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　２７　年　　６　月　　１　日現在

機関番号：　１２６１２
研究種目：　若手研究(B)
研究期間：　2012～2014
課題番号：　２４７００２８０

研究課題名（和文）統計的機械学習理論に基づく非線形多変量解析手法の開発研究

研究課題名（英文）Nonlinear multivariate analysis based on statistical machine learning theory

研究代表者
川野　秀一（KAWANO, SHUICHI）

電気通信大学・その他の研究科・准教授

研究者番号：５０６１１４４８

交付決定額（研究期間全体）：（直接経費）　　2,700,000円

研究成果の概要（和文）：多様かつ高次元な観測データからの有効な情報抽出を目的として，非線形統計モデリング手法の理論・方法論の開発研究に取り組んだ．特に，半教師あり学習モデルを軸として研究を進め，関数データやデータの異分布性を考慮に入れたデータ解析手法を提唱するとともに，スパース推定に基づくモデリング手法の評価・予測に関する一連の統計的方法論を開発することができた．開発したデータ解析手法は，生命科学などの様々な実データに応用した．

研究成果の概要（英文）：We developed nonlinear statistical methods to extract useful information from high-dimensional diverse data. In particular, we proposed semi-supervised methods that can treat functional data or labeled data and unlabeled data from different sampling distributions, and developed a series of procedures for evaluating and predicting statistical models based on sparse estimation. We applied the proposed methods to datasets in the various fields of research including life science.

研究分野：　統計科学

キーワード：　半教師あり学習　関数データ解析　モデル評価基準　スパース推定　正則化法　異分布性

１．研究開始当初の背景

計測技術・計算機環境の高度な発展は，諸科学のあらゆる分野で大量かつ多種多様な高次元データ（実数，離散値，順序カテゴリー，グラフ構造など）の収集と蓄積とを可能にし，データベースとして整備されつつあった．大量のデータから有効な情報を抽出するためには，統計モデルの利用が必須であり，これまで数多くの研究者が様々な統計モデルを開発研究してきた．一方で，大量に蓄積されたデータに対してラベルを割りつけること（例えば，データをあるカテゴリーに分類することなど）は人的・時間的・金銭的に高コストであるため，ごく一部の少量のデータにしかラベルが付与されず（これをラベルありデータと呼ぶ)，残りの大量のデータにはラベルが付与されない（これをラベルなしデータと呼ぶ）といった状況が研究開始当初取得されていた多くのデータにおいて生じていた．

このようなデータ取得状況下においては，これまでのラベルありデータのみに基づく統計モデルからだけではデータの背後に潜む重要な因子や規則性を見逃す恐れがある．そこで，ラベルありデータとラベルなしデータの両種のデータを用いることによってモデルの構築を行う学習法が機械学習の分野で注目を集め，この学習法は半教師あり学習法と呼ばれている．半教師あり学習法は実際問題への応用の重要性から，様々な分野で，多様なアプローチによって研究されており，特に，バイオインフォマティクスやテキスト分類などの分野で応用されつつあった．しかしながら，研究開始当初において，多様かつ高次元なデータを念頭に置いた半教師あり学習モデルを非線形化と正則化の両者の観点から定式化し，さらに客観的なモデルの評価も含むモデリング手法に関する理論および方法論については，十分に研究は進んでいるとは言えなかった．

２．研究の目的

データの高次元性・多様性を十分に考慮に入れた非線形統計的モデリング手法，特に，半教師あり学習法に対する統計的モデリング手法の理論・方法論の確立を確固たる数理基盤の上で構成することを研究の目的とした．具体的には，以下の 3 点

(1)．多種多様なデータ形式（実数値，文字，関数, グラフ構造データなど）に対する統計的モデルを構築することが重要になること．

(2)．従来の半教師あり学習問題においては，ラベルありデータが出現するデータの密度分布は, ラベルなしデータが出現するデータの密度分布と同じ分布であると仮定されているが, 実際に取得されるデータでは, 各々得られるデータの密度分布の様相が異なる可能性も十分に考えられるため, このような状況を考慮に入れた新たな半教師ありモデルの構築が大切になること．

(3)．ラベル付けが行われる際何らかの誤りが生じ，誤ったラベルが割り付けられ，さらに, 説明変数に関するデータが欠落する恐れもあり, その処理や評価が重要となること．

に着目し，従来の半教師あり学習法の枠を超えて，情報融合を行うことが可能な頑健かつ柔軟な統計的モデリング手法のための統計科学的理論の構築，数理的推測論の整備，および多様・高次元データ解析のための方法論の研究を目的として研究を推進した．

３．研究の方法

基底展開法によるモデルの非線形化，正則化法によるロバストな推定方式に基づいた非線形多変量解析手法について研究を行った．モデルとしては半教師あり学習モデルを中心に考え，関数データ解析，共変量シフト下での統計理論を利用することにより，様々なデータ形式を取り扱うことが可能な統計モデルの開発研究を行った．また，正則化項内に含まれるモデルの複雑さの程度を調整するチューニングパラメータの選択を構築したモデルの選択問題として捉え，新たなモデル評価基準の導出を行った．

高次元データのモデリングでは，構築される統計モデルが大量のパラメータを有するため，パラメータの自動削減が可能なスパース推定法が重要となる．スパース推定法は，正則化項にスパース正則化項を用いることにより実行されるが，通常の正則化項と同様にその正則化項内にはチューニングパラメータが含まれている．このパラメータの選択に際して，情報量やベイズアプローチに基づいた新たな選択方法を研究した．開発したモデリング手法は，理論的・数値的に検証し，実問題への適用を通して有効性を検証した．

４．研究成果

(1)．得られるデータ集合が関数の場合における半教師あり学習法を開発することができた. 離散観測データをガウス型基底関数展開法により関数データ化し, 得られた関数データに基づき関数ロジスティック識別･判別手法を半教師あり学習の枠組みに拡張することによりモデルを定式化した．情報量・ベイズ理論の双方の観点から構築したモデルを評価するための基準を導出した. 人工データによる数値実験を通して, 代表的な半教師あり学習法や関数サポートベクター

マシーン等と比較・検証することにより，提案手法の有効性を示すことができた．また，生命科学分野のマイクロアレイデータ解析に提案手法を適用し，現象の解明に寄与することができた．

(2)．ラベルあり・なしデータの異分布性を考慮に入れた半教師あり学習法を開発することができた．具体的には，ロジスティックモデルを想定し，その異分布性を捉えるために共変量シフト下における密度比推定のアイデアを用いた．半教師あり学習法でよく用いられる半教師ありサポートベクターマシーンやトランスダクティブサポートベクターマシーンと数値的に比較し，ベンチマークデータによる性能評価により，提案手法の有効性を示すことができた．

(3)．スパース推定の中でも，lasso，ブリッジと呼ばれる推定方法に着目し，各々の推定方式に対してモデル評価基準を導出することができた．また，ブリッジ推定については，正則化項をパラメータ毎に変化させる適応型ブリッジ推定法を提案し，新たなモデリング手法を確立することができた．さらに，一般化線形モデル，他のスパース推定法への拡張研究を現在推進中である．

(4)．多種多様なデータ形式の一つとして，稀にしか観測されない極値データと呼ばれるものがある．この極値データを統計的に解析する研究分野は極値統計学と呼ばれている．極値統計学における確率分布（極値分布）の中でも一般化パレート分布に着目し，ベイズアプローチによってパラメータ推定の不適解問題を解消する方法論を提唱した．他の推定法との精度比較を人工データや実データの解析を通して行い，提案手法の有効性を確認した．この研究によって，極値統計学の応用適用範囲を広げることができた．

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　５　件）

① Kim, D., Kawano, S. and Ninomiya, S. (2014) Adaptive basis expansion via l1 trend filtering, Computational Statistics, 29, 1005-1023, DOI (10.1007/s00180-013-0477-7), 査読有.

② Kawano, S. (2014) Selection of tuning parameters in bridge regression models via Bayesian information criterion, Statistical Papers, 55, 1207-1223, DOI (10.1007/s00362-013-0561-7), 査読有.

③ Kawano, S. (2013) Semi-supervised logistic discrimination via labeled data and unlabeled data from different sampling distributions, Statistical Analysis and Data Mining, 6, 472-481, DOI (10.1002/sam.11204), 査読有.

④ Kawano, S. (2012) Adaptive bridge regression modeling and selection of the tuning parameters, Bulletin of Informatics and Cybernetics, 44, 29-39, URL(http://catalog.lib.kyushu-u.ac.jp/handle/2324/1495409/Kawano_Final.pdf), 査読有.

⑤ Kawano, S. and Konishi, S. (2012) Semi-supervised logistic discrimination for functional data, Bulletin of Informatics and Cybernetics, 44, 1-15, URL (http://catalog.lib.kyushu-u.ac.jp/handle/2324/1495407/KawanoKonishi_Final.pdf), 査読有.

〔学会発表〕（計　１２　件）

① 嶋村海人，川野秀一，小西貞則，モデル平均化法による Bayesian lasso 回帰モデリング，日本計算機統計学会第 28 回シンポジウム，2014 年 11 月，沖縄科学技術大学院大学（沖縄県・国頭郡）．

② 川野秀一，藤澤洋徳，高田豊行，城石俊彦，スパース正則化に基づく主成分回帰モデリング，2014 年度統計関連学会連合大会，2014 年 9 月，東京大学(東京都・文京区)．

③ 二宮嘉行，川野秀一，LASSO による変数選択のための AIC，2014 年度統計関連学会連合大会，2014 年 9 月，東京大学(東京都・文京区)．

④ Kawano, S. and Fujisawa, H., Sparse principal component regression for simultaneous dimension reduction and variable selection, The 3rd Institute of Mathematical Statistics Asia Pacific Rim Meeting, 2014 年 6 月，台北（台湾）．

⑤ Ninomiya, Y. and Kawano, S., AIC-type information criterion for LASSO, The 3rd Institute of Mathematical Statistics Asia Pacific Rim Meeting, 2014 年 6 月，台北（台湾）．

⑥ 松井秀俊，川野秀一，ロジスティックモデルに対するスパース正則化，日本計算機統計学会第 28 回大会，2014 年 5 月，中央大学（東京都・文京区）．

⑦ 川野秀一，Bridge 回帰モデリングにお

けるモデル選択問題，日本計算機統計学会第 27 回シンポジウム，2013 年 11 月，崇城大学ホール（熊本県・熊本市）．

⑧ 堀畑雄一，川野秀一，極値分布に対するベイズ型モデル評価，2013 年度統計関連学会連合大会，2013 年 9 月，大阪大学（大阪府・豊中市）．

⑨ Kawano, S., Tuning parameter selection in bridge regression modeling, The 2013 Joint Statistical Meetings, 2013 年 8 月, Montréal (Canada).

⑩ 堀畑雄一，川野秀一，一般化パレート分布におけるベイズ推定法とその評価，2013 年度応用統計学会年会，2013 年 5 月，パルセいいざか（福島県・福島市）．

⑪ 川野秀一，Bridge 回帰モデリングとその評価，2012 年度統計関連学会連合大会，2012 年 9 月，北海道大学（北海道・札幌市）．

⑫ Kawano, S., Weighted logistic regression modeling for semi-supervised classification, The 2012 Joint Statistical Meetings, 2012 年 7 月, San Diego (USA).

〔図書〕（計　0　件）

〔産業財産権〕
○出願状況（計　0　件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　0　件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
取得年月日：
国内外の別：

〔その他〕
ホームページ等
http://kjk.office.uec.ac.jp/Profiles/68/0006701/profile.html

６．研究組織
(1)研究代表者

川野　秀一（KAWANO, SHUICHI）
電気通信大学・その他の研究科・准教授
研究者番号：５０６１１４４８

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：