様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　　元 年　　５ 月 ２０ 日現在

機関番号：１０１０１
研究種目：挑戦的萌芽研究
研究期間：2015～2018
課題番号：１５Ｋ１２１４８

研究課題名（和文）クラウドソーシングを用いた統計的因果推論基盤の構築

研究課題名（英文）Development of a Framework for Statistical Causal Analysis Using Crowdsourcing

研究代表者
小山　聡（Oyama, Satoshi）

北海道大学・情報科学研究科・准教授

研究者番号：３０３４６１００

交付決定額（研究期間全体）：（直接経費）　　2,800,000円

研究成果の概要（和文）：本研究では，クラウドソーシングによって多くの人々の能力を活用してオープンデータの分析を行うための基盤技術の研究を行った．とくに，相関関係ではなく因果関係を分析するタスクに着目し，クラウドソーシングで人間の知識を活用して因果関係の影響を与える変数の候補を発見し，探索的にオープンデータの分析を行うフレームワークを開発した．世界銀行や日本政府のオープンデータなどを用いて，各要素技術およびフレームワークの評価を行った．

研究成果の学術的意義や社会的意義
企業や官公庁の公開するオープンデータを活用して，社会的な意思決定やビジネス上の決定を行うことが今後ますます重要になってくると考えられる．その際に，相関関係と因果関係を区別することは重要であるが，そこには人間の背景知識の利用が不可欠である．本研究は，インターネット上で不特定多数に仕事を依頼できるクラウドソーシングを用いて，多くの人々の知識を活用して因果分析を行うフレームワークを提案しており，オープンデータ活用への一助になることが期待できる．

研究成果の概要（英文）：In this research, we investigated fundamental technologies to analyze open data using the ability of many people by crowdsourcing. In particular, we focused on the task of analyzing causality but not correlation and developed a framework for analyzing open data in an exploratory manner. Human knowledge was incorporated by crowdsourcing to discover candidate variables that can affect causality. The elemental technologies and the framework were evaluated using the open data of the World Bank and the Japanese government.

研究分野：人工知能

キーワード：クラウドソーシング　オープンデータ　因果分析

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

近年データのオープン化の推進により様々な統計データがWeb上で入手可能になってきている．また単に個々のデータを公開するだけでなくRDF（Resource Description Framework）によりデータを機械可読な形で記述し互いにリンクするLinked Open Dataの取り組みも進められている．これにより，計算機が自動的に関連するデータを発見し，それらを組み合わせて分析を行うことが可能になると期待されている．しかし，リンクされるデータの組合せが増えることは，例えば「学力テストの成績」と「自動車の保有台数」のように，直接は関連のない変数間に強い相関を発見する可能性も高くなる．これらは，いわゆる見せかけの相関であり，真の原因となる交絡変数（例えば「所得」）が別にあり，その影響によって相関が生じていると考えられる．データ分析の目的は多くの場合，分析結果をもとに何らかのアクションを起こすことであるが，因果関係（例えば「所得」を増やしたとき「学力テストの成績」が上がるか）が分からなければ，効果的な意思決定は困難となる．

因果関係を特定する直接的な方法は，ランダム化比較実験を行うことであるが，オープンデータで扱われる社会的な問題などに対しては，そのような実験は経済的・倫理的な理由などから実際には困難であることが多い．このような課題を解消するために，観測データから変数間の因果関係を推定する統計的因果推論と呼ばれる方法が研究されてきた．これまでの研究の多くでは，因果関係に影響を与える変数は，観測データの中に全て含まれているという仮説を置いていた．しかし，オープンデータの環境においては，事前に全ての関連するデータを取得して分析することは困難である．

データ分析の自動化は一般に考えられているほど容易ではなく，さまざまな段階で人手での作業が必要であることが認識されている．一方，インターネット上で不特定多数に仕事を依頼できるクラウドソーシングが，人工知能研究の様々な場面で用いられるようになってきている．クラウドソーシングは画像分類などの，出力が単純なタスク以外にも，自然言語で回答を行うようなタスクに対しても用いられる．クラウドソーシングの利点に一つとしては，集合知と呼ばれる，多くの人々の多様な知識や意見に容易にアクセスできることが挙げられる．

２．研究の目的

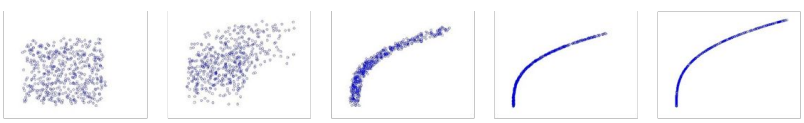
本研究では，クラウドソーシングによって多くの人々の能力を活用してオープンデータの分析を行うための基盤技術の研究を行った．とくに，オープンデータの環境で因果分析を行う際には，事前に関連のあるデータをすべて取得して分析を行うことは困難であり，分析の対象を特定するためには，人間の知識を導入することが不可欠である．本研究はクラウドソーシングにより多くの人の知識を導入し，計算機による自動的な処理と組み合わせることで，オープンデータ環境での因果分析を可能にすることを目的とする．

３．研究の方法

本研究においては，人間の知識を活用するために，商用のクラウドソーシングプラットフォームである Lancers（https://www.lancers.jp/）を利用した．　評価実験においては，CauseEffectPairs（CEP，Mooij et al. 2016）と呼ばれる，因果分析のベンチマークデータセット（https://webdav.tuebingen.mpg.de/cause-effect/）に加えて，日本政府（https://www.e-stat.go.jp/en）と世界銀行（https://data.worldbank.org/）のオープンデータを利用して研究を行った．

４．研究成果

（１）因果分析フレームワークで用いることができる既存手法を特定するために，二変数に対する既存の因果発見モデルを評価する方法を開発した．評価対象としてANMモデル（Hoyer et al.，2009），PNLモデル（Zhang et al.，2008），IGCIモデル（Janzing et al.，2012）を用い，CauseEffectPairs（CEP）データセットについて評価した．その際には，異なる決定率でモデルの予測制度を比較し総合的な比較を行った．さらに，各モデルの効率性を，計算時間の観点からも比較した．とくに，未知の交絡因子が各モデルに与える影響を評価するための，シミュレーションモデルを提案した（図１）．実データとしては，CEPデータから取得した異なるデータで共通の変数を持つものを組み合わせて，共通原因および選択バイアスのデータを生成した．これらのデータにより，各手法が未知の交絡因子に受ける影響を定量的に評価することが可能となった．



(a) a/b=0.1　(b) a/b=1　(c) a/b=10　(d) a/b=100　(e) a/b=1000

図 1 交絡の強さを制御したシミュレーションデータ

実験の結果, 全てのデータに対して因果関係の判定を行う場合 IGCI モデルが最も優れていることが分かった．また，計算速度においても IGCI が優れた性能を示した．交絡因子の影響について分析した結果，シミュレーションデータに対しても，実データに対しても，各モデルは交絡因子の影響を完全に防ぐことはできなかった．

（2）ある変数が別の二つの変数の共通原因であるための必要条件を用いて，共通原因の候補を発見する方法を提案した．2 つの変数 X と Y の間に相関がある場合，3 つ目の変数が共通原因になっている場合（図 2(a)）や選択バイアスの場合（図 2(b)）がある．データの内在次元を推定する方法を用いれば，理論的には，三つの変数が共通原因の関係にある場合は次元数が 1 に，選択バイアスの関係にある場合は次元数が 2 になることを示した．具体的には Grassberger らによる相関次元推定（Grassberger et al．2004）を用いてデータのフラクタル次元を計算し，その値が閾値より 1 に近ければ，共通原因の可能性があると判定し，そうでなければ，共通原因ではないと判定する．つぎに，シミュレーション実験を行い，データの関数形やノイズの種類を変えて評価を行った．さらに，CEP データセットから取得した実データに対して実験を行い，提案手法が共通原因の候補を特定できることが示した．この手法はマルコフ同値類を区別できないので，3 つ目の変数が共通原因であるための必要条件のみを与えるが，（3）で示すワークフローにおいて，交絡変数の候補を絞り込むことに用いることができる．

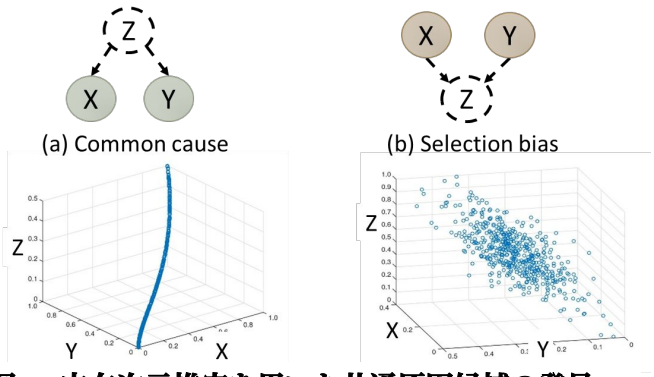


図 2 内在次元推定を用いた共通原因候補の発見

（3）データから因果関係を推定するこれまでの研究の多くでは，交絡因子と呼ばれる因果関係に影響する変数は，観測データの中に含まれているという仮説を置いていた．しかし，オープンデータの環境においては，事前に全ての関連するデータを取得して分析することは困難である．そこで，交絡変数の可能性があるデータを随時取得していく，探索的なデータ分析を行うフレームワークを提案した（図 3）．

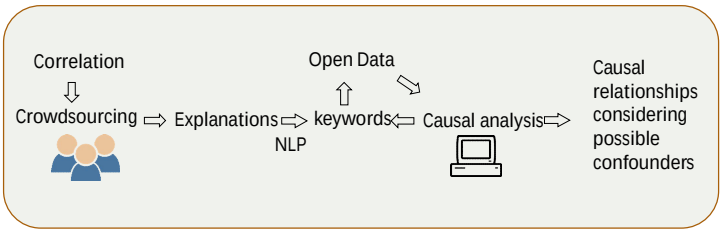


図 3 探索的因果分析フレームワーク

このフレームワークにおいては，オープンデータの中で相関が観測される変数の組に対して，クラウドソーシングでその理由についての説明を募集する．得られた説明文から，自然言語処理を用いて，交絡因子の可能性がある変数名の候補を抽出する．その際には，接続標識「ため」に着目する方法（乾ら，2002）と単語の生起確率に基づく方法を利用する．オープンデータからこれらの変数名に対応するデータを取得し，元のデータと組み合せて因果分析を行う．

取得したデータに対する因果分析は，図 4 のようなワークフローに従って実施する．新たに取得した変数に対し，（2）に示した内在次元推定の方法および既存手法を用いて，交絡因子の可能性があるか否かを検証する．もし交絡因子の可能性がある場合，（1）での評価の結果，高い精度と速度を示した IGCI モデルを用いて，因果関係の方向を推定する．

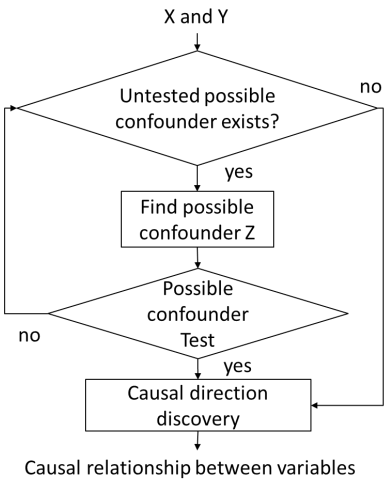


図 4 因果分析ワークフロー

因果分析のケーススタディとして，世界銀行のオープンデータにおいては，「GDP」と「二酸化炭素排出量」の関係を，日本政府のオープンデータにおいては，「大学進学率」と「負債」の関係を扱った．これらの変数はそれぞれのデータにおいて強い相関を示していたが，クラウドソーシングで各 50 人の作業者

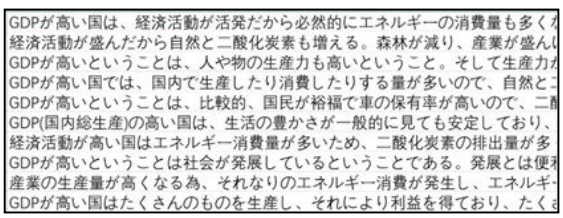

GDPが高い国は、経済活動が活発だから必然的にエネルギーの消費量も多く
経済活動が盛んだから自然と二酸化炭素も増える。森林が減り、産業が盛ん
GDPが高いということは、人や物の生産力も高いということ。そして生産力
GDPが高い国では、国内で生産したり消費したりする量が多いので、自然と
GDPが高いということは、比較的、国民が裕福で車の保有率が高いので、二
GDP(国内総生産)の高い国は、生活の豊かさが一般的に見ても安定しており、
経済活動が高い国はエネルギー消費量が多いため、二酸化炭素の排出量が多
GDPが高いということは社会が発展しているということである。発展とは便
産業の生産量が高くなる為、それなりのエネルギー消費が発生し、エネルギー
GDPが高い国はたくさんのものを生産し、それにより利益を得ており、たく

図 5 クラウドソーシングで得た説明文の例

からその理由についての説明文を募集した．図 5 に世界銀行のデータに対して収集した説明文の例を示す．一人当たりの説明文の平均文字数は 94.26 であり，クラウドソーシングによって，かなり詳細な説明を取得できていることが分かる．

収集した説明文から，交絡変数の候補となる名詞を抽出し，オープンデータから対応するデータを取得した．世界銀行のデータに対して，図 4 のワークフローに従って因果分析を行った結果を図 6 に示す．この中から，さらに可能性の高い候補を絞り込んで行くことは今後の研究課題であるが，いくつか因果関係の候補として検討の価値のあるものが抽出されていることが分かる．

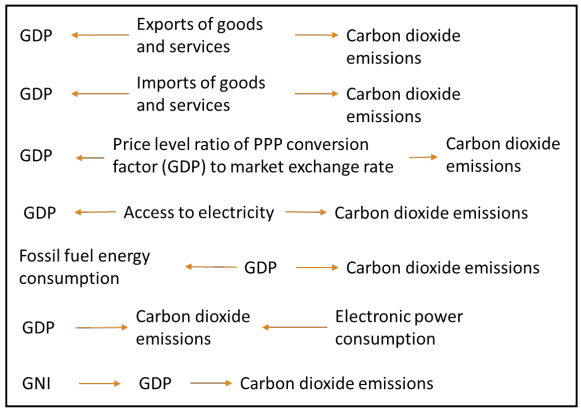


図 6 抽出された因果関係の候補

（4）現在，オープンデータの取組みによって様々な統計データが行政等によって公開されているが，これらのデータは画像や PDF の形式で与えられるものが少なくなく，データ分析などでの再利用を妨げている．そこで，クラウドソーシングを用いて，グラフ画像として与えられたレガシーな統計データを機械可読な表形式に変換する枠組みを提案した．その際，作業者に表だけを作成させるのではなく，グラフ画像をスプレッドシート上でグラフとして視覚的に再現させるタスク設計を行った．このタスク設計により，データの入力誤りに気付き易くなる効果に加えて，再現されたグラフオブジェクトのプロパティとして項目名や系列といったデータの構造を容易に取り出し，作業結果の統合などに利用することが可能となる．観光白書や情報通信白書といった日本政府のオープンデータに含まれるグラフ画像に対し，クラウドソーシングを用いて EXCEL でグラフを再現するタスクを実行し，その後プログラムで複数の作業者の結果を統合して最終的なデータ表を出力する評価実験を行い，提案手法の有効性を検証した．

＜引用文献＞

① Patrik O. Hoyer, Dominik Janzing, Joris M. Mooij, Jonas Peters, and Bernhard Scholkopf: Nonlinear Causal Discovery with Additive Noise Models, Advances in Neural Information Processing Systems (NIPS), 2009.
② Kun Zhang and Aapo Hyvarinen: Distinguishing Causes from Effects Using Nonlinear Acyclic Causal Models, NIPS 2008 Causality Workshop, 2008.
③ Dominik Janzing, Joris Mooij, Kun Zhang, Jan Lemeire, Jakob Zscheischler, Povilas Daniusis, Bastian Steudel, and Bernhard Scholkopf: Informationgeometric Approach to Inferring Causal Directions, Artificial Intelligence, Vol. 182, pp. 1-31, 2012.
④ Joris M. Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Scholkopf: Distinguishing Cause from Effect Using Observational Data: Methods and Benchmarks, Journal of Machine Learning Research, Vol. 17, No. 32, pp. 1–102, 2016.
⑤ Peter Grassberger and Itamar Procaccia: Measuring the Strangeness of Strange Attractors, The Theory of Chaotic Attractors, pp. 170-189, 2004.
⑥ 乾 孝司，乾 健太郎，松本 裕治: 接続標識「ため」に基づく文書集合からの因果関係知識の自動獲得，情報処理学会論文誌，Vol. 45, No. 3, pp. 919 - 933, 2004.

5．主な発表論文等

〔雑誌論文〕（計 2 件）

① Jing Song, Satoshi Oyama, and Masahito Kurihara: Tell Cause from Effect: Models and Evaluation, International Journal of Data Science and Analytics, Vol. 4, No.2, pp. 99-112, Springer,2017, 査読有.
DOI: 10.1007/s41060-017-0063-0
② Satoshi Oyama, Yukino Baba, Ikki Ohmukai, Hiroaki Dokoshi, and Hisashi Kashima: Crowdsourcing Chart Digitizer: Task Design and Quality Control for Making Legacy Open Data Machine-Readable, International Journal of Data Science and Analytics, Vol. 2, No. 1, pp. 45-60, 2016, 査読有.
DOI: 10.1007/s41060-016-0025-y

〔学会発表〕（計 8 件）

① Jing Song, Satoshi Oyama, and Masahito Kurihara: Identification of Possible Common Causes by Intrinsic Dimension Estimation, 2019 IEEE International Conference on Big Data and Smart Computing (IEEE BigComp 2019), pp. 2019, 査読有（Acceptance rate for

regular papers: 22.94%）.
DOI: 10.1109/BIGCOMP.2019.8679343

② Jing Song, Satoshi Oyama, and Masahito Kurihara: A Framework for Crowd-based Causal Analysis of Open Data, 2018 IEEE International Conference on Systems, Man, and Cybernetics (IEEE SMC 2018), pp. 2188-2193, 2018, 査読有（Acceptance rate: 57.40%）.
DOI: 10.1109/SMC.2018.00376

③ Jing Song, Satoshi Oyama, and Masahito Kurihara: Application of Intrinsic Dimension Estimation in Confounder Identification, 情報処理北海道シンポジウム 2017, 2017, 査読無.

④ Jing Song, Satoshi Oyama, Haruhiko Sato, and Masahito Kurihara: Evaluation of Causal Discovery Models in Bivariate Case Using Real World Data, 2016 IAENG International Conference on Data Mining and Applications (ICDMA 2016), pp. 291-296, 2016, 査読有（Acceptance rate: 50.39%）, Best Student Paper Award.

⑤ Satoshi Oyama, Yukino Baba, Ikki Ohmukai, Hiroaki Dokoshi, and Hisashi Kashima: Making Legacy Open Data Machine Readable by Crowdsourcing, Third AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2015), 2015, 査読有.

⑥ Satoshi Oyama, Yukino Baba, Ikki Ohmukai, Hiroaki Dokoshi, and Hisashi Kashima: From One Star to Three Stars: Upgrading Legacy Open Data Using Crowdsourcing, 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA 2015), 2015, 査読有（Acceptance rate for full papers: 12%）.
DOI: 10.1109/DSAA.2015.7344801

⑦ Jing Song, Satoshi Oyama, Haruhiko Sato, and Masahito Kurihara: Is the Direction with Biggest Covariance Partial to the Cause? 情報処理北海道シンポジウム 2015, 2015, 査読無.

⑧ Hisashi Kashima, Satoshi Oyama, and Yukino Baba: Crowdsourcing for Big Data Analytics, 19th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2015), 2015.チュートリアル講演.

〔図書〕（計 1 件）

① 鹿島 久嗣, 小山 聡, 馬場 雪乃: ヒューマンコンピュテーションとクラウドソーシング, 講談社, 117 ページ, 2016.

６．研究組織

(1)研究協力者
研究協力者氏名：宋 静
ローマ字氏名：SONG, Jing