Vol.2020-DPS-184 No.1
Vol.2020-EIP-89 No.1
2020/9/10

IoT 機器のプライバシー保護のための決定木分散学習システム

山下優衣[1]　田谷昭仁[1]　戸辺義人[1]

概要：社会の IT 化が進みさまざまな IoT 機器が普及している．IoT 機器が独立して学習するより，データや学習器を共有することで性能向上が期待できるが，個々の持っているオリジナルデータを共有することはプライバシーの観点から懸念される．そこで，データを直接共有せずに個々の端末で学習を行い作成した学習器のみを共有することでプライバシーの保護を考える．本論文では学習方法にランダムフォレストを用い，決定木を共有し学習をするアルゴリズムの提案をする．

キーワード： IoT，プライバシー保護，決定木，分散学習

Distributed Learning System with Decision Tree for Privacy-Preserving of IoT Devices

YUI YAMASHITA[†1]　AKIHITO TAYA[†1]
YOSHITO TOBE[†1]

Abstract: Recently, various Internet of Things (IoT) devices have spread in our daily lives with the advancement of Information Technology (IT) in society. This paper considers machine-learning-available IoT devices. It is necessary to cooperate with each other rather than learn independently in order to improve its performance. However, sharing local data directly is difficult because those data may contain private information. Therefore, this paper proposes an algorithm preserving users' privacy by sharing only machine learning models trained on each device without doing local data. This paper applies the cooperative learning method to decision tree, which is suitable for IoT devices because of their poor computational resources.

Keywords: IoT, Privacy-Preserving, Decision Tree, Distributed Learning

1. はじめに

社会の IT（Information Technology）化が進み，私たちの生活に身近なアプリケーションやデバイスだけでなく，産業においても IoT（Internet of Things）が普及している．最近では，エアコンや照明の自動調整，音声コントロールなどが可能な生活家電を用いたスマートホーム，環境問題をはじめとするさまざまな問題をセンサなどから得られる情報を用いて対処する都市のあり方であるスマートシティなどに注目が集まっている．これらはセンサから取得した情報を収集，分析することでさまざまな機能が利用でき，便利で快適な暮らしを実現することを目指している．

本稿では，1 台の機器で機械学習を行うのではなく，複数台の IoT 機器間での協調学習を検討する．協調学習の手法を用いた研究の 1 つに複数の機器間でデータの共有を行い，人の行動予測フレームワークを作成しているものがある [1]．このフレームワークは，複数人のスマートフォンから得られたセンシングデータを共有して一般的な行動モデルを抽出することにより，個人レベルの行動予測の精度が向上している．しかし，このように汎用モデルを作成することは，行動予測の精度が向上する一方で，プライバシーを含む可能性のある個人のセンシングデータを共有しなければならないという点が課題となる．

そこで，プライバシーを含む個々の学習データを共有しない協調学習が求められる．プライバシー保護を目的とした分散学習アルゴリズムに FL（Federated Learning）が提案されている [2]．FL はローカルデータを共有せずに，複数の分散された機器でニューラルネットワークモデルを構築し，そのモデルを共有することでデータ漏洩を防ぐ．しかし，深層学習は計算が複雑であるため，スマートウォッチや Raspberry Pi などの計算リソースが不十分な IoT 機器には適していない. このような機器で処理を実行するために，RF（Random Forest）などの比較的簡単な学習方法を適用するのが望ましい．

以上の背景から，ローカルデータではなく FL のような学習モデルのみを共有し，その学習方法として RF を使用した協調学習アルゴリズムを提案する．提案アルゴリズムでは，各端末で RF によりローカルデータから複数の決定木を学習し，その一部を他の隣接端末と交換する．交換後の決定木の集合をモデルとして予測等に利用する．

本稿では，第 2 章で関連研究について述べる．第 3 章では検討モデルについて述べ, 第 4 章では提案アルゴリズム，第 5 章では評価結果および考察，第 6 章では本論文の結論

1 青山学院大学
Aoyama Gakuin University

として，今後の課題と発展について述べる．

2. 関連研究

FL はデータをローカルに保持しながら分散機械学習を行う方法であり，携帯電話ユーザーの感情や行動の学習，自動車への歩行者の行動パターンの適応，ウェアラブルデバイスを用いた健康状態の予測など，多くの場面で使用されている [3]．さまざまなアプリケーションにおいて重要な役割を果たしているが，課題も存在する． その 1 つが通信コストである [4]．一般的な FL は，CNN（Convolutional Neural Network）のような複雑な深層学習モデルを使用しており，このようなモデルの更新には数百万のパラメータが含まれる場合がある．更新の次元が高くなると通信コストが高くなり，トレーニングが困難になる可能性がある．

IoT を利用した健康管理システムのビッグデータ分析に RF を使用している研究がある [5]．分析を行う際，ヘルスデータの分類に RF を使用した提案手法を用いると，既存の手法と比較して提案手法が最も良い精度を達成している．RF は高次元データを処理でき計算が比較的高速なため，さまざまな問題に対して優れたパフォーマンスを発揮することができる．さらに，過学習を引き起こす可能性も低いという利点がある [6]．また，RF はウェアラブルデバイスによる行動分類にも使用されている [7]．特徴を正規化する必要がなく，特徴選択手順が必要ないため，RF が分類手法に選択されている．

本稿では，プライバシーデータをローカルに保持する協調学習を行うために，提案アルゴリズムに RF を適用する．

3. モデル

本章では，提案アルゴリズムを適応する想定モデルについて述べる．提案アルゴリズムでは，機器同士がマルチホップネットワークで接続されていることを想定する．この接続方法では，各機器は接続されている機器とのみ情報を共有することが可能である．この想定モデルの概要を図 1 に示す．スマートフォンなどの IoT 機器は，RF によってモデルをトレーニングするためのローカルデータを所有している．ローカルデータはプライバシー情報が含まれる可能性があるため共有できず，代わりに機器同士はプライバシー情報を含まない予測モデルである決定木を他の端末と共

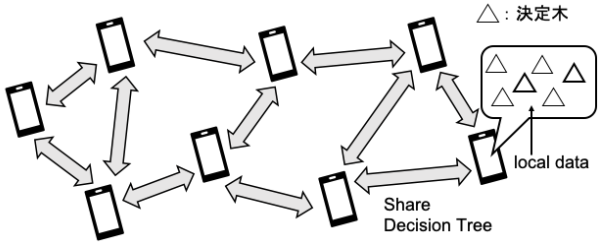

図 1 接続方法のモデル

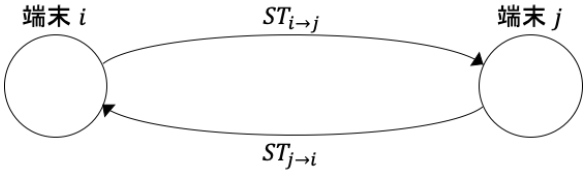

図 2 木の共有モデル

有することが可能となる．

スマートフォンなどの IoT 機器iが所持しているローカルデータをD_i，使用する端末数をNと定義したとき，各端末iはD_iを用いて RF によりローカルの予測モデルM_iをトレーニングする．M_iはn本の弱学習器である決定木で構成される．以下では，学習した決定木を木と呼び，端末iのp番目の木を$T_{i,p}$と表し，M_iは以下のように定義される．

$$M_i = \{T_{i,1}, T_{i,2}, T_{i,3}, \dots, T_{i,n}\} \quad (1)$$

nは全端末で共通とする．各端末は隣接端末間で情報を共有することが可能であるが，ローカルデータであるD_iは交換することができない．次章において，D_iを交換することなく協調学習を行うアルゴリズムを提案する．

4. 実装

本研究は，ローカルデータを共有せず木を交換することにより高精度なモデルの取得することを目標とする．本章では，プライバシーを保護する協調学習のアルゴリズムについて述べる．

端末iが端末jと隣接しているとき，2 端末i，j間でいくつかの木を交換する．交換する木の本数をmとしたとき，端末iは端末jへ渡す木を予測モデルM_iからm本選択する．端末iから端末jへ共有する木の集合を$ST_{i\to j}$と表し，以下のように定義する．

$$ST_{i\to j} \leftarrow \left\{T_{i,a_1}, \dots, T_{i,a_m} \middle| a_p(\leq n)\ are\ randomly\ sampled\ integer\right\} \quad (2)$$

木の共有の流れを図 2 に示す．端末iにおける隣接端末数をJ_iとすると，端末iは接続されている全ての端末同士と木を共有する必要があるため，J_i台全ての隣接端末に対して$ST_{i\to j}$を決定する．また，木を共有するとき，端末iはM_iからJ_i台の隣接端末より受け取る木の総数と等しい本数の木を削除する．この木の削除は，接続されている端末数や交換する木の本数が増えても，各端末のメモリが増えないようにするために行う．端末iから削除する木の本数をl_i，削除する木の集合をDT_iと表し，以下のように定義する．

$$l_i = m \times J_i \quad (3)$$

$$DT_i \leftarrow \left\{T_{i,b_1}, \dots, T_{i,b_{l_i}} \middle| b_p(\leq n)\ are\ randomly\ sampled\ integer\right\} \quad (4)$$

交換後，端末iが持つ木の本数は再びn本となり，新しく予測モデルM'_iを生成する．

$$M'_i \leftarrow M_i \cup \left(\bigcup_{j=1}^{J_i} ST_{j\to i}\right) / DT_i \quad (5)$$

アルゴリズム 1. Distributed Random Forest

1: **for** $i \in N$ **do**
2: $M_i \leftarrow \{T_{i,1}, T_{i,2}, T_{i,3}, \dots, T_{i,n}\}$
3: end for
4: while L do
5: $ST_{i\to j} \leftarrow \{T_{i,a_1}, T_{i,a_2}, \dots, T_{i,a_m} \mid a_p \ are\ sampled\ integer\}$
6: send $ST_{i\to j}$
7: $DT_i \leftarrow \{T_{i,b_1}, T_{i,b_2}, \dots, T_{i,b_{l_i}} \mid b_p \ are\ sampled\ integer\}$
8: delete DT_i
9: $M'_i \leftarrow$ append $S_{j\to i}$
10: end while

A ∪ BおよびA/Bはそれぞれ A, B の和集合と差集合を表す．最後に, 端末iは新しいモデルM'_iを RF の予測モデルとして用いる．この交換はL回繰り返すことを想定する．

アルゴリズム 1 に提案アルゴリズムの概要を示す. 始めに，各端末iは決定木を学習し，予測モデルM_iを作成する．作成したM_iからランダムにm本の木をランダムに選択し，隣接端末へ渡す$ST_{i\to j}$を決定する．$ST_{i\to j}$を渡した後，削除するl_i本の木をランダムに選択し，DT_iを定義する．$ST_{j\to i}$を受け取る前にM_iからDT_iを削除し，最後に$ST_{j\to i}$を受け取り新たなモデルM'_iを作成する．交換はすべての隣接端末J_iに対して実行する. この交換の流れのシーケンス図を図 3 に示す.

5. 結果

本章では，提案アルゴリズムの評価結果および結果の考察について述べる．

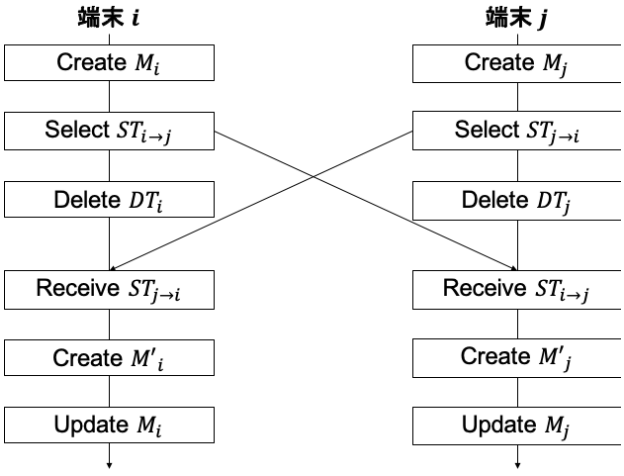

図 3 交換のシーケンス図

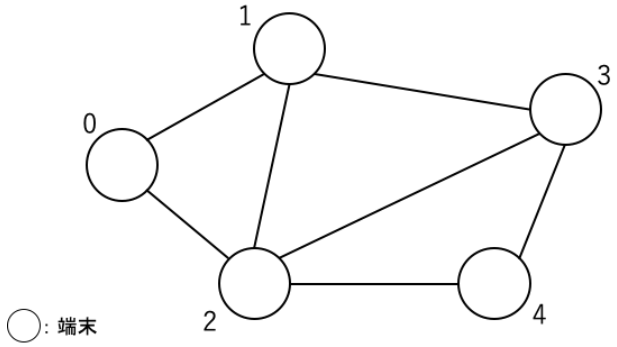

図 4 マルチホップネットワークの
シミュレーションモデル

5.1 マルチホップネットワーク

本節では，隣接端末のみで通信が可能なマルチホップネットワーク接続での評価結果を述べる．評価に用いた 5 端末間でのマルチホップネットワークのトポロジーを図 4 に示す．評価には，機械学習アルゴリズムの有名な画像データセットである MNIST [8]を使用した．各端末の予測精度は, 協調学習を行っていない一般的な RF（RF w/ local data）と提案アルゴリズムによる協調 RF（Coop. RF w/ exchange）を比較し評価した．また，プライバシー，ネットワーク接続，メモリリソースなどを考慮しない場合の参考値として理想的な状況の精度の評価も行った．参考値には，全端末で学習された木を全て集めた場合の予測（All trees）と，全データを用いた RF の予測（RF w/ all data）を適用した．RF の実装には Python ライブラリである scikit-learn を用いた．

表 1 にシミュレーションに用いたパラメータを示す．各端末は 1,000 のトレーニングデータを持っており，これらのデータを用いて$n = 100$本の決定木を学習する．そして，

表 1 シミュレーションパラメータ

	RF w/ local data	Coop. RF w/ exchange	All trees	RF w/ all data
トレーニングデータ数	1,000	5×1,000	5×1,000	5,000
テストデータ数	1,000	1,000	1,000	1,000
木の総数	100	100	500	100
木の深さ	5	5	5	5

Vol.2020-DPS-184 No.1
Vol.2020-EIP-89 No.1
2020/9/10

表 2 端末ごとの木の情報

端末	0	1	2	3	4
隣接端末数	2	3	4	3	2
決定木の総数(交換前)	100	100	100	100	100
交換する木の本数	20	30	40	30	20
決定木の総数(交換後)	100	100	100	100	100

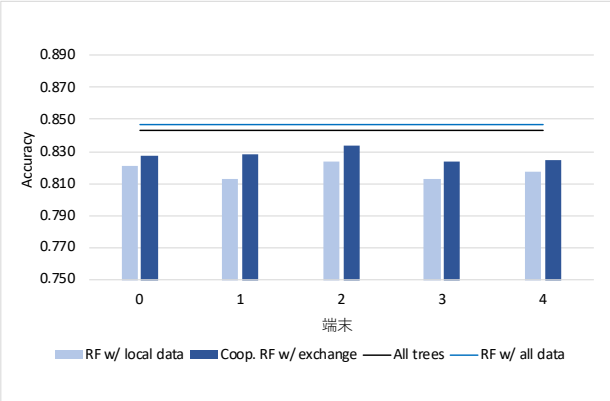

図 5 交換が 1 度のみの場合の結果

100 本の木から各隣接端末へ$m = 10$本の木を選び交換を行う．結果を確かめるためのテストデータ数はすべての場合で 1,000 とする．交換前後の予測結果は RF w/ local data と Coop. RF w/ exchange とし，図 5 の棒グラフに示す．まず 1 つ目の評価実験ではM_iの取得後，交換は 1 度だけ行った．All trees は，各端末で学習したすべての木をサーバ上で共有することを想定している．したがって端末数と作成した木の本数の積（i.e. $5 \times 100 = 500$）の決定木を集め予測を行う．また，RF w/ all data は，プライバシーを考慮せずにすべてのローカルデータをサーバ上で共有し, RF を行ったことを想定しており，学習する木の本数は提案アルゴリズムのローカルモデルと同じ本数の$n = 100$とする．また，All trees と RF w/ all data のトレーニングデータ数はすべてのローカルデータを使用することを想定しているため，5,000 とする．各端末において予測および交換に用いた決定木の本数を表 2 に示す. 交換前の木の本数を$n = 100$とすると，端末ごとに隣接端末数が異なるため交換する木の本数は各端末で異なるが，交換後の木の本数は全端末で再び$n = 100$となる．

表 1 および表 2 のパラメータを使用して行った評価結果を図 5 に示す．各端末 0，1，2，3，4 の RF w/ local data の精度は順に 82.1%，81.3%，82.4%，81.3%，81.7%となり，

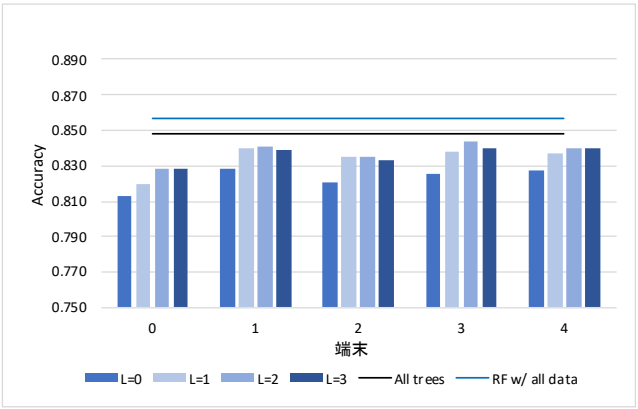

図 6 交換を繰り返したときの結果

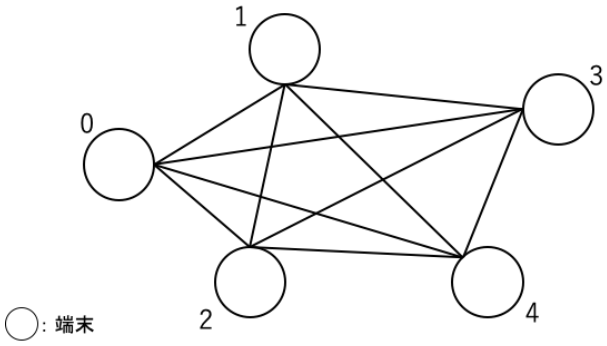

図 7 完全グラフに基づいた接続のシミュレーションモデル

Coop. RF w/ exchange では順に 82.7%, 82.8%, 83.4%, 82.4%, 82.4%という結果であった．また，比較対象となる All trees では 84.4%，RF w/ all data では 84.6%という精度が得られた．以上の結果は，協調学習を行うことによって精度が向上し，ローカルデータを共有せずに学習器のみを共有した際の精度がすべてのローカルデータを使用して学習したときの精度に近づいていることを示している．つまり，弱学習器のみの共有でも十分な精度を得ることができると考えられる．

次に, 交換を複数回繰り返したときの結果を図 6 に示す．図 6 では，交換なし（$L = 0$），1 回交換後（$L = 1$），2 回交換後（$L = 2$），3 回交換後（$L = 3$）の結果および図 5 と同様に All tree と RF w/ all data の結果を示している. 各端末 0, 1, 2, 3, 4 における$L = 0$の精度は順に 81.3%, 82.8%, 82.1%, 82.5%，82.7%であり，3 回交換を繰り返した$L = 3$での精度は順に 82.9%，83.9%，83.3%，84.0%，84.0%となった．また，All trees，RF w/ all data はそれぞれ 84.8%，85.6%であった．$L \leq 2$のとき，木の交換をすることで精度が向上するが，$L = 3$のときいくつかの端末で精度が下がっている．これは交換を繰り返すことによって 5 端末間の精度が同程度の数値に近づくためである．

Vol.2020-DPS-184 No.1
Vol.2020-EIP-89 No.1
2020/9/10

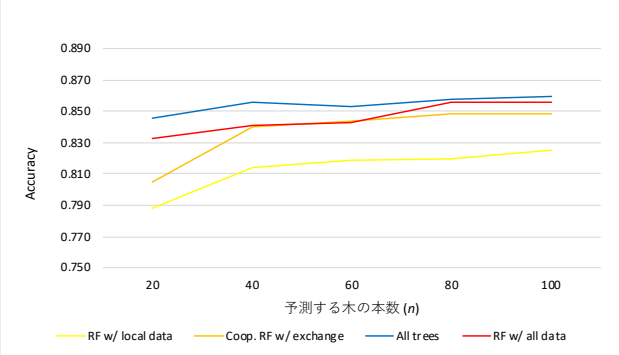

図 8　予測する木の数を変更した場合の結果

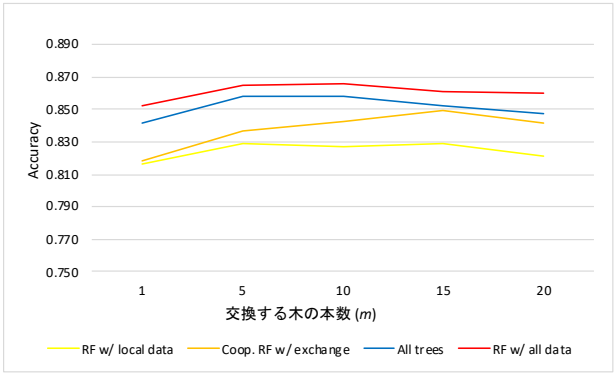

図 9　交換する木の本数を変更した場合の結果

5.2 パラメータ変更による評価

本節では，さまざまなパラメータでのパフォーマンスについて述べる．このシミュレーションでは，端末が他の全端末と接続されている状況を想定し，ネットワークトポロジを図 7 に示すような完全なグラフとして表す．

1 つ目のシミュレーションとして，各端末における予測する木の本数を変更したときの結果を図 8 に示す．図 8 では，協調学習を行っていない一般的な RF の 5 端末の平均（RF w/ local data），提案アルゴリズムによる協調 RF の平均（Coop. RF w/ exchange），全端末で学習された全ての木を用いた予測（All trees）と，全データを用いた RF（RF w/ all data）を示す．トレーニングデータ，テストデータ，木の深さは表 1 に示した値と同様とし，交換する木の本数は $m = 5$とした．また，予測の木の本数は 20 から 100 に設定した．この条件下で Coop. RF w/ exchange の結果は，それぞれ 80.5%，84.0%，84.4%，84.9%，84.9%となった．図 8 のグラフより提案スキームである Coop. RF の精度は予測の木の本数が増加するにつれ高くなり，ローカルデータのみで予測を行う RF w/ local data よりも速く，全データを用いた予測である RF w/ all data に近づいていることがわかる．

2 つ目のシミュレーションでは，1 組の隣接端末のペア間で交換する木を変更した場合の結果を図 9 に示す．1 つ目のシミュレーションと同様に，トレーニングデータ，テストデータ，木の深さは表 1 の値とし，予測した木の総数も表 1 と同様とした．交換する木の本数は，1，5，10，15，20 に設定し，協調学習を行う Coop. RF w / exchange の結果は，それぞれ 81.8%，83.7%，84.3%，85.0%，84.7%となった．これらの結果は，端末間で交換する木の本数が多いほど，より高い精度を達成できることを示している．

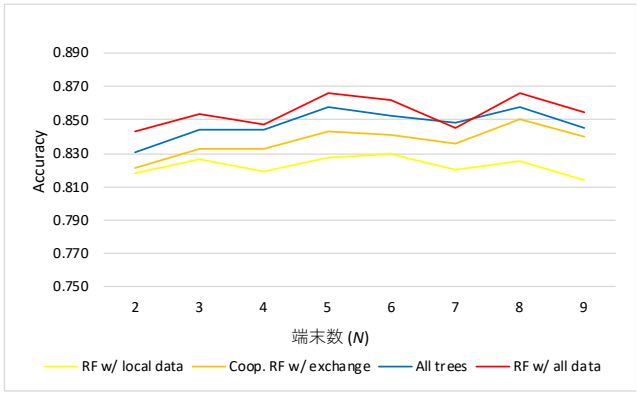

図 10 端末数を変更した場合の結果

表 3　端末数変更時の使用パラメータ

	RF w/ local data	Coop. RF w/ exchange	All trees	RF w/ all data
トレーニングデータ数	1,000	1,000×N	1,000×N	1,000×N
テストデータ数	1,000	1,000	1,000	1,000
木の総数	100	100	100×N	100
木の深さ	5	5	5	5

N: the number of devices

最後のシミュレーションは，さまざまな端末数で評価したもので，その結果を図 10 に示す．この評価で用いるパラメータは表 3 の通りである．各端末での交換する木の本数は，他の評価と同様に$m = 10$であるが，トレーニングデータ数および全端末で学習した木を用いる All trees での予測した木の本数は，端末数により異なる．図 10 より RF w/ local data と Coop. RF w/ exchange との差は端末数が増えるにつれ，大きくなっていることがわかる．この結果から，マルチホップネットワーク接続においてもより多くの IoT 機器間の通信による精度が保証されることが期待できる．

6. むすび

本稿では, RF を用いた協調学習のアルゴリズムを提案した．提案アルゴリズムは，ローカルデータを共有する代わ

りに学習器である決定木を交換することで，協調学習を行いプライバシーの保護を検討した．提案アルゴリズムを 5 端末間におけるマルチホップネットワークで評価した結果，ローカルデータを共有せずに学習器のみを共有する提案アルゴリズムで，すべてのローカルデータを使用した場合と同程度のパフォーマンスが実現された．

今後の展望として以下が挙げられる．まず，今回の評価実験において，アルゴリズムはとても単純な状況下でシミュレーションを行ったため実用性が確認されていない．マルチホップネットワーク接続におけるシミュレーションでは 5 端末，完全グラフに基づいた接続方法ではパラメータを変更し最大で 9 端末での評価を行ったが，実用性を確かめるためには端末数を増やした評価をする必要があると考える．また，データセットの変更も検討する．今回はデータセットに MNIST を使用したが，MNIST ではなくカメラから得られた画像データやウェアラブル機器を使用した生体情報など，実験で取得したデータを利用していくことを考える．次に，交換を繰り返す際にモデルの更新方法の改良が挙げられる．本稿での木の交換は，木を削除し相手の端末から木を受け取ることで新しくモデルを作成しているが，Gradient Boosting のように新しくモデルを作る際に差分を使用することで，さらに精度を向上させることが期待できる．

参考文献

[1] Do, T.M.T. and Gatica-Prez, D.. Where and what: Using smartphones to predict next locations and applications in daily life. Pervasive and Mobile Computing, 2014, vol. 12, p. 79-91.
[2] Konečný, J. et al.. Federated Learning: Strategies for Improving Communication Efficiency. NIPS Workshop on Private Multi-Party Machine Learning, 2016.
[3] Li, T. et al.. Federated Learning: Challenges, Methods, and Future Directions. IEEE Signal Processing Magazine, 2020, vol. 37, issue 3, p. 50-60.
[4] Lim, W.Y.B. et al.. Federated Learning in Mobile Edge Networks: A Comprehensive Survey. IEEE Signal Processing Magazine (Early Access), 2020.
[5] Lakshmanaprabu, S.K. et al.. Random forest for big data classification in the internet of things using optimal features. International Journal of Machine Learning and Cybernetics, 2019, vol. 10, p. 2609-2618.
[6] Belgiu, M. and Dragut, L.. Random forest in remote sensing: A review of applications and future directions. ISPRS Journal of Photogrammetry and Remote Sensing, 2016, vol. 114, p. 24-31.
[7] Pavey, T.G. et al.. Field evaluation of a random forest activity classifier for wrist-worn accelerometer data. Journal of Science and Medicine in Sport, 2017, vol. 20, issue 1, p. 75-80.
[8] “THE MNIST DATABASE of handwritten digits”. http://yann.lecun.com/exdb/mnist/, (参照 2020-08-17).