의료 인공지능 성능 향상을 위한 GAN 기반 희소 질병 데이터 합성

정예림[1], 김소연[2], 이일구[3]
[1]성신여자대학교 융합보안공학과 학부생
[2]성신여자대학교 미래융합기술공학과 박사과정
[3]성신여자대학교 융합보안공학과, 미래융합기술공학과 교수
[1,2,3]{20221130, 220237014, iglee}@sungshin.ac.kr

GAN-Based Synthesis of Sparse Disease Data for Improving Medical AI Performance

Ye-Rim Jeong[1], So-Yeon Kim[2], Il-Gu Lee[1,2]
[1]Dept. of Convergence Security Engineering, Sungshin Women's University
[2]Dept. of Future Convergence Technology Engineering, Sungshin Women's University

요 약

최근 디지털 헬스케어 기술과 서비스가 널리 활용되면서 의료 인공지능 성능 향상에 대한 관심이 높아지고 있다. 그러나 양성 데이터 대비 질병 데이터가 희소하여 학습 과정에서 과적합이 발생하거나 질병 예측 모델의 성능이 떨어진다는 한계가 있다. 본 논문에서는 데이터가 균질하지 않은 상황에서 생성형 인공지능 모델을 사용하여 합성 데이터를 생성하는 방안을 제안한다. 실험 결과에 따르면, 종래 방법 대비 제안한 방법의 정확도가 약 5.8% 향상되었고, 재현율이 약 21% 개선되었다.

1. 서론

최근 의료 산업이 디지털화되고 비대면 원격 진료가 보편화되면서 인공지능 기술을 활용한 디지털 헬스 의료 서비스가 널리 활용되고 있다. 그러나 희귀 질환은 진단과 치료를 위한 데이터가 충분하지 않고, 민감정보의 재식별 문제 때문에 기술 활용에 제약이 있다. 학습 데이터가 부족하면 과적합 문제가 발생하고, 데이터의 균질성이 부족하면 모델의 성능이 저하된다[1]. 희소 데이터의 비균질성 문제를 해결하기 위한 종래의 SMOTE(Synthetic Minority Over-sampling Technique) 기술은 희소 데이터의 최근접 이웃을 이용하여 데이터를 생성한다. 그러나 이상치 샘플을 사용하여 생성한 데이터도 이상치가 될 수 있다는 한계가 있다. 이러한 문제를 해결하기 위해 최근에는 개인정보 재식별 가능성을 최소화하면서도 원본 데이터의 특성을 충분히 반영하는 합성 데이터 생성 기술에 대한 논의가 진행되고 있다. 본 논문에서는 GAN(Generative Adversarial Network)의 적응적 학습을 통해 합성 데이터를 생성하는 특징을 활용하여 희소 질병 데이터를 증강하여 의료 인공지능 성능을 향상하는 방식을 제안한다.

본 논문의 기여점은 다음과 같다.

• GAN 기반으로 희소 질병 데이터를 합성하여 의료 인공지능 성능을 향상시키는 방법을 제안한다.

• 희소 데이터 증강 방법의 효율성을 평가하는 프레임워크를 제안한다.

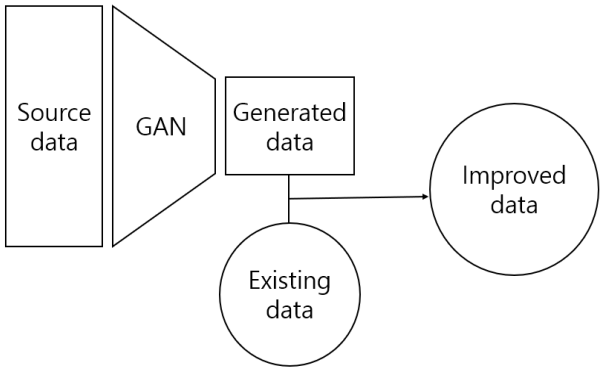
본 논문은 다음과 같이 구성된다. 2장에서는 제안 모델을 설명하고, 3장에서는 실험 결과를 분석하고, 4장에서 결론을 맺는다.

2. GAN 기반 희소 질병 데이터 합성

디지털 헬스 의료 분야에서 인공지능을 활용한 질병 탐지를 할 때, 희소 데이터로 학습을 진행하면 데이터 균질성의 부족으로 인한 오탐율이 증가할 가능성이 있다. 따라서 그림 1과 같이 희소 데이터를 GAN 모델로 학습시켜 생성한 합성 데이터로 모델의 성능을 개선하는 방법을 제안한다.

소스 데이터 (source data)는 원본 데이터를 의미하며, 생성 데이터 (generated data)는 생성형 모델을 통해 만들어진 합성데이터이다. 기존 데이터 (existing data)는 기존에 존재하는 데이터를 의미하며, 이는 소스 데이터와 동일하다. 개선 데이터 (improved data)는 생성 데이터와 기존 데이터를 합

쳐서 균질성을 개선한 데이터이다.

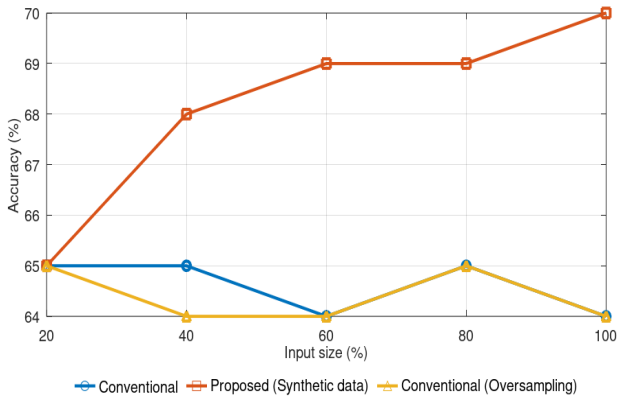


(그림 1) GAN 기반의 합성 데이터 생성 모델 구조도

3. 성능 평가

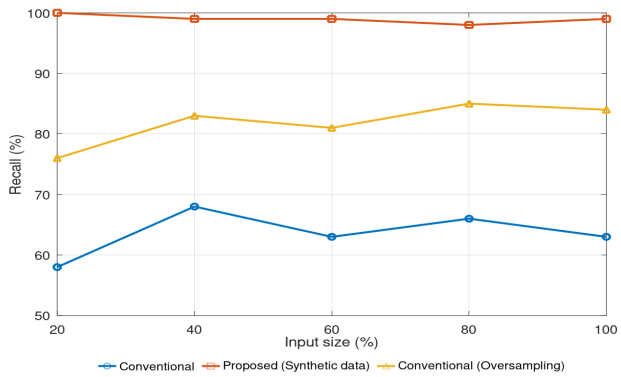
본 연구에서는 심장마비 예측 데이터 세트를 활용하였다[2]. 이는 26개의 특징과 8,764개의 데이터로 구성되며, 양성 데이터와 음성 데이터가 약 1:1.8의 비율로 존재한다. 실험은 원본 데이터 세트와 각각 오버샘플링 기법과 GAN 기반으로 증폭한 데이터 세트로 심장 위험율 (heart risk rate)를 예측하는 방법으로 진행했다. 제안 방식과 오버샘플링 기법 기반 데이터 세트는 희소 데이터를 증폭하여 1:1의 비율을 이루도록 조정하였다. 제안하는 방식의 평가는 실험은 100번 진행하여 나온 값의 평균으로 계산하여 정확도와 재현율을 비교하였다. x축은 데이터 세트의 크기를 의미하며, y축은 각각 정확도와 재현율을 의미한다.

그림 2는 데이터 세트에 따른 정확도를 나타낸다. 정확도는 모델이 올바르게 분류한 데이터의 비율을 나타내는 지표이다. 그림 3은 데이터 세트에 따른 재현율을 나타낸다. 재현율은 양성 데이터 중 모델이 올바르게 양성으로 예측한 비율을 나타낸다.



(그림 2) 모델별 정확도

그림 2와 그림 3에서 알 수 있듯이, 제안 모델은 오버샘플링 기법에 비하여 정확도가 약 5.8% 개선되고, 재현율이 약 21% 개선된 것을 확인할 수 있다. 또한 입력 데이터 크기와 상관 없이 제안 모델은 개선된 성능을 보인다. 따라서 희소 데이터를 증폭할 때, 오버샘플링 기법보다 생성형 모델이 효율적임을 실험적으로 확인했다.



(그림 3) 모델별 재현율

4. 결론

디지털 헬스케어 시장이 커지면서 인공 지능 기반의 의료 데이터 학습 정확도 향상하는 기술이 중요해지고 있다. 질병 데이터가 희소하면, 질병을 예측하는 인공지능 모델의 학습 성능이 떨어진다는 한계가 있다. 본 논문에서는 GAN 기반으로 희소 질병 데이터를 합성하여 의료 인공지능의 성능을 향상시키는 방법은 제안했다. 실험 결과에 따르면, 데이터 증강을 위해 많이 사용되는 오버샘플링 기법에 비해 정밀도와 재현율이 개선되었다. 후속 연구로 합성 데이터 생성 과정에서 민감 정보를 보호하는 기술을 연구할 계획이다.

ACKNOWLEDGEMENT

본 논문은 2024년도 산업통상자원부 및 한국산업기술진흥원의 산업혁신인재성장지원사업 (RS-2024-00415520)과 과학기술정보통신부 및 정보통신기획평가원의 ICT혁신인재4.0 사업의 연구결과로 수행되었음 (No. IITP-2022-RS-2022-00156310)

참고문헌

[1] Hao, Weituo, et al. "Towards fair federated learning with zero-shot data augmentation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.

[2] Heart Attack Risk Prediction Dataset. Sourav Banerjee. https://www.kaggle.com/datasets/iamsouravbanerjee/heart-attack-prediction-dataset