様式Ｆ－１９

科学研究費助成事業（学術研究助成基金助成金）研究成果報告書

平成　２４　年　６　月　５　日現在

機関番号：13901
研究種目：若手研究(B)
研究期間：2008～2011
課題番号：20700251
研究課題名（和文）
機械学習と最適化理論の横断的研究
研究課題名（英文）transversal study of machine learning and optimization

研究代表者
金森 敬文　（KANAMORI TAKAFUMI）
名古屋大学・情報科学研究科・准教授
研究者番号：60334546

研究成果の概要（和文）：統計科学や計算機科学の境界領域である機械学習の分野では，大規模データに潜む確率的な構造を，精度よく推論するための方法が研究されている．一方最適化理論の分野では，コンピュータを高度に用いて，様々な意思決定に現れる数理的な問題を，効率的に計算するための方法が研究されている．本研究では，機械学習と最適化理論を包括する理論基盤を構築し，高度な推論・計算アルゴリズムを開発することを目的としている．

研究成果の概要（英文）：My target of this study is to provide a transversal study of machine learning and optimization theory. We apply machine learning techniques to optimization problems with noisy data. Inversely, highly developed optimization algorithms are available to conduct statistical analysis of real-world high-dimensional data. Unifying machine learning and optimization is promising for the advanced information processing.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2008 年度	900,000	270,000	1,170,000
2009 年度	800,000	240,000	1,040,000
2010 年度	800,000	240,000	1,040,000
2011 年度	800,000	240,000	1,040,000
総　計	3,300,000	990,000	4,290,000

研究分野：総合領域
科研費の分科・細目：情報学・統計科学
キーワード：統計的学習理論

１．研究開始当初の背景

高度情報化に伴い，さまざまな科学技術の分野でデータが高次元化，複雑化している．例えば，医学・遺伝子データ，創薬分野の化学物質の合成に関するデータ，ウェブ上のさまざまなタイプのデータ，地質学・地震のデータなど，無視できない不確実性を含む大規模データを扱う必要性が，社会的に著しく増大している．このため，大規模データを高速に処理し，適切な情報を抽出するための統計解析の方法を開発することは，極めて重要である．統計科学，とくに機械学習の分野において，統計的推論と高速な計算手法の融合を目指して研究が進められ始めている．その成果として，サポートベクターマシンやブースティングなどの学習アルゴリズムが提案され，広く実データに応用されている．

一方で，最適化理論の分野において「不確実性のもとでの最適化」という問題が，とくに近年強く意識されている．最適化の分野では従来，確定している関数を最適化するための計算アルゴリズムが研究されていた．しかし現実の問題では，例えば株価の変動や測定誤

差のような，特定することが困難な不確実性が存在する．そのため，最適化すべき関数が確定しないという問題が生じる．不確実性のもとでの最適化では，そのような状況における最適化問題の定式化と，計算アルゴリズムの開発が研究されている．

以上のような研究の背景から，機械学習と最適化理論との間には「推論と計算の融合」という共通の問題が存在することが分かる．本研究においては，機械学習と最適化の理論的な融合をより深化させ，腱固な数学的基礎のもとに学習アルゴリズムや最適化アルゴリズムを開発することを目的としている．

２．研究の目的
機械学習と最適化理論との間には「推論と計算の融合」という共通の問題が存在する．本研究においては，機械学習と最適化の理論的な融合をより深化させ，腱固な数学的基礎のもとに学習アルゴリズムや最適化アルゴリズムを開発することを目的としている．機械学習と最適化の融合として，以下の３つの方向性を考えている．

(1) 最適化から機械学習へのアプローチ：
大規模なデータを統計的に解析するとき，汎用的な最適化ツールを利用するだけでは十分な計算速度が得られないことが多い．したがって，機械学習の特定の問題を効率的に処理するための最適化法が必要になる．我々は，最適化や金融工学の分野で近年研究されている条件付きバリュー・アット・リスク(CVaR)というリスク尺度とサポートベクターマシンの関連に着目して，CVaR の効率的な最適化法やその一般化に関する研究を行い，同時にサポートベクターマシンや機械学習への応用を探っていくことを目指している．また，パラメトリック最適化を分位点回帰推定に応用して，効率的に条件付き確率分布を推定するための学習アルゴリズムの開発を進める．

(2) 機械学習から最適化へのアプローチ：
不確実性のもとでの最適化問題では，不確実性をどのように最適化問題に組み込むかを考える必要がある．代表的な方法のひとつとしてサンプルを用いる最適化がある．サンプルを用いる方法では，無数にある不確実性の一部をサンプリングして，計算可能な最適化問題を構成する．この方法は機械学習におけるデータからの推論と数学的に等価である．したがって機械学習の研究成果である，推定結果の精度保証のための理論を援用することで，不確実性のもとでの最適化問題の解に対する精度保証を与えることが可能になる．我々はさらに機械学習の理論的な研究成果を応用して，少ないサンプル数で，最適値や最適解の精度を保証するための効率的な最適化アルゴリズムを開発するための研究をすすめる．

(3) 機械学習と最適化の融合：
機械学習では，損失関数を最小化することで，所望の推定量や判別機械を構成するという方法が一般的である．これは最適化理論の主要な応用と言うこともできる．本研究では機械学習と最適化理論のより深い理論的融合を目指す．統計的学習理論では，最悪の状況で最善の手を打つというミニマックスの考え方で推定量を構成することもある．これは最適化やゲーム理論のミニマックス定理と共通の枠組で議論することができる．我々はこの方向に沿って研究をすすめ，最適化法と学習アルゴリズムとの対応関係を明かにし，「推論と計算」を両立するアルゴリズムを開発するための理論的な基盤を整備することを目指している．

３．研究の方法
本研究では，機械学習と最適化理論を横断する理論的な枠組を構築することを計画している．このような研究を進めるために，国内外のさまざまな研究会や国際会議に参加して，他の研究者と議論する機会を多くもつことが重要と考えている．研究目標は主に３つのテーマからなるが，いずれにおいても最終的には，理論的な成果のみならず，実用的な学習・最適化アルゴリズムの構築と実装を目指す．

(1) 2008 年度
本研究では，機械学習と最適化理論を結ぶ理論的な枠組を構築することを計画している．このような横断的な研究を進める上では，国内外のさまざまな分野の研究者と議論する機会が非常に重要となる．主に最適化や機械学習の国際会議などに積極的に出席し，その分野の研究者と情報交換を図る機会を増やすことを計画している． また国内外の研究者を招いて講演を依頼するなどの活動を通して，情報交換をおこなっていく方針である．

理論的研究だけではなく，効率的な学習・最適化アルゴリズムの開発も本研究の大きなテーマである．この目的を達成するためには，豊富な計算機資源によるプログラム開発環境が必要不可欠である． 多くの統計解析で使用されているフリーの統計解析ソフト「R」などによりプログラム開発をおこない，ソフトウェアを公開していくことを計画している．とくに初年度は，理論研究をサポートするための数値実験などを中心に，数値解析の研究も進めることを計画している．

以下，個々の研究テーマについての研究計画・方法を具体的に述べる．

まず第１の研究テーマとして，さまざまリスク尺度やパラメトリック最適化を機械学習へ応用することを考える．最適化や金融工学で現れる条件付きバリュー・アット・リスク(CVaR)とサポートベクターマシンとの関連を一般化し，より判別精度の高い学習アルゴリズムを構築することを目指す．具体的には一般化 CVaR の統計的な意味を明確にして，機械学習に応用する．また，説明変数に測定誤差が加わっているデータの学習に対してロバスト最適化を応用して，安定した学習アルゴリズムを開発することを目標とする．さらに，すでに国際会議で発表されている我々の研究を発展させ，パラメトリック最適化を機械学習に積極的に導入する．これを通して，クロス・バリデーションなどの計算コストを劇的に軽減するための学習アルゴリズムの開発を進める．

第２の研究テーマは，不確実性のもとでの最適化である．国際会議での我々の発表を踏まえて研究を進める．現段階では，サンプルを用いた最適化問題を機械学習の観点から考察して，最適解と最適値の精度保証を与えている．しかし不確実性を表すパラメータの次元が高いとき，単純なサンプリングを用いた方法では，要求される精度を達成するためのサンプル数が莫大になる．このため最適化問題が実行可能ではなくなる傾向がある．このような「次元の呪い」の問題を回避するために，不要なサンプルを間引くための基準を開発する．また，単純なサンプリングではなく重点的サンプリングを行うことで，計算量の増加を回避するための方法を開発することを計画している．

第３の研究テーマとして，機械学習と最適化のゲーム理論的融合を掲げる．モーメント法と最大エントロピー法が等価であることは，ミニマックス原理により以前から知られているが，この関係をゲーム理論の観点から一般化しようという試みが，数理統計学や機械学習の研究者によって進められている．しかし，ミニマックス原理を通してさらに最適化との関連を探ろうという研究はまだ多くはないと我々は考えている．機械学習と最適化の興味深い関連を示すいくつかの論文が出版されているが，状況はまだ混沌としており，統一的な理解には至っていない．このような状況において，まず関連文献の調査を行って現状を精確に把握して，将来の理論的基盤の構築や実用的なアルゴリズム開発のための足掛かりとしたい．

(2) 2009 年度以降
研究は上述の３つの研究テーマに大きく分かれている．それらの進展具合に応じて 2009 年度以降の見直しを検討する．

理論的で基礎的な，第３の研究テーマについては，研究の進展具合を見積ることは難しい．進展が著しくなければ，他の研究者にも研究協力者として参画を依頼し，新しい視点から議論を深めていくことを計画している．

第１，第２の研究テーマについては，かなりの進展が期待される．これらの研究成果を踏まえて，2009 年度以降は実用的な学習アルゴリズムや最適化手法を開発していくことを計画している．さらに提案されたアルゴリズムを実装し，現実の問題に応用することまで視野に入れている．学習理論の分野における実データとしては，とくにテキストデータの判別を想定している．各テキストデータに対して適切な分類を行うことにより，テキストデータのウェブ検索などの効率を劇的に向上させることが可能になる．ウェブ上のテキストデータは非常に大規模であり，統計的な精度を保持しながら効率的な計算をおこなうことは非常に重要な課題と考える．このような課題に対して，充実した計算機環境を用いて，実用的な学習アルゴリズムを実装することを計画している．

４．研究成果
(1)2008 年度
主に密度比の推定とその応用，分位点回帰分析，ブースティングに関する研究を行った．

密度比とは，ふたつの確率密度関数の比によって表される関数であり，共変量シフトのもとでの学習や外れ値検出などの応用されている．本研究では２乗誤差関数にもとづく方法を提案した．この方法は既存の方法と比べて計算効率が優れており，推定量と交差検証法の計算が解析的に実行できる点に特徴がある．さらに密度比推定を独立成分分析へ応用した．提案手法の推定精度は他の方法と比較して優れていることが明らなになった．これらの成果は複数の論文にまとめられ，出版されている．

分位点回帰分析を用いた条件付き密度関数の推定のための方法を提案した．分位点回帰は回帰分析の分位点関数を推定する方法であり，パラメリック最適化の手法を組み合わせることによって，条件付き密度関数の推定が可能になった．この成果は国際的な評価が極めて高い英語論文誌上で出版されている．

ブースティングとは，学習アルゴリズムを組み合わせること，高精度の予測を行う手法である．多値判別においてミスラベルを考慮したときのロバストなブースティング法を開発し，理論的な性質について研究を行った．この研究成果は，国際的な評価が極めて高い英文論文誌上で発表されている．

(2)2009 年度
主に密度比の推定の理論とその応用，また最適化手法の機械学習への応用について研究を行った．

密度比に関する研究では，２乗誤差関数に基づく方法を提案し，そのソフトウェアを開発して公開している．また新たな応用として特徴抽出や２標本問題を取り上げた．さらに推定精度に関して理論的な研究を推進した．その結果，提案した方法が，他の方法と比較して，統計モデルが必ずしも正しくない状況において優れた性能を有することが明らかにされ，理論と実践の両面において特段の進展を遂げることとなった．これらの成果は複数の論文にまとめられ，国際的な評価が極めて高い英語論文誌にて出版されている．

金融工学において重要な指標として利用されている条件付きバリューアットリスクを，機械学習における２値判別の問題に応用した．とくに入力を観測する際にノイズが加わる状況を考慮した推論問題へ拡張を行い，理論と応用の両面で成果を得た．また最適化手法と機械学習の密接な関連についても考察し，両分野にまたがる研究成果として今後の進展が期待される．これらの成果は，英語文論として出版されている．

(3)2010 年度
主に密度比の推定の理論とその応用，最適化計算への幾何学的方法の応用について，研究を行った．また機械学習アルゴリズムの統計的信頼性に関する研究をスタートさせた．

密度比の研究に関して，2009 年度に提案した推定方法の理論的な推定精度に関する詳細について，研究を進めた．また実問題への応用として，密度比推定を外れ値検出に用い，成果を得た．これらの結果は，複数の英語論文にまとめられ，国際雑誌に出版されている．

機械学習アルゴリズムの推定結果に対する統計的信頼性を評価するための方法であるマルチスケールブートストラップについて研究を行った．これにより，従来の方法では正確に推定することが困難だった予測誤差を，精度よく推定することが可能となった．これらの結果は英語論文としてまとめられ，国際会議で報告されている．

最適化問題に対する解法のひとつである準ニュートン法の解析に対して，幾何学的な方法を導入した．また従来のアルゴリズを拡張し，さらに統計的な方法を用いて，アルゴリズムの頑健性を理論的に評価した．これらの結果は英語論文としてまとめられ，国際会議で報告されている．

(4)2011 年度
主に，密度比の推定の理論とその応用，また最適化理論に関する研究を行った．

密度比を外れ値検出の問題に応用し，さまざまな工学的応用について考察した結果は，英語論文として出版されている．また密度比モデルのもとでの２標本検定について考察した．その研究結果は，複数の論文として出版されている．この結果は，密度比モデルのもとでの情報量限界を理論的に示すものであり，重要な進展と考える．さらに本年度は，密度比に関するこれまでの研究成果を，英文書籍としてまとめた．これはケンブリッジ大学出版から出版されている．密度比推定を体系的に論じた書籍は他に例がなく，今後の機械学習，数理統計学の進展にとって重要な貢献となっている．

また，最適化と機械学習・統計学の両分野に深く関わる研究として，グループテストの検定統計量の効率的な計算法について，研究を行った．この成果は英語論文として出版されている．

さらに，ランダムなノイズが存在する場合の最適化問題についても考察し，統計的学習理論の方法を最適化に応用して解析した．この結果，最適解の統計的性質を明らかにした．これは最適化に関する学術雑誌に英語論文として出版されている．

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 17 件）
① T. Kanamori, A. Takeda, Worst-Case Violation of Sampled Convex Programs for Optimization with Uncertainty. Journal of Optimization Theory and Applications, vol. 152, Issue 1, pp.171-197, 2012. 査読有.

② T. Kanamori, T. Suzuki, M. Sugiyama, f-divergence estimation and two-sample homogeneity test under semiparametric density-ratio models. IEEE Transactions

on Information Theory, Vol. 58, Issue 2, pp. 708-720, 2012. 査読有.

③T. Kanamori, T. Suzuki, M. Sugiyama, Statistical analysis of kernel-based least-squares density-ratio estimation. Machine Learning, vol. 86, Issue 3, pp. 335-367, 2012. 査読有.

④ H. Shimodaira, T. Kanamori, M. Aoki, K. Mine, Multiscale Bagging and its Applications. IEICE Transactions on Information and Systems, Volume E94-D No.10, pp.1924-1932, 2011. 査読有

⑤ T. Kanamori, Deformation of Log-Likelihood Loss Function for Multiclass Boosting. Neural Networks, vol. 23, pp. 843-864, May, 2010. 査読有.

⑥T. Kanamori, T. Suzuki, M. Sugiyama, Theoretical Analysis of Density Ratio Estimation. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, vol. E93-A, no. 4, pp. 787-798, April, 2010. 査読有.

⑦T. Kanamori, S. Hido, M. Sugiyama, A Least-squares Approach to Direct Importance Estimation. Journal of Machine Learning Research. 10(Jul):1391-1445, 2009. 査読有.

⑧I. Takeuchi, K. Nomura, T. Kanamori, Nonparametric Conditional Density Estimation Using Piecewise-Linear Path Following for Kernel Quantile Regression. Neural Computation, vol. 21, num. 2, pp. 533-559, 2009. 査読有.

⑨T. Takenouchi, S. Eguchi, N. Murata, T. Kanamori, Robust Boosting Algorithm against Mislabelling in Multi-Class Problems. Neural Computation, vol. 20, num. 6, pp. 1596-1630, 2008. 査読有.

〔学会発表〕(計 46 件)
①M. Yamada, T. Suzuki, T. Kanamori, H. Hachiya, and M. Sugiyama, Relative density-ratio estimation for robust distribution comparison, Neural Information Processing Systems, Granada, Spain, 2011.12.13

②T. Kanamori and A. Ohara, A Bregman extension of quasi-Newton updates, Information Geometry and its Applications, Germany, 2010.08. 2

③Sugiyama, M., Hara, S., von Bünau, P., Suzuki, T., Kanamori, T., & Kawanabe, M., Direct density ratio estimation with dimensionality reduction SIAM International Conference on Data Mining Columbus, Ohio, USA, 2010.05. 29

④T. Kanamori, T. Suzuki, M. Sugiyama, Condition Number Analysis of Kernel-based Density Ratio Estimation, Numerical Mathematics in Machine Learning (NUMML2009), Montreal, Canada, 2009.06.18

⑤T. Kanamori, M. Sugiyama, and S. Hido, Efficient Direct Density Ratio Estimation for Non-stationarity Adaptation and Outlier Detection, Neural Information Processing Systems, Vancouver, B.C., Canada 2008.12.16

〔図書〕(計 4 件)
①M. Sugiyama, T. Suzuki, T. Kanamori, Cambridge University Press, Density Ratio Estimation in Machine Learning, 2012, 119-297.

②金森 敬文, 竹之内 高志, 村田 昇, 共立出版, パターン認識 (R で学ぶデータサイエンス 5), 2009, 1-153.

６．研究組織
(1)研究代表者
金森 敬文 (TAKAFUMI KANAMORI)
名古屋大学・情報科学研究科・准教授
研究者番号：60334546

(2)研究分担者
(　　)
研究者番号：

(3)連携研究者
(　　)
研究者番号：