SCIS 2022 2022 Symposium on
Cryptography and Information Security
Osaka, Japan & Online, Jan. 18 – 21, 2022
The Institute of Electronics,
Information and Communication Engineers

大規模データにも対応した秘密計算階層型クラスタリング
Privacy Preserving Hierarchical Clustering for Large Scale Data

三品気吹 * 五十嵐大 * 濱田浩気 * 菊池亮 *
Ibuki Mishina Dai Ikarashi Koki Hamada Ryo Kikuchi

キーワード 秘密計算，機械学習，教師無し学習，階層型クラスタリング

あらまし

秘密計算ではデータを暗号化したまま計算するため，プライバシーを保護したまま安全にデータ分析を行う方法として注目されている．中でも近年のデータ分析手法の主流である機械学習を秘密計算上で実現する研究は活発に行われている．本稿では教師無し学習に分類される手法の一つである「階層型クラスタリング」の秘密計算上での実現に取り組む．

階層型クラスタリングは，全データ間の距離を計算し，近いものから順番に仲間分けしていくクラスタリング手法である．データ間の距離を全て計算するため，平文のアルゴリズムでも計算量がレコード数 n に対して $n^2 \log n$ と重く，更に秘密計算ではデータ数を秘匿しながら行う処理があるため n^3 になってしまう．データ数に対する処理時間の増加が非常に大きいため，秘密計算階層型クラスタリングの既存手法 [1] では，処理速度の都合により数百件件程度までのクラスタリングが現実的な範囲であった．

しかし，本稿では Lance-Williams の更新式 [2] と呼ばれる効率的なクラスタ間の距離計算の手法を用いて，階層型クラスタリングのボトルネックとなる距離計算のアルゴリズムを改善した．これにより，3000 件のクラスタリングを 1 時間半程度で実行できるようになり，数千件程度のデータセットでも現実的な時間で処理できるようになった．また，以前の手法では 100 件のクラスタリングに約 44 秒かかっていたものが，本稿の提案手法では 7 秒で処理できるようになり，比較的小規模なデータでも 6 倍程度の高速化を実現した．

* NTT 社会情報研究所，東京都武蔵野市緑町 3-9-11, NTT Social Informatics Laboratories, 3-9-11 Midoricho, Musashino-shi, Tokyo-to

参考文献

[1] 三品気吹，五十嵐大，濱田浩気，菊池亮,“秘密計算によるプライバシー保護階層型クラスタリング,” CSS2021

[2] G. N. Lance, W. T. Williams,“A General Theory of Classificatory Sorting Strategies: 1. Hierarchical Systems,” The Computer Journal, Volume 9, Issue 4, February 1967, Pages 373–380