

論文要旨

共変量に基づくマッチング手法への ランダムフォレストの応用と実データ解析

生物統計情報学コース

49-236605

三島 遼

【緒言】 日常診療では、患者の病態などを考慮して治療が選択されるため、治療を受けた集団の比較が治療効果の妥当な比較にならない。交絡と呼ばれるこの問題に対処するため、Propensity Score Matching (PSM) が広く用いられている。PSM は傾向スコアの近さに基づいてマッチングを行う方法である。傾向スコアの近さは共変量の類似性に直結しないため、バイアスの増大や精度の低下を招く場合があると指摘されている。これに対して、Coarsened Exact Matching (CEM) は、カテゴリ変換した共変量に基づいて厳密なマッチングを行う方法である。CEM による交絡調整の成否は共変量のカテゴリ変換の適切さに依存するが、臨床的な事前知識のみに基づいてカテゴリを決定することは困難である。

【目的】 本論文の目的は、CEM のマッチング条件として参照する共変量およびそのカテゴリ変換を Random Forest を用いてデータ駆動的に決定するという手法を提案することである。

【方法】 提案法は次の手続きで構成される：まず、傾向スコアの推定値を目的変数とした Random Forest を構築する。次に、Random Forest を構成する分類木の中から代表的な 1 つを選択する。最後に、得られた分類構造を CEM のマッチング条件として応用する。提案法の性能を評価するため、複数のシナリオでシミュレーションを行った。また、実データへの適用例として、血糖降下薬であるビグアナイドと SGLT2 阻害薬を比較する解析を行った。

【結果】 シミュレーションにおいて、CEM が解析対象者数を大幅に減らした一方で、提案法はそれを回避しつつバイアスを低減した。とくに、傾向スコアとリスクスコアの相関が 0.7 程度の状況において、推定値の信頼区間が PSM を用いるより約 10% 狭い傾向にあった。実データ解析では、血糖降下薬の処方経験が無い 2 型糖尿病患者について、初回処方後 3 ヶ月時点での HbA1c[%] の変化量 [ポイント] に大きな差は認められなかった (ビグアナイド -1.33、SGLT2 阻害薬 -0.70、提案法により調整した群間差および 95% 信頼区間 -0.132 (-0.364, 0.100))。

【結論】 Random Forest を用いて、CEM のマッチング条件をデータ駆動的に決定できることが明らかとなった。提案法によって、従来の CEM に比べて解析対象者数を維持しつつ、PSM よりも狭い推定値の信頼区間を得られる可能性が示唆された。