Journal of KIIT. Vol. 19, No. 1, pp. 43-54, Jan. 31, 2021. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2021.19.1.43

클래스 불균형 문제에 연합학습 적용을 위한 최적화 기법 연구

이현수*, 홍성은**[1], 방준일**[2], 김화종***

Study of Optimization Techniques to Apply Federated Learning on Class Imbalance Problems

Hyeonsu Lee*, Seongeun Hong**[1], Junil Bang**[2], and Hwajong Kim***

이 논문은 2018년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임(No. 2018-0-00261) 또한 2019년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2019007059)

요 약

고도로 발달된 개인 정보 식별 기술에 의해 개인에 대한 식별이 용이해지면서, 정보 사회에서 정보 주체의 권리를 보장할 수 있는 다양한 방안이 요구되고 있다. 연합학습은 이러한 요구에 의해 제안된 기계학습 방식으로 데이터를 비공개로 유지하면서 기계학습 알고리즘을 훈련하기 위한 특정 접근 방식이다. 본 논문에서는 개인정보보호 이슈에 민감한 의료 산업에 연합학습을 적용할 때 발생할 가능성이 있는 문제점을 파악하기 위해 망막 환자 데이터셋을 실제 의료기관이 데이터를 보유하고 있는 환경처럼 데이터 분포를 불균형하게 분할했다. 여기서 발생하는 클래스 불균형 문제에 다양한 학습 최적화 기법을 적용한 실험을 진행한 결과, 언더 샘플링 및 TopkAvg 기법을 적용한 실험에서 F1 score 0.96을 달성했으며, 학습 시간도 단축시켰다.

Abstract

Recently, as highly advanced personal identification technology has made it easier to identify individuals, various measures are required to guarantee the rights of information subjects in the information society. Federated learning is a machine learning approach proposed by these needs, a specific approach to educating machine learning algorithms while keeping the data private. In this paper, in order to identify problems that may arise when applying federated learning to the medical industry, which is sensitive to privacy issues, a retinal patient data set, was disproportionately distributed like the environment in which the actual medical institution holds the data. As a result of experiments applying various learning optimization techniques to class imbalance problems that occur here, F1 score 0.96 was achieved in experiments with under sampling and TopkAvg techniques, and the learning time was also shortened.

Keywords
imbalance data, federated learning, sampling, optimization

* 강원대학교 컴퓨터정보통신공학과 석사과정
- ORCID: https://orcid.org/0000-0001-9186-2762
** 강원대학교 컴퓨터정보통신공학과 박사과정
- ORCID[1]: https://orcid.org/0000-0002-7469-2439
- ORCID[2]: https://orcid.org/0000-0003-0582-1572
*** 강원대학교 컴퓨터정보통신공학과 교수(교신저자)
- ORCID: https://orcid.org/0000-0002-3822-390X

· Received: Nov. 29, 2020, Revised: Jan. 18, 2021, Accepted: Jan. 21, 2021
· Corresponding Author: Hwa-Jong Kim
Dept. of Computer and Communications Engineering, Kangwon National Univ., Gangwondaehak-gil 1, Chuncheon-si, Gangwon-do, Korea
Tel.: +82-33-250-6323, Email: hjkim3@gmail..com

Ⅰ. 서 론

인공지능은 본질적으로 성능 향상을 위해 다양하고 많은 양의 데이터셋을 필요로 한다. 이러한 요구사항을 충족시키기 위해서는 다양한 기기 또는 기관에서 발생하는 데이터를 수집해야 하는데, 이 과정에서 개인 정보 보호 문제가 발생한다. 특히, 고도로 발달된 개인 식별 기술에 의해 개인에 대한 식별이 용이해지면서 정보 사회에서 정보 주체의 권리를 보장할 수 있는 다양한 방안이 요구되고 있다.

연합학습(Federated learning)은 데이터를 비공개로 유지하면서 기계학습 알고리즘을 훈련하기 위한 특정 접근 방식이다[1]. 각각 고유한 로컬 및 데이터를 보유하고 있는 여러 분산 장치 또는 서버에서 기계학습 알고리즘을 훈련하는 것을 목표로 함으로써, 데이터 수집 과정에서 발생하는 개인 정보 보호 문제를 해결할 수 있다는 장점이 존재한다[1]. 이러한 연합학습의 장점은 의료기관에서 극대화된다.

일반적으로 헬스 케어, 질병 분류 알고리즘의 경우, 많은 양의 데이터를 필요로 한다. 하지만 단일 의료기관은 소속된 지역의 환자 인구 통계, 사용된 도구 또는 임상 전문화에 의해 데이터 자체가 편향될 수 있다[2][3]. 따라서 다른 의료기관의 데이터도 함께 수집하여 데이터를 다양화함으로써 편향성 문제를 해결해야 한다. 그러나 의료기관의 데이터는 환자의 개인 정보이므로 환자의 개인 정보 활용 동의와 윤리적 승인 없이는 사용 및 공유가 어렵다는 단점을 지니고 있다[4]. 연합학습은 이러한 민감한 개인 정보를 기관 간에 직접 공유할 필요 없이 수회에 걸친 반복 학습 과정에서 단일 기관의 모델보다 더 많은 데이터를 학습한 공유 모델을 얻을 수 있다[2][3][5].

의료 관련 데이터는 일반적으로 클래스 간 데이터 분포가 균일하지 않아 클래스 불균형 문제를 쉽게 마주하게 된다. 의료 분야에서 찾고자하는 데이터의 타겟 수는 매우 극소수인 케이스가 많다. 이러한 불균형 데이터셋에서는 정확도(Accuracy)가 높더라도 재현율 또는 민감도(Recall)는 급격히 작아지는 현상이 발생한다. 예를 들어 1000개의 망막 이미지 중 5개가 질병과 관련된 이미지라면, 모든 데이터를 정상으로 예측해도 정확도가 99.5% 가 나오기 때문이다. 따라서 임상 의사결정 지원 시스템처럼 의료 서비스에 인공지능을 접목하고자 한다면 이와 같은 클래스 불균형 문제를 해결할 필요가 있다.

본 논문은 의료 산업에 연합학습을 적용할 때 발생할 가능성이 있는 문제점을 파악하기 위해 데이터 불균형 문제를 해결하기 위한 관련 연구를 소개했다. 그리고 해당 연구를 의료 산업 문제에 적용할 때 발생할 수 있는 문제점을 분석하면서 본 연구의 필요성과 제안 방법을 기술했다. 제안 방법을 적용하기 위해 실험 설계 단계에서 Retinal optical coherence tomography images dataset(Retinal OCT Images)[6]을 실제 각 의료기관이 데이터를 보유하고 있는 환경처럼 데이터 분포를 불균형하게 분할하여 진행했다. 해당 단계로 인해 발생한 클래스 불균형 문제를 해결하기 위해 제안 기법을 적용하여 기존 방법과 비교함으로써 연합학습에 최적화된 기법을 탐색했다. 추가로 기존 연합학습 모델 가중치 업데이트 방식인 전체 모델 가중치 평균 계산(FedAvg) 대신에 로컬 학습 모델 중 성능이 전체 로컬 모델 성능 평균보다 우수한 모델들의 가중치만 취하여 평균을 계산하는 방식(TopkAvg)을 제안하여 성능을 비교하는 실험을 진행하였고, 이에 대한 결과를 기술했다.

Ⅱ. 관련 연구 및 제안 방법

2.1 관련 연구

[1]의 연구에서는 인공지능 모델 학습에 필요한 데이터 수집 과정에서 발생하는 개인 정보 보호 문제 등을 해결하기 위해 연합학습 프레임워크를 제안했다. 연합학습이란 그림 1과 같이 훈련 데이터를 한 곳에 집중하여 사용하지 않고 여러 위치에 분산하여 기계학습을 수행하는 접근 방식이다. 기존 기계학습 모델(중앙 집중형 모델)과 달리, 데이터 소유 권한이 있는 사용자가 직접 데이터를 처리하여 모델을 훈련시키는 방식으로, Party라고 칭하는 연합학습에 참가하는 모델들의 학습 가능한 가중치를, Aggregator라고 칭하는 서버에서 집계하여 더 우수한 모델을 생성 및 배포하는 과정을 통해 학습을 진행한다.

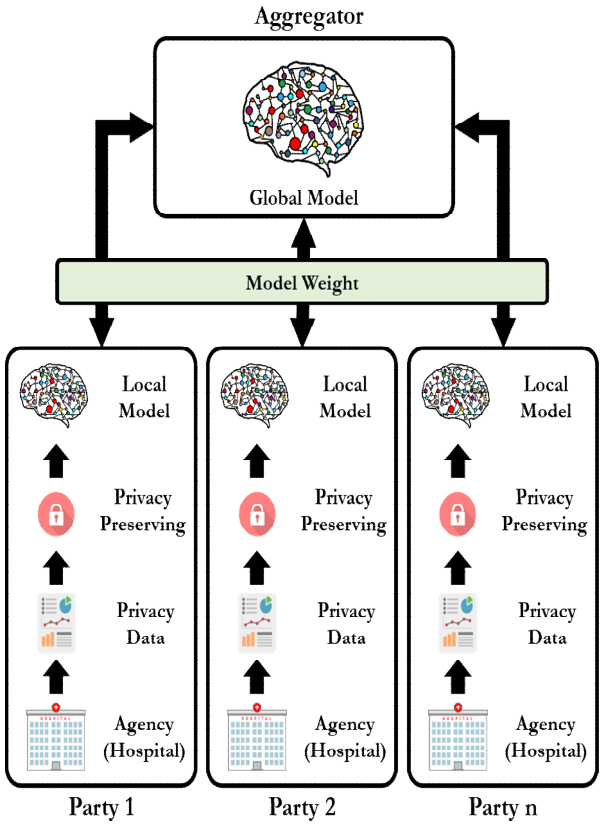

그림 1. 연합학습 프로세스
Fig. 1. Federated learning process

서로 다른 사용자가 소유하고 있는 데이터를 직접 공유하지 않고 모델의 학습 가능한 가중치 또는 가중치를 공유하는 방식이기 때문에 개인 정보 보호 문제를 해결하면서 상대적으로 적은 데이터로 최적화된 모델을 개발할 수 있다는 장점을 지니고 있다. 또한 각 참가자의 데이터 분포와 데이터셋의 크기에 대한 정의가 존재하지 않기 때문에 다양한 데이터 크기, 분포를 가진 참가자들이 공유 모델에 유연하게 접근함으로써 일반화된 모델을 구축하는 데 도움이 된다.

연합학습은 사용되는 기계학습 기술, 데이터 유형 및 운영 컨텍스트에 따라 다른 전략이 선호된다. 공유 모델의 가중치를 갱신하기 위해 다양한 방법이 연구되었으며, 가장 보편적으로 사용하고 있는 방법이 연합 평균 알고리즘이다.

연합 평균 알고리즘은 Aggergator의 모델(글로벌 모델) 가중치를 업데이트 하기 위해 Party의 모델(로컬 모델)이 훈련한 신경망 가중치를 정기적으로 평균화하는 것으로 구성된다[7]. 평균화된 모델 가중치는 추가 학습을 위해 로컬 모델의 업데이트로 사용된다. 각 로컬 데이터를 통한 학습은 그림 1과 같이 글로벌 모델이 업데이트 됨에 따라 모든 참가자에게 점진적으로 공유된다.

하지만 연합 평균 알고리즘은 불균형한 분포를 띄고 있는 데이터셋에 대해서 심각한 성능 저하를 나타낸다[8]. 이는 연합 학습 모델에 참가하는 Party 간의 데이터 균형 격차가 벌어질 경우, 가중치 격차도 함께 증가한 결과라고 할 수 있다.

이에 따라, 연합학습에서 발생할 수 있는 불균형 데이터 문제를 해소하기 위한 다양한 연구가 활발히 진행되고 있다.

[8]의 연구에서는 공유 가능한 균형 데이터셋을 모든 Party에 할당하는 방식을 제안했다. 해당 접근법은 실제로 불균형한 데이터셋에 대해 더 정확한 모델을 생성할 수 있지만, 여러 가지 단점이 존재한다.

먼저, 공유 데이터셋에 대한 과적합 문제다. 공유 데이터셋이 존재하는 경우, 데이터 보유량이 적은 Party에 대해선 해당 Party가 가지고 있는 데이터에 대한 학습이 진행되지 않고, 공유 데이터셋에 대해서만 학습이 진행되기 때문에 과적합 현상이 발생할 수 있다.

두 번째로, 공유 데이터셋의 존재는 연합학습의 목적에 부합하지 않다. 연합 학습 모델은 서로 다른 Party가 소유하고 있는 데이터를 직접 공유하지 않고 모델의 학습 가능한 가중치 또는 가중치를 공유하는 방식으로 개인 정보 보호 문제를 해결한다. 하지만, 공유 데이터셋은 이러한 개인 정보 보호 문제를 야기할 뿐만 아니라, 의료 데이터같이 외부로의 유출이 어려운 상황에서는 생성 자체가 불가능할 수 있다.

따라서 본 연구에서는 의료 데이터처럼 데이터의 불균형, 개인 정보 활용 동의 및 윤리적 승인 등 기관 간에 데이터 사용 및 공유가 어려운 환경을 실험 환경으로 설정하고, 여기에 연합학습을 최적화하여 적용하기 위해 다양한 클래스 불균형 해소 기법들을 비교 실험하였다.

[9]의 연구에서는 Aggregator와 Party 사이에

Mediator 개념을 도입하여 클래스 불균형 문제를 해소하는 Astraea 프레임워크를 제안했다. Astraea 프레임워크에서 각 Party가 가지고 있는 데이터는 전체 데이터 클래스 분포 비율에 따라 샘플에 대한 무작위 시프트, 회전 등을 통해 병렬로 증강되어 훈련에 사용된다. 모든 Party의 데이터 증강이 완료되면 Mediator에 데이터 분포가 균일하게 되도록 일정량의 Party를 할당하여 부분학습을 수행한다. 최종적으로 모든 Mediator에 대한 FedAvg를 수행하여 가중치를 수행하는 방식으로 진행되며, 기존 FedAvg와 비교했을 때 EMNIST, CINIC-10 불균형 데이터 셋에 대한 정확도가 각각 5.59%, 5.89% 향상되었다.

하지만 데이터 증강 과정에서 Aggregator는 원시 데이터에 접근하여 전체 Party가 가지고 있는 데이터의 분포 비율 계산을 통해 각 클래스 샘플에 대한 무작위 생성을 함으로써 개인정보보호 문제를 야기할 수 있다는 점과 전체 데이터에서 차지하는 특정 클래스의 비율이 극히 작다면 무작위 생성의 기준이 되는 샘플과 유사한 데이터가 과도하게 생성되어 과적합이 발생할 수 있다는 단점이 존재한다.

따라서 본 연구에서는 원시 데이터 접근으로 발생하는 개인정보보호 문제와 과도한 데이터 증강으로 인한 과적합 문제를 피하기 위해 클래스 불균형 문제를 해결하는 과정에서 오버 샘플링 기법 대신 언더 샘플링 기법을 채택하였으며, 실험에서 두 기법 간의 성능 차이를 비교 서술하고자 한다.

[10], [11], [12]의 연구에서는 연합학습에서 각 Party의 원시 데이터에 대한 접근 없이 클래스 분포를 추정하기 위한 모델을 설계하여 개인정보보호 문제를 해결하고, 클래스 불균형으로 인한 모델 성능 저하를 완화했다.

[10]은 연합학습에 참가하는 각 Party 모델 가중치의 사후 분포를 추정하고, 이를 통해 훈련 데이터에서 적은 비율을 차지하는 클래스 데이터를 오토인코더를 사용하여 오버 샘플링하는 근사 베이즈 계산 기반 가우시안 혼합 모델(Federated ABCGMM)을 제안하였다. 제안 모델은 중환자실(ICU) 내 환자 사망률 예측 데이터 셋(PhysioNet2012)을 통해 성능 검증 실험을 진행하였다. 전체 데이터 셋을 학습한 모델 성능을 기준으로 각 Party의 데이터만 학습한 모델, 각 Party에서 가우시안 혼합 모델을 통해 생성되는 오버샘플링을 포함하여 학습한 모델, 제안 방식으로 생성되는 오버샘플링을 포함하여 학습한 모델의 성능을 비교한 결과, 제안 방식이 F1 점수 측면에서 기준 모델에 근접한 성능을 도출하였다.

[11]은 연합학습 과정에서 Party를 선택할 때, Party 간의 데이터 분포가 상이함으로 인해 발생하는 전역 모델의 수렴 속도 저하 문제를 해결하기 위하여 Kullbak-Leibler Divergence(KLD)를 통해 각 Party 간 클래스 불균형을 평가하여 데이터 분포를 추정하며, 강화학습을 활용해 클래스 불균형을 최소화하는 장치 선택 알고리즘을 제안하였다. 제안된 알고리즘은 클래스 분포를 학습하고 가장 균형 잡힌 Party 조합을 선택한다. 이를 기반으로 CIFAR10 데이터 셋에 대하여 탐욕 알고리즘 및 Party 무작위 선택과 비교 실험한 결과, 제안 모델이 탐욕 알고리즘 및 무작위 선택에 비해 더 빠른 성능 수렴 속도와 더 높은 정확도를 달성하였다.

[12]는 연합학습의 클래스 불균형을 감지하고 그에 대한 영향을 완화하기 위해 연합학습 훈련 진행 과정에서 각 Party의 데이터 구성을 추론할 수 있는 모니터링 방식을 제안하였다. 훈련 데이터의 구성을 연합학습 라운드별로 코사인 유사도를 통해 추정하며, 클래스 불균형이 지속적으로 감지되면 클래스 불균형을 완화하기 위해 설계한 Ratio Loss를 적용하는 방식으로 진행된다. FEMNIST, MNIST, CIFAR10, Rer2013 4개의 불균형 데이터 셋(Non-IID)에 대한 데이터 분포 유사도 측정 결과, 평균 0.98 이상으로 모니터링의 감지 효과를 보여주었으며, 이를 바탕으로 새로 설계된 손실 함수를 적용하여 클래스 불균형 문제를 완화함으로써 전역 모델의 수렴 성능을 개선하였다.

[11]과 같이 적절한 Party 조합을 사용하여 전체 데이터의 클래스 불균형을 최소화하거나, [10], [12]와 같이 데이터 분포를 유추하여 클래스 불균형 문제를 완화하는 기법을 적용한다면 모델 성능이 개선된다는 점을 미루어보아, [11]의 장치 선택 알고리즘과 [12]의 모니터링 방식을 응용하여 연합학습에 참가하는 각 Party의 모델 가중치 정보를 집계하

는 과정에서 Party 별 모델 성능을 모니터링하여 적절한 수의 모델 가중치만을 집계하고 평균을 취한다면 성능을 개선할 수 있을 것으로 예상하며, 이에 대한 실험을 진행하였다.

[13]의 연구는 의료 영상을 통한 임상 결정 지원 알고리즘을 구현하기 위해 Retinal optical coherence tomography images dataset(Retinal OCT Images)[6]을 그림 2와 같이 전이 학습을 적용하여 신경망을 훈련시키는 프레임워크를 적용하였다.

실험 결과 정확도 93.4%, 민감도 96.6%, 특이도 94.0%로 망막에서 발생할 수 있는 다양한 질병 이미지를 분류하는 데 있어 전문가와 비교 가능한 성능을 보여주었으며, 추가로 흉부 X-ray images dataset을 사용한 소아 폐렴 진단에서도 우수한 성능을 보여줌으로써 제안하는 인공지능 시스템의 일반화 가능성을 확인하였다.

하지만 실험 데이터를 클래스 간 데이터 불균형 문제에 적용해보기 위해 전체 데이터에서 무작위 샘플링하여 학습한 결과, 성능이 저하됨을 확인하였다. 이러한 현상은 많은 양과 다양한 종류의 데이터를 확보하기 어려운 단일 기기 또는 단일 기관에서 로컬 모델로 학습할 때 발생할 가능성이 높다. 따라서 해당 연구는 각 기기 또는 기관 간의 데이터 공유가 어려워 충분한 양의 데이터를 확보하지 못하는 현실에서는 그대로 적용하기 어렵다고 할 수 있다.

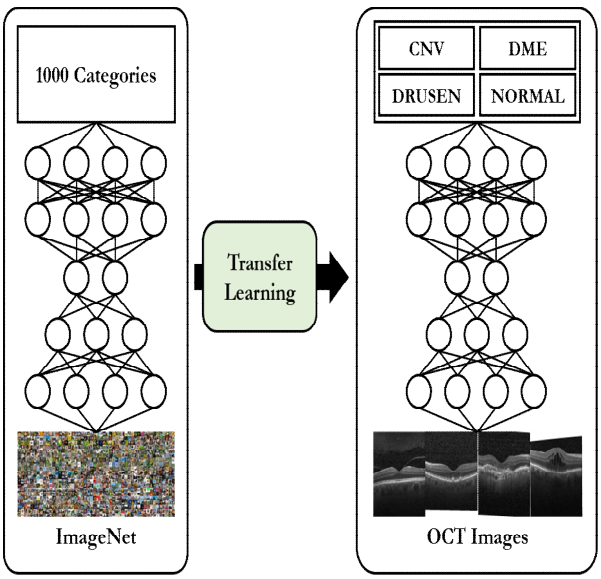

그림 2. [13]의 연구에서 제안하는 프로세스
Fig. 2. Process suggested by [13]'s research

본 논문에서는 일반적으로 불균형한 데이터 분포를 가진 실제 데이터 상황을 가정하여 Retinal OCT Images를 다양한 크기와 분포를 가지는 하위 데이터셋으로 분할하고, 불균형 데이터 문제가 가져오는 성능 저하를 개선하기 위해 언더 샘플링 기법과 TopkAvg 기법을 적용하였다.

2.2 제안 방법

2.2.1 Under Sampling

대부분의 현실 데이터에는 클래스 불균형 문제가 자주 발생한다. 클래스 불균형이란 어떤 데이터에서 각 클래스가 가지고 있는 데이터의 양에 차이가 큰 경우, 즉 데이터가 불균형한 분포를 띄고 있는 것을 의미한다. 임상 의사결정 지원 시스템같은 의료 관련 서비스는 다수의 데이터가 존재하는 정상 탐지가 아닌 소수의 데이터가 존재하는 이상 탐지에 초점을 맞추는데, 일반적으로 데이터가 불균형한 분포를 가지는 경우, 전체 샘플을 정상이라고 판별하는 현상 등 기계학습 모델의 학습이 제대로 이루어지지 않을 확률이 높다. 이러한 현상을 해소하기 위해 데이터 샘플링, 클래스 가중치 등 다양한 방법이 존재한다.

언더 샘플링(Under sampling)은 데이터 샘플링 기법 중 하나로 그림 3과 같이 불균형한 분포를 띄고 있는 데이터셋에서 높은 비율을 차지하는 클래스의 데이터 수를 감소시킴으로써 데이터 불균형을 해소하는 방법이다[14].

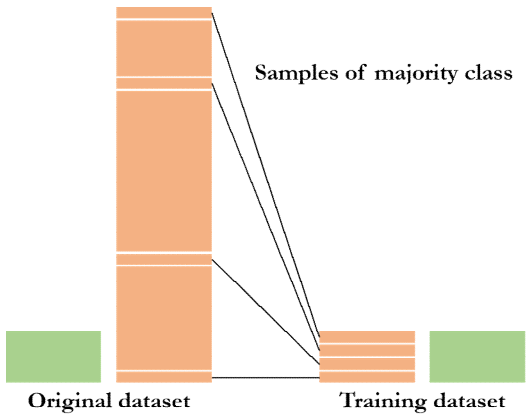

그림 3. 언더 샘플링 개념
Fig. 3. Under sampling concept

이 방법은 다른 클래스 불균형 해소 기법에 비해 상대적으로 빠른 시간 내에 결과를 도출해 낼 수 있지만, 학습에 사용되는 전체 데이터 수를 급격하게 감소시킬 가능성이 존재해 성능이 저하될 수 있다는 단점을 가지고 있다. 따라서 실무에서는 데이터의 특성, 확보 데이터 양에 따라 상이하겠지만 대부분 다량의 데이터 확보에 효과적인 오버 샘플링 기법을 적용하고 있다[15]. 하지만 연합 학습은 모델 가중치 공유를 통한 수 회에 걸친 반복 학습 과정에서 단일 Party가 보유하고 있는 데이터보다 더 많은 데이터를 학습할 수 있으므로 언더 샘플링의 단점을 극복할 수 있다. 따라서 본 논문에서는 연합학습에서 사용하는 데이터셋의 클래스 불균형 문제를 해결하기 위해 언더 샘플링 기법을 제안한다.

2.2.2 Top k Average

연합학습에서는 모델 가중치를 갱신할 때, Party의 모델 가중치 정보를 집계하여 평균을 취하는 방법으로 진행된다[7]. 이 과정에서 전체 Party의 모델 가중치가 집계되기 때문에 Party 수가 증가할수록 연산량이 증가하기 때문에 Aggregator에서 모델 가중치를 업데이트하는데 소요되는 시간도 증가할 수 있다. 따라서 본 논문에서는 해당 문제점을 개선하면서 성능도 유지하기 위해 각 Party의 모델 가중치 정보를 집계하는 과정에서 그림 4와 같이 모델 성능 기준, 전체 Party 성능의 평균 이상 모델 가중치만 집계하여 평균을 취하는 방식(TopkAvg)을 제안한다.

Party별 모델 훈련이 완료되는 시점에서 각 모델에 대한 가중치와 메트릭을 집계한다. 메트릭을 기준으로 모델 가중치를 정렬하여 전체 Party 절반에 해당하는 수만큼의 모델을 새로운 모델 가중치 계산에 사용한다. 업데이트된 모델 가중치는 기존 연합학습 과정과 동일하게 각 Party의 모델 가중치 갱신에 사용된다.

III. 실험 설계

3.1 데이터 전처리

3.1.1 데이터셋

본 논문에서는 [5]에서 제공하는 Retinal optical coherence tomography images dataset(Retinal OCT Images)을 사용하여 기존 연합학습 모델과 제안 모델의 성능 및 학습 안정성을 비교하였다. Retinal OCT Images는 환자 망막의 고해상도 단면을 캡처한 이미지 데이터로 그림 5와 같이 망막 이미지와 병명으로 구성되어 있다.

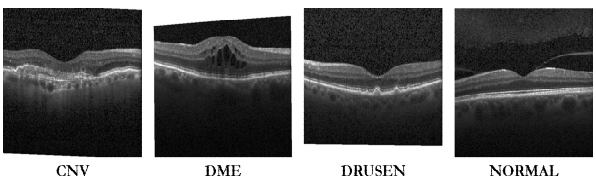

그림 5. Retinal OCT Images 클래스별 예시
Fig. 5. Examples by retinal OCT images class

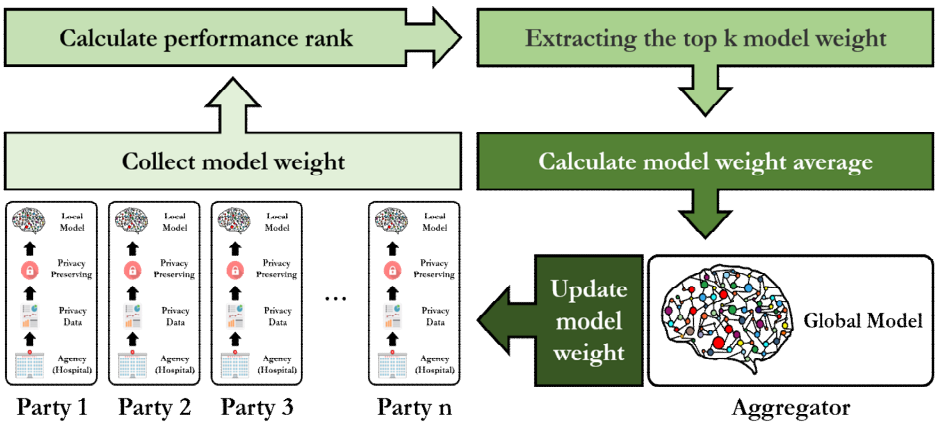

그림 4. TopkAvg 개념
Fig. 4. TopkAvg concept

각 이미지의 파일명은 {병명}-{무작위 환자 ID}-{해당 환자의 이미지 번호}로 구성되어 있으며, 원본 크기는 512×496이고 RGB 3개의 채널로 구성되어 있다. 원본 이미지를 그대로 학습에 사용하면 데이터 불러오기 및 학습 과정에서 속도에 대한 성능이 현저히 낮아지기 때문에 학습 성능이 저하되지 않는 적정값인 100 x 100으로 크기를 재조정했으며 채널은 동일하게 설정하였다.

레이블은 choroidal neovascularization(CNV), Diabetic macular edema(DME), Multiple drusen (DRUSEN), Normal retina(NORMAL) 총 4개로 구성되어있다. 훈련 데이터는 CNV 8,616개, DME 11,348개, DRUSEN 37,205개, NORMAL 26,315개의 이미지를 사용했으며 테스트 데이터로 각 레이블당 242 개의 이미지를 사용하였다.

레이블별 훈련 데이터 분포는 한 불균형한 형태를 보이며, 테스트 데이터 분포는 레이블별로 동일한 개수의 이미지를 사용하였으므로 균일한 형태를 보인다.

3.1.2 데이터 분할

환자의 정보를 관리하고 있는 병원 또는 기관끼리 데이터를 공유하지 않는 실제 환경과 유사한 환경에서 실험을 진행하기 위해 훈련 데이터를 조건에 맞게 분할하였다.

연합학습에 참여하는 각 Party를 환자의 데이터를 보유하고 있는 의료기관이라 가정하고, Retinal OCT Images dataset을 그림 6과 같이 환자를 기준으로 무작위 분할하였다. 테스트 데이터는 로컬 모델 간의 성능을 비교하기 위해 분할하지 않고 동일하게 사용하였다. 위와 같은 분할 방법으로 인해 특정 Party는 특정 환자의 데이터만 보유할 수 있으므로 참가자별 훈련 데이터 개수가 다를 수 있고, Party 간 훈련 데이터 분포가 극명한 차이를 보일 수 있으며, Party 간의 훈련 데이터 중복은 허용되지 않는다.

3.2 실험 설정

본 논문에서는 실제 환경에서 자주 발생하는 클래스 불균형 문제에 연합학습을 적용하기 위한 최적화 기법을 연구하기 위해 다양한 데이터 클래스 분포와 클래스 불균형 문제 해소 기법을 적용하여 비교 실험을 진행했다.

클래스 간 약한 불균형 데이터 분포를 가졌으나 각 클래스가 충분한 양의 데이터를 가진 분할 데이터셋과 어떠한 최적화 기법도 적용하지 않은 4개의 로컬 모델을 가진 연합학습 모델을 베이스라인으로 설정하였다. 이에 대한 다양한 최적화 기법을 적용하여 비교하였으며, 추가로 연합학습에 참가하는 Party 수에 따른 성능 비교도 함께 진행하였다.

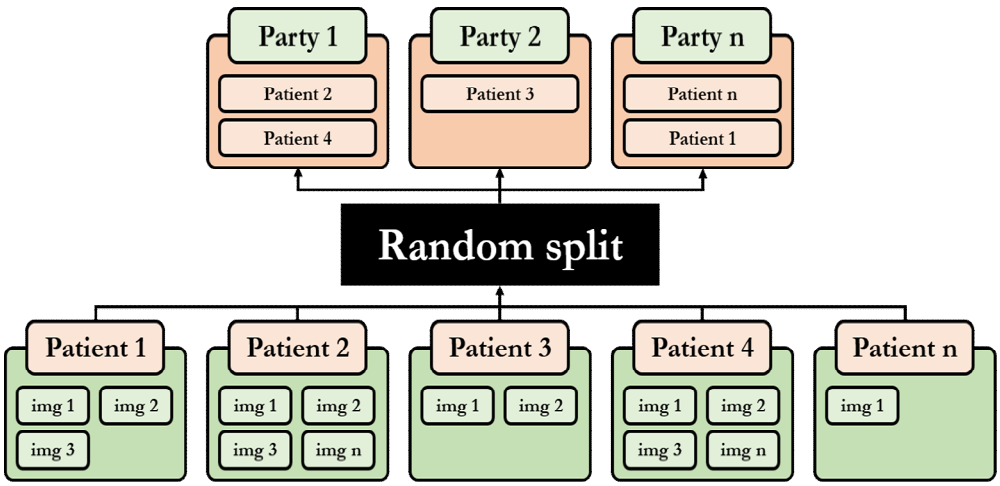

그림 6. 데이터 분할 프로세스
Fig. 6. Data split process

또한, 불균형 클래스 문제에 따른 성능 변화를 알아보기 위해 다양한 데이터 클래스 분포를 가진 분할 데이터셋을 생성했으며, Aggregator에서 로컬 모델의 가중치를 집계하는 방법에 대한 비교도 함께 진행했다.

실험은 Google Cloud Platform(GCP)에서 Intel Skylake CPU 16개, 메모리 용량 104GB, NVIDIA Tesla P100 GPU 1개를 사용한 VM 인스턴스 환경에서 진행했으며, 패키지 버전은 Python 3.6.12, tensorflow-gpu 1.15.0, keras-gpu 2.1.6, cuda 11.1, cudnn 7.6.5를 사용했으며, 연합학습에 참여하는 각 참가자들의 학습 환경을 동일하게 구성하기 위해 시스템 자원을 동일한 비율로 할당했다.

학습에 사용한 모델은 사전 학습된 VGG16의 출력층에 컨볼루션 층을 연결한 전이 학습 구조며, 하이퍼 파라미터는 연합학습의 경우에 고유한 로컬 및 데이터를 보유하고 있는 여러 분산 장치와 각 장치에서 모델 가중치를 집계하여 새로운 모델 가중치를 계산하는 Aggregator로 나뉘기 때문에 표 1과 같이 글로벌 하이퍼 파라미터와 로컬 하이퍼 파라미터 두 가지가 존재한다.

표 1. 연합학습 모델 하이퍼 파라미터
Table 1. Hyper-parameter of federated learning model

Hyper-parameter	Name	Value
global	max_timeout	36000
	num_parties	4~8
	rounds	10~100
local	lr	0.001
	epochs	10
	batch_size	128

max_timeout은 Aggregator에서 모델 가중치를 집계에 대한 제한 시간이다. 설정한 시간 내에 모든 참가자들의 모델 가중치가 집계되지 않으면 에러가 발생한다. num_parties는 모델 학습에 참가하는 참가자의 수다. 설정한 수만큼 로컬 모델이 연결되지 않으면 학습이 실행되지 않는다. rounds는 로컬 모델 학습 반복 횟수다. 해당 수치만큼 모델 가중치 집계와 업데이트가 진행된다. lr는 로컬 모델의 학습율이고, epochs는 한 번의 round 당 로컬 모델이 훈련하는 횟수며, batch_size는 각 로컬 모델에 대한 배치 크기다.

Ⅳ. 실험 결과

대부분의 이미지 분류 모델은 메트릭으로 올바르게 예측된 데이터의 수를 전체 데이터의 수로 나눈 값인 정확도를 사용한다. 하지만 클래스 불균형 문제에 정확도를 사용해 성능을 평가할 경우, 관점에 따라 잘못된 결과가 나올 가능성이 높다.

따라서 본 논문에서는 모델의 평가 지표로 F1 score를 사용했다. F1 score는 실제 참인 데이터를 참이라고 예측한 데이터 개수의 비율을 나타내는 재현율 또는 민감도와 참으로 예측한 데이터 중 실제로 참인 데이터 개수의 비율을 나타내는 정밀도 두 값의 조화평균으로, 불균형 데이터에 있어 정확도보다 정확한 평가를 도출할 수 있다[16].

4.1 클래스 분포

표 2는 데이터 클래스 분포에 따른 F1 score 측정 결과다. 균일한 분포로 분할한 데이터셋을 사용하여 학습(balance)을 진행한 결과, party1, party2, party4에서 최고 성능 0.95, 학습 시간 2051.48초로 기존 중앙 집중형 모델과 동일한 성능을 도출했음을 보여준다. 반면에 균일하지 않은 분포로 분할한 데이터셋을 사용하여 학습(imbalance)을 진행한 결과, party4에서 최고 성능 0.87을 달성했으나, party1과 party2의 경우, F1 score가 0.5 이하로 나오는 비정상적인 학습 결과가 도출되었다.

표 2. 데이터 분포에 따른 성능
Table 2. Performance according to data distribution

Distribution	Party name	Best performance
balance	party1	**0.95**
	party2	**0.95**
	party3	0.94
	party4	**0.95**
imbalance	party1	0.50
	party2	0.25
	party3	0.74
	party4	0.87

이러한 현상은 많은 양과 다양한 종류의 데이터를 확보하기 어려운 현실에서 발생할 가능성이 높다. 따라서 다양하고 충분한 양의 데이터를 확보하

지 못할 가능성이 높은 현실 문제에 연합학습을 적용하기 위해서는 클래스 불균형 문제를 해소할 필요가 있다.

4.2 최적화 기법

연합학습에서 클래스 불균형 문제를 해소하는 다양한 최적화 기법들의 비교를 위해 불균형 분할 데이터셋을 사용하여 학습을 진행했다. 클래스 불균형 최적화 기법은 오버 샘플링, 언더 샘플링, 클래스 가중치를 사용했으며, 결과는 표 3과 같다.

표 3. 클래스 불균형 최적화 기법에 따른 성능
Table 3. Performance according to the class imbalance optimization technique

Distribution	Party name	Best Performance
class weight	party1	0.50
	party2	0.25
	party3	0.90
	party4	0.80
over sampling	party1	0.94
	party2	0.94
	party3	0.94
	party4	0.94
under sampling	party1	0.94
	party2	0.94
	party3	0.94
	party4	**0.95**

오버 샘플링을 적용한 경우, 모든 party에서 최고 성능 0.94의 준수한 성능을 보였으나, 소수의 데이터를 가진 클래스 데이터 비율을 다수의 데이터를 가진 클래스 데이터 비율에 맞춤에 따라 총 데이터 개수가 증가하여 기존 모델 학습보다 훨씬 많은 시간 학습 시간 17,240.10초가 소요된다는 문제점이 발생하였다.

언더 샘플링을 적용한 경우, party4에서 최고 성능 0.95로 기존 모델 학습과 동일한 성능을 도출하였으며, 학습 시간 또한 1,566.50초로 기존보다 빠른 결과를 보여주었다. 이는 연합 학습에서 모델 가중치 공유를 통한 수 회에 걸친 반복 학습 과정이 단일 Party가 보유하고 있는 데이터보다 더 많은 데이터를 학습할 수 있게 해주어 언더 샘플링의 단점을 극복한 결과라 할 수 있다.

클래스 가중치를 적용한 경우, party3에서 최고 성능 0.90을 달성했으나 party1과 party2에서 불균형 데이터셋 성능과 유사하게 비정상적인 학습 결과가 도출되었다.

4.3 Party 수 및 학습 반복 횟수

총 데이터 개수가 동일하다는 가정 하에, 추가 실험으로 표 4와 같이 Party 수 및 학습 반복 횟수 따른 성능을 비교했다. 언더 샘플링 기법을 적용한 연합학습에 참가하는 참가자 수를 8로, 전역 하이퍼파라미터의 rounds를 100으로 증가시켜 실험을 진행했다.

표 4. Party 수 및 학습 반복 횟수에 따른 성능
Table 4. Performance according to the number of parties and the number of repetitions of learning

Distribution	Party name	Best Performance
num_parties=8, rounds=10	party1	0.94
	party2	0.94
	party3	0.94
	party4	0.94
	party5	0.93
	party6	0.94
	party7	0.94
	party8	0.93
num_parties=4, rounds=100	party1	0.95
	party2	**0.96**
	party3	0.95
	party4	**0.96**
num_parties=8, rounds=100	party1	0.95
	party2	0.95
	party3	0.94
	party4	0.95
	party5	0.94
	party6	0.95
	party7	0.95
	party8	0.95

Party 8개, 학습 반복 횟수 10회로 설정하여 실험한 결과, 대부분의 Party가 최고 성능 0.94로 준수한 성능을 보였으며, 학습 시간은 1,572.79초로 Party가

4개인 경우보다 약간의 추가 시간이 소요되었다. 이는 Party 수가 증가함에 따라 집계하는 모델 가중치 수도 증가하여 새로운 모델 가중치를 계산하고 전달하는 과정에서 발생하는 차이라고 할 수 있다.

학습 참가 Party 4개, 학습 반복 횟수 100회로 설정하여 실험한 결과, 모든 Party의 최고 성능이 0.95 이상으로 학습 반복 횟수 10회보다 좋은 성능을 보였으며, 학습 시간은 15,694.32초가 소요됐다.

Party 8개, 학습 반복 횟수 100회로 설정하여 실험한 결과, 대부분의 Party 최고 성능이 0.95 이상으로 좋은 성능을 보였으며, 학습 시간은 15,737.70초가 소요됐다.

학습 참가자 수 및 학습 반복 횟수에 따른 실험 결과, 매우 적은 데이터를 가지고 있는 단일 Party라도 연합학습에 참가하여 모델 가중치 공유를 통해 더 많은 데이터를 학습함으로써 준수한 성능을 낼 수 있음을 확인했다. 또한, 학습 반복 횟수가 증가함에 따라 조금 더 좋은 성능을 도출할 수 있었으나, 학습 시간 개선이 필요함을 확인했다.

4.4 모델 가중치 집계 방법

본 논문에서는 학습에 소요되는 시간을 개선하면서 성능도 유지하기 위해 각 참가자의 모델 가중치 정보를 집계하는 과정에서 모델 성능을 기준으로 전체 Party 성능의 평균 이상 모델 가중치만 집계하여 평균을 취하는 방식(TopkAvg)을 제안한다. 그림 7, 8은 학습 참가 Party 수를 각각 4, 8개로 설정하여 제안 방식을 적용한 학습 결과다.

Party가 4개인 경우, 학습 성능은 기존과 동일하게 최고 성능 0.96, 학습 시간 1,561.22초로 기존보다 약 4초 감소했다.

Party가 8개인 경우, 학습 성능은 기존보다 조금 더 높은 최고 성능 0.96, 학습 시간 15,721.96초로 기존보다 약 16초 감소하였다.

이는 모델 가중치를 갱신하는 과정에서 전체 Party의 가중치를 집계하는 방식보다 우수한 성능을 나타낸 Party 모델의 가중치만 집계하여 계산하는 방법이 성능을 개선하거나 유지할 수 있고, 학습 시간 또한 개선할 수 있음을 보여준다.

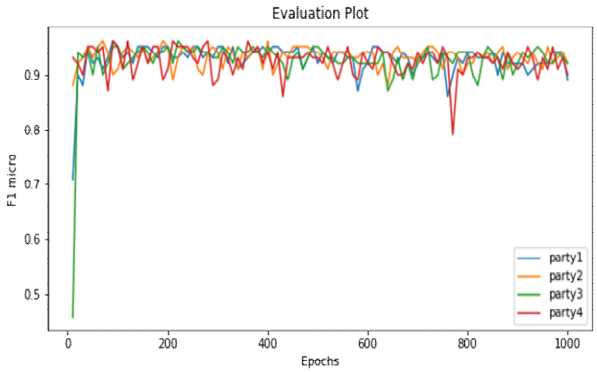

그림 7. TopkAvg 집계 방식 성능(num_parties=4, rounds=100)

Fig. 7. Performance of TopkAvg aggregation method (num_parties=4, rounds=100)

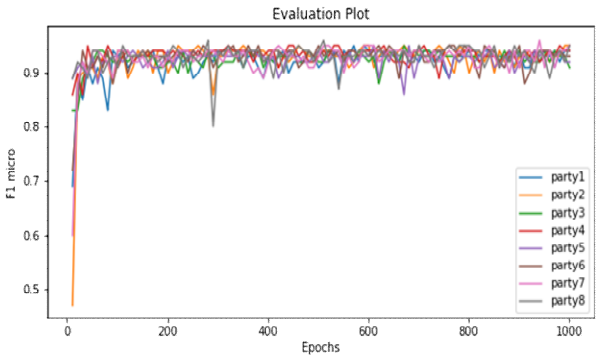

그림 8. TopkAvg 집계 방식 성능(num_parties=8, rounds=100)

Fig. 8. Performance of TopkAvg aggregation method (num_parties=8, rounds=100)

본 논문에서 TopkAvg에 대한 실험은 전체 Party에 대한 모델 가중치와 F1 score를 집계한 후, 전체 Party의 F1 score 평균 이상에 해당하는 모델 가중치만 사용하여 평균 계산을 취함으로써 모델 가중치를 업데이트하는 방식으로 진행하였다. 추가로, Aggregator와 Party를 연결해주는 Connection 과정에서 각 Party의 로컬 모델 메트릭을 기준으로 연결 우선 순위를 지정하여 모델 가중치를 집계한다면 학습 시간을 더욱 개선할 수 있을 것으로 예상된다.

V. 결 론

본 논문에서는 실제 환경에서 자주 발생하는 클래스 불균형 문제에 연합학습을 적용하기 위한 최적화 기법을 연구하기 위해 다양한 데이터 클래스 분포와 클래스 불균형 문제 해소 기법 및 모델 가

중치 집계 기법을 적용한 실험을 진행하였다.

클래스 불균형에 대한 최적화 기법 탐색 실험에서는 오버 샘플링, 언더 샘플링, 클래스 가중치를 사용했으며, 언더 샘플링을 적용한 경우 최고 성능 0.95로 기존 모델 학습과 동일한 성능을 도출하였으며, 학습 시간 또한 1,566.50초로 기존보다 빠른 결과를 보여주었다. 이는 연합 학습에서 모델 가중치 집계를 통해 생성된 공유 모델이 단일 참가자가 보유하고 있는 데이터보다 더 많은 데이터를 학습할 수 있어, 언더 샘플링의 단점인 데이터 축소를 극복한 결과라 할 수 있다.

모델 가중치 집계 기법 탐색 실험에는 기존 FedAvg 방식과 본 논문에서 제안하는 TopkAvg 방식을 비교했으며, TopkAvg 방식이 기존보다 더 높은 성능(0.96)과 학습 시간 개선(15,721.96초)이라는 결과를 도출했다. 이는 공유 모델 가중치를 갱신하는 과정에서 우수한 성능을 나타낸 Party의 모델 가중치만 집계하여 계산하는 방법이 성능과 학습 시간 또한 개선할 수 있음을 보여준다. 이를 통해 Aggregator와 Party의 연결 과정에서 우선 순위 지정을 통한 모델 가중치 집계를 구현한다면 학습 시간을 더욱 개선할 수 있을 것으로 예상된다.

References

[1] Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, and Dzmitry Huba, et al., "Towards federated learning at scale: system design", arXiv preprint arXiv:1902.01046, 2019.

[2] Wenqi Li, Fausto Milletar`ı, Daguang Xu, and Nicola Rieke, et al., "Privacy-preserving federated brain tumour segmentation", International Workshop on Machine Learning in Medical Imaging MLMI 2019: Machine Learning in Medical Imaging, Vol. 11861, pp. 133-141, Oct. 2019.

[3] Micah J Sheller, G Anthony Reina, Brandon Edwards, Jason Martin, and Spyridon Bakas, “Multi-Institutional Deep Learning Modeling Without Sharing Patient Data: A Feasibility Study on Brain Tumor Segmentation”, International MICCAI Brainlesion Workshop BrainLes 2018: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Vol. 11383, pp. 92-104, Jan. 2019.

[4] Hye Kyeong Ko, "A Study Personal Information Protection Technique for XML-based Electronic Medical Record", Journal of KIIT, Vol. 12, No. 5, pp. 185-191, May 2014.

[5] Micah J. Sheller, Brandon Edwards, and G. Anthony Reina, et al., "Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data", Nature Scientific Reports, Vol. 10, Article number: 12598, Jul. 2020.

[6] Daniel Kermany, Kang Zhang, and Michael Goldbaum, "Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification", Mendeley Data, Vol. 2, Jan. 2018.

[7] H. Brendan McMahan Eider Moore Daniel Ramage Seth Hampson and Blaise Aguera y Arcas, "Communication-efficient learning of deep networks from decentralized data", Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, PMLR, Vol. 54, pp. 1273-1282, Feb. 2017.

[8] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra, "Federated Learning with Non-IID Data", arXiv preprint arXiv:1806.00582, 2018.

[9] Moming Duan, Duo Liu, Xianzhang Chen, Yujuan Tan, Jinting Ren, Lei Qiao, and Liang Liang, "Astraea : Self-balancing Federated Learning for Improving Classification Accuracy of Mobile Deep Learning Applications", arXiv preprint arXiv: 1907.01132v2, May 2020.

[10] Seok-Ju Hahn and Junghye Lee, "Privacy-preserving Federated Bayesian Learning of a Generative Model for Imbalanced Classification of Clinical Data", arXiv preprint arXiv:1910.08489v3, Aug. 2020.

[11] Lixu Wang, Shichao Xu, Xiao Wang, and Qi Zhu, "Addressing Class Imbalance in Federated Learning", arXiv preprint arXiv:2008.06217v2, Dec. 2020.

[12] Miao Yang, Akitanoshou Wong, Hongbin Zhu, Haifeng Wang, and Hua Qian, "Federated learning with class imbalance reduction", arXiv preprint arXiv:2011.11266, Nov. 2020.

[13] Daniel S. Kermany, Michael Goldbaum, and Wenjia Cai, et al., "Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning", Cell Press Journals, Vol. 172, No. 5, pp. 1122-1131, Feb. 2018.

[14] X. Liu, J. Wu, and Z. Zhou, "Exploratory Undersampling for Class-Imbalance Learning", IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), Vol. 39, No. 2, pp. 539-550, Apr. 2009.

[15] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique", Journal of Artificial Intelligence Research, Vol. 16, pp. 321-357, Jun. 2002.

[16] Juri Opitz and Sebastian Burst, "Macro F1 and Macro F1", arXiv preprint arXiv:1911.03347, 2019.

저자소개

이 현 수 (Hyeonsu Lee)

2019년 8월 : 강원대학교 컴퓨터정보통신공학과(공학사) 수료
2019년 9월 ~ 현재 : 강원대학교 컴퓨터정보통신공학(석사과정)
관심분야 : 머신러닝, 빅데이터, 데이터 마이닝, 데이터 임베딩

홍 성 은 (Seongeun Hong)

2015년 : 강원대학교 컴퓨터정보통신공학부(공학석사)
2015년 ~ 2019년 : 강원대학교 컴퓨터정보통신공학과(공학박사) 수료
관심분야 : 빅데이터, 데이터 마이닝, 기계학습, 딥러닝

방 준 일 (Junil Bang)

2020 8월 : 강원대학교 컴퓨터정보통신공학과(공학석사)
2020년 9월 ~ 현재 : 강원대학교 컴퓨터정보통신공학과(박사과정)
관심분야 : 데이터마이닝, 머신러닝, 빅데이터, 데이터 임베딩

김 화 종 (Hwajong Kim)

1984년 3월 : KAIST 전기및전자공학과(공학석사)
1988년 3월 : KAIST 전기및전자공학과(공학박사)
1988년 3월 ~ 현재 : 강원대학교 컴퓨터정보통신공학과(정교수)
관심분야 : 데이터 통신, 컴퓨터네트워크, 네트워크 프로그래밍, 빅데이터