Journal of the Korean Data & Information Science Society
2011, **22**(5), 931–940

한국데이터정보과학회지

벌점 부분최소자승법을 이용한 분류방법[†]

김윤대[1] · 전치혁[2] · 이혜선[3]

[123]포항공과대학교 산업경영공학과

접수 2011년 8월 16일, 수정 2011년 9월 14일, 게재확정 2011년 9월 21일

요 약

분류분석은 학습표본으로부터 분류규칙을 도출한 후 새로운 표본에 적용하여 특정 범주로 분류하는 방법이다. 데이터의 복잡성에 따라 다양한 분류분석 방법이 개발되어 왔지만, 데이터 차원이 높고 변수간 상관성이 높은 경우 정확하게 분류하는 것은 쉽지 않다. 본 연구에서는 데이터차원이 상대적으로 높고 변수간 상관성이 높을 때 강건한 분류방법을 제안하고자 한다. 부분최소자승법은 연속형데이터에 사용되는 기법으로서 고차원이면서 독립변수간 상관성이 높을 때 예측력이 높은 통계기법으로 알려져 있는 다변량 분석기법이다. 벌점 부분최소자승법을 이용한 분류방법을 실제데이터와 시뮬레이션을 적용하여 성능을 비교하고자 한다.

주요용어: 로지스틱 회귀분석, 벌점함수, 부분최소자승법, 분류분석

1. 서론

분류분석은 새로운 표본을 범주 중의 하나에 분류하기 위해 학습표본 (training set)을 바탕으로 분류규칙을 생성한다. 또한 분류규칙의 성능을 비교하기 위해 학습표본과 시험표본 (test set)으로 나누어 실제 범주와 추정된 범주를 비교한다. 대표적인 분류 방법으로는 로지스틱 회귀분석, 판별분석, 트리 구조 등이 있다. 데이터마이닝 영역에서 다루어지는 데이터들은 대부분 차원이 높고 상호 상관성이 높은 복잡한 구조를 갖고 있어서 기존의 선형관계식이나 트리구조 분류에 의해서는 분류성과에 한계가 있다. 따라서 이러한 문제를 해결할 수 있는 새로운 분류 방법이 요구되고 있다.

이와 같은 문제를 해결하기 위한 다변량분석 방법중의 하나는 다중공선성의 문제를 감소시키고 종속변수와의 상관성을 고려한 잠재변수 도출에 의한 데이터 차원을 축소하여 예측력을 높인 부분최소자승법 (Partial Least Squares : PLS)이다 (Wold 등, 1984). 또한 PLS를 이용한 분류 규칙에 대한 연구도 활발하게 이루어지고 있는데, Barker와 Rayens (2003)는 PLS와 선형판별분석 (Linear Discriminant Analysis : LDA)을 결합한 분류 방법을 제안하였으며, Preda (2007) 등은 개선 모형을 제안하였다. PLS를 응용한 또 다른 분류 방법으로 로지스틱 회귀분석을 PLS에 접목한 방법이 제안되었으며 (Nguyen와 Rocke, 2002), 생물정보학 등의 분야에서도 널리 적용되고 있다 (Fort와 Lambert-Lacroix, 2005).

독립변수의 다중공선성을 해결하기 위한 또 다른 방법으로 능형회귀 (ridge regression), LASSO (Least Absolute Shrinkage and Selection Operator) 회귀방법 등의 축소방법 (shrinkage method)이

[†] 이 논문은 한국연구재단(NRF) 연구비에 의해 지원받았음(2010-0003628).
[1] (790-784) 포항시 남구 효자동 산 31번지, 포항공과대학교 산업경영공학과, 석사과정.
[2] (790-784) 포항시 남구 효자동 산 31번지, 포항공과대학교 산업경영공학과, 교수.
[3] 교신저자:(790-784) 포항시 남구 효자동 산 31번지, 포항공과대학교 산업경영공학과, 연구교수.
E-mail: hyelee@postech.ac.kr

사용된다. 일반적인 선형회귀분석에서 독립변수 간에 다중공선성이 존재할 경우, 추정계수들의 분산값이 커지고 그에 따라 예측된 종속변수값의 분산도 커지고 불안정할 수 있으며, 이러한 경우에 축소방법을 사용한다.

축소방법은 결과값이 어느 정도 편향되는 것을 감수하는 대신, 결과값의 분산을 줄이는 효과를 얻을 수 있다. 벌점의 크기를 조절함으로써 축소의 강도를 조절하고, 교차타당성 (cross-validation)을 통해 최적화할 수 있다. 주성분분석 (Principal Component Anaylsis : PCA)방법에서도 축소방법을 적용하여 희박한 적재값 (sparse loading)을 갖는 주성분을 구함으로써 변수선택, 분산설명력을 향상시킬 수 있음을 제안하였다 (Zou 등, 2006). 서포트벡터머신 알고리즘을 사용한 비모수적 접근방법으로 다중인자 차원축소를 적용하기도 한다 (이제영과 이종형, 2010).

본 연구에서는 위의 PLS와 벌점 함수를 사용하여 분류분석의 성능을 향상시킬 수 있는 방법을 제안하고자 한다. 실제 데이터와 시뮬레이션 데이터를 이용하여 기존의 로지스틱 회귀분석이나 판별분석 등의 분류 규칙과 성능을 비교하고, 제안된 방법의 개선점을 제시한다.

2. 관련연구

본 절에서는 본 연구와 관련있는 로지스틱 회귀분석, PCA를 이용한 분류방법과 PLS를 이용한 분류방법을 설명하고자 한다.

2.1. 로지스틱 회귀분석

일반적으로 선형회귀모형에서 종속변수는 연속적인 값을 가지는 양적 변수로 가정한다. 그러나 독립변수에 따른 질병의 유무를 판단하고자 하거나 품질의 등급을 매기고자 할 때, 종속변수가 범주형 변수로서 그 값이 0, 1 등의 값을 갖는 경우가 있다. 이 경우 선형회귀분석을 적용한다면 독립변수와 종속변수의 관계를 제대로 설명하기 힘들다.이러한 경우 회귀모형은 종속변수를 직접적으로 설명하는 것이 아니라, 그 종속변수의 값이 1일 확률 (P로 표현)을 설명하는 것을 고려한다. 이 확률이 독립변수에 대하여 선형적일 것으로 가정하면 다음과 같은 식이 성립한다.

$$P = \beta^T X \tag{2.1}$$

그러나 이 경우 결과값이 확률값이므로 0에서 1 사이의 값을 가져야 하는데, X값의 증감에 따라서 0 이하이거나 1 이상의 경우가 발생할 수 있다. 따라서 이러한 문제를 해결하기 위해 로짓 변환 (logit transformation)이 개발되었다 (McCullagh와 Nelder, 1989). 즉, 다음과 같은 로지스틱 함수를 사용한 후, 로짓 변환을 통해 선형회귀식으로 변환시킨다.

$$P = \frac{\exp(\beta^T X)}{1 + \exp(\beta^T X)} \tag{2.2}$$

$$logit(P) = \ln[\frac{P}{1-P}] = \beta^T X \tag{2.3}$$

이분 로지스틱 회귀모형에서의 분류규칙은 최소오분류비율에 따라 임계치를 정하고 그 확률이 임계치 이상일 경우 1로 분류하고 그렇지 않을 경우 0으로 분류한다.

이 외에도 곰페르츠, 프로빗, 노밋 (Normit) 변환 등이 로짓 변환과 함께 많이 사용되고 있으며, 경우에 따라 알맞은 방법을 사용할 수 있다 (Berger, 1981).

2.2. 요인분석을 이용한 판별분석 (PCA-DA)

Mallet 등 (1996)은 고차원데이터 특히 스펙트럼 데이터의 차원축소 방법 및 특성선택 (feature selection)을 이용한 판별분석 성과비교를 하였으며, 주성분에 의한 판별분석 및 벌점판별 분석 등을 시도하였다. Kemsley (1996)는 데이터 축소 혹은 변환을 위한 방법으로 PLS가 PCA대신 적용하여 분류성과가 개선됨을 보여주었다. 데이터 축소를 이용한 분류분석은 생명과학 및 정보분석학의 많은 분야에서 다양하게 개발 연구되고 있다.

주성분 판별분석은 다음과 같은 과정을 통해 이루어진다.

1단계. 주어진 데이터의 독립변수를 주성분 분석하여 설정한 주성분의 개수만큼 새로운 독립변수를 생성한다.
2단계. 생성된 k개의 주성분으로 판별 분석을 수행한다.
3단계. 수행된 판별 분석의 결과를 통해 데이터를 분류한다.

이 방법은 차원 감소의 측면에서 뛰어난 효과를 얻을 수 있고, 독립변수들 간에 다중공선성이 존재할 경우에도 유용하게 사용될 수 있다. 그러나 PCA의 특성상 주성분이 독립변수의 분산 설명력은 높은 반면 종속변수의 설명력과는 무관할 수 있어서 예측력이 약할수 있는 단점이 있다. 또한 판별분석을 사용할 경우 독립변수들이 정규분포를 따르는 것을 가정하기 때문에 PCA를 통해 얻어진 새로운 독립변수를 적용할 때 문제가 발생할 수 있다.

2.3. PLS를 이용한 분류분석

Nguyen과 Rocke (2002)는 마이크로어레이 데이터를 이용한 암의 유무 진단시 PLS를 이용한 분류방법을 제시하였다. PCA 방식으로 데이터 차원 축소를 수행하고 축소된 변환값으로 로지스틱 판별분석, 이차 판별분석을 적용하였고, PLS에 의한 판별분석의 성과가 향상됨을 보여주었다. PLS 방법은 p개 독립변수들의 분산과 더불어 종속변수와의 공분산을 최대로 설명하는 변환을 가능하게 함으로써 PCA보다 나은 예측성과를 기대할 수 있다. PLS를 이용한 분류 방법은 다음과 같다.

1단계. 주어진 데이터의 범주형 종속변수를 로지스틱 회귀분석을 통해 확률값으로 변환시킨다.
2단계. PLS 알고리즘에 적용하여 최적 잠재변수를 선택하고 그에 따라 축소된 차원의 선형 변환값을 얻는다.
3단계. 선형변환값을 이용하여 판별분석에 적용한다.

PLS는 주성분 선형모형이나 다중회귀모형보다 강건한 방법으로서 학습표본에 따라 모수 추정에 큰 차이가 없다는 장점을 갖고 있다 (Geldadi와 Kowalski, 1986). 또한 PLS 회귀분석이 여러개의 종속변수에서도 사용할 수 있기 때문에 종속변수의 범주가 여러개일때 한 번에 계산이 가능하다.

경우에 따라 PLS를 통한 차원 감소의 폭을 더 크게 해야 할 경우도 있으며, 분류 모형은 잠재변수의 수가 적으면 적을수록 계산 속도가 빨라지기 때문에 차원 감소를 효율적으로 할 수 있는 방법이 필요하다. PLS 알고리즘에서 가중 벡터를 결정하는 데 있어서 좀더 효율적인 값을 찾을 경우 PLS 분류 방법은 개선될 수 있다.

2.4. 벌점 PLS를 이용한 분류분석

일반적으로 PLS 기법은 차원 감소의 효과뿐만 아니라, 다중공선성하에서 독립변수의 계수를 구하거나 예측값을 구하는 데 있어서 뛰어난 성능을 보인다. 데이터의 다중공선성의 정도를 알아보기 위해 분

산팽창계수 (Variance Inflation Factor : VIF)를 주로 사용한다. j번째 회귀계수에 대한 분산팽창계수는 다음과 같이 정의된다 (전치혁 등, 2004).

$$VIF_j = \frac{1}{1 - R_j^2} \tag{2.4}$$

여기서 R_j^2는 X_j를 종속변수로 하고 나머지 변수를 독립변수로 하는 회귀모형에서의 결정계수이며, 분산팽창계수가 10을 넘으면 다중공선성이 있다고 볼 수 있다.

이러한 PLS의 장점에 벌점 함수의 장점인 편향성과 분산의 상반관계를 조절하여 안정된 회귀 계수를 얻을 수 있다. 특히 잡음이 포함된 데이터를 분석하는 데 있어서 이 방법이 뛰어난 성능을 보이는 것으로 알려져 있다 (Kramer 등, 2008). PLS 회귀분석의 알고리즘인 NIPALS (Nonlinear iterative partial least squares)에서 사용되는 가중 벡터, w를 구하는 식은 다음과 같다 (Wold 등, 1984; Wold 등, 2001).

$$w \leftarrow \arg max_w \frac{w^T X^T y y^T X w}{w^T w} \tag{2.5}$$

식 (2.5)의 해는 독립변수와 종속변수간 공분산을 최대화 설명하면서 상호적으로 직교인 가중 벡터를 얻게 한다. 벌점 PLS 에서는 이 식에 벌점함수를 삽입함으로써 가중 벡터의 값을 변화시킨다. 이것을 식으로 표현하면 다음과 같다.

$$w = \arg min_w [(y - Xw)^T (y - Xw) + \lambda P(w)] \tag{2.6}$$

여기서 $P(w)$는 w로 표현되는 벌점 함수이다. $P(w)$항이 추가됨으로써 좀더 안정된 회귀 계수를 얻을 수 있으며, $P(w)$항에 따라 사용자가 원하는 벌점 함수를 적용할 수 있다. 또한, λ는 조율 파라미터로서 가중치를 축소시키는 역할을 한다. 이는 일반적으로 교차타당성을 통하여 최소의 평균제곱합오차 (Mean square error)를 가져오는 값으로 결정한다. 위 식을 기존의 NIPALS 알고리즘의 $w_i = X_i^T y$대신 넣음으로써 벌점 PLS 알고리즘을 수행할 수 있다 (Kramer 등, 2008).

3. 제안방법

앞에서 언급한 것과 같이 PLS 회귀분석은 화학, 생명공학, 마케팅, 공정분석 등 여러 다양한 분야에서 예측모형으로 이용되고 있다. 본 연구에서는 분류분석에 적용하기 위하여 로지스틱 회귀분석을 이용하여 확률값인 종속변수로 변환하여 PLS를 적용한다.

또한 일반적인 PLS 회귀분석에서의 가중 벡터를 구하는 단계에서 벌점 함수를 사용함으로써, 독립변수 간에 다중공선성이 존재할 경우의 초기 가중 벡터를 구하는 부분에 더 효율적인 계수를 구하고자 한다. 이를 통하여 개선된 벌점 PLS를 이용한 분류 방법을 제안하고자 한다.

3.1. 능형 벌점 PLS를 이용한 분류방법

능형 벌점 함수를 이용한 PLS 분류 방법의 알고리즘은 다음과 같다.

1단계. 주어진 X와 y의 데이터를 이용하여 로지스틱 회귀분석을 수행하고, 종속변수에 대한 로짓값 $logit(P)$를 구하여 새로운 y로 놓는다.

2단계. 다음과 같은 방법으로 능형 벌점 가중 벡터를 구한다.

$$u = y \tag{3.1}$$

$$w = (X^T X + \lambda I)^{-1} X^T y \tag{3.2}$$

$$w \leftarrow \frac{w}{\|w\|} \tag{3.3}$$

$$t = Xw \tag{3.4}$$

3단계. 일반적인 PLS 회귀분석의 알고리즘에 맞추어 2단계에서 구한 잠재변수 t를 대입한다.
4단계. PLS 알고리즘에 따라 잠재변수의 수에 맞추어 2단계부터 반복 수행한다.
5단계. 벌점 PLS를 통하여 구한 확률값을 로지스틱 분류규칙에 맞추어 분류한다.

$$logit(P) > 0\text{이면 } i\text{번째 객체를 '1'로 분류}$$
$$logit(P) \leq 0\text{이면 } i\text{번째 객체를 '0'으로 분류}$$

위의 식 (3.2)에서 λ는 능형회귀모형의 조율 파라미터로서 2.4절에서 언급한 바와 같이 학습표본의 교차타당성 (cross-validation)으로 그 값을 결정한다. 식 (3.3)에서 $||w||$는 w의 노름 (norm)을 나타내며 결과로 얻어지는 새로운 w의 노름은 1이된다. 5단계에서 $logit(P) > 0$인 경우는 확률값이 0.5 이상인 것을 의미하며, 분류임계치는 오분류를 최소화하는 확률값을 선택할 수 있다.

3.2. LASSO 벌점 PLS를 이용한 분류방법

LASSO의 경우에는 능형보다 많은 차원 감소가 가능한데, 이는 LASSO 계수를 구하는 과정에서 잠재변수를 설명해 주는 변수 중 설명력이 약한 변수의 계수를 0으로 만들어 주기 때문이다. LASSO의 경우는 차원 감소나 추정치에 있어서는 능형보다 나은 성능을 보이지만, 2차계획법 (quadratic programming)을 사용해야 하기 때문에 추정치를 구하는데 시간이 많이 소요되는 단점이 있다. LASSO 벌점 PLS는 3.1 절의 2단계에서 가중 벡터 구하는 방법이 다음과 같으며 나머지 과정은 동일하다.

$$w = argmax_w[(y - Xw)^T(y - Xw) + \lambda||w||] \tag{3.5}$$

LASSO 방법에서도 λ를 변환시킴으로써 가중 벡터의 값을 조절할 수 있으며, 최적의 추정값을 가지는 경우의 λ를 찾는 것이 중요하다.

4. 실험결과

각 방법의 예측 정확도를 알아보기 위하여 주어진 데이터에 대하여 로지스틱 회귀분석, PLS 판별 분석, 그리고 능형과 LASSO 방법을 이용한 벌점 PLS 방법의 예측력을 비교해 보았다. 분류분석의 성능 평가는 오분류율을 척도로 사용하였다.

λ값을 선정시에는 주어진 데이터의 70%를 학습표본으로 사용하고, 나머지 30%를 시험표본으로 사용하여 15번 반복하여 오분류율을 최소화하는 잠재변수의 수와 λ값을 선정하고, 제안방법의 비교분석 시에는 100번의 재표본추출하여 (resampling) 시험표본의 평균 오분류율을 비교하였다.

4.1. 실제데이터 실험

UC Irvine (http://archive.ics.uci.edu/ml/)에서 제공하는 Hill-Valley 데이터를 사용하였다. Hill-Valley 데이터는 주어진 독립변수들로 언덕 (hill)인지 골짜기 (valley)인지를 분류하는 데이터로서 독립변수는 총 100개, 관측수는 606개로 이루어져 있다. 종속변수는 이분형 데이터로서, 값이 0일 경우에는 '골짜기'이며 1일 경우에는 '언덕'을 나타내며 분산팽창계수가 모두 10이상으로 다중공선성을 갖는 데이터이다. UC Irvine (http://archive.ics.uci.edu/ml/)에서 제공하는 데이터 중 다공선성이 없는 Wine 데이터에 대해서도 적용해보았는데, 분산팽창계수가 낮은 다공선성의 문제가 없는 데이터의 경우 벌점PLS의 성능이 PLS-DA, PCA-DA보다는 좋았지만 일반적인 분류분석 방법인 로지스틱방법으로도 충분히 분류가 되는 것을 확인하였다.

제안된 방법에서는 λ의 값을 0.001, 0.005, 0.01, 0.03, 0.05, 0.07, 0.1, 0.3, 0.5, 0.7을 사용하였다. 잠재변수의 개수는 1로 설정하여 실험하였다. λ가 변함에 따라 오분류율이 어떻게 변하는지 측정하였다.

그림 4.1에서 능형 PLS는 λ값이 0.3 일 때, LASSO PLS는 λ값이 0.5 일 때 오분류율이 가장 낮은 값을 나타내었다. λ값이 0.03보다 작을 경우에는 오분류율의 변동이 심한 것은 원 데이터의 다중공선성 문제로 계수의 분산이 커져서 이와 같은 결과가 나오는 것으로 볼 수 있다. 이러한 경우 벌점을 줌으로써 오히려 오분류율이 감소됨을 알 수 있다.

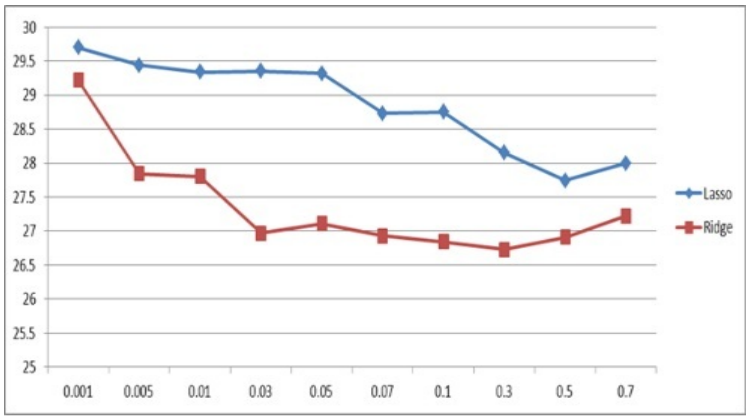

그림 4.1 λ에 따른 벌점PLS방법의 오분류율 변화 (Hill-Valley 데이터)

잠재변수 수와 λ값에 따른 오분류율 변화를 수행한 결과 PLS에서는 종속변수의 분산중 한 개의 잠재변수가 99%를 설명하므로 벌점 PLS에서 1개의 잠재변수를 선택하였다. PCA-DA에서는 3개의 잠재변수를 사용하였다. 벌점함수를 이용한 PLS 제안방법의 결과는 그림 4.2에서 보는바와 같이 능형 PLS, LASSO PLS가 로지스틱 회귀분석, PLS, PCA-DDA과 비교하여 오분류율이 낮은 좋은 성과를 보였다.

제안 방법인 능형 벌점 PLS 방법이 PLS를 이용한 분류 방법이나 PCA-DA에 비하여 낮은 오분류율을 나타내며, 100번의 리샘플링을 통한 시험표본의 오분류율 평균을 구해 보면 표 4.1과 같다.

표 4.1 각 분류방법에 따른 오분류율 평균값 (Hill-Valley 데이터)

	로지스틱	PLS 분류	PCA-DA	능형 PLS	LASSO PLS
오분류율	31.33	49.41	49.31	26.73	27.74
(표준편차)	(4.00)	(2.96)	(4.48)	(4.35)	(4.22)

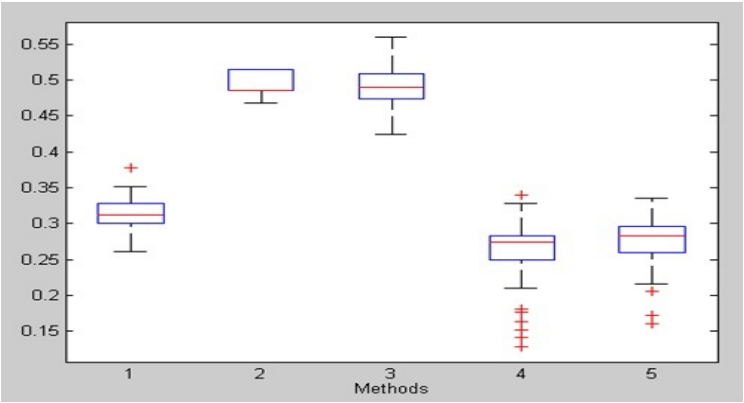

그림 4.2 분류방법에 따른 오분류율 비교 (Hill-Valley 데이터)
(1) 로지스틱 (2) PLS 분류 (3) PCA-DA (4) 능형 PLS (5) LASSO PLS

세 가지 방법의 오분류율 차이를 알아보기 위하여 t-test해 본 결과, 제안된 벌점 PLS와 로지스틱 방법과 PLS 방법, PCA-DA간에 유의한 차이를 보였다. 제안한 두 방법 중에서 능형 PLS가 LASSO PLS보다 낮은 오분류율을 보였으나 유의한 차이는 없었다.

4.2. 시뮬레이션 실험

제안된 방법이 일부 독립변수 간 다중공선성을 가질 때에도 강건한 결과를 내는지 알아보기 위하여 인공 데이터를 사용해 보았다. 각 독립변수가 평균이 0, 분산이 1인 정규분포를 따르도록 생성한 후, $y = \beta^T X + \epsilon$을 따르도록 종속변수를 생성하였다. 그 후 종속변수의 크기에 따라 범주를 구분하였다. 각각의 독립변수 사이의 공분산을 조절함으로써 전체 중 절반의 독립변수의 VIF 값이 10을 넘도록 하여 다중공선성이 존재하도록 하였다. 데이터는 총 10개의 독립변수로 구성되어 있으며 관측치의 개수는 300개로 이루어져 있다. 범주는 이분형 데이터이며, 종속변수의 관측치 비율은 1:1로 나타난다. 실험 방법은 4.1절과 같은 방법으로 진행하였다.

λ값을 결정하기 위하여 그 값들을 변화시켜 가며 오분류율을 측정해 보았다. 잠재변수는 1로 설정하였다. λ는 0.001, 0.005, 0.01, 0.03, 0.05, 0.07, 0.1, 0.3, 0.5, 0.7로 설정하였다.

그림 4.3에서 제안된 벌점 PLS에서 모두 λ가 0.03 일 때에서 오분류율의 감소가 시작되어 0.05, 0.07 부근에서 가장 낮은 오분류율을 보였다. 그 후 λ가 0.1 이상에서부터는 다시 오분류율이 증가하는 것을 보였다. LASSO의 경우에는 0.05에서, 능형의 경우는 0.07에서 가장 오분류율이 낮았으므로 이 실험에서는 위와 같은 값으로 λ를 설정하였다. 실험 결과 오분류율의 분포는 그림 4.4와 같다.

벌점함수를 이용한 PLS 제안방법의 결과는 그림 4.4에서 보는바와 같이 능형 PLS, LASSO PLS가 로지스틱 회귀분석, PLS, PCA-DA과 비교하여 오분류율이 낮은 좋은 성과를 보였다.그림에서 제안 방법인 능형 벌점 PLS 방법이 PLS 분류방법이나 PCA-DA에 비하여 낮은 오분류율을 나타내었다. 이 값들의 평균을 구해 보면 표 4.2와 같이 나타난다.

표 4.2 각 분류방법에 따른 오분류율 평균값 (시뮬레이션 데이터)

	로지스틱	PLS 분류	PCA-DA	능형 PLS	LASSO PLS
오분류율	18.57	27.51	31.03	17.86	17.83
(표준편차)	(3.87)	(4.73)	(4.05)	(3.95)	(3.87)

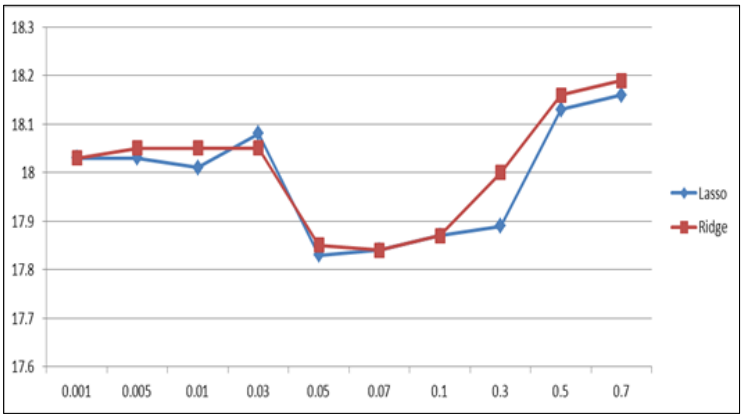

그림 4.3 λ에 따른 벌점 PLS의 오분류율 변화 (시뮬레이션 데이터)

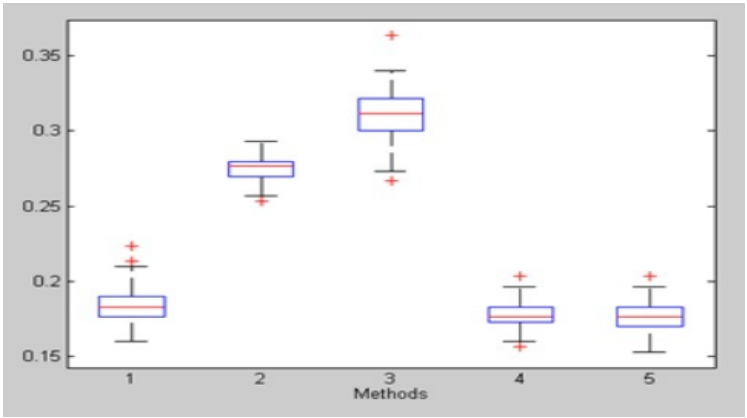

그림 4.4 분류방법에 따른 오분류율 비교 (시뮬레이션 데이터)
(1) 로지스틱 (2) PLS 분류 (3) PCA-DA (4) 능형 PLS (5) LASSO PLS

능형 PLS와 LASSO PLS의 오분류율 차이는 유의할 정도로 크게 나지는 않았으나, 대체로 LASSO PLS가 조금 더 낮은 오분류율을 가지는 것으로 나타났다. 시뮬레이션 데이터에 대해서도 제안한 벌점 PLS방법이 기존의 로지스틱, PLS, PCA-DA에 비해 나은 성능을 보였다.

제안된 벌점 PLS방법은 다중공선성이 있을 경우 분류분석에서 오분류율을 줄일 수 있는 방법이 될 수 있다.

5. 결론

본 연구에서는 벌점 PLS 기법을 이용한 개선된 분류분석방법을 제안해 보았다. 벌점함수의 도입은 다공선성이 높은 데이터의 경우 편향성을 줌으로써 분산의 감소를 통해 기존의 방법보다 더 나은 예측력을 갖게 할 수 있는 방법이다.

실제 데이터와 시뮬레이션 실험 결과 제안된 벌점 PLS 방법이 기존의 PLS, PCA-DA와 같은 대표적인 차원 감소 분류 방법보다 더 나은 예측력을 가지는 것으로 나타났다. 데이터의 종류나 다중공선성의 차이에 따라 예측력이 조금씩 다르게 나타날 수 있으나, 제안 방법이 우수함을 알 수 있다. 또한 잠재변수의 수을 변화시킴으로써 예측 정확도와 측정 비용간의 최적 지점을 찾아, 사용자가 원하는 가장 최적화된 예측 분석을 할 수 있을 것으로 기대된다.

PLS 회귀분석은 원래 종속변수 및 독립변수들이 연속형인 데이터에 대하여 적용 가능한 분석 방법이다. 그러나 분류분석을 위한 많은 데이터에는 범주형 변수가 포함되어 있다. 본 연구에서는 종속변수를 제외한 모든 독립변수가 연속형인 경우 분류분석에 적용할 수 있는 알고리즘을 담고 있다. 따라서 독립변수 중에 범주형을 포함하는 경우 벌점 PLS를 적용할 수 있는 방법에 대한 연구가 필요할 것이다.

또한 이번 연구에서는 PLS와 로지스틱, 축소방법을 결합한 방법에 대하여 다루었으나, 최근접이웃방법 (nearest neighborhood)나 SVM (Support Vector Machine)과 같은, 다른 여러 가지 분류 방법의 장점을 살려 결합해 보는 것 또한 새로운 연구과제가 될 것으로 판단된다.

참고문헌

이제영, 이종형 (2010). 서포트 벡터 머신 알고리즘을 활용한 연속형 데이터의 다중인자 차원축소방법 적용. <한국데이터정보과학회지>, **21**, 1271-1280.

전치혁, 정민근, 이혜선 (2004). <공학응용통계>, 홍릉출판사, 서울.

Barker, M. and Rayens, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics*, **17**, 166-173.

Berger, R. (1981). Comparison of the Gompertz and logistic equations to describe plant disease progress. *Phytopathology*, **71**, 716-719.

Fort, G. and Lambert-Lacroix, S. (2005). Classification using partial least squares with penalized logistic regression. *Bioinformatics*, **21**, 1104-1111.

Geldadi, P. and Kowalski, B. (1986). Partial least-squares regression: A tutorial. *Analytica Chemica Acta,* **185**, 1-17.

Kemsley, E. K. (1996). Discriminant analysis of high-dimensional data: A comparison of principal components analysis and partial least squares data reduction methods. *Chemometrics and Intelligent Laboratory Systems,* **33**, 47-61.

Kramer, N., Boulesteix, A. and Tutz, G. (2008). Penalized Partial Least Squares with applications to B-spline transformations and functional data. *Chemometrics and Intelligent Laboratory Systems*, **94**, 60-69.

Mallet, Y., Coomans, D. and de Vel, O. (1996). Recent developments in discriminant analysis on high dimensional spectral data. *Chemometrics and Intelligent Laboratory Systems,* **35**, 157-173.

McCullagh, P. and Nelder, J. (1989). *Generalized linear models*, second edition, Chapman and Hall/CRC, Boca Raton.

Nguyen, D. and Rocke, D. (2002). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, **18**, 39-50.

Preda, C., Saporta, G. and Leveder. C. (2007). PLS classification of functional data. *Computational Statistics,* **22**, 223-235.

Wold, S., Sjöström, M. and Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory System*, **58**, 109-130.

Wold, S., Rube, H., Wold, H. and Dunn, W.J. (1984). The collinearity problem in linear regression. The Partial Least Squares (PLS) approach to generalized inverses. *SIAM Journal of Scientific and Statistical Computations,* **5**, 735-743.

Zou, H., Hastie, T. and Tibshirani, R. (2006). Sparsep principal component analysis. *Journal of Computational and Graphical Statistics*, **15**, 265-286.

Journal of the Korean Data & Information Science Society
2011, **22**(5), 931–940

한국데이터정보과학회지

A new classification method using penalized partial least squares[†]

Yundae Kim[1] · Chi-Hyuck Jun[2] · Hyeseon Lee[3]

[123]Department of Industrial and Management Engineering, POSTECH

Received 16 August 2011, revised 14 September 2011, accepted 21 September 2011

Abstract

Classification is to generate a rule of classifying objects into several categories based on the learning sample. Good classification model should classify new objects with low misclassification error. Many types of classification methods have been developed including logistic regression, discriminant analysis and tree. This paper presents a new classification method using penalized partial least squares. Penalized partial least squares can make the model more robust and remedy multicollinearity problem. This paper compares the proposed method with logistic regression and PCA based discriminant analysis by some real and artificial data. It is concluded that the new method has better power as compared with other methods.

Keywords: Classification, logistic regression, partial least squares, penalized function.

[†] This research was supported with Basic Science Research Program through the National Research Foundation of Korea (NRF) from the Ministry of Education, Science and Technology (2010-0003628).

[1] Graduate student, Department of Industrial and Management Engineering, POSTECH, Pohang 790-784, Korea.

[2] Professor, Department of Industrial and Management Engineering, POSTECH, Pohang 790-784, Korea.

[3] Corresponding author: Research professor, Department of Industrial and Management Engineering, POSTECH, Pohang 790-784, Korea. E-mail: hyelee@postech.ac.kr