CTGAN기반 데이터 증강 비율 최적화 연구

성다훈[1], 임유진[2]
[1]숙명여자대학교 IT공학과 석사과정
[2]숙명여자대학교 인공지능공학부 교수
ekgns324@sookmyung.ac.kr, yujin91@sookmyung.ac.kr

A Study on the Optimization of Data Augmentation Ratio using CTGAN

Da-Hun Seong[1], Yujin Lim[2]
[1]Dept. of Information Technology Engineering, Sookmyung Women's University
[2]Div. of Artificial Intelligence Engineering, Sookmyung Women's University

요 약

머신러닝과 딥러닝 모델의 사용이 급증함에 따라 충분한 데이터 확보의 중요성이 부각되고 있다. 이에 따라 생성 모델을 통한 데이터 증강 기술이 주목받고 있으나, 증강 데이터를 활용했을 때 학습의 성능 분석은 아직 부족하다. 따라서 본 연구에서는 데이터 증강 시나리오에 따라 증강 비율별 합성 데이터의 유용성을 조사하고자 한다. 본 연구에서는 테이블 데이터를 증강하는 것에 초점을 맞추었으며, 이를 위해 테이블 데이터를 합성할 때 유용한 성능을 보이는 딥러닝 모델 CTGAN을 활용하였다. 실험에서 데이터를 증강하는 두 가지 다른 시나리오를 고려한 결과, 두 시나리오에서 모두 실험에서 설정한 증강 비율까지의 합성 데이터가 유용한 결과를 보임을 확인할 수 있었다.

1. 서론

최근 IT 분야에서 머신러닝과 딥러닝의 인공지능 모델의 활용은 빠르게 확장되어왔다. 그러나 이러한 인공지능 모델의 성능을 극대화하려면 충분한 양의 데이터가 필수적이다. 이는 모델의 일반화 능력과 예측 정확도를 향상시키는 데 중요한 역할을 한다. 그러나 현실에서는 모델을 학습시킬 충분한 양의 데이터를 확보하기 어려운 경우가 많다. 이러한 한계를 극복하기 위해 생성 모델을 통한 데이터 증강 기술이 많은 연구와 관심을 받고 있다. 데이터 증강은 기존 데이터를 활용하여 새로운 데이터를 합성하는 방법이다. 특히 테이블 데이터를 증강하는 모델 CTGAN(Conditional Table GAN)[1]은 데이터 분포를 고려하여 실제 데이터와 비슷한 가상 데이터를 생성하는 데 효과적으로 활용되고 있다. 그러나 데이터를 합성할 때 원본 데이터의 특성을 해치지 않기 위해서는 합성 데이터를 어느 정도 증강해야 하는지, 원본 데이터와 합성 데이터의 비율을 어느 정도로 설정해야 하는지 파악하기 어렵다. 따라서 본 연구에서는 CTGAN을 활용하여 데이터를 증강할 때, 데이터 증강 시나리오를 전체 데이터셋 사이즈가 작을 때 증강시키는 경우와 전체 데이터셋에서 중요도가 높으나 희소한 이상 데이터를 증강하는 경우로 나누어 실제 데이터와 합성 데이터의 비율에 따른 성능 변화를 파악하고자 한다. 이때 데이터 증강의 결과로 생성된 합성 데이터는 실제 데이터의 특성을 유지해야 한다. 이를 평가하기 위해 실제 데이터와 합성 데이터를 각각 4가지 머신러닝 모델에 적용하여 나오는 F1 점수와 자카드 유사도(Jaccard Similarity)를 활용하여 분석할 것이다.

2. 생성적 적대 신경망(GAN, Generative Adversarial Network)

데이터 증강 모델은 증강할 데이터의 종류(음성, 이미지, 시계열, 테이블 데이터(tabular data) 등)에 따라 다양하며 본 연구에서는 테이블 데이터 증강 모델에 집중할 것이다. 테이블 데이터를 증강하는 모델은 크게 베이지안 네트워크(Bayesian Network)를 활용한 방법(CLBN[2], PrivBN[3] 등)과 딥러닝 모델인 생성적 적대 생성망(GAN)[4]을 활용한 방법(MedGAN[5], VeeGAN[6], TableGAN[7], CTGAN 등)이 있다. 그러나 베이지안 네트워크기반 방법은 고차원 데이터와 같은 복잡하고 비선형적인 데이터 패턴을 모델링하기 어렵다는 한계가 있어, 최근에는 GAN을 활용한 딥러닝 기반의 데이터 증강 모델 연

구가 활성화되고 있다.

GAN은 생성자(Generator)와 판별자(Discriminator)의 구조를 가져 다양한 유형의 데이터를 생성할 수 있다. 생성자는 랜덤한 노이즈나 입력 데이터를 받아 실제 데이터와 유사한 데이터를 생성한다. 이렇게 생성된 데이터는 처음에는 랜덤하고 무질서하지만 훈련 과정을 통해 실제 데이터와 더욱 유사한 패턴을 학습하게 된다. 판별자는 생성자에서 만들어진 데이터와 실제 데이터를 구분하려고 노력하며 입력된 데이터가 실제인지 생성된 데이터인지 판별하는 역할을 한다. GAN의 핵심 아이디어는 생성자와 판별자가 서로 경쟁하며 학습하는 것이다. 생성자는 더욱 실제와 유사한 데이터를 생성하려고 노력하고, 판별자는 생성자가 만든 합성 데이터와 실제 데이터를 구분하려고 노력한다. 이 경쟁 과정에서 생성자는 점차적으로 더 나은 데이터를 생성하게 된다. 본 연구에서는 GAN을 활용한 테이블 데이터 증강 모델 중에서도 데이터 유용성 측면에서 가장 높은 성능을 보이는 모델인 CTGAN을 활용하고자 한다.

3. CTGAN 모델

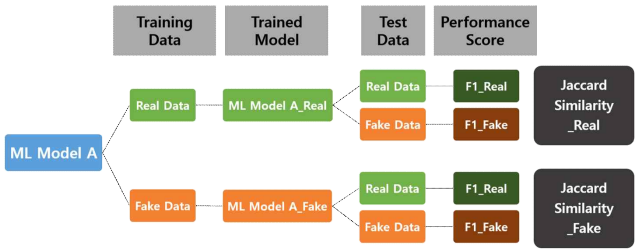
조건부 적대적 테이블 생성망(Conditional Tabular GAN, CTGAN)은 GAN을 기반으로 하는 딥러닝 모델로, 연속과 불연속 데이터가 혼재되어있는 테이블 데이터에서 데이터를 생성하는 테이블 데이터 증강 모델이다. 일반적인 적대적 생성망은 다수의 클래스 데이터 위주로 학습되기 때문에 그 희소한 값을 잘 학습하지 못해, 데이터 재현 시 다수의 클래스만 재현하는 문제가 발생하게 된다. 그러나 조건부 적대적 생성망은 이러한 문제에 대응할 수 있다. 조건부 적대적 생성망을 통해 희소한 클래스를 생성하도록 범주형 열을 조건으로 넣어주어, 생성자가 학습 과정에서도 조건으로 희소한 범주형 속성값에 노출되도록 조절 가능하기 때문이다.

데이터 증강 모델을 사용하여 합성 데이터를 만들어 데이터를 증강하는 것과 희소 데이터를 단순히 대량으로 복제하여 수를 늘리는 것(오버샘플링)은 차이가 있다. 오버샘플링은 기존 데이터의 특성만 반영하여 과적합 모델로 이어질 수 있다는 한계가 있다. 그러나 CTGAN 데이터 증강 모델은 이웃 정의(Defining Neighborhoods)를 통해 희소한 데이터 사이의 일반적인 공통점을 식별하여 실제자료와 유사한 분포를 가진 맥락있는 합성 데이터를 생성할 수 있는 장점이 있다.

4. 실험 설계

본 연구에서는 CTGAN 모델을 사용하여 테이블 데이터를 증강할 때, 데이터 증강 시나리오별 데이터 증강 비율에 따른 합성 데이터의 유용성을 분석하고자 한다. 실험을 위하여 Kaggle의 credit 데이터셋[8]을 사용하였다. 이는 신용 카드 거래 내역을 나타내며, 대부분의 거래가 정상 거래이나 상대적으로 희소한 사기 거래 내역이 존재한다.

실험은 테이블 데이터의 증강 시나리오에 따라 구분하였다. 먼저 시나리오 1인 원본 데이터의 사이즈가 작은 경우에는, 합성 데이터를 원본 데이터의 0.5배, 1배, 2배, 3배, 5배로 각각 증강하여, 원본 데이터와 합성 데이터의 5가지의 비율(1:0.5, 1:1, 1:2, 1:3, 1:5)을 구성하였다. 다음으로 원본 데이터에서 이상 데이터만 증강하는 시나리오 2에서는, 정상 데이터와 이상 데이터의 비율을 기존의 1:0.001에서 다음과 같도록(1:0.005, 1:0.01, 1:0.05, 1:0.1) 이상 데이터를 각각 5배, 10배, 50배, 100배 증강하였다.



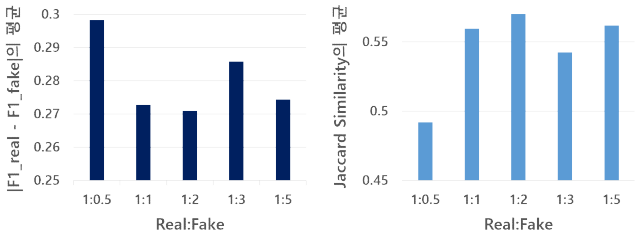
(그림 1) 머신러닝 모델 학습 및 테스트 구조

이렇게 구성한 데이터의 유사성을 평가하기 위해서는 먼저 머신러닝 모델별로 F1 점수와 자카드 유사도를 구했다. 머신러닝 모델은 Decision Tree, LogisticRegression, MLPClassifier, RandomForestClassifier 4가지를 활용했다. (그림 1)에서 보는 바와 같이, 우선 실제 데이터와 합성 데이터를 각각 학습 데이터와 테스트 데이터로 나누고 실제 데이터와 합성 데이터의 각 학습 데이터를 4개의 머신러닝 모델에 개별적으로 학습시켰다. 이후 학습된 모델을 실제 데이터와 합성 데이터의 각 테스트 데이터에 적용하여 각각의 F1_Real과 F1_Fake 점수, 그리고 자카드 유사도를 구했다. 최종적으로 합성 데이터의 증강 비율별 성능 평가는, 한 시나리오 내에서 합성 데이터의 증강 비율별 4가지 머신러닝 모델의 |F1_Real-F1_Fake| 점수의 평균과 자카

드 유사도의 평균을 비교하였다. 실제 데이터와 합성 데이터의 유사성은 |F1_Real-F1_Fake| 점수의 경우 작을수록 높은 것이며, 자카드 유사도의 경우에는 점수가 클수록 높다.

5. 실험 결과

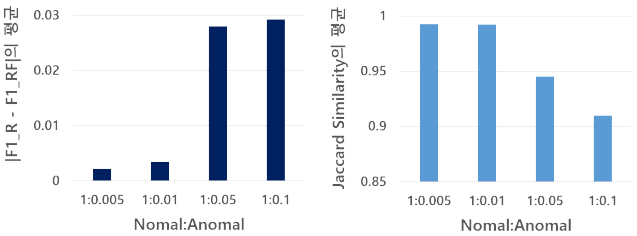
테이블 데이터의 증강 시나리오에 따른 실험 결과는 다음과 같다. 먼저 시나리오 1에서 전체 데이터를 늘리는 경우, 실제 데이터와 합성 데이터의 비율별로 |F1_Real-F1_Fake|의 점수와 자카드 유사도의 평균 점수를 비교하면 (그림 2)와 같다.



(그림 2) 시나리오 1. 전체 데이터 증강

증강 비율별 |F1_Real-F1_Fake|의 평균 점수를 보면, 합성 데이터를 원본 데이터의 2배로 증가시켰을 때 그 차이가 가장 작아 합성 데이터와 원본 데이터의 유사도가 가장 높고, 0.5배로 증가시켰을 때는 가장 낮은 유사도를 보였다. 자카드 유사도의 평균 점수의 경우에도 마찬가지로 합성 데이터를 원본 데이터의 2배로 늘렸을 때 유사도가 가장 높게 나왔으며, 0.5배로 늘렸을 때 가장 낮게 도출되었다. 그러나 |F1_Real-F1_Fake|의 경우 최고와 최저 성능의 차이가 0.03 미만이며, 자카드 유사도의 경우 0.01 미만으로 그 차이가 근소하여 본 실험의 시나리오 1에서는 합성 데이터를 실제 데이터의 5배까지 증강한다 하더라도 데이터의 특성에 큰 변화를 발생시키지 않는 것을 확인할 수 있다.

다음으로 시나리오 2에서 이상 데이터를 증강했을 때의 실험 결과는 (그림 3)과 같다.



(그림 3) 시나리오 2. 이상 데이터 증강

이 경우 원본 이상 데이터가 충분하지 않아 두 데이터 간 유사성을 평가하기 어려웠다. 따라서 실험에서 설정한 정상 데이터와 이상 데이터의 비율별로 합성한 각 이상 데이터 F를 전체 데이터 R에 더하여 RF 데이터를 만들었으며, R과 RF의 유사성을 비교하였다. |F1_R - F1_RF|의 평균 점수를 보면 이상 데이터를 기존의 5배로 합성하여 전체 데이터와의 변화가 가장 적은 1:0.005의 경우가 가장 유사도가 높고, 이상 데이터를 100배 늘린 1:0.01의 경우가 유사도가 가장 낮게 나왔으나 그 차이는 근소하다. 자카드 유사도의 경우에도 비슷한 결과가 도출되었다. 그리고 |F1_R - F1_RF|의 평균과 자카드 유사도의 평균 점수 모두 전반적으로 이상 데이터의 합성 배율을 최대 100배 가까이 증강시킨다 하더라도 데이터의 특성 변화에 유의미한 영향을 미치지 않는 것을 확인하였다.

<표 1> 시나리오 2에서의 비율별 |F1_RF-F1_R| 점수

		F1_RF - F1_R			
		1:0.005	1:0.01	1:0.05	1:0.1
Real	DecisionTree Classifier	0.0007	0.0006	0.0006	0.0004
	Logistic Regression	0.00009	0	-0.0005	-0.0007
	MLP Classifier	0.0076	0.0003	0.0008	0.0007
	RandomForest Classifier	0.0003	0.0003	0.0003	0.0002

<표 1>은 시나리오 2에서의 비율별 |F1_RF-F1_R| 점수이다. 시나리오 2에서는 원본 데이터와 이상 데이터를 증강하여 원본 데이터에 더했을 때의 데이터를 비교했으므로 두 데이터 간 모델 성능을 각 F1 점수를 통해 비교해볼 수 있다. 실제 데이터로 학습한 Logistic Regression에서 정상 데이터와 이상 데이터의 비율을 1:0.05와 1:0.1로 증강하였을 때를 제외하고는 모두 머신러닝 모델에 적용했을 때 증강한 데이터를 더한 쪽의 성능이 향상됨을 확인할 수 있다.

종합하자면 두 시나리오에서 모두 설정한 합성 데이터의 증강 비율별 차이가 크지 않고, 전반적으로 유사한 품질의 데이터를 합성할 수 있음을 확인할 수 있었다. 이렇게 데이터를 증강하게 되었을 때 두 가지 데이터 증강 시나리오에 따른 장단점은 다음과 같다. 먼저 전체 데이터를 늘린 경우, 장점은

합성 데이터를 통해 데이터의 다양성이 증가하게 되고, 이를 학습 모델에 적용하면 적은 데이터를 사용했을 때보다 일반화 능력이 향상되어 모델의 성능이 향상될 수 있다. 그러나 단점으로는 많은 데이터를 생성할수록 증강 데이터를 생산할 때와 증강된 데이터를 모델에 적용할 때 계산 및 저장 오버헤드가 증가한다는 것이며, 특히 테이블 데이터의 경우 열의 개수가 많을수록 메모리를 많이 차지해 학습에 부담이 될 수 있다. 다음으로 이상 데이터를 증강했을 때의 장점은 모델이 이상 데이터에 대한 가중치를 더하여, 보다 정확한 패턴을 학습하게 되고 이를 학습 모델에 적용 시 이상치 탐지 성능을 개선할 수 있다는 것이다. 그러나 단점으로는 이상 데이터의 증강 비율을 늘릴수록 원본 데이터에서의 정상 데이터와 이상 데이터의 비율이 달라져 이와 관련된 모델의 성능이 저하될 수 있다.

6. 결론

본 연구에서는 데이터를 증강하여 학습 데이터 부족 문제를 해결할 때, 데이터 증강 시나리오에 따른 성능을 분석하였다. 본 실험을 통해 시나리오 1에서는 5배, 시나리오 2에서는 100배 증강까지 유의미한 성능을 확인할 수 있었다. 다만 이는 데이터의 특성에 따라서 달라질 수 있으므로 데이터를 증강할 때 실험을 통해 증강한 데이터와 합성한 데이터의 유사성을 확인한 후 사용하는 것이 중요하다. 그럼에도 본 논문에서는 |F1_Real-F1_Fake| 비율 점수에 대한 비교를 통해서 증강 데이터 비율 증가에 대한 효과를 검증하고 유의미한 결론을 도출하였다.

데이터를 인위적으로 늘리는 것보다 실제 많은 양의 데이터를 활용하는 것이 인공지능 모델의 정확도가 더 높을 것이나, 본 논문에서는 데이터 증강의 다양한 시나리오에 대한 연구 결과를 제시함으로써 실제 환경에서의 응용 가능성에 대한 통찰력을 제공할 것으로 기대한다. 본 실험은 credit 데이터셋을 활용하여 실험해서 얻은 결과이므로, 향후 다양한 테이블 데이터셋을 기반으로 성능 분석 연구를 진행하고, 딥러닝 모델에도 적용하여 데이터의 유사성 평가를 분석할 예정이다. 또한 모델 학습의 예측력을 높이기 위한 방법으로 생성 모델을 활용하는 것 외에도 준지도 학습에서 가상의 레이블을 자동으로 생성하여 지도하는 학습 기술인 의사 레이블(pseudo label)을 이용하는 방안까지 확장하여 고려할 것이다.

사사문구

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2021R1F1A1047113).

참고문헌

[1] L. Xu, M. Skoularidou, A. C-I and K. Veeramachaneni, “Modeling Tabular Data Using Conditional GAN,” Proceedings of Advances in Neural Information Processing Systems (NIPS), Dec. 8-14, Vancouver, 2019.

[2] C. Chow and C. Liu, “Approximating Discrete Probability Distributions with Dependence Trees,” Proceedings of IEEE Transactions on Information Theory, 14, 3, pp. 462-467, 1968.

[3] J. Zhang, G. Cormode, C.-M. Procopiuc, D. Srivastava and X. Xiao, “Privbayes: Private Data Release via Bayesian Networks,” Proceedings of ACM Transactions on Database Systems, 42, 4, pp. 1-41, 2017.

[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, W.-F. D, S. Ozair and Y. Bengio, “Generative Adversarial Nets,” Proceedings of Advances in Neural Information Processing Systems (NIPS), Dec. 08-14, Montreal, 2014.

[5] E. Choi, S. Biswal, B. Malin, J. Duke, W.-F. Stewart and J. Sun, “Generating Multi-label Discrete Patient Records Using Generative Adversarial Networks.” Proceedings of Machine Learning for Healthcare Conference (PMLR), Aug. 18-19, Boston, 2017.

[6] A. Srivastava, L. Valkov, C. Russell, M.-U. Gutmann and C. Sutton, “Veegan: Reducing Mode Collapse in Gans Using Implicit Variational Learning,” Proceedings of Advances In Neural Information Processing Systems (NIPS), Dec. 4-9, California, 2017.

[7] N. Park, M. Mohammadi, K. Gorde, S. Jajodia, H. Park and Y. Kim, "Data Synthesis based on Generative Adversarial Networks," Proceedings of ArXiv, Preprint ArXiv:1806.03384, 2018.

[8] Kaggle의 credit 데이터셋, https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud