Enterprise Platform and Integration Concepts: Research Group of Prof. Dr. Hasso Plattner

Master Project Winter Term 2020/21

A Benchmark Suite for Causal Inference or *"How to model a complex world?"*

Motivation

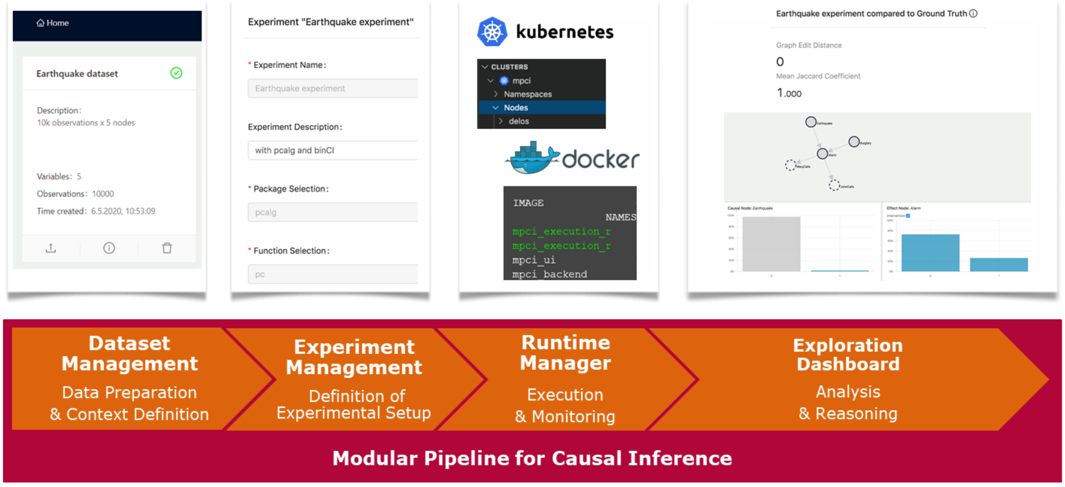
The questions that motivate most data analyses in an enterprise context are of causal nature, e.g., what are the causes and effects of events in manufacturing processes under observation? Nevertheless, the statistical methods commonly used today to answer those questions are of associational nature. But "Correlation does not imply causation!" , and misinterpretation often results in incorrect deduction [1]. For example, a constant conjunction of a raw material and an error in many past instances of a manufacturing process may let you falsely assume that a causal relationship exists. Hence, the avoidance of this raw material may not affect the error.

Based upon causal graphical models, the mathematical theory of causal calculus for modeling interventions introduces a framework for causal inference from observational data [2]. The variety of implemented methods for causal inference and the complexity of data characteristics in real-world scenarios introduces the demand of a modular pipeline that enables to plug in existing methods from different libraries into a single system to compare and evaluate the results, also under the aspect of improving existing algorithms.

In complex real-world settings, underlying causal structures are mostly unknown such that evaluation of methods for causal inference is difficult. Therefore, such a pipeline should incorporate a benchmark suite to generate synthetic data from known causal structures to enable a comprehensive evaluation. Hence, it remains to answer "How to model a complex world?" .

Framework

In this master project, you will extend an existing modular pipeline (see Figure 1: A schematic overview of the current state of the modular pipeline) that enables the application of machine learning techniques for causal inference in a real-world context. In order to provide data scientists a framework that ensures the comparability and reproducibility of experiments, you are going to understand and synthetically model data characteristics omnipresent in real-world scenarios.



Earthquake dataset

Description: 10k observations x 5 nodes

Variables: 5

Observations: 10000

Time created: 6.5.2020, 10:53:09

Experiment "Earthquake experiment"

* Experiment Name: Earthquake experiment

Experiment Description: with pcalg and binCI

* Package Selection: pcalg

* Function Selection: pc

```
IMAGE
                NAMES
mpci_execution_r
mpci_execution_r
mpci_ui
mpci_backend
```

Earthquake experiment compared to Ground Truth

Graph Edit Distance: 0

Mean Jaccard Coefficient: 1.000

Figure 1: A schematic overview of the current state of the modular pipeline

Therefore, you will deep dive into the core engine and extend the pipeline with a data generation component. You will evaluate existing methods for causal structure learning and causal calculus in these *"complex world models"* . As part of the journey, you will have the opportunity to deepen your knowledge about tools for data science, improve machine learning skills, and influence the performance of an end-to-end pipeline for causal inference.

Project Goals

Several goals can be addressed during the course of this project, depending on the number of participants, interests, expertise, and progress during the project

1. Deep dive into concepts of causal inference, in particular causal structure learning
2. Understand requirements on causal structures in real-world settings and how to evaluate methods of causal inference
3. Extend an existing tool for data scientists with a benchmark suite, to enable a comparison in different synthetic experimental settings with regards to performance and quality
4. Analyze methods for causal structure learning and causal inference to derive key data characteristics in your synthetic benchmark suite

Technology & Skills

The core of the work is based upon an existing pipeline for causal inference and previously developed extensions (e.g., using GPUs or independence tests for heterogeneous data) of the causal structure learning algorithms. Prior understanding of the fundamentals of machine learning techniques (e.g., having attended the lecture "Causal Inference – Theory and Applications in Enterprise Computing" , "Computational Statistics" or equivalent) is recommended as well as knowledge in one of the following areas (Python, NodeJS, R, Kubernetes).

References

[1] Matthews, Robert. "Storks deliver babies (p= 0.008)." Teaching Statistics 22.2 (2000): 36-38.

[2] Pearl, Judea. "Causal inference in statistics: An overview." Statistics surveys 3 (2009): 96-146.

Organization

- Programming project including design, implementation, test and documentation
- Group work
- Interim and final presentations

Contact

You are welcome to contact us via email:

Christopher Hagedorn (Christopher.Hagedorn@hpi.de),

Johannes Huegle (Johannes.Huegle@hpi.de),

Dr. Michael Perscheid (Michael.Perscheid@hpi.de)