様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２８　年　　６　月　２１　日現在

機関番号：　８２６３６
研究種目：　研究活動スタート支援
研究期間：　2014～2015
課題番号：　２６８８００３１

研究課題名（和文）クラスタリング法を中心とした教師なし学習の統計理論の構築

研究課題名（英文）Statistical theory of unsupervised learning with a focus on clustering methods

研究代表者
寺田　吉壱（Terada, Yoshikazu）

国立研究開発法人情報通信研究機構・脳情報通信融合研究センター　脳情報通信融合研究室・研究員

研究者番号：１０７３８７９３

交付決定額（研究期間全体）：（直接経費）　　1,600,000円

研究成果の概要（和文）：本研究では，関数データに対する教師なし又は半教師付き分類問題に対して，関数データの高次元性に着目した新しい方法の開発と理論的性質の解明を行った．また，高次元小標本データに適したクラスタリング法として提案したdistance vector clusteringに関して理論的性質の解明を行った．これらに加えて，シンプルで強い仮定を必要としないfMRIデータに対する脳活動領域の特定法を提案した．

研究成果の概要（英文）：In this research, I studied unsupervised classification and binary classification from only positive and unlabeled functional data (PU classification for functional data). Some important properties of the functional data clustering method proposed by Chiou and Li (2007) were derived, and a simple classification algorithm for functional PU learning problem was developed. Moreover, I proved that the distance vector clustering works well under several important high-dimension low-sample size settings. In addition, the simple voxelwise statistical inference for the underlying hemodynamic response function based on the difference-based estimator was developed. Under mild regularity conditions, it was shown that the proposed test statistics based on the difference-based HRF estimator follow chi-squared distributions under null hypotheses for several important hypotheses.

研究分野：　教師なし学習の統計理論，fMRIデータ解析

キーワード：　関数データ解析　高次元データ解析　fMRIデータ解析

１．研究開始当初の背景

クラスタリング法は，教師無し学習の代表的な方法として，様々な分野で用いられている．しかし，統計理論構築には高度な数学的知識が必要となるため，多くの方法について統計的性質は明らかにされていない．これは実データ解析において非常に大きな問題となる．データの増加に伴い何らかの結果に収束しない方法では，同一の分布から生成した２つのデータに対してその方法を適用しても全く異なる結果が得られてしまい分析結果の解釈は本来行うことはできない．しかし，統計的性質が解明されていない方法が多いことから，クラスタリング法の統計理論の構築が完成したとは言えない．そこで，実データ解析において有用であると考えられる方法を中心に，まず大標本理論において一致性を証明する必要がある．さらに，近年コンピュータの普及等に伴い遺伝子データ，fMRI データ等の超高次元データが得られ，このようなデータにクラスタリング法を適用する場合が多い．そのため，高次元データなどの枠組みに適したクラスタリング法の開発を行う必要がある．

２．研究の目的

研究 (1) 大標本理論の枠組みにおけるクラスタリング法の開発と理論研究

多くの場合，データはある分布からランダム発生していると仮定する．この場合，データに対して最適なクラスタリングを得ることではなく，背後にある未知の分布に対して最適な結果を得ることが分析の目的となる．独立同一分布サンプリングの下では，サンプルサイズが大きくなると，データのもつ分布の情報が増えるため，クラスタリング結果も背後の分布に対して最適な結果が得られる事が期待される．大標本理論の枠組みでのクラスタリング法の研究では，サンプルサイズが無限大に発散すれば，背後の分布に対して最適な結果に収束するという性質が重要となる．そのため，研究(1)の目的は，**主要なクラスタリング法の一致性もしくは不一致性を明らかにすることである**．

研究 (2) 高次元データの枠組みにおけるクラスタリング法の開発と理論研究

研究（1）では，変量数を固定し，サンプルサイズの増加にともないクラスタリングがどのような性質をもつかという部分に焦点を当てていた．しかし，近年，遺伝子データや fMRI データのようにサンプルサイズよりも次元数（変量数）のほうが大きいデータが多く得られるようになった．このようなデータに対して，大標本理論の枠組みは適切ではない．研究（2）では，サンプルサイズ n を固定し次元数 p が大きくなる枠組みの高次元小標本理論においてクラスタリング法の理論的性質を明らかにする．

研究 (3) fMRI データ等の実データに即した方法が必要と感じられる場合の研究

さらに，fMRI データ等の実データに即した方法が必要と感じられる場合は，適宜データに適した方法の開発や理論研究を行う．ニーズに合った解析法の開発や理論研究を行うことで，脳科学領域で重要な発見を促すとともに研究(1, 2)をより実用的な研究に昇華させ統計関連領域で break through となる理論の構築を目指す．

３．研究の方法

研究（1）：本研究では，当初 L1 正則化を用いて階層的なクラスタを構成するクラスタリング法に対する理論研究を進める予定であった．しかし，この方法は 1 次元のデータに対しては良い性質をもつが，各次元ごとのクラスタ構造が異なる場合には上手く機能しないことを理論的に明らかにした．具体的には，この方法は各次元ごとに別々にクラスタリング法を適用していると解釈できるため，ほとんどの場合クラスタが上手く構成されない．そのため，本研究では近年注目されている関数データに対するクラスタリング問題に取り組んだ．関数データは，Karhunen-Loève 展開により，本質的には無限次元の確率変数から構成されていると考えることができる．したがって，その背後に非常に多くの情報を保持していると考えることができる．この性質に基づいたクラスタリング法や半教師付き分類法の開発や理論的性質の解明を試みた．

研究（2）：本研究では，自らが提案した高次元小標本データに適したクラスタリング法である distance vector clustering についてより詳細にその理論的性質の解明を試みた．ここで，Distance vector clustering は，高次元空間においては距離の“近さ”ではなく“値”に意味 (クラスタ情報) があることに着目した方法であり，距離行列又は内積行列をデータ行列とみなし，その行列に対して従来法を適用する方法である．

研究（3）：研究を進めていく中で，fMRI データ解析に対する理論的性質の解明が進んでいないという問題点にたどり着いた．そのため，fMRI データ解析の中でも最も重要な血流動態反応関数の推定と脳活動領域の特定に関して新しい方法の開発と理論的性質の解明を試みた．

４．研究成果

研究（1）：関数データは，Karhunen-Loève 展開により，本質的には無限次元の確率変数から構成されている．そのため，潜在的

な関数データの無限次元性を引き出すことができれば，高次元データ解析と同様に，perfect な分類の達成が期待される．実際に，Delaigle and Hall (2012) では，通常の教師あり判別問題において完璧な分類が達成可能な方法を提案している．そこで，高次元データと関数データの高次元性 の相違点を通して，L2 距離に基づく関数データの分類が何故上手く機能しないかを明確にし，Chiou and Li (2007)で提案されている関数データのクラスタリング法によって高次元小標本データに対するクラスタリング法のように漸近的に perfect なクラスタリングが達成可能な条件を明らかにした．また，2 値判別問題において，一方のクラスの一部の対象にしかラベルしか観測されていない状況における半教師付き判別問題 (PU learning) に対して，関数データの射影とクラスタリングを組み合わせた方法を提案した．提案手法は，従来の多変量データに対する PU Learning の方法と異なり，クラスの混合率の推定を必要としない．関数データが連続的に観測されている場合に，適当な正則条件の下で，提案手法により完璧な分類が漸近的に達成可能であることを証明した．

図１では，一見してもクラスの分離が困難である 2 つの数値実験データと提案手法の適用例を示している．図の 3 行目の提案手法の適用結果から，一方のクラスの一部にしかラベルが観測されていない状況でも提案手法が上手く Label が観測されていないデータの分離を行えていることがわかる．

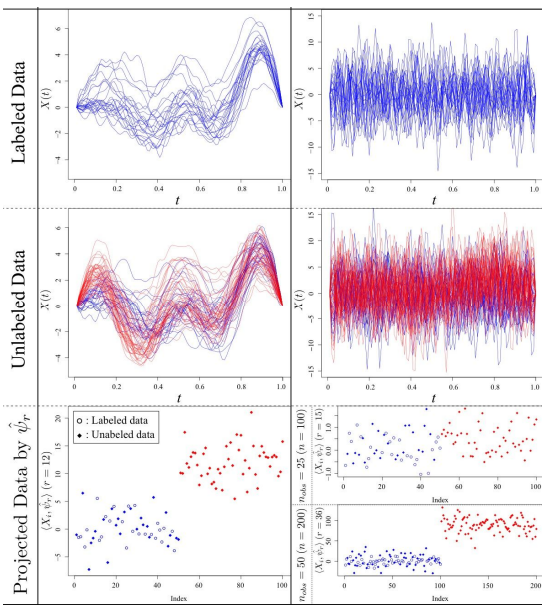

図１：２つの数値実験データと提案手法の適用例；Label が観測されているデータ (Labeled Data) と Label が観測されていないデータ（Unlabeled Data）と提案手法により射影した関数データの点はそれぞれ真のクラスタ構造で色付けされている．

研究（2）：これまでの distance vector clustering の理論的性質は Hall et al. (2005) の高次元小標本の枠組みに基づいていた．そこで，Jung and Marron (2009)や Qiao et al. (2010)などの Hall et al. (2005)とは異なる条件の下であっても提案手法が上手く機能することを理論的に示した．また，これらの成果をまとめて投稿中である．

研究（3）：fMRI データ解析の中でも最も基本的な脳活動領域の特定に必要となる血流動態反応関数 (HRF) の推定と検定に関する理論的研究を行った．具体的には，シンプルで計算が容易な 1 階差分推定量が理論的に良い性質（一致性と漸近正規性）をもつことを弱い仮定の下で示し，脳活動領域を特定するための新たな検定統計量を構成した．提案手法は, SPM 等の解析ソフトのように正規性やドリフト項の除去可能性などの強い仮定を必要としないため，実際の fMRI データに対しても SPM 等の結果よりも良い（妥当な）結果を得られている．

図 2 は，顔画像をランダムなタイミングで提示している際に計測した私の脳の fMRI データから顔画像に反応して活動した領域を特定した brain activity map である．右が既存手法の結果であり，左が提案手法の結果である．人の脳には，Fusiform Face Area (FFA) と呼ばれる顔を認識する脳領域があることが知られている．図 2 では，FFA の領域を大雑把に丸で示しており，提案手法では既存手法よりも上手く FFA の活動を捉えられていることがわかる．

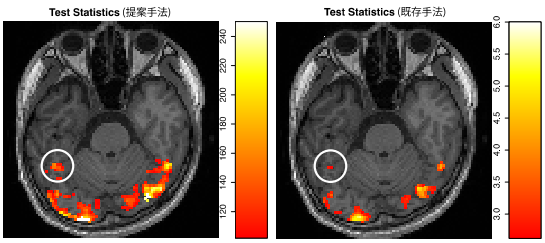

図 2 顔画像を提示した際の脳活動（左：提案手法，右：既存手法）

またこれらの研究に加えて，Groenen 教授とともに Symbolic data に対する多次元尺度構成法（MDS）に関してまとめた chapter を作成している．これに際して，symbolic MDS に関する R package を作成し公開している．

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 1 件）

① Patrik J.F.Groenen and Yoshikazu Terada, Symbolic Multidimensional Scaling, Econometric Institute Research Papers, 査読無, EI 2015-15, 2015,pp.1-22.

〔学会発表〕（計 8 件）

① Yoshikazu Terada 、 Brain activity detection based on the difference-based HRF estimator、The Third CiNet Conference: Neural mechanisms of decision making: Achievements and new directions、2016 年 02 月 05 日、CiNet（大阪府、吹田市）

② 寺田　吉壱、fMRI データに対する血流動態反応関数のセミパラメトリック推定とその応用、データ科学シンポジウム 2015（科研費）「欠測データ解析とモデル選択：生体情報データの統計モデル」、2016 年 01 月 22 日、大阪大学（大阪府、豊中市）

③ 寺田　吉壱、fMRI データに対するシンプルで強い仮定を必要としない脳活動領域の特定法、第 18 回情報論的学習理論ワークショップ（IBIS2015)、2015 年 11 月 26 日、つくば国際会議場（茨城県、つくば市）

④ Yoshikazu Terada, On the difference-based estimator of Hemodynamic Response Function (HRF)、IASC Satellite Conference 2015: Statistical Computing for Data Science、2015 年 08 月 03 日、Atlantico Buzios Convention & Resort, Buzios RJ(Brazil)

⑤ 寺田　吉壱、Asymptotic Properties of Difference-Based Estimation of Hemodynamic Response Function、2015 年度統計関連学会連合大会、2015 年 09 月 08 日、岡山大学（岡山県、岡山市）

⑥ 寺田　吉壱、Ulrike von Luxburg、非重み付きグラフに対する graph embedding とその理論的性質、研究集会「大規模統計モデリングと計算統計」、2015 年 02 月 06 日、東京大学（東京都、目黒区駒場）

⑦ 寺田　吉壱、Ulrike von Luxburg、Unweighted graph に対する機械学習の限界と可能性～Random geometric graph の観点から～、第 17 回情報論的学習理論ワークショップ（IBIS2014)、2014 年 11 月 17 日、名古屋大学（愛知県、名古屋市）

⑧ 寺田　吉壱、Local Ordinal Embedding、統計数学セミナー、2014 年 11 月 11 日、大阪大学（大阪府、豊中市）

〔図書〕（計　0　件）

〔産業財産権〕
○出願状況（計　0　件）

○取得状況（計　0　件）

〔その他〕
ソフトウェア
R package smds: Symbolic Multidimensional Scaling,https://cran.r-project.org/web/packages/smds/index.html

６．研究組織
(1)研究代表者
寺田　吉壱（TERADA, Yoshikazu）
国立研究開発法人 情報通信研究機構・脳情報通信融合研究センター・脳情報通信融合研究室・研究員
研究者番号：10738793

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：