様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　６ 年　 ５ 月 ２７ 日現在

機関番号：１４４０１
研究種目：若手研究
研究期間：2020～2023
課題番号：２０Ｋ１９７５６

研究課題名（和文）大規模・複雑データに対するクラスタリング法の開発とその理論的性質の解明

研究課題名（英文）Development of clustering method for large and complex data and its theoretical properties

研究代表者

寺田　吉壱（Terada, Yoshikazu）

大阪大学・大学院基礎工学研究科・准教授

研究者番号：１０７３８７９３

交付決定額（研究期間全体）：（直接経費）　　2,300,000円

研究成果の概要（和文）：本研究では，大規模なクラスタリングにおける汎用的な計算量削減方法の開発とその理論的性質の解明，及び，柔軟なグループ構造を階層的に捉えることが可能なconvex clusteringに対する高速なアルゴリズムの開発を行なった．本研究で開発した手法を用いることで，100万を超えるデータ点に対して，ノートPCを用いた場合でも1分以内に複雑なクラスタリング法を実行することが可能となり，大規模かつ複雑なデータに対しても高速に背後のクラスタ構造を推定することができるようになった．

研究成果の学術的意義や社会的意義
近年のデータの大規模化・複雑化に伴い，データからグループ構造を発見するためのクラスタリング法の重要性が増している．しかし，これまで大規模データに対しては，単純なクラスタ構造しか捉えられないクラスタリング法しか適用ができなかった．本研究成果により，クラスタリング法を必要とする任意の分野において，短時間かつ容易に，大規模データから複雑なクラスタ構造を推定することが可能となった．本研究を応用することで，様々な応用分野において，新たな知見の発見などが期待できる．

研究成果の概要（英文）：In this study, we developed a general computational cost reduction method for large-scale clustering and showed its theoretical properties. Additionally, we developed a fast algorithm for convex clustering that can flexibly capture hierarchical group structures. By using the proposed methods, it is possible to perform complex clustering techniques on over a million data points within one minute, even using a laptop. This enables the rapid estimation of underlying cluster structures in large and complex data.

研究分野：教師なし学習

キーワード：クラスタリング　高速化

様　式　C－１９、F－１９－１（共通）

１．研究開始当初の背景

クラスタリング法は，大規模・複雑なデータから背後のグループ（クラスタ）構造を獲得するために有用である．K-means 法など単純な方法に関しては，サンプルサイズが 100 万を超えるデータに対しても，例えば，mini-batch k-means 法によって適用が可能である．一方で，k-means 法は複雑なクラスタ構造を捉えるには不十分であることが知られている（図 1）．そこで，実データ解析には，より柔軟なクラスタ構造を捉えることのできる spectral clustering などの方法が望ましい．しかし，これらの方法はその計算コストの高さから大規模なデータに対しての適用は推奨されておらず，計算量の緩和が課題となっている．実際に，Python の機械学習ライブラリ scikit-learn では，サンプルサイズが 1 万を超えるデータに対してはこれらの方法の適用を推奨していない．

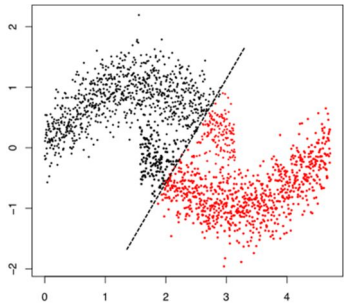

図 1 k-means法の適用結果

２．研究の目的

上述のような背景から，本研究では，複雑なクラスタ構造を捉えることが可能で，かつ，大規模なデータに対しても高速に実行可能なクラスタリング法を開発することが目的である．そして，大きく分けて 2 つの目的がある．1 つ目は「(1) クラスタリング法の高速かつ汎用的な近似方法の開発」，2 つ目は「(2) 柔軟な既存のクラスタリング法に対して，そのアルゴリズムの改良による高速化」である．

３．研究の方法

(1) クラスタリング法の高速かつ汎用的な近似方法の開発

本研究では，様々なクラスタリング法に適用が可能な汎用的な近似法を開発する．Yan et al. (2009; KDD) では，spectral clustering (SC) に対する k-mean 法に基づく近似法 (KASP) を提案している．KASP は， クラスタ数を多く設定した k-means 法を大規模データに適用し，得られたクラスタ中心をデータの代表点とし，代表点に対して SC を適用する．そして，推定された代表点のラベルを代表点に属する元のデータ点のラベルとする方法である．大きな利点は，その簡便性（計算コストの低さ）と汎用性である．実際に，SC 以外の方法に対してもこのアプローチによって近似することができる．一方で，先行研究における理論解析は不十分であり，KASP によって SC が近似できる理論的な保証は与えられていない．実際に，代表点の経験分布が背後のデータ分布とズレるため，KASP は深刻なバイアスをもっていることが理論的に示される．そこで，これらのズレを修正するための方法を開発する．

(2) 柔軟な既存のクラスタリング法に対して，そのアルゴリズムの改良による高速化

本研究では， 柔軟なクラスタ構造を捉えることが可能な convex clustering に対する高速なアルゴリズムの開発を行う．Convex clustering (CC) は k-means 法と階層的クラスタリングの両方の性質を備えたクラスタリング法であり， 重みを適切に選択することにより， 柔軟なクラスタ構造を捉えることが可能である．一方，既存のアルゴリズムの計算コストは高く，大規模データへの適用は困難である．そこで，MM (Majorization Minimization) アルゴリズムに基づく高速なアルゴリズムの開発と，木構造の重みを利用した CC の動的計画法に基づく高速なアルゴリズムの開発を行う．

４．研究成果

(1) データの背後の分布(母集団分布)の構造を壊さないようなデータの代表点の計算方法である密度保存ベクトル量子化法 (Density-Preserving Vector Quantization; DPVQ) を提案した．DPVQ は，単純な重み付き k-means 法であり，大規模なデータに対しても容易に適用可能である．また，提案手法によって生成された代表点の経験分布が漸近的に 母集団分布に収束することを証明した． 提案手法により生成した(サンプルサイズよりも少ない)代表点に対して， クラスタリング法を適用し，その結果を元の データに反映させることで，大幅に計算コストを削減することができる．

一方で，DPVQ は密度推定を必要とするため，高次元データに対しては不安定となる．そこで，既存の KASP の代表点に重みを付与することで，KASP の問題点が解決できることを示した（この方法は，2 次の VQAP と呼ぶ.）．しかし，spectral clustering をはじめ，いくつかのクラスタリング法は，経験損失最小化として定式化できないため，重みを考慮することは自明ではない．そのため，spectral clustering の統計理論を用いることで，重み付き spectral clustering を開発し，近似法の適用を可能とした．

KASP の改良は，その性質から密度の低い点も代表するため，DPVQ よりも効率が悪いという問題点がある． これを解消するため，r(<2)次のベクトル量子化を用いた代表点を利用した方法

(VQAP)を提案した．また，r次のベクトル量子化を高速に解くためのアルゴリズムを提案した．
以上の研究成果より，代表点を用いた汎用的かつ高速な近似法の開発とその性質の解明を達成することができた．以下の図 2 は，各代表点の作成方法が，データ分布の密度に対してどのような代表点を生成しているかを表している（赤文字が提案手法）．DPVQ が最も効率よく背後の分布を代表していることが分かる．また，提案手法（VQAP）を用いることで，27 次元 100 万点のデータに対して，ノート PC を用いて，1 分以内に spectral clustering の結果を得ることができる．また，提案手法では，既存手法のようなバイアスは生じないため，既存手法（KASP）よりも大幅にクラスタリングの精度を大幅に改善することができた．

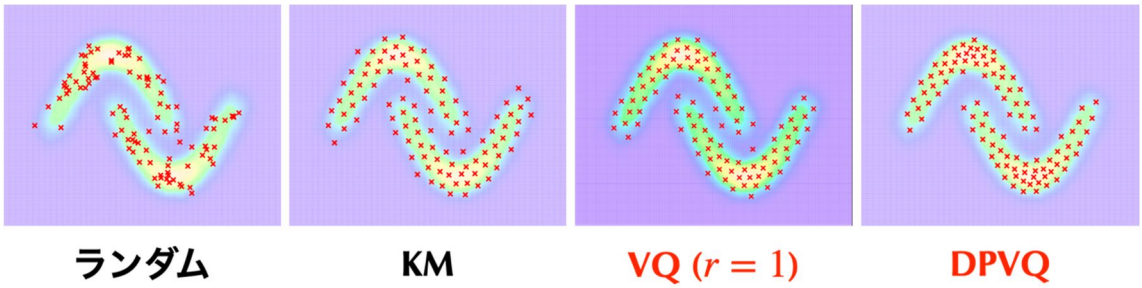

図 2 各手法により生成した代表点とデータ分布（密度関数）の関係

(2) 本研究では，まず重みが定数の場合の L1 convex clustering に対する高速なアルゴリズム (C-PAINT) の開発を行った．C-PAINT は，動的計画法に基づくアルゴリズムであり，L1 convex clustering をサンプルサイズに対して線形な時間で実行することができる．実際に，一千万点のデータに対してノート PC を用いても 2 分弱で解を得ることができるため，非常に大規模なデータにも適用できる．一方で，C-PAINT では，重みを定数としているため，柔軟なクラスタ構造が捉えられないという問題点がある．そこで，fused lasso は重みが木構造をもつときに動的計画法により効率良く最適化できることに注目し，重みが木構造をもつ場合の L1 convex clustering に対して，大規模データに適用可能な非常に効率的なアルゴリズム（TGCC）を提案した．この方法を用いれば，100 万点のデータに対しても，1 分ほどで cluster path と呼ばれる解の軌道と階層的クラスタ構造を得ることができる．また，重みを木構造まで一般化することができたため，複雑なクラスタ構造を捉えることができるようになった．実際に，重みとして最小全域木を用いた場合，古典的な階層的クラスタリング法である最短距離法（SLC）と同等の情報を用いているが，TGCC は最短距離法の問題点（鎖現象）を解消し，柔軟なクラスタ構造を捉えることが可能となっている（図 3）．さらに，対象と変数と同時にクラスタリングする biclustering への TGCC の拡張と変数選択を同時に実行できる sparse clustering への TGCC の拡張を行なった．これに関連して，非線形なクラスタリング法である kernel k-means 法を拡張した sparse kernel k-means 法を提案した．これらに加えて，一般の重みを伴った L2 convex clustering に対して，Majorization Minimization algorithm に基づく効率の良いアルゴリズム (CCMM) を開発した．

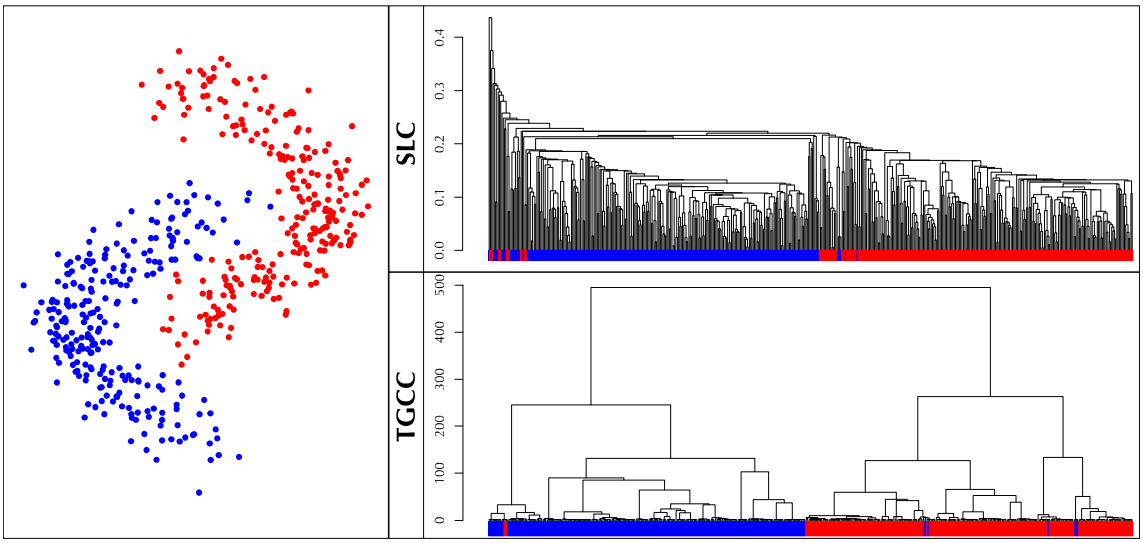

図 3 two moon dataと最短距離法と提案手法（TGCC）の適用結果

５．主な発表論文等

〔雑誌論文〕　計9件（うち査読付論文　8件／うち国際共著　1件／うちオープンアクセス　4件）

１．著者名 Terada Yoshikazu、Shimodaira Hidetoshi	４．巻 75
２．論文標題 Selective inference after feature selection via multiscale bootstrap	５．発行年 2022年
３．雑誌名 Annals of the Institute of Statistical Mathematics	６．最初と最後の頁 99～125
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s10463-022-00838-2	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Hirose Kei、Terada Yoshikazu	４．巻 1
２．論文標題 Sparse and Simple Structure Estimation via Prenet Penalization	５．発行年 2022年
３．雑誌名 Psychometrika	６．最初と最後の頁 1～26
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s11336-022-09868-4	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Morikawa Kosuke、Nagao Hiromichi、Ito Shin-ichi、Terada Yoshikazu、Sakai Shin’ichi、Hirata Naoshi	４．巻 226
２．論文標題 Forecasting temporal variation of aftershocks immediately after a main shock using Gaussian process regression	５．発行年 2021年
３．雑誌名 Geophysical Journal International	６．最初と最後の頁 1018～1035
掲載論文のDOI（デジタルオブジェクト識別子） 10.1093/gji/ggab124	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Bingyuan Zhang, Jie Chen, Yoshikazu Terada	４．巻 161
２．論文標題 Dynamic visualization for L1 fusion convex clustering in near-linear time	５．発行年 2021年
３．雑誌名 Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence	６．最初と最後の頁 515～524
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Poignard Benjamin、Terada Yoshikazu	4．巻 14
2．論文標題 Statistical analysis of sparse approximate factor models	5．発行年 2020年
3．雑誌名 Electronic Journal of Statistics	6．最初と最後の頁 3315～3365
掲載論文のDOI（デジタルオブジェクト識別子） 10.1214/20-EJS1745	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Terada Yoshikazu、Ogasawara Issei、Nakata Ken	4．巻 14
2．論文標題 Classification from only positive and unlabeled functional data	5．発行年 2020年
3．雑誌名 The Annals of Applied Statistics	6．最初と最後の頁 1724～1742
掲載論文のDOI（デジタルオブジェクト識別子） 10.1214/20-AOAS1404	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Terada Yoshikazu、Hirose Ryoma	4．巻 129
2．論文標題 Fast generalization error bound of deep learning without scale invariance of activation functions	5．発行年 2020年
3．雑誌名 Neural Networks	6．最初と最後の頁 344～358
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.neunet.2020.05.033	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Guan Xin、Terada Yoshikazu	4．巻 144
2．論文標題 Sparse kernel k-means for high-dimensional data	5．発行年 2023年
3．雑誌名 Pattern Recognition	6．最初と最後の頁 109873～109873
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.patcog.2023.109873	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Daniel J. W. Touw, Patrick J. F. Groenen, Yoshikazu Terada	4．巻 －
2．論文標題 Convex Clustering through MM: An Efficient Algorithm to Perform Hierarchical Clustering	5．発行年 2023年
3．雑誌名 arXiv:2211.01877	6．最初と最後の頁 1～27
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

〔学会発表〕　計13件（うち招待講演　11件／うち国際学会　5件）

1．発表者名 寺田吉壱，山本倫生
2．発表標題 ベクトル量子化による大規模クラスタリングの近似法とその性質
3．学会等名 科研費シンポジウム「データサイエンスと周辺領域の双方向的理解への挑戦」（招待講演）
4．発表年 2022年

1．発表者名 寺田吉壱，山本倫生
2．発表標題 代表点を用いた大規模クラスタリングの近似法とその性質
3．学会等名 科研費シンポジウム「大規模複雑データの理論と方法論～新たな発展と関連分野への応用～」（招待講演）
4．発表年 2022年

1．発表者名 Yoshikazu Terada, Masaki Sasaki
2．発表標題 On weak convergence of recovered functional data
3．学会等名 15th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2022)（招待講演）
4．発表年 2022年

１．発表者名
Yoshikazu Terada, Michio Yamamoto

２．発表標題
Regularized functional subspace clustering

３．学会等名
CSDA & EcoSta Workshop on Statistical Data Science (SDS 2022)（招待講演）

４．発表年
2022年

１．発表者名
Yoshikazu Terada, Michio Yamamoto

２．発表標題
Fast Approximation for large-scale clustering

３．学会等名
The 11th Conference of the IASC-ARS The Asian Regional Section of the International Association for Statistical Computing（招待講演）（国際学会）

４．発表年
2022年

１．発表者名
寺田吉壱，山本倫生

２．発表標題
クラスタリングにおける汎用的な計算コスト削減法について

３．学会等名
2021年度日本分類学会シンポジウム

４．発表年
2021年

１．発表者名
寺田吉壱、山本 倫生

２．発表標題
クラスタリングにおける汎用的な計算コスト削減法について

３．学会等名
2020年度統計関連学会連合大会

４．発表年
2020年

１．発表者名
寺田吉壱、山本 倫生

２．発表標題
大規模なクラスタリングにおける計算量削減法について

３．学会等名
第5回 統計・機械学習若手シンポジウム（招待講演）

４．発表年
2020年

１．発表者名
Yoshikazu Terada

２．発表標題
A statistical theory of clustering

３．学会等名
Forum "Math-for-Industry" (FMfI) 2023（招待講演）（国際学会）

４．発表年
2023年

１．発表者名
寺田吉壱

２．発表標題
クラスタリング法の統計理論と応用

３．学会等名
第43回情報計測オンラインセミナー（招待講演）

４．発表年
2023年

１．発表者名
Yoshikazu Terada

２．発表標題
On some properties of reconstructed trajectories from sparse longitudinal data

３．学会等名
The 15th Scientific Meeting of the Classification and Data Analysis Group（招待講演）（国際学会）

４．発表年
2023年

１．発表者名 Yoshikazu Terada, Hidetoshi Matsui
２．発表標題 On smoothing for spatial functional data
３．学会等名 The 6th International Conference on Econometrics and Statistics（招待講演）（国際学会）
４．発表年 2023年

１．発表者名 Yoshikazu Terada, Hidetoshi Matsui
２．発表標題 Dynamic prediction for variable-domain functional data
３．学会等名 The 12th Conference of the IASC-ARS (IASC-ARS2023)（招待講演）（国際学会）
４．発表年 2023年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関			
オランダ	Erasmus University Rotterdam			