응용통계연구 제 15권 1호, 2002년, pp. 33-43

범주형 자료의 결측치 추정방법 성능 비교

신형원[1] 손소영[2]

요 약

범주형 데이터의 결측치 추정을 위하여 최빈 범주법, 로지스틱 회귀분석, 연관규칙과 같은 다양한 방법이 연구되어왔다. 본 연구에서는 이러한 방법의 추정값을 결합하는 신경망 융합과 투표융합 방법을 제안하고 이의 성능을 시뮬레이션을 이용하여 비교하였다. 실험에 사용된 데이터의 특성을 나타내는 인자로는 (1) 입출력 변수간의 연결함수, (2) 데이터의 크기, (3) 노이즈의 크기 (4) 결측치의 비율, (5) 결측발생 함수를 사용하였다. 분석결과는 다음과 같다. 데이터의 크기가 작고 결측 발생 비율이 높으면 최빈 범주법, 연관규칙, 신경망 융합의 성능이 높게 나타났으며 데이터의 크기가 작고 결측발생 확률이 결측이 안된 나머지 변수에 높은 의존관계가 있으면 로지스틱 회귀분석, 신경망 융합의 성능이 높게 나타났다. 데이터의 크기가 크고, 결측치의 비율이 낮으면서, 노이즈가 크고 결측발생 확률이 결측이 안된 나머지 변수에 높은 의존관계가 있으면 신경망 융합의 성능이 높게 나타났다.

주요용어:범주형, 결측치, 신경망, 투표, 융합

1. 연구배경

최근 들어 신제품 개발을 위한 소비자 선호도 분석이나 각종 선거의 결과를 예측 하기 위하여 설문지나 전화 인터뷰를 이용하여 자료수집을 하고 있다. 이러한 방식으로 수집된 자료는 대부분 주어진 항 목에서 답을 고르는 형식인 범주형 성격을 가지고 있으며 다양한 분석을 위해 이용되고 있다 (Park and Lee, 1998). 분석된 결과의 이용가치는 분석방법과 더불어 분석의 바탕이 되는 조사자료의 질이 중요하다. 조사자료의 질을 높이려면 표본오차와 함께 비표본 오차(non-sampling error)도 다루어야 한다. 비표본 오차란 표본 추출된 조사대상자로부터 정보를 얻지 못하여 발생하며 이는 조사단위로부터 대답을 얻지 못하여 발생하는 무응답, 부정확한 측정도구나 대답, 또는 자료를 기록하거나 편집할 때 발생한다 (Park and Lee, 1998). 이중에서 대부분을 차지하는 무응답은 자료를 수집할 때 필연적으로 발생하는 결측치로써, 특히 범주형 자료에 대한 기존의 처리방법은 단순히 무응답으로 인한 빈칸을 제외 시키는 경향이 많다. 무응답이 발생하는 현상(mechanism) 대한 가정(Rubin, 1976)은 만일 주어진 변수 X, Y에서 Y 변수에 결측이 발생할 확률이 X, Y 의 값에 영향을 받지 않는다면 missing completely at random (MCAR) 이라 한다. 또한, Y 변수에 결측이 발생할 확률이 X에 영향을 받으며 Y 에는 영향을 받지 않으면 missing at random

1) 서울 특별시 서대문구 신촌동 134, 연세대학교 컴퓨터 과학과 산업시스템 공학과.
E-mail: won3@yonsei.ac.kr
2) (339-800) 서울 특별시 서대문구 신촌동 134, 연세대학교 컴퓨터 과학과 산업시스템 공학과.
E-mail: sohns@yonsei.ac.kr

(MAR) 이라 한다 (Little and Rubin, 1997). 이러한 가정을 따르는 경우는 결측 현상을 무시해도, 정확한 추정이 가능하다(Lee and Kim., 1997).

결측치가 포함된 범주형 자료에 대하여 결측값을 대체(imputation)하는 다양한 연구가 있어왔다. Lee and Kim (1997)은 출생 및 사망신고의 실태를 조사하기 위한 설문조사의 결측치를 대체하기 위하여 최빈범주법, 조건부확률법, 최우추정법을 사용하여 비교하였다. 비교결과, 최빈범주법이 가장 정확했으며 그 이유는 대체하고자 하는 결측발생 변수의 특정 수준(level)에 응답이 치우쳐 있기 때문으로 파악하였다. 시뮬레이션을 바탕으로 결측치 대체방법의 정확성을 비교한 논문으로 Hedderley and Wakeling (1995)에서는 연속형 변수를 대상으로 몇 가지 결측치 대체 방법의 성능을 시뮬레이션을 이용하여 비교하였다. 실험에 사용한 인자는 데이터의 수, 입력변수의 수, 데이터에 포함된 노이즈의 크기, 결측의 비율이며 결측치 대체 방법으로는 EM 알고리즘, row column substitution, prinqual, mean substitution 방법을 비교하였다. 실험결과 means substitution 방법이 다른 방법에 비하여 낮은 정확성을 보인 반면 데이터에 노이즈가 많을 때는 오히려 높은 성능을 보이는 등 다양한 결과를 제시하였다. Schenker and Taylor (1996)은 노이즈의 크기, 데이터의 크기, 결측치의 비율, 결측치의 분포를 실험인자로 하여 fully parametric imputation, predictive mean matching imputation, local residual draw imputation 방법에 따른 결측값 대체 성능을 비교하였다. 분석결과, 평균제곱오차 관점에서 일반적으로, local residual draw imputation 방법이 가장 우수한 것으로 나타났다.

이상과 같이 기존의 많은 결측치 대체를 위한 경험적 연구와 시뮬레이션 연구가 진행되어 왔다. 그러나 많은 연구가 연속형 데이터를 대상으로 하고 있으며 범주형 데이터에 대한 결측대체 방법에 대한 연구는 상대적으로 부족하였다. Ragel and Cremilleux (1999)는 결측치가 발생한 범주형 변수와 나머지 변수간의 관계를 연관규칙의 지지도(support value)를 기준으로 파악하여 지지도가 가장 높은 변수로부터 결측치의 값을 대체하였다. 그러나 이상의 연구는 특정 데이터를 바탕으로 경험적 연구(empirical study)를 바탕으로 한 결과이므로 다른 데이터에도 일반적으로 적용되는 성능분석이라 할 수 없다. 따라서 본 연구에서는 데이터의 특성을 나타낼 수 인자와 수준을 설정하여 시뮬레이션을 통한 실험을 수행하고자 한다. 또한 기존의 결측치 대체 방법은 개별 모형만을 사용하여 결측치를 대체하였으나 본 실험에서는 MAR 가정을 따르는 결측치를 포함한 데이터에 대하여, 개별 모형의 대체결과를 융합(fusion)하는 방법을 제안하고 이의 성능을 비교하고자 한다. 본 논문의 구성은 다음과 같다.

2장에서는 본 연구에서 사용한 결측치 대체를 위한 5가지 방법을 소개하였으며 3장에서는 다양한 조건 하에서 대체방법에 따른 성능비교를 위하여 사용한 실험의 인자와 수준을 기술하였다. 4장에서는 실험결과를 분석하고 결과의 해석을 하였으며, 끝으로 5장에서는 분석결과를 정리하고 향후 연구방향을 제시하였다.

주요용어:범주형, 결측치, 신경망, 투표, 융합

2. 결측치 대체를 위한 모형

범주형 자료의 결측치 대체를 위한 다양한 모형 중, 본 연구에서는 비교적 사용이 편하여 자주 쓰이는 최빈 범주법과 결측이 발생한 과정에 대한 모수를 추정하는 로지스틱 회귀분석법, 최근 들어 범주형 자료의 결측값 대체에 관심을 모으는 연관규칙법을 소개하고 이상의 세가지 분석방법의 융합된 형태로 투표방법과 신경망 방법을 기술하였다. 각각의 방법을 자세히 살펴보면 다음과 같다.

2.1. 최빈 범주법(modal category model)

하나의 관측치에서 결측이 발생한 변수를 y, 결측이 발생하지 않은 변수들을 x_p 라 하였을 때 결측된 관측치의 x_p 변수들과 값이 같은 다른 관측치들의 최빈(most frequency) y 범주를 결측된 항목에 삽입하는 방법이다 (Lee and Kim, 1997). 즉, 전체 n 개의 관측치중 i 번째 관측치의 y 변수에서 결측이 발생하였을 때, i 번째 관측치의 x 변수들과 동일한 값을 갖는 관측치들을 나머지 $n-i$ 개의 관측치중에서 찾고, 이들 관측치의 y 값 중에서 최대 빈도를 갖는 범주를 겨측이 발생한 i 번째 관측치의 y 변수 값으로 대체하는 것이다.

2.2. 로지스틱 회귀분석(logistic regression)

로지스틱 회귀분석은 식(2.1) 과 같이 하나 또는 그 이상의 설명변수(x_p)로 범주형인 종속변수(y)의 값을 예측하고자 할 때 사용하는 방법으로 결측치가 있는 변수를 종속변수로 하고 결측이 없는 변수를 설명변수로 하여 추정한다. 하나의 관측치에 두 개 이상의 결측이 있을 때는 결측이 없는 설명변수만으로 하나를 먼저 대체하고 대체된 값을 포함한 설명변수로 다른 결측값을 추정한다(Neter et al., 1996)

$$E(Y) = \frac{exp(\beta_0 + .. + \beta_p x_p)}{1 + exp(\beta_0 + .. + \beta_p x_p)} \tag{2.1}$$

미지의 모수인 β_0, β_p는 보통 최우추정법(Maxmun Likelihood Estimation)이나 반복적인 가중 최소 자승법(Iterative Weighted Least Squares Method)을 이용하여 추정이 된다.

2.3. 연관규칙(association rule)

연관규칙 탐사는 대상이 되는 둘 이상의 항목 또는 변수간의 연관성을 파악하여 의사결정에 유용한 정보를 제공하는 것을 목적으로 한다. 연관규칙 중 가장 널리 쓰이는 apriori 방법은 전체 데이터에 포함된 모든 항목을 이진(binary)형태로 변환하여 자주 거래되는 항목을 중심으로 다른 항목과 동시에 발생되는 비율(지지도)을 식 (2.2)과 같이 구한다. 이 값이 사전에 정의한 최소 지지도(minimum support)보다 크면 연관규칙을 생성하고, 그렇지 않으면 연관성이 없는 것으로 파악한다. 같은 과정으로 변수의 수를 확장해 가면서 고빈도 항목에 대한 지지도를 계산한다.

$$S = \frac{freq(X \cap Y)}{N} > threshold \tag{2.2}$$

전체 트랜젝션(N) 중 항목 X와 Y를 동시에 포함된 개수

이와 같이 생성된 연관규칙을 이용하여 하나의 관측치에서 결측이 발생한 변수를 y, 결측이 발생하지 않은 변수들을 x_p 라 하였을 때, y 의 범주별로 x_p 와 지지도 값이 가장 높은 관측치들의 최빈 y 범주를 결측에 대체한다 (Ragel and Cremilleux, 1999; Lee, 1999; Ng and Lee, 1998).

2.4. 투표융합 (voting fusion)

대체된 결측치 값이 대체방법에 따라 각기 다를 수 있다. 따라서 더욱 정확한 대체를 위하여 각 방법의 각기 다른 대체결과를 투표로 결정하는 방법이다. 예를 들어 세가지 대체방법의 결과가 각각 1,2,2 였다면 투표융합방법은 2로 결정된다.

2.5. 신경망 융합(neural network fusion)

신경망은 일반적으로 예측 능력에 높은 정확성을 가지고 있으며 비선형 모형에 적합하다고 평가되고 있는 패턴 추출의 방법 중 하나이다. 이를 결측치 대체 융합에 사용하기 위하여 결측이 없는 완전한 데이터를 학습용 자료와 검증용 자료로 나누어 개별 대체방법의 대체값을 역전파(back-propagation) 신경망의 입력으로 사용하고 결측이 발생하지 않은 실제 Y 값을 신경망의 출력으로 하여 개별 대체방법과 Y 값과의 관계를 학습한 후, 결측이 있는 데이터에 적용하는 방법이다. 이상의 다섯 가지 결측치 대체방법의 성능을 다양한 조건의 결측 자료에서 비교하기 위하여 3 장에서는 결측이 발생하는 형태와 데이터의 크기등에 따라 실험하기 위한 디자인을 기술하였다.

3. 실험디자인

본 장에서는 범주형 자료의 특성을 (1)입출력 변수간 연결함수 (2)데이터의 크기 (3)노이즈의 크기 (4)결측치의 비율 (5) 결측발생 형태로 나누고 각 시나리오별로 최빈 범주법, 로지스틱 회귀분석, 연관규칙 대체, 투표융합, 신경망 융합 방법을 이용하여 결측치를 대체하고 실제값과 대체된 값이 일치하는 정분류율을 반응변수로 하여 분석하고자 한다. 시뮬레이션에 사용된 각 요인(facor)과 요인별 수준(level)을 살펴보면 다음과 같다.

3.1. 입출력 변수간 연결함수의 종류

데이터의 입력과 결측이 있는 출력간의 연결 관계는 무한한 종류의 함수가 가능하나, 식 (3.1)과 같은 로지스틱 함수를 사용하고 선형인 $f_1(x)$ 와 비선형인 $f_2(x)$로 가정하였다. 입력변수의 수는 3개이며 각각 4개의 범주를 균일하게 (uniform) 갖도록 하였다. 함수에 사용된 계수는, 요인에 함수의 선형성과 비선형성 외에 다른 요소는 포함되지 않도록 하기 위하여, 두 함수의 표준편차를 동일하게 하는 범위 내에서 임의로 선택되었다.

$$P_i(x) = \frac{exp(f_k(x))}{1 + exp(f_k(x))} \tag{3.1}$$

where P_i= 관측치가 범주 i 에 속할 확률, k=함수의 종류

$f_1(x) = 0.5x_{11} - 0.5x_{12} - 0.23x_{13} + 0.28x14 + 0.1x_{21} - 0.7x_{22} - 0.24x_{23} + 0.57x_{24} - 0.31x_{31} - 0.4x_{32} + 0.52x_{33} - 0.4x_{34}$

$f_2(x) = \sin(x_{11} - x_{12})^2 - 0.5\sqrt{x_{13} + 3x_{14}} \cos \frac{x_{21}}{x_{22}+1} - (x_{33} + 0.5x_{34}) \sin(2x_{23} + 2x_{24} + x_{31} + x_{32})$

출력변수는 식 (3.1)의 P_i 를 이용하여 다음과 같이 2개의 범주를 갖도록 범주형 변수로 생성한다.

$$T_i = \begin{cases} 1 \\ 0 \end{cases}$$

where $T_i \sim Bin(1, P_i)$

3.2. 데이터의 크기

Hedderley and Wakeling (1995)은 결측치를 대체하는데 있어서, 데이터의 크기는 중요한 인자이며 변수의 크기에 비하여 관측치 개수와 변수의 개수의 관계에 따라, 즉 관측치의 숫자가 많고 변수의 수가 적은 데이터 (long and narrow) 인가 또는 관측치가 적고 변수의 수가 많은 데이터 (wide and shallow) 인가에 따라 결측 대체 방법들의 성능이 달라지는 것으로 보았다. 이들의 연구에서는 입력변수 개수의 최소 12배에서 250배에 해당하는 데이터 크기 수준(level)을 사용하였다. 또한 Schenker and Taylor (1996)의 시뮬레이션 연구에서는 2개의 입력변수를 사용하여 데이터 크기의 수준을 100과 500으로 설정하였으며 이는 입력변수 개수의 50배, 250배에 해당한다. 본 연구에서는 이보다 수준 차이를 더 크게 하였을 때 데이터의 크기가 미치는 영향을 알아보기 위하여 각각 4개의 범주를 갖고 있는 3개의 입력변수를 사용하고 추정 해야할 12개 모수들 수의 5배, 100배에 해당하는 60과 1200을 사용하였다.

3.3. 입출력 변수간 연결함수에 포함된 노이즈의 크기

Schenker and Taylor (1996)의 연구에서는 표준정규분포로부터 분산이 1과 4를 가지는 두 수준을 입출력 연결함수에 첨가하였다. 또한 Hedderley and Wakeling (1995)의 연구에서는 t 분포를 따르는 노이즈와 정규분포를 따르는 노이즈의 두 수준을 사용하였다. 본 연구에서는 Schenker and Taylor (1996)의 연구와 같이 동일한 분포를 사용함으로써 노이즈의 크기를 분명히 알 수 있도록 표준 정규분포를 사용하였다. 노이즈의 두 수준은 로지스틱 선형인 경우, 평균이 0이고 표준편차가 0.2와 0.8인 정규분포를 사용하였다. 노이즈의 표준편차로 정한 0.2 와 0.8 이라는 숫자는 임의로 정한 것이나, 이 숫자가 본래 데이터에 어떤 영향을 미치는가를 보이기 위하여 표 3.1 과 같이, 연결 함수 f(x)와 노이즈가 더해진 $f(x) + 0.2 \times Normal(0,1)$, $f(x) + 0.8 \times Normal(0,1)$ 의 상관관계를 나타내었다.

표 3.1: 연결 함수의 종류와 노이즈의 크기간 조합에 따른 피어슨 상관계수.

	$f_1(x)$	$f_1(x)+0.2\times Normal(0,1)$	$f_1(x)+0.8\times Normal(0,1)$
$f_1(x)$	1	0.96	0.68
	$f_2(x)$	$f_2(x)+0.2\times Normal(0,1)$	$f_2(x)+0.8\times Normal(0,1)$
$f_2(x)$	1	0.96	0.68

3.4. 결측치의 비율

결측치의 비율 인자는 연속형 변수의 결측값 대체를 위한 기존의 시뮬레이션 연구에서 많이 사용된 인자로써 기존 연구의 결측비율 수준을 살펴보면 다음과 같다. -Fraley(1999) : 5, 10, 15, 20 % -Hedderley and Wakeling (1995) : 5, 35, 65 % -Schenker and Taylor (1996) : 25, 50 % -Atkinson and Cheng (2000) : 10, 20, 30, 40 % 이와 같이 결측의 비율은 다양한 수준으로 실험되어 결측대체 방법에 영향을 미치는 것으로 나타났으나, 이상의 기존연구는 연속형 변수의 결측에 한정되어 있다. 따라서 본 연구에서는 범주형 변수에 대하여 결측치의 비율을 Fraley (1999)의 연구에서 사용된 최소 결측비율인 5%와 Schenker and Taylor (1996)의 연구에서 사용된 최대 결측비율인 50% 를 사용하였다. 로지스틱 함수로부터 발생된 결측비율이 5%일 때와 50% 일대의 결측 발생함수는 식 (3.2) 와 같다.

$$Q_i(x) = \frac{exp(g_k(x)+s(x))}{1+exp(g_k(x)+s(x))} \tag{3.2}$$

where Q_i= 관측치가 결측될 확률 $s(x)$=표준 정규분포로부터 발생된 랜덤 노이즈, k=함수의 종류

결측이 5%일 때의 결측발생 함수

$g_1(x) = 1.2X_{11}+0.4X_{12}+0.8X_{13}+0.9X_{14}+0.8X_{21}+1.4X_{22}+1.5X_{23}+1.1X_{24}+2X_{31}+0.8X_{32}+0.6X_{33}+0.3X_{34}$

결측이 50%일 때의 결측발생 함수

$g_2(x) = 0.05x_{11}+0.32x_{12}-0.38x_{13}-0.09x_{14}+0.08x_{21}-0.14x_{22}+0.15x_{23}+0.11x_{24}-0.2x_{31}+0.07x_{32}+0.06x_{33}-0.03x_{34}$

결측의 발생은 식 (3.2)의 결측 발생확률 Q_i를 이용하여 다음과 같이 M_i 값이 2개의 범주를 갖도록 한 뒤, 만일 M_i 값이 0이면 i 번째 관측치의 Y 변수를 결측 시키며, 1이면 결측 시키지 않고 그대로 나둔다.

$$M_i = \begin{cases} 1 \\ 0 \end{cases}$$

$where M_i \sim Bin(1, Q_i)$

이상과 같은 과정으로 결측을 발생 시켰을 때에는, 함수 $g(x)$에 사용된 계수에 따라 $exp[g(x)+s(x)]$의 기대값이 0에 가까우면 Q_i 의 기대값은 0.5 가 되므로 M_i 의 범주별 비율은 0.5:0.5 가 된다. 반대로 Q_i 의 기대값이 0.95이면 M_i 의 범주별 비율은 0.95:0.5 가 된다.

이 범주별 비율은 전체 데이터 중 결측이 발생한 비율로 사용된다. 한편 함수 $g(x)$에 사용된 계수는 결측치의 비율이 5%, 50% 가 나오도록 임의로 선택되었다. 이와 같이 함수 $g(x)$를 사용한 경우는 y 변수의 결측 발생확률이 x 변수에는 의존하며, y 변수에는 의존하지 않는 MAR (missing at random) 가정을 따른다.

3.5. 결측발생 형태

Spence and Domoney (1974) 의 연구에 의하면 결측치가 랜덤하게 발생했느냐 또는 어떤 패턴을 가지고 발생했느냐는 결측치 대체 방법에 영향을 미칠 수 있다는 연구 결과를 보였다. 따라서 Y 에 결측이 발생할 확률이 X 변수들과 강한 함수관계를 가지고 있는 경우와 약한 함수관계를 가지고 경우에 따른 결측대체 방법의 성능을 비교하였다. Y 에 결측이 발생할 확률이 결측이 없는 X 변수들과 강한 상관관계를 가진다면 결측된 Y 와 동일한 X 값을 가지는 결측이 없는 관측치(observation)의 수가 적어 결측을 대체하는데 참조할 수 있는 데이터의 수가 적어진다. 이러한 실험 데이터의 결측발생 특성이 결측 대체방법에 미치는 영향을 알아보기 위하여 식 (3.2)의 g(x) 에 두 수준의 노이즈 s(x) 를 더하였다. 50%의 결측 발생 함수에는 노이즈로써 평균이 0이고 표준편차가 0.3과 1.5인 정규분포를 사용하였고, 5% 결측발생 함수에는 정규분포의 평균이 0이고 표준편차가 2와 10이 되는 노이즈를 더해 주었다. 이와 같이 함수마다 다른 크기의 노이즈를 더해주는 것은 두 함수 g1(x), g2(x) 모두 본래의 결측 발생함수에서 동일하게 2배와 10배의 표준편차 차이가 나도록 하기 위함이다. 2배의 표준편차 차이가 나는 것은 Y 변수의 결측 발생확률이 X변수와 상대적으로 강한 상관관계를 갖는 것을 의미하며, 10배의 차이가 나는 것은 상대적으로 약한 상관관계를 갖는 것을 의미한다. 결측 발생 함수에 더해진 노이즈의 크기에 따른 상관관계의 차이는 표 3.2에 나타난 바와 같다.

표 3.2: 결측 발생 함수의 종류와 노이즈의 크기간 조합에 따른 피어슨 상관계수.

	$g_1(x)$	$g_1(x) + 2 \times Normal(0,1)$	$g_1(x) + 10 \times Normal(0,1)$
$g_1(x)$	1	0.5	0.08
	$g_2(x)$	$g_2(x) + 0.3 \times Normal(0,1)$	$g_2(x) + 1.5 \times Normal(0,1)$
$g_2(x)$	1	0.5	0.08

3.6. 결측 대체방법

실험에 사용된 결측 대체 방법들은, 많은 연구에서 오랜 기간 사용해온 방법인 최우추정법을 이용한 로지스틱 회귀분석, 최근 들어 범주형 결측자료의 대체방법으로써 주목을 받는 연관규칙, 비교적 대체방법이 쉬워 널리 쓰이는 최빈범주법을 사용하였다. 또한 본 연구에서 제시하는 방법으로써, 각 방법의 결측값에대한 대체값을 신경망의 입력으로 사용하여 결합된 예측을 하는 신경망 융합방법을 사용하고 이를 각 대체 방법들의 결측치에 대한 대체값들을 단순히 투표융합 (majority voting) 하는 방법과 비교하였다. 역전파 신경망

사용시, 구조는 사전 실험을 통하여 2개의 은닉층에 각각 2개와 3개의 은닉노드를 할당하고, 로지스틱 활성함수를 사용하였으며 학습률은 0.05 를 사용하였다.

이상과 같이 데이터의 특성을 나타내는 5개의 실험인자에 따른 다섯 가지의 결측 대체 방법의 결측치 대체 정확성을 비교함에 있어 검정하고자 하는 가설은 다음과 같다.

실험가설 Ha1: 입력과 출력의 함수가 선형이면 로지스틱 회귀분석이 다른 방법에 비하여 결측 대체 정확성이 높다. Ha2: 데이터의 크기가 작고 결측치의 비율이 크면 최빈범주법 또는 연관규칙이 다른 방법에 비하여 결측대체 정확성이 높다. Ha3: 데이터의 크기가 작고 결측발생 관계가 강한 함수 관계를 가지면 로지스틱 모형이 다른 방법에 비하여 결측대체 정확성이 높다.Ha4: 데이터의 크기가 크고, 결측치의 비율이 작고, 결측발생 상관관계가 강하면서 노이즈가 크면 신경망 융합과 투표 융합이 다른 방법에 비하여 결측대체 정확성이 높다.

Ha1 의 설정이유는 입력과 출력변수간의 연결함수가 선형이면 로지스틱 선형 회귀분석이 정확할 것으로 기대되기 때문이며, Ha2 는 데이터의 크기가 작으면서 결측치의 비율이 높으면 로지스틱 회귀분석과 같은 모수적인 대체방법은 모수추정에 사용할 수 있는 관측치의 숫자가 적어지므로 낮은 자유도(degree of freedom) 에 의하여 정확성이 떨어질 것으로 예상 되기 때문이다. 또한 데이터의 크기가 작으면서 Y 변수에 결측이 발생할 확률이 X와 강한 관계를 가지고 있으면 최빈 범주법이나 연관규칙은 알고리즘의 특성상, 결측된 관측치와 동일한 X 값을 가지는 완전한(complete) 데이터가 존재할 확률이 적어 참조할 수 있는 데이터의 수가 적어지므로 로지스틱 회귀분석보다 낮은 성능을 보인다 (Ha3). Ha4의 설정 이유는 신경망의 경우, 추정해야 할 모수의 수가 많으므로 데이터의 크기가 크고 결측치의 비율이 작을 때 높은 성능 보이며, 개별 대체방법의 융합된 예측을 함으로 큰 노이즈에 대하여 상대적으로 강건할 것으로 기대된다. 또한 Ha3 의 결측발생 함수 조건을 가지고 있으면 최빈범주법, 연관규칙의 단점을 보완하여 융합효과가 나타날 것으로 예상된다.

이상의 시뮬레이션을 위한 실험은 5×2^5 디자인이며 반복은 랜덤 노이즈의 초기값을 달리하여 5회를 함으로써 총 실험횟수는 800회 이다. 또한 실험의 반응변수로써, 0 또는 1의 값을 갖는 출력변수에 결측이 발생했을 때, 전체 데이터 중 원래의 관측치가 1인 경우에 1로 대체시킨 경우와 원래 관측치가 0인 경우에 0으로 대체시킨 경우의 비율(분류정확성)을 사용하였다.

4. 실험결과

위와 같은 $Ha1 \sim Ha4$의 가설들에 대해 표 4.1 과 같이 유의수준 5% 에서 분산분석을 하여 가설 검정한 결과, 모든 주효과와 $Ha1 \sim Ha4$에 포함된 모든 교호작용이 유의하게 나타났다. 표 4.1 의 분산분석 표에는 $Ha1 \sim Ha4$ 의 검정에 사용된 교호작용 만을 나타내었다.

F1: 입출력 변수간의 연결함수 F2: 데이터의 크기F3: 노이즈의 크기F4: 결측치 비율F5: 결측발생 함수F6: 대체방법

가설 $Ha1 \sim Ha4$ 의 관점에서 데이터의 특성에 따른 적합한 결측치 대체방법을 선택하기 위하여 고차 교호작용을 중심으로 던칸(Duncan) 검정을 하였다. 유의수준 5% 에서 던

표 4.1: 결측 발생 함수의 종류와 노이즈의 크기간 조합에 따른 피어슨 상관계수.

Source	DF	SS	MS	F-value	P-value
$F1 \times F6$	4	2126.59	531.64	155.11	0.0001
$F1 \times F3 \times F4 \times F5 \times F6$	4	385.50	96.37	28.12	0.0001

칸 검정결과, 입출력 변수간의 함수가 선형일 경우, 신경망 융합이 다른 네 가지 방법에 비하여 높은 성능을 나타냈으며 나머지 네 가지 방법간에는 유의한 차이가 없었다 (Ha1). 이는 입출력 변수간 함수관계가 선형일 경우, 로지스틱 회귀분석 외에 최빈 범주법이나 연관규칙에도 대체에 정확성을 더 해주는 것으로 볼 수 있다. 한편, 데이터의 크기가 작고, 결측치의 비율이 높으면 로지스틱 회귀분석은 다른 네 가지 방법에 비하여 예측성능이 떨어짐을 보였다. 이는 3장의 가설설정 이유에서 언급 하였듯이, 자유도의 부족에 기인한 것으로 보인다 (Ha2). 또한 데이터의 크기가 작고, Y 변수의 결측 발생 확률이 X 변수들과 강한 함수 관계를 가질 때는 신경망 융합과 로지스틱 회귀분석이 높은 성능을 보였다 (Ha3). 이러한 결과는 최빈 범주법이나 연관규칙이 Y 변수에 결측이 발생 되었을 때의 X 변수들의 값과 동일한 값을 가진 완전한 데이터가 적을 경우 대체 성능이 매우 저하되는 것으로 볼 수 있다. 또한 투표융합은 개별 대체방법의 다수결 투표에 의한 것이므로 세가지 대체방법 중 두 가지의 성능이 나쁠 경우 더불어 저하되는 것으로 볼 수 있다. Ha4의 측면에서는 데이터의 크기가 크고, 노이즈가 크면서, 결측치 비율은 작고, 결측 발생 상관관계가 강하면 신경망 융합이 가장 우수한 성능을 보였다. 이는 데이터의 크기가 크고 결측치 비율이 적어 모수 대체에 사용할 수 있는 데이터의 수가 많은 상태에서, 노이즈가 많고 결측 발생관계가 강하여 결측치 대체가 어려운 조건에서 융합효과가 나타나는 것으로 볼 수 있다. 이밖에 가설에는 포함되는 않았으나 a.데이터의 크기가 작고 b.노이즈가 크면서 c.결측발생 비율이 높고 d.결측 발생 상관관계가 강할 때는 로지스틱 회귀분석이 가장 낮은 성능을 보였다. 이 결과는 가설설정 단계에서 예상하지 못하였으나 Ha2 의 확장된 결과를 나타낸다. 결측치의 대체가 가장 열악한 $a \sim c$의 조건을 가지고 있는 상태에서, d 의 조건에 의하여 결측이 발생된 관측치의 X 변수들 값과 관련이 적은 결측 안된 관측치의 X 변수들 값으로 대체된 모수는 불안정 하기 때문으로 보인다.

5. 결론

본 연구에서는 최빈 범주법, 로지스틱 회귀분석, 연관규칙, 투표 융합, 신경망 융합을 이용하여 범주형 결측 데이터의 다양한 특성에 따른 결측치 대체 성능을 비교하였다. Monte-Carlo 시뮬레이션 결과, 실험에 사용된 6개 인자는 모두 결측치 대체 성능에 유의한 영향을 미치는 인자로 나타났으며, 고차 교호작용을 중심으로 분석결과를 정리하면, 데이터의 크기가 상대적으로 작으면서 결측치 비율이 전체 데이터의 50% 이상이면 로지스틱 회귀분석의 사용을 피하는 것이 바람직하다. 또한 데이터의 크기가 위와 같이 작으면서 한 변수에서 결측이 발생할 확률이 결측이 없는 나머지 변수들과 높은 의존관계가 있으면 최빈 범주법이나 연관규칙의 사용을 피하는 것이 좋다. 이상의 두 경우 모두에서 신경망 융합은 우

수한 성능을 보였으며 이밖에 신경망 융합은 데이터의 크기가 크고, 노이즈가 많이 포함되어 있으며, 결측치의 비율이 작고 결측발생 상관관계가 강할 경우 다른 네 가지 방법과 비교하여 유의하게 높은 성능을 보였다. 투표융합은 유의한 모든 교호작용에서 신경망 융합보다 낮은 성능 보였으나 개별 대체 방법들 (최빈 범주법, 로지스틱 회귀분석, 연관규칙) 의 평균적 성능 보이고 있어 특이하게 우수하거나 열등한 교호작용이 나타나지 않았다.

이상의 연구결과에서 신경망 융합은 범주형 결측치의 대체에서 높은 성능을 가지고 있음을 알 수 있으며 투표융합은 특이하게 우수하지는 않으나 안정적 성능을 가지고 있음을 알 수 있었다. 또한 데이터와 결측 발생의 특성에 따라 최빈 범주법, 로지스틱 회귀분석, 연관규칙 중 우수한 성능을 보이는 경우를 파악하였고, 최빈 범주법과 연관규칙은 성능이 유사하게 변함을 알 수 있었다.

이상의 연구결과는 하나의 관측치에서 여러 결측이 발생하는 다 결측(multiple missing) 데이터에서는 결과가 달라질 수 있다. 따라서 향후 연구방향으로, 다 결측 범주형 데이터에 대한 연구가 필요하며 연구에 사용된 5가지 결측 대체 방법 외에 구조방정식을 이용한 결측 대체 방법 Tang and Bentler (1998) 의 성능 또한 검증이 필요하다.

참고문헌

[1] Derringer, G. and Suich, R. (1980). *"Simultaneous optimization of several response variables"*, Journal of Quality Technology, 13, 1-45.

[2] Harrington, E. C., Jr. (1965). *"The desirability function"*, Industrial Quality Control, 21, (10), 494-498.

[3] Park, Sung. H. (1981). *"Simultaneous optimization techniques for multi-purpose response functions"*, Journal of Military Operations Research Society of Korea, 7, 118-138.

[4] 박성현, 박준오. (1997). *"Simultaneous optimization of multiple response using weighted desirability function"*, 한국품질경영학회지, 25(1), 56-68.

[2001년 4월 접수, 2001년 9월 채택]

Comparing Accuracy of Imputation Methods for Categorical Incomplete Data

Hyung Won Shin[1] So Young Sohn [2]

ABSTRACT

Various kinds of estimation methods have been developed for imputation of categorical missing data. They include category method, logistic regression, and association rule. In this study, we propose two fusions algorithms based on both neural network and voting scheme that combine the results of individual imputation methods. A Monte-Carlo simulation is used to compare the performance of these methods. Five factors used to simulate the missing data pattern are (1) input-output function, (2) data size, (3) noise of input-output function (4) proportion of missing data, and (5) pattern of missing data. Experimental study results indicate the following: when the data size is small and missing data proportion is large, modal category method, association rule, and neural network based fusion have better performances than the other methods. However, when the data size is small and correlation between input and missing output is strong, logistic regression and neural network based fusion algorithm appear better than the others. When data size is large with low missing data proportion, a large noise, and strong correlation between input and missing output, neural networks based fusion algorithm turns out to be the best choice.

Keywords: Categorical; Missing Data; Neural Network; Voting; Fusion

1) Computer Science Industrial Systems Engineering, Yonsei University, Seoul, 110-745, Korea
2) Computer Science Industrial Systems Engineering, Yonsei University, Seoul, 110-745, Korea