수정된 적응 최근접 방법을 활용한 판별분류방법에 대한 연구

맹진우¹ · 방성완² · 전명식³

¹고려대학교 통계학과, ²고려대학교 통계학과, ³고려대학교 통계학과

(2010년 8월 접수, 2010년 9월 채택)

요 약

비모수적 판별분류방법인 k -Nearest Neighbors Classification(KNNC) 방법은 널리 사용되고 있지만 고정된 이웃의 개수를 사용하며 또한 집단변수의 정보를 활용하지 않음으로서 자료의 국소적 특징을 반영하지 못하는 단점이 있다. Adaptive Nearest Neighbors Classification(ANNC) 방법과 Modified k -Nearest Neighbors Classification(MKNNC) 방법은 각각 이러한 단점들을 보완하기 위해 제안된 방법이다. 본 연구에서는 ANNC 방법과 MKNNC 방법의 장점을 결합한 Modified Adaptive Nearest Neighbors Classification(MANNC) 방법을 제안하였다. 나아가, 제안된 방법의 활용 가능성을 살펴보고자 실제자료에 대한 분석과 모의실험을 통해 기존의 방법들과 비교하였다.

주요어: 판별분류분석, adaptive nearest neighbors, k -nearest neighbors, modified adaptive nearest neighbors.

1. 서론

여러 개의 서로 상관된 연속변수(관별변수)와 소속집단을 나타내는 반응변수로 구성된 다변량 자료에 대한 판별분류분석에서 k -Nearest Neighbor Classification(KNNC) 방법은 각 개체에 가장 가까운 k 개 이웃들의 소속집단 중에서 가장 빈도가 높은 집단으로 해당 개체를 분류하는 방법이다. 이러한 KNNC 방법은 다변량 정규성 등의 모수적 모형이 만족되지 않을 때에도 강건성(robustness)을 지니는 비모수적 판별분류방법으로 널리 활용되고 있다. 하지만 KNNC 방법은 자료전체에 대하여 고정된 k 개의 이웃을 선정하므로 각 개체가 지니는 국소적 특징을 고려하지 못한다. 따라서 각 개체에 따라 소속 집단 결정에 사용하는 이웃의 개수 k 를 변화시키는 방법이 통계적 효율성을 더 높일 것으로 예상되며, 이에 대한 관련 연구로는 Friedman (1994), Hastie와 Tibshirani (1996) 등을 들 수 있다. 한편, Jhun 등 (2007)은 결측치 대체방법에 있어 k -최근접(k -nearest neighbors) 방법의 대안으로 자료의 국소적 특징을 고려한 적응 최근접(adaptive nearest neighbors) 방법을 제안하였고, 이와 유사하게 판별분류에서 Adaptive Nearest Neighbors Classification(ANNC) 방법이 연구되었다 (전명식과 최인경, 2009).

그러나 기존의 판별분류분석방법들은 각 개체에 대하여 이미 알려진 소속집단을 나타내는 반응변수가 주어져 있음에도 불구하고 각 개체들의 집단 정보를 활용하고 있지 않다. 이러한 기존 연구들의 제한점을 보완하기 위하여, Song 등 (2007)은 개체들의 소속집단 정보를 활용한 새로운 거리행렬을 이용하여 k 개의 최근접 이웃들 중 가장 유익한(informative) 개체를 선택함으로써 해당 개체를 분류하는 방법을 제안하였으며, Parvin 등 (2008)은 집단 정보와 거리 정보를 함께 활용하여 근사성을 측정함으로써 근

본 연구는 고려대학교 정경대학 교수 연구비에 의하여 수행되었음.

³교신저자: (136-701) 서울시 성북구 안암동 5가, 고려대학교 통계학과, 교수. E-mail: jhun@korea.ac.kr

그림 2.1. 타당도 계산의 예제

사성이 높은 k 개의 최근접 이웃을 선택하는 Modified k -Nearest Neighbors Classification(MKNNC) 방법을 제안하였다.

본 논문에서는 KNNC 방법의 단점을 보완한 ANNC 방법과 MKNNC 방법의 장점을 결합한 수정된 적응 최근접 방법을 활용한 판별분류방법(Modified Adaptive Nearest Neighbors Classification; MANNC)을 제안하고자 한다. 제안된 MANNC 방법은 각 개체가 지니는 국소적 특징을 고려하여 각 개체에 따라 소속집단 결정에 사용하는 최근접 이웃의 개수 k 를 변화 시키는 장점과 이미 알려진 소속집단의 정보를 활용한다는 두 가지의 장점을 동시에 지니게 되어 통계적 효율성을 높일 수 있을 것으로 기대된다.

2. MANNC 방법

p 개의 연속형 판별변수와 n 개의 개체로 구성된 $n \times p$ 자료행렬 $X = (x_{ij})$, $i = 1, 2, \dots, n$, $j = 1, 2, \dots, p$ 를 고려하자. 여기서 x_{ij} 는 i 번째 개체의 j 번째 변수에 대한 연속형 관찰값이고, n 개의 개체를 $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ (단, $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})'$)으로, 각 개체가 소속된 집단을 G_k ($k = 1, 2, \dots, g$)로 표기하자. 전명식과 최인경 (2009)은 KNNC 방법이 자료의 국소적 특징을 고려하지 못하는 제한점을 보완하기 위해 고정된 이웃의 개수 k 대신 적응적으로 이웃의 개수를 선택하는 ANNC 방법을 제시하였고, 모의실험과 실제자료의 분석을 통해 효율성을 보였다. 그러나 ANNC 방법은 개체들의 판별분류를 위해 개체들 사이의 거리만을 사용하여 근사성을 측정하기 때문에 이미 알려져 있는 소속집단의 정보를 활용하지 못하는 제약이 있다. 따라서 본 논문에서는 개체들 사이의 근사성을 측정할 때, 개체들 사이의 거리는 물론 Parvin 등 (2008)에 의해 제시된 각 개체의 타당도(validity)를 동시에 이용하는 MANNC 방법을 제안하고자 한다. 어떤 개체의 타당도란 자신과 같은 집단의 개체들이 자신 주위에 얼마나 존재하는 지에 대한 척도이며, i 번째 $\mathbf{x}_i$ 개체의 타당도는

$$V(i) = \frac{1}{h} \sum_{k=1}^h S(r(\mathbf{x}_i), r(N_k(i))), \quad i = 1, 2, \dots, n$$

그림 2.2. ANNC 방법과 MANNC 방법의 비교

와 같이 구해진다. 여기서 h 는 타당도 계산에 사용할 이웃의 개수, $r(\mathbf{x}_i)$ 는 $\mathbf{x}_i$ 의 집단변수, $N_k(i)$ 는 $\mathbf{x}_i$ 의 k 번째 최근접 이웃을 나타내고 S 는 $r(\mathbf{x}_i)$ 와 $r(N_k(i))$ 가 일치하면 1, 아니면 0인 함수이다.

타당도 계산의 이해를 돕기 위해 그림 2.1에서는 $h = 5$ 를 사용하여 두 개체 A와 B의 타당도를 계산하였다. 개체 A와 가장 가까운 5개의 개체들이 타당도 계산을 위해 선택되어지며 이 중에서 4개의 개체가 개체 A와 집단이 일치하므로 개체 A에 대한 타당도는 0.8이 된다. 같은 방법으로 개체 B의 경우 타당도는 0.2로 계산된다.

이제 i 번째 개체 $\mathbf{x}_i$ 와 j 번째 개체 $\mathbf{x}_j$ 사이의 거리를 d_{ij} 로, 개체 $\mathbf{x}_i$ 와 j 번째로 가까운 개체 사이의 거리를 $d_{i(j)}$ 로 표기하자. 타당도가 높은 개체가 최근접 이웃으로 선택되어질 가능성을 높이기 위해 거리의 역수에 타당도를 곱한 크기 $n \times n$ 인 근사성 척도행렬 W 를 정의하고, 개체 $\mathbf{x}_i$ 와 개체 $\mathbf{x}_j$ 사이의 근사성을 w_{ij} 로, 개체 $\mathbf{x}_i$ 와 j 번째로 근사성이 높은 개체 사이의 근사성을 $w_{i(j)}$ 로 나타내자. ANNC 방법이 개체 $\mathbf{x}_i$ 와 가장 가까운 거리 $d_{i(1)}$ 에 대한 다른 개체들과의 거리 d_{ij} ($i \neq j$)의 비(ratio)를 이용하여 이웃집단을 구성한 것과 마찬가지로 MANNC 방법은 개체 $\mathbf{x}_i$ 와 가장 높은 근사성 $w_{i(1)}$ 에 대한 다른 개체들과의 근사성 w_{ij} ($i \neq j$)의 비(ratio)를 이용하여 이웃집단을 선택하게 된다. 이와 같은 MANNC 방법은 ANNC 방법의 장점인 자료의 국소적 특징을 고려하여 각 개체에 따라 소속집단 결정에 사용하는 이웃의 개수 k 를 적응적으로 부여하는 것과 MKNNC 방법의 장점인 이미 알려져 있는 개체들의 집단정보를 활용하여 개체들 사이의 근사성을 측정하는 특징을 모두 가지게 된다.

2.1. MANNC 알고리즘

MANNC 방법은 다음과 같은 5단계의 과정으로 이루어진다. 이는 전명식과 최인경 (2009)이 제안한 ANNC 방법과 Parvin 등 (2008)이 제안한 MKNNC 방법을 활용한 것이다.

단계 1. 각 개체에 대한 타당도 $V(i)$, $i = 1, \dots, n$ 을 계산한다.

단계 2. 개체들 간의 거리 d_{ij} ($i, j = 1, 2, \dots, n$)에 조정계수(tuning parameter) δ 를 더한 새로운 거리 $d_{ij}^\delta = d_{ij} + \delta$ 를 구한다.

단계 3. 조정된 거리 d_{ij}^δ 와 타당도를 이용하여 크기 $n \times n$ 인 근사성 척도행렬 W 를 구한다.

$$W = (w_{ij}) = \left(V(j) \times \frac{1}{d_{ij}^\delta} \right), \quad i, j = 1, 2, \dots, n.$$

단계 4. 임계치(thresholding factor) q 를 정하고, 개체 $\mathbf{x}_i$ ($i = 1, \dots, n$)과 가장 큰 근사성 $w_{i(1)}$ 와 w_{ij} ($i \neq j$)와의 감소비가 q 보다 크거나 같은 개체들로 이루어진 이웃집단 $N(i)$ 를 정한다.

$$N(i) = \left(\mathbf{x}_j : \frac{w_{ij}}{w_{i(1)}} \geq q \right), \quad 0 < q \leq 1.$$

단계 5. 개체 $\mathbf{x}_i$ ($i = 1, \dots, n$)를 이웃집단 $N(i)$ 에 속한 개체들 중에서 가장 빈번하게 관측되는 집단으로 분류한다.

기존의 ANNC 방법과 제안된 MANNC 방법의 주요 차이점은 단계 3에 있다. 즉, 개체들 간의 근사성을 측정할 때 ANNC 방법은 개체들 간의 거리에만 의존하는 반면, MANNC 방법은 타당도를 이용하여 개체들의 집단 정보를 함께 이용한다. 단계 4에서는 개체들 간의 근사성이 높은 개체들을 선택하는 과정에서 그 개체와의 근사성 감소비가 임계치 q 보다 크거나 같은 개체들만을 이웃으로 선택함으로써, 해당개체 주변의 밀도나 구조를 포함한 국소적 성격을 반영한다. 이 때 q 가 크면 적은 수의 개체들을 이웃으로 선택하게 되고 q 가 작으면 많은 수의 개체들을 이웃으로 선택하게 된다. 그런데 어떤 개체의 이웃들을 선택하고자 할 때, 해당 개체와 가장 큰 근사성 $w_{i(1)}$ 에 사용된 d_{ij} 가 0에 가까우면 다음으로 가까운 개체의 근사성 감소비 $w_{i(2)}/w_{i(1)}$ 는 무한대로 증가하므로 거리 d_{ij} 들에 대하여 조정계수 δ 를 단계 2에서 이용하였으며, 통상적으로 거리행렬의 중앙값을 사용한다. 이는 기존의 ANNC 방법에서 제시된 것과 유사하다.

그림 2.2는 ANNC 방법과 MANNC 방법을 이용한 개체 A와 B의 관별분류 과정을 나타내고 있다. 각 개체의 괄호안의 값은 $h = 3$ 을 이용하여 계산되어진 타당도이며, 두 개체 A와 B에서 실선의 원은 ANNC 방법의 최근접 이웃 선택 결과이고 점선의 원은 MANNC 방법의 최근접 이웃의 선택 결과이다. 먼저 ANNC 방법을 살펴보면 개체의 집단 정보를 활용하지 않아 가까운 개체들만을 근접 이웃으로 선택하지만, MANNC 방법의 경우 거리정보와 타당도를 함께 고려함으로써 가까운 개체들 중 타당도가 높은 개체들을 선택하는 결과를 볼 수 있다. 개체 A의 경우 ANNC 방법을 사용하여 분류하게 된다면 “+” 집단으로 오분류 하게 되지만 MANNC 방법의 경우 “o” 집단으로 정분류 하게 되며, 개체 B의 경우 ANNC 방법은 최근접 이웃으로 타당도가 0으로 계산된 개체를 선택하지만 MANNC 방법은 이 개체를 선택하지 않는 장점을 확인할 수 있다. 그림 2.2의 예제 분석으로부터 각 개체의 집단 정보를 활용한 MANNC 방법이 기존의 방법에 비해 그 성능이 우수할 것으로 기대되며, 특히 이상치가 존재하거나 집단별 자료의 겹침 현상이 심할 경우 기존의 ANNC 방법보다 MANNC 방법이 더 좋은 분류 성능을 보일 것으로 예상할 수 있다.

2.2. 모수의 선택

제안된 MANNC 방법과 기존의 방법들(KNNC, ANNC, MKNNC)은 각각 조정모수(tuning parameter)의 선택을 필요로 한다. 본 논문에서는 KNNC와 MKNNC 방법에서 이웃의 개수에 대한 모수인 k 는 1부터 자료의 50%까지를 고려하였고, ANNC와 MANNC 방법에서 이웃의 개수를 결정하는 임계치 q 는 ANNC 방법의 경우 1부터 2까지, MANNC 방법의 경우 0부터 1까지 0.001씩 등간격으로 선택하였다. MKNNC와 MANNC 방법에서 타당도 계산에 사용되는 이웃의 개수에 대한 모수 h 는 자료의

표 2.1. 판별분류 방법의 분석결과 비교(Haberman의 자료)

		KNNC		ANNC		MKNNC		MANNC	
		집단1	집단2	집단1	집단2	집단1	집단2	집단1	집단2
실제 집단	집단1	204	21	210	15	210	15	217	8
	집단2	56	25	60	21	59	22	63	18
정분류율		0.748		0.755		0.758		0.768	
선택된 모수		$k = 13$		$q = 1.954$		$k = 17$		$q = 0.894$	

표 2.2. 판별분류 방법에 따른 선택된 이웃의 개수와 타당도 및 거리(Haberman의 자료)

		KNNC			ANNC			MKNNC			MANNC		
		이웃 개수	타당도 (거리)	분류 결과	이웃 개수	타당도 (거리)	분류 결과	이웃 개수	타당도 (거리)	분류 결과	이웃 개수	타당도 (거리)	분류 결과
$\mathbf{x}_{76}$	13	0.490 (5.50)	오분류	6	0.501 (4.39)	오분류	17	0.509 (4.10)	오분류	3	0.544 (3.48)	정분류	
$\mathbf{x}_{86}$	13	0.501 (5.73)	오분류	14	0.502 (6.46)	정분류	17	0.529 (5.37)	오분류	40	0.717 (10.22)	정분류	
$\mathbf{x}_{107}$	13	0.490 (5.46)	오분류	11	0.478 (5.71)	오분류	17	0.660 (6.27)	정분류	6	0.626 (6.08)	정분류	

10%부터 50%까지를 10%씩 증가시키며 고려하였고, ANNC, MANNC 방법에서의 조정계수 δ 는 거리 행렬의 중앙값을 사용하였다. 본 논문의 실제 사례분석 및 모의실험에서는 각각의 판별분류방법에 따른 모수 k, q, h 의 선택을 위해 교차타당성(leave-one-out cross validation) 방법을 이용하였다.

2.3. 사례분석

본 논문에서 제시하는 MANNC 방법의 성능을 기존의 방법들과 비교하기 위하여 UCI Machine Learning Repository(<http://www.ics.uci.edu/~mllearn/MLRepository>)에서 제공되는 Haberman의 자료를 분석하였다. 이 자료는 1958년에서 1970년도 까지 University of Chicago's Billings Hospital에서 유방암 수술을 받은 306명 환자에게 대한 생존 자료로, 3개의 판별변수(수술 당시의 나이, 수술 받은 년도, 발견된 양성 임파선의 개수)와 집단 변수(survival status)로 이루어져 있다. 집단 변수가 1이면 수술 후 5년 이상 생존, 2이면 5년 이내에 사망을 의미한다.

표 2.1은 교차타당성 방법으로 선택된 모수에 대한 판별분류방법들의 분류결과를 나타내고 있으며, 이로부터 MANNC 방법이 기존 방법들에 비해 그 성능이 우수함을 알 수 있다. MANNC 방법을 기존 방법들과 자세히 비교하기 위해 3개의 개체 $\mathbf{x}_{76}$, $\mathbf{x}_{86}$, $\mathbf{x}_{107}$ 에 대하여 선택된 이웃의 개수, 이웃 개체들의 타당도 평균과 거리평균을 살펴보았으며, 그 결과는 표 2.2에 있다. 표 2.2로부터 KNNC 방법과 ANNC 방법의 경우 최근접 이웃으로 선택된 개체들의 타당도가 MKNNC 방법이나 MANNC 방법에 비해 낮은 것을 알 수 있고, KNNC 방법과 MKNNC 방법의 경우 최근접 이웃의 수가 k 개로 고정되어 있는데 반해 ANNC 방법과 MANNC 방법은 개체의 국소적 특징을 고려하여 최근접 이웃의 수를 각 개체에 따라 다르게 선택하고 있는 것을 확인할 수 있다. 특히 76번 개체의 경우를 살펴보면 다른 방법들에 비해 먼 거리에 있는 개체들을 선택하지만 타당도가 높은 3개의 개체만을 선택함으로써 다른 방법에서 모두 오분류된 개체를 정분류하는 것을 알 수 있다. 이와 같이 높은 타당도를 가지는 개체 중 가까운 이웃개체들을 국소적 특성을 반영해 선택하는 MANNC 방법을 사용함으로써 실제 판별분류 성능을 향상시킬 수 있는 것을 확인할 수 있다.

표 3.1. 판별분류 방법의 정분류를 비교(모형 1)

판별분류방법	$L = 3$		$L = 5$		$L = 7$	
	평균	표준편차	평균	표준편차	평균	표준편차
KNNC	0.690	0.049	0.622	0.053	0.573	0.048
ANNC	0.706	0.044	0.641	0.050	0.592	0.045
MKNNC	0.726	0.045	0.668	0.047	0.629	0.045
MANNC	0.727	0.038	0.675	0.039	0.638	0.038

그림 3.1. 판별분류 방법의 정분류를 비교(모형 1)

3. 모의실험

여러 다양한 자료의 판별분류에서 제안된 MANNC 방법과 기존의 방법들을 비교하기 위해, 자료의 겹침 정도, 집단의 수, 판별변수의 차원을 변화시키면서 그 성능을 비교하는 모의실험을 하였다.

모형 1. 첫 번째 모의실험은 두 집단의 평균 벡터를 이동시켜 가면서 자료의 겹침 정도에 따라 4가지 판별분류 방법의 성능이 어떻게 변화하는지 알아보기 위한 것으로, 이는 정운경과 백장선 (2007)이 군집 타당성분석 기법의 성능비교에서 사용한 모의실험 방법과 유사하다. 집단 1(G_1)과 집단 2(G_2)의 자료는

$$G_1 : \mathbf{x} \sim N_2(\boldsymbol{\mu}_1, \Sigma), \quad G_2 : \mathbf{x} \sim N_2(\boldsymbol{\mu}_2, \Sigma)$$

의 분포 가정 하에서 각각 크기가 50인 랜덤포본을 생성하였다. 두 집단 모두 동일한 공분산 행렬 $\Sigma = (\sigma_{ij})$; $\sigma_{ii} = 2$, $\sigma_{ij} = 0.25 (i \neq j)$ 을 가지며, 두 집단의 평균 벡터는

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix} + L \times \begin{pmatrix} 0.1 \\ 0.1 \end{pmatrix}, \quad \boldsymbol{\mu}_2 = \begin{pmatrix} 2 \\ 2 \end{pmatrix} - L \times \begin{pmatrix} 0.1 \\ 0.1 \end{pmatrix}, \quad L = 0, 1, 2, \dots, 8$$

표 3.2. 판별분류 방법의 정분류율 비교(모형 2)

판별분류방법	$g = 3$		$g = 4$		$g = 6$	
	평균	표준편차	평균	표준편차	평균	표준편차
KNNC	0.584	0.044	0.523	0.046	0.516	0.054
ANNC	0.615	0.037	0.558	0.045	0.554	0.045
MKNNC	0.625	0.035	0.577	0.039	0.574	0.040
MANNC	0.634	0.033	0.585	0.037	0.582	0.040

그림 3.2. 판별분류 방법의 정분류율 비교(모형 2)

와 같이 L 을 증가시키에 따라 $(1, 1)'$ 의 방향으로 이동시켰다.

언급된 4가지 판별분류 방법의 정분류율을 비교하기 위하여 100회의 독립반복시행을 하였으며, $L = 3, 5, 7$ 에서의 결과가 표 3.1과 그림 3.1의 박스 플롯에 요약되어 있다. 표 3.1과 그림 3.1로부터 자료의 겹침 정도가 심해질수록 MANNC 방법의 성능이 기존의 방법에 비해 우수함을 알 수 있다. 다시 말해, 집단별 자료가 비교적 잘 나누어져 있는 경우 집단의 정보를 이용하지 않는 것이 큰 손실이라고 보기에는 힘들지만, 이상치가 존재하거나 집단의 겹침 정도가 심한 자료에서는 이미 알려져 있는 집단의 정보를 이용하지 않는다면 높은 성능을 기대하기 어려울 것으로 판단된다.

모형 2. 두 번째 모의실험은 3개 이상의 집단에서 4가지 판별분류 방법의 성능을 비교하기 위한 것으로, 집단 수 $g = 3, 4, 6$ 을 고려하였다. 3 집단의 경우 모평균은 각각 $\mu_1 = (0, 0)'$, $\mu_2 = (1, 1)'$, $\mu_3 = (0, 2)'$, 4 집단의 경우 $\mu_1 = (0, 0)'$, $\mu_2 = (0, 2)'$, $\mu_3 = (2, 0)'$, $\mu_4 = (2, 2)'$, 6 집단의 경우 $\mu_1 = (0, 0)'$, $\mu_2 = (0, 2)'$, $\mu_3 = (2, 0)'$, $\mu_4 = (2, 2)'$, $\mu_5 = (1, 1)'$, $\mu_6 = (3, 1)'$ 이다. 모든 집단은 동일한 공분산 행렬 $\Sigma = (\sigma_{ij})$; $\sigma_{ii} = 1.5$, $\sigma_{ij} = 0.25 (i \neq j)$ 을 가지며,

$$G_1 : \mathbf{x} \sim N_2(\mu_1, \Sigma), \dots, G_g : \mathbf{x} \sim N_2(\mu_g, \Sigma)$$

표 3.3. 판별분류 방법의 정분류율 비교(모형 3)

판별분류방법	$p = 3$		$p = 4$		$p = 8$	
	평균	표준편차	평균	표준편차	평균	표준편차
KNNC	0.665	0.053	0.665	0.054	0.706	0.049
ANNC	0.691	0.047	0.704	0.043	0.749	0.040
MKNNC	0.718	0.042	0.731	0.038	0.762	0.043
MANNC	0.727	0.042	0.740	0.037	0.773	0.042

그림 3.3. 판별분류 방법의 정분류율 비교(모형 3)

의 분포 가정 하에서 집단별로 크기가 50인 랜덤포본을 생성하였고, 100회 독립반복시행을 하였다.

판별분류 방법에 따른 정분류율은 표 3.2와 그림 3.2의 박스 플롯에 요약되어 있으며, 본 논문에서 제시하는 MANNC 방법이 기존의 다른 판별분류 방법들에 비해 좋은 성능을 나타내고 있음을 알 수 있다.

모형 3. 세 번째 모의실험은 3차원 이상의 자료에서 4가지 판별분류 방법의 성능을 비교하기 위한 것으로, 두 집단 자료에서 판별변수 $\mathbf{x}$ 의 차원 수 $p = 3, 4, 5$ 를 고려하였다. 집단 1(G_1)의 모평균 $\boldsymbol{\mu}_1 = 0 \times \mathbf{1}_p$, 집단 2(G_2)의 모평균 $\boldsymbol{\mu}_2 = 2 \times \mathbf{1}_p$ 이며, $\mathbf{1}_p = (1, \dots, 1)'$ 는 모든 원소의 값이 1인 p 차원 벡터이다. 두 집단의 공분산 행렬은 모형2에서와 동일하며,

$$G_1 : \mathbf{x} \sim N_p(\boldsymbol{\mu}_1, \Sigma), \quad G_2 : \mathbf{x} \sim N_p(\boldsymbol{\mu}_2, \Sigma)$$

의 분포 가정 하에서 각각 크기가 50인 랜덤포본을 생성했고, 이를 100회 독립반복시행을 했다. 표 3.3과 그림 3.3의 박스 플롯은 모형 3에서의 판별분류분석 방법에 따른 정분류율을 나타내고 있으며, 모의실험에서 고려된 모든 차원에서 제안된 MANNC 방법의 성능이 가장 높은 것을 알 수 있다.

4. 결론

비모수적 판별분류방법인 k -Nearest Neighbors Classification 방법은 자료의 위치적 특성을 고려하지 않고 모든 개체에 동일한 이웃 개체 k 개를 선택하는 제약과 판별자료가 가지고 있는 집단변수의 정보를 전혀 이용하지 않는다는 단점을 가지고 있다. 본 연구에서는 자료의 국소적 특징을 반영하여 적절한 수의 이웃개체를 선택하는 ANNC 방법과 개체간의 근사성 측정에서 집단변수의 정보를 활용하는 MKNNC 방법의 장점을 결합한 Modified Adaptive Nearest Neighbors Classification(MANNC) 방법을 제안하였으며, 실제사례 분석을 통해 제안된 MANNC 방법이 기존의 방법들보다 높은 정분류율을 나타내는 것을 확인할 수 있었다. 또한 다양한 모형에 대한 모의실험을 통해 집단별 자료의 겹침현상이 심해질수록 MANNC 방법의 성능이 기존의 방법들에 비해 높은 성능을 보이는 것과 집단의 수가 증가하고 판별변수의 차원이 증가하더라도 기존의 방법들에 비해 높은 성능을 유지할 수 있는 가능성을 확인할 수 있었다.

참고문헌

- 전명식, 최인경 (2009). Adaptive nearest neighbors를 활용한 판별분류방법, <응용통계연구>, **22**, 479-488.
- 정윤경, 백장선 (2007). 고차원(유전자 발현)자료에 대한 군집 타당성 분석 기법의 성능비교, <응용통계연구>, **20**, 167-181.
- Friedman, J. (1994). *Flexible Metric Nearest Neighbor Classification*, Technical report, Stanford University.
- Hastie, T. and Tibshirani, R. (1996). Discriminant adaptive nearest neighbor classification, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **18**, 607-616.
- Jhun, M., Jeong, H. C. and Koo, J. Y. (2007). On the use of adaptive nearest neighbors for missing value imputation, *Communications in Statistics: Simulation and Computation*, **36**, 1275-1286.
- Parvin, H., Alizadeh, H. and Minaei-Bidgoli, B. (2008). MKNN: Modified k -nearest neighbor, *Proceedings of the World Congress on Engineering and Computer Science*, 22-24.
- Song, Y., Huang, J., Zhou, D., Zha, H. and Giles, C. L. (2007). IKNN: Informative k -nearest neighbor pattern classification, *PKDD 2007*, 248-264.

On the Use of Modified Adaptive Nearest Neighbors for Classification

Jinwoo Maeng¹ · Sungwan Bang² · Myoungshic Jhun³

¹Department of Statistics, Korea University; ²Department of Statistics, Korea University

³Department of Statistics, Korea University

(Received August 2010; accepted September 2010)

Abstract

Even though the k -Nearest Neighbors Classification(KNNC) is one of the popular non-parametric classification methods, it does not consider the local features and class information for each observation. In order to overcome such limitations, several methods have been developed such as Adaptive Nearest Neighbors Classification(ANNC) and Modified k -Nearest Neighbors Classification(MKNNC). In this paper, we propose the Modified Adaptive Nearest Neighbors Classification(MANNC) that employs the advantages of both the ANNC and MKNNC. Through a real data analysis and a simulation study, we show that the proposed MANNC outperforms other methods in terms of classification accuracy.

Keywords: Adaptive nearest neighbors, classification analysis, k -nearest neighbors, modified adaptive nearest neighbors.

This research was supported by a Korea University grant.

³Corresponding author: Professor, Department of Statistics, Korea University, Seoul 136-701, Korea.

E-mail: jhun@korea.ac.kr