https://doi.org/10.5392/JKCA.2019.19.05.553

통계모형의 정확도에 기반한 비식별화 데이터의 품질 측정

Data Quality Measurement on a De-identified Data Set Based on Statistical Modeling

전희주*, 이현지**, 연규필***, 김동례****
동덕여자대학교*, 동국대학교**, 호서대학교***, ㈜이지서티****

Heuiju Chun(hjchun@dongduk.ac.kr)*, Hyun Jee Yi(dasbinich_2016@naver.com)**,
Kyupil Yeon(kpyeon1@hoseo.edu)***, Dongrae Kim(drkim@easycerti.com)****

요약

본 연구에서는 개인정보 비식별화 데이터의 통계적 유용성에 대한 품질 측정 방안에 대하여 통계 모형화에 따른 예측 정확도 측면에서 고찰하였다. 4차 산업혁명 시대에서 정보통신기술을 통한 혁신에는 반드시 빅데이터의 효과적인 활용이 필수적이지만, 개인정보 이슈는 적극적인 빅데이터 활용에 제약이 되고 있다. 이를 해결하기 위해 비식별화 가이드라인이 제정되었으며 다양한 개인정보 비식별화 방법이 활용되면서 개인정보의 실질적인 재식별 가능성은 매우 낮아졌다. 반면에 강력한 비식별화는 데이터의 유용성을 떨어뜨리는 부작용이 나타날 수 있다. 그 동안은 재식별 불가능한 비식별화 방법이 연구의 주를 이루어 왔다면 본 연구에서는 대표적인 비식별 방법인 KLT 모형에 의한 비식별화 데이터에 대한 통계적 유용성 측면의 품질 측정에 대하여 연구하였다. 비식별화 데이터에 대한 통계적 예측모형의 정확도에 기반하여 비식별화 된 데이터의 통계적 유용성이 어느 정도 훼손되는지에 대하여 사례분석을 수행하였다. 또한, 비식별 자료에 어느 정도의 비식별화 되지 않은 자료가 추가되어야 예측모형의 정확도를 회복하는 지를 살펴봄으로써 비식별화된 자료의 데이터 유용성 정도에 대한 새로운 측정지표를 제안하였다.

■ 중심어 : | 개인정보 | 데이터 품질 | 비식별화 | 예측모형 | KLT모형 |

Abstract

In this study, the method of quality measurement for the statistical usefulness of de-identified data was examined in terms of prediction accuracy by statistical modeling. In the era of the 4th industrial revolution, effective use of big data is essential to innovation through information and communication technology, but personal information issues are constrained to actively utilize big data. In order to solve this problem, de-identification guidelines have been established and the possibility of actual re-identification of personal information has become very low due to the utilization of various de-identification methods. On the other hand, strong de-identification can have side effects that degrade the usefulness of the data. We have studied the quality of statistical usefulness of the de-identified data by KLT model which is a representative de-identification method, A case study was conducted to see how statistical accuracy of prediction is degraded by de-identification. We also proposed a new measure of data usefulness of the de-identified data by quantifying how much data is added to the de-identified data to restore the accuracy of the predictive model.

■ keyword : | Personal Information | Data Quality | De-identification | Predictive Model | KLT-Model |

* 본 연구는 2015년 SW·컴퓨팅산업 원천기술개발 사업 연구과제로 수행되었습니다(연구과제번호 2015-0-00579).

접수일자 : 2019년 03월 04일
심사완료일 : 2019년 04월 18일
수정일자 : 2019년 04월 18일
교신저자 : 연규필, e-mail : kpyeon1@hoseo.edu

I. 서 론

1. 연구배경

1.1 개인정보 비식별화

빅데이터의 활용은 정보통신기술을 통한 산업 혁신에 필수적인 요소이다. 그러나 빅데이터의 적극적인 활용에는 개인정보 보호라는 걸림돌이 존재한다. 따라서 개인정보를 보호하면서 빅데이터의 유통 및 활용을 위해 개인정보 비식별화 기술이 부상하고 있다[1].

빅데이터 활용의 중요성을 이미 인식하고 있는 정·관계에서도 관계부처 합동으로 개인정보 비식별 조치 가이드라인을 제정하여 비식별 조치 기준 및 지원·관리체계에 대한 안내를 하고 있다[2]. 이 가이드라인에서는 정보주체를 알아볼 수 없도록 비식별 조치를 적정하게 한 비식별 정보는 개인정보가 아닌 것으로 추정하여 빅데이터 분석에 활용 가능하다고 보고 있다.

비식별화의 대상이 되는 개인정보는 개인 식별 정보로서 개인을 직접 식별하거나 유추하여 알 수 있는 모든 정보가 그 대상이 된다. 개인 식별 정보는 고유식별자, 준식별자 및 민감정보로 구분할 수 있다. 식별자는 개인 또는 개인과 관련한 사물에 고유하게 부여된 값이나 이름을 의미한다. 가령, 성명, 주민등록번호, 운전면허번호 등이 그 예이다. 준식별자는 해당 데이터만으로는 직접적으로 특정 개인을 식별할 수는 없지만 다른 정보와 결합하여 개인을 식별할 수 있는 정보로서, 성별, 우편번호, 혈액형, 학교명, 가족 정보 등이 준식별자의 대표적인 예이다. 민감정보는 개인정보보호법 제23조에서 '사상·신념, 노동조합·정당의 가입·탈퇴, 정치적 견해, 건강, 성생활 등에 관한 정보, 그 밖에 정보주체의 사생활을 현저히 침해할 우려가 있는 개인정보'로 규정하고 있다. 민감정보의 예로는 감염병병, 의료기록, 가족력, 장애등급, 소득 등과 같은 속성정보를 들 수 있다.

1.2 개인정보 비식별화 기술

개인정보 보호와 빅데이터의 활용은 서로 상반되는 목적을 가지고 있다. 정보의 비식별화 정도가 심할수록 비식별화된 데이터의 이용가치는 떨어지고, 데이터의 이용가치를 극대화하기 위해 낮은 수준의 비식별화를 한다면 개인정보 보호에 심각한 훼손을 초래할 수 있다. 즉 개인정보 보호는 개인정보 제공의 공급자 측면에서 다루어진 반면 개인정보 유통과 활용은 개인정보 사용자인 수요자 측면에서 서로 상응되는 측면에서 다루어진다[3]. 따라서 개인정보 수요와 공급측면에서 적정 수준의 개인정보 비식별화를 통해 개인정보의 재식별을 불가능하게 하면서 비식별화된 데이터의 유용성을 최대한 유지하는 것이 중요하다.

그 동안의 대부분의 연구는 식별정보 및 민감정보의 비식별화 및 재식별 방지에 초점을 맞추어 진행되어 왔다. 이러한 비식별화를 위한 전통적인 기법은 데이터 마스킹, 가명처리, 데이터 범주화, 데이터 값 삭제, 총계처리 등이 있다. 그러나 모든 속성정보에 대한 이러한 단순 익명화 방법의 적용은 개인정보 보호라는 목적은 달성할 수 있지만 데이터 활용성은 크게 떨어뜨릴 수 있다. 따라서 데이터 유용성을 적게 훼손시키면서 익명화 요구사항을 지킬 수 있는 프라이버시 모델이 사용되고 있다.

대표적인 프라이버시 모델은 KLT 모델로서, k-익명성(k-anonymity), ℓ-diversity(ℓ-다양성), t-근접성(t-closeness)을 만족하도록 데이터를 비식별화하는 방법이다. k-익명성은 공개된 데이터에 대한 연결공격을 방어하기 위해 제안된 프라이버시 보호 모델로서, 데이터 집합에서 특정 속성이 동일한 값을 가지는 레코드를 적어도 k이상으로 함으로써 다른 정보와의 결합이 쉽지 않게 하는 것이다. 미국 교육부 '프라이버시 보호 기술지원센터'의 안전도 기준에는 k=3을 안전도를 보장하는 최소한의 수준으로 보고 있으며 $5<k<10$ 일 경우 안전도가 높은 수준으로 간주하고 있다. 다만, 이는 절대적인 기준이 아니며 적절한 k 값은 해당 데이터에 대한 전문가 집단의 판단에 의해 정해진다. ℓ-다양성은 k-익명성의 취약점을 보완한 프라이버시 보호 모델로서, 주어진 데이터 집합에서 함께 비식별되는 레코드들은 적어도 ℓ개의 서로 다른 민감 정보를 가져야 한다. t-근접성은 ℓ-다양성 모델의 단점을 보완하기 위한 모델로서, 동질 집합과 전체 데이터 집합에서 특정 정보의 분포가 크게 차이가 나지 않도록 하는 것이다.

2. 연구목적

개인정보 비식별화는 필연적으로 원본 데이터의 왜

곡을 초래한다. 다만, 개인정보 보호를 위한 데이터의 왜곡이 항상 데이터 유용성을 떨어뜨리는 것은 아니다. 단순한 요약 통계치 만을 활용하고자 한다면 비식별화된 데이터만으로도 충분할 것이지만 통계 모델링을 통한 예측모형의 구축이 목적이라면 비식별화는 그 결과치에 더 많은 영향을 미칠 수 있다. 따라서 비식별화 데이터의 유용성은 데이터 활용의 목적에 따라 달라진다고 할 수 있다.

개인정보 비식별화에 대한 연구는 주로 제도적 측면에 대한 고찰이나 비식별화 기술적 방법론을 위주로 이루어져왔다[4-8]. 이러한 연구가 주를 이룬 이유는 데이터 유용성 보다는 개인정보 보호가 더 우선시되었기 때문으로 여겨진다.

비식별화 데이터의 품질을 주제로 한 해외 연구로는 [9-11] 등에서 원본 데이터와 비식별화 데이터 간의 상이도(dissimilarity)나 정보 손실율(information loss) 등의 측도를 품질 평가 지표로 사용하여 비식별화 데이터의 품질에 대하 연구를 수행하였다. 한편, 비식별화 데이터의 품질과 관련한 국내 연구로는 강동현 등[12]이 k-익명성과 마이크로어그리게이션(microaggregation) 프라이버시 모델로 비식별화된 데이터의 활용성을 판단하는 상이도 모델을 제시한 바 있다.

비식별화 데이터 유용성에 대한 기존 연구가 원본데이터와 비식별데이터 간의 상이한 정도를 측정하는 것이었다면, 본 연구에서는 비식별화 데이터의 유용성을 통계모형 구축을 통한 예측 정확도 측면에서 살펴보았다. 비식별 전후 데이터에 기반한 통계 예측모형을 비교 분석하여 비식별화 데이터의 데이터 품질을 평가하는 방법을 제시하고자 한다.

II. 비식별화 절차 및 방법론

1. 비실별화 절차

일반적인 비식별화 절차는 다음과 같다. 첫째, 데이터에 대한 이해 단계로서 개인정보 여부를 판단하고 식별자, 준식별자 및 민감정보를 구분함으로써 식별자 삭제, 민감정보의 중요도(또는 위험성)에 대한 판단, 비식별화 이후의 데이터 활용 방향 예측과 같은 작업이 진행된다. 비식별화로 인한 데이터의 손실량, 데이터의 활용성, 위험성에 대한 종합적인 고려가 요구되는 단계이다. 둘째, 비식별화 방법 결정단계이다. 이 단계에서는 KLT 모델이나 MAS(Multi-level abstraction and Synchronization) 모델같은 프라이버시 모델을 결정하고 안전도 수준(가령 KLT모델에서의 k,ℓ,t 값들)을 결정한다. 또한, 프라이버시 모델을 구현하는 기술적인 알고리즘을 선택하고 필요시 선택된 기술에 따른 추가 작업이 이루어진다. 셋째, 비식별화 수행단계로서 선택한 비식별화 알고리즘이 구현되는 프로그램을 수행하는 단계이다. 넷째, 비식별화 데이터에 대한 데이터 품질 평가 단계이다. 평가 결과에 따라 두 번째 단계로 이동하여 프라이버시 모델에서의 안전도 수준을 다시 세팅할 필요가 있다. 과도하게 변형된 데이터는 활용성이 떨어지므로 품질평가가 만족스럽지 못할 경우 두 번째 단계로 돌아가는 것이다. 데이터 품질평가 시에는 DM(discernibility metric), LM(loss metric), RCE(reconstruction error), Query error rate 등 다양한 데이터 품질 평가 측도가 사용될 수 있다[13]. 다섯째, 비식별화 적정성 평가 단계이다. 이 단계에서는 외부인이 포함된 전문가 평가단이 비식별 적정성을 평가한다. KLT 모델을 사용하는 경우 각 파라미터의 요구수준에 맞게 비식별되었는지를 평가한다. 평가결과가 부적정으로 나오면 두 번째 단계로 이동하여 비식별 조치를 재수행하게 된다. 전문가 평가단에 의한 평가는 정성적 평가와 정량적 평가로 구분할 수 있다. 정성적 평가로는 식별자 삭제 여부, 준식별자 및 민감정보 분류의 적정성, 사용한 비식별 기술/제품의 적정성 등을 판단한다. 정량적 평가로는 k,ℓ,t 값을 측정하여 데이터의 민감성 또는 위험성 대비 k,ℓ,t 값이 적절하게 반영되었는지 판단한다. 여섯째, 비식별화된 데이터의 제공 및 사후관리 단계로서 비식별화 적정성 및 데이터 품질 평가를 통과한 데이터를 제공하고 빅데이터 분석 등에 활용하며 사후관리를 하게 된다.

2. 비식별화 방법론

2.1 개인정보 비식별 조치 방법

정부 관계부처 합동으로 제정되어 2016년 7월부터 시행된 개인정보 비식별 조치 가이드라인에서는 비식

별 조치 방법론을 크게 다섯 가지 범주로 구분하여 제시하고 있다[2]. 즉, 개인정보 비식별화를 위한 일반적인 방법으로서 가명처리(Pseudonymization), 총계처리(Aggregation), 데이터 삭제(Data reduction), 데이터 범주화(Data suppression), 데이터 마스킹(Data masking)으로 구분하여 세부적으로 17가지의 처리 기술을 소개하고 있다.

가명처리는 개인식별이 가능한 데이터를 직접적으로 식별할 수 없는 다른 값으로 대체하는 기법으로서 실무적으로는 휴리스틱 가명화, 암호화, 교환방법 등을 사용한다. 총계처리는 데이터 전체 또는 부분을 집계(총합, 평균 등)하여 개별값을 요약값으로 대체하는 것으로서 실무적으로는 총계처리, 부분총계, 라운딩, 재배열 등의 방법을 적용한다. 데이터 삭제는 개인 식별정보의 전부 또는 일부를 삭제 처리하는 것으로서 실무적으로는 식별자 삭제, 식별자 부분삭제, 레코드 삭제, 식별요소 전부삭제 등의 방법을 적용한다. 데이터 범주화는 특정 정보를 해당 그룹의 대표값으로 변환(범주화)하거나 구간값으로 변환(범주화)하여 개인 식별을 방지하는 것으로서 실무적으로는 감추기, 랜덤라운딩, 범위 방법, 제어라운딩 등의 방법이 사용된다. 데이터 마스킹은 데이터의 전부 또는 일부분을 공백이나 노이즈 등으로 변환하는 것으로서 실무적으로는 임의 잡음 추가, 공백과 대체 등의 방법을 적용한다.

이러한 비식별 조치 방법론들은 상황에 따라 재식별 가능성이 존재하기 때문에 비식별화 기술에 대한 개별화 가능성, 연결 가능성, 추론 가능성 등의 위험성을 다시 검토하여야 한다[13]. 비식별화에 대한 적절성 평가에 주로 사용되는 계량적 방법은 KLT 모델이다. 또한, KLT 모델은 정부에서 마련한 개인정보 비식별조치 가이드라인에서 비식별화 적절성 평가에 대한 기준으로 채택하고 있는 계량적 방법이기도 하다. 따라서 다음 절에서 프라이버시 보호 모델인 KLT 모델을 중심으로 설명한다.

2.2 프라이버시 보호 모델

프라이버시 보호 모델은 비식별 데이터의 재식별 가능성을 검토하는 계량적인 방법으로 KLT 모델이나 MAS 모델 등이 있다. 여기서는 비식별조치 가이드라인에서 채택하고 있는 KLT 모델에 대하여 설명한다.

KLT 모델은 비식별 데이터가 만족해야 할 성질로서 k-익명성, ℓ-다양성, t-근접성을 요구한다. k-익명성은 비식별 데이터가 다른 데이터와 결합되었을 때 개인이 식별될 수 있는 위험성을 방지하기 위한 기준으로서, 주어진 데이터 집합에서 같은 속성값을 갖는 레코드가 적어도 k개 이상 존재할 것을 요구하는 기준이다. k-익명성을 보장하는 대표적인 비식별화 알고리즘으로는 LeFevre et al. [15]이 제안한 준식별자 계층격자 탐색 방법이 있다. k 값을 크게 설정할수록 재식별 가능성을 떨어뜨리지만 데이터의 활용성도 감소하게 된다. 따라서, 데이터의 활용목적을 고려하여 전문가의 판단하에 적정한 k 값을 설정한다. ℓ-다양성은 k-익명성이 갖는 단점, 즉 동질성 공격 및 배경지식에 의한 공격을 방어하기 위한 모델로서, 주어진 데이터 집합에서 구별되지 않는 레코드들의 모임인 동질 집합에서는 민감정보에 대하여 적어도 ℓ개 이상의 서로 다른 속성값을 가져야 할 것을 요구하는 기준이다[16]. ℓ-다양성을 만족하더라도 동질집합내에서 민감정보의 값이 특정값에 쏠려있다거나 유사한 범주의 값으로 이루어져있다면 프라이버시 노출 위험이 존재한다. 따라서 이러한 쏠림공격이나 유사성 공격에 대비될 수 있는 방법이 t-근접성 방법이 제안되었다[17]. t-근접성은 한 민감정보의 동질집합에서의 분포와 전체 데이터에서의 분포간의 차이가 t 이하일 때를 일컫는다. 분포간의 거리는 통상 EMD (Earth Mover's Distance)로 측정된다.

Ⅲ. 비식별화 데이터 품질 측정

1. 분석 데이터

1.1 기초 데이터

본 연구에 사용된 자료는 국민건강보험공단의 2009년 건강검진자 209,455명의 자료이다. 국민건강보험공단에서는 국민건강정보DB를 기반으로 표본코호트DB를 구축하고 있다. 표본코호트DB는 자격DB, 진료DB, 건강검진DB, 요양기관DB를 포함하고 있는데, 2009년 자격DB를 기준으로, 진료DB와 건강검진DB를 연결하여 분석을 위한 원본DB를 구축하였다. 다음 표는

2009년 건강검진자를 대상으로 향후 5년 동안 뇌졸중 발생유무에 대한 예측모형을 구축하기 위한 데이터 구성 변수들이다. 원본DB에는 혈압, 가족력, 콜레스테롤, 흡연유무, 요단백, 과거병력(고혈압, 심장병, 당뇨병, 고지혈증, 암) 등의 변수가 포함되어 있으나, 기초 분석 결과 뇌졸중 발병유무와의 연관성이 있는 7개의 설명변수만을 선택하여 분석에 사용하였다. 이 중에서 X_5와 X_7은 비식별화를 하지 않고 사용하였으며, 다른 설명변수들을 대상으로 비식별 전·후의 예측모형 결과에 미치는 영향을 살펴보았다.

표 1. 원본DB 구성 변수들

변수유형	변수명	내용	비식별화 대상여부
범주형	Y	뇌졸중 발병유무 (0:미발병, 1:발병)	비대상
범주형	X_1	성별(1:남성, 2:여성)	대상
범주형 (순서형)	X_2	연령그룹(5세단위)	대상
범주형 (명목형)	X_3	시도구분(서울,경기, ...)	대상
범주형 (순서형)	X_4	소득분위(0,1,2,..., 10)	대상
범주형 (명목형)	X_5	뇌졸중 과거병력 유무	비대상
연속형	X_6	(공복)혈당량	대상
연속형	X_7	체질량지수(BMI)	비대상
⋮	⋮		

1.2 비식별화 데이터

비식별 조치는 KLT 프라이버시 모델에서 $k=4$, $\ell=4$로 설정하였으며, ㈜이지서티의 비식별조치 솔루션인 Identity Shield 프로그램을 사용하여 수행하였다. 이 프로그램에서는 LeFevre et al.[15]이 제안한 준식별자 계층격자 탐색 방법(Incognito)으로 각 준식별자에 대하여 사전에 설정된 $k-$익명성을 만족하도록 일반화 방법을 사용하여 레코드들의 준식별자 값들을 변환한다.

성별(X_1) 변수는 원래 값이 1 또는 2이지만 비식별 조치로 약 5% 정도인 10,451개의 관측값(레코드)이 구간화되어 [1:2] 값을 갖게 되었다.

연령그룹(X_2) 변수는 5세 단위 그룹을 나타내는 정수값을 갖는 자료인데, 비식별 조치 후 전체 209,455 관측치 중 약 0.3%가 [0:8], [5:18], [16:18] 등으로 구간화 되어 비식별되었다.

시도구분(X_3) 변수는 시도지역을 나타내는 범주형변수로서 11(서울), 26(부산), ..., 49(제주) 값을 가지고 있는데, 비식별 조치 후 약 0.7% 관측치가 [11:49], [24,47], ... , [48:49] 등으로 구간화 되었다.

소득분위(X_4) 변수는 세대단위 보험료 부과에 따른 소득분위를 나타내는 범주형변수로서 0(0분위), 1(1분위), ..., 10(10분위) 값을 가지는데, 비식별 조치 후 약 0.7% 관측값이 [0:1], [0:10], ..., [6:7] 등으로 구간화 되었다.

연속형변수인 혈당량(X_6)은 민감정보로 간주하여 비식별 조치를 적용하였다. 이는 실제 민감정보라기 보다는 본 연구 목적상 연속형 자료의 비식별화가 예측모형 구축에 미치는 영향을 알아보기 위하여 설정한 것이다. 모든 관측치가 구간화되어 [0:295], [100:104], ..., [59:117] 등으로 비식별 처리 되었다.

구간화로 비식별 처리된 값들은 예측모형 구축에 사용되기 위해서 하나의 값으로 대체되어야 한다. 따라서, 명목형 변수의 비식별 처리된 값은 해당 구간의 범주 안에서 가장 빈도가 높은 그룹으로 처리하였고, 연속형 변수의 비식별 처리된 값은 해당 구간의 중앙값으로 대체하였다. 이런 과정을 거쳐 최종적으로 원본DB에 대응되는 비식별DB를 구축하였다.

2. 비식별화 데이터 품질 측정방안

2.1 예측모형

통계적 예측모형의 성능을 기반으로 비식별화 데이터의 유용성을 판단하기 위하여 전술한 변수들을 토대로 뇌졸중 유무(Y)를 예측하는 로지스틱회귀모형을 구축하였다. 로지스틱회귀모형은 다음 식과 같이 설명변수들의 선형결합으로 반응변수의 확률값을 예측하는 것으로 볼 수 있다.

$$\log\left(\frac{\Pr(Y=1|X_1,\cdots,X_p)}{1-\Pr(Y=1|X_1,\cdots,X_p)}\right)=\beta_0+\beta_1X_1+\cdots+\beta_pX_p$$

회귀모수들의 최우추정치(maximum likelihood estimates)를 $\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_p$라고 하면 설명변수 값들이

$x_1, x_2, \cdots, x_p$로 주어질 때 $Y=1$일 확률은

$$\hat{P}(Y=1|\boldsymbol{x}) = \frac{\exp(\hat{\beta}_0+\hat{\beta}_1 x_1+\cdots,\hat{\beta}_p x_p)}{1+\exp(\hat{\beta}_0+\hat{\beta}_1 x_1+\cdots,\hat{\beta}_p x_p)} \quad (1)$$

로 주어진다.

2.2 예측모형 평가

통계적 예측모형의 성능을 평가하는 측도는 여러 가지가 있다. 여기서는 분류 정확도(accuracy), 민감도(sensitivity), 특이도(specificity), 정밀도(precision)를 고려하기도 한다. 식 (1)에 의해 추정된 사후확률의 값을 0.5를 기준으로 하여 개체를 분류하였다. 즉 $\hat{P}(Y=1|\boldsymbol{x}) \geqq 0.5$ 이면 $Y=1$로 분류하고 그렇지 않으면 $Y=0$으로 분류하였다. 검증용 데이터에 대하여 실제 Y값과 예측된 Y값의 분류표가 다음과 같이 주어졌다고 하자.

분류표		예측	
		1	0
실제	1	a	b
	0	c	d

이때 모형의 성능을 나타내는 각 지표는 다음과 같이 구할 수 있다.

- 정확도(accuracy) = (a+d)/(a+b+c+d)
- 민감도(sensitivity)= a/(a+c)
- 특이도(specificity)= d/(b+d)
- 정밀도(precision) = a/(a+b)

2.3 비식별 데이터 품질 측정

원본DB의 209,455 관찰치(observation) 중에서 적어도 하나의 변수가 비식별 처리된 개체들을 DB1이라고 하자. 원본DB에서 DB1을 뺀 나머지 개체들(비식별 무관 개체들)을 7:3으로 랜덤하게 나누어 각각 DB2, DB_test 라고 하자. DB_test는 모형 평가를 위한 검증용 데이터(test data)로 사용한다.

비식별 데이터의 유용성 검증을 위해서 다음과 같이 두 가지 비교 방안을 제시한다.

첫째, 검증용 데이터를 제외한 전체 데이터를 훈련용 자료로 사용하되 비식별 처리된 DB1이 비식별 전의 참값을 가지고 훈련용 자료에 포함될 때와 DB1을 모형구축에 사용하지 않을 때의 통계적 예측모형의 성능을 비교함으로써 비식별 자료가 예측모형 구축에 얼마나 영향을 미치는지 가늠할 수 있을 것이다. 여기서 두 모형의 성능 비교는 검증용 자료에 기반하여 수행된다.

둘째, 비식별 처리된 DB1만을 훈련용 자료(training data)로 사용하는 방안이다. DB1의 비식별 전 참값을 이용하여 구축된 예측모형과 비식별 처리 후의 값(구간화 된 경우 최빈값 또는 중앙값으로 대체한 후)에 기반하여 구축된 예측모형을 검증용 자료에 대하여 평가지표를 구하여 비교한다. 또한 DB1에 비식별 무관 자료를 추가해 나가면서 테스트 자료에 대한 평가지표를 비교한다. 즉, DB2(비식별 무관 자료)에서 랜덤하게 DB1의 크기(레코드 수)와 동일한 양의 비식별 무관 자료를 추출하여 훈련용 자료에 더한다. 비식별 무관 자료를 2배, 3배, 4배 늘려가면서 훈련용 자료를 구성하고 각 단계마다 검증용 자료에 대한 모형평가 지표를 구하여 비교한다. 비식별 무관 자료가 추가될수록 비식별 처리된 자료(DB1)의 모형구축에의 영향력을 감소할 것이다. 검증용 자료에 대한 예측정확도 측면에서 어느 정도의 비식별 무관 자료가 훈련용 자료로 추가되어야 비식별 처리된 자료의 영향력이 소멸하는지를 살펴봄으로써 비식별 자료의 통계모형화 측면에서의 유용성 유지여부를 판단할 수 있을 것이다. 이러한 과정에서 통계모형으로는 전술한 로지스틱회귀모형을 사용하였다.

Ⅲ. 분석결과

1. 모형 적합 결과 비교

1.1 전체 자료의 유용성 검증

검증용 자료를 제외한 전체 자료를 사용하되 비식별 처리된 DB1이 비식별 전·후의 값을 가지고 훈련용 자료에 포함될 때의 두 로지스틱회귀모형을 비교한다. [표 2]는 DB1이 비식별 전과 후의 값으로 훈련용 자료에 포함됐을 때 추정된 회귀계수를 각각 나타내고 있다. 또한, [표 3]은 각 모형의 검증용 자료에 대한 분류

성능을 나타낸다.

비식별 처리된 DB1의 크기가 원본DB에 비해 매우 작아서 구축된 두 모형의 차이가 거의 없고, 검증용 자료에 대한 분류 성능도 매우 비슷함을 알 수 있다.

표 2. 비식별 전후의 추정된 회귀계수

변수	비식별전		비식별후	
	회귀계수	유의확률	회귀계수	유의확률
절편	-4.6294	0.0000	-5.1736	0.0000
성별	-0.0983	0.0433	-0.1288	0.0228
연령대	0.4300	0.0000	0.4802	0.0000
공복혈당	0.0007	0.3252	0.0049	0.0003
소득	0.0225	0.0063	0.0067	0.4873
시도_부산	-0.0993	0.3531	-0.0822	0.4886
시도_대구	-0.0409	0.7294	0.1133	0.4062
시도_인천	-0.0869	0.4729	-0.0349	0.8005
시도_광주	-0.0447	0.7548	0.1127	0.5346
시도_대전	-0.3615	0.0115	-0.1724	0.3300
시도_울산	-0.5664	0.0001	-0.3576	0.0510
시도_경기	0.1143	0.1356	0.1012	0.2193
시도_강원	0.1276	0.3601	0.4883	0.0062
시도_충북	-0.0164	0.9030	0.1632	0.3334
시도_충남	-0.1276	0.3275	0.1043	0.5162
시도_전북	-0.0106	0.9354	0.1651	0.3106
시도_전남	0.0229	0.8617	0.3085	0.0537
시도_경북	0.3089	0.0079	0.3747	0.0049
시도_경남	0.0430	0.6945	0.1664	0.1732
시도_제주	-0.0567	0.7817	0.3689	0.2713
과거병력_해당	2.2242	0.0000	2.2161	0.0000
과거병력_모름	-0.4240	0.0000	-0.4329	0.0000
가족력_해당	0.3092	0.0013	0.3040	0.0044
가족력_모름	0.2052	0.0011	0.2410	0.0012
BMI	0.0463	0.0000	0.0444	0.0000

표 3. 검증용 자료의 분류성능

모형	정확도	민감도	특이도	정밀도
비식별전	0.7581	0.7425	0.7737	0.7664
비식별후	0.7567	0.7690	0.7444	0.7505

1.2 비식별 처리된 자료의 유용성 검증

[표 4]는 전술한 두 번째 비교 방안에 의한 모형구축 결과로 도출된 두 모형의 분류 성능을 나타낸다. 1단계는 DB1만을 이용하여 비식별 전·후의 값을 이용하여 구축된 모형의 성능을 나타낸다. 2단계부터는 DB1의 크기와 동일한 양의 비식별 무관 자료를 훈련용 자료에 추가로 포함시킨 경우이다. 즉, 각 단계마다 DB1과 같은 크기의 비식별 무관 자료를 추가하면서 훈련용 자료에서 점점 비식별 처리된 자료의 비중이 적어지도록 구성한다.

[표 4]에서 1단계는 비식별화된 자료인 DB1만을 훈련용 자료로 사용한 것이고, 2단계부터는 DB1과 같은 크기의 비식별 무관 자료를 추가해가면서 비식별화된 자료와 비식별 무관 자료의 비율을 각각 1:1, 1:2, 1:3, 1:4로 하여 구성한 것이다.

표 4. 두 모형의 분류성능

단계	모형	정확도	민감도	특이도	정밀도
1단계	비식별전	0.7240	0.5517	0.8962	0.8417
	비식별후	0.7194	0.5462	0.8925	0.8356
2단계	비식별전	0.7570	0.6895	0.8246	0.7972
	비식별후	0.7566	0.6886	0.8246	0.7970
3단계	비식별전	0.7667	0.7293	0.8040	0.7882
	비식별후	0.7655	0.7296	0.8013	0.7859
4단계	비식별전	0.7704	0.7483	0.7924	0.7828
	비식별후	0.7678	0.7459	0.7897	0.7800
5단계	비식별전	0.7711	0.7544	0.7878	0.7805
	비식별후	0.7711	0.7538	0.7884	0.7808

DB1만을 훈련용 자료로 사용했을 때는 모든 평가 지표 측면에서 참값이 사용된 비식별전의 모형이 우수하다. 훈련용 자료에 비식별 무관 자료가 점점 추가되면서 비식별 전·후의 모형이 성능 면에서 차이가 줄어들면서 5단계가 되면 거의 차이가 없어진다. 이것은 비식별 처리된 자료의 4배 정도의 비식별 무관 자료가 추가되면 비식별화된 값들의 영향력이 거의 사라지는 것을 의미한다.

2. 분석결과 시사점

일부의 자료가 비식별화 처리되었을 때 전체 데이터의 유용성에 비식별 처리 전의 참값을 이용한 모델링 결과와 비식별 자료를 제외하고 모델링한 결과를 비교함으로써 판단할 수 있다. 두 모형의 비교에서 정확도(accuracy)를 사용하는 경우 유용성 측도로서

$$\frac{\text{비식별후 모형의 정확도}}{\text{비식별전 모형의 정확도}}$$

를 사용할 수 있을 것이다. [표 3]의 예에서는 0.7567/0.7581=99.8이 된다.

한편, 비식별 처리된 자료만의 유용성 측도로는 앞의

두 번째 비교방안에서 제시한 대로 비식별 처리된 자료의 영향력이 무시할 정도로 작아질 때까지 추가해야 할 비식별 무관자료의 크기로 정의할 수 있을 것이다. [표 4]의 예에서는 5단계에서 두 모형의 성능 지표가 거의 일치 했으므로 약 4배의 비식별 무관 자료가 추가되어야 비식별 처리된 값들의 영향력이 없어지는 것을 볼 수 있다. 물론 여기서 1단계에서의 비식별 전·후 모형 평가지표의 차이 크기로 비식별 처리가 얼마나 데이터 유용성을 훼손했는지 가늠해 볼 수 있을 것이다. 따라서 1단계에서의 정확도 차이와 그 차이를 상쇄하는데 필요한 비식별 무관 데이터의 필요량을 비식별 데이터에 대한 품질 지표로서 사용할 수 있을 것이다. 1단계에서의 정확도 차이가 클수록 비식별화 데이터의 품질은 더 낮아졌다고 볼 수 있고, 그 차이를 상쇄하는데 필요한 비식별 무관 데이터의 양이 많을수록 비식별화 데이터의 유용성도 떨어진다고 간주할 수 있다.

Ⅳ. 결 론

본 연구에서는 비식별화 데이터의 유용성에 대하여 예측 모형의 성능지표를 이용하여 검증하는 방안을 제시하였다. 비식별화된 자료를 포함하는 전체 데이터의 유용성은 비식별 자료를 제외했을 때의 모델링 결과를 토대로 판단할 수 있고, 비식별 처리된 자료만의 유용성 검증은 비식별 무관 자료를 훈련용 자료에 추가해가면서 모형 성능 측면에서 비식별 처리의 영향력이 미미해지는 단계를 파악함으로써 이루어질 수 있음을 사례분석으로 보여주었다.

본 연구는 예측모형 성능 비교 기반의 비식별화 데이터 품질 측정 방안을 제안하고 사례분석 결과를 제시했다는데 의미가 있다. 다만, 다양한 비식별화 방법론에 따른 데이터의 유용성 측정 비교연구는 다루지 못했다는 한계가 있으며 이는 향후 추가적인 연구 과제로 진행하고자 한다.

한편, 비식별화 되는 개체들의 속성 분포가 모집단의 분포와 비슷할수록 예측모형에의 영향력은 더 작아질 것으로 예상되는 바, t-근접성과 비식별 처리의 예측모형에의 영향력 간의 관계에 대하여 추후 연구과제로 삼고자 한다.

참 고 문 헌

[1] 양현철, 이영주, 김신곤, "개인정보 비식별화기술 적용 수준이 빅데이터 활성화에 미치는 영향," 정보화연구, 제13권, 제3호, pp.395-404, 2016.

[2] 국무조정실 등, *개인정보 비식별 조치 가이드라인*, 2016.

[3] 이영환, 전희주, 윤정연, "데이터 산업에서 창업 활성화를 위한 데이터 거래소 제안 : 금융거래소형 데이터 거래소를 중심으로," 한국창업학회지, 제10권, 제2호, pp.28-49, 2015.

[4] 김동국, 이혁, "빅데이터 기반의 개인정보 비식별화 동향," 한국인터넷정보학회지, 제16권, 제2호, pp.15-22, 2015.

[5] 이현승, 송지환, *개인정보 비식별화기술의 쟁점 연구*, 소프트웨어정책연구소, 2016.

[6] 임형진, "빅데이터 환경에서의 개인정보 비식별 처리 방법 분석," 전자금융과 금융보안, 제8호, pp.9-37, 금융보안원, 2017.

[7] 엄수현, 이인경, 이우기, "빅데이터 기반 개인정보 비식별화 동향," 정보화연구, 제15권, 제4호, pp.545-552, 2018.

[8] 김근령, 이대희, "보건의료 빅데이터 활용에 관한 법적 검토-개인정보보호를 중심으로-," 과학기술법연구, 제24권, 제3호, pp.57-90, 2018.

[9] D. Rebollo-Monedero, J. Forné, M. Soriano, and J. P. Allepuz, "k-Anonymous microaggregation with preservation of statistical dependence," Information Sciences, Vol.342, pp.1-23, 2016.

[10] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, and S. Martínez, "Enhancing Data Utility in Differential Privacy via Microaggregation- based k-Anonymity," The International Journal on Very Large Data Bases, Vol.23, No.5, pp.771-794, 2014.

[11] D. Sánchez, J. Domingo-Ferrer, S. Martínez, and J. Soria-Comas, "Utility-preserving differentially private data releases via individual ranking microaggregation," Information Fusion, Vol.30, pp.1-14, 2016.

[12] 강동현, 오현석, 용우석, 이원석, "비식별 데이터의 유사성 보존에 관한 연구," 한국정보처리학회 추계학술발표대회 논문집, 제24권, 제2호, pp.285-288, 2017.

[13] H. Lee, S. Kim, J. W. Kim, and Y. D. Chung, "Utility-preserving anonymization for health data publishing," BMC Medical informatics and Decision Making, Vol.17, No.1(104), 2017.
[14] 김동한, "개인정보 비식별화 기술 동향 및 전망," Weekly ICT Trend 주간기술동향, 제1809호, 정보통신기술진흥센터, pp.14-24, 2017.
[15] K. LeFevre, D. DeWitt, and R. Ramakrishnan, "Incognito: Efficient full-domain k-anonymity," *In Proceedings of the 2005 ACM SIGMOD international conference on Management of data (SIGMOD '05)*, pp.49-60, 2005.
[16] A. Machanavajjhala, J. Gehrke, and D. Kifer, "ℓ-Diversity: Privacy beyond k-anonymity," *22nd International Conference on Data Engineering, 2006.*
[17] N. Li, T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy beyond k-anonymity and l-diversity," *IEEE 23rd International Conference on Data Engineering*, 2007.

저 자 소 개

전 희 주(Heuiju Chun) 정회원

- 1989년 2월 : 고려대학교 통계학과(경제학사)
- 1998년 12월 : North Carolina 주립대학교 통계학과(통계학박사)
- 2012년 3월 ~ 현재 : 동덕여자대학교 정보통계학과 부교수

〈관심분야〉 : 빅데이터, 보험통계, 일반화선형모형

이 현 지(Hyun Jee Yi) 준회원

- 2018년 2월 : 동덕여자대학교 독일어과(문학사)
- 2018년 9월 ~ 현재 : 동국대학교 일반대학원 통계학과 석사과정

〈관심분야〉 : 데이터마이닝, 빅데이터

연 규 필(Kyupil Yeon) 정회원

- 1995년 2월 : 서울대학교 계산통계학과(이학사)
- 2006년 2월 : 서울대학교 통계학과(이학박사)
- 2014년 4월 ~ 현재 : 호서대학교 빅데이터경영공학부 부교수

〈관심분야〉 : 선형모형, 기계학습, 빅데이터

김 동 례(Dongrae Kim) 정회원

- 2000년 2월 : 한국외국어대학교 응용전산학과(이학석사)
- 2010년 8월 : 숭실대학교 일반대학원 컴퓨터공학(박사수료)
- 2002년 8월 ~ 현재 : ㈜이지서티 부사장

〈관심분야〉 : 개인정보보호기술, 비식별조치기술