한국지능정보시스템학회논문지 제13권 제2호 2007년 6월(pp.15~26)

Modified Bagging Predictors를 이용한 SOHO 부도 예측*

김승혁
에스큐테크놀리지
(ksh@sqtech.net)

김종우
한양대학교 경영대학 경영학부
(kjw@hanyang.ac.kr)

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 기업부도 예측 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 대출 데이터 집합이 사용되었다. 먼저, 기업부도 예측 모델에서 우수하다고 연구되어진 인공신경망과 의사결정나무 추론 기법을 적용하여 보았지만 만족할 만한 성과를 이끌어내지 못하여, 기존 기업부도 예측 모델 연구에서 적용이 미비하였던 Bagging Predictors와 이를 개선한 Modified Bagging Predictors를 제시하고 이를 적용하여 보았다. 연구결과, SOHO 부도 예측에 있어서 본 연구에서 제시한 Modified Bagging Predictors가 인공신경망과 Bagging Predictors 등의 기존 기법에 비해서 성과가 향상됨을 알 수 있었다.

논문접수일 : 2006년 06월 게재확정일 : 2007년 05월 교신저자 : 김종우

1. 서론

과거 외환 위기를 전후로 기업의 규모와 상관없이 많은 기업이 부도의 위기를 겪게 되었고, 이에 따라 많은 국내 금융기관들이 대출금을 회수하지 못하는 불상사가 발생하여 국내 경제가 급속히 와해되고 경제난이 가속화되기에 이르러, 국가 경제가 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 있다. 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 기업부도 예측 모델이 기존의 통계분석에 비해서 우수함이 많은 연구들을 통해서 입증이 되어 기업부도 예측 모델에 많이 활용되고 있다(Breiman, 1984; Breiman, 2001; Bauer and Kohavi, 1999; Gentry et al., 1983; Wilson and Sharda 1994). 하지만 기존의 국내외 기업부도 예측 모델에 관한 연구는 재무재표 데이터를 중심으

* 이 논문은 한양대학교 일반연구비 지원으로 연구되었음(HY-2006-G).

로 하는 대기업에 대한 연구에 치중되어 중소기업이나 소상공인에 대한 연구가 상대적으로 미비한 편이다. 물론 중소기업의 재무재표 데이터 및 비재무에 관련된 데이터가 신뢰성면에서 대기업에 비해 떨어지는 특징이 있기는 하지만, 중소기업과 소상공인이 국가경제에서 차지하는 비중이 결코 작지 않음을 생각하면 아쉬운 부분으로 남고 있다. 특히 SOHO만의 기업 부도 예측 모델에 대한 연구는 전무하다시피한 작금의 상황에서 이에 대한 연구는 필요하다. 실제로 SOHO에 포함되는 개인사업체의 수가 전체 사업체의 85.6%(약 275만, 2005년 기준(통계청))를 차지하는 현실에도 불구하고, 금융기관들의 대출 심사 시 대기업 및 중소기업과는 달리 SOHO에 대한 대출심사는 아직은 체계화 되지 못한 체 부도확률, 등급, 종합점수, 개요정보점수, 실적정보점수, 신용정보점수 등의 단편적인 요소를 사용하고 있는 것이 작금의 상황이고 이에 따라 잘못된 대출로 인한 금융기관의 부실화를 초래할 위험성이 크다.

본 연구에서는 SOHO 기업만의 부도 예측 모델을 구축하여 향후 기업부도 연구에 유용한 결과를 도출하고자 하였다. 연구에 쓰인 방법으로는 기업부도 예측 연구에서 우수하다고 입증된 인공신경망을 비롯하여 의사결정나무 추론(Decision tree induction)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. 본 논문의 구성은 다음과 같다. 제 2장에서는 관련연구에 대하여 살펴보고 제 3장에서는 Bagging Predictors와 본 논문에서 제시하는 Modified Bagging Predictors에 대하여 소개한다. 제 4장에서는 실험을 통한 분석을 통해 본 연구의 효용성을 입증하였으며, 마지막으로 제 5장에서는 연구에 대한 결과 요약 및 향후 연구 방향에 대하여 제시하였다.

2. 관련 연구

2.1 기업부도 예측 모형에 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. 통계기법을 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-최근접 이웃방법(k-nearest neighbor), 귀납적 추론(ID3)과 인공신경망과의 성과 비교를 통해 인공신경망의 우수함을 연구하였다. 이건창, 김명종, 김혁(1994)은 MDA(Multiple Discriminant Analysis), 귀납적 학습방법, 인공신경망과의 성과를 비교하였으며, 이재식과 한재홍(1995)은 인공신경망을 이용하여 중소기업의 도산 예측에 있어서, 재무정보를 보완할 수 있는 비재무정보의 유용성을 검증하였고, 신경식(2000)은 입력 변수군을 달리하는 다수의 인

공신경망 모델을 구축하고 통합하여 예측력 향상을 이끌어내었으며, 김진백과 이준섭(2000)은 인공신경망과 사례기반추론(Case-Based Reasoning, CBR) 기법을 사용하여 모델을 개발하고 현금흐름 지표가 기존에 주로 사용된 일반재무비율 변수에 근거한 부실 예측 모델에 추가적인 역할을 할 수 있는지를 평가하였고, 김경재과 한인구(2001)는 기존 신경망에 퍼지집합의 개념을 적용하여 신경망 학습에 사용될 자료를 퍼지화하고 이를 신경망에 학습시켰으며, 홍승현과 신경식(2003)은 유전자 알고리즘을 이용하여 입력 변수군을 도출하여 인공신경망을 적용한 연구를 하였다. 최근 연구에서는 모형의 예측력을 향상시키기 위하여 통합 인공신경망 또는 인공지능기법의 복합방법론을 제시하는 연구들이 주로 발표되고 있는 것이 특징이다.

2.2 SOHO

SOHO란, Small Office Home Office의 약어로서, 공간적으로는 집 또는 소규모 사무실 또는 소규모 점포형 독립사업을 영위하는 독립사업자를 말한다. 직업적으로는 개인의 아이디어를 활용하여 수익을 창출하는 프리랜서를 의미한다. 그러나 우리나라에서의 SOHO란 프리랜서 뿐만 아니라 소규모 인터넷 사업자 및 자영업 등을 포함한 복합적 의미로 사용되고 있으며, 표준화된 개념은 정의 되어 있지 않은 상태이다. SOHO와 관련된 개념으로 소기업과 소상공인이 있다. 법적으로는 소기업은 종업원수 10인 미만(제조, 건설, 운송, 광업은 50인 미만)의 기업을 의미하며(중소기업 기본법 시행령 제 8조), 소상공인은 종업원수 5인 미만(제조, 건설, 운송, 광업은 10인 미만)의 사업자를 의미한다(소기업 및 소상공인 지원을 위한 특별조치법 시행령 제 2조).

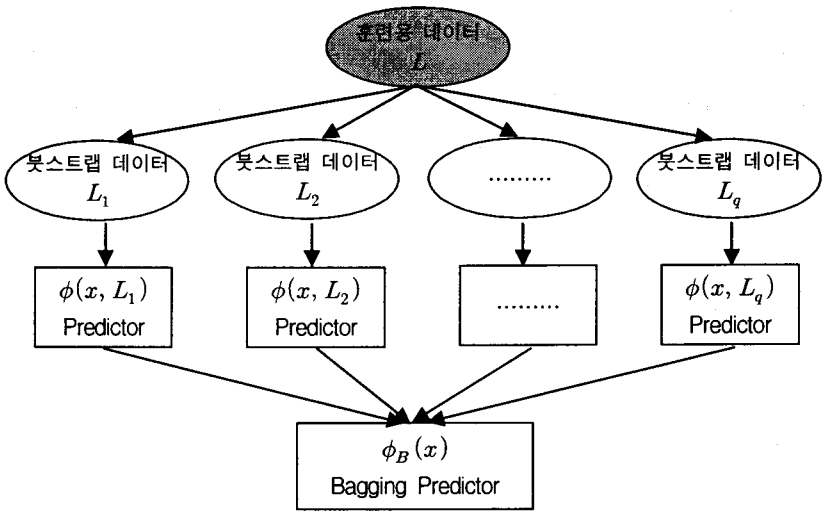
최근 시중은행은 정책당국의 가계대출 억제조치 등으로 새로운 대출대상처의 발굴 필요성이 대두됨에 따라서 SOHO를 새로운 고객층으로 인식하고 대출마케팅을 강화하고 있다. 그러나 현재 시중은행의 SOHO 대출 기준 및 운영방법은 은행별로 차이가 나는 등 정확한 개념은 정립되어 있지 않은 상황이다. 은행에 따라서는 법인을 제외한 중소 개인사업자만을 SOHO 대출 대상자로 하는 은행이 있는가 하면, 자산규모 10억원 이하의 소기업, 소상공인을 대상으로 하는 은행도 존재한다. 본 연구에서는 국내 A은행의 SOHO 대출 데이터를 사용하였는데, 이 은행의 경우, SOHO의 기준을 개인사업자와 총자산 5억원 이하인 소규모 중소기업을 포함한다.

3. Modified Bagging Predictors

3.1 Bagging Predictors

Bagging이란 여러 개의 predictor를 만들어 이것을 이용하여 통합 predictor를 얻어내는 방법을 말한다(허준 외, 2005). 다시 말하면, 지도학습에서 주어진 훈련용 데이터를 복원추출(bootstraping : resampling randomly with replacement)하여 여러 개의 훈련용 데이터 집합으로 만들고, 각각의 데이터 집합에 대하여 predictor를 생성하여 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 predictor에



[그림1] Bagging Predictors

x 입력변수 벡터에 대한 예측값을 의미한다. $\{L_k\}$는 L과 같은 분포에서 뽑힌 q개의 독립적인 관측치로 구성된 훈련용 데이터 집합의 순열(실제로는, L에서의 bootstrap sample을 사용)이라고 하자.

Bagging의 목적은 $\{L_k|, k=1,\cdots,q\}$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. Bagging은 L에서의 작은 변화가 φ에서 큰 변화를 가져올 때, 즉 모델이 불안정적인 경우 효과적이며, 그렇지 않은 경우는 효과가 미미하다(Breiman, 1996). 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. [그림 1]은 Bagging을 도식화한 그림이다.

3.2 Modified Bagging Predictors의 개념 및 절차

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(Preformance evaluating data)를 나눈 다음, 붓스트랩(Bootstrap) 방법으로 훈련용 데이터에서 랜덤하게 데이터를 복원추출하여 다수의 모델을 만들고, 만들어진 모델들을 성능 평가용 데이터에 적용하여 모델들의 예측도 정확도를 측정한 후, 예측 값을 평균하여 평균 이상의 예측 정확도를 가지는 모델들만을 선택해 voting하는 기법을 말한다.

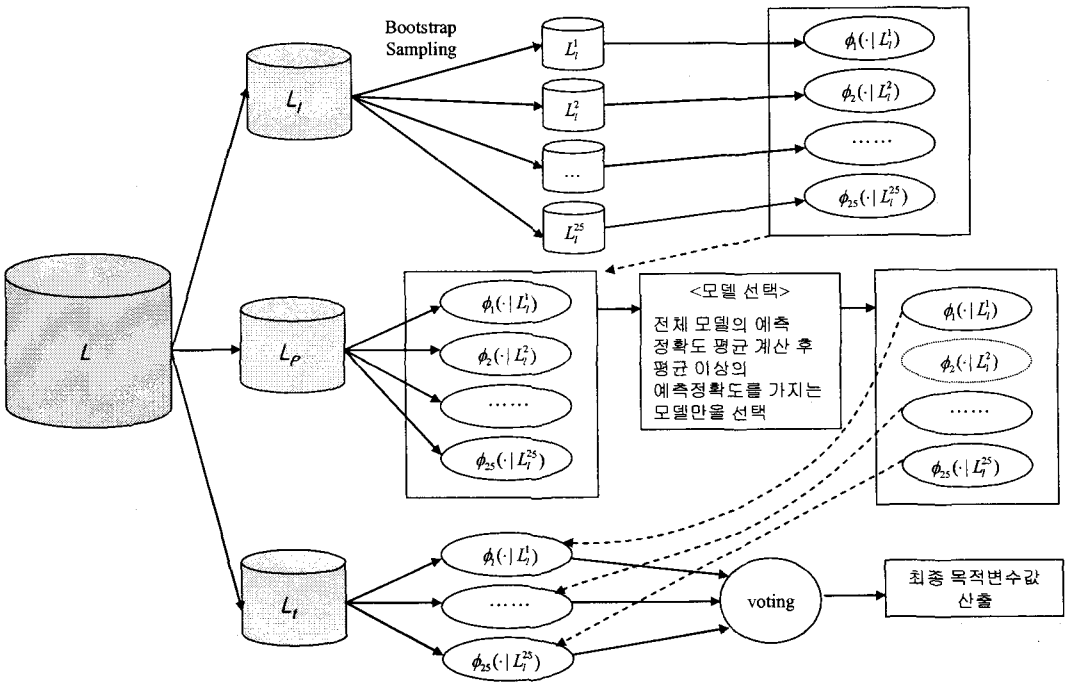
[그림 2]는 본 연구에서 제시하는 Modified Bagging Predictors의 적용 절차를 보여준다. Modified Bagging Predictors에서는 전체 데이터 집합을 훈련용 데이터(Training data) L_l과 시험용 데이터(Testing data) L_t로 나누는 기존의 방식 외에 성

능 평가용 데이터(Performance evaluating data) L_p를 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 다시 사용함으로써 왜곡된 예측도 순위가 나오는 것을 미연에 방지하고자 따로 성능 평가용 데이터를 사용하였다.

다음은 목표변수가 이진변수인 경우의 구체적인 Modified Bagging Predictors 기법의 실행 절차를 설명하고 있다.

1단계 : 전체 데이터 집합을 $L=\{(y_n, x_n), n=1, 2, \cdots, N\}$이라 정의했을 때, y_n은 목적변수이고, x_n은 설명 변수 벡터를 의미하고, N은 데이터의 레코드수를 의미한다. 그리고 훈련용 데이터 집합을 $L_l=\{(y_l, x_l), l=1, 2, \cdots, M\}$, 훈련용 데이터 집합을 통해 만들어진 모델을 평가하기 위한 성능 평가용 데이터 집합을 $L_p=\{(y_p, x_p), p=1, 2, \cdots, P\}$라고 정의하고, 시험용 데이터 집합을 $L_t=\{(y_t, x_t), t=1, 2, \cdots, T\}$라고 정의한다(단 $N=M+P+T$이다). 본 연구에서의 데이터 집합의 분리는, 원 데이터 집합을 훈련용 데이터는 60%, 성능 평가용 데이터는 20%, 시험용 데이터는 20%로 나누었다.

2단계 : 본 단계에서는 훈련용 데이터 집합을 사용하여 복원 추출(Bootstrap) 작업 과정을 통해 다수의 데이터 부분 집합을 생성시키고 이를 활용하여 다수의 모델을 생성한다. 훈련용 데이터 집합 L_l에서



[그림 2] Modified Bagging Predictors의 방법 및 절차

생성된 bootstrap sample의 집합을 $L_l^B = \{L_l^1, L_l^2 \cdots, L_l^q\}$이라고 정의했을 때, 이를 통해 $\Phi = \{\phi_1(\cdot|L_l^1), \cdots, \phi_q(\cdot|L_l^q)\}$를 생성한다(본 연구에서 $q=25$로 설정하였다). 즉, 본 연구에서는 훈련용 데이터를 90% 비율로 복원추출(Bootstrap) 작업을 하여 25개의 모델을 생성하였다.

3단계 : 생성된 모델들을 성능 평가용 데이터에 적용하여 각 모델들의 예측 정확도를 측정한다. 즉, 성능 평가용 데이터 집합 L_p에, 앞서 만들어진 $\phi_1(\cdot|L_l^1), \cdots, \phi(\cdot|L_l^q)$를 적용시킨 후 각 모델의 예측 정확도 $\{V_i|, i=1, \cdots, q\}$를 구한다. 다음 식 (1)은 모델 $\phi_i (i=1, \cdots, 25)$의 정확도 V_i를 구하는 식이다.

$$Vi = Accuracy\{\phi_i(\cdot|L_l^1)|L_p\} = \frac{|\{(y_p, x_p)|\phi_i(x_p|L_l^1) = y_p, (y_p, x_p) \in L_p\}|}{P} \quad (1)$$

4단계 : 측정된 각각 모델들의 예측 정확도의 평균을 구한 후 평균 이상의 예측값만 가지는 모델들을 추려내어 시험용 데이터에 적용한다. 즉, 식 $\nabla = (\sum_{i=1}^{q} V_i/q)$를 이용하여 평균 예측 정확도를 구한 후 $V_i > \nabla$인 ϕ_i만으로 이루어진 Φ의 부분집합 Φ'를 생성하였다. 즉,

$$\Phi' = \{\phi_i | V_i > \nabla, i = 1, \cdots q, \phi_i \in \Phi\} \quad (2)$$

5단계 : Φ'에 포함된 모델들을 활용하여 최종 예측자를 생성하였다. 즉 최종 예측자 ϕ''는 다음과 같이 정의된다. 즉, $(y_i, x_i) \in L_t$에 대하여, 목적변수값 j로 예측한 Φ'에 속한 Predictor의 개수를 $N_j = |\{\phi_k | \phi_k(x_i) = j, j = 0, 1, \phi_k \in \Phi'\}|$이라 정의하면,

$$\phi''(x_i) = j, \text{ 단 } N_j = \underset{i \in \{0,1\}}{MAX} N_i \quad (3)$$

6단계 : 성능 평가를 위하여 5단계에서 정의된 ϕ''를 L_t에 적용하여 예측 정확도를 계산한다. 즉 Modified Bagging Predictors의 최종 정확도는 다음 식 (4)와 같이 계산된다.

$$V' = \frac{|\{(y_t, x_t)|\phi''(x_t) = y_t, (y_t, x_t) \in L_t\}|}{T} \quad (4)$$

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

4. 실험

4.1 SOHO 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 4년 동안의 SOHO 관련 데이터를 제공받아 활용하였다. 이 중 실제 부도난 회사의 레코드는 976개이며, 부

도가 나지 않은 회사의 레코드는 976개로 50 : 50의 같은 비율을 가지고 있다. <표 1>은 본 연구에 사용된 SOHO 데이터 집합의 변수에 대해 보여주고 있다. 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. 이재식과 한재홍(1995)은 중소기업도산예측에 있어서 재무변수뿐만 아니라 비재무변수를 포함 활용하여 재무정보만을 사용하였을 때보다 예측력을 10% 향상 시켰다는 연구 결과를 제시하였다. 비재무 변수를 살펴보면 배우자 직업, 보유차종, 주택종류 등 SOHO 사업자의 특성에 맞는 개인적인 변수들이 포함되어 있다.

<표 1> SOHO 데이터 집합의 변수[1)]

변수 (총 37개)	변수명
재무 변수 (23개)	유형자산비중, 재무안정성, 유형자산회전율, 총자산회전율, 외상구매비중, 1인당 매출액, 현금매출비중 등
비재무 변수 (14개)	표준산업코드, 동업계종사기간, 경영(사업)경력, 배우자직업, 보유차종, 주택종류 등

4.2 실험

실험에서는 기법간 성능 비교를 하였다. SOHO 데이터 집합에 대하여 인공신경망, CART, CART를 이용한 Bagging Predictors, CART를 이용한 Modified Bagging Predictors를 적용하였을 때의 예측 정확도와 오분류율(error rate)를 정리한 것이 <표 2>와 <표 3>이다. <표 2>는 각 실험 횟수별 예측 정확도에 대한 것이고, <표 3>와 [그림 3]은 <표 2>의 결과를 평균한 최종 예측 정확도에 대한 것이다.

실험을 위해서는 SPSS의 데이터마이닝 도구인 클레멘타인을 사용하였다. 인공신경망과 CART에 대해서 모두 클레멘타인이 제공하는 디폴트 옵션을 사용하였다. 이 경우 CART에서는 불순도 기준으로 Gini 척도를 사용하였고, 정지 기준으로는 퍼센트 기준을 사용하여, 부모마디 최소 레코드 수(%)를 2로, 자식마디 최소 레코드 수(%)를 1로 하였다. 실험 결과 <표 3>에서 볼 수 있듯이, 인공신경망은 63.71%의 평균 예측 정확도를 보였고 CART는 68.08%의 평균 예측 정확도를 보이며 CART가 인공신경망에 비해 4.37%의 예측 정확도의 우위를 보였다. 이것은 기존 연구에서 인공신경망의 우수성에 반하는 결과로서 SOHO 데이터의 불안정성에 의한 것으로 잠정 추론된다. Bagging Predictors의 예측 정확도는 69.15%, Modified Bagging Predictors의 예측 정확도는 70.55%로서 인공신경망에 비해서 각각 5.44%, 6.84%의 예측 정확도 우위를 보였으며, CART 단독 기법에 비해서도 각각 1.07%, 2.47%의 예측 정확도 향상을 보였다. 또한 본 연구에서 제시한 Modified Bagging Predictors는 Bagging Predictors에 비해 1.4%의 예측 정확도 향상을 보였다. 결과를 분석해보면 Modified Bagging Predictors는 인공신경망, CART 뿐만 아니라 그 기법의 근간이 된 Bagging Predictors에 비해서도 성능이 좋다는 것을 알 수 있다.

결과에 대한 통계적 검증을 위하여, <표 2>에서의 각 회에 따른 4개 기법의 예측 정확도를 이용하여 대응 이집단 비율 검정(paired two sample test for proportions)을 수행하였다. 즉, 실험 횟수가 10회로 작으므로 분포 등과 무관하게 검정할

1) 해당 기업의 정보 보호를 위해 상세한 변수 리스트는 제시하지 않음.

<표 2> SOHO 부도 예측 정확도(세부)

실험횟수	인공신경망	CART	Bagging	Modified Bagging	Bagging -CART	Modified Bagging-Bagging
1회	63.97%	66.84%	67.10%	69.71%	0.26%	2.61%
2회	63.41%	64.41%	67.67%	68.42%	3.26%	0.75%
3회	63.42%	68.42%	70.53%	72.11%	2.11%	1.58%
4회	63.85%	66.15%	66.67%	66.67%	0.52%	0.00%
5회	63.97%	68.41%	68.67%	71.28%	0.26%	2.61%
6회	63.73%	68.77%	72.54%	71.54%	3.77%	-1.00%
7회	66.50%	71.03%	70.78%	73.05%	-0.25%	2.27%
8회	63.76%	68.52%	67.99%	70.90%	-0.53%	2.91%
9회	61,76%	68.48%	69.51%	70.80%	1.03%	1.29%
10회	62.75%	69.75%	70.00%	71.00%	0.25%	1.00%

<표 3> SOHO 부도 예측 정확도(평균)

(10회 반복)

전체 데이터수	인공신경망	CART	Bagging	Modified Bagging	Bagging -CART	Modified Bagging-Bagging
1,952개	63.71% (error rate : 36.29%)	68.08% (error rate : 31.92%)	69.15% (error rate : 30.85%)	70.55% (error rate : 29.45%)	1.07%	1.4%

<표 4> SOHO 부도 예측에 관한 실험 결과에 대한 기법 간 정확도에 대한 대응 이집단 비율 분석결과

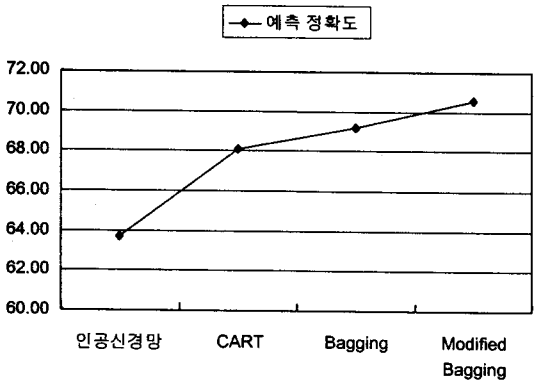
(유의수준)

	CART	Bagging	Modified Bagging
인공신경망	0.005**	0.005**	0.005**
CART	-	0.041*	0.005**
Bagging	-	-	0.018*

** 유의수준 1% 이내, * 유의수준 5% 이내.

수 있도록, 2개의 대응 집단 비율 검정방법은 비모수적인 방법인 Wilcoxon의 부호순위 검정을 이용하였다. 부호순위 검정 결과는 <표 4>와 같다. 검정결과가 유의수준 0.01(1%) 또는 유의수준 0.05(5%)에서 유의 한 것을 알 수 있다. 이는 전체적으로 Modifed Bagging Predictors가 다른 3개의 기법에 비해 효과적임을 나타내고 있다.

<표 2>를 보면 실험 횟수 별 각 기법들간의 예측 정확도 결과를 볼 수 있다. 자세히 살펴보면, 1회, 5회, 7회, 8회처럼 CART 단독 기법에 비해서 Bagging Predictors가 미미한 성능 향상을 보이거나 성능 감소가 있을 때 Modified Bagging Predictors의 성능 향상이 다른 때보다 더 크다는 것을 알 수 있다. 또 6회처럼 Bagging Predictors가 CART에 비해 크게 성능 향상이 있을 때는 오히려 Modified Bagging Predictors가 Bagging Predictors에 비해 성능이 감소되는 것을 알 수 있다. 이를 통해 Bagging Predictors의 성능 향상이 큰 경우에서는 Modified Bagging Predictors의 효과가 미미하거나 감소될 수 있음을 지적할 수 있다.



[그림 3] SOHO 부도 예측에 관한 실험 결과

실험 결과를 정리하면, SOHO 데이터가 가지는 데이터 수의 한계, 재무 정보의 신뢰성 한계, 다수의 변수 수 등으로 인해서 안정적인 의사결정나무가 생성되지 못할 것으로 생각되어 Bagging Predictors 기법을 적용하여 실험에서와 같이 성능향상을 꾀할 수 있었다. 또한 Bagging Predictors 기법의 적용 과정에서, 생성되는 모델들의 예측 정확도가 차이가 큰 것을 발견하고, 이에 착안하여 성능이 좋은 모델들만을 선별하여 최종 예측에 사용하는 Modified Bagging Predictors를 설계하여 Bagging Predictors보다 나은 성능을 확인할 수 있었다.

5. 결론

본 연구에서는 기존 기업 부도 예측 모델 구축 연구에서 적용이 미비하였던 Bagging Predictors를 개선한 Modified Bagging Predictors를 활용하여, 대기업 및 중소기업에 한정되어진 기업 부도 예측에 관한 연구의 틀을 벗어나 국가 경제에서 비중 있는 역할을 담당하는 SOHO를 위한 부도 예측에 관한 연구를 수행하였다. 연구 결과 기존 대기업 위주의 기업 부도 예측 모델 구축 시 인공신경망이 효과적이었던데 비하여, SOHO 부도 예측 모델 구축에서는 CART나 Bagging Predictors에 비해서 좋은 성능을 나타내지 못하였다. 이것은 SOHO 데이터가 입력변수의 수가 많고, 사례의 수가 부족한 특성에 기인한 것으로 보인다. 이러한 SOHO 데이터의 특성으로 생성되는 분류 모델의 안정성이 떨어지는 경우 Bagging Predictors가 유용하며, 본 연구에서 제시한 Modified Bagging Predictors가 기존의 Bagging Predictors에 비해서 예측 정확도가 높음을 확인할 수 있었다.

본 연구의 한계점은 다음과 같다. 국내 특정 A 은행에서 얻은 SOHO 대출 데이터 집합을 통해서 연구되어진 본 연구는 그 기업 수나 데이터의 정확도 면에서 한계를 가지고 있다. 따라서 보다 정확하고 많은 SOHO 기업으로 구성된 데이터집합을 사용한 연구가 필요하다고 생각된다. 또한 기업 부도 예측 모형에 사용될 수 있는 보다 다양한 통계분석 기법과 인공지능 기법과의 추가적인 성능 비교가 필요하다.

참고문헌

[1] 김경재, 한인구, "퍼지신경망을 이용한 기업 부도예측", *한국지능정보시스템학회논문지*, 7권 1호(2001), 135~147.

[2] 김영태, 이현철, "기업도산 예측과 재무비율 정보의 유용성에 관한 실증연구", *한남대학교 산업경영연구*, (2001), 45~56.

[3] 박성현, 조신섭, 김성수, *한글 SPSS Ver. SPSS 10K*, ㈜데이터솔루션, 2003.

[4] 신경식, "다수의 인공신경망 모형을 통합한

기업부도 예측모형에 관한 연구", *경영논총*, 18권 1호(2000), 57~69.

[5] 신현정, "앙상블 학습알고리즘의 일반화 성능 비교 : OLA, Bagging, Boosting", *정보과학회논문지*, 97호(2000).

[6] 이건창, 김명종, 김혁, "기업도산예측을 위한 귀납적 학습지원 인공신경망 접근방법: MDA, 귀납적 학습방법, 인공신경망 모형과의 성과비교", *경영학연구*, 23권 2호(1994), 109~144.

[7] 이군희, "모형의 평가와 앙상블을 이용한 데이터마이닝에 관한 연구", *서강경영논총*, 9권(1998).

[8] 이영섭, 오현정, 김미경, "데이터마이닝에서 배깅, 부스팅, SVM 분류 알고리즘 비교 분석", *응용통계연구*, 18권 2호(2005), 343~354.

[9] 이재식, 한재홍, "인공신경망을 이용한 중소기업도산예측에 있어서의 비재무정보의 유용성 검증", *한국전문가시스템학회지*, 1권 1호(1995), 123~134.

[10] 정수연, "Logit Model을 이용한 기업부도예측 결정요인에 관한 연구", *한국감정평가원*, (2003), 163~173.

[11] 조영임, *인공지능시스템*, 홍릉과학출판사, 2003.

[12] 허준, 김종우, "오차 패턴 모델링을 이용한 Hybrid 데이터마이닝 기법", *한국경영과학회지*, 30권 4호(2005), 27~43.

[13] 허준, 정규상, 허수희, 최희경, *Clementain 7 매뉴얼*, ㈜데이터솔루션, 2003.

[14] 홍승현, 신경식, "유전자 알고리즘을 활용한 인공신경망 모형 최적입력변수의 선정 : 부도예측 모형을 중심으로", *한국지능정보시스템학회논문지*, 9권 1호(2003), 227~247.

[15] Altman, E., "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy", *The Journal of Finance*, Vol.23, No.4(1968), 589~609.

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

[17] Bauer, Eric, Ron Kohavi, "An Empirical Comparisons of Voting Classification Algorithms : Bagging, Boosting, and Variants", *Machine Learning*, Vol.36(1999), 105~139.

[18] Beaver, W. H, "Financial Ratios as Predictors of Failure", *Journal of Accounting Research*, Vol.4(1966), 71~111.

[19] Berry, Michael J. A., Gordon S. Linoff, *Data Mining Techniques*, Wiley, 2004.

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

[21] Breiman, L., "Bagging Perdictors", *Machine Learning*, Vol.24, No.2(1996), 123~140.

[22] Breiman, L., "Using Iterated Bagging to Debias Regressions", *Machine Learning*, Vol.45(2001), 261~277.

[23] Dettling, Marcel, "BagBoosting for Tumor Classification with Gene Expression Data", *Bioinformatics*, Vol.20(2004), 3583~3593.

[24] Dudoit, Sandrine, Jane Fridlyand, "Bagging to Improve the Accuracy of a Clustering Procedure", Bioinformatics, Vol.19(2003), 1090~1099.

[25] Gentry, J. A., Newbold, P. and Whitford, D. T., "Classifying Bankrupt Firms with Funds Flow Components", *Journal of Accounting Research*, Spring(1985), 146~160.

[26] Gombola, M. J. and Ketz, J. E., "Financial Ratio Patterns in Retail and Manufacturing Organizations", *Financial Management*, Summer(1983), 45~56.

[27] Ha, Kyoungnam, Sungzoon Cho, Douglas Maclachlan, "Response Models based on

Bagging Neural Networks", *Journal of Interactive Marketing*, Vol.19(2005), 17~30.

[28] Han, Jiawei, Micheline Kamber, *Data Mining : Concepts and Techniques*, Morgan Kaufmann, 2001.

[29] Hebb, D. O., *The Organization of Behavior : A Neuropsychological Theory*, New York Wiley, 1949.

[30] Hecht-Nielsen, R, *Neurocomputing*, Addison-Wesley, 1990.

[31] Jo, H, I. Han, H. Lee, "Bankruptcy Prediction Using Case-based Reasoning, Neural Networks, and Discriminate Analysis", *Expert Systems with Applications*, Vol.13, No.2(1997), 97~108.

[32] Kim, Hyunjoong, Dongjun Chung, "Improving Bagging Predictors", *Proceedings of the Autumn Conference of Korea Statistical Society*, (2005), 141~146.

[33] Leung, Kelvin T., D. Stott Parker, "Empirical Comparisons of Various Voting Methods in Bagging", *SIGKDD 2003*, (2003), 595~600.

[34] McCulloch, W. S. and Pitts, W., "A Logical Calculus of the Ideas Immanent in Nervous activity", *Bulletin of Mathematical Biophysics*, Vol.5(1943), 115~133.

[35] Michie, D., D. J. Spiegelhalter, and C. Taylor, Machine Learning, Neural and Statistical Classification, Ellis Horwood, 1994.

[36] Odom, M. D. and R. Sharda, "A Neural Network Model for Bankruptcy Prediction", *Proceedings of the IEEE International Conference on Neural Networks*, San Diego, CA, (1990), 163~168.

[37] Rosenblatt, F., "The Perceptron : A Probabilistic Model for Information Storage and Organization in the Brain", *Psychological Review*, Vol.65(1958), 386~408.

[38] Skurichina, Marina, Robert P. W. Duin, "The Role of Combining Rules in Bagging and Boosting", *SSPR&SPR 2000, LNCS 1876*, (2000), 631~640.

[39] Skurichina, Marina, Robert P. W. Duin, "Bagging, Boosting and the Random Subspace Method for Linear Classifiers", *Pattern Analysis & Applications*, Vol.5, No.2(2002), 121~135.

[40] Tsymbal, Alexey, Seppo Puuronen, "Bagging and Boosting with Dynamic Integration of Classifier", *PKDD2000, LNAI1910*, (2000), 116~125.

[41] Tam, K. Y. and M. Y. Kiand, "Managerial Applications of Neural Networks : The Case of Bank Failure Predictions", *Management Science*, Vol.38, No.7(1992), 926~947.

[42] Wilson, R. L. and R. Sharda, "Bankruptcy Prediction Using Neural Networks", *Decision Support Systems*, Vol.11, No.5(1994), 545~557.

Abstract

SOHO Bankruptcy Prediction Using Modified Bagging Predictors

Seung Hyuk Kim* · Jong Woo Kim**

In this study, a SOHO (Small Office Home Office) bankruptcy prediction model is proposed using Modified Bagging Predictors which is modification of traditional Bagging Predictors. There have been several studies on bankruptcy prediction for large and middle size companies. However, little studies have been done for SOHOs. In commercial banks, loan approval processes for SOHOs are usually less structured than those for large and middle size companies, and largely depend on partial information such as credit scores. In this study, we use a real SOHO loan approval data set of a Korean bank. First, decision tree induction techniques and artificial neural networks are applied to the data set, and the results are not satisfactory. Bagging Predictors which has been not previously applied for bankruptcy prediction and Modified Bagging Predictors which is proposed in this paper are applied to the data set. The experimental results show that Modified Bagging Predictors provides better performance than decision tree inductions techniques, artificial neural networks, and Bagging Predictors.

Key words : Bankruptcy Prediction, Data Mining, Bagging Predictors, Artificial Neural Networks, Decision Tree Induction

* SQ Technology
** School of Business, Hanyang University