The Korean Journal of Applied Statistics (2018) DOI: http://dx.doi.org/10.5351/KJAS.2018.31.4.507
31(4), 507–516

On the use of weighted adaptive nearest neighbors for missing value imputation

Yunjin Yum[a] · Dongjae Kim[a,1]

[a]Department of Biomedicine · Health Science, The Catholic University of Korea

(Received June 4, 2018; Revised June 21, 2018; Accepted July 12, 2018)

Abstract

Widely used among the various single imputation methods is k-nearest neighbors (KNN) imputation due to its robustness even when a parametric model such as multivariate normality is not satisfied. We propose a weighted adaptive nearest neighbors imputation method that combines the adaptive nearest neighbors imputation method that accounts for the local features of the data in the KNN imputation method and weighted k-nearest neighbors method that are less sensitive to extreme value or outlier among k-nearest neighbors. We conducted a Monte Carlo simulation study to compare the performance of the proposed imputation method with previous imputation methods.

Keywords: k-nearest neighbors, adaptive nearest neighbors, weighted nearest neighbors, missing value, imputation

1. 서론

결측치란 계획된 대로 특정 피험자에게서 특정 변수를 특정 시점에 얻지 못한 경우를 말한다. 관측자료나 실험자료의 생성 과정에서 결측치의 발생은 매우 통상적이다. 결측치가 발생하면 편의를 일으키는 등 분석 결과에 큰 영향을 미칠 수 있기 때문에 알맞은 결측치 대치법을 선택하여 결측값을 추정하는 것은 굉장히 중요하며, 현재 대치법에 관한 연구가 활발하게 진행되고 있다. 대표적인 결측값 대치법으로 단일대치법과 다중대치법이 있다. 한 개의 결측값을 한 개의 값으로 대치하는 방법을 단일대치라 하며, 단일대치법에는 last observation carried forward, 평균 대치법(mean imputation), 핫덱 대치법(Hot-Deck imputation) 등이 있다. 그러나, 단일대치법은 추정량의 표준오차를 작게 만들어 제1종 오류가 증가할 가능성이 있다. 한편, 한 개의 결측값에 대하여 최소 두 번 이상 대치를 실시하여 여러 개의 대치된 자료를 생성하여 대치하는 방법을 다중대치라 한다. 하지만, 다중대치법은 계산이 복잡하고 시간이 오래 걸린다는 단점이 있다 (Little과 Rubin, 1987; Kang, 2013).

여러 가지 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 활

[1]Corresponding author: Department of Biomedicine · Health Science, The Catholic University of Korea, 222 Banpo-dero Seocho-gu, Seoul 06591, Korea. E-mail: djkim@catholic.ac.kr

용되고 있다 (Park 등, 2011). 하지만, KNN 대치법은 고정적인 k개의 이웃을 사용하므로 결측치가 발생하는 위치에 따른 국소적 특징을 간과할 수 있다. 따라서 이러한 KNN 대치법의 단점을 보완하고 확장한 적응 최근접 이웃(adaptive nearest neighbors; ANN) 대치법 (Jhun 등, 2007), 순차 적응 최근접 이웃(sequential ANN; SANN) 대치법 (Park 등, 2011) 등이 제안되었다. 한편, Lim과 Kim (2015)은 가중 k-최근접 이웃 방법을 이용한 통계적 매칭기법의 장점을 적용하여, k개의 최근접 이웃들에 극단값이나 이상값이 있는 경우에 이들의 영향에 덜 민감하고 정확도를 높일 수 있는 가중 k-최근접 이웃(weighted KNN; WKNN) 대치법을 제안하였다.

본 논문에서는 KNN 대치법을 보완한 ANN 대치법과 WKNN 대치법의 장점을 결합한 가중 적응 최근접 이웃(weighted ANN; WANN) 대치법을 제안하였다. 제안하는 방법은 자료에서 각 개체의 국소적 특징을 고려하여 개체에 따라 결측치 대치에 활용하는 최근접 이웃의 개수 k를 변화시키고 최근접이웃들 중에 극단값이나 이상값이 있는 경우, 이들의 영향에 덜 민감하면서도 정확도를 높인다는 두 가지 장점을 모두 지닌다. 2장에서는 WANN 대치법을 소개하고 단순 예제를 통해 제안한 방법을 설명하였으며, 3장에서는 다양한 설정에서의 모의실험을 통하여 기존의 ANN, KNN, WKNN 대치법과 제안된 WANN 대치법의 성능을 비교하였다. 마지막으로 4장에서는 결론 및 고찰을 통하여 마무리하였다.

2. 제안하는 방법

새롭게 제안하는 가중 적응 최근접이웃 대치법은 연속형 자료에서 결측된 개체의 특성에 따라 이웃의 개수를 유연하게 조정하는 ANN 대치법 (Jhun 등, 2007)의 장점과 최근접 이웃들에 극단값이 있는 경우에 덜 민감한 WKNN 대치법 (Lim과 Kim, 2015)을 결합한 방법이다.

먼저 유사성 거리는 Euclidean distance를 사용하여 측정하고 거리 행렬의 중앙값을 더해 수정된 거리 행렬을 새롭게 정의한다. 수정된 거리의 비와 임계치 q를 이용하여 개체의 이웃의 수 k를 정한다.

계산된 거리를 이용하여 k개의 최근접 이웃들의 가중치를 산출한다. 이때, 데이터에 따라 거리의 분포가 달라지므로 변화에 대해 유연성 있는 가중치 계산을 위해 가장 가까운 $k+1$개의 이웃들의 거리를 0과 1 사이의 값으로 변환한다. 여기서 $d_{(1)}$은 결측된 개체와의 거리가 가장 가까운 거리로 0의 거리를 갖고, $d_{(k+1)}$은 $k+1$번째 거리로 1의 거리를 갖게 된다.

$$d_i = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}, \qquad d_i^{'} = \frac{d_i - d_{(1)}}{d_{(k+1)} - d_{(1)}}, \quad i = 1, 2, \ldots, k+1.$$

기존의 WKNN 대치법에서 사용한 커널함수 중 첨도가 큰 Triweight 함수와 작은 Epanechnikov 함수, 그리고 중간인 Triangular 함수를 새롭게 추가하여 사용하였다. 커널함수의 첨도가 크면 최근접이웃들에 대한 가중치간의 차이를 크게 하는 효과가 있다. 또한 커널함수에 대입하여 가중치를 계산하면 $k+1$번째 이웃의 관측치는 1의 거리를 갖게 되어 가중치가 0이 된다. 여기서 가중치의 합은 1이 된다.

$$\text{Triweight} : W_l = \frac{\frac{35}{32}\left(1 - d_i^{'2}\right)^3}{\sum_{i=1}^{k}\frac{35}{32}\left(1 - d_i^{'2}\right)^3}, \quad l = 1, 2, \ldots, k+1,$$

$$\text{Epanechnikov} : W_l = \frac{\frac{3}{4}\left(1 - d_i^{'2}\right)}{\sum_{i=1}^{k}\frac{3}{4}\left(1 - d_i^{'2}\right)}, \quad l = 1, 2, \ldots, k+1,$$

$$\text{Triangular} : W_l = \frac{\left(1 - d_i^{'}\right)}{\sum_{i=1}^{k}\left(1 - d_i^{'}\right)}, \quad l = 1, 2, \ldots, k+1.$$

2.1. 가중 적응 최근접 이웃 대치법

세 가지 커널함수에 따라 세 가지 대치법 Triweight-WANN, Epanechnikov-WANN, Triangular-WANN을 제안하였고 그 단계는 다음을 따른다.

단계 1: n개의 관측개체와 p개의 변수를 가지고 있는 원 자료 행렬을 D라고 할 때, 자료행렬 D를 D_m과 D_c로 나눈다. 여기서 D_m은 적어도 하나의 결측치가 포함되어 있는 r개의 관측치를 갖는 자료행렬이고, D_c는 원 자료 행렬인 D에서 결측치가 포함되어 있지 않은 $n-r$개의 관측개체로 구성된 행렬이다.

$$D=\begin{pmatrix} x_{11}^* & x_{12} & x_{13} & \cdots & x_{1p} \\ x_{21} & x_{22} & x_{23}^* & \cdots & x_{2p} \\ x_{31} & x_{32} & x_{33} & \cdots & x_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2}^* & x_{n3} & \cdots & x_{np} \end{pmatrix}, \quad x_{ij}^* : \text{missing}(i=1,2,\ldots,n,\ j=1,2,\ldots,p), \quad D=\begin{pmatrix} D_m \\ D_c \end{pmatrix}.$$

단계 2: 결측치가 포함되어 있는 행렬 D_m에서 a번째 행 x_a와 $x_b \in D_c$의 각 행들 간의 Euclidean distance를 계산한다. 단, x_a에서 결측치는 제외하고 관측치만 고려한다.

$$d_{ab}=\sqrt{\sum_{j=1}^{p}(x_{aj}-x_{bj})^2}.$$

단계 3: 수정된 거리 행렬

$$d_{ab}^*=[d_{ab}+\text{median}(d_{ab},\ b=1,2,\ldots,n-r)]$$

를 계산한 후, $d_{ab}^*/d_{a(1)}^*$ 값이 임계치 $q\,(\geq 1)$이하인 k_a를 x_a의 이웃으로 선택한다.

단계 4: 단계 2에서 계산된 거리 값(d_{ab})을 통해 k_a+1개의 최근접 이웃 개체들을 찾고 이들에 대한 거리를 0과 1사이의 값으로 변환한다.

단계 5: 변환된 거리는 커널함수에 반영하여 k_a+1개 최근접이웃들에 대한 가중치를 계산한다. D_c에서 선택된 k_a개의 최근접 이웃들과 가중치를 각각 곱해서 더한 가중평균값을 구하고, 이 값으로 $x_a \in D_m$에 있는 결측치 x_{aj}^*를 대치한다.

제안된 WANN 대치법에서 임계치 q는 결측치 대치에 이용하는 최근접 이웃의 개수 k를 선택하는 기준으로 두 거리 비 $d_{ab}/d_{a(1)}$의 최대 허용 한계치가 된다. 즉, 거리의 비가 q보다 작거나 같은 개체들을 최근접 이웃들로 선택함으로써 자료의 국소적 특징을 반영하게 된다. 여기서 거리행렬의 중앙값을 더하여 결측이 발생한 개체와 가장 근접한 개체의 거리가 0에 가까운 경우를 피해 거리 비가 무한히 커지게 되어 어떤 개체도 이웃으로 선택될 수 없는 것을 막는 것이다.

2.2. 예제

다음은 개체의 수 $n=10$ 그리고 차원 $p=4$일 때 자료를 인위적으로 생성하여 WANN 대치법을 적용한 예이며, Table 2.1은 그 결과를 나타낸 것이다.

결측치가 포함되어 있는 행렬 D_m에서 a번째 개체 x_a와 D_c의 b번째 개체 x_b들 간의 Euclidean distance(d_{ab})를 계산한다. 먼저 D_m의 첫 번째 행인 x_9을 살펴보자. 거리 행렬의 중앙값 8.544를 d_{9b}에

Table 2.1. Example

	i	변수				d_{ab}		d^*_{ab}		k_a		d'_{ab}		가중치	
		$X1$	$X2$	$X3$	$X4$	d_{9b}	d_{10b}	d^*_{9b}	d^*_{10b}	k_9	k_{10}	d'_{9b}	d'_{10b}	w_{9b}	w_{10b}
D_c	1	2	4	7	5	5.385	2.449	13.929	7.398	**1.000**	**1.000**	0.000	0.000	0.626	0.626
	2	5	8	3	9	10.050	5.745	18.594	10.694	1.335	1.445	1.505	1.837		
	3	7	1	5	7	9.381	5.000	17.925	9.949	1.287	1.345	1.289	1.422		
	4	4	6	2	4	7.681	3.162	16.225	8.111	**1.165**	**1.096**	0.741	0.397	0.057	0.374
	5	9	5	1	3	8.485	7.280	17.029	12.229	1.223	1.653	1.000	2.693		
	6	5	4	1	2	8.602	4.899	17.146	9.848	1.231	1.331	1.038	1.366		
	7	6	2	4	8	9.950	5.099	18.494	10.048	1.328	1.358	1.473	1.477		
	8	3	4	8	7	6.782	4.243	15.326	9.192	**1.100**	1.242	0.451	1.000	0.316	
D_m	9	$x^*_{9,1}$	7	9	1										
	10	3	$x^*_{10,2}$	5	4										

Table 3.1. Coefficient of variance of the variables for each data ($\rho = 0, 0.3$)

Data	Coefficient of variance					
	$X1$	$X2$	$X3$	$X4$	$X5$	$X6$
A1	0.5	0.5	0.5	0.5	0.5	0.5
B1	0.1	0.2	0.3	0.4	0.5	0.6
B2	0.1	0.3	0.5	0.7	0.9	1.1
B3	0.1	0.4	0.7	1.0	1.3	1.6
C1	0.1	0.1	0.4	0.4	0.7	0.7
C2	0.1	0.1	0.6	0.6	1.1	1.1
C3	0.1	0.1	0.8	0.8	1.5	1.5
D1	0.1	0.1	0.1	0.5	0.5	0.5
D2	0.5	0.5	0.5	1.0	1.0	1.0
D3	0.5	0.5	0.5	1.5	15	1.5
E1	0.5	0.5	1.5	1.5	1.5	1.5
E2	0.5	0.5	0.5	0.5	1.5	1.5

더하여 수정된 거리 d^*_{9b}를 계산한다. $d^*_{9(1)} = 13.929$이므로 거리 비(k_a) $d^*_{9b}/13.929$를 계산하고 임계치 q를 1.17로 설정하면, 거리 비가 q보다 작거나 같은 개체들은 x_1, x_4, x_8로 3개의 최근접 이웃이 선택된다. 다음으로 d_{9b}를 결측된 개체와 가장 가까운 거리인 $d_{9(1)} = 5.385$와 $(3+1)$번째 거리인 $d_{9(4)} = 8.485$를 이용하여 변환한다. 변환된 거리 d'_{9b}를 Triweight 함수에 대입하여 가중치를 계산한다. 마지막으로 계산된 가중치를 D_c의 첫 번째 열과 곱하여 $x^*_{9,1}$을 가중 평균값인 $(2\times0.626)+(4\times0.057)+(3\times0.316) = 2.4$로 대치한다. 이와 같은 방법으로 x_{10}에 대해서도 계산하면, $x^*_{10,2}$은 $x'_{10,2} = 4.7$로 대치된다.

3. 모의실험 계획 및 결과

Lim과 Kim (2015)은 커널함수 Triweight, Epanechnikov에 따라 Triweight-WKNN, Epanechnikov-WKNN 대치법을 제안하였다. 본 논문에서는 커널함수 Triweight, Epanechnikov, Triangular를 사용하여 Triweight-WANN, Epanechnikov-WANN, Triangular-WANN 대치법을 새롭게 제안하였고 기존의 WKNN 대치법에 새롭게 Triangular함수를 사용하여 Triangular-WKNN을 추가로 제안하였다. 새롭게 제안한 세 가지 방법과 기존 방법인 ANN, KNN, WKNN의 세 가지 방법들의 성능을 비교하기 위

하여 모의실험을 시행하였다. 300개의 개체에 대해 숫자형 변수가 6개인 다변량정규분포를 따르는 자료를 생성하였고 상관관계가 없는 무상관관계와 상관계수가 0.3일 때를 고려하였다. 이때, 유사성 거리에 따라 최근접 이웃들을 선택하므로 변동계수를 이용하여 산포도에 따라 각 결측치 대치법들을 비교하였다. Table 3.1은 상관관계에 따른 6개의 변수에 대한 변동계수를 나타낸 표이다. A자료는 각 변수의 변동계수가 모두 0.5로 같은 경우이고 자료 B는 변수마다 변동계수가 모두 다른 경우로 B1은 0.1, B2는 0.2, B3는 0.3씩 증가한다. 다음으로 자료 C1, C2, C3는 두 개의 변수마다 변동계수가 0.3, 0.5, 0.7씩 증가하도록 하였다. 또한 자료 D는 세 개의 변수끼리 변동계수를 동일하게 하여 D1은 0.4, D2는 0.5, D3는 1.0씩 증가시켰다. 마지막으로 자료 E1은 0.5의 변동계수를 갖는 변수가 2개 1.5의 변동계수를 갖는 변수를 4개로 설정하였고, 반대로 E2는 변동계수 1.5를 갖는 변수를 2개 0.5를 갖는 변수를 4개로 설정하였다. 그리고 완전임의결측 가정 하에 5%와 10%의 결측을 발생시킨 후 ANN, KNN, WKNN, 제안한 방법을 통해 결측값을 추정하였다.

결측값을 추정한 후에 대치법들의 성능을 비교하기 위하여 정규화 제곱근 평균제곱오차(normalized root mean squared error; NRMSE)를 이용하였다.

$$\text{NRMSE} = \frac{1}{x_{\max} - x_{\min}} \left\{ \sum_i \sum_j \frac{(x_{ij} - x'_{ij})^2}{N} \right\}^{\frac{1}{2}},$$

여기서 x_{ij}는 실제값, x'_{ij}은 대치된 값이고, $x_{\max}$와 $x_{\min}$은 실제 값들의 최대값과 최소값, N은 결측치의 총 개수를 나타낸다. NRMSE의 값이 0에 가까울수록 결측값을 더 정확하게 추정했음을 의미한다. KNN 대치법과 WKNN 대치법의 k는 1부터 50까지 1씩 증가시켰다. ANN과 WANN 대치법의 임계치 q는 1과 2 사이의 값에서 좋은 추정을 하여 1부터 2까지 0.005씩 동일한 간격으로 선택하였다. 여기서 이웃의 수 k와 q의 척도가 다르기 때문에 각 각의 데이터에 대하여 가능한 k와 q값들에 대한 NRMSE의 최소값을 사용하여 대치 방법들의 성능을 비교하였다. 이와 같은 과정을 독립적으로 100회 반복하였으며, Table 3.2는 모의실험의 결과를 정리한 것이다.

먼저 무상관관계를 갖는 자료를 살펴보면, 결측률에 상관없이 모든 자료에서 ANN과 새롭게 제안한 Epanechnikov-WANN과 Triangular-WANN이 가장 우수한 성능을 보였다. 또한 모든 자료에서 WKNN의 NRMSE 값이 가장 컸다. 변동계수가 0.5로 모든 변수에 대해 동일한 경우 결측 5%, 10%에서 ANN이 가장 작은 NRMSE 값을 가졌고 새롭게 제안한 Epanechnikov-WANN, Triangular-WANN이 그 뒤를 따랐다. 다음으로 각 변수들의 변동계수가 모두 다른 자료를 살펴보면, 0.1씩 증가한 경우, 결측 5%에서 ANN, 제안한 방법 Epanechnikov-WANN과 Triangular-WANN, KNN, 제안한 방법 Triweight-WANN순으로 좋았다. 10% 결측일 때는 Triweight-WANN이 KNN보다 좋았다. 변동계수의 차이가 0.2인 경우 결측 5%와 10%에서 ANN이 가장 우수한 성능을 보였고 새롭게 제안한 방법들이 모두 Epanechnikov-WANN, Triangular-WANN, Triweight-WANN의 순서로 그 다음으로 우수한 성능을 보였다. 각 변수의 변동계수의 증가 폭이 0.3으로 커진 경우에는 결측 5%와 10%일 때 모두 새롭게 제안한 방법들 중에서 Epanechnikov-WANN이 가장 우수한 성능을 보였고 ANN과 제안한 나머지 두 방법들이 그 뒤를 따랐다. 두 개의 변수끼리 같은 변동계수를 갖는 자료에서 결측 5%일 때 차이가 0.3인 경우, ANN이 가장 작은 값을 가졌고 새롭게 제안한 세 가지 방법 모두 다른 대치법들보다 좋은 성능을 보였다. 하지만, 차이가 0.5인 경우에는 새롭게 제안한 Epanechnikov-WANN이 가장 작은 값을 가졌으며 그 뒤를 ANN이 따랐다. 차이가 0.7일 때는 0.3일 때와 같은 결과를 보였다. 결측 10%에서는 세 자료에서 모두 새롭게 제안한 방법 중에서 Epanechnikov-WANN이 가장 작은 값을 가졌다. 그 뒤를 ANN과 새롭게 제안한 Triangular-WANN, Triweight-WANN이 이었다. 즉, 자료의 변동계수가 다양하고 그 변화의 폭이 클 때 제안한 방법들 중에서 Epanechnikov함

Table 3.2. Average of the minimum NRMSE based on 100 independent trials

Data	Corr	Missing	ANN	KNN	WKNN			WANN		
					epan	triang	triw	epan	triang	triw
A1	0	5%	0.203024	0.204736	0.206258	0.206982	0.209510	0.203558	0.203849	0.204492
		10%	0.187363	0.188312	0.189103	0.189628	0.191517	0.187621	0.187844	0.188613
	0.3	5%	0.180410	0.179764	0.180693	0.180985	0.182151	0.180920	0.181094	0.181294
		10%	0.166551	0.165710	0.166086	0.166231	0.166961	0.166599	0.166670	0.166774
B1	0	5%	0.176234	0.177209	0.178353	0.178951	0.180297	0.176674	0.177005	0.177372
		10%	0.156355	0.157303	0.157763	0.158170	0.159345	0.156374	0.156592	0.156994
	0.3	5%	0.158756	0.158821	0.159892	0.160347	0.161226	0.159317	0.159607	0.159797
		10%	0.139954	0.139575	0.140052	0.140187	0.140683	0.140197	0.140297	0.140384
B2	0	5%	0.161194	0.162644	0.163572	0.164281	0.165755	0.161430	0.161758	0.162107
		10%	0.139572	0.140616	0.141134	0.141509	0.142519	0.139661	0.139857	0.140162
	0.3	5%	0.154233	0.154329	0.155166	0.155836	0.156966	0.154617	0.155030	0.155209
		10%	0.137062	0.136891	0.137282	0.137617	0.138303	0.137248	0.137438	0.137552
B3	0	5%	0.168182	0.169810	0.170351	0.171248	0.172902	0.168068	0.168524	0.169010
		10%	0.145578	0.146760	0.146993	0.147474	0.148763	0.145492	0.145745	0.146289
	0.3	5%	0.154238	0.154572	0.155624	0.156111	0.157121	0.154921	0.155252	0.155396
		10%	0.135622	0.135719	0.136192	0.136471	0.137208	0.135995	0.136149	0.136268
C1	0	5%	0.162957	0.164742	0.165814	0.166452	0.167930	0.163242	0.163612	0.164023
		10%	0.143257	0.144209	0.144777	0.145227	0.146477	0.143247	0.143473	0.143835
	0.3	5%	0.155863	0.156882	0.157810	0.158155	0.159110	0.156524	0.156806	0.157105
		10%	0.139848	0.140144	0.140581	0.140784	0.141551	0.140185	0.140291	0.140496
C2	0	5%	0.165634	0.167675	0.168424	0.169239	0.171028	0.165619	0.166079	0.166528
		10%	0.141425	0.142636	0.142705	0.143299	0.144463	0.141154	0.141513	0.142051
	0.3	5%	0.144521	0.144745	0.145592	0.145920	0.146967	0.145049	0.145240	0.145410
		10%	0.125919	0.126034	0.126540	0.126723	0.127329	0.126239	0.126294	0.126469
C3	0	5%	0.157434	0.158700	0.159309	0.159957	0.161412	0.157501	0.157823	0.158188
		10%	0.140450	0.141346	0.141517	0.141902	0.142949	0.140335	0.140565	0.140902
	0.3	5%	0.153561	0.153939	0.154450	0.155050	0.155954	0.153768	0.154167	0.154340
		10%	0.139364	0.139177	0.139257	0.139531	0.140081	0.139325	0.139514	0.139614
D1	0	5%	0.161930	0.163569	0.164482	0.164981	0.166401	0.162161	0.162506	0.162886
		10%	0.144697	0.145650	0.146119	0.146532	0.147598	0.144749	0.145001	0.145455
	0.3	5%	0.150324	0.150715	0.151516	0.151918	0.152768	0.150746	0.151054	0.151274
		10%	0.133990	0.133922	0.134355	0.134535	0.134988	0.134284	0.134407	0.134517
D2	0	5%	0.176733	0.177854	0.179371	0.179997	0.181784	0.177324	0.177581	0.178046
		10%	0.161323	0.162578	0.163359	0.163655	0.164860	0.161716	0.161922	0.162449
	0.3	5%	0.170369	0.170657	0.172019	0.172414	0.173844	0.171266	0.171506	0.171778
		10%	0.153931	0.153483	0.154080	0.154255	0.154991	0.154346	0.154429	0.154557
D3	0	5%	0.167354	0.168852	0.169978	0.170575	0.171790	0.168027	0.168408	0.168689
		10%	0.152360	0.153246	0.153788	0.154121	0.155048	0.152601	0.152813	0.153102
	0.3	5%	0.161373	0.161885	0.162872	0.163083	0.163934	0.162018	0.162282	0.162537
		10%	0.145889	0.146024	0.146638	0.146763	0.147319	0.146359	0.146437	0.146605
E1	0	5%	0.179704	0.181048	0.182161	0.182742	0.184230	0.180145	0.180413	0.180744
		10%	0.163793	0.164708	0.165294	0.165712	0.166949	0.163984	0.164173	0.164566
	0.3	5%	0.172637	0.172580	0.173410	0.173887	0.175061	0.173094	0.173380	0.173472
		10%	0.161373	0.160997	0.161288	0.161482	0.162101	0.161424	0.161564	0.161630
E2	0	5%	0.154801	0.156036	0.157069	0.157779	0.159084	0.155183	0.155544	0.155883
		10%	0.136834	0.137636	0.138462	0.138780	0.139617	0.137137	0.137351	0.137683
	0.3	5%	0.150860	0.150571	0.151833	0.152354	0.153360	0.151704	0.152115	0.152220
		10%	0.137124	0.136604	0.137109	0.137289	0.137830	0.137602	0.137704	0.137854

NRMSE = normalized root mean squared error; ANN = adaptive nearest neighbors; KNN = k-nearest neighbors; WKNN = weighted KNN; WANN = weighted ANN.

epan = epanechnikov kernel function; triang = triangular kernel function; triw = triweight kernel function.

수를 쓴 Epanechnikov-WANN 대치법의 성능이 가장 우수함을 알 수 있다. 세 개의 변수가 같은 변동계수를 갖고 그 차이가 0.5인 자료에서 결측 5%일 때 ANN, 제안한 방법들 중에서 Epanechnikov-WANN과 Triangular-WANN, KNN, Triweight-WANN순으로 좋았고 나머지 두 자료에서는 제안한 방법 Triweight-WANN이 KNN보다 작은 값을 보였는데 이 결과는 결측률 10%일 때 세 자료에서 모두 동일하다. 이는 같은 변동계수를 갖는 변수가 많아지면 ANN이 새롭게 제안한 Epanechnikov-WANN보다 좋은 성능을 보인다는 것을 알 수 있다. 마지막으로 무상관관계에서 여섯 개의 변수 중 2개가 0.5이고 4개가 1.5인 경우에는 결측 5%, 10%일 때 ANN과 더불어 새롭게 제안한 세 가지 대치법들이 Epanechnikov-WANN, Triangular-WANN, Triweight-WANN 순서로 다른 대치법들 보다 모두 좋은 성능을 보였고 여섯 개의 변수 중에서 2개가 1.5, 나머지가 0.5인 자료에서 결측 5%일 때도 같은 결과를 보였다. 하지만, 결측 10%에서는 KNN이 Triweight-WANN보다 작은 값을 가졌고 이는 변동계수가 모두 동일한 경우와 같은 결과이다.

다음으로 상관계수가 0.3인 자료를 살펴보면 대체적으로 ANN과 KNN이 가장 작은 값을 가졌는데 각각의 대치법들의 NRMSE 값의 차이는 무상관관계일 때보다 작았다. 먼저 변동계수가 0.5로 모두 동일할 때, 결측률에 상관없이 KNN이 가장 작은 값을 가졌다. 결측 5%에서 ANN, Epanechnikov-WKNN 그리고 새롭게 제안한 Epanechnikov-WANN이 그 뒤를 이었다. 결측 10%에서는 Epanechnikov-WKNN과 Triangular-WKNN이 ANN보다 좋은 성능을 보였다. 변동계수가 모두 같고 0.3의 상관관계를 가질 때에는 KNN과 Epanechnikov-WKNN 대치법이 좋은 성능을 보였다. 변수들의 변동계수가 모두 다른 자료에서 결측 5%일 때, 변동계수가 0.1, 0.3씩 증가할 때는 ANN과 KNN이 가장 작은 값을 가졌고, 새롭게 제안하는 세 가지 방법 Epanechnikov-WANN, Triangular-WANN, Triweight-WANN이 모두 WKNN보다 작은 값을 가졌다. 0.2로 증가하는 경우에는 Epanechnikov-WKNN이 제안한 Triweight-WANN보다 작은 값을 가졌다. 결측 10%일 때 0.1씩 증가하는 경우에 KNN이 가장 작았고 그 뒤를 ANN, Epanechnikov-WKNN이 따랐다. 0.2씩 증가하는 자료에서는 ANN, 새롭게 제안한 Epanechnikov-WANN, Epanechnikov-WKNN, 제안한 Triangular-WANN이 그 뒤를 따랐다. 하지만 0.3으로 폭이 증가하면 새롭게 제안한 Triangular-WANN이 Epanechnikov-WKNN보다 작은 값을 가졌다. 변수들의 변동계수가 모두 다르고 그 값이 큰 경우에는 새롭게 제안한 방법들이 대체적으로 성능이 좋아짐을 알 수 있다. 두 개의 변수가 같은 변동계수를 가지고 그 차이가 0.3, 0.5, 0.7씩 증가하는 자료를 살펴보면 먼저 결측 5%인 경우에 세 자료에서 ANN이 가장 우수함을 보였다. 0.3, 0.7씩 차이가 나는 경우에는 새롭게 제안한 Epanechnikov-WANN과 Triangular-WANN 그리고 KNN이 그 뒤를 따랐고, 0.5씩 차이가 나는 경우에는 KNN과 새롭게 제안한 세 가지 대치법들이 모두 WKNN보다 좋은 결과를 보였다. 결측 10%일 때 0.3, 0.5씩 증가하는 자료에서 결측 5%일 때 0.5씩 차이가 나는 경우와 같은 결과를 보였다. 하지만, 차이가 0.7인 경우에는 KNN이 가장 좋은 결과를 보였고 그 뒤를 Epanechnikov-WKNN, 제안한 방법 중 Epanechnikov-WANN 그리고 ANN이 뒤를 따랐다. 두 개의 변수가 동일한 변동계수를 갖는 자료에서 결측률이 높고 변동계수의 증가 폭이 커지면 ANN 대치법의 성능이 다른 대치법들보다 떨어지는 것을 볼 수 있다. 다음으로 세 개의 변수마다 변동계수가 0.4, 0.5, 1씩 차이가 날 때 결측 5%에서 모두 같은 결과를 보였는데, ANN과 KNN이 가장 작은 값을 가졌고 새롭게 제안한 세 가지 대치법들은 모두 WKNN보다 작은 값을 가졌다. 결측 10%에서, 변동계수의 차이가 0.4인 경우에는 KNN, ANN, 새롭게 제안한 Epanechnikov-WANN, Epanechnikov-WKNN순으로 좋은 결과를 보였고, 차이가 0.5인 경우에는 Epanechnikov-WKNN이 Epanechnikov-WANN보다 작은 값을 가졌다. 또한 차이가 1씩 날 때에는 결측이 5%일 때와 같은 결과를 보였다. 마지막으로 변동계수가 0.5인 변수가 2개 1.5인 변수가 4개인 자료와 0.5인 변수 4개, 1.5인 변수가 2개인 자료에서 결측 5%일 때는 KNN, ANN, 제안한 방법 Epanechnikov-WANN순으로 가장 우수함을 보였고 결측 10%에서는 KNN, Epanechnikov-WKNN, ANN이 가장 우수함을 보였다.

전반적으로 무상관관계를 갖는 자료와는 다르게 0.3의 상관관계를 가질 때에는 방법론들 간에 순위 변동이 많은 것을 볼 수 있다. 또한, 두 상관관계의 자료에서 결측 10%일 때보다 5%일 때 더 큰 NRMSE 값을 가지는데, 이는 결측의 수가 증가하면서 NRMSE 분모의 차이값이 커지기 때문이라고 볼 수 있다.

4. 결론 및 고찰

본 논문에서는 ANN 대치법과 WKNN 대치법을 결합한 WANN 대치법을 제안하였다. 제안한 방법과 WKNN 대치법은 각각 커널함수에 따라 세 가지 방법으로 나누고, ANN, KNN 대치법과 정규화 제곱근 평균제곱오차를 이용하여 비교하였다.

모든 변수들이 무상관관계를 갖는 다변량 정규분포에서 ANN과 새롭게 제안한 Epanechnikov-WANN, Triangular-WANN, Triweight-WANN이 대체적으로 우수한 성능을 보였다. 특히 Epanechnikov-WANN과 Triangular-WANN은 변동계수에 상관없이 모든 자료에서 KNN과 WKNN보다 좋은 성능을 보였다. 동일한 변동계수를 갖는 변수가 많고 그 증가 폭이 작을 때는 대체적으로 ANN이 가장 좋은 성능을 보였고 새롭게 제안한 Epanechnikov-WANN과 Triangular-WANN은 그 다음으로 우수한 성능을 보였다. 결측률이 높고 같은 변동계수를 갖는 변수가 적고 그 차이가 증가할수록 제안한 방법 중 Epanechnikov커널함수를 이용한 WANN 대치법이 ANN과 다른 대치법들에 비해 가장 작은 NRMSE값을 가졌다. 즉, 변수들의 변동계수가 다양하고 그 증가 폭이 클수록 개체의 국소적 특징을 반영하여 개체 수를 선택하고 Epanechnikov 커널함수를 이용하여 가중치를 산정하는, 본 논문에서 제안한 Epanechnikov-WANN 방법이 더 효율적이다.

변수들의 상관계수가 0.3인 경우에는 대체적으로 ANN 대치법과 KNN 대치법이 우수한 성능을 보였지만 무상관관계를 갖는 경우보다 NRMSE의 차이가 작았고 무상관관계를 갖는 경우와는 다르게 Epanechnikov-WKNN 대치법도 좋은 성능을 보였다. 특히, 결측률이 높을 때 같은 변동계수를 갖는 변수가 많으면 대체적으로 Epanechnikov-WKNN이 ANN보다 좋은 성능을 보였다. 하지만, 동일한 변동계수를 갖는 변수가 적고 그 차이가 클 때는 새롭게 제안한 방법들 중에서 Epanechnikov-WANN이 대체적으로 모든 WKNN보다 좋은 성능을 보였다. 따라서 변수들이 0.3의 상관관계를 갖는 경우에는 변동계수에 따른 대치법들 간에 차이가 명확하지 않았다.

상관계수에 관계없이 모든 자료에서 제안한 방법들과 WKNN 방법들을 비교하였을 때 항상 Epanechnikov, Triangular, Triweight의 순서로 우수한 성능을 보였다. 따라서 첨도가 작은 커널함수를 사용하여 가중치간의 차이를 작게 하는 것이 더 정확한 추정을 할 수 있다.

본 논문에서 제안하는 가중 적응 최근접 이웃 대치법은 조정된 이웃의 수를 사용하여 개체들의 국소적 특징을 반영하고 커널함수를 이용하여 가중치를 계산함으로써 각 변수들의 산포도를 고려할 수 있고 변동계수가 클 때에도 좋은 성능을 보인다. 즉, 최근접 이웃들의 극단값이나 이상값에도 높은 정확도를 보임으로써 앞으로 유용한 결측치 대치법이 될 것으로 기대한다.

References

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

Jhun, M., Jeong, H., and Koo, J. (2007). On the use of adaptive nearest neighbors for missing value imputation, *Communications in Statistics: Simulation and Computation*, **36**, 1275–1286.

Kang, S. (2013). *Medical Statistics Needed for Drug Development* (2nd ed), Freeacademy, Seoul.

Lim, C. and Kim, D. (2015). On the use of weighted k-nearest neighbors for missing value imputation, *The Korean Journal of Applied Statistics*, **28**, 23–31.

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

Park, S., Bang, S., and Jhun, M. (2011). On the use of sequential adaptive nearest neighbors for missing value imputation, *The Korean Journal of Applied Statistics*, **24**, 1249–1257.

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*, **17**, 520–525.

Yun, S. (2004). Imputation of missing values, *Journal of Preventive Medicine and Public Health*, **37**, 209–211.

가중 적응 최근접 이웃을 이용한 결측치 대치

염윤진[a] · 김동재[a,1]

[a]가톨릭대학교 의생명 · 건강과학과

(2018년 6월 4일 접수, 2018년 6월 21일 수정, 2018년 7월 12일 채택)

요 약

결측치를 대치하는 여러가지 단일대치법 중에서 다변량 정규성 등의 모수적 모형이 만족되지 않을 때에도 강건성(robustness)을 지니는 k-최근접 이웃 대치법(k-nearest neighbors; KNN)이 널리 활용된다. KNN대치법에서 자료의 국소적 특징을 반영한 적응 최근접 이웃(adaptive nearest neighbors; ANN) 대치법과 k개의 최근접 이웃들 중 극단값이나 이상값이 있는 경우 이들의 영향에 덜 민감한 가중 k-최근접 이웃(weighted KNN; WKNN) 대치법의 장점을 결합한 가중 적응 최근접 이웃(weighted ANN; WANN) 대치법을 제안하였다. 또한 모의실험을 통하여 기존의 방법들과 제안한 방법을 비교하였다.

주요용어: k-최근접 이웃, 적응 최근접 이웃, 가중 최근접 이웃, 결측치, 대치법

[1]교신저자: (06591) 서울시 서초구 반포대로 222, 가톨릭대학교 의생명 · 건강과학과.
E-mail: djkim@catholic.ac.kr