**DPFL-FUSION : Differentially Private Federated Learning – Fake features Utilizing SHAP-based Importance for Optimized Noise**

CUAI 8기 CV3팀

김대현(AI학과), 김태환(AI학과), 나상현(소프트웨어학부), 박서윤(경영학부)

**[요약]**

본 연구는 연합 학습 환경에서 프라이버시 보호와 예측 성능 향상을 동시에 달성하기 위해, SHAP 기반 피처 중요도 평가와 연속적 매핑 함수를 통한 정밀 노이즈 할당, 그리고 가짜 피처를 활용한 노이즈 흡수 전략을 결합한 Adaptive DPFL 프레임워크를 제안한다. 제안 기법은 각 피처의 중요도에 따라 노이즈를 차등 주입하여 DP 보장을 유지하면서도 모델 업데이트의 민감도를 효과적으로 조절하며, 실험 결과 기존 방법 대비 학습 안정성과 일반화 성능에서 우수함을 확인하였다. 이러한 결과는 안전한 분산 학습 시스템 구축에 기여할 수 있음을 시사한다.

**1. 서 론**

최근 빅데이터, 딥러닝, 분산 컴퓨팅 기술의 발전과 함께, 다양한 분야에서 머신러닝 모델의 학습 및 활용이 급증하고 있다. 이에 중앙집중형 학습의 한계와 개인정보 보호·데이터 보안에 대한 요구가 증가하면서, 각 클라이언트가 로컬에서 모델을 학습한 후 파라미터만 공유하는 연합 학습(FL)이 주목받고 있다. FL은 개별 디바이스의 프라이버시를 보장하면서 글로벌 모델 성능을 향상시키지만, 공유되는 모델 업데이트(그래디언트 또는 파라미터)만으로 민감 정보가 유추될 위험이 있어 추가 보안 메커니즘이 필요하다[1].

이를 보완하기 위해 도입된 차등 프라이버시(DP)는 데이터에 의도적 노이즈를 주입하여 특정 데이터의 존재나 기여도를 은닉하는 강력한 기법이다. 다만, 기존 DP 기법은 모든 피처나 파라미터에 동일한 노이즈를 부여하여 경우에 따라 모델 예측 성능에 악영향을 줄 수 있으므로, 보다 세밀한 노이즈 할당 전략이 요구된다. 한편, 모델 해석 가능성에 대한 관심 증대로 SHAP(Shapley Additive exPlanations)가 각 피처의 기여도를 정량적으로 평가하는 도구로 활용되며, DP 적용 시 피처별 중요도에 따른 노이즈 주입 전략 수립에 유용하다.

본 연구에서는 실제 모델 예측에 결정적인 영향을 미치지 않는 영역을 마련하기 위해, 기존 데이터에 포함되지 않은 인위적 가짜 피처를 도입한다. 이를 통해 노이즈가 우선적으로 할당될 영역을 확보함으로써, 실제 중요한 피처에 전달되는 노이즈 양을 효과적으로 여 학습 안정성과 일반화 성능을 향상시키고자 한다.

이와 같은 배경 하에, 본 연구는 SHAP 기반 피처 중요도 평가, 연속적 매핑 함수를 통한 정밀 노이즈 할당, 그리고 가짜 피처 활용에 의한 노이즈 흡수라는 세 가지 핵심 기여를 바탕으로, 분산 학습 환경에서 효과적인 프라이버시 보호와 예측 성능 향상을 동시에 달성할 수 있는 Adaptive DPFL 프레임워크를 제안한다. 이를 통해 FL 시스템의 안전성과 실용성을 크게 향상시키며 다양한 응용 분야에서의 활용 가능성을 제고하고자 한다.

**2. 본 론**

**2.1** **SHAP 기반 피처 중요도 평가 및 적응형 노이즈 주입:**

기존 연구에서는 분산이나 민감도와 같은 전통적 지표를 활용하여 피처 중요도를 산출해왔다면, 본 연구에서는 SHAP (SHapley Additive exPlanations) 기법을 도입하여 각 피처가 모델 출력에 미치는 기여도를 정량적으로 평가한다. SHAP 값은 각 피처의 기여도를 공정하고 세밀하게 반영할 수 있으므로, 중요한 피처와 덜 중요한 피처를 명확히 구분하여 다음과 같은 두 가지 측면에서 활용된다.

**(1) 적응형 노이즈 부여**

로컬 에포크 진행 시 각 미니배치 내의 개별 샘플 그래디언트는 미리 정해진 클리핑 임계값 이하로 제한된다. 에포크 종료 시 파라미터에 노이즈를 주입할 때, SHAP으로 산출한 피처 중요도 정보를 활용하여 중요한 피처에는 상대적으로 적은 노이즈, 덜 중요한 피처에는 더 많은 노이즈를 할당한다.

또한, SHAP 값 자체가 원본 데이터에 민감할 수 있으므로, 추가적인 DP 보호를 위해 SHAP 값에도 Gaussian 노이즈를 부여하였다.

**(2) 정량적 평가의 장점**

SHAP 기법은 피처가 출력에 미치는 기여도를 Shapley value 관점에서 정밀하게 산출할 수 있으므로, 기존 민감도나 분산 기반 평가보다 보다 공정하고 세밀한 피처 중요도 분석을 가능하게 한다. 이를 통해 산출된 피처 중요도 를 바탕으로 적응형 스케일러 를 산출하는데, 후술할 DP 보장 증명 파트에서 의 정규화(즉, 조건 만족) 가 전체 노이즈 에너지와 민감도 보정에 결정적인 역할을 함을 보인다.

**2.2 연속적 매핑 함수를 통한 정밀 노이즈 할당:**

단순한 비율 기반 노이즈 주입 방식 대신, 본 연구에서는 SHAP으로 산출한 각 피처의 중요도를 norm 기반 정규화를 통해 스케일을 정규화한 후 지수, 역수, 시그모이드 등 다양한 연속 매핑 함수를 활용하여 정밀하게 노이즈 할당량을 결정하는 방법을 제안하였다. 이러한 접근법은 정규화를 통해 피처 중요도의 분포를 균일하게 만든 후, 연속적인 함수로 변환하여 노이즈 주입량을 보다 세밀하게 조절할 수 있도록 설계되었으며, 결과적으로 모델의 예측 성능과 프라이버시 보호 간의 균형을 최적화할 수 있는 장점을 제공한다.

**2.3 가짜 피처를 통한 노이즈 흡수:**

본 연구에서는 기존에 DPFL에 시도된 바 없는 새로운 전략으로, 가짜 피처를 활용하여 모델 예측에 직접 기여하지 않는 영역에 노이즈를 우선적으로 할당하는 방법을 제안한다. 이를 통해 실제 중요한 피처에 전달되는 노이즈의 양을 효과적으로 감소시키며, 추가적인 정규화 효과를 도모하여 학습 안정성과 전체 모델의 일반화 성능을 향상시킬 수 있는 가능성을 제시한다.

**2.4 DP 보장 유지 확인**

**(1) 차등 프라이버시 기본 정의**

다음은 본 연구에서 가우시안 DP 메커니즘과 SHAP 중요도 기반으로 노이즈를 할당한 것이 어떻게 -DP 를 충족하는지 DP-SGD 프레임워크를 기반으로 증명한다.

먼저, 가우시안 노이즈 적용 시 DP가 보장됨을 확인한다.

두 인접 데이터셋 와 가 단 하나의 데이터만 다른 경우에 대해, 알고리즘 이 -DP를 만족한다는 것은 모든 출력 집합 에 대해

가 성립함을 의미한다[2]. 이는 한 샘플의 변화가 알고리즘의 출력 분포에 미치는 영향을 제한하여 프라이버시를 보장하는 핵심 조건이다.

**(2) 그래디언트 클리핑과 민감도 제한**

로컬 업데이트 과정에서 각 샘플의 그래디언트 ​는 클리핑을 통해 임계값 C 이하로 제한되고, 이를 통해 모델 파라미터의 업데이트 를 구한다.

그래디언트 클리핑으로 인해 해당 샘플의 그래디언트가 최대 로 제한된다. 의 최대값은 이므로 한 샘플 차이가 평균에 미치는 영향은 이고, 최종 업데이트에서는 학습률이 곱해지므로 평균 그래디언트에 미치는 최대 영향, 즉 민감도 는 다음과 같이 제한된다.

(는 학습률, 는 배치 크기)

민감도가 제한되면 이를 토대로 가우시안 노이즈의 표준편차를 제한할 수 있다. (Theorem 3.22 in [3])

이를 통해 업데이트 f(D)에 노이즈를 추가하면 다음과 같다. [1]

이 조건을 만족하면 가우시안 메커니즘이 한번의 업데이트에 대해 -DP를 만족한다.

**(3) Adaptive Noise Injection 및 정규화**

각 피처에 대해, DP로 보호된 SHAP 값을 기반으로 적응형 스케일러 ​를 산출한다. 이때, 중요한 피처일수록 ​가 작게, 덜 중요한 피처일수록 ​가 크게 결정된다.

이때, 는 다음과 같이 정의된다.

이때 은 피처의 개수이다. 이 정규화 과정의 결과로 가 성립한다.

따라서 차원 노이즈 벡터 의 제곱 기대값은

즉, 별도의 노이즈 벡터 를 활용해도 기존에 계획한 노이즈 에너지()와 동일함을 알 수 있다.

만약 가 정규화되지 않는다면 실제 적용되는 노이즈 에너지가 이론적 민감도 대비 과대 혹은 과소하게 되어 DP 보장이 위협받을 수 있음을 주의해야 한다.

**(4) Moments Accountant를 통한 최종 DP 보장**

업데이트 단계에서 적용된 DP와 SHAP 기반 피처 중요도 평가에 따른 DP는 단순 합산되는 것이 아니라, MA등의 고급 합성 기법을 통해 합산된다[4]. 구체적으로, 업데이트 DP에서 산출된 프라이버시 비용 와 SHAP-DP에서 산출된 비용 는

와 같이 결합되어, 전체 로컬 업데이트가 -DP를 만족함을 보장한다.

**(5) 가짜 피처의 영향**

입력 벡터가 일 때, 가짜 피처를 추가하면 다음과 같이 확장된다.

이때 여기서 추가된 개의 차원은 모두 0이므로, 두 이웃 데이터셋 와 에서 가짜 피처 부분의 값은 항상 동일하며, 이로 인해 해당 부분은 민감도에 기여하지 않는다.

민감도에 영향을 주지 않기 때문에, 전체 DP 보장에 추가적인 프라이버시 손실을 발생시키지 않는다.

**2.5 실험**

텍스트, 스크린샷, 라인, 그래프이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.

그림1. [MNIST] SHAP vs Variance

텍스트, 스크린샷, 그래프, 라인이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.

그림2. [MNIST] Naïve vs Continuous Functions

텍스트, 스크린샷, 그래프, 라인이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.

그림3. [MNIST] Fake Features 0 vs 5 vs 10 vs 20

텍스트, 스크린샷, 그래프, 라인이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.

그림4. [MNIST] Previous ([5]Talaei, 2024) vs Ours

대조군은 [5]의 분산 기반 단순 적응형 노이즈 모델이다.

SHAP을 활용할 때, 연속 매핑함수를 활용할 때, 가짜 피처를 증강하여 학습할 때 모두 뚜렷한 성과가 나타났다.

**3. 결 론**

본 연구는 DPFL 환경에서 프라이버시 보호와 예측 성능 향상을 위해 적응형 프레임워크 DPFL-FUSION를 제안한다. 제안된 프레임워크는 피처 민감도를 고려한 노이즈 주입 및 정규화를 통해 DP 보장을 유지하면서 모델 성능을 개선함을 확인하였다. 향후 다양한 응용 분야에서 안전한 연합 학습 시스템 구축에 기여할 것으로 기대된다.

**참고 문헌**

[1] Zhu, Ligeng, Zhijian Liu, and Song Han. "Deep leakage from gradients." NeurIPS. 2019.

[2] Abadi, Martin, et al. "Deep learning with differential privacy." ACM SIGSAC Proceedings. 2016.

[3] Dwork, Cynthia, and Aaron Roth. "The algorithmic foundations of differential privacy." Foundations and Trends® in Theoretical CS. 2014.

[4] Geyer, Robin C., Tassilo Klein, and Moin Nabi. "Differentially private federated learning: A client level perspective." arXiv preprint. 2017.

[5] Talaei, Mahtab, and Iman Izadi. "Adaptive Differential Privacy in Federated Learning: A Priority-Based Approach." arXiv preprint. 2024.