モデルとプライバシーを保護するアンサンブル決定木向け秘匿推論プロトコル

稲毛 康太[1,a)]　橋本 昌宜[1]　上野 嶺[2]　粟野 皓光[1]　本間 尚文[2]

概要：本論文では，入力データと機械学習モデルを保護し， 効率的な計算の実現を目的とした FHE ベースの秘匿推論プロトコルを提案する．高い推論精度の実現と， 計算量および通信量を削減するため，アンサンブル決定木を提案プロトコルに採用した．効率的な推論が可能なアンサンブル決定木の構築指針も合わせて提示した．特徴は秘密鍵がモデル所有者により生成，保持され，決定木のモデルパラメータを暗号化したままでユーザが推論可能なプロトコルである点にある．モデル所有者は推論計算を信頼できないクラウドサーバーへ外部委託することが可能となる．暗号化された決定木の構造や閾値が，モデル所有者から見た第三者に漏洩しないことをセキュリティモデルによる議論で確認した．実験結果として， 提案プロトコルは計算委託によりモデル所有者側の計算を 90 %以上削減できることを確認した また，決定木アンサンブルが推論精度の向上や， 計算量や通信量の削減に役立つことを実験的に示した．

キーワード：秘匿推論プロトコル，準同型暗号，決定木アンサンブル，プライバシー保護， 機械学習モデル保護

An oblivious inference protocol for decision tree ensembles with model and privacy protection

Kota Inage[1,a)]　Masanori Hashimoto[1]　Rei Ueno[2]　Hiromitsu Awano[1]　Naofumi Homma[2]

Abstract:
This paper proposes an oblivious inference protocol aiming to achieve efficient computation and privacy of user data and machine learning model parameters. This work focuses on decision tree ensemble (DTE) as a means of higher inference accuracy and less computation and communication, and explicitly proposes a model construction guideline for efficient inference execution. A novel feature is that the secret key is generated and held by the model owner and the model owner can outsource most inference computation to untrustable cloud servers. Our discussion on the security model reveals that the proposed homomorphic evaluation of DTE prevents the model leakage to a third party. Experimental results show that the proposed protocol can reduce the computation at the model owner side by more than 90% thanks to the outsourcing. Also, it is confirmed that DTE helps not only improve the inference accuracy but also reduce the calculation time and communication amount.

Keywords: oblivious inference protocol, homomorphic evaluation, decision tree ensemble, privacy preservation, machine learning model protection

1. 序論

今日の情報化社会において機械学習の発展はめざましく，様々な分野で機械学習モデルが構築されている．機械学習モデルは希少なデータを用い大規模計算によって構築された知的財産である．有用な機械学習モデルが構築さ

[1] 大阪大学大学院情報科学研究科情報システム工学専攻
Department of Information Systems Engineering, Graduate School of Information Science Technology, Osaka University
[2] 東北大学電気通信研究所
Research Institute of Electrical Communication, Tohoku University
[a)] k-inage@ist.osaka-u.ac.jp

れると，クラウド上でユーザデータの推論を実行するサービスを提供したい企業が出てくる．推論の外部委託は機械学習の力を活用する望ましい形態であるが，クラウドサーバー上にモデルが配置されている場合，モデルやそのパラメータが盗まれるリスクにさらされることがある．モデルパラメータを盗む方法として，OS のバグを活用する方法，Spectre や Meltdown といったサイドチャネル攻撃が報告されている [1].

機械学習モデルが盗まれる危険性に対して，モデルや入力データを暗号化し保護した状態で推論できる秘匿推論と呼ばれる技術の開発が進められている．秘匿推論では機械学習モデルを持つモデル所有者，入力データを持つデータ所有者の二者間で相互に情報が漏れないように推論を行う．準同型暗号による決定木モデルの秘匿推論について議論が盛んであり，本研究でも他の機械学習と比べ，本質的に計算量が少ない決定木モデルの秘匿推論について扱う．

一方で，従来の決定木秘匿推論プロトコルには 2 つの問題点がある．第一の問題はモデルの保護が不十分であるという点である．具体的には，決定木の閾値は暗号化されていても木構造については暗号化されておらず，クラウドサーバーへの計算委託をするとモデルの情報が漏洩すること，また，データ所有者が秘密鍵を持つためデータ所有者とクラウドサーバーが結託すればモデルの情報がわかってしまうことが挙げられる．第二の問題点は，モデルアンサンブルに対応していない点である．モデルアンサンブルは弱学習器を多数用いることで精度を向上させる技術である．後に詳しく述べるが従来の決定木秘匿推論プロトコルは木の深さに強く依存するため，弱学習器として浅い決定木を用いることで計算量や通信量の削減が期待できる．

そこで本研究では，モデル所有者が秘密鍵を持ち，決定木構造を暗号化することで計算委託を可能にするとともに，モデルアンサンブルを適用することで計算量や通信量を削減した 3 者間の秘匿推論プロトコルを提案する．さらに，計算量や通信量を最小限に抑えるため，アンサンブルモデル構築方法についてガイドラインを示す．

1.1 関連研究

従来の決定木秘匿推論プロトコルとして，Tai らは加法準同型暗号による線形関数を用いて決定木分類計算を行うプロトコル [2] を提案した．しかし，決定木分類において入力値と閾値のいずれかしか暗号化されておらず，モデル漏洩の危険がある．

その後，Lu ら は XCMP という大小比較プロトコルを提案し，これを Tai のプロトコルに適用することで，入力値と閾値がともに暗号化され，かつ相互通信を必要としないプロトコルを実現できることを示した [3]．ただし決定木構造は暗号化されておらず，また計算量コストが高い完全準同型暗号を用いているため計算効率が Tai のプロトコルに比べ良くないという課題がある．

2. 準備

2.1 準同型暗号 (HE)

準同型暗号 (HE) は平文の多項式関数を暗号文で評価できる公開鍵暗号である．本研究では HE の一種である，多項式環 FHE を用いる [4]．平文，暗号空間は多項式環 $\mathbb{Z}_t[X]/\langle X^m+1\rangle$，$\mathbb{Z}_q[X]/\langle X^m+1\rangle$ によって定義される．t，q は $t<q$ となる整数であり，m は 2 の累乗の整数である．HE の暗号化，復号化をそれぞれ $\mathsf{Enc}_{\mathsf{pk}}(x)$，$\mathsf{Dec}_{\mathsf{sk}}([x])$ と表す．ここで，pk と sk は公開鍵と秘密鍵である．次に，x の暗号文を $[x]$ $(=\mathsf{Enc}_{\mathsf{pk}}(x))$ と表す．また平文の定数係数 p の暗号文を $[\![p]\!]$ とする．

平文 x，y の暗号文をそれぞれ $[x]$，$[y]$ とする．YASHE や BGV 方式 [5] といった多項式環 FHE を用いれば，x，y を暗号化したまま加算や乗算の準同型評価が可能である．本稿では，準同型加算，乗算それぞれに対して平文に対する算術同様 $+$，$\times$ という算術記号を用いる．この表記法により，準同型加算と乗算をそれぞれ次のように記す．

$$[x+y] = [x]+[y], \tag{1}$$

$$[x\times y] = [x]\times[y] \tag{2}$$

2.2 秘匿大小比較プロトコル

提案プロトコルは XCMP[3] が持つ望ましい特性から，XCMP を秘匿大小比較プロトコルとして利用する．以下，XCMP の提案プロトコルに適した特徴を説明する．

XCMP は多項式環 FHE によって暗号化された 2 つの整数の準同型比較を行う．入力値と閾値がともに暗号化されていても通信を必要とせず大小比較が可能であるという利点を持つ．一方，他の多くの大小比較プロトコル (例 [6]) は入力値と閾値の片方だけが暗号化されている場合でも，通信を複数回必要とする．これらの XCMP の利点によって，提案プロトコルは限られた通信回数でのクラウドサーバーへの計算委託を可能とする．しかし，他の大小比較プロトコルよりも計算量が大きい．本研究では計算量の改善に対しアンサンブル計算が有用であることを明らかにする．

XCMP への入力値は多項式環によって暗号化された 2 つの整数であり，出力値は比較結果を示す暗号文である．入力値 x，閾値を y として，XCMP は比較結果を出力暗号文の定数係数として与える．すなわち，

$$\mathsf{XCMP}([x],[y]) = \begin{cases} [\![1]\!] \text{ if } (x<y), \\ [\![0]\!] \text{ if } (x\geq y) \end{cases} \tag{3}$$

なお，XCMP による結果を示す暗号文の係数は，定数係数を除けばランダムに与えられている．

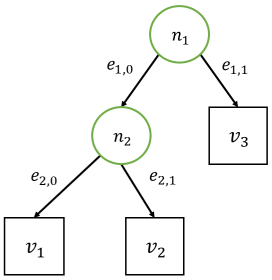

図 1　決定木の例
Fig. 1　An example of decision tree

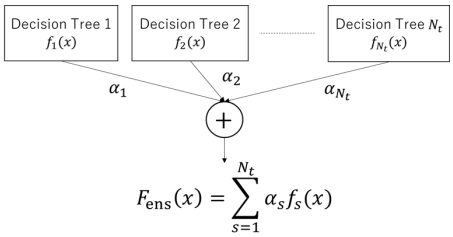

図 2　アンサンブル学習の概要
Fig. 2　Overview of ensemble learning

2.3　決定木 (DT)

2.3.1　定義と表記

決定木 (DT) は代表的な機械学習モデルの一つであり，分類と回帰に用いられる．図 1 に決定木の例を示す．決定木は，決定ノード，葉ノード，およびノード間の接続を表す一連の有向エッジの 3 つの要素から定義される．

決定木について N_d，N_p，および N_c をそれぞれ決定ノード数，葉ノード数，および推論クラス数とする．決定ノードを $n_1, n_2, \ldots, n_i, \ldots n_{N_d}$ と示し，それぞれ閾値 y_i を持つ．n_1 は親ノードを持たない根ノードである．また，$v_1, v_2, \ldots, v_j, \ldots, v_{N_p}$ は葉ノードを示し，推論結果を表す N_c 次元のベクトル $\mathbf{z}_j = (z_{j,1}, z_{j,2}, \ldots, z_{j,w})$ を持つ．$z_{j,w}$ は j 番目の葉ノードの w 番目のクラスを示している．決定木は完全 2 分木とし，決定ノード n_i における左，右方向の子ノードへの，二方向のエッジを $e_{i,0}$, $e_{i,1}$ と示す．

特徴量の総数を N_i, 推論に用いる特徴量を $u_1, u_2, \ldots, u_k, \ldots, u_{N_i}$ と示す．決定木推論においては最初に特徴量から入力値として $\mathbf{x} = (x_1, x_2, \ldots, x_{N_d})$ を選択する．

大小比較において，入力値 $x_1, x_2, \ldots, x_{N_d}$ と閾値 $y_1, y_2, \ldots, y_{N_d}$ をそれぞれ比較し，比較結果ベクトル $\mathbf{b} = (b_1, b_2, \ldots, b_{N_d})$ を生成する．b_i は $x_i \geq y_i$ であれば 1，それ以外は 0 となる．その後, パス $(e_{1,b_1}, e_{2,b_2}, \ldots, e_{N_d,b_{N_d}})$ に対応した葉ノード v_σ を決定する．こうして，決定木は推論結果として葉ノードにおける値ベクトル $\mathbf{z}_\sigma$ を出力する．

2.3.2　暗号化された入力と閾値による秘匿パス同定

本研究では，入力値と閾値はプライバシーとモデル保護のために暗号化されるため，単純なパスの同定（またはトレース）は実行できない．Lu らによるパスコストとエッジコストの概念，ならびに XCMP を利用したパス同定方法を紹介する [3].

葉ノード v_j に対して計算されたパスコストを pc_j，エッジ $e_{i,a}$ に対応したエッジコストを $\mathsf{ec}_{i,a} (a \in \{0, 1\})$ とする．パスコスト pc_j は整数によって与えられ，葉ノード v_σ が選択されたときのみ $pc_j = 0$ となる．エッジコスト $\mathsf{ec}_{i,a}$ は 0 または 1 として与えられる．決定木推論においてはまず，前述した比較結果ベクトル $\mathbf{b}$ が計算される．提案プロトコルの秘匿推論においては，XCMP による比較結果ベクトルより，$\mathbf{b}_{\text{encrypted}} = (\llbracket b_1 \rrbracket, \llbracket b_2 \rrbracket, \ldots, \llbracket b_{N_d} \rrbracket)$ を得る．その後，暗号化されたエッジコスト $\llbracket ec_{i,0} \rrbracket$, $\llbracket ec_{i,1} \rrbracket$ がそれぞれ $\llbracket ec_{i,0} \rrbracket = \llbracket b_i \rrbracket$ and $\llbracket ec_{i,1} \rrbracket = \llbracket 1 \rrbracket - \llbracket b_i \rrbracket$ $(= \llbracket 1 - b_i \rrbracket)$ と与えられる．さらに，暗号化されたパスコスト $\llbracket pc_j \rrbracket$ が v_j へのパスにおけるエッジコストの和として求められる．図 1 における例では，パスコストは以下のように与えられる．

$$\llbracket \mathsf{pc}_1 \rrbracket = \llbracket \mathsf{ec}_{1,0} \rrbracket + \llbracket \mathsf{ec}_{2,0} \rrbracket, \tag{4}$$

$$\llbracket \mathsf{pc}_2 \rrbracket = \llbracket \mathsf{ec}_{1,0} \rrbracket + \llbracket \mathsf{ec}_{2,1} \rrbracket, \tag{5}$$

$$\llbracket \mathsf{pc}_3 \rrbracket = \llbracket \mathsf{ec}_{1,1} \rrbracket \tag{6}$$

2.4　アンサンブル学習

多くの弱学習器を利用し強学習器を構築するアンサンブル学習は，推論モデルの機能と推論精度を向上させる，単純でありながら強力な枠組みである．決定木推論の場合には弱学習器は浅い決定木となる．バギング，ブースティング，スタッキング [7] などいくつかのアンサンブル方法があり，それらの実装が一般に公開されている，例えば Adaboost[8]，ランダムフォレスト [9]，XGBoost [10] などである．

図 2 にアンサンブル学習の概要を示す．N_t 個の弱学習器 (決定木等) による，推論結果は以下のように求められる．

$$F_{\text{ens}}(\mathbf{x}) = \sum_{s=1}^{N_t} \alpha_s f_s(\mathbf{x}) \tag{7}$$

$\mathbf{x}$ は入力データベクトル，$f_i(\mathbf{x})$ は s 番目の学習器の分類結果，そして α_s は s 番目の学習器に対応する重み係数を示す．N_c クラス分類には，N_c 個の α_s と $f_i(\mathbf{x})$ が用意される．

3.　提案手法

本章では 3 者間の決定木秘匿推論プロトコルを提案する．まずプロトコル構築への基本的な考えと，解決しなければならない 3 つの課題を明確化する．次に 3 つの課題に対する解決案を示す．最後に提案プロトコルの計算複雑度を分析し，セキュリティモデルについて議論する．

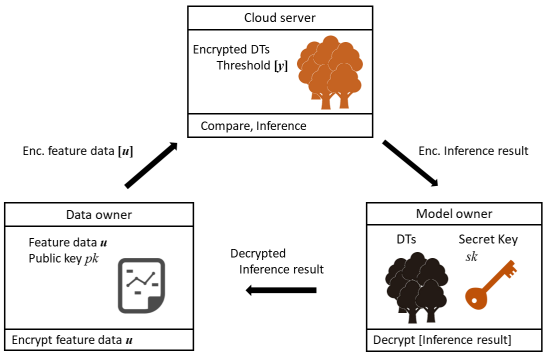

図 3　提案プロトコルのベースライン
Fig. 3　Baseline of the proposed protocol

3.1　プロトコルのベースラインと課題

図 3 に提案プロトコルのベースラインを示す．先行研究プロトコルと異なる点はモデル所有者が鍵を持つことである．木構造やパラメータを含めた決定木アンサンブル (DTE) モデルは，モデル所有者の鍵によって暗号化される．したがって DTE モデルは漏洩の危険なしで外部（クラウドサーバー等）に置くことができる．

一方，提案プロトコルでは暗号化した DTE モデルを用いてクラウドに秘匿計算を委託する必要がある．まず単一決定木において考えると，先行研究プロトコルでは入力データか閾値のどちらかは平文であったり [11]，木構造が暗号化されていない問題があった [2]．暗号化された DT モデルに対する秘匿計算を可能としなければならない（課題 1）．

提案プロトコルにおける秘匿推論では，計算結果もまた暗号化されている．データ所有者が推論結果を知るためには，秘密鍵を持つモデル所有者が復号しなければならない．しかしながらこの際に，モデル所有者は推論結果を知ってしまう．そこでモデル所有者が推論結果を取得できない仕組みが必要となる（課題 2）．

最後の課題は，どのようにアンサンブル計算（式 (7)）をプロトコルに効率的に組み込み，実装するかである．準同型乗算は計算量やノイズを増やすことから可能な限り避けつつ，効率的にマルチクラスでの推論を実行したい（課題 3）．

3.2　課題への解決案

図 4 に上記の課題を解決する提案プロトコルを示す．課題 1，2，3 に対する解決案はそれぞれ青，赤，緑で色付けされている．全体のプロセスは 3.3 章で説明する．本章では 3 つの課題とその解決案についてそれぞれ説明する．

3.2.1　解決案 1:暗号化された決定木モデルによる秘匿計算

図 4 のプロセス（P3）に当たる，暗号化された DT に対するパス同定方法を提案する．提案手法では図 5 に示すように決定木構造をアフィン写像として表現し，それを決定木構造マップと呼ぶことにする．決定木構造マップを用いて，準同型評価により XCMP の暗号化された結果から，暗号化されたパスコストを求める．決定木構造マップにおける配列の列はそれぞれ葉ノードへのパスを示し，行は決定ノードを示す．図 5 では $1,2,3$ 行はそれぞれ葉ノード v_1, v_2, v_3 に対応し，$1,2$ 列は決定ノード n_1, n_2 に対応する．$[1],[0]$ または $[-1]$ の値は，対象となるパス内の i 番目のノードにおいて，左右のいずれのエッジ $((e_{i,0}),(e_{i,1}))$ が含まれるか，またはそのノードがパスに含まれないということを示している．決定木構造マップの定数係数ベクトルは，対象となるパスに含まれる右エッジの総数を表している．図 5 では決定木構造から $1,2,3$ 行目の要素がそれぞれ $[0],[1],[1]$ となる．

次に，決定木構造マップを用いたパスコストの計算方法を説明する．図 5 の 1，2 番目のノードにおける XCMP の比較結果をそれぞれ $[\![b_1]\!]$，$[\![b_2]\!]$ とする．例として，2 行目に表現されるパスについてパスコスト計算を考える．決定木構造マップを用いると，v_2 へのパスコストは以下のように計算される．

$$[\![\mathsf{pc}_2]\!] = [\![b_1]\!] \times [1] + [\![b_2]\!] \times [-1] + [1] \tag{8}$$
$$= [b_1] + [1 - b_2] \tag{9}$$
$$= [\mathsf{ec}_{1,0}] + [\mathsf{ec}_{2,1}] \tag{10}$$

ここで，上記の式の最初の 2 つの項は，比較結果 $[\![b_i]\!]$ 行列との内積に対応し，3 番目の項は，定数係数ベクトル中の値である．また $\mathsf{ec}_{i,0} = b_i$, $\mathsf{ec}_{i,1} = 1 - b_i$ はそれぞれ左，右エッジを示し，正しくパスコストが求められていることが分かる．v_1, v_3 へのパスコストも同様に計算される．このように，準同型評価により暗号化されたままパスコストを求めることができる．

3.2.2　解決案 2:乱数を使用してモデル所有者が推論結果を取得できないようにする

モデル所有者が推論結果を知ってしてしまうという課題 2 への解決案として，パスコストをマスキングする．それぞれのパスコストに対してランダム値 r_j を生成し，準同型加算をしてマスクする．関連する部分は図 4 において赤色で示されている．ランダム値を加えることで，モデル所有者はマスキングされたパスコストを $pc_j + r_j$ を復号化しても，推論結果を得ることができない．復号してデータ所有者に送られた後，データ所有者は元のパスコストを取得するために，r_j を取り除く．この流れを成立させるため，データ所有者はランダム値を生成し．モデル所有者の公開鍵を用いて暗号化する．それらはクラウドサーバーにおける秘匿計算で用いるために，クラウドサーバーへ送られる．同じアイデアを DTE の推論結果 F_{ens} の秘匿化にも適用する．

3.2.3　解決案 3:効率的な暗号化されたアンサンブル計算

DTE による秘匿推論を実現するために，以下の暗号化

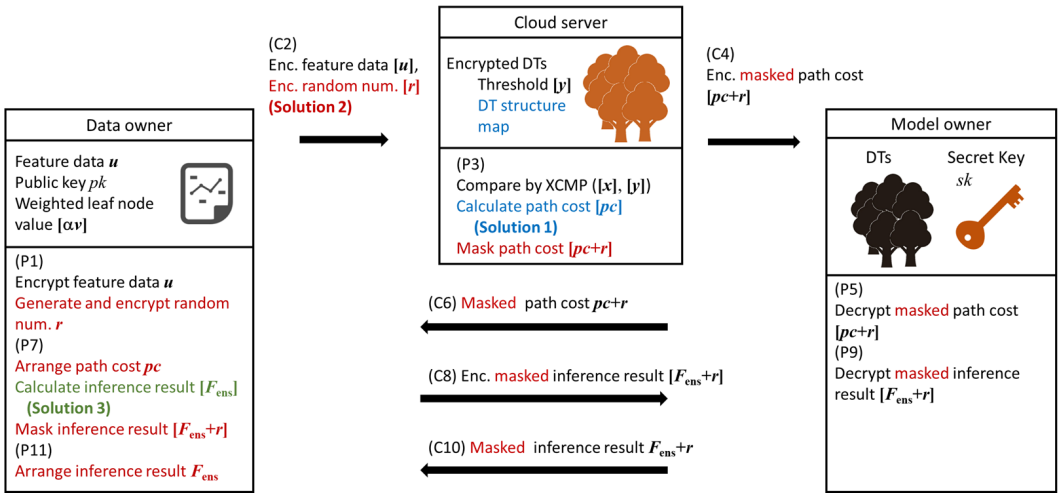

図 4　提案プロトコル

Fig. 4　Proposed protocol

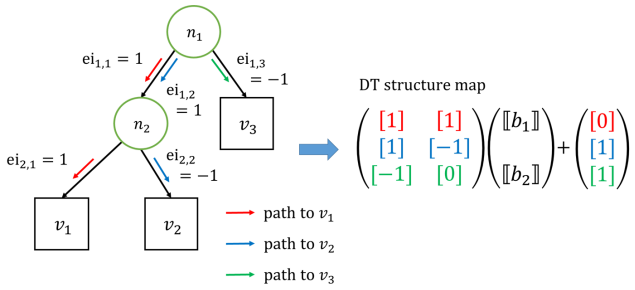

図 5　解決案 1:暗号化された決定木構造

Fig. 5　Solution 1: Encrypted DT structure

されたアンサンブル計算を行う．

$$[F_{\text{ens}}(\mathbf{x})] = \sum_{s=1}^{N_t} [\alpha_s] \times [f_s(\mathbf{x})] \tag{11}$$

$f_s(\mathbf{x})$ は s 番目の決定木の推論結果，α_s はその重みを示す．この $f_s(\mathbf{x})$，α_s は学習時点で定まっているため，モデル所有者は前もって重みと推論結果の乗算を行い，暗号化してデータ所有者に送ることができる．具体的には以下となる．

$$[F_{\text{ens}}(\mathbf{x})] = \sum_{s=1}^{N_t} [g_s(\mathbf{x})] \tag{12}$$

$g_s(\mathbf{x})$ は s 番目の決定木における重みがかけられた推論関数であり，その出力は次の通りである．

$$g_s(\mathbf{x}) = [\alpha_s \times z_{\sigma_s,1,s}, \alpha_s \times z_{\sigma_s,2,s}, \ldots, \alpha_s \times z_{\sigma_s,N_c,s}] \tag{13}$$

式 (13) は，モデル所有者は前もって $\alpha_s \times \mathbf{z}_{j,s}$ を計算して暗号化し，データ所有者へ $[\alpha_s \times \mathbf{z}_{j,s}]$ を送ることを示している．したがって，図 4 においてデータ所有者は重み付けられた葉ノードの値ベクトルを持つことになる．このように，アンサンブル計算において準同型乗算を排除し，データ所有者によって $[F_{\text{ens}}(\mathbf{x})]$ の計算を行う．

3.3　提案プロトコル

本章では図 4 に示した提案プロトコルについて，プロセスごとに説明する．N_d, N_p, N_t, N_c, N_i はそれぞれ決定木の決定ノード数，パス数，決定木数，推論クラス数，入力データ数である．

(P0) 事前準備:

- モデル所有者はクラウドサーバーへ暗号化された閾値 $[y_{i,s}]$ と決定木構造マップと送る．
- モデル所有者は $[\alpha_s \times \mathbf{z}_{j,s}]$ を用意し，データ所有者へ送る．

(P1+C2) 暗号化:

- For $k \in \{1, 2, \ldots, N_i\}$，データ所有者は特徴量 u_k を暗号化し，クラウドサーバーへ $[u_k]$ を送る．
- For $j \in \{1, 2, \ldots, N_p\}$ and $s \in \{1, 2, \ldots, N_t\}$，データ所有者はランダム値 $r_{j,s}$ を生成，暗号化し，$[r_{j,s}]$ をクラウドサーバーへ送る．

(P3+P4) 準同型パスコスト評価:

- For $i \in \{1, 2, \ldots, N_d\}$ and $s \in \{1, 2, \ldots, N_t\}$，クラウドサーバーは特徴量 $[u_k](1 \le k \le N_i)$ から DTE への入力値 $[x_{i,s}]$ を選択する．XCMP を用いて閾値 $[y_{s,i}]$ と大小比較し，比較結果 $[\![b_{i,s}]\!]$ を得る．
- For $j \in \{1, 2, \ldots, N_p\}$ and $s \in \{1, 2, \ldots, N_t\}$，$[\![b_{i,s}]\!]$ と決定木構造マップを用いてパスコスト $[\![\mathsf{pc}_{j,s}]\!]$ を求める．その後，ランダム値 $[r_{j,s}]$ を準同型加算し．モデル所有者へ $[pc_{j,s} + r_{j,s}]$ を送る．

(P5+P6) 復号:

- For $j \in \{1, 2, \ldots, N_p\}$ and $s \in \{1, 2, \ldots, N_t\}$，モデル所有者は $[pc_{j,s} + r_{j,s}]$ を復号し，$pc_{j,s} + r_{j,s}$ をデータ所有者へ送る．

(P7+P8) アンサンブル計算:

- For $j \in \{1, 2, \ldots, N_p\}$ and $s \in \{1, 2, \ldots, N_t\}$，データ所有者は $r_{j,s}$ を $\mathsf{pc}_{j,s} + r_{j,s}$ から取り除く．$\mathsf{pc}_{\sigma_s,s} = 0$

であれば σ_s 番目のパスに決定される．

- データ所有者は $\sum_{s=1}^{N_t}[\alpha_s \times \mathbf{z}_{\sigma_s,s}]$ より $[F_{\mathrm{ens}}(x)]$ を計算し，ランダム値 $[r]$ を生成，加算してモデル所有者へ $[F_{\mathrm{ens}}(x)+r]$ を送る．

(P9+P10) アンサンブル結果の復号: モデル所有者は $[F_{\mathrm{ens}}(x)+r]$ を復号し，$F_{\mathrm{ens}}(x)+r$ をデータ所有者へ送る．

(P11) 推論結果の取得: データ所有者は $F_{\mathrm{ens}}(x)+r$ から r を取り除き，推論結果 $F_{\mathrm{ens}}(x)$ を得る．

3.4 提案プロトコルのセキュリティ

提案プロトコルにおいてはステークホルダーはセミオネストであり，モデル所有者とクラウドの結託がないと仮定する．本章では漏洩する情報から提案プロトコルの安全性を評価し，モデル所有者とクラウドの結託がないとする仮定が必要な理由を説明する．

漏洩する情報は以下の通りである．

データ所有者:

- パス数 N_p
- 決定木数 N_t

クラウドサーバー:

決定ノード数 N_d

- パス数 N_p
- 決定木数 N_t

モデル所有者:

- マスクされたパスコスト $pc_{j,s}+r_{j,s}$
- マスクされた推論結果 $F_{\mathrm{ens}}+r$

データ所有者は重み付けされた葉ノードの値 $[\alpha_s \times \mathbf{z}_{j,s}]$, クラウドは特徴量 $[u_k]$，入力データ $[x_{i,s}]$，ランダム値 $[r_{j,s}]$，閾値 $[y_{j,s}]$，決定木構造マップと，その他いくつかの暗号化された情報を所有している．

決定木アンサンブルに関してデータ所有者は N_p, N_t，クラウドは N_d, N_p, N_t の情報を得ているが，これらの情報から決定木構造を正確に知ることはできない．また，モデル所有者が持つ $pc_{j,s}+r_{j,s}$，$F_{\mathrm{ens}}+r$ にはランダム値が加算されているため，パスコストや推論結果がモデル所有者に漏れることはない．

データ所有者やクラウドサーバーが持つ暗号化された情報については，モデル所有者が秘密鍵を持つために復号されることはなく安全である．以上より提案プロトコルの安全性が示された．ただし，モデル所有者とクラウドの結託があると仮定すると，モデル所有者の秘密鍵を用いてデータ所有者の特徴量 u_k，入力データ $x_{i,s}$ といった値が漏洩してしまう．これがモデル所有者とクラウドの結託がないよう仮定する理由である．

3.5 計算複雑度

最後に，求められる推論精度を保ちつつ，計算量や通信量を最小化するための指針を示す．DTE モデルの構築においては，決定木の深さ d と決定木の数 N_t がパラメータとなる．決定木は完全二分木とする．表 1 は単一決定木と決定木アンサンブルにおける計算量，通信量のオーダー表記であり，モデル構築の指針で利用する．

表 1 提案プロトコルの計算量と通信量
Table 1 Amounts of calculation and communication of the proposed protocol

	Single DT	DTE
Computation	$\mathcal{O}(2^{2d})$	$\mathcal{O}(N_t \times 2^{2d})$
Communication	$\mathcal{O}(2^d)$	$\mathcal{O}(N_t \times 2^d)$

表 1 において，計算量や通信量は決定木の深さ d に深く依存している．単一決定木，決定木アンサンブルにおける決定木の深さをそれぞれ d_s, d_e とすると，$d_s > \log_2 N_t + d_e$ の時，決定木アンサンブルの方が単一決定木よりも計算量，通信量ともに少なくなる．このことから決定木アンサンブルを扱う際は d，N_t，そして推論精度を考慮に入れる必要がある．

モデル構築指針として，まず求められる推論精度が達成できる，様々なパラメータ d，N_t の組を用意する．その後，計算量を重視する場合には $N_t \times 2^{2d}$ を，通信量を重視する場合には $N_t \times 2^d$ を最小化するパラメータ d，N_t に決定する．この指針の妥当性については，4.2 章で議論する．

4. 実験

4.1 実験設定

実験環境として，Intel Core i7-6700 CPU@3.40GHz, 16GB RAM，OS は ubuntu 18.04.02 を使用する．BGV 方式を実装した完全準同型暗号ライブラリ HElib と数論計算ライブラリ NTL を用いて，C++（コンパイラ g++-7.5.0）によって提案プロトコルを実装した．

HElib における暗号パラメータは，正しい準同型計算が可能で，一般な 128-bit セキュリティを満たすよう以下の値に設定した．

$$m = 16384,\ q = 1031,\ bits = 60$$

m, q は暗号空間，$bits$ は BGV 方式で利用可能な乗算回数に関連するパラメータである．セキュリティレベルとして 222.608-bit セキュリティを達成している．

アンサンブル学習においては python 向けオープンソース機械学習ライブラリ scikit-learn を用いる．様々なアンサンブル方法があるが，本実験には AdaboostClassifier と AdaboostRegressor を用いる．

4.2 計算時間，通信量の評価

実験には UCI repository[12] で公開されており，先行研究 [3] でも利用されている (1) heart disease (HD)，(2)

表 2　データセットと DT/DTE の構造
Table 2　Data set and DT/DTE structures

Data set	Single DT		DTE	
	(N_i, N_d, d)	Accuracy	$(N_i, N_d, d) \times N_t$	Accuracy
HD	$(13, 7, 3)$	80.0%	$(13, 1, 1) \times 3$	86.6%
HG	$(13, 24, 5)$	86.9%	$(13, 3, 2) \times 6$	87.1%
SP	$(57, 35, 6)$	89.8%	$(57, 1, 1) \times 7$	90.4%

表 3　先行研究プロトコルの計算時間（ms）
Table 3　Computation time of conventional protocol [3] (ms)

Data set	Data owner		Model owner
	Encryption	Decryption	Inference
HD	27.82	8.6	365.83
HG	33.64	16.97	1247.94
SP	192.12	83.04	1791.04

表 4　提案プロトコルの計算時間 (ms)
Table 4　Computation time of proposed protocol (ms)

Data set	Data owner	Cloud server	Model owner
	Encryption	Inference	Decryption
HD	70.55	272.74	30.04
HG	91.38	2437.02	88.58
SP	270.72	645.49	50.64

表 5　通信量の比較
Table 5　Comparison of communication amount

Data Set	Conventional [3]	Proposed
HD	16E	20E
HG	50E	74E
SP	70E	44E

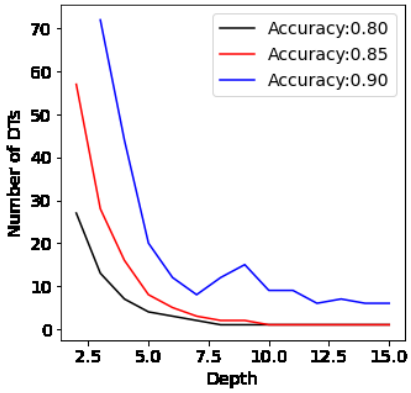

図 6　同じ推論精度を実現する決定木数と深さの組み合わせ (MNIST データセット)
Fig. 6　Combinations of Depth (d) and number of DTs (N_t) for the same accuracy

表 6　精度 85%達成に必要なパラメータ
Table 6　Parameters for 85% accuracy

(N_t, d)	(57, 2)	(28, 3)	(16, 4)	(8, 5)
$N_t \times 2^{2d}$	912	1792	4,096	8,192
$N_t \times 2^{d}$	228	224	256	256

表 7　MNIST データセットの計算時間
Table 7　Calculation time for MNIST (ms)

(N_t, d)	Data owner	Cloud server	Model owner
	Encryption	Calculation	Decryption
(57, 2)	5153.65	23580.62	873.60
(28, 3)	5130.21	40579.61	873.98
(16, 4)	5008.55	94632.78	990.22
(8, 5)	4009.03	128475.55	737.37

Spambase (SP)，(3) Boston housing (HG) の 3 つのデータセットを用いる．表 2 は単一決定木と決定木アンサンブルによって同等の精度を達成した際のパラメータを示している．

4.2.1　計算時間

表 3 と表 4 は先行研究プロトコル [3] と提案プロトコルについて，暗号化，推論，復号化における計算時間を示している．対応するステークホルダーもあわせて記されている．表 3 と表 4 を比較すると，提案プロトコルによってモデル所有者の計算時間が HD で 92%，HG で 93%，SP で 97%減少した．ランダム値の暗号化によってデータ所有者の計算時間の増加が見られるが，モデル所有者の計算時間改善に比べればわずかである．

4.2.2　通信量

表 5 は先行研究 [3] と提案プロトコルにおける通信量を示している．E は一つの暗号文のサイズである．それぞれの通信回数はプロトコルから $N_i + N_p$，$N_i + N_t \times 2N_p + 1$ となる．表 5 ではデータセットによって提案プロトコルの通信時間の方が小さい場合と大きい場合の両方が確認できる．これは決定木アンサンブルのパラメータ (N_t, d) に依存しており，慎重に設定しなければならない．次章では大きな木の数や深さを必要とするデータセットを用いて，パラメータ (N_t, d) による計算時間や通信量の比較を行う．

4.3　モデル構築指針の評価

図 6 は精度 0.80，0.85，0.90 の達成に必要なパラメータ (N_t, d) の組を示している．横軸が必要な木の深さ，縦軸が必要な木の数である．3.5 節における議論から，様々なパラメータ (N_t, d) において計算量 $\mathcal{O}(N_t \times 2^{2d})$ と通信量 $\mathcal{O}(N_t \times 2^{d})$ を最小とする (N_t, d) を選ぶことが最善であること確認する．

表 6 は，精度 85%を達成するパラメータ組 (N_t, d) に対する $N_t \times 2^{2d}$ と $N_t \times 2^{d}$ の値を示している．この見積もりでは $(N_t, d) = (57, 2)$ の時に計算時間が最小，$(N_t, d) = (28, 3)$ の時に通信時間が最小となっている．この見積もりが正しいかどうか，提案プロトコルを実行して確認する．

表 7，表 8 に 4 つのパラメータ組 (N_t, d) に対する提案プロトコルの計算時間，通信量を示す．表 7，表 8 か

表 8　MNIST データセットの通信時間
Table 8　Amount of communication of MNIST

(N_t, d)	(57, 2)	(28, 3)	(16, 4)	(8, 5)
Communication	686E	674E	770E	770E

ら，見積もり通り $(N_t, d) = (57, 2)$ の時に計算時間最小，$(N_t, d) = (28, 3)$ の時に通信量最小となっている．3.5 節に示したモデル構築指針を用いて効率的な実行が可能な決定木アンサンブルが構築できることを確認した．

5. 結論

本研究では決定木アンサンブルをベースとし，準同型暗号を用いた秘匿推論プロトコルを提案した．モデル所有者が推論計算を委託できるように，提案プロトコルではモデル所有者が秘密鍵を持ち，決定木構造を暗号化している．また提案プロトコルは分類精度向上や計算時間，通信量削減のために効果的なアンサンブル秘匿計算を採用した．提案プロトコルをセキュリティ，計算・通信量や分類精度の観点から評価した．モデルパラメータの安全性を確認し，モデル所有者に必要な計算時間を最大 97%削減した．さらに，MNIST データセットによる評価によって決定木アンサンブルの優位性を実証した．提案プロトコルは，モデルパラメータと構造を開示せずにモデル所有者がモデル配布が可能な最初の試みであり，秘匿推論の適用範囲の拡大に貢献する．

謝辞

本研究は，JST, CREST, JPMJCR19K5 の支援を受けたものである．

参考文献

[1] Batina, L., Bhasin, S., Jap, D. and Picek, S.: CSINN: Reverse Engineering of Neural Network Architectures Through Electromagnetic Side Channel, *28th USENIX Security Symposium (USENIX Security 19)*, Washington, USENIX, pp. 515–532 (2019).

[2] Tai, R. K., Ma, J. P., Zhao, Y. and Chow, S. S.: Privacy-preserving decision trees evaluation via linear functions, *European Symposium on Research in Computer Security*, Berlin, Springer, pp. 494–512 (2017).

[3] LU Wen-jie; ZHOU, J.-J. S. J.: Non-interactive and output expressive private comparison from homomorphic encryption, *: Proceedings of the 2018 on Asia Conference on Computer and Communications Security*, New York, Association for Computing Machinery, pp. 67–74 (2018).

[4] Gentry, C.: Fully homomorphic encryption using ideal lattices, *Proceedings of the forty-first annual ACM symposium on Theory of computing*, New York, Association for Computing Machinery, pp. 169–178 (2009).

[5] Brakerski, Z., Gentry, C. and Vaikuntanathan, V.: (Leveled) fully homomorphic encryption without bootstrapping, pp. 1–36 (2014).

[6] Joye, M. and Salehi, F.: Private yet efficient decision tree evaluation, *IFIP Annual Conference on Data and Applications Security and Privacy*, Berlin, Springer, pp. 243–259 (2018).

[7] Alpaydin, E.: *Introduction to machine learning*, MIT press (2020).

[8] Hastie, T., Rosset, S., Zhu, J. and Zou, H.: Multi-class adaboost, *Statistics and its Interface*, Boston, International Press of Boston, pp. 349–360 (2009).

[9] Liaw, A., Wiener, M. et al.: Classification and regression by randomForest, *R news*, Vol. 2, No. 3, pp. 18–22 (2002).

[10] Chen, T. and Guestrin, C.: Xgboost: A scalable tree boosting system, *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794 (2016).

[11] Tueno, A., Kerschbaum, F. and Katzenbeisser, S.: Private evaluation of decision trees using sublinear cost, *Proceedings on Privacy Enhancing Technologies*, Vol. 2019, No. 1, pp. 266–286 (2019).

[12] Asuncion, A. and Newman, D.: UCI machine learning repository (2007).