특집논문 (Special Paper)
방송공학회논문지 제27권 제6호, 2022년 11월 (JBE Vol.27, No.6, November 2022)
https://doi.org/10.5909/JBE.2022.27.6.872
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

서로 다른 특성의 시계열 데이터 통합 프레임워크 제안 및 활용

황 지 수[a)], 문 재 원[a)‡]

Introduction and Utilization of Time Series Data Integration Framework with Different Characteristics

Jisoo Hwang[a)] and Jaewon Moon[a)‡]

요 약

IoT 산업 발전으로 다양한 산업군에서 서로 다른 형태의 시계열 데이터를 생성하고 있으며 이를 다시 통합하여 재생산 및 활용하는 연구로 진화하고 있다. 더불어, 실제 산업에서 데이터 처리 속도 및 활용 시스템의 이슈 등으로 인해 시계열 데이터 활용 시 데이터의 크기를 압축하여 통합 활용하는 경향이 증가하고 있다. 그러나 시계열 데이터의 통합 가이드라인이 명확하지 않고 데이터 기술 시간 간격, 시간 구간 등 각각의 특성이 달라 일괄 통합하여 활용하기 어렵다. 본 논문에서는 통합 기준 설정 방법과 시계열 데이터의 통합 시 발생하는 문제점을 기반으로 두 가지의 통합 방법을 제시하였다. 이를 기반으로 시계열 데이터의 특성을 고려한 이질적 시계열 데이터 통합 프레임워크를 구성하였으며 압축된 서로 다른 이질적 시계열 데이터의 통합과 다양한 기계 학습에 활용할 수 있음을 확인하였다.

Abstract

With the development of the IoT industry, different types of time series data are being generated in various industries, and it is evolving into research that reproduces and utilizes it through re-integration. In addition, due to data processing speed and issues of the utilization system in the actual industry, there is a growing tendency to compress the size of data when using time series data and integrate it. However, since the guidelines for integrating time series data are not clear and each characteristic such as data description time interval and time section is different, it is difficult to use it after batch integration. In this paper, two integration methods are proposed based on the integration criteria setting method and the problems that arise during integration of time series data. Based on this, integration framework of a heterogeneous time series data was constructed that is considered the characteristics of time series data, and it was confirmed that different heterogeneous time series data compressed can be used for integration and various machine learning.

Keyword : Time Series integration, Multivariate Time Series, Classification, Sensor Data

Ⅰ. 서 론

최근 4차 산업은 지능화, 자동화로 진화하며 이를 위한 IoT 산업도 확대되고 있다. 헬스케어, 보안, 금융, 스마트 홈, 스마트 팩토리 등의 다양한 산업군에서 IoT가 활용되며, 이는 센서 데이터의 증가로 이어진다[1]. 센서 데이터는 실시간 상황 정보를 표현하는 데이터로 주로 시간 정보를 담고 있는 시계열 데이터 형태로 기록된다.

과거에는 정형화된 목적으로 수집된 데이터를 활용하는 것만 고려하였지만, 현재는 서로 다른 목적으로 수집된 데이터들 중 필요한 데이터들만 선택하여 적시 적소에 활용하고자 하는 필요성이 대두되고 있다[2]. 공장 내 안전한 근무 환경 정도를 예측하는 서비스 예를 들면 미세먼지, 일산화탄소, 오존, 이산화질소 등의 기본 공기 질 데이터를 중심으로 근접 지역의 화재 감지 정보, 날씨 정보, 근무자의 행동 정보 등을 함께 고려한다면 보다 정확한 서비스를 제공할 수 있을 것이다. 다시 말하면 기본 단일 시계열 데이터 뿐 아니라 관련 서로 다른 이질적 형태의 시계열 데이터 정보 또한 함께 활용되어야 한다. 더불어, 실제 산업에서는 데이터 저장 능력, 학습 처리 속도 등 각기 다른 활용 시스템 특성에 따라 수집된 원데이터의 크기를 압축 후 통합하여 활용하는 추세이다.

통합 시계열 데이터를 활용하기에 앞서 서로 다른 도메인의 이질적 시계열 데이터 통합이 선행되어야 하지만, 이질적 시계열 데이터간 통합은 쉽지 않다. 여러 산업에서 불특정한 형태로 수집된 데이터들에 관한 통합 표준 및 가이드라인이 존재하지 않으므로 일회성 연구에 맞춰 단발적으로 통합하는 방법 외에는 쉽게 통합 및 활용하는 것은 어렵기 때문이다.

또한 시계열 데이터는 서로 다른 환경에서 각각의 특성과 형태를 갖도록 저장되므로 통합 시 고려할 문제점이 있다. 대표적인 문제로는 데이터 손실과 결측 값 생성으로 인한 누락 데이터 처리, 불 규칙한 시간 간격을 보유한 데이터에 대한 통합 등이 있다. 소실된 데이터 정보의 중요도에 따라 기계 학습 성능이 좌우될 수 있으며 결함이 있거나 불 균일한 시간 간격으로 기술된 시계열 데이터는 균일한 시간 간격을 기준으로 동작하는 기계 학습의 성능 저하를 초래한다. 통합 시계열 데이터를 올바르게 활용하기 위해서는 통합과 동시에 결측 값 제거와 균일한 시간 주기로 조정하는 데이터 처리가 필수적이다.

따라서 본 연구에서는 서로 다른 이질적 특성을 지닌 시계열 데이터를 통합하는 프레임워크를 제안하고 이를 바탕으로 분류 문제를 해결하도록 활용 방법을 제시한다. 또한, 실제 산업의 시계열 데이터 활용도에 따라 서로 다른 크기로 압축된 데이터에 제안하는 통합 프레임워크를 이용하여 원 데이터와의 성능을 비교하였고 이를 통해 제안하는 프레임워크의 활용도를 증명하였다.

Ⅱ. 관련 연구

보다 나은 기계학습 성능을 위해 이질적 데이터들을 결합하여 응용하는 연구가 시도되고 있다. 데이터만으로는 진위 판단이 어려우므로 관련 소셜 데이터를 결합하여 잘못된 뉴스를 탐지하는 방법이 제안되었다[3]. 또한 금융 기업 조기 경보 모델의 성능을 높이고자 각 금융기관에 대한 민원 데이터와 뉴스 기사 데이터를 거시 경제 데이터와 재무 데이터에 결합시킨 통합 데이터를 응용하였다[4]. 이 연구에서는 두 개의 비정형 데이터와 두 개의 정형 데이터를 병합하여 모델 성능을 개선하였다. 다른 예로 작물을 재배하는 환경의 데이터(온도, 습도, 일사량 등)와 작물의 생장 데이터를 통합하여 작물의 성장 정도를 예측한다.

이처럼 많은 산업에서 데이터의 종류 및 형태에 국한되지 않고 데이터를 결합하여 활용하고 있으며, 이와 같이 다양한 데이터 통합을 통한 모델 성능 개선과 관련된 실험은 계속 진행되었다. 하지만, 여러 산업에서 불특정한 시계열 데이터를 통합하고 활용하는 방법에 대한 연구는 아직 미흡하다. 따라서, 본 논문에서는 시계열 데이터에 집중하여

a) 한국전자기술연구원(KETI)
‡ Corresponding Author : 문재원(Jaewon Moon)
E-mail: jwmoon@keti.re.kr
Tel: +82-10-9282-7975
ORCID: https://orcid.org/0000-0001-7451-6411
※ 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).
· Manuscript September 15, 2022; Revised November 15, 2022; Accepted November 15, 2022.

통합과 관련된 특성을 살펴보기 위하여 서로 다른 이질적 시계열 통합하고 활용하는 실험을 진행한다.

1. 시계열 데이터 변수 유형

본 논문에서는 범주 척도(Categorical Scale)와 양적 척도(Quantitative Scale)로 분류되며 각각을 다시 세분화한 스탠리 스미스 스티븐스의 측정 척도 이론의 구분법을 기준으로 삼았다[5]. 범주 척도는 측정 값이 아닌 그룹으로 묶이는 값이며 명목 척도(Nominal Scale)와 서열 척도(Order Scale)가 있다. 양적 척도는 측정 값으로 값의 크기 차이를 숫자로 연산이 가능한 척도이며 구간 척도(Interval Scale), 비율 척도(Ratio Scale)가 있다. 표 1은 측정 척도에 따라 가능한 연산 방법을 나타낸다[6].

표 1. 데이터 유형에 따른 연산 방법
Table 1. Method of operation according to data type

Operation	Nominal	Ordinal	Interval	Ratio
Equality	✓	✓	✓	✓
Order		✓	✓	✓
Add / Subtract			✓	✓
Multiply / Divide				✓
Mode	✓	✓	✓	✓
Median		✓	✓	✓
Arithmetic mean			✓	✓
Geometric mean				✓

명목 척도는 데이터 유형이 숫자나 고유명칭으로 지정한 경우를 의미하며 결측 치 연산 시 이전/이후 스텝 값인 동일한 정보 (Equality)를 활용하거나 KNN 알고리즘 등을 사용하여 해당 데이터에 가장 자주 발생하는 정보(Mode)로 보간 가능하다.

서열 척도는 순위 혹은 크기를 나타내며 예를 들어 '금메달', '은메달', '동메달' 정보는 순위가 존재하므로 데이터 유형은 서열 척도이다. 결측 치 연산은 Equality, Mode 값을 활용하거나 순서 비교(Order) 중앙값(Median)을 선택해 보완할 수 있다.

구간 척도는 순서와 함께 측정된 값 사이에 차이가 존재하여 명확한 간격 정보를 포함하는 형식이며, 결측 값 보완을 위해 명목 및 서열 척도가 쓰는 방법 이외에도 덧셈 뺄셈 및 산술평균(Mean) 연산을 활용할 수 있다. 결측 치를 처리할 때 기술된 정보가 나열된 주변 데이터의 의존성이 존재할 경우에는 곱하기, 나누기, 선형 방법론을 활용하여 데이터 정리를 할 수 있다.

비율 척도는 순서, 간격 정보와 비율 정보를 모두 갖는 데이터 형태이다. 이는 정보들 간의 비율 비교가 가능한 유형으로 예를 들어, 사과 무게가 100g, 200g인 데이터가 존재할 시 200g의 사과는 100g의 사과 무게보다 2배라고 표현할 수 있다. 이 유형은 구간 척도 연산자 외 곱셈 및 나눗셈을 이용해 데이터를 보간 할 수 있다.

2. 데이터 통합을 고려한 시계열 데이터 특성

시계열 데이터는 시간 흐름에 따라 정보 값이 나열된 데이터로 시간 정보를 배제하고 순서 정보만 활용한다면 중요한 정보를 손실하게 된다. 그러므로 데이터 정보 뿐 아니라 이를 기술하는 시간 정보를 고려한 데이터 정제가 선행되어야 한다[7][8].

시계열 데이터가 기술된 시간 간격은 정보가 저장되는 시간차를 의미한다. 시계열 데이터의 기술된 시간 간격을 고려하지 않고 통합할 시 문제가 발생한다. 그림 1은 서로 다른 주기의 데이터를 통합한 예시이다. 12시간마다 측정된 미세먼지 데이터와 일별 기록되는 오존 데이터를 통합한다고 가정하자. 그림 (C-1) 처럼 통합 데이터의 주기를 시간 간격이 중복되는 가장 작은 단위인 24 시간을 기준으로 하면 (a) 데이터의 2022년 8월 21일 12시의 정보는 손실된다. 만약 더 작은 단위인 12 시간 주기로 통합하면 오존 데이터는 기술된 주기에 정보가 없는 경우가 발생해 (C-2)와 같이 인위적인 결측 치가 생긴다. 다른 예로 5분 주기인 데이터와 7분 주기의 데이터를 일괄 결합 시 통합 데이터의 주기는 [5, 7, 10, 14 …] 등의 불규칙한 빈도로 기술되어 인위적인 결측 값을 생성한다. 이러한 시계열 데이터의 불균일한 기술 주기와 인위적으로 생성된 결측 값은 시간 정보를 순차적으로 처리하는 기계학습 알고리즘의 성능을 저하시킨다. 따라서, 시계열 데이터 통합 시에는 기술된 주기를 고려하여 데이터를 처리한 후 결합해야 한다.

데이터의 기술된 시간 구간은 시간이 저장된 기간을 의미하며 데이터 길이와 연관된다. 예를 들어서 2022년 8월 21일부터 24일까지 수집된 미세먼지 데이터와 8월 19일부

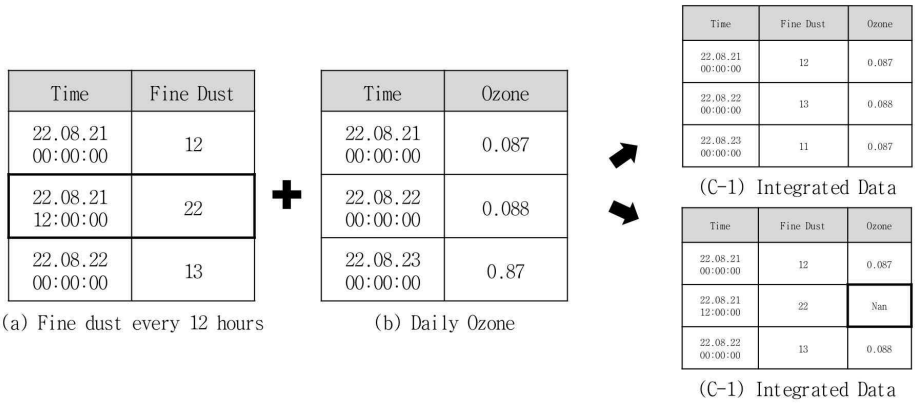

그림 1. 서로 다른 주기의 데이터 통합 예시
Fig. 1. Examples of data integration with different cycles

터 22일까지 수집된 오존 데이터를 총합하면 통합 데이터의 시간 구간은 2022년 8월 19일부터 24일로 구성된다. 이러한 경우 각 데이터의 기간이 겹치지 않는 부분은 정보가 존재하지 않아 결측 값이 발생한다. 이처럼 시간 구간도 서로 다른 시계열 통합 시 고려되어야 하는 요소이다. 결합 데이터의 시간 구간을 설정하는 기준 중 일반적으로는 각 데이터의 모든 데이터 시간에 대한 전체 범위를 선택하여 통합한다. 이는 초기 각 데이터에 대한 유실률은 적지만, 통합된 데이터의 결과에는 비 공통 시간 구간에 대해 인위적인 결측 값이 생성될 수 밖에 없다. 다른 방법은 각 데이터의 공통 시간 구간을 기준으로 통합하여 각 데이터의 비 공통 시간 구간의 정보를 제거해 추가로 생성될 수 있는 결측 값의 개수를 최소화하는 방법이다.

3. 데이터 재 표현

데이터 형태 일부를 변형 혹은 기존 데이터의 특성을 지닌 새로운 데이터로 변환하는 것을 데이터 재 표현이라고 정의하였다. 이미지 데이터는 픽셀 단위로 수치 변환하여 특징데이터를 추출하고 자연어는 기준 단위를 바탕으로 Feature vector 형태로 임베딩하는 과정을 통해 데이터 특징을 재현한다. 일반적으로 시계열 데이터는 산업군의 특성에 따라 각각 서로 다른 특성을 지니고 있으며 데이터 샘플링 및 재 추출 절차 후 활용한다.

이와 관련하여 센서 시계열 데이터의 재 표현 문제가 있을 수 있다. 첫 번째는 데이터 불균형 이슈이다. 예를 들어 반도체 웨이퍼 품질 분류 시 일반적으로 불량이 일어나기 전 정비를 통해 결함을 예방하므로 수집된 데이터는 정상 상태의 데이터가 큰 비중을 차지한다[9]. 그러므로 정보가 많은 클래스의 데이터 비율은 낮추고 정보가 적은 클래스의 비율을 높이는 샘플링 기법을 통해 데이터 비율을 균등하게 조정하여 해결한다. 또한 수집기의 에러, 데이터의 특성에 따라 시간 간격이 불규칙하게 수집된 경우로 불규칙한 데이터를 보정하기 위해 선형 보간 법과 단순이동 평균법을 활용해 시계열 데이터의 시간 간격을 균일하게 리 샘플링 할 수 있다[10].

데이터의 불균형과 시간 간격의 불규칙은 기계 학습 성능을 떨어뜨린다. 이 외에도 비 선형 데이터를 선형 데이터로 변형하여 활용하는 케이스처럼 데이터의 특성이 특정 학습에 적합하지 않은 형태인 경우 데이터 재 표현 과정이 필요하다.

3.1 결합을 고려한 데이터 재 표현 방법

데이터 재 표현하는 방법에는 각 데이터의 특성을 고려한 다양한 방법이 있다. 이번 장에서는 시계열 데이터의 기본적인 재 표현 방법인 샘플링(Sampling) 과 학습을 통한 차원 변환 방법을 소개하겠다.

샘플링의 방법에는 업 샘플링과 다운 샘플링이 있다. 시계열 데이터의 업 샘플링은 데이터 기술 시간 간격을 기존 간격보다 짧게 하여 더 자주 데이터를 기술하는 것으로 기술 시간을 추가로 생성하기 때문에 결측 치가 발생할 수 있다. 반대로 다운 샘플링은 기술된 시간 간격을 길게 시간

조정하여 데이터 양을 줄이는 방식으로 원본 데이터의 일부가 유실된다.

시계열 데이터의 차원 변환을 통한 재 표현 방법에는 Towards Universal Representation of Time Series (TS2Vec)[11], Long Short-Term Memory AutoEncoder (LSTM Auto-Encoder)[12][13][14] 등 다양한 방법이 존재한다. 본 논문에서는 데이터 결합으로 인해 발생한 데이터 결함 보정을 고려하여 LSTM AutoEncoder 모델을 활용하였다. LSTM AutoEncoder 는 시간 정보 처리가 가능한 LSTM과 데이터의 특징을 학습하여 복원하는 AutoEncoder로 구성된 모델이다. LSTM은 은닉층에서 다음 은닉층으로 정보를 전달하는 순환 신경망인 vanilla RNN의 단점을 개선한 고도화된 모델이며 AutoEncoder는 입력 데이터의 주요 정보를 학습하여 데이터 특성을 담은 벡터로 차원을 축소하는 Encoder와 이 벡터를 기반으로 원본 데이터와 유사한 데이터로 재표현하는 Decoder 로 이루어진 데이터 복원 알고리즘이다.

본 연구에서는 통합 데이터의 결측 치 정리 등 데이터 보정을 위한 데이터 재 표현 방법을 적용하였고 이때 샘플링 방법과 학습을 통한 차원 변환 방법을 이용했다.

Ⅲ. 시계열 데이터 통합 프레임워크 제안

1. 시계열 통합 방법 프레임워크

그림 3은 제시하는 시계열 통합 방법으로 서로 다른 특성을 갖는 N개의 데이터들의 통합 과정을 보여준다. 통합 방법으로는 데이터 차원 변환을 기반한 통합(Integration By

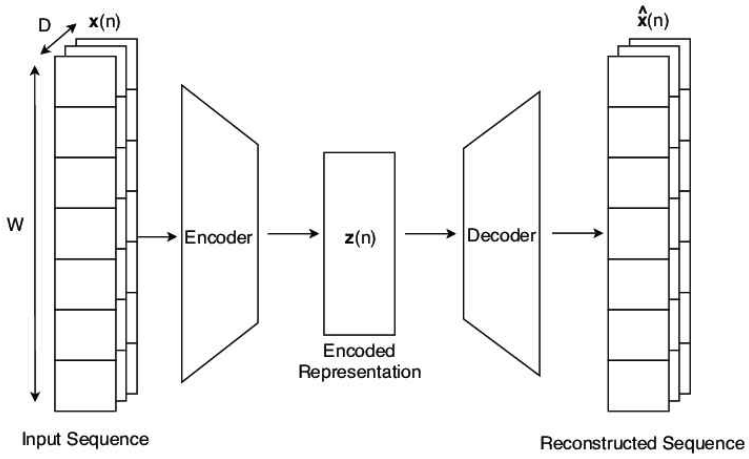

그림 2. LSTM Autoencoder[15]
Fig. 2. LSTM Autoencoder[15]

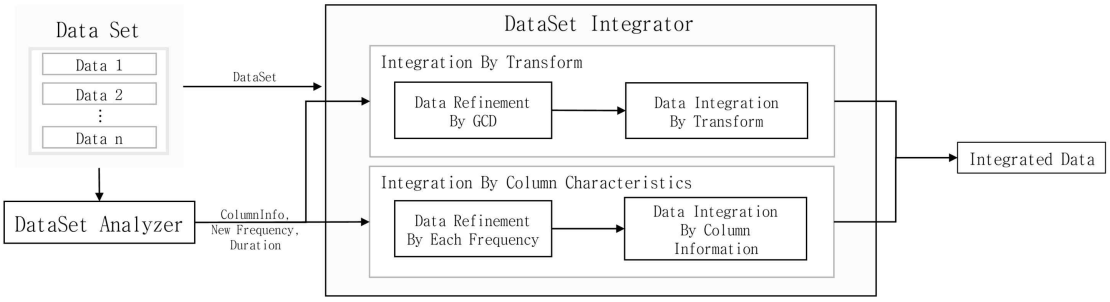

그림 3. 제안하는 시계열 데이터 통합 구조도
Fig. 3. Suggested time series data integration structure diagram

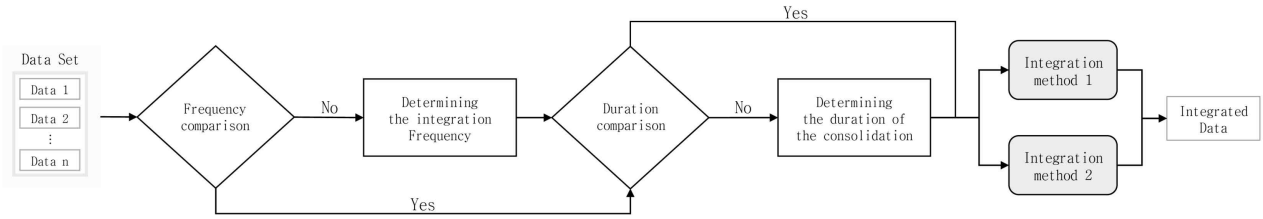

그림 4. 통합 기준의 시간 간격 및 시간 구간 결정 도식화
Fig. 4. Schematic for determining time intervals and time durations for the integration criteria

Transform) 과 데이터 특성을 기반한 통합(Integration By Column Characteristics) 방법을 제시하였다.

기본적으로 통합을 위한 관련 정보가 필요하다. 데이터 셋 분석기(DataSet Analyzer)는 본격적인 데이터 통합을 위해 필요한 통합 전 데이터 정보를 생성한다. N개의 개별 데이터를 입력 받아 각 데이터의 변수 단위 특성 정보(ColumnInfo: 데이터 정보 유형, 변수 주기 등)를 생성하고 데이터 통합 기준 시간 주기 (New Frequency: 새로운 통합 데이터의 기술 주기) 와 통합 시간 구간을 결정한다.

그림 4는 시계열 데이터 통합 프레임워크 중 데이터 셋 분석기의 기능을 도식화한 것이다. 이를 통해 통합에서 기준이 되는 시간 주기, 시간 구간을 생성한다. 입력 데이터의 주기 및 구간을 비교하는 방법은 먼저 N개의 데이터의 주기와 구간이 서로 일치/불일치 한지 확인한다. 비교한 결과 일치하면 일치한 정보 값이 기준 값으로 설정된다. 그러나, 불일치하다면 새로운 통합 주기 및 구간 결정 과정을 거쳐 다음 단계로 이어간다. 예를 들어서, 2021년 1월 1일부터 2021년 12월 31일까지 일년간 수집된 두 센서 데이터가 하나는 1시간을 기준으로 수집되었고 또 다른 하나는 2시간을 기준으로 정보가 생성되었다고 가정하면 이 두 데이터의 주기는 불일치하므로 통합 기준 주기 결정 단계를 거쳐 기간 비교 단계를 진행하며 두 데이터의 기간은 동일하기

표 2. 시간 주기 결정 방법의 예시
Table 2. Example of Time Frequency Determination Method

Integration Reference Time Frequency Setting Method	Time Frequency
Minimum of specified time frequencies	3 minutes
Maximum of a specified time frequencies	10 minutes
Most frequent appearance time frequency	5 minutes
Average of specified time frequencies	6 minutes
Greatest Common Divisor of specified time frequencies	1 minutes
Other User Inputs	30 minutes

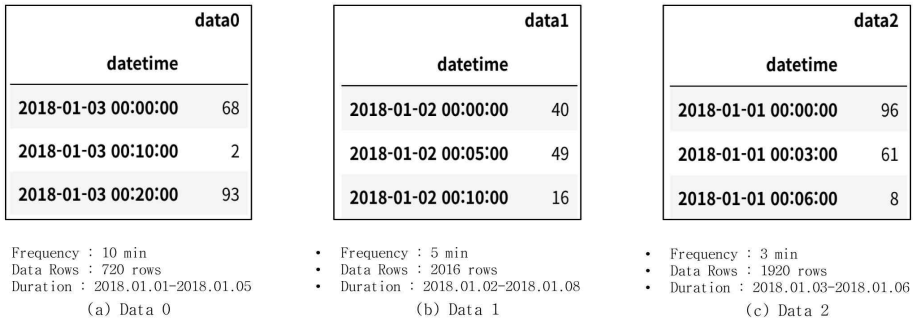

그림 5. 서로 다른 특성의 데이터 예시
Fig. 5. Examples of data with different characteristics

때문에 따로 통합 기간 결정 단계를 거치지 않고 통합하는 절차를 밟는다. 주기와 기간을 판정한 후 본격적으로 데이터 통합을 진행한다.

새로운 통합 데이터 기준 주기 설정 방법은 기술된 주기 중 최소 값, 기술된 주기 중 최대 값, 기술된 주기 중 최다 빈도 출연 주기 값, 기술된 주기들의 평균 값, 기술된 주기들에 대한 최대 공약수로 자동 설정될 수 있으며 임의의 사용자 입력으로 설정할 수 있도록 하였다. 통합 데이터의 시간 구간은 입력 N개 데이터에 따라 각각의 시작 시간 N개 ~ 끝나는 시간 N 개 중 선택하여 구성될 수 있으므로 최대 N의 제곱만큼의 시간 구간 경우의 수가 발생할 수 있다. 일반적으로 입력 데이터들의 전체 구간 혹은 각 데이터가 존재하는 공통 시간 구간을 결합 시간 구간으로 지정할 수 있다.

그림 5는 서로 다른 주기와 범위를 갖는 3개의 서로 다른 시계열 데이터로 표 2와 같이 통합 주기를 결정할 수 있다.

2. 데이터 특성 기반 통합 방법

첫 번째 방법은 데이터 변수 별 특성을 분석한 정보에 의거하여 업/다운 샘플링 방법을 활용한 데이터 재 표현을 통해 결측 치를 정리한 후 통합하는 방식이다.

그림 6은 변수 특성 기반으로 데이터를 통합하기 위한 업/다운 샘플링과 그에 따른 결측 치 처리 방법에 관한 도식화이다. 업/다운 샘플링 선택은 입력 데이터의 각 변수 별 기본 특성을 파악하여 생성한 변수의 시간 주기 값, 변수의 주기성을 바탕으로 진행된다. 변수의 주기성이 있다면 데이터 기술 주기(Column Freq)와 통합 기준 주기(Integration Freq)를 비교하고 주기성이 없다면 비 주기 임으로 데이터 기술 평균 주기(Column Average Freq)와 통합 기준 주기(Integration Freq)를 비교한다. 데이터 기술 주기(데이터 기술 평균 주기)가 더 클 경우 업 샘플링, 작을 경우 다운 샘플링으로 선정한다. 그림 4의 세 개의 데이터로 업/다운 샘플링 방법을 정해보자. 먼저, 그림 4의 세 개의 데이터는 각각 10분, 5분, 3분으로 주기성이 있으므로 각 데이터 별 기술 주기와 통합 주기를 비교하여 업/다운 샘플링 방식을 정한다. 통합 할 때 통합 주기를 평균 값인 6분으로 기준 삼는다면 (a), (b)는 다운 샘플링을 (c) 데이터는 업 샘플링으로 진행해야 한다.

그 다음으로 결측 값을 다루기 위한 업/다운 샘플링의 연산 방법 선택 과정이 시작된다. 변수의 주기성이 있는 경우 데이터 변수 유형을 1차/2차로 분석해 이 정보를 기반으로 연산 방법을 선정하고 반대로 비 주기 정보를 표현한다면 데이터 변수 유형과 상관없이 결측 치를 0으로 처리한다.

본 연구에서는 통계학적으로 데이터 유형을 나누는 방법인 범주 척도, 양적 척도와 문자 내용이 수집된 시계열 데이터와 같이 범주 유형이 아니지만 숫자로 표현되지 않은 데이터들을 고려하여 데이터 변수의 1차 유형을 정하였다. 데이터 변수의 1차 유형은 카테고리(Category) 판별이며 카테고리가 아닐 시 1차 유형은 문자열(String)과 숫자형(Numeric)으로 나뉜다.

2차 유형은 1차 유형이 카테고리와 문자열 일 경우 명목 척도/서열 척도 중 결정되며 숫자형 일 경우 구간 척도/비율 척도 중에서 고를 수 있다. 최종적으로 업/다운 샘플링 연산

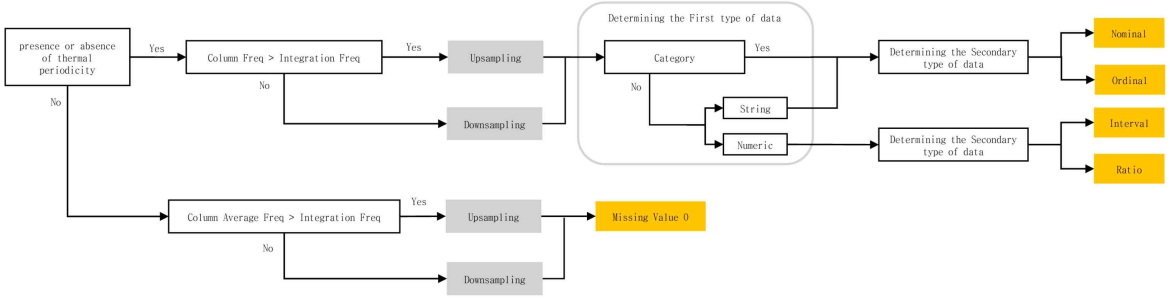

그림 6. Up/Down Sampling과 결측 치 처리 방법 도식화
Fig. 6. Schematic of Up/Down Sampling and Missing Value Handling Method

방법은 2차 유형인 명목/서열/구간/비율 척도에 따라 다르게 결정된다. 두 가지 예를 들어서 자세히 설명해보겠다.

시간별 출하되는 과일 종류의 정보를 담은 데이터를 활용해 과일을 분류하는 시스템에서 해당 데이터의 결측 값을 처리하는 과정을 가정해보자. 데이터는 '포도', '수박', '복숭아' 와 같이 측정 값이 아닌 고유명칭의 레이블이 저장된 데이터로 1차 유형은 카테고리이며 2차 유형은 명목 척도이다. 이러한 경우 연산이 불가하여 업/다운 샘플링은 이전 혹은 이후 스텝으로 채우기 방법, 또는 데이터 값의 최다 출연 데이터로 채우는 방법을 활용하여 진행할 수 있다.

두 번째로 품질에 따른 사과 수확량을 예측하는 시스템을 가정해보자. 해당 시스템에서는 일별 수확되는 사과의 무게를 수집한 데이터와 사과 수확에 영향을 미치는 온도 데이터, 사과의 품질 정보가 저장된 데이터를 활용한다. 1차 유형이 카테고리가 아니며 비율 척도인 사과 무게 데이터와 구간 척도인 온도 데이터는 산술평균 연산을 이용한 업/다운 샘플링을 통해 결측 값을 해결 할 수 있다. 사과 품질 데이터는 '좋음', '보통', '나쁨' 과 같이 사과의 품질성을 의미하는 레이블로 저장되어 1차 유형은 카테고리이며 2차 유형은 데이터의 서열을 비교할 수 있는 서열 척도이다. 따라서 품질 데이터의 업/다운 샘플링 연산은 비교연산, 중앙 값으로 처리 등의 방식을 이용하여 데이터의 결함을 조정할 수 있다.

3. 차원 변환 기반 통합 방법

해당 방식은 서로 다른 특성의 데이터 통합을 통해 규칙적인 주기로 조정함으로써 발생한 결함에 대해 차원을 변환하여 정돈된 시계열 통합 데이터를 생성하는 방식이다. 차원 변환 기반 통합 방법에서 통합 기준 주기는 임의의 사용자 입력으로 지정하고 데이터 유실을 방지하기 위해 입력 데이터들의 전체 시간 범위로 통합 기간을 선정했다.

이 방법의 목적은 차원 변환을 통해 누락 데이터를 해결하는 것으로 원리는 결측 값을 임의의 값으로 대체하는 데이터 정제 과정을 지나 데이터 특징을 학습하는 모델을 적용하여 데이터의 중요한 정보를 저장한 차원 변환 데이터로 표현하는 것이다. 임의의 값 결정 방식은 데이터 값의 의미에 따라 두 가지가 있다. 예를 들어 움직임 센서 데이터와 같이 0값이 없음을 뜻하는 데이터는 결측 치를 0으로 대체하며 온도 데이터와 같이 숫자 0에 대해 의미를 포함한 데이터는 해당 데이터 값의 범위를 벗어난 특정한 값으로 채워준다. 두 가지 방식이 가능한 이유는 데이터의 특성을 학습하여 차원을 변환하는 방법이기 때문이다.

본 연구에서는 LSTM AutoEncoder를 통해 차원 변환 데이터를 생성하였다. 데이터의 복원을 위해 데이터 특징을 학습하는 LSTM AutoEncoder의 특성을 이용하여 데이터 특징을 담은 latent vector 를 추출해 이를 새로운 차원의 데이터로 정의한 통합 데이터를 출력한다. 이때, 데이터의 특징을 학습하기 위한 Loss Function은 실제 값과 복원 값의 절대값 차이를 활용하는 L1 Loss를 활용하였다. LSTM AutoEncoder와 유사한 다른 모델을 사용하여 진행 할 수 있다. LSTM AutoEncoder는 비지도 학습 방법 중 하나인 Reconstruction을 활용하여 특징 학습 후 입력 데이터를 복원한다. 따라서, LSTM AutoEncoder 외에도 Transformer와 같이 Reconstruction이 가능한 구조의 모델을 사용하여 차원 변환 기반 통합을 할 수 있다. 그림 7은 3분, 7분, 10분 주기를 갖는 서로 다른 데이터의 차원 변환 기반 통합

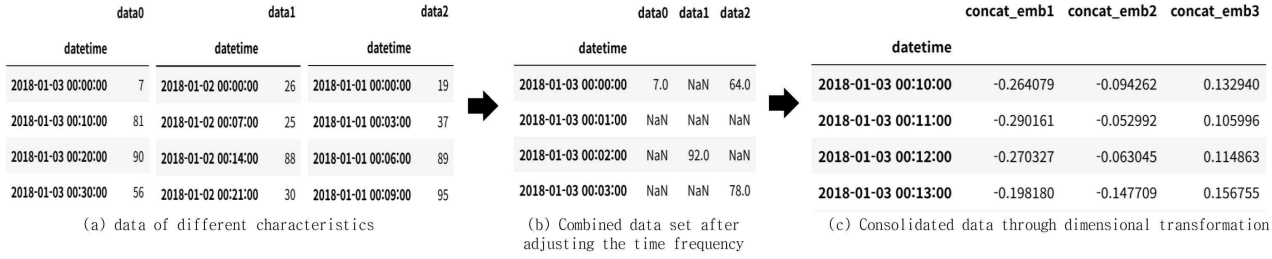

그림 7. 차원 변환을 통한 통합 데이터 예시
Fig. 7. Example of consolidated data through dimensional transformation

방법을 진행한 결과를 보여준다.

Ⅳ. 실험 개요 및 결과

1. 실험 개요

이번 장에서는 서로 다른 이질적 시계열 데이터를 제안하는 프레임워크를 이용해 이질적 시계열 데이터를 통합하고 분류 문제에 적용한 결과를 소개한다.

원 데이터를 그대로 활용한 경우와 각각 다르게 압축된 데이터를 제안한 프레임워크를 이용해 통합하여 활용한 경우 성능을 비교한다. 더불어, 저장 용량 문제로 기술 시간 주기를 늘려 데이터를 압축할 경우 삭제 혹은 다운 샘플링하는 방법을 적용하고 그 성능 차이도 비교하였다. 또한 통합 기준 주기를 다르게 하여 통합 데이터를 생성하고 분류 문제를 해결해보았다.

각각의 실험을 위해 같은 주기를 갖는 원 데이터를 서로 다른 주기를 갖는 개별 데이터로 변환한 후 실험 케이스의 조건에 따라 통합 데이터를 생성한다. 생성된 데이터를 활용하여 분류 문제를 해결하며 그 테스트 결과 정확도를 비교하였다.

2. 실험 데이터 셋 설정

UCI HAR(Human Activity Recognition) 는 9개의 독립변수와 1개의 종속 변수로 구성된 센서 데이터셋으로 30명의 실험자들의 걷기, 눕기, 계단 오르기, 계단 내려가기, 서있기, 앉기 등 6개 행동 시 센서에서 측정된 패턴 데이터이다. 각 데이터는 X, Y, Z에 대한 신체 가속도, 중력 가속도, 전체 가속도 크기가 기술된 9개의 독립 변수로 이루어지며 50Hz의 주기로 128번 동안 수집되었다. 트레이닝 데이터셋은 (7352, 9, 128), 테스트 데이터셋은 (2937, 9, 128) 형태로 이루어져 있다. 분류 문제의 성능 지표인 정확도 계산은 시간에 따라 나열된 9개 독립 변수 정보들이 알맞은 1개의 종속 변수로 올바르게 분류가 되는 지를 측정한다.

서로 다른 이질적 시계열 데이터 통합 및 분류 실험에 응용하기 위해 그림 8과 같이 UCI HAR 데이터를 서로 다른 특성의 개별 데이터로 분리 및 변환했다. 그림 8의 (a)는 (9, 128) 데이터가 7352개 모인 (7352, 9, 128)의 원본 UCI HAR 데이터 셋이며 (b)와 같이 2차원의 데이터프레임으로 변환한다. 이 후 (c)와 같이 데이터 프레임을 신체 가속도, 중력 가속도, 전체 가속도를 기준으로 각각 세개의 독립 변수를 갖는 데이터로 분할 한다. 이를 기반으로 실험 케이스에 따라 시간 간격을 변경하여 데이터를 재 표현한다. 세 가지의 서로 다른 이질적 시계열 데이터를 생성한 후 설정한 통합 시간 간격과 통합 시간 기간을 기준으로 (d)와 같은 통합 데이터를 추출하며 이 과정을 7352개의 데이터 반복 적용하여 분류 모델에 활용하도록 최종 데이터 셋을 생성한다.

그림 9는 그림 8의 (C)과정으로 UCI HAR 입력 데이터를 실험 케이스1에서 1/2/4단위로 시간 간격을 변형 및 분할하는 경우의 예시 그래프이다. 통합에 앞서 입력 데이터를 세 가지로 분할하고 1/2/4 단위로 시간 변경을 통해 재표현한 그래프로 각 데이터별 시간 간격이 다르며 이에

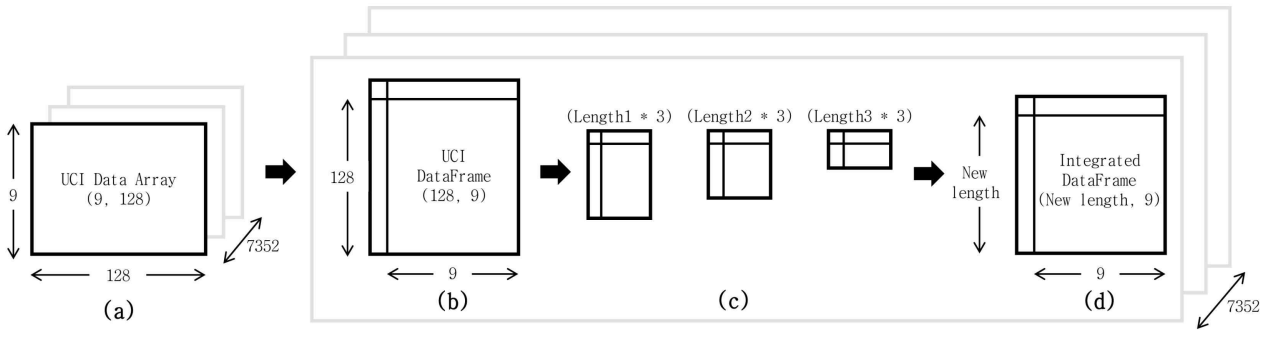

그림 8. 연구 활용 맞춤 데이터 셋 생성 과정
Fig. 8. Process of creating custom data set for research use

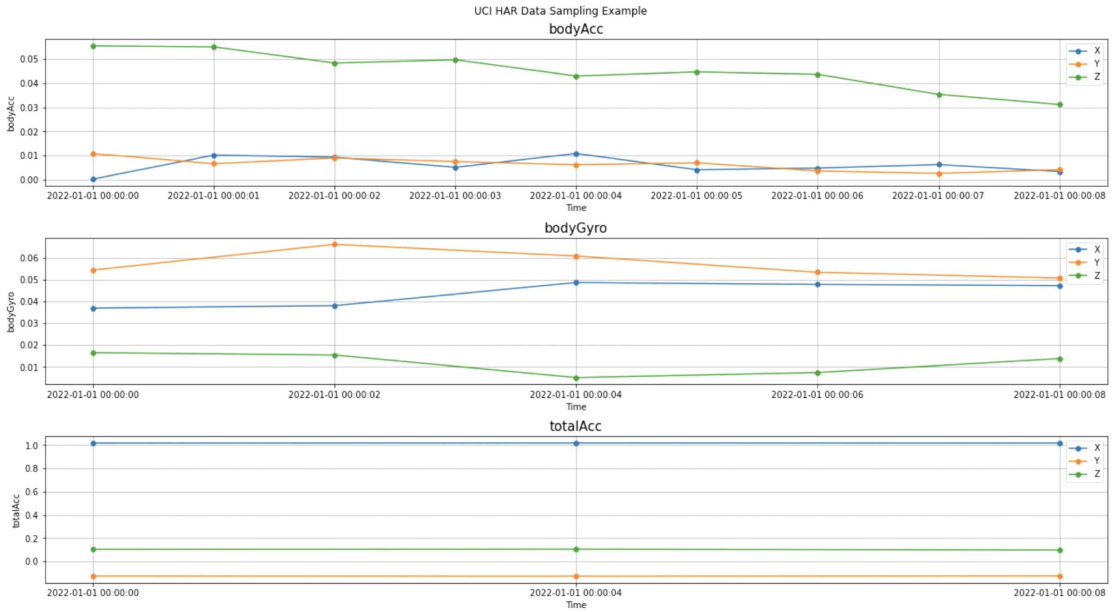

그림 9. 분할 및 재 표현 데이터 일부에 대한 그래프 예시
Fig. 9. Segmentation and Representation Data Graph Example

따른 데이터 특성이 다르다는 점을 알 수 있다.

개별 데이터 압축 시 기본적으로는 평균 샘플링 방식을 적용하였으며 데이터 특성을 기반한 통합(Integration By Column Characteristics) 방법을 활용하여 통합하였다. 별도로 데이터 압축 방법에 대한 평가를 위해 필요한 데이터만 선택하는 방법과, 평균 샘플링 방식을 비교하였다.

또한 통합 기준 주기를 다르게 했을 경우 분류 성능을 비교하였다. 새로운 통합 데이터의 기준 주기를 4가지로 나누어 지정하고 원 데이터와 압축 데이터를 제안하는 프레임워크를 통해 생성한 새로운 통합 데이터의 분류 성능을 비교하였다.

3. 실험 결과

분류 문제 해결을 위한 알고리즘으로는 Long Short-Term Memory(LSTM)과 1D Convolutional Neural Networks (1D-CNN) 알고리즘을 활용했다[16][17]. 1D-CNN 모델은 Convolutional Neural Network로 합성곱 연산을 사용한 딥러닝 모델인 CNN의 구조를 1차원 합성곱으로 연산 하는 신경망 모델이다. 3장에서 설명한 시계열 데이터에 특화된 LSTM 모델과 달리 CNN 모델은 비디오와 이미지 데이터에 특화된 신경망 모델이지만 1D-CNN은 1차원 시퀀스 데이터에 대한 합성곱 연산이 가능하여 시계열 데이터에서도 활용이 된다[18][19]. LSTM 모델 학습 파라미터의 num layers는 2, hidden size는 64, dropout은 0.1, learning rate는 0.0001으로 일괄적으로 설정하였다. 1D-CNN의 파라미터 output channels는 64, kernel size는 3, stride는 1, padding은 0, drop out은 0.1, learning rate는 0.0001으로 설정했다.

표 3는 원 데이터를 새로운 목표 주기로 일괄 삭제하여 샘플링한 데이터(Original Data 군)와 이미 압축된 파편적 데이터를 제안하는 프레임워크에 기반하여 새로운 목표 주기로 통합한 데이터(Partial Data 군)에 대해 각각의 분류 알고리즘을 적용하여 성능을 비교한 결과를 나타낸다. 이 때 모델 성능 결과는 원 데이터와 설정한 단위에 맞춰 압축된 데이터 별로 각각 모델을 따로 학습시켜 분류 실험을 진행하였기 때문에 모델 훈련 과정에서 각 데이터들의 특징이 다르게 반영된 결과이다. 원 데이터에 대한 분류 정확도는 LSTM 알고리즘을 활용할 경우 88.7%, 1D-CNN의 경우 88.1%가 나왔다. 이 경우의 데이터는 Original Time Frequency와 New Time Frequency가 1인 원본 데이터를

변함없이 사용한 경우로 원 데이터만으로 분류 실험을 진행했을 경우(Original Time Frequency 가 1인 경우)와 같은 결과(LSTM인 경우 88.7%, 1D-CNN인 경우 88.1%)가 나온다는 것을 알 수 있었다. 1/4 크기로 변환하기 위해 4 개 단위로 다운 샘플링한 데이터를 실험한 결과는 각각 71.8%, 86.4%였다. 이는 데이터를 압축할 경우 성능이 낮아지는 것을 의미한다.

제안하는 프레임워크에 대한 성능을 비교하기 위해 1/2/4 단위로 이미 압축된 데이터를 프레임워크를 활용하여 새롭게 4개 단위로 통합한 데이터로 분류 정확도를 비교하였다. 그 결과 LSTM과 1D-CNN에서 각각 75.2%, 88% 로 앞서 Original Data군에서 실험한 결과 보다 성능이 나아지는 것을 확인하였다. 추가로 4/1/8 단위로 압축된 데이터에 대해 동일한 실험 진행시에도 74.1%, 81.7%의 정확도를 보였으며 이는 절대 성능은 다소 떨어지지만 개별 데이터 압축에 따른 데이터 저장 용량 이득을 고려했을 때 합리적인 수준으로 판단된다. 즉 기본적으로 알고리즘 성능은 데이터의 양에 비례하지만 제안하는 프레임워크를 활용한다면 압축 데이터의 결합으로도 원 데이터와 상등한 성능을 유지할 수 있음을 확인하였다.

표 4는 초기 데이터 압축 시 필요 데이터만 선택하는 방식과 평균 값으로 다운 샘플링하여 대표 값을 설정하는 방식에 따른 성능을 비교한 결과이다. 설정한 시간 간격을 기준으로 필요 데이터만 선택하는 것 보다 산술 평균 샘플링으로 압축한 데이터는 보다 많은 데이터 정보를 반영하여 압축했기 때문에 성능이 좋다는 것을 확인할 수 있었다.

통합 기준 주기에 따른 성능 비교 결과는 표 5와 같다. Original Data에 대해 서로 다른 주기로 데이터를 압축하고

표 3. 실험 케이스 1의 결과
Table 3. Result of Experimental Case No. 1

Data	Original Time Frequency			New Time Frequency	Data Modification Method	Classification Model	Accuracy
Original Data	1			1	No Method	LSTM	0.887
						1D-CNN	0.881
				4	Sampling (drop)	LSTM	0.718
						1D-CNN	0.864
Partial Data	1	2	4	4	Integration by Column Characteristics	LSTM	0.752
						1D-CNN	0.880
	4	1	8			LSTM	0.741
						1D-CNN	0.817

표 4. 실험 케이스 2 의 결과
Table 4. Result data of Experimental Case No. 2

Data	Sampling Method	Original Time Frequency			New Time Frequency	Classification Model	Accuracy
Partial Data	Drop	1	2	4	2	1D-CNN	0.864
					4		0.862
	Mean				2		0.875
					4		0.880

표 5. 실험 케이스 3 의 결과
Table 5. Result data of Experimental Case No. 3

Data	Original Time Frequency			New Time Frequency	Classification Model	Accuracy
Original Data	1			1	1D-CNN	0.881
				2		0.884
				4		0.864
				8		0.787
Partial Data	1	2	4	1		0.855
				2		0.875
				4		0.880
				8		0.864

기계 학습 적용시 압축 량이 커질수록 정확도는 낮아지는 경향을 보였다. 반면 이미 압축되어 서로 다른 주기를 갖는 Partial Data를 제안하는 프레임워크를 활용하여 통합한 후 분류 문제에 적용한 경우에도 Original Data에 비해 같은 주기 대비 더 나은 성능을 보임을 확인할 수 있었다. 특히 새로운 기술 주기가 클수록 더 나은 성능을 보였다. 이는 제안하는 프레임워크를 활용 시 개별 데이터가 압축된 상태임에도 유사한 성능을 얻을 수 있음을 확인할 수 있다.

실험 결과를 통해 데이터 통합 시 원본 데이터의 품질이 좋고 시간 간격이 균일할 수록 정확도에 긍정적인 영향을 미치는 것은 사실이지만, 최종 활용하려는 통합 데이터 시간 주기가 커진다면 제안하는 프레임워크 활용 시 더 나은 결과를 얻을 수 있음을 확인하였다. 데이터의 기술 시간 주기는 데이터의 저장 용량을 줄이기도 하지만, 절대 학습 시간을 줄일 수 있어 실제 산업에서 얻는 이득이 크다. 또한 제안하는 프레임워크는 서로 다르게 압축된 서로 다른 주기의 데이터를 기계학습 가능한 형태로 결합하고, 원 데이터에 준하는 분류 성능을 얻을 수 있음을 확인할 수 있었다.

Ⅴ. 결 론

본 논문에서는 시계열 데이터 통합의 중요성을 살펴보고 특성을 고려한 시계열 통합 프레임워크를 제안하였다. 통합 시간 간격 및 통합 시간 구간을 결정하는 단계와 데이터 재 표현 방법 결정을 위한 데이터 특성을 정의하였으며 데이터 주기성과 변수 별 유형에 따른 결측 값 처리 방향을 제시하였다. 더불어, 차원 변환을 통한 통합 방법을 바탕으로 데이터 특징을 추출하여 누락 데이터를 정리하는 기법도 소개하였다.

제안한 통합 프레임워크가 문제없이 시계열 데이터를 재 표현하고 통합하는 것을 검증하기 위해 실험을 진행하였으며 그 결과 데이터를 압축할 경우 데이터 정보를 최대한 반영하도록 샘플링 하는 것이 중요하며, 서로 다른 데이터의 최종 통합 시간 주기가 성능에 영향을 미치는 것을 확인하였다. 또한 제안하는 프레임워크가 서로 다른 기술 주기로 압축된 데이터를 통합하며 통합된 데이터로 분류 문제 해결시 원 데이터에 준하는 정확도를 보임을 확인할 수 있었다. 그러나 본 논문에서 실험한 데이터 종류가 한정적이며 분류 문제에 대해서만 검증을 진행하였기 때문에 향후 다양한 문제에 대해서 여러 서로 다른 데이터를 활용하여 실험을 진행해야 할 것이다.

통합 데이터 관련 연구가 활발히 진행되고 있으나 여전히 정형화된 데이터에 비해 연구가 부족하며 통합 방법에 대한 표준을 확립하고 더 많은 시계열 데이터 사례에 대해 연구 및 실험을 진행해야 할 것이다. 따라서 본 논문에서 검증한 분류 문제 외에도 제안하는 통합 프레임워크를 이용하여 다양한 시계열 기반 문제들을 해결하고 부족한 부분을 보완하여 고도화 할 예정이다.

참 고 문 헌 (References)

[1] Dong-Gyu Jeong.(2017).A Study on IoT-Related Industry Trend. Korea Institute of Information Technology Magazine, 15(1),31-37. https://www.dbpia.co.kr/Journal/articleDetail?nodeId=NODE07187891

[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. doi: https://doi.org/10.1109/BIGCOMP.2017.7881688

[3] Yoonjin Hyun, Namgyu Kim.(2018).Text Mining-based Fake News Detection Using News And Social Media Data.The Jounal of Society for e-Business Studies,23(4),19-39. http://www.jsebs.org/jsebs/index.php/jsebs/article/view/338

[4] Seoha Song, Junhong Kim, Hyungseok Kim, Jaeseon Park, Pilsung Kang.(2019).Development of Early Warning Model for Financial Firms Using Financial and Text Data : A Case Study on Insolvent Bank Prediction.Journal of the Korean Institute of Industrial Engineers, 45(3),248-259. doi: https://doi.org/10.7232/JKIIE.2019.45.3.248

[5] Stevens, S. S. (1946). On the theory of scales of measurement. Science, 103(2684), 677-680. doi: https://doi.org/10.1126/science.103.2684.677

[6] Matthew Renze. Nominal, Ordinal, Interval, and Ratio Data. https://matthewrenze.com/articles/the-four-subtypes-of-data-in-data-science/ (accessed June 15. 2019).

[7] 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. https://psycnet.apa.org/record/2007-00569-003

[8] Eden Kim, Seok-gap Seok, Seung-cheol Son, & Byeong-tak Lee. (2021). Technical Trends of Time-Series Data Imputation. Electronics

and Telecommunications Trends, 36(4), 145 - 153.
doi: https://doi.org/10.22648/ETRI.2021.J.360414

[9] Won Seok Lee, Hyun Hee Kang.(2020).Interpretable convolutional neural network model for yield prediction in semiconductor fabrication.Journal of the Korean Data And Information Science Society,31(5),691-720.
doi: https://doi.org/10.7465/jkdi.2020.31.5.691

[10] Kang-hyeon Shin, Kyo-hong Jin.(2021).Irregularly-Sampled time Series Correction Method for Anomaly detection in Manufacturing Facility. Proceedings of the Korean Institute of Information and Commucation Sciences Conference,25(2),85-88.
https://koreascience.kr/article/CFKO202132348514233.page

[11] Yue, Z., Wang, Y., Duan, J., Yang, T., Huang, C., Tong, Y., & Xu, B. (2022, June). Ts2vec: Towards universal representation of time series. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 36, No. 8, pp. 8980-8987).
doi: https://doi.org/10.48550/arXiv.2106.10466

[12] 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.
doi: https://doi.org/10.1007/s00521-019-04678-9

[13] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.
doi: https://doi.org/10.1162/neco.1997.9.8.1735

[14] Xue, J., Huang, Q., Wu, S., & Nagao, T. (2022). LSTM-Autoencoder Network for the Detection of Seismic Electric Signals. IEEE Transactions on Geoscience and Remote Sensing, 60, 1-12.
doi: 10.1109/TGRS.2022.3183389

[15] Detecting Mobile Traffic Anomalies Through Physical Control Channel Fingerprinting: A Deep Semi-Supervised Approach - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/LSTM-Autoencoder-for-Anomaly-Detection_fig2_336594630 [accessed 13 Nov, 2022]
doi: 10.1109/ACCESS.2019.2947742

[16] Du, Q., Gu, W., Zhang, L., & Huang, S. L. (2018, November). Attention-based LSTM-CNNs for time-series classification. In Proceedings of the 16th ACM conference on embedded networked sensor systems (pp. 410-411).
doi: https://doi.org/10.1145/3274783.3275208

[17] Zhao, B., Lu, H., Chen, S., Liu, J., & Wu, D. (2017). Convolutional neural networks for time series classification. Journal of Systems Engineering and Electronics, 28(1), 162-169.
doi: https://doi.org/10.21629/JSEE.2017.01.18

[18] Wibawa, A. P., Utama, A. B. P., Elmunsyah, H., Pujianto, U., Dwiyanto, F. A., & Hernandez, L. (2022). Time-series analysis with smoothed Convolutional Neural Network. Journal of big Data, 9(1), 1-18.
doi: https://doi.org/10.1186/s40537-022-00599-y

[19] Youngjun Jang, Jiho Kim, Hongchul Lee. (2022). A Proposal of Sensor-based Time Series Classification Model using Explainable Convolutional Neural Network. Journal of the Korea Society of Computer and Information , 27(5), 55-67.
doi: 10.9708/jksci.2022.27.05.055

저 자 소 개

황 지 수

- 2019년 2월 : 국민대학교 전자공학부 학사
- 2021년 5월 ~ 현재 : 한국전자기술연구원 연구원
- ORCID : https://orcid.org/0000-0002-8310-9727
- 주관심분야 : 데이터 사이언스, 시계열 데이터 분석

문 재 원

- 2002년 2월 : 성균관대학교 전기전자컴퓨터공학과 학사
- 2004년 2월 : 서울대학교 전기컴퓨터공학부 석사
- 2004년 1월 ~ 2007년 6월 : ㈜ 삼성전자 통신연구소 선임연구원
- 2007년 7월 ~ 2007년 9월 : ㈜ SKTelecom 네트워크 연구소 매니저
- 2009년 10월 ~ 현재 : 한국전자기술연구원 책임연구원
- 2019년 2월 : 성균관대학교 데이터사이언스학 박사
- ORCID : https://orcid.org/0000-0001-7451-6411
- 주관심분야 : 데이터 사이언스, 시계열 데이터 기계 학습, 엣지 컴퓨팅