学位論文審査の要旨

博士の専攻分野の名称　博士 (情報科学)　　氏名　宋　静

審査担当者　主　査　准教授　小山 聡
副　査　教　授　栗原 正仁
副　査　教　授　山本 雅人
副　査　教　授　川村 秀憲
副　査　教　授　小野 哲雄

学位論文題名

A Study on Causal Discovery Considering Confounders
（交絡因子を考慮した因果発見に関する研究）

今日、大量のデータが世の中に蓄積されており、それらを意思決定に利用することが期待されている。相関関係と因果関係は異なり、意思決定を行う際には因果関係の特定が必要な場合が多い。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、そのような実験は実際には困難であることが多い。これまでの、観測データから因果関係を特定する研究の多くでは、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。一方今日では、インターネット上でデータを公開し、それらを関連付けて利用するオープンデータの考えが重視されている。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り、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解析が必要となる。

本論文は交絡因子を考慮しながら因果発見を行う方法についての研究をまとめたものであり、大きく三つの貢献が認められる。貢献の一つは、未知の交絡因子の存在が既存の因果発見の方法に与える影響を体系的に分析し評価したことである。そこでは、異なる決定率の影響やアルゴリズムの効率性といった、従来考慮されていなかった側面に着目し、シミュレーションモデルを提案して体系的な分析を可能にしている。二つ目の貢献は、データの内在次元推定の方法を用いて、交絡因子の可能性がある変数を特定する方法を提案したことである。これは、フラクタル次元というデータの幾何学的な性質を因果発見に適用した初めての研究であり、データの関数形やノイズの分布についての仮定を必要としない。三つ目の貢献は、オープンデータ環境で探索的に因果発見を行うフレームワークを提案したことである。これにより、相関があるが因果関係が分からない変数の組に対して、交絡因子の可能性がある変数をオーブンデータから取得し、分析を行うことが可能となる。

本論文の構成は以下のとおりである。

第 1 章では、本研究の背景と目的が述べられており、因果発見において交絡因子の影響を考慮することの重要性が論じられている。

第 2 章では、以降の章の理解に必要な基礎知識と関連研究について記述されている。

第 3 章では、二変数に対する既存の因果発見モデルを評価する方法について述べられている。まず、評価対象とする ANM モデル、PNL モデル、IGCI モデルと実験で用いる CauseEffectPairs(CEP) データセットについて説明されている。その後、異なる決定率でモデルの予測制度を比較し総合的

な比較を行っている。さらに、各モデルの効率性を、決定にかかる時間の観点から比較している。最後に、未知の交絡因子が各モデルに与える影響を、データの生成モデルを用いたシミュレーションと、CEP データから取得した実データによって評価している。

第 4 章では、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ている。まず、データの内在次元を推定する方法が紹介され、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 になることが示されている。次に、シミュレーション実験が行われ、データの関数形やノイズの種類を変えて評価が行われている。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が行われ、提案手法が共通原因の候補を特定できることが示されている。

第 5 章では、オープンデータ環境において探索的な因果発見を行うためのフレームワークが提案されている。まずフレームワークの概要が述べられ、その後、クラウドソーシングを用いた因果関係に関する説明文の生成、自然言語処理を用いた変数の候補となるキーワードの抽出、取得したデータに対する因果分析といった、フレームワークの各要素について説明されている。次に、世界銀行および日本政府のオープンデータを対象とした実験が示され、キーワード抽出や因果分析の精度について評価が行われている。最後に得られた因果関係の候補について議論されている。

第 6 章では、本論文の結果が総括され、今後の研究課題が議論されている。

これを要するに、著者は、複数の因果分析モデルへの交絡因子の影響について体系的な分析を行い、内在次元推定に基づき共通原因候補を特定する新たな方法を提案するとともに、これらを用いて、オープンデータ環境で交絡因子を考慮しながら探索的に因果発見を行うフレームワークを設計したものであり、データ科学に対して貢献するところ大なるものがある。よって著者は、博士 (情報科学) の学位を授与される資格あるものと認める。