The Korean Journal of Applied Statistics (2015) DOI: http://dx.doi.org/10.5351/KJAS.2015.28.1.009
28(1), 9–21

Weighted L_1-Norm Support Vector Machine for the Classification of Highly Imbalanced Data

Eunkyung Kim[a] · Myoungshic Jhun[a] · Sungwan Bang[b,1]

[a]Department of Statistics, Korea University
[b]Department of Mathematics, Korea Military Academy

(Received September 18, 2014; Revised November 12, 2014; Accepted January 13, 2015)

Abstract

The support vector machine has been successfully applied to various classification areas due to its flexibility and a high level of classification accuracy. However, when analyzing imbalanced data with uneven class sizes, the classification accuracy of SVM may drop significantly in predicting minority class because the SVM classifiers are undesirably biased toward the majority class. The weighted L_2-norm SVM was developed for the analysis of imbalanced data; however, it cannot identify irrelevant input variables due to the characteristics of the ridge penalty. Therefore, we propose the weighted L_1-norm SVM, which uses lasso penalty to select important input variables and weights to differentiate the misclassification of data points between classes. We demonstrate the satisfactory performance of the proposed method through simulation studies and a real data analysis.

Keywords: Imbalanced data, lasso, linear programming, ridge, support vector machine.

1. 서론

범주형 자료에 대한 분류분석(classification analysis)은 실생활의 많은 분야에서 유용하게 활용되고 있다. 예를 들어, 분류분석은 의료분야에서 환자가 특정 암에 걸릴 가능성을 진단하거나 효과적인 치료법을 판단하는데 활용되고, 금융분야에서 은행의 여신담당자가 대출자의 대출 여부 결정하는데 이용되며, 통신분야에서는 이동통신 고객의 유지 또는 이탈 여부를 판별하는데 활용되는 등 의사결정을 필요로 하는 많은 분야에서 그 활용성이 지속적으로 증가하고 있다. Cortes와 Vapnik (1995), Vapnik (1998) 등에 의해 제안된 SVM(support vector machine)은 높은 분류 정확도와 유연성을 바탕으로 여러 다양한 분야에서 널리 활용되고 있는 분류분석 기법중 하나이다. 이러한 SVM의 적합식을 살펴보면 경첩(hinge) 손실함수에 릿지(ridge) (Hoerl과 Kennard, 1970) 형태의 2-norm 벌칙함수를 이용하여 분류함수를 추정하도록 공식화 되어있다. 따라서 SVM은 L_2-norm SVM으로도 불리며, 분류함수의 추정

Bang's research was supported by 2014 research fund of Korea Military Academy (20140516). Jhun's research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (NRF-2013R1A1A2A10007545).
[1]Corresponding author: Department of Mathematics, Korea Military Academy, 574 Hwarang Rd, Nowon-Gu, Seoul 139-799, Korea.

에서 발생할 수 있는 다중공선성과 과대적합의 문제를 해결함으로써 분류분석에서 높은 정확도를 나타내고 있다.

분류분석이 실생활에서 효과적으로 활용되기 위해서는 범주형 반응변수에 대한 분류 정확도가 높아야 할 뿐만 아니라, 분류함수에 대한 해석이 가능해야 한다. 예를 들어, 특정 질병을 연구하는 의사들에게는 이 질병에 영향을 미치는 중요요인에 대한 분석이 필요하고, 은행에서 대출심사 결과가 부적격으로 판정되었을 경우에는 고객에게 부적격의 이유에 대한 충분한 설명이 필요하다. 일반적으로 L_2-norm SVM은 분류분석에서 높은 정확도를 나타내고 있으나, 릿지 벌칙함수의 특성으로 인하여 모든 입력변수(input variable)를 분류모형에 포함하는 경향이 있다. 따라서 고차원 자료의 분류분석에서 L_2-norm SVM은 잡음변수들을 분류모형에서 제거하지 못하고, 이로 인하여 분류 정확도가 감소되고 모형의 해석이 어려워진다. 이러한 제한사항을 보완하기 위하여 Zhu 등 (2003)은 릿지 벌칙함수 대신 라소(lasso) (Tibshirani, 1996) 형태의 벌칙함수를 적용하는 L_1-norm SVM을 제안하였으며, 분석자료에 잡음변수가 포함되어 있는 경우 분류 정확도와 모형의 간결성 측면에서 L_2-norm SVM에 비해 높은 성능을 나타내는 것을 보였다.

한편, 이항 범주형 자료의 분류분석에서는 집단(class)별 개체수가 상이한 불균형 자료(imbalanced data)를 자주 접하게 된다. 일반적으로 SVM은 불균형 자료의 분류분석에서 비교적 강건한 결과를 제공하나, 다른 분류기법들과 마찬가지로 불균형의 정도가 심각해짐에 따라 분류함수의 추정에서 집단 간 편향이 발생하고, 이에 따라 개체수가 작은 소수집단(minority class)에 대한 분류 정확도가 현저하게 감소하게 된다. 그러나 실제의 많은 사례에서는 다수집단(majority class)의 오분류보다 소수집단의 오분류가 더 중요하게 다루어진다. 불균형 자료의 분류분석에서 소수집단의 분류 정확도를 향상시키기 위하여 많은 연구가 진행되었으며, 대표적으로 오분류 비용을 차등 적용하는 방법과 과대추출(over sampling) 및 과소추출(under sampling)을 통해 인위적으로 집단별 개체수를 균형되게 조정하는 방법 등이 연구되었다 (Kubat과 Matwin, 1997; Japkowicz, 2000; Chawla 등, 2002; Barandela 등, 2003; Kim과 Jeong, 2004; Cohen 등, 2006; Garcia 등, 2007; Ganganwar, 2012; Lee와 Lee, 2014).

SVM을 불균형 자료의 분류분석에 적용할 때 발생할 수 있는 분류함수의 집단 간 편향을 줄이고 소수집단의 분류 정확도를 향상시키기 위해서도 많은 연구가 진행되었다. Akbani 등 (2004)과 Han 등 (2005)은 소수집단의 새로운 개체를 생성하는 SMOTE 기법 (Chawla 등, 2002)을 SVM에 적용하는 방법을 연구하였으며, Tang 등 (2009)은 다수집단에서 중복되거나 서포트 벡터로써 역할하지 않는 개체를 효과적으로 제거하는 과소추출 방법을 제안하였다. 또한, Veropoulos 등 (1999), Lin 등 (2002), 그리고 Akbani 등 (2004)은 가중치를 이용하여 집단별로 오분류 비용을 차등 적용하는 가중(weighted) L_2-norm SVM(WL_2-norm SVM)을 제안하였으며, 배깅이나 부스팅 알고리즘을 적용한 SVM (Liu 등, 2006; Wang과 Japkowicz, 2009)에 관해서도 연구가 진행되었다. 그러나 이러한 기존의 연구들은 릿지 형태의 벌칙함수를 사용하는 L_2-norm SVM에 기반하므로 불필요한 잡음변수가 포함되는 불균형 자료에서는 그 활용이 제한된다. 따라서, 본 논문에서는 불균형 자료의 분류분석에서 소수집단에 대한 분류 정확도를 향상시킴과 동시에 범주형 반응변수에 관련되어 있는 중요한 입력변수만을 선택하여 간결한 모형을 제공하는 새로운 SVM 기법을 연구하였다. 분류함수의 추정에서 잡음변수를 제거하기 위하여 L_1-norm SVM을 활용하였으며, 소수집단에 대한 분류 정확도의 향상을 위해서는 가중치를 이용하여 집단별로 오분류 비용을 차등 적용하는 방법을 이용하여 가중 L_1-norm SVM을 제안하였다. 본 논문의 구성은 다음과 같다. 2절에서는 먼저 L_2-norm SVM과 L_1-norm SVM을 소개하고, 불균형 자료의 분류분석을 위한 가중 L_1-norm SVM을 제안하였다. 3절과 4절에서는 모의실험과 실제자료의 분석을 통해 기존의 분류기법과 제안한 가중 L_1-norm SVM의 성능을 비교하였으며, 제안한 방법론의 활용 가능성을 보였다. 마지막으로 5절에서는 결론과 더불어 차후 연구방향을 제시하였다.

2. 불균형 자료의 분류분석을 위한 가중 L_1-norm SVM

2.1. L_2-norm SVM과 L_1-norm SVM의 비교

개체의 소속집단을 나타내는 범주형 반응변수 $y_i \in \{-1, 1\}$와 p차원 입력변수 $\boldsymbol{x}_i \in R^p$로 이루어진 훈련자료 $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$에 근거하여, 선형 분류함수 $f(\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\beta} + \beta_0$를 추정하는 문제를 고려하자. 여기서 새로운 개체의 입력값 $\boldsymbol{x} \in R^p$가 주어졌을 때 이 개체의 소속집단은 $y = \text{sign}(f(\boldsymbol{x}))$로 분류된다. 분류함수의 추정에서 다중 공선성 및 과대적합의 문제를 해결하기 위하여 손실함수와 벌칙함수를 동시에 고려하는 추정법에 관하여 많은 연구가 진행되었으며, 대표적인 것으로 L_2-norm SVM을 들 수 있다 (Cortes와 Vapnik, 1995; Vapnik, 1998). L_2-norm SVM은 훈련자료들의 마진(margin)을 최대로 하는 최적화 식 (2.1)

$$\left(\hat{\beta}_0, \hat{\boldsymbol{\beta}}\right)^{L_2\text{-SVM}} = \arg\max_{\boldsymbol{\beta}, \beta_0} \frac{1}{||\boldsymbol{\beta}||_2}$$
$$\text{subject to } \; y_i \left\{\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}\right\} \geq 1 - \xi_i, \quad \xi_i \geq 0, \; \forall\, i \;\text{ and }\; \sum_{i=1}^n \xi_i \leq s \tag{2.1}$$

을 통해 분류함수의 계수를 추정하며, 최적화 식 (2.1)은 손실함수에 릿지 형태의 벌칙함수가 적용된 적합식 (2.2)로 표현 가능하다.

$$\left(\hat{\beta}_0, \hat{\boldsymbol{\beta}}\right)^{L_2\text{-SVM}} = \arg\min_{\beta_0, \boldsymbol{\beta}} \sum_{i=1}^n \left[1 - y_i\left(\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}\right)\right]_+ + \lambda\,||\boldsymbol{\beta}||_2^2\,, \tag{2.2}$$

여기서 경첩 손실함수 $[t]_+ = \max(t, 0)$이고 $\lambda > 0$는 훈련자료의 오차와 벌칙항간의 균형을 맞추어 과대적합을 방지하는 조율모수로 식 (2.1)의 s와 일대일로 대응된다.

일반적으로 L_2-norm SVM은 높은 수준의 분류 정확도와 유연성을 바탕으로 여러 다양한 분야의 분류분석에서 널리 사용되고 있다. 그러나 릿지 벌칙함수의 특성으로 인하여 L_2-norm SVM은 불필요한 잡음 변수들의 제거에는 효율적이지 못하므로 고차원의 설명변수를 가지는 자료의 분류분석에는 부적절하다. 분류함수의 추정에서 잡음 변수의 선택은 모형의 해석을 어렵게 할 뿐만 아니라 분류 정확도를 감소시키며, 중요한 입력변수의 잘못된 제거는 추정을 심각하게 왜곡시키는 결과를 제공한다. 분류함수의 추정에서 중요한 입력변수의 동시적인 선택을 위하여 Zhu 등 (2003)은 식 (1.2)의 릿지 벌칙함수 대신 라소 형태의 벌칙함수를 적용하는 L_1-norm SVM을 제안하였으며, 그 적합식은

$$\left(\hat{\beta}_0, \hat{\boldsymbol{\beta}}\right)^{L_1\text{-SVM}} = \arg\min_{\beta_0, \boldsymbol{\beta}} \sum_{i=1}^n \left[1 - y_i\left(\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}\right)\right]_+ + \lambda||\boldsymbol{\beta}||_1 \tag{2.3}$$

와 같다. L_2-norm으로 표현되는 릿지 벌칙함수와 마찬가지로 L_1-norm의 라소 벌칙함수는 회귀계수 β_j $(j = 1, \ldots, p)$를 0 방향으로 축소 추정함으로써 추정량의 분산을 감소시키는 이점이 있다. 더불어 식 (2.3)의 조율모수 λ가 충분히 클 때 라소 벌칙함수는 L_1-norm의 특성으로 인하여 불필요한 입력변수의 회귀계수를 0으로 정확하게 추정함으로써 잡음변수를 모형에서 제거하게 된다.

2.2. 가중 L_1-norm SVM(WL_1-norm SVM)

두 집단으로 이루어진 범주형 자료의 분류분석에서 SVM은 분류함수 근처의 훈련개체, 즉 서포트 벡터(support vector)만을 분류함수의 추정에 사용하므로 마진의 범위를 넘어서고 분류함수로부터 멀리 있는 훈련개체는 추정에 아무런 영향을 주지 않는다. 따라서 여러 다양한 분류기법들과 비교할 때

SVM은 두 집단 간의 훈련 개체수가 적당히 불균형된 자료의 분류분석에서는 비교적 강건한 결과를 제공한다. 그러나 불균형의 정도가 심각해짐에 따라 SVM도 다른 일반적인 분류기법들과 마찬가지로 분류함수의 추정에서 집단 간의 심각한 편향이 발생하게 된다.

설명의 편의를 위해 +집단을 개체수가 작은 소수집단이라고 하고, −집단을 개체수가 많은 다수집단이라고 하자. 불균형 자료의 분류분석에서 SVM을 포함한 기존의 분류기법은 개체의 오분류에 대한 집단별 차이를 고려하지 않기 때문에 분류함수가 다수집단에 편향되게 추정됨으로써 소수집단의 분류 정확도가 심각하게 감소하게 된다. 그러나 희귀병 환자의 판별, 이동통신에서 이탈고객 선별, 신용거래에서 사기거래자 적발, 금융거래에서 신용불량자 적발 등 실제자료의 분류분석에서는 소수집단의 오분류가 다수집단의 오분류보다 더 심각한 결과를 초래하는 경우가 많다. 예를 들어, 특정 암에 걸린 10명의 인원들로 구성된 소수집단과 암에 걸리지 않은 990명의 인원들로 구성된 다수집단을 고려할 때, 소수집단의 오분류는 환자의 치료시기를 놓치게 함으로써 생명에 크게 영향을 미치므로 다수집단의 오분류 보다 더 중요하게 다루어져야 한다. 그러나 집단별 중요도의 차이를 고려하지 않고 모든 인원을 정상으로 분류하면 전체 정확도는 99%라고 하더라도 암환자의 분류 정확도는 0%로 이러한 분류함수는 아무런 의미를 갖지 못한다. 분류함수의 추정에서 집단별로 오분류에 대한 중요도를 구분하기 위하여 Veropoulos 등 (1999)과 Akbani 등 (2004)은 L_2-norm SVM에 집단별로 가중치를 적용하는 WL_2-norm SVM을 제안하였으며, 그 적합식은

$$\left(\hat{\beta}_0, \hat{\boldsymbol{\beta}}\right)^{WL_2\text{-SVM}} = \arg\min_{\boldsymbol{\beta}, b} \frac{1}{2}||\boldsymbol{\beta}||_2^2 + C^+ \sum_{\{i|y_i=+1\}}^{n} \xi_i + C^- \sum_{\{i|y_i=-1\}}^{n} \xi_i$$
$$\text{subject to } \; y_i \left\{\beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta}\right\} \geq 1 - \xi_i, \quad \xi_i \geq 0, \forall\, i \tag{2.4}$$

와 같다. 여기서 C^+와 C^-는 각각 소수집단(+)과 다수집단(−)의 오분류에 대한 비용을 나타내고, $C^+ = C^- = 1/(2\lambda)$일 때 적합식 (2.4)는 적합식 (2.2)와 동일함을 알 수 있다.

불균형 자료의 분석에서 적합식 (2.4)의 WL_2-norm SVM은 소수집단(+)의 오분류 비용 C^+를 다수집단(−)의 오분류 비용 C^-보다 상대적으로 크게 부여함으로써 소수집단(+)의 분류 정확도를 높일 수 있다. 그러나 L_2-norm SVM에서와 마찬가지로 WL_2-norm SVM은 릿지 형태의 벌칙함수를 사용하므로 중요한 입력변수들의 동시적인 선택에는 그 사용이 제한된다. 분류분석을 필요로 하는 대부분의 실제자료는 집단 간의 개체수가 불균형적으로 구성되어 있으며, 분류함수의 정확도와 더불어 모형의 해석력이 중요시 된다. 예를 들어 고객의 신용평가 자료는 일반적으로 상당히 불균형적이며, 고객의 심사 결과가 부적격 판정이 나올 경우 고객에게 부적격의 이유를 설명해야 하므로 해석력이 중요시된다. 또한 WL_2-norm SVM은 집단별 중요도에 따라 오분류의 비용을 달리하여 분류함수가 다수집단으로 편향되지 않도록 하지만 개별 훈련개체의 중요도에 대한 고려는 하지 못한다. 따라서 본 논문에서는 불균형자료의 분류분석에서 중요한 입력변수의 선택 기능을 만족시키는 분류기법을 개발하기 위하여 L_1-norm SVM에 훈련개체별로 가중치를 적용하는 가중(weighted) L_1-norm SVM(WL_1-norm SVM)을

$$\left(\hat{\beta}_0, \hat{\boldsymbol{\beta}}\right)^{WL_1\text{-SVM}} = \arg\min_{\beta_0, \boldsymbol{\beta}} \sum_{i=1}^{n} c_i \left[1 - y_i \left(\beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta}\right)\right]_+ + \lambda||\boldsymbol{\beta}||_1 \tag{2.5}$$

와 같이 제안한다. 여기서 c_i는 i번째 훈련개체의 오분류에 대한 비용을 나타낸다. 적합식 (2.4)에서는 집단 단위로 오분류 비용이 적용되는 반면에, 적합식 (2.5)는 훈련개체 각각의 중요도에 따라 오분류 비용을 구분할 수 있으므로 불균형 자료의 분류분석을 위하여 다수집단의 훈련개체를 과소추출하거나 소수집단의 훈련개체를 과대 추출 또는 생성하는 다양한 방법론 (Japkowicz, 2000; Chawla 등, 2002; Bang과 Jhun, 2014)과의 결합이 용이하다. 예를 들어, 소수집단의 새로운 개체가 생성되었다고 하면,

Table 3.1. Confusion matrix

	Positive prediction (소수집단으로 예측)	Negative prediction (다수집단으로 예측)
Positive class (실제 소수집단)	TP (True Positive)	FN (False Negative)
Negative class (실제 다수집단)	FP (False Positive)	TN (True Negative)

적합식 (2.5)에서는 소수집단 내에서 기존의 훈련개체와 생성된 훈련개체의 오분류에 대한 비용을 차등 적용할 수 있다.

2.3. 가중 L_1-norm SVM의 계산 알고리즘

적합식 (2.4)의 WL_2-norm SVM은 릿지 형태의 벌칙함수를 사용하므로 그 최적화 식은 이차계획법(quadratic programming)으로 공식화 되는 반면에 제안한 WL_1-norm SVM의 적합식 (2.5)는 선형계획법(linear programming)으로 공식화 될 수 있다. 이를 위해 먼저 n개의 slack 변수를 $\xi_i = [1 - y_i(\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta})]_+$ $(i = 1, \ldots, n)$라고 하자. 또한 분류함수의 회귀계수를 양수 부분 $\beta_j^+ \geq 0$과 음수 부분 $\beta_j^- \geq 0$을 이용하여 $\beta_j = \beta_j^+ - \beta_j^-$ $(j = 0, 1, \ldots, p)$로 표현하면 WL_1-norm SVM의 적합식 (2.5)는 선형계획법의 최적화 식 (2.6)과 동일함을 쉽게 보일 수 있다.

$$\begin{aligned} &\arg\min \sum_{i=1}^{n} c_i\xi_i + \lambda \sum_{j=1}^{p} \left(\beta_j^+ + \beta_j^-\right) \\ &\text{subject to } \; y_i \left(\beta_0^+ - \beta_0^- + \sum_{j=1}^{p} x_{ij}\left(\beta_j^+ - \beta_j^-\right)\right) \geq 1 - \xi_i, \quad \xi_i \geq 0, \; \forall\, i \\ &\qquad\qquad \beta_j^+ \geq 0, \; \beta_j^- \geq 0, \; \forall\, j. \end{aligned} \tag{2.6}$$

따라서 분류함수의 계산 속도를 고려할 때 선형계획법으로 공식화되는 WL_1-norm SVM이 이차계획법으로 공식화되는 WL_2-norm SVM에 비해 더욱 효율적이라고 할 수 있다.

본 논문에서는 WL_1-norm SVM의 최적화 식 (2.6)의 최적해를 구하기 위해 R 프로그램 (R Core Team, 2014)의 lpSolve 패키지 (Berkelaar 등, 2014)에 포함되어 있는 lp 함수를 사용하였다. 또한 WL_2-norm SVM의 최적해는 quadprog 패키지 (Turlach와 Weingessel, 2013)에 포함되어 있는 solve.QP 함수를 사용하였다. WL_2-norm SVM과 본 논문에서 제안한 WL_1-norm SVM의 최적화 식에 대한 R 코드는 차후 연구에 도움이 되도록 요청 시 제공할 것이다.

3. 모의 실험

3.1. 성능 평가(Performance measures)

이항 범주형 자료의 분석에서 분류기법들의 성능은 새로운 입력개체를 실제 집단으로 얼마나 잘 분류해 내는가를 나타내는 분류 정확도로 평가되며, Table 3.1의 오차행렬(confusion matrix)을 이용하면 다양한 종류의 분류 정확도를 측정할 수 있다. 오차행렬에서 TP는 실제 소수집단의 개체를 소수집단으로 올바르게 예측한 개체수를 의미하고, TN은 실제 다수집단의 개체를 다수집단으로 올바르게 예측한 개체수를 의미한다. 반면에 FN은 실제 소수집단의 개체를 다수집단으로 잘못 예측한 개체수를 의미하며, FP는 실제 다수집단의 개체를 소수집단으로 잘못 예측한 개체수를 나타낸다.

일반적으로 분류기법의 성능평가에서 가장 많이 사용되는 지표는 모든 훈련개체에 대한 전체 정확도(overall accuracy)로

$$\text{Overall accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}} \tag{3.1}$$

와 같이 정의되며, 이는 모든 개체 중에서 올바르게 분류된 개체의 비중을 나타낸다. 그러나, 집단별 자료의 개체수가 불균형인 경우에는 분류함수가 다수집단 쪽으로 강하게 편향되어 추정될 수 있으므로 전체 정확도만으로 분류기법의 성능을 평가하기에는 제한이 된다. 따라서 불균형 자료에서는 소수집단과 다수집단을 각각 독립적으로 평가하는 방법이 대안이 될 수 있으며, 소수집단에 대한 분류 정확도를 나타내는 민감도(sensitivity)와 다수집단에 대한 분류 정확도를 나타내는 특이도(specificity)는 각각 다음과 같다.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{3.2}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}. \tag{3.3}$$

또한 Kubat과 Matwin (1997)은 불균형 자료의 분류분석에서 민감도와 특이도의 기하평균(g-mean)

$$G\text{-mean} = \sqrt{\text{Sensitivity} \times \text{Specificity}} \tag{3.4}$$

을 전체 개체에 대한 평가지표로 사용하였다. 본 논문에서는 분류기법들의 예측력을 평가하기 위하여 전체 정확도(overall accuracy), 민감도(sensitivity), 특이도(specificity), 기하평균(g-mean)을 이용하였다.

3.2. 모의 실험

본 논문에서 제안하는 WL_1-norm SVM과 기존의 방법인 L_2-norm SVM, WL_2-norm SVM, 그리고 L_1-norm SVM의 성능을 비교하기 위하여 2가지 모형에 대한 모의실험을 시행하였다. WL_2-norm SVM과 제안 방법인 WL_1-norm SVM에서 사용된 오분류 비용은 집단별로 부여하였다. 즉, 다수집단($-$)과 소수집단($+$)의 개체수를 각각 N^-와 N^+로 나타낼 때, 다수집단($-$)의 오분류 비용 c_i는 $C^- = N^+/(N^+ + N^-)$로, 소수집단($+$)의 오분류 비용 c_i는 $C^+ = N^-/(N^+ + N^-)$로 부여하였다.

각각의 모의실험에서 모형적합을 위해 총 1,000개의 훈련자료(training data)를 생성하였으며, 이때 소수집단의 비율을 5%(소수집단 50개, 다수집단 950개), 10%(소수집단 100개, 다수집단 900개), 20%(소수집단 200개, 다수집단 800개)로 하여 불균형의 정도를 달리하였다. 또한, 조율모수 λ를 선택하기 위해 크기가 1000인 검증자료(validation data)와 분류기법들의 분류 정확도를 평가하기 위해 크기가 10,000인 평가자료(test data)를 각각 독립적으로 생성하였다. 각 분류기법의 예측력을 평가하기 위하여 전체 정확도, 민감도, 특이도, 그리고 기하평균을 계산하였으며, 변수선택의 성능을 평가하기 위하여 중요한 입력변수 중에서 유의한 변수로 올바르게 선택된 개수(NC; number of correctly selected input variable)와 잡음변수 중에서 유의한 변수로 잘못 선택된 개수(NIC; number of incorrectly selected input variable)를 각각 계산하였다. 이러한 과정을 100번 독립적으로 반복하였으며, 각 평가지표에 대한 100번의 평균을 표에 나타내었다.

(1) 실험모형 1

실험모형 1에서는 먼저 입력변수 $\boldsymbol{x} = (x_1, \ldots, x_{12})^T$를 다변량 정규분포 $N_{12}(\mathbf{0}, \Sigma)$로부터 생성하였다. 여기서 공분산 행렬 Σ의 (i, j)번째 원소는 $\text{Cov}(x_i, x_j) = 0.5^{|i-j|}$이다. 이항 범주형 반응변수 Y는 로지

Table 3.2. Simulation results for Example 1

Percentage of Minority class	Method	Test classification accuracy				Input variable selection	
		Overall accuracy	Sensitivity	Specificity	G-mean	NC	NIC
20%	L_2-norm SVM	0.840 (0.008)	0.712 (0.018)	0.968 (0.005)	0.830 (0.009)	3.00	9.00
	WL_2-norm SVM	0.880 (0.004)	0.877 (0.013)	0.883 (0.012)	0.880 (0.004)	3.00	9.00
	L_1-norm SVM	0.841 (0.007)	0.714 (0.018)	0.968 (0.005)	0.831 (0.009)	3.00	8.14
	WL_1-norm SVM	0.884 (0.004)	0.880 (0.013)	0.887 (0.011)	0.884 (0.004)	3.00	3.30
10%	L_2-norm SVM	0.789 (0.013)	0.591 (0.027)	0.986 (0.004)	0.763 (0.017)	3.00	9.00
	WL_2-norm SVM	0.876 (0.006)	0.868 (0.020)	0.884 (0.019)	0.876 (0.006)	3.00	9.00
	L_1-norm SVM	0.791 (0.014)	0.597 (0.030)	0.986 (0.004)	0.767 (0.018)	3.00	8.30
	WL_1-norm SVM	0.883 (0.004)	0.879 (0.018)	0.886 (0.016)	0.883 (0.005)	3.00	2.47
5%	L_2-norm SVM	0.736 (0.019)	0.477 (0.038)	0.994 (0.002)	0.688 (0.027)	3.00	9.00
	WL_2-norm SVM	0.869 (0.007)	0.858 (0.024)	0.870 (0.019)	0.869 (0.007)	3.00	9.00
	L_1-norm SVM	0.741 (0.020)	0.487 (0.041)	0.994 (0.002)	0.695 (0.029)	3.00	8.34
	WL_1-norm SVM	0.878 (0.008)	0.873 (0.023)	0.883 (0.021)	0.878 (0.008)	3.00	2.90

The numbers in parentheses are standard deviations.

스틱 모형 $P(Y=1) = \exp(f(\boldsymbol{x}))/(1+\exp(f(\boldsymbol{x})))$, $P(Y=-1) = 1 - P(Y=1)$에 의해 생성되었으며, 이때 실제 분류함수로는

$$f(\boldsymbol{x}) = 3x_1 - 1.5x_2 + 2x_5 \tag{3.5}$$

을 이용하였다. 선형 분류함수 (3.5)는 총 12개의 입력변수 중에서 3개의 유의한 입력변수 x_1, x_2, x_5를 포함하고 있다.

Table 3.2에는 각각의 분류기법에 대한 모의실험의 결과가 소수집단에 대한 훈련개체의 비율별로 정리되어 있다. 이로부터 오분류 비용을 균일하게 적용하는 L_2-norm SVM과 L_1-norm SVM의 경우 소수집단에 대한 훈련개체의 비율이 낮아질수록 소수집단의 분류 정확도인 민감도가 급격히 낮아져 전체 정확도와 기하평균이 낮아지는 것을 확인할 수 있다. 반면에 집단별 오분류 비용을 차등적으로 적용한 WL_2-norm SVM과 WL_1-norm SVM은 소수집단의 비율에 크게 영향을 받지 않으며, 가중치를 적용하지 않은 방법론에 비해 전체 정확도와 기하평균의 측면에서 그 성능이 향상되었음을 알 수 있다. 이는 분류분석에서 그 중요성이 상대적으로 낮은 다수집단의 정확도 즉, 특이도가 다소 감소하였으나, 중요하게 고려되는 소수집단의 정확도인 민감도가 크게 개선되었기 때문이다. 특히, 제안 방법인 WL_1-norm SVM의 성능이 기존의 방법인 WL_2-norm SVM에 비해 4가지 지표 모두에서 향상된 것을 알 수 있다.

Table 3.3. Simulation results for Example 2

Percentage of Minority class	Method	Test classification accuracy				Input factor selection		Input variable selection	
		Overall accuracy	Sensitivity	Specificity	G-mean	NC	NIC	NC	NIC
20%	L_2-norm SVM	0.840 (0.010)	0.736 (0.021)	0.975 (0.007)	0.847 (0.010)	3.00	6.00	6.00	21.00
	WL_2-norm SVM	0.882 (0.005)	0.860 (0.012)	0.910 (0.013)	0.885 (0.005)	3.00	6.00	6.00	21.00
	L_1-norm SVM	0.842 (0.010)	0.739 (0.023)	0.975 (0.007)	0.849 (0.011)	3.00	6.00	5.97	19.30
	WL_1-norm SVM	0.886 (0.004)	0.866 (0.010)	0.911 (0.011)	0.888 (0.004)	3.00	5.47	5.75	12.70
10%	L_2-norm SVM	0.796 (0.016)	0.648 (0.031)	0.988 (0.004)	0.800 (0.018)	3.00	6.00	6.00	21.00
	WL_2-norm SVM	0.872 (0.008)	0.845 (0.021)	0.907 (0.019)	0.875 (0.007)	3.00	6.00	6.00	21.00
	L_1-norm SVM	0.799 (0.017)	0.653 (0.033)	0.988 (0.005)	0.803 (0.019)	3.00	5.97	5.87	18.95
	WL_1-norm SVM	0.879 (0.006)	0.861 (0.016)	0.903 (0.018)	0.882 (0.006)	3.00	4.96	5.04	9.38
5%	L_2-norm SVM	0.752 (0.012)	0.564 (0.037)	0.994 (0.003)	0.748 (0.024)	3.00	6.00	6.00	21.00
	WL_2-norm SVM	0.855 (0.011)	0.816 (0.030)	0.905 (0.027)	0.859 (0.010)	3.00	6.00	6.00	21.00
	L_1-norm SVM	0.755 (0.022)	0.571 (0.041)	0.993 (0.003)	0.752 (0.027)	3.00	5.98	5.75	19.00
	WL_1-norm SVM	0.869 (0.009)	0.847 (0.027)	0.897 (0.027)	0.881 (0.008)	3.00	4.50	4.16	7.33

The numbers in parentheses are standard deviations.

또한, 변수 선택에 있어서도 제안 방법인 WL_1-norm SVM은 실제 입력변수 3개를 모두 선택하였으며, 분류분석 기법들 중에서 잡음변수를 가장 적게 선택하였음을 확인할 수 있다.

(2) 실험모형 2

실험모형 2에서는 9개의 잠재변수 $\boldsymbol{z} = (z_1, \ldots, z_9)^T$를 다변량 정규분포 $N_9(\mathbf{0}, \Sigma)$로부터 먼저 생성하였다. 여기서 공분산 행렬 Σ의 (i,j)번째 원소는 $\mathrm{Cov}(x_i, x_j) = 0.5^{|i-j|}$이다. 그 후에 독립적으로 잠재변수 w를 표준정규분포로부터 생성하고, 입력요인(input factor) $\boldsymbol{x} = (x_1, \ldots, x_9)^T$를 $x_j = (1/\sqrt{2})(z_j + w)$, $j = 1, \ldots, 9$와 같이 생성하였다. 이항 반응형 반응변수 Y는 로지스틱 모형 $P(Y = 1) = \exp(f(\boldsymbol{x}))/(1 + \exp(f(\boldsymbol{x})))$, $P(Y = -1) = 1 - P(Y = 1)$에 의해 생성되었으며, 이때 실제 분류함수로는

$$f(\boldsymbol{x}) = x_3 + x_3^2 + x_3^3 + 3x_6 + 1.5x_6^2 + 2x_9 - 1 \tag{3.6}$$

을 이용하였다. 비선형 분류함수 (3.6)은 연속형 입력요인의 3차 다항식을 입력변수로 이용한 가법모형이다. 총 9개의 요인 중에서 세 개의 요인 x_3, x_6, x_9이 중요한 요인으로 활용되었으며, x_3은 3차 다항식, x_6은 1, 2차 다항식, x_9는 1차 다항식만이 유의하여 최종적으로 6개의 입력변수가 분류함수에 포함

Table 4.1. Description of input factors and input variables

No	Input factor	Information of input factor	Type	Number of input variables
1	INCOME	년소득	continuous	3
2	AGE	연령		3
3	LOAN_CNT	기존대출 총건수		3
4	NEW_LOAN1_CNT	최근 1년내 신규대출건수		3
5	LOAN_AMT	현재 보유대출 금액		3
6	CARD_RATE	신용카드 사용 비율		3
7	CARD_CNT	신용카드 개수		3
8	CARD_AMT	신용카드 이용금액		3
9	CA_USAGE	현금서비스 소진율		3
10	SRT_DELQ_CNT	단기연체 경험건수		3
11	LONG_DELQ_CNT	장기연체 경험건수		3
12	MAX_DELQ_PERIOD	최장 연체일수		3
13	JOB	직업	categorical	2
14	OCC_AREA	지역		6
				44

되었다.

Table 3.3에는 각각의 분류기법에 대한 모의실험 결과가 소수집단에 대한 훈련개체의 비율별로 정리되어 있다. 모의실험 1과 마찬가지로 오분류 비용을 균일하게 적용하는 L_2-norm SVM과 L_1-norm SVM의 경우 소수집단에 대한 훈련개체의 비율이 낮아질수록 소수집단의 분류 정확도인 민감도가 급격히 낮아진 반면, 집단별 오분류 비용을 차등 적용한 WL_2-norm SVM과 WL_1-norm SVM은 소수집단의 비율이 낮아지더라도 민감도가 크게 하락하지 않았다. 또한, WL_2-norm SVM과 WL_1-norm SVM이 가중치를 적용하지 않은 방법론에 비해 민감도를 크게 개선하였으며, 그로 인해 전체 정확도와 기하평균의 측면에서 그 성능이 향상되었음을 알 수 있다. 특히, 제안 방법인 WL_1-norm SVM의 성능이 기존의 방법인 WL_2-norm SVM에 비해 우수한 성능을 나타내는 것을 확인할 수 있다. 변수 선택에 있어서도 제안 방법인 WL_1-norm SVM이 분류분석 기법들 중에서 잡음변수를 가장 많이 제거하였다. 비록 WL_1-norm SVM이 중요한 입력변수를 6개 중에서 평균적으로 5.75개, 5.04개, 4.16개 선택하였으나, 이는 실험모형이 입력요인의 3차 다항식으로 구성되어 있어서 한 요인에서 파생된 입력변수들 간의 상관관계가 매우 높기 때문인 것으로 판단된다 (Wang 등, 2006). 본 연구의 모의실험 과정에서 하나의 입력변수가 선택되고 나면 상관성이 높은 나머지 입력변수는 변수선택에서 제외되는 경향을 확인할 수 있었으며, Table 3.3으로부터 제안 방법인 WL_1-norm SVM이 실험모형에서 활용된 3가지 중요한 입력요인을 모두 선택하였음을 알 수 있다.

4. 실제자료 분석

이번 절에서는 대출승인 자료(loan approval data)를 활용하여 제안하는 WL_1-norm SVM과 기존 방법들의 성능을 비교 평가하였다. 이 자료는 2011~2012년 사이에 국내 은행의 대출 승인에 관한 데이터로 대출자 2,000명에 대한 14개의 입력요인과 우량 또는 불량을 나타내는 이항 범주형 반응변수로 구성되어 있다. 전체 자료는 대출자의 대출 실행 후 1년 간 정상적인 상환이 이루어진 1602명의 우량 대출자와 3회 차 이상의 연체가 발생한 398명의 불량 대출자로 구성되어 있으며, 두 집단 간의 개체수가 상당히 불균형적이다. 분류분석에서 사용한 입력요인은 대출 심사 시 대출 신청인이 제출한 신청서와 크레

Table 4.2. Simulation results for credit approval data

Method	Test classification accuracy				Number of selected factor	Number of selected variable
	Overall accuracy	Sensitivity	Specificity	G-mean		
L_2-norm SVM	0.897 (0.010)	0.718 (0.056)	0.941 (0.011)	0.821 (0.030)	14.00	43.98
WL_2-norm SVM	0.894 (0.009)	0.865 (0.038)	0.901 (0.010)	0.882 (0.018)	14.00	44.00
L_1-norm SVM	0.900 (0.009)	0.741 (0.057)	0.940 (0.011)	0.834 (0.029)	13.27	28.26
WL_1-norm SVM	0.900 (0.009)	0.902 (0.036)	0.900 (0.013)	0.901 (0.026)	12.37	24.73

The numbers in parentheses are standard deviations.

딧뷰로(credit bureau)로부터 수집하였으며, 이는 Table 4.1에 정리되어 있다. 비선형 분류함수의 추정을 위하여 표준화된 연속형 입력요인의 3차 다항식을 입력변수로 이용하였으며, 범주형 입력요인은 가변수(dummy variables)형태로 변환하여 입력변수로 활용하였다.

제안하는 WL_1-norm SVM과 기존의 방법인 L_2-norm SVM, WL_2-norm SVM, 그리고 L_1-norm SVM을 적용하여 대출승인 자료를 분석하였으며, 이때 모형의 적합 및 평가를 위해 전체 자료의 1/4을 훈련자료로, 1/4을 검증자료로, 그리고 나머지 1/2을 평가자료로 활용하였다. 분류기법들의 예측력 평가를 위해 전체 정확도, 민감도, 특이도, 그리고 기하평균을 계산하였으며, 변수선택의 성능을 평가하기 위해 14개의 입력요인 중 유의한 요인으로 선택된 입력요인의 개수와 44개의 입력변수 중 유의한 변수로 선택된 입력변수의 개수를 각각 계산하였다. 이러한 과정을 100번 독립적으로 반복하였으며, 각각의 평가지표에 대한 100번의 평균을 Table 4.2에 나타내었다.

Table 4.2로부터 집단 간의 오분류 비용을 차등 적용하기 위해 가중치를 이용하는 WL_2-norm SVM과 WL_1-norm SVM이 가중치를 이용하지 않는 방법론에 비해 비록 특이도가 다소 감소하지만, 금융분야에서 중요하게 다루어지는 민감도 측면에서 아주 우수한 성능을 나타내고 있으며, 이로 인하여 민감도와 특이도의 기하평균 또한 높게 나타고 있는 것을 확인할 알 수 있다. 특히, 제안 방법인 WL_1-norm SVM의 성능이 기존의 WL_2-norm SVM보다 우수하게 나타나는 것을 알 수 있다. 대출 신청자의 승인여부 결정에서 불량 신청자를 대출함으로써 발생하는 손실이 우량 신청자를 거절함으로써 발생하는 손실보다 상대적으로 크기 때문에 이 경우 정확도와 특이도는 좋은 평가지표가 될 수 없다. 이러한 관점에서 제안 방법인 WL_1-norm SVM이 기존의 방법론들에 비해서 그 성능이 매우 우수하며, 입력요인 및 입력변수의 선택에 있어서도 가장 간결한 모형을 제공하므로 실제 불균형 자료의 분류분석에서 그 활용가능성이 높다고 할 수 있겠다.

마지막으로 제안 방법인 WL_1-norm SVM을 이용하여 대출승인 자료 전체를 분석하였으며, 그 결과 연령과 지역을 나타내는 요인은 3차 다항식의 입력변수 모두가 분류함수에서 제외되었다. 이러한 두 요인은 일반적으로 금융의 승인결정에서 활용되지 않으므로 분류함수에서 제외되는 것이 타당한 것으로 판단된다.

5. 결론

일반적으로 L_2-norm SVM은 높은 수준의 분류 정확도와 유연성을 바탕으로 의학, 금융, 통신 등 여러 다양한 분야의 분류분석에서 널리 사용되고 있다. 그러나 릿지 벌칙함수의 특성으로 인하여 L_2-norm

SVM은 불필요한 잡음 변수들의 제거에 효율적이지 못하고, 또한 모형의 적합에서 집단별 오분류에 대하여 동일한 비용을 적용하므로 불균형 자료의 분류분석에는 부적절하다. 이러한 제한사항을 보완하기 위하여 본 논문에서는 라소 형태의 벌칙함수를 사용하여 변수선택의 기능을 유지함과 동시에 집단별 오분류 비용을 차등 적용함으로서 불균형 자료의 분류분석에 적합한 WL_1-norm SVM을 제안하였다. 모의실험과 실제자료의 분석을 통해 제안한 WL_1-norm SVM이 잡음변수가 포함되어 있는 불균형 자료의 분류분석에서 기존의 방법들 비해 분류 정확도와 모형의 간결성 측면에서 그 성능이 우수함을 확인하였다.

기존의 WL_2-norm SVM이 집단별로 가중치를 적용한 것에 반해, 본 논문에서 제안한 WL_1-norm SVM은 훈련개체별로 가중치를 적용하여 훈련개체 각각의 중요도를 구분할 수 있다. 따라서 다양한 샘플링 방법들에 의해 생성된 새로운 훈련개체를 이용할 경우, 원자료의 훈련개체와 생성된 훈련개체에 대한 가중치를 차등적으로 적용할 수 있다. 차후에는 샘플링 방법들과 결합하여 분류 정확도와 모형의 해석력을 향상시키는 새로운 SVM 기법이 개발되기를 기대해 본다.

References

Akbani, R., Kwek, S. and Japkowicz,, N. (2004). Applying support vector machines to imbalanced datasets, In *Proceedings of European Conference of Machine Learning*, **3201**, 39–50.

Bang, S. and Jhun, M. (2014). Weighted support vector machine using k-means clustering, *Communications in Statistics-Simulation and Computation*, **43**, 2307–2324.

Barandela, R., Sanchez, J., Garcia, V. and Rangel, E. (2003). Strategies for learning in class imbalance problems, *Pattern Recognition*, **36**, 849–851.

Berkelaar, M. and others (2014). lpSolve: Interface to Lp_solve v. 5.5 to solve linear/integer programs. R package version 5.6.10. http://CRAN.R-project.org/package=lpSolve.

Chawla, N., Bowyer, K., Hall, L. and Kegelmeyer, W. (2002). SMOTE: Synthetic minority over-sampling technique, *Journal of Artificial Intelligence Research*, **16**, 321–357.

Cohen, G., Hilario, M., Sax, H., Hugonnet, S. and Geissbuhler, A. (2006). Learning from imbalanced data in surveillance of nosocomial infection, *Artificial Intelligence in Medicine*, **37**, 7–18.

Cortes, C. and Vapnik, V. (1995). Support vector networks, *Machine Learning*, **20**, 273–297.

Ganganwar, V. (2012). An overview of classification algorithms for imbalanced datasets, *International Journal of Emerging Technology and Advanced Engineering*, **2**, 42–47.

Garcia, V., Sanchez, J. S., Mollineda, R. A., Alejo, R. and Sotoca, J. M. (2007). The class imbalance problem in pattern classification and learning, In *Proceedings of the 5th Spanish Workshop on Data Mining and Learning*, 283–291.

Han, H., Wang, W. Y. and Mao, B. H. (2005). Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In *Proceedings of the International Conference on Intelligent Computing*, **3644**, 878–887.

Hoerl, A. and Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics*, **12**, 55–67.

Japkowicz, N. (2000). The Class imbalance problem; Significance and Strategies, In *Proceedings of the 2000 International Conference on Artificial Intelligence : Special Track on Inductive Learning*, **1**, 111–117

Kim, J. and Jeong, J. (2004). Classification of class-imbalanced data: Effect of over-sampling and under-sampling of training data, *The Korean Journal of Applied Statistics*, **17**, 445–457.

Kubat M. and Matwin S. (1997). Adressing the curse of imbalanced training sets: One-sided selection, In *Proceedings of the Fourteenth International Conference on Machine Learining*, 179–186.

Lee, H. and Lee, S. (2014). A comparison of ensemble methods combining resampling techniques for class imbalanced data, *The Korean Journal of Applied Statistics*, **27**, 357–371.

Lin, Y., Lee, Y. and Wahba, G. (2002). Support vector machines for classification in nonstandard situations, *Machine Learning*, **46**, 191–202.

Liu, Y., An, A. and Huang, X. (2006). Boosting prediction accuracy on imbalanced datasets with SVM ensembles, In *Proceedings of the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, **3918**, 107–118.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Tang, Y., Zhang, Y., Chawla, N. and Krasser, S. (2009). SVMs modeling for highly imbalanced classification, *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, **39**, 281–288.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society, Series B*, **58**, 267–288.

Turlach, B. and Weingessel, A. (2013). quadprog: Functions to solve quadratic programming problems. R package version 1.5-5. http://CRAN.R-project.org/package=quadprog.

Vapnik, V. N, (1998). *Statistical Learning Theory*, Wiley, New York.

Veropoulos, K., Campbell, C. and Cristianini, N. (1999). Controlling the sensitivity of support vector machines, In *Proceedings of the International Joint Conference on AI*, 55–60.

Wang, B. X. and Japkowicz, N. (2009). Boosting support vector machines for imbalanced data sets, *Knowledge and Information Systems*, **25**, 1–20.

Wang, L., Zhu, J. and Zou, H. (2006). The doubly regularized support vector machine, *Statistica Sinica*, **16**, 589–615.

Zhu, J., Rosset, S., Hastiem, T. and Tibshirani, R. (2003). 1-norm support vector machine, *Neural Information Proceeding Systems*, **16**, 49–56.

불균형 자료의 분류분석을 위한 가중 L_1-norm SVM

김은경[a] · 전명식[a] · 방성완[b,1]

[a]고려대학교 통계학과, [b]육군사관학교 수학과

(2014년 9월 18일 접수, 2014년 11월 12일 수정, 2015년 1월 13일 채택)

요 약

SVM은 높은 수준의 분류 정확도와 유연성을 바탕으로 다양한 분야의 분류분석에서 널리 사용되고 있다. 그러나 집단별 개체수가 상이한 불균형 자료의 분류분석에서 SVM은 다수집단으로 편향되게 분류함수를 추정하므로 소수집단의 분류 정확도가 심각하게 감소하게 된다. 불균형 자료의 분류분석을 위하여 집단별 오분류 비용을 차등 적용하는 가중 L_2-norm SVM이 개발되었으나, 이는 릿지 형태의 벌칙함수를 사용하므로 분류함수의 추정에서 불필요한 잡음변수의 제거에는 효율적이지 못하다. 따라서 본 논문에서는 라소 형태의 벌칙함수를 사용하고 훈련개체의 오분류 비용을 차등적으로 부여함으로서 불균형 자료의 분류분석에서 변수선택의 기능을 지니는 가중 L_1-norm SVM을 제안하였으며, 모의실험과 실제자료의 분석을 통하여 제안한 방법론의 효율적인 성능과 유용성을 확인하였다.

주요용어: 불균형 자료, 라소, 선형계획법, 릿지, 서포트 벡터 머신.

본 논문은 육군사관학교 화랑대연구소의 2014년도(20140516) 연구활동지원에 의해 출간되었으며(방성완), 2013년도 정부(교육부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임(NRF-2013R1A1A2A10007545)(전명식).

[1]교신저자: (139-799) 서울시 노원구 화랑로 574, 육군사관학교 수학과. E-mail: wan1365@hanmail.net