2A-01

遺伝的アルゴリズムを用いた因果分析の解釈性の向上

Improving Interpretability of Causal Analysis Using Genetic Algorithm

浅沼 爽汰[1]　山本 純一[1]　菅原 収吾[1]　渡部 佳織[1]
NEC ソリューションイノベータ[1]

1. はじめに

LiNGAM[1]などをはじめとした因果探索手法は多く提案されているが,解釈性を考慮して因果を間引いていく手法は少ない.因果モデルの解釈性を向上させる場合,探索された因果モデルから weight や p 値に応じて因果を絞り込む方法や,因果モデルの一部を切り出す方法が考えられる.しかし,これらの方法では,因果モデルを構造方程式モデリング(SEM)で検証すると,絞り込み前後や切り出し前後で SEM の適合度が悪化してしまう．すなわち，探索後の因果モデルから，適合度を保ちつつ，解釈可能な形に因果を間引く方法は明らかになっていない.

本稿では,組合せ最適化手法である遺伝的アルゴリズム(以下,GA とする）を使用し,多変量データの解釈困難な因果モデルから,モデルの適合度を維持,向上しながら解釈可能な形に因果を間引いていく手法を提案する.

2. 遺伝的アルゴリズム

GA とは,生物の適応進化に関する自然淘汰説に着想を得た計算手法であり[2],組合せ最適化などに使用される．遺伝子で表現した個体を複数用意し,何らかのスコアの高い個体を優先的に選択し,交叉･突然変異などを繰り返すことで解を探索するアルゴリズムである.

よく使用される選択方式と交叉方式に,エリート選択方式と２点交叉手法がある． エリート選択方式は,個体群の中で一定割合のスコアの良い個体を残す方法である． ２点交叉手法は,個体内で遺伝子をランダムに２箇所指定し,２個体の指定した箇所の間の遺伝子を図１のように交換する手法である.

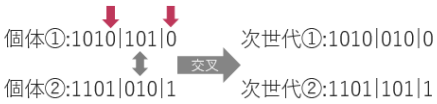

図１ ２点交叉手法のイメージ

3. 従来手法

因果分析は，因果モデルを推定する因果探索と推定された因果モデルを検証する因果推論で構成され，因果を間引く工程はほとんど考えられていない.しかし,多変数の因果分析では，因果探索を行うと膨大な因果が見つかり，解釈が困難となる．そのため,因果モデルの一部を抜粋したり,因果推論で weight や p 値,SEM の適合度を参考にして，因果を強制的に削除することで，仮説に近いものを探索するという方法がとられる.

しかし，このような方法では，因果を間引くことはできても，因果推論による検証に耐えうる因果モデルを見つけることが難しい．これは，因果の組合せの数が多く，様々な論文[3]で提唱されている SEM のモデル適合度を満足するような因果モデルを発見することができないためである.したがって，因果の組合せを効率的に試して検証する方法が必要となる.

4. 提案手法

従来手法の課題を解決するために，因果の組合せを効率的に試す方法として GA を使用するアルゴリズム「遺伝的因果探索（Genetic Causality Discovery; GCD)」を提案する.

提案手法 : Genetic Causality Discovery
1. 初期モデルを作成（実験：LiNG-SEMs）
2. モデルを遺伝子に見立て個体を作成
3. 個体ごとの適合度をSEMで算出
4. 適合度をもとに選択・交叉・突然変異を行い3.に戻る
5. 適合度が一番良い個体がn回同じ値の場合終了(実験：n=5)

この最適化アルゴリズムでは，因果の有無を１つの遺伝子に見立て，因果の集合を個体に対応させ（図２），GA より因果の組合せを発生させる．発生させた因果モデルの検証には SEM を使用し，SEM のモデル適合度指標を用いたスコアを計算する．スコアが最適なものになるまで,検証を繰り返す.

図２ 提案手法のモデルと遺伝子の関係

4.1 初期モデルの生成方法

LiNGAM は因果モデルに循環がない場合に使用できるが,一般には非循環を仮定することはできない.そのため,GCD では,LiNGAM の非循環制約を緩和した LiNG-SEMs[4]を使用する.

1 NEC Solution Innovators Ltd.

4.2 スコア関数の定義

スコア関数は，SEM のモデル適合度の最重要指標である CFI(Comparative Fit Index)がおよそ 0.95 に収束するように最小二乗法で定義し，RMSEA(Root Mean Square Error of Approximation)が 0.06 付近に，因果数が有意な相関のあった説明変数ペアの数に制約されるようにラグランジュ未定乗数法で制約条件を追加した．CFI と RMSEA の判定値は,モデル適合度を判断するのによく使われる値である[3].

4.3 不安定循環の削除

因果モデルに循環が発生すると，安定な循環と不安定な循環に分けることができる．例えば,要素 A,B,C のうち,A→B，B→C が正の weight を持ち，C→A が負の weight を持つとする．この場合，A が増えれば B が増え，B が増えれば C が増えるが,C が増えると A が減り，A が増えたり減ったりするため不安定になる．そのため，３つ weight の中で 1.0 以上の weight と負の weight が共存する場合，その因果を削除した．

4.4 遺伝子発生方式

GCD では,個体発生方式として,事前試行の結果,最も最適化効率が高かったエリート選択方式と２点交叉手法の組合せを用いた.エリート選択方式では,50%の個体を残す設定にした．

4.5 終了条件

最良スコアの個体が 5 回連続で選ばれたとき終了とした．

5. 実験１

妥当な因果の組合せを見つけるのに全組合せを試行する方法は,変数の数が多いと膨大な時間がかかってしまう．そこで,GCD によって処理時間の短縮ができるかを検証した．

5.1 実験方法

因果数を固定しながら,全組合せパターンを網羅した場合と GCD との実行時間を測定する．因果数に応じた実行時間の変化を観察する．

5.2 実験データ

標準化済みの心理尺度アンケートデータ（７件法，62 変数，n=564）を使用した．

5.3 実験結果

因果数８の場合を超えると，全組合せ網羅よりも GCD の方が,妥当な因果モデルを発見するまでの実行時間が短いことが分かった（図３）．

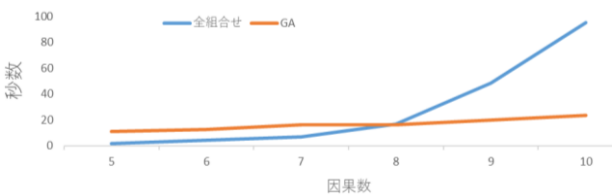

図 3 実行時間（青線：全組合せ,オレンジ線：GA）

6. 実験２

たとえ妥当な因果モデルが発見できたとしても，初期モデルより因果数が増えてしまっては，より解釈を困難にしてしまう．そこで,GCD によって因果数が減少することを確認する．

6.1 実験方法

GCD を実行しながら，各ステップにおいて因果数とスコア関数の値を記録し，スコアが減少すること，および因果数も減少していくことを確認する.実験データは,実験１と同じものを使用した．

6.2 実験結果

GCD によって，スコアが低下し，因果数が 50 以上からから 40 以下に減少した（図４）．

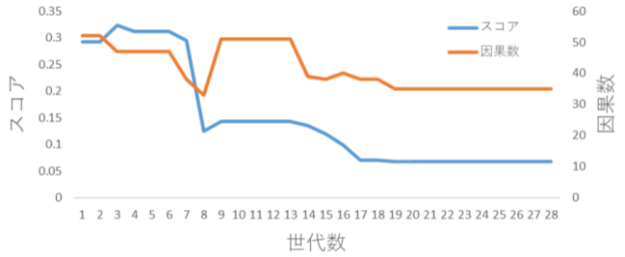

図 4 因果数の減少（青：スコア,オレンジ：因果数）

7. 結論

実験１の結果より,GCD は因果の全組合せを探索するよりは妥当な因果モデルの発見が高速であった．実験２の結果から，GCD は初期モデルよりも因果数を減らすことができるこが示された.これにより,多変数の因果探索結果が複雑すぎて解釈困難であっても,GCD が適合度を損なわずより解釈可能な因果モデルを提供できることが示唆された．

参考文献

[1] Shohei Shimizu, Patrik O.Hoyer, Aapo Hyvarinen, Antti Kerminen, A Liner Non-Gaussian Acyclic Model for Causal Discovery, Journal of Machine Learning Research 7 (2006).

[2] 喜多 一,遺伝的アルゴリズムによる最適化の現状,若手研究者・学生向けに最新技術をわかりやすく紹介する講演会「確率的アルゴリズムによる情報処理」

[3] Yamamoto, J. I., Fukui, T., Nishii, K., Kato, I., & Pham, Q. T. Digitalizing Gratitude and Building Trust through Technology in a Post-COVID-19 World—Report of a Case from Japan. Journal of Open Innovation: Technology, Market, and Complexity, 8(1), 22.Mark. Complex 2022

[4] Gustavo, L, Peter, S.J.R, Patrik, O.H, Discovering Cyclic Causal Models by Independent Components Analysis, (2008)