디노이징 셀프 어텐션 네트워크 기반 정형 데이터 결측값 보간 기법

이도훈, 김한준*
서울시립대학교

yesdhlee95@uos.ac.kr, *khj@uos.ac.kr

Structured Data Missing Value Imputation Based Denoising Self-Attention Network

Do-hoon Lee, Han-joon Kim*
University of Seoul

요 약

기존 머신러닝 기반 정형 데이터 결측값 보간 기법은 수치형 변수에 한해 적용되거나, 변수별로 예측 모형을 만들기 때문에 비효율적이다. 이에 본 논문은 수치형 변수와 범주형 변수가 혼합되어 있는 정형 데이터에 직접적으로 적용 가능한 결측값 보간 모델 DSAN(Denoising Self-Attention Network)를 제안한다. DSAN 은 디노이징 기법과 셀프 어텐션을 결합하여 견고한 특징 표현 벡터를 학습하고, 멀티 태스크 러닝 기법을 통해 결측된 다중 변수에 대한 대체값을 병렬적으로 예측한다. UCI 테이블 데이터 세트를 이용하여 결측값 보간 실험을 수행하여 제안 모델의 유효성을 입증한다.

Ⅰ. 서 론

대표적인 데이터 기반 의사결정 방법론의 기반 기술인 머신러닝은 기본적으로 완전한 고품질 데이터를 요구하지만, 실제 주어지는 입력데이터는 다양한 이유에 의해 결측값을 포함하게 된다. 결측값은 정보 손실을 야기하여 분석에 대한 신뢰성 하락과 모델 성능 하락의 주요 원인이 된다. 일반적으로 데이터 결측 문제를 해결하기 위해, 데이터 보간(Data Imputation) 기법을 활용하여 결측값을 적절한 대체값으로 채워 넣는다.

본 논문은 정형 데이터에 대해 직접적으로 활용 가능한 결측값 보간 모델 DSAN(Denoising Self-Attention Network)을 제안한다. DSAN 은 디노이징 기법과 셀프 어텐션을 결합하여 결측 입력 데이터에 대해 견고한 특징을 학습한다. 이후 멀티 태스크 러닝 기법을 통해 각 변수 별 대체값을 병렬적으로 예측한다.

Ⅱ. 관련 연구

2.1 결측값 보간 기법(Missing Value Imputation)

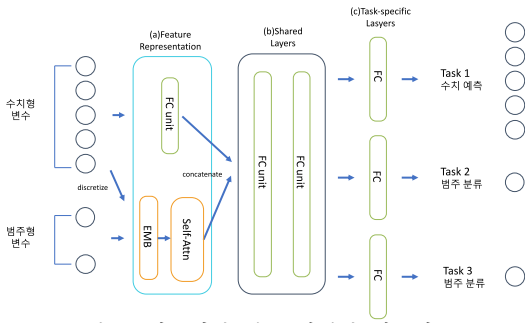
결측값 보간 기법은 적절한 대체값을 추정하여 결측값을 대체하는 기법을 말한다. 평균값이나 최빈값과 같은 기초 통계량 기반의 보간은 데이터가 크고 복잡해질수록 유효성이 떨어지며, 이를 해결하기 위해 머신러닝 기반의 결측값 보간 기법이 연구되고 있다. 머신러닝 기반 기법은 예측 모델 기반의 방법론과 생성 모델 기반의 방법론을 주로 사용한다.

2.2 정형 데이터를 위한 어텐션 메커니즘

최근 어텐션 메커니즘을 정형 데이터 도메인에 적용하기 위한 연구가 진행되고 있다. TabNet[1]은 순차적 어텐션 메커니즘을 활용하여 각 의사결정 단계에서 중요한 특징을 선택함으로써 해석 가능하고 효율적인 학습을 가능하도록 하였다. TabTransformer [2]는 셀프 어텐션을 기반으로 한 Transformer 의 인코더를 사용하였다. 셀프 어텐션을 통해 범주형 변수에 대한 문맥적 임베딩(Contextual Embedding)을 학습하여, 높은 예측 정확도를 달성했다.

Ⅲ. 디노이징 셀프 어텐션 네트워크

본 논문이 제안하는 디노이징 셀프 어텐션 네트워크(DSAN)의 구조는 [그림 1]과 같으며, 멀티 태스크 러닝 기법을 통해 각 변수 별 대체값을 예측한다. DSAN 은 역할에 따라 크게 3 가지 모듈인 특징 표현 모듈, 공유 레이어 모듈, 개별 태스크 레이어 모듈로 구성되어진다.



[그림 1] 디노이징 셀프 어텐션 네트워크 구조

특징 표현 모듈은 데이터를 입력 받아 대체값 예측을 위해 유용한 특징을 학습한다. 범주형 변수의 범주 값과 이산화한 수치형 변수 값에 대해 임베딩 후 셀프 어텐션을 통해 변수 간 연관성 정보를 포함하는 문맥적 임베딩(Contextual Embedding)을 학습한다. 또한 이산화하지 않은 수치형 변수 값을 FC(Fully Connected) 유닛을 통해 병렬적으로 특징을 학습하여 정보 손실을 보완한다. 병렬적으로 학습한 FC 유닛은 문맥적

임베딩과 결합(concatenate)하여 공유 레이어 모듈의 입력으로 활용된다. 이후 공유 레이어 모듈에선 각 변수 예측에 필요한 공유 파라미터를 학습하고, 개별 태스크 레이어 모듈에서 각 변수의 대체값 예측에 필요한 독립적인 파라미터를 학습한다. 수치형 변수의 경우 원래 값을 복원하도록 학습하고, 범주형 변수의 경우 원래 범주 값으로 분류하도록 학습하게 된다.

DSAN 은 학습과정에서 디노이징 기법[3]을 적용하여 일부 값을 제거한 입력 데이터를 받은 후 원래 값을 정답으로 맞추도록 학습한다. 디노이징 기법으로 결측 패턴에 대한 견고한 특징 표현을 학습하기 위해, DSAN 은 특정 변수의 결측 여부에 대한 임베딩 벡터를 학습한다. 이는 특정 변수의 결측 여부와 다른 변수의 관측값 간 연관성을 학습하게 하여, 모델의 예측 성능을 높일 수 있었다.

Ⅳ. 실험 및 결과

4.1 실험 데이터 및 방법

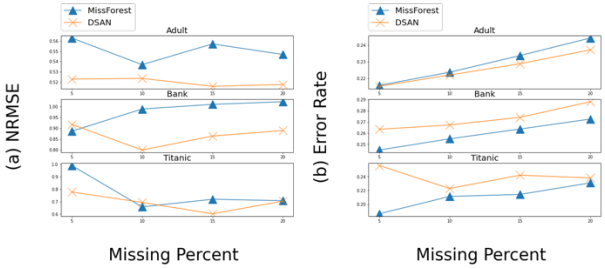
DSAN 의 유효성을 보이기 위해 UCI 정형 데이터 세트인 'Adult', 'Bank', 'Titanic' 데이터로 결측값 보간 실험을 수행한다. 실험은 세 단계로 이루어지며, 첫째 단계에선 완전한 데이터에 대해 MCAR(Missing Completely At Random)[4] 방식으로 5~20%의 결측값을 생성한다. 둘째 단계에선 결측값 보간을 수행하여 원래 값과의 오차를 측정한다. 이 때 수치형 변수의 경우 NRMSE 를, 범주형 변수의 경우 Error Rate 을 평가 척도로 삼는다. 셋째 단계에선 정제한 데이터를 이용해 이진 분류 모델을 학습하여 성능을 분석한다. 이 때 분류 모델의 성능 평가 지표로 AUC-ROC 를 활용한다.

[표 1] 실험 데이터 셋

데이터 셋	범주형 변수 개수	수치형 변수 개수	레코드 개수
Adult	9	6	30162
Bank	10	7	11162
Titanic	4	5	712

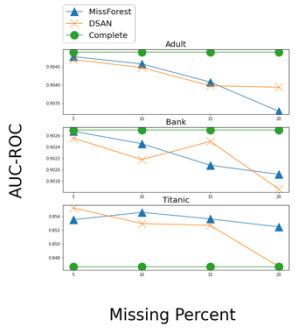
4.2 실험 결과

결측값 보간 모델 MissForest[5]를 비교 모델로 하여 실험 결과를 분석한다. [그림 2]의 (a)는 수치형 변수에 대한 오차를 의미하며 낮을수록 원래 값에 근사한 값으로 보간했음을 의미한다. DSAN 이 비교 모델 대비 평균 7% 향상된 성능을 보이는 것을 확인할 수 있다. [그림 2]의 (b)는 범주형 변수에 대한 오차를 의미하며, 제안 모델인 DSAN 이 비교 모델 대비 약 1% 낮은 성능을 보였다. 하지만 레코드 개수가 가장 큰 'Adult' 데이터 셋의 경우 DSAN 의 성능이 더 좋았으며, 이는 레코드 수가 클 수록 제안 모델이 우수한 성능을 보이는 것으로 해석할 수 있다.



[그림 2] 실험 결과(NRMSE, Error Rate)

[그림 3]의 경우 정제한 데이터를 학습한 이진 분류 모델의 성능을 나타낸다. 원래 데이터를 학습한 이진 분류 모델 성능 대비 평균 0.1% 미만의 차이를 갖으며, 이는 DSAN 이 정제한 데이터가 이 후 수행하는 작업에서 유의미한 결과를 낼 수 있음을 의미한다.



[그림 3] 실험 결과(AUC-ROC)

Ⅴ. 결론

본 논문은 DSAN 기반의 정형 데이터 결측값 보간 기법을 제안하였다. 수치형 변수 보간의 경우 비교 모델 대비 7% 향상된 성능을 보였으며, 범주형 변수의 경우 데이터 레코드 수가 커질수록 높은 성능을 보이는 것을 확인하였다. 또한 보간한 데이터를 학습한 이진 분류 모델이 원래 데이터 학습 대비 0.1% 미만의 차이를 보이며 유효성을 입증하였다.

ACKNOWLEDGMENT

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

참 고 문 헌

[1] S. O. Arik and T. Pfister, "TabNet: Attentive Interpretable Tabular Learning", arXiv preprint arXiv:1908.07442, 2019.

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

[3] P. Vincent, H. Larochelle, Y. Bengio and P.A. Manzagol, "Extracting and Composing Robust Features with Denoising Autoencoders", International Conference on Machine Learning, 2008.

[4] D. B. RUBIN, "Inference and missing data", Biometrika, Vol. 63, pp 581-592, 1976.

[5] DJ. Stekhoven and P. Buhlmann, "MissForest-non-parametric missing value imputation for mixed-type data", Journal of Machine Learning Research, Vol.28, No.1, pp 41-75, 1997