분류를 위한 차등 프라이버시 보호 연합 학습

조승현, 이상화, 김수민, 김주희, 채명수*, 임성수

충남대학교, *주식회사 노타

{pmcsh04, sanghwal22, teddy1223, aeon48}@naver.com, *myungsu.chae@nota.ai, sungsu@cnu.ac.kr

Differentially Private Federated Learning for Classification

Seunghyeon Cho, Sanghwa Lee, Soomin Kim, Juhui Kim, Myungsu Chae*, Sungsu Lim

Chungnam National University, *Nota Incorporated

요 약

인공지능의 발전으로 다양한 데이터가 수집 및 분석되면서 프라이버시 침해의 우려가 커지고, 프라이버시 보호가 보장되는 학습 모델의 수요가 급증하고 있다. 본 논문은 프라이버시 보호가 보장되는 인공지능(Privacy-Preserving AI)의 주요 기술인 차등 프라이버시와 연합 학습을 신경망 학습 과정에 접목하여 기계학습의 주요 문제 중 하나인 다중 분류에 적용해보았다. 차등 프라이버시 보호가 보장되는 전송을 수행하는 연합 학습이 그렇지 않은 경우와 비교하여 성능 차이가 작도록 학습할 수 있음을 확인하였으며, 더욱 안전한 연합 학습을 통한 신경망 학습 및 인공지능 문제해결이 가능함을 확인하였다.

Ⅰ. 서 론

산업 전반에 걸쳐서 디지털 전환(Digital Transformation)이 가속화되고 있다. 아날로그 데이터가 디지털 데이터로 변환되고 수집·가공·결합되어 삶의 질 향상과 사회문제 해결을 위한 인공지능(AI) 학습에 쓰이고 있다. 이처럼 AI 기술의 발전은 대량의 데이터를 바탕으로 이루어지기 때문에 데이터 주체자의 프라이버시 문제가 급부상하고 있다. AI 서비스를 사용하기 위해 개인 정보를 밝혀야 하며, 정보가 유출되는 프라이버시 침해의 우려가 발생한다. 프라이버시 보호가 보장되는 AI를 개발하지 않는다면 해킹 공격으로 프라이버시가 침해되는 위험성을 배제할 수 없다.

본 논문에서는 프라이버시 문제를 해소하는 최신 AI 기술을 구현하였고, 이를 통한 AI 서비스의 우수성을 프라이버시 보호와 정확도 두 가지 관점에서 입증하였다. 기계학습의 대표적인 문제 중 하나인 분류 문제해결을 위해 차등 프라이버시(Differential Privacy) 보호가 보장되는 연합 학습(Federated Learning) 모델을 구현하고 실험을 통해 성능을 입증하였다. 금융 데이터의 개인별 특징을 통해 대출상품의 종류를 알아낼 수 있는지 다중 분류를 통해 예측하는 모델을 학습했고, 차등 프라이버시를 적용한 연합 학습이 적용하지 않은 경우보다 더 우수함을 확인하였다.

Ⅱ. 본론

차등프라이버시는 개인 정보 비식별화 기술 중 하나이다. 비식별화란 식별 가능한 데이터의 수정을 통해 특정 개인 식별이 어렵게 하는 것을 말한다. 비식별화 판별의 기준으로 개별화 가능, 연결 가능, 추론 가능의 세 가지가 꼽히며, 이러한 기준에서 가장 뛰어난 기술의 하나가 차등 프라이버시라고 알려져 있다[1]. 차등 프라이버시의 주요 개념은 다음과 같다. 함수 f에 대해 D_1, D_2는 레코드 하나만 다른 두 데이터베이스, t는 임의의 실수라고 하자. 만약 특정 $\varepsilon > 0$에 대해 식 (1)이 항상 성립한다면 함수 f는 프라이버시 수준(level) ε으로 차등 프라이버시가 보호된다고 말한다.

$$\Pr(f(D_1)=t)/\Pr(f(D_2)=t) \le \exp(\epsilon) \quad (1)$$

차등 프라이버시가 보호되는 비식별화 함수 f 설계를 위해 노이즈(noise) 추가, 표본 추출 등의 방법이 쓰이며, Apple의 사용자 행동 패턴 파악[2]을 비롯한 개인 정보 보호를 고려한 AI 서비스에 사용되고 있다.

연합 학습은 기계학습을 중앙 서버가 아닌 각 클라이언트에서 데이터를 처리 후 갱신하는 기법이다[3]. 기존의 데이터를 모아서 학습하는 방법은 민감한 개인 정보가 이동하기 때문에 개인 정보가 침해될 우려가 있다. Google에서 개발한 연합 학습은 데이터 대신 각 기기(device)에서 학습한 모델 또는 가중치를 중앙에서 취합하는 방식으로 분산 학습을 수행하고 정보 유출을 막는다. 널리 쓰이는 연합 학습 방법인 FedAvg[3]의 방식은 다음과 같다. 각 기기는 신경망 학습 과정의 손실(loss) 값을 줄이기 위해 경사 하강법(gradient descent)으로 가중치를 갱신하고, 주기적으로 가중치를 평균값으로 동기화 및 재배포한다. 더 나은 연합 학습을 위한 통신 구조 및 갱신 방식에 관한 연구가 꾸준히 진행되고 있다[4].

차등 프라이버시 보호 연합 학습은 정보 유출을 효과적으로 방지하기 위해 차등 프라이버시 개념을 기반으로 집계 전 클라이언트에서 가중치에 노이즈를 추가하는 프레임워크로 구성된다[5]. 이러한 과정을 통해 중간 갱신값에 의한 프라이버시 노출을 방지하면서 연합 학습을 보다 안전하게 수행할 수 있다. 본 논문에서 구현한 차등 프라이버시 보호 연합 학습의 개념도는 그림 1과 같다. 기기가 각자의 모델 가중치 $\omega_1, \omega_2, \omega_3$를 보내서 평균을 취합하는 대신 목표로 하는 프라이버시 수준이 보장되는 노이즈 η_1, η_2, η_3을 각각 첨가하여 전송 및 취합한다. 평균값을 다시 각 기기로 배포한 후 학습 과정을 반복하는 방식으로 신경망 학습을 통한 학습을 진행하였다. 이를 통해 보안이 강화된 연합 학습 기반 AI 학습을 구현할 수 있었으며, 다중 분류 문제를 위한 학습에 적용 및 탐구하였다.

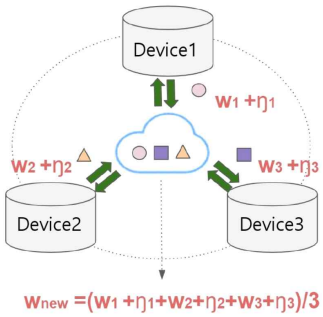

그림 1. 전송 가중치에 차등 프라이버시를 적용한 연합 학습

학습 모델은 다음과 같다. 다중 분류를 위해 신경망 학습 모델인 다층 퍼셉트론(Multi-Layer Perception, MLP)을 기준 모델(이하 Baseline)로 학습하였다. 활성화 함수(activation function)로는 ReLU와 softmax가 쓰여서 각각 비선형성을 반영하고 확률 기반의 학습을 가능하도록 하였다. 기존의 신경망 학습 모델에 대해 프라이버시 보호를 보장하기 위해서 차등 프라이버시와 연합 학습 기술을 도입한 모델을 구현하고 학습하였다. 중앙 서버 1개와 클라이언트 20개가 별(star) 구조로 통신하는 연합 학습 모델(이하 FL), 통신 과정에서 차등 프라이버시 보호가 프라이버시 수준 0.5 이내 보장되는 차등 프라이버시 보호 연합 학습 모델(이하 DP-FL)을 Baseline 모델과 비교하여 개인 정보 보호를 고려한 신경망 학습이 가능하고 다중 분류 문제에 있어서 좋은 정확도와 손실을 보임을 평가하였다. 요약하면 다음의 세 가지 학습 모델을 사용하였다.

- **Baseline**: 프라이버시 보장이 없는 MLP를 활용한 신경망 학습
- **FL**: 클라이언트 20개가 포함된 프로토콜을 활용한 연합 학습
- **DP-FL**: 차등 프라이버시 수준 0.5 이내로 전송하는 연합 학습

성능 평가를 위하여 신경망 학습 반복 횟수에 따른 분류의 정확도(accuracy)와 손실을 비교해보았다. 또한, 차등 프라이버시와 연합 학습 기술이 병행하였을 경우의 효과를 측정하기 위하여 차등 프라이버시를 적용하지 않은 연합 학습 결과와 비교하여 우수성을 평가하였다. 전 과정의 구현은 Python을 통해 진행하였으며, 학습을 위한 신경망 모델로는 keras 라이브러리의 Sequential 모델을 활용하였고, 차등 프라이버시 보호 연합 학습을 위해서 Sherpa.ai의 shfl 라이브러리를 사용하였다. 평가 기준인 정확도는 전체 예측치 중 맞춘 정도를 비율로 나타내었고, 손실은 softmax 계산에 대한 평가로 cross-entropy를 통해 평가하였다.

실험 설계로 금융 빅데이터 개방시스템(CreDB)이 제공하는 신용 정보 데이터를 사용했다[6]. 이 데이터는 신용 정보를 표본 추출한 뒤 비식별 조치하여 무료 배포하는 데이터로 구현한 학습 방법의 유효성을 검증하기 위해 활용하였다. 신경망 모델에 적용하기 위하여 데이터에서 제공되는 차주 정보, 대출 정보, 연체 정보에 대한 세 개의 테이블을 조인하고 불필요한 속성을 제거하였다. 또한, 전처리 및 정규화하는 과정을 거쳤다. 분류 문제는 개인의 나이, 성별, 업권코드, 대출금액, 연체유형코드, 연체사유코드, 등록사유코드, 연체등록금액, 개설사유코드, 카드유형코드 등 10개의 특징(feature)을 통해 대출상품의 레이블(label)을 예측하는 것으로 정의하였으며, 총 6,065개의 개인 신용 정보 데이터에 대해 70%인 4,245개를 훈련용 데이터, 30%인 1,820개를 평가용 데이터로 활용하였다.

실험 결과는 훈련 데이터로 100~200 epochs 만큼 훈련하고 10 epochs 마다 손실과 정확도를 확인하였다. Baseline 모델의 경우 적은 반복 수에 대해서도 빠르게 수렴하며 정확도 0.893, 손실 0.286의 결과를 보였다. 이 결과는 그림 2, 그림 3에 표현한 FL, DP-FL의 결과와 비교할 수 있다. FL은 파란색 선, DP-FL은 빨간색 선, 반복 실험에 따른 신뢰 구간(confidence interval)은 음영으로 나타냈다. 학습 반복의 수가 적었을 때 차등 프라이버시를 적용하지 않은 연합 학습 모델인 FL이 더 빠르게 좋은 성능을 보이지만 반복 수가 증가하면서 차등 프라이버시 보호 연합 학습 모델인 DP-FL도 우수한 정확도와 손실을 보였다. FL은 100 epochs 이후 정확도 0805, 손실 0.761을 보였으며, DP-FL은 100 epochs 이후 정확도 0.770, 손실 1.556으로 다소 떨어지는 결과를 보였다. 하지만 DP-FL을 100 epochs 추가하여 200 epochs 학습한다면 정확도 0.807, 손실 0.789로, 차등 프라이버시 수준 0.5에서 반복 수 2배인 경우 FL과 성능이 유사했다. Baseline, FL, DP-FL 순으로 프라이버시 보호가 보장되는 정도가 더욱 강력해지며, 성능 저하도 반복 수를 늘리면 보완 가능함을 확인하였다.

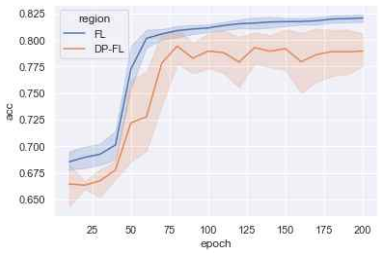

그림 2. 차등 프라이버시 보호 연합 학습의 정확도 평가

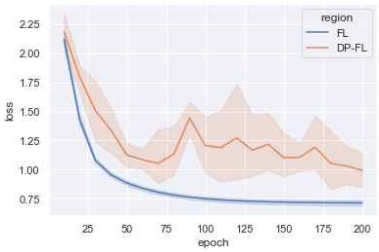

그림 3. 차등 프라이버시 보호 연합 학습의 손실 평가

Ⅲ. 결론

본 논문은 차등 프라이버시 보호가 보장되는 연합 학습 모델을 구현하고 다중 분류 문제에 적용하여 그 성능을 확인하였다. 그 결과 최근 주목받는 연합 학습에 대해서 프라이버시 수준이 이론적으로 보장되는 차등 프라이버시를 접목하고 유사한 성능을 나타낼 수 있음을 확인하였다. 본 논문은 졸업프로젝트 및 산학협력 프로젝트를 수행한 결과로, 프라이버시 보호가 보장되는 인공지능이라는 최신 주제를 학습 및 구현한 것이다. 후속 연구 주제로 더욱 다양한 딥러닝 모델 및 인공지능 문제로 적용하고, 더욱 실용적인 AI 서비스 개발에 접목할 수 있도록 발전시키고자 한다.

ACKNOWLEDGMENT

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2019R1F1A1063231).

참 고 문 헌

[1] Dwork, C., “Differential Privacy: A Survey of Results,” Proc. of ICALP, pp. 1-12, 2006.

[2] Differential Privacy Team, Apple, “Learning with Privacy at Scale,” Technical Report, Dec. 2017.

[3] McMahan, B., Moore, E., Ramage, D., Hampson, S., Arcas, B. A. y. “Communication-Efficient Learning of Deep Networks from Decentralized Data,” Proc. of AISTATS, pp. 1273-1282, 2017.

[4] Lee, J., Oh, J., Lim, S., Yun, S., Lee, J., “TornadoAggregate: Accurate and Scalable Federated Learning via the Ring-Based Architecture,” arXiv:2012.03214, Dec. 2020.

[5] Rodríguez-Barroso, N., Stipcich, G., Jiménez-López, D., Ruiz-Millán, J. A., Martínez-Cámara, E., González-Seco, G., Luzón, M. V., Veganzones, M. A., and Herrera, F., “Federated Learning and Differential Privacy: Software Tools Analysis, The Sherpa.ai FL Framework and Methodological Guidelines for Preserving Data Privacy,” Information Fusion, Volume 64, pp. 270-292, Dec. 2020.

[6] 개인신용정보 표본DB 이용자 매뉴얼, 한국신용정보원, Jun. 2019.