Journal of the Korean Institute of Industrial Engineers
Vol. 49, No. 3, pp. 248-257, June 2023.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2023.49.3.248

제조품질 향상을 위한 데이터 전처리 프로세스

서호진[1] · 김도현[2] · 변재현[1†]

[1]경상국립대학교 산업시스템공학부 / [2]명지대학교 산업경영공학과

Data Pre-processing for Manufacturing Quality Improvement

Hojin Seo[1] · Dohyun Kim[2] · Jai-Hyun Byun[1]

[1]Department of Industrial and Systems Engineering, Gyeongsang National University
[2]Department of Industrial and Management Engineering, Myongji University

Improved manufacturing data acquisition systems such as sensors and fast communication systems have made it possible to collect various types of data that were previously unavailable. However, manufacturing data may be contaminated with errors during the data collection process due to such problems as noise or process environment. Utilization of these error-contained data can waste resources and render analysis results useless. To help data scientists and quality engineers dealing with manufacturing quality data, a guideline is proposed for appropriate pre-processing of manufacturing quality data in six steps. Two case study data are used for illustration. The proposed approach is compared with six other methods and shows advantageous in terms of the F1 score. This paper is expected to help quality practitioners and data scientists applying machine learning methods to manufacturing quality data.

Keywords: Manufacturing Quality Data, Data Pre-processing, Big Data, Machine Learning, Quality Improvement

1. 서 론

데이터를 기반으로 의사결정을 하는 일은 수십 년 동안 품질 개선의 핵심이었다. 최근 센서 등 향상된 데이터수집 장치, 통신시스템의 개발로 다양한 종류의 데이터를 신속하게 얻을 수 있게 되었으며, 공정에서 생성되는 데이터의 양도 폭발적으로 증가하였다. 공정의 운영과정에서 축적된 데이터를 분석하여 얻은 정보와 현재 가동 중인 장비에서 수집한 결과를 비교하여 현재 장비가 어떤 상태인지를 판별하고, 이상이 있으면 수리하고 필요시 교체를 진행한다(Kim *et al.*, 2018). 수집한 데이터를 이용하여 분석한 결과는 품질특성의 특이한 경향이나 변화에 관한 정보를 실시간으로 제공하므로 공정개선을 위한 시간을 단축하고 품질비용을 절감하는 데 중요하다. 더군다나 데이터 분석 결과의 시각화는 모든 이해관계자가 공급망 전체에 걸쳐 제품의 특성과 생산공정의 상태를 신속하게 파악하는 데 도움을 준다(Lyle, 2017).

데이터를 활용하는 주목적은 현장에서 수집되는 데이터를 분석하여 의사결정에 유용한 정보를 획득하는 것이다. 하지만 데이터수집 과정에서 노이즈나 공정 환경 등의 문제로 데이터가 불완전할 수 있는데, 이들은 전처리 과정을 통해 처리되어야 한다. 문제가 있는 공정데이터를 전처리 과정을 거치지 않은 상태에서 분석하게 되면 자원이 낭비되고, 분석 결과가 쓸모없게 되기 때문이다. 데이터를 통해 의미 있는 결론을 도출하기 위해서는 모델을 개선하기보다 데이터를 개선하는 것이 더 중요하다. 사례연구에 의하면 머신러닝 방법을 개선하여 더 나은 모델을 구하는 것보다 데이터 자체를 개선하였을 때 분석 성능이 좋아진다(Ng, 2021).

앞으로도 제조공정에서 대량으로 수집될 데이터의 품질 문

본 논문은 교육부와 한국연구재단의 재원으로 지원을 받아 수행된 3단계 산학연협력 선도대학 육성사업(LINC 3.0)의 연구 결과입니다.
† 연락저자 : 변재현 교수, 52828 52828 경남 진주시 진주대로 501 경상대학교 산업시스템공학부, 055-772-1692, Fax: 055-772-1699, E-mail: jbyun@gnu.ac.kr
2023년 3월 13일 접수; 2023년 4월 16일 수정본 접수; 2023년 4월 18일 게재 확정.

제는 계속해서 발생할 것으로 예상된다. 제조공정의 품질을 높이기 위해 일하는 데이터 분석가나 엔지니어는 데이터의 품질을 높이기 위한 전처리 기법을 익히는 것만으로는 불충분하다. 이들을 적용할 순서를 단계별로 정하고, 각 단계에서 적용할 세부 기준에 대한 고찰하는 것이 필요하다. 본 연구에서는 제조 데이터 분석 모델의 성능 향상을 위한 데이터 전처리 방법인 데이터 통합, 결측치와 이상치의 처리, 피처엔지니어링, 변수변환, 데이터 불균형 처리 등 해당 기법을 검토하고, 그 결과에 기반하여 적절한 방법을 선택할 수 있도록 데이터 전처리 프로세스를 위한 지침(guideline)을 제시하고자 한다.

본 논문은 다음과 같이 구성되어 있다. 제2장에서는 본 연구와 관련 있는 전처리 기법을 소개한다. 제3장은 단계별 데이터 전처리 순서와 단계별로 데이터 분석가가 맞게 될 상황에 맞는 데이터 전처리 지침을 제안하고, 다양한 데이터 유형에 따른 전처리 프로세스 적용 방법은 제4장에 기술한다. 제5장은 2가지 사례를 통해 본 연구에서 제안한 프로세스의 성능을 평가하고 그 결과를 해석하며, 제6장에서는 결론과 추후 연구방향을 제시한다.

2. 전처리 기법 소개

2.1 데이터 통합

여러 프로세스에서 생성되는 대용량 데이터 세트를 체계적으로 연결하여 분석하는 것이 필요하다. 다양한 소스에서 나온 데이터를 단일화하면 더 가치 있는 통합데이터 세트를 만들 수 있고, 이것은 효과적 의사결정을 위한 기초가 된다(Hall and Llinas, 1997).

데이터 통합을 위해서는 같은 것을 뜻하면서도 표기가 다른 단어를 하나의 단어로 통일하고, 데이터의 단위 표시를 일치하며, <Figure 1>과 같이 데이터 소스 간의 서로 다른 시간 단위를 공동의 시간 단위로 통합하는 작업이 필요하다. 그 후, 분리된 데이터 세트를 공통의 데이터열을 기반으로 결합해야 한다.

2.2 결측치 처리

결측치는 데이터에 값이 없는 것으로서 입력이 빠진 데이터를 뜻하는데, 공정 내에서 센서 등의 고장이나 오류로 인하여 발생할 수 있다. 결측치가 있으면 변수 간의 관계가 왜곡되어 모델의 정확성이 떨어질 수 있으므로 삭제하거나 다른 값으로 대체해야 한다.

본 논문에서는 결측치를 처리하기 위해, 기존에 진행된 연구결과를 살펴보았다. Scheffer(2002)에 의하면, 결측치의 비율이 50%가 넘어가면 이들을 대체하더라도 예측성능 향상을 기대하기 어렵다고 밝혔으며, Graham(2009)은 결측치를 제거하면 편향이 발생할 수 있지만, 결측치 비율이 5% 미만일 경우 삭제하여도 문제가 되지 않는다고 하였다. Schmitt *et al.*(2015)은 대부분의 데이터 세트에 결측치가 존재한다는 점과 효과적으로 결측값을 대체하는 방법을 탐색하기 위해 KNN(K-Nearest Neighbors), Fuzzy K-means, MICE(Multiple Imputations by Chained Equations), bPCA(Bayesian Principal Component Analysis), SVD(Singular Value Decomposition)를 이용하는 방안에 관해 연구하였다. García-Laencina *et al.*(2015)은 결측치를 중앙값, KNN, Expectation Maximization 기법으로 대체한 것들을 모델에 학습시켰는데, 이 중 KNN으로 결측치를 대체하였을 때 최적의 결과가 나왔다. Cho *et al.*(2022)은 UCI(University of California at Irvine)에서 제공하는 SECOM(Semiconductor Manufacturing) 데이터 세트의 결측치를 Linear Interpolation, Poly Interpolation, MICE, KNN, MissForest 기법으로 대체하여 모델을 학습시켜 성능을 비교하였고, KNN 방법이 가장 좋다는 결과를 확인하였다.

2.3 이상치 처리

이상치는 피처 또는 레이블이 관측된 범위에서 많이 벗어난 아주 작거나 큰 값을 가진 데이터인데, 공정 내에서 측정기의 고장이나 부적합, 데이터 입력 및 처리 과정에서 발생한 오류 등으로 인하여 발생한다. 이상치가 존재하면 변수 간 관계성을 왜곡할 수 있다. 예를 들어, 관계가 있는 변수들의 상관성이 파악되지 않거나, 서로 연관이 없는 변수들이 관계가 있다는 결과가 나올 수 있다. 데이터를 분석하여 의사결정을 수행하기 전에 이상치를 식별하여야 한다.

이상치는 통계 기법인 사분위수를 이용하거나 LOF(Local Outlier Factor), iForest(isolation Forest), 마할라노비스 거리(Mahalanobis Distance) 등을 활용하여 검출하고, 검출된 이상치는 결측치와 마찬가지로 삭제하거나 다른 값으로 대체하여 처리한다.

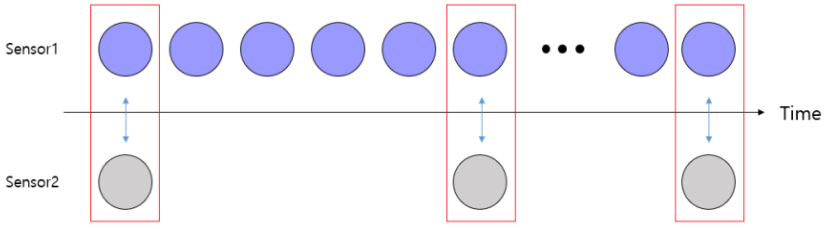

Figure 1. Time Unification

이상치에 의해 모델이 왜곡되는 것을 최대한 예방하기 위해 모델링을 할 때 이상치의 영향을 최대한 줄여주는 강건회귀(Robust Regression)와 같은 알고리즘이 제안되었다. 회귀 문제에 많이 활용되는 SVR(Support Vector Regression)은 목적함수를 새롭게 정의하여 이상치의 영향력을 최대한 줄이는 방법이다(Jun, 2008).

2.4 데이터 변환

데이터 변환이란 분석의 효율성을 높이기 위해 데이터 분포의 모양을 보고 변환하거나, 데이터의 범위를 조정하거나, 수치형 데이터를 범주화하기 위하여 이용하는 분포변환, 정규화, 표준화, 이산화 등을 말한다.

분포변환은 데이터 분포의 왜도(Skewness)나 첨도(Kurtosis)가 지나치게 큰 경우 제곱근이나 로그 등을 활용하여 정규분포 형태로 만드는 것이다. 정규화와 표준화는 데이터의 범위를 조정하는 것이다. 데이터 분석 모델은 범위가 큰 변수에 의해 영향을 받아 편향될 수 있다. 데이터 알고리즘을 활용하기 전에 정규화나 표준화를 통해 데이터의 범위를 균일하게 변환하면 알고리즘의 학습속도와 정확도를 높일 수 있다(Escobar *et al.* 2020). 이산화는 수치형 데이터를 범주형 데이터로 변환하는 것으로서, 머신러닝 모델의 계산속도를 높일 수 있다. 가장 간단한 이산화 기법은 구간화(Binning)인데, 이 방법은 연속형 변수의 값을 몇 개의 구간으로 나누는 것이다.

2.5 피처엔지니어링

피처엔지니어링은 많은 변수를 포함하는 고차원 데이터(high-dimensional data)를 전처리하는 것이다. 변수가 많으면 이를 분석하기 위하여 훨씬 많은 관측치가 필요한데, 사실은 많은 경우에 분석 목적과는 관련 없는 변수들이 많이 포함되어 있다. 관련 없는 변수들이 포함되면 모델의 성능이 떨어지므로 최대한 목적에 맞는 적절한 변수들로 데이터를 구성하여야 한다.

피처엔지니어링은 피처 선택(Feature Selection)과 피처 추출(Feature Extraction)의 두 가지로 나누어진다. 피처 선택은 주어진 피처 중에서 모델링에 유용한 피처들만 선택하는 과정으로서, 대표적인 방법으로 사전에 다양한 통계량을 통해 모델에 유용하지 않을 것 같은 피처를 제거하는 필터링 기법과 모델을 반복적으로 수행해가면서 가장 적합한 피처 집합을 업데이트하는 래퍼 기법이 있다. 피처 추출은 기존의 피처들을 결합하여 더 유용한 피처를 생성하는 것인데, 대표적인 방법으로 고차원의 피처 공간을 새로운 저차원의 피처 공간으로 투영하여 차원을 축소하는 주성분분석(PCA; Principal Component Analysis), 선형판별분석(LDA; Linear Discriminant Analysis), 비음수행렬분해(NMF; Non-Negative Matrix Factorization) 등이 있다.

2.6 데이터 불균형 처리

제조공정에서 양품은 많고 부적합품은 아주 적게 나오는 등 분류 데이터 세트에서 클래스 간 불균형이 심하면, 모델이 제대로 학습하지 못하고 분석 결과가 크기가 큰 클래스에 편향되어 나타나는 현상이 발생한다. 품질 분류 모델에서 사용하는 데이터는 대부분 양품이므로 이러한 품질데이터를 있는 그대로 학습하면 양품만 예측하는 모델이 생성될 수 있다. 예를 들어, 998개의 양품과 2개의 불량품으로 구성된 품질데이터를 학습할 경우, 모두 양품만 예측하는 모델이 생성된다. 그 결과 모델의 전체적인 정확도는 높지만, 소수의 클래스를 제대로 분류하지 못하는 문제가 발생한다. 클래스 불균형 문제를 해결하기 위해서는 리샘플링이 필요하다(Zhang *et al.*, 2019). 리샘플링 기법으로는 <Figure 2>처럼 소수 범주의 데이터를 증가시키는 오버샘플링(Over Sampling)과 <Figure 3>과 같이 다수 범주의 데이터를 소수 범주 데이터 수에 맞게 줄이는 언더샘플링(Under Sampling)이 있다.

데이터 불균형을 처리하기 위하여, Chawla *et al.*(2002)은 다수의 데이터 세트에 클래스 불균형 문제가 있음을 파악하여 이를 해결하기 위해 다수 클래스를 언더샘플링 하는 방법과 소수 클래스를 오버샘플링 하는 방법을 조합하여 더 나은 방법을 탐색하였다. Kim and Kwahk(2022)은 OpenML, Kaggle, UCI 등에서 제공하는 33가지 데이터 세트를 대상으로 Adasyn (Adaptive Synthetic Sampling), SMOTE(Synthetic Minority Oversampling Technique), NCR(Neighborhood Cleaning Rule), Tomek Link, RUS(Random Under Sampling), CNN(Condensed

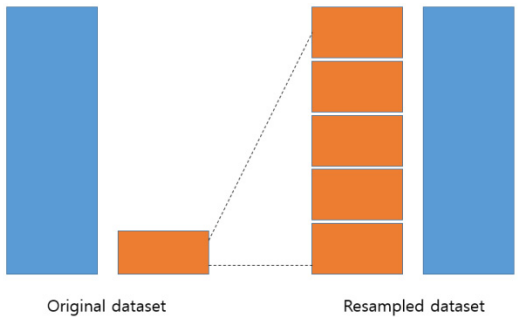

Figure 2. Over Sampling

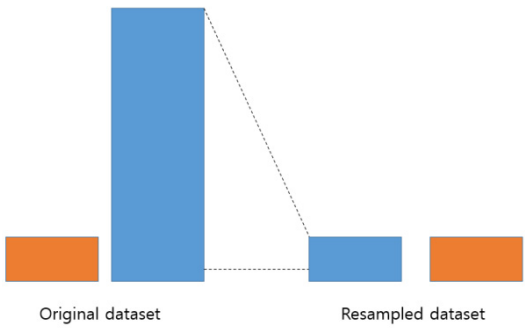

Figure 3. Under Sampling

Nearest Neighbor), ENN(Edited Nearest Neighbor)을 적용하여 데이터 불균형을 처리하는 모델의 성능을 확인하였는데, 33개의 데이터 세트 중 17개의 데이터 세트에서 Adasyn으로 데이터 불균형을 처리하였을 때 가장 좋은 결과가 나왔다.

3. 데이터 전처리 프로세스 적용 지침

데이터를 전처리를 할 수 있는 기법은 여러 개가 있고, 전처리를 활용한 사례도 많이 있지만, 다양한 전처리 방법을 적용하는 순서에 관한 연구는 아직 없다. 본 논문에서는 제조공정의 품질데이터를 효과적으로 전처리할 수 있는 프로세스를 제시하고자 한다.

데이터를 분석하기 위해 준비하는 전처리의 첫 번째 단계는 데이터 통합이다. 여러 저장 장치에 있는 데이터들을 통합하여 하나의 데이터 세트로 만들어야 한다. 그다음에는 결측 데이터와 이상 데이터를 차례로 처리한다. 이상 데이터가 있으면 데이터 변환이나 피처엔지니어링 등 데이터 전처리 기법에 악영향을 줄 수 있다. 네 번째 단계는 데이터 변환이다. 데이터 변환을 통해 피처의 범위를 일정하게 설정하고 수치형 데이터를 이산화하여 다른 데이터 전처리 속도 및 데이터 분석 속도를 향상할 수 있다. 다섯 번째는 피처엔지니어링이다. 앞선 데이터 전처리 기법들을 적용한 후에 분석에 필요 없는 피처를 선별하고 제거하여 성능과 해석력을 높이거나 데이터의 노이즈나 상관관계가 높은 피처들이 존재하는 경우, 모델의 성능을 높이기 위해 새로운 소수의 피처를 추출한다. 마지막으로 데이터 불균형 처리이다. 데이터 불균형은 리샘플링을 통해 데이터를 복제하거나 제거하는 기법으로 데이터 전처리 기법들이 모두 적용된 후 활용하면 더 좋은 성능을 발휘할 수 있다.

<Figure 4>에 6가지 데이터 전처리 기법을 이용하는 순서를 제시하였다. 5장의 사례에서는 전처리 된 데이터를 로지스틱 회귀, 랜덤포레스트, 서포트벡터머신을 활용하여 분석하였다. 각 전처리 과정에서 제시한 기준은 보수적으로 설정하였으며, 추후 현장 데이터 분석가와 엔지니어의 실무 경험을 바탕으로 보완이 필요하다. 또한, 데이터 전처리 프로세스를 적용하더라도 머신러닝을 활용한 데이터 분석 결과가 좋지 않을 경우, 다시 데이터 전처리를 진행하여야 한다.

데이터 전처리를 시작하기 전에 만일 분석에 필요한 데이터가 여러 파일로 흩어져 있다면, 이들을 하나의 데이터 세트로 통합하여야 한다. 데이터 전처리를 위한 첫 번째 단계인 데이터 통합 과정의 순서는 <Figure 5>와 같다. 다양한 공정 데이터 세트가 존재할 때는 의미는 같지만, 표기가 다른 단어와 측정 단위부터 통일한다. 그 후 데이터 측정 주기 및 시간을 통일한 후, 공동의 데이터열을 기준으로 연결하면 된다.

두 번째 단계인 결측치 처리 과정은 <Figure 6>과 같다. 우선 전체 데이터 대비 결측치가 차지하는 비율을 파악하여야 한다. 2.2절을 참고하여 결측치의 비율이 5% 미만일 경우 결측치를 삭제하고 다음 단계로 넘어간다. 어떤 피처의 결측치 비율이 50% 이상이면 그 피처는 삭제한다.

둘째, 결측치의 유형을 파악하여야 한다. 결측치의 유형은 1) 무작위로 발생하는 완전 무작위 패턴(Missing Completely At Random; MCAR), 2) 어떤 변수의 결측치가 다른 변수들이 특정 범위의 값을 갖는 경우에 발생하는 부분적 무작위 패턴(Missing At Random; MAR), 3) 결측치가 발생한 이유가 결측치

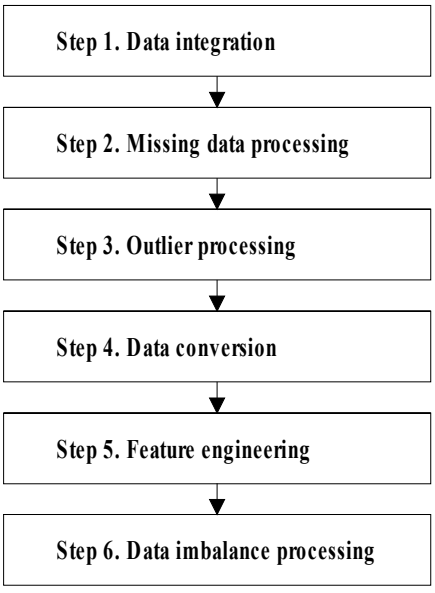

Figure 4. Data Pre-processing Steps

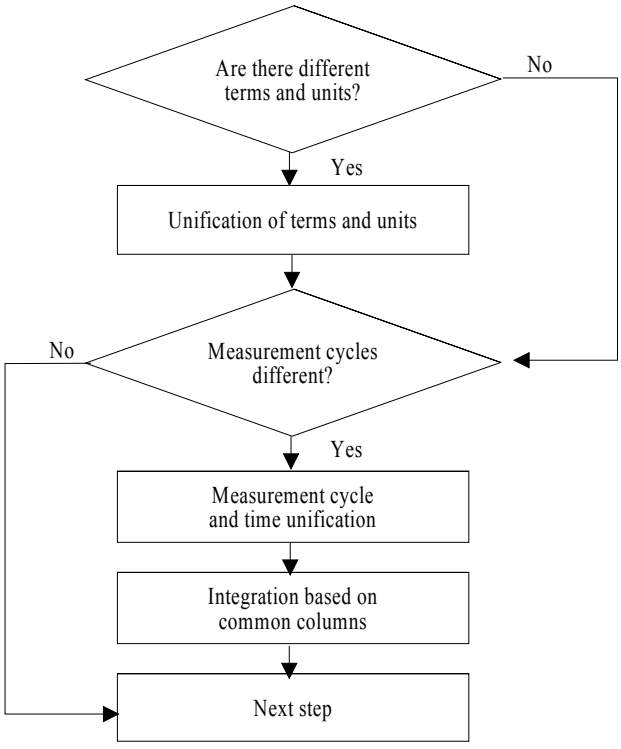

Figure 5. Data Integration

가 발생한 값 자체와 관련되어 나타나는 비무작위 패턴(Missing Not At Random; MNAR)으로 구분된다(Krishnamurthi, *et al.*, 2020, Kwak, *et al.*, 2017). MCAR의 경우에는 랜덤한 값으로, MAR은 다른 변수를 이용하여 예측 모델을 생성하고 이를 바탕으로 결측치를 대체한다. 다만, MNAR의 경우, 결측치가 누락된 변수와 관련된 정보를 포함하기 때문에 결측치 대체는 전체 데이터를 분석한 결과에 큰 영향을 미칠 수 있으므로 전문적인 지식과 기술에 근거하여 결측치를 대체한다.

결측치를 대체하기 위해서, MCAR, MAR의 경우, 2.2절에서 보듯이 가장 좋은 성능을 보인 KNN을 활용하는 것이 적절하다. MNAR의 경우에는 보다 주의를 기울여 대체해야 하는데, 이를 위해 KNN 기반의 NS-KNN, KNN-TN이 제안되었다(Lee and Styczynski, 2018; Shah, *et al.*, 2017) .

세 번째 단계인 이상치 처리 과정은 <Figure 7>과 같다. 우선 이상치를 식별해야 한다. 2.3절의 내용을 바탕으로 하여 단변량 이상치는 사분위수 범위를 이용하여 파악하고, 다변량 이상치를 탐색하기 위해서는 직관적인 해석이 가능하고 알고리즘이 간결하면서 높은 성능을 보여주는 iForest를 활용할 수 있다. 둘째, 이상치의 비율을 살펴보아야 한다. 이상치의 비율이 5%보다 낮으면 이상치를 삭제하고 다음 단계로 넘어간다. 이상치의 비율이 5%보다 높으면, 이상치의 원인을 파악할 수 있는지를 먼저 확인한다. 이상치의 원인이 밝혀지지 않은 경우, 해당 데이터의 특성, 분석 목적, 분석 알고리즘 등에 따라서 이상치를 반드시 대체해야 하는지를 판단하여 (1) 이상치를 대체하지 않아도 된다면 이것을 제거하고 다음 단계로 가고, (2) 대체해야 한다면 이를 정상적 패턴 데이터로 바꾸고 다음 단계로 이동한다. 이상치의 원인이 센서 또는 측정 장비의 오류, 외부충격 등으로 밝혀지면, 이들 원인을 제거하거나, 프로세스를 통제하거나, 프로세스가 이런 원인에 강건하여지도록 만드는 조처를 한다. 이후에 이상치 대체 필요성을 판단하여 위의 (1) 번과 (2) 번 활동을 시행한다.

네 번째 단계인 데이터 변환 과정은 <Figure 8>과 같다. 우선, 데이터의 분포를 살펴보아야 하는데, 분포의 왜도(skewness)나 첨도(kurtosis)가 정규분포에서 많이 벗어날 경우, 제곱근이나 로그 등 변수변환 방법을 이용하여 정규분포에 가깝게 되는지 확인해야 한다. 둘째, 피처별로 값의 범위를 파악한다. 피처 간의 범위에 차이가 크면 범위를 일정하게 맞추어주는 정규화나 표준화를 진행하여야 한다. 셋째, 추후 활용할 데이터 분석 알고리즘이 대표적인 의사결정나무 방법인 CART와 같이 수치화 데이터보다 범주형 데이터가 더 적합한 알고리즘을 이용하고자 하면 구간화(Binning) 등 방법으로 데이터를 이산화해야 한다.

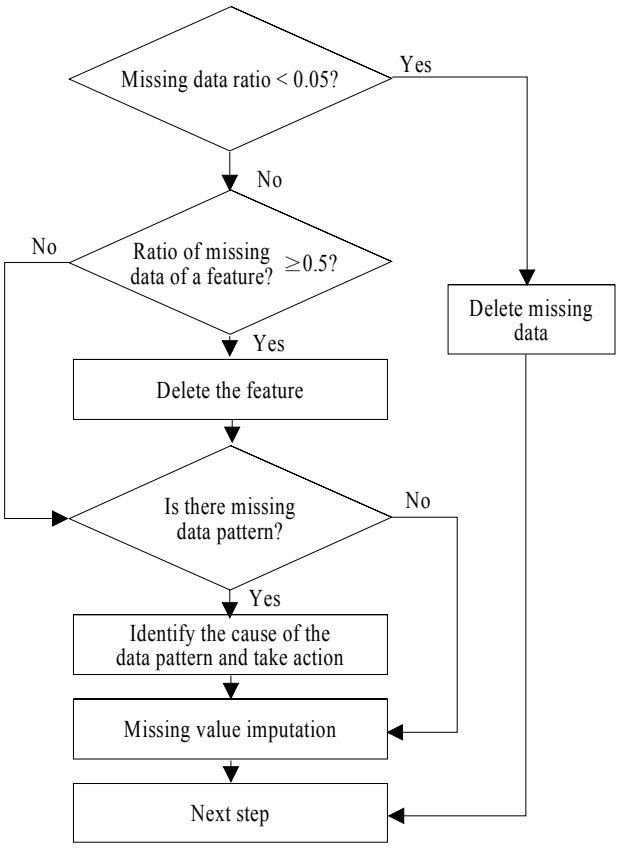

Figure 6. Missing Data Processing

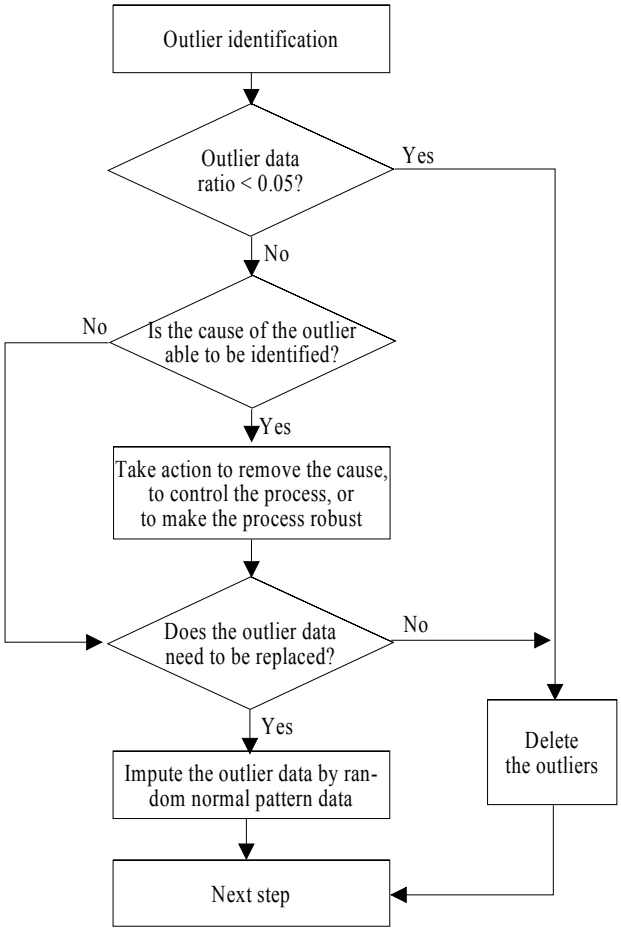

Figure 7. Outlier Processing

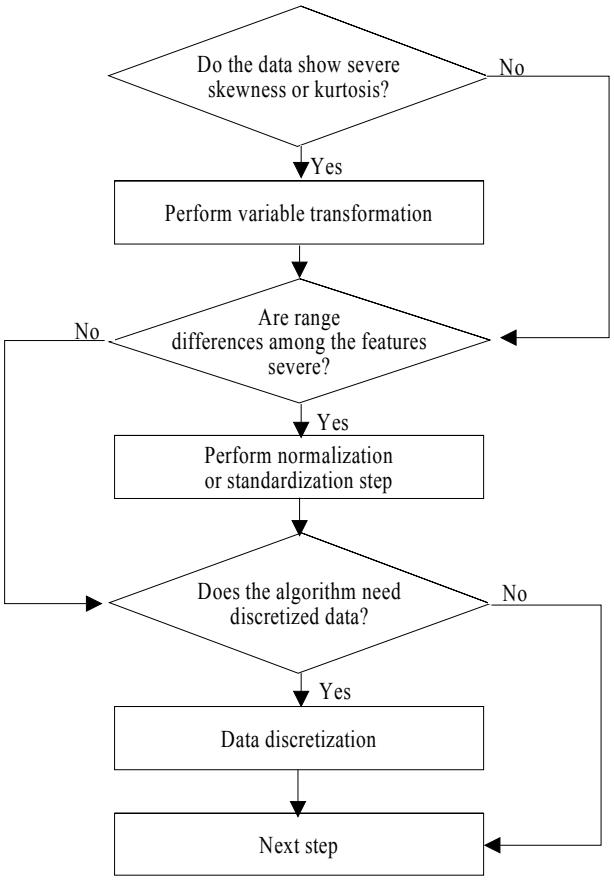

Figure 8. Data Conversion

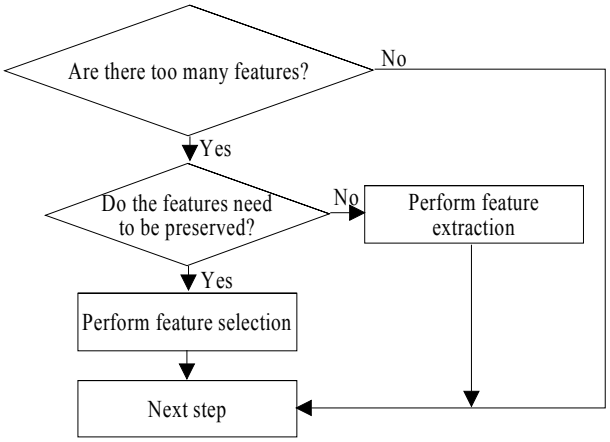

Figure 9. Feature Engineering

다섯 번째 단계인 피처엔지니어링 과정은 <Figure 9>와 같다. 피처의 개수가 많을 경우, 피처 선택을 하여 필요 없는 피처를 삭제하거나, 피처 추출 방법을 이용하여 피처들을 저차원으로 축소하여야 한다. 피처를 보존하고자 하면 피처 선택을, 그럴 필요가 없을 때는 피처 추출을 이용한다.

마지막 단계는 데이터 불균형 처리 과정이다. 데이터 불균형을 처리하는 과정은 <Figure 10>에 나타내었다. 우선, 클래스 비율을 측정하여야 한다. 클래스 간의 불균형이 크면 리샘플링

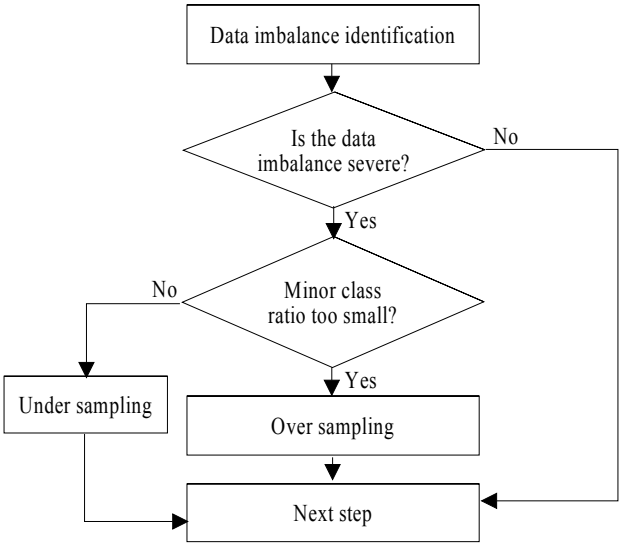

Figure 10. Data Imbalance Processing

을 통해 균형을 맞추어 주어야 한다. 마이너 클래스 비율이 매우 작으면 오버샘플링을 하고, 마이너 클래스와 메이저 클래스의 비율 차이가 크지 않을 때는 언더샘플링을 활용한다. 이때, 오버샘플링은 2.6절에서 성능이 가장 좋은 것으로 판단된 Adasyn을, 언더샘플링을 할 때는 Tomek-Link를 활용한다.

데이터 분석 과정은 초기에 순차적으로 진행할 수 있다. 하지만 이러한 순차적 진행만으로 의미 있는 정보를 얻고 데이터 분석 과정을 종료하기는 쉽게 않고, 실제로는 전 단계를 반복해야 하는 경우가 많다. 전처리한 데이터를 이용한 모델링의 성능이 좋지 않으면 데이터 전처리를 다시 실행한 후에 분석을 수행해야 하기 때문이다.

4. 다양한 데이터 유형에 따른 전처리 프로세스 적용 방법

데이터는 크게 수치형과 범주형 데이터로 구분되며, 수치형은 다시 연속형과 이산형 데이터로 나누어진다. 연속형 데이터는 범위 내의 어떤 값이든 가질 수 있는 데이터이며, 길이, 무게, 온도, 전류, 시간, 속도, 점도 등과 같이 주로 측정을 통해서 얻어지는 데이터이다. 반면, 이산형 데이터는 가질 수 있는 값의 집합이 유한하며, 제조공정을 거치고 난 후 나오는 품질특성인 가공 표면의 흠집, 크랙(crack), 기공(void) 수 등이 이에 해당한다. 지금까지 제조품질 데이터에서 주로 접하게 되는 수치형 데이터 특히 연속형 데이터에 초점을 맞추어서 전처리 프로세스를 기술하였다. 그러나 제조 데이터 분석 시 연속형 이외의 데이터 형태를 종종 접하게 된다. 따라서 본 장에서는 다른 데이터 형태의 전처리 단계에 대해서 살펴보고자 한다. 우선 전처리 프로세스에서는 이산형 데이터도 수치형이므로 연속형 데이터 분석 방법과 특별한 차이를 둘 필요는 없다. 다만, 이산형 데이터는 연속형 데이터와 달리, 유한한 값을 가지

는 데이터로서, 이미 이산화가 되어 있으므로 네 번째 단계인 데이터 변환 중 이산화 단계는 필요하지 않다. 이를 제외한 모든 단계 즉, 데이터 통합, 결측치 처리, 데이터 변환, 피처엔지니어링, 데이터 불균형 처리는 3장의 방법을 적용할 수 있다.

반면, 범주형 데이터에 대해서는 데이터 통합 이후 결측치 처리 이전에 별도의 변환 과정을 추가할 필요가 있다. 범주형 데이터는 크게 명목형 데이터와 순서형 데이터로 분류된다. 명목형 데이터는 범주형 데이터 중에서도 순서를 가지지 않는 데이터로서, 가공 후 표면에 나타나는 원형, 직선형, 곡선형 등 여러 형태의 불량 유형이 이에 해당한다. 명목형 데이터를 변환하는 방법으로는 다양한 방법이 존재한다. 가장 대표적인 방법이 원-핫 인코딩(One-Hot Encoding)으로서, 범주형 변수를 0과 1로 이루어진 이진 벡터로 변환하는 것이다. 범주별로 신규 변수를 생성하고, 해당 범주에 해당하는 위치에 1을 부여하고, 나머지는 0으로 채우게 된다. 원-핫 인코딩은 범주형 데이터를 변환하는 유용한 방법이지만 범주가 많아지면 사용하기 어렵다. 변수의 범주가 아주 많은 경우에 사용할 수 있는 방법이 타겟 인코딩(Target Encoding)이다. 타겟 인코딩은 범주형 변수의 어떤 범주에 대응하는 목표 변수의 값들이 여러 개 있을 때, 이러한 값들의 평균으로 변환한다(Micci-Barreca, 2001).

순서형 데이터는 명목형 데이터와 달리 순서가 있는 데이터로서, 가공된 제품의 표면 상태를 상, 중, 하로 평가하거나, 제품/서비스를 경험한 고객의 만족도를 파악하기 위한 설문조사 결과가 대표적인 예이다. 순서형 데이터도 명목형 데이터와 마찬가지로 결측치 처리 전에 변환 과정을 거치게 되는데, 순서형 데이터를 변환할 때는 각 범주를 순서대로 정렬하고, 각 범주에 대해 번호를 부여하여 값을 변환하게 된다. 예를 들어, 고객 만족도 조사에서 아주 만족 5점, 만족 4점, 보통 3점, 불만족 2점, 아주 불만족 1점을 부여하는 것이다. 이렇게 변환된 데이터는 수치형 데이터로 변환되므로 이후의 전처리 과정은 3장에서 제시된 방법으로 진행된다.

5. 사례연구

본 장에서는 두 가지 사례를 대상으로 제3장에서 제안한 데이터 전처리 프로세스를 적용하여 머신러닝 모델의 성능을 평가하고자 한다. 본 연구에서 적용한 머신러닝 모델은 일반적으로 가장 많이 활용하는 로지스틱 회귀, 랜덤포레스트, 서포트벡터머신이다. 본 연구에서 제시한 데이터 전처리 프로세스의 성능을 파악하기 위해 <Table 1>과 같이 7가지의 데이터 전처리 방법을 적용하였다. 사례연구에 사용한 데이터 세트는 이미 통합되어 있으므로 데이터 통합 과정은 생략하였다.

첫 번째 방법은 본 연구에서 제시한 데이터 전처리 프로세스의 5가지 단계를 모두 적용한 방법이며, 두 번째 방법은 데이터 불균형 처리, 세 번째는 피처엔지니어링, 네 번째는 데이터 변환, 다섯 번째는 이상치의 처리를 각각 제외한 방법이다. 여섯 번째는 모든 전처리 과정을 진행하지만, 이상치/결측치 처리에서는 가장 많이 활용하는 평균 대체를, 데이터 불균형 처리는 가장 자주 활용하는 SMOTE 기법을 활용한 것이다. 마지막 방법은 어떤 전처리도 하지 않은 조합이다. 첫 번째와 여섯 번째 방법의 차이점은 결측지/이상치의 처리와 데이터 불균형을 다루는 데 있다. 첫 번째 방법은 결측치와 이상치의 비율에 따라 이들을 삭제하거나 KNN을 활용하여 대체하지만 여섯 번째 방법은 결측치나 이상치를 모두 평균으로 대체하는 것이다. 또한, 데이터 불균형 처리에서도 첫 번째 방법은 마이너 클래스 비율에 따라 Adasyn과 Tomek Link를 활용하는 데 반해, 여섯 번째 방법은 오버샘플링에서 가장 많이 이용되는 SMOTE를 채택한 것이다.

5.1 데이터 세트

본 논문에서 분석에 활용할 데이터는 <Table 2>에 나타내었는데, Case 1은 사출성형 데이터, Case 2는 반도체 공정데이터이다.

Table 2. Case Study Data Sets

Data set	Instances	Features	Label
Case 1: Mold	6,737	36	1
Case 2: SECOM	1,567	590	1

Table 1. Data Pre-processing Approaches

No	Missing data processing	Outlier processing	Data conversion	Feature engineering	Data imbalance processing
1	O	O	O	O	O
2	O	O	O	O	X
3	O	O	O	X	O
4	O	O	X	O	O
5	O	X	O	O	O
6	mean replacement	mean replacement	O	O	SMOTE
7	O	X	X	X	X

(1) 사출성형 데이터 세트

중소벤처기업부에서 만든 인공지능 중소벤처 제조 플랫폼(Korea AI Manufacturing Platform; KAMP)에서 제공하는 사출성형 데이터 세트는 사출공정에서 수집된 데이터로서 Table 2와 같이 36개의 피처, 1개의 레이블, 6,737개의 인스턴스로 구성되어 있다(Ministry of SMEs and Startups of Korea, 2020). 사출성형 공정의 품질을 나타내는 레이블은 양·불량을 구분하기 위하여 '0'과 ' - 1'로 표시하는데, 0은 양품(Pass), -1은 부적합품(Fail)을 의미한다. 모든 인스턴스 값이 단일 값으로 나타나는 9개의 피처는 분석에서 제외하였다.

(2) 반도체 데이터 세트

반도체 데이터는 UCI(University of California, Irvine)에서 제공되는 SECOM(Semiconductor Manufacturing) 데이터 세트를 사용한다(Dua and Graff, 2019). SECOM 데이터 세트는 반도체 산업의 웨이퍼 제조공정에서 가져온 590개의 피처, 1개의 레이블, 1,567개의 인스턴스가 포함되어 있다. 반도체 제조공정의 품질을 나타내는 레이블은 양·불량이며 ' - 1'은 양품, '1'은 불량을 표시한다. 데이터 전처리에 앞서 각 피처의 데이터 결측치 비율을 파악하였다. 또한, 값이 같게 나타난 116개의 피처는 제외한 후에 분석하였다.

5.2 분석 방법

(1) 분석 모델

로지스틱 회귀, 랜덤포레스트, 서포트벡터머신 방법을 사용하여 제조 데이터의 품질을 분류하였다. 모델의 성능은 정확도와 F1 점수로 평가하였다.

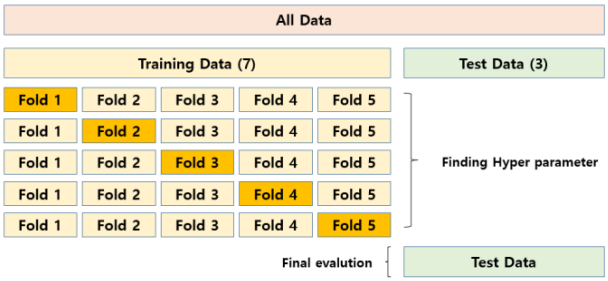

Figure 11. K-fold Cross Validation

(2) 데이터 세트 분할

학습데이터와 평가데이터의 과적합을 방지하고자 교차검증(Cross Validation) 기법의 하나인 K-겹 교차검증(K-Fold Cross Validation)을 사용하였다. 교차검증을 활용하여 C(cost function), Gamma 등 사용자가 설정할 수 있는 초매개변수(Hyper-parameter)를 다양하게 조합하여 모델들을 학습시킨 후 성능평가를 통해 가장 성능이 좋은 조합을 선택하였다. <Figure 11>과 같이 학습데이터와 평가데이터를 7:3의 비율로 분할하고, 학습데이터를 5개의 겹으로 나누어 교차검증을 시행하였다.

(3) 초매개변수 설정

모델의 최적 성능을 찾기 위해 설정한 하이퍼 파라미터는 <Table 3>과 같다. 로지스틱 회귀의 경우에는 과적합을 조정할 수 있는 C(Cost function)를 활용하였다. C가 클수록 훈련데이터를 정확하게 예측하고, 작을수록 과적합을 방지할 수 있다. 랜덤포레스트는 과적합을 방지하기 위해 트리의 최대 깊이인 Max_depth와 분할을 위해 필요한 최소 샘플 수인 Min_sample_split를 활용하였다. 서포트벡터머신을 사용할 때

Table 3. Hyper-parameter Range for Models

Method	Hyper parameter	Range
Logistic Regression	C	[0.01, 0.1, 1, 10]
Random Forest	Max_depth	[2, 5, 8, 11]
	Min_sample_split	[2, 5, 8, 11]
Support Vector Machine	C	[0.1, 0.5, 1.0]
	Gamma	[0.1, 0.5, 1.0]

Table 4. Molding Data Analysis Results

No	Logistic regression		Random forest		Support vector machine	
	accuracy	F1 score	accuracy	F1 score	accuracy	F1 score
1	**0.996**	**0.526**	**0.997**	**0.625**	**0.997**	**0.700**
2	0.998	0.444	0.995	0.170	0.992	0.200
3	0.990	0.222	0.995	0.421	0.993	0.417
4	0.994	0.455	0.995	0.421	0.995	0.421
5	0.993	0.516	0.992	0.348	0.991	0.400
6	0.993	0.348	0.995	0.522	0.995	0.500
7	0.992	0.200	0.992	0.190	0.994	0.236

Table 5. SECOM Data Analysis Results

No	Logistic regression		Random forest		Support vector machine	
	accuracy	F1 score	accuracy	F1 score	accuracy	F1 score
1	**0.923**	**0.414**	**0.926**	**0.476**	**0.923**	**0.444**
2	0.919	0.053	0.924	0.105	0.903	0.122
3	0.897	0.281	0.899	0.347	0.908	0.369
4	0.912	0.316	0.919	0.333	0.919	0.379
5	0.870	0.344	0.883	0.337	0.881	0.282
6	0.904	0.271	0.913	0.339	0.918	0.373
7	0.849	0.027	0.866	0.112	0.864	0.059

는 과적합을 방지하기 위해 C(Cost function)와 Gamma를 활용하였다. 여기서 Gamma는 클수록 정확하고, 작을수록 과적합을 방지할 수 있다.

5.3 분석 결과

2개의 데이터 세트를 대상으로 7가지 전처리 방법을 적용한 후, 3개의 데이터 알고리즘을 적용하여 분석함으로써 총 42개의 결과를 비교하여 정리하였다.

<Table 4>는 사출성형 데이터 세트의 성능을 비교한 것이다. 정확도는 실제 양품을 모두 맞춘 2번 조건의 로지스틱 회귀가 가장 높았지만, F1 점수가 상대적으로 낮았다. 본 연구에서 제안한 데이터 전처리 프로세스를 수행한 후에 서포트벡터머신을 적용한 것의 F1 점수가 0.700으로 가장 높았으며, 정확도도 0.997로 상대적으로 높았다. <Table 5>는 반도체 제조 데이터 세트의 성능을 비교한 결과이다. 본 논문에서 제안한 전처리 프로세스를 거친 후 랜덤포레스트를 적용한 것이 정확도와 F1 점수 측면에서 성능이 가장 좋았다. 특이한 점은 데이터 불균형 처리를 적용하지 않은 2번 방법이 부적합품을 올바르게 분류하지 못하여 다른 방법에 비하여 F1 점수가 특히 낮았다. 부적합품이 나타나는 것을 제대로 알아내지 못하여 부적합품을 양품으로 오판하여 고객에게 납품하면 그 수가 적더라도 고객만족도에 악영향을 미칠 수 있다.

6. 결론 및 추후 연구 과제

데이터 분석 모델에 관한 연구는 많지만 데이터 전처리를 체계적으로 적용하는 방법에 관한 연구는 미흡한 실정이다. 데이터를 통해 의미 있는 결론을 도출하기 위해서는 모델 개선보다 데이터를 개선하는 것이 더 중요하다. 본 연구에서는 제조 품질데이터를 전처리하여 분석 성능을 높이는 데 초점을 두고 현장의 데이터 분석가나 엔지니어들을 위한 데이터 전처리 프로세스를 제안하였다.

우선 데이터 전처리에 사용되는 다양한 기법들을 살펴보고, 이들을 비교, 분석하여 특징을 정리한 다음, 기존 연구내용을 종합하여 효율적으로 공정데이터를 전처리할 수 있는 지침을 제시하였다. 데이터 전처리 프로세스 단계를 데이터 통합, 이상치 및 결측치 처리, 데이터 변환, 피처엔지니어링, 데이터 불균형 처리 순서로 정하고, 각 단계의 활용방법과 세부기준을 제시하였으며, KAMP에서 제공한 사출성형 데이터 세트와 UCI의 SECOM 데이터 세트를 활용하여 본 연구에서 제안한 데이터 전처리 프로세스의 성능을 확인하였다. 제안된 데이터 전처리 프로세스 적용하여 분석한 결과 F1 점수가 각각 0.700, 0.476으로 전처리하지 않았을 때의 0.236, 0.112 보다 우수한 성능을 나타내었다.

본 논문의 전처리 단계에서 제시한 기준은 보수적으로 설정하였으며, 추후 현장 데이터 분석가와 엔지니어의 실무 경험을 바탕으로 보완하는 것이 필요하다. 또한 기업이 실제로 이러한 지침을 효과적으로 활용하기 위해서는 기계부품, 반도체, 화학 등 제조공정의 유형, 센서 등 수집장치, 온도, 습도, 진동, 전자파, 분진 등 환경 변수의 영향, 디지털 전환 수준 등 다양한 요소를 고려하여 제조 현장에 적합한 데이터 전처리 프로세스를 개발하여 적용해야 할 것이다. 또한 본 논문에서 예시한 사례에서 레이블 데이터가 양·불량 판정 결과만 다루었는데, 다양한 유형의 품질 데이터를 대상으로 전처리를 하는 후속 연구가 이루어지기를 기대한다.

참고문헌

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002), SMOTE: synthetic minority over-sampling technique, *Journal of Artificial Intelligence Research*, **16**, 321-357.

Cho, E., Chang, T. W., and Hwang, G. (2022), Data Preprocessing Combination to Improve the Performance of Quality Classification in the Manufacturing Process, *Electronics*, **11**(3), 477.

Dua, D. and Graff, C. (2019), *UCI Machine Learning Repository*, Irvine, CA.

Escobar, C. A., McGovern, M. E., and Morales-Menedez, R. (2021), Quality 4.0: A Review of Big Data Challenges in Manufacturing, *Journal of Intelligent Manufacturing*, **32**, 2319-2334.

García-Laencina, P. J., Abreu, P. H., Abreu, M. H., and Afonoso, N. (2015), Missing Data Imputation on the 5-year Survival Prediction of Breast Cancer Patients with Unknown Discrete Values,

Computers in Biology and Medicine, **59**, 125-133.
Graham, J. W. (2009), Missing Data Analysis: Making It Work in the Real World, *Annual Review of Psychology*, **60**, 549-576.
Hall, D. L. and Llinas, J. (1997), An Introduction to Multi Sensor Data Fusion, *Proceedings of the IEEE*, **85**(1), 6-23.
Jun, S. H. (2008), An Outlier Data Analysis using Support Vector Regression, *Journal of Korean Institute of Intelligent Systems*, **18**(6), 876-880.
Kim, J. H. and Kwahk, K. Y. (2022), Class Imbalance Resolution Method and Classification Algorithm Suggesting Based on Dataset Type Segmentation, *Journal of Intelligence and Information System*, **28**(3), 23-43.
Kim, S. D., Cho, H. C., Kim, C. Y., Yang, T. H., and Kim S. H. (2018), Publish-Subscribe Model Based Efficient Data Acquisition Architecture for Industrial Wireless Sensor Networks, *Proceedings of Symposium of the Korean Institute of Communications and Information Sciences*, 1336-1337.
Krishnamurthi, R., Kumar, A., Gopinathan, D., Nayyar, A., and Qureshi, B. (2020), An Overview of IoT Sensor Data Processing, Fusion, and Analysis Techniques, *Sensors*, **20**(21), 60-76.
Kwak, S. K. and Kim. J. H. (2017), Statistical Data Preparation: Management of Missing Values and Outliers, *Korean Journal of Anesthesiology*, **70**(4), 407-411.
Lee, J. Y. and Styczynski, M. P. (2018), NS-kNN: A Modified k-Nearest Neighbors Approach for Imputing Metabolomics Data, *Metabolomics*, **14**(12), 153.
Lyle, M. (2017), From Paper and Pencil to Industry 4.0: Revealing the Value of Data Through Quality Intelligence, *Quality Magazine*, **56**(10), 25-29.
Micci-Barreca, D. (2001), A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems, *ACM SIGKDD Explorations Newsletter*, **3**(1), 27-32.
Ministry of SMEs and Startups of Korea (2020), Molding AI Dataset, Korea AI Manufacturing Platform (KAMP), Available at: https://kamp-ai.kr.
Ng, A. (2021), A chat with Andrew on *MLOps: From Model-centric to Data-centric AI*, Available on-line: https://www.youtube.com/watch?v=06-AZXmwHjot=1607s.
Scheffer, J. (2002), Dealing with Missing Data, *Research Letters in the Information and Mathematical Sciences*, **3**(1), 153-160.
Schmitt, P., Mandel, J., and Guedj, M. (2015), A Comparison of Six Methods for Missing Data Imputation, *Journal of Biometrics and Biostatistics*, **6**(1), 1-6.
Shah, J. S., Rai, S. N., DeFilippis, A. P., Hill, B. G., Bhatnagar, A., and Brock, G. N. (2017), Distribution Based Nearest Neighbor Imputation for Truncated High Dimensional Data with Applications to Pre-clinical and Clinical Metabolomics Studies, *BMC Bioinformatics*, **18**(1), 114.
Zhang, X., Li, R., Zhang, B., Yang, Y., Guo, J., and Ji, X. (2019), An Instance-Based Learning Recommendation Algorithm of Imbalance Handling Methods, *Applied Mathematics and Computation*, **351**, 204-218.

저자소개

서호진: 경상국립대학교에서 산업공학 학사와 석사학위를 받았고, 현재 해성디에스에서 근무하고 있다. 관심분야는 품질빅데이터 분석, 실험계획법이다.

김도현: KAIST에서 산업공학 학사, 석사 및 박사학위를 취득하였고, 현재 명지대학교 산업경영공학과에서 부교수로 근무하고 있다. 관심분야는 데이터 마이닝, 통계적 기계학습, 데이터 분석공학이다.

변재현: 서울대학교에서 산업공학 학사, KAIST에서 산업공학 석사 및 박사학위를 취득하였고, 현재 경상국립대학교 산업시스템공학부에서 교수로 근무하고 있다. 관심분야는 실험계획법, 품질경영, 데이터 분석공학이다.