데이터 마이닝 기법의 성과측정시 표본추출 및 표본구성비의 영향에 관한 실증적 연구

김 광용
숭실대학교 경영학부 교수, gygim@saint.soongsil.ac.kr ☎ 820-0597

본 연구의 목적은 이원화된 위험을 분류하는데 사용된 여러 가지 데이터마이닝(data mining) 기법들의 성과를 측정·비교하는데 있어서, 표본추출(sampling error)의 영향, 표본의 구성비 영향, 기존의 전통적 위험예측치의 문제점등을 살펴보고, 새로운 위험예측치를 제시하여 실증적으로 비교, 검증하는 것을 연구의 주목적으로 하고 있다.

Keywords: 데이터마이닝, 표본추출, 표본구성, 성과비교, ROC

I. 서 론

최근 인공신경망(neural network)이나 사례기반추론(case based reasoning)같은 인공지능 기법을 부도예측이나 사기적발 등 주로 이원화된 위험분류에 사용하여 통계적인 기법 또는 전문가의 판단과 비교한 후, 인공지능기법이 위험예측력에서 우수하여 활용가치가 매우 높다는 연구가 많이 이루어져왔다.[22, 56, 57] 특히 이러한 여러 가지 데이터마이닝 기법은 최근 주목받고 있는 데이터베이스를 이용한 데이터베이스 마케팅이나 데이터 웨어하우징과 같은 새로운 비즈니스의 중요한 활용도구로서 더욱 더 그 가치가 중요시되고 있다.[57]

그러나 과거의 데이터마이닝 기법의 예측력을 측정하여 비교하는데 주로 이용된 표본설계나 전통적인 측정치는 다음과 같은 문제점을 있어 그 실증적 가치에 대한 의문이 제기되고 있다. 첫째, 모집단에서 추출한 적은 수의 표본을 무작위로 분석표본과 검증표본으로 분류하여 계산된 예측력이 추출 표본별로 차이(표본추출오차)가 매우 커서 실증적 비교의 신뢰가 떨어진다는 것이다.[10, 22, 23] 둘째, 표본의 추출시 모집단의 구성비에 기준하여 표본의 구성비가 balanced sampling(부실:건전=50:50)과 representative sampling(모집단의 구성비와 동일)사이에서 예측력의 편차가 커서, 데이터마이닝 기법의 실무적용에서의 유용성이 의심된다는 것이다.[31] 마지막으로 현재까지 가장 많이 사용되고 있는 전통적인 위험예측측정치(옳게 예측한 경우/모든 경우)는 모집단의 사전확률을 고려하지 못하며, 특히 이원화 이상의 위험분류에서는 예측력의 상대적 정확도를 측정하지 못하는 문제가 있다. 이러한 문제점을 극복하고자 1종오류별, 2종오류별, 또는 이러한 오류들에 가중치를 부여한 효율적 측정치(weighted efficiency measure)등이 사용되었지만, 근본적인 모형성과의 비교를 위한 측정치의 기준이 현재까지 정립이 되어있지 않은 형편이다.[54]

따라서 본 연구에서는 전통적인 위험예측치에 대한 대안으로 자연과학 연구영역에서 위험분류예측치로 많이 사용되지만 사회과학 연구분야에서는 비교적 생소한 ROC(Receive Operating Curve)의 사용을 제안하고, 다양한 표본 설계 하에서 전통적인 예측치들과 ROC의 예측력을 비교·분석하여 ROC의 사용에 대한 실증적 검증을 하고자 한다. 또한 다양한 표본 설계를 시도하여 데이터마이닝 기법들의 성과 비교에 객관적인 기준을 제시하고자 한다. 특히 본 연구는 자료의 특성별로 여러 가지 데이터마이닝 기법의 장점을 통합한 통합모형의 유용성이 어디에서 비롯되는지에 대한 이론적, 실증적 검증을 거침으로써 바람직한 통합모형의 개발 방향을 제시하고자 한다.

II. 연구내용 및 연구방법

1. 연구내용

본 연구의 주된 내용은 데이터마이닝으로부터 개발된 각 모형들의 예측력을 객관적으로 비교하고 검증할 수 있는 기준 및 새로운 위험측정치을 제공하고자 하는 것이다. 이러한 객관적인 기준으로 고려해야 할 것은 크게 나누어 3 가지로 볼 수 있다. 첫

째는 모집단으로부터 표본을 추출하여 무작위로 표본을 분류하는 것에 관한 연구이다. 모집단으로부터 추출된 표본을 무작위로 분석표본(training sample)과 검증표본(holdout sample)으로 분류하여 모형개발은 분석표본에서 하고 예측력은 검증표본에서 계산하는 것이 데이터마이닝 연구의 주된 표본추출 및 분석방법이다. 그러나 표본의 수가 적다면 이렇게 무작위로 두 표본을 분류할 때마다 검증표본에서 계산된 예측력은 차이가 많아 표본 추출 오차가 발생한다. 이미 많은 연구가 적은 표본수를 이용하여[1] 단 1회의 표본분류를 한 후 검증표본에서 데이터마이닝 기법의 예측력 비교를 하였으나, 다양한 표본추출을 시도했던 소수의 기존 연구는 그 편차가 매우 심하다는 것을 실증적으로 보여왔다.[10, 22, 23] 따라서 본 연구에서는 다양한 수의 표본수(15개부터 1000개까지)를 이용하여 여러 개의 표본분류(1회부터 30회)를 시도한 후, 표본추출오차를 실증적으로 검증하여 표본 분류의 문제점을 확인하고, 더 나아가 적절한 수의 표본의 분류에 대한 기준을 제시하고자 한다. 특히 최근 중소기업 부도에서 중시되는 정성적 정보의 특성을 고려하여, 자료의 특성별(정성적 비재무정보, 정량적 재무정보)로 표본추출오차를 검증하고자 한다.

둘째, 표본설계시의 구성비에 관한 연구이다. 중소기업에서 실제 발생하는 부도비율은 대개 1% 미만이며 카드나 보험사기, 신용평가 등 다른 이원화 위험분류도 비슷한 형편이다. 그러나 기존의 데이터마이닝에 관한 많은 연구는 부도와 건전의 구성비가 동일한 balanced sampling을 많이 사용하였다.[31] 물론 여기에는 많은 학자의 견해가 대립되고 있지만[2], 적어도 모집단의 구성비와 표본의 구성비가 다를 때 어떻게 예측력에 영향을 주는지에 관한 체계적인 연구가 필요하며, 특히 각 데이터마이닝 기법별로 어떤 차이가 나는지에 관한 연구가 중요하다.[23, 31, 56] 이러한 연구는 각 데이터마이닝 방법의 장점을 통합하는 통합모형의 개발에 이론적 근거를 제시하기 때문에 더욱 더 필요한 연구라 할 수 있다. 따라서 본 연구에서는 표본의 구성비를 다양하게 변화(50:50부터 모집단의 구성비까지)시키면서 각 데이터마이닝 및 통합모형의 예측력을 비교함으로써 각 구성비별 예측력에 미치는 영향을 조사하여 바람직한 데이터마이닝 기법의 개발방향을 제시하며, 더 나아가 바람직한 통합모형의 이론적 근거를 제시하고자 한다.

마지막으로 위험예측력의 측정치(measure)에 관한 연구이다. 기존의 많은 연구는 데이터마이닝 기법의 성과비교에서 전통적인 위험예측치(옳게 분류한 경우/ 모든 경우)를 주로 사용하였다. 그러나 중소기업 부도와 같이 부도의 경우가 희귀하고, 부도 발생시의 비용이 높은 경우 전통적인 예측치는 효과적이지 못한 경우가 많다. 따라서 많은 연구들이 1종오류, 2종오류별로 각 모형을 비교하거나, 또는 1종오류와 2종오류를 동시에 고려한 측정치를 개발하여 사용하였다. [31] 그러나 이러한 측정치들은 여전히 표본 구성비의 영향을 받아 사전확률(prior probability)에 대한 고려가 없으며, 특히 위험분류가 이원화 이상인 경우에는 분석이 불가능하거나, 이원화로 변환하여 분류하더라도 예측력의 강도를 측정하지 못하는 단점이 발생한다.

따라서 본 연구에서는 자연과학(특히 의학 및 생물학)분야의 위험분류에서 많이 사용되는 ROC(Receiver Operating Curve)의 도입을 제안하며[27], ROC의 실증적 유용성에 대한 검증을 하고자 한다. ROC는 이원화 분류의 경우 1종오류와 2종오류를 동시에 고려하여 그래프로 보여주는 유용성을 가지고 있으며, 특히 ROC 아래의 면적(일명 θ)으로 위험예측력을 측정할 수 있는 측정치로서의 요건도 가지고 있다. 특히 사전확률을 고려하여 표본의 구성비에 대한 비율을 고려한 위험예측력을 보여 줄 수 있으며, 마지막으로 이원화 이상의 위험분류에서 위험의 예측력 강도도 고려할 수 있는 매우 유용한 위험측정치라 하겠다. 본 연구에서는 앞에서 기술한 다양한 표본추출 및 설계단계의 실증적 검증을 할 때, 기존의 여러 예측치와 더불어 ROC를 사용함으로써 향후 사회과학 분야에서의 ROC 사용에 대한 이론적, 실증적 타당성을 검증하고자 한다. 특히 통합모형의 개발시 이원화 이상의 위험분류에서의 ROC 유용성도 검증하고자 한다.

현재 고려하고 있는 본 연구의 범위로는 데이터마이닝 기법 중 NN(인공신경망), CBR(사례기반추론), MDA(다중판별분석), ID3, 통합모형이며, 실증적 검증은 중소기업부도예측의 경우로 한정하고자 한다. 특히 표본 설계와 관련된 위험분류는 이원화된 위험분류만을 고려하고 있으며, 자료의 특성별 비교는 정성적 비재무정보와 정량적 재무정보만을 이용하고자 한다. 그 외 위험분류 기준치(cut-point)의 영향, 여러 가지 다른 데이터마이닝 기법이나 다양한 문제영역 등은 본 연구의 범위에 들어있지 않다.

2. 연구방법

본 연구의 주된 방법은 중소기업의 건전 및 부도자료를 수집하여 표본을 추출하고 무작위로 여러 개의 분석표본과 검증표본으로 분류하여, 다양한 표본 구성비를 가

1) Altman(1968)은 부실 및 건전기업의 표본이 각각 15-20개 정도라도 신뢰할 수 있는 예측모형이 가능하다고 하였다

2) Hair et al.(1995)은 판별분석의 경우, 판별력의 증대를 위하여 각 집단의 구성비가 동일한 것이 바람직하다는 견해를 표명했다

진 여러 분석표본에서 데이터마이닝 기법을 이용하여 부도예측 모델을 개발하고, 검증표본에서 기존의 전통적인 예측치 및 ROC를 계산하여 위험예측력을 측정하고 비교, 분석하는 것이다.

(1) 자료수집

본 연구를 위한 자료는 기업신용대출을 하고 있는 일반 회사들의 자료를 이용할 것이다. 보험회사, 신용보증기금 같은 여러 회사들은 자산의 운용을 위해서 일반기업들에게 신용 대출을 하고 있다. 특히 이러한 회사들은 재무비율과 같은 정량적 자료이외에도 기업형태 또는 경영능력같은 정성적 요소들도 많이 활용하는 것으로 알려져 있다. 정량적 자료항목으로 고려되어지는 것은 자기 자본비율, 고정장기 적합율, 차입금의 속도, Cash Flow/총부채, 경상수지 비율, 금융비용/매출액, 경영자본 영업 이익률, 매출채권 회전율, 매출액 증가율, 매출액 등이다. 정성적 자료항목들은 기업형태, 거래신뢰도, 대외신뢰도, 업계지위, 경영 및 기술력, 업종유망성, 판매전망, 수익성 전망, 자금전망 등이다.[14, 32] 본 연구를 위하여 약 200여 개의 부도를 포함한 총 2,000여 개의 자료수집을 하고자 한다.

(2) 표본설계

먼저 표본추출오차의 영향을 살펴보기 위하여, Altman[14]이 주장한 15-20개를 최저표본수로 하여 표본수를 10개씩 늘려가면서 표본을 추출하여 총 20개의 표본을 추출한다. 이렇게 추출된 각 표본을 분석표본과 검증표본으로 무작위 분류하는 것을 각 20회씩 실시한다. 따라서 각 표본수 별로 여러 가지 데이터마이닝 모델을 개발하고 예측력을 계산하여 표준편차를 계산하면, 각 표본수 별로 20개의 표준편차가 계산될 것이다. 이 표준편차를 분석하여 단 1회의 분석표본 및 검증표본 분류로 최소한의 안정된 예측력을 얻기 위하여서는 기본 표본수가 몇 개 이상은 되어야 한다는 결론을 얻고자 한다.

표본설계의 영향을 알아보기 위하여 앞에서 설정된 표본추출오차가 없는 기본 표본수를 중심으로 표본을 추출하고 표본을 단 1회의 시행으로 분석표본과 검증표본으로 나눈다. 그 다음 분석표본을 건전과 부도가 50:50의 비율로 나누어지는 balanced sampling을 시작으로 건전의 비율을 10%씩 증가시켜 최종적으로 모집단의 구성비와 비슷한 구조가 이루어지는 시점(representative sampling)까지 분석표본을 분류한다. 이렇게 다양한 구성비를 가진 여러 분석표본에서 데이터마이닝 기법 및 통합모형을 이용한 부도예측 모형을 개발하고 동일한 검증표본에서 예측력을 계산한다. 이러한 예측력을 비교함으로써 표본구성비에 의한 각 데이터마이닝의 기법의 특성을 비교분석하고 특히 통합모형과도 비교분석함으로써 통합모형의 예측력이 어떻게 생성되고 있는지에 대한 실증적 검증을 할 것이다. 표본구성비별로 차이가 존재하는 지에 대한 분석기법은 상호작용을 고려한 분산분석을 실시할 것이다.

또한 기존의 많은 학자들이 주장한 통계적 기법의 사용시 balanced sampling의 효과에 대한 실증적 검증도 병행하고 아울러 인공지능을 사용한 데이터마이닝 기법에도 이러한 효과가 있는지에 대한 검증을 하고자 한다. 마지막으로 데이터마이닝기법이 실제 비즈니스에서 사용될 경우에 예측력 증가를 위한 바람직한 표본구성비에 대한 방향도 분석될 것으로 기대한다.

(3) 여러 가지 위험측정치의 성과비교

본 연구에서 검증될 전통적인 부도위험예측치는 다음과 같다. [31,53]

	건전으로예측	부도로 예측
실제로 건전	TP (True Positive)	FN (False Negative)
실제로 부도	FP (False Positive)	TN (True Negative)

①전통적인위험예측치(PredictionAccuracy)
= (TP+TN)/(TP+TN+FP+FN)
② Sensitivity = TP/(TP+FN)
③ Specificity = TN/(FP+TN)
④ Positive Predictive Value
= TP/(TP+FP)
⑤ Negative Predictive Value
= TN/(FN+TN)
⑥ Type I Error = FN/(TP+FN)
= 1 - Sensitivity
⑦ Type II Error = FP/(FP+TN)
= 1 - Specificity
⑧ Weighted Efficiency
=① * ② * (TP/(TP+TN+FP+FN))

이러한 전통적 또는 기존의 연구에서 개발되어 사용된 위험예측치 이외에 본 연구에서 새로이 제시하여 검증할 위험 예측치는 ROC (Receiver Operating Curve)이다.

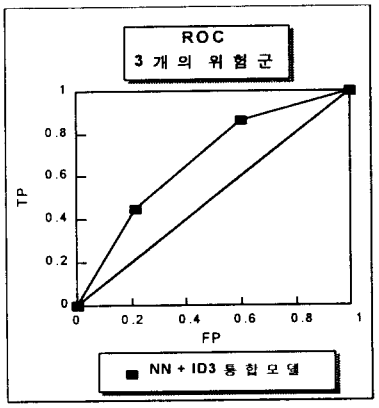
(4) 새로운 위험예측치: ROC(Receiver Operating Curve)

부도 경우가 건전 경우에 비해서 매우 희귀하므로 부도예측율 문제는 부도 그 자체에 대해서는 매우 작은 사전확률(prior

probability)을 갖고 있다. 따라서 부도예측의 경우 확률만 고려하여 볼 때, 건전하다고 예측하는 것이 맞을 확률이 높으므로 건전 경우를 부도라고 잘못 예측하는 것은 부도를 건전하다고 잘못 예측하는 것보다 더 큰 실수가 될 것이다. 그러나 여러 가지 데이터마이닝 기법의 유효성을 측정하는데 많이 이용된 전통적인 예측정확도(옳게 예측한 경우/총 경우)는 이러한 사전확률을 고려하지 않고 있다.

특히 위험분류군이 2원화 이상인 경우에는 전통적인 예측정확도는 예측 정확도의 강도가 전혀 고려되지 않은 단점이 있다. 예를 들어 종래의 예측정확도는 5개의 위험분류군을 가진 경우 실제 "부도"경우를 "요주의 관찰"로 예측한 것이나 "건전"으로 예측한 것이 모두 같이 틀린 예측으로 간주된다. 그러나 실제 "건전"이라는 예측이 "요주의 관찰"보다 더욱 잘못된 예측이나 이러한 예측정확성의 강도를 종래의 예측정확도 측정치는 고려하지 못한다는 것이다. 그러나 ROC는 이러한 사전확률 및 예측정확도의 강도를 고려할 수 있는 측정치이기 때문에 본 연구에서는 모델의 예측력 검증시 이미 의학이나 생물학 연구영역의 위험분류 예측력측정에서 사용이 보편화된 ROC를 사용하고자 한다.

<그림 2> ROC의 예 (3개의 위험분류군의 경우)



예를 들어 3개의 위험군을 가진 NN과 ID3를 통합한 모델에 관한 <그림 2>의 ROC는 NN, MDA 또는 통합모델과 같은 위험분류도구(risk classifier)가 생성한 위험추정치(연속값 또는 서열값)를 각 위험을 분류하는 최적 임계치를 기준으로 여러 가지 위험군(risk category)으로 나눈 후 Sensitivity (true positive: TP) 대 (1 - Specificity) (false positive: FP)의 비율을 각 위험군(Risk category) 별로 계산해서 그 값들을 연결하여 그리는 것이다. 이상적인 ROC 곡선은 왼쪽위로 가장 가깝게 붙어있는 것이다. 예를 들어 MDA의 부도확률추정치를 세 개의 위험군(건전, 보통, 부도)으로 나눌 때, 최적의 임계치는 0.33 과 0.66의 두 개가 되고, 각 임계치 별로 부도와 건전을 분류하는 2개의 교차 테이블(Cross Table)을 완성하여 Sensitivity(TP) 대 (1-Specificity)(FP)의 비율을 계산하여 그 값을 연결하면 ROC 가 완성된다. NN의 경우는 결과추정치가 확률값이 아니고 순서의 의미를 갖고있는 서열치이므로 반복시행(trial and error)을 통하여 검증표본을 최적으로 분류하는 두 개의 임계치를 찾아서 같은 방법으로 ROC를 완성한다

ROC의 아래면적인 θ(Theta)는 무작위로 고른 비정상 경우가 무작위로 고른 정상인 경우보다 비정상으로 의심되는 정도가 더 큰 확률을 나타내고 있다. 즉 θ 값이 1이면 완벽한 예측, 0.5이면 무작위 추측(그림 2의 대각선), 0.5보다 작으면 무작위 추측보다 작은 것을 나타낸다. 극단적인 예로, 99개의 건전 경우와 1개의 비건전경우의 자료를 무작위로 뽑을 때 종래의 예측도 측정치는 99%의 정확도를 기본 예측률로 갖으나 ROC 측정치는 0.5에 가까운 값을 갖게 된다. 또한 ROC 아래의 면적 중 대각선위의 왼쪽하단면적보다 오른쪽상단면적이 크면 부도의 경우를 건전의 경우보다 잘 예측하는 것이며, 그 반대로 오른쪽상단면적보다 왼쪽하단면적이 크면 건전의 경우를 부도의 경우보다 잘 예측하는 것을 나타낸다. Mayer & Riedinger (1995)는 보험사기 적발모형에 ROC를 사용하였으며, Hanley & McNeil (1982)은 ROC의 이론과 통계학적 특성에 대한 상세한 설명을 했다.

III. 기존 문헌연구

기존의 데이터마이닝 기법을 사용하여 위험분류를 한 연구 중 표본출이나, 표본설계, 또는 측정치의 문제가 있을 수 있음을 인식하여 나름대로 이러한 한계점을 단편적으로나마 연구한 논문들은 <표 2>에 따로 모아 정리하였다. 그 중 국내문헌 중 피종허(1995)나 정기웅과 홍관수(1995)는 표본추출이나 표본설계를 나름대로 다양하게 시도하였으나, 체계적으로 이러한 영향들이 예측력에 어떻게 미치는 지를 연구하지는 못하였다. 다만 표본추출의 영향이 존재한다는 실증적 점검만을 하였다. 우춘식외 2인(1997)은 표본 추출의 영향을 체계적으로 분석하고 적은 수의 표본분류는 많은 예측력오차가 있어 무작위 추출을 통한 예측력의 신뢰도에 의문을 제기하였다.

특히 Jain & Nug(1997)는 표본설계에서 좀더 체계적으로 다양한 표본구성비를 만들고 분석한 후, balanced sample이 반드시 높은 예측력을 보장하지 않는 다는 것을 실증적으로 보여주고 representative sampling의 접근이 필요함을 역설하였으나 표본추출오차의 영향은 분석하지 못하였으

며, 자료별 특징도 분석하지 않았고 특히 에서 위험분류예측치로 많이 사용되지만 사

<표 2> 기존연구의 정리

논문	사용된 표본수	표본 분류	데이터 마이닝 기법	구성비 (건전:부도)	측정치
이건창 외 2(1994)	검증:100,70,40 분석:66,96,126	3회	MDA, NN, ID3, Hybrid(NN+ID3)	50:50	PA
이건창(1995)	검증:100,70,40 분석:66,96,126	3회	MDA, ACLS, NN, HYNEN,ARTMAP	50:50	PA Type I, II
장정곤(1995)	검증: 100,분석: 100	1회	NN, MDA	50:50	PA
정기웅&홍관수(1995)	검증: 40, 분석: 40	4회	NN, MDA	50:50,75:25	PA
피종허(1995)	검증: 36,18,5 분석: 36,54,67	10회	NN	50:50	PA
노시천(1996)	검증: 32, 분석: 68	1회	MDA	50:50	PA
이재식&한재홍(1996)	검증: 60, 분석: 60	1회	NN, CBR	50:50	PA
우춘식 외2인(1997)	검증: 35, 분석: 30	20회	LOGIT, AHP	50:50	PA
Liang(1990)	검증: 20, 분석: 30	6회	DA, ID3,hybrid	50:50	PA
Schkada et al(1991)	검증: 52, 분석: 50	1회	NN	50:50	PA
Tam(1991)	검증: 44, 분석: 118	1회	DA, ID3, NN	50:50	PA
Altmon et al(1992)	검증:302, 분석: 808	1회	DA, NN	50:50	PA
Cinar et al(1992)	검증(1): 116, 검증(2): 404, 분석: 200	1회	Logit, NN	50:50	PA
Cronan et al(1992)	검증:100, 분석: 375	1회	RPA, ID3	50:50	PA
Tam & Kiang(1992)	검증: 44, 분석: 118	1회	DA,Logit, NN, ID3	50:50	TypeI,II, PA
Hansen et al(1993)	80	1회	Logit,ID3, NEWQ	50:50	Type I,II
Wilson&Sharda(1994)	검증: 40, 분석: 88	3회	NN, DA	50:50,80:20 90:10	PA
Lee et al(1996)	검증:100,70,40 분석:66,96126	1회	MDA,ID3,Hybrid (MDA+NN)(ID3+NN)	50:50	PA
Major & Riedinger (1995)	분석: 15,000 검증: 15,000	1회	Ststaistic, ES	50:50	**ROC**
Jain & Nag(1997)	검증: 231 분석: 230	1회	NN, Logit	50:50,45:55 40:60,35:65 30:70	Type I,II Weighted efficiency

* PA: 전통적인 예측치

기존의 전통적예측치에 의한 결론으로 신뢰도가 떨어지는 문제점이 있었다. Major & Riedinger (1995)는 데이터마이닝관련된 연구에서는 처음으로 ROC 측정치를 이용하였으며 실제 회사에서 이루어지고 있는 보험사기적발모형을 소개하였다.

IV. 연구결과의 기대효과

본 연구는 데이터마이닝으로부터 개발된 각 모형들의 예측력을 객관적으로 비교하고 검증할 수 있는 표본설계 기준과 새로운 측정치를 제공하고자 하는 목적을 갖고 있다. 따라서 본 연구에서는 전통적인 위험예측치에 대한 대안으로 자연과학 연구영역 회과학 연구분야에서는 비교적 생소한 ROC(Receive Operating Curve)의 사용을 제안하고, 다양한 표본 설계하에서 ROC의 예측력을 전통적인 예측치들과 실증적으로 비교·분석하여 ROC의 사용에 대한 타당성을 검증하고자 한다. 특히 추출된 표본을 무작위로 분류할 때 발생할 수 있는 표본추출오차의 문제점과 표본의 구성비에 따른 예측력의 차이를 지적하고 실증적으로 검증함으로써, 기존 데이터마이닝 기법의 성과비교에 대한 연구보고에서 발생할 수 있는 오류를 미연에 방지하고, 데이터마이닝 연구의 객관적인 표본설계 기준을 제시함으로써 학문의 질적 향상에 기여하고자 한다.

또한 이러한 연구는 최근 중소기업 부도예측에서 그 중요성이 더욱 더 강조되고 있는 정성적 정보의 영향을 좀더 자세히 살펴

보고자 자료를 정성적 비재무정보, 정량적 재무정보, 또는 모든 정보로 나누어 분석함으로써 자료의 특성상 예기되는 부도위험예측의 차이를 실증적으로 검증하고자 한다.

또한 각 개별적 데이터마이닝 기법을 통합모형과 분석, 비교하고 실증적으로 검증함으로써, 여러 가지 데이터마이닝의 장점을 흡수한 통합모형의 높은 예측력이 어디에서 나오는지를 실증적으로 검증할 수 있어, 바람직한 통합모형의 개발방향에 이론적, 실증적 기여를 할 것으로 기대된다.

마지막으로 데이터마이닝 기법을 실제 모집단에 적용할 경우 어떤 표본설계방법이 가장 우수한 예측력을 보이는지에 대한 기준을 제시하여 데이터마이닝 기법의 비즈니스 사용에 대한 실용적인 개발지침을 제시하고, 더 나아가 중소기업 신용평가에 데이터마이닝을 이용하는 신기법의 정착으로 기업경쟁력을 확보하는데 일조를 하고자 한다.

V. 참고문헌

[1] 김충섭, 남기정, "중소기업 신용평가의 질적요인에 관한 실증적 연구", 보증월보, 제17호(1997), pp.29-72

[2] 노시천, "우리나라 중소기업의 부실화원인 및 그 예방대책에 관한 실증연구," 성균관대학교 박사논문, 1996.

[3] 박상천, Lam P.S., Gupta, A, "Rule Extraction from Neural Naetworks: Enhancing the Explanation Capability," 한국전문가시스템학회지, 제1권, 제2호(1995)

[4] 이건창, "기업도산 예측을 위한 통계적 모형과 인공지능모형간의 예측력 비교에 관한 연구: MDA, 귀납적학습방법, 인공신경망," 한국경영과학회지, 제18권, 제2호, (1993). pp.57-81

[5] 이건창, 김명종, 김혁, "기업도산 예측을 위한 귀납적 학습지원 인공신경망 접근방법: MDA, 귀납적학습방법, 인공신경망, 모형과의 성과비교," 한국경영학회지, 제23권, 제3호, (1994). pp109-143

[6] 이건창, "효과적인 의사결정을 위한 2단계하이브리드 인공신경망 접근방법에 관한 연구," 한국경영정보학지, 제5권, 제1호, (1995). pp36-51

[7] 이건창, 한인구, 김명종, "통계적모형과 인공지능모형을 결합한 기업신용평가 모형에 관한 연구," 한국경영과학회지, 제21권, 제1호, (1996). pp.81-101

[8] 이재식, 한재홍, "사례기반추론을 이용한 중소기업 도산 예측에 있어서의 비재무정보의 활용," 한국전문가시스템학회, (1996). 243-252.

[9] 이재식, 한재홍, "NN을 이용한 중소기업도산예측에 있어서의 비재무정보의 유용성 검증", 한국전문가시스템학회지, 제1권, 제1호(1995), pp.123-134

[10] 우춘식, 김광용, 강성범, "LOGIT 분석과 AHP 분석을 이용한 부도예측모형의 비교연구," 재무관리연구, 제14권, 제2호(1997), pp.229-252

[11] 장정곤, "중소기업 부실예측을 위한 통계적 접근과 신경망 접근에 관한 비교 연구," 성균관대학교 박사논문, 1995.

[12] 정기웅, 홍관수, "신경망기법을 이용한 기업부실예측에 관한 연구," 재무관리연구, 제12권, (1995), pp.1-23.

[13] 피종허, "한정된 데이터 하에서 인공신경망을 이용한 기업도산예측 - 섬유 및 의류 산업을 중심으로," 고려대학교 석사논문, 1995.

[14] Altman, E. I., "Financial Ratios, discriminant analysis, and prediction of corporate bankruptcy," *The Journal of Finance,* Vol.23(1968), pp.589-609.

[15] Arizne, B., & Narasimha, P. N. "An experimental investigation of predictive accuracy of induction and regression," *Expert Systems with Applications,* Vol.7(1994), pp.535-544.

[16] Braun, H., & Chandler, J. "Predicting stock market behavior through rule induction: An application of the learning-from-example approach," *Decision Sciences,* 18(1987), pp.415-429.

[17] Carter, C., & Catlett, J. "Assessing credit card applications using machine learning," *IEEE Expert*, Vol.2(1987), pp,71-79.

[18] Chandler, J. S., Liang, T., & Han, I.. "An empirical investigation of some date effects on the classification accuracy of Probit and ID3," *Journal of Contemporary Accounting,* Vol,9(1992), pp,306-328.

[19] Chung, H. M., & Silver, M. S. "Rule-based expert systems and linear models: An empirical comparison of

learning-by-examples methods," *Decision Sciences,* Vol,23(1992), pp,687-707.

[20] Cronan, T. P., Glorfeld, L. W., & Preey, L. G. "Production system development for expert systems using a recursive partitioning induction approach," *Decision Sciences*, Vol,22(1991), pp,812-845.

[21] Deng, P. "Automating knowledge acquisition and refinement for decision support: A connectionist inductive inference model," *Decision Science, Vol,*24(1993), pp,371-393.

[22] Gim, G. "Hybrid Systems for Robustness and Perspicuity: Symbolic Rule Induction combined with a Neural nets or a Statistical Models," *Doctoral dissertation*, Georgia State University. (1995).

[23] Gim, G. & Whalen, T. "Second Order Logical System for Risk Classification in a Newly Developed Country," *International Journal of Uncertainty, Fuzziness, and Knowledge Based System*, Vol,4 No5(1996), pp,421-430

[24] Gim, G., & Whalen, T. "Dimensions of knowledge: Facts or skills, words or numbers," *The Proceedings of North American Fuzzy Information Proceeding Society.* (1994), pp,447-448.

[25] Gim, G., Whalen, T, & Schott, B. "Control of Error in Fuzzy Logic Modeling," *International Journal of Fuzzy Sets and System,* Vol.80, .No.1(1996), pp.23-35

[26] Goul, M., Henderson, J. C., & Tonge, F. M. "The emergence of AI as a reference for DSS research," *Decision Science*, Vol.23 (1992), pp.1263-1276.

[27] Hanley, J., & McNeil, B. "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, Vol.143(1982), pp.29-36.

[28] Hansen, J.V., Koehler, G.J., Messier, W. F., & Mutchler, J. F., "Developing knowledge structures," *Decision Support Systems*, Vol.10(1993), pp. 235-243.

[29] Hansen, J. V., McDonald, J. B., & Stice, J. D. "Artificial intelligence and generalized qualitative-response models: An empirical test on two audit decision making domains," *Decision Sciences,* Vol.23(1992), pp.708-723.

[30] Jacobstein, N., & Kitzmiller, C. T. "Integrating symbolic and numeric methods in knowledge-based systems. In L. S. Kowalik & C. T. Kitzmiller (Eds.)," *Coupling symbolic and numerical computing in expert systems* ,1988, pp.3-14 New York: Elsevier Science.

[31] Jain B.A. and Nag B.N., "Performance Evaluation of Neural Network Decision Models," Journal of Management Information Systems, Vol.14, No.2(1997), pp.201-216

[32] Jones, F. L. "Current Techniques in bankruptcy Prediction," *Journal of Accounting Literature*, Vol.6(1987), pp.131-164.

[33] Kandel, A., & Langholz, G. " Hybrid architectures for intelligent systemsm," *Boca Raton,* 1992, FL: CRC press.

[34] Kattan, M. W., Adams, D. A., & Parks, M. S. "A comparison of machine learning with human judgment," *Journal of management Information Systems,* Vol.9, No.4(1993), pp.37-57.

[35] Kiang, M. Y., Chi, R. T., & Tam, K. Y. "DKAS: A distributed knowledge acquisition system in a DSS," *Journal of Management Information Systems*, Vol.9, No.4(1993), pp.59-82.

[36] Kononenko, I., & Bratko, I. "Information-based evaluation criterion for classifier's performance," *Machine Learning*, Vol.6(1991), pp.67-80.

[37] Lee K.C., Han I., & Kwon Y., "Hybrid Neural Network Models for Bankruptcy Prediction," *Decision Support System,* Vol.18, No.1(1996), pp.63-72

[38] Liang, T. P. "A composite approach to inducing knowledge for ES design," *Management Science*, Vol.38(1992), pp.1-17.

[39] Liang, T. P., Chandler, J. S., & Han, I. "Integrating statistical and inductive learning methods for knowledge acquisition," *Expert Systems with Applications*, Vol.1(1990), pp.391-401.

[40] Liang, T. P., Chandler, J. H., Han, I., & Roan, J. "An empirical investigation of some data effects on the classification accuracy of Probit, ID3, and neural networks," *Contempory Accounting Research*, Vol.9(1992), pp.306-328.

[41] Meisser, W. F., & Hansen, J. V. "Inducing rules for expert systems development: An example using default and bankruptcy data," *Management Science,* Vol,34(1988), pp.1403-1415.

[42] Minger, W. F. "Rule induction with statistical data," *Journal of the Operational Research Society*, Vol,38(1987), pp.347-351.

[43] Quinlan, J.R "Discovering rules by induction from large collection of examples," In D. Michie (Eds.), *Expert systems in the micro electronic age*. 1979. Edinburgh, Scotland: Edinburgh University Press.

[44] Quinlan, J.R. "Induction of decision trees," *Machine Learning*, Vol,1(1986), pp.81-98.

[45] Rumelhart, D. E., Hinton,G., & Williams, R. "Learning internal representation by error propagation," In D. Rumelhart and J. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (1986), pp.318-362. Cambridge: MIT Press.

[46] Salchenberger, L.M., Cinar, E.M., & Lash, N.A., "Neural Networks: A New Tool for Predicting Thrift Failures," *Decision Sciences*, Vol.23(1992), pp. 899-916.

[47] Shaw, M. J., & Gentry, J. A. "Inductive learning for risk classification," *IEEE Expert,* Vol,2(1990), pp.47-53.

[48] Shavlik, J. W., Mooney, R. J., & Towell, G. G. "Symbolic and neural learning algorithms: An experimental comparison," *Machine Learning*, Vol,6(1991), pp. 111-143.

[49] Spiegehalter, D. J. "A statistical view of uncertainty in expert systems," In W. J. Gale (Eds.), *Artificial intelligence & statistics* , 1986, pp.17-56. Princeton: AT & T Bell laboratories.

[50] Tam, K.Y., & Kiang, M.Y. "Managerial applications of neural networks: The case of bank failure predictions," *Management Science*, Vol,38(1992), pp.926-947.

[51] Tam, K. Y. "Neural network models and the prediction of bank bankruptcy," *Omega: International Journal of Management Science,* Vol,19(1991), pp.429-445.

[52] Trippi, R. R., & Turban, E. "Neural networks in finance and investing: Using artificial intelligence to improve real-world performance," 1993, Chicago: Probus.

[53] Weiss, S. M. & Kulikowski, C. A, "Computer systems that learn classification and prediction methods from statistics, neural nets, machine learning, and expert systems," 1991, San Francisco: Morgan Kaufmann.

[54] Whalen, T., & Gim, G. "Hybrid neural-statistical classification system for potential medical malpractice claims," *Proceedings of the Second International Decision Science Institute Conference,* 1993 , pp.216-219.

[55] Wilson, R.L., & Sharda, R., "Bankruptcy prediction using neural networks," *Decision Support Systems*, Vol.11(1994), pp. 545-557.

[56] Wong B.K. , Bodnovich T.A., & Selvi Y., "Neural Network Applications in Business: A review and Analysis of the literature(1988-1995)," *Decision Support System,* Vol.19, No.4(1997), pp.301-320

[57] Zopounidis C., Doumpos M., & Matsatsinis N.F., "On the Use of Knowledge Based DSS in Financial Management:A Survey," *Decision Support Systems,* Vol.20, No.2(1997), pp.259-277