스마트팩토리 데이터를 활용한 클래스 불균형문제 해결 방법 비교

최은선, Kong Vungsovanreach, 조완섭, 김재성, 이경희, 손호선, 최성곤*

충북대학교

{tmxk147, kv.sovanreach, shon0621, lee.kyunghee}@gmail.com,

{wscho, comkjsb, sgchoi*}@cbnu.ac.kr

Comparison of Model for Class Imbalance Problem using Smart Factory Data

Eun-Seon Choi1, Kong Vungsovanreach, Wan-Sup Cho, Jae-Sung Kim, Kyung Hee Lee, Ho Sun Shon,

Seong Gon Choi*

Chungbuk National University

요 약

스마트팩토리는 ICT 기술을 전체 생산 공정에 적용해 실시간으로 수집된 대량의 데이터를 분석 · 관리할 수 있는 지능형 생산공장으로 생산성과 효율성을 높일 수 있다는 점에서 제조업계의 큰 이슈로 떠오르고 있다. 이에 따라 다양한 제조 산업에서도 스마트공장에 대한 연구가 진행 중이다. 다만 제조 산업 별로 스마트팩토리의 전개 양상이 상당히 다르므로 각 기업이 속한 산업에 효과적인 도입 전략이 필요하다. 본 연구에서는 조립 가공산업에서 사용되는 기계의 데이터를 사용하여, 기계 데이터의 특성인 클래스 불균형 문제를 해결하기 위한 모델을 구축하고 모델 간 성능 비교를 통해 후행연구로 진행될 스마트팩토리 플랫폼 구축의 기반을 다지고자 한다 .

Ⅰ. 서 론

클라우스 슈밥(Klaus Schwab)의 제4차 산업혁명 선언이후 빅데이터(Big data)는 그 가치가 더 크게 확장되고 있으며 최근에는 생산 제조 기술과 정보 통신 기술(ICT)이 융합된 스마트 팩토리가 국내에서 큰 주목을 받고 있다[1]. 주요 선진국은 경제에서 제조업의 중요성을 새롭게 인식하면서 자국의 제조업 경쟁력을 높이기 위해 다양한 노력을 전개하고 있다. 독일은 Industry 4.0 슬로건 하에 국가 차원에서 제조업 경쟁력을 높이기 위해 정부와 기업, 학계가 협업하며 혁신 방안을 모색하고 있다. 미국은 GE를 중심으로 주요 ICT 기업이 협업하여 클라우드 기반의 공장운영 플랫폼을 개발하면서 기업 주도로 제조업 혁신을 추진하고 있다. 각국의 제조업 혁신 노력과 함께 사물인터넷(IoT) 기술, Big Data 저장 및 분석기술, 산업용 로봇 기술 또한 발전하고 있다. 사물인터넷(IoT) 기술은 공장 설비의 각종 제어기 데이터와 센서류 데이터를 인터넷과 연결하여 수집할 수 있도록 지원하며, 하둡을 비롯한 Big Data 저장 및 분석기술이 발달하면서 대규모 데이터를 저렴한 비용으로 저장, 분석하는 것이 용이해졌다. 또한 산업용 로봇 기술의 발전으로 로봇 생산 비용과 로봇의 작업 범위가 개선되면서 로봇을 적용한 공장 자동화가 점차적으로 진전되고 있는 상황이다.

기존의 산업 현장에서 발생된 대용량 데이터를 사용하여 빅데이터 분석을 진행하면 제품 생산 공정의 효율성과 품질 및 신뢰성 향상을 기대할 수 있다. 이처럼 스마트팩토리의 이점을 취한 기업체들은 경쟁력 확보 및 시장의 우위를 차지할 수 있을 것으로 전망된다. 다만 자동차/전자/부품 같은 조립가공 산업과 금속/화학/에너지 같은 프로세스장치 산업에서 스마트팩토리의 전개 양상이 상당히 다르므로 각 기업이 속한 산업에 효과적인 도입 전략이 필요하다. 이처럼 각 산업 별로 적합한 모델을 구축하기 위해서는 반드시 데이터의 특성을 고려해야한다. 특히 정상데이터와 비정상데이터의 비율이 불균형하게 분포하는 기계 데이터를 사용하는 모델에서는 불균형 데이터 문제를 해결하지 않을 경우 모델의 성능 저하를 야기할 수 있다.

따라서 본 연구에서는 Kaggle에서 제공하는 기계데이터를 텐서플로우에 적용하여 모델을 구축하였다. 모델을 구축하며 불균형 데이터 문제를 해결하는 방법으로 클래스 가중치를 적용하는 모델과 SMOTE를 적용한 모델을 제시하고, 이 두 모델의 성능을 비교한다. 본 논문의 구성은 다음과 같다. 2절에서는 연구 방법에 대해 제시하고 3절에서는 연구의 결과를 제시하며 4절에서는 결론을 도출한다. 기존연구는 생산 공정에서 수집된 데이터를 통해 불량에 영향을 미치는 핵심적인 변수를 XGBoost와 SHAP를 사용하여 도출했다[2]. 다른 연구에서는 센서를 통해 데이터를 수집하고, 머신러닝과 딥러닝 기법을 적용하여 제품의 불량 사전 예측 방법을 제안했다[3].

Ⅱ. 연구 방법

본 연구에서 사용한 데이터는 예측모델 및 분석 대회 플랫폼인 Kaggle (https://www.kaggle.com/ivanloginov/the-broken-machine)에서 제공하는 생산 기계데이터로 59 개의 컬럼을 가지고 있으며, 컬럼명은 비공개로 제공하고 있다. 또한 ytrain.csv 파일로 기계 고장에 대한 상태를 제공한다. 본 연구에서의 연구방법은 다음 그림 1과 같다.

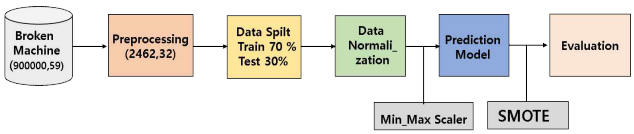

그림 1. 기계 데이터의 고장 여부 예측 프로세스

종속변수는 기계 고장 여부이며 나머지 변수들을 사용하여 예측 모델을 구축하였다. 전체 데이터는 900,000만 건으로 보다 정확한 예측 모델 구축을 위해 NULL값이 포함된 행을 삭제하여 2,462건의 데이터가 선택되었다. 또한, 데이터에 대한 설명이 존재하지 않아 알 수 없는 범주형 컬럼들을 제거하여 32건의 컬럼이 선택되었다. 데이터는 분석도구인 Python3.8을 사용하여 분석되었다. 기계 고장여부 예측를 위해 데이터를 70:30 비율

로 Train, Test 로 분리하였다. Train은 1747건, Test는 715건이다. 모델 구축은 텐서플로우(Tensorflow)를 사용했다. 모델 1은 텐서플로우만을 사용하여 구축하였으며, 불균형 데이터 해결을 위해 Weight balancing을 적용하는 모델 2와 SMOTE를 적용한 모델 3을 구축하였다. 텐서플로우는 구글이 2015년 공개한 머신러닝 오픈소스 라이브러리로 모델 저장 및 공유가 쉽다는 장점이 있다. Weight balancing은 training set의 각 데이터에서 loss를 계산할 때 특정 클래스의 데이터에 더 큰 loss 값을 갖도록 하는 방법이다. SMOTE(Synthetic Minority Over-sampling Technique)란 오버샘플링 기법 중 합성데이터를 생성하는 방식으로 많이 사용되는 기법 중 하나이다. SMOTE를 사용하여 1:1로 클래스를 조정한 데이터는 각 2,466건으로 오버샘플링 되었다. 예측 모델은 평가 지표를 바탕으로 성능을 평가한다.

Ⅲ. 연구 결과

텐서플로우만을 사용하여 기계 고장 여부 예측 모델(모델 1)을 구축하였다. 그림 2는 텐서플로우만을 사용하여 구축한 모델의 정밀도와 재현율을 나타낸다. 모델1의 경우 Train의 정밀도가 validation의 정밀도에 비해 과도하게 높게 측정되었으며, 재현율의 경우 매우 낮게 측정된 것을 확인할 수 있었다.

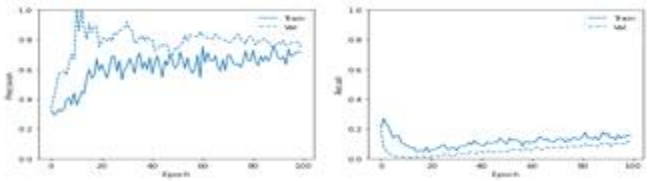

그림 2. 모델 1의 에포크 수에 따른 정밀도와 재현율

클래스간 불균형 문제를 해결하고, 모델의 성능을 높이기 위해 클래스 가중치를 적용한 모델(모델 2)을 구축하였다. 클래스 가중치를 적용한 모델의 정밀도와 재현율은 그림 3과 같다

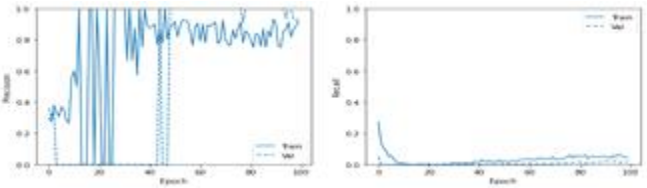

그림 3. 모델 2의 에포크 수에 따른 정밀도와 재현율

클래스간 불균형 문제를 해결하고, 모델의 성능을 높이기 위해 SMOTE를 적용한 모델(모델 2)을 구축하였다. SMOTE를 적용한 모델의 정밀도와 재현율은 그림 4와 같다.

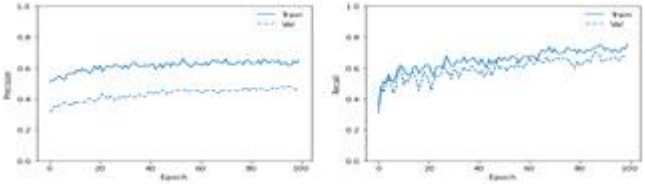

그림 4. 모델 3의 에포크 수에 따른 정밀도와 재현율

표1은 각 모델 별 평가 결과를 확인한 것이다. Model 1의 재현율은 0.0306, 정밀도는 0.3043, F1 Score는 0.0556이다. Model 2의 재현율은 0.0044, 정밀도는 0.5, F1 Score는 0.0556이다. Model 3의 재현율은 0.3144, 정밀도는 0.2963, F1 Score는 0.3051이다. 평가 결과 SMOTE를 사용하여 클래스간 불균형을 해결한 모델의 성능이 가장 높게 나온 것을 확인했다.

표 1. 각 모델별 평가지표 비교

	Model 1	Model 2	Model 3
Recall	0.0306	0.0044	0.3144
Precision	0.3043	0.5000	0.2963
F1 Score	0.0556	0.0087	0.3051

Ⅳ. 결론

우리는 Kaggle에서 제공하는 생산 기계 데이터를 사용하여 고장 여부 예측을 위한 모델을 구축하였다. 기계 데이터의 특성으로 나타나는 불균형 데이터 해결을 위해 Weight balancing와 SMOTE를 사용했다. 모델 평가 결과 SMOTE를 사용하여 클래스 불균형을 해결한 모델이 가장 높은 성능을 도출하는 것을 확인했다. 연구의 한계로 데이터의 각 컬럼들이 비공개로 제공되어 컬럼 별 상관관계를 확인하는 것에 어려움이 있었으며, 전처리 과정에서 NULL값 제거로 인한 데이터의 손실이 컸다. 향후 연구에서는 제조 산업에 특화된 예측모델을 구축하고 데이터 수집과 분석, 모니터링을 제공하는 스마트팩토리 플랫폼 구축을 진행하는 연구를 진행할 것이다.

ACKNOWLEDGMENT

이 논문은 2020년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. 2020R1A6A1A12047945, No. 2020R1I1A1A01065199)

참 고 문 헌

[1] 사공운(Woon Sagong), 이승철(SeungCheol Lee), 장용훈(Yonghun Jang), and 박창현(Changhyeon Park). "스마트 팩토리를 위한 하둡 에코 시스템 및 머신러닝 기반의 고무 공정 데이터 분석." 정보과학회 컴퓨팅의 실제 논문지 26.12 (2020): 519-527.

[2] 최은선(Eun-Seon Choi), 전유정(Yu-Jung Janu), 김진실(Jin-Sil Kim), 김재성(Jae-Sung Kim), and 조완섭(Wan-Sup Cho). "반도체 제조공정 변수 중요도 도출에 대한 연구." 한국통신학회 학술대회논문집 2020.8 (2020): 450-451.

[3] 한무명초, 이충권, and 김양석. "제조 공정에서 센서와 머신러닝을 활용한 불량예측 방안에 대한 연구." Entrue Journal of Information Technology 17.1 (2019): 89-98.

[4] Mart´ın Abadi, Ashish Agarwal, Paul Barham. "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems." Entrue Journal of Information Technology 17.1 (2019): 89-98.