The Korean Journal of Applied Statistics (2015) DOI: http://dx.doi.org/10.5351/KJAS.2015.28.1.023
28(1), 23–31

On the Use of Weighted k-Nearest Neighbors for Missing Value Imputation

Chanhui Lim[a] · Dongjae Kim[a,1]

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

(Received October 6, 2014; Revised November 10, 2014; Accepted November 10, 2014)

Abstract

A conventional missing value problem in the statistical analysis k-Nearest Neighbor(KNN) method are used for a simple imputation method. When one of the k-nearest neighbors is an extreme value or outlier, the KNN method can create a bias. In this paper, we propose a Weighted k-Nearest Neighbors(WKNN) imputation method that can supplement KNN's faults. A Monte-Carlo simulation study is also adapted to compare the WKNN method and KNN method using real data set.

Keywords: Missing value, Imputation, k-nearest neighbors.

1. 서론

통계적 분석을 할 때 결측치가 발생하는 것은 매우 통상적이다. 결측치란 특정 피험자에게서 특정 변수를 특정 시점에 측정하여 관측치를 얻어야 하는데 얻지 못한 경우를 말한다. 결측치의 발생은 분석을 어렵게 할 뿐만 아니라, 편의 발생으로 인해 분석 결과에 크게 영향을 미친다. 결측치를 처리하는 가장 단순한 방법은 결측치가 있는 변수를 모든 분석에서 제거하고, 결측치가 없는 변수들만 분석하는 방법이다. 하지만 이 방법은 편향이 발생할 수 있고, 결측비율이 높아지면 표본의 크기가 감소하여 검정력이 줄어드는 단점이 있다. 그래서 지금까지 결측치 처리에 관하여 많은 방법론이 연구되어 왔다. 단일대치법은 각각의 결측치들을 각각 하나의 다른 값으로 대치하는 방법으로 Last observation carried forward(LOCF), Baseline observation carried forward(BOCF), Regression method, Hot-deck imputation 등이 있다. 하지만 단일대치법은 추정량의 표준오차를 작아지게 하는 방향으로 편의를 일으키게 할 가능성이 있다. 즉, 제1종 오류가 증가할 가능성이 있다. 다중대치법은 각각의 결측치들을 각각의 어떤 값으로 대치한 후, 마치 그 대치한 값들이 실제 관측한 값들인 것처럼 분석하는 방법이다. 하지만 다중대치법은 계산이 복잡하고 시간이 오래걸린다는 단점이 있다 (Kang, 2013; Yun, 2004).

여러 다양한 종류의 단일대치법들 중에서 Dixon (1979)과 Troyanskaya 등 (2001)에 의해 제안된 k-최근접이웃(k-Nearest Neighbors; KNN) 대치법은 결측이 발생한 개체와 가장 가까운 거리에 있는 k개의 이웃 개체들을 활용하여 결측치를 대치하는 방법이다. 이는 다변량 정규성 등의 모수적 모형이 만족되지 않을 때에도 강건성(robustness)을 지니며 그 계산 알고리즘이 간단하다는 장점을 바탕으로 널리 활용되고 있다 (Park 등, 2011).

[1]Corresponding author: Department of Biomedicine · Health Science, The Catholic University of Korea, Banpo-Dong, Seocho-Gu, Seoul 137-701, Korea. E-mail: djkim@catholic.ac.kr

그러나 KNN 대치법은 고정적인 수 k개의 평균값으로 결측치를 대치하므로 그 편차가 심하면, 편의를 일으킬 수 있다. 따라서 KNN의 국소적 특정이 고려되지 않는 단점을 보완한 k-means clustering (Jang, 2004), Adaptive Nearest Neighbors (Jhun 등, 2007), Sequential Adaptive Nearest Neighbors (Park 등, 2011) 대치법들이 제안되기도 하였다. 한편, Kim (2010)은 가중 k-최근접이웃방법을 이용한 통계적 매칭기법을 제안하였다.

본 논문에서는 가중 k-최근접이웃방법을 이용한 통계적 매칭기법의 장점을 KNN 대치법에 적용하여 k개의 최근접이웃들 중 극단치나 이상치가 있는 경우, 이들의 영향에 덜 민감하면서도 정확도를 높일 수 있는 가중 k-최근접이웃(Weighted k-Nearest Neighbors; WKNN) 대치법을 제안하고자 한다. 2장에서는 WKNN 대치법을 제안하고, 3장에서는 실제자료를 이용한 모의실험을 통해 기존의 KNN 대치법과 제안된 WKNN 대치법의 결과를 비교하였다. 또한 4장에서는 결론을 제시하였다.

2. 제안한 방법

본 논문에서 제안하는 가중 k-최근접이웃 대치법은 숫자형으로 이루어진 자료에서 유사성 거리를 반영하여 계산된 가중치를 이용하고, 결측치를 가중평균값(weighted mean)으로 대치하는 방법이다. 이는 k개의 최근접이웃들에 대한 거리에 반비례하여 거리가 가까운 최근접이웃들에 대해서는 큰 가중치를 부여하고, 상대적으로 거리가 먼 최근접이웃들에 대해서는 작은 가중치를 부여하는 것이다. 그리고 가까운 이웃들을 찾기 위해서 거리를 계산하고, 그 거리에 따른 가중치는 커널함수(Kernel function)를 이용하여 계산한다.

유사성 측정을 위한 거리함수는 Euclidean distance와 Hellinger distance를 사용한다. 일반적으로 Euclidean distance는 두 지점의 단순한 거리를 계산하고, Hellinger distance는 두 확률분포 사이의 유사도를 측정하는 방법이다. 그러나 Hellinger distance는 음수가 아닌 값만 사용해야 한다는 단점을 가지고 있다.

$$\text{Euclidean} : d_{(p,q)} = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}\ , \qquad \text{Hellinger} : d_{(p,q)} = \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{n}\left(\sqrt{p_i} - \sqrt{q_i}\right)^2}\ .$$

거리의 계산이 이루어지면 이를 바탕으로 k개의 최근접이웃들에 대한 가중치를 산정한다. 이때 거리의 분포는 데이터에 따라 달라지므로 데이터의 변화에 대해 유연성 있는 가중치 계산이 필요하다. 따라서 가장 가까운 $k+1$개의 최근접이웃을 찾아 이들에 대한 거리를 0과 1사이의 값으로 변환한다. 즉, 가장 가까운 관측치는 0의 거리를 갖게 되고, $k+1$번째 관측치는 1의 거리를 갖게 된다. $d_{(1)}$은 가장 가까운 거리이고, $d_{(k+1)}$은 $k+1$번째로 가까운 거리이다.

$$d_i' = \frac{d_i - d_{(1)}}{d_{(k+1)} - d_{(1)}}, \quad i = 1, 2, \ldots, k+1.$$

변환된 거리는 커널함수에 반영하여 $k+1$개 최근접이웃들에 대한 가중치를 계산한다. 마지막 $k+1$번째 관측치는 1의 거리를 갖게 되어 가중치는 0이 되도록 하며($w_{k+1} = 0$), 가중치들의 합은 1이다. 본 논문에서는 커널함수 중 첨도가 작은 Triweight 함수와 첨도가 큰 Epanechnikov 함수를 사용한다. Triweight 함수는 분포가 뾰족한 형태로 $k+1$개 최근접이웃들에 대한 가중치간의 차이를 가장 크게 할 수 있다. 즉, 가까이 있는 관측치에는 더 큰 가중치를 부여하고 멀리 있는 관측치에는 아주 작은 가중치를 부여하는 것이다. 여기서 커널함수를 이용하여 각 최근접이웃들에 대해 가중치를 계산하면 다음과

같다.

$$\text{Triweight} : W_j = \frac{\frac{35}{32}\left(1-d_i'^2\right)^3}{\sum_{i=1}^{k}\frac{35}{32}\left(1-d_i'^2\right)^3}, \quad 0 \le d_i' \le 1,\ j = 1,2,\ldots,k+1,$$

$$\text{Epanechnikov} : W_j = \frac{\frac{3}{4}\left(1-d_i'^2\right)}{\sum_{i=1}^{k}\frac{3}{4}\left(1-d_i'^2\right)}, \quad 0 \le d_i' \le 1,\ j = 1,2,\ldots,k+1.$$

본 논문에서는 거리계산하는 방법과 커널함수에 따라 4가지 방법(Euclidean-Triweight, Euclidean-Epanechnikov, Hellinger-Triweight, Hellinger-Epanechnikov)의 대치법을 제안한다.

n개의 관측개체와 p개의 변수를 가지고 있는 원자료 행렬을 D라고 할 때, 자료행렬 D를 D_m과 D_c로 나눈다. 여기서 D_m은 적어도 하나의 결측치가 포함되어 있는 r개의 관측치와 p개의 변수를 갖는 자료행렬이고, D_c는 원자료 행렬인 D에서 결측치가 포함되어 있지 않은 $n-r$개의 관측개체와 p개의 변수로 구성된 행렬이다.

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

결측치가 포함되어 있는 행렬 D_m에서 i번째 행 x_i와 D_c의 각 행들간의 거리를 계산한다. 단, x_i에서 결측치는 제외하고 관측치만 고려한다. 계산된 거리 값을 통해 가까운 $k+1$개의 개체들을 찾고 이들에 대한 거리를 0과 1사이의 값으로 변환한다. 가중평균값으로 대치하기 위해 변환된 거리를 커널함수에 대입하여 가중치를 계산한다. D_c의 관측치와 가중치를 각각 곱해서 더한 가중평균값(Weighted Mean)을 구하고, 이 값으로 D_m에 있는 결측치 x_{ij}^*를 대치한다.

예를 들어, D가 다음과 같고, $k = 2$, 거리함수는 Euclidean, 커널함수는 Triweight을 이용하여 결측치를 대치한다고 하자.

$$D = \begin{pmatrix} 2 & 4 & 9 & 6 & 5 \\ \cdot & 8 & 1 & 9 & 7 \\ 6 & 7 & 1 & 8 & 9 \\ 9 & \cdot & 6 & 2 & 3 \\ 1 & 4 & 8 & 5 & 6 \\ 9 & 2 & 6 & \cdot & 2 \\ \cdot & 3 & 9 & 4 & 6 \\ 8 & 2 & 5 & 3 & 2 \\ 5 & \cdot & 3 & \cdot & 7 \end{pmatrix} \text{를} \quad D_m = \begin{pmatrix} \cdot & 8 & 1 & 9 & 7 \\ 9 & \cdot & 6 & 2 & 3 \\ 9 & 2 & 6 & \cdot & 2 \\ \cdot & 3 & 9 & 4 & 6 \\ 5 & \cdot & 3 & \cdot & 7 \end{pmatrix} \text{과} \quad D_c = \begin{pmatrix} 2 & 4 & 9 & 6 & 5 \\ 6 & 7 & 1 & 8 & 9 \\ 1 & 4 & 8 & 5 & 6 \\ 8 & 2 & 5 & 3 & 2 \end{pmatrix} \text{로 나눈다.}$$

D_m에서 1번째 행 $x_1 = (\ \cdot\ 8\ 1\ 9\ 7\ )$과 D_c의 각 행들간의 거리를 Euclidean distance로 계산하면, x_1과 가장 가까운 행은 D_c의 두 번째 행이며, 그 거리는 2.45이다. 따라서 x_1은 D_c의 두 번째, 세 번째, 첫

번째, 네 번째 행의 순으로 가깝다.

$$\text{Distance} = \begin{pmatrix} (8-4)^2+(1-9)^2+(9-6)^2+(7-5)^2 \\ (8-7)^2+(1-1)^2+(9-8)^2+(7-9)^2 \\ (8-4)^2+(1-8)^2+(9-5)^2+(7-6)^2 \\ (8-2)^2+(1-5)^2+(9-3)^2+(7-2)^2 \end{pmatrix} = \begin{pmatrix} \mathbf{9.64} \\ \mathbf{2.45} \\ \mathbf{9.06} \\ 10.63 \end{pmatrix} = \begin{pmatrix} d_{(3)} \\ d_{(1)} \\ d_{(2)} \\ d_{(4)} \end{pmatrix}.$$

k는 2로 정하였으므로, $d_{(1)} = 2.45$와 $d_{(3)} = 9.64$를 이용하여 거리를 변환하고 변환된 거리를 Tri-weight 함수에 대입하여 가중치 w_{1v}를 계산한다. D_m에 있는 결측치 x_{11}^*을 대치하기 위해 x_1의 가중치와 D_c의 1열을 곱하여 가중평균값을 구한다.

$$\text{Distance}' = \begin{pmatrix} (9.64-2.45)/(9.64-2.45) \\ (2.45-2.45)/(9.64-2.45) \\ (9.06-2.45)/(9.64-2.45) \\ (10.63-2.45)/(9.64-2.45) \end{pmatrix} = \begin{pmatrix} 1 \\ \mathbf{0} \\ \mathbf{0.92} \\ 1.14 \end{pmatrix}, \quad \text{Weight} = \begin{pmatrix} 0 & \mathbf{0.996} & \mathbf{0.004} & 0 \end{pmatrix},$$

$$x'_{11} = \sum_{v=1}^{4}(w_{1v} \times x_{v1}) = \begin{pmatrix} 0 & \mathbf{0.996} & \mathbf{0.004} & 0 \end{pmatrix} \times \begin{pmatrix} 2 \\ 6 \\ 1 \\ 8 \end{pmatrix} = 5.98.$$

따라서 D_c의 두 번째와 세 번째 행의 가중평균값인 5.98로 결측된 부분을 대치한다. 이와 같은 방법으로 모든 결측치에 대해서 거리 Distance, 변환된 거리 Distance′, 가중치 Weight를 계산하고 가중평균값을 구하여 D_m을 $D_{imputed}$로 대치한다.

$$\text{Distance} = \begin{pmatrix} 9.64 & \mathbf{8.83} & \mathbf{8.43} & \mathbf{2.45} & 8.43 \\ \mathbf{2.45} & 10.30 & 10.39 & 10.25 & \mathbf{3.00} \\ \mathbf{9.06} & 9.27 & 9.38 & \mathbf{1.73} & 6.48 \\ 10.63 & \mathbf{2.00} & \mathbf{1.41} & 5.83 & \mathbf{6.16} \end{pmatrix}, \quad \text{Distance}' = \begin{pmatrix} 1 & \mathbf{0.94} & \mathbf{0.88} & \mathbf{1.18} & 1.149 \\ \mathbf{0} & 1.14 & 1.13 & 2.08 & \mathbf{0} \\ \mathbf{0.92} & 1 & 1 & \mathbf{0} & 1 \\ 1.14 & \mathbf{0} & \mathbf{0} & 1 & \mathbf{0.91} \end{pmatrix},$$

$$\text{Weight} = \begin{pmatrix} 0 & \mathbf{0.996} & \mathbf{0.004} & 0 \\ \mathbf{0.002} & 0 & 0 & \mathbf{0.998} \\ \mathbf{0.011} & 0 & 0 & \mathbf{0.989} \\ \mathbf{0.477} & 0 & \mathbf{0.523} & \mathbf{0} \\ 0 & \mathbf{0.995} & 0 & \mathbf{0.005} \end{pmatrix},$$

$$D_c = \begin{pmatrix} 2 & 4 & 9 & 6 & 5 \\ 6 & 7 & 1 & 8 & 9 \\ 1 & 4 & 8 & 5 & 6 \\ 8 & 2 & 5 & 3 & 2 \end{pmatrix}, \quad D_{imputed} = \begin{pmatrix} \mathbf{5.98} & 8 & 1 & 9 & 7 \\ 9 & \mathbf{2} & 6 & 2 & 3 \\ 9 & 2 & 6 & \mathbf{3.03} & 2 \\ \mathbf{1.48} & 3 & 9 & 4 & 6 \\ 5 & \mathbf{6.97} & 3 & \mathbf{7.97} & 7 \end{pmatrix}.$$

3. 모의실험의 계획 및 결과

기존의 방법과 제안된 방법을 비교하기 위한 모의실험에는 효모세포주기분석(Cellcycle)의 자료를 이용하였다(http://genome-www.stanford.edu/cellcycle). Cellcycle 자료는 34개의 변수와 7744개의 개체로 구성되어 있고, 34개의 변수 중 26개는 숫자형 변수이다. 이중 숫자형 변수 6개 RAT1, RAT2,

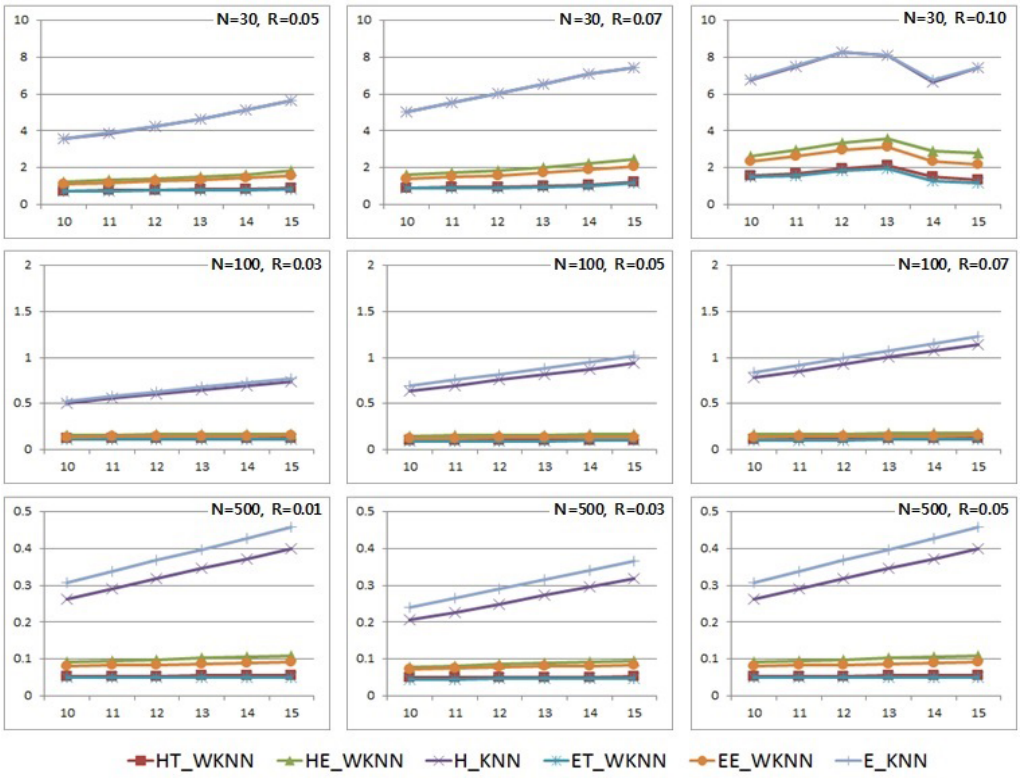

Figure 3.1. Result of NRMSE method

RATIN, RAT2N, CRT1, CRT2를 이용하여 모의실험을 수행하였다. 실제자료에서 적절한 크기의 표본을 만들고, 임의로 결측치를 발생시킨 후 대치하여 그 결과를 비교하였다. 이때 '정규화 제곱근 평균제곱오차(Normalized Root Mean Squared Error; NRMSE)'와 '실제자료의 검정결과 일치성'을 통해 KNN과 WKNN을 비교평가하였다.

정규화 제곱근 평균제곱오차는 실제 값과 대치된 값의 차이를 전체 결측치에 대해 계산하는 방법으로

$$\text{NRMSE} = \frac{1}{x'_{max} - x'_{min}} \left\{ \sum \frac{(x_{ij} - x'_{ij})^2}{M} \right\}^2$$

와 같이 정의된다. 이때 x_{ij}는 실제값, x'_{ij}는 대치된값, $x'_{\max}$은 대치된 값들 중에서 최대값, $x'_{\min}$은 대치된 값들 중에서 최소값을 나타내고, M은 결측치의 총 개수를 나타낸다.

모의실험을 하기 위해 6개의 변수로 구성된 Cellcycle 자료에서 크기가 30, 100, 500인 소표본, 중간표본, 대표본을 추출하였다. SAS 프로그램을 이용해서 균일분포에서 난수를 생성하였고, 결측비율을 소표본에서는 5%, 7%, 10%, 중간표본에서는 3%, 5%, 7%, 대표본에서는 1%, 3%, 5%로 정해서 결측치를 만들었다. k는 $10, 11, 12, 13, 14, 15$의 값에 따라 비교하였으며 100번 반복 수행하여 NRMSE들의 평균을 구하였다. 그 결과는 Figure 3.1과 같다. 그래프의 x축은 k, y축은 정규화 제곱근 평균제곱오차(NRMSE)이다. 최근접이웃의 갯수(k), 결측비율(R), 표본의 크기(N)에 따라 NRMSE를 비교하였다. 이때, NRMSE가 작을수록 더 정확한 방법이다.

KNN 방법의 NRMSE를 보면, Euclidean distance KNN(E-KNN) 방법과 Hellinger distance KNN(H-KNN) 방법이 모든 상황에서 비슷하게 나타났다. k가 10일때 KNN의 NRMSE는 WKNN의 NRMSE보다 $N = 30$인 경우에는 약 4배, $N = 100$인 경우에는 약 5배, $N = 500$인 경우 약 4배였다. 그리고 k, 결측비율, 표본크기에 상관없이 KNN 방법보다 WKNN 방법이 더 정확하였다.

WKNN 방법들만 비교하면 $N = 30$일 때, $R = 0.10$인 경우를 제외하고 k가 증가할수록, 결측비율이 증가할수록, NRMSE가 증가하였다. $N = 100$, $N = 500$인 경우에도 k와 결측비율은 NRMSE와 비례하였다. 하지만 변화하는 폭이 크지는 않았다. 그리고 표본의 크기가 커지면 NRMSE는 감소했다. 그러므로 WKNN 방법은 k가 작을수록, 결측비율이 작을수록, 표본의 크기가 클수록 더 정확하며, 표본이 큰 경우에는 NRMSE는 결측비율과 k에 민감하지 않은 것으로 나타났다.

WKNN 방법들을 비교했을 때, NRMSE의 결과는 모든 경우에서 Euclidean-Triweight-WKNN, Hellinger-Triweight-WKNN, Euclidean-Epanechnikov-WKNN, Hellinger-Epanechnikov-WKNN의 순서로 정확한 것을 확인할 수 있다. 세부적으로 보면 Triweight 함수가 Epanechnikov 함수보다 정확했다. 따라서 가중치를 더 크게 차이를 줄수록 정확하다고 할 수 있다. 그리고 Euclidean distance가 Hellinger distance보다 더 정확했다. 하지만 Hellinger distance는 음수를 사용하지 못하는 단점이 있으므로 Euclidean distance를 사용하는 것이 더 유용하다. 그러므로 정규화 제곱근 평균제곱오차 판별 방법에서 가장 좋은 방법은 Euclidean-Triweight-WKNN 방법이라고 볼 수 있다.

'실제자료의 검정결과의 일치성'으로 평가하기 위하여 6개의 변수로 구성된 Cellcycle 자료에서 임의로 크기가 100인 두 표본을 만들었다. 여기에서 두 표본의 독립 T-검정결과는 변수 RAT1와 RAT1N는 '두 군의 평균 차이가 없다'는 귀무가설이 기각되었고, 변수 RAT2, RAT2N, CRT1, CRT2는 귀무가설이 채택되었다. SAS프로그램에서 균일분포를 이용하여 결측비율을 3%, 5%, 7%로 정해서 결측치를 랜덤으로 생성하였고, k는 10의 값으로 고정하여 대치하였다. 1000번 반복 수행하여 검정결과가 원래의 결과와 일치한 수의 비율을 Table 3.1에 정리하였다. 여기에서는 대치법을 이용하지 않은 경우도 고려하였다.

귀무가설이 기각된 변수와 채택된 변수를 비교해보면, KNN 방법에서 귀무가설이 기각되었던 RAT1과 RAT1N이 다른 변수들에 비해 일치비율이 떨어졌다. 이 차이는 결측비율이 커질수록 더 크게 나타났다. 하지만 WKNN 방법에서는 귀무가설의 기각유무에 관계없이 결과가 거의 비슷하게 나타났다. 변수 CRT1에서만 Unimputation과 KNN이 WKNN보다 더 정확했는데, 변수 CRT1를 확인해보면 대부분의 변수가 1의 값을 가진다. 그 이유는 이 변수가 숫자형 변수보다는 명목형 변수의 성질이 있기 때문이다. 본 논문에서는 숫자형 변수의 경우만 가정했기 때문에 이러한 결과가 나온 것이다.

모든 방법들은 결측비율이 커질수록 정확도는 떨어졌지만, 변수 CRT1을 제외하고 결측비율에 상관없이 WKNN 방법, KNN 방법, Unimputation 방법의 순으로 실제자료의 검정 결과와 더 많이 일치했다. 그리고 KNN 방법에서 Euclidean distance와 Hellinger distance에는 차이가 거의 나타나지 않았다. WKNN 방법들을 비교하면 Hellinger distance가 Euclidean distance보다 더 정확했지만 차이는 크지 않았다. 그리고 Triweight 함수가 Epanechnikov 함수보다 더 정확했다. 따라서 '실제자료의 검정결과 일치성'에서 가장 좋은 방법은 Hellinger-Triweight WKNN 방법이었다.

4. 결론 및 고찰

본 논문에서는 KNN 대치법의 단점을 보완하기 위해 WKNN 대치법을 제안하였다. 그리고 제안한 WKNN 대치법은 distance와 커널함수에 따라 4가지로 나누고, KNN 대치법은 distance에 따라 2가지로 나눠서 총 여섯 방법을 모의실험을 통해 비교해 보았다.

Table 3.1. Equal ratio

T-test result of real data			Reject	Reject	Accept	Accept	Accept	Accept
R	Distance	Method	RAT1	RAT1N	RAT2	RAT2N	CRT1	CRT2
3%	Unimputation		0.871	0.850	0.902	0.896	1.000	0.947
	Euclidean	WKNN of T	0.953	0.967	0.998	1.000	0.967	1.000
		WKNN of E	0.947	0.961	0.996	0.998	0.967	0.999
		KNN	0.871	0.851	0.980	0.967	1.000	0.997
	Hellinger	WKNN of T	0.971	0.967	0.998	1.000	0.930	1.000
		WKNN of E	0.964	0.955	0.997	0.999	0.940	0.999
		KNN	0.871	0.851	0.980	0.965	1.000	0.997
5%	Unimputation		0.765	0.761	0.864	0.845	1.000	0.930
	Euclidean	WKNN of T	0.919	0.939	0.998	0.994	0.947	1.000
		WKNN of E	0.915	0.920	0.990	0.983	0.952	0.997
		KNN	0.765	0.766	0.961	0.937	1.000	0.994
	Hellinger	WKNN of T	0.943	0.942	0.998	0.996	0.899	1.000
		WKNN of E	0.931	0.922	0.991	0.984	0.921	0.997
		KNN	0.765	0.766	0.962	0.932	1.000	0.993
7%	Unimputation		0.689	0.678	0.843	0.834	1.000	0.914
	Euclidean	WKNN of T	0.891	0.914	0.990	0.988	0.928	1.000
		WKNN of E	0.887	0.892	0.972	0.961	0.935	0.998
		KNN	0.690	0.687	0.935	0.912	0.999	0.986
	Hellinger	WKNN of T	0.917	0.926	0.991	0.988	0.876	1.000
		WKNN of E	0.900	0.900	0.976	0.960	0.907	0.997
		KNN	0.690	0.686	0.934	0.911	1.000	0.986

최근접이웃의 개수, 결측비율, 표본의 크기에 상관없이 KNN 방법보다 WKNN 방법이 더 뛰어나다는 것을 확인하였다. 4가지 WKNN 방법들 중 Euclidean distance와 Hellinger distance 비교 결과 두 거리는 큰 차이가 없었고, Hellinger distance는 음수 값을 사용하지 못하는 단점이 있으므로 Euclidean distance를 사용하는 것이 더 효율적이라고 할 수 있겠다. 또한 커널함수는 가중치의 차이를 더 크게 준 Triweight 함수가 더 정확했다. 따라서 KNN 방법 보다는 WKNN 방법이 더 뛰어나고, 그 중 Euclidean-Triweight WKNN 방법이 가장 좋은 것으로 나타났다. 검정의 기각여부로 판단했을 때, 변수가 CRT1인 경우를 제외하고 WKNN 대치법이 더 정확했다. 이는 변수가 숫자형 변수보다는 명목형 변수의 성질을 가지기 때문이다.

앞으로 숫자형 변수뿐만 아니라 명목형 변수에도 적용할 수 있는 WKNN 방법을 고안해야 할 필요가 있다. 또한 본 논문에서는 KNN과 제안한 WKNN 방법을 비교하였는데, KNN의 단점을 다른 방식으로 보완했던 ANN 대치법, SANN 대치법 그리고 그 외에 다른 대치방법들과도 비교할 필요가 있다.

References

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

Jang, H. J. (2004). On the use of clustering method for missing value imputation, Korea University, M.S. Thesis.

Jhun, M. S., 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.

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

Kim, H. K. (2010). A study on statistical matching technique using the weighted k-nearest neighbor method, Dongguk University: Ph.D. thesis.

Park, S. H., Bang, S. W. and Jhun, M. S. (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. C. (2004). Imputation of missing values, *Journal of Preventive Medicine and Public Health*, **37**, 209–211.

Weighted k-Nearest Neighbors를 이용한 결측치 대치

임찬희[a] · 김동재[a,1]

[a]가톨릭대학교 의생명 · 건강과학과

(2014년 10월 6일 접수, 2014년 11월 10일 수정, 2014년 11월 10일 채택)

요 약

통계적 분석을 할 때 결측치가 발생하는 것은 매우 통상적이다. 이러한 결측치를 대치하는 방법은 여러가지가 있으며, 기존에 사용되는 단일대치법으로 k-nearest neighbor(KNN) 방법이 있다. 하지만 KNN 방법은 k개의 최근접 이웃들 중 극단치나 이상치가 있을 때 편의를 일으킬 수 있다. 본 논문에서는 KNN 방법의 단점을 보완하여 가중 k-최근접이웃(Weighted k-Nearest Neighbors; WKNN) 대치법을 제안하였다. 또한 모의실험을 통해서 기존의 방법과 비교하였다.

주요용어: 결측치, 대치법, k-최근접이웃.

[1]교신저자: (137-701) 서울 서초구 반포동, 가톨릭대학교 의생명 · 건강과학과. E-mail: djkim@catholic.ac.kr