碩士學位論文

주성분 로지스틱 회귀분류법의 분류율 향상에 관한 연구

韓國外國語大學校 大學院
統 計 學 科
金 基 浩

碩士學位論文

주성분 로지스틱 회귀분류법의 분류율 향상에 관한 연구

指導 李 碩 浩 教授

이 論文을 碩士學位請求論文으로 提出합니다.

2013年 3月

韓國外國語大學校 大學院
統 計 學 科
金 基 浩

이 論文을 金基浩의 碩士學位論文으로 認定함

2013年　月　日

審 査 委 員　_____ (인)

審 査 委 員　_____ (인)

審 査 委 員　_____ (인)

韓國外國語大學校 大學院

요약

본 연구에서는 로지스틱 회귀분류법을 응용한 이항분류법을 제안한다. 제안한 방법은 로지스틱 회귀분류법 사용시 본래의 변수를 그대로 이용하는 대신 이를 대표하는 주성분을 사용하되, 변수선택법에 의거하여 사용되는 주성분을 자동으로 선정하도록 하는 방법을 취하였다. 제안한 방법의 우수성을 컴퓨터 모의실험 및 실제자료들을 이용하여 타 방법론들과 정량적으로 평가 및 비교하였다.

목 차

표 목 차

그 림 목 차

1 서 론

통계적 분류법(statistica classification)은 자료의 주어진 특성(attributes)을 기반으로 자료를 분류하는 규칙을 생성하여 새로운 자료를 생성한 규칙에 의거하여 분류하는 통계적 방법이다. 자료를 두 개의 범주로 분류하는 경우를 이분분류(binary classification)라 부르며, 이는 다양한 응용학문에서 널리 활용되고 있다. 제안된 다양한 분류법들 가운데 좋은 분류법을 판단하는 기준은 새로운 자료에 대한 오분류율(test misclassification rate)으로, 작은 테스트 오분류율을 주는 방법이 선호된다. 본 연구에서 다루게 될 로지스틱 회귀분류법(logistic regression classification)은 통계적 이분분류법들 중 가장 널리 쓰이는 방법들 중 하나이다. 로지스틱 회귀분류법 이외에도 선형판별분석(linear discriminant analysis, LDA)이나 데이터마이닝 분야에서 널리 쓰이는 support vector machine (SVM) 및 k-nearest neighbor (kNN) 등의 방법이 있다. 통계적 분류법은 supervised learning의 일종으로, 이에 대한 자세한 설명은 표준적인 기계학습 관련 서적에 잘 기술되어 있다 (Hastie et al., 2009; Murphy, 2012).

로지스틱 분류법은 반응변수로써 이분형 변수(binary variable)를 가지는 로지스틱 회귀분석(logistic regression)을 분류문제에 적용한 것이다. 따라서 분류함수(classifier)를 생성하는데 쓰이는 자료(training data)가 자료의 개수보다 더 많은 설명변수를 가지게 되면 분류함수는 유일하게 결정되지 않는다. 또한 자료의 개수가 변수의 수보다 현저하게 크지 않다면, 이로부터 추정되는 분류함수를 새로운 자료(test data)에 적용할 경우 테스트 오분류율이 커짐은 이론적으로 잘 알려져 있다. 따라서 고차원의 자료를 고려하는 경우, 혹은 테스트 오분류율을 줄이기 위해서는 분류함

수를 찾기 위한 목적함수(objective function)에 벌점함수(penalty function)를 더함으로써 분류함수를 유일하게 찾을 수 있으며, 새로운 자료에 대한 최적의 분류함수를 얻을 수 있다. 이런 벌점함수의 도입은 분류함수 추정량의 자유도(degree of freedoms)를 축소하는 효과를 주며,이는 설명변수의 차원축소(demension reduction)와 연관된다.

회귀분석에서 예측력(prediction) 향상과 관련된 차원축소방법은 크게 두가지로 분류할 수 있다. 첫째, 설명변수 중 반응변수(분류문제의 경우, 범주표지변수)와 연관성이 높은 변수만 취하고 연관성이 낮은 설명변수는 모형에서 제외하는 방법이다. 이는 가능한 모든 조합의 모형 중에서 가장 높은 예측력을 주는 모형을 취함으로써 최적의 설명변수 집합을 찾는 방법이다. 하지만, 설명변수가 많은 경우(차원수가 높은 경우에 해당)에는 너무 많은 조합수가 존재하므로 실용적이지 못하다. 따라서 이를 손쉽게 하기 위해, 전진선택법(forward selection) 또는 후진소거법(backward elimination) 등의 방법이 이용된다. 하지만 이러한 모형선택은 이산적인 방법(discrete process)로써 안정적이지 못하다는 단점이 알려져 있다. 더욱 향상된 변수선택법으로써 희소벌점함수(sparsity-inducing penalty function)를 이용한 모형선택이 대안으로 이용되고 있다. 대표적인 희소벌점함수로는 Lasso (Tibshirani, 1996) 및 SCAD (Fan and Li, 2001)가 많이 이용된다. 두번째 방법으로는 차원축소법을 이용하여 설명변수 공간의 차원을 줄이는 방법이다. 대표적으로 주성분분석(principal component analysis, PCA)을 이용하여 설명변수 공간을 주요한 몇개의 주성분으로 대표하도록 하고, 이를 회귀분석에 설명변수로 대체하는 방법으로 주성분회귀분석(principal component regression, PC regression)이 있다. 주성분회귀분석은 일반적

으로 예측력을 높이는 것으로 알려져 있으나, 주요 주성분이 반응변수와 연관성이 적은 경우 예측력 향상을 보장할 수 없다는 단점이 있다.

회귀분석에서 사용되는 차원축소법은 로지스틱 회귀분류법에도 적용될 수 있다. 본 연구에서는 오분류율을 줄이는 방법으로써 주성분을 이용한 로지스틱 회귀분류법을 이용하되, 사용되는 주성분은 주요주성분만을 이용하는 대신 희소벌점함수를 이용하여 주성분을 선택하는 방법을 제안한다. 이는 주성분의 일부분을 사용하여 분류함수를 만들게 되므로 차원축소의 효과를 기대할 수 있으며 동시에 희소벌점함수를 이용하여 주성분을 선택함으로 주요주성분이 아닌 주성분이라 할 지라도 반응변수와 연관성이 높은 경우 분류함수 설정에 이용됨으로써 예측력의 향상을 기대할 수 있다.

2 이론적 배경

2.1 로지스틱 회귀모형

로지스틱 회귀모형은 반응변수가 이분형으로 주어진 회귀모형으로, 반응변수 값이 가지는 확률이 정해진 설명변수들에 의한 함수식으로 모형화한다. 이분형 자료는 변수값으로 2개의 범주를 가지는 경우를 의미한다. 이분형 자료를 가지는 예는 다양한 응용분야에서 나타난다. 예를 들어, 텍스트 마이닝에서 스팸메일의 여부, 의학분야에서 암의 존재여부 및 특정 소인에 대한 양/음성 반응, 전자상거래에서 특정 상품의 구매여부 등 다양

한 예들에서 이분형 자료를 관찰할 수 있다.

이분형 질적자료(binary qualitative variable)를 정량화 하기 위해 이를 확률변수 Y로 표기하고, 이분형 자료가 가지는 2개의 범주 중 관심 대상인 범주를 1, 다른 범주를 0으로 표기하자. 그리고 이분형 자료가 관심의 범주에 속할 확률을 $\Pr(Y=1)=\theta$로 나타내자. 이는 $\Pr(Y=0)=1-\theta$를 의미한다. 로지스틱 회귀모형은 확률 θ가 설명변수 $X=(X_1,\ldots,X_p)$의 함수로 주어짐을 가정하여 확률모형 $\theta(x)=\Pr(Y=1|X=x)$를 찾는다. 본 연구에서 이분형 반응변수를 두개의 범주에 대응시키고 관련 변수들을 설명변수로 두게 되면, 이분형 분류문제에 로지스틱 회귀모형을 적용할 수 있다. 이를 로지스틱 회귀분류법이라 부른다.

로지스틱 회귀분류법은 이분형 범주를 이분형 반응변수로 간주하여 로지스틱 회귀모형을 이용해 분류규칙(classification rule)을 만드는데 목적이 있다. 이분형 변수의 확률은 설명변수로 모형화 되며, 그 확률모형은 이항분포를 따르는 것으로 가정한다.

로지스틱 회귀모형에서 설명변수로 $\mathbf{x}=(x_1,\ldots,x_p)^T$가 있다고 가정하자. 이 때, 반응변수의 분포는 베르누이 분포(Bernoulli distribution)이며 이를

$$\Pr(Y=y)=\theta(\mathbf{x})^y(1-\theta(\mathbf{x}))^{1-y},\quad y\in\{0,1\}$$

로 둘 수 있다. 베르누이 분포는 이항분포(binomial distribution)에서 시행회수가 $n=1$인 경우에 대응하는 분포이다. 베르누이 분포를 따르는 확률변수의 평균과 분산은 다음과 같다.

$$E(Y|\mathbf{X}=\mathbf{x})=\theta(\mathbf{x}),\qquad Var(Y|\mathbf{X}=\mathbf{x})=\theta(\mathbf{x})(1-\theta(\mathbf{x})). \tag{1}$$

반응변수의 기대값을 설명변수의 선형식으로 표현하는 선형회귀모형(linear regression model)과 동일한 방식으로 성공확률 $\theta(\mathbf{x})$를 모형화 하는 것은 적절치 않다. 식 (1)에서 볼 수 있듯이 베르누이 확률변수의 기대값은 성공확률이며 이는 $0 \leq \theta(\mathbf{x}) \leq 1$이라는 제약을 가지게 되는데, 이를 설명변수의 선형식으로 가정하게 되면 확률값이 가지는 범위를 벗어나기 때문이다. 또한 선형회귀모형의 등분산성(constant variance)을 만족시킬 수 없는데, 이는 식 (1)에서 볼 수 있듯이 분산이 기대값의 함수로 나타나기 때문이다. 이러한 문제를 해결하기 위해 일반적으로 확률의 로짓변환(logit transformation)을 설명변수의 선형함수로 표현한다. 즉,

$$\log\left(\frac{\theta(\mathbf{x})}{1-\theta(\mathbf{x})}\right) \;=\; \alpha + \beta_1 x_1 + \cdots + \beta_p x_p \;=\; \alpha + \mathbf{x}^T\boldsymbol{\beta}. \tag{2}$$

여기서, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$이다. 식 (2)을 통해 성공확률은

$$\theta(\mathbf{x}) \;=\; \frac{\exp(\alpha + \mathbf{x}^T\boldsymbol{\beta})}{1+\exp(\alpha + \mathbf{x}^T\boldsymbol{\beta})} \;=\; \{1+\exp(-\alpha - \mathbf{x}^T\boldsymbol{\beta})\}^{-1}$$

로 재표현되며, 이는 0과 1사이값을 취하는 확률의 제약식을 만족함을 쉽게 알 수 있다.

로지스틱 회귀모형에서의 모수의 추정은 일반적으로 최대우도법(maximum likelihood method)을 이용한다. n개의 자료가 있는 경우, n 쌍의 반응변수와 설명변수를 $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$로 표기하자. 모수에 대한 우도함수(likelihood function)는 다음과 같이 주어진다.

$$L(\alpha, \beta_1, \ldots, \beta_p) \;=\; \prod_{i=1}^{n} \Pr(Y = y_i | \mathbf{X} = \mathbf{x}_i) \;=\; \prod_{i=1}^{n} \theta(\mathbf{x}_i)^{y_i}(1-\theta(\mathbf{x}_i))^{1-y_i}.$$

위의 우도함수에 대한 로그우도함수는 아래와 같다.

$$\begin{aligned}\log(L) &= \sum_{i=1}^{n} \log(\Pr(Y = y_i|\mathbf{X} = \mathbf{x}_i)) \\ &= \sum_{i=1}^{n} \left[y_i(\alpha + \mathbf{x}_i^T\boldsymbol{\beta}) + \log\left\{1 + \exp(\alpha + \mathbf{x}_i^T\boldsymbol{\beta})\right\}\right]. \qquad (3)\end{aligned}$$

로그우도함수 (3)을 최대화하는 모수$(\alpha, \beta_1, \ldots, \beta_p)$는 해석적(analytic)으로 구할 수 없으며, Newton-Raphson 방법 등을 이용한 반복 알고리듬을 통해 수치적(numerical)으로 구한다. 이는 SAS, R 등 통계분석 프로그램에 구현되어 있다(SAS의 GENMOD procedure 및 R의 glm함수). 로지스틱 회귀모형으로부터 구한 모수 추정량의 특성 및 모수의 해석은 표준적인 통계교과서에 잘 기술되어 있다 (Montgomery et al., 2006).

로지스틱 회귀모형은 분류문제로 전환할 수 있다. 식 (3)을 최대화하는 추정량 $\hat{\alpha}$ 및 $\hat{\boldsymbol{\beta}}$를 찾은 후, 새로운 자료 $\mathbf{x}$에 대하여 함수 $f(\mathbf{x}) = \hat{\alpha} + \mathbf{x}^T\hat{\boldsymbol{\beta}}$를 구성한다. 새로운 자료 $\mathbf{x}$에 대하여 로지스틱 회귀모형을 기반으로한 분류규칙(classification rule)은 아래와 같다.

$$\hat{y} = \begin{cases} 1 & \text{if } f(\mathbf{x}) \geq 0 \\ 0 & \text{if } f(\mathbf{x}) < 0 \end{cases}. \qquad (4)$$

여기서 $f(\mathbf{x})$를 분류함수라 부른다.

2.2 벌점회귀모형 및 희소회귀모형

선형회귀분석에서 자료의 개수(n)보다 설명변수의 개수(p)가 더 큰 경우, 회귀계수의 추정량은 유일하게 결정되지 않는다. 또한 $n > p$일지라도, n이 p보다 월등히 크지 않은 경우 회귀계수 추정량의 분산은 매우 크게 나타난

다. 이는 추정된 회귀계수의 신뢰성이 낮아지게 되고 결과적으로 모형으로부터 높은 예측력을 기대하기가 어렵다. 따라서 이러한 단점을 보완하기 위해 L_2 벌점함수(L_2 penalty function)를 도입하여 예측력을 높이는 방법으로 ridge 회귀모형을 많이 이용한다 (Hoerl and Kennard, 1970a,b). Ridge 회귀모형은 회귀계수 추정량에 약간의 편차(bias)을 주는 대신 분산을 크게 줄임으로써 예측력을 향상 시킨다. Tibshirani (1996)는 편의추정량(biased estimator)을 얻기 위한 방법으로 ridge regression에서의 L_2 벌점함수 대신, L_1 벌점함수를 부가한 lasso 회귀모형을 제안하였다. Lasso 회귀모형 또한 추정량의 분산을 줄이는 효과 및 추정량에 편차를 부여하는 면에서 ridge 회귀모형과 동일한 효과를 얻는다. 하지만, ridge 회귀모형에 의한 추정량은 추정값으로 0을 가지지 못하는 반면, lasso 회귀모형으로 부터 얻게되는 추정량은 추정값으로 쉽게 0을 가지게 된다. 이는 결과적으로 변수선택(variable selection)의 효과를 주게 되며, 모형선택(model selection) 기법으로 최근에 많이 활용되고 있다.

아래와 같은 선형회귀모형을 가정하자.

$$y_i = \alpha + \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i. \quad (i = 1, 2, \ldots, n)$$

최소제곱추정량(ordinary least square estimator, OLS)은 잔차제곱합(residual sum of squares, RSS)을 최소로 하는 모수값으로 아래와 같이 얻는다.

$$\underset{\alpha,\beta_1,\cdots,\beta_p}{\arg\min} \sum_{i=1}^{n} \left(y_i - \alpha - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 .$$

Lasso 회귀추정량은 아래와 같이 특정 제약조건 하에서 잔차제곱합을 최

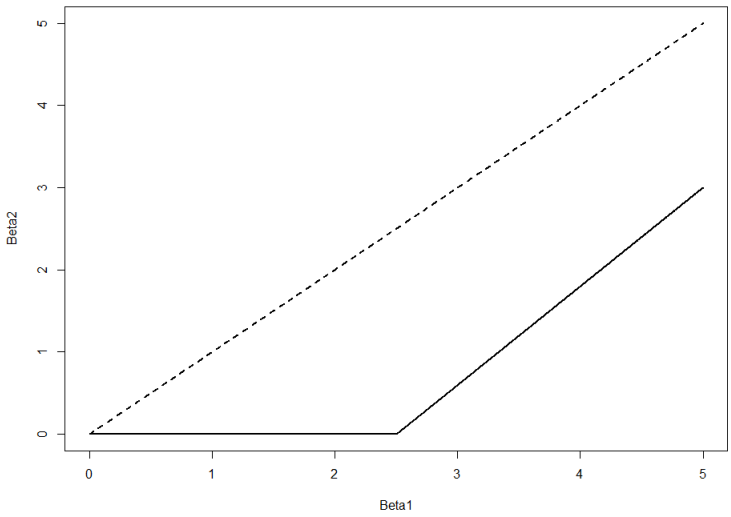
소로 하는 모수값이다.

$$\underset{\alpha,\beta_1,\cdots,\beta_p}{\arg\min} \sum_{i=1}^{n} \left(y_i - \alpha - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \quad \text{subject to } \sum_{j=1}^{p} |\beta_j| \leq t. \tag{5}$$

여기서 t는 양수로써 모수의 크기를 조절하는 조절 모수(regularization parameter)이다. t가 충분히 크면 모수에 대한 크기 제약이 없기 때문에 최소제곱추정량과 동일한 추정량을 준다. 반면에 t가 작아지면 추정량의 크기가 작아지므로 중요하지 않은 변수의 회귀계수는 0으로 회귀하게 되며, $t = 0$이면 모든 추정치는 0이 된다. t가 작아지는 경우 lasso 회귀계수는 정확히 0값인 추정치를 얻을 수 있게 된다. < 그림1 >은 최소제곱추정량값에 따른 해당 lasso 추정량의 변화를 그림으로 나타낸 것이다. 최소제곱추정량(가로축)에 비해 lasso 추정량(세로축)은 일괄적으로 작게 추정되며 최소제곱추정량값이 일정 수준 이하의 값을 가지게 되면 lasso 회귀계수는 정확이 0으로 주어진다. < 그림1 >에서 두 관계를 실선으로 표시하였고 점선은 최소제곱추정량과의 비교를 위한 참조선이다.

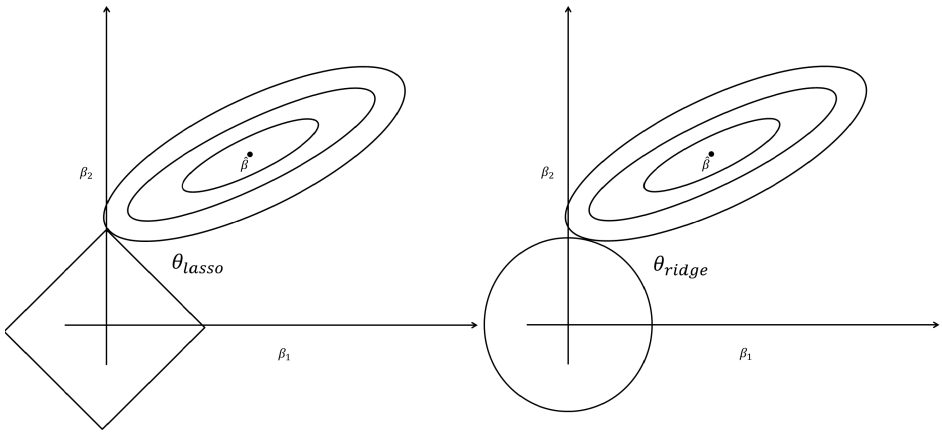
Lasso 회귀계수 추정치가 정확이 0이 나타나게 되는 메커니즘의 이해를 돕기 위해 $p = 2$인 경우를 살펴보자. < 그림2 >에서는 lasso(왼쪽) 및 ridge(오른쪽) 회귀계수가 얻어지는 과정에 대한 기하학적 묘사를 주고 있다. 두 방법 모두 동일한 RSS를 가지고 있으며 이는 그림 상에서 타원형태의 등고선으로 표시되어 있다. 등고선의 가장 낮은 위치를 점으로 표시하였고 이는 제약식 없이 얻어지는 최소제곱추정량($\hat{\boldsymbol{\beta}}$)에 대응한다. Lasso 회귀모형의 경우, $|\beta_1| + |\beta_2| \leq t$의 영역 하에서 RSS의 최소값을 찾는다. 이 영역을 왼쪽 그림에서 묘사한 바와 같이 마름모 형태로 표현할 수 있다. 이러한 제약영역은 모서리가 축 위에 놓이게 되어 있으며 그 영역이

그림 1: 최소제곱회귀추정량 및 lasso 추정량의 비교그림



뾰족하게 튀어 나와 있다. $\hat{\boldsymbol{\beta}}$는 마름모 영역에 놓이지 않으므로 lasso 회귀계수가 될 수 없다. 제약영역에서 RSS의 최소값을 탐색하기 위해 등고선의 높이를 최소제곱추정량으로부터 점차 위로 변화시킨다. 이때 제약영역의 점 중에서 증가하는 등고선을 최초로 만나게 되는 점이 lasso 회귀계수추정량이다. 그림에서 보듯 제약영역의 꼭지점에 닿을 확률이 높음을 알 수 있다. 꼭지점은 축의 일부분의 값이 0에 대응하는 점에 놓여 있으므로, lasso 회귀계수추정량은 그 일부분이 0이 될 가능성이 매우 높게 된다. 비교를 위해 < 그림2 >의 오른쪽에 ridge 회귀계수추정량을 얻는 과정을 같이 묘사하였다. Ridge 회귀모형에서는 제약조건이 $\beta_1^2 + \beta_2^2 \leq t$로 주어지며 이는 그림상에 보듯이 원형의 제약영역으로 표시된다. Ridge 회귀계수추정량은 RSS의 등고선을 높히면서 최초로 제약영역에 닿는 점에 대응한다. 제약영

그림 2: lasso및 ridge 추정량의 일반적인 형태



역이 원형으로 주어지므로 축위에 놓인 점이 ridge 회귀계수가 될 확률은 매우 낮음을 알 수 있다.

식 (5)에 주어진 Lasso 회귀분석을 위한 최소화 문제는 벌점함수를 이용한 목적함수의 최소화 문제로 표현할 수 있다. 이는 아래와 같다.

$$\underset{\alpha,\beta_1,\cdots,\beta_p}{\arg\min} \sum_{i=1}^{n} \left(y_i - \alpha - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

여기서 $\lambda(\geq 0)$는 조절모수로써 식 (5)에서의 조절모수인 t와 비슷한 역할을 한다. λ가 0에 가까우면 벌점함수의 역할이 축소되어 자유롭게 $\boldsymbol{\beta}$가 추정되므로 최소제곱추정량과 동일한 추정량을 얻게 되며, $\lambda \to \infty$이면 모든 회귀계수가 0으로 수렴하게 된다.

Lasso 회귀분석은 희소회귀모형(sparse regression)의 한 종류이다. Lasso

회귀분석의 추정량의 성질을 개선한 또 다른 희소회귀모형으로써 SCAD 벌점함수(Fan and Li, 2001)를 이용한 희소회귀모형 또한 많이 이용된다. 이에 대한 자세한 성질 및 추정방법은 표준적인 기계학습 교재에 잘 기술되어 있다 (Hastie et al., 2009; Murphy, 2012).

2.3 주성분 회귀모형

회귀분석에서 설명변수들 사이에 높은 상관관계가 존재하게 되면 다중공선성(multicollinearity) 현상이 발생한다. 다중공선성이 존재하는 경우, 회귀계수 추정량의 분산은 커지므로 회귀계수 안정성을 보장할 수 없다. 이러한 문제점을 해결하기 위하여 주성분 회귀모형을 사용한다. 주성분 회귀모형은 설명변수들에 대한 주성분분석을 통해 수행된다. 주성분분석은 다차원 상의 점으로 표현되는 데이터의 분산을 최대로 하는 선형결합(linear combination)을 순차적으로 찾아준다. 설명변수의 총변동을 주성분분석을 통해 각각의 주성분 방향으로 분해할 수 있다. 주요주성분은 본래의 설명변수의 특성을 최대한 유지하면서 낮은 차원에서의 묘사가 가능하므로 차원축소의 주요 방법으로 많이 이용된다. 하위주성분은 설명변수 간에 강한 상관성을 가지는 차원을 나타낸다. 따라서 주성분회귀모형에서는 하위주성분을 제거하고 주요주성분만을 설명변수로 고려하여 회귀분석을 수행하므로 다중공선성의 문제를 해결할 수 있다. 주성분 회귀모형은 ridge 회귀모형과 더불어 다중공선성을 해결하는 방법으로 많이 이용되고 있다. 주성분 회귀모형은 주성분의 일부만 이용함으로써 주성분 회귀모형을 통해 얻은 주성분 회귀계수는 편의추정량이 되지만, 추정량의 분산을 줄이는

효과가 있어 예측력을 높인다는 점에서 ridge 회귀모형과 유사하다.

크기가 p인 대칭행렬(symmetric matrix) $\mathbf{A} = (a_{ij})$ $(i, j = 1, 2, \ldots, p)$를 고려하자. 음이 아닌 스칼라 λ와 $\mathbf{0}$이 아닌 벡터 $\mathbf{v}$가

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$$

를 만족하면 λ를 $\mathbf{A}$의 고유치(eigenvalue), $\mathbf{v}$를 고유벡터(eigenvector)라 한다. 위의 관계를 만족하는 고유벡터는 p개가 존재하며 각기 대응하는 고유치를 가지고 있다. p개의 고유치를 크기 순으로 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$라 하고 대응되는 고유벡터를 $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_p$라 하면, 행렬 $\mathbf{A}$는 아래와 같이 분해가 된다.

$$\mathbf{A} = \sum_{k=1}^{p} \lambda_k \mathbf{v}_k \mathbf{v}_k^T = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T. \tag{6}$$

여기서 $\mathbf{\Lambda} = \mathrm{diag}(\lambda_k)_{k=1,2,\ldots,p}$ 이고 $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_p)$이다. $\mathbf{P}$는 직교행렬(orthogonal matrix)로써 $\mathbf{P}^T\mathbf{P} = \mathbf{P}\mathbf{P}^T = \mathbf{I}_p$를 만족한다. 식 (6)에서 차원 p보다 작은 K에 대해 행렬 $\mathbf{A}$를 다음과 같이 근사시킬 수 있다.

$$\mathbf{A} \approx \sum_{k=1}^{K} \lambda_k \mathbf{v}_k \mathbf{v}_k^T.$$

위에서 설명한 주성분분석을 이용하여 주성분 회귀모형을 설명한다. p개의 설명변수로 이루어진 $(n \times p)$ 크기의 자료행렬 $\mathbf{X} = (x_{ij}) = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p)$을 고려하자. $\mathbf{X}$의 열은 자신의 표본평균을 빼어 중심화(centering)가 되어 있다고 가정한다. 또한 반응변수 $\mathbf{y} = (y_1, \ldots, y_n)^T$도 중심화 되어 있다고 가정하자. 설명변수의 공분산행렬은 $\mathbf{A} = \mathbf{X}^T\mathbf{X}/(n-1)$로 표현된다. 공분산행렬에 대한 고유벡터를 $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p$라 하자. 각 고유벡터 $\mathbf{v}_k$에 대하여

대응되는 주성분(principal component)은 다음과 같이 얻는다.

$$\mathbf{z}_k = \mathbf{X}\mathbf{v}_k. \quad (k = 1, 2, \ldots, p)$$

크기 n인 벡터 $\mathbf{z}_k = (z_{1k}, z_{2k}, \ldots, z_{nk})^T$는 p차원 공간 상의 점으로 표현되는 n개의 자료의 설명변수를 고유벡터 $\mathbf{v}_k$ 방향으로 직교사영(orthogonal projection)된 점의 좌표이다. 주성분분석에서 $\mathbf{v}_k$를 k번째 주성분방향(PC loading)으로 부르며, 해당 $\mathbf{z}_k$를 k번째 주성분이라 부른다. 즉, i번째 자료의 k번째 주성분은 $z_{ik} = \tilde{\mathbf{x}}_i^T \mathbf{v}_k$로 주어지게 되며, 여기서 $\tilde{\mathbf{x}}_i$는 $\mathbf{X}$의 i번째 행이다. 전체 주성분은 다음의 행렬계산으로 쉽게 얻어진다.

$$\mathbf{Z} = (z_{ik}) = \mathbf{X}\mathbf{V}.$$

여기서 $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_p)$ 이다. 여기서 $\mathbf{Z}^T\mathbf{Z} = \mathbf{V}^T\mathbf{X}^T\mathbf{X}\mathbf{V} = (n-1)\mathbf{V}^T\mathbf{A}\mathbf{V} = (n-1)\mathbf{V}^T\mathbf{V}\mathbf{\Lambda}\mathbf{V}^T\mathbf{V} = (n-1)\mathbf{\Lambda}$로 주어짐을 확인할 수 있다. 중심화 되어 있는 설명변수와 반응변수로 이루어진 자료에 대한 선형회귀모형은 절편(intercept)이 없는 아래의 모형으로 표현 가능하며, 이를 주성분행렬($\mathbf{Z}$)과 주성분방향행렬($\mathbf{V}$)로 표현하면 아래와 같다.

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ &= \mathbf{Z}\mathbf{V}^T\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ &= \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}. \end{aligned} \tag{7}$$

여기서 $\boldsymbol{\gamma} = \mathbf{V}^T\boldsymbol{\beta}$이다. 즉 설명변수 대신 주성분을 설명변수로 하는 회귀분석의 결과는 본래의 회귀분석과 동등함을 알 수 있다. 새로운 회귀식 (7)의 모수 $\boldsymbol{\gamma}$의 추정치는

$$\hat{\boldsymbol{\gamma}} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{y}$$

로 주어지며, 이를 이용하여 본래의 회귀계수의 추정치는

$$\hat{\boldsymbol{\beta}} = \mathbf{V}\hat{\boldsymbol{\gamma}}$$

를 통해 회복할 수 있다. 주성분 회귀모형은 식 (7)에서 설명변수의 변동의 대부분을 설명하는 주요주성분을 사용하여 회귀분석을 수행하는 방법이다. 즉, 전체 p개의 주성분을 이용하는 대신, 처음 K개의 주성분만을 설명변수로 고려한다. 이를 이해하기 위해 우선 $\boldsymbol{\beta}$ 추정량의 분산을 계산하면 아래와 같다.

$$Var(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} = \sigma^2\mathbf{V}\boldsymbol{\Lambda}^{-1}\mathbf{V}^T = \sigma^2\sum_{k=1}^{p}\frac{1}{\lambda_k}\mathbf{v}_k\mathbf{v}_k^T.$$

만일 하위고유치값이 0에 가깝다면 추정량의 분산은 매우 커지게 됨을 알 수 있다. 따라서 고유치의 크기가 0에 가까운 하위주성분을 제거한다면 추정치의 분산을 줄일 수 있다. 참고로 0이 아닌 고유치의 개수는 자료행렬 $\mathbf{X}$의 rank보다 클 수 없음이 알려져 있다. 이를 $r = \text{rank}(\mathbf{X})$라 표기하자. r은 $\min\{n, p\}$보다 작거나 같다. 주성분 회귀모형에서 상위주성분을 취하는 일반적인 기준은 설명변수의 변동의 90% 이상 설명하는 최소 개수의 상위주성분을 취하는 것이다. 고유치가 설명변수의 변동량을 나타내므로, 다음과 같이 K를 정한다.

$$K = \min\left\{k : \frac{\sum_{l=1}^{k}\lambda_l}{\sum_{l=1}^{p}\lambda_l} \geq 0.9\right\}.$$

$\mathbf{Z}_{(k)} = (\mathbf{z}_1, \cdots, \mathbf{z}_K)$를 $\mathbf{Z}$의 처음 K개의 열로 이루어진 부분행렬이라 하자. 여기서 K는 위의 기준으로 선정된 값이라 하자. 이를 이용한 주성분 회귀모형

$$\mathbf{y} = \mathbf{Z}_{(K)}\boldsymbol{\gamma}_{(K)} + \boldsymbol{\epsilon}$$

의 추정량은

$$\hat{\boldsymbol{\gamma}}_{(K)} = (\mathbf{Z}_{(K)}^T \mathbf{Z}_{(K)})^{-1} \mathbf{Z}_{(K)}^T \mathbf{y}$$

이다. 여기서 $\mathbf{Z}_{(K)}^T \mathbf{Z}_{(K)} = \mathbf{V}_{(K)}^T \mathbf{X}^T \mathbf{X} \mathbf{V}_{(K)} = \boldsymbol{\Lambda}_{(K)} = \text{diag}(\lambda_1, \cdots, \lambda_K)$ 임을 확인할 수 있다. 이로부터 얻어지는 주성분 회귀추정량은 $\hat{\boldsymbol{\beta}}^{PCR} = \mathbf{V}_{(K)} \hat{\boldsymbol{\gamma}}_{(K)}$로 주어지며, 추정량의 분산은

$$\begin{aligned} Var(\hat{\boldsymbol{\beta}}^{PCR}) &= \mathbf{V}_{(K)} Var(\hat{\boldsymbol{\gamma}}_{(K)}) \mathbf{V}_{(K)}^T \\ &= \sigma^2 \mathbf{V}_{(K)} \boldsymbol{\Lambda}_{(K)}^{-1} \mathbf{V}_{(K)}^T \\ &= \sigma^2 \sum_{k=1}^{K} \frac{1}{\lambda_k} \mathbf{v}_k \mathbf{v}_k^T \end{aligned}$$

가 되어 하위고유치가 사라짐으로 분산축소의 효과를 얻을 수 있음을 알 수 있다. 주성분 회귀분석에 대한 더 자세한 내용은 Montgomery et al. (2006)을 참조하길 바란다.

3 벌점 주성분 로지스틱 회귀분류법

본 연구에서는 이항분류(binary classification)문제에서 주성분 회귀분석의 개념을 로지스틱 회귀분류에 도입하는 것이다. 주요주성분을 이용하는 대신, 변수선택법을 도입하여 주성분을 자동으로 선정하도록 하여 분류율을 높히는 주성분을 선별하여 취하고 그렇지 않은 주성분은 모형에서 제외하도록 한다. 이를 통해 로지스틱 회귀분류의 분류율을 향상시키도록 한다.

크기 $(n \times p)$인 자료행렬 $\mathbf{X} = (\mathbf{x}_1^T, \mathbf{x}_2^T, \cdots, \mathbf{x}_n^T)$는 p개의 변수를 가진 n개의 자료로 구성되었다고 가정하자. 즉, $\mathbf{x}_i$ $(i = 1, 2, \ldots, n)$은 i번째 관측치의

설명변수의 모임으로 해석할 수 있다. $\mathbf{X}$의 각 열은 중심화 되었음을 가정하고 full-column rank 혹은 full-row rank라 가정하자. i번째 자료는 두개의 범주(class) 중에 하나에 속하는 것으로 가정하고 이를 $\{0,1\}$로 표기한다. 자료행렬을 주성분으로 변환하게 된다면 이는 설명변수의 선형변환으로 표현되며 이를 다음과 같이 표기할 수 있다.

$$\mathbf{Z} = \mathbf{X}\mathbf{V}, \quad \mathbf{V} \in \mathbb{R}^{p\times r}.$$

여기서 $r = \mathrm{rank}(\mathbf{X})$이다. 따라서 $\mathbf{z}_i$를 $\mathbf{Z}$의 i번째 행으로 두면 이는 i번째 자료에 대응하는 주성분에 대응한다. 이로부터 $\mathbf{z}_i = \mathbf{V}^T\mathbf{x}_i$의 관계를 가짐을 알 수 있다. 따라서 주성분을 이용한 로지스틱 회귀분류의 분류함수는 다음과 같이 주성분으로 표현할 수 있다.

$$f(\mathbf{x}) \;=\; \alpha + \mathbf{x}^T\boldsymbol{\beta} \;=\; \alpha + \mathbf{z}^T\boldsymbol{\gamma}.$$

여기서 $\boldsymbol{\gamma} = \mathbf{V}^T\boldsymbol{\beta} \in \mathbb{R}^r$이다. 본 연구에서는 분류함수를 얻기 위해 로지스틱 회귀분류법과 동일한 목적함수로써 음의 베르누이 우도함수(negative Bernoulli likelihood function)를 최소화하는 방법을 취한다. 여기에 설명변수로써 주성분을 이용하고 희소회귀모형의 방법으로 사용되는 주성분을 선별하는 과정을 취한다. 즉, 회귀계수를

$$(\hat{\alpha}, \hat{\boldsymbol{\gamma}}) = \arg\min_{\alpha,\boldsymbol{\gamma}} L(\alpha, \boldsymbol{\gamma}) + \mathsf{pen}_\lambda(\boldsymbol{\gamma})$$

로 구하고 이를 이용하여 분류함수를 구성한다. 여기서

$$L(\alpha, \boldsymbol{\gamma}) = -\sum_{i=1}^{n} \left[y_i(\alpha + \mathbf{z}_i^T\boldsymbol{\gamma}) + \log\left\{1 + \exp(\alpha + \mathbf{z}_i^T\boldsymbol{\gamma})\right\}\right]$$

이며 $\text{pen}_\lambda(\boldsymbol{\gamma})$는 희소회귀를 위한 벌점함수이다. 이를 통해 얻은 추정량을 이용하여 얻게 되는 분류함수는

$$f(\mathbf{x}) = \hat{\alpha} + \mathbf{x}^T \mathbf{V} \hat{\boldsymbol{\gamma}} \tag{8}$$

이며 이를 이용한 분류규칙은 식 (4)와 동일하다. 희소회귀를 위한 벌점함수는 주성분에 대응하는 회귀계수 $\boldsymbol{\gamma}$의 일부분을 정확히 0으로 추정하게 하여 주성분을 선별하는 역할을 한다. 이러한 분류법을 벌점 주성분 로지스틱 회귀분류법(penalized principal component logistic regression classification, PPCLR)이라 명명한다. 식 (8)에 주어진 분류함수를 이용한 분류법은 다음의 성질을 가진다.

1. **(분류와 연관된 주성분 선택)** 본 방법을 통한 주성분의 선정시, 주요 주성분이 아니라 하더라도 분류율에 영향을 주는 하위주성분도 모형에 포함되어 분류규칙에 기여를 한다. 이는 주요주성분만을 사용하는 주성분 회귀분석법과는 달리 영향력 있는 하위주성분을 버리지 않음으로 분류율 향상에 도움을 주게 된다.

2. **(차원축소)** 선형판별분석(linear discriminant analysis, LDA)은 변수의 차원을 축소하여 분류율이 최대화 되는 부분공간을 변수공간 내에서 찾는다. 본 분류법 또한 선형판별분석과 유사한 부분공간을 찾게 되는데, 이 부분공간은 변수들의 주성분 방향벡터(고유벡터)에 의해 생성된다.

3. **(분류공간의 차원제약)** 부분공간의 차원은 선택되는 희소회귀분석법의 절차에 의해 선택된 주성분의 개수로써 자동으로 결정된다. 선

택되는 주성분의 개수의 제한은 없다. 반면, 동일한 차원축소를 이용하는 선형판별분석은 일반적으로 범주의 개수보다 적게 얻어진다. 즉, 분류문제에서 범주가 G개 인 경우, 선형판별분석에서 취하는 부분공간의 크기는 $G-1$보다 작거나 같다 (Hastie et al., 2009). 본 연구에서 다루는 이항분류 문제에서는 $G=2$이므로 공간의 차원은 1이 된다. 이에 반해 벌점 주성분 로지스틱 회귀분류법에 의해 얻어지는 부분공간은 G보다 많을 수도 있고 따라서 더 많은 정보를 이용함으로, 복잡한 형태의 자료가 있는 경우, 혹은 고차원자료에서의 분류문제에서 선형판별분석에 비해 분명한 이점을 지닌다.

4. **(해석의 용이함)** 벌점 주성분 로지스틱 회귀분류법에 의한 부분공간은 주성분에 의해 표현된다. 주성분은 변수공간의 변동을 직교방향(orthogonal direction)으로 나누어 설명한다. 즉, 부분공간은 변수공간의 변동의 직교방향에 따라 놓이게 되므로, 해당 부분공간은 주성분분석과 동일한 해석의 이점을 지닌다.

본 연구에서는 2가지의 벌점함수(Lasso, SCAD)를 고려하였다. 벌점함수의 형태는 아래와 같다.

- Lasso

$$\mathsf{pen}_{\lambda}^{Lasso}(\boldsymbol{\gamma}) = \lambda \sum_{j=1}^{p} |\gamma_j|.$$

- SCAD

$$\mathsf{pen}_{\lambda}^{SCAD}(\boldsymbol{\gamma}) = \sum_{j=1}^{p} p_{\lambda}(|\gamma_j|; a).$$

여기서 $p_\lambda(x;a)$는 다음과 같다.

$$\begin{aligned} p_\lambda(x;a) &= 2\lambda x I(x \le \lambda) - \frac{x^2 - 2a\lambda x + \lambda^2}{a-1} I(\lambda < x \le a\lambda) \\ &\quad +(a+1)\lambda^2 I(x > a\lambda). \end{aligned}$$

SCAD 벌점함수에서는 조절모수 λ이 외에 또 다른 조절모수 a가 주어져 있다. 본 연구에 사용시 a는 Fan and Li (2001)에서 제안한 값을 이용하였다. Lasso 벌점함수를 이용할 때의 벌점 주성분 로지스틱 회귀분류법을 PPCLR-L, SCAD 벌점함수를 이용할 경우 PPCLR-S라 구분하여 부르기로 한다.

4 모의 실험

본 장에서는 3장에서 소개한 PPCLR-L 그리고 PPCLR-S을 모의실험에 적용하고 기존의 다른 분류법들과 성능을 비교한다.

4.1 모의 실험 설계

본 논문에서는 통계프로그램 R을 사용하여 다음과 같이 모의실험 상황을 설정하였다. 변수의 개수가 p인 두 개의 다변량정규분포로부터 두개의 범주를 가지는 자료를 생성한다. 생성된 변수를 $\mathbf{x}_1$ 와 $\mathbf{x}_2$라 하면

$$\mathbf{x}_1 \sim MVN_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \qquad \mathbf{x}_2 \sim MVN_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$$

이다.

두 그룹의 공분산 행렬이 같은 경우($\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$)와 두 그룹의 공분산행렬이 다른 경우($\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$)로 나누어 모의실험을 진행한다. 공분산 행렬이 같은 경우는 일반적으로 선형판별분석에서 흔히 가정하는 경우이며, 공분산행렬이 다른 경우는 선형판별분석에서 공통공분산 추정이 좋지 않게 되는 경우이다. 주성분 추정시 공통공분산(pooled covariance)을 이용하게 되므로 높은 오분류율이 나타날 것으로 예상된다. Kondylis and Whittaker (2008)에서 사용한 공분산행렬 $\mathbf{\Sigma} = (\Sigma_{ij})$를 다음과 같이 고려하자.

$$\Sigma_{ij} = \frac{1}{|i-j|+1}, \quad (i, j = 1, 2, ..., p)$$

위 공분산은 변수간에 높은 양의 상관성을 가진다. 위의 공분산행렬에 대한 고유치분해는

$$\mathbf{\Sigma} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T = \sum_{k=1}^{p} \lambda_k \mathbf{v}_k \mathbf{v}_k^T$$

로 표현되며, 여기서 $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \cdots, \lambda_p)$는 고유치를 원소로 하는 대각행렬이고 $\mathbf{V} = (\mathbf{v}_1, \cdots, \mathbf{v}_p)$는 해당 고유백터를 열로 하는 직교행렬이다. 이를 이용하여 $\mathbf{x}_1$ 와 $\mathbf{x}_2$를 아래의 조건에서 생성한다.

(S1) 두 그룹이 동일한 공분산 행렬을 가지며 두그룹의 평균이 상위 주성분 방향으로 놓인 경우:

$\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}$, $\boldsymbol{\mu}_1 = \boldsymbol{\mu}$, $\boldsymbol{\mu}_2 = -\boldsymbol{\mu}$이고

$$\boldsymbol{\mu} = \frac{\sum_{k=1}^{K} \sqrt{\lambda_k}\mathbf{v}_k}{\sqrt{\sum_{k=1}^{K} \lambda_k}}.$$

여기서 K는 공분산 전체 변동의 90%를 설명하는 상위 주성분의 최저개수에 해당한다.

(S2) 두 그룹이 동일한 공분산 행렬을 가지며 두 그룹의 평균이 하위 주성분 방향으로 놓인 경우:

$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, $\boldsymbol{\mu}_1 = \boldsymbol{\mu}$, $\boldsymbol{\mu}_2 = -\boldsymbol{\mu}$이고

$$\boldsymbol{\mu} = \frac{\sum_{k=K+1}^{p} \sqrt{\lambda_k}\mathbf{v}_k}{\sqrt{\sum_{k=K+1}^{p} \lambda_k}}$$

여기서 K은 **(S1)**에서와 동일하게 정의된다.

(S3) 두 그룹이 상이한 공분산 행렬을 가지며 두 그룹의 평균이 특정 2개의 주성분 방향으로 놓인 경우:

$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}$이고, $\boldsymbol{\Sigma}_2$를 다음과 같이 고유치를 교환하여

$$\boldsymbol{\Lambda}_\pi = \mathrm{diag}(\lambda_p, \lambda_{p-1}, \lambda_3, \cdots, \lambda_{p-2}, \lambda_2, \lambda_1)$$

를 만들고 $\boldsymbol{\Sigma}_2 = \mathbf{V}\boldsymbol{\Lambda}_\pi\mathbf{V}^T$로 한다. 각 그룹의 평균은

$$\boldsymbol{\mu} = \frac{\sqrt{\lambda_1 + \lambda_p}\mathbf{v}_1 + \sqrt{\lambda_2 + \lambda_{p-1}}\mathbf{v}_2}{\sqrt{\lambda_1 + \lambda_2 + \lambda_{p-1} + \lambda_p}}$$

를 사용하여 $\boldsymbol{\mu}_1 = \boldsymbol{\mu}$, $\boldsymbol{\mu}_2 = -\boldsymbol{\mu}$로 설정한다.

(S4) 두 그룹이 상이한 공분산 행렬을 가지며 두 그룹의 평균이 공통공분산의 제일주성분 방향에 놓인 경우:

$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}$로 하고 $\boldsymbol{\Sigma}_2 = \mathbf{V}\boldsymbol{\Lambda}_\pi\mathbf{V}^T$로 하되 $\boldsymbol{\Lambda}_\pi = \mathrm{diag}(\lambda_{\pi_k})$로 둔다. 여기서 π는 순서 $\{1, 2, \cdots, p\}$에 대한 임의의 순열(permutation)을 의미하고 π_k은 주어진 순열에서 k번째 원소를 의미한다. 두 공분산 $\boldsymbol{\Sigma}_1$과 $\boldsymbol{\Sigma}_2$의 공통공분산을 $\boldsymbol{\Sigma}^* = p_1\boldsymbol{\Sigma}_1 + p_2\boldsymbol{\Sigma}_2$로 정의하자. 여기서 p_1과 p_2는 각각 그룹1과 그룹2가 뽑힐 확률이다. 첫번째 고유백터를 $\mathbf{v}_1^*$라 하자. 각 그룹의 평균은 $\boldsymbol{\mu}_1 = \boldsymbol{\mu}$, $\boldsymbol{\mu}_2 = -\boldsymbol{\mu}$로 두고, 여기서 $\boldsymbol{\mu} = \mathbf{v}_1^*$이다.

위의 4가지 상황(**S1**∼**S4**)에 대하여 각 그룹의 자료를 동등하게 $n/2$개씩 생성한다. 따라서 **(S4)**의 경우 $p_1 = p_2 = 1/2$이다. 표본의 개수(n)와 변수 개수(p)는 아래와 같이 설정하였다.

- 자료 개수가 변수 개수보다 많은 경우: $p = 10, n = 100, 200, 400, 800, 1600$
- 변수 개수가 자료 개수보다 많은 경우: $n = 100, p = 100, 200, 400, 800, 1600$

본 연구에서 제안한 2가지 방법(PPCLR-L, PPCLR-S)과 비교할 4가지 이분분류법은 아래와 같다.

- SVM(support vector machine) : radial basis kernel를 커널함수로 이용. kernlab 패키지 이용.
- PLDA(penalized linear discriminant analysis) : 벌점함수로 ridge penalty를 이용 (Hastie et al., 1995). mda 패키지 이용.
- PLR(penalized logistic regression classification) : 벌점함수로 ridge penalty를 이용 (Friedman et al., 2008). glmnet 패키지 이용.
- PCLR(principal component logistic regression classification) : 변수의 주요 주성분(총변동 90%설명)을 설명변수를 이용. glmnet 패키지 이용.

고려하는 방법들 중 PCLR을 제외한 모든 방법은 각기 모형선택(model selection)을 위한 조절모수를 포함하고 있다. 이는 10-fold cross validation(CV)을 통해 최적의 모수를 선택하였으며 가급적 사용하는 패키지에서 제공되는 기본옵션을 이용하였다.

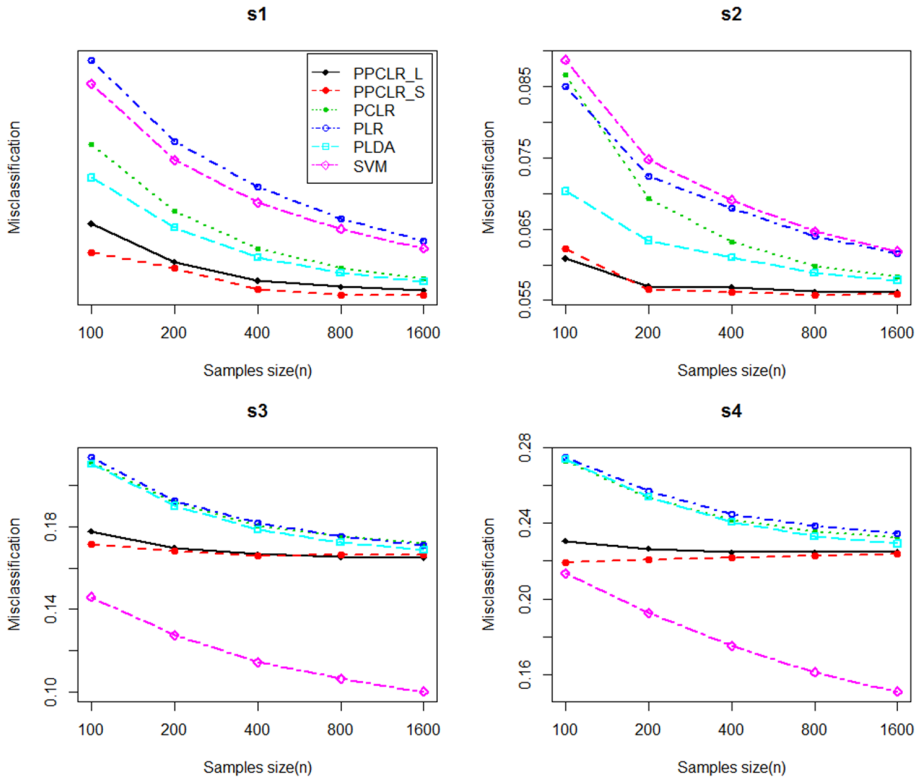
4.2 모의 실험 결과

설명한 두가지 방법론과 비교할 4가지 이분분류법을 통해 분류함수를 얻었다. 4가지 상황에 대하여 1,000회 자료를 반복생성하여 계산된 오분류율의 평균을 < 그림3 > 및 < 그림4 >에 수록하였다. < 그림3 >은 고정된 변수수($p = 10$)에 대하여 상대적으로 큰 표본수(n)를 상정하여 표본의 크기를 달리 하였을 때의 오분류율을 보여준다. 이를 통해 오분류율은 자료의 크기가 커짐에 따라 오분류율의 평균은 작아짐을 확인할 수 있다. 상황 **S1** 및 **S2**에서 PPCLR-L 및 PPCLR-S가 다른 방법론에 비해 낮은 오분류율을 보여주고 있음을 알 수 있으며, 반면에 상황 **S3** 및 **S4**에서는 SVM이 더 좋은 성능을 보여주고 있을을 확인할 수 있다. SVM이 공분산행렬에 대한 특정한 가정없이 수행되므로 보다 나은 성능을 기대할 수 있으며, 결과에서 확인할 수 있다. 하지만 SVM 다음으로는 PPCLR-L 및 PPCLR-S이 가장 낮은 오분류율을 보인다.

< 그림4 >는 고정된 표본수에 대하여 상대적으로 큰 변수수를 가정한 고차원 자료의 경우로써, 변수의 크기를 달리 하였을 때의 오분류율을 보여주고 있다. 위의 $n > p$ 상황과 유사하게 SVM은 $(n, p) = (100, 100)$ 경우에서 가장 낮은 오분류율을 보여주지만, p가 점차 커지면서 오분류율이 급증하는 것을 확인할 수 있다. **S1** 및 **S2**에서는 PPCLR-L, PPCLR-S, 그리고 PLDA가 서로 비슷한 오분류율을 보이면서 다른 방법들에 비해 낮은 오분류율을 보여준다. 하지만 변수수가 매우 큰 경우($p = 1600$)에는 PLDA의 우월성이 사라짐을 볼 수 있다. 반면에 **S3** 및 **S4**의 경우에는 PPCLR-L 및 PPCLR-S가 상대적으로 좋은 성능을 보여줌을 알 수 있다.

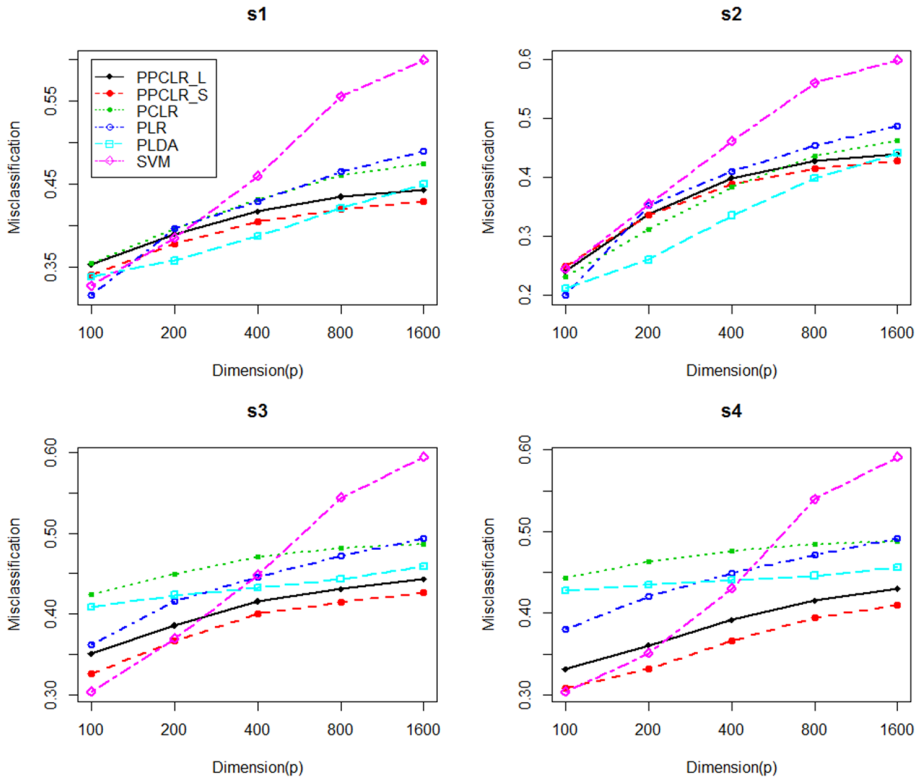
모의실험 결과를 종합적으로 해석해 볼 때, 본 연구에서 제안한 PPCLR-

그림 3: $p = 10$인 경우, n에 따른 오분류율의 평균



L 및 PPCLR-S은 고차원 자료에서 보다 좋은 효율을 보이는 것을 확인할수 있다. 또한 PPCLR-S가 PPCLR-L보다 조금 더 좋은 결과를 보여주는 것을 확인할 수 있는데, 이는 회귀분석의 모수추정이론에서 SCAD 추정량이 Lasso 추정량보다 좋은 성질(변수선택 및 추정량의 일치성에 대한 oracle 성질)을 가지게 되는데 이러한 성질이 분류문제에서 예측력에 좋은 영향을 주어 나타나는 현상으로 생각된다.

그림 4: $n = 100$인 경우, p에 따른 오분류율의 평균



5 자료 분석

본 논문에서 제시한 PPCLR 방법의 비교로서 이 전 장에서 컴퓨터 모의실험을 이용, 비교분석한 방식과 동일하게 실제 응용분야에서 나타나는 자료를 이용하여 비교분석을 실시하였다. UCI machine learning repository (Bache and Lichman, 2013), KEEL webpage [http:sci2s.urg.es/keel] (Alcalá-Fdez et al., 2011)에서 제공하는 data repository, 그리고 GitHub [http://github.com]에서 제공하는 datamicroarray package로 부터 11개의

자료를 수집하였다. 11개의 데이터셋은 물리학, 재료공학, 텍스트마이닝, 유전학 등 다양한 학문분야에서 과학연구 목적으로 실험/수집된 자료로써 각 자료에 대한 설명을 아래에 기술하였다.

1. Inosophere : Goose Bay, Labrador 시스템에 의하여 수집된 레이더 데이터이다. 지구대기의 전리층을 분석하였다.

2. Sonar : 광산과 암석에서 다양한 각도에서 얻은 신호를 분석하여, 이들을 분류한 자료.

3. Spambase : 4597개의 e-mail 메시지를 분석하여서 spam 메일과 아닌 경우를 분류한 자료.

4. Spectheart : cardiac Single Proton Emission 전산화 단층 촬영(SPECT) 영상 진단 자료이며, 0 정상 1환자로 분류.

5. WDBC : 유방 질량의 fine needle aspirate 데이터. 세포 핵의 특징을 통하여 종양의 양성, 악성 확인 자료.

6. Chin : 유방암 환자의 가슴에서의 genome copy number abnormalities (CNAs)을 연구하여 암발생 유무 확인 자료.

7. Chowdary : RNA를 사용한 암발생 유무확인 자료.

8. Gravier : 유방 암의 전이를 예측하기 위하여 168명의 환자를 연구한 자료.

9. Gordon : 폐암발생을 진단하기 위하여 181명의 환자를 연구한 자료.

10. Singh : 수술한자 235 중에서 102명의 좋은 샘플로서 52명의 전립선암 환자와 50명의 전립선암이 아닌 사람을 표본으로 한 전립선암 연구 자료.

11. Shipp : 임파선암 치료제의 효과를 알기위하여, 58명의 환자중에 32 치료환자, 26명의 난치성 질환환자 표본으로 임파선암 연구자료.

위의 자료는 모두 두개의 범주로 나누어지는 이분형 자료로 이분분류문제에 적합한 자료들이다. 각 자료에 대한 자료수(n), 변수수(p), 및 자료의 특징을 [표 1]에 정리하였다.

각 데이터에 대해 본 논문에서 제시한 2개의 방법과 4가지 비교대상 방법을 적용하여 오분류율을 계산하였고 그 결과를 [표 2]에 정리하였다. 비교를 수월하게 하기 위해, 각 자료에 대해 가장 작은 오분류율을 굵은 글씨체를 사용하여 강조하였다. 비교대상 중 PLDA를 위해 사용한 mda 패키지는 변수의 개수가 2,000이상인 경우에 수행이 되지 않아 [표 2]에서 일부 데이터의 경우는 오분류율을 수록하지 않았다. 자료의 개수가 변수의 개수보다 큰 경우는 SVM 방법이 가장 낮은 오분류율을 보여주고 있다. 반면, 본 논문에서 제시한 PPCLR-L 및 PPCLR-S은 변수의 개수가 자료의 개수보다 큰 경우에 가장 낮은 오분류율을 보여준다. 이는 컴퓨터 모의실험의 결과와 매우 유사하다. 이러한 사실은 PPCLR 방법이 고차원 자료의 이분분류 문제에 좋은 대안이 될 수 있음을 시사한다고 볼 수 있다.

표 1: 11개의 실제자료에 대한 설명

자료명	자료수	변수수	자료타입/응용분야	자료출처
Ionosphere	351	34	signal/physics	UCL
Sonar	208	60	signal/meterial science	KEEL
Spambase	4,597	57	text/textmining	KEEL
Spectheart	297	44	image/medical science	KEEL
WDBC	569	30	image/medical science	KEEL
Chin	118	22,215	microarray/medical science	GitHub
Chowdary	104	22,283	microarray/medical science	GitHub
Gravier	168	2,905	microarray/medical science	GitHub
Gordon	181	12,533	microarray/medical science	GitHub
Singh	102	12,600	microarray/medical science	GitHub
Shipp	58	6,817	microarray/medical science	GitHub

6 결론 및 향후 연구과제

본 연구를 통해, 주성분을 사용한 로지스틱 회귀모형과 벌점함수를 사용한 변수선택을 이분분류문제에 응용하여 새로운 분류법을 제시하였다. 제시

표 2: 11개의 실제자료를 이용한 각 방법론에 따른 오분류율

Data	PPCLR-L	PPCLR-S	PCLR	PLR	PLDA	SVM
Ionosphere	14.53	15.10	14.27	14.81	14.25	**5.13**
Sonar	20.67	21.63	24.05	25.00	22.07	**18.24**
Spambase	7.22	7.33	33.89	9.25	11.46	**6.74**
Spectheart	**18.35**	18.73	20.90	**18.35**	24.00	20.61
WDBC	3.34	4.04	9.66	3.69	4.56	**2.64**
Chin	11.86	**11.02**	30.53	14.41	-	14.55
Chowdary	3.85	**2.88**	21.27	3.85	-	3.00
Gravier	**23.81**	26.19	31.51	25.00	-	26.8
Gordon	**0.00**	1.66	1.08	1.66	-	2.22
Singh	**6.86**	12.75	14.91	8.82	-	13.64
Shipp	10.39	**6.49**	8.75	**6.49**	-	21.96

한 방법론의 성능을 확인하기 위해 다양한 상황에서의 컴퓨터 모의실험을 수행하였으며 다양한 응용분야에서 나타나는 실제자료를 분석하여 기존의 다른 방법론들과 정량적 비교를 수행하였다.

다양한 분야에서 생성되는 자료들은 많은 특성을 지니며 자료간의 상

관성이 높은 고차원자료들이 빈번하게 나타나는 만큼, 이분분류문제에서 변수간의 다중공선성 문제가 보이는 경우는 빈번하다. 다중공선성 문제로부터 기인하는 분류예측의 나빠짐을 해결하기 위하여 주성분회귀분석의 개념 및 변수선택의 개념을 분류문제에 도입하는 것은 매우 자연스럽다. 본 연구에서는 이를 자연스럽게 결합하였으며 모의실험 및 실제자료를 이용하여 분류율이 향상될 수 있음을 실증적으로 보였다.

본 연구에서는 이분분류문제를 중점적으로 다루었다. 범주의 개수가 2개가 아닌 3개 이상인 경우의 분류문제에서도 로지스틱 회귀분류법을 적용할 수 있다. 이에 대한 확장 가능성은 본 저자의 최근 논문에 기술하였으며 (Kim and Lee, 2013), 이는 앞으로 추후 연구과제이다.

참고문헌

Alcalá-Fdez, J., Fernandez, A., Luengo, J., Derrac, J., Garćıa, S., Sánchez, L. and Herrera, F.. (2011). KEEL data-mining software tool: data set repository, integration of algorithms and experimental analysis framework, *Journal of multiple-valued logic and soft computing*, **17**, 255-287.

Bache, K., and Lichman, M. (2013) UCI machine learning repository [http://archive.ics.uci.edu/ml] Irvine, CA: University of California, School of Information and Computer Science.

Fan, J., and Li, R. (2001) Variable selection via non concave penalized

likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**, 1348–1360.

Friedman, J., Hastie, T., and Tibshirani, R. (2008) Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**, 1–22.

Hastie, T., Buja, A., and Tibshirani, R. (1995) Penalized discriminant analysis. *The annals of Statisitcs*, **23**, 73–102.

Hastie, T., Tibshirani, R., and Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

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

Hoerl, A. E. and Kennard, R. W. (1970b) Ridge regression: Iterative estimation of the baising parameter. *Technometrics*, **12**, 77–88.

Kim, K., and Lee, S. (2013) Logistic regression classification by principal component selection *Submitted*.

Kondylis, A., and Whittaker, J. (2008) Spectral preconditioning of Krylow spaces: combining pls and pc regression. *Computational Statistics & Data Analysis*, **52**, 2588–2603.

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2006) *Introduction to linear regression analysis, the 4th Edition*. Wiley.

Murphy, K. P. (2012) *Machine Learning: A Probabilistic Perspective*. The MIT press.

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

A Study on Improving Misclassification Rate of Principal Component Logistic Regression Classification

Abstract

We propose new binary classification methods by modifying logistic regression classification. Main idea of our proposals is to select the principal components using variable selection procedures, rather than to select the original variables. We describe the resulting classifiers and discuss their properties. The performance of our proposals are illustrated numerically and compared with other existing classification methods using synthetic and real datasets.