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 5 月 7 日現在

機関番号：14401
研究種目：若手研究(B)
研究期間：2010～2012
課題番号：22700147

研究課題名（和文）　離散構造を利用した超高次元データ解析法とその応用

研究課題名（英文）　Data analysis for high-dimensional data with discrete structures and its applications

研究代表者
河原　吉伸（KAWAHARA YOSHINOBU）
大阪大学・産業科学研究所・助教
研究者番号：00514796

研究成果の概要（和文）：本研究は、極めて高次元なデータにおける組合せ的計算を伴う問題において、問題（組合せを評価する関数）が持つ離散構造を利用する事により、大域的な最適性を持つ解の効率的探索や，離散構造を事前知識とした解析を行うための計算に関する方法論の構築と、その応用における有用性の検証を目的とするものである．本研究では特に、連続関数における凸性に対応する劣モジュラ性と呼ばれる集合関数の離散構造に着目し、従来は厳密に計算する事が困難だと考えられてきた問題（NP 困難問題など）への大域的／近似的最適解の計算のためのいくつかの基盤的アルゴリズムの構築を行った．また問題に内在する離散構造を事前知識として利用する事で、効率的・高精度な組合せ的計算を実現する方法に関して研究を行った．そして、遺伝子データ解析などのいくつかの重要な応用へ適用・検証を行い、応用的な有用性についても確認を行った．

研究成果の概要（英文）： The main goal of this research is to develop methodologies for efficient searches of solutions with global optimality or for incorporating discrete structures as prior knowledge in combinatorial analyses with high-dimensional data. Moreover, we aim at its evaluation of the usefulness in applications where combinatorial computation plays a central role. We developed some fundamental algorithms to calculate global/approximate optimal solutions in problems that are known to be difficult to solve (such as, NP-hard problems). Also, we studied a framework for the calculation where we can incorporate such discrete structures as prior knowledge into data analysis for efficient and accurate combinatorial calculation. Furthermore, we applied the developed methods to several applications and investigate the usefulness of the framework.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２０１０年度	1,700,000	510,000	2,210,000
２０１１年度	700,000	210,000	910,000
２０１２年度	600,000	180,000	780,000
総　計	3,000,000	900,000	3,900,000

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、データマイニング、組合せ最適化

１．研究開始当初の背景

連続関数における凸性に対応する離散構造である劣モジュラ性は、組合せ最適化分野において、1970 年代頃から理論構築が進められてきた重要な概念である．特に近年では，機械学習・データマイニング分野などにおける多くの基本的・応用的問題の中に劣モジュラ性が現れる事が知られるようになってきており，機械学習における主要な国際会議 NIPS においても、研究開始当時頃から専門のワークショップが開催され始めた．

一方で、機械学習やデータマイニング分野では，近年その応用分野の広がりから、超高次元データにおける組合せ的計算を必要とする問題を扱う重要性が増している．このような計算のためには、近年盛んに研究が進められている HPC(ハイパフォーマンス・コンピューティング)技術などの計算機パワー自体の工場を目指す必要がある一方で、本質的な計算自体の効率化を行う事が不可欠であると言える．

このような学術的背景から、超高次元データにおける組合せ計算を伴う解析において、問題が持つ（劣モジュラ性を中心とした）離散構造を利用する事による大域的最適性解の効率的な探索や、こういった問題への離散構造の事前情報としての利用のための総合的なアルゴリズム体系の構築を行う事が、学術的観点から極めて重要であると考えた事が本研究の提案に至った最大の理由であった．また、従来高次元な場合には計算困難と考えられてきた組合せ的問題の中には，その評価関数や利用可能な問題構造が劣モジュラ性を有する場合が数多く存在する．そのような応用への開発した解析法の適用により，現実的時間内にそれらが計算可能となれば，応用的に多くの知見が得られ、多大なインパクトがあると考えられる．

２．研究の目的

本研究の目的は、大きく次の２つの課題の解決を行う事である．

(1) 離散構造を利用した高次元データ解析法の構築

劣モジュラ性を有する集合関数に関して、種々の問題形式（劣モジュラ関数最大化や離散 DC 計画問題など）の厳密解を効率的に解くためのアルゴリズム体系構築を行う．また問題に内在する事前情報を、劣モジュラ性などの離散凸性に基づき計算に取り込む事で、計算の効率化や高精度かを実現する枠組みの構築を行う．この際、従来から主に採用されてきた非組合せ的な近似手法（l1 正則化による連続最適化など）との理論的関係を明らかにし、実験的な性能の評価も実施する．

(2) 高次元データにおける組合せ的計算を伴う問題への応用

従来は組合せ的に解く事が困難と考えられてきた応用の中には，評価関数が劣モジュラ性を有する場合が多く存在する．本研究では、そのような応用を発見し、(課題 1) で得られた手法を適用して有用性の評価を行う．その際、問題特有の離散的な事前知識を用いた方法についてもその有用性を検討する．

３．研究の方法

上述のように、本研究は大きく２つの課題の解決を目的として、理論的研究と手法提案を行ってきた．そのために、(Sub-1)離散構造に基づくアルゴリズム体系の構築、(Sub-2)近似的な従来手法と提案手法との関連性の解明、(Sub-3)各応用への適用のための効率化、そして(Sub-4)専門領域研究者との応用的研究、の 4 つのサブテーマを軸に研究を進めた．主に、(Sub-1)及び(Sub-2)が(課題 1)、そして(Sub-3)及び(Sub-4)が(課題 2)を解決するためのものである．

(1) 理論的基礎の構築と検証

主に本研究の初年度と２年度目前半においては、(Sub-1)と(Sub-2)を中心とした理論的基礎に関する研究を進めた．
より具体的には、まずいくつかの応用的に重要な計算困難な集合関数最適化（劣モジュラ最大化や離散 DC 計画問題など）において、大域的な最適性を持った解の探索を行うためのアルゴリズム構築を行った．
また同じく応用的に重要な、(パラメータを持つ) 制約を持つ NP 困難な組合せ最適化問題に対して、(事前に指定しない) いくつかのパラメータに対する厳密解を多項式時間で計算するアルゴリズムに関する研究を行った．このようなアルゴリズムを、特にクラスタリングや最密グラフ問題などを含むものに関して検討を行い、人工データを用いた経験的な検証も行った．

(2) 効率化・拡張と応用への適用

２年度目後半以降は基本的に、(1)で扱うような問題に対するアルゴリズムの効率化・拡張に関する研究(Sub-3)や、組合せ計算が重要となる応用に関する研究(Sub-4)を行った．
まず(Sub-3)に関しては、特に問題に内在する離散構造を、事前情報として利用する事で効率的な計算を可能とする方法について議論を行った．
さらに、この方法や(1)で開発したアルゴリズムを、組合せ計算が重要となる高次元データ解析を伴う応用へ適用し、その経験的な有用性の検証を行った．

４．研究成果

(1) 組合せ的計算によるデータ解析アルゴリズムの構築

上述のように本研究では、いくつかの NP 困難な問題への厳密解法を開発した．まず、応用的にも重要な劣モジュラ最大化への厳密解法（切除平面法）を提案した．また一般の集合関数最適化に対して、劣モジュラ関数による分解（DC 分解．劣モジュラ関数の差への分解）を与え、これに基づく厳密解法（分枝限定法）を提案した．これらのアルゴリズムは、機械学習分野におけるトップの国際会議である NIPS（Ann. Conf. on Neural Information Processing Systems）で各々論文が採録されるなど、国際的にも高い評価を得ている．また上述のような、パラメータを持つ制約付きの組合せ的問題（特に、クラスタリングとサイズ制約付き劣モジュラ最小化（最密グラフ問題などを含む））に対して、パラメータを指定しない場合に多項式時間で大域解を与えるアルゴリズムを開発した．そして、これらのアルゴリズムは経験的にも優れた性質を持つ事を実験により確認した．これらの成果も、機械学習分野におけるトップ国際会議である ICML（Int'l Conf. on Machine Learning）や NIPS へ各々論文が採録されるなど、国際的にも高い評価を得ている．

(2) 組合せ的計算が重要な応用への適用

本研究では、劣モジュラ性を用いたアルゴリズムを、組合せ的計算が重要となる応用へ適用し、その効率性や経験的性能について検証を行った．
まずその一つとして、特徴選択を用いたポートフォリオ選択に対して適用を行った．劣モジュラ性に基づいて行った定式化を元に、同問題へ貪欲法を適用する事で図 1 のように、従来法と比べて、より少ない銘柄数で同様の性能を実現する事が可能である事を示した．また、遺伝子データ解析におけるゲノムワイド相関解析に対し、遺伝子間の構造（遺伝子配列や遺伝子間相互作用など）を、劣モジュラ性に基づき取り入れた解析方法を開発し、これを実際のデータに適用した．この方法は，従来法と比べて極めて速い計算が可能であるにも関わらず（図２参照)、従来法と同等以上の性能を実現する事が可能である事が経験的にも確認された．この成果は、バイオインフォマティクスのトップ国際会議である ISMB/ECCB（Ann. Int'l Conf. on Intelligent Systems for Molecular Biology & European Conf. on Computational Biology）にも論文が採録され、国際的にも高い評価を受けた．

また(1)と(2)に限らず、国内では本研究課題

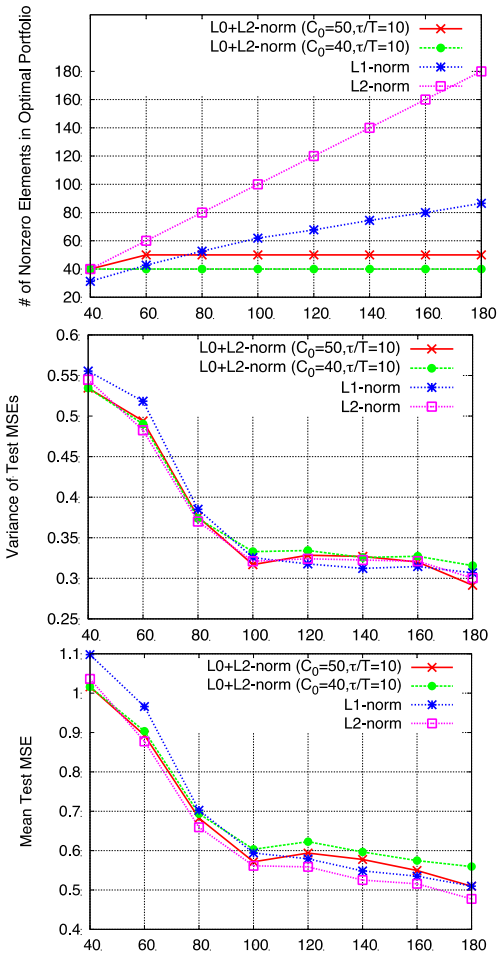

図１．提案手法(L0+L2)と従来手法(L1, L2)の選択された銘柄数(上)と、対応する分散(中)及び自乗誤差

に関連した研究を行っている研究者は基本的に研究代表者のみであるため、数多くの招待講演で話す機会や解説記事を書く機会が得られ、多くの関連する研究者との情報共有を行えた事も学術的な成果の一つであると思われる．

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６件）

① C. Azencott, D. Grimm, M. Sugiyama, Y. Kawahara and K. Borgwardt, Efficient network-guided multi-locus association mapping with graph cuts, Bioinformatics (Proc. of ISMB/ECCB'13), (査読有) (採録決定).

② A. Takeda, M. Niranjan, J. Goto and Y. Kawahara, Simultaneous pursuit of out-of-sample performance and sparsity in tracking portfolio, Computational Management Science, Vol. 10, No. 1, pp. 21-49, 2013 (査読有).
(DOI: 10.1007/s10287-012-0158-y)

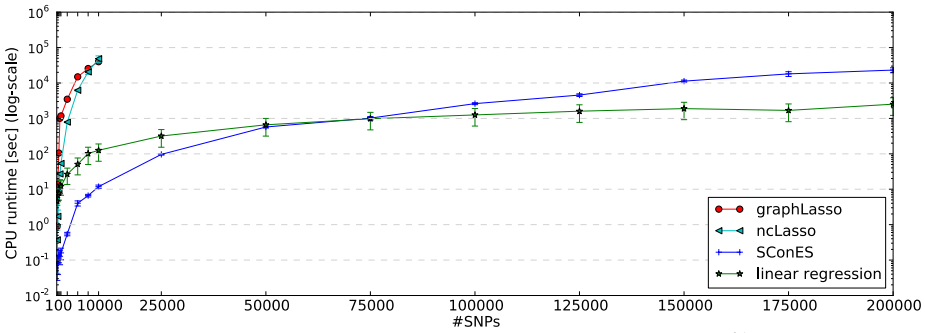

図２．提案手法（SConES）と既存手法の SNP 数に対する計算時間の比較

③ Y. Kawahara & T. Washio, Prismatic algorithm for discrete D.C. programming, Advances in Neural Information Processing Systems 24, pp. 2106-2114, 2011（査読有）.

④ K. Nagano, Y. Kawahara and K. Aihara, Size-constrained submodular minimization through minimum norm base, Proc. of the 28th Int'l Conf. on Machine Learning (ICML'11), pp. 977-984, 2011（査読有）.

⑤ Y. Kawahara, K. Nagano &Y. Okamoto, Submodular fractional programming for balanced clustering, Pattern Recognition Letters, Vol. 32, No. 2, pp. 235-243, 2011（査読有）. 10.1016/j.patrec.2010.08.008

⑥ K. Nagano, Y. Kawahara and S. Iwata, Minimum average cost clustering, Advances in Neural Information Processing Systems 23, pp. 1759-1767, 2010（査読有）.

〔学会発表〕（計 11 件）

① 杉本和正，河原吉伸，鷲尾隆，乱択アルゴリズムを用いた特徴選択，第 26 回人工知能学会全国大会，2012 年 6 月 14 日，山口.

② 河原吉伸，鷲尾隆， 離散 DC 計画のためのプリズム法とその応用，第 14 回情報論的学習理論ワークショップ(IBIS'11)，2011 年 11 月 9 日，奈良.

③ 岸本卓也，猪口明博，河原吉伸，鷲尾隆，劣モジュラ最適化に基づいたグラフ系列のクラスタリング，第 25 回人工知能学会全国大会，2011 年 6 月 1 日，盛岡.

④ Q. Liu, Y. Kawahara and T. Washio, Analyzing optimal marketing strategies over customers networks，第 25 回人工知能学会全国大会，2011 年 6 月 1 日，盛岡.

⑤ 永野清仁，河原吉伸，合原一幸，最小平均費用クラスタリング，第 13 回情報論的学習理論ワークショップ(IBIS'10)，2010 年 11 月 4 日，東京.

（招待講演）

⑥ 河原吉伸，機械学習における組合せ最適化の最近の話題：離散凸性の利用を中心として，日本オペレーションズ・リサーチ学会関西支部研究実践者交流会，2012 年 11 月 10 日，大阪.

⑦ Y. Kawahara, Challenges on combinatorial computation for large data using discrete structures, CompView Final Symposium, 2011 年 12 月 5 日，東京.

⑧ 鷲尾隆，稲積孝紀，清水昌平，鈴木譲，山本章博，河原吉伸，関数モデル上の統計的因果推論研究の現状，第 83 回人工知能基本問題研究会，2011 年 11 月 27 日，東京.

⑨ 河原吉伸，劣モジュラ性を用いた機械学習の新展開，第 23 回 RAMP シンポジウム，2011 年 10 月 24 日，大阪.

⑩ 河原吉伸，正則化による疎表現推定における劣モジュラ性の利用と最適化，圧縮センシングとその周辺(2)，2011 年 7 月 23 日，京都.

⑪ 河原吉伸，劣モジュラ性を用いたデータ生成過程の学習，第 13 回情報論的学習理論ワークショップ(IBIS'11)，2010 年 11 月 4 日，東京.

〔解説記事〕（計２件）

① 河原吉伸，機械学習における劣モジュラ性の利用と組合せ論的アルゴリズム，オペレーションズ・リサーチ，Vol. 58, No. 5, 2013（掲載予定）.

② 河原吉伸，永野清仁，鷲尾隆，劣モジュラ性を用いた知能情報処理への新展開，人工知能学会誌，Vol. 27, No. 3, pp. 252-260, 2012.

〔ホームページ〕

http://www.ar.sanken.osaka-u.ac.jp/~kawahara/jp/

６．研究組織

(1)研究代表者

河原　吉伸（KAWAHARA YOSHINOBU）

大阪大学・産業科学研究所・助教

研究者番号：00514796

(2)研究分担者

該当なし

(3)連携研究者

該当なし