Journal of KIIT. Vol. 21, No. 4, pp. 11-20, Apr. 30, 2023. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2023.21.4.11

통계 기반 특징 선택과 적대적 생성 학습을 이용한 사출기의 이상 감지

시종욱*, 정지수**, 정민수***, 김성영****

Anomaly Detection of Injection Molding using Statistics-based Feature Selection and Generative Adversarial Learning

Jongwook Si*, Jisu Jeong**, Minsu Jeong***, and Sungyoung Kim****

요 약

제조 산업 분야에서 자동화 시스템을 구축하고 있고 이에 따른 불량품에 대한 이상 감지는 포함되어야 하는 기술이다. 제조 산업 분야 중 금형 제조에서는 사출기를 통해 제품을 대량으로 생산하는 과정을 거친다. 따라서, 사출기를 통해 금형에 대한 불량품 생산을 감지할 수 있다. 본 논문에서는 사출기에 부착된 센서들의 데이터를 이용한 이상 감지 방법을 소개한다. 데이터 특징 선택에서는 통계 기반으로 유의미한 특징만을 선별하고 적대적 생성 학습을 통한 이상 감지 모델을 학습하고 추론한다. 데이터는 사출기에 부착한 센서를 통해 1분 단위로 수집하였으며, 기존 모델들에 대하여 실험하고 분석하였다. 그 결과로, 특징 선택은 사출기 데이터에 대하여 성능을 향상시킬 수 있음을 확인하였다.

Abstract

The automation system is being built in the manufacturing industry, and anomaly detection in defective products should be included. Among the manufacturing industry, mold manufacturing goes through a process of producing a large amount of products through injection molding. So, it is possible to detect the production of defective products for the mold through injection molding. In this paper, we introduce a method for anomaly detection using the data of sensors attached to the injection molding. In data feature selection, only meaningful features are selected based on a statistic, and anomaly detection models through generative adversarial learning are trained and inferred. Data were collected every minute through sensors attached to injection molding, and existing models were tested and analyzed. As a result, it was confirmed that the feature selection can improve performance of the injection molding.

Keywords

anomaly detection, time series data, injection molding, feature selection, GAN

* 금오공과대학교 컴퓨터·AI융합공학과 박사과정
- ORCID: http://orcid.org/0000-0003-2092-2769
** 금오공과대학교 수리빅데이터학과 학사과정
- ORCID: http://orcid.org/0000-0001-8148-5182
*** 금오공과대학교 컴퓨터공학과 학사과정
- ORCID: http://orcid.org/0000-0001-7992-5484
**** 금오공과대학교 컴퓨터공학과 교수(교신저자)
- ORCID: http://orcid.org/0000-0002-7722-6759

· Received: Feb. 08, 2023, Revised: Feb. 21, 2023, Accepted: Feb. 24, 2023
· Corresponding Author: Sungyoung Kim
Dept. of Computer Engineering, Kumoh National Institute of Technology, 61 Daehak-ro (yangho-dong), Gumi, Gyeongbuk, [39177] Korea
Tel.: +82-54-478-7530, Email: sykim@kumoh.ac.kr

Ⅰ. 서 론

최근 스마트팩토리가 큰 주목을 받으며 최적의 비용과 시간으로 생산 효율을 꾸준히 높이고 있다. 이러한 스마트팩토리에서는 자동화 시스템 구축을 위해 이상 감지 기술은 필수적이다. 하지만, 불량 제품에 대한 데이터는 정상 제품과 비교하면 매우 적어 고장이나 불량 등을 예측하기 어려운 것이 현실이다.

금형은 제조 분야에서 제품을 생산하는 데에 주로 사용되는 도구이며 사출, 프레스 등이 제조 과정에 함께 사용된다. 일반적으로 사출 공정에서 사출 금형을 이용해 제품을 대량으로 생산하며 일상생활에서 사용하는 대부분의 플라스틱 제품들이 사출 금형을 통해 만들어질 정도로 널리 사용되고 있다. 이러한 금형은 사출기를 통해 만들어진다. 따라서 사출기를 통하여 금형에 대한 불량품 생산을 감지할 수 있다. 생산 제조 공정에 부착된 센서 데이터를 기반으로 금형에 대한 불량 발생 시점과 수치를 분석하여 이상 패턴을 감지할 수 있다. 그리고 불량품 감지의 정확도 향상을 통해 생산성 또한 극대화할 수 있다. 이러한 이상 패턴을 감지하기 위해 데이터 공학에 주목할 필요가 있다.

데이터 공학이란 데이터를 수집, 가공, 저장 그리고 분석을 위한 시스템을 설계하고 구축하는 것이다. 데이터 공학에서는 수집된 데이터에 대하여 영향을 끼치는 속성을 찾기 위해 탐색적 데이터 분석을 진행할 수 있다. 데이터 분석은 다양한 방면에서 이해와 관찰이 필요한 데이터 분석 방법론으로, 가설검증을 적용하기 전에 사람에게 직관적인 정보를 얻게 하는 과정을 의미한다. 또한, 데이터 분석을 통해 시각화하고 불량에 영향을 끼치는 특징들을 제공하는 기술도 제안되었다[1].

본 논문에서는 사출기 데이터에 주목하여 이상 감지의 높은 이상 감지의 성능을 달성하는 것이 목적이다. 데이터는 사출기에 온도, 습도 등의 다양한 센서들을 부착하여 획득한다. 그리고 통계 기반의 특징 선택으로 차원을 줄임과 동시에 적대적 생성 학습을 이용하여 높은 정확도로 이상을 감지하는 방법을 소개한다.

Ⅱ. 관련 연구

이상의 정의나 종류, 감지 방법에 대하여 이상 감지은 여러 카테고리로 분류할 수 있다.

Probability 기반의 방법[2][3]은 확률 분포를 이용하여 이상을 감지하는 것이다. 정상 데이터의 확률 분포와 다르게 이상 데이터는 낮은 확률 분포에서 나타난다고 가정한다. 따라서, 통계적 추론에 의한 데이터의 분포를 잘 생성하여야 한다. COPOD[2]와 ECOD[3]는 모두 누적 분포 함수를 기반으로 데이터의 분포를 구성한다. 이렇게 만들어진 분포들에 대하여 Tail 부분에 위치하는 확률에 대하여 이상으로 판단한다. 두 연구 모두 복잡한 파라미터 튜닝을 진행하지 않아도 되는 장점을 지닌다.

Proximity 기반의 방법[4][5]은 정상적인 데이터의 집합으로부터 멀리 떨어진 특이치를 찾는 방식이다. LOF[4]은 Local의 밀도를 이용하여 이상을 감지하는 방법이다. 모든 데이터를 고려하지 않고 중심 데이터에 대한 Local 영역에 존재하는 데이터만을 가지고 감지한다. 따라서, Local을 정의할 수 있는 이웃의 개수에 대한 실험이 성능에 영향을 미치는 특징이 있다. ROD[5]는 전체 데이터 공간을 여러 Sub-Space 조합으로 분해하고 3D 벡터의 Rodrigues 회전 공식에 기반하여 이상을 감지하는 방법을 제안하였다.

Outlier Ensembles 기반의 방법[6][7]은 이상을 감지하는 여러 조건에 대한 모델을 종합하여 감지할 수 있다. Isolation Forest[6]는 여러 Decision Tree를 생성하여 판단한다. Decision Tree를 통해 정상을 찾기 위해서는 Tree의 깊이가 깊어진다. 즉, 이상 데이터는 Tree의 상단부에서 감지할 수 있다는 가정이 필요하다. HOBS[7]는 히스토그램을 기반으로 이상치를 찾는 방법을 제안하였다. 클러스터링, 근접 이웃해 기반 방법보다 속도가 빠름을 보였다.

Linear Model은 종속 변수와 독립 변수와의 관계가 선형 결합으로 표현할 수 있는 형태를 의미한다. 정상과 이상을 구분할 수 있는 경계를 이용해 이상 감지에 적용이 가능하다[8][9]. OCSVM[8]은 입력 데이터를 특징 공간으로 변환하는데 이때, 정상 데이터를 원점으로부터 거리가 멀어지도록 생성한다. 그리고 원점과의 거리를 기준으로 초평면을 만들며,

이 초평면을 기준으로 이상을 감지한다. Shyu et. al.[9]은 주성분 분석을 기반으로 이상을 감지하는데, 정상 데이터에 대한 주성분과 부성분에 주목한다. 전체 분산에 대한 주성분과 0.2 미만의 부성분의 고윳값을 통해 이상 감지가 가능함을 보였다.

Neural Network를 기반으로도 이상 감지가 가능하다[10]-[12]. DeepSVDD[10]와 AnoGAN[11]은 모두 정상 데이터만을 이용해 이상 데이터를 감지하는 Semi-supervised 학습 방식이다. DeepSVDD[10]은 OC-SVM[8]과 같이 데이터에 대한 새로운 특징 공간으로 변환한다. 변환 간에 가장 작은 원을 만드는 방향으로 딥러닝 모델을 학습한다. 정상 데이터를 포함하는 가장 작은 경계를 기준으로 이상을 감지할 수 있다. AnoGAN[11]에서 Generator는 노이즈가 실제에 가까운 데이터를 만들어 내도록 정상 데이터의 매니폴드를 학습하는 모델을 생성한다. Image Space와 Latent Space 사이의 차이를 이용하여 이상을 판단하는데, 이상 데이터의 경우 정상 데이터보다 그 차이가 크다는 가정이 필요하다.

이렇게 이상 감지는 방법에 따라 여러 카테고리로 분류할 수 있다. 그리고 이상 감지에는 고장 진단이나 이상 징후를 판단하는 것도 포함되며, 이와 관련한 연구가 활발히 진행 중에 있다[14]-[17].

III. 사출기의 이상 감지

제안하는 이상 감지 방법에서는 크게 특징 선택와 모델 부분으로 나뉜다. 특징 선택에서는 Label을 이진화하고 결측치 및 Object Type을 제거한다. 그리고 유효한 특징을 추출하고 정규화를 수행하여 데이터를 전처리한다. 이 데이터를 이상 감지 모델을 통해 이상을 감지할 수 있다.

3.1 유효한 데이터의 특징 선택

3.1.1 결측치 및 Object type 제거

결측치는 데이터에 값이 누락된 것을 의미하며 'NA' 또는 'Null'으로 표현한다. 실생활에서 측정되는 데이터에는 대부분 값이 정상적으로 모두 기록되어 있는 경우가 적기 때문에 결측치를 포함한 데이터를 다룰 일이 훨씬 더 많다. 따라서 데이터 분석에서 큰 영향을 미친다. 결측치를 포함하게 된다면 편향이 생길 수 있으며 데이터 손실을 불러올 수 있으므로 결과가 많이 달라질 수 있다. 또한, 데이터 분석 모형은 수치형 데이터만을 입력으로 받을 수 있다. 따라서 수치형 데이터만을 남기기 위하여 Object type을 모두 제거하며, 그림 1(a)은 남은 98개의 특징 들을 나타낸다.

```
Index(['CycleTime', 'uvlamp_cycletime', 'InjTime', 'RecoveryT', 'Cushion',
       'InjSpeed', 'InjPress', 'TotalShot', 'ThroatTemp', 'HPTransPrs',
       'HPTransPos', 'Oily', 'Moisture', 'Ebi', 'Eai', 'INPUT', 'paintTemp',
       'supportTemp', 'supportHumidity', 'pportRTemp', 'supportRHumidity',
       'ApplyS', 'SprayS', 'PatternS', 'ApplyP', 'SprayP', 'PatternP', 'TEMP',
       'HUMIDITY', 'RTEMP', 'RHUMIDITY', 'DRYING', 'DP001', 'DP002', 'DP003',
       'DP004', 'DP005', 'DP006', 'DP007', 'DP008', 'DP009', 'Temp1', 'Temp2',
       'Temp3', 'Temp4', 'Temp5', 'Temp6', 'Temp7', 'Temp8', 'Temp9',
       'UvTemp1', 'UvTemp2', 'UvTemp3', 'UvTemp4', 'UvTemp5', 'UvTemp6',
       'UvTemp7', 'UvTemp8', 'UvTemp9', 'UvTemp10', 'LABORATOR', 'SPTEMP',
       'SPHUMIDITY', 'HP1ACTIONTEMPP', 'HP1ACTIONTEMPS', 'HP1ENDTEMP',
       'HP2ACTIONTEMPP', 'HP2ACTIONTEMPS', 'HP2ENDTEMP', 'WH1ACTIONTEMPP',
       'WH1ACTIONTEMPS', 'WH1INTEMP', 'WH1OUTTEMP', 'WH1INFLUX',
       'WH2ACTIONTEMPP', 'WH2ACTIONTEMPS', 'WH2INTEMP', 'WH2OUTTEMP',
       'WH2INFLUX', 'WH3ACTIONTEMPP', 'WH3ACTIONTEMPS', 'WH3INTEMP',
       'WH3OUTTEMP', 'WH3INFLUX', 'WH4ACTIONTEMPP', 'WH4ACTIONTEMPS',
       'WH4INTEMP', 'WH4OUTTEMP', 'WH4INFLUX', 'CTACTIONTEMPP',
       'CTACTIONTEMPS', 'CTPRESSURE', 'CTOUTTEMP', 'OUTTEMP', 'OUTHUMIDITY',
       'MI_TEST_RESULT', 'PAMCALE_RESULT', 'RESULT'],
      dtype='object')
전처리 전 column 수 : 98
```

(a) 전체 특징 목록

(a) List of features after preprocessing

```
Index(['UvTemp2', 'SPTEMP', 'InjSpeed', 'HPTransPrs', 'InjPress', 'WH4OUTTEMP',
       'CTACTIONTEMPS', 'HP2ACTIONTEMPS', 'HP1ACTIONTEMPS', 'SprayS', 'SprayP',
       'HPTransPos', 'RecoveryT', 'TotalShot', 'Temp9', 'Cushion', 'DP002',
       'HP2ENDTEMP', 'ApplyS', 'WH4INFLUX', 'Moisture', 'DP001', 'OUTHUMIDITY',
       'OUTTEMP', 'WH2INFLUX', 'WH3INTEMP', 'uvlamp_cycletime', 'WH4INTEMP',
       'UvTemp1', 'INPUT', 'DP003', 'RESULT'],
      dtype='object')
회귀 분석 후 column 수 : 32
```

(b) 회귀 분석 후 특징 목록

(b) List of features after regression analysis

```
Index(['Cushion', 'DP001', 'DP003', 'HP2ENDTEMP', 'HPTransPos', 'HPTransPrs',
       'InjPress', 'InjSpeed', 'Moisture', 'OUTTEMP', 'RecoveryT', 'SPTEMP',
       'Temp9', 'UvTemp1', 'UvTemp2', 'WH2INFLUX', 'WH3INTEMP', 'WH4INFLUX',
       'WH4INTEMP'],
      dtype='object')
처리 후 column 수 : 19
```

(c) 최종 선택된 특징 목록

(c) Final selected list of features

그림 1. 유효한 특징 선택에 대한 단계적 결과

Fig. 1. Step-by-step results for valid feature extraction

3.1.2 회귀 분석

회귀 분석은 두 개 이상의 특징 간의 관계를 볼 수 있는 통계학적인 방법이다. 이는 결과에 영향을 미치는 독립변수들과 결과변수인 종속변수 간의 관계를 규정하기 위한 분석 방법 중 하나이다.

독립변수 중 영향을 가장 많이 끼치는 속성을 찾기 위해 선형회귀를 통한 Stepwise 기법을 적용한다. Stepwise 기법은 전진선택법에서 한 번 선택된 변수는 계속 모형에 존재한다는 단점과 후진 선택법의 한번 모형에서 제외된 변수는 다시 해당 모형

에 포함될 수 없다는 단점을 보완한 방법이다.

기존 모형에 포함된 변수들의 집합 Y, 모형에 포함되지 않은 변수들의 집합 $\hat{Y}$, 유의수준 α에 대하여 모형에 적합시키지 않은 변수(X_k)를 기존 모형에 추가한다($X_k \in \hat{Y}$). 모든 X_k에 대한 회귀계수 β_k를 구하고 β_k에 대한 t-통계량을 계산한다. 그리고 t-통계량에 대응되는 p-value를 구한다. 이후, 설정한 최소 p-value 값과 α와 비교한다. 만약 최소 p-value가 α보다 작다면 최소 p-value에 해당하는 변수를 Y에 포함한다.

추가된 변수를 포함하여 현재 Y에 존재하는 모든 변수를 이용하여 회귀 모형을 적합하고 절편항이 제외된 추정된 회귀변수에 대해 Max p-value를 구한다. Max p-value가 사전에 정의된 α보다 크거나 같으면 해당 변수를 제외하고 다시 처음 스텝으로 돌아가서 이전에 진행했던 스텝들을 수행한다. Max p-value가 사전에 정의된 α보다 작으면 제외하는 변수 없이 처음 스텝으로 돌아가서 다시 이전 스텝들을 진행한다. 만들어진 모형에 대한 f-검정은 (1)과 같이 나타내며 $\beta \neq 0$임을 검정하는 것이다. 따라서 하나의 회귀계수를 제외한 모든 계수가 0이라도 나머지 하나의 회귀계수가 0이 아니라면 이 모형은 유의하다고 판단한다.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \epsilon_i \quad (1)$$

선형 모형에서는 적합한 특징이 늘어날수록 R^2의 값은 늘어나게 된다. 따라서, 크게 유의미한 특징이 속해있지 않은 부적절한 모형임에도 불구하고 적합한 특징이 많아질수록 설명력은 좋아지는 문제가 발생할 수 있다. 이러한 문제를 방지하기 위해 본 논문에서는 R^2가 아닌 Adjusted-R^2을 사용한다. 기존 모형에 새로운 특징이 적합할수록 패널티를 적용함으로써, 모형에 대한 설명력을 측정하고 모형에 선택된 각각의 특징이 종속변수의 예측 변수로 적합함을 확인한다. 회귀 분석의 결과로 그림 1-(b)와 같이 32개의 특징이 선택되었다.

3.1.3 정규화 검정

선형회귀를 통해 선택된 속성 중 공통으로 영향을 주는 특징을 찾아야 한다. 변수 선택 부분에서는 상대적으로 많은 정상의 데이터를 이상 데이터의 개수만큼 UnderSampling 한다.

샘플링된 데이터는 각 속성의 모집단을 대표하는 대표성을 가져야 한다. 샘플링하는 과정에서 귀무가설 H_0는 “정상 데이터의 전체 평균인 모평균과 샘플링 된 정상 데이터의 표본평균이 같다”, 대립가설 H_1는 “정상 데이터의 전체 평균인 모평균과 샘플링 된 정상 데이터의 표본평균이 같지 않다”가 된다. 이를 검증하기 위해 t-test를 사용하여 랜덤으로 샘플링된 정상 데이터와 모집단을 비교한다. 그리고 각 속성을 대표할 수 있는 p-value값이 α보다 큰 샘플만 뽑아서 사용한다. 이상 데이터의 경우 전체 데이터를 사용하므로 위의 과정을 생략한다.

3.1.4 독립 T 검정

각 속성에 대해 뽑힌 샘플들은 이후 적용할 통계기법인 t-test는 일반적으로 데이터의 정규분포라 가정하기 때문에 독립 T-검정을 시행한다.

독립 T 검정은 두 집단에 대한 평균의 차이를 검증하기 위해 시행하게 된다. 귀무가설 H_0는 “선택된 집단에 따라 유의미한 평균 차이가 존재하지 않는다”, 대립가설 H_1은 “선택된 집단에 따라 유의미한 평균 차이가 존재한다.”가 된다. 여기서 집단에 따른 평균의 차이가 없다면 이는 해당 종속변수가 유의하게 영향을 끼치지 않는다고 판단한다. 따라서 대립가설이 채택되어야 유의미한 영향을 준다고 판단할 수 있다. 위 가설을 토대로, 독립 t-test를 통해 측정된 p-value를 확인한다. 그리고 α보다 낮은 값이 나오는 특징에 대하여 최종 특징으로 선택한다. 이는 정상과 이상에 대한 공변인이며, 결과에 영향을 끼치는 유의미한 특징으로 판단할 수 있다.

본 논문에는 신뢰도를 높이기 위해 샘플링 과정을 500번 반복하여 시행한 뒤 가장 빈번히 선택된 특징들을 선택한다. 회귀 분석에서 선택된 특징에 이어서 정규화 및 독립 T 검정을 통해 본 논문에서는 그림 1(c)와 같이 19개의 특징을 최종 선택한다. 그림 2는 최종 선택된 특징 간의 상관관계를 나타낸다.

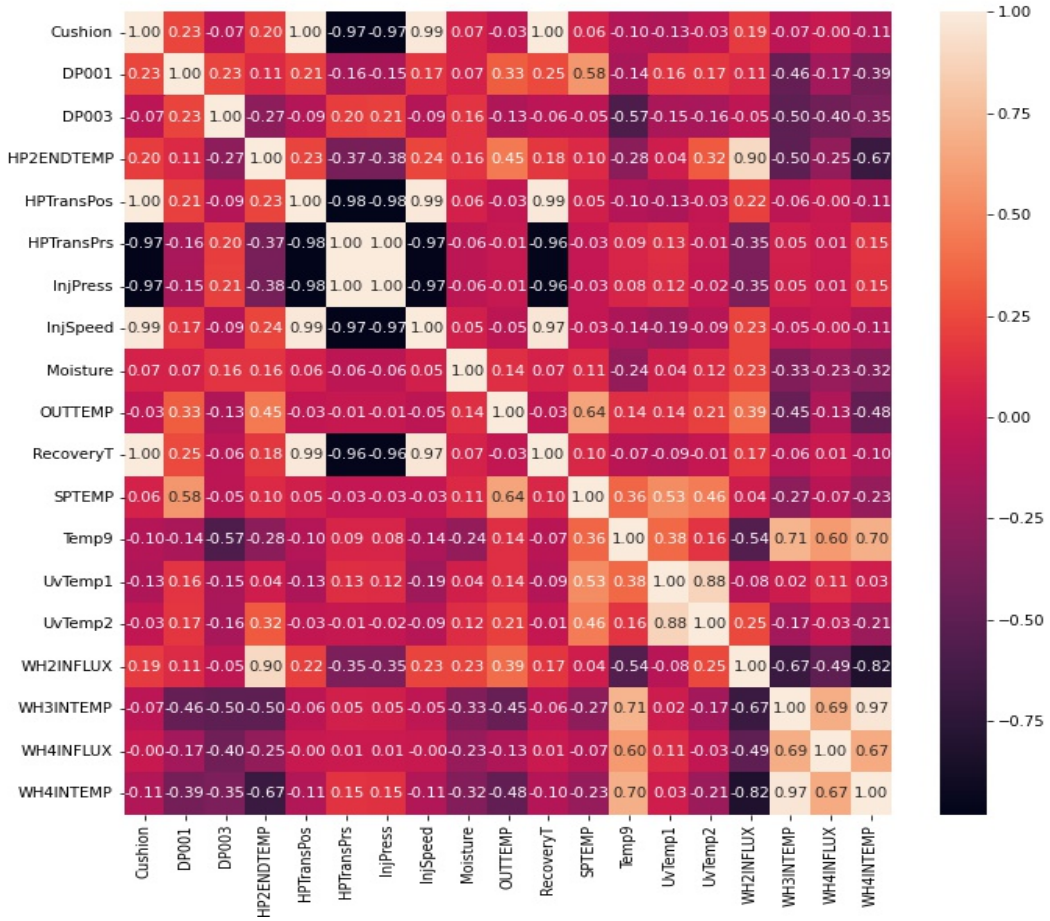


그림 2. 최종 선택된 특징 간의 상관관계
Fig. 2. Correlation between the final selected features

3.2 적대적 생성 방식의 모델 학습

본 논문에서는 사출기에 부착된 센서에서 획득한 다변량의 시계열 데이터에 대하여 이상을 감지하는 것이 목적이다. 하지만, 이 다변량의 데이터는 고차원의 특징을 지니기 때문에 높은 정확도를 달성하기가 힘들다. 이를 해결하기 위해 본 논문에서는 GAAL[11]을 이용하여 적대적 생성 방식의 모델을 학습하고 평가한다. GAAL[11]은 고차원의 데이터에 대한 희소성에 의해 이상 데이터를 쉽게 분리할 수 없는 단점을 해결할 수 있다. 그림 3은 본 논문에서 제안하는 전반적인 이상 감지의 시스템 구조를 나타낸다.

Generator는 Discriminator를 속일 수 있도록 실제 같은 Fake 데이터를 만들어 내는 것이 목적이다. Generator가 생성한 Fake 데이터를 잠재적 이상 데이터라 하며, 노이즈를 입력받기 때문에 훈련 초기에는 기존 데이터들의 분포와 관련 없는 형태를 보인다. 하지만, 학습이 진행되면 실제 데이터의 생성 메커니즘을 보이기 때문에 이상 데이터에 대한 데이터 편향을 극복할 수 있는 장점이 있다.

Discriminator는 정상 데이터와 이상 데이터를 구분하는 분류기 역할을 한다. 학습 초기에는 잠재적 이상 데이터와 이상 데이터는 Outlier 영역이지만 생성 메커니즘과 관계가 적어 Discriminator는 낮은 성능을 보인다. 학습이 충분히 진행되면 Discriminator는 잠재적 이상 데이터와 이상 데이터 사이에 분포되는 정상 데이터들에 대한 정확한 판단을 내릴 수 있다.

Generator는 k개의 Sub-Generator들로 구성된다. 각각의 Sub-Generator들은 Input Layer, Hidden Layer, Output Layer로 구성되며 모든 Layer는 FC Layer이다. Generator의 경우에는 입력 유닛은 3.1에서 진행한 특징 선택에 의해 선별된 특징들의 개수만큼 입

력으로 한다. 잠재적 이상 데이터는 이상 데이터와 차원이 같으므로 Output Layer 또한 입력 유닛의 크기로 지정한다. 그리고 모든 연산에서 활성화 함수는 ReLU를 사용한다. Discriminator는 Input Layer와 Output Layer로만 구성되며 Input Layer의 유닛 개수는 $\lceil \sqrt{n'} \rceil$ 이다. Output Layer의 유닛 개수는 하나이며, 이는 0에서 1 사이의 실수 값을 지닌다. 이 값을 통해 입력 데이터에 대한 정상과 이상을 판단할 수 있다. 활성화 함수의 경우 Input Layer에서는 ReLU, Output Layer에서는 Sigmoid를 사용한다.

Generator는 Stochastic Gradient Descent(SGD), Discriminator는 Stochastic Gradient Ascent(SGA) 방식으로 학습을 진행한다. k개의 Sub-Generator들은 전체 데이터를 k등분하여 각 블록에 대한 데이터들에 대해서만 데이터를 생성하고 Discriminator는 이를 분류한다. 학습 초기 δ Epoch 동안에는 Generator와 Discriminator는 동시에 각 모듈의 목적에 따라 학습을 진행한다. δ Epoch가 지난 후에는 Generator는 충분히 데이터 생성에 대한 메커니즘이 학습되었다고 가정하고 학습을 중지한다. 그리고 2δ Epoch 동안은 Discriminator가 정상과 이상을 잘 감지할 수 있도록 초점을 맞추어 Generator가 생성한 잠재 이상 데이터 사이에서 입력 데이터에 대한 정상과 이상을 감지하도록 한다.

IV. 실험 및 분석

4.1 데이터 세트

본 논문에서는 온도, 압력, 속도, 수분, 차압 등을 측정 가능한 다양한 센서들을 사출기에 부착하고 사출기 작동 후 1분 단위로 측정된 데이터를 수집하였다. 사출기에는 온도 센서에 대하여 52가지의 특징, 압력 센서에 대하여 18개, 습도 센서에 대하여 7개, 유량 센서에 대하여 4개, 시간 센서에 대하여 4개, 위치 센서에 대하여 2개, 그리고 기타 센서에 대한 특징이 존재한다. 이러한 특징들 중 주로 온도, 압력, 위치에 대한 열이 주로 유효한 속성을 가지는 특징을 보인다. 본 논문에서는 데이터의 통계 기반 특징 선택을 통해 전체 100개의 특징에서 19개의 특징을 선택하였다.

전체 수집한 데이터는 9650개의 행과 100개의 열로 구성된다. Train 세트와 Test 세트의 비율을 8:2로 하여 각각 7,720개의 행과 1,930개의 행으로 나눈다. 이때, 전체 데이터는 시계열 데이터이므로 순서가 섞이지 않게 나누어야 한다. 따라서 0번부터 7,719까지의 인덱스의 행은 Train 세트, 그 이후부터는 Test 세트로 나누어진다. 그림 3은 Train 세트와 Test 세트에 대하여 PCA를 통해 2차원으로 축소하여 나타낸 그림이다.

그림 4(a)는 Train 세트로 파란색 원은 정상, 노란색 세모는 이상을 나타내는데, 정상 데이터 사이에 이상 데이터가 포함되어 있기 때문에 감지하기가 어렵다. 그림 4(b)는 Test 세트에 대하여 나타낸 것인데 Train 세트와 비교하였을 때 분포가 다른 것이 특징이다. 초록색 원이 정상, 빨간색 세모가 이상을 나타내는데, 이상 데이터가 정상 데이터와 차이가 없음을 알 수 있다.

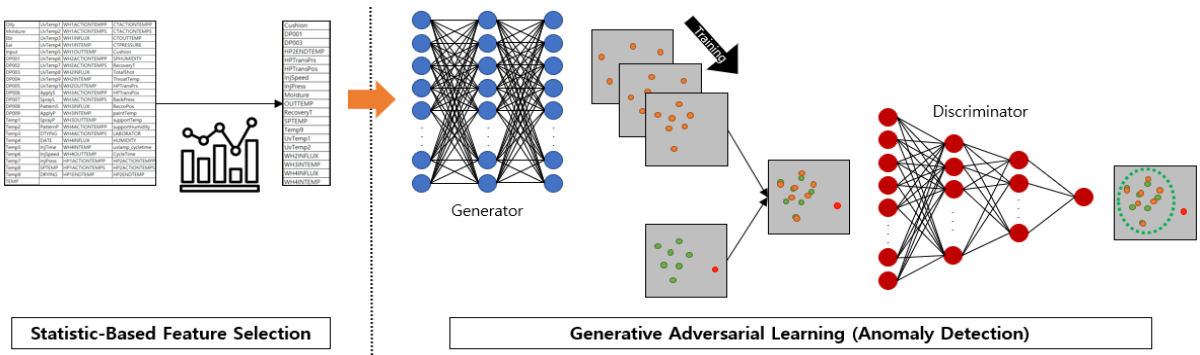
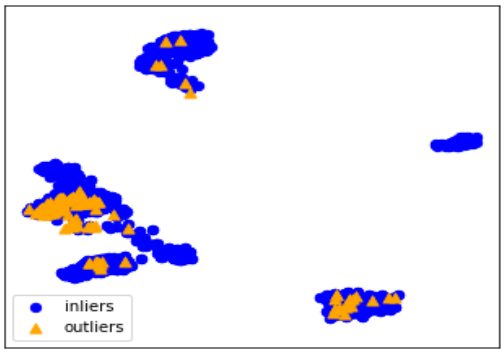


그림 3. 이상 감지 시스템의 전체 구조
Fig. 3. Structure of overall anomaly detection system



(a) 학습데이터에 대한 시각화 결과
(a) Visualization results of train data

(b) 실험데이터에 대한 시각화 결과
(b) Visualization results of test data
그림 4. 학습 및 실험 데이터에 대한 시각화
Fig. 4. Visualization of train and test data

4.2 성능 평가 및 비교

본 논문에서 제안하는 방법과 관련 연구에 대한 실험들은 Ubuntu 18.04 LTS의 운영체제 환경에서 진행한다. 그리고 Geforce RTX 3090을 이용해 모델을 학습하고 평가한다. Python은 3.6.9, Tensorflow-gpu는 2.6.0 버전을 사용한다. 특징 선택 과정에서 설정한 α는 0.05로 고정하여 실험한다.

PyOD[13]는 다변량 시계열 데이터에 대하여 Outlier를 감지할 수 있는 Python 라이브러리이며, 여러 알고리즘을 쉽게 사용할 수 있는 장점이 있다. PyOD[13]을 사용하여 GAAL[12]을 포함한 다양한 이상 감지 연구[2-11]에 대하여 성능 분석 및 비교를 진행한다. 관련 연구에 대한 모든 실험에서는 학습 시에 Contamination을 10%로 설정하며 기본값을 기준으로 모델을 학습하고 실험 데이터에 대하여 평가한다. 또한, Random Seed는 42로 고정한다.

이상 감지 성능 평가에는 모두 Macro F1-Score를 사용하여 비교 및 분석을 한다. 표 1은 Sub-Generator의 개수인 k와 StopEpoch를 의미하는 δ에 대하여 성능을 비교한 결과를 나타낸다. 전체적으로 분석하였을 때 δ가 25인 경우 k에 관계없이 0.8로 최고 성능을 달성하였다. 그리고 δ가 30인 경우 25일 때의 결과보다 근소하거나 같은 결과를 보인다. 이는 Nash Equilibrium에 도달하였다고 평가된다. Sub-Generator의 개수가 많아질수록 학습 시간 및 추론 시간이 길어지기 때문에, k가 1인 Single Mode를 최적으로 판단한다.

표 1. k와 δ의 변화에 따른 실험
Table 1. Experiments with changes in k and δ

k \ δ	5	10	15	20	25	30
1	0.506	0.731	0.789	**0.800**	**0.800**	**0.800**
5	0.636	0.784	0.628	0.793	**0.800**	0.798
10	0.661	0.778	0.665	0.788	**0.800**	0.798
15	0.661	0.778	0.686	0.792	**0.800**	0.798
20	0.661	0.778	0.698	0.793	**0.800**	0.798
25	0.674	0.775	0.713	0.795	**0.800**	0.798
30	0.661	0.778	0.704	0.793	**0.800**	**0.800**

표 2는 위의 실험에 따라 정상과 이상을 기준으로 한 Precision, Recall, F1-Score를 나타낸 것이다. 정상 및 이상 데이터의 비율이 매우 큰 상황에서 이상 데이터에 대한 성능이 높은 것을 알 수 있다.

표 2. 제안 방법의 성능 평가
Table 2. Performance evaluation of proposed method

	Pre.	Rec.	F1.	Support
Normal	0.980	0.981	0.981	1835
Abnormal	0.630	0.611	0.620	95
Macro F1			**0.800**	

표 3은 관련 연구에서 소개한 여러 카테고리의 이상 감지 연구들에 대하여 실험한 결과를 나타낸다. 그리고 특징 선택의 유무에 따른 성능 변화를 보인다. 본 데이터의 특성상 이상 데이터의 개수가 정상 데이터에 비하면 매우 적다. 따라서, 쉬운 모델을 사용하여 평가한다면 이는 하나의 결과로 밀어버리는 위험한 결과를 보여준다. 모두 정상 데이터로 감지한다면 F1-Score의 결과는 0.487, 이상 데이터로 감지한다면 0.047의 결과를 보인다.

표 3. 이상 감지 연구들에 대한 특징 선택의 유무에 따른 성능 비교
Table 3. Comparison of F1-score by feature extraction in anomaly detection researches

Related works		Feature selection	
Method	Category	X	O
COPOD[2]	Probability-based	0.467	0.481
ECOD[3]	Probability-based	0.446	0.422
LOF[4	Proximity-based	0.051	0.162
ROD[5]	Proximity-based	NaN	0.487
I-Forest[6]	Outlier ensembles	0.312	0.047
HBOS[7]	Outlier ensembles	0.239	0.047
OCSVM[8]	Linear model	0.407	0.047
Shyu *et. al.*[9]	Linear model	0.087	0.047
DeepSVDD[10]	Neural network	0.352	0.326
AnoGAN[11]	Neural network	0.487	0.312
Ours	Neural network	0.487	**0.800**

제안하는 방법의 최고 성능인 0.8을 제외하고 0.5 이상의 결과는 존재하지 않았다. Probability 기반의 방법들을 보면 COPOD[2]는 특징 선택시 1.4% 향상을 보였으나 ECOD[3]는 2.4%의 하락을 보였다. Proximity 기반의 방법들에서는 LOF[4]의 경우에는 특징 선택을 한 경우 11.2%의 성능 향상을 보였다. 하지만 사출기 데이터에 대한 이상 감지를 하기에는 매우 낮은 수치이다. ROD[5]는 특징 선택을 했을 때 모두 정상으로 평가하였지만 사용하지 않으면 결과를 볼 수 없다. 이는 98개의 특성에 대하여 Sub-Space 조합으로 분할하는 과정이 계산 비용이 높기 때문이다. Outlier Ensembles와 Linear Model이 기반인 연구들[6-9]은 특징 선택을 진행하면 모두 이상 데이터로 판단하는 상황이 발생한다. 이는 중요한 데이터의 특성만 사용하는 과정에서 정보가 손실되기 때문으로 판단된다. Deep-SVDD[10]와 Ano-GAN[11]의 경우 또한 특징 선택 시 성능이 하락하는 결과를 보인다. 이는 특성을 제거함으로써 모델에 대한 과적합 현상이 발생하고 몇 개의 이상 데이터에 대해서만 올바르게 감지하기 때문이다. GAAL[12]을 기반으로 한 본 논문에서 제안하는 방법은 특징 선택을 했을 때, 31.3%의 성능 향상을 보였다. 통계 기반으로 선택하여 관련이 높은 특징만을 이용하고, 잠재적 이상 데이터를 생성하는 방법으로 데이터 편향 현상을 해결한 것이 성능 향상의 큰 요인이다.

V. 결론 및 향후 과제

본 논문에서는 사출기에 부착된 센서로부터 데이터를 획득하고 통계 기반의 데이터 특징 선택을 이용하였다. 여러 특성 중 상관관계가 높은 특징들만 추출하여 성능을 향상시킬 수 있음을 보였다. 그리고 기존 이상 감지 모델 중에서 적대적 생성 학습을 기반으로 이상 데이터를 추가로 생성하는 방법으로 큰 성능 향상을 이루었다. 하지만, 본 논문에서 제안하는 방법은 기존의 이상 감지 모델에 대해서는 특징이 줄어 오히려 더 낮은 성능을 보인다.

수집한 데이터는 이상을 감지하는 것이 어려운 만큼 더 높은 성능을 달성해야 할 필요가 있다. 따라서 더 많은 데이터를 수집하고 이상 감지를 위한 딥러닝 모델 구조에 변경과 더 의미있는 특징 추출 과정 등의 연구가 진행될 예정이다.

References

[1] J. Jeong, M. Jeong, J. Si, and S. Kim, "Development of Data Analysis System for Determining Product Quaility of Injection Molding Machine Results", Proc. Of KIIT Conference, Jeje, Korea, pp. 331-332, Dec. 2022.

[2] Z. Li, Y. Zhao, N. Botta, C. Ionescu, and X. Hu, "COPOD: copula-based outlier detection", IEEE international conference on data mining, Sorrento, Italy, pp. 1118-1123, Nov. 2020. https://doi.org/10.1109/ICDM50108.2020.00135.

[3] Z. Liu, Y. Zhao, X. Hu, N. Botta, C. Ionescu, and G. Chen, "ECOD: Unsupervised Outlier Detection Using Empirical Cumulative Distribution Functions", IEEE Transactions on Knowledge and Data Engineering (Early Access), pp. 1-13, Mar. 2022. https://doi.org/10.1109/TKDE.2022.3159580.

[4] M. M Breunig, H. Kriegel, R. Ng, and J. Sander, "LOF: identifying density-based local outliers", ACM sigmod record, Vol. 29, No. 2, pp. 93-104, Jun. 2000. https://doi.org/10.1145/342009.335388.

[5] Y. Almardeny, N. Boujnah, and F. Cleary, "A Novel Outlier Detection Method for Multivariate

Data", IEEE Transactions on Knowledge and Data Engineering, Vol. 34, No. 9, pp. 4052-4062, Sep. 2022. https://doi.org/10.1109/TKDE.2020.3036524.

[6] F. T. Liu, K. M. Ting, and Z. Zhou, "Isolation Forest", IEEE International Conference on Data Mining, Pisa, Italy, pp. 413-422, Dec. 2008. https://10.1109/ICDM.2008.17.

[7] M. Goldstein and A. Dengel, "Histogram-based Outlier Score (HBOS): A Fast Unsupervised Anomaly Detection Algorithm", Annual German Conference on Artificial Intelligence, Saarbrucken, Germany, pp. 59-63, Sep. 2012.

[8] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the Support of a High-Dimensional Distribution", Neural Computation, Vol. 13, No. 7, pp. 1443-1471, Jul. 2001. https://doi.org/10.1162/0899766017 50264965.

[9] M. L. Shyu, S. C. Chen, K. Sarinnapakorn, and L. Chang, "A novel anomaly detection scheme based on principal component classifier", Proc. Of IEEE Foundations and New Directions of Data Mining Workshop, pp. 172-179, Jan. 2003.

[10] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft, "Deep One-Class Classification", Proc. Of International Conference on Machine Learning, Stockholm, Sweden, pp, 4390-4399, Mar. 2018.

[11] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery", International Conference on Information Processing in Medical Imaging, Boone, NC, USA, pp. 146-157, Jun. 2017.

[12] Y. Liu, Z. Li, C. Zhou, Y. Jiang, J. Sun, M. Wang, and X. He, "Generative adversarial active learning for unsupervised outlier detection", IEEE Transactions on Knowledge and Data Engineering, Vol. 32, No. 8, pp. 1517-1528, Mar. 2019. https://doi.org/10.1109/TKDE.2019.2905606.

[13] Y. Zhao, Z. Nasrullah, and Z. Li, "PyOD: A Python Toolbox for Scalable Outlier Detection", Journal of Machine Learning Research 20, arXiv preprint arXiv:1901.01588, pp. 1-7, May 2019.

[14] J. Yang, Y. Lee, and I. Koo, "Timely Sensor Fault Detection Scheme based on Deep Learning", Journal of The Institute of Internet, Broadcasting and Communication, Vol. 20, No. 1, pp. 163-169, Feb. 2020. https://doi.org/10.7236/JIIBC.2020.20.1.163.

[15] J. Si and S. Kim, "Prediction of Temperature Abnormality based on Stacked LSTM for Control System of Thermo-Hygrostat", Journal of Korean Institute of Information Technology, Vol. 20, No. 10, pp. 47-52, Oct. 2022. https://doi.org/10.14801/jkiit.2022.20.10.47.

[16] J. Si and S. Kim, "Traffic Accident Detection in First-Person Videos based on Depth and Background Motion Estimation", Journal of Korean Institute of Information Technology, No. 19, No. 3, pp. 25-34, Mar. 2021. https://doi.org/10.14801/jkiit.2021.19.3.25.

[17] J. Kim, J. Seon, and S. Yoon, "Classification Method based on Graph Neural Network Model for Diagnosing IoT Device Fault", Journal of The Institute of Internet, Broadcasting and Communication, Vol. 22, No. 3, pp. 9-14, Jun. 2022. https://doi.org/10.7236/JIIBC.2022.22.3.9.

저자소개

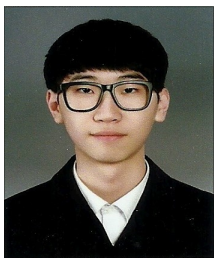
시 종 욱 (Jongwook SI)

2020년 8월 : 금오공과대학교 컴퓨터공학과(공학사)
2022년 2월 : 금오공과대학교 컴퓨터공학과(공학석사)
2022년 3월 ~ 현재 : 금오공과대학교 컴퓨터·AI융합공학과 대학원 박사과정
관심분야 : 컴퓨터비전, 영상분석, 인공지능, 이상 감지, 영상복원

정 지 수 (Jisu Jeong)

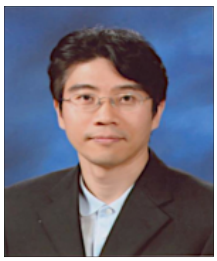
2019년 3월 ~ 현재 : 금오공과
대학교 수리빅데이터학과
학사과정
관심분야 : 데이터공학, 빅데이터

정 민 수 (Minsu Jeong)



2018년 3월 ~ 현재 : 금오공과
대학교 컴퓨터공학과 학사과정
관심분야 : 기계학습, 인공지능

김 성 영 (Sungyoung Kim)



1994년 2월 : 부산대학교
컴퓨터공학과(공학사)
1996년 2월 : 부산대학교
컴퓨터공학과(공학석사)
2003년 8월 : 부산대학교
컴퓨터공학과(공학박사)
2004년 ~ 현재 : 금오공과대학교
컴퓨터공학과 교수
관심분야 : 영상처리, 컴퓨터비전, 기계학습, 딥러닝