F-051

部分集合列挙による特定集団に固有な因果構造の発見プラットフォーム
Causal Discovery Platform in Heterogeneous Populations by Subpopulation Enumeration

藤重　雄大[1)]　鈴木　浩史[1)]　小柳　佑介[1)]　上村　健人[1)]　岩下　洋哲[1)]
Yuta Fujishige　Hirofumi Suzuki　Yusuke Koyanagi　Kento Uemura　Hiroaki Iwashita

Shuang Chang[1)]　浅井　達哉[1)]　丸橋　弘治[1)]　鹿島　久嗣[2)]
Chang Shuang　Tatsuya Asai　Koji Maruhashi　Hisashi Kashima

1　はじめに

近年、マーケティング、製造、医療など様々な実世界の分野において、AI を用いた意思決定支援の試みが増加している。意思決定では本来変更したいものの直接変更できない対象変数を、他の変数を代わりに操作することで変更することを目的としており、それは相関関係の確認だけでは達成できない。そのため、「A が B を引き起こす」というような原因と結果の関係を表す因果関係 [1] に着目する必要がある。

この代替変数の操作においては、対象変数の強い原因である変数を操作することが効率的である。しかし，多くの場合因果関係は事前にわからないため，データから推定する必要がある．

さらに、実用上は以下のような特定の条件下においてのみ成り立つ因果関係を考慮することが重要である [2, 3, 4, 5]。

- マーケティング：効果的なプロモーションは顧客ごとに異なるため、ある特徴を持つ顧客に特化した購入につながる要因を発見する必要がある。
- 製造：製品ごとに不良の原因が異なるため、特定の条件下で故障する要因を発見する必要がある。
- 医療：すべての患者に共通する要因ではなく、患者の状態に合わせた病気の原因を発見する必要がある。

本論文では、特定集団に固有な因果構造を探索・発見するプラットフォームを提案する。図 1 は、本プラットフォームの処理フローを示す。与えられたサンプルに対して、条件列挙部は、対象変数と高い相関を持つ説明変数郡からなる積項として定義される条件をすべて列挙し、それぞれの条件を満たすサンプルの部分集合を構築する。その後、因果探索部は、それぞれの部分集合に対して因果探索技術を適用し、各条件における変数間の因果関係を算出する。このような条件と因果関係の組は膨大な数になりうるため、情報統合部はサンプル部分集合の類似性と因果関係の類似性に基づいて結果をクラスタリングする。最後に、因果可視化部は、対象変数に大きな影響を与える変数群と、それらの間の因果関係を可視化する。

本プラットフォームの貢献は、条件を網羅的に検索し、データ全体の因果関係との差異を可視化することで、特定の部分集団にのみ存在する固有の因果関係を発見することである。

1）　富士通株式会社
2）　京都大学

2　関連研究

提案プラットフォームの条件列挙部では、相関関係を発見するために、エマージングパターンマイニング技術を用いる。データベースや知識発見の分野では、頻出パターン [6]、エマージングパターン [7]、最適パターン [8] などを高速に発見するデータマイニングの研究が盛んに行われている。これらの手法では、データ中の説明変数の全ての組合せを探索し、頻出する組合せや教師ラベルと強い相関を持つ組合せを網羅的に見つけ出す。

因果探索部は、条件列挙部が算出した条件に基づき抽出されたデータサンプルの部分集合における因果関係を探索する。構造方程式モデルで変数間の因果構造を表現し、部分サンプル集合からそのパラメータを推定することで最も尤度の高い構造を同定する。これまでに変数の種類や因果方程式が異なる様々な因果モデルと、それらの推定アルゴリズムが提案されている [9, 10, 11]。これらの手法は全て、与えられたデータに対して一つの因果構造を推定する。さらに近年、異なる因果関係をもつ集団がデータ内に混在する設定下における因果構造推定を理論的に分析した [4, 5]。本プラットフォームでは、異なる因果関係をもつ複数集団を扱うために、特徴的な因果関係を持つ可能性のある部分集団ごとに因果探索手法を適用し、データ全体ではなく、条件ごとの様々な因果関係を推定している。

因果関係の発見を目的としたプラットフォーム [12, 13] もこれまでにいくつか提案されているが、これらは特定のアプリケーション向けに設計されている。我々のプラットフォームは一般的な使用を想定しており、特定の条件下で出現する特徴的な因果関係を網羅的に発見することができる。

3　提案プラットフォーム

本節では、提案プラットフォームの各モジュールを図 1 に沿って説明する。

3.1　条件列挙部

条件列挙部では、特徴的な因果関係を持つ可能性のあるデータサンプルの部分集合（以下、部分サンプル集合という）を表す条件を全て列挙する。愚直に全ての条件候補に対して因果関係を計算することは、計算量の観点で非現実的である。一般に 2 つの変数が因果関係を持つとき、それらは相関関係も持っているため、提案手法では列挙する条件を特徴的な相関関係へと緩和し、効率的なパターンマイニングアルゴリズムで列挙している。今回の実装では、Wide Learning [14, 15] を用いて、エマージングパターン [7] を効率よく列挙している。

3.2　因果探索部

因果探索部は、条件列挙部で列挙された各条件に対応する部分サンプル集合に対して因果探索アルゴリズムを適用することで、それらに対して因果関係を出力

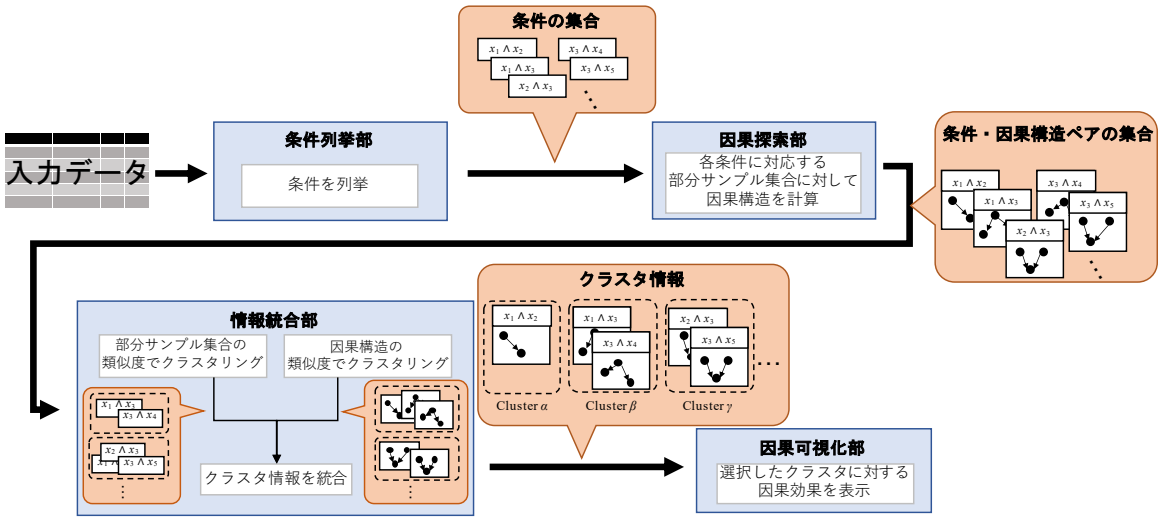

図 1　提案プラットフォームのフローチャート

する。今回の実装では、因果探索アルゴリズムとして、DirectLiNGAM [10] を採用している。

3.3　情報統合部

因果探索部の出力は、ユーザーが確認できないほど多いことがあるため、情報統合部では、類似する出力をクラスタリングすることで要約している。因果構造と部分サンプル集合の 2 つの類似性に対して独立してクラスタリングを実行する。これらはそれぞれ、因果構造、および部分サンプル集合間の距離を表す。今回の実装では、因果構造を表す DAG の隣接行列のフロベニウスノルムの逆数を因果関係類似度、Jaccard 距離の逆数を部分サンプル集合間の類似度として、階層的クラスタリングアルゴリズムにてそれぞれクラスタリングを行っている。

3.4　因果可視化部

因果可視化部は、情報統合部で計算された各クラスタの因果関係を可視化する。様々な条件下における各説明変数から対象変数への因果効果の分布を表示し、特徴的な因果関係を把握できる。

また、因果可視化部では、データに潜む未知の因果関係を探索したいユーザーのために、各条件下での因果グラフを表示できる。4 節では、オープンデータに対して提案プラットフォームを適用した例を示す。

4　実験

本プラットフォームの実装例として、以下の 2 つの UCI データセットを用いた実験を実施した。

- Wine quality [16]
- Student performance [17]

4.1　Wine quality

Wine quality は、ポルトガル北部のワインサンプルに関するデータセットである。図 2 と図 3 は、Wine Quality を対象変数として、各変数からの因果効果を可視化した結果である。このデータに対して、条件列挙部は 1325 の条件を抽出し、情報統合部が 32 の条件に統合した。図 2 は、「fixed acidity(酒石酸濃度) が高い」という条件に対応するクラスタの結果である。residual suger(残留糖分濃度) の項目を見ると、残留糖分濃度はデータセット全体の品質にマイナスの影響を与えるが、この条件では残糖が増えると品質が上がるという因果関係が抽出された。これは「残留糖分濃度を上がると一般にワインの品質は下がるが、酒石酸濃度が高い場合は残留糖分濃度をあげることが品質の向上する傾向がある」ことを示唆している。

図 3 は、「alcohol(アルコール度数) が低い ∧ total sulfur dioxide（総亜硫酸濃度）が高い」という条件に対応するクラスタの結果である。データセット全体で算出された因果関係とは対照的に、「塩化ナトリウム濃度が低くなると品質が高くなる」という因果関係が出力されている。このことから、「一般的に低品質になりがちな低アルコールワインでは亜硫酸塩の添加による品質管理が重要であるが、塩分濃度を下げることも重要である」ことが示唆された。

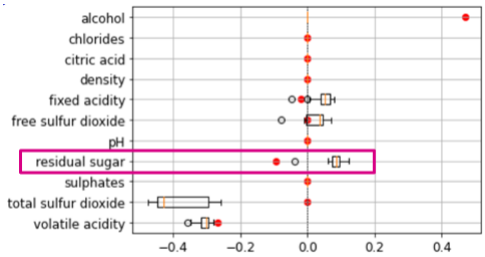

図 2　条件「fixed acidity(酒石酸濃度) が高い」に関する結果

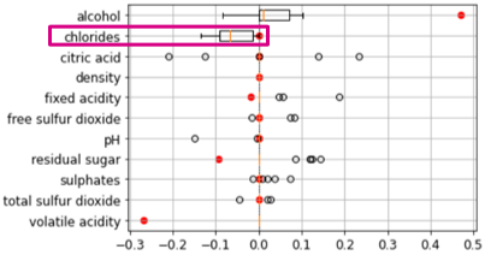

図 3　条件「alcohol(アルコール度数) が低い ∧ total sulfur dioxide（総亜硫酸濃度）が高い」に関する結果

4.2 Student Performance

Student Performance は、ポルトガルの学校における生徒に関する情報に関するものであり、G3(ポルトガル語の授業の 3 学期成績) を対象変数として、提案プラットフォームを適用した。本データに対して、条件列挙部では 769 の条件を抽出し、情報統合部にてそれを 10 個のクラスタへとまとめた。図 4 は、「Gabriel Pereira School に通っている ∧ G1(1 学期成績)≤11.5」という条件に対応するクラスタの因果関係を、可視化した結果である。ただし、データセット全体の因果関係には存在しないが、着目している条件に出現するの因果関係に対応する辺を赤色で強調表示している。「G2」から「absences」へのエッジは、負の重みを持つため、「2 学期に成績が上がると欠席回数が減る」ことを示唆している。また、「Fjob_at_home_0」から「absences」へのエッジは、「父親が在宅勤務でない場合、欠席数が減少する」ことを示唆している。これらの辺は赤色であり、データ全体で観察される因果関係の傾向とは異なっている。このように、本プラットフォームを利用することで、条件ごとに存在する独自の因果関係を発見することができる。

5 結論

多種の因果構造を含むデータに対して、様々な条件下で因果関係を包括的に抽出し、可視化するプラットフォームを提案した。また、オープンデータセットを用いて本プラットフォームの実現性を実証し、その有効性を示した。

今後は、患者の状態に応じて疾患と遺伝子の因果関係を発見するバイオインフォマティクスや、特定の条件下での酸化還元反応の原因を特定するマテリアルインフォマティクス（MI）など、実世界の問題に本プラットフォームを適用する予定である。

参考文献

[1] J. Pearl, *Causality: Models, Reasoning, and Inference*, Cambridge University Press, 2000.

[2] J.E. Brand, J.S. Thomas, Causal effect heterogeneity, In *Handbook of Causal Analysis for Social Research*, pp. 189–213, Springer, 2013.

[3] Y. Xie, Population heterogeneity and causal inference, Proc. the National Academy of Sciences, 110(16), pp. 6262–6268, 2013.

[4] B. Saeed, S. Panigrahi, C. Uhler, Causal structure discovery from distributions arising from mixtures of DAGs, Proc. 37th international conference on machine learning (ICML), Online, PMLR 119, 2020.

[5] A. Markham, R. Das, M. Grosse-Wentrup, A distance covariance-based kernel for nonlinear causal clustering in heterogeneous populations, Proc. 1st conference on causal learning and reasoning (CLeaR), pp. 1–17, 2022.

[6] R. Agrawal, R. Srikant, Fast algorithms for mining association rules, Proc. 20th international conference on very large data bases (VLDB), pp. 487–499, 1994.

[7] G. Dong, J. Li, Efficient mining of emerging patterns: Discovering trends and differences, Proc. 5th international conference on knowledge discovery and data mining (KDD), pp. 43–52, 1999.

[8] S. Morishita, J. Sese, Transversing itemset lattices with statistical metric pruning, Proc. 19th SIGMOD-SIGACT-SIGART symposium on principles of database systems (PODS), pp. 226–236, 2000.

[9] J. Peters, J. M. Mooij, D. Janzing, B. Schölkopf, Causal Discovery with Continuous Additive Noise Models, Journal of Machine Learning Research, vol. 15, pp. 2009–2053, 2014.

[10] S. Shimizu, T. Inazumi, Y. Sogawa, A. Hyvärinen, Y. Kawahara, T. Washio, P. O. Hoyer, K. Bollen, DirectLiNGAM: A Direct Method for Learning a Linear Non-Gaussian Structural Equation Model, Journal of Machine Learning Research, vol. 12, pp. 1225–1248, 2011.

[11] K. Uemura, S. Shimizu, Estimation of Post-Nonlinear Causal Models Using Autoencoding Structure, Proc. 45th International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3312–3316, 2020.

[12] S. Budd, A. Blaas, A. Hoarfrost, K. Khezeli, K. D'Silva, F. Soboczenski, G. Mackintosh, N. Chia, J. Kalantari, Prototyping CRISP: A Causal Relation and Inference Search Platform applied to Colorectal Cancer Data. Proc. 3rd Global Conference on Life Sciences and Technologies (LifeTech), pp. 517–521, 2021.

[13] C. Cintas, R. Raghavendra, V. Akinwande, A. Walcott-Bryant, C. Wayua, K. Weldemariam, Decision Platform for Pattern Discovery and Causal Effect Estimation in Contraceptive Discontinuation. Proc. 29th International Joint Conference on Artificial Intelligence (IJCAI), pp. 5288–5290, 2020.

[14] H. Iwashita, T. Takagi, H. Suzuki, K. Goto, K. Ohori, H. Arimura, Efficient constrained pattern mining using dynamic item ordering for explainable classification, arXiv, CoRR abs/2004.08015, 2020.

[15] Hello, Wide Learning!, `https://widelearning.labs.fujitsu.com/en/`

[16] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties, Journal of In Decision Support Systems, vol. 47, pp547-553, 2009.

[17] P. Cortez and A. Silva, Using Data Mining to Predict Secondary School Student Performance, Proc. 5th Future Business Technology Conference (FUBUTEC), pp. 5–12, 2008.

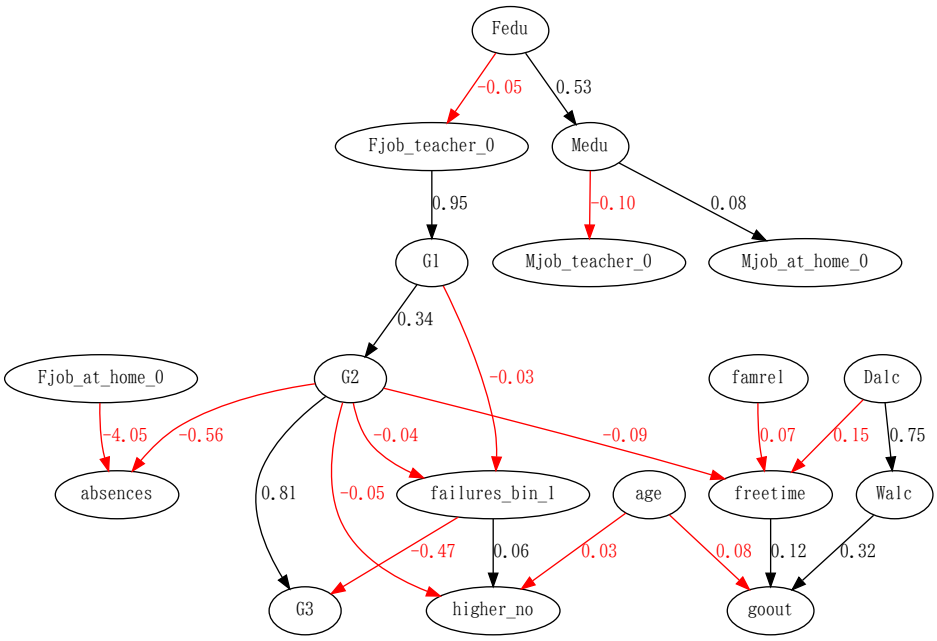

図 4　条件「Gabriel Pereira School に通っている ∧ G1(1 学期成績)≤11.5」に関する因果グラフ