ở Bảng 3. Mô hình kết hợp có hiệu suất tốt nhất bao gồm ba mô hình (CART + GENLIN) với hệ số tổng hợp SI chỉ có 0. Tiếp đó mô hình kết hợp gồm (CART + GENLIN + CHAID) có hệ số tổng hợp SI = 0,562 là mô hình kết hợp có hiệu suất tốt thứ 2. Hai mô hình kết hợp còn lại là (CART + GENLIN + CHAID + LR) và (CART + GENLIN + CHAID + LR + ANNs) có hiệu suất thấp nhất.

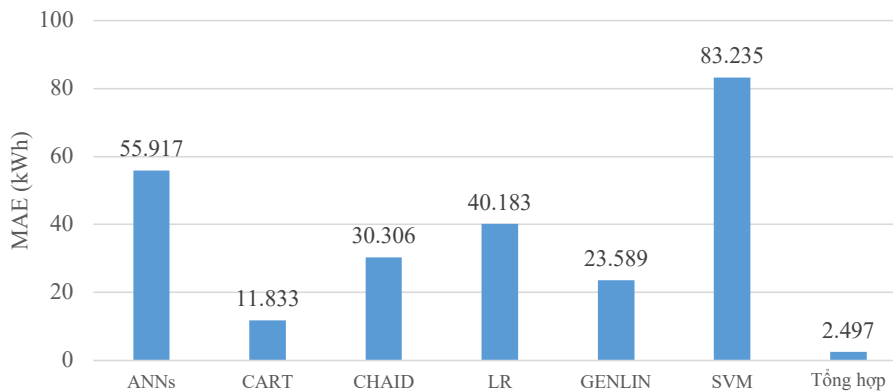
Bảng 3. Tổng hợp kết quả thử nghiệm của các mô hình tổng hợp

Model	MAPE (%)	MAE (kWh)	RMSE (kWh)	SI
Kết hợp 5 mô hình tốt nhất	12,314	25,002	30,825	1,000 (4)
Kết hợp 4 mô hình tốt nhất	9,135	19,437	25,553	0,754 (3)
Kết hợp 3 mô hình tốt nhất	7,078	14,934	20,591	0,562 (2)
Kết hợp 2 mô hình tốt nhất	1,146	2,497	5,187	0,000 (1)

Hình 4-6 thể hiện các chỉ số thống kê đã nêu ở mục 2.8 nhằm đánh giá hiệu quả của các mô hình dự đoán bao gồm ANNs, CART, CHAID, LR, GENLIN, SVM và mô hình kết hợp của 2 mô hình tốt nhất (CART + GENLIN). Mô hình kết hợp (Ensemble) có các chỉ số thống kê tốt hơn hẳn so với các mô hình đơn như phần trăm sai số trung bình tuyệt đối (MAPE) chỉ có 1,146%, sai số trung bình tuyệt đối (MAE) là 2,497 kWh và sai số toàn phương trung bình (RMSE) là 5,187 kWh.

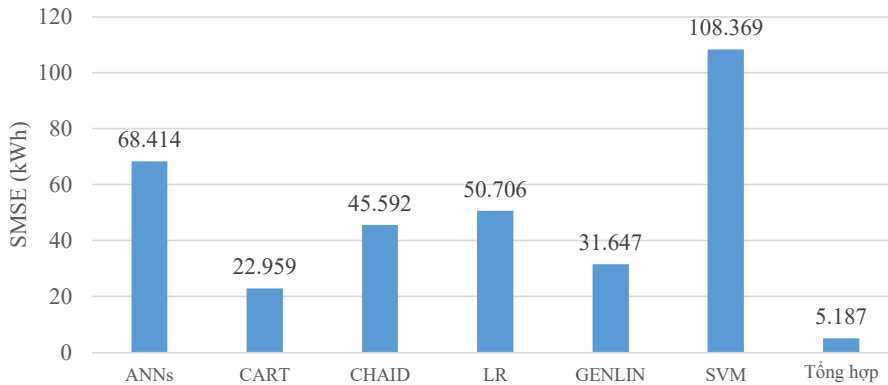


Hình 4. Biểu đồ phần trăm sai số trung bình tuyệt đối (MAPE) của các mô hình



Hình 5. Biểu đồ sai số trung bình tuyệt đối (MAE) của các mô hình

Các kết quả này thu được thông qua phương pháp k -fold ($k = 10$), bằng cách lấy trung bình của 10 lần chạy kiểm nghiệm. Từ Hình 4-6 ta thấy mô hình kết hợp là mô hình tốt nhất. Điều này chứng minh rằng mô hình kết hợp là mô hình phù hợp nhất cho việc dự đoán mức tiêu thụ năng lượng.



Hình 6. Biểu đồ sai số toàn phương trung bình (RMSE) của các mô hình

5. Kết luận

Bài báo này trình bày đề xuất các mô hình dựa vào thuật toán về trí tuệ nhân tạo bao gồm mạng nơ-ron thần kinh (ANNs – Artificial neural network), máy hỗ trợ véc tơ (SVMs), cây phân loại và hồi quy (CART), hồi quy tuyến tính (LR), hồi quy tuyến tính tổng quát (GENLIN), tự động phát hiện tương tác Chi-squared (CHAID) được sử dụng trong chương trình SPSS của IBM nhằm áp dụng trong việc dự đoán mức tiêu thụ năng lượng trong tòa nhà chung cư. Từ các mô hình đơn lẻ đó, tiếp tục xây dựng mô hình tổng hợp để tối ưu hóa dự đoán, khắc phục các nhược điểm của những mô hình riêng lẻ. Dữ liệu được sử dụng trong việc dự đoán bao gồm 200 khảo sát ở nhiều chung cư tại TP. Hồ Chí Minh. Đồng thời, nghiên cứu đã áp dụng thuật toán xác thực chéo 10 lần để giảm thiểu sai số trong quá trình huấn luyện mô hình.

Nghiên cứu đã chỉ ra rằng mô hình đơn có hiệu quả tốt nhất trong quá trình dự đoán mức độ tiêu thụ năng lượng là mô hình cây phân loại và hồi quy (CART) với chỉ số thống kê tốt hơn hẳn so với các mô hình đơn khác, đồng thời có hệ số tổng hợp SI tốt nhất trong các mô hình đã đề xuất ($SI = 0$). Mô hình tổng hợp có hiệu suất tốt nhất là mô hình được tổng hợp giữa 2 mô hình đơn bao gồm: CART + GENLIN ($SI = 0$).

Lời cảm ơn

Nghiên cứu này được tài trợ bởi Quỹ Phát triển khoa học và công nghệ Quốc gia (NAFOSTED) trong đề tài mã số 102.05-2018.07.

Tài liệu tham khảo

- [1] Mitterer, C., Künzle, H. M., Herkel, S., Holm, A. (2012). Optimizing energy efficiency and occupant comfort with climate specific design of the building. *Frontiers of Architectural Research*, 1(3):229–235.
- [2] Denmark, E., Trade, M. (2017). *Vietnam energy outlook report*. Bộ Công Thương, Việt Nam.
- [3] Davis, E. W., Heidorn, G. E. (1971). [An algorithm for optimal project scheduling under multiple resource constraints](#). *Management Science*, 17(12):B803–B816.
- [4] Trung, N. T., Toi, P. V. (2018). [Nghiên cứu, đề xuất suất tiêu thụ điện năng cho văn phòng làm việc: Áp dụng tính toán cho văn phòng làm việc tại Hà Nội](#). *Tạp chí Khoa học Công nghệ Xây dựng (KHCN XD)-ĐHXD*, 12(2):59–64.
- [5] Pérez-Lombard, L., Ortiz, J., Pout, C. (2008). [A review on buildings energy consumption information](#). *Energy and Buildings*, 40(3):394–398.

- [6] Zhao, H.-x., Magoulès, F. (2012). [A review on the prediction of building energy consumption](#). *Renewable and Sustainable Energy Reviews*, 16(6):3586–3592.
- [7] Lượng, N. Đ., Nga, T. T. V., Hiệp, N. H., Giang, H. M., Minh, N. B. (2018). [Ứng dụng BIM để mô phỏng lượng nhiệt bức xạ mặt trời tác động lên một tòa nhà văn phòng ở thành phố Hà Nội](#). *Tạp chí Khoa học Công nghệ Xây dựng (KHCN XD)-ĐH XD*, 12(1):83–88.
- [8] IBM (2010). *PASW Modeler*. IBM Cororation, USA.
- [9] Nilsson, N. J. (1965). *Learning machines: Foundations of trainable pattern classifying systems*. McGraw-Hill, New York.
- [10] Liao, S.-H., Chu, P.-H., Hsiao, P.-Y. (2012). [Data mining techniques and applications—A decade review from 2000 to 2011](#). *Expert Systems with Applications*, 39(12):11303–11311.
- [11] Li, K., Su, H., Chu, J. (2011). [Forecasting building energy consumption using neural networks and hybrid neuro-fuzzy system: A comparative study](#). *Energy and Buildings*, 43(10):2893–2899.
- [12] Tso, G. K. F., Yau, K. K. W. (2007). [Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks](#). *Energy*, 32(9):1761–1768.
- [13] Ekici, B. B., Aksoy, U. T. (2009). [Prediction of building energy consumption by using artificial neural networks](#). *Advances in Engineering Software*, 40(5):356–362.
- [14] Breiman, L. (2017). *Classification and regression trees*. Routledge.
- [15] Kass, G. V. (1980). [An exploratory technique for investigating large quantities of categorical data](#). *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 29(2):119–127.
- [16] Biggs, D., De Ville, B., Suen, E. (1991). A method of choosing multiway partitions for classification and decision trees. *Journal of Applied Statistics*, 18(1):49–62.
- [17] Neter, J., Wasserman, W., Kutner, M. H. (1986). Applied linear statistical models. *Journal of the American Statistical Association*, 81.
- [18] Nelder, J. A., Wedderburn, R. W. M. (1972). [Generalized linear models](#). *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384.
- [19] Vapnik, V. (1995). *The nature of statistical learning theory*. Springer-Verlag, New York.
- [20] Clementine, S. P. S. S. (2007). *12.0 Algorithm Guide*. Integral Solutions Limited, Chicago, USA.
- [21] Adeodato, P. J. L., Arnaud, A. L., Vasconcelos, G. C., Cunha, R. C. L. V., Monteiro, D. S. M. P. (2011). [MLP ensembles improve long term prediction accuracy over single networks](#). *International Journal of Forecasting*, 27(3):661–671.
- [22] Van Wezel, M., Potharst, R. (2007). [Improved customer choice predictions using ensemble methods](#). *European Journal of Operational Research*, 181(1):436–452.
- [23] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *International Joint Conference on Artificial Intelligence (IJCAI)*, Montreal, Canada, 14(2):1137–1145.
- [24] Học, T. Đ., Hậu, Đ. T., Tài, L. T. (2019). Các nhân tố ảnh hưởng đến khả năng tiêu thụ điện trong căn hộ chung cư. *Tạp chí Xây dựng*, 9:31–35.