Estimating causal effect using Random Forest based propensity score: when the propensity score is affected by unobserved interaction of covariates. (Random Forest による説明変数の交互作用を考慮した傾向スコアの推定)

慶應義塾大学大学院理工学研究科　中村知繁
慶應義塾大学理工学部　南 美穂子

1　はじめに

統計的因果推論における平均処置効果の推定法は，マッチング法 (Rubin, 1985) や逆確率重み付け (Hirano, 2003)，2 重に頑健な推定量（Robins, 1994）など傾向スコアを用いるものが多い．実際の解析においては，傾向スコアの真値は観測できないため，データから適切に推定する必要がある．傾向スコアの推定を誤った場合には，平均処置効果の推定にバイアスが生じるということは様々な文献で報告されている（Kang and Schefer, 2007）．これらの問題へと対処するために，近年，傾向スコアに対するモデルの誤特定に対して頑健な傾向スコアの推定法が提案されている (Imai and Ratkovic, 2014; Zhao, 2015; Wang and Zhou, 2016)．これらの手法の頑健性はシミュレーションなどで確認されているが，一方で傾向スコアに対する陽なモデルを構築する必要があるため，モデルの誤特定の種類によっては，推定された平均処置効果の一致性が保証されない．

2　ランダムフォレストを用いた因果的効果の推定

本報告では，以上を踏まえて，RandomForest（Breiman, 2001）を用いてノンパラメトリックに傾向スコアを推定する方法について報告する．Random Forest(RF) は，決定木/回帰木のアンサンブルによって構成されるバギング法の一種であり，予測やクラスタリングの問題に対して様々な分野で優れた結果を残している（Denil et al., 2014）．また，学習のパラメータをうまく設定することにより，over-fitting を抑制することができることも知られている (Louppe, 2014)．RF を傾向スコアの推定に用いる利点は，これらに加えて傾向スコアに対する回帰モデルを陽に記述する必要がないという点である．例えば傾向スコアに対するモデルで，説明変数の次元が大きく，変数間に交互作用がある場合，それらを正確にモデリングするのは実際の解析の場面では困難である．しかし，RF は説明変数のみを入力すれば，それらの交互作用や高次元項までを踏まえた予測モデルを構築することが可能である（Baiu, 2012）．また，傾向スコアが極端に 0 または 1 に近づいた結果，平均処置効果の推定結果が不安定になるという問題も，RF のパラメータ調整によって，特徴空間の分割の粗さを変化させることにより調節できる可能性がある．

当日の発表では，これまでに報告されている理論的な結果を踏まえた上で，傾向スコアを RF を用いて推定する方法について述べ，いくつかのシナリオのもとでのシミュレーションを行った結果を報告する．

参考文献 (抜粋)

L. Breiman, Random forests, Machine Learning, 45, (2001) 5 – 32.
E. Scornet, On the asymptotics of random forests, Journal of Multivariate Analysis, 146, (2016) 72 – 83.
G. Biau, Analysis of a random forests model, Journal of Machine Learning Research, 13, (2012) 1063 – 1095