응용통계연구 제 19권 1호, 2006년, pp. 183–193

L_1-거리와 L_1-데이터뎁스를 이용한 분류방법의 비교연구*

백수진[1] 김진경[2] 황진수[3]

요 약

L_1-데이터뎁스를 이용한 분류방법($L1DDclass$)과 관측치들 사이의 L_1-거리를 이용한 분류방법($L1DISTclass$)의 특징을 살펴보고, 이 두 방법을 결합한 새로운 분류방법($DnDclass$; Distance and Data-depth based classification)의 효용성을 소개하고자 한다. 모의실험을 통해 세가지 분류방법의 결과를 비교하고 제안된 분류방법이 다양한 경우에 더 효과적일 수 있다는 사실을 확인한다.

주요용어: L_1-거리, L_1-데이터뎁스, 분류방법

1. 서론

최근 여러 분야에서 자료의 효과적인 분석을 위해서 다양한 분류(군집)분석방법들이 제안, 이용되고 있다. 그간 많이 이용되어온 전통적 분류(군집)분석방법으로는 K-평균(K-means), 계층적 군집법(Hierarchical clustering), 차별분석(Discriminant analysis), CART (Classification And Regression Tree) 등이 있다. 그런데 이 방법들은 로버스트하지 못한 경향이 있는데, 이를 보완하기 위해서 여러 방향에서 노력이 이루어지고 있다. 그간 제안된 로버스트한 분류(군집)분석 방법으로 PAM(Partitioning Around Medoids), SVM(Support Vector Machine), 신경망(Neural Nets) 등이 있다.

그 중 PAM은 잡음(noise)의 수준이 높거나 고차원 자료에서 문제점을 드러내는 약점이 있다. 이를 보완하기 위해 관측치들 간의 거리를 이용하는 분류(군집)분석방법인 $SILclass$가 제안되었고, Jornsten, Vardi와 Zhang(2002)은 군집들간에 분산의 차이가 많은 경우 오분류하는 경향이 있는 $SILclass$의 단점을 극복하기 위해서 로버스트한 자료분석 도구인 데이터 뎁스(Data Depth, Liu *et. al.*(1999) 참고)의 개념을 이용한 새로운 분류(군집)분석방법($DDclass$)을 제안하였다.

$SILclass$에서 군집의 수를 결정하는데 이용하는 실루엣 너비는 각 관측치들 간의 거리에 의해서 결정된다. 따라서 군집들이 비슷한 퍼짐의 정도를 갖고 있는 경우에는 효과적인데 반해, 군집들간에 퍼짐의 정도가 다양한 경우에는 퍼짐의 정도가 큰 군집에 포함된

* 이 논문은 인하대학교 교내연구비로 지원되었음 INHA-31630-01

1) (122-701) 서울특별시 은평구 녹번동 5, 국립보건연구원 질병관리본부 호흡기세균팀, 연구원
E-mail: sjbaek@stat.inha.ac.kr
2) (교신저자) (402-751) 인천광역시 남구 용현동 253, 인하대학교 자연과학대학 수학통계학부, 교수
E-mail: jkkim@stat.inha.ac.kr
3) (402-751) 인천광역시 남구 용현동 253, 인하대학교 자연과학대학 수학통계학부, 교수
E-mail: jshwang@stat.inha.ac.kr

관측지는 상대적으로 밀집되어 모여있는 근처의 군집에 더 밀착된 것처럼 간주되는 오분류의 경향이 나타난다. 한편 $DDclass$에서 군집의 수를 결정하는데 이용하는 상대적 데이터 뎁스(Relative Data Depth :ReD)는 Vardi와 Zhang(2000)에 의해 제안된 L_1-데이터 뎁스($L_1 - DD$)의 개념을 각 군집에 대한 상대적인 개념으로 확장한 것이다. $L_1 - DD$는 관측치들 사이의 방향 벡터에 의해 결정되는데, 이 방향벡터는 거리와 무관하기 때문에 군집간의 퍼짐의 정도에 영향을 잘 받지 않는 특징이 있다. 그러나 거리에 대한 정보를 이용하지 않으므로 분류시에 효율이 떨어지는 경우가 발생한다.

최근 Jornsten(2004)은 $SILclass$와 $DDclass$의 이러한 장단점을 고려하여 군집수를 결정하는데 $SILclass$의 실루엣너비와 $DDclass$의 상대적 데이터 뎁스를 모두 고려하는 새로운 군집분석방법($DDclust$)을 제안하고, 시뮬레이션 연구를 통하여 $DDclust$가 기존의 두 방법보다 더 나은 결과가 얻어질 수 있음을 보여주었다. 또한 Jornsten(2004)은 분류단계에서 L_1-거리와 $L_1 - DD$를 결합하는 기준을 언급하였으나, 그 두 수치가 취하는 값의 범위가 너무 다양하기 때문에 생길 수 있는 어려움을 설명하고 이 분류기법을 비교연구의 대상으로 삼지 않았다. 이 논문에서는 이러한 문제점을 극복할 수 있도록 자료에 대하여 간단하게 적용시킬 수 있는 표준화방법을 제시하고, L_1-거리와 $L_1 - DD$를 결합한 분류방법(Data-depth and Distance classification; $DnDclass$)을 기존의 분류방법들과 비교하였다. 이 논문에서는 군집방법까지를 포함하는 $SILclass$나 $DDclass$와 구별하고자 L_1-거리를 이용한 분류방법을 $L1DISTclass$로, $L_1 - DD$를 이용한 분류방법을 $L1DDclass$ 라 표현하기로 한다.

여러 종류의 데이터 뎁스들 중에 $L_1 - DD$를 이용한 것은 우선 다른 데이터 뎁스들은 자료의 차원이 높아지면서 계산에 시간이 많이 걸리는데 반해 이 $L_1 - DD$는 고차원의 자료에 대해서도 계산이 간편한 장점이 있고, 군집의 바깥에 존재하더라도 그 상대적인 위치에 따라서 데이터 뎁스값이 0이 아닌 다양한 값을 갖기 때문에 상대적으로 좀 더 가까운 군집을 결정할 수 있기 때문이다. 또한 다양한 데이터 뎁스를 이용한 분류기법들을 비교한 Ghosh와 Chaudhuri(2005)를 참고하면 여러 종류의 데이터 뎁스 중에 $L_1 - DD$가 그 성능에 있어서도 대체적으로 만족스러운 결과를 보여주는 것으로 나타났다. 한편 $L_1 - DD$에 의한 분류기는 자료의 shift, scaling, rotation에 대해서는 같은 결과를 주지만 affine invariant 하지 않은 단점을 갖고 있다. Ghosh 와 Chaudhuri(2005)에는 이를 개선하기 위한 표준화시키는 방법 등이 제안되어 있는데, 표준화 여부가 성능에는 별 영향을 주지 않은 것을 볼 수 있다. 이 논문에서는 다른 분류기법과의 성능의 비교를 연구하기 위한 것이므로 표준화하는 번거로움을 따르지 않았음을 밝혀둔다.

제 2절에서는 L_1-거리와 $L_1 - DD$를 결합하는 분류방법, $DnDclass$를 정의하고, 그 방법을 적용하기 전에 필요한 표준화 방법을 소개한다. 제3절에서는 다양한 시뮬레이션 자료에 대하여 L_1-거리와 $L_1 - DD$를 모두 고려하는 이 새로운 분류방법 $DnDclass$이 얼마나 효율적으로 분류하는가를 비교분석한다.

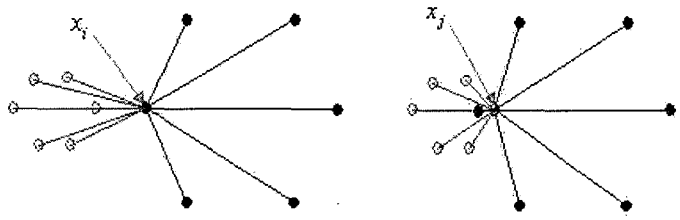
2. L_1-거리와 $L_1 - DD$를 이용한 분류방법들

R^p에서 N개로 구성된 $x_1, \ldots, x_N$ 의 데이터와 $\eta_1, \ldots, \eta_N$ 의 양수가 있다고 하자. 여기서 η_i는 가중치 혹은 x_i들의 다중도(multiplicity)로서 데이터가 중복되지 않으면, η_i는 1이다. 그리고 $I(k)$는 k군집의 x_i들의 라벨이다. 여기서 클래스 라벨을 가진 데이터를 훈련자료(training set; TR)라 하고, 라벨이 알려지지 않은 데이터를 테스트자료(test set; TE)라 한다.

2.1. L_1-거리를 이용한 분류방법 : $L1DISTclass$

각 군집 k에 대하여 $\bar{d}(x_i|k)$는 x_i로부터 k 군집내 다른 모든 관측치들까지 거리의 평균이다. 분류방법으로서의 $L1DISTclass$는 TE에 있는 각 관측점의 x_i에 대하여 TR에 의해 정해진 군집들 중 가장 작은 $\bar{d}(x_i|\cdot)$를 주는 군집에 분류하는 것이다.

$\bar{d}$는 각 집단의 퍼짐 정도에 영향을 받는다. 그러므로 자신이 속한 집단의 퍼짐 정도가 크면 $\bar{d}$의 값은 커져서 오분류될 수 있다. 그림 2.1에서는 퍼짐의 정도가 다를 경우에 $L1DISTclass$가 어떻게 분류하는가를 보기 위해서 각각 6개로 이루어진 두 개의 군집을 고려하였다. (a)는 퍼짐이 큰 군집에 속한 관측치 x_i가 $L1DISTclass$에 의해서 근처에 위치한 퍼짐이 작은 군집에 오분류되는 예를 보여준다. 한편 그림 2.1의 (b)는 두 군집이 서로 겹쳐있을 때 두 군집에 겹쳐 있는 곳에 위치하는 퍼짐이 작은 군집에 속한 관측치 x_j가 $L1DISTclass$에 의해서 제대로 자신의 군집에 분류되는 예를 보여준다.



(a) (b)

그림 2.1: L1DISTclass에 의한 오분류와 올바른 분류의 예

2.2. $L_1 - DD$를 이용한 분류방법 : $L1DDclass$

$L1DDclass$는 군집의 퍼짐 정도에 별 영향을 받지 않는 데이터 뎁스의 개념을 이용한 분류방법이다. 여기서 이용한 $L_1 - DD$는 고차원에서도 계산이 간편할 뿐만 아니라, 데이터 바깥에서도 중심으로부터 떨어져 있는 상대적인 위치에 따라 0이 아닌 다양한 값들을 갖는 장점이 있다. 이는 여러 군집들간에 $L_1 - DD$ 값을 비교할 때 의미있게 사용될 수 있다.

$L_1 - DD$는 L_1-중앙값(multivariate L_1-median)으로부터 계산되는데, 주어진 군집 k의 L_1-중앙값 $y_0(k)$는 다음과 같이 정의한다.

$$y_0(k) = \arg\min_y \sum_{j \in I(k)} \eta_j \|x_j - y\|$$

k번째 군집에 대하여 관측치 x_i의 $L_1 - DD$는 x_i가 군집 k의 L_1-중앙값 되기 위해서 x_i에 추가로 주어져야 하는 최소의 (확률)질량값을 1에서 빼준 값이다. 따라서 x_i가 중심으로부터 멀리 위치할수록 x_i가 L_1-중앙값이 되기 위해 필요한 (확률)질량값이 증가하므로 $L_1 - DD$는 0에 가까워진다. 한편 $L_1 - DD$를 다음과 같이 약간 다른 각도에서도 정의할 수 있다.

R^p에 있는 한점 z에서 관측치 x_i까지 단위벡터를 $e_i(z) = \frac{(x_i - z)}{\|x_i - z\|}$라 하면, z에 대한 k번째 군집에 있는 모든 관측치들 사이의 방향벡터의 평균은

$$\bar{e}(z|k) = \frac{\sum_{i \in I(k), x_i \neq z} \eta_i e_i(z)}{\sum_{i \in I(k)} \eta_i}$$

로 정의된다. 이 평균방향벡터를 이용하여 관측치 z에 대한 k번째 군집에서의 $L_1 - DD$를 정의하면 다음과 같다.

$$D(z|k) = 1 - \max[0, \|\bar{e}(z|k)\| - f(z|k)]$$

여기에서 $f(z|k)$는 군집 k에 있는 z와 같은 위치에 놓인 관측치들의 비율이다.

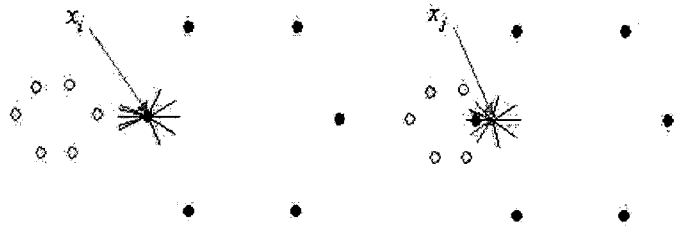
그림 2.2는 군집에서 관측치의 상대적 위치에 따라 $L_1 - DD$가 어떻게 계산되는지 보여준다. 즉, (a)를 보면 데이터들의 중앙에 위치한 z_1은 $\bar{e}(z_1) \simeq 0$이므로 z_1의 $L_1 - DD$는 1에 가까워진다는 사실을 알 수 있다. 한편, (b)에서와 같이 데이터들로부터 떨어진 곳에 위치한 z_2는 $\bar{e}(z_2) \simeq 1$이므로 z_2의 $L_1 - DD$는 0에 가까워진다.

(a) (b)

그림 2.2: 관측치의 위치에 따른 L1-DD 값의 변화

분류방법으로서의 $L1DDclass$는 TE에 있는 각 관측점의 x_i에 대하여 TR에 의해 정해진 군집들 중 가장 큰 $D(x_i|\cdot)$을 주는 군집에 분류하는 것이다.

$L_1 - DD$를 결정짓는 $\bar{e}$는 $L1DISTclass$의 $\bar{d}$와 달리 군집의 퍼짐의 정도에 거의 영향을 받지 않고 다만 방향에만 영향을 받는다. 따라서 $L1DISTclass$에 의해서 오분류되었던 관측치가 $L1DDclass$에 의해서는 제대로 분류될 수도 있고, 반대로 $L1DISTclass$에 의해서 제대로 분류되었던 관측치가 $L1DDclass$에 의해서는 오분류될 수도 있다. 그림 2.3은 그림 2.1에서 $L1DISTclass$에 의해서 분류되었던 관측치들이 $L1DDclass$에 의해서 어떻게 다르게 분류되는가를 보여준다. 즉, $L1DISTclass$에 의해서 잘못 분류되었던 x_i는 제대로 분류되는 반면에, $L1DISTclass$에 의해서 제대로 분류되었던 x_j는 잘못 분류될 수 있음을 보여준다.



(a) (b)

그림 2.3: L1DDclass에 의한 올바른 분류와 오분류의 예

2.3. L_1-거리와 $L_1 - DD$를 결합한 분류방법 : $DnDclass$

그림 2.1과 그림 2.3에서 설명한 것과 같이 군집간의 퍼짐의 정도가 다를 경우 두 군집의 경계 근처에 위치하는 관측치들에 대하여 $L1DISTclass$는 퍼짐이 작은 군집으로 분류하는 경향이 있고, $L1DDclass$는 퍼짐이 큰 군집에 분류하는 경향이 있다. 따라서 이 둘의 결합된 형태의 새로운 분류방법은 고려해 볼만하다고 할 수 있다. 이 논문에서는 Jornsten(2004)이 언급한 바와 같이 $\bar{d}$와 $L_1 - DD$를 이용하여 새로운 분류분석방법(Distance and Data-depth Classification ; $DnDclass$)을 제안하고 그 의 특징을 연구하고자 한다.

테스트자료의 관측치들 x_j에 대해서 $DnDclass$는 다음과 같이 $\widehat{k_j}$의 군집으로 분류한다.

$$\widehat{k_j} = \arg\min_k \{(1-\lambda)\bar{d}(j|TR^k) - \lambda D(j|TR^k)\}, \quad \lambda \in [0,1]$$

여기서 TR^k는 라벨k의 훈련자료에서의 관측치들이다. 또한 λ는 두 기준, $\bar{d}$와 $L_1 - DD$을 결합하는 값이다.

여기서 한가지 언급할 것은 Jornsten(2004)에 설명되어 있는 바와 같이 $L_1 - DD$는 0과 1 사이의 값을 취하는데 반해 $\bar{d}$는 데이터 단위에 따라 취하는 값의 범위가 다양하다. 따라서 가장 적절한 λ의 값은 관측치들의 단위에 따라 너무 그 변화가 너무 급격하게 나타난다. 따라서 이러한 문제점을 피하기 위해서는 관측치들에 적당한 표준화를 시키는 과정이

표 3.1: 시뮬레이션에 사용되는 이변량 정규분포

모집단	평균	분산	모집단	평균	분산
[1]	(0,0)	(1,0,0,1)	[4]	(3,3)	(9,0,0,9)
[2]	(2,2)	(4,0,0,4)	[5]	(0,0)	(1,0,0,9)
[3]	(2,2)	(1,0,0,1)	[6]	(7,0)	(9,0,0,1)

필수적이라 할 수 있다. 이 논문에서는 R에서 제공하는 표준화의 방법 중에 관측치들이 각 좌표축을 기준으로 전체범위가 ± 1 정도의 값을 갖도록 zsc와 $max1$을 이용하여 표준화 하는 과정을 거쳤다.

3. 모의실험을 통한 분류방법들의 비교

이 절에서는 시뮬레이션을 통하여 각 분류방법의 특징을 확인하고자 한다. 시뮬레이션 방법은 관측치의 90%를 이용하여 10%를 분류하는 과정을 실행하였다. 각 분류방법의 성능을 비교하기 위해서 1000번 표본을 추출해서 이 과정을 반복시행하고, 거기서 얻어진 오분류횟수들을 상자그림(Box-plot)을 이용하여 정리 비교하였다. 이 논문에서 분류방법들의 차이를 효과적으로 보여주기 위해서 고려하는 분포들은 이변량정규분포이고, 이를 정리하면 표 3.1과 같다. 여기서 평균은 평균벡터를 행벡터로 표현한 것이고, 분산은 분산행렬을 행벡터로 풀어서 표현한 것이다. 즉, 분산이 $(9,0,0,1)$이라 함은 분산 행렬이 $(9,0)$와 $(0,1)$의 두 개의 행벡터로 이루어졌음을 의미한다.

첫째, 위 절에서 우리는 군집간의 퍼짐의 정도가 다를 경우 두 군집의 경계 근처에 위치하는 관측치들에 대하여 $L1DISTclass$는 퍼짐이 작은 군집으로 분류하는 경향이 있고, $L1DDclass$는 퍼짐이 큰 군집에 분류하는 경향이 있다는 사실을 예측을 할 수 있었다. 따라서 이 두 방법을 결합한 $DnDclass$의 경우 λ값을 0의 값으로부터 1의 값까지 점점 크게 변화시키면 경계 근처의 자료들이 퍼짐이 큰 군집으로 분류되는 경향이 나타나리라 예상할 수 있다.

이를 확인하기 위해서 분산이 다른 두개의 모집단, 즉, 모집단 [1]로부터 270개(○), 모집단 [2]로부터 430개(+), 총 700개의 관측값을 생성하고, 이 중 임의로 70개를 선택해서 테스트 자료라 하였다. 그림 3.1의(a)는 테스트 자료의 군집 분포를 나타내고, (b)는 $L1DISTclass$ 경우의 분류상황을, (c)-(e)는 각각 $\lambda = 0.25, 0.5, 0.75$일 때의 분류상황을, (f)는 $L1DDclass$ 경우의 분류상황을 보여주고 있다. 여기서 보여지는 바와 같이 두 집단의 경계점의 값들이 λ의 값이 커지면서 점차로 분산이 더 큰 모집단 [2]로 분류되는 것을 알 수 있다.

둘째, 위의 서론에서 언급한 바와 같이 처음 Jornsten, Vardi와 Zhang(2002)에 의해서 L_1-DD를 이용한 분류방법이 제안된 것은 두 집단간에 퍼짐의 차이가 큰 경우에 L_1-거리에 의한 분류방법이 오분류하는 경향이 많기 때문에 이를 보완하기 위한 것이었다. 따라서 퍼짐의 정도가 비슷한 그룹들로 이루어진 경우에는 λ 값이 작은 경우가, 퍼짐의 차이가 많은 그룹들로 이루어진 경우에는 λ 값이 큰 경우가 더 효과적으로 분류할 것으로 예측할 수 있다.

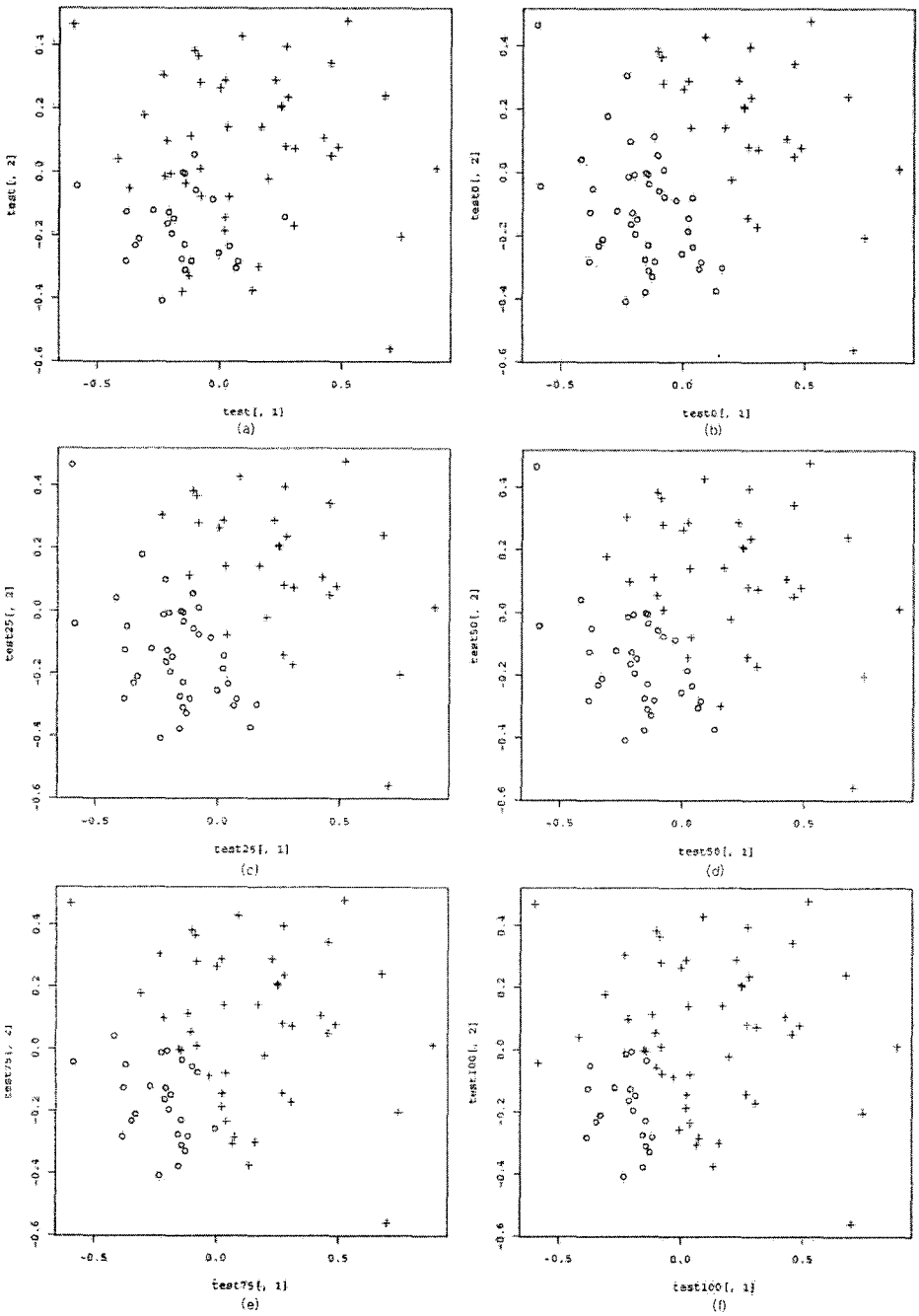


그림 3.1: λ값의 변화에 따른 분류상황의 변화

이를 확인하기 위해서 다양한 경우에 시뮬레이션을 시행하였고, 몇가지 특징있는 경우를 그림 3.2로 정리하였다. 각 경우에 두 개의 모집단에서 각각 200개의 관측치를 생성하였고, 그 10%인 40개의 관측치를 테스트자료로 하여 $\lambda = 0, 0.1, 0.25, 0.50, 0.75, 0.9, 1$에 대해서 분류를 시행하고 이를 1000번 반복한 후 상자그림을 그린 것이다. $\lambda = 0$은 $L1DISTclass$를 나타내고 이를 $L1Dist$로, $\lambda = 1$은 $L1DDclass$를 나타내고 이를 $L1DD$로 표현하였다.

그림 3.2의 (a)는 모집단 [1](○)과 [3]의 분포를, (b)는 모집단 [1](○)과 [4]의 분포를 (c)는 모집단 [5](○)와 [6]의 분포를 대략적으로 보여주는 그림이다. (a)와 (c) 는 두 모집단은 퍼짐의 정도가 같은 경우이고, (b)는 두 모집단은 퍼짐의 정도가 상당히 차이가 있는 경우이다. 여기서 우리는 (a)와 (c)의 경우는 λ가 작은 경우가 , (b)의 경우는 λ가 큰 경우가 상대적으로 더 효과적이리라 예상할 수 있다. 그러나 그 아래의 상자그림들을 보면 (c)의 경우는 예상한 바와 같이 λ가 작을수록 오분류율이 낮아지는 분포를 보이나, (a)의 경우는 다양한 λ에 대해서 오분류율이 거의 같은 분포를 보이고, (b)의 경우는 λ가 0.5인 경우에 가장 좋은 오분류율을 갖는 것을 볼 수 있다. 따라서 여기서 우리가 알 수 있는 것은 처음에 예측한 것처럼 퍼짐이 비슷한 그룹에 대해서라고 항상 λ가 작은 경우가 효과적이고, 퍼짐의 차이가 많은 경우에서는 λ가 큰 경우가 항상 효과적이라고 단정하기 어렵다는 것이다.

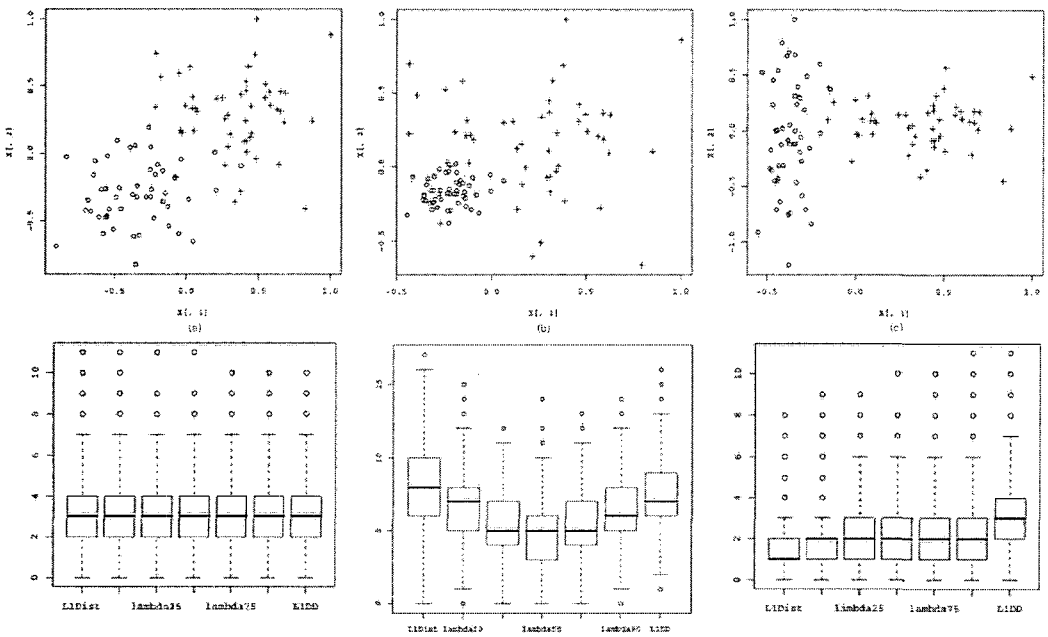


그림 3.2: 다양한 상황에서의 분류방법들의 성능비교

그림 3.2에서 (a)와 (c)가 모두 퍼짐의 정도가 같은 그룹으로 이루어져 있음에도 불구하고 서로 다른 결과가 나온 이유를 살펴보는 것은 분류방법들의 특징을 알아보는데 도움이 된다. (c)의 경우에 두 그룹의 경계에 있는 하나의 관측치를 분류하는 과정을 생각해보자.

이 관측치로부터 각 그룹에 속한 관측치들까지의 벡터들을 고려해보면 모집단 [5]에 속한 관측치들까지의 벡터들은 아주 넓은 각도에 걸쳐서 분포하는데 반해, 모집단[6]에 속한 관측치들까지의 벡터들은 주로 좁은 각도에 걸쳐서 분포하게 된다. 넓은 각도의 단위벡터들의 평균은 좁은 각도의 단위벡터들의 평균보다 작은 길이를 갖게 된다. λ의 값이 클수록 그 벡터들의 길이의 평균보다는 단위벡터들의 평균의 길이만을 고려하게 되므로 모집단 [6]에 속하면서 경계주변에 있는 많은 관측치들이 모집단 [5]로 오분류되는 경향이 나타나게 되는 것이다. 따라서 $L1DISTclass$가 다른 어떤 λ 값을 갖는 경우보다 더 나은 성능을 보여주게 되는 것이다.

마지막으로 그림 3.2의 (b)의 경우에 오분류되는 상황을 생각해보자. 앞에서 언급한 바와 같이 두 그룹의 퍼짐의 정도가 차이가 많을 경우 경계근처에 위치한 관측값들은 λ의 값에 따라 그 값이 클수록 퍼짐의 정도가 큰 그룹에 분류되는 경향이 있다. 따라서 경계근처에 있는 관측치들이 퍼짐이 큰 그룹에 많이 속해 있을수록 λ 값이 클 때 더 나은 분류가 이루어지게 된다. 이를 살펴보기 위해서 그림 3.2의 (b)의 경우와 같은 분포를 갖는 두 개의 모집단을 대상으로, 퍼짐이 큰 모집단에서 점점 더 많은 관측치를 생성하면서 분류방법들이 어떤 성능의 변화를 보이는지 알아보았다. 즉, 모집단 [1]에서보다 모집단 [4]로부터 점점 더 많은 관측치를 생성하면서 각 분류방법의 성능이 어떻게 변화하는지 시뮬레이션을 시행하고, 이를 그림 3.3에 정리하였다.

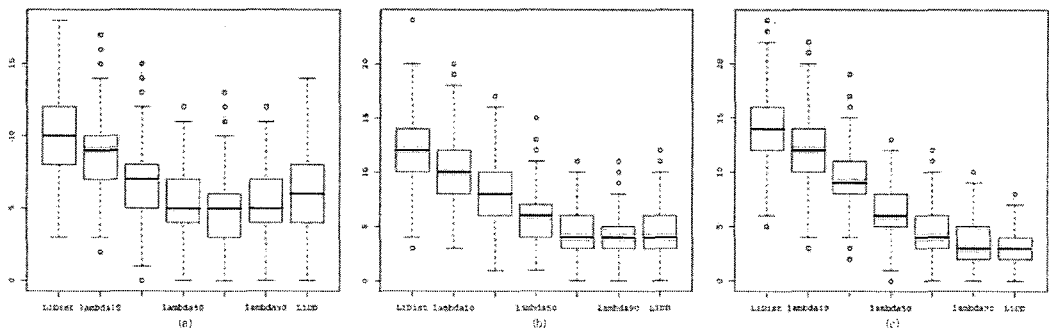


그림 3.3: 군집들의 관측치 갯수 변화에 따른 분류방법들의 성능 변화

그림 3.3의 (a)-(c)는 전체 400개의 관측치 중 모집단 [4]에서의 관측치의 수를 250,300,350으로 증가시킬 때 그림 3.2의 (b)의 상자그림이 어떻게 변화하는지 보여주는 것이다. 예측한 바와 같이 퍼짐이 큰 모집단에 속한 관측치의 갯수가 많을수록, 따라서 경계근처에 있는 관측치들이 퍼짐이 큰 그룹에 더 많이 속해 있을수록 λ 값이 클 때 더 나은 분류가 이루어지게 되는 것을 확인할 수 있다.

4. 결론

위의 절에서 확인한 바와 같이 몇 개의 특이한 경우가 아닌 많은 상황에서는 적당한 L_1-거리와 $L_1 - DD$를 모두 이용하는 $DnDclass$가 양극단의 $L1DISTclass$나 $L1DDclass$보다

더 나은 성능을 갖는 것을 볼 수 있다. 따라서 주어진 자료에서는 Cross Validation을 통해서 적당한 λ값을 결정한 후 선택된 λ 값을 이용한 *DnDclass*를 분류기법으로 제시하는 것이 더 바람직할 것으로 판단된다.

또한, 퍼짐의 정도가 서로 다른 그룹들을 분류하는 경우에 경계점에 위치한 관측치들은 λ의 값이 클수록 퍼짐이 큰 그룹으로 분류되는 경향이 있다. 따라서 큰 군집에 속하는 관측치의 갯수가 많을수록 더 큰 λ 값을 갖는 분류방법이 더 효과적인 것을 확인할 수 있다.

참고문헌

Alon, U., Barkai, N., Notterdam, D. A., Gish, K., Ybarra, S., Mack, D. and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceeding of the National Academy of Sciences*, 96, 6745-6750.

Christmann, A. (2000). Classification Based on the SVM and on Regression Depth. *Statistical data analysis based on the L1norm and related methods.* Birkhauser, Statistics for industry and technology. Y. Dodge editor.

Ghosh, A. K. and Chaudhuri, P. (2005). On Maximum Depth and Related Classifiers. *The Scandinavian Journal of Statistics*, 32, 327-350.

Jornsten, R. (2004). Clustering and Classification Based on the L1 data depth. *Journal of Multivariate Analysis*, 90, 67-89.

Jornsten, R., Vardi, Y. and Zhang, C-H. (2002). A Robust Clustering Method and Visualization Tool Based on Data Depth. *Statistical data analysis based on the L1norm and related methods.* Birkhauser, Statistics for industry and technology. Y. Dodge editor.

Liu, R., Parelius, J. and Singh, K. (1999). Multivariate analysis by data depth : descriptive statistics, graphics and inference (with discussion). *The Annals of Statistics*, 27, 783-858.

Tibshirani, R., Walther, G., Botstein, D. and Brown, P. (2001). Cluster validation by prediction strength. Technical report, Stanford University, Department of Biostatistics.

Vardi, Y. and Zhnag, C-H. (2000). The multivariate L-1 median and associated data depth. *Proceeding of the National Academy of Sciences*, 97, 1423-1426.

[2005년 7월 접수, 2006년 1월 채택]

Comparison Studies of Classification Methods based on L_1-Distance and L_1-Data Depth*

Soojin Baek[1] Jeankyung Kim [2] Jinsoo Hwang [3]

ABSTRACT

We consider a new classification method($DnDclass$) combining two classification rules based on L_1-distance($L1DISTclass$) and L_1-data depth($L1DDclass$). To investigate characteristics and to evaluate the performance of these classification methods, we use simulation data in various settings. Through this simulation study, we can confirm that the new method, $DnDclass$, performs relatively well in many cases.

Keywords: L_1-distance, L_1-data depth, classification

* This research was partially supported by Inha University Grant, INHA-31630-01.

1) Researcher, Division of Bacterial Respiratory Infections, Center for Infectious Disease Research, Korea Center for Disease Control & Prevention, Korean National Institute of Health, 5, Nokbeon-dong Eunpyung-gu, 122-701, Seoul, Korea.
E-mail: sjbaek@stat.inha.ac.kr

2) (Corresponding author) Professor, Department of Statistics, Inha University, 253 Yonghyun-Dong, Nam-Gu, 402-751, Incheon, Korea.
E-mail: jkkim@stat.inha.ac.kr

3) Professor, Department of Statistics, Inha University, 253 Yonghyun-Dong, Nam-Gu, 402-751, Incheon, Korea.
E-mail: jshwang@stat.inha.ac.kr