Benchmarking Causal Discovery Algorithms on Medical Data

Giuseppe Concialdi
gconci2@uic.edu

Vaibhav Jolly
vjolly2@uic.edu

Aditya Ranganathan
aranga22@uic.edu

Judhajit Roy
jroy20@uic.edu

1 INTRODUCTION

Causal discovery is the task of identifying the causal relationships among a set of variables. Causal networks are represented as graphs in which each variable constitutes a vertex and the causal connections are depicted as directed edges. Causal discovery algorithms allow modeling the underlying causal relationships between variables by making assumptions on the nature of the data [4]. The output of these algorithms is a structural causal model that makes it possible to perform causal reasoning. This is of paramount importance in several fields like marketing, biology, social sciences, finance, and many more. The causal framework has recently gained more attention in the machine learning domain, where causality could be beneficial to improve the quality and robustness of the predictions [7]. Additionally, causal models are useful for counterfactual inference. A change in the algorithm can be seen as an intervention [5]. Instead, with purely observational data it is often impractical to perform interventions or manipulation. The best way to assess the effect of intervention would be to perform an A/B test but this is both expensive and time-consuming.

2 PROBLEM DESCRIPTION

In our research, we will focus our attention on discovering causal relations in medical-related datasets. In this field, causal reasoning is crucial to alleviate the problem of data scarcity[1]. The lack of high-quality annotated datasets and the problems related to the collection and availability of these information makes it difficult to find large datasets of this kind. Furthermore, when analyzing medical data, it is pivotal to understand the relationships among the analyzed variables. We will compare several causal discovery algorithms on medical datasets and assess the performance and the quality of the outcomes.

3 DATASETS

We will focus our analysis on the following datasets:

- **Wisconsin breast cancer diagnostic dataset** (WBCD), consisting of 569 datapoints and 32 attributes [8].
- **Thyroid dataset** [6] consisting of 9,172 datapoints and 31 attributes.

Both dataset are from the UCI repository [2]. The features in WBCD describe the characteristics of the cell nuclei present in the image, while the features in the Thyroid dataset consist of data points merged from both hyperthyroidism and hypothyroidism databases. Both datasets are in the form of CSV files. The WBCD dataset features include the *ID* number and the categorical column *Diagnosis*, which consists of 357 Benign instances and 212 Malignant instances. The dataset also consists of real-valued features computed for each cell nuclei including *radius*, *texture*, *perimeter*, etc. The Thyroid dataset consists of features categorized under different factors such as hyperthyroid and hypothyroid conditions, therapy, and treatment and includes *age*, *sex*, *sick*, *pregnant*, *tumor*, *TSH*, *T3*, *T4*, etc. Furthermore, we will also be generating synthetic datasets to the ensure presence of counterfactuals within the data.

4 RESEARCH PLAN

Our research will be conducted in two stages. In the first stage, we will start with the baseline implementation of the causal discovery algorithms. In the second stage, the performances of the algorithms will be assessed and compared for each dataset.

4.1 Initial approaches

Our plan is to compare several causal discovery algorithms on the given medical datasets. These algorithms can be categorized into various classes of causal discovery algorithms like constraint-based methods, score-based methods, hybrid methods, and more. One of the differentiating factors of the algorithms is the underlying assumptions made. These assumptions could be *acyclicity*, *causal sufficiency*, *causal faithfulness*, etc. [4]. We want to start experimenting with some of the standard algorithms inclusive of PC, GES, and LINGAM as well as a recent algorithm NOTEARS[10] as our initial approach. The algorithms will be implemented on the original and synthetically generated datasets to identify causal relationships and build a structural graph to represent the relations between the nodes/variables in the graph.
We plan to evaluate the performance based on different metrics such as Area Over the Curve (AOC), Equal-Error-Rate (E-ER), No-False-Positives-Error-Rate (NFP-ER), No-False-Negatives-Error-Rate (NFN-ER) metrics [4]. All these metrics are based on the True Positive Rate (TPR) and False Positive Rate (FPR).

4.2 Desired outcomes

After implementing the initial algorithms, we move on to implement some of the recent algorithms like *DAG-graph neural networks* (DAG-GNN) [9], and *causal generative neural networks* (CGNN) [3]. The project aims to provide an unbiased evaluation of all the algorithms, by benchmarking their performance and comparing them on multiple datasets, both real and carefully designed synthetic datasets.

REFERENCES

[1] Daniel C Castro, Ian Walker, and Ben Glocker. 2020. Causality matters in medical imaging. *Nature Communications* 11, 1 (2020), 1–10.
[2] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml
[3] Olivier Goudet, Diviyan Kalainathan, Philippe Caillou, Isabelle Guyon, David Lopez-Paz, and Michèle Sebag. 2017. Causal generative neural networks. *arXiv preprint arXiv:1711.08936* (2017).

[4] Christina Heinze-Deml, Marloes H Maathuis, and Nicolai Meinshausen. 2018. Causal structure learning. *Annual Review of Statistics and Its Application* 5 (2018), 371–391.

[5] Karamjit Singh, Garima Gupta, Vartika Tewari, and Gautam Shroff. 2018. Comparative benchmarking of causal discovery algorithms. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*. 46–56.

[6] Feyzullah Temurtas. 2009. A comparative study on thyroid disease diagnosis using neural networks. *Expert Systems with Applications* 36, 1 (2009), 944–949.

[7] Matthew J Vowels, Necati Cihan Camgoz, and Richard Bowden. 2021. D'ya like DAGs? A survey on structure learning and causal discovery. *ACM Computing Surveys (CSUR)* (2021).

[8] William H Wolberg, W Nick Street, and Olvi L Mangasarian. 1992. Breast cancer Wisconsin (diagnostic) data set. *UCI Machine Learning Repository [http://archive.ics. uci. edu/ml/]* (1992).

[9] Yue Yu, Jie Chen, Tian Gao, and Mo Yu. 2019. DAG-GNN: DAG structure learning with graph neural networks. In *International Conference on Machine Learning*. PMLR, 7154–7163.

[10] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. 2018. Dags with no tears: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems* 31 (2018).