경영정보학연구
제15권 제2호
2005년 6월

Support Vector Machine을 이용한 기업부도예측*

박 정 민**, 김 경 재***, 한 인 구****

Bankruptcy Prediction using Support Vector Machines

Jung-min Park, Kyoung-jae Kim, Ingoo Han

There has been substantial research into the bankruptcy prediction. Many researchers used the statistical method in the problem until the early 1980s. Since the late 1980s, Artificial Intelligence (AI) has been employed in bankruptcy prediction. And many studies have shown that artificial neural network (ANN) achieved better performance than traditional statistical methods. However, despite ANN's superior performance, it has some problems such as overfitting and poor explanatory power.

To overcome these limitations, this paper suggests a relatively new machine learning technique, support vector machine (SVM), to bankruptcy prediction. SVM is simple enough to be analyzed mathematically, and leads to high performances in practical applications. The objective of this paper is to examine the feasibility of SVM in bankruptcy prediction by comparing it with ANN, logistic regression, and multivariate discriminant analysis. The experimental results show that SVM provides a promising alternative to bankruptcy prediction.

Keywords : Support Vector Machine, Bankruptcy Prediction, Artificial Neural Network, Logistic Regression, Multivariate Discriminant Analysis

* 이 논문은 2004년도 한국학술진흥재단 지원에 의하여 연구되었음(KRF-2004-003-B00069).
** 하나은행 리스크관리본부
*** 동국대학교 경영대학 정보관리학과(교신저자)
**** 한국과학기술원 테크노경영대학원

Ⅰ. 서 론

기업의 부도는 주주나 채권자는 물론 종업원, 고객, 소비자 모두에게 경제적 손실을 초래하고, 사회적 부를 감소시킨다. 따라서 기업의 부도가능성을 예측하는 활동은 이해관계자들에게 예측가능한 손실을 최소화할 수 있는 정보를 제공한다는 점에서 의의가 있다.

많은 연구자들에 의해 부도예측을 위한 연구는 꾸준히 이어지고 있다. 초기에는 통계적 기법을 이용하여 부도예측모형을 개발하였으나, 1980년대 이후부터는 인공지능기법을 이용한 연구가 활발하게 진행되었다. 특히, 인공신경망을 이용한 연구는 예측력이 우수하여 가장 많이 사용되고 있다. 그러나, 인공신경망을 사용할 경우 입력자료의 분포를 추정하기 위해 다량의 학습데이터가 필요하고, 과도적합문제(overfitting)로 인해 일반화의 어려움이 있을 뿐 아니라, 지역적 최소값(local minima)을 피하기 위한 초기화 작업이 경험에 의존하고, 기본적으로 '암상자 모형'이라서 각 변수의 가중치 등 모형을 해석하기 어렵다는 점 등이 한계로 지적되어 왔다.

본 연구에서는 이에 대한 해결방안으로 최근 각광 받고 있는 support vector machine(SVM)을 부도예측에 적용하고자 한다. SVM은 Vapnik [1995]에 의해 제안된 학습이론으로 분류문제를 해결하기 위한 최적의 분리 경계면(hyperplane)이라는 개념을 사용한다. SVM이 주목 받는 이유는 첫째, 명백한 이론적 근거에 기반하므로 결과 해석이 용이하고, 둘째, 실제 응용에 있어서 인공신경망 수준의 높은 성과를 내고, 셋째, 적은 학습자료만으로 신속하게 분류학습을 수행할 수 있기 때문이다. 또한 SVM은 기존의 학습 알고리즘이 경험적 위험 최소화 원칙(empirical risk minimization)을 구현하는 것인데 비해 구조적 위험 최소화 원칙(structural risk minimization)에 기반하므로 과대적합문제를 어느 정도 피할 수 있다.

본 연구는 총 5장으로 구성하였다. Ⅰ장에서는 연구의 의의와 목적을 정의하고, Ⅱ장에서는 SVM에 관해 알아보도록 한다. 이어서 Ⅲ장에서는 부도예측을 위한 SVM모형을 제안하고 MDA, 로지스틱 회귀분석, 인공신경망 등의 방법론들과 비교분석을 한다. 그리고, Ⅳ장에서는 Ⅲ장의 연구 결과를 분석하고, Ⅴ장에서는 본 연구의 시사점과 결론을 제시한다.

Ⅱ. Support Vector Machine (SVM)

SVM은 Vapnik[1995]에 의해 개발된 분류기법으로, 입력공간과 관련된 비선형문제를 고차원의 특징공간에서의 선형문제로 대응시켜 나타내기 때문에 수학적으로 분석하는 것이 수월하다[Hearst et al., 1998]. 또한, SVM은 조정해야 할 파라미터의 수가 많지 않아 비교적 간단하게 학습에 영향을 미치는 요인들을 규명할 수 있다. 그리고 구조적위험을 최소화함으로써 과대적합문제에서 벗어날 수 있으며, 볼록함수를 최소화하는 학습을 진행하기 때문에 전역최적해(globally optimal solution)를 구할 수 있다는 점에서 인공신경망보다 우월한 기계학습기법으로 주목 받고 있다.

최근 몇 년간 SVM을 사용한 다양한 연구가 진행되었다. SVM은 문서분류, 영상인식, 문자인식 등에서 뛰어난 일반화 성능을 보여주었다 [Joachims, 1998; Osuna et al., 1997]. 또한, SVM을 재무분야에 적용한 연구도 있는데, 주로 시계열 예측 및 분류에 관한 것이다[Tay and Cao, 2002; Kim, 2003]. 본 연구와 가장 유사한 연구로는 SVM을 사용하여 채권신용등급을 예측한 연구를 들 수 있다[Huang et al., 2004]. 이 연구들에서 SVM은 일반화에 있어서 인공신경망이나 판별분석 등의 다른 분류기법들과 비교하여 비슷하거나 더 우수한 성능을 나타낸 것으로 보고되었다. 본 연구에서는 이러한 연구 배경을 토대

로 채권신용등급문제와는 다른 재무적 특성을 지닌 부도예측에 SVM을 적용하여 보기로 한다. 이를 위해 SVM에 대하여 간단히 설명하고자 한다.

SVM에서는 모형구축용 데이터들을 서로 다른 두 개의 클래스로 분류할 때 분류의 기준이 되는 분리 경계면(hyperplane)을 학습 알고리즘을 이용하여 찾는다[이수용, 이일병, 2002]. 따라서, SVM은 입력벡터 x를 고차원의 특징공간(high-dimensional feature space)으로 사상(mapping)시킨 후 두 클래스 사이의 마진(margin)을 최대화시키는 분리 경계면를 찾는 것을 목적으로 한다. 이러한 최대마진 분리 경계면(maximum margin hyperplane)은 두 클래스 사이의 거리를 최대로 분리시킨다. 이 때 최대마진 분리 경계면에 가장 근접한 모형구축용 데이터를 서포트 벡터(support vector)라고 부른다. 선형분리문제에서, 독립변수가 3개인 경우 분리 경계면은 식 (1)과 같다.

$$y = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 \qquad (1)$$

여기서 y는 출력값이고, x_i는 변수값, 그리고 4개의 w_i는 학습 알고리즘에 의해 학습된 가중치이다. 상기 식에서 가중치 w_i는 분리 경계면을 결정하는 파라미터이다. 이 때 최대마진 분리 경계면은 서포트 벡터를 사용해서 식 (2)와 같이 나타낼 수 있다.

$$y = b + \sum \alpha_i y_i x(i) \cdot x \qquad (2)$$

여기서, y는 모형구축용 데이터 $x(i)$의 분류값이고, $\cdot$는 내적(dot product)이다. 벡터 x는 모형검증용 데이터를 나타내고, 벡터 $x(i)$는 서포트 벡터(최대마진 분리 경계면에 가장 근접한 모형구축용 데이터)를 나타낸다. 이 식에서, b와 α_i는 분리 경계면을 결정하는 파라미터이다. 서포트 벡터를 찾아내고, 파라미터 b와 α_i를 결정하는 것은 선형적으로 제약된 이차계획문제(linearly constrained quadratic programming)를 푸는 것과 같다.

앞에서 언급한 바와 같이, SVM은 입력변수를 고차원의 특징 공간으로 이동시킴으로써 비선형 분류문제를 선형모형으로 근사시킨다. 비선형 분류문제에서 사용될 식 (2)의 고차원 버전은 식 (3)과 같이 간단하게 나타낼 수 있다.

$$y = b + \sum \alpha_i y_i K(x(i), x) \qquad (3)$$

상기 식에서 함수 $K(x(i), x)$는 커널함수라고 정의된다. 커널함수는 원래 데이터를 고차원 공간으로 사상시킴으로써 특징공간 내에 선형으로 분리가능한 입력 데이터셋을 만든다. 이 때 사용될 수 있는 커널함수는 여러 가지가 있으며 어떤 커널함수를 선택하는 것이 바람직한가는 문제에 따라 상이하고, 이는 SVM을 적용하는데 있어서 가장 중요한 요소 중의 하나이다. 일반적으로 많이 사용되는 커널함수로는 다항식 커널(polynomial kernel)과 가우시안 RBF(Gaussian radial basis function)를 들 수 있다:

가우시안 RBF :

$$K(x, y) = \exp\left(-\frac{1}{\delta^2}(x-y)^2\right) \qquad (4)$$

다항식 커널: $K(x, y) = (xy+1)^d$ (5)

여기서 d는 다항식 커널의 차수이고, δ^2은 가우시안 RBF 커널의 대역폭이다.

분리가능한 문제에 있어서 상기 식의 계수 α_i의 하한은 0이다. 분리가 불가능한 문제에서 SVM은 계수 α_i의 하한 이외에 상한 C를 추가함으로써 일반화된 결과를 얻을 수 있다[Kim, 2003].

Ⅲ. 실증연구

3.1 자료수집 및 변수선정

부도예측모형을 구축하기 위하여 1335개의

건전기업과 1335개의 부도기업 등 총 2670개 기업의 재무데이터를 수집하였다. 건전기업의 데이터는 자산규모 10억 이상 70억 이하에 속하는 국내 비외감 중공업 기업의 1999년과 2000년의 재무자료를 기준으로 하였고 이에 대응하는 부도기업의 데이터는 역시 자산규모 10억 이상 70억 이하의 국내 비외감 중공업 기업의 데이터를 기준으로 하였다. 부도기업의 경우에는 일반적으로 건전기업보다 매년 발생하는 데이터의 건수가 적으므로 1996년부터 2000년까지의 부도기업 데이터를 사용하였다.

총 2670개의 재무데이터 중에서 모형구축용 데이터로는 부도기업과 건전기업을 50:50의 비율로 선정하여 총 데이터의 80%를 사용하였고, 나머지 20%는 모형검증용 표본으로 사용하였다. 인공신경망의 경우에는 모형구축용 데이터로 총 데이터의 60%를 사용하고, 모형시험용 데이터로 20%, 그리고 나머지 20%를 모형검증용 데이터로 사용하였다. 보다 일반화된 연구결과를 얻기 위하여 본 연구에서는 상호검증방법(cross-validation method)을 사용하였다[Weiss and Kulikowski, 1991]. 따라서 본 연구에서는 전체 데이터를 5개의 데이터셋으로 나누어 4개의 데이터셋으로 모델을 구성하고, 나머지 하나의 데이터셋으로 모형을 검증하는 과정을 5번 반복하였다. 이에 따라 모형구축용 데이터로는 총 10,680개, 모형검증용 데이터로는 총 2,670개의 데이터를 사용한 효과를 얻을 수 있었다.

본 연구에서는 재무모형을 구축하기 위해 모형에 사용할 데이터로는 회계데이터로부터 파생된 재무비율을 사용하였다. 모형에 사용될 변수선정과 관련하여 먼저 총 164개의 재무비율을 전처리하였다. 전처리과정은 먼저 이상치 제거를 위하여 각 재무비율별 분포의 양측 1%의 데이터를 제거하고, 결측치는 각 비율의 평균값으로 대체하였다. 그 후 단일변량분석의 과정인 *t-test*를 통하여 111개의 건전 또는 부실기업의 분류에 유의한 재무비율들을 가려내었다($P < 0.01$). 단일변량분석을 통해 선택된 변수들에 대해 다시 다변량분석의 과정으로 로지스틱 회귀분석의 단계별 변수선정방법(stepwise method with forward selection)을 통하여 최종적으로 15개의 변수를 선정하였다($P < 0.05$). 선정된 변수의 목록과 통계치는 <표 1>과 같다.

<표 1> 선정된 변수와 통계치

변 수 명	범 위	평 균	표준편차	Wald 값	P 값
금융비용대부채비율	17.021	7.082	3.551	37.373	0.000
매출원가비율	51.000	81.879	8.167	4.876	0.027
자기자본비율	107.192	24.183	16.821	58.059	0.000
금융비용부담율증가분	0.259	-0.007	0.035	4.289	0.038
매입채무회전율	191.440	13.373	21.388	13.219	0.000
지급여력도	294.118	38.553	36.620	4.087	0.043
분식계수	12.454	0.964	1.547	5.666	0.017
기업경상이익율	156.723	17.054	23.104	178.444	0.000
현금흐름대전기총부채	3.021	0.094	0.323	16.127	0.000
총자산변동계수	111.592	20.390	20.739	54.823	0.000
자산대비금융비용증가율	16.961	0.262	2.845	12.665	0.000
자산대비영업외비용증가율	24.991	-0.080	3.903	28.431	0.000
매출원가곱매출원가평균증가비	1172.787	142.358	152.350	30.616	0.000
매출대비재고자산증가율	51.987	1.597	7.146	9.660	0.002
총자본회전율곱매출액증가비	12.352	1.991	1.837	18.751	0.000

<표 2> 선정된 변수내역

변 수 명	변 수 내 역
금융비용대부채비율	당기 이자비용과 사채이자의 합이 당기 부채총계와 전기 부채총계의 합에서 차지하는 비
매출원가비율	당기 매출원가와 당기 매출액의 백분율
자기자본비율	당기 자본총계와 당기 자산총계의 백분율
금융비용부담율증가분	전기 대비 당기 금융비용부담율의 증가액
매입채무회전율	당기 매출액을 전기와 당기의 매입채무액으로 나눈 값
지급여력도	지급여력액과 당기매출액의 백분율
분식계수	분식금액추정액을 지급여력액으로 나눈 값
기업경상이익율	당기 경상이익, 이자비용, 사채이자의 합과 전기와 단기의 자산총계의 평균값의 백분율
현금흐름대전기총부채	영업활동후현금흐름을 전기 부채총계로 나눈 값
총자산변동계수	3년 간의 총자산변동율
자산대비금융비용증가율	당기 금융비용과 전기 금융비용의 차액과 전기와 당기 평균자산총계의 백분율
자산대비영업외비용증가율	당기 영업외비용과 전기 영업외비용의 차액과 전기와 당기의 평균자산총계의 백분율
매출원가곱매출원가평균증가비	당기 매출원가와 전기 매출원가의 비율을 매출원가비율과 곱한 값
매출대비재고자산증가율	당기 재고자산합계와 전기 재고자산합계의 차액과 당기 매출액의 백분율
총자본회전율곱매출액증가비	총자본회전율과 전기 대비 당기 매출액 비율을 곱한 값

<표 2>에서는 선정된 각 변수의 설명을 정리하였다.

3.2 SVM 모형구축

전술한 바와 같이 SVM 모형의 구축에 있어서 어떤 커널함수를 사용하느냐는 가장 중요한 문제 중의 하나이다. 본 연구에서는 SVM의 커널함수로서 가장 널리 사용되는 다항식 커널과 가우시안 RBF를 사용하였다. Tay and Cao[2001]에 의하면 SVM의 성능에 있어서 커널함수의 상한 C와 커널 파라미터 δ^2, d가 중요한 역할을 한다고 보고되었다. 따라서 적절한 상한 C와 커널 파라미터 δ^2, d를 선정하기 위해, 선행연구에서 SVM의 파라미터에 대해 제시된 일반적인 가이드를 따라 본 연구에서도 일정 범위 내에서 다양한 값을 대입하여 다양한 모형을 생성시켰다. SVM 모형의 구축은 LIBSVM[Chang and Lin, 2001]을 사용하였다.

3.3 인공신경망 모형구축

본 연구에서 SVM에 대한 비교대상으로서 인공신경망, 로지스틱 회귀분석, MDA를 수행하였다. 각 기법들에 대한 설계는 일반적으로 알려진 범위 내에서 가장 우수한 성능을 나타내는 방향으로 진행하였다.

전술한 바와 같이 인공신경망은 부도예측 분야에서 가장 많이 사용되어 온 방법론이므로 본 연구에서는 인공신경망의 일반적인 작동원리에 관해서는 그 설명을 생략하기로 한다. 한편, 인공신경망 모형의 설계에 관해서는 변수선정과정, 신경망 구조 설계 등에 있어서 아직까지 일반적인 원칙이 없다. 따라서 본 연구에서도 기존

연구에서 사용된 방법과 같이 다양한 실험조건을 가장 좋은 신경망 구조를 선택하였다.

일반적으로 인공신경망의 성능에 영향을 미치는 요인으로서 은닉층의 수, 은닉층의 노드 수, 학습횟수 등이 알려져 있다. Hornik[1991]에 따르면 은닉층의 수는 하나만으로도 분류문제를 포함한 대부분의 문제에서 만족할만한 결과를 얻을 수 있다. 따라서, 본 연구에서도 은닉층이 하나인 3층 퍼셉트론을 사용하였다. 은닉층의 노드 수는 경험적으로 입력노드 수와 출력노드 수의 합을 n이라 할 때 n/2, n, 2n을 많이 사용하지만, 모든 경우에 적합하다고 할 수는 없다. 은닉노드의 수는 인공신경망 구조를 설계하는데 있어서 중요한 요소이며 그 결정은 데이터 의존적인 경우가 많다. 모형구축용 데이터를 분류하는 경우에는 은닉노드의 수가 많을수록 바람직하지만, 모형검증용 데이터에서는 은닉노드의 수가 많은 것이 반드시 바람직한 것은 아니다[Patuwo et al., 1993]. 따라서 은닉노드의 수를 많게 하거나 적게 하는 두 가지 방법 모두 득실이 있으므로 본 연구에서는 은닉노드의 수를 8, 12, 16, 24, 32로 구분하여 실험해보았다.

학습횟수는 너무 적으면 학습이 제대로 이루어지지 않고, 너무 많아도 모형구축용 데이터에 과대적합되어 모형검증용 데이터의 예측력이 떨어지는 경우가 많다. 본 연구에서는 학습이 적당히 이루어지도록 모형시험용 데이터의 평균오차가 최소값을 기록한 후 50,000회가 지나면 학습이 멈추도록 하였다. 편의를 위해 인공신경망의 학습률(learning rate)은 0.1, 모멘텀은 0.1로 고정하였다.

3.4 로지스틱 회귀분석 모형구축

로지스틱 회귀분석은 독립변수가 등간척도나 비율척도이고, 종속변수가 명목척도인 경우에 사용하는 계량분석방법이다. 로지스틱 회귀분석을 부도예측에 사용할 경우 기업의 설명변수의 관찰치벡터를 X_i로 하고, 그 계수 β_i를 추정한다면 기업의 부실확률은 로지스틱 함수에 의해 식 (6)과 같이 유도된다.

$$Y_i = \frac{1}{1+e^{-P}} \quad (6)$$

여기서, $P = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_i X_i$ 이다. 본 연구에서는 로지스틱 회귀분석 모형구축을 위해 SPSS 소프트웨어를 사용하였고, 판별점은 0.5를 기준으로 하였다.

3.5 다변량판별분석(MDA) 모형구축

다변량판별분석은 로지스틱 회귀분석과 마찬가지로 등간척도나 비율척도로 측정된 독립변수와 명목척도인 종속변수를 이용한 분석기법으로 선험적으로 정의된 두 개 이상의 집단들을 가장 잘 판별할 수 있는 둘 이상의 독립변수의 선형조합을 찾아내는 과정을 포함한다. 다변량판별함수는 식 (7)과 같은 형태이다.

$$Z = W_1 X_1 + W_2 X_2 + W_3 X_3 + \cdots + W_n X_n \quad (7)$$

여기서 Z는 판별점수이고, W는 판별계수이고, X는 독립변수를 말한다. 다변량판별분석을 위해 본 연구에서는 SPSS 소프트웨어를 사용하여 수행하였다.

Ⅳ. 연구결과

본 연구에서는 SVM의 실험결과를 각 커널함수와 파라미터에 따라 정리해보고, 추가적으로 인공신경망, 로지스틱 회귀분석, MDA의 실험결과와 비교해보고자 하였다.

선형 SVM의 장점 중 하나는 조정해야 할 파라미터가 상수 C 이외에는 존재하지 않는다는 점이다. 그러나 선형 SVM으로 분리되지 않는 모형구축용 데이터인 경우에는 계수 α_i의 상한

인 C가 예측력에 영향을 미친다. 비선형 SVM인 경우에는 커널 파라미터도 조정해야 한다. 전술한 바와 같이 본 연구에서 사용한 커널함수는 가우시안 RBF와 다항식 커널이다. 따라서 본 연구에서는 상한 C와 커널 파라미터를 변경하면서 실험을 진행하였다.

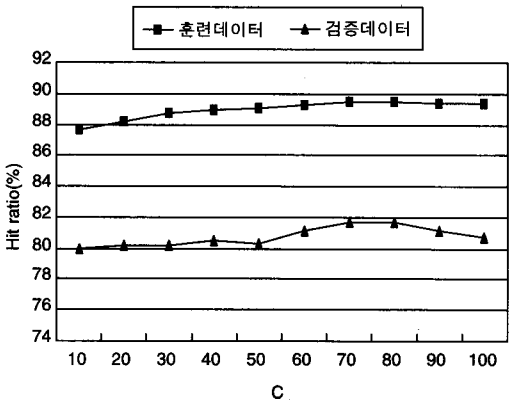
가우시안 RBF에서는 C 이외에 커널 파라미터로 δ^2을 고려해야 한다. Tay and Cao[2001]에 따르면, 적절한 δ^2의 범위는 1에서 100사이이고, C의 값으로 적합한 범위는 10에서 100사이라고 한다. 이를 참고하여 본 연구에서도 C와 파라미터의 값을 적합한 범위 내에서 세분화하여 실험하였고, 실험의 결과가 의미 있는 것을 위주로 정리하였다. ε은 0.001로 고정하였다. <표 3>은 각 파라미터에 대한 SVM의 예측력을 나타낸 것이다. <표 3>에 나타난 바와 같이 δ^2이 5이고, C가 80인 경우 가장 우수한 예측정확성을 나타내었다.

<표 3> 가우시안 RBF 사용시 모형검증용 데이터의 SVM 결과

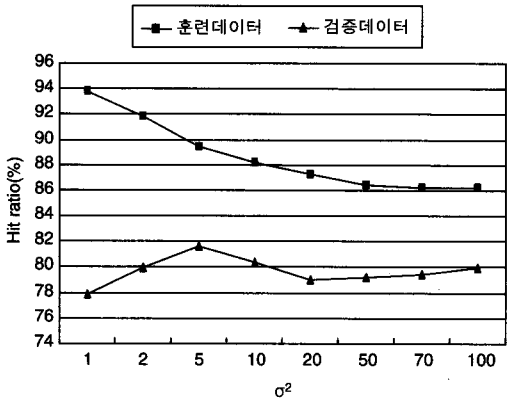
δ^2	C	Set 1	Set 2	Set 3	Set 4	Set 5	평균
1	20	85.02	87.08	90.64	84.64	79.59	85.39
	40	85.58	87.45	91.01	84.46	79.03	85.51
	60	85.02	86.14	89.89	83.71	77.90	84.53
	80	85.77	86.33	88.95	84.08	77.90	84.61
	100	85.39	85.96	88.76	83.90	77.90	84.38
2	20	86.70	86.14	92.13	84.64	80.34	85.99
	40	86.33	86.14	91.39	83.90	80.15	85.58
	60	85.96	86.33	91.57	84.64	79.78	85.66
	80	85.77	87.08	90.82	84.27	79.96	85.58
	100	85.39	87.08	91.01	84.27	79.78	85.51
5	20	84.27	86.14	91.95	83.15	80.15	85.13
	40	85.96	86.52	92.32	83.33	80.52	85.73
	60	86.14	85.96	92.51	84.27	81.09	85.99
	80	86.33	86.52	92.32	84.08	81.65	86.18
	100	86.14	86.70	92.13	83.71	80.71	85.88
10	20	83.71	85.02	91.39	82.21	79.03	84.27
	40	84.27	84.83	91.76	82.77	79.96	84.72
	60	84.27	85.77	91.76	83.15	79.78	84.94
	80	84.64	85.77	91.95	83.15	80.34	85.17
	100	85.21	86.14	91.76	83.52	79.78	85.28
20	20	83.71	84.64	90.64	82.58	79.40	84.19
	40	84.08	85.02	90.45	82.21	79.40	84.23
	60	84.08	84.83	91.20	82.21	79.21	84.31
	80	83.90	85.39	91.95	82.58	79.03	84.57
	100	84.08	85.39	91.95	82.58	79.03	84.61

<그림 1>과 <그림 2>는 δ^2과 C를 각각 고정하였을 때 C와 δ^2의 변화에 따른 모형구축용 데이터와 모형검증용 데이터의 성과의 추이를 나타낸 것이다. <그림 1>에서 알 수 있듯이 δ^2이 일정할 때 C가 증가할수록 모형구축용 데이터는 과대적합되는 경향을 보였다. 즉, δ^2이 일정할 때 C가 증가할수록 모형구축용 데이터의 학습성과는 매우 점점 높아지지만 모형검증용 데이터에서의 성과는 일정부분 증가하다가 다시 감소하는 추세를 보였다. <그림 2>에서는 C가 일정할 때 δ^2이 증가하면 모형구축용 데이터가 과소적합되는 양상을 나타내었다. 즉, C가 일정할 때 δ^2이 증가할수록 모형검증용 데이터의 성과는 일정한 추세를 보이지 않으나 모형구축용 데이터에서는 성과가 지속적으로 하락하는 경향을 나타냈다. 이는 Tay and Cao[2001], Kim[2003]의 연구 결과와 유사한 것이다.

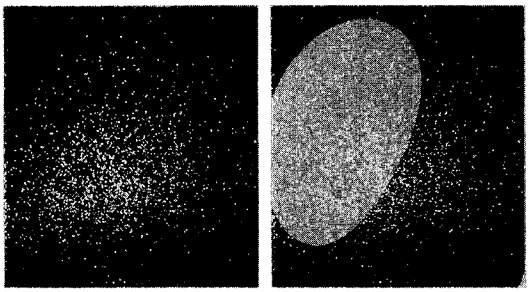
<그림 3>(a)는 SVM 사용 전후의 모형구축용 데이터의 패턴의 예를 나타내고 (b)는 SVM 사용 전후의 모형검증용 데이터의 패턴의 예를 나타낸다. 그림에서 각 점은 각 기업의 기하학적 위치를 의미하며, 부도기업과 건전기업의 두 가지 클래스를 상이한 두 가지의 색상으로 구분하였다. SVM 실행 전에는 분리 경계면이 표시되어 있지 않지만 SVM 실행 후에는 상이한 두 가지 색상의 곡면을 통해 분리 경계면을 표현하고 있으며 이 경계를 통해 예측결과가 부도기업과 건전기업의 두 가지 클래스로 분리된다.



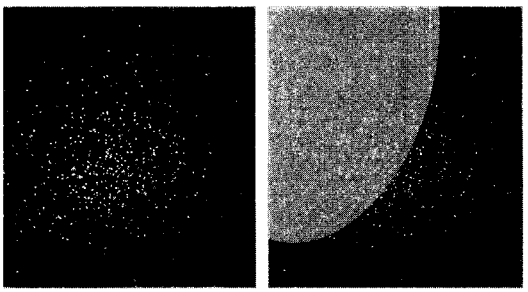
<그림 1> δ^2이 5일 때 C의 변화에 따른 SVM 결과



<그림 2> C가 80일 때 δ^2의 변화에 따른 SVM 결과



(a) SVM 실행 전과 후의 모형훈련용 데이터 패턴



(b) SVM 실행 전과 후의 모형검증용 데이터 패턴

<그림 3> SVM 실행 전과 후의 데이터 패턴의 예

<표 4> 다항식 커널 사용시 모형검증용 데이터의 SVM 결과

d	C	Set 1	Set 2	Set 3	Set 4	Set 5	평균
1	20	83.52	83.90	89.89	81.84	79.40	83.71
	40	84.08	84.64	89.89	82.02	79.21	83.97
	60	83.71	84.64	90.07	82.21	79.03	83.93
	80	83.52	84.83	90.26	82.40	78.84	83.97
	100	83.33	84.83	89.89	82.58	78.84	83.90
2	20	83.90	83.90	90.82	83.33	79.21	84.23
	40	84.46	84.27	90.82	82.40	78.65	84.12
	60	84.08	83.90	91.01	81.84	78.84	83.93
	80	84.27	84.64	91.01	82.40	78.84	84.23
	100	83.90	85.21	91.20	82.40	78.84	84.31
3	20	82.21	82.21	87.83	83.52	76.78	82.51
	40	83.90	83.52	89.14	82.96	78.46	83.60
	60	83.90	84.08	90.07	83.33	79.59	84.19
	80	83.90	84.08	90.26	83.52	79.96	84.34
	100	84.27	84.27	90.07	83.52	79.96	84.42
4	20	76.78	74.72	80.15	74.91	68.16	74.94
	40	79.21	77.15	84.08	79.40	71.91	78.35
	60	83.90	79.96	85.96	81.09	74.53	81.09
	80	82.21	81.09	86.70	82.77	75.84	81.72
	100	82.40	81.27	87.45	83.33	76.22	82.13
5	20	61.42	59.93	65.54	55.81	55.43	59.63
	40	65.92	65.54	70.97	64.04	58.80	65.06
	60	69.85	68.35	73.60	68.73	61.99	68.50
	80	73.03	70.97	76.22	71.54	64.61	71.27
	100	74.72	72.10	77.53	73.41	65.73	72.70

<표 4>는 SVM을 실행함에 있어서 커널함수로 다항식 함수를 사용한 경우의 결과를 나타낸다. d가 3일 때에 비교적 우수한 결과를 나타내었으나 전반적으로 가우시안 RBF를 사용한 경우보다 예측률이 떨어지는 경향을 볼 수 있었다.

<표 5>는 인공신경망 실험결과를 나타낸다. 은닉노드의 수가 32일 때 검증데이터의 예측정확성은 85.06%로 가장 높다.

<표 6>은 모든 모형의 학습결과를 나타낸 것이다. 기법 별로 실험을 진행한 결과를 예측정확성을 기준으로 정리하였다. 각 기법은 MDA, 로지스틱 회귀분석, 인공신경망, SVM이고, 인공신경망과 SVM은 여러 가지 파라미터를 조정하여 실험한 결과 중 가장 우수한 결과를 비교하였다.

결과 분석을 위하여 <표 6>에서 검증데이터의 결과만을 모아 기법 별 최고 성과를 보여주는 <표 7>을 구성하였다.

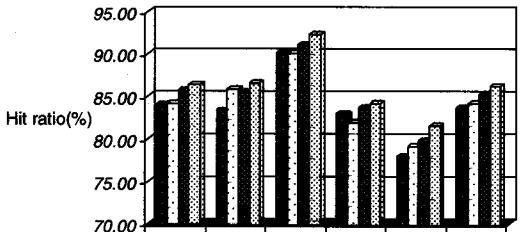
<표 5> 인공신경망 결과

은닉노드	데이터	Set 1	Set 2	Set 3	Set 4	Set 5	평균
8	모형훈련용	87.89	88.11	84.24	84.80	86.02	86.21
	모형검증용	84.27	85.21	90.64	83.15	80.15	84.68
12	모형훈련용	86.86	86.70	83.65	83.83	85.39	85.29
	모형검증용	84.46	85.58	90.82	82.21	79.78	84.57
16	모형훈련용	86.86	87.14	84.05	84.58	86.08	85.74
	모형검증용	84.83	84.83	90.45	82.58	79.21	84.38
24	모형훈련용	87.86	87.20	83.52	84.80	85.83	85.84
	모형검증용	85.96	84.46	90.82	82.02	79.40	84.53
32	모형훈련용	86.64	87.86	84.33	85.17	86.74	86.15
	모형검증용	85.58	85.39	91.01	83.52	79.78	85.06

<표 6> 전체 결과

실 험	MDA		로지스틱 회귀분석		ANN		SVM	
	훈련용 데이터	검증용 데이터	훈련용 데이터	검증용 데이터	훈련용 데이터	검증용 데이터	훈련용 데이터	검증용 데이터
Set 1	83.71	83.90	84.83	84.08	86.64	85.58	87.69	86.33
Set 2	84.32	83.33	84.93	85.77	87.86	85.39	87.50	86.52
Set 3	82.72	90.07	82.96	90.07	84.33	91.01	86.61	92.32
Set 4	84.69	82.96	85.35	82.02	85.17	83.52	88.90	84.08
Set 5	85.63	77.90	86.75	79.21	86.74	79.78	89.42	81.65
평균	84.21	83.63	84.96	84.23	86.15	85.06	88.02	86.18

<표 7> 기법별 최고 성과(모형검증용 데이터)



	Set 1	Set 2	Set 3	Set 4	Set 5	평균
MDA	83.90	83.33	90.07	82.96	77.90	83.63
Logit	84.08	85.77	90.07	82.02	79.21	84.23
ANN	85.58	85.39	91.01	83.52	79.78	85.06
SVM	86.33	86.52	92.32	84.08	81.65	86.18

<표 7>의 결과에서 나타난 것과 같이 예측력은 SVM이 가장 높았고 인공신경망, 로지스틱 회귀분석, MDA 순이었다. 각 set 별 결과로부터 항상 SVM의 결과가 우수하다는 것을 알 수 있다. 특히, Set 2에서는 인공신경망의 결과가 통계적 기법인 로지스틱 회귀분석보다 낮은 것을 볼 수 있는데, 이 때에도 SVM은 다른 기법들에 비해 가장 우수한 결과를 나타내었다. 상기 기법들의 예측력 차이의 유의성을 검증하기 위하여 McNemar Test를 실시하였다. 그 결과는 <표 8>과 같다.

<표 8>에서 나타난 것과 같이 로지스틱 회귀분석과 MDA는 통계적으로 유의한 차이를 나타내지 않았다. 두 통계적 기법과 인공신경망을 비교한 결과, 인공신경망이 MDA와는 1% 수준에서 유의하였고, 로지스틱 회귀분석과는 10% 수

<표 8> McNemar 값

	로지스틱 회귀분석	ANN	SVM
MDA	2.500	8.053***	24.397***
로지스틱 회귀분석		2.901*	15.482***
ANN			5.461**

주) *: 10% 수준에서 유의, **: 5% 수준에서 유의, ***: 1% 수준에서 유의

준에서 유의하였다. SVM은 MDA 및 로지스틱 회귀분석과 1% 수준에서 유의한 차이를 나타내었고, 인공신경망과도 5% 수준의 유의한 차이를 나타내었다. 이를 통해 부도예측에 있어서 SVM이 기존의 기법들 보다 우수한 예측정확성을 나타냄을 알 수 있다.

V. 결 론

본 연구에서는 최근 패턴인식 및 분류문제와 관련하여 활발하게 연구되고 있는 SVM을 기업부도예측에 적용하여 보았다. SVM은 전술한 바와 같이 통계적 이론에 기반하여 설명력이 우수하고, 구조적위험최소화접근에 따라 과대적합문제에서 벗어날 수 있으며, 볼록함수를 최소화하는 학습을 진행하기 때문에 유일한 최적해를 구할 수 있다는 점이 장점이다. 특히, 본 연구에서는 부도예측분야에 있어서 MDA, 로지스틱 회귀분석, 인공신경망과 비교하여 SVM의 적용가능성을 확인하고자 하였다. 실험 결과, 연구 데이터에 대해서 SVM은 상기 기법들보다 우수한 예측력을 보였으며, MDA 및 로지스틱 회귀분석, 인공신경망과의 예측력 차이가 통계적으로도 유의한 것으로 나타났다. 이와 같이 SVM은 인공신경망과 비슷한 수준의 높은 예측력을 나타낼 뿐만 아니라 인공신경망의 한계점으로 지적되었던 과대적합, 국소최적해와 같은 한계점들을 완화하는 장점을 기반으로 향후 재무분야의 분류문제에 있어서 유용할 것으로 생각된다.

〈참 고 문 헌〉

[1] 이수용, 이일병, "Fuzzy 이론과 SVM을 이용한 KOSPI 200 지수 패턴분류기," 한국증권학회 제4차 정기학술발표회, 2002, pp. 787-809.

[2] Chang, C.-C. and Lin, C.-J., LIBSVM: a library for support vector machines, Technical Report, Department of Computer Science and Information Engineering. National Taiwan University, Available at http://www.csie.edue.tw/~chlin/papers/libsvm.pdf, 2001.

[3] Hearst, M.A., Dumais, S.T., Osman, E., Platt, j. and Scholkopf, B., "Support vector machines," *IEEE Intelligent System*, Vol. 13, No. 4, 1998, pp. 18-28.

[4] Hornik, K., "Approximation Capabilities of Multilayer Feedforward Networks," *Neural Networks*, Vol. 4, 1991, pp. 251-257.

[5] Huang, Z., Chen, H., Hsu, C-J., Chen, W-H. and Wu, S., "Credit Rating Analysis with Support Vector Machine and Neural Networks: A Market Comparative Study," *Decision Support Systems*, Vol. 37, No. 4, 2004, pp. 543-558.

[6] Joachims, T., "Text Categorization with Support Vector Machines," *Proceedings of the European Conference on Machine Learning (ECML), 10th European Conference on Machine Learning*, 1998, pp. 137-142.

[7] Kim, K.J., "Financial Time Series Forecasting Using Support Vector Machines," *Neurocomputing*, Vol. 55, No. 1-2, 2003, pp. 307-319.

[8] Osuna, E., Freund, R. and Girosi, F., "Training Support Vector Machines: An Application to Face Detection," *Proceedings of Computer Vision and Pattern Recognition*, 1997, pp. 130-136.

[9] Patuwo, E, Hu, M.H. and Hung, M.S., "Two-group Classification Using Neural Networks," *Decision Science*, Vol. 24, No. 4, 1993, pp. 825-845.

[10] Tay, F.E.H. and Cao, L.J., "Application of Support Vector Machines in Financial Time Series Forecasting," *Omega*, Vol. 29, 2001, pp. 309-317.

[11] Tay, F.E.H. and Cao, L.J., "Modified Support Vector Machines in Financial Time Series Forecasting," *Neurocomputing*, Vol. 48, 2002, pp. 847-861.

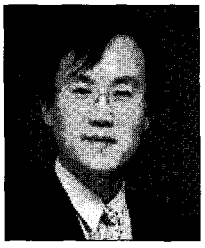
[12] Vapnik, V., "The Nature of Statistical Learning Theory," *Springer*, 1995.

[13] Weiss, S. and Kulikowski, C., "Computer Systems That Learn," *Morgan Kaufmann Publishers, Inc.*, 1991.

◆ 저자소개 ◆

박정민 (Park, Jung-min)
현재 하나은행 리스크관리본부 신용관리팀에 재직 중이다. 서강대학교를 졸업하고 KAIST에서 경영공학석사를 취득하였다. 주요 관심분야는 기업/개인 신용평가, 회계/재무 정보시스템, 인공지능 및 데이터마이닝 등이다.

김경재 (Kim, Kyoung-jae)
현재 동국대학교 경영대학 정보관리학과 교수로 재직 중이다. 중앙대에서 경영학사를, KAIST에서 경영정보시스템을 전공하여 공학석사와 박사를 취득하였다. 주요 관심분야는 데이터마이닝, 지능형 에이전트, 고객관계관리, 지식경영 등이다.



한인구 (Ingoo Han)
현재 한국과학기술원 테크노경영대학원 교수로 재직 중이다. 서울대학교 국제경제학사, KAIST 경영과학석사를 취득하였고, University of Illinois at Urbana-Champaign에서 회계정보시스템을 전공하여 경영학박사를 취득하였다. 주요 관심분야는 지능형 신용평가시스템, 인공지능을 이용한 재무예측, 지식자산 가치평가, 정보시스템 감사 및 보안 등이다.

◈ 이 논문은 2004년 5월 20일 접수하여 1차 수정을 거쳐 2004년 11월 10일 게재확정되었습니다.