Prediction of cyanobacteria blooms in the lower Han River (South Korea) using ensemble learning algorithms

Jihoon Shin, Seonghyeon Yoon, Yoonkyung Cha*

Department of Environmental Engineering, University of Seoul, Seoul, Korea, email: ykcha@uos.ac.kr (Y.K. Cha)

Received 5 February 2017; Accepted 13 April 2017

ABSTRACT

We developed a prediction model for cyanobacterial blooms in the lower Han River, South Korea, using decision tree algorithms. Decision tree is a type of machine learning method that can overcome missing values or outlier problems. Despite its simple application, it can accurately predict complex natural phenomena. To improve the robustness of the model, we used ensemble methods such as Bagging, AdaBoost, and Random Forest, and the performance of each method was compared against that of a single decision tree. The indicators of cyanobacterial blooms, namely chlorophyll-a concentration and cyanobacteria cell count, were classified into either the non-exceedance or the exceedance class according to administrative guidelines or criteria, and used as the response variables. Since the cyanobacteria cell count in the exceedance class was much smaller than that in the non-exceedance class, the synthetic minority over-sampling technique (SMOTE) was used to mitigate the imbalance between classes. The prediction abilities for chlorophyll-a and cyanobacteria were evaluated based on multiple indices, including area under curve (AUC). The result showed that the performance of ensemble models improved by 1.7%–11.1% and 1.5%–4.9% compared with that of the single model for chlorophyll-a and cyanobacteria, respectively. The implementation of SMOTE to mitigate the imbalance cyanobacteria cell count data enhanced AUC by 4.3%–6.7%. The results of the variable importance analysis indicated that water temperature, flow, and month were essential factors for the prediction of the cyanobacteria classes.

Keywords: Classification tree; Ensemble; Cyanobacteria bloom; Lower Han River; Data imbalance