研　究　報　告　書

「組合せ的計算に基づく超高次元データからの知識発見」

研究タイプ：通常型

研究期間：平成22年10月～平成26年3月

研 究 者：河原吉伸

1．　研究のねらい

近年のデータ取得技術の向上や情報処理技術応用の多様化を背景に，ますます高次元化・大規模化するデータを扱う知的解析/情報処理において，組合せ的計算に伴う計算量の爆発は，多くの場面その計算的ボトルネックとなってしまう．本研究ではこのような問題において，組合せ的な構造（特に，集合関数における凸性である劣モジュラ性）を用いたアルゴリズムを開発・適用する事で計算効率の著しい向上を計り，従来は難しかったような知的解析・情報処理への新たな枠組みの確立を目指すものである．これにより，これまでの解析や情報処理では難しかったような，新たな科学的，または社会的知見の獲得へとつながるアルゴリズム体系の実現を目的とする．

（集合関数最適化などとして定式化される）組合せ的な計算は大きく，(1)効率的に解ける問題のクラス（劣モジュラ最小化などで定式化される問題）と，(2)NP困難問題など効率的に解く事が困難な問題のクラス（劣モジュラ最大化や制約付き劣モジュラ最小化などで定式化される問題）とに分類される．本研究では，前者(1)に対しては，劣モジュラ性を用いて計算効率を著しく向上させる事で，従来より更に大規模なサイズへ適用可能なアルゴリズムの構築を行う．これにより，従来のアプローチでは難しかったようなサイズを扱う事が可能となり，新たな応用的知見や有用性の創出へとつながる事が期待される．また後者(2)に対してはまず，(a)組合せ最適化分野で知られる数理的保証のある近似的解法を実際のデータ解析へ適用する事により，汎用性・実用性を両立する効率的なアプローチを実現する．またそれとは逆に，(b)大域的な解の探索を行うための方法について，問題（組合せを評価する関数）が持つ劣モジュラ性などの組合せ的構造を利用する効率的なアルゴリズムを開発する．特に(b)については，人工知能やその他周辺分野の未解決問題への統一的なアプローチへ向けた一つの試みであり，高い学術的意義を有する課題と考えられる．

2．　研究成果

（1）概要

本研究では，上述のねらいの下，理論的解析を含めたアルゴリズム構築から，種々の人工データ・実データを用いた実験的検証までを含めた研究を遂行してきた．まず効率的に解ける問題に対する高速アルゴリズムの構築に関連して，劣モジュラ性を用いた構造正則化学習における高速アルゴリズムを開発した．この方法は汎用性が高く，遺伝子データ解析や脳画像データ解析など，多くの応用への展開へとつながった．そして劣モジュラ性に基づく近似的解法の適用としては例えば，株式ポートフォリオ選択の劣モジュラ最大化としての定式化と貪欲法の適用を行った．この応用では，従来手法と比べて，ポートフォリオに含まれる銘柄数が小

さいにもかかわらず，同等の性能を実現するポートフォリオの構成が可能である事を確認した．さらに計算困難な問題において大域的な解を探索する方法に関しては，劣モジュラ性に基づく大域最適化のアルゴリズムに加えて，いわゆる多項式時間内に大域解の一部の情報を計算可能なアルゴリズムなど，理論的にも実用的にも有用な方法論がいくつか得られた．特にこれらに関しては，機械学習分野で最も主要な国際会議でも継続して複数論文が採録されるなど，国際的にも高い評価を受けている．

またその他の事項として，本課題に関連する基礎的事項は，他の国内研究者等への理解促進のため学術雑誌における解説記事などの情報発信の活動も行った．

（2）詳細

研究テーマ A「劣モジュラ性を用いた知的情報処理のためのアルゴリズム開発」

(A-1) 劣モジュラ性を用いた高速な構造正則化学習アルゴリズム開発

データ解析や知的情報処理の場面では，往々にして用いるデータ中に利用可能な構造情報が存在する．例えば，ソーシャルネットワークのような 2 要素間の関係を持つデータを扱う場合や，変数の属性にグループ構造のようなものが存在する場合などである．構造正則化学習は，このようなデータ中の組合せ的構造を機械学習の各手法に取り組むための枠組みであるが，大規模なデータへの適用には十分な計算効率を実現するのが難しいという問題点があった．[論文 1]では，多くの構造正則化学習の問題が劣モジュラ性を用いて極めて効率的に計算可能である事を示し，これによる高速なアルゴリズムを導出した．ただし[論文 1]自体は，構造正則化中の計算よりもより広い範囲へ適用可能な最適化の枠組みを議論したものである．この枠組みは，後述のような応用においてもその有用性を確認しており，今後も様々な応用への適用展開を考えている．

(A-2) 計算困難問題における大域的解探索のための劣モジュラ性に基づくアルゴリズム開発

応用的に重要となる組合せ的計算の多くは，上述の構造正則化のように効率的に解ける問題のクラスに入らない．このような問題に対しては一般に，多項式時間で計算可能な近似アルゴリズムを適用するなどして，近似解の探索を行う事が多い．機械学習分野などでは，連続最適化問題へと緩和して近似的な計算を行う事もしばしばである．しかし応用によっては，近似的な解ではなく，大域的な解の探索を行う事が本質的に重要なケースも多い．理論的にはNP困難な問題は，変数の数の増加と共に計算時間が指数関数的に増加する．しかし実用的には計算が可能な分枝限定法などのアルゴリズムが知られている．本研究では，一般の集合関数最適化を表現可能な離散DC計画問題に対して，劣モジュラ性に基づく分枝限定法を導出し，ある程度のサイズまで大域解の計算が可能である事を示した([論文4])．また問題に何らかのパラメータが存在する際に(制約付きの劣モジュラ最小化など)，そのパラメータを固定しなければ，いくつかのパラメータに対する大域解が多項式時間で計算可能なアルゴリズムに関しても導出し，実用的には極めて高速に，かつ多くのパラメータに対する解が得られる事を確認している．

研究テーマ B「開発したアルゴリズムの応用への適用」

上述の研究テーマ A で開発したアルゴリズムの一部は，実用的な応用に対しても適用し，その有用性の検証を行った．多くは(A-1)で述べた構造正則化学習であるが，1.研究のねらいで

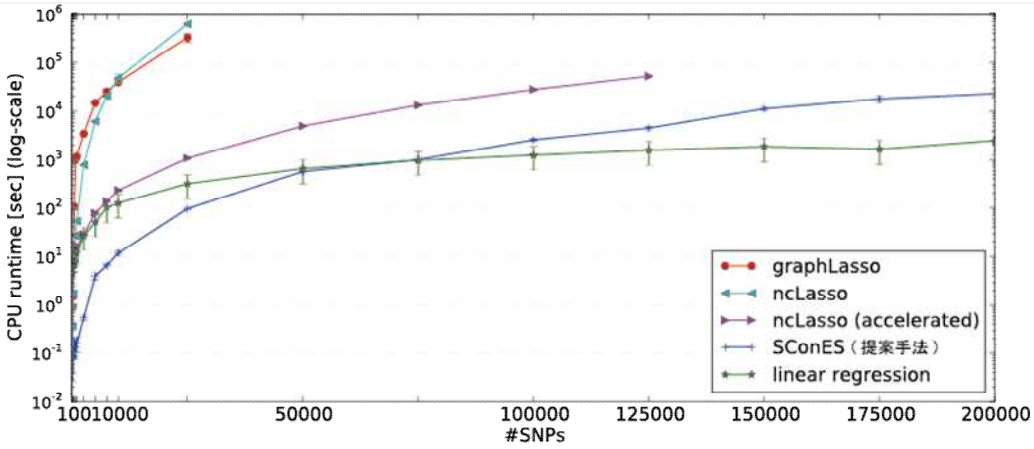

図 2 提案手法と従来手法の計算時間比較の例（青色の SConES が提案手法．緑色の線はベースラインの線形回帰によるもの）．

述べた 2(a)のような，組合せ最適化分野で知られる近似解法の適用についても，例えば株式ポートフォリオ選択などでその有用性を確認した．

まず[論文 2]では，遺伝子データ解析において重要な問題であるゲノムワイド相関解析で，遺伝子塩基間でのネットワーク上の構造(例えば塩基配列の順序や，遺伝子間相互作用)を構造正則化として用いる事で，極めて高速で性能の高い解析アルゴリズムが得られた．図 1 は計算時間を比較した例であり，数十万変数程度でも数分で計算が可能である．またその他には，MRI 画像を用いたアルツハイマー病の症状判定や，コンピュータ・ビジョンなどへ対しても，上記の計算を用いた構造正則化学習を適用しその有用性を確認している．また，株式ポートフォリオ選択を劣モジュラ最大化として定式化し，貪欲法を適用する事で，図 2 に示すように，少ない銘柄数により，他の手法と同等の性能を持つポートフォリオを実現できる事を確認した．

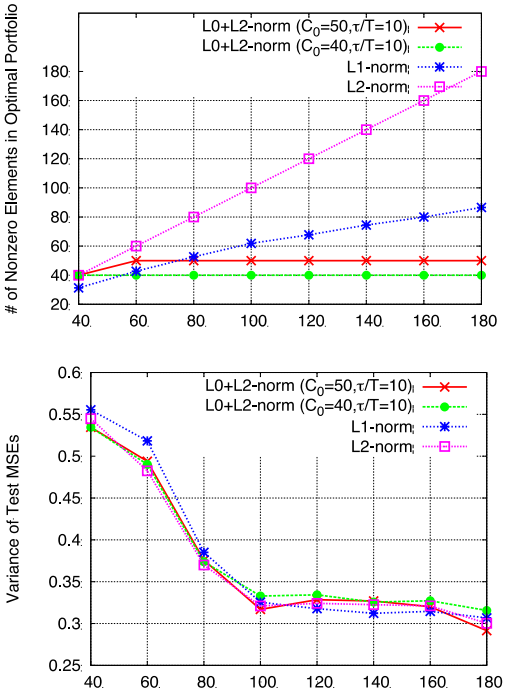

図 1 選択された銘柄数(上)と分散リスク(下)．各図とも横軸がベースの銘柄数．又赤色・緑色の線が提案法．

またその他にも，クラスタリングなどの基本的問題に対しても，劣モジュラ性などの組合せ的構造を用いる事で有用なアルゴリズムの構築が可能である事を確認している([論文 5]など)．

3. 今後の展開

本研究の成果は，上述の研究テーマ B のように応用性の高いものもあるが，アルゴリズム構築やその理論的解析といった基礎的成果が主である．今後はここで得られた理論的基盤をベースに，応用的に重要となる，または周辺情報分野の基礎問題などに必要となる理論的計

算基盤やアルゴリズム構築のための枠組み等の研究を進めると共に，更なる応用へも展開を進めていく予定である．特に本研究で議論した(A-2)の枠組みは，人工知能における論理推論や統計物理問題におけるスピングラス問題のように，情報分野で未解決な計算困難問題に対する一つのアプローチの基礎としても重要であると考えており，これらとの関係や，実用的な方法論の構築などに取り組む予定である．

4．　評価

（1）自己評価

研究開始当初目標としていた内容は，上記の(A-2)とその応用が中心であった．研究が進むにあたり，その他のテーマへと研究が広がり，当初は想定していなかったような成果が得られるようになってきた．特に，当初は想定していなかった(A-1)における理論的成果とそれに伴う高速アルゴリズムが得られ，(B)に述べたような，様々な応用への展開へとつながった．当初から想定していた(A-2)に関しては，研究提案時に想定していたような実用性を持つレベルまでの高速なアルゴリズムの構築には至らなかったが，方法論としての枠組みとしてや，理論的にも将来の研究につながる成果が得られたと考えている．これらの成果をベースに，学術的のみならず，周辺の情報分野や種々の応用へとつながる構想もあり，また現時点の成果に対しても学術的に高い評価も得られ，今後につながる研究成果が得られたと考えている．

（2）研究総括評価（本研究課題について、研究期間中に実施された、年２回の領域会議での評価フィードバックを踏まえつつ、以下の通り、事後評価を行った）。

劣モジュラ性を利用して大域的な最適性を持った組合せ(厳密な最適解)を見つける効率的なアルゴリズムを構築するという研究である。

効率的に解ける問題に対する高速アルゴリズムの構築では、劣モジュラ性を用いた構造正則化学習における高速アルゴリズムを開発している。これは、遺伝子データ解析や脳画像データ解析など、多くの応用への展開へとつながっている。また、劣モジュラ性に基づく近似的解法を株式ポートフォリオ選択に適用し、ポートフォリオに含まれる銘柄数が従来よりも小さいにもかかわらず、同等の性能を実現するポートフォリオの構成が可能であることを確認している。計算困難な問題において大域的な解を探索する方法に関しては、大域最適化のアルゴリズムに加え、多項式時間内に大域解の一部の情報を計算可能なアルゴリズムなど、理論的にも実用的にも有用な方法論がいくつか得られている。特にこれに関しては、国際的にも高い評価を受けており、高く評価したい。

5．　主な研究成果リスト

（1）論文（原著論文）発表

1. K Nagano and Y. Kawahara, “Structured convex optimization under submodular constraints,” Proc. of the 29th Ann. Conf. on Uncertainty in Artificial Intelligence (UAI'13), 459-468, 2013
2. C. Azencott, D. Grimm, M. Sugiyama, Y. Kawahara and K. Borgwardt, “Efficient network-guided multi-locus association mapping with graph cuts,” Bioinformatics, Vol.29, No.13, pp.i171-i179, 2013
3. 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, 21-49, 2013.
4. Y. Kawahara and T. Washio, "Prismatic algorithm for discrete D.C. programming problem," Advances in Neural Information Processing Systems, Vol.24, 2106-2114, 2011.
5. Y. Kawahara, K. Nagano and Y. Okamoto, "Submodular fractional programming for balanced clustering," Pattern Recognition Letters, Vol.32, No.2, 235-243, 2011.

（2）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

1. 河原吉伸, "構造的な事前情報を用いた機械学習： 構造正則化と劣モジュラ性," 情報処理, Vol.52, No.7, 734-740, 2013 (解説記事).
2. 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, 267-274, 2013 (解説記事).
3. 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, 252-260, 2012 (解説記事).