JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 25, No. 6: 749~756, Jun. 2021

머신러닝을 활용한 제품 특성 예측모델의 성능향상 방법 연구

김종훈[1] · 오하영[2*]

The methods to improve the performance of predictive model using machine learning for the quality properties of products

Jong Hoon Kim[1] · Hayoung Oh[2*]

[1]Graduate Student, Applied Data Science, Sungkyunkwan University, Seoul, 03063 Korea
[2*]Associate professor, College of Computing & Informatics, Sungkyunkwan University, Seoul, 03063 Korea

요 약

제조 생산공정에는 다양한 센서를 통해 실시간으로 양질의 데이터가 데이터베이스에 축적되고 있다. 이와 함께 통계적으로 접근하기 까다로운 데이터에 대해서 높은 수준의 정확도로 예측모델을 구축할 수 있는 머신러닝이 보급되면서 '4차 산업화 시대'를 맞이하고 있다. 본 논문에서는 이러한 제조업계의 흐름에 따라 업계의 주요 관심사인 제품의 품질특성을 예측하는 머신러닝 모델의 성능을 향상하는 방법을 제시한다. 머신러닝 모델의 성능을 향상하는데 일반적으로 사용되는 샘플 크기의 증가, Hyper-Parameter의 최적화 및 적절한 알고리즘 선택의 효과를 검증한다. 그리고, 새로운 성능향상 방법을 제시하고, 그 효과를 검증해본다. 논문에서 제시한 방법을 통해서 제조업에서는 더욱 향상된 성능의 예측모델을 구축, 품질예측과 관리에 크게 이바지할 수 있을 것이다.

ABSTRACT

Thanks to PLC and IoT Sensor, huge amounts of data has been accumulated onto the companies' databases. Machine Learning Algorithms for the predictive model with good performance have been widely utilized in the manufacturing process. We present how to improve the performance of machine learning predictive models. To improve the performance of the predictive model, typical techniques such as increasing the sample size, optimizing the hyper parameters for the algorithm, and selecting a proper machine learning algorithm for the predictive model would be shown. We suggest some new ways to make the model performance much better. With the proposed methods, we can build a better predictive model for predicting and controlling product qualities and save incredibly large amount of quality failure cost.

키워드 : 머신러닝, 머신러닝 모델, 예측성능, 예측모델 성능향상

Key Words : Machine learning, Predictive performance, Machine learning model, Model performance

Received 6 March 2021, **Revised** 16 April 2021, **Accepted** 21 April 2021

*** Corresponding Author** Hayoung Oh(E-mail:hyoh79@skku.edu, Tel:+82-2-740-1781)
Associate professor, College of Computing & Informatics, Sungkyunkwan University, Seoul, 03063 Korea

Open Access http://doi.org/10.6109/jkiice.2021.25.6.749 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

1.1. 선행연구

디지털 제조 분야 글로벌 선두 그룹인 지멘스 그룹의 조 케저 회장은 다보스 어젠다2021 '성장을 위한 제조업 재구상' 세션에서 제조업 혁신의 '게임체인저'로 머신러닝 기술을 꼽았다. 케저 회장은 "향후 5년간 제조 분야에서 가장 큰 변화는 머신러닝이 결정할 것이다." 라고 말했다.[1]

한국은 최근 경제성장의 정체로 인한 고용률 하락이 산업계 전반에서 나타나고 있다. 고용정보원은 보고서를 통해 4차 산업혁명에 따른 기술혁신과 빠른 성장이 예측되는 제조업 분야는 인력 수요가 계속해서 발생할 것으로 전망했다.[2]

'2003년 SARS 발발, 2008년 골드만 삭스로 촉발된 글로벌 금융위기, 2020년 COVID-19 등의 전 세계적인 충격 속에서도 디지털화와 AI를 통한 혁신을 추진한 기업들의 매출과 이익은 모두 증가했다'라는 것이 보스턴 컨설팅 그룹의 최근 연구 'The Rise of the AI-Powered Company in the Postcrisis World'를 통해서 증명되었다.[3]

최근 ICT 기술의 급속한 발전과 이를 활용한 광범위한 양질의 데이터 축적을 기반으로 통계적 분석의 한계를 극복할 수 있는 좋은 대안으로써 머신러닝은 제조업에서 큰 주목을 받고 있다. 또한, 회귀와 분류의 영역에서 셀 수 없을 정도의 다양한 알고리즘이 제공되어 개별 데이터의 특성에 맞는 성능이 높은 예측모델을 선택할 수 있다. 이에 본 논문에서는 머신러닝 예측모델의 성능을 향상할 수 있는 다양한 방법을 제안한다.

현재 가장 일반적으로 사용되고 있는 방법인 ①샘플 크기의 증가, ②적절한 머신러닝 알고리즘의 선택, ③ Hyper-Parameter의 최적화를 통해서 예측모델의 성능 향상을 확인한다.[4] 일반화된 성능향상 방법 외 새로운 접근을 본 연구에서 시도하고자 한다.

1.2. 연구 동향

기존 예측모델의 성능 연구를 고찰해 본 결과, 표 1에서 나타난 것과 같이 Linear Regression, Random Forest, SVM, KNN, DNN이 가장 폭넓게 사용되고 있음을 확인하였다. 해당 5가지 알고리즘을 본 연구에서 사용할 것이다. 머신러닝 예측모델의 성능향상 방법은 많은 연구에서 다뤄지고 있지만, 제조업에서는 고객의 요구 특성에 대한 예측성능이 곧 시장에서의 고객 만족 및 품질 비용으로 직결될 수 있는 사안이므로 기존의 방법 외 추가적인 성능향상 방법에 대한 요구가 매우 강하다. 이러한 제조업에서의 경향을 반영하여 본 논문에서는 머신러닝을 이용하여 제조공정에서의 품질특성을 예측하는 모델을 수립하고, 이의 예측성능을 향상하는 다양한 방법을 제시한다.

본 논문은 I장 서론을 포함하여 5개의 장으로 구성되어 있다. Ⅱ장에서는 연구에 사용된 머신러닝 알고리즘에 대한 소개, Ⅲ장에서는 연구대상과 데이터 전처리를 설명하고, Ⅳ장은 머신러닝 예측모델을 수립하여 기존에 알려진 방법과 함께 일반에 잘 알려지지 않은 방법을 통해서 예측성능을 측정, 비교하였다. Ⅴ장에서는 결론과 시사점을 기술한다.

Table. 1 Tasks of Past Studies

Task	Goal	Model
Regression	The Deformation Prediction[5]	Random Forest, DNN
Classification	Sillicon Wafer Micro-Cracks Prediction[6]	SVM
Classification	Surface Defect Prediction[7]	KNN
Regression	Hot Deformation Prediction[8]	DNN, SVM
Classification	Fine Dust PM10 Prediction [9]	Logistic Regression SVM, DNN, Random Forest
Regression	Fruit Size Prediction[10]	Linear Regression

Ⅱ. 머신러닝 알고리즘

2.1. Linear Regression Model

입력과 출력의 인과관계를 선형 방정식의 형태로 나타낸 것이다. 모델링의 대상이 되는 데이터(입력)가 정규성을 가지면 유용하다. 모델은 원인과 결과를 통계적으로 명확하게 설명하므로, 데이터를 통해서 그 데이터가 생성된 공정의 이해도 가능하게 된다. 다만, 데이터가 정규성, 등분산성에 어긋나거나 혹은 비정형 데이터가 포함되었으면 사용하기 매우 어렵다. 또한, 평균과

분산을 사용하므로 이상치의 영향을 크게 받는다. 이외 모델에 포함된 설명변수 간의 독립성이 어긋나면 예측력을 신뢰할 수 없다. [11]

2.2. Bagging : Random Forest

Random Forest는 Decision Tree를 바탕으로 Bagging의 개념을 적용한 알고리즘이다. 아래의 그림 1에서 살펴볼 수 있듯이 데이터셋에서 임의 복원 샘플링을 통해 다수의 데이터셋(부분집합)을 생성한다. 이것을 Boostrap 샘플링이라고 한다. 샘플링한 다수의 데이터셋으로 각각 예측모델을 수립하여 분류의 경우에는 다수결투표, 회귀의 경우에는 각 모델의 예측값들에 대한 평균값을 최종값으로 출력한다. 이러한 과정을 Aggregating이라고 한다. 이러한 Boostrap 샘플링과 Aggregating을 함께 사용한 알고리즘이 Bagging이다.[12] Bagging을 적용한 대표적인 예시가 Random Forest이다. Bagging은 모델의 편의는 유지되고, 특정 데이터에 Overfitting되는 것을 방지하며 분산을 감소시킬 수 있는 장점이 있다.

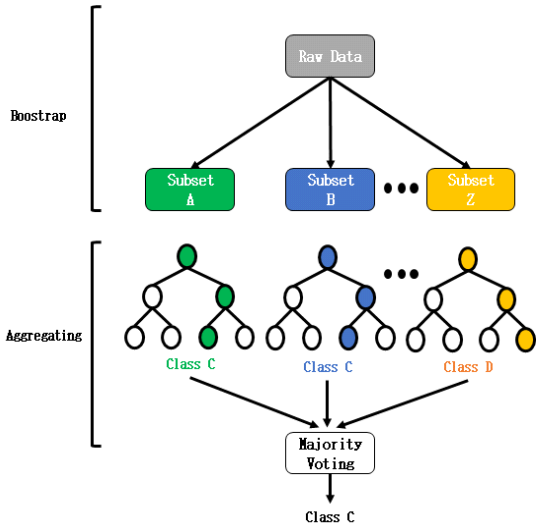

Fig. 1 Bagging : Random Forest Algorithm

2.3. SVM(Support Vector Machine)

SVM은 그림 2에서 볼 수 있듯이, 현재의 데이터를 보다 고차원(N차원)으로 매핑시킨 뒤에 N-1차원의 초평면으로 분류하는 기법이다. 특히, 비선형 분류 문제에서는 이를 높은 차원의 공간으로 변환시켜 선형적인 구분을 가능케 하는 결정경계를 쉽게 찾을 수 있도록 커널함수를 사용한다. 또한, 이러한 과정에서 잘못 분류된 데이터에 대해서는 일종의 Penalty를 부여하여 손실함수를 조정할 수 있다.

비교적 작은 데이터셋에서도 예측성능이 우수하다. 그러나, 데이터의 용량이 큰 경우에는 처리시간이 길어지는 단점도 수반한다. SVM은 Random Forest와 비교해서 Hyper-Parameter의 최적화가 까다롭다.

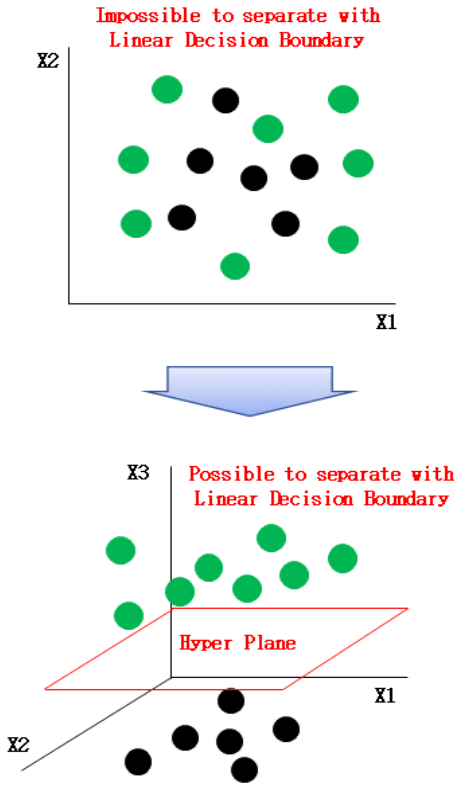

Fig. 2 SVM(Support Vector Machine)

2.4. KNN(K-Nearest Neighbors)

KNN은 그림 3과 같이 새로운 데이터가 입력되면, 그 데이터에서부터 K개의 가장 가까운 이웃들을 찾는다. 그 K개의 이웃이 어떤 클래스에 속하는지 파악하고, 다수결에 따라서 새로운 데이터의 클래스를 분류하는 기법이다. KNN의 경우에는 학습 과정이 없이 입력 데이터와 학습 데이터의 거리 측정을 통해 예측을 시행한다. 즉, 예측 모델을 별도로 수립하는 과정이 없다.

KNN은 학습 데이터 내에 포함된 노이즈 혹은 이상치의 영향을 크게 받는 것으로 알려져 있으므로 전처리 과정에서 이에 대한 충분한 검토와 제거가 필수적이다.

또한 높은 차원의 데이터셋에는 성능이 낮다고 알려져 있으므로, Feature Selection을 통해서 중요한 특성들의 선별이 선행되어야 한다.

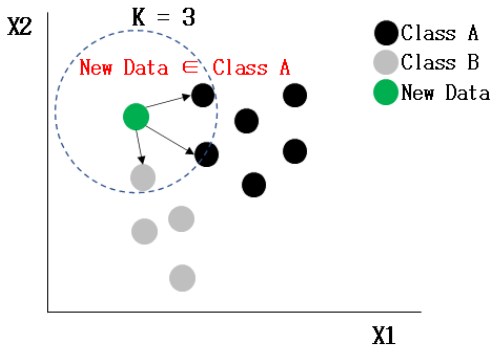

Fig. 3 KNN(K-Nearest Neighbors)

2.5. DNN(Deep Neural Networks)

DNN은 그림 4에서 보는 것과 같이 Input, Hidden, Output의 3단 구조가 기본적이다. Input Layer는 데이터를 입력하는 층이다. 이것이 Hidden Layer로 전달되어 주요한 특성들이 추출되며, Output Layer에서 클래스별 예측확률이 결정된다.

DNN에서는 앞서 언급한 Layer의 수와 각 Layer의 노드의 수를 적절하게 선정하는 것이 중요하다. 일반적으로는 복잡한 이미지의 경우에는 비교적 많은 수의 Layer와 Node를 쌓는 것이 유리하고, 정형적인 데이터의 경우에는 작은 수의 Layer와 Node를 갖도록 설계한다.

DNN은 머신러닝 알고리즘과 비교하면 고려해야 할 Hyper-Parameter가 매우 많다. 주요한 것으로는 Layer의 개수, Layer 별 Node의 수, 각 Layer 별 활성화 함수, Regularization, 최적화 함수 및 Learning Rate 등이 있다.

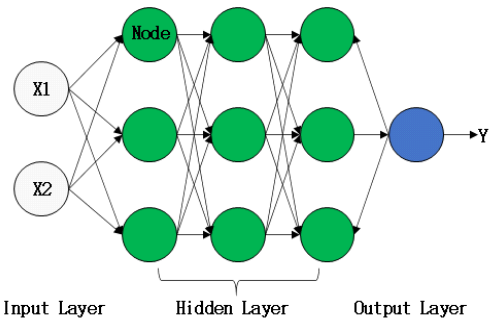

Fig. 4 DNN(Deep Neural Networks)

Ⅲ. 연구 방법

본 논문의 목적은 머신러닝 예측모델의 성능을 향상하는 효과적인 방법을 찾는 것이다. 제조공정의 데이터베이스에 축적된 10개월간의 데이터를 사용할 것이다. 전처리에서는 결측치의 제거, 영분산 변수의 제거, 중요변수의 선택 등이 시행되었다. 전처리 된 데이터셋을 이용하여 다양한 성능향상 방법의 효과를 검증한다.

3.1. 연구대상

본 논문에서는 건축용 자재 제조업체 LG Hausys사의 건축구조재의 제조공정 데이터를 사용하였다. 공정의 생산속도, 투입되는 화학물질의 배합비 및 열처리 온도 등 총 31개의 변수(30개의 설명변수, 1개의 반응변수)로 구성되어 있다. 데이터의 수집 기간은 2020년 2월 5일부터 2020년 12월 3일까지이다.

3.2. 데이터 전처리 과정

3.2.1. 결측치 제거

총 750개의 결측치가 존재하며, 특정 기간 동안 주요 설명변수 10개에서만 결측치가 발생했다. 그래서 결측치가 많은 변수를 제거하지 않고 결측치가 존재하는 행을 제거하였다.

3.2.2. 영분산 변수의 제거

총 30개의 설명변수 중에서 분산이 거의 없는 4개의 변수를 찾아서 제거하였다.

3.2.3. 중요 변수의 선택

26개의 설명변수에 대해서 상관분석 및 그래프 분석을 시행하였다. 또한 Domain 전문가들의 의견을 반영하여 19개의 주요 변수를 선정하였다.

3.2.4. 데이터셋의 분리

데이터셋은 학습데이터와 테스트셋으로 분리하였다. 이 논문에서는 7:3의 비율로 나누었다. 그리고, 모델의 성능평가를 엄격하게 하려고 분리 과정에서 원래 데이터셋에 대한 복원추출은 허용하지 않았다.

Ⅳ. 실험 및 결과

4.1. 평가 지표

예측모델의 성능을 평가하는 지표는 회귀분석에서는 결정계수(R^2), 평균제곱오차(MSE), 평균절대오차(MAE)를 많이 사용한다. 본 논문에서는 R^2, MSE와 같이 데이터의 제곱을 통해 구한 통계량이 관측값과 차원이 달라서 직관적인 이해가 어려우므로, 절댓값을 채용하여 평균적인 오차의 정도를 쉽게 파악할 수 있는 MAE를 평가 지표로 선정하였다.

반응변수가 명목형일 경우에는 Accuracy(정확도)를 사용하였다. (4)Accuracy는 반응변수의 Label(관측값)에 대해서 모델이 예측한 Label(예측값)이 일치하는 비율을 의미한다. [13]

$$R^2(y_{true}, y_{pred}) = 1 - [\ \Sigma(y_{true} - y_{pred})^2 / \Sigma(y_{true} - \bar{y})^2\ ] \quad (1)$$

$$MSE(y_{true}, y_{pred}) = 1 / n_{samples}\ \Sigma\ (y_{true} - y_{pred})^2 \quad (2)$$

$$MAE(y_{true}, y_{pred}) = 1 / n_{samples}\ \Sigma\ |y_{true} - y_{pred}| \quad (3)$$

$$\text{Accuracy} = \text{관측값과 일치하는 예측값의 수} / \text{예측값의 수} \quad (4)$$

4.2. 머신러닝 알고리즘 예측모델의 성능향상

전처리가 완료된 학습데이터를 이용해 Linear Regression, Random Forest, SVM, KNN, DNN 알고리즘을 적용한 5개의 예측모델을 구축하였다.

① Hyper-Parameter 최적화 이후 머신러닝 알고리즘별 예측모델의 성능을 비교하였다. 그리고, 가장 성능이 우수한 모델로 ② 샘플 크기의 증가, ③ 반응변수의 데이터 타입 변경, ④ Domain 지식 기반의 이상치 제거, ⑤ Feature Creation의 방법을 통해서 성능향상을 검증하였다.

4.2.1. 머신러닝 알고리즘 별 Hyper-Parameter 최적화를 통한 성능 비교

학습 데이터의 10%를 분리하여 Hyper-Parameter 최적화를 위해 사용하였다. Hyper-Parameter의 최적화는 Grid Search 방식을 사용하였다. 그리고 알고리즘별 최적화된 예측모델은 테스트셋을 통해 그 성능을 검증하였다.

그림 5는 Hyper-Parameter 최적화 전/후의 머신러닝 알고리즘별 예측모델의 성능을 보여준다. 일부 알고리즘에서 Hyper-Parameter 최적화의 효과가 매우 미미한 수준 악화하였다. 이는 한정된 범위에서의 Grid Search 방식을 통한 최적화로 야기된 것으로 판단된다. 연구보다 더 넓은 범위에서 Hyper-Parameter의 최적화가 진행되었다면, 다른 결과가 발생했을 것으로 판단한다. 실험의 결과를 바탕으로, Hyper-parameter의 최적화는 예측모델의 성능향상에 좋은 방법으로 선택될 수 있다고 판단한다.

머신러닝의 알고리즘별 MAE를 비교하면, SVM이 가장 우수한 것으로 나타났다. 일반적으로 알려진 것처럼, 데이터셋의 성격에 따라서 적절한 알고리즘이 선정되어야 예측모델의 성능도 보장할 수 있다.

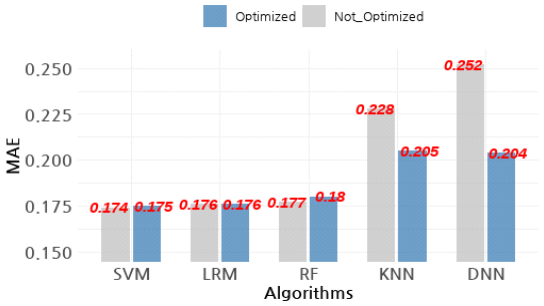

Fig. 5 Performance by Hyper-Parameter Optimization

4.2.2. 샘플 크기에 따른 성능 비교

머신러닝 알고리즘별 성능 비교에서 가장 우수한 것으로 나타난 SVM 모델을 이용하여 샘플 크기에 따른 성능을 실험하였다.

그림 6에서는 샘플의 크기를 각각 10개부터 552개로 점진적으로 증가하며 수립한 각 예측모델의 성능을 보여준다. 샘플 크기가 증가함에 따라 MAE는 지속적으로 감소함을 확인할 수 있다.

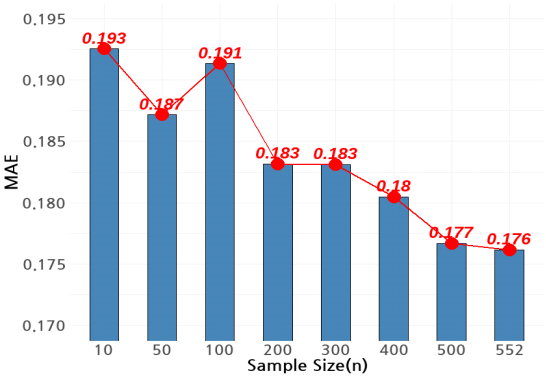

Fig. 6 Performance(MAE) By Sample Size

4.2.3. 반응변수 데이터 타입 변경 통한 성능향상

연속형 반응변수를 가진 SVM 모델로 예측한 값들에 제품의 합/불을 판단하는 기준을 적용하여 성능을 정확도로 환산하였다. 표 2에서 확인할 수 있듯이 정확도는 81.4%로 나타났다. 분류 모델의 모델링 전에 반응변수의 데이터 타입을 연속형에서 명목형으로 변환하여, 예측모델을 만들었다. 이 모델을 이용하여 성능을 평가한 결과, 정확도 84.4%가 발현되었다.

제조공정에서 품질특성의 예측 목적은 상황에 따라서 최적화 혹은 관리의 두 가지로 구분할 수 있다. 최적화는 연속형 데이터 타입에서 관리는 명목형 데이터 타입에서 선호되는 방법이다. 예측의 목적에 맞게 분류 혹은 회귀모델을 선택하는 것이 비교적 좋은 성능을 얻는 방법이다.

Table. 2 Performance by Data Type

	Continuous Type	Categorical Type
Sample Size (n) Tested	237	237
Samples Matched to True	193	200
Accuracy	81.4%	84.4%

4.2.4. Domain 지식 기반의 이상치 제거를 통한 성능향상

제조공정의 데이터는 센서를 통한 수집 및 작업자의 입력과정에서 다양한 오류가 포함될 수 있다. 이러한 오류는 이상치로 나타날 수 있다. 이러한 이상치에 대한 적절한 처리 없이 데이터셋을 사용하면 모델의 성능은 저하될 수밖에 없다.

이에 본 연구에서는 Domain 전문가들이 기술적 지원을 받아서 이상치를 선별하였다. 공정의 변수들은 생산되는 제품군에 따라서 다른 분포를 한다. 해당 분포를 기반으로 발생확률이 낮은 값들을 탐색하여, 이를 이상치라고 판단하고, 제거하는 것이다.

그림 7은 이상치 제거를 통해서 MAE가 37.7% 개선된 것을 보여준다. 예측모델의 성능향상 방법 중에서 가장 효과적인 것으로 보인다.

4.2.5. Feature Creation을 통한 성능향상

본 논문의 대상이 되는 제조공정에서는 반응변수에 열처리가 매우 큰 영향을 주는 것으로 알려져 있다. 열처리가 이뤄지는 3개의 영역에서의 온도 차이를 새로운 변수로 생성했다. 그림 8은 기존 데이터셋으로 수립한 모델과 새롭게 추가된 변수를 추가한 데이터셋으로 수립한 모델의 성능을 비교하였다. 유의미한 변화는 없었다. 이에 대해서는 앞으로 기술적, 수학적 측면에서 Feature Creation을 통해 다양한 파생변수를 생성하고, 이에 관한 추가적인 연구를 통해서 효과에 대한 검증이 이뤄져야 할 것으로 판단한다.

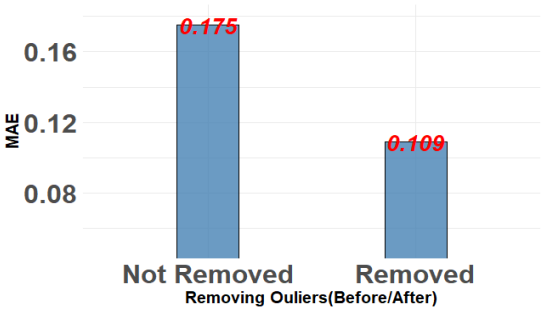

Fig. 7 Performance by Removing Outliers

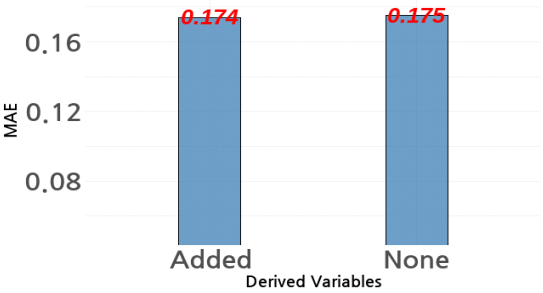

Fig. 8 Performance by Feature Creation

Ⅴ. 결 론

일반적으로 알려진 머신러닝 예측모델의 성능향상 방법으로 ① 적절한 머신러닝 알고리즘의 선택 ② Hyper-Parameter의 최적화 ③ 샘플 크기의 증가를 제조공정의 데이터셋에 적용하여 유효성을 검증해보았다. 또한, 새롭게 발견한 성능향상 방법으로 ① Domain 지식을 바탕으로 한 이상치의 제거 ② 반응변수의 데이터 타입의 전환 ③ Feature Creation을 살펴보았다.

반응변수의 데이터 타입의 전환을 통한 성능향상은 '관리'라는 의미에 대한 깊은 이해가 필요한 방법이다.

관리는 목표에 대한 현재 활동의 달성 여부를 살펴보는 행위로 현상과 목표의 일치 여부에 대한 비교만 가능하면 된다. 즉, 높은 수준의 해상도가 필요 없는 행위이다. 연속형의 데이터 타입은 목적한 구체적인 값을 달성하기 위한 회귀모델에는 적절하지만, 관리의 측면에서는 경제성에 어긋나는 형태의 데이터 타입이다. 이러한 원리를 이용하여 해상도를 양보하고 관리를 위한 성능지표를 향상한 것이다.

본 논문에서의 이상치의 탐색 방법은 흔히 알고 있는 전체 데이터의 분포에서 통계적 접근으로 찾아낸 이상치가 아닌 Domain에 대한 기술적 접근을 선택하였다. 수집한 전체 데이터의 통계적 분포보다는 이를 제품군으로 세분화하고, 각 제품군이 가지는 개별적인 분포를 바탕으로 통계적 확률이 낮은 개별 데이터를 이상치로 구분하였다.

Feature Creation의 경우에는 Domain의 기술 및 경험에 대한 충분한 검토를 통해서 다양한 파생변수들이 제안되어 평가되어야 하지만, 제한된 일부의 파생변수만으로 효과 검증이 진행되었다. 그래서, 앞으로 이에 관한 추가적인 연구가 필요한 것으로 보인다.

제조업에서도 PLC와 IoT 센서 등에서 실시간으로 축적되는 양질을 데이터를 활용하려는 적극적인 움직임이 진행되고 있다. 특히 빅데이터를 활용하는 적절한 도구로 머신러닝을 도입하여 품질을 예측하는 과제 활동은 일반화되었다. 머신러닝을 활용하여 개선의 방향성을 찾는 것은 비교적 낮은 예측성능에서도 충분하다. 그러나, 고객이 체감하는 중요 품질특성과 관련된 항목에 대한 예측은 비교적 높은 수준의 성능을 요구하게 된다. 왜냐하면, 이는 Claim이나 Complaint 등과 같은 품질 실패 비용과 직결되기 때문이다. 이에 기존에 알려진 성능향상 방법과 함께 논문에서는 새롭게 발견한 방법을 제시한다. 제시한 방법을 충실하게 적용한다면, 최적화나 관리의 목적에 부합되는 만족스러운 성능의 예측모델을 얻을 수 있을 것으로 기대한다.

REFERENCES

[1] Machine Learning is the keyword in the manufacturing industry for the next 5 years. Maeil Business News [Internet]. Available: https://www.mk.co.kr/news/economy/view/2021/01/100880/.

[2] The zero percent of the employment rate in the manufacturing industry is coming. Asia Economy [Internet]. Available: https://www.asiae.co.kr/article/2021011106524868927.

[3] The 10 ways to improve the manufacturing industry. AITIMES [Internet]. Available: http://www.aitimes.com/news/articleView.html?idxno=128731.

[4] J. Brownlee. How To Improve Deep Learning Performance. Machine Learning Mastery [Internet]. Available: https://machinelearningmastery.com/improve-deep-learning-performance/.

[5] S. K. Kim, S. Y. Hwang, and J. H. Lee, “Application of Multi-Layer Perceptron and Random Forest Method for Cylinder Plate Forming,” *Journal of the Society of Naval Architects of Korea*, vol. 57, no. 5, pp. 297-304, 2020.

[6] S. Y. Kim and G. B. Kim, “Classification Performance Analysis of Sillicon Wafer Micro-Cracks Based on SVM,” *Journal of Korean Society for Precision Engineering*, vol. 33, no. 9, pp. 715-721, 2016.

[7] C. H. Kim, S. H. Choi, W. J. Joo, and G. B. Kim, “Classification of Surface Defect on Steel Strip by KNN Classifier,” *Journal of Korean Society for Precision Engineering*, vol. 23, no. 8, pp. 80-88, 2006.

[8] S. H. Song, “Study on Hot Deformation of Inconel 601 Using DeepNeural Network and Support Vector Regression,” *Journal of the Korean Society of Mechanical Technology.*, vol. 21, no. 1, pp. 96-101, 2019.

[9] S. H. Jeon and Y. S. Son, “Prediction of fine dust PM10 using a deep neuralnetwork model,” *The Korean Journal of Applied Statistics*, vol. 31, no. 2, pp. 265-285, 2018.

[10] A. B. M. S. Rahman, V. Ragu, M. B. Lee, J. W. Park, Y. Y. Cho, M. H. Lee, and C. S. Shin, “An Analysis Study Based on Linear Regression Model for Changes of Fruit Size over Plum Diseases,” *Journal of Knowledge Information Technology and Systems*, vol. 13, no. 5, pp. 509-519, 2018.

[11] H. Y. Lee, “Research Methodology,” 2nd ed. Seoul, Korea: CRbooks, pp. 407-411, 2014.

[12] P. Sarkar. Bagging and Random Forest in Machine Learning. knowledgehut [Internet]. Available: https://www.knowledgehut.com/blog/data-science/bagging-and-random-forest-in-machine-learning.

[13] J. Jordan. Evaluating a machine learning model. Jeremy Jordan [Internet]. Available: https://www.jeremyjordan.me/evaluating-a-machine-learning-model/.

김종훈(Jong Hoon Kim)

금오공과대학교 생산기계 학사 졸업 (2001)
성균관대 데이터사이언스 석사과정 재학 (2021)
※관심분야 : Machine Learning, Deep Learning, Big Data

오하영(Hayoung Oh)

Sungkyunkwan University Professor (2019~)
Ajou University Professor (2016~2019)
Soongsil University Professor (2013~2016)
Ph.D. in computer engineering at Seoul National University (2013)