네트워크 지능화를 위한 인공지능 해커톤 코드 설명자료

- 분야 #2: 미디어 서비스-

대회 참가자 (장종환)

2022년 05월 15일

1. 라이브러리 및 데이터 (Library & Data)

- Pandas를 사용해 제공받은 csv형태의 데이터들을 불러와 데이터프레임 생성
- Numpy를 사용해 다차원 배열 계산
- Matplotlib, seaborn 사용해 시각화
- Torch 딥러닝 프레임워크를 사용해 Anoamly detection 모델 구현

2. 데이터 전처리 (Data Cleansing & Pre-Processing)

- 4종류(INFO, LOGIN, MENU, STREAM)의 데이터프레임을 하나로 합치고, 컬럼들을 서버별로 통합, 최종적으로 13종의 컬럼 생성

ex) login_svr1 = LOGIN-01-Request + LOGIN-01-Success + LOGIN-01-Fail

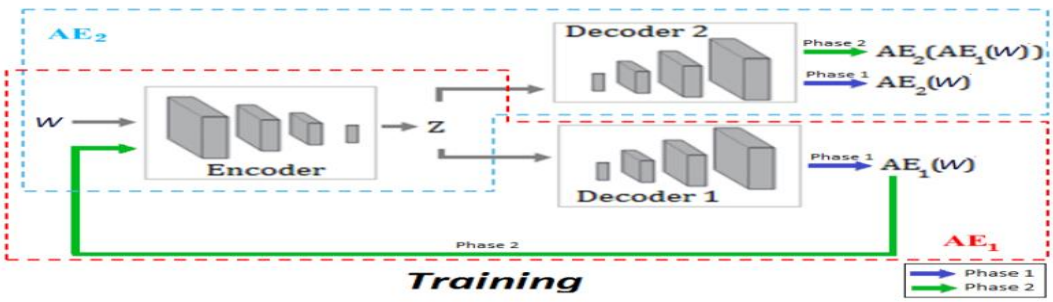
- train(2017 1.1 - 12.31), test(2018 1.1 - 12.31) 데이터셋 분리
- 각 컬럼이 가지는 값의 범위가 크게 다르므로 정규화 필요
- normalize 함수를 만들어 정규화. 정규화 방법은 train 데이터셋의 최솟값, 최댓값을 이용하여 0~1의 범위에 들어오도록 하는 것(test 셋은 1초과한 값이 있음)
- 정규화 후 ewm(지수가중함수)을 적용하여 시계열 데이터 특성인 무작위적인 변화로 생기는 효과를 줄이고자 함(data smoothing)
- window_size = 12로 설정하여 12개의 데이터(1시간 데이터)를 묶어서 예측하도록 재구성
- 최종 train, test shape => (105120, 12, 13)

3. 변수 선택 및 모델 구축 (Feature Engineering & Initial Modeling)

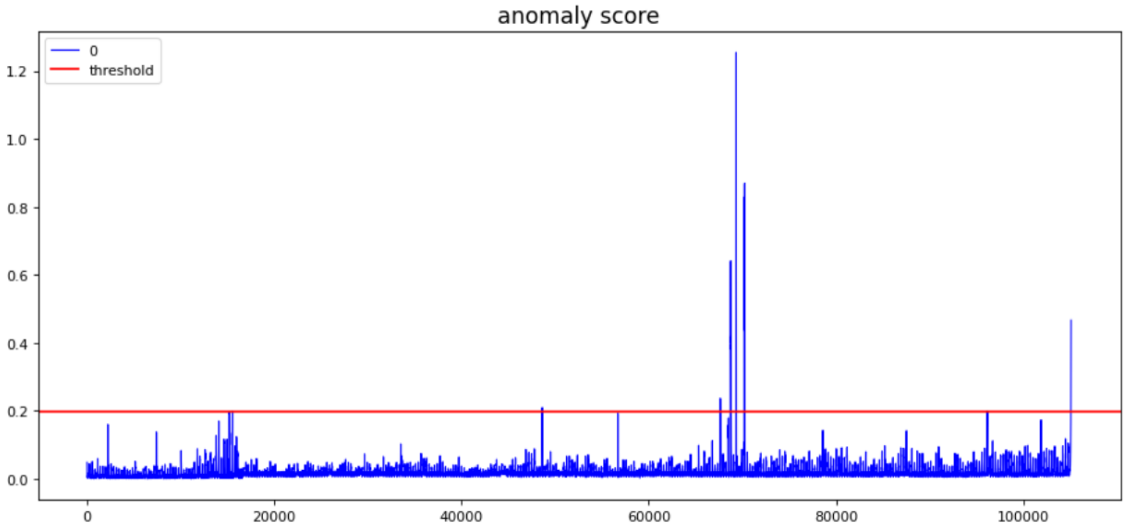
- 모델의 Task : 길이가 12인 time window $W_t = \{x_{t-11}, \dots, x_{t-1}, x_t\}$ 를 input으로 t시점의 normal/abnormal 여부를 예측
- 딥러닝 학습과 추론에는 Pytorch를 사용
- 사용 알고리즘은 USAD(Unsupervised anomaly detection on multivariate time series)
- USAD는 학습이 쉽고 안정적인 결과를 낼 수 있는 Auto encoder의 장점과 abnormal information을 강제할 수 있는 GAN의 장점을 결합한 모델
- USAD는 encoder, decoder d1, decoder d2를 기반으로 하나의 encoder를 공유하는 두 개의 autoencoder로 구성

4. 모델 학습 및 검증 (Model Tuning & Evaluation)

- USAD는 두 단계를 거쳐서 학습이 진행
  - Phase 1 : encode가 input W_t를 latent space Z로 압축한 후, decoder가 Z를 W_t와 동일하게 복원하는 Autoencoder를 학습
  - Phase 2 : 적대적 학습을 위해 각각의 AE는 다음의 역할 수행
    - AE1 : input을 잘 복원하면서 AE2를 잘 속이는 모델 학습
    - AE2 : input을 잘 복원하면서 AE1 이 복원한 데이터와 input을 잘 구별하는 모델 학습
  - AE2 도입 및 적대적 학습을 통해 정상 데이터와 유사한 이상치를 탐지하는 것을 가능하게 하며, AE 구조를 통해 안정적인 학습이 가능



- 학습이 완료된 두 AE를 기반으로 output으로 anomaly score를 산출
  - $Anomaly\ Score = \alpha||W_t - AE_1(W_t)||_2 + \beta||W_t - AE_2(AE_1(W_t))||_2$
- 이상을 식별하기 위한 적절한 임계값(threshold)을 결정하기 위해 anomaly score의 분포를 확인한 후, $threshold = Q3 + 5*(Q3 - Q1)$를 초기 임계값으로 설정하고 여러 번의 실험을 통해 조정한 결과 최적의 임계값으로 0.19 선택. Test 데이터셋을 통해서 anomaly score를 계산했을 때 해당 임계값을 넘을 경우 이상으로 판별



- 최종적으로 158개의 이상 탐지, F2_Score 0.7234 기록

5. 결과 및 결언 (Conclusion & Discussion)

- USAD 모델은 noise가 작은 경우 reconstruction 후 비정상 데이터가 잘 탐지가 안될 수 있다는 auto encoder의 문제점을 두개의 decoder을 두어 noise를 확대시킴으로서 해결한 모델. 이번 데이터셋 처럼 train data와 test data의 차이가 크지 않은 형태일 때, 정상과 비정상 데이터가 굉장히 유사한 형태를 가질 때 효과적
- 결측치를 단순하게 0으로 대체하는 방법을 사용했는데, 적절하지 않다고 생각함. 평균 대치나 회귀분석을 통해 결측치를 예측하는 등 다른 방식으로 적용해볼 필요가 있음
- 추후에 test데이터들의 label이 주어진다면 실험을 통해, 여러 파라미터들을 최적화하여 모델의 성능을 보완할 수 있음.

■ 최적화 해야 할 파라미터들

- Windows size : input의 sequence 길이
- z_size : AE의 특징 벡터 z 차원
- α, β : anomaly score의 계수($\alpha + \beta = 1$)