韓國IT서비스學會誌
第13卷 第4號
2014年 12月, pp.359-375

Journal of Information Technology Services
http://dx.doi.org/10.9716/KITS.2014.13.4.359

K-nn을 이용한 Hot Deck 기반의 결측치 대체*

권 순 창**

Imputation of Missing Data Based on Hot Deck Method Using K-nn*

Soonchang Kwon**

■ Abstract ■

Researchers cannot avoid missing data in collecting data, because some respondents arbitrarily or non-arbitrarily do not answer questions in studies and experiments. Missing data not only increase and distort standard deviations, but also impair the convenience of estimating parameters and the reliability of research results.

Despite widespread use of hot deck, researchers have not been interested in it, since it handles missing data in ambiguous ways. Hot deck can be complemented using K-nn, a method of machine learning, which can organize donor groups closest to properties of missing data. Interested in the role of k-nn, this study was conducted to impute missing data based on the hot deck method using k-nn.

After setting up imputation of missing data based on hot deck using k-nn as a study objective, deletion of listwise, mean, mode, linear regression, and svm imputation were compared and verified regarding nominal and ratio data types and then, data closest to original values were obtained reasonably. Simulations using different neighboring numbers and the distance measuring method were carried out and better performance of k-nn was accomplished.

In this study, imputation of hot deck was re-discovered which has failed to attract the attention of researchers. As a result, this study shall be able to help select non-parametric methods which are less likely to be affected by the structure of missing data and its causes.

Keyword : Missing Data, Imputation, K-nn, Hot Deck

Submitted : April 25, 2014 1st Revision : December 10, 2014 Accepted : December 14, 2014

* 본 연구는 인천대학교 교내 연구비지원에 의하여 연구되었음.

** 인천대학교 무역학부 교수

1. 서 론

사회과학과 심리학 및 인공지능, 기계학습, 의학 분야의 연구조사와 실험에서 응답자들이 자의적 또는 비자의적으로 응답하지 않아 발생하는 결측치를(King, 2001) 피해 갈 수 없을 것이다. 대부분 분석에서 제외되는 결측치로 인한 정보 손실로 모수 추정의 편이가 발생해 검정력이 약화된다. 연구의 신뢰성을 높이려면 결측치 본연의 의미를 잃지 않도록(Batista and Monard, 2002; Christobel and Sivaprakasam, 2012; Finch, 2010) 결측치 구조에 따른 처리가 이루어져야만 한다(Zhang, et al., 2006). 결측치 대체 기술은 회귀 모형을 이용한 모수적 대체(parametric imputation)와 기계학습 기반의 비모수적 대체를 중심으로 발전해 왔다. 결측치 분포에 대한 사전 지식과 함께 모델링이 가능하면 모수적 대체가 효과적이지만, 그렇치 않은 경우는 비모수적 대체가 좋은 대안이 될 수 있다(Somasundaram and Nedunchezhian, 2011).

결측치 분포에 대한 불확실한 모수 추정은 또 다른 연구 주제가 될 수 있으며, 평균과 같은 특정한 값으로 이용하는 결정적 대체법은 샘플 집단의 분산을 축소시킨다. 반면에 hot deck 대체는 응답값 중에서 하나를 임의로 선택하여 대체하는 것으로 모수 추정에 필요한 명시적인 모델 없이 여러 가지 변수 형태에 쉽게 적용할 수 있으며, 대체 후에도 원자료의 분포를 유지할 수 있어 널리 사용되어 왔다. 그러나 기증자 풀 구성에 있어 확률적 처리 과정의 복잡성과 함께 정확도에 영향을 주는 기증자 선택 과정이 모호해 연구자들의 관심을 끌지 못했다(Somasundaram and Nedunchezhian, 2010).

k-nn(k-nearest neighbors)은 간단한 기계 학습이지만 결측치 속성과 가까운 기증자 그룹을 구성할 수 있어 hot deck 대체 문제점 해결에 도움을 줄 수 있다. 이러한 k-nn의 역할에 관심을 갖고 결측치 발생 원인과 구조에 영향을 덜 받는 비모수적 결측치 대체 방안의 제시를 연구 목표로 설정하였다.

본 연구를 진행하면서 연구 목표 달성을 위해 명목형과 연속형 변수를 대상으로 k-nn 대체와 주로 사용되고 있는 listwise 제거, 임의(random imputation), 평균, 최빈값, LR(Linear Regression) 모형 및 svm(support vector machine) 대체 효과를 시뮬레이션을 통해서 비교 검증하였다.

본 연구의 구성은 제 2장에서 결측치의 문제점과 구조 및 처리에 대하여 기술하고, 제 3장은 hot deck과 k-nn 처리과정, 제 4장에서는 실험 방법과 사용된 자료, 제 5장은 결론으로 각각 이루어졌다.

2. 결측치

2.1 결측치 문제

자료 집합에서 결측치 비율이 낮으면 이를 무시할 수 있으나, 높은 경우에 해당 레코드 제외시 이에 따른 정보의 손실과 편이(bias)가 발생한다. 따라서 결측치를 연구 설계와 처리에서 중요하게 다루어야 할 문제라고 언급하지만(Bennett, 2001), 복원이 불가능한 것처럼 보여 대부분의 연구에서 제외된다. 다수의 전문학술지에서 조차 최대 50% 이상 응답하지 않은 결측치를 확인했지만, 이에 대한 처리가 불명확하였다(Acock, 2005; King et al., 2001; Peng et al., 2006; Saunders et al., 2006; Schlomer et al., 2010). 자료의 일관성과 연구 결과의 신뢰성을 높이려면 결측치 처리과정이 분석에 포함되어야 한다.

2.2 결측치 구조

Rubin은 결측치 발생의 임의성 여부에 따라 MCAR(Missing Completely at Random), MAR (Missing at Random) MNAR(Missing Not At Random)으로 구분해서 결측치 처리에 대한 지침을 제공하였다(Rubin, 1976).

결측치 연구에서 빠지지 않고 등장하는 MCAR에서는 자료 집합을 직사각형 행렬로 보고 결측치

는 특정 속성에서 임의로 발생한다고 본다. 결측치는 결측치를 포함한 자료집합의 모든 속성과 관련이 없다는 매우 엄격한 가정으로(Acock, 2005; Bennett, 2001; Roth, 1994; Rubin, 1976) 연구 설계에서부터 결측치가 고려되지 않는 한 드물게 발생한다. 이러한 가정에 따라 결측치를 포함한 자료도 전체 자료집단의 랜덤 샘플로 인정한다.

직관적으로 MCAR와 유사하지만 최소한 완전한 임의가 아니라는 MAR에서는 결측된 속성과는 관련이 없고 자료집합 내의 다른 속성에 조건적으로 의존한다고 본다(Graham, et al., 2003; Schafer and Graham, 2002). 이러한 현실적인 가정에 따라 결측치와 관측치 간의 관계에서 결측치를 예측할 수 있는 자료 모델에 초점을 맞출 수 있다.

결측치 발생 확률이 임의가 아니라는 MNAR에서는 결측치가 결측된 속성에 의존함에 따라 무응답을 무시할 수 없는 가장 난해한 상태이다. MNAR 하에서는 관심의 대상이 되는 결측치와 결측치 구조를 동시에 충족시키는 모델 발견이 쉽지 않아(Carpenter, 2010) 통상적으로 결측치 모델은 추론을 통해서 개념적으로 입증된다(He et al., 1999).

3가지 결측치 구조 중에서 MCAR만이 경험적으로 검증될 수 있어 Little(1988)은 다수의 통계적 MCAR 검증기를 개발하였다. 많은 학자들이 MAR 형태를 지지하지만 이를 내용적으로 확인할 수 없으며, MNAR 역시 관측되지 않은 자료에 의존함으로 검증이 불가능하다. 따라서 결측치 구조 보다는 결측치 발생 차원에서 임의가 아닌 경우에 내재된 편이가 있는가를 식별하고(Baraldi and Enders, 2010) 이에 따른 타당한 분석이 이루어져야 할 것이다.

2.3 결측치 처리

지난 30년에 걸친 일반적인 결측치 처리 방법을 요약하면(Allison, 2002; Little, 1988; Little and Rubin, 2002; Schafer, 1997; Van Buuren, 2012) 제거와 대체라 할 수 있다(Cheng et al., 2013). 연구 모델에서 독립변수 또는 종속 변수에 나타날 수 있는 결측치를 포함한 레코드의 제거와 평균 또는 가장 그럴듯한 값으로 결측치를 대체하는 2가지 기준에 따라 자료집합을 구성한다. 각각의 접근 방식과 문제가 될 수 있는 상황을 살펴보면 다음과 같다.

Listwise 제거는 샘플이 충분히 크고 MCAR면 검정력에 문제가 없으나, 그렇지 않은 경우 샘플의 감소와 함께 주요 레코드 속성까지 제외됨에 따라 유의도와 검정력에 영향을 줄 수 있다. Pairwise 제거는 MAR 가정시 listwise 제거와 달리 적당한 공변량 구조와 함께 수집된 모든 정보를 활용할 수 있어야(Anderson et al., 2013) 왜곡 현상을 줄일 수 있다. 그러나 해당 레코드에서 결측치 속성만 제거시 샘플 그룹별로 크기가 달라져 다수의 공분산이 형성되 회귀모형에 적합하지 않고, 상관관계와 자유도 계산도 분명치 않아 통계 검증력이 약화된다. 자료가 MCAR를 충족하지 못하면 편이가 발생한다.

평균 대체는 무 결측치 값에 근거한 해당 속성의 평균값으로 결측치를 대체하는 것으로 MAR와 MNAR의 경우에 분산과 공분산이 저평가되고 변수들의 결측치 비율이 불일치시 평균에 편이가 발생해 Type II 에러가 증가된다(Allison, 2002; Devane et al., 2004).

회귀 모형대체는 결측치가 없는 데이터를 대상으로 작성된 회귀 모형으로 결측치를 예측한다. MCAR 또는 MAR에 대해 편이가 없는 추정치를 제공하지만 분산과 공분산에 있어 편이가 발생한다(Graham et al., 2003).

Vapnik(1998)에 의해 개발된 지도학습(supervised learning)의 일종인 SVM은 선형판별 함수를 사용할 수 있도록 복잡한 경계를 공유하는 주어진 자료를 고차원 공간에 매핑시켜 분류와 회귀 및 대체에 이용할 수 있다. SVM은 해석이 용이하고, 경험적 위험보다는 구조적 위험을 최소화해서 과대적합 문제를 완화할 수 있으며, 인공신경망 수준의 높은 정확도와 함께 샘플 수가 작아도 분류가 가

능하기 때문에 주목 받고 있다(Gunn, 1998; Kim and Ahn, 2011). 그러나 복잡한 처리 알고리즘과 kernel 함수 선택의 필요성, 이산화 과정 및 quadratic programming 처리에 요구되는 메모리와 시간 등은 svm의 단점으로 지적된다(Suykens, 2003).

유사한 데이터 사례로 결측치를 채우는 패턴 매칭 대체는 문헌에서 hot deck과 cold deck으로 설명하고 있으며, 특별한 프로그램을 필요로 하지 않아 설문 자료에서 주로 사용되어 왔다(Roth, 1994). Hot deck 대체에서는 결측치 레코드와 일치하는 다른 사례의 속성 값으로 결측치를 대체하는 것으로, listwise 제거 또는 평균 대체보다 적은 편이를 갖는 것을 관찰했다(Bennett, 2001; MacCallum et al., 2002). 기증자의 응답값으로 결측치를 대체함에 따라 변수의 수가 늘어남에 따라 같은 그룹에 속하는 기증자를 찾기가 힘든 점과 연속형 변수의 이산형 변환에 따른 정보 손실, 기증자 선택 방법에 따라 여러 가지 편이와 분산의 차이가 발생한다(Pettersson, 2012). Cold deck 대체는 결측치가 속하지 않은 다른 샘플 그룹에서 기증자를 선택함에 따라 외부 정보의 가용성에 의존하고, hot deck과 동일한 문제점을 갖고 있다.

제거는 주로 사용되는 것으로 유용한 정보를 잃게 되며 평균 방식의 대체는 데이터 분산에 상당한 왜곡을 초래할 수 있어 자료가 MCAR 상태인 경우에만 사용할 수 있다. 따라서 MAR와 MNAR에서는 회귀모형, 베이지안, 의사 결정 트리, 클러스터링 알고리즘, svm, k-nn 방식으로 편의와 오차 범위를 축소하면서(Baraldi and Enders, 2010) 결측치를 대체하고 있는 추세이다.

3. Hot Deck과 K-nn 처리 과정

3.1 Hot Deck 과정

샘플 조사에서 개별적인 자료보다는 샘플의 평균, 상관관계 및 회귀 계수와 같은 모수 추론이 주요 관심사가 된다. 따라서 대체의 목적은 결측치에 대한 최상의 예측보다는 그럴듯한 값으로 누락된 값을 채운 후에 연구 집단의 모수에 대한 추론을 하는 데 있다(Little and Rubin, 2002).

Hot deck 대체는 결측치가 없는 완전한 하위 샘플 그룹에서 무작위로 속성 값을 선정하여 결측치를 대체하는 random hot deck, 특정한 순서에 의해 정렬된 샘플 그룹에서 순서대로 속성 값을 이용하는 sequential hot deck과 가중치를 두어 속성 값을 선정하는 weighted hot deck으로 구분된다. Random hot deck은 샘플 그룹의 속성 값을 무작위로 선택하여 대체함에 따라 대체 후에도 표본의 분포가 그대로 유지될 수 있다는 장점이 있으나 결측치와 응답패턴이 무관한 경우는 흔치 않아 정확도가 떨어지게 되며, sequential hot deck은 동일한 레코드의 반복적인 사용에 따른 자료 분포의 왜곡으로 추정치 분산이 감소된다(Andridge and Little, 2010).

Hot deck 대체는 기증자의 속성 값으로 수여자의 누락된 속성 값을 대체하는 것으로 기증자와 수여자를 연결하는 방법은 일반적으로 세 단계로 구현된다. 첫 번째 단계에서, 데이터를 그룹으로 분할하고, 두 번째 단계에서, 누락 된 데이터를 유사 그룹의 레코드와 연관시킨다. 세 번째 단계에서, 결측치가 없는 유사 그룹 속성의 평균 또는 최빈값으로 누락 된 값을 대체한다.

3.2 K-nn과 기증자 분류

결측치와 유사한 하위 그룹의 분류가 관건인 hot deck 대체에서는 주어진 결측치에 근접한 기증자 분류에 k-nn을 사용할 수 있다. k-nn은 학습 샘플을 저장한 후 새로운 데이터가 올 때까지 분류하지 않고 기다린다. 이러한 lazy 학습 알고리즘으로 인해 학습시간 보다 예측시간이 더 소요되지만 간단하고 정확도가 높고 다양한 분류 문제에 적용되고 있다(Christobel and Sivaprakasam, 2012; Viswanath and Sarma, 2011; Yan, 2013). 이웃으로부터 배워야하기 때문에 이웃 간의 거리 측정법과 새로운 샘플 분류에 사용되는 이웃의 수에 의해

성능이 좌우된다(He et al., 1999). 이웃들 간의 거리는 euclidean, manhattan, minkowski 방식 등으로 계산되며, 가장 가까운 이웃의 평균 패턴에 따라 가중치를 설정해서 분류하는 weighted k-nn이나 bootstrapping 방식으로 분류 정확도를 개선할 수 있다(Baraldi and Enders, 2010; Viswanath and Sarma, 2011).

3.3 K-nn을 이용한 Hot Deck 기반의 결측치 대체 과정

k-nn을 이용한 hot deck 기반의 결측치 대체 과정은 원자료 집합에서 임의의 결측치 생성, 대체, 평가의 3단계 과정을 거치면서 진행된다.

• 결측치 생성 단계

random seed를 발생해서 임의의 결측치를 원자료에서 생성한다.

• 결측치 대체 단계

이 단계에서는 listwise 제거, random, 평균, 최빈값, k-nn, 선형회귀와 svm 방식에 의한 대체가 이루어진다.

본 연구에서 제안된 과정은 5-fold cross-validation 평균으로 구한 k 값으로 기증자 그룹을 선정한 후 대체한다. 이웃간의 거리 측정에 사용된 euclidean, manhattan, minkowski 수식은 다음과 같다.

$Euclidean\ distance$

$$D(p, q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

$Manhattan\ distance$

$$d(P, Q) = \| P - Q \| = \sum_{i=1}^{n}|p_i - q_i|$$

$P : p_1 \cdots p_n$

$Q : q_1 \cdots q_n$

$Minkowski\ distance$

$$(\sum_{i=1}^{n}|x_i - y_i|^p)^{1/p}$$

$P : x_i \cdots x_n$

$Q : y_i \cdots y_n$

• 평가 단계

대체전과 후의 평균, 표준편차, 상관관계와 t-test로 대체 방법별 차이를 비교하였다. 명목척도는 분류 정확도 Precision과 재현율인 Sensitivity와의 조화평균으로 구해지는 F-score, 비율척도는 모델의 기대 값과 실제 값의 차이를 종합한 RMSE로 정확도를 측정하였다. 정확도 측정에 사용된 수식은 다음과 같다.

$$F_{score} = 2TP/(2TP + FP + FN)$$

$TP : ture\ positive,\ FP : false\ positive,$

$FN : false\ negative$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{x_i} - x_i)^2}$$

$\hat{x_i}$: 기대값, x_i : 실제값

척도별 결측치 대체 실험 의사 코드는 <Table 1>, <Table 2>와 같다.

〈Table 1〉 Pseudo-Codes for Experiment of Nominal/Interval Scale Missing Data Imputation

```
Begin
 X = Original Data
 random seed generation
 for 5 to 50 increased by 5 : missing value generation ratio
   missing X = Randmom(X×i×0.5)
   ListDeletion(missingX) :
   RandomImputation(missingX) #random imputation
   MeanImputation(missingX) #mean imputation
   ModeImputation(missingX) #mode imputation
for 1 to 30 increased by 1 : for calculating optimal k
 for 1 to k-fold : 5-fold cross validation
```

```
    k-nn analysis with euclidean(weighted),
       manhattan(weighted), minkowski(weighted)
    find optimal k
  k-nn train accuracy with optimal k value
  k-nn imputation accuracy with optimal k value
    mean, sd, correlation, F-score after imputation
  compare k-nn, ListDeletion,
    Random, Mean, Mode Imputation
End
```

〈Table 2〉 Pseudo-Codes for Experiment of Ratio Scale Missing Data Imputation

```
Begin
  X = Original Data
  random seed generation

#missing value generation ratio
  for 5 to 50 increased by 5
    missingX = Randmom(X×i×0.5) #missing ratio 5%

from 1 to 30 increased by 1 : for calculating optimal k
  for 1 to k-fold : 5-fold cross validation
  #find optimal k
  k-nn analysis : euclidean, manhattan, minkowski

  k-nn train accuracy with optimal k value
  k-nn imputaion accuracy with optimal k value
  linear regression imputaion
  svm imputation
  compare k-nn imputaion, linear regression,
    svm imputation
End
```

4. 실험 설계 및 결과 분석

4.1 실험 설계

Hot deck 기반의 k-nn을 이용한 결측치 대체와 제거, 임의, 평균, 최빈값, LR, SVM 대체와 비교하면서 그동안 연구자의 주의를 끌지 못했던 결측치의 구조와 원인에 따른 적절한 처리 방법의 선택에 도움을 줄 수 있도록 설계하였다. 이웃의 수와 거리 측정 척도로 시뮬레이션을 하면서 어느 정도의 결측치 비율까지 이 기법을 적용할 수 있는 가를 검증하였다.

4.1.1 데이터

잘 알려진 명목, 등간, 비율척도로 구성된 자료를 k-nn의 대체 성능 평가에 사용하였다. 수집된 데이터에 임의로 결측치를 생성하고, 제안된 방법에 따라 대체한 후에 차이를 분석하였다.

(1) 실험 자료

● 명목척도

본 연구에서는 Martin이 공개한 부도 예측에 필요한 6개 속성과 175개 레코드로 구성된 Qualitative Bankruptcy(Martin et al., 2013) 자료 중 Bankruptcy와 Non-Bankruptcy로 분류되는 명목척도 <Table 3>의 Class 속성에 결측치를 할당하였다.

〈Table 3〉 Experiment Data Set

Data Set	No of Attributes	No of Records	missing values	type of scales
Qualitative Bankruptcy	6	175	Class	Nominal
	Other attributes value with missing data P : Positive, A : Average, N : Negative Attribute value of the missing values B : Bankruptcy, NB : Non-Bankruptcy			
Assertiveness	10	1005	AS10	interval
	Record Attribute value : 5 points likert scale			
iris	5	150	Sepal width	ratio
	Other attributes value with missing data norminal 1, ratio 3			
Boston Housing Price	14	506	MEDV	ratio
	Other attributes value with missing data norminal 1, ordinal 2, ratio 10			

● 등간척도

5점 리커트 척도와 1005개의 레코드로 구성된 http://personality-testing.info/_rawdata의 Assertiveness 성격 테스트 자료를 대상으로 하였다. Assertiveness 자료 중 <Table 3>의 AS10 항목에 결측치를 할당하였다.

• 비율척도

4개의 속성과 3종류의 iris로 구성된 150개의 Fisher's iris 자료와 보스턴 주변 지역의 주택 가격 중앙값과 주택 가격에 영향을 주는 13개의 측정지표로 구성된 506개 Boston Housing Price 자료를 대상으로 하였다. Iris 자료는 Sepal width 항목에 Boston Housing Price 자료는 보스턴 주변 지역의 주택 가격 중앙값 MEDV에 <Table 3>과 같이 결측치를 할당하였다.

(2) 분석 프로그램

Listwise 제거, 평균, 최빈값, random 대체와 k-nn에서 거리 측정 알고리즘 구현에 python과 python으로 작성된 기계학습 라이브러리 scikit-learn을 사용하였다.

(3) 결측치 생성과 K-NN 이웃의 수 결정

결측치 비율을 5%에서 50%까지 5%씩 증가하면서 원 자료에서 임의의 결측치를 추출했으며, 결측치 이외의 자료는 학습용으로 구분했다. 이웃의 수는 1-30개의 범위로 한정하였다.

(4) 대체 비교 실험

제 3.3절의 결측치 대체 의사 코드로 결측치를 대체한 후에 결측치와 원본 자료간의 차이를 t-test로 비교하고 정확도는 F-score와 RMSE로 평가하였다.

4.2 실험 결과 분석

4.2.1 명목/등간척도 속성

결측치가 없는 분석 자료 Qualitative Bankruptcy 척도의 Class 속성의 평균은 0.60, 표준편차는 0.49, Assertiveness 척도 AS10 속성의 평균은 1.76, 표준편차는 1.01로 <Table 4>에 제시되었다. 이 값을 기준으로 대체 방법별 차이를 비교하였다.

(1) Listwise 제거

Listwise 제거시 결측비율별 평균과 표준편차 및 t-test 결과는 <Table 5>에 제시되었다. Class의 평균은 0.60~0.64, 표준편차는 0.48~0.49, AS10의 평균은 1.76~1.83, 표준 편차는 1.02~1.07 범위에서 각각 분포되었다. AS10은 결측치 비율이 높을수록 원 자료 모수와의 차이가 확대되었다.

〈Table 4〉 Analysis Group Mean and SD

	Class	AS10
mean	0.60	1.76
sd	0.49	1.01

〈Table 5〉 Deletion of Listwise

	Class					A10				
missing %	5	10	15	20	25	5	10	15	20	25
mean	0.60	0.60	0.59	0.60	0.60	1.76	1.78	1.79	1.78	1.79
sd	0.49	0.49	0.49	0.49	0.49	1.02	1.02	1.03	1.03	1.04
	Class					A10				
missing %	30	35	40	45	50	30	35	40	45	50
mean	0.60	0.62	0.62	0.61	0.64	1.81	1.80	1.82	1.83	1.82
sd	0.49	0.49	0.49	0.49	0.48	1.05	1.06	1.07	1.07	1.06

Meaningful differences exist between the original data and substitute date from the missing values in the level of * 0.1, ** 0.05, *** 0.01.

(2) 대체 방법별 t-test 결과

명목/등간척도 결측치 비율과 대체 방법에 따른 결과는 <Table 6>~<Table 13>에 제시되었다.

Class 속성 분석 <Table 6> 결과를 보면 결측치 비율에 따른 결측치 자료와 원자료 간의 의미있는 차이가 random 대체에서는 없었으나, 평균과 최빈값 대체의 경우 25%부터 의미 있는 차이가 발생하였다. 이러한 차이는 결측치 비율이 높아지면서 평균 차는 더 확대되고 표준 편차는 축소되는 것으로 나타났다. t-test 결과 random 대체의 경우 결측치 비율에 따른 유의한 차이가 없었으며 원 자료와 유사한 평균과 표준편차를 갖는 것으로 분석되었다. Normal k-nn과 weighted k-nn 대체의 경우 <Table 9>와 같이 결측치 비율에 따른

차이가 없었으며, 최적 이웃의 수 k는 15% 결측치 비율의 weighted manhattan을 제외하고는 <Table 7>과 같이 동일하였다. 결측치 비율이 높아질수록 상관관계는 <Table 8>처럼 하락해서 random 대체의 경우 50% 결측치에서 0.38에 불과하였다. 반면에 k-nn 방식은 결측치 비율이 50%에서도 0.83 이상을 유지하였다.

〈Table 6〉 Results of t-test for Class Mean, Mode, and Random Imputation

%	mean mode		random	
	m	sd	m	sd
5	0.61	0.49	0.6	0.49
10	0.64	0.48	0.59	0.49
15	0.67	0.47	0.59	0.49
20	0.67	0.47	0.6	0.49
25	0.70*	0.46	0.6	0.49
30	0.74***	0.44	0.59	0.49
35	0.75***	0.43	0.66	0.47
40	0.75***	0.43	0.63	0.48
45	0.77***	0.42	0.63	0.48
50	0.81***	0.40	0.65	0.48

Meaningful differences exist between the original data and substitute date from the missing values in the level of * 0.1, ** 0.05, *** 0.01.

〈Table 7〉 Optimal k of Class Imputation

%	normal k-nn	weighted k-nn	
	euclidean manhattan minkowski	euclidean minkowski	manhattan
5	9	7	7
10	9	3	3
15	3	4	5
20	12	5	5
25	7	7	7
30	13	5	5
35	10	7	7
40	17	9	9
45	8	2	2
50	15	9	9

〈Table 8〉 Correlation of Class imputation

%	mean mode	random	normal k-nn			weighted k-nn		
			e	ma	mi	e	ma	mi
5	0.98	0.95	0.98	0.98	0.98	0.99	0.99	0.99
10	0.92	0.92	0.96	0.96	0.96	0.98	0.98	0.98
15	0.85	0.85	0.95	0.95	0.95	0.97	0.95	0.97
20	0.86	0.86	0.92	0.92	0.92	0.92	0.92	0.92
25	0.81*	0.76	0.92	0.92	0.92	0.92	0.92	0.92
30	0.73***	0.73	0.87	0.87	0.87	0.92	0.92	0.92
35	0.70***	0.70	0.88	0.88	0.88	0.92	0.92	0.92
40	0.70***	0.68	0.90	0.90	0.90	0.90	0.90	0.90
45	0.67***	0.65	0.86	0.86	0.86	0.81	0.81	0.81
50	0.60***	0.38	0.83	0.83	0.83	0.84	0.84	0.84

Meaningful differences exist between the original data and substitute date from the missing values in the level of * 0.1, ** 0.05, *** 0.01.

〈Table 9〉 Result of t-test for Class k-nn Imputation

%	normal k-nn						weighted k-nn					
	euclidean		manhattan		minkowski		euclidean		manhattan		minkowski	
	m	sd	m	sd	m	sd	m	sd	m	sd	m	sd
5	0.6	0.49	0.6	0.49	0.6	0.49	0.59	0.49	0.59	0.49	0.59	0.49
10	0.59	0.49	0.59	0.49	0.59	0.49	0.59	0.49	0.59	0.49	0.59	0.49
15	0.59	0.49	0.59	0.49	0.59	0.49	0.58	0.49	0.59	0.49	0.58	0.49
20	0.57	0.49	0.57	0.49	0.57	0.49	0.58	0.49	0.58	0.49	0.58	0.49
25	0.57	0.49	0.57	0.49	0.57	0.49	0.56	0.50	0.56	0.50	0.56	0.50
30	0.57	0.49	0.57	0.49	0.57	0.49	0.58	0.49	0.58	0.49	0.58	0.49
35	0.55	0.50	0.55	0.50	0.55	0.50	0.58	0.49	0.58	0.49	0.58	0.49
40	0.56	0.50	0.56	0.50	0.56	0.50	0.55	0.50	0.55	0.50	0.55	0.50
45	0.53	0.50	0.53	0.50	0.53	0.50	0.57	0.50	0.57	0.50	0.57	0.50
50	0.52	0.50	0.52	0.50	0.52	0.50	0.53	0.50	0.53	0.50	0.53	0.50

Meaningful differences exist between the original data and substitute date from the missing values in the level of * 0.1, ** 0.05, *** 0.01.

AS10 분석 결과 <Table 10>을 보면 결측치 비율에 따른 차이가 random 대체에서는 없었으나, mode 대체는 10%, mean 대체는 35%부터 유의한 차이를 보여 주었다. 표준 편차는 평균 대체에서 가장 낮았으며, random 대체에서는 원자료 표준편차에 근접한 것으로 나타났다. Normal k-nn을 이용한 대체 결과 <Table 13>에서 보면 20% 결측

치부터 의미 있는 차이가 발생해서 결측치 비율이 높아질수록 평균과 표준편차가 감소하였다. Weighted k-nn 대체는 25%부터 유의한 차이가 발생했지만 normal k-nn과 유사한 결과를 보여 주었다. 최적 이웃의 수 k는 euclidean과 minkowski 방식은 동일했지만 manhattan과는 결측치 비율에 따라 <Table 11>처럼 차이가 있었다. 상관관계는 <Table 12>에 제시된 것처럼 결측치 비율에 반비례했으며, 평균과 k-nn 대체에서 70% 이상의 상관관계를 유지한 반면에 50% 결측치 수준에서 random 대체의 상관관계는 0.49에 불과하였다.

〈Table 10〉 Results of t-test of AS10 Mean, Mode, and Random Imputation

%	mean		mode		random	
	m	sd	m	sd	m	sd
5	1.77	0.99	1.72	1.00	1.77	1.02
10	1.78	0.95	1.68*	0.98	1.74	0.99
15	1.79	0.94	1.64***	0.97	1.76	1.02
20	1.81	0.92	1.61***	0.97	1.76	1.03
25	1.81	0.88	1.56***	0.93	1.73	1.00
30	1.82	0.85	1.52***	0.91	1.71	0.99
35	1.84**	0.82	1.49***	0.89	1.74	0.99
40	1.86**	0.79	1.46***	0.87	1.80	1.03
45	1.89***	0.76	1.44***	0.86	1.82	1.06
50	1.88***	0.73	1.38***	0.81	1.75	1.02

Meaningful differences exist between the original data and substitute date from the missing values in the level of * 0.1, ** 0.05, *** 0.01.

〈Table 11〉 Optimal K Value of AS10 Attribute

%	normal k-nn		weighted k-nn	
	euclidean minkowski	manhattan	euclidean minkowski	manhattan
5	27	28	26	27
10	20	13	20	30
15	30	30	29	18
20	21	19	24	29
25	25	26	27	25
30	24	20	30	23
35	22	28	29	22
40	21	24	21	24
45	22	29	27	29
50	13	24	19	25

〈Table 12〉 Correlation to AS10 Imputation

%	mean	mode	random	normal k-nn			weighted k-nn		
				e	ma	mi	e	ma	mi
5	0.98	0.97	0.95	0.98	0.98	0.98	0.97	0.98	0.97
10	0.94	0.91*	0.89	0.94	0.93	0.94	0.94	0.93	0.94
15	0.92	0.89***	0.88	0.91	0.92	0.91	0.89	0.90	0.89
20	0.90	0.86***	0.83	0.90	0.90	0.90	0.89	0.89	0.89
25	0.86	0.80***	0.77	0.84	0.84	0.84	0.84	0.83	0.84
30	0.83	0.76***	0.70	0.80	0.81	0.80	0.80	0.80	0.80
35	0.80**	0.73***	0.63	0.79	0.78	0.79	0.78	0.77	0.78
40	0.77**	0.70***	0.54	0.77	0.75	0.77	0.77	0.75	0.77
45	0.74**	0.69***	0.55	0.74	0.73	0.74	0.71	0.71	0.71
50	0.70**	0.63***	0.49	0.70	0.70	0.70	0.68	0.68	0.68

Meaningful differences exist between the original data and substitute date from the missing values in the level of * 0.1, ** 0.05, *** 0.01.

(3) 정확도 검증 결과

정확도 검증 결과는 <Table 14>와 <Table 15>에 제시되었다.

Class 속성 대체 성능은 k-nn, 평균, 최빈값, random 순으로 3가지 거리 측정 방식에 따른 유의한 차이는 없었지만 normal k-nn보다 weighted k-nn 정확도가 약간 높은 것으로 나타났다.

AS10 속성 결측치 대체 성능은 random, 최빈값, 평균 순이며, 평균 대체는 35% 최빈값 대체는 10%부터 차이가 발생하고 35% 이상부터는 모두 차이가 발생하였다. mode 대체는 45% random은 10%에서 비교적 높은 것으로 분석되었다. k-nn 대체는 결측치 비율이 높아도 F-score에는 큰 변화가 없었으며, 평균, 최빈값, random 대체에 의한 것 보다 25% 이상 높았다. Manhattan 거리 측정 방식이 정확도가 약간 높은 것으로 분석되었다.

이상의 결과를 살펴보면 최빈값 대체의 경우 정확도는 높았지만 분산과 표준편차도 동시에 상승하였고 이웃의 수 k에 따라 정확도에도 차이가 존재하는 것으로 확인되었다. k-nn 기반의 hot deck

〈Table 13〉 Result of t-test for AS10 K-nn Imputation

%	normal k-nn						weighted k-nn					
	eucli dean		manha ttan		minko wski		eucli dean		manha ttan		minko wski	
	m	sd	m	sd	m	sd	m	sd	m	sd	m	sd
5	1.74	1.00	1.74	1.00	1.74	1.00	1.75	1.00	1.75	1.00	1.75	1.00
10	1.72	0.97	1.72	0.97	1.72	0.97	1.72	0.97	1.72	0.97	1.72	0.97
15	1.71	0.96	1.71	0.96	1.71	0.96	1.73	0.98	1.72	0.98	1.73	0.98
20	1.68*	0.96	1.68*	0.96	1.68*	0.96	1.69	0.96	1.69	0.96	1.69	0.96
25	1.65**	0.92	1.64***	0.92	1.65**	0.92	1.66**	0.92	1.65**	0.92	1.66**	0.92
30	1.62***	0.91	1.60***	0.90	1.62***	0.91	1.63***	0.91	1.62***	0.91	1.63***	0.91
35	1.63***	0.89	1.62***	0.88	1.63***	0.89	1.64***	0.89	1.62***	0.89	1.64***	0.89
40	1.62***	0.86	1.61***	0.86	1.62***	0.86	1.64***	0.87	1.62***	0.87	1.64***	0.87
45	1.65***	0.85	1.64***	0.85	1.65***	0.85	1.65***	0.86	1.64***	0.86	1.65***	0.86
50	1.58***	0.83	1.56***	0.82	1.58***	0.83	1.59***	0.84	1.57***	0.83	1.59***	0.84

Meaningful differences exist between the original data and substitute date from the missing values in the level of * 0.1, ** 0.05, *** 0.01.

대체가 평균, random, 최빈값 대체에 비해 정확도가 높은 것으로 분석되었다.

〈Table 14〉 F-score of Class Attribute Imputation

Category	without k-nn		normal	weighted	
%	mean mode	radom	euclidean manhattan minkowski	euclidean minkowski	manhattan
5	0.68	0.56	0.78	0.9	0.9
10	0.46	0.61	0.83	0.89	0.89
15	0.35	0.52	0.85	0.89	0.85
20	0.52	0.66	0.81	0.8	0.8
25	0.47*	0.55	0.84	0.84	0.84
30	0.39***	0.57	0.79	0.87	0.87
35	0.41***	0.57	0.82	0.89	0.89
40	0.47***	0.61	0.87	0.87	0.87
45	0.47***	0.63	0.84	0.8	0.8
50	0.44***	0.40	0.82	0.83	0.83

Meaningful differences exist between the original data and substitute date from the missing values in the level of * 0.1, ** 0.05, *** 0.01.

4.2.2. 비율척도 자료

비율척도 자료의 평균과 표준편차 및 k-nn, LR, svm 대체의 t-test 결과는 <Table 16>~<Table 26>에 제시되었다. Iris의 평균은 3.05, 표준 편차는 0.43, boston의 평균은 22.53, 표준편차는 9.19로 각각 분포되었다.

(1) Listwise 제거

Listwise 제거시 결측비율별 평균과 표준편차는 <Table 17>에 제시되었다. Sepal width의 평균은 3.04~3.11, 표준편차는 0.42~0.44, Boston Housing Price MEDV의 평균은 22.52~23.05, 표준 편차는 9.15~9.62범위에서 각각 분포되었다. 결측치 비율이 높을수록 원 자료 모수와 차이가 발생하는 것으로 분석되었다.

(2) 대체 방법별 t-test 결과

<Table 18>의 iris 분석 결과를 보면 LR과 SVM 대체에서 결측치 비율에 따른 의미 있는 차이가 없었으나, 결측치 비율이 증가함에 따라 평균은 증가하고 표준편차는 감소하였다. SVM 대체 상관관계가 LR보다 높았으며 LR은 40%, SVM은 50% 결측치 비율까지 0.9 이상의 높은 상관관계를 보여주었다. 최적 이웃의 수 k는 <Table 19>처럼 결

〈Table 15〉 F-score of AS10 Imputation

Category	without k-nn			normal			weighted		
method / %	mean	mode	random	euclidean	manhattan	minkowski	euclidean	manhattan	minkowski
5	0.22	0.32	0.32	0.49	0.57	0.49	0.52	0.57	0.52
10	0.21	0.29*	0.39	0.53	0.52	0.53	0.53	0.51	0.53
15	0.19	0.31***	0.35	0.48	0.48	0.48	0.46	0.47	0.46
20	0.16	0.33***	0.37	0.51*	0.50*	0.51*	0.49	0.46*	0.49
25	0.20	0.29***	0.35	0.51**	0.49***	0.51**	0.50**	0.47**	0.50**
30	0.23	0.27***	0.36	0.46***	0.48***	0.46***	0.46***	0.46***	0.46***
35	0.17**	0.32***	0.36	0.49***	0.48***	0.49***	0.48***	0.48***	0.48***
40	0.17**	0.33***	0.37	0.52***	0.51***	0.52***	0.52***	0.50***	0.52***
45	0.15***	0.37***	0.36	0.51***	0.49***	0.51***	0.51***	0.49***	0.51***
50	0.17***	0.32***	0.31	0.51***	0.48***	0.51***	0.49***	0.47***	0.49***

Meaningful differences exist between the original data and substitute date from the missing values in the level of * 0.1, ** 0.05, *** 0.01.

〈Table 16〉 Mean Add SD of Ratio Scale Data

	iris	boston
mean	3.05	22.53
sd	0.43	9.19

〈Table 17〉 Deletion of Listwise

var	iris Sepal width					Boston Housing Price MEDV				
missing %	5	10	15	20	25	5	10	15	20	25
mean	3.05	3.06	3.06	3.05	3.05	22.52	22.65	22.67	22.61	22.61
sd	0.42	0.43	0.43	0.43	0.43	9.15	9.17	9.27	9.23	9.24
	iris Sepal width					Boston Housing Price MEDV				
missing %	30	35	40	45	50	30	35	40	45	50
mean	3.04	3.07	3.10	3.11	3.09	22.75	22.93	22.75	22.97	23.05
sd	0.44	0.43	0.43	0.43	0.42	9.21	9.35	9.24	9.42	9.62

Meaningful differences exist between the original data and substitute date from the missing values in the level of * 0.1, ** 0.05, *** 0.01.

〈Table 18〉 Result of T-test for Iris LR, svm Imputation

%	LR			svm		
	m	sd	cor	m	sd	cor
5	3.06	0.43	0.99	3.06	0.43	0.99
10	3.06	0.42	0.99	3.06	0.42	0.98
15	3.06	0.40	0.95	3.05	0.40	0.96
20	3.07	0.40	0.95	3.06	0.40	0.95
25	3.06	0.39	0.93	3.05	0.39	0.94
30	3.07	0.38	0.91	3.06	0.38	0.93
35	3.08	0.37	0.90	3.07	0.38	0.92
40	3.07	0.36	0.90	3.08	0.37	0.91
45	3.08	0.36	0.89	3.08	0.36	0.91
50	3.08	0.36	0.89	3.10	0.36	0.90

Meaningful differences exist between the original data and substitute date from the missing values in the level of * 0.1, ** 0.05, *** 0.01.

〈Table 19〉 Optimal K Value for Iris and Boston

%	iris			boston		
	euclidean	manhattan	minkowski	euclidean	manhattan	minkowski
5~50	11	13	11	4	2	4

측치 비율에 관계없이 동일했으며, <Table 20>의 k-nn을 이용한 대체 결과를 보면 euclidean과 manhattan 방식은 40%까지 minkowski는 50%까지 0.9 이상으로 나타나 결측치 비율에 덜 민감한 것을 알 수 있었다.

<Table 21>의 Boston t-test 결과에 의하면 결

측치 비율에 따른 의미 있는 차이는 0.01수준에서 LR에서 없었지만, SVM에서는 20%부터 차이가 발생해 결측치 비율이 증가할수록 평균과 표준편차 범위가 확대되었다. 상관관계는 50% 결측치까지 LR에서 0.9 이상을 유지했으나 SVM에서는 5% 결측치일 때 최대 0.68이며 40% 이상부터는 음의 상관관계로 나타났다.

〈Table 20〉 Results of T-test for Iris K-nn Imputation

%	euclidean			manhattan			minkowski		
	m	sd	cor	m	sd	cor	m	sd	cor
5	3.06	0.43	0.99	3.06	0.43	0.99	3.06	0.43	0.99
10	3.06	0.42	0.97	3.05	0.43	0.99	3.06	0.42	0.98
15	3.05	0.40	0.94	3.05	0.43	0.95	3.05	0.40	0.96
20	3.06	0.40	0.94	3.07	0.39	0.95	3.06	0.40	0.95
25	3.05	0.39	0.93	3.06	0.39	0.93	3.05	0.39	0.94
30	3.06	0.38	0.92	3.07	0.38	0.91	3.06	0.38	0.93
35	3.07	0.37	0.91	3.08	0.37	0.90	3.07	0.38	0.92
40	3.07	0.35	0.90	3.07	0.37	0.90	3.08	0.37	0.91
45	3.08	0.35	0.89	3.08	0.36	0.89	3.08	0.36	0.91
50	3.09	0.35	0.89	3.08	0.36	0.89	3.10	0.36	0.90

Meaningful differences exist between the original data and substitute date from the missing values in the level of * 0.1, ** 0.05, *** 0.01.
Original data : mean 3.054, sd 0.432.

〈Table 21〉 Results of T-test for Boston LR, svm Imputation

%	LR			svm		
	m	sd	cor	m	sd	cor
5	22.57	9.17	1.00	22.04	12.66	0.68
10	22.47	9.06	0.98	23.47	15.64	0.57
15	22.49	9.02	0.97	22.85	13.20	0.67
20	22.34	8.82	0.96	24.95**	22.40	0.53
25	22.29	8.87	0.95	19.43*	35.36	0.28
30	22.14	8.58	0.93	23.49	22.62	0.40
35	22.20	8.600	0.93	14.60***	54.84	0.37
40	22.17	8.59	0.92	11.79***	35.64	-0.02
45	22.12	8.55	0.92	25.75**	28.40	0.34
50	21.97	8.62	0.91	28.66***	40.89	-0.18

Meaningful differences exist between the original data and substitute date from the missing values in the level of * 0.1, ** 0.05, *** 0.01.

k-nn을 이용한 대체 결과 <Table 22>를 보면 25% 결측치까지 0.93 이상의 상관관계를 유지했으며 결측치 비율이 50%로 상승해도 완만하게 감소하면서 0.83 이상을 보여주고 있다.

〈Table 22〉 Results of T-test for Boston K-nn Imputation

%	euclidean			manhattan			minkowski		
	m	sd	cor	m	sd	cor	m	sd	cor
5	22.52	9.19	0.99	22.53	9.23	1.00	22.52	9.19	0.99
10	22.43	9.01	0.98	22.44	9.06	0.98	22.43	9.00	0.98
15	22.54	8.91	0.97	22.46	8.89	0.97	22.54	8.91	0.97
20	22.52	8.81	0.94	22.45	8.80	0.94	22.52	8.81	0.94
25	22.51	8.88	0.93	22.40	8.84	0.93	22.51	8.88	0.93
30	22.29	8.55	0.88	22.22	8.55	0.89	22.29	8.55	0.88
35	22.41	8.53	0.86	22.28	8.48	0.89	22.41	8.53	0.86
40	22.38	8.34	0.85	22.26	8.38	0.88	22.38	8.34	0.85
45	22.43	8.30	0.84	22.31	8.34	0.87	22.43	8.30	0.84
50	22.38	8.15	0.83	22.31	8.31	0.86	22.38	8.15	0.83

Meaningful differences exist between the original data and substitute date from the missing values in the level of * 0.1, ** 0.05, *** 0.01.
Original data : mean 3.054, sd 0.432.

(3) 정확도 검증 결과

<Table 23>과 <Table 24>에서 Sepal width RMSE는 LR 0.24~0.35, SVM 0.22~0.32, euli-dean과 minkowski는 0.28~0.37, manhattan은 0.28

〈Table 23〉 Results of Iris LR, svm Imputation of RMSE

	5		10		15		20		25	
	L	S	L	S	L	S	L	S	L	S
train data	0.30	0.25	0.30	0.24	0.29	0.23	0.29	0.23	0.29	0.23
missing data	0.27	0.22	0.24	0.25	0.35	0.32	0.31	0.30	0.32	0.30
	30		35		40		45		50	
	L	S	L	S	L	S	L	S	L	S
train data	0.29	0.23	0.29	0.24	0.30	0.24	0.30	0.24	0.31	0.26
missing data	0.33	0.29	0.31	0.28	0.31	0.29	0.30	0.28	0.29	0.28

~0.36 범위에 걸쳐 있었다. k-nn 대체와 달리 LR에서는 10%, SVM에서는 5%까지는 학습용 RMSE가 결측치 자료보다 높은 것으로 나타났다.

〈Table 24〉 Results of iris k-nn imputation of RMSE

	5			10			15			20			25		
	e	ma	mi	e	ma	mi	e	ma	mi	e	ma	mi	e	ma	mi
train data	0.25	0.26	0.25	0.25	0.25	0.25	0.23	0.24	0.23	0.24	0.25	0.24	0.24	0.24	0.24
missing data	0.28	0.28	0.28	0.31	0.31	0.31	0.37	0.36	0.37	0.33	0.32	0.33	0.32	0.32	0.32
	30			35			40			45			50		
	e	ma	mi	e	ma	mi	e	ma	mi	e	ma	mi	e	ma	mi
train data	0.24	0.25	0.24	0.24	0.26	0.24	0.24	0.26	0.24	0.26	0.27	0.26	0.27	0.28	0.27
missing data	0.31	0.31	0.31	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.29	0.29	0.29

Boston Housing Price <Table 25>와 <Table 26>에 의하면 MEDV RMSE는 euclidean 4.91~8.13, manhattan은 5.21~7.54, minkowski 5.21~8.13, LR은 5.4~6.2, SVM은 25.4~89.0의 범위에 걸쳐 있었다. k-nn과 LR 방식에서는 학습용 자료의 RMSE가 결측치 자료보다 낮았지만 SVM에서는 반대 현상이 발생해 LR, k-nn에 비해 높은 것으로 분석되었다.

〈Table 25〉 Results of Boston LR, svm Imputation of RMSE

	5		10		15		20		25	
	L	S	L	S	L	S	L	S	L	S
train data	4.5	54.0	4.5	54.0	4.5	33.6	4.3	42.7	4.3	70.9
missing data	6.0	40.6	6.0	40.6	5.7	25.4	5.9	43.0	5.7	68.1
	30		35		40		45		50	
	L	S	L	S	L	S	L	S	L	S
train data	4.0	33.6	4.0	73.7	4.1	59.2	4.2	37.8	4.0	58.3
missing data	6.2	37.8	5.9	89.0	5.6	60.8	5.4	40.0	5.5	62.1

〈Table 26〉 Results of Boston K-nn Imputation of RMSE

	5			10			15			20			25		
	e	ma	mi	e	ma	mi	e	ma	mi	e	ma	mi	e	ma	mi
train data	4.35	3.04	4.35	4.56	3.04	4.56	4.40	2.93	4.40	4.22	3.02	4.22	4.32	3.04	4.32
missing data	4.91	5.21	4.91	5.11	5.21	5.11	5.89	5.72	5.89	6.79	7.01	6.79	6.73	6.70	6.73
	30			35			40			45			50		
	e	ma	mi	e	ma	mi	e	ma	mi	e	ma	mi	e	ma	mi
train data	3.91	2.82	3.91	3.96	2.82	3.96	4.11	2.83	4.11	4.28	3.04	4.28	4.22	3.11	4.22
missing data	8.13	7.54	8.13	7.99	7.20	7.99	7.76	7.07	7.76	7.49	6.82	7.49	7.34	6.57	7.34

4.2.3 분석결과 종합

명목/등간, 비율 척도 대체 분석 결과를 살펴보면 결측치 비율이 상승하면 평균과 최빈값 대체의 경우 원 자료 모수와 차이가 발생하고 상관관계도 낮아짐을 알 수 있다. 랜덤 대체의 경우 평균과 표준 편차는 원자료에 근접했으나 상관관계가 급락해 역시 원자료와 차이를 보여 주었다. 반면에 k-nn의 경우는 결측치 비율에 관계없이 원자료 모수와의 차이가 적고 상관관계와 정확도 및 F-score가 일정 수준을 유지하면서 높게 나왔다.

비율척도 iris 자료 분석 결과를 보면 SVM의 대체 정확도가 가장 높았으며, 20% 이하 결측치 비율만 제외하고는 k-nn이 LR보다 높은 것으로 분석되었다. SVM과 k-nn의 RMSE 범위는 LR보다 좁았으며, 거리 측정 방식에 따른 유의한 차이는 존재하지 않았다. Boston 자료에서 LR과 k-nn은 결측치 비율에 영향을 덜 받았지만 SVM의 오차가 급격히 증가한 점은 주목할 만하다.

이상의 결과를 살펴보면 k-nn 기반의 hot deck 대체가 평균, random, 최빈값 대체에 비해 오차가 작았으며, 결측치 비율과 자료 형태에 관계없이 SVM에 필적할 수 있는 예측력을 보여주었다.

개별적인 정확한 값보다는 그럴듯한 값으로 대체하는 비모수적 기법이 샘플 모수 추론에 더 합리적이라 할 수 있다. 결측치 속성이 2개 이상일

경우는 결측치가 없는 완전한 자료를 갖고 위의 과정을 되풀이하면 될 것이다.

5. 결론 및 연구의 한계점

연구자가 심혈을 기울여 설문지를 작성해도 응답자들이 응답하지 않아 발생하는 결측치를 피해 갈 수 없을 것이다. 결측치는 모수 추정의 편이와 함께 표준 오차를 왜곡시켜 연구 결과의 신뢰성을 저해하는 요인이 되고 있어 결측치의 구조에 따라서 적절한 통계적 검증 과정을 거쳐야 할 것이다.

Hot deck 대체는 실제로 널리 사용되고 있지만 관련 이론이 부족한 실정으로(Andridge and Little, 2010) 이를 어떻게 적용하는 가에 대한 출발점을 제시하였다.

본 연구에서 소개한 k-nn을 이용한 결측치 처리는 1) 다른 대체 기법과 달리 최적 방법에 대한 합의가 이루어지지 않은 hot deck 대체를 새롭게 조명했으며, 2) 불확실한 모수 추정 보다는 결측치 구조에 비교적 영향을 덜 받는 비모수적 대체 방법으로 가능한 자료를 더 확보할 수 있어 연구의 신뢰성을 높일 수 있다.

본 연구를 통해서 분석에 사용될 수 없는 자료의 재수집 비용 절감과 편이(bias)가 최소화된 직사각형 형태의 행렬 자료를 생성해 표준 분석 프로그램과 기법을 쉽게 적용할 수 있으며, 이웃의 수와 거리 측정 방법 및 결측치 비율을 고려한 결측치 처리 방법의 고안에 기여할 수 있는 점을 연구 효과로 들 수 있을 것이다.

본 연구의 한계점 및 향후 연구방향은 다음과 같다.

첫째, k-nn 모델은 설명력이 부족해 결측치 속성이 2개 이상일 경우에 의사결정트리와 같은 기법으로 보완할 필요성이 있다.

둘째, 대체된 자료는 실제 자료가 아니기 때문에 추정된 분산의 불확실성을 반영할 수 있는 추가적인 노력이 필요하다.

셋째, 성능에 영향을 주는 속성 선택의 어려움과 최적 이웃 k를 정하는 것이 쉽지 않다. 여러 개의 k값 중에서 가장 성능이 우수한 것을 선택하지만 과적합으로 인해 대체 정확도가 하락할 수 있다.

k-nn의 설명력과 lazy 학습 성능 개선, k-nn의 모수를 최적화할 수 있도록 유전자 알고리즘이나 신경망 등을 결합한 hybrid 방식에 관한 추후 연구를 필요로 한다.

References

Acock, A.C., “Working with missing values”, *Journal of Marriage and Family*, Vol.67, No.4, 2005, 1012-1028.

Allison, P.D., “Missing data : Quantitative applications in the social sciences”, *British Journal of Mathematical and Statistical Psychology*, Vol.55, No.1, 2002, 193-196.

Anderson, A.B., R. Basilevsky, and D.P.J. Hum, “Missing data : a review of the literature”, *Handbook of survey research*, Vol.4, 1983, 415-494.

Andridge, R.R. and R.J.A. Little, “A Review of Hot Deck Imputation for Survey Non-response”, *International Statistical Review*, Vol.78, No.1, 2010, 40-64.

Baraldi, A.N. and C.K. Enders, “An introduction to modern missing data analyses”, *Journal of School Psychology*, Vol.48, No.1, 2010, 5-37.

Batista, G.E. and M.C. Monard, “A Study of K-Nearest Neighbour as an Imputation Method”, *HIS*, Vol.87, 2002, 251-260.

Bennett, D.A., “How can I deal with missing data in my study?”, *Australian and New Zealand Journal of Public Health*, Vol.25, No.5, 2001, 464-469.

Carpenter, J., “Statistical modelling with missing data using multiple imputation Session 2 : Multiple Imputation”, 2010.

Cheng, X. and D. Cook, and H. Hofmann, "MissingDataGUI : A Graphical User Interface for Exploring Missing Values in Data", 2013.

Christobel, Y.A. and P. Sivaprakasam, "Improving the performance of K-nearest neighbor algorithm for the classification of diabetes dataset with missing values", *International Journal of Computer Engineering and Technology*, Vol.3, No.3, 2012, 16-23.

Devane, D.C., M. Begley, and M. Clarke, "How many do I need? Basic principles of sample size estimation", *Journal of Advanced Nursing*, Vol.47, No.3, 2004, 297-302.

Finch, W.H., "Imputation Methods for Missing Categorical Questionnaire Data : A Comparison of Approaches", *Journal of Data Science*, Vol.8, 2010, 361-378.

Graham, J.W., P.E. Cumsille, and E. Elek, *Fisk Methods for handling missing data*, Handbook of psychology, 2003.

Gunn, S.R., "Support vector machines for classification and regression", *ISIS technical report*, Vol.14, 1998.

He, H., W. Graco, and X. Yao, "Application of genetic algorithm and k-nearest neighbour method in medical fraud detection", *Simulated Evolution and Learning*, Springer Berlin Heidelberg, 1999, 74-81.

Jonsson, P. and C. Wohlin, "An evaluation of k-nearest neighbour imputation using likert data", *Software Metrics, 2004. Proceedings 10th International Symposium on IEEE*, 2004.

Kim, K. and H. Ahn, "Optimization of Support Vector Machines for Financial Forecasting", *Journal of Intelligence and Information System*, Vol.17, No.4, 2011, 241-254.

King, G. et al., "Analyzing incomplete political science data : An alternative algorithm for multiple imputation", *American Political Science Association*, Vol.95. No.1, 2001.

Little, R.J.A., "A test of missing completely at random for multivariate data with missing values", *Journal of the American Statistical Association*, Vol.83, No.404, 1988, 1198-1202.

Little, R.J.A. and D.B. Rubin, "Statistical Analysis with", 2002.

MacCallum, R.C. et al., "On the practice of dichotomization of quantitative variables", *Psychological methods*, Vol.7, No.1, 2002, 19.

Martin, A.T., M. Akshmi, and V.P. Venkatesan, "An Analysis on Qualitative Bankruptcy Prediction Rules using Ant-Miner", *International Journal of Intelligent Systems and Applications*, Vol.6, No.1, 2013.

Peng, C.J. et al., "Advances in missing data methods and implications for educational research", *Real data analysis*, 2006, 31-78.

Pettersson, N., "Real donor imputation pools", *Proceedings of the Workshop of the Baltic-Nordic-Ukrainian network on survey statistics*, 2012.

Roth, P.L., "Missing data : A conceptual review for applied psychologists", *Personnel Psychology*, Vol.47, No.3, 1994, 537-560.

Rubin, D.B., "Inference and missing data", *Biometrika*, Vol.63, No.3, 1976, 581-592.

Sarma, H.T. et al., "An improvement to k-nearest neighbor classifier", *arXiv preprint arXiv : 1301.6324*, 2013.

Saunders, J.A. et al., "Imputing missing data : A comparison of methods for social work researchers", *Social work research*, Vol.30, No.1, 2006, 19-31.

Somasundaram, R.S. and R. Nedunchezhian,

"Evaluation of Three Simple Imputation Methods for Enhancing Preprocessing of Data with Missing Values", *International Journal of Computer Applications (0975-8887)*, Vol.21, No.10, 2011, 14-19.

Schafer, J.L., *Analysis of incomplete multivariate data*, CRC press, 1997.

Schafer, J.L. and J.W. Graham, "Missing data : our view of the state of the art", *Psychological methods*, Vol.7, No.2, 2002, 147.

Schlomer, G.L., S. Bauman, and N.A. Card. "Best practices for missing data management in counseling psychology", *Journal of Counseling Psychology*, Vol.57, No.1, 2010.

Suykens, J.A., "Advances in learning theory : methods, models, and applications," Vol.190, IOS Press, 2003.

Van Buuren, Stef, *Flexible imputation of missing data*, CRC press, 2012.

Vapnik, V.N., *Statistical Learning Theory*, Wiley, New York, 1998.

Viswanath, P. and T.H. Sarma, "An improvement to k-nearest neighbor classifier", *Recent Advances in Intelligent Computational Systems (RAICS)*, IEEE, 2011.

Yan, X., "Weighted K-Nearest Neighbor Classification Algorithm Based on Genetic Algorithm", *TELKOMNIKA Indonesian Journal of Electrical Engineering*, Vol.11, No.10, 2013.

Zhang, C., Q.Y. Zhu, X.J. Zhang, and S. Zhang, "Clustering-based missing value imputation for data preprocessing", *In Industrial Informatics*, IEEE International Conference on, IEEE, 2006, 1081-1086.

.

About the Authors

Soonchang Kwon (sckwon@incheon.ac.kr)

Soonchang Kwon is a professor of the Division of International Trade at the Incheon National University. He's got the Ph.D. from Oregons State University and his main research area is hybrid intelligent system and is now being carried out Financial Prediction through machine learning and pattern recognition.