한국컴퓨터정보학회 동계학술대회 논문집 제30권 제1호 (2022. 1)

품질 및 조건 기반 시계열 데이터 선별 활용 방법

문재원[O], 유미선[*], 오승택[*], 금승우[*], 황지수[*], 이지훈[*]
[O]한국전자기술연구원 정보미디어연구센터,
[*]한국전자기술연구원 정보미디어연구센터
e-mail: jwmoon@keti.re.kr

Methods for screening time series data according to data quality and statistical status

JaeWon Moon[O], MiSeon Yu[*], SeungTaek Oh[*], SeungWoo Kum[*], JiSoo Hwang[*], JiHoon Lee[*]
[O]Dept. of Information and Media Center, Korea Electronics Technology Institute,
[*]Dept. of Information and Media Center, Korea Electronics Technology Institute

● 요 약 ●

본 논문에서는 불완전한 시계열 데이터를 활용하기 전 데이터를 선별하여 활용하는 방법을 소개한다. 시계열 데이터의 품질은 수집 네트워크와 수집 기기의 시간적 변화와 같은 가변적 상황에 의존적이므로 불규칙적으로 이상 혹은 누락 데이터가 발생한다. 이때 에러를 포함하였다는 이유로 일괄적으로 데이터를 제거하여 활용하지 않거나, 혹은 누락 데이터의 구간을 조건 없이 복원하여 활용한다면 원하지 않는 결과를 초래할 수 있다. 제안하는 방법은 시계열 데이터의 구간에 대한 누락 데이터의 통계적 정보를 축출하고 이에 기반하여 활용 목적과 활용 가능한 품질의 기준에 부합하지 않는다면 활용 불가능한 데이터라고 판별하고 미리 분석 등의 데이터 활용 시 자동 제외하는 구조를 제안하고 실험하였다. 제안하는 방법은 활용 목적과 상황에 적응적으로 누락 값을 포함하는 데이터의 빠른 활용 판단이 가능하며 보다 나은 분석 결과를 얻을 수 있다.

키워드: 시계열 데이터 (Timeseries data), 누락 데이터 보완 (Missing Data Imputation), 적응적 데이터 처리 (Adaptive data processing)

I. Introduction

시계열 데이터는 시간 스탬프에 따라 기술된 정보 값의 집합으로 이루어진 데이터로, 대부분의 IoT 데이터는 시계열 데이터 형태를 갖는다. IoT 기기 보급 확대로 생성되는 시계열 데이터가 큰 폭의 증가율로 늘어나고 있어 이러한 수집 데이터를 기계학습 및 데이터 마이닝으로 자동 분석 예측하여 활용하고자 하는 수요도 함께 높아지고 있다. 데이터 활용 기술을 이용하여 원하는 결과를 얻기 위해서는 데이터가 무결하다는 전제가 필요하다.

그러나 실제 환경에서는 다양한 이유로 빈번하게 누락 되거나 이상 데이터가 발생하고 있다. 누락 데이터는 숫자, 문자 등으로 데이터를 정의할 수 없거나, 존재하지 않는 데이터를 포괄적으로 일컫는다. 누락 값을 다수 포함하고 있는 데이터를 기반으로 분석이나 학습을 진행한다면 오히려 잘못된 의사 결정을 내리는 결과를 초래한다.

본 논문에서는 누락 값을 포함하고 있는 데이터들을 활용하기 전 활용 가능 여부를 적응적으로 판단하고 선별된 데이터에 대해서만 선택하여 활용하는 방법에 대해 제안한다.

II. 누락 데이터 처리

1. 누락 데이터 표현

누락 값은 "-999"와 같은 극단적인 값을 표기하거나 "NaN", "NA"와 같이 정해진 문자를 표현하는 등의 다양한 방법으로 표현될 수 있다. 그러나 표준화되지 않은 누락 데이터 표기법은 데이터가 기록된 후 정상 데이터와 비정상 데이터를 명확하게 판단하기 어렵다. 그러므로 파이썬 Pandas를 비롯한 데이터를 처리하는 대표적인 라이브러리들은 이러한 누락 데이터를 단순성과 가능상의 이유로 "NaN" 혹은 "NA" 등으로 표기한다. 누락 된 값을 포함하는 데이터를 사용할 경우 대부분의 데이터 분석 및 학습 솔루션 사용시 동작하지 않기 때문에 여러 가지 보간 방법을 이용하여 누락 값을 제거한 후 사용하게 된다.

2. 누락 데이터 제거

각각의 행이 독립적으로 기술되어 있는 테이블 형식의 데이터의 경우 누락 값을 처리하는 방법으로는 누락 값을 포함하는 행을 일괄

삭제하는 방법이 가장 널리 쓰이고 있으며 간단하면서도 명확한 방법이다. 그러나 시간의 흐름에 의존하는 시계열 데이터는 데이터가 기술된 시간 정보 또한 중요하며 연속적 데이터들에 대해 의존적이기 때문에 임의대로 특정 행을 삭제하게 된다면 데이터의 연속성을 보장하기 어렵다. 누락 데이터를 삭제하는 경우 데이터의 양이 줄어들기 때문에 누락 된 데이터 주변을 함께 일괄 삭제할 경우 데이터의 많은 부분이 함께 삭제될 수 있다. 그러므로 이처럼 누락 데이터를 부분적으로 제거하는 방법은 데이터의 양(길이)이 충분할 경우 활용될 수 있는 방법이다.

3. 누락 데이터 보완

누락 데이터를 보완하는 경우 데이터의 양을 줄이지 않는 선에서 해당 데이터의 활용은 가능하게 되므로, 꾸준히 다양한 방법이 연구 및 적용되고 있다. 다변량 데이터 보완을 위한 Generative Adversarial Networks (GAN) 방법[1], XGBoost 등 [2] 여러 기계학습 방법 또한 소개되었다. 다양한 시계열 데이터의 누락 데이터 보완 방법이 활용되고 있지만, 빠르고 손쉬운 누락 데이터 보완을 위해 ImputeTS [3], 싸이킷런 라이브러리[4]와 같은 기본적인 누락 데이터 통계처리 기법 또한 아직 널리 활용되고 있다.

시계열 데이터가 어느 이상의 임계치를 넘어선 양의 누락 데이터를 포함할 경우 보간으로 인해 품질이 낮은 데이터 오히려 오염된 데이터를 생산하게 되어 복구하는 의미가 없을 수 있다. 그러므로 보간 가능한 정도 수준의 누락 값을 갖는 데이터를 부분적으로 선별하여 그 이후 복구하여 활용할 수 있는 방법이 필요하다.

4. 기존 기술의 문제점

시계열 데이터를 학습 및 분석에 활용하기 전 데이터가 누락 값을 포함하는 경우 아예 누락 값을 포함한 시간 구간 동안의 데이터를 일괄 삭제하고 남은 데이터를 활용하거나 혹은 발생하는 구간에 대해서 부분적으로 보간하여 누락 값을 없애는 방법을 사용하고 있다. 데이터를 일괄 삭제할 경우 누락 데이터에 대한 오염을 방지한 완벽한 데이터셋을 얻을 수 있으나 누락 값의 위치에 따라 삭제하는 정도가 커서 활용할 수 있는 데이터 크기가 작아지는 단점이 있다. 반대로 일괄 보간할 경우 근접 데이터나 과거 데이터를 바탕으로 누락 값을 임의로 복구할 경우 데이터를 최대한 보존할 수 있으나 정확한 데이터가 아니므로 무리한 보간 한다면 데이터의 품질이 좋지 않아 분석 및 학습의 결과를 오염시킬 수 있다.

III. 데이터 선별 보완 방법

1. 제안 방법

본 논문에서는 다량의 시계열 데이터에 대해 기준에 따라 데이터 내의 누락 데이터 상태를 판별하고, 기준 이상으로 오염된 데이터를 선택적으로 제거함으로써 기준에 부합하는 데이터만을 활용할 수 있도록 하는 방법을 제안하고 적용하였다. 제안하는 방법은 누락 값을 포함한 데이터에 대해서 누락 값 삭제 파라미터에 따라 복구 가능한 정도를 각 피처별로 판단하며, 복구 등으로 활용 가능한 피처만 선별한다. 그리고 이와 같이 선택된 데이터에 대해서만 적극적으로 데이터를 보완하고 활용하는 방법을 제안한다.

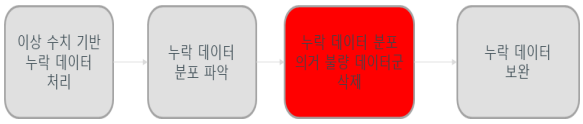

Fig. 1. 누락 데이터 처리 플로우

제안하는 방법은 그림 1과 같다. 먼저 입력된 데이터는 데이터 자체의 이상 수치 정보를 기반하여 최대한 누락 해야 하는 데이터 포인트를 선별하고, 기존 누락 데이터와 함께 이상 수치 기반 누락 데이터를 처리한다. 예를 들어 -50에서 50도까지 측정 가능한 온도 시계열 데이터에 100의 값을 포함하는 경우 이는 오히려 누락시켜야 하는 정보이기 때문에 후처리를 이용해 누락 값으로 대체한다. 이후 처리된 누락 데이터의 전체적인 분포를 파악하고 누락 데이터 분포에 따른 불량 데이터군을 삭제한다. 이 때 마지막으로 보완이 가능할 정도의 데이터만 자동 선별하여 선별된 데이터에 대해서만 누락을 보완한다. 선별하는 기준으로는 연속적 누락 값 및 전체 데이터의 누락 값에 대한 분포를 개수/비율/시간길이의 기준에 의거하여 선별한다. 요약하면, 초기 누락 데이터뿐만 아니라 누락이 되어야 하지만 누락 되지 않은 데이터까지 포함하여 확장된 누락 데이터를 만들고, 누락의 전체 및 연속성에 대한 분포를 파악 후 기준에 의거하여 선별된 데이터의 피쳐들에 대해서만 보완하고 활용한다.

2. 공기질 데이터 클러스터링

본 논문에서 제안하는 방법에 대해서 실제 사용 케이스를 고려하였다. 실제 어플리케이션에서 시계열 데이터를 분석 활용할 때는 고정적인 실험 환경과 달리 입력으로 서로 다른 상황의 가변적인 품질을 갖는 데이터를 활용하게 된다. 예를 들어 서로 다른 실내의 공기질에 대한 이산화 탄소의 일별 분포 패턴에 대해서 N개의 군집으로 클러스터링하기 위한 시나리오를 고려해보자. 클러스터링을 위해 각 장소에서 최근 수집된 데이터를 실제 애플리케이션에서 클러스터링 결과를 활용한다고 가정하면 각 장소의 활용 시간 구간은 계속 달라진다. 또한, 시계열 데이터의 품질도 앞서 기술한 이유로 계속 가변적으로 변한다. 클러스터링과 같은 비지도 학습은 데이터 자체의 구성에 따라 결과를 도출하므로 데이터의 품질에 특히 더 의존적이기 때문에 이를 선별해야 한다.

본 논문에서는 40개의 서로 다른 장소에서 1일 동안 수집된 환경 데이터에 대해 클러스터링해 보았다. 40개의 데이터는 피쳐별로 서로 다른 누락 데이터 분포를 보이고 있으며, 해당하는 방법을 활용한다면 각 피쳐별로 활용 가능한 데이터의 개수가 가변한다. 본 논문의 결과에서는 데이터를 선별하지 않고 데이터의 누락 값을 모두 보완한 데이터를 바탕으로 클러스터링한 결과와 제안하는 방법을 활용하여 데이터를 선별한 후 해당 데이터에 대해서만 적응적으로 누락 데이터를 처리하고 클러스터링한 결과를 비교하였다. 클러스터링 방법으로는 자기조직화 지도 (Self-Organizing Map) 방법을 적용하였고 4개의 군으로 나뉘도록 파라미터를 설정하였다.

3. 결과

각각의 방법을 적용한 클러스터링 결과는 그림 2 및 그림 3과 같다. 실험 결과는 온도 피쳐에 대해서만 기술 하였으며, 다른 피쳐에 대해서는 누락 데이터의 분포가 다르기 때문에 다른 결과가 도출된다. 실내 온도의 변화 패턴에 대한 클러스터링의 결과는 일반적으로 실내 온도가 하루를 단위로 높고 낮아지는 패턴이나 혹은 비슷하게 유지되는 패턴을 주 패턴으로 포함할 것이라고 기대할 수 있다.

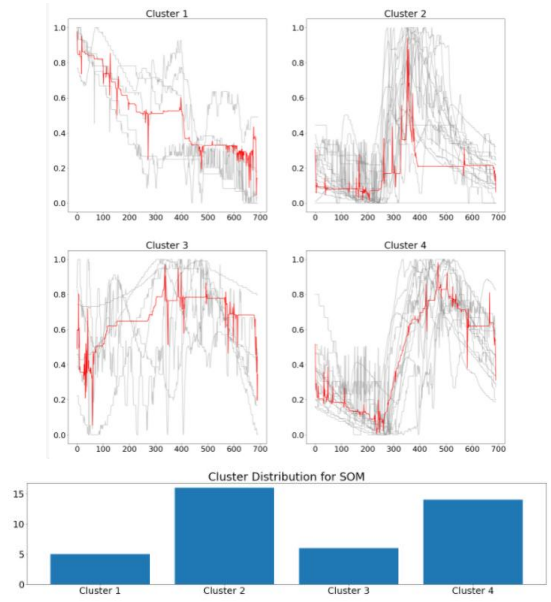

Fig. 2. 전체 데이터 클러스터링 활용

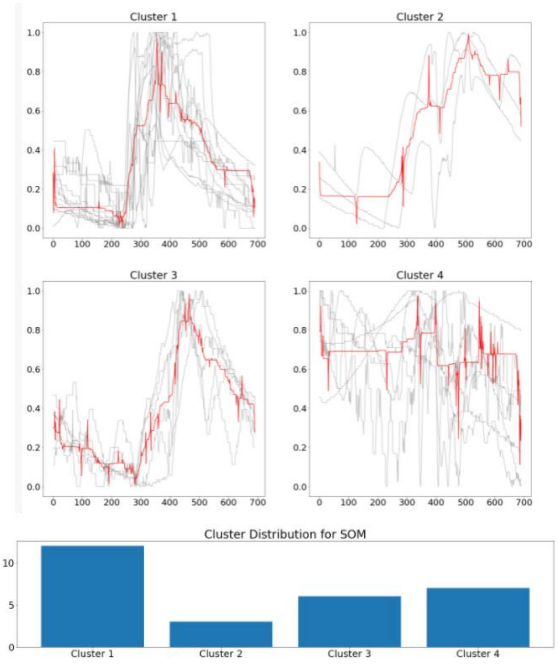

Fig. 3. 제안하는 방법 적용

그림 2와 같이 전체 데이터를 클러스터링에 모두 활용할 경우, 품질이 떨어지는 데이터까지 모두 클러스터링에 활용하므로 클러스터링 결과에 다수의 잡신호를 포함하게 되어 명확한 클러스터링 결과를 제공하기 어렵다. 그러나 그림 3과 같이 기준을 넘어선 누락 데이터를 포함한 데이터를 제거하고 보완 가능한 범위의 데이터에 대해서만 클러스터링할 경우 처음에 가정한 결과와 비슷한 결과를 얻을 수 있음을 확인하였다. 실험을 위해서 연속된 누락 값이 3 point 이상 10 point 이하인 데이터에 대해 자동 선별하였으며 40개의 데이터 중 29개의 데이터만 선별되어 활용되었다.

IV. 결론

대용량 데이터를 기반으로 분석 및 학습 등의 데이터 처리를 수행할 경우 연구자가 개별 데이터의 품질을 일일이 점검하여 선별하기 어렵다. 데이터 품질 적 결함을 보완하기 위해 원 데이터의 상태를 고려하지 않고 누락 데이터를 보완하여 알방적 처리를 한다면 오히려 그릇된 의사 결정을 할 수 있다. 그러므로 데이터의 품질을 적응적이며 일관되게 판별하고 이후 보완 및 활용하는 방법이 필요하다. 그러나 기존 기술들은 누락 데이터의 보완 기술에 대한 연구는 주로 이루어졌지만 품질이 떨어지는 데이터를 선택적으로 제거하는 기술에 대한 논의는 거의 이루어지지 않았다. 본 논문에서는, 개별 데이터의 최대 누락 정도를 파악하기 위해 데이터를 전처리하고, 이후 데이터의 전체 누락 정도와 연속적 누락 개수 등의 품질 기준에 의거 하여 적응적으로 선별되도록 자동화 모듈을 구성하였다. 이러한 자동 선별된 데이터들에 대해서 누락 값 보완 후 활용하는 경우보다 나은 분석 결과를 얻을 수 있는 것을 확인하였다.

ACKNOWLEDGMENT

이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2021-0-00034, 파편화된 데이터의 적극 활용을 위한 시계열 기반 통합 플랫폼 기술 개발)

REFERENCES

[1] Luo, Y., Cai, X., Zhang, Y., Xu, J., & Yuan, X. (2018, December). Multivariate time series imputation with generative adversarial networks. In Proceedings of the 32nd International Conference on Neural Information Processing Systems (pp. 1603-1614).

[2] Zhang, X., Yan, C., Gao, C., Malin, B., & Chen, Y. (2019, June). XGBoost Imputation for Time Series Data.

In 2019 IEEE International Conference on Healthcare Informatics (ICHI) (pp. 1-3). IEEE.

[3] Moritz, S., & Bartz-Beielstein, T. (2017). imputeTS: time series missing value imputation in R. R J., 9(1), 207.

[4] https://scikit-learn.org/stable/modules/impute.html