Journal of Korean Society on Water Environment, Vol. 39, No. 1, pp. 1-8 (January, 2023)
pISSN 2289-0971 eISSN 2289-098X https://doi.org/10.15681/KSWE.2023.39.1.1

머신러닝 CatBoost 다중 분류 알고리즘을 이용한 조류 발생 예측 모형 성능 평가 연구

김준오[a] · 박정수[b,†]

국립한밭대학교 건설환경공학과

Evaluation of Multi-classification Model Performance for Algal Bloom Prediction Using CatBoost

Juneoh Kim[a] · Jungsu Park[b,†]

Department of Civil and Environmental Eng, Hanbat National University

(Received 22 November 2022, Revised 22 December 2022, Accepted 22 December 2022)

Abstract

Monitoring and prediction of water quality are essential for effective river pollution prevention and water quality management. In this study, a multi-classification model was developed to predict chlorophyll-a (Chl-*a*) level in rivers. A model was developed using CatBoost, a novel ensemble machine learning algorithm. The model was developed using hourly field monitoring data collected from January 1 to December 31, 2015. For model development, chl-a was classified into class 1 (Chl-$a \leq 10$ μg/L), class 2 ($10 <$ Chl-$a \leq 50$ μg/L), and class 3 (Chl-$a > 50$ μg/L), where the number of data used for the model training were 27,192, 11,031, and 511, respectively. The macro averages of precision, recall, and F1-score for the three classes were 0.58, 0.58, and 0.58, respectively, while the weighted averages were 0.89, 0.90, and 0.89, for precision, recall, and F1-score, respectively. The model showed relatively poor performance for class 3 where the number of observations was much smaller compared to the other two classes. The imbalance of data distribution among the three classes was resolved by using the synthetic minority over-sampling technique (SMOTE) algorithm, where the number of data used for model training was evenly distributed as 26,868 for each class. The model performance was improved with the macro averages of precision, rcall, and F1-score of the three classes as 0.58, 0.70, and 0.59, respectively, while the weighted averages were 0.88, 0.84, and 0.86 after SMOTE application.

Key words : Algal bloom, CatBoost, Ensemble Machine learning, SMOTE, Water quality prediction

[a] 석사과정(Master Student), juneohkim@naver.com, https://orcid.org/0000-0001-7325-4018
[b] Corresponding author, 조교수(Assistant Professor), parkjs@hanbat.ac.kr, https://orcid.org/0000-0002-9780-6988

1. Introduction

기후 변화는 인류가 직면한 중요한 환경 문제이다. 지구의 기온은 계속적으로 상승할 것으로 전망되고 있고 우리나라 역시 기온의 상승이 지속되고 있다(Jung et al., 2002; Lee et al., 2012; NIMR, 2009; Solomon, 2007). 지구 온난화로 인한 기온상승과 점 · 비점오염원에 의한 영양염류의 유입 등의 요인으로 인해 하천의 부영양화로 상수원이 오염되고 있고 매년 하천 조류의 발생 빈도와 범위가 증가하고 있어 국민의 불안감도 높아지고 있다(Kwak, 2021).

효율적인 하천오염 방지와 수질관리를 위해서 수질의 변화와 상태를 실시간으로 추적하고 예측하는 것이 필요하고 최근에는 실시간으로 전송되는 데이터를 기반으로 하여 다양한 예측 모형을 이용한 수질 변화 예측 연구도 진행하고 있다. 머신러닝 모형은 입력 변수의 특성 등을 고려하여 모형이 목표하는 결과를 출력할 수 있어 시계열 데이터 분석, 언어 분석, 이미지 분석 등 넓은 분야에서 사용하고 있고 수질 분야에서의 활용도 점차 늘어가고 있다(Breiman, 2001; Chen and Guestrion, 2016; Lee et al., 2020; Lim and An, 2018; Uddameri et al., 2020).

Ensemble 머신러닝 모형은 weak learner로 불리는 개별 모형의 결과를 종합하여 예측의 정확도를 높일 수 있도록 구성된 모형으로 random forest (RF)와 gradient boosting decision tree (GBDT) 등이 대표적인 ensemble 머신러닝 모형이다(Sutton, 2005; Zhang et al., 2018). ensemble 머신러닝은 회귀(regression) 및 분류(classification) 모형에 모두 사용 가능하며 모형 구축을 진행할 때 충분한 양의 데이터를 확보하여 사용할 수 있는 경우 높은 예측성능을 보여 최근까지도 가장 널리 활용되는 머신러닝 모형 중 하나이다(Hollister et al., 2016).

CatBoost는 범주형 자료의 분석에 좋은 성능을 보이는 ensemble 머신러닝 모형으로 최근 수질 분야에서도 활용이 늘고 있다(Dorogush et al., 2018; Prokhorenkova et al., 2018). Zhao et al. (2022)은 CatBoost 알고리즘(algorithm)을 사용하여 다중 스펙트럼 이미지로 남조류 농도를 예측하는 모형을 구축하였으며, Nasir et al. (2022)은 수질의 오염도에 따라 수질이 좋은 경우와 오염된 경우의 2가지 경우로 분류하고 CatBoost를 활용하여 이를 예측하는 모형을 만들었다. Xin and Mou (2022)는 CatBoost 알고리즘을 수질 분류에 적용하여 수질에 영향을 미치는 요소 들을 찾고 요소들의 가중치를 구해 수질 분류 및 예측하는 연구를 수행하였다.

본 연구에서는 CatBoost 알고리즘을 이용하여, 우리나라 대표적인 취수원 중 하나인 대청호의 Chlorophyll-*a* (Chl-*a*)를 농도 범위별로 3개의 class로 분류하고 이를 예측하는 분류 모형을 구축하였으며, 대표적인 ensemble 머신러닝 알고리즘인 XGBoost (XGB) 및 LightGBM (LGBM)과 모형의 성능을 비교하였다. 또한, 분류 모형 구축 시 각 class에 속하는 입력자료의 불균형 해소가 모형 성능에 미치는 영향을 분석하기 위해 synthetic minority over-sampling technique (SMOTE) 알고리즘을 이용하여 각 class 별 입력자료 수를 균등하게 구성하여 성능의 변화를 확인하였다.

2. Materials and Methods

2.1 연구대상 지역 및 분석자료

연구대상 지역인 대청호는 국내 충청북도에 위치한 유역면적 3,204 km^2, 총저수량 14.9억m^3 의 대한민국에서 소양호와 충주호에 이어 세 번째 규모의 저수지로 금강 본류에 위치하고 있다. 대청호는 금강 유역의 최대 상수원으로써 대전과 청주를 비롯한 청주를 비롯한 충청지역에 매일 약 100만m^{3}의 생 · 공용수를 공급해주고 있으며 금강 중 · 하류 지역의 홍수조절 및 수력발전 등의 다목적으로 사용되고 있다(Fig. 1) (K-water, 2022).

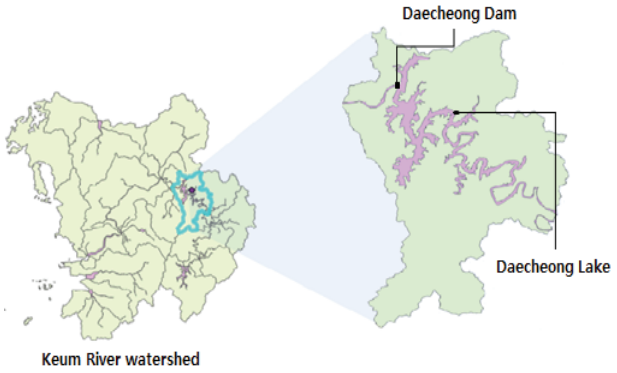


Fig. 1. Research site.

본 연구에서는 환경부 국립환경과학원의 물 환경정보시스템에서 제공되는 자동측정망 대청호지점(Site No. s03003)에서 2015년 1월 1일부터 2020년 12월 31일까지 측정된 시간별 총 47,372회의 측정 자료를 사용하였다(NIER, 2022).

2.2 입력자료 구축

측정된 수질 항목 중 수온(temperature, Temp), 수소이온농도(pH), 전기전도도(electronic conductivity, EC), 용존산소(dissolved oxygen, DO), 총질소(total nitrogen, TN), 총인(total phosphorus, TP) 의 6개 항목을 모형의 독립변수로 활용하였으며, 측정자료를 Chl-*a* 농도에 따라 분류한 3단계의 class를 종속변수로 사용하여 모형을 구축하였다. Chl-*a*의 분류 기준은 world health organization (WHO)의 “Thresholds of risk associated with potential exposure to cyanotoxins”를 참고하였다. WHO에 따르면 Chl-*a*를 low risk, moderate risk, high risk의 3단계로 나누고 있다. low risk (class 1)는 Chl-*a*가 10μg/L 이하, moderate risk (class 2)는 10μg/L 초과 50μg/L 이하, high risk (class 3)는 50μg/L 초과로 구분하고 있으며, 이를 근거로 class 분류를 진행하였다.

분석에 사용된 측정자료는 각 측정 항목별로 Temp 2.8%, pH 2.5%, EC 3.3%, DO 3.2%, TN 10.2%, TP 13.7%, Chl-*a* 3.5%의 결측값을 포함하고 있었으나 결측값이 발생한 시기가 대부분 강우가 발생하지 않는 10월부터 3월로 측정값의 변동이 크지 않는 구간으로 scikit-learn K-Neighbor (KNN)을 활용하여 결측값을 보정 하였다(Pedregosa et al., 2011).

KNN 알고리즘은 결측이 발생 된 측정일에서 가까운 k개의 자료를 이용하여 결측값에 대한 보정을 수행하며, 본 연구에서는 k 값을 3으로 지정하여 보정을 수행하였다.

모형의 구축을 위해 2015년 1월 1일부터 2019년 12월 31일까지의 자료를 학습(training) 데이터로, 2020년 1월 1일부터 12월 31일까지의 자료를 검증(testing) 데이터로 구성하여 training과 testing에 적용된 데이터의 비율을 각각 82% 및 18%로 구성하였다.

2.3 모형 구축

Ensemble 머신러닝 모형 중 하나인 CatBoost 모형은 gradient boosting machine (GBM)을 기반으로 한 머신러닝 알고리즘으로 ordered boosting을 사용하여 기존 GBM 알고리즘에서 각 단계 별 모형의 구축 시 해당 시점의 예측 대상이 되는 target 변수를 포함하여 모형의 과적합 (over fitting) 가능성이 높아지는 문제를 해결하였으며, 범주형 (categorical) 변수처리에 유용한 모형이다(Prokhorenkova et al., 2018).

본 연구에서는 대표적인 조류 발생 지표중 하나인 Chl-*a* 농도를 기준으로 3단계의 class를 구분하고 이를 종속변수로 하는 분류모형을 구축하였으며, 모형의 구축을 위해 범주형 모형의 구축에 우수한 성능을 보이는 것으로 알려진 CatBoost 모형을 이용하였다. 또한 우수한 성능으로 최근까지 널리 활용되고 있는 대표적인 ensemble 모형인 XGB와 LGBM을 이용하여 모형을 구축하여 성능을 비교하였다. XGB는 단일 모형인 weak learner로 구성되어 있으며, 전 단계의 weak learner의 결과를 다음 단계 weak learner의 연산에 활용하여 모형의 성능을 단계적으로 향상시키는 모형으로, 분류와 회귀 두 영역 모두에서 좋은 예측성능을 보여 물환경 분야에서도 지속적으로 활용되고 있다(Chen and Guestrin, 2016; Kim et al., 2020; Shin et al., 2021). LGBM은 Microsoft에서 개발한 GBDT 기반 앙상블 모형 중 하나로 모형을 구축할 때 사용되는 자료와 변수의 양을 감소시켜 연산 작업의 시간을 줄이고 성능을 향상시키는 내부 알고리즘을 사용한다. 이때 사용하는 알고리즘이 Gradient based one side sampling (GOSS)와 Exclusive feature bundling (EFB)이다. GOSS 알고리즘은 모형 적용 시 정보획득량 (information gain)의 절대값이 큰 순서대로 입력자료를 배열하고, 상대적으로 정보획득량이 큰 자료와 작은 자료에 서로 다른 가중치를 적용하여 선별적으로 입력 변수를 모형에 활용하는 알고리즘이며, EFB 알고리즘은 고차원의 변수들을 하나의 변수로 묶어 모형 구축에 사용되는 입력 변수를 줄이는 알고리즘이다(Ke et al., 2017; Ma et al., 2018). CatBoost, XGB, LGBM 세 모형의 최적화는 다양한 조건의 hyperparameter의 조합을 모형에 적용하여 최적의 성능을 얻을 수 있는 hyperparameter를 선정하는 gird search 방식을 이용하였으며, open source library인 scikit-learn을 이용하여 알고리즘을 구현하였다(Pedregosa et al., 2011). (Table 1)

Table 1. Hyperparameters used for the model optimization

Model	Parameter	Range
CatBoost	iterations	10, 30, 50, 100
	depth	3, 4, 5, 6, 7
	learning_rate	0.01, 0.03, 0.05, 0.1, 0.2
	l2_leaf_reg	2, 3, 5, 6
XGB	n_estimators	10, 30, 50, 100
	learning_rate	0.01, 0.03, 0.05, 0.1, 0.2
	max_depth	3, 4, 5, 6, 7
	min_child_weight	1, 3, 6, 10
LGBM	n_estimators	10, 30, 50, 100
	max_depth	3, 4, 5, 6, 7
	learning_rate	0.01, 0.03, 0.05, 0.1, 0.2
	num_leaves	100, 300, 500

2.4 SMOTE

SMOTE는 대표적인 오버 샘플링(Over sampling) 기법 중 하나이다. 분류 모형의 구축 시 낮은 비율로 존재하는 그룹 데이터를 KNN 알고리즘 등을 이용하여 새롭게 생성하여 class 간 데이터 불균형을 해결할 수 있는 방법이다(Chawla et al., 2002). 본 연구에서는 class 간 데이터의 불균형을 해결하기 위해 SMOTE를 이용하여 class 1～class 3의 자료 수에 대한 보정을 수행하였다.

2.5 모형 평가 기준

CatBoost 다중분류 모형의 성능 비교를 위해 혼동행렬 (Confusion Matrix)과 정밀도(Precision), 재현율(Recall) 및 F1 score 3개의 지표를 이용하여 성능을 비교 하였다(Eq 1, 2, 3).

$$\mathrm{Pr}ecision = \frac{TP}{TP+FP} \tag{1}$$

$$Recall = \frac{TP}{TP+FN} \tag{2}$$

$$F1\ Score = 2 \times \frac{\mathrm{Pr}ecision \times Recall}{\mathrm{Pr}ecision + Recall} \tag{3}$$

혼동행렬은 분할표 또는 오차 행렬(Error Matrix)로도 잘 알려져 있으며(Stehman, 1997). 머신러닝 모형을 구축하여 예측을 수행했을 때 실측값(observation)과 예측값(model prediction) 간의 분포를 행렬 형태로 나타내어, 머신러닝 분류 모형의 성능 평가에 널리 사용되는 지표이다. 아래 Table 2의 True는 실측값과 예측값이 일치하는 경우를 말하고, False는 실측값과 예측값이 일치하지 않는 경우를 나타내며, 실제 positive (P)를 P로 예측한 경우 True Positive (TP), 실제 N을 P로 예측한 경우 False Positive (FP) 실제 N을 N으로 예측한 경우 True Negative (TN)로 분류한다.

Table 2. Components of Confusion Matrix

Confusion Matrix		Model prediction	
		Positive (P)	Negative (N)
Observation	Positive (P)	True Positive (TP)	False Negative (FN)
	Negative (N)	False Positive (FP)	True Negative (TN)

모형 성능 평가에 사용된 지표의 평균값은 매크로 평균(macro average)과 가중평균(weighted average) 두 가지로 계산할 수 있다. macro average는 모든 class의 precision, Recall 별로 더하고 총 class 값으로 나눈 값으로 각 class에 속하는 데이터 수의 차이가 크지 않아 class 별 자료의 균형(balance)이 맞을 때 사용되는 평가지수이다. Weighted average는 전체 데이터 수 중 각 class에 속하는 데이터 수를 가중치로 하여 계산하는 방식으로, 데이터의 각 class에 속하는 데이터의 수의 차이가 커서 class 별 데이터가 불균형(imbalanced data)을 이루고 있는 경우에 macro average와 비교를 통해 데이터의 불균형에 따른 모형의 성능 차이를 확인할 수 있다.

3. Results and Discussion

3.1 입력자료의 특성 분석

본 연구에서 사용한 입력자료의 특성을 Table 3에 제시하였다. Class 별 수질 특성을 비교시 다른 수질항목에 비해 Temp의 차이가 상대적으로 큰 것으로 확인할 수 있었다. Class 1, 2와 3의 Temp 평균은 각각 15.3℃, 24.4℃, 27.8°C였으나, class 1의 경우 최소 2.2°C～최대 34.6°C class 2의 경우 최소 4.4°C～최대 34.8°C의 넓은 온도 범위를 보이는 반면 조류 발생이 가장 높은 구간인 class 3의 경우 최소 23.0°C～최대 33.9°C의 분포 범위를 보여 수온이 높아 상대적으로 조류 발생이 많은 하절기에 측정된 class임을 확인할 수 있다(Fig. 2).

Table 3. Characteristics of input variables according to classes

Class	Valiables	Average	Min.	25%	50%	75%	Max.	Standard deviation
Class 1	Temp	15.3	2.2	8.4	14.5	21.5	34.6	7.4
	pH	7.8	6.6	7.4	7.6	8.8	10.5	0.7
	EC	154.9	70.0	147.0	155.1	165.0	229.0	16.7
	DO	9.9	4.0	8.9	9.9	11.2	15.0	1.7
	TN	1.53	0.29	1.26	146.0	1.75	3.59	0.43
	TP	0.009	0.003	0.006	0.008	0.012	0.083	0.006
	Chl-*a*	4.6	0.1	3.0	4.1	6.0	10.0	2.2
Class 2	Temp	24.4	4.4	22.6	25.6	28.4	34.8	6.1
	pH	8.6	6.7	7.9	8.8	9.4	10.5	0.9
	EC	155.6	72.0	151.0	158.0	164.0	230.0	20.0
	DO	9.5	1.6	8.2	9.4	11.0	16.5	1.9
	TN	1.33	0.00	1.04	1.32	1.57	3.21	0.47
	TP	0.017	0.003	0.011	0.016	0.021	0.072	0.009
	Chl-*a*	17.8	10.0	12.7	15.2	20.4	50.0	17.8
Class 3	Temp	27.8	23.0	25.9	28.3	29.2	33.9	2.1
	pH	9.4	8.0	9.1	9.5	9.7	10.5	0.4
	EC	153.6	105.0	154.0	158.0	161.0	225.0	18.7
	DO	11.5	6.6	9.8	11.3	12.9	18.6	2.2
	TN	1.51	0.89	1.28	1.52	1.69	2.31	0.28
	TP	0.022	0.003	0.018	0.021	0.027	0.065	0.008
	Chl-*a*	76.6	50.1	56.5	66.6	81.8	281.9	34.7

3.2 CatBoost 모형 성능 분석

본 연구에서는 CatBoost 다중분류 알고리즘을 이용하고 grid search를 통해 모형의 최적화를 수행하여 하천 Chl-*a*를 예측하는 모형을 구축하였다. 최적화된 모형의 hyperparameter는 Table 4에 제시된 바와 같다. 분류 모형의 구축을 위해 Chl-*a* 값에 따라 데이터를 3단계의 class로 분류하였고 각 class 별 데이터의 수는 class 1 33,255개, class 2 13,491개, class 3 626개로 구성되었으며, 이중 모형의 training에 사용된 데이터 수

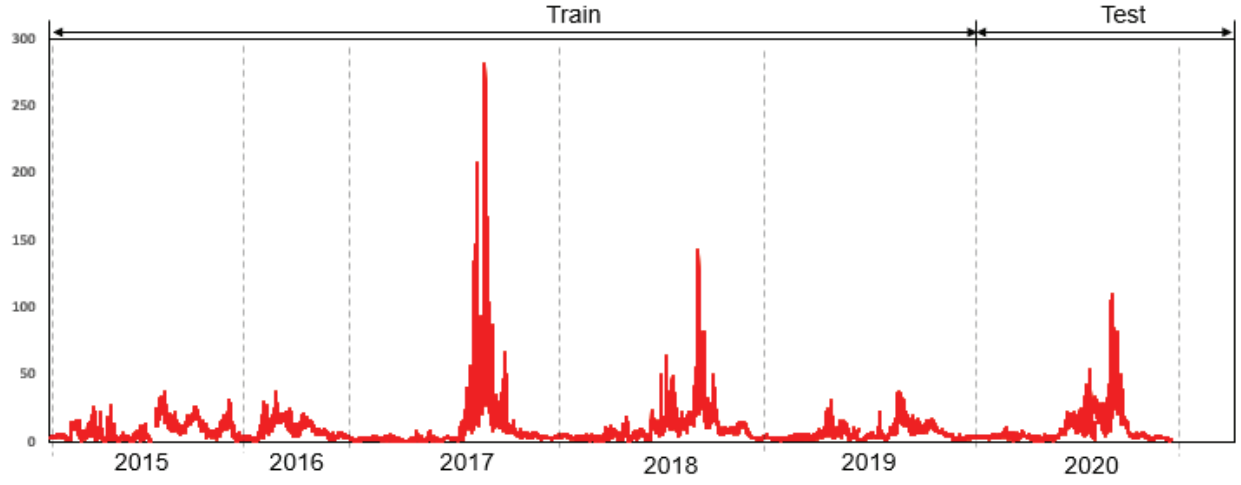


Fig. 2. Chl-*a* data used for model training and testing.

Table 4. Hyperparameters in optimized models

Model	Parameter	Optimum value
CatBoost	iterations	10
	depth	7
	learning_rate	0.05
	l2_leaf_reg	6
XGB	n_estimators	30
	learning_rate	0.01
	max_depth	5
	min_child_weight	10
LGBM	n_estimators	30
	max_depth	6
	learning_rate	0.2
	num_leaves	100

Table 5. CatBoost performance figures

Class	Precision	Recall	F1-score
Class 1	0.93	0.94	0.94
Class 2	0.79	0.80	0.80
Class 3	0.00	0.00	0.00
Macro average	0.58	0.58	0.58
Weighted average	0.89	0.90	0.89

는 class 1, class 2, class 3가 각각 27,192개, 11,031개, 511개이고, 모형 성능의 testing에 사용된 데이터 수는 class 1, class 2, class 3가 각각 6,063, 2,460개, 115개였다.

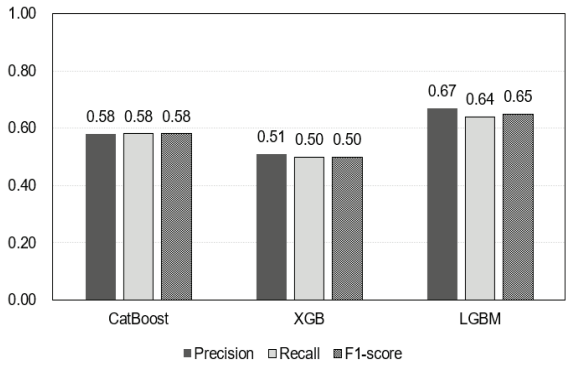
구축된 모형의 testing 데이터에 대한 Precision, Recall, F1-score는 class의 자료 수를 가중치로 고려한 weighted average의 경우 각각 0.89, 0.90, 0.89로 분석되었으며, 각 class 성능의 단순 평균값인 macro average의 경우 0.58, 0.58, 0.58로 분석되어, class 별로 자료의 불균형이 있어, class 3의 성능이 좋지 않아 macro average가 weighted average에 비해 낮은 성능을 보이는 것으로 산정된 것을 확인할 수 있었다. 각 class 별 성능은 class 1은 Precision, Recall, F1-score가 모두 0.9 이상의 좋은 성능을 보였으며, class 2는 Precision, Recall, F1-score가 각각 0.79, 0.80, 0.80으로 class 1에 비해 낮은 성능을 보였다(Table 5).

모형의 구축 결과 자료 수가 충분한 class 1과 2의 좋은 성능을 얻을 수 있었으나, 상대적으로 자료 수가 제한적인 class 3의 경우 모형이 class를 분류하지 못하는 것을 확인할 수 있었으며, 향후 추가적인 모니터링으로 충분한 자료가 확보되면 모형의 성능향상이 가능할 수 있을 것으로 판단된다.

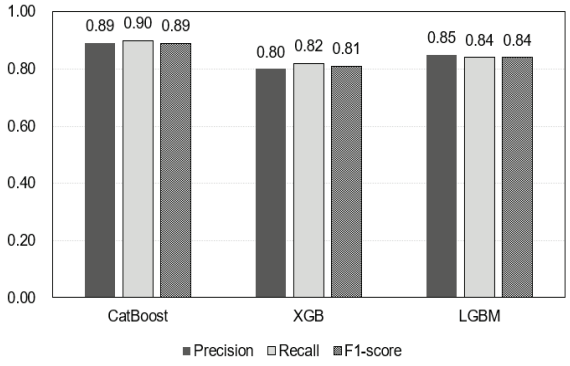
3.3 모형 성능 비교

본 연구에서는 CatBoost와 대표적인 ensemble 머신러닝 알고리즘인 XGB, LGBM 세 모형의 다중분류 모형 성능을 비교하였다(Fig. 3). 모형별 Precision, Recall, F1-score의 weighted average는 각각 CatBoost 0.89, 0.9, 0.89, XGB 0.80, 0.82, 0.81, LGBM 0.85, 0.84, 0.84였으며(Fig. 3(b)) 각 class의 자료수를 고려하지 않은 산술평균인 macro average는 각각 CatBoost 0.58, 0.58, 0.58, XGB 0.51, 0.50, 0.50, LGBM 0.67, 0.64, 0.65로 weighted average 보다 낮은 값을 보여(Fig. 3(a)), CatBoost, XGB, LGBM 모두 class 별로 자료의 불균형에 의한 영향을 받는 것으로 분석되어, 모형의 성능에 충분한 자료의 확보가 중요함을 확인할 수 있었다.

Class 별로는 class 1에 대해 CatBoost, XGB, LGBM의 Precision은 각각 0.93, 0.86, 0.93 Recall은 각각 0.94, 0.9, 0.86, F1-score는 각각 0.94, 0.88, 0.89로 모형에 따른 차이가 크지 않았으나 class 2의 경우 CatBoost, XGB, LGBM 의 Precision는 각각 0.79, 0.66, 0.64, Recall은 각각 0.80, 0.61, 0.79, F1-score는 각각 0.80, 0.63, 0.71로 CatBoost가 상대적으로 우수한 성능을 보이는 것을 확인할 수 있었다. 모형 구축에 사용된 class 중 자료의 수가 가장 작은 class 3의 경우 CatBoost와 XGB는 모두 Precision과 Recall이 0.00으로 나타나 모형이 분류를 정확히 하지 못하는 결과를 보였으나, LGBM은 Precision, Recall, F1-score가 각각 0.44, 0.28, 0.34로 분석되어, 입력자료의 수가 많지 않은 상황에서 상대적으로 안정적인 성능을 가지는 것을 확인할 수 있었다. Class 분류에 사용된 3가지 모형의 성능에 대한 비교를 통해 입력자료 수의 불균형이 모형의 성능에 영향을 미치는 것을 확인할 수 있었다.



(a) Macro average



(b) Weighted average

Fig. 3. Performance comparison by model.

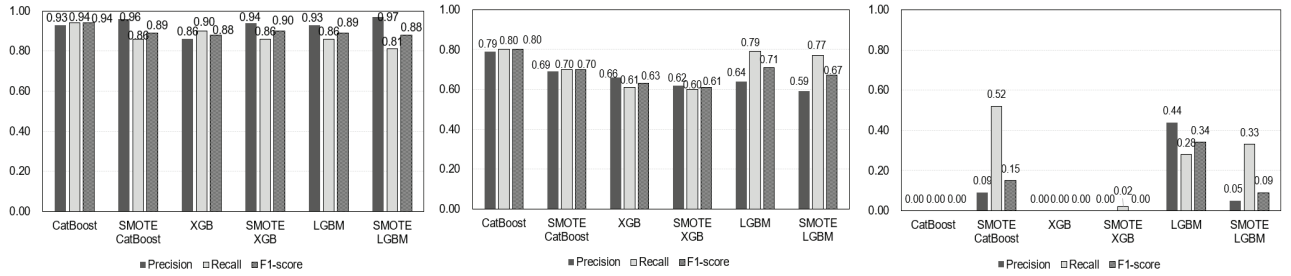
3.4 입력자료 불균형 해소에 따른 모형 성능 분석

구축된 머신러닝 모형의 성능분석 결과 입력자료의 수가 작은 class 3에서 모형 성능이 낮아지는 것으로 분석되어 class의 불균형이 모형의 학습에 영향을 주었음을 확인할 수 있었다. 따라서 모형의 학습에 사용된 입력자료의 class 간 자료 수의 균형을 맞추기 위해 SMOTE 알고리즘을 사용하여 각각의 class의 학습데이터 수를 26,868개로 동일하게 구성하고, SMOTE를 수행하지 않은 모형 구축 시와 동일한 hyperparameter 범위 조건에서 최적화를 수행하여 최적 모형을 구축하였다. SMOTE를 적용한 모형의 최적 hyperparameter를 Table 6에 제시하였다.

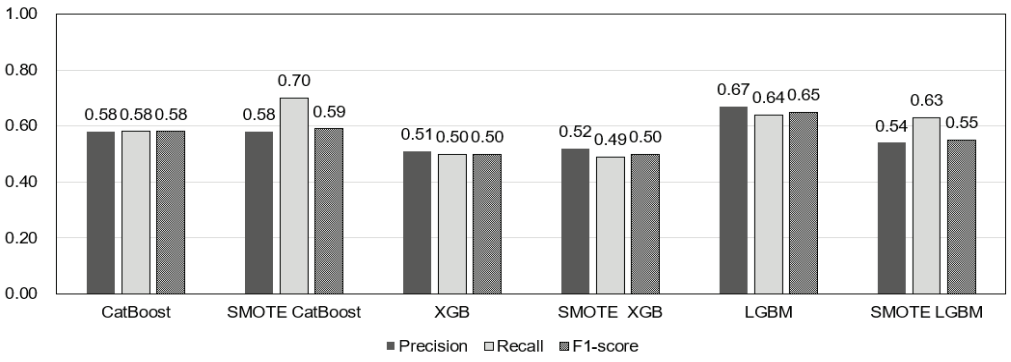
SMOTE 알고리즘의 적용 전후 CatBoost, XGB, LGBM 모형의 성능을 Fig. 4에 비교하였다. CatBoot와 XGB의 경우 Class 3에 해당하는 자료를 분류하지 못했으나, CatBoost의 경우 SMOTE 적용 후 class 3의 Precision, Recall, F1-score가 0.09, 0.52, 0.15로 개선되는 것을 확인하였다(Fig. 4(c)). XGB의 경우 Recall 값이 0.02로 약간의 개선이 있었으나 큰 차이

Table 6. Hyperparameters in optimized models using SMOTE

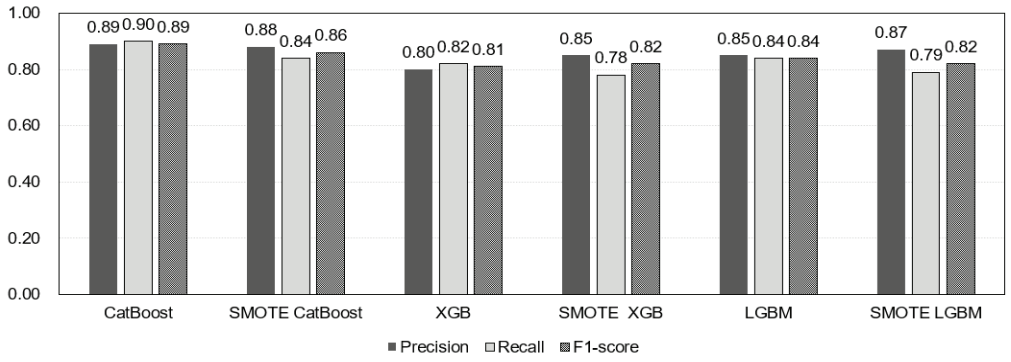
Model	Parameter	Optimum value
CatBoost	iterations	100
	depth	7
	learning_rate	0.01
	l2_leaf_reg	6
XGB	n_estimators	30
	learning_rate	0.01
	max_depth	6
	mind_child_weight	10
LGBM	n_estimators	100
	max_depth	3
	learning_rate	0.05
	num_leaves	100



(a) Class 1 (b) Class 2 (c) Class 3



(d) Macro average



(e) Weighted average

Fig. 4. Comparison of performance after SMOTE application.

가 없었다. LGBM은 SMOTE 적용 전 0.44, 0.28, 0.34에서 적용 후 0.05, 0.33, 0.09로 Recall은 개선되었으나 Precision 및 F1-score의 경우 오히려 성능이 낮아지는 경향을 보였다.

세 모형의 weighted average는 유사한 값을 보였으며 CatBoost의 SMOTE 적용 전후 Precision, Recall, F1-score 값은 각각 0.89, 0.90, 0.89 및 0.88, 0.84, 0.86으로 분석되는 등 SMOTE 적용 전후의 weight average의 변화는 크지 않은 것으로 확인되었다(Fig. 4(e)). 하지만, macro average의 경우 Precision, Recall, F1-score 값이 SMOTE 적용 전후 각각 0.58, 0.58, 0.58에서 0.58, 0.70, 0.59로 특히 Recall 값이 상대적으로 크게 개선되는 것을 확인할 수 있었다(Fig. 4(d)).

4. Conclusion

본 연구에서는 CatBoost 다중분류 알고리즘을 이용하여 구축된 데이터 세트의 Chl-*a*를 3단계의 class로 분류하고 예측하는 모형을 구축하였고 구축된 모형의 성능을 대표적인 ensemble 머신러닝 알고리즘인 XGB 및 LGBM을 이용해 구축된 모형과 성능을 비교하였다. CatBoost 모형은 Precision, Recall, F1-score가 class 1에 대하여 각각 0.93, 0.94, 0.94, class 2에 대하여 0.79, 0.8, 0.8로 분석되었다. weighted average와 macro average는 각각 0.89, 0.90, 0.89 와 0.58, 0.58 0.58로 학습과 테스트에 사용할 충분한 경우 안정적인 성능을 보이나, 자료의 수가 제한적인 class 3의 성능이 상대적으로 낮아 macro average 값이 낮아지는 경향을 확인할 수 있었다.

CatBoost, XGB, LGBM 모형의 비교 결과 class 1, 2와 weighted average 성능은 CatBoost 모형이 다른 두 모형에 비해 좋은 성능을 보였고 class 3과 macro average는 LGBM가 상대적으로 좋은 성능을 보이는 것으로 분석되었다.

머신러닝 모형의 성능은 입력자료의 구성 및 특성에 영향을 받게 된다. 분류 모형의 경우 모형의 training에 사용된 자료 중 해당 class에 포함되는 자료의 특성을 반영하여 모형이 구축되므로 해당 class의 특성을 학습할 수 있는 충분한 자료의 확보가 필요하며, 특정 class에 자료 수가 적게 포함되는 등 자료의 불균형이 있는 경우 모형의 성능이 저하될 수 있다. 본 연구에서는 SMOTE 알고리즘을 활용하여 데이터의 불균형을 해소하고 그 결과가 모형의 성능에 미치는 영향을 분석하였다. CatBoost의 경우 SMOTE 적용 전에는 class 3를 정확히 분류하지 못하는 결과를 보였으나 SMOTE 적용 후 precision, Recall, F1_score의 성능이 개선되었고 특히 재현율이 0에서 0.52로 향상되는 것을 확인하였다. Weight average와 macro average의 경우 SMOTE를 적용하기 전의 weighted average의 변화는 크지 않았으나 자료 수에 대한 가중치를 적용하지 않은 macro average는 SMOTE 전보다 개선되었고 특히 Recall 값의 상대적으로 크게 개선되는 것을 확인하였다. 향후 지속적인 모니터링으로 분류 모형의 구축시 각 class별로 모형이 학습할 수 있는 충분한 데이터를 확보하는 등 입력자료 불균형의 해결을 통해 모형의 성능향상이 가능할 것으로 판단된다.

Acknowledgment

1. 이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2022R1F1A1065518) (50%).
2. 본 결과물은 환경부의 재원으로 한국환경산업기술원의 환경시설 재난재해 대응기술개발사업의 지원을 받아 연구되었습니다 (2022002870001) (50%).

References

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

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique, *Journal of Artificial Intelligence Research*, 16, 321-357.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system, *In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining,* 785-794.

Dorogush, A. V., Ershov, V., and Gulin, A. (2018). CatBoost: Gradient boosting with categorical features support, *arXiv preprint arXiv*:1810.11363.

Hollister, J. W., Milstead, W. B., and Kreakie, B. J. (2016). Modeling lake trophic state: A random forest approach, *Ecosphere*, 7(3), e01321.

Jung, H. S., Choi, Y., Oh, J. H., and Lim, G. H. (2002). Recent trends in temperature and precipitation over South Korea, *International Journal of Climatology*, 22, 1327-1337.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree, *Advances in Neural Information Processing Systems*, 30.

Kim, Y., Choi, H., and Kim, S. (2020). A study on risk parity asset allocation model with XGBoo, *Journal of Intelligence and Information Systems*, 26(1), 135-149.

Kwak, J. (2021). A study on the 3-month prior prediction of Chl-a concentraion in the Daechong lake using hydrometeorological forecasting data, *Journal of Wetlands Research*, 23(2), 144-153. [Korean Literature]

K-water. (2022). *Mywater*, http://www.water.or.kr/ (Aug 4, 2022).

Lee, K. M., Baek, H. J., Park, S. H., Kang, H. S., and Cho, C. H. (2012). Future projection of changes in extreme temperatures using high resolution regional climate change scenario in the Republic of Korea, *Journal of the Korean Geographical Society*, 47(2), 208-225. [Korean Literature]

Lee, S. M., Park, K. D., and Kim, I. K. (2020). Comparison of machine learning algorithms for Chl-a prediction in the middle of Nakdong river (focusing on water quality and quantity factors), *Journal of Korean Socitey of Water and Wastewater*, 34(4), 277-288. [Korean Literature]

Lim, H. S. and An, H. U. (2018). Prediction of pollution loads

in Geum river using machine learning, *Proceedings of the Korea Water Resources Association Conference*, Korea Water Resources Association, 445. [Korean Literature]

Ma, X., Sha, J., Wang, D., Yu, Y., Yang, Q., and Niu, X. (2018). Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning, *Electronic Commerce Research and Applications*, 31, 24-39.

Nasir, N., Kansal, A., Alshaltone, O., Barneih, F., Sameer, M., Shanableh, A., and Al-Shamma'a, A. (2022). Water quality classification using machine learning algorithms, *Journal of Water Process Engineering*, 48, 102920.

National Institute of Environmental Research (NIER). (2022). *Water environmental information system*, https://water.nier.go.kr/web (Aug 4, 2022).

National Institute of Meteorological Research (NIMR). (2009). *Climate change in the Korean peninsula, present and future*, National Institute of Meteorological Research, Seoul. [Korean Literature]

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., and Dubourg, V. (2011). Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research,* 12, 2825-2830.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features, *Advances in Neural Information Processing Systems*, 31.

Shin, J. I., Park, J. S., and Shon, J. G. (2021). Prediction of semiconductor exposure process measurement results using XGBoost, *In Proceedings of the Korea Information Processing Society Conference*, Korea Information Processing Society, 505-508. [Korean Literature]

Solomon, S. (2007). *The physical science basis: Contribution of working group I to the fourth assessment report of the intergovernmental panel on climate change*, Intergovernmental Panel on Climate Change (IPCC), Climate change 2007, 996.

Stehman, S. V. (1997). Selecting and interpreting measures of thematic classification accuracy, *Remote Sensing of Environment*, 62(1), 77-89.

Sutton, C. D. (2005). *Classification and regression trees, bagging, and boosting*, Handbook of statistics, 24, 303-329.

Uddameri, V., Silva, A. L. B., Singaraju, S., Mohammadi, G., and Hernandez, E. A. (2020). Tree-based modeling methods to predict nitrate exceedances in the Ogallala aquifer in Texas, *Water*, 12, 1023.

Xin, L. and Mou, T. (2022). Research on the application of multimodal-based machine learning algorithms to water quality classification, *Wireless Communications and Mobile Computing*, 2022, 1-13.

Zhang, D., Qian, L., Mao, B., Huang, C., Huang, B., and Si, Y. (2018). A data-driven design for fault detection of wind turbines using random forests and XGboost, *IEEE Access,* 6, 21020-21031.

Zhao, X., Li, Y., Chen, Y., and Qiao, X. (2022). A method of cyanobacterial concentrations prediction using multispectral images, *Sustainability*, 14(19), 12784.