연합학습 환경에서 클라이언트 선택의 최적화 기법

박민정 [1], 손영진 [2], 채상미 [3*]
[1] 금오공과대학교 경영학과 교수
[2] 이화여자대학교 경영학부 박사과정
[3] 이화여자대학교 경영학부 교수

mjpark@kumoh.ac.kr, teumdal@ewhain.net, smchai@ewha.ac.kr

요 약

연합학습은 중앙 서버에서 데이터를 수집하는 방식이 아닌 로컬 디바이스 또는 클라이언트에서 학습을 진행하고 중앙 서버로 모델 업데이트만 전송하는 분산 학습 기법으로 데이터 보안 및 개인정보보호를 강화하는 동시에 효율적인 분산 학습을 수행할 수 있다. 그러나, 연합학습 대부분의 시나리오는 클라이언트의 서로 다른 분포 형태인 non-IID 데이터를 대상으로 학습함에 따라 중앙집중식 모델에 비하여 낮은 성능을 보이게 된다. 이에 본 연구에서는 연합학습 모델의 성능을 개선하기 위하여 non-IID 의 환경에서 참여 후보자 중에서 적합한 클라이언트 선택의 최적화 기법을 분석한다.

1. 서론

최근 스마트 기기, 엣지 컴퓨팅, 사물인터넷 등 디지털 기술의 발전으로 인하여 다양한 산업 분야에서 방대한 양의 데이터가 생성되고 있다. 이러한 데이터를 클라우드 서버에서 모두 수집하고 처리하기 위해서는 긴 분석 시간과 비용 등이 수반됨에 따라, 비효율성이 발생할 수 밖에 없다. 또한, 사용자의 개인정보보호 중요성에 대한 인식과 보안 위협 요소가 증가하는 환경적인 변화를 고려하여 연합학습이 등장하였다. 연합학습은 여러 단말기가 서로 데이터를 공유하지 않고 개인 단말기에서 학습을 수행하는 분산형 학습 기법이다 [1,2]. 즉, 사용자가 데이터를 소유하고 있는 단말기에서 직접 데이터를 처리하여 모델을 학습하는 구조로 중앙 서버에 데이터의 취합 과정이 불필요하다. 따라서, 연합학습 환경에서는 민감 데이터를 비롯한 학습 데이터를 중앙 서버에 모두 취합하지 않고도 모델을 학습할 수 있다. 이는 중앙집중형 방식의 학습보다 사용자의 개인정보를 포함한 데이터에 대한 보안성을 유지할 수 있다는 점에서 이점을 갖고 있다 [3]. 특히, 연합학습은 프라이버시 보존을 기반으로 다양한 불균형 데이터의 처리 성능 개선, 모델의 크기와 복잡성 감소를 통한 모델 효율성을 개선하기 위한 연구가 활발히 이루어지고 있다. 그럼에도 불구하고 현존하는 연합학습 대부분의 시나리오는 클라이언트의 종속적인 이질적 분포 데이터를 사용하여 학습하게 되며, 해당 결과는 중앙집중식 모델에 비하여 저하된 성능을 보임에 따라 개선 필요성이 제기되고 있다 [4,5].

2. 연합학습 환경에서의 클라이언트 선택 기법

연합학습 모델의 학습 결과는 클라이언트 간 시스템 다양성, 보유한 통신 자원, 데이터 분포의 이질성 등에 따라 영향을 받는다. 특히, 데이터 분포의 이질성은 학습에 참여하는 클라이언트들의 데이터가 독립적이고 서로 다른 분포를 가지고 있음을 의미한다.

대표적으로 연합학습 환경에서는 데이터를 업로드할 클라이언트를 임의로 선택하는 무작위 방식을 (random selection) 채택하는 경우가 존재한다. 그러나 이와 같은 무작위 클라이언트 선택 방식은 FedAvg 방법을 기반으로 이루어지며 모든 클라이언트의 업데이트를 동일하게 처리함에 따라, 각 클라이언트가 모델 학습에 선택될 확률이 동일하다. 또한, 각 클라이언트가 가진 데이터 분포의 이질성, 데이터 가치 등을 고려하지 않기 때문에 노이즈가 많은 이상 데이터가 학습에 포함되거나 데이터의 손실 문제 등이 발생한다 [4,7]. 스마트폰의 실제 데이터를 대상으로 수행한 연구 결과에 따르면, 데이터 분포 이질성으로 인하여 연합학습 모델의 정확성이 최대 9.2% 감소하고 수렴 시간이 약 2.64 배 증가한 것으로 밝혀졌다 [6].

Greedy 알고리즘에 기반한 클라이언트 선택 방법은 휴리스틱 방법을 통하여 각 클라이언트의 특징을 평가하여 등급이 높고 상대적으로 적은 비용이 요구되는 클라이언트를 선택하는 방식이다 [8,9]. 이 과정에서 각 클라이언트는 글로벌 모델을 학습하고 연합학

습의 모델을 평가하기 위하여 로컬 데이터의 하위 집합(subset)을 사용한다 [9]. Greedy 기반 방식은 각 라운드에서 가장 빠르게 학습을 수행한 클라이언트가 아닌 평균적 기여도가 가장 큰 클라이언트를 선택한다[8,9]. 또한, greedy 방식의 클라이언트 선택 기법은 무작위 선택 방식과 동일하게 데이터 품질을 고려하지 않는 다는 점에서 모델의 정확도가 점진적으로 낮아진다 [10].

클러스터링 선택 (clustering selection)은 클라이언트가 보유한 리소스, 데이터, 특징, 유사성, 경사 손실 등의 다양한 속성을 유사성에 따라 클라이언트를 선택하여 이들을 대상으로 클러스터링함에 따라, 전체 모델의 효율성을 높이고 모델 학습 성능을 향상시키는 방식이다 [11].

3. 결론 및 향후 연구방안

연합학습은 비효율적 통신 방식, 클라이언트 및 데이터의 통계적 이질성, 열악한 데이터 품질, 프라이버시 문제 등의 해결을 통한 모델의 성능 개선이 필요하며 이를 위해서는 이에 기여할 수 있는 적합한 클라이언트를 효율적으로 선택하는 것은 필수적이다. 이를 위하여 본 연구에서는 현재까지 연구된 클라이언트 선택 기법을 검토하였다. 그 결과, 각 방식은 모두 클라이언트가 보유한 데이터의 품질, 가치를 고려하지 못하는 동시에 이 과정에서 발생하는 공정성의 문제를 해결하여야 함이 밝혀졌다. 따라서, 본 연구는 향후, 이와 같은 기존 기법의 한계점을 개선하는 동시에 모델 성능 향상에 기여하는 클라이언트 선택의 최적화 방식 연구 필요성을 제시하였다.

참고문헌

[1] Yu, H., Liu, Z., Liu, Y., Chen, T., Cong, M., Weng, X., Niyato, D., Yang, Q.: “A fairness-aware incentive scheme for federated learning.” In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2020, 393-399.

[2] Zhan, Y., Li, P., Guo, S., Qu, Z.: “Incentive mechanism design for federated learning: Challenges and opportunities.” IEEE Network 35, 2022, 310-317.

[3] Ghosh, A., Chung, J., Yin, D., & Ramchandran, K. “An efficient framework for clustered federated learning.” Advances in Neural Information Processing Systems, 33, 2020, 19586-19597.

[4] Fu, L., Zhang, H., Gao, G., Zhang, M., & Liu, X. “Client selection in federated learning: Principles, challenges, and opportunities.” IEEE Internet of Things Journal, 2023.

[5] 이채은, & 이웅희. “Non-IID 데이터 분산 환경에서의 연합학습 참여 기기 선택 기법 연구.” 차세대융합기술학회논문지, 6(11), 2022, 2063-2075.

[6] C. Yang, M. Xu, Q. Wang, et al. “Flash: Heterogeneity-aware federated learning at scale.” IEEE Transactions on Mobile Computing, 1–18, 2022

[7] Tan, X., Ng, W. C., Lim, W. Y. B., Xiong, Z., Niyato, D., & Yu, H. “Reputation-Aware Federated Learning Client Selection based on Stochastic Integer Programming.” IEEE Transactions on Big Data., 2022.

[8] Mohammed, I., Tabatabai, S., Al-Fuqaha, A., El Bouanani, F., Qadir, J., Qolomany, B., & Guizani, M. “Budgeted online selection of candidate IoT clients to participate in federated learning.” IEEE Internet of Things Journal, 8(7), 5938-5952. 2020.

[9] Ji, S., Jiang, W., Walid, A., & Li, X. “Dynamic sampling and selective masking for communication-efficient federated learning.” IEEE Intelligent Systems, 37(2), 27-34. 2021.

[10] Nishio, T., & Yonetani, R. “Client selection for federated learning with heterogeneous resources in mobile edge.” In ICC 2019-2019 IEEE international conference on communications (ICC) 1-7. 2019.

[11] Wang, S., & Chang, T. H. “Federated matrix factorization: Algorithm design and application to data clustering.” IEEE Transactions on Signal Processing, 70, 1625-1640. 2022.