2025 Vol.2

PRIVACY REPORT

개인정보 심층분석

합성데이터 기술

KISA 한국인터넷진흥원

PRIVACY REPORT

2025년
개인정보
심층분석

CONTENTS

합성데이터 생성 기술 동향 및 발전 방향

김승환
인하대학교 데이터사이언스학과 I 교수

1. 서론

기업 또는 단체가 보유한 개인정보를 안전하게 활용하기 위해 가명 혹은 익명 데이터로 처리하여 이용한다. 익명 데이터는 개인정보 보호법에 저촉받지 않아 기간, 공간으로부터 자유롭게 분석을 수행할 수 있는 장점이 있지만 익명화 단계에서 정보손실이 크기 때문에 유용성이 떨어지는 문제를 가지고 있다.

잘 생성된 합성데이터는 가상의 데이터이기 때문에 기본적으로 익명성을 만족하지만, 모든 합성데이터가 익명성을 만족하는 것은 아니다. 데이터를 학습하는 과정에서 과적합으로 인해 특정 레코드를 거의 유사하게 재현하거나 레코드 단위 변조를 통해 합성데이터를 생성한 경우, 익명성을 만족하는 합성데이터로 인정받을 수 없다.

합성데이터가 생성된 후에는 어떤 과정을 통해 만들어졌는지 알 수 없으므로 원본 데이터로부터 합성데이터가 생성되는 전 과정을 평가하여 익명성을 증명해야 한다. 또한, 익명성은 충분히 만족하지만, 유용성이 떨어지는 경우 합성데이터 활용 목적을 달성할 수 없으므로 유용성과 안전성을 동시에 만족하도록 합성데이터를 생성하는 것이 중요하다.

이 글에서는 합성데이터 생성 방법론의 발전 과정을 살펴보고 안전하고 유용한 합성데이터를 생성하는 기술에 대해 고찰하고자 한다.

2. 합성데이터 생성 기술 발전

합성데이터 생성 기술은 데이터 형태에 따라 달라진다. 데이터 형태는 정형, 비정형(이미지, 텍스트, 소리 등) 형태가 있을 수 있고, 레코드 간 독립 여부에 따라 생성 기술이 나뉜다.

정형 데이터의 합성데이터 생성 방법으로 통계 모형 기반, 딥러닝 기반 생성 기술이 이용되는데 통계 모형 기반 생성 기술을 좀 더 많이 이용한다. 통계 모형 기반 생성 기술은 원본 데이터의 다차원 분포를 추정하고 추정된 분포에서 난수 혹은 샘플링 기법을 이용해 합성데이터를 생성하는 방법으로

속도가 빠르고 유용성과 안전성을 동시에 만족하는 합성데이터를 비교적 쉽게 만들 수 있는 장점이 있다. 하지만, 비정형 데이터 생성은 거의 불가능하다.

딥러닝 기반 생성 기술은 임의로 합성데이터의 분포 생성하고 이를 원본 데이터의 분포 특성에 반복 근사하는 알고리즘을 이용한다. 적절한 딥러닝 모형과 충분한 데이터, 학습 시간이 주어질 경우, 원본 데이터의 분포 특성과 유사한 합성데이터를 생성할 수 있다. 현실적으로는 딥러닝 모형 선택과 모형의 조절 모수 등에 따라 유용성과 안전성이 크게 변화되는 문제점이 있어 기술적으로 어려운 측면이 있다. 하지만 이미지와 같은 비정형 데이터 생성에서 뛰어난 성능을 보이기 때문에 비정형 합성데이터 생성에 많이 이용된다.

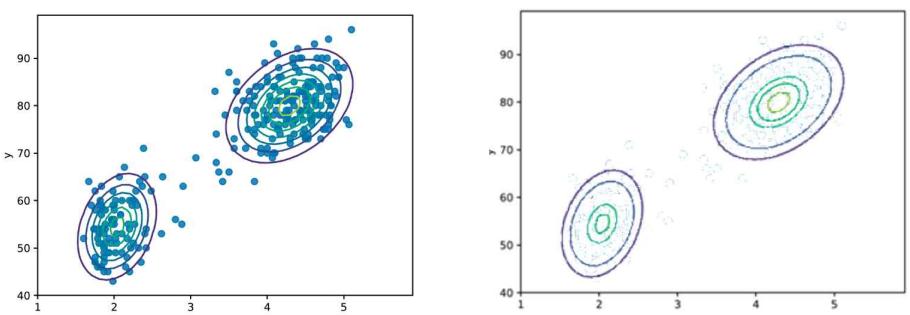
① 통계 기반 생성 기술

통계 모형 기반 기술은 원본 데이터의 다차원 분포를 추정하고 해당 분포로부터 가상의 데이터를 생성하는 기법으로 정형 데이터일 경우, 생성 시간이 짧고 비교적 유용성이 높은 데이터를 생성할 수 있다. 반면, 이미지와 같은 비정형 데이터를 통계 모형 기반 생성 기술로 생성하는 것은 거의 불가능하여 이용하지 않는다.

합성데이터 시초는 설문 자료 등에서 발생하는 결측값을 대체하기 위하여 하버드대의 통계학과 교수였던 Donald B.Rubin 교수가 1981년에 제시한 다중 대체법(Multiple Imputation)에 근간을 두고 있다. 다중 대체법은 결측된 변수의 값에 대한 예측 모형을 세우고 결측값을 여러 개의 예측값으로 대체하여 분석하는 기법이다.

Rubin 교수는 1993년에 자신이 제안한 다중대체법을 이용해서 합성데이터를 생성하는 방법을 최초로 제안하였다.[1] 합성데이터 생성 목적과 별개로 정규분포, 베타분포, 감마분포 등 알려진 확률분포의 매개변수를 추정하여 유사한 특성을 가진 난수 데이터를 생성하는 방식은 통계학에서 매우 흔한 기법이다.

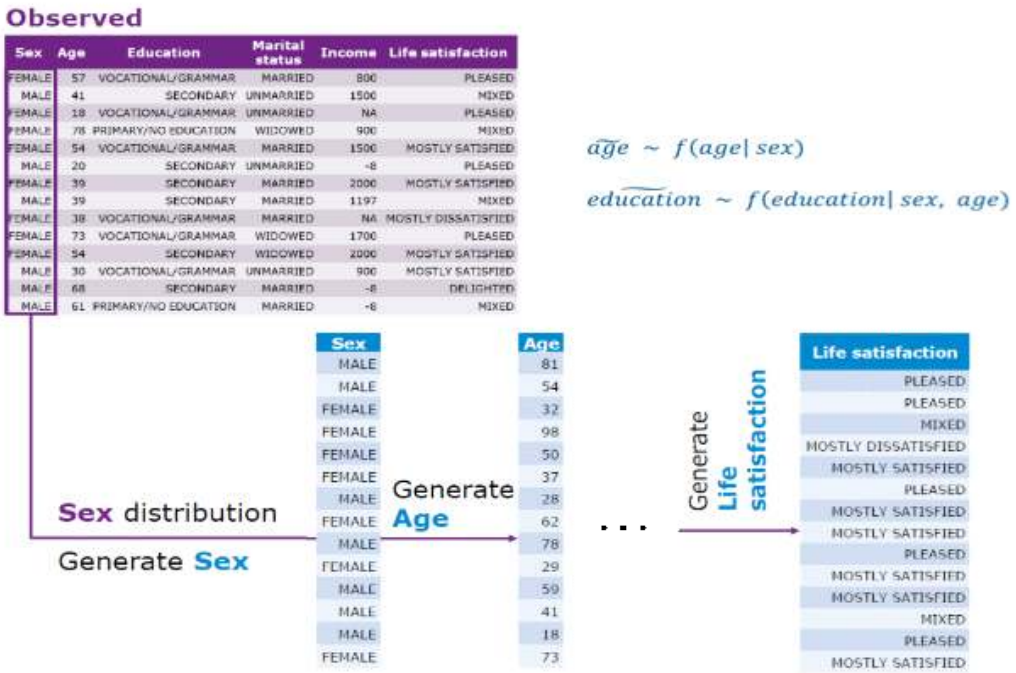
그림1 다차원 데이터의 분포 추정



1) Donald B. Rubin, Statistical Disclosure Limitation, (1993.), p.461-468.

또한, 다차원 분포를 재현하기 위해 베이지안, MCMC(Markov Chained Monte Carlo) 샘플링 기법으로 난수를 발생하여 다차원 분포의 난수 발생이 가능하다는 점을 이용해 합성데이터를 생성할 수 있다. 복잡한 분포 모양을 학습하기 위해 하나의 분포가 아닌 여러 분포를 혼합하는 Mixture 모형, 특정 분포로 가정하지 않고 비모수적으로 Kernel Density를 추정하는 방법이 합성데이터 생성에 응용되면서 유용성 높은 합성데이터를 생성할 수 있게 되었다.

그림2 Synthpop 조건부 표본 추출 과정2)



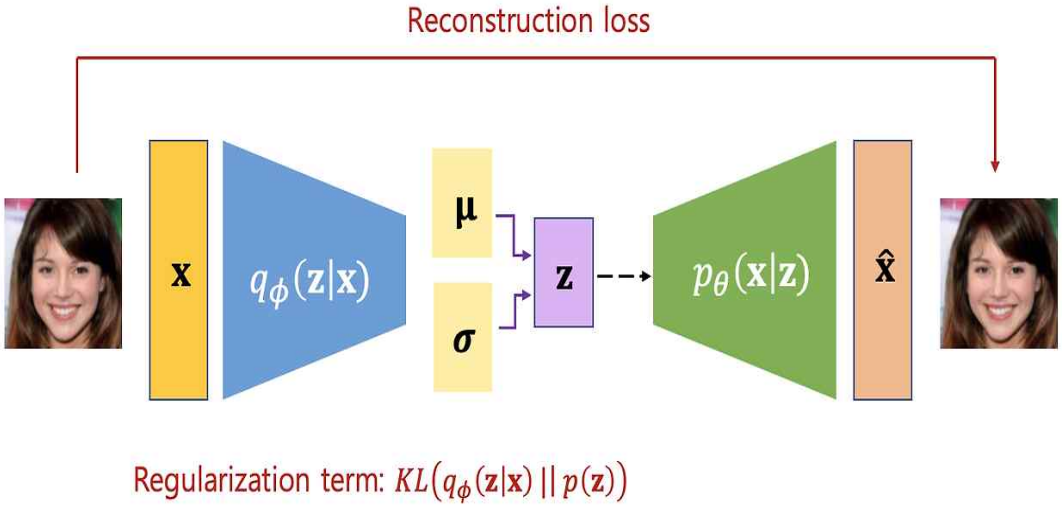
이러한 방법론은 모두 레코드 간 독립을 가정한 방법으로, 시계열 자료와 같이 레코드가 종속된 경우는 종속성을 제거하는 전처리를 하거나 적절한 시계열 모형으로 적합 후 난수를 통해 생성하는 모형을 이용해야 한다.

② 딥러닝 기반 생성 기술

딥러닝 기반 생성 기술은 VAE(Variational Auto Encoder), GAN(Generative Adversarial Network)이 시초로 볼 수 있다. VAE는 오토 인코더를 변형하여 인코더와 분포 매개변수 잠재 공간을 연결하고 그 결과를 디코더로 보내 이미지를 재구성하는 아이디어다. 시도 자체는 의미 있었으나 해상도가 떨어지고 세부적인 표현에 한계가 있어 현실적으로 활용되지 않지만, 이후 딥러닝 기반 생성 기술에 기초가 되는 모형이다.

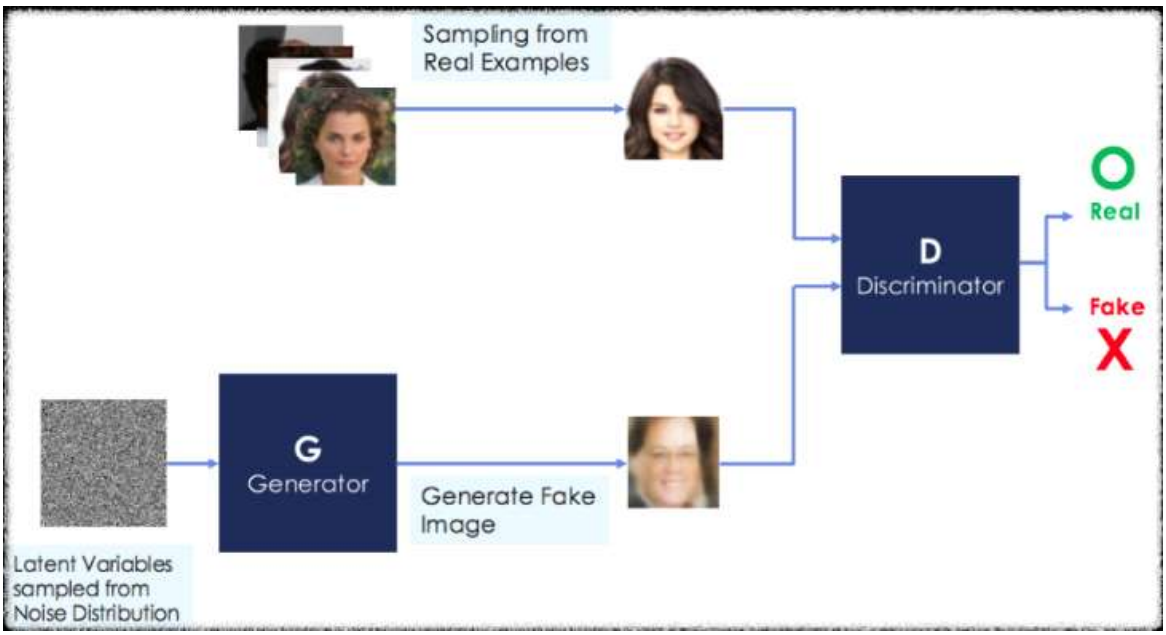
2) Beata Nowok, Introduction to synthpop, Administrative Data Research Centre – Scotland

그림3 VAE 개념도[3)]



GAN 기본 모형은 VAE보다 개선된 결과를 제공하지만, 여전히 유사한 문제점을 가지고 있다. 이후, DCGAN, ProGAN, StyleGAN 등으로 고해상도 이미지 생성이 가능하도록 발전되어 StyleGAN은 합성데이터 프로젝트에 많이 이용되는 방법론이다.

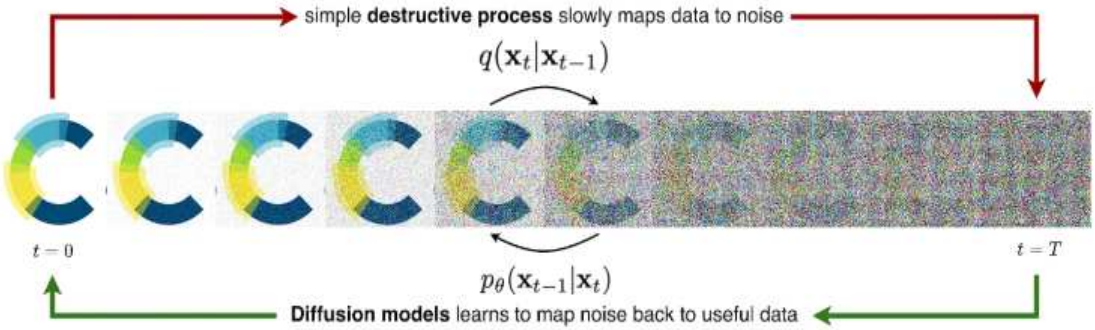
그림4 GAN 원리[4)]



이후, Diffusion 계열의 Stable Diffusion 모형이 StyleGAN을 능가하는 성능을 보여 많이 이용되고 있다.

3) DeepCampus [VAE], beta-VAE, (2023.05.)
4) 컴케빈의 IT 전문자식창고, "AN 실제 이미지와 구별할 수 없는 가짜 이미지 생성하는 생성적 적대 신경망 개념과 동작원리 및 유형 이해, (2024.07.04.)

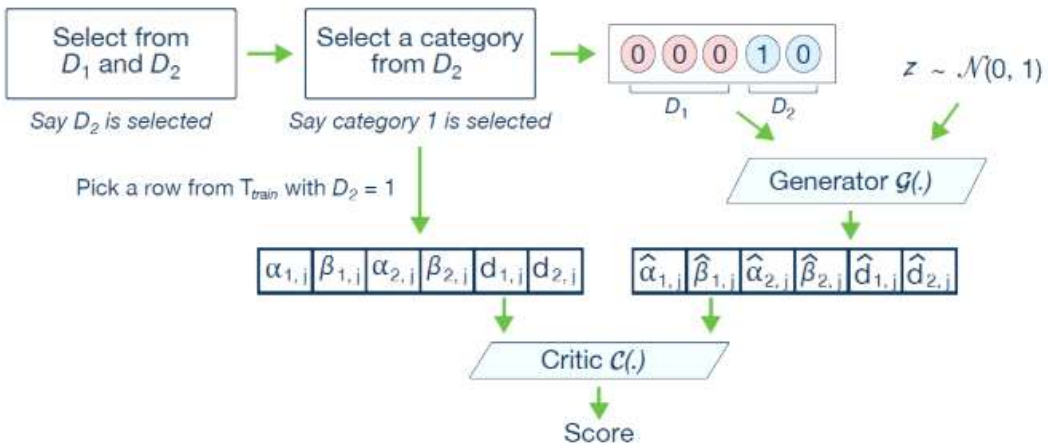
그림5 Diffusion 모형 원리[5)]



정형 데이터를 딥러닝 기법으로 합성하는 시도도 많이 발전하고 있다. 초기에는 정형 데이터가 수치형, 범주형으로 이루어지는 복잡한 구조를 고려하지 못하여 유용성을 만족하지 못하다가 통계 모형 기반 생성 방법의 조건부 확률분포 개념을 도입하면서 어느 정도 유용성을 확보하는 생성 기술인 CTGAN[6)] 알고리즘들이 발표되었다.

즉, 컬럼의 범주형 변수 선택에 따라 조건부 알고리즘을 채택하였다. 하지만, 이들 알고리즘도 통계 모형 기반 생성 기술에 비해 유용성이 좋지 못해 적극적으로 활용되지는 않았다.

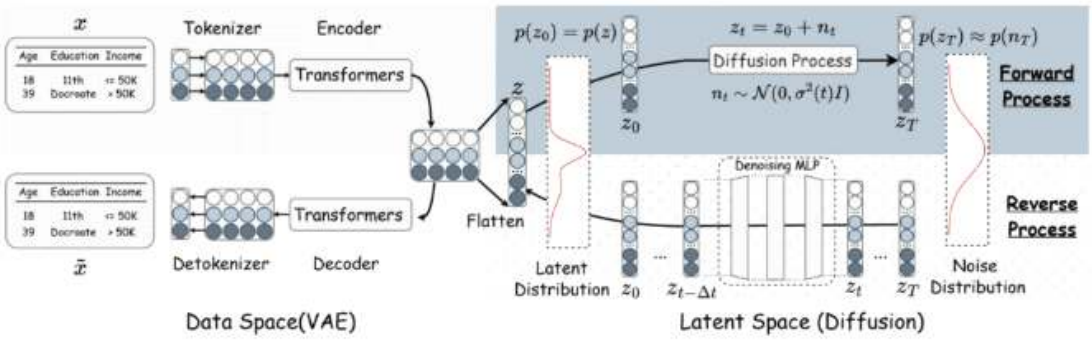
그림6 CTGAN 개념도



최근에는 VAE와 Diffusion 기법을 정형 데이터에 적용하는 시도로 유용성을 확보하는 TabSyn 알고리즘이 개발되었다. 필자도 실제 프로젝트에 적용해 본 결과, 유용성이 많이 개선되었지만, 동시에 안전성 확보에 어려움이 있어 좀 더 연구가 필요하다는 결론을 얻었다.

5) Charlie Harris, Diffusion Models for Molecule Design, (2024.06.05.)
6) Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, Kalyan Veeramachaneni, Modeling Tabular data using Conditional GAN, (2019.10.28.)

그림7 TabSyn 개념도[7)]



3. 합성데이터 생성 기술의 도전 과제

① 시계열 생성 기술

앞서 알아본 방법론은 레코드 간 독립성을 가정하고 만든 방법론이다. 하지만, 시계열 형태의 자료를 재현하고자 하는 수요가 꽤 있다. 이 경우, 위 방법론을 이용하면 안 된다. 이용을 위해서는 시계열 독립성을 제거하는 변환을 하고 이용할 수 있다.

시계열 자료를 재현하는 기술은 난도가 높은 분야로, 통계적으로는 XARIMA 모형, 딥러닝 RNN을 이용한 TimeGAN[8)] 기법 등이 있으나 유용성을 확보하기 위해서는 보다 많은 연구가 필요한 분야다.

② 딥러닝 기반 생성 기술

딥러닝 기반 생성 기술은 통계 모형 기반 생성 기술에 비해 아직 해결해야 할 문제가 많다. 현재 수준으로는 안전성보다는 유용성 확보가 먼저 필요하다. 유용성이 확보되지 않은 합성데이터는 안전하지만, 합성데이터로서 가치가 없다.

딥러닝 기반 생성 기술 알고리즘은 원본 분포와 합성데이터의 분포를 비교하면서 생성 알고리즘을 반복적으로 수정하기 때문에 시간이 오래 걸리고 학습률, 모형의 구조 등 하이퍼 파라미터에 따라 학습 결과가 달라져 같은 데이터를 서로 다른 사람이 해도 제각각의 결과를 얻어 유용성을 확보하기 어렵다. 그 때문에 학습의 불안정성을 해결할 수 있는 연구가 필요하다.

의료 이미지와 같이 특정 분야에 세밀한 이미지를 재현하는 문제는 어렵다. 일반인 육안으로 합성

7) Hengrui Zhang, et. al., Mixed-Type Tabular Data Synthesis with Score-based Diffusion in Latent Space, (2024.05.11.)
8) 김수은 외 2명, "경제 시계열 데이터의 확장을 위한 TimeGAN", 아시아태평양융합연구교류논문지, (2024.), p. 209-220.

이미지의 유용성을 판단할 수 없는 경우가 많다. 생성된 이미지가 전문가의 관점에서는 실제로 존재할 수 없는 형태로 나타날 수 있으며, 이러한 오류는 비전문가가 구분하기 어렵다. 이러한 문제는 데이터 생성 후, 육안으로 검색하여 제거하는 방식으로 해결하고 있다.

딥러닝 기반 합성데이터 생성 기술은 손실함수를 최소화하는 방향으로 학습을 진행하는데 이 과정에서 상대적으로 손실함수를 최소화하기 쉬운 사례를 집중적으로 생성하는 Mode Collapse 문제가 있다. 이러한 문제는 특정 그룹(성별, 인종, 질병 등)의 데이터를 많이 생성하거나 적게 생성하는 문제가 있다. 그 때문에 단순히 FID(Frechet Inception Distance) 등의 유용성 평가 방식 외에도 Precision, Recall과 같은 평가 방식을 추가로 고려해야 한다.[9)]

즉, FID, Precision, Recall 가 모두 일정 수준 이상이 되도록 생성해야 한다는 의미다. 마지막으로 원본 데이터를 레코드 단위로 변조하는 방식의 생성을 탐지하는 것이다. 합성데이터를 익명 데이터로 인정받기 위해서는 합성데이터 레코드와 원본 레코드 간의 연결 가능성이 없어야 한다.[10)]

그림8 잘못된 합성데이터 생성 예

9) Tuomas Kynkäänniemi, et_al., Improved Precision and Recall Metric for Assessing Generative Models, arXiv:1904.06991v3 [stat.ML], (2019.10.30.)
10) 개인정보보호위원회, 한국인터넷진흥원, 합성데이터 생성 활용 안내서, (2024.02.)

• 원본데이터

나이	BMI	수축기_혈압	이완기_혈압	혈당	...
58	21.103	110	87	133	...
71	19.504	122	75	72	...
48	32.807	101	74	170	...
34	28.401	111	89	120	...

• 합성데이터

나이	BMI	수축기_혈압	이완기_혈압	혈당	...
59	21.101	113	83	132	...
74	19.503	124	77	73	...
46	32.802	103	72	172	...
35	28.4.07	112	90	124	...

레코드 단위 변조는 합성 레코드와 원본 레코드의 연결 가능성이 존재하기 때문에 익명 데이터로 볼 수 없다. 이러한 경우는 원본 데이터를 가명 처리한 가명 데이터로 볼 수 있다.

4. 결론

본 원고에서는 합성데이터 생성 기술의 발전 과정에 대해 알아보고 각 기술에 대한 미래 도전 과제를 정리해 보았다. 합성데이터는 개인정보가 포함된 데이터를 원본과 같은 형식으로 완전히 개방할 수 있는 익명 데이터이므로 가치가 있다. 예를 들어. 개인 단위 교통카드 이용 데이터를 전 국민이 분석할 수 있도록 개방하면 집단 지성으로 훨씬 많은 데이터 활용을 이끌어낼 수 있을 것이다. 이처럼 많은 유용한 합성데이터 사례가 소개되길 바란다.

참고 문헌 |

1. Donald B. Rubin, Statistical Disclosure Limitation, (1993.)
2. Beata Nowok, Introduction to synthpop, Administrative Data Research Centre - Scotland
3. DeepCampus [VAE], beta-VAE, (2023.05.)
4. 컴케빈의 IT 전문지식창고, “AN 실제 이미지와 구별할 수 없는 가짜 이미지 생성하는 생성적 적대 신경망 개념과 동작원리 및 유형 이해, (2024.07.04.)
5. Charlie Harris, Diffusion Models for Molecule Design, (2024.06.05.)
6. Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, Kalyan Veeramachaneni, Modeling Tabular data using Conditional GAN, (2019.10.28.)
7. Hengrui Zhang, et. al., Mixed-Type Tabular Data Synthesis with Score-based Diffusion in Latent Space, (2024.05.11.)
8. 김수은 외 2명, “경제 시계열 데이터의 확장을 위한 TimeGAN”, 아시아태평양융합연구교류논문지, (2024.)
9. Tuomas Kynkäänniemi, et_al., Improved Precision and Recall Metric for Assessing Generative Models, arXiv:1904.06991v3 [stat.ML], (2019.10.30.)
10. 개인정보보호위원회, 한국인터넷진흥원, 합성데이터 생성 활용 안내서, (2024.02.)

합성데이터 생성 기술별 리스크 분석 및 완화방안

임종호
연세대학교 응용통계학과 | 교수

1. 데이터 합성과 노출위험 제어

데이터는 오늘날 사회·경제 전반에서 디지털 혁신을 가능케 하는 핵심 자산으로 자리 잡았다. 학계에서는 데이터 공유를 통해 연구의 재현성과 검증 가능성이 확보되고, 다분야 간 융합 연구가 활성화할 수 있다. 정부와 공공기관은 국가통계와 공공데이터를 개방·공유함으로써 증거 기반의 정책 결정 및 평가를 지원하고, 사회적 문제 해결을 위한 데이터 활용을 극대화할 수 있다. 기업과 산업 분야에서는 데이터 공유가 새로운 비즈니스 모델 발굴과 인공지능·빅데이터 기반 서비스 개발, 나아가 국제 경쟁력 강화의 원동력이 된다. 그러나 경제적 가치가 높은 데이터에는 개인 혹은 조직의 민감한 정보를 포함하는데, 이러한 데이터 공유의 확대는 해당 민감정보가 재식별되거나 노출될 수 있는 위험을 높이는 양면성을 지닌다. 따라서 데이터 공유의 필요성과 그로부터 파생되는 위험을 균형 있게 관리하는 것은 지속 가능한 데이터 활용 체계의 필수적 과제이다.

이러한 배경에서 최근 합성데이터(synthetic data) 생성 기술이 중요한 대안으로 부각되고 있다. 일반적으로 합성데이터는 '데이터 공유 및 문제 해결을 위한 목적으로 설계된 데이터 모형 또는 알고리즘을 활용하여 생성된 가상 데이터[11]를 의미한다. 이는 원시데이터(original data)의 민감정보를 직접 노출하지 않으면서도 데이터 분석 및 AI 모형 학습에 활용할 수 있다는 점에서 노출위험 제어(disclosure risk control)의 효과적 수단으로 사용되고 있다. 특히, 전통적 통계조사뿐 아니라 인공지능 학습데이터, 행정 데이터 등 광범위한 데이터 환경에서 합성데이터는 개인정보 보호와 활용 가치를 균형 있게 달성하는 핵심 기술로 주목받고 있다.

본 보고서는 이러한 합성데이터의 특성을 토대로, 생성 기술별 노출위험 제어 수준을 검토하고, 그 결과를 토대로 노출위험을 완화(緩和)할 방안을 탐색하는 데 목적이 있다. 이를 통해 합성데이터가 지닌 활용 가치를 극대화함과 동시에, 개인정보 보호를 보장하는 안전한 데이터 생태계 구축의 방향을 제시하고자 한다.

11) Jordon, J., Houssiau, F., Cherubin, G. & Cohen, S., Synthetic Data – what, why and how?, (2022.)

2. 합성데이터 노출위험 제어 원칙

노출위험 제어는 1970년대 마이크로데이터와 표 형태의 데이터를 민감정보를 노출하지 않고 공개해야 했던 국가 통계청과 공공데이터 기관의 필요로 시작되었으며, 초기에는 재식별 위험(identity disclosure risk)[12][13]에 집중되었으나, 이후 속성 노출위험(attribute disclosure risk), 추론위험(inference disclosure risk) 등으로 확대되었고, 최근에는 데이터 단위가 아니라 대규모 언어모델이나 생성형 모델 단위까지 노출위험 제어의 대상에 대한 범위가 넓어지고 있다.

이러한 노출위험 제어의 원칙은 합성데이터에도 동일하게 적용된다. 다만 합성데이터는 일반적으로 가명 데이터보다 익명 데이터에 가까운 형태로 생성하는 것을 목표로 하며, 이 과정에서 데이터 프라이버시 보호와 함께 활용 가치(유용성) 확보 역시 중요하다. 따라서 합성데이터 생성 기술별 노출위험 수준을 검토하기에 앞서, 그 기본 원칙을 정리하는 것이 필요하다. 이에 합성데이터를 중심으로 한 노출위험 제어 원칙을 표1에 정리하였다.

표1 합성데이터와 노출위험 제어 원칙

원칙	내용
노출위험 최소화	합성데이터 단위가 특정 개인·조직으로 연결되지 않도록 보장하고, AI 모델이 학습데이터를 그대로 재현하지 않도록 노출위험을 최소화
데이터 유용성과 안정성의 균형	원시데이터의 정보를 제대로 보존하지 않으면 활용도가 매우 낮아지고, 원시데이터와 유사하면 노출위험이 증가하기에 적절한 균형점을 설정하여 데이터 합성
맥락·목적 기반 차등적 제어	데이터 공개·활용 맥락(교육용, 연구용, 상업용 / 제한적접근, 공개 배포)에 따라 노출위험 수준을 차등적으로 적용
모집단 기반 제어	샘플 레벨이 아니라 모집단 레벨에서 노출위험 수준이 제어

* 출처: UK Data Service[14]와 Statistics Canada[15]가 발간한 보고서 내용 중심으로 저자가 재구성

합성데이터에서 노출위험 제어의 기본 전략은 표1에서 정리된 첫 번째 원칙에 정리된 것처럼 '위험 최소화'로 '위험 제거'가 아니다. 합성데이터는 원시데이터를 기반으로 생성되기 때문에, 생성된 데이터가 원시데이터의 통계적 특성과 구조를 반영하는 과정에서 직·간접적으로 민감정보를 포함할 가능성을 완전히 배제하기는 어렵다. 따라서 합성데이터에 대해 식별 위험을 전혀 존재하지 않는 것으로 간주하는 것은 타당하지 않다. 이러한 특성으로 인해 합성데이터를 법적·제도적으로 익명 데이터로 간주할 수 있는지가 중요한 쟁점이 되며, 그 적용과 해석에는 실제 데이터 생성 방식과 활용 맥락 등 구체적 조건을 반영할 필요가 있다. 이러한 맥락에서 노출위험 최소화, 데이터 유용성과 안정성의 균형, 맥락·목적 기반 차등적 제어와 같은 기본 원칙이 합성데이터 논의에서 특히 중요하다.

12) Dalenius, T., Towards a methodology for statistical disclosure control, (1977.)
13) Sweeney, L., k-Anonymity: A model for protecting privacy, (2002.)
14) UK Data Service, Handbook on Statistical Disclosure Control for Outputs, (2025.)
15) Statistics Canada, Statistical Disclosure Control for Public Use Microdata Files, (2016.)

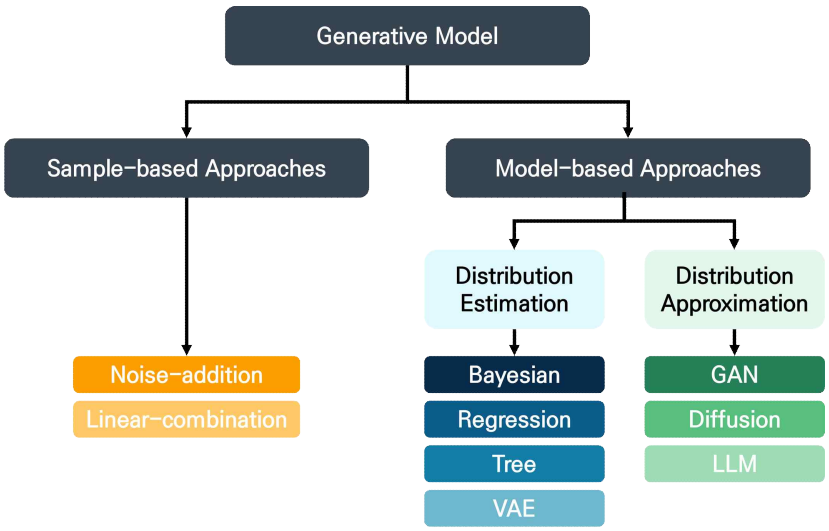
한편, 합성데이터 생성 과정에서 종종 간과되는 원칙 중 하나는 노출위험을 모집단 레벨에서 평가하고 제어해야 한다는 점이다. 예를 들어, 이메일·전화번호·증명사진과 같이 모집단 전체에서 개인을 식별하고 구분할 수 있는 정보는 식별 및 노출위험이 매우 크다. 반면 성별·연령·소득·거주지와 같은 속성의 조합은 원시데이터 수준에서는 특정 개인을 유일하게 구분할 수 있지만, 모집단 전체에서는 동일한 특성을 가진 집단이 존재하기 때문에 상대적으로 노출위험이 낮다. 이러한 특성을 고려하면, 데이터 합성 이전 단계에서 재식별 위험이 큰 샘플을 사전에 제거하거나 전처리하는 절차가 필요하다. 본 보고서는 합성데이터 생성 기술별 노출위험 수준을 검토하는 것을 목표로 하지만, 합성 전 과정의 관점에서 보면 사전 샘플 제어를 통해 위험 수준을 낮추는 방안을 함께 검토하는 것이 중요하다. 이를 위해서는 개별 데이터 단위에서의 노출위험 측정이 선행되어야 한다. 비정형데이터에 관한 연구는 아직 부족하지만, 정형·반정형 데이터와 관련해서는 상당한 연구가 축적되어 있으며, 이에 대해서는 국가통계연구원(구 통계개발원)이 2022년에 발간한 과학적 방법 중심 보고서[16]를 참조할 만하다.

3. 데이터 합성 기술과 노출위험 제어

① 데이터 합성 기술별 특성 및 노출위험 수준

재표집(resampling) 기법 이외에 데이터 모형에 기반한 합성데이터를 활용하여 노출위험을 제어하려는 시도는, 1993년 *Journal of Official Statistics*의 특집호 "Statistical Disclosure Limitation"에서 소개된 결측대체(imputation)·다중대체(multiple imputation) 기반 아이디어를 중심으로 체계화되기 시작하였다. 이 접근은 표본 단위의 직접적 수정이 아니라, 모형이 학습한 분포적 구조를 통해 새로운 자료를 생성함으로써 원시데이터의 민감정보를 직접 노출하지 않으면서도 통계적 유용성을 유지하려는 데에 초점을 둔다.

그림1 데이터 합성 기술 분류



* 출처: 저자 작성[17]

16) 임종호, 김현태, 정동훈, 개별 관측치 단위(극단값 등)에서의 통계적 노출위험 및 제어방법론 개발 연구, (2022.)
17) 김수영, 정동훈, 김현태, 임종호, A Survey on Tabular Data Synthesis: Generation, Evaluation, and Benchmarking Experiments, (2024.)

합성데이터를 통한 노출위험 제어는 그림1과 같이 모형 기반 방법과 샘플 기반 방법으로 발전해 왔다. 모형 기반 방법은 다시 분포 추정과 분포 근사로 나뉘며, 전자는 원시데이터의 분포와 추정을 활용하는 형태로 베이지안 방법 등 통계적 기법 등이 주로 해당한다. 후자는 분포를 근사해 재현하는 방식으로 GAN, DM (diffusion model), LLM 계열이 대표적이다. 한편, 샘플 기반 방법은 데이터에 노이즈를 추가하는 형태인 차등보호(differential privacy, DP) 기법과 mixup처럼 원시데이터를 선형 결합하는 기법 등이 대표적이다. 표2는 이러한 데이터 합성 기술들을 유용성, 노출위험 제어 수준, 최신 동향 관점에서 비교·분석한 것을 보여주고 있다.

표2의 결과를 종합하면, 데이터 유용성과 안전성을 동시에 충족하는 방법은 존재하지 않는다. 이는 양자 간 상충관계에 따른 자연스러운 결과이며, 최근 연구는 여러 합성 기법을 결합해 두 측면을 함께 개선하려는 방향으로 발전하고 있다. 다만 이러한 시도들은 아직 성능이 충분히 검증되지 않아, 실제 현장 적용에는 제한적이다.

표2 대표적 데이터 합성 기법 비교

구분	합성 기술	유용성	노출위험 수준 제어	내용 & 최신 기술 동향
모형 기반 분포 추정	Bayesian	★☆	★★☆	• 높은 수준으로 노출위험 제어가 가능하지만, 비정형 데이터나 빅데이터에 취약함 • 베이지안 합성+MCMC의 효율적 구현 사례가 증가하고 있음
	Tree	★★	★☆	• 구현이 용이하기에 많이 활용되고 있으나, 최근에는 상대적으로 최선호 기법은 아님 • 최근에는 랜덤포레스트 기반 합성 방법이 다양한 형태로 고도화되고 있음
모형 기반 분포 근사	GAN	★	★☆	• 다양한 형태로 합성데이터 생성 기술이 개발되었지만, 대체로 성능이 좋지 않음 • 효과적 노출위험 수준 제어를 위하여 노출위험 제어 중심의 방법들[18]이 연구되고 있음
	DM	★★☆	★☆	• 현시점 가장 주목받는 합성 기술로 정형·비정형·시계열 등 다양한 데이터에서 높은 유용성을 보장하기에 주목을 받고 있음 (예시: DDPM[19], LDM[20], SSSD[21] 등) • 구조적 특성상 노출위험 수준이 높은 것은 단점이기에, 최신에는 DP 등과 결합한 기법 등이 연구되고 있음
	LLM	★★	★	• 복잡한 상관 구조를 반영할 수 있기에 다양한 데이터에 적용할 수 있는 장점이 있으나, 학습 데이터 암기 등 노출위험은 매우 큰 편임 • 최근, 이러한 단점을 극복하기 위하여 DP 기술 등 PETs 기술을 접목한 기법들이 개발되고 있음

18) Shateri, M., Messina, F., Labeau, F. & Piantanida, P., Preserving Privacy in GANs Against Membership Inference Attack, (2023.)
19) Ho, J., Jain, A. & Abbeel, P., Denoising diffusion probabilistic models, (2020.)
20) Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B., High-Resolution Image Synthesis with Latent Diffusion Models, (2022.)

샘플 기반	mixup	★★	★☆	• 구현이 쉽고 특정 구조의 정보를 보존하기에 유리하기에 다방면으로 많이 활용되고 있으나 노출위험 수준은 매우 높은 편임 • 최근에는 불균형 문제를 해결하는 데 주로 활용되고 있음
	DP	★	★★☆	• 데이터 프라이버시 보호 측면에서는 유용하지만, 데이터가 가진 정보 보존 측면에서는 가장 열악함 • 최근에는 단독 기법으로 쓰이는 경우는 매우 드물고, 기존의 데이터 합성 기술과 결합한 형태 (예: PATE-GAN[22], DPDM[23], DP-LLM[24] 등)로 많이 활용 및 연구되고 있음

* 출처: 최신 데이터 합성 관련 서베이/리뷰 논문들[25][26]과 저자 연구 결과를 기반으로 작성

② 차등보호 기술 활용과 한계

차등보호 기술은 수학적 근거에 기반한 노출위험 제어, 데이터 합성과의 결합 가능성, 재식별 위험 차단이라는 장점으로 강력한 프라이버시 보호 수단으로 평가된다. 그러나 본질적으로 노이즈를 추가하는 형태이기에, 보호 수준이 높아질수록 데이터의 정확성과 활용성은 감소한다. 합성데이터 응용에서도 훈련 데이터 민감도로 인해 일관되고 강건한 데이터 생성이 어렵다는 한계가 있다.

예를 들어, 미국 2020년 인구센서스 사례가 이를 잘 보여준다. 미국의 인구조사국은 차등보호 기술을 적용해 통계를 산출했지만, 소지역·소집단에서 큰 오차와 왜곡이 발생해 품질 저하, 이해관계자 반발, 소송으로 이어졌다. 이는 충분한 검증 없이 기술을 도입할 경우 발생할 수 있는 위험을 보여준다. 이러한 위험을 줄이고 차등보호 기술의 장점을 기존 합성 데이터 기법에 안정적으로 접목하려면, 훈련 데이터 민감성 완화, 노이즈 규모와 배분의 최적화, 고차원·시계열·네트워크 등 복잡한 데이터 구조에 대한 적용 방안 개발 등 지속적인 연구와 검증이 필요하다.

4. 맺음말

합성데이터는 데이터 공유와 활용을 가능케 하면서도 민감정보의 노출위험을 줄일 수 있는 중요한 대안으로 자리 잡고 있다. 하지만 기술의 특성상 원시데이터의 통계적 구조와 정보를 반영하는 과정에서 완전한 위험 제거는 불가능하며, 유용성과 안전성 간의 균형이 핵심 과제로 남는다. 특히 최근 확산되고 있는 딥러닝 기반 합성 기법은 다양한 데이터 유형에서 높은 성능을 보이지만, 구조적 특성상 노출위험을 내포하고 있어 지속적인 보완 연구가 필요하다.

21) Alcaraz, J.M.L. & Nils, S., Diffusion-based time series imputation and forecasting with structured state space models, (2023.)
22) Jordon, J., Yoon, J., Van Der Schaar, M., PATE-GAN: Generating synthetic data with differential privacy guarantees, (2019.)
23) Dockhorn, T., Cao, T.A.L. & Mandt, S., Differentially Private Diffusion Models, (2022.)
24) Yu, D., Bagdasaryan, E. & Shmatikov, V., Differentially Private Fine-tuning of Language Models, (2021.)
25) Croitoru et al., Diffusion models: A comprehensive survey of methods and applications, (2023.)
26) Drechsler, J., Reiter, J.P., 30 years of synthetic data, (2024.)

합성데이터 생성 기술은 발전 단계에 따라 서로 다른 장단점을 보인다. 전통적 통계 모형은 해석 가능성과 안정성이 높지만 복잡한 데이터 구조를 충분히 반영하기 어렵다. 머신러닝 기반 기법은 데이터 유용성을 개선하는 데 효과적이나, 과적합 시 원본 데이터 유사 샘플을 생성할 위험이 따른다. 반면 생성형 모델은 정형·비정형·시계열 데이터를 아우르는 높은 재현성을 제공하지만, 메모리제이션(memorization) 현상으로 인한 노출위험 수준이 높다. 따라서 차등보호 기술을 활용하는 등 특정 목적에 맞는 합성 기법을 선택하고, 이에 따른 위험 관리 전략을 병행하는 것이 필수적이다.

원시데이터의 통계적 특성과 구조적 복잡성을 충실히 재현하면서도 일관성과 강건성을 확보해야 한다. 동시에 이러한 품질 기준이 다양한 활용 분야에서 일관되게 평가·검증될 수 있는 체계 마련이 요구된다. 이를 통해 합성데이터는 연구·산업·정책 현장에서 신뢰받는 자원으로 자리매김하고, 개인정보 보호와 데이터 활용을 조화롭게 달성하는 데 기여할 수 있을 것이다.

참고 문헌 |

1. Jordon, J., Houssiau, F., Cherubin, G. & Cohen, S., Synthetic Data – what, why and how?, (2022.)
2. Dalenius, T., Towards a methodology for statistical disclosure control, (1977.)
3. Sweeney, L., k-Anonymity: A model for protecting privacy, (2002.)
4. UK Data Service, Handbook on Statistical Disclosure Control for Outputs, (2025.)
5. Statistics Canada, Statistical Disclosure Control for Public Use Microdata Files, (2016.)
6. 임종호, 김현태, 정동훈, 개별 관측치 단위(극단값 등)에서의 통계적 노출위험 및 제어방법론 개발 연구, (2022.)
7. 김수영, 정동훈, 김현태, 임종호, A Survey on Tabular Data Synthesis: Generation, Evaluation, and Benchmarking Experiments, (2024.)
8. Shateri, M., Messina, F., Labeau, F. & Piantanida, P., Preserving Privacy in GANs Against Membership Inference Attack, (2023.)
9. Ho, J., Jain, A. & Abbeel, P., Denoising diffusion probabilistic models, (2020.)
10. Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B., High-Resolution Image Synthesis with Latent Diffusion Models, (2022.)
11. Alcaraz, J.M.L. & Nils, S., Diffusion-based time series imputation and forecasting with structured state space models, (2023.)
12. Jordon, J., Yoon, J., Van Der Schaar, M., PATE-GAN: Generating synthetic data with differential privacy guarantees, (2019.)
13. Dockhorn, T., Cao, T.A.L. & Mandt, S., Differentially Private Diffusion Models, (2022.)
14. Yu, D., Bagdasaryan, E. & Shmatikov, V., Differentially Private Fine-tuning of Language Models, (2021.)
15. Croitoru et al., Diffusion models: A comprehensive survey of methods and applications, (2023.)
16. Drechsler, J., Reiter, J.P., 30 years of synthetic data, (2024.)

AI 시대의 새로운 데이터 패러다임, 합성데이터

이진규
NAVER Corporation I 정보보호최고책임자 및 개인정보보호책임자

1. 들어가며: AI 시대의 새로운 데이터 패러다임, 합성데이터

① 고품질 데이터 확보와 데이터의 저주

인공지능(AI) 시대에 대규모 언어모델(LLM) 등 알고리즘은 더 이상 소수 기술기업의 전유물이 아닌 국가와 산업의 명운을 가르는 핵심 동력이 되었다. 이러한 알고리즘의 생성과 운용을 가능하게 해주는 것은 다름 아닌 데이터인데, AI 기술은 우리가 얼마나 양질의 데이터를 확보할 수 있는가에 달렸다고 해도 과언이 아니다. 그런데, 문제는 AI 시대의 연료라 할 수 있는 데이터가 적잖은 문제를 내포하고 있다는데 AI 시대의 딜레마가 있다. 더 많은 데이터를 확보하려는 경쟁에 매몰되어 소위 '데이터의 저주(curse of data)'에 걸릴 수 있다는 사실을 망각하게 되는 것이다.

표1 데이터의 저주에 해당하는 주요 사례

구분	핵심 내용	사례 / 영향
데이터 편향성	현실의 불공정성을 그대로 반영하거나 증폭	- 과거 채용 데이터 기반으로 학습한 AI가 특정 성별 내지 인종에 대해 차별적 결과를 야기[27] - 피부암 데이터가 부족한 인종에 대해 의료 AI가 오진
개인정보 민감성	유출 시 위험, 또는 활용의 제약	- 금융 거래기록, 의료영상, 개인 식별정보 등은 AI학습에 필수지만, GDPR, HIPAA 등 강력한 규제로 활용에 제약 - 사실과 구분하기 힘든 허위 주장(hallucination)과 결합한 개인정보 누설이 사생활 및 안전에 위협으로 작용[28]
데이터 확보 비용	고품질 데이터 확보에 막대한 비용과 시간 소요 (→독점 심화)	- 자율주행차: 수억 km의 실제 주행 데이터 필요 - 희귀 질환 AI: 전 세계 흩어진 환자 데이터 수집 필요 - 거대 기업만 데이터 축적 가능하여 스타트업이나 연구기관이 AI 모델 제작에 한계 존재

② 합성데이터의 부상 - 정의, 생성 기술 및 중요성

이처럼 실제 데이터가 가진 '삼중고(三重苦)'—편향성, 프라이버시, 비용—를 해결할 대안으로

27) 저자는 알고리즘과 빅데이터가 사회적 불평등을 심화시키는 방식을 다양한 사례를 통해 데이터 편향성의 위험을 경고했다. (O'Neil, C., Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown, (2016.))

28) 허위 사실과 결합한 개인정보 누설로 명예를 실추당한 언론인이 시민단체(NOYB)의 도움을 받아 노르웨이 개인정보 감독당국에 고발장을 제출했다. (NOYB, AI hallucinations: ChatGPT created a fake child murderer, (2025.))

합성데이터(Synthetic Data)가 수면 위로 떠오르고 있다. 합성데이터는 실제 데이터가 아닌, 알고리즘에 의해 인공적으로 생성된 가상의 데이터다. 생성적 적대 신경망(GANs, Generative Adversarial Networks)과 같은 기술은 한 네트워크(생성자)가 실제 같은 데이터를 만들어내면 다른 네트워크(판별자)가 그 진위를 판별하며 서로 경쟁하는 과정을 통해, 원본의 통계적 특성과 놀라울 정도로 유사하면서도 완전히 새로운 개별 데이터를 만들어낸다.[29)]

합성 데이터는 데이터의 양을 증가시키면서도, 데이터의 질과 패러다임을 바꾸어냈다. 데이터 확보와 관련한 저주를 풀어낼 수 있는 촉매제로서 기능하면서, AI 개발의 민주화를 이끌어내는 방법으로 기대를 모으고 있다. 실제 환자의 의료영상이 아닌 가상의 뇌 MRI 이미지를 생성하여 희귀 질환 진단 모델을 훈련시키거나 가상의 금융 거래 기록을 생성하여 사기 탐지 시스템(FDS, Fraud Detection System)을 고도화하는 것이 가능해지는 것이다.[30)]

아래에서 AI 학습용 합성데이터의 활용 사례, 활용 가치 및 한계, 향후 과제 등을 순서대로 살펴보기로 한다.

2. AI 학습용 합성데이터의 활용 사례 분석

합성데이터의 가치는 사고실험이나 실험실 수준에서 적용가능한 이론 중심의 체계에 머물지 않고 이미 다양한 산업 현장에서 실질적 성과를 창출하고 있다. 데이터의 확보 곤란으로 인해 발생했던 여러 문제를 합성데이터가 어떻게 풀어내고 있는지 실제 사례를 확인해 본다.

① 의료 부문 (희귀 질환 진단 및 의료 AI 모델의 고도화)

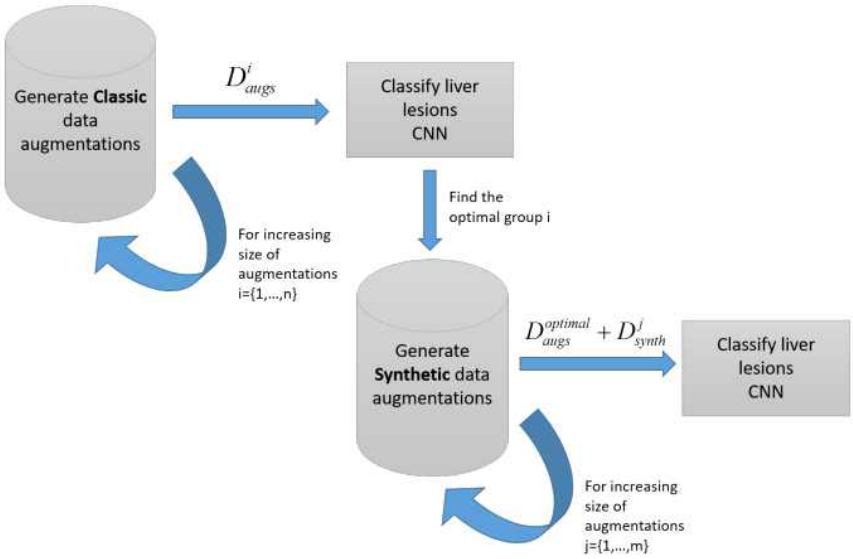
의료 분야에서 개인 데이터는 민감성이 가장 문제시되며, 희소성도 문제를 악화시킨다. 희귀 질환의 경우 데이터 부족으로 모델을 생성하는 것이 불가능한 경우가 많았다. 합성데이터는 개인식별성을 배제하여 민감성 문제를 해소하면서도 데이터 희소성도 동시에 해결한다.

스탠퍼드 대학 연구팀은 실제 뇌종양 환자 MRI 데이터에 GAN기술을 적용하여 대량의 고해상도 합성 뇌종양 MRI 이미지를 생성했다. 연구팀은 이를 학습한 AI와 실제 데이터만 학습한 모델을 비교했는데, 전자의 경우 종양 탐지 정확도가 유의미하게 향상되는 결과를 보였다. 이는 합성데이터가 단순히 데이터의 양을 늘리는 데 그치지 않고, 모델의 정확성을 향상시키는데 기여할 수 있음을 보여주는 사례다.[31)]

29) GAN의 기본 개념을 제시한 논문으로, 두 신경망의 제로섬 게임(zero-sum game)을 통해 데이터를 생성하는 혁신적인 아이디어를 제안했다. (Goodfellow, I. J., et al, Generative adversarial nets. In Advances in neural information processing systems, (2014.), p. 2672-2680.)

30) 가트너는 생성형 AI를 주요 기술 트렌드로 선정하며, 2025년까지 합성데이터가 실제 데이터를 추월하여 AI 학습의 주요 데이터 소스가 될 것으로 예측했다. (Gartner, Gartner Top Strategic Technology Trends for 2022: Generative AI, (2021.))

그림1 간 병변 ROI 분류 과제에서 합성 데이터 증강 효과 평가를 위한 실험절차 흐름도



이와 같이 의료 영역에서의 합성데이터 활용은 환자의 민감한 개인정보인 의료정보를 외부에 노출하지 않고 안전한 연구 환경을 제공한다. 또한, 특정 연령, 성별, 인종 등의 특성을 고려할 때 실제 데이터셋에서 부족이 발생할 수 밖에 없는 경우, 필요한 데이터를 보충하면서도 모델의 정확성을 향상시켜 의료 AI의 성능을 발전시키면서도 진단 결과에 대한 신뢰성을 증대할 수 있다는 점에서 매우 의미있는 방식이라 할 수 있다. 다만, 실제 의료 영상에서 관찰 가능한 비정형적 병변(atypical lesion)을 완벽히 재현하는 것에는 아직 기술적 한계가 존재한다. 이로 인해 만약 생성모델이 실제 데이터에 없는 가상의 이상 징후를 만들어내는 경우 오진으로 이어질 수 있으므로, 반드시 의료 전문가가 임상적 유효성을 검증하는 절차가 필요하다.

② 금융 부문(사기 탐지 시스템 고도화)

유럽 중앙은행 보고서에 의하면 가장 높은 사기 거래 비율은 카드 결제 분야에서 확인되는데, '23년도 상반기 기준으로 전체 신용카드 지불액의 약 0.031%, 거래 건수의 0.015%가 사기에 해당하는 것으로 나타났다.[32] 이는 금융 데이터의 극심한 불균형을 보여주는 것으로, AI 모델은 정상 거래를 주로 학습하고 신종 사기 패턴을 탐지하지 못할 가능성이 크다는 것을 보여주는 것이다.

글로벌 신용카드사 아메리칸 익스프레스(American Express)는 합성데이터를 활용하여 신종 사기 거래 패턴을 대량으로 생성하고, 이를 통해 자사의 사기 탐지 시스템(FDS)의 정확도를 크게 향상시

31) Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., & Greenspan, H., GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. Neurocomputing, (2018.), p. 321-331.
32) European Banking Authority, 2024 Report on Payment Fraud, (2024.), p.10

켰다.[33] 금융 거래는 개인의 민감한 신용과 관련한 개인정보로서 높은 수준의 보안이 요구되는 정보인데, 합성데이터는 개별 거래의 상세한 데이터를 포함하지 않으면서도 전체 데이터의 통계적 분포를 유지하여 실제 사기 거래를 탐지하는데 유용하다. 특히, 기존에 알려진 사기 유형에 대응하는 것을 뛰어넘어, 새로운 신종 사기유형을 탐지하는데 있어서도 풍부한 데이터를 제공함으로써 안전한 금융거래에 도움을 준다. 다만, 금융시장의 특성상 예측 불가능한 거시 경제 지표나 인간의 비합리적 의사결정(소위 '투심')까지 데이터로 반영하는 것은 가능하지 않다. 또한, 금융계에서 희귀하게 발생하는 '블랙스완(black swan)'과 같은 상황까지 예측하여 모델을 학습시키는 것이 가능한지도 여전히 불분명하다.

③ 이미지 분석(자율주행)

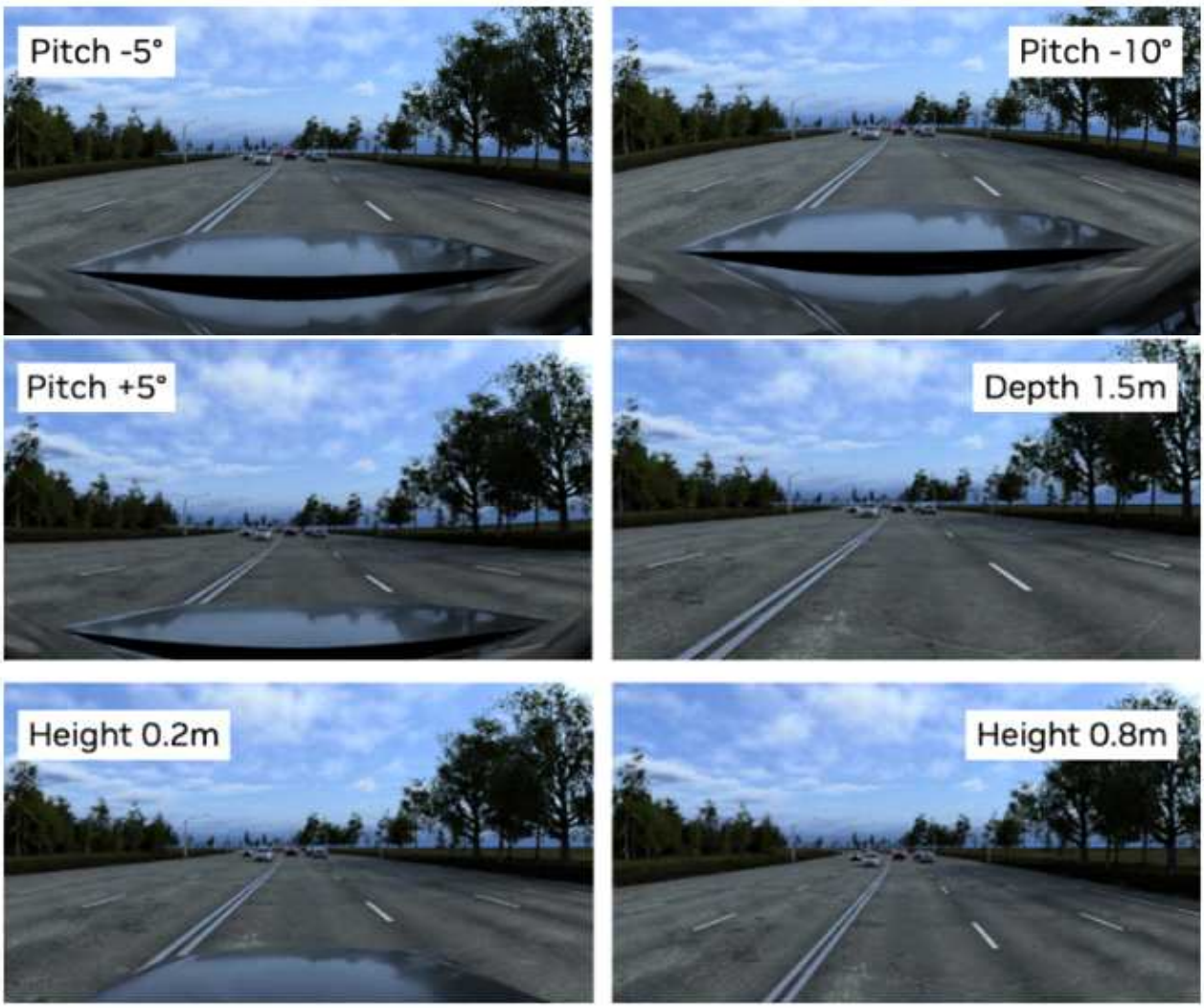
자율주행차의 안전성은 다양한 돌발 상황(Edge Case)에 얼마나 효과적으로 대처하는지에 달려 있다고 해도 과언이 아니다. 일반적 주행 환경에서의 주행 보조(driving assistance)와 달리 돌발 상황에서의 안전성을 확보하기 위해선 다양한 예외적 상황에 대한 데이터를 다수 확보해야 하기 때문이다. 무단횡단하는 보행자, 도로에 갑자기 진입하는 취객, 도로 파손, 예상치 못한 장애물 등과 같은 상황을 실제 데이터로 확보하는 것은 매우 어렵거니와, 만약 확보한다 하더라도 그 수량은 제한적일 수밖에 없는 것이 현실이다.

웨이모(Waymo) 등 자율주행 기술 기업은 엔비디아(NVIDIA)가 구축한 고도로 정교한 가상 시뮬레이션 환경에서 수백만 킬로미터에 달하는, 생성된 주행 데이터를 활용한다. 이 가상 환경에서는 다양한 기상 조건, 조명, 교통 상황을 자유자재로 조합하며 AI의 인지 모델을 극한의 상황까지 훈련시킨다. 때로는 자율주행차의 인식(perception) 알고리즘 향상을 위해 카메라나 라이다 같은 센서의 위치나 높이가 달라지는 차량을 고려하여, Novel View Synthesis(NVS) 기술을 도입해 기존 장면 데이터를 기반으로 원래 수집되지 않은 새로운 시점의 이미지를 합성한다. 합성 데이터와 NVS 기술을 활용하면, 센서 구성에 따른 데이터 수집의 비용과 시간을 크게 줄일 수 있으며, 한 번 수집된 데이터를 다양한 차량 타입과 관점에 적용 가능하게 되어, 자율주행차량 시스템 배치 시 유연성과 효율성이 크게 향상된다.

그림2 인식 검증용으로 다양한 피치·깊이·높이 조건으로 생성된 카메라 이미지셋

합성데이터는 데이터 수집 비용과 시간을 획기적으로 절감한다. 특히, 모든 객체가 자동으로 라벨링(Labeling) 된 상태로 데이터가 생성되므로, 수작업 라벨링에 소요되는 막대한 인적, 물적 자원을 아낄 수 있다는 장점도 있다. 그러나, 시뮬레이션과 현실 세계 간의 차이, 즉 'Sim-to-Real Gap'

33) Fortune, Why American Express is trying technology that makes deepfake videos look real, (2020.)



문제는 여전히 해결이 필요한 지점이다. 가상 환경의 물리 법칙이나 빛의 반사, 그림자 등이 현실과 미세하게 다를 경우, AI는 가상 환경에만 과적합(overfitting)되어 실제 도로에서는 제대로 작동하지 않을 수 있다. 이 간극을 줄이는 것이 합성데이터 기반 자율주행 기술의 핵심 과제라 할 수 있다. 도메인 랜덤화(domain randomization), 정밀 시뮬레이션 기반 직접 전이(direct transfer)와 같은 방식이 이러한 Sim-to-Real Gap 문제를 해결하기 위한 대안으로 제시된다. 전자는 시뮬레이션 단계에서 물리 법칙, 조명, 센서 노이즈 등 다양한 환경 요소를 일부러 무작위로 바꿔서 훈련해, 실제 환경에서도 잘 적응하도록 만드는 방법이며, 후자는 실제 차량의 특성(무게, 마찰, 센서 특성 등)을 정밀하게 측정해 시뮬레이션에 반영하고, 이를 통해 학습한 모델을 바로 실제 차량에 적용하는 방법이다.

표2 산업별 합성데이터 활용 사례 비교

	의료 영상	금융	이미지 인식
주요 적용 분야	희귀 질환 진단, 의료 AI 모델 개발	FDS, 알고리즘 트레이딩	자율주행, 스마트 팩토리, 로보틱스
핵심 해결 문제	데이터 희소성, 환자 프라이버시	데이터 불균형, 고객 개인(거래)정보 보호	돌발 상황 데이터 부족, 라벨링 비용
주요 활용 가치	희귀 질환 데이터 확보, 의료 데이터 공유 및 연구 활성화	신종 사기패턴 예측, 예측 모델 효율성 증대	위험 상황 시뮬레이션, 개발 비용 및 시간 단축
대표적 한계	미세 병변 재현 및 임상 유효성 검증 한계	생성 데이터의 현실성 문제	현실과 가상 환경의 괴리, 실제 물리 현상 재현 곤란

3. 합성데이터의 활용 가치와 한계

① 다각적 활용 가치

합성데이터는 소수의 빅테크 기업이 독점하던 고품질 데이터를 스타트업, 대학, 연구기관도 활용할 수 있게 되어 AI 기술 개발의 진입장벽을 낮춘다. 이를 통해 공정한 경쟁을 촉진하고, AI 생태계 전반의 혁신을 가속할 수 있다. 데이터 접근성을 확대하고 AI 개발 민주화를 확산할 수 있는 것이다.

프라이버시 측면에서 합성데이터는 개인 식별 정보를 제거하는 '비식별화' 조치를 넘어서, 개인정보가 원천적으로 존재하지 않는 익명 데이터를 사용함으로써 규제에 적극적으로 대응할 수 있다. 개인정보가 유출되지 않는 환경을 만들어 안전성을 확보한다. 개인 프라이버시가 침해되지 않는 데이터는 규제 준수 및 데이터 보호 비용을 획기적으로 줄여준다.

모델의 강건성(robustness)과 공정성(fairness) 측면에서 실제 데이터에 부족한 소수자 그룹(예: 특정 인종, 희귀 질) 데이터를 생성하고, 데이터 불균형을 해소한다. 이를 통해 인공지능 모델이 특정 사회 그룹에 대해 불리한 결정을 내리는 편향성을 완화하고 사회적 공정성을 제고할 수 있다.

비용 및 시간을 절약하는데 있어 합성 데이터는 큰 기여를 할 수 있다. 데이터 수집을 위한 물리적 제약(예: 실제 도로 주행)과 데이터 가공 및 라벨링에 필요한 막대한 인적 자원을 절감하여 AI 개발 주기를 획기적으로 단축하고 효율성을 극대화한다. 현실에서 매우 드문 빈도로 발생하는 일들이지만, 만약 발생하는 경우 사람의 생명과 재산을 앗아갈 수 있는 상황을 가상으로 생성하여 그러한 일이 실제 발생하지 않도록 예방할 수 있다. AI 개발주기에서의 비용과 시간을 절약할 뿐만 아니라, AI가 적용되는 현장에서 발생할 수 있는 사회적 비용을 줄여줄 수 있다는 점에서도 합성데이터의 활용 가치는 매우 크다 할 것이다.

② 한계

합성데이터 역시 일반적인 데이터셋에 언급되는 문제점을 다수 내포하고 있다. 우선, 품질과 신뢰성 측면에서 “Garbage In, Garbage Out”의 문제를 그대로 가지고 있다. 원본이 되는 실제 데

이터가 편향되어 있거나 품질이 낮을 경우, 합성데이터는 그 편향을 그대로 학습하거나 오히려 증폭시킬 위험이 있다.

합성데이터의 가장 큰 숙제는 현실 세계의 복잡성, 노이즈, 예측 불가능성을 얼마나 충실히 재현하는가 하는 '충실도(Fidelity)' 문제와 본질적으로 같다. 현실세계의 복잡성을 데이터라는 제한된 형식으로 표현하는 데는 반드시 제약이 발생할 수밖에 없다. 특히, 인간의 감정이나 개인 내지 그룹 간의 상호작용과 같이 데이터화 하기 어려운 영역에서는 이러한 제약이 더욱 도드라질 수밖에 없다.

합성데이터의 품질을 제대로 평가할 수 있는 방법론이나 도구가 아직 확립되지 않은 것도 합성데이터의 채택을 어렵게 하는 요인 가운데 한가지이다. 생성된 데이터가 원본과 얼마나 유사한지를 어떻게 평가할 수 있다는 말인가? 이미지의 경우, 평가단으로 하여금 유사도를 제시하도록 하여 평가를 진행할 수는 있으나 실제 AI 학습에서 이러한 유사도가 의도한 결과를 만들어낼 수 있는지는 전혀 다른 이슈이다.

마지막으로, 생성모델 자체의 저작권 귀속 문제나 딥페이크 기술처럼 사회적으로 악용되는데 합성데이터가 사용되는 경우 윤리적 문제가 발생할 수 있다.[34)]

4. 나가며

합성데이터는 데이터 부족, 편향, 프라이버시 문제로 성장의 한계에 부딪힌 AI 기술에 새로운 돌파구를 열어줄 것으로 기대된다. 그러나, 합성데이터가 모든 문제를 해결할 수 있는 것은 아니다. 원본 데이터의 품질에 의존하며, 현실 세계를 제대로 담아내는 데도 한계가 있다. 이를 제대로 평가할 수 있는 방법론은 아직 부족하며, 악용을 막아내는 것도 여전한 숙제다.

AI 혁신의 가능성을 극대화하고, 부작용을 막아내면서 사회의 여러 분야에서 합성데이터가 가진 가치를 제대로 끄집어낼 수 있도록 보다 많은 응용 사례를 만들어내는 동시에, 사회적으로 합의된 평가 방법론을 만들어내는 데 집중해야 할 때이다.

34) Chesney, R., & Citron, D., Deep fakes: A looming challenge for privacy, democracy, and national security. Lawfare Research Paper Series, (2019.), p.1-19.

참고 문헌 |

1. O'Neil, C., Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown, (2016.)
2. NOYB, AI hallucinations: ChatGPT created a fake child murderer, (2025.)
3. Goodfellow, I. J., et al, Generative adversarial nets. In Advances in neural information processing systems, (2014.)
4. Gartner, Gartner Top Strategic Technology Trends for 2022: Generative AI, (2021.)
5. Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., & Greenspan, H., GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. Neurocomputing, (2018.)
6. European Banking Authority, 2024 Report on Payment Fraud, (2024.)
7. Fortune, Why American Express is trying technology that makes deepfake videos look real, (2020.)
8. Chesney, R., & Citron, D., Deep fakes: A looming challenge for privacy, democracy, and national security. Lawfare Research Paper Series, (2019.)

PRIVACY REPORT

개인정보 이슈 심층분석

「2025 개인정보 이슈 심층분석 보고서」는
개인정보보호위원회의 출연금으로 수행한 사업의 결과물입니다.

한국인터넷진흥원의 승인 없이 본 보고서의 무단전재나 복제를 금하며,
인용 출처 「2025 개인정보 이슈 심층분석 보고서」를 밝혀주시기 바랍니다.

본 보고서의 내용은
한국인터넷진흥원의 공식 견해가 아님을 알려드립니다.

발행일 2025년 9월
발행처 한국인터넷진흥원 개인정보제도팀
발행처 전라남도 나주시 진흥길 9

KISA 한국인터넷진흥원