차분 프라이버시를 적용한 연합학습 연구

이주은[1], 김영서[2], 이수빈[1], 배호[3,4]
[1]이화여자대학교 인공지능융합전공 석사과정
[2]이화여자대학교 인공지능융합전공 석박사통합과정
[3]이화여자대학교 사이버보안학과 교수
[4]이화여자대학교 인공지능융합전공 교수
jel@ewhain.net, yskim0411@ewha.ac.kr, capablebear@ewha.ac.kr, hobae@ewha.ac.kr

Research on Federated Learning with Differential Privacy

Jueun Lee[1], YoungSeo Kim[2], SuBin Lee[1], Ho Bae[3,4]
[1,2]Dept. of Artificial Intelligence Convergence, Ewha Womans University
[3]Dept. of Cyber Security, Ewha Womans University
[4]Dept. of Artificial Intelligence Convergence, Ewha Womans University

요 약

연합학습은 클라이언트가 중앙 서버에 원본 데이터를 주지 않고도 학습할 수 있도록 설계된 분산된 머신러닝 방법이다. 그러나 클라이언트와 중앙 서버 사이에 모델 업데이트 정보를 공유한다는 점에서 여전히 추론 공격(Inference Attack)과 오염 공격(Poisoning Attack)의 위험에 노출되어 있다. 이러한 공격을 방어하기 위해 연합학습에 차분프라이버시(Differential Privacy)를 적용하는 방안이 연구되고 있다. 차분 프라이버시는 데이터에 노이즈를 추가하여 민감한 정보를 보호하면서도 유의미한 통계적 정보 쿼리는 공유할 수 있도록 하는 기법으로, 노이즈를 추가하는 위치에 따라 전역적 차분 프라이버시(Global Differential Privacy)와 국소적 차분 프라이버시(Local Differential Privacy)로 나뉜다. 이에 본 논문에서는 차분 프라이버시를 적용한 연합학습의 최신 연구 동향을 전역적 차분 프라이버시를 적용한 방향과 국소적 차분 프라이버시를 적용한 방향으로 나누어 검토한다. 또한 이를 세분화하여 차분 프라이버시를 발전시킨 방식인 적응형 차분 프라이버시(Adaptive Differential Privacy)와 개인화된 차분 프라이버시(Personalized Differential Privacy)를 응용하여 연합학습에 적용한 방식들에 대하여 특징과 장점 및 한계점을 분석하고 향후 연구방향을 제안한다.

1. 서론

연합학습(Federated Learning, FL)[1]은 중앙 집중식 저장 없이 다양한 기기나 분산된 머신러닝 방법으로, 여러 클라이언트가 개별 데이터로 학습하여 얻은 모델 업데이트 정보를 중앙 서버로 전송하면, 서버가 그 정보를 집계하여 학습하는 방식이다.

이 FL 방식은 클라이언트와 중앙 서버 간에 원본 데이터(Original Data)를 공유하지 않는다는 점에서 민감한 데이터가 유출될 위험을 줄일 수 있다는 이점이 있다. 이러한 이점으로 FL은 금융, 헬스케어 등 다양한 산업에서 활발하게 사용되고 있다[2].

그러나 이 방식은 크게 2가지 문제점이 있다[3].

첫째, FL은 추론 공격(Inference Attack) 위험에 노출되어 있다. 구체적으로, FL에서 클라이언트와 중앙 서버 사이에 원본 데이터가 아닌 모델 업데이트 정보만 공유하더라도, 악의적인 공격자가 모델 업데이트 정보에 포함된 기울기나 가중치를 통하여 클라이언트의 정보를 유추해낼 위험이 있다[4].

둘째, FL은 오염 공격(Poisoning Attack) 위험에 노출되어 있다. 즉, 클라이언트 중에 공격자가 숨어 있는 경우, 공격자가 잘못된 레이블 데이터로 유도한 모델 업데이트 정보를 중앙 서버로 전송함으로써 전역적 모델을 악의적으로 조작할 위험이 있다[5].

이러한 FL 공격 위험을 제거하는 방법론으로는 대표적으로 차분 프라이버시(Differential Privacy, DP)[6]를 적용하는 방법이 있다[7].

DP는 데이터 집합에 노이즈를 추가하여 개별 클라이언트의 정보를 보호하면서도 유의미한 통계적 정보 쿼리는 공유할 수 있도록 하는 기법이다. 이는 D_1과 D_2는 구성 요소가 하나만 다른 데이터 집합이라면 알고리즘 K에 D_1과 D_2를 각각 적용할 때 결과가 S∈range(K)이고 다음 부등식이 성립한다면 ε, δ가 모두 양수일 때 알고리즘 K가 (ε,δ)-차분 프라이버시를 만족한다[1]고 할 수 있다.

1) $\Pr[K(D_1)\in S] \le e^{\epsilon} \times \Pr[K(D_2)\in S]+\delta$

DP는 어느 시점이나 부분에 노이즈를 추가하는지에 따라 전역적 차분 프라이버시(Global Differential Privacy, GDP)[8]와 국소적 차분 프라이버시(Local Differential Privacy, LDP)[9]로 구분할 수 있다.

GDP는 중앙 집계자가 개별 데이터를 결합하여 데이터 집합 전체에 노이즈를 추가하는 방식이다. 이 방법은 적은 노이즈로도 전체 데이터 집합 전체의 DP를 보장할 수 있다는 장점이 있으나, 중앙 집계자의 높은 신뢰성이 요구된다는 단점이 있다.

LDP는 각 클라이언트가 자신의 데이터에 노이즈를 추가하는 방식이다. 이 방법은 중앙 집계자가 필요하지 않아 신뢰성 문제가 발생하지 않는다는 장점이 있지만, 각 클라이언트가 개별 데이터에 노이즈를 추가하기 때문에 전체 노이즈가 GDP 방식에 비해 커지게 되어 유의미한 쿼리를 공유하기 위해서는 더 많은 클라이언트가 필요하다는 단점이 있다.

또한, LDP나 GDP의 세부 방식은 크게 3가지로 나눌 수 있다. 고전적인 DP 방식과 고전적인 DP 방식을 응용하여 발전시킨 개인화된 차분 프라이버시(Personalized Differential Privacy, PDP), 그리고 적응형 차분 프라이버시(Adaptive Differential Privacy, ADP)가 있다[10], [11].

고전적인 DP 방식은 모든 데이터의 소유자에게 동등한 프라이버시 보호 수준을 제공하는 것을 목표로 하여 모두에게 비슷한 수준으로 무작위 노이즈를 추가함으로써 데이터를 보호하는 방식이다.

PDP는 데이터 소유자가 프라이버시 보호 수준을 개별적으로 설정할 수 있도록 하는 방식이다.

ADP는 데이터 집합이 시간이 지남에 따라 자동으로 업데이트될 때, 이에 따른 프라이버시 보호 수준도 적응적으로 유지될 수 있도록 하는 방식이다.

본 논문에서는 FL에 DP를 적용한 최신 연구 동향을 GDP를 적용한 방법론과 LDP를 적용한 방법론으로 나누어 검토하고, 각 세부 방식에 나누어 그 특성을 분석한다. 또한, 이를 토대로 향후 발전시킬 수 있는 연구 방향성을 제시하고자 한다.

2. 차분 프라이버시를 적용한 연합학습

2.1. 전역적 차분 프라이버시를 적용한 연합학습

전역적 차분 프라이버시를 적용한 연합학습 연구로는 [12], [13], [14], [15], [16], [17] 등이 있다.

이 중 Robin C.Geyer et al.[12]은 처음으로 FL에 GDP를 적용한 알고리즘을 제안하였다. 이 방식은 고전적인 DP를 FL에 적용한 방식이다. 이는 노이즈

<표 1> 전역적 차분 프라이버시를 적용한 연합학습

세부방식	고전적인 GDP	GDP+ADP
인용번호	[12],[13],[14],[15],[16]	[17]
특징	모든 데이터를 동등하게 보호	모든 데이터를 보호할 필요 없다고 가정
트레이드오프 문제 접근방식	정확도와 프라이버시 보호 사이 균형점 탐색	보호하지 않아도 되는 데이터에 노이즈 추가 안 함
장점	복잡한 계산이나 구성이 불필요	중요한 데이터만 보호한다는 아이디어로 낭비되는 프라이버시 보호 예산을 최소화할 가능성 생김
한계점	실제 산업에서는 클라이언트, 데이터 별로 요구하는 프라이버시 보호 수준이 같지 않다는 점을 간과	1. 데이터 별 프라이버시 요구 수준 계산 필요 2. 요구 수준 계산할 때 고려하는 요소가 적어 정확한 계산 어려움

를 많이 추가할수록 프라이버시 보호 수준은 높아지지만 모델 정확도는 낮아진다는 DP의 전형적인 트레이드오프(trade-off) 한계를 내포한다.

이후 이런 한계점을 극복하고자 Kang Wei et al.[15]이 Noising before Model Aggregation Federated Learning(NbAFL)을 제안하였다. 이는 클라이언트 측 파라미터에 노이즈를 추가하여 집계한 후 그 결과에도 노이즈를 추가하는 방식으로, 트레이드오프 문제에서 균형점을 개선할 수 있는 파라미터 값을 제안하였다. 그러나 실제 산업에서는 각 클라이언트의 프라이버시 보호 요구 수준이 다르다는 사실을 고려하지 않아 모든 클라이언트를 동등하게 보호하였고, 프라이버시 보호에 필요한 비용에 대하여도 고려하지 않았다는 한계점이 있다.

이러한 비용 문제에 대하여 지적한 대표적인 논문으로는 2022년 제안된 Rui Hu et al.[17]이 있다. 해당 논문에서는 트레이드오프를 개선하고 DP 적용 시 발생하는 비용을 줄이고자 Federated Learning with Sparsified Model Perturbation(Fed-SMP)을 제안하였다. Fed-SMP는 클라이언트의 모델 업데이트 정보에서 핵심 가중치 일부에만 노이즈를 추가하는 방식이다. 이처럼 불필요한 노이즈를 줄임으로써, 동일한 프라이버시 수준에서 모델 정확도가 높아졌다. 각 데이터마다 필요한 수준의 노이즈를 적용하고자 한 것이니 ADP가 적용되었다 볼 수 있으나, 해당 방식에서 노이즈를 추가할 파라미터를 선택하는 과정이 무작위로 고르는 것이거나 절댓값만 확인하여 적용하는 것이 전부라는 점에서 한계가 있다.

2.2. 국소적 차분 프라이버시를 적용한 연합학습

2020년부터 중앙 집계자가 필요하지 않은 LDP를 FL에 적용하는 연구가 활발히 진행되기 시작하였다. FL에 LDP를 적용한 논문으로는 [18], [19], [20], [21], [22], [23], [24] 등이 있다.

처음 FL에 LDP를 적용한 것은 2020년 Wang, Y. et al.[18]으로, FedLDA를 제안하였다. 이 알고리즘

<표 2> 국소적 차분 프라이버시를 적용한 연합학습

세부방식	고전적인 LDP	LDP+ADP	LDP+PDP
인용번호	[18],[20]	[24]	[19],[21],[22],[23],[24]
특징	모든 클라이언트, 데이터를 동등 보호	데이터 분포, 시계열에 따라 데이터 차등 보호	클라이언트 프라이버시 요구에 따라 개인화 보호
트레이드 오프 문제 접근방식	전반적으로 불필요한 노이즈 최소화	데이터의 특성이나 시점에 따라 필요한 수준으로만 노이즈 추가	클라이언트, 기기 특성 등에 따라 최적화된 수준으로 노이즈 추가
장점	복잡한 계산 불필요	각 데이터를 필요한 수준으로만 보호한다는 아이디어로 프라이버시 보호 예산을 최소화할 가능성 생김	클라이언트별 요구하는 수준에 따라 보호한다는 아이디어로 프라이버시 보호 예산을 최소화할 가능성 생김
한계점	데이터 분포가 다르거나 클라이언트별로 특성이 다르다면 낭비되는 프라이버시 보호 예산 발생	1. 데이터 별 노이즈 수준에 대해 지속적인 계산이 필요 2. 주로 단일 요소만 고려하여 정확한 계산 어려움	주로 단일 요소만 고려하여 정확한 계산이 어려움

은 사전 지식을 활용하여 노이즈를 조정하는 방식인 RRP라는 LDP를 개발함으로써 FL 환경에서의 Latent Dirichlet Allocation(LDA)에 적용한 방식이다. 해당 방식은 적절한 노이즈의 크기를 구할 수 있도록 사전 지식을 활용한다는 점에서, 과도하게 노이즈가 적용되는 것을 막을 수 있다는 장점이 있다. 그러나 이 방식은 모든 클라이언트의 프라이버시 보호 수준이 같다는 가정을 하였다.

이후 Kang Wei et al.[19]은 실제 환경에서는 클라이언트 별로 요구하는 프라이버시 요구 수준이 다르다는 점을 지적하면서 User-Level Differential Privacy(UDP)를 제안하였다. UDP는 사용자가 모델 업데이트 전에 선택한 ε, δ 값을 기반으로 공유된 모델에 노이즈를 추가하는 방식이다. 따라서 FL에 PDP를 적용한 것이라 볼 수 있으나, 사용자가 자신의 데이터 보호 수준을 알아야 한다는 한계가 있다.

이후 제안된 방법론으로 Fang Dong et al.[24]이 있다. 해당 논문에서는 Personalized and Adaptive Differentially Private Federated Meta Learning Mechanism(PADP-FedMeta)을 제안하였다. 이는 실제 산업에서는 데이터가 서로 독립적이지도, 고르게 분포되어 있지도 않은 경우가 많으나 과거 FL에 DP를 적용할 때는 IID(Independent and Identically Distributed)를 가정하였다는 사실을 지적하며 제안되었다. 구체적으로, 각 클라이언트에 대하여 개인화된 모델 파라미터를 학습할 수 있으면서도 업데이트 결과에 따라 노이즈 크기를 조절할 수 있는 연합 메타 학습을 도입한 방식이다. 그러나 이 방식은 데이터의 민감도에만 초점을 맞춘다는 한계가 있다.

2.3. 향후 연구 방향

FL에 DP를 적용하려는 연구가 발전하면서 최근 PDP와 ADP를 FL에 적용하고자 하는 연구가 주를 이루고 있다. 그러나 이러한 연구들은 데이터의 민감도만을 고려하여 계산한다는 점에서 한계가 있다.

따라서 PDP와 ADP를 FL에 적용할 때, 클라이언트의 프라이버시 보호 요구 수준을 더욱 정확하게 계산할 메커니즘이 필요하다. 이를 위해 데이터의 특성, 클라이언트의 속성 등을 종합적으로 고려하여 프라이버시 보호 수준을 조절하는 PDP와 ADP를 설계해 FL에 적용하는 연구가 이루어져야 한다.

3. 결론

FL에 존재하는 추론 공격, 오염 공격을 방어하기 위해 FL에 DP를 적용하고자 하는 연구가 활발하게 진행되고 있다. 최근에는 GDP보다는, 중앙 집계자를 필요로 하지 않아 신뢰성 이슈가 발생하지 않는 LDP를 FL에 적용하는 연구가 더욱 활발하게 진행되고 있는 것으로 나타났다. 특히, PDP와 ADP를 적극적으로 활용한 LDP를 적용하여 클라이언트 별로 필요한 수준의 프라이버시 보호 예산을 편성하려는 시도가 2022년을 기점으로 빈번하게 관찰되었다.

그러나 현재 연구에서는 FL에 PDP나 ADP를 적용함에 있어 한 가지 변수만을 고려한다는 문제점이 있다. 이로 인해 개별 클라이언트의 프라이버시 보호 요구가 잘못 계산될 가능성이 남아 있다. 이를 해결하기 위해 앞으로 FL에 적용할 PDP와 ADP는 데이터의 민감도, 분포, 모델의 집계 횟수, 클라이언트의 속성 등 다양한 변수를 종합적으로 고려하는 방향으로 연구가 이루어져야 할 것이다.

4. 사사

이 논문은 2024년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.RS-2022-00155966, 인공지능융합혁신인재양성(이화여자대학교))

참고문헌

[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] HAO, Meng, et al. Efficient and privacy-enhanced federated learning for industrial artificial intelligence.IEEE Transactions on Industrial Informatics, 2019, 16.10: 6532-6542.

[3] LYU, Lingjuan; YU, Han; YANG, Qiang. Threats to federated learning: A survey.arXiv preprint arXiv:2003.02133, 2020.

[4] ZHANG, Chen, et al. A survey on federated learning.Knowledge-Based Systems, 2021, 216: 106775.
[5] BAGDASARYAN, Eugene, et al. How to backdoor federated learning. In:International conference on artificial intelligence and statistics. PMLR, 2020. p. 2938-2948.
[6] DWORK, Cynthia. Differential privacy. In: International colloquium on automata, languages, and programming. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006. p. 1-12.
[7] LI, Qinbin, et al. A survey on federated learning systems: Vision, hype and reality for data privacy and protection.IEEE Transactions on Knowledge and Data Engineering, 2021, 35.4: 3347-3366.
[8] WANG, Huazheng, et al. Global and local differential privacy for collaborative bandits. In:Proceedings of the 14th ACM Conference on Recommender Systems. 2020. p. 150-159.
[9] ARACHCHIGE, Pathum Chamikara Mahawaga, et al. Local differential privacy for deep learning.IEEE Internet of Things Journal, 2019, 7.7: 5827-5842.
[10] JORGENSEN, Zach; YU, Ting; CORMODE, Graham. Conservative or liberal? Personalized differential privacy. In:2015 IEEE 31St international conference on data engineering. IEEE, 2015. p. 1023-1034.
[11] PHAN, NhatHai, et al. Adaptive laplace mechanism: Differential privacy preservation in deep learning. In:2017 IEEE international conference on data mining (ICDM). IEEE, 2017. p. 385-394.
[12] GEYER, Robin C.; KLEIN, Tassilo; NABI, Moin. Differentially private federated learning: A client level perspective.arXiv preprint arXiv:1712.07557, 2017.
[13] MCMAHAN, H. Brendan, et al. Learning differentially private recurrent language models.arXiv preprint arXiv:1710.06963, 2017.
[14] TRIASTCYN, Aleksei; FALTINGS, Boi. Federated learning with bayesian differential privacy. In:2019 IEEE International Conference on Big Data (Big Data). IEEE, 2019. p. 2587-2596.
[15] WEI, Kang, et al. Federated learning with differential privacy: Algorithms and performance analysis.IEEE transactions on information forensics and security, 2020, 15: 3454-3469.
[16] ZHANG, Xinwei, et al. Understanding clipping for federated learning: Convergence and client-level differential privacy. In:International Conference on Machine Learning, ICML 2022. 2022.
[17] HU, Rui; GUO, Yuanxiong; GONG, Yanmin. Federated learning with sparsified model perturbation: Improving accuracy under client-level differential privacy.IEEE Transactions on Mobile Computing, 2023.
[18] WANG, Yansheng; TONG, Yongxin; SHI, Dingyuan. Federated latent dirichlet allocation: A local differential privacy based framework. In:Proceedings of the AAAI Conference on Artificial Intelligence. 2020. p. 6283-6290.
[19] WEI, Kang, et al. User-level privacy-preserving federated learning: Analysis and performance optimization.IEEE Transactions on Mobile Computing, 2021, 21.9: 3388-3401.
[20] SHI, Lu, et al. HFL-DP: Hierarchical federated learning with differential privacy. In:2021 IEEE Global Communications Conference (GLOBECOM). IEEE, 2021. p. 1-7.
[21] ZHOU, Hao, et al. PFLF: Privacy-preserving federated learning framework for edge computing.IEEE Transactions on Information Forensics and Security, 2022, 17: 1905-1918.
[22] WU, Chuhan, et al. A federated graph neural network framework for privacy-preserving personalization.Nature Communications, 2022, 13.1: 3091.
[23] SHEN, Xiaoying, et al. Pldp-fl: Federated learning with personalized local differential privacy.Entropy, 2023, 25.3: 485.
[24] DONG, Fang, et al. PADP-FedMeta: A personalized and adaptive differentially private federated meta learning mechanism for AIoT.Journal of Systems Architecture, 2023, 134: 102754.