로지스틱회귀 모델을 이용한 익명처리 데이터의 유용성 측정

성민경, 이진규, 정성룡, 김윤중*, 김동례*

한국정보통신기술협회, ㈜이지서티*

[mksung, jklee, srjeong]@tta.or.kr, [yjkim, drkim]@easycerti.com

Data Usefulness Measurement on an Anonymized Dataset Using Logistic Regression Model

Sung Min Kyoung, Lee Jin Kyu, Jeong Seong Ryong,

Kim Yun Joong*, Kim Dong Rye*

Telecommunications Technology Association, EASYCERTI Co., Ltd.*

요 약

데이터는 4차산업혁명의 원동력이 되는 중요한 자원이지만 개인정보 처리 문제로 인해 활용에 큰 제약을 받고 있다. 이를 해결하기 위하여 정부는 가명·익명처리를 통한 비식별 처리 데이터 활용을 법·제도 개편을 통해 가능하게 하고 있다. 본 논문에서는 대표적인 익명화 기법인 K-익명성 모형에 의한 익명처리 데이터의 유용성을 로지스틱회귀 모델의 예측 정확도 측정을 통해 알아본다. 또한, 예측에 사용되는 데이터를 변경하면서 측정을 수행하여 데이터 유용성 측정에 대한 모델을 제안한다.

Ⅰ. 서 론

빅데이터는 제4차 산업혁명의 원동력이라 할 수 있을 만큼 필수적인 요소이다. 빅데이터는 기존산업 발달, 신산업 도출, 새로운 지식 축적 등 다양한 분야에 활용될 수 있으나 개인정보 보호의 문제로 유통·활용이 매우 제한되고 있다. 이에 대한 대안으로 데이터 비식별 처리를 통해 데이터를 가명화 또는 익명화하여 활용하는 방안이 각국의 정부로부터 제시되고 있다. 비식별 데이터의 활용을 통해 개인정보 보호와 빅데이터 활용 이라는 두 가지 상충되는 목적을 동시에 달성할 수 있다.

데이터 비식별화는 개인을 구분/구별할 수 있는 정보의 일부 또는 전체를 삭제·수정하여 개인의 식별성을 낮추는 기술이다. 비식별 수준이 클수록 개인정보 보호는 강해지고 데이터 유용성은 낮아지며, 비식별 수준이 낮을수록 개인정보 보호는 약해지고 데이터 유용성은 높아진다. 즉, 데이터 활용 상황, 환경, 목적에 따라 비식별 수준이 결정되어야 하며, 유용성 측정은 해당 과정에서 필수적으로 수행되어야 하는 작업이다.

본 논문에서는 빅데이터의 안전한 활용을 위한 비식별 처리 데이터의 유용성을 로지스틱 회귀 모델의 예측 정확도를 통해 측정한다. 또한, 학습 및 추론(예측)에 활용되는 데이터의 속성(컬럼) 변경에 따른 분석을 통해 데이터 유용성 측정에 대한 모델을 제안한다.

Ⅱ. 배경지식

1. 기존연구

비식별 처리 데이터의 유용성 측정을 위한 기존 연구는 크게 두 가지 방식으로 나뉜다. 첫 번째 방식은 비식별된 데이터 자체의 변경 정도를 원본데이터와 비교하는 방법이다[1,2]. 이 방식은 공식화된 수식을 통해 정량적인 수치를 얻을 수 있지만 데이터를 활용하는 상황을 반영하지 못하는 단점이 있다. 두 번째 방식은 비식별 데이터를 로지스틱 회귀 등 학습/추론 모델에 적용하여 원본데이터의 예측 정확도와 비교하는 방식이다[3,4]. [3]에서는 로지스틱 회귀 모델을 이용한 통계모형 정확도에 기반한 비식별 데이터의 품질을 측정하였다. [4]는 본 연구의 선행 연구로 다양한 비식별화 수준과 모델에 따른 로지스틱회귀 모델 정확도를 측정하였다.

2. 데이터 비식별 처리 모델

데이터 비식별 처리는 가명처리와 익명처리로 나뉘며, 본 논문에서는 익명처리 모델을 활용하므로 이와 관련된 KLT 및 차분프라이버시 모델에 대해서 소개한다.

K-익명성은 데이터에 포함된 개인을 적어도 데이터 내의 다른 K-1명과 구분이 되지 않게 데이터를 수정·삭제 하는 모델이다[5]. K-익명성을 만족하는 데이터집합 내에서 개인 식별 확률은 1/K 이하로 보장된다.

L-다양성은 같은 데이터 값을 가진 그룹에 속한 개인들의 민감한 개인정보 속성의 값이 서로 다른 L개 이상 존재하게 데이터를 수정·삭제 하는 모델이다[6]. L-다양성 모델을 만족하는 데이터집합은 최소 L-익명성을 만족하여 일반적으로 K-익명성 모델에 비해 더 높은 개인정보 보호 수준을 가진다.

T-근접성은 같은 데이터 값을 가진 그룹에 속한 개인들의 민감한 개인정보 속성분포와 전체 데이터의 민감한 개인정보 속성 분포를 비교하여 두 분포의 차이가 T이하가 되게 데이터를 수정·삭제·삽입 하는 모델이다[7].

차분프라이버시는 개별 질의에 대한 응답에 잡음을 삽입하여 응답하며, 통계 데이터를 배포할 때는 전체 데이터에 일부 잡음을 삽입하여 배포함으로써 개인정보를 보호하는 모델이다[8]. 각각의 상황에서 데이터 사용자는 정해진 수학적 모델을 준수하는 수준만큼의 잡음이 포함된 결과를 받게 된다.

Ⅲ. 본 문

1. 실험목적

본 실험의 목적은 익명 처리된 데이터의 유용성을 로지스틱회귀 모델의 예측 정확도를 통해 측정하는 것이다. 원본데이터의 예측 정확도를 기반으로 하여 낮은 수준의 익명화로부터 높은 수준의 익명화 데이터의 예측 정확도 변화를 살펴본다. 또한, 전체 데이터 중 학습에 사용되는 컬럼을 변경시키면서 예측 정확도를 측정하여 그 결과를 살펴봄으로써 새로운 데이터 유용성 측정 모델을 제안한다.

2. 실험설계

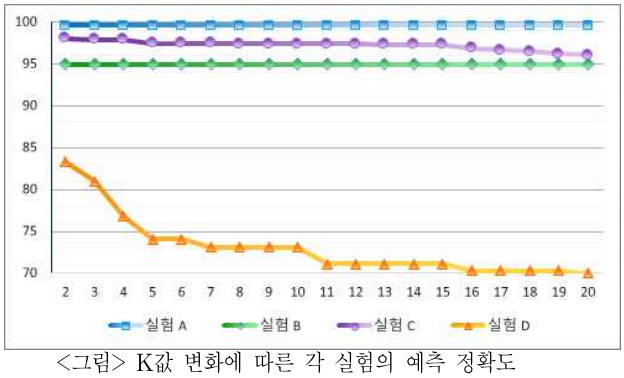
본 실험에서는 UCI Machine Learning의 Adult Dataset을 이용한다[9]. Adult Dataset은 나이, 교육정도, 성별, 결혼여부 등에 따른 급여의 상관관계를 나타낸 데이터이다. 데이터는 15개의 컬럼으로 구성되어 있으며, 'Salary'를 로지스틱회귀 모델의 종속변수로, 'Salary'를 제외한 14개 컬럼을 로지스틱회귀 모델의 독립변수로 설정한다. 14개 컬럼 중 9개 컬럼('Age', 'Workclass', 'Education', 'Relationship', 'Martial-status', 'Occupation', 'Race', Sex', 'Native-country')을 '식별가능정보'로 설정하여 비식별 처리 하였다. 종속변수 'Salary'컬럼은 '>50K'와 '<=50K' 두 가지 값을 가진다(원본데이터의 '>50K'와 '<=50K' 비율은 3:1). 익명화 모델은 K-익명성을 적용하였으며(K값은 2~20), ARX툴[10,11]을 이용하여 익명 데이터를 생성하였다.

텐서플로우 라이브러리를 통한 로지스틱회귀 모델을 이용하여 학습데이터와 테스트데이터를 8:2 비율로 나누어 'Salary'를 예측하였다. 비식별 처리된 데이터의 유용성을 측정하기 위해 다음과 같은 4가지 실험을 수행하였다.

실험 A	원본데이터 14개 컬럼 전체로 학습
실험 B	비식별대상 9개만 컬럼 원본 상태로 학습 (나머지 5개 컬럼 학습에 이용하지 않음)
실험 C	비식별 처리된 9개 컬럼+원본 5개 컬럼 학습
실험 D	비식별 처리된 9개 컬럼만 학습 (나머지 5개 컬럼 학습에 이용하지 않음)

3. 실험결과

<그림>은 K값 변화에 따른 실험 A ,B ,C ,D의 예측 정확도를 나타내고 있다. 실험 A와 B는 비식별 처리되지 않았으므로 K값 변화에 따른 예측 정확도 차이가 없다. 실험 C의 결과를 통해 2~20 사이의 K 값에서 95% 이상의 예측 정확도가 나왔다. 이는 비식별 대상 9개 컬럼만 원본 상태로 학습한 실험 B 보다 높은 수치로 비식별 처리 대상 컬럼이 있더라도 원본데이터가 보존된 컬럼이 일부 있다면 예측 정확도가 더 상승함을 알 수 있다. 또한 실험 D의 결과를 통해 비식별 처리된 데이터만으로 학습할 경우 예측 정확도가 매우 떨어짐을 알 수 있다. 이를 통해, 비식별 수준(여기서는 K값) 보다는 학습에 사용된 원본데이터 포함 여부가 더 큰 영향을 미침을 볼 수 있다.



<그림> K값 변화에 따른 각 실험의 예측 정확도

4. 유용성 측정모델 제안

본 실험 결과를 통해 로지스틱회귀 모델을 이용한 비식별 처리 데이터 유용성 측정모델을 제안한다. 우선 실험 A와 B를 통해 원본데이터 전체컬럼으로 학습한 모델과 비식별 처리 대상 컬럼들의 원본으로 학습한 모델의 예측 정확도를 측정하여 비교를 위한 기준점으로 삼는다. 그 후 실험 C를 통해 비식별 처리된 컬럼들과 비식별 처리되지 않은 컬럼들을 이용하여 예측 정확도를 측정하고, 실험 D를 통해 비식별 처리된 컬럼들만 학습에 이용하여 예측 정확도를 측정한다. 실험 A와 C의 비교를 통해 비식별 처리된 데이터의 유용성을 간접 측정할 수 있으며, 실험 B와 D의 비교를 통해 비식별 처리된 컬럼들의 유용성 정도를 측정할 수 있다.

Ⅳ. 결 론

본 연구에서는 로지스틱회귀 모델의 예측 정확도를 이용하여 비식별 처리 데이터의 유용성을 측정해 보았으며, 실험을 통한 유용성 측정 모델을 제안하였다. 학습에 사용되는 데이터가 변화함에 따른 예측 정확도 결과 분석을 바탕으로 유용성을 측정하는 한 방법을 보여주었다.

본 연구는 그동안 비식별 처리된 데이터 자체의 변경정도 측정에 치우쳤던 유용성 측정기법과는 다르게 데이터 활용 측면에서 유용성을 측정하는 방법을 사용하였다는 것에 의미가 있다. 다만, 다양한 타입의 데이터와 비식별 기법에 대한 내용을 다루지 못한 한계가 있다.

향후연구는 다음과 같다. 다양한 데이터 분포를 가진 데이터에 대한 결과 분석으로 데이터 분포에 따른 적절한 비식별 기법 적용방식을 선정하는 방법론을 연구해 볼 수 있다. 또한, 각 컬럼별 특성을 고려하여 각 컬럼에 대한 비식별 처리 수준을 결정하는 방법을 연구할 수 있다.

ACKNOWLEDGMENT

본 연구는 대용량 정형 데이터 대상 개인정보 가명/익명화를 위한 자동처리 기술 연구과제로 수행되었습니다.(2021-0-00634)

참 고 문 헌

[1] D. Robllo-Monedero, J. Forne, M. Soriano, and J. P. Allepuz, "k-Anonymous microaggregation with preservation of statistical dependence," Information Science, Vol.342, 2016.

[2] D. Sanchez, J. Domingo-Ferrer, S. Martinez, and J. Soria-Comas, "Utility-preserving differentially private data releases via individual ranking microaggregation," Information Fusion, Vol.30, 2016.

[3] 전희주, 이현지, 연규필, 김동례, "통계모형의 정확도에 기반한 비식별화 데이터의 품질 측정", 한국콘텐츠학회논문지 19(5), 2019.

[4] 성민경, 강지성, 유정승, "비식별 조치 데이터가 기계학습에 미치는 영향 분석", 한국정보과학회 학술발표논문집, 2019.

[5] L. Sweeney, k-anonymity: A model for protecting privacy, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, Vol. 10, No. 5, pp. 557-570, 2002

[6] A. Machanavajjhala, J. Gehrke, D. Kifer and M. Venkitasubramanian, ℓ-diversity: Privacy beyond k-anonymity, in Proceedings of the IEEE International Conference on Data Engineering, 2006

[7] N. Li, T. Li and S. Venkatasubramanian, t-Closeness: Privacy beyond k-anonymity and ℓ-diversity, in Proceedings of the IEEE International Conference on Data Engineering, 2007

[8] C. Dwork, Differential Privacy. In: van Tilborg H.C.A., Jajodia S. Encyclopedia of Cryptography and Security. Springer, Boston, MA, 2011.

[9] UCI Machine Learning Repository, www.archive.ics.uci.edu

[10] ARX Data Anonymization Tool, arx.deidentifier.org

[11] Fabian Prasser, Florian Kohlmayer, Ronald Lautenschlaeger, Klaus A. Kuhn., ARX - A comprehensive tool for anonymizing biomedical data., in Proceedings of the AMIA 2014 Annual Symposium, November 2014, Washington D.C., USA.