Modified Bagging Predictors를 이용한 SOHO 부도 예측

SOHO Bankruptcy Prediction Using Modified Bagging Predictors

김승혁[a] 김종우[b]

[a] 한양대학교 대학원 경영학과
서울 성동구 행당동 17번지
Tel: +82-2-2220-1067, Fax: +82-2-2220-1169, E-mail:fanta0507@naver.com

[b] 한 양대학교 경영대학 경영 학부(교신저자)
서울 성동구 행당동 17번지
Tel: +82-2-2220-1067, Fax: +82-2-2220-1169, E-mail:kjw@hanyang.ac.kr

Abstract

본 연구에서는 기존 Bagging Predictors에 수정을 가한 Modified Bagging Predictors를 이용하여 SOHO에 대한 부도예측 모델을 제시한다. 대기업 및 중소기업에 대한 기업부도예측 모델에 대한 많은 선행 연구가 있어왔지만 SOHO만의 기업부도예측 모델에 관한 연구는 미비한 상태이다. 금융기관들의 대출 심사 시 대기업 및 중소기업과는 달리 SOHO에 대한 대출심사는 아직은 체계화 되지 못한 채 신용정보점수 등의 단편적인 요소를 사용하고 있는 것이 현실이고 이에 따라 잘못된 대출로 인한 금융기관의 부실화를 초래할 위험성이 크다. 본 연구에서는 실제 국내은행의 SOHO 데이터 집합이 사용되었다. 먼저, 기업부도예측 모델에서 우수하다고 연구되어진 인공신경망과 의사결정나무 추론기법을 적용하여 보았지만 만족할 만한 성과를 이끌어내지 못하여, 기존 기업부도예측 모델 연구에서 적용이 미비하였던 Bagging Predictors와 이를 개선한 Modified Bagging Predictors를 제시하고 이를 적용하여 보았다. 연구결과, SOHO 부도예측에 있어서 본 연구에서 제시한 Modified Bagging Predictors가 인공신경망과 Bagging Predictors 등의 기존 기법에 비해서 성과가 향상됨을 알 수 있었다. 제시된 Modified Bagging Predictors의 유용성을 확인하기 위해서 추가적으로 다수의 공개 데이터 집합을 활용하여 성능을 비교한 결과 Modified Bagging Predictors가 기존의 Bagging Predictors에 비해 일관적으로 성과가 향상됨을 알 수 있었다.

Keywords:

기업부도예측; 데이터마이닝; Bagging Predictors; 인공신경망; 의사결정나무

1. 서론

외환 위기 를 전후 로 기업의 규모와 상관없이 많은 기업이 부도의 위기를 겪게 되었고, 이에 따라 많은 국내 금융기관들이 대출금을 회수하지 못하는 불상사가 일어나 국내 경제가 급속히 와해되고 경제난이 가속화되기에 이르러, 국내 경제는 한치 앞을 바라 볼 수 없는 상황에 처해졌었다. 이렇듯 기업에 대한 정확한 부도 예측 모델의 구축은 해당 기업이나 금융기관뿐만 아니라 일반 국민에게 있어서도 중요한 사안이라 말할 수 있을 것이다. 기업에 대한 부도 예측 모델에 관한 연구는 Beaver(1966)와 Altman(1968)등에 의해 처음 시작된 이후 국내외적으로 많은 연구가 있어왔다. 최근 들어서는 인공신경망(Artificial Neural Networks)등의 인공지능 기법을 이용한 기업 부도 예측 모델이 기존의 통계분석에 비해서 우수함이 많은 연구들을 통해서 입증이 되어 기업 부도 예측 모델에 많이 활용되고 있다[18,20,21,22].

하지만 기존의 국내외 기업 부도 예측 모델에 관한 연구는 재무재표 데이터를 중심으로 하는 대기업에 대한 연구에 치중되어 중소기업에 대한 연구가 상대적으로 미비한 편이다. 물론 중소기업의 재무재표 데이터 및 비재무에 관련된 데이터가 신뢰성면에서 대기업에 비해 떨어지는 특징이 있기는 하지만, 중소기업이 국가경제에서 차지하는 비중이 결 코 작지 않 음을 생각하면 아쉬운 부분으로 남고 있다. 특히 SOHO만의 기업 부도 예측 모델에 대한 연구는 전무하다시피한 작금의 상황에서 이에 대한 연구는 필요하다.

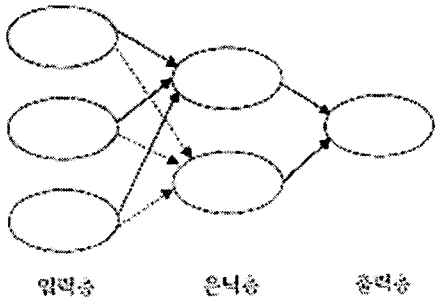
SOHO란 'Small Office Home Office'의 약자로서,

우리나라 전체 사업자수의 99.8%(약 300만)를 차지하는 중소기업 중, 90%(약 270만)를 차지하는 소규모 자영업을 말하는데, 금융기관들의 대출 심사 시 대기업 및 중소기업과는 달리 SOHO에 대한 대출심사는 아직은 체계화 되지 못한 채 부도확률, 등급, 종합점수, 개요정보점수, 실적정보점수, 신용정보점수 등의 단편적인 요소를 사용하고 있는 것이 작금의 상황이고 이에 따라 잘못된 대출로 인한 금융기 관의 부실화를 초래할 위험성 이 크다.

이에 본 연구에서는 SOHO 기업만의 부도 예측 모델을 구축하여 향후 기업부도 연구에 유용한 결과를 도출하고자 하였다. 연구에 쓰인 방법으로는 기업 부도 예측 연구에서 우수하다고 입증된 인공신경망을 비롯하여 의사결정나무(Decision tree)의 한 종류인 CART와의 성과 비교뿐 아니라, 기존의 기업 부도 예측 연구에서 적용이 미비하였던 Bagging Predictors와 이를 일부 수정한 Modified Bagging Predictors를 통한 기업 부도 예측 모델을 구축하여 성과 비교 연구를 하였다. 또한 새롭게 제시된 Modified Bagging Predictors의 신뢰성을 확보하기 위하여 SPSS 교육용 데이터 집합과 유선통신에 관련된 데이터집합, 또한 UCI 데이터 저장소[23]에서 다수의 데이터 집합을 활용하여 추가로 성과를 비교하 였다.

2. 기업 부도 예측 모델에 관한 문헌연구

기업의 부도 예측 모델에 관한 연구는 경영학 분야에서 국내외적으로 활발하게 연구되어 왔으며 다양한 통계 기법과 최근 들어 의사결정나무 및 인공신경망, 유전자 알고리즘(Genetic Algorithm) 등의 인공지능 기법이 적용 연구되고 있다. 통계 기법을 이용한 부도 예측 연구에는 Raja et al.(1980)에서의 판별분석 사용, Gentry and Whitford(1985)에서의 판별분석, 로짓분석, 프로빗분석의 사용, Gombola and Ketz(1983)에서의 요인분석 등의 연구가 있으나 선형적 통계기법의 한계로 1980년대 후반부터는 인공신경망, 유전자 알고리즘과 같은 인공지능 기법들이 기업 부도 예측 모델에 많이 사용되고 있다. Odom and Sharda(1990)는 판별분석과 인공신경망의 성과를 비교하여 인공신경망의 우수성을 증명하였으며, Tam and Kiang(1992)는 판별분석, 로지스틱 회귀분석, k-최근접이웃방법(k-nearest neighbor), 귀납적 추론(ID3)과 인공신경망과의 성과 비교를 통해 인공신경망의 우수함이 연구되어졌다. 이건창·김명종·김혁(1994)은 MDA(Multiple Discriminant Analysis), 귀납적 학습방법, 인공신경망과의 성과를 비교하였으며, 이재식·한재홍(1995)은 인공신경망을 이용하여 중소기업의 도산예측에 있어서, 재무정보를 보완할 수 있는 비재무정보의 유용성을 검증하였고, 신경식(2000)은 입력 변수군을 달리하는 다수의 인공신경망 모델을 구축하고 통합하여 예측력 향상을 이끌어내었으며, 김진백·이준섭(2000)은 인공신경망과 사례기반추론(Case-Based Reasoning : CBR) 기법을 사용하여 모델을 개발하고 현금흐름지표가 기존에 주로 사용된 일반재무비율 변수에 근거한 부실 예측 모델에 추가적인 역할을 할 수 있는지를 평가하였고, 김경재·한인구(2001)는 기존 신경망에 퍼지집합의 개념을 적용하여 신경망 학습에 사용될 자료를 퍼지화하고 이를 신경망에 학습시켰으며, 홍승현·신경식(2003)은 유전자 알고리즘을 이용하여 입력 변수군을 도출하여 인공신경망을 적용한 연구를 하였다.



<그림 1> 기본적인 지도학 습 인공신경망의 구조

3. 연구기법

3.1. 인공신경망(Artificial Neural Networks)

인공신경망은 인간의 뉴런(Neuron) 개념에서 아이디어를 얻어 개발된 인공지능 기법이다. <그림 1>은 가장 일반적인 Feed-Forward 인공신경망의 지도학습(Supervised learning)의 구조이 다. 인공신 경망 알고리즘은 은닉층의 활성함수에 따라 MLP(Multi-Layer Perception: 다층 퍼셉트론)와 RBF(Radial Basis Function: 방사형 기저 함수)로 구분된다. 입력층을 통해서 들어온 각종 데이터들은 은닉층의 입구에서 선형 결 합으로 연결되 고 이 선형 결합의 값이 커질수록 뉴런(Neuron)이 활성화되고 반대의 경우 비활성화가 된다. 이 활성화 값의 범위를 제한하기 위한 함수를 활성함수라고 한다. 은닉층에서 마지막으로 출력층으로 신호를 보내고, 신호를 받은 출력층에서 이들을 결합하여 최종적으로 반응 결과를 생성하게 된다. 은닉층에서는 학습률(learning rate)과 모멘트(moment) 값을 기반으로 활 성 함수에 나온 값들을 계 속적으로 목표(Target)값과 비교하여 가중치를 변경시켜 나가는데 이러한 반복적인 과정을 역전파 알고리즘(Back Propagation algorithm)이라고 한다 [8].

3.2. 의사결정나무 (Decision Tree)

의사결정나무분석은 예측과 분류를 위해 보편적이고 강력한 성능을 보이며 나무구조로 규칙을 표현하기 때문에 이해하기가 쉽다. 의사결정나무분석의 나무형성 알고리즘은 다양하지만, 가장 보편적인 것으로 CART (classification and regression trees)와 CHAID(chi-squared automatic interaction detection)이고, 좀 더 새로운 알고리즘은 C4.5나 C5.0을 들 수 있다. 본 연구에서는 이중 CART를 사용하여 연구를 진행하였다. CART는 1984년 Breiman에 의해 발표되어 machine-learning 실험의 시초가 되고 있다. CART는 이진분리를 하는 의사결정나무 추론 알고리즘이다. 이진분리라는 것은 부모 노드로부터 자식노드를 분리 할 때 항상 2개로만 분리를 한다 는 것을 말한다. 이런 이진 분리는 여러 개의 분리가 되는 것에 비해 정확도 등이 떨어질 수 있으나 반대로 가지 분류가 비교적 간단하여 해석이 편하다는 장점이 있기도 하다. CART는 두가지의 특징이 있다. 먼저 목표 변수에 연속형의 변수를 사용 할 수 있다. C5.0의 경우 오직 범주형 변수만이 목표 변수 로 오는 것에 비하면 모델 적용에 확장성 이 더 크다는 것을 알 수 있다. CART는 Gini 지수를 이용하여 데이터의 불순도를 측정, 분류를 하게 된다 [8].

3.3. Bagging Predictors

Bagging이란 여러 개의 predictor를 만들어 이것을 이용하여 통합 predictor를 얻어내는 방법을 말한다. 다시 말하면, 지도학습에서 주어진 훈련용 데이터를 복원추출(bootstraping : resampling randomly with replacement)하여 여러 개의 훈련용 데이터 집합으로 만들고, 각각의 데이터 집합에 대하여 모델링을 하여 이를 결합하는 방법이라고 할 수 있다.

전체 훈련용 데이터 집합을 $L=\{(y_n,x_n),n=1,\cdots,N\}$ 이라 하면, 여기서 y_n은 목표변수 값, x_n은 입력변수 벡터이다. y를 목표 변수라 하면, y는 범주형거나 수치형일 수 있다. 또한 $\varphi(x,L)$ 는 훈련용 데이터 집합 L을 사용해서 생성된 p redictor에 x 입 력변수 벡터에 대한 예측값을 의미한다. $\{L_k\}$ 는 L과 같은 분포에서 뽑힌 N개의 독립적인 관측치로 구성된 훈련용 데이터 집합의 순열(실제로는, L에서의 bootstrap sample을 사용)이라고 하자.

Bagging의 목적은 $\{L_k\}$를 이용하여 $\varphi(x,L)$ 보다 좋은 predictor를 얻는 것이다. y가 수치형일 경우 $\varphi(x,L)$ 를 $\varphi(x,L_k)$ 의 평균으로 대치시키고, y가 범주형일 경우에는 투표방식(voting)을 적용한다. L에서의 작은 변화가 φ 에서 큰 변화를 가져올 때는 Bagging이 효과적이고 unstable한 모델에 유용하며, stable한 모델 에는 효과적이지 못하다 [13]. 요약하자면, Bagging은 예측력 및 정확도의 향상을 위해서 하나의 데이터 집합으로부터 한 가지 로직만을 추출하는 것이 아닌 다양한 로직을 추출하고 이를 결합하여 오분류된 부분을 보강하고, 이를 통한 예측력 및 정확도의 향상을 가져오는 방법이다.

4. 실험설계

4.1. Modified Bagging Predictors의 개념

다음은 Modified Bagging Predictors에 대한 개념에 대한 설명이다. 데이터 마이닝 작업 시 임의의 데이터 집합이 10개의 레코드를 가지고 있고, 각각 랜덤하게 샘플링된 데이터를 임의의 기법(예를 들면 의사결정나무 추론 기법 중 하나인 CART)을 사용하여 만들어진 A, B, C, D, E의 5개 모델이 있다고 가정하자. 그리고 각각의 모델을 시험용 데이터를 이용하여 측정한 결과 모델 A의 예측 정확도가 70%(7개) 그리고 오분류(error rate) 또는 예측을 잘못한 것이 30%(3개)였고 나머지 B, C, D, E 모델은 70%, 60%, 60%, 50%의 예측 정확도를 가지고 있다고 가정한다. 이와 같은 결과만 놓고 봤을 때, 모델A과 모델B, 모델D와 모델E는 같은 예측 정확도를 가지는 성능을 가진다고 생각될 수 있을 것이다. 하지만 다음 <그림 2>와 같은 경우처럼 레코드 하나하나를 자세히 살펴보면 반드시 그렇지만은 않다는 것을 알 수 있다.

Bagging

Modified Bagging

번호	실제값	모델 A (70% 예측 정확도)	모델 B (70% 예측 정확도)	모델 C (60% 예측 정확도)	모델 D (50% 예측 정확도)	모델 E (50% 예측 정확도)	Bagging (80% 예측 정확도)	Modified Bagging (90% 정확도)
1	Yes	Yes	Yes	No	No	Yes	Yes	Yes
2	Yes	No	No	No	No	Yes	No	No
3	Yes	Yes	Yes	Yes	No	No	Yes	Yes
4	Yes	Yes	Yes	No	No	No	No	Yes
5	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes
6	No	No	No	Yes	No	Yes	No	No
7	No	No	Yes	No	Yes	No	No	No
8	No	No	Yes	No	No	No	No	No
9	No	Yes	No	No	No	Yes	No	No
10	No	Yes	No	No	No	No	No	No

<그림 2> Bagging과 Modified Bagging의 비교1

<그림 2>에서 보면 모델A과 모델 B는 70%의 같은 예측 정확도를 가지고 있지만 두 모델 둘다

틀린 부분(번호 2번)이 있고, 모델A는 맞추고 모델B는 틀린 경우(번호 7,8번), 모델A는 틀리고 모델B는 맞춘 경우(번호 9,10)가 있다. 위와 같은 경우를 가지고 추론해보면 각각의 모델이 예측 정확도의 높고 낮음에 상관없이 같은 사례라도 올바르게 예측하는 경우가 있는가 하면 틀리게 예측하는 경우가 있을 수 있다는 것이다. 만약 <그림 2>에서 각각의 모델이 올바르게 예측한 것만을 선택하게 된다면 이론적으로는 100%의 예측 정확도를 가질 수 있을 것이다.

<그림 2>에서 다섯 개의 모델을 Bagging Predictors, 즉 voting 기법을 사용하여 가장 많이 득표한 값을 채택 하게 된다면 <그림 2> 우측 Bagging 예측 정확도에서 보듯이 80%로 예측의 정확도가 올라가게 된다. 하지만 예를 들어 <그림 2>에서 5개 모델 중 예측 정확도가 60%이상인 모델 A, B, C만을 선택하여 voting하게 된다면(Modified Bagging Predictors), <그림 2> 가장 우측부분의 결과에서 보듯이 90%의 예측 정확도를 가질 수 있는데 이는 기존 다섯 개의 모델에 비해 예측도가 향상되는 결과를 볼 수 있으며 Bagging Predictors에 비해서도 10%의 예측 정확도 향상을 가져오는 것을 볼 수 있다.

번호	실제값	Bagging (모델 5개 사용하여 voting)				Modified Bagging (모델 3개 사용하여 voting)			
		Yes개수	No개수	최다득표	맞음/틀림	Yes개수	No개수	최다득표	맞음/틀림
1	Yes	3	2	Yes	맞음	2	1	Yes	맞음
2	Yes	1	4	No	틀림	0	3	No	틀림
3	Yes	3	2	Yes	맞음	3	0	Yes	맞음
4	Yes	2	3	No	틀림	2	1	Yes	맞음
5	Yes	4	1	Yes	맞음	3	0	Yes	맞음
6	No	2	3	No	맞음	1	2	No	맞음
7	No	2	3	No	맞음	1	2	No	맞음
8	No	1	4	No	맞음	1	2	No	맞음
9	No	1	4	No	맞음	1	2	No	맞음
10	No	1	4	No	맞음	1	2	No	맞음

<그림 3> Bagging과 Modified Bagging의 비교2

<그림 3>은 <그림 2>에 있는 5개의 모델 예측값들을 가지고 Bagging Predictors와 Modified Bagging Predictors 기법을 실제로 각각 적용하여 보았을 때의 결과를 가정한 것이다. 번호 2, 즉 2번 레코드에서는 Bagging에서 사용한 5개의 모델 중 1개의 모델에서 'Yes'로 예측하였고 4개 모델에서는 'No'로 예측하여 voting을 통 해 ' No'로 예측되었으며, Modified Bagging 에서는 사용한 3개 모델 중 'Yes'로 예측한 모델은 없고 나머지 3개가 'No'로 예측하여 voting을 통해 'No'로 예측해 두 기 법이 다 실 제값과 틀림을 알 수 있다. 하지만 4번 레코드에서는 Bagging에서는 voting을 통한 값이 'No'로 예측하여 실제값과 틀리지만 Modified Bagging에서는 voting을 통한 값이 'Yes'로 예측되어 실제값과 같아서, 결과적으로 Modified Bagging이 Bagging에 비해서 한 레코드를 더 잘 예측한 것을 알 수 있다. 다음 식은 <그림 3>을 정리한 것이다.

Bagging 예측정확도 (모델 5개를 사용하여 voting) = 맞은 개수÷실제값개수 × 100 = 8÷10 × 100 = 80%

Modified Bagging 예측정확도 (상위 예측 정확도 모델 3개 사용하여 voting)

= 맞은개수÷실제값 개수 × 100 = 9÷10 × 100 = 90%

본 연구에서 제시하는 Modified Bagging Predictors란, 원 데이터 집합을 훈련용 데이터(Training data)와 시험용 데이터 (Testing data)로 나누는 것 이외에 만들어진 모델의 예측도를 알기 위하여 성능 평가용 데이터(Performance evaluating data)를 추가로 나눈 다음, 붓스트랩(Bootstrap) 방법으로 훈련용 데이터에서 랜덤하게 데이터를 복원추출하여 다수의 모델을 만들고, 만들어진 모델들을 성능 평가용 데이터에 적용하여 모델들의 예측 정확도를 측정한 후, 예측값을 평균하여 평균 이상의 예측 정확도를 가지는 모델들만을 선택해 voting하는 기법을 말한다.

4.2. 데이터 집합

본 연구에서는 SOHO 부도 예측 모델 구축을 위해서, 국내 A은행으로부터 총 입력 변수는 37개(재무변수 23개, 비재무 변수 14개), 총 레코드는 1,952개로 구성된 2001년부터 2004년까지 4년동안의 SOHO 관련 데이터를 제공받아 활용하였다. 이 중 실제 부도난 회사의 레코드는 976개이며, 부도가 나지 않은 회사의 레코드는 976개로 50:50의 같은 비율을 가지고 있다. <표 1>은 본 연구에 사용된 SOHO 데이터 집합의 변수에 대해 보여주고 있다.

<표 1> SOHO 데이터 집합의 변수

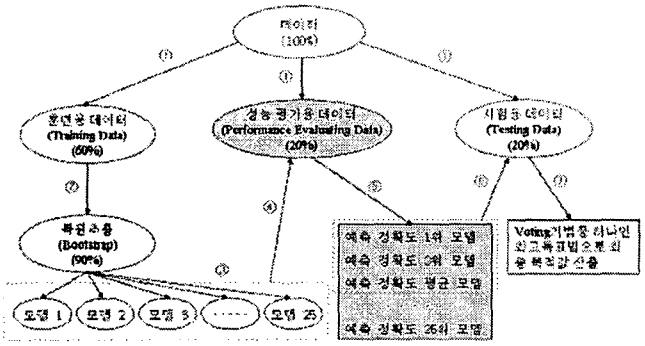
변수 (총37개)	변수 명
재무 변수 (23개)	유형자산비중, 재무안정성, 유형자산회전율, 총자산회전율, 외상구매비중, 1인당매출액, 현금매출비중, 총수신3개월평균(말잔), 총수신12개월증가율(평잔), 총여신9개월증가율(말잔), 총여신6개월평균(말잔), 총여신6개월변동계수(말잔), 카드현금서비스9개월평균, 카드신용판매6개월평균, 총수신/총여신 12개월평균(말잔), 총여신3개월평균(말잔)/총여신3개월변동계수(말잔), 총여신6개월평균(말잔)/총여신6개월변동계수(말잔), 최근신용카드개설일로부터의기간, 최근30일이내조회건수, 최근183일이내조회건수, 최근365일이내신용카드개설조회건수, 최근신용카드개설일로부터의기간/최근365일이내신용카드개설조회건수
비재무 변수 (14개)	표준산업코드, 동업계종사기간, 경영(사업)경력, 전문분야(업무), 배우자직업, 보유차종, 주택종류, 주택크기, 주택시가, 거주기간, 임차시임차형태, 입지조건_업세권여부, 종업원수, 실적정보구분

기존 기업 부도 예측에 관한 연구에서는 주로 재무변수만을 사용하였지만 중소기업 및 SOHO는 대기업과는 달리 재무 정보의 신뢰성이 약할 뿐만

아니라 재무정보가 충분치 못하기 때문에 비재무정보를 포함해서 연구를 할 필요성이 있다. 이재식·한재홍(1995)은 중소기업도산예측에 있어서 재무변수뿐만 아니라 비재무변수를 포함 활용하여 재무정보만을 사용하였을 때보다 예측력을 10% 향상 시켰다는 연구를 하였다. 비재무 변수를 살펴보면 SOHO 사업자의 특성에 맞는 더욱 개인화된 특징을 보여주고 있다.

4.3. 실험방법

본 연구에서는 두 가지의 실험을 하였다. 첫 번째 실험은 SOHO 기업 데이터 집합을 가지고 기존 기업부도 예측 연구에서 우수하다고 알려진 인공신경망을 적용한 후, CART, CART를 이용한 Bagging Predictors, CART를 이용한 Modified Bagging Predictors를 SOHO 데이터 집합에 적용하여 예측 정확도를 측정하였다. 두 번째 실험에서는 본 연구에서 새롭게 제시하는 Modified Bagging Predictors 성능에 대한 신뢰성을 검증하기 위하여 기존 SOHO 데이터 이외에 데이터 마이닝 연구에 많이 활용되는 UCI 데이터 저장소에 추출한 목적(target) 변수가 이진 변수(binary variable) 형태인 Hepatitis Domain 데이터 집합과 Credit Approval 데이터 집합, Wisconsin Breast Cancer Database 데이터 집합에 적용하여 보고, 이외에도 SPSS 학습용 데이터인 german 데이터집합, 그리고 실제 기업에서 기업의 활동을 위해 사용된 유선통신가입 유무에 관한 데이터 집합에 적용해 보았다. 이를 통해 CART, CART를 이용한 Bagging Predictors, CART를 이용한 Modified Bagging Predictors의 예측 정확도를 비교 평가하여 Modified Bagging Predictors가 전반적으로 Bagging Predictors에 비하여 성능이 좋은지, 또한 어떤 데이터 집합 유형에서 성능이 좋은지에 관한 연구를 진행하였다.



<그림 4> Modified Bagging Predictors의 방 법 및 절차

<그림 4>는 본 연구에서의 두 가지 실험에 적용한 M odified Bagging Predictors의 방법 및 절차에 대해 보다 명확하게 설명하고자 하기 위한 것이다. <그림 4>에서 진하게 색으로 표시된 부분이 기존 Bagging Predictor에는 없는 추가된 부분으로 Modified Bagging Predictors의 핵심 부분이라 말할 수 있다. Modified Bagging Predictors에서는 전체 데이터 집합을 훈련용 데이터(Training data)와 시험용 데이터(Testing data)로 나누는 기존의 방식 외에 성능 평가용 데이터(Performance evaluating data)를 시험용 데이터와 같은 비율로 추가적으로 나누었다. 성능 평가용 데이터의 용도는 훈련용 데이터에서 붓스트랩 기법으로 생성된 모델들의 예측도 순위를 평가하기 위함이다. 훈련용 데이터로 모델들의 예측도 순위 평가 시, 모델 구축에 사용된 데이터를 사용한 원 인으로 인한 왜곡 된 예측 도 순위 가 나오는 것을 미연에 방지하고자 따 로 성능 평가용 데이터를 사용하였다. 다음은 구체적인 Modified Bagging Predictors 기법의 실행 절차를 설명하 고 있다 .

(1단계) 원 데이터 집합(100%)을 훈련용 데이터(60%), 성능 평가용 데이터(20%), 시험용 데이터(20%)로 나눈다 .

(2~3단계) 훈련용 데이터를 90%의 비율로 복원추출(Bootstrap) 작업을 하여 25개의 모델을 생성한다.

(4~5단계) 생성된 모델들을 성능 평가용 데이터에 적용하여 각 모델들의 예측 정확도를 측정한다.

(6단계) 측정된 각각 모델들의 예측 값의 평균을 구한 후 평균 이상치의 예측값만 가지는 모델들을 추려내어 시험용 데이터에 적용 한다.

(7단계) 각 모델들을 시험용 데이터에 적용하여 나온 결과를 voting 기법 중 하나인 최고 득표법을 적용하여 최종 목적값 을 산출한다.

Bagging Predictors와 Modified Bagging Predictors를 위한 붓스트랩을 통한 다수 모델 생성은 Breiman (1996)에서 시도한 25회로 설정하였고 이를 통해 25개의 모델을 얻을 수 있었다. 25회 이상의 횟수를 통한 붓스트랩 작업은 Bagging의 성능 향상에 큰 영향을 끼치지 않는다는 것이 Breiman의 연구 결과로 입증되었다 [13].

5. 실험 및 결과 분석

5.1. SOHO 부도 예측에 관한 실험

첫번째 실험에서는 SOHO 데이터 집합을 통한 SOHO 부도 예측 모델 구축 및 기법간 성능 비교를 하였다. 우선 기존 기업 부도 예측에 관한 많은 연구에서 우수성이 입증된 인공신경망만의 단독 기법을 SOHO 데이터 집합에 적용하여 보았다.

다음으로는 의사결정나무의 한 기법인 CART를 적용하였다. 실험 결과, 인공신경망은 63.97%의 예측 정확도를 보였고 CART는 66.84%의 예측 정확도를 보이며 CART가 인공신경망에 비해 2.87%의 예측 정확도의 우위를 보였다. 이것은 기존 연구에서 인공신경망의 우수성에 반하는 결과로서 SOHO 데이터의 불안정성에 의한 것으로 잠정 추론된다. Bagging Predictors와 Modified Bagging Predictors의 연구는 인공신경망에 비해 성능이 좋았던 CART를 활용하였다. 각 기법의 적용 후 예측 정확도 수치와 오분류(Error rate)는 다음 표와 같다.

<표 2> SOHO 부도 예측에 관한 실험 결과

전체 데이터수	인공 신경망	CART	Bagging Predictors (총 25개 모델 사용)	Modified Bagging Predictors (총 25개 모델중 예측정확도 상위 모델 사용)
1,952개	63.97% (error rate: 36.03%)	66.84% (error rate: 33.16%)	67.1% (error rate: 32.9%)	69.71% (error rate: 30.29%)

Bagging Predictors와 Modified Bagging Predictors는 인공신경망에 비해서는 각각 3.13%, 5.74%의 예측 정확도 우위를 보였으며, CART 단독 기법에 비해서도 각각 0.26%, 2.87%의 예측 정확도 향상을 보였다. 특히 Modified Bagging Predictors가 Bagging Predictors에 비해 2.61%의 예측 정확도 향상을 보였으며, 이 결과를 분석해보면 Modified Bagging Predictors는 인공신경망, CART 뿐만 아니라 그 기법의 기초가 된 Bagging Predictors에 비해서도 성능이 좋다는 것을 알 수 있으며 따라서 SOHO의 부도 예측 모델 구축에 유용하다는 것을 알 수 있었다.

5.2. Modified Bagging Predictor의 성능 평가에 관한 실험

SOHO 부도 예측에 관한 실험을 통해 SOHO 데이터 집합에서 Modified Bagging Predictors의 유용성을 입증할 수 있었으나 이러한 결과가 SOHO 데이터 집합에만 한정된 결과인지 아니면 일반적으로 설명될 수 있는 결과인지를 입증하기 위해 데이터 마이닝 연구에서 많이 활용되는 UCI 데이터 저장소에 등록되어있는 세 개의 데이터 집합(Hepatitis Domain, Credit Approval, Wisconsin Breast Cancer Database)과 SPSS 교육용 데이터 집합(german), 또한 기존 데이터 집합의 레코드가 상대적으로 적다는 한계를 극복하기 위한 실제 기업에서 사용되었던 유선통신 가입 유무 데이터 집합을 활용하여 두 번째 실험을 하였다. 다음 <표 3>은 두 번째 실험에서 사용된 데이터 집합에 대한 요약이다. 5개의 데이터 집합은 데이터의 목적(target) 변수가 모두 이진 변수(binary variable) 형태인 데이터이다. 각 데이터 집합의 특징 및 내용은 <표 3>을 참고한다.

<표 3> Modified Bagging Predictors 성능 평가에 관한 실험에서 사용된 데이터 집합.

데이터	입력변수	목적 변수	비고
Hepatitis Domain	19개	사망/생존	환자들의 간염으로 인한 생존 여부에 관한 데이터 집합.
Credit Approval	15개	신용도 긍적적/ 신용도 부정적	Credit approval에 사용된 데이터 집합.
Wisconsin Breast Cancer Database	10개	양성/음성	Wisconsin 대학 병원에서의 환자들의 유방 암 판정에 관한 데이터 집합.
german	20개	좋음/나쁨	독일 신용 관련자료 데이터 집합.
유선통신	13개	이탈/유지	유선통신 회사 고객들의 가입 유지 여부에 관한 데이터 집합.

<표 4> Modified Bagging Predictors 성능 평가에 관한 실험에서 사용된 데이터 집합의 실험 결과

데이터	전체 데이터수	CART	Bagging Predictors (총 25개 모델 사용)	Modified Bagging Predictors (총 25개 모델중 예측정확도 상위 모델 사용)
Hepatitis Domain	155개	76.67% (error rate: 23.33%)	76.67% (error rate:23.33%)	80% (error rate: 20%)
Credit Approval	690개	80.39% (error rate: 19.61%)	83.01% (error rate: 16.99%)	83.66% (error rate: 16.34%)
Wisconsin Breast Cancer Database	699개	92.52% (error rate: 7.48%)	92.52% (error rate: 7.48%)	93.2% (error rate: 6.8%)
german	1,000개	73.68% (error rate: 26.32%)	75.26% (error rate: 24.74%)	76.84% (error rate: 23.16%)
유선통신	4,574개	83.26% (error rate: 16.74%)	83.26% (error rate: 16.74%)	83.26% (error rate: 16.74%)

<표 4>는 5개의 데이터 집합에 대한 CART, Bagging Predictors, Modified Bagging Predictors의 예측 정확도를 보여준다. <표 4>에서 볼 수 있듯이 유선통신 데이터 집합을 제외하고는 Modified Bagging Predictors가 CART 단독 적용이나 Bagging Predictors에 비하여 예측 정확도 향상의 크기는 차이가 있지만 일관적으로 성능이 향상됨을 알 수 있다. 다만 유선통신 데이터 집합에서는 해당 데이터 집합의 레코드 수가 변수 수에 비해 많고 데이터 집합 자체가 어느 정도 마사징(Massagging)되어 클린(clean)한 상태라 Bagging Predictors나 Modified Bagging Predictors의 적용의 효과가 미미하였고 따라서 성능 개선 효과가 없는 것으로 보인다.

6. 결론

대기업 및 중소기업에 한정되어진 기업 부도 예측에 관한 연구의 틀을 벗어나 국가 경제에서 비중 있는 역할을 담당하는 SOHO만의 부도 예측 모델 구축의 중요성은 상당히 크다 할 수 있다. 본 연구에서는 기존 기업 부도 예측 모델 구축 연구에서 적용이 미비하였던 Bagging Predictors 및 본 연구에서 새롭게 제시하는 Modified Bagging

Predictors를 적용하여 의미 있는 결과를 도출할 수 있었으며 향후 연구에 있어서도 많은 시사점을 준다 할 수 있겠다. 기업 부도 예측 데이터 집합의 특성이 다른 데이터 마이닝의 경우보다 데이터 수가 적고 입력 변수의 수가 많아서 생성되는 모델이 unstable한 특징을 갖는데, 이 경우 Bagging Predictors나 본 연구에서 제시한 Modified Bagging Predictors가 좋은 성능을 보임을 알 수 있었다.

추후 연구 과제로는 보다 다양한 데이터 집합에 Modified Bagging Predictors의 적용과 성능 비교가 필요하다. 또한 Modified Bagging Predictors가 좋은 성능을 제공할 수 있는 데이터 집합의 특성을 도출하는 작업이 필요하다

7. 참고문헌

[1] 김경재, 한인구, "퍼지신경망을 이용한 기업부도예측",한국지능정보시스템학회논문지,제7권 제1호, 2001.

[2] 신경식, "다수의 인공신경망 모형을 통합한 기업부도 예측모형에 관한 연구",경영논총, 제18권 제1호, 2000.

[3] 신현정, "앙상블 학습알고리즘의 일반화 성능 비교:OLA, Bagging, Boosting", 정보과학회논문지, 제97호, 2000.

[4] 이건창, 김명종, 김혁, "기업도산예측을 위한 귀납적 학습지원 인공신경망 접근방법:MDA, 귀납적 학습방법, 인공신경망 모형과의 성과비교", 경영학연구, Vol. 23, No. 2, 1994, pp. 109-144.

[5] 이군희, "모형의 평가와 앙상블을 이용한 데이터마이닝에 관한 연구", 서강경영논총, 제9권, 1998.

[6] 이영섭, 오현정, 김미경, "데이터 마이닝에서 배깅, 부스팅, SVM 분류 알고리즘 비교 분석", 응용통계연구, 제18권 2호, 2005.

[7] 이재식, 한재홍, "인공신경망을 이용한 중소기업도산예측에 있어서의 비재무정보의 유용성 검증", 한국전문가시스템학회지 창간호, 1995.

[8] 허준, 정규상, 허수희, 최희경, "Clementain 7 매뉴얼', 2003.

[9] 홍승현, 신경식, "유전자 알고리즘을 활용한 인공신경망 모형 최적입력변수의 선정: 부도예측 모형을 중심으로", 한국지능정보시스템학회논문지, 제9권 제1호, 2003.

[10] Altman, E., "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy", The Journal of Finance, Vol.23, No4, 1968, pp 589-609.

[11] Altman, E., "Corporate Financial Distress - A Complete Guide to Predicting, Avoiding and Dealing with Bankryptcy", John Wiley & Sons, New York, 1983.

[12] Breiman L., Friedman J.H., Olshen R.A. and Stone C.J., "Classification and Regression Trees (CART)", Chapman &Hall/CRC, 1984.

[13] Breiman L., "Bagging Perdictors", Machine Learning, Vol. 24, No. 2, 1996, pp. 123-140.

[14] Breiman L., "Using Iterated Bagging to Debias Regressions", Machine Learning, Vol. 45, 2001, pp. 261-277.

[15] Gentry, J. A., Newbold, P. and Whitford, D. T., "Classifying Bankrupt Firms With Funds Flow Components", Journal of Accounting Research, Spring, 1985, pp. 146-160.

[16] Gombola M. J. and Ketz, J.E., "Financial Ratio Patterns in Retail and Manufacturing Organizations", Financial Management, Summer, 1983, pp. 45-56.

[17] Hyunjoong Kim, Dongjun Chung, "Improving Bagging Predictors", Korea Statistical Society, Proceedings of the autumn Conference, 2005, pp. 141-146.

[18] Jo, H, I. Han and H. Lee, "Bankruptcy prediction using vase-based reasoning, neural networks, and discriminate analsis", Expert Systems with Applications, Vol. 13, No.2, 1997, pp. 97-108.

[19] Kelvin T. Leung, D. Stott Parker, "Empirical Comparisons of Various Voting Methods in Bagging", 2003.

[20] Odom, M. D. and R. Sharda, "A neural network model for bankruptcy prediction", In Proceedings of the IEEE International Conference on Neural Networks, San Diego, CA, 1990, pp 163-168.

[21] Tam, K. Y. and M. Y. Kiand, "Managerial applications of neural networks: The case of bank failure predictions", Management Science, Vol. 38, 1992, pp. 926-947.

[22] Wilson, R. L. and R. Sharda, "Bankruptcy prediction using neural networks', Decision Support Systems, Col. 11, 1994, pp. 545-557.

[23] _, www.ics.uci.edu/~mlearn/databases, 2006.