



한강 수위 예측을 위한 데이터 품질 진단 및 개선

Data Quality Assessment and Improvement for Water Level Prediction of the Han River

최지현 · 강진엽 · 안현*

한신대학교 소프트웨어융합학부

Ji-Hyun Choi · Jin-Yeop Kang · Hyun Ahn*

Division of Software Convergence, Hanshin University, Gyeonggi 18101, Korea

[요 약]

최근 급격한 기후 변화 및 온난화로 인한 부작용으로 전 세계적으로 홍수 재해의 빈도 및 피해 규모가 증가하고 있다. 국내의 경우, 한강 수위는 대한민국 수도인 서울의 홍수 재해를 예방하기 위한 주요 관리 대상이다. 본 논문에서는 기계학습 기반의 한강 수위 예측을 개선하기 위해 관련 데이터 품질을 종합적으로 진단하고 이를 개선하기 위한 전처리 방안을 제안한다. 구체적으로는 결측치 처리와 교차 상관 분석을 통해 데이터를 완전성, 유효성, 그리고 정확성 측면에서 개선한다. 또한, 제안한 데이터 개선 방법이 한강 수위 예측 성능에 미치는 영향을 분석하기 위해 랜덤 포레스트와 LightGBM을 이용하여 성능 평가를 수행한다.

[Abstract]

As a side effect of recent rapid climate change and global warming, the frequency and scale of flood disasters are increasing worldwide. In Korea, the water level of the Han River is a major management target for preventing flood disasters in Seoul, the capital of Korea. In this paper, to improve the water level prediction of the Han River based on machine learning, we perform a comprehensive assessment of the quality of related dataset and propose data preprocessing methods to improve it. Specifically, we improve the dataset in terms of completeness, validity, and accuracy through missing value processing and cross-correlation analysis. In addition, we conduct a performance evaluation using random forest and LightGBM to analyze the effect of the proposed data improvement method on the water level prediction performance of the Han River.

Key word : Data preprocessing, Data quality assessment, LightGBM, Random forest, Water-level prediction.

<http://dx.doi.org/10.12673/jant.2023.27.1.133>



This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Received 21 January 2023; Revised 31 January 2023

Accepted (Publication) 21 February 2023 (28 February 2023)

*Corresponding Author ; Hyun Ahn

Tel: +82-31-379-0663

E-mail: hyunahn@hs.ac.kr

I. 서론

최근 급격한 기후 변화 및 온난화로 인한 부작용으로 전 세계적으로 홍수 재해의 빈도 및 피해 규모가 증가하고 있다. 국내의 경우, 한강 수위는 대한민국 수도인 서울의 홍수 재해를 예방하기 위한 주요 관리 대상이다. 이와 관련하여, 한강 본류에 위치한 팔당댐은 한강에 가장 근접한 댐으로서 한강 수위를 적절하게 관리하는 중요한 역할을 수행한다. 이러한 팔당댐의 운영에는 데이터에 기반한 여러 가지 예측 결과들이 영향을 미치는데 그중에서 주요 한강 다리에서 측정되는 수위에 대한 예측 결과는 중요한 의미를 가진다.

이와 관련하여 인공지능 관련 국내 경진대회 플랫폼인 DACON에서는 2022년 팔당댐 홍수 안전운영에 따른 한강 수위 예측 AI 경진대회를 개최한 바 있다[1]. 본 경진대회에서 다양한 참가자들이 최신의 기계학습 기법들을 적용하여 정확한 수위 예측 모델을 개발하고자 했지만, 데이터 품질에 대한 진단과 개선에 대한 이슈는 커뮤니티 내에서 활발하게 논의되지는 않았다. 그러나, 경진대회 자체의 목적보다 과거에서부터 앞으로 계속 생성될 한강 수위 관련 데이터 품질을 지속적으로 관리하는 것이 더 중요하다고 판단된다.

본 논문에서는 한강 수위 데이터셋에 데이터 품질 진단을 수행하고, 도출된 문제점을 해결하기 위한 데이터 개선 방안을 제안한다. 구체적으로는 결측치 처리, 유효하지 않은 값 처리, 그리고 교차 상관 분석 기반 시계열 시프트를 통해 데이터셋에서 발견된 데이터 완전성, 유효성, 그리고 정확성 문제들을 개선한다. 또한, 제안한 데이터 개선 방법이 한강 수위 예측 성능에 미치는 영향을 분석하기 위해 랜덤 포레스트와 LightGBM을 이용하여 성능 평가를 수행한다. 결론적으로 제안한 데이터 전처리 방법이 모든 측면에서 데이터를 개선하지는 않지만, 예측 성능 측면에서 긍정적인 기여를 할 것으로 기대된다.

본 논문의 구성은 다음과 같다. 2장에서는 수위 예측 관련 연구에 대해 소개한다. 3장에서는 한강 수위 데이터셋에 대한 품질 진단 결과에 대해 설명한다. 4장에서는 품질 진단 결과에 기반한 데이터 개선 및 전처리 방법을 제안한다. 5장에서는 실험을 통해 제안 방법의 효과를 검증하며, 마지막으로 6장에서는 결론 및 향후 연구 방향 소개로 글을 맺는다.

II. 관련 연구

국내외적으로 홍수와 같은 자연재해를 예방하기 위해 기계 학습 기술을 이용한 수위 및 강수량 예측과 관련된 연구들이 수행되었다[2]-[6]. 특히, 대규모 관측 데이터를 학습하여 우수한 예측 성능을 보이는 기계학습 및 딥러닝 기법들이 활발하게 적용되고 있다. 관련 연구[3]에서는 입력 지연 인공신경망(Input Delay Neural Network)과 같은 동적 인공신경망 모델을 적용하

여 실시간 도시 홍수 예측을 위한 성능 평가를 수행하였다. 이와 유사하게 관련 연구[2]에서는 시계열 예측에 적합한 순환 신경망(recurrent neural networks) 계열의 딥러닝 모델을 활용하여 미국의 트리니티강(Trinity River)의 침수 센서 데이터를 학습하고, 침수 가능성을 예측하였다.

수위 예측에는 딥러닝 모델 적용 외에도 예측에 필요한 다양한 데이터 수집 및 활용이 요구된다. 이와 관련하여 관련 연구 [6]에서는 태화강 상류의 강수량, 수위 및 댐 방류량, 그리고 하류의 조위 및 강수량 자료 등을 종합적으로 활용하여 태화교 수위를 예측하는 연구를 수행하였다. 본 연구에서는 한강 수위 예측을 위해 제공된 공개 데이터셋의 품질을 진단하여 한강 수위 예측 성능을 개선하기 위한 데이터 전처리 방법을 제안한다.

III. 데이터 품질 진단

3-1 데이터셋 설명

본 연구에서는 한강홍수통제소[7]에서 제공한 팔당댐 및 주요 한강 다리의 수위 관련 데이터를 이용한다. 데이터셋은 2012년~2022년 기간의 5월부터 10월 사이에 측정된 10분 단위의 시계열이며, 예외적으로 2022년의 경우 5월 1일부터 7월 18일까지의 데이터를 포함한다. 다음 표는 데이터셋 내의 변수에 대한 설명을 나타낸다.

표 1. 데이터셋 내의 변수에 대한 설명

Table 1. Descriptions of the variables in the dataset

Variable	Description
ymdhm	Timestamp (year, month, day, hour, minute)
swl	Water level of Paldang Dam (Elevation above sea level, unit: <i>m</i>)
inf	Inflow of Paldang Dam (m^3/s)
sfw	Water storage volume of Paldang Dam ($10K m^3$)
ecpc	Available water storage volume of Paldang Dam ($1M m^3$)
tototf	Outflow of Paldang Dam (m^3/s)
tide_level	Tide level of Ganghwa Bridge (<i>cm</i>)
wL_1018662	Water level of Cheongdam Bridge (<i>cm</i>)
fw_1018662	Flow rate of Cheongdam Bridge (m^3/s)
wL_1018680	Water level of Jamsu Bridge (<i>cm</i>)
fw_1018680	Flow rate of Jamsu Bridge (m^3/s)
wL_1018683	Water level of Hangang Bridge (<i>cm</i>)
fw_1018683	Flow rate of Hangang Bridge (m^3/s)
wL_1019630	Water level of Haengju Bridge (<i>cm</i>)
fw_1019630	Flow rate of Haengju Bridge (m^3/s)

3-2 데이터 품질 진단 방법

한강 수위 데이터셋의 품질을 진단하고 개선 방안을 도출하기 위해 한국데이터베이스진흥원에서 개발한 데이터 품질 진단 가이드라인[8]을 적용한다. 가이드라인에서 정의한 데이터

품질 기준은 데이터 완전성, 유일성, 유효성, 일관성, 그리고 정확성으로 구성되며 이에 대한 설명은 표 2와 같다.

표 2. 가이드라인에서 정의한 데이터 품질 기준[8]
Table 2. Data quality criteria defined by the guideline [8]

Quality criterion	Description
Completeness	There must be no missingness of mandatory variables
Uniqueness	Data items must be unique and not duplicated
Validity	Data items must meet the valid value range for each variable
Consistency	The structure, value, and expression of data times must be retained consistently
Accuracy	Data items must accurately represent objects in the real world

표 2의 품질 기준을 한강 수위 데이터에 적용한 결과는 다음과 같다.

- 완전성(completeness): 변수별로 결측치가 존재하므로 만족하지 못한다.
- 유일성(uniqueness): 중복된 데이터 항목이 없으므로 만족한다.
- 유효성(validity): 일부 변수에서 유효값 범위를 벗어난 측정치가 포함되어 있으므로 만족하지 못한다.
- 일관성(consistency): 모든 측정치들이 일관된 데이터 유형으로 저장되어 있으므로 만족한다.
- 정확성(accuracy): 팔당댐 변수와 다리 변수 사이의 시차(time lag)가 데이터에 반영되지 않았으므로 만족하지 못한다.

데이터 품질 진단 결과, 한강 수위 데이터셋의 경우 데이터 완전성, 유효성, 그리고 정확성을 만족하지 못하는 것으로 분석된다. 이에 따라 다음 장에서는 데이터 전처리 과정을 통한 데이터 품질 개선에 대해 설명한다.

IV. 데이터 전처리 및 개선

3장에서 도출한 한강 수위 데이터셋의 품질 문제를 개선하기 위해 본 장에서는 결측치 처리, 유효하지 않은 값 제거, 그리고 교차 상관 분석 기반 시프트를 통한 품질 개선 결과에 대해 설명한다.

4-1 결측치 처리

데이터 완전성은 데이터베이스 테이블의 필수 컬럼(예: 기본 키)에 NULL 값이 없는 경우 만족된다. 한강 수위 데이터의 경우, 필수 컬럼이 별도로 지정되어 있지 않으므로 모든 컬럼에 대해 완전성을 진단해야 한다. 이는 데이터셋의 결측치 분석 결과로 설명할 수 있으며, 표 3은 데이터셋의 변수별 결측치 수와

비율을 나타낸다.

표 3. 변수별 결측치 수와 비율
Table 3. The number and percentage of missing values in each variable

Variable	# of missing values	Missing rate (%)
ymdhm	0	0
swl	743	0.27
inf	743	0.27
sfw	743	0.27
ecpc	743	0.27
tototf	743	0.27
tide_level	4927	1.78
wl_1018662	59	0.02
fw_1018662	16380	5.93
wl_1018680	59	0.02
fw_1018680	196849	71.24
wl_1018683	59	0.02
fw_1018683	1279	0.46
wl_1019630	59	0.02
fw_1019630	59	0.02

그림 1은 데이터셋의 전체 결측 정보 시각화 결과이다. 팔당댐 변수 그룹(swl, inf, sfw, ecpc, tototf)과 다리별 수위 변수 그룹(점두사: wl)의 경우 각각 결측치 수와 결측 시기가 일치하므로 동일한 데이터 취득 체계에 의해 결측이 발생한 것으로 추정할 수 있다. 또한 잠수교 유량(fw_1018680)의 경우, 결측 비율이 71.24%이며, 2015년부터 완전 결측된 변수이다.

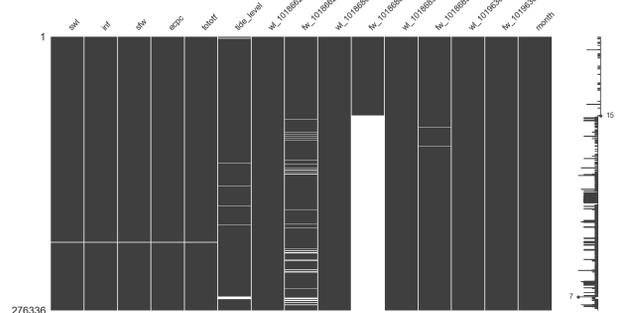


그림 1. 변수별 결측치 시각화
Fig. 1. Missing value visualizations by variable

데이터셋의 변수별 결측 현황을 토대로 다음과 같이 결측치들을 처리하였다. 먼저, 잠수교 유량의 경우 설명 변수로서 결측 비율이 높으므로 변수 삭제를 실시하였다. 팔당댐 변수 그룹과 다리별 수위 변수 그룹의 경우에는 결측 비율이 낮고 산발적 패턴을 보이므로 선형보간(linear interpolation)을 이용하여 처리하였다. 행주대교 유량(fw_1019630)의 경우, 다리별 수위와 비선형(non-linear) 관계인 동시에 단조(monotonic) 관계를 가지므로 다항회귀(polynomial regression)를 이용하여 결측치 대체를 수행하였다.

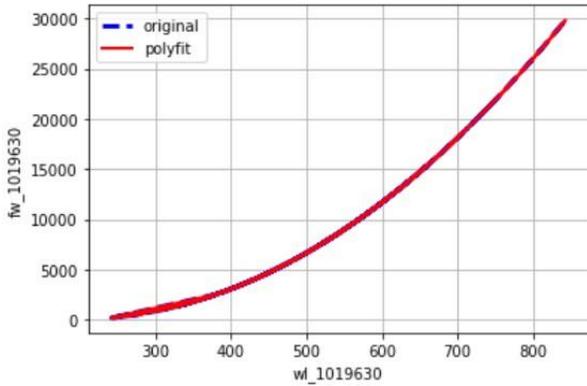


그림 2. 다항회귀 결과 예: 행주대교 유량
Fig. 2. Example of polynomial regression result: Flow rate of Haengju Bridge

강화대교 조위(tide-level)는 주요 한강 다리 수위에 영향을 주는 변수이며, 조석에 따라 약 12시간의 주기를 가지고 있다. 그러므로, 조위 변수의 결측치는 기본적으로 측정된 이전 주기의 조위 값으로 대체하였으며, 연속된 주기에 걸친 결측 등의 문제로 대체가 불가능한 경우 동일 시간대의 중앙값으로 대체하였다.

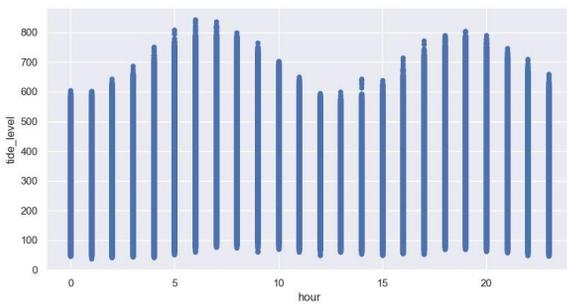


그림 3. 시간별 조위 분포
Fig. 3. Tide level distributions by hour

4-2 유효하지 않은 값 처리

데이터셋의 팔당댐 변수 그룹의 경우 모두 수치형 변수로서 $(0, \infty)$ 의 유효값 범위를 가진다. 그러나, 일부 변수들이 음수 값을 포함하므로, 표 4와 같이 데이터 유효성을 위반하는 사례들이 존재한다. 이를 해결하기 위해, 해당 값들을 0으로 처리하였다.

표 4. 변수별 음수 값의 개수
Table 4. The numbers of negative values by variable

Variable	# of negative values
inf	8994
sfw	1
tototf	5

4-3 교차 상관 분석 기반 시프트

교차 상관 분석(cross correlation analysis)은 두 시계열 간의 인과 관계를 파악하기 위한 분석 기법이다. 한 시계열에서 $\pm k$ 까지 시차가 적용된 시계열들을 각각 생성하고 대상 시계열과의 상관 계수를 모두 측정한다. 이를 통해 인과관계를 가지는 두 시계열 사이의 가장 상관관계가 강한 시차를 찾을 수 있다. 본 연구에서는 그림 4와 같이 팔당댐과 주요 한강 다리 사이의 물의 방향과 물리적 거리에 따른 시차를 모델링하기 위해 교차 상관 분석을 적용하여 시계열 시프트(shift)를 수행한다.

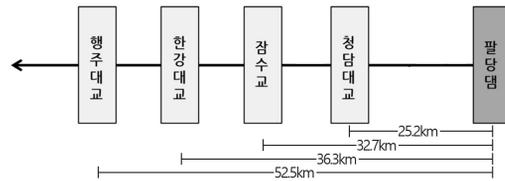


그림 4. 팔당댐과 한강 주요 다리 간의 공간 관계
Fig. 4. Spatial relationships between Paldang Dam and major bridges of the Han River

교차 상관 분석 예로서 그림 5는 팔당댐 방류량(tototf)과 청담대교 수위(wl_1018662) 변수 간의 교차 상관 분석 결과를 나타낸다. 방류량을 선행 변수로, 청담대교 수위를 후행 변수로 하고, 15시간 시차를 적용할 때 두 시계열의 상관 계수가 가장 높게 측정되었다. 이와 같은 방식으로 다른 변수에도 가장 상관 계수가 높은 시차로 시프팅을 수행하였다.

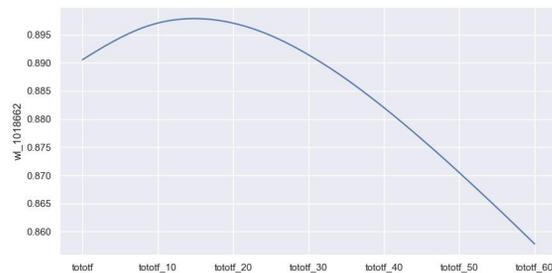


그림 5. 댐 방류량(tototf)과 청담대교 수위(wl_1018662) 변수 간의 교차 상관 분석 결과
Fig. 5. Cross correlation results between dam discharge (tototf) and Cheongdam Bridge water level (wl_1018662) variables

V. 실험 및 결과

4장에서 제안한 결측치 처리, 음수 제거 그리고 교차 상관 분석 기반의 시프트를 통해 데이터의 완전성, 유효성, 정확성 등

을 개선하였다. 본 장에서는 데이터 품질 개선의 효과를 검증하기 위해 주요 한강 다리 4개의 수위를 예측하는 모델을 생성하고 데이터 품질 개선 유무에 따른 성능 평가를 실시한다.

5-1 학습 알고리즘

본 실험에서는 수위 예측 모델을 생성하기 위해 두 가지의 앙상블 기반 학습 알고리즘을 활용한다. 첫 번째는 랜덤 포레스트(random forest [9]) 알고리즘으로 배깅(bagging) 계열의 앙상블 기법이다. 무작위 표본 추출을 통해 독립적인 다수의 의사결정나무(decision tree)들을 생성하여 종합적인 예측을 수행하므로 과적합에 강점을 가진다.

두 번째는 GBM(Gradient Boosting Machine) 계열의 학습 알고리즘인 LightGBM[10]이다. 해당 알고리즘은 기존의 GBM 계열의 특징인 균형 트리 분할 방식이 아닌 리프 중심 트리 분할 방식을 사용하는 모델로서 데이터 샘플과 특징 공간을 줄이는 연산을 수행한다. 소규모의 데이터에 대해 학습할 경우 과적합 위험이 있지만, 균형 트리 분할 방식에 비해 손실을 줄일 수 있고 적은 연산으로도 우수한 성능을 보인다. 모델 학습 관련 초매개변수 설정은 표 5와 같다.

표 5. 알고리즘별 초매개변수 설정
Table 5. Hyperparameter settings by algorithm

Algorithm	Hyperparameter	Value	Description
Random Forest	n_estimators	100	Number of trees
	max_depth	None	Maximum tree depth
	min_samples_split	2	Minimum number of samples required to split
	max_samples	None	Fraction of the original dataset is given to any individual tree
LightGBM	n_estimators	100	Number of trees
	max_depth	None	Maximum tree depth
	num_leaves	31	Maximum tree leaves
	subsample	1.0	Subsample ratio of the training instance

5-2 실험 결과

한강 수위 데이터 품질 개선을 통한 예측 성능 차이를 확인하기 위해 두 가지 경우로 학습 데이터를 구성하였다. 첫 번째는 데이터 품질 개선을 수행하지 않은 경우이다(Baseline). 음수 제거 및 변수 시프트를 수행하지 않았으며, 결측치의 경우 간단하게 평균 대체(mean substitution) 기법으로 처리하였다. 두 번째는 본 논문에서 제안한 데이터 품질 개선을 수행한 경우이다(Proposed). 음수 제거 및 교차 상관 분석 기반의 변수 시프트를 수행하였으며, 변수별 적절한 결측치 처리 방법을 적용하였다. 결론적으로 두 가지 데이터셋에 두 학습 알고리즘을 조합하여 총 4 가지의 모델을 학습하여 성능 평가를 수행하였다.

성능 평가지표로서 결정계수(R-squared, 이하 R^2)와 평균 제곱근 오차(Root Mean Squared Error, 이하 RMSE)를 적용하였다. 결정계수는 회귀모델이 데이터를 얼마나 잘 설명하는지를 나타내는 지표이며, 대상 변수와 회귀선의 변동(variation)을 비교하여 측정한다. 먼저, 대상 변수의 총 변동(total sum of squares, SS_{tot})은 관측값 y_i 와 관측값 평균 \bar{y} 과의 편차를 제곱한 값들의 총합이다.

$$SS_{tot} = \sum_{i=1}^n (y_i - \bar{y})^2 \tag{1}$$

잔차제곱합(residual sum of squares, SS_{res})은 관측값 y_i 와 모델의 예측값 \hat{y}_i 과의 차이를 제곱한 값들의 총합이다.

$$SS_{res} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \tag{2}$$

결과적으로 결정계수는 다음 수식과 같다.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{3}$$

모델의 예측 정확도를 나타내는 RMSE는 다음 수식과 같이 계산된다.

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}} \tag{4}$$

총 4개의 모델에 대해 성능 평가를 수행한 결과는 표 6과 같다. 데이터 품질을 개선한 경우, 랜덤 포레스트와 LightGBM 모델 모두 결정계수가 미세하게 증가하고 예측 오차는 각각 약 0.0745, 0.0677 씩 감소했으므로, 모델 성능은 약 2.1%, 1.7% 정도 개선되었다. 결론적으로 데이터 완전성, 유효성, 정확성을 개선하기 위한 결측치 처리, 음수 제거, 그리고 변수 시프트를 통해 수위 예측 모델의 성능을 개선한 것으로 분석된다.

표 6. 성능 평가 결과
Table 6. Performance evaluation results

Model	R^2	RMSE
RFa)_Baseline	0.9976	3.5856
RF_Proposed	0.9977	3.5111
LGBMb)_Baseline	0.9974	3.9340
LGBM_Proposed	0.9975	3.8663

a) RF: Random Forest
b) LGBM: LightGBM

VI. 결 론

본 논문에서는 한강 수위 예측 모델을 개선하기 위해 데이터셋에 데이터 품질 진단 기준을 적용하여 문제점을 발견하고 이를 개선하기 위한 전처리 방안을 제시하였다. 이를 검증하기 위해, 랜덤 포레스트와 LightGBM 학습 알고리즘을 이용하여 데

이더 개선을 수행한 경우와 그렇지 않은 경우로 나누어 모델 성능 평가를 수행하였다. 그 결과, 결측치 처리와 음수 제거, 그리고 교차 상관 분석을 통한 시프트로 이루어진 데이터 전처리 과정들이 예측 모델 성능을 개선하였다. 그러나, 개선의 정도는 미세하다고 평가되며, 그 이유로는 데이터셋에 포함된 다리별 유량의 설명력이 높기 때문에, 데이터 전처리 과정이 모델 성능에 큰 영향을 미치지 않은 것으로 분석된다.

향후에는 데이터 품질 기준과 연계되는 데이터 전처리 기법들에 대해 폭넓게 조사하여 통합적으로 데이터 품질을 진단 및 개선하기 위한 이론적 체계에 대해 연구하고자 한다.

Acknowledgments

본 연구는 과학기술정보통신부의 지원을 받아 수행되었음 (과제번호: NRF-2021R1F1A1063160).

References

[1] DACON [Internet]. Available: <https://dacon.io>

[2] Q. -K. Tran and S. -K. Song. "Water level forecasting based on deep learning: A use case of Trinity River-Texas-The United States," *Journal of KIISE*, Vol. 44, No. 6, pp. 607-612, June, 2017.

[3] G. -Y. Han, H. -J. Geum, and H. -I. Kim. "Building a data

driven model for real-time urban flood reduction," *Water for future*, Vol. 51 No. 11, pp. 13-19, Nov. 2018.

[4] S. Jung, et al. "Prediction of water level in a tidal river using a deep-learning based LSTM model," *Journal of Korea Water Resources Association*, Vol. 51, No. 12, pp. 1207-1216, 2018.

[5] S. A. A. Ojeda, A. S. Geoffrey, and C. P. Elmer, "Multivariate time series imaging for short-term precipitation forecasting using convolutional neural networks," in *Proceedings of 2020 International Conference on Artificial Intelligence in Information and Communication*. Fukuoka, Japan, pp. 33-38, 2020.

[6] M. Lee, et al. "Water level prediction in Taehwa River basin using deep learning model based on DNN and LSTM," *Journal of Korea Water Resources Association*, Vol. 54. No. S-1, pp. 1061-1069, 2021.

[7] Han River Flood Control Office [Internet]. Available: <https://www.hrfo.go.kr/>.

[8] *Data quality assessment procedure manual*, Korea Database Agency, 2009.

[9] T. K. Ho, "Random decision forests," in *Proceedings of 3rd international conference on document analysis and recognition*, Montreal: QC, pp. 278-282, 1995.

[10] G. Ke, et al., "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in neural information processing systems*, pp. 3149-3157, 2017.



최지현 (Ji-Hyun Choi)
2019년 3월 ~ 현재 한신대학교 소프트웨어융합학부 학사과정
※ 관심분야 : 데이터 공학, 빅데이터 분석



강진엽 (Jin-Yeop Kang)
2018년 3월 ~ 현재 한신대학교 소프트웨어융합학부 학사과정
※ 관심분야 : 데이터 공학, 빅데이터 분석



안현 (Hyun Ahn)
2021년 3월 ~ 현재 한신대학교 소프트웨어융합학부 조교수
2018년 3월 ~ 2021년 2월 경기대학교 컴퓨터공학부 연구교수
2017년 2월 경기대학교 컴퓨터과학과 이학박사
※ 관심분야 : 데이터 공학, 프로세스 마이닝, 시계열 분석