기계학습 방법을 이용한 기업부도의 예측[†]

박동준¹ · 윤예분² · 윤민³

¹부경대학교 통계학과 · ²간사이대학교 수리계획공학연구소

접수 2012년 4월 30일, 수정 2012년 5월 17일, 게재확정 2012년 5월 25일

요약

기업도산에 대한 분석과 관리는 기업의 성과와 성장능력을 평가하는 재무관리 분야에서 중요하게 인식되어 왔다. 결국, 기업도산 예측에 대한 효과적인 모형이 필요하게 된다. 본 논문은 서포트 벡터 기계의 한 종류인 토탈 여유도 알고리즘을 이용하여 기업도산 예측을 위하여 새로운 접근 방법을 서술한다. 몇 개의 실제 자료를 통하여 제안한 방법들이 도산 위험의 평가에서 기존의 방법들보다 개선됨을 확인할 수 있었다.

주요용어: 도산기업예측, 서포트 벡터 기계, 소프트 여유도 알고리즘, 토탈 여유도 알고리즘.

1. 서론

건전한 재무투자자의 관점에서 적절한 기업도산의 위험을 평가하는 것은 중요하다. 본 논문의 목적은 서포트 벡터 기계 (Support Vector Machines; SVMs)를 기업도산의 예측에 적용하는 것이다. 특히, 저자들이 제안한 토탈 여유도 (total margin) 서포트 벡터 기계가 기업도산 예측에 있어서 좋은 수행능력을 제공한다는 것을 보일 것이다.

서포트 벡터 기계는 원래 두 범주를 가지는 패턴 분류문제에서 개발되었다 (Bartlett과 Mangasarian, 1992; Cherkassky와 Mulier, 1998; Cristianini와 Shawe-Taylor, 2000; Vapnik, 1998). 이 방법의 주된 아이디어는 원래의 자료집합을 어떤 비선형 함수에 의하여 변환된 특징공간에서 한 범주의 자료집합을 커널에 근거한 최대 여유도 (maximal margin) 선형 분류기 (classifier)를 이용하여 다른 범주의 자료집합을 가능한 멀리 분리하는 것이다 (박해정, 2011; Glover, 1990; Mangasarian, 1968). 만약 자료가 잡음에 의해 오염되어 있다면, 완전분리는 일반화에 있어서 좋은 수행력을 얻을 수 없을 것이다. 이러한 어려움을 극복하기 위하여, 오분류된 자료 점들의 여유성 (slackness)을 고려한 소프트 여유도 (soft margin) 알고리즘이 제안되었다 (임주열 등, 2010; Bartlett과 Shawe-Taylor, 1999; Cortes와 Vapnik, 1995). 이 후에 본 저자들은 분류된 자료 점들이 얼마나 정확하게 분류가 되었는가의 척도로서 잉여변수 (surplus variable)를 소개하였다. 여유도를 최대화 하기 위하여 토탈 여유도 알고리즘은 공식화되었는데, 이는 여유변수들을 최소화하고 동시에 잉여변수들을 최대화하는 것이다 (Yoon 등, 2004; Nakayama 등, 2009).

다음 절에서 기업부도 예측을 위한 기계학습 방법으로 서포트 벡터 기계를 소개하고, 3절에서 수치예제들을 통하여 토탈 여유도 알고리즘의 수행 능력을 평가 할 것이다. 마지막으로 기계학습 방법을 이용한 성능의 효율성을 논의하고 추후의 문제점에 대한 제안을 할 것이다.

[†] 이 논문은 2011학년도 부경대학교의 지원을 받아 수행된 연구임 (PK-2011-36).

¹ (608-737) 부산시 남구 대연3동 599-1번지, 부경대학교 통계학과, 교수.

² (564-8680) 일본 오사카 수이타-시 3-3-35번지, 간사이대학교 수리계획공학연구소, 부교수.

³ 교신저자: (608-737) 부산시 남구 대연3동 599-1번지, 부경대학교 통계학과, 조교수.

E-mail: myoon@pknu.ac.kr

2. 서포트 벡터 기계

두 집합을 가지는 훈련집합을

$$S = ((\mathbf{x}_1, y_1), \dots, (\mathbf{x}_\ell, y_\ell)) \in (X \times \{-1, 1\})^\ell,$$

와 같이 정의하자. 여기서 $X (\subset \mathbb{R}^n)$ 는 입력공간이다. 서포트 벡터 기계의 주요 아이디어는 어떤 비선형 함수를 이용하여 주어진 자료집합을 특징공간으로 변환을 한 후에, 이 공간상에서 커널 (kernel)을 사용한 최대 여유도를 가지는 선형 분류함수를 찾는 것이다. 여기서 여유도는 훈련집합의 점들과 판별함수에 의하여 얻어진 초평면 $f(\mathbf{x}) = 0$ 사이의 최소 거리로 정의된다. 특징공간상에서 판별함수는

$$f(\Phi(\mathbf{x})) = \mathbf{w}^T \Phi(\mathbf{x}) + w_0, \quad (2.1)$$

와 같이 정의되고 각 패턴 $\mathbf{x}_i$ 는 $f(\Phi(\mathbf{x}_i))$ 의 부호에 의하여 분류된다 (석경하, 2010; Glover, 1990; Mangasarian, 1968).

실질적으로, 자료들이 잡음 (noise)에 의하여 오염되어 있다면, 비록 고차원의 특징공간에서도 선형으로 분리할 수 없을 것이다 (김철웅과 윤민, 2003). 또한 완전분리가 되는 경우는 과적합 (over-fitting)의 문제때문에 예측능력에서 좋지 않은 결과를 보인다 (Shawe-Taylor 등, 1998). 이러한 어려움을 극복하기 위하여, 그림 2.1에서 보는 바와 같이 오분류된 점들에 대한 허용가능한 오차로서 여유변수가 정의된다.

그림 2.1 여유변수들

그림 2.2 잉여변수들

$\mathbf{w}$ 와 w_0 를 결정하기 위하여 소프트 여유도 알고리즘은 아래와 같이 정의된다 (Cortes와 Vapnik, 1995; Vapnik, 1998):

$$\begin{aligned} & \underset{\mathbf{w}, w_0, \xi_i}{\text{minimize}} && \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^{\ell} \xi_i \\ & \text{subject to} && y_i \left(\mathbf{w}^T \Phi(\mathbf{x}_i) + w_0 \right) \geq 1 - \xi_i, \\ & && \xi_i \geq 0, \quad i = 1, \dots, \ell. \end{aligned} \quad (2.2)$$

1차계획 문제 식 (2.2)를 라그랑즈 쌍대문제 식 (2.3)으로 바꾸면

$$\begin{aligned}
 & \underset{\alpha_i}{\text{maximize}} && \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \\
 & \text{subject to} && \sum_{i=1}^{\ell} y_i \alpha_i = 0, \\
 & && 0 \leq \alpha_i \leq C, \quad i = 1, \dots, \ell,
 \end{aligned} \tag{2.3}$$

와 같이 주어지고, 여기서 C 는 여유변수의 모수이다.

위의 수식에서, 적절한 커널 $\langle \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j) \rangle \equiv K(\mathbf{x}_i, \mathbf{x}_j)$ 을 사용하는 것이 일반적이다. 이러한 커널들의 전형적인 예들은 Schölkopf와 Smola (2002)에서 소개한 바와 같이 q 차 다항함수 커널과 본 논문에서 수치예제들의 실험에서 사용하는 가우시안 (Gaussian)함수 커널등이 있다 (황창하와 신사임, 2010; Haykin, 1998).

$$\begin{aligned}
 K(\mathbf{x}, \mathbf{x}') &= (\mathbf{x}^T \mathbf{x}' + 1)^q, \\
 K(\mathbf{x}, \mathbf{x}') &= \exp \left(- \sum_{i=1}^n \frac{(\mathbf{x}_i - \mathbf{x}'_i)^2}{2\sigma_i^2} \right).
 \end{aligned}$$

더우기, 그림 2.2 에서 보는 것과 같이 분리 초평면에 의해 정분류된 자료들이 얼마나 잘 분류가 되었는가의 척도로써 잉여변수 η_i 가 정의된다. 여유도를 최대화 함으로써, 즉, 여유변수들을 최소화 하고 잉여변수들을 최대화 하는 두 가지를 동시에 고려하는, 토탈 여유도 알고리즘을 아래와 같이 얻게된다 (Yoon 등, 2004):

$$\begin{aligned}
 & \underset{\mathbf{w}, w_0, \xi_i, \eta_i}{\text{minimize}} && \frac{1}{2} \mathbf{w}^T \mathbf{w} + C_1 \sum_{i=1}^{\ell} \xi_i - C_2 \sum_{i=1}^{\ell} \eta_i \\
 & \text{subject to} && y_i (\mathbf{w}^T \Phi(\mathbf{x}_i) + w_0) \geq 1 - \xi_i + \eta_i, \\
 & && \xi_i \geq 0, \quad \eta_i \geq 0, \quad i = 1, \dots, \ell.
 \end{aligned} \tag{2.4}$$

식 (2.4)에 대응되는 라그랑즈 쌍대문제는

$$\begin{aligned}
 & \underset{\alpha_i}{\text{maximize}} && \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j \langle \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j) \rangle \\
 & \text{subject to} && \sum_{i=1}^{\ell} y_i \alpha_i = 0, \\
 & && C_2 \leq \alpha_i \leq C_1, \quad i = 1, \dots, \ell
 \end{aligned} \tag{2.5}$$

식 (2.5)로 주어지고, 여기서 C_1 과 C_2 는 각각 여유변수와 잉여변수의 모수이다. 여기서, $C_1 > C_2$ 의 조건은 ξ_i 와 η_i 들 중에서 적어도 하나는 0임이 보장되어야 하기 때문에 C_1 과 C_2 는 $C_1 > C_2$ 인 경우를 선택해야한다.

다음 절에서 수치 예제를 통한 토탈 여유도 알고리즘의 수행능력을 평가할 것이다.

3. 기업부도의 예측

이 절에서 2절에서 소개한 소프트 여유도 알고리즘과 토탈 여유도 알고리즘을 실제 자료를 기초로 기업도산을 예측한다.

예제 2.1

전통적으로 패턴 분류 문제에 공통적으로 사용하는 피셔 선형 판별분석과 소프트 여유도 알고리즘, 토탈 여유도 알고리즘의 분류율을 비교한다.

Tokyo Shoko Research Ltd. (TSR)에서 제공된 자료를 사용하여 표 3.1에서 나타난 조건들을 만족하는 도산기업들과 건전기업들을 예측하였다. 표 3.1에서 보는 바와 같이 다섯 개의 재무지표를 입력변수로 사용하였다.

표 3.1 자료 집합의 요약

범 주	건설 산업에 속하는 기업 (일본의 총무성의 기준에 따르는 분류)	
규 모	중,소규모의 크기	
주 식	동경 주식거래 시장에 상장되지 않음	
기 간	1999.04 - 2000.03	
기업의 수	건전 기업의 수	24480
	도산 기업의 수	375
재무 지표	1. 자본금/총자산 2. 총부채/총자산 3. 당기 순이익/매출액 4. 이자비용/매출액 5. 매출액/총자산	

수치예제를 위한 자료집합들은 건전기업과 도산기업에서 각각 표 3.2에서 나타내었다.

표 3.2 자료 집합

	훈련집합		테스트 집합	
	도산기업 수	건전기업 수	도산기업 수	건전기업 수
자료집합 1	50	300	100	100
자료집합 2	100	400	100	100

소프트 여유도 알고리즘에서의 모수는 $C = 1, 10, 100, 1000$ 로 선택하였고, 토탈 여유도 알고리즘에서는 위의 값들과 조합으로 $C_1 = 1, 10, 100, 1000$ 와 $C_2 = 0.01, 0.1, 1, 10, 100, 1000$ 를 선택하였다. 가우시안 커널함수의 반지름 σ_i 는 i 번째 재무지표의 표준편차로 사용하였다.

표 3.3 피셔의 선형 판별분석을 이용한 분류율

	훈련집합		테스트 집합	
	도산기업	건전기업	도산기업	건전기업
자료집합 1	47.7	84.2	76.5	71.9
자료집합 2	43.5	82.9	69.2	74.4

표 3.3은 피셔의 선형 판별분석을 이용한 분류율을 나타내었는데, 도산기업과 건전기업 모두가 판별율이 나쁘지 않게 나타났다. 그러나 판별분석은 두 모집단의 분산-공분산 행렬이 같은 다변량 정규분포를 가진다는 가정을 만족해야 한다. 그러나 Eisenbis (1978)는 판별분석을 이용할 때 다변량 정규분포의 가정이 일반적인 경영 경제 문제에 만족하지 않음을 지적하고 비판적인 평가를 하였고 본 논문에서와 같이 기업도산 문제의 예측에는 위의 가정을 만족하는 경우는 거의 없다.

한편 표 3.4에서 소프트 여유도 알고리즘을 이용한 분류율의 결과를 나타냈는데, 모수 C 가 작은 값들에 대하여 건전기업의 판별율에 비하여 도산기업의 분류율이 상대적으로 아주 낮게 나타났다.

표 3.4 소프트 여유도 알고리즘을 이용한 분류율
(1) 도산 기업들
(a) 자료 집합 1

C	1	10	100	1000	10^8
훈련집합	17.4	45.2	73.2	92.6	100
테스트 집합	1.8	15.5	23.0	26.9	30.4

(b) 자료집합 2

C	1	10	100	1000	10^4
훈련집합	23.5	51.1	69.3	87.6	100
테스트 집합	9.9	25.9	25.2	34.1	38.4

(2) 건전 기업들
(a) 자료 집합 1

C	1	10	100	1000	10^8
훈련집합	99.9	99.4	99.47	99.5	100
테스트 집합	99.8	96.4	90.6	84.9	85.7

(b) 자료 집합 2

C	1	10	100	1000	10^4
훈련집합	99.3	98.2	98.6	98.8	100
테스트 집합	97.5	91.9	90.0	83.2	81.9

일반적으로, 소프트 여유도 알고리즘은 C 의 값이 충분히 큰 경우에 대하여 훈련 자료집합을 완전 분리 하려는 경향이 있다. 이것은 토탈 여유도 알고리즘의 경우도 마찬가지이다. 이런 성질때문에 C 값이 커지는 경우에 훈련집합의 자료에 대하여 분류율이 테스트 집합에 대한 분류율보다 상대적으로 좋게 나타난다. 그러나 이러한 현상들은 본 논문에서 사용된 기업 도산에 대한 예측의 결과에서는 관측되지 않았다. 이는 도산 기업의 수가 건전기업의 수와 비교해서 상대적으로 너무 작기 때문에, 상대적으로 C 값이 작은 경우에 대하여 대부분의 도산기업은 오분류되고, 테스트 집합의 분류율의 결과는 분류율이 좋지않게 된다. 따라서 도산기업들에 대한 분류율을 개선하기 위하여 상대적으로 큰 C 값을 선택하였다. 그러나, 저자들의 기대와는 반대로 상대적으로 큰 C 값에서 조차 도산기업에 대한 분류율은 기껏해야 40% 정도였다. 결국, 소프트 여유도 알고리즘에서 도산기업들을 잘 예측하기에 적절한 C 값을 선택하기는 어렵다. 훈련 자료집합에서 여유변수와 잉여변수들을 함께 고려하는 토탈 여유도 알고리즘은 이러한 어려움을 극복하는 것으로 보인다. 토탈 여유도 알고리즘을 이용한 결과들은 표 3.5에서 나타나는 것과 같이 C_1 의 값을 고정하고 C_2 의 값이 커짐에 따라서 도산기업에 대한 분류율이 더 나아지게 된다는 것을 알 수 있다. 소프트 여유도 알고리즘에서 $C = 1$ 을 가지는 것과 동일하게 $C_1 = 1$ 일 때, 잉여변수의 모수 $C_2 (= 0.1)$ 를 고려하면 도산영역이 더 커지게된다. 결국, 표 3.5에서 분류율은 표 3.4에서 보다는 나은 결과를 얻는다. 더우기, 이 결과들에서 모수 C_1 의 임의의 값에 대하여 C_2 의 값이 충분히 크다면, 도산 기업들에 대한 피셔의 선형 판별분석의 분류율보다 토탈 여유도 알고리즘이 더 나은 분류율을 보인다.

표 3.5 토탈 여유도 알고리즘을 이용한 분류율
(1) 도산 기업들
(a) 자료집합 1

	$C_1 = 1$		$C_1 = 10$		$C_1 = 100$		$C_1 = 1000$			
C_2	0.01	0.1	0.1	1	0.1	1	10	1	10	100
훈련집합	18.6	85.0	62.0	95.6	95.0	100	93.8	100	100	94.4
테스트 집합	2.7	74.5	30.9	89.0	53.9	82.0	89.6	76.1	88.5	88.6

(b) 자료집합 2

	$C_1 = 1$		$C_1 = 10$		$C_1 = 100$		$C_1 = 1000$			
C_2	0.01	0.1	0.1	1	0.1	1	10	1	10	100
훈련집합	30.1	81.6	63.3	96.0	95.2	99.7	95.4	100	100	95.6
테스트 집합	10.6	71.5	39.3	92.0	60.5	80.9	94.4	79.2	86.7	91.8

(2) 건전 기업들
(a) 자료집합 1

	$C_1 = 1$		$C_1 = 10$		$C_1 = 100$			$C_1 = 1000$		
C_2	0.01	0.1	0.1	1	0.1	1	10	1	10	100
훈련집합	99.9	69.9	98.5	49.6	77.4	49.7	49.3	48.3	41.3	50.9
테스트 집합	99.4	68.0	89.3	46.4	69.4	47.1	48.5	46.1	37.2	48.4

(b) 자료집합 2

	$C_1 = 1$		$C_1 = 10$		$C_1 = 100$		$C_1 = 1000$			
C_2	0.01	0.1	0.1	1	0.1	1	10	1	10	100
훈련집합	98.7	71.2	96.2	43.6	75.1	46.6	40.1	46.3	34.9	44.4
테스트 집합	96.5	66.4	88.6	42.9	69.2	43.8	36.6	47.4	33.0	42.3

예제 2.2

두 번째 자료는 한국금융협회에서 기업의 부도에 관한 자료로서 총 3964개 기업들에서 재무구조를 나타내는 9개의 변수들에 대하여 얻어진 자료이다. 이 자료들에서 건전기업을 수는 3644개이고 부도 기업은 319개였다. 각 변수들에 관한 특징은 아래의 표 3.6에 표현하였으며 아홉 개의 재무지표들을 입력변수로 사용하였다.

표 3.6 기업부도 자료의 요약

	반응변수	독립변수	관측치
기업의 자산 건전성 특징들	정상기업	유동성 비율	3963
		부채 비율	
		차입금의존 비율	
	부도기업	단기 차입금의존 비율	
		현금유동성/총부채	
		현금유동성/대출비율	
		대출비용/EBIT	
		이자비용/매출	
		이자보상비율	

여기서 EBIT (earning before interest tex)는 이자비용과 세금비용을 지불하기 전의 수익을 나타낸다. 분석을 위하여 전체 자료로부터 가우시안 함수 커널에서 $\sigma_i = 0.25$ 로 고정하고 자료에서 훈련집합과 테스트 집합을 각각 7:3과 8:2로 추출하여 얻어진 결과가 표 3.7과 표 3.8에 각각 나타나 있다. 이들 각각의 경우에 로지스틱 회귀분석을 이용한 결과 정분류율이 각각 67.349%, 67.994%의 정분류율을 얻었다. 표 3.7에서 보여진 바와 같이 $C_1 = 1.0$ 과 $C_2 = 0.50$ 의 경우에 분류율이 69.329%

로 얻어져 가장 좋은 결과를 얻었다. 전체적으로 토탈 여유도 알고리즘을 이용한 경우에 소프트 여유도 알고리즘을 이용한 경우보다 다소 나은 분류율을 얻을 수 있었고 마찬가지로 표 3.8의 $C_1 = 1.0$ 과 $C_2 = 0.50$ 인 경우에 분류율이 70.466%로서 가장 높았다. 이 사실로 보아 동일한 자료에서 훈련집합과 테스트 집합을 나누고 비율에 따라 최적의 모수가 바뀌지 않음을 암시한다. 표 3.7과 표 3.8을 보면 토탈 여유도 알고리즘을 이용한 경우가 전체적으로 소프트 여유도 알고리즘을 사용한 경우보다 다소 오분류율이 감소됨을 알 수 있었다.

표 3.7 가우시안함수 커널을 이용한 소프트 여유도 알고리즘과 토탈 여유도 알고리즘의 분류율의 비교 (7:3)

C_1	소프트 여유도 알고리즘		토탈 여유도 알고리즘	
			C_2	
1.0	68.844		0.05	68.766
			0.10	68.875
			0.50	69.329
5.0	68.172		0.10	68.219
			0.50	68.329
			1.00	66.829
10.0	68.000		0.10	68.141
			0.50	68.532
			1.00	66.704

표 3.8 가우시안함수 커널을 이용한 소프트 여유도 알고리즘과 토탈 여유도 알고리즘의 분류율의 비교 (8:2)

C_1	소프트 여유도 알고리즘		토탈 여유도 알고리즘	
			C_2	
1.0	68.977		0.05	69.000
			0.10	68.977
			0.50	70.466
5.0	67.745		0.10	67.721
			0.50	68.605
			1.00	67.884
10.0	67.489		0.10	67.489
			0.50	68.047
			1.00	67.187

4. 결론

본 논문에서는 보다 정확한 기업도산의 예측을 위하여 토탈 여유도 알고리즘을 이용한 방법을 제안하였다. 제안된 방법의 성능을 평가하기 위하여, 두 종류의 예제에서 제공되는 도산기업과 건전기업의 분류율을 비교하였다. 그리고 기업도산의 예측에 있어서 토탈 여유도 알고리즘의 효율성을 보였다. 본 논문에서 사용한 알고리즘은 C_1 과 C_2 의 값에 의존적이기는 하지만 예제에서 보여진 정분류율에서 소프트 여유도 알고리즘보다 다소 나은 수행능력을 보인다는 것을 확인할 수 있었다. 다만 최적은 C_1 과 C_2 를 찾기 위한 노력은 하나의 과제로 보인다. 또한 본 분석에서 사용한 가우시안함수 커널에 대한 최적의 반경 σ_i 를 찾는 것도 또한 중요한 문제라고 생각된다. 그리고 첫번째 예제에서는 일년간의 재무정보를 기초하여 부도율의 예측을 수행하였으나, 추후에 다년간의 기간을 통한 예측과 재무 지표들의 선택에 관해서 좀 더 조사해야할 필요가 있다고 생각된다.

참고문헌

- 김철웅, 윤민 (2003). 서포트 벡터 기계에서 잡음 영향의 효과적 조절. <응용통계연구>, **16**, 261-271.
- 박혜정 (2011). 온라인 서포트벡터기계를 이용한 온라인 비정상 사건탐지. <한국데이터정보과학회지>, **22**, 207-215.
- 석경하 (2010). 최소제곱 서포트벡터기계 형태의 준지도 분류. <한국데이터정보과학회지>, **21**, 461-470.
- 임주열, 백장선, 김민수 (2010). 서포트벡터머신과 정칙화판별함수를 이용한 비디오 문자인식의 분류성능개선. <한국데이터정보과학회지>, **21**, 689-697.
- 황창하, 신사임 (2010). 커널기계 기법을 이용한 일반화 이분산자기회귀모형 추정. <한국데이터정보과학회지>, **21**, 419-425.
- Bartlett, P. and Shawe-Taylor, J. (1999). Generalization performance of support vector machines on other pattern classifiers. In *Advances in Kernel Methods-Support Vector learning*, edited by Schölkopf, B., Burges, C. J. C., and Smola, A., The MIT Press, London, 43-54.
- Benenett, K. P. and Mangasarian, O. L. (1992). Robust linear programming discrimination of two linearly inseparable sets. *Optimization Methods and Software*, **1**, 23-24.
- Cherkassky, V. and Mulier, F. (1998). *Learning from data concepts, theory, and methods*, John Wiley & Sons, New York.
- Cristianini, N. and Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*, Cambridge University Press, Cambridge.
- Cortes, C. and Vapnik, V. (1995). Support vector networks. *Machine Learning*, **20**, 273-297.
- Eisenbis, R. A. (1978). Problem in applying discriminant analysis in credit scoring models. *Journal of Banking and Finance*, **2**, 205-219.
- Glover, F. (1990). Improved linear programming models for discriminant analysis. *Decision Sciences*, **21**, 771-785.
- Haykin, S. (1998). *Neural networks a comprehensive foundation*, 2nd ed., Prentice Hall, New Jersey.
- Mangasarian, O. L. (1968). Multi surface method of pattern separation. *IEEE Transaction on Information Theory*, **14**, 801-807.
- Nakayama, H., Yun, Y. B. and Yoon, M. (2009) *Sequential approximate multiobjective optimization using computational intelligence*, Springer, Heidelberg.
- Schölkopf, B. and Smola, A. J. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*, The MIT Press, London.
- Shawe-Taylor, J., Bartlett, P. L., Williamson, R. C. and Anthont, M. (1998). Structural risk minimization over data-dependent hierarchies. *IEEE Transaction on Information Theory*, **44**, 1926-1940.
- Smola, A. J. and Schölkopf, B. (1998). *A tutorial on support vector regression*, NeuroCOLT2 Technical Report, NeuroCOLT, London.
- Yoon, M., Yun, Y. B. and Nakayama, H. (2004). Total margin algorithms in support vector machines. *IEICE Transactions on Information and Systems*, **87**, 1223-1230.
- Vapnik, V. N. (1998). *Statistical learning theory*, John Wiley & Sons, New York.

Prediction of bankruptcy data using machine learning techniques[†]

Dongjoon Park¹ · Yeboon Yun² · Min Yoon³

¹³Department of Statistics, Pukyong National University

²Faculty of Environmental and Urban Engineering, Kansai University

Received 30 April 2012, revised 17 May 2012, accepted 25 May 2012

Abstract

The analysis and management of business failure has been recognized to be important in the area of financial management in the evaluation of firms' performance and the assessment of their viability. To this end, effective failure-prediction models are needed. This paper describes a new approach to prediction of business failure using the total margin algorithm which is a kind of support vector machine. It will be shown that the proposed method can evaluate the risk of failure better than existing methods through some real data.

Keywords: Prediction of failure company, soft margin algorithm, support vector machines, total margin algorithm.

[†] This work was supported by the Pukyong National University Research Fund in 2011(PK-2011-36).

¹ Professor, Department of Statistics, Pukyong National University, Busan 608-737, Korea.

² Associate professor, Faculty of Environmental and Urban Engineering, Kansai University, Osaka 3-3-35, Japan.

³ Corresponding author: Assistant professor, Department of Statistics, Pukyong National University, Busan 608-737, Korea. E-mail: myoon@pknu.ac.kr