스마트팩토리융합캡스톤디자인2

개인 보고서

주제: CNC 툴 마모 인식과 모니터링을 위한

MFCC 데이터 기반 앙상블 모델 개발-



Student ID: 2019314991

이름: 서재원

전공: 기계공학부

날짜 : 2023. 12. 14

**목차**

**1. 선행연구 자료조사3**

1.1. RUL 예측 (Remaining Useful-Life Prediction)3

1.2. 이상탐지(Anomaly Detection)5

1.3. XAI 및 Knowledge distillation6

**2. 데이터 전처리 및 EDA9**

2.0 데이터 구성 분석 및 압축9

2.1. 기본 plotting9

2.2. 이상 영역 가정10

2.3. 가설 검정10  
**3. 모델 및 알고리즘 개발16**

3.1. 데이터 라벨링16

3.2. 변수 선택 및 Scaling 및 PCA 진행16

3.3. 모델 선택 17

3.4. 데이터 불균형 문제를 위한 Sampling기법 적용17

3.5. 앙상블18

**4. 최종 결론19**

4.1. 최종 모델 성능 평가19

4.2. 고찰19

4.3. Future works19

**1. 선행연구 자료조사**

**1.1. RUL 예측 (Remaining Useful-Life Prediction)**

RUL은 기계의 수리 또는 교체가 필요하게 되기 전까지 남아 있는 사용 시간 또는 예상 수명을 예측하는 알고리즘, 시스템으로 이 시스템을 활용하여 교체시기를 예측할 수 있음. 수집 데이터가 CNC 진동, 전류, 전압 및 제조 데이터이므로 CNC 오픈데이터셋 기반으로 작성된 RUL 예측 관련 논문을 탐색함.

한국 논문으로는 다음과 같이 조사하였다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

한국 논문 중 많은 논문들이 기본 머신러닝을 통하여 공구 수명을 예측하였다. 첫번째 논문은 머신러닝의 기초인 선형회귀방법을 사용하여 예측한 논문이었다. 그리고 데이터셋의 크기 등을 고려하여 전이학습한 모델을 탐색하였다. 여러가지 논문이 있었고 그중 두번째 논문이 Pre-trained model을 구축한 논문으로 transfer learning의 가능성을 확인하였다.

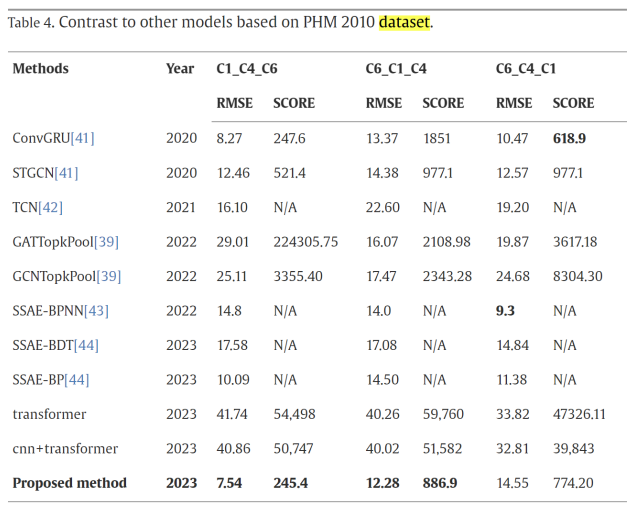
외국 논문 탐색은 RUL 관련 Review 논문 및 최신 논문을 중점적으로 탐색하였다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

또한 오픈 데이터셋 중 PHM 2010이 수집데이터와 유사하여 이를 기반으로 모델을 구축한 모델을 탐색하였다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명 

위 표는 PHM 2010 dataset으로 구축된 기존 모델들의 목록과 논문에 기재된 논문의 방법론을 비교한 것으로 graph attention network와 deep adaptive transformer를 사용하여 성능 높임을 확인함.

대부분 RUL 예측 논문에서는 BI-LSTM까지 사용한 것이 대부분, ANN, CNN+LSTM으로 모델 구축하는 경우가 대다수이다. 최신 논문 중에는 transformer 기반 모델이 있었다. 하지만 transformer 같은 경우 아래 논문과 같이 예측관련해서는 성능이 좋지 않다고 알려져 있다.

**1.2. 이상탐지(Anomaly Detection)**

RUL은 이상탐지기반 고장진단 결과를 기반하는 경우가 많음. 이에 이상탐지 관련 알고리즘 연구가 필요, 자료 조사를 진행함. 여러가지 모델이 있으나 아래의 모델을 선정하여 자세히 탐구하였다.

DAMP(Discord Aware Matrix Profile)텍스트, 폰트, 화이트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

위 논문은 DAMP의 기반이 된 논문으로 기존 모델들의 문제점을 지적하였다. 1. 라벨링 문제, 2. 알고리즘 문제 등 다양한 문제들을 지적하면서 통계기반 모델을 제안함. 이는 후에 Rule-base 모델 제작시 필요할 것으로 판단, 알고리즘을 분석하였다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 폰트, 문서이(가) 표시된 사진

자동 생성된 설명

위는 DAMP 사이트 및 관련 코드 중 일부이다. 여기서 기준점에 대하여 앞뒤로 보면서 통계치 및 이상치를 판단함을 파악함. 더욱 자세히 코드 및 알고리즘을 파악하기 위하여 ChatGPT를 활용함. 아래는 ChatGPT를 이용하여 파악한 내용이다.

요약: DAMP는 시계열에서 디스코드를 효율적으로 식별하기 위해 설계된 알고리즘이다. 이 때. 이 알고리즘은 무차별 대입, MASS 알고리즘 및 룩어헤드 메커니즘의 조합을 사용하여 검색을 최적화함으로써 성능을 높였다. 함수는 또한 적절한 매개변수를 선택하는 데 도움이 되도록 광범위한 오류 처리와 사용자 피드백을 제공한다.

주요 용어

1) 디스코드:

큰 시계열 내에서 동일한 길이의 모든 다른 부분 시퀀스와 최대한 다른 부분 시퀀스를 의미한다. 즉, 다시 말해, 나머지 데이터와는 맞지 않는 패턴을 의미한다.

2) MASS(Mueen’s Algorithm for Similarity Search) 알고리즘:

시계열 데이터 내에서 부분 시퀀스와 주어진 쿼리 사이의 유사도를 빠르게 계산하기 위한 알고리즘으로 FFT를 사용하여 시계열과 쿼리 사이의 거리 프로필을 효율적으로 계산함.

3) contains\_constant\_regions 함수 :

이 함수는 시계열에 일정하거나 거의 일정한 영역이 포함되어 있는지 확인하는 함수로 MASS 알고리즘을 효율적으로 계산하기 위한 도움 함수이다.

4) 룩어헤드(Lookahead) 메커니즘

알고리즘의 성능을 향상시키기 위해 미래의 데이터나 상태를 미리 검토하는 기법이다. 미 메터니즘을 통하여 불필요한 연산을 줄이고 더 효율적인 결정을 내릴 수 있도록 도와준다. 작동 과정은 미래의 데이터 검토- 프루닝으로 이뤄진다.

5) 프루닝(Pruning)

특정 부분 시퀀스가 현재까지의 디스코드 점수보다 낮은 점수를 가질 가능성이 있다면 계산을 건너뛰어 계산 효율성을 높인다.

**1.3. XAI 및 Knowledge distillation**

RUL와 이상탐지는 기존 모델이 많이 개발되어 있다. 하지만 XAI(eXplainable AI)와 모델 경량화 기법 중 하나인 Knowledge distillation기반 RUL 및 이상탐지는 많이 개발되어 있지 않다. 그리고 산업에서 경량화와 설명가능한 인공지능은 필요하기 때문에 이를 접목한 논문이 있는지 탐색하였다.

**1.3.1. XAI**

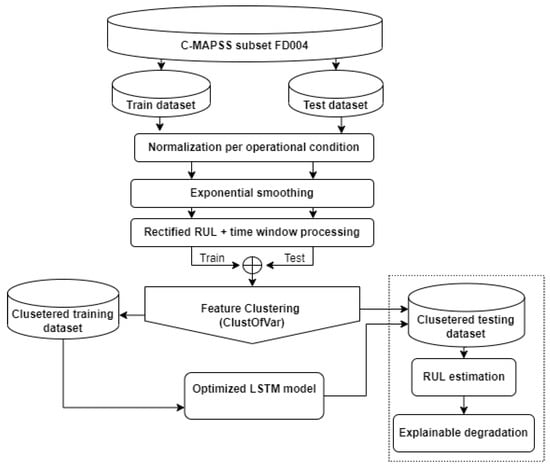
텍스트, 스크린샷, 폰트, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

위 논문은 1DCNN, BILSTM, LSTM을 결합하여 모델을 구축하였고 SHAP(Shapley Additive exPlanation)를 사용하여 RUL 예측에 크게 영향을 미친 기능을 분석할 수 있게 하여 XAI를 접목하였다.

텍스트, 스크린샷, 폰트, 라인이(가) 표시된 사진

자동 생성된 설명

위 논문은 옆의 Flowchart와 같이 Clustering과 LSTM을 활용하여 XAI이 가능하도록 하였다.

이렇게 XAI는 클러스터링과 SHAP을 이용하여 어느 부분에서 혹은 어느 변수에서 영향이 컸는지 확인해줌으로써 구현하였음을 조사했다.

**1.4.2. Knowledge distillation(지식 증류)**

딥러닝에서 지식 증류는 큰 모델(Teacher Network)로부터 증류한 지식을 작은 모델(Student Network)로 transfer하는 일련의 과정으로 큰 모델의 지식을 모두 담아 성능은 유지하고, 모델 크기는 줄여 추론 속도는 높이는 방식이다.

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

위 두 논문이 Knowledge distillation을 이용한 논문으로 각각 큰 모델과 작은 모델의 성능이 비슷했다.

**2. 데이터 전처리 및 EDA**

**2.0 데이터 구성 분석 및 압축**

데이터는 크게 진동데이터, 전류 및 전압 데이터가 있다.

1) 진동데이터

X,y,z, 각각 축에 대한 진동데이터가 있으며 raw데이터는 sampling rate이 5000Hz이다. 우리는 연산 속도 및 컴퓨터 성능을 고려하여 1Hz로 압축된 데이터를 사용하였다.

2) 전류 및 전압 데이터

전류 및 전압 데이터의 column은 Active\_Power1,2,3과 Apparent\_Power\_Va1,2,3, Line\_Current \_L12,3, 그리고 Line\_Voltage\_V12, Line\_Voltage\_V23, Line\_Voltage\_V31이렇게 있다. Sampling rate은 1Hz이다. 또한 각각 column 이름과 의미를 파악했다. Acitve\_Power는 유효 전력으로 동력을 돌리는 일(Work)을 행하는 전력을 의미한다. 즉, 툴에게 가해지는 일들을 의미한다. Apparent\_Power는 피상 전력으로 전압 및 전류의 실효값의 곱을 의미한다. Line\_Current와 Line\_Voltage는 Line에 가해지는 전류와 전압을 의미한다.

최종적으로 우리가 분석, 학습에 이용한 데이터로 위 두 데이터를 합친 csv 파일을 사용하였다.

**2.1. 기본 plotting**

친필, 텍스트, 폰트, 라인이(가) 표시된 사진

자동 생성된 설명텍스트, 폰트, 친필, 스크린샷이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 그래프, 라인이(가) 표시된 사진

자동 생성된 설명

위 3개의 그래프에서 왼쪽은 Active\_Power1,2,3와 Apparent\_Power\_Va1,2,3, 그리고 Line\_Current\_ L1,2,3이다. 이 그래프를 통하여 결측치가 조금 있으나 이는 간단하게 해결할 수 있고, 다른 범위보다 큰 부분이 있음을 확인하였고 이 부분을 유심히 볼 필요가 있다고 판단했다. 또한 오른쪽 위 그래프는 Line\_Voltage그래프로 많은 부분에서 결측치가 있음을 확인하였다. 마지막으로 오른쪽 아래 그래프는 1hz으로 압축된 진동데이터이다.

**2.2. 이상 영역 가정**

이상 영역(고장 영역)을 설정하기 위하여 여러가지 가설을 설정하였다. 그 중 툴 교체 시점 전후 일정시간대의 집단끼리(교체전 t부터 교체직전까지, 교체직후 t까지) 유의미한 차이가 있을 것이라고 가정하여 아래 방법론을 통하여 분석하였다.

**2.3. 가설 검정**

**2.3.0. 두 집단이 유의미하게 다르다는 것을 증명하기 위한 추가 가설 설정**

이상영역을 교체하기 직전의 2분으로 정의하고 5개의 집단을 정의하여 비교, 분석한다.

stat\_list\_front = [pd.DataFrame() for i in range(len(selected\_columns))]

stat\_list\_second = [pd.DataFrame() for i in range(len(selected\_columns))]

stat\_list\_third = [pd.DataFrame() for i in range(len(selected\_columns))]

stat\_list\_back = [pd.DataFrame() for i in range(len(selected\_columns))]

stat\_list\_back\_second = [pd.DataFrame() for i in range(len(selected\_columns))]

이렇게 5개의 집단을 설정하였는데 front는 교체 1분후부터 120초, second은 다음 120초, third은 second직후 120초, back은 교체 직전 1분부터 120초 구간, back\_second은 back 직전부터 120초 구간을 모집단으로 한다. 그리고 교체 시간대는 교체주기 데이터를 기반으로 Tool\_number 62는 28개, Tool\_66은 31개의 모집단을 생성하였다.

**2.3.1. 통계량 기반 특징 추출 및 분석**

통계량 추출이란 두 집단에 대하여 기존의 데이터를 분석할 때 raw 데이터로 분석하는 것이 아닌 통계량을 통하여 통계분석을 하기 위하여 통계량을 추출하여 분석하는 것이다. 특히 mean, max, IQR 등 기본적인 통계량뿐만 아니라 시계열 데이터의 특징을 고려하기 위하여 모집단의 그래프 개형과 관련된 통계량을 추가로 추출하여 분석을 진행하였다. 아래 코드는 통계량을 추출하는 코드로, 기본 통계량(Max, Min, Mean, Rms, Std, q1, q3, IQR)와 그래프 개형과 관련된 통계량(Skewness, Kurtosis, Crest factor, Impulse factor, Shape factor) 총 15개의 통계량을 추출한다. 이를 통하여 앞서 말한 가설을 검정하였다.

오른쪽**텍스트, 소프트웨어, 멀티미디어 소프트웨어, 멀티미디어이(가) 표시된 사진

자동 생성된 설명**텍스트, 스크린샷, 소프트웨어이(가) 표시된 사진

자동 생성된 설명 코드는 두 모집단에 대하여 Column별 p-value를 추출하는 코드이다. 여기서 나오는 P-value값이 일정 수준(0.05)보다 작으면 두 집단에 대하여 특정 통계량을 변수를 기준으로 유의미하게 다름을 보여준다. 즉, 두 집단이 다름을 간접적으로 보여준다. 이를 비교한 결과는 아래와 같다.

텍스트, 스크린샷, 폰트, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

위 3가지 결과는 위부터 second와 third, first와 back, back와 back\_second 이렇게 비교하여 p\_value를 추출한 결과이다. 이를 0.05를 기준으로 작으면 diff, 크면 Nope으로 나오게 하여 다시 비교한 결과는 아래와 같다.

텍스트, 스크린샷, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

여기서 대부분 NOPE으로 유의미하게 다르다고 한부분이 없다. 하지만, z축 진동데이터에서 보면 frontback에서 Kurt가 diff이고 나머지는 Nope을 확인하였다. 즉, z축 진동데이터가 교체 전후에서 그래프 개형이 다르고 나머지 부분에서는 비슷하기에 이를 기준으로 이상영역을 정의할 수 있음을 확인하였다.

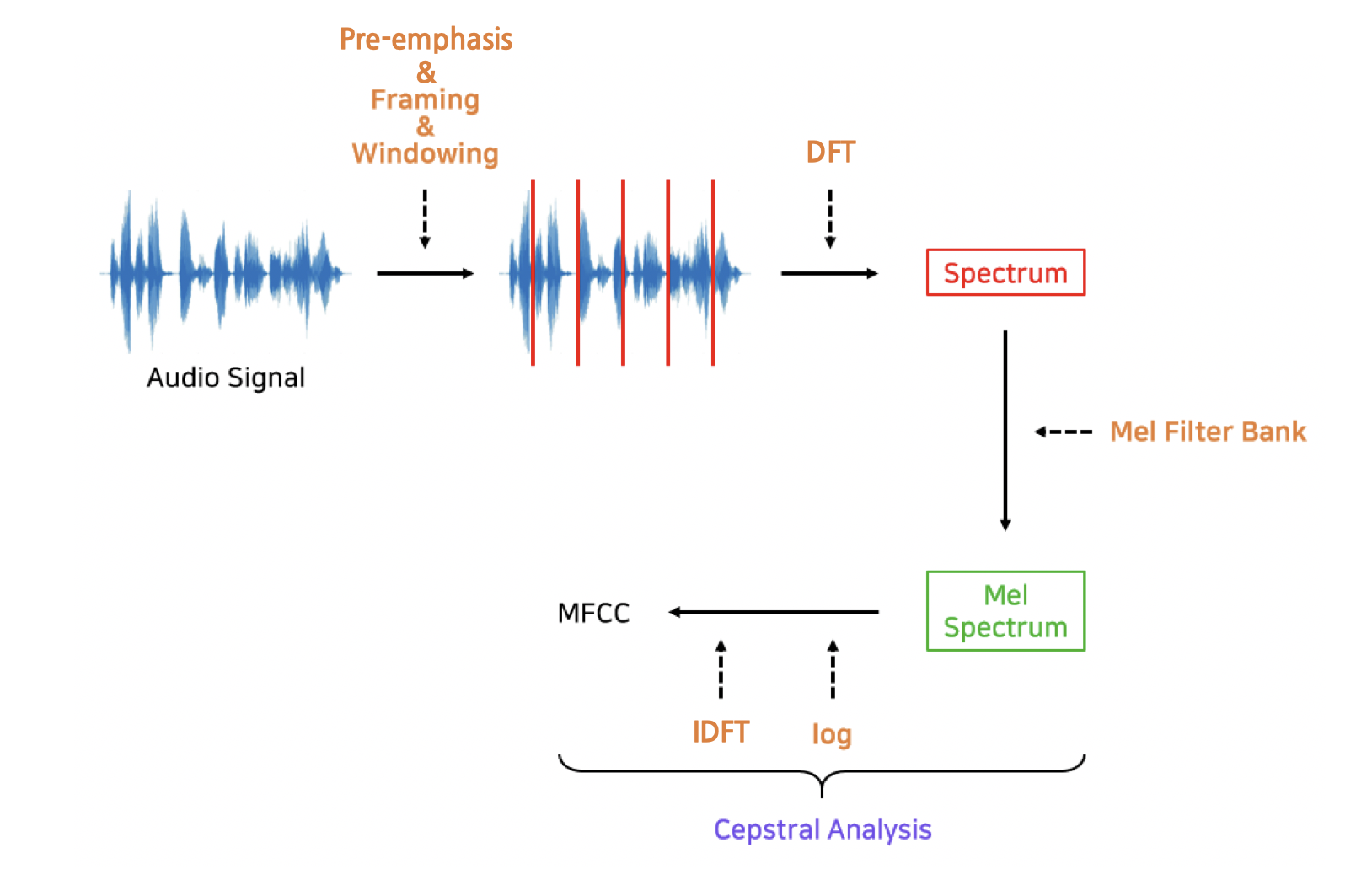
이 과정을 66에 대해서도 반복진행한 결과는 아래와 같다.

텍스트, 스크린샷, 소프트웨어, 디자인이(가) 표시된 사진

자동 생성된 설명

여기서도 대부분 NOPE으로 유의미하게 다르다고 한부분이 없다. 하지만, z축 진동데이터에서 보면 frontback에서 Kurt가 diff이고 나머지는 Nope을 확인하였다. 즉, z축 진동데이터가 교체 전후에서 그래프 개형이 다르고 나머지 부분에서는 비슷하기에 이를 기준으로 이상영역을 정의할 수 있음을 확인하였다.

**2.3.2. MFCC 평균값 기반 분석**



출처: <https://ratsgo.github.io/speechbook/docs/fe/mfcc>

MFCC란 Mel-Frequency Cepstral Coefficients의 약자로 위의 과정을 거쳐 Mel-Spectrum에서 나온 Cepstral Coefficients들을 의미한다. 즉, 이 MFCC를 통하여 진동데이터를 음성적인 관점에서 특징을 추출하여 이미지처럼 할 수 있다.  MFCCs의 shape은 num\_frames × nfilter으로 되어 있다. 그러나 이를 각각 t-test하는 것도 좋으나 Raw진동데이터에서 1초마다 MFCC의 평균들을 대표값으로 보고 평균을 내고 1초마다 20개(n\_coefficient)의 값을 나오게 하고 이를 x,y,z축마다 진행하였다. 먼저 Raw데이터로 1초마다 MFCC의 평균값들을 뽑았다. 코드는 아래와 같다.

텍스트, 전자제품, 스크린샷, 소프트웨어이(가) 표시된 사진

자동 생성된 설명텍스트, 스크린샷, 소프트웨어, 디스플레이이(가) 표시된 사진

자동 생성된 설명텍스트, 스크린샷, 디스플레이, 소프트웨어이(가) 표시된 사진

자동 생성된 설명텍스트, 스크린샷, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

바로 위에 있는 3개의 코드 중 왼쪽 코드는 위에서 뽑은 코드를 저장하는 코드로 x,y,z별로 MFCC의 평균값을 1초마다 뽑을 수 있다. 오른쪽 코드는 위의 5개 집단을 뽑는 코드와 이를 토대로 x,y,z별 t-test를 진행하는 코드이다.

이 데이터를 기반으로 front-last(교체 전후), front\_second(교체 직후), second\_third(교체직후 120초 후의 두집단), back\_two(교체 직전 두집단)에 대하여 t-test를 진행하였다. 62에 대한 결과는 아래와 같다.

텍스트, 스크린샷, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

이를 위의 과정과 비슷하게 특정 값(0.05)보다 낮으면 파란색, 높으면 빨간색으로 시각화 하였다.

스크린샷, 다채로움, 직사각형, 라인이(가) 표시된 사진

자동 생성된 설명

위 그림에서 만약에 MFCC에 대하여 맨 위는 파란색, 아래는 빨간색인 경우 MFCC로 파생된 변수에 대하여 교체 전후는 유의미하게 다른 분포를 가지고 나머지는 비슷하기에 파생변수로 선택할 수 있다. 62에 대해서는 X축의 6와 Z축의 6에 대하여 적용할 수 있다고 판단하였다.

이와 같은 과정을 66에 대해서도 적용한 결과는 아래와 같다.

스크린샷, 다채로움, 라인, 직사각형이(가) 표시된 사진

자동 생성된 설명

66의 경우에는 X의 경우에는 2,3,7,8,14에서 적용할 수 있고 Y의 경우는 2,3,14, Z의 경우는 0,17에서 사용할 수 있음을 확인하였다.

**3. 모델 및 알고리즘 개발**

**3.1. 데이터 라벨링**

Tool\_66 데이터에 대하여 교체 직전 20분을 이상영역으로 선정하여 1로 나머지는 0으로 라벨링하여 진행하였다. 이 라벨링에 대하여 Train, test를 8:2로 분리하여 라벨링을 분석한 결과는 아래와 같이 Class imbalance문제가 있다.

원, 도표, 스크린샷이(가) 표시된 사진

자동 생성된 설명

**3.2. 변수 선택 및 Scaling 및 PCA 진행**

이번 실험에서는 Tool\_66에 대한 데이터만 사용하기로 하였다. 이에 변수는 MFCC에서 X와 Y의 2,3,14, Z의 0,17 총 8개 변수를 추출하여 사용하였다.

또한 MinMaxScaling을 적용하여 8개의 변수별 중요도를 같게 하였다. 마지막으로 8개의 성분을 바로 사용하지 않고 PCA를 진행하여 차원 축소를 진행, 차원의 저주를 방지하였다. 위 과정을 적용한 결과는 다음과 같다.

텍스트, 스크린샷, 디스플레이, 폰트이(가) 표시된 사진

자동 생성된 설명 텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

**3.3. 모델 선택**

모델은 이상을 1로 정상을 0으로 예측하게 하도록 하고 시계열 데이터 특성을 고려하지 않고 각각 초당 이상과 정상을 예측하도록 하였다. 총 5가지 모델을 선정하였다. 이중 3가지는 이상탐지에 특화되어 있는 Isolation Forest, LOF(Local Outlier Factor), EllipticEnvelope 모델과 LightGBM모델 2개를 선정하였다.

여기서 이상탐지에 특화되어 있는 3가지의 모델은 비지도학습으로 라벨링을 주어지지 않고 이상 비율만 알려준 뒤 학습을 하였다. 또한 LightGBM모델의 경우는 지도학습이지만 Class Imbalance 문제를 초래할 수 있기 때문에 UnderSampling과 OverSampling으로 해결하고자 했다.

**3.4. 데이터 불균형 문제를 위한 Sampling기법 적용**

LightGBM으로 이진분류할 시 Class Imbalance문제가 발생하기 때문에 UnderSampling과 Over Sampling을 통하여 클래스별 데이터 수를 맞추어야 한다. 이에 UnderSampling은 Random UnderSampler를 사용하였다. OverSampling은 SMOTE, ADASYN 둘다 적용, 비교하여 더 성능이 좋은 알고리즘인 ADASYN을 사용하였다.

도표, 원, 스크린샷, 텍스트이(가) 표시된 사진

자동 생성된 설명

도표, 스크린샷, 원, 텍스트이(가) 표시된 사진

자동 생성된 설명

위의 두 그래프와 같이 결과적으로 UnderSampling한 결과 데이터 총수는 53770이고 이중 50%는 정상, 나머지는 이상인 데이터로 구성하도록 하였다. OverSampling의 데이터 총수는 634598이고 이중 50.4%는 정상데이터, 49.6%는 이상데이터로 구성하도록 하였다.

**3.5. 앙상블**

위 5가지 모델에 관하여 각각 성능이 다르고 해결하고자 하는 방향성도 다르기 때문에 5가지 모델을 Voting을 기반으로 하여 아래 코드와 같이 앙상블하였다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

여기서 lof, isforest, ee는 각각 이상탐지를 위한 모델인 LOF(Local Outlier Factor), Isolation Forest, EllipticEnvelope 모델이고, lgb\_u은 undersampling한 데이터를 가지고 학습시킨 LightGBM모델, lgb\_o는 oversampling한 데이터를 가지고 학습시킨 LightGBM모델이다.

Predict\_1,2,3으로 이상탐지를 진행한후 평균을 내고 가중치를 곱한 값과 Predict\_4,5으로 이진분류를 진행한후의 평균에 가중치를 곱하여 최종적으로 이상인 확률로 predict를 도출한다. 그리고 이 predict에 threshold를 적용하여 최종적인 이상탐지 예측값을 도출한다.

**4. 최종 결론**

**4.1. 최종 모델 성능 평가**

**최종 모델 성능은 아래와 같다.**

텍스트, 스크린샷, 도표, 직사각형이(가) 표시된 사진

자동 생성된 설명 **텍스트, 폰트, 스크린샷, 타이포그래피이(가) 표시된 사진

자동 생성된 설명**

**4.2. 고찰**

성능이 엄청 좋지 못한 이유에는 다음이 있다.

1) 데이터 라벨링 문제: 이상영역을 교체 전 20분으로 설정하였는데 이부분이 전부다 이상이진 않을 것이다. 이는 라벨링 문제를 초래하며 성능에 악영향을 준다.

2) 모델의 앙상블 문제: 이진분류와 이상탐지를 동시에 적용하였는데 둘이 서로 좋은 영향을 주지 않고 안 좋은 쪽으로 영향을 주면 문제가 발생할 수 있다.

**4.3. Future works**

1) 앞으로 성능을 더욱 높이기 위하여 딥러닝 모델을 연구, 도입할 예정

2) MFCC 데이터를 평균이 아닌 이미지나 Time Series 데이터로 전처리하여 딥러닝에 적용