J. Soc. Korea Ind. Syst. Eng Vol. 43, No. 1 : 123-131, March 2020
https://doi.org/10.11627/jkise.2020.43.1.123
ISSN : 2005-0461(print)
ISSN : 2287-7975(online)

An Ensemble Model for Machine Failure Prediction

Kang Min Cheon* · Jaekyung Yang**†

*Hyosung Information System
**Dept. of Industrial and Information Systems Engineering, Jeonbuk National University

앙상블 모델 기반의 기계 고장 예측 방법

천강민* · 양재경**†

*효성인포메이션시스템
**전북대학교 산업시스템공학과

There have been a lot of studies in the past for the method of predicting the failure of a machine, and recently, a lot of researches and applications have been generated to diagnose the physical condition of the machine and the parts and to calculate the remaining life through various methods. Survival models are also used to predict plant failures based on past anomaly cycles. In particular, special machine that reflect the fluid flow and process characteristics of chemical plants are connected to hundreds or thousands of sensors, so there are not many factors that need to be considered, such as process and material data as well as application of derivative variables. In this paper, the data were preprocessed through time series anomaly detection based on unsupervised learning to predict the abnormalities of these special machine. Next, clustering results reflecting clustering-based data characteristics were applied to produce additional variables, and a learning data set was created based on the history of past facility abnormalities. Finally, the prediction methodology based on the supervised learning algorithm was applied, and the model update was confirmed to improve the accuracy of the prediction of facility failure. Through this, it is expected to improve the efficiency of facility operation by flexibly replacing the maintenance time and parts supply and demand by predicting abnormalities of machine and extracting key factors.

Keywords : Ensemble Model, Anomaly Detection, STL and GESD, Clustering, Failure Prediction

1. 서 론

설비의 이상치를 예측하는 방법들은 과거에도 많은 연구들이 있었다. 잔존수명과 같이 현재 설비의 이상 상태를 분석하여 수명을 환산하거나 생존 모델을 활용하여 과거 이상 주기를 판단해 설비 수명을 예측하기도 한다. 하지만 센서가 많고 앞뒤 연계공정이 많은 특수 설비에 대한 수명 예측이나 이상 탐지를 예측하기에는 고려해야 될 요소가 많기 때문에 예측이 쉽지 않다. 여기서 이상치 탐지는 정상 데이터 분포를 크게 벗어나는 데이터 샘플을 탐지하는 것을 의미한다[11]. 최근에는 이상치를 탐지하는 방법을 이용하여 설비운용에서 이상 상태 감지나 네트워크 침입 탐지 등을 실시간으로 수행한다[14, 15]. 또한 거리 기반, 밀도 기반, 클러스터링 기반, 트리 기반 방법 등 다양한 이상치 탐지 방법이 개발되어 왔다[2, 5]. 하지만 단편적인 알고리즘 위주의 예측이 대다수를 이뤄왔다. 예를 들어 클러스터링 기반의 이상 탐지 방법은 기존 데이터셋의 클러스터링 중심점을 기준으로 신규 데이터와의 중심점간의 거리가 멀어지는 경우 이상으로 판단한다. 하지만 이 경우는 센서들의 미세한 변동에는 반응

Received 9 March 2020; Finally Revised 24 March 2020;
Accepted 26 March 2020
† Corresponding Author : jkyang@jbnu.ac.kr

하지 않는다는 단점이 있다. 순환신경망(RNN, Recurrent Neural Network)과 같은 방법이 대안이 될 수도 있으나, 본 연구에서는 화학 공정의 설비 이상 예측을 위해 성능이 우수한 CART(Classification and Regression Tree)의 부스팅(Boosting) 형태인 EGB(Extreme Gradient Boosting) 모델을 적용하였다. 또한 예측모델의 설명력을 높이기 위해 클러스터링 결과를 데이터셋에 포함시키는 앙상블 기법을 적용하였다.

본 연구에서는 폴리에틸렌(Polyethylene) 제품을 생산하는 공장 설비를 대상으로 고장을 예측한다. 해당 공정은 초고압법 방법을 통해 고밀도 폴리에틸렌을 생산하며, 초고압 공법에 사용되는 초고압왕복동압축기(Hyper Compressor) 설비를 대상으로 고장 예측에 대한 연구를 진행했다. 초고압왕복동압축기는 초고압 운전으로 인해 Bearing, Packing Cup, Poppet Valve, Cooling Oil Sight Glass 등 의 부품이 마모되거나 파손되는 고장이 발생한다. 그런데 그 고장의 형태가 부품에 대한 마모나 파손에 대해 미리 예측하지 못한 설비의 긴급고장(ESD : Emergency Shut Down)이 주를 이룬다. 이로 인해 생산량 감소 및 유지보수 비용, 유지보수 기간이 증가되는 어려움을 겪고 있다. 따라서 본 연구에서는 최소 2주~4주 전에 설비의 이상을 예측하고 이상 원인을 확인 할 수 있는 설비 이상 예측 모델을 제시하였다.

2. 기존 연구

이상 탐지(Anomaly Detection)란 오래전부터 연구되었던 데이터 분석 기법 중의 한 가지이며 데이터에서 예상 행동을 준수하지 않는 패턴을 찾는 문제를 말한다[4]. 패턴의 이상 탐지 알고리즘은 다양한 방법들이 존재하며 가장 기본적인 방법론으로는 수리적 모형이나 규칙기반, 군집화 계열, 패턴을 찾아내는 탐색형, 확률에 기반한 베이지안 계열 등의 통계 모형들이 있다. 최근에는 통계적 방법뿐만 아니라 기계학습을 이용하여 이상 탐지를 하는 방법들이 연구되고 있다. 하지만 이러한 기술의 특성상 학습을 위한 많은 훈련 데이터가 필요한데 비정상(Anomaly)으로 규정할 수 있는 데이터가 없거나 매우 부족하기 때문에 학습을 위한 데이터를 확보하기가 매우 어렵다[8]. 이 밖에 기계학습이 아닌 이상 탐지 방법으로 강건한 마할라노비스 거리(Robust Mahalanobis Distance, RMD)를 이용하여 이상치를 탐지하기도 한다. 다중 속성을 가지는 자료의 형태는 공분산 행렬에 의해 특성 지어지는데, 마할라노비스(MD)는 이것을 고려하는 잘 알려진 척도이며, 다변량 정규 분포 자료에서는 MD^2을 이용하여 이상치를 탐지한 연구가 있었다[7]. 또 비지도학습 알고리즘인 k-Means를 사용한 이상치 탐지는 각 점들이 군집에 할당이 되고, 군집의 중심들이 확정 된 이후에 각 점들과 할당된 군집의 중심 사이의 거리를 계산 한 후, 가장 큰 거리를 이상치로 간주한다. 해당 연구에서는 시뮬레이션으로 발생시킨 이상치의 개수만큼 가장 큰 거리를 가지는 관측치부터 순서대로 나열하여 이상치를 구분한다[9, 12]. 다음으로 Isolation Forest 방법은 훈련에 분류 정보가 필요하지 않은 무감독 학습 방법이다. 정상치에 대한 관심보다는 이상치들을 완전히 고립시키는 다른 유형의 모형 기반 방법이다. 이러한 특성은 이상치가 정상치보다 더 쉽게 고립되게 되어, 데이터로 나무를 형성했을 때, 이상치는 나무의 뿌리에 가까운 곳에서 고립되고, 이러한 나무(Isolation Tree 또는 iTree)의 특성을 이용하여 이상치를 탐지할 수 있게 된다. 주어진 데이터에 대해 iTree들의 앙상블을 쌓은 후 짧은 통과 길이를 가지는 관측치를 이상치로 분류한다[13].

본 연구에서도 희박한(Sparse) 이상치 데이터의 학습에 적합한 의사결정나무 계열의 앙상블 방법인 EGB 알고리즘을 사용하여 고장 유무를 예측한다.

3. 연구 방법

본 논문에서는 설비의 고장을 단순히 설비 상태 이상을 예측하기보다는 고장의 원인을 파악하고 설비 고장이 발생 시점을 예측하고자 했다. 본 연구대상 공정의 설비인 초고압왕복동압축기는 해당 공정에서 연속 생산에 중요한 역할을 하며 단순 이상이 발생한다고 설비를 점검할 수 없는 현실이다. 따라서 본 논문에서는 설비의 이상과 고장 예측이 가능한 모델을 생성하기 위해 아래의 <Table 1>과 같은 방법을 사용하였다.

<Table 1> Machine Failure Prediction Procedure

Order	Procedures
1	Environment Analysis(Interview with Machine and Process Personnel)
2	EDA(Exploratory Data Analysis) on Raw Data
3	Preprocess(STL and GESD Anomaly Detection)
4	Clustering(K-Means)
5	Data Set(Data set+Clustering Result) Integration
6	Composing Training Set(Normal 70% vs Abnormal 30%)
7	Model Training
8	Forecasting Risk Scores & Feature Importance Evaluation
9	Model Evaluation(Test Set)

위 <Table 1>의 프로세스는 본 연구에서 새롭게 개발한 화학 공정 설비의 고장 발생 시점과 위험 스코어 예측 모델을 생성하기 위한 방법론이다.

공정 및 설비 담당자와의 인터뷰를 통해 화학공정의 특성을 이해하는 환경분석과 데이터에 대한 기본적인 통계분석(EDA)을 실시하였다. 이 후 전처리와 학습데이터셋을 구성하였다. 전처리 단계에서는 시계열 데이터 특성이 고려된 STL(Seasonal-Trend Decomposition Procedure Based on Loess) 기법을 활용하였다. 제시된 STL과 GESD(Generalized Extreme Studentized Deviate) 기법은 전처리에 적용되어 설비 센서 측정 시계열 데이터에서 이상치(Anomaly)를 탐지하는 적절한 필터 역할을 수행한다. 또한, 클러스터링 모델의 결과를 파생속성으로 학습데이터셋에 포함하여 학습 모델의 설명력을 향상시켰다. 마지막으로 본 연구의 제안 방법론의 학습방법인 EGB 알고리즘을 통해 생성된 예측모델의 학습과정에서 계산되는 Information Gain 값을 기반으로 중요 속성(Key Features)을 확인하였고 예측정확도 확인을 통해 모델의 성능을 검증하였다.

3.1 전처리(Preprocess)

고압 압축 설비의 주요 속성으로는 유량, 온도, 압력, 가스 상태, 진폭, 재료, 촉매 및 연계(1차 압축기 및 Reactor & Intercooler 등) 설비의 센서 데이터를 포함하여 400여 개의 속성으로 구성되어있다.

아래 <Table 2>는 미가공 데이터셋과 <Table 3>은 일부 속성의 기초 통계값이다.

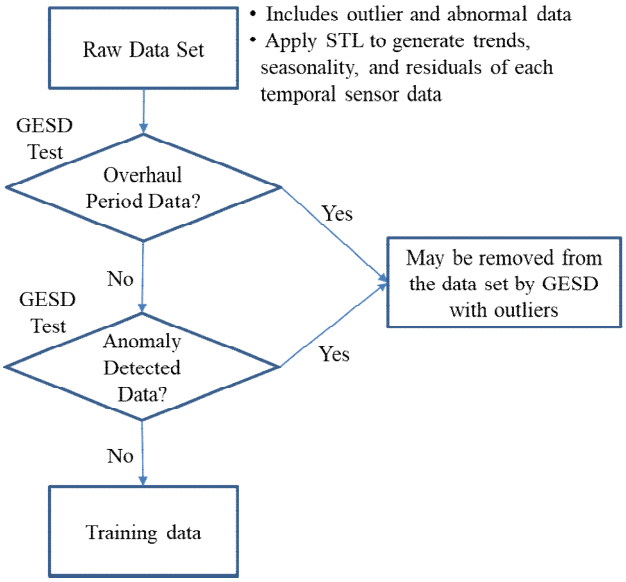
<Table 2> Raw Data Example of Several Features

Ethylene Pressure	Cylinder Temp.	Bearing Temp.	Plunger Temp.	Flux	Class
254.6	61.5	63.4	54.4	339.3	0
254.6	61.6	63.4	54.4	340.0	0
254.7	61.9	63.4	54.5	332.8	0
254.8	62.0	63.5	54.5	334.7	0
254.0	62.4	63.5	54.6	336.3	0
254.2	62.5	63.6	54.6	334.7	0
254.8	62.6	63.6	54.6	330.0	0
255.8	62.8	63.6	54.7	329.4	0
255.8	62.9	63.6	54.7	328.0	0
255.6	63.0	63.6	54.8	328.3	0
:	:	:	:	:	:
255.7	63.7	63.8	55.1	325.2	0

<Table 3> Basic Statistics of Several Features

Div.	Ethylene Pressure	Cylinder Temp.	Bearing Temp.	Plunger Temp.	Flux
Min.	122.6	-2.888	18.11	0	0
1st Qu.	221	61.113	59.7	55.73	267.5
Median	229	63.811	61.63	57.94	286.3
Mean	232.2	62.958	61.76	57.17	284.1
3rd Qu	247.9	66.445	64.07	59.35	303
Max.	266.2	105.561	72.47	91.09	595.5

데이터 수집기간은 2011년 01월부터 2019년 07월까지이며, 데이터는 약 20억 건이다. 데이터 전처리는 EDA와 설비담당자 인터뷰를 통해 방향을 설정하였으며 기본적으로 결측치는 제거하였으나 일부 는 선형보간법을 사용하여 보정하였다.



<Figure 1> Data Preprocessing by STL and GESD

또한 이상치 데이터를 제거하기 위한 전처리는 <Figure 1>과 같이 수행하였다. 설비운용에는 정상운용(Normal), 계획정비(PSD : Plan Shut Down), 이상정비(ESD : Emergency Shut Down)로 구분되어 있기 때문에 일반적인 방법으로 설비운용 상태에 대해 자동으로 판단하여 이상치로 구분하기는 어렵다. 또한 장기간 데이터의 경우 기록에 대한 정확성 문제로 인해 설비운용 구분에 대한 방법을 별도로 설정해야 한다. 그렇지 않으면 설비 정비(Overhaul) 전, 후 이상 상태에 대해 정상적인 데이터인지 판단하기 어렵고 변동의 폭이 크기 때문에 고장 증상으로 볼 수 있기 때문이다. 본 논문에서는 이러한 구분을 위해 STL을 통한 이상치 탐지 방법을 적용함으로써 정상운용과 설비정비를 구분하였다. STL은 시계열 데이터를 세 가지로 분해하기 위해 개발 된 알고리즘으로 구성 요소로는 추세, 계절 및 잔차로 구성되어 있다[19]. 이때 GESD 방법은 잔차의 이상치(최대 이상치 개수, r)를 점진적으로 제거하면서, t-Test를 통해 차이를 검정하고 신뢰구간을 설정한다. 식 (1)과 식 (2)는 GESD의 이상치 제거 기준인 R_i 및 λ_i을 보여주고 있다[16].

$$R_i = \frac{\max_i |x_i - \bar{x}|}{s} \quad (1)$$

$$\lambda_i = \frac{(n-i)t_{n-i-1,p}}{\sqrt{(n-i-1+t^2_{n-i-1,p})(n-i+1)}}, \quad (i=1,2,\cdots,r) \tag{2}$$

$\bar{x}$와 s는 표본 평균와 표본 표준 편차를 나타내며, $t_{v,p}$는 v의 자유도를 갖는 t 분포로부터의 백분율이다. 여기서 p는 식 (3)과 같다.

$$p = 1 - \frac{\alpha}{2(n-i-1)} \tag{3}$$

위 식 (1)의 이상치 최대 수 r은 $R_i > \lambda_i$일 때 가장 큰 i로 결정된다. STL 알고리즘의 매개속성(이상 감지 범위와 정상범위, 비정상범위) 설정은 계획정비(PSD) 일자와 매칭하여 명확하게 구분이 가능한 범위로 설정하였다. 이러한 STL 및 GESD 알고리즘을 통해 계획정비(PSD) 기간에 발생한 이상치 데이터에 대한 전처리 작업을 실시하여 고장 예측 시 이상치를 제거하였다.

3.2 학습데이터셋(Training Data Set) 구성

3.2.1 클러스터링 결과 파생 속성 추가

설비 이상 탐지 모델의 구현을 위해 학습데이터셋 구성은 매우 중요한 요소이다. 하지만 설비의 노후화나 설비 정비 여부에 따라 설비 데이터 변동 가능성이 있기 때문에 설비 센서 데이터에만 의존하는 것은 데이터 자체에서 의미 있는 정보를 찾기가 어려운 점이 있다. 따라서 본 연구에서는 보다 데이터의 설명력을 높이기 위해 미가공 데이터에 k-Means 클러스터링을 적용한 결과, 기존 데이터 셋에 클러스터 특성 속성을 학습데이터셋에 포함하였다. 이때, k-Means 클러스터링 알고리즘은 사용자가 클러스터의 개수를 미리 지정해야 하는데 엘보우 기법(Elbow Method)을 이용하여 적절한 클러스터 개수를 설정했다. 엘보우 기법은 최적의 클러스터 개수를 구하기 위한 알고리즘으로 알려져 있으며, 한 개의 클러스터를 추가했을 때, 추가하기 전보다 특정 범위 값을 넘어서는 더 좋은 결과가 나타나지 않으면 이전 클러스터의 개수를 최적의 클러스터 개수로 설정한다[10].

3.2.2 학습구간 설정

본 연구에서는 분류 기반의 학습 모델이 사용되었으며 정상 구간과 이상 구간에 대한 설정은 모델 결과에 큰 영향을 준다. 이는 단순 설비의 이상 예측이 아닌 고장 시점 예측과도 연결된다. 데이터 전처리와 클러스터링이 포함된 데이터셋을 기준으로 학습 구간 설정이 필요하다. 또한 과거 데이터에 대한 즉 설비의 과거 상태에 대해서 전 기간 알 수 없기 때문에 고장 이력을 중심으로 학습 데이터 구간을 설정한다. 학습 데이터 구간 설정은 크게 2가지로 설정하였으며, Case_1은 정상 구간 30일과 비정상 구간 10일, Case_2는 정상 구간 90일과 비정상 구간 20일로 설정하였다.

3.3 주요속성 추출 및 위험 스코어 예측

본 연구에서는 설비 위험도를 예측하기 위해 앙상블 모델인 EGB(Extreme Gradient Boosting) 알고리즘을 사용하였다. EGB 알고리즘은 다양한 데이터 분석 대회에서 좋은 성능을 보여주고 있는 의사결정나무 기법으로[1, 17], 과적합을 방지하는 장치를 가지고 있다[3]. EGB 알고리즘은 CART 알고리즘으로 선행 학습 후 성능 개선을 위해 CART 알고리즘의 에러에 대해 반복 학습한다. t단계에서의 예측을 위한 함수는 (4)과 같다[18].

$$f_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = f_i^{(t-1)} + f_t(x_i) \tag{4}$$

x_i는 입력 데이터이고 $f_t(x_i)$는 t단계에서의 예측 치이다. $f_i^{(t)}$와 $f_i^{(t-1)}$는 각각 t단계, t-1단계까지의 예측치를 결합한 결과이다. 모델링의 빠른 계산과 과적합 방지를 위해 우도식(Goodness Function)을 식 (5)와 같이 사용한다.

$$obj^{(t)} = \sum_{k=1}^{n} l(\overline{y_i}, y_i) + f_i^{(t-1)} + \sum_{k=1}^{n} \Omega(f_i) \tag{5}$$

l은 loss 함수이고 n은 데이터 개수이다. Ω는 정규화 항으로 (6)으로 정의 된다.

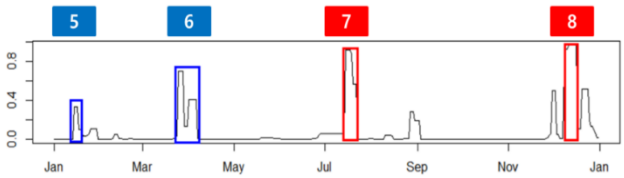
$$\Omega(f) = \gamma^T + \frac{1}{2}\lambda\|\omega\|^2 \tag{6}$$

ω는 Leaf 노드의 스코어고, λ는 정규화 매개속성이다. γ는 Leaf 노드의 분할을 위해서 사용되는 최소 loss값이다[6]. EGB 알고리즘은 예측치를 계산하면서 활용된 속성의 중요도를 판별해준다. CART 모델의 노드에 포함되는 속성 테스트를 Information Gain 값으로 하게 되는데, 이 과정에서 상대적으로 중요하지 않은 속성은 모델에 포함이 되지 않아 자연적으로 중요 속성이 선택되는 결과가 나타난다. 설비데이터와 파생 속성과 설비 고장 유무(Class 속성)을 포함한 학습데이터셋을 기반의 학습 모델을 통해 고압 압축 설비의 위험상황 탐지 및 예측을 실시하였다. 설비 고장 위험 스코어 예측에서 중요한 포인트는 설비의 이상 전에 전조 시그널의 확인 가능 여부이다. 따라서 학습 시 모델의 위험 시그널에 대한 결과에서 전조 현상이 잘 보여지고 있는가에 대한 확인이 필요하다. <Table 4>는 아래 표와 같이 예측 모델의 고장 이력(예시)을 학습하고 있다.

<Table 4> Data Set Composition Example by Machine Failure History

No.	Div.	Event Time	Machine
1	Training	2012-07-**	A
2		2013-06-**	A
3		2014-02-**	A
4		2014-06-**	A
5	Test	2015-02-**	B
6		2015-04-**	C
7		2015-06-**	A
8		2015-11-**	A

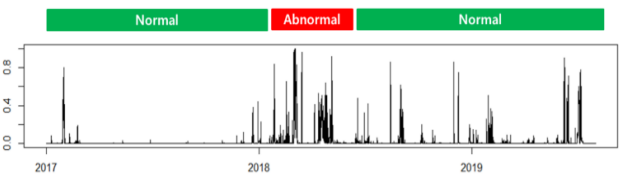
아래 <Figure 2>는 학습된 설비 위험 예측 모델에테스트 데이터를 예측한 결과이다. 테스트 데이터 적용 결과 해당 설비의 고장 시점인 '7'과 '8' 구간에 설비의 위험 예측에 대해서 위험 스코어가 상승되어 있는 것을 확인할 수 있다. 추가적으로 타 설비의 고장인 '5'와 '6' 시점에서도 설비 위험도가 일부 상승 되는 것을 확인할 수 있는데 이는 연계 설비의 고장에 대해서도 일부 예측이 가능 한 것을 확인 할 수 있다.



<Figure 2> Machine Risk Prediction Model Test Results

3.4 모델 검증

학습데이터셋 기반으로 설비의 위험 예측 모델을 생성하고 테스트 데이터를 기반으로 모델을 검증 하였다. 검증 방법으로는 테스트 데이터를 정상구간과 비정상 구간으로 나누고 EGB 모델의 예측 값과의 비교를 통해 예측 정확도를 확인하였다. <Figure 3>은 설비 위험 예측 모델의 정상과 비정상 구간을 분류한 예시이다. 모델 예측치에 대한 평가는 ROC Curve의 AUC(Area under the Curve) 값으로 판단하였다.



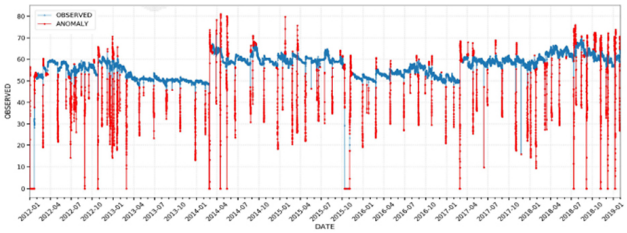
<Figure 3> Example of Classification of Normal and Abnormal Intervals in the Machine Risk Prediction Model

4. 실험 결과

본 연구에서 제시한 방법을 통해 화학 공정의 고압압축 설비의 위험을 예측하고 선택된 주요 속성을 통해 설비 고장 원인 파악에 이용하였다.

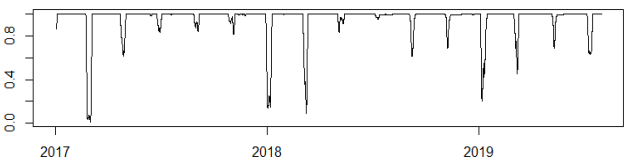
4.1 전처리 결과

아래의 <Figure 5>는 제시된 방법론 중 전처리 과정인 센서 데이터의 STL 및 GESD 적용 결과를 보여주고 있다.



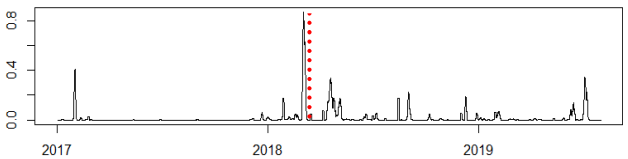
<Figure 5> Sensor Data STL and GESD Result

이러한 STL 및 GESD 알고리즘을 통해 계획정비(PSD) 기간에 발생한 이상치 데이터와 기타 이상치에 대한 전처리 작업을 실시하여 노이즈(Noise)를 제거하였다. 노이즈를 제거하지 못 할 경우 잘못된 학습이 이루어지고 모델의 결과도 상당부분 차이를 보인다. 아래 <Figure 6>과 <Figure 7>에서는 데이터셋 전처리 전, 후에 대한 예측 스코어를 나타내고 있다. 테스트셋 전처리 전의 경우 상당기간 이상스코어가 최대값인 1(고장)로 학습된 결과를 확인 할 수 있다.



<Figure 6> Prediction Score before Preprocess

반면에 전처리 후 예측 스코어의 경우, 2018년 3월(점선)의 설비 고장 징후를 사전에 예측하여 모델이 정상적인 설비 위험 예측이 가능함을 확인 할 수 있다.



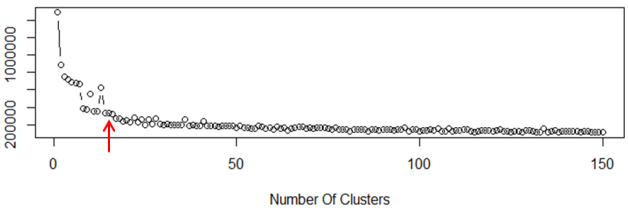
<Figure 7> Prediction Score after Preprocess

이와 같은 결과는 전처리 시 이상치 제거 작업의 중요성에 대해 확인 할 수 있으며, 정상 구간이 이상 구간으로 학습 되거나 이상치 구간이 고장 위험구간으로 학습되는 경우 데이터 정합성이 현저하게 떨어지는 것을 확인 할 수 있다.

4.2 학습 데이터셋 구성 결과

4.2.1 클러스터링 결과 파생 속성 추가

학습 모델의 예측 정확도와 설명력을 향상시키기 위해 학습데이터셋 구성 시 비지도학습 기반의 클러스터링 결과를 파생속성으로 추가하였다. 아래 <Figure 8>은 학습 셋의 k-Means 클러스터링 개수를 정하기 위한 엘보우 기법 적용 결과이다. 이 파생 속성은 각 인스턴스가 클러스터에 속하는지를 나타내는데, 각 클러스터는 속성의 특징에 따라 어떤 특징을 나타내는가를 판별하고 설명하는 역할 뿐 만 아니라 고장 유무를 분류하는 성능 향상에 중요한 역할을 할 수 있다.



<Figure 8> Choosing Clustering k

<Table 5>은 클러스터링을 적용한 학습데이터셋의 일부이다.

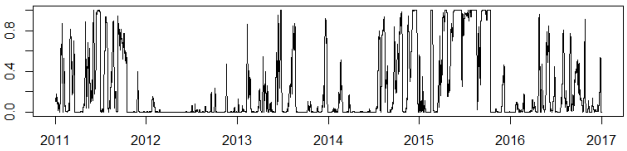
<Table 5> Training Set Example

V1	V2	V3	·	V80	Cluster
478.62	9.2	10.96	·	50.39	2
478.34	9.15	10.95	·	50.83	2
478.44	9.19	10.96	·	51.14	2
479.28	9.19	10.95	·	51.38	4
479.22	9.19	10.95	·	51.72	4
478.96	9.18	10.95	·	51.67	5

위 <Table 5>와 같이 학습데이터셋에 클러스터링 결과값을 센서 데이터와 통합하였지만 실제 정수형 값을 의미하는 것은 아니다. 따라서 명목형(Category) 속성으로 분류해야 하며 해당 값을 그대로 모델에 적용하게 되면 의미 없는 결과를 도출하게 된다. 명목형 속성인 Cluster를 더미(Dummy)속성으로 변환한 결과 클러스터링 14개 속성으로 더미 변환되었다.

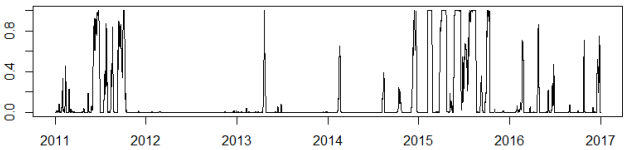
4.2.2 학습 구간 설정

분류 기반의 학습 모델이 사용되었기에 정상 구간과 비정상 구간에 대한 학습데이터셋 분류 설정은 모델 결과에 큰 영향을 준다. 설비의 과거 상태에 대해 모든 기간을 알 수 없기 때문에 고장 이력을 중심으로 학습 데이터 구간을 설정 하였다. 학습 데이터 구간 설정은 크게 2가지로 설정하였으며, Case_1은 정상 구간 30일과 비정상 구간 10일, Case_2는 정상 구간 90일과 비정상 구간 20일로 설정 하였다. 아래 <Figure 9>와 <Figure 10>에서는 학습 구간에 따른 스코어를 비교한 결과이다.



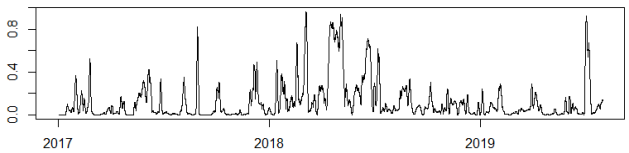
<Figure 9> Learning Result for Case_1

학습데이터셋의 2011년도부터 2017년까지의 결과를 확인해보니 학습 기간에 따른 스코어 차이가 상당 부분 있는 것으로 확인되었다. Case_1 경우 Case_2에 비해서 고장 위험 예측 스코어가 더 높은 경우가 많음을 확인 할 수 있다.

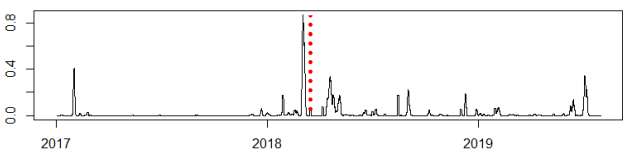


<Figure 10> Learning Result for Case_2

아래 <Figure 11>와 <Figure 12>에서와 같이 테스트 결과 역시 Case_1 경우 Case_2에 비해서 스코어가 상향되었으며, 2018년 3월(점선) 이상 발생 외에도 상당 부분 시그널이 발생하는 것을 볼 수 있었다.



<Figure 11> Test Result for Case_1



<Figure 12> Test Result for Case_2

4.3 주요 속성 추출 및 위험스코어 예측 결과

4.3.1 주요 속성 추출

본 연구에서는 모델인 EGB(Extreme Gradient boosting) 알고리즘을 사용하여 설비 위험의 주요 속성을 판별하였다. 비교를 위해 Case_2(정상 구간 90일과 비정상 구간 20일)와 Case_3(정상 구간 90일과 비정상 구간 30일)로 구분하였다. 아래<Table 6>는 Case_3에 대한 결과이며 클러스터링 속성 포함 여부에 따른 추출된 주요 속성을 비교하였다.

<Table 6> Key Feature Selection Results for Case_3

Div.	Include Cluster		Exclude Cluster	
Rank	Feature	Gain	Feature	Gain
1	V04	0.404	V04	0.405
2	V10	0.133	V10	0.135
3	V19	0.122	V19	0.122
4	V09	0.087	V09	0.090
5	V03	0.085	V03	0.085
.	.	.	.	.
20	V20	0.002	V20	0.001
21	Cluster3	0.001		
22	Cluster10	0.001		
23	Cluster8	0.001		
24	Cluster4	0.001		

Case_3의 경우 클러스터링 속성 3, 10, 8, 4의 속성이 중요속성으로 포함이 되었다. 아래 <Table 7>는 Case_2에 대한 결과이며 클러스터링 속성 포함 여부에 따른 추출된 주요 속성을 비교하였다.

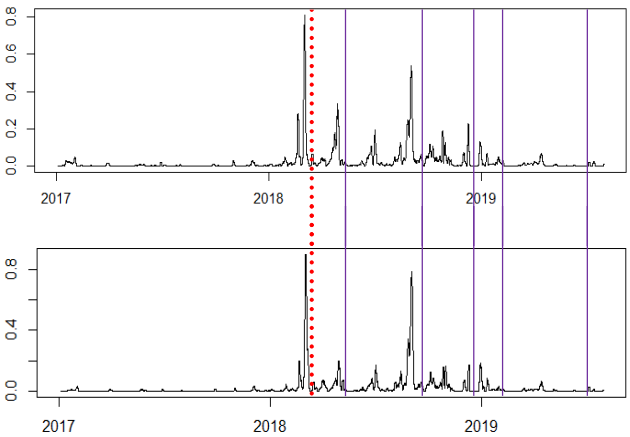
<Table 7> Key Feature Selection Results for Case_2

Div.	Include Cluster		Exclude Cluster	
Rank	Feature	Gain	Feature	Gain
1	V03	0.320	V03	0.329
2	V10	0.188	V10	0.195
3	V04	0.118	V04	0.109
4	V20	0.047	V14	0.056
5	V01	0.041	V20	0.045
6	V02	0.040	V09	0.044
7	V11	0.040	V02	0.044
8	V09	0.037	V01	0.044
9	V13	0.034	V11	0.040
10	Cluster4	0.034	V13	0.033
11	V19	0.033	V19	0.026

Case_2의 경우 클러스터 4의 속성이 10번째 중요속성으로 포함이 되었다. Case_3의 결과에 비해 높은 중요도를 보이고 있으며 다른 속성 중요도 순서 역시 변화가 보였다. 이는 클러스터링의 파생 속성의 영향도가 모델의 전체 속성 중요도에 영향을 미치는 것으로 판단된다.

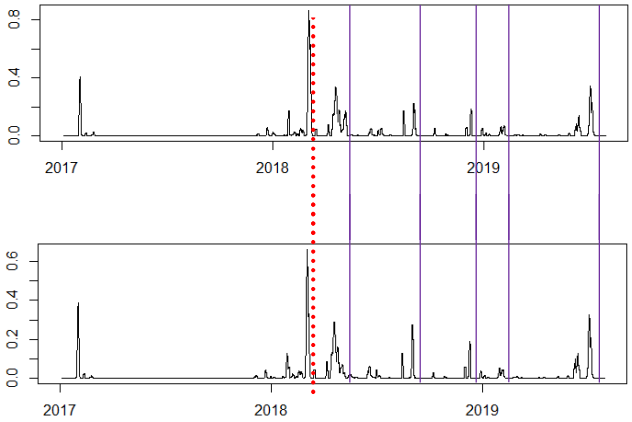
4.3.2 위험스코어 예측

고장 위험 예측 모델의 학습 구간은 2011년 1월부터 2016년 12월까지 총 6년으로 설정하였으며, 테스트 구간은 2017년 1월부터 2019년 7월까지로 설정하였다. 학습구간 설정 값은 비정상구간과 정상구간으로 분리하였고, 설비 고장 발생 여부에 따라 총 4개 구간으로 2015년 2월, 4월, 6월, 8월로 구성하였으며, 정상구간은 총 5개 구간으로 2012년 6월, 2013년 9월, 2015년 2월, 2015년 10월, 2016년 12월로 구성하였다. 테스트 구간의 설비 고장 이력은 2018년 3월이다. 본 연구에서는 학습구간 설정 간의 비교, 클러스터 속성 적용 여부에 따른 위험스코어 비교를 통해 테스트셋의 설비 고장에 대해 사전에 포착이 가능한지를 확인하였다.



<Figure 13> Comparison of 3-day Moving Average Prediction Score According to Whether Cluster Attribute is Included (lower) or Not (upper) for Case_3

위 <Figure 13>과 <Figure 14>는 시간당 예측 결과값에 3일 이동평균을 적용한 결과이다. 이동평균을 적용한 결과 시간당 예측 결과에 대한 해석이 용이해진 것으로 확인되었다. 그래프 상의 점선은 설비의 실제 고장 시점을 나타내고 있으며 실선은 연계설비의 고장 시점을 나타낸다. 예측결과 제시된 모델은 연계 설비의 고장에 대해서 영향을 받는 것으로 판단되며, 단독 설비의 고장 위험 시그널을 통해 연계 설비의 고장 시그널에 대해서 일부 파악이 가능한 것으로 보인다.



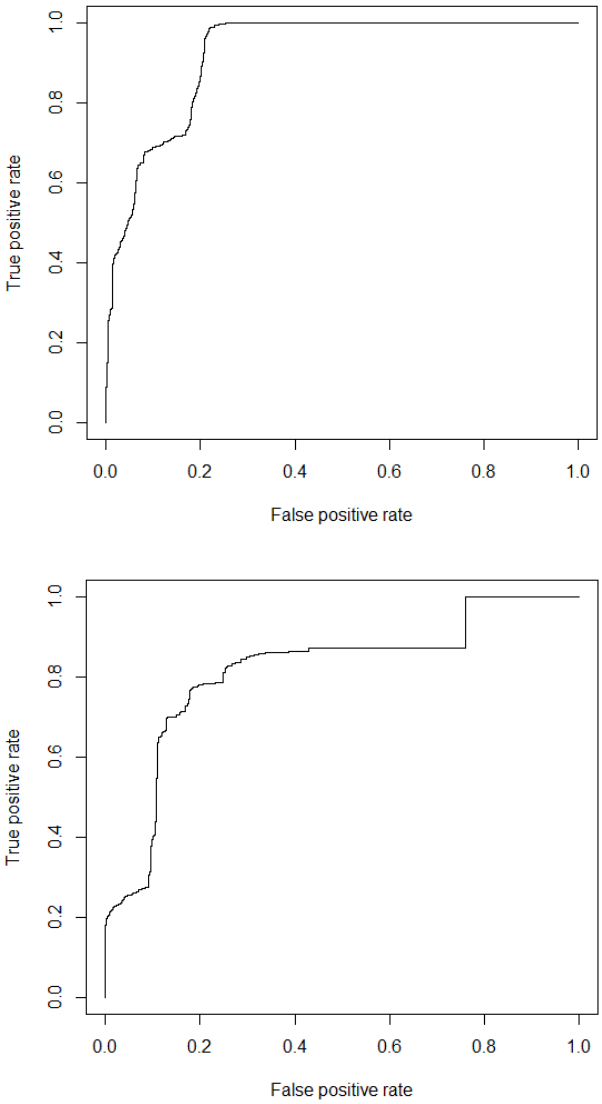
<Figure 14> Comparison of 3-day Moving Average Prediction Score According to Whether Cluster Attribute is Included (lower) or Not (upper) for Case_2

학습 구간 및 클러스터링 적용 유무에 따른 4가지 위험 예측 모델은 2018년 3월(점선)의 설비 고장을 사전에 예측하고 있다. 설비 고장이 발생하기 20일~40일 전에 스코어가 상승하는 것을 확인 할 수 있으며 시그널이 상승 될 경우 설비의 상태가 이상하다거나 설비 점검 시기가 근접해 있음을 인지 할 수 있을 것으로 판단된다.

4.3.3 위험스코어 예측 정확도

본 연구에서는 설비 위험을 예측하기 위해 4가지 케이스 비교를 통해 설비의 위험에 대한 예측 모델이 성능을 확인하였다. 또한 실제 스코어의 예측 정확도에 대한 비교를 통해 가장 정확도가 높은 케이스와 모델의 검증을 실시하였다. 검증 방법으로는 두 개의 범주를 갖는 속성을 예측하는 분류 모델에 대한 성능 평가 방법으로 ROC를 이용하였다. ROC는 x축의 값이 0일 때, y축의 값은 0이 되며, x축이 증가할수록 y축이 증가한다. 완전히 랜덤하게 자료를 분류한 경우라면, ROC 곡선은 원점을 통과하는 기울기 1인 직선이 된다. 분류 모델의 성능이 랜덤한 예측 보다 좋은 경우 그 분류기의 ROC 곡선은 원점을 통과하는 기울기 1인 직선보다 위에 위치한다. x축의 값을 고정한 경우 y축 값이 큰 ROC 곡선이 성능이 좋다고 할 수 있다. 모든 경우의 결과를 완벽하게 예측하는 모델인 경우 AUC는 1, 무작위로 예측한 모델과 별 차이가 없는 경우 AUC는 0.5의 값을 가지게 된다[6].

<Figure 15>는 Case_2와 Case_3모델의 ROC 곡선을 비교한 결과이다. 비교 결과는 <Table 7>에서와 같이 비정상 30일, 클러스터 속성 포함 학습 모델이 AUC 0.923로 가장 높은 값을 가지는 모델로 확인되었다.



<Figure 15> ROC Curve of Two Models(Case_3-Upper, Case_2-Lower)

<Table 7> Comparing AUC Results from Prediction Models

Rank	Condition	AUC
1	Abnormal 30Days, Cluster Feature Included	0.923
2	Abnormal 30Days, Cluster Feature Excluded	0.907
3	Abnormal 20Days, Cluster Feature Excluded	0.867
4	Abnormal 20Days, Cluster Feature Included	0.815

학습시 비정상에 대한 기간과 클러스터링 적용 여부에 따라서 모델 정확도에 대해 일부 차이가 발생하였음을 확인 할 수 있었다. 이는 클러스터 파생 속성 데이터셋이 그렇지 않은 데이터셋(Only Sensor Data) 에서 가지고 있지 않은 정보를 가지고 있기에 이런 결과가 도출되었다고 판단된다.

5. 결 론

본 연구에서는 화학공장의 특수설비인 고압 압축기를 대상으로 설비의 위험 또는 고장을 예측하였다. 제안하는 예측모델은 설비 상태에 대한 객관적인 정보가 없거나 부정확한 과거 이력 정보들에 대한 신뢰도를 높이기 위해서 적용한 STL 기반의 전처리 기법을 적용하였다. 이 전처리 기법을 통해 센서 시계열 데이터의 이상치를 적절하게 필터링 하였음을 확인 할 수 있었다. 또한 높은 예측 성능의 EGB 모델을 적용하였고 더 높은 예측 정확도와 모델의 설명력 향상을 위해 클러스터링 결과의 파생 속성 적용 방법을 제시하였다. 이 예측 모델은 고장 이력을 중심으로 설계된 학습데이터셋을 기반으로 설비 고장에 대한 정도와 시점에 대해 예측이 가능한 것을 보여 주었다.. 또한 제시된 예측모델은 설비 이상에 대한 주요 속성 추출을 통해 고장의 원인을 파악하는데 도움을 줄 수 있었다.

Acknowledgement

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(2017R1D1A1B03034475).

References

[1] Adam-Bourdarios et al., The Higgs boson machine learning challenge, NIPS 2014 Workshop on High-energy Physics and Machine Learning, 2015.

[2] Aggarwal, C., Outlier analysis, Springer, Switzerlnd, 2017.

[3] Babajide Mustapha, I. and Saeed, F., Bioactive molecule prediction using extreme gradient boosting, *Molecules*, 2016, Vol. 21, No. 8, p. 983.

[4] Chandola, V., Banerjee, A., and Kumar, V., Anomaly detection : A survey, *ACM Computing Surveys(CSUR)*, 2009, Vol. 41, No. 3.

[5] Domingues, R., Filippone, M., Michiardi, P., and Zouaoui, J., A comparative evaluation of outlier detection algorithms : Experiments and analyses, *Pattern Recognition*, 2018, Vol. 74, pp. 406-421.

[6] Fan et al., Comparison of Support Vector Machine and Extreme Gradient Boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates : A case study in China, *Energy Conversion and Management*, 2018, Vol. 164, pp. 102-111.

[7] Hadi, A.S., Identifying multiple outliers in multivariate data, *Journal of the Royal Statistical Society : Series B (Methodological)*, 1992, Vol. 54, No. 3, pp. 761-771.

[8] In to the data, http://intothedata.com/02.scholar_category/anomaly_detection/.

[9] Jianliang, M., Haikun, S., and Ling, B., The application on intrusion detection based on k-Means cluster algorithm, *2009 International Forum on Information Technology and Applications*, 2009, Vol. 1, pp. 150-152.

[10] Joseph, M.P., A PD Validation Framework for Basel II Internal Ratings-Based Systems, Credit Scoring and Credit Control IV, 2005.

[11] Knorr, E.M. and Ng, R.T., Finding intensional knowledge of distance-based outliers, in Proceedings of 25th International Conference on Very Large Databases, 1999.

[12] Laskov, P., Dussel, P., Schafer, C., and Rieck, K., Learning intrusion detection : supervised or unsupervised?, *In International Conference on Image Analysis and Processing*, 2005, pp. 50-57.

[13] Liu, F.T., Ting, K.M., and Zhou, Z.-H., Isolation-based anomaly detection, *ACM Transactions on Knowledge Discovery from Data*, 2012, Vol. 6, No. 1, pp. 1-39.

[14] Markou, M. and Singh, A., Novelty detection : a review-part 1: statistical approaches, *Signal Processing*, 2003, Vol. 83, No. 12, pp. 2481-2497.

[15] Markou, M. and Singh, A., Novelty detection : a review-part 2 : neural network based approaches, *Signal Processing*, 2003, Vol. 83, No. 12, pp. 2499-2521.

[16] Origins of the NIST/SEMATECH e-Handbook of Statistical Methods in the Work of Mary Natrella, http://www.itl.nist.gov/div898/handbook/, 2003.

[17] Phoboo, A.E., Machine learning wins the Higgs challenge, 2014, No. BULNA-2014-265.

[18] Punnoose, R. and Ajit, P., Prediction of employee turnover in organizations using machine learning algorithms, International Journal of Advanced Research in Artificial Intelligence, 2016, Vol. 5, No. 9, pp. 22-26.

[19] Robert et al., STL : A Seasonal-Trend Decomposition Procedure Based on Loess, *Journal of Official Statistics*, 1990, Vol. 6, No. 1, pp. 3-73.

ORCID

Kang Min Cheon | http://orcid.org/0000-0001-8877-2870

Jaekyung Yang | http://orcid.org/0000-0002-4904-1351