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

平成　２９　年　　６　月　　６　日現在

機関番号：　１２６０５
研究種目：　挑戦的萌芽研究
研究期間：　2014～2016
課題番号：　２６５６０１６５

研究課題名（和文）数理計画法による特徴選択の厳密化

研究課題名（英文）Exact methods for variable selection via mathematical programming

研究代表者
宮代　隆平（Miyashiro, Ryuhei）

東京農工大学・工学（系）研究科（研究院）・准教授

研究者番号：５０３７６８６０

交付決定額（研究期間全体）：（直接経費）　　2,900,000円

研究成果の概要（和文）：インターネット等から得られる大量のデータを分析し，そこから背後に含まれる法則を推定する方法の一つに回帰分析がある．本研究は，回帰分析の中でも，データにおけるどの要素が重要なのかを求める「特徴選択（あるいは変数選択）」と呼ばれる研究課題に対して，より高速に精度の高い推定を行えるようにするための数学的モデルの構築を行うことを目的とした．研究成果として，様々な種類の特徴選択問題に対し，計算の高速化や高精度化を達成することに成功した．

研究成果の概要（英文）：Regression analysis is a method to extract a hidden model from a large number of observations (samples). In this research, we concentrated on constructing algorithms for feature selection (variable selection) problems. Our algorithms are based on mathematical programming, which aims getting true optimal solutions. We have developed several integer-programming based algorithms, and have shown that the proposed algorithms produced better solutions than ones given by heuristics algorithms of previous researches.

研究分野：　数理計画法

キーワード：　特徴選択　数理工学　アルゴリズム　統計　ＯＲ

１．研究開始当初の背景

機械学習における特徴選択，統計学における変数選択は，多数の観測データから，現象の回帰モデルを推定する研究課題である．より良い回帰モデルの選択は，観測する現象のより正確な予測を可能とするため，特徴選択は理論・応用の両面において重要なテーマである．しかし研究の開始当初では，特徴選択にはもっぱらステップワイズ法などのヒューリスティクスが用いられており，得られるモデルに厳密な意味での最適性は無い状況であった．それに対し本研究課題代表者らの研究により，従来のヒューリスティクスによって得られたモデルは，予測精度がかなり悪い場合があることが判明していた．

また，特徴選択において答えとなる回帰モデルに含まれるパラメータ数が既知の場合は，数理計画法を利用して厳密に最適な回帰モデルが求まることはすでに知られており，その前提での研究が最近も存在する．しかしながら，現実的には最適な回帰モデルに含まれるパラメータ数が既知であるという状況は稀であり，それが未知の場合は数理計画法では扱えない状況であった．そのため，特徴選択はもっぱらステップワイズ法に代表されるヒューリスティクスで行われているが，それらによる特徴選択は厳密性（最適性）の保証が無く，その欠点は近年の論文でも指摘されていた．

２．研究の目的

上記のような背景に応じて，本研究では大規模データに対する厳密に最適な回帰モデルを求めるアルゴリズムの開発を行う．単なるデータの大規模化への対応ではなく，回帰モデルの厳密化については，統計学・機械学習において提案されている従来の手法では考慮することが難しい．そこで本研究は数理計画法を利用したアプローチにより，得られる回帰モデルの精度向上という観点から特徴選択という課題にアタックし，この研究分野への貢献を目指す．

ステップワイズ法に代表される既存の特徴選択手法は，あくまでもヒューリスティクスであり，選ばれた回帰モデルが情報量規準の意味で最適である保証は全く無い．また近年は Lasso に端を発する，データの大規模化を考慮した L1 型正則化法や L0 型正則化法という手法が，機械学習の分野で流行している．しかしながら，Lasso（およびそれらの後継アルゴリズム）は変数選択の際の一致性が無いことが指摘されており，出力として得られる回帰モデルの厳密性については注力されていない状況である．

これに対して本研究の代表者らは，数理計画法を利用した厳密に最適な回帰モデルを求める特徴選択アルゴリズムを開発した．ただし現状では，このアルゴリズムは小規模なデータしか扱えない．しかし小規模データにおいても，従来のステップワイズ法よりかなり良い回帰モデルが得られるケースが存在した．本研究では，計算時間や数値的不安定性など，このアルゴリズムの問題点を克服し，大規模データに対しても厳密な最適性を保証可能な特徴選択アルゴリズムを開発する．

また，UCI Machine Learning Repository などのベンチマーク問題集は，世界中の研究者が研究の実験対象として用いているが，ごく一部の問題を除くと最適な回帰モデルは知られておらず，いわば比較対象としての最終到達地点が不明な状況にある．本研究では，それらのベンチマーク問題に対しても厳密に最適な特徴選択を行い，この分野の標準問題に対するベンチマーク解の提供も目標とする．

３．研究の方法

本研究の代表者らが開発した，小規模データに対して有効な数理計画を用いた特徴選択アルゴリズムは，研究開始前の段階では扱えるデータサイズに限界がある．研究目的である大規模データに対する特徴選択の厳密化を達成するため，平成 26 年度は主に「中規模データに対する計算機実験，特に数値不安定性の解析」「従来の特徴選択アルゴリズムの調査および実装」「小規模データに対する厳密な回帰モデルの求解」「Special Ordered Set type 2 による高速化」を行う．それ以降については，平成 26 年度の研究進捗状況をふまえつつ，平成 27 年度に「数値不安定性の克服」「中規模データに対する厳密な回帰モデルの求解」「数理計画法における探索ルールの構築」，平成 28 年度に「大規模データに対する厳密な回帰モデルの求解」「超大規模データに対する回帰モデルの高精度化」の順序で研究を遂行し，本研究課題の目標達成を試みる．

(1) 中規模データに対する計算機実験，特に数値不安定性の解析

予備的な計算機実験により，開発した特徴選択アルゴリズムは小規模データに対しては有効だが，中規模以上のデータでは数値的不安定性により計算の続行が困難になる場合が多いことが判明している．この問題点を解決するために，さらに多くの中規模データに対し計算機実験を行い，数値的不安定性をもたらす原因を究明する．特に，現状では数値的不安定性の主要因は，分枝限定法の内部で実行されている内点法の部分に含まれると想定しており，これを確認することを第一とする．

(2) 従来の特徴選択アルゴリズムの調査および実装

開発するアルゴリズムの有効性を示すため，比較対象として従来のヒューリスティクスによっても回帰モデルを求める必要がある．頻繁に用いられるステップワイズ法やL1正則化法なども，最新のアルゴリズムと以前のものでは性能が全く違うため，最先端のアルゴリズムを調査し実装を行う．また，それら従来法で最適な回帰モデルが得られないデータを収集し分析することにより，データ自体に含まれる計算困難な構造の発見を試みる．

(3) 小規模データに対する厳密な回帰モデルの求解

UCI Machine Learning Repository に掲載されているベンチマーク問題のうち，小規模なデータについて計算機実験を行い，厳密に最適な回帰モデルを求める．また従来法による回帰モデルとの比較・検証を行い，提案手法の有効性を確認する．実験によって得られた最適な回帰モデルは，ある程度の個数がそろった段階で国際会議または論文の形で発表を行い，研究成果の周知に努める．

(4) Special Ordered Set type 2 (SOS type2) による高速化

応募者らが開発した小規模データに有効なアルゴリズムでは，内部でSpecial Ordered Set (SOS) type 1 と呼ばれる数理計画法の技術を用いているが，これを SOS type 2 と呼ばれる，より複雑だが効率的なテクニックに実装し直すことにより，この研究段階におけるアルゴリズムの高速化を行う．

(5) 数値不安定性の克服

平成 26 年度に解析を行うアルゴリズムの数値的不安定性に対して，その克服を行う．現段階では，不安定性の原因として「内点法における初期実行可能内点の設定が洗練されていない」「分枝限定法における緩和問題の解き直しのタイミングが悪い」のどちらかを想定している．前者については，特徴選択問題専用の内点法のウォームスタートの開発を，後者については内点法ベースの分枝限定法中における，Taylor 展開を利用した双対単体法の導入を対策として考えている．

(6) 中規模データに対する厳密な回帰モデルの求解

UCI Machine Learning Repository に掲載されているベンチマーク問題のうち，中規模なデータについて計算機実験を行い，厳密に最適な回帰モデルを求める．従来法との比較および成果の発表に関しては，小規模データでの対応と同様に行う．

(7) 数理計画法における探索ルールの構築

特徴選択を離散最適化問題の観点から見ると，「(離散的な変数のみ考えると)実行不能解が存在しない」という大きな特徴がある．この特徴を生かすと，数理計画法の分枝限定アルゴリズムの実行途中において，探索ルールにかなり自由度を設けることができる．これを利用して，特徴選択問題に有効な分枝限定法の探索ルールを新しく構築し，計算時間の大幅な高速化を図る．

(8) 大規模データに対する厳密な回帰モデルの求解

UCI Machine Learning Repository に掲載されているベンチマーク問題のうち，大規模なデータについて計算機実験を行い，厳密に最適な回帰モデルの求解を試みる．特に，同ベンチマーク集に含まれるデータの中でも，被引用回数が非常に多いデータに対して計算資源を集中させる．

(9) 超大規模データに対する回帰モデルの高精度化

データが超大規模な場合には，厳密に最適な回帰モデルを求めることが難しい場合も想定される．しかし，数理計画法を利用したアルゴリズムは，厳密に最適な回帰モデルを求めるためだけでなく，計算時間を限定してアルゴリズムを実行することにより，ヒューリスティクスな手法として用いることが可能である．そこで超大規模データに対しては，単に計算時間を限定するだけでなく，数理計画法における分枝限定法の初期解探索ルールを変更することにより，より効率的なヒューリスティクスとして働くようにアルゴリズムの改善を行う．これにより，従前のヒューリスティクス法では発見できなかった，より高精度な回帰モデルの発見を試みる．

４．研究成果

本研究では，研究方法の欄にあげた解決すべき項目についてそれぞれ研究を行い，またそれらを相互に結びつけ，結果として以下の研究成果を得た．研究開始当初の背景に記入した通り，数理計画法による特徴選択問題の厳密化という研究の観点は先駆的であり，それらに対し得た成果については，「５．主な発表論文等」の欄に記述した通り，国際英文論文誌などで発表を行った．特徴選択については昨今の研究動向からますます重要視されているため，本研究課題の成果についても特徴選択に関する他研究者からの引用が見込まれる状況である．

(1) Mallows’Cp を規準とした特徴選択問題の混合整数二次計画法によるモデル化

本研究課題の開始前の段階では，特徴選択の最適化指標としてAIC, BIC，自由度調整済決定係数などを中心に考えていたが，古くから知られているMallows’Cpと呼ばれる指標についても，本研究で混合整数二次計画問題としての定式化を与えた．混合整数二次錐計

画問題ではなく混合整数二次計画問題としての定式化を与えたことにより，分枝限定法におけるホットスタートが可能となり，また分枝限定法の初期段階においても非常に質の良い解が得られることが実験でも確認できた．

(2) ロジスティック回帰における区分線形近似モデルの構築

ロジスティック回帰における特徴選択問題については，目的関数が対数関数を含むことから，従来の数理計画モデルで扱うことは困難であった．これに対し，本研究では区分線形近似を導入することにより，任意の精度でロジスティック回帰における対数損失関数を近似できるモデルの導出に成功した．また，中規模程度の問題に対して十分な精度の解が得られることが実験で確かめられた．

(3) 逐次ロジットモデルにおける区分線形近似モデルの提案

逐次ロジットモデルにおける特徴選択問題については，先行研究で二次関数による目的関数の近似が行われていた．しかしながら，この近似手法は計算は高速であるものの，ばらつきを多く含むようなデータに対しては誤差が大きくなってしまうことに注目し，本研究では区分線形近似を導入することにより，より高精度で近似を行う事の出来るモデルを構築した．また，中規模程度の問題に対して先行研究より十分に精度の高い解が得られることを実験で確かめた．

(4) 多重共線性を抑える特徴選択に対する条件数制約のモデル化

特徴選択問題では，AIC や BIC などの情報量規準を最小化することを目指すが，扱うインスタンスによっては得られた答えが最適であっても，多重共線性を含んでしまうことがある．多重共線性を含むような解は，予測精度が低下することが知られており，これを排除することが望ましい．多重共線性の大きさは「対応する相関係数行列の条件数の大小」，あるいは「分散拡大要因の大小」のどちらかで計測されることが多いが，このどちらも（いわゆる線形の）数理計画法では定式化が難しいものであるため，これまでの数理計画法による特徴選択の研究では扱うことができていなかった．

これに対して本研究では，まず条件数で計測される多重共線性に対して，条件数を一定値以下に抑えるという制約が半正定値制約として定式化できることを発見し，それにより条件数制約付の特徴選択問題を混合整数半正定値問題として定式化することに成功した．

(5) 多重共線性を抑える特徴選択に対する分散拡大要因に関する制約のモデル化

上記(4)の理由から，多重共線性を抑えることのできる特徴選択の手法が求められていた．これに対して本研究では，多重共線性の指標として最もよく用いられる分散拡大要因について，それを一定値以下に抑える特徴選択問題が混合整数二次計画問題として定式化できることを証明した．また実験により，従来のステップワイズ法より質の良い解が得られることを確認した．

(6) 特徴選択のための，混合整数二次錐計画モデルの安定化および高速化

本研究課題開始前に得られていた特徴選択のための数理計画モデルは混合整数二次錐計画問題であり，これは計算が数値的に不安定であること，また分枝限定法におけるホットスタートが効かないことから，大規模な問題を解けないという弱点を有していた．これに対し，本研究では数理計画モデルを改良することにより，先行モデルの約 10 倍～100 倍の高速化に成功した．これは数理モデルに含まれる行列部分を等価変形し縮小したことと，特徴選択の暫定解による分枝限定法の効率的な枝刈りを実装することにより達成した．

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８件）

(1) R. Tamura, K. Kobayashi, Y. Takano, R. Miyashiro, K. Nakata, T. Matsui, Best subset selection for eliminating multicollinearity , Journal of the Operations Research Society of Japan, No. 60, 2017, 印刷中.（査読有）

(2) T. Sato, Y. Takano, R. Miyashiro, Piecewise-linear approximation for feature subset selection in a sequential logit model, Journal of the Operations Research Society of Japan, No. 60, 2017, pp. 1－14.（査読有）
http://www.orsj.or.jp/~archive/pdf/e_mag/Vol.60_01_001.pdf

(3) T. Sato, Y. Takano, R. Miyashiro, A. Yoshise, Feature subset selection for logistic regression via mixed integer optimization, Computational Optimization and Applications, No. 64, 2016, pp. 865－880.（査読有）
doi:10.1007/s10589-016-9832-2

(4) R. Miyashiro, Y. Takano, Mixed integer second-order cone programming formulations for variable selection, European Journal of Operational Research, No. 247, 2015, pp. 721－731.（査読有）
doi:10.1016/j.ejor.2015.06.081

(5) 小林健, 高野祐一, 宮代隆平, 中田和秀, 多重共線性を考慮した回帰式の変数選択：混合整数半正定値計画法を用いた解法, 最適化アルゴリズムの進展：理論・応用・実装, 京都大学数理解析研究所講究録, No. 1931, 2015, pp. 169－183.（査読無）
http://www.kurims.kyoto-u.ac.jp/~kyodo/kokyuroku/contents/pdf/1931-14.pdf

(6) R. Miyashiro, Y. Takano, Subset selection by Mallows' Cp: a mixed integer programming approach, Expert Systems with Applications, No. 42, 2015, pp. 325－331.（査読有）
doi:10.1016/j.eswa.2014.07.056

(7) 宮代隆平, 高野祐一, 混合整数 2 次錐計画法による回帰式の変数選択, オペレーションズ・リサーチ, No. 59, 2014, pp. 732－738.（査読無）
http://www.orsj.or.jp/archive2/or59-12/or59_12_732.pdf

(8) 宮代隆平, 高野祐一, 混合整数二次錐計画法を用いた回帰式の変数選択, 最適化の基礎理論と応用：京都大学数理解析研究所講究録, No. 1879, 2014, pp. 222－229.（査読無）
http://hdl.handle.net/2433/195609

〔学会発表〕（計１２件）

(1) 田村隆太, 小林健, 高野祐一, 宮代隆平, 中田和秀, 松井知己, 分散拡大要因を考慮した変数選択問題とその混合整数二次計画法による定式化, 日本オペレーションズ・リサーチ学会 2017 年春季研究発表会, 2017 年 3 月 17 日, 沖縄県市町村自治会館（沖縄県・那覇市）

(2) 田村隆太, 小林健, 高野祐一, 宮代隆平, 中田和秀, 松井知己, 多重共線性を除去するための最良部分集合選択, 日本オペレーションズ・リサーチ学会 2017 年春季研究発表会, 2017 年 3 月 17 日, 沖縄県市町村自治会館（沖縄県・那覇市）

(3) Ryuta Tamura, Ken Kobayashi, Yuichi Takano, Ryuhei Miyashiro, Kazuhide Nakata, Tomomi Matsui, A mixed integer semidefinite programming approach for variable selection avoiding multicollinearity, The Fifth International Conference on Continuous Optimization (ICCOPT 2016), August 10, 2016, National Graduate Institute for Policy Studies（東京都・港区）

(4) 佐藤俊樹, 高野祐一, 宮代隆平, 区分線形近似を用いた逐次ロジットモデルの変数選択, 日本オペレーションズ・リサーチ学会 2016 年春季研究発表会, 2016 年 3 月 17 日, 慶應義塾大学（神奈川県・横浜市）

(5) 佐藤俊樹, 高野祐一, 宮代隆平, 区分線形近似を用いた逐次ロジットモデルの変数選択, 情報処理学会 第 78 回全国大会, 2016 年 3 月 11 日, 慶應義塾大学（神奈川県・横浜市）

(6) Ryuta Tamura, Ryuhei Miyashiro, Yuichi Takano, Improvement on a stepwise method in variable selection problem in linear regression, The 2015 4th ICT International Student Project Conference (ICT-ISPC 2015), May 24, 2015, Tokyo University of Agriculture and Technology（東京都・小金井市）

(7) 小林健, 高野祐一, 宮代隆平, 中田和秀, 多重共線性を考慮した回帰式の変数選択問題に対する汎用解法, 日本オペレーションズ・リサーチ学会 2015 年春季研究発表会, 2015 年 3 月 27 日, 東京理科大学（東京都・新宿区）

(8) 佐藤俊樹, 高野祐一, 宮代隆平, 吉瀬章子, 混合整数最適化によるロジスティック回帰モデルの変数選択, 日本オペレーションズ・リサーチ学会 2015 年春季研究発表会, 2015 年 3 月 27 日, 東京理科大学（東京都・新宿区）

(9) 小林健, 高野祐一, 宮代隆平, 中田和秀, 多重共線性を考慮した回帰式の変数選択―混合整数半正定値計画問題を用いた解法―, 京都大学数理解析研究所研究集会 最適化アルゴリズムの進展：理論・応用・実装, 2014 年 9 月 25 日, 京都大学（京都府・京都市）

(10) 高野祐一, 宮代隆平, Mallows の Cp 規準による回帰式の変数選択：混合整数二次計画法を用いた解法, FIT2014 第 13 回情報科学技術フォーラム, 2014 年 9 月 4 日, 筑波大学（茨城県・つくば市）

(11)　小林健，高野祐一，宮代隆平，中田和秀，多重共線性を考慮した回帰式の変数選択―混合整数半正定値計画法を用いた解法―，日本オペレーションズ・リサーチ学会 2014 年秋季研究発表会，2014 年 8 月 29 日，北海道科学大学（北海道・札幌市）

(12)　Ryuhei Miyashiro, Yuichi Takano, Mixed integer second-order cone programming formulations for variable selection , SIAM Conference on Optimization 2014, May 19, 2014, San Diego (CA, USA)

６．研究組織

(1)研究代表者

宮代　隆平（MIYASHIRO, Ryuhei）

東京農工大学・大学院工学研究院・准教授

研究者番号：５０３７６８６０