# 서로 다른 시계열 데이터들간 통합 활용을 고려한 해시 함수 기반 학습 모델 관리 플랫폼

유미선[O], 문재원[*]
[O]한국전자기술연구원,
[*]한국전자기술연구원
e-mail: {altjs543, jwmoon}@keti.re.kr[O*]

# Learning model management platform based on hash function considering for integration from different timeseries data

Miseon Yu[O], Jaewon Moon[*]
[O]Korea Electronics TechnologyInstitute,
[*]Korea Electronics TechnologyInstitute

● 요 약 ●

IoT 기술의 발전 및 확산으로 다양한 도메인에서 서로 다른 특성의 시계열 데이터가 수집되고 있다. 이에 따라 단일 목적으로 수집된 시계열 데이터만 아니라, 다른 목적으로 수집된 시계열 데이터들 또한 통합하여 분석활용하려는 수요 또한 높아지고 있다. 본 논문은 파편화된 시계열 데이터들을 선택하여 통합한 후 딥러닝 모델을 생성하고 활용할 수 있는 해시함수 기반 학습 모델 관리 플랫폼을 설계하고 구현하였다. 특정되지 않은 데이터들을 기반하여 모델을 학습하고 활용할 경우 생성 모델이 개별적으로 어떤 데이터로 어떻게 생성되었는지 기술되어야 향후 활용에 용이하다. 특히 시계열 데이터의 경우 학습 데이터의 시간 정보에 의존적일 수밖에 없으므로 해당 정보의 관리도 필요하다. 본 논문에서는 이러한 문제를 해결하기 위해 해시 함수를 이용해서 생성된 모델을 계층적으로 저장하여 원하는 모델을 쉽게 검색하고 활용할 수 있도록 하였다.

**키워드:** 시계열 데이터(Time-Series data), LSTM, 파편화된 시계열 데이터(partial time series data)

## I. Introduction

날씨, 공기질 등을 측정하기 위해 사용되는 센서 데이터들은 시간의 흐름에 따라 취득되며 이러한 데이터를 시계열 데이터(time-series data)라고 한다. 특히, IoT 환경에 대한 발달 등으로 더 다양한 도메인에서의 시계열 데이터가 수집되고 있다.[1] 시계열 데이터는 타데이터와 달리 예측 및 활용을 위해서 시간의 개념을 고려해야 한다. 또한, 시간 축에 대한 의존성이 높고 데이터를 통한 직관적 의미 파악이 어렵다. 즉, 시계열 데이터에대한 전문적 지식 없이 적절한 학습 모델을 생성하고 활용하는 난이도가 높다. 이러한 이유로전문적 지식을 함양하지 않은 일반 사용자들도 쉽게 사용할 수 있는 시계열 데이터 처리 플랫폼에 대한 수요가 높아지고 있다. 실제로 마이크로소프트(Microsoft)의 클라우드컴퓨팅 서비스 애저(Azure)에서는 Time Series Insights[2] 라는 시계열 전용 플랫폼을 개발하여 IoT 환경에서 수집되는 데이터들에 대한 그래픽 시각화 도구를 제공하고 있다. 이외에도 많은 기업, 기관에서 관련 서비스를 제공하고 있다. 하지만, 기존 서비스들은 스마트 팩토리(smart factory)와 같은 특정 도메인에 편향되어 있는 경우가 많다. 다양한 도메인으로 파편화된 데이터를 통합하여 모델을 생성하고 응용하는 서비스를 제공하기 위해서는 추가적인 연구 개발이 필요한 상황이다.

학습 모델 제공 플랫폼이 특정되지 않은 데이터들을 기반하여 모델을 학습하고 활용할 경우 개별적 생성 모델이 어떤 데이터로 어떻게 생성 되었는지가 반드시 기술되어야 한다. 특히 시계열데이터의 경우 학습 데이터의 시간 정보에 의존적일 수밖에 없으므로 모델 생성 시 해당 정보의 관리도 필요하다. 상황에 따라 생성되는 각각의 모델들은 서로 연관성이 없고 독립적이라고 가정한다면, 다양하게 생성되는 서로 다른 모델들에 대한 검색/활용을 위한 합리적인 방법 제시가 필요하다.

본 논문은 서로 다른 데이터, 시간 구간으로 생성된 학습 모델 정보를 이용해 sha-1 해시 방식으로 고유값을 생성하고 생성된 모델

파일의 저장 구조를 계층화하여 관리하는 방식을 제안한다. 해당 방법론은 파편화된 시계열 데이터들을 선택하여 통합하고 이를 이용해 생성한 모델을 구조적으로 저장한다. 결국, 모델 검색 및 활용이 용이해진다. 본 논문은 해당 플랫폼의 구현 과정에서 생긴 요구 조건과 그에 대한 방법론을 함께 기술하였다.

## II. 해시 함수 기반 모델 관리 기법

### 1. 모델 관리 요구 사항

파편화되어있는 서로 다른 데이터들을 선택하여 새로운 통합 데이터를 기반으로 분석, 활용하면 새로운 의미를 도출할 수 있다.시계열 데이터는 같은 도메인이라도 어디서 어떻게 수집된 데이터인지에 따라서 구성이 달라진다. 예로 같은 공기질 데이터라도 서울에서 수집된 것은 5분주기, 인천에서는 10분주기로 수집되었을 수 있다. 통합데이터 생성에서 이 두 가지 데이터를 10분주기로 통합할 수도 있고, 30분주기로 통합할 수도 있다. 즉, 플랫폼 사용자가 요청할 수 있는 통합 데이터의 경우의 수는 매우 많고, 통합 데이터 하나를 구분할 수 있는 특징도 시간 주기, 수집 기간 등 여러 가지가 있다.

통합 데이터로 학습 후 모델을 생성하면 후에 해당 모델로 예측값을 추론할 때도 학습 시와 같은 통합 데이터의 인풋 형식이 대응되어야 한다. 즉, 각 통합 데이터를 인풋으로 사용하는 모델들이 무엇인지 명시되어야 한다. 추가적으로 시계열 데이터의 경우 시간 정보에 의존적이다. 같은 모델 설계라고 하여도 시간 및 다른 파라미터들에 따른 고유의 특징들을 가지고 있다. 따라서 같은 통합 데이터를 활용한 모델이라도 시간 정보,학습 파라미터 등으로 모델이 구분되어질 필요가 있다. 이러한 상황으로 각각의 모델들을 쉽게 구분하고 검색하여 응용할 수는 관리 기법이 요구된다.

### 2. 모델 관리 기법

각 모델들의 특징을 구분하고 이를 쉽게 검색/활용하기 위해서 생성된 모델 파일들을 디렉토리 계층 구조로 저장한다. 그러나 각 통합 데이터들을 나타내는 구분 특징들과 각 모델을 구분할 수 있는 모델 파라미터의 모든 값들로 계층 구조를 나누면 디렉토리의 깊이가 매우 깊어져 모델 검색이 복잡해질 수 있다. 본 연구는 이를 쉽게 관리할 수 있도록 통합 데이터의 정보를 활용해 통합 데이터 아이디를 생성하고 모델의 파라미터들을 활용해 모델 파라미터 아이디(model parameter id)를 생성한다. 통합 데이터 아이디는 데이터 결합에 사용했던 통합 데이터 정보의 고유값을, 모델 파라미터 아이디는 각 모델 파라미터들의 고유값을 나타낸다. 고유값으로는 SHA-1 해시(hash) 방식에서 해당되는 정보를 입력으로 했을 때, 출력되는 해시 값을 할당한다. 즉,다량의 깊이로 나타내야 했던 각 정보의 특징을 한 줄의 아이디로 압축하여 나타냄으로써 검색/활용의 효율성을 높인다.

해시 생성은[표 1]에서 보는 것과 같이sha256, MD5, sha-1, 랜덤 생성 등의 다양한 방식이 존재한다. 이 중에서 MD5는 파일 구별 용도로는 속도가 빨라서 이상적이지만, 16bytes의 짧은 해시 값을 생성하므로 충돌의 가능성이 크게 존재한다. [3] sha256 방식은 32bytes의 출력으로 통합 데이터의 고유 값을 표현하기에는 사이즈가 매우 크고 상대적으로 해시 생성이 복잡하다. 따라서 사이즈가 20bytes로 적당하고 해시 생성이 sha256에 비해 비교적 간단한 sha-1방식을 선택하게 되었다. 통합 데이터 아이디의 경우 아이디 생성 모듈에 통합 데이터의 데이터베이스 이름, 테이블 이름들을 오름차순으로 정렬하여 합친 문자열이 입력으로 들어가면 그에 따른 20bytes의 해시 값이 생성된다. 통합된 테이블들의 종류, 개수가 같은 경우는 항상 같은 값이 생성된다. 따라서 이 해시 값은 통합 데이터의 고유 값이라고 말할 수 있다.

Table 1. 고유값 생성 방식 비교

| | sha256 | MD5 | sha-1 | random |
|---|---|---|---|---|
| 문자열 파라미터 사용 | O | O | O | X |
| 고유성 부여 | O | △ | O | X |
| 크기 (size) | 32bytes | 16bytes | 20bytes | 16bytes |

## III. LSTM 기반 모델 생성 및 관리 기법 구현

### 1. 주요 설정 파라미터

본 플랫폼은 모델 학습 방법으로 Vanilla LSTM(RNN), Stacked LSTM, BiDirectional LSTM,CNNLSTM, ConvLSTM의 5가지 LSTM 학습 방법을 고려하였다. 본 연구는 해당5가지의 학습 방법에서 공통으로 요구되는주요 파라미터를 정의하였다. 해당 파라미터는 [표 2]을 통해 확인할 수 있다. [표2]와 같이 LSTM 모델 생성 과정 전반에는 다양한 파라미터가 필요하다. 이런 많은 파라미터 값들을 따로 정의하면 각 단계마다 필요한 파라미터 값을 정의하고 관리해야 한다. 이는 코드 상의 가독성을 떨어트리고 코드 개발의 복잡성을 증가시킨다. 본 플랫폼에서는 모든 파라미터들을 통합 관리할 수 있도록 파라미터 전용 오브젝트(Object)를 정의하였다. 이를 통해 많은 파라미터들을 통합하여 관리함으로써 코드 상의 편의를 도모하였다.

Table 2. 주요 설정 파라미터 설명

| ML 파라미터 | 설명 |
|---|---|
| target feature | 예측하려는 타켓 피처 |
| training features | 트레이닝에 활용하려는 피쳐들 |
| input width time | 예측에 활용하려는 과거의 시간 구간 |
| output width time | 예측하려는 미래의 step 시간 구간 |
| reference frequency | 다수의 데이터를 결합하여 재생성 하는 기준이 되는 frequency |
| learning method | LSTM 모델 종류 (ex. vanilla LSTM) |
| Train/Validation split ratio | train과 validation에 쓰일 데이터를 나누는 비율 |
| output preparation method | 미래의 시간 구간 데이터의 처리 방법으로 (step, mean, min, max 중 택 1) |
| scaler method | scaler 종류 선택 (robust scaler, robust scaler and log, log 중 택 1) |

## 2. 모델 생성 플랫폼 구현

본 플랫폼에서 생성된 모델에서 새로운 예측 값을 추론할 때의 인풋은 앞선 훈련 과정에서의 통합 데이터 인풋과 조건이 같아야 한다. 또한 사용자가 생각하는 예측 목표 피처 및 훈련 피처(training features)들이 적절히 대응된 모델을 검색할 수 있어야한다. 훈련과 예측 과정에의 적절한 모델 대응을 위해 본 플랫폼은 모델 과정에서 생성되는 파일을 [그림 1]과 같은 구조로 저장한다. 본 구조는 각 모델을 잘 표현할 수 있는 주요 파라미터 순으로 나누어 졌다.

본 과정에서 데이터 전처리에 필요한 스케일러(scaler) 파일의 경우 같은 통합 데이터라고 한다면 같은 스케일러 형식을 공유하기 때문에 통합 데이터 아이디로만 구분 지어 저장하면 된다. 그러나, 모델 파일의 경우에는 더 많은 구분 값(즉, 파라미터)들이 존재하기 때문에 더 깊게 나뉜다. 상위에는 스케일러와 동일하게 통합 데이터 아이디로 구분한다. 그 후에는 해당 모델이 궁극적으로 예측하고자 하는 피처 즉, 목표 피처로 구분이 된다. 그 후에는Stacked LSTM과 같은 모델형식(model method)이름으로 하위 구분이 된다. 마지막으로는 같은 모델 형식이라도 다양한 파라미터들이 존재한다. 따라서, 모델 파라미터 아이디로 한번 더 구분한다.

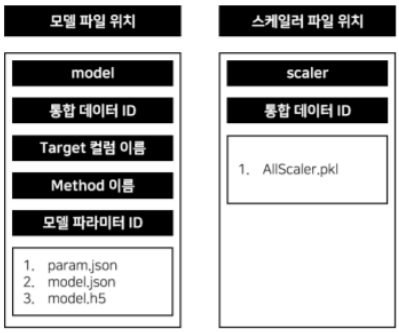


Fig. 1. 파일 저장 위치 구조도

## 3. 모델 예측 플랫폼 구현

모델 추론과정은 [그림 2]와 같이 데이터 전처리 모듈에 테스트 데이터 정보를 입력하고 이에 기반하여 테스트 데이터가 준비되면 인퍼런스(inference)모듈은 인퍼런스 모델을 활용하여 결과를 도출한다. 이 과정에서 같은 스케일링, 전처리 방법을 거쳐 예측 값을 도출하고 이 값을 다시 원 값에 준하도록 역 스케일링 등을 진행하여야 현재 상태에 준하는 값을 예측할 수 있다. 본 플랫폼은 해시 함수 기반 모델 관리 기법을 통해 해당 과정을 효율적으로 구현하였다.

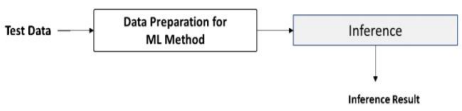


Fig. 2. 모델 추론 과정

[그림 3]을 통해, 본 플랫폼의 모델 예측 서비스 흐름을 알 수 있다. 첫번째로, 사용자가 통합 데이터와 함께 서버에 해당 페이지를 요청하면, 서버는 통합 데이터를 이용해 통합 데이터 아이디를 생성한다. 이 통합 데이터 아이디는 통합 데이터를 sha-1 방식으로 인코딩(encoding)한 문자열이다. 서버는 이 아이디와 함께 기존에 생성된 모델 파일에 대한 전체 폴더 구조를 json화 하여 제공한다. 사용자의모니터에는 사용자가 미리 선택한 통합 데이터에 대응되는 기존 모델들의 검색 페이지가 출력된다. 이 페이지에서 사용자는target feature, model method, 모델파라미터 아이디 순으로 조건을 선택하고 그에 대한 모델 파라미터 정보를 요청한다. 사용자는 서버 인퍼런스 엔진에 동작 요청과 함께 해당 모델 정보를 보내주면, 그에 맞는 예측 결과를 받을 수 있다.

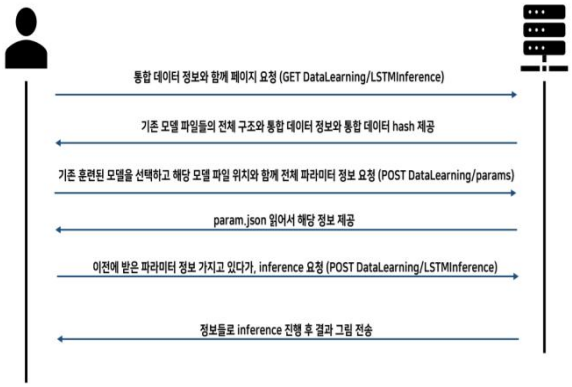


Fig. 3. LSTM 모델 예측 플로우

## IV. 결과

구현된 모델 생성 과정에 대한UI는 [그림 4]와 같다. 사용자는 통합 데이터 정보를Integration Info 칸에서 확인하고 확인 버튼을 눌러 통합 데이터 생성을 진행한다. 후에, Learning Parameter 칸에서 모델 생성을 위한 파라미터 값들을 입력한다. 마지막으로 모델 훈련 시작 버튼을 클릭하면 서버에서 모델 훈련이 시작되고, 모델훈련이 완료되면 모델 저장 위치와 훈련RMSE 그래프가 출력된다. 이를 통해 사용자는 필수 파라미터 값만 입력하고, 다른 과정에 대한 관여 없이 원하는 모델을 훈련시키고 생성할 수 있다.

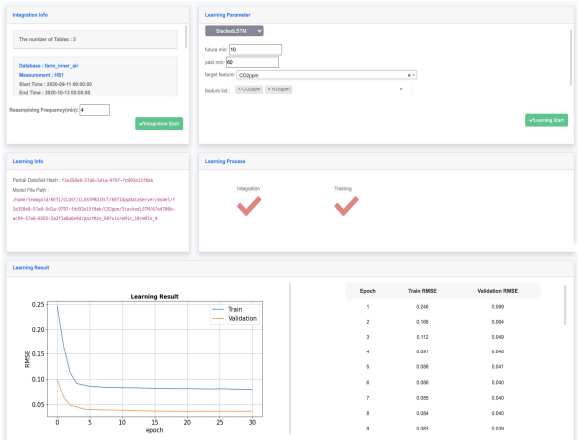


Fig. 4. 모델 생성 UI

## Ⅴ. 결론

본 논문은 파편화된 시계열 데이터를 통합하여 모델 생성/추론 과정에 필요한 해시 함수를 이용한 모델 관리 기법을 설계하였다. 또한 이 기법에 기반한 모델 생성/추론 플랫폼을 구현하였다. 이 과정에서 다양한 LSTM 모델에서 공통적으로 추출되는 통합 파라미터를 정의하였다. 또한 모델 생성/추론 플랫폼을 구현하여 제안된 해시 기법을 적용한 모델 파일 저장 방식으로 사용자가 손 쉽게 모델을 검색하고 활용할 수 있음을 검증하였다. 본 연구에서 제안하는 방법은 비전문가들도 다양한 도메인에 파편화된 시계열 데이터를 통합하고 이를 이용해 딥러닝 모델을 쉽게 생성/응용할 수 있도록 한다.

## ACKNOWLEDGEMENT

이 논문은 2022년도 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] Ismail Fawaz, H., Forestier, G., Weber, J. etal. Deep learning for time series classification: a review. Data Min Knowl Disc33, 917–963 2019.

[2] “Time Series Insights”, Microsoft Azure, https://azure.microsoft.com/ko-kr/services/time-series-insights/

[3] Sung-Min Hwang, Seog-Gyu Kim, “Design of System for Avoiding upload of Identical-file using SA Hash Algorithm”, Journal of the Korea Society of Computer and Information, 19(10), 81-89, 2014.