# 연합학습에서의 공정성 간의 관계

유상윤[O1] 문수묵[1]
[1]서울대학교 전기정보공학부
sangyoonyu@snu.ac.kr, smoon@snu.ac.kr

# Relationship between Fairness in Federated Learning

Sangyoon Yu[O1] Soo-Mook Moon[1]
[1]Department of Electrical and Computer Engineering, Seoul National University

**요 약**

기계학습에서의 공정성 문제는 지속적으로 대두되어왔다. 최근에는 연합학습 환경에서의 공정성에 대한 연구들이 점차 활발히 진행되고 있다. 연합학습 환경에서는 기존 공정성에 대한 문제를 포함해서 새로운 공정성에 대한 필요성 또한 존재하는데, 본 연구에서는 연합학습 환경에서의 공정성 두 가지인 '클라이언트간 성능 분포의 공정성'과 '그룹 공정성'에 대해서 논의하고 '클라이언트간 성능 분포의 공정성'을 FedDyn을 방법을 적용한 q-FedDyn을 통해 개선하였다. 이후 두가지 공정성에 대한 q-FedDyn의 성능과 클라이언트들 간의 데이터 분포에 따른 차이점을 분석하였다.

## 1. 서 론

기계학습은 최근 발전을 통해 비약적인 발전을 거듭하고 있다. 최근에는 많은 데이터를 기반으로 학습하기 위해 개인들의 데이터를 기업들이 수집하는 문제가 발생한다. 이에 따라 개인정보보호와 탈중앙 분산컴퓨팅에 대한 관심이 증가해, 연합 학습(FL: Federated Learning)의 개념[1]과 중요성이 대두되고 있다. 기존 기계학습의 경우 중앙서버가 각 사용자들의 데이터를 수집해 모델을 학습하는 데 반해, 연학학습은 로컬의 모델만을 받아서 학습하기 때문에, 기존의 인공지능 학습이 가지던 데이터 프라이버시 침해문제를 해결할 수 있는 방법으로 여겨진다.

기계학습에서의 공정성은 지속적으로 다루어지고 있는 주제이다. 중앙화된 학습에서 공정한 학습을 하는 방법은 많이 제시되었지만 연합학습 환경에서는 기존 기계학습과 달리 새롭게 발생되는 불공정 문제들이 존재한다. 이는 연합학습의 경우 전체 데이터 분포를 보고 판단을 할 수 없고, 학습에 있어 모델에게 학습한 데이터 샘플 수에 비례하는 무게를 주고 글로벌 모델 학습을 진행하기 때문에 공정성 측면에서 불공정한 모델이 만들어질 수 있기 때문이다.

이에 본 연구에서는 연합학습과 공정성에 대해서 논의하고, 서로 다른 공정성 간의 관계 정리하고자 한다. 이후 기존 연구의 방법론을 개선한 방법론을 제시하고, 이를 통해 서로 다른 두 공정성을 모두 개선하는 상황과 이 때 개선된 방법론의 성능을 실험을 통해 확인하고자 한다.

* 이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No.2020R1A2B5B02001845)

## 2. 배경지식

### 2.1 연합학습

연합학습은 로컬 클라이언트들이 서버를 통해, 또는 서버를 통하지 않고 협력하여 데이터가 탈중앙화된 상황에서 글로벌 모델을 학습하는 기술이다[1]. 이때 로컬 클라이언트들로는 사물 인터넷 기기, 스마트폰 등의 엣지 디바이스들 또는 병원 등의 데이터를 보호해야되는 주체가 해당된다.

연합학습은 데이터 프라이버시 향상과 커뮤니케이션 효율성 측면에서 장점을 가진다. 데이터를 공유하지 않으며 학습해 차등 개인정보 보호(Differential Privacy)를 통해 데이터 프라이버시를 향상할 수 있다. 데이터 프라이버시가 향상되므로 연합학습을 통해 환자들의 임상 데이터와 같은 개인정보 보호가 굉장히 중요한 분야들에서 데이터 유출 없이 학습이 가능하다. 또한 기존 방식으로 서버에서 학습하기 위해서는 로컬 디바이스 데이터를 모두 서버로 전송해야 하므로 네트워크 트래픽과 스토리지 비용이 증가하는 반면, 연합학습에서는 클라이언트에서 학습 후 로컬 모델의 일부 업데이트 정보만을 전송하므로 커뮤니케이션과 스토리지 비용이 상당히 줄어들게 된다.

### 2.2 기계학습 공정성

기계학습에서의 공정성은 여러 가지를 개념에서 사용되고 있다[2]. 가장 많이 통용되는 개념은 그룹 공정성으로 형량을 결정하는 모델이 인종에 따라서 편향된 결정을 내리는 등의 경우를 그룹 공정성에 반하는 것으로 볼 수 있다[3]. 이외에도 '공정성'은 여러 가지 개념에 통용되는 표현으로, 연합학습 환경에서만 적용되는 공정성

**Algorithm 1** q-FedDyn

**Input** $K, T, q, 1/L, w^0, p_k, k = 1, ..., m$
**for** t = 1,2,...T **do**
  **for** k = selected clients **do**
    $\theta_k^t = arg\min_{\theta} L_k(\theta) - < \nabla L_k(\theta), \theta > + \frac{\alpha}{2}||\theta - \theta^{t-1}||^2$
    $\nabla L_k(\theta_k^t) = \nabla L_k(\theta_k^{t-1}) - \alpha(\theta_k^t - \theta^{t-1})$
    $\Delta\theta_k^t = L(\theta_k^t - \theta_k^{t-1})$
    $m_k^t = qL_k^{q-1}(\theta^{t-1})||\Delta\theta_k^{t-1}||^2 + LL_k^q(\theta^{t-1})$
  **end for**
  $\Delta_k^t = L_k^q(\theta_k^t)\Delta\theta_k^t$
  $h^t = h^{t-1} - \alpha(\Sigma\Delta\theta_k^t)$
  $\theta^t = \theta^{t-1} + \frac{\Sigma\Delta_k^t}{\Sigma m_k^t} - \frac{1}{\alpha}h^t$
**end for**

그림 1 q-FedDyn 알고리즘

의 정의가 존재한다. 이는 클라이언트 간 성능 분포의 공정성으로, 일부 클라이언트의 학습 참여를 저해시키는 등의 문제를 발생시킬 수 있다.

## 3. 연합학습과 공정성

### 3.1 클라이언트 간 성능 분포의 공정성

성능 분포의 공정성은 연합학습 환경에서 제안된 공정성이다[4]. 연합학습 환경에서는 클라이언트들의 각자 학습을 진행하고 전체 데이터에 대해서 높은 성능을 유지하는 것이 학습 목표이기 때문에, 전체 데이터에 대한 성능은 높더라도 각 클라이언트들의 데이터에 대한 성능은 불균일하게 분포될 수 있다. 이는 클라이언트들의 입장에서는 불공정하므로 학습 참여를 저해시킬 수도 있고, 특정 그룹에게 모델이 사용되지 못하는 문제점을 발생시킬 수 있다.

### 3.2 그룹 공정성

전통적으로 통용되는 공정성 개념을 기계학습에 적용한 개념이다. 연합학습 환경에서 또한 그룹 공정성 문제가 발생한다. 이는 연합학습 환경에서 특히 해결하기가 어려운데, 전체 데이터셋의 분포와 로컬 데이터의 분포가 달라 각 로컬에서 공정성 방법을 적용하여서 전체 모델이 공정해지지 않을 수 있기 때문이다. 예시로 학습 시 배치에 해당하는 데이터 샘플링을 그룹 별로 조절하는 방법을 들 수 있다[5]. 각 로컬에서 불공정한 그룹과 전체 데이터에서 불공정한 그룹이 다를 수 있기 때문에 단순히 적용하는 것에는 어려움이 있다. 연합학습 환경에서의 그룹 공정성은 최근 활발히 연구되고 있는 주제로, 기존 기계학습에서의 공정성 방법론과 차이가 있다.

### 3.3 공정성 간의 관계

클라이언트들이 그룹에 대하여 구분될 수 있는 경우, 두 공정성이 연관될 수 있음을 예상할 수 있다. 그룹 공정성이 훼손되는 이유는 다양하게 존재하지만, 그 중 한 가지는 특정 그룹에 대한 학습이 다른 그룹들에 비해서 모델에 덜 반영된 경우 발생한다. 이때 q-ffl을 적용하여 클라이언트들의 데이터, 즉 그룹의 데이터가 모델에 공정하게 반영되도록 하면 이를 해결할 수 있다.

| Seperate | Accuracy | ΔAcc |
|---|---|---|
| FedDyn | 0.866 | 0.070 |
| q-FedDyn(q=1) | 0.878 | **0.029** |
| q-FedDyn(q=2) | **0.887** | 0.041 |
| q-FedAvg(q=1) | 0.878 | 0.051 |

표 1 방법론별 성능(Accuracy)과 공정성(ΔEO), Seperate 한 데이터 분포

| IID | Accuracy | ΔAcc |
|---|---|---|
| FedDyn | **0.892** | **0.061** |
| q-FedDyn(q=1) | 0.883 | 0.067 |
| q-FedDyn(q=2) | 0.871 | 0.062 |
| q-FedAvg(q=1) | 0.873 | 0.065 |

표 2 방법론별 성능(Accuracy)과 공정성(ΔEO), IID 한 데이터 분포

## 4. q-FedDyn

### 4.1 q-FFL

q-ffl은 q를 조절하여 클라이언트 간 성능 분포의 공정성을 조절할 수 있는 방법론이다[4]. 학습 목표를 아래와 같이 재설정하여 클라이언트 간 성능 분포를 균일하게 조정한다. q를 높일 경우 공정성을 더 얻을 수 있고, q=0인 경우 일반적인 FedAvg의 학습 목표와 같아진다.

$$\min_w f_q(w) = \sum_{k=1}^{m} \frac{p_k}{q+1} F_k^{q+1}(w)$$

### 4.2 q-FedDyn

q-ffl의 경우 공정성에 맞게 학습 목표가 정의되어있으나 q-FedAvg가 q-FedSGD을 휴리스틱하게 변형한 것이기 때문에 이 학습 목표에 수렴하지 않을 수 있다. 이를 해결 하기 위해서 [6]에서 제안된 dynamic regularization을 q-ffl에 그림 1과 같이 적용하여 q-FedDyn을 제안하였다. q-FedDyn을 사용할 경우 통해 학습 목표에 학습이 수렴되어 성능과 공정성 모두 좋아지는 모습을 실험을 통해 확인하였다.

## 5. 실험

### 5.1 실험 설정

본 연구에서는 CelebA 데이터셋을 사용하였다[7]. Smiling을 타겟으로 학습을 시켰고, 공정성을 측정하기

위해서 성별을 보호 변수로 설정하였다. 학습의 난이도를 높이고, 데이터셋 자체의 불공정함을 증가시켜 방법론의 유효성에 대해서 비교 분석하기 위해서 전체 CelebA 데이터셋에서 여자 샘플은 50%, 남자 샘플은 5%를 남긴 후 실험을 진행하였다. 공정성 지표로는 Δ Acc(highest acc among groups - lowest acc among groups)를 사용해 측정하였다.

클라이언트의 데이터 분포의 경우 2가지로 나누어서 실험을 진행하였다. Seperate는 각 클라이언트가 남자 또는 여자 샘플만 들고 있는 상황이고, IID는 각 클라이언트가 전체 데이터 분포와 균일하게 데이터를 가지고 있는 상황을 의미한다.

100개의 클라이언트를 가정하였고, 학습 라운드 별로 10개의 클라이언트가 랜덤하게 학습에 참여한다. 실험 환경은 AMD Ryzen Threadripper 1950X(CPU), Nvidia 2080 super GPU 2개를 사용해 진행되었다.

### 5.2 q-FedDyn 성능

q-FedDyn이 q-FedAvg에 비교하여 성능이 높은 것을 표 1,2를 통해 확인할 수 있다. 추가적으로 공정성이 개선되는 Seperate한 분포에서는 공정성이 개선된 것을 표 2를 통해 확인할 수 있다. 이때 q가 클수록이 아닌, 적정 q에서 가장 공정한 것을 확인할 수 있었다.

### 5.2 공정성 간의 관계

Seperate 하게 데이터가 분포된 경우, 클라이언트들이 그룹을 대표하기 때문에 클라이언트 들간의 공정성과 그룹 공정성이 긍정적으로 연관되어 q-ffl을 적용할 경우 공정성이 개선된 모습을 표 1을 통해 확인할 수 있다. IID 하게 데이터가 분포된 경우, client들 간의 공정성이 그룹 공정성과 달라지기 때문에 공정성이 개선되지 않고 오히려 나빠지는 모습을 표 2를 통해 확인할 수 있다. 데이터 분포에 따라 두 공정성을 같이 해결되거나, Trade-Off 관계에 있는 모습을 실험을 통해 확인할 수 있다.

## 6. 결 론

본 연구에서는 공정성 간의 관계 대해서 다루고, 특정 상황에서 성능과 공정성 두 가지를 한 방법을 통해 모두 성취할 수 있음을 보였다. 이 때 사용할 수 있는 개선된 q-ffl인 q-FedDyn을 제안하였다. 이를 통해 앞으로 연합학습에서의 공정성을 고려할 때 두 가지 공정성의 상관관계를 데이터 분포에 따라서 달리 해야되는 것을 알 수 있다.

## 참 고 문 헌

[1] Communication-Efficient Learning of Deep Networks from Decentralized Data, Brendan McMahan. 2017
[2] Fairness in Machine Learning: A Survey, Simon Caton. 2020
[3] Machine bias: There’s software used across the country to predict future criminals. And its biased against blacks, Julia Angwin. 2016
[4] Fair Resource Allocation in Federated Learning, Tian Li. 2020
[5] FairBatch: A Batch Selection for Model Fairness, Yuji Roh. 2021
[6] Federated Learning Based on Dynamic Regularization, Durmus Alp Emre Acar. 2021
[7] Deep Learning Face Attributes in the Wild, Ziwei Liu. 2015