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

# 秘密計算による k-means 法と k-means++法
# K-Means Clustering and K-Means++ in secure computation

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，k-means

## あらまし

秘密計算はデータを暗号化したまま計算する技術である．そのため，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おり，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．これまでの研究の多くは，ロジスティック回帰やニューラルネットワークといった「教師あり学習」に分類されるものであったが，本稿では「教師無し学習」に分類される手法の一つである「k-means 法 [1]」の秘密計算上での実現に取り組む．

クラスタリング手法には，k-means 法以外でも有名なものとして「階層型クラスタリング」があるが，データ数の二乗に比例して計算量が増える階層型クラスタリングと比較して，データ数に比例して計算量が増える k-means 法は大きなサイズのデータも高速で処理することができる．そのため，小規模なデータで詳細なクラスタリングを行う場合は階層型，大規模なデータに対して大まかなクラスタリングを行う場合は k-means 法という使いわけをするのが一般的である．

暗号などを用いたプライバシー保護 k-means クラスタリングの先行研究はいくつかあるが [2]，暗号化したままでは計算が困難な部分は平文で計算をしていたり，近似を用いて計算しているため本来の結果が得らないといった課題がある．本稿の提案手法では，データを一度も平文にすることなく k-means クラスタリングを行う．

また k-means 法ではランダムな初期値を用いて学習を行うが，クラスタリング結果の初期値依存性が高いため，単純な初期化方法では安定した結果が得にくいことが知られている．そのため，平文では一般的に k-means++[3] という初期化方法が用いられる．k-means++では初期値を生成する際に，完全なランダムではなく，データ間の距離に応じた重み付けをすることによって，初期値が近くに密集してしまうのを防ぐ．基本的には完全にランダムな初期値を用いるより，k-means++で生成した初期値を用いるほうが収束が速く，良好なクラスタリング結果が得やすいことが知られている．しかし，プライバシー保護 k-means クラスタリングの先行研究では，k-means++も秘密計算で行っているものが無いため，本稿では，k-means 法だけでなく k-means++法についても秘密計算上で実現するアルゴリズムを提案する．

iris データを用いた実験では，平文で k-means クラスタリングを行った結果と一致することが確認でき，また処理時間に関しては 1000 件のクラスタリングを 3 秒程度で行うことができた．

* 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] Hartigan, John A., and Manchek A. Wong. ”Algorithm AS 136: A k-means clustering algorithm.” Journal of the royal statistical society. series c (applied statistics) 28.1 (1979): 100-108.

[2] Vaidya, Jaideep, and Chris Clifton. ”Privacy-preserving k-means clustering over vertically partitioned data.” Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. 2003.

[3] Arthur, David, and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. Stanford, 2006.