Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.34, NO.2, Apr. 2024 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2024.34.2.323

오픈 데이터 환경에서 개인정보 노출 위험 측정을 위한 통계적 방법론 연구*

김 시 은,[1†] 엄 익 채[2‡]
[1,2]전남대학교 (대학원생, 교수)

A Statistical Methodology Study for Measuring Privacy Disclosure Risk in Open Data Environment*

Sieun Kim,[1†] Ieck-chae Euom[2‡]
[1,2]Chonnam National University (Graduate student, Professor)

요 약

최근 합성데이터는 실제 데이터의 패턴과 특성을 유지하면서도 프라이버시 유출을 예방할 수 있는 기술로 각광받고 있다. 그에 따라 합성데이터 활용에 대한 기술적 및 제도적 연구가 활발하게 진행되고 있지만, 명확한 표준과 지침이 부재하여 합성데이터 기술의 적극적 활용이 어렵다. 본 연구는 합성데이터의 개인정보 노출 위험 정량화를 위한 전초적 연구로 통계적 방법론을 활용하여 개인정보 노출 위험 정량화 지수를 도출하고 유럽 개인정보 보호 규정(General Data Protection Regulation)를 준수하기 위한 구체적인 적용 방안을 제시한다. 오픈 데이터 환경에서 본 연구의 개인정보 노출 위험 지수를 통해 개인정보 노출 위험을 인식하고 데이터 활용성과의 균형을 조절할 수 있을 것으로 기대한다.

ABSTRACT

Recently, Syntheic data has been in the spotlight as a technology that can protect personal information while maintaining the patterns and characteristics of actual data. Accordingly, technical and institutional research on synthetic data is actively being conducted, but it is difficult to actively use synthetic data due to the lack of clear standards and guidelines. This study is a preliminary study for quantifying the disclosure risk of synthetic data, and derives a privacy disclosure risk index through statistical methodology and suggests specific application measures to comply with the General Data Protection Regulation(GDPR). It is expected that the disclosure risk and the balance of data utility can be controlled through the privacy disclosure risk index of this study in an open data environment.

Keywords: Synthetic data, Open data, De-identification, Data privacy, Disclosure risk assessment

I. 서 론

최근 다양한 산업 분야에서는 크롤링, Open API, 로그, 센서, DBMS, FTP 등의 기술로 대량의 데이터를 수집하고 있다. 다양한 경로로 수집된 데이터는 부가가치 창출을 위해 인공지능 기술인 머

Received(01. 11. 2024), Modified(02. 23. 2024), Accepted(02. 26. 2024)
* 본 논문은 2023년도 한국정보보호학회 호남지부 학술대회에서 발표한 우수논문을 개선 및 확장한 것임.
* 이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구(RS-2023-00242528, 50%)와 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 지역지능화혁신인재양성사업(IITP-2024-00156287, 50%)의 지원을 받은 연구임.
† 주저자, sinnie613@jnu.ac.kr
‡ 교신저자, iceuom@jnu.ac.kr(Corresponding author)

신러닝과 딥러닝 기술에 사용된다. 하지만 인공지능 기술 활성화를 위해서 학습데이터 불균형, 라벨 부족, 오픈 데이터 부족 등의 문제 해결이 필요하다.

데이터 3법의 개정과 금융 데이터 규제의 완화로 개인정보가 포함된 데이터 활용이 점차 증가하고 있다. 그러나 민감정보를 포함한 데이터의 활용시 개인정보의 비식별화가 필수적이며, 다양한 가명·익명 처리 기법들이 고려되고 있다. 최근에는 프라이버시 보호와 데이터 유용성 확보를 위한 합성데이터 기술이 주목받고 있다. 합성데이터는 전통적인 비식별화 기술과 달리, 원본 데이터와의 일대일 대응성이 없으며, 개인정보 노출 위험을 줄이면서 유용성을 유지한다. 가트너사의 '2023년 인공지능 하이프사이클'에 따르면 합성데이터는 혁신적인 성과를 낼 수 있는 기술이며 2025년에는 AI 모델에서 원본 데이터보다 합성데이터가 더 많이 사용될 것으로 전망하고 있다.

합성데이터 생성을 위해 GAN(Generative Adversarial Networks), VAE(Variational Autoencoder)등의 생성 모델이 주로 사용되며 데이터의 특성을 학습하여 데이터를 재현하기 때문에 개인정보 노출 위험이 있다. 이에 더해 익명 데이터가 추가 정보와 결합되면 연결 공격이나 추론 공격 등 개인정보 노출이 발생할 수 있다. 그러나 기존 연구들은 합성데이터의 유용성, 즉 원본 데이터와 통계적 유사성과 원본 데이터의 특정 분석(분류, 회귀 등)을 재현 능력을 보이는 연구가 대부분이다. 또한 합성데이터의 개인정보 노출 위험을 평가한 연구는 위험이 사후적으로 이루어지거나, 특정 공격에 국한하여 평가한다. 이처럼, 합성데이터에 평가에 대한 표준이나 지침의 부재는 합성데이터 연구 고도화와 적극적 활용에 제약 사항이다. 이에 본 연구는 국내외 비식별 처리 관련 지침 분석을 통해 합성데이터가 가질 수 있는 주요 위험을 도출하고 개인정보 노출 위험 정량화를 위한 통계적 방법론을 연구한다.

II. 관련 연구

본 장에서는 본 논문에서 사용되는 용어의 정의, 개인정보 비식별 관련 국내외 지침, 기존 연구들의 합성데이터 평가 방법을 분석한다.

2.1 용어 정의

본 논문에서 사용되는 비식별화 정도에 따른 개인정보 유형과 개인 식별성 정의는 다음과 같다.

- 비식별화 : 정보 주체와 식별속성의 집합 간 연계를 제거하는 과정[5]
- 개인정보(personal data) : 살아있는 개인에 관한 정보로 성명, 주민등록번호, 영상 등 개인을 알아볼 수 있는 정보[2]
- 가명 정보(pseudonymous data) : 개인정보를 일부 또는 정보를 삭제, 대체하는 등 가명 처리를 통해 추가정보 없이는 개인을 알아볼 수 없는 정보[2]
- 익명 정보(anonymous data) : 시간 · 비용 · 기술 등을 합리적으로 고려할 때 다른 정보를 사용하더라도 더이상 개인을 알아볼 수 없는 정보[2]
- 추가 정보(auxiliary information) : 개인정보의 전부 또는 일부를 대체하는 가명처리 과정에서 생성 또는 사용된 정보로서 특정 개인을 알아보기 위하여 사용·결합 될 수 있는 정보[2]
- 합성데이터(synthetic data) : 원본과 최대한 유사한 통계적 성질을 보이는 가상의 데이터를 생성하기 위해 개인정보의 특성을 분석하여 새로운 데이터를 생성하는 기법[5]
- 개별화(single out) : 데이터 주체를 고유하게 식별하는 것으로 알려진 일련의 특성을 관찰하여 데이터 집합에서 데이터 주체에 속하는 레코드를 분리하는 행위[2]
- 연결가능성(linkability) : 데이터 주체에 관한 레코드를 별도의 데이터 집합에서 동일한 데이터 주체에 관한 레코드와 연관(연결)할 수 있는 데이터 집합에 대한 속성[2]
- 추론(inference) : 하나 이상의 속성의 값을 사용하거나 외부 데이터 소스를 연관시킴으로써 무시할 수 없는 정도의 확률로 알려지지 않은 정보를 추론하는 행위[2]

2.2 개인정보 비식별 처리 관련 국내외 지침

2.2.1 국내 지침 현황

2016년 국무조정실, 행정자치부 등이 '개인정보 비식별 조치 가이드라인'[1]을 마련하여 개인정보 정보 비식별 조치 기준과 비식별 정보의 활용 범위 등을 제시하였다. 해당 지침에서 비식별 정보란 데이터 셋에 대해 지침에 따라 적정하게 비식별 조치 된 정보

를 말한다. 비식별 조치란 데이터 셋에서 개인을 식별할 수 있는 요소를 전부 또는 일부 삭제하거나 대체하는 등의 방법을 사용하여 개인을 알아볼 수 없도록 하는 조치'를 의미한다. 해당 지침의 적정성 평가는 k-익명성을 적용하여 측정한다. 2020년 개인정보 보호법이 개정되어 데이터 경제 활성화의 토대가 마련되었다. 후속 조치로 개인정보보호위원회에서 '가명정보처리 가이드라인'[2]을 발행하였다. 해당 가이드라인에서는 가명정보와 익명정보에 따른 처리 가이드라인을 구별 하지 않는다. Fig.1은 가명정보처리 가이드라인의 개인정보 가명·익명 처리 기술 종류를 정리 및 재구성한 표이다. 2022년 개정된 지침의 적정성 검토는 처리 환경과 항목별 위험도 분석을 통해 이루어진다. 2024년 2월 개정된 지침에는 비정형데이터 가명처리 기준에 대한 내용이 추가되었다.

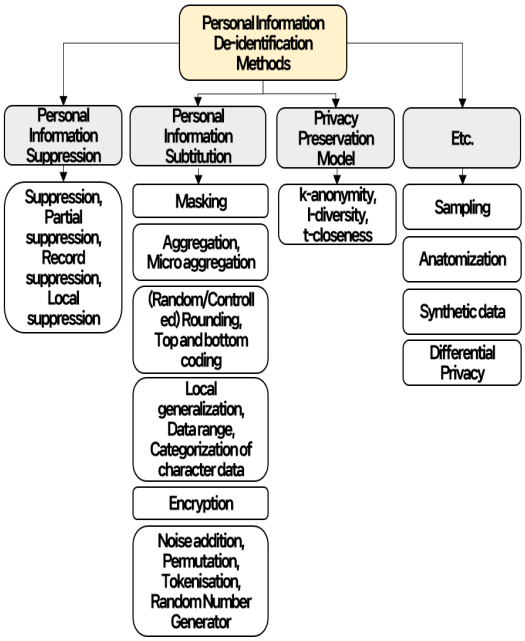

Fig. 1. Classification of De-identification Technology

2.2.2 국외 지침 현황

유럽연합 개인정보 보호 규정인 GDPR(General Data Protection Regulation)은 데이터의 가명처리와 익명 처리를 명확하게 구분하고 있다.[3] 해당 규정 제4조 5항에 따르면, 가명화(pseudonymisation)는 개인정보를 추가적인 정보 없이는 특정 개인에게 귀속될 수 없도록 처리하는 것으로 정의한다. 익명화(anonymisation)는 개인을 식별할 수 없는 방법으로 정보를 처리하는 것으로 정의한다. 개인정보 보호 원칙은 식별 가능한 개인에 관련되지 않은 정보, 정보주체가 식별 가능하지 않도록 익명 처리된 개인정보에는 적용되지 않는다. 비식별화 관련 표준으로는 ISO/IEC 20889(비식별 처리 기술 표준)이 존재한다.[5] 익명화 관련 지침으로는 EDPB(European Data Protection Board) 전신인 Article 29 Data Protection Working Party의 익명 처리 기법에 관한 의견서(Working Party Opinion 05/2014 on Anonymisation Techniques)가 있다.[4] 위 표준과 지침을 종합하면 익명 정보는 Single out, Linkability, Inference 위험을 가질 수 있으며 개인정보 재식별의 기준이 된다.

2.2.3 국내외 합성데이터 안정성 평가 지침 현황

2.2.1과 2.2.2에서는 합성데이터의 안정성 평가 지침의 부재로 개인정보 비식별 처리 관련 국내외 지침을 살펴보았다. 개인정보보호위원회의 '가명정보처리 가이드라인'에 따르면 '불특정 제3자에게 제공하는 경우(공개 등) 익명정보로 처리하는 것을 원칙으로 함'으로 기재되어 있지만, 익명정보의 안정성 및 적정성 평가 내용이 존재하지 않는다. 하지만, [4]에 따르면 익명정보는 세가지 노출 위험을 가질 수 있으며, 이 위험은 개인정보 재식별 기준이 된다.[5] 합성데이터는 익명 처리 기술 중 하나이며 개인정보 노출 위험이 존재한다.

2.3 합성데이터 평가 방법 분석

합성데이터 관련 연구들을 생성, GAN의 키워드를 통해 수집하였다.[6-15] 합성데이터 평가 방법에 중점을 두고 관련 연구 분석을 수행하였다. 각 연구는 연구 목적에 따라 평가 지표를 각기 채택하여 사용하였다. Table 1은 각 연구에서 사용한 평가 방법을 공통 범주로 묶어 비교 분석한 표이다.

합성데이터는 데이터 유용성, 개인정보 노출 위험 두 가지 측면에서 평가할 수 있다. 광역 유용성(넓은 범위 측정법)은 전체 데이터 셋의 일반적인 특징(두 데이터 세트 간의 전반적인 분포 유사성)을 평가한다. 특정 유용성(좁은 범위 측정법)은 합성데이터가 원본 데이터에서 수행된 특정 분석(데이터 요약이나 훈련된 모델의 계수 등)을 재현하는 능력을 평가한

Table 1. Classification of Evaluation Methods for Synthetic Data in Related Research

Author	Domestic/ International	Evaluation Metrics (○: Single Measure, ◎: Multiple Measure)								
		Global Utility			Specific Utility			Disclosure Risk		
		Statistical Fitness	Distance based	Basic statistics	Classification	Regression	Other metrics	Privacy risk	Distance based	Algorithm based
Dong-Suk Hong et al.(2021)	Domestic	○			◎					
Hyunwoo Jeong et al. (2023)	Domestic		○			○		◎		
Sun Chang et al.(2023)	International	◎		◎	◎		○	◎		
Bourou et al.(2021)	International	◎			◎					
Sungbin An et al.(2023)	Domestic		◎			○		◎		◎
Espinosa et al.(2023)	International		◎	○	○					
Boris van et al.(2023)	International				○				○	○
Ricardo et al.(2022)	International	○				○				○
Matthieu et al.(2023)	International				○			○		
Claire et al.(2022)	International				○	○		○		
Sieun Kim et al. (2023)	**Domestic**	**Evaluation through Unified Procedure**								

다. 개인정보 노출 위험은 데이터 셋에서 개인의 식별이 가능한 정보가 무단으로 노출되거나 공개될 수 있는 위험이다. 데이터 유용성과 개인정보 노출 위험은 반비례 관계(trade-off)를 보인다.

먼저, 유용성 측면, 개인정보 노출 위험 측면으로 분류하고 유용성 평가를 광역 유용성, 특정 유용성으로 분류하였다. 둘째로, 관련 연구의 광역 유용성 평가 방법을 통계적 적합도 테스트, 거리 및 분포 기반 측정, 상관관계 및 기초통계량의 세 가지 범주로 묶어 분류하였다. 셋째로, 관련 연구의 특정 유용성 평가 방법을 분류 성능 평가, 회귀 분석 오차 평가, 기타 측정법의 세 가지 범주로 묶어 분류하였다. 마지막으로 관련 연구의 개인정보 노출 위험 평가 방법을 프라이버시 위험 평가, 거리 및 유사도 평가, 기타 측정법의 세 가지 범주로 묶어 분류하였다. 단일 측정법을 사용하였으면 ○ 기호로, 복수 측정법을 사용하면 ◎ 기호로 표기하였다.

광역 유용성을 평가하는 통계적 적합도 테스트 범주에는 KS-Test(Kolmogorov-Smirnov test), CS-TEST(Chi square test), Student t-test 측정법이 해당한다. 거리 및 분포 기반 측정 범주에는 pMSE, s-pMSE, Jenson-Shannon 발산, Wasserstein 거리, Hellinger 거리, Kullback - Leibler 발산 측정법이 해당한다. 상관관계 및 연관성 측정 범주에는 Pearson 상관계수, correlation ratio, Theil's U 계수, 쌍별 상관, Cramers'V, 기초통계량 등의 측정법이 해당한다.

특정 유용성을 평가하는 분류 성능 평가 범주에는 AUC, 재현율, 정확도, ROC, f1-점수 등이 포함된다. 회귀 분석 평가 범주에는 회귀계수, 신뢰 구간 중첩도 측정법이 포함된다.

마지막으로 개인정보 노출 위험 측면은 프라이버시 위험 평가, 거리 및 유사도 평가, 기타 측정법의 세 가지 범주로 묶어 분류하였다. 프라이버시 위험 평가 범

주에는 차분 프라이버시(Differential Privacy), WEAP(Within Equivalence Class Attribution Probability), TCAP(Targeted Corrected Attribution Probability), 신원 노출 위험도, 멤버십 공격 측정법이 포함된다. 거리 및 유사도 기반 평가 범주에는 DCR(Distance to Closest Record), NNDP(Nearest Neighbor Distance Point), Maximum mean discrepancy 측정법이 포함된다. 모델 및 알고리즘 기반 평가 범주에는 a-정밀도, b-재현율, 독창성 점수 측정법이 포함된다.

Fig. 2는 관련 연구 분석을 통한 도출한 문제점과 연구 범위를 도식화한 그림이다. 관련 연구 분석을 통해 도출한 문제점은 다음과 같다. 첫째, 합성데이터 평가 방법에 대한 명확한 기준이 없다. 합성데이터의 평가 방법에 대한 명확한 제도적 기준이 없고, 관련 연구들도 각기 다른 방식으로 평가를 진행하였다. 이는 합성데이터의 품질과 안전성에 대한 일관된 평가 기준의 부재를 의미하며, 이로 인해 합성데이터의 신뢰성과 효용성에 대한 객관적인 판단이 어려워진다.

둘째, 대부분의 연구에서 유용성 측면에 대해 평가가 이루어졌지만, 비교적 개인정보 노출 위험 측면에 대한 평가는 미흡했다. 합성데이터는 실제 개인정보를 바탕으로 생성되기 때문에 개인정보 노출의 위험이 분명히 존재한다. 이러한 위험을 효과적으로 관리하고 최소화하기 위해서는 합성데이터의 안전성을 체계적으로 평가하는 것이 필요하다.

셋째, 개인정보 노출 위험을 정량화한 연구가 부족하다. k-익명성(k-anonymity)과 같은 프라이버시 보호 모델은 데이터 셋의 속성을 식별자, 준 식별자, 민감정보 등으로 구분한다. 하지만, 어떤 정보가 민감하고 그렇지 않은지 명확한 구분이 어렵다. 또 어떤 변수가 공격자에 의해 식별될 수 있는지 알 수 없고 다양한 공격 유형에 대해 충분히 대비할 수 없다. 차분 프라이버시 기술 또한 생성모델 훈련을 통해 생성

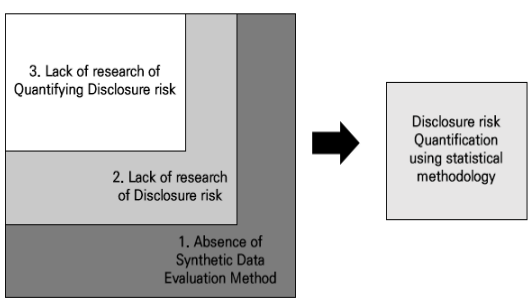

Fig. 2. Schematic of Research Scope

되는 합성데이터의 개인정보 노출 위험의 크기를 미리 정량화할 수 없고 사후적으로 생성된 데이터의 개인정보 노출 위험을 측정해야 하는 한계점을 갖는다.

본 연구는 통계적 방법론을 통해 개인정보 노출 위험 정량화하여 특정 유형의 공격에 국한하지 않는 포괄적인 위험 평가를 진행하고자 한다. 기존 연구 분석을 통해 도출한 세가지 문제점을 보완하고 단일 방법론에 따라 개인정보 노출 위험을 정량화하는 동시에 합성데이터가 가지는 유용성을 측정한다. 합성 데이터 유용성은 훈련데이터 공격과 평가데이터 공격의 개인정보 노출 위험 차이를 통해 측정한다.

III. 합성데이터 개인정보 노출 위험 정량화 방법론

합성데이터를 만드는 과정은 외부자에게는 블랙박스 형태이다. 즉, 합성데이터를 소유한 공격자도 이 메커니즘에 직접 접근하거나 쿼리(질의) 할 수 없다. 다시 말해, 합성데이터는 어떻게 만들어지는지 외부에서는 정확히 알 수 없는 상태로 유지된다. 해당 방법론은 개인정보 보호 위험을 보수적으로 평가하기 위해 공격자가 합성데이터 세트를 완전히 소유한 상황을 가장 강력한 위협으로 본다. 또한 공격자가 공격 대상인 원본 레코드 일부에 대해 부분적이지만 정확한 추가 정보를 보유하고 있다고 가정한다.

우선 원본 데이터 셋(X_{ORI})은 모집단에서 독립적으로 추출된 N개의 레코드로 이루어져 있다고 가정한다. 무작위 추출을 통해 원본 데이터 대표성을 유지할 수 있고 전반적인 평가가 가능하게 된다. 원본 데이터 세트를 훈련데이터(X_{train})와 평가데이터(X_{test})로 분할한다. 이때 훈련데이터와 평가데이터는 상호 배타적이다. 훈련데이터는 생성모델을 통해 합성데이터(X_{syn})를 합성한다. 훈련데이터, 평가데이터, 합성데이터는 동일 수의 속성을 가지고 있으며, 레코드 수는 상이하다.

본 연구의 합성데이터 노출 위험 평가 정량화 방법론은 세 단계로 구성된다. Fig. 3은 본 연구의 통계적 방법론 절차를 도식화한 그림이다. 첫 번째는 개인정보 공격 실행 단계, 두 번째는 공격 성공 여부를 측정하는 공격 평가 단계, 세 번째는 개인정보 노출 위험을 정량화하는 위험 측정 단계로 구성된다.

공격 실행 단계는 세 가지 다른 공격을 실행한다. 첫째, '무작위 대입 공격'을 실행한다. 무작위 추측을 생성하여 훈련데이터 공격력을 측정할 수 있는 기준

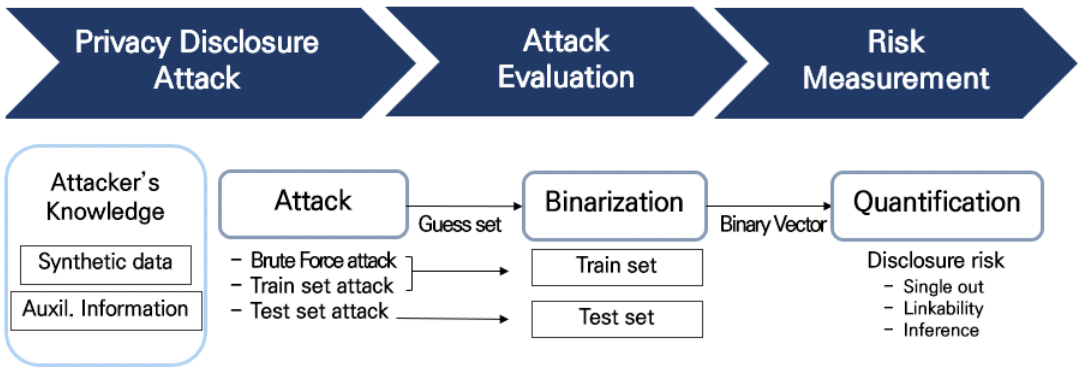

Fig. 3. Methodology Process of Measuring Disclosure Risk

이 된다. 둘째, 공격자가 합성데이터를 사용하여 훈련데이터 레코드의 개인정보를 추측하는 공격인 '훈련데이터 공격'을 실행한다. 마지막으로, 원본 데이터 레코드(즉, 특정 정보)의 구체적인 개인정보 위험을 전체 모집단에 고유한 일반 위험(즉, 일반 정보)과 구별하기 위해 실행하기 위해 평가데이터에서 실행하는 '평가데이터 공격'을 실행한다. 위 세 가지 공격은 공격 대상 원본 레코드에 대해 N_A개 추측 집합 $g=\{g_1,\cdots,g_{N_A}\}$을 만든다. 예시로 "원본 데이터 세트에 60세 남성이고 32315 지역에 거주하는 사람이 단 한 명 존재한다."와 같은 추측을 할 수 있다.

'무작위 대입 공격'은 데이터 셋의 속성 범위를 알기 위해 진행하며, 합성데이터만 사용해 무작위로 추측한다. '훈련데이터 공격'과 '평가데이터 공격'은 합성데이터와 추가 정보를 활용해 추측을 생성한다. 둘 다 동일 공격 알고리즘을 사용한다. 합성데이터와 평가데이터는 독립이다. 따라서 공격자가 평가데이터 레코드에 대한 공격이 성공하면 이는 일부 평가 레코드에 대한 것이라기보다 모집단에 공통된 패턴 및 상관관계 때문이다. 따라서 두 공격 성공률 차이는 훈련데이터를 사용해서 생성기를 학습시킬 때 발생하는 개인정보 노출 위험으로 볼 수 있다. 즉, 평가데이터보다 훈련데이터에 대한 추측 성공률이 높으면 합성데이터 생성하면서 발생하는 개인정보 노출로 보고 이를 측정한다.

공격 평가 단계에서는 앞 공격 실행 단계에서 생성한 추측을 원본 데이터와 비교한다. 이때 원본 데이터 참값과 비교하여 같다면 공격 성공으로 판단하고, 상이하다면 공격 실패로 판단한다. 이 단계의 결괏값은 $y=\{y_1,\cdots,y_{N_A}\}$ 비트로 구성된 벡터이다. 여기서 i번째 추측 g_i에 대해 공격 성공이면 $y_i=1$이고, 공격 실패 시 $y_i=0$ 이다.

위험 추정 단계는 공격 종류별로 다르게 진행된다. 각 공격의 결과 y_i가 다른 공격과 독립적이라는 가정하에 y는 베르누이 시행으로 모델링 될 수 있다. 개인정보 노출 위험 $\hat{p}$을 공격자가 공격에 성공할 확률로 정의한다. 베르누이 시행의 결과는 성공 확률에 대한 신뢰 구간인 이항 비율 신뢰 구간을 사용하여 나타낼 수 있다. 여기서 실제 공격 성공률 $\hat{p}$의 추정치 p와 신뢰도 수준 α를 고려한 신뢰 구간 $\hat{p} \in p \pm \delta_p$은 윌슨 점수 구간(Wilson Score Interval)을 통해 추정한다. 윌슨 점수 구간은 비율에 대한 신뢰 구간을 계산할 때 중앙값을 조정하고, 이진 데이터의 변동성을 더 잘 반영하여 비율의 분산을 더 정확하게 추정한다. '무작위 대입 공격', '훈련데이터 공격', '평가데이터 공격'에 대한 공격 성공률을 $p_{rand} \pm \delta_{rand}$, $p_{train} \pm \delta_{train}$, $p_{test} \pm \delta_{test}$ 형태로 평가한다.

'무작위 대입 공격'을 기준선으로 설정하여 공격의 강도는 '훈련데이터 공격' 성공률과 '무작위 대입 공격'의 성공률의 차로 정의한다. 공격이 기준선인 '무작위 대입 공격'보다 약한 경우, 공격 실패로 본다. '평가데이터 공격'의 성공률은 '훈련데이터 공격'의 성공률과 동일 평가방식으로 평가한다. 학습된 레코드인 합성데이터가 평가데이터보다 더 많은 정보가 포함된 경우 $p_{train} \geq p_{test}$이다. 훈련데이터 공격과 평가데이터 공격의 성공률로 특정 개인정보 위험은 다음과 같이 유도된다.

수식 (1)의 분자는 훈련데이터 레코드가 공격 대상이었을 때 공격 성공의 수가 평가데이터 레코드가 공격 대상이었을 때 공격 성공의 수를 초과하였을 때

해당한다. 분모는 모든 평가데이터 셋에 대한 공격이 성공했을 때 ($p=1$) 최댓값을 나타낸다. 분모는 정규화 인자로 작용해서 분자의 차이를 맥락 화하는 데 도움이 된다. 일반 정보는 익명 데이터의 유용성을 제공하고, 특정 정보를 통해 공격자가 일부 개인의 프라이버시를 침해할 수 있다.[16] 합성데이터의 개인정보 노출 위험이 존재하는 이유는 데이터의 유용성 때문이다. 예를 들어, 100개 추측 세트에서 훈련데이터에 대해 90개 추측 성공, 평가데이터에 대해 80개 성공을 했다고 했을 때 각각 공격 성공률은 0.9, 0.8이다. 90개의 성공 추측 중 80개는 데이터 셋이 가지는 유용성 때문에 성공했다고 볼 수 있다. 나머지 10개 성공 추측은 개인정보 노출로 볼 수 있다. 평가데이터의 공격 성공률 0.8은 100개 추측이 모두 성공한 경우($p=1$)일 때 보다 0.2 높고 훈련데이터에 대한 공격 성공률보다 0.1 높다. 이는 개인정보 노출 위험 0.5로 계산된다.

$$\text{개인정보노출위험}(P)=\frac{p_{train}-p_{test}}{1-p_{test}} \quad (1)$$

훈련데이터와 평가데이터를 목표로 한 공격 성공률이 똑같다면, 공격자가 합성데이터에 접근하여 훈련데이터에 대한 정보를 얻는데 아무런 이득이 없다. 공격 성공은 합성데이터의 일반적인 유용성으로 설명된다. 하지만 훈련데이터의 성공률이 평가데이터의 성공률을 초과하면 이는 합성데이터에서 개인정보가 노출되었고 본다.

IV. 공격 유형에 따른 위험 측정 방법

이번 장에서는 2.2의 국내외 비식별화 조치 관련 지침 분석을 통해 도출한 익명 데이터가 가질 수 있는 개인정보 노출 위험인 개별화, 연결가능성, 추론 위험에 대한 구체적 적용 방법을 보인다. 위험별로 공격 실행 단계, 공격 평가 단계는 다르게 진행하고 위험 평가 단계는 동일하게 진행한다.

4.1 개별화 위험 측정

개별화(Single out) 공격은 합성데이터를 기반으로 훈련데이터의 한 사람을 개별화할 수 있도록 추측을 생성해 내는 공격이다. 직관적으로 합성데이터에서 드물게 나타나거나 독특한 값을 갖는 속성은 원본 데이터에서도 동일하게 드물게 나타나거나 독특한 값을 갖는 속성일 확률일 것이다. 즉, 합성데이터에 접근하는 것은 아무런 정보 없이 추측하는 것보다 더 의미 있는 추측을 생성해 낼 수 있다.

공격 실행 단계에서는 Single out 추측은 단일 속성 조건자, 복합 속성 조건자의 두 가지로 나눠 생성한다. 단일 속성 조건자는 단일 속성을 사용하여 조건자(Predicate)를 생성하고, 복합 속성 조건자는 여러 속성을 조합하여 조건자를 생성한다. 합성데이터의 모든 고유 속성값들이 조건자가 될 수 있다.

단일 속성 조건자는 다음과 같은 방식으로 조건자를 생성한다. 범주형 속성 또는 결측값의 경우 데이터 셋에 한 번만 나타나는 속성값을 사용하여 조건자를 생성한다. 수치형 속성의 경우 해당 속성의 최댓값과 최솟값을 사용하여 최솟값보다 작은 값, 최댓값보다 큰 값을 기준으로 조건자를 생성한다. 즉, 이상값(Outlier)을 활용해 조건자를 생성한다. 공격 실행 단계에서 생성된 조건자 집합에서 N_A개를 무작위로 선택하여 추측으로 사용한다.

복합 속성 조건자는 합성데이터에서 무작위로 선택된 레코드 의 단일 속성 조건자의 결합으로 생성한다. 무작위로 선택된 레코드 $\tilde{x}$는 속성$\{a_1,\cdots,a_d\}$으로 구성되어 있다. 조건자는 논리연산자 AND로 결합하여 사용한다. a_i가 범주형 속성 또는 결측값인 경우 $\tilde{x}[a_i]=a_i$를 조건자로 사용한다. a_i가 수치형 속성의 경우, $\tilde{x}[a_i]$가 $X_{syn}[a_i]$의 중앙값보다 '크거나 작다', '작거나 크다'로 표현한다. 등호의 부호는 a_i가 $X_{syn}[a_i]$의 중앙값보다 크거나 같은지에 따라 결정하여 조건자의 공격 성공률을 높인다. 공격 실행 단계에서 합성데이터에서 이러한 각 조건자를 평가해 X_{syn}의 단일 레코드에만 존재할 때만 추측 집합에 추가한다. 합성데이터에서 무작위로 추출한 레코드와 속성을 사용해 합성데이터에 과적합하지 않는 모든 파라미터 공간에서 공격 추측을 생성한다. 단일 속성 조건자 생성과 복합 속성 조건자 생성뿐만 아니라 무작위 추측 공격의 결과는 N_A개 조건자 집합이 된다. 원본 데이터의 Single out 조건자들을 잘 나타내는지 평가하기 위해 원본 데이터에서도 평가된다. 공격 성공 여부 결과는 다음과 같다.

$$y_i=\begin{cases}1\ g_i\ is\ correct\\ 0\ g_i\ otherwise\end{cases} \quad (2)$$

4.2 연결가능성 위험 측정

연결가능성(Linkability)는 데이터 A,B가 원본 데이터 레코드 집합의 특성들을 가지고 있고 합성데이터도 이 특성들을 가지고 있다고 가정한다. Linkability 공격 대상 레코드는 에서 무작위로 N_A개 추출한 원본 레코드의 모음 T이다. 공격자는 데이터 A와 B의 속성값인 $T[:,A]$, $T[:,B]$에 대해 알고 있다고 가정한다. 공격자는 $T[:,A]$인 레코드와 $T[:,B]$ 레코드와 정확하게 연결을 실행한다.

공격을 위해 공격자는 $T[:,A]$의 모든 레코드에 대해 $X_{syn}[:,A]$에서 k번째 가장 가까운 합성 레코드를 찾는다. 결과 인덱스들은 $I^A=(I_i^A,...,I_{N_A}^A)$이다. I_i^A는 특성 세트 A의 부분 공간에서 i번째 대상의 가장 가까운 이웃인 k합성 레코드의 인덱스 세트이다. 레코드 간의 거리를 측정하기 위해 Gower 계수를 사용한다.

범주형 속성의 경우 일치하는 경우 1이고 그렇지 않으면 0이다. 수치형 속성의 경우 거리는 L_1정규식 $|x_i-x_j|\le 1(\forall x_i,x_j\in x)$을 사용한다. 마찬가지로 평가데이터에서 가져온 공격 대상 레코드에 합성데이터를 사용하여 공격을 진행한다. 결과 인덱스는 I_{test}^A, I_{test}^B이다. 무작위 대입 공격을 이용해 우연히 링크를 찾을 확률을 측정한다. I_{rand}^A, I_{rand}^B는 $[0,n_{syn}-1]$ 범위에서 무작위로 균등하게 얻는다. 가까운 이웃 집합들이 같은 합성데이터 레코드를 가리키고 있는지 확인한다. 합성데이터 레코드는 공격자가 원본 데이터에서 대상 개인에 대한 이전에 연결되지 않은 정보를 함께 연결할 수 있게 한다. 연결이 성공하면 공격 성공으로 볼 수 있다. 이때 공격 성공 여부 결과는 다음과 같다.

$$y_i(I_i^A,I_i^B)=\begin{cases}1\text{ if }I_i^A\cap I_i^B\neq\varnothing\\0\ otherwise\end{cases} \tag{3}$$

4.3 추론 위험 측정

추론(Inference) 위험은 공격자가 공격 대상이 되는 원본 레코드의 속성(추가 정보)을 알고 있다고 가정한다. 공격자는 공격 대상이 되는 레코드의 비밀 속성(민감정보)을 정확하게 추론해 내는 것을 실행한다. 추론 공격은 근접 이웃 탐색(Nearest Neighbor search)이 핵심이다. 공격 대상 레코드에 대해 공격자는 추가 정보의 속성에 의해 정의된 하위 공간에서 가장 가까운 합성 레코드를 찾는다. 근접 이웃 알고리즘은 Linkability 위험 측정과 동일하게 진행된다. 가장 가까운 합성데이터 레코드의 비밀 속성값이 공격자의 추측으로 사용된다. 그런 다음 평가데이터의 대상에 대해서도 진행한다. 마지막으로, 우연히 추론 공격에 성공했을 확률을 측정하기 위해 비밀 속성값의 가능한 값들을 랜덤하게 사용해 얻은 추측들로 무작위 대입 공격을 진행한다.

주어진 비밀속성에 대해 공격자의 추측이 올바른 경우 공격자가 성공적인 추론을 했다고 할 수 있다. 추측한 내용을 원본 데이터에 있는 비밀의 실제 값과 비교하여 공격자가 올바른 추론을 한 횟수를 계산한다. 비밀속성 s_i이 범주형 변수일 때 정확한 값을 복구해야 공격 성공이다. 수치형 변수일 경우 추측이 실제 값에서 구성 가능한 허용 오차 δ 내에 있으면 추론이 정확하다고 평가한다. 훈련데이터 공격과 평가데이터 공격에서 똑같은 허용 오차 δ를 사용한다. 공격 성공 여부 결과는 다음과 같다.

$$y_i(s_i,g_i,\delta)=\begin{cases}\dfrac{|s_i-g_i|}{s}\le\delta\ i\ numerical\\1\qquad\qquad i\ categorical, s_{1=}g_i\\0\qquad\qquad i\ categorical, s_{1\neq}g_i\end{cases} \tag{4}$$

V. 결 론

본 연구에서는 합성데이터 개인정보 노출 위험 측정 정량화 방법론과 그에 따른 절차를 설계하고 위험 유형에 따른 구체적인 적용 방법을 보인다. 이때, 유럽 GDPR 정책 자문기구인 29조 작업 처리반이 규정한 익명 데이터가 가질 수 있는 세 가지 위험(Single out, Linkability, Inference)에 대한 구체적 적용 방법을 보였다. 본 연구의 개인정보 노출 위험 정량화를 위한 통계적 방법론은 위 세 가지 위험뿐만 아니라 다른 위험 평가에도 적용할 수 있다.

현재까지 합성데이터 평가 방법에 대한 국제적이고 통일된 제도적 기준은 존재하지 않는다. 합성데이터의 사용과 평가는 상대적으로 새로운 분야이며, 이 분야에서의 표준화 노력은 아직 초기 단계이다. 인공지능 학습용 데이터 품질검증 방법 등 TTA 표준이 승인되었으나, 데이터의 정확성과 유효성 검증에 대한 표준이며, 개인정보 노출 위험도는 다루지 않는다. 합성데이터 분야의 발전과 함께 제도적인 표준의 필

요성은 점차 증가하고 있다. 이는 합성데이터의 신뢰성을 보장하고, 데이터 관리자들에게 명확한 지침을 제공하는 데 중요하다. 따라서, 향후 합성데이터 평가를 위한 국제적인 표준이 개발될 가능성이 크며, 이는 데이터 과학과 인공지능 분야에서 중요한 발전이 될 것이다.

본 방법론은 기업의 데이터 관리자가 개인정보가 포함된 데이터를 제공하는 상황, 다른 기업의 데이터와 결합하는 상황, 인공지능 학습용 데이터로 사용하는 상황에서 개인정보 노출 위험을 효과적으로 제어하는 데에 도움이 된다. 이 방법론의 장점은 다음과 같다. 첫째, 포괄적 위험 평가가 가능하다는 것이다. Single out, Linkability, Inference 등 다양한 개인정보 노출 위험을 평가함으로써, 합성데이터가 개인정보 보호 규정을 준수하면서도 연구 목적에 적합한지 종합적으로 판단할 수 있다. 둘째, 합성데이터의 신뢰성을 보장한다. 데이터 관리자는 실제 데이터의 통계적 특성을 잘 반영하면서도 개인정보 보호 기준을 충족하는지 판단할 수 있다. 데이터 유형과 상황에 맞춰 개인정보 보호 수준을 조정할 수 있다.

본 연구는 합성데이터의 개인정보 노출 위험을 정량화하는 방법론을 제시함으로써, 실무적으로 기업 및 연구 기관이 개인정보 보호 규정 준수를 위한 내부 지침 마련과 개인정보 노출 위험 평가 도구의 개발하도록 독려한다. 또한, 합성데이터의 신뢰성과 안전성을 높이는 방안을 모색하고, 데이터의 유효성을 강화하는 새로운 통계적 접근 방식을 제안한다. 학술적으로는 데이터 유형과 상황에 맞는 위험 평가 방법론의 발전을 촉진하며, 합성데이터 평가의 정밀성을 높이는 데 중점을 둔다. 이는 합성데이터 사용의 효율성과 안전성을 보장하는 데 기여한다.

향후 연구로는 개인정보 보호 위험의 정량화 단계에서 명확한 '위험 기준점'에 대한 연구를 진행할 예정이다. 위험 수치가 특정 기준 이상일 때 '위험이 크다'라고 판단할 수 있는 구체적인 임곗값이 설정되지 않아 데이터 관리자가 언제 추가적인 보호 조치를 취해야 하는지 명확히 판단하기 어려울 수 있다. 이러한 명확한 기준의 부재는 법적 기준과 일치성 확보, 다양한 적용 맥락에서의 위험 수준 평가, 데이터 보호 규정 및 요구사항에 부합하는 정교한 위험 평가 방법 개발에 있어 중요하다. 따라서, 이 분야의 추가 연구를 진행할 예정이다.

References

[1] Office for Government Policy Coordination, Ministry of Government Administration and Home Affairs, Korea Communications Commission, Ministry of Science, ICT and Future Planning, "Guideline for Personal Information De-identification Measures", Jun. 2016.

[2] Personal Information Protection Commission, "Pseudonymous information processing guidelines", Apr. 2022.

[3] The European Parliament and of the Council "The protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)", Regulation (EU) 2016/679, Apr. 2016.

[4] Article 29 Working Party, "A critical appraisal of the Article 29 Working Party Opinion 05/2014 on data anonymization techniques", International Data Privacy Law, vol 5, no.1, pp.73-87, Jan. 2014.

[5] ISO/IEC 20889:2018, "Privacy enhancing data de-identification terminology and classification of techniques", International Standard under systematic review, pp. 46, Nov. 2018.

[6] Hong Dong-Suk and Cheol Baik. "Generating and Validating Synthetic Training Data for Predicting Bankruptcy of Individual Businesses.", Journal of Information & Communication Convergence Engineering, 19(4), pp.228-233, Dec. 2021.

[7] Hyunwoo Jung, Younsang Cho, Geonwoo Ko, Jae-ik Song, and Donghyeon Yu, "Comparison study of synthetic data generation methods for credit card transaction data.", Journal of the Korean Data And Information

Science Society, 34(1), pp. 49-72, Jan. 2023.

〔8〕 Sun, Chang, Johan van Soest, and Michel Dumontier. "Generating synthetic personal health data using conditional generative adversarial networks combining with differential privacy.", Journal of Biomedical Informatics, vol. 143, no. C, pp. 104404., Jul. 2023.

〔9〕 Seongbin An, et al. "A comparison of synthetic data approaches using utility and disclosure risk measures.", The Korean Journal of Applied Statistics, 36(2), pp. 141 - 166, Apr. 2023.

〔10〕 Bourou, Stavroula, et al. "A review of tabular data synthesis using GANs on an IDS dataset.", MDPI Information, vol.12, no. 09, pp.375, Sep. 2021.

〔11〕 Espinosa, Erica, and Alvaro Figueira. "On the Quality of Synthetic Generated Tabular Data.", MDPI Mathematics, vol. 11, no. 15, pp. 3278, Jul. 2023.

〔12〕 Van Breugel, Boris, et al. "Membership inference attacks against synthetic data through overfitting detection.", Proceedings of the 26th International Conference on Artificial Intelligence and Statistics(AISTATS), Spain, pp. 3493-3514, Apr. 2023.

〔13〕 Ricardo Muñoz-Cancino, Cristián Bravo, Sebastián A. Ríos, and Manuel Graña, "Assessment of Creditworthiness Models Privacy-Preserving Training withSynthetic Data.", Hybrid Artificial Intelligent Systems: 17th International Conference(HAIS 2022), Spain, pp. 60 - 72, Sep. 2022.

〔14〕 Matthieu Meeus, Florent Guepin, Ana-Maria Creţu, and Yves-Alexandre de Montjoye, "Achilles' Heels: Vulnerable Record Identification in Synthetic Data Publishing.", Computer Security - ESORICS 2023: 28th European Symposium on Research in Computer Security, Netherlands, pp. 182 - 198, Sep. 2023.

〔15〕 Little, Claire et al. "Comparing the Utility and Disclosure Risk of Synthetic Data with Samples of Microdata.", Privacy in Statistical Databases:International Conference(PSD 2022), France, pp. 234 - 249, Sep. 2022.

〔16〕 Dwork, Cynthia, and Aaron Roth. "The algorithmic foundations of differential privacy.", Foundations and Trends® in Theoretical Computer Science, vol. 9. no. 3-4, pp. 211 - 407, Aug. 2014.

〈저자소개〉

김 시 은 (Sieun Kim) 학생회원
2022년 2월: 한동대학교 ICT융합학과, 수학통계학과 졸업
2024년 2월: 전남대학교 데이터사이언스대학원 석사 졸업
〈관심분야〉 데이터사이언스, AI, 프라이버시 강화기술, 데이터 프라이버시, 합성데이터

엄 익 채 (Ieck-chae Euom) 종신회원
2003년 8월: 전남대학교 컴퓨터정보학부 학사 졸업
2015년 2월: 한국과학기술원 소프트웨어대학원 석사 졸업
2019년 2월: 전남대학교 정보보안협동과정 박사 졸업
2019년 10월~현재: 전남대학교 시스템보안연구센터 소장, 데이터사이언스대학원 교수
〈관심분야〉 제어시스템보안, 스마트그리드 보안, 원자력 보안, 취약점 분석, 차세대인프라 보안, 스마트시티·공장 보안, AI기반 이상징후 탐지, 지능형 보안