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　４ 年　 ６ 月 ２１ 日現在

機関番号：１２６０１
研究種目：若手研究
研究期間：2019～2021
課題番号：１９Ｋ２０３３６

研究課題名（和文）大規模部分空間クラスタリングのための凸最適化スキームの構築とその理論保証

研究課題名（英文）Construction of convex optimization schemes for large-scale subspace clustering and its theoretical guarantees

研究代表者
松島　慎（Matsushima, Shin）

東京大学・大学院情報理工学系研究科・准教授

研究者番号：９０７２１８３７

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：本研究の目的は大規模部分空間クラスタリングのための凸最適化スキームの構築とその理論保証である。
本研究ではS5Cと呼ばれる大規模部分空間クラスタリングのためのアルゴリズムを開発した。従来はデータ数の二乗以上の計算量が必要である学習を、本提案アルゴリズムではデータ数に比例する計算量で達成できることを実験的にも理論的にも示した。本研究成果は機械学習の最も重要な国際会議の一つであるThirty-Third Conference on Neural Information Processing Systemsに採択され、バンクーバーにて発表を行った。

研究成果の学術的意義や社会的意義
部分空間クラスタリングは理論保証が難しいクラスタリング手法において理論的に美しい結果が知られているという意味で有意義な手法である。本研究ではこの部分空間クラスタリングが大規模データにも適用可能であることを示し、実験的にも有効性を検証した点に学術的な意義がある。部分空間クラスタリングに関する研究が機械学習の主要な国際会議に採択されたことで、機械学習の発展には実用性だけではなく理論的な美しさも重要であるという理念の補強の一助となったことにも社会的な意義があると考えられる。

研究成果の概要（英文）：The goal of this project is to establish a convex optimization scheme for large-scale subspace clustering and its theoretical guarantees.
This research project has developed an algorithm called S5C for large-scale subspace clustering. This algorithm empirically and theoretically achieves linear complexity to the data size, whereas previous methods required more than quadratic complexity to the data size. This result has been accepted in the thirty-third conference on neural information processing systems which is one of the most important international conferences on machine learning and presented in Vancouver.

研究分野：数理情報学

キーワード：機械学習　凸最適化　クラスタリング　部分空間クラスタリング

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

データ分析技術者の社会的なニーズなどから見られるように、与えられた複雑な高次元データが持つであろう性質や構造に応じて数理モデルを設計し、データの持つ隠れた特性や構造を明らかにするという機械学習の側面は社会的にますます重要になってきている。特に、画像認識、文書のトピック推定やバイオインフォマティクスに現れるような高次元データはそれぞれのデータが未知の低次元の部分空間に存在すると考えられることが多い。このようなデータの性質を利用して、隠れた低次元部分空間の構造を明らかにするデータ分析手法として部分空間クラスタリングがある。部分空間クラスタリングの特徴としては、応用面での有効性もさることながら、理論的に厳密な原理がある点、アルゴリズムが簡潔に記述可能であることなどが挙げられる。部分空間クラスタリングの手法は二段階のステップで構成される。一段階目では、データ間の類似度を表現する類似度行列を凸最適化問題を解くことにより学習する。二段階目では、得られた類似度行列に基づきスペクトルクラスタリングを行う。K-means 法などの低次元のユークリッド座標で直接クラスタが学習できる場合と異なり、部分空間クラスタリングは類似度行列の学習を介することで、部分空間という複雑な構造に基づくクラスタリングを厳密な理論保証の下で可能にする。

大規模な部分空間クラスタリングは一段階目のステップで大規模な類似度行列を学習する必要がある。これは本質的にデータ数の２乗の規模のパラメータを推定する問題になるため、現在の計算機では１００万を超えるような大規模なデータに適応するのは困難である。また、二段階目のステップにおいては大規模な類似度行列の固有ベクトルを求める必要があり、さらなる計算量的な困難がある。この問題意識の中で、データ数に対して線形時間で学習可能な部分空間クラスタリングアルゴリズムが近年盛んに研究されているが、多くの文献は一段階目の類似度行列学習のアルゴリズムであるか、全データを利用せずに無作為に抽出した標本から構造を推定するアプローチをとっている[1,2,3]。前者の場合は二段階目のステップにおいて依然として多大な計算時間を必要とし、後者の場合は、多くのデータは学習に影響を与えないため、本質的に大規模データを利用しているとは言えない。結果として大規模な部分空間クラスタリングとしての性能はいまだに十分とは言えず、理論的な評価を与えられた線形時間で動作する部分空間クラスタリング手法も存在しない。

本研究課題における学術的問いは「部分空間クラスタリング手法はいかに効率的に大規模データを扱うことができるか」である。大規模な部分空間クラスタリングは上述の通り線形時間で学習可能なアルゴリズムが提案されているが、従来のアルゴリズムは、以下に示す２つの重要な疑問点に答えられていない。一つは、データがメモリ容量を超えた場合にも効率よく大規模データを扱えるのか、もう一つは大規模データを扱う効率的なアルゴリズムが、部分空間を正しく推定できることを適当な条件の下で保証できるか、という点である。本研究では、これらの疑問点を(3)において後述するアプローチを用いることで解消する。これらの疑問が肯定的に解決された場合、部分空間クラスタリングが確立されたデータマイニング手法としてデータ分析技術者に利用され、より広い現実社会の各分野での知識発見に貢献すると考えられる。

２．研究の目的

本研究の目的は「大規模部分空間クラスタリングのための汎用計算機に特化された凸最適化スキームの構築とその理論保証」である。すなわち次に掲げる２つの項目に関する研究を行う。さらに本研究の課題の全体像を以下に示す。

●大規模部分空間クラスタリングのための汎用計算機に特化された凸最適化スキームの構築

部分空間クラスタリングを大規模なデータに適用する場合、データやパラメータがメモリ容量を超えてアルゴリズムの動作が急激に遅くなるなどの問題が生じる。そこで、本研究では現在の汎用計算機の特質であるメモリ階層構造やマルチコアプロセッサを効率的に利用するアルゴリズムを申請者が確立してきた Cached Loops のスキームにより開発することで、実際に大規模な部分空間クラスタリングを安価な汎用計算機で可能にすることを目的とする。

●部分空間クラスタリングの大規模アルゴリズムの理論保証

部分空間クラスタリングの手法は真の各クラスタの部分空間の推定に関する理論的性能保証がある点が大きな利点である。本研究では、計算量のオーダーがデータ数の線形に抑えられるアルゴリズムによって、真の各クラスタの部分空間を正しく推定する確率が１に近づく条件の理論解析を目的とする。さらに、性能保証だけでなく、提案アルゴリズムの計算時間や必要なメモリ容量に関しても、理論的な上限の保証を与えることを目的とする。

３．研究の方法

●大規模部分空間クラスタリングのための汎用計算機に特化された凸最適化スキームの構築

データ全体を継続的かつ反復的に読み込みながら、学習のために重要なデータのみを選択的に RAM に記憶させることで、RAM 容量を超えるデータから効率的に学習できる部分空間クラスタリングアルゴリズムを構築し、スパース性や低ランク性などの解の特質を利用することでいかに大規模なデータから効率的な学習が可能であるかを明らかにする。具体的には、一般的な計算機の RAM 容量の 10〜100 倍程度のデータを学習すること、すなわち、数十〜数百 GB のデータを一般的な計算機を用いて扱うことができるアルゴリズムの構築を目指す。さらに実際にアルゴリズムの実装を行い、実験的にも大規模な画像認識タスク、文書分類タスクに応用し、既存のアルゴリズムとの性能比較を行いながら有効性を確認する。

●部分空間クラスタリングの大規模アルゴリズムの理論保証

計算量のオーダーがデータ数の二乗以上であるアルゴリズムを用いれば、データ数が増大するにつれ、真の各クラスタの部分空間を正しく推定する確率が１に近づくことが、様々な条件の下で証明されている[4,5,6]。大規模アルゴリズムの構築のためには類似度行列の近似計算が不可欠であるが、この近似が部分空間の推定に及ぼす影響の理論解析を行う。具体的には、既存の文献で課されている様々な条件を検討しながら、上述の事象がどのような場合に示されるかを検討する。実験的には少ないデータで全体のデータを代表させることで、各クラスタの部分空間が正しく推定できることから、上述の主張が証明できない場合は実験と理論にどのようなギャップがあるかを明らかにする。

４．研究成果

本研究の目的は大規模部分空間クラスタリングのための凸最適化スキームの構築とその理論保証である。
本研究では S5C と呼ばれる大規模部分空間クラスタリングのためのアルゴリズムを開発した。従来はデータ数の二乗以上の計算量が必要である学習を、本提案アルゴリズムではデータ数に比例する計算量で達成できることを実験的にも理論的にも示した。本研究成果は機械学習の最も重要な国際会議の一つである Thirty-Third Conference on Neural Information Processing Systems に採択され、バンクーバーにて発表を行った。

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計1件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 Shin Matsushima, Maria Brbic
２．発表標題 Selective Sampling-based Scalable Sparse Subspace Clustering
３．学会等名 Conference on Neural Information Processing Systems（国際学会）
４．発表年 2019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