Accuracy assessment of extreme learning machine in predicting soil compression coefficient

Hai-Bang Ly1, Panagiotis G. Asteris², Binh Thai Pham1*

1University of Transport Technology, Hanoi 100000, Vietnam
2Computational Mechanics Laboratory, School of Pedagogical and Technological Education, Heraklion, GR 14121, Athens, Greece

Received 7 February 2020; Received in revised form 18 April 2020; Accepted 25 June 2020

ABSTRACT

The compression coefficient (Cc) is an important soil mechanical parameter that represents soil compressibility in the process of consolidation. In this study, a machine learning derived model, namely extreme learning algorithm (ELM), was used to predict the Cc of soil. A total of 189 experimental results were used and randomly divided to construct the training and testing parts for the development and validation of ELM. Monte Carlo approach was applied to take into account the random sampling of samples constituting the training dataset. A number of 13 input parameters reflecting the experiment were used as the input variables to predict the output Cc. Several statistical criteria, such as mean absolute error (MAE), root mean square error (RMSE), correlation coefficient (R) and the Monte Carlo convergence estimator were used to assess the performance of ELM in predicting the Cc of soil. The results showed that ELM had a strong capacity to predict the Cc of soil, with the R value > 0.95. The convergence of results, as well as the capability of ELM were fully investigated to understand the advantage of using ELM as a predictor.

Keywords: Compression coefficient; extreme machine learning; Monte Carlo simulations.

©2020 Vietnam Academy of Science and Technology

1. Introduction

Machine learning, which is a part of the artificial intelligence approach, has been effectively applied in various fields to solve a lot of problems in the real world (Armaghani et al., 2016; Collins and Moons, 2019; Ly et al., 2019c; Michie et al., 1994; Mohamad et al., 2018, 2016; Thanh et al., 2020). This approach is based on the computational mathematic algorithm to analyze the relationship between input and output variables; therefore, it can solve a complex problem with large data (Dao et al., 2020a; Khandelwal et al., 2018; Khosravi et al., 2019; Le et al., 2020; Ly et al., 2019b; Pham et al., 2020b, 2020a). Compared with traditional approaches like linear regression or weighted methods, machine learning is known as more subjective and effective approaches for better performance (Chen et al., 2019; Ly et al., 2019c).
In geotechnical engineering, machine learning has also been applied effectively in solving both classification and regression problems. Dao et al. (Dao et al., 2020b) developed and compared deep learning and various machine learning models (quadratic discriminant analysis, Fisher's linear discriminant analysis, and multi-layer perceptron neural network) to solve a classification problem of spatial landslide prediction, and stated that machine learning and deep learning are potential approaches for quick and accurate prediction of landslides. In another work, machine learning models have been developed and applied to predict important properties of soil materials as regression problems such as soil compression coefficient using HHO-ANN and GOA-ANN ensembles (Moayedi et al., 2020a) and LCA-ANFIS hybrid model (Moayedi et al., 2020b), soil aggregate stability (Rivera and Bonilla, 2020), soil shear strength (Bui et al., 2019), soil consolidation coefficient (Pham et al., 2019a). In general, these mentioned studies show a great potential application of machine learning models for the prediction of geotechnical problems, including the prediction of soil properties.

In this study, the main aim is to use Extreme Learning Machine (ELM) for the prediction of an important property of soil, namely compression coefficient (Cc), which can help in quick and accurate determination of this parameter for reducing the cost and time for construction. ELM is a well-known machine learning model that has been applied effectively in solving real-world problems (Huang et al., 2006; Nizar et al., 2008), but its application is still limited in geotechnical problems. For this, a total of 189 experimental results were used and randomly divided to construct the datasets for the training and testing of the model. Monte Carlo approach was applied to evaluate the effect of the random sampling strategy of samples on the performance of the prediction model. Various statistical criteria, such as mean absolute error, root mean square error, the correlation coefficient (R) and the Monte Carlo convergence estimator were used to validate the predictive capability of the model. Matlab software and application was used for data processing and modeling.

2. Data collection and preparation

In this study, a number of 189 experimental data were collected from the national highway project in Hai Phong and Ninh Binh (Pham et al., 2019b). The samples were composed of soft clayed soil, and the tests were performed in the laboratory to determine 13 input variables and one output variable. The input variables contained the depth of sample, clay content in soil, moisture content, bulk, and dry soil densities, specific gravity, void ratio, porosity, saturation degree, liquid and plastic limits, plasticity, and liquidity indexes. The target variable was the Compression Coefficient of soil (denoted as Cc), which reflects the compressibility of soil in the consolidation process. It is considered as one of the important soil physicomechanical parameters of soil usually utilized in the determination of the primary consolidation settlement (Das and Sobhan, 2013). For the description of the input and output variables, the readers could refer to the work of Das and Sobhan (Das and Sobhan, 2013). The statistical analysis of the dataset used in this study can be found in our previous work (Pham et al., 2019b).

In this work, the dataset was divided into two parts: the training and testing one. The training dataset contained 132 samples (70%) and used for the training phase of ELM algorithm, whereas the testing dataset contained 57 samples (30%), served to verify the performance of ELM model. In addition, the whole dataset was scaled in the range of [0, 1], as suggested in many machine learning
problems to reduce numerical bias. Eq. (1) was used to scale the dataset:

\[ \chi_{\text{scaled}} = \frac{2(\chi - \overline{\chi})}{\mu - \chi} - 1 \quad (1) \]

where \( \lambda \) and \( \mu \) are the minimum and maximum values of given variables, and \( \chi \) is the value of the variable to be scaled.

3. Methods used

3.1. Extreme learning machine (ELM)

Extreme learning machine (ELM) algorithm is basically a feedforward neural network and often used for classification, regression, clustering, sparse approximation, compression, and feature learning with a single layer or multiple layers of hidden nodes. The basis of ELM lied in the generation of hidden nodes, which can be randomly assigned, and no update is performed during the simulation process. Otherwise, the parameters of hidden nodes (weights that connect input variables to the hidden nodes and their biases) were not tuned or can be inherited from the ancestors in several cases. According to the creators, ELM can produce a good performance and the ability to learn much faster than backpropagation networks. For the detailed development of ELM algorithm, the readers are referred to the literature (Huang et al., 2011, 2006, 2004). To validate the performance of ELM model, various statistical indexes such as mean absolute error (MAE), root mean square error (RMSE), and correlation coefficient (R) were used. Detail description of these indexes is presented in previously published works (Ly et al., 2019a; Mohamad et al., 2012; Moemeni et al., 2015; Nguyen et al., 2019; Pham et al., 2019b).

3.2. Monte Carlo approach

Monte Carlo (MC) method is widely applied in problems related to civil engineering to take into account the variability of parameters (Dao et al., 2020a; Ly et al., 2019d; Qi et al., 2020). The main idea of MC lies in the computation of the target using as many realizations as possible, which repeats the random sampling of input variables (Dao et al., 2020c; Nguyen et al., 2020; Pham et al., 2020c). The concept of using the MC method consists in propagating the variability of input parameters on the outputs using a predefined model and using statistical analysis of the output results to investigate the sensitivity or robustness of such a predefined model. By doing so, the input space reflects the random combinations of variables, where each variable obeys a probability density function, represented by its own variability. With the variability of the input space, the obtained outputs possess a distribution of response and the corresponding statistical properties. Statistical analysis needs to be performed to assess the accuracy of the model or the sensitivity of inputs. The higher the number of combinations in MC simulation, the more reliable the outputs. Nevertheless, to achieve a reasonable time consuming and reach the appropriate number of MC simulations, statistical convergence analysis of the mean value of a random variable is defined by the following equation:

\[ \text{NMC} \Rightarrow \text{Conv.(R|NMC)} = \frac{1}{\overline{R}} \frac{1}{\text{NMC}} \sum_{j=1}^{\text{NMC}} R_j \quad (2) \]

where \( \overline{R} \) is the mean value of the random variable \( R \), NMC is the number of Monte Carlo simulations, \( R_j \) is the value of the \( j \)th observation of \( R \).

4. Results and Discussion

4.1. Convergence of results

The prediction convergence results in the function of RMSE, MAE, and \( R \) are plotted in Fig. 1. A fluctuation up to 15% around the average values of RMSE and MAE was
observed, whereas that of the case of R was smaller (i.e., <0.5%). The predicted output stabilized in the 5% range after about 20 Monte Carlo simulations and smaller than 3% when NMC was higher than 300. Considering the R criterion, the NMC required for the stabilization of results below 1% was 20, whereas to obtain smaller fluctuation (i.e., <0.5%), a number of 300 NMC was needed. This result showed that ELM somehow outperformed support vector machine (SVM) algorithm in predicting the Cc of soft soil, as the NMC required to achieve a 1% converged result was 20, compared with 50 in the literature (Pham et al., 2019b).

4.2. Prediction capability

The probability density distributions of RMSE, MAE, and R values are presented in Fig. 2. It is observed that all distributions were asymmetric. The highest probability values of RMSE, MAE, and R were observed at 0.1467, 0.1166, and 0.8873, respectively. With respect to R criterion, values between R = 0.4 to R = 0.7 were noticed (Fig. 3c), showing that ELM exhibited a strong fluctuation in predicting Cc of soft soil. Similar observations were noticed with RMSE and MAE (Fig. 3a, b). The average values of RMSE, MAE, and R over 1000 simulations were 0.1743, 0.1450, and 0.8568, respectively. In conclusion, from overall statistical analysis, ELM algorithm was an excellent predictor to predict Cc. As in ELM, the weights and biases were randomly generated, an optimization algorithm might be needed to stabilize and enhance the prediction performance.

Figure 1. Normalized statistical convergence over 1000 NMC using ELM for: (a) RMSE; (b) MAE; (c) R

Figure 2. Probability distribution over 1000 NMC using ELM for the prediction of Cc for: (a) RMSE; (b) MAE; (c) R
4.3. Typical prediction results

In this section, a typical result of the prediction of Cc using ELM is presented. Obviously, the experimental and predicted Cc were highly correlated for both the training part (Fig. 4a) and the testing dataset (Fig. 4b). Regarding the testing data, lower values of error were found (error mean = -0.0517, error Std = 0.0728), whereas the training part exhibited errors of mean = -0.0725 and Std = 0.0982. The distribution of error with respect to the training and testing datasets was shown in Fig. 5. The maximum error was found at 20% in the testing set and 30% in the training set.

Validation results by the linear fit lines, its equations, and the R values are given in Fig. 6 for the training and testing datasets. Two equations relating the predicted and actual values of Cc were obtained as \( y = 1.1x + 0.034 \) for the training part and \( y = x + 0.0520 \) for the testing one. The values of R were computed as 0.9158 and 0.9566 for the training, testing datasets, respectively, showing an excellent accuracy of ELM in predicting Cc of soft soil.

The performance of ELM used in this study was compared with previously published work presented in Pham et al. (Pham et al., 2019b) and observed that the performance of ELM \((R^2 = 0.915)\) is slightly lower than the performance of Artificial Neural Network (ANN) \((R^2 = 0.9499)\), Support Vector Machines (SVM) \((R^2 = 0.9841)\), and Adaptive Network-based Fuzzy Inference System (ANFIS) \((R^2 = 0.9906)\).
Figure 5. Histograms and error distribution of ELM algorithm in predicting Cc of soft soil for (a) training dataset; (b) testing dataset

Figure 6. Linear fit lines, fit equations, and correlation graphs of ELM in predicting Cc of soft soil for (a) training dataset; (b) testing dataset

5. Conclusions

In this study, a popular machine learning model, namely ELM was developed and applied to predict the compression coefficient (Cc)-a critical soil mechanical parameter. A total of 13 input parameters and one output parameter extracted from 189 experimental soil results were used to generate the training and testing parts for building and validating the model. Monte Carlo approach and other statistical indexes such as RMSE, MAE, and R were applied to validate the performance of the models.

The analysis results showed that ELM showed a strong capacity to predict the Cc of soil, with the value of R up to 0.9158 (training dataset) and 0.9566 (testing dataset). Monte Carlo results also proved that the ELM model has a good convergence under the effect of the random sampling strategy. However, it is also noticed that the ELM was applied in a small dataset (189 samples) collected in the local
site of Vietnam; thus, a bigger dataset with a diversity of soil properties should be used to validate the further application of ELM. The results of this study might help in quick and accurate prediction of the Cc and other vital parameters for reducing cost and time for practical construction.

Acknowledgments
This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 105.08-2019.03.

References


Vietnam Journal of Earth Sciences, 42(3), 228–236


