Evaluation of Soil Liquefaction Potential due to Earthquake using Intelligent Classification Algorithm in Orange Software

Hadi Fattahi[1] Fateme Jiryaee[2]

1. Introduction

Liquefaction is a phenomenon whereby a granular material transforms from a solid state to a liquefied state as a consequence of an increase in pore water pressure. Liquefaction of saturated granular soils due to seismic loading is a major concern for geotechnical engineers, since it may induce uncontrolled lateral spreading of soil masses, potentially causing considerable damages to civil engineering structures. Therefore, assessing liquefaction potential is an imperative task in earthquake geotechnical engineering. There are different methods available for determining the liquefaction potential of soil. Most of these methods depend on some limit states that separate the non-liquefaction region from the liquefaction region established empirically using in situ field observations from sites where test data are available. Among them, standard penetration tests (SPTs), cone penetration tests (CPTs), flat dilatometer tests (DMTs), the shear wave velocity technique (SWV), and self-boring pressure meter (SBPT) are the most commonly used in situ tests for liquefaction potential prediction. However, the high uncertainty in earthquake environments and soil characteristics make it difficult to choose a suitable empirical equation for regression analysis. Therefore, many scholars and experts have attempted to develop scientifically derived analytical models that are simpler and easier to implement, and more accurate than traditional empirical equations for soil liquefaction analysis. Artificial neural network (ANN) models have been used for prediction of liquefaction potential as a classification problem. Although the ANN is found to be more efficient compared to statistical methods, it also has several inherent drawbacks such as over fitting, slow convergence speed, poor generalizing performance, arriving at a local minimum, etc. A support vector machine (SVM) is a machine-learning algorithm, founded by Vapnik (1995), which is gaining popularity due to its good performance and attractive features. The method has been applied to several areas. Application of SVMs in civil engineering is an emerging area and needs to be explored in other disciplines such as geotechnical engineering. SVM-based approaches for classification and prediction have been used in several other civil engineering disciplines and found to work better in comparison to a neural network approach.

In this study, in order to evaluate the potential of soil liquefaction on 79 samples from China Tangshan Earthquake Database, several intelligent classification models were constructed with the help of Orange software. Therefore, the performance of 5 intelligent classification methods (Logistic Regression, ANN, SVM, K-fold Nearest Neighbor (KNN) and Random Forest) were compared based on different criteria.

2. The most important methods used in this research

2.1. Support vector machine. A nonlinear mapping $\varphi(\cdot):R^n \to R^{n_h}$ is specified to map the input and output datasets $\{(\mathrm{x}_i, \mathrm{y}_i)\}_{i=1}^{N}$ (where N is the total number; x_i is inputs and y_i is the actual value) into a so-called high dimensional feature space (which may have infinite dimensions). Such a linear function, that is to say SVM function, is like Eq. (1),

$$f(x) = W^T \varphi(x) + \mathrm{b} \tag{1}$$

where $\varphi(x)$ is the inputs feature and both b and W are coefficients. In SVR approach, the coefficients (b and W) are predicted by minimizing the empirical risk as Eq. (2),

$$R_{emp}(f) = \frac{1}{N}\sum_{i=1}^{N} \Theta_{\varepsilon}\left(y_i, W^T\varphi(x) + \mathrm{b}\right) \tag{2}$$

that $\Theta_{\varepsilon}\left(y_i, W^T\varphi(x) + \mathrm{b}\right)$ is specified as Eq.(3),

$$\Theta_{\varepsilon}\left(y_i, W^T\varphi(x) + \mathrm{b}\right) = \begin{cases} \left|W^T\varphi(x_i) + \mathrm{b} - y_i\right| - \varepsilon, & \text{if } \left|W^T\varphi(x_i) + \mathrm{b} - y_i\right| \geq \varepsilon \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

The SVM focuses on minimizing the training error between the ε-insensitive loss function and the training data and finding the optimum hyper plane. Minimize:

$$\underset{w,b,\xi^*,\xi}{Min} \; R_{\xi}\left(W, \xi^*, \xi\right) = \frac{1}{2}W^T W + C\sum_{i=1}^{N}\left(\xi_i^*, \xi_i\right) \tag{4}$$

with the constraints

$$y_i - W^T\varphi(x_i) - \mathrm{b} \leq \varepsilon + \xi_i^*, \quad \xi_i^* \geq 0, \quad i = 1, 2, ..., N$$

$$-y_i + W^T\varphi(x_i) + \mathrm{b} \leq \varepsilon + \xi_i, \quad \xi_i \geq 0, \quad i = 1, 2, ..., N$$

ξ_i^* : training errors above ε, ξ_i : training errors below –ε and C is parameter to trade off these two terms.

[1] Corresponding author. Associate Professor in Rock Mechanics Engineering, Faculty of Earth Sciences Engineering, Arak University of Technology, Iran. Email: h.fattahi@arakut.ac.ir

[2] MSc Student, Faculty of Earth Sciences Engineering, Arak University of Technology, Iran

The first term of Equation (4), usage of the principle of maximizing the distance of two different training data, is utilized to penalize large weights, to maintain regression function flatness, and to regularize weight sizes. In addition, the parameter W (in Equation (1)) is achieved, after the quadratic optimization problem,

$$W = \sum_{i=1}^{N} \left(\beta_i^* - \beta_i\right) \varphi(x_i) \tag{5}$$

where β_i^*, β_i are the Lagrangian multipliers (obtained by solving a quadratic program). Finally, the SVM function is Eq. (6):

$$f(x) = \sum_{i=1}^{N} \left(\beta_i^* - \beta_i\right) K\left(x_i, x_j\right) + b \tag{6}$$

Here, $K\left(x_i, x_j\right)$ is called the function of kernel that is, $K\left(x_i, x_j\right) = \varphi(x_i)\varphi(x_j)$. In this study, the radial basis kernel function (RBF) $K\left(x_i, x_j\right) = \exp(-\left\|x_i - x_j\right\| / 2\sigma^2), \sigma > 0$ is utilized in the SVM model.

2.2. Artificial neural network. ANN firstly was introduced by McCulloch and Pitts (1943) who presented ability of this technique to calculate any logical functions. Processing of the data is executed with the help of many interconnected simple elements known as neurons which are placed in distinct layers of the network. Multi-layer perceptron, the most famous type of ANNs, consists of at least three layers: input, intermediate or hidden layers, and output. Difficulty level of the problem determines the number of the hidden layers and neurons. The neurons are linked from a layer to the next one, but this connection is not within the same layer. Once a series of inputs presents to the network, the input values are transmitted through the links to the second layer. In every link, the transmitted value is multiplied to the weight of the link. The weighted values are come together at a node in the hidden layer and a bias is summed to the weighted values in that particular node. Consequently, the achieved value transfer to an activation function and a signal is created. Using the departing links of hidden nodes, the results are transmitted to the output layer. Similar to hidden nodes, the input values of the output nodes are weighted, biased, summed, and transferred to the activation function. The created values of activation functions in output layers are the outputs of the network. Performance of an ANN depends on the architecture of the network which is the pattern of the connections existing between the neurons. The network should be trained with sufficient input–output patterns that are known as the training data. As the error reached specified error goal, training is finished and the optimum model is determined.

3. Case studies

It is generally known that the susceptibility of soil deposits to liquefaction is determined by a combination of various factors to which they may be subjected, such as soil properties, geological conditions, and ground motion characteristics. Therefore, it is widely recognized that determining liquefaction potential is a complex geotechnical engineering problem. It is worth noting that soil properties and geological conditions determine the resistance of the deposit to liquefaction, while earthquake characteristics control the seismic loading conditions. Among the factors listed herein, the three most important ones are the following: 1) the ground is a loose sandy deposit; 2) the ground water table is high and the ground is saturated; and 3) the earthquake intensity is sufficiently large and the duration of shaking is sufficiently long. Based on this, the following factors, such as 1) earthquake magnitude, M, 2) water table, d_w, 3) total vertical stress, σ_v, 4) effective vertical stress, σ_{v0}, 5) depth, d_s, 6) peak acceleration at the ground surface, α_{max}, 7) cyclic stress ratio, $\tau_{av}/\sigma_{v0}^{'}$, 8) mean grain size, D_{50}, and 9) measured CPT tip resistance, qc, are selected as the evaluating indices. The database used in this study is from the Tangshan earthquake. A few of the original records were omitted because of incomplete data. The database includes 79 CPT-based field liquefaction records.

4. Results and conclusion

One of the possible consequences of earthquakes in saturated areas is soil liquefaction and as a result the failure of foundations of buildings, types of infrastructure, bridges and many other disasters. In this study, in order to evaluate the potential of soil liquefaction on 79 samples from China Tangshan Earthquake Database, several intelligent classification models were constructed with the help of Orange software. Therefore, the performance of 5 intelligent classification methods (Logistic Regression, ANN, SVM, KNN, and Random Forest) were compared based on different criteria. The results showed that SVM, ANN, and Logistic Regression methods have a high ability to predict soil liquefaction class and among them the Logistic Regression method with AUC index (0.98) was selected as the best method. In addition, the study of the effectiveness of variables using three criteria of Information Gain, Information Gain Ratio and Gini Index, indicates that the variable measured CPT tip resistance is the most effective variable and is the first priority. The variables of cyclic stress ratio and peak acceleration at the ground surface are also important features.