ISSN 2383-6318(Print) / ISSN 2383-6326(Online)
KIISE Transactions on Computing Practices, Vol. 25, No. 10, pp. 511-516, 2019. 10
https://doi.org/10.5626/KTCP.2019.25.10.511

그래프 기반 준지도 학습을 이용한 속성값 전파 결측치 추정

(Missing Value Imputation with Attribute Value Propagation using Graph-based Semi-Supervised Learning)

신 유 경 † 신 현 정 ††
(Yukyung Shin) (Hyunjung Shin)

요 약 데이터의 레코드들 중에 하나 이상의 속성값이 없는 경우는 비일비재하다. 많은 경우에 있어서 데이터의 수 대비 결측치가 없는 완전레코드의 수의 비율이 적다. 이에 대하여 평균값, 최빈값, 그리고 중앙값 등으로 대체하는 통계적 방법이 가장 보편적으로 쓰이고 있다. 또한 기계학습에서도 k-최근접 이웃 탐색이나 의사결정나무 등을 활용한 결측치 추정방법들이 자주 활용된다. 전자는 각 속성의 대표하는 값으로 대체하는 전역적 방법인데 반해 후자는 해당 레코드와 유사한 레코드들의 속성값으로 대체하는 지역적 방법이라 할 수 있다. 그러나 한 속성의 값이 대부분 결측된 경우라면 두 방법 모두 활용하기 어렵다. 이러한 한계를 극복하기 위하여, 본 연구에서는 결측치의 속성과 상관성이 큰 이웃 속성들로부터 값을 추정하는 방법을 제안한다. 속성 간 상관성을 기반으로 하여 한 속성의 대부분의 값이 결측이 되더라도 활용할 수 있다. 제안 방법론으로는 속성들 간의 상관계수로 이루어진 상관 그래프를 만들고, 그래프 기반 준지도 학습을 적용한다. 결측치는 다른 속성값들로부터 상관계수에 비례하여 전파되어 추정된다. 본 논문에서 제안한 결측치 대체 추정 방법과 기존에 결측치 대체에 많이 사용하는 통계적 방법과 기계학습을 비교하여 실험을 진행하였다.

키워드: 준지도 학습, 그래프 이론, 기계학습, 결측치 대체

Abstract The number of data records without one or more attributes is very large. In many cases, few complete records are available without missing the data values. Statistical methods that replace the missing values with mean, mode and median are commonly used. In machine learning algorithms such as K-nearest neighborhood or decision tree, the missing values are replaced by estimation methods. The statistical method is a global method that replaces each attribute with a representative value, whereas the machine learning algorithm is a local method that replaces the attribute values similar to the records. However, it is difficult to use both methods for records that contain almost all the missing values. In order to overcome these limitations, in this paper, we propose a method to estimate values from neighborhood properties associated with large correlation with the missing attribute. It is based on correlation between attributes, and can be used even if the attributes carry

· 이 논문은 2019년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원(No. 2018-0-00440, 위험 상황 초기 인지를 위한 ICT 기반의 범죄 위험도 예측 및 대응 기술 개발), 한국연구재단의 지원(No. 2017R1E1A1A03070345) 및 아주대학교 일반연구비 지원을 받아 수행된 연구입니다.
· 이 논문은 제45회 한국소프트웨어종합학술대회에서 '결측치 대체를 위한 상관성 그래프 기반 속성값 전파'의 제목으로 발표된 논문을 확장한 것임

† 비 회 원 : 아주대학교 데이터사이언스학과
ykshin93@ajou.ac.kr
†† 정 회 원 : 아주대학교 산업공학과 교수(Ajou Univ.)
shin@ajou.ac.kr
(Corresponding author임)

논문접수 : 2019년 3월 29일
(Received 29 March 2019)
논문수정 : 2019년 7월 8일
(Revised 8 July 2019)
심사완료 : 2019년 7월 23일
(Accepted 23 July 2019)

정보과학회 컴퓨팅의 실제 논문지 제25권 제10호(2019. 10)

almost missing values. In this proposed method, a correlation graph representing correlation coefficients related to attribute values was constructed based on graph-based semi-supervised learning. Missing values were estimated in proportion to the correlation coefficient derived from related attributes. In this paper, the proposed method compared the statistical method and machine learning algorithm, which are generally used for missing value imputation.

Keywords: semi-supervised learning, graph theory, machine learning, missing value imputation

1. 서 론

컴퓨터의 발전으로 방대한 데이터를 저장, 처리하고 분석하여 가치있는 인사이트를 도출하기 위한 노력이 다방면에서 이뤄지고 있다. 하지만, 데이터가 생성되는 와중에 데이터의 성분이 누락되는 경우가 빈번하게 발생하여 제대로 된 정보를 도출하지 못하는 문제점이 있다. 의료 데이터 같은 경우, 환자의 질환에 따라 다양한 검사 항목과 각종 설문조사에 대한 속성들이 다양하게 존재한다. 그러나 환자가 특정 검사에 대한 무응답 또는 비용과 시간에 대한 부담으로 인한 불응 등으로 인하여 데이터가 누락된다. 의료 데이터 뿐만 아니라 태양광 발전량 데이터, 전력 데이터 등 날씨의 영향, 장비의 오작동, 통신 상태 불량 등의 문제로 인하여 데이터의 누락이 발생하는 경우가 비일비재하다.

일반적으로 누락된 데이터를 결측치라 한다[1]. 데이터를 분석하기에 앞서 결측치가 포함된 데이터를 여러 방법을 이용하여 처리하거나 대체한다. 하지만, 많은 경우에 있어서 결측치가 포함되어 있는 해당 레코드를 제거하고 데이터를 분석하는 경우가 많아 데이터를 활용하지 못하는 문제가 있어 적합한 방법을 고안해야 한다. 데이터 분석의 전처리 단계에서 효과적으로 결측치를 처리하는 방법에 대한 많은 연구들이 진행되고 있다[2]. 연구와 실증을 통해 결측치의 대체가 가능한 추정된 값의 정확성을 높여 데이터를 제대로 활용해야 할 필요성이 있다. 이에 따라, 본 연구에서는 결측치를 추정하는 문제에 대해 고찰해보고 다양한 데이터들의 효용을 높이고자 결측치를 처리하는 방법에 대해 제안하고자 한다.

2. 연구 배경

데이터의 누락이 발생했을 때 다양한 결측치 대체 방법을 이용한다. 결측치 대체 방법은 보편적으로 쓰이는 통계적 방법과 예측 모델을 활용한 기계학습 방법이 있다. 통계적 결측치 대체 방법은 결측치를 평균값과 중앙값, 최빈값 등으로 각 속성의 대표하는 값으로 대체하는 방법으로 속성의 전체적인 특징을 가지는 속성의 전역적 방법이다. 기계학습 결측치 대체 방법은 K-최근접이웃 탐색(K-Nearest Neighborhood, KNN) 방법과 의사결정나무 등 다양한 방법이 존재하며 결측치를 해당 레코드와 유사한 레코드들의 속성값으로 대체하는 방법으로 이는 속성의 지역적 방법에 해당된다[3]. 하지만 결측치를 대체하는 통계적 방법과 기계학습 결측치 대체 방법은 각 속성 간의 상관성을 고려하지 않고 대체하는 문제점이 있다. 데이터의 각 속성이 독립적이지 않을 경우, 상관관계가 높은 속성을 이용하여 결측된 성분의 예측이 가능하다. 데이터의 각 속성은 독립적이지 않으므로 각 속성 간 상관성을 고려하여 결측치를 추정하는 방법을 제안한다.

공통적인 한계점을 보완하기 위해 속성들 간의 상관계수로 이루어진 속성 상관 네트워크를 구성하여 이를 가중치로 두고 기계학습 방법을 예측 모델로 이용하여 통계적 방법을 결합한 결측치 대체 방법론을 제안한다. 통계적 방법은 평균값과 중앙값, 최빈값을 이용하며 기계학습 방법으로 그래프 기반 준지도 학습을 이용한다. 준지도 학습은 예측값을 전파하여 정해지지 않은 값의 예측이 가능하므로 그래프 기반 준지도 학습을 예측 모델로 구성한다[4]. 기계학습의 예측 모델은 다른 속성의 상관계수에 비례하여 전파된 값을 도출하고, 이를 통계적 방법과 결합하여 결측치 추정을 향상시켜 결측치 대체가 가능하다.

3. 제안 방법론

결측치 대체 방법에 추정값으로 이용할 예측 모델인 그래프 기반 준지도 학습은 일반적으로 식 (1)을 가지며 예측값 f를 최소화하여 최적화한 식 (2)을 가진다. n은 1부터 데이터가 가진 속성의 개수의 인덱스이며 W는 대칭 그래프의 가중치 행렬, D는 $diag(\sum_i w_{ij})$, L은 라플라시안 행렬로 $L=D-W$로 정의한다. I는 단위행렬, y는 예측값 벡터이며 이를 이용하여 속성을 전파한다.

$$\min_f (f-y)^T(f-y)+\mu f^T L f \quad (1)$$

$$f=(I+\mu L)^{-1}y \quad (2)$$

본 연구에서는 속성의 전역적 방법과 지역적 방법의 한계를 고려하기 위해 각 속성 간 상관성을 고려해야 한다. 이를 고려하기 위해 그래프 기반 준지도 학습의 가중치 행렬 W를 각 데이터의 상관관계를 이용하여 속성 상관 네트워크로 구성한다. 데이터 종류의 상관성을

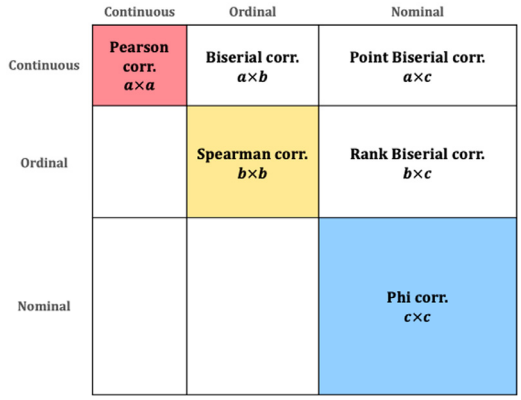


그림 1 Weight Network 구성
Fig. 1 Composition of Weight Network

고려하기 위해 다음과 같이 상관계수(γ)로 가중치 행렬 $W_{ij} := \gamma_{ij}$을 구성한다.

데이터 종류에 따라 상관관계를 정의하는 상관계수는 다르며 일반적으로 연속형 변수끼리의 상관성을 구하기 위해서는 피어슨(Pearson), 순서형 변수의 상관성은 스피어만(Spearman), 명목형 변수의 상관성은 파이(Phi) 상관계수를 쓴다[5]. 연속형과 순서형 변수의 상관성은 이연(biserial), 연속형과 명목형 변수의 상관성은 점이연(Point Biserial) 그리고 순서형과 명목형 변수의 상관성은 순위 이연(Rank Biserial) 상관계수를 이용하여 데이터 속성에 관한 상관 속성 네트워크를 구성한다[6,7]. 그림 1은 데이터 종류에 따른 상관계수를 정의한 그림이다. 전제 데이터의 속성의 개수가 n개이고 연속형 변수가 a개, 순서형 변수가 b개, 그리고 명목형 변수가 c일 때, $a+b+c=n$ 을 만족하는 속성 상관 네트워크이다.

기존의 그래프 기반 준지도 학습은 식 (3)과 같이 두 점 x_i와 x_j의 예측값이 유사한 경우에 전파하는 특징을 가진다.

$$\frac{1}{2}\sum_{i,j=1}^{n} w_{ij}(f(x_i)-f(x_j))^2 \tag{3}$$

식 (4)은 두 점 x_i와 x_j가 둘 다 을 갖지 않거나 한 점만 예측값을 가지고 있는 경우이다[8]. 결측치의 경우 값이 없는 성분에 속성의 추정값을 전파해야 하므로 식 (4)을 고려해야 한다.

$$\frac{1}{2}\sum_{i,j=1}^{n} w_{ij}(f(x_i)+f(x_j))^2 \tag{4}$$

또한, 기존 그래프 기반 준지도 학습은 가중치 행렬의 각 성분들이 0보다 클 경우에 사용한다. 하지만 상관성을 고려한 상관계수의 속성 상관 네트워크의 경우는 -1부터 1 사이의 값을 가진다. 두 가지의 문제를 고려하기 위해 기존의 그래프 기반 준지도 학습을 변형시킨 혼합 그래프 기반 준지도 학습을 이용한다. 식 (5)은 식 (3)과 식 (4)의 두 가지를 고려한 식이다. s_{ij}는 속성 상관 그래프의 부호에 따라 식 (3)일 경우는 1, 식 (4)일 경우는 -1이다.

$$\frac{1}{2}\sum_{i,j=1}^{n} w_{ij}(f(x_i)-s_{ij}f(x_j))^2 \tag{5}$$

식 (6)은 기존의 그래프 기반 준지도 학습을 변형시킨 혼합 그래프 기반 준지도 학습의 식이다. 식 (6)에서 $f^T Mf$ 는 $\frac{1}{2}\sum_{i,j=1}^{n} w_{ij}(f(x_i)-s_{ij}f(x_j))^2$를 만족한다.

$$\min_f (f-y)^T(f-y)+\mu f^T Mf \tag{6}$$

식 (7)은 식 (6)의 f를 최적화한 식으로 S는 s_{ij}에 따라 -1과 1을 가지는 부호행렬이다. M은 $L+(1-S)\odot W$로 정의하며, 속성 상관 그래프의 부호가 양수일 경우와 음수일 경우를 함께 고려한 식이다.

$$f=(I+\mu M)^{-1}y \tag{7}$$

식 (7)을 바탕으로 전역적 방법과 지역적 방법의 한계인 각 속성 간의 상관성을 고려할 수 있다. 다른 속성의 상관계수를 비례하여 전파된 식 (7)과 통계적 방법을 결합하여 식 (8)과 식 (9)와 같이 결측치 대체 방법론을 제안한다. 평균값, 최빈값, 중앙값은 각 속성의 전역성을 고려하는 값이며 f는 속성의 지역성을 가지는 혼합 준지도 학습 값으로 속성의 상관성을 고려하여 전파한 결측치를 결정하는데 필요한 추정값이다. 식 (8)에서 $\overline{X}$는 각 데이터 속성의 평균값 벡터이며, 연속형 변수인 속성의 결측치를 대체하는 식이다. 식 (9)는 $\hat{X}$는 각 데이터 속성의 최빈값/중앙값 벡터이며, 이산형 변수인 속성의 결측치를 대체하는 식이다.

$$\overline{f}=\overline{X}+f \tag{8}$$

$$\hat{f}=\hat{X}+f \tag{9}$$

4. 실 험

4.1 실험 데이터 및 실험 설정

본 실험에는 ML Repository에서 제공한 오픈 데이터 집합을 이용하여 4가지의 실험 데이터를 이용하여 실험을 진행하였다. 표 1은 실험에 사용할 데이터이며 4가지의 실험 데이터는 분류 예측 실험을 하였다.

실험은 총 2가지로 진행하였다. 첫 번째 실험은 제안 방법론인 SSLI 데이터와 원시 데이터(Raw Data), 완전 레코드(Filled-Record Data, FR Data)의 형태의 데이터를 가지고 분류 예측 정확성 비교 실험을 진행하였다. 첫 번째 실험의 목적은 결측치를 대체하는 것의 중요성과 결측치 처리 방법 중 하나인 데이터의 레코드나 속성을 제거하는 데이터 삭제로 인하여 데이터를 활용하

표 1 실험 데이터
Table 1 Experimental Data

Data	Number of Record/Attribute	Number of Filled-Record	Number of Missing Value
Arrhythmia	452 / 280	68	408
Pittsburgh Bridge	107 / 12	72	77
Audiology	226 / 69	140	95
Ozone	2534 / 73	1847	14937

지 못하는 문제점을 보이고자 제안 방법론인 SSLI을 적용하여 대체한 데이터와 비교하였다. 결측치를 처리하지 않은 원시데이터와 결측치를 포함하고 있는 레코드를 다 제거한 완전 레코드 형태의 데이터를 비교하였다.

두 번째 실험은 결측치 대체 비교 실험으로 기존의 통계적 방법인 평균값과 중앙값의 속성의 전역성을 가지는 대체 방법과 속성의 지역성을 가지는 기계학습 알고리즘의 KNN과 Mice를 이용하여 분류 예측 정확성 비교 실험을 진행하였다. Mice는 다른 변수를 예측 변수로 사용하기 때문에 실험에서는 랜덤 포레스트(Random Forest)을 이용하였다.

혼합 그래프 기반 준지도학습의 파라미터인 μ를 선정하기 위해 모든 데이터의 완전레코드를 이용하여 임의 결측(MAR)을 결측치로 두고 SSLI로 결측치 대체한 값과 원시값을 비교하여 가장 높은 정확성을 가지는 최적의 μ를 구하였다. 반복 실험은 10번 진행하였고, μ 값은 0.001, 0.01, 0.1, 1, 10, 100을 이용하여 선정하였다.

예측 모델의 성능을 비교하기 위해 모델은 입력(input), 히든(hidden), 출력(output) 레이어로 구성된 신경망인 3-MLP(Multi-Layer Perceptron) Classifier를 이용하였다. 훈련 데이터와 테스트 데이터는 7:3 비율로 나눴고, Arrhythmia, Audiology, Pittsburgh 데이터는 다중 클래스 분류 예측, Ozone 데이터는 이진 클래스 분류 예측 실험을 진행하였다. 정확성은 One-Vs-Rest 방법을 이용하여 반복 실험 10번을 진행하여 측정하였으며 모든 실험 데이터의 히든 레이어 노드 개수는 10을 설정하였다.

4.2 실험 결과

첫 번째 실험은 제안 방법론인 SSLI을 이용하여 대체한 SSLI 데이터와 원시 데이터, 완전 레코드 데이터를 가지고 분류 예측 정확성 비교 실험 결과를 도출하였다. 그림 2는 각 데이터에 대한 분류 예측 정확성을 비교한 그림으로 SSLI으로 대체한 결과는 원시 데이터와 완전 레코드 데이터만 가지고 데이터 분류 실험했을 때보다 높은 정확성을 가지는 것을 확인하였다. Ozone 데이터는 레코드의 수가 2534개와 완전레코드의 수가 1847개로 다른 데이터에 비해 훈련 가능한 데이터가 많아 SSLI으로 결측치 대체했을 경우의 약 94 %의 정확성과 완전레코드만 가지고 분류 예측 실험했을 경우에도 약 92%의 높은 정확성을 보였다.

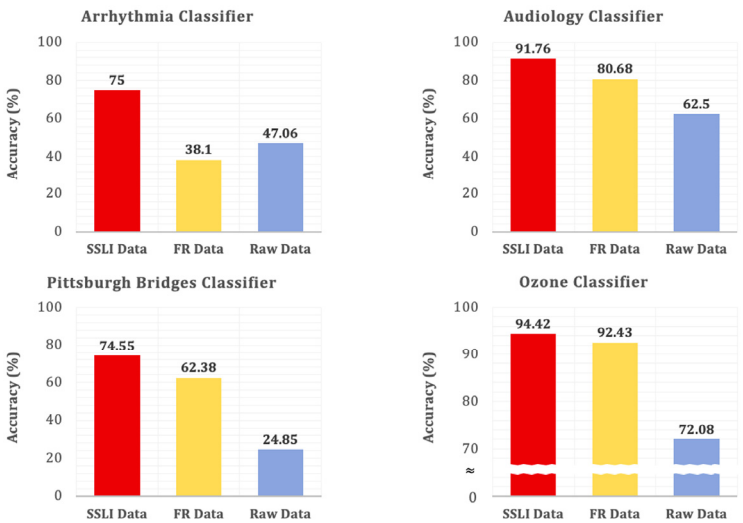


그림 2 첫 번째 실험: 결측치 대체의 중요성
Fig. 2 First Experiment: The Importance of Missing Value Imputation

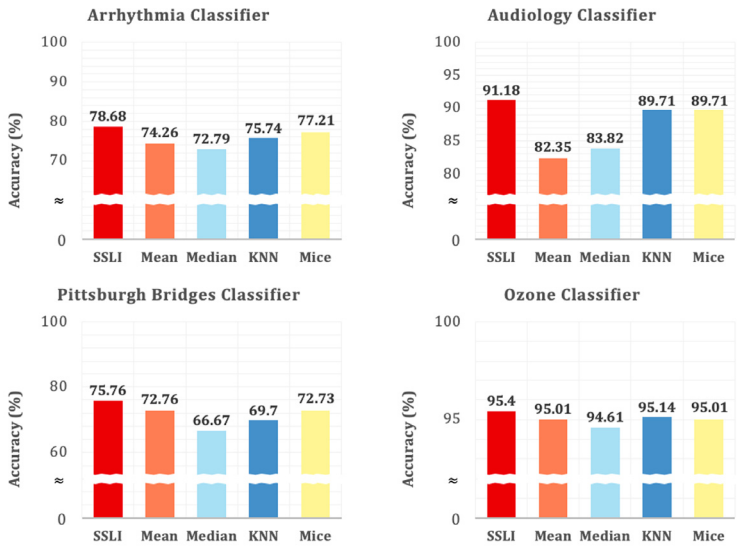


그림 3 두 번째 실험: 다양한 결측치 대체 방법의 분류 예측 정확성 비교

Fig. 3 Second Experiment: Comparing Classification Accuracy of Missing Value Imputation Methods

그림 3은 두 번째 실험으로 각 데이터에 대한 결측치 대체 방법에 따른 분류 예측 정확성을 비교한 그림이다. 결측치 대체 비교 실험으로 제안 방법론인 SSLI와 평균값, 중앙값, KNN, 그리고 Mice를 이용하여 분류 예측 정확성 비교 실험한 결과 SSLI로 결측치 대체한 데이터의 예측 성능이 가장 높은 것을 확인하였다. Arrhythmia 데이터의 분류를 위해 결측치를 SSLI으로 대체한 결과 다른 대체 방법들의 비해 약 78.68%, Audiology 데이터는 약 91.18%, Pittsburgh 데이터는 약 75.76%, 그리고 Ozone 데이터는 약 95.4%로 다른 결측치 대체 방법들에 비해 정확성이 높은 것을 확인하였다.

첫 번째 비교 실험은 결측치 처리 방법 중 데이터 삭제를 이용하여 데이터를 무분별하게 제거하거나 결측치가 포함된 데이터를 처리하지 않는 것보다 결측치를 대체하여 다양한 데이터를 효율적으로 활용하는 것이 중요하다는 것을 보였다. 두 번째 비교 실험의 경우, 기존 방법들은 한 속성의 전체 값으로 대체하거나 레코드의 특이성을 반영하지 않고 각 속성 간의 상관성을 고려하지 않은 대체 방법들보다 한계를 극복하여 제안한 대체 방법인 SSLI의 정확성이 훨씬 우수한 것을 확인하였다. 결측치 대체 방법 실험을 통해 한 레코드의 해당하는 각 속성들의 상관성과 레코드 간의 지역성이 함께 고려하는 것이 중요하다는 사실을 보여준다.

5. 결 론

본 연구에서는 제안한 결측치 대체 방법론은 속성의 전역성과 지역성을 함께 고려함과 동시에 한계점을 극복하였고 여러 결측치 대체 방법들과 비교 실험을 진행하였다. 실험 결과를 통해 제안 방법론이 기존의 대체 방법들보다 우수함을 보여 결측치 대체 방법으로 사용 가능하다. 결측치 대체 방법론을 이용하여 결측치 처리가 가능하고 완전 레코드 데이터의 비율을 높여 데이터 효용이 가능하며 데이터 마이닝을 통해 더욱 더 정확한 정보를 도출할 수 있다.

본 논문의 제안 방법론 중 속성 상관 네트워크는 데이터 속성의 상관성을 기반으로 속성 간 선형 관계일 경우를 구성하였다. 제안 방법론에서 결측치는 선형 관계에서 비례된 추정값 전파의 성능이 뛰어나 결측치 대체가 가능하지만 비선형 관계일 경우는 전파가 안 되는 한계점을 가지고 있다. 또한 수치적인 상관계수의 상관관계일 경우만 결측치 대체가 가능하여 수치가 아닌 다양한 데이터의 도메인 지식(Domain Knowledge)의 상관관계를 파악하여 상관 속성 네트워크를 구성하는 것은 한계가 있다. 수치가 아닌 특수한 데이터 도메인 지식의 상관관계 네트워크를 구성하여 결측치 대체하는 것이 성능이 더 뛰어날 것으로 판단된다. 따라서 선형 관계에서의 수치적 상관관계 파악이 아닌 선형과 비선형 관계를 함께 고려하고 다양한 도메인 지식의 상관성을 고려한 네트워크를 구성하는 연구가 필요하다.

추후 연구로는 비선형 관계의 결측치 추정이 가능하고 수치에 국한되지 않은 특수한 연관성에 기반한 연구가 진행할 예정이다. 실제 많은 산업체에서는 데이터에

대한 상관성 측면에서 수치적인 상관관계로 규정하지 않고 그 데이터만의 특수한 연관성에 기반을 두고 있는 경우가 많다. 산업체에서 이용하는 특수한 데이터나 음성, 이미지 데이터를 수치가 아닌 상관관계나 연관성에 대한 측면에서 다룰 예정이다.

References

[1] P. E. McKnight, K. M. McKnight, S. Sidani, and A. J. Figueredo, *Missing data: A Gentle Introduction*, Guilford Press, 2007.

[2] P. Schmitt, J. Mandel, and M. Guedj, "A Comparison of Six Methods for Missing Data Imputation," *Journal of Biometrics & Biostatistics*, Vol. 6, No. 1, pp. 1-6, 2015.

[3] G. E. Batista and M. C. Monard, "An Analysis of Four Missing Data Treatment Methods for Supervised Learning," *Applied Artificial Intelligence*, Vol. 17. No. 5-6, pp. 519-533, 2003.

[4] W. Tong and R. Jin, "Semi-Supervised Learning by Mixed Label Propagation," *Proc. of the 22nd National Conference on Artificial Intelligence*, Vol. 1, pp. 651-656, 2007.

[5] H. Cramér, Mathematical Methods of Statistics (PMS-9), Princeton University Press, Vol. 9, 2016.

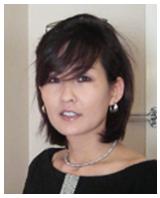
[6] G. Glass and K. Hopkins, Statistical Methods in Education and Psychology, *Psyccritiques*, 1996.

[7] G. V. Glass, "A Ranking Variable Analogue of Biserial Correlation: Implications for Short-cut Item Analysis," *Journal of Educational Measurement*, Vol. 2, No. 1, pp. 91-95, 1965.

[8] A. B. Goldberg, X. Zhu, and S Wright, "Dissimilarity in Graph-based Semi-Supervised Classification," *Artificial Intelligence and Statistics*, pp. 155-162, 2007.

신 유 경

2017년 덕성여자대학교 수학과, 컴퓨터학과 졸업(학사). 2019년 아주대학교 데이터사이언스학과 졸업(석사). 관심분야는 데이터 마이닝, 머신러닝, 딥러닝, 빅데이터 보안, 바이오메디컬 인포매틱스



신 현 정

2005년 서울대학교 산업공학과 졸업(박사) 2004년~2005년 Max-Planck-Institute for Biological Cybernetics 독일(연구원) 2005년~2006년 Friedrich-Miescher-Laboratory, Max-Planck-Institute 독일(수석연구원). 2006년 서울대학교 의과대학 연구교수. 2006년~현재 아주대학교 산업공학과, 데이터사이언스학과, 융합시스공학과 교수. 관심분야는 머신러닝, 데이터 마이닝, 바이오메디컬 인포매틱스