

Ứng dụng thuật toán học máy LightGBM cho bài toán hồi quy ước lượng khả năng chịu tải của giàn thép sử dụng phân tích trực tiếp

Application of LightGBM algorithm for regression problem of predicting load-carrying capacity of steel trusses using direct analysis

> TS MAI SỸ HÙNG

Khoa Công trình thủy, Trường Đại học Xây dựng Hà Nội

TÓM TẮT

Kỹ thuật học máy dựa trên trí tuệ nhân tạo đang phát triển hết sức nhanh chóng và thể hiện hiệu quả to lớn trong nhiều lĩnh vực của cuộc sống bao gồm thiết kế công trình. Các kỹ thuật học máy thường được xây dựng dựa trên lý thuyết và cấu trúc phức tạp, đòi hỏi người sử dụng phải có trình độ hiểu biết nhất định để sử dụng hiệu quả chúng. Ngoài ra, đối với mỗi dạng bài toán khác nhau, hiệu quả của các thuật toán học máy cũng thay đổi. Nhằm cung cấp cho các kỹ sư thiết kế một tài liệu tham khảo hữu ích về ứng dụng học máy trong thiết kế công trình, trong bài báo này, tác giả giới thiệu thuật toán học máy LightGBM (light gradient boosting machine) cho bài toán hồi quy ước lượng khả năng chịu tải của giàn thép. Phân tích trực tiếp có xét đến phi tuyến tính phi đàn hồi được sử dụng để xác định khả năng chịu tải của công trình. Trên cơ sở đó, tập dữ liệu huấn luyện sẽ được xây dựng với biến đầu vào là tiết diện của thanh giàn và đầu ra là hệ số chịu tải của công trình. Một cấu thép phẳng 113 thanh được xem xét để thể hiện hiệu suất làm việc của LightGBM. Kết quả tính toán cho thấy LightGBM có độ chính xác cao trong việc ước lượng khả năng chịu tải của kết cấu giàn phi tuyến và có thể áp dụng hỗ trợ công tác thiết kế hàng ngày.

Từ khóa: Học máy; giàn; LightGBM; phân tích trực tiếp.

ABSTRACT

Machine learning techniques based on artificial intelligence are developing very rapidly and showing great effectiveness in many areas of life including building design. Machine learning techniques are often built on complex theories and structures, requiring users to have a certain level of understanding to use them effectively. In addition, for each different type of problem, the efficiency of machine learning algorithms also changes. In order to provide design engineers with a useful reference on the application of machine learning in building design, in this paper, the author introduces the LightGBM (light gradient boosting machine) machine learning algorithm for the problem. Regression estimates the load-carrying capacity of steel trusses. Direct analysis taking into account nonlinear inelasticity is used to determine the load-carrying capacity of the building. On that basis, the training dataset will be built with the input variable being the cross-section of the truss rod and the output being the load-bearing factor of the building. A 113-bar planar steel bridge is considered to demonstrate the performance of LightGBM. Calculation results show that LightGBM has high accuracy in estimating the load-carrying capacity of nonlinear truss structures and can be applied to support daily design work.

Keyword: Machine learning; trusses; LightGBM; direct design.

1. ĐẶT VẤN ĐỀ

Đặc tính ứng xử phi tuyến từ cả vật liệu và hình học của kết cấu thép nói chung và giàn thép nói riêng cần được xem xét trong quá trình thiết kế công trình. Các phương pháp truyền thống xét đến ứng xử này thông qua thiết kế 2 bước với bước thứ nhất là tính toán nội lực trong các cấu kiện trên cơ sở sử dụng phân tích đàn hồi tuyến tính. Sau đó, sự an toàn của từng cấu kiện được kiểm tra riêng lẻ thông qua các công thức cho sẵn trong tiêu chuẩn thiết kế mà các hệ số được xây dựng có xét đến độ an toàn của công trình

chịu ứng xử phi tuyến. Các phương pháp này không mô tả được ứng xử của công trình khi chịu tải cũng như không xác định được khả năng chịu tải lớn nhất của công trình. Để khắc phục nhược điểm này, các phương pháp phân tích trực tiếp đang ngày được sử dụng rộng rãi [1-6]. Trong cách tiếp cận này, ứng xử của công trình bao gồm cả ứng xử phi tuyến hình học và phi tuyến vật liệu sẽ được xác định một cách trực tiếp trong quá trình tính toán và qua đó xác định được khả năng chịu tải của công trình. Chính vì vậy, phương pháp này cho phép loại bỏ việc kiểm tra an toàn cho từng

cấu kiện riêng lẻ như trong các phương pháp thiết kế truyền thống. Ngoài ra, các nghiên cứu cho thấy phân tích trực tiếp còn có thể tiết kiệm vật liệu [7]. Tuy nhiên, các phương pháp phân tích trực tiếp sử dụng thời gian tính toán nhiều hơn rất nhiều lần so với phân tích đàn hồi tuyến tính. Chính vì vậy, việc vận dụng chúng vào các bài toán phức tạp còn khá hạn chế, đặc biệt là đối với các bài toán đòi hỏi số lần phân tích kết cấu rất lớn như thiết kế tối ưu sử dụng metaheuristic hay tính toán xác suất hư hỏng của công trình theo các phương pháp mô phỏng.

Các loại mô hình học máy có thể xem là một phương án tiềm năng để giải quyết vấn đề trên. Trong các mô hình học máy, mối quan hệ phi tuyến ở mức độ cao giữa dữ liệu đầu vào và dữ liệu đầu ra được biểu diễn thông qua các mô hình được huấn luyện dựa trên tập dữ liệu cho trước được tạo ra bằng phân tích trực tiếp, các dữ liệu đầu vào và đầu ra dùng để huấn luyện mô hình học máy được lấy từ kết quả phân tích nâng cao như đã đề cập ở trên. Lúc này, các ứng xử phi tuyến của công trình có thể được ước lượng mà không cần thực hiện quá trình phân tích phi tuyến tốn thời gian. Một số mô hình học máy nổi tiếng có thể kể đến như là: Kriging [8], artificial neural network (ANN) [9], support vector machines (SVM) [10], polynomial response surface method (RSM) [11], v.v... Gần đây, Microsoft đã giới thiệu một thuật toán học máy rất mạnh mẽ là LightGBM dựa trên nền tảng của các thuật toán cây ra quyết định [12]. Thuật toán này hiệu quả cho cả bài toán hồi quy cũng như phân loại với ưu điểm mạnh nhất là tốc độ xử lý rất nhanh và độ chính xác cao. Do là thuật toán mới, việc áp dụng LightGBM vào các lĩnh vực khác nhau còn khá hạn chế và mới mẻ. Do vậy, cần nhiều nghiên cứu chuyên sâu nhằm phân tích và đánh giá toàn diện hiệu quả của thuật toán này đối với các lớp bài toán cụ thể.

Trên cơ sở đó, trong bài báo này, tác giả sẽ giới thiệu nội dung thuật toán LightGBM và việc ứng dụng thuật toán này trong việc giải quyết bài toán hồi quy ước lượng khả năng chịu tải của giàn thép sử dụng phân tích trực tiếp. Cấu giàn thép phẳng 113 thanh được sử dụng để minh họa cho nghiên cứu. Phần tiếp theo của bài báo như sau: nội dung cơ bản của thuật toán LightGBM được trình bày trong phần 2; phần ba là các thông tin cơ bản của cấu giàn thép 113 thanh; phần 4 trình bày kết quả tính toán; và cuối cùng là kết luận.

2. THUẬT TOÁN LIGHTGBM

Nguyên lý căn bản của thuật toán LightGBM được phát triển từ mô hình cây ra quyết định tăng cường độ dốc Gradient Tree Boosting (GTB) do Friedman đề xuất [13, 14]. Các kết quả nghiên cứu cho thấy rằng mô hình LightGBM có thể xử lý hiệu quả các bài toán về phân loại, hồi quy và xếp hạng trong học máy. Để hiểu về nguyên lý thuật toán LightGBM, trước hết chúng ta cùng tìm hiểu lý thuyết căn bản của thuật toán GTB. Nguyên lý chung của GTB là thu được câu trả lời cuối cùng bằng cách kết hợp nhiều cây quyết định và bằng cách cộng kết quả của tất cả các cây quyết định. Quá trình này đã được cải thiện để đạt được mức tăng cường độ dốc cực cao trong thuật toán XGBoost. Sự khác biệt giữa XGBoost và GTB là ở cách phân chia cây và cách xác định giá trị của nút lá. Ý tưởng cốt lõi là tiến hành mở rộng Taylor bậc hai của hàm mất mát để phù hợp với GTB và giới thiệu thuật ngữ thông thường của cây một cách thông minh, để công thức của khai triển Taylor bậc hai có thể được đơn giản hóa và giải quyết bằng phương pháp phân tích. do đó, một phương pháp tách cây mới và phương pháp xác định giá trị nút lá đã ra đời. LightGBM được tối ưu hóa hơn nữa trên cơ sở cải tiến công thức GTB của XGBoost.

Nguyên lý cơ bản được sử dụng trong thuật toán GTB là việc kết hợp các cây mô hình học tập cơ bản yếu (tức là có độ sai số cao) thành một cây mô hình học tập mạnh hơn theo kiểu tuần tự.

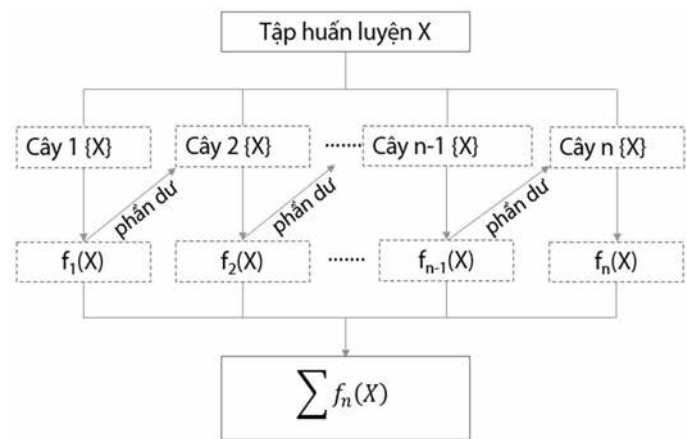
Chúng ta có thể xem quá trình học của thuật toán GTB như minh họa trong Hình 1. Cây học sau được xây dựng nhằm mục đích ước lượng các giá trị của phần sai số của cây học ngay trước. Cuối cùng thì giá trị ước lượng sẽ là $\sum f_n(X)$. Mô hình cuối cùng sẽ có dạng:

$$f(x) = \sum_{i=1}^M \beta_i h(x; \theta_i) = f_{M-1} + \rho_M \beta_M h(x; \theta_M), \quad (1)$$

Trong đó x là mẫu và hàm $h(x; \theta_i)$ là cây ra quyết định thứ i . Các tham số khác được tính như sau:

$$\theta_M, \beta_M = \argmin_{\theta, \beta} \sum_{j=1}^N \left\| -\frac{\partial L(f_{M-1}(x_j), y_j)}{\partial f_{M-1}(x_j)} - \beta h(x_j; \theta) \right\|^2, \quad (2)$$

$$\rho_M = \argmin_{\rho} \sum_{j=1}^N L(f_{M-1}(x_j) + \rho h(x_j), y_j)$$



Hình 1. Mô hình thuật toán GTB

Để nâng cao hiệu suất làm việc của mô hình GTB, trong mô hình XGBoost một thành phần được thêm vào trong hàm mất mát (loss function) và đơn giản hóa việc tách nút trong cây. Trong GTB, chúng ta cần xác định độ dốc tương ứng với từng mẫu $\{(x_j, g_j)\}_{j=1}^N$, còn trong XGBoost ta cần xác định độ dốc và hệ số Hess tương ứng với từng mẫu dạng: $\{(x_j, g_j, h_j)\}_{j=1}^N$. Lúc này việc tách nút của cây XGBoost và giá trị tại các nút được tính dựa trên (g_j, h_j) như sau:

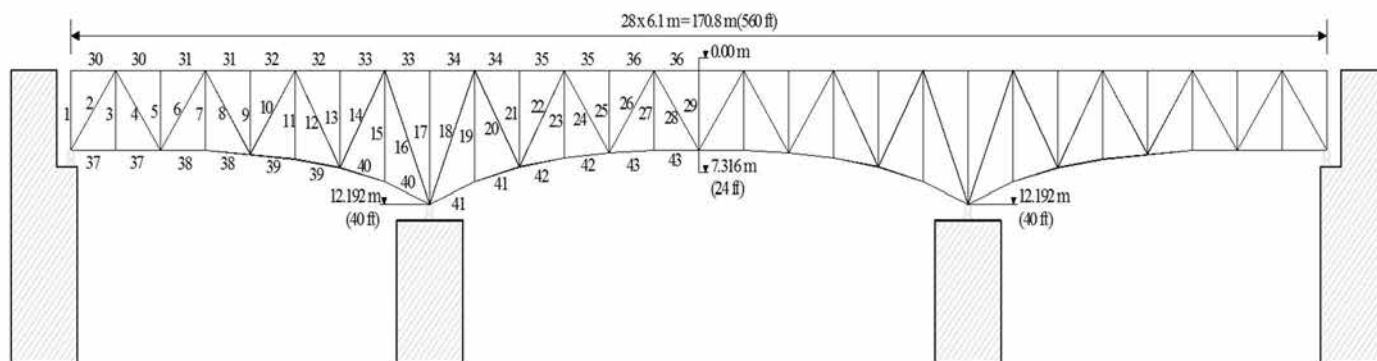
$$Gain = \frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma, \quad (3)$$

$$W_j = -\frac{G_j}{H_j + \lambda}$$

Trong đó γ và λ là 2 hệ số mô tả độ phức tạp của cây bằng cách điều chỉnh số nút lá và trọng số của tổng bình phương các giá trị nút lá. G_L và H_L là tổng của g_j và h_j của tất cả mẫu ở phía bên trái, G_R và H_R là tổng của g_j và h_j của tất cả mẫu ở phía bên phải. W_j là giá trị tại nút. G_j và H_j là tổng của g_j và h_j của tất cả các mẫu được chia trong khu vực của nút thứ j .

Mô hình LightGBM được tối ưu hóa hơn nữa trên cơ sở thuật toán XGBoost được mô tả ở trên. Những tối ưu hóa này được thực hiện để giảm thời gian tính toán, nhưng chúng cũng có thể đóng vai trò ngăn chặn hiện tượng mô hình quá khớp (vì dữ liệu gốc bị nhiễu, một số xử lý thô có thể làm tăng khả năng quá khớp của mô hình). Thời gian tính toán của mỗi lần tách nút được tính như sau:

$$cost_{time} = feature_{num} \times sample_{num} \times point_{num}, \quad (4)$$



Hình 2. Cấu giàn thép phẳng 113 thanh

Trong đó $point_{num}$ thể hiện thời gian (giây) cần thiết, $feature_{num}$, $sample_{num}$ và $point_{num}$ là số đặc điểm, số mẫu và số điểm sử dụng để phân chia cây. Như vậy, thời gian tính toán sẽ phụ thuộc vào 3 vấn đề chính là: (1) cỡ mẫu, (2) số đặc điểm và (3) số điểm tiềm năng dùng để phân chia cây. Trên cơ sở đó LightGBM sẽ tối thiểu hóa các chi phí này được trên 3 kỹ thuật cơ bản: GOSS (Gradient unilateral sampling) nhằm giảm số mẫu, (2) EFB (Feature binding technology) nhằm giảm số đặc tính của mẫu và (3) Hist (Histogram algorithm) nhằm giảm số điểm lựa chọn.

GOSS giữ tất cả các đối tượng có độ dốc lớn và thực hiện lấy mẫu ngẫu nhiên trên các đối tượng có độ dốc nhỏ. Để bù lại ảnh hưởng đối với phân phối dữ liệu, khi tính toán mức tăng thông tin, GOSS giới thiệu một hệ số nhân cho các trường hợp đối tượng có độ dốc nhỏ. Cụ thể, GOSS trước hết sắp xếp các thể hiện dữ liệu theo giá trị tuyệt đối của độ dốc và chọn các $a \times 100\%$ đối tượng hàng đầu. Sau đó, nó lấy mẫu ngẫu nhiên $b \times 100\%$ đối tượng trong phần còn lại của dữ liệu. Sau đó, GOSS khuếch đại dữ liệu được lấy mẫu với độ dốc nhỏ theo hằng số $(1-a)/b$ khi tính toán mức tăng thông tin. Bằng cách đó, các đối tượng chưa được đào tạo sẽ được tập trung nhiều hơn mà không làm thay đổi quá nhiều đặc điểm phân phối của dữ liệu gốc.

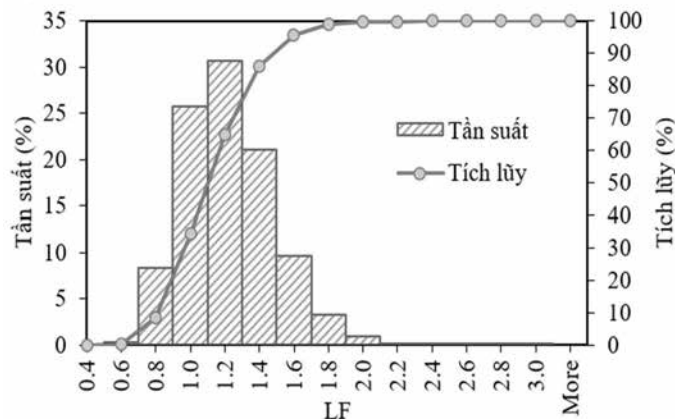
EFB được xây dựng dựa trên đặc điểm là các dữ liệu nhiều chiều thường phân bố rất thưa thớt theo các chiều. Sự thưa thớt của không gian đặc trưng cung cấp cho chúng ta khả năng thiết kế một cách tiếp cận gần như không mất dữ liệu để giảm số lượng đặc trưng. Cụ thể, trong một không gian đối tượng thưa thớt, nhiều đối tượng loại trừ lẫn nhau, nghĩa là chúng không bao giờ đồng thời nhận các giá trị khác không. Chúng ta có thể kết hợp các tính năng bị loại trừ một cách an toàn vào một tính năng duy nhất (gọi là gói tính năng bị loại trừ). Bằng thuật toán quét tính năng được thiết kế cẩn thận, chúng ta có thể tạo biểu đồ tính năng giống nhau từ các gói tính năng cũng như từ các tính năng riêng lẻ. Theo cách này, độ phức tạp của việc xây dựng biểu đồ thay đổi từ $O(\#data \times \#tổng\ đặc\ trưng)$ thành $O(\#data \times \# đặc\ trưng\ kết\ hợp)$, trong đó $\# đặc\ trưng\ kết\ hợp \ll \#tổng\ đặc\ trưng$. Điều này giúp tăng tốc đáng kể quá trình đào tạo cây ra quyết định mà không ảnh hưởng đến độ chính xác.

3. ÁP DỤNG TÍNH TOÁN

3.1 Mô hình cầu giàn thép phẳng

Trong phần này, một cầu giàn thép phẳng 113 thanh với kích thước và sơ đồ của cầu trong Hình 2 được nghiên cứu. Các thanh giàn được chia thành 43 nhóm tiết diện khác nhau có giá trị trong khoảng [3870.96, 22580.6] (mm²). Tổ hợp được xem xét là (1.25TT+1.75HT). TT và HT là tĩnh và hoạt tải tác dụng được lấy giá

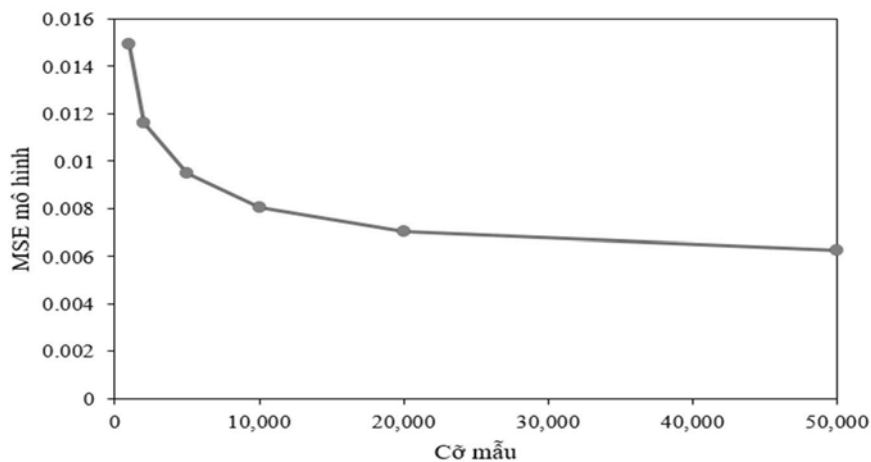
trị bằng 130 và 80 (kN) tại các nút giàn ở đường xe chạy phía trên. Vật liệu thép được sử dụng là thép A992 có $F_y = 344.7\text{ MPa}$ và $E = 200\text{ GPa}$. Sự hư hỏng của công trình được đánh giá thông qua việc so sánh giữa khả năng chịu tải cực hạn R của công trình và tải trọng tác dụng S. Để đơn giản hóa, ta đặt $LF=R/S$. Lúc này, công trình sẽ an toàn nếu $LF \geq 1.0$ và ngược lại. Do vậy, trong nghiên cứu này, thay vì xem xét ước lượng S, ta sẽ ước lượng LF. 50000 mẫu ngẫu nhiên của giàn thép được tạo ra để dùng làm cơ sở dữ liệu cho việc huấn luyện mô hình. Sự phân bố LF của 50000 mẫu được thể hiện trong Hình 3. Dựa trên hình 3 ta có thể thấy rằng tỉ lệ phá hủy của giàn thép là khoảng 34.33%.



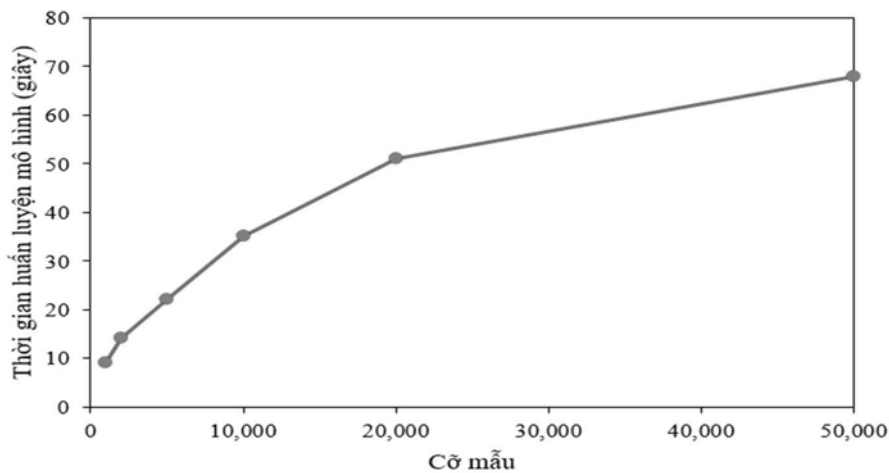
Hình 3. Phân bố của 50000 dữ liệu

3.2 Kết quả tính toán và thảo luận

Hình 4 thể hiện hiệu suất ước lượng giá trị LF của thuật toán LightGBM với kích cỡ của tập huấn luyện lần lượt là 1000, 2000, 5000, 10000, 20000 và 50000. Số mẫu cho tập kiểm tra được cố định là 5000 mẫu. Tương ứng với mỗi cỡ mẫu, mô hình huấn luyện được thực hiện độc lập 10 lần khác nhau để tìm ra giá trị trung bình MSE đối với tập kiểm tra thể hiện trong hình. Kết quả cho thấy MSE của mô hình khá nhỏ (< 0.01) khi cỡ mẫu từ 5000 trở lên. Điều này cho thấy thuật toán LightGBM đạt độ chính xác khá cao và độ chính xác tăng lên khi cỡ mẫu huấn luyện tăng lên. Bên cạnh đó, Hình 5 trình bày thời gian huấn luyện trung bình của thuật toán với các cỡ mẫu khác nhau. Ta có thể thấy rằng, thời gian sử dụng để huấn luyện của thuật toán LightGBM là khá ngắn. Cụ thể, với cỡ mẫu nhỏ (< 5000), thời gian huấn luyện của mô hình dưới 25 (giây). Khi cỡ mẫu lớn lên đến 20000 thì thời gian huấn luyện cũng chỉ là 51 (giây). Còn với cỡ mẫu rất lớn là 50000, thời gian huấn luyện của mô hình là 68 (giây). Thời gian huấn luyện này hoàn toàn hiệu quả cho các công việc thực tế.



Hình. 4. MSE của mô hình đối với tập kiểm tra



Hình. 5. MSE của mô hình đối với tập huấn luyện

4. KẾT LUẬN

Bài báo giới thiệu tổng quát thuật toán học máy hiện đại LightGBM. Hiệu suất của thuật toán được đánh giá thông qua ví dụ về cầu giàn thép phẳng 113 thanh với 43 biến thiết kế là diện tích tiết diện của thanh giàn. Biến đầu ra của mô hình huấn luyện là hệ số chịu tải của kết cấu được tính toán sử dụng phân tích trực tiếp nhằm xét đến tính chất phi tuyến tính phi đàn hồi của công trình. Mô hình LightGBM ước lượng giá trị LF với độ chính xác khá cao khi sai số MSE của mô hình đối với tập kiểm tra dưới 0.01 khi cỡ mẫu huấn luyện lớn hơn 5000. Bên cạnh đó, thời gian huấn luyện của mô hình LightGBM cũng khá ngắn khi chỉ tiêu tốn dưới 35 (giây) cho cỡ mẫu huấn luyện dưới 10000. Kết quả tính toán này cho thấy thuật toán LightGBM có độ chính xác cao và thời gian tính toán nhanh, hoàn toàn có thể ứng dụng vào các bài toán thiết kế kết cấu thép phi tuyến.

TÀI LIỆU THAM KHẢO

- [1] V.H. Truong, P.C. Nguyen, S.E. Kim. An efficient method for optimizing space steel frames with semi-rigid joints using practical advanced analysis and the micro-genetic algorithm. *Journal of Constructional Steel Research* 2017; 128: 416-427.
- [2] V.H. Truong, H.M. Hung, P.H. Anh, T.D. Hoc. Optimization of steel moment frames with panel-zone design using an adaptive differential evolution. *Journal of Science and Technology in Civil Engineering (STCE)-HUCE* 2020; 14(2): 65-75.
- [3] H.A. Pham, V.H. Truong, T.C. Vu. Fuzzy finite element analysis for free vibration response of functionally graded semi-rigid frame structures. *Applied Mathematical Modelling* 2020; 88:

852-869.

- [4] V.H. Truong, S.E. Kim. Reliability-based design optimization of nonlinear inelastic trusses using improved differential evolution algorithm. *Advances in Engineering Software* 2018; 121: 59-74.
- [5] V. H. Truong, S.E. Kim. An efficient method for reliability-based design optimization of nonlinear inelastic steel space frames. *Struct Multidisc Optim* 2017; 56: 331-351.
- [6] V.H. Truong, H.A. Pham, T.H. Van, S. Tangaramwong. Evaluation of machine learning models for load-carrying capacity assessment of semi-rigid steel structures. *Engineering Structures* 2022; 273: 115001.
- [7] AASHTO LRFD. Bridge design specifications. 4th Ed. 2012.
- [8] Y. Zhang, S. Hu, J. Wu, Y. Zhang, L. Chen. Multi-objective optimization of double suction centrifugal pump using Kriging metamodels. *Advances in Engineering Software* 2014; 74: 16-26.
- [9] V.H. Truong, G. Papazafeiropoulos, Q.V. Vu, V.T. Pham, Z. Kong. Predicting the patch load resistance of stiffened plate girders using machine learning algorithms. *Ocean Engineering* 2021; 240: 109886.
- [10] V.H. Truong, H.A. Pham. Support vector machine for regression of ultimate strength of trusses: A comparative study. *Engineering Journal* 2021; 25(7): 157-166.
- [11] B. Keshtegar, P. Hao, Y. Wang, Q. Hu. An adaptive response surface method and Gaussian global-best harmony search algorithm for optimization of aircraft stiffened panels. *Applied Soft Computing* 2018; 66: 196-207.
- [12] G. Ke và cộng sự. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA: 1-9.
- [13] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *Annals of Statistics* 2000; 28(2): 337-407.
- [14] J. Friedman. Greedy function approximation: A gradient boosting machine. *Ann. Statist.* 2001; 29(5) 1189-1232.