Journal of Korean Institute of Intelligent Systems
Vol. 30, No. 3, July 2020, pp. 228-235
http://dx.doi.org/10.5391/JKIIS.2020.30.3.228

ISSN(Print) 1976-9172
ISSN(Online) 2288-2324

JKIIS

오토인코더를 이용한 데이터 비식별화

Data De-identification using Autoencoder

김승환* · 전성해**†

* 인하대학교 IT융합기술연구소 연구교수, ** 청주대학교 빅데이터통계학과 교수
* Research professor, Division of Software Convergence, Inha University
** Professor, Department of Big Data and Statistics, Cheongju University

요 약

비식별 처리된 개인정보를 포함한 데이터를 제3자에게 제공하는 것이 본격적으로 가능하게 된 데이터법이 국회를 통과하면서 비식별 처리에 대한 중요성이 더욱 증가하고 있다. 비식별 처리는 기본적으로 데이터에서 특정 개인을 식별할 수 있는 가능성을 일정 수준 이하로 낮추는 방법인데 익명성, 다양성, 근접성 등에 기반한 모형을 사용한 마스킹과 범위 변환 방법이 널리 사용된다. 이 방법들은 이해가 쉬운 장점이 있으나 정보손실을 작게 유지하면서 데이터를 변환시키는 데에는 어려움이 있다. 마스킹, 범위화 이외에도 다양한 비식별화 방법에 대한 연구가 이루어지고 있다. 본 논문에서는 오토인코더 딥러닝을 사용하여 원자료에 대한 식별성을 낮추어 개인정보를 최대한 보호하면서 동시에 정보손실은 최소화할 수 있는 데이터 비식별화 방법을 제안한다. UCI 기계학습 데이터를 이용하여 제안 방법의 성능평가를 수행한다.

키워드 : 데이터 비식별화, 정보손실, 오토인코더, 개인정보, 데이터 잡음

Abstract

The importance of de-identification processing is increasing as the data law, which made it possible to provide data including de-identified personal information to third parties, passed the National Assembly. De-identification processing is basically a method of reducing the possibility of identifying a specific person in data below a certain level. Masking and range conversion methods based on anonymity, diversity, and proximity have been widely used. These methods have the advantage of being easy to understand, but have difficulty in converting data while keeping information loss small. In addition to masking and categorization, various methods of de-identification have been researched. In this paper, we propose a data de-identification method that can protect personal information as much as possible while minimizing the identity of the original data by using autoencoder deep learning. To verify the performance of the proposed method, we make experiments using UCI machine learning data.

Key Words : Data de-identification, information loss, autoencoder, privacy, data noise

Received: Apr. 07, 2020
Revised : Jun. 17, 2020
Accepted: Jun. 17, 2020
† Corresponding author
(shjun@cju.ac.kr)

1. 서 론

정부는 2016년 개인정보가 포함된 데이터라 할지라도 데이터 비식별화(de-identifi- cation) 처리를 하면 별도의 동의 없이 제 3자에게 데이터 제공할 수 있게 하는 개인정보 비식별 조치 가이드라인을 발표하였다. 이에 많은 기업들이 정부의 가이드라인에 따라 데이터를 타 기관에 제공하였는데 이는 적절한 데이터의 사용은 아니었다. 특히 비식별화 처리 과정에서 발생하는 데이터의 손실이 높아서 실제로 비식별화된 데이터의 활용에서 많은 문제가 발생하였다[1,2]. 최근 개인정보 관련 법 개정안이 2020년 1월 9일 국회 법제사법위원회를 통과하여 비식별화 된 데이터를 합법적으로 제 3자에게 제공할 수 있게 되었다. 따라서 데이터 비식별화에 대한 다양한 수요에 대처할 수 있는 연구가 필요하게 되었다. 비식별화 처리에서 주로 활용되는 기법은 데이터 마스킹(masking)과 범주화 기법이다[1,3,4]. 마스킹은 정보의 일부 혹은 전부를 가릴 수도 있고, 랜덤 노이즈(random noise)를 추가하여 원본 데이터를 훼손하는 방법 등을 포함한다. 일반적으로 자료

의 일부를 가리는 방법은 널리 활용되고 있지만 난수를 이용하여 노이즈를 추가하는 방법은 제대로 활용되지 않고 있다. 노이즈를 사용할 경우, 원 데이터에 노이즈를 추가하는 방법을 결정하는 것이 쉽지 않고 또한 추가된 노이즈에 의한 데이터의 정보손실이 발생하기 때문이다. 그러나 노이즈를 추가하는 방법은 데이터 형태를 유지할 수 있고 원 자료 역시 노이즈를 포함하고 있기 때문에 이 방법이 적절하게 수행될 경우 범주화 방법보다 더 우수한 결과를 기대할 수 있다. 최근에는 베이지안 및 딥러닝 기법을 이용한 데이터 비식별화에 대한 연구도 이루어지고 있다[5,6]. 구글에서는 RAPPOR (Randomized Aggregatable Privacy Preserving Ordinal Responses) 알고리즘을 이용한 노이즈 추가 방법에 대한 연구를 수행하였다[7]. 구글은 개인별 정보인 나무(tree)를 확인할 수 없게 하고 대신에 데이터 전체 정보인 숲(forest)을 허용하는 비식별화를 수행하기 위하여 확률화 응답(randomized response)에 의한 노이즈를 사용하였다. 본 논문의 2장에서는 관련 연구에 대한 소개를 하고 제안방법은 3장에서 다룬다. 제안 방법의 성능평가는 4장에서 수행하고 마지막으로 결론 및 향후 연구과제는 5장에서 나타낸다.

2. 관련 연구

2.1 확률화 응답모형

가장 먼저 고려할 수 있는 잡음추가 방법은 확률화 응답모형(randomized response model)을 이용하는 것이다. 확률화 응답모형은 표본조사에서 사용되는 방법으로 응답자가 민감한 질문에 응답을 회피하거나 거짓으로 응답하는 것을 방지하기 위하여 고안된 방법이다[8,9]. 예를 들어, 마약 경험 여부에 대한 질문에 대하여 응답자가 솔직한 응답을 해주지 않을 수 있기 때문에 아래와 같은 확률장치를 이용하여 응답을 받는다.

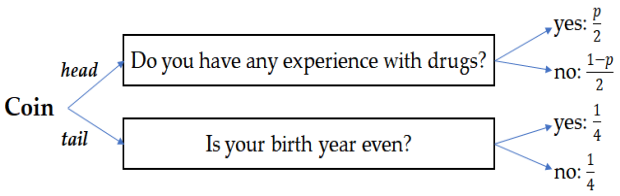


그림 1. 확률화 응답모형
Fig. 1. Randomized response model

응답한 결과가 "Yes" 일 때는 마약 경험이 있는 경우와 생일이 짝수인 경우 두 가지 이므로 민감한 질문에 대한 프라이버시를 만족하면서 응답을 구할 수 있다. 만약, 위에서 "Yes"라고 대답한 사람이 100명 중에서 28명이라고 하면 p는 다음과 같이 계산된다.

$$P(yes) = \frac{p}{2} + \frac{1}{4} = 0.28$$

따라서 p는 6%가 된다. 이 장치를 이용하여 개인정보에 잡음을 추가할 경우, 정보의 식별성이 낮아지기 때문에 정보가 노출되어도 그 정보가 어떤 것을 의미하는지 확신할 수 없다.

2.2 데이터 비식별화

일반적으로 가장 널리 사용되는 비식별처리 방법은 k-익명성(k-anonymity)을 보장하도록 처리하는 방법이다[10-15]. k-익명성은 특정 개인을 식별할 수 있는 가능성을 1/k 이하로 낮추기 위해 삭제, 범주화 등의 처리를 하는 방법으로 식별자(identifier)는 삭제하고 준식별자(quasi-identifier) 조합에 의한 식별성을 방지하기 위해 준식별자 조합이 같은 사람 수가 k명 이상 되도록 데이터 셋을 변형한다. 유일한 준식별자 조합에 대해 적어도 k명 이상의 레코드가 존재하면 해당 레코드가 누구의 정보인지 식별할 수 있는 확률은 1/k 보다 작아진다. 이 방법은 모집단 고유성(population uniqueness) 문제가 발생하기 때문에 이를 방지하기 위해서는 k를 크게 해야 한다[10]. 하지만 k가 커지면 정보손실이 증가한다. 준식별자에 대한 마스킹, 범주화 방법 이외에 준식별자에 난수를 더해 식별 가능성을 낮출 수 있다. 그림 2는 무작위화와 범주화를 설명하는 그림이다. 무작위화는 준식별자에 난수를 더하는 방법이고 일반화는 준식별자를 몇 개의 범위로 합치는 방법이다.

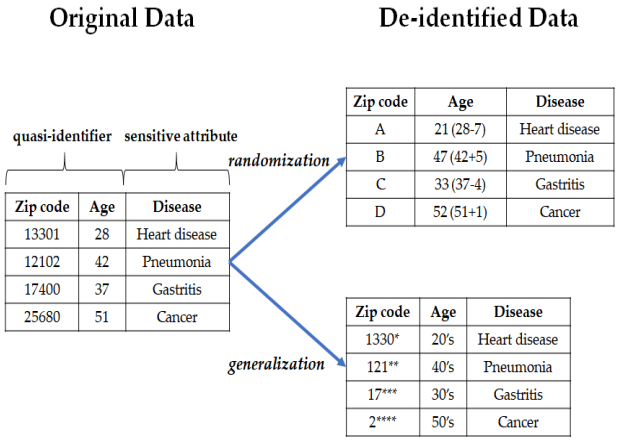


그림 2. 비식별화를 위한 임의화와 일반화
Fig. 2. Randomization and generalization for de-identification

무작위화는 데이터 형태를 그대로 유지하는 장점이 있지만 난수의 분산이 크면 원 자료의 통계적 특성을 훼손할 수 있기 때문에 적절한 분산을 더해야 한다. 그리고, 분산이 더해지면 통계적 설명력은 작아진다.

2.3 오토인코더

차원축소 기법인 오토인코더는 처음 발표된 이후 오랜 시간이 지났지만 최근 딥러닝의 사용과 함께 다시 주목받고 있다[16-19]. 인코더(encode)와 디코더(decode)로 이루어진 오토인코더는 입력과 출력이 같은 비지도

학습(unsupervised learning) 모형으로 이미지에서의 잡음 제거, 비정상 신호탐지 등에 활용되는 딥러닝(Deep Learning) 모형이다[20-22]. 그림 3처럼 좌측의 k개의 입력 뉴런이 p(p<k)개의 뉴런으로 인코드되고 다시 k개로 디코드 출력되는 과정을 거쳐 입력과 출력이 같은 값이 되도록 학습한다.

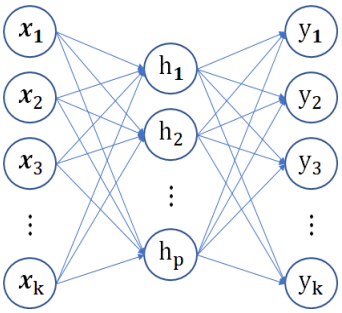


그림 3. 오토인코더 구조
Fig. 3. Autoencoder structure

하지만 인코드 과정에서 차원축소가 발생하기 때문에 정보손실이 발생한다. p가 클수록 정보손실이 작고, p가 작을수록 정보손실이 커진다. 학습은 주어진 모형에 대해 입력과 출력의 최소제곱오차(mean squared error, MSE) 또는 교차엔트로피(cross entropy)가 최소화되는 방향으로 진행된다. 오토인코더의 은닉층은 입력변수들의 비선형적 관계를 유지시키는 역할을 하고 은닉층 노드는 차원축소 크기를 결정하므로 노드수가 많으면 정보손실이 작아진다. 은닉층이 없고 층과 층 사이에 비선형 활성화 함수가 없으면 인코더는 통계학의 주성분분석(principal component analysis, PCA)이 된다. 즉, 오토인코더는 비선형 PCA 모형이다[20,21]. 오토인코더는 이미지에서 잡음을 제거하여 원본 이미지에 가까운 선명한 결과를 제공한다. 이는 이미지 픽셀을 오토인코더를 통과해 복원하면 차원축소로 인해 규칙성이 없는 노이즈는 없어지고 규칙성이 존재하는 숫자 모양만 복원되는 원리이다.

3. 오토인코더를 이용한 데이터 비식별화

개인정보로 구성된 k개의 변수를 오토인코더의 입력에 넣고 입력과 출력이 같게 학습하면 출력 결과는 입력과 유사하게 생성된다. 개인정보는 노이즈가 포함된 통계적 확률변수(Random Variable)이다. 오토인코더는 차원축소 과정을 거쳐 노이즈를 제거하고 최대한 입력된 정보의 통계적 성질은 보존하여 출력으로 전달한다. 이와 같은 오토인코더의 특성 때문에 출력에서 얻은 결과는 식별성은 감소하면서도 정보손실은 작게 된다. 예를 들어, 나이가 27살로 입력되었을 때, 출력에서는 28.2살로 출력되어 식별 가능성이 작아지고, 원자료의 잡음이 제거되어 변수들의 다차원 관계를 보존하는 방향으로 출력을 생성한다.

3.1 입력변수 준비

오토인코더의 입력은 숫자로 이루어지기 때문에 범주형 자료(categorical data)는 원핫 인코딩(one-hot encoding)을 해야 한다. 예를 들어, 성별은 Male→[1, 0], Female→[0, 1]로 변환한다. 수치형 자료는 다음 식과 같이 정규화(normalization) z_N과 표준화(standization) z_S 변환을 한다[23].

$$z_N = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

$$z_S = \frac{x - \bar{x}}{s_x} \tag{2}$$

3.2 오토인코더 모형 설계

입력변수가 범주형이면 원-핫 인코딩의 결과만큼 노드의 개수가 생성된다. 따라서 범주형과 수치형 변수를 모두 고려한 n개의 노드가 입력층을 생성한다. 그림 4는 본 논문에서 제안하는 데이터 비식별화를 위한 오토인코더의 구조를 나타낸다. 입력은 $(Z_1, Z_2, \ldots Z_n)$이고 출력은 $(Z_1', Z_2', \ldots, Z_n')$이다. 입력층과 출력층 사이에 은닉층이 있다.

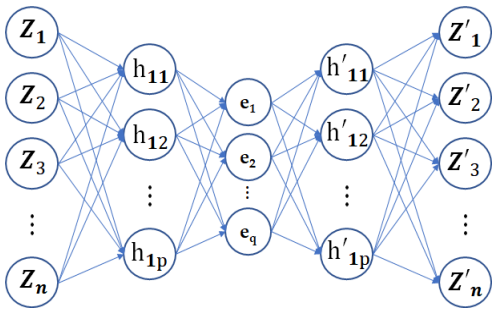


그림 4. 제안된 오토인코더 구조
Fig. 4. Proposed autoencoder structure

여기서 h와 e는 각각 은닉층과 인코더를 나타낸다. p와 q는 n보다 작은 값이다. 총 n개의 컬럼 변수를 입력과 출력으로 학습하는 그림 4의 오토인코더 모형에서 은닉층 수와 각 은닉층의 은닉노드 수 그리고, 인코더 층의 노드 수는 사용자가 결정한다. 은닉층과 은닉노드가 많을수록 정보손실이 작아 입력에 대한 출력의 식별성은 높아진다. 반대의 경우에는 식별성이 낮아지고 정보손실이 증가한다.

3.3 오토인코더 학습

표준화 및 원핫 인코딩된 입력 Z_i가 로 들어가서 식 (3)에 의해 은닉층(h)과 인코더(e)를 연산한다. 식 (3)에서 $S(\cdot)$는 활성함수(activation function)다. 일반적으로 활성함수로 시그모이드(sigmoid) 함수를 많이 사용하지만 비식별 문제에서는 선형함수도 사용 가능하다. 그림 4에서 제안 모형은 n차원 정보를 포함하는 (1,n) 차원

행렬을 p차원 정보를 포함하는 (1,p) 차원 행렬로 압축한다.

$$h_1 = S(z\,W_1 + b_1), e = S(h_1\,W_2 + b_2) \quad (3)$$

다음으로 e로 부터 식 (4)와 같이 $Z_i^{'}$을 구한다. 이때 $Z_i^{'}$은 다시 (1,n) 행렬이 된다.

$$h_1^{'} = S(e\,W_2^{'} + b'_2),\ Z_i^{'} = S(h_1^{'}\,W_1^{'} + b_1^{'}) \quad (4)$$

따라서 Z_i와 $Z_i^{'}$의 최소제곱오차(mean squared error, MSE)를 최소화하는 모수 W와 b값을 구한다. 충분한 학습에 의해 W와 b가 결정되면 개인정보가 포함된 입력변수 Z_i에 대하여 비식별화된 출력 $Z_i^{'}$을 얻는다.

3.4 최종 비식별 자료 생성

오토인코더 출력 $Z_i^{'}$는 잡음을 제거하는 특징을 가지고 있기 때문에 원 데이터에 비하여 비식별화된 데이터의 분산이 작아지는 문제를 가지고 있다. 즉, 원 데이터를 Z라고 하고 오토인코더 출력을 Z′이라고 하면 원 데이터 Z는 참값 T에 오차 e가 더해진 형태다. 식 (5)와 (6)처럼 오토인코더의 출력을 Z′라고 할 때, Z′은 오차가 제거되어 분산이 거의 0(zero)에 가깝게 된다.

$$Z = T + e, Z^{'} \approx T \quad (5)$$

$$V(Z) = V(e) = \sigma^2, V(Z^{'}) \approx 0 \quad (6)$$

은닉층의 수와 은닉 노드의 수가 많아지면 오토인코더가 과적합(over-fitting)되어 오차가 적게 제거된다. 충분히 압축되어 오차가 거의 제거되었을 때 오차가 제거된 오토인코더의 출력을 그대로 비식별 데이터로 사용할 경우 데이터의 분산이 작아지면서 원 데이터를 이용하여 분석한 것 보다 설명력이 좋아지는 문제가 발생한다. 이와 같은 문제를 해결하기 위해 Z′에 정규 난수γ를 추가하면 식 (7)과 같은 Z″을 얻을 수 있다.

$$Z^{''} = Z^{'} + \gamma, \gamma \sim N(0, V(e)) \quad (7)$$

이 때 Z″의 분산은 Z의 분산과 유사하므로 분산이 축소되는 문제가 해결된다. 최종적으로 원 데이터 Z에 대한 비식별 데이터인 Z″를 얻게 된다.

4. 실험 및 결과

제안 모형의 성능평가를 위하여 본 논문에서는 UCI 머신러닝 저장소로부터 붓꽃(Iris) 데이터와 미국 성인 소득 (adult salary) 데이터를 사용하였다[24]. Iris 데이터는 붓꽃의 외형을 결정하는 4개의 설명변수들(sepal length, sepal width, petal length, petal width)과 붓꽃의 3개 종(Setosa, Versicolor, Virginica)을 나타내는 1개의 반응변수(species)로 이루어진다. 그림 5는 4개의 설명변수에 대하여 반응변수의 종을 확인할 수 있는 산점도 행렬(plot matrix)를 보여 준다.

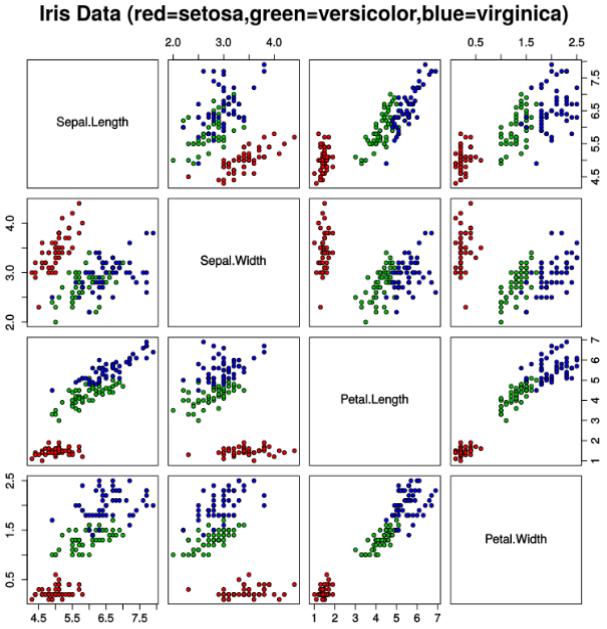


그림 5. Iris 데이터의 산점도 행렬
Fig. 5. Plot matrix of Iris Data

그림 5로부터 4개 설병변수들 사이에 상관관계가 있고 붓꽃의 종에 따라서도 차이가 있음을 확인할 수 있다. Iris 데이터는 꽃 개체에 대한 자료이지만 원 데이터에 대한 오토인코더 비식별 처리를 수행하야 제안 방법의 성능평가를 수행하였다. 모든 설명변수들은 식 (1)에 의해 표준화 되고 반응변수는 3개의 컬럼으로 원핫 인코딩하여 총 7개의 Z_i를 생성하였다. 오토인코더 모형은 두 개의 은닉층에 대하여 각각 4개와 3개의 은닉노드를 만들어 서로 대칭이 되도록 디자인하였고 총 1,000회의 역전파(back propagation) 알고리즘을 수행하여 (W, b)를 구하였다. 출력값 $Z_i^{'}$은 표준화 및 원핫 인코딩된 결과이므로 이를 다시 원래 스케일로 회복하는 변환을 통해 최종적으로 (150, 5)의 결과행렬을 구하였다. 학습은 구글 코랩 서버(Google Colab server)에서 5초 동안 진행되었다.

제안 방법에 의해 원 데이터와 통계적 상관관계는 그대로 유지되면서 원 데이터가 가지고 있는 오차 부분이 제거된 결과를 얻게 된다. 그림 6은 Sepal Length와 Sepal Width를 각각 X축과 Y축으로 지정하여 원 데이터와 오토인코더를 통과한 비식별화된 데이터 결과를 산점도를 통하여 나타내었다.

비식별 처리 후에도 원 데이터가 가지고 있는 다차원 상관구조는 그대로 보존되고 있지만 원 데이터의 잡음

이 제거되어 분산이 작아졌음을 알 수 있다. 이는 오토인코더가 차원을 축소하는 과정에서 확률변수가 가지고 있는 오차부분을 제거하기 때문이다. 이와 같이 오차가 과도하게 제거되는 문제 해결을 위해 오토인코더 출력에 평균이 0이고 분산이 σ^2인 서로 독립인 정규분포(normal distribution)를 따르는 난수를 생성하여 각 변수에 더하여 산포를 추가하여 얻은 결과는 그림 7과 같다. 난수의 분산은 i번째 변수자료와 오토인코더 출력 데이터 간 차이의 분산이다.

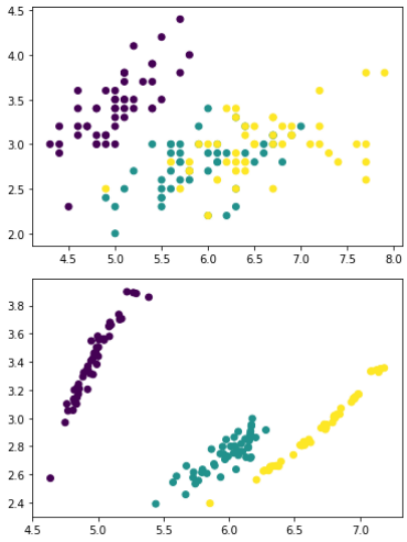


그림 6. 원본 데이터(위)와 비식별화된 데이터(아래)의 산점도

Fig. 6. Plots of Original (above) and de-identified (below) data

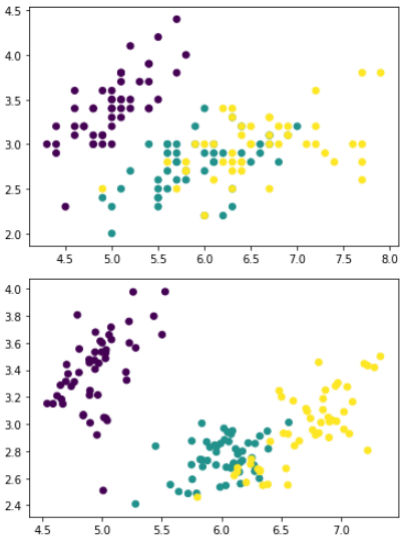


그림 7. 산포 추가 후 원본 데이터(위)와 비식별화된 데이터(아래)의 산점도

Fig. 7. Plots of Original (above) and de-identified (below) data after adding noises

결과적으로 원 데이터와 비슷한 산포 결과를 얻을 수 있게 되었다. 이와 같이 본 논문의 목표는 비식별화된 데이터의 식별 가능성은 낮아지고 동시에 원 데이터와 비슷한 통계적 성질을 그대로 유지하는 것이다. 표 1과 2는 원 데이터와 비식별화된 데이터를 이용하여 로지스틱 회귀모형(Logistic regression)을 통해 4개의 설명변수들이 반응변수를 어느 정도 예측할 수 있는지 분석하여 얻게 된 혼동행렬(confusion matrix)다. 혼동행렬의 결과를 통하여 원 데이터와 비식별 처리된 데이터의 결과가 서로 비슷하게 나타남을 알 수 있다.

표 1. 원본 데이터의 로지스틱 회귀분석 결과

Table 1. Logistic regression result of original data

Label		Estimation		
		Setosa	Versicolor	Virginica
Real	Setosa	50	0	0
	Versicolor	0	47	3
	Virginica	0	1	49

표 2. 비식별 데이터의 로지스틱 회귀분석 결과

Table 2. Logistic regression result of de-identified data

Label		Estimation		
		Setosa	Versicolor	Virginica
Real	Setosa	50	0	0
	Versicolor	0	48	2
	Virginica	0	1	49

본 논문의 추가적인 성능평가를 위하여 미국 성인 소득데이터를 사용하였다[24]. 반응변수는 미국성인 48,842명의 개인 급여액을 50K 이하와 50K 초과로 구분한 범주형 변수이고 설명변수는 나이(age), 직업구분(work class), 교육년수(education), 결혼상태(marital status), 인종(race), 성별(sex), 이자소득(capital gain), 부채(capital loss), 주당근무시간(hours per works), 출신국가(native country)로 총 11개의 변수를 사용하였다. 연속변수인 나이, 교육년수, 이자소득, 부채, 근무시간을 표준화 하고, 범주형 자료인 직업구분, 결혼상태, 인종, 성별, 출신국가, 소득을 가변수 처리하면서 총 컬럼은 11개에서 72개로 증가하였다.

오토인코더 모형은 72개의 컬럼 값을 갖는 입력층과 20개의 은닉노드로 이루어진 1개의 은닉층으로 구성되었다. 활성화 함수는 선형함수를 사용하였고, MSE를 최소화하는 알고리즘으로 학습률이 0.02인 Adam 최적화(optimizer) 학습을 수행하였다. 원 데이터 $Z_{48842,72}$를 이용하여 오토인코더 출력 데이터인 $Z'_{48842,72}$를 구하였다. 오토인코더의 분산축소 문제를 해결하기 위하여 두 데이터 행렬의 차이인 e 행렬에서 72개 열에 대한

분산을 구하고 식 (6)을 사용하여 $Z''_{48842,72}$ 행렬을 구하였다. 그림 8은 e에서 72개 열에 대한 표준편차의 히스토그램을 나타낸다. 표준편차는 최대 0.15 정도로 퍼져 있음을 알 수 있다. e의 표준편차가 작으면 정보손실은 작지만 식별성이 커지고, 반대로 크면 식별성은 작지만 정보 손실이 커지게 된다.

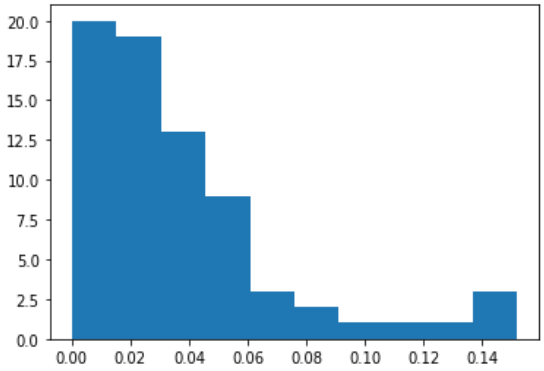


그림 8. 원 데이터와 비식별 데이터 간 차이의 표준편차
Fig. 8. Standard deviation of difference between original and de-identified data

Z″를 다시 원 데이터 스케일로 역표준화하고 원핫 인코딩된결과를 역변환하면 (48842, 11)개의 비식별화된 결과 데이터를 얻게 된다. 표 3에서 연속자료인 age, edunum 등이 다른 값으로 변화되었음을 확인할 수 있다.

표 3. 비식별 결과
Table 3. De-identified result

Variable	Original data	De-identified data
age	39	38.16498947
workclass	State-gov	State-gov
edunum	13	13.0773325
marital_status	Never-married	Never-married
race	White	White
sex	Male	Male
capital_gain	2174	0
capital_loss	0	181.5397491
hours_per_works	40	33.55421066
native_country	United-States	United-States
salary_class	<=50K	<=50K

범주형 데이터는 데이터 형태의 특성에 의해 변형이 심하지는 않지만 표 4에서 나타난 결과처럼 변할 수 있다.

비식별화된 결과 데이터가 유용성을 가지려면 원 데이터와 통계적 분석 결과가 유사해야 한다. 이를 위해 본 연구에서는 Salary class 변수에 대한 의사결정나무 모형 분석 결과를 표 5와 6에서 나타낸다.

표 4. 범주형 자료의 변형
Table 4. Taansformation of categorical data

원본범주 \ 비식별범주	Female	Male
Female	16181	11
Male	0	32650

표 5. 원 데이터의 Salary-class 추정 정확도
Table 5. Estimation accuracy of Salary-class: Original data

Label		Estimation	
		<= 50K	> 50K
Real	<= 50K	11598	835
	> 50K	1563	2283

표 6. 비식별 데이터의 Salary-class 추정 정확도
Table 6. Estimation accuracy of Salary-class: de-identified data

Label		Estimation	
		<= 50K	> 50K
Real	<= 50K	11319	1114
	> 50K	1717	2129

원 데이터에 대한 추정 정확도는 85%이고 비식별화된 데이터도 83%로 두 데이터가 비슷한 결과를 나타내고 있다. 2% 정도의 차이를 보이는 것은 오토인코더가 노이즈를 완전하게 제거하지 못하기 때문에 오토인코더의 잔여 분산과 정규 난수를 추가한 분산이 더해져 실제 분산보다 약간 더 커졌기 때문이다. 정규 난수의 분산을 약간 작게 하면 원자료와 비슷한 분산을 가지는 비식별화된 데이터도 만들 수 있다.

5. 결론 및 향후 연구

본 연구는 비식별 처리 방법의 하나로서 잡음추가 모형을 제안하였다. 일반적으로 수집된 데이터에는 이미 잡음이 섞여 있는데 여기에 비식별 처리를 위하여 추가적으로 잡음을 섞는 것은 비식별화된 데이터의 분석 측면에서 어려움이 있다. 본 논문에서는 이와 같은 문제점을 해결하기 위해 오토인코더를 사용하여 통계적 특성을 보존하면서도 비식별 처리효과를 얻을 수 있는 연구를 수행하였다. 제안 방법에 의한 실험결과에서 오토인코더가 비식별 처리에 활용될 수 있다는 가능성을 확인할 수 있었다. 최근에는 원자료의 통계적 분포를 추정하고 이를 통해 새로운 자료를 생성하는 재현자료(synthetic data)에 대한 연구가 진행되고 있다. 재현자료 생성기술로 베이지안적 접근을 사용할 수 있지만, 변이형 오토인코더(variational autoencoder), 적대적생성망(Generative Adversarial Networks) 등의 딥러닝 생

성모형을 통해 원자료의 분포를 학습하고 가짜 정보를 난수로 생성하는 것도 가능하다. 비식별 정보는 법적으로 개인정보로 취급되지만 재현자료는 통계적 분포로부터 만들어진 정보이기 때문에 개인정보가 아니다. 원자료의 분포를 정확하게 재현하는 재현자료 생성이 가능하게 된다면 데이터를 활용하고자 하는 많은 분야에서 법적 제약 없이 데이터 분석이 가능하게 된다. 이에 대한 연구는 우리의 향후 연구 과제로 남긴다.

References

[1] S. W. Kim, S. Jun, "Big Data Integration using Data De-identification," Journal of The Korean Institute of Intelligent Systems, Vol. 29, No. 3, pp. 235-241, 2019.

[2] S. Jun, "A Big Data Preprocessing using Statistical Text Mining," Journal of The Korean Institute of Intelligent Systems, Vol. 25, No. 5, pp. 470-476, 2015.

[3] S. Garfinkel, "De-Identification of Personal Information," NISTIR, 8053, 2015.

[4] H. R. Kim, "De-identification and Privacy protection for Statistical Purpose," Journal of The Korean Official Statistics, Special Issue, pp. 35-51, 2016.

[5] F. Dernoncourt, J. Y. Lee, O. Uzuner, P. Szolovits, "De-identification of patient notes with recurrent neural networks," Journal of the American Medical Informatics Association, Vol. 24, No. 3, pp. 596-606, 2017.

[6] P. Nousi, S. Papadopoulos, A. Tefas, L. Pitas, "Deep autoencoders for attribute preserving face de-identification," Signal Processing: Image Communication, Vol. 81, pp. 115699, 2020.

[7] Erlingsson, U., Pihur, V. and Korolova, A. (2014). RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response, Proceedings of the 21st ACM Conference on Computer and Communications Security, ACM, Scottsdale, Arizona.

[8] S. L. Lohr, (2019) Sampling: Design and Analysis, second edition, Boca Raton, FL, CRC Press, 2019.

[9] S. K. Thompson, Sampling 3rd Edition, Hoboken, NJ, John Wiley & Sons, 2012.

[10] 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.

[11] M. J. Park, H. J. Kim, "Statistical disclosure control for public microdata:present and future," The Korean Journal of Applied Statistics, Vol. 29, No. 6, pp. 1041-1059, 2016.

[12] Y. N. Shin, M. G. Chun, "Personal Information Protection for Biometric Verification based TeleHealth Services," Journal of The Korean Institute of Intelligent Systems, Vol. 20, No. 5, pp. 659-664, 2010.

[13] C. Caballero-Gil, J. Molina-Gil, J. Hernández-Serrano, O. León, M. Soriano-Ibañez, "Providing k-anonymity and revocation in ubiquitous VANETs," Ad Hoc Networks, Vol. 36, Part 2, pp. 482-494, 2016.

[14] A. Gionis, T. Tassa, "k-Anonymization with Minimal Loss of Information," IEEE Transactions on Knowledge and Data Engineering, Vol. 21, Iss. 2, pp. 206-219, 2009.

[15] N. Man, X. Li, K. Wang, "A Privacy Protection Model Based On K-Anonymity," Advances in Engineering Research, Vol. 153, pp. 15-19, 2018.

[16] D. H. Ballard, "Modular learning in neural networks," Proceedings of the sixth National conference on Artificial intelligence, Vol. 1, pp. 279-284, 1987.

[17] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P. Manzagol, "Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion," Journal of Machine Learning Research 11, 3371-3408, 2010.

[18] P. Baldi, K. Hornik, "Neural Networks and Principal Component Analysis: Learning from Examples Without Local Minima," Neural Networks, Vol. 2, pp. 53-58, 1989.

[19] D. E. Rumelhart, G. E. Hinton, R. J. Williams, "Learning representations by back-propagating errors," Nature, Vol. 323, pp. 533-536, 1986.

[20] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, Cambridge, MA, MIT Press, 2016.

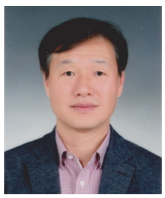
[21] R. Vidal, Y. Ma, S. Sastry, Generalized Principal Component Analysis, London, Springer, 2016.

[22] S. Theodoridis, Machine Learning A Bayesian and Optimization Perspective, London UK, Elsevier, 2015.

[23] K. P. Murphy, Machine Learning: A Probabilistic Perspective, Cambridge MA, MIT Press, 2012.

[24] UCI ML Repository, UC Irvine Machine Learning Repository, Available: http://archive.ics.uci.edu/ml, 2019, [Accessed: October 15, 2019]

저 자 소 개



김승환 (Seungwhoun Kim)

1989년: 충북대학교 토목공학 공학사
1991년: 인하대학교 통계학과 이학석사
1997년: 인하대학교 통계학과 이학박사
2014년: SK 에너지, SK M&C 재직
2015년 ~ 현재: 인하대학교 소프트웨어융합공학 연계전공 연구교수

관심분야 : Big Data, Statistical Algorithm, Machine Learning, Software Convergence
ORCID ID : 0000-0001-9303-2068
Phone : +82-32-860-8423
E-mail : swkim4610@inha.ac.kr

전성해 (Sunghae Jun)

1993년: 인하대학교 통계학과 이학사
1996년: 인하대학교 통계학과 이학석사
2001년: 인하대학교 통계학과 이학박사
2007년: 서강대학교 컴퓨터공학과 공학박사
2013년: 고려대학교 정보경영공학과 공학박사
2003년 ~ 현재: 청주대학교 소프트웨어융합학부 빅데이터통계학전공 교수

관심분야 : Artificial Intelligence, Statistical Learning, Big Data
ORCID ID : 0000-0003-1961-0055
Phone : +82-43-229-8205
E-mail : shjun@cju.ac.kr