Journal of Korean Society of Industrial and Systems Engineering
https://doi.org/10.11627/jksie.2023.46.3.139

Vol. 46, No. 3 : 139-147, September 2023
ISSN : 2005-0461 / EISSN : 2287-7975

Application and Comparison of Data Mining Technique to Prevent Metal-Bush Omission

Sang-Hyun Ko* · Dongju Lee**

*Myungha Tech, **Department of Industrial & Systems Engineering, Kongju National University

메탈부쉬 누락예방을 위한 데이터마이닝 기법의 적용 및 비교

고상현* · 이동주**†

*명하테크, **공주대학교 산업시스템공학과

The metal bush assembling process is a process of inserting and compressing a metal bush that serves to reduce the occurrence of noise and stable compression in the rotating section. In the metal bush assembly process, the head diameter defect and placement defect of the metal bush occur due to metal bush omission, non-pressing, and poor press-fitting. Among these causes of defects, it is intended to prevent defects due to omission of the metal bush by using signals from sensors attached to the facility. In particular, a metal bush omission is predicted through various data mining techniques using left load cell value, right load cell value, current, and voltage as independent variables. In the case of metal bush omission defect, it is difficult to get defect data, resulting in data imbalance. Data imbalance refers to a case where there is a large difference in the number of data belonging to each class, which can be a problem when performing classification prediction. In order to solve the problem caused by data imbalance, oversampling and composite sampling techniques were applied in this study. In addition, simulated annealing was applied for optimization of parameters related to sampling and hyper-parameters of data mining techniques used for bush omission prediction. In this study, the metal bush omission was predicted using the actual data of M manufacturing company, and the classification performance was examined. All applied techniques showed excellent results, and in particular, the proposed methods, the method of mixing Random Forest and SA, and the method of mixing MLP and SA, showed better results.

Keywords : Machine Learning, Metal-Bush Assembly, Simulated Annealing

1. 서 론

메탈부쉬 조립 공정은 회전하는 구간에서 안정감 있는 압착과 이음 발생을 감쇄해주는 역할을 하는 메탈부쉬를 삽입하여 압착하는 공정이다. 메탈부쉬 조립공정에서는 메탈부쉬 누락, 미압입, 압입불량으로 인한 메탈부쉬의 두경 및 두고 불량이 발생되고 있다. 이러한 불량 원인 중 공정에 부착된 센서들의 신호를 이용하여 메탈부쉬 누락으로 인한 불량을 예방하고자 하며, 추후 미압입, 압입불량의 원인으로 인한 불량도 본 연구를 확장하여 예측하고자 한다.

메탈 부쉬란 얇은 철판 겉면에 테프론 코팅처리를 한 것으로써 주로 트인 원형인데, 그 이유는 압착이 되어야 하므로 터진 부분이 필요하기 때문이다. <Figure 1>에 다양한 크기의 메탈부쉬들이 주어져 있으며, <Figure 2>에는 메탈부쉬가 자동차 시트(Car Seat)의 부품 중 하나인 프론트 링크(Front Link)부품에 조립된 모습이 주어져 있다.

Received 25 July 2023; Finally Revised 2 August 2023; Accepted 7 August 2023
† Corresponding Author : djlee@kongju.ac.kr

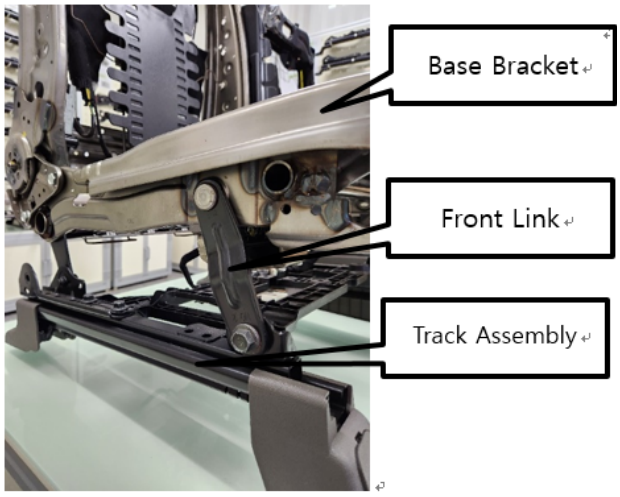
<Figure 1> Metal Bushes in Different Sizes

<Figure 2> Assembled Metal Bush in Front Link

<Figure 3>에는 자동차 시트의 하단 부위의 모습이 보여지고 있는데, 베이스 브랫킷(Base Bracket)과 트랙 어셈블리(Track Assembly)를 연결하는 프론트 링크의 모습을 확인할 수 있다. 또한, 이러한 프론트 링크에 메탈부쉬(Metal Bush)가 체결되어 있는 것을 확인할 수 있다.

프론트 링크는 트랙 어셈블리로부터 시트 프레임(Seat Frame)을 띄워주는 회전 지렛대 역할을 하고 있어, 탑승자의 안전과 직결되어 있으며, 탑승자가 시트의 height 기능을 사용할 경우 시트의 움직임이 부드러운지 아닌지의 여부는 메탈부쉬의 조립상태에 상당한 영향을 받고 있다. 이렇듯, 메탈 부쉬는 중요한 역할을 담당하는 부품이며, 메탈 부쉬의 누락 여부와 메탈 부쉬 압입 품질은 일관성 있게 유지되어야 한다.

제조현장에서 흔히 발생하듯이, 메탈부쉬 누락 불량의 경우에도 불량데이터의 확보가 어려워 데이터불균형 문제가 발생한다. 데이터 불균형이란 클래스 별 속하는 데이터의 수의 차이가 많은 경우를 의미하는데, 분류 예측을 할 때 문제가 될 수 있다.



<Figure 3> Front Link that Connects Base Braket and Track Assembly in Car Seat

본 연구에서는 메탈부쉬 누락 여부를 종속변수로 하고, 좌/우 로드셀 값, 전류, 전압 데이터를 독립변수로 하여 메탈부쉬 조립공정에서 메탈부쉬 누락을 예측하고자 한다.

본 논문의 구성은 다음과 같다. 이어지는 제2장에서는 선행 연구들을 살펴보고, 제3장에서는 분류예측을 위해 사용된 기법들에 대해 소개하였다. 제4장에서는 M사의 부쉬누락과 관련된 데이터에 대해 설명하고, 제3장에서 소개한 기법들을 적용한 결과를 보여주었다. 마지막으로, 제5장에서는 결론과 향후연구방향에 대해 논하고자 한다.

2. 문헌 연구

데이터 불균형으로 발생하는 문제를 해결하기 위한 샘플링 기법으로는 오버샘플링(oversampling), 언더샘플링(undersampling), 복합샘플링 기법이 있다. 소수 클래스의 데이터 수를 증가시키는 방법이 오버샘플링 방법이며, 반대로 다수 클래스의 데이터 수를 감소시키는 방법이 언더샘플링 방법이며, 오버샘플링과 언더샘플링을 모두 적용시키는 방법이 복합샘플링이다. 오버샘플링은 데이터 수가 증가하므로 모형 구축에 시간이 더욱 소모되고, 과적합(Overfitting)이 생길 수 있으며, 언더샘플링은 정보손실이 생길 수 있다.

오버샘플링 기법으로는 ROS (Random Over Sampling), SMOTE(Synthetic Minority Oversampling Technique)[1], Borderline SMOTE[2], ADASYN(Adaptive Synthetic Sampling)[4] 등이 있고, 복합샘플링으로는 SMOTE + ENN, SMOTE + Tomek 등이 있다.

다양한 분야에서 샘플링 기법을 적용하여 데이터불균

형 문제에 대처하는 연구가 이뤄지고 있다.

Han and Joe[3]는 데이터불균형 문제를 가지고 있는 Kaggle의 신용카드사기 데이터를 활용하여 오버샘플링 기법 중 하나인 SMOTE와 Gradient Boosting 트리 알고리즘인 Light GBM을 적용한 해법을 제시하였다. Lee[7]는 가계금융 복지조사 데이터를 활용하여 가계부채 상환 연체 가능성이 높은 가계를 탐지하기 위해 이항 로지스틱 회귀모형과 의사결정나무모형을 적용하였다. 특히, 데이터불균형 문제를 해결하기 위해 오버샘플링과 언더샘플링을 적용하였는데, 오버샘플링을 적용한 경우 분류모형의 민감도가 향상되었다. Shin et al.[11]은 데이터불균형 문제를 해결하기 위한 데이터 증강 기법으로 사용되는 생성적 적대 신경망(GAN)에서 생성 데이터의 품질 향상을 위해 유사도를 통한 거리 학습 방법을 변형한 해법을 제시하고, 소수 클래스들의 학습 부족 문제를 개선하였다. Oh and Lee[9]는 신용예측 데이터의 불균형 문제를 해결하기 위해 생성적 적대 신경망(GAN)을 적용하고, KNN기반의 오버샘플링 기법보다 우수함을 보였다. Moon and Kim [8]은 환경데이터를 활용한 적기 서리 예측 모델에서 데이터 불균형 문제를 해결하기 위해 SMOTE 방법을 적용하고, 분류기법으로는 AdaBoost, SVM, ANN, Random Forest를 적용하였다. 적용기법 중 Random Forest가 가장 우수한 결과를 보여 주었다.

한편, 제조현장에서도 자주 발생하는 데이터 불균형 문제를 해결하기 위해 샘플링 기법을 적용한 연구들이 행해졌다.

Park et al.[10]은 생분해성 섬유 방사공정 데이터를 이용한 인장 강신도 예측 문제에서 데이터불균형 문제를 해결하기 위한 데이터 증강 기법을 제시하였다. 이 기법은 각 변수에 대해 상관계수의 크기와 데이터 불균형을 고려하여 우선순위를 결정하고, 그에 따른 증강 비율로 소수 클래스의 데이터를 증강하는 기법이다.

Lee[6]는 사출성형공정에서의 가공불량을 예측하는 문제에서 데이터불균형을 해결하기 위해 다양한 오버샘플링 기법과 복합샘플링 기법을 적용하였으며, 예측기법으로는 다층퍼셉트론(MLP, Multilayer Perceptron)과 장단기메모리(LSTM, Long Short Term Memory)를 적용하였다. 샘플링 방법과 적용된 기법들의 하이퍼 파라미터 최적화를 위해 담금질 모사(SA, Simulated Annealing)를 적용하였다.

담금질 모사기법은 Kirkpatrick et al.[5]이 제안하였는데, 전역최적해(Global Optimum)에 근사한 해를 구하기 위해 자주 사용되는 최적화 기법이다.

분류 예측을 위해서는 다양한 데이터마이닝 기법들이 적용될 수 있는데, 본 연구에서는 로지스틱회귀모형, 랜덤포레스트(Random Forest), MLP(Multilayer Perceptron)을 적용하였다.

본 연구에서는 Lee[6]의 연구를 개량한 방법론을 메탈부쉬 조립공정에서의 메탈부쉬 누락 예측을 위해 적용하였다. 메탈부쉬 누락 예측을 위해 다양한 데이터마이닝 기법을 적용하고, 최적의 방법을 탐색하였는데, SA기법과 Random Forest를 결합한 방법론을 제안하고 Random Forest의 하이퍼파라미터값 선정, 샘플링기법의 선정과 샘플링기법의 파라미터 값의 선정을 동시에 고려하였다. 또한, SA기법과 MLP를 결합한 방법론을 제안하였는데, Lee[6]의 연구에서는 한 개의 변수를 이용하여 MLP의 은닉층(Hidden Layer)의 노드 수를 설정하였는데, 본 연구에서는 개별 은닉층의 노드 수를 각각 변수로 설정하여 최적 은닉층의 수를 구하였다.

3. 적용된 데이터 마이닝 기법

본 연구에서 고려한 기법으로는 로지스틱 회귀 분석, Random Forest, 다층퍼셉트론이 있다. 또한, 제안하는 기법으로는 Random Forest와 SA를 혼용한 기법과 다층퍼셉트론과 SA를 혼용한 기법이 있다.

3.1 분류 예측을 위한 기존의 기법

로지스틱 회귀 분석(Logistic Regression Analysis):

로지스틱 회귀 분석은 선형회귀분석과 유사하게 종속변수와 하나 이상의 독립변수간의 관계를 파악하는데 사용되는데, 어떤 사건이 발생할 확률을 추정하여 분류 및 예측 분석에 자주 사용된다. 종속변수가 발생하거나 발생하지 않을 승산비(odds)의 자연로그값을 로짓(logit)이라고 하는데, 로지스틱 회귀는 종속변수를 독립변수에 관한 로짓변수로 변환하여 최대 우도 추정기법을 적용한다.

로짓은 다음과 같다

$$\log\left(\frac{p}{1-p}\right)=\beta_0+\beta_1X_1+\beta_2X_2...+\beta_nX_n$$

여기서 p는 어떤 사건이 발생할 확률이며, X_i는 독립변수 i이고, β_i는 X_i의 계수이다.

Random Forest:

랜덤으로 일부 독립변수만을 선택하여 의사결정나무를 다수 만들고, 여러 개의 의사결정나무에서의 예측값을 기준으로 가장 자주 나온 예측값을 최종 예측값으로 선정하는 방법으로 Kaggle에서 우승하는데 여러 번 사용된 기법이다. Random Forest는 여러 개의 결과를 합쳐 최종결과를 도출하는 의사결정 나무의 앙상블기법이다.

다층퍼셉트론(MLP, Multilayer Perceptron):

지도학습에 사용되는 인공 신경망 중 하나로, 분류 및 회귀 문제에 적용되는 기법이다. MLP는 퍼셉트론으로 이뤄진 층(layer) 여러 개를 순차적으로 연결하였는데 입력 벡터가 있는 층을 입력층, 최종 출력값이 있는 층을 출력층이라 부르며, 입력층과 출력층 사이에 존재하는 중간층을 은닉층(hidden layer)이라고 한다. 여러 개의 은닉층이 있는 인공 신경망을 심층 신경망(Deep Neural Network)이라고 부른다. 하나 이상의 비선형 은닉층은 복잡한 패턴을 추출하는데 유용하며, 복잡한 패턴을 학습하기 위해서는 노드의 개수와 은닉층의 수는 중요하다.

3.2 분류 예측을 위한 제안하는 기법

SA+RF(Random Forest):

분류 예측을 위해서 사용된 기법은 Random Forest이며, 오버샘플링과 복합샘플링 기법, 샘플링 기법들의 파라미터값, Random Forest의 하이퍼 파라미터 값의 최적화를 위해서 SA가 사용되었다.

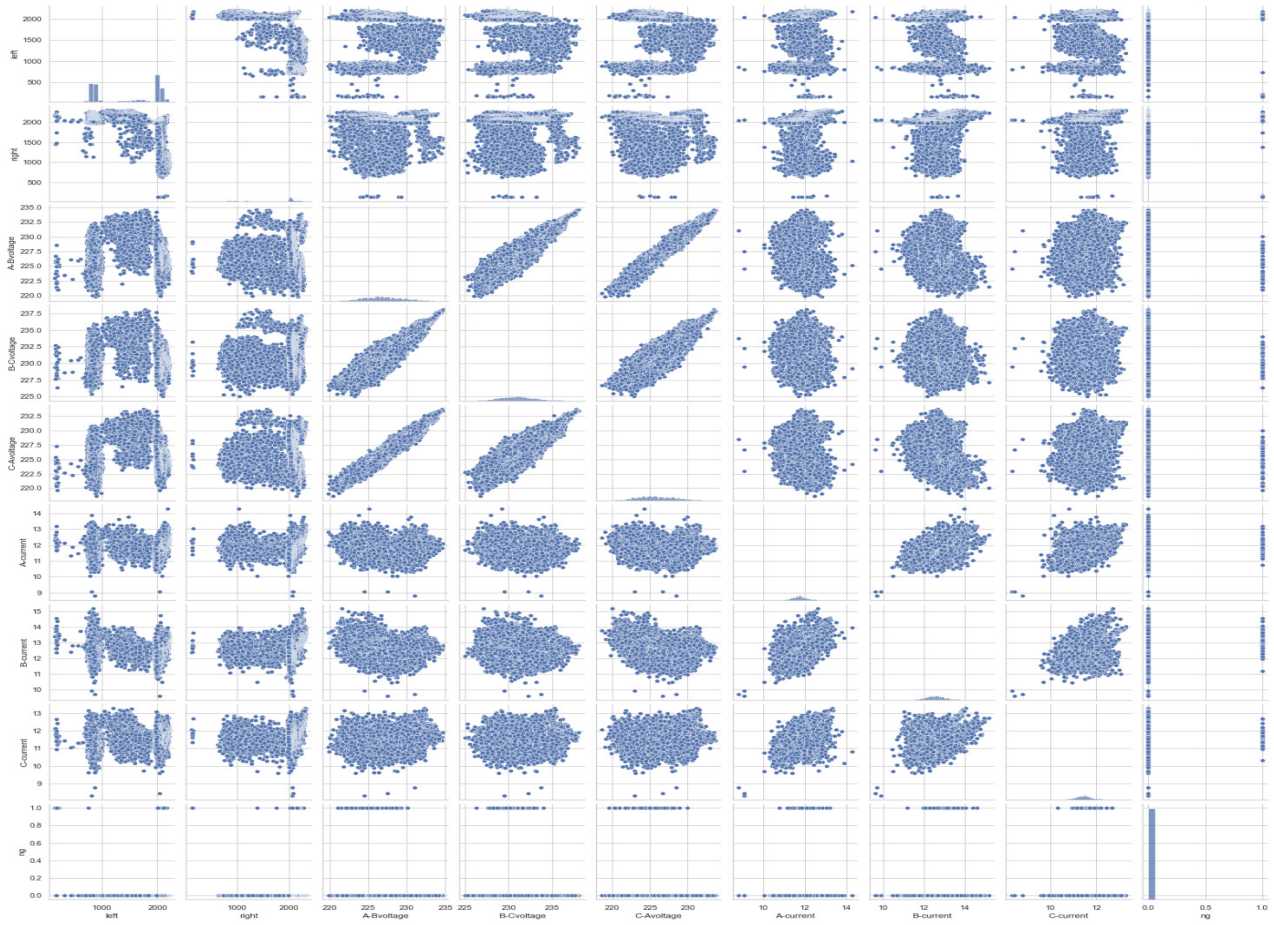
SA+MLP:

분류 예측을 위해서 사용된 기법은 MLP이며, 오버샘플링과 복합샘플링 기법, 샘플링 기법들의 파라미터값, MLP의 하이퍼 파라미터 값과 각 은닉층의 노드 수의 최적화를 위해서 SA가 사용되었다.

4. 실험 및 결과

분석에 사용한 데이터는 M사의 제품 DE와 제품 RG3-LH에 메탈부쉬를 조립하는 공정에 대한 데이터이다. 좌로드셀값(left), 우로드셀값(right), 3곳의 위치로 부터의 전류(A-current, B-current, C-current), 3곳의 위치로 부터의 전압(A-BVoltage, B-CVoltage, C-AVoltage)으로 구성된 8개의 독립변수와 부쉬누락여부(ng)에 해당하는 1개의 종속변수로 이뤄져 있다. 이때 종속변수는 부쉬누락여부(0: 양품, 1: 부쉬누락불량)이 사용되었다. 8개의 독립변수는 모두 연속변수이며, 종속변수는 이진변수이다.

DE 제품 조립의 양품 관측치 수는 19,404개, 불량품 관



<Figure 4> Scatter Plot of Each Pair of Variables for Product DE

측치 수는 42개이며, RG3-LH 제품 조립의 양품관측치 수는 16,011개, 불량품 관측치 수는 46개로 데이터 불균형이 심하다.

제품 DE에 대한 각 변수쌍별 산점도는 <Figure 4>에 주어져 있다.

분류를 위한 기법으로는 로지스틱 회귀, Random Forest, 다층퍼셉트론(MLP, Multi-layer Perceptron)이 적용되었다. 또한, 제안한 기법으로 Simulated Annealing과 Random Forest를 혼합한 SA+RF와 Simulated Annealing과 MLP를 혼합한 SA+MLP가 적용되었다.

70%의 데이터는 훈련(training)용으로, 나머지 30%의 데이터는 테스트(test)용으로 사용되었는데, 양품, 불량품 각각에 대해 해당 비율로 랜덤하게 선택되었다. 파이썬으로 코딩되었으며, 다층퍼셉트론의 경우에는 Keras package의 Sequential 모델을 이용해 구축하였다.

4.1 로지스틱 회귀분석

제품 DE와 RG3-LH에 대해 후진제거법으로 변수를 선정하였는데 기준은 각 변수의 $p \leq 0.05$이며 분산팽창계수(VIF, Variation Inflation Factor)는 5 이하가 되도록 하였다.

제품 DE의 경우 후진제거법으로 선택된 최종모형의 결과는 <Table 1>과 같다.

<Table 1> Final Logistic Regression Model for Product DE

Pseudo R-Squ.: 0.3180
Log-Likelihood(AIC): -141.42

	Coef.	Std err	z	P>|z|
Const	-15.047	6.681	-2.252	0.024
Left	-0.006	0.001	-9.047	0.000
Right	-0.007	0.001	-10.014	0.000
B-current	1.210	0.390	3.102	0.002
C-current	1.122	0.549	2.042	0.041

최종모형의 VIF(Variation Inflation Factor)는 left: 1.371, right: 1.298, B-current: 1.112, C-current 1.043으로 모두 5보다 작으므로 다중공선성 문제는 없다고 볼 수 있다. 전체모형과 최종모형을 비교해 볼 때 Log-Likelihood(AIC)가 -141.26에서 -141.42로 약간 좋아졌으며, C-statistic은 0.826에서 0.826으로 동일하고, 0.7보다 크므로 좋은 모형이라고 판단된다.

제품 RG3-LH의 최종모형은 <Table 2>와 같다.

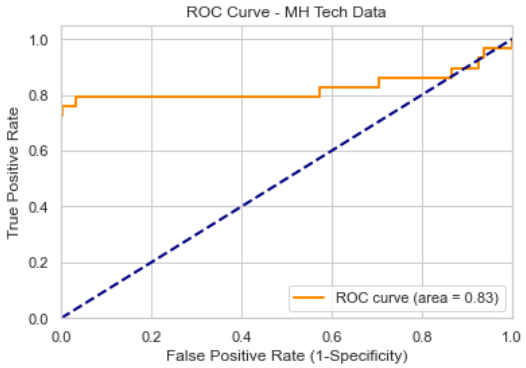
독립변수로는 좌로드셀값, 우로드셀값, B위치에서의 전류, C-A로의 전압이 선택되었고, p-value는 B위치에서의 전류, C-A로의 전압이 0.066, 0.062로 0.05보다 약간 크나 변수로 선정하는 것이 모형의 적합도를 높여주므로 선택하였다. 선택된 4개의 변수 VIF는 left: 2.157, right: 2.331, B-current: 1.442, C-A Voltage: 1.444로 모두 5보다 작으므로 다중공선성이 없다.

<Table 2> Final Logistic Regression Model for Product RG3-LH

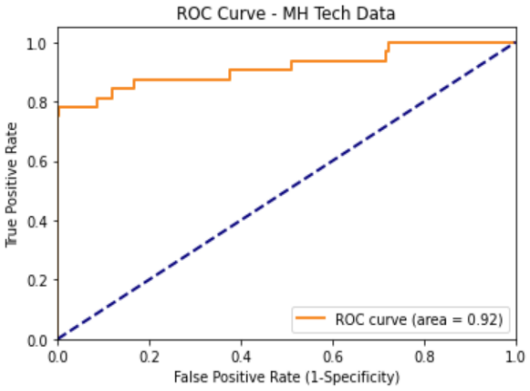
Pseudo R-Squ.: 0.6588
Log-Likelihood(AIC): -74.901

	Coef.	Std err	z	P>|z|
Const	-5.799	25.556	-0.227	0.820
Left	-0.003	0.001	-2.713	0.007
Right	-0.011	0.001	-9.575	0.000
C-A voltage	0.156	0.085	1.838	0.066
B-current	-1.465	0.784	-1.868	0.062

양품과 불량품을 결정하는 임계값(thresthold value)는 0.1부터 0.9까지 중 정확도(Accuracy)가 가장 높은 값을 선택하였는데, 제품 DE는 0.1이 제품 RG3-LH는 0.5가 선택되었다.



<Figure 5> ROC Curve of Logistic Regression Model for Product DE



<Figure 6> ROC curve of Logistic Regression Model for RG3-LH

전체모형과 최종모형을 비교해 볼 때 Log-Likelihood (AIC)가 -73.850에서 -74.901로 약간 좋아졌으며, C-statistic은 0.9197에서 0.9155로 약간 나빠졌으나 0.7보다 크므로 좋은 모형이라고 판단된다. <Figure 5>와 <Figure 6>에 ROC curve가 주어져 있는데, 좋은 예측결과를 보여준다는 것을 알 수 있다.

4.2 Random Forest

훈련데이터 중 20%는 모델의 성능과 하이퍼파라미터(hyper-parameter)들의 검정(validation)용으로 사용되었으며, k-겹 교차타당성(k-fold cross validation)이 적용되었는데, 이때 k=5가 사용되었다.

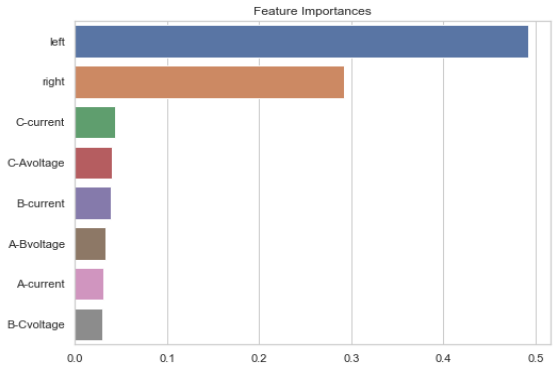
고려한 하이퍼파라미터들의 값들과 최적값은 <Table 3>과 같다. Grid Search를 적용하여 최적값을 선택하였다.

<Table 3> Considering Values and Optimal Values for Hyper-parameters of Random Forest

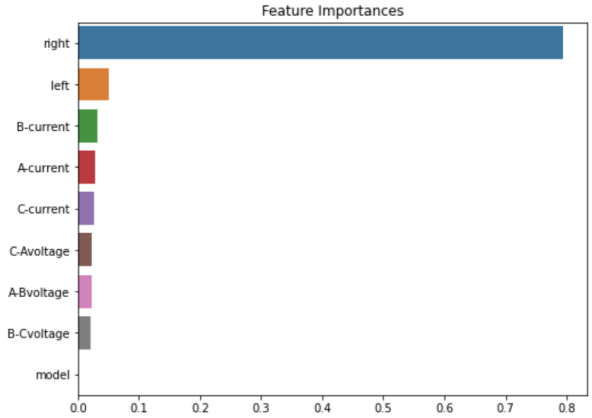
Hyper-parameter	Optimal Value		Considered Values
	DE	RG3-LH	
n_estimators	2000	1000	1000,2000,3000
max_depth	200	100	100,200,300
min_samples_split	2	2	2,3
min_samples_leaf	2	1	1,2

n-estimators는 결정트리의 개수이며, max_depth는 트리의 최대깊이, min_samples_split는 노드를 분할하기 위한 최소한의 샘플 데이터수, min_samples_leaf는 리프노드가 되기 위해 필요한 최소한의 샘플 데이터수이다.

지니의 평균값 감소를 이용한 변수 중요도 그래프는 <Figure 7>, <Figure 8>과 같다. 제품 DE의 경우에는 좌로드셀값, 우로드셀값, C위치에서의 전류 등의 순이었으며, 제품 RG3-LH는 우로드셀값, 좌로드셀값, B위치에서의 전류 등의 순이다.



<Figure 7> Feature Importance for Product DE



<Figure 8> Feature Importance for Product RG3-LH

4.3 다층퍼셉트론(MLP)

분류를 위한 인공신경망 기법으로 MLP가 사용되었다. 최적화기법으로는 Adam, 학습률은 0.001, 손실함수(loss function)로는 이진교차엔트로피(binary cross entropy), 에포크(epoch)은 100, 배치크기(batch size)는 32가 사용되었다. MLP는 4개의 은닉층이 있으며, 각 은닉층의 노드 수는 32, 64, 32, 16이 사용되었다. 활성화함수로는 은닉층에는 ReLU를, 마지막 신경망은 sigmoid 함수를 사용하였다. 과대적합을 방지하기 위해 2번째와 3번째 은닉층에 드롭아웃을 0.3으로 설정하였다.

4.4 SA + Random Forest

데이터 불균형이 심하므로 오버샘플링과 복합샘플링기법이 적용되었다. 오버샘플링 기법으로는 ROS (Random Over Sampling), SMOTE[1], Borderline -SMOTE[4], ADASYN(Adaptive Synthetic Sampling)이 사용되었고, 복합샘플링으로는 SMOTE+ENN, SMOTE+Tomek이 사용되었다.

SA를 이용하여 구하고자 하는 최적조합은 샘플링방법, 샘플링에서 달성하고자 하는 소수클래스의 비율, Random Forest의 4가지 하이퍼파라미터이다.

소수클래스의 비율로는 0.1, 0.2, 0.3, 0.4, 0.5가 고려되었고, Random Forest의 하아퍼파라미터 값으로 고려된 값을 다음과 같다. 결정트리의 개수(n-estimators)로는 200, 400, …, 2000, 트리의 최대깊이(max_depth)로는 10, 20, …, 90, 노드를 분할하기 위한 최소한의 샘플 데이터수(min_samples_split)로는 2, 3, 4, 리프노드가 되기 위해 필요한 최소한의 샘플 데이터수(min_samples_leaf)로는 1,2,4가 고려되었다. 이때, 가능한 조합의 수는 6×5×10×9×3×3= 24,300이다.

SA에서 사용한 파라미터들은 최대반복횟수 = 200으로 설정하였으며, 해의 수렴에 영향을 미치는 냉각스케줄은 $T_k = \alpha \times T_{k-1}$이다. 여기서 k는 반복횟수를 의미한다. T_k는 반복횟수 k에서의 온도이다. α = 0.95, 초기온도 T_0 = 1, 최소온도 T_F = 0.01로 설정하였다.

4.5 SA + MLP

SA의 파라미터 설정, 샘플링 방법, 소수 클래스 비율은 4.4절과 동일하다. 또한, 다층퍼셉트론의 하이퍼파라미터인 배치 크기(Batch size)는 20, 30, …, 120을 고려하였고, 각 은닉층의 노드 수로는 10, 12, 14,…, 42를 고려하였다. 이때 가능한 조합의 수는 6×5×10×17×17×17×17=25,056,300 이다.

4.6 결과

3개의 기존 기법들과 2개의 제안한 기법들을 테스트 데이터 셋에 적용하였을 때의 혼동행렬은 <Table 4>에 주어져 있다.

분류가 얼마나 잘 분류되었는지 확인하는 성과척도로는 Accuracy(정확도), Precision(정밀도), Recall(재현율), F1 score가 있는데, 이들을 계산하는 식은 다음과 같다.

$$\text{Accuracy} = \frac{TP+FN}{TP+TN+FP+FN}$$

$$\text{precision} = \frac{TP}{TP+FP}$$

$$\text{recall} = \frac{TP}{TP+FN}$$

$$\text{F1 score} = 2 \times \frac{recall \times precision}{recall + precision}$$

여기서, TP는 True Positive, FN은 False Negative, FP는 False Positive, TN은 True Negative를 의미한다.

F1 score는 데이터불균형이 존재하는 경우 성과척도로 종종 사용되므로, F1 score를 위주로 사용된 기법의 우수성을 판단하였다.

Accuracy, Precision, Recall, F1 Score는 <Table 5>과 <Table 6>에 각각 주어져 있다.

제안하는 SA+RF, SA+MLP기법에서 좋은 결과를 보여준 조합은 <Table 7>, <Table 8>에 각각 주어져 있다.

<Table 4> Confusion Matrix for Each Method

Sampling Method		DE			RG3-LH		
			Predictive Values			Predictive Values	
			Positive	Negative		Positive	Negative
Logistic Regression	Actual Values	Positive	5818	3	Positive	4804	0
		Negative	5	8	Negative	5	9
Random Forest	Actual Values	Positive	5819	2	Positive	4804	0
		Negative	4	9	Negative	4	10
MLP	Actual Values	Positive	5820	1	Positive	4804	0
		Negative	4	9	Negative	4	10
SA+ RF	Actual Values	Positive	5820	1	Positive	4804	0
		Negative	4	9	Negative	3	11
SA+MLP	Actual Values	Positive	5821	0	Positive	4804	0
		Negative	4	9	Negative	3	11

<Table 5> Accuracy, Precision, Recall, and F1 Score for Product DE

	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.999	0.727	0.615	0.667
Random Forest	0.999	0.818	0.692	0.750
MLP	0.999	0.900	0.692	0.783
SA+RF	0.999	0.900	0.692	0.783
SA+MLP	0.999	1.000	0.692	0.818

<Table 6> Accuracy, Precision, Recall, and F1 Score for Product RG3-LH

	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.999	1.000	0.643	0.783
Random Forest	0.999	1.000	0.714	0.833
MLP	0.999	1.000	0.714	0.833
SA+RF	0.999	1.000	0.786	0.880
SA+MLP	0.999	1.000	0.786	0.880

<Table 7> Best Combinations for SA+ RF by Product Type

	SA+RF	
	DE	RG3-LH
Sampling Method	ROS	ROS
Minority Ratio	0.3	0.5
n_estimators	400	1600
max_depth	60	70
min_samples_split	3	3
min_samples_leaf	2	2

<Table 8> Best Combinations for SA+ MLP by Product Type

	SA+ MLP	
	DE	RG3-LH
Sampling Method	ADASYN	ROS
Minority Ratio	0.4	0.4
Batch size	100	70
# of units in hidden layer 1, 2, 3, 4	30, 16, 40, 34	18, 36, 34, 24

실험결과를 요약하면 다음과 같다.

- 로지스틱 회귀분석. 랜덤 포레스트, 다층퍼셉트론은 정확도(Accuracy)가 아주 높으며, 데이터 불균형을 고려할 때 자주 사용되는 F1 Score에서도 0.5 이상의 값을 나타내었다.
- 샘플링 방법과 제안하는 SA 기법들은 일부의 실제불량을 예측할 수 있었다. 하지만, DE는 불량갯수 13개 중 9개, RG3-LH는 14개중 11개의 실제 불량만 불량으로 예측할 수 있었다. 즉, 불량을 모두 예측할 수 있는 방법은 없었으며, 좀 더 많은 데이터와 추가 학습이 필요하다.
- 샘플링 기법들 중 우수한 결과를 나타내는 기법은 ROS와 ADASYN이었다.
- 제안하는 SA는 Random Forest와 MLP 모두에서 F1 Score가 가장 좋은 결과를 보여 주었으며, 특히, SA+MLP는 모두 0.8 이상이었다.
- 복합샘플링보다는 언더샘플링이 더 나은 결과를 보여주었다.

5. 결론 및 연구과제

회전하는 구간에서 안정감 있는 압착과 이음 발생을 감쇄해주는 역할을 하는 메탈부쉬를 삽입하여 압착하는 메탈부쉬 조립 공정에서 메탈부쉬의 누락 여부를 다양한 기법들을 적용하여 예측하였다.

메탈부쉬 누락불량 예방을 위해서는 현재의 결과대로 현장에 적용가능하다고 판단되나, 압입불량 발생시의 데이터를 추가적으로 수집하여 압입불량 예측이 가능하다면 이를 추가하여 현장에 적용하는 것이 적절할 것으로 보인다.

또한, 제조데이터에서 자주 발생하는 데이터 불균형 문제를 해결하기 위해 언더샘플링, 복합샘플링을 적용하고, 샘플링과 관련된 파라미터와 부쉬누락예측을 위해 사용된 기법들의 하이퍼 파라미터 최적화를 위해 SA를 적용하였다.

적용된 모든 기법들은 우수한 결과를 보여주었으며, 특히 제안한 기법들인 Random Forest와 SA를 혼용한 방법, MLP와 SA를 혼용한 방법들은 좀 더 우수한 결과를 보여주었다. 하지만, 몇몇 부쉬누락 불량의 경우 정확히 예측하지 못하는 경우가 발생하고 있다. 추가적인 데이터 확보와 센서 도입을 통해 예측의 정확도를 더욱 향상시킬 필요가 있다.

한편, 메탈부쉬 조립불량은 조립불량 데이터 확보가 아주 어려워 다루지 않았다. 조립불량이 발생한 상태에서의 각 센서로부터의 값들을 확보한다면 조립불량 예측도 가능하리라 판단된다.

제조공정에서의 가공불량이나 조립불량과 같은 데이터는 데이터 불균형 문제가 심각하다. 이러한 문제에 대처하기 위한 다양한 연구들이 필요하며, 제조현장의 데이터를 이용한 예측에 적합한 기법에 대한 연구도 필요하다.

References

[1] Chawla, N.V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P., SMOTE: Synthetic Minority Over-sampling Technique, *Journal of Artificial Intelligence Research*,

2002, Vol. 16, pp. 321-357.

[2] Han, H., Wang, W.-Y., and Mao, B-H, Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning, *Proceedings of ICIC 2005*: Advances in Intelligent Computing, 2005, pp. 878-887.

[3] Han, Y.J. and Joe, I.W., Imbalanced Data Improvement Techniques Based on SMOTE and Light GBM, *KIPS Trans. Comp. and Comm. Sys.*, 2022, Vol. 11 No.12, pp. 445-452

[4] He, H., Bai, Y., Garcia, E., and Li, S., ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning, *International Joint Conference on Neural Networks*, 2008, Vol. 3, pp. 1322-1328.

[5] Kirkpatrick, S., Gelatt, C.D., and Vecchi, M.P., Optimization by Simulated Annealing, *Science*, 1983, Vol. 220, No. 4598, pp. 671-680.

[6] Lee, D.J., Simulated Annealing for Overcoming Data Imbalance in Mold Injection Process, *J of KSIE*, 2022, Vol. 45, No. 4, pp. 233-239.

[7] Lee, J.H., Machine Learning Applications to Households Insolvency with Imbalanced Data, *J of Consumer Studies*, 2019, Vol. 30, No. 6, pp. 97-118.

[8] Moon, A.K. and Kim, H.S., Microclimate-Based Frost Prediction Model Resolving the Class Imbalance, *J of Korea Ins. of Comm. And Inf. Sci.*, 2022, Vol. 47, No.10, pp. 1704-1715.

[9] Oh, S.M. and Lee, J.H., Virtual Data Generation Techniques for Imbalance problem of credit prediction data, *Proceedings of KICS*, Yongpyung, 2023.02.08-10, pp. 874-875.

[10] Park, S.C., Kim, D.Y., Seo, K.B., and Lee, W.J., The Development of Biodegradable Fiber Tensile Tenacity and Elongation Prediction Model Considering Data Imbalance and Measurement Error, *KIPS Trans. Softw. and Data Eng.*, 2022, Vol. 11, No. 12, pp. 489-498.

[11] Shin, H.J., Lee, S.B., and Lee, K.C., Improving the Quality of Generating Imbalance Data in GANs through an Exhaustive Contrastive Learning, *J of KIISE*, 2023, Vol. 50, No.4, pp. 295-305. 2023.

ORCID

Dongju Lee | http://orcid.org/0000-0001-6650-9270
Sang-Hyun Ko | https://orcid.org/0009-0008-4088-6930