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

令和　元 年　 ６ 月　 ７ 日現在

機関番号：１４５０１
研究種目：若手研究(B)
研究期間：2015～2018
課題番号：１５Ｋ１６０６４

研究課題名（和文）変数選択結果の対数線形モデルを用いた再解析による安定性向上と多重変数集合の抽出

研究課題名（英文）A log-linear modeling of unstable variable selection: improving stability and extracting multiple sets of variables

研究代表者
北園　淳（Kitazono, Jun）

神戸大学・工学研究科・工学研究科研究員

研究者番号：００７３３６７７

交付決定額（研究期間全体）：（直接経費）　　3,100,000円

研究成果の概要（和文）：多くの変数の中から，データの分類などに有用となる少数の変数を取り出す手法を，変数選択と呼ぶ．この変数選択において，データサンプルが少し変わっただけで選択される変数が大きく変わってしまうという，安定性の問題が指摘されている．本研究では，この変数選択の安定性の問題の解決を目的とした．変数選択によって得られたばらついた結果の中から，真に有用な変数を定める枠組の構築に取り組んだ．

研究成果の学術的意義や社会的意義
変数選択は，医学・工学・理学等多くの場面で使用されている．例えば，遺伝子解析では，変数選択によって，特定の病気に将来なりやすいかどうか識別するのに有用な遺伝子を選び出すことが可能である．変数選択の安定性の問題を解消し，信頼性を高めることにより，より確度の高い病気リスク診断などへの発展が期待される．また，本研究の過程で開発した可視化やクラスタリングの手法は，変数選択に限らず，多くのデータ分析に適用可能なものもあり，幅広い分野での応用が期待される．

研究成果の概要（英文）：Variable selection is a process of selecting useful variables for pattern recognition tasks, such as classification and regression. Recent studies showed that variable selection is often unstable: even a small change in data results in huge difference of selected variables. This study aimed at resolving this instability of variable selection. We developed a framework for identifying truly useful variables by finding regularity among varied combinations of selected variables.

研究分野：知能情報学、機械学習、理論神経科学

キーワード：変数選択　安定性　対数線形モデル　可視化　クラスタリング

様　式　C－１９、F－１９－１、Z－１９、CK－１９（共通）

１．研究開始当初の背景

近年の計測技術・計算機技術の発展に伴い，得られるデータの高次元化が進んでいる．しかしながら，高次元データそのままでは，データの解釈やその後の解析が困難な場合が多い．そこで，高次元データの中から，重要な少数の変数を取り出す変数選択と呼ばれる手法の開発が進んでいる．この変数選択によって例えば，DNA マイクロアレイのデータから，特定の病気に将来なりやすいかどうか識別するのに有用な遺伝子を選択することが可能である．この変数選択は，遺伝子解析に限らず，工学・理学・医学等多くの場面で使用されている．

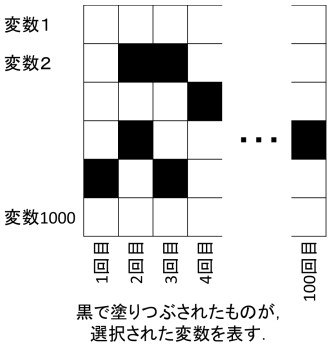

図１：変数選択の安定性

しかしながら，この変数選択において，データサンプルが少し変わっただけで選択される変数が大きく変わってしまうという，変数選択の安定性の問題が指摘されている（Ein-Dor et al.，2005; Kalousis et al.，2007；プリチャード & 江口，2009）．例えば図１のように，手元にある 100 サンプルのうち，ある１サンプルを除いた 99 サンプルを用いて変数選択を行うことを考える．このとき，どの１サンプルを除くかによって，選択される変数が大きく変わってしまうのである．これは，既存の変数選択の結果が必ずしも信用できず，脆弱であることを意味している．

変数選択が不安定になる要因としては，まず，データが高次元で候補となる変数の組が大量に存在するため，A：本来有用でない変数が偶然有用に見えてしまうというものが挙げられる．図１の例でいえば「各回に選ばれている変数の中には実際には有用でないものが紛れ込んでおり，データのばらつきに依存して偶然どの変数が選ばれるかが毎回変わる」という状況である．別の要因としては，B：有用な変数の組が多重に存在するというものが挙げられる．図１の例でいえば「各回に選ばれている変数の組それぞれが有用であり，データの微妙な差異によってどの組が選ばれるかが毎回変わる」という状況である．もちろんこれら二つの要因両方によって，変数選択結果が不安定になっている状況も考えられる．

２．研究の目的

本研究では，上述の変数選択の安定性の問題の解決を目的とした．具体的には，変数選択の結果得られたばらついた結果をさらに解析し，真に有用な変数を定める枠組みを構築することを目的とした．

３．研究の方法

（１）網羅的な変数の選択・有用性の定量的な評価

変数選択のばらつきを定量的に評価する枠組みの構築に取り組んだ．マルコフ連鎖モンテカルロ法の一つである交換モンテカルロ法（Hukushima & Nemoto，1996）を用いて，識別精度の高くなる変数の組み合わせを網羅的に選び出す手法の開発に取り組んだ．さらに，それらの変数の組が実際に有用かどうか，Null 分布と比較することで判定する方法の開発を行った．

（２）可視化

ばらついた変数選択の結果はそのままでは解釈が難しいため，何らかの方法で情報を抽出することが重要である．そこで，変数同士の関係性，変数とサンプルの関係性を可視化・分析するための手法開発を行った．

（3）クラスタリング

ばらついた変数選択の結果を解釈するには，クラスタリングが有効であると考えられる．しかしながら，通常，変数選択の適用先は高次元であるため，スペクトラルクラスタリングなどの既存の手法ではあまり良い性能が期待できない場合がある．そこで，次元圧縮手法である t-SNE を前処理に用いてクラスタリングする手法の開発・検証を行った．

（4）低次の対数線形モデルによる検証・推定の高速化

対数線形モデルを用いて，従来の変数選択手法におけるばらつきを確率分布として表現することで，真に有用な変数を取り出すことが可能になると考えらえる．そこで，対数線形モデルの中でも低次のモデルを用いた検証を行った．また，変数選択の結果は，多くの場合，高次元小サンプルになる．そこで，そのような場合に有効でかつ高速な推定手法，Minimum Probability Flow（Sohl-Dickstein et al., 2011）について検証を行った．

4．研究成果

（1）網羅的な変数の選択・有用性の定量的な評価

交換モンテカルロ法を用いることで、識別精度の高くなる変数を網羅的に選び出す手法の開発に成功した（雑誌論文④）．提案法では，ある変数の組を用いた場合の交差検証誤差をエネルギー関数とするボルツマン分布を考える．このボルツマン分布からサンプリングすることで，エネルギー関数値が低い，すなわち，誤差が小さくなる変数の組を効率的に探索することができる．また，マルコフ連鎖モンテカルロ法の中でも，特に交換モンテカルロ法を用いることの利点は，主に以下の二点である．

1．交換モンテカルロ法では，局所最小に囚われずに，効率的にサンプリングを行うことができる．これによって，有用な変数の組を一部だけでなく，網羅的に探し出すことが可能となる．またこの特徴は，特に，有用な変数の組が多く存在し，しかもそれらの組が互いに全くことなるような場合に特に重要になると考えられる．

2．交換モンテカルロ法のサンプリング結果に，多重ヒストグラム法（Ferrenberg & Swendsen, 1989）を適用することで，エネルギー密度，つまり，誤差のヒストグラムを得ることができる．これにより，ある誤差を与える組合せがどの程度あるのか，知ることが可能になる．また，このヒストグラムを Null 分布（ランダムに識別した場合）と比較することで，対象とする変数集合が真に有用なのかどうか判定することが可能になる．

（2）可視化

高次元データを可視化する手法の開発を行った（雑誌論文③，学会発表③）．提案法は，可視化のための次元圧縮手法として知られる t-distributed stochastic neighbor embedding (t-SNE, van der Maaten & Hinton, 2008)をベースとする．提案法では，データ点の実効的な次元が，データ点ごとに異なる，つまり，非一様性を考慮することで，高次元構造と低次元構造が混じったデータにおいても，良い可視化性能が得られる．提案法をばらついた変数選択結果に適用することで，変数同士の関係性を可視化することが可能となる．

また，変数選択同士だけでなく，サンプルと変数との関係性も併せて可視化するために，変数とサンプルを同じ空間上にマップすることが可能な可視化手法を開発した（雑誌論文②，学会発表②）．

（3）クラスタリング

従来可視化に用いられていた次元圧縮手法である t-SNE をベースに，新たなクラスタリング手法を開発した（雑誌論文①，学会発表①）．これにより，局所的な構造にもとづき（埋め込まれた多様体構造に沿って），選択された変数の組同士の類似度，および，どのサンプルの識別にどの変数が有効かといった関係性を，可視化のみならず，グループ分けという形で，明らかにすることが出来ると期待される．

（4）低次の対数線形モデルによる検証・推定の高速化

低次の対数線形モデルを用いた解析について検証を行った．また，高次元小サンプルの設定における推定に重要となる Minimum Probability Flow (MPF)法について，その性能の検証を行った．MPF 法のテストベッドとして，MPF 法を適用することが比較的容易な次元圧縮法の一つである t-SNE を用い，MPF 法によって，実際に高速に推定が可能となることを確認した（学会発表⑤）．また，MPF 法の利点・欠点も精査した．

＜引用文献＞

Ein-Dor, L., Kela, I., Getz, G., Givol, D., and Domany, E. (2005) Bioinformatics, 21(2), 171178.

Ferrenberg, A. M., and Swendsen, R. H. (1989). Computers in Physics, 3(5), 101-104.

Hukushima, K. and Nemoto, K. (1996), J. Phys. Soc. Jpn., 65, 1604-1608.

Kalousis, A., Prados, J. and Hilario, M. (2007) Knowl. Inf. Syst., 12(1), 95-116.

Maaten, L. V. D and Hinton, G. (2008), J. Machine Learning Research, 9, 2579-2605.

Sohl-Dickstein, J., Battaglino, P. B., and DeWeese, M. R. (2011). Physical review letters, 107(22), 220601.

プリチャード真理，江口真透．(2009) 日本統計学会誌，38(2)，199-212.

５．主な発表論文等

〔雑誌論文〕（計 4 件）

① N. Rogovschi, J. Kitazono, N. Grozavu, T. Omori and S. Ozawa. t-Distributed stochastic neighbor embedding spectral clustering. Proceedings of 2017 International Joint Conference on Neural Networks，査読有，2017，pp. 1628–1632. DOI: 10.1109/IJCNN.2017.7966046

② N. Murata, J. Kitazono, S. Ozawa. Multidimensional Unfolding Based on Stochastic Neighbor Relationship. Proceedings of the 9th International Conference on Machine Learning and Computing，査読有，2017，pp. 248–252.

③ J. Kitazono, N. Grozavu, N. Rogovschi, T. Omori, S. Ozawa. t-Distributed Stochastic Neighbor Embedding with Inhomogeneous Degrees of Freedom. Doya K., Ikeda K., Lee M., Liu D. (eds) Neural Information Processing. ICONIP 2016. Lecture Notes in

Computer Science. 査読有, 2016, Vol. 9949, pp. 119 – 128. DOI: 10.1007/978-3-319-46675-0_14

④ K. Nagata, J. Kitazono, S. Nakajima, S. Eifuku, R. Tamura, M. Okada. An exhaustive search and stability of sparse estimation for feature selection problem. IPSJ Transactions on Mathematical Modeling and Its Applications, 査読有, Vol. 8, 2015, pp. 23–30. DOI: 10.2197/ipsjtrans.8.25

〔学会発表〕(計 5 件)

① N. Rogovschi, J. Kitazono, N. Grozavu, T. Omori and S. Ozawa. t-Distributed Stochastic Neighbor Embedding Spectral Clustering. 2017 International Joint Conference on Neural Networks, 2017.

② N. Murata, J. Kitazono, S. Ozawa. Multidimensional Unfolding Based on Stochastic Neighbor Relationship. The 9th International Conference on Machine Learning and Computing, 2017.

③ J. Kitazono, N. Grozavu, N. Rogovschi, T. Omori, S. Ozawa. t-Distributed Stochastic Neighbor Embedding with Inhomogeneous Degrees of Freedom. The 23rd International Conference on Neural Information Processing, 2016.

④ J. Kitazono. Machine Learning for Making Data Understandable. Cyber-Physical System for Smarter World 2016, 2016.

⑤ 北園淳, 大森敏明, 小澤誠一. Minimum Probability Flow による t 分布型確率的近傍埋め込み法の高速化. 第 59 会システム制御情報学会研究発表講演会, 2015.

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属されます。