

令和 6 年 6 月 11 日現在

機関番号：24405

研究種目：若手研究

研究期間：2020～2023

課題番号：20K20247

研究課題名（和文）機械学習による自動的変数選択をもちいた新しい治療効果推定手法の提案

研究課題名（英文）Variable selection for causal inference approach using machine learning methods

研究代表者

加葉田 大志朗 (Kabata, Daijiro)

大阪公立大学・大学院医学研究科・特任助教

研究者番号：40793435

交付決定額（研究期間全体）：（直接経費） 2,600,000円

研究成果の概要（和文）：本課題はDouble/debiased Machine Learning (DML) 推定量における機械学習による局外関数推定時の変数選択に焦点を置いて、(1) 結果変数適応型の機械学習手法を用いたDML推定量の提案と理論的妥当性の評価、(2) 広範な状況下における従来法との結果の際の評価、(3) 臨床研究データへの応用を目的として取り組んだ。上記成果をまとめた論文を4編執筆し、それぞれ査読付き国際学術誌に採録された。

研究成果の学術的意義や社会的意義

既存のデータベースなどを用いた研究は倫理的な制約がなく、研究対象者が被り得るリスクが小さいというメリットがある。一方で、分析結果の妥当性や再現性に懸念が残ることも多く、より適切な解析の実施が求められる。特に膨大な情報を扱うデータベース研究では、どのような特性情報を利用するかによって、その結果は大きく変化し得る。

本課題では膨大な特性情報の中から、適切な治療効果の推定に役立つ情報の選定方法を提案することによって、上記の課題の克服への寄与を目指した。本課題で得られた成果は、データベースを用いた解析の妥当性や再現性を向上させ、データベースを利活用したエビデンス構築に寄与するものであると考える。

研究成果の概要（英文）：This project focused on variable selection during nuisance function estimation with machine learning in Double/debiased Machine Learning (DML) estimators. It involved: (1) proposing a DML estimator using outcome-adaptive machine learning methods and assessing its theoretical validity, (2) evaluating the results under various conditions compared to typical methods, and (3) applying the proposed estimator to the clinical research data. We compiled the results into four papers, each of which was accepted by peer-reviewed international scientific journals.

研究分野：生物統計

キーワード：因果推論 機械学習 観察研究 データベース

様式 C - 19、F - 19 - 1 (共通)

1. 研究開始当初の背景

近年非実験的に集積されたデータベースの利活用が関心を寄せている。研究を目的にせず集積されたデータベースから知見を見出す際にはあらゆる妥当性の評価が必要となる。特に関心の処置効果の推定を目的とする場合には、比較群間で均質でなく、かつ関心の結果変数に関連する因子の存在下では、処置効果推定値に交絡によるバイアスが生じうる。交絡の影響を低減させるための手法論として、近年では因果推論手法に関する研究が活発に行われている。

特にデータベースのように高次元な情報の存在下では、膨大な数の因子が、交絡を生じさせる候補因子として分析において考慮することができる。この膨大な候補因子の特徴を分析に反映させるために、近年では機械学習手法を用いた分析手法も利用されるようになった。特に因果推論手法においては、局外関数の推定に機械学習手法を応用することで、従来の因果推論の枠組みを残したまま、機械学習手法を用いた高次元データの活用が可能となる (Setoguchi et al. 2008; Lee, Lessler, and Stuart 2010)。ただし因果推論において局外関数の推定を機械学習手法を用いて行うことで、正則化や過適合によるバイアスが処置効果推定値に生じることも知られている。この課題を克服するために Chernozhukov et al. (Chernozhukov et al. 2018) は、機械学習手法によって推定した処置変数と結果変数の局外関数を用いた推定式を用いると同時に、局外関数の推定に cross-fitting を応用した Double/debiased Machine Learning 推定量 (以下、DML 推定量) を提案している。

一方、DML 推定量を用いた場合にも、局外関数の推定に用いる機械学習手法が異なれば、えられる結果に差異が生じうる。特に、局外関数推定時に交絡を生じさせうる候補因子が十分な重みをつけて考慮されているかどうかは、その局外関数推定値に基づいて算出される処置効果推定値の妥当性に大きく関わる。特に処置変数に関する局外関数を推定する際に機械学習を用いた場合、データ内容によっては過度に処置変数にのみ関連する因子の重みが強く反映された結果を導く恐れがあり、処置効果推定値のバイアスを十分に低減できない恐れがある。この課題を改善するために、Shortreed and Ertefaie (Shortreed and Ertefaie 2017) は、結果変数適応型 Lasso 推定量を提案し、より結果変数に関連の強い因子を優先的に重みづけた局外関数推定を提案している。この点において、従来の DML 推定量は処置変数に関する局外関数推定においても処置変数を良く説明する機械学習手法の利用を前提としているが、結果変数適応型の機械学習手法のような異なる特性を持つ機械学習手法を応用することの是非については検討の余地がある。

2. 研究の目的

本課題は DML 推定量における機械学習による局外関数推定時の変数選択に焦点を置いて、(1) 結果変数適応型の機械学習手法を用いた DML 推定量の提案と理論的妥当性の評価 (2) 広範な状況下における従来法との結果の際の評価 (3) 臨床研究データへの応用を目的とした。

3. 研究の方法

上記目的を達成するために、DML 推定量の局外関数推定に結果適応型 Lasso を応用することの漸近的性質を確認し、複数のデータ生成過程に基づく性能評価をモンテカルロシミュレーションによって実施した。特にデータベース研究でよくみられる状況、特に処置変数あるいは結果変数のいずれかのみに関係する因子の存在下を想定したデータ生成過程を用いた。さらに、局外関数の推定のために複数の機械学習手法が候補となる場合、それらを同時に組み込み、いずれかの候補が適当な変数選択ができていない場合に適切な推定値が得られるよう拡張した DML 推定量を提案し、同様にモンテカルロシミュレーションによってその性能を評価した。

また臨床で非実験目的に集積されたデータベースに、提案手法を適用すると同時に、他手法との結果の差異を評価した。本課題では大学病院に搬送された脳卒中症例を対象として、血液抗凝固治療の処方歴を処置変数とし、搬送時の脳内出血量との関連性を評価した。

4. 研究成果

目的 (1) について、処置変数に関する局外関数推定において結果変数適応型 Lasso を用いた場合にも、DML 推定量は不偏性を持つことを示すことができた。目的 (2) について、膨大な交絡因子の存在下では DML 推定量は従来の機械学習を応用した因果推論手法よりも優れた性能を持つこと、そして処置変数に関連の強い因子の存在下では従来の DML 推定量よりも結果適応型 Lasso を組み入れた DML 推定量でより良い性能が確認された。これらの結果は *Communications in Statistics: Simulation and Computation* に採録された (Kabata and Shintani 2021)。

また複数の機械学習手法を同時に考慮した DML 推定量を提案するために、層別化アプローチに基づく DML 推定量を提案した。提案法では結果変数に関する局外関数の特定化の誤りがある場合でも、処置変数に関する局外関数がいずれかの機械学習手法による推定値によって適切に特定できていれば、バイアスの小さい推定値が得られることを示した。これらの結果は *Communications in Statistics: Simulation and Computation* に採録された (Kabata and Shintani 2023)。

目的(3)について、脳卒中発症症例における血液抗凝固治療の処方歴の影響評価を行った観察研究において、従来法による分析に加えて、目的(1)および(2)で提案した結果適応型機械学習手法を応用したDML推定量を用いた分析を行った。本観察研究においては、提案法を用いた場合にも従来法と類似した結果が得られた。これらの結果は *Journal of Clinical Neuroscience* に採録された (Gon, Kabata, and Mochizuki 2022b, 2022a)。

<引用文献>

- Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. "Double/Debiased Machine Learning for Treatment and Structural Parameters." *The Econometrics Journal* 21 (1): C1-68.
- Gon, Yasufumi, Daijiro Kabata, and Hideki Mochizuki. 2022a. "Response to Correspondence Concerning 'Association between Kidney Function and Intracerebral Hematoma Volume.'" *Journal of Clinical Neuroscience*. <https://doi.org/10.1016/j.jocn.2022.02.025>.
- Gon, Yasufumi, Daijiro Kabata, and Hideki Mochizuki. 2022b. "Association between Kidney Function and Intracerebral Hematoma Volume." *Journal of Clinical Neuroscience: Official Journal of the Neurosurgical Society of Australasia* 96 (February): 101-6.
- Kabata, Daijiro, and Mototsugu Shintani. 2021. "Variable Selection in Double/Debiased Machine Learning for Causal Inference: An Outcome-Adaptive Approach." *Communications in Statistics: Simulation and Computation*, November, 1-14.
- Kabata, Daijiro, and Mototsugu Shintani. 2023. "On Propensity Score Misspecification in Double/Debiased Machine Learning for Causal Inference: Ensemble and Stratified Approaches." *Communications in Statistics - Simulation and Computation*, October, 1-11.
- Lee, Brian K., Justin Lessler, and Elizabeth A. Stuart. 2010. "Improving Propensity Score Weighting Using Machine Learning." *Statistics in Medicine* 29 (3): 337-46.
- Setoguchi, Soko, Sebastian Schneeweiss, M. Alan Brookhart, Robert J. Glynn, and E. Francis Cook. 2008. "Evaluating Uses of Data Mining Techniques in Propensity Score Estimation: A Simulation Study." *Pharmacoepidemiology and Drug Safety* 17 (6). <https://doi.org/10.1002/pds.1555>.
- Shortreed, Susan M., and Ashkan Ertefaie. 2017. "Outcome-Adaptive Lasso: Variable Selection for Causal Inference." *Biometrics* 73 (4). <https://doi.org/10.1111/biom.12679>.

5. 主な発表論文等

〔雑誌論文〕 計4件（うち査読付論文 4件/うち国際共著 0件/うちオープンアクセス 0件）

| | |
|--|--------------------------|
| 1. 著者名 Kabata Daijiro, Shintani Mototsugu | 4. 巻 Published online |
| 2. 論文標題 Variable selection in double/debiased machine learning for causal inference: an outcome-adaptive approach | 5. 発行年 2021年 |
| 3. 雑誌名 Communications in Statistics - Simulation and Computation | 6. 最初と最後の頁 1-14 |
| 掲載論文のDOI (デジタルオブジェクト識別子) 10.1080/03610918.2021.2001655 | 査読の有無 有 |
| オープンアクセス オープンアクセスではない、又はオープンアクセスが困難 | 国際共著 - |

| | |
|--|--------------------------|
| 1. 著者名 Kabata Daijiro, Shintani Mototsugu | 4. 巻 Published online |
| 2. 論文標題 On propensity score misspecification in double/debiased machine learning for causal inference: ensemble and stratified approaches | 5. 発行年 2023年 |
| 3. 雑誌名 Communications in Statistics - Simulation and Computation | 6. 最初と最後の頁 1~11 |
| 掲載論文のDOI (デジタルオブジェクト識別子) 10.1080/03610918.2023.2279022 | 査読の有無 有 |
| オープンアクセス オープンアクセスではない、又はオープンアクセスが困難 | 国際共著 - |

| | |
|--|-----------------------|
| 1. 著者名 Gon Yasufumi, Kabata Daijiro, Mochizuki Hideki | 4. 巻 96 |
| 2. 論文標題 Association between kidney function and intracerebral hematoma volume | 5. 発行年 2022年 |
| 3. 雑誌名 Journal of Clinical Neuroscience | 6. 最初と最後の頁 101~106 |
| 掲載論文のDOI (デジタルオブジェクト識別子) 10.1016/j.jocn.2021.12.022 | 査読の有無 有 |
| オープンアクセス オープンアクセスではない、又はオープンアクセスが困難 | 国際共著 - |

| | |
|--|-----------------------|
| 1. 著者名 Gon Yasufumi, Kabata Daijiro, Mochizuki Hideki | 4. 巻 99 |
| 2. 論文標題 Response to correspondence concerning " Association between kidney function and intracerebral hematoma volume " | 5. 発行年 2022年 |
| 3. 雑誌名 Journal of Clinical Neuroscience | 6. 最初と最後の頁 390~391 |
| 掲載論文のDOI (デジタルオブジェクト識別子) 10.1016/j.jocn.2022.02.025 | 査読の有無 有 |
| オープンアクセス オープンアクセスではない、又はオープンアクセスが困難 | 国際共著 - |

〔学会発表〕 計1件（うち招待講演 0件 / うち国際学会 0件）

| |
|--|
| 1. 発表者名 加葉田大志朗, 新谷元嗣 |
| 2. 発表標題 Variable selection in double/debiased machine learning for causal inference: An outcome-adaptive Lasso approach |
| 3. 学会等名 2019年度 関西計量経済学研究会 |
| 4. 発表年 2020年 |

〔図書〕 計0件

〔産業財産権〕

〔その他〕

-

6. 研究組織

| | 氏名 (ローマ字氏名) (研究者番号) | 所属研究機関・部局・職 (機関番号) | 備考 |
|-------|-----------------------------------|-----------------------|----|
| 研究協力者 | 新谷 元嗣 (Shintani Mototsugu) | | |
| 研究協力者 | 権 泰史 (Gon Yasufumi) | | |

7. 科研費を使用して開催した国際研究集会

〔国際研究集会〕 計0件

8. 本研究に関連して実施した国際共同研究の実施状況

| | |
|---------|---------|
| 共同研究相手国 | 相手方研究機関 |
|---------|---------|