연합학습 환경에서 멤버십 추론·모델 포이즈닝·적대적 예제 공격 통합 평가 및 방어 기법 비교

Unified Evaluation and Defense against Membership Inference, Model Poisoning, and Adversarial Examples in Federated Learning

홍석원[1] · 최지현[1] · 유예찬[2] · 최석환[2]
Seok-Won Hong, Ji-Hyun Choi, Ye-Chan Yu and Seok-Hwan Choi

[1]연세대학교 전산학과
E-mail: {sw.hong, jihyun.choi}@yonsei.ac.kr
[2]연세대학교 소프트웨어학부
E-mail: {zkfla2773, sh.choi}@yonsei.ac.kr

요 약

연합학습(FL)은 원천 데이터를 중앙에 수집하지 않고 참여자가 로컬 데이터로 학습한 업데이트만 공유해 민감한 헬스케어 환경에 적합하다. 그러나 멤버십 추론(MIA), 업데이트 조작 기반 모델 포이즈닝, 적대적 예제 공격이 가능하다. 본 논문은 PGD 기반 적대적 학습과 절사평균 기반 집계 완화를 결합한 방어를 제안하고 MNIST·CIFAR10·UCI MHEALTH(웨어러블/센서 데이터)에서 공격을 동일 흐름으로 정량 평가하였다. MHEALTH에서 MIA와 포이즈닝 공격 성공률이 감소하고, 공격 조건 강건성이 향상됨을 확인하였다.

키워드 : 연합학습, 적대적 학습, 이상치 완화 집계, 멤버십 추론 공격, 모델 포이즈닝

1. 개요

원격 진료 및 비대면 건강관리 서비스의 확산으로 개인 건강 데이터 활용 수요가 증가하고 있다. 다만 헬스케어 데이터는 민감도가 높아 중앙 서버로 원천 데이터를 집적하여 학습하는 방식은 개인정보 유출 위험 및 규제 준수 관점에서 제약이 존재한다. 이에 따라 데이터가 존재하는 위치에서 학습을 수행하고 모델 업데이트만 공유하는 연합학습(Federated Learning, FL)이 대안으로 주목받고 있다[1].

그러나 FL 환경에서도 보안 위협은 여전히 존재한다. 대표적으로 모델의 출력 또는 손실 값을 이용해 특정 샘플이 학습에 포함되었는지를 추정하는 멤버십 추론 공격(MIA)은 프라이버시 침해로 이어질 수 있다[2]. 또한 악성 참여자가 로컬 업데이트를 조작하여 서버 집계 결과를 오염시키는 모델 포이즈닝 공격은 학습 무결성을 저해한다. 더불어 입력을 미세하게 변형하여 오분류를 유도하는 적대적 예제 공격은 실제 서비스의 안정성을 저하시킨다[3].

본 논문에서는 위 세 가지 위협을 동일한 실행 흐름에서 통합 평가하고, 추론 단계와 학습 단계 위협을 동시에 고려한 방어 구성을 제안한다. 구체적으로 로컬 학습에는 PGD 기반 적대적 학습을 적용하고, 서버 집계에는 절사평균 기반 이상치 완화 집계를 적용한다. 제안 방법의 효과는 MNIST, CIFAR10과 더불어 웨어러블/디바이스 센서 기반 공공 데이터인 UCI MHEALTH에서 정량적으로 검증한다.

감사의 글 : 본 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-대학ICT연구센터(ITRC)의 지원(IITP-2025-RS-2023-00259967) 및 2025년도 교육부 및 강원특별자치도의 재원으로 강원RISE센터의 지원을 받아 수행된 지역혁신중심 대학지원체계(RISE) (2025-RISE-10-006)의 연구결과로 수행되었음

2. 본문

2.1 보안 강화형 연합학습 및 방어 구성

본 연구는 FedAvg 기반의 연합학습 절차를 따른다. 서버는 라운드마다 일부 클라이언트를 선택하고, 선택된 클라이언트는 자신의 로컬 데이터로 모델을 학습한 뒤 업데이트를 서버로 전송한다. 서버는 클라이언트별 데이터 크기 등을 고려하여 업데이트를 결합하고 글로벌 모델을 갱신한다.

Baseline은 표준 로컬 학습과 FedAvg 집계를 사용한다. Defense는 다음 두 요소를 결합한다. 첫째, 로컬 학습 단계에서 PGD로 생성한 적대적 샘플을 포함하여 학습하는 적대적 학습을 적용한다. 이는 입력 교란에 대한 모델의 민감도를 낮추어 추론 단계 강건성을 높이는 대표적 방법이다. 둘째, 서버 집계 단계에서 절사평균(trimmed mean)을 적용한다. 절사평균은 파라미터 좌표별로 극단값(이상치)을 일정 비율 제거한 뒤 평균을 계산함으로써, 부호 반전(sign-flip)과 같이 비정상적으로 큰 영향력을 갖는 업데이트를 완화하는 효과가 있다. 즉, Defense는 “추론 단계(적대적 예제)”와 “학습 단계(모델 포이즈닝)” 위협을 동시에 완화하는 구성을 목표로 한다.

2.2 공격 및 성능 평가

표 1에 제시한 평가지표는 Clean(%), MIA(%), Adv ASR(%), Robust(%), AGG ASR(%)로 구성되며, 각 지표는 다음과 같이 산출하였다. Clean(%)는 공격을 적용하지 않은 정상 입력에 대한 테스트 정확도(accuracy)이

표 1. Baseline과 Defense의 성능 비교(단위: %)
Table 1. Performance Comparison between Baseline and Defense (Unit: %)

Dataset	Setting	Clean (%)	MIA (%)	Adv ASR (%)	Robust (%)	AGG ASR (%)
MNIST	Baseline	98.6400	0.1202	96.9700	3.0300	81.2600
	Defense	98.4800	0.0801	27.6800	72.3200	2.0800
CIFAR10	Baseline	84.8100	0.0401	100.0000	0.0000	48.3100
	Defense	74.6700	0.0200	65.4800	34.5200	38.3500
MHEALTH	Baseline	89.1791	0.0520	64.1791	35.8209	34.3284
	Defense	87.5000	0.0080	50.3731	49.6269	19.9627

다. MIA(%)는 손실(loss) 기반 임계값 방식의 멤버십 추론 공격을 적용한 뒤, 오탐률(FPR)을 0.1%로 고정했을 때의 재현율(TPR), 즉 TPR@FPR=0.1%로 정의하였다. Adv ASR(%)는 PGD 기반 적대적 예제 공격을 수행했을 때의 공격 성공률로서, 공격 후 오분류율(%)로 계산하였다. Robust(%)는 동일한 PGD 공격 조건에서의 분류 정확도(적대적 강건 정확도)이며, 본 실험 정의에서는 Robust(%) = 100 − Adv ASR(%)의 관계를 갖는다. AGG ASR(%)는 모델 포이즈닝(aggregation-level) 시나리오에서 악성 클라이언트가 로컬 업데이트의 부호를 반전시키는 sign-flip 공격을 적용했을 때의 공격 성공률(오분류율, %)로 정의하였다. 따라서 AGG ASR(%)가 낮을수록 모델 포이즈닝에 대한 저항성이 높으며, 이에 대응되는 모델 포이즈닝 조건 정확도는 100 − AGG ASR(%)로 환산할 수 있다. 본 논문에서는 공격 저항성은 ASR 계열 지표가 낮을수록, 정확도 계열 지표(Clean, Robust)는 높을수록 우수한 것으로 해석한다.

2.3 데이터셋 구성 및 결과 분석

MNIST와 CIFAR10은 공개 이미지 분류 데이터로, 분산 환경을 모사하기 위해 전체 학습 데이터를 클라이언트별 로컬 데이터로 분할하여 각 클라이언트가 자신의 로컬 데이터만으로 학습을 수행하도록 구성하였다. UCI MHEALTH는 웨어러블/디바이스 센서 기반 활동 인식 공공 데이터로, 사용자(피험자) 단위의 데이터 로컬리티가 유지되도록 분할하여 클라이언트별 로컬 학습을 구성하였다. 센서 신호는 일정 길이의 윈도우로 분할하고 정규화를 적용한 뒤 분류 입력으로 사용하였다.

표 1은 데이터셋별 Baseline과 Defense의 성능을 Clean(%), MIA(%), Adv ASR(%), Robust(%), AGG ASR(%) 관점에서 비교한 결과이다. 먼저 MNIST의 경우 Defense 적용 후 Clean은 98.64%에서 98.48%로 거의 유지되었고, MIA는 0.1202%에서 0.0801%로 감소하였다. 또한 적대적 예제에 대한 공격 성공률(Adv ASR)은 96.97%에서 27.68%로 크게 감소하였으며, 이에 따라 PGD 조건 정확도(Robust)는 3.03%에서 72.32%로 크게 향상되었다. 모델 포이즈닝 측면에서도 AGG ASR이 81.26%에서 2.08%로 감소하여 sign-flip 기반 모델 포이즈닝에 대한 저항성이 크게 향상됨을 확인하였다.

CIFAR10에서는 Defense 적용 시 MIA가 0.0401%에서 0.0200%로 감소하였고, Adv ASR은 100.00%에서 65.48%로 낮아지면서 Robust는 0.00%에서 34.52%로 향상되었다. 또한 AGG ASR 역시 48.31%에서 38.35%로 감소하여 모델 포이즈닝 저항성이 개선되었다. 다만 Clean이 84.81%에서 74.67%로 감소하여, 본 데이터셋에서는 강건성 향상과 정상 정확도 간의 trade-off가 상대적으로 크게 나타났다.

UCI MHEALTH에서는 Baseline에서 Clean이 89.18%로 높게 형성되었고, Defense 적용 후 Clean은 87.50%로 1.68%p 감소하는 수준이었다. 반면 보안 관련 지표는 개선되는 경향을 보였다. MIA는 0.0520%에서 0.0080%로 감소하여 멤버십 추론 관점의 노출 위험이 낮아졌고, Adv ASR은 64.18%에서 50.37%로 감소하면서 Robust는 35.82%에서 49.63%로 향상되었다. 또한 AGG ASR은 34.33%에서 19.96%로 감소하여(환산 시 모델 포이즈닝 조건 정확도는 65.67%에서 80.04%로 증가), 센서/디바이스 기반 데이터에서도 모델 포이즈닝에 대한 저항성이 함께 개선됨을 확인하였다.

3. 결론

본 논문에서는 연합학습 환경에서의 학습 단계 위협(모델 포이즈닝), 추론 단계 위협(적대적 예제), 프라이버시 위협(MIA)을 통합 평가하고, PGD 기반 적대적 학습과 절사평균 기반 이상치 완화 집계를 결합한 방어 구성을 제안하였다. MNIST와 UCI MHEALTH에서 제안 방법은 공격 성공률을 낮추고 강건 정확도를 향상시키는 경향을 보였으며, 특히 UCI MHEALTH에서 MIA 지표 및 모델 포이즈닝 관련 지표가 동시에 개선되어 센서/디바이스 기반 디지털 헬스케어 시나리오에 대한 적용 가능성을 확인하였다. 다만 CIFAR10에서는 강건성 향상과 함께 Clean 정확도 저하가 관찰되어, 데이터셋 특성에 따른 적대적 학습 강도 및 집계 파라미터 조정이 필요하다. 또한 MIA 평가는 손실 기반 단순 공격으로 구성하였으므로, 향후 shadow model 기반 공격 등 다양한 공격자 모델로 확장 평가할 계획이다.

참 고 문 헌

[1] MCMAHAN, Brendan, et al. Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics. Pmlr, 2017. p. 1273-1282.

[2] SHOKRI, Reza, et al. Membership inference attacks against machine learning models. In: 2017 IEEE symposium on security and privacy (SP). IEEE, 2017. p. 3-18.

[3] MADRY, Aleksander, et al. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083, 2017.