Support Vector Machine을 이용한 기업부도예측

박정민[a]·김경재[b]·한인구[c]

[a] 한국과학기술원 테크노경영대학원
서울특별시 동대문구 청량리동 207-43
Tel: +82-2-958-3685, E-mail:paspark@kgsm.kaist.ac.kr

[b] 동국대학교 정보관리학과
서울특별시 중구 필동 3가 26
Tel: +82-2-2260-3324, E-mail:kjkim@dongguk.edu

[c] 한국과학기술원 테크노경영대학원
서울특별시 동대문구 청량리동 207-43
Tel: +82-2-958-3613, E-mail:ighan@kgsm.kaist.ac.kr

Abstract

기업부도예측은 전통적인 통계적 기법을 사용한 이래 1980년대 후반에는 인공지능기법을 적용하여 많은 연구가 진행되어 왔다. 대표적인 인공지능기법으로서 인공신경망이 많이 이용되어 왔으며 이를 부도예측 분야에 응용한 결과, 기존의 통계적 기법보다 우수한 예측성과를 얻을 수 있었다. 그러나 인공신경망은 우수한 예측력에도 불구하고, 학습과정에서의 과도적합현상으로 인한 일반화 어려움이나 결과 해석의 어려움이 한계로 지적되어 왔다. 이에 본 연구에서는 인공신경망 수준의 높은 예측력을 나타내면서 동시에 과대적합 문제를 해결하고 우수한 설명력을 제공하는 것으로 알려진 Support Vector Machine(SVM)을 기업부도예측모형에 적용한다. SVM은 벡터공간에서 임의의 비선형 경계를 찾아 두 개의 집합을 분류하는 방법으로서, 이미 많은 분야의 분류문제에 있어서 좋은 결과를 나타내고 있다. 본 연구에서는 기업부도예측모형으로서 SVM의 적합성을 판단하기 위하여 다변량판별분석, Logit, CBR, 인공신경망을 사용하여 그 성과를 비교한다. 본 연구의 결과 SVM이 기업부도예측모형에 있어서 다른 기법들에 비해 우수한 성과를 나타내었다.

Keywords:

SVM; 부도예측; MDA; Logit; ANN; CBR

1. 서론

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

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 입력 패턴의 분포를 추정하기 위해 다량의 학습데이터가 필요하고, 과도적합문제로 인해 일반화의 어려움이 있을 뿐 아니라, 지역적 최소값을 피하기 위한 초기화 작업이 경험에 의존하고, 결과를 해석하기 어렵다는 점 등이 한계로 지적되어 왔다.

본 연구에서는 이에 대한 해결방안으로 최근 각광받고 있는 SVM을 부도예측에 적용하고자 한다. SVM은 1998년 Vapnik에 의해 제안된 학습이론으로 분류문제를 해결하기 위해 최적의 분리경계면(hyperplane)을 제공한다. 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 분별학습을 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 과대적합을 피할 수 있다.

본 연구는 총 5장으로 구성하였다. 1장에서는 연구의 의의와 목적을 정의하고, 2장에서는 기존의 부도예측모형과 SVM에 관해 알아보도록 한다. 이어서 3장에서는 부도예측을 위한 SVM모형을 제안하고 MDA, Logit, CBR, 인공신경망 등의 방법론들과 비교분석을 한다. 그리고, 4장에서는

3장의 연구 결과를 분석하고, 5장에서는 본 연구의 시사점과 결론을 내린다.

2. 선행연구

기업부도예측에 관해서는 이미 상당한 연구가 이루어졌다. 기존 연구들을 방법론에 따라 통계적기법을 사용한 모형과 인공지능기법을 사용한 모형으로 나누어 살펴보고, 본 연구에서 제안하고자 하는 Support Vector Machine에 대해 알아보겠다.

2.1 통계적기법을 사용한 부도예측모형

통계적 모형을 이용하여 부도예측에 관한 실증분석을 수행한 연구는 Beaver[4]의 단일변량분석을 시초로 한다. 이 연구는 부실기업과 건전기업의 두 집단 간에 '영업활동에 의한 현금흐름/총부채비율' 및 '당기순이익/총자산비율'이 뚜렷한 차이가 있음을 보였으나, 재무비율의 평균치의 차이는 소수의 극단치에 의해 크게 영향을 받으며, 극단치를 제외할 경우 평균치의 차이가 줄어들 수 있다는 비판을 받았다. 또한 Beaver의 연구에서는 평균치의 차이가 통계적으로 유의한지에 대한 검증이 이루어지지 않았다.

단일변량분석연구는 기본적으로 한 변수에 의해 부도예측을 하는 경우 다른 변수의 영향을 고려하지 못한다는 한계가 있다. 이러한 단일변량분석의 한계를 극복하기 위해 다수의 변수를 하나의 모형에 통합하는 다변량분석 연구가 이루어졌다. 대표적으로는 Altman[1]이 사용한 다변량판별분석(Multiple Discriminant Analysis)을 들 수 있다. 그 후 Altman은 같은 기법으로 표본의 수를 늘리고, 자료를 최신화하여 연구를 계속하여 예측력을 향상시켰다. 그러나, 판별분석은 변수들의 분포가 다변량정규분포이고, 각 집단의 분산 및 공분산구조가 동일하다는 기본 가정을 충족시켜야 하는 통계적인 한계점을 안고 있다.

이러한 한계점을 완화하기 위하여 확률판별함수(probability discriminant function)가 제시되었다. 확률판별함수는 기업의 부도확률 또는 비부도확률을 퍼센티지로 제시하여 판별분석과는 달리 단정적인 평가를 회피하도록 하면서 적중률을 높일 수 있다. 이러한 확률모형에는 Logit 또는 Probit이 있다. 대표적인 연구로는 부실예측을 위해 Logit을 사용한 Ohlson[18]의 연구를 들 수 있다. 그 결과 예측력이 96.12%로 기존의 판별분석모형보다 높은 예측력을 나타내었다.

2.2 인공지능기법을 사용한 부도예측모형

1980년대 후반부터는 인공신경망, 귀납적 학습방법, CBR, 유전자 알고리즘 등 인공지능기법이 부도예측 분야에 응용되기 시작하였다. 인공신경망을 부도예측에 적용한 초기 연구로는 Odom and Sharda[17]의 연구를 들 수 있다. 이들은 판별분석과 인공신경망모형을 적용하여 부도예측률을 비교하였으며, 인공신경망모형이 판별분석에 비해 우수한 결과를 나타내었다.

Tam and Kiang[21]은 인공신경망을 이용하여 은행의 부도여부를 예측하고, 그 결과를 판별분석, Logit, k-최근접이웃방법(k-nearest neighbor), 귀납적 추론(ID3)의 결과와 비교하였다. 연구 결과, 인공신경망에 의한 모형이 예측성과, 적응력 등에서 다른 방법에 비해 우수한 결과를 나타내었다. 또한, Fletcher and Goss[10]는 Logit과 인공신경망 모형의 부도 예측률을 비교하였는데, 연구 결과는 다른 연구에서와 마찬가지로 인공신경망의 예측률이 우수한 것으로 나타났다.

국내 연구로는 이건창[28]이 판별분석, 귀납적학습방법, 인공신경망을 부도예측에 적용하였다. 그 결과 국외 연구에서와 같이 인공신경망이 가장 우수한 기업부도예측모형임을 실증분석하였다. 인공지능기법을 이용한 기존의 부도예측 연구들을 표 1에 요약·정리하였다.

이처럼 많은 연구들을 통해 입증된 인공신경망의 우수한 예측정확성에도 불구하고, 인공신경망의 결과는 설명력이 부족하여 예측결과의 원인을 설명하기 어렵다는 한계점과 함께 기존의 연구가 주로 소량의 자료를 대상으로 하였기 때문에 일반화의 가능성도 떨어진다는 한계점이 제기되어 왔다[13]. 또한, 인공신경망모형을 구축함에 있어 과도적합의 문제와 인공신경망 구조의 설계를 위해 많은 시간과 노력이 필요하다는 단점도 제기되었다[2].

표 1. 인공지능기법을 이용한 기업부도예측 주요 선행 연구

참고문헌	AI 기법	샘플사이즈	비교 통계기법	결과
[17]	ANN	129개	판별분석	ANN이 우수함
[5]	ANN	233개	Logit	ANN과 Logit의 결과가 비슷함
[21]	ANN	202개	판별분석, Logit, k-최근접이웃방법, 귀납적 추론	ANN이 우수함
[8]	ANN	282개	판별분석	ANN이 우수함
[10]	ANN	36개	Logit	ANN이 우수함
[9]	GANNA	190개	Logit	비슷함. Large size에서는 Logit이 우수함

참고문헌	AI 기법	샘플사이즈	비교 통계기법	결과
[24]	ANN	129개	판별분석	ANN이 우수함
[6]	ANN	342개	판별분석, Logit, Probit	ANN 결과가 판별분석, Logit, Probit의 결과와 유사함
[3]	ANN	237개	판별분석, Logit	전체 및 평가데이터에서는 ANN이 우수하지만 검증데이터에서는 비슷한 결과를 나타냄
[16]	ANN	166개	판별분석, ID3	Hybrid ANN 모델의 성과가 가장 우수함
[26]	GRG2	220개	Logit	ANN이 우수함

2.3 Support Vector Machine (SVM)

SVM은 1998년 통계학자인 Vapnik에 의해 개발된 학습기법으로, 입력공간과 관련된 비선형문제를 고차원의 특징공간의 선형문제로 대응시켜 나타내기 때문에 수학적으로 분석하는 것이 수월하다[11]. 또한, SVM은 조정해야 할 파라미터의 수가 많지 않아 비교적 간단하게 학습에 영향을 미치는 요소들을 규명할 수 있다. 그리고 구조적위험을 최소화함으로써 과대적합문제에서 벗어날 수 있으며, 볼록함수를 최소화하는 학습을 진행하기 때문에 global 최적해를 구할 수 있다는 점에서 인공신경망보다 성능이 좋은 기계학습기법으로 주목받고 있다.

최근 몇 년간 SVM을 사용한 다양한 연구가 진행되었다. 그 예로서 SVM은 문서분류, 영상인식, 문자인식 등에서 뛰어난 일반화 성능을 보여주었다[14,19]. 또한, SVM을 재무분야에 적용한 연구도 있는데, 주로 시계열 예측 및 분류에 관한 것이다[23,15]. 본 연구와 가장 유사한 연구로는 SVM을 사용하여 채권신용등급을 예측한 연구를 들 수 있다[25]. 이 연구에서 SVM은 일반화에 있어서 인공신경망이나 판별분석과 같은 다른 분류기법과 비교하여 비슷하거나 더 우수한 성능을 나타낸 것으로 보고하였다. 본 연구에서는 이러한 연구 배경을 토대로 SVM을 부도예측에 직접 적용하여 보기로 한다. 이를 위해 SVM에 대하여 간단히 설명하고자 한다.

SVM에서는 훈련데이터들을 서로 다른 두 개의 클래스로 분류할 때 분류의 기준이 되는 분리 경계면(hyperplane)를 학습 알고리즘을 이용하여 찾는다[27]. 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 사이를 최대로 분리한다. 최대마진 분리 경계면에 가장 근접한 훈련 데이터를 support vector라고 부른다.

선형분리문제에서, 애트리뷰트가 3개인 경우 분리 경계면은 식(1)과 같다.

$$y = w_0 + w_1x_1 + w_2x_2 + w_3x_3 \quad (1)$$

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

$$y = b + \sum \alpha_i y_i \mathbf{x}(i) \cdot \mathbf{x} \quad (2)$$

여기서, y_i는 훈련데이터 x(i)의 분류값이고, ·는 도트 프로덕트(dot product)이다. 벡터 x는 검증데이터를 나타내고, 벡터 x(i)는 support vector를 나타낸다. 이 식에서, b와 α_i는 분리 경계면을 결정하는 파라미터이다. support vector를 찾아내고, 파라미터 b와 α_i를 결정하는 것은 선형적으로 제약된 이차계획(QP) 문제(linearly constrained quadratic programming)를 푸는 것과 같다.

앞에서 언급한 바와 같이, SVM은 입력변수를 고차원의 특징 공간으로 이동시킴으로서 비선형 분류 문제를 선형모델로 구현한다.

비선형 분류문제에서 식(2)의 고차원 버전은 다음의 식(3)과 같이 간단하게 나타낼 수 있다.

$$y = b + \sum \alpha_i y_i K(\mathbf{x}(i), \mathbf{x}) \quad (3)$$

상기 식에서 함수 $K(\mathbf{x}(i),\mathbf{x})$는 커널함수라고 정의된다. 커널함수는 원래 데이터를 고차원 공간으로 사상시킴으로써 특징공간 내에 선형으로 분리가능한 입력 데이터셋을 만든다. 어떤 커널함수를 선택하는 것이 바람직한가는 문제에 따라 다르며, SVM을 적용하는 데 있어서 가장 중요한 요소이다. 일반적인 커널함수의 예로는 다항식 커널(polynomial kernel)과 가우시안 RBF(Gaussian radial basis function)를 들 수 있다:

$$\text{가우시안 RBF} : K(x,y)=\exp(-1/\delta^2(x-y)^2) \quad (4)$$

다항식커널함수: $K(x,y)=(xy+1)^d$ (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[15].

3. 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 데이터 건수가 적으므로 1996년부터 2000년까지의 부도기업 데이터를 사용하였다.
총 2670개의 재무데이터 중에서 모형구축용 표본으로는 부도기업과 건전기업을 50:50의 비율로 선정하여 총 80%을 사용하였고, 나머지 20%는 검증용 표본으로 사용하였다. 인공신경망의 경우에는 훈련데이터로 60%를 사용하고, 시험데이터로 20%, 그리고 나머지 20%를 검증용 데이터로 사용하였다.
모형에 사용할 재무비율을 선정하기 위하여 총 164개의 재무비율을 전처리하였다. 먼저 이상치 제거를 위하여 분포의 양측 1%에 데이터를 제거하고, 1차 통계분석을 통하여 결측치는 각 비율의 평균값으로 대체하였다. 그 후 단일변량 분석의 과정인 t-test를 통하여 건전 또는 부실에 유의한 변수들을 가려내었다. 단일변량검증을 통해 선택된 변수들에 대해 다시 다변량 분석의 과정으로 Logit의 단계적 변수선정방법을 통하여 최종적으로 15개의 변수를 선정하였다. 선정된 변수의 목록과 설명은 표 2와 같다.

표 2. 선정된 변수리스트

변수명	변수내역
금융비용대부채비율	당기 이자비용과 사채이장의 합이 당기 부채총계와 전기 부채총계의 합에서 차지하는 비
매출원가비율	당기 매출원가와 당기 매출액의 백분율
자기자본비율	당기 자본총계와 당기 자산총계의 백분율
금융비율부담율증가분	전기 대비 당기 금융비용부담율의 증가액
매입채무회전율	당기 매출액을 전기와 당기의 매입채무액으로 나눈 값
지급여력도	지급여력액과 당기매출액의 백분율
분식계수	분시금액추정액을 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 전전기 매출원가의 비율을 매출원가비율과 곱한 값
매출대비재고자산증가율	당기 재고자산합계와 전기 재고자산합계의 차액과 당기 매출액의 백분율
총자본회전율곱매출액증가비	총자본회전율과 전기 대비 단기 매출액 비율을 곱한 값

3.2 SVM

본 연구에서는 SVM의 커널함수로서 가장 널리 사용되는 다항식 커널과 가우시안 RBF를 사용하였다. SVM의 성능에 있어서 커널함수의 상한 C와 커널 파라미터 δ^2, d가 중요한 역할을 한다고 보고되었다[22]. SVM의 파라미터에 대해 제시된 일반적인 가이드를 따라 본 연구에서도 보고된 범위 내에서 다양한 값을 대입하여 모형을 변경시켰다. 실험은 LIBSVM[7]를 사용하였다.

3.3 ANN

본 연구에서 SVM에 대한 비교대상으로서 인공신경망, CBR, Logit, 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 모델의 설계에 관해서는 아직까지 체계적인 원칙이 없다. 인공신경망의 설계는 art에 가깝기 때문에 본 연구에서도 기존 연구의 결과를 바탕으로 몇 가지 실험을 통하여 가장 좋은 아키텍처를 선택하였다.
일반적으로 인공신경망의 성능에 영향을 미치는 요인으로서 은닉층의 수, 노드의 수, 학습횟수 등이 알려져 있다. Hornik[12]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 입력노드수와 출력노드수의 합을 n이라 할 때 n/2, n, 2n을 사용하지만, 모든 경우에 적합하다고 할 수는 없다.

은닉노드의 수는 인공신경망 아키텍쳐를 구성하는데 중요한 요소일 뿐만 아니라 데이터 의존적이다. 훈련데이터를 분류하는 경우에는 은닉노드의 수가 많을수록 바람직하지만, 검증데이터에서는 반드시 바람직한 것은 아니다[20]. 본 연구에서는 은닉노드의 수를 8, 12, 16, 24, 32로 구분하여 실험해보았다.
학습횟수는 너무 적으면 학습이 제대로 이루어지지 않고, 너무 많아도 훈련데이터에 과대적합되어 검증데이터의 예측력은 떨어진다. 적합한 학습횟수를 선택하기 위하여 본 연구에서는 learning epoch를 50, 100, 200, 300으로 나누어 실험하였다. 사용한 인공신경망의 학습률(learning rate)는 0.1, 모멘텀은 0.1로 고정하였다.

3.4 CBR(Case-Based Reasoning)

CBR은 문제해결을 위하여 가장 유사한 과거의 사례를 통해 새로운 문제를 해결하려는 기법이다.
CBR에서 관련된 사례를 추출하기 위해 근접이웃방법(Nearest Neighbor Method)을 사용하였다. 이 방법은 재무자료와 같은 숫자로된 자료에 쉽게 적용할 수 있기 때문에 많이 사용되는 추출방법이다. 본 연구에서는 최근접이웃의 수를 1에서 10까지 다양하게 설정하였다. 근접이웃방법을 사용하여 사례간 유사성을 추출하기 위하여 유클리디안 거리를 사용하였으며, 그 식은 다음과 같다:

$$D_{IR} = \sqrt{\sum_{i=1}^{n} w_i(f_i^I - f_i^R)^2} \quad (6)$$

여기서, D_{IR}은 f_i^I와 f_i^R 사이의 거리이고, f_i^I와 f_i^R는 입력값과 검색된 사례의 애트리뷰트 f_i의 값이고, n은 애트리뷰트의 수이고, w_i는 애트리뷰트 f_i의 가중치를 나타낸다.

3.5 Logit

Logit은 독립변수가 연속형 자료이고, 종속변수가 범주 혹은 명목척도인 경우에 분석하는 계량분석방법이다. Logit을 부도예측에 사용할 경우 기업의 설명변수의 관찰치벡터를 X_i로 하고, 그 계수 β_i를 추정한다면 기업의 부실확률은 로지스틱 함수에 의해 다음과 같이 유도된다.

$$Y_i = \frac{1}{1+e^{-P}} \quad (7)$$

여기서, $P = \beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_iX_i$이다.
본 연구에서는 Logit을 위해 SPSS 소프트웨어를 사용하였고, 판별점은 0.5를 기준으로 하였다.

3.6 MDA

다변량판별분석은 Logit과 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 다음과 같은 형태이다.

$$Z = W_1X_1+W_2X_2+W_3X_3+\ldots+W_nX_n \quad (8)$$

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

4. 연구결과

본 연구에서는 SVM의 실험결과를 각 커널함수와 파라미터에 따라 정리해보고, 추가적으로 인공신경망, CBR, Logit, MDA의 실험결과와 비교해보고자 하였다.
선형 SVM의 장점 중 하나는 조정해야 할 파라미터가 상수 C 이외에는 존재하지 않는다는 점이다. 그러나 선형 SVM으로 분리되지 않는 훈련데이터인 경우에는 계수 α_i의 상한인 C가 예측력에 영향을 미친다. 비선형 SVM인 경우에는 커널 파라미터도 조정해야 한다. 본 연구에서 사용한 커널함수는 가우시안 RBF와 다항식 함수이다. 본 연구에서는 상한 C와 커널 파라미터를 변경하면서 실험을 진행하였다.
가우시안 RBF에서는 C 이외에 δ^2을 고려해야 한다. Tay and Cao의 연구에 따르면, 적절한 δ^2의 범위는 1에서 100사이이고, C의 값으로 적합한 범위는 10에서 100사이라고 한다[22]. 이를 참고하여 본 연구에서도 C와 파라미터의 값을 세분화하여 실험하였고, 의미있는 결과를 위주로 정리하였다. ϵ은 0.001로 고정하였다. 표 3은 각 파라미터에 대한 SVM의 예측력을 나타낸 것이다.
표 3에 나타난 바와 같이 δ^2이 2이고, C가 20인 경우 가장 우수한 예측정확성을 나타내었다.

표 3. RBF 커널 사용시 SVM 결과

δ^2	C	훈련데이터	검증데이터
1	20	90.8665	85.2336
	40	91.8033	85.0467
	60	92.4122	85.0467
	80	92.6932	84.8598
	100	93.0211	85.2336
2	20	88.5246	86.5421
	40	89.2272	86.1682
	60	89.6487	85.9813
	80	90.0703	85.6075
	100	90.8197	85.2336
5	20	86.4637	84.1121
	40	87.4005	86.1682
	60	87.4473	85.7944
5	80	87.9157	86.3551

δ^2	C	훈련데이터	검증데이터
	100	88.0094	86.1682
10	20	85.7611	83.5514
	40	86.1827	83.9252
	60	86.4169	84.1121
	80	86.6042	84.486
	100	86.8384	85.0467
30	20	85.0117	83.5514
	40	84.918	83.5514
	60	85.1054	83.5514
	80	85.2927	83.7383
	100	85.3396	83.7383

그림 1과 그림 2는 δ^2과 C를 각각 고정하였을 때 C와 δ^2의 변화에 따른 훈련데이터와 검증데이터의 추이를 나타낸 것이다. 그림 1에서 알 수 있듯이 δ^2이 일정할 때 C가 증가할수록 훈련데이터는 과대적합되는 경향을 보였다. 그림 2에서는 C가 일정할 때 δ^2이 증가하면 훈련데이터가 과소적합되는 양상을 나타내었다. 이는 Tay and Cao[22], Kim[15]의 연구 결과와 유사한 것이다.

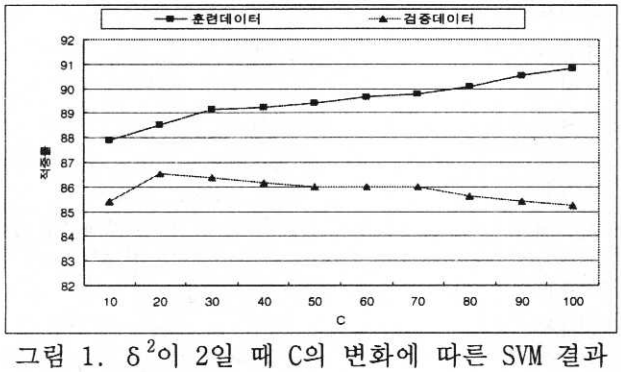


그림 1. δ^2이 2일 때 C의 변화에 따른 SVM 결과

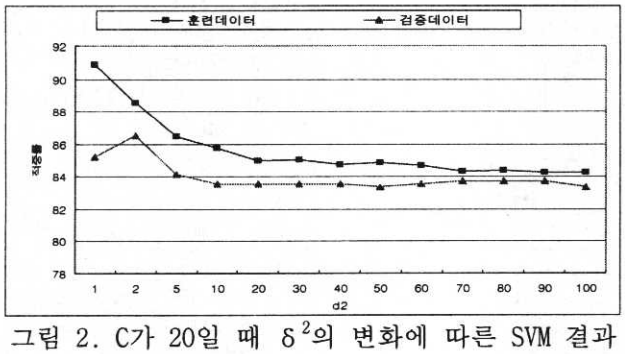
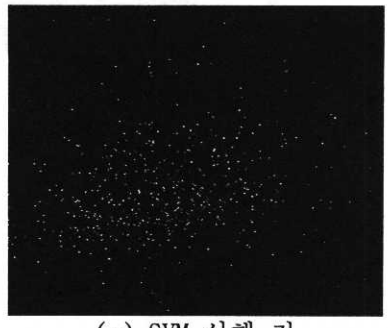
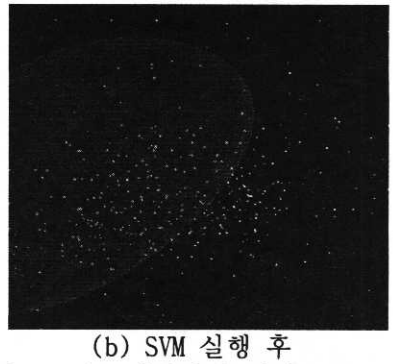


그림 2. C가 20일 때 δ^2의 변화에 따른 SVM 결과

그림 3의 (a)와 (b)는 SVM을 사용하기 전과 후 검증데이터의 패턴을 나타낸다. 그림 3(b)의 경계를 통해 두 개의 클래스로 분리된다.



(a) SVM 실행 전



(b) SVM 실행 후

그림 3. 검증데이터에 대한 SVM 결과

표 4는 SVM을 실행함에 있어서 커널함수로 다항식 함수를 사용한 경우의 결과를 나타낸다. RBF를 사용한 경우보다 예측률이 다소 떨어지며 δ^2가 일정할 때 차수가 증가할수록 예측력이 현저히 떨어지는 것을 알 수 있다. 또한 차수가 일정한 경우에는 δ^2가 증가할수록 예측력이 떨어지는 경향을 나타내고 있다.

표 4. 다항식 함수 커널 사용시 SVM 결과

δ^2	차수	훈련데이터	검증데이터
1	1	84.6838	83.5511
	2	86.0422	83.9252
	3	87.6815	85.7944
2	1	83.9813	83.7383
	2	82.3419	81.6822
	3	75.5035	77.0093
3	1	83.4660	83.7383
	2	75.4567	76.8224
	3	53.8642	53.4579

표 5는 인공신경망 실험결과를 나타낸다. 학습횟수가 늘어날수록 훈련데이터의 예측정확성이 높아지는 것을 알 수 있다. 가장 우수한 예측력을 나타낸 경우는 epoch가 200이고, 은닉노드수가 24 및 32일 때와 epoch가 300이고, 은닉노드수가 24인 경우이다. 이때 검증데이터의 예측정확성은 85.61%로 가장 높다.

표 5. 인공신경망 결과

학습횟수	은닉 노드	예측정확성(%)	
		훈련데이터	검증데이터
50	8	85.39	84.30
	12	84.64	84.11
	16	84.83	83.74
	24	84.46	84.11
	32	84.83	83.74
100	8	85.96	85.05
	12	84.83	85.42
	16	86.33	83.74
	24	85.39	85.23
	32	84.83	84.67
200	8	86.33	84.86
	12	86.52	85.23
	16	86.33	84.30
	24	85.96	85.61
	32	85.58	85.61
300	8	86.70	82.99
	12	86.89	84.86
	16	86.89	84.11
	24	86.89	85.61
	32	86.89	85.42

표 6은 CBR의 실험결과를 나타낸 것이다. 일반적으로 CBR에서 사용하는 근접이웃의 수는 흔히 1-10 사이에서 결정된다. 본 연구에서는 이들에 대한 실험을 수행하였고, 이 중 검증데이터에서 가장 높은 예측성과를 보인 근접이웃의 수는 3개이며, 이때 예측정확성은 83.93%이다.

표 6. CBR 결과

근접이웃의 수	예측정확성(%)
1	81.87
2	82.24
3	83.93
4	83.36
5	82.24
6	83.74
7	82.80
8	83.18
9	83.18
10	83.55

표 7은 본 연구에서 실험한 기법들의 최고의 성과들을 비교한 표이다. 예측력은 SVM이 가장 높았고 인공신경망, CBR, Logit, MDA 순이었다. 예측력 차이가 통계적으로 유의한가를 검증하기 위하여 McNemar 테스트를 실시하였다. 그 결과 표 8에 나타낸 바와 같이 SVM은 MDA, Logit 및 CBR과는 유의한 차이를 보였으나 인공신경망과는 그 차이가 유의하지 않았다.

표 7. SVM, ANN, Logit, MDA의 최고 예측정확성(%)

	MDA	Logit	CBR	ANN	SVM
훈련데이터	83.75	84.92	-	85.96	88.52
검증데이터	83.55	83.93	83.93	85.61	86.54

표 8. McNemar 값

	Logit	CBR	ANN	SVM
MDA	-♣	0.015	2.128	5.114**
Logit		0.000	1.488	4.024**
CBR			0.928	3.130*
ANN				0.489

*: 10% 수준에서 유의, **: 5% 수준에서 유의,
♣: 이항분포를 따름

5. 결론

본 연구에서는 최근 패턴인식 및 분류문제와 관련하여 활발하게 연구되고 있는 SVM을 기업부도예측에 적용하여 보았다. SVM은 통계적 이론에 기반하여 설명력이 우수하고, 구조적 위험 최소화 접근에 따라 과대적합문제에서 벗어날 수 있으며, 볼록함수를 최소화하는 학습을 진행하기 때문에 유일한 최적해를 구할 수 있다는 점에서 특히, 본 연구에서는 부도예측분야에 있어서 MDA, Logit, CBR, 인공신경망과 비교하여 SVM의 적용가능성을 확인하고자 하였다. 실험 결과, SVM은 상기 기법들보다 우수한 예측력을 보였으며, MDA 및 Logit, CBR과는 그 차이가 통계적으로도 유의한 것으로 나타났다. 선행연구들과 유사하게 본 연구에서도 인공신경망에 비해 SVM이 조금 더 높은 예측정확성을 나타내었으나 그 차이는 통계적으로 유의하지 않은 것으로 드러났다. 그럼에도 불구하고 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.

참고문헌

[1] Altman, E.I. (1968). "Financial Ratios, Discriminant Analysis and The Prediction of corporate Bankruptcy," *Journal of Finance*, pp. 589-609.

[2] Altman, E.I., Marco, G., and Varetto, F. (1994) "Corporate Distress Diagnosis Comparisons Using Linear Discriminant Analysis and Neural Networks," *Journal of Banking and Finance*, Vol. 18, No. 3, pp. 505-529.

[3] Barniv, R., Agarwal, A., and Leach, R. (1997). "Predicting the outcome following bankruptcy filing: a three-state classification using neural networks", *International Journal of Intelligent Systems in Accounting, Finance and Management*, Vol. 6, No. 3,

pp. 177-194.

[4] Beaver, W.H.(1966). "Financial Ratios and Prediction of Failure, Empirical Research in Accounting : Selected studies," *Journal of Accounting Research*, Vol. 5, pp. 71-111.

[5] Bell, T., Ribar, g., and Verchio, J. (1990). "Neural nets vs. logistic regression: a comparison of each model's ability to predict commercial bank failures", *Proceedings of the 1990 Deloitte & Touche/University of Kansas Symposium on Auditing Problems*, pp. 29-58.

[6] Boritz, E., Kennedy, D., and Albuquerque, A. (1995). "Predicting corporate failure using a neural network approach," *International Journal of Intelligent Systems in Accounting, Finance and Management*, Vol. 4, No. 2, pp. 95-112.

[7] Chang, C.-C., and Lin, C.-J. (2001). 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.

[8] Coates, P., and Fant, L., (1993). "Recognizing financial distress patterns using a neural network tool", *Financial Management*, pp. 142-155.

[9] Fanning, K., and Cogger, K., (1994). "A comparative analysis of artificial neural networks using financial distress prediction", *International Journal of Intelligent Systems in Accounting, Finance and Management*, Vol. 3, No. 3, pp. 241-252.

[10] Fletcher, D., and Goss, E., (1993). "Forecasting with neural networks: and application using bankruptcy data", *Information and Management*, Vol. 24, pp. 159-167.

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

[12] Hornik, K. (1991). "Approximation capabilities of multilayer feedforward networks," *Neural Networks*, Vol.4, pp. 251-257.

[13] Jo, H., and Han, I. (1996). "Integration of case-based forecasting, neural network, and discriminant analysis for bankruptcy prediction," *Expert Systems with Applications*, Vol. 11, pp. 415-422.

[14] Joachims, T. (1998). "Text categorization with support vector machines," *Proceedings of the European Conference on Machine Learning (ECML), 10th European Conference on Machine Learning*, pp. 137-142.

[15] Kim, K.J. (2003). "Financial time series forecasting using support vector machines," *Neurocomputing*, forthcoming.

[16] Lee, K., Han, I., Kwon, Y. (1996) "Hybrid neural networks for bankruptcy predictions," *Decision Support Systems*, Vol. 18, pp. 63-72.

[17] Odom, M., and Sharda, R., (1990). "A neural network model for bankruptcy prediction", *Proceedings of the International Joint Conference on Neural networks*, pp. II-163-II-168.

[18] Ohlson, J.A. (1980). "Financial Ratios and Probabilistic Prediction of Bankruptcy," *Journal of Accounting Research*, pp. 109-131.

[19] Osuna, E., Freund, R., and Girosi, F. (1997). "Training support vector machines: an application to face detection," *Proceedings of Computer Vision and Pattern Recognition*, pp. 130-136.

[20] Patuwo, E, Hu, M.H., and Hung, M.S. (1993). "Two-group classification using neural networks," *Decision Science*, Vol. 24, No. 4, pp. 825-845.

[21] Tam, K., and Kiang, M. (1992). "Managerial applications of neural networks: the case of bank failure prediction", *Management Science*, Vol. 38, No.7, pp. 926-947.

[22] Tay, F.E.H., and Cao, L. (2001). "Application of support vector machines in financial time series forecasting," *Omega*, Vol. 29, pp. 309-317.

[23] Tay, F.E.J., and Cao, L.J. (2002). "Modified support vector machines in financial time series forecasting," *Neurocomputing*, Vol. 48, pp. 847-861.

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

[25] Huanjg, Z., Chen, H., Hsu, C-J., Chen, W-H., Wu, S. (2003). "Credit rating analysis with support vector machine and neural networks: a market comparative study," *Decision Support Systems*, forthcoming.

[26] Zhang, G., Hu, M. Y., Patuwo, B.E., and Indro, D. C. (1999). "Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis," *European Journal of Operational Research*, Vol. 116, pp. 16-32.

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

[28] 이건창 (1993). "기업도산예측을 위한 통계적 모형과 인공지능 모형간의 예측력 비교에 관한 연구: MDA, 귀납적학습방법, 인공신경망," *한국경영학회지* 제 18권 제 2호, pp. 57-81.