KIPS Trans. Softw. and Data Eng.
Vol.11, No.11 pp.455~464
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)

https://doi.org/10.3745/KTSDE.2022.11.11.455

Boosting the Performance of the Predictive Model on the Imbalanced Dataset Using SVM Based Bagging and Out-of-Distribution Detection

Jong Hoon Kim† · Hayoung Oh††

ABSTRACT

There are two unique characteristics of the datasets from a manufacturing process. They are the severe class imbalance and lots of Out-of-Distribution samples. Some good strategies such as the oversampling over the minority class, and the down-sampling over the majority class, are well known to handle the class imbalance. In addition, SMOTE has been chosen to address the issue recently. But, Out-of-Distribution samples have been studied just with neural networks. It seems to be hardly shown that Out-of-Distribution detection is applied to the predictive model using conventional machine learning algorithms such as SVM, Random Forest and KNN. It is known that conventional machine learning algorithms are much better than neural networks in prediction performance, because neural networks are vulnerable to over-fitting and requires much bigger dataset than conventional machine learning algorithms does. So, we suggests a new approach to utilize Out-of-Distribution detection based on SVM algorithm. In addition to that, bagging technique will be adopted to improve the precision of the model.

Keywords : Imbalanced Dataset, Predictive Performance, Bagging, Out-of-Distribution(OoD) Detection

SVM 기반 Bagging과 OoD 탐색을 활용한 제조공정의 불균형 Dataset에 대한 예측모델의 성능향상

김 종 훈† · 오 하 영††

요 약

제조업의 공정에서 생성되는 데이터셋은 크게 두 가지 특징을 가진다. 타겟 클래스의 심각한 불균형과 지속적인 Out-of-Distribution(OoD) 샘플의 발생이다. 클래스 불균형은 SMOTE 및 다양한 샘플링 전략을 통해서 대응할 수 있다. 그러나, OoD 탐색은 현재까지 인공신경망 영역에서만 다뤄져 왔다. OoD 탐색의 적용이 가능한 인공신경망은 제조공정 데이터셋에 대해서 만족스러운 성능을 발현하지 못한다. 원인은 제조공정의 데이터셋이 인공신경망에서 일반적으로 다루는 이미지, 텍스트 데이터셋과 비교해서 크기가 매우 작고, 노이즈가 심하다는 것이다. 또한 인공신경망의 과적합(overfitting) 문제도 제조업 데이터셋에서 인공신경망의 성능을 저하하는 원인으로 지적된다. 이에 현재까지 시도된 바 없는 SVM 알고리즘과 OoD 탐색의 접목을 시도하였다. 또한 예측모델의 정밀도 향상을 위해 배깅(Bagging) 알고리즘을 모델링에 반영하였다.

키워드 : 불균형 데이터, 예측성능, 배깅(Bagging), Out-of-Distribution(OoD) 탐색

1. 서 론

제조업에서는 수치형 타입 위주의 데이터를 기반으로 하고 있으며, 데이터셋의 크기 면에서도 인공신경망에서 많이 다루는 이미지나 텍스트에 비교가 안 될 정도로 매우 작다. 또한 그 원인을 명확하게 단정하기 어려운 노이즈를 다양하게 포함하고 있다. 그리고, 양품과 불량으로 구분되는 타겟 클래스는 일반적으로 극단적인 불균형을 나타낸다. 이러한 특성을 가진 데이터셋에 대해서 인공신경망은 다른 머신러닝 알고리즘과 비교해서 상대적으로 낮은 분류성능을 보였다[1]. 앞서 설명한 이유로 제조업체의 기술, 개발 부문에서는 주로 SVM, Random Forest, KNN 등의 머신러닝 알고리즘을 인공신경망보다 예측모델의 개발에 선호하는 경향을 보여왔다.

제조업에서는 시장경쟁력 유지를 위해서 다품종 소량생산과 함께 지속적인 신규 품종의 개발이 진행되고 있다. 예측모델의 입장에서 클래스 분류가 불가능한 OoD 샘플이 지속적으로 발생하고 있다[2]. 이러한 제조업 데이터셋의 특성을 고려하면, 제조업에서 예측모델의 현실적인 활용은 OoD 탐색 성능에 좌우된다[3]. 과거에 생성된 데이터셋에 대한 예측성능이 우수하다고 현실에 곧바로 적용할 수 없다는 이야기이

† 비 회 원 : LX하우시스 책임
†† 종신회원 : 성균관대학교 인공지능융합학과 교수
Manuscript Received : November 23, 2021
First Revision : February 3, 2022
First Revision : April 5, 2022
Accepted : April 7, 2022
* Corresponding Author : Hayoung Oh(hyoh79@skku.edu)

다. 그러나, 현재까지 이루어진 다양한 OoD 탐색 관련 연구는 인공신경망의 영역에 국한되는 한계를 가진다.

본 연구에서는 인공신경망을 기반으로 활용되어 오던 OoD 탐색을 일반적인 머신러닝 영역으로 확대 적용하기 위해 SVM 알고리즘을 선택하였다. SVM 알고리즘은 OoD 탐색을 적용하는데 필요한 예측결과에 대한 예측확률을 제공하기 때문이다. 또한 예측모델의 정밀도 향상을 위해 배깅을 도입하여 SVM 알고리즘 기반 앙상블 모델을 구현하였다.

제조공정에서 채취된 데이터셋에서 일반적으로 발견되는 클래스 불균형문제는 SMOTE를 적용하여 클래스별 샘플 수의 불균형을 완화하도록 시도하였다[4]. 소수 클래스에 대한 단순 오버샘플링과 SMOTE의 성능을 비교 검증하여 SMOTE 도입의 타당성을 제시하였다.

2. 선행 연구

OoD 탐색 분야에 큰 발자취를 남긴 연구들이 다수 있다. 그러한 연구 중에서 본 논문에 주요하게 참조한 논문 3개를 살펴보고자 한다.

먼저 인공신경망 OoD 탐색 연구의 시초라고 할 수 있는 흔히 'Baseline OoD Study'로 잘 알려진 논문이다. D. Hendrycks와 K. Gimpel이 ICLR 2017에서 발표하였다. 두 저자는 인공신경망의 분류 문제에서 예측모델의 OoD 샘플에 대한 출력층의 예측확률이 학습에 사용된 훈련셋의 샘플과 비교하여 큰 차이가 발생한다는 것을 논문에서 밝혔다[5]. 그들은 이러한 성질을 이용한 OoD 탐색이 다양한 데이터셋에서 매우 효과적임을 검증하였다.

두 번째 연구는 'ODIN(Out-of-Distribution Detector for Neural Networks)'으로 잘 알려진 S. Liang, Y. Li 및 R. Srikant의 ICRL 2018년 논문이다. 이 연구에서는 기존의 'Baseline OoD Study' 논문에서 제시한 탐색 방법의 성능개선을 위해서 Temperature Scaling과 Input Preprocessing의 접목을 제안하였다[6]. 그리고, 이미지 및 자연어처리 분류 문제에 적용하여 'Baseline OoD Study'에서 제시한 탐색법보다 우수한 성능을 검증해 보였다.

세 번째 연구는 2018년 ICLR에 발표된 'Training Confidence-calibrated Classifiers for Detecting Out-of-Distribution Samples'이다. 기존의 OoD 탐색 연구에서는 대부분 인공신경망 출력층의 소프트맥스(Softmax) 함수의 예측확률에 대한 보정에 초점에 둔 것에 비해서 이 연구에서는 분류기의 훈련을 통해서 분류성능을 향상하는 방법을 제시한 것이 큰 차이라고 할 수 있다[7].

데이터셋의 클래스 불균형 개선에 관한 연구로는 2006년 PAKDD에서 발표된 'Boosting Prediction Accuracy on Imbalanced Datasets with SVM Ensembles'을 살펴보았다. 데이터셋의 클래스 불균형을 SVM 기반의 앙상블 기법과 SMOTE를 혼합하여 크게 개선하였다[4].

3. 배경지식

3.1 Support Vector Machine

Fig. 1은 현재의 데이터를 보다 고차원(N차원)으로 매핑시킨 뒤에 N-1 차원의 초평면으로 분류하는 SVM의 중요한 알고리즘을 설명한다. 특히, 비선형 분류 문제에서 고차원 공간으로 매핑을 통해 선형 결정경계를 쉽게 찾을 수 있도록 커널함수를 사용한다[8]. 또한, 이러한 과정에서 잘못 분류된 데이터에 대해서는 벌점을 부여하여 손실함수를 교정한다.

SVM은 샘플이 많은 데이터셋뿐만 아니라, 비교적 적은 데이터셋에서도 예측성능이 우수하다. 그러나, 100,000개 이상의 샘플을 가지는 용량이 큰 데이터셋의 경우에는 처리시간이 길어지고, 메모리 관리도 어려워지는 단점을 수반한다. 그리고, Random Forest, Gradient Boosting과 같은 전처리가 거의 또는 전혀 필요 없는 트리 기반 알고리즘과 달리 모든 특성에 대한 표준화가 요구되며, 하이퍼 파라미터의 최적화가 이뤄진 경우에만 우수한 성능을 기대할 수 있다.

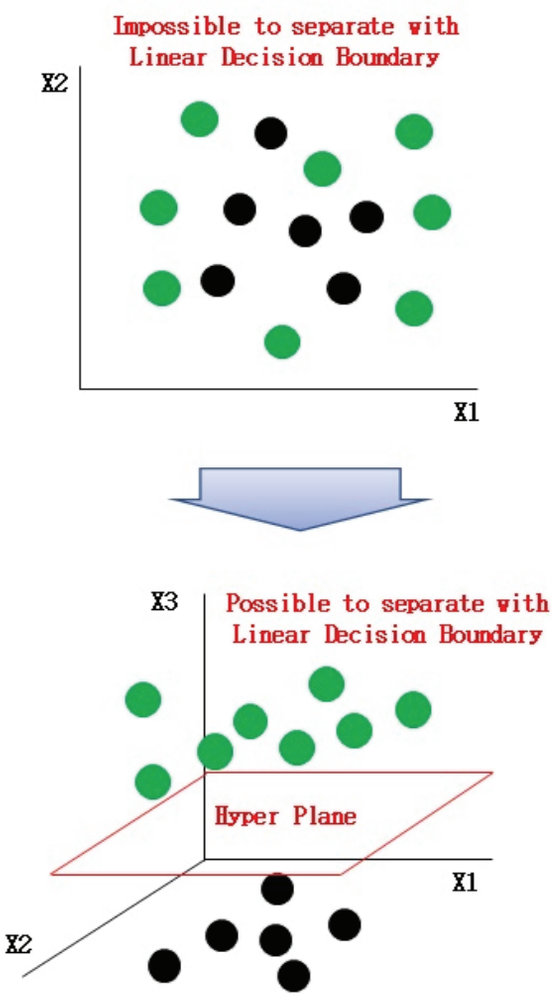

Fig. 1. The Binary Classification with Hyper Plane in SVM

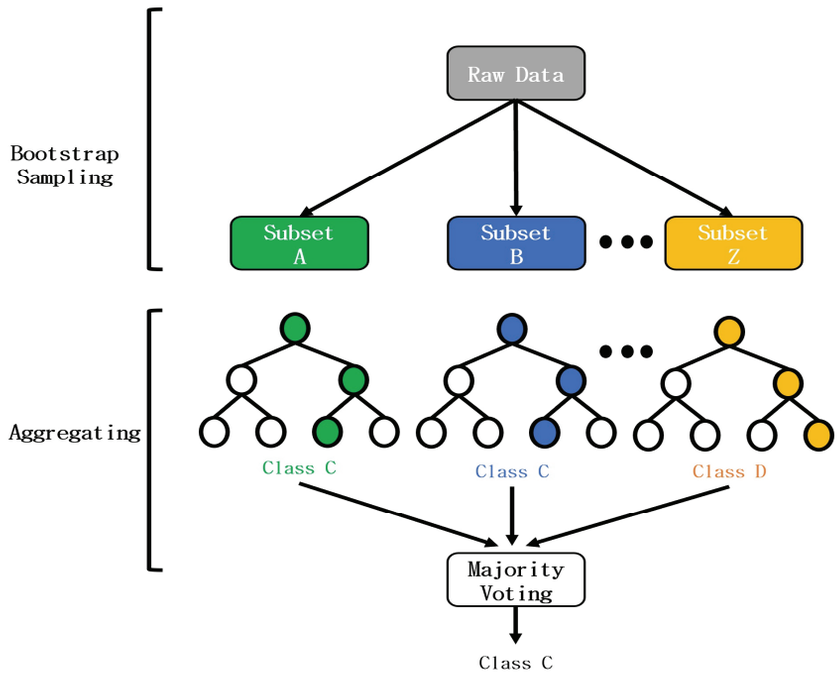

Fig. 2. Bootstrap Sampling and Aggregating in Bagging

3.2 Bootstrap Sampling and Aggregating

Fig. 2는 배깅을 간략하게 설명하고 있다. 데이터셋에서 임의 복원 샘플링을 통해 다수의 부분집합을 생성한다. 이것을 부트스트랩 샘플링(Bootstrap Sampling)이라고 한다.

샘플링한 부분집합으로 각각 예측모델을 수립하여 분류의 경우에는 다수결 투표, 회귀의 경우에는 각 모델의 예측값들에 대한 평균값을 최종값으로 출력한다. 이러한 과정을 결합(Aggregating)이라고 한다.

부트스트랩 샘플링과 결합을 함께 사용한 알고리즘이 배깅이다. 배깅은 모델의 편의는 유지되고, 특정 데이터에 과접합되는 것을 방지하며 분산을 감소시키는 장점이 있다[9].

3.3 Out-of-Distribution 탐색

분류기는 학습과정을 통해서 형성된다. 그러나, 학습과정에 포함되지 않은 샘플에 대해서는 적절한 분류성능을 확보할 수 없는 한계가 있다[10]. Fig. 3에 도식화되어 있듯이 훈련셋의 영역에 포함되지 않은 샘플을 OoD 샘플이라고 한다. 종종 이상치와 OoD 샘플을 혼동하는 경우가 발생한다. 이상치는 훈련셋에 포함된 샘플이지만, 타겟의 각 클래스의 분포에서 벗어난 샘플을 의미한다.

OoD 탐색은 OoD 샘플을 훈련셋에 포함된 학습 샘플(In-Distribution Samples)과 구분할 수 있는 방법론을 통칭하는 표현이다. 일반에 잘 알려진 Anomaly Detection의 한 가지 부류로 이해할 수도 있다.

OoD 탐색은 일반적으로 인공신경망에서 소프트맥스 함수를 사용하는 출력층에서 도출하는 예측확률을 바탕으로 작동

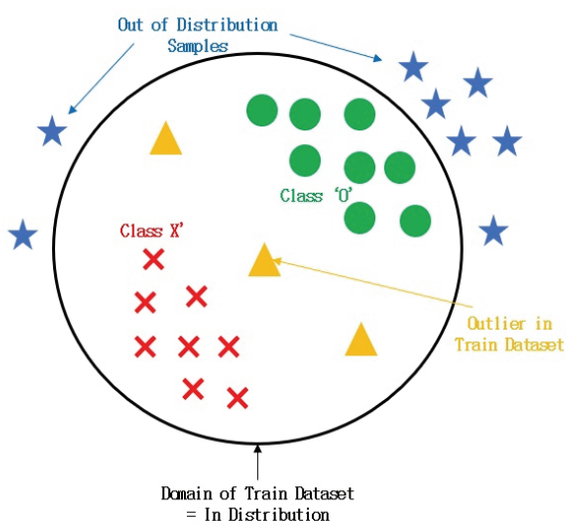

Fig. 3. In- and Out-of-Distribution Samples

한다. 훈련셋에 포함된 샘플에 대해서는 예측모델이 학습과정을 통해서 충분한 학습을 했기 때문에 비교적 높은 예측확률이 발생한다. OoD 샘플은 학습이 전혀 이뤄지지 않은 새로운 패턴의 데이터이므로, 예측모델에서 예측확률을 매우 낮은 수준으로 생성할 수밖에 없다.

3.4 Synthetic Minority Oversampling Technique

제조공정에서 수집된 데이터셋에서는 일반적으로 심각한 불균형이 관찰된다. 타겟은 합격과 불량으로 판정된다. 양산

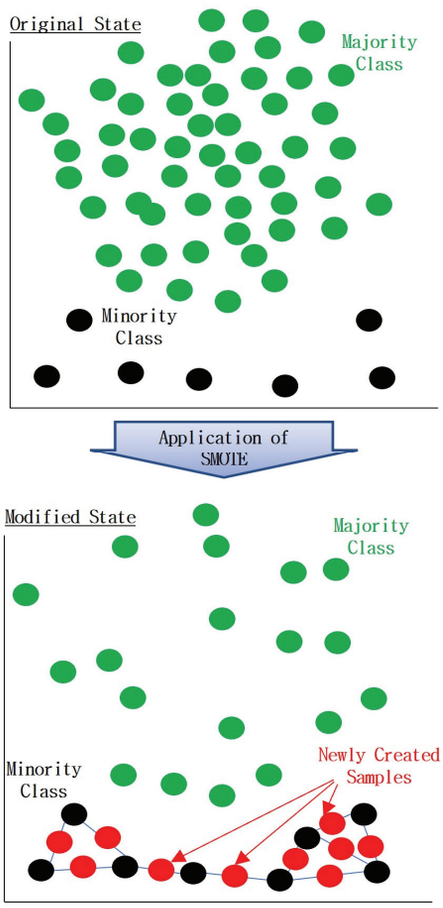

Fig. 4. The Sample Augmentation Method with SMOTE

제품들은 높은 수준의 수율(합격률)을 가지므로 불균형은 자연스러운 현상이다. 클래스 불균형 상태에서는 소수 클래스의 정보량이 적어서 예측모델이 다수 클래스에 편향된 학습을 시행하게 된다[11]. 불량을 양성(positive) 클래스로 설정했을 경우, 정확도는 높으나 재현율 및 정밀도가 저하되는 결과를 초래한다.

클래스 불균형문제를 해결하기 위한 가장 대표적인 방법이 SMOTE이다. SMOTE는 Fig. 4와 같이 소수 클래스의 임의 샘플을 선택하고, 해당 샘플의 인접한 K개 이웃과의 직선 위에 새로운 샘플을 생성한다[4]. 이와 비교하여, 일반적인 오버샘플링은 데이터셋의 샘플을 몇 번이고 복원추출하므로, 불균형 상태는 완화되지만, 예측모델의 성능을 개선할 수 있는 새로운 정보를 발생시키지 못한다.

4. 연구 방법

4.1 연구 절차

본 논문의 목적인 불균형 데이터셋에 대한 머신러닝 예측모델의 성능향상 방법을 찾기 위해 제조공정의 데이터베이스에 축적된 약 8개월간의 데이터셋을 사용하였다. 전처리에서는 결측치의 제거, 영분산 특성의 제거, 중요 특성의 선택 등이 시행되었다. 전처리 된 데이터셋을 이용하여 다양한 성능향상 방법의 효과를 검증하였다. 상세한 연구 절차는 Fig. 5에 소개되어 있다.

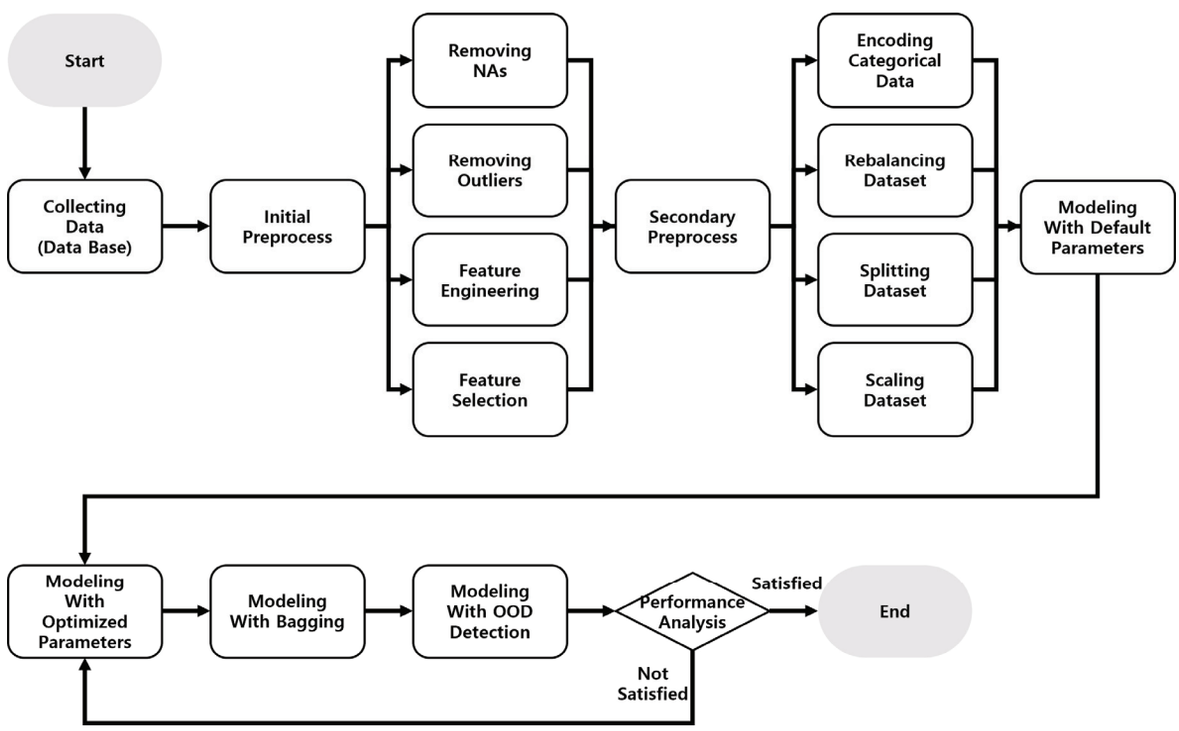

Fig. 5. Proposed Study Scheme

4.2 연구 대상

본 논문에서는 국내 건축 구조재 제조공정 데이터를 사용하였다. 공정의 생산속도, 투입되는 화학물질의 배합비 및 열처리 온도 등 총 52개의 특성과 1개의 타겟, 그리고 592개 샘플로 구성되어 있다. 데이터의 수집 기간은 2021년 2월 2일부터 2021년 10월 8일까지 약 8개월간이다. 그리고, 동일한 제품이지만, 생산설비(공정조건)가 다른 OoD 샘플 46개를 모델의 성능평가를 위해 전처리 과정에서 호출하여 테스트셋에 합병하는 방식을 취하였다.

4.3 데이터 전처리 과정

데이터의 전처리에서는 ①결측치 제거, ②영분산 특성 제거, ③탐색적 데이터 분석, ④중요 특성 선택, ⑤훈련셋과 테스트셋으로 분리, ⑥클래스 불균형에 대한 완화 작업, ⑦특성의 표준화를 실시하였다.

4.4 예측모델 구축

전처리를 완료한 훈련셋으로 SVM 알고리즘 기반의 총 3개의 예측모델을 만들었다. 첫 번째, 기본설정(default) 상태의 하이퍼 파라미터를 이용하여 예측모델을 만든다. 두 번째, 그리드 탐색 방법을 통해서 4개의 하이퍼 파라미터를 최적화하여 예측모델을 만든다. 세 번째, 최적화된 하이퍼 파라미터를 기반으로 배깅과 OoD 탐색기법을 함께 적용한 예측모델을 구축한다. 앞서 기술한 3가지 방법으로 수립된 예측모델의 성능을 비교 평가하였다.

5. 예측모델(성능향상 방법) 비교

5.1 평가 지표

본 연구에서는 3가지 예측모델을 수립하여 그 성능을 비교하였다. 성능 비교를 위한 평가 지표는 정확도, 정밀도, 재현율 및 특이도의 4가지를 사용하였다. 연구대상인 데이터셋은 심각한 클래스 불균형문제를 안고 있다. 이럴 경우, 분류 모델의 성능평가에 일반적으로 사용되는 정확도만으로는 올바른 비교가 어려우므로 다양한 평가 지표를 채용하였다.

Table 1의 분류 모델의 평가를 위한 정오표(Confusion Matrix)를 참조, 아래와 같이 정확도, 정밀도, 재현율, 그리고 특이도를 정의할 수 있다[13].

- 정확도(Accuracy) = (TP+TN)/(TP+FP+FN+TN) (1)
- 정밀도(Precision) = TP/(TP+FP) (2)
- 재현율(Recall) = TP/(TP+FN) (3)
- 특이도(Specificity) = TN/(TN+FP) (4)

Table 1. The Matrix for Classification Performance Evaluation

		Actual Value	
		Positive	Negative
Predicted Value	Positive	**T**rue **P**ositive (TP)	**F**alse **P**ositive (FP)
	Negative	**F**alse **N**egative (FN)	**T**rue **N**egative (TN)

Table 2. Predictive Performance by Default Settings

Default		Actual	
		Good	Bad
Predicted	Good	150	0
	Bad	17	16

5.2 하이퍼 파라미터의 기본설정값으로 모델 구축

예측성능을 비교하기 위한 대조군으로 머신러닝의 하이퍼 파라미터 기본설정값으로 예측모델을 수립하였다. 머신러닝 알고리즘은 SVM을 사용하였으며, 하이퍼 파라미터들은 별도 설정 없이 알고리즘에서 제공하는 기본설정값을 사용하였다. 하이퍼 파라미터들의 기본설정값을 구체적으로 살펴보면, kernel = 'radial', cost = 1.0, gamma = 0.0417, epsilon = 0.1이다.

Table 2는 테스트셋을 이용하여 기본설정값 예측모델의 성능을 평가한 결과이다.

5.3 그리드 탐색를 통한 하이퍼 파라미터 최적화

SVM 예측모델 성능에 크게 영향을 미치는 하이퍼 파라미터는 Table 3에 나열된 4개이다. Table 3에 제시된 4개 하이퍼 파라미터값의 범위에서 8,000가지 조합을 만들었다. 8,000가지 조합을 이용하여 각각 예측모델을 생성하고, 성능평가를 통해 가장 우수한 조합을 선정하였다. kernel = 'polynomial', cost = 8.0, gamma = 0.05, epsilon = 0.01일 때, 최고의 성능을 나타냈다.

Table 3. Hyper-Parameter Value Range for Grid Search

Hyper Parameter	Value Range for Grid Search
kernel	radial, polynomial
cost	1.0 ~ 10.0 (step = 1.00)
gamma	0.01 ~ 0.2 (step = 0.01)
epsilon	0.01 ~ 0.2 (step = 0.01)

Table 4. Performance by Hyper-Parameter Optimization

Hyper-Parameter Optimization		Actual	
		Good	Bad
Predicted	Good	156	0
	Bad	11	16

Table 4는 그리드 탐색를 통해서 최적화된 예측모델을 테스트셋으로 성능을 평가하여 정리한 정오표이다. 하이퍼 파라미터 최적화를 통해서 구현한 예측성능은 정확도 = 0.940, 정밀도 = 0.593, 재현율 = 1.000, 그리고 특이도 = 0.934이다. 하이퍼 파라미터의 최적화는 기본설정값을 사용했을 경우와 비교하여 정확도가 3.3%P, 특이도는 3.6%P 개선되었다. 정밀도는 무려 10.8%P 개선된 것으로 확인되었다.

머신러닝 예측모델의 성능개선을 위해서 일반적으로 하이퍼 파라미터의 최적화를 시도한다. 본 연구에서 검증된 하이퍼 파라미터의 최적화 효과는 그러한 활동이 매우 타당함을 보여주는 좋은 근거라고 할 수 있다.

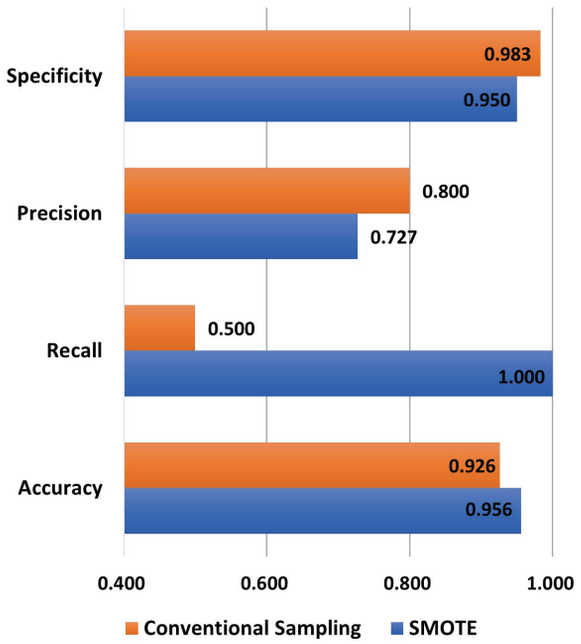

Fig. 6. Performance by both SMOTE and Conventional Sampling Method

5.4 SMOTE의 효과

클래스 불균형 완화를 위한 전통적인 접근법은 소수 클래스에 대한 오버샘플링 혹은 다수 클래스에 대한 다운샘플링이다[11]. 특정 클래스에 대한 샘플링을 통해서 클래스의 균형을 산술적으로 맞출 수는 있으나, 예측성능 개선에 활용되는 새로운 정보를 생성할 수는 없다. 불균형 상태와 비교해서 성능은 향상되지만, 극적인 개선은 기대할 수 없다[11].

이에 반해서 SMOTE는 아래 Fig. 6에서 확인할 수 있는 것과 같이 하이퍼 파라미터가 최적화된 예측모델에서 소수 클래스에 대한 단순 오버샘플링 기법과 비교해서 전반적으로 우수한 성능을 보여준다.

SMOTE는 클래스 불균형문제를 완화하는 좋은 방법이지만, 주의해야 할 사항이 있다. 그것은 SMOTE로 새롭게 생성된 소수 클래스의 샘플 중에서 다수 클래스 영역을 침범하는 샘플이 발생한다는 것이다. 이러한 샘플들은 분류 모델의 결정경계를 왜곡시키는 문제를 초래한다. 이것은 Fig. 7A 원래 데이터셋과 Fig. 7B SMOTE를 적용한 후의 데이터셋에서 'Good'(다수)과 'Bad'(소수) 클래스의 샘플 분포를 통해서 명확하게 확인할 수 있다. 이러한 현상은 앞서 소개한 SMOTE 알고리즘을 통해서 이해할 수 있다. 소수 클래스의 임의 샘플에 인접한 이웃의 선정과정에서 다수 클래스의 샘플이 선정되어 다수 클래스 영역에 새로운 샘플이 생성되면서 결정경계의 왜곡이 발생하게 된다.

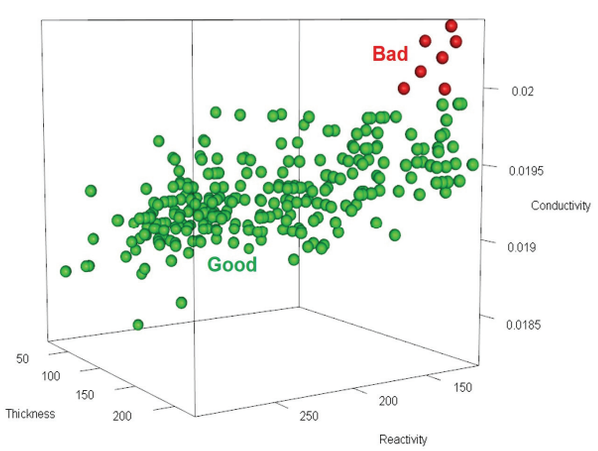

Fig. 7A. Distribution of Samples from the Raw Dataset

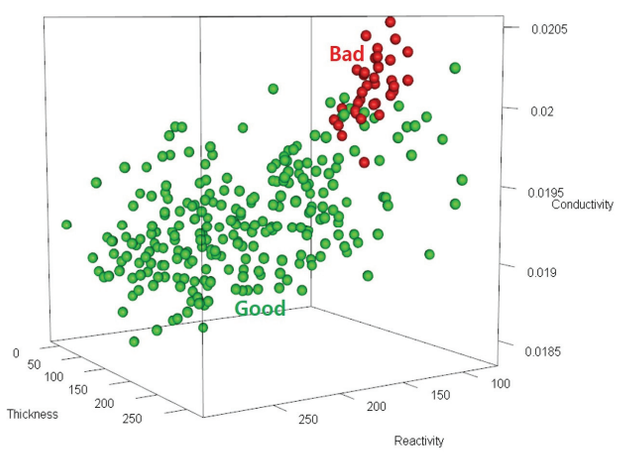

Fig. 7B. Distribution of Samples Modified with SMOTE

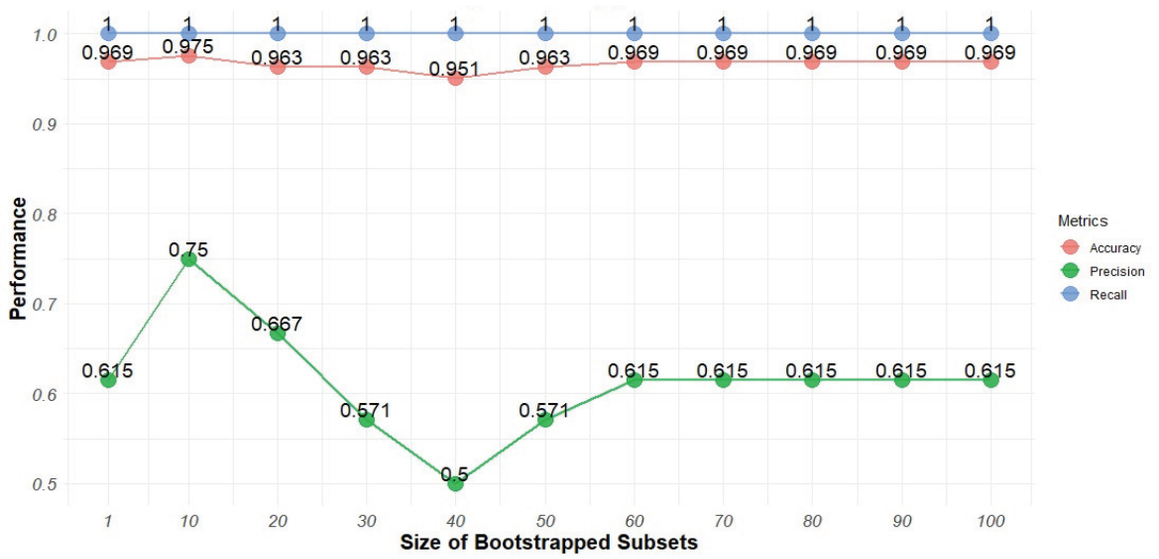

Fig. 8. Performance by Bootstrapped Subset Size

5.5 배깅의 적용

배깅을 적용하면서 반드시 고려해야 할 것이 있다. 그것은 부트스트랩 샘플링을 통해서 생성하려는 부분집합의 개수이다[12]. 앞에서 기술한 바와 같이 SVM은 연산속도에서 다른 머신러닝 알고리즘과 비교해서 열위하다. 그래서 우수한 성능을 발현하면서도 연산량이 작은 데이터셋을 선택하는 작업이 요구된다. 이를 위해서 부트스트랩 샘플링을 통해서 생성된 부분집합의 개수에 따른 배깅 예측모델의 성능을 살펴보았다. Fig. 8은 부분집합 개수에 따른 배깅 예측모델의 성능을 보여준다. 일반적으로 부분집합 개수가 많을수록 좋은 성능을 보여줄 것이라 기대하지만, 비교적 적은 개수인 10개에서 가장 우수한 성능이 발현되었다.

훈련셋에 대해서 부트스트랩 샘플링을 적용하여 10개의 부분집합을 형성하였다. 그리고, 각 부분집합을 이용하여 10개의 SVM 예측모델을 수립하였다. 하이퍼 파라미터는 앞서 기술한 최적화의 값을 채용하였다. 하나의 입력에 대해서 예측모델에서 발생하는 예측값이 10개이므로, 이에 대해서 다수결 투표를 통해서 최종 예측값을 결정한다.

부트스트랩 샘플링에서는 훈련셋의 100%에 해당하는 샘플을 복원 추출하였다. 목적은 훈련셋과 같은 크기를 유지하면서도 복원추출을 통해 다양성을 극대화하려는 것이다. 그러한 다양한 부분집합으로 생성된 모델들의 결합은 단일 예측모델에 비해서 좋은 성능을 발현한다[14]. 이것은 Equation (5)과 (6)에서 확인할 수 있는 바와 같이 부분집합의 개수가 증가하면, 각 부분집합에서 산출된 통계량의 평균으로 모수를 추정할 때 표준오차가 감소한다는 중심극한정리로 잘 설명된다[15]. 표준오차의 감소는 신뢰구간의 감소를 가져오며, 이는 추정의 정밀도가 개선됨을 의미한다. 머신러닝의 측면에서 설명하자면, 배깅은 모델의 정확도와 관련된 오차인 편의(Bias)는 증가시키지 않으면서, 정밀도와 관련된 오차인 분산만 감소시키는 효과를 발생시킨다[16]. 일반적으로 편의와 분산은 트레이드오프의 관계가 있는 것으로 알려져 있다[17].

- Value of Population : $Y \sim N(\mu, \sigma^2)$ (5)
- Mean of Subsets : $\overline{Y} \sim N(\mu, \frac{\sigma^2}{\sqrt{n}})$ (6)

5.6 Out-of-Distribution 탐색의 적용

머신러닝 예측모델의 구축과정에서 훈련셋이 포함할 수 있는 샘플의 양은 매우 제한적이다. 즉, 일상에서는 훈련셋의 분포에서 벗어난 샘플이 일반적일 것이다[10]. 그러므로, 머신러닝 모델이 가지는 이러한 한계를 보완할 수 있는 도구로써 OoD 탐색은 반드시 검토되어야 한다. 본 연구에서는 'Baseline OoD Study'에서 제안했던 출력층의 소프트맥스 함수의 예측확률을 이용하여 OoD 샘플을 구분하는 알고리즘을 도입하였다[18].

SVM 배깅 예측모델에 OoD 탐색 기능을 부가하여 Fig. 9에서 확인할 수 있는 의사코드(Pseudo Code)로 구현되는 새로운 모델을 만들었다. 일반적인 SVM 예측모델과 다른 점은 각 입력에 대한 출력의 예측확률이 임계값(Threshold)보다 작을 경우, 이를 OoD 샘플로 분류하는 것이다. OoD 샘플을 구분하는 기준인 임계값은 'Bad' 클래스 예측확률의 평균을 적용하였다. 이에 대한 설정 근거는 Equation (7), (8)에서 살펴볼 수 있듯이, SMOTE를 통해 데이터셋의 클래스 불균형을 완화하였음에도 'Bad' 클래스의 비율이 약 29%에 수준에 불과하므로 예측확률은 다수 클래스보다 낮다. 즉, 다수 클래스에 비해 불충한 정보를 제공하는 소수 클래스의 예

```
Algorithm : Bagging + OOD Detection Algorithms
Input:trainSet
Input:testSet
Input:HP:Hyper Parameters from Optimized Model
Input:TH:Threshold For In-and Out-of-Distribution
Input:Results:Data Frame for Predicted Results

for i = 1 in 1:50 do
    Subset i ← Bootstrap Sampling(trainSet)
    Model i ← SVM(HP, Subset i)
    Pred.Probability i ← Model i(testSet)
    Pred.Label i ← Model i(testSet)
    Results.append(Pred.Label i,Pred.Probability i)
Return (Results)

BAD ← Results[Pred.Label =='BAD']
TH ← BAD[Pred.Probability]

for i = 1 in 1:nrow(Results_Mean) do
    if  Results[Pred.Probability]i > TH
        Final_Results i ← Results i
    else
        Final_Results i ← Out_of_Distribution
Return (Final_Results)

Output: Final_Results
```

Fig. 9. Pseudo Code for the Hybrid Algorithm of Bagging and OoD Detection based on SVM

Table 5. Performance by Bagging With OoD Detection

Bagging with OoD Detection		Actual	
		Good	Bad
Predicted	Good	132	0
	Bad	2	8

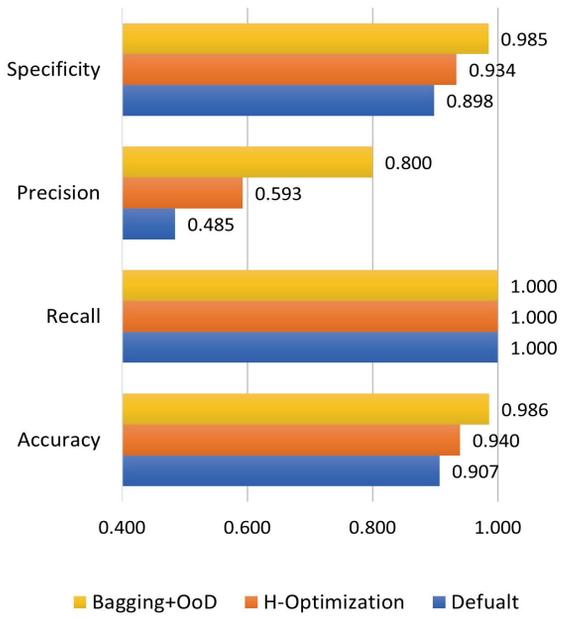

Fig. 10. The Performance of Three Classification Models

측확률은 낮을 수밖에 없으며, OoD 샘플은 이보다 훨씬 낮은 예측확률이 발생하리라 합리적으로 추정할 수 있다. 본 연구 과정에서 생성된 임계값 TH = 0.776이다.

일반적인 SVM 예측모델은 입력에 대한 각 클래스의 예측확률을 비교하여 가장 큰 확률을 가진 클래스를 결과로 선택한다. 문제는 선택된 클래스의 확률과 선택되지 않은 클래스의 확률의 차이가 매우 작을 수 있다는 것이다[19]. 이것을 예측 불확실성이라고 한다. 가장 높은 예측확률의 클래스를 선택하는 것은 타당하지만, 클래스 간 예측확률의 차이가 작다면, 결과의 신뢰성을 의심해야 한다. 가시화된 예측결과뿐만 아니라, 그 결과에 대한 예측확률도 살펴볼 필요가 있다는 것이다.

- Pp(Bad Sample) ≤ Pp(Good Sample) (7)
- Pp(OoD Sample) < Pp(Bad Sample) (8)
- Threshold = Mean[Pp(Bad Samples)] (9)

※Pp : Prediction Probability

SVM 배깅 예측모델에 OoD 탐색 기능이 부가된 알고리즘의 성능을 검증한 결과는 Table 5의 정오표로 확인할 수 있다. 예측모델의 성능을 객관적으로 평가하기 위해서 OoD 샘플을 별도 준비하여 기존의 테스트셋에 병합하였다. 그러한 테스트셋으로 예측모델의 성능을 평가한 결과, 정확도 = 0.986, 정밀도 = 0.800, 재현율 = 1.000, 그리고 특이도 = 0.985가 발현되었다. Fig. 10은 본 연구에서 제시한 예측모델들의 성능을 비교한 것이다. 3가지 방법으로 구축한 예측모델 중에서 OoD 탐색 기능을 추가한 모델은 정확도, 정밀도, 재현율, 특이도 모두에서 가장 우수했다. 특히 정밀도에서 다른 예측모델에 비해서 월등한 성능을 보여주었다.

6. 결 론

머신러닝 예측모델의 성능향상을 위한 가장 일반적인 방법은 하이퍼 파라미터의 최적화이다. 본 연구에서도 하이퍼 파라미터 최적화는 기본설정값으로 생성된 예측모델에 비해서 개선된 성능을 보여주었다.

하이퍼 파라미터 최적화 모델의 성능은 우수하지만, 앙상블 모델과 비교해서 강건성은 다소 열위에 있다. 이에 본 연구에서는 SVM 알고리즘을 기반으로 배깅 기법을 활용하여 앙상블 모델을 구성하고, 앙상블 모델의 최종결괏값의 예측확률에 대해서 OoD 탐색을 적용하였다. 배깅을 통해 10개 부분집합에서 생성된 각 예측모델의 다양성을 흡수하였고, OoD 탐색을 통해서 이상치와 OoD 샘플을 검출하는 능력을 부여하였다.

Fig. 10에서 확인할 수 있듯이 배깅 + OoD 탐색 모델은 기대한 성능을 보여주었다. 다만, 정확도, 재현율 및 특이도와 비교하여 정밀도가 다소 저조하였다. 이는 OoD 탐색을 통해서 2개 클래스에 대한 잘못 예측된 샘플 수는 많이 감소하였지만, 이와 함께 올바르게 예측된 샘플 중에서 예측확률이 낮은 것들도 OoD로 분류되어 제거되었기 때문에 발생한 것이다. 발생 원인은 SMOTE와 관련된 것으로, 소수 클래스의 샘플을 합성하는 과정에서 K-최근접 이웃 알고리즘에 의해 소수 클래스 샘플 인근에 있는 다수 클래스 샘플을 이용하여 신규 샘플을 생성하였기 때문이다. 이것은 SMOTE로 새롭게 합성된 샘플을 재분류하거나, 혹은 SMOTE의 하이퍼파라미터인 소수 클래스의 비율(p) 혹은 소수 클래스 샘플 주위의 최근접 이웃의 수(K)를 적절하게 설정하게 되면 완화할 수 있을 것으로 판단된다. 본 연구에서는 비율(p) = 0.20으로 적용하여 오류율을 감소시켰다.

본 연구를 통해서 새롭게 제시하는 알고리즘은 제조공정의 데이터셋을 활용하여 완제품의 품질을 실시간으로 예측하는 시스템에 적용되어 장기적으로 예측성능이 검증될 것이다.

본 연구에서 제안한 알고리즘은 높은 성능과 함께 OoD 샘플에 대한 분류능력을 통해 생산 현장에 적용할 수 있는 신뢰성을 확보하였다. 그러나, 지속적인 신제품의 개발 및 양산투입, 생산성 향상을 위한 공정조건의 변동은 지도학습 기반의 예측모델에 지속적인 숙제를 던져준다. 큰 변화가 있을 때마다 데이터셋을 수집하고 모델을 업데이트해야 한다. 이에 본 논문의 후속 활동으로, 제조공정에서 실시간 예측모델로 사용할 수 있을 정도로 성능이 우수하면서도 구조가 단순하여 빠르게 모델을 업데이트할 수 있는 알고리즘인 오토인코더를 연구할 계획이다.

References

[1] E. A. Zanaty, "Support Vector Machines (SVMs) versus Multilayer Perceptron (MLP) in data classification," *Egyptian Informatics Journal*, Vol.13, Iss.3, pp.177-183, 2012.

[2] S. Bulusu, B. Kailkhura, P. K. Varshney, B. Li, and D. Song, "Anomalous example detection in deep learning: A survey," *IEEE Access*, Vol.8, pp.132330-132347, 2020.

[3] Y. B. Hur, E. H. Yang, and S. J. Hwang, "A simple framework for robust out-of-distribution detection," *IEEE Access*, Vol.10, pp.23086-23097, 2022.

[4] Y. Liu, A. An, and X. Huang, "Boosting prediction accuracy on imbalanced datasets with SVM ensembles," *10th Pacific-Asia Conference, Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp.107-118, 2006.

[5] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *International Conference on Learning Representations*, 2017.

[6] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of Out-of-Distribution image detection in neural networks," *International Conference on Learning Representations*, 2018.

[7] K. M. Lee, H. L. Lee, K. B. Lee, and J. W. Shin, "Training confidence-calibrated classifiers for detecting Out-of-Distribution samples," *International Conference on Learning Representations*, 2018.

[8] K. Hansson, S. Yella, M. Doughherty, and H. Fleyeh, "Machine learning algorithms in heavy process manufacturing," *American Journal of Intelligent Systems*, Vol.6, No.1, pp.1-6, 2016.

[9] F. Mohareb, O. Papadopoulou, and E. Panagou, "Ensemble-based support vector machine classifiers as an efficient tool for quality assessment of beef fillets from electronic nose data," *Analytical Methods*, Vol.8, Iss.18, pp.3711-3721, 2016.

[10] D. Hendrycks, M. Mazeika, and T. Dietterich, "Deep anomaly detection with outlier exposure," *International Conference on Learning Representations*, 2019.

[11] C. Jian, J. Gao, and Y. Ao, "A new sampling method for classifying imbalanced data based on support vector machine ensemble," *Neurocomputing*, Vol.193, Iss.C, pp.115-122, 2016.

[12] M. Farrash and W. Wang, "How data partitioning strategies and subset size influence the performance of an ensemble?," *IEEE International Conference on Big Data*, pp.42-49, 2013.

[13] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," *International Journal of Data Mining & Knowledge Management Process*, Vol.5, No.2, pp.1-11, 2015.

[14] C. M. Bishop, "Neural networks for pattern recognition," Oxford University Press, pp.365, 1996.

[15] H. Y. Lee, "Research methodology," 2nd ed. Seoul, Korea: CRbooks, pp.234-235, 2014.

[16] S. M. Nzuva, L. Nderu, and T. Mwalili, "Ensemble model for enhancing classification accuracy in intrusion detection systems," *International Conference on Electrical, Computer and Energy Technologies*, 2021.

[17] C. Ayuya, "Ensemble learning on bias and variance," Updated on January 20, 2021, Section [Internet], https://www.section.io/engineering-education/ensemble-bias-var/

[18] D. H. Yang, K. M. Ngoc, I. S. Shin, K. H. Lee, and M. G. Hwang, "Ensemble-based out-of-distribution detection," *Electronics*, Vol.10, Iss.5, 2021.

[19] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," *Neural Information Processing Systems*, 2018.

김 종 훈

https://orcid.org/0000-0001-7481-5497
e-mail : npainkiller@naver.com
2001년 ~ 2016년 LG 마이크론/전자 차장
2016년 ~ 현 재 LX하우시스 책임
2022년 성균관대학교 데이터사이언스 융합학과(석사)
관심분야 : 머신러닝, 제품 특성 실시간 예측, 객체 인식

오 하 영

https://orcid.org/0000-0002-7362-5138
e-mail : hyoh79@skku.edu
2013년 ~ 2016년 숭실대학교 IT융합대학 교수
2016년 ~ 2019년 아주대학교 소프트웨어학과 교수
2019년 ~ 현 재 성균관대학교 인공지능융합학과 교수
관심분야 : 소셜정보망 분석, 추천시스템, 데이터분석 및 인공지능