Benchmarking Causal Discovery Algorithms on Medical Data - Progress Report

Giuseppe Concialdi
gconci2@uic.edu

Judhajit Roy
jroy20@uic.edu

Aditya Ranganathan
aranga22@uic.edu

Vaibhav Jolly
vjolly2@uic.edu

ABSTRACT

Causal Discovery or Causal Structure Discovery is the problem of identifying causal relationships from given data. Our study aims to examine the workings of various Causal Discovery algorithms like PC, GES, LiNGAM and NOTEARS. We compare the causal structures generated by each of the algorithms on the Wisconsin breast Cancer Dataset, and the Thyroid dataset from the UCI repository, and evaluate the performance of the structures using evaluation metrics like ROC and AUC.

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 modelling the underlying causal relationships between variables by making assumptions on the nature of the data [5]. 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 [15]. Additionally, causal models are useful for counterfactual inference. A change in the algorithm can be seen as an intervention [10]. 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 RELATED WORK

Experiments using synthetic and real data are the fundamental component of evaluating causal discovery algorithms. Synthetic data are typically drawn from graph structures that were created at random or by models created specifically for this task. While such synthetic data experiments can demonstrate the increased quality of suggested methodologies, they occasionally misrepresent the difficulties that arise in real-world situations. However, there aren't many real-world datasets available for testing causal discovery methods. Few causal discovery methods can be evaluated using the ground truth if available in the data, like the cause-effect pairs in [6]. However, usually labelled data is expensive or unavailable in a real-life scenario. Datasets that contain only pair-wise data are not complex enough to evaluate real causal discovery analysis. Other health-care datasets like in [13] contain causal relations among multiple variables and are commonly used for the evaluation; however, few pairs of ground-truth causal relations are known/labelled by domain experts and the evaluation is not systematic [14]. Therefore, it is necessary to develop causal discovery benchmarks for real-world evaluation.

Generally, causal discovery requires controlled experiments employing interventions but in many cases, this is too expensive, unethical or impossible to perform [10]. In this framework, it is crucial to develop methods that allow the identification of causal relationships from purely observational data. There exist several classes of causal discovery algorithms based on the assumptions made on top of the data.

Constraint-based. These methods learn causal Bayesian networks with conditional independence tests by analyzing the probabilistic relations exploiting the Markov assumption of the network. **PC** [11] is the most popular algorithm of this kind and it assumes acyclicity, causal sufficiency and causal faithfulness. It is an iterative algorithm that starts with a complete undirected graph and performs several conditional independence tests to learn the structure of the underlying DAG.

Score-based. These algorithms assign a score to each candidate network for measuring the quality of the fitness with the dataset. Greedy equivalence search (**GES**) algorithm [2] is a prominent example of score-based algorithms. It makes the same assumptions as PC and employs a penalized likelihood score. It starts from an empty graph and its greedy approach selects, at each iteration, the edge that yields the maximum improvement on the score.

Linear Non-Gaussian Acyclic Model. This model assumes acyclicity and causal sufficiency. **LiNGAM** [8] employs a technique called independent component analysis and the assumption of non-Gaussianity of the disturbance variables, together with the assumption of linearity and causal sufficiency, allows the causal model to be completely identified.

Acyclicity penalty. The recent method DAGs with **NO TEARS** (Non-combinatoric Optimization via Trace Exponential Augmented lagRangian Structure learning) [18] rethink the combinatoric graph search problem as a continuous optimization problem. The acyclic penalty enforces acyclicity as a constraint of the optimization.

3 CAUSAL DISCOVERY

3.1 Formal 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.

3.2 Data description and cleaning

We focused our analysis on the following datasets:

- **Wisconsin breast cancer diagnostic dataset** (WBCD), consisting of 569 datapoints and 32 attributes [16].
- **Thyroid dataset** [12] consisting of 9,172 datapoints and 31 attributes.

Both datasets are from the UCI repository [3]. 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.

WBCD preprocessing. To perform the causal discovery on the selected dataset, it is necessary that all the features are represented as numerical values. Several causal discovery algorithms cannot handle categorical data, therefore the preprocessing of these attributes is needed. The preprocessing of the WBCD was fairly straightforward: all the variables are encoded as numerical values except for the *diagnosis* attribute. This categorical variable employs the string "M" to represent a malignant tumor and the value "B" to encode the presence of a benign tumor. We performed a simple binarization of these features by mapping "M" $\rightarrow 1$ and "B" $\rightarrow 0$. Additionally, we found out that the column *Unnamed: 32* does not contain any value, therefore, we dropped this feature.

Thyroid dataset preprocessing. The thyroid dataset has many categorical attributes encoded as strings, therefore it requires several steps of preprocessing. Firstly, we collected all the variables that have only two levels. These variables can be easily binarized by performing an arbitrary mapping of the categories to $\{0, 1\}$. Several variables are indicators for the presence or the absence of a specific feature. These variables encode the presence with "t" (true) and the absence with "f" (false). The attribute *sex* employs "F" for females and "M" for males. Finally, the feature *binaryClass* has the category "N" for negative and "P" for positive. All these attributes were binarized. However, the dataset contains the attribute *referral source* that has more than two levels. There are several techniques to deal with categorical attributes, like one-hot encoding or binary encoding. For the scope of our task, we decided to simply encode this variable with distinct integer values. We performed the following mapping: "SVHC" $\rightarrow 0$, "SVI" $\rightarrow 1$, "STMW" $\rightarrow 2$, "SVHD" $\rightarrow 3$, "other" $\rightarrow -1$. Finally, we deleted both the variables *TBG* and *TBG measured* because they did not contain any observation.

3.3 Description of initial solutions

As mentioned, our main task is to compare several causal discovery algorithms on medical datasets and evaluate the model's performance against the ground truth. We've handpicked a few different algorithms to run the experiment with. These include **PC(Peter-Clark)**, **GES()**, **LINGAM** as well as a recent algorithm **NOTEARS**[19]. PC[7] and LiNGAM[9] are constraint-based structure learning algorithms that assume there are no hidden confounders. The PC algorithm identifies all the causal graphs that are consistent with the available data and conditional dependencies, while the LiNGAM model extends upon this by determining a unique DAG for a given dataset. The assumptions considered for PC are:

- **No assumptions are made on the distribution of the data**
- **There are no unobserved confounders for any variables in the graph**
- **Acyclicity**
- **Requires the faithfulness assumption**

while the assumptions considered for LiNGAM are:

- **Linearity**
- **Acyclicity**
- **No hidden common causes**
- **Non-Gaussian continuous error variables (except at most one)**

We implement the PC algorithm using the casual-learn library and the DirectLiNGAM method from the LiNGAM library. The first step of PC is to estimate the skeleton of the DAG, and then for each edge, the constraint is tested. If a separation set is found, the edge is then deleted. The next step is to orient the unshielded triples (3 nodes a, b, and c with ab, bc but a and c are not connected.) If node b is not in the separation set(a,c), the unshielded triples abc is oriented into an unshielded collider a $\rightarrow$ b $\leftarrow$ c (v-structure), else b is marked as a non-collider on abc. In the third step, the partially directed graph is checked to see if further edges can be oriented while avoiding new unshielded colliders.
DirectLiNGAM improves upon the ICA-LiNGAM by converging to the right solution in a fixed number of steps equal to the number of variables if the data strictly follows the model.
GES is a score-based structure learning algorithm that assumes there are no hidden confounders. In the first step, the GES algorithm begins with an empty CPDAG (Completed Partially Directed Acyclic Graph) and greedily adds edges in the 'forward phase'. The addition of the edges is based on the fit measured by the score function, and the edge that most improves fit is added to maximize the increase in score, till it can't be further increased. This entails the calculation of the DAG from the CPDAG, adding directions to the edge, and calculating the CPDAG from the new DAG. In the 'backward phase', once the score can no longer increase, the edges are greedily eliminated until an optimum score is obtained.
The goal of Causal Discovery is to estimate the directed acyclic graphs (DAGs), which is a challenging problem as the analysis is to be done in a combinatorial search space that is of the size of the number of nodes. NOTEARS simplify this problem by circumventing the combinatorial paradigm and formulating it as a continuous optimization problem on 2D vectors that scale cubically. We implemented NOTEARS on the two datasets using the Causalnex library.
In the analysis for the Breast Cancer dataset, on the cleaned we first normalize the features after min-max scaling them. Then, we

generate the structural model based on these features, considering "diagnosis" as the target/outcome. On the structure, we perform edge pruning to remove all edges with weight below the defined threshold. This reduces the complexity of my graph and avoids false positive connections. Once the structural model is finalized we use it to create a Bayesian Network. Bayesian Network is a directed acyclic graph (DAG), consisting of nodes that represent random variables and edges that depict the causal connections between variables, in a conditional probability distribution. Since the algorithm works only on discrete probability distributions, we had to discretize the variables using the Decision Tree Discretizer. The Decision Tree Discretizer categorizes all the variables in buckets formed by deciding the depth of the tree. Further, we evaluate the DAG using the receiver operating characteristic curve (ROC) curve and also find the Area Under the Curve (AUC). We found it was easier to evaluate the models on NOTEARS using metrics ROC and AUC, when compared to the older models that were implemented like PC, GES and DirectLiNGAM. We were able to obtain the Bayesian Information Criterion (BIC) score from the causal-learn library for the GES models that were generated. we faced issues when evaluating the models generated from the Casual Discovery Toolbox library, due to the R library connections that were required to compute the model in the backend.

4 EXPERIMENTAL RESULTS

We ran experiments on the data using the algorithms and obtained the causal graphs. The representation shows how the variables are related to the outcome. The DirectLiNGAM graph was generated on the Breast Cancer Dataset with the edges having low-threshold values being dropped. The structure shows the relations between the various features and the outcome. Based on our understanding

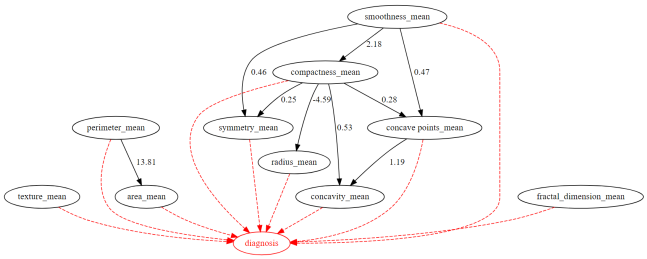

Figure 1: Causal graph using LiNGAM on WBCD

of the causal graphs, we concluded that NOTEARS gives a better representation of the causal relations between the features in the dataset. Based on the structural graph, we implemented the Bayesian Network and calculated the ROC and AUC values. The graph shows that the causal structural model was able to easily classify the outcome based on the nodes.

5 IDEAS FOR NEXT STEPS

As per suggestions, we have identified a larger and more recent dataset "Cardiovascular Disease Dataset". This is an open-source dataset on Kaggle. It contains 70,000 patient records and 11 features out of which 34,979 patients are diagnosed with cardiovascular disease and 35,021 are not. We will also be generating a synthetic

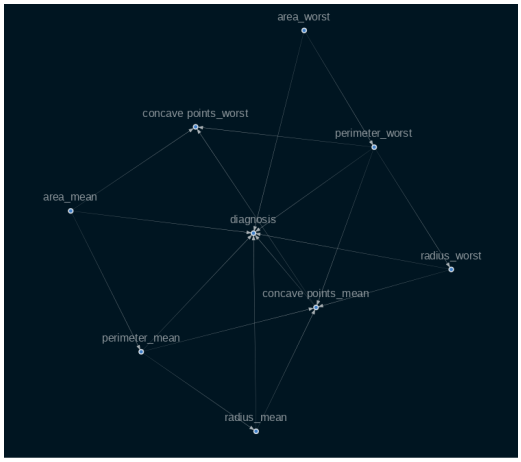

Figure 2: Causal Graph using NOTEARS on WBCD

Figure 3: ROC for WBCD Dataset using NOTEARS

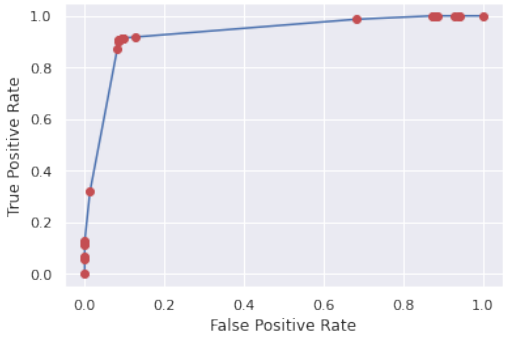

Figure 4: ROC for Thyroid Dataset using NOTEARS

dataset to compare the performance of the casual structures generated among the different algorithms. We plan to implement some more recent algorithms like *DAG-graph Neural Networks* (DAG-GNN) [17], and *Causal Generative Neural Networks* (CGNN) [4]. The aim is to provide an unbiased evaluation of the algorithms, by benchmarking their performance and comparing them on multiple 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] David Maxwell Chickering. 2002. Optimal structure identification with greedy search. *Journal of machine learning research* 3, Nov (2002), 507–554.

[3] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[4] 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).

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

[6] Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. 2016. Distinguishing cause from effect using observational data: methods and benchmarks. *The Journal of Machine Learning Research* 17, 1 (2016), 1103–1204.

[7] Chandramouli Rathnam, Sanghoon Lee, and Xia Jiang. 2017. An algorithm for direct causal learning of influences on patient outcomes. *Artificial intelligence in medicine* 75 (2017), 1–15.

[8] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. 2006. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research* 7, 10 (2006).

[9] Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, and Kenneth Bollen. 2011. DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *The Journal of Machine Learning Research* 12 (2011), 1225–1248.

[10] 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.

[11] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. 2000. *Causation, prediction, and search*. MIT press.

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

[13] Ruibo Tu, Cheng Zhang, Paul Ackermann, Karthika Mohan, Hedvig Kjellström, and Kun Zhang. 2019. Causal Discovery in the Presence of Missing Data. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 89)*, Kamalika Chaudhuri and Masashi Sugiyama (Eds.). PMLR, 1762–1770. https://proceedings.mlr.press/v89/tu19a.html

[14] Ruibo Tu, Kun Zhang, Bo Bertilson, Hedvig Kjellstrom, and Cheng Zhang. 2019. Neuropathic pain diagnosis simulator for causal discovery algorithm evaluation. *Advances in Neural Information Processing Systems* 32 (2019).

[15] 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).

[16] 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).

[17] 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.

[18] Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. 2018. DAGs with NO TEARS: Continuous Optimization for Structure Learning. https://doi.org/10.48550/ARXIV.1803.01422

[19] 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).