Journal of KIIT. Vol. 23, No. 3, pp. 27-38, Mar. 31, 2025. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2025.23.3.27

기업부도예측 부스팅 모델의 성능 최적화

김명종*, 김윤후**, 김재남***

Performance Optimization of Bankruptcy Prediction Boosting Model

Myoung-Jong Kim*, Yun-Hu Kim**, and Jae-Nam Kim***

요 약

대부분의 분류모델들은 손실함수 극소화를 학습 알고리즘의 목적함수로 채택하고 있지만, 이러한 분류모델은 범주 불균형 문제에서 다수 범주의 특이도에만 초점을 맞추고 소수 범주의 민감도를 무시하는 불균형 학습을 진행하기 때문에 소수 범주에 대한 분류성능이 저하된다. 본 연구의 GMOPT는 기하평균(GM) 극대화를 학습 알고리즘의 목적함수로 설정하고 GM 극대화를 위한 최적화 작업을 수행함으로써 다수 및 소수 범주에 대한 균형 학습을 진행한다. GMOPT의 성능 검증을 위하여 한국, 러시아, 폴란드 3개 국가의 부도예측 데이터를 구성하였으며 전통적인 부스팅 모델을 벤치마크 모델로 활용하여 성능을 비교한 결과 GMOPT는 범주 불균형 문제에서 다수 범주와 소수 범주에 대한 균형 학습을 진행할 수 있으며, GM과 ROC곡선 하의 면적(AUROC) 측면에서 탁월한 성능개선 효과를 보였다.

Abstract

Most classification models adopt the minimization of a loss function as the objective function in their learning algorithms. However, such models often exhibit imbalanced learning in the presence of class imbalance, focusing solely on the specificity of majority classes while neglecting the sensitivity of minority classes. This results in degraded classification performance for minority classes. The proposed GMOPT in this study addresses this issue by setting the maximization of the Geometric Mean(GM) as the objective function of the learning algorithm and performing optimization tailored for GM maximization. To validate the performance of GMOPT, bankruptcy prediction datasets from three countries－South Korea, Russia, and Poland－were utilized. Performance comparisons were conducted using traditional boosting models as benchmark models. The results demonstrated that GMOPT effectively performs balanced learning for both majority and minority classes in class imbalance scenarios and significantly improves performance in terms of GM and Area Under Receiver Ope-rating Characteristic Curve(AUROC) metrics.

Keywords
bankruptcy prediction, class imbalance, boosting, GMOPT

* 부산대학교 경영학과 교수(교신저자)
- ORCID: https://orcid.org/0000-0002-4316-8087
** 퍼듀대학교 컴퓨터사이언스 석사과정
- ORCID: https://orcid.org/0009-0005-0795-9841
*** 부산대학교 경영학과 학사
- ORCID: https://orcid.org/0009-0007-0608-008X

· Received: Dec. 13, 2024, Revised: Jan. 21, 2025, Accepted: Jan. 24, 2025
· Corresponding Author: Myoung-Jong Kim
School of Business at Pusan National University,
2, Busandaehak-ro 63beon-gil, Geumjeong-gu, Busan, Korea
Tel.: + 82-51-510-3154, Email: mjongkim@pusan.ac.kr

Ⅰ. 서 론

제 4차 산업혁명의 도래로 인한 경영 패러다임의 변화에 대응하기 위하여 글로벌 기업들은 4차 산업혁명의 핵심 정보기술을 경쟁력 강화의 도구로 활용하여 물류 최적화, 고객관계관리, 이상거래탐지 등 다양한 분야에서 적용하고 있다. 특히, 인공지능 기반의 핵심기술이 가장 적극적으로 활용되고 있는 산업은 금융분야이다. 예로, 비자카드는 신경망 네트워크를 활용하여 Visa 첨단 승인 시스템으로 연평균 250억 달러 규모의 부정 결제를 방지하고 유통업체 및 소비자를 위한 글로벌 지불결제 생태계를 제공하고 있다고 발표했다. Bank of America는 지능형 가상 비서 시스템인 에리카를 개발하여 챗봇의 역할뿐만 아니라 고객의 금융거래 패턴에 대한 맞춤형 솔루션을 도출하여 인공지능 기반의 고객관리 시스템으로 고객의 요구에 신속하게 대응하고 있다.

금융분야에서 인공지능을 의사결정 솔루션으로 활용하는 다른 분야는 부도예측, 카드연체 등의 분류 및 예측 문제이다. 그러나 금융분류모델의 성능을 개선하기 위해서는 범주 불균형 문제에 효과적으로 대처할 필요가 있다. 범주 불균형 문제는 데이터가 특정 범주에 편중된 표본분포의 왜곡문제로 정의된다. 예로, 기업의 부도는 발생 비율이 매우 낮은 사건으로 국내 외부감사 법인의 장기평균 부도율은 대략적으로 3%~5% 수준으로 부도 기업의 발생 빈도는 매우 희귀한 반면, 대다수의 기업들은 정상 기업의 범주에 속하게 된다. 범주 불균형 문제에 대하여 대다수 분류모델은 소수 범주)에 대한 정확도(민감도, Sensitivity)를 무시하고 다수 범주에 대한 분류 정확도(특이도, Specificity)에 편중된 불균형 학습을 진행하는 경향이 있다. 특히 범주 간 불균형 비율이 높을수록, 다수 범주가 소수 범주의 경계영역을 침투함에 따라 소수 범주의 경계영역이 소멸되어 모든 데이터를 다수 범주로 분류하기 때문에 소수 범주에 대한 분류 능력을 상실하게 된다[1]. 이러한 불균형 학습은 금융분류모델의 설계 목적에도 부합하지 못한다. 예로, 부도기업 예측문제에서 부도기업을 정확하게 예측하는 민감도는 정상기업을 정확하게 예측하는 특이도보다 중시되는데 특이도에 초점을 맞추어 민감도를 무시하는 불균형 학습은 분류모델이라는 의미를 실질적으로 상실하게 한다

범주 간 불균형 문제를 완화하기 위한 대표적 기법으로 데이터 샘플링, Cost-sensitive 기법, 앙상블 학습 등이 활용되고 있다. 그러나 이러한 기법들을 금융실무에 활용하기 위해서는 금융기관의 규제적 환경을 고려할 필요가 있다. 현재 금융기관의 신용리스크 관리시스템의 관리규범으로서 바젤은행감독위원회가 제안한 바젤협약은 차주의 위험 특성을 정확하게 반영하기 위하여 데이터 유효성 및 평가모델의 일관성 문제를 강조하고 있다[2]. 데이터 샘플링은 임의적인 데이터 조작을 통하여 균형 표본을 구성하는 방법으로 임의적 조작으로 차주의 위험 특성에 대한 정보를 왜곡하여 신용 데이터의 유효성 문제가 제기될 위험이 발생할 수 있다. Cost-sensitive 기법은 오분류 비용함수 설정에 주관성이 개입되기 때문에 평가모델의 객관성을 훼손할 수 있다. 이에 따라 금융실무에서는 데이터 샘플링 및 Cost-sensitive 기법의 활용도가 매우 저조한 실정이다. 반면, 앙상블 학습은 데이터의 임의적인 조작으로 인한 데이터 유효성 문제와 평가모델의 주관성 문제에서 자유롭기 때문에 현재 앙상블 기반의 분류모델이 금융 실무에서 다양하게 적용되고 있다.

앙상블을 포함한 대다수의 분류모델이 범주 불균형 문제에 대하여 성능이 저하되는 주요 원인은 목적함수 및 성능지표의 설정 문제와 관련되어 있다. 대부분의 분류모델은 목적함수로서 손실함수 극소화를 설정하고 있다. 예로, 회귀분석이나 인공신경망은 손실함수를 오차 제곱으로 정의하고 미분 또는 경사하강법을 적용하여 손실함수를 극소화함으로써 간접적으로 정확도를 극대화하는 학습 방식을 선택하고 있다. 그러나 이러한 손실함수는 균형분포를 전제로 데이터의 범주를 고려하지 않고 산술평균에 기초하여 예측오차를 측정하기 때문에 다수 범주에 편중된 불균형 학습을 진행하여 정확도를 극대화한다. 이러한 문제로 인하여 불균형 범주에서 정확도는 더 이상 적합한 성능지표가 아니라는 문제가 제기되어 왔다.

최근에는 범주 불균형의 성능지표로서 기하평균 정확도(GM, Geometric Mean Accuracy), F1-Score, AUROC(Area Under Receiver Ope-rating Characteristic Curve) 및 balanced accuracy 등과 같은 대체적 성능지표가 활용되고 있다[3].

분류모델에서 손실함수를 목적함수로 설정하는 문제에 대한 비판이 제기됨에도 불구하고 손실함수를 대체하여 성능지표를 목적함수로 설정하지 못하는 이유는 대부분의 성능지표가 이산확률함수로 정의되며, 이에 따라 누적분포함수(CDF, Cumulative Distribution Function)는 미분이 거의 불가능하거나 미분을 하더라도 미분값이 0으로 측정되기 때문에 미분이나 경사하강법 등과 같은 최적화 기법을 적용할 수 없기 때문이다. 결과적으로 범주 불균형 문제에서 다수 범주와 소수 범주에 대한 균형 학습을 진행하기 위해서는 성능지표를 미분가능한 목적함수로 정의하고 이에 대한 최적화를 통하여 분류모델의 성능을 개선할 필요가 있다.

본 연구는 금융 분야의 범주 불균형 문제를 해결하기 위한 방안으로 GM 최적화(GMOPT, GM Optimization algorithm)를 제안한다. 본 연구의 GMOPT는 전통적인 분류모델과 비교하여 다음과 같은 차이점이 있다. 첫째, GMOPT는 전통적인 분류모델의 목적함수로 채택되어 왔던 손실함수 극소화를 대체하여 GM 최적화를 새로운 목적함수로 채택하고 있다. 둘째, 이산확률함수인 GM을 중심극한정리를 도입하여 미분가능한 다변량 정규분포함수로 치환하였다. 셋째, 치환된 GM에 대하여 통계적 미분 방법인 가우시안 경사하강법(Gaussian gradient descent method)을 적용하여 성능을 최적화한다.

본 연구에서는 2015년에서 2018년까지 국내 비금융기업을 대상으로 500개 부도 기업과 10,000개 기업-년도별 정상 기업 표본을 대상으로 GMOPT를 적용하였다. 또한 러시아와 폴란드의 해외 기업 부도 자료를 이용하여 GMOPT의 일반성을 추가적으로 검증하였다. GMOPT의 성능검증을 위하여 AdaBoost, GBM 및 XGBoost 등 부스팅 계열의 모델을 벤치마크 모델로 활용하였다. 성능 평가의 일반성을 확보하기 위하여 10-Fold 교차검증을 3회 반복하는 반복 교차타당성 검증을 진행하였다. 분석결과 GMOPT는 벤치마크 모델과 비교하여 범주 불균형 데이터에서 다수 범주와 소수 범주에 대한 균형 학습을 진행할 수 있으며, 부스팅 모델의 유의적인 성능개선에 기여하고 있음을 확인하였다.

본 연구는 다음과 같이 구성되어 있다. 2장에서는 분류모델의 성능지표와 범주 불균형 문제의 해결을 위하여 제안된 기법들의 실무적 적용가능성을 검토한다. 3장에서는 본 연구에서 제안하고 있는 GMOPT에 대해서 설명한다. 4장에서는 데이터의 수집, 변수 선정, 모델 설계 과정에 대하여 설명한다. 5장에서는 벤치마크 모델과 GMOPT의 성능을 비교하고, 6장에서는 본 연구의 결론을 간략하게 요약하고 향후 연구 방향을 제시한다.

II. 분류모델의 성능지표 및 범주 불균형 문제의 해결 기법

2.1 모델의 성능지표

표 1은 이진 분류의 정오행렬표(Confusion matrix)와 다양한 성능지표를 보여준다. 가장 보편적으로 활용되는 성능지표는 정확도이다.

표 1. 이범주 모델의 성능능지표
Table 1. Performance Metrics for binary classification models

		Predicted class	
		True	False
Actual class	True	True Positive (TP)	False Negative (FN)
	False	False Positive (FP)	True Negative (TN)

Sensitivity (Recall, TPR): $TP\ /\ (TP\ +\ FN)$
Specificity (TNR): $TN\ /\ (FP\ +\ TN)$
Type 1 Error (FPR): $FP\ /\ (FP\ +\ TN)$
Type 2 Error (FNR): $FN\ /\ (TP\ +\ FN)$
Precision: $TP/(TP+FP)$
Accuracy:
$(TP\ +\ TN)\ /\ (TP\ +\ FN\ +\ FP\ +\ TN)$
GM: (Sensitivity * Specificity)$^{1/2}$
AUROC:
$(1+TPR-FPR)/2 = (TPR+TNR)/2$
F1-Score: 2×((precision×Recall)/(precision+Recall))

하지만, 범주 불균형 문제에서 대부분의 분류모델은 특이도 개선에 초점을 맞추어 정확도를 개선하지만, 민감도를 무시하기 때문에 정확도는 더 이상 적합한 성능지표가 되지 못한다. 이에 따라 범주 불균형 문제에서 GM과 AUROC와 같은 대체적 성능지표가 활용된다[3]. GM은 다수 범주의 특이도와 소수 범주의 민감도의 기하평균으로 정의된다. AUROC는 ROC곡선(Receiver Operation Characteristic Curve)을 기반으로 하는 성능지표이다. ROC곡선은 FPR을 X축으로 TPR을 Y축으로 설정하여 FPR(False Positive Rate)의 변화에 따른 TPR(True Positive Rate)의 변화를 도시한 곡선이다. AUROC는 ROC 곡선의 하단 면적으로 1에 가까울수록 분류성능이 우수한 것으로 평가된다.

2.2 범주 불균형 문제 해결 기법

현재까지 범주 불균형 문제 해결을 위하여 데이터 샘플링, Cost-sensitive 기법과 앙상블 기법 등의 다양한 기법들이 제안되어 왔다[4]. 하지만, 불행하게도 은행 등 금융기관에 대한 규제로 인하여 금융실무에서는 거의 활용되지 못하고 있다. 특히, 바젤은행감독위원회가 제안한 바젤 규제안에서는 금융기관의 신용데이터는 차주의 위험 특성을 정확하게 반영하여야 하며 신용위험 측정모델은 차주의 신용위험을 일관성있게 측정되어야 한다고 규정하고 있다. 이에 따라 임의적인 데이터 조작, 오분류 비용에 대한 주관적 추정 또는 임의적인 예측모델의 결합 등이 원인이 되어 신용 데이터 및 신용위험 측정모델의 유효성 문제가 제기되는 경우 바젤 규제안의 승인요건을 충족하지 못하게 되며, 금융기관의 국제 결제, 대출 및 예금 등의 업무가 제한되기 때문에 해당 기관의 금융신인도가 하락하고 금융경쟁력이 악화되는 결과를 초래할 수 있다. 이에 따라 금융기관들은 학계에서 제안된 다양한 범주 불균형 기법들을 금융실무에 용이하게 적용하지 못하고 있다. 이러한 현실적 상황을 고려하여 최근 바젤은행감독위원회는 AI기반의 신용위험 관리시스템에 대한 새로운 규제안을 준비 중이다.

데이터 샘플링은 불균형 데이터를 균형 데이터로 구성하는 방법이다. 언더샘플링(Undersampling)은 주어진 데이터에서 다수 범주의 비율을 낮추어 균형분포를 만들어내는 방법으로 다수 범주의 데이터를 랜덤하게 제거하는 방식인 RUS(Random Under-Sampling)이 있다. 반면, 오버샘플링(Oversam-pling)은 소수 범주의 데이터를 인위적으로 생성하여 균형표본을 재구성하는 방법이다. 최근접 이웃 알고리즘(K-Nearest neighbor)을 활용하는 SMOTE (Synthetic Minority Oversampling Technique)와 데이터의 최근접 이웃의 소수 범주 비율에 따라 데이터 개수를 다르게 생성하는 ADASYN(Adaptive Synthetic Sampling Approach)이 대표적으로 사용되고 있다. 데이터 샘플링은 사용의 편리성 때문에 범주 불균형 문제에 대한 활용도가 높지만, 데이터에 대한 임의적인 조작으로 정보 왜곡과 같은 데이터 품질 문제를 발생시킬 수 있다[4][5]. 이러한 데이터의 유효성 문제로 금융실무에서는 활용도가 미약한 편이다.

Cost-sensitive 기법은 소수 범주에 더욱 높은 가중치를 부여함으로써 소수 범주의 오분류에 더욱 민감하게 반응한다. Cost-sensitive 기법은 범주별로 서로 상이한 오분류 비용으로 비용 행렬을 구성하는 방법과 손실함수에 오분류 비용함수를 추가하여 학습 알고리즘을 변경하는 방법이 이용된다[6]. Cost-sensitive 기법은 오분류 비용(손실)함수 추정에 주관성이 개입되기 때문에 모델의 객관성이 훼손될 수 있다는 문제가 있다. 특히 금융실무에서는 Cost-sensitive 기법의 오분류 비용 추정방식과는 다른 방법으로서 담보자산별로 기대회수율을 측정하고 기대회수율을 반영하여 손실율을 측정하고 있다. 결과적으로 Cost-sensitive 기법은 금융실무에서 활용도가 매우 저조한 실정이다.

앙상블 기법은 복수의 분류모델의 학습결과를 결합하여 성능을 개선하는 방법이다. 이 중 부스팅 기법은 다수 약분류자(Weak learner)들의 학습 결과를 결합하여 최종 예측결과를 도출하는 앙상블 기법이다. 부스팅 알고리즘은 약분류자를 학습시키는 과정에 있어서 오분류된 데이터에 대하여 상대적으로 높은 가중치를 부여하며, 높은 가중치가 부여된 데이터는 후행 분류자의 학습 데이터에 포함될 확률이 높아지는데 이는 오분류된 데이터에 대하여 더 많은 학습기회를 제공하며 분류자의 다양성을 확보하게 한다.

금융실무와 관련하여 앙상블 학습은 데이터에 대한 임의적인 조작이나 오분류 비용의 주관성, 임의적인 모델결합 등의 문제를 배제할 수 있다는 장점으로 인하여 부도예측 등의 금융 문제에 광범위하게 적용되고 있다[4]. 그러나 부스팅 알고리즘은 손실함수 최소화를 목적함수로 설정하고 있기 때문에 범주 불균형 문제 해결에 한계를 가진다. 부스팅 알고리즘은 데이터의 범주를 구분하지 않고 오분류 관측치의 학습에 초점을 맞추기 때문에 민감도를 개선하기보다는 정확도 개선에 더욱 도움이 되는 특이도 개선에 편향된 불균형 학습을 진행한다[5].

부스팅 알고리즘의 불균형 학습 문제를 해결하기 위하여 부스팅 알고리즘과 데이터 샘플링 또는 Cost-sensitive 기법을 결합한 하이브리드 부스팅 기법이 제안되어 왔다. 데이터 샘플링 기반으로는 SMOTEBoost[8], MSMOTEBoost[9], RUSBoost[10] 및 DataBoost-IM[11] 등이 있다. Cost-sensitive 기반의 부스팅 알고리즘으로는 AdaCost[12], CSB1과 CSB2[13], RareBoost[14] 및 AdaC1, AdaC2, AdaC3[15] 등이 대표적으로 활용되고 있다. 하이브리드 부스팅 알고리즘은 전통적인 부스팅 알고리즘과 비교하여 탁월한 성능개선의 효과를 보여주었지만, 이들은 여전히 임의적인 데이터 조작이나 주관적인 오분류 비용 추정과 같은 문제점을 가지고 있다.

III. GMOPT

부스팅 모델의 불균형 학습 문제를 개선하기 위하여 제안된 GMOPT의 학습 알고리즘은 다음과 같다. 다수 범주(Y= -1)와 소수 범주(Y = +1)의 분류문제에 적용된 AdaBoost를 가정해보자. AdaBoost의 약분류자의 예측 결과값을 확률변수 X라 하고 x^- 및 x^+는 각각 다수 범주 확률변수 X^-와 소수 범주 확률변수 X^+의 개별 관측치라 하자. 강분류자는 약분류자의 예측결과값과 약분류자의 가중치 벡터(w)를 선형결합한 판별함수(W^TX)와 최적의 임계점 C를 기준으로 범주를 분류하게 된다.

n^-와 n^+를 각각 다수 범주 표본수와 소수 범주 표본수라 할 때 임계점 C에서 GM은 식 (1)과 같이 측정된다. 식 (1)에서 임계점 C는 특이도($SPE(X^-,w)$)와 민감도($SPE(X^-,w)$)의 차이가 극소화되는 지점으로 임계점 C에서 GM은 극대화된다. 임계점 C를 기준으로 특이도와 민감도는 개별 관측치의 $f(x^-,w)$와 $f(x^+,w)$의 누적합을 범주별 표본수로 나누어 측정된다. 그러나, $f(x^-,w)$와 $f(x^+,w)$는 식 (1)에 정의된 것처럼 0 또는 1의 값을 가지는 이산확률함수로서 누적분포함수는 계단형 함수로 표현되기 때문에 미분할 수 없거나 미분하더라도 미분값이 0이 되어 미분이나 경사하강법 등을 통한 최적화가 불가능하다.

$$
\begin{aligned}
&GM(x,w) = \left(SPE(X^-,w)\times SEN(X^+,w)\right)^{1/2} \quad (1)\\
&= \left(\frac{1}{n^-}\sum^{n^-} f(x^-,w) \times \frac{1}{n^+}\sum^{n^+} f(x^+,w)\right)^{1/2}\\
&f(x^-,w) = \begin{cases} 1 & \text{if } w^Tx^- \le C) \\ 0 & Otherwise \end{cases}\\
&f(x^+,w) = \begin{cases} 1 & \text{if } w^Tx^+ \ge C) \\ 0 & Otherwise \end{cases}
\end{aligned}
$$

이산확률함수로 정의된 GM의 기대값 $E(GM(X,\ w))$은 식 (2)와 같은 연속확률함수로 정의할 수 있다.

$$
\begin{aligned}
&E(GM(X,w)) \quad (2)\\
&= \left(E(SPE(X^-,w)) \times E(SEN(X^+,w))\right)^{1/2}\\
&= \left(\int I(f(X^-,w) \le C) dP(X^- \mid Y=-1) \right.\\
&\quad \left. \times \int I(f(X^+,w) > C) dP(X^+ \mid Y=+1) \right)^{1/2}\\
&f(X^-,w) = W^TX^-, f(X^+,w) = W^TX^+
\end{aligned}
$$

그러나, 식 (2)에서 다수 범주 확률변수 X^-와 소수 범주 확률변수 X^+의 확률분포가 알려져 있지 않은 경우 $E(GM(X,\ w))$의 계산이 불가능하게 된다. 이러한 문제를 해결하기 위하여 본 연구에서는 중심극한정리에 따라 데이터가 충분한 경우 최종 분류자의 선형판별함수(W^TX)는 식 (3)과 같이 다변량 정규분포를 따른다고 가정한다. 식 (3)에서 μ^-와 μ^+는 각각 약분류자의 다수 범주와 소수 범주 데이터에 대한 예측 결과의 평균벡터이며, Σ^-와 Σ^+는 각각 다수 범주와 소수 범주 데이터에 대한 예측 결과의 공분산 행렬이다.

$$X^- \sim N(\mu^-, \Sigma^-) \rightarrow f(X^-, w) = W^T X^- \sim N(\omega^T \mu^-, \omega^T \Sigma^- \omega)$$
$$X^+ \sim N(\mu^+, \Sigma^+) \rightarrow f(X^+, w) = W^T X^+ \sim N(\omega^T \mu^+, \omega^T \Sigma^+ \omega) \quad (3)$$

식 (2)에 자연로그를 적용하여 식 (4)로 변환한 후 식 (4)를 미분하면 식 (5)에 정의된 것처럼 $\nabla GM(X,\omega)$를 측정할 수 있게 된다.

$$GM(X,w) = \frac{1}{2}\left(LN(SPE(X^-,\omega)) + LN(SEN(X^+,\omega))\right) \quad (4)$$

$$\nabla GM(X,\omega) = \frac{1}{2}\left(\frac{\nabla SPE(X^-,w)}{SPE(X^-,w)} + \frac{\nabla SEN(X^+,w)}{SEN(X^+,w)}\right) \quad (5)$$

식 (5)에서 분모에 해당하는 $SPE(X^-,w)$와 $SEN(X^+,w)$는 학습과정에서 임계점이 설정됨에 따라 시스템에서 자동적으로 계산되는 값으로 표준정규분포의 누적분포함수(Φ)를 이용하여 각각 식 (6) 및 식 (7)과 같이 정의할 수 있다. 여기에서 $\mu_{Z^-} = \omega^T \mu^-$, $\sigma_{Z^-} = \sqrt{\omega^T \Sigma^- \omega}$, $\mu_{Z^+} = \omega^T \mu^+$, $\sigma_{Z^+} = \sqrt{\omega^T \Sigma^+ \omega}$ 이다.

$$\Phi(SPE(X^-,w)) = P(W^T X^- \le C) = \Phi\left(\frac{\mu_{Z^-}}{\sigma_{Z^-}} \le C\right) = \int_{-\infty}^{C} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{C-\omega^T\mu^-}{\sqrt{\omega^T\Sigma^-\omega}}\right)^2\right) dc \quad (6)$$

$$\Phi(SEN(X^+,w)) = P(W^T X^+ > C) = (1 - P(W^T X^+ \le C)) = 1 - \Phi = 1 - \int_{-\infty}^{C} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{C-\omega^T\mu^+}{\sqrt{\omega^T\Sigma^+\omega}}\right)^2\right) dc \quad (7)$$

식 (5)의 분자에 해당하는 $\nabla(SPE(X^-, w))$와 $\nabla(SEN(X^+, w))$는 식 (6)과 식 (7)을 미분하여 각각 식 (8)과 식 (9)로 정의된다.

식 (8)과 식 (9)에 정의된 $\nabla(SPE(X^-,w))$와 $\nabla(SEN(X^+,w))$을 식 (5)에 대입하게 되면, $\nabla GM(X,w)$은 식 (10)과 같이 정의된다.

$$\nabla SPE(X^-,w) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{C-\mu_{Z^-}}{\sigma_{Z^-}}\right)^2\right) \times \left(\frac{\sigma_{Z^-}(\mu^- - C) + \left(\frac{\mu_{Z^-}-C}{\sigma_{Z^-}}\right)^2 \Sigma^- \omega}{(\sigma_{Z^-})^2}\right) \quad (8)$$

$$\nabla SEN(X^+,w) = -\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{C-\mu_{Z^+}}{\sigma_{Z^+}}\right)^2\right) \times \left(\frac{\sigma_{Z^+}(\mu^+ - C) + \left(\frac{\mu_{Z^+}-C}{\sigma_{Z^+}}\right)^2 \Sigma^+ \omega}{(\sigma_{Z^+})^2}\right) \quad (9)$$

$$\nabla GM(X,w) = \frac{1}{2}\left(\frac{1}{SPE(X^-,w)}\right) \times \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{C-\mu_{Z^-}}{\sigma_{Z^-}}\right)^2\right) \times \left(\frac{\sigma_{Z^-}(\mu^- - C) + \left(\frac{\mu_{Z^-}-C}{\sigma_{Z^-}}\right)^2 \Sigma^- \omega}{(\sigma_{Z^-})^2}\right) - \frac{1}{2}\left(\frac{1}{SEN(X^+,w)}\right) \times \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{C-\mu_{Z^+}}{\sigma_{Z^+}}\right)^2\right) \times \left(\frac{\sigma_{Z^+}(\mu^+ - C) + \left(\frac{\mu_{Z^+}-C}{\sigma_{Z^+}}\right)^2 \Sigma^+ \omega}{(\sigma_{Z^+})^2}\right) \quad (10)$$

식 (10)에 대하여 가우시안 경사하강법을 적용하여 결합가중치($\triangle\omega_k$)를 탐색하며 학습률 β와 결합하여 $w_k^{new} = (w_k^{old} - \beta \bullet \triangle w_k)$와 같이 새로운 결합가중치($w_k^{new}$)가 생성된다. 결합가중치는 정규화 과정($w_k^{new} / \sum_{k=1}^{K} w_k^{new}$)을 거쳐 모든 가중치의 합이 1인 되는 정규화 가중치로 변형된다. 이러한 가중치 탐색은 사전에 설정된 종료 조건을 만족할 때까지 반복되며 최적의 결합가중치(w^*)는 학습표본의 $GM(X,w)$이 극대화되는 임계점에서의 결합가중치로 설정된다.

Ⅳ. 연구 설계

4.1 데이터 수집 및 변수 선정

본 연구에서는 GMOPT의 성능을 검증하기 위해 한국과 더불어 공개데이터인 자료인 러시아, 폴란드 3개국의 기업 부도 데이터를 수집하였다.

4.1.1 한국 기업 부도 데이터

본 연구에서 활용된 국내 기업의 부도기업 표본은 2015부터 2018까지의 기간 중 당좌부도 발생, 회사정리/구조조정 절차 개시 및 은행연합회 신용정보등록 등에 해당하는 500개 외부감사 비금융기업이며, 건전 기업은 외부감사 비금융기업 중 부도 사유와 관련 없는 10,000개의 기업-년도별 자료로 구성하였다.

입력 변수는 한국상장협의회 TS2000에서 제공하는 재무자료를 기초로 구성하였다. 입력 변수 선정은 두 단계에 따라 진행하였다. 먼저 첫 번째 단계에서, 부도예측을 위한 변수로 일차적으로 30개의 재무비율을 수집하였다. 최종 변수를 선정하기 위하여 재무비율에 대한 요인 분석을 통하여 수익성, 부채상환능력, 레버리지, 자본구조, 유동성, 활동성, 규모의 7개 그룹으로 분류하였으며, 그룹별로 AUROC가 가장 높은 재무비율을 최종 입력 변수로 선정하였다. 7개 그룹별 입력 변수의 AUROC는 표 2에 제시되어 있다.

표 2. 국내 기업의 부도 예측 입력 변수의 AUROC
Table 2. AUROC of variables in Korea bankruptcy

Group	Variable	AUROC
Profitability	Ordinary income to total assets	54.3
Debt coverage	EBITDA to interest expenses	53.1
Leverage	Total debt to total assets	51.7
Capital structure	Retained earnings to total assets	51.3
Liquidity	Cash ratio	48.4
Activity	Inventory turnover ratio	33.4
Size	Total asset	23.7

4.1.2 러시아 기업 부도 데이터

러시아 부도 데이터는 선행연구[16]에서 사용한 데이터로 https://github.com/yzelenkov/classification_with_feature_selection에서 수집하였다. 수집된 표본은 456개의 부도기업과 2,001개의 정상기업으로 구성되어 불균형 비율은 약 1:4 수준이다.

부도예측에 활용되는 입력 변수는 재무 및 비재무 그룹으로 구성되어 있다. 재무 그룹은 레버리지, 유동성, 자본구조, 수익성, 활동성, 투자활동성, 투자수익률, 부채상환능력, 규모의 9개 그룹의 재무비율로 구성되어 있다. 비재무 그룹은 시장점유율, 채권청구비율, 외국의 경제제재 여부, 공급 신뢰도, 국가소유 여부, 연체금 미지급 여부 운영의 자율성의 7개 평가지표로 구성되어 있다[16]. 재무 및 비재무 그룹의 재무비율 및 평가지표에 대한 AUROC는 표 3에 기술되어 있다.

표 3. 러시아 기업의 부도 예측 입력 변수의 AUROC
Table 3. AUROC of variables in Russia bankruptcy

Group	Variable	AUROC
Leverage	Total debt / Total assets	75.3
Liquidity	Current asset / Total asset	63.2
Capital structure	Equity / Total debt	62.7
Profitability	Operating income / Total assets	53.5
Activity	Sales / Total assets	47.6
Cash flow	Cash flow / Equity	41.9
Investment profitability	Net Income / CAPEX	36.9
Debt coverage	EBITDA / Interest expense	33.3
Size	Equity	26.9
Qualitative group	Herfindahl-Hirschman Index	49.3
	claims to other companies	62.5
	sanctions of foreign States	51.4
	unreliable supplier	50.6
	Government control	49.7
	level of overdue	35.9
	coefficient of autonomy	35.7

4.1.3 폴란드 기업 부도 데이터

폴란드 기업의 표본은 개발 도상국 국가의 시장정보 데이터 베이스인 Emerging Markets Information

Service(EMIS)에서 수집한 것으로 2000년에서 2012년까지 발생한 부도 기업 410개와 5,500개의 정상 기업으로 구성되어 있다. 현재 폴란드 기업 부도 데이터는 UCI Machine Learning Repository에 공개되어 있으며 https://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data를 통해 수집 가능하다[17]. 최종 입력 변수는 레버지리, 수익성, 부채회전율, 활동성, 유동성, 규모, 자본구조, 성장성의 총 8개 그룹의 재무비율로 구성되어 있으며, 재무비율의 AUROC는 표 4에 제시되어 있다.

표 4. 폴란드 기업 부도예측에 입력변수의 AUROC
Table 4. AUROC of variables in Poland bankruptcy

Group	Variable	AUROC
Leverage	Total debt / Total asset	71.4
Profitability	Total expense / Sales	71.2
Debt turnover	Borrowings / Sales	70.3
Activity	Sales / Receivables	56.4
Liquidity	Current asset / Current liability	44.7
Size	Total assets	33.8
Capital structure	Working capital / Fixed assets	30.6
Growth	Sales growth	20.9

4.2 연구 모델 설계

본 연구에 활용된 연구모델은 python의 scikit-learn을 라이브러리로 활용하여 구축하였다. 본 연구에서는 의사결정나무를 약분류자로 활용한 AdaBoost, GBM 및 XGBoost를 벤치마크 모델로 선정하였다. 부스팅 모델에서 초모수의 최적화를 위하여 그리드 서치 기법을 이용하여 기저분류자의 depth는 1~10개, 기저분류자 수는 10~30개의 범위에서 학습표본에 대하여 정확도를 극대화하는 최적의 조합을 구성하였다. 그리드 서치의 결과 최적의 depth와 기저분류자 수 조합은 AdaBoost의 경우 2와 20, GBM의 경우 3과 15, XGBoost는 2와 17로 탐색되었다. 부스팅 모델의 학습의 종료조건으로 반복활동수를 10,000회로 설정하였다. 최적의 결합가중치는 학습표본에 대하여 특이도와 민감도의 차이가 극소화되는 임계점, 즉 학습표본을 대상으로 GM이 극대화되는 임계점에서의 결합가중치로 설정하였다.

V. 실증 분석 결과

표 5는 한국, 러시아, 폴란드 3개 국가의 부도 데이터에 대한 GMOPT와 벤치마크 모델의 성능을 비교한 결과로서 분류모델별로 3회의 10-Fold 교차검증을 반복하여 총 30회의 교차타당성 분석을 수행한 결과로서 특이도(SPE), 민감도(SEN), 정확도(ACC), F1-Score(F1), GM, AUROC 등의 성능지표를 제시하고 있다. 표 5의 모델 성능 분석 결과를 요약하면 다음과 같다.

표 5. 모델 성능 분석 결과
Table 5. Model performance results
패널 A. 한국기업 부도예측 (불균형 비율 1:20)
Panel A. Korean corporate bankruptcy (IR 1:20)

Metrics	Training set				Test set			
	Ada boost	GBM	XGB	GM OPT	Ada boost	GBM	XGB	GM OPT
SPE	0.99	0.99	0.99	0.80	0.99	0.99	0.99	0.80
SEN	0.05	0.07	0.07	0.80	0.04	0.05	0.06	0.73
ACC	0.95	0.96	0.96	0.80	0.95	0.95	0.95	0.79
F1	0.08	0.11	0.11	0.28	0.06	0.08	0.10	0.25
AUROC	0.52	0.53	0.53	0.80	0.52	0.53	0.53	0.76
GM	0.22	0.26	0.26	0.80	0.18	0.21	0.22	0.76

패널 B. 러시아기업 부도예측(불균형 비율 1:4)
Panel B. Russian Corporate Bankruptcy (IR 1:4)

Metrics	Training set				Test set			
	Ada boost	GBM	XGB	GM OPT	Ada boost	GBM	XGB	GM OPT
SPE	0.96	0.99	0.98	0.82	0.95	0.98	0.97	0.79
SEN	0.49	0.33	0.46	0.81	0.43	0.30	0.40	0.68
ACC	0.88	0.86	0.88	0.82	0.85	0.85	0.86	0.77
F1	0.48	0.45	0.53	0.38	0.41	0.38	0.44	0.30
AUROC	0.73	0.66	0.72	0.82	0.69	0.64	0.68	0.73
GM	0.69	0.57	0.67	0.82	0.63	0.54	0.62	0.73

패널 C. 폴란드기업 부도예측(불균형 비율 1:13)
Panel C. Polish corporate bankruptcy (IR 1:13)

Metrics	Training set				Test set			
	Ada boost	GBM	XGB	GM OPTt	Ada boost	GBM	XGB	GM OPT
SPE	0.99	0.99	0.99	0.90	0.99	0.99	0.99	0.88
SEN	0.48	0.32	0.44	0.90	0.42	0.29	0.39	0.76
ACC	0.96	0.95	0.96	0.90	0.95	0.95	0.95	0.87
F1	0.63	0.47	0.59	0.77	0.57	0.43	0.54	0.66
AUROC	0.73	0.66	0.72	0.90	0.70	0.64	0.69	0.82
GM	0.69	0.56	0.66	0.90	0.64	0.53	0.62	0.82

첫째, 산술평균에 기반한 정확도는 벤치마크 모델이 GMOPT 모델과 비교하여 상대적으로 우수한 성능을 가지는 것으로 확인되었다. 검증 표본에서 AdaBoost의 정확도는 한국(95%)/러시아(85%)/폴란드(95%)로 확인되었다. GBM와 XGBoost는 각각 한국(95%)/러시아(85%)/폴란드(95%), 한국(95%)/러시아(86%)/폴란드(95%) 수준의 정확도를 보여주었다. 반면 GMOPT는 한국(79%)/러시아(77%)/폴란드(87%)의 정확도를 보이고 있다.

둘째, 다수 범주에 대한 특이도 측면에서 벤치마크 모델은 GMOPT와 비교하여 우수한 것으로 분석되었다. 검증표본에 대하여 AdaBoost는 한국(99%)/러시아(95%)/폴란드(99%)의 특이도를 가지며 GBM의 경우 한국(99%)/러시아(98%)/폴란드(99%)이며 XGBoost의 경우 또한 한국(99%)/러시아(97%)/폴란드(99%)의 높은 수준의 특이도를 보였다. 반면 GMOPT는 한국(80%)/러시아(79%)/폴란드(88%)의 특이도 수준으로 벤치마크 모델과 비교하여 상대적으로 저조한 성능을 보여주었다.

셋째, 소수 범주에 대한 민감도는 GMOPT가 벤치마크 모델과 비교하여 우수한 성능을 보였다. 검증표본에 대하여 AdaBoost의 민감도는 한국(4%)/러시아(43%)/폴란드(42%)이며, GBM은 한국(5%)/러시아(30%)/폴란드(29%) 수준이며, XGBoost 또한 유사하게 한국(6%)/러시아(40%)/폴란드(39%)로 벤치마크 모델은 불균형 비율과 비례하여 매우 저조한 수준의 민감도를 가진다. 다수 범주에 대한 특이도 분석 결과와 비교하여 보면 손실함수 극소화를 학습 알고리즘으로 채택하고 있는 전통적인 부스팅 모델은 범주 불균형 문제에 대하여 특이도에 초점을 맞추어 정확도가 증가하지만, 민감도를 무시하기 때문에 부도 예측의 주 관심 대상인 부도 기업이라는 소수 범주에 대한 분류기능이 저하되었음을 보여주고 있다. 반면, GMOPT의 민감도는 한국(73%)/러시아(68%)/폴란드(76%) 수준으로 특이도와 유사한 수준을 보여주고 있다. 이러한 결과는 GMOPT는 학습 알고리즘으로 GM 최적화를 채택함으로써 특이도와 민감도가 동시에 고려된 균형 학습을 진행할 수 있기 때문에 금융산업의 범주 불균형 문제 해결에 보다 강건하고 적합성이 높은 모델임을 의미한다.

넷째, 소수 범주와 다수 범주를 동시에 고려하는 GM의 측면에서 GMOPT가 벤치마크 모델과 비교하여 개선된 성능을 보였다. GM측면에서 AdaBoost는 한국(18%)/러시아(63%)/폴란드(64%)의 수준이며, GBM의 경우 한국(21%)/러시아(54%)/폴란드(53%) 수준이다. 또한 XGBoost도 한국(22%)/러시아(62%)/폴란드(62%)로 매우 저조한 수준의 성능을 보여주고 있다. 반면 GMOPT는 한국(76%)/러시아(73%)/폴란드(82%) 수준으로 벤치마크 모델과 비교하여 탁월한 성능개선 효과를 보여주고 있다. 유사하게 다수 범주와 소수 범주를 동시에 고려하는 또 다른 대체적 성능지표로 활용되는 AUROC와 F1-Score 측면에서도 GM과 유사하게 탁월한 성능개선을 확인하였다. 이는 본 연구에서 GM 최적화 방식이 금융 비즈니스의 범주 불균형 문제 해결에 있어 부스팅 모델의 성능개선에 기여하고 있음을 의미한다.

표 6은 반복측정 분산분석을 이용한 GMOPT와 벤치마크 부스팅 모델의 성능차이에 대한 통계검증 결과를 요약한 것이다. 표 6에서 모든 데이터의 GM 및 AUROC에 대한 F-value는 1% 수준이하에서 유의한데 이는 최소 하나의 모델이 다른 모델보다 유의하게 높은 성능을 가지고 있음을 의미한다. 한국기업 부도데이터에 대한 모델간 성능차이를 확인하기 위한 사후검정으로 Duncan검정을 이용하여 쌍체비교를 수행한 결과, 벤치마크 부스팅 모델은 모두 그룹 a로 분류되어 벤치마크 모델들의 분류성능에 유의한 차이가 없는 것으로 분석되었다. 반면 상대적으로 우수한 분류성능을 가진 GMOPT는 그룹 b로 분류되었으며, 이는 벤치마크 모형과 비교하여 유의하게 높은 성능을 가지고 있음을 의미한다. 폴란드와 러시아 데이터에서 GBM은 a그룹, AdaBoost와 XGBoost는 b그룹으로 분류되어 b그룹이 a그룹보다 에서 유의적으로 높은 성능을 가지고 있는 것으로 분석되었다. c그룹으로 분류된 GMOPT는 a와 b 그룹의 벤치마크 모델보다 유의적으로 높은 성능을 가지는 것으로 나타났다. 본 연구에서 t검정을 대체하여 반복측정 분산분석을 이용하는 이유는 t검정으로 두 모델의 성능 비교를 여러 번 수행하면 제1종 오류의 확률이 증가하게 되는데 이를 Bonferroni의 inequality라 한다. 반복측정 분산분석에서는 이러한 오류를 보정하기 위하여 t값과 F값의 임계치를

상향 조정한 Tukey, Duncan 등 사후검정을 수행하여 5% 유의수준에서 성능 차이가 발생하는 모델별로 그룹화하여 모델별 성능차이을 명확하게 구분하기 때문이다[18].

표 6. 모델 성능에 대한 반복측정 분산분석 결과
Table 6. Repeated measure ANOVA results
패널 A. 한국기업 부도예측 (불균형 비율 1:20)
Panel A. Korean corporate bankruptcy (IR 1:20)

Metrics	Classification model					
	Ada boost	GBM	XGB	GM OPTt	F-value	P-Value
GM	0.18	0.21	0.22	0.76	471.2	0.0001
	a	a	a	b		
AUROC	0.52	0.53	0.53	0.76	1289.2	0.0001
	a	a	a	b		

패널 B. 러시아기업 부도예측 (불균형 비율 1:4)
Panel B. Russian corporate bankruptcy (IR 1:4)

Metrics	Classification model					
	Ada boost	GBM	XGB	GM OPT	F	P-Value
GM	0.63	0.54	0.62	0.73	63.5	0.0001
	b	a	b	c		
AUROC	0.69	0.64	0.68	0.73	33.6	0.0001
	b	a	b	c		

패널 C. 폴란드기업 부도예측 (불균형 비율 1:13)
Panel C. Polish corporate bankruptcy (IR 1:13)

Metrics	Classification model					
	Ada boost	GBM	XGB	GM OPTt	F	P-Value
GM	0.64	0.53	0.62	0.82	176.8	0.0001
	b	a	b	c		
AUROC	0.70	0.64	0.69	0.82	166.4	0.0001
	b	a	b	c		

VI. 결론 및 향후 과제

선행 연구에서 범주 불균형 문제의 해결을 위하여 데이터 샘플링 및 Cost-sensitive 기법등 다양한 해결기법이 제안되었지만, 바젤 협약과 같은 실무적 규제안에 따라 이러한 기법을 직접적으로 적용하는 경우 금융 차주의 위험 특성이 왜곡되는 데이터 유효성 문제가 발생하는 이유로 실무적 활용도가 부족하다. 부스팅 알고리즘은 범주 불균형 문제의 해결에 활용된 앙상블 기법의 한 종류로 다수의 학습된 약분류자들의 예측결과를 결합하여 최종 예측결과를 도출하는 방식으로 오분류 가능성이 높은 소수 범주의 학습에 초점을 맞출 수 있다는 장점이 있지만, 손실함수 극소화를 분류모델의 목적함수로 설정하기 때문에 범주 간 균형 학습을 수행하기 어렵다는 한계가 있다.

본 연구는 금융 분야의 대표적인 범주 불균형 문제로서 기업부도 예측문제를 효과적으로 해결하기 위해 GMOPT를 제안하였다. GMOPT는 GM 극대화를 분류모델의 목적함수로 설정하였으며, 이산확률함수인 GM을 중심극한정리를 기초로 미분 가능한 다변량 정규분포함수로 치환한 후 가우시안 경사하강법을 적용하여 최적화 과정을 수행한다. 한국, 러시아, 폴란드 3개 국가의 기업 부도 데이터를 대상으로 AdaBoost, GBM, XGBoost의 벤치마크 모델과 GMOPT의 성능을 비교하였다. 분석 결과, GMOPT는 다수 범주와 소수 범주에 대한 균형 학습을 통하여 범주 불균형 문제를 대하여 부스팅 모델의 성능을 유의적으로 개선할 수 있음을 확인하였다. 이러한 결과는 본 연구에서 제안하고 있는 GMOPT는 신용위험 식별능력의 성능개선을 통하여 금융기관의 재무건전성 제고에 기여할 수 있음을 의미한다.

본 연구에서는 금융실무에 대한 활용가능성을 고려하여 범주 불균형 문제를 효과적으로 해결하기 위한 새로운 GM 최적화 학습 알고리즘을 제안하고 있지만, 향후 연구 방향으로서 다음과 같은 점이 고려되어야 한다.

첫째, 본 연구는 금융기관의 규제적 환경을 고려하여 범주 균형 문제에 대한 GMOPT의 성능을 확인하지 못하였다. 본 연구에서 제안한 GMOPT는 범주 간 균형 데이터에서도 성능을 최적화할 수 있을 것으로 기대하기 때문에 다른 범주 불균형 해결 기법과 결합된 합성 GMOPT의 성능을 검토하는 연구를 진행하고자 한다.

둘째, 본 연구는 범주 불균형 문제에서 부스팅 알고리즘의 성능개선을 위한 GM 최적화 기법을 제안하였다. 그러나, GM 개선에 초점을 맞춘 부작용으로 모델의 정확도가 감소하고 있다. 이러한 문제의 해결을 위하여 분류 목적에 따라 정확도, F1-Score, AUROC 등 다양한 성능지표의 최적화 기법의 개발을 향후 연구에서 수행하고자 한다.

셋째, 본 연구는 범주 불균형 문제의 효과적 대처를 위하여 선행연구에서 범주 불균형 문제에 다양하게 적용되어왔던 부스팅 기법의 성능개선을 목적으로 성능지표의 최적화 기법을 제안하고 있다. 향후 연구에서는 인공신경망, 서포트벡터머신(Support vector machine), 딥러닝(Deep learning) 등 다양한 분류 기법을 대상으로 범주 불균형 문제에 대한 성능개선 효과를 검증하고자 한다.

References

[1] P. Kang and S. Cho, "EUS SVMs: ensemble of undersampled SVMs for data imbalance problems", ICONIP-2006: Neural Information Processing, pp. 837-846, 2006. https://doi.org/10.1007/11893028_93

[2] S. Blochwitz and S. Hohl, "Validation of Banks' Internal Rating Systems - A Supervisory Perspective", The Basel II Risk Parameters, pp. 243-262, 2005. https://doi.org/10.1007/3-540-33087-9_11.

[3] J. Du, C. M. Vong, C. M. Pun, P. K. Wong, and W. F. Ip, "Postboosting of classification boundary for imbalanced data using geometric mean", Neur-al Networks, Vol. 96, pp. 101-114, Dec. 2017. https://doi.org/10.1016/j.neunet.2017.09.004.

[4] H.-J. Kim and H.-D. Kim, "Predicting Loan Defa-ults with Gradient Boosting and Balanced Classifi-cation", Journal of KIIT, Vol. 12, No. 1, pp. 155-164, Jan. 2014. https://doi.org/10.14801/kiitr.2014.12.1.155.

[5] M. Galar, A. Fernandez, E. Barrenechea, H. Busti-nce, and F. Herrera, "A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches", IEEE Transactions on Systems, Man and Cybernetics Part C: Applic-ations and Reviews, Vol. 42, No. 4, pp. 463-484, Jul. 2012. https://doi.org/10.1109/TSMCC.2011.2161285.

[6] H. He and E. A. Garcia, "Learning from imbalan-ced data", IEEE Transactions on Knowledge and Data Engineering, Vol. 21, No. 9, pp. 1263-1284, Sep. 2009. https://doi.org/10.1109/TKDE.2008.239.

[7] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class imb-alanced data:Review of methods and applications", Expert Systems with Applications, 73, pp. 220-239, May 2017. https://doi.org/10.1016/j.eswa.2016.12.035.

[8] N. V. Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer, "SMOTEBoost: Improving prediction of the minority class in boosting", In Proceedings of the Knowledge Discovery and Data Mining Conference (KDD), Cavtat-Dubrovnik, Croatia, pp. 107-119, Sep. 2003. https://doi.org/10.1007/978-3-540-39804-2_12.

[9] S. Hu, Y. Liang, L. Ma, and Y. He, "MSMOTE: Improving classification performance when training data is imbalanced", In Proceedings of the 2nd International Workshop on Computer Science and Engineering, Qingdao, China, pp. 13-17, Oct. 2009. https://doi.org/10.1109/wcse.2009.756.

[10] C. Seiffert, T. Khoshgoftaar, J. V. Hulse, and A. Napolitano, "Rusboost: A hybrid approach to alleviating class imbalance", IEEE Transactions on Systems, Man, and Cybernetics - Part A: S-ystems and Humans, Vol. 40, No. 1, pp. 185-197, Jan. 2010. https://doi.org/10.1109/tsmca.2009.2029559.

[11] H. Guo and H. Viktor, "Learning from imbalan-ced data sets with boosting and data generation: The DataBoost-IM approach", SIGKDD Explorati-ons Newsletter, Vol. 6, No. 1, pp. 30-39, Jun. 2004. https://doi.org/10.1145/1007730.1007736.

[12] W. Fan, S. J. Stolfo, J. Zhang, and P. K. Chan, "Adacost: Misclassification cost-sensitive boosting", In Proceedings of the 16th International Conference on Machine Learning, Bled, Slovenia, Vol. 99, pp. 97-105, Jun. 1999.

[13] K. M. Ting, "A comparative study of

cost-sensitive boosting algorithms", In Proceedings of the 17th International Conference on Machine Learning, Stanford, CA, USA, pp. 983-990, Jun. 2000.

[14] M. Joshi, V. Kumar, and R. Agarwal, "Evaluating boosting algorithms to classify rare classes: Com-parison and improvements", In Proceedings of the IEEE International Conference on Data Mining, San Jose, CA, USA, pp. 257-264, Nov.-Dec. 2001. https://doi.org/10.1109/icdm.2001.989527.

[15] Y. Sun, M. S. Kamel, A. Wong, and Y. Wang, "Cost-sensitive boosting for classification of imba-lanced data", Pattern Recognition, Vol. 40, No. 12, pp. 3358-3378, Dec. 2007. https://doi.org/10.1016/j.patcog.2007.04.009.

[16] Y. Zelenkov, E. Fedorova, and D. Chekrizov, "Two-step classification method based on genetic algorithm for bankruptcy forecasting", Expert Systems with Applications, Vol. 88, pp. 393-401, Dec. 2017. https://doi.org/10.1016/j.eswa.2017.07.025.

[17] A. D. Pozzolo, O. Caelen, Y.-A. Borgne, S. Waterschoot, and G. Bontempi, "Learned lessons in credit card fraud detection from a practitioner perspective", Expert Systems with Applications, Vol. 41, No. 10, pp. 4915-4928, Aug. 2014. https://doi.org/10.1016/j.eswa.2014.02.026.

[18] E. R. Girden, "ANOVA: Repeated Measures", Sage University Paper.

저자소개

김 명 종 (Myoung-Jong Kim)

1991년 2월 : 성균관대학교 경영학(학사)
1993년 2월 : 성균관대학교 경영학(석사)
2002년 2월 : 한국과학기술원 경영공학(박사)
2010년 3월 ~ 현재 : 부산대학교 경영학과 교수
관심분야 : 핀테크, 딥러닝, 최적화

김 윤 후 (Yun-Hu Kim)

2024년 5월 : 미국퍼듀대학교 메인캠퍼스 Computer Science(공학사)
2024년 8월 ~ 현재 : 미국퍼듀대학교 메인캠퍼스 Computer Science 석사과정
관심분야 : AI, Natural Langage

김 재 남 (Jae-Nam Kim)

2024년 2월 : 부산대학교 경영학과 (경영학학사)
관심분야 : AI, Fintech, Risk Management