

Nghiên cứu và ứng dụng học máy dự báo hệ số cố kết của đất yếu tại một số khu vực ven biển Bắc Bộ

Nguyễn Đức Mạnh*, Hồ Sỹ An, Phạm Bá Khải,
Nguyễn Đình Trung, Lê Anh Đức

Trường Đại học Giao thông Vận tải

Ngày nhận bài 12/5/2020; ngày chuyển phản biện 15/5/2020; ngày nhận phản biện 22/6/2020; ngày chấp nhận đăng 1/7/2020

Tóm tắt:

Trong nghiên cứu này, hệ số cố kết (C_v) của đất yếu tại một số khu vực ven biển Quảng Ninh, Hải Phòng và Thái Bình được dự báo bằng phương pháp học máy - kỹ thuật trí tuệ nhân tạo thông qua 3 mô hình học máy: hồi quy véc tơ - Support Vector Regression (SVR); mạng thần kinh nhân tạo đa lớp tri giác - Artificial Neural Network Multilayer Perceptron (ANN MLP); hồi quy sườn bên - Ridge Regression (RR). Các mô hình này được xây dựng trên ngôn ngữ lập trình Python và thư viện hỗ trợ Scikit-learn. Số liệu sử dụng gồm 133 mẫu đất yếu thu thập từ các công trình thực tế, được lấy từ những độ sâu khác nhau (m) và được phân tích trong phòng thí nghiệm nhằm xác định các thông số: hàm lượng sét (%), hàm lượng bụi (%), giới hạn chảy (%), giới hạn dẻo (%), chỉ số dẻo (%), độ sệt, độ ẩm (%), khối lượng thể tích tự nhiên (g/cm^3), khối lượng thể tích khô (g/cm^3), khối lượng riêng (g/cm^3), độ rỗng (%), độ bão hòa (%), hệ số rỗng. Để dự báo C_v , 15 thông số đầu vào được phân tích tương quan. Sau khi loại bỏ các thông số không có quan hệ chặt chẽ với C_v , 6 thông số có quan hệ chặt chẽ được xác định gồm: giới hạn chảy, độ ẩm, khối lượng thể tích tự nhiên, khối lượng thể tích khô, độ rỗng, hệ số rỗng. Để dự báo và cho máy học, tiến hành xây dựng mô hình chung với 70% dữ liệu học và 30% dữ liệu kiểm tra. Hiệu suất của các mô hình được kiểm tra bằng các hệ số: giá trị trung bình của tổng các trị tuyệt đối của sai số - Mean Absolute Error (MAE); độ lệch tiêu chuẩn các sai số - Root Mean Square Error (RMSE); hệ số tương quan R - Correlation coefficient (R); và hệ số xác định - Coefficient of determination (R^2). Kết quả các mô hình nghiên cứu thể hiện hiệu suất của các mô hình học máy khác nhau với R^2 biến thiên từ 0,7899 đến 0,8737, đảm bảo quan hệ chặt. Nghiên cứu đã chỉ ra hiệu suất của mô hình ANN MLP là tốt nhất với kết quả các hệ số: $R^2=0,8737$, $MAE=0,2196$, $RMSE=0,2678$ và $R=0,9367$ tốt nhất trong 3 mô hình sử dụng.

Từ khóa: ANN MLP, đất yếu, hệ số cố kết, học máy, RR, SVR.

Chỉ số phân loại: 2.1

Đặt vấn đề

Hệ số cố kết (C_v) là một trong những thông số đất nền quan trọng dùng để dự báo lún cố kết nền đất yếu dưới tác dụng của tải trọng, đặc biệt cần thiết khi tính toán thiết kế xử lý nền đất yếu trong xây dựng giao thông [1, 2]. Nó tỷ lệ với tốc độ giảm của áp lực nước lỗ rỗng và phản ánh tốc độ lún cố kết của đất yếu [1, 2]. Thông thường, C_v được xác định trực tiếp bằng các thí nghiệm địa kỹ thuật trong phòng trên các mẫu đất được lấy từ các lỗ khoan thăm dò [3, 4]. Việc lấy và bảo quản mẫu đất trước khi thí nghiệm thường có nhiều khó khăn, dễ mất tính nguyên trạng, đặc biệt với các mẫu đất yếu khi lấy từ độ sâu lớn [5, 6]. Ảnh hưởng của các lớp cát hay cát pha mỏng nằm xen kẽ ở đất yếu càng làm cho việc xác định chính xác C_v trong phòng thí nghiệm thêm khó khăn và dễ sai số lớn [7]. Để thay thế cho các thí nghiệm trong phòng trên mẫu đất, các thí nghiệm hiện trường xác định thông qua tương quan thực nghiệm hay bán thực nghiệm đã được phát triển để dự đoán C_v [4, 8]. Những tương quan này cơ bản dựa trên

việc cải tiến phân tích hồi quy [9]. Tuy nhiên, phương pháp hồi quy có giới hạn như cấu trúc của mô hình dựa trên một hay một số phương trình tuyến tính hoặc phi tuyến [10, 11].

Trong cuộc cách mạng số ngày nay, học máy (Machine Learning) hay trí tuệ nhân tạo (Artificial Intelligent) đã được ứng dụng rộng rãi trong nhiều lĩnh vực bao gồm cả địa kỹ thuật, như để dự báo trượt lở đất [12-14], dự đoán lũ lụt [15], dự báo trữ lượng nước ngầm [16, 17] và dự đoán tính chất vật liệu đất [18-24]. Trong nghiên cứu này, thông số được lựa chọn dự đoán là C_v của đất yếu tại một số khu vực ven biển ở Quảng Ninh, Hải Phòng và Thái Bình bằng việc sử dụng ba phương pháp học máy hỗ trợ hồi quy véc tơ - Support Vector Regression (SVR); mạng thần kinh nhân tạo đa lớp tri giác - Artificial Neural Network Multilayer Perceptron (ANN MLP); hồi quy sườn bên - Ridge Regression (RR). Kết quả của các mô hình dự báo được kiểm tra bằng các phương pháp thống kê tiêu chuẩn như: giá trị trung bình của tổng các trị tuyệt đối của sai số - Mean Absolute Error (MAE); độ lệch tiêu

*Tác giả liên hệ: Email: nguyenducmanh@utc.edu.vn

Prediction of consolidation coefficient of soft soil using machine learning in some areas of North coast Vietnam

Duc Manh Nguyen*, Sy An Ho, Ba Khai Pham,
Dinh Trung Nguyen, Anh Duc Le

University of Transport and Communications

Received 12 May 2020; accepted 1 July 2020

Abstract:

The main object of this study is to accurately predict the consolidation coefficient (C_v) of soft soil in some areas of Quang Ninh, Hai Phong, and Thai Binh provinces using 3 machine learning methods: Support Vector Regression (SVR), Artificial Neural Network Multilayer Perceptron (ANN MLP), and Ridge Regression (RR). These methods are built on the Python programming language and the Scikit-learn library. The data for this study includes 133 soil samples that were collected from the construction site in North coast Vietnam. These samples at various depth (m) were analysed in the laboratory for the determination of clay content (%), silt content (%), liquid limit (%), plastic limit (%), plasticity index (%), liquidity index, moisture content (%), wet density (g/cm^3), dry density (g/cm^3), specific gravity (g/cm^3), porosity (%), degree of saturation (%), and void ratio. In addition, the authors have analysed the correlation of 15 parameters in predicting C_v . After removing parameters which have a weak correlation with C_v , there were 6 parameters that were strongly correlated with C_v including liquid limit, moisture content, wet density, dry density, porosity, and void ratio. For prediction C_v , the authors built a model with 70% of learning data and 30% of test data. The performance of the models was validated using Mean Absolute Error (MAE), Root Mean Square Error (RMSE) Correlation coefficient (R), and Coefficient of determination (R^2). Results of the models' study showed that the performance of the models using different methods is much different where R^2 -value varies from 0.7899 to 0.8737. The present study suggested that the ANN MLP model has the highest predictive capability in 3 models with the corresponding results of the coefficients: $R^2=0.8737$, $\text{MAE}=0.2196$, $\text{RMSE}=0.2678$, and $R=0.9367$.

Keywords: artificial neural network multilayer perceptron (ANN MLP), consolidation coefficient, machine learning, ridge regression (RR), soft soil, support vector regression (SVR).

Classification number: 2.1

chuẩn các sai số - Root Mean Square Error (RMSE); hệ số tương quan R - Correlation coefficient (R); và hệ số xác định - Coefficient of determination (R^2).

Dữ liệu và phương pháp sử dụng dự báo

Dữ liệu sử dụng

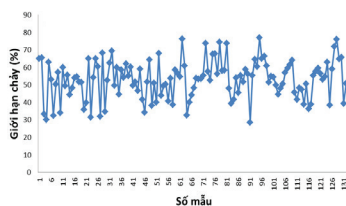
Trong nghiên cứu này, tổng cộng 133 mẫu đất yếu là bùn sét pha được thu thập từ các công trình thực tế tin cậy và một số mẫu thí nghiệm bổ sung kiểm chứng tại một số khu vực ven biển từ Quảng Ninh đến Nam Định, với đầy đủ các chỉ tiêu được xác định trong phòng thí nghiệm sử dụng làm cơ sở dữ liệu xây dựng mô hình dự báo. Các thông số của đất được xem xét sử dụng phân tích trong nghiên cứu này bao gồm: độ sâu lấy mẫu (m), hàm lượng hạt sét (%), hàm lượng hạt bụi (%), giới hạn chảy (%), giới hạn dẻo (%), chỉ số dẻo (%), độ sệt, độ ẩm (%), khối lượng thể tích tự nhiên (g/cm^3), khối lượng thể tích khô (g/cm^3), khối lượng riêng hạt (g/cm^3), độ rỗng (%), độ bão hòa (%), hệ số rỗng. Các thông số này được coi là các biến đầu vào độc lập tương ứng $X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}, X_{12}, X_{13}, X_{14}, X_{15}$ nhằm xác định biến đầu ra phụ thuộc C_v (Y). Dữ liệu thu thập được chia ngẫu nhiên thành hai phần là tập dữ liệu học (70%), dùng để tập luyện cho máy và tập dữ liệu kiểm tra (30%) nhằm đánh giá hiệu suất mô hình. Các mô hình này được xây dựng trên ngôn ngữ lập trình Python và thư viện hỗ trợ Scikit-learn.

Bảng 1. Dữ liệu đầu vào và đầu ra sử dụng trong nghiên cứu.

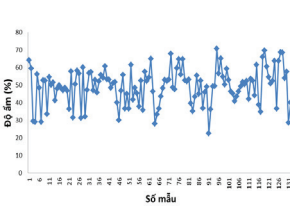
STT	Thông số	Giá trị lớn nhất	Giá trị nhỏ nhất	Trung bình	Độ lệch tiêu chuẩn
1	X1	35,40	1,2	11,04466	6,938565
2	X2	36,00	2	11,62752	6,949414
3	X3	53,00	4,5	27,28947	9,863118
4	X4	76,20	18	52,7218	12,09864
5	X5	77,00	28,57	53,09496	10,83319
6	X6	39,96	15,06	26,95699	4,659818
7	X7	47,15	9,40	26,13797	7,917158
8	X8	1,66	0,35	0,862932	0,200529
9	X9	70,65	22,55	49,19586	10,09728
10	X10	2,02	1,52	1,706767	0,088374
11	X11	1,65	0,91	1,152707	0,141143
12	X12	2,73	2,65	2,694361	0,017596
13	X13	66,30	38,88	57,21278	5,268267
14	X14	99,95	84,11	96,64316	2,856277
15	X15	1,967	0,636	1,370263	0,273962
16	Y	3,37	0,31	1,274286	0,713631

Phân tích phân phối 133 mẫu cho 15 thông số biến đầu vào (X_1 - X_{15}) và biến đầu ra (Y) cho thấy, các thông số biến thiên trong khoảng giá trị rất khác nhau (bảng 1 và hình đại diện 1, 2, 3, 4): độ sâu lấy mẫu trên (X_1) biến thiên từ 1,2 đến 35,4 m; độ sâu lấy mẫu dưới (X_2) biến thiên từ 2 đến 36 m; hàm lượng

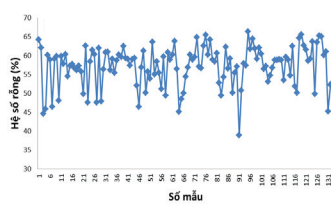
sét (X3) của các mẫu đất thì từ 4,5 đến 53%; hàm lượng bụi của các mẫu đất (X4) từ 18 đến 76,2%; giới hạn chảy (X5) từ 28,57 đến 77%; giới hạn dẻo (X6) từ 15,06 đến 39,96%; chỉ số dẻo (X7) từ 9,4 đến 47,15%; độ sét (X8) từ 0,35 đến 1,66; độ ẩm (X9) từ 22,55 đến 70,65%; khối lượng thể tích tự nhiên (X10) từ 1,52 đến 2,02 g/cm³; khối lượng thể tích khô (X11) từ 0,91 đến 1,65 g/cm³; khối lượng riêng hạt (X12) từ 2,65 đến 2,73 g/cm³; độ rỗng (X13) từ 38,88 đến 66,3%; độ bão hòa (X14) từ 84,11 đến 99,95%; hệ số rỗng (X15) từ 0,636 đến 1,967 và hệ số cố kết C_v (Y) biến thiên 0,31 đến 3,37 [25-30]. Độ lệch tiêu chuẩn lớn nhất thuộc về thông số hàm lượng hạt bụi (X4, độ lệch tiêu chuẩn 12,09864), nhỏ nhất khối lượng riêng hạt đất (X12, 0,017596).



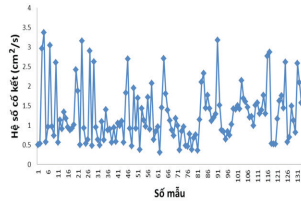
Hình 1. Biến thiên giá trị giới hạn chảy.



Hình 2. Biến thiên giá trị độ ẩm.



Hình 3. Biến thiên giá trị hệ số rỗng.



Hình 4. Biến thiên giá trị C_v từ thí nghiệm.

Công cụ và phương pháp sử dụng

Python là một ngôn ngữ lập trình thông dịch (interpreted), hướng đối tượng (object-oriented), và là một ngôn ngữ bậc cao (high-level) ngữ nghĩa động (dynamic semantics). Python hỗ trợ các module và gói (packages), khuyến khích chương trình module hóa và tái sử dụng mã. Trình thông dịch Python và thư viện chuẩn mở rộng có sẵn dưới dạng mã nguồn hoặc dạng nhị phân miễn phí cho tất cả các nền tảng chính và có thể được phân phối tự do [30]. Cùng với đó, Scikit-learn là một thư viện tài nguyên mở dùng cho học máy, hỗ trợ học máy có giám sát (supervised learning) và học máy không giám sát (unsupervised learning). Nó cũng cung cấp nhiều công cụ dùng cho làm khớp mô hình (model fitting), tiền xử lý dữ liệu (data preprocessing), lựa chọn mô hình và nhiều tiện ích khác [31-33].

Thuật toán Support Vector Regression (SVR)

Support Vector Regression (SVR) là một thuật toán nằm trong bộ thuật toán Support Vector Machine (SVM) dùng để giải quyết các vấn đề hồi quy [34-37]. Thay vì giảm thiểu lỗi trong quá trình huấn luyện, SVR cố gắng giảm thiểu lỗi tổng

quát bị ràng buộc để đạt được hiệu suất tổng thể. Ý tưởng về SVR dựa trên tính toán của hàm hồi quy tuyến tính, trong không gian đặc trưng chiều cao nơi dữ liệu đầu vào bằng hàm phi tuyến (non-linear function). SVR đã được áp dụng trong các lĩnh vực khác nhau như phân tích và dự đoán theo chuỗi thời gian và tài chính (lọc nhiễu và rủi ro), xấp xỉ các phân tích kỹ thuật phức tạp, lập trình và lựa chọn các hàm mất mát... SVR sử dụng các nguyên tắc tương tự cho phân loại (classification) và sử dụng thêm loại mới của hàm mất mát. Với một tập dữ liệu huấn luyện nhất định, được biểu thị trong một không gian vector, trong đó mỗi dữ liệu của mẫu là một điểm. Phương thức này là tốt nhất, tại đó có thể chia các điểm trong không gian thành hai lớp riêng biệt, tương ứng với (lớp) + và (lớp) - (phân loại nhị phân). Đặc trưng của siêu phẳng này được xác định bởi khoảng cách (được gọi là ranh giới) của điểm dữ liệu gần nhất của mỗi lớp với mặt phẳng này. Do đó, ranh giới càng rộng cho thấy mặt phẳng phân chia và phân loại càng chính xác. Mục tiêu của phương pháp SVR là tìm ra khoảng cách ranh giới tối đa. Trong nghiên cứu này, chúng tôi xác định các giá trị cho các tham số SVR thông qua quá trình thử - lỗi.

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 quy tuyến tính. Đặc điểm của SVR là cho ta một giải pháp thưa (sparse solution): nghĩa là để xây dựng được hàm hồi quy, ta không cần phải sử dụng hết toàn bộ tất cả các điểm dữ liệu trong bộ huấn luyện. Những điểm ở biên đóng góp vào việc xây dựng hàm hồi quy được gọi là Support Vector. Việc phân lớp cho tập dữ liệu mới sẽ chỉ phụ thuộc vào các Support Vector.

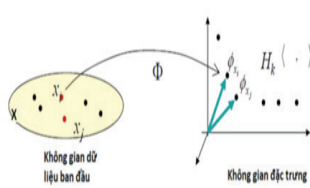
Dữ liệu học được đưa vào dạng $[(x_1, t_1), \dots, (x_n, t_n)] \subset R \times R$, trong đó n biểu thị không gian của các mẫu đầu vào. Dựa vào quan hệ giữa ε -SVR, mục tiêu là tìm được hàm f_x có độ lệch ε với mục tiêu t_j cho tất cả tập dữ liệu huấn luyện và đồng thời càng phẳng càng tốt. Do vậy hàm hồi quy cần có dạng: $y = f_x = w^T \Phi(x) + b$.

Trong đó: $w \in R^m$ là vector trọng số; T là ký 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 là hàm ánh xạ từ không gian đầu vào sang không gian đặc trưng (hình 5). Như vậy để tìm w và b ta phải tối thiểu hóa hàm lỗi chuẩn hóa: $\frac{1}{2} \sum_{i=1}^n \{y_i - t_i\} + \frac{\lambda}{2} \|w\|^2$ với λ là hằng số chuẩn hóa.

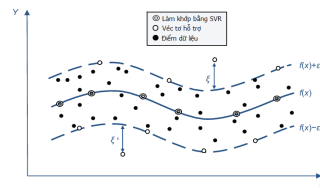
Để có được một giải pháp thưa, ta sẽ thay hàm lỗi trên bằng hàm lỗi ε -insensitive (hình 6).

Đặc điểm của hàm lỗi này là giá 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: $C \sum_{i=1}^n E_\varepsilon(y(x_i) - t_i) + \frac{1}{2} \|w\|^2$.

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 $\|w\|^2$.



Hình 5. Biến đổi không gian dữ liệu sang không gian đặc trưng (thủ thuật Kernel).

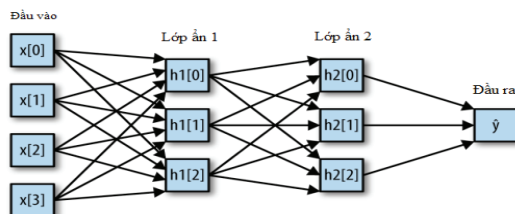


Hình 6. Sơ đồ nguyên lý thuật toán SVR sử dụng hàm lỗi ϵ -insensitive.

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 . 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 vector hỗ trợ, làm cho mô hình quá đơn giản. Ngược lại, nếu ϵ quá nhỏ thì có nhiều vector hỗ trợ, dẫn đến mô hình phức tạp, dễ bị quá khớp. Tham số phản ánh mối tương quan giữa các vector hỗ trợ nên cũng ảnh hưởng đến độ chính xác dự đoán của mô hình.

Mạng nơ ron nhân tạo (Artificial Neural Network Multilayer Perceptron - ANN MLP)

Mạng nơ ron nhân tạo (ANN) là một tập hợp của các nút được liên kết với nhau nhằm giải thích và giải quyết các vấn đề có mối quan hệ phức tạp giữa các biến độc lập đầu vào và biến phụ thuộc đầu ra [35]. Mạng perceptron đa lớp (MLP) là một trong những mạng nơ ron nhân tạo, hay một trong những kỹ thuật mạng nơ ron nhân tạo hiệu quả nhất trong mô hình và dự đoán. Do đó nó được sử dụng như một mô hình chuẩn của nhiều nhà nghiên cứu [36]. ANN MLP có khả năng mô phỏng các quá trình phi tuyến và phức tạp của thế giới thực. Trong nghiên cứu này, ANN MLP được sử dụng để phân tích và dự đoán hệ số cố kết C_v . Cơ bản, ANN MLP là một mô hình chuyển tiếp bao gồm một lớp đầu vào, một hay nhiều lớp ẩn và một lớp đầu ra như trên hình 7.



Hình 7. Minh họa một ANN MLP với 2 lớp ẩn.

Nói chung, số lượng của các nút đầu vào phụ thuộc vào các thông số được lựa chọn trong nguồn dữ liệu sử dụng, và số lượng nơ ron ẩn được xác định dựa trên tập dữ liệu dùng cho học máy trong từng trường hợp cụ thể. Số lượng lớp ẩn

được sử dụng cho tính toán và lớp đầu ra đại diện cho mục đích của mô hình. Mỗi nút của lớp ẩn phải kết nối với tất cả các nút của lớp đầu vào, và mỗi nút của lớp đầu ra phải được kết nối với toàn bộ nút trong các lớp ẩn. Thông qua các liên kết này, quá trình hoạt động của ANN MLP có thể chia ra hai bước: truyền thẳng và truyền ngược sử dụng thuật toán truyền ngược.

Đối với mục đích mô hình hóa hàm với một biến dự đoán, ANN MLP được sử dụng để tổng quát hóa một hàm phi tuyến $f: X \in R^D \rightarrow Y \in R^1$. Hàm f được thể hiện ngắn gọn thông qua phương trình sau [35]: $f(X) = b_2 + W_2 \times (f_A(b_1 + W_1 \times I))$.

Với $W1$ và $W2$ lần lượt là các ma trận trọng số của lớp ẩn và lớp đầu ra $b_1 = [b_{11}, b_{12}, \dots, b_{1N}]$; biểu thị vec tơ “bias” của lớp ẩn, b_2 là vec tơ “bias” của lớp đầu ra; f_A biểu thị hàm kích hoạt. Một trong những hàm kích hoạt phổ biến được sử dụng hiện nay là Sigmoid, Tanh và Relu.

Trong nghiên cứu này, chúng tôi sử dụng mạng nơ ron nhân tạo ANN MLP có 3 lớp ẩn với số nút lần lượt trong mỗi lớp ẩn là 16, 10, 14 và sử dụng hàm kích hoạt Sigmoid.

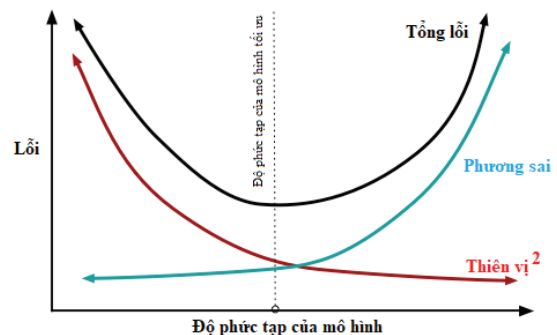
Ridge Regression (RR)

Mô hình Ridge Regression (RR) là phương pháp áp dụng khi bộ dữ liệu gặp vấn đề về đa cộng tuyến các biến độc lập x có mối liên hệ với nhau, và ảnh hưởng lên kết quả dự báo của y , hay giải quyết các vấn đề về Overfitting (mô hình áp dụng tốt cho dữ liệu học (training) nhưng không hoạt động tốt trên dữ liệu kiểm tra (test) mà mô hình hồi quy tuyến tính thông thường gặp phải) [36].

Phương trình tổng quát của linear regression:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \epsilon$$

Ta có thể thấy giá trị ϵ ở cuối phương trình. Đây là sai số của các phương trình hồi quy, là chênh lệch giữa kết quả dự báo và kết quả thực tế. Các sai số được chia thành 2 phần: Biased (thiên lệch), Variance (phương sai). Biased là trường hợp mô hình phân tích không khớp, không đem lại kết quả chính xác trên tập dữ liệu học tập (training). Variance là đối với dữ liệu thử (test). Mối quan hệ đánh đổi giữa Biased và Variance xét trên mức độ phức tạp của mô hình được minh họa ở hình 8.



Hình 8. Mối quan hệ đánh đổi giữa bias và variance.

dự đoán. Nó được xác định thông qua công thức:

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^m (y_{test,i} - y_{pred,i})^2}$$

trong đó m là số lượng mẫu kiểm tra, y_{test} là giá trị đúng đã xác định của y , y_{pred} là giá trị do mô hình dự đoán.

MAE là giá trị trung bình của tổng các trị tuyệt đối của sai số mà mô hình dự đoán. Nó được xác định thông qua công thức:

$$MAE = \frac{1}{m} \sum_{i=1}^m |y_{test,i} - y_{pred,i}|$$

trong đó, m là số lượng mẫu kiểm tra, y_{test} là giá trị đúng đã xác định của y , y_{pred} là giá trị do mô hình dự đoán.

Hệ số xác định R^2 là một thông số biến đổi từ 0 đến 1 nhằm thể hiện hiệu suất của mô hình dự đoán. Nó có thể được xác định thông qua công thức:

$$R^2 = 1 - \frac{\sum_{i=1}^m (y_{test,i} - y_{pred,i})^2}{\sum_{i=1}^m (y_{test,i} - \bar{y}_{test})^2}$$

trong đó m là số lượng mẫu kiểm tra, y_{test} là giá trị đúng đã xác định của y , y_{pred} là giá trị do mô hình dự đoán, $\bar{y}_{test}$ là giá trị trung bình của các mẫu kiểm tra.

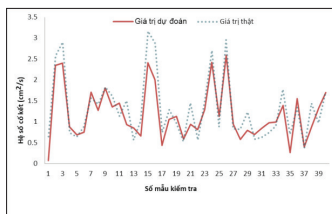
Hệ số tương quan R đánh giá tương quan giữa hai biến số trong quan hệ. Nó được thể hiện thông qua công thức sau:

$$R = \frac{\sum_{i=1}^m (y_{pred,i} - \bar{y}_{pred})(y_{test,i} - \bar{y}_{test})}{\sqrt{\sum_{i=1}^m (y_{pred,i} - \bar{y}_{pred})^2} \cdot \sqrt{\sum_{i=1}^m (y_{test,i} - \bar{y}_{test})^2}}$$

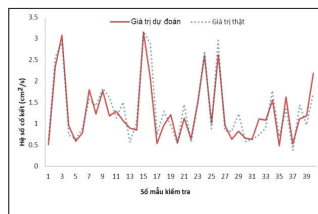
trong đó m là số lượng mẫu kiểm tra, y_{test} là giá trị đúng của y , y_{pred} là giá trị do mô hình dự đoán, $\bar{y}_{test}$ là giá trị trung bình của các mẫu kiểm tra, $\bar{y}_{pred}$ là giá trị trung bình của các mẫu dự đoán.

Kết quả dự báo hệ số cố kết và phân tích

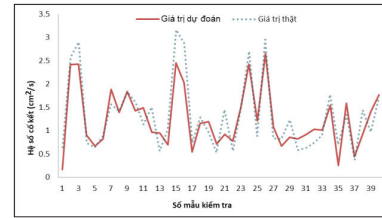
Sử dụng ba mô hình SVR, ANN MLP và RR như đã nêu trên, với bộ số liệu sau chuẩn hóa gồm 6 biến độc lập (X5, X9, X10, X13, X11, X15) có tương quan chặt và tiến hành dự báo theo sơ đồ mô hình học máy đã xây dựng như trình bày trên hình 12, kết quả dự báo hệ số cố kết C_v được so sánh với giá trị thực tế thí nghiệm (hình 13, 14 và 15).



Hình 13. Giá trị C_v dự đoán và giá trị thật khi sử dụng mô hình SVR.



Hình 14. Giá trị C_v dự đoán và giá trị thật sử dụng mô hình ANN MLP.



Hình 15. Giá trị C_v dự đoán và giá trị thật sử dụng mô hình RR.

Kết quả đánh giá hiệu suất các mô hình sử dụng trong nghiên cứu này thu được giá trị tốt với tất cả các mô hình dự báo sử dụng, cụ thể: khi dự báo C_v bằng mô hình SVR thì $RMSE=0,345$, $MAE=0,285$, $R^2=0,7899$, $R=0,9102$; với mô hình ANN MLP thì $RMSE=0,2678$, $MAE=0,2196$, $R^2=0,8737$, $R=0,9367$; còn mô hình RR có $RMSE=0,3359$, $MAE=0,2791$, $R^2=0,8011$, $R=0,9025$ (bảng 2).

Bảng 2. So sánh hiệu suất các mô hình.

STT	Mô hình	RMSE	MAE	R	R^2
1	SVR	0,345	0,285	0,9102	0,7899
2	ANN MLP	0,2678	0,2196	0,9367	0,8737
3	RR	0,3359	0,2791	0,9025	0,8011

Kết quả đánh giá hiệu suất dự báo của các mô hình sử dụng trong nghiên cứu này $RMSE$, MAE , R hay R^2 đều cho kết quả tốt, chứng tỏ độ tin cậy cao khi dự báo C_v thông qua một số chỉ tiêu vật lý dễ xác định khác đã có. Riêng với hệ số xác định R^2 , kết quả thay đổi từ 0,7899 đến 0,8737. Kết quả này cho thấy rằng, độ chính xác của cả 3 mô hình đã sử dụng trong dự đoán C_v của đất yếu khu vực ven biển Bắc Bộ khác nhau không lớn, và mô hình SVR có độ chính xác thấp nhất trong nghiên cứu này ($R^2=0,7899$), trong khi đó mô hình ANN MLP đạt hiệu suất cao nhất trong, dự đoán C_v ($R^2=0,8737$).

Kết luận

Hệ số cố kết (C_v) của đất yếu là thông số đất nền không thể thiếu, sử dụng trong phân tích lún khi đắp nền trên đất yếu. Nó thường được xác định bằng thí nghiệm nén cố kết mẫu đất trong phòng với chi phí khá tốn kém và phức tạp. Ba mô hình SVR, ANN MLP và RR thuộc kỹ thuật học máy hoàn toàn có thể sử dụng để dự báo tốt hệ số cố kết của đất yếu thông qua một số chỉ tiêu vật lý dễ xác định khác.

Mô hình ANN MLP có khả năng dự báo hệ số cố kết của đất yếu khu vực ven biển Quảng Ninh, Hải Phòng, Thái Bình là tốt nhất ($R^2=0,8737$), trong khi các mô hình còn lại cũng thể hiện khả năng dự đoán khá tốt.

Ứng dụng của phương pháp học máy có thể xem như một công cụ thay thế hiệu quả và đầy triển vọng nhằm giảm thời gian, chi phí và sai sót do phải thí nghiệm xác định C_v trên từng mẫu đất. Dựa trên kết quả của nghiên cứu này, có thể đề xuất rằng, cả 3 phương pháp học máy SVR, ANN MLP và RR là công cụ hữu dụng trong dự đoán C_v của đất yếu, tương ứng là 6 thông số vật lý quan trọng để dự đoán

C_v gồm: giới hạn chảy, độ ẩm, khối lượng thể tích tự nhiên, khối lượng thể tích khô, độ rỗng và hệ số rỗng.

LỜI CẢM ƠN

Nghiên cứu này được thực hiện với nguồn kinh phí được cấp từ Bộ Giáo dục và Đào tạo thông qua đề tài mã số B2020-GHA-03 do Trường Đại học Giao thông Vận tải chủ trì. Các tác giả xin trân trọng cảm ơn sự hỗ trợ của Vụ KHCN&MT (Bộ Giáo dục và Đào tạo), Trường Đại học Giao thông Vận tải và các đơn vị đã hỗ trợ số liệu phục vụ cho nghiên cứu.

TÀI LIỆU THAM KHẢO

- [1] K.H. Andersen, and K. Schjetne (2012), "Database of friction angles of sand and consolidation characteristics of sand, silt, and clay", *J. Geotech. Geoenviron. Eng.*, **139**, pp.1140-1155.
- [2] E. Conte, and A. Troncone (2006), "One-dimensional consolidation under general time-dependent loading", *Can. Geotech. J.*, **43**, pp.1107-1116.
- [3] T. Moriwaki, and K. Umehara (2003), "Method for determining the coefficient of permeability of clays", *Geotech. Test. J.*, **26**, pp.47-56.
- [4] A. Sridharan, and H. Nagaraj (2004), "Coefficient of consolidation and its correlation with index properties of remolded soils", *Geotech. Test. J.*, **27**, pp. 469-474.
- [5] M.S. Al-Zoubi (2008), "Coefficient of consolidation by the slope method", *Geotech. Test. J.*, **31**, pp.526-530.
- [6] A. Muthumani, L. Fay, M. Akin, S. Wang, J. Gong, and X. Shi (2014), "Correlating lab and field tests for evaluation of deicing and anti-icing chemicals: A review of potential approaches", *Cold Reg. Sci. Technol.*, **97**, pp.21-32.
- [7] C-Y. Yune, and C-K. Chung (2005), "Consolidation test at constant rate of strain for radial drainage", *Geotech. Test. J.*, **28**, pp.71-78.
- [8] C. Pistor, M. Yardimci, and S. Güçeri (1999), "On-line consolidation of thermoplastic composites using laser scanning", *Compos., Part A Appl. Sci. Manuf.*, **30**, pp.1149-1157.
- [9] I. Rizzo, G. Vedoya, S. Maurutto, M. Haidukowski, and E. Varsavsky (2004), "Assessment of toxigenic fungi on Argentinean medicinal herbs", *Microbiol. Res.*, **159**(2), pp.113-120.
- [10] M. Kanayama, A. Rohe, and L.A. van Paassen (2014), "Using and improving neural network models for ground settlement prediction", *Geotech. Geol. Eng.*, **32**, pp.687-697.
- [11] P. Psyllaki, K. Stamatou, I. Iliadis, A. Mourlas, P. Asteris, and N. Vaxevanidis (2018), *Surface treatment of tool steels against galling failure*, <http://dx.doi.org/10.1051/mateconf/201818804024>.
- [12] B.T. Pham (2018), "A novel classifier based on composite hyper-cubes on iterated random projections for assessment of landslide susceptibility", *J. Geol. Soc. India*, **91**, pp.355-362.
- [13] Q. He, Z. Xu, S. Li, R. Li, S. Zhang, and N. Wang (2019), "Novel entropy and rotation forest-based credal decision tree classifier for landslide susceptibility modeling", *Entropy (Basel)*, **21**, p.106, <http://dx.doi.org/10.3390/e21020106>.
- [14] J. Dou, A.P. Yunus, Y. Xu, Z. Zhu, C-W. Chen, and M. Sahana (2019), "Torrential rainfall-triggered shallow landslide characteristics and susceptibility assessment using ensemble data-driven models in the Dongjiang Reservoir Watershed, China", *Nat. Hazards*, **3**, pp.1-31, <http://dx.doi.org/10.1007/s11069-019-03659-4>.
- [15] K. Khosravi, B.T. Pham, K. Chapi, A. Shirzadi, H. Shahabi, I. Revhaug, I. Prakash, and D. Tien Bui (2018), "A comparative assessment of decision trees algorithms for flash flood susceptibility modeling at Haraz watershed, northern Iran", *Sci. Total Environ.*, **627**, pp.744-755, <http://dx.doi.org/10.1016/j.scitotenv.2018.01.266> [PMID: 29426199].
- [16] B.T. Pham, A. Jaafari, I. Prakash, S.K. Singh, N.K. Quoc, and D.T. Bui (2019), "Hybrid computational intelligence models for groundwater potential mapping", *Catena*, **182**, p.104101, <http://dx.doi.org/10.1016/j.catena.2019.104101>.

- [17] K. Khosravi, M. Sartaj, F.T-C. Tsai, V.P. Singh, N. Kazakis, A.M. Melesse, I. Prakash, D. Tien Bui, and B.T. Pham (2018), "A comparison study of DRASTIC methods with various objective methods for groundwater vulnerability assessment", *Sci. Total Environ.*, **642**, pp.1032-1049, <http://dx.doi.org/10.1016/j.scitotenv.2018.06.130> [PMID: 30045486].
- [18] D.C. Camilo, L. Lombardo, P.M. Mai, J. Dou, and R. Huser (2017), "Handling high predictor dimensionality in slope-unit-based landslide susceptibility models through LASSO-penalized Generalized Linear Model", *Environ. Model. Softw.*, **97**, pp. 145-156, <http://dx.doi.org/10.1016/j.envsoft.2017.08.003>.
- [19] B.T. Pham, T-A. Hoang, D-M. Nguyen, and D.T. Bui (2018), "Prediction of shear strength of soft soil using machine learning methods", *Catena*, **166**, pp.181-191, <http://dx.doi.org/10.1016/j.catena.2018.04.004>.
- [20] P.G. Asteris, and M. Nikoo (2019), "Artificial bee colony-based neural network for the prediction of the fundamental period of infilled frame structures", *Neural Comput. Appl.*, **3**, pp.1-11, <http://dx.doi.org/10.1007/s00521-018-03965-1>.
- [21] P.G. Asteris, A. Moropoulou, A.D. Skentou, M. Apostolopoulou, A. Mohebbkhan, and L. Cavaleri (2019), "Stochastic vulnerability assessment of masonry structures: concepts, modeling and restoration aspects", *Appl. Sci. (Basel)*, **9**, p.243, <http://dx.doi.org/10.3390/app9020243>.
- [22] H. Chen, P.G. Asteris, D. Jahed Armaghani, B. Gordan, and B.T. Pham (2019), "Assessing dynamic conditions of the retaining wall: Developing two hybrid intelligent models", *Appl. Sci. (Basel)*, **9**, p.1042, <http://dx.doi.org/10.3390/app9061042>.
- [23] A. Ashrafiyan, and R.-B. Mohammad (2019), *Prediction of the compressive strength of self-compacting concrete using surrogate models*.
- [24] V. Rodriguez-Galiano, M. Sanchez-Castillo, M. Chica-Olmo, and M. Chica-Rivas (2015), "Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines", *Ore Geol. Rev.*, **71**, pp.804-818, <http://dx.doi.org/10.1016/j.oregeorev.2015.01.001>.
- [25] B.M. Das (2007), *Principles of geotechnical engineering*, Thomson.
- [26] B. Sharma, and P.K. Bora (2003), "Plastic limit, liquid limit and undrained shear strength of soil reappraisal", *J. Geotech. Geoenviron. Eng.*, **129**, pp.774-777.
- [27] B.M. Das, and K. Sobhan (2012), *Principles of geotechnical engineering*, Cengage learning.
- [28] Manh Duc Nguyen, Binh Thai Pham, Tran Thi Tuyen, Hoang Phan Hai Yen, Indra Prakash, Thanh Tien Vu, Kamran Chapi, Ataollah Shirzadi, Himan Shahabi, Jie Dou, Nguyen Kim Quoc and Dieu Tien Bui (2019), "Development of an artificial intelligence approach for prediction of consolidation coefficient of soft soil: a sensitivity analysis", *The Open Construction and Building Technology Journal*, **13**, <http://dx.doi.org/10.2174/1874836801913010178>.
- [29] J. A. Knappett and R.F. Craig (2012), *Craig's Soil Mechanics*.
- [30] <https://www.python.org/doc/essays/blurp/>.
- [31] https://scikit-learn.org/stable/getting_started.html.
- [32] Sebastian Raschka/Vahid Mirjalili (2017), *Machine learning and deep learning with Python*, Kindle Edition.
- [33] Michael Bowles (2015), *Machine learning in Python*, John Wiley & Sons, Inc.
- [34] Binh Thai Pham, Le Hoang Son, Tuan-Anh Hoang, Duc-Manh Nguyen, Dieu Tien Bui (2018), "Prediction of shear strength of soft soil using machine learning methods", *CATENA*, **166**, pp.181-191, <https://doi.org/10.1016/j.catena.2018.04.004>.
- [35] Binh Thai Pham, Manh Duc Nguyen, Kien-Trinh Thi Bui, Indra Prakashd, Kamran Chapie, Dieu Tien Bui (2019), "A novel artificial intelligence approach based on multi-layer perceptron neural network and biogeography-based optimization for predicting coefficient of consolidation of soil", *CATENA*, **173**, pp.302-311, <https://doi.org/10.1016/j.catena.2018.10.004>.
- [36] <https://bigdatauni.com/vi/tin-tuc/tong-quan-ve-regression-phan-tich-hoi-quy.html>.
- [37] Debasish Basak (2007), *Srimanta Pal and Dipak Chandra Patranabis*, Support vector regression.