관계형 테이블 합성데이터를 위한 기존 평가 지표의 한계와 개발 방향

이수빈 [1], 배호 [2,3]
[1] 이화여자대학교 인공지능융합전공 석사과정
[2] 이화여자대학교 사이버보안학과 교수
[3] 이화여자대학교 인공지능융합전공 교수

sbin52@ewha.ac.kr, hobae@ewha.ac.kr

Limitations and Improvements of Evaluation Metrics for Relational Tabular Synthetic Data

Su-Bin Lee[1], Ho Bae[2,3]
[1]Dept. of Artificial Intelligence Convergence, Ewha Womans University
[2]Dept. of Cyber Security, Ewha Womans University
[31]Dept. of Artificial Intelligence Convergence, Ewha Womans University

요 약

합성데이터는 통계적 특성이 유사한 가상의 데이터로, 개인정보 보호 및 데이터 부족 문제를 해결하는 데 기여한다. 이를 관계형 데이터베이스로 확장한 관계형 테이블 합성데이터는 금융, 통신 등 다양한 응용 분야에서 사용되고 있으며 이에 대한 유용성과 안전성을 평가하는 다양한 지표들이 개발되어왔다. 그러나 현재 사용되는 평가지표는 단일 테이블이나 여러 테이블을 하나로 결합한 후 평가하는 경우가 많아 관계형 데이터의 복잡한 구조를 충분히 반영하지 못한다는 한계가 있다. 따라서 본 논문은 관계형 테이블 합성데이터 평가 시 기존 지표에 대한 한계를 분석하고, 데이터 간 관계 보존을 효과적으로 평가할 수 있는 포괄적 평가 지표의 필요성을 강조하며, 이를 위한 향후 지표 개발 방향성을 논의한다. 본 연구는 관계형 테이블 합성데이터의 신뢰성과 품질을 높이는 데 중요한 기여를 할 것으로 기대된다.

1. 서론

합성데이터[1]는 실제 데이터와 통계적 특성이 유사하여, 실제 데이터 분석 결과와 유사한 결과를 얻을 수 있도록 새롭게 생성해낸 가상의 데이터로 개인정보보호, 데이터부족, 그리고 머신러닝 모델에서 발생하는 편향문제를 해결[2]하기 위해 사용된다. 실제로 합성데이터는 의료[3], 금융[4] 등 여러 응용분야에서 실제 데이터를 대체할 수 있으며, 동시에 개인정보 유출위험을 줄이는 데 기여한다. 특히 테이블 합성 데이터의 경우 범주형, 연속형, 시간적 의존성이 포함된 복합적인 데이터 구조를 가질 수 있어 여러 산업에서 활용가능성이 높아지고 있다.

이러한 합성데이터의 품질과 신뢰성을 평가하기 위해 평가지표는 필수적이다. 대표적으로 데이터 유효성 측면에서 실제데이터와의 유사성을 평가하는 지표와 안정성 측면에서 합성데이터가 실제 개인을 식별할 수 없도록 적절히 처리되었는지 확인하는 평가 지표[5]들이 있다. 특히 테이블 합성데이터의 경우에는 앞서 제시된 원본 데이터의 복합적인 구조적 특성을 유지하기 위해 컬럼단위의 분포와 컬럼 간 상관관계를 평가하는 지표들이 있다.

이미지, 텍스트와 더불어 테이블 도메인에서 합성데이터 연구[6][7][8]가 활발히 진행되고 있으며, 이러한 데이터의 평가를 위한 평가지표 개발도 이어지고 있다. 특히, 최근에는 관계형 데이터베이스의 광범위한 활용을 바탕으로, 관계형 테이블 합성 데이터 생성 연구[9][10]가 시작되었고, 앞으로 더 활발한 연구가 기대된다. 그러나 관계형 테이블 합성 데이터에 대한 포괄적이고 체계적인 평가지표 연구는 아직 미비하다. 관계형 데이터는 여러 테이블 간의 복잡한 상관관계를 포함하므로, 이를 평가하기 위해서는 기존 지표보다 더욱 정교하고 구체적인 지표가 필요하다.

따라서 본 논문에서는 관계형 합성데이터 평가에 있어서 기존 지표가 지닌 한계를 분석하고, 이를 통해 관계형 데이터의 복잡성을 반영할 수 있는 포괄적인 평가지표 개발의 필요성과 방향성을 논의하고자 한다.

2. 관계형 테이블 합성데이터 평가의 한계

2.1 기존 테이블 합성데이터 평가지표

테이블 합성데이터에 대한 기존 평가지표는 다양한 연구[에서 제시되어왔으며 주로 데이터의 유효성 과 안정성 두가지 측면을 중심으로 개발되어왔다.

유용성	머신러닝 성능	RF-utility, Machine Learning Efficacy-based, Utility Evaluation
	변수 간 유사성	Marginal Distribution, Pairwise correlation, Leave-One-Out
	기타	Log Cluster, Support Coverage
안전성	거리기반	ε-Identifiability
	재식별 정도	Attribute Disclosure, Identity Disclosure

표 1 테이블 합성데이터 평가지표 종류

데이터의 유용성 평가 측면에서는 머신러닝 모델의 성능을 기반으로 평가하는 방법들이 대표적으로 사용된다. 예를 들어 RF-utility [11], Machine Learning Efficacy-based Metric[12], Utility Evaluation[13]은 합성데이터를 사용하여 학습한 머신러닝 모델과 실제데이터로 학습한 모델 간의 성능차이를 비교한다. 이때 성능차이가 적을수록 합성데이터가 실제 데이터의 특성을 잘 반영하고 있음을 의미한다.

또한 변수 간의 유사성을 평가하는 방법도 제안되어 왔다. 각 변수의 분포(단일 변수, Marginal)나 변수 간 상관관계(쌍 변수, Pairwise)[14]를 측정하여 원본 데이터와 합성 데이터 간의 유사성을 평가한다. PCD(Pairwise Correlation Difference) [14]는 데이터셋 단위로 원본 데이터와 합성 데이터의 상관행렬 차이를 분석하는 쌍별 상관관계 차이를 사용하여 변수 간의 관계를 평가한다. Leave-One-Out(LOO)[15]는 특정 변수를 제외한 나머지 변수들로 해당 변수를 예측하는 방식으로, 이 방법은 전체 분포를 완전히 표현할 수 있기 때문에 전체 분포와의 일치를 평가한다. 이러한 지표들은 주로 개별 테이블 데이터를 대상으로 하며, 실제 데이터의 구조적 특성이 잘 보존되었는지를 확인하는 데 사용된다.

이외에도 클러스터링을 통해 실제 데이터와 합성데이터의 잠재구조가 얼마나 유사한지 평가하는 지표인 Log Cluster[14], 원본데이터의 다양한 카테고리가 합성데이터에 어느정도 반영되었는지 평가하는 지표인 Support Coverage[12] 등 이외 다양한 지표가 테이블 합성데이터를 평가하기 위해 사용되고 있다.

안전성 측면에서는, 원본 데이터와 합성 데이터 간의 유사도가 개인정보 유출 위험을 판단하는 기준으로 사용된다. 예를 들어, [13]은 테이블 간의 유클리드 거리나 하우스도르프 거리를 측정하여 합성 데이터가 원본 데이터를 얼마나 보호하고 있는지를 평가한다. ε-Identifiability[11]에서는 레코드 단위로 원본과 가장 가깝게 생성된 레코드 간의 거리를 비교하여 재식별 위험을 측정한다. 이는 합성 데이터가 원본 데이터와 너무 유사할 경우, 개인정보가 노출될 위험이 크다는 점을 반영한 평가 방법이다.

이외에도 Attribute Disclosure[14]는 특정속성을 알고 있는 상태에서 k 개의 가장 가까운 이웃을 찾아 다수결투표로 알려지지 않은 속성을 추론하는 정확도를 통해 합성데이터가 원본데이터를 얼마나 보호하는지를 평가한다. Identity Disclosure[14]역시 모델학습에 사용된 실제데이터와 합성데이터를 비교한 후 특정 개인이 합성 데이터에 포함되었는지 여부를 통해 원본 데이터를 재식별할 수 있는 정도를 평가한다.

이처럼 테이블 합성 데이터에 대한 평가 지표들은 데이터의 유사성뿐만 아니라, 개인정보 보호 측면에서도 중요한 역할을 한다. 그러나 이러한 지표들은 주로 단일 테이블 데이터를 대상으로 개발된 것이기 때문에, 복잡한 관계를 포함하는 관계형 테이블 합성데이터에 동일하게 적용하기에는 한계가 존재한다.

2.2 관계형 테이블 합성데이터 기존 평가 방법

관계형 테이블 합성데이터는 여러 테이블 간의 관계를 표현하는 데이터 구조로, 기본 키와 외래 키를 통해 선형 관계, 다중 자식 관계, 다중 부모 관계, 그리고 다중 부모 및 자식 관계 등[10] 테이블 간의 다양한 부모-자식 관계를 형성한다. 이러한 관계형 데이터는 단순히 개별 테이블의 열과 행을 모델링하는 것이 아니라, 테이블 간의 복잡한 구조를 유지해야 한다[16]. 따라서 관계형 데이터의 특성들을 반영한 새로운 평가 지표가 필요하다.

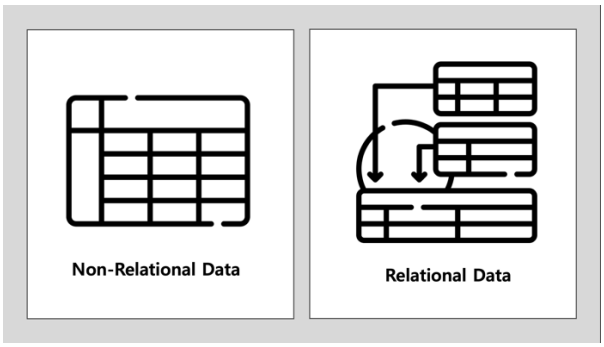

그림 1 비관계형데이터 vs 관계형데이터

초기 연구[17]에서는 관계형 합성데이터를 평가하기 위해 기존의 테이블 데이터 평가 방법을 적용하였다. 각 열의 데이터 분포가 실제 데이터와 합성 데이터 간에 유사한지를 평가하기 위해 Kolmogorov-Smirnov 테스트와 같은 통계적 방법을 사용하여 분

포 일관성을 확인하고, 공분산 평가를 통해 데이터 간의 상관관계를 분석하여, 합성 데이터가 원본 데이터의 의존성을 얼마나 잘 반영했는지 측정하였다. 그러나 이러한 방법은 관계형 데이터의 복잡한 부모-자식 관계를 충분히 반영하지 못하며 실제로 원본 데이터의 부모-자식 관계가 합성 데이터에서 깨졌을 때도 기존 평가 방법으로는 탐지하기 어렵다.

최근 연구에서는 관계형 데이터의 복잡한 구조를 보다 정교하게 평가하기 위한 지표들이 제안되었다. 예를 들어, [18]에서는 거리 측정(Distance to Closest Record, DCR)를 사용하여 관계형 합성 데이터를 평가한다. DCR 은 각 관측치 간의 거리로 정의된 거리행렬을 생성하여 최소거리를 계산한 후 합성 데이터가 원본 데이터와 얼마나 가까운지를 측정하는데, 이는 근본적으로 기존의 거리기반 지표와 유사한 방식이다.

또한 [10]에서는 부모-자식 관계를 갖는 합성 테이블을 비정규화하여 하나의 테이블로 결합한 후, 이 테이블을 기반으로 실제 데이터와 합성데이터를 구분하는 이진분류기의 AUROC 성능을 평가지표로 사용하였다. 이 지표의 성능이 낮을수록 두 데이터를 구분하기 어렵다는 의미로 실제 데이터와 합성데이터가 유사하고 관계형 데이터 간의 중요한 구조적관계를 잘 보존되었음을 나타낸다. [16]에서도 원본데이터와 합성데이터를 사용하여 각각 학습 시킨 후 원본데이터에서의 AUROC 및 F1-score 같은 성능 지표를 통해 데이터의 유사성을 비교한다. 그러나 이는 AUROC 성능이 낮게 나왔을 때 데이터 간의 유사성이 부족한 이유가 테이블 간의 관계 손상 때문인지, 아니면 단순한 데이터 분포 차이 때문인지를 명확하게 파악하는 것이 어렵다.

안전성 측면에서는 Qδ 통계 분위수[18]를 기반으로 분포 차이를 분석함으로써 합성 데이터가 원본 데이터와 너무 유사해지지 않도록 학습을 중단시키는 일반화 지표로 사용되었다. 최근접 이웃 거리 기반 지표[16]를 통해 합성 데이터가 원본 데이터와 너무 가깝지 않은지를 평가하여 개인정보 보호 여부를 확인할 수 있다

이처럼 관계형 테이블 합성 데이터를 평가하기 위한 새로운 지표들이 제안되었음에도 불구하고, 여전히 기존의 평가 방법에서 크게 벗어나지 않고 있으며, 이러한 지표들이 실제 관계형 데이터의 복잡한 관계를 충분히 반영하는지에 대해서는 의문이 제기될 수 있다.

3. 향후 평가 지표 개발 방향성

관계형 합성데이터의 평가 지표 개발에 있어, 향후 방향성은 기존의 테이블 기반 평가 방식이 가지는 한계를 보완하고, 데이터 간의 복잡한 관계를 보다 직관적으로 반영할 수 있는 새로운 접근법이 필요하다. 기존 방식은 주로 각 테이블의 분포 유사성이나 독립적인 열 데이터를 평가하는 데 중점을 두었지만, 관계형 데이터는 테이블 간의 상호 의존성을 효과적으로 평가할 수 있는 방법이 상대적으로 부족하다. 이러한 한계를 극복하기 위해, 원본 데이터와 합성 데이터를 그래프 구조로 모델링하여 추가적으로 두 데이터 간의 구조적 유사성을 평가하는 지표를 가중치로 사용하도록 제안할 수 있다.

각 테이블을 노드로, 테이블 간의 관계를 엣지로 모델링한 후, 두 그래프 간의 유사성을 평가하는 방식이다. 이를 통해 테이블 간의 복잡한 관계를 시각화하고, 원본 데이터와 합성 데이터 간의 관계적 구조를 보다 명확하게 비교할 수 있다. 이러한 방식은 관계형 데이터가 가지는 구조적 복잡성을 보다 효과적으로 반영할 수 있으며, 특히 기존의 분포 기반 평가 지표가 다루기 어려웠던 테이블 간의 관계 보존 여부를 명확하게 평가할 수 있다.

대표적인 방법으로 Graph Edit Distance[19]와 Graph Isomorphism[20]이 있다. Graph Edit Distance 는 한 그래프를 다른 그래프로 변환하기 위해 필요한 최소 편집 횟수를 의미하며, 두 그래프 간의 구조적 유사성을 정량화하는 지표로 사용할 수 있다. 편집 거리가 작을수록 두 그래프가 유사한 구조를 가지고 있음을 나타낸다. Graph Isomorphism 두 그래프가 동일한 구조를 가지고 있는지 여부를 판단하는 방법으로, 관계형 데이터의 중요한 관계 구조를 평가하는 데 유용하다. 이 두 지표를 통해 원본 데이터와 합성 데이터 간의 관계 보존 정도를 더 정확하게 측정할 수 있다.

더불어, 기존의 관계형 데이터베이스에서 안전성 고려한 연구는 아직 충분하지 않으며, 이 분야에 대한 다양한 연구가 요구된다. 특히 관계형 데이터는 그 구조적 특성상 테이블 간 결합 과정에서 민감한 정보가 의도치 않게 노출될 위험이 크다. 예를 들어, 외래 키와 기본 키를 통해 형성된 테이블 간의 관계는 데이터 간의 연결 방식을 드러내며, 이로 인해 데이터가 재식별될 가능성이 있다. 따라서 관계형 합성 데이터에서는 데이터의 구조적 관계를 보호하는 것이 매우 중요하다. 다른 테이블과의 관계를 통해 재식별될 가능성을 평가하거나, 특정 구조적 패턴이 개인정보를 추론하는 데 악용될 수 있는지를 평가하는 새로운 지표가 필요하다.

일반적으로 안전성 평가 지표로 k-익명성(k-anonymity)[21]이나 l-다양성(l-diversity)[22] 개념이 사용되지만, 이러한 지표는 단일 테이블 환경에서만 유효한 경우가 많다. 다중 테이블 환경에서는 테이블 간의 결합을 통해 재식별 위험이 존재할 수 있다. 따라서, 다중 테이블 간 결합으로 인한 **개인정보 침해 가능성**을 평가하기 위해 결합 민감도(Joint Sensitivity)라는 새로운 개념을 도입할 수 있다. **결합 민감도**는 각 테이블의 민감도(sensitivity)를 기반으로, **외래 키와 기본 키** 등의 관계를 통해 연결된 테이블 간의 결합이 이루어졌을 때 **재식별 위험**이 얼마나 증가하는지를 정량적으로 평가하는 지표를 사용하여 관계형 데이터 전체에 대한 개인정보 보호 수준을 평가할 수 있다. 이렇듯, 다중 테이블 결합 시의 개인정보 유출 가능성까지 고려한 포괄적인 평가가 필요하다.

향후 이러한 요소들을 고려하여, 관계형 데이터의 구조적 특성과 안전성을 모두 반영하는 새로운 평가 지표들이 제안될 수 있다.

4. 결론

관계형 테이블 합성데이터의 평가에 있어서 기존 지표들이 가진 한계를 분석하고, 복잡한 데이터 구조와 안전성 측면에서 보다 포괄적인 평가 방법의 필요성을 제시하였다. 기존의 평가지표들은 주로 단일 테이블 데이터를 대상으로 데이터 유사성과 개인정보 보호 측면에서 개발되어 왔지만, 관계형 데이터의 복잡한 부모-자식 관계를 충분히 반영하지 못하는 한계가 있었다. 이러한 문제를 해결하기 위해 관계형 테이블 간의 복잡한 구조적 유사성을 반영할 수 있는 그래프 기반 평가 방법을 제안하고, 안전성을 고려한 새로운 지표 개발의 필요성을 강조하였다. 향후 연구는 관계형 데이터의 구조적 복잡성을 보다 효과적으로 반영하는 평가 지표 개발에 중점을 두어야 하며, 이를 통해 관계형 테이블 합성데이터의 품질과 신뢰성을 높이는 데 기여할 수 있을 것이다.

5. 사사

이 논문은 2024 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No.RS-2022-00155966, 인공지능융합혁신인재양성(이화여자대학교))

참고문헌

[1] 개인정보보호위원회, "합성데이터 생성 참조모델," 개인정보보호위원회, 대한민국, 2023.
[2] M. Giuffrè, D. L. Shung, "Harnessing the power of synthetic data in healthcare: innovation, application, and privacy," *NPJ Digital Medicine*, vol. 6
[3] H. Murtaza, M. Ahmed, N. F. Khan, G. Murtaza, S. Zafar, A. Bano, "Synthetic data generation: State of the art in health care domain," *Computer Science Review*, vol. 48, pp. 100546, May 2023.
[4] M. Dogariu, B. Kim, L.-D. Ştefan, B.-A. Boteanu, C. Lamba, B. Ionescu, "Generation of Realistic Synthetic Financial Time-series," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 18, no. 3, Mar. 2022
[5] M. Goyal, Q. H. Mahmoud, "Privacy Mechanisms and Evaluation Metrics for Synthetic Data Generation: A Systematic Review," *Electronics*, vol. 13, no. 17, pp. 3509, Sept. 2024.
[6] Z. Zhao, A. Kunar, R. Birke, L. Y. Chen, "CTAB-GAN: Effective Table Data Synthesizing," in *Proceedings of The 13th Asian Conference on Machine Learning*, PMLR, vol. 157, pp. 97-112, Nov. 2021.
[7] M. Vero, M. Balunović, M. Vechev, "CuTS: Customizable Tabular Synthetic Data Generation," in *Proceedings of the 41st International Conference on Machine Learning*, PMLR, vol. 235, pp. 49408-49433, July 2024.
[8] T. Sattarov, M. Schreyer, D. Borth, "FedTabDiff: Federated Learning of Diffusion Probabilistic Models for Synthetic Mixed-Type Tabular Data Generation," *arXiv preprint*, arXiv:2401.06263, Jan. 2024.
[9] K. Cai, X. Xiao, G. Cormode, "PrivLava: Synthesizing Relational Data with Foreign Keys under Differential Privacy," *Proceedings of the ACM on Management of Data*, vol. 1, no. 2, pp. 142:1-142:25, 2023.
[10] M. Park, S. Kang, "Row Conditional-TGAN for generating synthetic relational databases," *Proceedings of the 2021 International Conference on Data Mining and Applications*, pp. 78-85, 2021.
[11] M. Miletic, M. Sariyar, "Challenges of Using Synthetic Data Generation Methods for Tabular Microdata," *Applied Sciences*, vol. 14, no. 14, pp. 5975, 2024.
[12] V. S. Chundawat, A. K. Tarun, M. Mandal, M. Lahoti, P. Narang, "TabSynDex: A Universal Metric for Robust Evaluation of Synthetic Tabular Data," *arXiv preprint*, arXiv:2207.05295, July 2022.
[13] M. Hernandez, G. Epelde, A. Alberdi, R. Cilla, D. Rankin, "Synthetic Tabular Data Evaluation in the Health Domain Covering Resemblance, Utility, and Privacy Dimensions," *Methods of Information in Medicine*, vol. 62, no. 1, pp. 11-22, 2023.
[14] A. Goncalves, P. Ray, B. Soper, J. Stevens, L. Coyle, A. P. Sales, "Generation and Evaluation of Synthetic Patient Data," *BMC Medical Research Methodology*, vol. 20, no. 1, pp. 1-13, May 2020.
[15] S. C. Yang, B. Eaves, M. Schmidt, K. Swanson, P. Shafto, "Structured Evaluation of Synthetic Tabular Data," *arXiv preprint*, arXiv:2403.10424, March 2024.
[16] C. A. Mami, A. Coser, A. T. P. Boudewijn, M. Volpe, M. Whitworth, D. Panfilo, S. Saccani, "Generating Realistic Synthetic Relational Data through Graph Variational Autoencoders," *Proceedings of NeurIPS 2022*, Dec. 2022.
[17] N. Patki, R. Wedge, K. Veeramachaneni, "The Synthetic Data Vault (SDV)," in *Proceedings of the 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 399-410, Oct. 2016.
[18] A. V. Solatorio, O. Dupriez, "REaLTabFormer: Generating Realistic Relational and Tabular Data using Transformers," *arXiv preprint*, arXiv:2302.02041, Feb. 2023.
[19] A. Sanfeliu, K. S. Fu, "A distance measure between attributed relational graphs for pattern recognition," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 13, no. 3, pp. 353–362, 1983.
[20] H. Whitney, "Congruent Graphs and the Connectivity of Graphs," *American Journal of Mathematics*, vol. 54, pp. 150–168, 1932.
[21] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy," *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, vol. 10, no. 5, pp. 557-570, 2002.
[22] A. Machanavajjhala, D. Kifer, J. Gehrke, M. Venkitasubramaniam, "ℓ-Diversity: Privacy Beyond k-Anonymity," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, pp. 3-23, March 2007.