반도체디스플레이기술학회지 제22권 제3호(2023년 9월)
Journal of the Semiconductor & Display Technology, Vol. 22, No. 3. September 2023.

Savitzky-Golay 필터와 미분을 활용한 LSTM 기반 지하수 수위 예측 모델의 성능 비교

송근산[*]·송영진[*†]
[*†]건양대학교 의공학부

Performance Comparison of LSTM-Based Groundwater Level Prediction Model Using Savitzky-Golay Filter and Differential Method

Keun-San Song[*] and Young-Jin Song[*†]
[*†]Division of Biomedical Engineering, Konyang University

ABSTRACT

In water resource management, data prediction is performed using artificial intelligence, and companies, governments, and institutions continue to attempt to efficiently manage resources through this. LSTM is a model specialized for processing time series data, which can identify data patterns that change over time and has been attempted to predict groundwater level data. However, groundwater level data can cause sen-sor errors, missing values, or outliers, and these problems can degrade the performance of the LSTM model, and there is a need to improve data quality by processing them in the pretreatment stage. Therefore, in pre-dicting groundwater data, we will compare the LSTM model with the MSE and the model after normaliza-tion through distribution, and discuss the important process of analysis and data preprocessing according to the comparison results and changes in the results.

Key Words : Water resource management, Artificial intelligence, LSTM, Time series data, Groundwater level data, Data pre-processing

1. 서 론

수자원 관리에 있어서 인공지능을 활용하여 데이터 예측을 수행하고 기업, 정부, 기관 등이 이를 통해 자원을 효율적으로 관리하려는 시도는 지속적으로 이루어지고 있으며 예측 결과를 바탕으로 생산량, 수요 예측, 자원 분배 등을 최적화하여 비용을 절감하고 생산성을 향상시키는 방법이 지속적으로 연구되고 있다. 디지털 혁신에 따른 인프라에서 중요시되는 자원에는 물이 포함되어있으며[1] 도시성장을 위한 공공영역의 원동력으로 신기술 경쟁 등 필요성에 의한 노력이 심화될 전망이다[2]. LSTM은 시계열 데이터를 처리하는데 특화된 모델로, 시간에 따라 변화하는 데이터 패턴을 잘 파악할 수 있고 지하수 수위 데이터는 시간에 따른 연속적인 변화를 갖기 때문에 LSTM이 이를 효과적으로 다룰 수 있고 수위가 이전 수위에 영향을 받는 장기 의존성의 문제를 해결하기 위해 적합하기 때문에 제안되고 있는 모델이기도 하다[3, 4]. 지하수 수위 예측은 여러 변수(온도, 강수량, 지형 등)의 영향을 받을 수 있고 LSTM은 다양한 입력 변수를 동시에 처리하고 이들 간의 복잡한 상호작용을 파악하여 정확한 예측을 도출하기 위해서도 사용된다. 그러나 지하수 수위 데이터는 센서 오류, 누락값 또는 이상치가 발생할 수 있으며 이러한 문제들은 LSTM 모델의 성능을 저하시킬 수 있으며, 전처리 단계에서 이를 처리하여 데이터 품질을 향상시킬 필요성을 가지고 있다. 뿐만아니라 지하수 수위 데이터를 알맞은 시계열 형태로 변환하여 주입시키는 것

[†]E-mail: songjin@konyang.ac.kr

이 정확한 예측을 위한 핵심요소라고 할 수 있으며 후보정과 전처리 등을 통하여 안정적이고 정규화된 데이터를 입력하는 것이 시계열 데이터 예측에서 가장 중요한 부분이라고 할 수 있다. 따라서 본 논문에서는 지하수 데이터 예측을 함에 있어서 원신호를 LSTM모델에 학습시킨 것과 정규화 후에 모델에 학습시킨 것을 MSE(평균제곱오차)와 산포도를 통해 비교하고 비교결과에 따른 분석 및 데이터 전처리에 중요한 과정과 그에 따른 결과의 변화에 관해 논의하고자 한다. 본 논문에서 인공지능을 위한 프레임워크는 Python 언어 기반의 Anaconda사를 통해 배포되는 TensorFlow를 backend로 도입하였으며[5] 이를 위한 프론트엔드로는 Keras[6]를 사용하였다. 또한, 모델의 학습속도를 향상시키고자 NVIDIA CUDA 아키텍처[7]를 사용하였다.

2. 본 문

2.1 현장 설치

본 논문에서 필요로 하는 지하수의 수위데이터 측정 및 수집을 위해 실제 지하수를 사용하고 있는 현장에 설치하여 실험을 진행하기로 하였고 이에 따라 지하수 수위 측정 기능을 포함한 지하수 단말기를 관정에 위치한 측정공을 통해 케이블을 도하하여 설치하도록 한다. Fig. 1은 관정에 관측공을 삽입 시공하고 있는 장면을 촬영한 것이다.

Fig. 1. Observation hole construction.

2.2 LSTM 모델 정의

LSTM을 통해 지하수 수위데이터를 예측하기 위해 주어진 시퀀스내에서 장기의존성을 갖는 주기적 패턴으로 데이터를 구분 지을 수 있어야 하고 잡음이나 충분하지 않은 시퀀스를 갖는 일부데이터를 정규화를 통해 제거하거나 복잡한 특성을 제거하여 시계열 데이터 패턴을 학습하는데 활용이 적합한 데이터 전처리를 해야한다. 본 논문에서 연구한 장비는 측정된 지하수 수위를 바탕으로 향후 수위의 변화를 미리 예측하여 펌프를 동작시킬 수 있다.

아래 Fig. 2는 LSTM모델의 구조를 그림으로 나타낸 것이다.

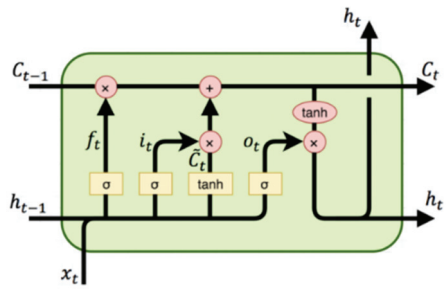

Fig. 2. LSTM Model Structure.

아래 Table 1은 학습에 사용한 LSTM 모델의 신경망 구조를 표로 나타낸 것이다.

Table 1. LSTM Neural Network Specifications

title	detail
window size	N(500 samples)
sampling frequency	100 Hz
features	24
Cost Function	Softmax Cross Entropy
Optimizer	Adam Optmizer
LSTM layers	2
hidden layers	32
Activation Function	ReLU
Weight regulation	L2 regularization
batch size	1500
Loss function	Softmax cross entropy with logits

Fig 3 은 Python Tensorflow Keras 환경에서 데이터를 학습시킨 화면이다.

Fig. 3. Model learning and test screen capture using Python.

3. 연구방법

3.1 Raw Data 학습

Fig. 4은 실제 수미현장에 시제품을 설치한 후에 3일동안 지하수 데이터를 측정한 것을 그래프로 나타낸 것이다. y축의 단위는 m 이고 지하수 수위 변화를 나타낸다. x축의 단위는 index로 데이터가 들어온 순서대로 기록된 것이며 측정 주기는 10분으로 설정되었다.

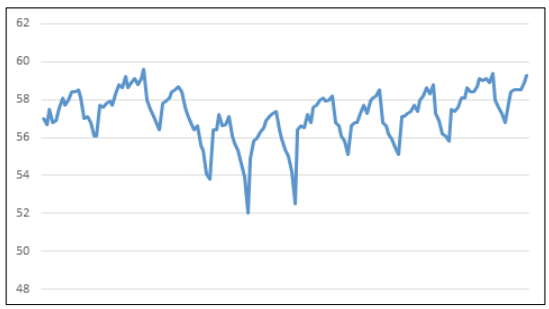

Fig. 4. Raw data of groundwater level in Soomi area.

실제 당진 수미 현장에서 측정한 3일간의 지하수 수위 데이터를 가지고 LSTM 모델을 이용하여 80 %는 훈련데이터로 사용하고 20 %를 테스트 데이터로 사용하였고 정확도를 확인하기 위해 MSE를 측정 후 산포도를 통해 시각적으로 나타내었다.

3.2 지하수 수위 변화량 데이터 학습

시계열데이터에 있어서 저주파대역의 데이터가 포함되어 있을 경우 예측과정에서 모델은 이것을 제거한 후의 데이터를 학습할 만한 능력이 없기 때문에 전처리 과정에서 저주파대역을 제거해야만 할 필요성이 있다. 수위데이터에 있어서 저주파는 장기적인 수위의 변화를 나타내므로 이를 간단하게 제거하기 위해서는 수위의 변화량만을 얻는 방법을 취할 수 있다. 즉 수위의 등락 데이터만 추출하기 위하여 원신호를 미분하여 아래 Fig. 5와 같이 정규화된 데이터를 얻고 이를 LSTM에 학습시켜 결과를 확인하였다. 수위데이터를 정규화한 결과 데이터를 가지고 LSTM 모델을 이용하여 80 %는 훈련데이터로 사용하고 20 %를 테스트 데이터로 사용하여 정확도를 확인하기 위해 MSE를 측정하였다.

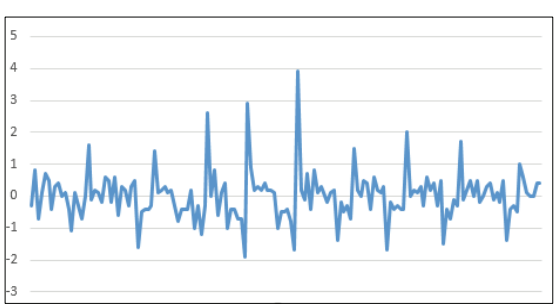

Fig. 5. Differential data of groundwater level in Soomi area.

3.3 Savitzky-Golay 필터 적용 데이터 학습

이어서 고주파데이터를 제거하기 위한 방법으로써 Savitzky-Golay 필터를 사용하여 데이터를 전처리 후 사용하기로 한다. 지하수 데이터에서 고주파데이터는 센서나 케이블, 통신 상에 발생할 수 있는 잡음으로써 진폭의 범위가 센서의 오차범위 내에서 이루어지기 때문에 전처리에서 필수적으로 제거해야 할 요소로는 여겨지지 않는다. 이에 따라 원신호를 Savitzky-Golay필터를 통하여 고주파신호를 제거하고 아래 Fig. 6과 같은 데이터를 얻었다.

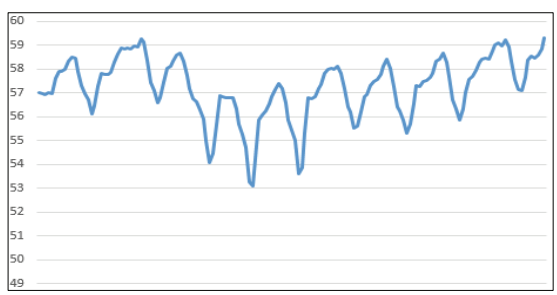

Fig. 6. Results of Savitzky-Golay Filter on Groundwater Level in Soomi area.

이처럼 Savitzky-Golay 필터를 이용해 획득한 데이터를 가지고 LSTM 모델을 이용하여 80 %는 훈련데이터로 사용하고 20 %를 테스트 데이터로 사용하여 정확도를 확인하기 위해 MSE를 측정하였다.

3.4 변화량 데이터에 Savitzky-Golay 필터 적용 후 학습

저주파 제거를 위한 미분과 고주파 제거를 위한 Savitzky-Golay 필터를 모두 적용시켰을 때 저주파제거를 위해 미분하여 유효한 결과를 얻었을 경우 고주파 제거는 저주파를 제거하지 않고 Savitzky-Golay필터를 적용하는 것은 저주파를 제거한 후에 Savitzky-Golay필터를 적용하였을 때 결과가 어떻게 달라지는지 평가하기 위함이다. 따라서 미분을 통해 저주파데이터를 제거한 이후 Savitzky-Golay필터를 적용하기로 하고 해당 결과데이터는 Fig. 7과 같다.

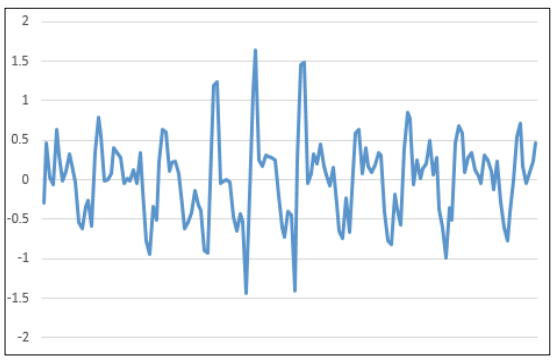

Fig. 7. Results of applying Savitzky-Golay filter after differentiating groundwater level in Soomi area.

위 데이터를 가지고 LSTM 모델을 이용하여 80 %는 훈련데이터로 사용하고 20 %를 테스트 데이터로 사용하여 정확도를 확인하기 위해 MSE를 측정하였다.

4. 연구결과

4.1 Raw Data 학습 결과

전처리 과정을 거치지 않고 측정한 MSE는 Train MSE의 경우 539.5221로 측정 되었으며 Test MSE의 경우 581.0892로 측정되었다. 아래 Fig. 8은 산포도와 같이 산포도와 인접한 기울기를 그래프로 나타낸 것이다.

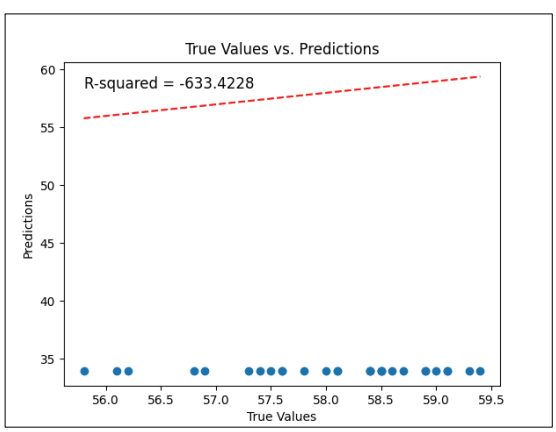

Fig. 8. Groundwater Level Raw Data and Prediction Data Distribution in Soomi Area.

이 때 원 신호의 표준 편차가 1.3189 인 것을 고려하였을 때 MSE가 상대적으로 매우 크게 나왔으며 일반적으로 MSE가 1이상의 값을 갖는 경우에는 모델의 예측이 실제 값과 크게 다른 오차를 갖는다는 의미로써 모델이 데이터를 잘 설명하지 못하고 있고 좋은 예측 결과를 얻기 위해서는 모델의 개선이나 다른 방법의 고려가 필요하다는 의미로 해석될 수 있다. 또한 실제 수위에 대한 데이터 예측이 어려운 상황이므로 전처리 과정이 필요한 것으로 확인된다. 지하수 수위 데이터는 일간 수위 변화 이외에 전체적인 수위가 등락하는 저주파 대역의 신호를 포함하여 이를 전처리 과정없이 LSTM 예측에 사용할 경우 앞서 확인한 결과와 같이 예측에 어려움을 겪을 수 있다. 또한 R-squared가 -633.4228 이므로 이는 예측이 정확하게 이루어지지 않았으며 해당 모델은 유효하다고 판단할 수 없음을 이야기할 수 있다.

4.2 지하수 수위 변화량 데이터 학습 결과

저주파대역을 제거하기 위한 미분 전처리 후 MSE를 측정한 결과 Train MSE는 0.4450로 측정되었고 Test MSE는 0.4132로 측정되었다. 전처리 과정을 거쳐 미분데이터로 활용하였을 때 MSE가 0에서 1사이의 값으로 나타나기 때문에 모델이 예측에 유효한 범위 내로 들어왔음을 확인할 수 있었다. 크게 개선되었음을 확인할 수 있고 이를 산포도로 나타내었을 때도 산포도 기울기가 유의한 것을 확인할 수 있다. 아래 Fig. 9는 산포도와 같이 산포도와 인접한 기울기를 그래프로 나타낸 것이다.

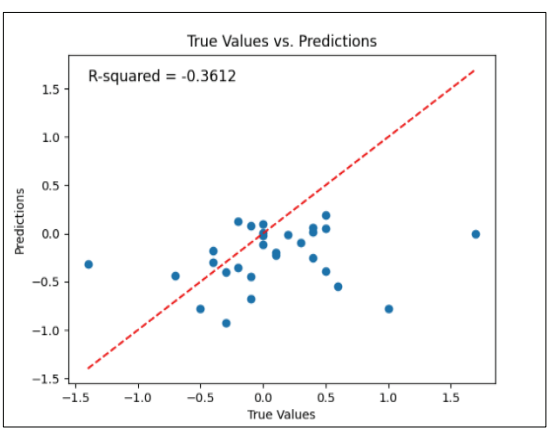

Fig. 9. Groundwater Level Differential Data and Prediction Data Distribution in Soomi Area.

MSE가 유효한 범위 내로 들어온 것에 비해서 이 때의 R-squared는 -0.3612로써 0에서 1사이의 값이 아닌 음수의 값을 지니므로 이는 예측이 정확하게 이루어지지 않았으며 해당 모델은 R-squared결과값을 놓고 보았을 때 유효하다고 할 수 없다.

4.3 Savitzky-Golay 필터 적용 데이터 학습 결과

이어서 Savitzky-Golay필터를 적용한 후에 데이터를 학습시키고 해당 결과를 확인하도록 한다. 해당 필터의 적용은 센서나 케이블에 의한 고주파 잡음을 제거하기 위함이다. 아래 Fig. 10은 산포도와 같이 산포도와 인접한 기울기를 그래프로 나타낸 것이다.

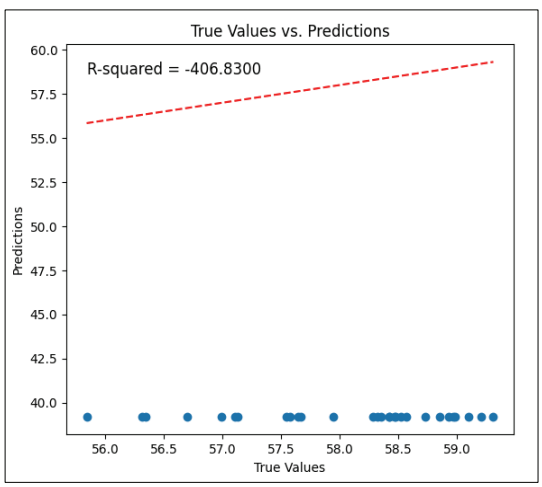

Fig. 10. Results of Savitzky-Golay Filter on Groundwater Level Data and Prediction Data Distribution in Soomi Area.

MSE를 측정한 결과 Train MSE는 323.3609로 측정되었고 Test MSE는 355.0203로 측정되었다. 고주파데이터를 Savitzky-Golay 필터를 통과시켜 제거한 결과의 MSE는 0에서 1사이의 값으로 나타나지 않았으므로 이는 유효한 모델 학습 결과로 판단할 수 없으며 이를 통해 예측이 불가능하다는 것을 의미한다. 또한 R-squared는 -406.8300으로 0에서 1사이의 값이 아니고 음수이므로 정확도를 표현한 R-squared 또한 예측 결과가 유효하지 않다는 것을 의미한다.

4.4 변화량 데이터에 Savitzky-Golay 필터 적용 후 학습 결과

미분을 통해 수위 변화량 데이터를 획득한 후에 이에 Savitzky-Golay필터를 적용한 데이터를 학습시키고 해당 결과를 확인하도록 한다. 아래 Fig. 11은 산포도와 같이 산포도와 인접한 기울기를 그래프로 나타낸 것이다.

MSE를 측정한 결과 Train MSE는 0.0898로 측정되었고 Test MSE는 0.0867로 측정되었다. 이러한 MSE의 결과값은 충분히 유효한 범위 내에 있으며 기존에 측정되었던 다른 전처리 결과값이나 원신호에 비해 예측 정확도가 우수하다는 것을 의미한다. 또한 R-squared값은 0.5061로 0에서 1사이의 유효한 범위 내에 포함되어 있으므로 미분 후에 Savitzky-Golay 필터를 적용하여 지하수 데이터를 예측하는 것이 유효한 의미를 갖는다고 해석할 수 있다.

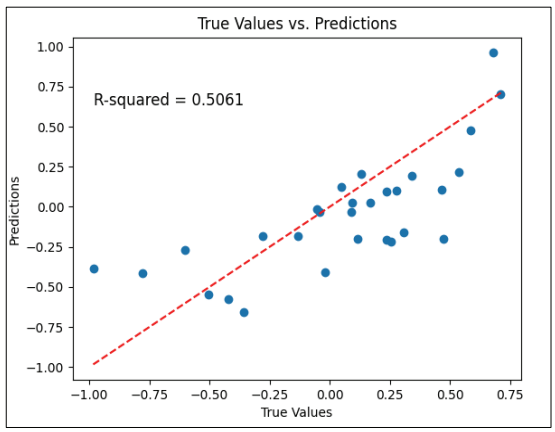

Fig. 11.

5. 결론 및 고찰

지하수 데이터는 많은 잡음을 포함하고 있을 뿐 아니라 주변의 많은 환경에 영향을 받는 굉장히 민감한 데이터이다. 그러나 지하수의 특성상 사람들이 자주 사용하는 시간대에 사용량이 급증하고 사람들의 물 사용량이 줄어드는 시간대가 되면 물 사용량이 급격하게 줄어들게 되는데 이는 일별로 시간에 따른 일정한 데이터의 일관성을 가지고 있어 본 논문처럼 고주파와 저주파를 제거하는 것만으로 시계열 데이터의 가치를 가지며 예측이 가능하다. 그러나 데이터의 특성상 잡음을 많이 포함하고 있으며 LSTM 예측 정확도에 영향을 끼치기 쉬운 복잡성을 갖고 있다는 문제가 있어 이를 위해서는 다양한 필터를 적용하여 유효한 전처리과정을 구성하고 준수해야할 필요성이 있다. 본 논문에서는 MSE와 산포도, R-squared 등을 통해 전처리된 데이터가 동일한 모델에 대해 얼마나 정확한 예측 결과를 갖는지 평가하였고 미분을 통해 저주파데이터를 제거하고 Savitzky-Golay 필터를 이용해 고주파데이터를 제거하여 전처리 하였을 때 일반적인 LSTM 모델을 사용하여 유효한 예측 결과를 얻을 수 있다는 결론에 도달하였다. 향후 논문에서는 더 많은 필터를 사용하고 다양한 현장에 대하여 MSE와 R-squared를 평가하는 것을 자동화 하여 실제 수위와 최대한 유사한 예측 값을 얻을 수 있는 다변량 수위 예측 알고리즘을 개발하고자 한다.

참고문헌

1. Son, H.S., “Governance study of intelligent information society by digital transformation”, Study of Public Law, Series 49 No. No. 3, pp. 205, 2021.
2. Sun, J.W., “Legislative tasks related to the 'AI Intelligent Government' to promote the Korean version of the New Deal”, Digital New Deal Issue Brief, pp. 2, 2020.
3. Zhang, Z., Zhu, Y., Zhang, X., Ye, M., and Yang, J., “Developing a Long Short-Term Memory (LSTM) based model for predicting water table depth in agricultural area.” Journal of Hydrology, Vol. 561, pp. 918-929, 2018.
4. Tran, Q., and Song, S., “Water level forecasting based on deep learning: A use case of trinity river-Texas-The United States.” The Journal of Korean Institute of Information Scientists and Engineers, Vol. 44, No. 6, pp. 607-612, 2017.
5. Helmus, J., “TensorFlow in Anaconda”, https://www.anaconda.com/bolg/developer-blog/tensorflow-in-anaconda, 2018.
6. Chollet, F., “Deep Learning with Python. Manning Publications”, p. 384., 2017.
7. NVIDIA., NVIDIA Tesla P100 Whitepaper. NVIDIA, p. 45. 2016.

접수일: 2023년 8월 23일, 심사일: 2023년 9월 6일,
게재확정일: 2023년 9월 12일