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　元 年　 ９ 月　 ４ 日現在

機関番号：１４４０１
研究種目：若手研究(B)
研究期間：2016～2018
課題番号：１６Ｋ１６０２４

研究課題名（和文）ネットワークデータと関数データに対する教師なし学習を中心とした解析法の理論と応用

研究課題名（英文）Theory and application of unsupervised learning for Network data and functional data

研究代表者
寺田　吉壱（Terada, Yoshikazu）

大阪大学・基礎工学研究科・助教

研究者番号：１０７３８７９３

交付決定額（研究期間全体）：（直接経費）　　1,800,000円

研究成果の概要（和文）：近年，計測技術の発展によって，データの複雑化や大規模化が進んでいる．そして，教師なし学習の重要性が再認識されている．本研究では，ネットワークデータと関数データを主に扱い，教師なし学習に関連する解析法の開発，理論的性質の解明，実社会への応用を行った．具体的には，(1) グラフ分割に基づくクラスタリング法の理論的性質の解明，(2) 関数データに対する半教師付き判別法の提案とその理論的性質の解明及び実社会の問題への応用 (3) 教師なし学習によって得られた結果に対する信頼度の計算法の提案と理論的性質の解明 (4) 関数データの特徴を活かした部分空間クラスタリング法の提案，の4つの研究を行った．

研究成果の学術的意義や社会的意義
本研究では，実社会への応用を想定し，応用上重要な問題に対して，新しい教師なし学習法の開発や理論的性質の解明を行っている．例えば，研究(1)では教師なし分類法において金字塔と呼べる広く用いられているクラスタリング法に関して，これまで明らかとなっていなかった重要な理論的性質を解明している．さらに，本研究では，理論研究にとどまらず，実社会の問題への応用を実際に行っている．実際に，研究(2)ではスポーツ医学の分野において，提案手法を適用することで怪我のリスクのある選手の特定に成功している．

研究成果の概要（英文）：With recent advances in computer and measurement technologies, big and complicated data have been common in various application fields, and thus the importance of unsupervised learning has been recognized. In this research, I dealt with the following 4 research topics related to unsupervised learning for the complicated data: (1) I studied theoretical properties of graph-partitioning clustering method, (2) I developed a new semi-supervised learning method for functional data with theoretical guarantees and used the proposed algorithm to identify handball players who are at-risk for anterior cruciate ligament (ACL) injury based on ground reaction force data, (3) I developed a general approach via multiscale bootstrap to selective inference with theoretical guarantees, (4) I developed a new regularized subspace clustering algorithm for functional data which is based on a cluster-separation criterion in the finite-dimensional subspace.

研究分野：統計科学，機械学習

キーワード：関数データ解析　グラフ分割　クラスタリング　教師なし学習　半教師なし学習　機械学習

様　式　C－１９、F－１９－１、Z－１９、CK－１９（共通）

１．研究開始当初の背景

近年，インターネットの普及や計測技術の発展によって従来の多変量（対象 x 変数）データでは表現しきれないデータが多く得られている．例えば，SNS における登録者間の関係はネットワーク（グラフ）によって表現される．また，計量化学分野における近赤外線分光法に関連するデータ（図 1），細胞状態を表すラマンスペクトルデータ，運動に関連する軌道データなどは連続的・断続的に記録されるデータである．このような複雑なデータに対しては，既存のデータ解析手法を適用することは困難であるが，データの構造を生かし上手く情報を引き出すことができれば多くの知見を得ることができる．さらに，データの大規模化に伴い，得られたデータから有用な情報を得るための教師なし学習をはじめとする探索的なデータ解析が重要となっている．そのため，教師なし学習を中心とした解析法の開発，解析法の理論的性質の解明，実社会への応用は重要な課題である．

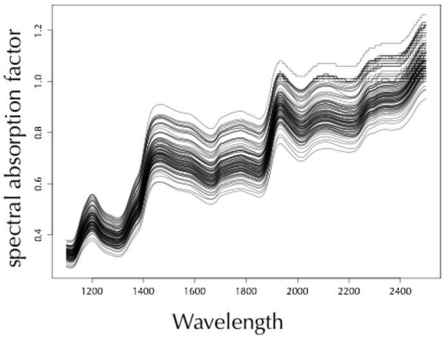

図 1 小麦サンプルに対する近赤外線分光法データ

（a）ネットワークデータに対しては，データの背後に様々な確率モデルを考えることができる．重要なモデルの１つとして，得られたグラフの背後に幾何的な構造を仮定する場合がある(random geometric graph の仮定)．Random geometric graph はネットワークデータのみならず多変量データの局所的な構造を捉える際にも重要な役割を担う．一方で，random geometric graph を仮定した状況でのクラスタリング法等の解析法の理論的性質の解明は発展途上である．（b）図１に示したような連続的に変化するデータに対しては，背後のデータ発生機構として，実数空間上の確率分布を考えるよりも，ある（有界な）領域や区間上のランダムな関数（もしくは，確率過程）を考える方が自然である．ある領域や区間上で連続的・断続的に観測されたデータをランダムな関数や確率過程の実現値として捉えたデータ解析は関数データ解析と呼ばれ，統計科学分野において盛んに研究が進められている．一方で，教師なし分類問題に対する研究はまだ十分に進んでいるとはいえない．また，応用分野においての認知度はまだ高いとは言えず幅広い応用可能性を秘めている．

２．研究の目的

上述の研究背景から，ネットワークデータと関数データを主に扱い，クラスタリング法などを始めとする教師なし学習に関連する解析法の開発，一致性などの理論的性質の解明，実社会への応用を目指す．具体的には，(1) グラフクラスタリングの理論的性質の解明，(2) 関数データに対する半教師付き判別問題 (3) 教師なし学習によって得られた結果に対する信頼度の計算 (4) 関数データの特徴を活かした教師なし分類問題，の 4 つの研究に取り組む．

(1) 近年，ネットワークデータに対しても多変量データのクラスタリングにおいてもグラフ（ネットワーク）分割に基づく spectral clustering (SC) がクラスタリング法の主流となっている．SC は normalized cut (Ncut)とよばれるグラフ分割法をベースとしている．しかし，normalized cut は NP 困難な問題であるため，SC は Ncut を緩和した方法となっている．一方で，Dhillon et al. (2007) によって Ncut の局所解を効率的に得るアルゴリズムが提案された．SC に対する理論的性質の解明は十分に進んでいるのに対して，Ncut に対する理論的性質の解明は十分とは言えない．そこで，random geometric graph に対する Ncut の理論的性質の解明を試みる．

(2) 実データ解析においては，必ずしも対象のグループラベルが完全に得られるとは限らない．例えば，スポーツ医学の分野において，動きのダイナミクスから選手生命に関わる怪我のリスクのある選手を特定することが重要である．しかし，リスクのある選手の中で実際に怪我が起こるのは一部で，すべてのリスクのある選手が怪我をするわけではない．したがって，学習用データとして,positive なデータとラベルをつけることのできないデータ(unlabeled data）のみが得られている．このような状況における判別問題は PU classification (classification from only positive and unlabeled data)と呼ばれ，機械学習分野を中心に盛んに行なわれている．そこで，関数データに対する PU 判別問題に関する解析手法の開発と詳細な理論解析を行う．さらに，実データ解析への応用を目指す．

(3) 仮説をデータから選択した影響を考慮していない従来の統計的推測の妥当性は保証されない．近年，データに基づく仮説の選択の影響を適切に扱った統計的推測は，selective inference と呼ばれ，統計科学分野で注目を集めている (Taylor and Tibshirani, 2015)．階層的クラスタリング法における各クラスタに関する検定 (pvclust; Suzuki and Shimodaira, 2006) では，予め仮説を用意するのではなく，データから得られたクラスタに対して検定を行うことが多い．このような場合は，近年注目されている選択的推測(selective inference)を行う必要があるが，仮説領域や選択領域が複雑であり既存の方法は適用することができなかった．そこで，一般的な状況でも適用可能な選択的推測の方法を提案し，その理論的性質を明らかにする．

(4) 教師あり判別問題に対しては，Delaigle and Hall (2012)によって関数データの実数空間への射影によって良い判別性能が得られることが明らかになっている．そして，教師なし分類問題に対しても部分空間を用いた方法が多数提案されている．しかし，既存の部分空間クラスタリングの方法は関数データの特徴を活かしきれておらず十分な性能を得ることができていなかった．そこで，関数データの高次元性を上手く活かした方法の提案を試みる．

３．研究の方法

(1) サンプルサイズを n とすると，Dhillon et al．(2007)では，Ncut が n 次元空間上の重み付き k-means 法と同等であることを示している．しかし，この空間はデータに依存する空間であり理論的性質を考えることが困難である．そこで，SC の理論的性質を明らかにした von Luxburg et al．(2008)と同様の設定の下で，サンプルサイズが無限大に発散した際にどの様な空間における重み付き k-means 法に収束するのかを明らかにする．また，Levrard (2015)によって Hilbert 空間上の k-means 法に関する詳細な理論があたえられていることから，この結果を重み付き k-means 法に対するものへと拡張することで，Ncut の詳細な理論を与える．

(2) 多変量データに対する PU learning においては，unlabeled data における positive label のデータの割合（混合率）の推定が重要な役割を担う．しかし，関数データに対してはこの量を推定することは困難である．一方で，関数データは Karhunen-Loéve （KL） 展開によって無限次元の確率変数から構成されていることがわかる．そのため，この高次元性を利用することで，混合率の推定を必要としない方法の開発と理論的性質の解明を目指す．そして，床反力データから膝に関わる特定の怪我（ACL 損傷）のリスクがあるスポーツ選手を特定する問題に提案手法を適用する．

(3) 仮説領域がある空間上の一般の領域で表現される場合の統計的推測問題を Efron et al．(1998)では領域の問題と呼んでいる．例えば，階層的クラスタリング法によって得られたクラスタに信頼度を与える問題などはこの問題の一例である．最も単純な信頼度の計算方法としてブートストラップ確率の計算が挙げられるが，頻度論的な信頼度（p 値）としては精度が低い．より精度の高い頻度論的な信頼度を計算するためには領域の幾何的な情報が必要となる．一方で，多くの応用において，仮説領域の陽な表現を得ることは困難である．本研究では，領域の問題を selective inference に対するものへと拡張し，マルチスケール・ブートスットラップ法 (Shimodaira，2002，2004)を用いることで，一般的な状況で近似的に不偏な selective inference を行う方法を提案する．

(4) 既存の関数データに対するクラスタリング法では，教師あり判別で完全に分離可能なデータに対して適用しても，分類構造を得ることはできなかった．そこで，関数データの高次性に着目することで関数データの背後に隠れたクラスタ構造を反映した部分空間を求める新しい部分空間クラスタリング法を提案し，その理論的性質を明らかにする．

４．研究成果

(1) 本研究では，Ncut が漸近的に母集団分布に対するある再生核 Hilbert 空間 (RKHS)上の重み付き k-means 法の解に収束することを示した．また，真の degree 関数が既知の場合に，推定量の母集団分布に対する損失が $O(1/n)$の早いレートで母集団分布に対する損失の最適値に収束することを示した．さらに，RKHS 上の母集団分布に対する重み付き k-means 法は，データ空間における母集団分布に対する Ncut と同等であることを示した． これにより，Ncut によるクラスタリング結果は母集団分布に対する Ncut（真の損失関数）の意味で最適な分割に収束することが明らかになった．Ncut の relaxation である normalized spectral clustering はこのような最適性を持たず一般に異なる解に収束することを示した．本研究成果は，機械学習の top conference である ICML2019 に採択された．

(2) 本研究では，Delaigle and Hall (2012)と同様に KL 展開に基づく関数データの潜在的な無限次元性に着目することで，PU classification に適した関数データの射影に基づく距離関数を提案し，この距離関数の推定値に基づくシンプルな関数データに対する PU classification の方法を提案している．提案手法の大きな利点として，混合率の推定を必要としないことが挙げられる．さらに，通常の判別問題とほぼ同じ正則条件の下 (Delaigle and Hall，2012)，誤判別率が漸近的に 0 となるような完璧な分類が達成可能であることを示した．実際に，図 2 では Kalivas (1995)の NIR データに対して一部のクラスラベルを隠すことで PU learning の疑似データを作成し，提案手法を適用した結果であり，PU learning の枠組みにおいて提案手法によりよい分類ができていることが確認できる．さらに，大阪大学医学系研究科小笠原一生先生との共同研究を行い，床反力データから膝に関わる特定の怪我(ACL 損傷)の危険性があるスポーツ選手を特定する問題に提案手法を適用することで，

実際に怪我の危険性のある選手の特定に成功した．この内容は，ＮＨＫニュース「おはよう日本」の番組内で紹介された．

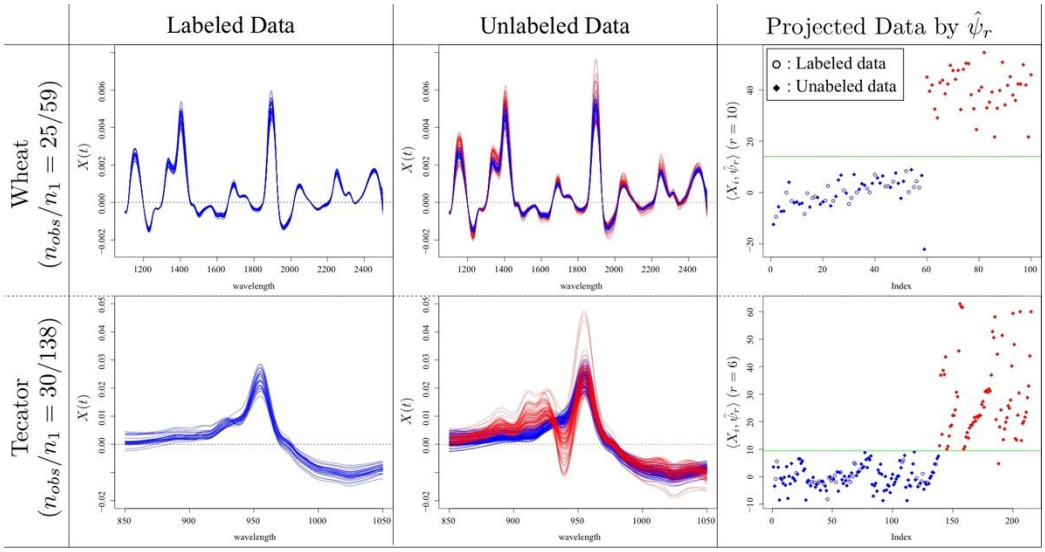

図 2 NIR データ(Kalivas，1995)に対する提案手法の適用結果

(3) 本研究では，一般的な状況で，マルチスケール・ブートストラップ法を用いて，通常のブートストラップ法と同じ計算量で，近似的に不偏な selective inference を行う方法を提案した．また，仮説領域の境界が滑らかな場合を想定した通常の漸近理論（大標本理論）と仮説領域が滑らかでない状況を想定した nearly flat surface の漸近理論の両理論において提案手法の精度に対して理論的保証を与えた．更に，計算コストの高いダブル・ブートストラップ法による方法が，同等の精度をもつことを示した．これにより，マルチスケール・ブートストラップ法による提案手法は，少ない計算コストで精度の良い selective inference が実行できることが分かる．

(4) 本研究では，データをクラスタ間の分離度が最も大きくなるような部分空間へ射影しクラスタリングを行う方法を提案した．さらに，本研究では，適当な正則条件の下で，データの背後に平均関数の差の意味でクラスタ構造があるならば，提案手法によって推定したクラスタ構造の誤判別率が 0 に（確率）収束するということを示している．

５．主な発表論文等

〔雑誌論文〕（計 4 件）

① Terada, Y. and Yamamoto, M. (2019). Kernel Normalized Cut: a Theoretical Revisit, *In Proceedings of International Conference on Machine Learning (ICML2019), eds., K. Chaudhuri and R. Salakhutdinov*, **PMLR 97**, pp. 6206-6214.（査読あり）
http://proceedings.mlr.press/v97/terada19a.html

② Shimodaira, H. and Terada, Y. (2019). Selective inference for testing trees and edges in phylogenetics, *Frontiers in Ecology and Evolution*, **7:174**, 1-15. （査読あり）
DOI: 10.3389/fevo.2019.00174

③ Terada, Y. and Shimodaira, H. (2017). Selective inference for the problem of regions via multiscale bootstrap, *arXiv:1711.00949*. （査読なし）
https://arxiv.org/abs/1711.00949

④ Hirose, K. and Terada, Y. (2018). Simple structure estimation via prenet penalization, *arXiv:1607.01145*. （査読なし）
https://arxiv.org/abs/1607.01145

〔学会発表〕（計 20 件）

① Terada, Y. and Shimodaira, H. (2018). Selective inference for the problem of regions via multiscale bootstrap, The 27th South Taiwan Statistics Conference and CIPS Annual Meeting 2018, National Cheng Kung University, Tainan, Taiwan.

② Terada, Y. and Yamamoto, M. (2018). Subspace clustering for functional data, The 2nd International Conference on Econometrics and Statistics (EcoSta2018), the City

University of Hong Kong, Hong Kong.

③ Terada, Y. (2017). Semi-supervised classification for functional data and its applications, The 10th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2017), University of London, London, United Kingdom.

６．研究組織

(1) 研究代表者

寺田 吉壱（TERADA, Yoshikazu）
大阪大学・基礎工学研究科・助教
研究者番号：10738793