https://doi.org/10.5392/JKCA.2021.21.11.135

혼합형 데이터 보간을 위한 디노이징 셀프 어텐션 네트워크

Denoising Self-Attention Network for Mixed-type Data Imputation

이도훈*, 김한준, 전종훈*****

서울시립대학교 전자전기컴퓨터공학부*, 서울시립대학교 전자전기컴퓨터공학부**,
명지대학교 융합소프트웨어학부***

Do-Hoon Lee(yesdhlee95@uos.ac.kr)*, Han-Joon Kim(khj@uos.ac.kr),
Joonghoon Chun(jonghoonchun@gmail.com)*****

요약

최근 데이터 기반 의사결정 기술이 데이터 산업을 이끄는 핵심기술로 자리 잡고 있는바, 이를 위한 머신러닝 기술은 고품질의 학습데이터를 요구한다. 하지만 실세계 데이터는 다양한 이유에 의해 결측값이 포함되어 이로부터 생성된 학습된 모델의 성능을 떨어뜨린다. 이에 실세계에 존재하는 데이터로부터 고성능 학습 모델을 구축하기 위해서 학습데이터에 내재한 결측값을 자동 보간하는 기법이 활발히 연구되고 있다. 기존 머신러닝 기반 결측 데이터 보간 기법은 수치형 변수에만 적용되거나, 변수별로 개별적인 예측 모형을 만들기 때문에 매우 번거로운 작업을 수반하게 된다. 이에 본 논문은 수치형, 범주형 변수가 혼합된 데이터에 적용 가능한 데이터 보간 모델인 Denoising Self-Attention Network(DSAN)를 제안한다. DSAN은 셀프 어텐션과 디노이징 기법을 결합하여 견고한 특징 표현 벡터를 학습하고, 멀티태스크 러닝을 통해 다수개의 결측치 변수에 대한 보간 모델을 병렬적으로 생성할 수 있다. 제안 모델의 유효성을 검증하기 위해 다수개의 혼합형 학습 데이터에 대하여 임의로 결측 처리한 후 데이터 보간 실험을 수행한다. 원래 값과 보간 값 간의 오차와 보간된 데이터를 학습한 이진 분류 모델의 성능을 비교하여 제안 기법의 유효성을 입증한다.

■ 중심어 : | 머신러닝 | 딥러닝 | 데이터 품질 | 결측값 | 데이터 정제 | 어텐션 |

Abstract

Recently, data-driven decision-making technology has become a key technology leading the data industry, and machine learning technology for this requires high-quality training datasets. However, real-world data contains missing values for various reasons, which degrades the performance of prediction models learned from the poor training data. Therefore, in order to build a high-performance model from real-world datasets, many studies on automatically imputing missing values in initial training data have been actively conducted. Many of conventional machine learning-based imputation techniques for handling missing data involve very time-consuming and cumbersome work because they are applied only to numeric type of columns or create individual predictive models for each columns. Therefore, this paper proposes a new data imputation technique called 'Denoising Self-Attention Network (DSAN)', which can be applied to mixed-type dataset containing both numerical and categorical columns. DSAN can learn robust feature expression vectors by combining self-attention and denoising techniques, and can automatically interpolate multiple missing variables in parallel through multi-task learning. To verify the validity of the proposed technique, data imputation experiments has been performed after arbitrarily generating missing values for several mixed-type training data. Then we show the validity of the proposed technique by comparing the performance of the binary classification models trained on imputed data together with the errors between the original and imputed values.

■ keyword : | Machine Learning | Deep Learning | Data Quality | Missing Values | Data Imputation | Attention |

* 사사의 글: 이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No.2020-0-00121, 데이터 품질 평가기반 데이터 고도화 및 데이터셋 보정 기술 개발)과 과학기술정보통신부 및 정보통신기획평가원의 대학 ICT 연구센터지원사업으로 수행되었으며 (IITP-2021-2018-0-01417), 또한 2018년도 정부(교육부)의 재원으로 한국연구재단의 지원(No. NRF-2018R1D1A1A02086148, 시멘틱 텍스트 큐보이드 기반 자가증식 기술 및 딥러닝 학습 기술)을 받아 수행된 연구임

접수일자 : 2021년 09월 23일 심사완료일 : 2021년 10월 19일
수정일자 : 2021년 10월 19일 교신저자 : 김한준, e-mail : khj@uos.ac.kr

I. 서 론

빅데이터 시대가 되면서 데이터 기반 의사결정 방법론이 데이터 산업의 핵심으로 자리 잡게 되었다. 이를 위한 기반 기술로서 머신러닝은 주어진 학습데이터를 학습하여 특정 사물 또는 상황을 인식하거나 예측하는 학습모델을 생성한다. 그런데 기본적으로 고성능 학습모델을 생성하기 위해서는 고품질의 학습데이터가 필요하다. 하지만 실세계 데이터는 다양한 이유에 의해 결측값이 포함되게 마련이며, 이 결측값에 의한 정보 손실은 예측 모델의 성능 한계의 주요 원인이 된다. 실제로 고객 데이터를 이용하는 추천, 광고 분야의 경우 많은 결측값이 존재하며, 이를 해결하기 위해 여러 기법을 사용하여 전처리를 수행한다. 일반적으로 데이터 보간(Data Imputation) 기법을 활용하여 데이터 결측값 문제를 해결한다.

과거 통계 기반의 데이터 보간 기법은 평균값, 최빈값과 같은 기초 통계량을 활용하는데, 이는 구조가 복잡한 대용량 데이터에 적용하기에는 신뢰도가 매우 떨어져 사용하기 어렵다. 그래서 최근 데이터 보간의 신뢰도를 높이고자 머신러닝 기술이 활용되고 있다. 즉 결측값을 대체하는 값을 예측하는 학습모델을 구축하기 위해 머신러닝 기술이 활용되는 것이다. 그러나 대부분의 머신러닝 기반 데이터 보간 기법들은 수치형 변수에만 적용되거나, 변수별로 예측모델을 구축하기 때문에 번거로운 작업을 수반하게 된다. GAIN[1]과 같은 심층 생성 모델(Deep Generative Model) 기반의 보간 기법은 범주형 변수에 대한 대체값을 생성하지 못하여 범주형 변수가 혼합된 데이터에는 적용하기 어렵다. 반면 MissForest[3], DataWig[5]와 같이 예측 모델 기반의 데이터 보간 기법은 변수별로 예측 모델을 생성하기 때문에, 혼합형 데이터에 적용 가능하지만 비효율적이다.

본 논문은 수치형과 범주형 변수가 공존하는 혼합형 데이터에 대해 별도의 가공 과정 없이 바로 활용 가능한 데이터 보간 기법인 Denoising Self-Attention Network(DSAN)를 제안한다. DSAN은 셀프 어텐션(Self-Attention) 기반 특징 표현 기법과, 디노이징 기법을 결합하여 입력 데이터 내 결측 데이터에 대해 견고한(robust)한 특징을 학습한다. 이후 멀티태스크 학습(Multi-task Learning) 개념을 반영하여 변수별 대체값을 예측하도록 학습한다.

본 논문의 구성은 다음과 같다. 2장은 관련 연구로 데이터 보간 기법과 정형 데이터에 이용되는 어텐션 메커니즘에 대해 소개한다. 3장은 본 논문에서 제안하는 데이터 보간 모델인 DSAN에 대해 설명한다. 4장은 제안 모델의 유효성을 입증하기 위해, 여러 혼합형 데이터를 이용해 실험한 결과를 제시하고, 마지막 5장은 결론 및 향후 연구를 서술한다.

II. 관련 연구

1. 데이터 보간 기법

머신러닝 기반 의사결정 방법론에 활용되는 데이터는 그 품질이 매우 높아야 하는데, 많은 경우 결측값이 다수 포함되어 있어 고신뢰도의 학습모델을 구축하는데 어려움을 겪는다. 데이터 보간 기법은 결측값을 대체값으로 채워 넣기 위해 통계 또는 머신러닝 알고리즘을 사용한다[2]. 통계 기반의 보간 기법은 평균값, 최빈값과 같은 기초 통계량을 대체값으로 사용하는 방식과 회귀모델을 만들어 대체값을 예측하는 방식을 포함한다. 기초 통계량 기반 보간 기법은 쉽고 빠르게 적용할 수 있는 장점이 있지만, 데이터의 크기가 커질수록 유효성이 떨어지는 문제점이 있고, 회귀모델 기반 보간 기법은 수치형 변수에만 적용되고 그 정확도가 높지 못하는 문제점을 안고 있다. 이를 보완하기 위해 최근 k-최인접이웃(k-Nearest Neighbors), 지지벡터머신(Support Vector Machine), 랜덤포레스트(Random Forest), 인공신경망(Neural Networks)과 같은 머신러닝 기반의 보간 기법들이 연구되고 있다[2]. 머신러닝 기반 보간 기법은 결측값이 존재하는 변수에 출현한 관측값을 가지고 학습모델을 만들어 대체값을 추정하는 형태가 일반적이다. 대표적인 머신러닝 기반 보간 기법인 MissForest[3]는 각 관측 변수를 이용하여 결측 변수별로 Random Forest 기반의 모델을 구축하여 대체값을 예측한다. Neural Networks 기반 보간 기법의 경우 GAIN[1], HIVAE[4]와 같은 생성 모델을 이용하

여 결측값에 대한 대체값을 생성한다. 하지만 생성 모델은 연속적인 수치형 변수만을 생성하기 때문에, 영상 보간 문제에는 적합하지만 일반적인 테이블 데이터와 같이 범주형 변수가 존재하는 혼합형 데이터의 경우 적합하지 않다. 이 경우 DataWig[5]와 같이 변수별로 분류모델을 사용해서 해당 변수 도메인 내 적절한 대체값을 예측하여 활용한다.

2. 정형 데이터 학습을 위한 어텐션 기법

어텐션(Attention) 메커니즘은 인공신경망 분야에서 최근 가장 활발하게 연구되고 있는 중요한 기법이다. 초기 어텐션 기법은 자연어 처리 분야에서 시작하여, 최근에는 컴퓨터 비전[6], 정형 데이터 학습[7][8] 등 인공신경망을 응용한 대부분 분야에 적용되고 있다. 어텐션의 기본 아이디어는 특정 사물에 대한 사람의 인식 과정에서 유래한다. 예를 들어, 우리가 특정 이미지를 바라볼 때, 불필요한 영역은 무시하고 일부 영역에만 집중적으로 관찰하여 주어진 이미지의 주요 정보를 얻는 경향이 있다. 이와 유사하게 어텐션 기법은 신경망이 입력 데이터에 대해 정답과 관련 있는 특정 영역에만 집중하도록 학습하여 더 나은 인식 결과를 얻게 한다. 구체적으로, 어텐션 함수 A 는 입력으로 주어진 쿼리 q에 대해 모든 키 k_i와의 유사도를 구한다. 여기서 키 k_i는 정답과 관련 있는 특정 부분을 찾기 위해 탐색하는 정보에 해당한다. 각 키 k_i와의 유사도를 계산한 값 $a(k_i, q)$을 소프트맥스(Softmax)와 같은 분포 함수를 통해 합이 1인 어텐션 가중치로 변환하여 키 k_i와 맵핑된 값 v_i에 반영한다(수식 1 참조).

$$A(q, K, V) = \sum_i p(a(k_i, q)) \cdot \quad v_i \tag{1}$$

이러한 어텐션 메커니즘이 최근에는 정형데이터 도메인에 적용되는 연구가 활발히 진행되고 있다. 예를 들어, TabNet[7]은 정형데이터에 잘 작동하는 의사결정 트리(Decision Tree)의 구조를 모방하기 위해 어텐션 메커니즘을 활용하였다. TabNet은 순차적 어텐션을 사용하여 각 의사결정 단계에서 중요한 특징(Feature)을 선택함으로써 해석 가능하고 효율적인 학습을 가능하게 하였다. TabTransformer[8]는 셀프 어텐션 메커니즘을 기반으로 한 트랜스포머(Transformer)[9]의 인코더를 사용하였다. 셀프 어텐션을 통해 범주형 변수에 대한 문맥적 임베딩(Contextual Embedding)을 학습하여 높은 예측 정확도를 성취하였다.

3. 데이터 보간을 위한 디노이징 기법

디노이징(Denoising)은 입력 데이터에 일부 손상을 주어 학습함으로써 견고한 특징 표현을 얻는 기법이다. 이런 학습 방식을 통해 학습한 오토인코더를 디노이징 오토인코더(Denoising AutoEncoder)[10]라고 하며, 일반적으로 오토인코더보다 상대적으로 더 나은 성능을 보인다.

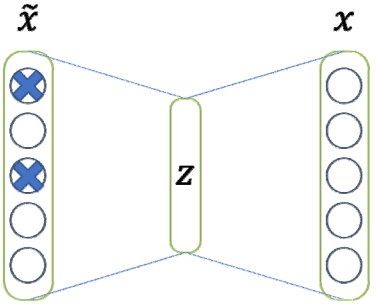


그림 1. Denoising AutoEncoder

[그림 1]은 디노이징 오토인코더의 학습 원리를 보여주며, 이는 손상된 입력 $\tilde{x}$을 받아들여 손상되지 않은 입력 x을 복원하도록 학습한다. 이러한 디노이징 기법은 노이즈한 입력에 대해 우수한 성능을 갖기 때문에, 이 기법은 자연스럽게 결측값 데이터 보간 영역에 적용되었다 [11]. 결측값은 하나의 노이즈한 입력 형태로 볼 수 있으며, 다양한 도메인 데이터에 대한 보간에 있어 디노이징 기법이 활용될 수 있다[12].

표 1. 입력 데이터(혼합형 결측 데이터) 예시

age	job	martial	education	default	···	duration	campaign	pdays	poutcome	deposit
59	admin.	married	secondary	no	···	1042	1	NaN	0	yes
56	admin.	married	secondary	no	···	1467	1	−1	0	yes
41	technician	married	secondary	no	···	NaN	1	−1	0	yes
55	services	married	secondary	no	···	579	1	−1	0	yes
54	admin.	NaN	NaN	no	···	673	2	−1	0	yes

표 2. 출력 데이터(정제된 데이터) 예시

age	job	martial	education	default	···	duration	campaign	pdays	poutcome	deposit
59	admin.	married	secondary	no	···	1042	1	**−1**	0	yes
56	admin.	married	secondary	no	···	1467	1	−1	0	yes
41	technician	married	secondary	no	···	**1389**	1	−1	0	yes
55	services	married	secondary	no	···	579	1	−1	0	yes
54	admin.	**married**	**tertiary**	no	···	673	2	−1	0	yes

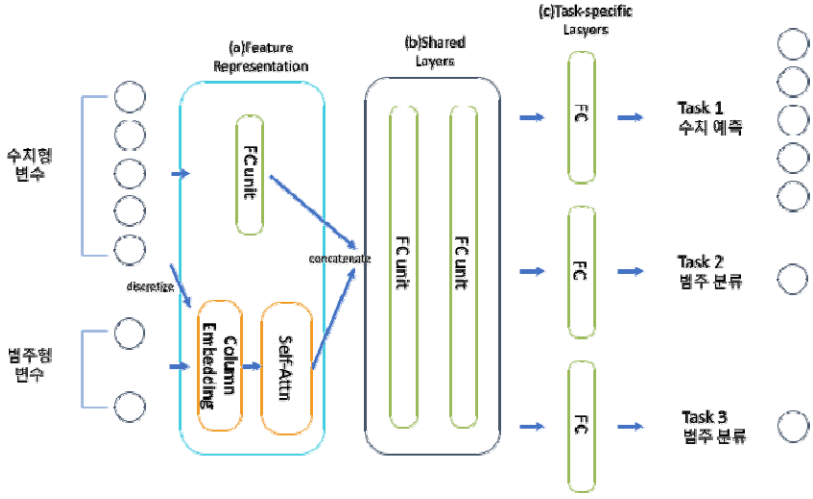


그림 2. Denoising Self-Attention Network

Ⅲ. 디노이징 셀프 어텐션 네트워크

1. 문제 정의

본 논문은 결측값이 포함된 혼합형 데이터에 대해 대체값을 예측하는 것을 목표로 한다. 혼합형 데이터란 수치형 변수와 범주형 변수가 혼합되어 있는 데이터를 의미하며, [표 1]은 입력되는 혼합형 데이터의 예시를 보여준다. 이는 본 연구의 성능평가에서 활용된 'Bank' 데이터셋의 일부로서, 7개 수치형 변수와 10개 범주형 변수를 포함한다.

각 입력 데이터 레코드는 $x = (x_{cont}, x_{cat})$와 같이 표현될 수 있다. $x_{cont} \in R^n$는 입력 데이터 레코드에서 관

측되는 n차원의 수치형 변수 벡터를 의미하며, $x_{cat} = (x_1, x_2, \cdots, x_k)$은 입력 데이터 레코드에서 관측되는 k개의 범주형 변수를 의미한다. 결측값 보간 기법의 목표는 불완전한 입력 데이터의 결측값 $X_{missing}$에 대해 적절한 대체값 $X_{missing}^{imp}$을 예측하여 [표 2]와 같은 완전한 데이터 $X_{complete}$로 변환 정제하는 것이다.

2. 디노이징 셀프 어텐션 네트워크 (DSAN)

본 논문이 제안하는 디노이징 셀프 어텐션 네트워크의 구조는 [그림 2]와 같으며, 멀티태스크 러닝 기법을 결합하여 혼합형 데이터를 복원하는 구조이다. DSAN은 크게 3가지 모듈로 구성된다. 첫 번째 특징 표현 모듈에서 혼합형 데이터를 입력받아 각 변수에 대한 임베딩 특징 벡터를 학습하고, 셀프 어텐션 레이어를 통해 변수 간 연관성을 학습한다. 두 번째 공유 레이어 모듈은 각 변수 예측에 필요한 공유 파라미터를 학습하며, 마지막 태스크 개별 레이어 모듈은 각 변수 예측에 필요한 독립적인 파라미터를 학습한다.

2.1 특징 표현 모듈

DSAN은 먼저 혼합형 데이터의 각 변수에 대해 칼럼 임베딩을 적용하여 일정 차원의 임베딩 행렬 $E \in R^{(n+k)\times d}$를 계산한다. 칼럼 임베딩(Column Embedding)은 각 칼럼의 이산값을 d 차원의 연속형 실수 벡터 $e_i = \phi(x_i) \in R^d$ 로 표현하는데, 이때 수치형 변수의 경우 이산화의 과정이 필요하다. 따라서 수치형 변수의 경우 이산화 함수 f_{disc}를 통해 이산 값으로 변환 후 임베딩 함수 ϕ 를 적용하며, 수식 2과 같이 임베딩 행렬 E를 얻는다.

$$E = \phi(f_{disc}(x_{cont}), x_{cat}) \quad (2)$$

임베딩 행렬 E는 이후 셀프 어텐션 레이어의 입력이 되며, 각 변수 간 연관성 정보를 포함하는 문맥적 임베딩 행렬 $H \in R^{(n+k)\times d}$을 학습한다. 이 때 어텐션은 멀티 헤드 어텐션을 사용하여, 다수의 특징 표현을 병렬적으로 학습하게 하였다[9]. 수치형 변수를 이산화하는 과정에서 일부 정보 손실을 야기하게 되며, 이를 보완하기 위해 이산화하지 않은 수치형 변수값을 FC(Fully Connected) 유닛을 통해 병렬적으로 학습한다. FC 유닛은 FC 레이어, 레이어 정규화(Layer Normalization), ReLU함수로 구성되며, 이는 이어지는 공유 레이어 모듈에도 적용된다. 이때 FC유닛의 노드는 임베딩 차원 d의 노드를 갖도록 하여 차원의 통일성을 주었다. 병렬적으로 학습한 FC 유닛은 수치형 변수에 대한 정보를 유지하도록 도와주며, 차후에 셀프 어텐션을 통해 계산된 문맥적 임베딩 행렬과 결합하여 공유 레이어 모듈의 입력으로 활용된다. 이때 문맥적 임베딩 행렬 H를 펼치고(flatten) FC 유닛에서 계산된 벡터를 결합하여 최종적으로 공유 레이어로 넘어가는 특징 벡터는 $(n+k+1)\cdot d$차원을 갖게 된다.

2.2 디노이징 기법

앞서 언급한 바와 같이, 우리는 결측 입력에 대한 견고한 특징 표현 파라미터를 학습하기 위해 디노이징 기법을 적용한다. 이를 적용하기 위해 우선 기존 입력에 대해 일정 비율의 변수들을 무작위로 선택하여 제거한다. 무작위로 선택된 변수가 수치형 변수면 0으로 초기화하고, 범주형 변수인 경우 해당 칼럼이 결측임을 나타내는 특이값으로 채워 넣는다. 예를 들어 노이즈를 주입하기로 선택된 3번째 칼럼이 범주형 변수일 경우 "Col3:NA"와 같이 해당 변수의 값이 결측임을 나타내는 특이값으로 초기화한다. 이때 DSAN은 결측 여부를 나타내는 특이값에 대해서도 임베딩 벡터를 학습하게 하여 결측값에 의한 정보 손실을 보완하고, 임의로 생성되는 다양한 결측 패턴에 대한 특징 표현을 학습하게 하였다. 이후 연결되는 셀프 어텐션 레이어는 특정 변수의 결측 여부와 다른 변수의 관측값 간 연관성을 학습하며, 후에 태스크 개별 레이어 모듈에서 결측 처리한 원래값을 정답으로 학습하여, 적절한 대체값을 예측하도록 하였다.

2.3 공유 레이어 및 태스크 개별 레이어 모듈

DSAN은 멀티태스크 러닝 기법을 활용하여, 수치형 변수 벡터와 범주형 변수별로 예측 작업을 수행한다. 모델 $f = (f_1, f_2, \ldots, f_T)$는 $|T| = 1 + k$개의 태스크를 병렬적으로 학습하며, 각 f_t는 태스크 t에 해당하는 독립 파라미터 θ_t를 가진 모델을 의미한다. 각 태스크가

서로의 보조 태스크(Auxiliary Task)의 역할을 하여 정칙화(Regularization) 효과를 주며 이를 일반화해서 표현하면 수식 3와 같다.

$$f_t(\tilde{x};\theta_{sh},\theta_t): X \rightarrow Y_t \quad (3)$$

공유 파라미터 θ_{sh}를 통해 공통으로 도움이 되는 특징을 학습하고, 독립적인 파라미터 θ_t를 통해 각 변수별로 예측에 필요한 특징을 학습한다. 수치형 변수 벡터 x_{cont}에 대해 수식 4과 같이 결측 입력 $\tilde{x}$를 입력으로 수치형 벡터 관측값 $\hat{x}_{cont}$을 복원하도록 학습하며, 이때 θ_{cont}는 수치형 변수 벡터를 복원할 때 필요한 특징들을 학습하는 파라미터 집합이다. 각 범주형 변수 x_i $i=(1,2,\cdots,k)$에 대해서는 수식 5과 같이 결측 입력 $\tilde{x}$를 입력으로 범주형 변수 i 내의 범주값일 확률 $\hat{y}_i$를 예측하도록 학습한다. 이때 θ_i는 i번째 범주형 변수의 분류 태스크를 수행할 때 필요한 특징들을 학습하는 파라미터 집합이다. 수식 5의 σ함수는 확률값으로 대응하기 위한 활성화 함수이며 이진 분류의 경우 시그모이드(Sigmoid) 함수를 이용하고, 다중 분류의 경우 소프트맥스 함수를 이용한다.

$$\hat{x}_{cont} = f_{cont}(\tilde{x};\theta_{sh},\theta_{cont}) \quad (4)$$

$$\hat{y}_i = \sigma(f_i(\tilde{x};\theta_{sh},\theta_i)) \quad (5)$$

관측값에 대하여 손실 함수를 계산하기 위해 입력 변수별 결측 정보를 나타내는 이진 벡터 $m=(m_{cont},m_{cat})$을 사용한다. 즉, 해당 변수가 관측값을 가지면 1로, 결측이면 0으로 표현한다. 이 때 디노이징 기법을 적용하기 위해 임의로 제거한 변수 값의 경우, 해당 변수에 대해 관측값과 같이 1로 표현한다. 수치형 변수 벡터의 경우 수식 6 과 같이 평균제곱오차(Mean Squared Error) 손실함수 L_{cont}를 통해 학습되며, 범주형 변수의 경우 수식 7과 같이 교차 엔트로피(Cross Entropy) 손실함수 $L_{cat,i}$를 통해 학습된다. DSAN은 수식 8과 같이 각 손실 함수의 총합을 최소화하는 파라미터 집합 Θ을 추정한다.

$$L_{cont} = \sum((x_{cont}-\hat{x}_{cont})^2 \odot m_{cont}) \quad (6)$$

$$L_{cat,i} = -m_i \cdot \sum y_i \cdot \log(\hat{y_i}) \quad (7)$$

$$\mathrm{argmin}_\Theta \; L_{cont} + \sum_{i=1}^{k} L_{cat,i} \quad (8)$$

Ⅳ. 실험 및 결과

1. 실험 데이터 및 방법

우리는 제안 기법의 유효성을 보이기 위해 [표 3]에서 제시한 3개 혼합형 데이터셋을 가지고 보간 실험을 수행하였다. 본 연구에서 사용된 실험 데이터는 UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/index.php)에 적재된 'Adult', 'Bank', 'Titanic' 데이터셋이며, 이 데이터셋들은 수치형 변수와 범주형 변수를 포함하고 있다.

표 3. 실험 데이터셋

데이터셋	범주형 변수 개수	수치형 변수 개수	레코드 개수
Adult	9	6	30162
Bank	10	7	11162
Titanic	4	4	712

데이터 보간 실험은 3 단계로 이루어진다. 1 단계는 결측값이 존재하지 않는 완벽한 테이블 데이터에 대해 MCAR(Missing Completely At Random) [13] 방식으로 5 ~ 20%의 결측값을 생성한다. 2 단계는 결측 처리된 값에 대해 결측값 보간을 수행하고, 보간된 값 X^{imp}과 본래값 X^{true}의 오차를 측정한다. 여기서, 오차에 대한 평가 척도로서, 우리는 수치형 변수의 경우 NRMSE(식 8 참조)를, 범주형 변수의 경우 Error Rate(식 9 참조)를 각각 사용한다.

3 단계는 본래 데이터와 보간된 데이터를 이용해 간단한 이진 분류를 수행하고 모델 성능을 비교 분석한다. 보간된 데이터를 이용한 분류기의 성능과 원래 데이터를 이용한 분류기 성능의 차이가 적을수록, 주어진 입력 데이터를 보다 효과적으로 보간하였다고 평가할 수 있으며, 이때 우리는 분류기 성능 평가척도로서 AUC-ROC 값을 사용한다.

2. 실험 결과

본 논문이 제안한 데이터 보간 기법의 유효성을 입증하기 위해서, 우리는 MissForest[3]를 비교 기법으로 설정하였다. MissForest는 혼합형 데이터에 적용 가능한 랜덤 포레스트 기반의 데이터 보간 기법이며, 각 변수별로 모델을 만들어 대체값을 추정한다. 우리가 제안하는 모델은 신경망 모델을 사용하고 단일 모델로 대체값을 추정함에 있어 차이가 있다. 실험에 이용된 모델은 $d=32$의 칼럼 임베딩을 수행하였으며, 8개 헤드의 멀티 헤드 어텐션을 이용하였다. 앞서 언급한 바와 같이 특징 표현 모듈의 FC 유닛의 경우 임베딩 차원과 같은 32개의 노드를 갖도록 설정하였다. 공유 레이어의 경우 입력되는 특징 벡터의 차원과 같은 $(n+k+1)\cdot d$ 개의 노드를 갖는 2층의 FC모듈을 사용하였다. 이후 태스크 개별 모듈의 경우 수행하는 태스크에 맞추어 수치형 변수 예측 FC 레이어는 n개, 범주형 변수 예측 FC 레이어는 각 변수 도메인 크기(범주 수)에 맞는 노드를 갖는다. 이때 이진 분류 태스크의 경우 단일 노드를 갖는 FC 레이어가 된다. 모델 학습에 있어 학습률(learning rate)은 0.003로 설정하였고, Adam 알고리즘[14]을 사용하여 30 에포크(epoch) 만큼의 학습을 수행하였다. [표 4]는 제안 기법과 비교 기법에 대한 데이터 보간 실험 결과를 보여주며, 결측률을 5% 단위로 증가시키면서 실험을 수행하였다. 또한 특정 레코드에 편향되는 것을 방지하기 위해 5-겹 교차검증(5-fold Cross Validation)을 수행하였으며, 각 실험 결과 수치는 5-겹 교차검증을 통해 측정된 성능지표에 대한 평균값이다. 각 실험 결과를 시각화한 [그림 3-그림 5]의 x축은 결측률(Missing Percent)이며, y축은 해당 지표를 의미한다.

표 4. 데이터 Imputation 실험 수행 결과

데이터셋	평가 척도	모델	5% 결측	10% 결측	15% 결측	20% 결측
Adult	NRMSE	MissForest	0.5629	0.5369	0.5571	0.5468
		DSAN	0.5228	0.5235	0.5158	0.5175
	ErrorRate	MissForest	0.2159	0.2237	0.2337	0.2442
		DSAN	0.2153	0.2221	0.2288	0.2371
	AUC-ROC	complete data	0.9049			
		MissForest	0.9048	0.9046	0.9041	0.9033
		DSAN	0.9047	0.9045	0.9040	0.9039
Bank	NRMSE	MissForest	0.8859	0.9890	1.0107	1.0228
		DSAN	0.9178	0.7986	0.8637	0.8893
	ErrorRate	MissForest	0.2448	0.2548	0.2636	0.2726
		DSAN	0.2634	0.2673	0.2741	0.2881
	AUC-ROC	complete data	0.9027			
		MissForest	0.9027	0.9025	0.9021	0.9019
		DSAN	0.9026	0.9022	0.9025	0.9017
Titanic	NRMSE	MissForest	0.9922	0.6578	0.7192	0.7078
		DSAN	0.7785	0.6937	0.6017	0.7043
	ErrorRate	MissForest	0.1860	0.2112	0.2141	0.2309
		DSAN	0.2561	0.2227	0.2420	0.2381
	AUC-ROC	complete data	0.8466			
		MissForest	0.8535	0.8546	0.8536	0.8525
		DSAN	0.8552	0.8530	0.8527	0.8467

[그림 3]은 수치형 결측값 보간에 대한 NRMSE(수식 8 참조)척도 값을 보여준다. NRMSE는 수치형 변수에 대해 원래값과 대체값의 오차이며, 값이 낮을수록 원래값에 근사한 대체값을 예측하였음을 의미한다. 그림에서 보는 바와 같이 제안 기법 DSAN이 비교 기법 MissForest 대비 평균 7% 우수한 성능을 보였다. 이는 신경망 기반 모델이 트리 기반 모델에 비해 수치형 예측에 강점이 있음을 의미한다.

$$NRMSE = \sqrt{\frac{E((X_{missing}^{true} - X_{missing}^{imp})^2)}{Var(X_{missing}^{true})}} \quad (8)$$

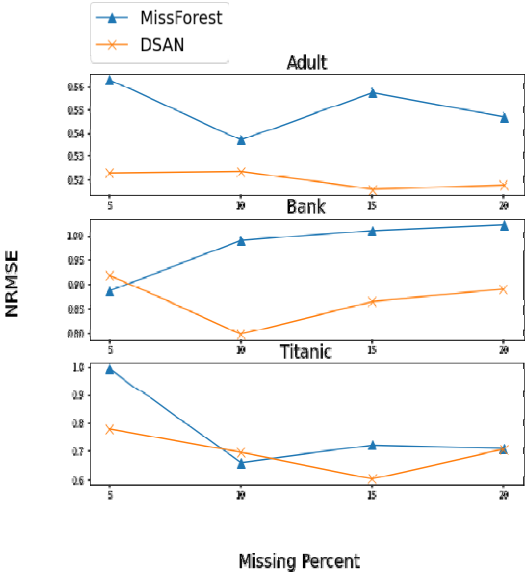


그림 3. NRMSE 척도 기준 성능 비교

[그림 4]는 정제된 범주형 변수값과 본래 범주형 변수값의 오차율 Error Rate(수식 9 참조)을 시각화한 것이다. 변수별로 오차율을 계산한 이후, 평균값을 계산하여 모델의 성능을 평가하였다. 전체적으로 제안 기법이 MissForest 대비 다소 저조한 성능을 보이지만, 가장 용량이 큰 데이터셋인 'Adult' 데이터의 경우에는 DSAN이 약 0.7%의 나은 성능을 보였다. 또한 DSAN 기법은 데이터셋의 레코드 수가 클수록 보간 성능이 우수한 것으로 나타났으며, 이는 제안 기법이 빅데이터의 품질 개선에 효과적으로 활용될 수 있음을 시사한다. 이는 신경망 기반 모델이 학습하는 파라미터 수가 크고 복잡하므로, 데이터의 양이 많아질 때 더욱 복잡한 패턴을 잘 학습할 수 있기 때문이다.

$$Error\ Rate = \frac{| X_{missing}^{true} \neq X_{missing}^{imp} |}{| X_{missing} |} \quad (9)$$

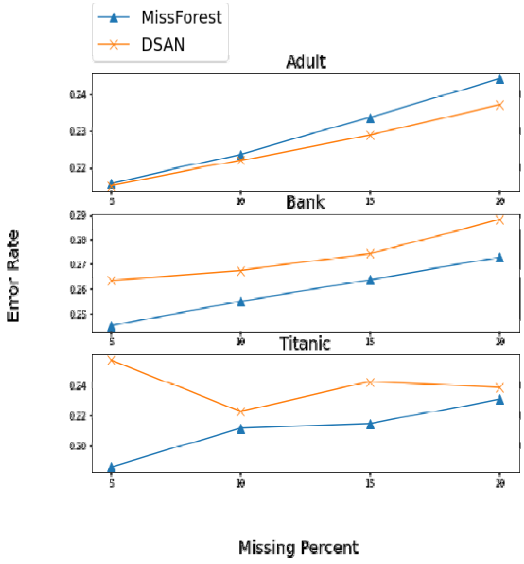


그림 4. Error Rate 척도 기준 성능 비교

[그림 5]의 경우 보간된 데이터를 이용하여 간단한 이진 분류 모델의 성능을 비교한 그래프이다. 'Adult' 데이터의 경우 근로자 수입이 5만 달러를 초과하는지 예측하도록 학습을 수행하였고, 'Bank' 데이터의 경우 고객의 예금 상품 가입 여부, 'Titanic'의 경우 승객의 생존 여부를 예측하도록 학습을 수행하였다. 결과적으로 결측값이 존재하지 않는 완전한 본래 데이터를 이용한 이진 분류 모델 성능과 비교하여, 제안 기법에 의해 보간된 데이터를 사용하여 구축한 분류 모델의 성능이 평균 0.1% 미만인 것으로 나타났다. 다시 말해서, 결측치가 다수 포함된 데이터셋이 제안 기법에 의해 개선된 데이터가 학습데이터로서 예측 모델을 구축하는 데 활용될 수 있음을 보여주는 것이다.

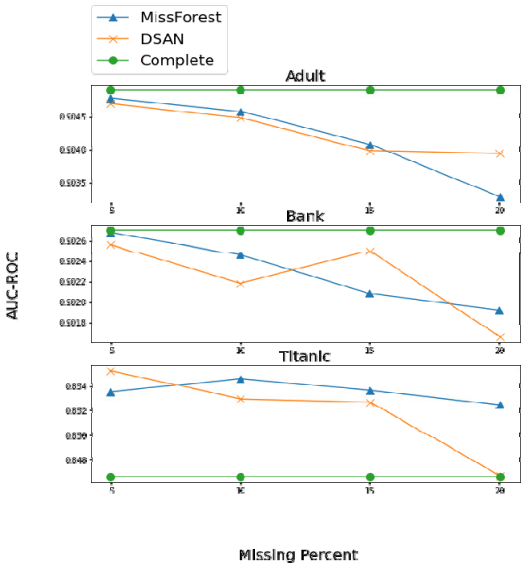


그림 5. AUC-ROC 척도 기준 성능 비교

V. 결론

본 논문은 결측값이 존재하는 혼합형 데이터의 결측값 보간 문제를 해결하기 위한 디노이징 셀프 어텐션 네트워크(DSAN)를 제안하였다. 제안 기법의 데이터 보간 성능의 유효성을 입증하기 위해 3개의 혼합형 학습 데이터셋에 대하여 결측값을 생성한 후 보간하는 실험을 수행하였다. 본래값과 추정한 대체값 간의 오차를 비교 분석하여 유효성을 입증하였으며, 수치형 변수값의 보간 실험 결과를 통해, 비교 기법인 MissForest 대비 제안 기법이 평균적으로 7% 향상된 성능을 보였다. 범주형 변수값의 보간 실험 결과 MissForest 대비 평균적으로 저조한 성능을 보였지만, 데이터 레코드 수가 증가할수록 DSAN의 보간 성능이 우수함을 보였으며, 실험 데이터셋 중 가장 레코드 수가 많은 'Adult' 데이터셋의 경우 DSAN이 MissForest 대비 다소 향상된 성능을 보였다. 또한 제안 기법을 통해 보간된 데이터를 학습한 분류 모델의 성능과 본래 데이터를 학습한 분류 모델의 성능 차이가 0.1% 미만으로서 거의 차이가 없음을 확인하였다. 이는 제안 기법이 고품질 학습 데이터를 구축하여 고성능의 모델 학습에 기여할 수 있음을 기대할 수 있다. 특히 수집하지 못해 추정해야 하는 결측값이 다수 존재하고, 실시간으로 이벤트 로그 데이터가 쌓여 대용량의 데이터를 취급하는 전자상거래(e-commerce) 추천, 모바일 광고 분야에 활용될 수 있을 것으로 판단된다.

향후 연구로는 상대적으로 부족한 범주형 변수 정제 성능을 올리기 위한 모델 구조 개선과 멀티 태스크 러닝 특성상 일부 성능을 떨어뜨리는 태스크에 대한 최적화 연구가 필요하다.

참 고 문 헌

[1] J. Yoon, J. Jordon, and M Schaar, "GAIN: Missing Data Imputation using Generative Adversial Nets," International Conference on Machine Learning, pp.5689-5698, 2018.

[2] W. Lin and C. Tsai, " Missing value imputation: a review and anlaysis of the literature (2006-2017)," Artificial Intelligence Review, Vol.53, No.2, pp.1487-1509, 2020.

[3] D. J. Stekhoven and P. Buhlmann, "MissForest-non-parametric missing value imputation for mixed-type data," Bioinformatics, Vol.28, No.1, pp.112-118, 2012.

[4] A. Nazabal, P. Olmos, Z. Ghahramani, and I. Valera, "Handling Incomplete Heterogeneous Data using VAEs," Pattern Recognition, Vol.107, 2020.

[5] F. Biessmann, T. Rukat, P. Schmidit, P. Naidu, S. Schelter, A. Taptunov, D. Lange, and D. Salinas, "Datawig: Missing Value Imputation for Tables," Journal of Machine Learning Research, Vol.20, 2019.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehgani, M. Minderer, G. Heigold, S. Gelly, J. Uszkreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," arXiv preprint arXiv:2010.11929, 2020.

[7] S. O. Arik and T. Pfister, "TabNet: Attentive

Interpretable Tabular Learning," Proceedings of the AAAI Conference on Artificial Intelligence, Vol.35, No.8, pp.6679-6687, 2021.

[8] X. Huang, A. Khetan, M. Cvitkovic, and Z. Karnin, "TabTransformer: Tabular Data Modeling Using Contextual Embeddings," arXiv preprint arXiv:2012.06678, 2020.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All you need," Advances in Neural Information Processing Systems, pp.5998-6008, 2017.

[10] P. Vincent, H. Larochelle, Y. Bengio and P. A. Manzagol, "Extracting and Composing Robust Features with Denoising Autoencoders," Proceedings of the 25th International Conference on Machine Learning, pp.1096-1103, 2008.

[11] N. Abiri, B. Linse, P. Eden, and M. Ohlsson, "Establishing Strong Imputation Performance of a Denoising Autoencoder in a wide range of missing data problems," Neurocomputing, Vol.365, pp.137-146, 2019.

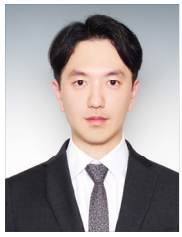
[12] L. Gondara and K. Wang, "Mida: Multiple imputation using denoising autoencoders," Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp.260-272, Springer, 2018.

[13] D. B. RUBIN, "Inference and missing data," Biometrika, Vol.63, No.3, pp.581-592, 1976.

[14] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

저 자 소 개

이 도 훈(Do-Hoon Lee) 준회원



- 2020년 : 서울시립대학교 건축공학과/통계학과 (공학사/이학사)
- 2020년 ~ 현재 : 서울시립대학교 전자전기컴퓨터공학부 석사과정

〈관심분야〉 : 머신러닝, 딥러닝, 빅데이터 기술

김 한 준(Han-Joon Kim) 정회원

- 1994년 : 서울대학교 계산통계학과 (이학사)
- 1996년 : 서울대학교 전산과학과 (이학석사)
- 2002년 : 서울대학교 컴퓨터공학부 (공학박사)
- 2002년 ~ 현재 : 서울시립대학교 전자전기컴퓨터공학부 정교수

〈관심분야〉 : 머신러닝, 빅데이터 분석, 텍스트 마이닝, 데이터베이스, 정보검색

전 종 훈(Jonghoon Chun) 정회원

- 1986년 : Computer Science, University of Denver(학사)
- 1988년 : Computer Science, Northwestern University(공학석사)
- 1992년 : Computer Science, Northwestern University(공학박사)
- 1992년 ~ 1995년 : University of Central Oklahoma Department of Computing Science 조교수
- 1995년 ~ 현재 : 명지대학교 융합소프트웨어학부 정교수

〈관심분야〉 : 빅데이터, 데이터베이스, 정보검색, 지능형 소프트웨어, 의료정보시스템