Comparison of ten machine learning algorithms for classification of the tangential model parameters for various soil texture classes

*Dang Quoc Thuyet[2], Hirotaka Saito[1], Yuji Kohgo[1]

1. Institute of Agriculture, Tokyo University of Agriculture and Technology, 2. Graduate School of Agricultural and Life Sciences, The University of Tokyo

The tangential model (TANMOD) is one of the few soil water retention curve (SWRC) models that can be applied in both unsaturated and saturated soils accounting for volume changes in the entrapped air in soil pores. Understanding the behavior of the model parameters and their relation with soil texture classes is essential for interpreting soil hydraulic phenomena and expanding the model application. Machine learning has recently emerged as robust algorithms to efficiently model or classify multi-dimension data. The advantage of machine learning techniques as compared to other traditional statistical methods is that the model can handle non-linear problems and does not rely on pre-defined equations. In this study, we aimed to compare ten different machine learning algorithms to classify TANMOD parameters to assess the underlying relationship between the parameters and the soil texture classes.

The TANMOD was obtained by fitting 399 SWRC from 10 USDA soil texture classes in the UNSODA soil database. The model parameters consist of three coordinates (S_{re}, s_e), (S_{rm}, s_m), and (S_{rf}, s_f), three tangential slopes, c_e, c_m, and c_f, along the curve. Ten common supervised machine learning algorithms i.e. Linear Discriminant Analysis (LDA), Classification and Regression Trees (CART), Bagging of Classification and Regression Trees (TB), Decision Trees and Rule-Based Models using Quinlan's C5.0 algorithm (C5.0), k-Nearest Neighbors (kNN), Naive Bayes (NB), Neural Network (NNET), Random Forest (RF), Stochastic Gradient Boosting (Generalized Boosted Modeling) (GBM), and Support Vector Machines with a linear kernel (SVM) were used to classify TANMOD parameters to 10 USDA soil texture classes.

The results showed that RF was the best classification algorithm among supervised machine learning algorithms. The accuracy of the classification of RF was 62.6 %. The maximum tangential slope, c_m, was the most important parameter whereas the coordinate s_e and S_{re} were less important for the classification. The uncertainty and effect of Train/Test data ratio on the classification of RF were also examined. Train/Test data ratio during random splitting of original data could vary from 80%/20% to 50%/50% that did not significantly influence the prediction accuracy of RF. Finally, the TANMOD parameters not only have their own physical meaning but also can be used to relate to the soil particle size distribution and/or the soil texture class. We recommend the RF algorithm to reveal the underlying relationship of the parameters and soil texture class. The relationships are beneficial for the application of any SWRC models.

Keywords: machine learning

