How useful is causal discovery when making predictions?

MATH-516 Applied Statistics

Linda Mhalla

2024-04-22

Introduction: Local causal discovery

Local causal discovery methods focus on identifying causal relationships within a specific subset of variables in complex systems
$\rightarrow$ targets localized areas, reducing complexity and computational demands

Key Methods:

- Constraint-based approaches: utilize conditional independence tests within a confined set of variables, e.g., the PC algorithm applied locally
- Score-based approaches: evaluate a subset of models based on a scoring criterion that measures how well the model fits the data for a particular subset of variables

Applications:

- Biological Systems: particularly useful in studying gene regulatory networks where full network analysis is unfeasible
- Economic Models: applied to study localized economic interactions without needing to model the entire economic system

Introduction: Local causal discovery

For a target variable T, local causal discovery methods aim at learning its *Markov blanket*, i.e., the direct causes (parents), direct effects (children), and spouses (direct causes of the direct effects) of T, yielding its "neighbourhood" in the causal DAG

$\Rightarrow$ not interested in the causal DAG of the entire system of $(T, \mathbf{X})$ but only on the causal DAG around T

Popular constraint-based methods:

- IAMB (Incremental Association Markov Blanket) (Tsamardinos et al., 2003): incrementally adds and then removes variables to find the Markov blanket, using conditional independence tests
- HITON-PC (Aliferis et al. 2003): identifies the Markov blanket of a target variable using conditional independence tests

Introduction: Local causal discovery

Popular constraint-based methods:

- Max-Min Markov Blanket (mmmb) (Tsamardinos et al., 2003): forward selection on a maximum-minimum conditional dependence on T + backward elimination

Note: All rely on the faithfulness condition

Data and aim

LUCAS0 data: artificially generated **lung cancer data** set for studying causal discovery

- along with the binary variable `Lung_cancer`, 11 binary features/covariates are available, e.g., `Smoking`, `Anxiety`, `Allergy`, etc

Pre-defined training and test data provided in repository for comparability (`lucas_train.csv` and `lucas_test.csv`)

The goal of this project is to investigate

- *how stable (local) causal discovery is and*
- *how far it can pay off also in terms of prediction accuracy*

Tasks

To model the effect of the other variables given for each patient/control in the data set on lung cancer (target variable T), perform classification using logistic regression, where – instead of using all of the 11 variables – you reduce the set of features/covariates by

1. performing **standard variable selection**, such as, e.g., by using a L_1 penalty (LASSO)
2. using the **PC algorithm** for causal discovery and selecting only variables in the Markov blanket around T
3. using **a local causal discovery method** (such as implemented in the `R` packages `bnlearn` and `MXM`) to get the Markov blanket around T and keeping only the corresponding variables
4. selecting **the true variables** `Smoking` , `Genetics`, `Fatigue`, `Coughing` and `Allergy` in the actual Markov blanket around T

and compare the prediction performance on the test set

Tasks

Do this

a) for **different samples sizes** ("small", "medium", and "large" corresponding to different shares of the provided training set)

b) discussing also the **stability of the (local) causal discovery**, and

c) adding a **brief description of the chosen local discovery method** to the report