효율적 연합학습을 위한 참여자 기여도 기반 클라이언트 선택 방법

정영환, 최원기, 이상신*

한국전자기술연구원

{cjstntjd, cwk1412}@keti.re.kr, *sslee@keti.re.kr

A client selection method based on participant contribution for efficient federated learning

Jeong Young Hwan, Choi Won Gi, Lee Sang Shin*

Korea Electronics Technology Institute

요 약

최근 의료, 제약 등 다양한 민감 도메인에서 생성되는 데이터를 기반으로 인공지능 모델을 훈련하고자 하는 시도가 증가하면서 연합학습(Federated Learning) 기술이 주목받고 있다. 그러나 다양한 장치 단말에 분산된 데이터를 통해 공통된 모델을 훈련하기 위해서는 무작위 클라이언트에 대한 한정된 정보를 기반으로 모델을 훈련해야 하는 도전적인 문제를 해결해야한다. 이러한 한계를 극복하고 지역화된 로컬 데이터 기반으로 고성능의 AI 모델을 훈련하기 위해, 이 논문은 참여자 기여도 기반 클라이언트 선택 방법을 제안한다. 제안하는 방법은 연합학습 참여자의 데이터 및 훈련 품질이 전역 모델에 기여하는 정도를 정량적으로 수치화하여 훈련 절차를 조율해 고성능의 전역 모델을 신속하게 훈련한다. 이와 같은 방식은 단순 연합학습 훈련 체계와 비교했을 때 참여자의 데이터 및 훈련 품질을 고려하면서 동시에 공정성을 반영하여 훈련 속도 및 정확도 측면에서 우수함을 보여준다.

Ⅰ. Introduction

최근 산업, 의료, 제약 등 다양한 분야에서 AI 기술을 도입하기 위한 시도가 증가하고 있다. 허나, 일반적으로 이러한 도메인에서 도입된 단말 데이터는 설치 환경, 환자 진단 기록, 생산 노하우 등을 내포하고 있는 경우가 많아 데이터 프라이버시 및 보안이 중요하여 실 데이터는 장치 단말에만 한정되어 적재되고, 제한적으로 접근 및 활용되고 있어, 기존의 중앙 집중식 모델 훈련 방식과 같이 장치 단말의 적재 데이터가 로컬을 떠나 클라우드와 같은 환경으로 전송되고 학습되는 경우 프라이버시 문제가 발생할 수 있어 주의가 요구된다. 연합학습(Federated Learning)은 다수의 장치 단말에 적재된 분산된 로컬 데이터 환경에서 데이터 공유 없이 협력적으로 공통의 모델을 학습하는 기계학습 방법론이다. 이러한 방법론은 민감 데이터를 유출하지 않고 공통된 모델의 가중치 만을 상호교환하면서 공통의 모델을 협력적으로 학습하므로 데이터의 기밀성을 유지하면서 로컬 데이터의 특징을 학습할 수 있다는 점에서 다양한 응용분야로 확장될 수 있다[1]. 그러나 이러한 연합학습 기술을 실질적 환경에 적용하기 위해서는 극복해야 할 과제들이 여전히 남아있다. 먼저, 연합학습의 훈련 참여자는 다양한 환경에서 데이터를 편향적으로 수집하므로 클라이언트 다양성이 확보되지 않는 경우 전역 모델의 일반화 성능이 감소할 수 있다. 또, 서버는 참여자의 로컬 데이터에 대한 정보를 제한적으로 취득하므로 무작위 참여자의 무제한 접근을 수용하는 경우 악의적 참여자에 의해 전역 모델의 성능이 감소하거나 발산할 수 있다. 따라서 연합학습을 통한 공통 모델 훈련 시나리오에서 참여자의 지역화된 원본 데이터에 대한 정보 없이 고성능의 전역 모델을 훈련하기 위해서는 참여자의 훈련에 대한 기여를 평가하고 절차에 반영하여 훈련 절차를 조율하는 과정이 필수적이다.

최근 XAI(eXplainable AI)로 불리는 설명 가능한 인공지능 기법은 모델의 블랙박스적인 특성을 해소하고 모델의 예측 성능을 사람이 이해할 수 있도록 표현하는 기술로 모델 신뢰도 관점에서 다양한 분야에 적용가능해 크게 주목 받고 있다. 이러한 XAI 기술을 활용하면 모델의 동작 방식, 중요 입력 변수, 의사결정 논리, 은닉된 특징 정보에 따른 중요도를 식별할 수 있어 연합학습과 같이 제한된 클라이언트 정보를 기반으로 블라인드 모델 훈련을 수행하는 시나리오에 도입되는 경우 서버측 의사 결정에 큰 도움을 줄 수 있다. 이에 본 논문에서는 중앙 서버에서 XAI 기반으로 기여에 따른 클라이언트 선택을 조율하여 연합학습 모델 훈련 성능을 향상시키는 PCCS (Participant Contribution based Client Selection scheme)을 제안한다. 제안하는 방식은 연합학습 모델 훈련 시나리오에서 참여자의 보유 데이터 특성이 전역 모델에 기여하는 정도를 SHAP(SHapley Additive exPlanations) 기반으로 정량적으로 수치화하고 공정성을 반영한 기아 지수(Starvation Index)와 통합하여 훈련 참여자의 클라이언트 선택 우선순위를 책정해 참여자 선택에 부분적으로 반영 및 훈련한다. 이러한 방식은 기존의 기본 연합학습 훈련 시나리오와 달리, 무작위 참여자의 로컬 데이터 및 학습 품질을 훈련 절차에 반영하면서 동시에 공정성을 확보하므로, 동일한 훈련 환경에서 기본 훈련 체계 대비 전역모델의 훈련 속도 및 정확성이 향상됨을 확인할 수 있다.

Ⅱ. PCCS (Participants Contribution based Client Selection Scheme)

이 장에서는 PCCS의 동작 방식을 설명한다. PCCS는 무작위 훈련 참여 시나리오에서 연합학습 모델 훈련을 가속하고, 전역 모델의 일반화 성능을 향상시켜 고품질 AI 모델을 훈련하기 위해 고안되었다. 기존 연합학습의 기본 훈련 방식은 서버와 연결된 참여 클라이언트 집합에서 랜덤하게 일정 비율의 참여자를 선정하여 전역모델의 복사본과 훈련 수행 가이드를 제공한다. 선정 클라이언트는 전역모델의 복사본을 로컬 연산자원과 데이터를 활용해 수행 가이드를 기반으로 업데이트하고 서버 측으로 전송한다. 서버는 수신된 로컬 모델을 데이터 규모에 기반해 가중합하여 전역모델을 업데이트하고 위 절차를 반복해 모델을 훈련한다. 하지만 이러한 방식은 참여자의 편향된 데이터, 악의적 참여자의 데이터 증강 등에 의해 모델

이 발산할 수 있어 한계가 존재한다. 따라서 전역모델의 성능을 동일 훈련 반복에서 극대화 하기 위해 PCCS는 훈련 참여자의 학습 품질에 따른 전역모델의 기여도를 우선순위 방식에 통합한다. 그림 1은 PCCS의 시스템 구성도이다.

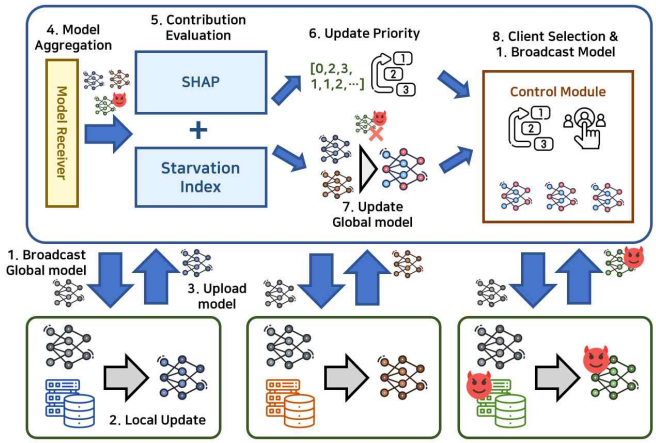

그림 1. PCCS 시스템 구조도

PCCS는 다음과 같이 동작한다. 먼저, 서버측 전역모델을 참여자가 수신하고 로컬 업데이트를 수행해 업로드 한다. 그 다음 서버는 수신된 로컬 모델을 집계하고 SHAP를 통해 참여자 조합에 따른 특정참여자의 평균 기여도 Φ_i 계산을 다음과 같이 수행한다.

$$\Phi_i = \sum_{S \subseteq P/\{i\}} \frac{|S|!(|P|-|S|-1)!}{|P|!}\left[F(S\cup\{i\})-F(S)\right] \quad (1)$$

여기서 P는 참여자 전체집합, $S \subseteq P/\{i\}$는 참여자 i를 제외한 부분집합, $F(S)$는 참여자가 협력하여 업데이트한 전역모델의 성능을 의미한다. 해당 값을 기반으로 일정 기여 이상을 수행한 참여자는 사용자 설정에 따라 가산된 우선순위를 차등적으로 부여받고 일정 기여 이하의 참여자는 우선순위를 감소시키며 동시에 전역모델 집계에서 제외한다. 한편, 전역모델의 훈련 데이터 다양성과 참여자 공정성을 확보하기 위해 기아 지수를 우선순위$P_i[t]$에 다음과 같이 반영한다.

$$P_i[t+1] = \begin{cases} P_i[t] + \alpha_t \times \Phi_i \,(\text{if } client\ is\ selected) \\ (P_i[t]+1) \qquad (Otherwise) \end{cases} \quad (2)$$

여기서 α_t는 사용자가 설정한 기여도 가중치이다. 이를 바탕으로 서버는 그 다음 전역 훈련에 참여할 클라이언트S_{t+1}를 다음과 같이 선정한다.

$$S_{t+1} = \left\{x \in S : x \in S_{priority} = \left\{x_1, \ldots, x_{\frac{n}{2}}\right\}\right\} \cup \left\{S \setminus S_{priority} : x \sim Uniform(S \setminus S_{priority}, \frac{n}{2})\right\} \quad (3)$$

여기서 S는 연결된 참여자 전체 집합이고, $S_{priority}$는 우선순위에 의해 정렬된 단위 훈련 참여자의 상위 절반이다. 이와 같은 클라이언트 선택 방법론을 통해 차기 전역반복의 훈련 참여 클라이언트 집합 S_{t+1}은 학습 품질에 따른 전역모델의 기여도와 공정성을 통합하고 참여자 다양성을 동시에 반영하여 선택되고 훈련될 수 있다.

Ⅲ. Simulation Result

이 장에서는 PCCS의 성능을 평가한다. 성능 평가 요소는 크게 2가지로 훈련 속도 및 훈련 정확도로 구성된다. 먼저, 훈련 속도는 100번의 전역반복을 수행할 때 얼마나 빨리 일정 정확도(최종 정확도의 80% 수준)에 도달하는 전역반복 횟수이다. 두 번째로 훈련 정확도는 테스트 데이터에 대한 일치 여부이다. 시뮬레이션을 위해 SMILES 화학구조를 통해 CYP2C19의 억제를 예측하는 CYP P450 2C19 억제 데이터를 활용한다. 각 클라이언트는 500개의 데이터를 가진 20개의 클라이언트로 구성되어 있으며 그 중 1개의 클라이언트는 악성 클라이언트로 동일데이터의 복사본을 500개 갖는다. 매 전역반복마다 서버는 5개의 클라이언트를 뽑고, 훈련 참여자의 식별, 호환성 검증과 로컬 데이터의 특성을 추출을 위한 클라이언트 유효성 검증 및 약물-타켓 분석/전처리 도구를 참여자에게 제공해 훈련한다. 이때 각 클라이언트는 5번의 로컬 업데이트를 수행하고 총 100번의 전역 반복을 수행한다.

그림 2는 연합학습 시나리오에 PCCS를 도입한 경우 발생하는 성능의 차이를 보여준다. 그림에서 알 수 있듯이 제안하는 방법은 로컬 클라이언트 데이터로 단순 훈련한 경우 및 FedAVG 방식과 비교했을 때 훈련 속도 측면에서는 각각 [6.46x, 3.23x] 빠르고, 훈련 정확도 측면에서는 각각 [16.72%, 11.86%] 더 높다.

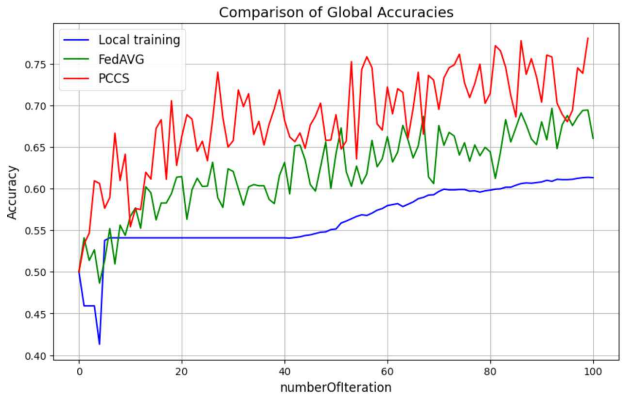

그림2. PCCS 훈련 성능 비교

이는 PCCS가 훈련 참여자의 훈련 품질에 따른 전역모델 기여를 산출하고 이를 참여자 선택 절차에 반영했기 때문이며 우선순위 기반으로 클라이언트의 공정한 참여를 통해 훈련 데이터 다양성을 확보하고, 악성 참여자를 배제하여 보다 안정적인 훈련을 수행했기 때문이다.

Ⅳ. Result

본 논문에서는 연합학습 시나리오에서 참여자 기여도 기반으로 클라이언트 선택을 수행해 전역모델 훈련 성능을 향상시키는 PCCS를 제안했다. 제안하는 방법은 참여자 기여도, 공정성, 다양성을 통합하여 클라이언트 선택에 반영하여 단일 클라이언트 훈련, 기본 연합학습 훈련 시나리오 대비 향상된 성능을 보여준다. 추후 연구에서는 참여자 데이터 특성을 고려한 참여자 맞춤형 훈련 시나리오를 통해 전역모델 훈련의 가속을 시도할 계획이다.

ACKNOWLEDGMENT

본 연구는 보건복지부 및 과학기술정보통신부의 재원으로 연합학습 기반 신약개발 가속화 프로젝트사업 지원에 의하여 이루어진 것임(과제 고유번호 : RS-2024-00459866)

참 고 문 헌

[1] Li, L., Fan, Y., Tse, M., & Lin, K. Y., A review of applications in federated learning. Computers & Industrial Engineering, 149, 106854, 2020.

[2] Veith, H., Southall, N., et al., Comprehensive characterization of cytochrome P450 isozyme selectivity across chemical libraries. Nature biotechnology, 27(11), 1050-1055, 2009.

[3] McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A., Communication-efficient learning of deep networks from decentralized data. In Artificial intelligence and statistics (pp. 1273-1282). PMLR, 2017.