The Korean Journal of Applied Statistics (2016) DOI: http://dx.doi.org/10.5351/KJAS.2016.29.5.961
29(5), 961–975

Hierarchically penalized support vector machine for the classification of imbalanced data with grouped variables

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

[a]Research Center, Korea Credit Bureau; [b]Department of Statistics, Korea University;
[c]Department of Mathematics, Korea Military Academy

(Received June 9, 2016; Revised July 6, 2016; Accepted July 7, 2016)

Abstract

The hierarchically penalized support vector machine (H-SVM) has been developed to perform simultaneous classification and input variable selection when input variables are naturally grouped or generated by factors. However, the H-SVM may suffer from estimation inefficiency because it applies the same amount of shrinkage to each variable without assessing its relative importance. In addition, when analyzing imbalanced data with uneven class sizes, the classification accuracy of the H-SVM may drop significantly in predicting minority class because its classifiers are undesirably biased toward the majority class. To remedy such problems, we propose the weighted adaptive H-SVM (WAH-SVM) method, which uses a adaptive tuning parameters to improve the performance of variable selection and the weights to differentiate the misclassification of data points between classes. Numerical results are presented to demonstrate the competitive performance of the proposed WAH-SVM over existing SVM methods.

Keywords: adaptive tuning parameter, hierarchical penalization, imbalanced data, support vector machine, variable selection

1. 서론

두 집단의 개체수가 상이한 불균형 자료(imbalanced data)는 이상거래탐지(fraud detection), 희귀병 진단(medical diagnosis of rare diseases), 이동통신 이탈탐지(churn) 등 실제 분류분석의 사례에서 자주 접하는 자료형태이다. 불균형 자료의 분류분석에서는 일반적으로 개체수가 많은 다수집단(majority class)보다 개체수가 작은 소수집단(minority class)의 오분류 손실이 더 크며, 그로 인해 소수집단의 분류 정확도에 대한 중요성이 더 강조된다. 그러나 불균형 자료의 분석에서 일반적인 분류기법을 적용할 경우 전체 정확도를 향상시키기 위해 분류함수를 다수집단으로 편향되게 추정하므로 소수집단의 분류 정확도가 현저히 감소하게 된다. 소수집단의 분류 정확도를 향상시키기 위한 대표적인 방법에는 가중치를 이용하여 소수집단의 오분류 비용을 증가시키는 오분류 비용의 차등적용 방법과 균형된 자료로 만들

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by (1) the Ministry of Science, ICT & Future Planning (NRF-2015R1C1A1A02036473) for S. Bang and (2) the Ministry of Education (NRF-2013R1A1A2A10007545) for M. Jhun.
[1]Corresponding author: Department of Mathematics, Korea Military Academy, 574 Hwarang-ro, Nowon-gu, Seoul 01805, Korea. E-mail: wan1365@gmail.com

기 위해 개체수를 인위적으로 조정하는 샘플링 방법이 있다. Veropoulos 등 (1999), Lin 등 (2002), 그리고 Akbani 등 (2004)은 가중치를 이용하여 오분류 비용을 차등 적용하는 방법에 대하여 연구하였으며, Kubat과 Matwin (1997), Japkowicz (2000), Chawla 등 (2002), Tang 등 (2009)은 과소추출과 과대추출을 통해 개체수를 균형있게 조정하는 샘플링 방법에 대한 다양한 연구를 진행하였다. 각 방법들의 장·단점과 성능을 이론적으로 비교분석하기는 어렵다. 이들의 제한사항을 살펴보면, 오분류 비용의 차등적용 방법은 소수집단의 개체수가 극히 적을 경우 예측편향으로 인한 과적합과 분포 왜곡이 발생할 수 있으며 (Domingos, 1999; Akbani 등, 2004), 과소추출 방법은 원 자료가 가지고 있는 정보를 손실하게 되어 분류정확도가 하락할 수 있다 (Chawla 등, 2002; Tang 등, 2009). 반면에, 과대추출 방법은 과적합과 계산시간의 부하를 줄 수 있는 등 (Chawla 등, 2002; Kotsiantis 등, 2006)의 이슈가 있다.

Cortes와 Vapnik (1995), Vapnik (1998) 등에 의해 제안된 support vector machine(SVM)은 높은 분류 정확도와 유연성을 바탕으로 분류분석에서 널리 사용되고 있는 기법 중 하나이다. 이항 범주형 반응변수 $y_i \in \{-1, 1\}$와 입력변수 $\boldsymbol{x}_i \in R^p$로 이루어진 훈련자료 $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$에 근거하여, 선형 분류함수 $f(\boldsymbol{x}) = \beta_0 + \boldsymbol{x}^T\boldsymbol{\beta}$를 추정하는 문제를 고려하자. SVM은 훈련자료들의 마진(margin)을 최대로 하는 최적화 식

$$\left(\hat{\beta}_0, \hat{\boldsymbol{\beta}}\right)^{L_2\text{-SVM}} = \arg\min_{\beta_0, \boldsymbol{\beta}} \frac{1}{||\boldsymbol{\beta}||_2} \quad \text{subject to} \quad y_i\left\{\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}\right\} \geq 1 - \xi_i,\ \xi_i \geq 0,\ \forall i, \text{ and } \sum_{i=1}^n \xi_i \leq s \tag{1.1}$$

을 통해 분류함수를 추정하며, 이때 ξ_i는 여유변수(slack variables)이고 $s \geq 0$는 축소추정의 정도를 나타내는 조율모수이다. 최적화 식 (1.1)은 손실함수에 릿지 형태의 벌칙함수가 적용된 적합식

$$\left(\hat{\beta}_0, \hat{\boldsymbol{\beta}}\right)^{L_2\text{-SVM}} = \arg\min_{\beta_0, \boldsymbol{\beta}} \sum_{i=1}^n \left[1 - y_i\left(\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}\right)\right]_+ + \lambda||\boldsymbol{\beta}||_2^2 \tag{1.2}$$

으로 표현 가능하므로 L_2-norm SVM이라 불리기도 한다. 여기서 경첩 손실함수 $[t]_+ = \max(t, 0)$이고 $\lambda > 0$는 훈련자료의 오차와 벌칙항간의 균형을 맞추어 과대적합을 방지하는 조율모수로 식 (1.1)의 s와 일대일로 대응된다.

SVM은 분류함수 근처의 훈련개체인 서포트 벡터(support vector)만을 분류함수의 추정에 사용하므로 분류함수로부터 멀리 떨어진 다수집단의 많은 개체는 함수의 추정에 영향을 주지 않는다. 따라서 다른 분류 기법들과 비교할 때 SVM은 불균형 자료의 분류 분석에서 비교적 강건한 성능을 나타낸다. 그러나 불균형의 정도가 심해짐에 따라 SVM 또한 다른 분류기법들과 마찬가지로 소수집단에 대한 분류 정확도가 크게 감소하게 된다. Veropoulos 등 (1999)과 Akbani 등 (2004)은 불균형 자료의 분석에서 소수집단에 대한 분류 정확도를 향상시키기 위하여 L_2-norm SVM에 집단별로 가중치를 적용하는 WL_2-norm SVM을 제안하였으며, 그 적합식은

$$\left(\hat{\beta}_0, \hat{\boldsymbol{\beta}}\right)^{WL_2\text{-SVM}} = \arg\min_{\beta_0, \boldsymbol{\beta}} \frac{1}{2}||\boldsymbol{\beta}||_2^2 + C^+ \sum_{\{i|y_i=+1\}}^n \xi_i + C^- \sum_{\{i|y_i=-1\}}^n \xi_i \tag{1.3}$$

$$\text{subject to} \quad y_i\left\{\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}\right\} \geq 1 - \xi_i,\ \xi_i \geq 0,\ \forall\ i$$

와 같다. 여기서 C^+와 C^-는 각각 소수집단(+)과 다수집단(−)의 오분류에 대한 비용을 나타내고, $C^+ = C^- = 1/(2\lambda)$일 때 적합식 (1.3)은 적합식 (1.2)와 동일하게 된다. 불균형 자료의 분류분석에서 적합식 (1.3)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.

불균형 자료와 더불어 분류분석에서는 고차원 자료를 자주 접하게 되며, 이러한 고차원 자료의 분석에서는 분류함수의 예측력과 해석력의 향상을 위해 잡음변수를 제거하고 중요한 입력변수만을 모형에 포함해야 한다. 고차원 자료의 분류분석에서 L_2-norm SVM과 WL_2-norm SVM은 릿지 벌칙함수의 특성으로 인하여 변수들의 동시적인 선택이 불가능하며, 이로 인해 분류 정확도가 감소하고 모형의 해석이 어려워진다. 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

$$\left(\hat{\beta}_0, \hat{\boldsymbol{\beta}}\right)^{L_1\text{-SVM}} = \arg\min_{\beta_0, \boldsymbol{\beta}} \sum_{i=1}^{n} \left[1 - y_i \left(\beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta}\right)\right]_+ + \lambda ||\boldsymbol{\beta}||_1 \tag{1.4}$$

와 같이 제안하였고, Zou (2007)는 식 (1.4)의 라소 벌칙함수에 적응적 조율모수(adaptive lasso) (Zou, 2006)를 적용한 AL_1-norm SVM을

$$\left(\hat{\beta}_0, \hat{\boldsymbol{\beta}}\right)^{AL_1\text{-SVM}} = \arg\min_{\beta_0, \boldsymbol{\beta}} \sum_{i=1}^{n} \left[1 - y_i \left(\beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta}\right)\right]_+ + \lambda \sum_{j=1}^{p} \left|\hat{\beta}_j^{L_2}\right|^{-r} |\beta_j| \tag{1.5}$$

와 같이 제안하였다. 여기서 $\hat{\beta}_j^{L_2}$는 L_2-norm SVM의 추정값이고 $r > 0$은 사전에 지정된 상수이다. 라소 벌칙함수는 회귀계수 β_j $(j = 1, \ldots, p)$를 0 방향으로 축소 추정함과 동시에 조율모수 λ가 충분히 클 때 불필요한 입력변수의 회귀계수를 0으로 정확하게 추정함으로써 잡음변수를 모형에서 제거하게 된다. 또한 불균형 자료의 분석에서 소수집단에 대한 분류 정확도를 향상시키기 위해 Kim 등 (2015)은 L_1-norm SVM에 훈련개체별로 가중치(weight)를 적용하는 WL_1-norm SVM을

$$\left(\hat{\beta}_0, \hat{\boldsymbol{\beta}}\right)^{WL_1\text{-SVM}} = \arg\min_{\beta_0, \boldsymbol{\beta}} \sum_{i=1}^{n} c_i \left[1 - y_i \left(\beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta}\right)\right]_+ + \lambda ||\boldsymbol{\beta}||_1 \tag{1.6}$$

와 같이 제안하였다. 여기서 c_i는 i번째 훈련개체의 오분류에 대한 비용을 나타내며, 집단별로 오분류 비용이 적용되는 WL_2-norm SVM과 달리 적합식 (1.6)은 훈련개체 각각의 중요도에 따라 오분류 비용을 구분할 수 있으므로 불균형 자료의 분류분석을 위하여 다수집단의 훈련개체를 과소추출하거나 소수집단의 훈련개체를 과대 추출 또는 생성하는 다양한 방법론과의 결합이 용이하다.

본 논문에서는 두 집단의 개체수가 상이한 불균형 자료에서 고차원의 입력변수들이 그룹화 되어 있거나 특정 요인(factor)에서 의해 파생되어진 경우를 고려하였다. 이러한 자료구조에서는 개별 입력변수 뿐만 아니라 그룹(group) 또는 요인의 중요성도 함께 고려되어야 한다. 예를 들어, 범주형 입력변수가 여러 개의 가변수(dummy variable) 형태로 모형의 적합에 활용되거나, 다항식(polynomial) 형태의 가법모형(additive model)에서 기저함수(basis function) 등이 이에 해당된다. 여기서는 "그룹"과 "요인"의 용어를 동일한 의미로 사용하였다. 입력변수들이 그룹화 되어있는 경우에는 개별 입력변수의 선택보다는 입력변수들의 그룹특성을 대표하는 공통요인의 선택이 더 중요하다 (Yuan과 Lin, 2006). SVM 방법론에서 그룹별 변수선택에 관한 연구로는 F_∞-norm SVM (Zou와 Yuan, 2008a)과 H-SVM (Bang 등, 2016) 등이 있으며, 특히 H-SVM은 그룹별 변수 선택에만 중점을 둔 다른 방법들에 비해 그룹과 그룹 내 입력변수의 동시적인 선택이 가능하여 예측력과 모형의 간결성 측면에서 그 성능이 우수한 분류분석 기법이다. 그러나 고차원 불균형 자료의 분류분석에서 H-SVM의 직접적인 활용은 제한적이다.

따라서 본 논문에서는 소수집단의 분류 정확도를 향상시키기 위하여 H-SVM의 적합식에 오분류 비용을 차등 적용하고, 벌칙항의 조율모수를 적응적으로 부여하여 그룹과 그룹내 입력변수의 선택에서 효율적인 WAH-SVM 기법을 제안하였다. 본 논문의 구성은 다음과 같다. 2절에서는 먼저 분류분석에서 그룹화된 입력변수의 선택을 위한 기존의 F_∞-norm SVM과 H-SVM 방법론을 소개하고, 이어서 H-SVM에 적응적 조율모수와 오분류에 대한 개체별 가중치를 적용한 WAH-SVM을 제안하였다. 3절과 4절에서는 모의실험과 실제자료 분석을 통해 기존의 분류기법과 제안한 WAH-SVM의 성능을 비교하였으며, 제안 방법론의 활용가능성을 보였다. 마지막으로 5절에서는 결론과 더불어 차후 연구방향을 제시하였다.

2. 고차원 불균형 자료의 분류분석을 위한 WAH-SVM

2.1. 기존 SVM 방법론의 비교분석: F_∞-norm SVM과 H-SVM

p개의 입력변수가 K개의 그룹으로 나누어져 있는 경우 입력변수는 $\boldsymbol{x} = (\boldsymbol{x}_{(1)}^T, \ldots, \boldsymbol{x}_{(K)}^T)^T$와 같이 표현 가능하며, k $(k = 1, \ldots, K)$번째 그룹은 p_k개의 입력변수 $\boldsymbol{x}_{(k)} = (\boldsymbol{x}_{k1}, \ldots, \boldsymbol{x}_{kp_k})^T$로 표현된다. 이 경우 분류함수는 $f(\boldsymbol{x}) = \beta_0 + \sum_{k=1}^K \boldsymbol{x}_{(k)}^T \boldsymbol{\beta}_{(k)}$으로 나타낼 수 있고, 여기서 $\boldsymbol{\beta}_{(k)} = (\beta_{k1}, \ldots, \beta_{kp_k})^T$는 k번째 그룹의 계수 벡터이다.

분류함수의 추정에서 입력변수들이 그룹화 되어 있는 경우에는 개별 입력변수 뿐만 아니라 그룹의 중요성도 함께 고려되어야 한다. 그러나 식 (1.4)의 L_1-norm SVM은 입력변수들을 개별적으로 선택하는 방법론이므로 그룹변수들의 동시적인 선택에서는 그 유용성이 떨어진다. 그룹별 변수선택을 위해 Zou와 Yuan (2008)은 sup-norm 벌칙함수를 이용한 F_∞-norm SVM을 제안하였으며, 그 적합식은

$$\left(\hat{\beta}_0, \hat{\boldsymbol{\beta}}\right)^{F_\infty\text{-SVM}} = \arg\min_{\beta_0, \boldsymbol{\beta}} \sum_{i=1}^n \left[1 - y_i \left(\beta_0 + \sum_{k=1}^K \boldsymbol{x}_{i,(k)}^T \boldsymbol{\beta}_{(k)}\right)\right]_+ + \lambda \sum_{k=1}^K ||\boldsymbol{\beta}_{(k)}||_\infty \tag{2.1}$$

와 같다. 여기서 벌칙함수는 $||\boldsymbol{\beta}_{(k)}||_\infty = \max\{|\beta_{k1}|, \ldots, |\beta_{kp_k}|\}$ $(k = 1, \ldots, K)$으로 정의된다. 그룹내 입력변수 중 최대값을 통제하는 벌칙함수의 특성으로 인해 F_∞-norm SVM은 유의한 그룹 변수를 동시에 선택할 수 있다. 그러나 이 경우 중요 그룹으로 선택되면 그룹 내 입력변수들의 중요도에 상관없이 모든 입력변수를 선택하는 한계가 있다.

이러한 F_∞-norm SVM의 제한사항을 보완하기 위하여 Bang 등 (2016)은 그룹의 정보를 나타내는 변수 $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K)^T$와 그룹 내 입력변수의 정보를 위한 변수 $\boldsymbol{\theta}_{(k)} = (\theta_{k1}, \ldots, \theta_{kp_k})^T$를 이용하여 계수 β_{kj}를 $\beta_0 = \theta_0$, $\beta_{kj} = \gamma_k \theta_{kj}$ $(k = 1, \ldots, K,\ j = 1, \ldots, p_k)$으로 재모수화 하였으며, 이를 바탕으로 계층적 벌칙함수를 이용한 Hierarchically penalized SVM(H-SVM)을

$$\left(\hat{\boldsymbol{\gamma}}, \hat{\theta}_0, \hat{\boldsymbol{\theta}}\right)^{H\text{-SVM}} = \arg\min_{\boldsymbol{\gamma}, \theta_0, \boldsymbol{\theta}} \sum_{i=1}^n \left[1 - y_i \left(\theta_0 + \sum_{k=1}^K \gamma_k \boldsymbol{x}_{i,(k)}^T \boldsymbol{\theta}_{(k)}\right)\right]_+ + \sum_{k=1}^K \gamma_k + \lambda \sum_{k=1}^K ||\boldsymbol{\theta}_{(k)}||_1 \tag{2.2}$$

와 같이 제안하였다. 여기서 γ_k는 계층의 첫 번째 단계에서 k번째 그룹에 속하는 모든 β_{kj} $(j = 1, \ldots, p_k)$를 제어하는 그룹효과를 반영하고, θ_{kj}는 계층의 두 번째 단계에서 k번째 그룹에 속한 변수들의 영향력 차이를 반영하게 되어, 결과적으로 그룹과 그룹내 변수를 동시에 선택할 수 있게 된다. 즉, F_∞-norm SVM은 중요한 그룹으로 선택된 경우 그룹 내 모든 입력변수를 선택하게 되는 단점이 있는 반면, H-SVM은 중요한 그룹의 선택 뿐만 아니라 그룹 내에서 중요한 입력변수를 동시에 선택하는 장점을 가지고 있다.

2.2. 오분류 가중치와 적응적 조율모수 이용한 H-SVM(WAH-SVM)

앞 절에서 소개된 SVM 방법론의 벌칙함수에는 입력변수 또는 그룹에 라는 동일한 조율모수를 부여하고 있다. 이처럼 입력변수의 상대적인 중요도에 관계없이 동일한 강도로 계수를 축소 추정하는 경우에는 추정의 효율성이 감소될 수 있다 (Fan과 Li, 2001; Yuan과 Lin, 2006). 따라서 변수선택의 성능을 향상시키기 위해 입력변수의 영향력이 강하다면 계수 추정 시 약한 벌점을 주고, 반대로 입력변수의 영향력이 약하다면 강한 벌점을 주는 적응적(adaptive) 조율모수를 고려할 수 있을 것이다. 입력변수들이 그룹화 되어 있는 경우 그룹별 변수선택의 성능을 향상시키기 위해 Bang과 Jhun (2012)은 조율모수를 적응적으로 부여한 Adaptive F_∞-norm SVM(AF_∞-norm SVM)을 제안하였으며, 그 적합식은

$$\left(\hat{\beta}_0, \hat{\boldsymbol{\beta}}\right)^{AF_\infty\text{-SVM}} = \arg\min_{\beta_0,\boldsymbol{\beta}} \sum_{i=1}^{n} \left[1 - y_i \left(\beta_0 + \sum_{k=1}^{K} \boldsymbol{x}_{i,(k)}^T \boldsymbol{\beta}_{(k)}\right)\right]_+ + \lambda \sum_{k=1}^{K} \left|\left|\hat{\boldsymbol{\beta}}_{(k)}^{L_2}\right|\right|_\infty^{-r} \left|\left|\boldsymbol{\beta}_{(k)}\right|\right|_\infty \quad (2.3)$$

와 같다. 여기서 조율모수는 기존 λ 대신 그룹별로 그 중요도를 달리하는 적응적 조율모수 $\lambda||\hat{\boldsymbol{\beta}}_{(k)}^{L_2}||_\infty^{-r}$를 사용하였으며, $r > 0$은 사전에 지정된 상수이다.

H-SVM은 입력변수가 그룹화 되어 있는 경우에 개별적인 변수선택 보다는 그룹 및 그룹 내의 변수선택을 동시에 할 수 있는 SVM 방법론이다. 그러나, H-SVM은 적합식 (2.2)에서 보는 바와 같이 모든 입력변수에 동일한 조율모수 λ를 사용하므로 입력변수들의 중요도에 상관없이 동일한 강도로 계수들을 축소 추정하게 된다. 중요한 입력변수를 과도하게 축소시킬 경우 최종 모형에서 제외되거나 추정에서 왜곡되는 경우가 발생할 수 있으며, 그로인해 분류 정확도가 감소 될 수 있다. H-SVM의 보다 효율적인 추정을 위해 본 논문에서는 먼저 Zou (2006)와 유사한 방법으로 적응적 조율모수를 부여하여 입력변수별로 그 중요도를 구분하는 Adaptive H-SVM(AH-SVM)을

$$\left(\hat{\boldsymbol{\gamma}}, \theta_0, \hat{\boldsymbol{\theta}}\right)^{AH\text{-SVM}} = \arg\min_{\boldsymbol{\gamma},\theta_0,\boldsymbol{\theta}} \sum_{i=1}^{n} \left[1 - y_i \left(\theta_0 + \sum_{k=1}^{K} \gamma_k \boldsymbol{x}_{i,(k)}^T \boldsymbol{\theta}_{(k)}\right)\right]_+ + \sum_{k=1}^{K} \gamma_k + \lambda \sum_{k=1}^{K} \sum_{j=1}^{p_k} \left|\hat{\theta}_{kj}^{L_2}\right|^{-r} |\theta_{kj}| \quad (2.4)$$

와 같이 제안하고자 한다. 여기서 $\hat{\theta}_{kj}^{L_2}$는 L_2-norm SVM의 추정값이고 $r > 0$은 사전에 지정된 상수이다. 적합식 (2.4)의 AH-SVM은 중요하지 않은 변수에는 강한 벌점을 부여하고 반대로 중요한 변수에는 약한 벌점을 부여함으로써 추정과 변수선택에서 그 효율성을 향상시키게 된다. 그러나 AH-SVM을 불균형 자료의 분류분석에 그대로 적용하면 다른 방법론들과 마찬가지로 편향된 분류함수를 추정하게 된다. 따라서 본 논문에서는 소수집단의 분류 정확도를 향상시키기 위해 Kim 등 (2015)이 오분류 비용을 차등적으로 적용하기 위해 사용한 가중치를 이용하여 Weighted AH-SVM(WAH-SVM)을

$$\left(\hat{\boldsymbol{\gamma}}, \theta_0, \hat{\boldsymbol{\theta}}\right)^{WAH\text{-SVM}} = \arg\min_{\boldsymbol{\gamma},\theta_0,\boldsymbol{\theta}} \sum_{i=1}^{n} c_i \left[1 - y_i \left(\theta_0 + \sum_{k=1}^{K} \gamma_k \boldsymbol{x}_{i,(k)}^T \boldsymbol{\theta}_{(k)}\right)\right]_+ + \sum_{k=1}^{K} \gamma_k + \lambda \sum_{k=1}^{K} \sum_{j=1}^{p_k} \left|\hat{\theta}_{kj}^{L_2}\right|^{-r} |\theta_{kj}| \quad (2.5)$$

와 같이 최종 방법론으로 제안하고자 한다. 여기서 c_i는 i번째 훈련개체의 오분류에 대한 비용을 나타낸다. 소수집단의 오분류에는 다수집단에 비해 상대적으로 큰 가중치를 부여하여 소수집단의 분류정확도를 향상시킬 수 있으며, 소수집단 또는 다수집단 내에서도 서로 다른 가중치 부여가 가능하다.

2.3. WAH-SVM의 계산 알고리즘

WAH-SVM의 적합식 (2.5)는 비선형 계획법(nonlinear programming; NLP)으로 공식화되므로 입력변수가 많아질수록 계산상의 어려움이 존재하며 시간적 측면에서 비효율적이다. 따라서 γ_k와 θ_{kj}는 선형 계획법(linear programming; LP)으로 구성되는 다음의 반복 알고리즘을 이용하여 추정하게 된다.

단계 0. 초기값 $\hat{\gamma}_k^{(0)}$을 지정하고 $t=1$이라 하자. 본 연구에서는 $k=1,\ldots,K$에 대하여 $\hat{\gamma}_k^{(0)}=1$로 지정하였다.

단계 1. t번째 반복에서 $\tilde{\boldsymbol{x}}_{i,(k)}=\hat{\gamma}_k^{(k-1)}\boldsymbol{x}_{i,(k)}$ $(k=1,\ldots,K)$로 정의하고 $(\hat{\theta}_0^{(t)},\hat{\boldsymbol{\theta}}^{(t)})$를 다음과 같이 추정한다.

$$\left(\hat{\theta}_0^{(t)},\hat{\boldsymbol{\theta}}^{(t)}\right)=\arg\min_{\theta_0,\boldsymbol{\theta}}\sum_{i=1}^{n}c_i\left[1-y_i\left(\theta_0+\sum_{k=1}^{K}\tilde{\boldsymbol{x}}_{i,(k)}^T\boldsymbol{\theta}_{(k)}\right)\right]_+ +\lambda\sum_{k=1}^{K}\sum_{j=1}^{p_k}\left|\hat{\theta}_{kj}^{L_2}\right|^{-r}|\theta_{kj}|. \quad (2.6)$$

단계 2. $u_{ik}=\boldsymbol{x}_{i,(k)}^T\hat{\boldsymbol{\theta}}_{(k)}^{(t)}$ $(k=1,\ldots,K)$로 정의하고 $\boldsymbol{\gamma}^{(t)}$를 다음과 같이 추정한다.

$$\hat{\boldsymbol{\gamma}}^{(t)}=\arg\min_{\boldsymbol{\gamma}\geq 0}\sum_{i=1}^{n}c_i\left[1-y_i\left(\theta_0^{(t)}+\sum_{k=1}^{K}u_{i,k}\gamma_k\right)\right]_+ +\sum_{k=1}^{K}\gamma_k. \quad (2.7)$$

단계 3. $\hat{\theta}_0^{(t)}$, $\hat{\boldsymbol{\theta}}^{(t)}$, $\hat{\boldsymbol{\gamma}}^{(t)}$가 수렴하면 반복 알고리즘을 종료하고 계수를 $\hat{\beta}_0=\hat{\theta}_0^{(t)}$와 $\hat{\boldsymbol{\beta}}_{(k)}=\hat{\gamma}_k^{(t)}\hat{\boldsymbol{\theta}}_{(k)}^{(t)}$으로 추정한다. 그렇지 않으면 $t\leftarrow t+1$로 두고 단계 1부터 다시 반복한다.

식 (2.6)과 (2.7)의 최적화 문제는 각각 적응적 조율모수를 사용한 라소와 Nonnegative garrote의 형태로서 여유변수(slack variable)를 이용하여 선형 계획법으로 공식화 될 수 있다 (Breiman, 1995; Kim 등, 2015). 또한 Wang 등 (2009), Zhou와 Zhu (2010), Bang 등 (2016)에서 보인 바와 같이, 단계 1에서 단계 3을 반복함에 따라 식 (2.5)의 목적함수 값이 감소하기 때문에 항상 수렴된 추정값을 얻을 수 있다. 본 논문에서는 선형계획법 문제의 최적해를 구하기 위해 R 프로그램 (R Core Team, 2014)의 lpSolve 패키지 (Berkelaar 등, 2014)에 포함되어 있는 lp 함수를 사용하였다. 또한 L_2-norm SVM의 최적해는 quadprog 패키지 (Turlach와 Weingessel, 2013)에 포함되어 있는 solve.QP 함수를 사용하였다. 입력변수의 차원과 훈련자료의 개체수가 커짐에 따라 Rmosek 패키지 (Friberg, 2013)나 FIRST (Hwang 등, 2009) 등과 같은 대규모(large scale) 최적화 문제(optimization problem)에 적합한 계산 알고리즘을 활용할 수 있을 것이며, 나아가 효율적인 계산을 위하여 solution path (Zhu 등, 2003)로 구현할 수 있을 것이다.

3. 모의 실험

이항 범주형 자료의 분류분석에서 분류 정확도를 비교 평가하기 위한 지표로 Kim 등 (2015)에서 사용한 전체정확도(overall accuracy), 민감도(sensitivity), 특이도(specificity) 그리고 기하평균(g-mean)을 사용하였다. 불균형 자료의 경우에는 일반적으로 널리 사용되는 전체정확도 보다 소수집단의 예측력을 평가할 수 있는 민감도와 기하평균이 더 의미 있는 평가지표가 될 수 있다. 본 논문에서 제안하는 WAH-SVM의 성능을 평가하기 위해 L_1-norm SVM, F_∞-norm SVM, H-SVM에 적응적 조율모수 또는 오분류 가중치를 적용한 방법론들과 그 성능을 비교하였다. 그리고 불균형 자료의 분류분석에서 오분류 비용에 대한 가중치는 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 선택

Table 3.1. Simulation results for Example 1 (percentage of minority class: 10%)

Weight (Y/N)	Adaptive (Y/N)	Method	Test classification accuracy(%)				Group selection		Input variable selection	
			Overall accuracy	Sensitivity	Specificity	G-mean	NC	NIC	NC	NIC
N	N	L_1	80.0 (1.2)	61.9 (2.7)	98.0 (0.4)	77.9 (1.6)	3.0	5.0	10.0	27.7
		F_∞	80.1 (1.3)	62.4 (2.7)	97.8 (0.5)	78.1 (1.6)	3.0	5.0	10.0	30.0
		H	80.0 (1.2)	61.8 (2.7)	98.2 (0.5)	77.9 (1.5)	3.0	4.0	10.0	21.9
	Y	AL_1	80.2 (1.3)	61.9 (2.7)	98.5 (0.4)	78.0 (1.6)	3.0	4.6	10.0	13.1
		AF_∞	80.4 (1.2)	62.6 (2.6)	98.3 (0.5)	78.4 (1.6)	3.0	4.8	10.0	28.8
		AH	80.2 (1.2)	61.9 (2.6)	98.6 (0.4)	78.1 (1.6)	3.0	2.8	10.0	8.9
Y	N	WL_1	87.0 (0.8)	85.0 (2.0)	89.0 (1.7)	87.0 (0.8)	3.0	4.5	9.8	14.2
		WF_∞	86.5 (0.7)	84.0 (1.9)	89.0 (1.5)	85.6 (0.7)	3.0	4.8	10.0	28.9
		WH	87.9 (0.6)	86.6 (1.9)	89.2 (1.4)	87.9 (0.6)	3.0	0.5	9.9	4.0
	Y	WAL_1	87.8 (0.7)	86.1 (1.9)	89.5 (1.4)	87.7 (0.7)	3.0	2.6	9.9	4.6
		WAF_∞	87.7 (0.6)	85.9 (1.9)	89.5 (1.3)	87.7 (0.7)	3.0	2.4	10.0	16.8
		WAH	88.0 (0.6)	86.7 (1.9)	89.4 (1.4)	88.0 (0.6)	3.0	0.3	9.9	1.2

The numbers in parentheses are standard deviations.
NC = number of correctly selected input variable; NIC = number of incorrectly selected input variable.

하기 위해 크기가 1,000인 검증자료(validation data)와 분류기법들의 분류 정확도를 평가하기 위해 크기가 10,000인 평가자료(test data)를 각각 독립적으로 생성하였다. 각 분류기법의 예측력을 평가하기 위하여 전체정확도, 민감도, 특이도, 그리고 기하평균을 계산하였으며, 변수선택의 성능을 평가하기 위하여 중요한 입력변수 중에서 유의한 변수로 올바르게 선택된 개수(number of correctly selected input variable; NC)와 잡음변수 중에서 유의한 변수로 잘못 선택된 개수(number of incorrectly selected input variable; NIC)를 각각 계산하였다. 이러한 과정을 100번 독립적으로 반복하였으며, 각 평가지표에 대한 100번의 평균을 모의실험의 결과를 정리한 각각의 표에 나타내었다.

3.1. 실험모형 1

실험모형 1에서는 먼저 잠재변수 $\boldsymbol{z} = (z_1, \ldots, z_8)^T$를 다변량 정규분포 $N_8(\mathbf{0}, \Sigma)$로부터 생성하였다. 여기서 공분산 행렬 Σ의 (i,j)번째 원소는 $\mathrm{Cov}(z_i, z_j) = 0.5^{|i-j|}$이다. 그리고 독립적으로 40개의 확률변수 w_{kj} $(k = 1, \ldots, 8;\ j = 1, \ldots, 5)$를 표준정규분포로부터 생성하여 총 40개의 입력변수 $x_{kj} = (1/\sqrt{2})(z_k + w_{kj})$를 생성하였다. 이항 범주형 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 계산되었으며, 이때 실제 분류함수로는

$$f(\boldsymbol{x}) = [1.2x_{11} - 0.8x_{12} + 1.6x_{13}] + [x_{21} - 0.9x_{22} - 1.1x_{23} - 1.3x_{24} + 3x_{25}] + [2.5x_{61} + 1.4x_{62}] \quad (3.1)$$

을 이용하였다. 선형 분류함수 (3.1)은 총 8개의 그룹 중에서 3개의 그룹 x_1, x_2, x_6을 중요한 그룹으로 포함하고 있으며, x_1 그룹에서는 3개의 입력변수를, x_2 그룹에서는 5개의 입력변수 모두를, 그리고 x_6 그룹에서는 2개 입력변수를 각각 포함하고 있다.

먼저 제안한 WAH-SVM의 성능을 집단별 가중치와 적응적 조율모수의 적용여부에 따른 다양한 SVM 방법론들과 비교하기 위해 소수집단의 비율이 10%인 불균형 자료에서 모의실험을 진행하였으며, 그 결과는 Table 3.1에 정리되어 있다. 가중치의 적용여부에 따른 분류정확도를 보면, 집단별로 오분류 비용을 차등 적용한 가중치 적용 방법이 가중치를 적용하지 않은 방법에 비해 다수집단의 분류 정확도인 특

Table 3.2. Simulation results for Example 1 (percentage of minority class: 5%, 10%, 20%)

Percentage of Minority class	Method	Test classification accuracy(%)				Group selection		Input variable selection	
		Overall accuracy	Sensitivity	Specificity	G-mean	NC	NIC	NC	NIC
20%	H	84.5 (0.7)	72.4 (1.7)	96.5 (0.7)	83.6 (0.8)	3.0	2.8	10.0	17.0
	AH	84.9 (0.7)	73.3 (1.7)	96.7 (0.5)	84.0 (0.8)	3.0	1.9	10,0	6.3
	WH	86.6 (0.6)	88.3 (1.4)	88.9 (1.0)	88.8 (0.6)	3.0	0.7	10.0	5.7
	WAH	88.7 (0.5)	88.1 (1.2)	89.3 (1.0)	88.7 (0.5)	3.0	0.4	10.0	1.6
10%	H	80.0 (1.2)	61.8 (2.7)	98.2 (0.5)	77.9 (1.5)	3.0	4.0	10.0	21.9
	AH	80.2 (1.2)	61.9 (2.6)	98.6 (0.4)	78.1 (1.6)	3.0	2.8	10.0	8.9
	WH	87.9 (0.6)	86.6 (1.9)	89.2 (1.4)	87.9 (0.6)	3.0	0.5	9.9	4.0
	WAH	88.0 (0.6)	86.7 (1.9)	89.4 (1.4)	88.0 (0.6)	3.0	0.3	9.9	1.2
5%	H	75.6 (1.9)	51.9 (4.0)	98.8 (0.5)	71.6 (2.7)	3.0	4.8	10.0	25.9
	AH	75.8 (2.3)	51.4 (4.8)	99.1 (0.4)	71.6 (3.3)	3.0	3.5	9.9	11.2
	WH	86.5 (1.2)	84.3 (3.2)	88.7 (2.0)	86.4 (1.2)	3.0	0.4	9.6	3.5
	WAH	86.9 (1.1)	84.6 (3.1)	89.1 (2.0)	86.8 (1.1)	3.0	0.2	9.6	1.1

The numbers in parentheses are standard deviations.

NC = number of correctly selected input variable; NIC = number of incorrectly selected input variable.

이도가 다소 하락하긴 하였으나, 전체정확도, 민감도, 기하평균이 향상되었으며, 특히 소수집단의 정확도인 민감도가 크게 향상된 것을 알 수 있다. 변수 선택 측면에도 가중치를 적용한 방법들이 그렇지 않은 방법들에 비해 잡음그룹과 변수의 제거 능력이 탁월한 것으로 나타났다. 다음으로 적응적 조율모수의 사용여부에 따른 분류 정확도를 보면, 동일한 조율모수를 사용하는 방법에 비해 적응적 조율모수를 사용하는 방법론이 4가지 지표 모두 높은 수준을 나타내는 것을 알 수 있다. 마지막으로, 제안 방법인 WAH-SVM의 경우 기존의 SVM 방법론들에 비해 소수집단의 분류 정확도인 민감도와 기하평균 측면에서 가장 우수한 성능을 보였으며, 잡음그룹과 잡음변수의 제거에도 그 성능이 탁월하였다.

Table 3.2에는 소수집단의 비중을 달리 적용한 경우의 모의실험 결과가 정리되어 있다. 오분류 비용에 대한 가중치를 적용하지 않은 H-SVM과 AH-SVM의 경우 소수집단에 대한 훈련개체의 비율이 낮아질수록 소수집단의 분류 정확도인 민감도가 급격히 낮아진 반면, 가중치를 적용한 WH-SVM과 WAH-SVM의 경우에는 민감도 하락이 크지 않음을 알 수 있다. 또한 제안 모형인 WAH-SVM의 분류 정확도가 가장 우수하게 나타났으며, 변수 선택에서도 잡음그룹과 변수를 제거하는 능력이 가장 탁월한 것을 확인할 수 있다.

3.2. 실험모형 2

실험모형 2에서는 9개의 잠재변수 $\boldsymbol{z} = (z_1, \ldots, z_9)^T$와 확률변수 w를 표준정규부포로부터 독립적으로 생성한 후, 이로부터 36개의 입력변수 $x_{kj} = \{(1/\sqrt{2})(z_k + w)\}^j$ $(k = 1, \ldots, 9;\ j = 1, \ldots, 4)$를 생성하였다. 이항 범주형 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 계산되었으며, 이때 실제 분류함수로는

$$f(\boldsymbol{x}) = \left[x_{31} + x_{32} + x_{33} - \frac{1}{5}x_{34}\right] + [3x_{61} + 1.5x_{62}] + \left[\frac{2}{3}x_{91} + \frac{1}{3}x_{93}\right] \tag{3.2}$$

을 이용하였다. 비선형 분류함수 (3.2)는 연속형 요인의 4차 다항식을 입력변수로 이용한 가법모형이다. 총 9개의 요인 중에서 세 개의 요인 x_3, x_6, x_9가 중요한 요인으로 활용되었으며, 최종적으로 8개의 입력변수가 분류함수에 포함되었다.

Table 3.3. Simulation results for Example 2 (percentage of minority class: 10%)

Weight (Y/N)	Adaptive (Y/N)	Method	Test classification accuracy(%)				Group selection		Input variable selection	
			Overall accuracy	Sensitivity	Specificity	G-mean	NC	NIC	NC	NIC
N	N	L_1	77.6 (1.2)	61.3 (2.7)	98.7 (0.4)	77.7 (1.6)	3.0	6.0	7.2	25.4
		F_∞	77.7 (1.3)	61.5 (2.7)	98.6 (0.5)	77.9 (1.6)	3.0	6.0	8.0	38.0
		H	77.6 (1.2)	61.1 (2.7)	98.8 (0.5)	77.6 (1.5)	3.0	4.9	7.2	20.3
	Y	AL_1	77.6 (1.3)	61.4 (2.7)	99.0 (0.4)	77.8 (1.6)	3.0	5.2	6.0	12.9
		AF_∞	77.8 (1.2)	61.5 (2.6)	98.7 (0.5)	77.8 (1.6)	3.0	5.9	8.0	27.5
		AH	77.7 (1.2)	61.5 (2.6)	98.9 (0.4)	77.7 (1.6)	3.0	2.9	7.2	8.4
Y	N	WL_1	86.3 (0.8)	84.0 (2.0)	89.3 (1.7)	86.6 (0.8)	3.0	5.3	4.9	10.5
		WF_∞	85.8 (0.7)	81.7 (1.9)	91.3 (1.5)	86.3 (0.7)	3.0	5.8	8.0	27.2
		WH	86.9 (0.6)	84.3 (1.9)	90.2 (1.4)	87.2 (0.6)	3.0	0.9	6.0	3.3
	Y	WAL_1	86.6 (0.7)	84.3 (1.9)	90.7 (1.4)	86.9 (0.7)	3.0	2.6	4.8	4.2
		WAF_∞	86.5 (0.6)	82.4 (1.9)	91.7 (1.3)	86.9 (0.7)	3.0	4.4	8.0	21.6
		WAH	87.0 (0.6)	84.9 (1.9)	91.1 (1.4)	87.4 (0.6)	3.0	0.5	6.3	1.7

The numbers in parentheses are standard deviations.

NC = number of correctly selected input variable; NIC = number of incorrectly selected input variable.

Table 3.4. Simulation results for Example 2 (percentage of minority class: 5%, 10%, 20%)

Percentage of Minority class	Method	Test classification accuracy(%)				Group selection		Input variable selection	
		Overall accuracy	Sensitivity	Specificity	G-mean	NC	NIC	NC	NIC
20%	H	82.2 (1.0)	70.3 (2.1)	97.5 (0.6)	82.8 (2.5)	3.0	4.5	7.5	17.7
	AH	82.4 (1.2)	70.4 (1.2)	97.7 (0.6)	82.9 (1.2)	3.0	2.4	6,5	6.8
	WH	87.3 (0.5)	84.7 (0.4)	90.6 (1.3)	87.6 (0.4)	3.0	1.5	7.4	5.3
	WAH	87.5 (0.5)	84.9 (0.4)	91.0 (1.3)	87.8 (0.4)	3.0	0.6	6.9	2.1
10%	H	77.6 (1.2)	61.1 (2.7)	98.8 (0.5)	77.6 (1.5)	3.0	4.9	7.2	20.3
	AH	77.7 (1.2)	61.5 (2.6)	98.9 (0.4)	77.7 (1.6)	3.0	2.9	7.2	8.4
	WH	86.9 (0.6)	84.3 (1.9)	90.2 (1.4)	87.2 (0.6)	3.0	0.9	6.0	3.3
	WAH	87.0 (0.6)	84.9 (1.9)	91.1 (1.4)	87.4 (0.6)	3.0	0.5	6.3	1.7
5%	H	73.4 (2.6)	53.3 (4.8)	99.3 (0.3)	72.7 (3.2)	3.0	4.6	7.1	18.7
	AH	73.9 (2.9)	53.3 (5.4)	99.6 (0.3)	72.9 (3.7)	3.0	2.1	6.8	6.8
	WH	86.2 (0.8)	83.5 (2.6)	89.7 (2.5)	86.5 (0.7)	2.9	0.7	4.9	2.4
	WAH	86.4 (0.9)	83.6 (2.7)	90.6 (2.5)	86.8 (0.8)	2.9	0.3	5.3	2.5

The numbers in parentheses are standard deviations.

NC = number of correctly selected input variable; NIC = number of incorrectly selected input variable.

먼저 소수집단의 비율이 10%인 불균형 자료에서 모의실험을 진행하였으며, 각각의 SVM 방법론에 대한 성능은 Table 3.3에 정리되어 있다. 모의실험 1의 결과와 마찬가지로 오분류 비용을 차등 적용하기 위한 가중치와 적응적 조율모수의 사용으로 제안 방법인 WAH-SVM의 분류 정확도 및 변수선택의 성능이 크게 향상되었음을 알 수 있다.

소수집단의 비중을 달리 적용한 경우의 모의실험 결과는 Table 3.4에 정리되어 있으며, 이 결과 또한 모의실험 1과 유사하다. 가중치 적용방법인 WH-SVM과 WAH-SVM의 경우에는 소수집단의 비율이 낮아짐에도 불구하고 민감도 하락이 크지 않았다. 또한, 제안 모형인 WAH-SVM은 특이도는 다소 하락했으나, 민감도와 기하평균의 성능이 향상되었음을 알 수 있다. 변수 선택에 있어서도 제안 방법인

Table 4.1. Description of input factors and input variables of credit approval data

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

WAH-SVM이 분류분석 기법들 중에서 잡음변수를 가장 많이 제거하였다.

4. 실제자료 분석

4.1. 국내 대출승인 자료

이번 절에서는 대출승인 자료(loan approval data)를 활용하여 제안하는 WAH-SVM과 기존 SVM 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 1,602명의 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 크레딧뷰로(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 입력변수(input features)로 활용하였다.

제안하는 WAH-SVM의 성능을 비교 평가하기 위해 L_1-norm SVM, F_∞-norm SVM, H-SVM과 각각의 기법들에 가중치와 적응적 조율모수를 적용한 방법론을 사용하여 대출승인 자료를 분석하였으며, 이때 모형의 적합 및 평가를 위해 전체 자료의 1/4을 훈련자료로, 1/4을 검증자료로, 그리고 나머지 1/2을 평가자료로 활용하였다. 분류기법들의 분류 정확도 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 4.2를 보면, 우선 집단 간의 오분류 비용을 차등 적용하는 가중치 방법인 WL_1-norm SVM, WF_∞-norm SVM, WH-SVM이 가중치를 이용하지 않는 방법론에 비해 비록 특이도가 다소 감소하

Table 4.2. Simulation results for credit approval data

Method	Test classification accuracy(%)				Number of selected factors	Number of selected variables
	Overall accuracy	Sensitivity	Specificity	G-mean		
L_1	90.0 (0.9)	74.1 (5.7)	94.0 (1.1)	83.4 (2.9)	13.27	28.26
F_∞	90.5 (0.9)	74.9 (5.4)	94.3 (1.1)	84.0 (3.0)	13.29	42.00
H	90.3 (0.9)	73.0 (6.1)	94.5 (1.8)	82.9 (3.0)	8.43	18.43
AL_1	90.0 (0.9)	72.3 (5.8)	94.4 (1.2)	82.5 (3.0)	9.06	13.61
AF_∞	90.3 (0.9)	76.8 (5.0)	93.7 (1.1)	84.8 (2.5)	11.20	35.14
AH	90.3 (0.9)	74.4 (5.6)	94.3 (1.1)	84.2 (2.9)	4.57	10.14
WL_1	90.0 (0.9)	90.2 (3.6)	90.0 (1.3)	90.1 (2.6)	12.37	24.73
WF_∞	89.7 (0.9)	94.5 (3.1)	88.2 (1.3)	90.8 (1.8)	12.28	39.00
WH	91.3 (1.0)	94.9 (2.6)	88.9 (1.5)	91.8 (1.7)	3.57	7.57
WAL_1	90.1 (1.0)	90.8 (4.7)	90.0 (1.7)	90.3 (1.9)	7.25	10.33
WAF_∞	89.9 (0.9)	95.1 (2.0)	87.4 (1.4)	91.1 (0.9)	9.10	27.61
WAH	91.3 (1.1)	93.2 (2.3)	89.3 (1.8)	91.9 (1.5)	2.24	5.27

The numbers in parentheses are standard deviations.

지만, 금융분야에서 중요하게 다루어지는 민감도 측면에서 아주 우수한 성능을 나타내고 있는 것을 확인할 알 수 있다. 또한 가중치가 적용된 상태에서 적응적 조율모수를 추가 적용한 WAL_1-norm SVM, WAF_∞-norm SVM, WAH-SVM은 적응적 조율모수를 적용하지 않은 방법들에 비해 변수선택에 있어서 더 간결한 모형을 제공하고 있다. 마지막으로 제안 방법인 WAH-SVM은 다른 비교 방법론들에 비해서 분류 정확도가 가장 우수하며, 입력변수의 선택에 있어서도 가장 간결한 모형을 제공하므로 실제 불균형 자료의 분류분석에서 그 활용 가능성이 높다고 할 수 있겠다.

4.2. 이동통신 이탈 자료

이동통신 이탈 자료는 이동통신 고객 5,000명의 유지(retention)와 이탈(churn)의 이항 반응변수와 20개의 입력요인으로 구성되어 있으며, UCI Machine Learning Repository (http://archive.ics.uci.edu/ml)에서 제공하는 자료를 사용하였다. 이 자료는 소수집단인 이탈고객이 707명(15%)이고 다수집단인 유지고객이 4,293명(85%)인 불균형 자료이다. 총 20개 입력요인 중, 입력변수로 사용하기 어려운 주소, 전화번호 등을 제외한 16개 요인을 활용하였다. 연속형 입력변수 14개는 표준화하여 3차 다항식을 입력변수로 활용하였으며, 범주형 입력요인은 가변수(dummy variables) 형태로 변환하여 입력변수로 활용하였다. 따라서, 모형적합을 위해 총 44개 입력변수(input features)를 16개의 그룹(groups)으로 구성하였다 (Table 4.3).

전체 자료 중 1/4을 모형 적합을 위한 훈련자료로, 1/4을 조율모수 λ를 선택하기 위한 검증자료로, 그리고 나머지 1/2을 적합된 모형을 평가하기 위한 평가자료로 활용하였다. 분류기법들의 예측력 평가를 위해 전체 정확도, 민감도, 특이도. 그리고 기하평균을 계산하였으며, 변수선택의 선택을 평가하기 위해 16개의 입력요인 중 유의한 요인으로 선택된 입력요인의 개수와 44개 입력요인 중 유의한 변수로 선택된 입력변수의 개수를 각각 계산하였다. 이러한 과정을 100번 독립적으로 반복하였으며, 각각의 평가지표에 대한 100번의 평균을 계산하였다.

Table 4.4에는 이동통신 이탈 자료에 대한 WAH-SVM의 분석결과가 기존 방법들의 결과와 함께 정리되어 있으며, 국내 대출 승인자료의 분석에서와 유사한 시사점을 보이고 있다. 우선, 소수집단의 예측력

Table 4.3. Description of input factors and input variables of churn data

No	Input factor	Information of input factor	Type	Number of input variables
1	NVM	음성메시지 통화회수	Continuous	3
2	TDM	통화시간(오전)		3
3	TDC	통화회수(오전)		3
4	TDG	통화요금(오전)		3
5	TEM	통화시간(오후)		3
6	TEC	통화회수(오후)		3
7	TEG	통화요금(오후)		3
8	TNM	통화시간(저녁)		3
9	TNC	통화회수(저녁)		3
10	TNG	통화요금(저녁)		2
11	TIM	통화시간(외국)		3
12	TIC	통화회수(외국)		3
13	TIG	통화요금(외국)		2
14	NCSC	고객콜센터로의 전화회수		3
15	IP	국제통화사용여부	Categorical	1
16	VMP	음성메세지 사용여부		1
				44

Table 4.4. Simulation results for churn data

Method	Test classification accuracy(%)				Number of selected factors	Number of selected variables
	Overall accuracy	Sensitivity	Specificity	G-mean		
L_1	88.0 (0.6)	32.8 (6.0)	97.1 (1.0)	56.2 (5.0)	14.4	28.9
F_∞	87.9 (0.6)	30.0 (6.6)	97.4 (1.1)	53.7 (5.8)	14.2	40.5
H	87.9 (0.6)	29.8 (7.2)	97.5 (1.1)	53.4 (6.4)	9.3	21.6
AL_1	87.9 (0.7)	30.6 (7.7)	97.4 (1.0)	54.7 (9.1)	12.5	20.4
AF_∞	87.9 (0.7)	31.4 (7.6)	97.2 (1.1)	54.8 (6.8)	12.0	33.6
AH	87.9 (0.7)	31.8 (8.0)	97.3 (1.1)	54.7 (9.3)	9.1	17.8
WL_1	83.0 (1.3)	75.4 (3.6)	84.2 (1.7)	79.7 (1.7)	13.6	24.9
WF_∞	82.6 (1.5)	75.2 (3.8)	83.8 (1.9)	79.4 (1.8)	13.8	39.4
WH	83.1 (1.2)	76.5 (4.1)	84.2 (1.7)	80.2 (1.8)	6.6	12.8
WAL_1	83.5 (1,2)	75.5 (4.0)	84.8 (1.7)	80.0 (1.7)	10.9	15.7
WAF_∞	83.1 (1.5)	75.6 (4.1)	84.2 (2,1)	79.7 (1.8)	10.8	30.3
WAH	83.5 (1.2)	77.1 (4.5)	84.5 (1.9)	80.5 (1.8)	6.3	10.5

The numbers in parentheses are standard deviations.

개선을 위해 가중치를 달리 적용하는 WL_1-norm SVM, WF_∞-norm SVM, WH-SVM이 가중치를 이용하지 않는 방법론에 비해 소수집단의 예측력인 민감도가 크게 향상되었으며, 그로 인해 전체 예측력과 기하평균이 높게 나타남을 알 수 있다. 적응적 조율모수를 추가 적용한 WAL_1-norm SVM, WAF_∞-norm SVM, WAH-SVM은 동일한 조율모수를 적용하는 방법론에 비해 분류 정확도와 모형의 간결성 측면에서 우수했으며, 특히 제안 방법인 WAH-SVM이 가장 좋은 성능을 보이고 있는 것을 확인할 수 있다.

5. 결론

두 집단 간의 개체수가 상이한 불균형 자료의 분류분석에서 입력변수들이 그룹화 되어 있는 경우 라소 벌점함수를 사용한 WL_1-norm SVM은 입력변수들을 개별적으로 선택함으로 인해 그룹변수들의 동시적인 선택에서는 그 유용성이 떨어진다. 본 논문에서는 입력변수들이 그룹화 되어 있는 고차원 자료의 분류분석에서 그룹과 그룹내 입력변수의 동시적인 선택이 가능한 H-SVM을 응용하여 불균형 자료의 분석에 활용 가능한 WAH-SVM 방법론을 제안하였다. 이는 H-SVM에 적응적 조율모수를 적용하여 추정의 효율성을 향상시키고, 오분류 비용을 집단별로 차등적으로 적용하여 소수집단의 예측력을 개선한 모형이다. 본 논문에서는 모의실험과 실제자료의 분석을 통해 제안한 WAH-SVM이 입력변수가 그룹화 되어 있는 불균형 자료의 분류분석에서 기존의 방법들에 비해 분류 정확도와 변수선택 측면에서 그 성능이 우수함을 확인하였다.

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. (2012). On the use of adaptive weights for the F_∞-norm support vector machine, *The Korean Journal of Applied Statistics*, **25**, 829–835.

Bang, S., Kang, J., Jhun, M., and Kim, E. (2016). Hierarchically penalized support vector machine with grouped variables, *International Journal of Machine Learning and Cybernetics*, DOI:10.1007/s13042-016-0494-2.

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.

Breiman, L. (1995). Better subset regression using the nonnegative garrote, *Technometrics*, **37**, 373–384.

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

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

Domingos, P. (1999). Metacost: a general method for making classifiers cost-sensitive. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 155–164.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its Oracle properties, *Journal of American Statistical Association*, **96**, 1348–1360.

Friberg, H. A. (2013). Users Guide to the R-to-MOSEK Interface. URL http://rmosek.r-forge.r-project.org.

Hwang W., Zhang H., and Ghosal, S. (2009). FIRST: Combining forward iterative selection and shrinkage in high dimensional sparse linear regression, *Statistics and Its Interface*, **2**, 341–348.

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, E., Jhun, M., and Bang, S. (2015). Weighted L_1-norm support vector machine for classification of highly imbalanced data, *The Korea Journal of Applied Statistics*, **28**, 9–22.

Kotsiantis, S., Kanellopoulos, D., and Pintelas, P. (2006). Handling imbalanced datasets: a review, *GESTS International Transactions on Computer Science and Engineering*, **30**, 25–36.

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

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

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.

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, S., Nan, B., Zhou, N., and Zhu, J. (2009). Hierarchically penalized Cox regression with grouped variables, *Biometrika*, **96**, 307–322.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society, Series B*, **68**, 49–67.

Zhou, N. and Zhu, J. (2010). Group variable selection via a hierarchical lasso and its oracle property, *Statistics and Its Interface*, **3**, 557-574.

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

Zou, H. (2006). The adaptive lasso and its oracle properties, *Journal of the Royal Statistical Society, Series B*, **101**, 1418–1429.

Zou, H. (2007). An improved 1-norm SVM for simultaneous classification and variable selection. In *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics*.

Zou, H. and Yuan, M. (2008). The F_∞-norm support vector machine, *Statistica Sinica*, **18**, 379–398.

그룹변수를 포함하는 불균형 자료의 분류분석을 위한 서포트 벡터 머신

김은경[a] · 전명식[b] · 방성완 [c,1]

[a]코리아크레딧뷰로 연구소, [b]고려대학교 통계학과, [c]육군사관학교 수학과

(2016년 6월 9일 접수, 2016년 7월 6일 수정, 2016년 7월 7일 채택)

요 약

H-SVM은 입력변수들이 그룹화 되어 있는 경우 분류함수의 추정에서 그룹 및 그룹 내의 변수선택을 동시에 할 수 있는 방법론이다. 그러나 H-SVM은 입력변수들의 중요도에 상관없이 모든 변수들을 동일하게 축소 추정하기 때문에 추정의 효율성이 감소될 수 있다. 또한, 집단별 개체수가 상이한 불균형 자료의 분류분석에서는 분류함수가 편향되어 추정되므로 소수집단의 예측력이 하락할 수 있다. 이러한 문제점들을 보완하기 위해 본 논문에서는 적응적 조율모수를 사용하여 변수선택의 성능을 개선하고 집단별 오분류 비용을 차등적으로 부여하는 WAH-SVM을 제안하였다. 또한, 모의실험과 실제자료 분석을 통하여 제안한 모형과 기존 방법론들의 성능 비교하였으며, 제안한 모형의 유용성과 활용 가능성 확인하였다.

주요용어: 적응적 조율모수, 계층적 벌점화, 불균형 자료, 서포트 벡터 머신, 변수선택

이 논문은 2015년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업이며 (NRF-2015R1C1A1A02036473)(방성완), 2013년도 정부(교육부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임 (NRF-2013R1A1A2A10007545)(전명식).

[1]교신저자: (01805) 서울시 노원구 화랑로 574, 육군사관학교 수학과. E-mail: wan1365@gmail.com