

인공신경망 및 랜덤포레스트 기법을 활용한 기업 분식회계 탐지 성능 평가 연구

황동혁¹, 서영석¹

¹영남대학교 컴퓨터공학과

hdh0261@ynu.ac.kr, ysseo@yu.ac.kr

A Study on Accounting Fraud Detection using Neural Network and Random Forest

Dong-Hyeok Hwang¹, Yeong-Seok Seo¹

¹Dept. of Computer Science, Yeungnam University

요 약

ESG 경영이 중요해짐에 따라 기업의 분식 여부도 중요해졌다. 따라서 본 논문에서는 인공신경망과 랜덤포레스트를 활용하여 기업의 분식회계 여부를 판단 성능을 비교분석하고 그 유용성에 대해 평가하였다. 실제 기업 회계정보를 수집하여 실험을 수행하였고, 실험 결과 F1-Score 기준 랜덤포레스트의 RFECV 기법이 0.81로 분식 기업을, SMOTE 기법을 사용한 모델이 정상 기업을 탐지하였고 Accuracy 기준 랜덤포레스트의 RFECV 기법과 SMOTE 기법을 사용한 모델이 0.77로 가장 효과적인 탐지 성능을 보여주었다.

1. 서론

최근 ESG 경영의 중요성이 대두되고 있는 가운데 기업들의 분식회계의 여부도 중요해지고 있다.

분식회계란 회계장부를 고의로 꾸며 기업의 자산과 이익을 부풀리거나 부채를 줄여 장부를 조작하는 행위를 의미한다. 분식회계가 발생하면 투자자와 채권자는 투자하기 위한 정확한 판단을 내리지 못하게 된다. 이러한 행위를 막지 않고 방치한다면 투자자나 채권자는 더 이상 기업을 믿지 못하게 되어 ESG 경영에 큰 걸림돌이 된다.

따라서 본 논문에서는 기계학습의 대표기법 중 인공신경망(NN)과 랜덤포레스트(RF)를 활용한 데이터 분석을 통해 기업의 분식회계 여부 탐지 성능에 대해 비교분석을 하고 그 유용성에 대해 평가해보고자 한다. 실제 분식회계를 수행한 기업들의 회계정보를 수집하여 체계적인 분석을 진행함으로써 보다 현실적인 연구 결과를 제공하고자 하였다.

2. 실험

2.1. 데이터 수집 및 전처리

본 연구의 실험에서 데이터 셋은 금융감독원의 회계감리결과제계결과 및 전자공시시스템의 공시 중 과대계상에 해당하는 기업의 회계정보를 수집하였

다.[1] 2000년부터 2020년까지 207건의 분식회계 사례와 207건의 정상 회계 사례를 데이터 셋으로 사용하였으며, 데이터 셋의 입력변수는 선정된 기업들의 재무제표의 통일성과 결측값을 고려하여 435개로 선정하였다.

수집된 데이터 셋은 기업별 금액의 규모 차이에 의한 데이터 편차가 크기 때문에 MinMaxScaler 정규화 함수[2]를 적용하여 0과 1 사이의 소수값으로 변환하는 전처리 작업을 거친다.

2.2. 교차 검증을 위한 데이터 구성

전체 데이터 중 70%는 훈련 셋, 20%는 모델의 평가를 위한 시험 셋, 나머지 10%는 학습이 잘 이루어졌는지 검증하기 위한 검증 셋으로 구성하였다.

2.3. 모델 실험 및 결과

본 소절에서는 연구에서 사용한 NN과 RF 기법의 하이퍼파라미터 및 기법별 실험 결과들을 소개한다. 표 1은 NN과 RF의 5-fold 교차 검증을 수행하여 가장 우수한 성능을 보이는 하이퍼파라미터를 선발한 결과이다. 우수한 하이퍼파라미터는 볼드체로 표시하였으며 그 아래의 값은 사용한 하이퍼파라미터의 범위를 뜻한다.

<표 1> 클래스별 하이퍼파라미터

class	Unit		Activation Function	Loss Function	
NN	512, 512, 128, 32, 1		relu, sigmoid	binary crossentropy	
	estimators	max depth	max features	min samples leaf	min samples split
RGS	300 [100,200,300]	300 [100,200,300]	100 [100,200,300,400]	1 [1,2,3]	2 [1,2,3]
RRS	450 [400, 450, 500, 550]	317 range(100, 400)	270 range(100, 400)	1 [1,2,3,4,5]	2 [1,2,3,4,5]
	the number of selected data				
RRF	281				
RSM	302				

NN은 총 5개의 수준으로 구성하였으며 활성 함수로는 relu와 sigmoid, 손실 함수로는 binary crossentropy를 사용하였다. RF Grid Search(RGS)의 estimators는 100, 200, 300중 300, max depth는 100, 200, 300중 300, max features는 100, 200, 300, 400중 100, min samples leaf는 1, 2, 3중 1, min samples split은 1, 2, 3중 2를 사용하였다. RF Randomized Search(RRS)의 estimators는 400, 450, 500, 550중 450, max depth는 100과 400 사이에서 317, max features는 100과 400 사이에서 270, min samples leaf는 1과 5 사이에서 1, min samples split은 1과 5 사이에서 2를 사용하였다. RF RFECV(RRF)의 최적 입력변수의 개수는 281개로 선정하여 사용하였다. RF SMOTE(RSM)의 훈련 셋의 크기는 302로 선정하여 사용하였다.

다음 표 2는 위와 같은 모델들을 활용하여 5-fold 교차 검증 결과의 평균 성능 결과이다. 성능 평가지표로는 정밀도(Precision), 재현율(Recall), F1-Score와 정확도(Accuracy)를 사용했으며 가장 높은 결과를 보이는 수치를 볼드체로 표시하였다.

<표 2> 기법별 성능

class		Precision	Recall	F1-score	Accuracy
NN	Normal	0.73	0.73	0.73	0.73
	Fraud	0.73	0.74	0.73	
RGS	Normal	0.74	0.67	0.71	0.71
	Fraud	0.68	0.75	0.71	
RRS	Normal	0.76	0.65	0.70	0.71
	Fraud	0.67	0.78	0.72	
RRF	Normal	0.77	0.67	0.72	0.77
	Fraud	0.77	0.85	0.81	
RSM	Normal	0.80	0.74	0.77	0.77
	Fraud	0.74	0.80	0.77	

NN의 경우 평균 F1-Score는 0.73으로 정상 기업을, 0.73으로 분식 기업을 탐지하였다.

RF의 경우 RSM 기법을 적용한 RF의 F1-Score가 0.77로 정상 기업을 가장 잘 탐지하였고 RRF 기법의 F1-Score가 0.81로 분식 기업을 가장 잘 탐지하였다. 정확도 측면에서는 RRF 기법과 RSM 기법의 정확도가 동일한 0.77의 성능을 보였다.

성능지표로 바라보았을 때 정밀도는 RSM이 0.80로 정상 기업을, 재현율은 RSM이 0.80으로 분식 기업을, F1-Score은 RRF가 0.81로 분식 기업을, 정확도는 RRF, RSM가 동일한 0.77로 가장 높은 결과를 보여주었다.

본 실험 결과 NN에 비해 RF가 F1-Score 기준에서 0.04 정도의 향상된 결과를 보여주었다. 또한 RF 기법 내에서도 데이터 처리 방법에 따라 RGS와 RRS에 비해 RRF와 RSM을 적용하였을 시 F1-Score 기준에서 0.06의 정도 분식회계 기업 식별에 더 높은 결과를 보여주었다.

3. 결론 및 향후 연구

본 논문에서는 NN과 RF 기법을 활용하여 실제 기업의 분식회계를 보다 정확히 탐지하기 위한 최적의 모델을 평가하였다. 모델의 성능 평가를 위해 정상 기업의 실제 회계장부를 수집하였으며 MinMaxScaler 정규화 함수로 전처리를 실행했다. 성능을 평가하기 위해 정밀도, 재현율, F1-Score와 정확도를 사용하였으며 실험 결과 F1-Score 기준 RRF 기법과 RSM 기법이 F1-Score가 0.77로 동일하게 가장 높은 성능을 보였다. 반면 NN은 0.73으로 RF보다 낮은 성능의 경향을 보여주었다.

향후에는 보다 더 많은 분식 기업의 회계정보를 수집하여 본 연구결과의 일반화를 위해 추가적인 분석을 수행할 예정이다.

Acknowledgement

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2023R1A2C1008134).(Corresponding author: Yeong-Seok Seo).

참고문헌

- [1] 전자공시시스템, <https://dart.fss.or.kr/>, Available on Apr 16, 2023.
- [2] scikit-learn, <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html>, Available on Apr 16, 2023.