様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成　２５年　６月　５日現在

機関番号：12601
研究種目：若手研究（B）
研究期間：2010　～　2012
課題番号：22700138
研究課題名（和文）　行列の上のスパース正則化に基づく機械学習
研究課題名（英文）　Machine learning based on sparsity-inducing regularization for matrices

研究代表者
冨岡　亮太（TOMIOKA RYOTA）
東京大学・大学院情報理工学系研究科・助教
研究者番号：70518282

研究成果の概要（和文）：
本事業の成果は以下の通りである．

1. ベクトルのためのスパース正則化を実現するための最適化アルゴリズムである双対拡張ラグランジュ(DAL) 法を行列のスペクトルに対する正則化に拡張し，国際会議 ICML2010 で発表した．さらに，その収束性に関して国際論文誌 JMLR で発表，また MIT Press から出版された Optimization for Machine Learning の一部として発表するとともに，ウェブページでプログラムを公開してアルゴリズムの普及に務めた．
2. 非数値データに対応するために DAL 法のマルチカーネル学習への拡張を行い，候補カーネルの数が数千という非常に大規模な問題にもスケールする方法を提案した．この成果は国際論文誌 Machine Learning で発表した．
3. 行列を含む高次テンソルに対するスペクトルに基づく正則化法を提案し，その理論的な性能を解析した．この成果は国際会議 NIPS2011 で発表した．

研究成果の概要（英文）：
The outcomes of this research project can be summarized as follows:

1. I have extended the dual augmented Lagrangian (DAL) algorithm to deal with spectral regularization for matrices and proposed the M-DAL algorithm (ICML2010). The super-linear convergence of DAL and M-DAL algorithms was proven and published in JMLR. A review of DAL and related algorithms has been published as part of “Optimization for Machine Learning” (MIT Press). I have also made the code publicly available to promote its use in wider research communities.
2. In order to deal with non-numerical data, I have extended DAL to handle multiple kernel learning with thousands of kernels. This was published in Machine Learning Journal.
3. I have extended the framework to spectral regularization for higher-order tensors and analyzed its statistical performance. This was presented at NIPS2011.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2010 年度	1,000,000	300,000	1,300,000
2011 年度	1,300,000	390,000	1,690,000
2012 年度	800,000	240,000	1,040,000
年度			
年度			
総　計	3,100,000	930,000	4,030,000

研究分野：総合領域

科研費の分科・細目：情報学・知能情報学
キーワード：機械学習，スパース性，正則化，行列，テンソル

１．研究開始当初の背景
近年，ウェブ,バイオインフォマティクス，各種センサー技術の発展によりこれらのデータは増大の一途をたどっている．これらのデータの中に潜んでいる規則性を数学的に表現し，データに基づい てこれを推定する手法の体系は機械学習やデータマイニングと呼ばれ，ICML, NIPS, KDD, ICDM などの国際会議を中心として活発に研究が進められている．

従来の機械学習が個々の対象のベクトル表現に基づくものであったのに対し，本研究課題はある対象と別の対象のベクトル表現の間の関係を行列を用いて表現し，これを推定することを目指すものである.このような研究はこれまで個々の応用分野での散発的なものに留まっており，扱えるデータ が数値的なものに限られているなど，いまだ体系化されておらず未発展である．

２．研究の目的
本研究課題は，行列に基づく機械学習の理論的な枠組みを構築する．今までの機械学習で用いられてきたベクトルに比べて行列は要素の数が多いため,行列に基づく機械学習においては予測精度だけでなく．なぜそのような予測になるか説明できることが重要である．そこで本研究課題ではスパース正則化を用いた説明可能な学習モデルの体系を築く．このような枠組みの上で，従来のベクトルに基づく機械学習を理論とアルゴリズムの両面で行列に対して拡張する．さらに，この体系の中でテキストなどの非数値データを扱ったり，類似度などの補助情報 を活用したりするための拡張を行うとともに,大規模な行列データへ適用するための最適化アルゴリズムの開発 を行い,これをテキスト解析，バイオインフォマティクスなどに適用する．

３．研究の方法
正則化付き経験リスク最小化の枠組みに基づき（a）部分観測行列の未知部分の推定 (b) 行列上の教師付き判別（c）多入力多出力の関係の推定のすべてを共通の学習問題として定式化し，それらに適切な最適化アルゴリズムを創出する．さらにこれらの手法を非数値データに対して拡張する．これらの行列に基づく機械学習手法に関して統計的な性能を明らかにするとともに，実問題への適用を行う．

４．研究成果
まず，最適化アルゴリズムに関して，従来のベクトル型データに対して提案していた双対拡張ラグランジュ法を行列に対して拡張することに成功した．提案した M-DAL 法は従来の勾配射影法，加速付き近接勾配法や，内点法に対して 10 倍から 100 倍高速で，大規模なデータに対して容易に適用可能である（図１）．また，損失関数などを柔軟に変更することができるため，様々な問題に対して応用することができる．

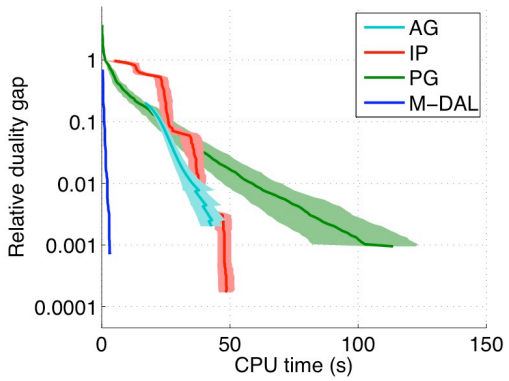

図 1: 提案した M-DAL 法と既存手法：内点法 (IP)，加速付き近接勾配法(AG)，勾配射影法 (PG)の比較. M-DAL は既存手法より 10 倍から 100 倍高速である．

次に非数値データへの拡張に関して，テキスト，画像などの非数値データを扱うため，マルチカーネル学習への双対拡張ラグランジュ法の拡張を行った．提案した SpicyMKL アルゴリズムは非常に多数の候補カーネルを効率的に扱うことができる．また，現実にマルチカーネル学習を行う際にはスパース性と汎化性能のトレードオフが重要になるが，これを解決するために ElasticnetMKL を提案した．

さらに，行列をより一般化したテンソル（多次元配列）に対するスペクトルに基づく正則化を提案し，世界ではじめて凸最適化に基づく効率的なテンソル分解のアルゴリズムを提案するとともに，その統計的な性能を解析し，明らかにした（図 2）．提案した凸最適化に基づくテンソル分解法は従来の非凸最適化に基づく方法に比べてモデルの特異性の問題がなく，安定に成分を求めることができることを実データ実験で示した（図 3）．

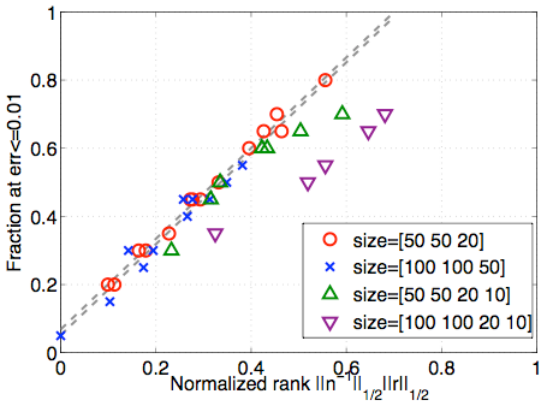

図 2：凸最適化に基づくテンソル分解の性能の理論予測（横軸）と実際の性能（縦軸）．理論は実験的な性能をよく予測することができる．

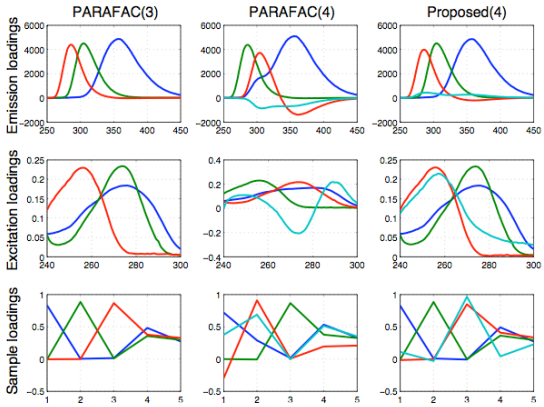

図 3：アミノ酸データに対して，凸最適化によって得られる成分（右列）と非凸最適化によって得られる成分（中央列）の比較．凸最適化を用いると安定して真の成分（左列）と同じパターンが得られる．一方，非凸最適化に基づく方法は非特異な状況では必ずしも真の成分と同じパターンが得られるとは限らない．

その他の実データへの応用として，デンマーク工科大学の M. Mørup 助教とともに，テンソル的なモデリングと確率的生成モデルを組み合わせて 2011 年 3 月の福島第一原子力発電所の事故を受けた各地の放射線量データを対象として，時間的空間的なカーブフィッティングを用いて核種を推定する問題に取り組んだ．このモデルは人工的に生成したデータに対しては正確に核種の種類数と半減期を推定することができることがわかった．実データに対してこの手法を適用したところ，いくつかの特徴的な核種が事故の初期に発電所内に存在したことを示唆することができた（図 4）．

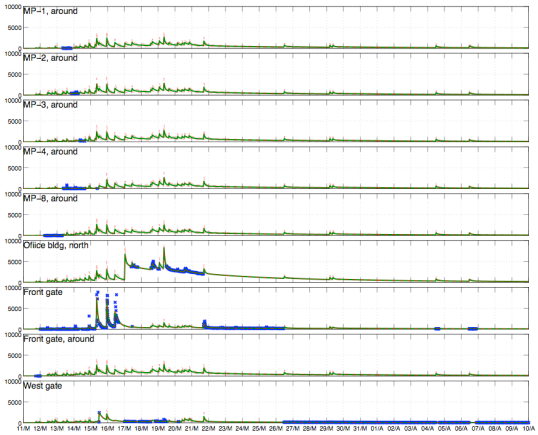

図 4：福島第一原子力発電所の線量データに対するカーブフィッティングの結果．

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 4 件）

1. R. Tomioka, T. Suzuki, and M. Sugiyama (2011) Super-linear convergence of dual augmented Lagrangian algorithm for sparse learning. Journal of Machine Learning Research, 12, pp. 1537-1586.
2. T. Suzuki and R. Tomioka (2011) SpicyMKL: A Fast Algorithm for Multiple Kernel Learning with Thousands of Kernels. Machine Learning, 85, pp. 77-108.
3. T. Takahashi, R. Tomioka, and K. Yamanishi (2012) Discovering emerging topics in social streams via link anomaly detection. IEEE Transactions on Knowledge and Data Engineering. Accepted.
4. S. Nakajima, M. Sugiyama, S. Babacan, and R. Tomioka. (2013) Global analytic solution of fully-observed variational Bayesian matrix factorization. J. Mach. Learn. Res., 14, pp. 1-37.

〔学会発表〕（計 6 件）

1. R. Tomioka, T. Suzuki, M. Sugiyama, and H. Kashima. A fast augmented Lagrangian algorithm for learning low-rank matrices. In Proc. the 27th Annual International Conference on Machine Learning (ICML2010), Omnipress, 2010.
2. T. Takahashi, R. Tomioka, and K. Yamanishi. (2011) Discovering emerging topics in social streams via link anomaly detection. In Proc. of the 11th International Conference on Data Mining (ICDM2011), pp. 1230-1235.
3. R Tomioka, T. Suzuki, K. Hayashi, and

H. Kashima (2011) Statistical Performance of Convex Tensor Decomposition. In Advances in Neural Information Processing Systems 24, pp. 972-980.
4. R. Tomioka and M. Mørup (2012) A Bayesian Analysis of the Radioactive Releases of Fukushima. In JMLUR Workshop and Conference Proceedings 22 (AISTATS 2012), pp. 1243-1251.
5. F. Kiraly and R. Tomioka (2012) A combinatorial algebraic approach for the identifiability of low-rank matrix completion. In Proc. 29th International Conference on Machine Learning, pp. 967-974.
6. S. Nakajima, R. Tomioka, M. Sugiyama, and S. Babacan (2012) Perfect dimensionality recovery by variational Bayesian PCA. In Advances in Neural Information Processing Systems 25, pp. 980-988.

〔図書〕(計１件)

1. R. Tomioka, T. Suzuki, and M. Sugiyama. (2011) Augmented Lagrangian methods for learning, selecting, and combining features. In Suvrit Sra, Sebastian Nowozin, and Stephen J. Wright, editors, Optimization for Machine Learning. MIT Press.

〔その他〕

ホームページ等

http://www.ibis.t.u-tokyo.ac.jp/ryotat/

双対拡張ラグランジュ法

http://www.ibis.t.u-tokyo.ac.jp/ryotat/dal/

６．研究組織

(1)研究代表者　冨岡亮太
(Tomioka Ryota)
東京大学・大学院情報理工学系研究科・助教
研究者番号：70518282