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.

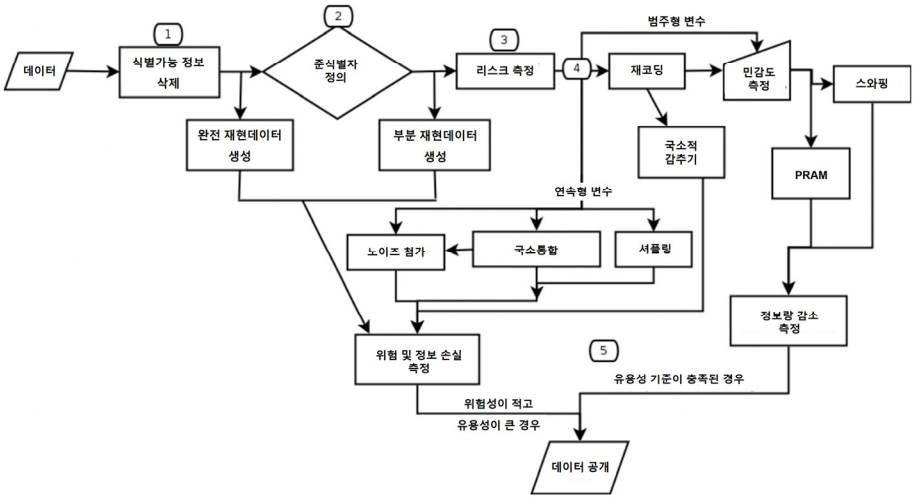
이와 관련하여 빅데이터의 활용 및 이종 데이터 결합 시 개인정보 노출 및 오·남용 방지를 위해서는 개인식별이 어렵도록 가공하는 가명·익명처리가 중요한 이슈로 대두되고 있다. 가공을 통해 개인정보 노출 정도를 조절하고 제한하는 기법을 총칭해 통계적 노출 제어[2](Statistical Disclosure Control)라 부르는데, 현재 통계적 노출 제어는 통계적인 기법과 딥러닝·AI 기법 등을 이용해 다양한 형태로 이루어지고 있다. 〈그림 Ⅰ-1〉은 통계적 노출 제어의 전체적인 과정을 간략히 나타낸 것이다. 이중 대중적으로 사용되고 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)