논문 21-46-11-20 The Journal of Korean Institute of Communications and Information Sciences '21-11 Vol.46 No.11
https://doi.org/10.7840/kics.2021.46.11.1941

통계적 이질성 문제 해결을 위한 데이터 분포 추정 기반 확률적 샘플링 기법을 적용한 연합학습 구현

김 선 욱◆, 이 현 수*, 방 준 일**, 홍 성 은**, 김 화 종°

Implementation of Federated Learning Using Probabilistic Sampling Techniques Based on Data Distribution Estimation to Solve Statistical Heterogeneity Problems

Seon Uk Kim◆, Hyeonsu Lee*, Junil Bang**, Sung Eon Hong**, Hwa Jong Kim°

요 약

연합학습 통계적 이질성이란 연합학습에 참여하는 다수의 사용자가 사용하는 디바이스, 동적 환경 및 시공간으로부터 수집된 데이터에서 IID(Independent Identically Distributed) 조건을 만족하지 못하고 불균형한 분포 특성(Non-Independent Identically Distributed)을 나타내는 것을 의미한다. 본 논문은 연합학습의 통계적 이질성 문제를 해결하기 위해 로컬 데이터 분포에 기반하여 글로벌 데이터 분포 추정하고, 확률적으로 데이터 샘플링을 수행하는 프로세스를 제안하고 직접 구현하여 성능을 비교한다. 로컬 데이터에 직접적인 접근 없이 로컬 데이터의 분포를 통해 전체 데이터의 분포를 추정하여 로컬 데이터의 분포를 조정한다. 공개된 연합학습 프레임워크에 프로세스 기능을 추가하는 형태로 구현하여 MNIST(Modified National Institute of Standards and Technology database) 데이터를 이용해 분류 모델을 학습시킨다. 일반 연합학습과 본 연구에서 제안한 샘플링 기법을 적용한 연합학습을 다양한 환경의 클라이언트에서 100라운드까지 수행한 후 성능을 비교한 결과, 평균 0.91의 Accuracy와 평균 0.98의 AUROC(Area Under the Receiver Operating Characteristic Curve)로 비슷한 수준의 성능을 보이지만 라운드당 약 9% 정도의 학습 시간 단축과 로컬 클라이언트 사이에서 발생하는 성능 이질성을 약 1.5% 감소시켰다.

키워드 : 연합학습, 클래스 불균형, 샘플링, 데이터 이질성
Key Words : Federated Learning, Class Imbalance, Sampling, Heterogeneous Data

ABSTRACT

The statistical heterogeneity means that data collected from devices, dynamic environments, time and space, used by a number of users participating in Federated Learning(FL) does not satisfy the IID (Independently Distributed) condition and shows an unbalanced distribution(Non-Independently Distributed). In this paper, we estimate global data distribution based on local data distribution, propose and implement a process that perform

※ 본 연구는 2019년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원(No. 2019007059, 시퀀스 데이터 분석 성능 향상을 위한 다차원 임베딩 기술 연구)을 받아 수행되었습니다.
◆ First Author : Kangwon Nat. University Department of Computer Science and Engineering, king950411@gmail.com, 학생회원
° Corresponding Author : Kangwon National University Department of Computer Science and Engineering, hjkim3@gmail.com, 종신회원
* Kangwon Nat. University Department of Computer Science and Engineering, Korea Pharmaceutical and Bio-Pharma Manufacturers Association Researh Scientist of KAICD(Korea AI Center for Drug Discovery and Development, lhs@kpbma.or.kr, 정회원
** Kangwon Nat. University Department of Computer Science and Engineering, tkfka965@gmail.com, 학생회원; sunkenh@gmail.com, 정회원
논문번호 : 202105-112-D-RE, Received May 15, 2021; Revised July 20, 2021; Accepted July 25, 2021

data sampling stochastically, and compare the performance to solve the statistical heterogeneity problem of FL. We estimate The distribution of total data through the distribution of local data without the direct access about local data. Then we adjust the distribution of local data. We implement process functions in the open-source framework and Train classification model using MNIST(Modified National Institute of Standards and Technology database) data. After experimenting basic FL and FL with sampling techniques proposed in this study which performed up to 100 rounds in various environments, we compare the performance. As a result, the accuracy of 0.89-0.91 and the accuracy of AUROC(Area Under the Receiver Operating Characteristic Curve) of 0.98-0.99 showed similar performance, but the learning time per round was reduced by about 9% The performance heterogeneity between local clients decreased by about 1.5%.

Ⅰ. 서 론

인공지능이 발전하기 위해서는 근본적으로 많은 양의 데이터가 요구된다. 따라서 다양한 방면에서 여러 사용자의 데이터를 수집해야 할 필요가 있다. 하지만, 최근 고도로 발달된 개인정보 식별 기술로 인해 데이터 수집 과정에서 개인정보보호와 관련된 갈등을 초래할 수 있는 문제점에 직면하게 된다. 이러한 문제점의 해결책으로 정보 사회에서 정보 주체의 권리를 보장할 수 있는 연합학습(Federated Learning) 기법이 제시되고 있다.

연합학습이란 개인정보보호 문제, 데이터 저장 공간 문제 등의 이유로 학습 데이터를 한 공간에 집중하여 사용하지 않고 여러 위치에 분산된 학습 데이터를 사용하여 각 장치에서 학습을 수행한 후 업데이트된 가중치를 공유하여 중앙에서 한 개의 모델을 지속적으로 업데이트하며 진행하는 기계학습의 새로운 접근 방식이다. 기존의 중앙 집중형 학습 방식과 달리, 개인 소유 데이터를 공유하지 않고 각 사용자가 직접 데이터를 처리하여 각 장치의 모델 즉, 로컬 모델을 훈련시킨다. 이를 기반으로 공유 모델(글로벌 모델) 학습에 대해 협업하여 더 우수한 모델을 만듦으로써 성능을 개선시킨다. 데이터를 소유하고 있는 사용자 간에 직접적으로 공유되는 데이터 없이 각 사용자의 로컬 모델 가중치를 공유하는 방식이기 때문에 기존 중앙 집중 형태의 모델에 존재하는 문제점을 해결할 수 있다는 장점이 존재한다. 또한, 공유모델을 공유받음으로써 각 장치에서 지니고 있지 않은 데이터를 활용한 학습의 효과도 얻을 수 있다.

본 연구는 Kim Seon Uk et al.[1]에서 연구 및 제안했던 데이터의 통계적 이질성 문제 해결을 위한 샘플링 기법을 바탕으로 실제 구현을 하였으며, 연합학습 프레임워크에 적용하여 전체적인 학습을 진행하였다.

연합학습 통계적 이질성이란 연합학습에 참여하는 다수의 사용자가 사용하는 디바이스, 동적 환경 및 시공간으로부터 수집된 데이터에서 IID(Independent Identically Distributed) 조건을 만족하지 못하고 불균형한 분포(Non-Independent Identically Distributed) 특성을 나타내는 것을 의미한다. 시공간적으로 동적인 환경에서 Non-IID 문제를 해결하기 위해서는 각 사용자의 다양한 디바이스 특징과 애플리케이션 요구사항을 최적화하여 반영해야 한다. 따라서, 연합학습의 글로벌 모델 뿐만 아니라 로컬 모델의 성능 및 학습 방법을 최적화할 수 있는 기술이 요구된다.

최근에는 통계적 이질성 문제를 해결하기 위해 글로벌 데이터 불균형, 로컬 데이터 불균형으로 나누어 문제를 분석하고 다양한 불균형 완화 기법을 적용하여 해결하는 연구가 활발히 진행되고 있다. 개인정보를 보호함과 동시에 클래스 불균형 완화를 위해 각 사용자의 데이터에 대한 직접적인 접근없이, 연합 학습을 사용함과 동시에 다양한 방법을 제안, 적용하여 통계적 이질성 문제를 완화하고자 했으며[2-4], 또 다른 연구에서는 데이터 증강을 통해 상대적으로 비율이 작은 분포의 클래스 데이터를 증가시킴으로써 통계적 이질성 문제를 해결하고자 했다[5]. 하지만 Major 데이터는 별도의 전처리 과정이 없기 때문에 모델에 그대로 적용되어 클래스 불균형 문제가 야기될 수 있다.

본 연구에서는 연합학습의 통계적 이질성 문제를 해결하기 위해 글로벌 및 로컬 데이터 분포를 추정하여 분포 정도에 따라 파라미터를 설정하여 데이터 증강 및 축소를 수행하는 프로세스를 제안하고, 직접 연합학습 프레임워크에 모델과 샘플링 기법을 구현하여 실험을 진행한다. 로컬과 서버 사이의 직접적인 데이터 공유없이, 로컬 데이터를 이용해 전체 데이터의 분포를 추정하여 추정한 결과만을 공유한다. 또한, 상이한 환경에서 이루어질 연합학습을 가정하여 로컬 학습을 진행할 클라이언트의 기기를 다양화하여 다양한 환경에서 진행되는 연합학습을 실험한다. 글로벌 분포

를 이용해 로컬에서는 데이터를 조정하고 학습을 진행하며, 벤치마크 데이터인 MNIST(Modified National Institute of Technology database) 데이터를 이용하여 분류 모델을 학습시키고 성능을 비교해 보았다. 본 연구에서 제안 및 구현한 샘플링 기법을 타 연구에도 활용한다면, 서로 다른 환경에서 연합학습에 참여한 각 로컬 클라이언트들로부터 수집되는 데이터로 인해 야기되는 클래스 불균형 문제를 최적화하고, 학습의 효율을 향상시킬 수 있을 것이다.

Ⅱ. 관련 연구

2.1 연합학습 프레임워크

연합학습은 로컬 데이터를 이용해 학습하는 다수의 클라이언트들과 해당 클라이언트들의 학습결과를 취합하는 서버로 구성되어 있으며, 학습의 각 라운드마다 학습결과를 송신/수신하는 통신이 발생한다. 서버와 클라이언트 사이에서 발생하는 통신 내용은 데이터를 공유하는 것이 아닌 로컬에서의 학습 결과인 가중치를 공유한다. 최근 다양한 연구에서 연합학습을 실험할 수 있는 프레임워크를 공개하고 있다[6-8].

Tensorflow Federated[6], IBM Federated Learning[7]의 프레임워크에서는 한 개의 기기에서 소스코드와 데이터를 가상으로 분리하여 여러 기기(서버 및 각 클라이언트)에서의 학습을 모방하여 실험을 진행하므로, 실제 통신이 이루어지지 않는다. 실제 통신이 없는 실험은 추후 연합학습을 이용해 실증 서비스 혹은 시스템을 구현할 때 문제가 발생할 수 있다.

Lee GH et al.의 연구[8]에서는 웹 프레임워크인 Django를 이용하여 웹을 기반으로 한 실제 통신을 통해 여러 기기에서 학습을 진행할 수 있는 연합학습 실험 프레임워크를 구현하였다. MNIST, MIMIC-III (Medical Information Mart for Intensive Care-III), ECG(Electrocardiogram)의 벤치마크 데이터를 이용해 학습시키고 성능을 비교하는 실험을 진행하였으며, 해당 프레임워크는 Github에 공개되어 있다[9]. 이 프레임워크는 가중치 전송, 가중치 수신, 라운드 정보 전송의 세가지 기능이 구현되어 있으며, REST API 기반의 통신을 이용해 GET, PUT의 HTTP Response를 통해 가중치와 라운드 정보를 송/수신한다. 본 연구에서는 Lee GH의 연구의 프레임워크 소스코드[9]를 참조하여 가중치 송/수신, 라운드 정보 수신 세 가지 기능 외에 로컬 데이터 분포를 전송하고, 글로벌 데이터 분포를 수신하는 2가지 새로운 기능을 추가하여 실제 다수의 기기에서 실험을 진행한다.

2.2 클래스 불균형 데이터 최적화

클래스 불균형은 데이터를 구성하고 있는 클래스마다 분포의 차이가 심화되는 경우를 의미한다. 이러한 불균형 문제는 인공지능 모델 학습 과정에서 전체 데이터 분포 중 적은 비율의 클래스 데이터(Minor data)가 전체 데이터 분포 중 큰 비율의 클래스 데이터(Major data)보다 학습 반영이 원활하게 이루어지지 않기 때문에 모델 성능 저하의 직접적인 원인이 될 수 있다. 한 개의 데이터 셋에서도 분포에 따라 특정 데이터의 양이 많고, 다른 특정 데이터의 양은 현저히 적은, 극단적인 비율의 격차가 있을 수 있으며, 다양한 환경에서 수집된 로컬 데이터의 경우에 이러한 클래스 불균형 문제가 심화될 수 있으며, 이는 성능에 큰 영향을 미친다. 특히, 데이터를 수집, 저장하는 로컬 클라이언트마다 상이한 환경을 가지는 연합학습에서 더욱 심화될 수 있으며, 로컬 데이터의 상이한 데이터 분포는 연합학습의 클라이언트마다 성능 차이가 심화될 수 있으며, 특정 클라이언트에서 성능이 저하되어 클라이언트 사이에서 성능의 격차가 심화되는 이질성(Heterogeneity) 문제가 발생할 수 있다. 결과적으로 이러한 이질성 문제는 전체 공유모델(글로벌 모델)의 성능 저하까지 이어지게 된다.

Lixu Wang et all.은 연합학습에서 발생하는 클래스 불균형을 해결하기 위해 로컬 데이터에 대한 클래스 불균형을 의미하는 Local Imbalance와 전체 데이터에 대한 클래스 불균형을 의미하는 Global Imbalance로 클래스 불균형을 정의하고, 이를 탐지, 최적화할 수 있도록 Ratio Loss라는 클래스 불균형에 따른 손실 함수를 제안하였고, 이를 통해 불균형한 클래스가 성능에 주는 좋지 않은 영향을 최소화하였다[2]. Miao Yang et al.은 다수의 클라이언트 중 라운드마다 학습에 참여할 클라이언트를 선택하여 데이터의 균형을 맞춤으로서 불균형한 클래스 문제를 최소화하였다[3]. Daliang Li, Junpu Wang은 특정 로컬 클라이언트에서는 높은 성능을, 다른 로컬 클라이언트에서는 낮은 성능을 나타내는 성능에서 발생하는 이질성 문제를 해결하고자 하였다[4]. 모두 접근 가능한 큰 공용 데이터와 로컬 클라이언트만 접근 가능한 작은 로컬 데이터로 나누어 서버에서 공용 데이터를 이용해 사전 학습(Pre-Train)을 완료한 글로벌 모델을 로컬 클라이언트로 전송해서 로컬 데이터를 이용해 클라이언트에 최적화함으로서 이질성 문제를 해결하고자 하였다. 하지만 데이터를 한 곳에 집중시켜야 하기 때문에 큰 공용 데이터를 필요로 하는 부분은 데이터 공유의 문제점이 있다.

2.3 취합 알고리즘

연합학습의 서버에서 클라이언트의 학습결과(가중치)를 취합하는 방법도 성능에 큰 영향을 끼친다. H. Brendan McMahan et al.은 모든 클라이언트의 학습결과를 평균화하는 Federated Averaging 알고리즘을 제안하였다[10]. Federated Averaging 알고리즘의 세부내용은 표 1과 같다. 각 라운드 t마다 학습에 참여한 m 개의 클라이언트는 각자 w^k를 로컬에서 학습하여 업데이트한 후 서버로 전송하고, 서버는 각 클라이언트로부터 받은 가중치 w^k의 평균을 계산하며 글로벌 가중치 w를 업데이트한다.

표 1. Federated Averaging 알고리즘
Table 1. Federated Averaging Algorithm

Federated Averaging Algorithm :
K Clients is participated, k is Client Number, B is the local batch size, E is the number of local epoch, η is learning rate

Server Calculate:
initialize w_0
for each round t =1,2, ... **do**
  $m \leftarrow$ maximize Available Clients at K
  $S_t \leftarrow$ random set of clients at m
  $n_k \leftarrow$ each client at S_t
  for each client $n_k \in S_t$ **in parallel do**
    $w_{t+1}^k \leftarrow \text{ClientUpdate}(k, w_t)$
    $w_{t+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{m} w_{t+1}^k$
ClientUpdate(k, w_t):
  for each local epoch i from 1 to E **do**
    for batch b $\in B$ **do**
      $w \leftarrow$ local train(b)
  return w to server

Ⅲ. 본 론

본 연구에서는 연합학습 과정에 불균형한 클래스를 조정하는 과정을 추가하여 불균형 문제를 해결하고자 한다. 전체 데이터의 분포를 추정하고, 로컬 데이터를 조정하여 학습 진행 전에 로컬 데이터에 대해 최적화를 진행함으로써 데이터의 클래스 불균형 문제로 인한 성능의 불균형 문제를 해결한다.

3.1 분포 추정 기반 확률적 샘플링 기법

연합학습에서 발생하는 클라이언트 사이의 성능 문제를 해결하기 위해서는 불균형 클래스 최적화가 필요하다. 그림 1에서는 로컬 데이터의 클래스 분포 취합을 통해 계산된 분포 조정 수치를 적용한 데이터 샘플링 프로세스를 설명한다. 클라이언트는 로컬로부터 수집된 데이터의 분포를 계산하여 서버로 전송한다. 서버는 전송받은 로컬 데이터의 분포를 통해 전체 데이터의 분포를 추정한다. 이 때, 분포 추정 계산은 연합학습에서 가중치를 취합하는 알고리즘과 동일한 방식을 채택하였다. 클라이언트는 전체 데이터 분포 추정의 결과값을 참조하여 로컬 데이터 분포를 조정할 파라미터를 계산한다. 추정 분포는 로컬 데이터의 분포 조정 기준이 될 파라미터로서, 클라이언트로 재전송되는 과정을 거친다.

클라이언트는 전달받은 분포 추정 파라미터를 통해 데이터 샘플링을 진행한다. Minor data는 Over-Sampling을 통해 데이터를 증강하고, Major data는 Under-Sampling을 통해 데이터를 축소시킴으로써, 로컬 데이터에 대한 클래스 분포를 조정한다. 여기서 로컬 데이터 조정에 대한 기준이 되는 파라미터는 데이터 특성에 맞게 설정한다. 예를 들어, 분류 모델 학습에 제안 기법을 적용한다면, 각 클래스별 데이터 분포를 추정하고, 회귀 모델 학습에 제안 기법을 적용한다면, 데이터의 평균, 분산 및 표준편차 등의 통계 분석을 통해 전체 데이터에 대한 분포를 추측한다. 본 연구에서 진행한 MNIST 데이터 분류 실험에서는 서버에서의 전체 데이터 분포 취합 과정에서는 Federated Averaging 알고리즘을 사용하였으며, 로컬 데이터의 각 숫자별 데이터의 분포를 통해 전체 데이터의 분포를 추정하고, 추정한 전체 데이터의 분포만큼 로컬 데이터의 분포를 조절한다.

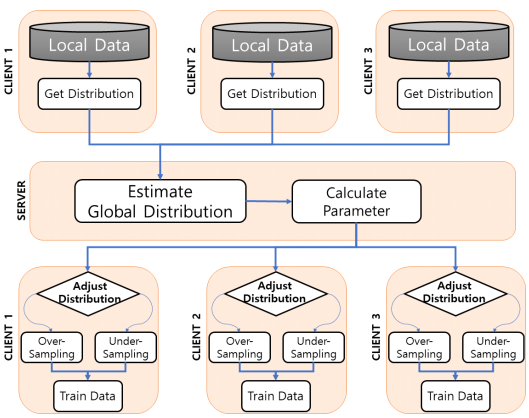

그림 1. 분포 추정을 통한 데이터 샘플링 프로세스
Fig. 1. Process of Sampling Data according to Estimating Distribution

3.2 샘플링 기법을 적용한 연합학습

본 연구에서 진행된 연합학습은 서버와 클라이언트가 일대다 관계로 구성되어 있고, 서버와 클라이언트는 같은 라운드동안 동일한 동작을 취하는 모델을 학습한다. 각 라운드는 샘플링 파트와 학습 파트로 구성되어 있으며, 라운드를 여러번 반복하며 모델을 업데이트해 나감으로서 학습이 진행된다.

그림 2에서는 본 연구에서 진행한 연합학습 각 라운드별 프로세스를 보여준다. 라운드가 시작되면 현재 라운드 정보를 클라이언트로 보내, 서버와 모든 클라이언트가 동일 선상에서 진행되는지 확인한다. 동일 라운드라면, 각 클라이언트에 있는 로컬 데이터의 분포를 계산하고, 서버로 분포를 전송한다. 서버에서는 모든 클라이언트의 로컬 데이터 분포 정보를 수신할 때까지 대기한다. 모든 로컬 클라이언트로부터 수신이 완료되면, 전체 데이터 분포(Global Distribution) 추정이 가능한 상태가 되는데, 이 때, 추정 분포를 통해 파라미터를 계산 및 설정한다. 설정된 파라미터는 로컬 데이터 분포를 조정할 때 사용된다. 로컬 데이터의 분포가 설정된 파라미터보다 낮다면, 파라미터에 해당하는 수치만큼 오버 샘플링을 수행하고, 높다면 파라미터에 해당하는 수치만큼 언더 샘플링을 수행한다. 이를 통해 균형 데이터 셋이 생성되며, 이후 모델 학습이 진행된다.

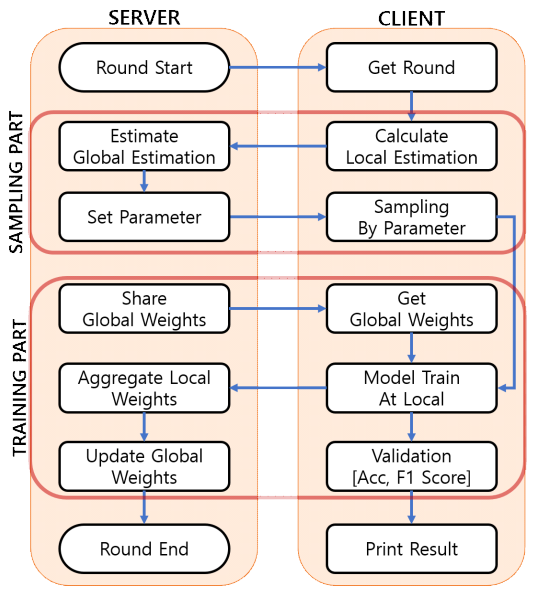

그림 2. 연합학습의 각 라운드별 프로세스
Fig. 2. Process of Federated Learning for Each Round

Ⅳ. 실 험

4.1 데이터 스플릿

연합학습에서는 클라이언트마다 데이터가 수집되는 환경 요인이 동일할 수 없으므로 크기, 분포가 모두 상이할 수 있다. 따라서, 본 연구의 실험에 사용할 MNIST 데이터를 각 숫자별 개수와 전체 데이터의 개수를 모두 랜덤하게 분할하여 각 로컬 클라이언트 기기에 할당한다. 70,000개의 MNIST 데이터 중 10,000개의 검증데이터, 60,000개의 학습 데이터로 나눈 후, 60,000개의 학습 데이터는 다시 각 클라이언트에 분할하여 저장한다. 각 클라이언트의 로컬 데이터는 100개에서 1500개 사이에서 랜덤하게 데이터를 할당하였으며, 숫자별 데이터 개수는 10개에서 150개 사이에서 랜덤하게 할당하여 상이한 분포를 띄고 불균형한 클래스를 가진 데이터를 가진 로컬 클라이언트를 가정하였다. 표 2는 랜덤하게 할당한 5개 클라이언트의 숫자별 데이터의 크기를 보여준다. 검증 데이터는 10,000개의 데이터 중 랜덤하게 8,000개를 추출하여 사용하였다.

모델 학습에서는 현재 서버에 저장되어 있는 이전 라운드에 대한 모든 클라이언트의 가중치를 취합하여 업데이트 된 글로벌 가중치를 수신하는 것으로 시작한다. 로컬 클라이언트는 수신된 글로벌 가중치를 적용하여 데이터 샘플링이 적용된 균형 데이터 셋으로 학습을 진행한다. 첫 라운드의 경우, 이전 글로벌 가중치가 존재하지 않으므로 적용하지 않고, 각 로컬 클라이언트의 모델 학습을 진행한다. 학습을 통해 얻은 가중치는 리스트의 형태로 저장하고, 검증 단계를 수행한다. 검증 결과는 JSON으로 변환되어 서버로 전송된다. 서버에서는 클라이언트의 각 가중치가 모두 수신될 때까지 대기하고, 수신이 완료되면 취합 알고리즘(Federated Averaging)을 통해 글로벌 가중치를 업데이트하고 서버에 저장한 후 라운드가 종료된다.

표 2. 클라이언트의 숫자별 MNIST 데이터 분포 예
Table 2. Example of MNIST Data Distribution for Number in each Client

CLIENT	0	1	2	3	4	5	6	7	8	9
Client 1	81	119	85	146	126	55	59	89	141	11
Client 2	135	45	105	120	23	87	11	14	85	44
Client 3	13	128	139	140	10	146	100	148	31	47
Client 4	132	63	149	35	37	89	32	22	142	23
Client 5	34	80	105	90	78	38	84	42	16	94
Total	395	435	583	531	274	415	286	315	415	219

로컬 클라이언트에서 목표 성능이 달성될 때까지 라운드를 반복하며 학습을 진행해 나가며, 목표 성능이 달성되면 서버로 학습 종료를 요청한다.

4.2 실험 환경

학습에 필요한 CPU와 GPU, 네트워크 상황 등의 상이한 환경요인을 가정하기 위해, 서버와 다수의 클라이언트는 각자 다른 기기에서 학습을 진행하였다. 또한, 성능 비교를 위해 본 연구에서 제안한 샘플링 기법을 적용한 연합학습과 적용하지 않은 연합학습, 그리고 연합학습이 아닌 일반 기계 학습을 통해 계산된 성능지표인 Accuracy와 F1 Score를 비교한다. 본 연구에서 제안한 샘플링 기법을 적용한 연합학습과 적용하지 않은 연합학습에서는 1개의 서버와 총 5개의 클라이언트가 학습에 참여하였다. 서버의 경우, CPU는 Intel사의 Xeon CPU E5-2630 v4를 사용했고, GPU는 Nvidia사의 GeForce GTX 1080Ti를 사용했으며, 리눅스 운영체제를 기반으로 한 환경에서 학습을 수행하였다. 클라이언트 1, 클라이언트 2, 클라이언트 3의 경우, CPU는 Intel사의 i7-6700K를 분할하여 사용했고, GPU는 GeForce GTX 1060 6GB를 분할하여 사용하였으며, 서버와 달리 윈도우 운영체제를 기반으로 한 환경에서 학습을 수행하였다. 클라이언트 3과 클라이언트 4의 경우, CPU는 Intel사의 8코어 9세대 i9 프로세서를 사용했고, GPU는 AMD사의 Radeon Pro 5500M을 사용했으며, Apple 사의 BigSur 운영체제를 기반으로 한 환경을 가진 기기 1개에서 학습을 진행하였다.

서버에서 Django를 통해 특정 포트를 Open하면, 각 클라이언트에서 서버의 Ip와 Open되어 있는 포트번호를 통해 접속하여 서버와 통신할 수 있다. 웹 기반으로서 통용되고 있는 Rest API기반 GET, PUT 통신(Http Response)을 사용해서 각자 가중치와 로컬 데이터의 분포와 같은 연합학습에 필요한 데이터를 송신, 수신한다.

로컬 데이터를 샘플링하기 위해 공유하는 데이터의 분포는 MNIST 데이터의 각 숫자별 데이터가 전체 데이터에서 차지하는 비중, 즉 로컬 데이터의 숫자별 분포를 공유하였으며, 서버에서는 로컬 데이터의 숫자별 분포를 Federated Averaging 알고리즘을 통해 취합하여 클라이언트로 전송한다.

학습에 사용된 모델은 1개의 입력 레이어와 128 유닛의 Rectified Linear Unit(ReLu) 활성함수를 사용하는 은닉 레이어, Softmax를 통해 결과를 출력하는 출력 레이어로 이루어진 단순한 신경망 모델이며, 최적화 함수로는 Stochastic Gradient Descent를 사용했고, 손실 함수는 Sparse Categorical Crossentropy를 사용했다.

4.3 실험 결과

클라이언트마다 각자 다른 검증 데이터를 이용해서 성능을 측정하므로 다른 값이 계산된다. 5개의 클라이언트와 1개의 서버를 사용하여 연합학습을 진행한 결과, 기존 연합학습의 경우, 클라이언트마다 0.88에서 0.92 사이의 정확도를, 0.98에서 0.99 사이의 AUROC를 나타내며, 100라운드를 진행하는 데에 약 2,260초가 소요되었으며, 학습을 1라운드 진행하는 데에 16-29초, 평균 22.6초 소요되었고, 라운드 정보 수신과 로컬 가중치 전송, 글로벌 가중치 수신의 3번의 통신이 발생하였다. 본 연구에서 제안한 샘플링 기법을 적용한 연합학습의 경우, 클라이언트마다 0.89에서 0.91 사이의 정확도를, 0.98에서 0.99 사이의 AUROC를 나타내었으며, 100라운드를 진행하는 데에 2,077초가 소요되었으며, 학습을 1라운드 진행하는 데에 14-26초, 평균 20.7초 소요되었고, 라운드 정보 수신, 로컬 데이터 분포 전송, 글로벌 분포 파라미터 수신, 로컬 가중치 전송, 글로벌 가중치 수신으로 총 5번의 통신이 발생하였다.

본 연구에서 제안한 샘플링 기법을 적용한 연합학습(Sampled FL)을 사용해 학습을 진행했을 때, 라운드마다 약 9% 정도의 시간 절약을 보였다. 서버와 클라이언트간 통신 횟수는 더 많지만, 시간이 절약된 것으로 보아 학습의 효율성이 개선되었다고 볼 수 있다. 또한, 클라이언트마다 다른 성능을 보이는 이질성(Heterogeneity) 역시 소폭 개선되었음을 확인하였다.

표 3. Basic FL과 Sampled FL의 실험결과
Table 3. Result of Experiment with Basic FL and Sampled FL

	Basic FL	Sampled FL
Time	22.6s/round	20.6s/round
Accuracy	0.88-0.92	0.88-0.91
AUROC	0.98-0.99	0.98-0.99
Heterogeneity	0.04	0.03

4.4 추가 실험

클라이언트의 개수가 성능에 미치는 영향을 실험해 보기 위해 클라이언트의 개수를 다르게 하여 연합 학습을 진행한 후 성능을 비교해 보았다. 클라이언트가 2개일 경우와 8개일 경우에서 추가 실험을 진행하였

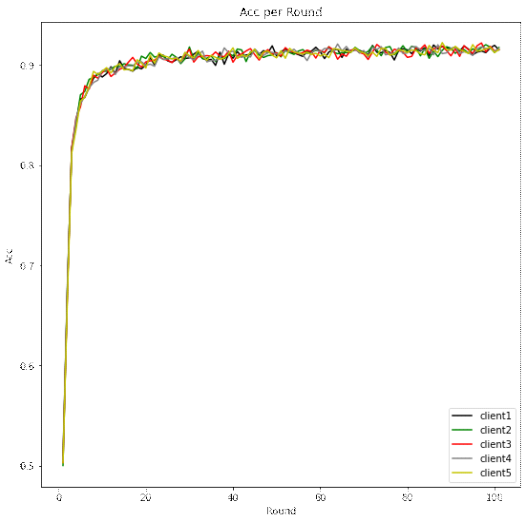

그림 3. 라운드 진행에 따른 5개의 클라이언트별 Accuracy
Fig. 3. Accuracy by 5 Clients according to Round progress

으며, 전체 데이터의 양은 동일하게 유지하면서 각 클라이언트에 할당된 데이터의 개수를 다르게 하였다. 서버는 모든 경우가 같은 환경에서 실험을 진행하였고, 클라이언트는 위 실험과 조금 다른 환경에서 진행되었다. 2개일 경우의 실험은 위의 일부 클라이언트의 환경과 동일하게 Window 운영체제 기반의 Intel사의 i7-6700K를 분할하여 사용했고, GPU는 GeForce GTX 1060 6GB를 분할하여 사용하였다. 8개일 경우에는 기존 5개 클라이언트의 환경에 1개 기기를 추가하였으며, CPU는 Intel사의 8코어 9세대 i9 프로세서를 사용했고, GPU는 AMD사의 Radeon Pro 5500M을 사용했으며, Apple 사의 BigSur 운영체제를 기반으로 구성되어 있다.

전체 데이터의 양은 동일하게 유지하고, 클라이언트의 개수만을 다르게하여 실험을 진행한 결과, Accuracy, AUROC는 각각 0.88-0.92와 0.98-0.99로 모두 비슷한 수준이다. 그러나, 학습에 소모된 시간에서 차이가 발생하였다. 클라이언트가 2개일 때는 라운드당 평균 15.34초, 총 1,534초, 8개일 때는 라운드당 평균 25.4초, 총 2,540초로, 클라이언트의 개수가 증가함에 따라 학습 시간도 증가함을 확인하였다. 또한 성능에서의 차이가 없는 것으로 보아, 클라이언트의 개수는 연합학습 모델의 성능에 직접적인 영향은 없는 것을 알 수 있다. 하지만, 전체 데이터의 양은 항상 고정되어 있지 않기 때문에 실제 데이터 수집과 학습, 검증이 동시에 이루어지는 연합학습의 특성상, 데이터의 특징에 따라 결과가 다를 수 있으므로, 다른 데이터에서도 성능에 영향이 없는지에 대해 추가 연구가 필요하다.

V. 결 론

본 논문에서는 연합학습에서 발생하는 문제 중 하나인 통계적 이질성 문제의 원인이 되는 클래스 불균형 문제를 해결하는 방안으로 데이터의 직접적인 접근 없이 확률적으로 분포를 추정을 통한 데이터 샘플링을 제안 및 구현하였고, MNIST 데이터를 이용해 분류 모델 학습을 실험하였다. 9%의 학습 시간 절약과 클라이언트 사이의 성능차이를 1%만큼 감소시켜 성능에서의 이질성 문제를 개선하였다. 또한, 클라이언트의 개수가 증가함에 따라 학습 시간이 증가하는 것을 확인함으로서 다른 환경에서 진행되는 연합 학습에 미치는 영향도 확인하였다.

데이터마다 다른 특징을 가지고 있으므로 추정된 전체 분포를 이용하여 서버에서의 효과적인 파라미터 계산법에 있어 추가적인 연구가 필요하다. 예를 들어, 본 연구의 이미지 분류 데이터와는 달리 변수를 예측하는 회귀 모델에 사용될 다른 통계적 특징을 띄는 데이터에서 동일한 파라미터 계산이 효과가 있을지 추가 실험이 필요하다.

서버에서의 파라미터 계산에 사용되는 로컬 데이터의 통계적 분포 또한 본 연구에서 사용된 각 클래스별 데이터의 분포 외에도 다양한 통계적 특징을 띄는 정보를 공유, 취합하여, 다양한 모델에 적용할 수 있다. 또한, 추후 연구에서는 의료 분야에 적용하기 위해 Physio Net 데이터를 사용하여 의료 데이터에 적용하여 학습을 진행해보고, 데이터로부터 추출되는 벡터를 다차원 임베딩을 적용하여 학습을 강화할 수 있다. 클라이언트 개수에 따른 연합학습 성능의 비교 또한 추가적인 실험을 진행해볼 예정이다. 클래스 불균형 문제가 포함된 연합학습 환경에서의 실험을 통해 샘플링 기법은 학습 시간 절약과 이질성 감소 효과를 확인하였다. 추후 연구에서 더욱 구체적인 실험을 통해 연구에 사용된 소스코드를 고도화할 예정이며, 연합학습에 참여하는 로컬 클라이언트들의 다양한 환경에서 수집되는 데이터들로 야기되는 클래스 불균형 문제를 해결할 수 있을 것이라 기대한다.

References

[1] S. U. Kim and H. J. Kim, “A study on probabilistic sampling techniques based on

data distribution estimation to solve federated learning statistical heterogeneity problems," in *Proc. Symp. KICS*, pp. 771-772, Yong Pyeong, Korea, Feb. 2021.

[2] L. Wang, S. Xu, X. Wang, and Q. Zhu, "Addressing class imbalance in federated learning," *AAAI 2021*, Aug. 2020.

[3] M. Yang, A. Wong, H. Zhu, H. Wang, and H. Qian, "Federated learning with class imbalance reduction," Retrived May., 9, 2021, from https://arxiv.org/abs/2011.11266, Nov. 2020.

[4] D. Li and J. Wang, "FedMD : Heterogenous federated learning via model distillation," Retrived May., 9, 2021, from https://arxiv.org/abs/1910.03581, Oct. 2019.

[5] S.-J. Hahn and J. Lee, "Privacy-preserving federated bayesian learning of a generative model for imbalanced classification of clinical data," Retrived May., 9, 2021, from https://arxiv.org/abs/1910.08489, Nov. 2019.

[6] *TensorFlow Federated API*, Google, Retrived May., 9, 2021, from https:// www.tensorflow.org/federated?hl=ko.

[7] *IBM Federated Learning Library*, IBM, Retrived May., 9, 2021, from https://ithub.com/IBM/federated-learning-lib.

[8] G. H. Lee and S. Y. Shin, "Federated learning on clinical benchmark data: Performance assessment," *J. Med. Internet Res.*, vol. 20, no. 10, Oct. 2020.

[9] G. H. Lee, *FL_Server*(2019), Retrived May., 9, 2021, from https://github.com/nanara1119/FL_Server

[10] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th AISTATS 2017. JMLR: W&CP*, vol. 54, Feb. 2017.

김 선 욱 (Seon Uk Kim)

2020년 2월 : 강원대학교 컴퓨터공학과 학사 졸업

2020년 3월~현재 : 강원대학교 컴퓨터공학과 석사과정

<관심분야> 연합학습, 딥러닝, 엣지 컴퓨팅, 데이터 불균형

[ORCID:0000-0002-4454-5462]

이 현 수 (Hyeonsu Lee)

2019년 8월 : 강원대학교 컴퓨터공학과 학사 졸업

2021년 2월 : 강원대학교 컴퓨터공학과 석사 졸업

2021년~현재 : 한국제약바이오협회 인공지능 신약 개발지원센터 연구원

<관심분야> 연합학습, 딥러닝, 머신러닝

[ORCID:0000-0001-9186-2762]

방 준 일 (Junil Bang)

2018년 : 강원대학교 컴퓨터공학과 학사 졸업

2020년 : 강원대학교 컴퓨터공학과 석사 졸업

2020년~현재 : 강원대학교 컴퓨터공학과 박사 재학 중

<관심분야> 딥러닝, 인공지능, 연합학습

[ORCID:0000-0003-0582-1572]

홍 성 은 (Seong Eon Hong)

2013년 : 강원대학교 컴퓨터공학과 학사 졸업

2015년 : 강원대학교 컴퓨터공학과 석사 졸업

2019년~현재 : 강원대학교 컴퓨터공학과 박사 수료 후 연구원

<관심분야> 딥러닝, 엣지 컴퓨팅, 연합학습

[ORCID:0000-0002-7469-2439]

김 화 종 (Hwa Jong Kim)

1984년 : KAIST 전기및전자공학과 공학석사 졸업
1988년 : KAIST 전기및전자공학과 공학박사 졸업
1988년~현재 : 강원대학교 컴퓨터공학과 정교수
2020년~현재 : 한국제약바이오협회 인공지능 신약 개발 지원센터 센터장

<관심분야> 데이터 통신, 컴퓨터 네트워크, 네트워크 프로그래밍, 빅데이터

[ORCID:0000-0002-3822-390X]