제3회 K-인공지능 제조데이터 분석 경진대회 보고서

프로젝트명	LSTM Auto-Encoder를 통한 불량률 발생 요건 사전예측
팀명	망 구 리
내용요약	Austempering 열처리 공정 데이터와 LSTM Auto-encoder를 활용하여 시계열 센서의 이상치를 탐지를 학습하고 불량률을 예측하는 모델을 개발하는 과정을 살펴봅니다. 공정에서 발생할 수 있는 문제를 예측하고 그에 맞는 데이터 전처리, 모델 구축, 모델 성능확인 및 분석 과정을 통하여 모델을 구현하고 활용할 수 있는 방안에 대하여 탐구합니다.

상기 본인(팀)은 위의 내용과 같이 제3회 K-인공지능 제조데이터 분석 경진대회 결과 보고서를 제출합니다.

2023 년 11 월 03 일

팀장 : 김 지 호 (서명)

팀원 : 강 창 민 (서명)

한국과학기술원장 귀중

□ 문제정의

○ Austempering 열처리 공정 중 불량품이 발생하는 경우

- Austempering 열처리 공정이란

* 일반적인 열처리 공정을 할 때, 공기중에서 빠른속도로 냉각하여 austenite의 구조가 변형되며 martensite 조직을 형성하나, 위 냉각공정을 빠르게 거칠 시, 제품이 팽창하여 변형, 파손등 문제점이 발생

* Austempering 열처리 공정은, 열처리를 가한 austenite 상태의 제품을 약 300℃ ~ 350℃로 가열된 염욕에서 냉각하여 위 열처리 공정보다 천천히 냉각할 수 있어 제품의 변형을 막을수 있으며, 염욕과정에서 일정한온도로 냉각과정을 거친 뒤 공기중에서 냉각하는 과정에서 생성되는 상, 하부 bainite 조직에 의하여 일반적인 열처리를 한 제품에 비하여 신축, 단면수축률, 충격치 등이 크고 인성이 풍부한 제품이 얻어진다.

- Austempering 열처리 공정 중 불량품이 발생하는 케이스 추정

* 염욕과정중 염수의 온도가 300℃ ~ 350℃로 유지되지 않는 경우
* 염욕과정을 통해 천천히 냉각하므로, 염욕과정이 진행되는 시간이 짧을 경우
* 열처리 공정의 공통사항으로, 특정 구간들의 온도가 공정에서 사용하는 온도 범위를 벗어나는 경우

○ 데이터셋과 비교하여 불량품이 발생하는 케이스 분석

- 염욕과정중 염수의 온도가 300℃ ~ 350℃로 유지되지 않는 경우

* 데이터셋에서 확인할 수 있는 염욕과정(솔트조 온도 1, 2 zone)의 최솟값 최댓값의 범위가 해당 범위 안쪽이므로, 문제 사항 없다고 판단

- 염욕과정을 통해 천천히 냉각하므로, 염욕과정이 진행되는 시간이 짧을 경우

* 염욕과정의 소요시간은 데이터셋에서 확인할 수 없는 사항임

- 공정 구간들의 온도가 공정에서 사용하는 온도 범위를 벗어나는 경우

* 열처리 공정의 연속된 공정 구간 온도를 데이터셋에서 확인가능, 학습하여 공정설비 가동 시 특정 연속된 구간의 온도가 양품의 제품을 얻어낼 수 있는 범위를 벗어날 경우, 경고나 알림등을 통하여 사전에 온도를 조절함으로써, 공정 도중 제품의 불량을 확인할 수 없는 열처리 공정의 취약점을 보강, 불량률을 감소시킬 수 있을 것으로 판단됨.

□ 제조데이터 정의 및 처리과정

○ 데이터셋 분석

- 데이터셋 정의

* data.csv : 열처리 공정 각 구간별 온도 및 cp값을 배정번호별로 수집한 초 단위 데이터가 기록된 확장자 .csv 파일

* quality.xlsx : 배정번호별 작업일 및 작업설비명, 제품의 양품, 불량품, 총수량이 기록된 확장자 .xlsl 파일

- 데이터 구성 및 구조

* data.csv : 공정과정이 초단위로 기록되어있는 TAG_MIN, 공정의 배정번호, 공정 과정인 건조 1존 OP, 건조 2존OP부터 소입4존 OP까지의 출력데이터 및 소입로 온도 1 Zone 부터 염욕과정인 솔트조 온도 2 ZONE까지의 연속된 온도데이터 및 제품에 고열을 가하기 시작하는 소입로CP 값, 소입로 CP모니터 값으로 이루어진 21 컬럼, 2939722 로우를 갖는 데이터프레임, 구성 확인 결과 TAG_MIN이 중복값 없이 PK 속성을 가지고 있을것으로 판단됨

* data.csv의 구성

순번	컬 럼 명	데이터 타입	최댓값	최솟값	평균	표준편차	결측치
1	TAG_MIN	object	-	-	-	-	-
2	배정번호	int64	148069	102410	-	-	-
3	건조 1존 OP	float64	87.2995	47.2532	69.8940	4.0148	1
4	건조 2존 OP	float64	47.5395	0.0001	20.4471	5.2171	1
5	건조로 온도 1 Zone	float64	102.4690	97.3421	100.0061	0.4360	116
6	건조로 온도 2 Zone	float64	101.8430	97.8706	100.0198	0.3624	148
7	세정기	float64	71.4901	60.6244	67.7186	1.6308	91
8	소입1존 OP	float64	100.0000	0.0009	75.6437	25.1608	4288
9	소입2존 OP	float64	77.2709	8.6200	54.8624	4.4291	-
10	소입3존 OP	float64	66.0150	0.0437	53.8603	2.6643	2
11	소입4존 OP	float64	87.3907	0.0062	71.0893	2.5570	3
12	소입로 CP 값	float64	0.9091	0.0051	0.4489	0.0189	1
13	소입로 CP 모니터 값	float64	0.0000	0.0000	0.0000	0.0000	147
14	소입로 온도 1 Zone	float64	877.2280	840.2980	859.2077	3.6477	130
15	소입로 온도 2 Zone	float64	866.0340	855.9290	860.0021	0.5578	128
16	소입로 온도 3 Zone	float64	870.1190	858.2800	860.0029	0.3518	157
17	소입로 온도 4 Zone	float64	882.1480	857.9920	860.0062	0.4552	170
18	솔트 컨베이어 온도 1 Zone	float64	298.5300	266.2300	283.9963	9.5128	106
19	솔트 컨베이어 온도 2 Zone	float64	291.6960	266.4260	279.9293	6.6116	142
20	솔트조 온도 1 Zone	float64	332.7170	328.1610	331.8062	0.7827	209
21	솔트조 온도 2 Zone	float64	333.1790	328.0730	332.1773	0.8733	203

* quality.xlsx : 배정번호 및 배정번호별 작업일, 공정명, 설비명, 양품수량, 불량수량, 총수량으로 이루어진 7컬럼 136로우를 갖는 데이터프레임. 구성 확인 결과 배정번호가 중복값 없이 PK 속성을 가지고 있을것으로 판단됨

* quality.xlsx의 구성(결측치 없음)

순번	컬 럼 명	데이터 타입	최댓값	최솟값	평균	표준편차	결측치
1	배정번호	int64	-	-	-	-	-
2	작업일	datetime64[ns]	-	-	-	-	-
3	공정명	object	-	-	-	-	-
4	설비명	object	-	-	-	-	-
5	양품수량	int64	-	-	-	-	-
6	불량수량	int64	-	-	-	-	-
7	총수량	int64	-	-	-	-	-

○ 데이터셋 품질지수 측정

- data.csv 데이터셋 nan값 처리

* fillna 메서드를 이용 bfill -> ffill 순으로 모든 nan 값을 채워준다.

- 품질지수 측정

* 데이터의 완전성, 유일성, 유효성, 일관성, 정확성, 무결성 품질지수를 일괄 계산하여 리스트로 반환해주는 data_quality_idx() 함수로 data.csv와 quality.xlsx의 품질지수 측정

```
data.csv 품질지수
완전성품질지수 : 1.0
유일성품질지수 : 1.0
유효성품질지수 : 1.0
일관성품질지수 : 1.0
정확성품질지수 : 1
무결성품질지수 : 1

quality.xlsx 품질지수
완전성품질지수 : 1.0
유일성품질지수 : 1.0
유효성품질지수 : 1.0
일관성품질지수 : 1.0
정확성품질지수 : 1.0
무결성품질지수 : 1
```

○ 데이터셋 전처리 과정

- data.csv 데이터셋

* data.csv 데이터셋을 fillna 메서드로 nan값을 채운 이후 .corr() 및 heatmap 으로 상관관계 파악하여 상관관계가 1.00으로 매우 높은 '소입로 CP 모니터 값' 컬럼 제거

* 타입이 object인 TAG_MIN 컬럼을 pandas.to_datetime 및 포맷을 이용하여 데이터타입 변경

* 데이터를 train과 test로 분리. 시계열 데이터이므로, 특정 시점이 아닌 배정번호 기준 train 및 test가 분리되어야 하므로 약 80% 지점(40650 번째)에 있는 배정번호 '141145'를 기준으로 train 및 test를 나누었으며 데이터셋 분리 이후 불필요한 '배정번호', 'TAG_MIN' 컬럼을 제거하는 것으로 전처리 종료

- quality.xlsx 데이터셋

* quality.xlsx 데이터셋에서 불필요한 '작업일', '공정명', '설비명', '양품수량' 컬럼 제거

□ **분석모델 개발** (HY헤드라인M 15, 줄간격 160)

○ LSTM Auto-Encoder 분석모델

- LSTM Auto-Encoder 모델개요 및 선정이유

* LSTM Auto-Encoder 모델이란 RNN 모델중 하나인 LSTM을 활용하여 데이터를 학습, 압축하고 이를 재구성하는 Encoder-Decoder 아키텍처의 일종으로, 데이터의 압축된 표현을 학습하는 비지도 학습 모델

* LSTM Auto-Encoder 모델은 LSTM-Encoder와 LSTM-Decoder로 구성되어 있으며 LSTM-Encoder에서는 다변량 데이터를 압축하여 feature로 변환하는 역할을 하고 LSTM-Decoder에서는 해당 feature를 다변량 데이터로 재구성 하는 방식으로 input된 다변량 데이터와 output된 다변량 데이터의 차이를 줄이도록 학습됨

* 시계열, 시퀀스 데이터 내에서 이상치를 제거하거나, 이상치를 탐지하는데 우수한 성능을 보임

* data.csv 데이터는 배정번호별 초 단위로 작성된 시계열 데이터이며, 각 공정 과정별 온도 및 출력등에 관한 데이터이므로 해당 분석모델이 적절할 것으로 판단됨

○ AI 분석 방법론(알고리즘) 구축 절차

- 데이터 전처리

* data.csv 데이터셋은 상기된 전처리 과정을 통해 얻은 train, test 데이터를 사용하며, 각 컬럼별로 학습하는 과정을 거치기 위하여 스케일링 및 train_test_split 메서드는 모델링 과정중에 진행

* quality.xlsx 데이터셋의 '총수량'과 '불량수량' 컬럼을 이용 '불량수량' 컬럼의 총 합산값을 '총 수량' 컬럼의 총 합산값으로 나누어 전체 불량률 산출

- 모델 구축

* get_model_lstmae() 함수로 모델링하며 learning late는 0.0001로 설정

* 인코더 - RepeatVector - 디코더 순으로 구축되어 있으며, 활성함수로 ‘relu’, 손실함수로 평균제곱오차(MSE) 채택

* 인코더 : 입력데이터.shape를 이용하여 시퀀스의 길이 및 요소를 받아 데이터를 feature로 출력

* RepeatVector : 인코더에서 출력된 feature를 반복하여 디코더의 입력데이터로 입력

* 디코더 : 인코더로부터 받은 feature를 다시 시퀀스로 변환하고 TimeDistreibuted를 사용하여 시퀀스의 각 요소를 복원하여 출력

* optimizer = ‘adam’으로 최적화

- 모델 학습

* 반복문을 이용하여 학습 진행하며, data.csv의 각 컬럼별로 StandardScaler를 통한 정규화 및 Train_test_split 과정 진행 후 학습 진행되게끔 설계

* 각 컬럼별로 정규화 및 학습/테스트 데이터셋이 분할되어 get_model_lstmae()의 LSTM기반의 인코딩, 디코딩 과정 진행

* 각 컬럼별 학습된 모델로 train 데이터셋 모델의 MSE_loss 및 test 데이터셋의 MAE_loss 계산

* THREDSHOLD = MAE_loss의 95% 지점으로 이 지점을 넘어가는 값을 이상치로 설정

* 지정하며 최종적으로 MAE_loss 및 THRESHOLD, 이상치 여부(_anomaly) 출력

- 모델 활용 예측

* 어느 한 지점에서의 온도 이상이 아닌, 연속된 공정들의 연속된 온도 이상치에 의해 불량품률이 증가함으로 가정

* data.csv 컬럼 중 ‘OP’ 컬럼은 공정의 온도가 내려갔을 때, 올라가는 출력의 %이므로 예측에서 제외

* 연속된 3개의 공정에서 이상치인 _anomaly가 확인되었을 경우 1점, 연속된 4개의 공정 또는 3개의 공정 2구간에서 _anomaly가 확인되었을 경우 2점, 연속된 5개의 공정 또는 연속된 4개의 공정 2구간 또는 연속된 3개의 공정 3구간에서 _anomaly가 확인되었을 경우엔 3점을 부여함

* 위 점수를 토대로 불량률 계산

2점 이상일 경우 불량률	0.010715445394523217
3점 이상일 경우 불량률	0.0006494209330014071
quality.xlsx 기준 전체 불량률	0.0003357431395151549

□ 분석결과 및 시사점

○ 분석 모델 성능(MAE_loss)

- MAE_loss란

* Mean Absolute Erorr : 모델의 예측값과 실제값 간의 차이의 절대값의 평균으로, MAE가 낮을수록 모델이 더 정확한 예측을 수행하고 있음을 의미함

```
tmp_df.sort_values(by='소입로 온도 1 Zone_loss')['소입
```

```
8153    0.057552
8154    0.058772
8155    0.064923
6504    0.066504
6027    0.066869
6029    0.067160
6505    0.068285
6503    0.068508
8152    0.068659
6502    0.068901
6028    0.069932
382     0.070548
6501    0.070662
383     0.070910
8156    0.072717
5119    0.073047
5118    0.073985
4736    0.073998
4737    0.075380
6506    0.077099
Name: 소입로 온도 1 Zone_loss, dtype: float64
```

```
tmp_df.sort_values(by='소입로 온도 2 Zone_loss')['소입로
```

```
6814    0.053645
6815    0.054486
6813    0.057003
6816    0.057563
6812    0.059226
6811    0.062523
6817    0.063675
6810    0.066448
6800    0.066953
6801    0.067063
6946    0.067104
6809    0.067287
6803    0.067322
6802    0.067485
6945    0.067523
6799    0.068549
6947    0.068806
6804    0.068822
6808    0.069405
6930    0.070262
Name: 소입로 온도 2 Zone_loss, dtype: float64
```

* 모델 예측 과정에 계산된 일부 컬럼의 MAE_loss 점수로 상당히 낮고 모델이 더 정확한 예측을 수행하고 있으므로 성능이 우수한 모델이다

○ 예측된 데이터 분석

- 모델학습을 통한 분석결과

* 위 점수체계를 토대로 불량률을 계산 해 보면,
3점이상일 경우 불량률 < quality.xlsx 기준 전체 불량률 < 2점이상일 경우 불량률
위 순서로 불량률이 높아지는 것을 확인할 수 있다.

* 즉 3점 이상이 되는 경우, 상술한 연속된 5개의 공정 또는 연속된 4개의 공정 2구간 또는 연속된 3개의 공정 3구간에서 이상치가 확인되었을 경우 경고메시지, 알림 등을 통하여 해당 구간의 온도를 이상치 에서 정상구간으로 조절함으로써 불량률을 줄일 수 있을 것으로 기대된다.

□ 중소제조기업에 미치는 파급효과

○ 공정의 시작 전 미리 예측하여 불량률을 감소시킬 수 있는 강점이 있다.

제조 공정의 일반적인 특성상, 제품의 품질 확인을 공정의 마지막에 할 수 밖에 없다. 제품의 생산과정 중간에 1차 품질확인, 2차 품질확인 식으로 공정이 있을 경우, 최종적인 불량률은 감소 할 수 있겠으나 각 공정별 추가 설비비용과 인건비에 대한 부담이 중소 제조기업엔 당연히 크게 다가 올 것이다. 이러한 중간 관리과정이 아닌, 공정 시작 시 전체적인 환경데이터를 체크하여 불량률 수치가 높을것으로 예상되는 구간의 환경을 공정 직전에 변경함으로써 불량률을 감소시킴으로써, 품질확인 공정단계를 추가할 필요 없이 불량품이 줄어듦에 따라 소모되는 비용은 줄어들고 생산량은 증가 할 것으로 기대된다.

○ Austempering 열처리 공정과 같이 일정시간단위로 기록된 각 구간별 환경의 데이터가 있는 제조 산업류라면 적용 가능 할 것으로 기대된다.

단순히 열처리 공정뿐 아닌, 시간단위로 기록되는 제조산업, 생산산업등에도 적용 가능할 것으로 기대된다. 식품생산공정의 온도관리, 기계화 농법의 작물 생장관리와 같이 시간단위로 환경에 대한 기록에 부담이 적은 산업들의 경우 해당 모델을 적용하여 단순 이상치 확인으로만 끝날 것이 아닌 생산량, 수확량등의 예측도 가능할 것으로 보이며, 다양한 산업에 접목 가능할 것으로 기대된다.