Journal of Korean Institute of Intelligent Systems
Vol. 33, No. 1, February 2023, pp. 29-36
http://dx.doi.org/10.5391/JKIIS.2023.33.1.29

ISSN(Print) 1976-9172
ISSN(Online) 2288-2324

머신러닝을 활용한 낙동강 유해남조류 발생 예측

A Study on the Predicting Harmful Cyanobacteria Algal Blooms Using Machine Learning Technology – The Nakdong River Case

송찬영*, 김주연**, 김예진***, 주해종****, 서재현*****†
Chan-Young Song, Ju-Yeon Kim, Ye-Jin Kim, Hae-Jong Joo, and Jae-Hyun Seo†

*한국외국어대학교 미디어커뮤니케이션 학부, **세종대학교 응용통계학과,
***서울여자대학교 수학과, ****동국대학교 컴퓨터공학과 교수,
*****†광주대학교 컴퓨터공학과 조교수
*Division of Media-Communication, HanKuk University of Foreign Studies
**Dept. of Applied Statistics, Sejong University
***Dept. of Mathematics, Seoul Women's University
****Professor, Dept. of Computer Engineering, Dongkuk University
*****†Assistant Professor, Dept. of Computer Engineering, Gwangju University

요 약

기후변화와 환경오염으로 인하여 낙동강 녹조 문제는 심화되고 있다. 따라서, 본 연구는 녹조 문제가 심각한 낙동강 지역의 유해남조류 세포수를 예측하는 모델을 개발한다. 주요 변수에는 과거수질자료, 댐 제원 정보와 풍속, 기상관측 데이터, 폐수처리장, 공장 현황 데이터를 결합하여 사용하였다. 특징 선택으로 Wrapper 기반의 Genetic Search를 사용하였으며, 예측 모델에는 Random Forest와 k-NN, SVM을 사용하였다. 예측 모델 성능은 Random Forest에서 0.880의 결정계수와 1.606의 RMSE로 가장 좋은 성능을 보였다. 본 연구는 pH, DO, BOD, COD 등의 변수를 제거하고, 새롭게 공장과 폐수처리장 위치 데이터를 써서 모델의 정확도를 높였다는 점에 의의가 있다.

키워드 : 녹조, 유해남조류, 머신러닝, 랜덤포레스트, *k*-NN, SVM

Received: Oct. 03, 2022
Revised : Oct. 17, 2022
Accepted: Nov. 01, 2022
†Corresponding author
(jhseo@gwangju.ac.kr)

Abstract

Because of climate change and environmental pollution, the problem of algae in the Nakdong River is getting worse. Therefore, this study develops a model to predict the number of harmful blue-green algae cells in the Nakdong River region where the problem of green algae is serious. For major variables, historical water quality data, dam information, wind speed, weather observation data, wastewater dispodal facility, and facility status data were combined and used. Wrapper based Genetic Search was used for feature selection, and Random Forest, *k*-NN, and SVM were used to predict models. The predictive model performance showed the best performance with a coefficient of determination of 0.880 and an RMSE of 1.606 in the Random Forest. This study is meaningful in terms of removing variables such as pH, DO, BOD, and COD and improving accuracy of the model by using facility and wastewater dispodal facility location data.

Key Words : Green Algae, Harmful-Cyanobacteria, Machine learning, Random Forest, *k*-NN, SVM

본 논문은 한국연구재단 기본연구 사업 (No. NRF-2020R1F1A1070363)과 2022년도 광주대학교 대학 연구비의 지원을 받아 수행되었음.

1. 서 론

지구온난화로 인한 기후변화와 환경오염으로 인하여 매년 여름마다 녹조 문제는 심화되고 있는 실정이다. 녹조를 일으키는 유해남조류세포는 청산가리의 100배가 넘는 마이크로시스틴이란 독소를 생성한다. 최근, 녹조 문제가 심각한 낙동강의 퇴적토와 논경지에서 마이크로시스틴이 나온 것으로 알려져 국민 건강과 국토 보존을 위협하고 있다. 녹조의 대표적인 발생 조건으로 영양분(특히, 인), 수온, 긴 체류 시간, 성층화 현상으로 인한 물의 안정화를 뽑을 수 있다[1]. 지구온난화로 기상이변으로 인한 수온 상승과 일부 지역에는 강수량이 적어 보의 체류시간도 많이 길어졌다. 또한, 공장이 내보내는 오염물질로 인하여 강물에 인이 있는 영양 염류가 포함되면서 한반도에 녹조화가 가속되고 있다.

녹조 원인 종의 출처와 발생 근원지를 알지 못하여 녹조에 대한 예방 및 대처가 미흡한 실정이다[2]. 녹조 예측이 가능하다면, 녹조 심화 예상 지역에 선제적 대응이 가능할 것이다. 따라서 본 연구에서는 녹조의 원인이 될 것으로 예상되는 변수들을 활용하여 녹조 발생 규모 예측 모델 개발을 목적으로 한다.

최근 유해남조류 예측 기술로 낙동강 지역의 보를 중심으로 진행되었다. 예측 모델은 주로 LSTM과 Random Forest를 중심으로 진행되었으며, pH와 DO(용존산소), BOD(생물화학적산소요구량), COD(화학적산소요구량)을 토대로 한 데이터에 연구별로 수문데이터 또는 기상데이터를 포함하여 유해남조류를 예측하는 연구들이 선행됐다.

하지만, 선행 연구에서 사용된 pH, DO, BOD, COD 변수는 생성된 유해남조류가 광합성을 하고 햇빛을 차단하였을 때 높아지는 변수로[3] 현상의 결과로 현상을 예측한다는 한계점이 있다. 따라서, 본 연구에서는 기존 연구에서 제시하지 않은 인근 지역의 폐수처리장과 공장의 수를 변수로 활용하여 예측 모델의 성능을 높이는 것을 목표로 한다.

2. 관련 연구

최근 연구 동향은 녹조 문제가 고조된 낙동강 지역의 보를 중심으로 한 연구가 많이 진행됐다. Kim, Min Seok 등[4]은 Chl-a를 예측하기 위해 LSTM을 포함한 딥러닝 기술을 이용하였다. 이를 위해 조류데이터(기온, 강수량, 풍속), 수문 데이터(수위, 총유입량, 총방류량)을 이용하였다. 이때, 조류 데이터는 MLP를 이용하여 특징점을 추출하고 기온 강수량 등 기상데이터와 수문데이터는 Bi-LSTM을 이용하여 특징점을 추출하였다. 그리고 Element-wise와 Product를 이용하여 Chl-a의 농도를 예측할 수 있도록 하였다.

Jung, W. S. 등[5]은 낙동강 지역의 8개 보 지점별 유해 남조류 발생의 주요 영향인자를 도출하고, 조류경보제 기반의 범주형 예측 모델을 개발하였다. 해당 연구에선 pH, DO, E.C, BOD, COD, T-P, Chl-a, 수온, SPI를 변수로 활용하였으며, 낙동강 중류 구간에서 DO와 E.C가 영향인자로 도출되었다. 중류 구간은 대규모 산업공단이 밀집되어 있는 곳으로 환경 기초시설의 배출량이 큰 영향을 끼치는 구간이다. 따라서, 본 연구의 결론으로 환

표 1. 유해남조류 예측 관련 선행 연구
Table 1. Prior research of harmful Cyanobacteria prediction

Author	Features	Algorithms	Research Purpose
Jung, W. S., Jo, B. G., Kim, Y. D., & Kim, S. E	*pH, E.C , DO, BOD, COD, T-P, Chl-a, Temperature, SPI*	Random Forest	Prediction of Harmful Cyanobacteria Algal Blooms in the Nakdong River' s Eight Weir area
Jung, W. S., Kim, S. E., and Kim, Y. D	*Outflow, pH, E.C , DO, BOD, COD, T-P, Chl-a, Temperature, SPI*	Random Forest	Prediction of Harmful Cyanobacteria Algal Blooms in The Main Stream of Nakdong River
Kim, S.-H., Park, J. H., and Kim, B. H.	Chl-a, TOC, PO4-P, T-P, N03-N, NH3-N, T-N, SS, COD, BOD, EC, DO, Water temperature, pH	ANN, RNN, LSTM	Prediction of Harmful Cyanobacteria Algal Blooms In YeongCheon dam Using Water Temperature Variables
Kim, Min Seok, Park, Hyung wook, Jo, Hyun Jung, and Kim Eun Joo.	Water temperature, pH , Do, T-P, Chl-a, Meteorogical data, Dam Control Management Infromation	LSTM	Prediction of Chl-a in The Nakdong River' s Weir
Jung, Y. J	pH, DO, BOD, COD, TOC SS, TN, TP, temperature	Xgboost, DecisionTreeRegressor, Random Forest	Test of Machine Learning Model for Water Quality Prediction

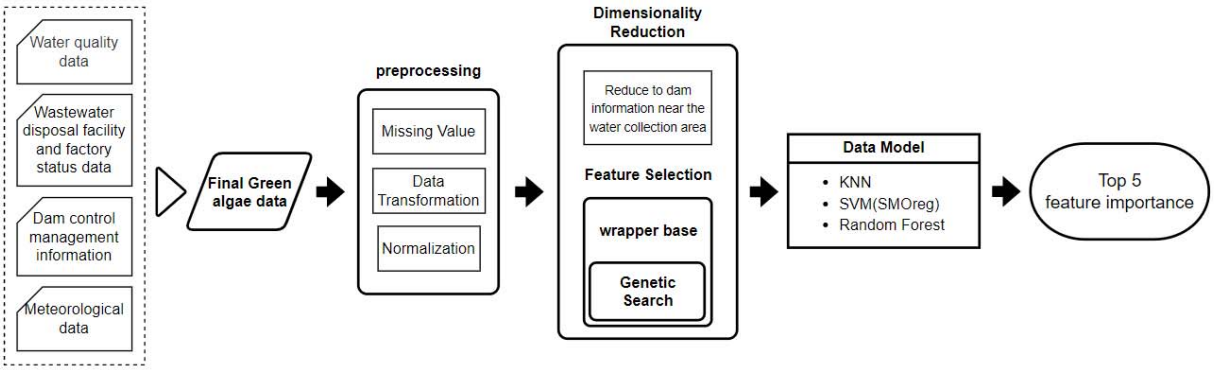

그림 1. 연구 순서도
Fig. 1. Research Flow Chart

경기초시설의 방류가 본류의 E.C를 증가하게 하고 남조류 발생을 촉진한 것으로 판단하였다.

Kim, S.-H. 등[6]은 보현산 댐과 영천 댐을 대상으로 유해 남조류 발생을 예측하는 연구를 진행하였다. 유해남조류 발생에 미치는 수질 인자로 수온과 총질수(T-N)이 공통적으로 높게 나왔으며, 인공신경망(ANN)을 이용한 유해남조류 발생 예측에서 결정계수 0.977의 값을 얻을 수 있었다.

Jung, W. S. 등[7]은 낙동강 본류 구간의 유해남조류 세포수와 변수(pH, 기온, spi, 가뭄지수 등)들을 의사결정나무를 이용하여 영향인자에 따른 남조류 발생조건을 정량적으로 분석하였다. 8개 보 모든 지점에서 기상학적 요인인 기온과 SPI 가뭄지수가 유의한 상관관계를 보였다고 결론지었다.

Jung, Y. J [8]은 한강, 낙동강, 금강, 영산강 4대강을 중심으로 유해남조류 세포수를 예측하는 실험을 진행하였다. 연구에서 수질데이터로 2015년 데이터를 이용하여 학습시킨 후 2020년 수질 농도를 예측 후 실험평가는 회귀모델 평가 지표인 RMSE를 사용하였다. 실험에는 pH, DO, BOD, COD, TOC, SS, TN, TP, 온도 총 9개의 파라미터를 사용하였고,. 실험 결과 DecisionTreeRegressor에서 RMSE 0.510, RandomForest엣 0.448, XGBoost에서 0.3991의 결과로 XGBoost에서 가장 우수한 성능을 보였다.

Table 1에서 녹조예측과 관련된 연구들을 연도, 데이터 셋, 사용 변수, 사용 알고리즘 별로 구분하여 비교한다

3. 연구 방법

3.1 사용 데이터셋

수집할 데이터는 조건인 영양분, 수온, 물의 안정화를 고려하여 구성한다. 첫째로, 유해남조류 세포수와 수온의 정보를 구한다. 영양염류 관련 변수로 대부분의 선행 연구에서는 pH와 DO를 변수로써 활용하였다. 하지만, pH는 유해남조류가 광합성을 하였을 때 높아지는 것으로[3] 녹조 현상을 예측할 때 사용할 변수로 부적절하다고 판단하였다. DO 또한 유해남조류가 햇빛을 차단하였을 때 용존 산소량이 적어지는 것으로 사용할 변수에서 제외하였다. 대신, 영양염류와 직접적인 관련이 되는 공단과 폐수처리장의 데이터를 수집한다.

유해남조류는 수중의 영양분, 성층화 현상으로 인한 물의 안정화, 높은 수온의 환경에서 발생할 확률이 높은 성질을 지닌다. 본 연구에선, 이런 특징을 알고리즘에 반영하기 위하여 유해남조류의 성질을 고려하여 다음과 같은 변수 설정을 하였다.

낙동강 지역의 수중에 영양분이 포함되는 주된 원인은 공장에서 배출하는 오염물질 때문이다. 공장의 개수와 폐수처리장의 개수를 알고리즘에 반영하도록 한다. 성층화 현상으로 인한 물의 안정화는 댐의 방류와 강수량에 크게 영향을 받는다. 댐의 제원 정보와 해당 채수 지역의 강우량과 해당 채수 지역에 영향을 주는 지역의 기상 데이터를 활용하여 각 알고리즘에 반영하도록 했다. 마지막으로 높은 수온은 과거 수질 자료에서의 수온 정보를 활용했다. 따라서, 각 모델의 기계 학습에서 유해남조류 특징에 기반하여 유해남조류가 발생하기 쉬운 환경을 분류하도록 하여 ml당 유해남조류 세포수를 예측한다.

3.2 데이터 수집 과정

먼저, 환경부 물환경 정보시스템(https://water.nier.go.kr/)이 제공하는 금강, 한강, 낙동강, 금강에서의 상수원 구간, 친수활동구간, 조류 관찰지점에서의 2016년부터 2022년 5월까지의 과거 수질 자료를 사용한다. 해당 데이터에서 수집된 변수는 목표 변수가 될 유해남조류세포수(cells)와 수온(temp), 채수지역(region)이 있다.

각 시도에서 제공하는 시도별 공장현황, 폐수처리장 데이터(https://data.go.kr) 중 한강, 낙동강이 지나는 지역의 데이터를 사용한다. 데이터 내의 공장과 폐수처리장 주소지를 Geo-coding을 통하여 위도와 경도로 변환한 뒤, 물환경 정보시스템에서의 채수지역과 각 공장 간의 거리를 Haversine Formula를 이용하여 계산한다. 채수 지역 3km, 5km, 7km, 10km 내의 공장의 개수를 계산하여 fac_3, fac_5, fac_7, fac_10 변수로 설정하고, 폐수처리장 개수 또한 ww_3, ww_5, ww_7, ww_10을 변수로 설정한다.

물의 방류량과 유입량을 측정하기 위하여 K-water (https://www.kwater.or.kr)에서 제공하는 댐 제원 정보를 사용한다. 데이터에는 댐수위(lowlevel), 강우량(prcptqy), 유입량(inflowqy), 총방류량(totdcwtrqy), 저수량(rsvwtqy), 저수율(rsvwtrt)의 정보가 있다. 채수지역을 기준으로 인근에 있는 댐 제원 정보를 채수지역과 매칭하였다.

마지막으로 기상청에서 제공하는 종관 기상 관측(https://data.kma.go.kr)을 사용한다. 일별로 측정된 풍속, 강수량, 평균기온 정보가 있다. 채수지역 인근의 기상관측소의 자료를 활용하여 강우량(rain), 풍속(wind), 평균 기온(av_temp) 자료를 수집하였다.

3.3 데이터 전처리

본 연구의 목표 변수가 될 유해남조류 세포수가 결측치인 행은 제거하고, 유해남조류 세포수(cells)에 log를 취하여 정규분포와 유사하게 만들어 더 정교한 예측을 가능하게 한다. 유해남조류 세포수 외의 각 결측치는 월평균 값으로 대체하였다. 댐 제원 정보는 각 채수위치 별 인근 댐정보로 차원을 축소하였다. 변수인 채수 위치, 계절, 댐은 라벨 인코딩을 하여 범주형 변수를 수치화 하였고, 외의 변수들은 Standard scale 과정을 거쳤다.

따라서, 최종적으로 선택된 변수는 유해남조류 세포수(log_cells), 채수 위치(region), 연도(year), 월(month), 일(day), 날짜(date), 계절(season), 수온(temp), 평균 기온(av_temp), 강우량(rain), 풍속(wind), 인근 댐(weir), 댐 유입량(inflowqy), 댐 수위(lowlevel), 댐 기준 강우량(prcptqy), 댐 저수량(rsvwtqy), 댐 저수율(rsvwtrt), 댐 총 방류량(totdcwtrqy), fac_3, fac_5, fac_7, fac_10, ww_3, ww_5, ww_7, ww_10이다. 전체적인 과정은 Figure 3과 같다.

3.4 사용 알고리즘

알고리즘을 적용할 기계학습 소프트웨어는 Weka를 사용한다. Weka는 Waikato 대학교에서 제작한 Jave 기반의 검증된 기계학습 소프트웨어이다.

본 연구에서 사용할 예측 알고리즘은 *k*-NN, SVM, Random Forest이다. 전통적인 시계열 분석 모델은

표 2. 유해남조류 세포수와 인자 간의 피어슨 상관계수
Table 2. Pearson's correlation coefficients between harmful cyanobacteria and factors

Features	Correlation coefficient
Water Temperature(temp)	0.630
Month(month)	0.374
Inflow amount of water(inflowqy)	0.188
Total discharge(totdcwtrqy)	0.187
Rainfall of dam region(prcptqy)	0.146
Rainfall(rain)	0.130
Wind velocity(wind)	0.124
Number of waste water disposal facilities within 7km(ww_7)	0.102
Number of factories within 3km(fac_3)	0.060
Number of waste water disposal facilities within 3km(ww_3)	0.059
Number of waste water disposal facilities within 5km(ww_5)	0.057
Number of factories within 5km(fac_5)	0.044
Number of factories within 7km(fac_7)	0.043
Number of factories within 10km (fac_10)	0.043
Water storage rate(rsvwtrt)	0.034
Average of air temperature(av_temp)	0.033
Number of factories within 10km(ww_10)	0.027
Storage capacity(rsvwtqy)	-0.027
Year(year)	-0.079
Season(season)	-0.129
Low level(lowlevel)	-0.207

선형 모델을 가정하여 정확도가 낮다는 문제점이 있고 [9], 비선형 기계학습 방법은 분류와 회귀에서 좋은 성과를 보여주고 있다[10]. 또한,, 본 연구의 데이터셋이 한 지역의 장시간 변화가 아닌 다양한 지역의 일정 기간 유해남조류 변화를 담은 것이기 때문에 타연구에서 진행됐던 시계열 기법보다 분류 기법이 더 적절하리라 판단하여 선정하였다.

먼저, *k*-NN 알고리즘을 설명하자면, 분류와 회귀에 사용되는 비모수 방식인 알고리즘이다. *k*-NN은 *k* 개의 다른 데이터의 레이블을 참조하여 유클리디안 거리(Euclidean Distance)를 사용하여 분류를 진행한다. *k*-NN은 수치 기반의 데이터 분류 작업에서 성능이 우수하다는 장점이 있다. 위의 전처리 과정에서 라벨인코딩 작업을 거쳐 범주형 데이터를 수치화 하였기 때문에 본 연구에서 사용할 데이

터는 날짜를 제외하면 모두 수치 데이터이기에 k-NN을 선정하였다.

SVM은 데이터들 간의 거리를 최대로 하는 초평면(hyperplane)을 계산하여 데이터를 분류하는 방법이다. SVM의 알고리즘은 결과 해석이 용이하고, 라벨링을 할 때 로지스틱 회귀나 판별 분석과 달리 SVM은 확률이 아닌 라벨을 직접 예측하기 때문에 선정하였다.

Random Forest는 다양한 데이터 및 상황에 적합한 지도 머신러닝의 종류로, 기능을 무작위로 선택한 여러 개의 결정트리를 형성한다. 여러 결정 트리들의 결과를 앙상블(ensemble)하여 가장 많이 나온 값을 최종값으로 선정하는 기법이다[11]. Random Forest는 과적합 문제를 회피하여, 모델 정확도를 개선한다는 장점이 있는데 본 연구에서 수온과 같은 상관계수가 높은 특정 변수에 의한 과적합 문제를 방지하고자 모델을 선정했다.

4. 실험 결과 및 고찰

4.1 유해남조류 상관분석

유해남조류에 선형적 영향을 미치는 변수를 파악하기 위하여 피어슨 상관 분석을 실시하였다. 본 연구에서 새롭게 추가된 N km 내의 공장개 수, 폐수처리장 개수는 굵은 글씨로 표시하였다.목표 변수인 유해남조류 세포수(log_cells)와 타 변수와의 상관계수를 분석한 결과(Table 2) 수온(temp)이 0.630으로 가장 높게 나타났고, 월(month), 댐 유입량(inflowqy), 댐 총 방류량(totdcwtrqy)가 각각 0.374, 0.188, 0.187로 높게 나타났다. 공장과 폐수처리장 관련 변수는 대체로 낮은 상관계수를 가진 것으로 나타났다.

4.2 유해남조류 예측 모델

본 연구에서 Ground truth는 기존 데이터에 로그값을 취한 유해남조류세포수(log_cells)이다. 이 목표변수가 의미하는 바는 ml당 마이크로시스티스(Microsystis), 아나베나(Anabaena), 아파니조메논(Aphanizomenon), 오실라토리아(Oscillatoria) 속(屬) 세포수의 합이다. 연구 데이터의 기간은 2016년부터 2022년 5월까지로, 총 4281 행의 데이터를 대상으로 진행하였다. Train data, Test data 비율을 8:2로 무작위로 나누어 진행하였고, 이때 Train data, Test data의 행의 수는 각각 3425행, 856행이다. 유해남조류 예측 모델로서 활용할 기법은 k-NN, SVM, Random Forest를 3가지를 사용하였다. 모델의 정확도를 평가할 지표로는 결정계수(Coefficient of Determination)와 RMSE를 사용하였다. RMSE와 결정계수를 구하는 수식은 각각 다음 식 (1) 식 (2)와 같다.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \quad (1)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} \quad (2)$$

식 (1)과 식 (2)에서 n은 최종 데이터의 총 개수, y_i는 ml당 유해남조류 세포수를 의미한다.

4.2.1 변수 선택 방법 및 예측 모델 모수 설정

기계 학습을 통한 유해남조류 예측에 앞서 각 모델을 변수를 선택하는 방법으로 Wrapper 기반의 Genetic search을 사용하였다. Genetic search는 자연에서의 생명체 진화 개념에 기초한 효율적인 방법으로 성공적으로 적용되어 왔다[12]. Genetic search는 매개 변수가 많은 경우에 적합하고, 연속 및 이산 기능과 다중 목표 문제를 최적화하기 적합하기 때문에 선정하였다.

본 논문에서 사용할 알고리즘 기법으로 k-NN, SVM, Random Forest의 모수는 각각 Table 3, Table 4, Table 5와 같다.

본 연구에서는 k-NN의 Linear NNS(Nearest Neighbor Search)를 사용하여 회귀를 목적으로 하

표 3. k-NN 실험에 사용된 파라미터
Table 3. k-NN parameters

Parameters	Values
k	12
Batchsize	100
DistanceWeighting	No distance weighting
NearestNeighbourSearch Algorithm	LinearNNSearch (EuclideanDistance)
NumDecimalPlaces	2

표 4. SVM 실험에 사용된 파라미터
Table 4. SVM parameters

Parameters	Values
SVMType	nu-SVR (regression)
cost	1
eps	0.001
Batchsize	100
KernelType	radial basis function: exp(-gamma*|u-v|^2)

표 5. Random Forest 실험에 사용된 파라미터
Table 5. Random Forest parameters

Parameters	Values
BatchSize	100
MaxDepth	Unlimited
NumDecimalPlaces	2
NumExecutionSlots	1
NumIterations	100
Seed	1

는 최근접 이웃법을 적용한다. 이때 거리 공식을 적용할 때, 거리가 멀리 떨어진 이웃에 가중치를 주지 않고 계산한다(No distance weighting). k-NN을 적용할 때의 거리 공식으로 식 (3)을 사용한다.

$$d(\mathbf{x}_i, \mathbf{x}_l) = \sqrt{(x_{i1} - x_{l1})^2 + (x_{i2} - x_{l2})^2 + \cdots + (x_{ip} - x_{lp})^2} \quad (3)$$

이때, x_i와 x_l은 서로 다른 변수를 의미한다.

본 연구에서 SVM 유형으로 nu-SVR을 사용한다. nu-SVR은 변형된 유형 중 하나로 결정함수는 식 (4)와 같다. 기존 목적 함수에 Slack Variable을 대입하여 식 (5)와 같이 표현하였다. 이때, ξ은 오차범위를 벗어난 거리이고, w와 b는 각 파라미터에 대해 편미분을 했을 때, 0이 되게 하는 값이다.

$$\hat{y}_i = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}) + b \quad (4)$$

$$C\sum_{i=1}^{n}(\xi_i + \hat{\xi}_i) + \frac{1}{2}\|\mathbf{w}\|^2 \quad (5)$$

Random Forest의 BatchSize는 100, MaxDepth는 제한 없음으로, NumIterations는 100으로 설정하여 각 환경에 따른 유해남조류 세포수를 분류할 수 있도록 한다.

4.2.2 유해남조류 예측 모델 결과

각 모델 별 선택된 변수는 다음(Table 6)과 같다. 본 연구에서 새롭게 추가된 N km 내의 공장 개수, 폐수처리장 개수는 굵은 글씨로 표시하였다. 3가지 모델 모두 수온(temp), 월(month)이 변수로서 선택되었으며, 이는 통상적인 유해남조류 발생 조건인 수온과 계절성을 반영한 것으로 보인다. 반면, 계절(season) 변수는 선택되지 않은 것으로 보아, 3개월 단위로 나눈 계절 요인보다 1년을 12개월로 나눈 월 변수가 더 모델에 적합하였다고 판단된다. 댐 제원정보와 폐수처리장 개수, 공장 개수 또한 세부 변수는 상이하지만, 모든 모델에 본 연구에 새롭게 추가된 N km내의 공장 개수, 폐수처리장 개수 포함된 것을 보아 녹조 현상을 예측하는데 유의하게 사용되었다고 판단된다.

모델 성능은 Random Forest가 0.880의 결정계수와 1.606의 RMSE로 가장 성능이 좋았다. 그 뒤로 k-NN, SVM 순으로 좋았으며 각각의 결정계수는 0.839, 0.744이고, RMSE는 1.852, 2.28로 나타났다. 변수 간의 선형성을 중심으로 보는 SVM은 해당 연구에서 선형 관계를 나타내는 변수가 적었기에 성능이 낮았던 것으로 판단된다. k-NN에 비해 Random Forest 모델이 성능이 좋았던 것은 데이터 간 유클리디안 거리(Euclidean distance)로 설명되는 것이 아닌 여러 조건을 고려할 수 있는 Random Forest의 특성상 녹조 발생 조건들을 고려할 수 있었다고 판단된다.

본 연구와 가장 유사한 지점을 대상으로 진행한 Jung, W. S.[4]은 Random Forest로 0.83의 결과를 얻은 것에 비해 본 연구가 제시하는 Feature 조합을 통한 Random Forest 결과 0.88로 유의한 차이가 있다는 것을 통해 본 연구가 제시하는 조합이

표 6. 예측 모델 결과
Table 6. Results of Prediction model

Model	Input Variables	Data period	Coefficient of Determination	RMSE
k-NN	weir, date, year, month, temp, rain, inflowqy, lowlevel, prcptqy, rsvwtqy, rsvwtrt, totdcwtrqy, **fac_5, fac_10**	2016.01~ 2022.05	0.839	1.852
SVM	month, temp, av_temp, lowlevel, **fac_3, ww_3**		0.744	2.28
Random Forest	year, month, temp, av_temp, rain, lowlevel, rsvwtqy, **fac_7, ww_3, ww_10**		0.880	1.606

유의하다고 할 수 있다.

5. 결론 및 향후 연구

본 연구는 낙동강의 유해남조류 세포수 예측 모델을 구현하였다. 선행 연구에서 증명된 수온과 pH, DO가 유해남조류를 예측하는데 주요 변수로써 작용한다는 사실 외에도 N km 내의 공장 수(fac_N), N km내의 폐수처리장 수(ww_N) 변수가 유의한 변수로써 활용된다는 사실을 도출하였다. 유해남조류 세포수 예측 결과 Random Forest로 0.880의 결정계수와 1.606의 RMSE를 얻을 수 있었고, *k*-NN과 SVM은 각각 0.803, 0.706의 결정계수와 2.041, 2.42의 RMSE를 얻을 수 있었다. 본 연구는 영양염류 관련 변수로 pH와 DO가 아닌 N km 내의 공장 수(fac_N) 와 N km 내의 폐수처리장 수(ww_N)를 사용하여 높은 결정계수의 모델을 도출했다는 데에 의의가 있다. 추후, Greedy Forward, Greedy Backward를 통한 변수 검색을 시도하고, 타 연구[7]에서 의의가 있었던 가뭄지수 변수를 고려하여 모델의 성능 개선을 목표로 삼겠다. 또한, 한강, 금강, 영산강과 같은 타 강으로 확장한 연구를 기대한다.

Conflict of Interest

저자는 본 논문에 관련된 어떠한 잠재적인 이해 상충도 없음을 선언한다.

References

[1] Water Sanitation and Health Team, "Management of cyanobacteria in drinking-water supplies," Available: https://www.who.int/publications/i/item/WHO-FWC-WSH-15.03, 2015, [Accessed: January 29, 2015].

[2] Kevin G Sellner, Gregory J Doucette and Gary J Kirkpatrick, "Harmful algal blooms: causes, impacts and detection," *Journal of Industrial Microbiology and Biotechnology*, vol. 30, no. 7, pp. 383-406, 2003.

[3] Kim, Beom-Cheol, "Causes and Countermeasures of Green Tide Phenomenon," *Water for future*, vol 50, no. 6, pp. 8-14, 2017.

[4] Kim, Min Seok, Park, Hyung wook, Jo, Hyun Jung, and Kim Eun Joo, "A Research of Algal Early Forecasting Method using Deep Learning," *Proceedings of the Korean Operations and Management Science Society Fall Conference*, Jeju, Republic of Korea, pp. 6071-6072, June, 2021.

[5] Jung, Woo Suk, Kim, Sung Eun, Kim, Young Do, "Analysis of influential factors of cyanobacteria in the mainstream of Nakdong river using random forest," *Journal of Wetlands Research*, vol. 23, no. 1, pp. 27-34, https://doi.org/10.17663/JWR.2021.23.1.27, 2021.

[6] Kim, Sang-Hoon, Park, Jun Hyung, Kim, Byunghun, "Prediction of cyanobacteria harmful algal blooms in reservoir using machine learning and deep learning," *Journal of Korea Water Resources Association*, vol. 54 no. 1, pp. 1167-1181, https://doi.org/10.3741/JKWRA.2021.54.S-1.1167, 2021.

[7] Jung Woo Suk, Jo, Bu Geon, Kim, Young Do, Kim, Sung Eun, "A study on the characteristics of cyanobacteria in the mainstream of Nakdong river using decision trees," *Journal of Wetlands Research*, vol. 21 no. 4, pp. 312-320, https://doi.org/10.17663/JWR.2019.21.4.312, 2019.

[8] Jung, Y. J., "A Study on the Prediction of Water Quality Concentration Using XGBoost," *Journal of Information Technology and Applied Engineering*, vol. 12, no. 2, pp. 27-33, 2022.

[9] Byun, Jun-Hyung, Kim, Ji-Ho, Choi, Young-Jin, Lee, Hong-Chul, "Predicting the Real Estate Price Index Using Machine Learning Methods and Time Series Analysis Model," *Housing Studies Review*, vol. 26, no. 1, pp. 107-133, https://doi.org/10.9708/jksci.2020.25.06.035, 2018.

[10] Byun, Jun-Hyung, Kim, Ji-Ho, Choi, Young-Jin, Lee, Hong-Chul, "Movie Box-office Prediction using Deep Learning and Feature Selection : Focusing on Multivariate Time Series," *Journal of the Korea Society of Computer and Information*, vol. 25, no. 6, pp. 35-47, https://doi.org/10.9708/jksci.2020.25.06.035, 2020.

[11] Kim, Hyun Il, Lee, Yeon Su, and Kim, Byunghyun, "Real-time flood prediction applying random forest regression model in urban areas," *Journal of Korea Water Resources Association*, vol. 54 no. 12, pp. 1119-1130, 2021.

[12] E.SH. Hou, N. Ansari, Hong Ren. "A genetic algorithm for multprocessor scheduling," *IEEE Transaction on Parallel and Distributed Systems*, vol. 5, no. 2, pp. 113-120. 1994.

저 자 소 개

송찬영(Chan-Younng Song)
2018~현재 : 한국외국어대학교
미디어커뮤니케이션 학부

관심분야 : 빅데이터, 머신러닝
ORCID Number : 0000-0002-1089-850X
E-mail : scy0208@daum.net

김주연(Ju-Yeon Kim)
2020년~현재: 세종대학교 응용통계학과

관심분야 : 데이터 사이언스, 베이즈 통계, 머신러닝
ORCID Number : 0000-0002-1266-0853
E-mail : jgys014@naver.com

김예진(Ye-Jin Kim)
2020년~현재: 서울여자대학교 수학과

관심분야 : 빅데이터 분석, 데이터 사이언스, 데이터 엔지니어링
ORCID Number : 0000-0001-5928-8852
E-mail : 06kyj20@naver.com

주해종(Hae-Jong Joo)
2008년: (미)Cumberland University 교육학과 박사 졸업
2010년: 명지대학교 컴퓨터공학과 공학박사
2018~현재: 동국대학교 컴퓨터공학과 교수

관심분야 : ICT융합기술, 데이터공학, 뉴미디어 품질평가, 드론응용SW
ORCID Number : 0000-0003-1086-5860
E-mail : hjjoo@dongguk.edu

서재현(Jae-Hyun Seo)
2008년: 광운대학교 컴퓨터과학과 공학석사
2016년: 광운대학교 컴퓨터과학과 공학박사
2017~2020년: 원광대학교 컴퓨터·소프트웨어공학과 조교수
2020.3~현재: 광주대학교 컴퓨터공학과 조교수

관심분야 : 인공지능, 최적화 알고리즘, 진화연산
ORCID Number : 0000-0002-1587-788X
E-mail : jhseo@gwangju.ac.kr