물 공급량 예측을 위한 데이터 마이닝 기법

신강욱* · 김연권

한국수자원공사

Data Mining for Water Supply Forecasting

Gang-Wook Shin* · Youn-Kwon Kim

K-water

E-mail : gwshin@kwater.or.kr / 201commando@hanmail.net

요 약

본 논문에서는 물 공급량 예측을 위한 다양한 알고리즘 적용에 있어서 데이터 마이닝의 효용성을 검토하고자 하였다. 물 공급분야에 있어서, 물 이용 지역의 특성에 따라 공급량과 이용 시간이 매우 상이한 특성을 나타낸다. 물 이용 지역은 주택지역, 상업지역, 산업단지지역 등 다양한 형태로 분류할 수 있고, 이에 따라 물 이용 시간의 상이함에 따른 물 공급패턴이 일정하지 않게 된다.

특히, 주택지역과 상업지역이 복합적으로 이루어진 경우, 물 이용 단위인 블록 단위에서의 물 특성이 불규칙적인 패턴을 나타낸다. 따라서, 각 블록 단위 특성에 적합한 물 이용량을 예측하여 효율적 물 공급 방안을 마련할 필요가 있다. 또한, 물 이용량 데이터 중 이상 데이타 감지와 이상 데이터 보정을 통하여 물 이용량 예측의 정확도가 향상된다. 따라서, 블록 단위의 물 이용량에 대한 원시데이타의 효율적인 데이터 마이닝 방안이 요구된다.

본 연구에서는 물 공급지역의 특성에 따른 물 공급 패턴을 분석하고, 이에 적합한 데이터 마이닝 기법을 제시하고 비교 분석하였다. 제안된 데이터 마이닝 기법은 딥러닝 예측모델을 적용하여 적합성을 검증하고, 이를 물 공급량 예측알고리즘에 폭넓게 활용될 수 있음을 확인하였다.

키워드

물공급, 예측, 딥러닝, 블록시스템

Ⅰ. 서 론

수요예측은 이동통신망, 전력 공급망, 경제 성장 등 다양한 분야에서 연구되고 응용되고 있다.[1] 물 분야에서의 수요예측은 물 이용량에 따른 생산량 결정에 활용된다. 일 단위 혹은 주 단위 물 공급을 위한 생산량 결정은 수요량 예측에 기반으로 이루어지며, 이는 안정적인 물 공급과 생산비 절감에 기여하는 바가 매우 높다. 이러한 물 수요 예측에 대한 연구는 시계열 모델, 칼만필터, 뉴럴모델 등 다양하게 진행되고 있다.[2-3]

또한, 딥러닝 알고리즘은 수요예측 뿐만 아니라 빅 데이터 처리를 통한 최적제어 및 의사결정시스템등 다양한 분야에 적용되고 있다.[4] 본 연구에서는 딥러닝 알고리즘 중 GRU(Gated Recurrent Units) 알고리즘을 활용하여 수요예측 기법의 적용성을 확인하고자 한다. 또한, 물 관리시스템으로부터 취득된 원시 데이타의 다양한 물 공급 패턴에 적합한 데이터 마이닝 기법을 제안하고 이를 적용하여 물 수요예측의 정확도를 향상하고자 한다.

Ⅱ. 물 공급시스템

일반적인 물 공급시스템의 계통도는 그림 1과 같이 표현할 수 있다. 물 공급시스템의 세부적인 구성은 하천으로부터 물을 취수하는 취수장, 이를 정수장으로 공급하는 가압장, 그리고 먹는 물을 생산하는 정수장, 그리고 정수된 물을 공급하는 배수지 등으로 구성된다. 본 연구에서는 물을 직접 이용하는 수용가에 물을 공급하는 배수지 계통에 대하여 고려한다. 특히, 배수지 유출량에 대한 예측과 배수지 계통의 블록시스템 공급량에 대한 예측을 고려한다.

* speaker

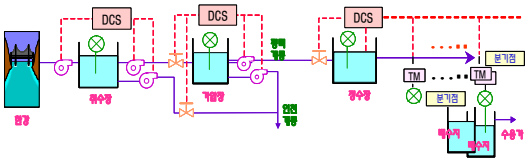
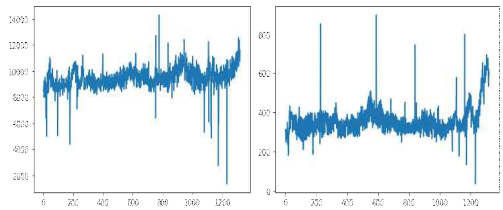


그림 1. 물공급시스템 계통도

배수지에서 수용가로 공급되는 계통은 해당 지역의 특성에 적합하도록 다수의 소규모 블록시스템으로 분할되어 있다. 본 연구에서 적용된 배수지에서는 7개의 블록시스템을 통하여 물 공급되도록 구성되어 있다. 현재 운영되고 있는 배수지와 샘플 블록에서의 실제 사용량은 그림 2와 같다. 대상 배수지 총 유출량은 그림 2(a)에서와 같이 일일 물 공급량은 평균 1만 ㎥이며 최대 약 1.2만 ㎥을 공급하고 있다. 또한, 대상 배수지 계통의 7개 블록 중 샘플 블록의 공급량은 그림 2(b)에 나타내었다. 일일 평균 약 400 ㎥, 최대 약 600 ㎥의 물을 공급하고 있음을 알 수 있다.



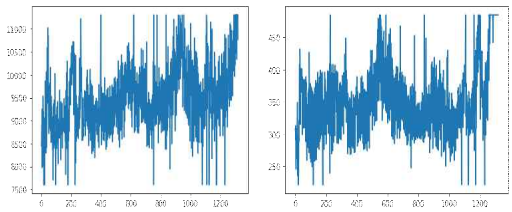
(a) 배수지 유출량 (b) 샘플 블록 공급량

그림 2. 원시 데이터

III. 데이터 마이닝

그림 2의 원시 데이타에서 다수의 이상 데이타가 존재함을 알 수 있고, 이에 대한 데이터 마이닝이 요구된다. 시계열 데이타의 데이터 마이닝은 정규분포를 이용하는 방법과 최소값부터 최대값까지 사분위 범위를 이용하는 방법이 가장 많이 사용되고 있다. 정규분포를 이용하는 방법은 정규분포에서 97.5 % 이상 혹은 2.5 % 이하에 포함되는 값, 즉 3-시그마 범위를 벗어나는 데이타를 이상치로 판별하는 방법이 있다. 그리고, 사분위 범위 IQR(InterQuartile Range)를 적용하는 방법은 Q1-1.5*IQR과 Q3+1.5*IQR의 범위를 벗어나는 데이터를 이상치로 판별하는 방법이다.

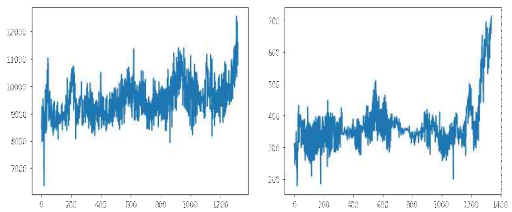
그림 2에서 대상 배수지와 샘플 블록에서의 물 공급량 추세를 살펴보면, 공급량이 최근에 증가하고 있음을 알 수 있다. 일반적인 IQR 방법을 적용하여 배수지 유출량과 블록 공급량에 대한 데이터 마이닝의 결과는 그림 3과 같다. 그러나, 최근에 공급량이 증가하고 있는 경우, 정상데이타를 이상치로 분류하는 오류가 발생되고 있음을 알 수 있다. 이는 정규분포를 이용하는 방법에서도 유사한 결과를 얻었다.



(a) 배수지 유출량 (b) 샘플 블록 공급량

그림 3. IQR 데이터 마이닝

제한된 범위에서 물 공급량이 증가하거나 감소하는 경우, 위 두 방식으로 이상 데이타를 감지하는 것이 부적절함을 알 수 있었다. 따라서, 본 연구에서는 물 공급 패턴분석 단위를 1개월 주기로 설정하였다. 이에 대한 데이터 변동성을 고려하여, 기존의 IQR 방법을 1개월 단위로 적응형 연산이 이루어지도록 제안하였다. 적응형 연산의 결과는 그림 4와 같다. 그림에서와 같이 최근 증가하는 물 공급 패턴이 잘 반영되었음을 알 수 있고, 타 방법에서의 오류를 대폭 개선할 수 있음을 알 수 있다.



(a) 배수지 유출량 (b) 샘플 블록 공급량

그림 4. 제안 데이터 마이닝

IV. 시뮬레이션

제안된 데이터 마이닝 기법에 대한 적용성을 검증하기 위하여 본 논문에서는 딥러닝 알고리즘 중 순환신경망인 GRU(Gated Recurrent Unit) 알고리즘을 적용하였다. 알고리즘 구조는 그림 5와 같이 대상배수지 데이터와 샘플 블록 데이터를 동시에 입력하도록 설계하였다. 이는 대상 배수지 총 유출유량에 대한 최적화를 위한 GRU 1과 블록 시스템에서의 최적화를 위한 GRU 2를 합성하도록 모델화하였다. 시뮬레이션은 원시 데이타를 입력한 결과와 제안된 데이터 마이닝을 거친 데이터 결과의 상호 예측률을 비교하였다.

각 GRU 세부 파라메타는 표 1과 같이 설정하였다. GRU 1과 GRU 2의 계층구조는 동일하게 설정하고, 과적합에 대한 최적화를 고려하여 Dropout을 0.1로 설정하였다. 입력 데이터는 2018년이후 약 44개월의 일일 운영데이타를 적용하였다.

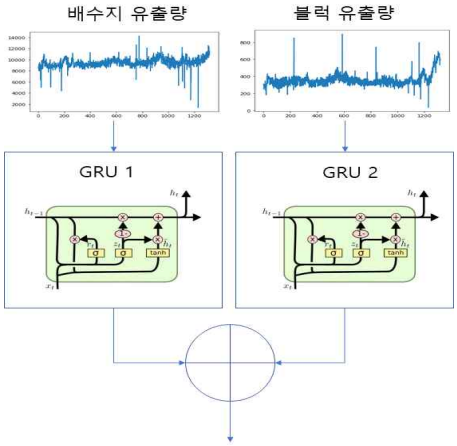


그림 5. Gated Recurrent Unit 합성 모델

표 1. GRU 파라메터

Network	Hyper Parameter
GRU 1	Hidden Layer : 4 Dropout : 0.1 Epoch : 100 Batch size : 20 Loss Function : MSE Optimizer : Adam
GRU 2	Hidden Layer : 4 Dropout : 0.1 Epoch : 100 Batch size : 20 Loss Function : MSE Optimizer : Adam
Merge	Merge node : 32 Output node : 1

시뮬레이션을 통한 예측결과에 대한 성능평가는 MAE(Mean Absolute Error)와 RMSE(Root Mean Squared Error) 지표를 다음과 같이 적용하였다.

$$MAE = \frac{1}{N}\sum_{k=1}^{N}|y_k - \widehat{y_k}| \quad (1)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(y_k - \widehat{y_k})^2} \quad (2)$$

표 2와 같이 MAE 지표의 경우 원시 데이타를 적용한 결과 대비 데이터 마이닝을 통한 시뮬레이션 결과는 48에서 36.6으로 약 20% 이상 개선되었음을 알 수 있다. RMSE와 R-square 지표에 있어서도 원시 데이타 대비 대폭 개선됨을 알 수 있었고, 이러한 시뮬레이션 결과를 통하여 제안된 데이터 마이닝의 효용성을 확인 할 수 있었다.

표 2. 딥러닝 예측결과

구 분	원시데이타	데이터 마이닝
MAE	48.0	36.6
RMSE	67.8	46.9
R-square	15.5	8.8

Ⅴ. 결 론

본 연구에서는 물 분야에서의 공급량 예측을 위한 효율적인 데이터 마이닝에 대하여 제안하였다. 제안된 데이터 마이닝의 효율성 검증을 위하여 딥러닝 예측알고리즘을 적용하였다. 시계열 예측기법 중 가장 성능이 우수한 GRU 알고리즘을 적용하였고, 모델 성능을 개선하기 위하여 GRU 합성모델을 설계하였다.

시뮬레이션 결과, 원시 데이타에서 발생되는 이상 데이타는 전체 수요 예측에 큰 영향을 미치는 것을 알 수 있었다. 제안된 데이터 마이닝 기법을 통하여 예측률이 약 20% 이상 향상되는 유효성을 확인할 수 있었고, 향후 시계열 데이터의 특성을 고려하여, 물 공급량 예측을 위한 데이터 마이닝 기법에 폭 넓게 적용될 수 있을 것으로 판단된다.

References

[1] J. L. Torres, A. Garcia, M. De Blas, and A. De Francisco, “Forecast of hourly average wind speed with ARMA models in Navarre (Spain),” *Solar Energy*, vol. 79, pp. 65-77, 2005.

[2] H. M. Al-Hama and S. A. Soliman, “Short-term electric load forecasting based on Kalman filtering algorithm with moving window weather and load model,” *Electric power systems research*, No. 68, pp. 47-59, 2004.

[3] G. S. Choi and G. W. Shin, “Short-term Water Demand Forecasting Based on Kalman filtering with Data Mining,” *Journal of Institute of Control, Robotics and Systems*, Vol. 15, No. 10, pp. 1506-1061, 2009.

[4] Amal M. and Ammar M., “A Survey on Deep Learning for Time-Series Forecasting,” *Machine Learning and Big Data Analytics Paradigms: Analysis, Applications and Challenges,* pp. 365-392, January 2021