Support Vector Machine을 이용한 부도예측모형의 개발*

- 격자탐색을 이용한 커널 함수의 최적 모수 값 선정과 기존 부도예측모형과의 성과 비교 -

민재형** · 이영찬***

Support Vector Bankruptcy Prediction Model with Optimal Choice of RBF Kernel Parameter Values using Grid Search*

Jae H. Min** · Young-Chan Lee***

Abstract

Bankruptcy prediction has drawn a lot of research interests in previous literature, and recent studies have shown that machine learning techniques achieved better performance than traditional statistical ones. This paper employs a relatively new machine learning technique, support vector machines (SVMs), to bankruptcy prediction problem in an attempt to suggest a new model with better explanatory power and stability. To serve this purpose, we use grid search technique using 5-fold cross-validation to find out the optimal values of the parameters of kernel function of SVM. In addition, to evaluate the prediction accuracy of SVM, we compare its performance with multiple discriminant analysis (MDA), logistic regression analysis (Logit), and three-layer fully connected back-propagation neural networks (BPNs). The experiment results show that SVM outperforms the other methods.

Keyword : Bankruptcy Prediction, Support Vector Machine, Grid Search, Kernel Function, Back-Propagation Neural Networks

논문접수일 : 2004년 7월 22일 논문게재확정일 : 2004년 12월 22일
* 이 논문은 2004년도 두뇌한국21사업에 의해 지원되었음.
** 서강대학교 경영대학 경영학과
*** 동국대학교 상경대학 전자상거래학과(교신저자)

1. 서 론

IMF 구제금융신청 이후 국제결제은행(Bank for International Settlement : BIS)의 신용위험 규제, 금융시장에서의 대출경쟁 심화 등으로 인해 국내 금융기관들은 신용위험의 계량화를 통한 효율적인 여신 관리에 주력하고 있다[7]. 은행, 보험회사, 증권사, 신용보증기관 등과 같은 금융기관들이 신용대출을 하게 될 경우 주된 관심사는 대출고객의 신용위험을 측정하여 고객의 채무불이행(default) 여부를 사전에 얼마나 정확하게 예측할 수 있느냐에 집중되어 있다. 특히 기업의 부도는 주주나 채권자는 물론 종업원, 고객, 정부 모두에게 경제적 손실을 초래하고 사회적 부를 감소시키기 때문에 금융기관의 신용위험노출(credit risk exposure)을 효과적으로 관리하기 위해서는 재무 및 운용위험과 비효율성 정도를 측정, 감시, 통제할 수 있는 분석기법을 기반으로 한 의사결정지원시스템의 도입이 무엇보다도 필요하다[1, 3, 29].

최근 들어 금융기관들은 대출고객의 잠재적 신용위험수준을 평가하기 위해 내부적으로 신용평점모형을 구축하고 이를 활용하고 있는데, 이러한 신용위험의 계량화는 여신부문에 있어 채무불이행 가능성을 예측하고 이에 따른 손실의 위험성을 정확하게 파악할 수 있으며, 손실위험 정도에 따라 조기에 대응책을 지시해 줄 수 있다는 이점이 있다[14, 51, 52, 76]. 이러한 관점에서 신용평점화는 넓은 의미에서 보편적인 판별 및 분류 문제로 이해될 수 있다.

이러한 신용의사결정문제를 지원하기 위해 그동안 선형확률 및 다변량 조건부 확률모형, 순환적 분할(recursive partitioning) 알고리즘, 인공지능, 다기준 의사결정, 수리계획법 등과 같은 다양한 방법론들이 제안되어 왔으며[10, 12, 15, 16, 20, 25, 29, 30, 35, 45, 55, 63, 64, 69, 71, 80], 특히 인공신경망을 이용한 연구는 예측력이 우수하여 가장 많이 사용되고 있다. 그러나 인공신경망을 이용할 경우 입력패턴의 분포를 추정하기 위해 다량의 학습자료가 필요하고, 과대적합으로 인해 일반화의 어려움이 있을 뿐만 아니라 국지적 최소값을 피하기 위한 사전처리 작업이 연구자의 경험이나 지식에 의존하며, 결과에 대한 해석이 어렵다는 점 등이 인공신경망 기법의 공통적인 한계로 지적되어 왔다[36, 47, 56, 65, 66, 75].

본 연구에서는 이러한 인경신경망 기법의 문제점을 해결하기 위한 방안으로 최근 다양한 분류문제에 도입되어 그 성과가 우수한 것으로 알려진 Support Vector Machine(SVM)을 부도예측에 적용하고자 한다[31, 32, 38, 43, 46, 70, 72, 74]. SVM은 Vapnik[73]이 제안한 학습이론으로 분류문제를 해결하기 위해 최적의 분리초평면(hyperplane)을 제공한다. SVM이 주목받는 이유는 첫째, 명확한 이론적 근거에 기반하기 때문에 결과 해석이 용이하고, 둘째, 실제 응용에 있어서 인공신경망 수준의 높은 성과를 내며, 셋째, 적은 양의 학습 자료만으로도 신속하게 분리 학습을 수행할 수 있기 때문이다. 또한 SVM은 기존의 학습 알고리즘이 경험적 위험 최소화 원칙(ERM : empirical risk minimization)을 구현하는 것인데 비해 구조적 위험 최소화 원칙(SRM : structural risk minimization)에 기반함으로 과대적합을 피할 수 있다[1, 11, 18, 37].

특히, 본 연구에서는 부도예측을 위한 SVM 모형을 구축하는데 있어 가장 중요한 커널함수의 모수 값을 최적으로 선택하기 위해 v-겹(v-fold) 교차타당성(cross-validation)을 이용한 격자탐색(grid search)이라는 새로운 기법을 도입하고자 하며[42], 이러한 격자탐색을 통해 선정된 최적 모수 값은 안정성과 예측력이 우수한 부도예측모형을 구축하는데 매우 유용할 것으로 기대된다. 또한 본 연구에서는 부도예측모형으로서 SVM의 적합성을 검증하기 위하여 기존의 문헌에서 예측력이 우수하다고 알려진 인공신경망과 다변량 판별분석 및 로짓분석과 같은 다변량 통계분석 기법과의 성과 비교도 함께 수행하고자 한다.

2. 이론적 고찰

2.1 기존의 부도예측기법

부도예측에 대한 과학적인 연구는 Beaver[8]에 의해 본격적으로 시작되었다. Beaver[8]는 단일변량 통계분석기법을 이용하여 부도기업과 건전기업 간의 차이를 가장 잘 구분하는 지표를 개발한 바 있으며, Altman[4]은 다변량 판별분석을 이용하여 개별적으로 관찰되던 재무비율을 종합하고 단순화하여 기계적이고 명확한 형태의 부도예측을 가능하게 하였다. 이후 여러 연구자들이 부도예측에 다변량 판별분석을 성공적으로 도입한 바 있다[5, 27, 30, 62]. 1970년대에는 선형확률 및 다변량 조건부 확률모형(Logit 모형과 Probit 모형)을 이용한 부도예측 연구가 이루어졌는데, 이러한 방법론들은 기업의 부도가능성을 확률로 추정하였다는 점에서 의의를 가진다[24, 58]. 그리고 1980년대에는 이진분류 나무모형에 기반한 순환적 분할(recursive partitioning) 알고리즘이 부도예측 연구에 많이 활용되었다[35, 68].

한편, 1980년대 후반부터는 인공신경망 기법과 귀납적 학습(inductive learning) 등의 인공지능 기법들이 부도예측 연구에 활발히 도입되었는데[21, 28, 44, 53, 54, 61, 67, 76, 78, 79], 다층 퍼셉트론 구조를 기반으로 한 인공신경망 기법의 예측성과와 판별분석, 로지스틱 회귀분석, k-최근접이웃방법, 귀납적 추론, 의사결정나무 등 다른 기법의 예측성과를 비교한 연구결과를 보면 대부분 인공신경망의 예측력이 상대적으로 우수한 것으로 나타났다[6, 9, 17, 19, 22, 33, 34, 45, 49, 57, 69, 77].

그러나 이러한 연구들을 통해 입증된 인공신경망의 높은 예측력에도 불구하고, 인경신경망은 학습과정 자체가 블랙박스이며, 과대적합으로 인해 실험결과의 일반화가 어렵다는 한계점이 지적되어 왔다. 또한 자료의 사전처리와 최적의 인공신경망 구조설계가 연구자의 기술(art)에 상당히 의존한다는 단점도 제기되고 있다[47, 65].

2.2 Support Vector Machines

SVM은 Vapnik[73]에 의해 개발된 학습기법으로, 입력 벡터를 고차원(high dimensional)의 특징공간(feature space)으로 이동시켜(mapping) 분리 경계가 매우 복잡한 문제를 선형판별함수의 사용이 가능한 단순한 문제로 변환시키기 때문에 수학적인 분석이 수월하고, 조정해야 할 모수(parameter)의 수가 많지 않아 비교적 간단하게 학습에 영향을 미치는 요소들을 규명할 수 있다는 장점을 갖고 있다. 또한 SVM은 구조적 위험을 최소화함으로써 과대적합문제에서 벗어날 수 있으며, 볼록함수를 최소화하는 학습을 진행하기 때문에 전역적 최적해(global optima)를 구할 수 있다는 점에서 인공신경망보다 성능이 우수한 기계학습기법으로 주목 받고 있다[11, 18, 37, 40].

SVM은 두 집단으로 구분된 입력 벡터를 가지는 훈련용 자료에 대해 집단을 분류할 때 기준이 되는 분리초평면(separating hyperplane)을 특수한 학습알고리즘을 이용하여 찾게 되는데, 이를 구체적으로 설명하면 다음과 같다[11, 18, 37, 73].

두 개의 집단 $y_i \in \{-1, +1\}$으로 분리된 입력벡터 $x_i = \{x_i^{(1)}, \cdots, x_i^{(n)}\}^T \in \mathbf{R}^n$를 가지는 훈련용 자료 $D = \{x_i, y_i\}_{i=1}^N$이 있다고 하자. Vapnik[73]이 최초 제안한 공식에 따르면 SVM은 다음의 조건을 만족한다.

$$\begin{cases} \mathbf{w}^T \phi(x_i) + b \geq +1, & \text{if } y_i = +1 \\ \mathbf{w}^T \phi(x_i) + b \leq -1, & \text{if } y_i = -1 \end{cases} \quad (1)$$

또는

$$y_i[\mathbf{w}^T \phi(x_i) + b] \geq 1, \; i = 1, \cdots, N \quad (2)$$

여기서 $\mathbf{w}$는 가중치 벡터를 나타내며, b는 편차(bias)를 나타낸다. 비선형 함수 $\phi(\cdot) : \mathbf{R}^n \rightarrow \mathbf{R}^{n_k}$은 입력벡터를 고차원의 특징공간으로 이동시키는(mapping) 역할을 수행한다. 식 (2)는 특징공간에서 분리초평면 $\mathbf{w}^T \phi(x) + b = 0$을 사이에 두고 반대쪽

에 두 개의 평행한 경계초평면(bounding hyperplane)을 만들게 되는데, 그 마진(margin)의 폭은 $\frac{2}{\|\mathbf{w}\|^2}$ 가 된다. 이러한 조건을 만족하는 가중치 공간에서 분류의 결과는 식 (3)과 같이 도출된다.

$$\mathrm{sgn}(\mathbf{w}^T\phi(x)+b) \tag{3}$$

그러나 선형 분리가 불가능한 문제가 대부분이므로 식 (4) 및 식 (5)와 같이 오분류(misclassification)를 허용할 수 있도록 오차항(ξ_i)을 도입한 후 가중치 벡터를 찾는 것이 일반적이다.

$$\underset{\mathbf{w},b,\xi}{\mathrm{Min}}\ \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}+C\sum_{i=1}^{N}\xi_i \tag{4}$$

subject to

$$\begin{cases} y_i(\mathbf{w}^T\phi(x_i)+b)\geq 1-\xi_i, & i=1,\cdots,N \\ \xi_i\geq 0, & i=1,\cdots,N \end{cases} \tag{5}$$

여기서, ξ_i는 제약식의 부등식에서 오분류를 허용하는데 필요한 여유변수(slack variable)이며, 목적함수에 있는 $C\in\mathrm{R}^+$는 마진폭에 대응하는 분류오차의 가중치이다. 식 (4)와 식 (5)로 구성된 최소화 문제는 선형 제약식을 가진 이차계획(quadratic programming : QP) 모형으로, 최적해는 라그랑지 승수 α_i를 이용하여 구하게 된다. 승수 α_i는 훈련용 데이터 각각에 곱해지는데, 만약 비음인 α_i가 존재한다면 이 승수에 대응하는 데이터를 support vector라고 한다.

한편, 위의 최소화 문제는 식 (6)의 목적함수와 식 (7)의 제약식을 갖는 쌍대 문제(dual problem)로 전환할 수 있는데, 이 경우 의사결정변수는 support vector, 즉, 라그랑지 승수가 되기 때문에 해석하기가 용이하다는 장점이 있다.

$$\underset{\boldsymbol{\alpha}}{\mathrm{Max}}\ \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{Q}\boldsymbol{\alpha}-\mathbf{e}^{\mathrm{T}}\boldsymbol{\alpha} \tag{6}$$

subject to

$$\begin{cases} 0\leq\alpha_i\leq C, & i=1,\cdots,N \\ \mathbf{y}^T\boldsymbol{\alpha}=0 \end{cases} \tag{7}$$

여기서, $\mathbf{e}$는 단위행렬을 나타내고, Q는 $N\times N$ 양(+)의 준정부호 행렬(positive semi-definite matrix)로 $Q_{ij}=y_iy_jK(x_i,x_j)$와 같은 등식으로 표현되며, $K(x_i,x_j)=\phi(x_i)^T\phi(x_j)$는 커널함수를 나타낸다. 따라서 훈련용 벡터 x_i는 커널함수 ϕ에 의해 고차원의 특징 공간으로 이동된다는 것을 알 수 있다. 그러나 고차원 문제를 주로 다루는 SVM에서 $\mathbf{w}$나 $\phi(x)$를 실제로 계산할 수 없기 때문에 매핑함수인 $\phi(x)$를 식 (8)과 같은 커널함수 $K(\cdot,\cdot)$로 연결시켜 주는 Mercer의 조건(Mercer's condition)을 이용하여 문제를 풀게 된다.

$$K(x_i,x_j)=\phi(x_i)^T\phi(x_j) \tag{8}$$

이와 같이 커널함수는 이론적으로는 가능하나 실제로는 풀기 힘든 매핑함수를 대신하여 원자료를 고차원 공간으로 이동시켜 특징공간 내에 선형으로 분리 가능한 입력자료 집합을 만들어 주는 역할을 수행한다. 이때 어떤 커널함수를 선택하는 것이 바람직한가는 문제의 종류에 따라 다르며, SVM을 적용하는데 가장 중요한 요소이다. 대표적인 커널함수로는 차수 d를 가지는 다항식 커널인 $K(x_i,x_j)=(\gamma x_i^Tx_j+r)^d$, $\gamma>0$과 RBF(radial basis function) 커널인 $K(x_i,x_j)=\exp\{-\gamma\|x_i-x_j\|^2\}$, $\gamma>0$이 있다. 여기서 $d,r\in N$와 $\gamma\in\mathbf{R}^+$는 상수이다. 따라서 식 (3)의 분류기(classifier)는 커널함수를 이용하여 식 (9)의 SVM 분류기로 최종 구성된다.

$$\mathrm{sgn}\left(\sum_i^N\alpha_it_iK(x,x_i)+b\right) \tag{9}$$

최근 부도예측[31, 32, 38, 72], 신용등급분석[43], 시계열예측[46, 70], 보험사기적발[74] 등 다양한 분야에 SVM을 적용한 연구가 발표되고 있는데, 이들 연구에서 SVM은 실험결과의 일반화에 있어 인공신경망이나 LVQ(learning vector quantization), 의사결정나무모형, 다변량 판별분석, 로지스틱 회귀분석, 사례기반추론 등과 같은 다른 분류기법과 비교하여 비슷하거나 더 우수한 성능을 나타내는 것으

로 보고되고 있다[31, 38, 46, 72]. 본 연구에서는 이러한 연구결과를 토대로 SVM을 부도예측에 적용하는 연구를 수행하였다.

3. 연구모형

3.1 자료수집과 사전처리

부도예측모형을 구축하기 위해 본 연구에서는 기술신용보증기금이 보유하고 있는 944개의 건전기업과 944개의 부도기업 등 총 1,888개 기업의 재무비율 자료를 수집하였다. 건전기업의 자료는 자산규모 10억 이상 70억 이하의 국내 비외감 제조중공업 기업의 2002년도 재무자료를 기준으로 하였고, 이에 대응하는 부도기업의 자료 역시 자산규모 10억원 이상 70억원 이하의 국내 비외감 제조중공업 기업의 자료를 기준으로 하였다. 그러나 일반적으로 부도기업은 건전기업보다 매년 발생하는 자료 건수가 적으므로 부도기업의 경우 1999년부터 2002년까지의 자료를 사용하였다. 또한 이상치를 제거하기 위해 분포의 양측 꼬리 1%를 초과하는 값들은 모두 1% 값으로 조정한 후 평균을 구하여 이 값으로 결측치를 대체하였다.

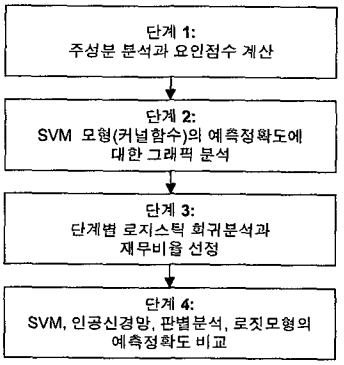
분석에 사용할 재무비율은 모두 [-1, 1] 사이의 값을 가지도록 단위를 조정하였다.[1] 이와 같이 자료를 정규화(normalize)하게 되면 분석에 사용되는 모든 재무비율의 분산이 동일한 범위 내에 있게 되므로 측정 단위에 따른 예측오차를 줄일 수 있게 된다[62]. 분석에 사용할 재무비율의 선정은 두 단계로 이루어졌다. 먼저 주성분 분석을 수행하기 위한 재무비율은 t-검정을 통해 선정하였으며, 최종 분석에 사용할 재무비율은 단계별 로지스틱 회귀분석을 이용하여 선정하였다.

1) 한국은행의 기업경영분석에 소개된 수익성, 안정성, 활동성, 유동성, 성장성, 현금흐름 관련 비율을 토대로 1차 t 검정과 기존 문헌[4, 5, 8, 30, 45, 48, 49, 50, 58, 60]의 고찰을 통해 유의하다고 판명된 비율들을 선정하였음.

SVM, 다변량 판별분석, 로지스틱 회귀분석의 경우 데이터 셋은 훈련용과 검증용의 두 가지로 구분되는데, 전체 데이터의 80%(1510/1888)는 훈련용(training) 데이터 셋으로 사용하고, 나머지 20%(378/1888)는 검증용(holdout)으로 사용하였다. 여기서, 검증용 데이터 셋은 모형구축 과정에는 사용하지 않는다. 한편, 인공신경망의 경우 데이터 셋은 훈련용, 시험용(cross-validation), 그리고 검증용의 세 가지로 구분되는데, 전체 데이터의 60%(1132/1888), 20%(378/1888), 20%(378/1888)를 각각의 용도로 사용하였다. 인공신경망의 경우, 학습과정에서 과대적합이 발생할 가능성이 크기 때문에 시험용 데이터를 사용하여 학습과정이 적절히 이루어졌는가를 확인하게 된다.

3.2 분석절차

본 연구는 [그림 1]과 같은 분석절차에 따라 진행된다.



[그림 1] 분석절차

단계 1에서는 t-검정을 통해 1차 선정된 38개의 재무비율 변수에 대해 주성분 분석을 수행하여 다차원적인 재무성과를 2개의 주성분으로 축약한다. 단계 2에서는 0.5를 상회하는 주성분 변수의 요인적재량을 이용하여 건전/부도 기업들의 요인점수를 계산하고, 이를 학습용 입력자료로 사용하여 두 종류의 커널함수(RBF, 다항식)에 따른 SVM의 학습능력을 비교한다. 단계 3에서는 단계별 로지스틱 회

귀분석을 이용하여 부도예측에 사용될 최종 재무비율 변수를 선정한다. 단계 4에서는 선정된 재무비율 변수를 이용하여 SVM, 인공신경망, 판별분석, 그리고 로지스틱 회귀분석의 예측정확도를 비교한다.

3.3 SVM의 적용

본 연구에서는 SVM의 기본 커널함수로 RBF (radial basis function)를 사용한다. RBF 커널함수를 이용한 SVM의 예측성과는 라그랑지 승수(support vector)의 허용오차 상한값인 C와 커널 모수(parameter)인 γ에 따라 영향을 받게 된다[42, 70]. 따라서 두 가지 모수가 적절하게 선택되지 않을 경우 과대적합이나 과소적합 현상이 발생할 수 있다. 최근 Hsu et al.[42]은 v-겹(v-fold) 교차타당성(cross-validation)과 격자탐색(grid search) 방법을 이용하여 SVM의 모수 선택에 대한 실용적인 방안을 제안한 바 있다. 본 연구에서는 이들이 제시한 방법을 부도예측에 도입하여 SVM의 최적 모수 값을 선택한다.

허용오차 상한값 C와 커널 모수 γ의 값을 최적으로 선택하는 목적은 분류기(classifier)로 하여금 부도여부에 대한 정보가 알려져 있지 않은 검증용 데이터를 대상으로 부도예측을 정확히 하도록 하는 것이다. 따라서 SVM의 모수가 훈련용 데이터의 예측정확도를 가장 높이도록 선택되면 과대적합이 발생할 가능성이 있으므로(훈련용 데이터는 부도여부에 대한 정보가 주어지므로) 적절한 모수 선택을 하려면 훈련용 데이터를 임의의 두 집합으로 분리한 후 그 중 하나는 부도여부에 대한 정보를 주지 않은 상태에서 학습을 시켜야 한다. 이러한 방법은 부도여부를 모르는 상태에서 새로운 데이터의 예측정확도를 향상시키는데 매우 유용하며, 이러한 방법을 개선한 것이 바로 v-겹(v-fold) 교차타당성(cross-validation) 방법이다. 즉, v-겹 교차타당성 방법에서는 훈련용 데이터를 동일한 표본크기를 갖는 v개의 부분집합으로 나눈 후 (v-1)개의 부분집합(부도여부에 대한 정보가 주어짐)을 훈련용 데이터로 사용하여 학습을 시킨 후, 나머지 하나의 부분집합(부도여부에 대한 정보가 주어지는 훈련용 데이터로 사용하지 않음)에 대해 검증하는 과정을 연속적으로 수행하게 된다. 교차타당성은 앞서 언급한 바와 같이 과대적합을 방지하기 위한 목적으로 사용되는데, 본 연구에서는 이러한 교차타당성 기법을 도입하여 C와 γ의 값을 최적으로 선택하고자 하며, C와 γ의 조합에 따른 교차타당성 분석과정을 격자탐색(grid search)이라고 한다[42].

구체적으로, C와 γ의 조합을 여러 개 만들고 이에 대해 교차타당성 분석을 수행하여 정확도가 가장 높게 나타나는 C와 γ의 값을 최종적으로 선택하게 된다. 본 연구에서는 C와 γ의 값을 지수적으로 증가시켜 가면서 조합을 만들었다. 예를 들어, C는 2^{-5}, 2^{-3},…, 2^{15}와 같이 증가시키고, γ는 2^{-15}, 2^{-13}, …, 2^{3}와 같이 증가시키면서 각각의 조합을 만들고, 각 조합에 대해 교차타당성 분석을 실시하여 정확도가 가장 높은 C와 γ를 찾는다. 이 경우 훈련용 데이터를 이용한 모수탐색 과정에 상당히 많은 시간이 소요된다는 단점이 있는데[2], 이러한 단점에도 불구하고 본 연구에서 격자탐색을 사용하는 이유는 다음과 같다. 첫째, 기존의 SVM 연구[43, 46, 70]에서 사용하고 있는 휴리스틱한 접근방법을 이용하여 C와 γ의 값을 찾는 것은 매우 소모적인 작업일 뿐만 아니라 최적해를 보장하지 못한다. 둘째, 선택할 모수가 두 개인 경우 격자탐색 기법은 비교적 신속한 처리속도를 보인다. 셋째, C와 γ는 상호 독립적이기 때문에 조합을 만들기가 용이하다[42]. 한편, 본 연구에서는 RBF 커널에 대한 최적 모수 값을 선택한 후 이 모수 값이 다른 커널에서도 유용한지의 여부를 확인하기 위해 다항식 커널에 대한 예측정확도도 함께 비교한다. SVM 실험도구로는 *LIBSVM*[13]을 사용하였다.

2) 자료의 크기 및 변수의 개수와 종류, 그리고 컴퓨터의 성능에 따라 차이가 있지만 경험적으로 볼 때 1,000개 이상 크기의 자료와 10개 이상의 변수를 사용하는 경우 격자탐색을 수행하는데는 20분 이상이 소요된다.

3.4 SVM의 예측성과 비교 기법

본 연구에서는 SVM의 유용성을 검증하기 위한 비교대상으로 인공신경망, 다변량 판별분석, 그리고 로지스틱 회귀분석을 이용하였다.

3.4.1 인공신경망

인공신경망은 부도예측 분야에서 가장 많이 사용되어 온 방법으로 통계적 기법이나 기타 인공지능 기법에 비해 예측력이 우수하다고 알려져 있다. 그러나 인공신경망 모형의 설계에 대한 일반적인 원칙은 아직까지 존재하지 않으며, 연구자의 경험과 지식에 크게 의존하기 때문에 본 연구에서도 기존의 연구 결과를 바탕으로 반복적인 실험을 통해 가장 좋은 아키텍쳐를 선택하였다. 일반적으로 신경망의 성능에 영향을 미치는 요인으로는 은닉층 수, 은닉노드 수, 학습 회수 등이 있는데[41, 56, 65], 어떤 값이 최적인지에 대한 일반적인 규칙은 없다. 다만, 분류 문제를 포함한 대부분의 문제에서 한 개의 은닉층으로도 만족할만한 결과를 얻을 수 있다는 선행 연구[41]를 토대로 본 연구에서도 은닉층이 하나인 3층 퍼셉트론을 사용하였다.

은닉층의 노드 수는 경험적으로 입력노드 수와 출력노드 수의 합을 n이라 할 때 $n/2$, n, $2n$ 등을 사용하지만, 이러한 노드 수가 모든 경우에 적합하다고 할 수는 없다. 은닉노드의 수는 인공신경망 아키텍쳐를 구성할 때 고려해야 할 중요한 요소일 뿐만 아니라 데이터에 매우 의존적이다. 예를 들어, 훈련용 데이터를 분류하는 경우에는 은닉노드의 수가 많을수록 바람직하지만, 검증용 데이터에서는 이것이 반드시 바람직한 것은 아니다[61]. 본 연구에서는 은닉노드의 수를 8, 12, 16, 24, 32와 같이 다양하게 변화시키면서 실험을 수행하였다. 학습 회수는 너무 적을 경우 과소적합 문제가 발생하고, 너무 많을 경우에는 과대적합 문제가 발생하므로 은닉노드와 마찬가지로 50, 100, 200, 300으로 다양하게 그 값을 변화시키면서 학습이 이루어지도록 하였다. 학습율과 모멘텀은 각각 0.1과 0.7로 고정하였고, 은닉노드와 출력노드에서의 전이함수는 양극 시그모이드 함수를 사용하였다. 인공신경망의 실험도구로는 *NeuroSolutions* 4.32를 사용하였다.

3.4.2 다변량 판별분석

다변량 판별분석은 사전에 정해진 집단(본 연구에서는 부도와 건전)을 가장 잘 판별해 내는 선형 판별함수를 도출하기 위한 통계적 기법이다. 선형 판별함수는 집단내 분산대비 집단간 분산비율을 최대로 하는 통계적 의사결정규칙을 생성하게 되는데, 이를 식으로 나타내면 식 (10)과 같다.

$$Z = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n \tag{10}$$

여기서, Z는 판별점수를 나타내며 $w_i (i = 1, 2, \cdots, n)$는 판별 가중치를 나타낸다. 그리고 $x_i (i = 1, 2, \cdots, n)$는 독립변수인 재무비율을 나타낸다. 각 기업에 대해 계산된 판별점수를 절사점(cut-off)과 비교하여 어떤 집단에 속하는지를 결정하게 된다.

다변량 판별분석은 독립변수가 다변량 정규분포를 따르고, 각 집단의 공분산행렬이 동일할 때는 유용하지만 부도기업의 경우 정규성에 대한 가정이 위배되는 경우가 많고, 집단별 공분산이 동일하다는 가정도 위배되는 경우가 많다. 특히, 독립변수간에 다중공선성이 존재할 경우 단계별 분석을 적용하게 되면 심각한 오류를 발생할 가능성이 높다[39].

3.4.3 로지스틱 회귀분석

로지스틱 회귀분석은 독립변수가 연속형 또는 범주형 자료이고, 종속변수가 범주형 또는 명목형 자료인 경우에 사용하는 통계분석기법이다[58]. 로지스틱 회귀분석을 부도예측에 사용할 경우, 특정 기업에 대하여 독립변수의 관찰치 벡터를 x_i로 하고, 가중치 w_i를 선택한다면 해당 기업의 부도확률은 식 (11)과 같이 계산된다.

$$\text{Probability of default} = \frac{1}{1+e^{-Z}} = \frac{1}{1+e^{-(w_0 + w_1 x_1 + \cdots + w_n x_n)}} \tag{11}$$

로지스틱 회귀분석은 독립변수가 다변량 정규분

포를 따르고 집단간 공분산행렬이 동일하다는 가정을 하지 않아도 된다는 장점이 있다.

본 연구에서 로지스틱 회귀분석은 다음과 같은 두 가지 목적으로 사용된다. 첫째, 분석에 사용될 재무비율을 선정하기 위해 단계별 로지스틱 회귀분석을 실시하였다. 이 경우 수집된 자료를 훈련용과 검증용으로 따로 분리하지 않고 모든 자료를 함께 이용하였다. 둘째, SVM의 유용성을 비교하기 위한 벤치마킹 대상으로 로지스틱 회귀분석을 수행하였다. 이 경우에는 수집된 자료를 훈련용과 검증용으로 분리하고, 검증용 데이터의 예측정확도를 SVM의 예측정확도와 비교하게 된다.

4. 실증분석

4.1 요인분석

본 연구에서는 최초에 수집한 재무비율 변수(65개)에 대한 독립표본 t-검정을 실시하여 통계적으로 유의한 38개의 재무비율 변수를 선정하고, 이에 대하여 요인분석(Varimax 직교회전)을 실시하였다. 요인추출은 주성분 분석을 이용하였으며, 요인의 개수는 2개로 한정하였다.[3] <표 1>은 요인분석 결과를 요약한 것으로 요인적재량이 0.5 이상인 재무비율은 모두 23개로 나타났다.

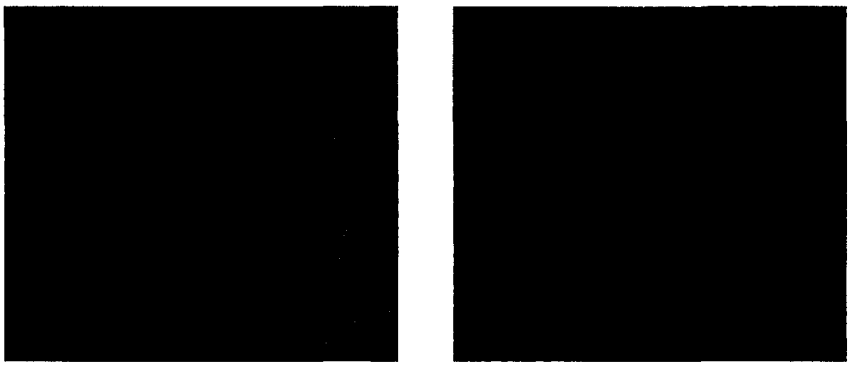
<표 1>의 두 요인을 구성하는 재무비율 중에서 적재량이 0.5 이상인 재무비율들을 이용하여 건전과 부도기업의 요인점수를 산출하고, 이를 [-1, 1] 사이의 값으로 정규화한 후 산점도로 나타낸 결과는 [그림 2]와 같다. 여기서 자주색 데이터 계열(왼쪽 하단에 분포)은 부도기업을, 하늘색 데이터 계열(오른쪽 상단에 분포)은 건전기업을 각각 나타낸다. [그림 2]에서 보는 바와 같이 건전기업은 부도기업에 비해 상대적으로 안전성 및 활동성, 그리고 수익성 및 현금흐름의 정규화 점수가 높음을 알 수 있다.

〈표 1〉 요인분석 결과

변 수	안정성/활동성 요인	수익성/현금흐름 요인
유형자산증가율	-0.058	0.091
총자산경상이익율*	0.541	0.698
총자산순이익율*	0.529	0.713
자기자본경상이익율	0.374	0.446
자기자본순이익율*	0.444	0.521
매출액경상이익율*	0.342	0.761
매출액순이익율*	0.327	0.771
변동비대매출액*	0.544	-0.313
EBITDA대매출액*	-0.246	0.714
감가상각율	0.400	0.177
차입금평균이자율	0.008	0.048
금융비용대총비용*	-0.761	-0.130
금융비용대매출액*	-0.770	-0.220
순금융비용대매출액*	-0.727	-0.199
이자보상비율*	0.408	0.604
손익분기점율	0.430	-0.210
자기자본비율	0.053	0.304
현금흐름대전기단기차입금	-0.103	0.431
현금흐름대단기차입금	-0.128	0.448
현금흐름대차입금*	-0.103	0.541
현금흐름대총부채*	-0.174	0.568
현금흐름대이자비용*	-0.136	0.519
고정비율	-0.346	-0.066
고정장기적합율	-0.214	-0.051
차입금대매출액*	-0.749	-0.279
총자산회전율*	0.889	0.066
자기자본회전율*	0.503	-0.112
자본금회전율*	0.523	-0.027
경영자산회전율*	0.866	0.067
고정자산회전율*	0.758	-0.157
유형자산회전율*	0.634	-0.178
재고자산회전율	0.141	0.266
매입채무회전율	0.017	0.157
총자본투자효율	0.485	0.451
설비투자효율*	0.557	-0.071
부가가치율	-0.396	0.485
지급여력도*	-0.593	0.062
경상수지비율*	0.027	0.566
고유값	8.639	6.192
설명된 분산	22.734	16.296
누적 설명된 분산	22.734	39.030

*) 요인적재량이 0.5 이상인 재무비율을 나타냄.

3) 본 연구에서 SVM 실험도구로 사용한 LIBSVM은 학습성과 및 예측성과를 2차원 형태로만 시각화 할 수 있기 때문에 다양한 재무비율 변수를 사용한 다차원 분석에 앞서 SVM의 학습 및 예측성과를 탐색적으로 살펴보기 위하여 요인의 수를 2개로 한정하였다.



(a) 훈련용 데이터의 산점도 (b) 검증용 데이터의 산점도

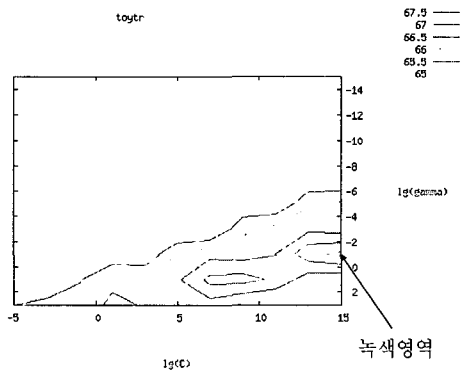
[그림 2] 주성분의 요인점수 산점도

본 연구에서는 [그림 2]의 산점도와 같이 분리경계가 복잡한 문제를 풀 때 RBF 커널을 이용한 SVM 모형과 다항식 커널을 이용한 SVM 모형의 학습성과 및 예측성과를 시각적으로 비교하였다.

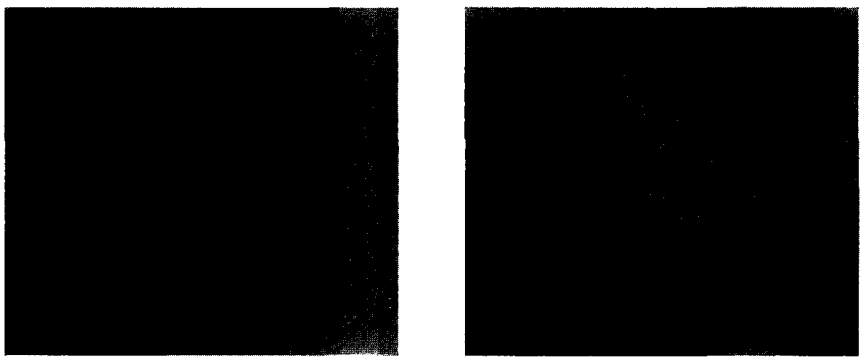
4.2 SVM 모형의 학습성과 및 예측성과 비교

RBF 커널과 다항식 커널을 이용한 SVM 모형의 학습성과와 예측성과를 시각적으로 분석하기에 앞서 본 연구에서는 먼저 허용오차 상한값인 C와 커널 모수 γ의 값을 각각 2^{15}와 2^{-1}로 설정하였다. 이 값들은 5-겹 교차타당성을 이용한 격자탐색 과정을 통해 도출되었는데, 이에 대한 자세한 설명은 4.4절에 기술하였다.

[그림 3]에서 보는 바와 같이 C와 γ값이 각각 2^{15}와 2^{-1}일 때 성과가 가장 우수함을 알 수 있다(녹색영역). 격자탐색을 통해 도출된 최적 모수값을 이



[그림 3] $C=2^{-5}, 2^{-3}, \cdots, 2^{15}$와 $\gamma=2^{-15}, 2^{-13}, \cdots, 2^{3}$의 조합에 대한 격자탐색

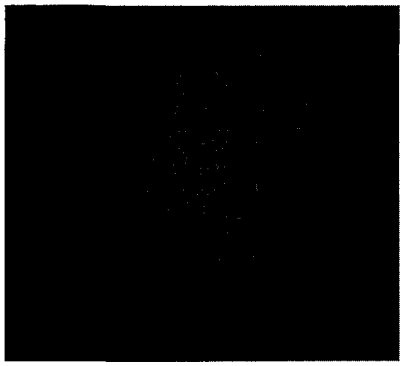


(a) 훈련용 데이터 (b) 검증용 데이터

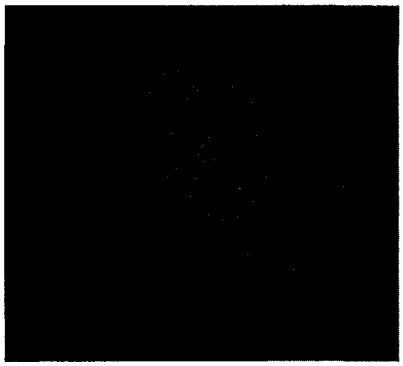
[그림 4] RBF 커널을 이용한 SVM 실행 결과

용하여 RBF 커널과 다항식 커널을 사용한 SVM 모형의 성과를 비교하였다. [그림 4]와 [그림 5]는 두 개의 요인점수를 입력벡터로 하는 분리문제에서 RBF 커널과 다항식 커널을 각각 이용한 경우의 성과를 시각적으로 나타낸 것이다.

먼저 훈련용 데이터와 검증용 데이터에 대해 RBF 커널을 이용한 SVM 모형을 실행한 결과는 [그림 4]와 같다. 여기서, 종속변수는 부도여부를 나타내는 명목변수이고, 독립변수는 두 개의 요인점수이다.



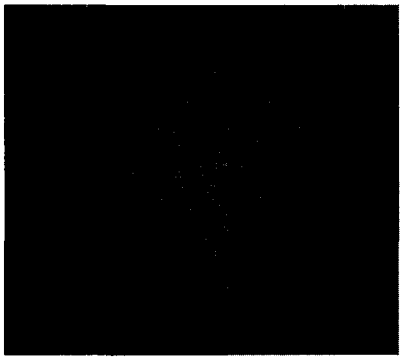
(a) 훈련용 데이터(d=1)



(b) 검증용 데이터(d=1)

(a) 훈련용 데이터(d=2)

(b) 검증용 데이터(d=2)



(a) 훈련용 데이터(d=3)

(b) 검증용 데이터(d=3)

[그림 5] 다항식 커널을 이용한 SVM 실행 결과

RBF 커널을 이용한 SVM 모형은 support vector를 결정하기 위해 고차원 특징공간(high dimensional feature space)으로 이동된 가우시안(Gaussian) 벡터에 대해 학습을 시작한다. 일단 support vector가 결정되면 분리초평면 경계에 있는 데이터에 모형이 수렴하면서 입력벡터를 분리하게 된다. [그림 4]에서 보는 바와 같이 SVM은 건전과 부도기업을 적절하게 분류하고 있음을 확인할 수 있으며, 이때 예측정확도는 훈련용 데이터의 경우 67.2185%이고 검증용 데이터의 경우 66.1376%이다.

다음으로 훈련용 데이터와 검증용 데이터에 대해 다항식 커널을 이용한 SVM 모형을 실행한 결과는 <그림 5>와 같다. 다항식 커널의 경우 허용오차 상한값 C와 커널 모수인 γ이외에 차수 d라는 추가적인 모수가 필요하다. 본 연구에서는 다항식 커널의 차수를 1, 2, 3으로 조정해 가면서 예측성과를 분석하였다.

[그림 5]에서 보는 바와 같이 다항식 커널함수를 사용한 경우, 과소적합 또는 과대적합 현상이 발견됨을 알 수 있다. 구체적으로, 차수가 1과 2일 때 훈련용 데이터의 예측성과는 각각 65.5629%와 66.8212%인데 비해 검증용 데이터의 예측성과는 각각 70.1058%와 68.5185%로 나타나 기계학습의 전형적인 과소적합 현상이 발생하고 있음을 알 수 있다. 한편, 차수가 3일 때는 훈련용 데이터의 예측성과와 검증용 데이터의 예측성과가 각각 67.8808%와 65.6085%로 나타났는데, 이를 RBF 커널을 이용한 경우와 비교해보면 훈련용 데이터의 예측성과는 다항식 커널이 RBF 커널보다 높게 나타났지만 검증용 데이터의 경우는 다항식 커널의 예측성과가 RBF 커널보다 낮게 나타나 과대적합 현상이 발견되고 있다. 따라서 다항식 커널함수를 사용한 SVM 모형은 최적의 차수 d를 찾기 위한 추가적인 노력이 요구된다는 것을 알 수 있다.

4.3 재무비율 변수의 선정

앞서 4.2절에서 38개 재무비율 변수에 대한 요인분석을 이용하여 RBF 커널과 다항식 커널을 이용한 두 가지 SVM 모형의 학습성과를 시각적으로 살펴보았다. 두 요인의 적재량을 이용하여 계산한 요인점수만을 독립변수로 사용했음에도 불구하고 SVM의 예측성과는 비교적 매력적이라는 것을 확인할 수 있었다. 이러한 시각적 분석결과에 근거하여 본 연구에서는 SVM, 인공신경망, 다변량 판별분석, 그리고 로지스틱 회귀분석 모형의 예측성과를 비교하고자 한다. 이를 위해 단계별 로지스틱 회귀분석을 이용하여 분석에 사용할 최종 재무비율 변수를 선정하였다. 선정된 11개 변수들과 그 산출식을 정리하면 <표 2>와 같다.

〈표 2〉 선정된 재무비율 변수 목록

변 수 명	산 출 식
자기자본경상이익율	경상이익 ÷ 자기자본
변동비대매출액	변동비 ÷ 매출액
금융비용대총비용	이자비용 ÷ (매출원가 + 판매비와관리비 + 영업외비용)
금융비용대매출액	이자비용 ÷ 매출액
손익분기점율	손익분기점에서의 매출액 ÷ 매출액
자기자본비율	자기자본 ÷ 총자산
현금흐름대단기차입금	영업활동후 현금흐름 ÷ 단기차입금
고정비율	고정자산 ÷ 자기자본
자본금회전율	매출액 ÷ 자본금
경영자산회전율	매출액 ÷ 경영자산
고정자산회전율	매출액 ÷ 고정자산

부도예측의 성과는 재무비율 변수가 무엇이냐에 따라 좌우된다고 해도 과언이 아니다. 기존 연구[1, 3, 45, 49, 50, 60]를 살펴보면 건전/부도 여부를 판별하는 데에는 통계적으로 유의하지만 경제적으로는 그 의미를 해석하기 힘든 변수들이 선정되는 경우가 많았다. 본 연구에서는 한국은행이 기업경영분석을 위해 사용하고 있는 재무비율을 중심으로 경제적으로 해석이 용이하면서도 기업의 재무건전성을 파악하는데 유용한 변수를 선정하는데 주력하였다.

4.4 SVM과 기존 부도예측모형의 예측정확도 비교

4.4.1 SVM 모형

SVM 모형에서 데이터 셋은 훈련용과 검증용으로 구분되며, 각각의 용도로 전체 데이터(1,888개)의 80%와 20%를 사용하였다.

SVM을 실행하기 위해서는 먼저 어떤 커널함수를 사용할 것인가를 결정해야 한다. 다음으로 SVM 모형의 모수인 C와 커널함수 모수를 선택해야 하는데, 선형 커널함수는 상한값 C이외에 별도로 선택해야 할 모수가 없다는 장점은 있으나 선형으로 분리가 불가능한 문제에서는 효과적이지 못하다는 단점이 있다[18]. SVM에서 선형으로 분리가 불가능한 문제를 해결하는데 사용되는 커널함수로는 RBF 커널과 다항식 커널이 대표적인데, 이 경우 커널함수의 모수가 존재하기 때문에 추가적으로 모수의 값을 선택해야 한다.

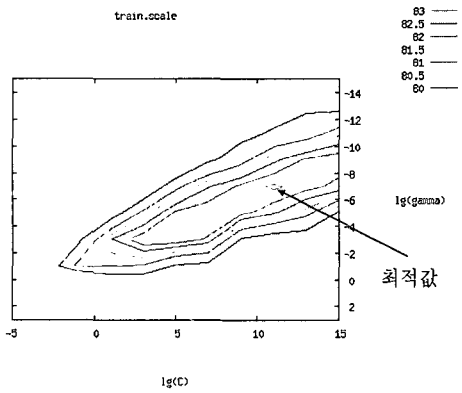
구체적으로, RBF 커널은 고차원의 특징공간으로 입력 벡터를 비선형적으로 이동시키기 때문에 선형분리가 불가능한 문제를 해결하는데 매우 유용하다. 그러나 다항식 커널은 차수 d라는 추가적인 모수의 선택이 필요하고, 학습에 걸리는 시간도 RBF 커널에 비해 상대적으로 더 많이 소요된다는 단점이 있다. 또한 예측정확도 측면에서도 다항식 커널은 RBF 커널과 비교하여 성과가 낮은 것으로 보고되고 있다[43, 46, 70]. 따라서 본 연구에서는 RBF 커널을 이용한 SVM 모형을 기본 모형으로 사용하였다.

RBF 커널함수는 SVM 모형의 허용오차 상한값 C와 모수 γ에 따라 예측성과가 달라지기 때문에 최적의 C와 γ의 값을 선정해야 하는 사전 작업이 필요하다. 본 연구에서는 앞서 기술한 바와 같이 5-겹 교차타당성을 이용한 격자탐색 기법을 도입하여 최적의 C와 γ의 값을 도출하고, 이를 이용하여 SVM 분석을 수행하였다. 11개 재무비율 변수로 구성된 훈련용 데이터를 대상으로 격자탐색을 수행한 결과, C와 γ의 값이 각각 2^{11}과 2^{-7}일 때 교차타당성의 예측성과가 가장 우수한 것으로 나타났으며, 이때의 예측성과는 83.1126%이다(<표 3>과 [그림 6] 참조).

최적의 C와 γ의 값이 구해지면 이를 이용하여 훈련용 데이터 전체에 대하여 학습을 다시 수행하고, 최종적인 SVM 분류기(classifier)를 생성한 후 검증용 데이터를 대상으로 예측정확도를 분석하게 된다. C와 γ의 값을 2^{11}과 2^{-7}으로 하는 RBF 커널함수를 사용하여 SVM 분석을 수행한 결과, 검증용 데이터에 대한 예측정확도는 83.0688%(부도 :

〈표 3〉 5-겹 교차타당성을 이용한 격자탐색 수행 결과

C \ γ	2^{3}	2^{1}	2^{-1}	2^{-3}	2^{-5}	2^{-7}	2^{-9}	2^{-11}	2^{-13}	2^{-15}
2^{-5}	47.2185	47.2185	74.4371	71.9205	57.2848	47.2185	47.2185	47.2185	47.2185	47.2185
2^{-3}	47.2185	62.9801	79.1391	77.2848	74.4371	67.8808	47.2185	47.2185	47.2185	47.2185
2^{-1}	48.1457	74.3709	81.2583	79.6689	77.7483	75.3642	68.1457	47.2185	47.2185	47.2185
2^{1}	62.5828	77.5497	81.0596	81.9868	79.2715	78.1457	75.6954	68.0795	47.2185	47.2185
2^{3}	62.5828	77.6159	80.9272	82.9139	81.1921	78.8079	78.0795	75.5629	67.8808	47.2185
2^{5}	62.5828	77.5497	79.8675	82.9139	82.5828	80.7285	78.4106	77.8808	75.5629	67.7483
2^{7}	62.5828	77.5497	79.7351	82.4503	82.8477	81.9205	79.7351	78.0795	77.6159	75.5629
2^{9}	62.5828	77.5497	79.7351	79.9338	82.7152	82.6490	80.6623	79.6026	78.0795	77.5497
2^{11}	62.5828	77.5497	79.7351	79.3377	82.1192	**83.1126**	81.8543	80.1325	79.3377	78.0132
2^{13}	62.5828	77.5497	79.7351	79.3377	81.2583	82.7152	82.5828	80.4636	79.8013	79.4702
2^{15}	62.5828	77.5497	79.7351	79.2053	79.9338	82.3179	82.8477	81.3245	79.6026	79.8013



[그림 6] 5-겹 교차타당성을 이용한 격자탐색 과정

82.0106%, 건전 : 84.1270%)로 나타났으며, 이때 훈련용 데이터에 대한 예측정확도는 88.01%이다. 한편, 본 연구에서는 RBF 커널함수에 사용된 C와 γ의 값을 동일하게 적용하여 다항식 커널(차수는 1, 2, 3으로 구분)을 사용했을 때의 예측정확도도 추가적으로 분석하였다. 분석결과를 정리하면 <표 4>와 같다.

〈표 4〉 SVM 분석 결과

커 널	C	γ	d	예측정확도(%)	
				훈련용 데이터	검증용 데이터
RBF	211	2-7	N/A	**88.0132**	**83.0688**
다항식*	211	2-7	1	80.3311	77.2487
			2	86.6225	83.8624
			3	88.4768	82.0106

*) 다항식 커널함수의 추가 모수인 r은 1로 고정하였음.

4.4.2 인공신경망 모형

인공신경망에서 데이터 셋은 훈련용과 시험용, 그리고 검증용으로 구분되며, 전체 데이터(1,888개)의 80%, 20%, 그리고 20%를 각각의 용도로 사용하였다. <표 5>는 인공신경망 모형의 실험 결과를 정리한 것이다.

<표 5>에서 보는 바와 같이 학습 회수가 늘어날수록 훈련용 데이터의 예측정확도가 높아지는 것을 확인할 수 있다. 가장 우수한 예측력을 나타낸 경우는 학습 회수가 300이고, 은닉노드 수가 24인 경우이며, 검증용 데이터에 대한 예측정확도는 82.5397%(부도 : 79.3651%, 건전 : 85.7143%)로 나타났다. 이때 훈련용 데이터에 대한 예측정확도는 85.2474%이다.

〈표 5〉 인공신경망 실험 결과

학습 회수	은닉노드 수	예측정확도(%)	
		훈련용 데이터	검증용 데이터
100	8	83.1272	80.6878
	12	82.9505	80.4233
	16	81.5371	80.1587
	24	81.0954	79.8942
	32	83.5689	81.2169
200	8	84.8057	80.9524
	12	86.7491	81.2169
	16	82.6855	81.2169
	24	85.1590	81.2169
	32	85.2474	81.7460
300	8	84.8940	82.0106
	12	86.1307	81.4815
	16	87.0141	80.6878
	24	**85.2474**	**82.5397**
	32	88.1625	81.4815

한편, <표 5>에서 인공신경망의 최고 예측정확도는 학습 회수가 300이고 은닉노드 수가 32일 때의 88.1625%로서, RBF 커널함수를 사용한 SVM과 거의 비슷한 수준이다. 그러나 인공신경망은 학습 회수가 증가함에 따라 훈련용 데이터의 예측정확도는 향상되지만 검증용 데이터의 예측정확도는 오히려 낮아지는 현상이 발생함을 알 수 있는데, 이는 훈련과정에서 과대적합이 이루어졌다는 것을 의미한다.

따라서 SVM과 인공신경망의 예측정확도에 대한 실험결과를 종합해 볼 때 SVM이 과대적합을 피할 수 있을 뿐만 아니라 소모적인 아키텍쳐 설계를 해야 하는 인공신경망에 비해 예측성과도 우수하다고 할 수 있다.

4.4.3 다변량 판별분석과 로지스틱 회귀분석 모형

다변량 판별분석과 로지스틱 회귀분석 모형에서

데이터 셋은 SVM과 동일하게 훈련용과 검증용으로 구분되며, 전체 데이터(1888개)의 80%와 20%를 각각의 용도로 사용하였다. 다변량 판별분석과 로지스틱 회귀분석의 예측성과를 정리하면 <표 6>과 같다.

〈표 6〉 다변량 판별분석과 로지스틱 회귀분석 실험 결과

모 형	예측정확도(%)	
	훈련용 데이터	검증용 데이터
다변량 판별분석	78.8079	79.1391
로지스틱 회귀분석	79.8676	78.3069

<표 6>에서 보는 바와 같이 로지스틱 회귀분석 모형이 다변량 판별분석 모형보다 검증용 데이터에서 약 0.8% 정도 예측성과가 더 높은 것을 알 수 있다.

4.4.4 예측성과의 비교

<표 7>은 본 연구에서 실험한 기법들의 최고 예측성과를 비교한 것이다. 예측정확도는 SVM, 인공신경망, 로지스틱 회귀분석, 다변량 판별분석의 순으로 나타났으며, SVM이 다른 기법에 비해 적게는 0.5%에서 많게는 4.8% 정도의 높은 예측성과를 나타내었다.

본 연구에서는 분석기법들간의 예측정확도 차이가 통계적으로 유의한 지를 검증하기 위해 검증용 데이터를 대상으로 McNemar 검정을 실시하였다. McNemar 검정은 비모수 통계분석기법으로 이진값을 가지는 명목형 변수에 대해 관련이 있는 두 집단간의 차이를 검정할 때 사용한다. 특히, McNemar 검정은 동일한 대상에 대한 처리 전·후의 측정치 비교에 매우 유용한 것으로 알려져 있다.

<표 8>에서 보는 바와 같이 예측정확도 측면에서 SVM은 다변량 판별분석과 로지스틱 회귀분석과는 통계적으로 유의한 차이를 나타내어 예측성과가 뛰어남을 확인할 수 있었으나 인공신경망과는 그 차이가 통계적으로 유의하지 않았다. 한편, 인공신경망, 다변량 판별분석, 로지스틱 회귀분석 모형 등 기존 부도예측모형의 예측정확도는 통계적으로 의미 있는 차이를 보이지 않는 것으로 나타났다.

5. 결 론

본 연구에서는 인공신경망이 가지고 있는 과대적합 문제를 해결하는 동시에 높은 예측력과 우수한 설명력을 제공하는 것으로 알려진 Support Vector Machine(SVM)을 기업부도예측문제에 적용하였다. SVM은 데이터를 고차원 공간으로 투사시켜 분리경계가 매우 복잡한 문제를 선형판별함수의 사용이 가능한 단순한 문제로 변환시키는 학습기법으로 최

〈표 7〉 SVM, 인공신경망, 다변량 판별분석, 로지스틱 회귀분석의 최고 예측정확도(%) 비교

	SVM	인공신경망	다변량 판별분석	로지스틱 회귀분석
훈련데이터	88.0132	85.2474	78.8079	79.8676
검증데이터	83.0688	82.5397	79.1391	78.3069

〈표 8〉 McNemar 검정 결과

	인공신경망	다변량 판별분석	로지스틱 회귀분석
SVM	1.750[1] (0.186)[2]	4.470(0.035)**	2.857(0.091)*
인공신경망		0.356(0.551)	0.103(0.749)
다변량 판별분석			0.595(0.440)

1) McNemar 통계량 값 2) p-값
*) 유의수준 10%에서 통계적으로 유의함.
**) 유의수준 5%에서 통계적으로 유의함.

근 들어 다양한 분류문제에 적용되고 있다. SVM이 주목 받는 이유는 첫째, 견고한 이론적 근거에 기반하므로 결과 해석이 용이하고, 둘째, 실제 응용에 있어서도 인공신경망 수준의 높은 성과를 나타내며, 셋째, 적은 양의 학습자료만으로도 판별학습을 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 있으며, 볼록 집합(convex set)을 실행가능영역으로 하는 최적화 기법을 사용하기 때문에 유일한 최적해를 구할 수 있다는 점에서 관심을 끌고 있다.

특히, 본 연구에서는 부도예측을 위한 SVM 모형을 구축하는데 있어 가장 중요한 커널함수의 모수 값을 최적으로 선택하기 위해 5-겹 교차타당성과 격자탐색이라는 기법을 사용하였으며, 격자탐색을 통해 선정된 최적의 모수 값을 이용하여 안정성과 예측력이 우수한 부도예측모형을 구현하였다. 그리고 부도예측모형으로서 SVM의 적합성을 평가하기 위하여 예측력이 우수하다고 알려진 인공신경망 모형, 그리고 다변량 판별분석 및 로지스틱 회귀분석 모형과 같은 통계분석모형과 예측성과를 비교하였다. 실증분석 결과, SVM은 향후 부도예측에 있어 인공신경망을 대체할 수 있는 유용한 기법임을 확인할 수 있었다.

한편, 본 연구에서는 RBF 커널을 SVM 모형의 기본 커널함수로 사용하였는데, 분류문제의 종류 및 복잡도에 따라 적합한 커널함수는 달라질 수 있으며, 예측정확도를 높이기 위한 관점에서만 본다면 본 연구에서 활용한 격자탐색 기법이 반드시 이를 보장해 준다고는 볼 수 없다. 따라서 향후 SVM이 다양한 문제에 활발히 적용되기 위해서는 커널함수의 종류 및 모수 선정에 따른 성과 비교 연구가 지속적으로 수행되어야 할 것이다. 또한 본 연구에서는 2차원 평면이라는 제한된 공간에 *LIBSVM*에서 제공하는 도구와 요인분석을 결합하여 SVM의 학습성과 및 예측성과를 시각적으로 표현하였다. 향후 요인의 수를 3개로 확장하여 3차원의 입체적인 형태로 SVM의 학습 및 예측성과를 나타낼 수 있는 도구가 개발된다면 부도예측과 관련하여 보다 의미있는 정보를 제공할 수 있을 것으로 기대한다.

참 고 문 헌

[1] 박정민, 김경재, 한인구, "Support Vector Machine을 이용한 기업부도예측", 「한국경영정보학회 추계학술대회 발표논문집」, 2003, pp.751-758.

[2] 한국은행, 「기업경영분석」, 2003.

[3] 홍태호, 신택수, "부실확률맵과 AHP를 이용한 기업신용평가시스템의 개발", 「한국경영정보학회 추계학술대회 발표논문집」, 2003, pp.719-726.

[4] Altman, E.I., "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy," *The Journal of Finance*, Vol. 23, No.4(1968), pp.589-609.

[5] Altman, E.I., G. Marco and F. Varetto, "Corporate distress diagonis comparisons using linear discriminant analysis and neural networks," *Journal of Banking and Finance*, Vol.18, No.3(1994), pp.505-529.

[6] Barniv, R., A. Agarwal and R. Leach, "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(1997), pp. 177-194.

[7] Basel Committee on Banking Supervision, *Credit Risk Modeling : Current Practices and Applications*, Basel Committee Publications, 1999.

[8] Beaver, W., "Financial ratios as predictors of failure," *Journal of Accounting Research*,

Vol.4(1966), pp.71-102.

[9] Bell, T., G. Ribar and J. Verchio, "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 Sysposium on Auditing Problems*, 1990, pp. 29-58.

[10] Bryant, S.M., "A Case-based Reasoning Approach to Bankruptcy Prediction Modeling," *International Journal of Intelligent Systems in Accounting, Finance and Management*, Vol.6, No.3(1997), pp.195-214.

[11] Burges, C.J.C., "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, Vol.2, No.2(1998), pp.955-974.

[12] Buta, P., "Mining for Financial Knowledge with CBR," *AI Expert*, Vol.9, No.2(1994), pp.34-41.

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

[14] Chen, M.-C. and S.-H. Huang, "Credit Scoring and Rejected Instances Reassigning through Evolutionary Computation Techniques," *Expert Systems with Applications*, Vol.24(2003), pp.433-441.

[15] Cielen, A. and K. Vanhoof, *Bankruptcy prediction using a data envelopment analysis*, Manuscript, Limburg University, Diebenpeek, 1999.

[16] Coakley, J.R. and C.E. Brown, "Artificial Neural Networks in Accounting and Finance : Modeling Issues," *International Journal of Intelligent Systems in Accounting, Finance and Management*, Vol.9, No.2 (2000), pp.119-144.

[17] Coates, P. and L. Fant, "Recognizing financial distress patterns using a neural network tool," *Financial Management*, Vol.22, No.3 (1993), pp.142-155.

[18] Cristianini, N. and J. Shawe-Taylor, *An Introduction to Support Vector Machines*, Cambridge, England : Cambridge University Press, 2000.

[19] Curram, S.P. and J. Mingers, "Neural Networks, Decision Tree Induction and Discriminant Analysis : An Empirical Comparison," *Journal of Operational Research Society*, Vol.45, No.4(1994), pp.440-450.

[20] Davis, R.H., D.B. Edelman and A.J. Gammerman, "Machine Learning Algorithms for Credit-Card Applications," *IMA Journal of Mathematics Applied in Business and Industry*, Vol.4(1992), pp.43-51.

[21] Desai, V.S., J.N. Conway and G.A. Overstreet Jr., "Credit Scoring Models in the Credit Union Environment Using Neural Networks and Genetic Algorithms," *IMA Journal of Mathematics Applied in Business and Industry*, Vol.8(1997), pp.324-346.

[22] Desai, V.S., J.N. Crook and G.A. Overstreet Jr., "A Comparison of Neural Networks and Linear Scoring Models in the Credit Union Environment," *European Journal of Operational Research*, Vol.95(1996), pp.24-37.

[23] Diakoulaki, D., G. Mavrotas and L. Papayannakis, "A multicriteria approach for evaluating the performance of industrial firms," *Omega*, Vol.20, No.4(1992), pp.467-474.

[24] Dimitras, A.I., R. Slowinski, R. Susmaga, and C. Zopounidis, "Business failure prediction using rough sets," *European Journal of Operational Research*, Vol.7, No.3 (1999), pp.263-280.

[25] Dimitras, A.I., S.H. Zanakis and C. Zopounidis, "A Survey of Business Failure with an Emphasis on Prediction Methods and Industrial Applications," *European Journal of Operational Research*, Vol.90, No.3(1996), pp.487-513.

[26] Drucker, H., D. Wu and V.N. Vapnik, "Support vector machines for spam categorization," *IEEE Trans. Neural Networks*, Vol.10, No.5(1999), pp.1048-1054.

[27] Eisenbeis, R.A., "Problems in applying discriminant analysis in credit scoring models," *Journal of Banking and Finance*, Vol.2 (1978), pp.205-219.

[28] Elmer, P.J. and D.M. Borowski, "An expert system approach to financial analysis : the case of S&L bankruptcy," *Financial Management Autumn*, Vol.17, No.3(1988), pp. 66-76.

[29] Emel, A.B., M. Oral, A. Reisman and R. Yolalan, "A Credit Scoring Approach for the Commercial Banking Sector," *Socio-Economic Planning Sciences*, Vol.37(2003), pp. 103-123.

[30] Falbo, P., "Credit-scoring by enlarged discriminant models," *Omega*, Vol.19, No.4 (1991), pp.275-289.

[31] Fan, A. and M. Palaniswami, "A new approach to corporate loan default prediction from financial statements," In Proceedings of the computational finance/forecasting financial markets conference, 2000, London (CD), UK.

[32] Fan, A. and M. Palaniswami, "Selecting bankruptcy predictors using a support vector machine approach," In Proceedings of *the International Joint Conference on Neural Networks*, 2000.

[33] Fanning, K. and K. Cogger, "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 (1994), pp.241-252.

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

[35] Frydman H.E., E.I. Altman and D. Kao, "Introducing Recursive Partitioning for Financial Classification : the case of Financial Distress," *The Journal of Finance*, Vol.40, No.1(1985), pp.269-291.

[36] Geman, S., E. Bienenstock and R. Doursat, R., "Neural Networks and the Bias/Variance Dilemma," *Neural Computation*, Vol.4(1992), pp.1-58.

[37] Gunn, S.R., "Support Vector Machines for Classification and Regression," Technical Report, University of Southampton, 1998.

[38] Häardle, W., R. Moro and D. Schäafer, "Predicting corporate bankruptcy with support vector machines," *Working Slide*, Humboldt University and the German Institute for Economic Research available on, 2003.

[39] Hair, J.F., R.E. Anderson, R.E. Tatham and W.C. Black, *Multivariate Data Analysis with Readings*, Prentice Hall, 1995.

[40] Hearst, M.A., S.T. Dumais, E. Osman, J. Platt and B. Scholkopf, "Support vector ma-

chines," *IEEE Intelligent System*, Vol.13, No.4(1998), pp.18-28.

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

[42] Hsu, C.-W., C.-C. Chang and C.-J. Lin, "A Practical Guide to Support Vector Classification," Technical Report, Department of Computer Science and Information Engineering, National Taiwan University, 2004. Available at http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.

[43] Huang, Z., H. Chen, C.-J. Hsu, W.-H. Chen, and S. Wu, "Credit rating analysis with support vector machine and neural networks : a market comparative study," *Decision Support Systems*, Vol.37(2004), pp. 543-558.

[44] Jensen, H.L., "Using Neural Networks for Credit Scoring," *Managerial Finance*, Vol. 18(1992), pp.15-26.

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

[46] Kim, K.J., "Financial time series forecasting using support vector machines," *Neurocomputing*, Vol.55, No.1-2(2003), pp.307-319.

[47] Lawrence, S., C.L. Giles and A.-C. Tsoi, "Lessons in Neural Network Training : Overfitting May be harder than Expected," In Proceedings of the Fourteenth National Conference on Artificial Intelligence, AAAI-97, AAAI Press, Mento Park, California, 1997, pp.540-545.

[48] Lee, G., T.K. Sung and N. Chang, "Dynamics of Modeling in Data Mining : Interpretive Approach to Bankruptcy Prediction," *Journal of Management Information Systems*, Vol.16(1999), pp.63-85.

[49] Lee, H., H. Jo and I. Han, "Bankruptcy Prediction Using Case-based Reasoning, Neural Networks, and Discriminant Analysis," *Expert Systems With Applications*, Vol.13(1997), pp.97-108.

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

[51] Lee, T.-S., C.-C. Chiu, C.-J. Lu and I.-F. Chen, "Credit Scoring Using Hybrid Neural Discriminant Technique," *Expert Systems with Applications*, Vol.23(2002), pp.245-254.

[52] Lopez, J.A. and M.R. Saidenberg, M.R., "Evaluating credit risk models," *Journal of Banking and Finance*, Vol.24, No.1-2(2000), pp.151-165.

[53] Malhotra, R. and D.K. Malhotra, "Differentiating Between Good Credits and Bad Credits Using Neuro-fuzzy Systems," *European Journal of Operational Research*, Vol. 136, No.2(2002), pp.190-211.

[54] Markham, I.S. and C.T. Ragsdale, "Combining Neural Networks and Statistical Predictions to Solve the Classification Problem in Discriminant Analysis," *Decision Sciences*, Vol.26, No.2(1995), pp.229-242.

[55] Martin, D., "Early Warning of Bank Failure : A Logit Regression Approach," *Journal of Banking and Finance*, Vol.1(1997), pp.249-276.

[56] Moody, J.E., "The Effective Number of Parameters : An Analysis of Generalization and Regularization in Nonlinear Learning

Systems," *NIPS*, Vol.4(1992), pp.847-854.

[57] Odom, M. and R.Sharda, "A neural network model for bankruptcy prediction," In Proceedings of *the International Joint Conference on Neural Networks*, 1990, II-163-II-168.

[58] Ohlson, J.A., "Financial ratios and probabilistic prediction of bankruptcy," *Journal of Accounting Research*, Vol.18, No.1(1980), pp.109-131.

[59] Oral, M. and R. Yolalan, "An empirical study on measuring operating efficiency and profitability of bank branches," *European Journal of Operational Research*, Vol.46(1997), pp.282-294.

[60] Park, C.-S. and I. Han, I., "A case-based reasoning with the feature weights derived by analytic hierarchy process for bankruptcy prediction," *Expert Systems with Applications*, Vol.23, No.1(2002), pp.255-264.

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

[62] Peel, M.J., D.A. Peel and P.F. Pope, "Predicting corporate failure-some results for the UK corporate sector," *Omega*, Vol.14, No.1(1986), pp.5-12.

[63] Reichert, A.K., C.C. Cho and G.M. Wagner, "An Examination of the Conceptual Issues Involved in Developing Credit-Scoring Models," *Journal of Business and Economic Statistics*, Vol.1(1983), pp.101-114.

[64] Roy, B., "The outranking approach and the foundations of ELECTRE methods," *Theory and Decision*, Vol.31(1991), pp.49-73.

[65] Sarle, W.S., "Stopped Training and Other Remedies for Overfitting," In Proceedings of *the 27th Symposium on the Interface of Computing Science and Statistics*, 1995, pp. 352-360.

[66] Smith, M., *Neural Networks for Statistical Modeling*, NY : Van Nostrand Reinhold, 1993.

[67] Srinivasan, V. and B. Ruparel, "CGX : an expert support system for credit granting," *European Journal of Operational Research*, Vol.45(1990), pp.293-308.

[68] Srinivasan, V. and Y.H. Kim, "Designing expert financial systems : a case study of corporate credit management," *Financial Management*, Vol.5(1998), pp.32-43.

[69] Tam, K.Y. and M.Y. Kiang, "Managerial Applications of Neural Networks : the Case of Bank Failure Predictions," *Management Science*, Vol.38, No.7(1992), pp.926-947.

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

[71] Troutt, M.D., A. Rai and A. Zhang, "The potential use of DEA for credit applicant acceptance systems," *Computers and Operations Research*, Vol.23, No.4(1996), pp.405-408.

[72] Van Gestel, T., B. Baesens, J. Suykens, M. Espinoza, D.E. Baestaens, J. Vanthienen, and B. De Moor, "Bankruptcy prediction with least squares support vector machine classifiers," In Proceedings of *the IEEE international conference on computational intelligence for financial engineering*, Hong Kong, 2003, pp.1-8.

[73] Vapnik, V., *Statistical Learning Theory*, Springer, New York, 1998.

[74] Viaene, S., R.A. Derrig, B. Baesens and G. Dedene, "A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection," *The Journal of Risk and Insurance*, Vol.69, No.3 (2002), pp.373-421.

[75] Weigend, A., "On overfitting and the effective number of hidden units," In Proceedings of *the 1993 Connectionist Models Summer School*, 1994, pp.335-342.

[76] West, D., "Neural Network Credit Scoring Models," *Computers & Operations Research*, Vol.27(2000), pp.1131-1152.

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

[78] Zhang, G.P., "Neural Networks for Classification : A Survey," *IEEE Transactions on Systems, Man, and Cybernetics- Part C : Applications and Reviews*, Vol.30, No.4 (2000), pp.451-462.

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

[80] Zopounidis, C. and M. Doumpos, "Developing a multicriteria decision support system for financial classification problems : the Finclas system," *Optimization Methods and Software*, Vol.8(1998), pp.277-304.