

A data repository for causal inference from observational data

CAUSE

CAUSE a repository for causal inference benchmark datasets: the repository contains variety of datasets for different causal inference tasks.

CAUSE also contains datasets for application of causal inference in machine learning along with reliable groundtruths.

CAUSE provides novel techniques to validate causal features when groundtruth is not available.

Abstract

Traditional supervised learning algorithms are designed to make predictions given existing observational data by capturing existing correlations and dependencies from the data. However, decision making tasks often require discovering underlying causal mechanisms. Therefore, to make decisions from observational data, one needs to design a system that capture causal relationships instead of the correlations.

Existing benchmark datasets for supervised learning do not contain groundtruths for causal inference tasks and therefore can not be leveraged.

CAUSE is a data repository which contains various existing benchmark datasets for different causal inference tasks with reliable groundtruths. Moreover, we propose novel techniques to evaluate causal features and relationships learned from the data when no groundtruth is available.

CAUSE consists of datasets for causal discovery, causal effect estimation and application of causal inference in machine learning such as recommendation systems and reinforcement learning.

References

Pearl, J., & Bareinboim, E. (2011, Dec). Transportability of causal and statistical relations: A formal approach. In *2011 IEEE 11th International Conference on Data Mining Workshops* (p. 540-547). doi: 10.1109/ICDMW.2011.169

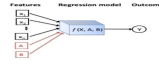


Traditional machine learning algorithms are optimized to perform prediction using available data and are not suitable to answer questions such as "How does type of school affect a child's achievements in feature?" which is the question typically asked for policy or decision making tasks. In order to answer these types of questions, models with the ability of discovering and exploring causal relationships amongst variables in the data are required. However, to verify the performance of these models, we need to come up with benchmark datasets designed for these tasks. This task is particularly challenging due to the following reasons: 1) Obtaining causal structure of variables in the data is a difficult task specially in the case of high dimensional data. 2) Only factual samples can be observed from the data and no information is available about counterfactuals.

In this paper we introduce **CAUSE**, a data repository that contains existing benchmark datasets for causal inference tasks. Moreover, to address the first challenge, we propose two novel techniques to obtain and evaluate causal features from the data when groundtruth is not available:



- **Translation:** we propose to translate our problem into a new domain for which groundtruth is available. We then use the available data to validate the performance of our model for the new problem and use the results to validate our model for the original task.
- **Relaxation:** Based on the theory of transportability (Pearl & Bareinboim, 2011), classifiers that capture causal relationships are more robust to different datasets.

CAUSE: A Data Repository for Causal Inference

Raha Moraffah, Lu Cheng, Ruocheng Guo, K Seluk Candan, Huan Liu
Arizona State University
{rmoraffa, huan.liu}@asu.edu

<h4>Traditional Supervised Learning vs Causal Inference</h4> <ul style="list-style-type: none">How does type of school affect child's achievements in feature?Traditional supervised learning  <ul style="list-style-type: none">These methods are designed to predict the outcome and not the effect of types of schools	<h4>Correlation vs Causation</h4> <ul style="list-style-type: none">From observational data, School's type (T) and outcome (Y) are correlated. But neither of them might cause the other oneObserved correlation could be due to the child's features 
<h4>Challenges and desiderata of causal inference</h4> <p>Challenges:</p> <ul style="list-style-type: none">Causal structure of variables is unknownNot all variables are confoundersCounterfactuals can not be observed in observational data <p>Example:</p> <p>Factual: John Went to X school and his college gpa is A.</p> <p>Counterfactual: What would have been John's gpa, had he not gone to X school?</p> <p>Desiderata:</p> <ul style="list-style-type: none">Overlap assumption: Common supportIgnorability assumption: No unmeasured confounder	<h4>Causal feature evaluation: datasets and methodologies</h4> <p>Methodologies:</p> <ul style="list-style-type: none">Available datasets for causal discovery:<ul style="list-style-type: none">Not available for all tasksExisting datasets are smallMapping the problem to another domain and use existing datasets:  <ul style="list-style-type: none">Relaxing the problem by entering strong assumptions:<ul style="list-style-type: none">Based on the theory of transportability, causal relationships are more robustcompare the performance of classifiers with different interventional distributions

This material is based upon work supported by, or in part by, the X grant Y.

Data Mining and Machine Learning LabARIZONA STATE UNIVERSITY

A detailed explanation of all datasets available in **CAUSE** repository along with the description of the tasks they are designed for is available in the attached draft paper.