응용통계연구 (2009)
22(3), 479–488

Adaptive Nearest Neighbors를 활용한 판별분류방법

전명식[1] · 최인경[2]

[1]고려대학교 통계학과, [2]고려대학교 통계학과

(2009년 2월 접수, 2009년 3월 채택)

요 약

비모수적 판별분류방법으로 널리 사용되는 k-Nearest Neighbors Classification(KNNC) 방법은 자료의 국소적 특징을 고려하지 않고 전체 자료에 대해 고정된 이웃의 개수 k를 사용하여 개체를 분류하는 방법이다. 본 연구에서는 KNNC의 대안으로 자료의 국소적 특징을 고려하는 Adaptive Nearest Neighbors Classificaion(ANNC) 방법을 제안하였다. 제안된 방법의 특징을 규명하기 위하여 실제 자료에 대한 분석을 통하여 제안된 방법의 응용 가능성을 제시하였으며, 나아가 모의실험을 통하여 기존의 방법과의 효율성을 비교하였다.

주요용어: 판별분류분석, Adaptive nearest neighbors, k-nearest neighbors.

1. 서론

판별분류분석(classification analysis)은 각 개체에 대하여 이미 알려진 소속집단을 나타내는 반응변수와 집단 간 차이를 식별하는데 사용되는 여러 개의 서로 상관된 연속변수(판별변수)를 가진 다변량 자료를 그 대상으로 한다. 이와 같은 판별분류분석에 대한 비모수적 방법으로 k-Nearest Neighbors Classification(KNNC) 방법은 각 개체에 가장 가까운 k개의 이웃들의 소속집단 중에서 가장 빈도가 높은 집단으로 해당개체를 분류하는 방법이다. 이러한 KNNC는 다변량 정규성 등의 모수적 모형이 만족되지 않을 때에도 강건성(robustness)을 지니는 널리 활용되는 방법이다.

그런데 KNNC는 각각의 개체가 지니는 국소적 특징을 고려하지 않고, 자료 전체가 지니는 특징에 따라 고정된 이웃의 개수 k를 선정하므로, 각 개체가 지니는 국소적 특징을 간과할 수 있다. 따라서 자료의 국소적 특징을 고려하여 각 개체에 따라 소속집단 결정에 사용하는 이웃의 개수 k를 변화시키는 방법이 통계적 효율성을 더 높일 것으로 예상할 수 있다. Friedman (1994)은 직사각형 안에 포함된 이웃들의 바깥 부분을 제외시킴으로써 적응시켜 나가는 방법을 제안하였으며, Hastie와 Tibshrani (1996)는 KNNC 방법을 사용함에 있어서 개체가 속한 지역정보를 포함하는 효과적인 거리공간(metric space)을 찾기 위해 국소선형 판별분석(local linear discriminant analysis)을 활용하였다. Jhun 등 (2007)은 결측치 대체방법에 있어 k-최근접(k-nearest neighbors) 방법의 대안으로 자료의 국소적 특징을 고려한 적응 최근접(adaptive nearest neighbors) 방법을 제안하였다. 본 연구에서는 이와 유사한 방법을 사용한 분류방법으로 Adaptive Nearest Neighbors Classification(ANNC)을 제안하고 KNNC 분류방법과의 비교에 주된 관심을 두고자 한다.

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

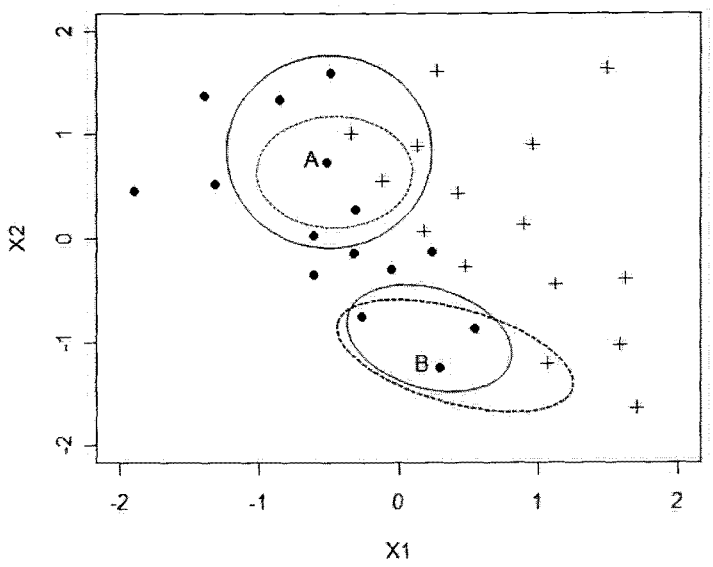

그림 **2.1.** 2차원 판별변수를 가지는 두 집단의 산점도

2. ANNC 방법

p개의 연속형 판별변수와 n개의 개체를 지닌 $n \times p$ 자료행렬 $X = (x_{ij}), i = 1, \ldots, n, j = 1, \ldots, p$가 주어졌다고 하자. x_{ij}는 i번째 개체의 j번째 변수에 대한 연속형 관찰값이다. 이제, n개의 개체를 $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$ (단, $\underline{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$)으로, 각 개체가 소속된 집단을 $G_k, k = 1, \ldots, g$로 표기하자. 여기서 개체들의 판별분류를 위하여 개체들 사이의 상사성을 측정하는데 이에 대한 척도로는 유클리드 거리, 마할라노비스 거리는 물론 상관계수 등의 다양한 방법들이 고려될 수 있다. 이제 i번째 개체 $\underline{x}_i$와 j번째 개체 $\underline{x}_j$사이의 거리를 d_{ij}로, i번째 개체 $\underline{x}_i$와 j번째로 가까운 개체 사이의 거리를 $d_{i(j)}$로 나타내자. KNNC 방법이 i번째 개체 $\underline{x}_i$와의 거리가 k번째로 가까운 개체와의 거리 $d_{i(k)}$보다 작거나 같은 개체들을 이웃으로 선택하는 반면, ANNC 방법은 각 i번째 개체 $\underline{x}_i$와 가장 가까운 개체와의 거리 $d_{i(1)}$에 대한 다른 개체들과의 거리 $d_{ij}(j \neq i)$의 비(ratio)를 이용하여 이웃집단을 구성하게 된다. 이 $d_{i(1)}$과 d_{ij}는 i번째 개체 $\underline{x}_i$가 속한 지역의 자료양상에 따라 달라지기 때문에 분류에 사용되는 이웃의 개수도 달라진다.

ANNC 방법을 이해하기 위하여 간단한 예를 들어보자. 그림 2.1은 2차원 판별변수를 갖는 표본크기가 각각 15개인 두 집단의 산점도이다. 이와 같은 자료에서 개체 A와 B에 대해 KNNC 방법을 사용함에 있어 최적 이웃의 개수 k가 3으로 정해졌다고 하자. 모든 개체가 고정된 이웃의 개수를 가지기 때문에 개체 A와 B 모두는 점선 내에 포함되는 가장 가까운 이웃 3개를 이용하여 판별분류를 수행하게 된다. 하지만 이렇게 고정된 개수의 이웃들을 사용하면 개체 A와 같이 밀집된 지역에 있는 개체는 이웃 개수가 너무 적을 경우 자료의 국소적 특징을 충분히 반영하지 못하게 되고 개체 B와 같이 밀집되지 않은 지역에 있는 개체는 이웃의 개수가 너무 많을 경우 지나치게 멀리 떨어져 있는 이웃의 부적절한 정보까지 사용하게 될 위험이 있다. 따라서 개체가 속한 지역의 밀집 정도에 따라 이웃의 개수를 유연하게 조절하는 방법(가령 실선에 포함된 이웃들을 사용하는 것)이 보다 타당해 보이며, 이를 수행하기 위한 ANNC

표 **2.1.** MBA 지원자들의 입학심사자료 및 결과

개체	GPA	GMAT	Status
1	2.96	596	1
2	3.01	453	3
...	...	...	...
...	...	...	...
83	3.24	467	1
84	3.03	414	3
85	2.54	446	2

알고리즘을 다음과 같이 제안한다.

2.1. ANNC 알고리즘

ANNC 방법은 다음과 같은 과정으로 이루어진다. 이는 Jhun 등 (2007)이 결측치 대체에서 제안한 방법과 매우 유사한 형태이다.

단계 1. 거리행렬 D를 정의한다.

$$D = (d_{ij}), \quad i, j = 1, \ldots, n.$$

단계 2. 조정계수(tuning parameter) δ를 이용해 거리행렬을 조정하여 새로운 거리행렬 D^{δ}를 만든다.

$$D^{\delta} = \left(d_{ij}^{\delta}\right) = (d_{ij} + \delta).$$

단계 3. 임계치(thresholding factor) q를 정하고 i번째 개체 $\underline{x}_i$와 가장 가까운 조정된 거리를 $d_{i(1)}^{\delta}$라고 하자. 그리고 $d_{i(1)}^{\delta}$와 조정된 거리 $d_{ij}^{\delta}(j \neq i)$와의 증분비가 q보다 작거나 같은 개체들로 이루어진 이웃집단 $N(i)$를 정한다.

$$N(i) = \left(\underline{x}_j : \frac{d_{ij}^{\delta}}{d_{i(1)}^{\delta}} \leq q\right).$$

단계 4. i번째 개체 $\underline{x}_i$를 $\underline{x}_i$의 이웃집단 $N(i)$에 속한 개체들에서 가장 빈번하게 관측되는 집단으로 분류한다.

제안된 ANNC 방법은 해당 개체로부터 가까운 개체들을 선택하는 과정에서 그 개체와의 거리 증분비가 임계치보다 작은 개체들만을 이웃으로 선택한다. 이렇게 함으로써 해당개체 주변의 밀도나 구조를 포함한 국소적 성격을 반영하는 것이다. 그런데 어떤 개체의 이웃들을 선택하고자 할 때, 해당 개체로부터 가장 가까운 이웃의 거리 $d_{i(1)}$가 0에 매우 가까우면 다음으로 가까운 개체의 거리 증분비 $d_{i(2)}/d_{i(1)}$는 무한대로 증가하므로 거리 d_{ij}들에 대하여 일종의 조정계수 δ를 단계 2에서 이용하였으며, 통상적으로 거리행렬의 중앙값을 사용한다. 다음으로 해당개체의 이웃 선택의 기준이 되는 조정된 거리 증분비 $d_{ij}^{\delta}/d_{i(1)}^{\delta}$의 최대 허용한계를 결정하는 임계치 q를 단계 3에서 고려하고, 증분비가 q이하인 개체들을 해당 개체의 이웃으로 선택한다. 따라서 q가 크면 많은 수의 개체들을 이웃을 선택하게 되고 q가 작으면 적은 수의 개체들을 이웃으로 선택하게 된다.

2.2. 사례분석

실제 사례를 이용하여 ANNC 과정을 살펴보자. 다음 표 2.1과 그림 2.2는 어느 대학 MBA 과정 지원자 85명의 입학심사자료(GPA, GMAT 점수)와 입학심사결과(Status) 자료이다 (Johnson과 Wichern,

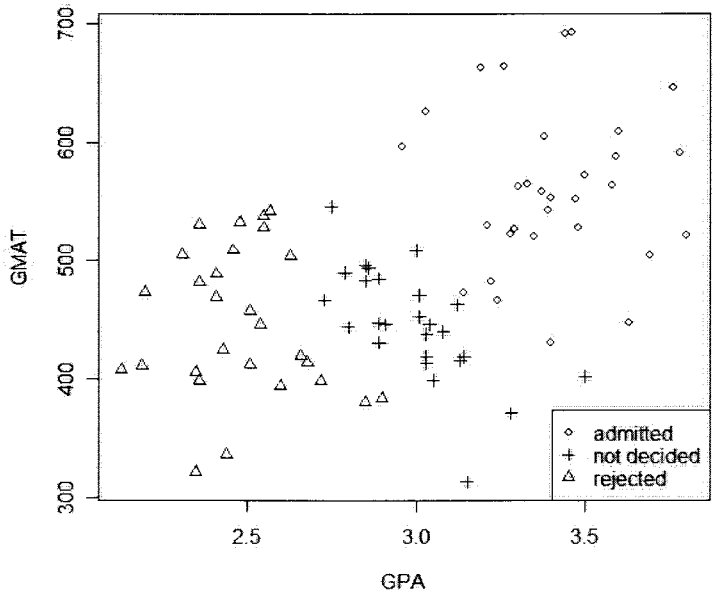

그림 **2.2.** MBA 지원자들의 입학심사 자료와 결과

2007, p.661). Status가 1이면 합격, 2이면 불합격 그리고 3이면 미정상태를 의미한다.

판별변수의 측정단위가 상이하기 때문에 각 변수를 그의 표준편차로 나누어 척도 변환 시킨 표준화 방법(standardized metric)을 이용하여 크기가 85×85인 거리행렬 $D = (d_{ij})$을 구하고 중위수 $\delta = 3.10$을 이용하여 조정된 거리행렬

$$D^{\delta} = (d_{ij} + 3.10)_{i,j=1,\ldots,85} = \begin{pmatrix} 3.10 & 5.22 & \cdots & 6.11 & 5.58 \\ 5.22 & 3.10 & \cdots & 4.33 & 3.85 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 3.58 & 3.85 & \cdots & 3.64 & 3.10 \end{pmatrix}$$

를 구하였다. 그리고 i번째 개체 $\underline{x}_i$에 대해 가장 가까운 개체와의 거리 $d^{\delta}_{i(1)}$와 다른 개체들 사이의 거리 $d^{\delta}_{ij}(j \neq i)$를 이용하여 증분비를 구한다.

$$Q = \left(\frac{d^{\delta}_{ij}}{d^{\delta}_{i(1)}} \right)_{i=1,\ldots,85, j \neq i} = \begin{pmatrix} - & 1.38 & \cdots & 1.61 & 1.47 \\ 1.57 & - & \cdots & 1.30 & 1.16 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1.71 & 1.18 & \cdots & 1.12 & - \end{pmatrix}$$

후보 임계치 q를 1에서 2까지 0.001간격으로 고려하고 각 고정된 임계치 값 q보다 증분비가 작거나 같은 개체들을 i번째 개체의 이웃집단으로 정한다. 예컨대 q가 1.150인 경우와 1.400인 경우 i번째 개체의 이웃집단 $N(i), i = 1, \ldots, 85$는 아래의 표 2.2와 같이 각각 구할 수 있다. 동일한 임계치를 사용하였지만 개체마다 이웃집단의 크기가 상이함을 알 수 있다. 한편 q가 커질수록 이웃으로 선택되는 개체수

표 **2.2.** 임계치에 따른 이웃집단

$N(i)$	q	
	1.150	1.400
$N(1)$	$\underline{x}_7$	$\underline{x}_7, \underline{x}_8, \cdots, \underline{x}_{79}, \underline{x}_{82}$
$N(2)$	$\underline{x}_1, \underline{x}_{30}, \underline{x}_{69}, \underline{x}_{75}, \underline{x}_{76}$	$\underline{x}_3, \underline{x}_4, \cdots, \underline{x}_{84}, \underline{x}_{85}$
$\vdots$	$\vdots$	$\vdots$
$N(85)$	$\underline{x}_{69}, \underline{x}_{70}, \cdots, \underline{x}_{83}, \underline{x}_{84}$	$\underline{x}_2, \underline{x}_3, \cdots, \underline{x}_{83}, \underline{x}_{84}$

표 **2.3.** ANNC와 KNNC의 판별분류분석 결과

		ANNC 판별결과			KNNC 판별결과		
		집단 1	집단 2	집단 3	집단 1	집단 2	집단 3
실제집단	집단 1	31	0	0	30	0	1
	집단 2	0	28	0	0	27	1
	집단 3	1	0	25	1	0	25

표 **2.4.** ANNC와 KNNC가 판별분석에 사용한 이웃의 개수

개체	ANNC		KNNC	
	이웃개수	분류결과	이웃개수	분류결과
$\underline{x}_2$	3	정분류	5	오분류
$\underline{x}_{28}$	5	정분류	5	정분류
$\underline{x}_{59}$	1	정분류	5	오분류
$\underline{x}_{66}$	2	오분류	5	오분류
$\underline{x}_{70}$	8	정분류	5	정분류

가 많아지기 때문에 가령 첫 번째 개체 $\underline{x}_1$의 경우 q가 1.150일 때는 한 개의 개체 $\underline{x}_7$만을 이웃으로 선택하였지만, q가 1.400일 때는 13개의 개체들 $\underline{x}_7, \underline{x}_8, \ldots, \underline{x}_{79}, \underline{x}_{82}$를 이웃으로 선택하였다.

각각의 개체는 그 이웃집단에 속한 개체들에서 가장 빈번히 관찰되는 집단으로 분류된다. 각 집단 별로 이웃의 수가 같은 경우 무작위 방법을 이용할 수 있다.

다음 표 2.3은 오분류율을 최소화하는 최적 임계치 1.125를 이용한 ANNC 방법의 판별분류결과와 같은 자료에 대하여 역시 오분류율을 최소화하는 최적의 이웃의 개수 5개를 이용한 KNNC 방법의 판별분류결과이다. 전체 85개의 개체 중 ANNC의 경우 1개의 개체가 오분류된 반면, KNNC 방법의 경우 총 3개의 개체가 오분류되었다. 표 2.4는 KNNC 방법에 의해 오분류된 3개의 개체 $\underline{x}_2$, $\underline{x}_{59}$, $\underline{x}_{66}$을 포함한 5개 개체들에 대해 ANNC 방법과 KNNC 방법이 사용한 이웃의 개수를 나타낸다. MBA 입학자료의 경우 ANNC 방법이 KNNC 방법보다 더 적은 수의 이웃들의 정보를 활용하면서 KNNC가 오분류한 개체들을 정분류함을 알 수 있다.

3. 모의실험

3.1. 실험모형

다음과 같은 다양한 4가지 모형 하에서 ANNC와 KNNC의 최적성능을 몬테칼로 모의실험을 통해 비교하였다. 독립반복시행횟수는 100회이며 각 집단의 표본크기는 모형에서 명시되어 있는 경우를 제외하고 모두 100으로 동일하다.

표 **3.1.** 모형 1~4로부터 얻은 오분류 개체 수

		ANNC	KNNC	KNNC − ANNC
모형 1	$d = 1.0$	56.93(0.612)	58.40(0.618)	1.47(0.185)
	$d = 1.5$	11.56(0.327)	12.24(0.340)	0.68(0.086)
모형 2	$s = 1.5$	31.84(0.523)	33.55(0.531)	1.71(0.157)
	$s = 3.0$	39.88(0.496)	41.55(0.511)	1.67(0.158)
모형 3	$p = 0.7$	48.26(0.470)	49.62(0.503)	1.36(0.196)
	$s = 0.9$	18.62(0.158)	19.49(0.121)	0.87(0.112)
모형 4	$g = 3.0$	93.71(0.783)	95.11(0.810)	1.40(0.216)
	$g = 4.0$	161.45(0.988)	162.71(1.003)	1.26(0.238)

모형 1. 집단 $1(G_1)$의 모평균 $\underline{\mu}_1 = (0,0)'$, 집단 $2(G_2)$의 모평균 $\underline{\mu}_2 = (d_1, d_2)'$으로 $\underline{\mu}_2$는 두 집단 간 거리가 $d = 1, 3$을 만족하도록 정해졌다. 두 집단 모두 동일한 모분산구조 $\Sigma = (\sigma_{ij}); \sigma_{ii} = 1, \sigma_{ij} = 0.25 (i \neq j)$를 갖는다.

$$G_1 : \underline{x} \sim N_2\left(\underline{\mu}_1, \Sigma\right), \qquad G_2 : \underline{x} \sim N_2\left(\underline{\mu}_2, \Sigma\right).$$

모형 2. 집단 $1(G_1)$의 모분산 Σ_1은 모형 1과 동일하며 집단 $2(G_2)$의 첫 번째 판별변수의 모분산 $\sigma_{11}^{(2)} = 1$이다. 두 번째 판별변수의 모분산 $\sigma_{22}^{(2)} = s = 1.5, 3.0$을 고려하였고, 집단 1의 모평균 $\underline{\mu}_1 = (0,0)'$, 집단 2의 모평균 $\underline{\mu}_2 = (0,2)'$이다.

$$G_1 : \underline{x} \sim N_2\left(\underline{\mu}_1, \Sigma_1\right), \qquad G_2 : \underline{x} \sim N_2\left(\underline{\mu}_2, \Sigma_2\right).$$

모형 3. 집단 $1(G_1)$과 집단 $2(G_2)$의 전체 표본크기는 200개로 고정되어 있고, 집단 1의 표본이 전체표본에서 차지하는 비율 $p = 0.7, 0.9$를 고려하였다. 집단 1의 모평균 $\underline{\mu}_1 = (0,0)'$, 집단 2의 모평균 $\underline{\mu}_2 = (0,2)'$이며 두 집단의 모분산 구조에 대한 가정은 모형 1과 동일하다.

$$G_1 : \underline{x} \sim N_2\left(\underline{\mu}_1, \Sigma\right), \qquad G_2 : \underline{x} \sim N_2\left(\underline{\mu}_2, \Sigma\right).$$

모형 4. 집단 수 $g = 3, 4$인 모형을 고려하였다. 집단 수 $g = 3$인 경우 모평균은 각각 $(0,0)'$, $(1,1)'$, $(0,2)'$이며 집단 수 $g = 4$인 경우 모평균은 $(0,0)'$, $(1,1)'$, $(0,2)'$, $(1,-1)'$이다. 모든 집단의 모분산 구조에 대한 가정은 모형 1과 같다.

$$G_1 : \underline{x} \sim N_2\left(\underline{\mu}_1, \Sigma\right), \ldots, G_g : \underline{x} \sim N_2\left(\underline{\mu}_g, \Sigma\right).$$

3.2. 모의실험 결과

표 3.1은 모형 1~4 하에서 얻은 ANNC 방법과 KNNC 방법의 오분류 개체수(number of misclassification: NOM)를 평균한 것으로 괄호 안은 이 평균에 대한 표준오차를 나타낸다. 독립적 모의시행에서 생성된 자료로부터 발생하는 변동요인을 제어하기 위해 각각의 모의시행에서 KNNC 방법에 의한 오분류 개체수와 ANNC 방법에 의한 오분류 개체수의 차이(KNNC − ANNC)를 구하였다. 모든 모형 하에서, 평균적으로 ANNC 방법에 의한 오분류 개체수가 KNNC 방법에 의한 오분류 개체수 보다 적은 것을 알 수 있다.

모든 개체가 고정된 개수의 이웃을 사용하는 KNNC 방법에 비해 ANNC 방법은 유연하게 이웃의 개수를 조정하며, 특히 자료가 밀집된 지역에서는 이웃의 개수를 크게 택하고 밀집되지 않은 지역에서는

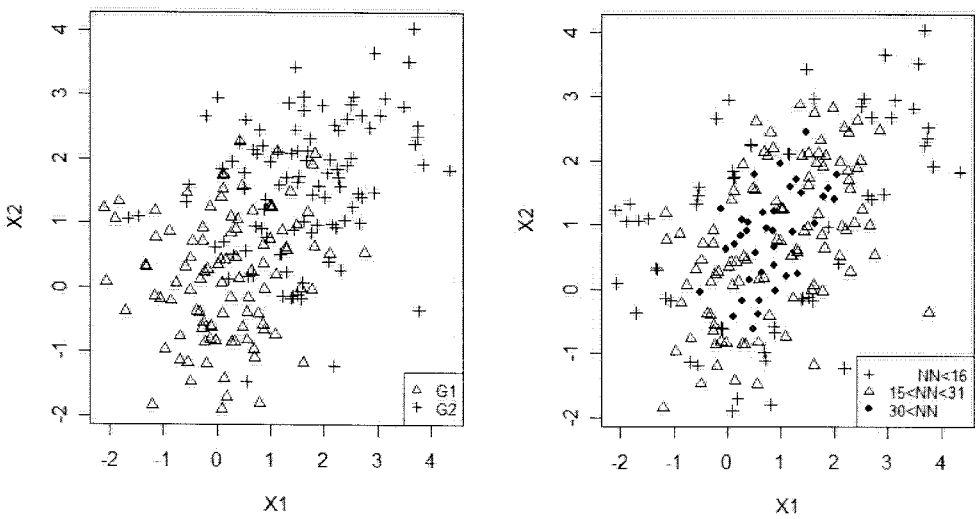

그림 **3.1.** 원자료와 각 개체별로 ANNC가 사용한 이웃개수(NN)의 분포

표 **3.2.** 모형 2에서 분산 s에 따른 집단 별 오분류(단위:%)

s	집단 1			집단 2		
	ANNC	KNNC	KNNC − ANNC	ANNC	KNNC	KNNC − ANNC
1.5	12.69	13.84	1.15	19.15	19.71	0.56
	(0.386)	(0.374)	(0.211)	(0.414)	(0.443)	(0.224)
3.0	9.68	11.50	1.82	30.20	30.05	−0.15
	(0.388)	(0.389)	(0.281)	(0.515)	(0.430)	(0.298)

적게 택하는 특징을 가지고 있다. 다음 그림 3.1의 왼쪽 그림은 모형 1 하에서 모의실험을 통해 생성된 자료 중 하나이다. 여기서 개체들이 가장 밀집되는 지역은 두 집단이 겹쳐 있는 부분이 된다. 그림 3.1의 오른쪽 그림은 이 자료에 ANNC 방법을 사용했을 때 각 개체들이 사용한 이웃의 개수(number of neighbors: NN)의 분포 형태를 보여준다. 가장 밀집된 지역에 속한 개체들은 30개 이상의 개체를 이웃으로 선택하였고, 밀집되지 않은 지역에 속한 개체들은 상대적으로 적은 수의 개체를 이웃으로 선택하였다. 동일한 자료에서 KNNC 방법이 최적 이웃의 개수 $k = 25$를 모든 개체에 동일하게 적용한 것에 비교해봤을 때, ANNC 방법이 개별 개체가 속한 지역의 자료구조를 보다 효과적으로 활용함을 확인할 수 있다.

표 3.2는 표 3.1의 모형 2의 오분류 개체수를 집단 별 오분류율로 재표현한 것이다. 집단 1의 두 판별변수 $x_1^{(1)}, x_2^{(1)}$와 집단 2의 첫 번째 판별변수 $x_1^{(2)}$의 분산은 모두 1이며, 집단 2의 두 번째 판별변수 $x_2^{(2)}$의 분산은 s로 주어졌다. 두 번째 판별변수의 분산 s는 1보다 크기 때문에 이는 두 집단이 겹쳐지는 부분에서 집단 1의 개체들이 상대적으로 더 밀집된다는 것을 의미하게 된다. 이제 표 3.2에서 집단 별 오분류율을 살펴보면, 두 번째 판별변수의 분산 s가 3.0일 때 집단 2의 ANNC와 KNNC의 매우 작은 차이 -0.15를 제외하면 대부분 ANNC 방법이 KNNC 방법보다 낮은 오분류율을 보이고 있다. 특히 분산 s가 커질수록 집단 1의 ANNC와 KNNC의 오분류율의 차이가 더 커짐을 알 수 있다. 이것은 두 집단의 개체들이 혼재되어 밀집된 지역에서 ANNC가 많은 수의 이웃을 사용하여 오분류율을 낮추고 있음을 보여준다.

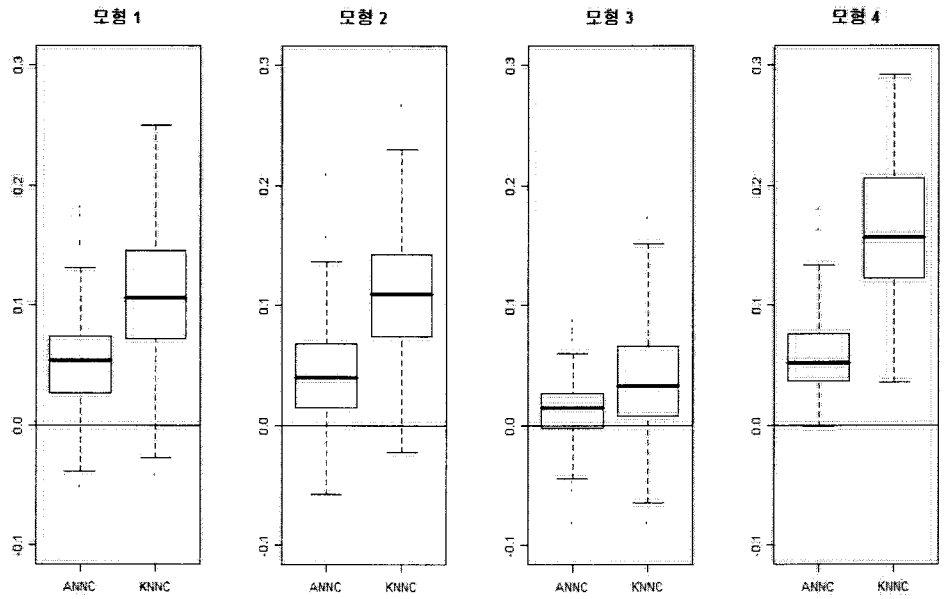

그림 **3.2.** 정분류된 개체들과 오분류된 개체들의 평균 이웃거리 차이

다음의 그림 3.2는 모의실험 과정에서 구한 정분류된 개체들의 평균 이웃거리에서 오분류된 개체들의 평균 이웃거리를 뺀 결과를 고려된 모든 모형에 대해서 박스-플롯으로 표현한 것이다. 이 값이 클수록 오분류된 개체들의 이웃거리가 정분류된 개체들의 이웃거리보다 평균적으로 짧다는 것을 의미한다. 그림 3.2에서, KNNC 방법의 평균 이웃거리의 차이가 ANNC 방법의 그것보다 큰 것을 확인할 수 있다. 즉, ANNC 방법의 경우 오분류되는 개체나 정분류되는 개체나 평균 이웃거리의 차이가 KNNC의 그것보다 크지 않으며, KNNC 방법의 경우 오분류되는 개체들의 평균 이웃거리가 정분류되는 개체들의 평균 이웃거리보다 ANNC에 비해 상대적으로 더 짧다는 것을 나타낸다. 밀집된 지역에 속한 개체들의 평균 이웃거리가 대체로 짧은 것을 감안하면 KNNC 방법이 이웃의 수를 모든 개체에 동일하게 적용함으로써 개체가 속한 지역의 특성을 ANNC 방법에 비해 충분히 반영하지 못함을 알 수 있다.

4. 결론

비모수적 판별분류방법으로 널리 사용되는 k-Nearest Neighbors Classification(KNNC) 방법이 전체 자료에 대해 고정된 이웃의 개수 k를 사용함으로서 발생하는 단점을 보완하기 위한 방법으로 Adaptive Nearest Neighbors Classificaion(ANNC) 방법을 제안하였다. 제안된 ANNC 방법은 분류를 위한 이웃의 선택에 있어서 자료의 국소적 구조나 밀집도 등을 반영하여 선택되는 이웃의 개수가 변화하며 결과적으로 오분류율을 감소시키는 것을 보일 수 있었다. 사례분석과 다양한 모형에 대한 모의실험을 통해 제안된 ANNC의 통계적 성질을 규명하였으며 나아가 기존의 KNNC의 대안이 될 수 있음을 보였다고 여겨진다.

참고문헌

Friedman, J. (1994). *Flexible metric nearest-neighbor classification*, Technical report, Standford University.

Hastie, T. and Tibshrani, 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.

Johnson, R. A. and Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*, Prentice Hall, New York.

Adaptive Nearest Neighbors for Classification

Myoungshic Jhun[1] · Inkyung Choi[2]

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

(Received February 2009; accepted March 2009)

Abstract

The k-Nearest Neighbors Classification(KNNC) is a popular non-parametric classification method which assigns a fixed number k of neighbors to every observation without consideration of the local feature of the each observation. In this paper, we propose an Adaptive Nearest Neighbors Classification(ANNC) as an alternative to KNNC. The proposed ANNC method adapts the number of neighbors according to the local feature of the observation such as density of data. To verify characteristics of ANNC, we compare the number of misclassified observation with KNNC by Monte Carlo study and confirm the potential performance of ANNC method.

Keywords: Adpative nearest neighbors, classification analysis, k-nearest neighbors.

[1]Corresponding author: Professor, Department of Statistics, Korea University, Seoul 136-701, Korea. E-mail: jhun@korea.ac.kr