Journal of the Korean Data Analysis Society (April 2023), 25(2), 749-762.
https://doi.org/10.37727/jkdas.2023.25.2.749

금융 분야의 범주 불균형 문제 해결을 위한 성과 최적화 기반의 부스팅 학습*

김명종[1], 안재현[2], 김윤후[3]

요 약

본 연구에서는 금융 분야의 범주 불균형 문제에 적용된 부스팅 계열 알고리즘의 성과 개선을 위하여 성과지표에 대한 직접적인 최적화 기법을 도입한 GMBoost(Geometric Mean-based Boosting) 기법을 제안한다. 본 연구에서는 기업 부실, 카드 연체 및 카드 사기와 같은 대표적인 범주 불균형 금융 문제를 대상으로 GMBoost의 성과를 비교했다. 성과 비교를 위해 벤치마킹 모형으로 부스팅 계열 알고리즘인 AdaBoost, GBM, XGBoost를 활용하였으며, 이들과 GMBoost간의 성과 차이를 비교하였다. 30회의 교차 검증을 통하여 분석한 결과, 첫째, 다수 범주의 특이도에 초점을 맞추어 학습을 진행하여 범주 불균형 문제에서 유의한 성과를 내지 못하는 기존 부스팅 모형과는 달리 GMBoost는 다수 범주의 특이도와 소수 범주의 민감도를 동시에 고려하는 균형적인 학습을 진행함으로써 범주 불균형 문제 해결에 효과적임을 확인하였다. 둘째, 기존 부스팅 모형과 비교하여 GMBoost는 GM 및 AUC 측면에서 우수한 예측 성과를 보여주고 있으며, GM 및 AUC에 대한 t-검정 결과에서도 유의적인 성과 차이를 보여주었다.

주요용어 : 범주 불균형, GMBoost, 성과지표 최적화, 가우시안 경사하강법

1. 서론

4차 산업혁명이 산업 전반에 급속도로 확산되고 4차 산업혁명의 핵심 기술이 기업 경쟁력 제고의 핵심 원천으로 인식됨에 따라 ICT(information communication technology)에 대한 기업들의 투자가 확대되고 있다. 금융 산업에서도 생체 인식, 클라우드 컴퓨팅, 모바일 컴퓨팅 등의 정보기술이 금융 서비스 혁신에 적용되고 있으며, 빅데이터와 인공지능 등의 지능정보기술은 부도예측, 신용평가, 카드사기적발 등 분류 및 예측 작업에 광범위하게 적용되고 있다(Park, Lee, Choi, 2009; Cho, Choi, Cho, 2022; Lee, Joo, Jeon 2017; Kwon, Park 2022)

그러나 금융 문제에 적용된 분류 모형의 성과 개선을 위해서는 범주 불균형 문제(class imbalance problem)가 필수적으로 고려되어야 한다(Galar et al., 2012). 범주 불균형 문제는 데이터 표본이 특정 범주에 현저하게 편향된 데이터 분포의 왜곡 문제로서 분류 모형의 성과를 저하시키는 부정적인 데이터 품질 문제이다. 금융 분야에서 부도예측, 카드사기, 카드연체 등 범주 불균형 문제가 빈

*본 연구는 과학기술정보통신부 및 정보통신기획평가원의 융합보안핵심인재양성사업의 연구 결과로 수행되었음(IITP-2023-2022-0-01201).
[1](교신저자) 46241 부산시 금정구 부산대학로 63번길 2, 부산대학교 경영학과교수. E-mail: mjongkim@pusan.ac.kr
[2]46241 부산시 금정구 부산대학로 63번길 2, 부산대학교 경영학과 학사과정. E-mail: bonjour8084@pusan.ac.kr
[3]미국 퍼듀대학교 메인캠퍼스 Department of Computer Science 학사과정. E-mail: kim3438@purdue.edu
[접수 2023년 2월 17일; 수정 2023년 3월 15일; 게재 확정 2023년 3월 18일]

번하게 관찰되며, 범주 불균형이 존재하는 경우, 다수 범주(majority class)는 소수 범주(minority class)의 경계영역(decision boundary)을 침투하여 분류 경계선(decision hyperplane)이 소수 범주 방향으로 편향된다. 이에 따라 다수 범주의 특이도(specificity)는 증가하지만, 소수 범주의 민감도(sensitivity)는 급격하게 감소하게 된다. 범주 균형을 전제로 개발된 분류 모형은 범주 불균형 표본에 대하여 소수 범주를 무시하고 다수 범주에 치중된 불균형 학습을 진행하게 되는데 특히, 범주 불균형이 심각한 경우 소수 범주에 대한 분류 기능을 상실하게 된다(Kang, Cho, 2006). 범주 불균형 문제는 성과지표의 유효성 문제로 연계되는데 산술평균 개념에 기초한 정확도(accuracy)는 가장 보편적으로 활용되는 성과지표이지만, 다수 범주의 특이도와 소수 범주의 민감도를 동시에 고려하지 못하기 때문에 범주 불균형 영역의 문제에 적용된 분류 모형의 성과를 정확하게 평가하지 못하게 된다. 이러한 문제점을 개선하기 위하여 기하평균 정확도(Geometric Mean Accuracy, GM), F1 Score, AUC(Area Under ROC Curve)와 같은 새로운 성과지표(performance metrics)들이 활용되고 있다(Du et al., 2017; Kubat, Holte, Matwin, 1997).

범주 불균형 문제에 적용된 분류 모형의 성과를 개선을 위하여 해결되어야 할 또 다른 난제는 대부분의 분류 알고리즘이 데이터의 균형 분포를 전제로 손실함수(loss function) 극소화를 학습과정의 목적함수(object function)로 채택하고 있는 문제이다. 정확도, GM, AUC와 같은 성과지표가 성과평가와 더불어 학습과정의 목적함수로 채택되어야 함에도 불구하고 이를 목적함수로 활용하지 못하는 이유는 성과지표는 기본적으로 이산확률함수(discrete function)로 측정되며 누적분포함수(cumulative distribution function, CDF)는 계단형 함수(stepping function) 혹은 비감소 함수(non-decreasing function)로 정의되기 때문에 직접적인 미분이 불가능하거나 미분이 가능하더라도 경사가 0으로 측정되어 경사하강법(gradient descent method)과 같은 최적화 알고리즘의 적용이 거의 불가능하기 때문이다. 이러한 문제를 우회하기 위하여 RMSE(root mean square error), 로지스틱 손실(logistic loss), 힌지 손실(hinge loss)과 같은 손실함수가 대체적인 목적함수로 채택되고 있지만, 이러한 손실함수는 범주 균형을 가정하기 때문에 범주 불균형 문제에 적용된 분류 모형의 성과를 저하시키는 요인이 된다.

본 연구에서는 금융 분야의 범주 불균형 문제에 적용된 부스팅 모형의 성과 개선을 위하여 GM 최적화 기반의 부스팅 모형인 GMBoost를 제안한다. 부스팅 학습은 기저분류자(base classifiers)의 예측 결과를 결합하여 강분류자(strong classifiers)의 최종 결론을 도출하는 학습 방법으로서, 분류 및 예측 연구에서 가장 효율적인 분류 기법의 하나로 평가받고 있다(Galar et al., 2012). 그러나, 부스팅 학습은 손실함수 극소화를 목적함수로 채택하고 있기 때문에 범주 간 특이도와 민감도를 동시에 고려하지 못하며, 오히려 다수 범주의 특이도에 초점을 맞춘 학습을 진행하는 경향이 있다.

GMBoost는 범주 불균형 문제에 대한 성과 개선을 위하여 다음과 같은 수정 알고리즘을 도입하였다. 첫째, 분류 모형의 목적함수를 손실함수 극소화에서 GM 극대화로 대체하여 범주 불균형 문제에 대한 성과지표의 유효성을 확보하였다. 둘째, 이산확률함수로 측정된 GM에 대하여 다변량 정규분포 가정(the assumption of multivariate normal distribution)을 도입하여 미분 가능한 정규분포함수로 치환하였다. 셋째, 치환된 GM에 대하여 확률개념 기반의 가우시안 경사하강법(Gaussian gradient descent method)을 적용하여 성과지표에 대한 직접적인 최적화를 수행하였다.

본 연구에서는 GMBoost의 성과 검증을 위하여 기업 부실(bankruptcy), 카드 연체(card insolvency) 및 카드 사기(card fraud)와 같은 금융 분야의 범주 불균형 문제를 선정하였다. AdaBoost, GBM,

XGBoost 모형을 성과 비교를 위한 벤치마킹 모형으로 활용하였으며, 각 금융 데이터에 대하여 30회 교차 검증(cross validation)을 수행한 주요 연구 결과는 다음과 같다. 첫째, 다수 범주의 특이도에 초점을 맞추어 진행하는 기존 모형과는 달리 GMBoost는 특이도와 민감도를 동시에 고려하는 균형 학습을 진행함으로써 범주 불균형 문제 해결에 효과적인 것으로 판정되었다. 둘째, 기존 모형과 비교하여 GMBoost는 GM 및 AUC 측면에서 우수한 예측 능력을 가지며, t-검정 결과 유의적인 성과 차이를 보여주었다.

본 연구는 다음과 같이 구성된다. 제 2 장에서는 분류 모형의 성과지표와 범주 불균형 문제를 해결에 적용된 앙상블 학습에 대한 선행연구를 고찰한다. 제 3 장에서는 AdaBoost, GBM, XGB 및 GMBoost 알고리즘에 대하여 기술한다. 제 4 장에서는 본 연구의 범주 불균형 데이터에 대하여 설명한다. 제 5 장에서는 기존 부스팅 알고리즘과 GMBoost에 대하여 30회 교차 검정을 수행한 연구 결과를 제시하고, 마지막으로 제 6 장에서는 본 연구의 결론과 향후 연구 방향에 대하여 소개한다.

2. 선행연구고찰

2.1 분류 모형의 성과지표

Table 1은 이범주 분류(binary classification)에 대한 정오행렬표(confusion matrix)로 각 셀(cell)에는 정분류 빈도(다수 범주(TN) 및 소수 범주(TP))와 오분류 빈도(제1종 오류(FP) 및 제2종 오류(FN))가 표기된다. <표 1>에서 정확도는 범주 균형 표본의 성과 평가에 일반적으로 활용되는 성과 측정치이지만, 범주 간 불균형 분포와 오분류 비용을 고려하지 못하는 문제로 인하여 범주 불균형 표본에 대하여 유효한 성과지표가 되지 못한다. 이러한 한계점을 보완하기 위한 성과지표로서 다수 범주의 특이도와 소수 범주의 민감도를 동시에 고려하는 GM 및 AUC와 같은 측정치가 이용되고 있다(Fawcett, 2006; Davis, Goadrich, 2006). GM은 다수 범주와 소수 범주의 특이도와 민감도의 기하평균으로 측정되며 특이도와 민감도가 동일한(또는 유사한) 임계점에서 극대값을 가진다. AUC는 TPR과 FPR의 상충 관계(trade-off)에 기반한 ROC 곡선 아래의 면적으로 모형의 성과를 단일 측정치로 제시하기 때문에 모형 간 성과 비교가 용이하다. 임의 모형의 AUC는 0.5로 범주 간 분류 능력이 없음을 의미하며 완벽 모형(perfect model)의 AUC는 1로서 다수 범주와 소수 범주를 완벽하게 분류함을 의미한다. 일반적인 모형의 AUC는 0.5보다 크고 1보다 작으며, 1에 근접할수록 모형은 우수한 분류 능력을 가진 것으로 해석된다.

Table 1. Confusion matrix in binary classification

		Predicted Class	
		Bankrupt	Non-bankrupt
Actual Class	Bankrupt	True Positive (TP)	False Negative (FN)
	Non-bankrupt	False Positive (FP)	True Negative (TN)

Sensitivity(TPR): $TP/(TP+FN)$
Specificity(TNR): $TN/(FP+TN)$
Type 1 Error(FPR): $FP/(FP+TN)$
Type II Error(FNR): $FN/(TP+FN)$
Accuracy: $(TP+TN)/(TP+FN+FP+TN)$
GM: (Sensitivity • Specificity)1/2
AUC: $(1+TPR-FPR)/2=(TPR+TNR)/2$

2.2 범주 불균형 문제 해결을 위한 앙상블 학습

금융 분류 및 예측 문제의 해결을 위하여 다변량 판별분석 및 로지스틱 회귀분석과 같은 통계적 기법을 비롯하여 의사결정트리, 인공신경망, 서포트벡터머신(support vector machine) 및 앙상블 학습(ensemble learning) 등과 같은 다양한 머신러닝 기법들이 활용되고 있다. 초기의 모형들은 다수 범주와 소수 범주의 관측치가 동일한 대응 표본(paired sampling)을 대상으로 손실함수 최소화로 설정된 목적함수에 초점을 맞추어 분류 문제 해결에 적용되어 왔다. 그러나, 균형 분포를 전제로 가정한 분류 모형은 범주 불균형 문제가 포함된 현실 세계의 데이터 환경을 반영하지 못하며, 특히 다수 범주와 소수 범주의 데이터 분포 및 오분류 비용을 고려하지 못하기 때문에 금융 분야의 현실 문제 해결에 거의 적용될 수 없다는 비판을 받아왔다.

범주 불균형 문제가 분류 모형의 성과를 저해하는 중요한 문제로 인식되면서 범주 불균형 문제에 대처하기 위한 다양한 기법들이 제안되었으며, 데이터 샘플링, cost-sensitive 및 앙상블 학습 등이 대표적으로 활용되고 있다. 데이터 샘플링 기법은 데이터의 분포를 조작하여 균형 표본을 구성하는 방법으로 소수 범주의 표본을 인위적으로 늘리는 오버 샘플링(over-sampling)과 다수 범주의 표본을 인위적으로 제거하는 언더 샘플링(under-sampling) 기법이 있다. 데이터 샘플링은 범주 불균형 문제 해결을 위하여 가장 광범위하게 적용되고 있는 기법이지만, 샘플링 과정이 인위적이며, 정보 손실이나 데이터의 중복으로 인한 데이터 품질 문제가 발생할 수 있다(Galar et al., 2012; He, Garcia, 2009).

Cost-sensitive 방법은 범주 별로 오분류 비용의 가중치를 고려하는 방법으로 1) 각 범주의 관측치에 서로 다른 가중치를 비용 행렬(cost matrix)에 배분하는 방법, 2) 손실 함수에 오분류 비용을 포함하여 학습 알고리즘을 변경하는 방법 및 3) 비용 행렬을 베이지안 의사결정 영역 결정에 포함하는 방법 등 다양한 형태로 적용되고 있다(Haixiang et al., 2017). 데이터 샘플링과 비교하여 연산이 용이하고 데이터의 삭제 및 생성과 같은 인위적인 데이터 전처리 과정을 포함하지 않는다는 장점이 있지만, 범주 별 오분류 비용에 대한 정확한 추정이 어렵다는 한계로 데이터 샘플링과 비교하여 소수의 연구에 활용되고 있다.

앙상블 학습은 분류 모형의 다양성을 확보하고 예측 성과를 개선할 수 있다는 장점으로 인하여 다양한 예측 및 분류 문제의 성과를 개선하기 위한 탁월한 기법으로 평가되고 있다(Mellor et al., 2015). 앙상블 학습의 대표적인 학습 기법인 부스팅 학습은 선행 분류자에서 오분류된 표본에 높은 가중치를 부여함으로써 오분류 표본에 강화된 학습기회를 제공한다. 범주 불균형 문제에서 소수 범주의 오분류 가능성이 높기 때문에 부스팅 학습은 소수 범주에 보다 많은 학습기회를 제공한다(Kim, Kang, Kim, 2015). 앙상블 학습을 금융 분야의 문제 해결에 적용한 연구로서 Nanni, Lumini(2019)는 호주, 독일 및 일본 기업을 대상으로 기업부도예측 및 기업신용평가를 위한 개별 분류모형과 앙상블 모형의 예측 성과를 비교한 결과 앙상블 모형의 성과가 우수함을 보고하였다. Barboza, Kimura, Altman(2017)은 998개 미국 기업(건전: 449, 부도: 449)을 대상으로 단일 모형과 비교하여 부스팅 및 배깅 학습이 정확도와 AUC 측면에서 우수한 예측 능력을 가지고 있음을 확인하였다. Zieba, Tomczak, Tomczak(2016)은 5,910개 폴란드 기업(건전: 4,500 부도: 410)으로 구성된 범주 불균형 데이터를 대상으로 10회의 교차검정을 실시한 결과 XGBoost(AUC 평균: 94.5%)가 개별 분류기보다 우수한 예측성과를 가지고 있음을 확인하였다.

최근 연구에서는 앙상블 알고리즘의 성과 개선을 위한 다양한 방법들이 제안되고 있다. 가장 활발하게 적용되는 방법으로서 데이터 샘플링과 결합된 앙상블 모형이 범주 불균형 문제 해결에 보다 효과적인 것으로 보고되고 있다. UlagaPriya, Pushpa(2021)은 데이터 샘플링과 앙상블 학습을 결합하여 분석한 결과 SMOTEBagging 및 SMOTEBoost이 RUSBoosting 및 RUSBagging보다 기업부도 예측에 강건함을 실증하였다. Le et al.(2018b)은 클러스터 기반 부스팅 모형인 CBoost를 제안하였다. 분석 결과 CBoost는 GMBoost(Kim, Kang, Kim, 2015), 클러스터 기반 under-sampling 모형(Lin et al., 2017), 클러스터 기반 oversampling 모형(Le et al., 2018a)보다 기업부도 예측에 우수함을 확인하였다. 또 다른 방법으로 범주 불균형 문제 해결을 위하여 앙상블 학습 자체의 알고리즘을 수정하는 연구가 진행되고 있다. Kim, Kang, Kim(2015)은 산술평균 기반의 예측 오차 및 정확도를 채택하고 있는 AdaBoost 알고리즘을 GM으로 대체한 부스팅 알고리즘을 제안하였으며, 제안된 알고리즘은 불균형 표본과 균형 표본 모두의 분류 성과를 개선할 수 있음을 실증하였다.

3. 학습 알고리즘

3.1. 부스팅 알고리즘

대표적 부스팅 알고리즘인 AdaBoost(Freund, Schapire, 1997)는 선행 분류자에서 오분류된 표본에 높은 가중치를 부여하여 새롭게 생성된 분류자에서는 오분류된 표본에 더욱 강한 학습 기회를 부여한다. 강분류자는 기저분류자들의 예측 결과를 선형결합하여 최종 예측 결과를 도출한다. 그러나 AdaBoost는 기저분류자의 결합과정에서 특정 기저분류자의 결합가중치가 과도하게 설정되는 문제로 인하여 기저분류자의 다양성을 확보하기 어렵고 강분류자에 대한 최종 결과도 특정 기저분류자의 결과에 편의된다는 단점이 존재한다.

이러한 단점을 개선하기 위한 GBM은 최적화 기법에서 자유로운 확장성을 제공하기 위한 부스팅 알고리즘으로 AdaBoost와 비교하여 강건하고 해석 가능하며 비교적 적은 데이터를 이용할 때에도 우수한 성과를 제공한다(Friedman, 2001). GBM은 각 기저분류자에서 경사하강법을 반복적으로 적용하여 잔차를 최소화하며 최종 결과는 각 기저분류자의 예측결과를 합산하여 산출한다. GBM은 손실함수의 변경이 용이하지만 손실함수가 미분가능해야 한다는 제약 사항이 존재하며 특히, GBM은 최적화 방법으로 Greedy 알고리즘을 채택하기 때문에 데이터의 크기가 커질수록 학습 속도가 느려진다는 문제가 있다.

XGBoost(Chen, Guestrin, 2016)는 GBM의 학습 속도를 개선하고, 확장성을 높인 알고리즘이다. XGBoost는 기저분류자로 CART(Classification And Regression Tree)를 사용하며 트리에서의 최적의 분기점(Split Point)를 찾는다. 분기점 탐색과정에서 일반적인 Greedy 알고리즘을 통한 탐색기법은 데이터의 크기가 커질수록 학습 시간이 길어진다는 문제가 발생하는데 이에 대하여 XGB는 최적에 근사한 분기점 탐색 알고리즘을 도입하여 Greedy 알고리즘에 비해 정확도는 떨어지지만, 분기점을 효율적이고 빠르게 찾아낸다. 특히 XGB는 데이터를 Block별로 나누어 저장함으로써 병렬연산을 통한 메모리의 효율적 사용과 학습 시간을 단축하는 효과도 제공한다.

이외에도 현재까지 다양한 부스팅 학습 기법이 제안되고 있으며, 부스팅 모형은 다양한 분류 및 예측 문제에 적용되어 탁월한 성과가 인정되고 있다. 하지만 부스팅 계열 알고리즘은 산술평균 개

넘의 예측오차 및 정확도를 학습 알고리즘으로 포함하기 때문에 범주 불균형 문제에 대하여 데이터 불균형과 오분류 비용을 고려하지 못한다는 단점이 있다. 범주 불균형 문제에서 부스팅 학습은 다수 범주의 특이도에 의존하여 정확도를 개선하려는 학습을 진행하게 되며, 결과적으로 분류 모형의 특이도는 높아지지만, 민감도는 급격하게 하락한다. 이에 따라 범주 불균형 문제를 효과적으로 해결하기 위해서는 부스팅 학습 알고리즘의 개선 또는 변경이 필요하다는 비판이 제기되고 있다(Galar et al., 2012).

3.2 GMBoost 알고리즘

GMBoost는 범주 불균형 문제를 효과적으로 해결하기 위하여 제안된 수정 부스팅 알고리즘으로 학습 절차는 다음과 같다. 다수 범주($Y=-1$)와 소수 범주($Y=+1$)로 구성된 이범주 분류문제(binary classification problem)에 대하여 부스팅 학습 모형의 기저분류자의 학습 결과를 가정해보자. 임의의 관측치 x를 다수 범주 관측치 x^-와 소수 범주 관측치 x^+로 구분하고, n^-와 n^+는 각각 다수 및 소수 범주의 관측치 수라 할 때 GM은 식 (1)과 같이 이산확률함수로 측정된다.

$$GM(x,w)=\left(\frac{1}{n^-}\sum^{n^-}SPE(x^-,w)\times\frac{1}{n^+}\sum^{n^+}SEN(x^+,w)\right)^{1/2} \tag{1}$$

$$SPE(x^-,w)=SPE(f(x^-,w))=\begin{cases}1 & \text{if } f(x^-,w)=w^Tx^-\le T)\\ 0 & Otherwise\end{cases}$$

$$SEN(x^+,w)=SEN(f(x^+,w))=\begin{cases}1 & \text{if } f(x^+,w)=w^Tx^+>T)\\ 0 & Otherwise\end{cases}$$

여기에서 임계점 T는 TPR과 FPR의 상충관계를 기초로 계산되며 TPR과 FPR의 차이가 최대화 되는 지점 ($\max(TPR-FPR)$)에서 최적의 임계점이 설정된다. 최적의 임계점에서 특이도와 민감도는 가장 유사한 값을 가지게 되며 GM은 극대화된다. 식 (1)에서 특이도($SPE(x^-,w)$)와 민감도($SEN(x^+,w)$)는 0 또는 1의 값을 가지는 비평활 손실함수(non-smoothed loss function)로서 미분이 불가능하거나 미분을 하더라도 0의 값을 가지기 때문에 경사하강법과 같은 최적화 기법의 적용이 불가능하다.

식 (1)을 연속확률함수로 변환하게 되면 GM의 기대값 $E(GM(x,w))$은 식 (2)와 같이 계산할 수 있다. 식 (2)에서 $P_{X,Y}(x,y)$는 확률변수 (X,Y)의 결합확률(joint probability)로서 결합분포 $P_{X,Y}(x,y)$의 사전적 분포가 알려져 있는 경우 GM의 기대값을 측정할 수 있지만, 분포가 알려지지 않은 경우 $E(GM(x,w))$의 계산이 불가능하게 된다.

$$E(GM(x,w))=\left(\int SPE(x^-,w)\times\int SEN(x^+,w)\right)^{1/2} \tag{2}$$

$$SPE(x^-,w)=P(w^Tx^-\le T)P_{X,Y}(x,y)dP(x,y)$$

$$SEN(x^+,w)=P(w^Tx^+>T)P_{X,Y}(x,y)dP(x,y)$$

본 연구에서는 데이터가 충분하게 확보되는 경우 표본의 분포는 정규분포를 따른다는 중심극한정리를 활용하여 결합확률 $P_{X,Y}(x,y)$가 다변량 정규분포를 따른다는 가정을 도입함으로써 GM의 기대값을 측정하고자 한다. 다변량 정규분포 가정에 따라 다수 범주 확률변수 X^-와 소수 범주 확

률변수 X^+에 대한 기저분류자의 선형 결합으로 산출되는 최종 분류자의 예측값(W^TX)도 식 (3)와 같이 다변량 정규분포로 정의된다. 여기에서, μ^-와 μ^+는 기저분류자에서 산출된 다수 범주와 소수 범주 관측치들의 예측 결과값의 평균벡터(mean vector)이며, Σ^-와 Σ^+는 다수 범주와 소수 범주 관측치들의 예측 결과값의 공분산 행렬(covariance matrix)을 의미한다.

$$X^- \sim N(\mu^-,\Sigma^-) \rightarrow W^TX^- \sim N(\omega^T\mu^-,\omega^T\Sigma^-\omega) \sim N(\mu_{Z^-},\sigma^2_{Z^-})$$
$$X^+ \sim N(\mu^+,\Sigma^+) \rightarrow W^TX^+ \sim N(\omega^T\mu^+,\omega^T\Sigma^+\omega) \sim N(\mu_{Z^+},\sigma^2_{Z^+}) \quad (3)$$

식 (2)에 자연로그를 취하여 식 (4)와 같이 변형하고 미분을 하게 되면 식 (5)와 같은 $\nabla GM(x,\omega)$를 정의할 수 있다.

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

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

표준정규분포의 $CDF(\Phi)$를 이용하여 $SPE(x^-,w)$와 $SEN(x^+,w)$는 식 (6) 및 (7)과 같이 정의된다. 여기에서 $\mu_{Z^-} = \omega^T\mu^-$, $\sigma_{Z^-} = \sqrt{\omega^T\Sigma^-\omega}$, $\mu_{Z^+} = \omega^T\mu^+$ 및 $\sigma_{Z^+} = \sqrt{\omega^T\Sigma^+\omega}$ 이다.

$$SPE(x^-,w) = P(W^TX^- \le T) = \Phi\left(\frac{\mu_{Z^-} - T}{\sigma_{Z^-}} \le 0\right) \quad (6)$$

$$SEN(x^+,w) = P(W^TX^+ > T) = (1 - P(W^TX^+ \le T)) = 1 - \Phi\left(\frac{\mu_{Z^+} - T}{\sigma_{Z^+}} \le 0\right) \quad (7)$$

식 (6)과 (7)에서 임계점 T에서 $SPE(x^-,w)$와 $SEN(x^+,w)$의 CDF는 각각 식 (8)과 식 (9)로 정의된다.

$$\Phi(SPE(x^-,w)) = \int_{-\infty}^{W^TX^- = T} \frac{1}{\sqrt{2\Pi}}\exp\left(-\frac{1}{2}\left(\frac{\omega^T\mu^- - T}{\sqrt{\omega^T\Sigma^-\omega}}\right)^2\right)dw \quad (8)$$

$$\Phi(SEN(x^+,w)) = 1 - \int_{-\infty}^{W^TX^+ = T} \frac{1}{\sqrt{2\Pi}}\exp\left(-\frac{1}{2}\left(\frac{\omega^T\mu^+ - T}{\sqrt{\omega^T\Sigma^+\omega}}\right)^2\right)dw \quad (9)$$

$\nabla(SPE(x^-,w))$와 $\nabla(SEN(x^+,w))$은 $\Phi(SPE(x^-,w))$및 $\Phi(SEN(x^+,w))$에 대한 미분한 결과로서 식 (10) 및 (11)과 같이 정의된다.

$$\nabla SPE(x^-,w) = \frac{1}{\sqrt{2\Pi}}\exp\left(-\frac{1}{2}\left(\frac{\mu_{Z^-} - T}{\sigma_{Z^-}}\right)^2\right)\left(\frac{T-\mu_{Z^-}}{\sigma_{Z^-}}\right)\left(\frac{\sigma_{Z^-}\mu_{Z^-} - \left(\frac{\mu_{Z^-} - T}{\sigma_{Z^-}}\right)\Sigma^-\omega}{(\sigma_{Z^-})^2}\right) \quad (10)$$

$$\nabla SEN(x^+,w) = -\frac{1}{\sqrt{2\Pi}}\exp\left(-\frac{1}{2}\left(\frac{\mu_{Z^+} - T}{\sigma_{Z^+}}\right)^2\right)\left(\frac{T-\mu_{Z^+}}{\sigma_{Z^+}}\right)\left(\frac{\sigma_{Z^+}\mu_{Z^+} - \left(\frac{\mu_{Z^+} - T}{\sigma_{Z^+}}\right)\Sigma^+\omega}{(\sigma_{Z^+})^2}\right) \quad (11)$$

식 (10)의 $\nabla(SPE(x^-,w))$와 식 (11)의 $\nabla(SEN(x^+,w))$을 식 (5)에 대입하면, $\nabla GM(x,w)$은 식 (12)와 같이 정의된다.

$$\nabla GM(x,w) = \frac{1}{2}\left(\frac{1}{SPE(x^-,w)}\right)\frac{1}{\sqrt{2\Pi}}\exp\left(-\frac{1}{2}\left(\frac{\mu_{Z^-}-T}{\sigma_{Z^-}}\right)^2\right)\left(\frac{T-\mu_{Z^-}}{\sigma_{Z^-}}\right)\left(\frac{\sigma_{Z^-}\mu_{Z^-}-\left(\frac{\mu_{Z^-}-T}{\sigma_{Z^-}}\right)\Sigma^-\omega}{(\sigma_{Z^-})^2}\right) \quad (12)$$
$$-\frac{1}{2}\left(\frac{1}{SEN(x^+,w)}\right)\frac{1}{\sqrt{2\Pi}}\exp\left(-\frac{1}{2}\left(\frac{\mu_{Z^+}-T}{\sigma_{Z^+}}\right)^2\right)\left(\frac{T-\mu_{Z^+}}{\sigma_{Z^+}}\right)\left(\frac{\sigma_{Z^+}\mu_{Z^+}-\left(\frac{\mu_{Z^+}-T}{\sigma_{Z^+}}\right)\Sigma^+\omega}{(\sigma_{Z^+})^2}\right)$$

식 (12)에 대하여 가우시안 경사하강법을 적용하여 결합 가중치 집합($\triangle\omega_k$)이 탐색되며 학습률(learning rate) β와 결합하여 $w_k^{new} = (w_k^{old} - \beta \bullet \triangle w_k)$와 같이 새로운 결합가중치($w_k^{new}$)을 생성하게 된다. 결합가중치는 정규화 과정($w_k^{new}/\sum_{k=1}^{K} w_k^{new}$)을 거쳐 후행 학습의 가중치로 활용된다. 가중치 탐색작업은 종료조건에 도달할 때까지 반복적으로 수행되며, 최종적으로 $GM(x,w)$을 극대화하는 최적의 결합가중치(w^*)가 탐색된다.

4. 연구방법론

4.1. 자료 수집

본 연구에서는 금융 분야의 범주 불균형 분류 문제로서 기업부도, 카드연체 및 카드사기 문제를 선정했으며, 수집된 데이터에 대한 개요는 Table 2에 요약되어 있다.

Table 2. The Summary of Data

Data	Independent Vars.	Minority	Majority	IR
Bankruptcy	7	500	10,000	1:20
Card Insolvency	23	6,636	23,364	1:3.52
Card Fraud	9	492	284,314	1:577.8

4.1.1 기업 부도 데이터(Bankruptcy Data)

기업 부도 예측은 비즈니스 분야의 대표적인 범주 불균형 사례로 본 연구의 불균형 비율과 유사하게 국내의 경우 10년 평균 장기부도율은 3~5% 수준으로 심각한 범주 불균형 문제를 내포하고 있다. 본 자료는 국내 기업을 대상으로 한 2015~2018년 실재 기업 자료로서 부도 예측을 위한 설명변수로서 총자산경상이익률, EBITDA/이자비용, 자기자본비율, 이익잉여금/총자산, 현금비율, 재고자산회전율 및 총자산 등 총 7개 재무비율을 활용하였다.

4.1.2 카드 연체 데이터(Card Insolvency Data)

금융사의 신용리스크 관리의 핵심적인 자료로서 본 연구의 카드 연체 데이터는 UCI Machine Learning Repository(http://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients)에서 수집된 자료로서 대만의 실제 데이터이다. 모형에 투입된 설명변수로 신용수준, 성별, 나이, 혼인여부, 교육수준,

과거 카드결제내역(정상지불, 연체 등), 과거 카드 사용금액 등 총 23개 변수가 사용되었다(Yeh, Lien, 2009).

4.1.3 카드 사기 데이터(Card Fraud Data)

신용 카드 거래의 증가와 더불어 카드 사기거래를 자동 탐지하기 위한 다양한 연구가 진행되고 있다. 본 연구의 데이터는 2013년 9월 유럽에서 거래된 284,807건의 카드 거래 데이터로 카드 사기 거래 빈도는 0.172%(492건)이다. 데이터는 머신러닝 알고리즘 경진 대회 사이트 Kaggle (https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud)에서 수집하였다. 설명변수는 거래 정보에 대한 28개의 변수를 주성분분석(Principal Component Analysis)를 통해 차원 축소된 9개 변수로 구성되어 있다(Dal Pozzolo et al., 2014).

5. 연구 결과

5.1. 모형 성과 분석

본 연구에 사용된 모든 부스팅 학습의 기저분류자로 의사결정나무를 사용하였다. 기저분류자의 최대 depth는 2이며, 기저분류자는 최대 20개까지 생성되는 것으로 설정하였다.

Table 3은 수집된 금융 데이터에 대한 부스팅 알고리즘과 GMBoost 모형의 성과를 검증한 결과를 제시하고 있다. 모형의 성과 검증를 위하여 10-fold 교차타당성 분석을 3회 실시하여 총 30회의 검증을 시행하였으며 성과지표로서 30회 시행에 대한 특이도(SPE), 민감도(SEN), 정확도(ACC), AUC, GM의 평균값을 사용하였다. Table 3의 모든 데이터에서 학습표본과 검증표본의 ACC, AUC 및 GM의 차이가 크지 않은 것으로 확인되어 과적합(overfitting)의 위험은 없는 것으로 판단된다.

30회 교차 타당성 분석을 실시한 주요 결과는 다음과 같다. 첫째, 세 가지 데이터 모두에서 기존 알고리즘(AdaBoost, GBM, XGBoost)은 GMBoost와 비교하여 높은 정확도(ACC)를 가진 것으로 확인되었다. 기업부도 데이터의 검증표본에서 기존의 부스팅 알고리즘은 약 96%의 정확도를 가지는 반면, GMBoost의 정확도는 79%로 약 17% 낮은 것으로 확인되었다. 카드연체 데이터에서 GMBoost는 기존 부스팅 알고리즘의 정확도 82%와 비교하여 약 12% 낮은 70%의 정확도를 가지는 것으로 나타났다. 카드사기 데이터에서는 GMBoost는 기존 알고리즘 99%에 비해 약 4% 낮은 95%의 정확도를 가지는 것으로 이러한 결과는 기존의 부스팅 알고리즘이 예측오차의 최소화를 목적함수로 설정하는 문제에 기인하는 것으로 범주 불균형 문제에서 정확도는 소수 범주의 낮은 민감도는 무시하고 다수 범주의 특이도에 의존하여 지나치게 높게 나타나기 때문이다. 예로 기업부도 데이터에서 기존 알고리즘의 다수 범주에 대한 특이도는 99%로 이러한 특이도에 의존하여 정확도는 96%로 높게 나타나고 있지만, 정작 소수 범주의 민감도는 5%~11%로 소수 범주에 대해서는 매우 낮은 분류 성과를 보이고 있다. 이와 같이 특이도와 민감도를 균형 있게 반영하지 못하는 문제는 카드연체(특이도 95~96%, 민감도 32%~37%) 및 카드사기(특이도 99%, 민감도 68~75%) 문제에서도 동일하게 발생하고 있다. 이러한 결과는 기존 부스팅 알고리즘을 범주 불균형 문제가 내재된 금융 비즈니스 문제 해결에 직접적으로 적용하는 경우 소수 범주에 대한 분류 성과가 심각하게 저하되어 분류 모형으로서의 의미를 상실할 수 있음을 보여주는 증거로서 예측오차 최소화에 초점을 맞

춘 기존 부스팅 알고리즘을 포함한 대부분의 분류 알고리즘이 범주 불균형 문제를 보다 효과적으로 해결하기 위해서는 산술평균 개념에 기초한 예측오차 극소화라는 목적함수가 변경될 필요가 있음을 보여주는 증거로도 해석된다.

둘째, 세 가지 불균형 데이터 모두에서 기존 알고리즘은 다수 범주의 특이도와 소수 범주의 민감도 차이가 크게 나타나고 있는 반면, GMBoost의 경우에는 특이도와 민감도가 유사한 수준으로 근접함으로써 특이도와 민감도의 차이(기업부도 5%, 카드연체 1%, 카드사기 7%)가 작게 나타나고 있으며, 결과적으로 GMBoost는 GM측면에서 기존 알고리즘과 비교하여 우수한 성과를 보이고 있다. 기업부도 데이터의 경우 기존 알고리즘의 GM은 약 21~33%로 나타나는 반면, GMBoost는 76%로 기존 알고리즘과 비교하여 약 43~55%의 성과 개선 효과를 보여주고 있다. 유사하게 카드 연체의 경우에도 기존 알고리즘의 GM(56~59%)과 비교하여 GMBoost는 70%로 약 10% 이상의 차이를 보여주고 있으며, 카드 사기의 경우에도 약 4~9%의 성과 차이를 보여주고 있다. AUC 측면에서도 GMBoost는 기존 알고리즘과 비교하여 우수한 분류 결과를 보여주고 있다. 이러한 결과는 GMBoost는 GM 최적화를 학습 목표로 설정하고 가우시안 경사하강법을 이용하여 성과지표인 GM을 직접적으로 최적화함으로써 금융 비즈니스의 불균형 데이터의 분류/예측 문제에 대하여 다수 범주와 소수 범주의 불균형 분포 및 비대칭적인 오분류 비용 문제를 효과적으로 해결할 수 있음을 보여주고 있다.

Table 3. Results of Cross-Validation for Performance Comparison

Panel A. Bankruptcy (IR 20:1)

Performance Metrics	Training Set				Test Set			
	AdaBoost	GBM	XGB	GMBoost	AdaBoost	GBM	XGB	GMBoost
SPE	0.99	0.99	0.99	0.80	0.99	0.99	0.99	0.79
SEN	0.14	0.07	0.07	0.79	0.11	0.05	0.06	0.74
ACC	0.96	0.96	0.96	0.80	0.95	0.95	0.95	0.79
AUC	0.57	0.53	0.53	0.80	0.55	0.53	0.53	0.76
GM	0.38	0.26	0.26	0.80	0.33	0.21	0.22	0.76

Panel B. Card Insolvency (IR 3.52:1)

Performance Metrics	Training Set				Test Set			
	AdaBoost	GBM	XGB	GMBoost	AdaBoost	GBM	XGB	GMBoost
SPE	0.95	0.96	0.95	0.72	0.95	0.96	0.95	0.70
SEN	0.37	0.33	0.36	0.70	0.36	0.32	0.35	0.71
ACC	0.82	0.82	0.82	0.71	0.82	0.82	0.82	0.70
AUC	0.66	0.64	0.66	0.71	0.65	0.64	0.65	0.70
GM	0.59	0.56	0.58	0.71	0.59	0.56	0.58	0.70

Panel C. Card Fraud (IR 577.8:1)

Performance Metrics	Training Set				Test Set			
	AdaBoost	GBM	XGB	GMBoost	AdaBoost	GBM	XGB	GMBoost
SPE	0.99	0.99	0.99	0.95	0.99	0.99	0.99	0.95
SEN	0.79	0.7	0.78	0.95	0.75	0.68	0.75	0.88
ACC	0.99	0.99	0.99	0.95	0.99	0.99	0.99	0.95
AUC	0.89	0.85	0.89	0.95	0.88	0.84	0.87	0.91
GM	0.89	0.85	0.88	0.95	0.87	0.82	0.87	0.91

Table 4. Results of t-test for the Performance Comparison Between Boosting Models and GMBoost

Performance Metrics	Bankruptcy			Card Insolvency			Card Fraud		
	AdaBoost	GBM	XGB	AdaBoost	GBM	XGB	AdaBoost	GBM	XGB
GM	32.41***	33.15***	31.73***	35.44***	41.09***	34.34***	2.23**	3.55***	2.34**
AUC	34.30***	42.10***	41.28***	19.32***	23.56***	18.71***	1.81*	3.20***	1.91*

***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively (two-tailed test).

Table 4는 GMBoost와 기존 알고리즘의 성과 차이에 대한 t-검정 결과를 제시하고 있다. 기업 부도 및 카드연체 데이터에서 GMBoost의 GM은 기존 알고리즘과 비교하여 1% 수준에서 유의한 성과차이를 보여주고 있으며, 카드 사기 데이터에서 GMBoost는 GBM과 1%, AdaBoost 및 XGBoost와 5% 수준에서 유의한 성과 차이가 나타나고 있다. AUC 측면에서도 기업 부도 데이터와 카드연체 데이터에서 GMBoost는 기존 알고리즘과 비교하여 1% 수준에서 유의한 성과 차이를 보여주고 있으며, 카드 사기 데이터에서는 GBM과 1% 수준에서 유의한 차이가 나타나고 있는 반면, AdaBoost 및 XGBoost와는 유의적인 성과차이가 발생하지 않는 것으로 분석되었다. 이러한 결과는 GMBoost가 데이터의 종류와 무관하게 분류 성과가 탁월하며, 데이터 불균형에도 불구하고 다수 범주와 소수 범주를 균형적으로 학습하는 보다 강건한 알고리즘이라는 것을 의미한다.

6. 결론

본 연구에서는 금융 분야의 범주 불균형 문제를 해결하고 앙상블 모형의 성과 개선을 위한 GMBoost 모형을 제안하였다. 부스팅 학습 알고리즘은 범주 균형에 대한 분류 및 예측 문제 해결에 탁월한 성과를 보여주었지만, 범주 불균형 문제가 내재된 금융 문제의 해결에는 상당한 한계점을 보여주었다. 이러한 한계는 부스팅 모형이 예측 오차 극소화를 목적함수로 설정하고 있는 점에서 기인하는 문제로서 이러한 문제에 대응하여 GMBoost는 GM 극대화를 목적함수로 설정하고, 다변량 정규분포의 가정을 도입하여 미분 가능한 함수로 치환한 후, 이를 가우시안 경사하강법을 이용하여 최적화함으로써 범주 불균형 문제의 분류 성과를 극대화할 수 있다. 금융 분야의 범주 불균형 문제에 적용하여 GMBoost의 분류 성과를 검증한 결과, GMBoost는 AdaBoost, GBM 및 XGBoost에 비교하여 특이도와 민감도 차이를 감소시켜 범주 간 균형적인 학습을 진행할 수 있는 장점을 가지며, 다양한 불균형 데이터에서 기존 알고리즘의 성과를 개선할 수 있음을 실증하였다.

본 연구는 예측 오차 극소화를 목적함수로 설정하고 있는 기존 알고리즘의 문제점에 대하여 성과지표를 직접적으로 최적화함으로써 예측 성과의 개선을 도모할 수 있다는 이론적 토대와 실증자료를 제시함으로써 인공지능/머신러닝 학습의 새로운 연구 방향을 제시하고 있다는 공헌점이 있다. 그러나 이러한 공헌점에도 불구하고 다음과 같은 한계점을 가지며, 이러한 한계점을 개선하기 위한 향후 연구 방향을 제시하고자 한다.

첫째, 본 연구에서는 범주 불균형 문제에 따른 성과지표로서 GM 최적화에 기반한 수정 부스팅 기법으로 GMBoost의 성과를 검증하였다. 그러나 범주 균형 문제에서 정확도를 목적함수로 채택한 부스팅 알고리즘이나 범주 불균형 문제에서 또 다른 성과지표인 AUC와 같은 다른 성과지표를 목적함수로 채택한 부스팅 알고리즘의 최적화 방안을 고려하지 못하였다. 따라서 향후 연구에서는 정확도 및 AUC 최적화 기반의 부스팅 기법에 대한 연구를 진행하고자 한다.

둘째, 본 연구에서는 언더 샘플링 또는 오버 샘플링과 같은 데이터 샘플링 기법과 결합된 부스팅 학습모형에 대하여 최적화된 부스팅 기법이 성과개선에 공헌할 수 있는지를 검토하지 못하였다. 데이터 샘플링을 통하여 범주 불균형 데이터에 대한 성과가 개선되고 있음을 보고한 연구 결과는 다수 존재하지만, 데이터 샘플링 기법은 범주 불균형 문제를 완화할 수 있지만, 궁극적으로 성과의 최적화를 보장하지 않는다. 본 연구에서 제안된 최적화 부스팅 기법은 데이터 샘플링과의 결합이 가능하기 때문에 범주 균형 및 범주 불균형 문제에 보다 효과적으로 적용할 수 있을 것으로 판단된다. 후속 연구에서는 범주 불균형 문제 해결을 위하여 제안된 다양한 알고리즘과의 결합 방안에 대한 연구를 진행하고자 한다.

References

Barboza, F., Kimura, H., Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, 83, 405-417. DOI: https://doi.org/10.1016/j.eswa.2017.04.006

Chen, T., Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd Association for Computing Machinery's Special Interest Group on Knowledge Discovery and Data Mining International Conference on Knowledge Discovery and Data Mining*, 785-794. DOI: https://doi.org/10.1145/2939672.2939785

Cho, Y., Cho, D., Choi, B. (2022). Applications of the classification algorithm for unbalanced time series data: Focusing on the corporate default model, *Journal of The Korean Data Analysis Society*, 24(2), 639-651. (in Korean) DOI: https://doi.org/10.37727/jkdas.2022.24.2.639

Dal Pozzolo, A., Caelen, O., le Borgne, Y.-A., Waterschoot, S., Bontempi, G. (2014). Learned lessons in credit card fraud detection from a practitioner perspective. *Expert Systems with Applications*, 41(10), 4915-4928. DOI: https://doi.org/10.1016/j.eswa.2014.02.026

Davis, J., Goadrich, M. (2006), The relationship between Precision-Recall and ROC curves, *Proceedings of the 23rd International Conference on Machine Learning - ICML* '06, 2, 33-240.

Du, J., Vong, C. M., Pun, C. M., Wong, P. K., Ip, W. F. (2017), Post-boosting of classification boundary for imbalanced data using geometric mean, *Neural Networks*, 96, 101-114. DOI: https://doi.org/10.1016/j.neunet.2017.09.004

Fawcett, T. (2006) An introduction to ROC analysis, *Pattern Recognition Letters*, 27(8), 861-874. DOI: https://doi.org/10.1016/j.patrec.2005.10.010

Freund, Y., Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1), 119-139. DOI: https://doi.org/10.1006/jcss.1997.1504

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 1189-1232. DOI: https://www.jstor.org/stable/2699986

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F. (2012), A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches, *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, 42(4), 463-484 DOI: https://doi.org/10.1109/TSMCC.2011.2161285

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73, 220–239. DOI: https://doi.org/10.1016/j.eswa.2016.12.035

He, H., Garcia, E. A. (2009), Learning from imbalanced data, *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263-1284. DOI: https://doi.org/10.1109/TKDE.2008.239.

Kim, M.-J., Kang, D.-K., Kim, H. B. (2015). Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction. *Expert Systems with Applications*, 42(3), 1074-1082. DOI:https://doi.org/10.1016/j.eswa.2014.08.025

Kang, P., Cho, S. (2006), EUS SVMs: Ensemble of Under-Sampled SVMs for Data Imbalance Problems, ICONIP 2006: *Neural Information Processing*, 837-846. DOI: https://doi.org/10.1007/11893028_93

Kwon, Y., Park, S. Y. (2022) Survival prediction of hospitality businesses during the pandemic, *Journal of The Korean Data Analysis Society*, 24(5), 1791-1809 (in Korean) DOI: https://doi.org/10.37727/jkdas.2022.24.5.1791

Kubat, M., Holte, R., Matwin, S.(1997) Learning when negative examples abound, *European Conference on Machine Learning 1997*, 146-153.

Lee, B. E., Joo, Y. S., Jeon. H. J. (2017) Rare bankruptcy event prediction with missing data, *Journal of The Korean Data Analysis Society*, 19(1), 129-139 (in Korean) DOI: https://doi.org/10.37727/jkdas.2017.19.1.129

Le, T., Son, L. H., Vo, M. T., Lee, M. Y., Baik. S. W. (2018a), A Cluster-based Boosting Algorithm for Bankruptcy Prediction in a Highly Imbalanced Dataset, *Symmetry*, 10(7) DOI:https://doi.org/10.3390/sym10070250

Le, T., Lee, M. Y., Park, J. R., Baik, S. W. (2018b), Oversampling Techniques for Bankruptcy Prediction: Novel Features from a Transaction Dataset, *Symmetry*, 10(4) DOI: https://doi.org/10.3390/sym10040079

Lin, W. C., Tsai, C. F., Hu, Y. H., Jhang, J. S. (2017), Clustering-based Undersampling in Class Imbalanced Data, *Information Sciences*, 409-410, 17-26. DOI: https://doi.org/10.1016/j.ins.2017.05.008

Mellor, A., Boukir, S., Haywood, A., Jones, S. (2015), Exploring issues of training data imbalance and mislabelling on random forest performance for large area land cover classification using the ensemble margin, *International Society for Photogrammetry and Remote Sensing*, 105, 155-168 DOI: https://doi.org/10.1016/j.isprsjprs.2015.03.014

Nanni, L., Lumini, A. (2019), An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring, *Expert Systems with Applications*, 36(2), PART 2, 3028-3033 DOI: https://doi.org/10.1016/j.eswa.2008.01.018

Park, Y., Lee, S. Y., Choi, B. (2009). A study of effective default forecasting model development for small and medium sized enterprises, *Journal of The Korean Data Analysis Society*, 11(3), 1363-1375 (in Korean)

UlagaPriya, K., Pushpa, S. (2021), A comprehensive study on ensemble-based imbalanced data classification methods for bankruptcy data, *Institute of Electrical and Electronics Engineers 6th international Conference on Inventive Computa- tion Technologies*, 800-804. DOI: https://doi.org/10.1109/ICICT50816.2021.9358744.

Yeh, I.-C., Lien, C. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2), 2473-2480. DOI: https://doi.org/10.1016/j.eswa.2007.12.020

Zieba, M., Tomczak, S. K., Tomczak, J. M. (2016), Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction, *Expert Systems with Applications*, 58, 93-101 DOI: https://doi.org/10.1016/j.eswa.2016.04.001

Performance optimization-based boosting algorithm for resolving class imbalance problems in finance

Myoung-Jong Kim[1], *Jae-Hyun Ahn*[2], *and Yun-Hu Kim*[3]

Abstract

In this paper, we propose a GMBoost (Geometric Mean-based Boosting) with a direct performance optimization technique to resolve class imbalance problem in financial field, such as bankruptcy, card insolvency, and card fraud. The conventional boosting models including AdaBoost, GBM, and XGBoost are used as benchmarking models for performance comparison. The main findings of 30 rounds of cross validations are as follows. First, the conventional boosting models largely depend on the specificity of majority, but ignore the sensitivity of minority. On the contrary, GMBoost proceeds with balanced learning that simultaneously considers the specificity and the sensitivity at the same time. Second, GMBoost outperforms the conventional boosting algorithms in terms of GM and AUC, and the results of the t-test for GM and AUC also showed that the performance of GMBoost is significantly different from those of the benchmarking models.

Keywords : class imbalance, GMBoost, direct optimization of performance metrics, Gaussian gradient descent method.

[1](Corresponding Author) Professor, Department of Business, Pusan National University, 2, Busandaehak-ro 63 beon-gil, Geumjeong-gu, Busan, Zip Code 46241, Republic of Korea. E-mail: mjongkim@pusan.ac.kr
[2]Undergraduate student, Department of Business, Pusan National University, 2, Busandaehak-ro 63 beon-gil, Geumjeong-gu, Busan, Zip Code 46241, Republic of Korea. E-mail: bonjour8084@pusan.ac.kr
[3]Undergraduate student, Department of Computer Science, Purdue University - Main Campus, 610 Purdue Mall, West Lafayette, IN 47907, USA. E-mail: kim3438@purdue.edu
[Received 17 February 2023; Revised 15 March 2023; Accepted 18 March 2023]