*Journal of the Korean Data & Information Science Society*
*2013,* **24**(6), 1113–1125

http://dx.doi.org/10.7465/jkdi.2013.24.6.1113
한국데이터정보과학회지

# 앙상블 SVM 모형을 이용한 기업 부도 예측[†]

최하나[1] · 임동훈[2]

[12]경상대학교 정보통계학과

접수 2013년 6월 24일, 수정 2013년 7월 10일, 게재확정 2013년 7월 26일

**요 약**

기업의 부도를 예측하는 것은 회계나 재무 분야에서 중요한 연구주제이다. 지금까지 기업 부도 예측을 위해 여러 가지 데이터마이닝 기법들이 적용되었으나 주로 단일 모형을 사용함으로서 복잡한 분류 문제에의 적용에 한계를 갖고 있었다. 본 논문에서는 최근에 각광받고 있는 SVM (support vector machine) 모형들을 결합한 앙상블 SVM 모형 (ensemble SVM model)을 부도예측에 사용하고자 한다. 제안된 앙상블 모형은 $v$-조각 교차 타당성 ($v$-fold cross-validation)에 의해 얻어진 여러 가지 모형 중에서 성능이 좋은 상위 $k$개의 단일 모형으로 구성하고 과반수 투표 방식 (majority voting)을 사용하여 미지의 클래스를 분류한다. 본 논문에서 제안된 앙상블 SVM 모형의 성능을 평가하기 위해 실제 기업의 재무비율 자료와 모의실험자료를 가지고 실험하였고, 실험결과 제안된 앙상블 모형이 여러 가지 평가척도 하에서 단일 SVM 모형들보다 좋은 성능을 보임을 알 수 있었다.

주요용어: 교차 타당성, 부도예측, 서포트 벡터 머신, 앙상블 서포트 벡터 머신, 재무비율.

## 1. 서론

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 (Altman, 1983; Bellovary 등, 2007).

지금까지 많은 연구자들에 의해 부도예측을 위한 연구가 꾸준히 진행되어왔다. 부도예측을 위한 통계적 방법으로는 다변량 판별분석 (multivariate discriminant analysis; Altman 1968, 1983), 로짓 분석 (logit analysis; Ohlson, 1980), 프로빗 분석 (probit analysis; Zmijewski, 1984)과 같은 모수적 방법 (parametric method)과 인공신경망 (artificial neural network; Zhang 등, 1999; Anandarajan 등, 2004), SVM (support vector machine; Min 과 Lee, 2005; Shin 등, 2005; Park 등, 2012)과 같은 비모수적 방법 (non-parametric method)이 사용되어 왔다. 모수적 방법은 입력변수에 대한 제약조건을 갖고 있고 이런 제약조건을 만족하는 경우 성능이 뛰어나는 반면, 비모수적 방법은 독립변수에 대해 엄격한 제한을 두지 않으면서 자료에 대한 풍부한 설명력을 갖고 있어 모수적인 방법에 비해 많이 사용되고 있다. 특히, SVM은 Vapnik (1995, 1998)에 의해 제안된 통계적 학습이론으로 인공신경망에서 지적된

[†] 이 논문은 2013년도 정부 (교육부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임 (No.2011-0010089)
[1] (660-701) 경남 진주시 가좌동 900번지, 경상대학교 정보통계학과, 석사학생.
[2] 교신저자: (660-701) 경남 진주시 가좌동 900번지, 경상대학교 정보통계학과, 교수 및 RINS.
E-mail: dhlim@gnu.ac.kr

과대적합 (over-fitting)문제와 지역적 최적해 (local optima) 존재를 해결할 수 있을 뿐만 아니라 높은 예측력을 제공함으로서 최근 각광 받고 있는 기계학습이다.

SVM의 성능은 분류모형을 결정하는 커널함수 (kernel function)를 포함하여 여러 가지 모수에 의해 영향을 많이 받는다. 따라서, 본 논문에서는 SVM 성능이 커널함수에 의존한다는 사실에 기인하여 하나의 커널함수를 갖는 SVM 모형을 단일 SVM 모형으로 간주하여 여러 개의 단일 SVM 모형을 결합하여 만든 앙상블 SVM 모형 (ensemble SVM model)을 부도예측에 사용하고자 한다.

일반적으로, 앙상블 모형이 단일모형보다 성능이 우수하지만 앙상블 모형을 구성하는 단일모형들 간에 높은 상관이 있는 경우, 단일모형보다 성능이 떨어지는 경우도 흔히 발생한다 (Eom 등, 2008; Kim과 Kang, 2012). 본 논문에서는 이를 해결하기 위해 $v$-조각 교차 타당성 ($v$-fold cross-validation)을 사용하여 좋은 성능을 갖는 상위 $k$개의 단일 모형으로 앙상블 모형을 구성하고 과반수 투표 방식 (majority voting)을 사용하여 미지의 클래스를 분류하고자 한다.

본 논문에서 제안된 앙상블 SVM 모형의 성능을 평가하기 위해 실제 기업의 재무비율 자료와 모의실험 자료를 가지고 분석하고자 한다.

본 논문은 다음과 같이 구성되어 있다. 제 2 절에서는 SVM에 대해 간략하게 살펴보고 제 3 절에서는 교차 타당성을 사용한 앙상블 SVM 모형과 평가 척도에 대해 살펴보고자 한다. 제 4 절에서는 실제 기업의 재무비율 자료와 모의실험을 통해 성능을 비교 평가하고 제 5 절에서 결론을 맺고자 한다.

## 2. SVM 모형

SVM을 이용한 분류모형을 설명하기 위해 다음의 학습자료 (learning dataset) $D$가 주어졌다고 가정하자.

$$D = \{(x_i, y_i)|x_i \in R^p,\ y_i \in \{-1, +1\}\}_{i=1}^{n}.$$

여기서 $y_i$는 $p$차원 벡터 $x_i$가 속하는 클래스를 나타내는 값으로 $+1$ 혹은 -1이다. SVM에서 $x_i$를 두 클래스로 분류하기 위한 최적 분리 초평면 (optimal separating hyperplane)은 각 클래스에 속하는 점들 중에서 서포트 벡터 (support vector)를 지나는 두 개의 평행인 초평면들 사이의 거리 즉, 마진 (margin)을 최대로 함으로서 결정된다. 따라서, 임의의 선형 초평면을 다음과 같이 표현하자.

$$w^T x + b = 0.$$

여기서 벡터 $w$는 초평면에 수직인 정규벡터 (normal vector)이고 $b$는 원점으로부터 거리 (offset)이다. 그리고 두 개의 평행인 초평면은 다음의 방정식에 의해 표현할 수 있다.

$$w^T x + b = +1,$$
$$w^T x + b = -1.$$

학습 자료가 선형분리 가능하면 두 개의 초평면 사이의 거리는 $2/||w||$이므로 구하는 초평면은 다음과 같은 조건 하에서 $||w||$을 최소화함으로서 얻을 수 있다.

$$y_i = +1\text{에 대하여 } w^T x + b \geq +1, \tag{2.1}$$

$$y_i = -1\text{에 대하여 } w^T x + b \leq -1. \tag{2.2}$$

식 (2.1)과 (2.2)을 다음과 같이 하나의 부등식으로 표현할 수 있다.

$$y_i(w^T x + b) \geq 1.$$

우리는 위의 최적화 문제를 SVM의 원문제 (primal problem)로 하여 다음과 같이 형식화할 수 있다.

$$\begin{aligned} &\text{minimize } \frac{1}{2}||w||^2 \\ &\text{subject to } y_i(w^T x + b) \geq 1. \end{aligned} \tag{2.3}$$

학습자료가 선형분리 불가능한 경우는 여유변수 (slack variable) $\xi_i$를 도입하여 다음과 같이 형식화할 수 있다.

$$\text{minimize } Q(w, b, \xi) = \frac{1}{2}||w||^2 + C\sum_{i=1}^{n} \xi_i$$

$$\text{subject to } \begin{cases} y_i(w^T x_i + b) \geq 1 - \xi_i, & i = 1, \ldots, n \\ \xi_i \geq 0, & i = 1, \ldots, n. \end{cases}$$

여기서 $C$는 마진의 최대화와 분류 오류율의 최소화 사이 트레이드-오프 (trade-off)를 결정하는 벌칙모수 (penalty parameter)이다.

라그랑즈 배수 (Lagrange multiplier) $\alpha_i$를 도입하여 식 (2.3)의 원문제를 다음과 같이 라그랑즈 쌍대문제 (dual problem)로 변환할 수 있다.

$$\text{maximize } Q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j {x_i}^T x_j$$

$$\text{subject to } \begin{cases} \sum_{i=1}^{n} y_i \alpha_i = 0 \\ C \geq \alpha_i \geq 0, \ i = 1, \ldots, n. \end{cases}$$

따라서, 최적의 분리 초평면을 다음과 같이 입력벡터 $x$의 결정함수로 나타낼 수 있다.

$$f(x) = sgn(\sum_{i=1}^{N} \alpha_i y_i {x_i}^T x + b). \tag{2.4}$$

여기서 $sgn(x)$는 $x$가 양수이면 1, 0이면 0, 음수이면 -1을 갖는 함수이고 $N$은 서포트 벡터의 수이다.

대부분의 학습자료는 선형적으로 분리가 가능하지 않다. 따라서, 커널 함수 $K(x_i, x_j)$을 도입하여 식 (2.4)을 다음과 같이 나타낼 수 있다.

$$f(x) = sgn(\sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b).$$

커널함수 $K(x_i, x_j)$는 선형 분리 불가능한 경우, 어떤 함수의 형태로 클래스를 구분할지를 결정하는 함수로 자주 사용되는 커널함수는 Table 2.1과 같다.

**Table 2.1** Three common kernel functions

| kernel functions | expressions |
|---|---|
| Linear | $K(x_i, x_j) = {x_i}^T x_j$ |
| RBF(radial basis function) | $K(x_i, x_j) = exp(-\gamma\|\|x_i - x_j\|\|^2)$ |
| Polynomial | $K(x_i, x_j) = (\gamma x_i^T \mathbf{x}_j + r)^d$ |

여기서 $\gamma$, $r$ 그리고 $d$는 커널의 형태를 결정하는 모수들이다.

## 3. 교차 타당성을 이용한 앙상블 SVM 모형

### 3.1. $v$-조각 교차 타당성

$v$-조각 교차 타당성 ($v = 5$)을 Figure 3.1에서 보는 것처럼 단계별로 설명하면 다음과 같다.

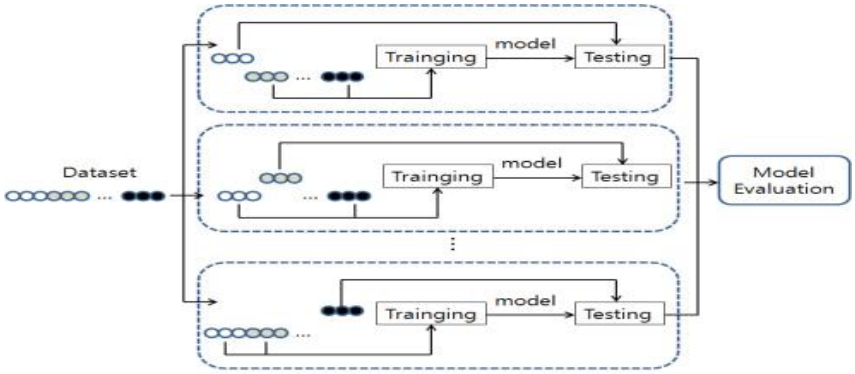


**Figure 3.1** Diagram for $v$-fold cross validation ($v = 5$)

단계 1. 전체 자료를 $v$개 조각으로 분할한다. 이때 조각의 크기는 같아야한다.

단계 2. 분할된 $v$개 조각 중에서 한 개 조각은 테스트용 자료 (testing data)로 사용하고 너머지 $v-1$개 조각은 훈련용 자료 (training data)로 사용한다.

단계 3. 단계 2의 훈련용 자료에서 SVM 모형을 추정한다.

단계 4. 단계 3에서 구한 추정된 SVM 모형을 단계 2의 테스트용 자료에 적용하여 정확도 (accuracy)를 계산한다. 정확도에 대한 정의는 제 3.3 절의 Table 3.2에 소개되어있다.

단계 5. 단계 2에서 분할된 모든 조각은 한번은 테스트용 자료로 사용되어야 한다. 따라서, 다른 조각에 대해서도 단계 2−단계 4를 반복 수행한다.

Figure 3.1에서 보는 것처럼 $v$-조각 교차 타당성을 통한 모형 평가 (model evaluation)는 자료 중 일부는 테스트용 자료로 사용되고 나머지 자료는 훈련용 자료로 번갈아 사용하면서 종합적으로 이루어진다. 본 논문에서 분류 모형의 정확도는 $v$-조각 교차 타당성 과정에서 얻어진 $v$개 조각에 대한 평균 정확도를 계산하고 위의 과정을 10번 반복하여 얻은 전체 평균 정확도를 가지고 계산한다.

### 3.2. 앙상블 SVM 모형

SVM 모형은 Figure 3.2와 같이 $v$-조각 교차 타당성에 의해 얻어진 $v$개의 학습용 자료에서 제 2 절에서 언급한 자주 사용되는 3가지 커널함수를 사용한 총 $n$개 ($n = 15$)의 SVM 모형을 생성한다. 총 $n$ 개 모형을 평가척도, 여기서는 정확도에 의해 상위 $k$개의 모형을 선택한 후 다수결의 투표 방법을 이용하여 미지의 클래스를 분류한다.

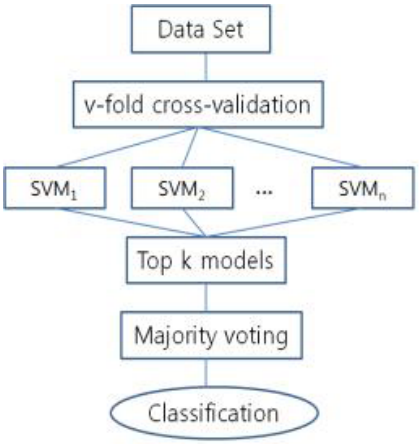


**Figure 3.2** Flow chart of ensemble SVM model

### 3.3. 모형 성능 평가 척도

앙상블 SVM 모형의 성능을 평가하기 위해 정확도, 오류율 (error rate), 민감도 (sensitivity), 특이도 (specificity), ROC (receiver operating characteristic) 곡선 (Egan, 1975), AUC (area under the curve; Cook, 2008) 등을 가지고 비교한다. 두 개의 집단을 $G_1$, $G_2$라 할때 분류모형에 대한 분류 결과에 대한 교차표는 Table 3.1과 같다.

**Table 3.1** Cross tables for actual group and classified group

| actual group | classified group | |
|---|---|---|
| | $G_1$ | $G_2$ |
| $G_1$ | $TP$ | $FN$ |
| $G_2$ | $FP$ | $TN$ |

Table 3.2는 평가척도들에 대해 Table 3.1의 분류 교차표에 있는 기호를 가지고 표현한 수식이다.

**Table 3.2** Performance measures for classification model

| measures | expressions |
|---|---|
| Accuracy | $\frac{TP+TN}{TP+FP+TN+FN}$ |
| Error rate | $\frac{FN+FP}{TP+FP+TN+FN}$ |
| Sensitivity | $\frac{TP}{TP+FN}$ |
| Specificity | $\frac{TN}{FP+TN}$ |

ROC 곡선은 분류모형의 (1-특이도)를 $x$축으로 하고 민감도를 $y$축으로 한 그래프이다. (1-특이도)를 오분류율 (false positive rate)이라하고 민감도를 정분류율 (true positive rate)이라고 한다. ROC 곡선은 분류모형을 이용하여 주어진 모든 자료에 대해 클래스에 속할 사후확률 (posterior probability)을 계산하고 계산된 사후확률을 정렬한 다음 임계값 (cut-point)이 변화함에 따라 오분류율과 정분류율의 변화를 그래프로 나타낸 것으로 ROC 곡선이 좌측 상단으로 더 위에 위치할수록 좋은 모형이다. AUC는 ROC 곡선 아래의 면적으로 c-통계량이라고도 하는데 어떤 모형의 ROC 곡선 아래의 면적이 다른 모형의 면적보다 크면 평균적으로 더 우수한 모형이라 할 수 있다.

## 4. 실험결과 및 논의사항

본 절에서는 실제 기업의 재무비율 자료인 Pietruszkiewicz 자료 (Pietruszkiewicz, 2004, 2008)와 모의실험을 통해 생성된 인공적인 자료를 가지고 앙상블 SVM모형의 성능을 여러 가지 평가 척도를 가지고 비교하고자 한다.

본 실험은 통계 프로그램 R을 사용하였고 SVM은 R 패키지 "e1071"을 사용하여 구현하였다 (Dimitriadou, 2005). SVM에서 모수 $C$와 $\gamma$값은 패키지 e1071의 기본값 즉, $C = 1$과 $\gamma = 1/$ (차원의 수)를 사용하였다.

### 4.1. 기업의 재무비율 자료

기업의 재무비율 자료는 기업의 부도 상태 여부를 나타내는 상태변수 (status)와 Table 4.1에서 보는 것처럼 30개의 재무비율을 나타내는 변수 (variable)로 구성되어 있다. 조사는 120개의 기업들에 대한 파산되기 2년에서 5년 전에 2번에 걸쳐 이루어 졌고 파산 기업 (failed company) 112개와 정상 기업 (non-failed company) 128개 총 240개 케이스로 구성되어있다.

**Table 4.1** Pietruszkiewicz data set

| Variable | Description |
|---|---|
| $X_1$ | cash/current liabilities |
| $X_2$ | cash/total assets |
| $X_3$ | current assets/current liabilities |
| $X_4$ | current assets/total assets |
| $X_5$ | working capital/total assets |
| $X_6$ | working capital/sales |
| $X_7$ | sales/inventory |
| $X_8$ | sales/receivables |
| $X_9$ | net profit/total assets |
| $X_{10}$ | net profit/current assets |
| $X_{11}$ | net profit/sales |
| $X_{12}$ | gross profit/sales |
| $X_{13}$ | net profit/liabilities |
| $X_{14}$ | net profit/equity |
| $X_{15}$ | net profit/(equity + long term liabilities) |
| $X_{16}$ | sales/receivables |
| $X_{17}$ | sales/total assets |
| $X_{18}$ | sales/current assets |
| $X_{19}$ | (365×receivables)/sales |
| $X_{20}$ | sales/total assets |
| $X_{21}$ | liabilities/total income |
| $X_{22}$ | current liabilities/total income |
| $X_{23}$ | receivables/liabilities |
| $X_{24}$ | net profit/sales |
| $X_{25}$ | liabilities/total assets |
| $X_{26}$ | liabilities/equity |
| $X_{27}$ | long term liabilities/equity |
| $X_{28}$ | current liabilities/equity |
| $X_{29}$ | EBIT (Earnings Before Interests and Taxes)/total assets |
| $X_{30}$ | current assets/sales |

Table 4.2는 5-조각 교차 타당성에 의해 얻어진 5개의 훈련용 자료에 대해 3가지 커널함수를 적용하여 얻어진 예측 정확도이다.

**Table 4.2** 5-fold cross validation classification accuracies for kernel functions

| No | kernel functions | | |
|---|---|---|---|
| | Linear | Polynomial | RBF |
| 1 | **0.8333** | 0.7083 | **0.8125** |
| 2 | 0.7917 | **0.8125** | 0.7708 |
| 3 | **0.8125** | 0.6875 | **0.8542** |
| 4 | 0.7292 | 0.7083 | 0.7917 |
| 5 | **0.8125** | 0.7708 | **0.8125** |
| Average accuracy | 0.7958 | 0.7375 | 0.8083 |

Table 4.2에서 No는 5개의 훈련용 자료에 붙여진 번호이고 각 훈련용 자료에 3개의 커널함수를 적용하여 총 15개의 SVM 모형을 얻었다. 총 15개 모형 중에서 정확도가 높은 상위 7개를 진한 글씨로 표시하였다. 평균 정확도는 Linear 커널함수를 사용한 경우는 0.7958, Polynomial 커널함수를 사용한 경우는 0.7375 그리고 RBF 커널함수를 사용한 경우는 0.8083으로 대체로 RBF 커널함수를 사용한 경우가 높은 정확도를 보였다.

Table 4.3은 Table 4.2에서 상위 7개 모형들의 분류 결과를 다수결 투표 시스템에 의해 결정하는 앙상블 SVM 모형과 제각기 커널함수를 사용하여 얻은 7개의 단일 SVM 모형들과의 여러 가지 평가척도 하에서 계산된 평가비교 수치들을 보여주는 표이다.

**Table 4.3** Comparison of models with kernel functions at different performance measures

| No | Kernel functions | Accuracy | Error rate | Sensitivity | Specificity |
|---|---|---|---|---|---|
| 1 | Linear | 0.7958 | 0.2042 | 0.7692 | **0.8211** |
| 3 | Linear | 0.7833 | 0.2167 | 0.7679 | 0.7969 |
| 5 | Linear | 0.7667 | 0.2333 | 0.7373 | 0.7951 |
| 2 | Polynomial | 0.7167 | 0.2833 | **1.0000** | 0.6531 |
| 1 | RBF | 0.7917 | 0.2083 | 0.8229 | 0.7708 |
| 3 | RBF | 0.8000 | 0.2000 | 0.8077 | 0.7941 |
| 5 | RBF | 0.8042 | 0.1958 | 0.8095 | 0.8000 |
| Ensemble | | **0.8208** | **0.1792** | 0.8416 | 0.8058 |

Table 4.3에서 보면 앙상블 모형은 정확도와 오류율에서 단일 모형보다 높은 성능을 보여주고 있다. 민감도에서 앙상블 모형이 0.8416으로 높은 수치임에도 불구하고 No 2에서 Polynomial 커널함수를 사용한 모형의 민감도가 1.0000으로 높은 수치를 보여 상대적으로 낮게 보이고 있으며 특이도에서도 No 1에서 Linear 커널함수를 사용한 모형보다는 낮은 수치를 보이나 그 밖에 다른 모형들보다는 높은 수치를 보여주고 있다. 그러나, 민감도에서 좋은 성능을 갖는 Polynomial 커널함수를 사용한 모형은 다른 평가척도 즉, 정확도, 오류율 그리고 특이도에서 최하위 수치를 보여주고 있음을 알 수 있다.

Figure 4.1은 Pietruszkiewicz 자료에서 앙상블 모형과 7개의 단일 모형에 대한 ROC 곡선이다.

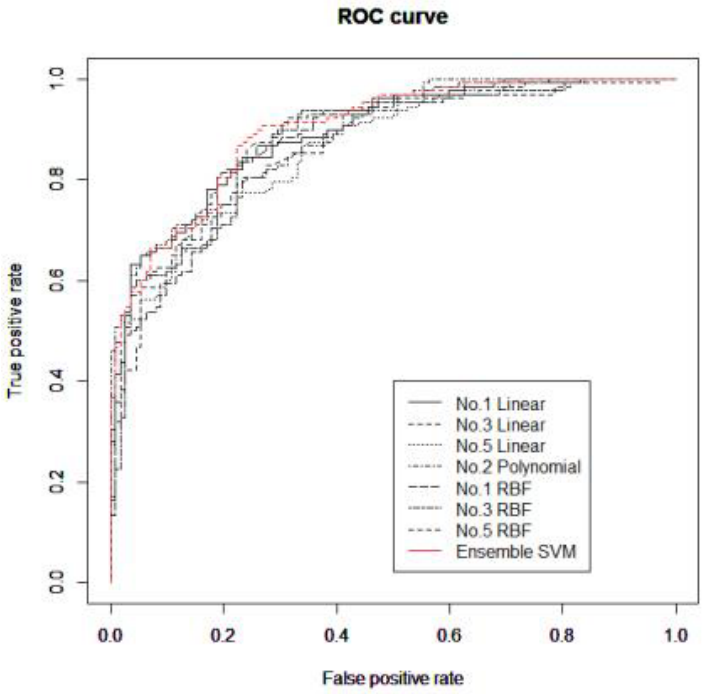


**Figure 4.1** ROC curves for Pietruszkiewicz data set

Figure 4.1을 보면 가시적으로는 앙상블 SVM 모형의 ROC 곡선이 단일 모형의 곡선들보다 전반적으로 상단에 위치해 좋은 성능을 갖는 것처럼 보이나 정확하게 구별하는 것은 쉽지않다. Table 4.4는 ROC 곡선의 아래 면적을 수치적으로 나타낸 AUC 값에 대한 표이다.

**Table 4.4** Performance comparison of AUC

| No | Kernel functions | AUC |
|---|---|---|
| 1 | Linear | 0.8902 |
| 3 | Linear | 0.8675 |
| 5 | Linear | 0.8633 |
| 2 | Polynomial | 0.8888 |
| 1 | RBF | 0.8800 |
| 3 | RBF | 0.8751 |
| 5 | RBF | 0.8842 |
| Ensemble | | **0.8991** |

Table 4.4로부터 앙상블 SVM 모형의 AUC가 0.8991로 다른 단일 모형보다 높은 AUC값을 갖고 있다.

### 4.2. 모의실험 자료

우리는 모의실험을 통해 이진분류 문제에 대해 다루고자 한다. 클래스를 나타내는 변수 $Y$는 0과 1을 갖는 베르누이 분포 (Bernoulli distribution)를 따르고, $Y = 1$인 표본 $X$는 평균이 $\mu_1$이고 공분산 $\Sigma$인 4차원 다변량 정규분포, $Y = 0$ 인 표본 $X$는 평균이 $\mu_2$이고 공분산 $\Sigma$인 똑같은 다변량 정규분포를 갖는다고 한다. 즉, 간단하게 표현하면 다음과 같다.

$$Y \sim Bernoulli(p),$$
$$Y = 1 \rightarrow X \sim N_4(\mu_1, \Sigma),$$
$$Y = 0 \rightarrow X \sim N_4(\mu_2, \Sigma).$$

여기서

$$p = 0.4,\ \mu_1 = \begin{pmatrix} 2 \\ 2 \\ 2 \\ 2 \end{pmatrix},\ \mu_2 = \begin{pmatrix} 6 \\ 6 \\ 6 \\ 6 \end{pmatrix},\ \Sigma = \begin{pmatrix} 3 & 1 & 1 & 1 \\ 1 & 3 & 1 & 1 \\ 1 & 1 & 3 & 1 \\ 1 & 1 & 1 & 3 \end{pmatrix}.$$

표본 1,000개를 생성하여 앙상블 SVM모형의 성능을 평가한다. Table 4.5는 5-조각 교차타당성에 의해 얻어진 5개의 훈련용 자료에 대해 3가지 커널함수를 적용하여 얻어진 예측 정확도이다.

**Table 4.5** 5-fold cross validation classification accuracies for kernel functions

| No | kernel functions | | |
|---|---|---|---|
| | Linear | Polynomial | RBF |
| 1 | 0.9550 | **0.9600** | 0.9550 |
| 2 | **0.9700** | 0.9550 | **0.9700** |
| 3 | 0.9550 | 0.9500 | 0.9550 |
| 4 | 0.9550 | 0.9300 | 0.9550 |
| 5 | 0.9500 | 0.9300 | 0.9550 |
| Average accuracy | 0.9570 | 0.9450 | 0.9580 |

Table 4.5에서 보면 총 15개의 모형 중에서 정확도가 높은 상위 3개를 진한 글씨로 표시하였다. 평균 정확도는 Linear 커널함수를 사용한 경우는 0.9570, Polynomial 커널함수를 사용한 경우는 0.9450 그리고 RBF 커널함수를 사용한 경우는 0.9580으로 Pietruszkiewicz 자료와 마찬가지로 RBF 커널함수를 사용한 경우가 높은 정확도를 보였다. 여기서 정확도가 높은 상위 3개를 선택한 이유는 다음 순위 즉, 4순위를 나타내는 정확도 0.9550을 가진 모형이 너무 많기 때문이다.

Table 4.6은 Table 4.5에서 예측 정확도가 높은 상위 3개 모형을 가지고 만든 앙상블 SVM 모형 결과와 3개의 단일 모형들과의 성능비교를 나타내고 있다.

**Table 4.6** Comparison of models with kernel functions at different performance measures

| No | Kernel functions | Accuracy | Error rate | Sensitivity | Specificity |
|---|---|---|---|---|---|
| 2 | Linear | 0.9400 | 0.0600 | 0.9535 | 0.9196 |
| 1 | Polynomial | **0.9420** | **0.0580** | 0.9492 | **0.9308** |
| 2 | RBF | 0.9410 | 0.0590 | 0.9521 | 0.9241 |
| Ensemble | | **0.9420** | **0.0580** | **0.9536** | 0.9242 |

Table 4.6에서 보면 앙상블 SVM 모형과 Polynomial 커널함수를 사용한 모형이 비슷한 성능을 갖는 것처럼 보인다. 정확도와 오류율면에서 위 두 모형은 같고 민감도에서는 앙상블 SVM 모형이 높으나 특이도에서는 역으로 Polynomial 커널함수를 사용한 모형이 좋게 나타났다. 그러나, 특이도에서 보면 앙상블 SVM 모형은 0.9242로 Polynomial 커널함수를 사용한 모형 다음으로 좋으나 민감도에서 보면 Polynomial 커널함수를 사용한 모형은 0.9492로 다른 커널함수 Linear와 RBF을 사용한 모형들보다도 성능이 떨어지는 것으로 나타났다.

Figure 4.2는 모의실험 자료에서 앙상블 모형과 단일 모형에 대한 ROC 곡선이다.

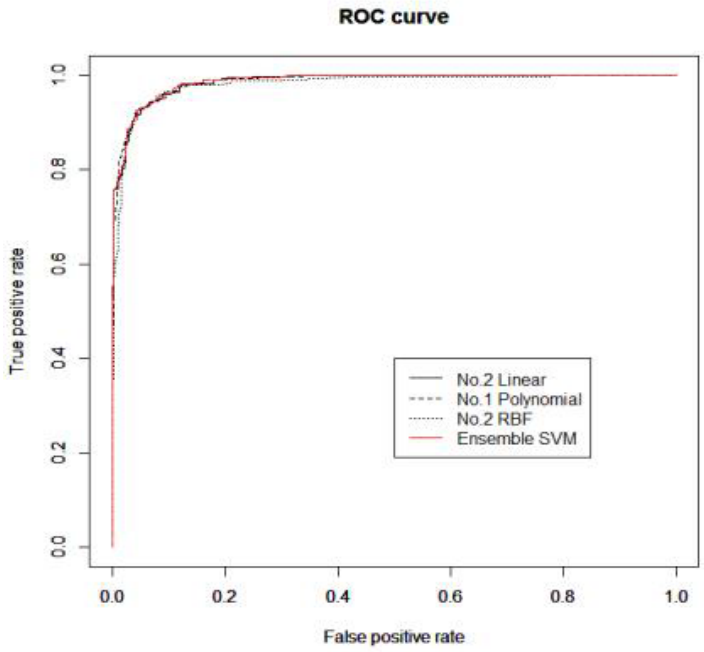


**Figure 4.2** ROC curves for simulated data set

Figure 4.2에서 보면 단일모형을 포함하여 모든 모형들이 비슷한 곡선모양을 보여 가시적으로 어느 모형이 좋고 나쁘고 판단하는 것은 쉽지 않다. Table 4.7은 모의실험 자료에서 앙상블 모형과 단일 모형에 대한 AUC의 값을 나타내고 있다.

**Table 4.7** Performance comparison of AUC

| No | Kernel functions | AUC |
|---|---|---|
| 2 | Linear | 0.9861 |
| 1 | Polynomial | 0.9863 |
| 2 | RBF | 0.9815 |
| Ensemble | | **0.9864** |

Table 4.7에 의해 앙상블 SVM 모형의 AUC가 0.9864로 다른 단일 모형보다 높은 AUC값을 보여주고 있다.

## 5. 결론

지금까지 기업 부도예측을 하는데 여러 가지 데이터마이닝 기법 들이 사용되었다. 그러나, 하나의 단일 모형을 가지고 복잡한 분류 문제에 적용하는데 한계를 갖고 있다. 최근에 SVM은 일반화 능력이 높기 때문에 많은 분류문제에 사용되고 있으나, 어떤 커널함수를 사용하느냐 따라 많은 성능 차이를 보인다. 따라서, 본 논문에서는 커널함수에 따른 성능편차를 줄이기 위해 여러 가지 커널함수에 따른 SVM 모형들을 결합한 앙상블 SVM 모형을 부도예측에 사용하였다.

일반적으로, 앙상블 모형은 단일모형들보다 성능이 높지만 간혹 단일모형 간 높은 상관 때문에 성능이 떨어지는 경우가 있다. 본 논문에서는 이를 해결하기 위해 $v$-조각 교차타당성을 사용하여 성능이 좋은 상위 $k$개의 단일 모형으로 앙상블 모형을 구성하고 과반수 투표 방식을 사용하여 미지의 클래스를 분류하였다.

본 논문에서 제안된 앙상블 SVM 모형의 성능을 평가하기 위해 실제 기업의 재무비율 자료와 모의실험 자료에서 평가척도인 정확도, 오류율, 민감도, 특이도, ROC 곡선 그리고 AUC을 가지고 실험하였다.

먼저, 기업 부도 자료에서 앙상블 SVM 모형은 정확도와 오류율에서 단일 모형들보다 좋은 성능을 보였다. 민감도와 특이도에서도 각각 Polynomial 커널함수와 Linear 커널함수를 사용한 모형보다는 낮은 수치를 보이나 그 밖에 다른 모형들보다는 높은 수치를 보였다. 민감도에서 좋은 성능을 보인 Polynomial 커널함수를 사용한 모형은 다른 평가척도에서는 매우 낮은 수치를 보임을 알 수 있었다. 모의실험 자료에서도 앙상블 SVM 모형은 Polynomial 커널함수 사용한 모형과 함께 좋은 성능을 보였다. 그러나, 민감도 평가척도에서 Polynomial 커널함수 사용한 모형은 다른 모형들보다 성능이 떨어지는 것으로 나타났다. 따라서 전반적으로 앙상블 SVM 모형은 좋은 성능을 갖고 있으면서 로버스트 (robust)한 결과를 보였다.

## References


Altman, E. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, **23**, 589-609.

Altman, E. (1983). *Corporate financial distress: A complete guide to predicting, avoiding and dealing with bankruptcy*, John Wiley and Sons, Inc., New York.

Anandarajan, M., Lee, P. and Anandarajan. A. (2004). Bankruptcy predication using neural networks. In *Business Intelligence Techniques: A Perspective from Accounting and Finance*, edited by M. Anandarajan, A. Anandarajan and C. Srinivasan, Springer-Verlag, Germany.

Bellovary, J., Giacomino, D. and Akers, M. (2007). A review of bankruptcy prediction studies: 1930 to present. *Journal of Financial Education*, **33**, 1-11.

Cook, N. R. (2008). Statistical evaluation of prognostic versus diagnostic models: Beyond the ROC curve. *Clinical Chemistry*, **54**, 17-23.

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D. and Weingessel. A. (2005). E1071: Misc functions of the department of statistics, Tu Wien. R package version 1.5-11.

Egan, J. (1975). *Signal decision theory and ROC analysis*, Academic Press, New York.

Eom, J. H., Kim, S. C. and Zhang, B. T. (2008). AptaCDSS-E: A classifier ensemble-based clinical decision support system for cardiovascular disease level prediction. *Expert Systems with Applications*, **34**, 2465-2479.

Kim, M. J. and Kang, D. K. (2012). Classifiers selection in ensembles using genetic algorithms for bankruptcy prediction. *Expert Systems with Applications*, **39**, 9308-9314.

Min, J. H. and Lee, Y. C. (2005). Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert Systems with Applications*, **28**, 603-614.

Ohlson, J. A. (1980). Financial ratios and the shirata's adjusted pattern according to the information probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 109-131.

Park, D., Yun, Y. and Yoon, M. (2012). Prediction of bankruptcy data using machine learning techniques. *Journal of the Korean Data & Information Science Society*, **23**, 569-577.

Pietruszkiewicz, W. (2004). *Application of discrete predicting structures in an early warning expert system for financial distress*, Ph.D. Thesis, Faculty of Computer Science and Information Technology, Szczecin University of Technology, Szczecin.

Pietruszkiewicz, W. (2008), Dynamical systems and nonlinear kalman filtering applied in classification. *Proceedings of 2008 7th IEEE International Conference on Cybernetic Intelligent Systems*, 263-68.

Shin, K. S., Lee, T. S. and Kim, H. J. (2005). An application of support vector machines in bankruptcy prediction model. *Expert Systems with Applications*, **28**, 127-35.

Vapnik. V. (1995). *The nature of statistical learning theory*, Springer-Verlag, New York.

Vapnik. V. (1998). *Statistical learning theory*, John Wiley and Sons, Inc., New York.

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*, **116**, 16-32.

Zmijewski, M. E. (1984). Methodological issues related to the estimated of financial distress prediction models. *Journal of Accounting Research*, **22**, 59-82.

*Journal of the Korean Data & Information Science Society*
*2013,* **24**(6), 1113–1125

http://dx.doi.org/10.7465/jkdi.2013.24.6.1113
한국데이터정보과학회지

# Bankruptcy prediction using ensemble SVM model[†]

Ha Na Choi[1] · Dong Hoon Lim[2]

[12]Department of Information Statistics, Gyeongsang National University

Received 24 June 2013, revised 10 July 2013, accepted 26 July 2013

**Abstract**

Corporate bankruptcy prediction has been an important topic in the accounting and finance field for a long time. Several data mining techniques have been used for bankruptcy prediction. However, there are many limits for application to real classification problem with a single model. This study proposes ensemble SVM (support vector machine) model which assembles different SVM models with each different kernel functions. Our ensemble model is made and evaluated by $v$-fold cross-validation approach. The $k$ top performing models are recruited into the ensemble. The classification is then carried out using the majority voting opinion of the ensemble. In this paper, we investigate the performance of ensemble SVM classifier in terms of accuracy, error rate, sensitivity, specificity, ROC curve, and AUC to compare with single SVM classifiers based on financial ratios dataset and simulation dataset. The results confirmed the advantages of our method: It is robust while providing good performance.

*Keywords*: Bankruptcy prediction, cross-validation, ensemble SVM, financial ratio, SVM.

[†] This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No.2011-0010089).

[1] Master of Science, Department of Information Statistics, Gyeongsang National University, Jinju 660-701, Korea.

[2] Corresponding author: Professor and RINS, Department of Information Statistics, Gyeongsang National University, Jinju 660-701, Korea. E-mail: dhlim@gnu.ac.kr