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　１３　日現在

機関番号：　１３９０３
研究種目：　基盤研究(B)（一般）
研究期間：　2014～2016
課題番号：　２６２８００８３

研究課題名（和文）ビッグデータ解析における最適保証スクリーニングの理論と応用

研究課題名（英文）Theory and application of optimality-guaranteed screening methods for big-data analysis

研究代表者

竹内　一郎（takeuchi, ichiro）

名古屋工業大学・工学（系）研究科（研究院）・教授

研究者番号：４０３３５１４６

交付決定額（研究期間全体）：（直接経費）　　12,500,000円

研究成果の概要（和文）：超大規模データには高度な機械学習アルゴリズムをそのまま適用することができないため，重要でないと予測されるデータをスクリーニングにより削除し，残された一部のデータのみを解析するという場合が多い.しかしながら，従来のスクリーニング手法はヒューリスティクスであり，重要な情報を誤って削除してしまうリスクがあった．本研究では最適保証のあるスクリーニングのための理論と応用に関する研究を行い、おおきく３つの研究成果を得た.

研究成果の概要（英文）：It is often difficult to apply advanced machine learning methods to big data. In such a case, a common approach is to screen out some features and/or instances before the data is fed into machine learning algorithms. Existing screening methods are heuristics in the sense that there is no guarantee that the features and/or instances screened out by the methods are truly irrelevant. In this study, we investigated theory and application of new approach called optimality-guaranteed screening (it is also called safe screening in the literature). We obtained three significant results. The first result is the application of optimality-guaranteed screening to machine learning problems in dynamic environments. The second result is the extension of the scope of optimality-guaranteed screening to the field of pattern mining. The third result is the development of a new method for optimality-guaranteed screening that enables us to screen out features and instances simultaneously.

研究分野：　機械学習

キーワード：　機械学習　ビッグデータ　スクリーニング　凸最適化　パターンマイニング

１．研究開始当初の背景

(1) 大規模・高次元データから知識を獲得する試みは，ビッグデータの名のもと，様々な分野で注目されている.例えば，医学生物学分野では，数百万を超える次元の網羅的遺伝情報が日常的に蓄積され，データ解析技術が重要な科学的発見を支える礎となっている．また、個々の因子の相互作用効果を考える場合には、超高次元なデータをあつかはなくてはならない．データからの知識獲得は，機械学習の分野を中心として古くから研究されてきたが,データの急激な大規模化・高次元化によって，従来の枠組を超えるアプローチが必要とされていた．ビッグデータ研究では，データ全体をメモリにロードできない場合もあり，従来の機械学習アルゴリズムをそのまま適用することができなかった.

(2) このような状況では，なんらかの方法によってデータの一部を削除し，残された一部のデータのみを詳しく解析するといった方針を採らざるを得ない．簡単なルールに基づいてデータの一部を削除することはスクリーニングと呼ばれているが,従来のスクリーニング手法はすべてヒューリスティクスに基づいていた．そのため，重要なデータを誤って削除してしまう可能性があり，最終的なモデルが最適である保証は得られなかった．網羅的遺伝情報解析では，スクリーニングにおいて重要な遺伝子を取り除いてしまう可能性があり，貴重な科学的発見を見逃してしまうリスクが懸念されていた．また、客観的な指標のないスクリーニングを行うことにより、本来は重要でなかった要素が重要であると誤検出されてしまう問題も起こりうる.

２．研究の目的

(1) 2010 年に El Ghoui らによって最近提案されたアプローチ（引用文献①）は，最適保証のあるスクリーニングが可能となることを示唆する画期的なもので，代表者を含む一部の研究者が効率化や汎用化のための拡張を試みている状況であった．最適性を保証しつつデータのサイズを減らすことはビッグデータ研究にとって本質的に重要であり，本研究課題では，理論・応用の両面において関連研究を発展させ，同分野をリードすることを目指すことを目的とした．代表者は 2013 年より最適保証スクリーニングの研究を初めており、その最初の成果は 2013 年の International Conference on Machine Learning で発表されている.

(2) 最適性を保証しつつデータのサイズを減らすことはビッグデータ研究にとって本質的に重要である.最適保証スクリーニングは凸最適化理論に基づいており，様々な問題設定において同様のアイデアが利用可能と考えられる.上述の El Ghoui らのアプローチは機械学習分野で注目を集めており,研究期間の３年間で関連研究が大きく進展する可能性があった．この状況を踏まえ，本研究では，同分野の研究を理論・応用の両面でリードすることを目的とした.

３．研究の方法

(1) 当時までの最適保証スクリーニング研究は高次元データから特徴をスクリーニングする問題に限られていた．本研究では，問題対象を拡げ,最適保証スクリーニング技術の新たな応用、最適保証スクリーニング技術のパターンマイニング問題への拡張特徴と事例の同時最適保証スクリーニング技術の開発の課題に取り組んだ.

(2) これまでの最適保証スクリーニング研究は理論に重点を置いており，実社会の問題に対する有効性が十分に検証されていない．医学生物学分野は探索的アプローチが最も成功している分野の１つであり，データ分析技術が科学的発見の礎となっている．本研究では医学生物学分野の網羅的遺伝情報解析を中心的な応用範囲と定め、最適保証スクリーニングの実証研究を行った.

４．研究成果

(1) １つ目の成果は最適保証スクリーニングを動的環境下での機械学習問題へ応用したことである．この方法は高速感度分析（Quick Sensitivity Analysis）と呼ばれ、機械学習に適用するデータベースに少量の更新があった場合に、計算コストをかけて機械学習モデルを再学習する必要はなく、簡単な計算で機械学習モデルの変化しうる範囲を適切にしることができるものである．この研究成果は機械学習における重要な技術であるクロスバリデーションの高速化などにも応用することができる．本研究では、事例の追加や削除が行われた場合、特徴の追加や削除が行われた場合、データベースの要素に変更があった場合それぞれに対応することができ、おおまかには、変更のあったサイズの線形時間で機械学習モデルが変化しうる範囲を求めることができる．本成果はデータマイニング分野の最難関国際会議である 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining（KDD2015）に採録され、国内外から注目を集めている（主な発表論文等①）.

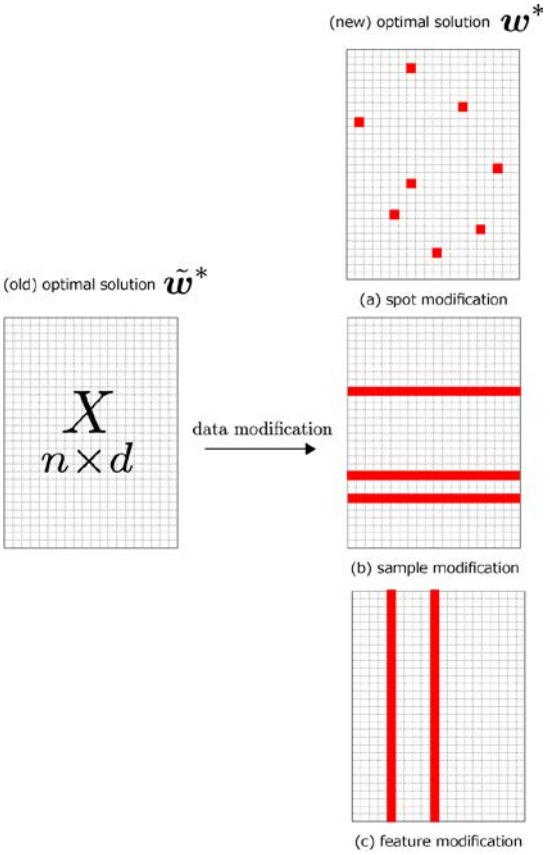

図 1 高速感度分析のイメージ図．データ変更が行われる前の最適モデルパラメータを利用してデータ変更後の最適モデルパラメータの存在範囲を高速に計算することができる．

(2) ２つ目の成果は最適保証スクリーニングの適用範囲をパターンマイニングの分野へ拡張したことである．パターンマイニング問題とは、例えば、複数のアイテムから頻出するアイテムの組み合わせを数え上げるような問題(頻出アイテムセットマイニング)である．本研究では、回帰モデルや分類モデルの係数がパターンから成るような予測パターンマイニングの問題を考え、その学習の高速化に最適保証スクリーニングを利用するためのアプローチを提案した．具体的には、パターン空間上で定義される木構造において、最適性を保証しつつ枝刈りをする方法を提案した．この方法はセーフプルーニング（safe pruning）と呼ばれ、アイテムセットの予測マイニングだけでなく、グラフの予測マイニングや系列の予測マイニングにも適用することができる．本成果はデータマイニング分野の最難関国際会議である 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining（KDD2016）に採録され、国内外から注目を集めている（主な発表論文等②）．

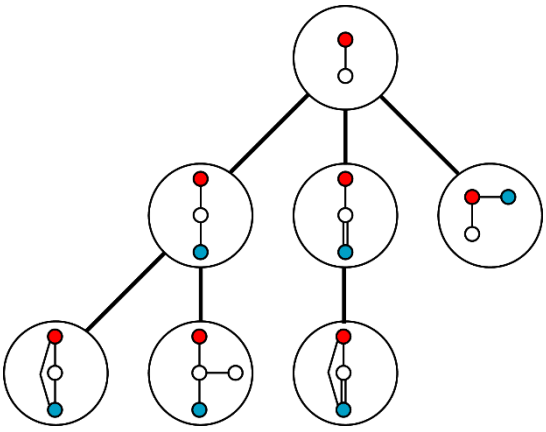

図 2 最適保証スクリーニングを利用したパターンマイニングのイメージ図．木構造で表現されたパターンのうち、最適な予測モデルでは利用されないことが保証されるパターンを枝刈りできる．

(3) ３つ目の成果は特徴と事例を同時に最適保証スクリーニングするための新たな方法を開発したことである．これまで、機械学習分野では、主にスパースモデルの特徴の最適保証スクリーニングが研究されてきた．一方、筆者は本研究が開始する直前の 2013 年に事例の最適保証スクリーニングを初めて提案した（引用文献②）．その後、様々な機械学習モデルにおける特徴のスクリーニング、事例のスクリーニングが個別に研究されてきたが、本研究では、特徴のスクリーニングと事例のスクリーニングを同時に行うための方法を初めて開発した．同時のスクリーニングすることの利点は相乗効果があることであり、特徴のスクリーニングを行うと事例のスクリーニング率が向上し、事例のスクリーニングを行うと特徴のスクリーニング率が向上することを解明した．このアプローチを実際の大規模機械学習問題に適用したところ、大幅な計算コストの削減がみられた．本研究の成果は機械学習の最難関国際会議の１つである 33rd International Conference on Machine Learning(ICML2016) に採録され、内外から大いに注目を集めている（主な発表論文等③）．

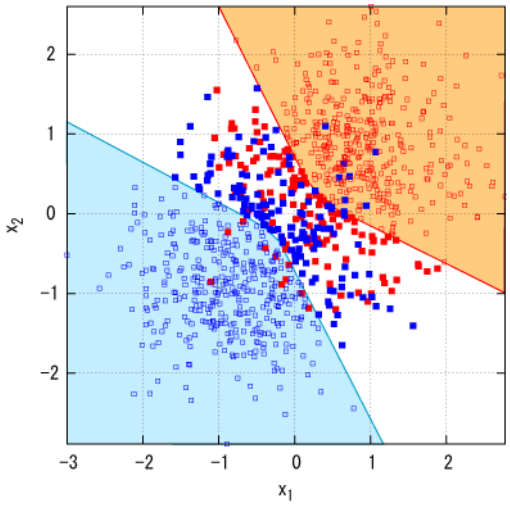

図 3事例スクリーニングのイメージ図. 2クラス分類問題において一部の事例を取り除いても最適な分類境界に影響を与えないが、最適保証スクリーニングではそのような事例を事前に見つけることができる.

＜引用文献＞

① El Ghaoui et al., Safe feature elimination in sparse supervised learning, arXiv:1009.4219. 2010.

② K. Ogawa, Y. Suzuki, I. takeuchi. Safe screening of non-support vectors in pathwise SVM computation. Proceedings of the 30th International Conference on Machine Learning (ICML2013). 2013.

５. 主な発表論文等

〔雑誌論文〕(計 4 件)

① S. Okumura, Y. Suzuki, I. Takeuchi. Quick sensitivity analysis for incremental data modification and its application to leave-one-out CV in linear classification problems. Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD2015). 2015.

② Nakagawa K., Suzumura S., Karasuyama M., Tsuda, K., Takeuchi I. Safe Pattern Pruning: An Efficient Approach for Predictive Pattern Mining. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD2016). 2016.

③ A. Shibagaki, M. Karasuyama, K. Hatano, I. Takeuchi. Simultaneous safe screening of features and samples in doubly sparse modeling. Proceedings of the 33rd International Conference on Machine Learning (ICML2016). 2016.

④ S. Suzumura, K. Ogawa, M. Karasuyama, M. Sugiyama, I. Takeuchi. Homotopy continuation approaches for robust SV classification and regression. Machine Learning. 2017, 1-30. DOI:10.1007/s10994-017-5627-7.

⑤ K. Toyoura, D. Hirano, A. Seko, M. Shiga, A. Kuwabara, M. Karasuyama, K. Shitara, I. Takeuchi. Machine-learning-based selective sampling procedure for identifying the low-energy region in a potential energy surface: A case study on proton conduction in oxides. Physical Review B. 93, pp. 054112: 1-11. 2016.

⑥ T. Takada, H. Hanada, Y. Yamada, J. Sakuma, I. Takeuchi. Secure approximation guarantee for cryptographically private empirical risk minimization. Proceedings of the 8th Asian Conference on Machine Learning (ACML2016). 2016.

〔学会発表〕(計 5 件)

① 奥村翔太・鈴木良規・小川晃平・新村祐紀.竹内一郎. 二次正則化分類学習のための Leave-one-out cross-validation の高速化. 電子情報通信学会第 25 回情報論的学習理論研究会. 2014 年 11 月 17 日.

② 竹内一郎.大規模機械学習のための事例と特徴のセーフスクリーニング. 第 18 回情報論的学習理論ワークショップ (IBIS2015). 2015 年 11 月 25 日 (招待講演).

③ I. Takeuchi. Safe Feature/Sample Screening and Its Applications to High-order Interaction Modeling and Quick Sensitivity Analysis. The First Korea-Japan Machine Learning Symposium. 2016 年 6 月 2 日 (招待講演).

④ 柴垣篤志, 烏山昌幸, 畑埜晃平, 竹内一郎. スパースモデルのための特徴と標本の同時セーフスクリーニング. 電子情報通信学会第 25 回情報論的学習理論研究会. 2016 年 7 月 6 日.

⑤ 中川和也, 鈴村真矢, 烏山昌幸, 津田宏治, 竹内一郎. パターンマイニング問題におけるセーフパターンプルーニングを用いたスパースモデルの学習. 電子情報通信学会第 26 回情報論的学習理論研究会. 2016 年 9 月 5 日.

⑥ 花田博幸, 柴垣篤志, 佐久間淳, 竹内一郎. 経験損失最小化問題における高速感度分析に関する一提案. 電子情報通信学会第 26 回情報論的学習理論研究会. 2016 年 9 月 6 日.

⑦ 烏山昌幸, 田村友幸, 小林亮, 竹内一郎, 中山将伸. ガウス過程に基づく粒界カスプ構造の確率的探索. 電子情報通信学会第27回情報論的学習理論研究会. 2016 年 11 月 16 日.

⑧ 梅津佑太, 中川和也, 津田宏治, 竹内一郎. 高次元分類問題のための Selective Inference. 電子情報通信学会第 27 回情報論的学習理論研究会. 2016 年 11 月 16 日.

⑨ 花田博幸, 高田敏行, 柴垣篤志, 佐久間淳, 竹内一郎. 区間データに対する経験損失最小化とそのプライバシー保護への応用. 電子情報通信学会第 27 回情報論的学習理論研究会. 2016 年 11 月 17 日.

⑩ 金森研太, 豊浦和明, 中島伸一, 世古敦人, 烏山昌幸, 桑原彰秀, 本多淳也, 設楽和希, 志賀元紀, 竹内一郎. ガウス過程と動的計画法を用いたプロトン伝導体の伝導度推定. 電子情報通信学会第 27 回情報論的学習理論研究会. 2016 年 11 月 17 日.

〔図書〕(計 0 件)

なし

〔産業財産権〕

○出願状況 (計 0 件)
なし

○取得状況 (計 0 件)
なし
〔その他〕

なし

６．研究組織

(1)研究代表者
竹内　一郎 (TAKEUCHI,Ichro)
名古屋工業大学・工学(系)研究科(研究院)・教授
研究者番号：40335146

(2)研究分担者
烏山　昌幸 (KARASUYAMA,Masayuki)
名古屋工業大学・工学(系)研究科(研究院)・助教
研究者番号：40628640

(3)研究分担者
畑埜　晃平 (HATANO,Kohei)
九州大学・学内共同利用施設等・准教授
研究者番号：60404026