ISSN(Print): 1225-7672 / ISSN(Online): 2287-822X
DOI https://doi.org/10.11001/jksww.2021.35.1.083

딥러닝과 앙상블 머신러닝 모형의 하천 탁도 예측 특성 비교 연구

Comparative characteristic of ensemble machine learning and deep learning models for turbidity prediction in a river

박정수*
Jungsu Park*

국립한밭대학교 건설환경공학과
Department of Civil and Environmental engineering, Hanbat National University

ABSTRACT

The increased turbidity in rivers during flood events has various effects on water environmental management, including drinking water supply systems. Thus, prediction of turbid water is essential for water environmental management. Recently, various advanced machine learning algorithms have been increasingly used in water environmental management. Ensemble machine learning algorithms such as random forest (RF) and gradient boosting decision tree (GBDT) are some of the most popular machine learning algorithms used for water environmental management, along with deep learning algorithms such as recurrent neural networks. In this study GBDT, an ensemble machine learning algorithm, and gated recurrent unit (GRU), a recurrent neural networks algorithm, are used for model development to predict turbidity in a river. The observation frequencies of input data used for the model were 2, 4, 8, 24, 48, 120 and 168 h. The root-mean-square error-observations standard deviation ratio (RSR) of GRU and GBDT ranges between 0.182~0.766 and 0.400~0.683, respectively. Both models show similar prediction accuracy with RSR of 0.682 for GRU and 0.683 for GBDT. The GRU shows better prediction accuracy when the observation frequency is relatively short (i.e., 2, 4, and 8 h) where GBDT shows better prediction accuracy when the observation frequency is relatively long (i.e. 48, 120, 160 h). The results suggest that the characteristics of input data should be considered to develop an appropriate model to predict turbidity.

Key words: Deep learning, Drinking water supply systems, Ensemble machine learning, Recurrent neural network, Turbidity management

주제어: 딥러닝, 정수 공급 시스템, 앙상블 머신러닝, 순환신경망, 탁도 관리

Received 7 December 2020, revised 31 December 2020, accepted 6 January 2021.
*Corresponding author: Jungsu Park(E-mail: parkjs@hanbat.ac.kr)

• **박정수 (조교수) / Jungsu Park (Assistant Professor)**
대전광역시 유성구 동서대로 125, 34158
125, Dongseo-daero, Yuseong-gu, Daejeon 34158, Republic of Korea

1. 서 론

최근 몇 년간 머신러닝(machine learning) 등 고도화된 자료 분석 기술의 환경분야 적용이 빠르게 증가하고 있다. 실시간 측정 센서 기술의 발달로 데이터의 취득이 용이해지고 하천 수질 및 상하수도 운영 등 물환경분야 관련 측정 자료의 데이터 베이스(DB: data base) 구축이 활발해지면서 이러한 DB를 기반으로 한 수질변화 예측 및 공정 운영 최적화 등에 다양한 머신러닝 알고리즘(algorithm)이 적용되고 있다. 초기 머신러닝 알고리즘 중 하나인 인공신경망(ANN: Artificial Neural Networks)이 하천 및 해안 등에서 녹조 발생의 지표인 클로로필-*a*(Chl-*a*) 예측에 사용되거나(Huang et al., 2015; Park et al., 2015; Wu et al., 2014), 서포트벡터머신(SVM: Support Vector Machine)이 하천 유량과 BOD, Chl-*a* 등 수질항목 예측에 활용되는 등 머신러닝은 물환경 관리를 위한 모형 구축에 다양하게 활용되어왔다 (Kisi, 2012; Liu and Lu, 2014; Park et al., 2015; Singh et al., 2011).

머신러닝 중 의사결정나무(decision tree) 알고리즘 기반의 앙상블(ensemble) 모형인 random forest(RF)와 이후 개발된 gradient boosting decision tree(GBDT) 모형은 분류와 회귀 방식 모두에 적용이 가능하고, 딥러닝 모형 등에 비해 상대적으로 모형의 구축이 복잡하지 않으면서도 수질 예측 등에 좋은 성능을 보여 점차 사용이 늘고 있다 (Hollister et al., 2016; Park et al., 2020; Zhang et al., 2018).

또한, 최근 몇 년간 딥러닝(deep learning) 알고리즘 중 자연어 처리 등 복잡한 데이터 분석에 뛰어난 성능을 보이는 순환신경망(RNN: Recurrent Neural Networks) 모형의 사용이 활발하게 이루어지고 있으며 시계열 자료의 분석과 예측에도 좋은 성능을 보여 물환경분야에서도 Chl-*a*와 하천 탁도 등 수질 항목의 예측에 활용되는 사례가 점차 늘고 있다 (Park and Lee, 2020; Shin et al., 2020; Zhou et al., 2018).

우리나라에서는 매년 하절기 태풍 및 집중호우 등에 의해 하천 탁도가 급격히 증가하는 현상이 반복되며 이러한 고탁수가 정수장에 유입될 경우 수처리를 위한 약품 투입 비용이 증가하고 및 수질 사고 발생의 우려가 커지게 된다. 따라서 이러한 고탁수의 발생을 사전에 예측하면 공정 운영의 안전성을 높일 수 있다.

본 연구에서는 최근 활발히 사용되고 있는 앙상블 머신러닝 알고리즘 중의 하나인 GBDT와 딥러닝 RNN 알고리즘 중 하나인 GRU를 이용하여 하천에서의 탁도 변화를 예측하는 모형을 구축하여 그 적용성을 확인하고 입력 자료의 특성이 모형의 성능에 미치는 영향을 비교하였다.

2. 재료 및 실험방법

2.1 모형 개요

2.1.1 GRU 모형

RNN은 이전 단계의 정보를 현단계의 연산에 적용하여 다음 단계의 변화를 예측하도록 구성된 딥러닝 알고리즘 중 하나이다 (Mikolov et al., 2011; Zaremba et al., 2014). 하지만, RNN은 연산을 위한 은닉층(hidden layer)의 증가 시 모형의 가중치 계산과 최적화를 위한 역전파(backpropagation) 과정에서 기울기 손실(vanishing gradient)이 발생하여 모형의 성능이 저하하는 문제가 있다 (Greff et al., 2016). LSTM은 forget gate, input gate, output gate의 3가지 내부 연산 알고리즘을 가지고 있으며 이전 단계의 정보를 장기간 저장하는 cell state와, 단기간에 활용하는 hidden state에서 선택적으로 연산에 사용될 값과 사용하지 않을 값을 결정하도록 모형을 구성하여 RNN의 기울기 손실 문제를 보완한 모형이다 (Greff et al., 2016; Hochreiter and Schmidhuber, 1997). 이후 개발된 GRU (Gated Recurrent Unit) 모형은 reset gate와 update gate의 2개의 gate 구조를 가지고 있으며 cell state 및 hidden state로 나누어진 LSTM과 다르게 cell state가 없고 hidden state만으로 구성되어 상대적으로 단순한 구조를 가지고 있는 모형이다 (Cho et al., 2014). GRU 모형의 각 gate의 역할은 다음과 같다 (Fig. 1).

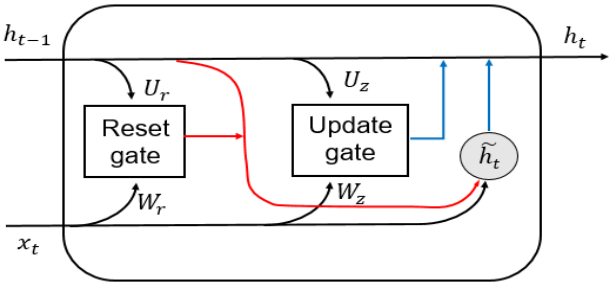


Fig. 1. A simple schematic of GRU cell structure.

1) Reset gate: 이전 단계 hidden state의 정보 중 현 단계의 연산에 사용하지 않을 정보를 제거하는 단계이다. Reset gate는 이전 단계 hidden state의 출력값(h_{t-1})과 현단계의 입력 자료(x_t)에 각각 가중치(U_r및 W_r)를 적용 후 sigmoid 함수를 이용하여 0~1 범위의 결과를 출력한다 (Cho et al., 2014).
2) Update gate: reset gate와 유사하게 이전 단계의 hidden state의 출력 값(h_{t-1})과 현단계의 입력 자료(x_t)에 각각 가중치(U_z및 W_z)를 적용 후 sigmoid 함수를 이용하여 0~1의 범위로 출력하여 현 단계의 연산에 과거 hidden state의 정보와 현단계의 입력 자료를 사용하는 비율을 결정한다 (Cho et al., 2014).

Cho et al. (2014)은 reset gate와 update gate를 각각 다음과 같이 제시하였다 (Eq. 1과 2).

$$r_j = \sigma\{(W_r x)_j + (U_r h_{t-1})_j\} \quad (1)$$

$$z_j = \sigma\{(W_z x)_j + (U_z h_{t-1})_j\} \quad (2)$$

where

σ: a logistic sigmoid function,

j: denotes the j-th element of a vector.

GRU 모형은 0~1사이의 update gate의 출력값을 통해 이전 단계 hidden state의 정보 중 현단계에서 사용할 정보를 정하여 최종 출력값 h_t를 계산한다 (Fig. 1).

2.1.2 GBDT 모형 개요

GBDT는 RF와 함께 널리 쓰이는 대표적인 앙상블 머신러닝 모형이다 (Zhang et al., 2018). RF는 bagging 방법을 이용하여 모형의 입력 자료를 무작위로 선정 후 다수의 의사결정나무를 생성한다 (Genuer et al., 2010). 생성된 의사결정나무는 각각 독립적으로 예측 결과를 산정하게 되며 최종적으로 각 의사결정나무에서 생성된 결과를 평균하여 RF 모형의 최종 결과값을 산출한다 (Breiman, 2001; Genuer et al., 2010). GBDT 모형은 각각의 의사결정나무를 독립적으로 생성하는 RF와는 다르게 이전 단계의 잔류 오차(residual errors)를 현단계의 의사결정나무의 생성에 적용하여 모형의 성능을 향상시키도록 구성되어 있다 (Chen and Guestrin, 2016; Friedman, 2001; Shin et al., 2020; Zhang et al., 2018). GBDT는 Eq. 3과 같이 각각의 의사결정나무에서의 실측값($y_{obs,i}$)와 모형의 예측값($y_{pred,i}$)의 차이를 계산하는 손실함수(L: Loss Function)와 모형의 구축시 생성된 K개의 의사결정나무에 대한 regulation 함수(Ω)로부터 산출된 목적 함수(J: Objective Function)을 최적화하여 모형의 최종결과를 산출한다 (Shin et al., 2020; Zhang et al., 2018).

$$J = \sum_{i=1}^{n} L(y_{obs,i}, y_{pred,i}) + \sum_{k=1}^{K} \Omega(f_k) \quad (3)$$

2.2 모형 구축

2.2.1 GRU 모형 구축

GRU 모형은 TensorFlow(version 2.3) 환경에서 Keras 기반의 알고리즘을 활용하여 구축하였으며, 프로그램 구성은 python(version3.5.4)를 이용하였다. 모형은 독립변수인 유량(Q)을 이용하여 종속변수인 탁도(T)를 예측하며, 전체 입력 자료의 80%를 모형의 학습(train)에 20%는 예측 결과의 검증(test)에 사용하도록 구성하였다. 모형의 time step은 1로 적용되어 전단계 유량 및 탁도값인 Q_{t-1}과 T_{t-1}로부터 시간 t에서의 탁도값인 T_t를 예측하게 된다 (Fig. 2). GRU 모형의 최적 hyper-parameter는 grid search 방식의 시행착오법(trial and error method)을 이용하여 결정하였다.

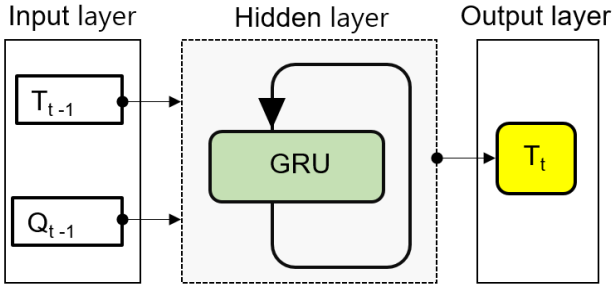


Fig. 2. A simple schematic of GRU model.

2.2.2 GBDT 모형 구축

GBDT 모형은 python open-source library인 Scikit-learn에서 제공되며, 가장 일반적으로 사용되는 GBDT 알고리즘인 XGBoost를 이용하여 구축하였다 (Chen and Guestrin, 2016; Pedregosa et al., 2011; XGBoost; Zhang et al., 2018). 모형의 최적화는 Scikit-learn의 grid search library를 이용 하였으며, Q를 독립변수로 하여 종속변

수 T를 예측하였다. GRU 모형과 마찬가지로 학습과 예측 결과의 검증에 사용된 자료의 비율은 0.8:0.2로 구성하였다.

2.3 입력 데이터 구축

본 연구에서는 입력 자료 측정 빈도에 따른 딥러닝과 앙상블 모형의 예측 특성을 비교하였으며, 이를 위해 미국지질조사국(USGS: United States Geological Survey)이 미국 California의 Russian River에 위치한 Guerneville 관측소(USGS site number: 11467000)에서 2014년 1월 1일부터 2019년 12월 31일까지 15분 간격으로 측정하여 공개한 Q 및 T 자료를 이용하였다(https://waterdata.usgs.gov/nwis) (Fig. 3). Russian Rivers는 미국 California 북쪽에서 남쪽 방향의 San Francisco 만으로 흐르는 California에서 2번째로 긴 강으로 총연장은 약 180 km, 유역면적은 약 3,850 km이다. Russian River 유역은 지중해성 기후 지역에 위치하고 있으며 봄부터 가을까지는 건조하고 겨울철부터 시작되어 봄까지 지속되는 우기에 주로 강우가 발생하며 연평균 강수량은 약 700 mm 연평균 기온은 약 15℃정도이다(USGS, 2011). 미국지질조사국 Guerneville 관측소는 Russian River 하류부에 위치하며 상류유역 면적은 약 3,465 km^2이다.

입력 자료의 측정 빈도가 모형의 예측 성능에 미치는 영향을 분석하기 위해 15분 간격 입력 자료의 2, 4, 8, 24, 48, 120 및 168시간 간격 평균을 구하여 총 7가지 유형의 입력 자료를 구축하고, 각 빈도별 자료의 평균, 최대, 최소 및 중앙값을 Table 1에 제시하였다. Ensemble 기반의 GBDT 모형은 입력 자료를 직접 사용할 수 있어 별도의 정규화를 하지 않았으며 GRU 모형은 입력 자료를 최소 0에서 최대 1 사이의 값으로 정규화하여 모형 구축에 이용하였다.

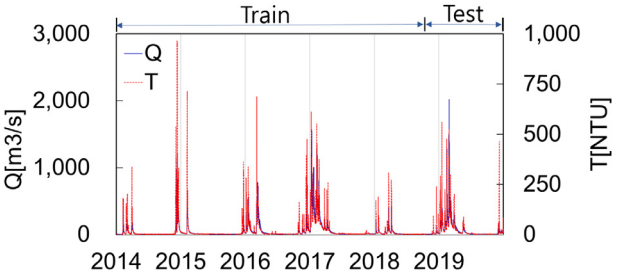


Fig. 3. Discharge and turbidity with 2 h observation frequency at the Guerneville station on the Russian River, California USA observed between 2014 and 2019.

2.4 모형 성능 비교 검정

GRU와 GBDT를 이용한 하천 탁도 예측 모형의 성능 비교를 위해 평균 제곱근 오차(RMSE: Root Mean Square Error)와 평균 제곱근 오차-관측값 표준편차비(RSR: Root Mean Squared Error- Observation Standard Deviation Ratio)를 이용하여 2가지 모형 각각에 대해

Table 1. Statistical characteristics of input variables at various observation frequencies

Frequency (hr)	Variable	Average	Max	Min	Median
2	T	18.76	965.00	0.21	3.15
	Q	54.89	2,017.59	0.10	7.95
4	T	18.76	960.00	0.25	3.14
	Q	54.89	2,015.29	0.10	7.96
8	T	18.76	951.88	0.29	3.13
	Q	54.89	2,012.81	0.13	7.97
24	T	18.76	604.58	0.36	3.12
	Q	54.89	1,895.09	0.15	7.84
48	T	14.05	361.67	0.38	3.15
	Q	54.89	1,417.71	1.13	7.92
120	T	14.03	290.00	0.36	3.17
	Q	54.79	1,152.39	1.65	8.36
168	T	14.08	241.65	0.43	3.21
	Q	54.98	983.10	1.83	7.81

2, 4, 8, 24, 48, 120 및 168시간 간격 입력 자료로 수행된 simulation 결과 산출된 예측값($T_{t,pred}$)을 같은 기간의 측정값($T_{t,obs}$)과 비교하였다 (Eq. 4와 5).

$$RSME = \sqrt{\frac{\sum_{t=1}^{n}(T_{t,obs} - T_{t,pred})^2}{n}} \quad .(4)$$

$$RSR = \frac{\sqrt{\sum_{t=1}^{n}(T_{t,obs} - T_{t,pred})^2}}{\sqrt{\sum_{t=1}^{n}(T_{t,obs} - \overline{T_{t,obs}})^2}} \quad (5)$$

where

$T_{t,obs}$: Observed turbidity at time t,

$T_{t,pred}$: Predicted turbidity at time t.

RMSE는 머신러닝 알고리즘의 성능을 평가하는데 널리 사용되는 지표중 하나로 숫자가 작을수록 모형의 결과가 실측값을 잘 예측함을 나타낸다. RSR은 0~1의 범위를 가지며 0에 가까운 값을 가질수록 모형의 성능이 더 좋은 것을 의미한다. RSR이 0.7 이하일 경우 구축된 모형이 실측값을 잘 예측한 것으로 판단하며 정해진 범위의 값을 계산하게 되므로 서로 다른 특성을 가진 모형의 상대적 비교가 가능한 장점이 있다 (Bennett et al., 2013; Moriasi et al., 2007).

3. 결과 및 고찰

3.1 탁도 예측 결과

GRU와 GBDT 모형이 성능에 입력 자료의 관측 빈도가 미치는 영향을 분석하기 위해 2, 4, 8, 24, 48, 120 및 168시간 관측 빈도로 구축된 7개의 입력 자료별 탁도 예측 simulation을 수행하였다. 검증자료에 대한 모형의 성능은 RSR과 RSME를 지표로 평가하였으며, RSR은 GRU 모형과 GBDT 모형이 각각 0.182~0.766 및 0.400~0.683의 범위를 RMSE는 GRU 모형과 GBDT 모형이 각각 9.679~34.943와 13.762~34.558의 범위를 가지는 것으로 분석되었다 (Table 2).

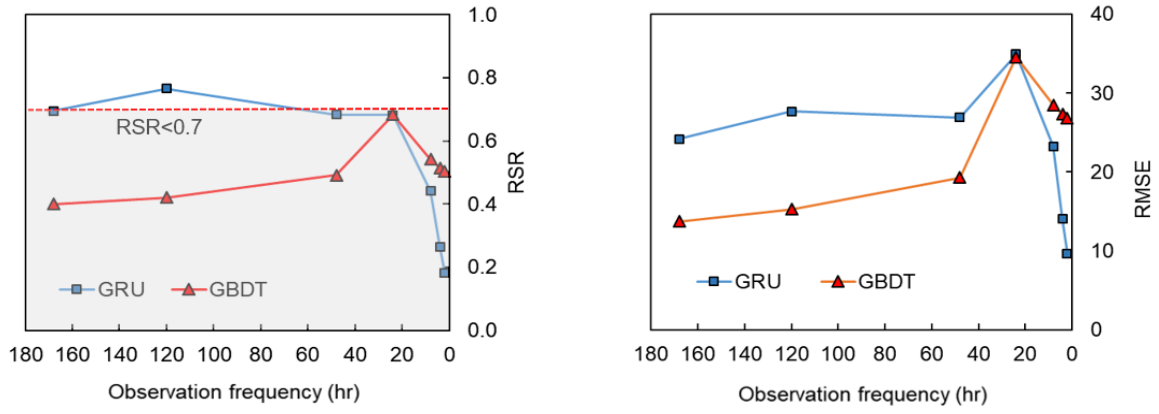
전체적으로 GRU 모형이 GBDT 모형에 비해 RSR과 RMSE의 분포가 큰 것으로 확인되었으며 GRU의 RSR이 120 시간 관측빈도에서 0.766으로 산정된 경우를 제외하면 두 모형 모두 모든 simulation 조건에서 RSR<0.7로 탁도값을 잘 예측하였다.

3.2 모형 성능 비교분석

GRU 모형의 경우 입력 자료의 관측 빈도가 높을수록 성능이 향상되는 경향을 보여 168시간 관측빈도에서 RSR과 RSME가 각각 0.696, 24.198로 가장 낮은 탁도 예측 성능을 보였으며 관측 빈도가 높아지면서 성능이 향상되어 2시간 관측 빈도에서 RSR과 RMSE가 각각 0.682 및 34.943으로 가장 좋은 성능을 보였다 (Table 2와 Fig. 4). GBDT 모형은 168시간 관측 빈도일 때 RSR과 RMSE가 각각 0.400 및 13.762로 가장

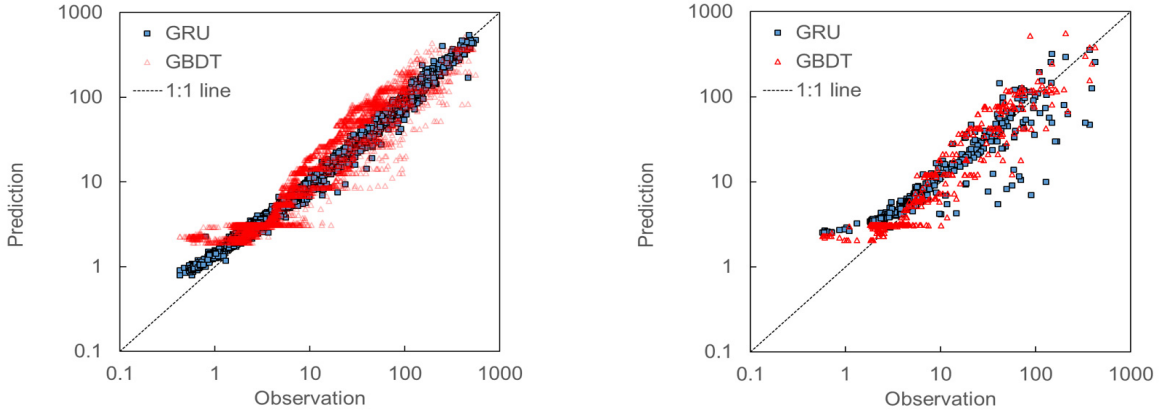
Table 2. Model simulation results at various observation frequencies

No.	Frequency (hr)	RSR		RMSE	
		GRU	GBDT	GRU	GBDT
1	2	0.182	0.504	9.679	26.798
2	4	0.265	0.515	14.045	27.312
3	8	0.442	0.542	23.251	28.532
4	24	0.682	0.683	34.943	34.558
5	48	0.684	0.492	26.893	19.299
6	120	0.766	0.422	27.674	15.295
7	168	0.696	0.400	24.198	13.762
Max		0.766	0.683	34.943	34.558
Min		0.182	0.400	9.679	13.762
Average		0.531	0.508	22.955	23.651



(a) RSR (b) RSME

Fig. 4. Model simulation results at various observation frequencies.



(a) 2 h (b) 24 h

Fig. 5. Comparison of simulation results between 2 h and 24 h observation frequencies.

좋은 성능을 보였으며, 이후 관측 빈도가 높아질수록 예측 성능이 낮아지는 경향을 보이다가 24시간 관측 빈도에서 RSR과 RMSE가 각각 0.683 및 34.558로 가장 낮은 예측 성능을 보인 후, 관측 빈도가 더 높아지면 다시 예측 성능이 향상되는 것으로 확인되었다.

관측 빈도가 24시간인 경우 GRU와 GBDT의 RSR이 각각 0.682와 0.683, RMSE는 각각 34.943과 34.558로 유사한 탁도 예측 성능을 보였으나 관측 빈도가 더 높은 2~8시간 간격 자료의 simulation 결과에서는 RSR이 GRU는 0.182~0.442, GBDT는 0.504~0.542로 GRU가 GBDT 보다 좋은 예측 성능을 보이는 것으로 확인되었다. 하지만 관측 빈도가 상대적으로 낮은 48~168시간 간격 자료의 simulation 결과에서는 반대로 GBDT가 GRU보다 좋은 예측 성능을 보였다.

GRU와 GBDT 모형 simulation 결과의 RSR과 RMSE를 비교한 결과 측정 빈도가 높은 경우 GRU 모형이 좋은 성능을 보였으나 측정 빈도가 낮은 경우 반대로 GBDT가 좀 더 좋은 예측 성능을 보였으며, 전체적으로는 앙상블 모형인 GBDT가 측정 빈도에 따른 RSR의 변동 폭이 적은 것으로 분석되었다. Fig. 5에 GRU와 GBDT의 탁도 예측값 분포를 비교하였다. GBDT는 측정 빈도에 관계없이 측정값(observation)과 예측값(prediction)이 유사한 분포를 보이는 반면 GRU는 24시간 측정 빈도에 비해 2시간 측정 빈도에서 측정값과 예측값의 오차가 줄어들어, 1:1 line에 근접하여 분포하였다.

3.3 입력 자료의 측정 빈도 등을 고려한 모형의 선정

하천에서의 고탁수 발생은 정수처리공정, 하천 수질 및 어류 서식 환경 등에 다양한 영향을 미친다

(Asrafuzzaman et al., 2011; Park et al., 2017; Suttle et al., 2004). 따라서 고탁수 발생 등 하천 탁도 변화의 예측은 정수장 운영 및 수질관리에 중요하며 이를 위해서는 현황에 대한 지속적인 모니터링이 필요하다.

우리나라의 대표적인 수질 모니터링 DB인 환경부 국립환경과학원의 물환경정보시스템(http://water.nier.go.kr)에는 전국 하천 및 호소에서 장기간에 걸쳐 주기적으로 측정된 수질 자료가 공개되어있다. 하천 및 호소 일반지점의 경우 월 1회, 주요지점의 경우 주 1회(연 48회) 빈도로 측정된 부유물질(SS) 측정 결과를 공개하고 있으며, 비점오염원 관리를 위해 운영중인 자동측정망의 경우 1시간 간격 탁도 측정 자료를 제공하는 등 수질측정의 목적 및 대상 지점에 따라 실시간 측정, 일일 측정 및 월간 측정 등 다양한 빈도의 측정 결과를 제공하고 있다.

본 연구에서는 딥러닝과 앙상블 머신러닝 알고리즘을 이용하여 하천 탁도 변화를 예측하는 모형을 구축하고, 입력 자료의 특성에 따른 영향을 분석하였다. 딥러닝 알고리즘인 GRU의 경우 앙상블 머신러닝 알고리즘인 GBDT에 비해 모형의 구성이 좀 더 복잡하며 입력 자료의 측정 빈도가 높은 경우 좋은 성능을 보여 주었으나 낮은 측정 빈도의 자료를 이용할 경우 성능이 떨어지는 경향을 보였다. 머신러닝은 모형자체의 한계 및 분석을 위한 하드웨어 성능의 한계 등으로 개발이 침체되는 시기도 있었으나, 2000년대 이후 SVM, RF 및 GBDT 등 다양한 알고리즘의 개발이 활발해지고 실시간 측정 센서 등을 통해 측정된 자료가 증가하면서 폭넓은 분야에 지속적으로 활용되어왔다 (Ben-Hur et al., 2005; Pal, 2005; Vapnik, 1995; Zhang et al., 2017). 특히 Hinton et al (2006)이 신경망의 기울기 손실 문제를 개선하면서 딥러닝 발전의 획기적인 전기를 제시한 이후 딥러닝의 급속한 발전이 이루어지고 있으며, 최근 수년간 환경분야에서도 RF 및 GBDT 등 머신러닝 알고리즘과 LSTM 및 GRU 등의 딥러닝 알고리즘의 활용이 점차 증가하고 있다 (Shin et al., 2020; Zhang et al., 2018; Zhou et al., 2018). Ensemble 기반 머신러닝 모형과 딥러닝은 서로 다른 연산 알고리즘 구조를 가지고 있으며 복잡한 모형이 항상 좋은 성능을 보이는 것은 아니므로 입력 자료의 특성에 따라 적정한 모형을 선정하는 것이 필요하다. 또한 최근에는 이미지 분석에 널리 사용되는 CNN (Convolutional Neural Network) 알고리즘과 LSTM을 함께 적용하는 등 서로 다른 알고리즘을 복합하여 각 모형의 장점을 이용해 모형의 성능을 높이는 연구도 점차 늘고 있다 (Islam et al., 2020; Kim and Cho, 2019).

다양하게 개발되는 머신러닝 모형을 하천 탁도 등 수질변화 예측 모형구축에 적용하고 좋은 성능을 확보하기 위해서는 양질의 실측 자료 확보와 확보된 자료의 특성에 맞는 모형의 선정이 필요하다. 향후 지속적인 연구로 입력 자료의 특성을 고려한 모형의 선정과 다양한 모형의 장점을 이용한 복합 모형의 구성 등을 통해 예측 모형의 성능을 향상시킬 수 있을 것이다.

4. 결 론

본 연구에서는 딥러닝 알고리즘인 GRU와 앙상블 머신러닝 알고리즘인 GBDT로 구축된 모형의 입력 자료 측정 빈도에 따른 하천 탁도 예측 특성을 비교하였다. 두 모형 모두 전체적으로 하천 탁도 변화를 잘 예측하였으며, 측정 빈도가 높은 경우 GRU 모형이 GBDT 모형에 비해 좋은 예측 성능을 보이나, GBDT 모형이 측정 빈도에 따른 영향을 적게 받는 것으로 분석되었다. GRU 모형은 입력 자료 측정 빈도가 높은 경우, 상대적으로 측정 빈도가 낮은 경우에 비하여 실측값과 모형 예측값의 오차가 작아지며 1:1 line에 근사하는 경향을 보였으며, GBDT는 측정 빈도에 따른 차이가 상대적으로 크지 않았다.

본 연구를 통해, 입력 자료의 특성이 딥러닝과 앙상블 머신러닝 알고리즘을 이용한 예측 모형의 성능에 미치는 영향을 분석하였으며, 입력 자료의 특성을 고려하여 적합한 모형을 선택하면 좀 더 좋은 예측 성능을 가지는 수질예측 모형의 구축이 가능함을 확인하였다.

사 사

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2020R1G1A1008377).

References

Asrafuzzaman, M., Fakhruddin, A., and Hossain, M.A. (2011).

Reduction of turbidity of water using locally available natural coagulants, ISRN Microbiol., 1-6.

Ben-Hur, A., Horn, D., Siegelmann, H.T., and Vapnik, V. (2001). Support vector clustering, J. Mach., 2(Dec), 125-137.

Bennett, N.D., Croke, B.F., Guariso, G., Guillaume, J.H., Hamilton, S.H., Jakeman, A.J., and Perrin, C. (2013). Characterising performance of environmental models, Environ. Modell. Softw., 40, 1-20.

Breiman, L. (2001). Random forests, Mach. Learn, 45, 5-32.

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*, 13-17 August, San Francisco, CA, USA. Association for computing Machinery.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation, 1078.

Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine, Ann. Stat., 29(5), 1189-1232.

Genuer, R., Poggi, J.M. and Tuleau-Malot, C. (2010). Variable selection using random forests, Pattern Recognit. Lett., 31, 2225-2236.

Greff, K., Srivastava, R.K., Koutník, J., Steunebrink, B.R., and Schmidhuber, J. (2016). LSTM: A search space odyssey, IEEE Trans. Neural Netw., 28(10), 2222-2232.

Hinton, G.E., Osindero, S., and The, Y.W. (2006). A fast learning algorithm for deep belief nets, Neural Comput., 18(7), 1527-1554.

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory, Neural Comput., 9(8), 1735-1780.

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

Huang, J., Gao, J., and Zhang, Y. (2015). Combination of artificial neural network and clustering techniques for predicting phytoplankton biomass of Lake Poyang, China, Limnol., 16, 179-191.

Islam, M.Z., Islam, M.M. and Asraf, A. (2020). A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images, Inform. Med. Unlocked, 100412.

Kim, T.Y., and Cho, S.B. (2019). Predicting residential energy consumption using CNN-LSTM neural networks, Energy, 182, 72-81.

Kisi, O. (2012). Modeling discharge-suspended sediment relationship using least square support vector machine, J. Hydrol., 456, 110-120.

Liu, M., and Lu, J. (2014). Support vector machine—an alternative to artificial neuron network for water quality forecasting in an agricultural nonpoint source polluted river?, Environ. Sci. Pollut. R., 21, 11036-11053.

Mikolov, T., Kombrink, S., Burget, L., Černocký, J., and Khudanpur. S., (2011). "Extensions of recurrent neural network language model", *In 2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 22-27 May, IEEE.

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, Am. Soc. Agric. Biol. Eng., 50, 885-900.

Pal, M. (2005). Random forest classifier for remote sensing classification, Int. J. Remote. Sens., 26(1), 217-222.

Park, J. and Lee, H. (2020). Prediction of high turbidity in rivers using LSTM algorithm, J. Korean Soc. Water Wastewater, 34, 35-43.

Park, H.S., Chung, S.W. and Choung, S.A. (2017). Analyzing the effect of an extreme turbidity flow event on the dam reservoirs in North Han River basin, J. Korean Soc. Water Environ., 33, 282-290.

Park, J., Park, J.H., Choi, J.S., Joo, J.C., Park, K., Yoon, H.C., Park, C.Y., Lee, W.H., and Heo, T.Y. (2020). Ensemble model development for the prediction of a disaster index in water treatment systems, Water, 12, 3195.

Park, Y., Cho, K.H., Park, J., Cha, S.M. and Kim, J.H. (2015). Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea Sci. Total Environ., 502, 31-41.

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, J. Mach. Learn. Res., 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.

Singh, K.P., Basant, N., and Gupta, S. (2011). Support vector machines in water quality management, Anal. Chim. Acta., 703, 152-162.

Suttle, K.B., Power, M.E., Levine, J.M., and McNeely, C. (2004). How fine sediment in riverbeds impairs growth and survival of juvenile salmonids, Ecol. Appl., 14(4), 969-974.

United States Geological Survey (USGS). (2011). Water-quality Data for the Russian River Basin, Mendocino and Sonoma Counties, California, 2005-2010, USGS, Report-data series 610.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. New York, Springer2-Verlag.

Wu, N., Huang, J., Schmalz, B. and Fohrer, N. (2014). Modeling daily chlorophyll a dynamics in a German lowland river using artificial neural networks and multiple linear regression approaches, Limnol., 15, 47-56.

XGBoost. Available online: https://xgboost.readthedocs.io/en/latest/build.html (February 15, 2020).

Zaremba, W., Sutskever, I. and Vinyals, O. (2014). Recurrent neural network regularization.

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, L., Tan, J., Han, D., and Zhu, H. (2017). From machine learning to deep learning: progress in machine intelligence for rational drug discovery, Drug Discov. Today, 22(11), 1680-1685.

Zhou, J., Wang, Y., Xiao, F., Wang, Y. and Sun, L. (2018). Water quality prediction method based on IGRA and LSTM, Water, 10, 1148.