응용통계연구 (2011)
24(6), 1249–1257
DOI: http://dx.doi.org/10.5351/KJAS.2011.24.6.1249

순차 적응 최근접 이웃을 활용한 결측값 대치법

박소현[1] · 방성완[2] · 전명식[3]

[1]고려대학교 통계학과, [2]육군사관학교 수학과, [3]고려대학교 통계학과

(2011년 9월 접수, 2011년 10월 채택)

요 약

비모수적 결측치 대치법인 k-최근접 이웃(k-Nearest Neighbors; KNN) 대치법을 개선한 적응 최근접 이웃(Adaptive Nearest Neighbor; ANN) 대치법과 순차 k-최근접 이웃(Sequential k-Nearest Neighbor; SKNN) 대치법의 장점들을 결합한 순차 적응 최근접 이웃(Sequential Adaptive Nearest Neighbor; SANN) 대치법을 제안하고자 한다. 이 방법은 ANN 대치법의 장점인 자료의 국소적 특징을 반영할 뿐 아니라, SKNN 대치법과 같이 결측값 대치가 이루어진 개체를 다음 결측값을 대치할 때 사용함으로써 효율성에 개선이 있을 것으로 기대한다.

주요용어: 적응 최근접 이웃, k-최근접 이웃, 순차 적응 최근접 이웃, 결측 자료, 대치법.

1. 서론

관측자료나 실험자료의 수집에서 무응답이나 실험의 오류 등으로 결측값이 발생하는 것은 매우 통상적이며, 이러한 결측값은 자료의 통계적 분석을 어렵게 할 뿐 아니라 그 분석 결과에도 크게 영향을 미친다. 통계적 분석과정에서 결측값의 처리에 관하여 많은 방법론이 연구되어 왔으며 (Little과 Rubin, 1987; 이진희 등, 2006; 이상은과 신기일, 2010), 그 중 대치법(imputation)은 가장 널리 사용되고 있는 방법 중 하나이다. 여러 다양한 종류의 대치법들 중에서 Dixon (1979)과 Troyanskaya 등 (2001)에 의해 제안된 k-최근접 이웃(k-Nearest Neighbors; KNN) 대치법은 결측이 발생한 개체와 가장 가까운 거리에 있는 k개의 이웃 개체들을 활용하여 결측값을 대치하는 비모수적 방법으로, 다변량 정규성 등의 모수적 모형이 만족되지 않을 때에도 강건성(robustness)을 지니며 그 계산 알고리즘이 간단하다는 장점을 바탕으로 널리 활용되고 있다. 그러나 KNN 대치법은 모든 개체에 대해 고정된 k개의 이웃을 사용하기 때문에 결측이 발생한 개체의 위치에 따른 국소적 특징을 간과할 수 있다.

이러한 KNN 대치법의 단점을 보완하기 위해 Jhun 등 (2007)은 자료의 국소적 특징을 고려하여 이웃의 개수 k를 변화시키는 적응 최근접 이웃(Adaptive Nearest Neighbors; ANN) 대치법을 제안하였다. Jhun 등 (2007)이 제시한 이웃의 개수를 적응적으로 부여하는 방안은 판별분류방법에도 적용되어 연구되었다 (전명식과 최인경, 2009; 맹진우 등, 2010). 한편, KNN 대치법은 결측값이 발생하지 않은 완전한 개체들만을 이용해서 이웃집단을 선정하기 때문에 결측값이 하나라도 발생한 개체가 지닌 관측된 변수들에 대한 정보는 이용할 수 없다. 이에, Kim 등 (2004)은 KNN 대치법의 대안으로 결측이 발생한 개체 내의 관측된 변수들의 값을 순차적으로 활용하는 순차 k-최근접 이웃(Sequential k-Nearest Neighbors; SKNN) 대치법을 제안하였다.

본 논문에서는 KNN 대치법의 단점을 보완한 ANN 대치법과 SKNN 대치법의 장점을 결합한 순차 적응 최근접 이웃(Sequential Adaptive Nearest Neighbors; SANN) 대치법을 제안하고자 한다. 제안된

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

SANN 대치법은 각 개체가 지니는 자료의 국소적 특징을 반영하여 개체에 따라 결측값 추정에 이용하는 이웃의 개수 k를 변화시킨다는 점과 결측값이 대치된 개체를 다음 대치에서 사용한다는 두 가지 장점을 모두 지니게 되어 보다 효율적으로 결측값을 대치할 수 있을 것으로 기대한다. 본 논문의 2장에서는 SANN 대치법의 알고리즘을 소개하고, 단순예제를 이용하여 기존의 ANN 대치법과의 차이점을 설명하였다. 3장에서는 모형 및 실제자료를 이용한 모의실험을 통해 기존의 KNN, SKNN, ANN 대치법과 제안된 SANN 대치법을 다양한 설정에서 비교하였으며 SANN 대치법의 활용 가능성을 보였다.

2. SANN 대치법

p개의 변수와 n개의 개체로 이루어진 $n \times p$ 자료행렬 $\boldsymbol{X} = (x_{ij})$와 결측값의 지시행렬 $\boldsymbol{R} = (r_{ij})$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, p$가 주어졌다고 하자. 여기서, x_{ij}가 관측 됐다면 $r_{ij} = 1$, 결측이라면 $r_{ij} = 0$으로 나타낸다. 이제, 자료행렬 $\boldsymbol{X}$의 행으로 표현되는 n개의 개체를 $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ (단, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$)라 하고 결측값이 있는 행들은 목표개체, 결측값이 없는 행들은 후보개체라고 한다. 또한 목표개체들의 집단을 M, 후보개체들의 집단을 C로 구분하여 표기하며, 집단 M과 집단 C에 속한 개체의 수를 각각 m, $n - m$이라고 하자. 또한 개체 $\boldsymbol{x}_i$, $1 \le i \le n$에서 결측이 발생한 변수를

$$V_i = \{j : x_{ij}\text{가 결측},\ 1 \le j \le p\}$$

로 정의한다. 이제 $\boldsymbol{x}_i$와 $\boldsymbol{x}_j$사이의 유클리디안 거리를 d_{ij}라고 표기하고, $\boldsymbol{x}_i$로부터 k번째로 가까운 개체와의 거리를 $d_{i(k)}$로 표기하자.

KNN 대치법이 자료의 국소적 특징을 반영하지 못한다는 한계를 보완하기 위해 Jhun 등 (2007)은 결측된 개체마다 주변의 밀집도 등을 반영하여 이웃의 개수를 다르게 선정하는 ANN 대치법을 제안하였다. 그러나 KNN과 ANN 대치법은 집단 M에 속한 개체들의 관측값들을 대치과정에 이용하지 못한다는 제약이 있다. 이를 보완할 수 있는 방법으로 Kim 등 (2004)은 결측값 대치가 완료된 개체를 다음 결측값을 대치할 때 이웃의 후보로 사용하는 SKNN 대치법을 제안하였다. 본 논문에서는 KNN 대치법을 보완한 ANN과 SKNN 대치법의 장점을 결합한 Sequential ANN(SANN) 대치법을 제안하고자 한다. 이 방법은 ANN 대치법이 목표개체가 속한 지역의 국소적 특징을 고려해서 이웃의 개수를 유연하게 조정하는 장점과 SKNN 대치법의 장점인 결측값이 대치된 개체를 그 다음 대치과정에서 후보집단으로 이용하는 효율성을 함께 지니게 된다.

SANN 대치법의 이해를 돕기 위하여 간단한 예를 들어보자. 크기가 12×4인 자료행렬을 고려하였으며, 모든 결측값은 두 변수 X_3과 X_4에서 발생시켰다. 그림 2.1은 결측이 발생하지 않은 변수 X_1과 X_2의 평면상에 12개의 개체를 나타낸 것으로, 목표개체 A, B, C, D 옆의 숫자는 결측이 발생한 변수의 수를 의미한다. 이제 목표개체 A와 C에서 결측된 두 변수 X_3과 X_4의 값을 대치하고자 하면, ANN 대치법은 점선 안에 있는 두 개의 이웃을 이용해서 결측값을 추정한다. 반면에, SANN 대치법은 결측 변수의 수가 적은 개체부터 대치를 함으로서, 이미 결측값 대치가 완료된 개체 B와 D도 이웃으로 사용할 수 있어 실선 안에 있는 세 개의 개체를 이용하여 결측값을 대치하게 된다. 이 방법은 이웃에 있는 개체 수가 제한적인 상황에서 먼저 결측값이 대치된 개체를 재활용함으로써 자료를 더 효과적으로 이용할 수 있다는 장점이 있다.

2.1. SANN 대치법 알고리즘

본 논문에서는 SANN 대치법을 수행하기 위한 알고리즘을 다음과 같이 제안한다. 이는 Jhun 등 (2007)에서 제안된 ANN 대치법과 매우 유사한 형태이다.

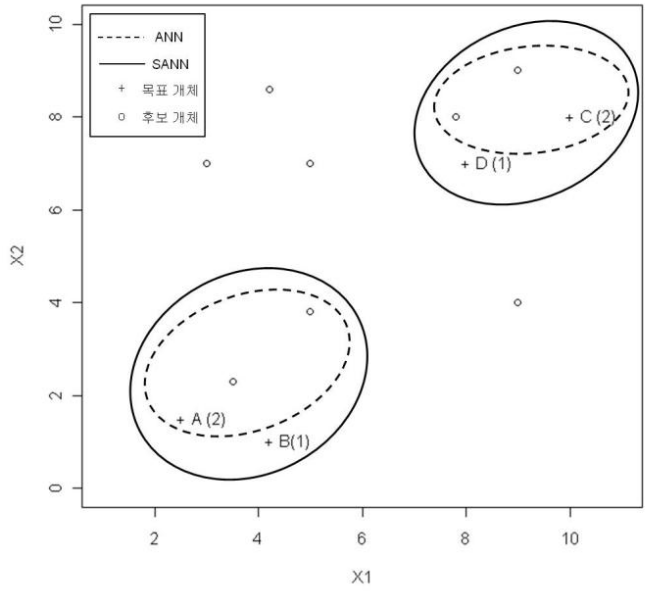

그림 **2.1.** ANN 대치법과 SANN 대치법 비교

단계 1. 집단 M에 속한 목표개체들을 결측이 발생한 변수의 수가 작은 순서대로 정렬한다.

단계 2. 단계 1에서 정렬된 순서에 따라 해당 목표개체의 결측값을 아래와 같이 대치한다.

(2-1) 목표개체 $\boldsymbol{x}_a \in M$와 후보개체 $\boldsymbol{x}_b \in C$에 대해서 $(n-m)\times m$ 가중 유클리디안 거리행렬 $D=(d_{ab})$를 정의한다.

$$d_{ab}=\left\{n_a^{-1}\sum_{j=1}^{p} r_{aj}(x_{aj}-x_{bj})^2\right\}^{\frac{1}{2}} \quad \text{그리고} \quad n_a=\sum_{j=1}^{p} r_{aj}.$$

(2-2) 거리행렬 D의 중앙값을 조정계수(tuning parameter) δ로 설정하고 수정된 거리행렬 D^*를 구한다.

$$D^*=(d_{ab}^*)=(d_{ab}+\delta).$$

(2-3) 임계치(thresholding factor) q를 정하고 목표개체 $\boldsymbol{x}_a \in M$에 대한 이웃집단을

$$\eta_a=\left\{\boldsymbol{x}_b \in C : \frac{d_{ab}^*}{d_{a(1)}^*} \le q\right\}, \quad q \ge 1$$

로 정의한다.

(2-4) 목표개체 $\boldsymbol{x}_a \in M$의 결측 변수 $x_{aj}(j \in V_a)$를

$$\hat{x}_{aj}=\sum_{b\in\eta_a} w_{ab}x_{bj}, \quad \text{단} \quad w_{ab}=\frac{1}{d_{ab}S_a},\ S_a=\sum_{b\in\eta_a} d_{ab}^{-1}$$

으로 추정하여 대체한다.

표 **2.1.** SANN 대치법의 과정

	i	변수				수정된 거리				가중치			
		X_1	X_2	X_3	X_4	d^*_{12b}	d^*_{13b}	d^*_{14b}	d^*_{15b}	w_{12b}	w_{13b}	w_{14b}	w_{15b}
	1	2	2	7	8	5.39	4.51	4.31	5.39	1.00			
	2	3	4	8	7	5.83	4.70	4.61	5.80				
	3	4	7	4	2	7.27	6.64	5.19	5.97				
	4	5	6	5	1	7.80	7.26	5.84	7.92				
	5	5	8	2	2	8.19	7.39	6.43	7.99				
C	6	6	7	6	3	8.76	7.39	6.61	8.63				
	7	7	6	3	4	9.07	7.82	6.77	8.69				
	8	9	7	5	4	9.10	8.21	7.14	8.86				
	9	8	8	3	3	9.30	8.24	7.21	8.91				
	10	9	4	9	8	9.46	9.14	7.21	8.91		0.55		
	11	10	4	9	8	9.52	9.17	7.91	9.53		0.45		
	12	4	2	?	9	·	9.17	8.13	9.91			1.00	
M	13	9	3	?	9	·	·	8.13	9.91				1.00
	14	4	3	?	?	·	·	·	10.04				
	15	10	2	?	?	·	·	·	·				
						δ: 4.12	3.70	3.61	4.39				

단계 3. 목표개체 $\boldsymbol{x}_a$의 결측값을 추정값으로 대체한 후에는 $\hat{\boldsymbol{x}}_a$을 집단 C에 포함시키고 단계 2를 반복한다.

제안된 SANN 대치법에서 이웃을 선택하는 기준이 되는 임계치 q는 다음 같은 의미로 해석될 수 있다. 목표개체 $\boldsymbol{x}_a \in M$와 후보개체 $\boldsymbol{x}_b \in C$ 사이의 거리를 d_{ab}라고 할 때, q는 두 거리의 비 $d_{ab}/d_{a(1)}$의 최대 허용한계치가 된다. 즉, 거리의 비가 q보다 작거나 같은 개체들을 후보개체 집단 C에서 선택하여 이웃 집단을 구성하는 것이다. 그렇게 함으로써 이웃을 선택할 때 목표개체의 국소적 특징을 고려하게 된다. 하지만 거리 비의 분모에 해당하는 목표개체와 가장 근접한 개체의 거리가 0에 가까워지면 거리 비가 무한히 커지게 되므로 어떤 개체도 이웃으로 선택될 수 없다. 이런 상황을 막기 위해 조정계수(δ)를 더한 수정된 거리(d^*_{ab})를 단계 (2-3)에서 이용했으며, 조정계수는 통상적으로 거리행렬의 중앙값을 사용한다.

2.2. 단순예제를 통한 SANN과 ANN 대치법의 비교

SANN 대치법과 ANN 대치법이 어떤 과정을 거쳐서 결측값을 추정하는지를 크기가 15×4인 자료행렬을 이용하여 살펴보았으며, 그 결과는 표 2.1과 표 2.2에 각각 나타나 있다. 결측된 변수가 하나인 목표개체 $\boldsymbol{x}_{12}$의 결측값부터 추정한다고 하자. 먼저, 목표개체와 후보 개체들간의 거리 d_{ab}를 구하고 조정계수 δ를 사용해서 수정된 거리(d^*_{ab})를 계산한다. 각 목표개체에 대한 δ는 거리행렬 D의 중위수를 이용하였으며, SANN 대치법에서는 매 추정 시 그 값이 변하는 반면, ANN 대치법에서는 항상 동일한 값을 사용한다. 따라서 이 예제에서 SANN 대치법의 첫 번째 목표개체 $\boldsymbol{x}_{12}$의 추정에 사용된 δ는 11×4인 거리행렬 원소들의 중위수이고, 두 번째 목표개체 $\boldsymbol{x}_{13}$의 추정에 사용된 δ는 12×3 거리행렬 원소들의 중위수가 된다. 거리 비(ratio)의 임계치를 $q = 1.05$로 설정한 결과, 각 목표개체에 대한 이웃은 SANN 대치법에서는 $\eta_{12} = \{\boldsymbol{x}_1\}$, $\eta_{13} = \{\boldsymbol{x}_{10}, \boldsymbol{x}_{11}\}$, $\eta_{14} = \{\boldsymbol{x}_{12}\}$, $\eta_{15} = \{\boldsymbol{x}_{13}\}$이, ANN 대치법에서는 $\eta_{12} = \{\boldsymbol{x}_1\}$, $\eta_{13} = \{\boldsymbol{x}_{10}, \boldsymbol{x}_{11}\}$, $\eta_{14} = \{\boldsymbol{x}_2\}$, $\eta_{15} = \{\boldsymbol{x}_{10}, \boldsymbol{x}_{11}\}$으로 선택되었다. 또한, 가중치는 각각 선택된 이웃들에 의해 알고리즘의 단계 (2-4)의 식에 의해 계산되었다. 그 결과 SANN 대치법에서는 결측값이 $\hat{x}_{12,3} = 7$, $\hat{x}_{13,3} = 9$, $\hat{x}_{14,3} = 7$, $\hat{x}_{14,4} = 9$, $\hat{x}_{15,3} = 9$, $\hat{x}_{15,4} = 9$로, 반면에 ANN 대치법

표 **2.2.** ANN 대치법의 과정

	i	변수				수정된 거리				가중치			
		X_1	X_2	X_3	X_4	d^*_{12b}	d^*_{13b}	d^*_{14b}	d^*_{15b}	w_{12b}	w_{13b}	w_{14b}	w_{15b}
	1	2	2	7	8	5.39	4.92	5.10	5.52	1.00			
	2	3	4	8	7	5.83	5.10	5.68	5.68			1.00	
	3	4	7	4	2	7.27	7.66	6.34	7.64				
	4	5	6	5	1	7.80	7.80	6.93	7.71				
	5	5	8	2	2	8.19	7.80	7.10	8.57				
C	6	6	7	6	3	8.76	8.23	7.27	8.63				
	7	7	6	3	4	9.07	8.61	7.71	8.63				
	8	9	7	5	4	9.10	8.65	7.71	9.25				
	9	8	8	3	3	9.30	9.55	8.40	9.63				
	10	9	4	9	8	9.46	9.58	8.63	9.63		0.55		0.47
	11	10	4	9	8	9.52	9.58	8.63	9.76		0.45		0.53
	12	4	2	?	9	·	·	·	·				
M	13	9	3	?	9	·	·	·	·				
	14	4	3	?	?	·	·	·	·				
	15	10	2	?	?	·	·	·	·				
					δ:	4.12	4.12	4.12	4.12				

표 **2.3.** SANN 대치법과 ANN 대치법의 비교

목표개체	SANN 대치법			ANN 대치법		
	선택된 이웃	d	평균절대오차	선택된 이웃	d	평균절대오차
$\boldsymbol{x}_{14}$	12	1	1.5	2	1.41	2
$\boldsymbol{x}_{15}$	13	1.41	0.5	10, 11	2.11	1

d: 목표개체와 선택된 이웃사이의 거리 평균

에서는 $\hat{x}_{12,3} = 7$, $\hat{x}_{13,3} = 9$, $\hat{x}_{14,3} = 8$, $\hat{x}_{14,4} = 7$, $\hat{x}_{15,3} = 9$, $\hat{x}_{15,4} = 8$로 추정되었다. 두 방법의 차이는 $\boldsymbol{x}_{14}$와 $\boldsymbol{x}_{15}$의 결측값을 대치하는 과정에서 확연히 드러난다. ANN 대치법은 C 집단의 개체들만을 이웃으로 이용한 반면, SANN 대치법에서는 C 집단의 개체들과 더불어 M 집단이었던 $\boldsymbol{x}_{12}$와 $\boldsymbol{x}_{13}$을 활용하였다. 이러한 결과, 목표개체와 추정에 사용되는 이웃 사이의 거리가 ANN 대치법보다 SANN 대치법을 사용했을 때 더 가깝다는 것을 표 2.3에서 확인할 수 있다. 또한, 평균절대오차(mean absolute error)가 SANN 대치법을 이용했을 때 더 작은 것을 보면, 보다 가까운 개체를 이용해서 결측값을 대치한 SANN 대치법이 결측값을 더 정확하게 추정했음을 알 수 있다.

3. 모의실험

제안된 SANN 대치법과 기존의 세 가지 대치법(KNN, ANN, SKNN)의 성능을 모의실험을 통해 비교하였다. 모의실험에서는 Jhun 등 (2007)에서 사용된 3가지 모형으로부터 생성된 자료와 미국 국립암연구소(The National Cancer Institute)의 anticancer 프로젝트로부터 제공된 NCI60 자료(http://genome-www.stanford.edu/nci60)를 사용하였다.

3.1. 모형을 이용한 모의실험

방법론들의 성능을 비교하기 위해 먼저 아래 3개의 모형들로부터 크기가 $n = 500$인 자료를 생성하였다. 이 때 변수의 차원은 $p = 4$와 8을 고려하였다. 자료의 중심을 $\boldsymbol{x}_c$라고 하면, 개체 $\boldsymbol{x}_i$가 결측될 확률

표 **3.1.** 100회 독립시행에 근거한 NRMSE의 최소값들의 평균 (모형 1, 2, 3)

모형	결측비율	p	KNN	SKNN	ANN	SANN
1	10%	4	0.6624 (0.0482)	0.6617 (0.0499)	0.6531 (0.0508)	0.6520 (0.0492)
		8	0.6167 (0.0499)	0.6126 (0.0508)	0.6123 (0.0489)	0.6086 (0.0499)
	30%	4	0.6900 (0.0448)	0.6817 (0.0458)	0.6851 (0.0482)	0.6749 (0.0488)
		8	0.6334 (0.0436)	0.6262 (0.0447)	0.6299 (0.0487)	0.6210 (0.0499)
2	10%	4	0.8467 (0.0489)	0.8458 (0.0498)	0.8383 (0.0481)	0.8356 (0.0502)
		8	0.7879 (0.0482)	0.7844 (0.0499)	0.7826 (0.0487)	0.7786 (0.0499)
	30%	4	0.8854 (0.0481)	0.8720 (0.0489)	0.8730 (0.0447)	0.8615 (0.0499)
		8	0.8236 (0.0487)	0.8085 (0.0489)	0.8140 (0.0508)	0.8011 (0.0489)
3	10%	4	0.6827 (0.0433)	0.6785 (0.0467)	0.6748 (0.0484)	0.6704 (0.0454)
		8	0.6330 (0.0466)	0.6276 (0.0478)	0.6183 (0.0495)	0.6143 (0.0506)
	30%	4	0.7339 (0.0438)	0.7223 (0.0478)	0.7345 (0.0506)	0.7138 (0.0506)
		8	0.6720 (0.0382)	0.6523 (0.0429)	0.6620 (0.0442)	0.6451 (0.0478)

괄호안의 값은 표준오차임.

은 $d_{ic}/\sum_{i=1}^{n} d_{ic}$에 비례하도록 함으로써, 자료의 중심으로부터 멀어질수록 더 많은 결측값이 발생하도록 하였다. 이는 자료의 밀도가 낮은 주변부에 더 많은 결측값이 발생하도록 설계한 것으로, 전체 결측값의 비율은 10%와 30%를 고려하였다.

모형 1. $(x_1, \ldots, x_p)^T \sim N_p(\boldsymbol{\mu}, \Sigma)$. 여기서 $\boldsymbol{\mu} = (0, \ldots, 0)^T$이고 $\Sigma_{jj} = \mathrm{Var}(x_j) = 0.5$, $j = 1, \ldots, p$, $\Sigma_{jk} = \mathrm{Cov}(x_j, x_k) = 0.3, 1 \le j < k \le p$이다.

모형 2. $(x_1, \ldots, x_p)^T \sim \sum_{g=1}^{5} p_g N_p(\boldsymbol{\mu}_g, I)$. 여기서, $\boldsymbol{\mu}_g = (\mu_g, \ldots, \mu_g)^T$, $g = 1, \ldots, 5$이고 $(\mu_1, \ldots, \mu_5) = (0, 0.5, -0.5, 2.0, -2.0)$, $(p_1, \ldots, p_5) = (0.7, 0.1, 0.1, 0.05, 0.05)$이다.

모형 3. $(x_1, \ldots, x_p)^T \sim t_p(\nu, \boldsymbol{\mu}, \Sigma)$. 여기서 $\nu = 5$, $\boldsymbol{\mu} = (0, \ldots, 0)^T$이고 $\Sigma_{jj} = \mathrm{Var}(x_j) = 0.5$, $j = 1, \ldots, p$, $\Sigma_{jk} = \mathrm{Cov}(x_j, x_k) = 0.3$, $1 \le j < k \le p$이다.

모형 1은 모든 두 변수들 간의 상관계수가 0.6인 다변량 정규분포의 경우이며, 모형 2는 무상관 관계를 갖는 다변량 정규분포의 혼합모형으로 개체들이 중심부에 밀집되어 있고 주변부에는 산재되어 있어 자료가 국소적인 특징을 가지도록 설계된 것이다. 모형 3은 모평균과 모분산 구조가 모형 1과 동일한 자유도가 3인 다변량 t분포로, 분포형태는 모형 1과 흡사하지만 자료의 주변부에 흩어진 개체가 많도록 설계되었다.

결측값의 추정 후에는 정규화 제곱근 평균제곱오차(Normalized Root Mean Squared Error; NRMSE)를 통해서 각 방법들의 성능을 비교하였다.

$$\mathrm{NRMSE} = \frac{1}{\mathrm{SD}(x_{ij})} \left\{ \sum_{i \in M} \sum_{j \in V_i} \frac{(\hat{x_{ij}} - x_{ij})^2}{N} \right\}^{\frac{1}{2}},$$

여기서 $\mathrm{SD}(x_{ij})$는 실제 값들의 표준편차이고 N은 추정된 결측값들의 개수이다. 만약 추정된 값이 실제값과 유사하다면 NRMSE는 0에 가까울 것이고, 그렇지 않으면 1에 가까울 것이다. KNN 대치법과 SKNN 대치법의 조정모수 k는 1부터 50까지 1씩 증가시켰으며, ANN과 SANN 대치법의 임계치 q는 1부터 2까지 0.005씩 등간격으로 선택하였다. 이웃의 수 k와 임계치 q의 척도가 다르기 때문에 본 논문에서는 가능한 k와 q값들에 대한 NRMSE의 최소값을 사용하여 대치법들의 성능을 비교하였다. 각각의 모형에 대하여 이와 같은 과정을 독립적으로 100회 반복하였으며, 모의실험의 결과는 표 3.1에 주어져 있다.

표 **3.2.** 100회 독립시행에 근거한 NRMSE의 최소값들의 평균 (NCI60 자료)

결측비율	KNN	SKNN	ANN	SANN
10%	0.7422 (0.0294)	0.7393 (0.0343)	0.7399 (0.0348)	0.7308 (0.0368)
30%	0.7579 (0.0265)	0.7441 (0.0290)	0.7536 (0.0289)	0.7398 (0.0297)

괄호안의 값은 표준오차임.

표 3.1에서 고려된 세 모형 하에서, SANN 대치법에 의한 결측치 추정이 기존 방법들(KNN, SKNN, ANN)에 의한 추정보다 평균적으로 그 성능이 우수함을 알 수 있다. 가령, 결측비율이 10%, 차원 수 $p = 8$인 경우의 "SANN 대치법의 최소 NRMSE들의 평균/SKNN 대치법의 최소 NRMSE들의 평균"이 모형 1에서 0.9936, 모형 2에서 0.9926, 모형 3에서 0.9788임을 확인할 수 있다. 이는 SANN 대치법이 SKNN 대치법보다 더 정확한 추정을 한다는 것을 의미한다. 또한, 모형 3에서의 NRMSE 비가 모형 1에서 보다 작은 것은 자료의 밀도가 낮은 지역에서 결측이 발생했을 때 SANN 대치법이 SKNN 대치법보다 더 효율적인 추정이 가능하기 때문이다. 한편, 모든 모형에서 결측비율이 높아질수록 "SANN 대치법의 최소 NRMSE들의 평균/ANN 대치법의 최소 NRMSE들의 평균"이 작아지는 것을 볼 수 있다. 이는 제안된 SANN 대치법이 결측값이 대치된 목표개체의 정보를 다음 결측값의 추정에 사용함으로써, 목표개체를 이웃으로 고려하지 않는 ANN 대치법보다 더 정확한 추정을 할 수 있기 때문이다.

3.2. NCI60 자료를 이용한 모의실험

NCI60 자료는 60개의 변수로 이루어진 9703개의 개체로 구성되어 있지만, 본 논문에서는 그 중 완전한 $n = 681$개의 개체와 $p = 8$개의 melanoma 변수만을 사용하였다. 3.1절의 모의실험과 동일한 절차에 의해 각 방법론들의 성능을 비교하였으며, 그 결과는 표 3.2에 주어져있다. 표 3.2에서 SANN 대치법이 기존의 대치법(KNN, SKNN, ANN)에 비해 그 성능이 우수함을 알 수 있다. 특히, 표 3.1의 모형을 이용한 모의실험 결과에서와 마찬가지로, 결측비율이 10%일 때 보다 30%일 때 기존 방법들보다 SANN 대치법의 성능이 더 뛰어남을 확인할 수 있다.

그림 3.1은 결측비율이 30%, 임계치 $q = 1.165$일 때, 목표개체 $\boldsymbol{x}_a \in M$에 대한 제곱근 평균절대오차(Root Mean Absolute Error; RMAE)를 이용하여 SANN 대치법과 ANN 대치법을 비교한 것이다.

$$\mathrm{RMAE}(a) = \sqrt{\sum_{j \in V_a} \frac{|\hat{x}_{aj} - x_{aj}|}{n_a}},$$

여기서 n_a는 목표개체 $\boldsymbol{x}_a \in M$에서 결측된 변수들의 개수이다. 그림 3.1에서 가로축인 레버리지는 개체와 자료의 중심과의 거리를 나타내고 세로축은 ANN 대치법과 SANN 대치법의 RMAE 차이를 나타낸다. 다시 말해, 세로축의 값이 0보다 크면 SANN 대치법이 ANN 대치법보다 정확성이 높다는 것을 의미한다. 그림 3.1로부터 레버리지가 커질수록, 즉 결측된 개체가 자료의 중심부와 멀리 떨어져 있을수록 SANN 대치법의 결측치 추정이 ANN 대치법 보다 우수함을 알 수 있다.

4. 결론

실제 자료에서 흔히 발생하는 결측값을 다루는 방법으로 널리 쓰이는 KNN 대치법은 알고리즘이 간단하고 자료의 분포에 대해 강건성을 가진다는 장점이 있지만 자료의 국소적 특징을 간과하고 결측값이 발생한 개체가 지닌 관측값들은 이용하지 않는다는 단점을 가진다. 본 논문에서는 이러한 단점을 보완하기 위하여, 자료의 국소적 특징을 고려하는 ANN 대치법과 결측값이 대치된 개체를 재활용하는 SKNN

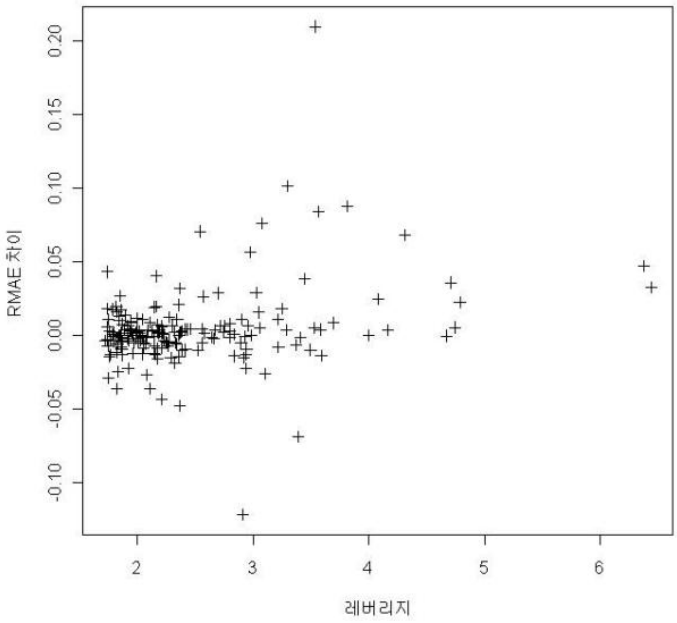

그림 **3.1.** ANN 대치법과 SANN 대치법의 RMAE 차이 비교

대치법의 장점을 결합한 SANN(Sequential Adaptive Nearest Neighbors) 대치법을 제안하였다. 다양한 모형 하에서의 모의실험을 통해서 결측비율이 높고 자료의 밀도가 낮은 지역에서 결측값이 발생한 경우, 기존의 방법들보다 SANN 대치법의 성능이 뛰어나다는 것을 확인하였고, 또한 실제사례 분석을 통해서 SANN 대치법이 기존의 방법에 비해 우수하다는 것을 보임으로써 제안된 방법의 활용가능성을 제시하였다.

참고문헌

맹진우, 방성완, 전명식 (2010). 수정된 적응 최근접 방법을 활용한 판별분류방법에 대한 연구, <응용통계연구>, **23**, 1093–1102.

이상은, 신기일 (2010). BLS 무응답 보정법을 이용한 대체법과 이월대체법에 관한 연구, <응용통계연구>, **23**, 909–921.

이진희, 김진, 이기재 (2006). 표본조사에서 공간변수를 이용한 결측 대체의 효율성 비교, <응용통계연구>, **19**, 57.

전명식, 최인경 (2009). Adaptive nearest neighbors를 활용한 판별분류방법, <응용통계연구>, **22**, 479–488.

Dixon, J. K. (1979). Pattern recognition with partly missing data, *IEEE Transactions on Systems, Man, and Cybernetics*, **9**, 617–621.

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.

Kim, K. Y., Kim, B. J. and Yi, G. S. (2004). Reuse of imputed data in microarray analysis increases imputation efficiency, *BMC Bioinformatics*, **5**, 160.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis With Missing Data*, Wiley, New York.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R. B. (2001). Missing value estimation methods for DNA microarrays, *Bioinformatics*, **7**, 520–525.

On the Use of Sequential Adaptive Nearest Neighbors for Missing Value Imputation

So Hyun Park[1] · Sungwan Bang[2] · Myoungshic Jhun[3]

[1]Department of Statistics, Korea University
[2]Department of Mathematics, Korea Military Academy
[3]Department of Statistics, Korea University

(Received September 2011; accepted October 2011)

Abstract

In this paper, we propose a Sequential Adaptive Nearest Neighbor(SANN) imputation method that combines the Adaptive Nearest Neighbor(ANN) method and the Sequential k-Nearest Neighbor(SKNN) method. When choosing the nearest neighbors of missing observations, the proposed SANN method takes the local feature of the missing observations into account as well as reutilizes the imputed observations in a sequential manner. By using a Monte Carlo study and a real data example, we demonstrate the characteristics of the SANN method and its potential performance.

Keywords: Adaptive nearest neighbors, imputation, k-nearest neighbors, missing data.

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