연구보고서

2023

07

데이터 가명·익명처리 기법의 현황과 대안:

재현데이터를 중심으로

김현태·장가영

KiRi 보험연구원

본 보고서에 수록된 내용은 집필자 개인의 의견이며 우리원의 공식 의견이 아님을 밝혀 둔다.

목 차

표 차례

그림 차례

Abstract

Current and Alternative Methods of Data Alias Processing(Pseudonymization) Techniques

Statistical disclosure control is a generic term for data processing methods for data privacy, including pseudonymization and anonymization techniques.

Although various disclosure control techniques have been proposed, they often present difficulties in achieving the desired objectives due to the loss of utility resulting from data integration or alteration in microdata. In light of these challenges, we propose synthetic data as an alternative approach to control disclosure risk. Synthetic data is generated by assuming a population that produces the observed data and employing statistical and machine learning models. We delve into the background, available techniques, recent theoretical developments, and case studies of synthetic data, emphasizing the importance of data quality assessment as a fundamental aspect of technical advancements and regulatory frameworks.

While synthetic data has started to be actively utilized in domains such as finance and healthcare in the United States and Europe, its adoption in South Korea faces practical challenges. With improved and more comprehensive regulatory guidelines to distinguish anonymous data from pseudonymous data, we believe that synthetic data can become a practically valuable and indispensable tool in South Korea's financial industry.

요 약

데이터 3법 시행 이후 보험을 포함한 금융업계 전역에서 데이터의 적극적 활용이 그 어느 때보다 중요해지고 있다. 데이터 활용을 위한 개인정보보호 과정에서 필요한 가명·익명처리를 포함한 데이터 가공 방법을 총칭해 통계적 노출 제어라 한다. 서로 상충하는 데이터의 효용과 정보보호를 적절히 수준에서 제어하는 것은 매우 중요한 문제이다.

현재 널리 사용되고 있는 마스킹기법은 그 종류가 다양하고 쉽게 적용할 수 있지만, 정보통합이나 변경에 따른 효용의 손실이 발생함으로써 분석의 목적에 맞게 사용하기가 어려운 측면이 있다. 최근 대안으로 제시된 차등적 정보보호기법 역시 쿼리와 분석의 종류에 의존하며 반복되는 쿼리에 대해 점차적으로 효율이 감소해 분석의 종류에 따라 매개변수의 값을 특정하기 어려운 단점이 있다. 암호화된 데이터의 연산을 통해 정보의 손실없이 개인정보를 보호할 수 있는 동형암호기법 역시 아직 분석의 복잡성에 따른 계산량의 문제, 그리고 쿼리의 종류와 범위에 따른 암호화 설계의 어려움이 존재한다.

본 보고서에서는 노출 제어 기법의 대안으로 재현데이터를 제안한다. 재현데이터는 관측된 원데이터를 생성하는 모집단을 가정하고, 통계적·기계학습모형을 통해 생성한 모의데이터이다. 재현데이터는 개인정보를 보호하면서도 원데이터와 비슷한 수준의 효용을 가지며, 익명데이터로 분류되어 개인정보 관련 규제로부터 자유롭다는 장점이 있다. 본 보고서에서는 재현데이터의 배경·기법·사례 등을 살펴보고, 실제 사용에 있어 기술과 규제의 기반이 되는 데이터 품질평가에 대한 중요성과 최근의 이론적 성과도 소개한다.

현재 재현데이터가 금융과 헬스케어 등의 영역에서 활발히 사용되고 있는 미국이나 유럽과 달리, 국내에서는 법적으로는 익명데이터이지만 실무를 위한 가이드라인 기준으로는 익명데이터로서 인정받는 것이 현실적으로 어려운 상황이다. 이를 위해 익명데이터의 정의에 좀 더 충실하고 포괄적인 새로운 가이드라인이 절실히 필요하다. 개선된 익명데이터 분류의 가이드라인이 뒷받침되어 재현데이터의 적극적 활용이 가능해진다면, 방대한 데이터가 수집되고 있는 우리나라의 금융데이터 산업에 큰 도움이 될 것으로 기대한다.

Ⅰ 서론

1. 연구 배경 및 목적

2018년 11월 신용정보법 개정안 발의 이후 2020년 1월 데이터 3법이 통과되었고, 2020년 8월부터 이 법이 시행됨에 따라 가명정보의 산업적 활용에 대한 법적 근거가 마련되었다. 이에 따라 개인식별이 어렵게 가공된 '가명정보'를 통계 작성, 공익적 기록 보존, 과학적 연구 등에 정보 소유자 사전동의 없이 사용할 수 있게 되었다. 즉, 연구자들이나 각종 산업 기관의 과도한 정보보호로 사용하지 못했던 개인정보를 기존 데이터에 결합하여 기존보다 더 많은 정보를 추출할 수 있게 된 것이다.

데이터 3법 개정의 핵심 내용은 다음과 같다. 첫째 개인정보보호법 개정을 통해 개인정보 범위가 해당 정보만으로는 특정 개인을 알아볼 수 없더라도 다른 정보와 쉽게 결합하여 알아볼 수 있는 정보로 명확화되었다. 이에 따라 기업 등에서 좀 더 구체적인 가이드라인을 통해 데이터를 활용할 수 있게 되었다. 또한 '익명정보(Anonymized data)'에 대한 개인정보보호법의 적용 제외를 명시하여 과거에는 개인정보로 간주하여 활용되지 못했던 정보들을 활용할 수 있는 가능성이 높아진 것은 큰 의미를 갖는다.

둘째, 신용정보법에서는 '가명정보(Pseudonymized data)'를 상업적 목적을 포함한 통계 작성, 연구, 공익적 기록 보존 목적으로 동의 없이 활용할 수 있도록 규정하여 빅데이터 분석에서 가명정보 활용이 이전보다 자유로울 수 있도록 영역을 확장했다. 특히 금융분야 마이데이터 산업의 도입을 허용해 금융분야 데이터 활용 혁신을 도모할 수 있도록 했다.

마지막으로 정보통신망법 주요 개정 내용은 개인정보보호 관련 사항을 '개인정보보호법'으로 이관하고 온라인상에서 수집된 개인정보보호 관련 규제와 감독 주체를 개인정보보호위원회로 이관한 것이다. 이는 법령 간 중복사항을 정리하여 활용성을 높인 점에서 의의가 크다고 하겠다.

〈표 Ⅰ-1〉 데이터 3법 개정 사항 요약

법률명	소관부처	내용
개인정보보호법	행정안전부	• 가명정보 개념 도입 및 동의 없이 사용 가능한 범위 구체화 • 가명정보 이용 시 안전장치 및 통제수단 마련 • 개인정보 관리·감독 체계를 개인정보보호위원회로 일원화
신용정보법	금융위원회	• 신용주체자의 본인 정보 통제 기능 강화 • 금융분야 빅데이터 분석 및 이용의 법적 근거 명확화 • 마이데이터(MyData) 도입 및 금융분야 규제 정비
정보통신망법	과기정통부	• 개인정보보호 관련 사항은 '개인정보보호법'으로 이관 • 온라인상 개인정보보호 관련 규제와 감독 주체를 방송통신위원회에서 '개인정보보호 위원회'로 변경

자료: 금융위원회(https://www.fsc.go.kr/index)

〈표 Ⅰ-1〉과 같은 법령 정비로 인해 기존보다 더 많은 정보를 활용하고 가공함에 따라 데이터를 활용한 경제, 즉 데이터 경제(Data Economy)[1]가 모든 산업에서 중요한 화두로 떠올랐다. 특히 보험회사의 경우, 상품개발, 인수심사, 요율개선, 위험관리, 마케팅 등 다양한 분야에서 데이터를 적극적으로 활용할 수 있을 것으로 기대되고 있다. 그 외 금융분야에서도 빅데이터를 활용한 서비스 혁신을 꾀하고 있으며, 이미 통신정보 납부, 고용정보 등의 비금융 데이터를 결합한 대안신용평가를 활용하여 새로운 은행 대출 심사 기준을 정립하는 등의 사례가 나타나고 있다.

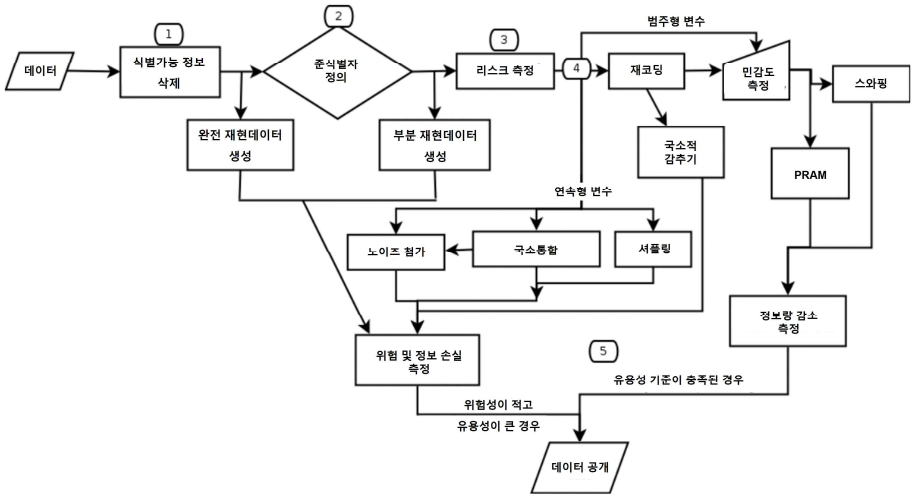
이와 관련하여 빅데이터의 활용 및 이종 데이터 결합 시 개인정보 노출 및 오·남용 방지를 위해서는 개인식별이 어렵도록 가공하는 가명·익명처리가 중요한 이슈로 대두되고 있다. 가공을 통해 개인정보 노출 정도를 조절하고 제한하는 기법을 총칭해 통계적 노출 제어[2](Statistical Disclosure Control)라 부르는데, 현재 통계적 노출 제어는 통계적인 기법과 딥러닝·AI 기법 등을 이용해 다양한 형태로 이루어지고 있다. 〈그림 Ⅰ-1〉은 통계적 노출 제어의 전체적인 과정을 간략히 나타낸 것이다. 이중 대중적으로 사용되고 있는 대표적인 방법은 마스킹(Masking)과 차등적 정보보호(Differential Privacy)이다. 그러나 기존의 마스킹이나 차등적 정보보호는 데이터의 중요한 내용을 숨기거나 뭉개버림으로써 의미 있는 정보의 손실을 야기하고 있어 새로운 대안으로 재현데이터(Synthetic data) 방법

1) 2011년 데이비드 뉴먼(David Newman)이 쓴 가트너 보고서("How to Plan, Participate and Prosper in the Data Economy")에 처음으로 나온 개념으로 데이터가 경제활동의 중요한 생산요소로 사용되는 경제구조를 말함

2) 통계청(2021a)에 따르면 통계적 노출 제어라는 용어와 비식별화 처리, 통계적 비밀 보호는 같은 의미로 통용됨

론이 제안되고 있다. '재현데이터'는 원데이터에 적절한 노이즈를 주거나 통계·빅데이터 알고리즘을 통해 원데이터의 구조적 특성을 유지하는 가상의 모의(Simulated) 데이터를 말한다. 재현데이터도 약간의 정보 손실을 감수해야 하지만 데이터가 뭉개지지 않는다는 점, 민감한 개인정보의 역추적(Re-identification)이 기술적으로 사실상 불가능해 익명데이터로 분류된다는 점, 그리고 이를 통해 자유롭게 공유, 전송, 사용이 가능하다는 장점이 있다.

〈그림 Ⅰ-1〉 통계적 노출 제어(Statistical Disclosure Control) 프로세스



자료: Meindl et al.(2013)

데이터 활용에 대한 기대가 어느 때보다 높아진 지금, 본 보고서는 데이터 통계적 노출 제어 기법에 대한 동향을 살펴보고 보다 효율적인 통계적 노출 제어 기법으로써 재현데이터를 소개하고자 한다. 구체적으로, 다양한 통계적 노출 제어 방법론들을 먼저 설명하고 한계점들에 대해 논할 것이다. 다음으로 기존 방법론의 정보 손실에 대한 해결책이 될 수 있는 재현데이터 방법론을 소개하고, 통계 소프트웨어인 R 패키지를 이용한 재현데이터 생성 방법을 예시로 제시하여 실무자들이 직접 재현데이터를 사용하는 데 도움이 될 수 있도록 할 것이다.

2. 연구 범위 및 방법

본 연구는 마이크로데이터에 적용되는 통계적 제어 방법론을 조사하고, 효과적으로 데이터의 통계적 노출 제어를 할 수 있는 대안으로 재현데이터 방법론을 제안한다.

마이크로데이터(MicroData)[3]란 통계조사 원데이터(Raw data)에서 개인정보, 입력 오류, 논리 오류 등을 수정한 개인, 가구, 사업체 등 조사단위별 관측 자료를 말한다. 데이터 수요자 입장에서는 마이크로데이터의 양이 많을수록 다양한 활용이 가능하지만 개인정보보호 문제로 인해 활용의 범위나 용도가 제한되는 경우가 많다. 따라서 데이터 제공자는 개인정보 노출 가능성을 최소화하기 위해 개인정보의 일부 또는 전체를 삭제·통합하거나 변수를 제한하고 데이터에 노이즈를 추가하는 등의 통계적 노출 제어 방법을 이용하게 된다.

연구 내용은 다음과 같다. 먼저 다음 절에서 가명처리, 익명처리 등의 용어를 정의하고 용어의 차이점을 설명하겠다. 다음 장인 본론에서는 가명처리 기법들을 검토하는데, 현재 가장 널리 사용되는 마스킹 기법을 변조(Perturbation)와 비 변조(Non-perturbation) 방식으로 나누어 살펴보고, 또 다른 가명처리 기법인 차등적 정보보호(Differential Privacy)를 검토한다.

다음으로 위 가명처리 기법들이 갖는 한계를 극복할 수 있는 대안으로 재현데이터 방법론을 소개한다. 선행 연구에 기반하여 재현데이터가 가지는 특징을 국내외 활용사례와 함께 검토한다. 현재 재현데이터 생성 방법은 통계적 방법과 딥러닝 방법으로 나뉠 수 있는데 본 보고서에서는 통계적 방법 위주로 그 생성이론을 정리하고, 추가로 자동차 보험 데이터를 이용하여 통계 소프트웨어인 R로 재현데이터를 생성해 볼 것이다.

이어 재현데이터에서 효용감소와 정보보호의 증가라는 상충관계를 설명하고 이와 연결해 재현데이터의 품질을 측정하는 척도와 최근의 학문적 성과에 대해 논한다.

마지막으로 재현데이터 사용을 위해 고려해야 할 점과 앞으로 해결해야 할 문제를 규제와 관련지어 검토하여 재현데이터 활용도를 높일 방안을 제시한다. 현재 재현데이터를 실무에서 바로 사용하기 위해 해결해야 할 이슈들이 있지만, 세계적으로 재현데이터 시장의 규모가 급속히 커지고 있어 금융을 포함한 전 산업에 있어 재현데이터의 광범위한 사용이

3) 통계청은 현재 MDIS(MicroData Integrated Service)를 통해 국민들이 다양한 통계자료를 편리하게 이용할 수 있도록 서비스하고 있으며 다양한 계층의 이용자를 위해 응답자 개인정보가 보호되는 범위 안에서 제한 없이 마이크로데이터를 제공하고 있음

곧 일반화될 것으로 기대한다.

3. 용어 정의[4)]

본론에 앞서, 용어에 대한 정확한 의미를 짚으려 한다. 먼저 가장 넓은 범위를 포괄하는 통계적 노출 제어(Statistical Disclosure Control)라는 용어는 개별정보보호를 구현하기 위한 방법들을 통칭한다. 가명처리 및 익명처리를 포함하여 본 보고서에서 자세히 다루려 하는 재현데이터를 생성해 원데이터를 대체하는 것도 통계적 노출 제어의 한 종류이다. 과거에는 비식별 처리, 비식별화라는 용어를 주로 사용하였으나 그 범위의 모호함 때문에 최근에는 사용하지 않는 추세이다.

데이터의 고유한 특성은 속성(Attribute)이라 하는데 속성은 식별자(Identifier)와 준식별자(Quasi-identifier)로 구분된다. 식별자는 주민등록번호, 이메일 주소, 휴대전화번호 등과 같이 그 자체로 특정 개인을 직접 식별하는 용도로 사용하는 속성을 뜻한다. 식별자는 노출 제어 과정 중 삭제되는 것이 일반적이다. 준식별자는 개인식별가능정보라고도 불리며 연령, 성별, 거주지역, 국적 등과 같이 해당 정보만으로는 직접적으로 특정 개인을 식별할 수 없지만, 다른 속성과 결합하여 특정 개인의 신원을 전부 또는 일부를 드러낼 수 있는 속성이다. 준식별자는 데이터에서 활용될 수 있지만 개인 식별 가능성이 높은 경우에는 가명처리된다.

가명처리와 익명처리의 차이를 아는 것도 중요하다. 자세한 내용은 V장에서 설명하겠고, 여기서는 쉽게 이해할 수 있는 수준에서 정의한다. 먼저 가명처리란 추가정보를 사용하지 않고는 특정 개인을 알아볼 수 없도록 데이터를 처리하는 것을 뜻하며, 가명처리한 개인정보는 가명정보라 한다. 즉, 어떤 정보에 성명, 주민등록번호, 계좌번호 등의 식별자를 포함하는 추가정보를 연결했을 때 특정 개인이 식별된다면 이 정보는 가명정보가 되는 것이다. 가명정보는 상업적 목적을 포함한 통계 작성, 산업적 연구를 포함한 연구, 공익적 기록보존 목적을 위해 사용할 경우 동의 없이 활용 가능하다. 하지만 규제상으로 볼 때 가명정보도 개인정보로 분류된다.

익명처리는 추가정보를 사용하더라도 더 이상 특정 개인을 식별할 수 없도록 처리하는 것

4) 개인정보보호법(https://www.law.go.kr/)

을 말한다. 익명처리된 정보를 익명정보라 하며 가명정보와는 달리, 정보 손실로 외부 데이터를 연결하는 것이 불가능하거나 연결해도 정확한 정보가 가려져 있는 상태이다. 익명정보는 더 이상 개인정보가 아니기 때문에 제한 없이 자유롭게 활용 가능하다는 점이 가명정보와 근본적으로 다르다. 〈그림 Ⅰ-2〉는 가명과 익명정보의 차이를 보여준다.

〈그림 Ⅰ-2〉 가명·익명정보 개념 및 활용 가능 범위

	개념	활용가능 범위
개인정보	특정 **개인에 관한 정보**, 개인을 **알아볼 수 있게 하는 정보**	**사전적**이고 **구체적인** **동의**를 받은 범위 內 활용 가능
가명정보	**추가정보**의 사용없이는 **특정 개인을 알아볼 수 없게 조치**한 정보	다음 목적에 **동의 없이** 활용 가능 (EU GDPR 반영) ① 통계작성 (상업적 목적 포함) ② 연구 (산업적 연구 포함) ③ 공익적 기록보존 목적 등
익명정보	**더 이상 개인을 알아볼 수 없게** (복원 불가능할 정도로) **조치**한 정보	개인정보가 아니기 때문에 **제한없이** **자유롭게 활용**

자료: 대한민국 정책브리핑(https://www.korea.kr/special/policyCurationView.do?newsId=148867915)

Ⅱ 데이터 노출 제어 기법 현황

이 장에서는 현재 많이 사용되는 데이터 가명·익명처리 방법인 마스킹(Masking)과 차등적 정보보호(Differential Privacy)에 대해 살펴보고 각각의 장단점에 대해 논의하겠다.

1. 마스킹(Masking) 기법

마스킹은 원자료의 적절한 변환을 통해 민감한 정보를 가리는 기법을 말하며, 마스킹된 자료란 변환된 자료와 변환에 관한 모든 정보를 의미한다. 이용자가 올바른 분석을 진행하기 위해서는 변환에 관한 정보를 보유하여야 하므로 함께 제공되어야 한다. 이해가 용이하다는 장점이 있지만, 마스킹 기법을 과도하게 사용하면 정보손실이 커지기 때문에 유의하여야 한다. 데이터마다 다른 특성을 가지고 있기 때문에 상황에 따라 적용되는 마스킹 방법 역시 달라져야 하는데, 보통 마스킹은 크게 변조(Perturbation), 비 변조(Non Perturbation) 방법론으로 나뉜다.

변조(Perturbation)는 데이터값을 억제하지 않고 실제 값 주변에 불확실성을 만들어 실제 원데이터값을 변경하는 방식인 반면, 비 변조(Non Perturbation)는 데이터 구조 자체를 왜곡하지 않고 특정 값을 대체하거나 억제해서 데이터의 세부사항을 가리는 방식이다. 변조와 비 변조 방법 모두 명목형 변수와 연속형 변수에 사용할 수 있다.

마스킹 기법은 확률론적 방법(Probabilistic method)과 결정론적 방법(Deterministic method)으로 나누어 생각할 수도 있다. 확률론적 방법은 확률 메커니즘 또는 무작위 난수 생성 메커니즘에 따라 달라지는데, 사용될 때마다 다른 결과가 생성된다는 특징이 있으며 이러한 방법의 경우 계속 같은 결과를 도출하고 싶으면, 난수 생성기에 시드(set.seed)를 설정해야 한다. 이에 반해 결정론적 방법은 지정된 특정 알고리즘을 정확히 따르기 때문에 동일한 데이터에 반복적으로 적용하더라도 동일한 결과를 도출하게 된다.

가. 변조(Perturbation)

변조는 데이터값을 억제하는 것이 아니라 특정 알고리즘이나 모형을 이용하여 실제 값을 변경하여 노출위험을 낮추는 방법이다. 즉, 원데이터의 값들이 변경되었기 때문에 준식별자 값의 빈도 수에 기초한 위험 측정은 변조를 적용한 후에는 더 이상 유효하지 않다.

〈표 Ⅱ-1〉 데이터 변조 예시

ID	원데이터			변조 이후 데이터		
	성별	지역	교육 수준	성별	지역	교육 수준
1	female	rural	high	female	rural	high
2	female	rural	high	female	rural	*low**
3	male	rural	low	*male**	rural	low
4	male	rural	low	female	rural	low
5	female	urban	low	*male**	urban	*high**
6	female	urban	low	female	urban	low

주: *는 원데이터에서 변조 이후 값이 변경된 부분을 뜻함
자료: SDC practical guide(https://sdcpractice.readthedocs.io/en/latest/index.html)

〈표 Ⅱ-1〉을 예로 들면, 데이터 변조 전과 후 모두, 모든 관측치는 $k = 3$ 수준에서 k-익명성을 위반하지만[5] 그럼에도 불구하고 각 관측치의 정확한 재식별 위험은 감소함을 알 수 있다. 침입자가 샘플에 유일하게 있는 조합인 {'male', 'urban', 'high'}를 원데이터와 맞추려 할 때, 원데이터에는 이러한 특성을 가진 개인이 포함되어 있지 않으므로 일치 관계를 찾아내기가 힘들다. 따라서 침입자는 해당 관측치의 다른 변수값이 올바른지 확신할 수 없게 된다.

변조의 장점은 수치를 없애버리는 것이 아니라 말 그대로 변화를 주는 것이기 때문에 비변조 방식에 비해 정보 손실이 덜하다는 것이다. 단점은 변조에 사용된 알고리즘이나 모형의 성능에 따라 변조된 데이터가 원데이터의 구조를 보존하는 정도가 달라지고, 외형상 원데이터와 흡사해 익명화 처리가 미흡하다는 인상을 줄 수 있다는 점이다.

5) 즉, 각 키가 데이터 세트에 두 번 이상 나타나지 않음. k-익명성에 대해서는 Ⅳ장에서 자세히 설명할 것임

1) PRAM(Post RAndoMization)[6]

PRAM은 범주형 변수에 주로 사용하며 제어하고자 하는 값에 의도적으로 노이즈를 첨가하는 방법이다. 재코딩과 국소적 감추기로는 노출 제어가 충분하지 않고 정보 손실이 큰 경우 이 방법을 사용할 수 있으며 일부 데이터 개체들에 미리 지정한 전이확률 행렬에 따라 다른 값으로 변환하는 기법이다.[7]

구체적인 행렬을 예시로 들어 PRAM을 설명하려 한다. 어떤 데이터셋에 여아를 출산한 산모가 30명, 남아를 출산한 산모가 50명, 사산아를 출산한 산모 2명에 대한 정보가 있다고 하자. 사산아를 낳은 산모의 수가 상대적으로 적기 때문에 추가정보와 결합 시 신원 노출이 일어날 수 있다. 이때 전이행렬 P를 목적에 맞는 수치로 지정해 PRAM을 적용할 수 있다. 전이행렬 P는 범주가 n개일 때 $n \times n$ 크기로 만들어지는 행렬 형태이며 원래 범주를 어느 정도 유지할 것인지 비율을 먼저 정한다. 이 비율은 행렬에서 대각 행렬 부분에 나타나는데 원데이터를 최대한 보존하고 싶다면 그 수치를 1에 가깝게 하면 된다. 데이터 수를 맞춰주기 위해서는 열별 합이 1이 되도록 지정해야 하며 대각행렬 비율 지정 후 나머지 범주에서 다른 범주로 얼마나 변경할지에 대한 비율 수치를 정한다. 이 행렬의 성분으로 들어가는 수치는 분석 목적과 정보보호 수준에 따라 자의적으로 정하게 된다. 즉, 프로그래밍 상에서 PRAM을 적용할 때, 확률생성 메커니즘에 따라 전이행렬이 다르게 생성되어 결과가 달라질 수 있다. 예를 들어 P를 다음과 같이 두자.

$$P = \begin{bmatrix} 1 & 0.05 & 0.17 \\ 0 & 0.8 & 0.03 \\ 0 & 0.15 & 0.8 \end{bmatrix}$$

이 행렬의 원소들은 자의적으로 정한 것이다. 먼저 첫째 열에서는 (1,1)번째 원소의 값을 1로 두고 나머지 (2,1)과 (3,1)의 원소 값을 0으로 지정해 열의 합을 1로 맞췄다. 나머지 대각원소인 (2,2)와 (3,3)은 0.8로 두어 원데이터에서 관측된 비율의 80%가 유지되도록 했다. 이 전이행렬의 전치(Transpose)를 원데이터 응답값 R_x와 내적해서 나온 다음 결과가 마스킹을 적용한 데이터가 된다.

---

6) Gouweleeuw, J. M., Kooiman, P., and De Wolf, P. P.(1998)

7) 김승현(2020)

$$P^T R_x = \begin{bmatrix} 1 & 0 & 0 \\ 0.05 & 0.8 & 0.15 \\ 0.17 & 0.03 & 0.8 \end{bmatrix} \begin{bmatrix} 30 \\ 50 \\ 2 \end{bmatrix} = \begin{bmatrix} 30+0+0 \\ 1.5+40+0.3 \\ 5.1+1.5+1.6 \end{bmatrix} = \begin{bmatrix} 30 \\ 41.8 \\ 8.2 \end{bmatrix} \approx \begin{bmatrix} 30 \\ 42 \\ 8 \end{bmatrix}$$

따라서 이 전이행렬 P를 이용해 데이터를 처리하게 되면 여아를 출산한 산모는 30명, 남아를 출산한 산모는 42명, 사산아를 출산한 산모는 8명으로 바뀌어 기록된다. 사산아 출산 산모가 2명에서 8명으로 늘어 정보 노출의 가능성이 줄어드는 것이다.

PRAM의 단점은 전이행렬의 개별 원소 값에 대한 결정이 자의적이고, {중학교 재학, 63세}와 같은 이상치 조합이 생성될 수 있다는 점이다. 따라서 PRAM을 적용할 때 특정 전이가 불가능하도록 설계가 되어야 한다. 예를 들어, 학교 재학 관측치의 경우 나이는 6세에서 18세 사이여야 하는 조건을 주고 PRAM을 수행해야 한다. 또는 나이별로 Strata를 구성하고 그 Strata 내에서만 PRAM이 이루어질 수 있도록 해야 한다.

2) 순위 스와핑(Rank Swapping)

순위 스와핑은 관측치 간 특정 변수 값을 교환하는 것에 기반을 두고 있는 변조 방법으로써 순서형 및 연속형 변수에 다 적용될 수 있다. 순위 스와핑에서는 대상이 되는 변수의 값들이 먼저 내림차순 또는 오름차순으로 정렬된다. 관측치 간 변수 값을 교환하려 할 때 순서대로 정렬된 데이터셋에서 교환 대상이 되는 행은 그 주변 몇몇 값들로만 바꾸어 변경에 제약 조건을 주는 것이다. 즉, 대상이 되는 변수 값은 원래 값 근처에 있는 동일하거나 유사한 값으로 교환되게 된다. 일반적으로 순위 스와핑은 정보 손실과 데이터 보호 사이 균형을 잘 지키면서 노출위험을 낮추는 좋은 방법이지만 다음과 같은 한계가 존재한다.

첫째, 대상 변수의 관측치 주변에 바꿀 수 있는 이웃 값이 많지 않거나 적은 범주만 있는 경우 또는 결측값이 많은 경우 순위 스와핑은 실행해도 값이 변경되지 않기 때문에 적절하지 않다. 예를 들어 '교육' 변수에 {초등, 중등, 고등, 대학}과 같이 소수의 범주만 있는 경우, 순위 스와핑보다는 다른 노출 제어 방법을 사용하는 것이 낫다.

둘째, 순위 스와핑은 침입자가 특정 변수 최고 또는 최저값이 누구에게 속하는지 아는 경우 적절하지 않다. 이는 값 자체가 변경되는 것이 아니라 원래 값이 모두 공개되기 때문에 순위 스와핑 이후 침입자는 이 변수 수준을 알 수 있게 되기 때문이다.

마지막으로 여러 변수에 동시에 순위 스와핑을 적용하게 되면 이후 변수 간 상관 구조

(Correlation structure)가 적절하게 유지되는지 확인해야 하는 작업이 필요하다.

3) 셔플링(Shuffling)

셔플링은 자료를 섞는 방식으로 스와핑과 유사하지만, 차이점은 변수에 대해 적절한 지도학습모형 혹은 분포추정기법을 사용하여 모형에서 추정된 값들의 순위에 따라 원변수를 다른 순서로 치환하는 방식을 취한다는 것이다.

셔플링은 대상이 되는 민감변수의 새로운 값을 나머지 변수들의 조건부 분포를 이용하여 생성한다. 예를 들어, 어떤 데이터에 '소득'과 같이 개인의 중요한 정보가 담겨있는 민감변수가 존재한다고 하고 이를 Y라 하자. 이때 나머지 변수인 '나이', '직업', '학력'을 독립변수로 하는 회귀모형을 적합하여 Y의 추정값을 산출할 수 있다. 이제 관측된 소득값들인 y_i를 순서대로 정렬해 $y_{(1)} < y_{(2)} < \cdots < y_{(n)}$로 두고 추정된 소득값들 $\hat{y}_i$에 대해서도 마찬가지로 순위를 모두 고려한다. 셔플링의 첫 단계로서, 가장 작은 추정값 $\hat{y}_{(1)}$이 자료의 i번째 레코드(혹은 i번째 관측값)에서 산출되었다고 하자. 그러면 i번째 원래의 Y값인 y_i를 $y_{(1)}$으로 치환한다. 두 번째 단계로서, $\hat{y}_{(2)}$가 자료의 j번째 경우에서 산출되었다고 하면 j번째 원래의 값인 y_j를 $y_{(2)}$로 치환한다. 이를 모든 순위에 대해 차례로 실행하면 원관측치들이 모두 치환이 되는데, 치환된 값들의 집합은 원래 값들의 집합과 동일하지만, 순서가 다르게 된다. 이런 의미에서 셔플링은 추정값의 순서에 기반해 원래 값을 재정렬하는 방식으로 이해할 수 있다. 따라서 주변분포에 대해서는 원데이터와 동일한 평균, 분산 등의 통계량을 얻을 수 있지만 공분산 구조는 유지되지 않는다는 단점이 있다. 위의 예를 확장하면 민감변수가 여러 개일 때 다변량(Multivariate) 회귀모형을 사용할 수 있으며 변수나 데이터의 종류에 따라 일반화선형모형(GLM) 혹은 다양한 코퓰라(Copula)를 적용한 모수·비모수적 분포추정 모형 역시 이용 가능하다.

4) 잡음추가(Noise Addition)

데이터의 원래 값에서 매우 작은 값을 더하거나 빼서 연속형 변수의 노출위험을 낮추는 방법이다. 평균이 0인 잡음을 추가하게 되면 연속형 변수에서 원데이터와의 정확한 일치를 방지하면서도 데이터 기본 형태는 대략적으로 유지할 수 있어 효과적인 노출 제어를

할 수 있다. 잡음을 추가할 때는 데이터 유형, 데이터의 사용 목적 및 잡음추가 전후 데이터의 분포적 특성(평균, 분산, 공분산, 상관관계 등)을 고려해야 한다. 다음은 잡음추가를 할 때 주의해야 할 사항들이다.

첫째, 제한 관계를 가지는 변수들에 잡음추가를 적용하면 그 구조에 오류가 생길 수 있다. 예를 들어 개인의 수입과 지출 데이터를 생각해보자. 수입과 지출의 세부 항목들은 총수입 또는 총지출에 합산되며, 원데이터에서 총수입 및 총지출은 그 세부 항목들의 합과 같아야 한다. 그러나 개별 세부항목과 총합에 잡음을 같이 추가하면 총계의 결과가 일치하지 않는다.

둘째, 극단값이 있는 데이터의 경우 잡음추가를 해도 극단값의 노출위험을 낮출 수가 없다. 이는 극단값의 경우 잡음추가 이후에도 여전히 극단값으로 감지되기 때문이다. 예를 들면 특정 지역에서 매우 높은 소득을 가지는 관측치는 이 소득값에 잡음을 첨가해도 여전히 해당 지역에서 가장 높은 소득으로 인식되어 재식별이 가능해질 수 있다. 만약 이러한 극단값의 노출 제어를 위해 분산이 큰 잡음을 추가하게 되면 데이터의 분포가 점점 원데이터와 다른 형태로 변해 정보 손실을 야기하게 된다.

셋째, 둘 이상의 변수에 잡음추가를 적용할 경우에는 데이터의 분포적 구조를 보존하기 위해 원데이터의 공분산에 비례하는 오차를 가지는 다변량(Multivariate) 잡음을 이용하는 것이 도움이 된다. 만약 개별 변수에 독립적으로 단변량 잡음을 추가한다면 주변 분포들의 성질은 유지가 되겠지만 변수 간 상관 구조가 왜곡되기 때문이다.

마지막으로, 변수가 특정 범위에만 존재할 때 이렇게 잡음을 더하는 방식은 적절하지 않을 수도 있어 알맞은 변형이 필요하다. 예를 들어, 양의 변수에 잡음을 더하면 음수값이 생성될 수 있는데 이 경우 음수로 생성된 모든 관측치를 0으로 대체할 수 있다. 그러나 특정 임계값 미만의 값이 잘리는 형태가 되기 때문에 데이터의 분포가 왜곡된다. 이 경우 대안으로서 변수에 기댓값 1과 적절한 양의 분산을 갖는 잡음을 곱하는 방식을 이용할 수 있다.

5) 국소통합(Microaggregation)[8)]

국소통합은 일반적으로 연속형 변수에 적용되는 변조 기법이다. 그러나 군집을 형성할 수 있고 군집값에 대한 집계값을 계산할 수 있는 경우에는 범주형 데이터에도 국소통합을 적용할 수 있다.

국소통합은 연속형 혹은 범주형 변수를 적절히 그룹화한 후 그룹별로 관측치의 총계 값을 구해 원데이터값을 대체하는 방법이다. 총계 값은 보통 각 그룹의 평균을 사용하고 회귀분석을 이용해 통계량을 구하는 등 다양한 방법으로 구할 수 있다. 범주형 변수에 국소통합을 사용하는 경우, 최빈값을 사용하는 것이 일반적이다. 또는 연속형 변수에 이상치가 있는 경우에도 중앙값 대체가 더 적절하다.

나. 비 변조(Non-Perturbation)

1) 재코딩(Recoding)

재코딩은 변수에 대한 범주 수 또는 연속형 변수 수를 줄이는 데 사용되는 결정론적 방법이다. 이 작업은 범주형 변수 범주들을 결합 또는 그룹화하거나 연속형 변수에 대한 구간을 생성하는 방식으로 수행된다. 재코딩은 특정 변수의 모든 관측치에 적용되며 노출위험이 있는 관측치에만 적용되는 것이 아니다. 재코딩은 글로벌 재코딩(Global Recoding)과 상·하위 코딩(Top and bottom Coding)으로 나눌 수 있다.

가) 글로벌 재코딩(Global Recoding)

글로벌 재코딩은 범주형 변수에서 범주들을 합하거나 연속형 변수에서 관측값들을 구간별로 나누는 방법이다. 이 방법은 희귀한 변수 값을 가진 관측치들의 노출위험을 줄일 수 있지만, 세부적인 정보들이 범주 안에 통합되어 버리기 때문에 정보의 손실을 피할 수 없다.

예를 들어, 〈표 Ⅱ-2〉와 같이 데이터셋에 5개의 지역 값(지역 1~5)이 있다고 가정해보자. 일부 지역에 속하는 관측치는 매우 적기 때문에 여기에 속하는 관측치들은 높은 재식별

8) 박민정(2016)

위험을 갖게 된다. 이 경우 일부 지역을 재코딩하여 노출위험을 낮출 수 있다.

〈표 Ⅱ-2〉 글로벌 재코딩 예시

ID	재코딩 이전 데이터				재코딩 이후 데이터			
	성별	지역	성적	빈도(f_k)	성별	지역	성적	빈도(f_k)
1	female	지역 1	A	1	female	북부	A	3
2	female	지역 2	A	2	female	북부	A	3
3	female	지역 2	A	2	female	북부	A	3
4	female	지역 3	B	2	female	동부	B	2
5	male	지역 3	B	1	male	동부	B	2
6	female	지역 3	B	2	female	동부	B	2
7	male	지역 3	C	2	male	동부	C	2
8	male	지역 4	C	2	male	남부	C	3
9	male	지역 4	C	2	male	남부	C	3
10	male	지역 5	C	1	male	남부	C	3

자료: SDC practical guide(https://sdcpractice.readthedocs.io/en/latest/index.html)

이를테면 다섯 개의 지역을 {'북부', '동부', '남부'} 세 개의 그룹으로 나누어 재코딩하는 식이다. 이렇게 하면 그 변수 범주 수가 5개에서 3개로 줄어들면서 침입자는 각 준식별자에 대해 최소 2 이상의 빈도를 찾을 수 있게 되어 1~3번 관측치, 4번과 6번 관측치, 5번과 7번 관측치, 8~10번 관측치를 구분할 수 없게 된다. 즉, 원데이터에 재코딩을 적용함으로써 관측치 준식별자 조합에 대한 빈도 수(f_k)를 높이고 노출위험을 감소시킬 수 있다.

여기서 주의할 점은 재코딩 이후 나누어지는 범주 수는 임의적이라는 것이다. 몇 가지 범주로 나눌지는 통계적 노출 제어를 행하는 당사자가 적절히 선택해야 하며, 데이터 사용의 목적에 따라 정보 손실을 최소화하도록 주의해야 한다. 예를 들어, 나이 변수(Age variable)에 대해, 분석가가 6~11세와 12~17세의 학교에 다니는 아동·청소년에 대한 지표를 계산함과 동시에 노출위험도 줄이기 위해 연령을 그룹화해야 하는 경우, 이 분석에 적용할 수 있는 연령 간격을 신중하게 만들어야 한다. 적절한 그룹화는 0~5세, 6~11세, 12~17세 등이 될 수 있는 반면, 0~10세, 11~15세, 16~18세 그룹화는 데이터 유용성

(Data Utility)을 해칠 수 있다. 그룹화할 때는 너비가 동일한 구간을 만드는 것이 보통이지만, 필요에 따라 변수 일부만 그룹화하고 나머지는 그냥 두어도 된다. 예를 들어, 20세 이상의 모든 연령에 그룹화를 적용하고, 20세 미만은 원자료 값을 유지하는 식이다.

나) 상·하위 코딩(Top and bottom Coding)

상·하위 코딩은 분포 또는 범주의 위쪽 또는 아래쪽, 즉 극단값들만 재코딩하는 방법이다. 순서형, 범주형, 연속형 변수에 다 적용할 수 있으며 특히 나이와 소득 관련 변수에서 극단값으로 인해 신원 노출이 일어날 수 있을 때 자주 사용된다. 예를 들어, A동에 자산 500억 원 이상의 인물이 1명 있다면, 'A동 거주, 자산 500억 원 이상'의 정보를 입력하면 그 인물이 특정된다. 외부 데이터와 매칭 성공 시 질병과 같은 세부적인 다른 정보들도 노출되어 문제가 발생할 수 있다.

상·하위 코딩은 극단값을 극단적이지 않은 관측값들과 합쳐서 범주형으로 제공한다. 예를 들어 99세라는 극단값은 '90세 이상'의 범주에 넣으면 된다. 일반적으로 범주 내 관측치가 적을수록 노출위험이 높아지므로 분포의 끝에 있는 값을 하나의 범주로 그룹화하면 극단적 관측치의 노출위험이 감소하게 된다.

그룹화할 관측치의 임계치(Threshold)를 결정하려면 변수의 전체 분포를 확인하여 범주 내 관측치 개수가 특정 빈도 아래로 떨어지는 지점을 찾아내야 하는데, 데이터의 사용 목적과 노출위험의 허용 정도를 고려해서 결정한다.

2) 국소적 감추기(Local suppression)

국소적 감추기는 범주형 변수에 사용하는 방법으로 재코딩 이후에도 노출위험이 있을 시 사용한다. 보통 데이터에서 몇몇 변수 조합들은 빈도가 낮게 나타나게 되는데 이는 그러한 조합을 가진 관측치들의 노출위험이 높아짐을 뜻한다. 이때 민감변수 전체를 삭제하는 것이 아니라 일부 노출위험이 높은 변수의 값들(즉, 테이블 형태의 데이터에서는 특정한 행들)을 삭제하는 국소적 감추기 방식을 취한다.

하지만 연속형 변수에서는 고유한 수치 값이 과도하게 많아지므로 국소적 감추기가 적절하지 않을 수도 있다. 이는 범주형 변수에서 범주의 수가 너무 많은 경우도 마찬가지이다.

이럴 때는 앞서 설명한 재코딩으로 적절한 수의 범주로 먼저 나눈 후 국소적 감추기를 수행할 수 있다.

연구에 따라 사용되는 중요 변수의 노출 제어 혹은 억제를 최대한 줄여야 하는 경우가 있다. 예를 들어, 나이가 중요한 변수인 연구에서 나이 변수에 국소적 감추기가 적용되면 중요한 정보의 손실이 매우 커지는데, 이러한 경우 변수별로 중요도를 측정해 최적의 억제 패턴을 지정하는 것이 좋다. 예를 들면, 나이 변수를 배제한 나머지 변수들에 국소적 감추기를 적용하여 노출위험을 줄이는 것이다. 이렇게 되면 국소적 감추기의 적용 횟수 자체는 많아지지만 나이 변수를 가능한 원래의 값으로 보존할 수 있다.

〈표 Ⅱ-3〉 마스킹 종류 및 적용되는 변수 형태

마스킹 방법	분류	적용되는 변수 형태
글로벌 재코딩	비 변조, 결정론적 방법	연속형, 범주형
상·하위 코딩	비 변조, 결정론적 방법	연속형, 범주형
국소적 감추기	비 변조, 결정론적 방법	범주형
PRAM	변조, 확률론적 방법	범주형
국소통합	변조, 확률론적 방법	연속형
노이즈 첨가	변조, 확률론적 방법	연속형
셔플링	변조, 확률론적 방법	연속형
순위 스와핑	변조, 확률론적 방법	연속형

〈표 Ⅱ-3〉은 앞서 언급한 마스킹 방법론을 정리한 것이다. 목적과 변수 형태에 따라 적용되는 마스킹 방법이 달라진다는 것을 명심하고 적절한 방법을 선택해야 한다.

2. 차등적 정보보호(Differential Privacy)

차등적 정보보호(Differential Privacy, 이하, 'DP'라 함)는 알고리즘이라기보다 수학적 개념이다. 데이터베이스 안에 여러 요소들이 있을 때, 침입자가 쿼리(Query)[9]를 통해 특정 요소에 대한 정보를 알아내는 것을 막아낸다는 개념으로 개인정보보호와 연관이 있고, 이를 통해 다양한 방법론이 파생될 수 있어 가명처리 기법의 일환으로 이해할 수 있다.

DP는 공학 분야에서 2006년 처음 제시된 개념[10]으로 어떤 원데이터에 대해 특정 관측값 하나를 제외했을 때의 쿼리 결과가 원데이터에서 얻은 결과와 충분히 유사하다면 정보보호가 되었다고 여기는 것에서 출발한다.

DP를 수학적으로 정의하기 위해 두 개의 데이터 D_1과 D_2를 고려하자. 두 데이터는 동일하고 다만 특정 관측값 하나만 차이가 난다고 가정한다. 즉 둘 중 어느 한 데이터는 이 관측값이 빠져있다. 이제 특정 결과를 얻기 위한 쿼리를 K라고 하면, 이 쿼리를 각 데이터에 적용한 결과인 $K(D_1)$와 $K(D_2)$가 동일한 분포 S에 속할 확률의 비율을 일정 수준보다 작게 하는 것이 DP의 정의이다. 이를 수식으로 표현하면 어떤 양수 ε에 대해

$$\frac{\Pr\left[\mathrm{K}(\mathrm{D}_1)\in \mathrm{S}\right]}{\Pr\left[\mathrm{K}(\mathrm{D}_2)\in \mathrm{S}\right]} \leq \exp(\varepsilon),\ \text{for } all\ S \subseteq \mathrm{range}(\mathrm{K})$$

이며 이 조건을 만족하면 해당 관측값은 ε 수준에서 정보보호가 보장된다고 말한다. 다시 말해 쿼리의 입장에서 두 데이터는 구분이 불가능한 것이다. 매개변수 ε는 노출허용의 범위를 결정하는데 ε이 작으면 두 데이터 분포 차이가 작아 강한 정보보호를 실현할 수 있지만, 유용성이 떨어지는 반면, ε가 크면 두 데이터 분포 차이가 커 정보보호 정도가 작으나 원본과 바뀐 데이터가 비슷해 유용성이 높아지게 된다. 따라서 이 상충관계를 잘 고려해서 쿼리 K에 대한 적절한 ε의 크기를 설정하는 것이 중요하다.

위의 조건을 만족하는 대표적인 기법으로 라플라스 메커니즘인데, 쿼리의 결과에 라플라스 분포에서 추출한 잡음(Noise)를 더하는 방식이다. 라플라스 메커니즘에서 잡음의 크기

9) 쿼리는 데이터베이스에 특정한 정보를 보여달라는 사용자의 요청을 말함. 넓은 의미에서 데이터 분석도 쿼리로 이해할 수 있음

10) Cynthia Dwork et al.(2006)

는 민감도(Sensitivity)와 프라이버시 모수(Privacy) 값과 관련이 있다. 민감도는 쿼리가 해당 데이터에서 얼마나 민감한가를 나타내며 민감도가 클수록 더 큰 잡음을 추가해야 원데이터와 변형된 데이터에 쿼리를 적용했을 때 두 데이터의 구분이 불가하게 만들 수 있다. 프라이버시 모수는 위 식에서 ε으로 개인정보보호 취약 수준을 나타내는 모수로 커질수록 개인정보보호 정도가 약해진다. 최근에는 라플라스 메커니즘 이외에도 가우시안 메커니즘(Gaussian mechanism), 딥러닝을 이용한 메커니즘 등 다양한 방법론들이 연구되고 있다.[11)]

현재 Google, Microsoft, Meta 등 글로벌 기업들은 개발자들이 DP를 활용할 수 있도록 관련 오픈소스를 제공하고 있다. 대표적으로 Google은 자사의 머신러닝 학습 및 개발 프레임워크 TensorFlow의 내부 모듈로서 TensorFlow Privacy를 2019년에 공개해 DP 기법을 통해 사용자 데이터를 보호할 수 있도록 지원하고, 지속적으로 개인정보보호를 위한 실험용 모듈을 공개하고 있다. 특히 이 Privacy 모듈은 멤버십 추론 공격을 방어하기 위해 개인정보 침해 대상이 되는 특정 샘플이 인공지능 학습용 데이터에 포함되어 있는지를 판정하는 분류기를 구축했다는 점에 의의가 있다.

DP 방법론은 마스킹 기법의 한계를 보완하는 획기적인 개념이지만 한계점이 존재한다. 첫째, 주어진 데이터에 반복적으로 적용되는 쿼리에 대해서는 그 효율이 떨어져 실제 사용에 있어 만족할만한 수준의 보호를 보장하기 위해서는 상당한 크기의 잡음을 더해 데이터의 효용을 훼손해야 하는 것으로 알려져 있다. 둘째, DP 메커니즘의 구현이 데이터의 종류 및 특성, 쿼리와 분석의 종류, 그리고 데이터를 보유한 조직의 개인정보보호 요구 사항에 따라 다르게 되어 일관성이 떨어진다는 점이다. 따라서 목적에 따라 적절한 DP를 구현하기 위해서는 상당한 비용과 시간이 필요할 수 있다. 셋째, 주어진 쿼리에 대해서도 그 복잡성에 따라 적절한 매개변수 ε의 값을 다르게 결정해야 하는데 이 또한 어려운 문제이다. 마지막으로, 크기가 작은 데이터에 DP 메커니즘을 적용하기가 어렵다는 점을 들 수 있다.

11) 자세한 내용은 Abadi, Martin, et al.(2016); Ji, Zhanglong, Zachary C. Lipton, and Charles Elkan(2014)를 참고 바람

3. 한계 및 대안의 필요성

이 장에서는 현재 널리 사용되는 가명·익명처리 기법들에 대해 살펴보았다. 마스킹 기법은 몇 가지 분명한 한계점이 존재한다. 우선, 마스킹된 데이터가 완전히 익명화되지는 않는다. 즉 마이크로데이터의 경우 일부 데이터 행에서 여전히 개인이 식별된다는 의미이다. 예를 들어, 마스킹 기법을 사용한 후에도 나이, 성별, 직업, 지역 등의 정보를 조합하면 여전히 개인을 식별할 가능성이 존재한다. 또 다른 한계점으로, 마스킹된 데이터는 원데이터보다 유용성이 떨어질 수 있다는 점을 들 수 있다. 이는 물론 모든 가명·익명처리 기법에 적용되는 단점이겠지만 마스킹의 수준이 높아질수록 원데이터의 중요한 정보들이 훼손되어 어느 시점에서는 더 이상 쿼리나 분석을 통해 의미있는 결과를 얻지 못하게 된다. 현재 국내의 익명데이터로 분류되는 정보들이 대표적인 예이다.

DP의 경우에도 다양한 방향으로의 이론적 확장과 일반화가 진행되고 있지만 위에서 서술한 한계점들로 인해 아직 광범위하게 사용되고 있지는 않다.

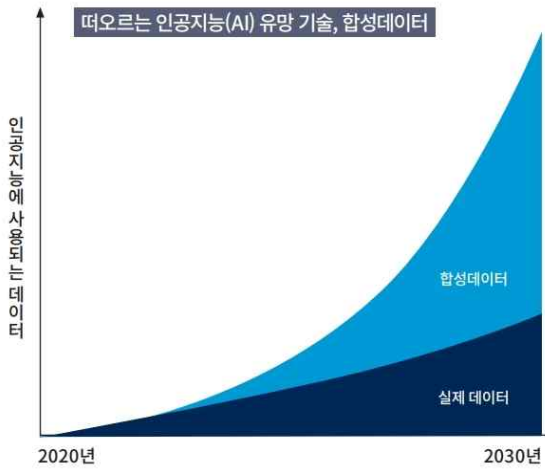
이러한 기존의 방법론들을 대신할 수 있는 가명·익명처리를 위한 기술들에 대해 개인정보보호위원회(2022. 4)의 가명정보 처리 가이드라인은 부록에서 표본 추출, 해부화, 동형비밀분산(혹은 동형암호기법)등을 제시하고 있다. 이 중 데이터의 훼손을 최소화하면서도 높은 수준의 개인정보보호가 가능한 기술은 동형암호기법과 재현데이터 정도이다. 나머지 기술들의 경우 원데이터의 무작위 복원 추출이나 데이터의 분리를 이용함으로 사용의 범위에 있어 제한적이다.

최근 주목을 받고 있는 동형암호기법은 나머지 방법론과 전혀 다른 방식으로 개인정보를 보호한다. 동형암호의 개념은 1970년대 논문으로 소개되었고 이후 2009년 크레이그 젠트리(Craig Gentry)의 연구에 기반해 가명·익명정보처리를 위한 기술로써 사용되고 있다. 동형암호기술은 원데이터를 암호화한 후 그 상태에서 다양한 쿼리와 분석을 진행하는 방식을 취한다. 암호화된 데이터를 이용해 분석하고 결과를 얻는다는 점에서 개인정보보호가 유지되면서도 데이터의 정보는 그대로 보존되는 장점을 가지지만, 아직 분석의 복잡성에 따른 계산량의 문제, 그리고 쿼리의 종류와 범위에 따른 암호화 설계의 어려움이 존재한다고 알려져 있다.

이러한 배경에서 본 보고서는 기존 방법론에 대한 대안으로 재현데이터에 대해 주목하고자 한다. 재현데이터(Synthetic data)[12]는 익명데이터로서 원데이터의 주요 통계적 특성

들을 유지하면서 생성해 낸 모의 데이터이기 때문에 특정 개인을 매칭하는 것이 불가능하고 데이터가 사외로 유출되는 경우에도 큰 충격이 없다. 방법론적으로도 민감변수와 식별변수를 선택적으로 배제할 수 있고 개발된 기법들을 쉽고 적은 비용으로 적용할 수 있다는 점에서 현실적인 대안으로 적합하다.

〈그림 Ⅱ-1〉 재현데이터의 미래



자료: Gartner(https://www.gartner.com/en)

미국의 정보 기술 연구 및 자문 회사 가트너는 2021년 6월 Maverick Research을 인용한 〈그림 Ⅱ-1〉에서 보이듯이 2030년경에는 모형의 학습에 사용되는 데이터 대부분이 합성(재현)데이터가 될 것으로 예측하고 있다. 재현데이터에 대한 자세한 내용은 다음 장에 이어 다루겠다.

12) 인공지능 분야에서는 합성데이터라고 부르며, 비정형데이터를 이용한 딥러닝모형의 학습에 널리 사용되고 있음

Ⅲ 재현데이터

본 장에서는 기존 가명·익명처리 기법의 대안으로 재현데이터를 제시한다. 재현데이터는 정형데이터뿐 아니라 이미지, 텍스트 등의 다양한 비정형 데이터에도 적용되고 있다. 본 장에서 재현데이터의 역사와 특징에 대해 논하고 재현데이터의 통계학적 및 딥러닝 기반 생성 방법론을 설명한다. 더불어 현재 국내외 재현데이터의 활용사례도 살펴보겠다.

1. 주요 방법론

재현데이터는 실제로 관측된 데이터(Real Data)를 생성하는 모형 혹은 모집단이 존재한다고 가정하고, 통계적 방법이나 기계학습 방법 등을 이용해 추정된 모형에서 새롭게 생성한 모의 데이터(Simulated Data)이다. 개인정보 노출을 막는 효과적인 방법이며 민감정보를 활용할 수 있어 연구자들이 재현데이터를 활용하여 보다 세밀한 분석을 진행할 수 있다는 이점을 갖는다. 다양한 규제로 인해 데이터의 적극적 활용에 제한이 가해진 현 상황에서 재현데이터는 노출 제어의 효과적인 방법론으로 부상하고 있다.

재현데이터는 하버드대학교 통계학과 루빈 교수가 미국 정부 기관 프로젝트로 빈곤층에 대한 과소평가와 같은 문제들을 해결하는 연구 과정에서 처음 제시하였다.[13] 루빈 교수는 모집단에서 관측되지 않은 자료를 결측값으로 간주해 다중 대체법(Multiple imputation)을 적용했고 이를 반복적으로 샘플링하여 다수의 데이터셋을 생성했는데, 이렇게 만들어진 데이터셋을 재현데이터(Synthetic data)로 지칭하였다. 이후 재현데이터 관련 연구는 다양한 방향으로 확장되어 왔다.[14]

13) Rubin, D. B.(1993)
14) 유성준·박나리(2020)

가. 분류

재현데이터는 재현 범위와 방식에 따라 완전 재현데이터(Fully Synthetic Data)와 부분 재현데이터(Partially Synthetic Data)로 분류할 수 있다.

1) 완전 재현데이터(Fully Synthetic Data)

완전 재현데이터는 측정된 실제 데이터를 기반으로 모든 변수를 재현하여 가상으로 생성된 데이터를 의미한다. 정보보호 측면에서 가장 강력한 보안성을 가지며 제공되는 데이터가 실제 데이터를 포함하지 않으므로 민감정보가 노출되지 않는 구조이다. 루빈 교수가 최초로 정의한 개념으로 다중대체 기법을 기반으로 하고 있다. 완전 재현데이터는 보통 민감정보이거나 공개 불가능한 내용을 포함하고 있는 부분을 모두 결측치(Missing value)라고 간주하고 다중 대체(Multiple imputation)[15]를 적용한다. 다음으로, 완성된 모집단을 모형에 적합하여 무작위 추출(Random sampling)한 후 재현데이터를 생성하는 방식을 취한다. 관측치별로 정보 노출 방지의 안정성을 확보할 수 있고 올바른 모형을 사용하면 데이터 구조가 원본과 비슷하게 유지될 뿐만 아니라 마스킹과 같은 다른 가명·익명처리 기법보다 정보 손실이 적다는 장점이 있다.

2) 부분 재현데이터(Partially Synthetic Data)

부분 재현데이터는 공개하려는 변수 중 일부만을 선택하여 재현데이터로 대체한 데이터를 말하며 Little[16]이 최초로 제안했다고 알려져 있다. 완전 재현의 경우 생성 데이터의 차원이 과도하게 커져 과적합이 일어나거나 변수 중요도가 반영되지 않을 수 있는데, 부분 재현을 통해 이러한 이슈들을 해결할 수 있다. 부분 재현 시 대체되는 변수들은 보통 민감변수(Sensitive variable) 혹은 식별 변수(Identifiable variable)가 되지만 사용자 임의대로 또는 분석 목적에 따라 선택 가능하다. 일반적으로 노출위험이 높거나 공개 불가능한 정보들을 임의로 선택하여 그 변수에서만 값을 대체하기 때문에 정보 손실이 적고 데이터 구조가 이전과 비슷하게 유지될 수 있다.

---

15) Li, P., Stuart, E. A., and Allison, D. B.(2015)

16) Little, R. J.(1993)

완전 재현데이터와 또 다른 차이점은 생산 가능한 데이터셋의 수가 다르다는 것이다. 완전 재현데이터는 데이터를 무한정 생산할 수 있지만, 부분 재현데이터는 선택한 일부 변수만 채우기 때문에 기존 데이터와 동일한 크기의 데이터를 생산하게 된다.

또 다른 재현데이터의 범주로서 복합 재현데이터(Hybrid Synthetic Data)를 들 수 있다. 이는 완전 재현과 부분 재현 방법 둘 다 차용해서 생성하는 데이터를 묶어서 부르는 단어로, 사실 복합 재현데이터의 정의와 범위는 아직 정확히 정해지지 않고 있다. 다만, 복합 재현데이터는 일부 변수들의 값을 재현데이터로 생성한 후 남은 실제 데이터를 이용해 또 다른 변수들의 값을 다시 재현데이터로 생성하는 방법을 지칭하거나, 이미지 데이터에서 배경은 실제 이미지를 쓰고 사물은 재현된 합성 이미지를 사용해 새로운 이미지를 생성하는 방식을 일컫기도 한다.

나. 생성이론

재현데이터는 다음과 같은 개념적 단계를 통해 생성한다.[17]

(1) 데이터 스키마(Data schema) [18] 정의: 원데이터의 구조, 데이터 유형, 관계 및 데이터에 적용되는 제약 조건 또는 규칙을 정하고 도식화한다.

(2) 원데이터의 통계적 특성 확인: 원데이터가 가지는 분포, 상관관계 및 기타 통계적 패턴 등의 정보를 미리 확인한다. 이 정보는 실제 데이터를 분석하거나 데이터에 대한 사전 지식을 조사하여 기록해둘 수 있고 추후 재현데이터 생성 과정에서 활용될 수 있다.

(3) 적합한 재현데이터 생성 방법 선택: 재현데이터를 생성하는 알고리즘이 다양하므로 데이터 스키마 및 통계적 특성에 따라 생성 방법론 채택이 달라져야 한다. 따라서 어떤 방법을 사용하여 어느 유형의 재현데이터를 생성할지 결정하는 과정이 필요하다.

(4) 재현데이터 생성(Generation): 선택한 방법을 사용하여 재현데이터를 생성하고, 그 과정에서 원데이터의 통계적 특성과 일치하도록 보장하는 단계이다.

(5) 재현데이터의 품질 평가(Evaluation): 생성된 재현데이터를 원데이터와 비교·평가하

17) https://gretel.ai; https://limoss.london/how-does-sdc-work

18) 데이터베이스를 구성하는 데이터 개체(Entry), 속성(Attribute), 관계(Relationship) 및 데이터 조작 시 데이터값들이 갖는 제약 조건 등에 관해 전반적으로 정의함을 나타내는 용어임. 즉, 데이터베이스를 어떻게 설계할지에 대한 계획을 짜서 구조와 제약 조건을 정하는 것임

는 과정이다. 통계적 검정, 시각화 또는 원데이터와의 단변량 및 다변량 비교를 통해 평가할 수 있다. 품질이 기준에 미치지 못하거나 필요한 특성을 충분히 반영하고 있지 않다면, 데이터 스키마를 수정하거나 다른 생성 방법을 사용하여 위 과정을 반복할 수 있다. 동시에 민감한 개인정보의 노출위험이 충분히 제어되었는지도 같이 평가해야 한다.

(6) 재현데이터 배포 및 사용: 재현데이터의 품질이 검증되면, 머신러닝 학습 모델 훈련 및 테스트, 데이터 증강, 민감한 정보를 제외한 데이터 공유 등 다양한 용도로 사용될 수 있다.

위의 단계 (4)에서 재현데이터의 생성은 Python과 R 등의 다양한 소프트웨어 툴을 이용하여 구현할 수 있는데 이 중 오픈소스로 공개된 솔루션도 있고 상용화되어 내부적으로 어떤 모형·알고리즘을 사용하는지 알 수 없는 경우도 있다. 〈표 Ⅲ-1〉은 재현데이터 생성 소프트웨어들을 조사한 결과이다. 표에 포함되지 않은 상용 소프트웨어들도 많은데 2022년 기준으로 재현데이터 생성 서비스를 하는 크고 작은 업체는 대략 100여 개 이상으로 추산된다.[19]

〈표 Ⅲ-1〉 재현데이터 생성 관련 소프트웨어

소프트웨어	관련 홈페이지 URL	내용
synthpop	https://www.synthpop.org.uk/	분류 회귀모형을 사용하여 재현데이터에 대한 변수를 생성함. 정교한 샘플링 디자인을 필요로 하거나 가구 및 구성원 정보 등과 같은 계층 또는 클러스터 구조를 가진 데이터를 처리하는 기능은 없으나 편의성이 높음
sms	https://cran.r-project.org/web/packages/sms/index.html	주어진 영역 내 매크로 데이터로부터 마이크로 데이터를 시뮬레이션하는 기능을 제공함. 계층적 구조의 데이터 처리는 불가능하나, Simulated Annealing을 단순화하여 제한된 영역에 대한 설명을 최적화하는 기능이 존재함
simPop	https://cran.r-project.org/web/packages/simPop/index.html	주체가 가진 속성값에 따라 다르게 적용되는 정책의 거시적인 효과를 예측하기 위한 복잡한 구조의 데이터 재현에 매우 유용함. 가구와 가구 구성원 정보 등 계층적 구조 처리가 가능함. IPF와 SA를 사용한 통계량 조정, 로지스틱 회귀를 통한 모델링 기능을 제공함

19) 업체들의 목록은 다음의 링크를 참고함(https://elise-deux.medium.com/new-list-of-synthetic-data-vendors-2022-f06dbe91784)

〈표 Ⅲ-1〉 계속

소프트웨어	관련 홈페이지 URL	내용
PoPGen	https://www.mobilityanalytics.org/popgen.html	Arixona State University의 SimTRAVEL Research Initiative에서 개발되었으며, 상대적으로 전수 정확도가 높은 반복비율갱신(Iterative Proportional Updating) 알고리즘으로 전수 인구 데이터 생성이 가능함
TRANSIMS	https://sourceforge.net/projects/transims/	미국 Los Alamos National Laboratiry의 연구원이 개발한 운송 분석 시뮬레이션 시스템임. 인구조사 마이크로 데이터를 기반으로 재현데이터를 생성함
Synthia	https://synthia-dataset.net/	비영리 연구기관인 RTI에서 개발한 웹 기반 재현데이터 생성 프로그램으로, 사용자 정의 변수를 사용하여 사용자가 정의한 학습 영역에 대한 재현데이터를 생성함
SDV	https://sdv.dev/	datacebo에서 관리하는 재현데이터 생성 공개소스코드로, 개별, 관계형, 시계열 데이터에 대한 재현데이터를 생성함. 생성된 데이터에 대한 평가 및 시각화 모듈도 제공함. 소스코드 공유 사이트 github에서 1,500개 이상의 star를 받을 정도로 높은 대중적 인지도를 가지고 있음
DataSynthesizer	https://github.com/DataResponsibley/DataSynthesizer	차등보호(Differential privacy) 기술을 활용한 재현데이터 생성 오픈소스코드로, Django를 활용한 UI 앱도 제공함
Gretel	https://gretel.ai/	미국 샌디에이고에 위치한 재현데이터 생성 스타트업으로 LSTM, DGAN, Diffusion 등의 딥러닝 방법을 활용하여 관계형, 시계열, 비정형, 이미지 재현데이터를 생성하는 서비스를 제공함.

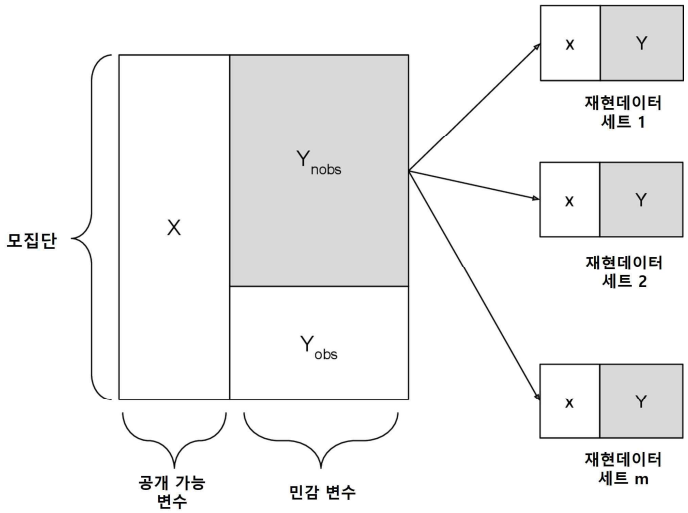
자료: 김승현(2020) 및 저자 조사

이제 대표적인 재현데이터 생성이론 몇 가지를 살펴보도록 하겠다. 재현데이터 생성 방법은 다양하게 발전되고 있고 아직 모든 경우에 잘 작동하는 표준화된 방법은 없다. 여기서는 통계적인 방식으로 재현데이터를 생성하는 이론을 주로 소개하고 추가로 최근 각광받고 있는 딥러닝 방법인 GAN(Generative Adversarial Network) 기반 생성 방법론도 간략하게 언급한다.

1) 모수적 다중대체 모형을 이용한 생성 방법론[20]

통계학에서 대체(Imputation)란 결측된 값을 적절한 값으로 바꾸는 것을 뜻하며 경우에 따라 하나의 값으로 바꾸는 단순대체를 사용할 수도 있고, 대체되는 값의 불확실성을 감안해 여러 개의 값으로 제공하는 다중대체를 사용하기도 한다. 대체에 대한 통계적 내용은 방대해 여기서 모든 내용을 다루기 어렵기 때문에 여기서는 다중대체를 중심으로 설명하겠다. 모수적 다중대체는 사후예측분포(Posterior predictive distribution)를 추정하여 결측값을 생성하며, 이 과정을 반복하여 여러 개의 대체값을 만들어내는 방식을 취한다. 대체를 이용한 재현데이터의 생성 방법을 소개하기 위해 아래 〈그림 Ⅲ-1〉을 이용해 데이터 내부 자료 구조를 나누어 설명해 보겠다.

〈그림 Ⅲ-1〉 데이터 분할



자료: Dandekar, A., Zen, R. A., and Bressan, S.(2009)

그림의 왼쪽 X는 민감하지 않은 공개 가능한 변수이고, Y는 노출을 최소화해야 하는 민감변수들을 나타낸다. 추가로 Y는 Y_{obs}과 Y_{nobs}로 나뉘는데, 전자는 관측된 값들이고 후자인 Y_{nobs}는 수집되지 못한 결측치로 정의된다. 따라서 수집된 혹은 관측된 전체 데이터

20) 박민정(2020)

는 $D = \mathrm{X}, \mathrm{Y}_{\mathrm{obs}}$이다. 재현데이터는 사후예측분포인 $P(Y_{nobs}|X, Y_{obs}) = P(Y_{nobs}|D)$를 추정하고 이로부터 값을 추출하여 Y_{nobs}를 채우는 방식으로 생성된다. 이를 반복하면 다른 값들의 Y_{nobs}로 채울 수 있어 다중대체의 효과를 지닌다. 채워진 Y_{nobs}값들은 재현데이터인 Y_{syn}으로 이해할 수 있으므로 사후예측분포 $P(Y_{syn}|D)$는 재현데이터 생성기를 의미한다고도 할 수 있다.

위의 다중대체 방법은 맥락에 따라 몇 가지 다른 형태로 변형될 수 있다.

(1) 만약 위 그림에서 Y_{obs}의 공개가 불가하다면 생성기인 $P(Y_{syn}|D)$를 이용하여 Y_{obs}, Y_{nobs} 둘 다 대체된 재현데이터 (X, Y_{syn})를 생성할 수 있다. 이는 앞서 소개한 부분재현데이터에 해당한다.

(2) 만약 위의 (1)에 추가로 X의 공개 역시 불가하다면 확장된 재현데이터 생성기인 $P(X_{syn}, Y_{syn}|D)$를 추정하고 이로부터 재현데이터 (X_{syn}, Y_{syn})를 생성할 수 있다. 이는 완전 재현데이터에 해당한다.

재현데이터 생성을 위한 대체에서 사후예측분포를 추정하는, 즉 생성기를 만드는 방법은 다양한데 기본적으로 다변량 데이터에 대해 분포추정을 할 수 있는 모든 모형이 사용 가능하다. 여기서는 모수적 방법에 국한하여 결합분포를 이용하는 방법과 각 주변분포에 대해 순차적으로 회귀모형을 적용하는 방법을 소개한다.

가) 결합모형을 이용한 분포 추정[21)]

생성하고자 하는 변수들이 벡터이고 이를 잘 설명할 수 있는 다변량 분포가 있다면 사후예측분포를 통해 재현변수들을 한 번에 벡터로 생성해 낼 수 있다. 만약 민감변수 벡터를 $Y_{nobs} = (Y_1, \cdots, Y_n)$이라고 하면 이 변수들의 분포를 동시에 고려한 결합분포(Joint distribution)를 추정해 이로부터 대체 값들을 벡터로 추출하는 것이다. 이는 수식

$$P(Y_{nobs}|D) = P(Y_1, Y_2, .., Y_n|X, Y_{obs}) = \int P(Y_{nobs}|X, Y_{obs}, \theta)P(\theta|X, Y_{obs})d\theta$$

으로 표현할 수 있다. 여기서 $P(\theta|X, Y_{obs}) = P(\theta|D)$는 데이터가 주어졌을 때 θ의 사후

21) 김정연·박민정(2019); 박민정·김항준(2016)

분포이다. θ는 변수들의 추정된 모수, 즉 모집단의 특성을 나타내는데, 예를 들어 $Y_{nobs} = (Y_1, \cdots, Y_n)$가 다변량 정규 분포를 따른다고 하면 θ는 평균 벡터와 공분산 행렬이 된다. 이 사후분포는 베이지안 통계의 방법론을 따라 추정하는 것이 보통이다. 다음으로 $P(Y_{nobs}|\theta, X, Y_{obs})$는 데이터와 모수 θ가 주어졌을 때의 조건부 다변량 분포이다.

실제 위의 수식을 이용할 때는, 먼저 추정된 $P(\theta|X, Y_{obs})$로부터 하나의 θ'을 추출한 후 이를 조건부 분포인 $P(Y_{nobs}|\theta', X, Y_{obs})$에 넣고 $Y_{nobs} = Y_{syn}$을 생성하는 과정을 반복한다.

그러나 결합분포를 사용하는 이 같은 방식은 생성할 변수의 종류가 다양하고 차원 n이 커짐에 따라 결합확률분포를 추정하기가 어려워지고 차원의 저주(curse of dimensionality)[22]로 인해 정확성이 떨어지기 때문에 아래 설명할 조건부 모형을 이용한 순차회귀 방법이 보다 널리 사용된다.

나) 순차회귀를 이용한 분포추정

이 방법은 개별 주변변수에 대해 순차적으로 회귀모형 혹은 다른 지도학습모형을 적용하는 방법으로 변수들이 서로 다른 형태(이산형, 범주형, 연속형)를 가지거나 한꺼번에 다변량 분포로 추정하는 것이 어려울 때 사용할 수 있으며, 특히 변수들이 순차적인 관련성이 있을 때 잘 작동한다. 실제 적용할 때에도 조건부 확률분포를 변수 중요도 순서에 따라 추정하여 사용할 수 있고, 개별 변수의 사후예측분포를 순차적 혹은 연쇄적으로 추정하기 때문에 직관적이고 적용이 쉽다. 통계학에서 순차회귀 다중대체는 Sequential Regression Multiple Imputation(SRMI) 혹은 Multiple Imputation by Chained Equations(MICE)라고 불린다.

순차회귀를 예를 들어 살펴보기 위해 어떤 데이터 D가 주어졌을 때 세 변수의 결합확률분포 $P(W, Y, Z|D)$를 추정하는 것이 목표라고 하자. 만약 세 변수들 사이에 적절한 인과관계가 존재하거나 중요도 또는 상관관계가 있어 $Z \rightarrow Y \rightarrow W$의 순서대로 변수들을 정렬할 수 있다면, 조건부 확률의 성질을 이용해

22) 차원의 저주란 차원이 증가하면서 모형의 성능을 동일하게 유지하기 위해 필요한 학습데이터 수가 기하급수적으로 증가하는 현상을 말함

$$P(W,Y,Z|D) = P(W|Y,Z,D)\,P(Y,Z|D) = P(W|Y,Z,D)P(Y|Z,D)P(Z|D)$$

로 쓸 수 있고, 추가로 Y가 주어졌을 때 W와 Z가 독립이라면

$$P(W,Y,Z|D) = P(W|Y,D)P(Y|Z,D)P(Z|D)$$

와 같이 더욱 단순화할 수 있다.

여기서 모수적 순차회귀 다중대체를 위해 $P(Z|D)$, $P(Y|Z,D)$, $P(W|Y,Z,D)$의 분포를 순서대로 추정했는데, 이 분포들은 모두 일변량의 조건부 분포이므로 회귀모형을 포함한 적절한 지도학습모형을 이용하여 추정할 수 있다. 다시 말해 다변량 분포의 추정 문제를 쪼개어 다수의 일변량 분포 추정으로 전환한 셈이다. 추정된 조건부 분포들이 주어지면 개별 변수들을 동일한 $Z \rightarrow Y \rightarrow W$ 순서대로 하나씩 랜덤하게 추출할 수 있고, 이를 반복하면 원하는 수 만큼의 재현데이터를 생성할 수 있다.

원칙적으로 W, Y, Z의 순서는 인과관계에 기반하여 정해야 하지만 현실에서는 인과관계가 명확하지 않은 경우가 많기 때문에 상관성이 높은 순서로 정하기도 한다. 따라서 이 기법에서 순서의 결정에는 자의적인 면이 있다.

만약 변수 간의 순서가 존재하지 않는다면, 개별 변수에 대해 해당 변수를 제외한 나머지 변수들과 D를 설명변수, 즉, 조건부로 두고 분포추정을 할 수 있으며 이때 변수들의 순서는 무시하면 된다.

2) 비모수적 다중대체 모형을 이용한 생성 방법론[23]

실제 재현데이터를 생성할 때 모수적으로 분포를 추정하는 것이 적절하지 않거나 어려운 경우가 많다. 혼합분포(Mixture distribution)를 활용하여 모수적 접근 방식의 한계를 어느 정도 극복할 수도 있으나 이러한 시도는 제한적이다.[24] 이런 경우 비모수적인 방법을 이용할 수 있다. 비모수적 방법은 모집단에 대한 특정 분포 가정을 하지 않는 통계적 방법으로 여기서는 재현데이터 생성에 사용되는 몇 개의 기법들을 소개한다.

먼저 소개할 기법은 CART(Classification And Regression Tree)이다. CART는 주어진 여

23) 박민정(2020)

24) Chib, S.(1996)

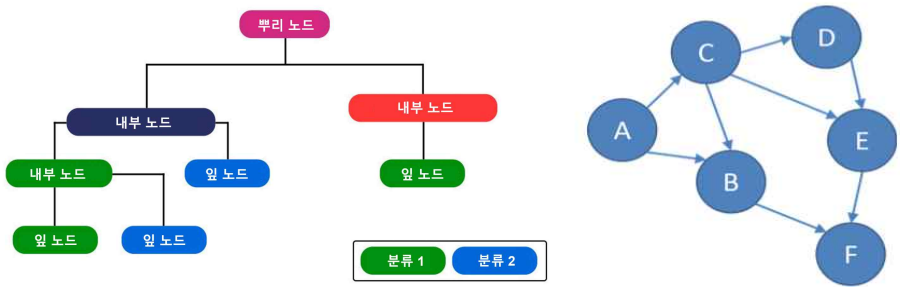
러 설명 변수에 기반하여 분할 규칙에 따라 의사결정나무를 만들고 반응 변수의 값을 예측하는 비모수적 방법이다. 각 분할은 뿌리노드(Root node)에서 시작해 각 변수별로 이진구조(Binary)로 쪼개는 방식을 순차적으로 적용해 진행되며 마지막에 말단노드(Terminal or leaf node)를 만든 후 종료된다. 분할의 기준은 보통 불순도(Impurity)[25]이다. CART는 연속형과 범주형 변수가 섞여 있는 데이터에 적합할 뿐만 아니라 분석 결과가 나무(Tree) 구조로 표현되어 이해가 쉽기 때문에 재현자료 생성에 자주 사용된다. 통계 소프트웨어 R의 패키지 synthpop 또한 method 옵션을 따로 지정하지 않으면 디폴트 옵션으로 CART 기법을 선택해 재현데이터를 생성한다. 보통 CART 알고리즘은 예측을 위해 개별 말단노드에서 관측된 반응 변수의 평균을 사용하지만 데이터 재현을 위해서는 비모수적으로 추정된 분포를 사용해 데이터를 생성하거나 무작위 복원추출법인 붓스트랩(Bootstrap)을 사용하는 것이 더 적절하다. synthpop에서는 베이지안 붓스트랩을 사용한다고 알려져 있다.

이외에도 CART를 확장한 Bagging이나 Random Forest, 그리고 서포트벡터머신(Support Vector Machine; SVM), 베이지안 네트워크(Bayesian Network) 등의 다른 비모수 방법론들도 재현데이터 생성에 사용될 수 있다.

베이지안 네트워크의 경우 개별 변수들을 노드로 두고 이들 간의 종속(인과)관계를 방향성이 있는 화살표로 연결한 네트워크 형태를 가지는 그래프 기반 모형이다. 원데이터로 학습된 베이지안 네트워크를 이용하면 다양한 쿼리에 대한 답을 얻을 수 있고, 각 노드에서 표본을 생성하는 과정을 통해 데이터 생성기로도 사용할 수 있다. 베이지안 네트워크는 Directed Acyclic Graph(DAG)라는 조건을 만족해야 하며, 개별 확률변수를 순차적으로 분할하는 CART와 달리 데이터 기저에 존재하는 다변량 분포를 종속관계에 따라 순차적으로 쪼갠 후 개별 조건부 확률분포를 범주형의 경우 전이행렬로, 연속형인 경우 회귀모형을 이용해 만드는 것이 보통이다. 〈그림 Ⅲ-2〉는 CART와 베이지안 네트워크의 예시를 보여준다.

25) 불순도는 다양한 범주(Factor)들의 개체들이 얼마나 포함되어 있는가를 의미함. 즉, 여러 가지의 클래스가 섞여 있는 정도를 말함. 회귀나무에서는 주로 지니계수(Gini Index)를 사용함

〈그림 Ⅲ-2〉 CART(좌) 및 베이지안 네트워크(우)의 예시



자료: https://www.geeksforgeeks.org/cart-classification-and-regression-tree-in-machine-learning/; Wei, J., Nie, Y., and Xie, W.(2020)

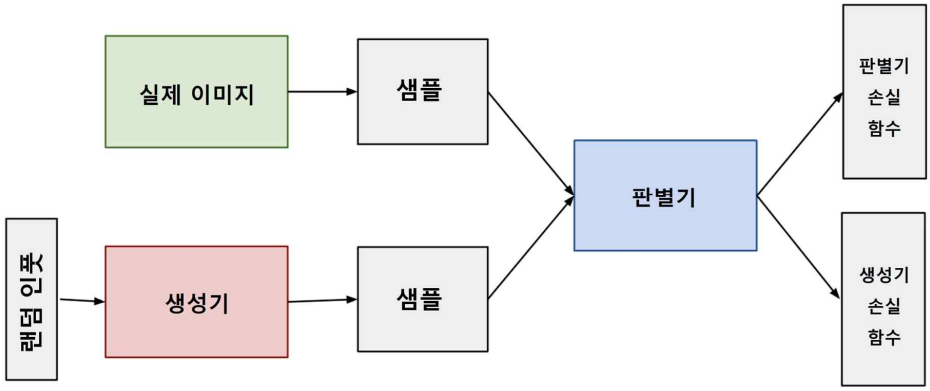
3) GAN 기반 생성 방법론

딥러닝 기술을 활용한 재현(합성)데이터는 컴퓨터 비전 등의 영역에서 AI학습용 데이터로 적극 활용되고 있다. 특히 이미지나 동영상 데이터에 대한 재현 기술이 많이 연구되고 왔으며 최근 정형데이터로의 확장도 이루어지고 있다. 본 보고서에서는 대표적인 딥러닝 기술인 GAN(Generative Adversarial Network)이 어떻게 재현데이터를 생성하는지 간략히 설명하려 한다.

회귀 및 분류 모형과 같은 많은 머신러닝 혹은 딥러닝 방법론들은 주어진 데이터를 바탕으로 특정 값을 도출하는 지도학습(Supervised learning)에 속한다. 지도학습모형을 훈련시키기 위해서는 모형을 통해 얻고 싶은 결과값이 포함된 데이터가 필요한데, Ian Goodfellow는 지도학습에 필요한 결과값 데이터를 모형이 자체적으로 만드는 생성모형(Generative model)인 GAN을 제안하였다.[26] 그가 모형을 설명할 때 사용한 비유를 인용하자면, GAN은 경찰과 위조지폐범 사이의 게임과 같다. 위조지폐범은 진짜 같은 화폐를 생성해 경찰을 속이려 하고, 경찰은 진짜 지폐와 가짜 지폐를 판별하려 한다. 이러한 경쟁적 학습이 지속되면 위조지폐범은 진짜와 매우 유사한 위조지폐를 만들 수 있게 되고, 경찰 또한 위조지폐를 판별하는 상당한 실력을 가지게 된다. 즉, 최종적으로 위조지폐범은 고품질의 위조지폐를 '생성'하게 된다. 이 비유에서 위조지폐범은 생성기(Generator)를, 경찰은 판별기(Discriminator)를 의미하며 이 두 모델이 서로 적대적 학습을 이어나가는 것이 GAN의 원리이다.

26) Goodfellow, I., Bengio, Y., and Courville, A.(2016)

〈그림 Ⅲ-3〉 GAN의 Generator와 Discriminator 관계



자료: Google developers(https://developers.google.com/machine-learning/gan/gan_structure?hl=ko)

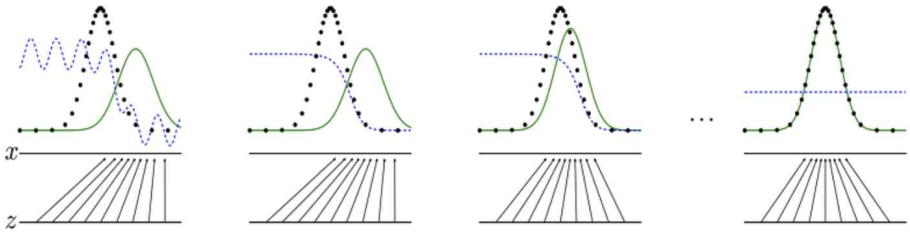
GAN은 〈그림 Ⅲ-3〉과 같이 데이터를 생성하는 생성기 G와 생성된 데이터가 진짜인지 판별하는 판별기 D를 학습시키는 과정을 반복한다. 이를 수식적으로 표현한 것이 아래의 손실 함수(Loss function)이다.

$$\min_{G}\max_{D} V(G,D) = E_{x \sim P_{data}(x)}[\log D(x)] + E_{z \sim P_z(z)}[\log (1 - D(G(z)))]$$

생성기는 데이터의 임의성(Randomization)을 주기 위해 정규분포 등에서 임의 생성한 값 z을 인자로 받아, 원데이터 x와 유사한 데이터 $G(z)$를 만들어 준다. 원데이터와 유사하게 만들도록 생성기를 학습하는 과정이 손실함수를 최소화($\min_{G}$)하는 과정이다. 원데이터와 만들어진 데이터 $(x, G(z))$를 각각 판별기 $D(\bullet)$에 넣어 판별하게 되는데, 판별력을 높이는 과정이 손실함수를 최대화($\max_{D}$)하는 과정에서 진행된다. 알고리즘의 목표는 진짜와 유사한 데이터를 만들어 내는 생성기이기에 최소화($\min_{G}$)를 최종적으로 진행하게 된다.

이렇게 손실함수를 이용하여 역전파(Backpropagation)의 방법을 통해 생성기와 학습기의 능력을 향상시키는데 Ian Goodfellow는 이 과정을 〈그림 Ⅲ-4〉를 통해 표현했다.

〈그림 Ⅲ-4〉 GAN 분포 학습 과정



자료: Goodfellow, Ian, et al.(2020)

〈그림 Ⅲ-4〉에서 굵은 점선은 원데이터의 확률 분포를 나타내고, 학습이 진행될수록 점차 이 선에 가까워지는 실선은 GAN의 생성기가 만든 확률 분포이다. 후반부로 갈수록 평행에 가까워지는 가는 점선은 판별기의 확률 분포를 나타낸다. 가장 왼쪽 그림이 학습 시작 시점에서의 상태이고 학습이 진행되면서 차례로 오른편의 상태로 변한다. 학습이 완료된 가장 오른쪽 상태가 되면 분류기가 생성된 것인지 원본인지 구분할 확률이 0.5에 수렴해, 생성모델이 실제 데이터와 거의 유사한 데이터를 만들어 냈음을 뜻한다.

이렇게 적대적 학습을 통해 얻게 된 생성모델 자체를 이용하여 재현데이터를 생성할 수 있다. 현재까지 GAN은 끊임없이 발전하여 이미지 데이터 분야에서는 DCGAN(Deep Convolutional GAN), medGAN(Choi et al. 2017), ehrGAN(Che et al. 2017) 등이 사용되고 있다. 뿐만 아니라 원하는 레이블에 대한 데이터를 생성하기 위해 학습데이터에 레이블을 포함하여 학습시키는 CGAN(Conditional-GAN)을 활용하여 정형데이터에서도 사용 가능한 GAN 기반 방법론에 대한 연구가 진행되고 있다.[27)]

27) Xu, Lei, et al.(2019)

2. 보험데이터를 이용한 재현데이터의 예시

가. 데이터 소개

여기서는 실제 데이터를 이용해 재현데이터를 생성하고 생성된 데이터와 원본 데이터를 비교한다. 대표적인 비모수적 재현데이터 생성 알고리즘인 통계 소프트웨어 R의 synthpop 패키지와 그래프를 그리기 위한 tidyverse 패키지를 사용한다.

사용할 데이터는 뮌헨 공과대학교(TUM)가 실제 회사로부터 데이터를 얻어 공개 가능한 수준으로 조작한 보험데이터[28]로 15,000행과 11개의 열로 이루어져 있다. 이 데이터셋은 자동차 보험가입 여부를 포함해 성별, 직업, 혼인 상태, 신용불량 여부, 통화 시간 등 자동차 보험가입에 대한 정보를 수집하기 위해 연락을 받은(Cold call) 고객들의 다양한 정보를 포함하고 있다.[29] 원래의 데이터셋은 자동차 보험가입 여부를 예측하기 위해 훈련용 데이터(Training data)와 검증용 데이터(Test data)로 나누어져 있지만 본 보고서는 예측이나 분류의 문제를 다루는 것이 아닌 재현데이터 생성에 초점을 두고 있으므로 자동차 보험가입 여부가 라벨링되어 있는 훈련용 데이터만을 사용했다.

재현데이터 생성 전 데이터 전처리 과정은 다음과 같다. 먼저 이전 마케팅 콜에 대한 결과를 담고있는 'Outcome'변수의 결측값이 약 76%로 관측되어 변수를 삭제했다. 마케팅 콜 시작 시각과 종료 시각 정보가 각각 담겨있는 'Call Start'와 'Call End' 변수에서 가장 중요한 정보는 통화 시간이므로 후자 변수에서 전자 변수를 뺀 'Call_time' 변수를 연속형으로 생성하고 'Call Start'와 'Call End' 두 변수는 삭제했다. 통화 기기 종류를 나타내는 Communication 변수(Cellular와 Telephone 범주 존재)에서 N/A로 처리된 관측치들(22.5%)에 대해 'Others'라는 새로운 범주를 부여했다. 그 외 정보 동의를 하지 않거나 부정확한 정보를 포함한 관측치 169건(4%)에 대해 행을 삭제해 최종적으로 3,820건 관측치를 재현했다.

28) 데이터 출처(https://www.kaggle.com/datasets/kondla/carinsurance)

29) 전체 변수 정보는 〈표 Ⅲ-2〉를 참고하길 바람

〈표 Ⅲ-2〉 자동차 보험 Training Dataset Description

변수명	설명	예시
Id	고유 ID 번호	'1' … '5000'
Age	고객 나이	'18', '20', … , '90'
Job	고객 직업	'admin.', 'blue-collar', 등
Marital	고객 혼인 상태	'divorced', 'married', 'single'
Education	고객 학력 수준	'primary', 'secondary', 등
Default	신용불량자 여부	'yes' - 1, 'no' - 0
Balance	연간 평균 잔고(달러 기준)	-2119, 589, …
HHInsurance	가계 보험 보유 여부	'yes' - 1, 'no' - 0
CarLoan	자동차 대부금 보유 여부	'yes' - 1, 'no' - 0
Communication	마케팅콜 수단	'cellular', 'telephone', 'N/A'
LastContactMonth	직전 마케팅콜 월	'jan', 'feb', 등
LastContactDay	직전 마케팅콜 일	'1' … '31'
CallStart	통화 시작 시각	12:43:15
CallEnd	통화 종료 시각	12:43:15
NoOfContacts	총 연락 횟수	'0'. '1'. '2'. …
PrevAttempts	이전 마케팅콜 결과	'failure', 'other', 'success', 'N/A'
CarInsurance	자동차 보험가입 결과	'yes' - 1, 'no' - 0

자료: https://www.kaggle.com/datasets/kondla/carinsurance

나. 재현데이터 생성

본 보고서는 재현데이터 생성을 간단하게 제시하고 유효한 데이터셋이 생성되었는지 확인하는 것이 목표이므로 재현될 변수들의 순서와 method를 디폴트 옵션으로 설정했다. 디폴트 옵션에서는 원데이터의 왼쪽에 있는 열부터 순차적으로 재현데이터를 생성하지만 필요하다면 패키지 내부 함수 syn의 옵션인 visit.sequence를 통해 변수들의 순서를 사용자가 지정할 수 있다. 생성 알고리즘의 선택은 옵션 'method ='으로 지정 가능한데 디폴트 옵션은 CART로서 범주형과 수치형 변수가 섞인 해당 데이터에 바로 적용 가능하다. CART 알고리즘에 변수들이 순차적으로 입력되면서 특정 조건을 만족하는지 여부에 따라 분류작업을 지속적으로 진행하며 지니 불순도(Gini Imputity)가 낮아지는 방향으로 구간을 정해 합리적인 관측치 집합을 만들어 낸다. 여기서는 최초분류변수로 Age를 선택하고 나머지는 CART 알고리즘을 따라 생성하였는데, 최초변수는 복원 랜덤 추출로 생성된다.

재현되는 데이터셋의 갯수는 syn 함수의 m으로 설정한다. 통계적 검증과 노출위험의 추정을 위해 다수의 재현데이터셋을 만들고자 한다면 m값을 조절하면 된다.[30] 본 예시에서는 m값을 1로 두고 한 세트의 재현데이터만를 생성해 원데이터와 직관적인 비교가 가능하게 하였다.

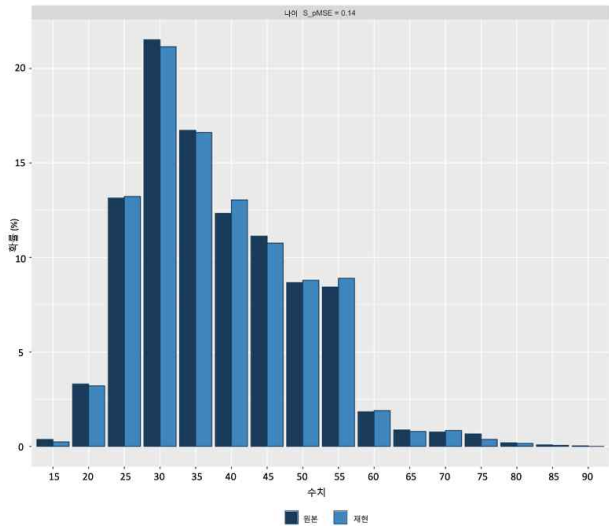
마지막으로 syn 함수를 이용해 생성한 재현데이터를 원데이터와 비교해 동일한 관측치가 있는지 확인한 결과, 총 9건이 원데이터와 중복되었는데 이는 전체 데이터의 0.23%이다. 따라서 이 9건의 관측치를 삭제하고 원데이터와 중복되지 않은 재현데이터 9건을 추가로 생성하여 원데이터와 동일한 크기인 3,820건을 최종 재현데이터로 채택했다.

다. 원데이터와 특성 비교

재현데이터에서 원데이터의 통계적 구조가 유지되는지 확인하기 위해 재현 전후 데이터의 단변량 및 다변량 비교 결과를 살펴보면 다음과 같다.

1) 단변량 비교

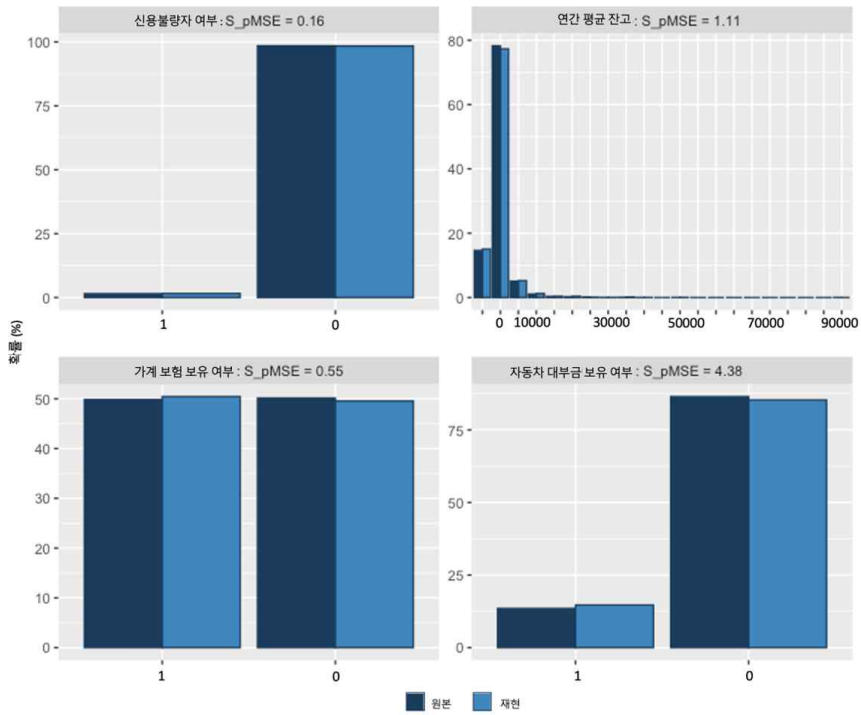
〈그림 Ⅲ-5〉 나이 변수 비교



30) Nowok, B., Raab, G. M., and Dibben, C.(2016)

〈그림 Ⅲ-5〉와 같이 랜덤 샘플링 기법으로 재현된 데이터는 원본과 그 데이터 분포가 유사하다. 실제로 이 두 분포의 차이를 확인해 보려 Kolmogorov-Smirnov Test[31]를 시행한 결과, 양측 검정 p-value가 약 0.99로 도출되어 재현이 잘 되었다고 볼 수 있다. 특히 원데이터에서는 최댓값이 95세로 노출이 일어날 확률이 높았으나 재현데이터에서는 최댓값이 86세로 나타나 원데이터나 외부의 데이터와 매칭을 어렵게 만들었다. 또 두 데이터에서 나이의 최솟값은 18세로 동일하다.

〈그림 Ⅲ-6〉 연속형 변수의 비교

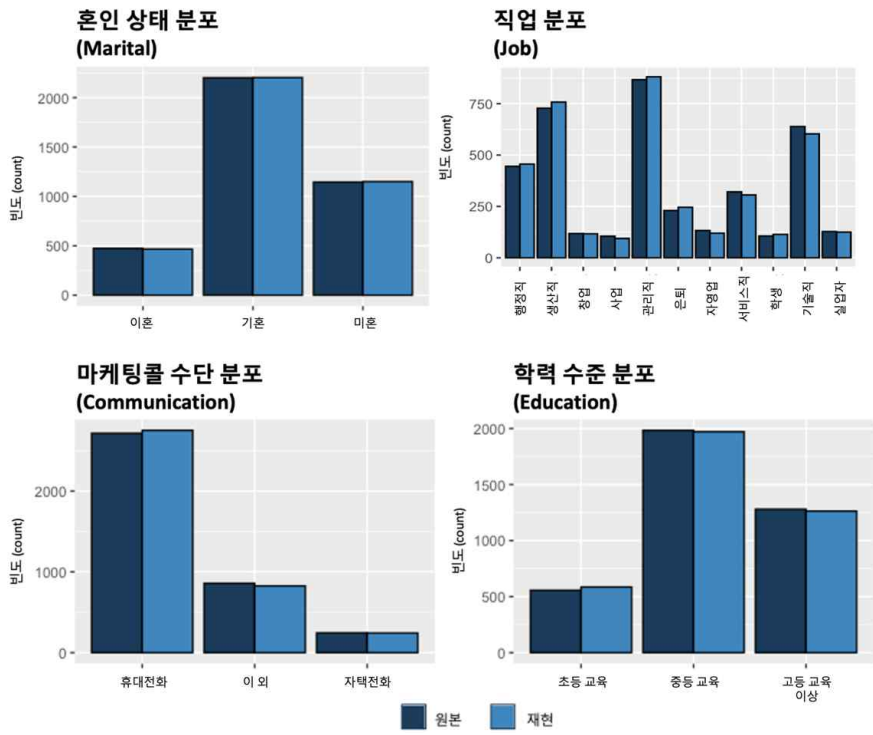


〈그림 Ⅲ-6〉은 신용불량자 여부(Default), 계좌 잔고(Balance), 가계 보험 보유 여부(HHInsurance), 자동차 대부금 보유 여부(CarLoan) 변수를 비교한 각각의 그래프로서 재현데이터와 원데이터의 분포가 유사함을 확인할 수 있다. 이를 수치적으로 보았을 때도

31) 데이터의 누적분포함수와 비교하고자 하는 분포의 누적분포함수 간의 최대 거리를 통계량으로 사용하는 가설검정 방법임. 귀무가설을 두 데이터의 분포가 동일한 것으로 설정하여 p-value가 α(주로 0.05, 0.01로 설정)을 으면 분포가 동일한 것으로 봄

마찬가지이다. 가계보험 보유(가입) 여부(HHInsurance) 변수의 경우, 보험가입(1)이 원데이터에서는 1,904건(49.8%), 재현데이터에서는 1,927건(50.4%)으로 나타났고 보험미가입(0)의 경우는 원데이터에서 1,916건(50.2%), 재현데이터에서 1,893건(49.6%)이 나타났다. 연속형 변수인 계좌 잔고(Balance)는 나이(Age) 변수와 같이 Kolmogorov-Smirnov Test를 통해 검정할 수 있는데 역시 p-value가 약 0.99로 두 데이터셋이 동일한 분포를 따른다는 결과를 보였다.

〈그림 Ⅲ-7〉 범주형 변수의 비교

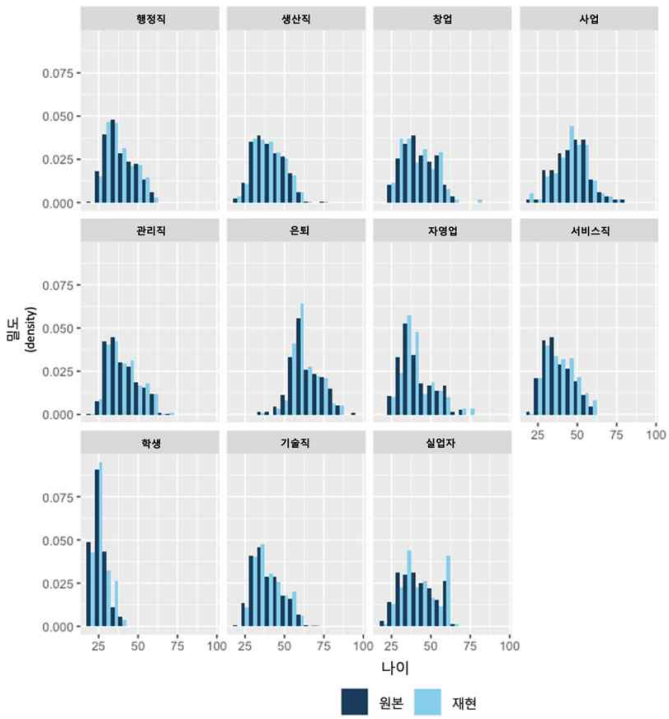


수치형으로 표현된 변수에 이어 〈그림 Ⅲ-7〉을 통해 범주형 변수 분포를 살펴보자. 혼인상태(Marital), 직업군(Job), 마케팅콜 수단(Communication), 학력(Education)의 원데이터와 재현데이터를 비교한 막대그래프이다. 범주의 수가 10개 이상인 직업군(Job)에서도 원본 분포와 재현데이터 분포가 유사한 것을 확인할 수 있으며 특히 재현 순서가 비교적 앞 순서였던 혼인 상태(Marital)는 각 범주의 관측치 수가 원데이터와 99% 이상 동일했다.

2) 다변량 비교

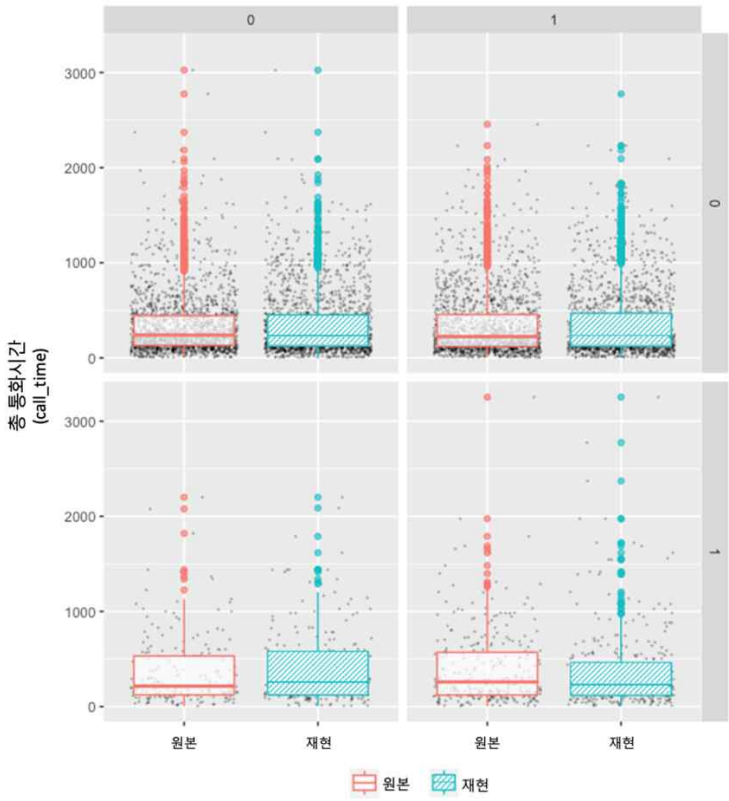
다변량 데이터에서 변수들의 상관관계는 매우 중요하다. 재현 전후 데이터가 유사한 상관관계를 유지하는지를 확인할 필요가 있지만, 모든 변수 조합을 고려하는 것은 지면상 어려우므로 여기서 예시로서 몇 가지 경우만 살펴보겠다.

〈그림 Ⅲ-8〉 직업군(Job)과 나이(Age) 이변량 분포 히스토그램



먼저 〈그림 Ⅲ-8〉은 이변량 분포를 히스토그램으로 나타낸 것이다. 연속형 변수인 나이(Age)를 가로 축으로 두고 범주 수가 많았던 직업군(Job)별로 각각의 범주에서 재현이 잘 되었는지를 보여준다. 예를 들어 학생 범주의 경우, 원데이터에서 20세 미만의 관측치는 12건이었고 재현데이터에서 20세 미만 관측치는 10건이었다. 그중 원데이터에서 18세로 관측된 행은 2건이었고 재현데이터에서 18세로 관측된 행 또한 2건으로 동일했다. 그 외 직업군 범주의 경우, 원데이터에서 20세 미만의 관측치가 발견되지 않았으며 재현데이터에서도 20세 미만 관측치가 없는 것으로 나타났다.

〈그림 Ⅲ-9〉 세 변수의 비교



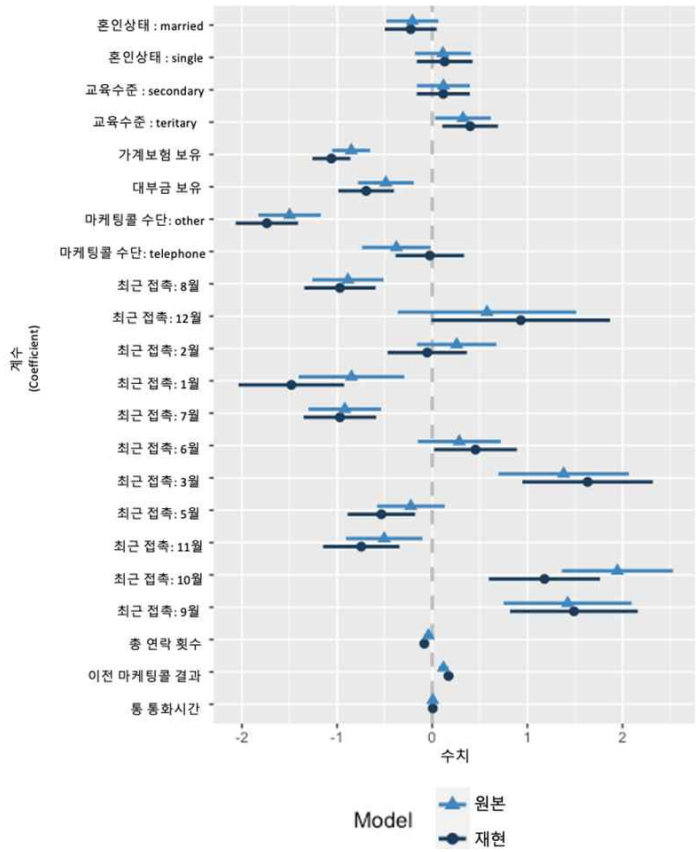
〈그림 Ⅲ-9〉는 마케팅콜 총 시간(call_time), 자동차 대부금 보유 여부(CarLoan), 가계 보험 보유 여부(HHInsurance)를 동시에 보여주는 박스 플롯이다. 원데이터 관측치 분포를 나타내는 눈금 없는 박스와 재현데이터 관측치 분포를 나타내는 눈금이 있는 박스가 각 범주마다 비슷한 범위 안에 놓여있다. 각 범주에 해당하는 데이터를 분할하여 총 4번의 Kolmogorov-Smirnov test를 시행한 결과, p-value가 모두 0.9에 가까운 수치를 기록하여 같은 분포를 따르고 있다는 결론을 도출했다.

3) 로지스틱회귀모형을 이용한 비교

엄밀한 의미에서 재현 전후 데이터를 비교하기 위해서는 가능한 모든 분석들을 시도하고 그 결과들을 비교해야 하겠지만 이는 현실적으로 불가능하다. 여기서는 대표적인 분석 모형인 회귀분석의 결과를 예시로 설명하겠다. 구체적으로, 변수를 다차원적으로 평가할 수 있는 로

지스틱회귀모형[32]을 적합해 두 데이터를 비교한다. 다중 회귀 모델에서 설명변수가 과도하게 많아지면 오히려 성능이 저하되기 때문에 본격적인 로지스틱회귀모형을 만들기 앞서 후진제거법(Backward elimination)을 이용해 설명력이 적은 변수를 순차적으로 제거하였다. 최종적으로 모델에 사용된 설명변수는 9가지로 혼인 상태(Marital), 교육 수준(Education), 자동차 대부금 보유 여부(CarLoan), 가계 보험 보유 여부(HHInsurance), 마케팅콜 연락 수단(Communication), 마케팅콜 통화 시간(Call_time), 총 연락 횟수(NoOfContacts), 이전 마케팅콜 결과(PrevAttempts), 직전 마케팅콜 월(LastContactMonth)이다. 종속변수로는 자동차 보험가입 여부(CarInsurance)를 설정하여 로지스틱회귀모형을 적합했다.

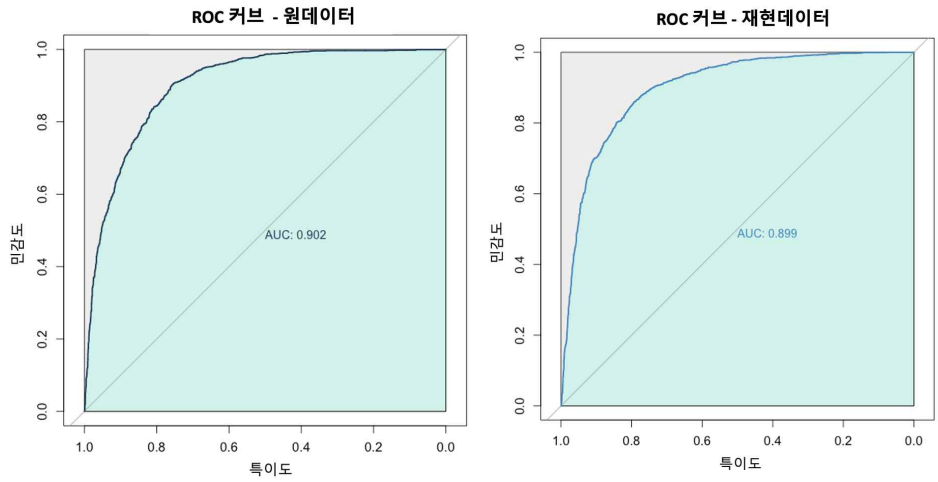
〈그림 Ⅲ-10〉 재현 전후 로지스틱회귀모형의 계수와 신뢰구간 비교



32) 종속변수(y)가 0 또는 1로 binary 변수일 때 사용하는 회귀분석 모델로 결과가 특정 분류로 나뉘기 때문에 일종의 분류 예측 모델임

〈그림 Ⅲ-11〉은 회귀분석 결과로 도출된 원데이터와 재현데이터 회귀계수와 신뢰구간을 비교한 결과이다. 원데이터와 재현데이터 범주형 변수 회귀계수의 신뢰구간이 Overlap되는 구간 평균은 약 0.71로 계산되었는데, 특히 범주형 변수의 회귀계수들의 Overlap 비율은 90% 이상의 정확성을 보였다.

〈그림 Ⅲ-11〉 재현 전후 데이터에 적합한 로지스틱회귀모형의 ROC 곡선 비교



적합된 두 개의 로지스틱회귀모형에서 도출된 ROC 곡선[33]은 〈그림 Ⅲ-11〉에서 보듯이 매우 유사하며, AUC[34]값 역시 원데이터는 0.902, 재현데이터는 0.899로 비슷하다. ROC 곡선을 도출할 때 보통 데이터를 훈련 집합과 검증 집합으로 나누는 것이 일반적이지만 여기서는 모형의 예측성능이 아니라 재현 전후 데이터의 유사성에 관심이 있으므로 검증 집합을 따로 두지 않고 전체 데이터를 훈련 집합으로 두고 진행하였다.

본 절에서는 재현 전후의 데이터를 다각도로 비교하였지만 여전히 제한적인 비교에 그친다는 점에서 재현데이터의 원데이터와의 유사성에 대해 일반적인 결론을 내리는 것은 어렵다. 사용자가 시도하는 쿼리나 분석의 종류와 범위는 매우 다양할 수 있어 이를 모두 조

33) ROC 곡선은 FPR(False Positive Rate)과 TPR(True Positive Rate)을 각각 x, y축으로 놓은 그래프로 임곗값(Threshold)을 바꿔가며 측정했을 때 FPR과 TPR의 변화를 나타낸 곡선임. 주로 이진 분류 모형의 성능평가에 사용되며 그래프가 좌상단에 근접할수록 좋은 성능을 보인다고 해석함

34) Area Under the ROC Curve의 약자로 ROC 곡선 아래의 면적을 뜻함. AUC 값은 ROC 곡선을 하나의 숫자로 요약한 값으로, 1(100%)에 가까울수록 모형의 성능이 좋다고 할 수 있음

사하는 것은 불가능하기 때문이다. 이런 이유로 특정 분석이 아니라 재현 전후 데이터를 분석 수준이 아니라 분포 수준에서 비교하는 것이 더 합당하다. 이는 재현데이터의 품질에 관한 중요한 대목으로 다음 장에서 좀 더 자세히 논하기로 한다.

3. 국내 및 해외 재현데이터 이용 사례

이 절에서는 금융과 관련된 국내외 몇 개의 사례들을 살펴보겠다. 이 외에도 다른 사례들이 있고 금융 외 다른 분야에서도 재현데이터의 사용 사례가 다수 있으나, 지면의 한계로 다 싣지 못함을 양해하기 바란다.[35]

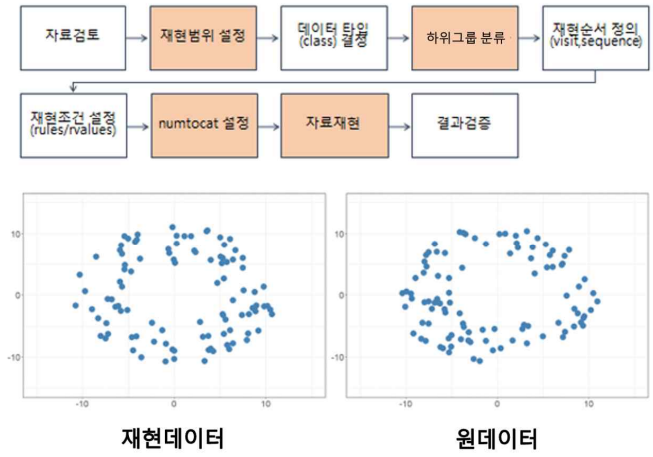
가. 국내 사례

1) 통계청 재현데이터 기초연구(2016~)

통계청은 마이크로데이터 공개에 따라 민감한 개인정보 노출위험을 줄이면서, 정보 손실을 최소화 하기 위한 방안으로 재현데이터 활용방안에 대한 탐색과 해외 사례 등에 대한 기초연구를 수행해 왔다. 구체적으로, 재현자료 생성을 위해 통계적 모형을 활용하는 방법론과 재현자료 노출위험 및 정보 손실 측정론에 대한 연구와 함께, 국내외 통계기관의 재현자료 생성 사례 및 통계데이터센터 DB에 대한 재현자료 시범 생성 결과 보고서를 발간한 바 있다.

35) 국내외 재현데이터 사용의 다른 사례들은 박민정(2020) 및 한국신용정보원(2020)를 참고하길 바람

〈그림 Ⅲ-12〉 통계청 K-통계시스템 구축 계획



자료: 통계청(2021b)

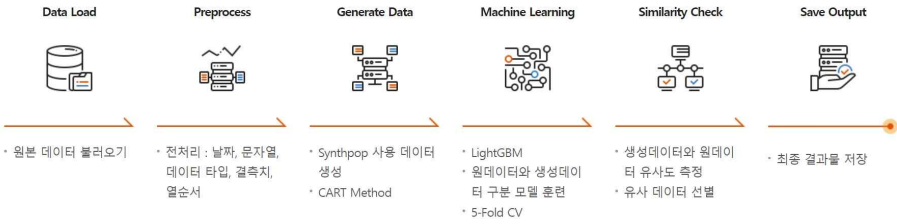
또한 〈그림 Ⅲ-12〉와 같이 2021년 공공데이터 활용을 증대하기 위한 통계청 K-통계 시스템 구축 계획을 발표했는데 새로운 데이터 보호 기술로 제안한 항목에 재현데이터가 포함되어 있는 것을 확인할 수 있다.

2) 코리아크레딧뷰로(KCB) 제주도 전입인구 특성 분석(2019)

코리아크레딧뷰로(이하, 'KCB'라 함)는 신용정보를 활용하여 개인신용평가를 시행하는 기업으로 신용평가 외에도 데이터의 제공과 활용을 도와주는 데이터 스토어를 운영하고 있다.

재현데이터와 관련해서, KCB는 2018년 이후 급증한 제주도 전출인구의 특성을 재현데이터로 만들어 분석·시각화하여 도내 인구 구조의 변화와 그 영향을 파악하기 위한 프로젝트를 진행한 바 있다. 재현데이터를 생성하는 과정에서 고차원으로 인한 데이터 부족 문제 해결을 위해 가능한 기준이 되는 변수 중 연속형은 축약된 명목형으로 바꿔 모델링하였고, 최종적으로 원데이터의 분포와 유사한 약 50만 건의 재현데이터를 생성해 제주도 전출인구 특성 분석을 진행하였다.

〈그림 Ⅲ-13〉 재현데이터 생성 순서



자료: KCB 데이터스토어(https://datastore.koreacb.com/support/utilizeCaseView8.do)

〈그림 Ⅲ-13〉은 이 프로젝트에서 재현데이터를 이용하여 성능평가를 했을 때의 진행 순서이다. Python에서 synthpop 패키지의 CART 방법론[36]을 적용하여 재현데이터를 생성했으며 머신러닝 모델 훈련 및 원데이터와 유사도 측정도 진행하였다. 그 결과 정확도와 AUC에서 각각 성능비 97.8%, 97.7%라는 우수한 평가수치를 보였다고 한다.

3) 신용정보원 부도예측을 위한 GAN 기반 재현데이터 생성 및 검증 보고서(2022)

신용정보원은 2022년 『부도 예측을 위한 인공지능 학습용 데이터 생성 및 검증 기법: GAN 기반 재현데이터를 중심으로』[37]라는 CIS 보고서를 발간했다. 인공지능 학습모형인 GAN을 적용해 원데이터의 통계적 특성을 유지한 재현데이터를 생성하고 그 결과를 평가한 내용을 담고 있으며, 생성된 재현데이터가 인공지능 학습데이터로 유용하게 활용될 수 있음을 시사했다.

내용을 간략히 요약하자면, 원데이터를 기반으로 한 재현데이터를 만들어 내기 위해 대출, 연체, 부도 정보를 포함한 실제 데이터를 준비하고 종속변수로는 부도 여부(Binary variable), 설명변수로는 신용공여 총 잔액, 원화 대출 총 기관 수, 연체율 등 신용정보를 사용했으며 Python 프로그래밍을 활용했다. 신용정보원 출처의 가명처리된 원데이터에서 적정 크기 표본 5만 행을 추출하고 이 샘플의 90%는 학습데이터로, 10%는 분류평가용 데이터로 분리해 사용했다. 학습데이터 중 부도 차주 레코드가 전체의 50%가 되도록 1차 재현데이터를 생성했고 이후 실제 데이터와 재현데이터 비율이 5:5가 되도록 구성한 새로

36) 추후 비모수적 생성 방법론에서 언급할 것임
37) 홍동숙(2022)

운 학습데이터를 생성했다고 한다. 그 이후 KS 통계량, AUC, 재현율을 평가지표로 사용하여 최종적으로 생성된 재현데이터를 평가한 결과, 재현데이터가 실제 데이터와 유사한 통계량을 보유하고 있음을 확인했고 데이터 불균형 문제에 따른 낮은 재현율이 개선되는 효과를 보여 실제 데이터를 대체할 수 있는 수준임을 보였다.

4) 국세청 재현데이터 도입 계획(2022)[38]

국세청 또한 2022년 국세 데이터 활용도를 높이기 위한 재현데이터 활용 계획을 밝혔다. 국세청의 데이터 센터는 방대한 양의 금융데이터를 보유하고 있으며 주요 정책 결정이 되는 소득자료를 포함하고 있다. 그러나 소득자료는 민감한 개인정보를 포함하고 있어 그동안의 국세청 데이터는 마스킹 등 전통적인 가명처리 기법을 적용하여 제공되어왔기 때문에 자료 훼손 또는 관계 왜곡으로 유용성이 저하되었다.

국세청은 이에 대한 대안으로 재현데이터를 선택하고 재현데이터 우선 구축대상은 활용빈도가 높은 종합소득세, 근로소득세 등 소득 분야라고 밝혔다. 추후 시범 구축된 재현데이터를 이용해 데이터 활용도 및 정보 노출위험을 테스트할 예정이다.[39]

나. 해외 사례

1) 미국 SIPP Synthetic Beta(SSB)(2013~2023년)

SIPP Synthetic Beta(이하, 'SBB'이라 함)는 가구 조사에서 비롯된 개별 관측치 수준의 마이크로데이터를 세금 및 사회보장 데이터와 통합한 결과물이다. 원데이터에는 Survey of Income and Program Participation 조사에 응한 응답자들의 설문기록이 담겨있으며 사회보장행정(SSA)/내부수입서비스(IRS) 양식 W-2 기록과 퇴직 및 장애 급여 수령에 대한 행정 기록 등 민감정보가 포함되어 있어 노출위험이 존재한다. 이러한 노출위험을 축소하기 위해 미국 인구조사국(Census Bureau)은 2013년부터 매년 누적된 데이터를 통합하여 새로운 버전의 부분 재현데이터를 가공하여 배포하고 있다.

---

38) 국세청(2021)

39) https://www.etnews.com/20220315000132

Census Bureau 소속 경제학자, 통계학자와 대학 연구원들이 협력하여 세금, 수입 등 개인정보가 담긴 데이터를 부분 재현했다고 알려져 있으며, 데이터 구조의 보존을 위해 변수 간 관계를 유지하도록 하고 각 관측치 기록을 대치하는 방식으로 연구를 진행했다. 총 9개의 SIPP 패널데이터(Panel data)가 1984년, 1990년, 1991년, 1992년, 1993년, 1996년, 2001년, 2004년, 2008년에 걸쳐 공개되었고 현재는 이 패널데이터 정보를 포함한 업데이트된 데이터가 주기적으로 업데이트되고 있다. 2018년에 업데이트된 버전 7.0에서는 누락되어 있던 잠재변수들도 재현데이터로 구현되어 채워졌고 데이터 내 논리적 불일치가 나타나는 부분들이 수정되었다. 600개 이상의 변수로 구성되어 있는 이 데이터는 주요 변수에 대한 통계적 특성이 원데이터와 매우 유사하다고 알려져 있고 가장 광범위하게 공개된 재현데이터로 평가받는다.

이 데이터는 공개 전에 노출위험에 대한 심사를 받았고 SSB의 기록과 외부 데이터를 연결해도 개인정보 파악이 불가능할 것이라는 판정을 받았다. SSB는 최종적으로 데이터 공개 검토 위원회와 IRS, SSA위원회의 승인을 받은 후 서버에 안전하게 등록되었다. 현재 SSB는 코넬 대학교의 가상 연구 데이터 센터(Virtual Research Data Center)에 있는 SDS(Synthetic Data Server)에 저장되어 있다. 연구자는 서버에서 무료계정을 만들고 키를 얻으면 SSB 데이터를 사용할 수 있다.

2) 영국 SynAE project(2018~2023년)[40]

이 프로젝트는 영국의 NHS(National Health Service) Digital이 제공하는 SUS(Secondary Uses Service) 데이터에서 추출한 Attendances&Emergency 데이터와 입원 환자 치료 데이터를 결합·가공하여 재현데이터로 구현한 시범사업으로서 현재까지도 추가된 정보가 업데이트되고 있다. 환자의 개인정보 노출을 막으면서도 데이터 공유를 확대하기 위한 취지로 시작된 이 프로젝트는 영국 정부의 NHS England 의무조항을 따르도록 수행되었으며 원데이터에 민감정보가 다수 포함되어 있어 ONS(Office for National Statistics)[41]에서 이 프로젝트를 검수하였다.

이 프로젝트에서 밝힌 데이터 가공 방법은 다음과 같다. 먼저 지역 인구통계학적 정보에

40) https://open-innovations.org/events/synae/

41) 영국에서 가장 규모가 큰 공식 통계의 독립 생산기관이자 공인된 국가 통계기관임

기반하여 지리적 정보를 제거하고 세분화된 연속형 변수를 밴드로 그룹화한다. 예를 들어, 연속형인 연령 변수를 '0~5', '5~10', '10~15'와 같이 5단위로 끊어 재코딩하는 것이다. 또 입원 일시, 시각과 같이 정확한 시간정보를 제거하고 이상치 정보는 추출해 마스킹 처리를 하였다. 추가로 고유한 값을 가진 관측치 및 일부 희귀 부분 집합도 제거하여 노출 위험을 줄였다.

이후 1차 가공된 데이터에 대해 R 패키지 'BNLearn'을 적용하여 재현데이터를 생성했다. 이는 베이지안 네트워크에 기반한 방식으로 계층 구조에서 일련의 조건부 확률을 기반으로 데이터 모델을 생성하는 것이다. 계층 구조의 상단(Top)의 네트워크 구조를 활용해 각 변수 분포에서 표본을 추출하면, 계층 구조 하부(Bottom)에 위치한 변수들의 표본 추출 가능 범위가 줄어들기 때문에 신빙성 있는 데이터를 만들어 낼 수 있다.

이 프로젝트에서 산출된 재현데이터는 매년 공개되고 있다. 건강·보건 관련 데이터 공개는 다양한 건강정보를 활용한 서비스 개발을 가능하게 하기에 보건의료 분야에서 재현데이터 생성에 대한 관심이 높은 편이다.

3) 스위스 보험회사 Die Mobiliar(2021년)

Die Mobiliar은 스위스의 보험회사로 재현데이터 솔루션을 제공하는 기업인 Statice와 협업을 통해 재현데이터를 이용한 고객이탈 예측모델을 활용하고 있다.[42] 애초에 Die Mobiliar는 보유한 실제 고객데이터를 기반으로 모델을 만드려고 했으나, 스위스 데이터 보호법의 규제로 인해 대신 Statice의 합성데이터 솔루션을 사내데이터에 적용해 재현데이터로 예측모델을 학습시켰다고 한다. 원데이터로 훈련한 모델과 재현데이터로 훈련한 모델을 비교했을 때, 재현데이터 훈련 모델이 원본의 95% 성능을 보인다고 알려져 있다. 이를 이용해 Die Mobiliar은 효과적인 데이터 익명화를 통해 소비자 개인정보를 보호하면서도 효과적인 마케팅 전략을 수립할 수 있었다. Die Mobiliar의 사례는 금융권에서의 재현데이터의 사용이 매우 큰 효과를 가져올 수 있음을 시사한다.

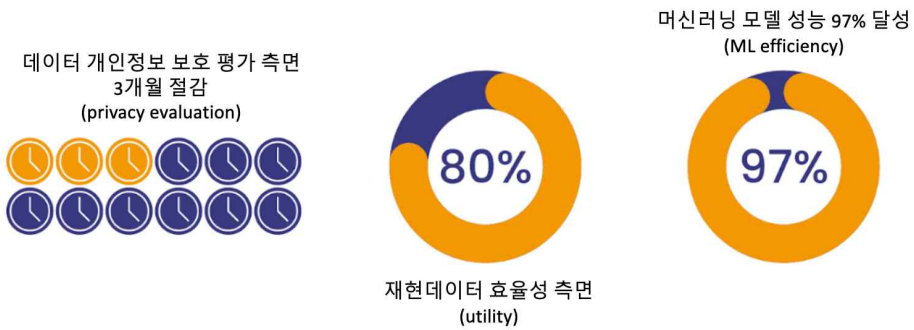
42) https://www.statice.ai/case-study/die-mobiliar

4) 독일 보험회사 Provinzial Rheinland(2021년)

재현데이터를 이용한 또 다른 해외 사례는 독일 보험회사 Provinzial Rheinland(이하, 'Provinzial'이라 함)에서 찾을 수 있다. Provinzial은 이미 소비자의 종류별 보험 수요를 예측하고 그에 맞는 제품을 추천하는 'Next best offer' 모델을 보유하고 있었으나, 이 마케팅 모델을 강화하기 위해 재현데이터를 활용하였다.

Provinzial은 데이터 가용성, 모델 사용, 개인정보보호 규정 세 가지 지표를 만족하는 재현데이터를 생성했고 성능을 테스트하는 과정을 진행하였고, 원데이터와 재현데이터를 비교한 결과 재현데이터의 80% 정도를 사용할 수 있음을 확인했으며 재현데이터로 학습시킨 모델의 성능이 원본 성능의 97% 이상을 충족했다고 한다. 이러한 결과는 원데이터와 재현데이터를 적절히 섞어 활용해 개인정보 노출 우려 없이 기존 머신러닝 모델을 발전시킨 것에 의의가 있다.

〈그림 Ⅲ-14〉 Provinzial사 모델 성능 결과



자료: Statice(https://www.statice.ai/case-study/provinzial-predictive-analytics-synthetic-insurance-data)

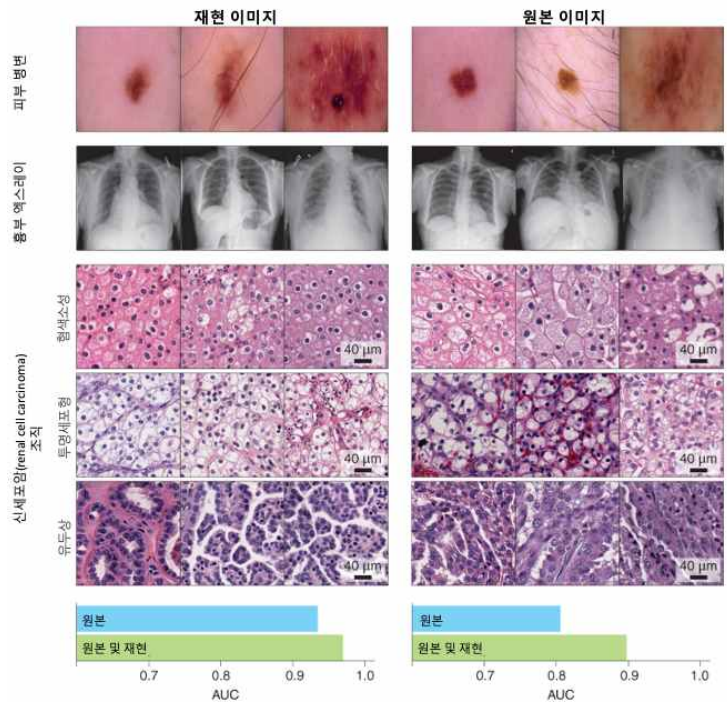
5) 해외 논문 사례[43]

AI 분야에서 재현데이터를 이용한 이미지 관련 연구가 활발히 진행되고 있는데, 이 중 헬스케어 분야에서 재현데이터를 활용한 최근 논문 사례를 하나 소개한다.

43) Chen, R. J. et al.(2021)

이 논문은 의료학습용 이미지 데이터 부족의 해결책으로 재현데이터를 제시하고 있다. 예를 들어 특정 희귀질환 관련 의료 이미지를 분석하고 싶은 경우, 그 데이터 수가 현저히 부족하기 때문에 데이터 불균형으로 인한 편향 문제가 발생할 수 있다. 또한 병원마다 사용하는 의료장비가 다르기 때문에 데이터 표현에 이질성이 생기면 분석 정확도가 떨어질 수도 있다. 저자는 수집된 원래의 영상으로부터 대규모 합성데이터를 생성해 이 문제들을 해결할 수 있음을 주장하였다.

〈그림 Ⅲ-15〉 의료 이미지 재현데이터 예시



자료: Chen, R. J. et al.(2021)

〈그림 Ⅲ-15〉은 해당 논문에 제시된 그림이다. 피부 병변, 흉부 엑스레이 및 신장 세포 암의 세 가지 유형에 대해 우측에는 실제 이미지를, 좌측에는 GAN으로 생성된 재현데이터 이미지를 보여준다. 그림의 하단에는 재현데이터를 추가한 경우 AUC값이 증가해 예측모형의 성능이 더 좋아짐을 보여준다. 또한 해당 논문에서는 Visual Turing test[44]를 이용해

44) 자세한 내용은 Geman, D. et al.(2015)을 참고하길 바람

이미지 재현데이터 채택 및 평가 문제를 해결할 수 있을 것이라 제안하였다.

이 연구 사례는 의료 영상 분야에서의 재현데이터 사용이 개인정보보호 문제를 해결해 줄 뿐 아니라 학습데이터의 편향 문제도 해결해 사용하고자 하는 모형의 분석과 예측의 성능을 개선하는 데에도 도움을 줄 수 있음을 시사한다.

4. 재현데이터의 활용과 한계점

재현데이터는 알고리즘을 이용하기 때문에 제한 없이 빠르게 생성할 수 있으며 다음의 장점이 있다.

첫째, 재현데이터는 개인정보보호 등의 규제로부터 자유롭고 익명데이터로 분류되므로 사용자들은 이를 자유롭게 공유·전송·거래할 수 있다. 원데이터의 민감변수는 데이터 사용자들에게 공개되지 않지만, 재현데이터는 민감정보까지 포함하여 공개된다. 따라서 외부의 추가정보를 이용하더라도 민감정보 유출 가능성이 최소화되고 기존의 가명·익명처리 기법과 같이 사용할 수 있다. 이러한 특성으로 인해 재현데이터 사용자는 데이터의 모든 속성을 확인할 수 있고, 이를 이용하여 다양한 분석을 시행할 수 있게 된다.

특히 금융데이터의 경우 재현데이터의 생성기를 금융사의 내부망에 탑재해 사용할 수 있어 원데이터의 외부 유출에 대한 우려가 없다. 더 나아가 회사 내 부서별로 재현데이터를 독립적으로 생성한 후 부서별 재현데이터를 병합·결합해 전사적인 재현데이터를 구축할 수도 있다. 재현데이터가 규제로부터 자유로운 익명데이터임을 고려할 때, 이와 같은 방식은 개인정보가 담긴 금융데이터에 대해 엄격한 현행 규제 하에서 사내 수집된 다양한 빅데이터의 활용을 극대화할 수 있는 현실적인 대안이라고 하겠다.

둘째, 재현데이터는 모든 분야에서 더 많은 양의 고품질 학습데이터를 구축하는데 사용될 수 있다. 재현데이터의 생성을 통해 불균형 데이터의 문제를 해결할 수도 있고, 데이터 편향이 존재하는 경우에도 보정된 재현데이터를 사용해 분류기나 다른 지도학습모형의 성능 향상에 도움을 줄 수도 있다.[45)]

45) 데이터의 양과 질을 개선하는 작업을 데이터 증강이라고 하기도 함

그러나 재현데이터 사용 시 주의해야 하는 점도 존재한다. 원데이터와 완전히 같게 복제한 것이 아니기 때문에 원데이터와 비교할 때 오차가 존재할 수 밖에 없다.[46] 반대로 생각하면 재현데이터의 생성과정에서 확률은 낮지만 우연히 원자료와 비슷한 값이 나올 가능성이 있으므로 이를 제어하는 장치가 필요하기도 하다. 또한 원데이터의 성격에 따라 재현 알고리즘이 다를 수 있다. 예를 들어 횡단면 연구인지, 시계열 혹은 패널형태인지, 위계가 있는지, 극단값이 있는지에 따라 적합한 알고리즘이 다를 수 있다. 다양한 상황에서 적절한 재현 알고리즘을 구현하는 문제는 앞으로의 연구 문제로 남아있다. 마지막으로 선택된 모형과 알고리즘에 따라 재현데이터의 품질이 달라지므로, 주어진 재현데이터의 품질을 원데이터와 비교하고 평가할 수 있는 척도 또한 필요하다. 재현데이터의 효용과 노출위험을 측정하는 방법에 대해서는 본 보고서의 IV장에서 설명한다.

46) 여러 세트의 재현데이터를 생성해 제공하면 분석 오차의 크기 측정과 튜닝을 통해 오차 크기 조절이 가능함

Ⅳ 재현데이터의 효용과 노출위험 측정

1. 노출위험

데이터 노출위험 종류는 다음과 같은 세 가지로 분류할 수 있다. 첫째, 신원 노출(Identity disclosure)의 위험이다. 이 위험은 데이터 침입자가 공개된 다른 데이터 레코드와 결합하는 경우 발생한다. 예를 들어 데이터 내부 특정 관측치들을 외부 정보와 연결했을때 극단적인 값을 가진 개인이 있다면 이를 식별해 낼 수 있다. 여기서 극단치는 매우 큰 숫자일 수도 있지만 매우 드문 범주가 될 수도 있다. 식별이 성공하면 침입자는 노출된 관측치의 다른 모든 정보 접근도 가능하게 된다. 둘째, 속성 노출(Attribute disclosure)의 위험이다. 속성 노출위험은 신원 노출 없이도 발생할 수 있는데, 예를 들어 어떤 병원에 입원한 56~60세의 모든 여성 환자가 암에 걸렸다는 것을 보여주는 데이터가 공개된 경우, 침입자는 특정 개인을 식별할 필요 없이 이 병원에 입원한 적이 있는 모든 56~60세의 여성 환자의 의학적 정보를 알 수 있게 된다. 즉, 개인의 특정 속성이 노출되는 것이다. 마지막으로 추론 노출(Inferential disclosure)위험인데 이는 예측 성능이 뛰어난 모형이 주어졌을 때 데이터에 기록된 속성을 모형에 대입하면 민감한 개인정보를 신뢰할 만한 수준으로 추론할 수 있게 되는 노출위험을 말한다.[47]

공개된 데이터가 안전한지를 확인하기 위해서는 이를 측정할 수 있는 기준이 있어야 한다. 이것이 위험 측정(Risk measurement)이라는 개념이다. 적절한 위험 측정이 가능하다면 데이터 관측치별 위험성을 파악하여 특정 관측치들에만 가명처리 기법을 적용할 수도 있고, 데이터 변수(테이블 데이터의 경우 칼럼)별로 통계적 노출 제어를 적용할 수도 있다. 전통적인 위험 측정 방법은 다음 세 가지이다. 먼저 k-익명성[48]은 노출 제어 이후 데이터에 각각의 익명화된 범주에 속하는 개체의 개수로 노출의 위험을 측정하는 방법이다. 각 익명화된 범주에 속하는 개체의 개수가 임계치인 k개 이상이면 노출의 위험이 낮다고 판단한다. 두 번째로 l-다양성[49]은 k-익명성 노출위험 측정 후에도 특정 동질집합에서

47) Park, M. J., and Kim, H. J.(2016)

48) Samarati and Sweeney(1998)

서로 같은 민감한 정보를 재식별이 가능하다는 단점을 보완한 방법이다. 주어진 데이터 집합에서 가명처리되는 개체들은 동질집합에서 적어도 l개의 서로 다른 민감한 정보를 가졌는지 확인하여 노출위험을 판단한다. 마지막으로 t-근접성[50]은 민감정보 분포와 전체 데이터의 민감한 정보 분포의 거리 차이가 t 이하이면 노출의 위험이 낮다고 판단하는 방법이다. 거리를 측정하는 방법으로는 〈그림 Ⅳ-1〉과 같이 Variational distance, Kullback-Leibler divergence, Earth Mover's distance 방법 등이 있다. 하지만 k-익명성, l-다양성, t-근접성의 경우 사용할 모수인 k, l, t의 값의 결정이 자의적일 수 밖에 없다는 점에 유의해야 한다.

〈그림 Ⅳ-1〉 t-closeness 거리 측정법

거리 측정 방법	식
Variational distance	$d(p,q)=\sum_{i=1}^{m}\frac{1}{2}\lvert p_i-q_i\rvert$
Kullback–Leibler divergence	$d(p,q)=\sum_{i=1}^{m}p_i\log\frac{p_i}{q_i}=H(p)-H(p,q),$ $where\ H(p)=\sum_{i=1}^{m}p_i\log p_i,\ H(p)=\sum_{i=1}^{m}p_i\log q_i$
Earth Mover's distance	$d(p,q)=Work(p,q,f)=\sum_{i=1}^{m}\sum_{j=1}^{m}d_{ij}f_{ij}$

자료: 김승현(2020)

2. 데이터의 효용과 노출위험의 상충관계

일반적으로 데이터 노출 제어 정도 혹은 정보보호의 강도가 높을수록 원데이터가 가지는 유용성 혹은 효용(Data utility)[51]이 떨어지게 된다. 이는 원데이터 정보에 손실이 생겼다는 의미로, 사용자가 의도한 데이터 기반 모형 구현을 통한 분석·예측이 실패하거나 왜곡될 가능성이 증가하게 된다. 반대로 데이터의 정보보호 수준을 낮춘다면 데이터의 효용은 높아지겠지만 반대로 노출위험은 그만큼 커진다. 이를 데이터의 효용-정보보호의 상충관

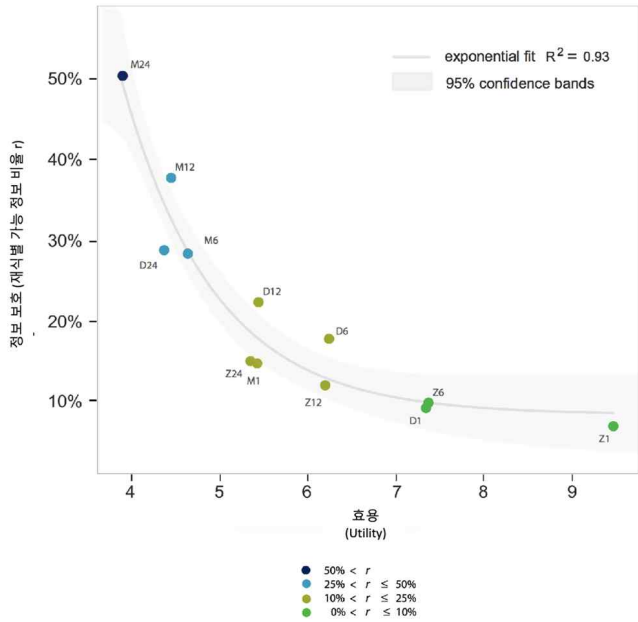
49) Machanavajjhala et al.(2007)

50) Li et al.(2007)

51) 데이터 효용의 측정에 대해서는 다음 절에서 자세히 설명함. 여기서 데이터의 효용은 원데이터가 가지고 있는 유용한 정보 정도의 원론적인 정의로 사용했음

계(Utility-Privacy trade-off)라 하는데 〈그림 Ⅳ-2〉는 이러한 관계를 보여주고 있다.

〈그림 Ⅳ-2〉 데이터 정보보호와 효용의 상충관계 예시



자료: Calacci, D. Berke, A., Larson, K., and Pentland, A.(2019)

〈그림 Ⅳ-2〉에서 가로축은 데이터 유용성을, 세로축은 데이터 정보보호의 수준 혹은 강도를 나타내며 이 둘의 상충관계는 아래로 볼록한 곡선로 표현된다. 그림 내에서 우하단의 데이터셋 Z1은 최대의 유용성을 가지지만 동시에 최하의 정보보호 수준을 보여주고 있어 노출 제어가 되지 않은 원데이터에 가깝고, 좌상단 데이터인 M24의 경우 반대로 최고의 정보보호 수준을 보여주지만 유용성이 최하임을 알 수 있다. 즉, Z1을 원데이터라고 할 때 여기 노출 제어의 수준을 높힘에 따라 커브를 따라 좌상의 방향으로 움직여 노출 제어의 수준을 최대로 끌어올리게 되면 M24의 위치로 이동하게 되는 것이다. 이처럼 유용성과 정보보호를 동시에 최대화할 수는 없다는 사실을 받아들인다면, 가장 적절한 데이터 노출 제어의 수준은 그림에서 커브의 중간지점에 존재하는 데이터에 상응한다는 것을 알 수 있다. 그러나 이를 위해서는 데이터의 유용성과 정보보호(혹은 노출위험)를 동시에 측정하는 작업이 선행되어야만 하는데 이 둘을 동시에 측정하는 연구는 최근에 와서야 등장하고 있다.

3. 효용과 노출위험의 측정 기법들

통계적인 관점에서 볼 때 잘 생성된 재현데이터는 원데이터가 속한 동일한 모집단으로부터 독립적으로 랜덤 추출된 또 다른 표본으로 이해할 수 있다. 따라서 좋은 재현데이터는 원데이터와 충분히 유사하여 그 효용을 보존하는 동시에 개인정보 노출위험을 최소화하기 위해 충분히 다른 것이어야 한다. 재현데이터의 유용성과 정보보호는 서로 상충관계에 있어 이 둘 사이의 균형을 달성하는 것은 데이터 재현(합성)의 중요한 목표이자 고려해야 하는 핵심사항이다.

사실 위의 상충관계는 재현데이터뿐 아니라 다른 비식별화 처리된 데이터에 대해서도 마찬가지이지만 여기서는 논의의 편의를 위해 재현데이터로 지칭하겠다. 재현데이터의 효용측정은 문헌에서 두 개의 다른 수준에서 논의되어왔다. 첫째, 개별 분석의 수준에서 재현데이터와 원데이터 간의 데이터 요약 또는 추정된 통계량을 비교하는 방식이 있다. 즉, 두 데이터에서 각각 계산된 분석 통계량의 값이 충분히 가까우면 재현데이터는 높은 효용을 갖는 것으로 간주되는 것이다. 그러나 이러한 분석 특정(Analysis-specific) 측정 방법은 사용자가 수행할 분석의 종류에 따라 측정된 효용의 정도가 다르게 된다. 이 방식은 사용자가 수행하는 분석의 종류나 범위가 알려져 있지 않는 것이 일반적이기 때문에 적용가능한 범위가 제한적이다. 이러한 단점을 극복하기 위해, 특정 분석이 아니라 표본의 수준에서 데이터 효용을 측정하는 척도 역시 연구되어왔다. 이는 재현 전후의 데이터 분포의 차이를 측정함으로써 모든 가능한 분석에 대한 데이터의 효용을 평가하는 것으로 글로벌(Global)척도라 부른다. 최근 소개된 경향점수평균제곱오차(Propensity Score Mean Squared Error; PMSE) 기법[52]이 대표적인 예이다.

PMSE는 분류기로부터 얻어지는 경향점수에 기반한다. 설명을 위해 테이블 형태의 정형 원데이터와 이로부터 생성된 재현데이터를 얻었다고 가정하자. 이 두 데이터를 상하로 이어붙인 후 새로운 변수(열)를 추가하고 각 행이 원데이터(0)인지 재현데이터(1)인지의 정보를 입력한다. 다음 단계로 통합된 데이터에 분류기를 적용해 각 행이 실제 속하는 데이터로 잘 분류가 되는지(0/1)를 확인하게 되는데, 여기서 분류기는 로지스틱회귀, CART, 혹은 다른 이진 분류기(Binary classifier)를 적절히 사용할 수 있다. 이제 n_0와 n_1을 각각 원데이터와 재현데이터의 크기라고 하고, 분류기가 각 행에 대해 재현데이터(1)에 속할

52) Snoke, J. et al.(2018)

확률을 추정한 값을 $\hat{p_i}$라고 하자. 즉 $\hat{p_i}$ 값이 1에 가까울수록 재현데이터로 분류될 가능성이 높아진다. 만약 재현이 성공적이라면 분류기는 재현 전후의 데이터를 구별하기 어렵게 되어 $\hat{p_i}$ 값이 $c = n_1/(n_0 + n_1)$ 에 가깝게 될 것이 자명하다. 이러한 성질을 이용하기 위해 PMSE는

$$pMSE = \frac{1}{n_0 + n_1} \sum_{i=1}^{n_0+n_1} \left(\hat{p_i} - c\right)^2.$$

으로 정의된다. 따라서 성공적인 재현데이터에 대해서 PMSE는 0에 가깝게 되고, 반대로 재현데이터와 원데이터의 차이가 많은 경우 PMSE의 값은 커지게 됨을 알 수 있다. PMSE가 0에 충분히 가까운지 아닌지를 구분하는 경계에 대한 논의는 해당 논문에서 확인할 수 있다.

그러나 이 기법에는 몇 가지 제한이 있다. 첫째, 재현데이터와 원데이터가 동일한 모집단 분포를 따르더라도 점근적 PMSE의 값이 매우 다를 수 있는데 이는 제안된 방법이 재현데이터 생성 방법에 크게 의존함을 뜻한다. 둘째, PMSE 기법은 분류기에 민감하다고 알려져 있다. 즉 동일한 재현데이터에 대해 어떤 분류기를 사용하느냐에 따라 PMSE 값이 상당히 다를 수 있다는 것이다. 마지막으로, PMSE 기법은 데이터의 효용만을 평가하며 개인정보 노출위험 대한 정보를 제공하지 않는다.

전통적인 노출위험 척도인 k-익명성, l-다양성, t-근접성도 모수의 결정이 임의적이라는 점과 데이터 효용을 계량적으로 측정할 수 없다는 약점이 있다. 좀 더 최근의 척도로는 앞서 기술한 차등적 정보보호(DP)를 들 수 있다. 데이터 정보보호 연구에서 유용한 도구로 인정되고 있지만 DP 역시 한계점이 존재한다. 첫째, 반복되는 쿼리에 대해 점차적으로 효율이 떨어진다. 둘째, 쿼리의 종류에 따라 적절한 매개변수 ε의 값을 결정하는 것은 현실적으로 매우 어려운 문제이다. 실제로 특정 매개변수 값이 경우에 따라 매우 다른 정보보호 수준을 의미할 수 있어, 쿼리(분석)의 종류나 범위가 알려져 있지 않은 것이 일반적인 상황에서 DP를 사용한 마이크로데이터의 노출 제어는 현실적으로 매우 어렵다. 이 두 단점은 앞에서도 기술하였지만, 여기서는 추가로 DP 기법 역시 노출위험을 제어한 후 감소한 데이터의 효용을 측정할 수 없다는 점을 지적한다.

4. DUPI: 분포 기반 재현데이터 품질평가 척도

위와 같이 현재 사용되는 척도들은 대부분 데이터의 효용만을 혹은 노출위험만을 측정하고 있는데, 서로 상충관계에 있는 두 측면을 동시에 측정하는 척도가 필요함은 자명하다. 이 둘을 같이 고려한 재현데이터의 성능을 품질이라고 할 수 있는데, 사실 재현데이터 연구에서 재현된 데이터의 품질 측정은 재현데이터의 생성기법만큼이나 중요한 주제이다. 새로운 노출 제어(가명·익명처리 혹은 재현데이터) 기법은 결국 노출 제어를 통해 산출된 데이터의 품질을 통해 평가받아야만 하기 때문이다.

이와 관련해 기억해야 할 중요한 내용은 재현 전후 데이터를 비교할 때 분석특정(Analysis-specific) 척도의 한계점이다. 앞서 보험데이터를 이용한 예시에서 우리는 보험가입 여부를 반응변수로 둔 로지스틱회귀모형을 포함해 다양한 비교를 하였다. 만약 사용자가 특정한 분석만을 진행할 것이라는 것을 미리 안다면 재현 전후 데이터의 효용과 노출위험을 쉽게 통제할 수 있을 것이지만, 일반적으로 우리는 재현된 데이터가 어떤 분석이나 쿼리에 이용될지 알 수 없다. 예를 들어, 동일한 보험데이터에서 다른 변수를 반응변수로 두고 일반화선형모형으로 이용할 수도 있고, 전혀 새로운 모형으로 실험을 할 수도 있을 것이다. 또한 앞서 살펴본 두 변수 사이의 공분산이나 상관계수 역시 저차원의 적률만을 비교하는 것이기 때문에 고차원의 종속성을 측정하지 못한다. 이를 통해 우리는 가능한 모든 쿼리나 분석에 대해 재현데이터의 효용을 측정하는 것은 불가능함을 알 수 있다. 더 나아가 동일한 재현데이터를 사용하더라도 쿼리나 분석의 종류와 범위에 따라 품질도 다르게 측정될 것이다. 이런 이유로 재현 전후 데이터를 비교할 때는 특정한 분석 결과가 어떻게 달라지느냐를 비교하는 것도 필요하지만, 더 근본적으로는 데이터의 분포가 어떻게 달라지는지를 측정할 필요가 있다. 엄밀하게 말해 분석의 결과가 유사한 것과 두 데이터가 유사한 분포를 가지는 것은 매우 다른 이야기일 수 밖에 없기 때문이다. 결론적으로, 재현데이터의 품질을 측정할 때 분포 수준에서 원데이터와 비교하는 것이 합당한 방법이며 이를 확인하는 과정이 반드시 필요하다.

이러한 맥락에서 최근 개발된 재현데이터의 효용과 노출위험을 동시에 측정하는 품질척도인 DUPI(Data Utility and Privacy Index)[53]를 소개한다. PMSE 기법과 마찬가지로 특정한 쿼리(분석)에 의존하지 않는 글로벌척도로서 제안된 DUPI는 재현 전후 데이터 사이

53) Jeong, D., Kim, J. H., and Im, J.(2022)

의 거리를 분포에 기반해 확률적으로 측정하는 방식을 이용한다. 기본적인 개념은 원데이터가 재현데이터와 동일한 다변량 모집단에서 독립적으로 추출된 표본일 경우 가장 이상적이라 볼 수 있다는 사실에서 출발한다. DUPI는 비모수적 방법으로 모집단의 분포를 특정하지 않는다는 장점이 있고 연속형과 범주형을 둘 다 포함하는 데이터에 적용가능하다.

원론적인 수준에서 볼 때 재현 전후 데이터 사이의 거리가 지나치게 가까우면 두 데이터의 유사성이 충분히 강해 재현데이터가 원데이터의 효용을 유지하지만 동시에 노출위험 역시 원데이터와 비슷한 수준으로 높아진다. 반대로 거리가 멀수록 재현데이터는 원데이터와의 유사성이 줄어 효용 측면에서 정보손실이 커지지만 그만큼 노출위험이 감소하게 된다. DUPI는 이러한 확률적 거리 개념을 이용하여 재현데이터의 효용과 노출위험을 동시에 측정 및 수치화하고, 재현된 데이터가 상충관계 사이에서 어느 지점에 위치하는지 좌표평면에 그림으로도 시각화할 수 있다. 또한 원데이터와 재현데이터가 동일모집단에서 추출되었다는 가정을 이용하여 재현데이터가 가질 수 있는 최적 위치 역시 제공한다.

DUPI의 이론을 간략히 살펴보기 위해 먼저 $X_n = \{x_1, \cdots, x_n\}$를 크기가 n인 원데이터로, $Y_m = \{y_1, \cdots, y_m\}$을 크기가 m인 재현데이터로 두자. 여기서 m은 n과 다를 수 있다. 이제 재현 전후 데이터 사이의 거리를 측정하기 위해 거리함수가 필요한데, 이를 위해 특정한 두 점 x와 y 사이의 거리를 계산하는 함수를 $d(x,y)$라고 하겠다. 거리함수는 유클리드 거리 혹은 범주형을 포함하는 경우 HEOM(Heterogeneous Euclidean-Overlap Metric) 거리 등을 사용할 수 있다. 다음으로, 어떤 고정된 점 c과 임의의 데이터 A 사이의 거리를 재기 위해 c와 A의 개별원소들 사이의 거리들을 모두 계산하고, 그 중 k번째로 작은 값을 $d_A^{<k>}(c)$로 둔다. 다시 말해

$$d_A^{<k>}(c) = \{d(a,c) : a \in A\}^{<k>}$$

로 정의한다. 이를 이용하면 재현 전후 데이터가 동일한 모집단에서 추출되었다는 귀무가설하에서 다음의 식을 증명할 수 있다.

$$\frac{1}{n}\sum_{i=1}^{n} P\left(d_{Y_m}^{<k>}(x_i) \le d_{X_{n\setminus i}}^{<k>}(x_i)\right) = \sum_{s=k}^{2k-1} \frac{\binom{s-1}{k-1}\binom{n+m-s-1}{m-k}}{\binom{n+m-1}{m}}$$

여기서 $X_{n\setminus i}$는 원데이터 X_n에서 i번째 관측값 x_i이 빠진 것이다. 만약 i번째 관측값이 포함되면 x_i는 자신과의 거리인 0을 산출하게 되기 때문이다. 즉, $d_{X_n}^{<1>}(x_i)=0$이 되는 상황을 피하기 위한 장치이다. 위의 이론적 결괏값을 $DUPI_0^{<k>}$로 정의하고, 이 값의 표본 추정량을 $DUPI^{<k>}$로 두면 지시함수 $I(\cdot)$를 이용해

$$DUPI^{<k>}=\frac{1}{n}\sum_{i=1}^{n} I\left(d_{Y_m}^{<k>}(X_i)\le d_{X_{n\setminus i}}^{<k>}(X_i)\right)$$

를 얻는다. 논문에서는 k값을 1로 두고 있는데 이는 큰 k값을 선택할 경우 데이터의 극단치 혹은 이상치에 민감하게 되고 $k=1$만으로도 재현 전후 데이터의 특징을 충분히 비교할 수 있기 때문이다. 따라서, 위의 식에 의해 $DUPI_0=DUPI_0^{<1>}$, $DUPI=DUPI^{<1>}$가 된다. 이렇게 계산된 $DUPI$ 값은 0과 1 사이의 값을 지니며 다음의 해석이 가능하다.

- $DUPI$값이 1에 가까우면 재현 전후 데이터가 지나치게 유사해 생성된 재현데이터가 효용은 높지만 노출위험 또한 크다.
- 반대로 $DUPI$값이 0에 가까우면 재현 전후 데이터가 지나치게 달라서 생성된 재현데이터의 노출위험은 최소화할 수 있지만 원데이터의 정보를 거의 보존하지 못한다.
- 이상적인 $DUPI$값은 $DUPI_0$로서 이 경우 귀무가설을 만족해 재현데이터가 원데이터와 동일한 모집단에서 독립적으로 생성된 표본으로 볼 수 있다.

이러한 성질을 이용해 논문은 재현데이터의 $DUPI$값이 효용과 노출위험의 관점에서 최적점인 $DUPI_0$에서 얼마나 떨어져 있는지에 대한 시각화 기법을 제안한다. 시각화를 위해 먼저

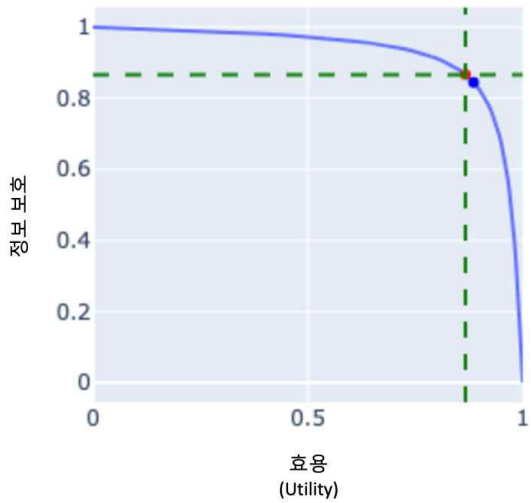
$$g(DUPI)=\begin{cases}\dfrac{DUPI}{2DUPI_0}, & 0\le DUPI\le DUPI_0,\\ \dfrac{DUPI-DUPI_0}{2(1-DUPI_0)}, & DUPI_0<DUPI\le 1.\end{cases}$$

를 정의하고, nuisance 모수인 τ를 이용해

$$UI(DUPI) = \frac{\arctan[\tau \times g(DUPI)]}{\arctan(\tau)}$$
$$PI(DUPI) = \frac{\arctan[\tau \quad - \quad \tau \times g(DUPI)]}{\arctan(\tau)}$$

로 변환을 하고, 이로부터 주어진 재현데이터의 (UI, PI) 좌표값을 도출해 낸다. 여기서 디폴트 τ값은 5로 지정되었다.

〈그림 Ⅳ-3〉 DUPI 시각화 예시



자료: Jeong, D., Kim, J. H., and Im, J.(2022)

〈그림 Ⅳ-3〉은 앞서 예시로 보인 보험데이터에 DUPI를 적용해 시각화한 결과이다. 그림에서 가로축은 데이터의 효용(Utility Index; UI)을, 세로축은 노출보호(Privacy Index; PI)를 뜻하며 두 축 모두 [0,1] 구간에 존재한다. 위로 볼록한 곡선은 주어진 데이터에 대해 가능한 모든 (UI, PI) 조합의 매핑인데, 곡선의 우측하단으로 갈수록 효용은 크지만 노출보호가 낮은 재현데이터를, 반대로 좌측상단은 높은 노출보호가 보장되지만 낮은 효용을 가지는 재현데이터를 나타낸다. 이 그림은 앞서 소개한 〈그림 Ⅳ-2〉와 비슷한 정보를 제공하지만 세로 축의 의미가 반대이므로 곡선 역시 서로 대칭적으로 보여진다.

위 그림에서 최적의 재현데이터는 점선으로 표시된 두 직선이 교차하는 지점에 있는 빨간

색 점에 위치하며 효용과 노출위험이 가장 적절히 상쇄되는 기준점이다. 곡선 위에 찍혀 있는 검은 점은 앞장에서 살펴본 보험데이터를 CART 알고리즘으로 재현한 데이터의 위치를 나타낸다. 이 점은 최적점에 비해 우하단에 위치해 있어 재현된 데이터가 최적데이터보다 비교적 높은 수준의 효용을 가지고 있지만 노출보호는 상대적으로 약하다는 것을 알 수 있다. 그럼에도 불구하고 재현데이터는 최적점과 가까운 곳에 위치해 있어 재현이 잘되었다고 평가할 수 있다. 그림에서 표현된 재현데이터와 기준점의 좌표를 도출하고 수치적 비교를 하는 과정은 원논문[54]을 참고하면 되겠다.

그러나 데이터의 크기가 커지면서 계산량이 지수적으로 증가하고, 재현데이터 품질을 표현하는 척도가 점 추정으로만 되어있어 이 추정량의 불확실성이 어느 정도가 되는지는 알 수 없다는 점에서 DUPI 방법론은 개선될 여지가 있다고 하겠다.

또 다른 한계점으로 최적점에서 어느 정도로 멀거나 가까울 때 노출위험이 없다고 할 수 있는지에 대한 수치적인 가이드라인이 없다는 점을 지적할 수 있다. 그러나 이는 현재 존재하는 모든 계량적인 기법과 척도에 공통적으로 해당되는 사항이다. 재현데이터를 원데이터와 비교했을 때 원데이터와 완전히 동일한 관측값이 없다면 일차적으로는 노출위험이 없다고 할 수 있다. 그러나 완전히 동일하지는 않지만 유사한 관측값은 존재할 것이고 이를 다양한 외부 데이터와 연결한다면 개인정보의 일부라도 노출이 될 수도 있다는 점에서, 모든 가능한 시나리오하에서 노출위험을 수리적으로 0으로 만드는 것은 불가능하다. 따라서 노출위험은 없애는 것이 아니라 최소화하는 것이며 이를 위해 전문가나 규제당국이 토론과 경험을 통해 허용 가능한 노출위험의 수준을 정하는 논의가 필요하다. 비슷한 예로서 금융사의 자산건전성을 관리·감독하는 국제적 규제인 바젤Ⅱ 어코드의 경우 위험자본의 양을 계산할 때 Value-at-Risk(VaR) 99.9% 수준을 요구하고 있다. 이 수치 역시 이론적인 근거로부터 도출된 것이 아니라 특정 위험을 100%의 확률로 차단하거나 피하는 것이 현실적으로 불가능하다는 것을 받아들여 바젤위원회가 금융사의 이윤추구와 소비자·투자자의 공익이라는 상충관계를 고려해 결정한 위험허용의 수준으로 볼 수 있다.

54) Jeong, D., Kim, J. H., and Im, J.(2022)

V 시사점 및 결론

1. 가명정보와 익명정보의 차이

현재 개인정보보호에 관한 제도와 규제는 나라마다 다르지만 세계적으로 관련 규제가 점차 세부적이고 엄밀하게 수정·보완되는 추세이다. 당분간 계속될 것으로 보이는 이러한 추세로 인해 앞으로 개인정보가 포함된 양질의 데이터를 확보하는데 드는 시간과 비용은 지속적으로 증가할 것이다.

데이터에 관한 대표적인 해외 규제로는 유럽(EU)의 개인정보보호법(General Data Protection Regulation; GDPR), 미국 캘리포니아의 소비자정보보호법(California Consumer Privacy Act; CCPA), 그리고 ISO 표준(ISO standard 29100:2011) 정도이다. 이들 규제에서 공통적으로 기술하는 가명·익명정보의 대략적인 내용은 다음과 같다.

첫째, 가명정보(Pseudonymized information or data)는 추가적인 외부정보 없이 개인의 재식별이 불가능한 정보나 데이터로서 규제 대상이 된다. 이는 반대로 외부정보(결합키나 다른 출처의 데이터 등)를 적절히 연결할 경우 개인의 재식별이 가능한 정보라는 의미이기도 하다. 가명정보의 사용범위는 보통 공공의 목적이나 학술적 연구에 제한되며 사용절차와 방식 역시 규제가 적용된다. 가명정보 역시 개인정보로 분류되기 때문이다.

둘째, 익명정보(Anonymized information or data)는 추가적인 외부정보를 이용하더라도 개인의 재식별이 불가능한 정보로서 관련 규제의 적용대상이 아니다. 익명정보는 더 이상 개인정보가 아니기 때문이다. 따라서 익명정보로 분류된 데이터는 자유로운 사용과 공유가 가능하다. 하지만 어떤 과정을 거쳐야 익명데이터가 될 수 있는지, 즉 익명처리 절차에 대한 구체적 가이드라인은 아직 없어 보인다.

국내에서도 유사한 개인정보보호 관련 데이터 규제가 있는데 대표적으로 개인정보보호법과 신용정보법을 들 수 있다. 논의에 앞서 각 규제별로 가명정보와 익명정보의 정의를 비교하면 다음과 같다(2023년 5월 기준).

가. 개인정보보호법(제2조)

- 가명처리: 개인정보의 일부를 삭제하거나 일부 또는 전부를 대체하는 등의 방법으로 추가 정보가 없이는 특정 개인을 알아볼 수 없도록 처리하는 것을 말한다.
- 가명정보: 가명처리함으로써 원래의 상태로 복원하기 위한 추가 정보의 사용·결합 없이는 특정 개인을 알아볼 수 없는 정보이다.

개인정보보호법에서는 익명처리나 익명정보에 관한 정의를 찾을 수 없으며, 대신 제58조의2(적용 제외)에 "이 법은 시간·비용·기술 등을 합리적으로 고려할 때 다른 정보를 사용하여도 더 이상 개인을 알아볼 수 없는 정보에는 적용하지 아니한다."고 기술하고 있다.

나. 신용정보법(제2조)

- 가명처리: 추가정보를 사용하지 아니하고는 특정 개인인 신용정보주체를 알아볼 수 없도록 개인신용정보를 처리하는 것을 말한다.
- 가명정보: 가명처리한 개인신용정보를 말한다.
- 익명처리: 더 이상 특정 개인인 신용정보주체를 알아볼 수 없도록 개인신용정보를 처리하는 것을 말한다.

개인정보보호법과 신용정보법은 방대하고 전문적인 내용을 담고 있으며 지속적으로 보완·개정되고 있어 실무에서 사용하는 데 어려움이 크다. 이를 위해 법에 기술된 내용의 이해를 높이고 안전한 가명·익명정보의 활용을 위해 구체적인 설명과 예시를 제시하는 지침서가 있는데, 개인정보보호법을 위한 가명정보 처리 가이드라인과 신용정보법을 위한 금융분야 가명·익명처리 안내서가 그것이다. 각 지침서에서 기술하는 가명정보와 익명정보의 정의는 다음과 같다. 기본적으로는 개인정보보호법과 신용정보법의 내용을 담고 있으나 부분적으로는 차이도 있다.

다. 가명정보 처리 가이드라인(2022. 4)[55]

- 가명처리: 개인정보의 일부를 삭제하거나 일부 또는 전부를 대체하는 등의 방법으로 추

55) 개인정보보호위원회(2022)

가정보가 없이는 특정 개인을 알아볼 수 없도록 처리하는 것(개인정보보호법과 동일)을 말한다.

- 가명정보: 가명처리를 거쳐 생성된 정보로써 그 자체로는 특정 개인을 알아볼 수 없도록처리한 정보(개인정보보호법과 동일)를 말한다.
- 익명정보: 시간·비용·기술 등을 합리적으로 고려할 때 다른 정보를 사용하여도 더 이상 개인을 알아볼 수 없는 정보를 말한다.

위의 익명정보의 정의는 개인정보보호법 제58조의2(적용 제외)를 원용하고 있다. 그리고 가이드라인에서 익명처리에 대한 정의나 절차는 없으나, 부록에 가명·익명처리 기술 중 재현데이터를 소개하고 있다.

라. 금융분야 가명·익명처리 안내서[56]

- 가명처리: 추가정보(예: 가명정보와 기존 식별자를 연결하는 매핑테이블 등)를 사용하지 아니하고는 특정 개인인 신용정보주체를 알아볼 수 없도록 개인신용정보를 처리하는 것을 말하는데, 그 처리 결과가 ① 어떤 신용정보주체와 다른 신용정보주체가 구별되는 경우, ② 하나의 정보집합물에서나 서로 다른 둘 이상의 정보집합물 간에 어떤 신용정보주체에 관한 둘 이상의 정보가 연계되거나 연동되는 경우, ③ 위와 유사한 경우로써 대통령령으로 정한 경우의 어느 하나에 해당하는 경우로서 법령에 따라 그 추가정보를 분리하는 등 특정 개인인 신용정보주체를 알아볼 수 없도록 개인신용정보를 처리한 경우를 포함한다(신용정보법 제2조 제15호와 유사하나 동일하지 않음).
- 가명정보: 가명처리한 개인신용정보를 말한다(신용정보법 제2조 제16호와 동일).
- 익명처리: 데이터값 삭제, 가명처리, 총계처리, 범주화 등의 방법으로 개인신용정보의 전부 또는 일부를 삭제하거나 대체함으로써 더 이상 특정 개인인 신용정보주체를 알아볼 수 없도록 개인신용정보를 처리하는 것을 말한다(신용정보법 제2조 제17호와 유사하나 동일하지 않음).
- 익명정보: 개인신용정보를 더 이상 특정 개인인 신용정보주체를 알아볼 수 없도록 익명처리한 정보를 말한다.

56) 금융위원회·금융감독원(2022)

위 안내서는 가명정보 처리 가이드라인에서와 마찬가지로 부록에 가명·익명처리 기술 중 재현데이터를 소개하고 있다. 여기서 주목할 부분은 익명처리의 구체적인 절차로 제시된 '데이터값 삭제, 가명처리, 총계처리, 범주화 등의 방법'인데, 다음 절에서는 이 조항을 재현데이터와 관련하여 논의하겠다.

2. 시사점

위에서 기술한 바와 같이 현재 국내 관련법에 의하면 익명정보란 다른 정보를 사용하더라도 현실적으로 더 이상 개인을 알아볼 수 없는 정보로 정의하고, 이 같은 익명정보에 대해서는 개인정보 관련법이 적용되지 않는다. 해외의 유사한 관련 규제에서도 익명정보는 비슷하게 정의되고 있다.

이 정의에 따르면 잘 생성된 재현데이터는 익명정보로 분류되어 개인정보법과 무관하게 자유롭게 공유·배포가 가능해 보인다. 하지만 실무선에서 준수해야 하는 지침서인 『금융분야 가명·익명처리 안내서』에는 익명처리를 "데이터값 삭제, 가명처리, 총계처리, 범주화 등의 방법으로 개인신용정보의 전부 또는 일부를 삭제하거나 대체함으로써 더 이상 특정 개인인 신용정보주체를 알아볼 수 없도록 개인신용정보를 처리하는 것"으로 기술하고 있다. 이는 상위 규정인 신용정보법에서 정의한 익명처리의 개념을 매우 구체적인 수준으로 바꾼 것으로, 익명데이터의 자격요건을 위해 원데이터의 상당한 훼손을 요구하고 있다. 물론 '대체'라는 단어를 통해 재현데이터의 활용 가능성을 열어두긴 했지만, 전반적인 문구상 재현데이터가 익명데이터로서 인정받는 것이 아직은 현실적으로 매우 어렵다는 것을 알 수 있다.

이와 비슷하게 가명정보 처리 가이드라인의 2장에 제시된 아래 〈그림 V-1〉에서도 익명정보는 가명정보에 추가로 마스킹을 적용하여 개인 식별을 할 수 없게 처리된 정보로 명시하고 있다. 동일 가이드라인의 부록에 재현데이터가 익명처리의 한 기법으로 소개되고 있음에도 불구하고 본문에 포함된 〈그림 V-1〉은 익명데이터를 오직 극도로 높은 수준의 마스킹을 통해서만 얻을 수 있다는 잘못된 인식을 보여주고 있어 익명정보로서의 재현데이터 사용이 현실적으로 어려움을 재확인해 준다.

〈그림 V-1〉 가명정보와 익명정보의 비교

자료: 개인정보보호위원회(2022)

지금까지의 논의를 요약하면, 개인정보보호법과 신용정보법에서 정의된 익명정보의 개념이 관련 지침서로 만들어지는 과정에서 판단이 들어가 실제 의도한 법적 의미가 상당 부분 퇴색되었고, 이에 따라 재현데이터와 같이 생성형 모형을 통해 만들어진 모의 데이터들이 익명정보로 인정받기 어려운 상황이 된 것으로 파악할 수 있다.

이 같은 간극을 극복하는 방안으로 관련법에서 정한 익명데이터의 정의에 좀 더 충실한 객관적이고 포괄적인 지침을 만드는 것이 필요하다. 즉 극단적인 마스킹을 적용하지 않더라도 '시간·비용·기술 등을 합리적으로 고려할 때 다른 정보와 연동이 되어도 개인을 알아볼 수 없는' 데이터라면 익명데이터로 분류될 수 있도록 지침서를 개선해야 한다. 물론 다른 외부 정보의 범위와 종류는 매우 넓고도 다양해 모든 노출 가능성을 다 고려할 수는 없겠지만, 적어도 '합리적인 수준'에서 이러한 노출위험이 없음을 확인할 수 있는 절차를 만들고 이를 만족하는 경우 익명데이터로 분류하는 방향으로의 모색은 가능하다고 생각된다. 이와 관련해 EU의 데이터 보호 작업반 의견서[57]가 도움이 될 수 있다. 의견서에서는 가명과 익명데이터의 차이가 재식별의 가능성에 있다고 보고 구체적으로 익명데이터가 다음 세 종류의 재식별의 위험으로부터 자유로워야 한다고 기술하였다.

① 특정할 수 있는 재식별 위험(Singling out): 데이터 내에서 개인을 특정할 수 있는 위험

② 연결을 통한 재식별 위험(Linkability): 데이터 내에서 혹은 다른 데이터를 이용해 특정

57) Opinion 05/2014 on Anonymisation Techniques, Article 29 Data Protection Working Party(2014); https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

개인에 속한 정보들을 다시 연결할 수 있는 위험

③ 추론을 통한 재식별 위험(Inference): 속성들의 추론을 통해 개인을 특정할 수 있는 위험

이 재식별의 위험을 염두에 두고 본 보고서는 익명데이터로 인정할 수 있는 '합리적인 수준'에 대한 논의를 위해 다음의 방향을 제시한다.

첫째, 데이터의 보안과 보호를 구별하고, '보호'를 위한 제도적·실무적 지원을 확장할 필요가 있다. 데이터 보안(Data security)을 개인정보 유출을 방지하기 위한 일련의 규제와 장치라고 한다면, 이와 대비되게 데이터 보호(Data protection)는 노출되었을 경우 따르는 위험을 최소화하는 작업이다. 현재 시행되고 있는 가명·익명처리의 방법과 처리된 데이터에 대한 관리는 모두 데이터 보안에 해당한다. 국내 데이터 보안의 수준은 전반적으로 매우 높지만, 역설적이게도 강한 보안으로 인해 가명처리나 결합 이후 얻어진 데이터가 노출이 되었을 때 어떤 위험들에 대해 얼마나 취약한지 알기 어려운 실정이다. 이는 승인된 가명데이터에 대한 엄격한 관리체계가 있기 때문이기도 하지만, 더 근본적으로는 가명데이터를 특정 개인을 알아보기 위해, 즉 재식별을 위해 사용하는 것이 법적으로 금지되어 있기 때문이다.[58] 가명처리된 데이터가 가진 재식별 혹은 노출위험의 종류와 범위에 대한 측정과 연구는 그 자체로도 현재 시행되고 있는 개인정보보호 처리의 수준을 평가하는데 중요한 역할을 하며, 동시에 익명데이터에 대한 제도적인 근거를 만드는 데에 꼭 필요한 요소이다. 따라서 가명데이터의 재식별 위험에 대한 연구가 활성화될 수 있도록 법적인 장치를 개선하는 것이 시급하며, 제도적 지원과 함께 데이터의 '보호'에 관련한 연구나 검증을 수행할 수 있는 전문가집단을 구성하고 운영하는 것도 제안할 수 있다.

둘째, 재식별 위험의 측정과 관리를 위해 현재 공공기관과 사기업에서 개인으로부터 수집하고 있는 모든 데이터들과 변수들의 목록을 확보하여 데이터베이스화하고 변수별로 노출위험을 계량화하는 것을 고려할 수 있다. 개별 변수나 개별 관측값에 대해 노출위험의 범위와 수준을 수치화 혹은 라벨링할 수 있다면 가명·익명처리 시 데이터의 노출위험을 측정하는 데 도움이 될 뿐만 아니라 확보된 데이터베이스를 이용해 가능한 외부정보의 종류와 범위에 대한 실질적인 이해가 높아질 것이다. 이러한 경험이 축적된다면 이종데이터의 결합 시 변수들의 새로운 조합에 따른 노출위험의 증가 역시 계량화할 수 있을 것으로

58) 개인정보보호법 제28조의5(가명정보 처리 시 금지의무 등) ① 제28조의2 또는 제28조의3에 따라 가명정보를 처리하는 자는 특정 개인을 알아보기 위한 목적으로 가명정보를 처리해서는 아니 된다. ② 개인정보처리자는 제28조의2 또는 제28조의3에 따라 가명정보를 처리하는 과정에서 특정 개인을 알아볼 수 있는 정보가 생성된 경우에는 즉시 해당 정보의 처리를 중지하고, 지체 없이 회수·파기하여야 한다.

생각된다.

셋째, 위의 두 제안을 뒷받침할 수 있는 재현데이터의 품질 평가 방법론에 대한 관심과 지원이 필요하다. 재현데이터의 품질은 원데이터와 비교했을 때 효용과 노출위험을 다 포괄하는 개념이다. 현재 학계에서는 재현데이터의 생성기법과 함께 품질측정에 관한 이론적·실증적 연구가 활발하고 앞서 기술하였듯이 이미 활용할 수 있는 기술도 존재한다. 그러나 위에서 설명했듯이 아직 이를 적용하고 검증할 수 있는 여건이 조성되어 있지 않다. 이 부분이 해소되고 전문가들이 동의할 수 있는 품질의 척도와 기준을 정할 수 있다면 이를 이용해 익명데이터의 분류를 더 객관적이면서 관련법에 충실한 방식으로 할 수 있을 것이다. 참고로 재현데이터의 품질 평가는 원데이터의 품질 평가와도 연결되고 이는 시장에서 유통·거래되는 다양한 데이터들의 경제적 가치를 평가하는 과정에도 도움이 될 수 있을 것이다.

물론 위의 방안들이 구현된다고 하더라도 여전히 개인정보의 노출 가능성은 존재할 수 있고, 정성적인 측면에서의 노출위험에 대한 전문가의 평가도 필요할 것이다. 하지만 이러한 새로운 방향으로의 모색이 어렵다고 해서 현재의 방식을 고수하는 것은 바람직하지 않다. 현재 국내에서는 복잡한 절차를 거쳐 소수의 기업과 연구기관만이 가명데이터에 접근할 수 있고, 접근 또한 물리적인 제약으로 인해 상당한 시간과 비용의 손실이 발생하고 있다. 나아가 가명처리 과정에서 발생하는 정보 손실로 인해 애초 의도한 목적에 사용하기 어려운 저품질의 데이터를 양산하는 위험도 상존한다. 이미 미국과 유럽에서는 금융, 헬스케어, 이미지와 동영상 처리 등의 영역에서 재현데이터를 광범위하게 사용함으로써 해당 산업의 경쟁력을 확보해 가고 있다. 이를 감안할 때 사회 전 영역에 걸쳐 방대한 데이터가 수집되고 있는 우리나라에서 익명정보에 대한 포괄적 지침의 부재로 인한 데이터 산업의 정체는 매우 안타까운 점이다.

3. 결론

본 보고서는 개인정보보호를 위한 가명·익명처리 기법들을 검토하고 대안으로서 재현데이터를 제시하였다. Ⅱ장에서는 기존에 쓰이고 있던 기존의 가명처리 기법을 짚어보았다. 마스킹의 경우 가장 보편적으로 쓰이는 통계적 노출 제어 기법이지만, 정보를 통합하거나 변경함에 따른 데이터 손실이 발생하여 분석에 필요한 목적에 맞게 사용하기가 어려운 측면이 있다. 다음으로 차등적 정보보호를 소개하였는데 차등적 정보보호는 그 자체로 방법론이라기보다는 수학적 개념에 가깝지만 이 아이디어를 차용하여 다양한 통계적 노출 제어 기법들이 나타나기도 하고 다른 기법과 통합해 기존의 정보보호를 강화하는 방식으로 확장되어 왔다. 그러나 반복되는 쿼리에 대해 효율이 감소하고, 쿼리의 종류에 따라 적절한 매개변수 값이 달라진다는 점, 데이터의 효용을 측정할 수 없다는 한계점을 가진다.

Ⅲ장에서는 최근 각광을 받고 있는 가명·익명처리 기법인 재현데이터에 대해 논했다. 재현데이터는 원데이터의 통계적 특성은 유지하되 새롭게 생성된 모의 데이터이기 때문에 외부 데이터와 매칭이 불가해 높은 수준의 개인정보보호가 가능하다. 특히 개인정보보호 규제상 익명데이터로 분류되기 때문에 가명데이터와 달리 규제로부터 자유롭게 사용하고 공유할 수 있다. 관련 이론을 설명한 후 실제 자동차 보험 데이터를 이용해 재현데이터를 생성하고 재현 전후 데이터를 분포적인 측면과 지도학습모형의 적합도를 통해 비교하였다. 추가로 국내외 재현데이터의 이용사례들을 수집해 소개하였다.

Ⅳ장에서는 재현된 데이터에서 공통적으로 관측되는 효용감소와 정보보호의 증가라는 상충관계를 설명하고 이와 연결해 재현데이터의 품질을 측정하는 척도와 최근의 학문적 성과에 대해 논하였다.

Ⅴ장에서는 국내외 가명·익명정보에 대한 규제를 살펴보고 관련 시사점을 논하였다. 특히 국내의 개인정보보호법과 신용정보법에서 정하는 가명·익명정보가 정의와 차이를 살펴보고, 현재 이 규제를 설명하는 지침서인 가명정보 처리 가이드라인과 금융분야 가명·익명처리 안내서가 상위 법규들의 의도한 것과 다르게 익명정보를 기술하고 있어 재현데이터의 실제 사용에 제도적 걸림돌이 되고 있음을 지적하였다. 이를 해소하고자 관련 상위 규정에서 정의하는 가명·익명정보의 개념에 더 충실한 새로운 가이드라인을 만들기 위한 몇 가지 제도적 방안을 제안하였다.

이미 미국과 유럽에서는 재현데이터를 익명데이터로 보고 전 산업 영역에서 광범위하게

사용해 국가적 경쟁력을 확보하고 있다. 사회 전 영역에 걸쳐 방대한 데이터가 수집되고 있는 우리나라에서 익명정보에 대한 좀 더 포괄적이고 전향적인 지침이 마련된다면 4차 산업혁명시대의 석유라 불리는 빅데이터의 이용을 극대화하고 사회 전 분야의 데이터 산업 활성화에도 큰 도움이 될 것으로 기대한다.

참고문헌

개인정보보호위원회(2022),『가명정보 처리 가이드라인』

국세청(2021),『국세 데이터 활용 안내서』

금융위원회·금융감독원(2022),『금융분야 가명익명처리 안내서』

김승현(2020),「통계적 노출 제어 방법 및 활용사례 연구: 재현데이터 방법론을 중심으로」,『주택금융리서치』, 9호, pp. 18~35

김정연·박민정(2019),「다중대체와 재현자료 작성」,『응용통계연구』, 32(1), pp. 83~97

박민정·김항준(2016),「마이크로데이터 공표를 위한 통계적 노출제어 방법론 고찰」,『응용통계연구』, 29(6), pp. 1041~1059

박민정(2016),『통계적 노출제어 최신동향 검토: 차등정보보호를 중심으로』, 2016년 하반기 연구보고서, 제 I 권

_____(2020),『통계데이터센터 DB를 활용한 재현자료 생성 방법 연구』, 통계청 통계개발원 데이터 정보보호연구

유성준·박나리(2020),「CART 기법을 이용한 개인신용정보 재현자료 생성기법」,『통계연구』, 25(1), pp. 1~30

통계청(2021a),『통계적 노출제어 가이드라인』

_____(2021b),『통계의 창』, 여름호

한국신용정보원(2020),『진짜 같은 가짜! 재현데이터의 개념 및 활용 사례』, 이슈리포트, 2019-8

홍동숙(2022),『부도 예측을 위한 인공지능 학습용 데이터 생성 및 검증 기법: GAN (Generative Adversarial Network) 기반 재현데이터 중심으로』, 신용정보원 CIS 보고서

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L.(2016), "Deep learning with differential privacy", In Proceedings of the 2016 ACM SIGSAC conference on computer and communications security,

pp. 308~318

Calacci, D., Berke, A., and Larson, K.(2019). "The tradeoff between the utility and risk of location data and implications for public good", arXiv preprint arXiv:1905.09350

Che, Z., Cheng, Y., Zhai, S., Sun, Z., and Liu, Y.(2017), "Boosting deep learning risk prediction with generative adversarial networks for electronic health records", In 2017 IEEE International Conference on Data Mining(ICDM), IEEE, pp. 787~792

Chen, R. J., Lu, M. Y., Chen, T. Y., Williamson, D. F., and Mahmood, F.(2021), "Synthetic data in machine learning for medicine and healthcare", *Nature Biomedical Engineering*, 5(6), pp. 493~497

Chib, S.(1996), "Calculating posterior distributions and modal estimates in Markov mixture models", *Journal of Econometrics*, 75(1), pp. 79~97

Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W. F., and Sun, J.(2017), "Generating multi-label discrete patient records using generative adversarial networks", In Machine learning for healthcare conference, PMLR

Dandekar, A., Zen, R. A., and Bressan, S.(2018), "A comparative study of synthetic dataset generation techniques", In Database and Expert Systems Applications: 29th International Conference, DEXA 2018, Regensburg, Germany, September 3–6, 2018, Proceedings, Part II 29, Springer International Publishing, pp. 387~395

Dwork, C., McSherry, F., Nissim, K., and Smith, A.(2006), "Calibrating noise to sensitivity in private data analysis". In Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3, Springer Berlin Heidelberg, pp. 265~284

Geman, D., Geman, S., Hallonquist, N., and Younes, L.(2015). "Visual turing test for computer vision systems", Proceedings of the National Academy of Sciences, 112(12), pp. 3618~3623

Goodfellow, I., Bengio, Y., and Courville, A.(2016), *Deep learning,* MIT press

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y.(2020), "Generative adversarial networks", Communications of the ACM, 63(11), pp. 139~144

Gouweleeuw, J. M., Kooiman, P., and De Wolf, P. P.(1998), "Post randomisation for statistical disclosure control: Theory and implementation", *Journal of official Statistics*

Jeong, D., Kim, J. H., and Im, J.(2022), "A new global measure to simultaneously evaluate data utility and privacy risk", IEEE Transactions on Information Forensics and Security, 18, pp. 715~729

Ji, Z., Lipton, Z. C., and Elkan, C.(2014), "Differential privacy and machine learning: a survey and review", arXiv preprint arXiv:1412.7584

Li, N., Li, T., and Venkatasubramanian, S.(2007), "t-closeness: Privacy beyond k-anonymity and l-diversity", In 2007 IEEE 23rd international conference on data engineering, IEEE, pp. 106~115

Li, P., Stuart, E. A., and Allison, D. B.(2015), "Multiple imputation: a flexible tool for handling missing data", Jama, 314(18), pp. 1966~1967

Little, R. J.(1993), "Statistical analysis of masked data", *Journal of official statistics-stockholm-*, 9, pp. 407~407

Machanavajjhala, A., Kifer, D., Gehrke, J., and Venkitasubramaniam, M.(2007), "l-diversity: Privacy beyond k-anonymity", ACM Transactions on Knowledge Discovery from Data(TKDD), 1(1), 3-es

Nowok, B., Raab, G. M., and Dibben, C.(2016). "synthpop: Bespoke creation of synthetic data in R", *Journal of statistical software*, 74, pp. 1~26

Park, M. J., and Kim, H. J.(2016), "Statistical disclosure control for public microdata: present and future", *The Korean Journal of Applied Statistics*, 29(6), pp. 1041~1059

Rubin, D. B.(1993), "Statistical disclosure limitation", *Journal of official Statistics*, 9(2),

pp. 461~468

Samarati, P., and Sweeney, L.(1998), “Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression”

Snoke, J., Raab, G. M., Nowok, B., Dibben, C., and Slavkovic, A.(2018), “General and specific utility measures for synthetic data”, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181(3), pp. 663~688

Templ, M., Meindl, B., Kowarik, A., and Chen, S.(2013), “Introduction to statistical disclosure control(sdc)”, Project: Relative to the testing of SDC algorithms and provision of practical SDC, data analysis OG

Wei, J., Nie, Y., and Xie, W.(2020), “The Study of the Theoretical Size and Node Probability of the Loop Cutset in Bayesian Networks”, Mathematics, 8(7), p. 1079

Xu, Lei, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni(2019), “Modeling tabular data using conditional gan” Advances in Neural Information Processing Systems, 32

개인정보보호법(https://www.law.go.kr/)

금융위원회(https://www.fsc.go.kr/index)

대한민국 정책브리핑(https://www.korea.kr/special/policyCurationView.do?newsId=148867915)

KCB 데이터스토어(https://datastore.koreacb.com/support/utilizeCaseView8.do)

Gartner(https://www.gartner.com/en)

Google developers(https://developers.google.com/machine-learning/gan/gan_structure?hl=ko)

SDC practical guide(https://sdcpractice.readthedocs.io/en/latest/index.html)

https://gretel.ai

https://limoss.london/how-does-sdc-work

https://elise-deux.medium.com/new-list-of-synthetic-data-vendors-2022-f06dbe91784

https://www.kaggle.com/datasets/kondla/carinsurance

https://www.etnews.com/20220315000132

https://open-innovations.org/events/synae/

https://www.statice.ai/case-study/die-mobiliar

https://www.geeksforgeeks.org/cart-classification-and-regression-tree-in-machine-learning/

도서회원 가입안내

회원	연회비	제공자료	
법인 회원	₩300,000원	- 연구보고서 - 기타보고서 - 연속간행물 · 보험금융연구 · 보험동향 · 해외 보험동향 · KOREA INSURANCE INDUSTRY	영문 연차보고서 추가 제공
특별 회원	₩150,000원		
개인 회원	₩150,000원		

* 특별회원 가입대상 : 도서관 및 독서진흥법에 의하여 설립된 공공도서관 및 대학도서관

가입 문의

보험연구원 도서회원 담당
전화 : (02)3775-9113 | 팩스 : (02)3775-9102

회비 납입 방법

무통장입금
- 계좌번호 : 국민은행 (400401-01-125198) | 예금주: 보험연구원

자료 구입처

서울 : 보험연구원 자료실(02-3775-9113 | lsy@kiri.or.kr)

저자약력

김현태 워털루대학교 보험계리학 박사 / 연세대학교 응용통계학과 교수
E-mail : jhtkim@yonsei.ac.kr

장가영 연세대학교 통계데이터사이언스학과 석사 / 연세대학교 응용통계학과 연구원
E-mail : gayoung1876@yonsei.ac.kr

연구보고서 2023-07

데이터 가명·익명처리 기법의 현황과 대안: 재현데이터를 중심으로

발 행 일 2023년 6월
발 행 인 안 철 경
발 행 처 보험연구원
주 소 서울특별시 영등포구 국제금융로 6길 38 화재보험협회빌딩
인 쇄 고려씨엔피

ISBN 979-11-93021-11-8 (정가 10,000원)
979-11-85691-50-3(세트)