SVM 전처리기를 활용한 코스닥기업 도산예측 성능 향상

강부식*, 조준희**

요약

기업 도산은 다양한 이해관계자에게 사회 경제적으로 큰 손실을 주게 된다. 따라서 기업 도산 및 부실화를 사전에 예측할 수 있다면 이에 대한 대비를 하거나 기업 부도 요인을 사전 제거함에 따라 기업 도산에 대한 손실을 최소화하는 것이 가능할 것이다. 기업의 도산을 예측하기 위한 다양한 통계적 모형 및 데이터 마이닝 모형이 제시되고 있다. 기업 도산 예측 모형의 주요 이슈 중 하나는 예측력을 높이는 것이다. 이 논문은 기업 도산예측에 주로 사용되며 예측성능이 비교적 높은 로짓모형, 의사결정나무모형, 신경망모형, SVM모형에 대해서 살펴본다. 표본기업은 코스닥기업 중 도산 및 정상기업으로 각각 49개 기업을 선정하였고, 설명변수로는 통계적으로 유의미한 9개의 재무변수를 이용하고, 5년간의 재무변수 자료를 사용하였다. SVM 모형을 전처리기를 사용한 경우 다양한 분류모형의 도산예측 성능을 크게 높일 수 있음을 실험결과로 보이고 있다.

Prediction Performance Improvement of KOSDAQ Business Bankruptcy using SVM Preprocessor

Boo-Sik Kang*, Jun-Hee Cho**

ABSTRACT

Business bankruptcy has had a lot of losses to various stakeholders. If we are able to predict bankruptcy or insolvency in advance, we can defend enterprises against bankruptcy or insolvency and minimize the loss. Various types of statistical and data mining models have been proposed for prediction of business bankruptcy. One of the main issues of the bankruptcy prediction models is to increase predictability. This research deals with logit, decision tree, neural networks, and SVM models those are mainly used to predict corporate bankruptcy. Sample companies were consisted of 49 bankrupt firms and 49 normal firms in KOSDAQ. Nine financial variables were selected by statistical test and five years of data were collected. In case of using SVM for preprocessing data, experimental results showed that classification models could improve the performance of the bankruptcy prediction.

Key Words : Bankruptcy Prediction, Support Vector Machine, Logit, Decision Tree, Neural Networks.

* 목원대학교 서비스경영학부(✉bookang@mokwon.ac.kr)
** 목원대학교 경영학과
· 제1저자(First Author) : 강부식 · 교신저자(Correspondent Author) : 조준희
· 접수일(2012년 7월 10일), 수정일(1차 : 2012년 8월 17일), 게재확정일(2012년 8월 20일)

Ⅰ. 서 론

기업의 부도는 다양한 기업의 이해관계자에게 사회 경제적으로 많은 손실을 초래한다. 사전에 기업의 부실화를 감지하거나 예측할 수 있다면 이에 대한 대비를 하거나 더 나아가 부도의 사전 요인을 제거함으로서 기업 부도에 따른 손실을 최소화 할 수 있을 것이다.

기업의 도산예측을 위한 많은 연구가 진행되어 왔으며, 도산예측 모형의 수립 및 예측성능 향상은 기업 도산예측 연구의 주요 이슈 중 하나라고 할 수 있다.

기업 도산예측 모형으로는 다양한 통계적 기법[1,2,3,4]이나 데이터 마이닝 기법[5,6,7]이 활용되고 있다. 최근에 SVM(Supprt Vector Machine) 기법은 도산 혹은 건전과 같은 이진값을 예측하는 데 있어 성능이 좋음을 보여주고 있다[6].

이 연구에서는 KOSDAQ 기업의 도산예측을 위해 로짓(Logit) 모형, SVM 모형, 의사결정나무 모형, 인공신경망 모형과 같은 데이터마이닝 기법의 예측성능을 살펴보고자 한다. 특별히 SVM 모형을 데이터 전처리기로 사용하였을 경우 도산예측 모형의 성능을 크게 향상 시킬 수 있음을 실험을 통해 살펴보고자 한다.

Ⅱ. 관련 연구

2.1 기업 부실예측 관련 연구

Beaver[1]는 5년간 재무비율을 가지고 단변량분석에 의한 기업부실 예측모형을 개발하였다. Altman[2]은 다변량판별분석을 이용한 Z-score 예측모형을 개발하였다. 장휘용[3]은 로짓모형을 이용하여, 강종만과 홍성희[4]는 다변량판별분석의 Z-score모형과 로짓모형을 비교분석하였다.

이건창, 김명종,김혁[5]은 귀납적학습방법과 인공신경망 모형을 결합한 귀납적 학습지원 인공신경망 기법을 소개하면서 다변량판별분석에 비해 예측력이 우수함을 실증적으로 보이고 있다. 민재형과 이영찬[6]은 예측 성능이 SVM, 인공신경망, 로지스틱 회귀분석 순으로 높음을 보였다.

조준희와 강부식[7]은 KOSDAQ 기업의 도산예측을 위해 로짓모형, 의사결정나무모형, 신경망모형을 비교분석하였다. [7]의 연구에서 의사결정나무모형이 비교적 예측성능이 높음을 실험결과 제시하였다.

이 논문에서는 기업 도산예측을 위해 많이 사용되고 있는 로짓모형, 의사결정나무모형, 신경망모형, SVM 기법을 적용하여 코스닥 기업의 도산예측율을 비교하고자 한다.

2.2 분류모형

2.2.1 로짓(Logit) 모형

로짓 모형은 종속변수가 이진값을 갖고, 설명변수가 k개인 경우에 적용할 수 있으며, 설명변수에 대한 정규분포와 공분산행렬에 대한 가정을 하지 않아 기업부실 예측분석 모형으로 사용할 수 있다[8]. 로짓모형에서 모든 설명변수들을 설명변수로 선택하는 경우 다중공선성 등의 문제가 발생할 수 있다[8]. 따라서 설명변수들 중에서 종속변수를 가장 잘 설명할 수 있는 변수들을 선택하여 모형을 구축할 필요가 있다. 대표적인 변수선택방법으로는 전진선택법, 후진소거법, 단계적방법 등이 있으며[8], LogitBoost방법[9]을 이용하여 학습자료에서 가장 적절한 변수들을 자동으로 선정하는 SimpleLogistic 방법[10] 등이 있다.

2.2.2 의사결정나무 모형

의사결정나무 모형의 장점은 통계지식이 없어도 누구나 쉽게 이해할 수 있으며, 변수간의 교호관계를 잘

나타내며, 변수의 종류에 관계없이 사용할 수 있고, 계산속도가 빠르며, 대형자료처리에 용이한 점을 들 수 있다[8]. 의사결정나무 모형의 대표적인 기법중 하나는 기계학습 분야에서 활발하게 사용되고 있는 C4.5[11]이다. C4.5는 하향식 반복적 분할정복 원리를 이용하여 의사결정나무를 생성한다. 학습데이터는 의사결정나무 뿌리에서부터 시작하여 분할된다. 분할을 위한 속성선택을 위한 기준으로는 엔트로피 척도가 사용되며, 정보이득값이 크도록 학습자료가 분할을 하게 된다. 의사결정나무를 생성한 후에는 과도적합 문제를 피하기 위해 나무 가지치기 작업을 하게 되며, 이후 최종 의사결정나무를 생성한다[7][11].

2.2.3 신경망 모형

신경망은 복잡한 구조를 가진 자료의 분류 및 예측에 사용되는 비선형모형 중 하나이다[8]. 여러 신경망 모형중에서 이 연구에서는 역전파 알고리즘을 사용하는 신경망을 이용한다.

신경망의 구성은 입력층, 은닉층, 출력층 3계층으로 구성된다. 역전파 신경망은 학습자료내의 각 변수들에 대한 정규분포 등의 사전 조건이 필요없이 각 자료들의 관계나 패턴을 학습을 통해 찾아낸다. 그러나 신경망 학습을 위해서는 은닉층 내의 은닉노드의 수, 학습률, 모멘텀 등의 파라미터를 설정하여야 하는 데, 최적 파라미터를 사전에 알 수가 없다. 따라서 파라미터를 조정해 가면서 시행착오식으로 학습이 가장 잘 되는 파라미터를 찾아 나가게 된다. 신경망의 과도적합 문제를 피하기 위해서는 신경망의 구조를 간단하게 하거나[12], 최소한 학습자료의 수가 신경망을 구성하는 노드간을 연결하는 링크의 수보다는 크도록 구성해야 한다[13].

2.2.4 SVM 모형

SVM 모형은 지지도 벡터(Support Vector)라 불리는 학습사례의 부분집합을 이용하여 종속변수를 가장 잘 분류하는 의사결정 경계인 최대마진 초평면(Hyperplane)을 표현하는 기법이다[14]. 본래의 좌표공간 $\mathbf{x}$에 있는 데이터를 선형 의사결정 경계를 사용할 수 있는 새로운 좌표공간 $\Phi(\mathbf{x})$로 변환시킴으로서 초평면을 찾게 된다. 비선형 SVM에서 변환함수로 다항식 커널함수, RBF 커널함수, 탄젠트 커널함수 등이 사용되며, 고차원의 데이터에도 잘 작동되는 특징이 있다 [14].

III. SVM 전처리기를 사용한 부도예측 모형

다음 <그림 1>은 기업 부도 예측을 위한 일반 모형 및 SVM 전처리기를 사용한 예측 과정을 표현 한 것이다.

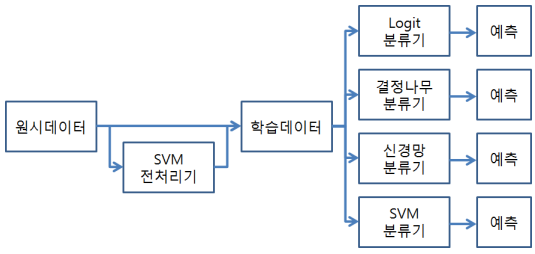

그림 1. 기업부도 예측모형
Fig. 1. Framework for Prediction of Business Bankruptcy

3.1 원시데이터 추출

일반적으로 기업부실이란 기업의 파산이나 해산 또는 만기가 도래한 채무에 대하여 지급능력을 상실한 지급불능상태 혹은 채무의 합계가 지급능력을 초과하는 채무초과상태 등을 의미한다[7]. 본 연구의 표본기업은 1999년부터 2005년까지 증권선물거래소(코스닥시장)의 상장폐지 기업 중에서 금융업, 건설업을 제외

한 12월 결산법인이며, 도산이전 5년간의 재무제표가 입수가능한 기업으로, 49개의 도산기업을 선정하였고, 이에 대응하는 정상기업은 도산기업과 동일업종, 비슷한 자산규모 등을 기준으로 선정한 [7]의 표본기업을 사용하였다.

[7]의 연구에서는 선행연구에서 제시한 유의미한 재무변수를 포함하여 1차적으로 52개 재무변수를 선정하였고, 표본기업 전체에 각각 항목이 존재하는 것 중에서 두 집단의 차이검정을 실시하여 두 집단을 구분하는 데 유의하다고 판정된 9개 재무변수를 <표 1>처럼 최종적으로 선정하였다. 본 연구는 선행연구 [7]의 데이터, 즉 부도기업 49개, 정상기업 49개의 9개 재무변수의 5년간 자료를 사용하였다.

표 1. 재무변수
Table 1. Financial variables

설명변수	재무비율
X3	유동비율
X7	총자산부채비율
X8	차입금의존도
X17	금융비용
X24	영업현금흐름 대 매출액
X25	영업현금흐름 대 자기자본
X26	영업현금흐름 대 부채
X32	순금융부담율
X33	투자수익율

3.2 학습데이터 구성

재무변수 9개의 5년간 자료를 사용하여 학습데이터를 구성하는 다양한 대안을 생각할 수 있다. [7]의 연구에서처럼, 1년전, 2년전과 같이 년간 데이터 각각을 이용하여 예측하거나 2년 데이터나 3년 데이터를 횡적으로 배열하여 구성하는 방법 등을 생각할 수 있다.

본 연구에서는 다음과 같은 형태로 학습데이터를 구성하였다.

1) 표본기업 1년전 재무변수 9개 자료를 설명변수로, 부도여부를 종속변수로 구성(ago1)

2) 표본기업 재무변수 9개의 1년전과 2년전의 2년간 평균을 설명변수로, 부도여부를 종속변수로 구성(avg2)

3) 표본기업 재무변수 9개의 1년전, 2년전, 3년전의 3년간 평균을 설명변수로, 부도여부를 종속변수로 구성(avg3)

4) 표본기업 재무변수 9개의 1~4년전의 4년간 평균을 설명변수로, 부도여부를 종속변수로 구성(avg4)

5) 표본기업 재무변수 9개의 1~5년전의 5년간 평균을 설명변수로, 부도여부를 종속변수로 구성(avg5)

6) 표본기업 재무변수 9개의 5년간 자료를 SVM 전처리 과정을 통해 변환하고 이를 설명변수로, 부도여부를 종속변수로 구성(svmp)

3.3 SVM 전처리 과정

SVM 전처리과정은 다음 순서로 진행한다.

먼저 재무변수 각각의 5년간 자료가 설명변수로, 부도여부를 종속변수로 하여 학습데이터를 구성한다. 예를 들어, X3(유동비율)의 5년간 자료가 설명변수가 되고, 부도여부가 종속변수가 된다.

다음에 SVM 학습을 통해 의사결정 경계식을 구한다. X3의 경우 식 (1)과 같은 경계식을 구하였다. 식 (1)에서 X3a는 1년전 데이터를, X3e는 5년전 데이터를 나타낸다.

$$0.0085*X3a - 0.0013*X3b - 0.0016*X3c - 0.0025*X3d + 0.003*X3e - 1.1361 = 0 \quad (1)$$

식 (1)에 X3의 5년간 자료를 대입하여 나온 결과값이 m이라면, 식 (2)와 같은 sign변환을 하면 변환된 X3의 값을 구하게 된다.

new X3=sign(m) (2)

모든 재무변수에 대해 위와 같은 SVM 전처리과정을 통해 새로운 재무변수의 값을 얻게 된다.

3.4 분류모형 학습

구성된 학습데이터를 이용해서 로짓모형, 신경망모형, 의사결정나무모형, SVM모형에 대해 학습을 한다. 본 연구에서는 공개 데이터마이닝 도구인 WEKA[15]를 이용하여 학습하였다. 로짓모형은 WEKA의 Logistic와 SimpleLogistics를, 신경망모형은 WEKA의 MultilayerPerceptron을, 의사결정나무모형으로 WEKA의 J48을, SVM모형으로 WEKA의 SMO를 사용하였다.

MultiLayerPerceptron의 파라미터 설정은 사전 실험결과 학습데이터의 예측성능이 비교적 높게 나오는 은닉노드 5, 학습률 0.2, 모멘텀 0.1, 학습회수 5000을 설정하였다. SMO에서는 여러 커널함수를 적용한 사전 실험결과 성능이 우수하게 나오는 Polynomial 커널함수를 사용하였다.

3.5 예측

연구의 학습자료는 부도 및 정상 기업 49개씩 모두 98개의 학습사례로 구성되어 있다. 예측 성능은 학습자료가 비교적 적을 때 효과적인 방법 중 하나인 10-겹 상호검증 방법을 사용하여 측정하였다.

Ⅳ. 실험 및 결과 분석

1999년부터 2005년까지의 코스닥의 부도기업 49개, 정상기업 49개의 9개 재무변수의 5년간 자료에 대해 부도기업 예측모형에 대해 실험하였다. 먼저 SVM 전처리과정을 적용하지 않은 3.2절의 1)~5)의 학습데이터에 대한 실험 결과 <표 2>와 같은 예측결과를 얻었다.

표 2. 분류모형 예측결과(전처리 이전)
Table 2. Prediction results of classifiers (before preprocessing)

학습 데이터	분류 모형				
	Logistic	Simple Logistic	J48	NN	SVM
ago1	83.67	85.71	92.86*	83.67	86.73
avg2	92.86	94.9*	93.88	87.76	92.86
avg3	93.88*	93.88*	92.86	86.73	93.88*
avg4	92.86	92.86	95.92	89.8	95.92*
avg5	91.84	94.9	93.88	95.92	96.94*

학습데이터에 따라 우수한 예측성능을 보이는 모형은 조금씩 다르게 나타났다. 최근 1년전 자료를 이용한 경우에는 의사결정모형(J48)이, 최근 2년간 평균 자료를 이용한 경우에는 로짓모형(SimpleLogistic)이, 최근 3년간 평균 자료를 이용한 경우에는 로짓모형(Logistic, SimpleLogistic)과 SVM모형이, 최근 4년간 평균자료와 5년간 평균자료를 이용한 경우에는 SVM모형의 예측성능이 가장 좋은 것으로 나타났다. 전체적으로 보면 최근 5년간의 평균자료(avg5)를 이용하여 SVM모형을 가지고 학습하고 예측한 경우에 예측성능이 96.94로 가장 높게 나타났다.

특이한 점 중의 하나는 다른 분류모형과는 달리 SVM모형은 평균자료에 사용하는 년도 자료가 많아질수록 예측성능이 좋아진다는 점이다. 다른 분류모형을 사용할 경우에는 가장 최적의 년도 자료의 구성을 찾아야 하지만 SVM모형은 5년 자료 전체를 사용하면 되기 때문에 많은 노력을 경감시킬 수 있음을 의미한다.

3.3절의 SVM전처리 과정을 적용한 3.2절 6)의 학습데이터(svmp)의 실험결과는 <표 3>과 같다. SVM전처리 과정을 거친 이후의 예측성능은 98.98로, 전처리 이전의 가장 높았던 예측성능 96.94보다 높음을 알 수 있다. 더욱 특징적인 점은 SimpleLogistic을 제외하고

모든 분류모형의 예측성능이 98.98로 높아졌다는 점이다.

표 3. 분류모형 예측결과(전처리 이후)
Table 3. Prediction results of classifiers (after preprocessing)

학습 데이터	분류 모형				
	Logistic	Simple Logistic	J48	NN	SVM
svmp	98.98*	96.94	98.98*	98.98*	98.98*

SVM의 전처리효과는 다음과 같이 추정해 볼 수 있다. 각 재무변수 5년간의 데이터가 있는 경우 5년간 데이터의 대표성을 어떻게 표현해서 학습 데이터를 구성할 것인가의 의사결정 문제가 있다. 본 연구에서 실험해본 것처럼 5년간 데이터의 평균을 구하여 학습데이터를 구성할 수도 있다. 가능한 또 다른 아이디어로는 판별 모형을 이용해서 나온 결과를 활용하는 것이다. 이 연구에서는 SVM을 활용하여 각 재무변수의 새로운 대표성을 구하고 이를 이용하여 각 판별모형을 이용한 부도 예측 실험을 실시하였다. 그 결과 여러 분류모형의 예측률을 높일 수 있음을 볼 수 있었다.

이는 SVM이 여러 분류모형의 전처리 모형으로서 효과적으로 사용될 수 있음을 시사한다.

Ⅴ. 결론

기업의 부도로 인한 사회 경제적 손실은 매우 크다. 사전에 부도에 대한 징후를 예견할 수 있다면 이에 대한 대비 및 원인 제거를 통해 손실을 최소화하는 것이 가능할 것이다. 기업의 부도예측을 위한 다양한 분류모형이 제시되어 왔으며, 분류모형의 주요 이슈중 하나는 기업 부도 예측성능을 높이는 것이다. 본 연구에서는 코스닥기업의 부도 예측을 위한 분류모형 및 예측모형에 대해 살펴보고, 선정된 재무변수의 5년간 자료를 가지고 예측성능을 비교해 보았다.

일반적인 기업부도 예측성능 비교에서 5년간 평균자료를 이용하고 SVM모형을 사용할 경우 예측성능이 높게 나타남을 알 수 있었다. SVM모형을 학습데이터 전처리 모형으로 사용할 경우 예측성능을 더욱 높일 수 있음을 보았다. 특히 본 연구에서 비교대상인 로짓모형, 의사결정나무모형, 신경망모형, SVM모형 등 모든 분류모형의 예측성능을 높임을 알 수 있었다.

SVM 모형은 본래의 좌표공간 **x**에 있는 데이터를 선형 의사결정 경계를 사용할 수 있는 새로운 좌표공간 $\Phi(\mathbf{x})$로 변환시킴으로서 초평면을 찾게 되는 데, 고차원의 데이터에도 잘 작동되는 특징이 있다 [14]. SVM 모형이 기업 부도 예측에 영향을 주는 재무변수의 본래 차원을 새로운 차원으로 변환함으로서 예측률을 높이는 것으로 판단된다.

본 연구에서 사용한 자료는 부도 및 정상기업 각각 49개씩 98개 사례를 이용하였다. 연구 결과의 일반화를 위해서는 더욱 많은 사례와 응용 분야에 대한 추가적인 실험을 통해 확인하는 과정이 필요하다.

참고문헌

[1] W.H. Beaver, "Financial Ratios as Predictors of Failure," Journal of Accounting Research, Vol. 4, pp. 71-111, 1966.

[2] 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.

[3] H. W. Jang, "비금융 상장기업의 부실예측모형," The Korean Journal of Financial Management, Vol. 5, No. 1, pp. 299-327, 1998.

[4] J. M. Kang and S. H. Hong, "부실예측모형의 적합성 분석," The Journal of Finance and Banking, Vol. 5, No. 1, pp. 83-110, 1999.

[5] K. C. Lee, M. J. Kim and H. Kim, "An Inductive Learning-Assisted Neural Network Approach to Bankruptcy Prediction: Comparison with MDA, Inductive Learning, and Neural Network Models,"

Korean Management Review, Vol. 23, No. 3, pp. 109-144, 1994.

[6] J. H. Min and Y. C. Lee, "Support Vector Bankruptcy Prediction Model with Optimal Choice of RBF Kernel Parameter Values using Grid Search," Journal of the Korean Operations Research and Management Science Society, Vol. 30, No. 1, pp. 55-74, 2005.

[7] J. H. Cho and B. S. Kang, "A Study on the Prediction of KOSDAQ Business Bankruptcy," Review of Business and Economics, Vol. 20, No. 1, pp. 141-160, 2007.

[8] T. R. Lee, J. Y. Koo, H. J. Park, K. H. Lee, and D. W. Choi, Data Mining, KNOU Press, 2004.

[9] J.T. Friedman, T. Hasti, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting," The Annals of Statistics, Vol.32, No. 2, pp. 337-374, 2000.

[10] N. Langwehr, M. Hall, and E. Frank, "Logistic Model Trees," Proceedings of the 16th Conference on Machine Learning, pp. 241-252, 2003.

[11] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, 1993.

[12] A.P. Engelbrecht, "A New Pruning Heuristic based on Variance Analysis of Sensitivity Information," IEEE Transactions on Neural Networks, Vol. 12, No. 6, pp. 1386-1399, 2001.

[13] P.H. Winston, Artificial Intelligence 3rd Ed. NY, Addison-Wesley, 1993.

[14] H. S. Yong, Y. M. Na, J. S. Park, H. U. Seung, M. S. Lee, S. J. Lee and R. Choi, Introduction to Data Mining, Infinitybooks, 2007.

[15] I.H. Witten, E. Frank, and M.A. Hall, Data Mining: Practical Machine Learning Tools and Techniques 3rd Ed., Morgan Kaufmann, 2011.

저자소개

강부식(Boo-Sik Kang)

1985년 경희대학교 산업공학과
1989년 KAIST 산업공학과(공학석사)
2000년 KAIST 산업공학과(공학박사)

1989년~2001년 KT 품질보증단, 연수원, 통신망연구소
2001년~현재 목원대학교 서비스경영학부 교수
※ 관심분야 : 지능정보시스템, 데이터마이닝, 고객관계관리, 서비스품질경영

조준희(Jun-Hee Cho)

1987년 목원대학교 경영학과
2002년 홍익대학교 대학원 경영학과 (경영학박사)

2004년~현재 목원대학교 경영학과 교수
※ 관심분야 : 기업창업론, 경영분석