科学研究費助成事業（基盤研究(S)）公表用資料
〔令和５(2023)年度 中間評価用〕

令和５年３月３１日現在

研　　究　　期　　間：２０２１～２０２５
課　　題　　番　　号：２１Ｈ０５０５２
研　究　課　題　名：圧縮秘匿計算による大規模データ処理

研究代表者氏名（ローマ字）：定兼 邦彦（SADAKANE Kunihiko）
所 属 研 究 機 関・部局・職：東京大学・大学院情報理工学系研究科・教授
研　究　者　番　号：２０３２３０９０

研究の概要：本研究では、「圧縮秘匿計算」という新概念を提案する。これは次のような概念である。(a) 秘匿計算：個人のプライバシーを保護するためにデータを暗号化したまま保存、計算する。(b) 圧縮索引：データに付加する補助情報を圧縮することで大規模データを省資源の計算機で高速に処理する。(c) 圧縮計算：データを圧縮することで冗長性を排除し、データからの学習・解析の性能・速度を向上させる。

研究分野：情報学、情報学基礎、情報学基礎理論
キーワード：秘匿計算、データ圧縮

１．研究開始当初の背景

人間中心社会 (Society 5.0) において様々な知識や情報を共有する際に発生する問題には次のものがある。

- 大量のデータを学習・解析する際の計算コスト
- 個人情報を共有する際のプライバシー

プライバシーを保護しつつ計算を行うためにはデータを匿名化・暗号化したまま学習・解析を行う必要がある。また、学習・解析をする際の計算コストとしては、計算時間と計算機のリソース（処理速度やメモリ量）がある。

２．研究の目的

プライバシーを保護しつつ計算を行う技術として秘匿計算 (secure computation) がある。秘匿計算とは、データを暗号化したまま計算を行う技術である。秘匿計算の応用としては、DNA配列のデータベースがある。多くの患者のDNA配列情報を集め、解析することで遺伝性疾患の治療に役立てることができる。しかし、DNA配列情報は「究極の個人情報」であるため、患者のプライバシーを保護するために暗号化しておく必要がある。通常の暗号化では、暗号化されているデータに対して計算を行う際には、いったん復号し、計算後に再暗号化する必要がある。つまり、計算を実行する側に秘密が漏れてしまう。一方、秘匿計算では、データを暗号化したまま計算でき、計算を実行する側は個々のデータの情報は得られず計算結果のみが得られる。つまり個人のプライバシーを保ったまま学習・解析処理が行える。

計算コストについては、スパコンを使えば解決するというものではない。大規模データを扱う場合には計算に必要なデータを高速に取り出すための「索引」と呼ばれるデータ構造が必須となる。しかし、従来の索引構造はサイズが大きく、大規模データに対しては適応できない。このような問題を解決するための技術として、簡潔データ構造(succinct data structure)がある。簡潔データ構造とは、データを圧縮したまま計算を行う技術である。これにより、大規模データを省メモリ量の計算機で高速に処理することが可能になっている。応用としては、DNA配列のアセンブリなどがある。

本研究では、秘匿計算のアルゴリズムを発展させ、簡潔データ構造と組み合わせることで、データを圧縮したまま、暗号化したまま高速に処理することが目的である。

３．研究の方法

本研究は大きく次の3つのテーマに分類できる：(1) 圧縮秘匿基盤、(2) 圧縮秘匿学習、(3)圧縮秘匿解析である。それぞれについて説明する。

(1) 圧縮秘匿基盤技術の開発
基盤技術である、秘匿計算が可能な簡潔データ構造の開発を行う。

(2) 圧縮秘匿学習の研究
自然言語等の非定型データを秘匿分析可能にする計算モデルを開発する。本研究では、プライバシーを含むデータ群を匿名化せずにそのまま暗号化し、復号することなく分析できる技術を構築する。

(3) 圧縮秘匿解析の研究
ゲノム解析に即した圧縮秘匿解析技術を開発するとともに、大規模個人DNA配列データベースに対するゲノムワイド関連解析へのそのような圧縮秘匿解析技術の展開による高度化、個人医療解析への同様の圧縮秘匿解析技術の展開による高度化を狙っていく。

４．これまでの成果

ソート（データを小さい順に並び替える）はデータの処理において最も重要な処理と言ってよく、多くの計算の内部で使われる。本研究で開発したソートアルゴリズムは、基数ソート（radix sort）に基づいている。これは、数を2進数で表現した際に、まず最下位のビットに基づきソートを行い、次に下から2番目のビットに基づきソートし、というように桁ごとにソートを行うものである。これをそのまま実行すると W ビットの数のソートは W 回のラウンドを必要とする。これを高速化するには1回のラウンドで複数ビットを処理する必要がある。1回に L ビットに基づきソートを行えばラウンド数は W/L に削減される。しかし秘匿計算においては、複数ビットに基づくソートを行う際には問題が生じる。それは、1回のラウンドあたりの通信量が増大してしまうという点である。既存手法では、1ラウンドあたり O(2^L N log N) ビットのオンライン通信量が必要であったが、これを O(NL) ビットに削減した。

また、入力文字列の全ての接尾辞をソートするアルゴリズムを開発した。通常の計算モデルでは、長さ n の文字列に対し線形（O(n)）時間で接尾辞をソートすることができるが、秘匿計算モデルでは効率的な（O(n^2) よりも高速な)アルゴリズムは存在しなかった。本研究では O(n log^2 n) 時間の秘匿計算アルゴリズムを与えた。ラウンド数は O(log^2 n)、 通信量は O(n log^3 n) ビットである。接尾辞ソートはDNA解析等の文字列処理では必須の処理である。文字列検索の索引である接尾辞配列を構築するには接尾辞をソートする必要があり、またこの索引を圧縮した圧縮接尾辞配列やFM-indexの構築でも接尾辞ソートが必要になる。これまでの索引構築は、まず平文のDNA配列に対し接尾辞ソートを行い平文のFM-indexを作成し、それを秘匿化する手法が取られているため、索引を構築するアルゴリズムの実行者にはDNA配列の情報が漏れてしまう。本研究のアルゴリズムは、秘匿化されたDNA配列から秘匿化された索引を構築できるものである。

決定木とは、機械学習のモデルの1つであり、データをクラスに分類する基準を表すものである。決定木の葉はクラスに対応し、内部ノードは2つの子を持ち、またノードにはデータを分類する基準が書かれている。クラスに分類したいデータが与えられたときに、決定木の根からノードの基準に従って進んでいき、葉に到達したときにその葉に対応するクラスにデータを分類する。秘匿決定木では、決定木の形を秘密にしたまま、秘匿化されたデータに対しその属するクラスを求めることができる。しかし既存手法では決定木を完全2分木にマップするため、最悪の場合には木の高さに対して指数的な領域を必要とする。本研究の手法では、木の高さに比例する領域が増えるだけである。計算機実験により、本手法は既存手法よりもオンラインの通信量が少ないことが示された。

５．今後の計画

引き続き、ビッグデータに対する効率的な秘匿計算アルゴリズムの開発を行う。個別のアルゴリズムの開発の他に、秘匿計算アルゴリズムの汎用的な開発法についても検討する。

また、秘匿計算のライブラリを開発し、無償で公開する。

６．これまでの発表論文等（受賞等も含む）

神保 洸貴, 定兼 邦彦. 秘匿接尾辞ソーティングとその応用. 暗号と情報セキュリティシンポジウム (SCIS), 1B2-2, 2023.

吉田 勇輝, 定兼 邦彦, 戸澤 一成. 秘密計算基数ソートの通信量の削減. 暗号と情報セキュリティシンポジウム (SCIS), 1B2-3, 2023.

Yamamoto Akito, Shibuya Tetsuo. Privacy-Preserving Statistical Analysis of Genomic Data Using Compressive Mechanism with Haar Wavelet Transform. Journal of Computational Biology, Vol. 30, pp. 176〜188, 2023. DOI: 10.1089/cmb.2022.0246

Mohammad Nabil Ahmed, Kana Shimizu. Private Evaluation of a Decision Tree based on Secret Sharing. Information Security and Cryptology – ICISC 2022, Revised Selected Papers. Vol. 13849, pp. 186-209, 2023.

Ono Shinji, Takata Jun, Kataoka Masaharu, I Tomohiro, Shin Kilho, Sakamoto Hiroshi. Privacy-Preserving Feature Selection with Fully Homomorphic Encryption. Algorithms, Vol. 15:229, 2022. DOI: 10.3390/a15070229

7．ホームページ等

https://researchmap.jp/sada