금융 합성데이터 생성 및 후처리를 통한 품질 개선

이동현[1], 홍승혁[1], 김호정[1], 김세영[1]
[1]인천대학교 컴퓨터공학부 학부생
bt12333r@gmail.com, sdbrandon0408@gmail.com, 2005kimo@inu.ac.kr,
sseyeong42@naver.com

Improving the Quality of Financial Synthetic Data through Generation and Post-processing

Dong-Hyeon Lee[1], Seung-Hyuk Hong[1], Ho-Jung Kim[1], Se-Young Kim[1]
[1]Dept. of Computer Science Engineering, Incheon National University

요 약

금융 데이터의 민감성으로 인해 합성 데이터가 주목받고 있으며, 본 연구는 유사성이 낮은 열에서 0 근처 값을 0으로 치환하는 SVS(Single-Value Substitution) 후처리 기법을 제안한다. 기존 방식과 비교해 SVS와 DSF(Dynamic Sample Filtering)를 결합하여 합성 데이터의 품질을 크게 향상시켰다. 특히, DSF와 SVS를 함께 적용했을 때 Column Shapes와 Overall 성능에서 가장 높은 결과를 얻었다. 이 연구는 후처리 기법의 결합이 합성 데이터의 신뢰성과 성능을 극대화하는 데 효과적임을 보여준다.

1. 서론

금융 데이터는 민감한 정보를 포함하고 있어 신중한 처리가 요구되지만, 개인정보 보호 문제로 실제 데이터를 사용할 수 없는 경우가 많다. 이를 해결하기 위한 대안으로 합성 데이터(synthetic data) 기술이 주목받고 있다.

본 연구에서는 기존 후처리 방식[1]에 더해, 단일 최빈값의 근삿값을 치환하는 후처리 기법을 제안한다. 이 기법을 SVS (Single-value substitution)로 정의하며, 기존 기법과의 비교 및 결합을 통해 생성 데이터의 품질을 향상시키고, 분석 및 예측 모델의 성능을 극대화하는 것을 목표로 한다.

2. CTGAN

CTGAN (Conditional Tabular GAN)은 테이블 형식의 데이터를 생성하는 모델로, 특히 범주형, 연속형 데이터가 섞인 복잡한 분포와 불균형 데이터를 효과적으로 처리한다. 일반적인 GAN 구조를 따르며, 생성자와 판별자가 상호 경쟁하면서 데이터를 학습한다.

$$L_D = -\left[\mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x|c)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z|c)))]\right]$$

$$L_G = -\mathbb{E}_{z\sim p_z(z)}[\log D(G(z|c))]$$

CTGAN은 비대칭적 분포나 범주형 변수 처리에 강점이 있으며, 다양한 분야에서 합성 데이터를 생성하는 데 적합한 모델이다.

3. 연구 방법

3.1. 데이터셋

본 논문에서는 AI-Hub[2]에서 제공하는 '금융 합성 데이터'의 카드 청구정보 데이터를 활용한다. 생성에 쓰인 원천데이터의 크기는 10,000개이며, 컬럼은 32개의 이산형 데이터, 15개의 연속형 데이터 컬럼으로 총 57개이다.

3.2. 합성 데이터 생성

CTGAN 모델을 사용하여 원본 데이터의 특성을 학습하고, 100 epochs로 모델을 훈련시켰다. 모델 학습 후 10,000개의 샘플을 생성하였다.

3.3. 합성 데이터 평가

본 연구에서는 Column Shapes와 Column Pair Trends의 평균을 기준으로 합성 데이터를 평가하였으며 column shapes은 74.04%. column pair Trend는 88.55%로 총 81.29%의 유사성을 보였다. Column Shapes는 각 열의 분포를 비교하여, 실제 데이터와 합성 데이터의 값들이 얼마나 유사하게 분포하는지를 평가한다.

$$D_{KS} = \sup_x |F_{\text{real}}(x) - F_{\text{synthetic}}(x)|$$

$$D_{TV} = \frac{1}{2}\sum_x |P_{\text{real}}(x) - P_{\text{synthetic}}(x)|$$

Kolmogorov-Smirnov (KS) Test 또는 Total Variation (TV) Distance를 통해 두 데이터의 분포 차이를 측정한다. 연속형 변수의 경우 KS Test를 사용하고, 범주형 변수는 TV Distance로 평가한다.

Column Pair Trends는 두 열 간의 관계나 상관관계를 평가하는 지표이다.

$$\rho_{X,Y} = \frac{\sum(X_i - \mu_X)(Y_i - \mu_Y)}{\sqrt{\sum(X_i - \mu_X)^2}\sqrt{\sum(Y_i - \mu_Y)^2}}$$

실제 데이터에서 두 열 간의 상관관계가 유지되는지를 분석하며, Pearson 상관계수와 같은 통계적 방법을 사용하여 열 간의 관계를 비교한다.

3.4. 생성 데이터 후처리

기존의 후처리 방식인 동적 샘플 필터링과 임계값 최적화 기법을 결합하여 합성 데이터의 품질을 개선한다. 동적 샘플 필터링은 합성된 데이터 중 특정 기준을 충족하지 못하는 샘플을 제거하는 방식으로, 원본 데이터의 특성을 잘 반영하도록 보정한다.

3.4. 단일 최빈값 치환 후처리

본 연구는 원천데이터에 특정 열에 단일값 0에 대해 비율이 높은 데이터들의 column shapes 점수가 낮은 것이 관찰되었다. 해당 열들의 생성데이터를 분석한 결과 원천데이터와 달리 최빈값의 근삿값이 존재했으며 0이 아닌 값의 비율이 극소한 원천데이터의 특성상 0의 근삿값들이 생성데이터의 품질을 낮춘다고 판단하였다. $P_{0,i}$은 열i에서 0이 차지하는 비율, X_i은 원천 데이터 값의 집합이며 threshold_i은 원천데이터에서 0이 아닌 값들 중 가장 작은 값의 50%에 해당하는 기준값이다.

$$P_{0,i} = \frac{\sum_{x \in X_i} 1(x = 0)}{n} \qquad \text{if } P_{0,i} > 0.7$$

$$Y_i = \begin{cases} 0 & \text{if } Y_i < \text{threshold}_i \\ Y_i & \text{otherwise} \end{cases} \qquad \text{threshold}_i = 0.5 \times \min(X_i \setminus \{0\})$$

원천데이터에서 최빈값의 비율이 높은 열에 대해 기준값을 설정하고 합성 데이터에서 0의 근삿값을 0으로 치환하는 새로운 후처리 기법을 제안한다.

4. 결과

기본 모델의 성능은 column shapes에서 74.04%, overall 성능은 81.29%로 상대적으로 낮은 수준을 보였다. 하지만 DSF를 적용한 후에는 성능이 모든 지표에서 고르게 상승하였으며, 특히 column pair trends는 91.73%로 향상되었다. 이는 DSF가 저품질 샘플을 효과적으로 제거하였음을 보여준다.

SVS 기법을 적용한 경우, column shapes가 89.90%로 급격히 상승하였으며, overall 성능도 89.13%로 개선되었다. 이는 특정 변수 샘플링이 데이터의 특성에 맞는 변수를 선택함으로써 데이터의 표현력을 극대화했음을 시사한다.

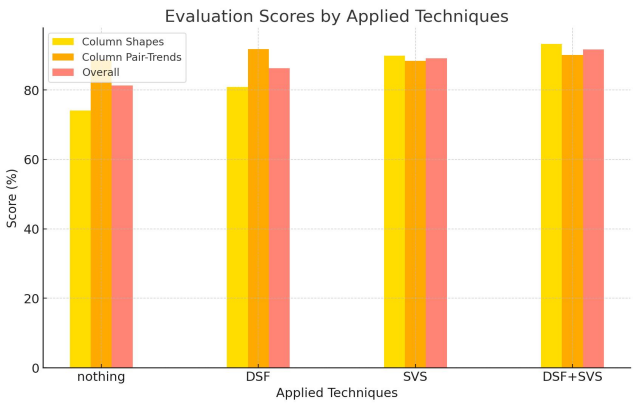
가장 큰 성능 향상은 DSF와 SVS를 결합했을 때 column shapes에서 93.24%, overall에서 91.68%로 모든 측정 지표에서 최고의 성능을 기록하였다. 이는 두 기법이 상호 보완적으로 작용하여 데이터 품질을 크게 향상시켰음을 의미한다.

결과적으로, DSF와 SVS를 결합하는 것이 데이터셋의 신뢰성과 성능을 극대화하는 방법임을 확인할 수 있었다. 이러한 결과는 데이터 신뢰도 향상을 위한 중요한 전략을 제안할 수 있다.

<표 1> 후처리 적용 후 데이터 평가 점수

applied techniques	column shapes	column pair-trends	overall
nothing	74.04%	88.55%	81.29%
DSF	80.89%	91.73%	86.31%
SVS	89.9%	88.36%	**89.13%**
DSF+SVS	93.24%	90.11%	**91.68%**

<그래프 1> 후처리 적응 후 데이터 평가 점수



※ 본 논문은 과학기술정보통신부 대학디지털교육역량강화 사업의 지원을 통해 수행한 ICT멘토링 프로젝트 결과물입니다.

참고문헌

[1] Andrea Lampis, Eugenio Lomurno, Matteo Matteucci, "Bridging the Gap: Enhancing the Utility of Synthetic Data via Post-Processing Techniques," arXiv preprint arXiv:2305.10118v2, 2023.

[2]금융합성데이터,https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=71792