HOKKAIDO UNIVERSITY

Title	A Study on Causal Discovery Considering Confounders [an abstract of dissertation and a summary of dissertation review]
Author(s)	宋, 静
Citation	北海道大学. 博士(情報科学) 甲第13511号
Issue Date	2019-03-25
Doc URL	http://hdl.handle.net/2115/74095
Rights(URL)	https://creativecommons.org/licenses/by-nc-sa/4.0/
Type	theses (doctoral - abstract and summary of review)
Additional Information	There are other files related to this item in HUSCAP. Check the above URL.
File Information	Jing_Song_abstract.pdf (論文内容の要旨)

Instructions for use

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

学　位　論　文　内　容　の　要　旨

博士の専攻分野の名称　　博士（情報科学）　　氏名　宋　静

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）

There are a lot of observational data in the real world in which many variables are correlated with each other. Correlation is not equal to causality. The best way to demonstrate a causal relationship between variables is to conduct a controlled randomized experiment. However, real-world experiments are often expensive, unethical, or even impossible. Many researchers working in various fields are thus using statistical methods to analyze causal relationships between variables. Many studies have been conducted to infer causality from raw observational data, but most of them have been based on the assumption that all the variables (including confounders) affecting the causal relationships have been known. Today, however, emphasis has been placed on open data. In the open data environment, it is difficult to consider all related data beforehand, and an exploratory analysis is required to acquire data that can be confounding. Therefore, in this study, first, we analyzed how the existing methods which determines the causal direction between variables are influenced by unknown confounders. Through assessing the existing methods, we found that the existing methods are susceptible to confounding in different degrees. We thus investigated how to decide whether a third variable is confounding for two observed variables. Finally, We studied on a framework to perform causal analysis while considering the possible confounders. We have three purposes for the study. Firstly, investigating a general assessment method for causal discovery methods, especially investigating their performance when the data is confounded. Secondly, investigating how to determine a possible common cause variable. Thirdly, investigating how to do causal analysis of open data while considering the possible confounders. The dissertation is organized as the following:

In Chapter 1, introduction and research purposes are stated.

In Chapter 2, the background of our research, including the definition of causality, the causal Bayesian network, the task of cause effect pairs and human computation are discussed. Some related work about causal analysis and human computation is introduced as well.

In Chapter 3, we gave an assessment of three existing causal discovery models: the additive-noise model (ANM), the post nonlinear (PNL) model, and the information geometric causal inference (IGCI) model. The ANM was learned by performing Gaussian processes for machine learning regression and tested the independence between the assumed cause and residuals using the Hilbert-Schmidt Independence Criterion (HSIC). The direction with the greater independence was determined to be the true causal direction. The PNL model was learned by using a particular type of constrained nonlinear ICA to extract the assumed cause and noise. The independence between the assumed cause and noise was tested in the two directions using HSIC. The direction with the greater independence was decided to be the true one. The IGCI decides the true causal direction based on the assumption of the complexity

loss between cause and effect. There are two applicable and explicit forms for IGCI: entropy-based methods and slope based methods. To give a relatively all-sided evaluation of the existing methods, we used three evaluation metrics: accuracy for different decision rates, area under ROC curve (AUC) and algorithm efficiency. Besides, we tested how the three models responded to spurious correlation caused by confounding using simulated and real world data. The experimental results showed that the existing methods are susceptible to confounding in different degrees.

In Chapter 4, we proposed using intrinsic dimension estimation as a necessary condition to determine a possible common cause for two variables. Simulated application showed that the proposed method worked well for both linear and non-linear functions. Testing using different types of noise showed that it generally worked well for different types of added noise. In particular, it worked better than a kernel-based conditional independence test for Poisson noise. Testing of how the estimated intrinsic dimension is affected by different types of distributions showed that the estimated dimension is nearly not affected by the type of distribution. Simulation of mixed pattern showed that the proposed method can still tell a possible common cause when it is mixed with causal relationship. Finally, experiments using variables from the CauseEffectPairs dataset showed that the proposed method can give correct inferred results for real world data.

In Chapter 5, we proposed a framework for exploratory causal analysis of open data. It contains three main parts: collecting explanations of correlation in open data using crowdsourcing market place, extracting keywords using natural language processing (NLP) methods and verification of the collected assumption using machine learning methods. The explanations were collected using general crowdsourcing platform. The topic words are learned using natural language processing methods. According to the extracted words, related variables were obtained. The data about related words was searched in the open data. Finally, the causal relationship was analyzed using machine learning methods. We did comparison experiments with causal discovery methods: PNL and BMLiNGAM. The PNL works under the no confounding assumption and the BMLiNGAM works under the confounding assumption.

In Chapter 6, we concluded our research and discussed our future work.