LSTM 알고리즘을 이용한 수도데이터 정제기법

유기현*, 김종립*, 신강욱*

*한국수자원공사 K-water융합연구원

A Study on the cleansing of water data using LSTM algorithm

Yoo Gi Hyun*, Kim Jong Rib*, Shin Gang Wook*

*K-water Research Institute

E-mail : ghy135@kwater.or.kr, kjr8963@kwater.or.kr, gwshin@kwater.or.kr

요 약

수도분야에서는 정수장 및 관말 관로 상의 전 공정에서 유량, 압력, 수질, 수위 등 다양한 데이터를 수집하고 있다. 수집되는 데이터는 각 정수장 DB에 저장되며, 권역별 DB에서 합쳐져 수자원공사 본사의 DB 서버에 최종 저장된다. 측정기기가 데이터를 측정하거나 여러 과정에 걸쳐 데이터가 통신될 때 다양한 이상 데이터가 발생할 수 있으며 크게 결측 데이터와 오측 데이터로 분류할 수 있다. 각각의 이상 데이터의 발생원인은 상이하다. 따라서 오측 및 결측 데이터를 검출하는 방식에는 차이가 있으나 실제 이를 정제하는 방식은 동일하다. 본 연구에서는 딥러닝 알고리즘의 일종인 LSTM(Long Short Term Memory) 방식을 적용하여 오·결측 데이터를 자동으로 정제할 수 있는 프로그램에 대하여 고찰한다.

ABSTRACT

In the water sector, various data such as flow rate, pressure, water quality and water level are collected during the whole process of water purification plant and piping system. The collected data is stored in each water treatment plant's DB, and the collected data are combined in the regional DB and finally stored in the database server of the head office of the Korea Water Resources Corporation. Various abnormal data can be generated when a measuring instrument measures data or data is communicated over various processes, and it can be classified into missing data and wrong data. The cause of each abnormal data is different. Therefore, there is a difference in the method of detecting the wrong side and the missing side data, but the method of cleansing the data is the same. In this study, a program that can automatically refine missing or wrong data by applying deep learning LSTM (Long Short Term Memory) algorithm will be studied.

키워드
LSTM, 수도데이터, 오결측 데이터, 딥러닝

Ⅰ. 서 론

수도시스템에서는 각 공정별 측정 데이터 수집을 위해 정수장 및 분기점에 RTU를 설치하여 운영하고 있다. 각 현장별 RTU에서 수집된 수도데이터는 권역별 본부DB를 거쳐 최종적으로 본사 DB에 저장된다. 측정기기 계측부터 각 구간별 통신과정에서 다양한 이상데이터가 발생할 수 있다. 본 연구에서는 수도데이터 공정 상 발생할 수 있는 각종 이상데이터의 특성에 대해 분석하고 LSTM 알고리즘을 통해 이를 정제할 수 있는 방안을 제시한다.

II. 본 론

수도 분야에서 발생할 수 있는 이상데이터는 오측과 결측 두 가지로 분류할 수 있다. 결측 데이터는 측정기기 및 통신상의 오류로 인해 데이터가 DB에 저장되지 않고 누락된 상태를 뜻한다. 오측 데이터는 측정기기 등의 오작동으로 참 값이 아닌 값이 DB에 저장되는 현상을 의미한다. 오측과 결측은 서로 상이한 현상으로, 이를 검출하는 방식에도 차이가 있다. 결측의 경우 DB에 비어있는 구간을 탐색함으로써 오측에 비해 상대적으로 검출이 용이하다. 그러나 오측은 각 공정상의 데이터가 다양한 변수들과 상관관계를 가지고 있기 때문에 결측과 같이 단순 탐색 알고리즘으로 한번에 검출하는 것이 불가능하다. 또한 오측의 정의를 "측정기기, 전송계통, 프로그램, 기타 문제로 인하여 명백히 잘못 계측된 것으로 판명된 신뢰할 수 없는 계측값을 측정한 상태"로 정의할 경우 참 값이 명확하지 않은 상태에서 어느 범위까지를 신뢰할 수 없는 값으로 판단할 지도 명확하지 않다. 따라서 이러한 검출 알고리즘을 시스템에 탑재하기 위해서는 각 공정 별 상관관계에 대한 파악이 선행되어야 한다.

측정기기에서 데이터가 측정되고 통신선로 상으로 전송되어 DB에 최종적으로 저장되기 까지 다양한 경로에서 이상데이터가 발생할 수 있다. 결측의 경우 DB 상에 결측되어 있는 구간을 찾는 것으로, Query 명령을 통해 손쉽게 검출할 수 있다. 한국수자원공사 B 정수장의 2015년 7월부터 12월까지 약 6개월 간의 결측 데이터를 추출하여 분석한 결과 평균적으로 1.05%의 결측률을 나타냈다.

표 1. 한국수자원공사 B정수장 결측률

시 간	총 데이터	정상 데이터	결측 데이터	결측율 (%)
'15.07	49,505,760	49,501,324	4,436	0.009
'15.08	51,056,155	51,053,937	2,218	0.004
'15.09	51,594,881	51,495,416	99,465	0.193
'15.10	52,389,666	51,793,729	595,937	1.138
'15.11	50,852,552	49,671,247	1,181,305	2.323
'15.12	54,317,328	52,953,110	1,364,218	2.512
합 계	309,716,342	306,468,763	3,247,573	1.059

3억여 개의 데이터가 취득되는 동안 약 3백만 건의 결측데이터가 발생되었고, 이러한 결측데이터를 자동으로 정제하기 위해서 통계적, 혹은 인공지능 알고리즘을 적용할 수 있다. 이상데이터를 정제하는 방법은 결측과 오측이 동일하다.

오측데이터는 결측데이터와 검출방법이 상이하다. DB 상의 빈 공간을 검색하거나, 실시간으로 계측기기로부터 값이 들어오지 않는 것을 감지하는 것은 어렵지 않다. 그러나 오측데이터의 발생 여부를 판단하기 위해서는 해당 데이터 뿐 아니라 연관되는 상관관계를 도출해야 되는 어려움이 있다. 정수장에서 취득되는 데이터는 크게 유량, 수질, 압력, 수위 네 가지로 구분할 수 있는데 각각의 특성이 크게 다르기 때문에 일괄적으로 오측여부를 판단할 수 없다. 따라서 효율적인 오측 검출을 위해서는 정제 대상인 데이터 각각에 대한 특성 분석과, 연관되는 태그 간의 상관성을 분석하는 것이 중요하다.

오측을 검출하기 위한 방법으로 데이터 간 상관관계를 이용하는 방법이 있다. 오측으로 의심되는 시점에 해당 데이터 뿐 아니라 관계되는 다른 데이터들의 변화도 같이 관찰하여 오측 여부를 판별한다.

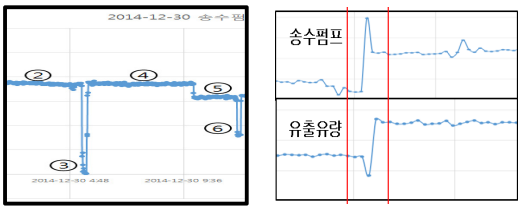

그림 1. 유량데이터 및 연관데이터

위의 그림을 통해 유출유량과 유출압력은 거의 유사한 패턴을 보이며 변화하는 것을 알 수 있다. 또한 유량의 경우 수도관을 따라 이동하기 때문에 몇 대의 펌프를 가동시키는 지에 따라 유량의 양이 결정된다. 따라서 오측이라고 추정되는 데이터가 있을 시, 바로 직전에 펌프 가동대수의 변동여부와 압력의 변동패턴을 동시에 관찰한다면 해당 데이터가 펌프 변화에 따른 정상적 변화인지 잘못 측정된 데이터인지 판단할 수 있다. 이러한 방식은 유량, 수질 및 압력 등 물리적 특성을 가지는 데이터 분석 시에는 정확도가 높지만 수질 등 화학적 특성 데이터 적용 시 정확도가 감소하는 단점이 있다. 특히 물의 탁한 정도를 나타내는 수치인 탁도의 경우 정수장에서 주입하는 특정 약품 외에도 외부 변수가 많아 일괄적인 분석이 힘들다. 따라서 본 연구에서 오측을 검출하기 위해 각 대상 데이터 별 정상 분포범위를 산정하여 이를 초과하는 경우 오측이라고 판단하였다.

오·결측 검출 및 정제 프로그램은 크게 검출모듈과 정제모듈로 구성되어 있다. 검출모듈에서는 입력되는 데이터에 대해 정상범위를 초과하는 경우 오측으로 판단하고 정제모듈로 전달한다.

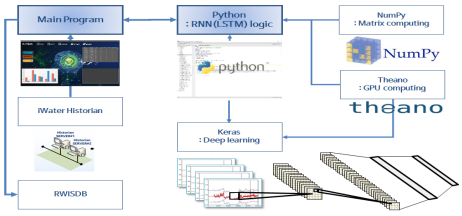

그림 2. 오·결측 검출 및 정제 프로그램 구조

오·결측 정제모듈은 Python 기반으로 구현되었으며, 딥러닝 연산을 위해 Theano 및 Keras 라이브러리를 사용하였다. 계측기기에서 측정되는 데이터는 시간에 흐름에 따른 시계열데이터이며 긴 시계열에 대한 분석이 수행되어야 하기 때문에 RNN 계통의 LSTM(Long Short Term Memory) 알고리즘을 사용하였다. 분석을 위해 B 정수장 253개 태그에 대해 3개월('16.7~9) 분을 학습하였으며 검증을 위해 '16년 12월에 측정된 데이터를 사용하였다.

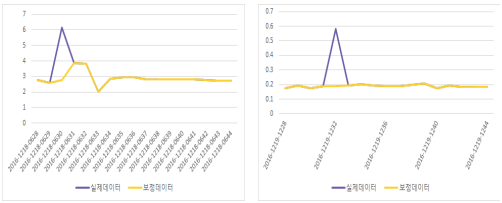

그림 3. LSTM 알고리즘 학습결과 검증

위의 그래프는 '16년 12월 19일 가압장 압력(좌)과 송수터널 후단압력(우)에서 발생한 오측 데이터를 정제한 모습이다. 검출된 오측이 딥러닝 정제 결과 직전 및 직후 값의 패턴과 유사하게 보정된 것을 알 수 있다. 현재 학습에 사용하는 데이터는 3개월이며, 향후 학습기간을 늘릴 경우 정제 정확도가 증가할 것으로 기대된다.

III. 결 론

본 논문에서는 수도데이터에서 발생할 수 있는 이상데이터인 결측과 오측의 특성 및 검출방법에 대한 연구를 진행하였다. 결측 데이터는 DB에서 값이 들어있지 않는 항목을 검색하는 것으로 쉽게 검출할 수 있다. 그러나 오측을 검출하기 위해서는 각각의 데이터 특성에 따라 연관되어 있는 상관태그를 도출하고 오측데이터 발생시점에 상관태그의 변화 추이를 분석하여 실제 오측여부를 판단할 수 있다. 본 연구에서 개발한 오·결측 검출 및 정제 프로그램에서는 오측 데이터 검출을 위해 각 데이터 별 정상범위를 산정하여 해당 범위를 초과할 시 오측으로 판단하였다. 오측으로 판별된 데이터는 딥러닝 기반의 정제모듈로 전달되어 LSTM 알고리즘을 통해 보정되어 DB에 저장된다. LSTM 분석을 위한 학습 시 B 정수장 253개 태그에 대해 3개월분의 데이터를 사용하였다. 향후 학습 데이터 분량을 증가시킬 시 정제모듈의 정확도가 향상될 것으로 예상되나, 동시에 학습 소요시간 역시 크게 증가되기 때문에 학습 소요시간과 알고리즘 정확도 사이의 최적치에 대한 추가분석이 필요하다. 기존에 정상범위 산정을 통해 오측데이터를 검출하는 방식을 활용하였지만 향후 인공신경망의 분류 알고리즘을 통해 오측, 결측, 정상 데이터를 분류하는 검출 신경망과 실제 정제가 이루어지는 정제 신경망으로 프로그램을 분리하여 구성하게 되면 검출모듈에 대한 신뢰성이 더 높아질 것으로 분석된다. 또한 2개의 인공신경망에 대한 충분한 연산속도를 보장하기 위해 GPU를 도입한다면 향후 수도데이터에 대한 이상데이터 관리효율이 획기적으로 높아질 것으로 기대된다.

참고문헌

[1] J.Quevedo, "Validation and Reconstruction of Flow Meter Data in the Barcelona Water Distribution Network,"Control Engineering Practice 18(6), pp. 640-651.

[2] Joon-Hong Seok, "Abnormal Data Refinement and Error Percentage Correction Methods for Effective Short-term Hourly Water Demand Forecasting," International Journal of Control, Automation, and Systems(2014) 12(6), pp.1-12.

[3] http://deeplearning.net/tutorial/lstm.html