시계열 데이터 특성 기반 품질 관리 방법 연구

이지훈, 문재원, 황지수
정보미디어연구센터, 한국전자기술연구원
jhlee31@keti.re.kr, jwmoon@keti.re.kr, jshwang34@keti.re.kr

Data Quality Management Method base on Seasonality from Time series Data

Jihoon Lee, Jaewon Moon, Jisoo Hwang
Information & Media Research Center, Korea Electronics Technology Institute

요 약

IoT 기기의 보급 및 확산으로 많은 산업군에서 이를 바탕으로 시계열 데이터를 획득하고 분석하려는 시도가 확대되고 있다. 시간의 흐름에 따라 저장된 데이터들은 주기에 따라 특정 패턴을 갖는 경우가 많으며 이러한 패턴을 파악한다면 주요 산업군의 의사 결정에 도움이 된다. 그러나 IoT 기기의 수집 오류 및 네트워크 환경에 의해 대부분의 시계열 데이터들은 누락 데이터, 이상 데이터를 갖고 있으며 이를 처리하지 않고 분석할 경우 오히려 잘못된 결과를 초래한다. 본 논문에서는 패턴 파악을 위해 '시간, 일, 주, 월, 년' 등 시간의 주기를 기준으로 데이터를 분할하며 이에 기반하여 데이터셋을 재구성하고 활용 가능한 데이터와 불가능한 데이터로 구분한다. 선별된 데이터셋은 클러스터링에 적용하였으며, 제안하는 방법을 적용할 경우 주기를 갖는 시계열 데이터를 활용하는 분석 및 학습에서 더 나은 결과를 보임을 확인하였다.

키워드: 시계열 데이터 품질, 시계열 데이터 분석, 클러스터링 기법

1. 서론

IoT 디바이스의 확산 보급으로 대용량 시계열 데이터가 생산되어 확산 보급됨에 따라 다양한 산업군에서 시계열 데이터에 대한 분석, 예측, 분류 기법을 적용하여 인사이트를 얻으려는 시도가 계속되고 있다. 국내 정부 및 공공기관을 중심으로 공공데이터 포털, 서울 열린 데이터 광장, 카드 빅데이터 플랫폼 등 중요 데이터를 개방하고 다수의 사용자가 여러 목적으로 활용할 수 있도록 하는 시도 또한 계속되고 있다. 또한 스마트 팜, 스마트 팩토리, 스마트 시티 등 여러 도메인에서 다양한 센서들을 활용하여 시계열 데이터를 수집하고 기계학습을 적용하여 생산성을 높이려 한다.

그러나 현존하는 시계열 데이터들은 대부분 그대로 사용하기에는 문제가 있다. [1] 시계열이 종종 비동기 적이거나 불규칙하게 샘플링이 되며, 중간에 시점이 누락된 불완전한 데이터 형태를 지닌 경우가 흔하게 발생한다. 이러한 데이터들은 대략적인 형태를 파악하여 모니터링하는 정도의 활용은 가능하지만 불완전한 탓에 정밀한 분석 및 학습 데이터로 활용하기에는 적절하지 않다. 그러므로 해당 데이터들에 대한 적절한 전처리를 통해 사용 가능한 데이터셋을 구성하는 작업이 필요하다.

그러므로 본 논문에서는 시계열 데이터의 주기를 기반하여 데이터의 품질을 판단하고 기준에 적합한 데이터를 선별하여 활용하는 방법을 제안한다. 제안하는 방법을 검증하기 위해 여러 기준에 기반하여 데이터를 선별하고 클러스터링 기법을 적용해 보았다.

2. 관련 기술

2-1. 시계열 데이터 활용 및 문제

시계열 데이터는 여러 분야에서 활발하게 연구되고 있다. 이상치 탐지는 [2] 시계열 데이터의 센서 및 네트워크 이상으로 생성되는 에러 데이터나 이상 상황으로 발생하는 비정상 데이터 구간을 탐지한다.

또한 시계열 데이터 간 유사성 및 패턴을 찾기 위해 데이터 분류와 클러스터링 기법 [3]도 활용되고 있다. 고차원에서 저차원으로 차원을 축소해 처리 비용을 감소하고 효과적으로 유사한 특징들을 추출하며, 이를 시각적으로 보여줌으로써 데이터에 대한 인지와 통찰력을 확보 및 유사 패턴을 쉽게 파악하는 데 사용한다.

이러한 시계열 데이터 연구들은 데이터가 무결하다는 가정하에 진행된다. 그러나 대부분 실제 상황에서 수집되는 데이터들은 불완전한 경우가 많다. 시간 데이터가 수집되지 못해 중간에 비어 있는 경우도 있지만 센서가 수집하는 범위를 벗어나는 데이터를 포함하고 있어서 값이 있더라도 오히려 데이터 분석에 혼란을 주는 오류 데이터를 다량으로 포함하는 경우도 있다. 이와 같은 데이터를 사용할 경우 의미 있는 결과를 도출하기 어렵고 오히려 잘못된 의사 결정을 할 수 있기 때문에 유의해야 한다.

2-2. 기존 시계열 데이터 품질 관리 방법

이러한 문제를 해결하기 위한 방법으로는 부분적인 유실 데이터를 문제없는 데이터인 것처럼 복원하여 활용하는 방법이 있다 [4]. 그러나 유실 데이터의 양이 많을 경우 무리하여 데이터를 복원하여 활용한다면 오히려 잘못된 결과를 초래할 수 있다. 이를 해결하기 위한 근본적인 해결 방법은 오류 데이터를 포함하는 데이터를 완전히 삭제하고 적합한 데이터 구간만을 활용하는 것이다. 그러나 삭제 시 많은 데이터가 사용하지 못하고 버려지기 때문에 원본 데이터가 부족한 경우 분석 및 활용이 불가할 수 있어 삭제에 대한 기준 제시가 필요하다.

2-3. 시계열 데이터의 주기성

시계열 데이터는 연속된 특징을 보이고 있으며, 이 연속된 데이터들은 시간의 흐름에 따라 반복되거나, 공통적인 패턴을 보일 수 있다. 시계열 데이터는 주기성을 갖는 경우가 많다. 그 주기들은 보통 '시간, 일, 주, 월, 년' 등의 단위를 기준으로 공통되고 반복되는 패턴을 보인다.

예를 들어 실외 온도의 경우 공전과 자전의 영향을 받기 때문에 하루 및 년 단위의 주기성을 동시에 갖는다. 또한 학교 실내의 이산화탄소 변화의 경우 일과로 인한 하루 및 일주일 단위 패턴을 가질 확률이 높으며, 외부 온도에 따라 실내 창문 개방 패턴이 달라지기 때문에 년 단위의 주기성을 가질 수도 있다. 이러한 패턴은 데이터의 분석 및 정제에 중요한 역할을 하며 데이터의 활용 시 반드시 고려해야 한다.

3. 품질 및 주기성 기반 데이터 선택 방법

3-1. 개요

본 논문에서는 시계열 데이터의 품질을 진단하고 주기성을 기반으로 활용 가능한 데이터를 선택하는 방법을 고안하였다. 누락 데이터가 없는 완전한 시계열 데이터라면 그대로 활용하더라도 문제없지만 대부분의 시계열 데이터는 불완전하여 전처리 없이 활용이 어렵다. 현재 사용하는 데이터가 불완전한 시계열 데이터라는 전제하에 그림 1 과 같이 프로세스를 진행한다.

첫 번째는 기존에 흔히 사용하는 방법으로, 데이터를 소비하기 전체 누락 데이터를 보완하고 활용한다. 이는 입력으로 활용하는 전체 데이터에 대해서 일괄적인 처리를 하여 프로세스 자체는 간단하지만, 장기간의 누락 데이터를 포함하여 활용이 어려울 정도의 저품질의 데이터도 일괄 처리한다는 단점이 있다.

두 번째 방법은 데이터의 주기성을 고려하여 분할한 후 누락 데이터를 처리하고 활용한다. 누락 데이터 처리 시 보통 주변 혹은 과거의 데이터 통계에 기반하여 값을 생성한다. 두 번째 방법의 경우에는 시계열 데이터의 전체 구간이 아닌 독립된 주기 동안의 값을 기반하여 누락 데이터를 복구하기 때문에 시계열 데이터의 구간별 독립성을 확보하고 데이터 처리 시간을 줄일 수 있다.

세 번째 방법은 데이터를 주기별로 분할한 후 분할한 여러 데이터셋 중에 적절한 데이터를 한 번 더 선별하는 과정을 거치는 것이다. 선별된 데이터에 대해서만 누락 값에 대한 보완을 진행하고 해당 데이터를 활용하게 된다. 세 번째

방법의 장점은 데이터의 절대량이 많을 때 불필요한 데이터를 걸러내어 비교적 품질이 좋은 데이터만을 선별함으로써 정확도를 높일 수 있다. 또한 데이터 선별 시 주기를 단위로 선택 혹은 제거하기 때문에 데이터 분석 및 학습 활용 시 문제가 되지 않는다.

본 논문에서는 세 번째 방법을 제안하고 그 결과를 클러스터링에서 활용하였다. 이를 위해서는 데이터를 선별하기 위한 누락 허용 정도 파라미터 값 (nanLimitInfo) 두가지를 정의하였다. 첫번째로는 독립된 주기의 데이터 안에서 허용할 수 있는 연속 누락 데이터의 최대 개수이며, 두번째 값은 전체 누락 데이터의 최대 개수다.

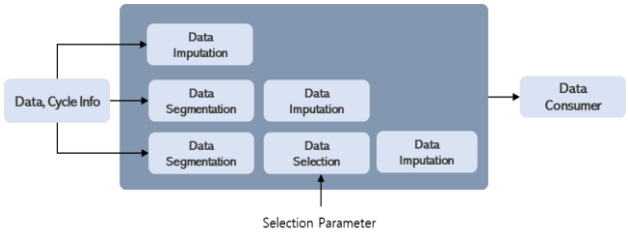

그림 1. Data Cycle & Imputation process

3-2. 주기 기반 데이터 분할

데이터가 가질 수 있는 기본 주기는 '시간, 일, 주, 월, 년' 단위로 설정하였다. 기준 주기가 설정되면 전체 데이터는 기준 주기에 의해 분할된다. 데이터들은 복합적인 주기를 가질 수 있기 때문에 반복하여 데이터 분할 선별 보완 가정을 거쳐 데이터를 준비할 수도 있다.

데이터를 분할하기 전 활용할 데이터가 균일하고 완전한 시간 인덱스를 지닌 상태인지 확인하는 절차가 필요하다. 대부분의 시계열 데이터들은 연속적으로 저장되고 있다고 해도 중간에 오류나 누락으로 인하여 시간대가 비어 있는 경우가 빈번히 존재한다. 그러므로 원본 데이터의 기술 주기가 일정 시간 간격이 될 수 있도록 정제한다. 시계열 데이터의 시간 인덱스가 완전해야 주기별로 데이터를 안정적으로 나누며 품질에 따른 데이터 선택이 가능하다. 예를 들어, 주기를 하루로 설정하여 1 시간 단위의 시간 인덱스 단위로 기술되는 데이터가 20 시 다음에 22 시가 나온다면, 21 시가 누락된 상태이며 이러한 누락 데이터를 처리하여 완전한 시간 인덱스를 생성해야 한다. 시간 인덱스가 완전하다는 조건하에 데이터를 분할하며 해당 방법으로 주기별로 데이터를 분할했을 시, 주기별로 분할된 데이터들의 개수는 동일하게 된다.

3-3. 데이터 선별 및 보완

주기별로 분류된 데이터들에 이상치, 누락 데이터가 다수 존재할 경우, 사용되는 데이터들의 품질이 낮을 수 있으므로 데이터를 일괄 제외하는 것이 좋다. 하지만 소량의 누락 데이터가 존재하는데도 제외한다면 사용할 수 있는 데이터가 현저히 줄어들 수 있다. 이러한 상황을 대비하기 위해, 허용할 정도의 누락 데이터 정보를 기술한 nanLimitInfo 에 의거하여 데이터의 품질 선별 정도를 조절함으로써 무분별하게 데이터를 제외하는 경우를 줄인다. 이를 통해 활용할 수 있는 데이터와 활용하지 못하는 데이터로 다시 분류한다. 본 논문에서는 독립 데이터에 대한 연속 누락데이터와 전체 누락데이터의 개수와 발생 확률을 그 기준으로 한다.

활용할 수 있는 데이터로 선별된 데이터들은 누락 데이터 값을 대치할 수 있는 알고리즘을 적용하여 데이터를 보완한다. 누락 데이터를 어느 정도까지 보완할 수 있는지에 대해서도 활용 어플리케이션에 따라 다르게 설정할 수 있다.

4. 실험 결과

본 논문에서 제안하는 방법을 검증하기 위해 특정 어린이집의 실내 공간에서 2021 년 3 월 말부터 8 월 말까지 152 일 동안 수집된 소음 데이터를 활용하였다. 그림 2 는 전체 데이터를 분할하기 위한 조건을 설정하는 사용자 인터페이스로 전체 데이터를 하루 주기로 분할하여 60 분 간격으로 리샘플링 하였다. 그리고 해당 인풋 데이터에 대해서 연속 및 전체 누락 데이터에 대한 처리 개수를 변화하면서 최종적으로 사용할 데이터를 선택하는 과정을 거쳤다. 그림 3 은 특정 처리 파라미터에 의해 나온 결과 (붉은 그래프: 활용 가능 데이터, 회색 그래프: 활용 불가능 데이터, 노란색 빈 그래프: 데이터가 존재하지 않는 경우)를 나타낸다.

어린이집의 특성상 오전 8~9 시경 소음이 커지기 시작하여 5~8 시경부터는 소음이 작아질 것이라고 가설을 세울 수 있다. 이를 바탕으로 첫 번째 (Type 1), 누락데이터와 상관없이 모든 데이터를 활용하여 하루 단위의 데이터로 분할하고 해당 데이터들을 사용하여 클러스터링을 진행하였다. 그리고 두 번째 (Type 2)로, 하루 단위의 데이터에 대해서 연속하는 누락 데이터는 1 이하, 전체 누락 데이터 개수가 3 이하인 데이터만을 선별하였다. 본 실험에서는 누락 데이터를 선형 보간법을 활용하여 보완하였으며, 선별된 데이터에 대해 입력 데이터를 9 개의 패턴으로 클러스터링을 진행하였다. 클러스터링 기법으로는 비지도 학습으로 경쟁 학습을 통해 연결 강도를 조정하여 유사 데이터를 군집하고 차원을 축소함으로써 2 차원상에서 시각적으로 쉽게 나타낼 수 있는 SOM(자기조직화 지도)을 활용하였다.

그 결과 그림 4 와 같이 서로 다른 클러스터링 결과를 생성함을

확인할 수 있었다. 모든 데이터를 허용하는 <Type 1>의 클러스터링 결과의 경우, cluster 3, 7 과 같이 밤~새벽 시간대에도 소음이 높은 이상한 현상을 보이고 있는 결과가 주요 패턴으로 확인되었으며 이는 잘못된 유실 값 보완으로 인해 일어난 현상이라고 추측된다. 이와 다르게 누락 데이터를 고려하여 선별한 데이터를 클러스터링 한 <Type 2>의 경우, 우세 패턴이 확연히 구분되며, 다수의 데이터로 구분된 cluster 3, 4, 8 은 가설과 동일하게 오전 10 시부터 오후 8 시 노이즈 점점 처졌다 낮아지는 패턴을 보이고 있음을 확인하였다. 이처럼 누락 데이터를 처리하지 않고 사용하면 불필요한 데이터로 인하여 불완전하고 정확하지 않은 결과가 나오게 되므로 적절한 데이터 선별을 통해 올바른 결과를 도출해야 한다.

그림 2. 실험 파라미터 조건

그림 3. 전체 데이터 주기별 분할

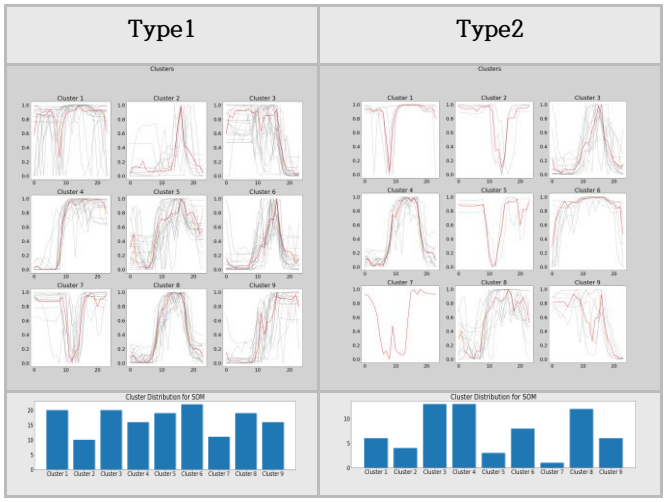

그림 4. Typ1, Type2 의 Clustering 패턴 및 개수 비교

5. 결론

시계열 데이터는 시간의 흐름에 따라 데이터가 지속적으로 축적되기 때문에 용량이 크며 이에 따른 다수의 이상 데이터를 포함하고 있다. 본 논문에서는 시계열 데이터 활용 전 저품질의 데이터는 활용에서 제외하기 위해 시간 주기 단위로 품질을 파악하여 선별하고 선택된 데이터만 활용하는 방법을 제안하였다. 이를 검증하기 위해 특정 공간의 실내 소음 데이터를 기반으로 클러스터링에 적용하였고, 이상 데이터를 제거한 데이터가 더 나은 결과를 보이는 것을 확인하였다. 해당 기술은 데이터 분석 전 학습 데이터를 준비하는 과정의 전처리에 활용되어 안정적인 데이터를 생성하는 데 도움이 될 것이며, 보다 정확한 분석 결과를 도출하거나 고품질의 학습 데이터를 준비하는 용도로 활용할 수 있을 것이다.

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] Tan Zhi-Xuan,1,2 Harold Soh,3 Desmond C. Ong1,4(2020). Factorized Inference in Deep Markov Models for Incomplete Multimodal Time Series

[2] Blázquez-García, A., Conde, A., Mori, U., & Lozano, J. A. (2021). A review on outlier/anomaly detection in time series data. ACM Computing Surveys (CSUR), 54(3), 1-33.

[3] MOHAMMED ALI 1,2, ALI ALQAHTANI 1,2, MARK W. JONES 1, AND XIANGHUA XIE 1 (2019). Clustering and Classification for Time Series Data in Visual Analytics: A Survey

[4] Cao, W., Wang, D., Li, J., Zhou, H., Li, L., & Li, Y. (2018). Brits: Bidirectional recurrent imputation for time series. Advances in neural information processing systems, 31.