산업빅데이터분석실제
프로젝트 최종결과 발표

2021. 12. 13

2020254011
윤 재 웅

데이터 분석 목적

분석 필요성

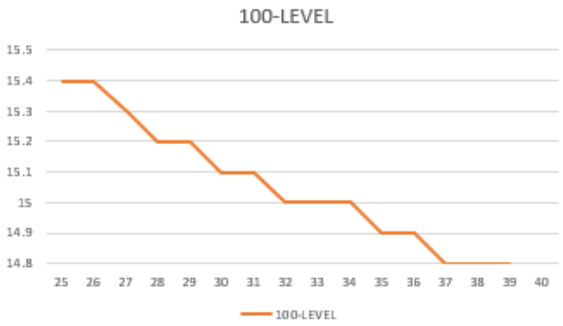
- *주력 장비의 성능을 끌어올리기 위한 연구 필요*
- *노광장비의 공정 수율을 올리기 위함*

- *머신러닝 기술을 접목하여 빠르고 정확한 분석을 위해 적용*

분석 대상의 문제점 및 필요성

- *노광 컨트롤러의 광원은 온도가 높아지면 출력이 감소하는 특성을 가짐*
- *변수에 따른 출력감소를 막으려면 전류의 적정량을 판단하고 공급해야 함*
- *이를 위해선 사전 분석이 필요*

노광용 컨트롤러



온도별 출력 변화 (무보상)

데이터 종류

데이터 종류

- *이용하고자 하는 데이터는 온도 및 광출력*
- *데이터를 수집하여 알고리즘을 이용한 수식으로 재탄생*
- *광원의 온도에 상관 없이 일정한 광출력 레벨 시각화 (오차율 10% 내외)*
- *광레벨(level), 온도(temp)별 광출력(power) 데이터 확보*
- *Train 210개 / Test 90개 = 70 : 30 의 비율*

Train 210개

unnamed	level	temp	power
1	100	25	15.4
2	100	26	15.4
3	100	27	15.3
4	100	33	15.0
5	100	34	15.0
⋮			
207	80	46	12.9
208	80	47	12.9
209	80	48	12.8
210	80	49	12.8

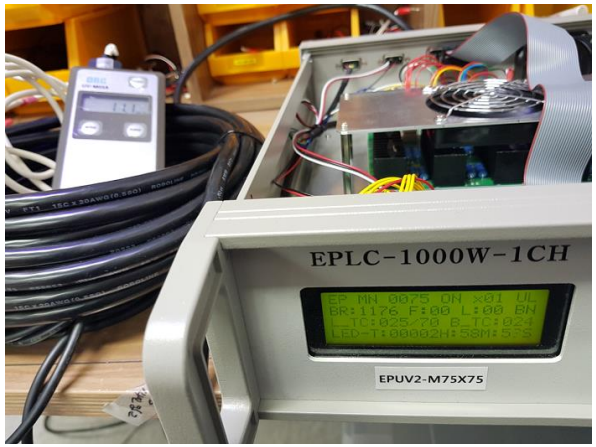
Test 90개

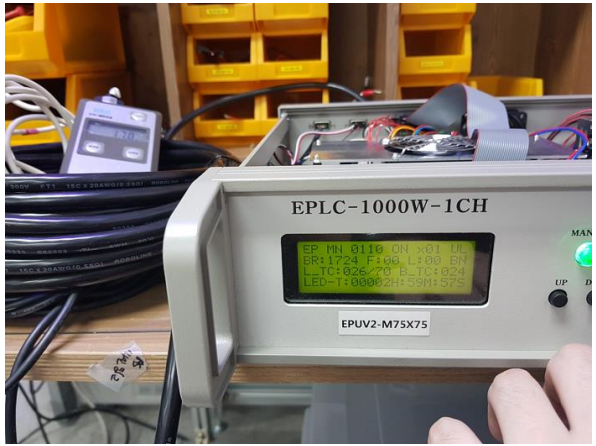
unnamed	level	temp	power
1	100	25	15.5
2	100	26	15.4
3	100	27	15.3
4	100	49	14.4
5	120	25	18.5
⋮			
87	80	45	12.9
88	80	46	12.9
89	80	47	12.9
90	80	48	12.8

데이터 종류

테스트 데이터 확보

- *딥러닝을 적용하기 위해선 다양한 데이터가 필요*
- *기존 장비를 이용하여 raw data를 확보하고 이를 통한 분석*
- *장비 작동에 핵심이 되는 부분 우선 연구 (온도값 및 광출력 수치)*
- *센싱 데이터는 온도만을 사용(온도의 영향이 90%, 타변수 통제)*





데이터 종류

실험 방법

- *사내 장비를 이용한 테스트 (UIT-250, EPLC-1000W, 50X50 노광광원, 전용ZIG 사용)*
- *UIT-250을 이용한 395nm 파장대 UV광출력 측정*
- *광원 ON 후 1초 안정상태 가진 다음의 출력을 측정*
- *온도, ADC, 컨트롤러 출력(A) 등의 영향에 의한 데이터를 표 및 그래프로 작성*

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 210 entries, 0 to 209
Data columns (total 4 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   unnamed  210 non-null    int64
 1   level    210 non-null    int64
 2   temp     210 non-null    int64
 3   power    210 non-null    float64
dtypes: float64(1), int64(3)
memory usage: 6.7 KB
column:    unnamed   Percent of NaN value: 0.00%
column:      level   Percent of NaN value: 0.00%
column:       temp   Percent of NaN value: 0.00%
column:      power   Percent of NaN value: 0.00%
```

데이터 탐색적 분석 결과

탐색적 분석 결과

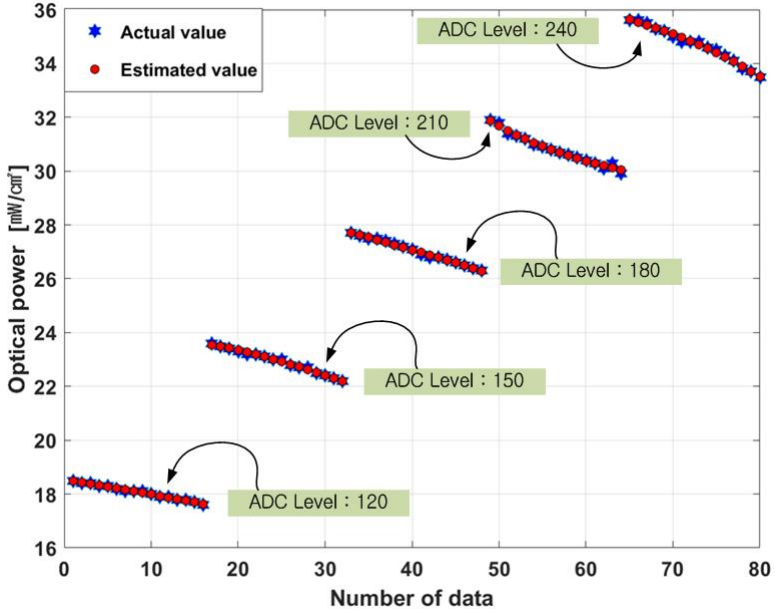
- *온도라는 조건에 따른 ADC, 전류 값의 변화 이해*
- *알고리즘에 의한 분석 후 광원의 출력 변화 이해*
- *Column의 개수가 적어서인지 탐색적 데이터 분석 코드 제작에 어려움*

```
"""

산업 빅데이터 분석 - 탐색적 데이터 분석
2020254011 윤재동

"""

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
plt.style.use('fivethirtyeight')
import warnings
warnings.filterwarnings('ignore')
%matplotlib inline

train_data = pd.read_excel('train.xlsx')
test_data = pd.read_excel('test.xlsx')

"level:광출력레벨 / temp:광원온도 / power:광출력"


"train_data.head()"
train_data.info()
"train_data.describe()"

for col in train_data.columns:
    msg = 'column: {:>10}\t Percent of NaN value: {:.2f}%'.format(col,
            100 * (train_data[col].isnull().sum() / train_data[col].shape[0]))
    print(msg)

"msno.matrix(df=data.iloc[:, :], gifsize=(8, 8), color=(0.8, 0.5, 0.2))"

"f, ax = plt.subplot(1, 2, figsize=(18, 8))"

"data['power'].value_counts().plot.pie(explode=[0, 0.1], autopct='%1.1f%%', ax=ax[0], shadow=True)"

"ax[0].set_title('Pie plot - power')"
"ax[0].set_ylabel('')"
"sns.countplot('power', data=data, ax=ax[1])"
"ax[1].set_title('Count plot - power')"
"plt.show()"
"train_data[['level', 'temp']].groupby(['level'], as_index=True).count()"
"pd.crosstab(train_data['level'], train_data['temp'], margins=True).style.background_gradient(cmap='summer_r')"

train_data['level'].value_counts()
fig, ax = plt.subplots(1, 1, figsize=(8, 6))
sns.countplot(x='level', data=train_data)
plt.show()
```

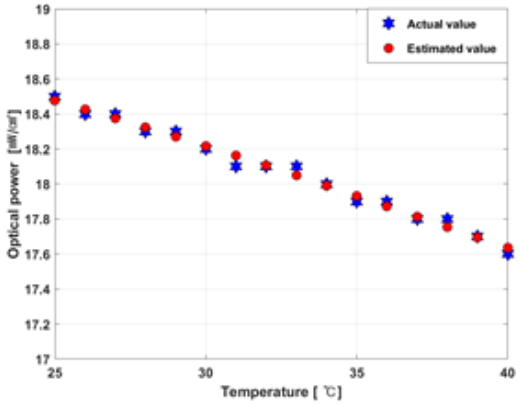
데이터 탐색적 분석 결과

탐색적 분석 결과

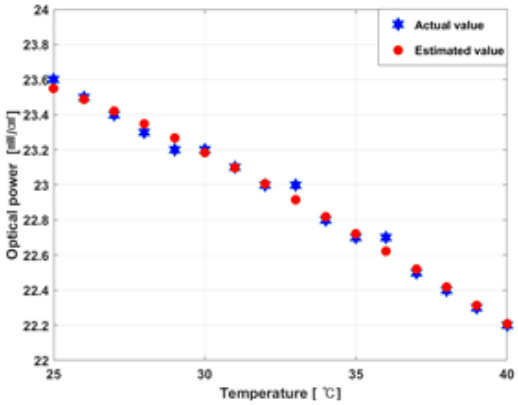
- *Train과 Test를 통한 광출력 추정 결과*
- *회귀모델에 의한 선형 추론*



레벨별 광출력 추정



(a) 120-LEVEL



(b) 150-LEVEL

Train과 Test 데이터

데이터 탐색적 분석 결과

탐색적 데이터 분석 해결 방안

- 분류와 군집 모델에 대한 적용 방법 고민

ㅇ 분류의 6가지 평가지표와 군집의 클러스터링에 해당하는 항목 선정에 애로사항

- 향후 온도와 광출력과 같은 단순한 데이터가 아닌 국가별 소득이나 계절별, 도시별 온도 등 적합한 정보의 선택

KNN 알고리즘

- 적절한 *k* 선택이 필요
- 훈련 데이터가 매우 크면(특성의 수, 샘플의 수가 클 경우) 예측이 느리고 잘 작동 하지 않음
- Nominal 속성과 누락 데이터는 추가 처리가 필요

K-Means

1. Step 1: 매개 변수(parameter) k 결정 (k > 0)
2. Step 2: 중심점을 시작하기 위해 k개의 점을 무작위로 선택
3. Step 3: 모든 점을 가장 가까운 중심에 할당하여 k 클러스터 형성
4. Step 4: 각 클러스터의 중심을 다시 계산 (각 클러스터의 평균 계산)
5. Step 5: 중심이 변하지 않을 때까지 step 3를 반복

감사합니다