Journal of Korean Society on Water Environment, Vol. 37, No. 5, pp. 335-343 (September, 2021)
pISSN 2289-0971 eISSN 2289-098X https://doi.org/10.15681/KSWE.2021.37.5.335

입력자료 군집화에 따른 앙상블 머신러닝 모형의 수질예측 특성 연구

박정수†

국립한밭대학교 건설환경공학과

The Effect of Input Variables Clustering on the Characteristics of Ensemble Machine Learning Model for Water Quality Prediction

Jungsu Park†

Department of Civil and Environmental Eng, Hanbat National University

(Received 17 August 2021, Revised 16 September 2021, Accepted 23 September 2021)

Abstract

Water quality prediction is essential for the proper management of water supply systems. Increased suspended sediment concentration (SSC) has various effects on water supply systems such as increased treatment cost and consequently, there have been various efforts to develop a model for predicting SSC. However, SSC is affected by both the natural and anthropogenic environment, making it challenging to predict SSC. Recently, advanced machine learning models have increasingly been used for water quality prediction. This study developed an ensemble machine learning model to predict SSC using the XGBoost (XGB) algorithm. The observed discharge (Q) and SSC in two fields monitoring stations were used to develop the model. The input variables were clustered in two groups with low and high ranges of Q using the k-means clustering algorithm. Then each group of data was separately used to optimize XGB (Model 1). The model performance was compared with that of the XGB model using the entire data (Model 2). The models were evaluated by mean squared error-observation standard deviation ratio (RSR) and root mean squared error. The RSR were 0.51 and 0.57 in the two monitoring stations for Model 2, respectively, while the model performance improved to RSR 0.46 and 0.55, respectively, for Model 1.

Key words : Clustering, Ensemble machine learning, Gradient boosting decision tree, Water quality prediction, Water supply system, XGBoost

† Corresponding auther, 조교수(Assistant Professor), parkjs@hanbat.ac.kr, https://orcid.org/0000-0002-9780-6988

1. Introduction

취수원의 안정적 수질관리를 위해서는 수질현황에 대한 지속적인 모니터링과 함께 수질의 변화에 대한 예측이 필요하다. 하천 및 저수지 등 취수원 수질은 유기물질 및 영양염류 등 다양한 오염원에 의해 영향을 받게 되며 수중의 부유사(suspended sediment)도 취수원의 수질과 수생태에 영향을 미치는 중요한 인자중 하나이다(Packman and MacKay, 2003; Singer et al., 2013). 또한 강우시 유량증가에 따른 부유사 농도(suspended sediment concentration, SSC)의 증가는 취수원 고탁수의 원인이 되며 정수처리 비용의 증가 및 수질사고 발생 등 정수처리공정에도 다양한 영향을 미치게 된다(Lin et al., 2004; Park and Lee, 2020).

최근 다양한 분야에서 머신러닝 알고리즘에 기반한 모형의 적용이 활발하게 늘어나고 있으며, 물환경분야에서도 이러한 고도화된 데이터 분석 알고리즘을 수질 예측 및 관리에 적용하기 위한 연구가 계속되고 있다(Haghiabi et al., 2018; Li et al., 2021; Muhammad et al., 2015). 대표적인 머신러닝 알고리즘인 인공신경망(artificial neural network, ANN) 뿐 아니라 support vector machine (SVM), ensemeble 머신러닝 알고리즘인 random forest (RF), 그리고 기존 ANN 모형의 한계를 개선하여 머신러닝 분야의 획기적인 발전을 이루어낸 딥러닝(deep learning) 모형중 시계열 자료의 분석에 좋은 성능을 보이는 순환신경망(recurrent neural network) 기반의 long short term memories (LSTM) 등 다양한 머신러닝 모형이 탁도 예측에 적용되는 등 관련분야 연구가 활발히 진행되고 있다(Park and Lee, 2020; Stevenson and Bravo, 2019; Wang et al., 2021).

Ensemble 머신러닝 모형은 weak learner로 불리는 여러 개의 모형을 함께 사용하여 예측성능을 높이는 방식을 사용하며 RF와 gradient boosting decision tree (GBDT) 등이 대표적인 ensemble 머신러닝 모형이다(Sutton, 2005; Zhang, Qian et al., 2018). 두가지 모형 모두 회귀분석(regression) 및 분류(classification) 두가지 방식 모두에 적용이 가능하고 충분한 입력자료를 확보할 경우 높은 예측성능을 보여 최근까지도 가장 널리 활용되는 머신러닝 모형중 하나이며, 수질분야에도 활용이 점차 늘고 있다(Hollister et al., 2016; Uddameri et al., 2020).

머신러닝 모형은 물리적 혹은 화학적 관계에 기반한 별도의 계수 등을 구하지 않아도 모형에 사용되는 독립변수와 복잡한 비선형관계(non-linear)를 가지는 종속변수에 대해서도 좋은 예측성능을 보이는 장점이 있다. 머신러닝 모형의 성능은 입력자료로 활용되는 항목의 구성과 측정빈도 및 적정한 전처리 등을 포함하는 feature engineering에 의해 많은 영향을 받게 되며, 모형의 성능을 최적화하기 위해서는 적정한 입력 변수의 구축이 중요하다(Park, 2021).

취수원으로 활용되는 하천 및 저수지 등에서의 부유사 농도(suspended sediment concentration, SSC)는 강우량, 유사(sediment) 발생원의 특성, 유사 발생원과 측정지점의 거리, 강우 발생 이전의 무강우 일수, 최대 강우강도 등 자연적 요인과 함께(Hicks et al., 2000; Park and Hunt, 2017; Warrick, 2015; Warrick et al., 2013) 건설공사, 농업 활동 등 인간활동 그리고 기후변화 등 다양한 환경인자에 영향을 받게 된다(Gray et al., 2016; Gray et al., 2015). 하천 유량(Q)는 SSC에 영향을 주는 가장 중요한 인자 중 하나이다. 하지만 SSC는 Q외에도 여러 가지 환경요인에 영향을 받으므로, 동일 장소에서 동일한 Q가 발생해도 연도, 계절 및 선행 강우조건 등에 따라 SSC가 큰 차이를 보이기도 하고, Q의 크기에 따라 구간별로 SSC와 Q의 상관관계가 다르게 나타나기도 한다(Walling, 1977; Warrick, 2015).

본 연구에서는 최근 까지도 널리 사용되는 대표적인 ensemble 머신러닝 모형중 하나인 Gradient boosting decision tree (GDBT)를 활용하여 Q를 독립변수로 이용하여 SSC를 예측하는 모형을 구축하였다. 모형의 구축에 입력자료의 특성을 반영하기 위해 자료의 특성에 따라 군집화(clustering)를 수행하는 머신러닝 비지도 학습(unsupervised learning) 알고리즘 중 하나인 k-평균 군집화(k-means clustering, KMC) 모형을 이용하여 Q에 따라 입력자료의 군집화를 수행하고, GBDT 모형을 이용하여 각각의 군집에 최적화된 SSC 예측 모형을 구축하였다. 또한 비교를 위하여 별도의 군집화를 수행하지 않고 전체자료를 입력자료로 이용하는 GBDT 모형을 구축하여 입력자료의 군집화 수행여부에 따른 모형 성능을 비교하여, 입력자료의 특성을 고려한 모형의 구축이 모형 성능에 미치는 영향을 분석하였다.

2. Materials and Methods

2.1 Data sources

미국 지질조사국(United States Georogical Survey, USGS)은 국토관리와 연구를 위해 미국 전역에 현장측정소를 설치하여 장기간에 걸쳐 유량과 SSC를 측정하고 그 결과를 공개하고 있으며, 본 연구에서는 USGS에서 운영하는 현장측정소 중 미국 Calfironia Reedwood Creek에 위치한 2개 지점(Blue Lake 및 Orick)의 Q 및 SSC 일일 측정자료를 활용하

Table 1. Research sites

Sites	Watershed area (㎢)	Location		USGS site number	Observation period
		Latitude	Longitude		
Blue Lake	175	40°54′22″	123°48′51″	11481500	Oct 1, 1972- April 30, 1992
Orick	717	41°17′58″	124°03′00″	11482500	March 19, 1970- April 30, 1992

였다(Table 1) (USGS, 2014). 미국 서부연안에 위치한 Redwood Creek은 지중해성 기후 지역에 속하며 10월경부터 우기가 시작되어 이듬해 봄까지 계속되고 이후 9월경까지 건기가 이어지는 강우 특성을 가지고 있다. Orick 지점은 강 하구로부터 약 6km 상류에 위치하며 하천은 Blue Lake에서 Orick을 거쳐 태평양으로 유입하게 된다(USGS, 2009).

2.2 Model development

본 연구에서는 ensemble 머신러닝 모형인 GBDT 모형을 이용하여 하천의 SSC를 예측하는 모형을 구축하였다. GBDT는 RF와 함께 대표적인 ensemble 머신러닝 모형 중 하나이다. RF는 의사결정나무(decision tree, DT)기반의 다수의 weak learner를 생성하고 각 weak learner에서 독립적으로 생성된 결과의 평균을 이용하여 예측값을 산정하는 반면, GBDT 모형은 전단계 weak learner의 예측값을 다음 단계의 weak learner의 구축에 활용하며, 실측값과 예측값 간의 잔차가 많은 입력자료에 더 높은 가중치를 주어 모형의 학습을 (training) 수행하여 모형의 성능을 향상시키도록 구성된 모형이다(Chen and Guestrin, 2016; Friedman, 2001; Zhang, Bouadi et al., 2018).

GBDT 모형은 예측의 대상이 되는 항목의 실측값($y_{obs,i}$)과 모형의 예측값($y_{pred,i}$)의 차이를 계산하는 손실함수(L: loss function)와, 개별 DT 모형(f_k)의 함수인 regulation 함수(Ω)로 구성된 objective 함수(J)를 최소화하는 방향으로 모형을 최적화한다(Eq. 1) (Chen and Guestrin, 2016; Shin et al., 2020; Zhang, Qian et al., 2018). 모형의 구축은 가장 널리 사용되는 GBDT 알고리즘 중 하나인 XGBoost regressor (XGB)를 이용하였으며, Q를 독립변수로 하여 종속변수 SSC를 예측하도록 구성하였다. 또한 일단위 자료의 차분을 적용하여 시간 t에 대해서 1일 전의 Q 및 SSC인 Q_{t-1}과 SSC_{t-1}을 입력자료로 추가하여 모형의 구축에 활용하였다. 모형의 최적화는 grid search 방법을 이용하였으며, 입력자료를 10개의 set으로 구분하여 cross validation을 수행하였다. 모형의 구축과 최적화 등은 python open source library인 Scikit-learn을 이용하여 실행하였다(Pedregosa et al., 2011).

$$J=\sum_{i=1}^{n} L(y_{obs,i}, y_{pred,i}) + \sum_{i=1}^{K} \Omega(f_k) \tag{1}$$

2.3 Clustering of input variables

XGB 모형에 사용된 입력자료의 군집화를 위해 비지도 학습 모형인 KMC을 이용하였다. KMC는 입력자료를 사전에 개수가 정해진 임의의 군집에 분류하고 각 군집의 평균값(μ_j)과 각 입력자료의값(x_i) 과의 차이를 유클리디언 거리(euclidean distance)를 이용하여 구하고 이를 최소화 할 수 있도록 최종적으로 분류되는 군집을 결정하는 모형이다(Ahmad and Dey, 2007; Ayub et al., 2016; Song, 2017) (Eq. 2). KMC는 python Scikit-learn library를 이용하여 수행되었다(Pedregosa et al., 2011).

$$\sum_{j=1}^{k} \sum_{x_i \in R} \| x_i - \mu_j \| \tag{2}$$

2.4 Model evaluation

구축된 XGB 모형을 이용한 SSC 예측성능의 평가는 평균제곱근 오차(root mean square error, RMSE)와 평균 제곱근 오차-관측값 표준편차비(mean squared error-observation standard deviation Ratio)를 이용하였다(Eq. 3 and 4). RMSE는 예측값과 실측값의 차이의 절대치를 비교하는 지수로 RMSE가 0에 가까울수록 모형의 예측성능이 좋음을 의미한다. RSR은 모형간 성능의 절대적인 비교가 가능한 지수로 0~1의 범위를 가지며 일반적으로 RSR이 0.7 이하인 경우 예측이 잘 수행된 것으로 판단하고, 0에 가까울수록 모형의 성능이 우수한 것을 의미한다(Bennett et al., 2013; Moriasi et al., 2007).

$$RSME=\sqrt{\frac{\sum_{t=1}^{n}(Y_{t,obs}-Y_{t,pred})^2}{n}} \tag{3}$$

$$RSR=\frac{\sqrt{\sum_{t=1}^{n}(Y_{t,obs}-Y_{t,pred})^2}}{\sqrt{\sum_{t=1}^{n}(Y_{t,obs}-\overline{Y_{t,obs}})^2}} \tag{4}$$

where

$Y_{t,obs}$: Observed value at time t,

$Y_{t,pred}$: Predicted value at time t,

$\overline{Y_{t,obs}}$: mean of observed values.

3. Results and Discussion

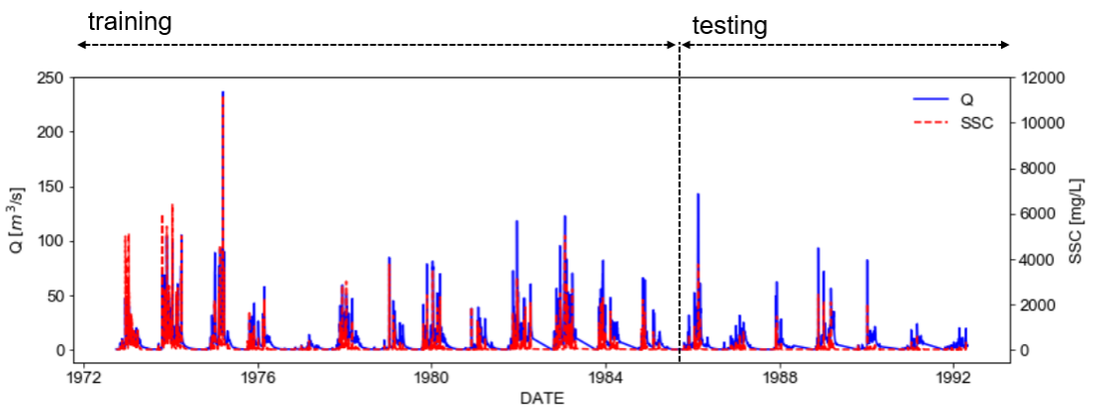
3.1 Characteristics of input variables and pretreatment of missing variables

모형의 구축에 사용된 입력자료의 기초 통계량을 Table 2에 제시하였다. 모형 구축에 사용된 측정값은 Blue Lake와 Orick에서 각각 15% 및 13%의 SSC의 결측치를 포함하고 있다. 머신러닝 모형의 구축시 입력자료에 결측치가 포함되는 경우 이를 제거하거나, 보간법이나 주변값들의 평균값을 이용해서 결측치를 추정하는 k nearest neighbors 등을 통해 결측치에 대한 전처리를 수행하게 되며, 입력자료의 특성을 고려한 적정한 전처리 방법의 선정이 필요하다.

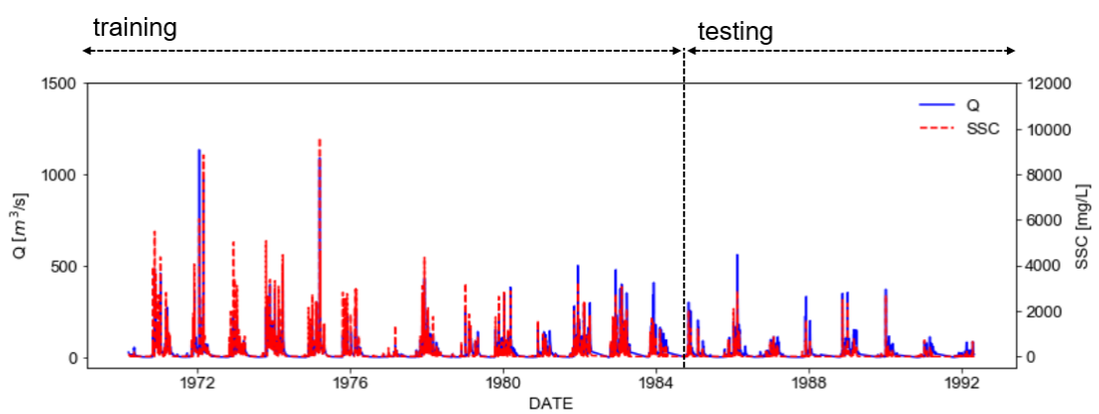
본 연구에 사용된 입력자료가 측정된 북부 California 지역은 10월부터 우기가 시작되고 이듬해 2월경까지 강우가 지속되게 되며, 이후 봄과 여름 동안은 강우가 거의 발생하지 않는 건기가 지속된다. SSC의 결측치는 대부분 이러한 건기인 3~9월중에 발생하였으며, 이시기는 강우가 발생하지 않아 낮은 Q와 SSC가 측정되는 시기로 본 연구에서는 별도의 결

Table 2. Characteristics of input variables

Site	Variables	Average	Min	Max	Standard deviation
Blue Lake	Q (m^3/s)	6.97	0.05	236.73	11.87
	SSC (mg/L)	117.44	0	11,200	427.69
Orick	Q (m^3/s)	30.60	0.06	1,135.51	56.33
	SSC (mg/L)	158.70	0	9,610	474.69



(a) Blue Lake



(b) Orick

Fig. 1. Training and testing data.

측치 보정을 수행하지 않고, 입력자료에서 결측값을 제거하고 사용하는 방법을 적용하였다. 또한 이러한 강우 특성을 고려하여 건기가 끝나고 새로운 우기가 시작되는 10월을 기준으로 모형의 training과 성능의 평가를 위한 testing에 사용되는 입력자료를 구분하여, Blue Lake는 1985년 10월 1일 이후의 자료를 Orick에서는 1984년 10월 1일 이후의 자료를 testing에 활용하였다(Fig. 1). 모형의 training과 testing에는 결측치를 제외하고 Blue Lake에서는 각각 4,271일 및 1,792일, Orick에서는 각각 4,853일 및 2,157일 간 측정된 값이 사용되어, training과 testing에 사용된 입력자료의 비율은 Blue Lake와 Orick에서 각각 0.70:0.30 및 0.69:0.31로 구성되었다.

3.2 Clustering of input SSC

KMC를 이용하여 모형의 구축에 사용된 training 자료를 Q가 낮은 군집과(Class 1), 높은 군집(Class 2)의 2개의 군집으로 구분하여 XGB 모형에 적용하였으며, 군집화를 하지 않은 전체자료를 이용하여 구축된 모형과 성능을 비교하여 입력자료의 특성을 고려한 모형의 구축이 모형 성능에 미치는 영향을 분석하였다(Table 3 and Fig. 2).

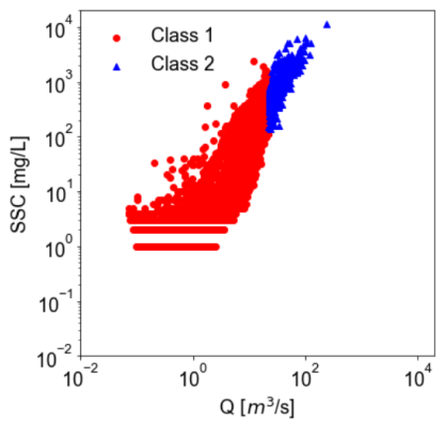
3.3 Model simulation result

Model 1. Separated model

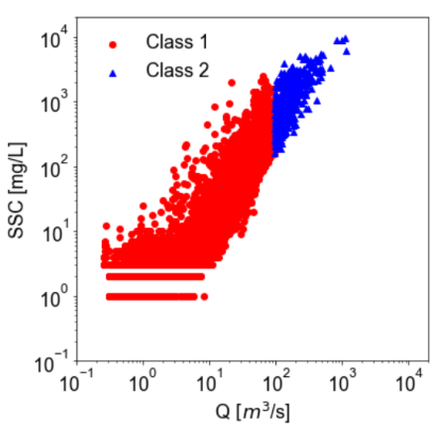
군집화를 통해 Q를 기준으로 구분된 낮은 Q 값을 가지는 Class 1과 높은 Q 값을 가지는 Class 2 각각에 대하여 별도의 training을 수행하여 모형을 구축하였다. 모형의 testing은 각 testing 자료가 해당되는 Class에서 구축된 모형을 적용하여 수행하였다.

Table 3. Clustering of input variables for the model training

Site	Blue lake		Orick	
Class	Class 1 (low range)	Class 2 (high range)	Class 1 (low range)	Class 2 (high range)
Max Q (m^3/s)	22.6	236.7	96.0	1135.5
Number of observation	3,923	348	4,393	460

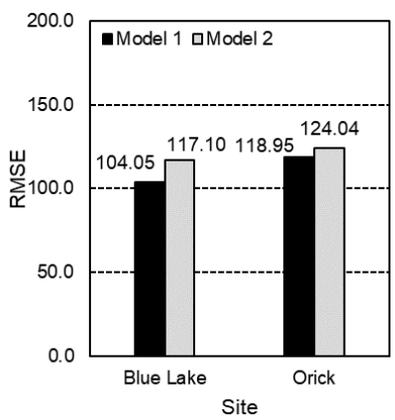


(a) Blue Lake

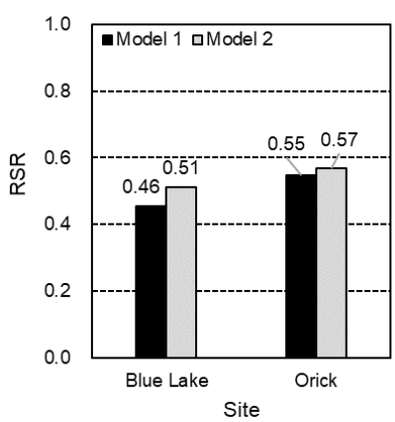


(b) Orick

Fig. 2. Distribution of clustered input variables for the model training.



(a) RMSE



(b) RSR

Fig. 3. A comparison of model evaluation results.

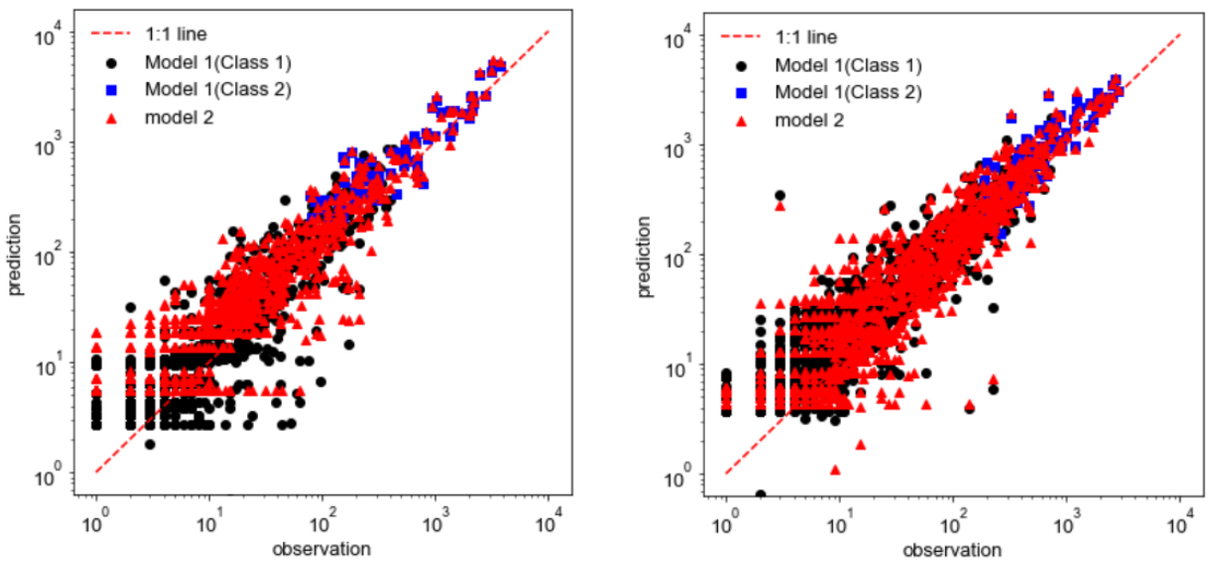
Model 2. Combined model

군집화를 통해 입력자료의 특성을 반영하여 모형을 구축한 Model 1과의 비교를 위해 Q에 따른 구분 없이 전체 입력자료를 모두 사용하여 training 및 testing을 수행하였다.

구축된 모형의 testing 결과 Blue Lake와 Orick 두측정지점 모두에서, 군집화를 통해 낮은 Q와 높은 Q 구간에 대하여 별도의 최적화를 수행한 Model 1이 전체자료를 이용하여 구축된 Model 2보다 더 좋은 성능을 보여, 입력자료 특성을 고려한 모형 구축을 통해 XGB 모형의 성능을 개선할 수 있는 것을 확인하였다(Fig. 3).

Q의 범위와 상관없이 전체 입력자료를 모두 이용하여 구축한 Model 2의 경우 Blue Lake와 Orick에서 RSR이 각각 0.51과 0.57로 분석되었으나, Model 1의 RSR은 Blue Lake와 Orick에서 각각 0.46 및 0.55로 분석되어 개선된 SSC 예측성능을 보여주었다. RMSE는 Blue Lake와 Orick에서 Model 2의 경우 각각 117.10과 124.04로 Model 1의 경우 각각 104.05와 118.95로 분석되어, RSR과 마찬가지로 두지점 모두에서 Model 1을 사용할 때 성능이 개선되었다.

모형의 결과를 시각적으로 확인하기 위해 구축된 Model 1과 Model 2의 testing 자료에 대한 실측값과 예측값을 비교하여 Fig. 4에 제시하였다. Fig. 4의 검은색 원은 Model 1의 낮은 Q 구간에 대하여, 파란색 사각형은 Model 1의 높은 Q 구간에 대해서 각각 최적화된 모형의 예측값과 실측값의 관



(a) Blue Lake (b) Orick

Fig. 4. A comparison of model predictions.

계를 나타내며, 빨간색 삼각형은 전체자료를 이용하여 구축한 Model 2의 모형을 통한 예측값과 실측값의 관계를 나타낸다. Blue Lake와 Orick 두측정지점 모두에서 낮은 Q와 높은 Q의 구간에 대하여 각각 최적화된 Model 1이 Model 2에 비해 1:1 선에 상대적으로 근접하여 분포하는 경향을 보이는 것을 시각적으로 확인할 수 있다.

3.4 Comparision with arbitrarily separated model

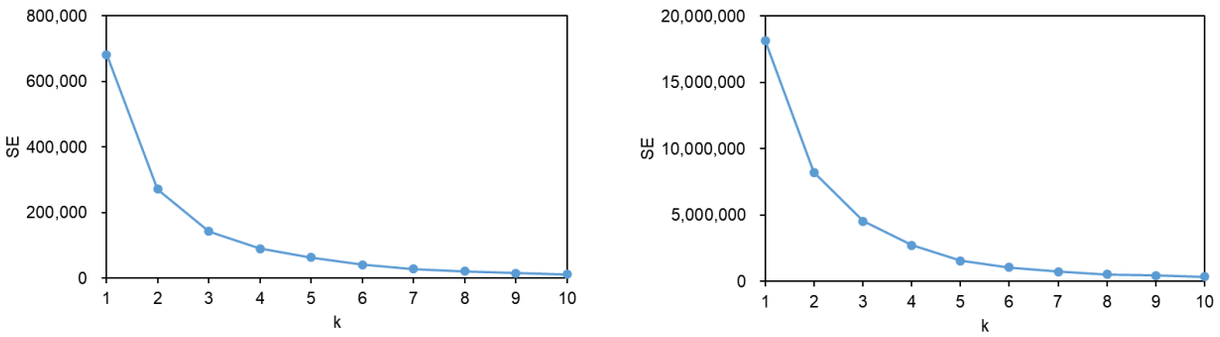
본 연구에서는 자료의 특성에 따라 입력자료를 구분하기 위해 KMC를 적용하였으며, KMC를 이용하지 않고 임의로 입력자료를 구분하여 구축된 모형과의 비교를 통해 KMC의 적용에 따른 모형 성능 개선 효과를 확인하였다.

Blue Lake와 Orick에서 각각 Q=2.6 ㎥/s 및 Q=10.5㎥/s를 기준으로 입력자료를 구분한 결과 각 지점에서 전체 training에 사용된 자료의 50%가 높은 Q와 낮은 Q 구간에 각각 분포하도록 구분이 되었다. 이후 Model 1과 유사한 방식으로 상위 50%와 하위 50% Q에 해당되는 구간에 대해서 각각 최적화를 수행하여 모형의 성능을 분석하였다. 분석결과 Blue Lake의 경우 RMSE와 RSR이 각각 112.35와 0.49로 분석되어, 전체자료를 사용한 Model 2에 비해서는 개선된 성능을 보였으나 KMC를 이용하여 군집화된 입력자료를 이용하여 구축된 Model 1에 비해서는 낮은 성능 개선효과를 보였다. Orick의 경우 RMSE와 RSR이 각각 124.73과 0.57로 전체 자료를 이용하여 구축된 Model 2와 유사한 모형 성능을 보였다. 모형의 training에 사용된 자료를 상위 50% 및 하위 50%로 구분하여 모형을 구축한 결과, 전체자료를 적용하는 모형에 비해 다소 성능이 개선되거나 거의 개선되지 않은 것으로 분석되어, KMC를 이용하여 입력자료를 군집화하여 모형을 구축하는 경우와 차이가 있음을 확인할 수 있었다.

3.5 Optimal clustering

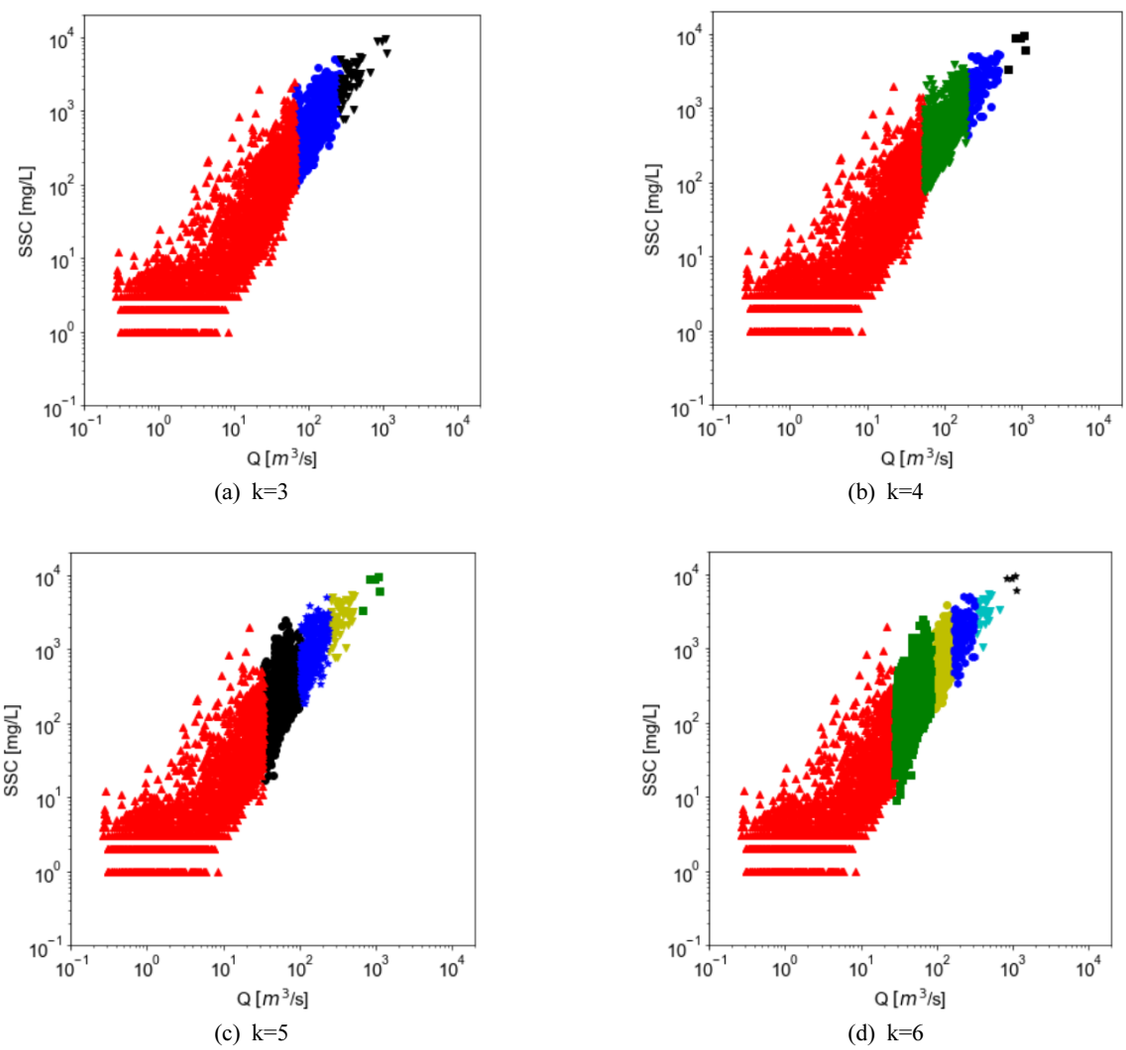
Elbow 알고리즘은 KMC를 이용하여 군집수 k를 늘려가면서 각 k에서의 오차의 제곱합(sum of squared error, SE)를 구하고 k의 증가에 따른 SE의 감소율이 적어지는 지점을 최적의 군집수로 결정하여 입력자료의 최적 군집수를 산정하는 방법이다(Park, 2018; Zhang, Bouadi et al.,2018). 모형 구축에 사용된 자료의 특성을 고려한 최적군집수를 확인하기 위해 elbow 알고리즘을 이용한 최적군집수 분석을 수행하였다. 이를 위해 KMC를 이용하여 training에 사용된 자료의 Q를 기준으로 1~10개로 군집수 k를 증가시켜가면서 SE의 변화를 분석하였다. 입력자료를 2개의 군집으로 구분한 경우 SE가 초기값의 절반 이하로 급격히 감소하였으며 이후 군집수가 증가함에 따라 SE가 지속적으로 감소하였으나 군집수 k=6 이후 SE 변화율이 크지 않아 최적의 군집수는 6개 내외 정도임을 확인할 수 있었다(Fig. 5).

KMC을 통해 모형 구축에 사용된 training 자료를 k=3~6개의 군집으로 나눈 결과 가장 낮은 Q 범위에 가장 많은 자료가 분류되는 것을 확인할수 있었다(Fig. 6). 각 군집별로 분류된 자료의 비율은 군집수 k에 따라 차이가 있었다. 군집수 k=3일 경우 Blue Lake와 Orick에서 각각 전체자료의 78% 및 85%가 가장 낮은 Q의 범위로 분류되었으며, 군집수가 커짐에 따라 그 비율이 줄어들어 k=6일 경우 Blue Lake와 Orick에서 각각 전체자료의 62% 및 67%가 가장 낮은 Q의 범위로 분류되었다. 가장 높은 Q의 범위에는 가장 작은 수의 자료가 분포하여 k=3일 경우 Blue Lake에는 134일, Orick에는 50일간의 측정자료가, k=6일 경우 Blue Lake에서는 1일 Orick에는 4일간의 측정자료가 가장 높은 Q구간으로 분류되었다. 군집수를 3개 이상으로 진행하는 경우 머신러닝 모형의 구축에는 자료가 충분하지 않아 본 연구에서는 추가적인 군집별 모형구축은 수행하지 않았다.



(a) Blue Lake (b) Orick

Fig. 5. Result of the elbow analysis.



(a) k=3 (b) k=4

(c) k=5 (d) k=6

Fig. 6. Distribution of the clustered training data in Redwood Creek at Orick, California USA.
*Note: Each color represents different cluster.

머신러닝 모형의 성능은 입력자료의 특성에 다양한 영향을 받게 된다. 본 연구에서는 군집화 모형을 이용하여 입력자료의 특성을 반영한 군집화를 통한 전처리를 수행하여 머신러닝 모형의 성능을 향상시킬수 있는 가능성을 확인하였다. 향후 입력자료의 다양한 특성의 반영을 통해 머신러닝 모형의 성능을 개선할 수 있는 지속적인 연구가 필요할 것으로 생각된다.

4. Conclusion

본 연구에서는 KMC을 이용하여 입력자료의 특성에 따른 군집화를 수행하고 XGB를 이용하여 SSC를 예측하는 모형(Model 1)을 구축하고 입력자료의 군집화가 모형성능에 미치는 영향을 분석하였다. 모형의 구축에는 미국 California Redwood Creek에 위치한 USGS 현장측정소 Blue Lake와 Orick 2개소에서 장기간 측정된 Q와 SSC 일일 측정자료를 활용하였다. 모형의 성능은 RMSE 및 RSR을 이용하여 평가하였다. 비교를 위하여 입력자료의 특성을 고려한 군집화를 적용하지 않고 전체 입력자료를 사용한 모형(Model 2)를 구축하여 예측 성능을 분석하였다.

모형의 수행 결과 입력자료의 특성을 고려하지 않은 Model 2는 Blue Lake와 Orick 각각에서 RSR이 0.51 및 0.57로 분석되었으며, 군집화를 통해 입력자료를 Q가 낮은 경우와 높은 경우의 2개 군집으로 구분하여 각각의 입력자료에 최적화시킨 Model 1의 경우 RSR이 Blue Lake와 Orick에서 각각 0.46과 0.55로 개선되는 것을 확인하였다. RMSE도 RSR과 마찬가지로 Model 1이 더 좋은 성능을 보이는 것으로 분석되어, 입력자료의 특성을 고려한 모형의 구축을 통해 머신러닝 모형의 성능이 개선되는 사례를 확인할 수 있었다. 향후 입력자료의 다양한 특성을 반영하여 머신러닝 모형의 성능을 향상시킬 수 있는 지속적인 연구가 필요할 것으로 생각된다.

Acknowledgement

본 논문은 2021년도 정부(국토교통부)의 재원으로 국토교통과학기술진흥원의 지원을 받아 수행된 연구입니다(21UGCP-B157942-02).

References

Ahmad, A. and Dey, L. (2007). A k-mean clustering algorithm for mixed numeric and categorical data, *Data & Knowledge Engineering*, 63, 503-527.

Ayub, J., Ahmad, J., Muhammad, J., Aziz, L., Ayub, S., Akram, U., and Basit, I. (2016). Glaucoma detection through optic disc and cup segmentation using k-mean clustering, 2*016 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube)*, 143-147.

Bennett, N. D., Croke, B. F., Guariso, G., Guillaume, J. H., Hamilton, S. H., Jakeman, A. J., Marsili-Libelli, S., Newham, L. T., Norton, J. P., and Perrin, C. (2013). Characterising performance of environmental models, *Environmental Modelling & Software*, 40, 1-20.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, Association for Computing Machinery, 785-794.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine, *Annals of statistics*, 29(5), 1189-1232.

Gray, A. B., Pasternack, G. B., Watson, E. B., Goni, M. A., Hatten, J. A., and Warrick, J. A. (2016). Conversion to drip irrigated agriculture may offset historic anthropogenic and wildfire contributions to sediment production, *Science of the Total Environment*, 556, 219-230.

Gray, A. B., Pasternack, G. B., Watson, E. B., Warrick, J. A., and Goñi, M. A. (2015). The effect of El Niño Southern Oscillation cycles on the decadal scale suspended sediment behavior of a coastal dry-summer subtropical catchment, *Earth Surface Processes and Landforms*, 40, 272-284.

Haghiabi, A. H., Nasrolahi, A. H., and Parsaie, A. (2018). Water quality prediction using machine learning methods, *Water Quality Research Journal*, 53, 3-13.

Hicks, D. M., Gomez, B., and Trustrum, N. A. (2000). Erosion thresholds and suspended sediment yields, Waipaoa river basin, New Zealand, *Water Resources Research*, 36, 1129-1142.

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

Li, L., Rong, S., Wang, R., and Yu, S. (2021). Recent advances in artificial intelligence and machine learning for nonlinear relationship analysis and process control in drinking water treatment: A review, *Chemical Engineering Journal*, 405, 126673.

Lin, W., Sung, S., Chen, L., Chung, H., Wang, C., Wu, R., Lee, D., Huang, C., Juang, R., and Peng, X. (2004). Treating high-turbidity water using full-scale floc blanket clarifiers, *Journal of Environmental Engineering*, 130(12), 1481-1487.

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., and Veith, T. L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations, *Transactions of the American Society of Agricultural and Biological Engineers*, 50(3), 885-900.

Muhammad, S. Y., Makhtar, M., Rozaimee, A., Aziz, A. A., and Jamal, A. A. (2015). Classification model for water quality using machine learning techniques, *International Journal of software engineering and its applications*, 9, 45-52.

Packman, A. I. and MacKay, J. S. (2003). Interplay of stream-subsurface exchange, clay particle deposition, and streambed evolution, *Water Resources Research*, 39(4), 1097.

Park, J. (2021). Comparative characteristic of ensemble machine learning and deep learning models for turbidity prediction in a river, *Journal of Korean Society of Water and Wastewater*, 35, 83-91. [Korean Literature]

Park, J. and Hunt, J. R. (2017). Coupling fine particle and bedload transport in gravel-bedded streams, *Journal of Hydrology*, 552, 532-543.

Park, J. and Lee, H. (2020). Prediction of high turbidity in rivers using LSTM algorithm, *Journal of Korean Society of Water and Wastewater*, 34, 35-43. [Korean Literature]

Park, R. K. (2018). An empirical comparison and verification study on the containerports clustering measurement using k-means and hierarchical clustering (average linkage method Using Cross-Efficiency Metrics, and Ward Method) and Mixed Models, *Journal of Korea Port Economic Association*, 34, 17-52. [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.

Shin, Y., Kim, T., Hong, S., Lee, S., Lee, E., Hong, S., Lee, C., Kim, T., Park, M. S., and Park, J. (2020). Prediction of chlorophyll-a concentrations in the Nakdong river using machine learning methods, *Water*, 12, 1822.

Singer, M. B., Aalto, R., James, L. A., Kilham, N. E., Higson, J. L., and Ghoshal, S. (2013). Enduring legacy of a toxic fan via episodic redistribution of California gold mining debris, *Proceedings of the National Academy of Sciences*, 110, 18436-18441.

Song, J. (2017). K-means cluster analysis for missing data, *Journal of Korean Data Analysis Society,* 19, 689-697. [Korean Literature]

Stevenson, M. and Bravo, C. (2019). Advanced turbidity prediction for operational water supply planning, *Decision Support Systems*, 119, 72-84.

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.

United States Geological Survey (USGS). (2009). *USGS(United States Geological Survey) Water-Data Report 2009*, 11482500 Redwood Creek at Orick, CA.

United States Geological Survey (USGS). (2014). *National Water Information System (NWIS)*. https://waterdata.usgs.gov/nwis (accessed Jun. 2014).

Walling, D. (1977). Assessing the accuracy of suspended sediment rating curves for a small basin, *Water Resources Research*, 13(3), 531-538.

Wang, Y., Chen, J., Cai, H., Yu, Q., and Zhou, Z. (2021). Predicting water turbidity in a macro-tidal coastal bay using machine learning approaches, *Estuarine, Coastal and Shelf Science*, 252, 107276.

Warrick, J. A. (2015). Trend analyses with river sediment rating curves, *Hydrological processes*, 29(6), 936-949.

Warrick, J. A., Madej, M. A., Goñi, M., and Wheatcroft, R. (2013). Trends in the suspended-sediment yields of coastal rivers of northern California, 1955–2010, *Journal of Hydrology*, 489, 108-123.

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.

Zhang, Y., Bouadi, T., and Martin, A. (2018). An empirical study to determine the optimal k in Ek-NNclus method, *5th International Conference on Belief Functions (BELIEF2018)*, 260-268.