전자의무기록 데이터에서의 적대적 생성 알고리즘 기반 결측값 대치 알고리즘 성능분석

조용연, 정민영, 황보율*
*국립암센터 인공지능사업팀
e-mail:{yyjo0430, jmy, yulhwangbo}@ncc.re.kr

Performance Evaluation of an Imputation Method based on Generative Adversarial Networks for Electric Medical Record

Yong-Yeon Jo, Min-Yeong Jeong, Yul Hwangbo*
Healthcare AI Team, National Cancer Center

요 약

전자의무기록 (EMR)과 같은 의료 현장에서 수집되는 대용량의 데이터는 임상 해석적으로 잠재가치가 크고 활용도가 다양하나 결측값이 많아 희소성이 크다는 한계점이 있어 분석이 어렵다. 특히 EMR의 정보수집과정에서 발생하는 결측값은 무작위적이고 임의적이어서 분석 정확도를 낮추고 예측 모델의 성능을 저하시키는 주된 요인으로 작용하기 때문에, 결측치 대체는 필수불가결하다. 최근 통상적으로 활용되어지던 통계기반 알고리즘기반의 결측치 대체 알고리즘보다는 딥러닝 기술을 활용한 알고리즘들이 새로이 등장하고 있다. 본 논문에서는 Generative Adversarial Network를 기반한 최신 결측값 대치 알고리즘인 Generative Adversarial Imputation Nets을 적용하여 EMR에서의 성능을 분석해보고자 하였다.

1. 서론

대용량의 데이터를 기반으로 개인 맞춤의학을 실현하고자하는 미래의학의 핵심 자원으로서 임상 현장에서 실시간으로 수집되는 다양한 종류의 의료 데이터들이 주목받고 있다. 전자의무기록 (Electric medical record, EMR)은 대표적인 의료 빅데이터 중 하나로서 질병의 원인을 파악하고 예후를 예측하는데 필요한 정보인 질병력, 약물력, 가족력 등의 기본 역학정보에서부터 맥박, 호흡수, 체온과 같은 활력징후 검사 정보에 이르기까지 다수의 유의미한 정보를 포함하고 있어 데이터에 내재된 해석적 가치가 매우 높다.

이러한 이점에도 불구하고 EMR은 시간의 흐름에 따라 수작업으로 기록되어지는 수집 과정의 특성상 환자의 사망, 장비 오작동, 수신자 응답 거부 등으로 발생되는 다량의 무작위적인 결측값이 존재할 수 있어 희소성이 크기 때문에 분석이 어렵다는 한계점이 있었다. [1] 빅데이터 분석에 활용되는 대부분의 알고리즘들은 데이터의 완결성을 기본으로 가정하기 때문에 EMR에서의 결측값 문제는 필수적으로 선결되어야할 문제라고 할 수 있다. [2]

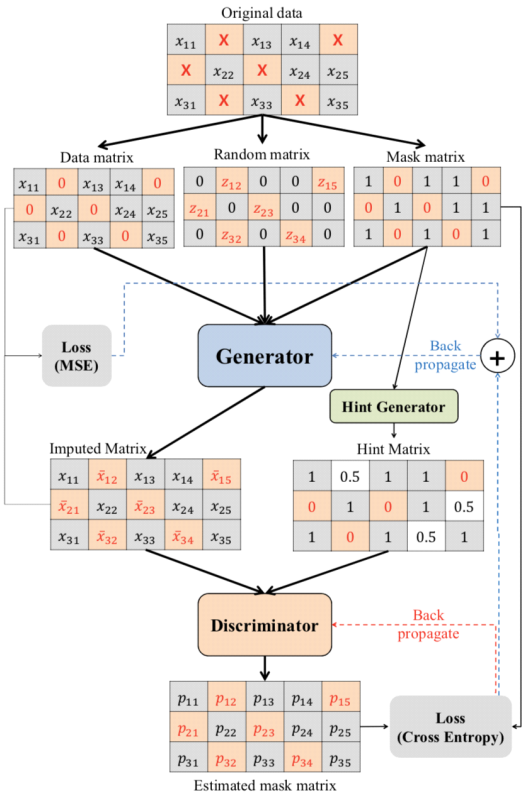
빅데이터에서 결측값을 다루는 전통적인 방법으로는 Naive Bayesian Classifier, K-Nearest Neighbor와 같이 알고리즘 내부에서 결측치를 대체하는 모델과 결측값 단순삭제, 평균값, 중앙값 대체, 임의의 난수값으로 대체하는 등 알고리즘 외부에서 독립적으로 결측치를 대체하는 방법이 쓰여왔다. [3] 최근 딥러닝 기술의 급속한 발전으로 Generative Adversarial Net (GAN) [4]과 같은 알고리즘이 이미지 생성 분야와 문장생성 등 다분야에서 기존 머신러닝의 성능을 뛰어넘는 결과를 보였다.

따라서 본 논문에서는 결측치 대체 문제를 해결하기 위해 GAN의 대표적인 결측값 대치 알고리즘인 Generative Adversarial Imputation Nets (GAIN) [5]을 사용하고, 이를 활용하여 EMR에서의 보다 효율적으로 학습할 수 있는 방법을 살펴본다. 실험결과, GAIN에 입력으로 사용되는 Random matrix의 초기값 보정을 통해 사망 예측에 대한 AUROC (Area Under the Receiver Operating Characteristic 와 Area Under the Precision Recall Curve (AUPRC)를 각각 0.7%와 0.9% 향상시켰다.

2. Generative Adversarial Imputation Nets (GAIN)

GAIN은 적대적 생성 알고리즘인 GAN을 골자로 결측값을 보간하는 딥러닝 모델이다. GAIN은 크게 Imputed matrix를 생성하는 Generator와 Generator가 생성된 값을 얼마나 잘 맞추는지 구분해내는 Discriminator 네트워크로 이루어져 있다. 원본 데이터에서 데이터만 추출한 Data matrix, 결측치 자리에 채울 난수값이 들어있는 Random matrix와 결측값 유무의 위치를 표시한 Mask matrix 총 세 개의 matrix를 Generator가 입력받으면 결측치가 채워진 Imputed matrix가 생성되며, Discriminator는 Mask matrix를 기반한 Hint Hint matrix와 앞단계에서 Generator가 만든 Imputed matrix를 입력받아 Estimated mask matrix를 만들어 그 차이에 대해 손실값을 Backpropagation하여 모델을 학습시키는 구조로 이

루어져있다.



<그림 1> GAIN의 기본 구조

3. 실험

3.1. 실험환경

실험은 Intel Core i7 2.6 GHz, 16GB의 MacBook Pro에서 수행되었다. 실험에 사용된 데이터는 The PhysioNet Computing in Cardiology Challenge 2012 (PhysioNet 2012)[1]을 사용하였다. 해당 데이터는 ICU 병상에서 12,000명의 환자에 대해서 입원 후 첫 48시간내에 측정된 레코드들로 구성되어 있다. 각 레코드는 최대 42개의 변수를 가지고 있다. 변수의 종류는 환자의 ID, 나이, 성별, 키, 몸무게, ICUtype이 고정변수로 존재하며, 37개의 time series 변수가 존재한다.

3.2 실험방법

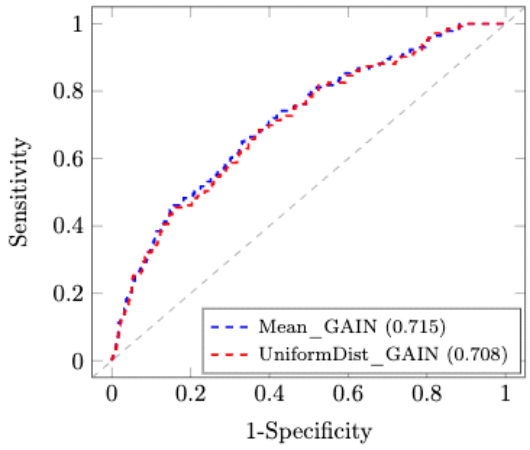
EMR 데이터의 많은 변수들은 연속변수이다. 기본 GAIN은 Random matrix에 연속균등분포를 가정하여 완전하게 임의적으로 산출된 수치를 입력한다 (UniformDist_init).

반면, EMR의 연속변수들은 정규분포를 따른다는 점에 착안하여, 우리는 Random matrix에 각 변수들의 정규분포상의 평균값을 입력한다 (Mean_init). 이를 통해 학습의 시작 시점에 실제 수치와 유사할 확률이 높은 값으로 시작함으로써 좀 더 빠른 학습속도와 실제 값에 더 가까운 값을 생성해낼 수 있도록 한다.

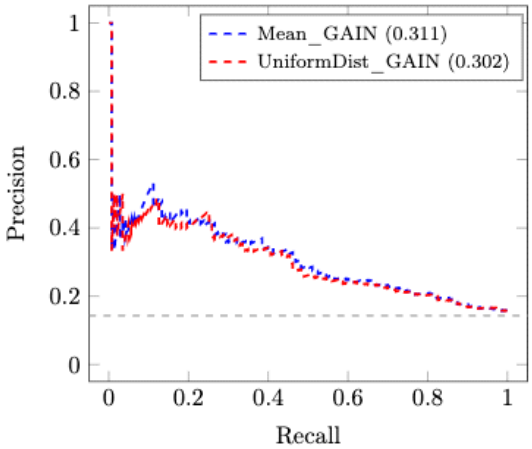
1) https://physionet.org/content/challenge-2012/1.0.0/

3.4. 실험결과

실험결과는 그림 2와 같다. 연속균등분포를 기반으로 초기 입력을 사용한 GAIN (UniformDist_GAIN) 모델에서 중환자실 환자의 사망률 예측에 대한 AUROC는 0.708로 측정되었다. 반면, EMR 데이터의 변수 특성을 고려하여 정규분포상의 평균값을 초기값으로 반영한 GAIN (Mean_GAIN)은 AUROC가 0.715로 측정되어 약 0.7%의 향상됨을 보였다 (그림 2-a). 또한 AUPRC도 Mean_GAIN이 Uniform_GAIN보다 0.9% 향상되었다 (그림 2-b). 실험결과 초기값이 예측 성능에 영향을 미칠 수 있음을 확인하였다.



(a) ROC curve



(b) PR curve

<그림 2> 결측치가 대체된 PhysioNet 2012에 대한 사망 예측 결과

4. 결론

본 논문에서는 GAN을 기반한 결측치 대체 알고리즘인 GAIN 모델을 이용하여 EMR의 결측치를 효율적으로 대체할 수 있는 방법을 적용하였다. 실험 결과, PyshioNet 2012에서 사망예측에 대한 성능이 향상됨을 확인하였다. 추후 연구로, 더 다양한 데이터와 다양한 generative 대체 알고리즘에 적용하여 초기값에 따른 성능의 변화를 살펴보고자 한다.

참고문헌

[1] Barnard, J. and Meng, X.-L. Applications of multiple imputation in medical studies: from aids to nhanes. Stat Methods Med Res, 1999.

[2] Kreindler, D. M. and Lumsden, C. J. The effects of the irregular sample and missing data in time series analysis. Nonlinear Dynamical Systems Analysis for the Behavioral Sciences Using Real Data, 2012.

[3] Purwar, A. and Singh, S. K. Hybrid prediction model with missing value imputation for medical data. Expert Systems with Applications, 2015.

[4] Goodfellow, I. J., et. al. Generative adversarial nets. NIPS, 2014.

[5] J. Yoon, J. Jordon, M. van der Schaar, GAIN: Missing Data Imputation using Generative Adversarial Nets. ICML, 2018.