서로 다른 특성의 다수 시계열 데이터 정합 방법

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

Integration method of heterogeneous time series data with different characteristic

Jisoo Hwang, Jaewon Moon, Jihoon Lee
Korea Electronics Technology Institute

요 약

다양한 산업 분야에서 생성되는 시계열 데이터는 그 특성상 데이터의 기술 방법 범위의 양과 질이 서로 다르며 이로 인해 서로 통합하여 활용하기가 쉽지 않다. 본 논문에서는 서로 다른 수집 주기와 길이를 갖는 시계열 데이터 간의 통합 방법을 제안한다. 여러 이질적 데이터를 함께 사용하기 위해 고려해야 할 시계열 데이터의 특성과 연관 기술을 소개하고 두 가지 시계열 데이터 통합 방법 및 필요한 파라미터를 제안한다. 제안하는 방법은 시계열 본연의 특성을 고려하여 데이터를 같은 차원으로 변환하거나 활용 목적을 고려하여 다른 차원을 변환하는 방법으로 이를 통해 통합하려는 데이터의 불균등 주기 문제를 극복할 수 있다.

1. 서론

4 차 산업의 발전으로 인한 보안, 의료, 경영, 과학 등의 성장은 다양한 데이터의 공급과 수요를 증가시켰다. 데이터 처리, 저장, 분석 관련 국내 동향을 살펴보면 과거에는 영상, 음성과 같은 정형화된 데이터를 중심으로 한 연구가 활발하게 진행되어 왔다. 하지만 현재 데이터 시장에서는 IoT 센서의 보급으로 인한 시계열 데이터의 공급량이 증가함에 따라 시계열 분석에 대한 요구가 확대되고 있다. 그러나 다양한 데이터의 분석 및 응용 연구를 수행하는 국외와 달리 국내는 주류 데이터 연구에 비해 시간을 기반한 복합 데이터 연구는 아직 미흡한 상태이다 [1].

다양한 산업 발전을 위해서는 서로 다른 도메인의 데이터를 통합하여 응용하는 기술이 필요하다 [2]. 예를 들어서 스마트 팜에서 환경에 따른 작물의 성장 영향력을 예측하고자 할 때는 작물이 자라는 환경의 온도, 습도, 조도와 같은 시계열 센서 데이터 정보만 아니라 사용자의 입력 혹은 영상으로 취득한 작물의 생장 정도의 데이터가 함께 필요하다.

서로 다른 도메인의 통합 데이터를 활용한 기계학습 연구가 활발히 진행되기 위해서는 앞서 데이터의 통합이 선행되어야 한다. 일회성의 연구를 위해서는 연구자가 상황에 맞게 각 데이터를 변환하고 통합하여 활용할 수 있지만, 여러 산업에서 생성되는 불특정 데이터를 실시간으로 통합 활용하려 할 경우, 그 방법에 대한 표준 및 가이드라인이 존재하지 않아 실제 활용은 어려운 상황이다.

특히 시간을 기준으로 기술된 시계열 데이터를 분석 및 학습에 활용하기 위해서는 데이터가 순차적이면서 균일한 간격으로 저장되어 있어야 한다. 그러나 다양한 산업군에서 수집된 시계열 데이터는 수집기의 에러 및 네트워크 환경 문제로 인한 누락 데이터가 빈번하게 발생한다. 결측 값이 정리되지 않은 상태의 데이터는 기계학습에서 원하는 결과를 얻지 못하기 때문에 필수적으로 분석 전 결측 값 처리가 필요하다. 더불어,

시간 정보에 의존적인 데이터이므로 기술된 시간의 빈도, 기술 시간 범위 등에 의거하여 전처리를 진행하고 시간 정보를 고려하는 기계학습 알고리즘을 선택해야 한다.

따라서 본 논문에서는 서로 다른 이종 시계열 데이터를 통합하는 두가지 방법을 제안한다. 데이터를 구성하는 각 피처들의 개별적 특성을 바탕으로 데이터를 변환하고 통합하는 방법과 심층신경망에 기반하여 데이터의 차원을 변환하는 방법으로, 제안하는 방법에 의거하여 서로 다른 데이터 기술 주기로 저장된 다수 데이터를 결합할 수 있다.

2. 본론

2.1 통합을 위한 시계열 데이터의 특성

시계열이란 시간 정보를 갖고 이를 기준으로 시간 순서에 맞춰 정보가 저장된 데이터를 의미한다. 시계열 데이터는 순서에 따른 정보로 시간의 영향을 받기에 시간 정보를 배제하고 분석할 시 전혀 다른 결과를 얻는다. 즉, 다른 정형 데이터와 다르게 시간의 순서가 중요하다. 따라서 시계열의 큰 특징인 시간 빈도, 데이터가 기술된 시간의 범위 또한 중요하므로, 다수 시계열 데이터 통합 시 시간 빈도와 범위를 모두 고려하여 데이터를 정리해야 한다. [3]

시간 빈도는 시간 정보가 저장되는 주기이다. 시계열은 수집하는 목적과 방식에 따라 데이터의 기술 빈도가 달라진다. 예를 들어 같은 공간에서 온도는 3 분에 한 번씩 제공이 되고 습도는 5 분에 한 번씩 제공이 되는 데이터가 있다고 가정해보자. 그렇다면 온도 데이터의 시간 빈도는 3 분이고 습도 데이터는 5 분이다. 이를 일괄 통합 시 데이터 시간 정보의 포인트는 [3, 5, 6, 9, 10 …] 분으로 기술된다. 이런 시간의 불 균일로 인해 시간 차례대로 데이터를 처리하는 기계학습 방법 활용 시 성능 저하를 초래한다. 즉, 두 개 이상의 데이터를 통합할 시 균일할 시간 빈도로 정리해야 한다. [4]

시간 범위는 시계열 데이터의 저장된 시작 시간과 끝나는 시간의 범위를 뜻한다. 앞서 예를 들었던 온도 데이터는 2022.01.01. ~2022.02.25. 일까지 저장되어 있고 습도 데이터는 2022.01.10. ~2022.03.06. 의 범위로 이루어졌다고 가정해보자. 이때 두 데이터를 통합하면 데이터는 2022.01.01. ~2022.03.06. 의 범위를 가질 것이며 각 데이터가 겹치지 않는 시간 범위에서는 정보가 존재하지 않기 때문에 결함이 생긴다. 분석 성능을 높이고 다양한 응용 어플리케이션에서 활용하기 위해서는 이와 같은 결함을 반드시 해결한 후 진행해야 한다. 일반적으로 서로 다른 데이터의 통합 시간 범위는 전체 시간 범위를 고려하거나 공통 시간 범위 중 하나를 선택한다. 전체 시간 범위를 고려할 경우 통합에 의한 결측 값의 보완이 필요하며, 공통 범위만 통합할 경우 공통 부분이 아니기 때문에 버려지는 데이터의 유실을 생각해야 한다. 본 논문에서는 데이터의 공통 범위를 고려하는 교집합 범위를 기준 삼아 통합 시간 범위로 설정하여 통합하였다.

2.2 시계열 데이터 통합 방법

이번 챕터에서는 제안하는 방법으로 데이터를 통합하는 과정과 그에 따른 필요한 정보를 설명한다.

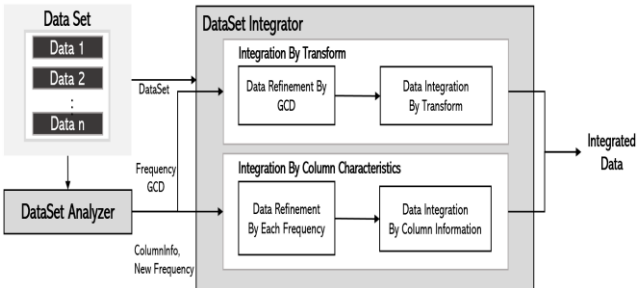

Figure 1.Data Integration Process

2.2.1 Data Integration Process

데이터의 통합을 위해서 다음과 같은 두가지 방법을 소개한다. Fig. 1은 서로 다른 주기를 갖는 N개의 데이터를 통합하는 구조를 나타낸 그림이다. 첫번째 방법은 데이터셋의 변환(Integration By Transform)을 통해 통합하며, 두번째 방법은 컬럼 별 데이터 특성에 따라 통합 (Integration by column characteristics) 한다.

DataSet Analyzer는 데이터를 통합하기 위한 필요 정보를 생산하는 모듈로, 주요 전달 정보로는 새롭게 기술될 통합 데이터의 빈도 정보 (GCD: 최대 공약수 기준 빈도, New Frequency: 신규 기준 빈도) 와 각 개별 데이터들의 컬럼 별 특성 정보 (Column Info) 가 있으며 자세한 설명은 2.2.2에 기술하였다. 통합 방법에 따라 N개의 Data와 각 통합 방법에 필요한 파라미터 정보를 바탕으로 통합이 진행된다.

2.2.2 컬럼 별 데이터 특성에 따른 통합 방법

해당 방법은 새로운 기술 주기를 중심으로 개별 데이터의 각 컬럼을 업 혹은 다운 샘플링하여 변환하고 한번에 통합하는 방법이다. 각 데이터의 컬럼들에 대해 데이터 타입, 데이터 포인트간 연속성을 기반해 데이터의 업샘플링과 다운샘플링 방법을 결정한다. 업샘플링은 데이터의 주기 수를 기존 주기 보다 늘려 더 자주 데이터가 발생하도록 만드는 방법이며 반대로 다운샘플링은 데이터의 주기 수를 줄이는 방법이다. 업샘플링의 경우 데이터를 늘리는 작업이기 때문에 필연적으로 누락 데이터가 발생하며 이에 대한 결측치 치환 방법이 함께 필요하다. Fig.2는 기존 기술 시간 빈도가 2시간인 데이터를 통합을 위해 1시간 주기로 업샘플링하거나 4시간 주기로 다운샘플링할 때의 결과를 나타낸 예시 그림이다.

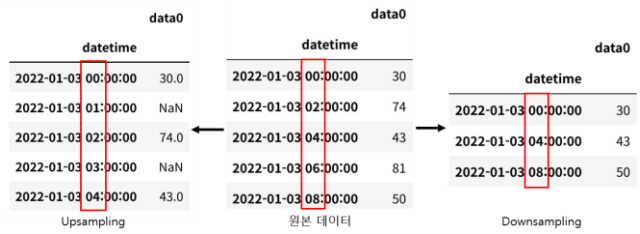

Figure 2. Upsampling & Downsampling

새로운 기술주기에 따라 원본 데이터는 업/다운 샘플링을 거치며 이에 따라 발생되는 결측치의 처리 방법은 컬럼의 특성을 파악하여 적합한 방법을 제안하거나 사용자의 결정 하에 진행되도록 설정하였다. Fig.3은 컬럼 별 데이터 특성에 따른 통합 방법으로 통합 데이터를 생성한 예시 그림이다. 2분씩 주기를 갖는 Data0 과 1시간씩 주기를 갖는 Data1, Data2를 공통 시간 범위를 기준으로 각 컬럼 별 특성에 따라 샘플링 후 통합한 결과이다.

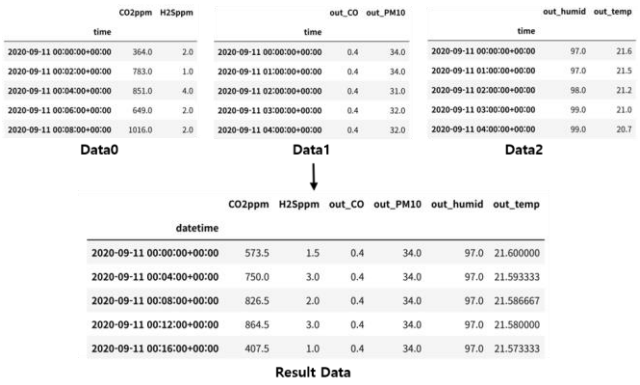

Figure 3. Data Integration By Column Characteristics

통합하려는 N개의 데이터들의 모든 컬럼에 대한 정보를 갖는 Column Info는 컬럼 이름, 데이터 기술 빈도, 데이터 타입, 각 데이터포인트 간 관계가 기술되며 이를 바탕으로 업/다운 샘플링 방법이 결정된다. 업/다운 샘플링의 기준이 되는 새로운 기술 주기는 통합하려는 데이터들의 기존 기술 주기들을 기반으로 최다 빈도 주기, 최소 주기, 최대 주기, 평균 주기 값 등의 여러가지 선택 방법 중 결정될 수 있다.

2.2.3 변환에 의한 통합 방법

두 번째 방법은 서로 다른 주기의 데이터를 아예 다른 차원의 데이터로 변형함으로써 데이터 기술 주기의 불일치를 극복하는 방법으로 Integration By Transform이라고 명명하였다. 해당 방법을 위해 여러 기법이 활용될 수 있지만, 본 논문에서 LSTM AutoEncoder 를 활용한 변형 통합을 설명하며 이와 유사한 다른 변환 모델을 사용할 수 있다.

LSTM AutoEncoder 기법은 입력 데이터의 특징을 학습한다는 특성을 활용한다. 그러나 LSTM AutoEncoder의 입력으로는 불규칙한 기술 주기의 데이터를 활용할 수 없다. 그러므로 통합을 원하는 N개의 데이터들은 강제적으로 규칙적인 주기에 맞춰 기술되는 준비과정이 필요하다. 서로 다른 주기의 데이터에 대해 유실값을 포함하더라도 규칙적인 주기로 기술되는 데이터 정제 (Data Refinement) 과정을 거치며, 이를 위한 새로운 기술 주기는 각 데이터의 기술 주기에 대한 최대공약수(GCD)로 강제 설정된다. 그 이유는 최대공약수 주기는 데이터의 유실을 최소화 하면서 기술 주기를 맞출 수 있는 값이기 때문이다. 각 데이터를 최대공약수 기술 주기에 의거하여 시간 빈도 조정 후 결합을 진행하면 기준 주기에 해당 정보가 존재하지 않는 부분은 결함이 생긴다. 이후 LSTM AutoEncoder를 활용하여 입력 데이터를 기반해 새로운 차원으로 변환된 데이터를 생성하여 결측 값을 처리하며 새롭게 변환한 통합 데이터를 출력한다.

LSTM AutoEncoder 는 Encoder-Decoder LSTM 구조로 정보에 순서가 담긴 데이터를 AutoEncoder 가 다룰 수 있게 하는 알고리즘이다. [5][6] LSTM 은 순서가 있는 데이터 즉, 시퀀스 데이터를 다룰 수 있는 알고리즘 vanilla RNN 의 기울기 소실, 기울기 폭주인 문제점을 보완하기 위해 만들어졌다. 은닉층에서 발생한 모든 결과를 출력층으로 전달하는 Feed Forward Neural Network 와 달리 RNN 은 은닉층에서 다음 스텝 은닉층으로 결과를 전달하는 순환 신경망이며 RNN 의 단점을 극복하고자 RNN 보다 고도화된 구조를 갖는 알고리즘이 LSTM 이다. AutoEncoder 는 입력한 데이터에서 중요한 정보를 추출하여 이 정보를 바탕으로 입력 데이터와 유사한 데이터를 재구성해 데이터를 효율적으로 복원하는 알고리즘이다. 데이터를 입력하면 차원을 축소해 불필요한 정보는 정리하고 필요한 정보만을 추출해 벡터로 변환하는 이 부분을 Encoder 라 명한다. 생성된 벡터에 의거하여 다시 데이터를 복원해 출력하는 부분을 생성 네트워크라 하며 이를 Decoder 라 한다.

본 논문에서의 LSTM AutoEncoder 사용 목적은 입력 데이터들의 주기를 균등하게 조정함으로써 발생한 결함을 변환을 통해 해결하기 위함이다. 먼저 최대공약수의 기술 주기로 인해 발생한 결합 데이터의 결함 부분은 0 으로 대체 후 LSTM

AutoEncoder 를 진행해 변환한다. 입력 데이터를 복원하기 위해 입력 데이터의 주요 특징을 학습하는 LSTM AutoEncoder 의 특성을 활용하여 결합한 데이터의 특징을 학습한 모델을 생성하고 이를 기반으로 최적의 특징을 추출한 latent vector 를 새로운 변수로 정의하여 생성한 다른 차원의 통합 데이터를 출력한다. Fig.4 는 주기가 10 분, 7 분, 3 분인 입력 데이터의 주기를 최대공약수 1 분을 기준으로 정제 및 결합 후 LSTM AutoEncoder 를 활용해 데이터를 변환한 결과이다.

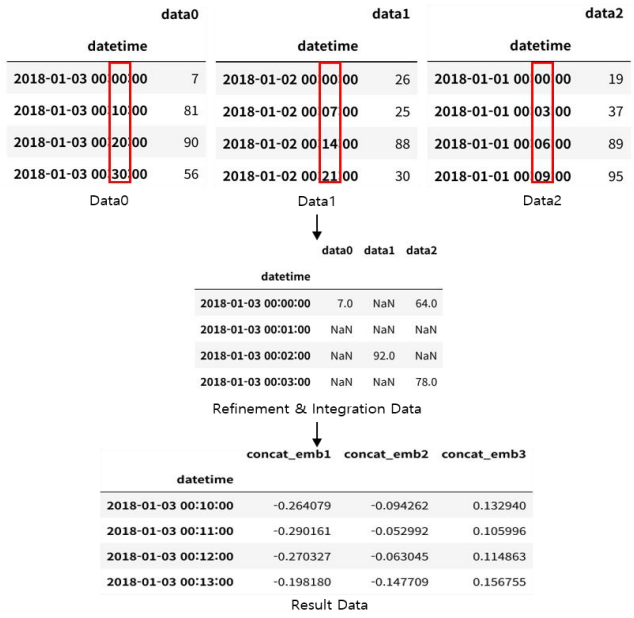

Figure 4. Data Integration By Transform

3. 결론

본 논문에서는 서로 다른 형태의 시계열 데이터 통합 방법에 대해 제안했다. 통합 방법을 진행하기에 앞서 통합에 필요한 정보 파라미터를 두 가지로 정의했다. 데이터 본연의 특성을 분석하여 정의한 정보는 기본적인 시계열 데이터 특성에 따른 통합 방향을 제시할 수 있다. 더불어, 결합한 데이터를 기계학습에 활용하기 전 데이터의 기술된 시간 빈도를 균일하게 정리해야 하는 필요성과 차원 변환을 통한 통합 데이터 생성 방법을 제안함으로써 다양한 시계열 데이터 통합 연구 길을 확대했다.

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] 동국대학교. 2018). 이종 빅데이터 통합 분석 메티러닝 기술개발. https://scienceon.kisti.re.kr/srch/selectPORSrchReport.do?cn=TRKO202000006791.

[2] Ahn, H., Chae, H., Jung, W., & Kim, S. (2017, February). Integration of heterogeneous time series gene expression data by clustering on time dimension. In 2017 IEEE International Conference on Big Data and Smart Computing (BigComp) (pp. 332-335). IEEE.

[3] 김에덴, 고석갑, 손승철, & 이병탁. (2021). 시계열 데이터 결측치 처리 기술 동향.

[4] Kreindler, D. M., & Lumsden, C. J. (2016). The effects of the irregular sample and missing data in time series analysis. In Nonlinear Dynamical Systems Analysis for the Behavioral Sciences Using Real Data (pp. 149-172). CRC Press.

[5] Jin, H. Y., Jung, E. S., & Lee, D. (2020). High-performance IoT streaming data prediction system using Spark: a case study of air pollution. Neural Computing and Applications, 32(17), 13147-13154.

[6] Sagheer, A., & Kotb, M. (2019). Unsupervised pre-training of a deep LSTM-based stacked autoencoder for multivariate time series forecasting problems. Scientific reports, 9(1), 1-16.