Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı 28, S. 290-294, Kasım 2021

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue 28, pp. 290-294, November 2021

Research Article

Assessment of Feature Selection for Liquefaction Prediction Based on Recursive Feature Elimination

Selçuk Demir[1*], Emrehan Kutluğ Şahin[2]

[1*] Bolu Abant Izzet Baysal University, Faculty of Engineering, Departmant of Civil Engineering, Bolu, Turkey, (ORCID: 0000-0003-2520-4395), selcukdemir@ibu.edu.tr

[2] Bolu Abant Izzet Baysal University, Faculty of Engineering, Departmant of Civil Engineering, Bolu, Turkey, (ORCID: 0000-0002-9830-8585), emrehansahin@ibu.edu.tr

(1st International Conference on Applied Engineering and Natural Sciences ICAENS 2021, November 1-3, 2021)

(**DOI:** 10.31590/ejosat.998033)

ATIF/REFERENCE: Demir, S. & Sahin, E. (2021). Assessment of Feature Selection for Liquefaction Prediction Based on Recursive Feature Elimination. *European Journal of Science and Technology*, (28), 290-294.

Abstract

This paper presents a machine learning model using a random forest (RF) algorithm with the recursive feature elimination (RFE) for the soil liquefaction prediction. The prediction model is tested on 253 CPT-based field data from different earthquakes. RFE, which is one of the feature selection methods, was adopted for eliminating irrelevant features in the mentioned dataset, and then the performance of the RFE-RF (i.e., the model determined by the RFE method) and the RF models with all variables were compared in terms of their performance matrices. The primary focus of this study is to investigate the effectiveness of the feature selection algorithm approach, therefore the raw data that is a benchmark dataset was used to compare the performance of the RFE-RF. The result showed that the RFE approach improved the overall accuracy of the liquefaction prediction.

Keywords: Liquefaction Prediction, Feature Selection, Machine Learning, Recursive Feature Elimination, Random Forest.

Sıvılaşma Tahmininde Özyinelemeli Özellik Seçmeye Dayalı Faktör Seçme Yönteminin Değerlendirilmesi

Öz

Bu çalışma, zemin sıvılaşması tahmini için özyinelemeli özellik seçimi (RFE) ile rastgele orman (RF) algoritması kullanan bir makine öğrenme modeli sunmaktadır. Tahmin modeli, farklı depremlerden elde edilen 253 CPT tabanlı saha verileri üzeri kullanılarak test edilmiştir. Söz konusu veri setindeki ihtiyaç fazlası özelliklerin elimine edilmesi için özellik seçim yöntemlerinden biri olan RFE benimsenmiştir. Ardından RFE-RF'nin (yani RFE yöntemiyle belirlenen modelin) ve bütün değişkenlerin kullanıldığı RF modelin performansları performans matrisleri açısından incelenmiş ve karşılaştırılmıştır. Bu çalışmanın önceliği, öznitelik seçim algoritması yaklaşımının etkinliğini araştırmaktır, bu nedenle RFE-RF'nin performansını karşılaştırmak için bir kıyaslama veri seti olan ham veriler kullanılmıştır. Sonuç olarak, RFE yaklaşımının kullanılmasının sıvılaşma tahmin modelinin genel doğruluğunu arttırdığı görülmüştür.

Anahtar Kelimeler: Sıvılaşma Tahmini, Özellik Seçimi, Makine Öğrenme, Özyinelemeli Özellik Seçimi, Rastgele Orman.

[*] Corresponding Author: selcukdemir@ibu.edu.tr

1. Introduction

It has long been recognized that soil liquefaction-induced earthquake hazards have enormously contribute to social and economic losses. The liquefaction phenomenon is often described in the literature as a transformation of cohesionless soils from solid to viscous state due to the generation of excess porewater pressures and negligible shear resistance under seismic loads (National Academies of Sciences, Engineering, and Medicine, 2006; Kumar et al., 2021). Liquefaction-related ground failures and hazards have been observed in many previous notable earthquakes (Niigata 1964, Alaska 1964, Kobe 1995, Kocaeli 1999, Christchurch 2010-2011). For this reason, the challenge for predicting soil liquefaction and its effects has been still studying by many engineers and researchers.

The literature consists of several methods to estimate the soil liquefaction potential. Typical methods for the evaluation of liquefaction potential at a specific site are laboratory tests and field studies (in situ tests). Due to the disadvantages of laboratory tests, such as representing actual field conditions and obtain undisturbed soil samples, the in-situ tests are much preferred for soil liquefaction evaluation. The commonly used liquefaction evaluation procedure developed by Seed and Idriss (1971) are based on different in situ tests including standard penetration test (SPT) (Cetin et al., 2004; Idriss and Boulanger, 2008), cone penetration test (CPT) (Robertson and Wride, 1998; Boulanger and Idriss, 2014) and shear wave velocity (Vs) (Andrus and Stokoe, 2000; Kayen et al., 2013). However, these methods have some limitations and require extensive resources.

In recent years, soft computing methods (e.g., artificial intelligence, machine learning) have been successfully applied in various geotechnical engineering topics. Most recent literature surveys point out that soft computing models based on in situ testing data have been increased and considered as an alternative approach for liquefaction prediction (Hoang and Bui, 2018). Artificial neural networks (ANN), support vector machines (SVM), and decision tree-based algorithms such as Random Forest (RF) are popular tools for data processing with different parameters. For example, researchers have employed the ANN model to estimate soil liquefaction (e.g., Erzin and Ecemis, 2015; Shahri, 2016). In addition, some researchers have used the SVM algorithm to determine prediction models for soil liquefaction (Samui et al., 2011; Xue and Yang, 2016). Similarly, Kohestani et al. (2015) have utilized an RF algorithm to predict the occurrence of soil liquefaction.

It is often desired to have a model that has the best predictive ability. However, the performance of prediction models reduces when models include noisy and redundant features. Especially, some of the learning models, such as support vector machines and neural networks, may have been significantly affected by irrelevant features (Kuhn and Johnson, 2019). This performance can be improved by removing the superfluous features using feature selection (FS) methods (Guyon and Elisseeff, 2003). FS is an effective technique in order to be able to decide which features are mostly irrelevant and should be eliminated from the original feature space. Recursive feature elimination (RFE) is one of the examples of the most preferred feature selection algorithms to analyze datasets and achieve the best model performance (Granitto et al., 2006; Sánchez-Maroño et al., 2007; Gregorutti et al., 2017).

This paper presents a machine learning model using an RF algorithm for soil liquefaction prediction. The RFE algorithm was applied to define only important and relevant features of the dataset. The performance of the RFE-RF and RF models were separately investigated in terms of their performance matrices. All computations in this paper were handled with the open-source software called R (Team R, 2020).

2. Material and Method

2.1. Data Used

The prediction model was constructed by using CPT data collected from 253 soil liquefaction cases that occurred in different countries (Boulanger and Idriss, 2014). Summary of statistical measures of the CPT data is summarized in Table 1. This dataset consists of the following ten independent variables, namely depth of the soil specimen (d); cone tip resistance (q_c); sleeve friction ratio (R_f); fine content (FC); depth of ground water table (dw); total and effective vertical stress (σ_v and σ'_v); maximum horizontal acceleration at the ground surface (a_{max}); cyclic stress ratio (CSR); and earthquake moment magnitude (M_w).

Table 1. Statistical measures of the CPT data

Variable	Min-Max	Mean	Median
d	1.4-11.8	4.45	4.1
q_c	0.94-45	5.9	4.79
R_f	0.03-2.91	0.76	0.61
FC	0-85	17.82	11
d_w	0.20-7.20	2.04	1.8
σ_v	24-210	81.19	74
σ'_v	19-147	57.65	53
a_{max}	0.09-0.84	0.32	0.28
CSR	0.07-0.70	0.28	0.25
M_w	5.9-9	6.98	6.93

2.2. Random Forest and Recursive Feature Elimination

Random Forest (RF) (Breiman, 2001) is a popular ensemble learning algorithm consisting of separately trained binary decision trees. The RF algorithm has many advantages as compared to other machine learning algorithms. For example, it can be used for both classification and regression problems. Moreover, RF is user-friendly and requires fewer hyperparameters, such as the number of decision trees, the number of suitable features for splitting, and the minimum size of the maximum depth of each tree.

Recursive Feature Elimination (RFE) is a feature selection algorithm proposed by Guyon et al. (2002) that basically works by using all features with a ranking according to their feature importance. The process of the RFE algorithm continues recursively by removing the least important features until the desired number of features remains. After that, the best accuracy performance is achieved by removing irrelevant features with RFE.

2.3. RFE with Cross-Validation (CV)

k-Fold CV as well as grid search technique were employed to find the optimal variables and to get a robust result for liquefaction prediction. In each model of RFE with CV, the subset of the optimal variables was determined by the highest accuracy. In this study, k was set as 10 and the simulation was repeated 3 times.

Figure 1 presents the result of the RFE method for achieving the best predictive features. It can be seen that accuracies increase gradually from 56.4% to a maximum value of 79% as the number of elements rises from one to six. Thereafter, the accuracy curve becomes flatter, and accuracies change negligibly, which means the remaining variables do not affect the accuracy of the model. Selected six variables (qc, amax, CSR, FC, σv, and σ′v) and their feature importance rankings are given in Figure 2. For further analysis, the RF model with these six variables (RFE-RF) and RF with ten variables were separately investigated for evaluating their performances.

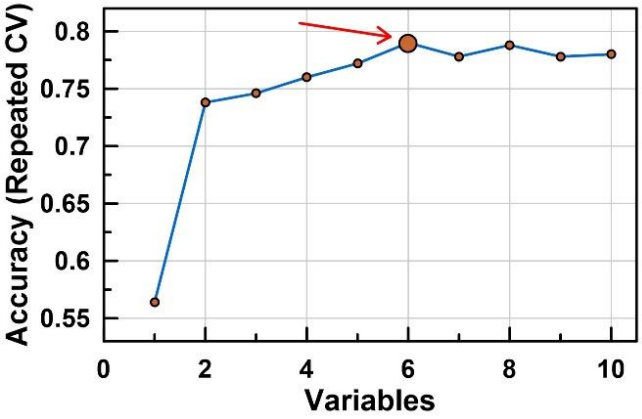

Figure 1. The accuracy curve of ten variables based on RFE

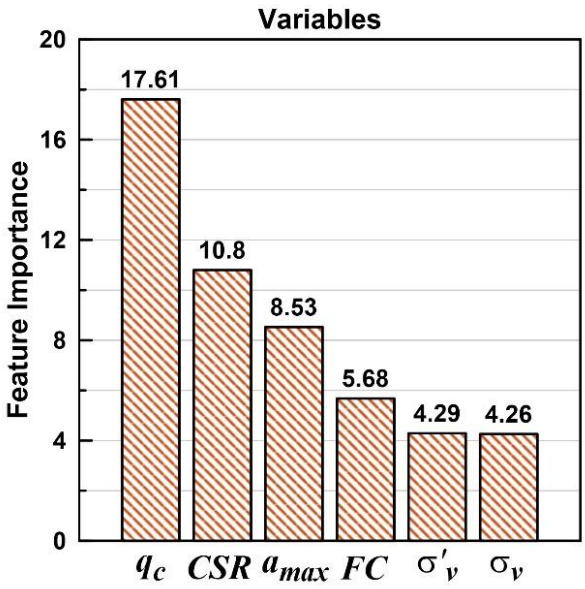

Figure 2. Variation of feature importance rankings of the selected variables

2.4. Parameter Optimization

In the RF algorithm, the number of trees (ntree) and the number of suitable features for splitting (mtry) are adjustable parameters related to the performance of classification accuracy. In this paper, accuracy results were used to find the optimum values for ntree and mtry regarding k-Fold CV. Figure 3 and Figure 4 show optimization results both for six and ten- featured model cases, respectively. According to the Figure 3 and Figure 4, the accuracy is highest when ntree =200 and mtry =2 for the six-featured model and ntree =100 and mtry =2 for the model having all ten features.

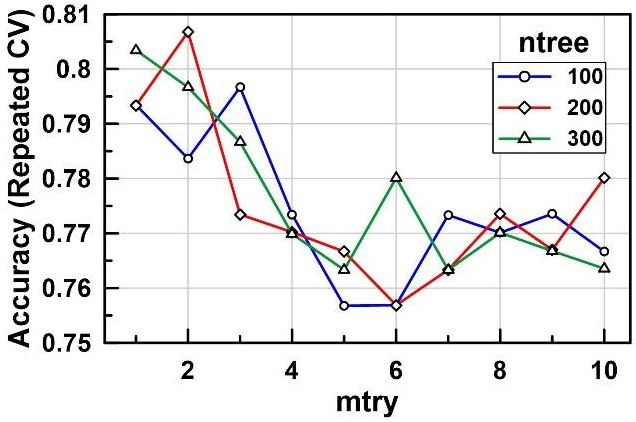

Figure 3. Parameter optimization for six-featured model

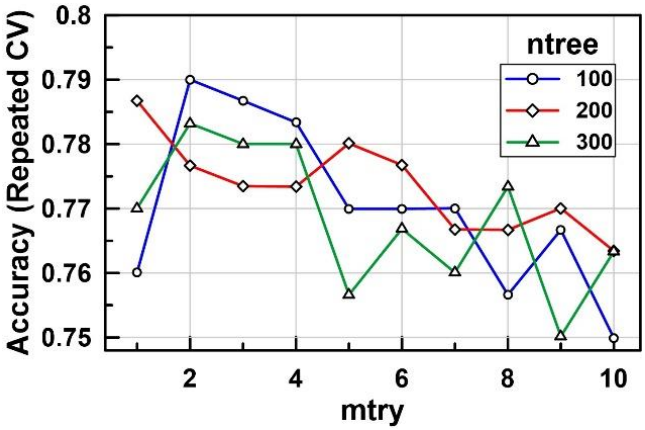

Figure 4. Parameter optimization for ten-featured model

3. Results

Different ratio options for splitting training and test data can be preferred based on machine learning methods. In the present study, training /test data ratio was used in the ratio of 70:30 with the aid of the stratified random sampling technique. The basic idea of this technique is that it partitions the entire dataset into relatively homogeneous groups of samples, in which the training and testing datasets of each fold contain roughly the same proportion of each class label. For calculating the performance of models, Accuracy, Kappa, Precision, Recall, and F1 metrics were used. Details of the used matrices are given in Table 2. The performance metrics are based on the Confusion Matrix (CM). The CM essentially places the resulting predictions into four groups; TP: True-Positive, FP: False-Positive, TN: True-Negative, FN: False-Negative.

The results of the performance matrices corresponding RFE-RF and RF models are compared in Figure 5. It is clearly seen that from Figure 5, the RFE-RF model outperforms compared to the RF model in terms of accuracy, kappa, precision, recall, and F1, respectively. The accuracy, kappa, precision, recall, and F1, were

obtained as 88.64%, 77.27%, 95.45%, 84.00%, and 89.36%, respectively for the RFE-RF model. So, when the RFE-RF model was compared to the RF model, results revealed that applying RFE increased the accuracy by 6%. Thus, it can be concluded from all examined results that applying the feature selection method with RFE enhanced the quality of the liquefaction prediction.

Table 2. Performance matrices

Metrics	Range	Formula
Accuracy	0-1.0	$\frac{(TN+TP)}{(TN+FN+TP+FP)}$
Kappa	-1.0-1.0	$\frac{Accuracy-RA}{1-RA}$
Precision	0-1.0	$\frac{TP}{TP+FP}$
Recall	0-1.0	$\frac{TP}{TP+FN}$
F1	0-1.0	$2\frac{\text{PrecisionxRecall}}{\text{Precision+Recall}}$
RA	$\frac{(TN+FP)(TN+FN)+(TP+FN)(TP+FP)}{Accuracy^2}$	

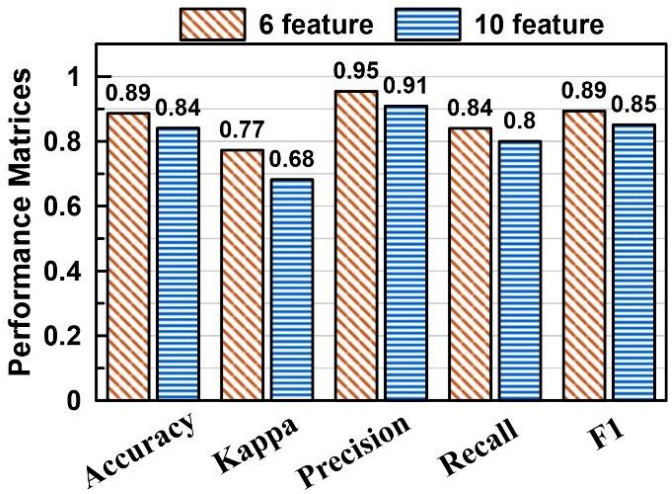

Figure 5. Analysis results of prediction models from different performance matrices

4. Conclusions

The main aim of the study was to predict soil liquefaction with a feature selection algorithm. Therefore, the recursive feature elimination algorithm was employed to select the most relevant features and to enhance the predictive performance of the model. According to the results of the study, the variables of R_f, d, d_w and M_w were found less important parameters among the other CPT parameters after applying RFE and the RFE-RF model exhibited better performance than the RF model. As a result, this study shows that removing redundant parameters improved the capability of the liquefaction prediction model.

References

Andrus, R. D., & Stokoe II, K. H. (2000). Liquefaction resistance of soils from shear-wave velocity. Journal of geotechnical and geoenvironmental engineering, 126(11), 1015-1025.

Boulanger, R. W., & Idriss, I. M. (2014). CPT and SPT based liquefaction triggering procedures. Center for Geotechnical Modelling, Civil and Environmental Engineering, UC Davis, CA. Report No. UCD/CGM-14/01.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Cetin, K. O., Seed, R. B., Der Kiureghian, A., Tokimatsu, K., Harder Jr, L. F., Kayen, R. E., & Moss, R. E. (2004). Standard penetration test-based probabilistic and deterministic assessment of seismic soil liquefaction potential. Journal of geotechnical and geoenvironmental engineering, 130(12), 1314-1340.

Erzin, Y., & Ecemis, N. (2015). The use of neural networks for CPT-based liquefaction screening. Bulletin of Engineering Geology and the Environment, 74(1), 103-116.

Granitto, P. M., Furlanello, C., Biasioli, F., & Gasperi, F. (2006). Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. Chemometrics and intelligent laboratory systems, 83(2), 83-90.

Gregorutti, B., Michel, B., & Saint-Pierre, P. (2017). Correlation and variable importance in random forests. Statistics and Computing, 27(3), 659-678.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of machine learning research, 3(Mar), 1157-1182.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine learning, 46(1), 389-422.

Hoang, N. D., & Bui, D. T. (2018). Predicting earthquake-induced soil liquefaction based on a hybridization of kernel Fisher discriminant analysis and a least squares support vector machine: a multi-dataset study. Bulletin of Engineering Geology and the Environment, 77(1), 191-204.

Idriss, I. M., & Boulanger, R. W. (2008). Soil liquefaction during earthquakes. Earthquake Engineering Research Institute.

Kayen, R., Moss, R. E. S., Thompson, E. M., Seed, R. B., Cetin, K. O., Kiureghian, A. D., ... & Tokimatsu, K. (2013). Shear-wave velocity–based probabilistic and deterministic assessment of seismic soil liquefaction potential. Journal of Geotechnical and Geoenvironmental Engineering, 139(3), 407-419.

Kohestani, V. R., Hassanlourad, M., & Ardakani, A. J. N. H. (2015). Evaluation of liquefaction potential based on CPT data using random forest. Natural Hazards, 79(2), 1079-1089.

Kuhn, M., & Johnson, K. (2019). Feature engineering and selection: A practical approach for predictive models. CRC Press.

Kumar, D., Samui, P., Kim, D., & Singh, A. (2021). A Novel Methodology to Classify Soil Liquefaction Using Deep Learning. Geotechnical and Geological Engineering, 39(2), 1049-1058.

National Academies of Sciences, Engineering, and Medicine. (2016). State of the art and practice in the assessment of earthquake-induced soil liquefaction and its consequences. Washington, DC: The National Academies Press. doi, 1017226, 23474.

Robertson, P. K., & Wride, C. E. (1998). Evaluating cyclic liquefaction potential using the cone penetration test. Canadian geotechnical journal, 35(3), 442-459.

Samui, P., Kim, D., & Sitharam, T. G. (2011). Support vector machine for evaluating seismic-liquefaction potential using shear wave velocity. Journal of applied geophysics, 73(1), 8-15.

Sánchez-Maroño, N., Alonso-Betanzos, A., & Tombilla-Sanromán, M. (2007, December). Filter methods for feature selection–a comparative study. In International Conference on Intelligent Data Engineering and Automated Learning (pp. 178-187). Springer, Berlin, Heidelberg.

Seed, H. B., & Idriss, I. M. (1971). Simplified procedure for evaluating soil liquefaction potential. Journal of the Soil Mechanics and Foundations division, 97(9), 1249-1273.

Shahri, A. A. (2016). Assessment and prediction of liquefaction potential using different artificial neural network models: a case study. Geotechnical and Geological Engineering, 34(3), 807-815.

Team, R. C. (2020). R: the R project for statistical computing. https://www.r-project.org/

Xue, X., & Yang, X. (2016). Seismic liquefaction potential assessed by support vector machines approaches. Bulletin of Engineering Geology and the Environment, 75(1), 153-162.