물류시스템에서 비정형 데이터를 활용한 비식별화 인공지능 모델 설계

(Design of de-identified artificial intelligence model using atypical data in logistics system)

정 철 우, 황 상 호, 김 성 호*
(재)경북IT융합산업기술원 연구개발부
(Cheol-Woo Jung, Sang-Ho Hwang, and Sungho Kim)
(Research Development Division, Gyeongbuk Institute of IT Convergence Industry Technology (GITC))

Abstract : In order to legally provide de-identified data, deprivation of personal information is necessary. However, conventional de-identification methods have many problems in data utilization. In this paper, we propose a de-identification method for personal information using the SeqGAN model. By using the proposed design model to change the personal information of the existing logistics data and the generated personal information, it becomes possible to utilize the pseudonymized personal information data in the logistics data.

Keywords : De-Identification, SeqGAN, Machine Learning, Bigdata

Ⅰ. 서 론

물류시스템은 기본적으로 방대한 데이터가 존재한다. 빅데이터 발전은 물류시스템에서 방대한 양의 데이터를 저비용·초고속 처리함으로써 더 많은 정보의 활용과 연계를 가능하게 하여 다목적 물류정보시스템(MPS), 온습도 관제시스템, 고품질 택배 시스템 활용 등과 같이 물류산업에서 혁신적 변화가 이루어지고 있다. 이와 같이 물류에서 빅데이터의 활용은 장래영향력이 매우 높은 기술로 평가되고 있으며, 국내 물류기업들은 광범위한 데이터 수집을 진행하고 있으며, 이를 활용한다면 물류산업에서 성장가능성은 매우 높을 것으로 기대된다.[1]

2016년 정부는 개인정보가 포함된 데이터에서 데이터 비식별화 처리를 하면 개인의 동의 없이 데이터를 제공할 수 있게 하는 '개인정보 비식별 조치 가이드라인'을 발표하였다. 데이터를 가지고 있는 많은 기업들은 정부의 가이드라인을 따라 데이터를 제공하였는데 이에 따르는 문제점이 많이 있었다. 특히 비식별화 처리 과정에서 발생하는 데이터의 손실이 높아서 실제로 비식별화된 데이터의 활용에서 많은 문제가 발생하였다. 최근 개인정보 관련 법 개정안이 2020년 법안을 통과하여 비식별화 된 데이터를 합법적으로 제 3자에게 제공할 수 있게 되었다. 따라서 데이터 비식별화에 대한 다양한 연구가 필요하게 되었다. 특히 물류시스템은 개인정보가 많이 포함된 데이터 중 하나이다. 이름, 주민등록번호, 휴대폰번호, 주소, 계좌번호 등과 같이 공개되었을 경우 개인에게 치명적인 문제를 일으킬 수 있는 정보들이 많이 있다. 따라서 물류시스템에서는 개인정보에 대한 비식별화 방안은 필수요소로 작용한다.

* 교신저자(Corresponding Author)
정철우, 황상호, 김성호 : (재)경북IT융합산업기술원
※ 본 연구는 중소벤처기업부의 규제자유특구혁신사업육성 지원에 의한 연구임 [P0020333]

Ⅱ. 비정형 데이터를 활용한 비식별화 인공지능 모델 설계

본 논문에서는 물류시스템에서 비정형 데이터의 개인정보의 비식별화 인공지능 모델을 설계한다. 비식별화 인공지능 모델은 물류시스템의 데이터베이스에서 개인정보만을 추출하여 비식별화를 진행하여 활용할 수 있도록 한다.

1. 비식별화 관련연구

기존의 개인정보 비식별화 방법은 크게 가명처리(pseudonymisation), 총계처리(Aggregation), 데

이터 값 삭제(Data Reduction). 범주화(Data Suppression), 데이터 마스킹(data masking) 등이 있다. 가명처리는 개인정보 중 식별요소를 다른 값으로 대체하는 방법이다.[2] 총계처리는 데이터의 총합 값을 보임으로서 개별 데이터의 값을 보이지 않도록 하며, 데이터 값 삭제는 데이터 세트에 구성된 값 중 필요 없는 값 또는 개인 식별 값을 삭제하는 방법이다. 범주화는 데이터의 값을 범주의 값으로 변환하여 명확한 값을 감추는 방법이며, 데이터 마스킹은 공개된 정보 등과 결합하여 개인을 식별하는 데 기여할 확률이 높은 개인 식별자가 보이지 않도록 처리하여 식별을 하지 못하게 하는 방법이다. 본 논문에서는 이 중 가명처리의 한 방법을 제안한다. 가명처리의 단점은 대체 값 부여 시 식별 가능한 고유 속성이 계속 유지되어야 한다는 단점이 존재한다.

가명처리의 기존 알고리즘은 시계열 데이터 망닝, 부분그래프 익명화, 휴리스틱익명화 등이 있다.[3] 시계열 데이터 마이닝은 동일한 속성 값을 가지는 데이터를 k개 이상으로 유지하여 데이터를 공개하는 방법으로서 지정된 속성이 가질 수 없는 값을 k개 이상으로 유지하여 개인정보 노출을 방지한다. 부분그래프 익명화는 소셜네트워크 데이터의 구조적 특징 중 하나인 부분 그래프에 의한 개인정보 노출을 방지하기 위한 익명화 기법으로 익명화를 이해서 그래프 수정을 통해 특정 부분 그래프가 전체 그래프에서 k개 이상 존재하게 만드는 기법이다. 마지막으로 휴리스틱익명화는 해당 값들을 몇 가지 정해진 규칙 혹은 유저의 판단에 따라 가공하여 개인정보를 숨기는 방법이다.

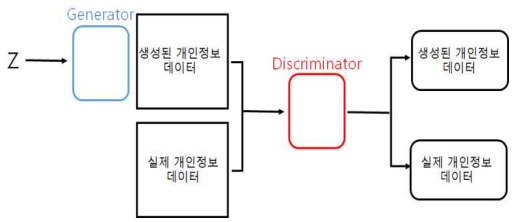

그림 1. SeqGAN 모델 개요

Fig. 1. SeqGAN model overview

2. 비식별화 인공지능 모델 설계

본 논문에서는 SeqGAN(Sequence Generative Adversarial Network)을 통하여 물류시스템에서 비정형 개인정보 데이터 비식별화를 수행한다. 그림 1은 SeqGAN의 모델 개요에 대해 소개한다.

Algorithm 1 Sequence Generative Adversarial Nets

Require: generator policy G_θ; roll-out policy G_β; discriminator D_ϕ; a sequence dataset $S = X_{1:T}$

Initialize G_θ, D_ϕ, with random weights θ, ϕ.
Pre-train G_θ using MLE on S
$\beta \leftarrow \theta$
Generate negative samples using G_θ for training D_ϕ
Pre-train D_ϕ via minimizing the cross entropy
repeat
  for g-steps **do**
    Generate a sequence $Y_{1:T} = (y_1, \cdots, y_T) \sim G_\theta$
    for t in 1 : T **do**
      Compute $Q(a = y_t; s = Y_{1:t-1})$
    end for
    Update generator parameters via policy gradient
  end for
  for d-steps **do**
    Use current G_θ to generate negative examples and combine with given positive examples S
    Train discriminator D_ϕ for k epochs
  end for
  $\beta \leftarrow \theta$
until SeqGAN converges

그림 2. SeqGAN algorithm 흐름도

Fig. 2. Algorithm of SeqGAN

SeqGAN은 기존의 GAN[4]과 동일하게 생성기(Generator)와 판별기(Discriminator)로 구성되어 있다. 생성기는 실제 데이터를 모방하여 새로운 데이터를 생성하고, 판별기는 생성기에 생성한 데이터가 얼마나 진짜 같은지를 판단한다. SeqGAN은 기존의 GAN에서 discrete한 단어 데이터(text data)에 특화된 모델로, 기존 생성기에 강화학습을 접목한 모델로 단어 생성(text Generation)을 실행하게 된다.

그림 2는 SeqGAN의 알고리즘이다.[5] 생성기(G_θ) 은 문장 생성을 위해 LSTM(Long-Short Term Memory)을 사용하며, 판별기(D_ϕ)는 실제 단어와 생성된 단어를 구분하기 위해 Text-CNN 모델을 사용한다. 생성기(G_θ)는 별개의 단어(token)를 생성하는 역할을 하며 판별기(D_ϕ)는 생성기(G_θ)가 생성한 단어를 판단하고 리워드를 주는 역할을 하게된다.

SeqGAN 모델을 학습시키기 위한 데이터는 물류시스템의 비정형 개인정보 데이터(이름, 휴대폰번호, 주소, 계좌번호 등)로 데이터를 수집하고 단어 분석을 위해 KoNLPy[6]의 kkoma 형태소 분석을 통하여 토근화(tokenization)을 진행한다. 학습을 진행하면 판별기는 학습데이터와 동일한 개인정보 데이터를 생성하고 검증을 수행하며 학습을 진행한다.

Ⅲ. 결 론

본 논문에서는 SeqGAN 모델을 사용하여 개인정보 비식별화 설계를 진행하였다. 제안한 설계 모델을 활용하여 기존의 물류데이터에서 비정형 개인정보 데이터와 생성된 개인정보를 변경하게 되면 가명처리가 된 개인정보 데이터를 물류데이터에서 활용할 수 있게 된다.

본 논문에서 제안하는 비식별화 알고리즘을 통해 공공데이터 제공 정부의 지침에 발맞추어 방대한 데이터를 공개하면서도 개인정보를 보호할 수 있는 기법으로 활용할 수 있을 것을 기대한다. 향후 연구에서는 제안하는 알고리즘 설계를 기반으로 구현 및 검증을 진행할 것이다.

참 고 문 헌

[1] 민연주, 장소영, 신민성, "물류산업부문 한국판 뉴딜 추진방안", 기본-PR-21-15, 2021.

[2] 이현승, 송지환, "개인정보 비식별화기술의 쟁점 연구", 연구보고서 2016-001, 2016.

[3] 미래창조과학부, 한국정보화진흥원, 빅데이터전략센터, "빅데이터 활용을 위한 개인정보 비식별화 기수활용 안내서", 2015.

[4] L. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, "Generative adversarial networks". Communications of the ACM, Vol. 63, pp.139-144, 2020.

[5] L. Yu, Z. W. Zhang, J. Wang, Y. Yu, "SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient". Proceedings of the AAAI conference on artificial intelligence, Vol. 31, No. 1, 2017.

[6] E.J Park, S.Z Cho, "KoNLPy:Korean natural language processing in Python". Annual Conference on Human and Language Technology, pp. 133-136, 2014.