응용통계연구 (2012)
**25**(2), 261–267
DOI: http://dx.doi.org/10.5351/KJAS.2012.25.2.261

# 신용평점화에서 벌점화를 이용한 절단값 선택

진슬기[1] · 김광래[2] · 박창이[3]

[1]서울시립대학교 통계학과, [2]고려대학교 통계연구소, [3]서울시립대학교 통계학과

(2011년 10월 14일 접수, 2011년 12월 20일 수정, 2011년 12월 26일 채택)

**요 약**

신용평점표(credit scorecard) 작성시 각 특성변수(characteristic variable)들을 몇 개의 속성(attribute)들로 나누고 각 속성에 적절한 가중치를 부여하게 된다. 이 과정을 성김화(coarse classification)라 한다. 특성변수들을 속성들로 나눌 때 그 기준이 되는 절단값(cutpoint)을 선택해야 한다. 본 논문에서는 벌점화(penalization) 기반의 절단값 선택법을 제안한다. 또한 여러가지 모의실험과 실제 신용자료의 분석을 통하여 제안된 방법과 기존의 절단값 선택법인 스플라인 분류 기계 (Koo 등, 2009)의 성능을 비교한다.

주요용어: 스플라인 분류 기계, 성김화, 신용평점표.

## 1. 서론

신용평점화(credit scoring)에서 신용평점표(credit scorecard)는 매우 중요한 역할을 한다. 금융기관은 불량과 우량의 그룹에 속하는 고객들의 특성을 이용해 작성한 신용평점표를 기반으로 대출 신청자의 대출신청서와 다른 금융기관으로부터의 정보 등을 이용하여 대출신청자의 신용위험도의 한 측도인 신용평점(credit score)을 산출한다. 이러한 평점표의 작성에 사용되는 분류 기법으로는 판별분석, 의사결정나무, 로지스틱 회귀 등이 있다. 신용평점화에서 사용되는 여러 가지 통계 모형에 대해서는 Hand와 Henley (1997)를 참조할 수 있다.

신용평점화에서 한 가지 중요한 문제는 나이와 수입 등과 같은 연속형 특성변수(characteristic variable)들을 몇 개의 속성(attribute)들로 분할하거나 수준의 개수가 많은 이산형 특성변수를 몇 개의 수준으로 재그룹화 하는 것이다. 이러한 과정을 성김화(coarse classification)라고 한다. 이를 통해 금융기관에서는 고객들에게 대출에 대한 결정 사항을 쉽게 이해시킬 수 있고 반응변수와 설명변수간에 비선형적인 관계를 허용함으로써 설명의 복잡도를 높이지 않고 예측력을 향상시킬 수도 있다. 본 논문에서는 연속형 특성변수를 속성들로 나눌 때 그 기준이 되는 절단값(cutpoint)을 선택하는 방법에 대하여 연구하고자 한다.

문헌상의 주요한 절단값 선택방법들로는 Hand와 Adams (2000)의 모의 담금질(simulated annealing), Koo 등 (2009)의 스플라인 분류 기계(classification spline machine) 등이 있다. 모의 담금질에서는 절단값의 개수를 입력값으로 주어야 하는데 현실적으로 최적의 절단값의 개수를 미리 아는 경우는 드물다. 반면 스플라인 분류 기계에서는 절단값 선택문제를 상수 스플라인 기저(basis)의 선택문제로 변환하여 단계적 선택법(stepwise selection method)에 의해 자동적으로 적절한 기저를 선택하게 된다. 그러나

이 논문은 2011년도 서울시립대학교 교내학술연구비에 의하여 연구되었음.
[3]교신저자: (130-743) 서울시 동대문구 전농동 90, 서울시립대학교 통계학과, 조교수. E-mail: park463@uos.ac.kr

단계적 선택법과 같은 최적부분집합 선택법(best subset selection)은 변수선택이 불연속적으로 이루어짐으로써 그 선택 결과가 매우 불안정하며 경우에 따라서는 계산량이 많을 수 있다 (Breiman, 1996).

본 연구에서는 상수 스플라인 기저 선택에 최근 많이 연구되는 벌점함수(penalty function)인 Tibshirani (1996)의 LASSO(least absolute shrinkage and selection operator)와 Fan과 Li (2001)의 SCAD(smoothly clipped absolute deviation)를 적용하여 절단값을 선택하고자 한다. 최적부분집합 선택법의 경우에는 자료가 약간만 변해도 선택된 변수들이 매우 달라질 수 있는 반면에 벌점화 기반의 방법은 추정계수가 자료의 연속함수이므로 변수 선택 측면에서 보다 안정적이다. 따라서 벌점화 기반의 선택법은 기저의 선택과 축소 추정을 동시에 함으로써 최적부분집합 선택법의 일종인 스플라인 분류기보다 더 안정적인 기저 선택과 예측력의 향상을 기대할 수 있다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 벌점화 기반의 절단값 선택법에 대하여 설명한다. 3장에서는 Koo 등 (2009)의 모의실험과 하재환과 박창이 (2009)의 신용자료에 대하여 스플라인 분류기계와 벌점함수에 의한 절단값 선택법의 예측성능을 비교하며, 마지막으로 4장에서는 결론을 기술한다.

## 2. 벌점화를 이용한 절단값 선택법

벌점화에 기반한 절단값 선택법에 대하여 설명하기 전에 우선 필요한 기호들을 정의하기로 한다. 설명변수를 $\boldsymbol{X} = (X^{(1)}, \ldots, X^{(p)}) \in \mathbb{R}^p$, 반응변수를 $Y \in \{0,1\}$라고 하고 $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$를 이에 대응하는 훈련자료라고 정의하자. 여기서 $\boldsymbol{x}_i = (x_i^{(1)}, \ldots, x_i^{(p)})$는 $i$번째 대출 신청자의 나이, 재직 기간, 수입과 같은 특성값들로 이루어진 벡터이고, $y_i$는 신용상태가 우량인 경우 1이고 불량인 경우 0의 값을 갖는다. 각 $j(=1,\ldots,p)$에 대하여 특성변수 $X^{(j)}$의 절단값의 개수는 $n_j$이고 그 변수가 취하는 값의 하한과 상한을 각각 $t_0^{(j)}$와 $t_{n_j+1}^{(j)}$로 나타내자. $n_j$개의 절단값 $t_1^{(j)}, \ldots, t_{n_j}^{(j)}$에 대하여 $t_0^{(j)} < t_1^{(j)} < \cdots < t_{n_j}^{(j)} < t_{n_j+1}^{(j)}$를 가정하면, $n_j$개의 가변수(dummy variable) $z_k^{(j)} = I(X^{(j)} \in (t_{k-1}^{(j)}, t_k^{(j)}])$, $k = 1, \ldots, n_j$들을 정의할 수 있다. 여기서 $I$는 지시(indicator) 함수를 나타낸다. 설명변수 $\boldsymbol{X}$가 주어졌을 때 반응변수 $Y$의 조건부 확률을 로지스틱 회귀 모형

$$\mathbb{P}(Y = 1 | \boldsymbol{X} = \boldsymbol{x}) = \frac{\exp(f(\boldsymbol{x}))}{1 + \exp(f(\boldsymbol{x}))}$$

을 이용하여 모형화할 수 있다. 여기서

$$f(\boldsymbol{x}) = \sum_{j=1}^{p} \sum_{k=1}^{n_j+1} \beta_{j,k-1} z_k^{(j)} \tag{2.1}$$

는 $\sum_{j=1}^p n_j$개의 절단값을 후보 절단값으로 갖는 후보모형을 나타낸다. 즉, 절단값 선택의 목적은 예측력이 높도록 적절한 절단값 $t_k^{(j)}$들을 선택하는 것이다.

구자용 등 (2005)과 Koo 등 (2009)에서 처럼 특성변수 $x^{(j)}$에 대하여 절단값 $t_k^{(j)}$를 매듭점(knot)으로 하는 상수 스플라인 기저 $B_{jk}(x^{(j)}) = (x^{(j)} - t_k^{(j)})_+^0 = I(x^{(j)} \geq t_k^{(j)})$를 고려하면

$$f(\boldsymbol{x}) = \gamma_0 + \sum_{j=1}^{p} \sum_{k=1}^{n_j} \gamma_{jk} B_{jk}\left(x^{(j)}\right) \tag{2.2}$$

로 표현된다. 여기서 $\gamma_0 = \sum_{j=1}^p \beta_{j,0}$, $\gamma_{jk} = \beta_{j,k} - \beta_{j,k-1}$, $k = 1, \ldots, n_j$이다. 따라서 절단값 선택문제는 상수 스플라인 기저의 선택문제로 바뀌게 된다. 스플라인 분류 기계에서는 Kooperberg 등 (1997)의

단계별(stepwise) 기저 선택법에 의하여 AIC(Akaike information criterion)를 최소화하는 모형을 선택한다. 보다 자세한 사항은 Koo 등 (2009)를 참조하기 바란다.

Brieman (1996)에서 지적된 바와 같이 단계적 선택법과 같은 최적부분집합 선택법은 변수선택이 불연속적으로 이루어짐으로써 그 선택 결과가 매우 불안정하며 경우에 따라서는 계산량이 많을 수 있다는 단점이 있다. 이에 대한 대안으로 벌점화를 통한 기저 선택법을 고려할 수 있는데 이는 계수 선택과 축소 추정을 동시에 함으로써 보다 안정적인 선택과 함께 예측력 또한 향상시킬 수 있는 장점이 있다. 로그-우도함수는

$$l(\boldsymbol{\gamma}) = \sum_{i=1}^{n} [y_i \log(f(\boldsymbol{x}_i)) + (1 - y_i) \log(1 - f(\boldsymbol{x}_i))]$$

와 같고, $\lambda > 0$는 조율모수(tuning parameter), $J_\lambda$는 벌점함수, $\boldsymbol{\gamma} = (\gamma_0, \gamma_{11}, \ldots, \gamma_{pn_p})^T$라 하자. 그러면 상수 스플라인 기저의 선택을 위하여 다음과 같은 벌점화 목적함수의 최소화를 고려할 수 있다.

$$\hat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma}}{\operatorname{argmin}} \left[ -l(\boldsymbol{\gamma}) + \sum_{j=1}^{p} \sum_{k=1}^{n_j} J_\lambda(\gamma_{jk}) \right].$$

본 논문에서는 대표적인 벌점함수인 Tibshirani (1996)의 LASSO

$$J_\lambda(\theta) = \lambda|\theta|$$

와 일차 도함수로 정의되는 Fan과 Li (2001)의 SCAD

$$J'_\lambda(\theta) = I(\theta \le \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda)$$

를 고려하기로 한다. 여기서 $a > 2$인 조율모수이나 Fan과 Li (2001)에서와 마찬가지로 $a = 3.7$로 고정하며 조율모수 $\lambda$는 로그-우도함수값에 대한 교차확인법(crossvalidation)으로 선택하기로 한다. 기본적으로 훈련자료의 모든 변수값을 후보 절단값으로 설정할 수 있는데, 고차원 자료에 대한 계산량의 부담을 줄이기 위하여 각 $x^{(j)}$에 대한 기저의 개수에 대한 적절한 상한 $M$을 두기로 한다. 후보 절단값으로는 상한 $M$의 범위안에서 훈련자료에서 $x^{(j)}$의 표본 분위수들을 이용할 수 있다.

## 3. 자료 분석

모의실험 및 실제 신용자료의 분석에 사용된 코드는 R을 이용하였으며, LASSO는 `glmpath` 패키지의 `glmpath` 함수, SCAD는 `ncvreg` 패키지의 `ncvreg` 함수를 이용하여 구현하였다.

### 3.1. 모의실험

벌점화 방법과 스플라인 분류 기계의 성능 비교를 위하여 다음의 세 가지 모의실험을 실시하였다.

- 모의실험 A

  $Y = 0$과 1을 각각 $N/2$개씩 생성한다. $Y = 0$인 경우 서로 독립적으로 $(X_1, X_2)^T$는 평균벡터가 $(-1, 1)^T$ 공분산행렬이 $\boldsymbol{I}$인 이변량 정규분포를 따르며, $Y = 1$인 경우에는 $(X_1, X_2)^T$는 평균벡터가 $(1, -1)^T$이고 공분산행렬이 $\boldsymbol{I}$인 이변량 정규분포를 따르도록 하였다. 나머지 17개 변수 $X_3, \ldots, X_{20}$은 $Y$값에 관계없이 서로 독립적으로 $N(0, 1)$에서 발생된 잡음변수이다. 이 경우 최적의 베이즈 분류기(Bayes classifier)는 두 설명변수 $X_1$과 $X_2$의 선형함수이다.

표 **3.1.** 모의실험의 결과: 평균 오분류율, ( )안은 표준오차

| 모의실험 | 스플라인 분류 기계 | LASSO 벌점화 | SCAD 벌점화 |
|---|---|---|---|
| A | 0.1718 (0.0064) | 0.1119 (0.0018) | 0.1195 (0.0032) |
| B | 0.2288 (0.0064) | 0.1739 (0.0041) | 0.1896 (0.0049) |
| C | 0.0997 (0.0067) | 0.0646 (0.0033) | 0.0682 (0.0057) |

- 모의실험 B

  모의실험 A와 마찬가지로 $Y$는 0과 1로 각각 $N/2$개씩 생성한다. $Y=0$이면 $X_1,\ldots,X_{10}$을 독립적으로 $N(1.5,2.5)$에서, $Y=1$일 때는 독립적으로 $N(1.5,3)$에서 생성한다. 이 경우에는 베이즈 분류기는 모든 설명변수 $X_1,\ldots,X_{10}$들로 이루어진 이차함수이다.

- 모의실험 C

  $X_1$과 $X_2$는 독립적으로 $U(-3,3)$에서, 나머지 $X_3,\ldots,X_{20}$은 독립적으로 $N(0,1)$에서 생성한다. $X_2>\tanh(\pi X_1)$이면 $Y=1$이고 그렇지 않으면 $Y=0$으로 $Y$를 생성한다. 이 경우 $X_1$과 $X_2$가 설명력이 있는 변수이고 나머지는 잡음변수이며, 베이즈 분류기는 $\text{sign}(X_2-\tanh(\pi X_1))$로 비선형 함수이다.

이 세 가지 모의실험은 Koo 등 (2009)의 모의실험을 약간 변형한 것이다. Koo 등 (2009)의 모의실험 A와 C에서는 잡음변수가 없는데, 이 경우 스플라인 분류 기계와 벌점화 방법의 예측력은 유의한 차이가 나지 않는 것으로 나타난다. 본 논문에서는 모의실험 A와 C에서 기저선택의 효율성의 차이를 명확히 보이기 위하여 잡음변수들을 더 추가하였다.

각 모의실험에서 자료의 크기가 각각 100과 1000인 훈련자료와 시험자료를 생성한 후, 훈련자료를 이용하여 LASSO와 SCAD 벌점함수를 사용한 절단값 선택법과 스플라인 분류 기계를 적합하고, 시험자료에 대하여 오분류율을 구하였다. 벌점화 방법에서 특성변수의 값들 중 서로 다른 값들의 개수가 9개 이상이면 10% 간격의 분위수들을 후보 절단값으로 하고, 9개 미만인 경우에는 최소값과 최대값을 제외한 나머지 값을 후보 절단값으로 하였다. 또한 벌점화 방법의 훈련 시 0과 1사이의 값을 갖도록 표준화된 조율모수를 100등분한 격자에 대하여 5-묶음(fold) 교차확인법으로 결정하였다. 실험의 변동성을 파악하기 위해 자료의 생성, 모형 적합, 시험자료에 대한 오분류율의 계산 등 전 과정을 50회 반복하였다.

표 3.1은 세 가지 모의실험에 대하여 50회의 반복에 대한 스플라인 분류기와 LASSO 및 SCAD 벌점화의 평균 오분류율을 보여준다. 괄호안의 숫자는 표준오차를 의미한다. 세 모의실험 모두에서 벌점화 방법이 스플라인 분류 기계 보다 작은 오분류율을 갖는 것을 확인할 수 있다. LASSO와 SCAD를 비교하면 LASSO가 SCAD에 비해 오분류율이 상대적으로 작은 것을 알 수 있다. SCAD는 일반적으로 모형선택 관점에서 일치성(consistency)이 성립하는 반면, LASSO는 설명변수들 간의 상관행렬이 특정 조건을 만족하는 경우에는 일치성이 성립하지 않는다는 사실이 알려져 있다 (Zou, 2006). 그러나 실제 자료분석시 흔히 SCAD가 LASSO에 비해 불필요한 변수들을 덜 선택하는 반면에 예측오차는 조금 더 크게 나타난다. 본 논문의 모의실험의 결과도 이러한 맥락에서 바라볼 수 있을 것이다.

### 3.2. 실제자료

하재환과 박창이 (2009)에서 분석된 국내 신용카드사와 은행의 신용대출 관련 자료에 대하여 스플라인 분류 기계와 벌점화 방법의 예측력을 비교하였다. 첫 번째 자료는 국내 어느 신용카드 회사의 카드론 자료로 자료의 크기는 941개이며 6개의 연속형 설명변수와 카드론 사용여부를 나타내는 반응변수를 분석에 사용하였다. 두 번째 자료는 국내 어느 은행의 대출 자료로 전체 자료수는 1921개, 연속형 설명변수

표 **3.2.** 실제 신용자료에서의 결과: 평균 오분류율, ( )안은 표준오차

| 자료 | 스플라인 분류 기계 | LASSO 벌점화 | SCAD 벌점화 |
|---|---|---|---|
| 신용카드 대출 | 0.2027 (0.0023) | 0.1780 (0.0017) | 0.1833 (0.0019) |
| 은행 대출 | 0.2725 (0.0018) | 0.2627 (0.0015) | 0.2687 (0.0018) |

는 21개이며 반응변수는 신용상태를 나타낸다.

우선 각 자료에 대하여 훈련 및 시험자료를 5:5의 비율로 임의로 분할한 후, 훈련자료를 이용하여 모형을 적합시키고 시험자료에 대하여 오분류율을 구하였다. LASSO와 SCAD의 조율모수는 모의실험과 동일한 방식으로 결정하였다. 실험의 변동성을 파악하기 위하여 임의 분할, 모형 적합 및 예측의 전 과정을 50번 반복하였다.

표 3.2는 두 신용자료에 대하여 50회의 임의 분할에 대한 스플라인 분류 기계와 LASSO 및 SCAD 벌점화 방법의 시험자료에 대한 평균 오분류율을 보여준다. 두 자료 모두에서 벌점화 방법의 예측성능이 스플라인 분류 기계보다 좋게 나타났다. 또한 모의실험에서와 유사하게 LASSO의 예측오차가 SCAD에 비해 더 작은 것을 볼 수 있다.

## 4. 결론

본 연구에서는 신용평점화에서 LASSO와 SCAD 벌점화에 기반한 절단값 선택법에 대하여 살펴보았다. 기존의 스플라인 분류 기계와 같은 최적부분집합 선택법은 기저의 선택이 불연속적이라서 불안정한 선택 결과를 주기 때문에 예측력이 저하될 수도 있다는 단점이 있다. 반면에 벌점화 기반의 선택법은 연속적으로 기저를 선택하는 동시에 축소추정을 하기 때문에 선택결과의 안정성과 예측력을 향상시킬 수 있다. 여러가지 모의실험과 실제 신용자료를 이용한 스플라인 분류 기계와 벌점화 방법의 비교를 통해 이러한 사실을 확인할 수 있었다.

스플라인 분류 기계의 경우에는 이산형 설명변수에 대한 재그룹화 기능이 있는 반면에 본 연구에서 제안한 벌점화 방법은 설명변수가 연속형인 경우만을 다루고 있다. 따라서 추후 연구과제로 연속형 설명변수뿐만 아니라 이산형 설명변수에 대한 성김화를 동시에 수행하도록 하는 벌점화 방법의 개발을 생각해 볼 수 있다. Yuan과 Lin (2006)의 그룹 LASSO 벌점화는 이산형과 연속형 변수를 동시에 다룰 수 있지만 이산형 변수의 수준에 대한 재그룹화 기능은 없기 때문에 신용평점화의 성김화 문제에 적합한 새로운 벌점화 방법의 개발이 필요할 것으로 예상된다.

## 참고문헌


구자용, 최대우, 최민성 (2005). 스플라인을 이용한 신용평점화, <응용통계연구>, **1**, 543–553.

하재환, 박창0l (2009). 선형판별분석에서의 변수 선택, *Journal of the Korean Data Analysis Society*, **11**, 381–389.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection, *Annals of Statistics*, **24**, 2350–2383.

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

Hand, D. J. and Adams, N. M. (2000). Defining attributes for scorecard construction in credit scoring, *Journal of Applied Statistics*, **27**, 527–540.

Hand, D. J. and Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review, *Journal of the Royal Statistical Society Series A*, **160**, 523–541.

Koo, J.-Y., Park, C. and Jhun, M. (2009). A classification spline machine for building a credit scorecard, *Journal of Statistical Computation and Simulation*, **79**, 681–689.

Kooperberg, C., Bose, S. and Stone, C. J. (1997). Polychotomous regression, *Journal of the American Statistical Association*, **92**, 117–127.

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

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

Zou, H. (2006). The adaptive lasso and its oracle properties, *Journal of the American Statistical Association*, **101**, 1418–1429.

# Cutpoint Selection via Penalization in Credit Scoring

Seul Ki Jin[1] · Kwang-Rae Kim[2] · Changyi Park[3]

[1]Department of Statistics, University of Seoul; [2]Institute of Statistics, Korea University
[3]Department of Statistics, University of Seoul

(Received October 14, 2011; Revised December 20, 2011; Accepted December 26, 2011)

**Abstract**

In constructing a credit scorecard, each characteristic variable is divided into a few attributes; subsequently, weights are assigned to those attributes in a process called coarse classification. While partitioning a characteristic variable into attributes, one should determine appropriate cutpoints for the partition. In this paper, we propose a cutpoint selection method via penalization. In addition, we compare the performances of the proposed method with classification spline machine (Koo *et al.*, 2009) on both simulated and real credit data.

Keywords: Classification spline machine, coarse classification, credit scorecard.

This work was supported by the University of Seoul 2011 Research Fund.
[3]Corresponding author: Assistant Professor, Department of Statistics, University of Seoul, 90 Jeonnong-Dong, Dongdaemun-Gu, Seoul 130-743, Korea. E-mail: park463@uos.ac.kr