連合学習における局所更新情報の類似度に基づくビザンチン攻撃検出法

大野, 賢太 / OHNO, Kenta

(出版者 / Publisher)
法政大学大学院理工学研究科
(雑誌名 / Journal or Publication Title)
法政大学大学院紀要. 理工学研究科編
(巻 / Volume)
65
(開始ページ / Start Page)
1
(終了ページ / End Page)
7
(発行年 / Year)
2024-03-24
(URL)
https://doi.org/10.15002/00030737

連合学習における
局所更新情報の類似度に基づくビザンチン攻撃検出法

BYZANTINE ATTACK DETECTION
VIA SIMILARITY OF LOCAL UPDATES IN FEDERATED LEARNING

大野 賢太
Kenta OHNO
指導教員 山岸昌夫

法政大学大学院理工学研究科応用情報工学専攻修士課程

We propose a method to detect Byzantine attacks in federated learning, as well as a method for identifying clients repeating Byzantine attacks. The Byzantine attack is an attack that a malicious client sending false update information to the central server, in order to degradate learning performance at the central server. Firstly, we propose a method that utilizes a sum of cosine similarities of local update information to evaluate the authenticity of the local update sent by a client, through comparison with local updates from other clients. Next, by using this method, we also propose a method to identify clients repeatedly sending false update information. Furthermore, we clarify the importance of client grouping for enhance applicability of the proposed methods in scenarios where each client has diverse datasets. Finally, numerical experiments demonstrate the effectiveness of the proposed methods.

Keywords *: Machine learning, Federated learning, Privacy-preserving, Byzantine attack, Cosine similarity*

1. はじめに

コンピュータービジョンや自然言語処理などのタスクでは，高い性能を実現するモデルを実現するために，大規模なデータセットをニューラルネットワークモデルに学習させる手法が用いられている [1, 2, 3]．既存の機械学習手法の多くは，学習に用いるデータセットがあらかじめ一箇所に集約されている，あるいはある単一の計算機から容易に全データサンプルにアクセスが可能である状況を想定している．しかしながら，この想定は，データサンプルがプライバシーに関連する情報を含んでいる際には成立しない．例えば，モバイルデバイスでのテキスト入力履歴や，診療情報や電子カルテ，製造現場における機密データなどは有益な情報を含んでいるが個人情報であり，集約してデータベースを作成することは．一般的には困難である．

この問題に対する一つの解決策として, 連合学習 (Federated Learning)[4, 5] が注目されている．連合学習は分散型の機械学習手法の一種であり，中央サーバと複数のクライアントが連携して，クライアントの持つローカルデータを共有することなくグローバルモデルを学習することができる．

一方で，「悪意を持ったクライアントが偽の情報を中央サーバに送りつける攻撃」であるビザンチン攻撃により，グローバルモデルが性能劣化を引き起こすリスクについても指摘されており [6]，ビザンチン耐性のある連合学習の開発が検討されている．例えば，各クライアントから送信された更新情報の中央値を利用してグローバルモデルを更新する手法 [7] を含め様々な手法が提案されている [6, 7, 8, 9, 10]．これらの先行研究はビザンチン攻撃の影響を抑えることを主眼とし，攻撃を行なっているクライアントの検出を目指していない．

本稿では，連合学習におけるビザンチン攻撃を検出する手法，および，ビザンチン攻撃を繰り返すクライアントを検出する手法を提案する．まず，更新情報のコサイン類似度の和を活用し，クライアントが送信してきた更新情報の真偽を他クライアントの更新情報との比較により判定する手法を提案している．この手法を活用し，偽の更新情報を繰り返し送りつけてくるクライアントを特定する手法を提案している．さらに，各クライアントが多様なデータセットを保持している状況下において，提案手法を適用するためにはクライアントのグループ化が重要であることを明らかにしている．最後に，数値実験により，提案手法の有効性を確認している．

(数学的表記) $\mathbb{R}, \mathbb{N}$ をそれぞれ実数全体，自然数全体の集合とする．集合 S に対して，その濃度を $|S|$ で表す．太字はベクトルまたは行列を表す．$(\cdot)^\top$ は転置を表す．$d(\in \mathbb{N})$ 次元ユークリッド空間を $\mathbb{R}^d$ で表す．$\mathbb{R}^d$ の内積を $\langle \mathbf{x}, \mathbf{y} \rangle := \mathbf{x}^\top \mathbf{y}$ $(\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d)$ とし，この内積から誘導されるノルムを $\|\mathbf{x}\| := \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$ で表す．

2. 連合学習

(1) 概要

連合学習は，データのプライバシー保護を目的とした分散型の機械学習手法の一種であり中央サーバと複数の

クライアントが連携して，クライアントの持つローカルデータを共有することなくグローバルモデルを学習を行う．具体的には以下の Step A〜C をまとめて 1 エポックとしてグローバルモデルの反復更新を行う．

Step A. 中央サーバは各クライアントにグローバルモデルのパラメータを送信する．

Step B. 各クライアントは, ローカルデータを用いてグローバルモデルの更新先を計算し, 中央サーバに送信する．

Step C. 中央サーバは, 各クライアントから送られてきた更新先の情報を集約しグローバルモデルを更新する．

以下，クライアント数を $K \in \mathbb{N}$, 各クライアントの損失関数を $F_k \colon \mathbb{R}^d \to \mathbb{R}$ $(k = 1, 2, \ldots, K)$ とする．中央サーバは各クライアントの損失関数の平均の最小化

$$\underset{\mathbf{w} \in \mathbb{R}^d}{\text{Minimize}} \ \frac{1}{K} \sum_{k=1}^{K} F_k(\mathbf{w})$$

を目指すものとする．また，初期のグローバルモデルのパラメータを $\mathbf{w}_0 \in \mathbb{R}^d$ とし，t 回目のグローバルモデルのパラメータを $\mathbf{w}_t \in \mathbb{R}^d$ とする．

連合学習アルゴリズム Federated Averaging [5]（以下, FedAvg）では，j 番目のクライアントは，Step B において，損失関数 F_k の勾配 ∇F_k を用いてグローバルモデル $\mathbf{w}_t$ の更新先

$$\mathbf{w}_{t+1}^{(k)} := \mathbf{w}_t - \alpha \nabla F_k(\mathbf{w}_t) \tag{1}$$

を計算し中央サーバへ送信する．中央サーバは，Step C において，以下の式を用いて，各クライアントからの情報を集約し，グローバルモデルのパラメータを次の式を更新する．

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \frac{\mu}{K} \sum_{k=1}^{K} \left(\mathbf{w}_t - \mathbf{w}_{t+1}^{(k)} \right) \tag{2}$$

ここで $\mu > 0$ は学習率である．この更新式 (2) は，(1) を用いて

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \frac{\alpha\mu}{K} \sum_{k=1}^{K} \nabla F_k(\mathbf{w}_t)$$

と表すことができる．

(2) ビザンチン攻撃

典型的なビザンチン攻撃として, 絶対値の大きい値を送信するもの, ローカルデータのラベルを偽ラベルに付け替えて更新先を計算し送信するもの，更新情報の符号を反転させて送信するもの (Sign-flipping 攻撃) などが検討されている [6][7]．例えば，FedAvg (2) を想定した Sign-flipping 攻撃では，攻撃を行うクライアント（攻撃者）の添字集合を $\mathcal{B}_\star \subset \{1, 2, \ldots, K\}$ とすると，各クライアントの更新先は

$$\mathbf{w}_{t+1}^{(k)} := \begin{cases} \mathbf{w}_t + \alpha \nabla F_k(\mathbf{w}_t) & \text{if } k \in \mathcal{B}_\star, \\ \mathbf{w}_t - \alpha \nabla F_k(\mathbf{w}_t) & \text{otherwise} \end{cases} \tag{3}$$

となる．本稿では，攻撃者ではないクライアントを正規クライアントと呼び, その添字集合を $\mathcal{N}_\star$ とする（正規クライアントは (1) で更新していることに注意されたい）．

FedAvg を含む多くの連合学習では, Step C の情報集約に平均を採用しており, この平均処理がビザンチン攻撃に対して影響を受けやすいことが知られている [6]．これを踏まえ，ビザンチン攻撃への耐性を実現するために，Step C の更新式 (2) において，各クライアントの更新情報

$$\mathbf{g}_t^{(k)} := \mathbf{w}_t - \mathbf{w}_{t+1}^{(k)} \in \mathbb{R}^d \quad (k = 1, 2, \ldots, K) \tag{4}$$

の平均値の代わりに，中央値を利用する方法 [7] やバッチ平均の幾何中央値を利用する方法 [6] が提案されている．

(中央値を利用する方法 [7]) Step C の更新式 (2) の代わりに,

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \mu \operatorname{med} \left\{ \mathbf{g}_t^{(1)}, \mathbf{g}_t^{(2)}, \ldots, \mathbf{g}_t^{(K)} \right\}$$

によりグローバルモデルの更新を行う．ここで，中央値 med は成分ごとに計算される．[*1]

(バッチ平均の幾何中央値を利用する方法 [6]) Step C の更新式 (2) の代わりに,

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \mu \operatorname{gmed} \left\{ \frac{1}{|\mathfrak{B}_l|} \sum_{k \in \mathfrak{B}_l} \mathbf{g}_t^{(k)} \ \middle| \ l = 1, 2, \ldots, m \right\}$$

によりグローバルモデルの更新を行う．[*2] ここで，K 個のクライアントを同サイズの m 個のバッチに分割し，各バッチに属するクライアントの添字集合を $\mathfrak{B}_1, \mathfrak{B}_2, \ldots, \mathfrak{B}_m \subset \{1, 2, \ldots, K\}$ としている．バッチサイズは $|\mathfrak{B}_1| (= |\mathfrak{B}_2| = \cdots = |\mathfrak{B}_m|)$ である．

この他にも，トリム平均を利用する方法 [7] やビザンチン耐性と学習の収束保証を両立する方法 [8]，2.3 節で説明する non-iid データの状況に対応した手法 [9, 10] などが提案されている．

(3) non-iid データ

本稿では，[3] の表現に従って，各クライアントのデータの背後にある確率分布がクライアントごとに異なっている状況を「non-iid(non-independent and identically distributed) データが与えられている」と表現する．逆に，背後にある確率分布が同一の状況を「iid データが与えられている」と表現する．連合学習では各クライアントがローカルにデータ収集を行っているため，クライアントの環境や収集方法によって，データに偏りが生じる状況（つまり non-iid データの状況）は十分に想定される．non-iid データは学習アルゴリズムの収束を遅くすること [9, 11] を含め様々な困難の要因として知られている [10, 10].

[*1] [7] では，グローバルモデルのパラメータ空間 $\mathcal{W} \subset \mathbb{R}^d$ を導入し，$\mathbf{w}_t (t = 1, 2, \ldots)$ が $\mathcal{W}$ 内にとどまるよう距離射影を用いる手法を提案している．本稿では，$\mathcal{W} = \mathbb{R}^d$ である状況を想定し，距離射影を省略している．

[*2] 与えられた $\{\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(m)}\} \subset \mathbb{R}^d$ に対して，幾何中央値 $\operatorname{gmed}\{\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(m)}\} \in \mathbb{R}^d$ は集合 $\operatorname{argmin}_{\mathbf{v} \in \mathbb{R}^d} \sum_{l=1}^{m} \|\mathbf{v} - \mathbf{v}^{(l)}\|$ に属するベクトルとして定義される [6, (6)].

3. 提案手法：コサイン類似度を用いたビザンチン攻撃検出法

本稿では，

- iid データが割り当てられており，
- 全クライアント内の攻撃者の割合が半分より少なく，
- 攻撃者は sign-flipping 攻撃 (3) を行う

状況を想定した上で，Step C において，中央サーバが受け取った各クライアントからの更新情報 $\mathbf{g}_t^{(j)} \in \mathbb{R}^d$ $(j = 1, 2, \dots K)$ の相互のコサイン類似度を基準に，各更新情報の真偽を判定する手法を提案する．

更新情報 $\mathbf{g}_t^{(j)}$ が攻撃ではない場合 $(j \in \mathcal{N}_\star)$，直感的には，他の攻撃ではない更新情報と方向が似ていることが期待できる．つまり，更新情報 $\mathbf{g}_t^{(j)}$ と更新情報 $\mathbf{g}_t^{(k)}$ $(k = 2, 3 \dots K)$ のコサイン類似度

$$\mathrm{S_C}\left(\mathbf{g}_t^{(j)}, \mathbf{g}_t^{(k)}\right) := \begin{cases} \frac{\langle \mathbf{g}_t^{(j)}, \mathbf{g}_t^{(k)} \rangle}{\|\mathbf{g}_t^{(j)}\| \|\mathbf{g}_t^{(k)}\|} & \text{if } \|\mathbf{g}_t^{(j)}\| \neq 0 \neq \|\mathbf{g}_t^{(k)}\|, \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

は，$k \in \mathcal{N}_\star$ では正となることが期待できる．同様の議論で，更新情報 $\mathbf{g}_t^{(j)}$ が攻撃である場合 $(j \in \mathcal{B}_\star)$，コサイン類似度 (5) は $k \in \mathcal{N}_\star$ では負となることが期待できる．そのため，コサイン類似度の和 $\sum_{k=1}^K \mathrm{S_C}\left(\mathbf{g}_t^{(j)}, \mathbf{g}_t^{(k)}\right)$ が，更新情報 $\mathbf{g}_t^{(j)}$ が攻撃であるか否かを測る指標となることが期待できる．この知見を用いて，閾値 $\tau \in \mathbb{R}$ に対して

$$\sum_{k=1}^K \mathrm{S_C}\left(\mathbf{g}_t^{(j)}, \mathbf{g}_t^{(k)}\right) < \tau \tag{6}$$

が成立する際には，更新情報 $\mathbf{g}_t^{(j)}$ が攻撃であると判定することを提案する．

また，上記条件 (6) が満たされたエポック（攻撃検出回数）の割合

$$R_j(t) := \frac{1}{t} \left| \left\{ \hat{t} \in \{1, 2, \dots, t\} \;\middle|\; \sum_{k=1}^K \mathrm{S_C}\left(\mathbf{g}_{\hat{t}}^{(j)}, \mathbf{g}_{\hat{t}}^{(k)}\right) < \tau \right\} \right| \tag{7}$$

が 1 に近いクライアント j は，攻撃を繰り返しているクライアントである．そのため，攻撃者の判定基準として，閾値 ρ を用いて $R_j(t) > \rho$ となるクライアント j を攻撃者と判定することを提案する．

(クライアントのグループ化による non-iid データへの適用) 上記の提案手法は，ある種の non-iid データに対しては，クライアントをグループへ分割する工夫を加えることで良好な判別性能を実現することが期待できる．具体的には，クライアント全体では non-iid データが与えられていたとしても，クライアントを m 個のグループへ分割し，各グループ内では iid データが与えられていると見做せる場合，上記の提案法をグループごとに適用することで，攻撃者の良好な判定が可能となると考えられる．すなわち，事前に，各クライアントが保持する各ラベルのデータ数を中央サーバへ送信し，その情報によりクライアントのグループへの割り当てを決定する．そして，それぞれのグループに属するクライアントの添字集合を $\mathcal{K}_1, \mathcal{K}_2, \dots, \mathcal{K}_m$ とし，$l = 1, 2, \dots, m$ に対して，クライアント $j \in \mathcal{K}_l$ からの更新情報が攻撃であるかどうかを

$$\sum_{k \in \mathcal{K}_l} \mathrm{S_C}\left(\mathbf{g}_t^{(j)}, \mathbf{g}_t^{(k)}\right) < \tau \tag{8}$$

により判定し，攻撃を繰り返しているクライアントであるかどうかを

$$\hat{R}_j(t) := \frac{1}{t} \left| \left\{ \hat{t} \in \{1, 2, \dots, t\} \;\middle|\; \sum_{k \in \mathcal{K}_l} \mathrm{S_C}\left(\mathbf{g}_{\hat{t}}^{(j)}, \mathbf{g}_{\hat{t}}^{(k)}\right) < \tau \right\} \right| \tag{9}$$

により判定する手法は良好な検出を実現すると考えられる．

4. 数値実験

本実験では MNIST データセット [12] を利用した手書き数字画像の分類問題を対象とした．MNIST データセットは「0」〜「9」の 10 種類の手書き数字画像から成る．60000 枚の学習データと 10000 枚のテストデータの合計 70000 枚の 28×28 のグレースケール画像で構成されている．本実験では，「0」と「1」の手書き数字画像のデータのみを使用した．

グローバルモデルは Softmax 回帰モデルとし，クライアント数は $K = 30$，クライアントの更新時のステップサイズは $\alpha = 1$, 学習率は $\mu = 0.01$，提案手法の閾値は $\tau = -2$ とした．

以下の 3 条件で性能を比較した．

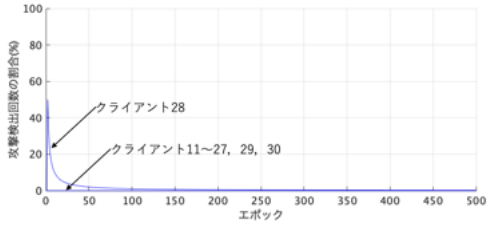
条件 A 全クライアントが攻撃者でないときに学習した場合
条件 B 攻撃者がいてもそのまま学習した場合
条件 C 攻撃者を特定し対処して学習した場合

条件 B と条件 C では，クライアントに含まれる攻撃者の人数は 10 人とし，攻撃者は sign-flipping 攻撃 (3) を行うこととした．条件 C では，$t = 500$ エポックの時点で (7) の $R_j(t) \times 100[\%]$（または (9) の $\hat{R}_j(t) \times 100[\%]$）が 98% を超えるクライアント j を攻撃者と判定する手法を用いた．以後，$R_j(t)$ を用いる手法を「提案法」と呼び，$\hat{R}_j(t)$ を用いる手法を「グループ化付き提案法」と呼ぶ．また，攻撃者と判定したクライアントに対して，501 エポック以降の学習では，中央サーバは更新情報の符号を反転して学習を進めることとした．

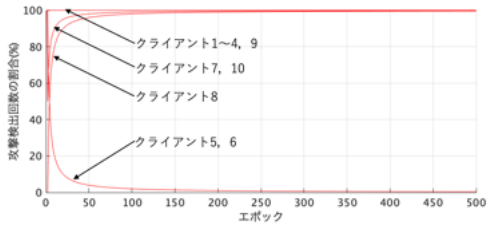
以下, それぞれの条件で得られた各エポック t のグローバルモデルのパラメータを $\mathbf{w}_t^{\mathrm{A}}, \mathbf{w}_t^{\mathrm{B}}, \mathbf{w}_t^{\mathrm{C}} \in \mathbb{R}^d$ と表記し，クライアントからの更新情報を $\mathbf{g}_t^{(k),\mathrm{A}}, \mathbf{g}_t^{(k),\mathrm{B}}, \mathbf{g}_t^{(k),\mathrm{C}} \in \mathbb{R}^d$ $(k = 1, 2, \dots, K)$ と表記する．

また，データの割り当てを 2 通りで行なった．

データ $\mathcal{D}_{\mathrm{iid}}$： iid データが割り当てられている状況を



(A) 正規クライアント.



(B) 攻撃者.

図 1　提案法による攻撃検出回数の割合（データ $\mathcal{D}_{\mathrm{iid}}$ の場合）.

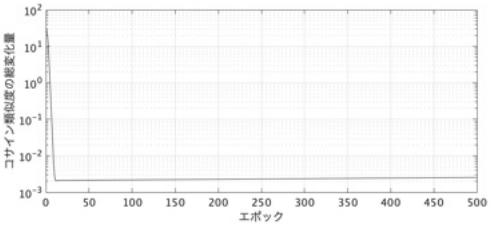
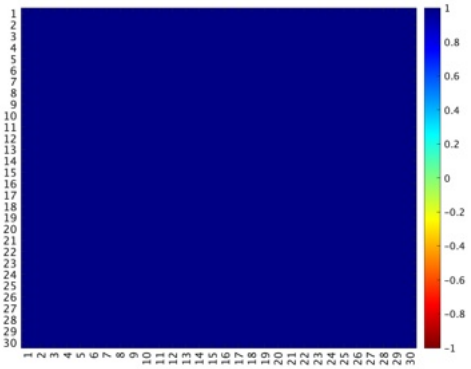
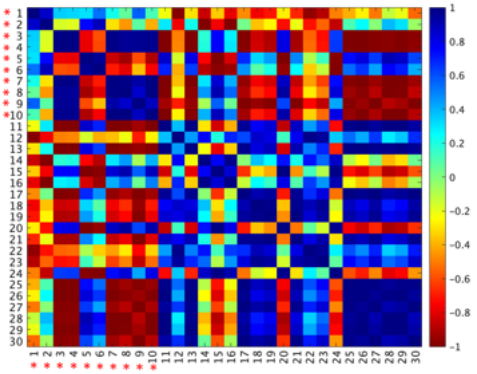


図 2　コサイン類似度の総変化量 ($\mathcal{D}_{\mathrm{iid}}$ の場合).



(A) 攻撃者がいない場合 (条件 A).



(B) 攻撃者がいる場合 (条件 B, 条件 C).

赤の*印は，攻撃者に対応する行（および列）を示している.

図 3　コサイン類似度行列

再現するため，「0」と「1」の手書き数字画像のデータから最後の 5 個を除いて 30 等分し，各クライアントに割り当てた．このデータの割り当ての際には，攻撃者を $\mathcal{B}_\star = \{1, 2, \ldots, 10\}$ とした.

データ $\mathcal{D}_{\mathrm{non-iid}}$： non-iid データが割り当てられている状況を再現するため，手書き数字画像データ「0」のみを持っているクライアント (1〜13),「0」と「1」を同数持っている持っているクライアント (14〜16),「1」のみを持っているクライアント (17〜30) からなるようにデータを割り当てた．以降，それぞれの添字集合を $\mathcal{K}_0, \mathcal{K}_{01}, \mathcal{K}_1$ とする．このデータの割り当ては，全体としては non-iid データであるが，$\mathcal{K}_0$ に属するクライアントだけに注目すると iid データが割り当てられていると見做すことができ，同様に $\mathcal{K}_{01}$，$\mathcal{K}_1$ に属するクライアントそれぞれについても iid データが割り当てられていると見做すことができる．4.2 節において，このグループを用いたグループ化付き提案法の性能を確認する．このデータの割り当ての際には，攻撃者を $\mathcal{B}_\star = \{1, 2, \ldots, 6, 14, 17, 18, 19\}$ とした.

(1)　データ $\mathcal{D}_{\mathrm{iid}}$ での実験結果と考察

まず，提案法の攻撃者の検出性能を確認する．図 1 は，条件 C において，提案法による攻撃検出回数の割合 (7) の推移を図示したものである．図 1 より，提案法は攻撃者 10 人中 8 人を特定できており，クライアント 5，6 を見逃したことが分かる.

見逃した理由を明確にするために，更新情報のコサイン類似度を観察した．図 2 は，条件 B（および条件 C）において，コサイン類似度の総変化量

$$\sum_{j=1}^{K}\sum_{k=1}^{K}\left|\mathrm{S_C}\left(\mathbf{g}_t^{(j)}, \mathbf{g}_t^{(k)}\right) - \mathrm{S_C}\left(\mathbf{g}_{t-1}^{(j)}, \mathbf{g}_{t-1}^{(k)}\right)\right| \qquad (10)$$

を図示したものである．図 2 より，15 エポック程度で総変化量が 3×10^{-3} 以下に減少しており，15 エポック以降はコサイン類似度がほぼ変化していないことが確認できる．また，図 3(A)(B) は，各条件のそれぞれの $t = 500$ エポックにおけるコサイン類似度行列 $\mathbf{S}_{\mathrm{C}}^{\mathrm{A}}(t), \mathbf{S}_{\mathrm{C}}^{\mathrm{B}}(t) \in \mathbb{R}^{K\times K}$ を図示したものであり，それぞれの j 行 k 列の成分は

$$[\mathbf{S}_{\mathrm{C}}^{\mathrm{A}}(t)]_{j,k} := \mathrm{S_C}\left(\mathbf{g}_t^{(j),\mathrm{A}}, \mathbf{g}_t^{(k),\mathrm{A}}\right)$$
$$[\mathbf{S}_{\mathrm{C}}^{\mathrm{B}}(t)]_{j,k} := \mathrm{S_C}\left(\mathbf{g}_t^{(j),\mathrm{B}}, \mathbf{g}_t^{(k),\mathrm{B}}\right)$$

である[*3]．図 3 より，攻撃者であるクライアント 5, 6 については，正の成分が多く含まれており，正規クライアントの行に類似していることが確認できる．このことから，クライアント 5, 6 が保持しているデータが「0」か「1」のどちらかを多く含んでおり，「iid データが割り当

[*3] コサイン類似度行列は対称行列であり，条件 C で得られるコサイン類似度行列 $\mathbf{S}_{\mathrm{C}}^{\mathrm{C}}(t)$ は $\mathbf{S}_{\mathrm{C}}^{\mathrm{B}}(t)$ と同一であることに注意されたい

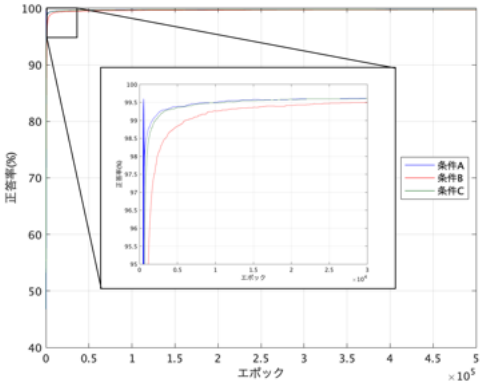


図 4　正答率の推移．条件 C では提案法を用いた．

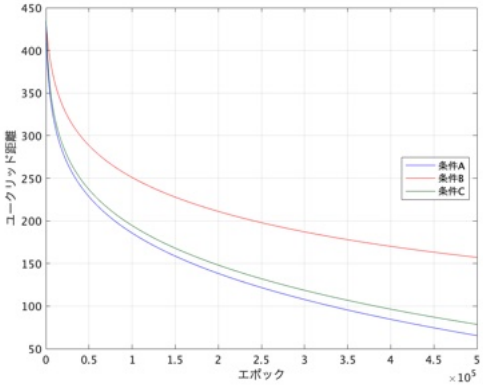


図 5　パラメータのユークリッド距離の推移．条件 C では提案法を用いた．

てられている」という仮定が十分に成立していなかったと予想される．

図 4 は条件 A,B,C の実験での正答率

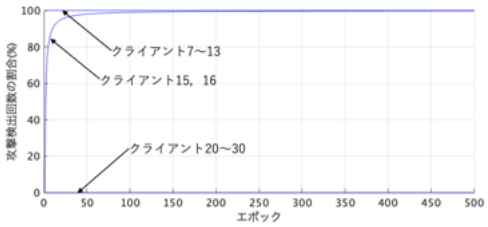
$$\left(\frac{\text{正解のデータの個数}}{\text{学習データの個数}} \times 100\ [\%]\right)$$

の推移を表している．図 4 より，条件 A,B,C で得られた正答率の推移にはほとんど差がないことがわかる．

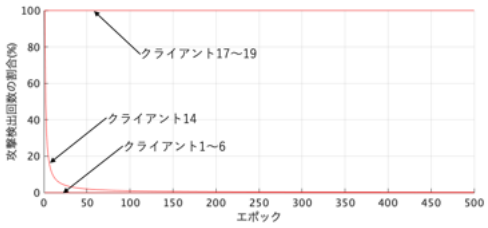
図 5 は，条件 A で得られた $\mathbf{w}^{\mathrm{A}}_{10^6}$ と各エポックで得られたパラメータベクトル $\mathbf{w}^{\mathrm{A}}_t, \mathbf{w}^{\mathrm{B}}_t, \mathbf{w}^{\mathrm{C}}_t$ それぞれとのユークリッド距離 $\|\mathbf{w}^{\mathrm{A}}_t - \mathbf{w}^{\mathrm{A}}_{10^6}\|, \|\mathbf{w}^{\mathrm{B}}_t - \mathbf{w}^{\mathrm{A}}_{10^6}\|, \|\mathbf{w}^{\mathrm{C}}_t - \mathbf{w}^{\mathrm{A}}_{10^6}\|$ の推移を図示したものである．図 5 より，最終的な結果として条件 A と条件 C でのユークリッド距離には 10 程度の差しかなく，条件 B では 100 程度の劣化が確認された．つまり，条件 C において，ビザンチン攻撃と攻撃者を検出し，適切に対処することで，ビザンチン攻撃が学習に与える悪影響を抑圧できることが確認できた．

(2)　データ $\mathcal{D}_{\mathrm{non-iid}}$ での実験結果と考察

条件 C で提案法を適用した際の攻撃検出回数の割合 (7) の推移を図 6 に示す．図 6 より，攻撃者と判定されたのはクライアント 7〜13, 15〜19 であった．この内，実際に攻撃者だったのは 17〜19 のみである．これより，non-iid データが割り当てられている状況では，提案手



(A) 正規クライアント．



(B) 攻撃者．

図 6　提案法による攻撃検出回数の割合 ($\mathcal{D}_{\mathrm{non-iid}}$ の場合)

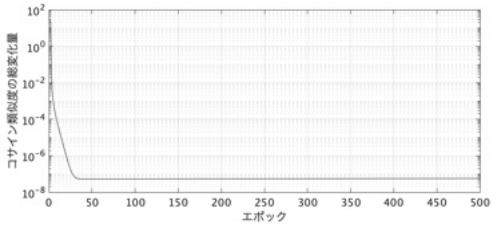


図 7　コサイン類似度の総変化量 ($\mathcal{D}_{\mathrm{non-iid}}$ の場合)．

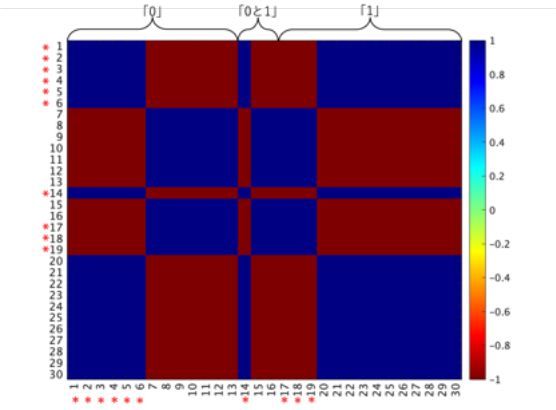
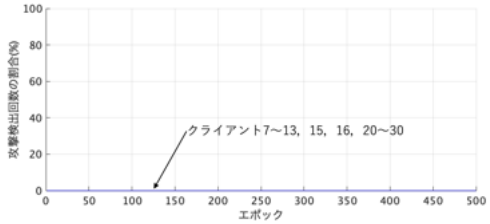


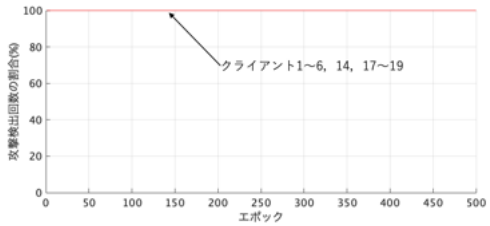
図 8　コサイン類似度行列 ($\mathcal{D}_{\mathrm{non-iid}}$ の場合)．赤の*印は，攻撃者に対応する行（および列）を示している．図上部の「0」，「0 と 1」，「1」はそれぞれ，$\mathcal{K}_0, \mathcal{K}_{01}, \mathcal{K}_1$ に対応する行（および列）を示している．

法の検出精度が低下することが確認できた．

なぜ誤検知が起こってしまったのかを明確にするため，更新情報のコサイン類似度を観察する．図 7 は，条件 B（および条件 C）において，更新情報のコサイン類似度の総変化量を図示したものである．iid データの場合と同様に，40 エポック以降は更新情報のコサイン類似度がほぼ変化していないことが確認できる．図 8 は，条件 B（および条件 C）の 500 エポックにおける更新情報のコサイン類似度行列を図示したものである．図 8 よ



(A) 正規クライアント.



(B) 攻撃者.

図 9　グループ化付き提案法による攻撃検出回数の割合 ($\mathcal{D}_{\text{non-iid}}$ の場合)

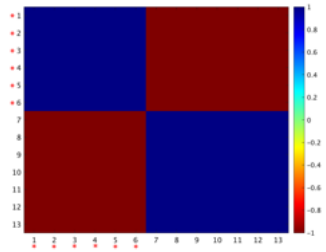
り，攻撃者（1〜6,14) に対応する行にも正の値が多く含まれており，さらには，正規クライアント（7〜13,15,16）に対応する行にも負の値が多く含まれていることがわかる．具体的には，「0」のみを持つ攻撃者の一例としてクライアント 6 に着目すると，「1」のみを持つ正規クライアント 20〜30 とのコサイン類似度がほぼ 1 になっていることが分かる．攻撃者と正規クライアントのコサイン類似度が負になることを想定していたが，non-iid データに対しては，その想定が成立していないことが分かる．

そこで，グループ化付き提案法を適用し，その有効性を確認する．条件 C でグループ化付き提案法を適用した際の攻撃検出回数の割合 (9) の推移を図 9 に示す．図 9 から攻撃者 10 人を全て特定できており，その有効性を確認できた．

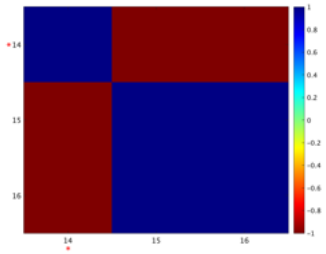
図 10 は，条件 C の 500 エポックにおける更新情報のコサイン類似度行列を $\mathcal{K}_0, \mathcal{K}_{01}, \mathcal{K}_1$ のグループごとに図示したものである．図 10 より，グループごとのコサイン類似度行列では，攻撃者に対応する要素に負の値が多く含まれており，グループ化付き提案法で正しく判別できることが確認できる．

5. おわりに

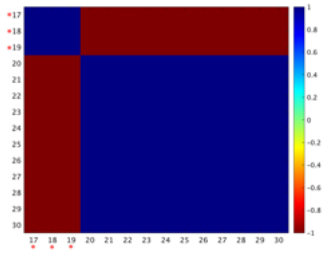
連合学習においてクライアントから集約した更新情報のコサイン類似度を用いてビザンチン攻撃を検出する方法を提案した．今後の課題として，多クラス分類問題での有効性の検証，既存の攻撃者の判定法との比較などが考えられる．



(A) 0 のみを持っているクライアント ($\mathcal{K}_0$ に対応).



(B) 0 と 1 を持っているクライアント ($\mathcal{K}_{01}$ に対応).



(C) 1 のみを持っているクライアント ($\mathcal{K}_1$ に対応).

図 10　コサイン類似度行列 ($\mathcal{D}_{\text{non-iid}}$ の場合)．赤の*印は，攻撃者に対応する行（および列）を示している．(A)(B)(C) は，それぞれ図 8 のコサイン類似度行列の対角ブロックであることに注意されたい．

参考文献

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, 2012.

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, 2015.

[3] 米谷竜, "連合学習入門 —基本的なアプローチと典型的な課題—," **精密工学会誌**, 2021.

[4] J. Konečnỳ, H. B. McMahan, F. Yu, A. Richtárik, P.and Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," *arXiv preprint arXiv:1610.05492*, 2016.

[5] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Aguera, "Communication-efficient learning of deep networks from decentralized data," *International Conference on Artifical Intelligence and Statistics*, 2016.

[6] J. Xu, Y. Chen, and L. Su, "Distributed statistical machine learning in adversarial settings: Byzantine gradient descent," pp. 1–44, 2017.

[7] D. Yin, Y. Chen, K. Ramchandran, and P. L. Bartlett, "Byzantine-robust distributed learning: Toward optimal statistical rates," *CoRR*, 2018.

[8] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," *Advances in Neural Information Processing Systems*, 2017.

[9] S. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. Suresh, "Scaffold: Stochastic controlled averaging for federated learning," *International Conference on Machine Learning*, 2020.

[10] L. Gao, H. Fu, L. Li, Y. Chen, M. Xu, and C. Z. Xu, "Feddc: Federated learning with non-iid data via local drift decoupling and correction," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[11] X. Li, K. Huang, Z. Z. W. Yang, and S. Wang, "On the convergence of fedavg on non-iid data," *International Conference on Learning Representations*, 2020.

[12] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, 1998.