데이터 오·결측 저감 정제 알고리즘

이종원[1] · 김호성[2] · 황철현[3] · 강인식[4] · 정회경[1]

[1]배재대학교 · [2]수자원공사· [3]데이터마루(주) · [4]한국영상대학교

Data Cleansing Algorithm for reducing Outlier

Jongwon Lee[1] · Hosung Kim[2] · Chulhyun Hwang[3] · Inshik Kang[4] · Hoekyung Jung[1]

[1]PaiChai University· [2]K-water · [3]DataMalu(co) · [4]Korea University of Media Arts

E-mail : starjwon@naver.com / collar@kwater.or.kr / chhwang@einssnc.com / hue114@hanmail.net /hkjung@pcu.ac.kr

요 약

본 논문에서는 기존 오·결측 데이터 분석 기법인 평균 대체법, 상관계수 수치분석, 그래프 상관성 분석 및 통계 전문가 분석 등 통계적 방법으로 대체 가능성을 조사하여 정수처리 공정에서 계측되는 각종 이상 데이터를 정제하기 위한 방법을 다양한 분석연구로 진행하였다. 또한 물 정보 데이터 오·결측 저감 정제 알고리즘의 신뢰성 및 검증에 있어 분위수 패턴과 딥러닝 기반의 LSTM 알고리즘으로 동작하는 시스템을 모델링하고, Keras, Theano, Tensorflow 등의 오픈 소스 라이브러리로 구현할 수 있는 체계를 연구하였다.

ABSTRACT

This paper shows the possibility to substitute statistical methods such as mean imputation, correlation coefficient analysis, graph correlation analysis for the proposed algorithm, and replace statistician for processing various abnormal data measured in the water treatment process with it. In addition, this study aims to model a data-filtering system based on a recent fractile pattern and a deep learning-based LSTM algorithm in order to improve the reliability and validation of the algorithm, using the open-sourced libraries such as KERAS, THEANO, TENSORFLOW, etc.

키워드

Cleansing Algorithm, CNN, Deep Learning, LSTM

Ⅰ. 서 론

본 논문에서는 데이터의 활용 및 품질관리를 위해 오측과 결측 데이터를 저감하기 위한 정제 알고리즘을 도출하였다. 물 관리 수도사업장은 다양한 물 정보데이터 태그들로 구성되어 있다. 수질, 유량, 압력, 수위 등은 물을 생산하는 정수장의 운영측면에서 매우 중요한 데이터이며 펌프, 밸브 등을 조작하기 위한 수많은 데이터가 유기적으로 운영되고 있다[1,2]. 상대적으로 검출이 용이한 결측 데이터와는 달리 오측데이터는 검출이 어려우며 이는 오측데이터의 특성과 연관이 있다. 실험자가 참값을 인지하고 있는 실험실 환경과는 달리 실제 현장에서 측정되는 데이터의 오측 여부를 판단하는 것은 대단히 어려운 일이다[3,4].

물을 생산하는 정수장의 송수유량을 예로 들면 유량 값이 갑자기 임계치를 벗어났을 때 이것이 유량계의 일시적 오류로 인한 측정값인지 펌프 조작으로 발생한 변화인지 여러 가지 데이터를 기반으로 판단하여야 한다. 따라서 오·결측으로 예상되는 값을 정제하기 위해 데이터 간 관계를 분석하고 통계적 기법과 딥러닝 알고리즘방법을 검토하였다.

이 연구를 통해 실시간으로 연계 및 개방되는 데이터에 대하여 품질을 향상하고, 각종 사고 및 문제 발생 시 정확한 데이터를 제공함으로써 신속한 현황파악 및 의사결정을 지원하고 단순한 데이

터의 품질관리를 신뢰성 있는 데이터 활용으로 변화추이 및 위험징후 파악에 걸리는 시간을 단축하여 선제적으로 대응하는 예방경영 체계를 강화할 수 있도록 물 관리 분야 오·결측 데이터 관리를 위한 데이터 정제 알고리즘을 제안하고자 한다.

II. 시스템 설계

2.1 제안 알고리즘 구성

본 장에서는 제안하는 오측 데이터의 검출 성능을 개선하기 위한 함수적 종속성을 가지지 않고 정밀도가 높은 물 정보 데이터 검출 방법에 대해 설명한다. 기존의 연구 결과 가운데 정밀도가 높고 주변 환경 특성을 가장 잘 반영하는 평균 대체법을 기존 방법으로 인용하였다. 또한 제안 알고리즘으로 단일 값이 시계열로 발생하는 IoT 태그의 데이터 특성을 고려하여 예측의 정확도를 높이는 방법으로 분위수 패턴과 딥러닝 기반의 LSTM을 이용하였다. 제안하는 알고리즘 구성은 그림 1과 같다. 우선 데이터로부터 수집된 데이터를 적재하는 '수집 데이터 저장부'와 검출과 예측을 수행하는 ' 정제 처리부', 정제 결과를 저장하는 '결과 데이터 저장부'이다.

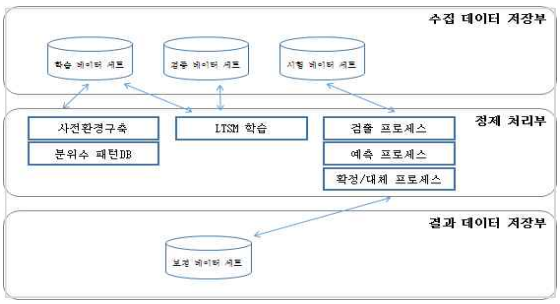

그림 1. 제안 알고리즘 구성

제안한 알고리즘의 처리 과정 흐름도는 그림 2와 같다.

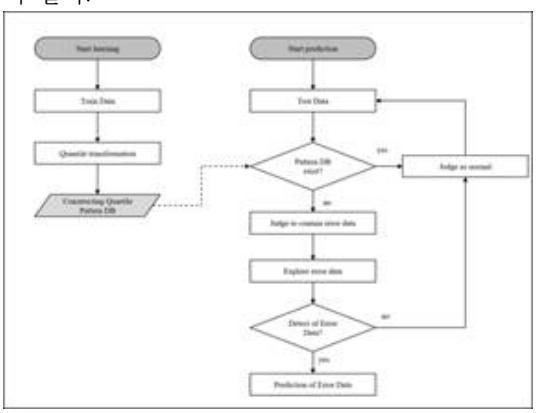

그림 2. 제안한 알고리즘 처리 흐름도

2.2 제안 알고리즘 실험

실험은 Window 환경의 PC급에서 수행되었으며 LSTM 예측을 위한 Python-Keras를 활용하였다. 또한 부족한 메모리 환경을 고려하여 별도의 DBMS를 설치하지 않고 모든 데이터는 file 형태로 기록하고 관리하였다. 실험에 사용된 데이터는 K-water에서 관리하는 현장에서 수집된 IoT 센서에서 추출한 데이터를 활용하였다. 기존 알고리즘과 비교하여 검출률과 오류율은 그림 3과 4와 같다.

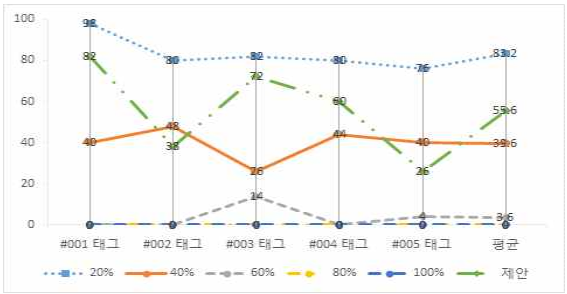

그림 3. 기존 방법과 제안 알고리즘의 검출률 비교

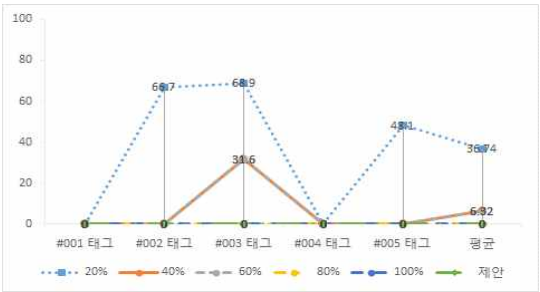

그림 4. 기존 방법과 제안 알고리즘의 오류율 비교

III. 결 론

본 논문에서는 함수적 종속성을 가지지 않고 정밀도가 높은 물 정보 데이터를 처리하기 위해 분위수를 대표값으로 이용하여 데이터를 단순화시키는 분위수 패턴을 이용한 오측데이터 검출방법을 제안하고 기존의 전통적인 방법과 비교하여 오측 데이터의 검출 성능이 개선되었는지 실험을 통해 확인하였다. 검출된 오·결측 데이터에 대한 보정값을 산출하기 위해 IoT 센서의 시계열 데이터 특성에 맞는 LSTM 알고리즘을 제안하고 실험을 통해 예측값의 정확도를 비교하였다. 또한 제안 방법의 실효성을 검증하기 위해 모든 실험은 실제 운영중인 물 정보 데이터를 활용하여 실험하였다. 그리고 실험 결과에 대한 분석을 수행하여 제안 방법이 기존의 전통적인 방법에 비해 오측데이터 검출 및 오·결측 데이터 예측에서 모두 우수하다는 결론을 도출할 수 있었다.

본 논문에서는 사업장(현장)의 오·결측 데이터를 자동으로 검출하고 정제할 수 있는 정제 프로그램을 제안하였다. 분위수 패턴을 이용한 오측 데이터 검출 방법은 기존의 단순 통계적 분석으로는 검출하기 힘들었던 오측 데이터 검출 성능을 향상 시켰고, 딥러닝 기반의 LSTM 알고리즘을 이용해서는 오·결측 데이터에 대한 예측 정확도를 높일 수 있다는 것을 실험을 통해 확인하였다.

References

[1] J. R. Kim, G. W. Shin, S. T. Hong, and G. H. Yoo, "A Study on Analysis of Errors and Data Quality Control Technique for Data Communication," *The Journal of Korean Institute of Communications and Information Sciences*, Vol. 2015, No. 6, pp. 37-38, 2015.

[2] J. R. Kim, G. W. Shin, S. T. Hong, and G. H. Yoo, "A study on detecting algorithm for outlier data in water supply." *The Journal of Korean Institute of Communications and Information Sciences*, Vol. 2016, No .6, pp. 327-328, 2016.

[3] R. R. R. Barbosa and A. Pras, "Intrusion Detection in SCADA Networks," *International Federation for Information Processing*, pp.163-166, 2010.

[4] S. H. Jung, C. S. Shin, and Y. Y. Cho, "A Novel of Data Clustering Architecture for Outlier Detection to Electric Power Data Analysis," *KIPS Transactions on Software and Data Engineering*, Vol. 6, No. 10, pp. 465-472, 2017.