Information Systems Review
Vol. 24, No. 1 February 2022 http://dx.doi.org/10.14329/isr.2022.24.1.039

기업부도 예측 앙상블 모형의 최적화

The Optimization of Ensembles for Bankruptcy Prediction

김 명 종 (Myoung Jong Kim) 부산대학교 경영대학 경영학과 교수, 교신저자
윤 우 섭 (Woo Seob Yun) 부산대학교 경영대학 경영학과 학사과정

요 약

본 연구에서는 범주 불균형 문제가 내재된 기업부도 예측 AdaBoost 앙상블 모형의 성과를 개선하기 위하여 GMOPTBoost 알고리즘을 제안한다. AdaBoost 알고리즘은 오분류 표본에 대하여 강건한 학습기회를 제공한다는 장점이 있지만, 산술평균 정확도에 기반하기 때문에 범주 불균형 문제를 효과적으로 해결하지 못한다는 한계점이 존재한다. GMOPTBoost는 가우시안 경사하강법(Gaussian gradient descent)을 적용하여 기하평균 정확도를 최적화하고 범주 불균형 문제를 효과적으로 해결할 수 있다는 장점이 있다. 본 연구에서는 첫째, 범주 불균형 문제가 예측 모형의 성과에 미치는 효과와 GMOPTBoost의 성과 개선 효과를 검증하기 위하여 5개의 범주 불균형 데이터를 구성하였으며, 둘째, 범주 균형 데이터에 대한 GMOPTBoost의 성과 개선 효과를 검증하기 위하여 데이터 샘플링 기법을 통하여 구성된 균형 데이터를 구성하였다. 30회의 교차타당성 분석의 주요 결과는 다음과 같다. 첫째, 범주 불균형 문제는 예측 성과에 부정적인 영향을 미친다. 둘째, GMOPTBoost는 불균형 데이터에 적용된 AdaBoost의 성과를 유의적으로 개선시키는 긍정적인 효과를 제공한다. 셋째, 데이터 샘플링 기법은 성과 개선에 긍정적인 영향을 미친다. 마지막으로 데이터 샘플링 기법을 적용한 범주 균형 데이터에서도 GMOPTBoost는 유의적인 성과 개선에 기여한다.

키워드 : 부실예측, 범주 불균형, 범주 균형, 데이터 샘플링, 기하평균 정확도, GMOPTBoost

Ⅰ. 서 론

기업의 부도는 기업, 기업의 이해관계자와 더불어 국민경제에 심각한 손실을 초래하는 중대한 문제로서 기업부도 예측 모형의 성과 개선은 선행연구의 주요 관심사가 되어왔다. 기업부도와 같은 예측 모형의 성과 개선에 있어서 대표적인 난제는 범주 불균형(class imbalance) 문제이다. 대부분의 예측 모형은 범주 간 데이터의 균형 분포를 가정하지만, 현실 문제에서 대부분의 표본이 특정 범주에 편중되어 분포하는 범주 불균형 문제가 자주 관찰된다. 비즈니스 분야의 대표적인 범주 불균형 사례로서 카드 사기 탐지(Somasundaram and Reddy, 2019), 회사채등급평가(Kwon *et al.*, 1997) 및 기업부실예측문제(Kim *et al.*, 2015) 등을 들 수 있다.

범주 불균형 문제는 소수 범주와 다수 범주에 속하는 표본 수에 현저한 차이가 발생하는 왜곡된

† 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터지원사업의 연구결과로 수행되었음(IITP-2020-0-01797).

분포(skewed distribution) 문제로 정의된다. 기업부도 예측 표본도 다수 범주(정상 기업)와 소수 범주(부도 기업) 사이에 왜곡된 분포를 보이는데 무디스(Moody's) 등 전문신용평가기관의 자료에 따르면, 국내 외부감사 법인의 예상 10년 장기 평균 기업 부도율은 약 3~5% 수준으로 심각한 범주 불균형이 존재하고 있음을 보고하고 있다. 범주 불균형 문제는 예측 모형의 성과 개선에 부정적인 영향을 미친다. 범주 불균형이 존재할 경우 기업부도 예측 모형은 정상 기업의 정확도에 의존하여 모형의 성과는 개선되지만, 경제적으로 중요한 의미를 가지는 부도 기업에 대한 정확도는 급격하게 하락하게 된다. 특히, 범주 불균형이 심화되는 경우 분류 모형은 모든 소수 범주 표본을 다수 범주로 예측하기 때문에 예측 모형의 유용성을 상실하게 된다. 범주 불균형 문제가 예측 모형의 성과 개선을 저해하는 주요한 원인으로 인식되면서 데이터 샘플링(data sampling), 알고리즘 기법(algorithm-level techniques) 및 앙상블 학습(ensemble learning) 등 범주 불균형 문제를 해결하기 위한 다양한 기법들이 제안되어 왔다.

범주 불균형의 성과 저하 문제와 더불어 범주 불균형에서 파생되는 문제로서 성과측정치의 유효성 문제가 제기되고 있다. 산술평균 정확도(arithmetic accuracy or predictive accuracy, 이하 정확도)는 분류 및 예측 모형의 성과 평가에 가장 보편적으로 활용되어 왔던 성과 측정치이지만, 범주 불균형 상황에서 정확도는 다수 범주의 정확도(특이도)에 편향되며, 소수 범주의 정확도(민감도)는 고려하지 못하는 문제점으로 인하여 성과 측정치로서의 유효성을 상실하게 된다(Barua *et al.*, 2014). 최근에는 범주 불균형 문제에서 소수 범주 정확도(민감도)와 다수 범주 정확도(특이도)를 균형적으로 고려할 수 있는 대체적 성과측정치로 기하평균 정확도(geometric mean, 이하 GM)이나 AUROC(Area under the ROC curve)와 같은 성과측정치들이 활용되고 있다.

본 연구에서는 범주 불균형 문제와 관련된 최근 연구에서 가장 성공적인 기법으로 평가되고 있는 AdaBoost 기반의 앙상블 모형의 성과 개선문제를 다루고자 한다. AdaBoost 앙상블 학습은 오분류 표본(misclassified sample)에 대한 학습을 강화하여 정확도를 개선한다는 장점에 기인하여 범주 불균형 문제에 활용되고 있지만, AdaBoost 앙상블 학습의 기본 알고리즘은 산술평균으로 측정된 오차(predictive error)와 정확도에 기초하기 때문에 소수 범주의 정확도를 개선하기 보다는 다수 범주의 정확도에 초점을 맞추어 학습을 진행한다. 선행연구에서는 AdaBoost 앙상블 학습이 범주 불균형 문제를 효과적으로 적용되기 위해서는 AdaBoost 알고리즘을 수정하거나 데이터 샘플링과 같은 다른 기법들과 결합하여 활용될 필요가 있음을 제시하고 있다.

본 연구에서는 범주 불균형 문제를 효과적으로 해결하기 위한 GM 최적화 기반의 부스팅 알고리즘인 GMOPTBoost(GM-based boosting algorithm)를 제안하고자 한다. GMOPTBoost는 AdaBoost 알고리즘의 수정 알고리즘으로 기업부도 예측 모형의 GM을 최적화하기 위하여 경사하강법(gradient descent method)을 활용한다. 본 연구에서는 두 가지 관점에서 GMOPTBoost의 성과 개선 효과를 검증하고자 한다. 첫째, 범주 불균형이 예측 모형의 성과에 미치는 영향과 GMOPTBoost가 상이한 불균형 비율(Imbalance Ratio, 이하 IR)로 구성된 데이터에 대하여 학습된 AdaBoost 앙상블 모형의 성과 개선에 공헌하는지를 검증하고자 한다. 둘째, 선행연구에서는 데이터 샘플링을 적용하여 균형 데이터(balanced data)를 구성함으로써 예측 모형의 성과를 개선할 수 있지만, 데이터의 균형 분포만으로 예측 성과가 최적화됨을 보장되지 않는다는 주장이 제기되고 있다 (Kuncheva, 2019). 이에 따라 본 연구에서는 GMOPTBoost가 균형 데이터에 대하여 학습된 AdaBoost 앙상블 모형의 성과를 추가적으로 개선하는지를 검증하고자 한다.

첫 번째 검증 목적을 위하여 2015~2018년의 4년 동안 500개의 부도 기업과 10,000개 기업-년도별

건전 기업 표본을 수집하여 범주 불균형 비율에 따라 5개의 하위 표본을 구성하였다. 두 번째 검증 목적을 위하여 4가지 데이터 샘플링 기법을 활용하여 균형 데이터를 구성하였다. 본 연구에서는 선행연구에서 유용하게 활용되는 30개의 재무비율을 7개의 재무비율 군으로 분류하여 AUROC (Area under Receiver Operating Characteristic)가 가장 우수한 재무비율 7개를 설명 변수로 선정하였다. 앙상블 모형에 사용되는 기저 분류자(Base classifier)로 3층 인공신경망(3-layered neural networks)을 구현하였으며, AdaBoost 앙상블 모형과 GMOPTBoost 앙상블 모형의 성과를 측정하기 위하여 총 30회의 교차 타당성(cross validation) 분석을 수행하였다. 주요 검증 결과는 다음과 같다. 첫째, 범주 불균형 문제는 모형 성과를 저하시킨다. 둘째, GMOPTBoost는 불균형 데이터에 대해 훈련된 AdaBoost 앙상블의 유의적인 성과 개선에 기여한다. 셋째, 데이터 샘플링이 적용된 균형 데이터에 대해서도 GMOPTBoost는 유의적인 성과 개선 효과를 제공한다.

본 연구는 다음과 같이 구성되어 있다. 제Ⅱ장에서는 기업부도 예측 문제에 내재된 범주 불균형의 문제점과 관련하여 예측 모형의 성과 저하 문제와 성과측정치의 유효성 문제 및 이러한 문제점들에 대한 해결방안에 대하여 기술한다. 제Ⅲ장에서는 본 연구에 활용되는 AdaBoost와 GMOPTBoost의 학습 알고리즘에 대하여 설명한다. 제Ⅳ장에서는 실증분석에 사용된 표본의 수집, 변수의 선정 및 모형 설계 과정에 대하여 소개한다. 제Ⅴ장은 실증 연구의 결과를 제시하고 마지막으로 제Ⅵ장은 본 연구의 결론 및 향후 연구 방향에 대하여 제시하고자 한다.

Ⅱ. 범주 불균형 문제

2.1 기업 부실 예측의 범주 불균형 문제

기업의 부실화로 인한 부정적인 영향을 사전적으로 예방하기 위하여 단별량 모델(Beaver, 1996)을 시초로 다변량판별분석(Altman, 1968), 의사결정트리(Messier *et al.*, 1998), 인공신경망(Odom and Sharda, 1990) 및 SVM(Support Vector Machine) (Shin *et al.*, 2005) 등의 다양한 통계 및 인공지능 기법들이 기업부도 예측 모형 개발에 적용되어 왔다. 2000년대 초반까지의 초기 선행 연구들은 다양한 기법을 적용한 기업부도 예측 모형 간의 성과 비교에 초점을 맞추어 진행되어 왔다. 이러한 선행연구의 주요 결론은 통계 모형과 비교하여 인공지능 모형의 성과가 우수하며, 단일 모형보다는 여러 모형을 결합한 하이브리드(hybrid) 모형 또는 앙상블(ensemble) 모형의 학습성과가 더욱 우수한 것으로 보고하고 있다(Kim and Kang, 2010; Odom and Sharda, 1990; Zhang *et al.*, 1999).

2000년대 초반 이후 연구들은 기업부도 예측 모형의 성과를 저해하는 요소로서 범주 불균형 문제 해결에 초점을 맞추어 진행되고 있다. 기업부도 예측 문제와 같은 예측 및 분류 문제에 적용된 대부분의 예측 모형은 범주 간 표본의 균형 분포를 기본적으로 가정하고 모형의 예측 오차를 낮추고 예측 정확성을 극대화하는 것을 목표로 한다. 그러나 범주 불균형 문제는 다양한 현실 문제에 존재하는 보편적인 특성으로 범주 불균형이 존재하는 경우 예측 모형은 다수 범주에 편향되어 다수 범주의 분류 정확성은 증가하는 반면, 소수 범주에 대한 분류 정확성이 급격하게 낮아진다. 특히, 범주 불균형이 증가할수록 다수 범주에 대한 예측 정확성은 증가하지만 소수 범주에 대한 예측 정확성은 급격하게 낮아져 예측 모형의 유용성을 상실하게 된다.

범주 불균형 문제 해결에 가장 효과적인 해결 기법으로 주목을 받고 있는 기법은 데이터 샘플링 기법(data sampling techniques)과 앙상블 학습(ensemble learning)이다. 데이터 샘플링은 크게 언더 샘플링(under-sampling)과 오버 샘플링(over-sampling)으로 분류된다. 언더 샘플링 기법은 다수 범주의 표본을 소수 범주 표본의 수만큼 축소하는

기법으로 다수 범주의 표본 중 임의적으로 선택된 표본을 제거하는 RUS(random under-sampling) 기법과 클러스터링(clustering) 기법을 적용하여 다수 범주의 표본을 군집(cluster)으로 할당한 후 cluster별로 다수 범주의 표본을 제거하는 CUS(cluster-based under-sampling) 기법이 대표적으로 활용되고 있다. 언더 샘플링은 다수 범주의 샘플을 제거하기 때문에 정보의 손실을 발생시킬 수 있다는 단점이 있지만, 이해하기 쉽고 구현이 편리하며 학습시간을 단축할 수 있다는 장점으로 인하여 다양한 분야에서 적용되고 있다(Lin *et al.*, 2009). 반면, 오버 샘플링은 소수 범주의 표본 수를 다수 범주의 표본 수만큼 추가하는 기법으로 임의적으로 선택된 소수 범주의 표본을 단순 복제하는 ROS(random over-sampling)와 k-neighborhood 기법을 적용하여 소수 범주의 표본을 생성하는 SMOTE(Synthetic minority oversampling technique) 등이 대표적으로 활용되고 있다. 오버 샘플링은 소수 범주의 표본을 확장하기 때문에 정보 손실의 위험은 없지만, 표본 수의 증가로 모형 구축 시간이 길어지며 과적합의 위험이 증가하는 문제가 발생할 수 있다(He and Garcia, 2009).

Zhou(2013)은 언더 샘플링과 오버 샘플링을 적용하여 기업 부실을 예측한 결과 오버 샘플링과 비교하여 언더 샘플링의 성과가 보다 우수함을 확인하였다. 국내 연구로서 김량형 등(2016)은 다양한 데이터마이닝 기법을 활용하여 SMOTE 기법의 효과를 검증한 결과 약 10% 이상의 예측 정확성을 개선하는 효과를 발견하였다. 안철휘, 안현철(2018)은 ROS, SMOTE 및 RUS 등의 다양한 샘플링 기법의 효과를 검증한 결과 ROS 모형의 성과가 더욱 강건함을 확인하였다. Kim *et al.*(2015)은 RUS 인공신경망과 비교하여 CUS 인공신경망이 부실 측에 효과적임을 검증하였다.

앙상블 학습의 대표적인 기법으로서 부스팅 알고리즘은 순차적 분류자 생성기법으로 이전 분류자에서 오분류된 관측치에 높은 가중치를 부여하여 오분류 표본에 대하여 강건한 학습을 진행할 수 있다는 장점으로 인하여 불균형 문제 해결에 효과적인 것으로 보고되고 있다(Mellor *et al.*, 2015). Nanni and Lumini(2009)는 호주, 독일 일본 기업을 대상으로 부도 예측 및 신용평가 앙상블 모형을 구성하여 개별 분류자와의 예측 성과와 비교한 결과 앙상블 학습의 성과가 더욱 우수함을 발견하였다. Kim and Upneja(2014)는 기업부도 예측 문제에 AdaBoost를 성공적으로 적용하였다. Zieba *et al.*(2016)은 XGBoost(Extreme Gradient Boost)를 폴란드 기업의 부도예측 문제에 적용한 결과 개별 분류 모형과 비교하여 우수한 예측 성과를 가지고 있음을 확인하였다. Barboza *et al.*(2017)은 단일 분류기와 비교하여 배깅(bagging), 및 부스팅(boosting) 등의 앙상블 학습이 더욱 정확한 예측 성과를 가지고 있음을 확인하였다.

최근에는 범주 불균형 문제의 해결 대안으로 데이터 샘플링 기법과 앙상블 학습을 결합한 SMOTEBoost(Chawla *et al.*, 2003) 및 RUSBoost (Seiffert *et al.*, 2008) 등의 부스팅 알고리즘이 제안되었으며, 범주 불균형 문제에 대하여 효과적인 결과를 보여주었다. Kim *et al.*(2015)은 산술평균 정확도에 초점을 맞추고 있는 AdaBoost의 학습 알고리즘을 수정하여 기하평균 정확도에 초점을 맞춘 GMBoost를 제안하였다. 기업부도 예측 문제를 대상으로 GMBoost는 범주 불균형 및 균형 기업부도 데이터 모두에서 가장 성능이 높은 것으로 나타났다. Le *et al.*(2018b)은 부실 예측을 위해 클러스터 기반 부스팅 알고리즘인 CBoost를 제안하였다. 실험 결과 제안된 CBoost는 GMBoost(Kim *et al.*, 2015), 클러스터 기반의 oversampling 기법인 SMOTEENN 모형(Le *et al.*, 2018a), 클러스터 기반 under-sampling 모형(Lin *et al.*, 2017) 보다 성능이 우수함을 보였다. UlagaPriya and Pushpa(2021)은 다양한 샘플링 기법과 앙상블 기법을 결합하여 기업부도 예측문제에 적용한 결과 SMOTEBagging 및 SMTEBoost 모형이 RUSBoosting 및 RUSBagging 보다 기업부도 예측에서 더욱 강건한 모형임을 실증하였다.

이와 같이 범주 불균형 문제의 해결을 위하여 다양한 데이터 샘플링 기법과 앙상블 기법이 적용되어 왔지만, 이러한 기법의 적용만으로는 예측 성과의 최적화가 보장되지 않는다는 지적이 제기되고 있다. 즉, 데이터 샘플링 기법을 적용하여 범주 간 균형 분포를 확보하더라도 또는 앙상블 학습을 적용하여 오분류 관측치에 대한 학습을 강화하여 정확도가 개선된다 하더라도 이는 직접적으로 예측 성과의 최적화를 의미하지 않는다(Kim *et al.*, 2015; Kuncheva, 2019). 본 연구에서는 범주 불균형 및 범주 균형 표본에 대하여 학습된 예측 모형의 성과를 최적화하기 위한 기법으로서 GMOPTBoost 알고리즘을 제안하고 있다.

2.2 범주 불균형의 성과 측정치 문제

예측 모형의 성과측정치는 모형의 평가뿐만 아니라 모형의 학습과정을 유도하는 중요한 역할을 한다. <표 1>은 전형적인 이진 분류(binary classification) 모형에서 주요한 성과 측정치를 보여주고 있다.

정확도는 분류 모형의 성과평가에 보편적으로 사용되는 성과 측정치이지만, 기업부도 예측 문제와 같은 범주 불균형 표본에서 정확도는 정상 기업의 분류 정확성에 의존하여 모형의 정확도는 지속적으로 높아지지만, 정작 경제적으로 중요한 의미를 가지는 부도 기업에 대한 정확도가 심각하게 저하되기 때문에 성과 측정치로서의 적합성에 대한 의문이 제기되고 있다. 범주 불균형 데이터에서 정확도의 한계점을 보완하기 위한 대체적인 성과측정치로서 다수 범주의 정확도(특이도)와 소수 범주의 정확도(민감도)를 동시에 고려할 수 있는 AUROC 및 GM과 같은 대체적인 측정치가 이용되고 있다(Davis *et al.*, 2006; Fawcett, 2006; Weng *et al.*, 2008).

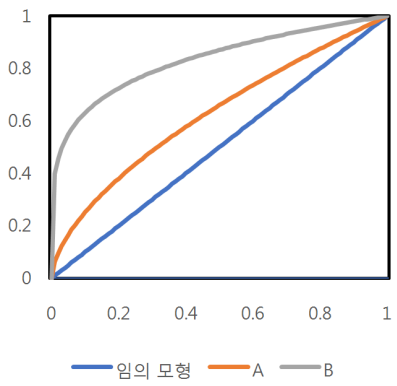
GM은 특이도와 민감도의 곱으로 측정하기 때문에 다수 범주와 소수 범주의 정확도를 동시에 고려한 성과 측정치이다. 최적화된 GM은 특이도와 민감도를 근사적으로 일치시키는 동시에 모형의 정확도를 개선하는 장점을 가진다.

ROC 곡선(Davis *et al.*, 2006)은 TPR과 FPR의 상충 관계(trade-off)를 그래프로 표현하며 AUROC는 ROC 곡선 아래 면적을 의미하며, ROC 곡선의 의미를 단일 측정치로 환산해주어 모형 간의 비교가 용이하다는 장점이 있다. ROC곡선의 사례로서 <그림 1>에서 임의 모형(random model)의 ROC 곡선은 정사각형의 대각선으로 표현되며, 임의 모형의 AUROC는 0.5로 범주 간 표본에 대한 실질적 분류 능력이 없음을 의미한다. 완벽한 모형(perfect model)의 ROC 곡선은 (0,0)-(0,1)-(1,1)을 연결한 정사격형으로 표현되며, 완벽 모형의 AUROC는 1이며 다수 범주와 소수 범주의 표본을 100% 완벽하게 분류한다. 일반적인 모형은 A, B와 같이 임의 모형과 완벽 모형 사이에서 결정된다. 모형의 AUROC는 1에 가까울수록 모형의 분류 능력이 우수한 것으로 해석되기 때문에, 모형 B는 모형 A에 비하여 분류 능력이 우수하다고 해석된다.

〈표 1〉 Confusion Matrix in Binary Classification

		예측 관측치	
		예측 부도	예측 건전
실제 관측치	실제 부도	True Positive (TP)	False Negative (FN)
	실제 건전	False Positive (FP)	True Negative (TN)

- 민감도(Sensitivity, TPR): TP / (TP + FN)
- 특이도(Specificity, TNR): TN / (FP + TN)
- 정확도(Arithmetic Accuracy or accuracy): (TP + FN) / (TP + FN + FP + TN)
- 기하평균 정확도(Geometric Mean: GM) = (민감도 특이도)$^{1/2}$



〈그림 1〉 ROC Curve

Ⅲ. 학습 알고리즘

3.1 AdaBoost 알고리즘

AdaBoost 알고리즘은 부스팅 알고리즘 중 가장 보편적으로 사용되는 알고리즘이다(Freund and Schapire, 1997; Schapire, 1990). AdaBoost 알고리즘은 기저 분류자를 순차적으로 생성하는 앙상블 학습으로 학습 표본에 대한 기저 분류자의 학습이 종료된 후 학습결과를 기반으로 기저 분류자의 학습성과로서 산술평균 오차에 기반한 오류율(e_k)을 계산하게 된다. 오류율을 기반으로 오분류 표본에 대하여 높은 가중치가 부여되고 정분류 표본은 낮은 가중치가 부여된다. 다음 차례로 생성된 기저 분류자에서는 오분류 학습 표본에 초점을 맞추어 학습을 진행하게 된다. 이러한 순차적 학습 과정은 기저 분류자의 오류율이 0.5를 초과하는 경우 종료된다.

최종적으로 k개의 기저 분류자가 생성되고 이를 가중 결합하여 최종 분류자의 최종 결과를 생성하게 된다. 이 때 기저 분류자의 결합 가중치는 각 개별 분류자의 산술평균 정확도 개념이 포함된 α_k를 활용하여 최종 분류자의 예측 값이 임계점보다 높으면 소수 범주로, 그렇지 않으면 다수 범주로 분류된다. 이러한 AdaBoost의 절차는 간략하게 <표 2>에 기술되어 있다.

<표 2>에 설명된 바와 같이, AdaBoost 알고리즘은 오분류 표본에 강화된 학습기회를 제공하는 장점이 있지만, 알고리즘 자체적으로 산술평균 기반의 오차 및 정확도 개념을 내포하고 있기 때문에 특이도와 민감도를 동시에 고려한 학습을 진행하지 못한다는 단점이 존재한다. 이러한 단점으로 인하여 소수 범주의 정확도를 개선하기 보다는 다수 범주의 정확도 개선에 초점을 맞추어 학습을 진행하게 된다. 결과적으로 특이도는 높아지지만, 민감도는 낮아져 범주 불균형 문제 해결에 한계점

〈표 2〉 AdaBoost algorithm

AdaBoost algorithm
Input: Training set $S=(x_i, y_i): i=1, \cdots, n;\ y_i \in (-1, +1)$
K: Number of base classifiers C_k: Base classifier
Output: the final classifier $C(x_i)=\begin{cases}-1, & \sum_{k=1}^{k} \alpha_k C_{k \le T} \\ +1, & \sum_{k=1}^{k} \alpha_k C_{k > T}\end{cases}$
1. Initialize $\omega_1(i)=\frac{1}{n}$ for $i=1, \cdots, n$
2. For k=1 to K
a. Calculate the error rate: $e_k=\frac{1}{n}\sum_{i=1}^{n} L(C_k(x_i), y_i),\ L(C_k(x_i), y_i)=\begin{cases}1, & C_k(x_i) \neq y_i \\ 0, & C_k(x_i)=y_i\end{cases}$
b. If e_k > 0.5 Then stop learning
c. Choose weight updating parameter: $a_k=\ln(((1-e_k))/e_k)$
d. Update sample weights: $\omega_{k+1}(i)=\omega_k(i)\exp(-\alpha_k Y_k C_k(X_i))$
e. Normalize $\omega_{k+1}(i)$ to be a proper distribution

을 가지게 된다. 이와 같은 한계점을 개선하기 위하여 AdaBoost 알고리즘을 수정한 새로운 알고리즘이 제안되기도 하며 데이터 샘플링 기법과 결합하여 활용되고 있다.

3.2 GMOPTBoost 알고리즘

GMOPTBoost 알고리즘은 범주 불균형 문제를 효과적으로 해결하기 위한 AdaBoost의 수정 알고리즘으로 GMOPTBoost 알고리즘의 학습 절차는 다음과 같다.

n개의 관측치를 학습한 기저 분류자의 학습 결과로서 $U=\{(x_i,\ y_i):i=1,\cdots,\ n\}$를 가정하자. 여기에서 $y_i\in(-1,\ +1)$는 범주이며 $x_i\in R^C$는 최종 분류자의 입력 벡터(입력변수)이며 C는 기저 분류자의 수이다. 확률변수 $(X,\ Y)$의 결합확률을 $P_{X,Y}(x,\ y)$라 하면 $(x_i,\ y_i)$는 결합확률 $P_{X,Y}(x,\ y)$의 i.i.d(independently identically distributed) 관측치이다. GMOPTBoost의 학습 목적은 GM을 최적화하는 선형분류함수 $f(w)=w^TX\rightarrow Y$를 찾는 것이다. x_i를 다수 범주 확률변수(X^-)의 관측치 x^-와 소수 범주 확률변수(X^+)의 관측치 x^+로 구분하게 되면, GM은 특이도 ($SPE(x^-,\ w)$)와 민감도 (($SEN(x^+,\ w)$))의 함수로 식 (1)과 같이 정의된다.

$$GM(x,\ w)=\sqrt{(SPE(x^-,\ w)\times SEN(x^+,\ w)} \quad (1)$$

GM의 기대값 $E(GM(X,\ w))$은 연속확률 함수의 누적분포함수(Cumulative Distribution Function, CDF)로 식 (2)와 같은 방식으로 계산할 수 있지만 결합분포 $P_{X,Y}(x,\ y)$의 분포가 정의되지 않았기 때문에 $E(GM(X,\ w))$의 계산은 불가능하게 된다.

$$E(GM(x,\ w))=\Big(\int SPE(x^-,\ w)P_{X,Y}(x,\ y)dP(x,\ y) \times\int SEN(x^+,\ w)P_{X,Y}(x,\ y)dP(x,\ y)\Big)^{1/2} \quad (2)$$

위험 최소화 원칙(risk minimization principal)에 따라 측정된 기하평균오류(geometric error) $GE(x,\ w)$의 근사치는(approximation)는 다음과 같이 계산된다.

$$GE(x,\ w)=1-GM(x,\ \omega)= 1-\left(\frac{1}{n^-}\Sigma^{n^-}SPE(x^-,\ w)\times\frac{1}{n^+}\Sigma^{n^+}SEN(x^+,\ w)\right)^{1/2} \quad (3)$$

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

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

여기에서 n^-와 n^+는 각각 다수 및 소수 범주의 사례 수이며, T는 임계점을 의미한다. 0-1 손실함수(loss function)인 $SPE(x^-,\ w)$와 $SEN(x^+,\ w)$는 비평활함수(non-smoothed function)로 미분이 거의 불가능하거나 미분을 하더라도 0의 미분값을 가지게 되므로 결과적으로 경사하강법의 적용이 불가능하게 된다.

본 연구에서 $GM(x,\ \omega)$에 경사하강법을 적용하기 위하여 결합확률 $P_{X,Y}(x,\ y)$가 정규분포를 따른다고 가정한다. 정규분포 가정에 따라 다수 범주의 확률변수 X^-와 소수 범주와 확률 변수 X^+와 기저 분류자의 선형 결합으로 구성된 최종 분류자의 예측 값(W^TX)도 다음과 같이 정규 분포를 따르게 된다.

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

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

여기에서, μ^-와 μ^+는 다수 및 소수 범주 입력변수의 평균벡터(mean vector)이며, Σ^-와 Σ^+는 다수 및 소수 범주 입력벡터의 공분산 행렬(covari-

ance matrix)이다.

정규분포 가정 하에서 식 (1)에 자연로그를 취하게 되면 식 (4)와 같이 변형된다.

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

식 (4)의 $GM(x, w)$에 대한 1차 도함수 $\nabla GM(x, \omega)$는 식 (5)와 같이 정의된다.

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

정규분포 가정하에서 $SPE(x^-, w)$와 $SEN(x^+, w)$는 식 (6) 및 식 (7)과 같은 확률 개념으로 정의된다.

$$SPE(x^-, w) = P(W^T X^- \leq T) = \Phi\left(\frac{\mu_Z^-}{\sigma_{Z^-}} \leq T\right) \quad (6)$$

$$\begin{aligned} SEN(x^+, w) &= P(W^T X^+ > T) \\ &= (1 - P(W^T X^+ \leq T)) \\ &= 1 - \Phi\left(\frac{\mu_{Z^+}}{\sigma_Z^+} \leq T\right) \end{aligned} \quad (7)$$

여기에서 Φ는 표준정규분포 누적분포함수(CDF)이며 $\mu_{Z^-} = \omega^T \mu^-$, $\sigma_{Z^-} = \sqrt{\omega^T \Sigma^- \omega}$, 및 $\sigma_{Z^+} = \sqrt{\omega^T \Sigma^+ \omega}$ 이다. 식 (6)과 식 (7)에서 임계점 T에서 $SPE(x^-, w)$와 $SEN(x^+, w)$의 CDF는 식 (8) 및 식 (9)와 같이 정의된다.

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

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

$\Phi(SPE(x^-, w))$의 1차 미분함수($\nabla(SPE(x^-, w))$)는 식 (10)과 같으며, $\Phi(SEN(x^+, w))$의 1차 미분함수($\nabla(SEN(x^+, w))$)는 식 (11)과 같이 정의된다.

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

$$\nabla SEN(x^+, w) = -\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{T - \mu_Z^+}{\sigma_Z^+}\right)^2\right) \left(\frac{\sigma_Z^+(\mu^+ - T) - \left(\frac{\mu_Z^+ - T}{\sigma_Z^+}\right)\Sigma^+ \omega}{(\sigma_Z^+)^2}\right) \quad (11)$$

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

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

최종적으로 $GM(x, w)$의 최대화 문제는 식 (13)와 같은 쌍대문제로 $GE(x,w)$의 최소화 문제로 정의될 수 있다.

목적 함수: $Min\ GE(x, w) = 1 - GM(x, w)$

제약 조건: $\sum_{k=1}^{K} \omega_k = 1$ (13)

식 (13)의 최적해는 가우시안 경사하강법을 이

〈표 3〉 GMOPTBoost Algorithm

GMOPTBoost algorithm
1. Input: the outputs of n pairs from base classifiers $U=\{(C_k(x_i), y_i): i=1, \cdots, n\}$
2. Estimate mean vector and covariance matrix of multivariate normal distributions and linear combination of multivariate normal distributions. $X^- \sim N(\ mu\ ^-, \Sigma^-)$, $X^+ \sim N(\ mu\ ^+, \Sigma^+)$, $W^T X^- \sim N(\omega^T \mu^-, \omega^T \Sigma^- \omega)$, $W^T X^+ \sim N(\omega^T \mu^+, \omega^T \Sigma^+ \omega)$
3. Initialize $\omega_1^k = a_k$ for all k and choose learning rate β.
4. For j =1 to J
a. Calculate partial derivatives of Specificity and Sensitivity: $\nabla SPE(\omega)$ and $\nabla SEN(\omega)$
b. Calculate $SPE(\omega)$ and $SEN(\omega)$ at threshold T.
c. Calculate the first derivatives of GM: $\nabla GM(\omega)$
d. Update the set of weights $\omega_{j+1}^k = \omega_j^k - \beta \nabla GM(\omega))$
f. Normalize $\omega k_{j+1}^k / \sum_{k=1}^K \omega k_{j+1}^k$
5. Output: the prediction of the final classifier $C(x_i) = \begin{cases} -1, \ \sum_{k=1}^k \alpha_k C_k \le T \\ +1, \ \sum_{k=1}^k \alpha_k C_k > T \end{cases}$

용하여 결합 가중치 집합($\triangle\omega_k$)이 탐색되며 학습률(learning rate) β와 결합하여 $\omega_k^{new} = (\omega_k^{old} - \beta \cdot \triangle\omega_k)$와 같이 새로운 결합 가중치 집합($\omega_k^{new}$)을 생성하게 된다. 결합 가중치 집합의 총합이 1이 되도록 $\frac{\omega_k^{new}}{\sum_{k=1}^K \omega_k^{new}}$로 정규화된 가중치로 변환되어 후행 학습의 가중치로 활용된다. 이러한 학습과정은 종료조건(stop condition)이 만족할 때까지 반복 수행되며, 최종적으로 $GM(x,w)$을 극대화하는 최적의 결합 가중치(ω^*)가 탐색된다. GMOPTBoost 알고리즘의 간략한 절차는 <표 3>에 기술되어 있다.

Ⅳ. 연구방법론

4.1 표본 수집

부도에 대한 정의는 법률, 경제 및 금융 등 이론적 측면에서 다양하게 정의할 수 있는 바, 본 연구에서는 국제결제은행 산하의 바젤위원회의 은행감독안(Basel Committee on Banking Supervision)에 따라 금융감독원에서 규정한 부도의 정의를 활용하였다. 금융감독원에서는 1) 원리금상환기일 기준 원리금 연체 91일 이상(90일 초과) 업체 2) 법적 부도 업체(최종부도, 청산, 파산절차 진행, 폐업), 정리대상기업(법정관리, 부도유예 협약 및 화의, Work-out 등) 3) 대손상각 등과 같은 특정차주에 대한 신용손실 사건이 발생하고, 원금, 이자 또는 수수료감면 및 유예 등으로 채무조정이 이루어진 업체 4) 국내신용평가회사의 신용등급이 CCC 이하인 차주를 부도로 정의하고 있다.

본 연구에 활용되고 있는 재무비율은 한국상장협의회의 TS2000에서 재무제표 자료를 기초로 산출하였다. 재무제표의 수집 기간과 관련하여 건전기업은 당해 년도(t기)의 자료를 활용하였다. 반면 부도 기업의 경우 부도일자와 재무제표 공시일을 비교하여 부도일자가 공시일보다 이른 경우 전전기(t-2기)의 재무제표를 활용하였으며, 부도일자가 재무제표 공시일보다 늦은 경우 전기(t-1기)의 재무자료를 활용하였다.

본 연구의 부도 정의에 따라는 2015년부터 2018년까지 외부감사 제조기업 중 부도 기업의 관측치로 542개 기업과 부도 사유에 해당하지 않는 건전기업 10,965개의 기업-년도별 관측치로 총 11,507개의 원천 표본(original sample)을 수집하였다. 원천 표본 중 재무 자료가 불충분하거나 수집된 재

무비율별로 상하위 1%에 해당하는 관측치를 이상치로 제거한 결과 부도 기업 507개와 건전 기업 10,027개의 관측치를 확보하였다. 추가적으로 본 연구에서는 범주 불균형이 앙상블 학습의 학습성과에 미치는 영향에 대한 분석을 수행하기 때문에 다양한 범주 불균형 비율을 생성기준으로 국내 기업의 장기평균부도율이 약 3~5% 수준이라는 점을 고려하여 최종 표본으로 부도 기업 500개 및 건전 기업 10,000개의 총 10,500개의 관측치를 구성하였다. 이와 같은 표본 수집 과정은 <표 4>에 요약되어 있다.

본 연구에서는 두 가지 검증 목적에 따라 두 단계에 따라 다음과 같이 표본을 구성하였다. 첫 번째 단계에서는 범주 불균형이 예측 모형의 성과에 미치는 효과와 범주 불균형 문제에서 GMOPTBoost의 성과개선 효과를 검증하기 위하여 상이한 범주 불균형 비율(A(1:1), B(2:1), C(4:1), D(10:1), E(20:1)에 따라 5개의 하위 표본을 구성하였다. 본 연구에서는 10-flod 교차타당성 분석을 검증 절차로 사용하기 때문에 각각의 교차타당성 분석의 학습 표본과 검증 표본은 <표 5>와 같이 구성된다.

두 번째 단계에서는 범주 균형 문제에서도 GMOPTBoost 알고리즘이 AdaBoost 알고리즘의 성과 개선에 기여할 수 있는지 검증하기 위하여 데이터 샘플링 기법(RUS, CUS, ROS, SMOTE)를 활용하여 균형 표본을 구성하였다. 범주 균형에서 10-fold 교차타당성 분석을 적용하기 위하여 먼저 학습 표본과 검증 표본을 9:1의 비율로 분할한 이후에 데이터 샘플링 기법을 학습표본에만 적용하였다. 데이터 샘플링 기법에 따른 학습 표본과 검증 표본은 <표 6>과 같이 구성된다.

〈표 4〉 표본 기업의 구성

수집기준	건전 기업	부도 기업	관측치
원본데이터	10,965	542	11,507
재무자료 불충분	10,452	528	10,900
이상치 제거	10,027	507	10,534
범주불균형	10,000	500	10,500

〈표 5〉 범주 불균형 비율에 따른 학습표본과 검증표본의 구성

Datasets (IR)	Training(in-the-sample)			Validation(out-of-sample)		
	Normal	Bankrupt	Total	Normal	Bankrupt	Total
A (1:1)	450	450	900	50	50	100
B (2:1)	900	450	1,350	100	50	150
D (10:1)	4,500	450	4,950	500	50	550
E (20:1)	9,000	450	9,450	1,000	50	1,050

〈표 6〉 범주 균형 데이터의 학습표본과 검증표본의 구성

Datasets	Training(in-the-sample)			Validation(out-of-sample)		
	Normal	Bankrupt	Total	Normal	Bankrupt	Total
RUS	450	450	900	1,000	50	1,050
CUS	450	450	900	1,000	50	1,050
SMOTE	9,000	9,000	18,000	1,000	50	1,050

4.2 변수 선정

부도 예측을 위한 변수 선정을 위하여 일차적으로 선행연구에서 사용된 30개의 재무비율을 수집하였다(Altman, 1968; Beaver, 1996; Kim *et al.*, 2015; Lin *et al.*, 2018). 수집된 30개의 재무비율간의 공통성을 기초로 소수의 요인(재무비율 그룹)으로 축약하고 기업의 부도 속성을 다양한 관점에서 측정하기 위하여 탐색적 요인 분석(exploratory factor analysis)을 시행하였으며, 요인회전방식으로 요인간 직각을 가정하는 베리맥스 회전(varimax rotation)을 활용하였다. 요인분석 결과는 <표 7>에 제시되어 있다.

요인분석에 따라 그룹화된 7개 재무비율 그룹에서 AUROC가 가장 높은 재무비율 7개를 최종 설명변수로 선정하였으며, 각 재무비율의 AUROC는 <Table 8>에 제시되어 있다.

최종적으로 선정된 재무비율들에 대해 다중공선성(multicollinearity)를 분석하기 위해 <표 9>에 제시한 것과 같이 분산팽창요인(variance inflation factor) 분석을 실시한 결과, 변수간 VIF 값은 2.5 이하로 다중공선성 문제는 관측되지 않았다.

〈표 7〉 재무비율에 대한 요인분석 결과

Group	Variable	요인적재 값	고유치	분산설명비율
수익성	총자산경상이익률	0.814	2.612	65.302
	총자산이익률	0.811		
	매출액경상이익율	0.801		
	매출액순이익률	0.795		
	금융비용/매출액	0792		
	금융비용/총부채	0.785		
	순금융비용/매출액	0.774		
	자기자본경상이익율	0.772		
	자기자본순이익율	0.745		
부채상환능력	EBITA/이자비용	0.851	2.600	65.009
	EBIT/이자비용	0.842		
	영업현금흐름/이자비용	0.836		
	부채상환계수	0.796		
	잉여현금흐름/이자비용	0.747		
	잉여현금흐름/총부채	0.731		
	잉여현금흐름/총부채	0.725		
레버리지	자기자본비율	0.826	2.599	64.984
	유동자산/총자산	0.812		
자본구조	이익잉여금/총자산	0.814	2.522	93.062
	이익잉여금/유동자산	0.796		
	이익잉여금/총부채	0.721		
유동성	현금비율	0.919	2.459	81.973
	유동비율	0.899		
	당좌비율	0.899		
규모	총자산	0.927	2.412	80.412
	매출액	0.888		
	고정자산총자산	0.874		

〈표 8〉 30개 재무비율의 AUROC 분석결과

Financial ratio		AUC		Financial ratio	AUC
수익성	**총자산경상이익률***	**54.3**	레버리지	**자기자본비율***	**51.7**
	총자산이익률	45.2		유동자산/총자산	49.3
	금융비용/매출액	47.8	자본구조	**이익잉여금/총자산***	**51.3**
	금융비용/총부채	42.1		이익잉여금/총부채	50.5
	순금융비용/매출액	47.3		이익잉여금/유동자산	48.4
	매출액경상이익율	48.2	유동성	**현금비율***	**48.4**
	매출액순이익률	51.1		당좌비율	47.0
	자기자본경상이익율	45.3		유동비율	43.5
	자기자본순이익율	47.9	활동성	**재고자산회전율***	**33.4**
부채상환능력	**EBITA/이자비용***	**53.1**		유동부채회전율	29.4
	EBIT/이자비용	49.2		매출채권회전율	25.3
	영업현금흐름/이자비용	44.7	규모	**총자산***	**23.7**
	잉여현금흐름/총부채	46.2		매출액	20.3
	잉여현금흐름/이자비용	51.0		고정자산	23.4
	잉여현금흐름/총부채	48.3			
	부채상환계수	47.2			

주) 1) *최종 선정된 7개 재무비율.

〈표 9〉 최종 7개 재무비율의 VIF 분석 결과

Group	Variable	VIF
수익성	총자산경상이익율	1.32
부채상환능력	EBITA/이자비용	2.15
레버리지	자기자본비율	1.78
자본구조	이익잉여금/총자산	2.52
유동성	현금비율	1.36
활동성	재고자산회전율	1.51
규모	총자산	1.35

4.3 연구모형 설계

본 연구에서는 기저 분류기(base classifier)로 3개의 층으로 구성된 인공신경망을 활용하였다. 인공신경망의 성과는 모수(hyper-parameter)의 설정에 따라 달라질 수 있기 때문에 본 연구에서는 일정 범위의 모든 경우의 실험을 통해 가장 좋은 성과를 보이는 모수를 탐색하는 그리드 탐색(grid search) 방식을 사용하여 최종적으로 은닉층의 노드, epoch, 정지조건, 학습률 등의 모수를 결정하였다.

이러한 방식에 따라 인공신경망은 7개의 입력 노드, 5개의 은닉 노드, 1개의 출력 노드로 구성하였으며 모든 계층에서 활성화 함수는 시그모이드(sigmoid) 함수를 활용하였다. 학습 알고리즘은 학습률이 0.1인 역전파(back propagation)이며 학습의 종료 조건으로 epoch를 10,000으로 설정하였다.

AdaBoost 알고리즘에서 반복적인 기저 분류자의 생성은 기저 분류자의 오류율이 0.5 이상인 경우 기저 분류자의 생성을 중단하도록 하였다.

GMOPTBoost 알고리즘은 GM을 최대화하는 결

합 가중치는 AdaBoost 알고리즘에서 산출된 결합 가중치를 초기값으로 하며 종료 조건으로 epoch를 10,000으로 설정하였다.

V. 연구 결과

5.1 범주 불균형 표본에 대한 성과 분석

본 연구에서는 기존의 AdaBoost와 제안한 GMOPTBoost 알고리즘의 성과 비교를 수행하기 위해 10-fold 교차타당성 검증을 3회 반복하여 총 30회의 검증을 수행하였다. <표 10>은 <표 5>의 범주 불균형 표본에 대한 30회 검증결과로서 정확도(ACC), 기하평균 정확도(GM), AUROC의 평균값과 두 모형의 성과 차이를 비교한 t-검정의 결과를 제시하고 있다.

주요 분석 결과는 다음과 같다. 첫째, AdaBoost 및 GMOPTBoost 앙상블의 모든 성과 측정치는 범주 불균형 비율이 증가함에 따라 감소되는 것으로 분석되었다. 이러한 결과는 범주 불균형 문제가 분류모형의 성과에 부정적인 영향을 미친다는 것을 의미한다. AdaBoost의 경우 범주 불균형 비율이 증가함에 따라 정확도(ACC)는 86.7%에서 69.4%, GM은 86.6%에서 70.1%, AUROC는 92.4%에서 76.0%로 감소하였다. GMOPTBoost 앙상블의 경우에도 ACC는 90.5%에서 75.4%, GM은 90.5%에서 75.2%, AUROC는 96.0%에서 82.6%로 감소하였다.

둘째, GMOPTBoost 앙상블은 AdaBoost 앙상블의 성과 개선에 긍정적으로 기여하는 것으로 분석되었다. AdaBoost 및 GMOPTBoost 앙상블의 성과에 대한 t-검정 결과 두 앙상블 모형의 성과는 1% 수준에서 유의적인 차이를 보이는 것으로 분석되었다. 특히 불균형 비율이 증가할수록 성과 개선의 효과가 증가하고 있는데 A(1:1) 표본에서 E(20:1) 표본으로 범주 불균형 비율이 높아질수록 AdaBoost 앙상블과 GMOPTBoost 앙상블의 성과 차이는 더욱 확대되고 있다. ACC는 4.3%에서 8.6%, GM은 4.3%에서 7.2%, AUROC는 3.9%에서 8.6%로 확대되는 것으로 분석되었다. 이러한 결과는 GMOPTBoost 앙상블이 AdaBoost 앙상블의 성과 향상에 유의적으로 기여하고 있음을 의미한다.

5.2 범주 균형 표본에 대한 성과 분석

<표 11>은 <표 6>의 균형 표본에 대하여 30회의 교차타당성 분석을 수행한 검증결과를 제시하고 있다. 주요 분석 결과는 다음과 같다. 첫째, 데이터 샘플링 기법을 활용하여 구성된 범주 균형 표본에 대하여 학습된 AdaBoost 및 GMOPTBoost 앙상블의 모두에서 예측 성과가 크게 개선되어 데이터 샘플링은 예측 성과 개선에 긍정적으로 기여하는 것으로 분석되었다. <표 10>에서 E(20:1)에 대한 AdaBoost 알고리즘의 ACC(0.694), GM(0.701), AUROC(0.760)으로 측정된 반면, 데이터 샘플링 기법 중 가장 낮은 성과를 보이는 CUS의 경우에도 ACC(0.747), GM(0.747), AUROC(0.857)로 성과

〈표 10〉 범주 불균형 표본에 대한 성과 분석

Datasets (IR)	AdaBoost			GMOPTBoost			t-test		
	ACC	GM	AUROC	ACC	GM	AUROC	ACC	GM	AUROC
A (1:1)	0.867	0.866	0.924	0.905	0.904	0.960	2.77*	2.79*	3.01*
B (2:1)	0.834	0.833	0.903	0.884	0.880	0.943	3.46*	3.30*	2.60*
C (4:1)	0.815	0.815	0.888	0.861	0.858	0.930	3.40*	3.25*	2.65*
D (10:1)	0.756	0.755	0.828	0.797	0.803	0.875	3.33*	2.86*	2.51*
E (20:1)	0.694	0.701	0.760	0.754	0.752	0.826	3.80*	2.69*	3.15*

주) 1) *1% 수준에서 유의.

〈표 11〉 범주 균형 표본에 대한 성과 분석

Datasets	AdaBoost			GMOPTBoost			t-test		
	ACC	GM	AUROC	ACC	GM	AUROC	ACC	GM	AUROC
RUS	0.78	0.782	0.861	0.793	0.794	0.874	4.04*	3.84*	3.91*
CUS	0.747	0.747	0.857	0.779	0.773	0.869	5.13*	4.07*	2.52*
ROS	0.751	0.761	0.848	0.780	0.786	0.872	2.76*	3.28*	3.25*
SMOTE	0.753	0.765	0.854	0.783	0.783	0.874	2.87*	2.44*	2.48*

주) 1) *1% 수준에서 유의.

개선이 이루어졌음을 알 수 있다. GMOPTBoost의 경우에도 <표 10>의 E(20:1)에 제시된 바와 같이 ACC(0.754), GM(0.752) 및 AUROC(0.826)와 비교하여 가장 낮은 성과를 보이는 CUS의 경우에도 ACC(0.779), GM(0.773), AUROC(0.869)로 성과가 개선된 것으로 분석되었다. 비록 본 연구의 본문으로 제시하지는 못하였지만, 범주 불균형과 범주 균형 표본에 적용된 AdaBoost와 GMOPTBoost의 성과 차이에 대한 t-검정 결과에서도 1% 수준에서 유의적으로 성과차이가 발생하고 있음을 확인하였다.

둘째, GMOPTBoost는 데이터 샘플링을 적용하여 구성한 범주 균형 표본에 대하여 학습된 AdaBoost 알고리즘의 성과를 유의적으로 개선하는 것으로 분석되었다. AdaBoost 및 GMOPTBoost 앙상블의 성과에 대한 t-검정 결과 두 앙상블 모형의 성과는 1% 수준에서 유의적인 차이를 보이는 것으로 분석되었다. 이러한 결과는 데이터 샘플링과 앙상블 학습의 적용하는 것만으로는 최적화된 성과가 보장되는 것은 아니라는 선행 연구의 주장과 일치하고 있으며, 본 연구에서 제안한 GMOPTBoost를 적용하여 유의적인 성능 개선이 이루어졌음을 의미한다.

Ⅵ. 결 론

본 연구는 비즈니스 분야의 대표적인 범주 불균형 문제인 기업부도 예측의 성과를 개선할 수 있는 GM 최적화 기반의 GMOPTBoost알고리즘을 제안하였다. 실험 결과, 본 연구에서 제안한 GMOPTBoost 알고리즘은 다양한 범주 불균형 표본에 대하여 학습된 AdaBoost 알고리즘의 산술평균 정확도, GM 및 AUROC 등의 유의적인 성과 개선에 기여하여 범주 불균형 문제 해결에 효과적임을 보였으며 특히, 범주 불균형이 심화될수록 GMOPTBoost 알고리즘의 성과 개선 효과가 확대되고 있음을 분석하였다. 범주 불균형 문제 개선 효과와 더불어 GMOPTBoost 알고리즘은 데이터 샘플링 기법을 적용한 범주 균형 데이터에 대하여 학습된 AdaBoost 앙상블 모형의 성과 개선에도 기여할 수 있음을 확인하였다.

본 연구의 공헌점은 다음과 같다. 첫째, 정확도, GM 및 AUROC와 같은 성과 측정치는 예측 모형의 궁극적인 학습목표이지만, 이러한 성과 측정치는 비평활함수의 특성을 가지기 때문에 최적화 과정을 직접적으로 적용하는 것은 불가능하다는 한계가 있다. 이러한 한계점을 극복하기 위하여 본 연구는 정규성을 가정함으로써 경사하강법에 기반한 GM의 최적화 알고리즘을 제안하고 실증 분석 자료를 제공하고 있는바, 본 연구는 성과 측정치에 대한 최적화 기법을 적용하기 위한 이론적 토대 및 실증 증거를 제공하고 있다.

둘째, 선행연구에서는 범주 불균형 문제에 대한 대표적인 해결기법으로 데이터 샘플링 기법과 앙상블학습을 추천하고 있지만 이러한 기법들의 적용만으로 성과의 최적화가 보장되지는 않는다. 본 연구에서는 데이터 샘플링 및 앙상블 학습과 더불어 최적화 기법을 추가적으로 적용하여 유의

적인 성과 개선을 유도하였다는 점에서 범주 불균형 문제에 대하여 보다 효과적인 해결 대안을 제시하고 있다.

향후의 연구 방향과 관련하여 본 연구는 다음과 같은 한계점을 가진다. 첫째, 본 연구에서는 기저 분류자로서 인공신경망을 활용하고 있지만, 향후 연구에서는 SVM, 로지스틱회귀분석 및 의사결정나무 등 다양한 기법의 적용되어 GMOPTBoost의 일반화 가능성이 검증되어야 한다. 둘째, 본 연구에서는 기업부도 예측과 같은 이범주 문제에 내재된 범주 불균형 문제에만 국한하여 GMOPTBoost 모형을 제안하였다. 하지만, 현실 문제에서는 신용등급평가, 회사채평가 등의 다범주 (multi-class) 불균형 문제가 존재하고 있다. 이러한 다범주 문제에 내재된 범주 불균형 문제를 해결하기 위하여 GMOPTBoost모형을 수정한 기법이 개발되어야 한다. 마지막으로 본 연구의 설명 변수는 재무 데이터에 국한되어 있다. 최근의 연구에서 재무 데이터와 비재무 데이터를 결합함으로써 모형의 성과가 더욱 개선되는 것으로 보고되고 있는 바 향후 연구에서는 다양한 데이터를 활용하여 예측 모형의 성과 향상을 고려할 필요가 있다.

참 고 문 헌

[1] 김량형, 유동희, 김건우, “데이터마이닝 기법을 이용한 기업부실화 예측 모델 개발과 예측 성능 향상에 관한 연구”, *Information Systems Review*, 제18권, 제2호, 2016, pp. 173-198.

[2] 안철휘, 안현철, “효과적인 기업부도 예측모형을 위한 ROSE 표본추출기법의 적용”, *한국콘텐츠학회논문지*, 제18권, 제8호, 2018, pp. 525-535.

[3] Altman, E. L., “Financial ratios, discriminant analysis and the prediction of corporate bankruptcy”, *The Journal of Finance*, Vol.23, No.4, 1968, pp. 589-609.

[4] Barboza, F., H. Kimura, and E. Altman, “Machine Learning Models and Bankruptcy Prediction”, Expert Systems with Applications, Vol.83, 2017, pp. 405-417.

[5] Barua, S., M. Islam, and X. Yao, “MWMOTE-Majority weighted minority oversampling technique for imbalanced data set learning”, *IEEE Transaction on Knowledge and Data Engineering*, Vol.26, No.2, 2014, pp. 405-424.

[6] Beaver, W., “Financial ratios as predictors of failure, empirical research in accounting: Selected studied”, *Journal of Accounting Research*, Vol.4, No.3, 1996, pp. 71-111.

[7] Chawla, N. V., A. Lazarevic, L. O. Hall, and K. W. Bowyer, “SMOTEBoost: Improving prediction of the minority class in boosting”, *Proceedings of 7th European Conference on Principles and Practice of Knowledge Discovery in Databases*, 2003, pp. 107-119.

[8] Davis, J. and M. Goadrich, “The relationship between precision-recall and ROC curves”, *Proceedings of the 23rd International Conference on Machine Learning*, 2006, pp. 233-240.

[9] Fawcett, T., “An introduction to ROC analysis”, *Pattern Recognition Letters*, Vol.27, No.8, 2006, pp. 861-874.

[10] Freund, Y. and R. E. Schapire, “A Decision theoretic generalization of online learning and an application to boosting”, *Journal of Computer and System Science*, Vol.55, No.1, 1997, pp. 119-139.

[11] He, H. and E. A. Garcia, “Learning from imbalanced data”, *IEEE Transactions on Knowledge and Data Engineering*, Vol.21, No.9, 2009, pp. 1263-1284.

[12] Kim, M. J. and D. K. Kang, “Ensemble with neural networks for bankruptcy prediction”, *Expert Systems with Applications*, Vol.37, No.4, 2010, pp. 3373-3379.

[13] Kim, M. J., D. K. Kang, and H. B. Kim,

"Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction", *Expert Systems with Applications*, Vol.42, No.3, 2015, pp. 1074-1082.

[14] Kim, S. Y. and A. Upneja, "Predicting restaurant financial distress using decision tree and adaboosted decision tree models", *Economic Modeling*, Vol.36, 2014, pp. 354-362.

[15] Kuncheva, L. I., Á. Arnaiz-González, J. F. Díez-Pastor, and L. A. D. Gunn, "Instance selection improves geometric mean accuracy: A study on imbalanced data classification", *Progress in Artificial Intelligence*, Vol.8, 2019, pp. 215-228.

[16] Kwon, Y. S., I. Han, and K. C. Lee, "Ordinal pairwise partitioning(OPP) approach to neural networks training in bond rating", *Intelligent Systems in Accounting, Finance and Management*, Vol.6, 1997, 23-40.

[17] Le, T., M. Y. Lee, J. R. Park, and S. W. Baik, "Oversampling techniques for bankruptcy prediction: novel features from a transaction dataset", *Symmetry*, Vol.10, No.4, 2018b, Available at https://doi.org/10.3390/sym10040079.

[18] Le, T., L. H. Son, M. T. Vo, M. Y. Lee, and S. W. Baik, "A cluster-based boosting algorithm for bankruptcy prediction in a highly imbalanced dataset", *Symmetry*, Vol.10, No.7, 2018a. Available at https://doi.org/10.3390/sym10070250.

[19] Lin, W. C., C. F. Tsai, Y. H. Hu, and J. S. Jhang, "Clustering-based undersampling in class imbalanced data", *Information Sciences*, Vol.409-410, 2017, pp. 17-26.

[20] Mellor, A., S. Boukir, A. Haywood, and S. Jones, "Exploring issues of training data imbalance and mislabelling on random forest performance for large area land cover classification using the ensemble margin", *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol.105, 2015, pp. 155-168.

[21] Messier, W. F. Jr. and J. V. Hansen, "Inducing rules for expert system development: An example using default and bankruptcy data", *Management Science*, Vol.34, No.4, 1998, pp. 1403-1415.

[22] Nanni, L. and A. Lumini, "A genetic encoding approach for learning methods for combining classifiers", *Expert Systems with Applications*, Vol.36, No.4, 2009, pp. 7510-7514.

[23] Odom, M. D. and R. Sharda, "A neural network model for bankruptcy prediction", *IJCNN International Joint Conference on Neural Networks Neural Networks*, Vol.2, 1990, pp. 163-168.

[24] Schapire, R. E., "The strength of weak learnability", *Machine Learning*, Vol.5, No.2, 1990, pp. 197-227.

[25] Seiffert, C., T. M. Khoshgoftaar, J. V. Hulse, and A. Napolitano, "RUSBoost: Improving classification performance when training data is skewed", *Proceedings of the 19th International Conference on Pattern Recognition*, 2008, pp. 1-4.

[26] Shin, K., T. Lee, and H. Kim, "An application of support vector machines in bankruptcy prediction", *Expert Systems with Applications*, Vol.28, 2005, pp. 127-135.

[27] Somasundaram, A. and S. Reddy, "Parallel and incremental credit card fraud detection model to handle concept drift and data imbalance", *Neural Computing and Applicatopms*, Vol.31, 2019, pp. 3-14.

[28] UlagaPriya, K. and S. Pushpa, "A comprehensive study on ensemble-based imbalanced data classification methods for bankruptcy data", *IEEE 6th international Conference on Inventive Computation Technologies(ICICT)*, 2021. pp. 800-804.

[29] Weng, C. G. and J. Poon, "A new evaluation

measure for imbalanced datasets", *Proceedings of the 7th Australasian Data Mining Conference*, Vol.87, 2008, pp. 27-32.

[30] Zhang, G., M. Y. Hu, B. E. Patuwo, and D. C. Indro, "Theory and methodology artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis", *European Journal of Operational Research*, Vol.116, 1999, pp. 16-32.

[31] Zieba, M., S. K. Tomczak, and J. M. Tomczak, "Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction", *Expert Systems with Applications*, Vol.58, 2016, pp. 93-101.

Information Systems Review
Volume 24 Number 1
February 2022

The Optimization of Ensembles for Bankruptcy Prediction

Myoung-Jong Kim* · Woo-Seob Yun**

Abstract

This paper proposes the GMOPTBoost algorithm to improve the performance of the AdaBoost algorithm for bankruptcy prediction in which class imbalance problem is inherent. AdaBoost algorithm has the advantage of providing a robust learning opportunity for misclassified samples. However, there is a limitation in addressing class imbalance problem because the concept of arithmetic mean accuracy is embedded in AdaBoost algorithm. GMOPTBoost can optimize the geometric mean accuracy and effectively solve the category imbalance problem by applying Gaussian gradient descent. The samples are constructed according to the following two phases. First, five class imbalance datasets are constructed to verify the effect of the class imbalance problem on the performance of the prediction model and the performance improvement effect of GMOPTBoost. Second, class balanced data are constituted through data sampling techniques to verify the performance improvement effect of GMOPTBoost. The main results of 30 times of cross-validation analyzes are as follows. First, the class imbalance problem degrades the performance of ensembles. Second, GMOPTBoost contributes to performance improvements of AdaBoost ensembles trained on imbalanced datasets. Third, Data sampling techniques have a positive impact on performance improvement. Finally, GMOPTBoost contributes to significant performance improvement of AdaBoost ensembles trained on balanced datasets.

Keywords: Bankruptcy Prediction, Class Imbalance, Class Balance, Data Sampling, Geometric Mean, GMOPTBoost

* Corresponding Author, Professor, School of Business, Pusan National University
** Undergraduate Student, School of Business, Pusan National University

◐ 저 자 소 개 ◑

Myoung-Jong Kim (mjongkim@pusan.ac.kr)

He is a professor of Division of Business in Pusan National University. He received a BS and MS degree from Sungkyunkwan University, and a PhD from Korea Advanced Institute of Science and Technology in Korea. He has published many papers related to the business applications of Artificial Intelligence. His main research interests are Data Mining and intelligent systems in accounting and finance fields.

Woo-Seob Yun (tpdbs4032@pusan.ac.kr)

He is an undergraduate student of Division of Business in Pusan National University. He is majoring in business administration and industrial mathematics software. He published paper related to corporate bankruptcy. His main research interests are artificial intelligence, quantitative economy, finance fields.

논문접수일 : 2021년 10월 29일 **게재확정일 : 2021년 12월 29일**
1차 수정일 : 2021년 12월 09일