（ 続紙 1 ）

京都大学	博士（情報学）	氏名	加藤 郁之
論文題目	A Study on Private and Secure Federated Learning （プライベートで安全な連合学習）		

（論文内容の要旨）

連合学習(FL (Federated Learning))は、異なる機関やデバイス間でのプライバシを考慮した共同学習を実現する新しい機械学習パラダイムとして登場している。基本的な枠組みは、一つの中央サーバが全体の訓練プロセスを調整し、多くのクライアントが自身のプライベートデータを訓練データとして持つ。これらが分散最適化アルゴリズムに従い、訓練データを直接共有せずに単一のグローバルモデルを共同で訓練する。連合学習では最適化プロセスで使用される勾配情報のみが共有されるため、クライアントは生のデータを共有する必要はなく，モデルを訓練するサーバはプライベートデータの管理コストから解放される。GDPRに代表される大規模データ分析におけるプライバシ規制の懸念が高まる中、連合学習は学界と産業界の双方から注目を集めている。

しかし、連合学習はプライバシを意識した機械学習スキームであるにもかかわらず、その分散型のアーキテクチャにより様々なプライバシリスクを抱えている。連合学習を使用したとしても訓練されたモデル自体に厳格なプライバシ保証がなく、公開されたモデルは訓練データに関する機密情報を漏洩することがある。さらに、訓練フェーズで分散した当事者間で交換される勾配情報からもプライベート情報が漏洩することが知られている。プライバシに関わる情報の漏洩に加え、セキュリティリスクも存在する。分散型の性質により，協調を行うサーバとクライアントは信頼が必要となる。しかし実際には既存のプロトコルから逸脱し、モデルの振る舞いを不適切に制御することがある。このように連合学習はプライバシとセキュリティの点で不十分なスキームである。

本論文では、プライベートで安全な連合学習を実現するため、先進的なプライバシ保護技術を検討し、特に差分プライバシ（DP）とTrusted Execution Environment（TEE）に焦点を当て、連合学習の不完全性を包括的に克服することを目指す。本論文の目標は、次の主要な研究問題に答えることである 「連合学習におけるプライバシとセキュリティを強化するために、差分プライバシや信頼実行環境などの先進的なプライバシ保護技術をどのように効果的に統合できるか？」この問いに答えるために、既存研究とユースケースから連合学習の主要なプライバシ/セキュリティ特性を抽出し、何を行うべきかを明確にする。次に、DPやTEE、または代替的なMPC技術を利用して、連合学習のセキュリティおよびプライバシの側面の弱点を克服するために、ULDP-FL（第3章）、OLIVE（第4章）、VLDP（第5章）という三つの革新的なFL/FA (Federated Analytics) フレームワークを設計した。

第3章では、DPを用いた連合学習で訓練されたモデルの厳格なプライバシ保護を検討している。特に、各参加クライアントがある規模の機関に相当するクロスサイロ連合学習の一般的な設定を目標とし、ユーザレベルのDP保証を提供する。ユーザレベルのDPは、通常のDPにおける単一レコードではなく、ユーザが保持するすべてのレコードに対する識別不可能性を保証するより実用的なDPの定義である。この設定の下で、既存のアルゴリズムは実用的でないプライバシ保証しか達成できないことを示し、より優れたプライバシと有用性のトレードオフを提供するアルゴリズムを提案している。提案された方法は、既存の事実上のDP-FedAVGに対してユーザごとの重み付けクリッピングを適用することで直接ユーザレベルのDPを保証する。さらに、より厳格な信頼モデルの下で実現可能な効用向上のための重み付け方法を提案し、それを達

成するためのMPCプロトコルを開発している。

第4章では、連合学習におけるサーバ側のTEEに焦点を当てている．これは，中央の信頼できないサーバに対して共有勾配のプライバシを保証し、差分プライベート連合学習のより高い実用性を提供することを可能にする。TEEは連合学習に高いレベルのセキュリティを提供するが、TEE自体にはメモリアクセスのリークという基本的な脆弱性があることが知られている。本研究では連合学習におけるTEE内での集約操作のメモリアクセスパターンの分析を通じて、スパース化勾配を使用した場合にプライバシリスクが発生する可能性を発見した。観測可能なメモリアクセスパターン情報を用いて、プライベートデータを暴露する新しい攻撃を設計し、実データを用いた実験によりその有効性を示した。この攻撃を防御するために、連合学習における集約演算の際に発生するメモリアクセスパターンが入力データに依存しないような忘却アルゴリズムを設計した。最後に、提案した忘却アルゴリズムを実データで評価し、その効率性を示した。

第5章では、連携分析(FA)タスクの出力の振る舞いを制御するために想定されるプロトコルから逸脱する悪意のあるクライアントに対する防御に取り組んでいる。特に、この種の攻撃に関する初期研究を開始するために、比較的単純なFAタスクである局所差分プライバシ（LDP）の下での頻度推定に焦点を当てる。LDPはクライアント側でデータの摂動を必要とするため、中央サーバはこのプロトコルを完全に制御することができず、悪意のあるクライアントがサーバ側における最終的な推定値を制御することを可能にする。本論文では、このLDPプロトコルを対象とした検証可能なLDPプロトコルを開発することで、ここでの攻撃を部分的に防ぐことができることを示した。本手法で提案するこのようなクライアント側の検証可能性は、将来的には連合学習を含むより複雑な連合タスクにおける攻撃を防ぐために拡張できる技術である。

第6章では、本論文における研究の包括的な社会的影響について論じている。さらに、実世界での連合学習の応用がすでに研究され始めているいくつかの具体的な分野における，本研究の利活用について論じている。

最後に，全体的な結論として、本論文は，厳密なプライバシ保証の欠如、脆弱性からの保護、信頼できる防御メカニズムの確立の難しさといった連合学習における基本的な問題を解決することにより、連合学習におけるプライバシとセキュリティの課題に対処するための高度なプライバシ保護技術を組み合わせた最先端の方法論を提示している。

（続紙 ２ ）

（論文審査の結果の要旨）

連合学習（FL）は、異なる機関やデバイスなどの間でプライバシを考慮した共同学習を可能にする新しい機械学習パラダイムとして知られる。大規模データ分析に対するプライバシ規制の懸念が高まる中、連合学習は学界と産業界の双方から注目を集めている。連合学習では、訓練データそのものではなく、最適化プロセスで使用される勾配情報のみが共有され、モデルを訓練する当事者はプライベートデータの管理コストから解放される。一方で，連合学習はプライバシとセキュリティの観点からまだ不十分なスキームであることも知られている。

本論文は，連合学習におけるプライバシとセキュリティを強化するための次の三つの課題に取り組んだ研究成果をまとめたものである。(1) 連合学習で訓練されたモデルに対する現実的かつ厳密なプライバシ保護．(2) Trusted Execution Environment (TEE)と呼ばれる特別なハードウェアを用いたセキュリティの高い連合学習の実現．(3) 連携分析(FA)における悪意のあるクライアントからの防御．具体的には、これら三つの各課題について以下の成果を上げている。

第一に，複数の組織による協調的な機械学習であるクロスサイロ連合学習において，訓練済みモデルに対してサイロを跨いだユーザレベルの差分プライバシを保証するための新しいアルゴリズムを開発し評価を行なった．また，提案アルゴリズムの有用性をさらに高めるための重み付け集約手法とそれを実現するためのプライベートなプロトコルを提案した．提案アルゴリズムは同等のプライバシ保証を達成するグループ差分プライバシに基づく従来の手法を大幅に上回るプライバシ・有用性のトレードオフを達成している．

第二に，TEEをサーバサイドに用いた連合学習において，アクセスパターンの漏洩というTEEの脆弱性を用いることで，連合学習がパラメータ交換の際にスパース化を利用する際にプライバシリスクが起こることを示した．実際にプライバシ攻撃が可能であることを示し，その攻撃に対する防御手法として，TEE内での安全な連合学習の集約操作のための新しいアルゴリズムを提案した．提案アルゴリズムがメモリアクセスパターンから何の情報も漏らさないことを理論的に示し，提案アルゴリズムが十分効率的であることを実験で確認した．

第三に，想定されたプロトコルから逸脱して連携分析タスクの出力の振る舞いを制御する悪意のあるクライアントに対する防御に取り組んだ．特に，局所差分プライバシを保証した頻度推定を行うプロトコルにおいて，悪意のあるクライアントからの攻撃の防御を行なった．検証可能なランダム化プロトコルを提案することで，悪意のあるクライアントからの攻撃が大幅に制限されることを示し，効率性に関しても問題ないことを実験で評価した．

以上、本論文は、連合学習のプライバシとセキュリティの強化に関する三つの重要な課題に対する解決手法をまとめたもので、学術上、および、実際上、寄与するところが少なくない。よって、本論文は博士（情報学）の学位論文として価値あるものと認める。また、令和６年１月１５日、論文内容とそれに関連した事項について試問を行った結果、合格と認めた。なお、本論文の令和７年３月２４日以降のインターネットでの全文公表についても支障が無いことを確認した。

要旨公開可能日：　令和６年　　６月　２４日以降