Prediction of Distribution Coefficient by Using the Random Forest and the Nested K-fold Cross Validation Method

Do-Hyeon Kim and Jun-Yeop Lee*
School of Mechanical Engineering, Pusan National University
2, Busandeahak-ro 63beon-gil, Geumjeong-gu, Busan 46241, Korea
Corresponding author: jylee@pusan.ac.kr

1. Introduction

According to the continuous use of nuclear energy, the amount of spent nuclear fuel waiting for the final disposal has been increasing progressively. Therefore, it is highly required to establish a disposal plan for those high-level wastes because of its remarkable radiotoxicity and potential hazards to the environment.

Sorption reaction is considered to be one of the major geochemical reactions hindering the migration process of various radionuclides in the geologic repository-relevant condition. In general, the sorption behavior of radionuclides can be explained with the distribution coefficient (K_d). Since the K_d is a conditional parameter and thus is dependent on various environmental conditions (*i.e.* pH, ionic strength, mineral type, solid-liquid ratio, etc.), establishing a model to predict the sorption behavior of radionuclide is known to be highly complicated.

The objective of the present work is to predict the distribution coefficient of various radionuclides onto the bentonites by using the machine learning-based random forest (RF) [1,2] method coupled with the nested K-fold cross validation approach.

2. Materials and Methods

The distribution coefficients employed in this study were taken from the JAEA-SDB [3]. Three types of bentonites (such as MX-80, SWy-2, and Kunigel V1) and 22 kinds of radionuclides (Am, Ac, Co, Cm, Cd, Cs, Cu, Na, Np, Ni, Nb, U, Sr, Sn, Pb, Pa, Pu, Po, I, Tc, Th, and Zr) were selected for further data processing and the establishment of the machine learning model. The database for machine learning calculation included 9 variables such as solid-liquid ratio (LS, unit: mL/g), ionic strength (IS, unit: mol/L), oxidation number of radionuclide (RX), acidity (pH), initial radionuclide concentration (C_0, unit: mol/L), cation exchange capacity (CEC, unit: meq/100 g), surface area (SA, unit: m^2/g), electronegativity (EN), and ionic radius (IR, unit: Å).

For the establishment of the machine learning model, the RF method, a supervised ensemble machine learning approach based on multiple decision trees and bagging, was employed in the present work. In the course of the calculation, various hyperparameters were adjusted to monitor the change in the coefficient of determination (R^2) and the root mean square error (RMSE). Particularly, the number of decision trees (N_T), the number of features selected to be used to divide each node (N_F), and two different types of random seeds were selected to be controlled during the calculation. The random seeds employed to distinguish the train/test sets and internal RF model calculation were referred to as RS_T and RS_M, respectively.

Additionally, the relative importance of input variables used in the calculation of the distribution coefficient was quantified by using the mean decrease in impurity (MDI) approach [4].

Furthermore, the nested K-fold cross validation (CV) method was employed to validate the normal RF model result owing to the presumable overfitting and bias of data problems expected in the normal RF model calculation. In the present work, the CV method with the double loop consisting of five inner and outer fold loops was utilized to provide further robustness and redundancy in the calculation result.

The computational codes to establish the normal RF model and to perform the nested K-fold CV were taken from the scikit-learn software package [2].

3. Results

The ranges of the hyperparameters controlled in the RF calculation were N_T = 5 – 1000 (set in multiples of five intervals), N_F = 2 – 9, RS_T = 0 – 10, and RS_M = 0 – 10. Note that the other parameters were fixed at their default values.

Figure 1 presents the range of R^2 values calculated with two different approaches such as the normal RF model and the nested K-fold CV. According to the result, the highest R^2 value among the entire R^2 results was obtained with the normal RF model calculation.

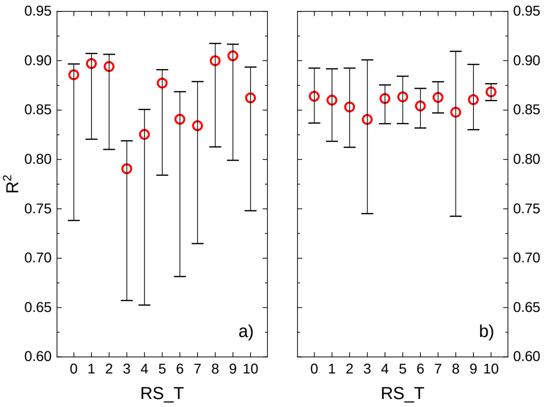

Fig. 1. The entire range of R^2 values calculated with various RS_T by using (a) the normal RF model and (b) the nested K-fold CV. Symbol presents the averaged R^2 value at given RS_T.

The maximum R^2 value derived with the normal RF model was $R^2 = 0.9175$ at RS_T = 8, RS_M = 5, $N_T = 105$, and $N_F = 3$. However, the result obtained with the nested K-fold CV shows a relatively low R^2 value compared with that determined with the normal RF model. Nevertheless, the relative deviation of averaged R^2 value for various RS_T and the range of R^2 values were significantly decreased, indicating that the robustness and stability of the calculation model were dramatically enhanced. The maximum of the averaged R^2 value determined with the nested K-fold CV was $R^2 = 0.8683$ at RS_T = 10.

Figure 2 presents the comparison results between the experimental log K_d values with those predicted with two different machine learning approaches.

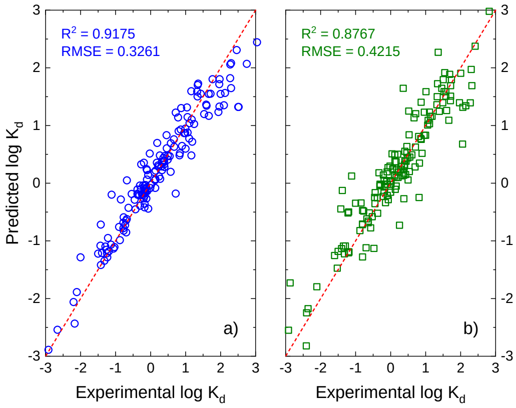

Fig. 2. Comparisons of experimental and predicted log K_d values evaluated with (a) the normal RF model with the highest R^2 value at RS_T = 8 and (b) the nested K-fold CV at the fourth outer fold loop at RS_T = 10

Because the nested K-fold CV method performs an additional validation process to avoid the overfitting of data, the R^2 value derived from the nested K-fold CV is somewhat lower than that obtained with the normal RF model. Still, according to the considerable robustness and stability of the model possibly expected from the nested K-fold CV method, it can be judged that the usefulness of the result calculated with the nested K-fold CV is sufficient even though the R^2 value is slightly low.

Table 1 shows the relative importance of the input variables employed in the present work determined by using the MDI approach and the nested K-fold CV.

Table 1: Relative importance of input variables

Variables	Relative importance (%)
pH	32.8
C_0*	14.3
IS	11.4
IR	10.4
LS*	10.2
EN	9.8
RX	7.3
SA	2.4
CEC	1.4

*log-scaled value was used in this study

Among 9 input variables, the pH contributed the largest influence on the K_d value prediction while the CEC and SA provided remarkably low influences. The tendency towards low importance of the CEC and SA is caused apparently by small divergences in the mineral properties since only types of bentonites such as MX-80, SWy-2, and Kunigel V1 were considered in this study.

4. Conclusions

The computational prediction model for K_d was constructed by adopting the machine-learning based RF model together with the JAEA-SDB. The model established in the present work enables the reliable estimation of K_d value under arbitrarily given condition. Although the nested K-fold CV provided somewhat lower R^2 value than that produced by the normal RF model, the nested K-fold CV was assessed to be an advisable way to avoid presumable overfitting and bias problems. Furthermore, the MDI approach suggested that the K_d is highly influenced by the pH and initial radionuclide concentration.

According to the result obtained in this study, the normal RF model and the nested K-fold CV approach were assessed to be a useful method for the reliable prediction of K_d under arbitrary conditions for various radionuclides considered in the deep-geologic disposal facility.

ACKNOWLEDGEMENT

This work was supported by the National Research Foundation of Korea grant funded by the Korea government [No. NRF-2021M2E1A1085204].

REFERENCES

[1] L. Breiman, Random forests, Machine Learning, Vol. 45, pp.5-32, 2001.

[2] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research Vol. 12, pp.2825-2830, 2011.

[3] Y. Sugiura, T. Suyama, and Y. Tachi, Development of JAEA sorption database (JAEA-SDB): Update of sorptional/QA Data in FY2019, Japan Atomic Energy, Ibaraki, Japan, 2020.

[4] G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts, Understanding variable importances in forests of randomized trees, Proceedings of the Advances in Neural Infromation Processing Systems, Dec.5-10, 2013, Lake Tahoe, NV, USA.