Desalination and Water Treatment
www.deswater.com

doi: 10.5004/dwt.2017.20986

84 (2017) 31–39
July

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

1. Introduction

An increasing number of cyanobacterial bloom events have been reported in freshwater and coastal systems worldwide. The main causes are believed to be eutrophication of water bodies and global warming [1]. Cyanobacterial blooms could result in highly turbid water, limiting the use of water resources and making water treatment prohibitively expensive. In particular, there is increasing concern over certain cyanobacteria species that can produce toxins harmful to human health and the ecosystem [2].

Given the enormous impact of cyanobacterial bloom events on water quality and the aquatic ecosystem, accurate predictions are essential for effective management of water resources. Nevertheless, cyanobacterial bloom events are difficult to predict since this is a complex natural phenomenon affected by numerous variables. Recently, there has been an increasing interest in using machine learning combined with observational data to predict complex phenomena [3–5]. Among machine learning methods, the decision tree approach can overcome the problems caused by missing values and outliers in the data. Owing to its easy and simple application, it has been widely used for predicting natural phenomena [6–8]. However, when a single decision tree alone is used, small variations in training data might cause significant changes in the model. Ensemble methods are believed to act as a countermeasure against this weakness [9]. These methods can generate more robust models than the single decision tree model by generating and integrating multiple

* Corresponding author.

models. In this study, cyanobacterial blooms were predicted in the lower Han River, South Korea, using the ensemble decision tree algorithms. The performance of each ensemble approach was compared with that of the single decision tree.

Since typical machine learning algorithms assume that the data sets being used are balanced, data imbalance could pose problems in classification. When data are imbalanced, the classifier of the decision tree can result in more inaccurate predictions for the minority class, which consists of a small number of data, than the majority class, during optimization of the overall performance [10–12]. In this study, a single class consisting of high chlorophyll-a concentration and cyanobacteria cell count that exceed the management guidelines or environmental standards for water quality is considered as the minority class, since cyanobacterial blooms are not frequently occurring phenomena [11]. Environmental scientists and managers, including us undertaking this study, are generally interested in the occurrences of the minority class. Nonetheless, the prediction power of the model for the minority class is relatively inaccurate because of data imbalance. Data preprocessing, cost-sensitive learning, and algorithm modifications have been used to overcome this issue. Data preprocessing, which is the most widely used method, was employed in this study [13].

The Han River is one of the largest rivers in South Korea, flowing from east to west of Seoul, the capital of the country. The river is 481.7 km long, with a basin area of 260,188 km^2. It serves as a major drinking water source for citizens of the Seoul metropolitan area [14] (Fig. 1). The average annual temperature and total annual precipitation during the years 2007–2015 were 12.8°C and 1,430 mm, respectively. The precipitation in summer alone (July–September) accounts for 61% of the total precipitation (Korea Meteorological Administration, http://www.kma.go.kr). The increase in the infrastructure around the Han River, such as bridge piers and the underwater bridge, has created stagnant water pools. Growing industrialization and urbanization add to the problem. Hence, the local water pollution is severe and cyanobacterial blooms are likely to occur because of eutrophication in the downstream area [15].

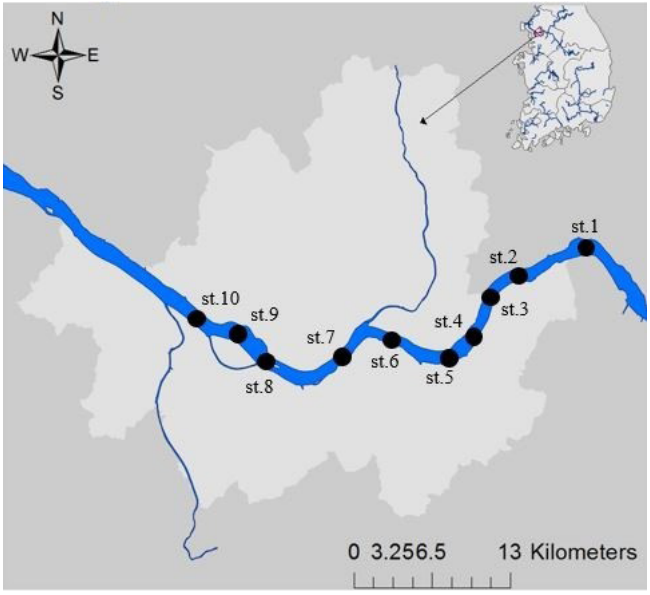


Fig. 1. Map of the lower Han River, the study site. Ten sampling sites were numbered from upstream to downstream; st.1: Gangbuk, st. 2:Amsa, st. 3: Guui, st.4: Pungnap, st. 5: Jayang, st. 6: Seongsu Bridge, st. 7: Hannam Bridge, st. 8 Hangang Bridge, st. 9 Mapo Bridge, and st. 10 Seongsan Bridge.

The section of the Han River flowing through Seoul is equipped with a cyanobacteria bloom alert system [15]. The cyanobacteria bloom alert system issues an "advisory" level if the cyanobacteria cell count exceeds 1,000 cell/mL, a "warning" level if it exceeds 10,000 cell/mL, and an "outbreak" level for counts higher than 1,000,000 cell/mL in two consecutive samplings. The current response for the occurrence of cyanobacteria blooms has been designed according to the above-mentioned levels. The harms resulting from cyanobacteria blooms could be prevented or mitigated by accurately predicting the occurrence of cyanobacteria cell counts exceeding the guidelines.

The objective of this study was to predict cyanobacteria blooms in the lower Han River using decision tree algorithms. The performance of single and ensemble decision trees was compared, and the effect of using preprocessing for addressing data imbalance issues was assessed.

2. Methods

2.1. Data description

Data for water temperature (°C), biochemical oxygen demand (mg/L), dissolved oxygen concentration (mg/L), chemical oxygen demand (mg/L), suspended solids (mg/L), total organic carbon (mg/L), total nitrogen (mg/L), total phosphorus (mg/L), chlorophyll-a (mg/m^3), and cyanobacteria cell count (cells/mL) were obtained from the Water Information System managed by the National Institute of Environmental Research. The weekly meteorological data from April 2007 to December 2015, such as daily precipitation (mm) and irradiance (MJ/m^2), were collected from the Korea Meteorological Administration, and the flow rate data (m^3/s) were collected from Water Resources Management Information System (WAMIS). In addition, weekly total precipitation and average weekly irradiance were calculated and added to the data set to observe the accumulation effect. The observational data for Seoul were used as the meteorological data.

2.2. Decision tree algorithms

The decision tree approach yields the most homogeneous binary splits to explain the variation of the response variable by "testing" the attributes of the data in various ways. Using the decision tree provides one major advantage: simple application. It can deal with both categorical (classification) and continuous-(regression-) response variables and can overcome missing data and outlier problems [7]. However, numerous variables with similar sorting capabilities might be present when constructing a single decision tree. Thus, small changes in data would cause large variations in the constructed model, the prediction performance of which is susceptible to instability [9]. Ensemble models can overcome problems that may arise when using a single decision tree,

as this involves constructing and integrating multiple models to obtain a result. Thus, we developed a single decision tree and various ensemble models to compare their performance.

2.2.1. A single decision tree model

For developing a single decision tree, we used the Classification and Regression Tree algorithm (CART), a method based on the probability theory and sophisticated statistics. Using the CART, the decision tree was generated from the starting node (root) to the terminal node (leaf). The Gini index was used to evaluate the impurity in each node, and dichotomy was applied based on the variables with the lowest Gini index. The size was optimized to the extent possible (the process was stopped only because of lack of data). Then, a tree with the optimal size was generated by pruning [16,17].

If pruning is not applied to the optimum-sized tree, the tree would provide poor prediction accuracy for test data or other data due to over-fitting to the training data. Therefore, pruning based on cost complexity was carried out to prevent over-fitting as follows:

$$R_\alpha(T) = R(T) + \alpha |T|$$

where $R_\alpha(T)$ refers to the cost complexity of tree T, and $R(T)$ refers to the training sample error cost. $|T|$ is the number of terminal nodes, and α is the complexity penalty of each node. Pruned subtrees of different cost complexities were generated during the pruning process. The tree with the lowest cost value among all subtrees was selected as the optimum tree.

2.2.2. Ensemble decision tree models

An ensemble model is used to develop a prediction model with better performance than a single model, as it generates and integrates multiple models [18]. Boosting, Bagging, and Random Forest, which have been widely used as ensemble models, were applied in this study.

2.2.2.1. Boosting[TS: Please check head level 4] The AdaBoost algorithm was used for boosting. AdaBoost stands for "adaptive boosting algorithm." It has been applied in various fields and has achieved outstanding results owing to its robust theoretical foundation, accuracy, and simplicity [16]. AdaBoost is an algorithm that generates a number of weak classifiers and integrates them to improve prediction accuracy. For example, a binary classification problem for data set D was solved using a weak classifier (h_1), generating inaccurate results slightly better than arbitrary speculation. AdaBoost assigns a weight to the error generated by h_1 in the existing D, and generates a new data set W_1 to improve h_1. The process to obtain h_2, which is the improved classifier, was repeated using data set W_1. Subsequently, AdaBoost integrated a number of weak classifiers, providing a stronger classifier than a single classifier.

2.2.2.2. Bagging Bagging is one of the ensemble methods. It is an algorithm that improves model performance by randomly generating a number of predictors to create an aggregated predictor. In other words, Bagging consists of bootstrap and aggregating. The bootstrap process extracts a random subset in the data set, and the aggregating process learns each classifier in the extracted subset and combines them to generate a strong classifier [17–19].

2.2.2.3. Random Forest Random Forest is one of the most powerful machine learning methods. As a variation of Bagging, this algorithm implements Bagging for the decision tree [20]. Similar to Bagging, Random Forest generates a number of random subsets from the training data set using bootstrap. Then, a number of independent decision trees are generated using each random subset to create aggregated trees. Accordingly, a voting process was performed using the value selected by numerous trees among the aggregated trees to derive a result [21].

2.3. Modeling procedure

2.3.1. Variable selection

We classified response variables, cyanobacteria cell count and chlorophyll-a concentration, into the non-exceedance class when they did not exceed the "advisory" level (1,000 cells/mL) of the cyanobacteria bloom alert system and the lake water quality standard of 14 mg/m^3; otherwise, they were classified into the exceedance class. Water quality data on water temperature, flow rate, sampling month (to account for the seasonal factor), sampling station (to account for the spatial factor), total precipitation over 7 d, and irradiance, which showed a significant correlation with the cyanobacteria cell count and chlorophyll-a concentration, were collected through correlation analysis and used as prediction variables.

2.3.2. Data preprocessing

Data imbalance can lead to poor prediction performance for the minority class when the number of samples is relatively small; for example, the cyanobacteria cell count and chlorophyll-a concentration may exceed the standard. The data set used in this study included 1,437 samples corresponding to the non-exceedance class and 1,656 samples corresponding to the exceedance class for chlorophyll-a, indicating no imbalance in the data set. In the case of the data set for cyanobacteria count, 2,893 and 200 samples belonged to the non-exceedance class and exceedance class respectively, indicating a high degree of imbalance.

We did not perform preprocessing on the chlorophyll-a data (which showed no imbalance), while the representative preprocessing method, synthetic minority over-sampling technique (SMOTE), was used to mitigate the imbalance problem in the cyanobacteria cell count data. The SMOTE is a combination of under-sampling and over-sampling methods. It has been used to generate random points where the difference between the points equals the difference between the selected minority sample and its neighbor, multiplied by a random value between 0 and 1 after selecting the minority sample in the feature space (instead of the entire data space), and randomly selecting a k value from the nearest neighbors.

The synthetic sample generated by this process exists as a random point on a straight line connecting the minority sample and its nearest neighbor, which makes the application of the synthetic sample more logical [22]. SMOTE was applied to the training data set for modeling the classes of cyanobacteria cell count. The under-sampling parameter and over-sampling parameter was set at 200 and 500, respectively. After the SMOTE application, the ratio between the non-exceedance and exceedance classes in the training data set was 0.62:0.38, indicating relaxed imbalance, compared with 0.93:0.07 before the application.

2.3.3. Model training and validation

To compare the performance of the decision tree algorithms, the entire data set was randomly divided into training data set and test data set in the ratio of 7:3. The training data set was used for calibrating the single and ensemble tree models and the test data set was used to validate the prediction performance of the models. The prediction performance was compared by repeating the above process 10 times to improve the reliability of the analysis (Fig. 2). The decision tree models were fitted in the R programming language [23] Packages *rpart*, *adabag*, *adaboost*, *randomForest*, and *DMwR* were used for implementing CART, Bagging, AdaBoost, Random Forest, and SMOTE, respectively. The parameter settings of each method are shown in Table 1.

2.3.4. Model performance evaluation and comparison

The prediction performance was evaluated based on multiple indices: accuracy, sensitivity, specificity, and area under curve (AUC) of receiver operating characteristic. The accuracy is the ratio of the number of total samples to the number of correctly classified samples. Sensitivity is the proportion of exceedance (positive) samples that are correctly predicted as the exceedance class, while specificity is the proportion of non-exceedance (negative) samples that are correctly predicted as the non-exceedance class. AUC evaluates the model performance at all possible combinations of sensitivity and specificity.

The relative importance among the predictors in each decision tree model was calculated based on the Mean Gini Decrease (MGD). The MGD is the average value of the difference in the Gini index (impurity) of the parent node and the child node. The decrease in the Gini value becomes larger as the impurity levels of the predictor used as classification criteria decrease. In other words, the predictor that results in a larger increase in MGD is interpreted as more important in predicting the class of response variable.

3. Results

3.1. Chlorophyll-a models

The single decision tree model, CART, showed a prediction accuracy of 71.7%, specificity of 63.3%, sensitivity of 78.9%, and AUC of 71.1% (Fig. 3(a)). As illustrated in Fig. 4, the validation results from the single model capture the environmental

Table 1
Parameter setting for each algorithms

Algorithm	Parameters
CART	Method = class
Random Forest	Number of trees to grow, n_{tree} = 1,000
Bagging	Number of iteration, m_{final} = 100
	Method = class
AadaBoost	Method = class
SMOTE	Number of nearest neighbors k = 100
	Number that drives under-sampling = 200
	Number that drive over-sampling = 500

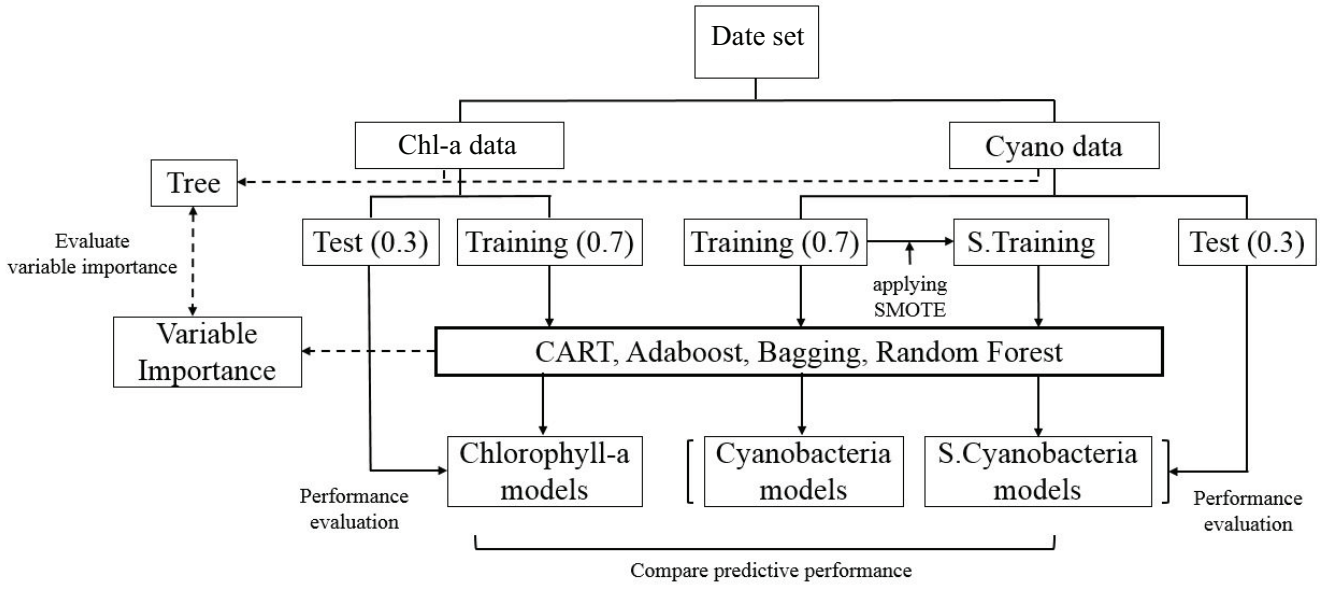
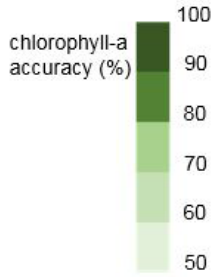


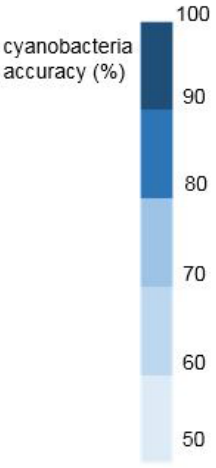
Fig. 2. Modeling procedure used in this study. The procedure above was iterated 10 times. S. cyanobacteria model: cyanobacterial model after SMOTE.

(a)



		CART	Random Forest	Bagging	AdaBoost
Chloro-phyll-a	SP	63.3%	77.8%	64.4%	76.0%
	SE	78.9%	86.6%	81.3%	82.2%
	ACC	71.7%	82.5%	73.7%	79.3%
	AUC	71.1%	82.2%	72.8%	79.1%

(b)



		CART	Random Forest	Bagging	AdaBoost
Cyano-bacteria before SMOTE	SP	98.5%	99.3%	98.6%	99.0%
	SE	67.4%	78.5%	72.4%	78.7%
	ACC	96.8%	98.0%	97.0%	97.8%
	AUC	84.0%	88.9%	85.5%	88.9%
Cyano-bacteria after SMOTE	SP	92.5%	96.9%	93.0%	96.7%
	SE	88.9%	90.8%	89.5%	89.7%
	ACC	92.2%	96.5%	92.8%	96.3%
	AUC	90.7%	93.8%	91.3%	93.2%

Fig. 3. The results of the performance evaluation for (a) chlorophyll-a concentration and (b) cyanobacteria cell count models. SP, specificity; SE, sensitivity; ACC, accuracy; and AUC, area under curve.

conditions under which chlorophyll-a concentration is likely to exceed the standard level. As compared with threshold values denoted in the tree, lower flow, higher temperature (or lower temperature in the case of higher flow), and lower precipitation, in combination, would induce high chlorophyll-a concentration that exceeds the standard. The predictors and conditions that determine whether the resultant class would be exceedance or non-exceedance varied by month (Fig. 4).

Ensemble decision tree models showed higher performance than the single model. Among them, Random Forest showed the highest prediction accuracy of 82.5%, followed by AdaBoost (79.3%), and Bagging (73.7%; Fig. 3(a)). Sensitivity was higher than specificity for all ensemble methods. Random Forest showed the most outstanding performance, with a specificity of 77.8% and sensitivity of 86.6%. AdaBoost and Bagging exhibited similar sensitivities. Meanwhile, specificity was 76.0% for AdaBoost and 64.4% for Bagging. Bagging did not show a significant improvement in specificity compared with CART. Similar to other indicators, the AUC for Random Forest was the highest, and the AUC for Bagging was the lowest (Fig. 3(a)).

For predicting chlorophyll-a classes, flow rate was the predictor with the highest importance in all ensemble models, while it was the predictor with the second highest importance in the single model (Fig. 5). Sampling month was ranked as the most or the second most important predictor in all models except for Random Forest, where sampling month was ranked as the least important predictor for chlorophyll-a classes.

3.2. Cyanobacteria cell count models

Overall, modeling performance of cyanobacteria was better than that of chlorophyll-a; the single decision tree model showed an accuracy of 96.8%, specificity of 98.5%, sensitivity of 67.4%, and AUC of 84.0% for predicting the classes of cyanobacteria cell count without applying SMOTE (Fig. 3(b)). Similarly, as seen in the chlorophyll-a model, the ensemble models showed improved performance as compared with the single model. Among the ensemble models, Random Forest indicated the highest prediction accuracy of 98.0%, and the remaining models showed similarly high prediction accuracies. In all ensemble methods, sensitivity was approximately 20% higher than specificity. Random Forest exhibited the highest specificity of 99.3%, while AdaBoost showed the highest sensitivity of 78.7%. The sensitivity of all ensemble methods was higher than that of the single model by at least

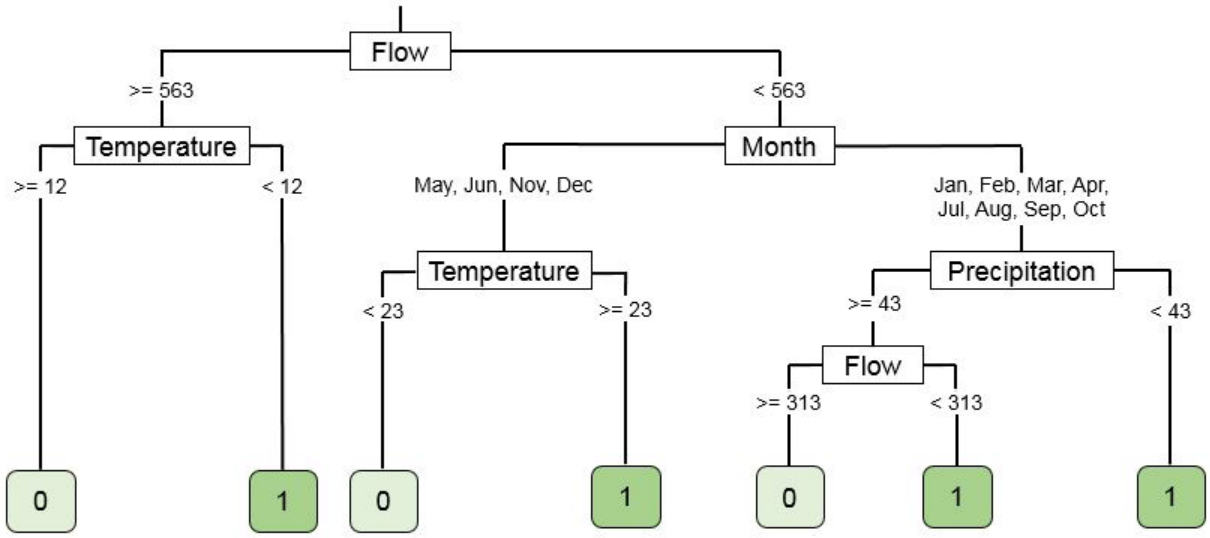


Fig. 4. Tree representation derived from single CART for chlorophyll-a. The tree was constructed using the entire data set.

		C.Precipitation	Flow	Irradiance	Month	Station.no	Temperature
Chloro-phyll-a	CART	4	2	3	1	6	5
	Random Forest	5	1	3	6	4	2
	Bagging	6	1	4	2	5	3
	Adaboost	6	1	5	2	4	3
Cyano-bacteria	CART	5	2	4	3	6	1
	Random Forest	5	1	4	3	6	2
	Bagging	5	2	6	3	4	1
	AdaBoost	5	3	4	6	2	1
Cyano-Bacteria after SMOTE	CART	6	2	4	3	5	1
	Random Forest	5	2	4	3	6	1
	Bagging	5	2	4	3	6	1
	AdaBoost	5	2	4	6	3	1

Fig. 5. Variable importance for predicting chlorophyll-a, cyanobacteria cell count, and cyanobacteria cell count after applying SMOTE. The numbers and colors in the table denote the importance of the variables by rank. For example, Rank 1 and the color black denote the variable with the highest importance. C. precipitation: weekly total precipitation.

10%. Similar to other evaluation indices, AUC of all ensemble methods was higher than that of the single model.

When SMOTE was applied to the cyanobacteria models, the specificity slightly decreased in all models (Fig. 3(b)). These decreases were more than compensated by the increases in sensitivity so that AUC, a combined result of specificity and sensitivity, was improved in all models. After SMOTE application, as before, all ensemble models exhibited improved performance for all evaluation indices than the single model, although the sensitivity difference between the single and ensemble models decreased.

Before SMOTE application, water temperature was ranked as the most important predictor in all but the Random Forest model, in which water temperature was the predictor with the second highest importance (Fig. 5). In contrast, after SMOTE application, water temperature and flow rate were unanimously indicated as the most and second most important predictors, respectively, by all models.

Fig. 6 shows the validation results of the single CART model after SMOTE application, depicting the environmental conditions under which cyanobacteria cell count is likely to exceed the advisory threshold value. The constructed tree structure for cyanobacteria was more complex than that for chlorophyll-a. Despite the complexity, the tree can be interpreted that the exceedance of advisory level was generally associated with high temperature, low flow, low precipitation, and high irradiance. In addition, cyanobacteria cell count was likely to exceed the advisory level during warm months (July–October) at the sampling stations downstream of the river (Stations 6–10) when other conditions were met.

4. Discussion and conclusions

We constructed prediction models for classifying chlorophyll-a and cyanobacteria into exceedance vs. non-exceedance classes using single and ensemble decision tree methods. The prediction performance of the single and ensemble methods was compared. The effect of using SMOTE, whereby the degree of imbalance for cyanobacteria data set was mitigated, on prediction performance was assessed.

Accuracy, the proportion of correctly classified positive and negative classes, is commonly used as an index of performance evaluation. The cyanobacteria models in this study showed slight decreases in model accuracy after SMOTE implementation. However, the use of accuracy might not be suitable when the data are imbalanced. Although the imbalanced data set leads to low predictability for the positive (minority) class, it can result in high accuracy [22]. In contrast, AUC accounts for both true positive rate and false positive error rate, constituting a reliable evaluation index especially when the data are imbalanced. The model performance is considered "good" if 70% < AUC ≤ 90%, and "excellent" if AUC > 90% [24].

In this study, the AUC of the single decision tree model for the chlorophyll-a model without data imbalance was 71.1%. When ensemble models were implemented, AUC improved by 1%–10%, indicating a "good" performance (Fig. 3(a)). The AUC of the single model for cyanobacteria with data imbalance was 84.0%. AUC was enhanced up to 5% by using ensemble models, indicating a "good" performance (Fig. 3(b)). The ability of predicting complex phenomena using a single decision tree model tends to be unsatisfactory, because an increasing number of splits involve larger variations and uncertainty of the model. In contrast, since the ensemble models are designed to learn subsets of data iteratively and perform the classification through majority voting, model variance is generally low and prediction performance appears to be superior compared with that of a single decision tree [9,25].

Ensemble approaches have received growing attention since the mid-2000s as they were proved to exhibit enhanced prediction performance [13]. Nevertheless, they share a similarity with single tree models, in that imbalanced data can lower the prediction power of both single and ensemble models for the minority class even though the overall accuracy is high [13,26]. In our study, when the data imbalance was not resolved, despite high accuracy of the models, the sensitivity, which calculates the proportion of correctly classifying the exceedance (minority) class for cyanobacteria cell count, was significantly low relative to specificity, the proportion of correctly classifying the non-exceedance (major) class (Fig. 3(b)). When the imbalance in cyanobacteria data was addressed by applying SMOTE, the specificity and accuracy of single and ensemble models slightly decreased in exchange for increasing sensitivity and AUC (Fig. 3(b)). Noticeably, the AUC results indicated that the performance of all models shifted from "good" to "excellent" after SMOTE application.

The effects of SMOTE application on changing specificity, sensitivity, and AUC are consistent with a previous finding. Ramezankhani et al. [27]applied SMOTE for the prediction of type 2 diabetes, which constituted only 510 samples among a total of 4,652 samples. After SMOTE implementation through oversampling of the minority class, they reported the effect of using SMOTE before decision tree on increasing sensitivity from 0.215 to 0.726, and decreasing specificity from 0.992 to 0.802 and accuracy from 0.907 to 0.794, while AUC was not used as an evaluation index.

Previous studies reported that cyanobacteria blooms are more likely to occur with increasing temperature [28–32].

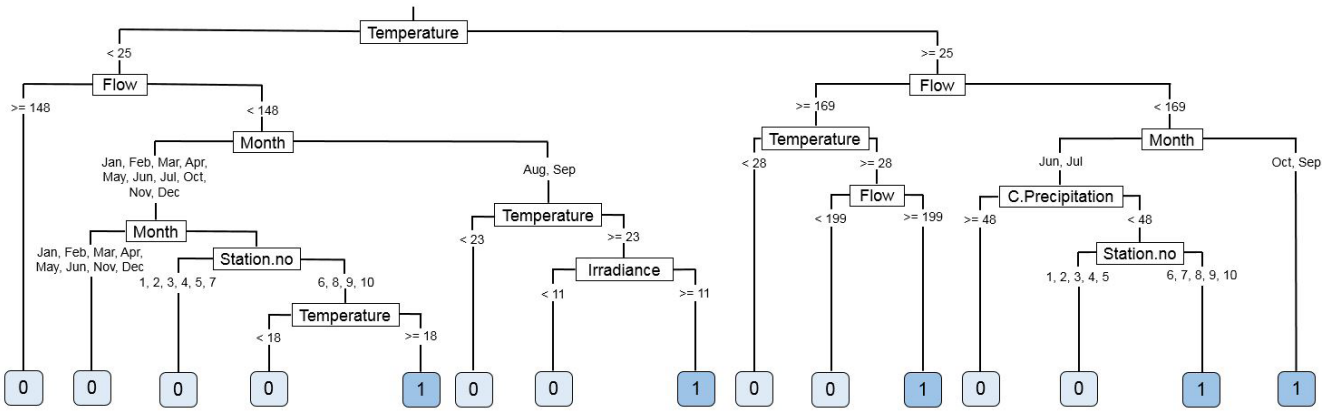


Fig. 6. Tree representation derived from single CART for cyanobacteria cell count. The tree was constructed using the entire data set.

High temperature directly affects the growth of cyanobacteria species positively, while indirectly facilitating cyanobacteria blooms by prolonging and strengthening thermal stratification [28,33]. In the case of flow rate, a cyanobacteria bloom is likely to occur when the discharge flow of the dam reduces because of decreasing flow rates. Hence, maintaining high flow rates, which cause the flushing effect, is considered as a measure to control such blooms [34,35]. A report from the Seoul Metropolitan Government informed that during the investigated years, 2008, 2012, 2014, and 2015, cyanobacteria bloom alerts for the lower Han River were issued in summer months (August–November) [36]. Cyanobacteria blooms occurred most severely in the year 2015, when the mean annual water temperature was highest and the discharge rate of Paldang Dam was lowest. In that year, the bloom alert persisted from August to November [36].

Based on the variable importance analysis, our results confirm that, regardless of the algorithms used, temperature and flow are the variables with the highest importance in predicting the level of cyanobacteria cell count (Fig. 5). Moreover, the graphical representation of tree derived from CART indicates that temperature and flow, located at the top of the tree, interplay with seasonal, spatial, or hydrological factors to induce exceedance or non-exceedance classes of cyanobacteria cell count (Fig. 6). The variance importance for chlorophyll-a was not as clear as that for cyanobacteria cell count.

Similar to cyanobacteria prediction, flow was a unanimously important variable in predicting chlorophyll-a classes, whereas the importance of temperature was lower and varied by model (Fig. 5). The tree for chlorophyll-a demonstrates that the exceedance of standard level is likely to occur either at high temperature (≥23°C) or low temperature (<12°C; Fig. 4). It implies that the dominance of other algal type than cyanobacteria, possibly diatoms, would be an explanation for high chlorophyll-a concentration at low temperature. Seasonality, rather than temperature, may play a critical role in predicting chlorophyll-a classes (Figs. 4 and 5). The tree for chlorophyll-a indicates that the predictor variable that determine chlorophyll-a classes may vary by month (Fig. 4).

In this study, we constructed the single and ensemble classification tree models that predict cyanobacteria blooms in the lower Han River. Although we used freely available data, such as the data provided by the Ministry of Environment and the Korea Meteorological Administration, the constructed models exhibited good to excellent prediction performance, promising to act as an aid tool for managing the cyanobacteria bloom alert system. We note that the presence of typically occurring issues for observational data, such as lack of sample size and imbalance between classes, can be resolved by using suitable techniques, in this study, ensemble approaches and SMOTE.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP; No. NRF-2016R1C1B1014395).

References

[1] H.W. Paerl, V.J. Paul, Climate change: links to global expansion of harmful cyanobacteria, Water Res., 46 (2012) 1349–1363.
[2] G.A. Codd, Cyanobacterial toxins, the perception of water quality, and the prioritisation of eutrophication control, Ecol. Eng., 16 (2000) 51–60.
[3] F. Recknagel, Applications of machine learning to ecological modelling, Ecol. Modell., 146 (2001) 303–310.
[4] N. Muttil, K. Chau, Neural network and genetic programming for modelling coastal algal blooms, Int. J. Environ. Pollut., 28 (2006) 223–238.
[5] N. Jung, I. Popescu, P. Kelderman, D.P. Solomatine, R.K. Price, Application of model trees and other machine learning techniques for algal growth prediction in Yongdam reservoir, Republic of Korea, J. Hydroinf., 12 (2010) 262–274.
[6] G.G. Moisen, Classification and Regression Trees, Encyclopedia of Ecology, Volume 1, Elsevier, Oxford, U.K, 2008, pp. 582–588.
[7] G. De'ath, K.E. Fabricius, Classification and regression trees: a powerful yet simple technique for ecological data analysis, Ecology, 81 (2000) 3178–3192.
[8] A. Peretyatko, S. Teissier, S. De Backer, L. Triest, Classification trees as a tool for predicting cyanobacterial blooms, Hydrobiologia, 689 (2012) 131–146.
[9] M. Rodrigues, J. de la Riva, An insight into machine-learning algorithms to model human-caused wildfire occurrence, Environ. Modell. Software, 57 (2014) 192–201.
[10] H. He, E.A. Garcia, Learning from imbalanced data, IEEE Trans. Knowl. Data Eng., 21 (2009) 1263–1284.
[11] S. Ertekin, J. Huang, L. Bottou, L. Giles, Learning on the Border: Active Learning in Imbalanced Data Classification, Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, ACM, Lisbon, Portugal, 2007, pp. 127–136.
[12] A. Estabrooks, T. Jo, N. Japkowicz, A multiple resampling method for learning from imbalanced data sets, Comput. Intell., 20 (2004) 18–36.
[13] B. Gong, J. Ordieres-Meré, Prediction of daily maximum ozone threshold exceedances by preprocessing and ensemble artificial intelligence techniques: case study of Hong Kong, Environ. Modell. Software, 84 (2016) 290–303.
[14] M.Y. Suh, B.H. Kim, K.S. Bae, Fluctuation of environmental factors and dynamics of phytoplankton communities in lower part of the Han River, Korean J. Ecol. Environ., 40 (2007) 395–402.
[15] T.K. Kim, J.H. Choi, K.J. Lee, Y.B. Kim, S.J. Yu, Study on introduction to predicting indicator of cyanobacteria dominance in algae bloom warning system of Hangang Basin, J. Korean Soc. Environ. Eng., 36 (2014) 378–385.
[16] X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A. Ng, B. Liu, S.Y. Philip, Top 10 algorithms in data mining, Knowl. Inf. Syst., 14 (2008) 1–37.
[17] C.D. Sutton, Classification and Regression Trees, Bagging, and Boosting, Handbook of Statistics, Elsevier, Vol. 24, 2005, pp. 303–329.
[18] L. Rokach, Ensemble-based classifiers, Artif. Intell. Rev., 33 (2010) 1–39.
[19] L. Breiman, Bagging Predictors, Machine Learning, Vol. 24, 1996, pp. 123–140.
[20] L. Breiman, Random Forests, Machine Learning, Vol. 45, 2001, pp. 5–32.
[21] A. Liaw, M. Wiener, Classification and Regression by randomForest, Vol. 2, 2002, pp. 18–22.
[22] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res., 16 (2002) 321–357.
[23] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2016, Available at: https://www.R-project.org/.
[24] M. Greiner, D. Pfeiffer, R. Smith, Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests, Prev. Vet. Med., 45 (2000) 23–41.

[25] T.G. Dietterich, Ensemble Learning, The Handbook of Brain Theory and Neural Networks, MIT Press, Cambridge, England, Second Edition, 2002, pp. 405–408.
[26] Y. Zhang, M. Bocquet, V. Mallet, C. Seigneur, A. Baklanov, Real-time air quality forecasting, part I: history, techniques, and current status, Atmos. Environ., 60 (2012) 632–655.
[27] A. Ramezankhani, O. Pournik, J. Shahrabi, F. Azizi, F. Hadaegh, D. Khalili, The impact of oversampling with SMOTE on the performance of 3 classifiers in prediction of type 2 diabetes, Med. Decis. Making, 36 (2016) 137–144.
[28] B.W. Ibelings, M. Vonk, H.F. Los, D.T. van der Molen, W.M. Mooij, Fuzzy modeling of cyanobacterial surface waterblooms: validation with NOAA-AVHRR satellite images, Ecol. Appl., 13 (2003) 1456–1472.
[29] K.D. Joehnk, J. Huisman, J. Sharples, B. Sommeijer, P.M. Visser, J.M. Stroom, Summer heatwaves promote blooms of harmful cyanobacteria, Global Change Biol., 14 (2008) 495–512.
[30] W.M. Mooij, S. Hülsmann, L.N. De Senerpont Domis, B.A. Nolet, P.L. Bodelier, P.C. Boers, L.M.D. Pires, H.J. Gons, B.W. Ibelings, R. Noordhuis, The impact of climate change on lakes in the Netherlands: a review, Aquat. Ecol., 39 (2005) 381–400.
[31] H.W. Paerl, J. Huisman, Blooms like it hot, Science, 320 (2008) 57–58.
[32] B.J. Robson, D.P. Hamilton, Summer flow event induces a cyanobacterial bloom in a seasonal Western Australian estuary, Mar. Freshwater Res., 54 (2003) 139–151.
[33] J. Elliott, I. Jones, S. Thackeray, Testing the sensitivity of phytoplankton communities to changes in water temperature and nutrient load, in a temperate lake, Hydrobiologia, 559 (2006) 401–411.
[34] K. Ha, M. Jang, G. Joo, Spatial and temporal dynamics of phytoplankton communities along a regulated river system, the Nakdong River, Korea, Hydrobiologia, 470 (2002) 235–245.
[35] K. Jeong, D. Kim, G. Joo, Delayed influence of dam storage and discharge on the determination of seasonal proliferations of *Microcystis aeruginosa* and *Stephanodiscus hantzschii* in a regulated river system of the lower Nakdong River (South Korea), Water Res., 41 (2007) 1269–1279.
[36] Water Environment Ecology Team, Han River (Water Recreational Activity Area) Algae Warning System Operation Result of 2016, Water Environment Research Department, Seoul Metropolitan Government, 2016.