https://doi.org/10.5392/JKCA.2018.18.08.525

효과적인 기업부도 예측모형을 위한 ROSE 표본추출기법의 적용

Application of Random Over Sampling Examples(ROSE) for an Effective Bankruptcy Prediction Model

안철휘, 안현철
국민대학교 비즈니스IT전문대학원

Cheolhwi Ahn(7089ach@kookmin.ac.kr), Hyunchul Ahn(hcahn@kookmin.ac.kr)

요약

분류 문제에서 특정 범주의 빈도가 다른 범주에 비해 과도하게 높은 경우, 왜곡된 기계 학습을 유발할 수 있는 데이터 불균형(imbalanced data) 문제가 발생한다. 기업부도 예측 문제도 그 중 하나인데, 일반적으로 금융기관과 거래하는 기업들의 부도율은 대단히 낮아서, 부도 사례보다 정상 사례의 빈도가 월등히 높은 데이터 불균형 문제가 발생하고 있다. 이러한 데이터 불균형 문제를 해결하기 위해서는 적절한 표본추출 기법이 적용될 필요가 있으며, 지금껏 소수 범주 데이터를 복원 추출함으로써 다수 범주 데이터와 비율을 맞추어 데이터 불균형을 해결하는 오버 샘플링(oversampling) 기법이 주로 활용되어 왔다. 그러나 전통적인 오버 샘플링은 과적합화(overfitting)가 발생할 위험이 높아질 수 있는 단점이 있다. 이러한 배경에서 본 연구는 효과적인 기업부도 예측 모형 학습을 위한 표본추출 기법으로 2014년에 Menardi와 Torelli가 제안한 ROSE(random over sampling examples) 기법을 제안한다. ROSE 기법은 학습에 사용될 사례를 반복적으로 새롭게 합성하여 생성(synthetic generation)하는 기법으로, 과적합화 문제를 회피하면서도 분류 예측 정확도 개선에 도움을 줄 수 있다. 이에 본 연구에서는 ROSE 기법을 가장 성능이 우수한 이분류기로 알려진 SVM(support vector machine)과 결합하여 국내 한 대형 은행의 기업부도 예측에 적용해 보고, 다른 표본추출 기법들과의 비교연구를 수행하였다. 실험 결과, ROSE 기법이 다른 기법에 비해 통계적으로 유의한 수준으로 SVM의 예측정확도 개선에 기여할 수 있음을 확인하였다. 이러한 본 연구의 결과는 부도 예측 외에 다른 사회과학 분야 예측문제의 데이터 불균형 문제 해결에도 ROSE가 우수한 대안이 될 수 있다는 사실을 시사한다.

■ 중심어 : | ROSE | 데이터 불균형 | 표본추출 | 부도 예측 |

Abstract

If the frequency of a particular class is excessively higher than the frequency of other classes in the classification problem, data imbalance problems occur, which make machine learning distorted. Corporate bankruptcy prediction often suffers from data imbalance problems since the ratio of insolvent companies is generally very low, whereas the ratio of solvent companies is very high. To mitigate these problems, it is required to apply a proper sampling technique. Until now, oversampling techniques which adjust the class distribution of a data set by sampling minor class with replacement have popularly been used. However, they are a risk of overfitting. Under this background, this study proposes ROSE(Random Over Sampling Examples) technique which is proposed by Menardi and Torelli in 2014 for the effective corporate bankruptcy prediction. The ROSE technique creates new learning samples by synthesizing the samples for learning, so it leads to better prediction accuracy of the classifiers while avoiding the risk of overfitting. Specifically, our study proposes to combine the ROSE method with SVM(support vector machine), which is known as the best binary classifier. We applied the proposed method to a real-world bankruptcy prediction case of a Korean major bank, and compared its performance with other sampling techniques. Experimental results showed that ROSE contributed to the improvement of the prediction accuracy of SVM in bankruptcy prediction compared to other techniques, with statistical significance. These results shed a light on the fact that ROSE can be a good alternative for resolving data imbalance problems of the prediction problems in social science area other than bankruptcy prediction.

■ keyword : | Random Over Sampling Examples | Data Imbalance | Sampling | Bankruptcy Prediction |

* 이 논문은 2017년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2017S1A5A2A03067632).
접수일자 : 2018년 07월 09일
수정일자 : 2018년 08월 21일
심사완료일 : 2018년 08월 21일
교신저자 : 안현철, e-mail : hcahn@kookmin.ac.kr

I. 서 론

데이터 불균형(data imbalance)문제란 이분류 문제에서 특정 범주(class)의 빈도가 다른 범주에 비해 과도하게 높은 경우를 말한다. 대부분의 분류 기계 학습알고리즘에서 학습의 왜곡을 피하기 위해서는 데이터 범주 비율이 1:1로 동등한 조건에서 학습을 진행하여야 한다. 그러나 사기 탐지(fraud detection) 등과 같은 현실문제에서는 일반적으로 정상인 데이터는 빈도가 높은 반면, 비정상 혹은 부정을 나타내는 데이터의 비중은 상당히 적은 경우가 많다. 때문에 데이터의 불균형을 해결하기 위한 다양한 표본추출(sampling) 기법들이 지금껏 사용되어 왔다.

대표적인 표본추출 기법 중 하나인 오버 샘플링(oversampling) 기법은 소수 범주 데이터를 복원 추출함으로써 다수 범주 데이터와 비율을 맞추어 데이터 불균형을 해결하는 표본추출 기법이다. 이러한 오버 샘플링 기법은 소수 범주 데이터 수에 맞춰 다수 범주 데이터를 무작위 추출하여 적은 수로 데이터의 균형을 맞추려는 언더 샘플링(undersampling) 기법에 비해 예측 정확도 개선에 더 효과적인 것으로 알려져 있다. 하지만, 전통적인 오버 샘플링 기법 하에서는 무작위 복원 추출로 인해 중복되는 많은 데이터가 생성되기 때문에 과적합화(overfitting) 문제가 발생할 위험이 증가하여, 안정적인 예측이 어려워지는 단점이 있다.

이러한 배경에서 본 논문에서는 전통적인 오버 샘플링 기법의 단점인 과적합화 문제를 회피하면서도, 분류 예측 정확도 개선에 크게 도움을 줄 수 있는 것으로 알려진 새로운 ROSE(Random Over Sampling Examples) 기법의 적용을 제안한다. 제안하는 ROSE 기법은 Menardi and Torelli[1]가 제안한 표본추출 기법으로, 부드러운 부트스트랩(smooth bootstrap)을 적용해 학습용 데이터를 새롭게 생성하는 원리를 따른다. 여기서 부트스트랩(bootstrap)이란 표본에 대해 재표본추출(resampling)을 적용하는 것을 말하며, 그 중에서도 부드러운 부트스트랩 기법은 표본에서 재표본추출된 관측치에 평균이 0이고 정규분포를 따르는 아주 작은 양의 무작위 잡음(random noise)을 더해 새로운 학습용 표본을 생성(synthesize)하는 원리를 따른다. 이러한 ROSE 기법은 지금껏 다양한 연구들을 통해 전통적인 오버 샘플링 기법보다 더 나은 예측정확도의 개선을 도모하는 것으로 알려져 왔다. 하지만 경영분야, 그 중에서도 기업부도 예측에 ROSE를 적용한 사례는 국내는 물론 해외에서도 찾아보기 어렵다. 이에 본 연구에서는 대표적인 데이터 불균형 사례 중 하나인 기업 부도예측 문제에 ROSE 기법을 적용하고, 그 유용성을 검증해 보고자 한다.

이후 논문의 구성은 다음과 같다. 2장에서는 부도예측과 표본추출 기법과 관련한 선행연구들을 살펴본다. 이어 3장에서는 본 연구에서 제안하는 핵심 알고리즘인 ROSE 기법을 소개하고, 4장에서는 ROSE 기법의 성능을 검증하기 위한 부도예측 데이터 수집과 실험 과정에 따른 결과 및 검증을 설명한다. 마지막 5장에서는 전반적인 연구의 결론 및 한계점과 향후 연구 방향에 대하여 논의한다.

II. 이론적 배경

1. 부도 예측

최초 기업부도 예측 선행연구는 Beaver[2]에 의해 처음으로 진행되었다. 당시 이 연구는 기업의 재무 비율의 평균 차이를 분석하였는데, 단변량 통계분석기법(univariate analysis)을 이용하여 수익성, 유동성 등과 관련된 재무 비율의 평균값의 차이를 통해 기업채무상환 위험도를 예측하는 연구를 진행하였다. 그러나 이 연구는 재무 비율 평균값의 차이가 통계적으로 유의한지에 대해 검증이 이루어지지 않았으며, 부도예측을 진행할 때 오직 하나의 변수만을 고려하고 다른 변수들은 고려하지 못했다는 한계점을 가지고 있었다.

이와 같은 한계점을 해결하기 위해 Altman[3]은 다수의 변수들을 하나로 통합하여 분석하는 기법인 다중판별분석(multivariate discriminant analysis)을 이용하여 기업부도 예측연구를 수행하였다. 본 연구는 개별적으로 영향력이 확인되던 변수들을 하나의 모형으로 통합하고 단순화시켜, 보다 정확하고 명확한 형태로 부도

예측 실험을 가능하게 하였다. 그렇지만 분석을 수행하기 위한 각 집단의 분산과 공분산의 구조가 동일해야 하는 조건을 만족시켜야 하는 한계점을 가지고 있었다.

이를 극복하기 위해 후속 연구들은 확률판별함수(probability discriminant function)를 제안하였다. 이러한 연구들은 기업의 부도 확률과 비부도 확률을 퍼센트(percent)로 제시함과 동시에, 임계치(보통 0.5)를 중심으로 기업의 부도여부를 1과 0으로 분류, 예측하여 의사결정에 활용할 수 있는 장점이 있다. 확률판별함수를 이용한 대표적인 연구는 Ohlson[4]이 진행한 로지스틱 회귀분석을 통한 기업부도 예측 연구이다. 이 연구는 시그모이드(sigmoid) 함수를 예측 모형에 적용하여 기업부도 예측을 위한 실험을 진행하였다. 로지스틱 회귀분석은 다중 판별 분석과 달리 0과 1 사이의 값을 산출하여, 확률적으로 해석을 할 수 있는 장점이 있다. 그러나 예측력이 높지 않다는 한계점을 가지고 있었다.

이를 보완하기 위해 Zmijewski[5]의 프로빗 분석(probit analysis)과 Edmister[6]가 진행한 다중회귀분석 등 다양한 통계 기법을 적용한 연구가 진행되었다. 1980년대 후반에는 인공신경망(artificial neural network, ANN), 사례기반추론(case based reasoning, CBR), 의사결정나무(decision tree, DT) 등 기계학습 기법을 이용한 연구가 진행되었는데, 높은 예측력을 가진다고 알려진 인공신경망이 기업부도 예측연구에서 많이 사용되었다. Odom and Sharada[7]는 인공신경망을 부도예측모델에 적용한 최초의 연구로 알려져 있다. 이 연구에서는 판별 분석과 인공신경망을 동시에 사용하여, 재무비율변수들을 독립변수로 두고 부도여부를 예측하였는데, 실험 결과 인공신경망의 예측정확도가 더 우수함을 확인하였다.

이후 부도예측 연구는 신경망과 통계적 방법론의 예측성과를 비교하는 형태의 연구로 발전하였다. Tam and Kiang[8]은 은행의 부도예측 여부를 대표적인 기계학습 기법인 인공신경망과 의사결정나무를 이용하여 구축하고, 이를 판별분석, 로지스틱 회귀분석, K-최근접이웃 방법(K-nearest Neighbor, K-NN) 등 3가지 통계 기반 기법들과 비교 분석하였다[7]. 그 결과 비선형 패턴을 반영할 수 있는 인공신경망이 예측정확도에서 가장 우수한 결과를 산출했음을 확인하였다.

이러한 Tam and Kiang[8]의 선구적 연구 이후, 많은 학자들이 인공신경망 기법의 변형 또는 개선을 통하여 부도예측 정확도를 개선하기 위해 노력하였다. 예를 들어, Serrano-Cinca[9]는 재무 분야를 중심으로 자기 조직화 신경망(Self-Organizing Feature Maps)의 활용 가능성을 연구하였다. 또한 Yang and Honavar[10]는 PNN(Probabilistic Neural Networks)을 부도예측에 적용하여 판별 분석과 역전파 신경망을 비교하였다. 연구 결과 PNN이 판별 분석보다 예측정확도에서 우수한 결과가 나왔음을 확인하였다.

한편 국내에서도 1997-1998년 IMF 사태로 인해 기업부도 문제가 사회문제로 대두되면서, 과학적인 기업부도예측에 대한 관심이 높아졌다. 이로 인해 국내 데이터를 활용한 기업부도예측 연구가 90년대 후반부터 2000년대 초반을 기점으로 활발하게 이루어졌다. 그 예로 김경재, 한인구[11]는 퍼지 신경망을 이용하여 국내의 기업부도 예측에 적용하였고, 이영찬[12]은 인공신경망과 SVM(Support Vector Machine)을 이용하여 기업부도 예측에 적용하고, 그 결과 SVM이 가장 높은 예측정확도를 보인다는 점을 보고하였다.

2. 데이터 불균형

일반적인 데이터의 비정상 범주의 빈도는 정상 범주의 빈도보다 매우 적다. 이와 같은 데이터 불균형 문제는 금융, 마케팅, 의료, 재난 관리 등 다양한 분야에서 발생한다. 데이터 불균형 문제는 최근 상당한 주목을 받아왔고, 문제를 해결을 위해 많은 연구가 진행되었다.

데이터 불균형문제를 해결하는 가장 기본적인 방법은 바로 표본추출을 이용하는 것이다. 대표적인 표본추출 기법으로는 소수 범주를 기준으로 다수 범주에 속해있는 데이터를 무작위로 비복원 추출하여, 소수 범주 데이터와 다수 범주 데이터를 1:1 비율로 맞추는 언더 샘플링 기법이 있고, 다수 범주를 기준으로 소수 범주를 무작위 복원 추출하여 다수 범주 데이터와 비율을 1:1 비율로 맞추는 오버 샘플링 기법이 있다. 데이터 불균형을 해결하기 위한 기존 연구들은 이러한 언더 샘플링과 오버 샘플링을 중심으로 수행되어 왔다. 다음의

표 1. 데이터 불균형 해결을 위한 기존 연구 동향

기본 접근방식	연구자	발표시기	연구 내용
언더 샘플링	강필성, 조성준 [13]	2006	언더 샘플링 기반 앙상블 SVM 기법 연구
	이재동, 이지형 [14]	2014	K-means 군집화 기법 기반 SVM 앙상블 연구
	김태훈, 안현철 [15]	2015	언더 샘플링과 k-RNN을 결합한 새로운 하이브리드 표본추출 기법 연구
오버 샘플링	Japkowicz [16]	2000	기본 표본추출 기법인 오버 샘플링과 언더 샘플링 비교 연구
	Chawla 등[17]	2002	k-NN 알고리즘을 활용한 오버 샘플링 기법인 SMOTE 기법 제안
	이재동, 이지형 [18]	2015	딥러닝을 적용한 DAE(Deep-Auto-Encoder) 기법 제안

[표 1]은 데이터 불균형을 해결하기 위한 언더 샘플링과 오버 샘플링 관련 국내·외 연구의 동향을 종합적으로 제시하고 있다.

국내 연구는 언더 샘플링 접근법을 중심으로 주로 연구가 수행되어 왔다. 강필성과 조성준[13]은 언더 샘플링 기법을 기반으로 하는 앙상블 SVM 모델을 제안하였다. 이 모델은 언더 샘플링이 갖는 정보 손실(information loss)의 문제를 앙상블을 통해 극복하려고 했다는 점에서 학술적으로 의미가 있다. 본 모델을 두 가지 인공 데이터에 적용해 본 결과, 소수 범주에 속하는 데이터 수가 매우 적고, 데이터의 불균형이 심한 경우에도 예측 정확도와 안정성 관점에서 효과적임을 확인하였다.

이재동과 이지형[14]은 K-means 군집화 기법에 기반한 SVM 앙상블 기법을 연구하였다. 본 연구의 제안 모델은 각 분류기의 성능을 측정하여, 군집 별로 분류 결과 및 데이터 소속도를 총합하고, 이를 기반으로 최종 분류 결과를 산출하도록 설계되었다[17]. 모형의 성능을 검증하기 위한 실증 분석 결과, 제안하는 분류 방법이 기존의 패턴 인식 분류 알고리즘 보다 불균형 데이터를 분류함에 더 좋은 성능을 나타냄을 확인하였다.

김태훈과 안현철[15]은 언더 샘플링과 k-RNN을 결합한 새로운 하이브리드 기법을 기업 부도 예측에 적용하였다. 이 연구에서는 로지스틱 회귀분석, 판별분석, 의사결정나무, SVM을 통한 모델 비교연구를 진행하였는데, 연구결과 4가지의 모델 중에서 SVM을 통한 예측 정확도가 더 높았음을 확인하였다. 이 연구는 부도 예측 분야의 데이터 불균형 문제를 대상으로 한 선도적인 국내 연구라는 점에서 의의가 있으나, 언더 샘플링에 한정하여 연구모형이 설계되었다는 점에서 한계를 갖는다.

반면 해외에서는 일찍이 오버 샘플링의 가능성에 주목하고, 이와 관련된 연구가 다수 발표되었다. Japkowicz[16]는 가상 데이터를 중심으로 기본적인 언더 샘플링 기법과 오버 샘플링 기법을 비교하는 연구를 수행하였다. 이 연구에서 저자는 오버 샘플링 기법이 언더 샘플링 기법 보다 더 효과적이라는 사실을 증명하였다[16].

Chawla 등[17]은 k-NN 기법으로 소수 범주 데이터를 오버 샘플링기법을 통하여 데이터 불균형문제를 해결하는 SMOTE(Synthetic Minority Over sampling TEchnique)를 제안하였다. SMOTE는 오버 샘플링기법을 샘플링 기법이다. 소수 범주의 데이터를 단순한 복원 추출이 아닌, 이웃 데이터 사이에 보간법을 적용하여 데이터를 추출함으로써 데이터 불균형 문제를 해결하고자 하는 기법이다. SMOTE 기법은 일반적인 오버 샘플링 기법과는 달리 소수 범주에 속해있는 데이터 간 거리를 기준으로 비복원 추출을 통해, 오버 샘플링 기법이 가지고 있던 과적합화 단점을 극복한 새로운 샘플링 기법이다.

국내에서도 이재동과 이지형[18]이 최근 딥러닝 기법을 적용하여 데이터 불균형 문제를 해결하는 DAE(Deep-Auto Encoder)를 제안하였다. 이 연구에서는 소수 범주에 속해있는 데이터 분포를 학습하고, 학습된 분포를 따라 데이터를 생성하는 오버 샘플링 기법을 제안하였는데, 실증분석 결과 데이터 불균형 문제가 완화됨을 확인하였다.

이처럼 두 기법을 기반으로 많은 연구들이 발표되었지만, 일반적으로 보다 많은 정보를 학습에 활용할 수 있는 오버 샘플링 기법이 언더 샘플링 기법보다 분류를

위한 기계학습 알고리즘의 예측 정확도 개선에 더 효과적인 것으로 알려져 있다[19]. 이러한 배경에서 본 연구는 오버 샘플링 기법에 기반한 ROSE 표본추출 기법을 활용하여, 부도 예측 문제의 데이터 불균형 문제를 완화시키고, 이를 바탕으로 부도 예측 모형의 정확도를 제고해 보고자 한다.

III. 제안 알고리즘

전술했듯이 이분류 데이터에 심각한 불균형이 존재하면 모델 추정 및 정확도 평가 단계에서 성능을 저하시키는 영향을 줄 수 있다. 불균형이 심각할 경우, 분류규칙은 다수 범주에 의해 압도되어 소수 범주 데이터가 무시된 상태로 학습이 이루어지게 되며, 이 경우 분류모형의 결과는 신뢰할 수 없게 된다[20].

금융기관이나 신용평가회사에서 금융 리스크 관리 시 필수적으로 활용하게 되는 부도예측 모형의 경우도 부도 사례의 빈도보다 비부도 사례의 빈도가 월등히 높아 데이터 불균형 문제의 해소가 중요하다. 이러한 배경에서 본 연구에서는 새로운 오버 샘플링 기반의 표본추출 기법인 ROSE 기법을 부도예측에 적용하는 것을 제안한다.

ROSE 기법은 Menardi and Torelli[1]가 고안한 새로운 표본추출 기법으로서, 이분류 문제의 데이터 불균형 문제를 해소하는 가장 효과적인 대안으로 최근 각광받고 있다[21]. ROSE 기법은 데이터 불균형 학습의 모형 추정과 정확도 문제를 동시에 해결하는 통일된 틀을 제공하고, 이른바 부드러운 부트스트랩(smooth bootstrap)을 기반으로 새로운 데이터를 생성하여 데이터 불균형 문제를 완화시킨다. 이를 통해 오버 샘플링 기법을 기반으로 하면서도, 과적합화 위험을 줄이는 장점이 있다[1][21].

부트스트랩은 표집하는 절차를 반복(resampling)하여 높은 통계적 유의수준 하에서 주어진 표본의 성질을 이해하고자 하는 방법으로서, 전통적인 오버 샘플링의 이론적 토대가 되는 방법론이다[22]. 부트스트랩은 표본 크기의 한계와 통계기법이 가지고 있던 비현실성을 극복함으로써, 많은 분야에서 사용되어 왔다. 하지만 부트스트랩은 데이터 생산 추정치에 대한 확률오차를 고려하지 않아 통계적인 유의성을 검증할 수 없는 한계점이 있다[22]. 부드러운 부트스트랩을 적용하면 이와 같은 문제를 해결할 수 있는데, 이는 각각의 재표집된 관측치에 평균 0의 값을 갖는 (또한 보통 정규분포를 따르는) 작은 잡음을 더해주는 부트스트랩 방법이다. Menardi and Torelli[1]는 이를 오버 샘플링 기법에 응용하여, ROSE 기법을 고안하였다.

ROSE 기법이 새로운 가상의 학습 표본을 생성하는 과정은 다음과 같다. 아래 설명에서 독립변수 벡터 $\mathbf{x}$에 대해 $P(\mathbf{x}) = f(\mathbf{x})$는 확률밀도함수이고, n_j는 종속변수의 상태를 의미하는 y_j $(j = 0,1)$의 크기(표본수)를 의미한다. n_j는 전체 표본수 n에 대해 $n_j < n$을 만족하며, $\mathbf{T}_n$은 학습용 데이터셋을 지칭한다.

1단계: $\frac{1}{2}$의 확률로 $\mathrm{y}_j(j=0,1)$ 값을 갖는 y를 선택한다.

2단계: $p_i = \frac{1}{n_j}$의 확률로 $y_i = y$인 $(\mathbf{x}_i, y_i)$를 $\mathbf{T}_n$으로부터 선택한다.

3단계: $K_{\mathbf{H}_j}(\cdot, \mathbf{x}_i)$로부터 $\mathbf{x}$를 표본추출한다. 이 때, $K_{\mathbf{H}_j}$는 $\mathbf{x}_i$를 중심으로 하고, 규모 조정계수(scale parameter)는 행렬 $\mathbf{H}_j$에 종속하는 확률 분포이다.

결국 ROSE 기법은 두 범주 중 한 범주에 해당되는 관측값 하나를 학습용 데이터셋으로부터 추출하고, 그 관측값 주위에 있는 이웃(이웃을 결정하는 범위는 $\mathbf{H}_j$에 의해 결정)에서 새로운 학습사례를 생성하는 원리를 갖고 있다. 위의 1단계에서 언급했듯이, ROSE 기법이 새로운 학습사례를 생성할 때 특정 범주를 선택할 확률은 1/2이므로, 각 범주 별로 생성되는 학습 표본의 개수는 거의 동수가 된다. 물론 ROSE가 확률적으로 생성할 범주를 선택하는 방법을 택하고 있어서 정확하게 1:1의 비율을 갖는 것은 아니다. 하지만, 충분한 학습표본을

생성할 경우 거의 1:1의 비율로 수렴하게 된다. 이렇게 함으로써, ROSE는 과적합의 위험을 피하는 동시에 데이터 불균형 문제도 해결할 수 있게 된다.

ROSE 기법은 새로운 학습용 데이터를 기존 관측치로부터 생성하는 원리를 가진다는 점에서, Chawla 등[17]이 제안한 SMOTE와 유사한 면이 있다. 하지만, 소수 범주 관측치들의 선형 결합을 통해 새로운 학습용 데이터를 생성하는 SMOTE와 달리, ROSE는 범주에 관계없이 관측치 주변의 데이터를 학습용 데이터로 생성함으로써, SMOTE 보다 관측치에 덜 종속적인 특징을 갖는다.

ROSE 기법을 실제 데이터 및 시뮬레이션 된 데이터에 적용한 연구들에 따르면, ROSE는 전통적인 언더 샘플링 기법이나 오버 샘플링 기법들에 비해 더 우수한 성능을 보였다[1][21]. 하지만, 이렇게 우수한 성능에도 불구하고 지금까지 부도 예측 문제에 ROSE의 적용을 시도한 연구는 찾아보기 어려웠다. 이에 본 연구에서는 ROSE 기법을 부도 예측을 위한 표본추출 기법으로 제안하고, 그 성능을 검증해 보고자 한다.

특히 본 연구에서는 기존 연구[12]에서 기업 부도 예측에서 상대적으로 우수한 성능을 보이는 것으로 나타난 분류 기법인 SVM을 ROSE 기법과 결합하여 적용해 보고, 이것이 부도 예측 모형의 예측정확도와 모형 설명력의 개선이 이루어지는지 확인해 보고자 한다.

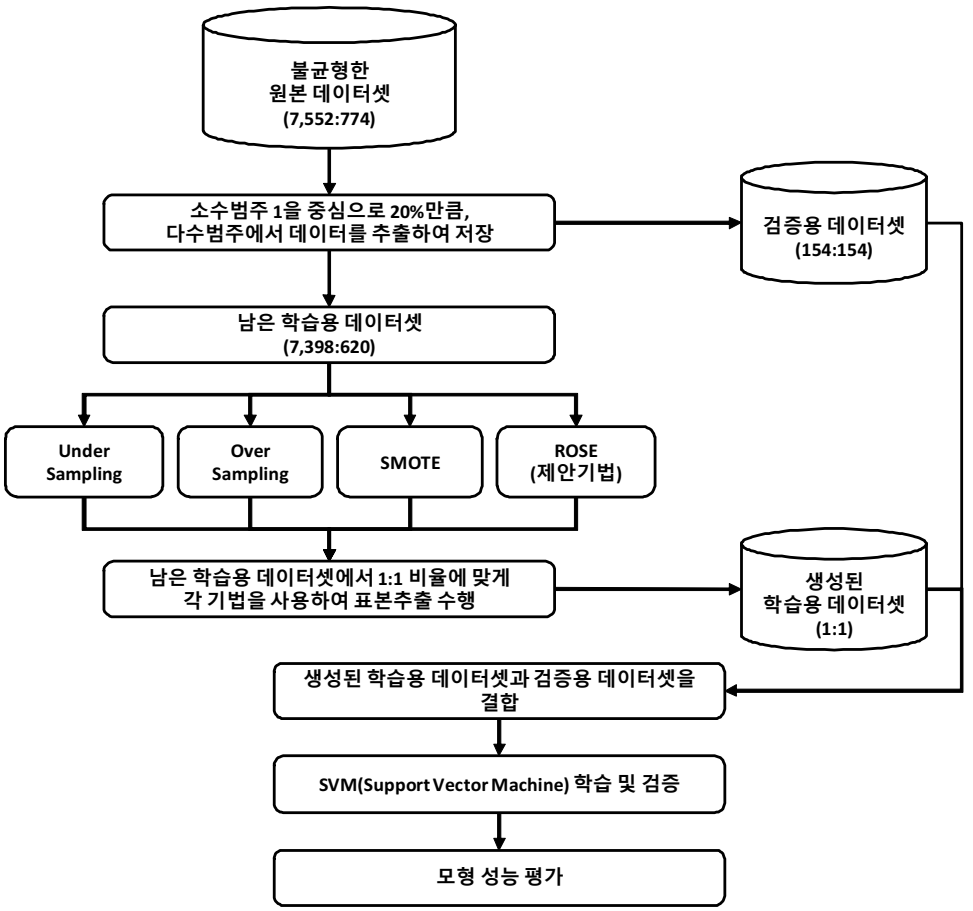

그림 1. 실증 분석 체계

IV. 실증 분석

본 연구에서 수행된 전체적인 실증 분석의 체계는 앞의 [그림 1]과 같다.

1. 실험 데이터

제안한 ROSE 기법의 유효성을 확인하기 위해 본 연구에서는 김태훈과 안현철[19]에서 사용했던 국내 한 대형 은행의 비외감 기업에 대한 부도 이력 데이터셋을 사용하였다. 본 실험에 사용한 데이터셋의 종속변수는 '부도 여부'로서, 0인 경우는 정상 기업을 1인 경우는 부도 또는 부채 비용 지불 기한이 3개월 이상 연체된 기업을 표시하였다. 부도 예측 모형 구축 시 산업의 영향을 통제하기 위해, 분석 대상을 중공업 기업으로 제한하였다. 실험 데이터는 774건(9.3%)의 부도 사례와 7,552건(90.7%)의 비부도 사례로 구성되어, 총 8,326건이 분석에 사용되었다.

실험에 앞서 데이터셋의 독립 변수가 많았기 때문에 전문가의 의견을 고려한 독립 표본 t 검정과 로지스틱 회귀분석에 의한 변수 선택을 사용하여 총 9개의 독립변수를 선정하였다. 최종 선정된 재무비율 변수들은 (1) 매출 변동률, (2) 매출원가, (3) 총자산에 대한 누적 이익, (4) 총 차입금 및 총자산 채권, (5) 현금 부채, (6) 영업 자본의 매출 변동성, (7) EBITDA 차입금, (8) 채권 상환 기간, (9)부채 현금흐름이다.

이렇게 완성된 전체 데이터셋은 소수 범주 데이터를 기준으로 8:2의 비율로 학습용 데이터셋과 검증용 데이터셋으로 구분되었다. 다음의 [표 2]는 각 데이터셋의 범주별 빈도를 정리하여 제시하고 있다.

표 2. 데이터셋 구분 (단위: 건)

종속변수	학습용 데이터셋	검증용 데이터셋
1(부도)	620	154
0(비부도)	7,398	154
합 계	8,018	308

2. 실험 설계

부도 예측에 사용할 분류 모형으로는 앞에서 설명한 바와 같이 이분류에 있어 가장 높은 정확도를 보이는 것으로 널리 알려져 있는 SVM을 적용하였다. 이러한 SVM의 예측 성능은 매개 변수에 의해 영향을 받지만, 커널 함수와 매개 변수를 결정하기 위한 절대적인 기준은 따로 제시되고 있지 않다. 이에 본 연구에서는 최적값을 찾기 위한 다양한 설정을 시도해 보고, 이 중 가장 최적의 매개변수 값을 찾아서 적용하였다. 구체적으로 고차원 사상(mapping)에 사용되는 커널 함수(kernel function)로는 선형, 다항식, 가우시안 RBF 함수가 적용될 수 있는데, 본 실험에서는 이 세 가지를 모두 적용해 보고 이 중 가장 높은 정확도를 산출하는 함수를 선택하였다. 그 밖에 SVM 성능에 매개 변수 C와 다항식 커널함수 사용 시 매개 변수 d, 그리고 가우시안 RBF 함수 사용 시 σ^2 도 영향을 미칠 수 있다[23]. 이에 C는 1, 10, 33, 55, 78, 100의 총 6가지 경우를, d는 1~5까지 5가지 경우를, 그리고 σ^2은 1, 25, 50, 75, 100의 5가지 경우를 모두 조합해 실험해 보고, 이 중 가장 우수한 결과를 보인 실험 결과를 선택하였다. SVM 실험은 오픈소스 실험도구인 LibSVM v3.21을 통해 수행하였다.

제안 알고리즘인 ROSE 기법의 성능을 정확하게 검증하기 위해, 본 연구에서는 전통적으로 많이 사용되어 온 표본추출 기법들인 기존 오버 샘플링(이하 Over), 기존 언더 샘플링(이하 Under), 그리고 SMOTE의 3가지 기법을 비교 알고리즘으로 사용하였다. 그림 1에 제시된 바와 같이, ROSE 및 비교 표본추출 알고리즘들은 검증용 데이터셋을 제외하고 난 후의 나머지 데이터, 즉 '남은 학습용 데이터셋'에만 한정하여 적용되었다.

알고리즘의 원리상, Over와 Under, 그리고 SMOTE의 경우에는 부도:비부도의 학습 데이터 비중이 정확하게 1:1로 조정되었다. 반면, 확률적으로 범주를 선택하여 학습용 데이터를 생성하는 ROSE의 경우에는 50.25:49.75의 비율로 데이터가 생성되었다. 다음의 [표 3]은 각 표본추출 기법 적용 후에 최종 생성된 학습용 데이터셋의 각 클래스별 빈도수가 어떤 비중으로 구성되어 있는지 구체적으로 나타내고 있다.

표 3. 표본추출 기법 적용 후의 학습용 데이터셋

표본추출기법	빈도수(비부도:부도)	비중
Under	1,240 (620:620)	1 : 1
Over	14,798 (7,399:7,399)	
SMOTE	3,096 (1,548:1,548)	
ROSE	8,019 (4,048:3,971)	50.25:49.75

3. 실험 결과

다음의 [표 4]는 제안 알고리즘인 ROSE와 3가지 비교 알고리즘을 적용했을 때의 SVM 예측 정확도를 종합적으로 제시하고 있다. 본 연구에서는 5-집단 교차검증(5-fold cross validation)을 수행하였는데, 그 결과 총 5개 데이터셋을 대상으로 한 각각의 실험결과와 전체를 종합한 결과가 해당 표에 함께 제시되어 있다.

실험 결과를 보면 ROSE 기법의 종합 예측정확도는 학습용 데이터셋에서 76.35%, 검증용 데이터셋에서 70.58%로서, 다른 표본추출 기법들과 비교했을 때 가장 우수한 예측정확도를 보임을 확인할 수 있다. 특히 학습용에서 괄목할 성과를 보임을 알 수 있는데, 이는 ROSE가 상당히 유의미한 학습용 표본을 생성하는 증거라 할 수 있다.

이어서, [표 4]에 제시된 알고리즘 간 예측정확도의 차이가 통계적으로 유의한 지 확인하기 위해, 이표본 비율 검정(two-sample test for proportions)을 수행하였는데, [표 5]가 그 결과를 제시하고 있다. 이 표를 통해 확인할 수 있듯이, ROSE는 Over, Under는 물론 유사한 원리의 SMOTE에 비해서도 99% 신뢰수준 하에서 통계적으로 유의한 성능 차이를 보이고 있음을 알 수 있다.

표 5. 이표본 비율 검정 수행 결과(Z-value)

	Under	SMOTE	ROSE
Over	−2.970*	−3.878*	−10.138*
Under		−3.920*	−7.203*
SMOTE			−3.298*

*99% 신뢰수준 하에서 통계적으로 유의

보다 정밀하게 제안 알고리즘의 성능을 검증하기 위해서, 본 연구에서는 예측 정확도 외에 민감도와 특이도를 종합적으로 고려하여 이분류 모형의 성능을 확인하는 ROC 곡선 분석을 추가로 수행하였다. 다음의 [그림 2]에서 각 알고리즘 별 ROC(Receiver Operating Characteristic) 곡선과 AUC(Area Under ROC Curve) 값을 함께 제시하고 있다.

이 그림에서 알 수 있듯이, ROSE 기법의 ROC 곡선은 다른 모형들에 비해 좌측 모서리에 더 가까이 붙어 있어, 판별능력이 상대적으로 더 우수함을 알 수 있다. 또한 이를 수치화 한 ROSE 알고리즘의 AUC 역시 0.795로서, 비교 알고리즘인 Under(0.743), Over(0.750), SMOTE(0.762)에 비해 명백하게 더 큰 값을 가짐을 알 수 있다. 이상 실험결과를 통해, 기업 부도예측에 있어 ROSE와 SVM을 적용했을 때, 우수한 예측 성과를 산출함을 확인할 수 있다.

표 4. 종합 실험 결과

표본추출 알고리즘	1		2		3		4		5		평균	
	학습용	검증용	학습용	검증용	학습용	검증용	학습용	검증용	학습용	검증용	학습용	검증용
Under	75.24%	71.10%	70.48%	68.83%	76.13%	71.75%	70.24%	66.23%	71.53%	68.51%	72.73%	69.29%
	RBF,C=10,σ^2=1		RBF,C=100,σ^2=25		RBF,C=10,σ^2=1		RBF,C=100,σ^2=25		RBF,C=1,σ^2=1			
Over	72.99%	71.75%	68.26%	67.86%	72.41%	71.75%	69.86%	68.18%	70.50%	69.16%	70.80%	69.74%
	RBF,C=1,σ^2=1		RBF,C=10,σ^2=100		RBF,C=1,σ^2=1		다항식,C=78,d=3		RBF,C=100,σ^2=25			
SMOTE	71.98%	71.43%	69.52%	69.16%	77.38%	71.75%	69.52%	67.53%	81.09%	69.81%	73.90%	69.94%
	RBF,C=1,σ^2=1		RBF,C=55,σ^2=25		RBF,C=10,σ^2=1		RBF,C=100,σ^2=100		RBF,C=33,σ^2=1			
ROSE	75.19%	71.10%	68.92%	67.86%	82.17%	75.00%	69.98%	69.16%	85.51%	69.81%	76.35%	70.58%
	RBF,C=1,σ^2=1		RBF,C=55,σ^2=50		RBF,C=33,σ^2=1		다항식,C=10,d=2		RBF,C=100,σ^2=1			

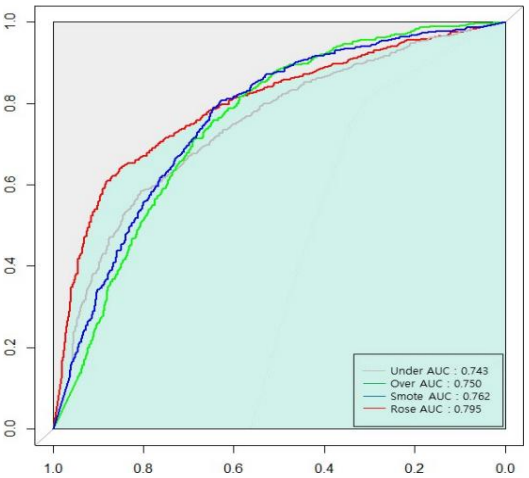

그림 2. ROC 곡선 분석 결과

V. 결론

본 연구에서는 현실에서 종종 발생하는 데이터 불균형 문제를 해결하고자 개발된 새로운 표본추출 기법인 ROSE 기법을 기업 부도예측 문제의 데이터 불균형 문제 해결을 위한 대안으로 제시하고, 이러한 ROSE 기법이 대표적인 이분류 기계학습 기법인 SVM과 결합되었을 때 전통적으로 사용되어 온 기존 표본추출 기법들보다 더 나은 성과를 도출하는지 실험을 통해 검증하였다. 비교 표본추출 알고리즘으로는 언더 샘플링, 오버 샘플링, 그리고 SMOTE를 사용하였는데, 실증 분석 결과 본 연구에서 제안한 ROSE 기법이 다수의 평가지표상으로 볼 때 가장 우수한 성능을 보임을 확인하였다.

데이터 불균형문제의 효과적인 학습을 위해 컴퓨터가 무작위로 추출하여 비율을 일정하게 만드는 일반적인 표본추출 기법들은 데이터의 손실, 중복 등 본질적으로 여러 단점을 안고 있다. 하지만 본 연구에서 사용한 ROSE 기법은 기존의 표본추출 기법들과 달리 새로운 학습 데이터를 창조적으로 생성함으로써 보다 효과적으로 데이터 불균형문제를 해결한다. 그리고 이러한 ROSE 기법의 장점이 부도 예측 문제에도 유효함을 본 연구를 통해 실증적으로 확인할 수 있었다.

학술적 측면에서 본 연구는 기업 부도예측 모형의 데이터 불균형 문제를 해결하기 위한 대안으로 ROSE를 제안하고, 그 적용 가능성을 실증분석을 통해 최초로 확인한 연구라는 점에서 그 의의를 갖는다. 본 연구를 계기로 하여, 앞으로 부도 예측 뿐 아니라 데이터 불균형이 발생할 수 있는 경영, 경제를 포함한 다양한 사회과학 분야의 예측 모형 개발에 ROSE의 적용이 검토될 수 있을 것으로 기대된다.

실무적 관점에서 본 연구는 위험 관리를 위해 부도예측 모형을 적극적으로 개발, 활용하고 있는 각종 금융기관 및 신용평가기업들에게 데이터 불균형 문제를 해결할 새로운 대안을 제시하고 있다는 점에서 의의를 갖는다. 특히 최근 빅데이터 및 인공지능 시대의 도래로 인해, 딥러닝과 같이 성능은 우수하지만 학습에 많은 양의 데이터를 필요로 하는 기계학습 알고리즘들이 선보이고 있는 상황에서, 부족한 소수 범주 데이터의 빈도를 효과적으로 늘려줄 수 있는 ROSE의 활용가치는 더 높아질 수 있을 것으로 기대된다.

이처럼 학술적, 실무적으로 유익한 발견이 본 연구를 통해 이루어졌지만, 본 연구의 결과는 다음과 같은 한계점을 가지고 있다.

첫째, 본 연구에서는 분류 알고리즘으로 SVM 하나만을 사용한 점을 지적할 수 있다. 물론 SVM이 부도예측을 대상으로 한 여러 기존 연구에서 가장 우수한 성능을 보인 검증된 기법이지만, 진정 ROSE가 다른 표본추출 기법들에 비해 더 우수한 지 검증하기 위해서는 타 기계학습 기법들에 대해서도 ROSE로 표본을 추출했을 때 더 우수한 성과를 보이는지 확인할 필요가 있다.

둘째, 본 연구에서 사용된 기업 부도 예측 데이터가 특정 산업군의 특정 대상(비외감 기업)으로 한정된 표본이라는 점 역시 또 다른 한계로 지적될 수 있다. 향후 연구에서는 보다 다양한 부도 예측 사례에 ROSE와 비교 알고리즘들을 적용해 보고, 다른 데이터에서도 ROSE가 더 우수한 성능을 보이는지 검증할 필요가 있다.

나아가 부도 예측 외에 다른 데이터 불균형 문제에 있어서도 ROSE가 괄목할만한 성과를 보이는지 확인할 필요가 있다. 앞서 언급했듯이 공학 분야의 여러 연구들에서 ROSE가 우수한 성능을 가진 표본추출 기법이라는 점이 검증된 바 있다. 그러나 경영학을 비롯한 기

타 인문사회과학 연구에서는 ROSE가 활용된 사례를 거의 찾아보기 어렵다. 때문에 부도 예측 외에 다른 경영/경제 혹은 인문사회과학 분야의 데이터 불균형 문제 해결에 ROSE를 적용해 보고, 그 성능을 확인하는 후속 연구가 추후 이루어져야 할 것이다.

참 고 문 헌

[1] G. Menardi and N. Torelli, “Training and assessing classification rules with imbalanced data,” Data Mining and Knowledge Discovery, Vol.28, No.1 pp.92-122, 2014.

[2] W. H. Beaver, “Financial ratios as predictors of failure, Journal of Accounting Research,” Vol.4, pp.71-111, 1966.

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

[4] J. A. Ohlson, “Financial ratios and the probabilistic prediction of bankruptcy,” Journal of accounting research, Vol.18, No.1, pp.109-131, 1980.

[5] M. E. Zmijewski, “Methodological issues related to the estimation of financial distress prediction models,” Journal of Accounting Research, Vol.22, pp.59-82, 1984.

[6] R. O. Edmister, “An empirical test of financial ratio analysis for small business failure prediction,” Journal of Financial and Quantitative Analysis, Vol.7, No.2, pp.1477-1493, 1972.

[7] M. D. Odom and R. Sharda, “A neural network model for bankruptcy prediction. In Proceedings of the International Joint Conference on Neural networks,” Vol.2, pp.163-168, 1990.

[8] K. Y. Tam and M. Y. Kiang, “Managerial applications of neural networks: the case of bank failure predictions,” Management Science, Vol.38, No.7, pp.926-947, 1992.

[9] C. Serrano-Cinca, “Self-organizing neural networks for financial diagnosis,” Decision Support Systems, Vol.17, No.3, pp.227-238, 1996.

[10] J. Yang and V. Honavar, “Feature subset selection using a genetic algorithm,” IEEE Intelligent Systems and their Applications, Vol.13, No.2, pp.44-49, 1998.

[11] 김경재, 한인구, “퍼지 신경망을 이용한 기업부도예측,” 지능정보연구, 제7권, 제1호, pp.135-146, 2001.

[12] 이영찬, “인공신경망과 Support Vector Machine의 기업부도예측 성과 비교,” 한국지능정보시스템학회 춘계학술대회논문집, pp.211-218, 2004.

[13] 강필성, 조성준, “데이터 불균형 해결을 위한 Under-Sampling 기반 앙상블 SVMs,” 대한산업공학회 춘계공동학술대회 논문집, pp.291-298, 2006.

[14] 이재동, 이지형, “데이터 불균형 문제 해결을 위한 K-means Clustering 기반 SVM앙상블 기법,” 한국정보과학회 한국컴퓨터종합학술대회 논문집, pp.297-799, 2014.

[15] 김태훈, 안현철, “A Hybrid Under-sampling Approach for Better Bankruptcy Prediction,” 지능정보연구, 제21권, 제2호, pp.173-190, 2015.

[16] N. Japkowicz, “The Class Imbalance Problem : Significance and Strategies,” In Proceedings of the International Conference on Artificial Intelligence, pp.111-114, 2000.

[17] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, Vol.16, pp.321-357, 2002.

[18] 이재동, 이지형, “데이터 불균형의 효과적인 학

습을 위한 딥러닝 기법," 한국지능시스템학회 춘계학술대회 학술발표논문집, 제25권, 제1호, pp.113-114, 2015.

[19] G. E. Batista, R. C. Prati, and M. C. Monard, "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data," ACM SIGKDD Explorations Newsletter, Vol.6, No.1, pp.20-29, 2004.

[20] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: one-sided selection," Proceedings of the Fourteenth International Conference on Machine Learning, pp.179-186, 1997.

[21] N. Lunardon, G. Menardi, and N. Torelli, *ROSE: A Package for Binary Imbalanced Learning,* r-project.org, 2014.

[22] B. Efron and R. Tibshirani, *An introduction to the bootstrap*, Chapman and Hall, 1993.

[23] F. E. J. Tay and L. J. Cao, "Modified support vector machines in financial time series forecasting," Neurocomputing, Vol.48, pp.847-861, 2002.

저 자 소 개

안 철 휘(Cheolhwi Ahn) 준회원

- 2016년 2월 : 전주대학교 정보시스템학과(공학사)
- 2017년 3월 ~ 현재 : 국민대학교 비즈니스IT전문대학원 비즈니스IT전공(석사)

<관심분야> : 비즈니스 애널리틱스

안 현 철(Hyunchul Ahn) 정회원

- 2006년 8월 : KAIST 테크노경영대학원 경영공학(박사)
- 2008년 3월 ~ 2009년 2월 : 성신여자대학교 경영학과
- 2009년 3월 ~ 현재 : 국민대학교 경영대학 부교수

<관심분야> : 비즈니스 애널리틱스, 추천시스템