From cause-effect pairs to causal graphs

Margarida Freitas de Sá Mendes
margarida.sa.mendes@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

November 2023

Abstract

Causal inference is a key focus in modern machine learning, aiming to unravel complex cause-and-effect relationships in intricate systems. This pursuit encompasses various methods, broadly classified into constraint-based, score-based, and parent-child distinguishing approaches. Constraint-based methods explore causal links through conditional independence, while score-based methods employ statistical scores to assess causality strength. Parent-child distinguishing methods aim to determine the direction of causality.

This thesis offers a comprehensive analysis of causal inference, focusing on the evaluation and comparison of scoring methods, including the widely used Bayesian Information Criterion (BIC) and Bayesian Dirichlet equivalent uniform (BDeu) score, and their variants using entropy and Kullback-Leibler divergence. The objective is to assess their performance across diverse causal discovery scenarios, involving different graph structures and noise levels. By examining these methods across various graph configurations, this thesis enhances our understanding of the strengths and limitations of causal inference techniques, enabling more precise and robust causal analysis in practical applications.
Keywords: Causal inference, Score-based methods, Bayesian Information Criterion (BIC), Bayesian Dirichlet equivalent uniform (BDeu).

1. Introduction

Causality, a concept deeply embedded in the fabric of human thought, has been a subject of profound fascination and debate. Understanding and unraveling causal connections between events and phenomena have shaped our comprehension of the world. Causal relationships are of paramount importance across scientific disciplines, enabling prediction and informed decision-making. The field of causal research has transformed from a conceptual framework to a well-defined mathematical construct, advancing methods to infer causation.

Causal discovery seeks to transcend mere correlation, delving into the complexities of causation. It grapples with the fundamental challenge of inferring causality from observable data, spurred by advancements in computational techniques and the age of big data[1].

Key challenges confront causal discovery, including data limitations, confounding variables, indirect effects, feedback loops, nonlinear relationships, and the scalability of data. Limited or incomplete data can result in inaccurate causal inferences, while confounders and intricacies like feedback loops and nonlinearity can obscure direct causal links. Efficient and scalable causal discovery algorithms are imperative for navigating complex, large-scale systems.

In the pursuit of causality, statistical methods offer valuable insights into variable relationships. However, they fall short of unraveling the intricate cause-and-effect mechanisms inherent to true causal connections. Causal inference methods have emerged to address this gap, aiming to infer causal graphs from data and delve into the underlying generating mechanisms[7].

Understanding causality, causal inference, and causal models relies on key foundational concepts. Directed Acyclic Graphs (DAGs) are graphical tools that depict relationships among variables, with directed edges denoting causal or probabilistic connections. Bayesian Networks, rooted in DAGs, capture conditional dependencies and assist in probabilistic inference. The Markov Blanket, a pivotal element in Bayesian Networks, isolates the minimum set of nodes that, when observed, makes a node conditionally independent of the rest. Reichenbach's Common Cause Principle explores correlated events, emphasizing the presence of a common underlying cause when two events show a correlation. These concepts serve as a vital framework for delving deeper into the realms of causal inference and causal models.

This thesis has four main goals. First, it ex-

plores the concept of causality, digging into what causes things to happen and how we can understand it. Second, it reviews different methods that help us figure out why things happen the way they do. Third, it introduces some new ways to measure causality, adding to the tools we have. Finally, it tests these methods using real-world data to see how well they work. By doing this, the thesis helps us better understand the strengths and weaknesses of these methods and how they can be used in practical situations.

2. Background

In the world of causal inference, understanding the fundamental concepts and models that underpin causal relationships is of paramount importance. This chapter delves into the essential building blocks of causal inference, offering a comprehensive overview of Directed Acyclic Graphs (DAGs), Structural Equation Models (SEMs), the Markov Blanket, and Reichenbach's Common Cause Principle. Besides, a diverse array of causal discovery methods, including Constraint-Based Methods, Score-Based Methods, and Methods designed to distinguish parents from children, provides the tools necessary for unraveling the complexities of causality.

Directed Acyclic Graphs (DAGs): A finite family of random variables is introduced, and their joint distribution is defined as $(\mathbf{X})$. A graph $G(\mathbf{V}, E)$ consists of nodes $\mathbf{V}$ and edges $E \subseteq \mathbf{V}^2$. These nodes and edges define the parent-child relationships within the graph, forming a structure to represent dependencies and causal connections. A critical concept introduced is the "partially directed acyclic graph" (PDAG), which lays the foundation for comprehending a "directed acyclic graph" (DAG) where all edges are directed, symbolizing causal relationships [2, 5]. A practical example of a DAG is illustrated in Figure 1.

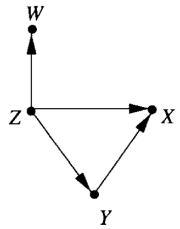

Figure 1: Example of a DAG [1]

Structural Equation Models: Structural equation models (SEMs) are explored, which consist of equations describing the relationships between variables. Each equation specifies how a variable depends on its parent variables and noise variables. The graphical representation of SEMs reveals how parent variables influence their direct effects. In the context of causal inference, the concept of "structural causal models" is introduced, where each variable is associated with a distinct equation, particularly useful for causal modeling [5].

Markov Blanket: The Markov Blanket concept is introduced, which plays a vital role in understanding the conditional independence relationships between variables in a Bayesian Network. The Markov Blanket of a variable consists of its parents, children, and the other parents of its children. Knowing a variable's Markov Blanket ensures that other variables do not provide additional information about it. This concept facilitates probabilistic modeling, causal discovery, and feature selection [2]. A visual example of a Markov Blanket is presented in Figure 2.

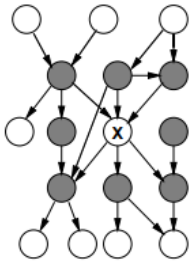

Figure 2: Example of a Markov Blanket [9].

Reichenbach's Common Cause Principle: The principle is introduced, outlining that if two variables, X and Y, are unconditionally dependent, there must be a causal connection. The principle presents three possibilities: a direct path from X to Y, from Y to X, or the existence of a common node Z that causally influences both X and Y, which aligns with the Markov property [2].

Causal Relationships and Interventions: Causal relationships are explained as connections where one variable influences or causes a change in another. The challenge of discovering these causal relationships, especially without direct experimental methods, is highlighted. Interventions are introduced as a method to simulate physical interventions using the mathematical operator "do(x)" to isolate the impact of a specific variable, allowing for causal inferences [19, 8].

The Cause-Effect Problem: This central problem is described as the task of determining whether X causes Y or vice versa when given observations from the joint distribution. The goal is to understand the causal relationship between variables in a way that can predict the effects of perturbations or actions on the system. It is crucial to address this problem without confounding variables to achieve valid causal inferences [3, 4].

Additive Noise Models: These models are dis-

cussed in the context of learning causal Directed Acyclic Graphs (DAGs) from observational data. Linear models with additive Gaussian noise are commonly used, but non-Gaussian data can help infer causal directions. Identifiability is emphasized, ensuring that the model's parameters can be learned from observations. Identifiability is essential for successful inference. The classic assumption of faithfulness is presented, emphasizing the importance of faithfulness in faithfully representing the data-generating process by the Bayesian Network (BN) [5].

These concepts lay the groundwork for understanding causal inference, modeling, and the challenges involved in identifying causal relationships from data.

In order to find the Markov Blanket, a few methods were developed over time, different approaches and typologies for these are now presented.

Constraint-Based Methods This section on "Constraint-Based Methods" discusses approaches used for causal inference and Markov Blanket (MB) learning. Constraint-based methods operate under two fundamental propositions.

The first proposition concerns the relationships between a node and its parents or children in a Bayesian Network (BN). It asserts that if a node is a parent or child of another node in the BN, these two variables are not independent given any subsets of the variables except themselves.

The second proposition delves into the relationships between a node and its spouses in a BN. It highlights that if one node is a spouse of another node, and they share a child, there exists a subset in which the two nodes are independent, but they become dependent when the child is included in the set [6].

These propositions serve as the basis for identifying conditional dependencies, which in turn, facilitate MB learning. To assess these dependencies, statistical independence tests are utilized. Five distinct types of tests are employed as selection criteria in constraint-based methods: the λ^2 test, the G^2 test, mutual information for discrete features, Fisher's Z test for continuous features with linear relations and additive Gaussian errors, and kernel-based tests for continuous features with non-linearity and non-Gaussian noise.

Within this category of methods, **"Simultaneous MB learning"** techniques are explored. These algorithms aim to simultaneously find the parents, children, and spouses of the class variable while learning its Markov Blanket. They follow a forward-backward strategy and employ various heuristics to efficiently identify these sets. Three prominent methods are discussed: the Grow-Shrink Markov Blanket (GSMB), the Incremental Association Markov Boundary (IAMB), and the Forward-Backward selection with Early Dropping ($FBED^k$). Each of these techniques employs distinct strategies and heuristics for the forward and backward phases.

GSMB is a two-phase algorithm. In the "growing phase," it expands the MB by adding variables that violate the MB property, which may lead to false positives. The subsequent "shrinking phase" aims to remove these extraneous variables. While GSMB can be effective, it may not perform well with small sample sizes and high-dimensional datasets [9].

IAMB is an enhanced version of GSMB, addressing the issue of false positives by using a dynamic heuristic in the forward phase. It adds features with the highest association to the class variable while conditioned on the current MB, reducing the risk of false positives. IAMB relies on assumptions that the data can be faithfully represented by a Bayesian Network and that reliable conditional independence tests are available [6].

To mitigate the high data requirements of IAMB, several variants have been introduced, such as IAMBnPC, Inter-IAMB, Inter-IAMBnPC, and Fast-IAMB, aiming to minimize the size of the Markov Blanket and enhance efficiency.

$FBED^k$ builds on the IAMB framework, introducing an early dropping strategy in the forward phase. It efficiently reduces the number of candidate variables while retaining the relevant ones, terminating when no features remain and subsequently revisiting dropped features in K additional runs of the forward phase [10].

In the context of Markov Blanket (MB) learning, a **"divide and conquer"** approach aims to address the data requirement issue posed by simultaneous MB learning methods. This approach splits the MB discovery process into two sub-problems:

- Learning the parents and children (PC(C)).
- Identifying the spouses (SP(C)).

The first sub-problem focuses on identifying the set of parents and children while using subsets of PC(C) as conditional sets. This minimizes the number of samples required, making it more data-efficient. For each feature X, it checks if there exists a subset S in PC(C) such that X is independent of C given S. Discarded features are not reconsidered, improving efficiency. However, this approach can become computationally expensive as the number of selected features grows.

Three methods following the "divide and conquer" philosophy are explored: Min-Max

Markov Blanket (MMMB), Parents and Children-based Markov Blanket (PCMB), and Simultaneous Markov Blanket (STMB) algorithms. They aim to identify conditional dependencies among variables and are grounded in the Standard forward-backward selection (SFBS) framework, guiding the identification of relevant Markov Blankets.

MMMB (Min-Max Markov Blanket): It starts by finding candidates for the parents and children set using Max-Min Parents and Children (MMPC) based on SFBS. In the forward phase, variables are sequentially added to PC(C) using a heuristic. The backward phase removes false positives. MMMB then recursively identifies the candidate set of the Markov Blanket (CMB) and distinguishes spouses based on specific conditional dependency tests. It efficiently filters out false positives using a property that only true spouses have [11].

PCMB (Parents and Children-based Markov Blanket): PCMB aims to find the true MB of a target variable under the assumptions of faithfulness and causal sufficiency. It utilizes two functions, GetPCD and GetPC, to solve the first subproblem of the divide and conquer approach. PCMB aims to minimize the number of nodes not belonging to PC(C) and employs a series of steps to iteratively refine the candidates set. It identifies spouses using specific conditional independence tests similar to MMMB [12].

STMB (Simultaneous MB): STMB attempts to improve computational efficiency by finding spouses without performing a thorough search for conditioning sets to identify the parents and children. It then removes false positives, the non-MB descendants, using the set of candidate spouses. While more efficient than some previous methods, STMB still relies on the entire set as a conditional set, making it less data-efficient [13].

These algorithms offer varying trade-offs between data efficiency, time efficiency, and the ability to distinguish parents from children, providing different tools for Markov Blanket learning in complex systems.

"MB learning with interleaving PC and spouse learning" methods are also explored. This approach alternates between the learning of parents and children (PC) and the learning of spouses. Two algorithms, Balance MB learning (BAMB) and Early-Exit MB learning (EEMB), are presented. BAMB integrates the learning of PC and spouses, while EEMB divides the process into learning and pruning phases. These methods aim to strike a balance between data and time efficiency [6].

Two algorithms, BAMB (Balance MB learning) and EEMB (Early-Exit MB learning), were developed as part of a method for Markov Blanket learning. These algorithms operate after learning the candidates for the parents and children (PC) and spouses (SP) sets. While both BAMB and EEMB aim to balance data and time efficiency, they have different approaches to the problem of identifying false positives.

BAMB (Balance MB learning): BAMB employs the IFBS framework to combine the learning of the PC set and the identification of spouses into a single procedure. It has three main steps: first, it finds the candidates for the PC and SP sets. Whenever a new feature is added to the candidate PC set, the algorithm starts identifying the spouses of the class variable C regarding that feature. It uses the discovered SP(C) set to remove false positives from CPC(C) and further prunes SP(C). BAMB's goal is to keep both sets as small as possible for the sake of data and time efficiency. However, BAMB may suffer from data inefficiency because false positives in the PC set can lead to the inclusion of many false spouses in SP(C), necessitating additional searches in the union of SP(C) and CPC(C) to remove false PC, making it less efficient both in terms of data and time [14].

EEMB (Early-Exit MB learning): EEMB, in contrast to BAMB, divides the learning process into two subroutines: one for learning and the other for pruning. This approach seeks to find a balance between data and time efficiency by intervaling the learning of both the PC and SP sets. While this approach may include some false PC, it prevents the entry of numerous false spouses into the SP set, which would result in a larger candidate set.

Both BAMB and EEMB offer strategies to handle the Markov Blanket learning process by attempting to optimize data and time efficiency. They present different trade-offs in terms of the size of candidate sets and computational costs, allowing users to choose the algorithm that best fits their specific needs and dataset characteristics.

Furthermore, the **"MB learning with relaxed assumptions"** is discussed, where assumptions such as faithfulness and causal sufficiency are relaxed. The Target Information Equivalence (TIE*) is introduced as a method for multiple MB learning when the faithfulness assumption is not met. TIE* involves generating new datasets to identify potential MBs, and the algorithm outputs all possible MBs of a class variable, even when the faithfulness assumption is violated [15].

Score-Based methods in causal inference rely on Bayesian Network (BN) learning algorithms to discover the Markov Blanket or the Parents-Children set. Unlike constraint-based methods that employ independence tests, score-based methods focus on finding the structure of a BN by maxi-

mizing a scoring function. The scoring function evaluates the fit between the BN's structure, represented as a Directed Acyclic Graph (DAG), and the given dataset. A greedy search method is applied to explore feasible solutions and identify the structure with the highest score. The problem of learning the BN structure is formulated as finding a DAG that maximizes the scoring function for the given dataset [6].

The scoring function must be decomposable, meaning it can express the score for a structure as a combination of local scores for each node and its parents in the graph. This decomposition enables score-based methods to distinguish parents from children during the Markov Blanket (MB) learning stage, which constraint-based algorithms cannot do.

One of the common strategies in score-based methods is "Divide and conquer MB learning," where the learning of the MB is divided into two sub-problems: finding parents and children (PC) and identifying spouses. The "Score-based Local Learning (SLL)" algorithm follows this approach. SLL starts by learning the PC set and then prunes it using symmetry checks. It proceeds to identify spouses of the class variable and uses symmetry checks to confirm the final spouses.

Another score-based algorithm is the "Greedy Search Algorithm (GES)," which divides its process into the Forward Equivalence Search (FES) phase and the Backward Equivalence Search phase. In the FES phase, GES greedily adds edges until a local maximum is reached, while the Backward Equivalence Search phase removes edges to reach another local maximum. The standard implementation of GES considers each variable as a root node in a DAG and iteratively adds parents, optimizing the scoring function to find the best structure [18].

Score-based methods offer an alternative approach to causal inference and MB learning, providing a framework for efficiently learning Bayesian network structures from data without relying on independence tests. These methods can be particularly valuable when the scoring function is well-designed and decomposition-friendly.

MB learning with interleaving PC and spouse learning In the realm of causal inference, it's often essential to distinguish between parents (direct causes) and children (direct effects) in order to predict the consequences of interventions and make informed decisions. This distinction becomes particularly valuable when dealing with datasets that exhibit variations in data distributions between the training and test sets. In such cases, the set of parents of a class variable can serve as a set of invariant features, allowing for robust predictions [6].

To achieve this distinction, two key strategies are employed: "Global Bayesian Network (BN) Structure Learning" and "Local BN Structure Learning."

Global BN Structure Learning: This approach begins by first learning the Markov Blanket of each feature, typically using existing causality-based feature selection methods. Then, it constructs an initial draft of the BN structure, starting as an undirected graph. The edges in the graph are subsequently directed, either through independence tests or scoring criteria. Several methods have been developed based on this approach, including "GGSL" (Global Graph Structure Learning) and "PSL" (Parallel Structure Learning), which are used to learn BN structures from data [6].

Local BN Structure Learning: In many real-world scenarios, the primary focus is on understanding the causal relationships related to a specific class variable, rather than the entire dataset. Algorithms designed for local BN structure learning focus on finding the Parents-Children (PC) set of a class variable and constructing the structure based on this set. The process then extends to identifying the Markov Blanket of features connected to the class variable, effectively building local structures for each feature. This iterative process continues until parents are fully distinguished from children or until further distinction is not possible.

Both of these strategies contribute to distinguishing parents from children, a crucial aspect of causal inference that enables more accurate and meaningful predictions in various domains, particularly when datasets exhibit distribution variations. These methods help researchers and practitioners make more informed decisions by understanding the causal relationships within their data.

3. Implementation

In summary, the implementation phase of my work involves selecting the appropriate toolbox for my research. After considering the PyCausal toolbox [16], I encountered some issues that made it unsuitable for my needs. As an alternative, I chose the GCastle toolbox, developed by HUAWEI Noah's Ark Lab [17], which offers a comprehensive set of tools for causal structure learning.

The GCastle toolbox boasts several features, including dataset generation (both simulated and based on real-world data), causal discovery, and graph evaluation. It supports four categories of algorithms for causal discovery: constraint-based, function-based, score-based, and gradient-based algorithms. The toolbox provides a graphical user interface (GUI) that simplifies task management and offers visualization capabilities for datasets, learned graphs, and evaluation metrics.

My plan involves extending the functionality of the GCastle toolbox. This extension includes developing a function for generating categorical data and introducing variations of existing scores to evaluate causal discovery on different graphs. These variations incorporate adaptations of the BIC and BDeu scores for categorical data and introduce new scoring criteria based on entropy and Kullback-Leibler divergence.

For continuous variables, two variants of the BIC score have been implemented within the toolbox:

- BIC by R^2: This variant calculates the BIC score using the coefficient of determination (R^2). It assesses how well independent variables (parents) explain the variances in the dependent variable (Y), considering both model fit (R^2) and model complexity (the number of parent variables, k).
- BIC by Scatter: This implementation computes the BIC score based on a scatter matrix. It takes into account the variance of the target variable (Y) and the effect of parent variables (Pa) by modifying the variance to calculate the BIC score, considering the number of parent variables (k).

In this section of my work, I've detailed the implementation of scoring methods specifically designed for categorical variables. These methods include adaptations of the BIC score and the BDeu score. Additionally, I've introduced a new scoring method that integrates entropy into the evaluation process. Here's a summary of the key points:

BIC for Categorical Data:

I've successfully adapted the BIC scoring method to handle categorical data. The process involves initializing data matrices (r_i and q_i), calculating term0 as the penalty for adding a parent configuration to the target variable, iterating through possible parent configurations, calculating term2, iterating over states (k) for the target variable, and computing the final BIC score as the sum of term0 and term1.

BDeu for Categorical Data:

Similar to the BIC score, I've effectively adapted the BDeu scoring method to categorical data. The process includes initializing data matrices (r_i and q_i), calculating term0 as the penalty for adding a parent configuration to the target variable, iterating through possible parent configurations, calculating term2, iterating over states (k) for the target variable, and computing the final BDeu score as the sum of term0 and term1.

Scores Using Entropy:

I've introduced a new scoring method that incorporates entropy. This method combines log-likelihood, entropy, and a weight factor (alpha) to compute the final score. The process involves calculating unique combined configurations, counting unique parent configurations, computing conditional probabilities, checking for parents of the target variable, and determining the final score using the weighted combination of log-likelihood and entropy.

Scores using KL divergence:

I've also introduced the concept of Kullback-Leibler (KL) divergence and explained how it quantifies the dissimilarity between probability distributions. KL divergence is applied in my scoring methods as a means to evaluate the fit of a Bayesian Network structure to the data. It is used similarly to entropy in my scoring methods. I've provided equations for KL divergence and detailed the processes for computing both BIC and BDeu scores using KL divergence, extending the toolbox's capabilities for evaluating Bayesian Network structures.

4. Results & discussion

To conduct a thorough assessment of the implemented scoring methods, a structured approach was adopted. This approach involved the use of three distinct groups of causal graphs, each designed to represent varying levels of complexity commonly encountered in causal inference:

- **Group 1:** Graphs with three nodes and three edges.
- **Group 2:** Graphs with four nodes and three edges.
- **Group 3:** Graphs with four nodes and five edges.

Within each of these groups, three unique graphs were created, each characterized by different noise scales: 1, 10, and 100. The combination of diverse graph structures and noise levels enabled a comprehensive evaluation of the implemented scoring methods across a range of scenarios and conditions.

BIC Scores In the comprehensive evaluation of the implemented BIC scores across different graph structures and noise scales, several key findings emerge. Let's break down the observations for each scenario:

In the context of Group 1, encompassing scenarios with increasing noise scale (G1_1, G1_10, G1_100), the evaluation of the Categorical BIC score reveals a consistent pattern. In G1_1, it demonstrates the capacity to accurately pinpoint certain causative links while simultaneously committing errors in the recognition of others. As the noise scale escalates in G1_10, this pattern persists, showcasing a blend of correct and incorrect

identifications. When the noise reaches its maximum in G1_100, the Categorical BIC continues to unveil causative relationships, but the intricacy of the problem becomes evident, highlighting the challenges of distinguishing genuine causal connections amidst increased noise levels.

Entropy BIC score reveals a recurring trend. In G1_1, it showcases the ability to accurately identify certain causal links while simultaneously making errors in the recognition of others. This pattern persists in G1_10, where the noise scale increases, demonstrating a mixture of correct and incorrect identifications. As the noise reaches its maximum in G1_100, the Entropy BIC continues to exhibit competence in recognizing specific causal relationships but encounters difficulties in others. Meanwhile, the Divergence BIC score follows a similar trajectory. In G1_1, it effectively reveals some causal connections while making errors in identifying others. This pattern remains consistent in G1_10, and in G1_100, it maintains its success in recognizing particular causative links but grapples with challenges posed by others. The complexity of the problem becomes particularly evident in G1_100, emphasizing the intricate task of distinguishing genuine causal connections within the context of heightened noise levels.

Group 2 presents a notable pattern where the categorical, entropy, and divergence scores consistently reveal the same edges in G2_1. This pattern is marked by the simultaneous discovery of some genuine causal connections and the inclusion of numerous incorrect causal identifications. G2_10 and G2_100 maintain this consistent pattern of identifying the same edges across all score variants, but they differ primarily in the sequence of edge discovery. These results underscore the trade-offs and complexities in accurately distinguishing true causal relationships within different noise scales and graph complexities.

In Group 3, specifically within G3_1 and G3_10, a strikingly consistent pattern emerges across all score variants. These patterns involve the concurrent discovery of the same causal relations while also featuring the presence of multiple incorrect identifications. The prevalence of incorrect identifications adds a layer of complexity when attempting to discern the genuine causal structure amidst the noise. G3_100 continues to maintain this unwavering pattern, with each score variant discovering edges in the same order, a testament to the challenges and intricacies involved in identifying true causal connections across different levels of noise and graph structures.

Bdeu Scores Across various groups and noise scales, the BDeu scores consistently discover all true causal links, but they are also prone to identifying a significant number of incorrect causal connections. The primary distinction between score variants lies in the order of edge discovery and the frequency of score assignments to different configurations.

In summary, the BIC derivatives of the three different score variants (categorical, entropy, divergence) appear to offer a better chance of determining the true causal relationships in the underlying Directed Acyclic Graph (DAG). However, it's important to note that as the graph complexity and noise scale increase, the scores tend to both discover fewer correct relations and misidentify more edges. This indicates the need for careful consideration when selecting the appropriate score variant based on the context and characteristics of the causal inference problem at hand.

5. Conclusions

In this thesis, we delved into the realm of causal inference, a fundamental challenge in machine learning and data analysis. Unraveling the hidden causes and effects within complex systems requires sophisticated methods due to the intricate web of relationships between variables.

Our exploration led us through three primary categories of causal discovery methods: constraint-based, score-based, and those discerning parents from children in causal relationships. Each category offered a unique perspective on causality.

For our experiments, we selected the GES (Greedy Equivalence Search) algorithm as our foundation. It allowed us to assess different scoring methods across diverse graph structures and noise levels. Within these scoring methods, we focused on the Bayesian Information Criterion (BIC) and the Bayesian Dirichlet equivalent uniform (BDeu) score, each featuring three distinct variants: the standard categorical score, an entropy-based score, and a score rooted in Kullback-Leibler (KL) divergence.

Our comprehensive investigation unveiled essential insights into causal inference. BDeu, in its quest for causation, often identified numerous edges, posing a challenge in distinguishing true causal connections from spurious ones. In contrast, BIC emerged as a more precise and cautious scorer, simplifying the identification of genuine causal links.

Furthermore, our exploration highlighted the influence of graph size and noise scale on score performance. The delicate balance between graph complexity and noise level added complexity to the causal discovery process, emphasizing the importance of selecting scores carefully in consideration of these contextual factors.

References

[1] J. Pearl, Causality. Cambridge: Cambridge University Press, 2009.

[2] J. Peters, D. Janzing, and Scholkopf Bernhard, Elements of causal inference: Foundations and learning algorithms. Cambridge, MA: The MIT Press, 2017.

[3] I. Guyon, A. R. Statnikov, and B. B. Batu, Cause effect pairs in machine learning. Cham, Switzerland: Springer, 2019.

[4] J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf, "Distinguishing cause from effect using observational data: methods and benchmarks," arXiv [cs.LG], 2014.

[5] J. Peters, J. Mooij, D. Janzing, and B. Schölkopf, "Causal discovery with continuous additive noise models," arXiv [stat.ML], 2013.

[6] K. Yu et al., "Causality-based feature selection: Methods and evaluations," ACM Comput. Surv., vol. 53, no. 5, pp. 1–36, 2021.

[7] P. O. Hoyer, J. Mooij, J. Peters, and B. Schölkopf, "Nonlinear causal discovery with additive noise models," Nips.cc. [Online]. Available: https://papers.nips.cc/paper/2008/file/f7664060cc52bc6f3d620bcedc94a4b6-Paper.pdf. [Accessed: 02-Jan-2023].

[8] "The do-calculus revisited Judea Pearl keynote lecture," Ucla.edu. [Online]. Available: https://ftp.cs.ucla.edu/pub/stat_ser/r402.pdf. [Accessed: 02-Jan-2023].

[9] D. Margaritis and S. Thrun, Bayesian network induction via local neighborhoods. 2000. In Advances in neural information processing systems. 505–511.

[10] G. Borboudakis and I. Tsamardinos, "Forward-backward selection with early dropping," arXiv [cs.LG], 2017.

[11] I. Tsamardinos, C. F. Aliferis, and A. Statnikov, "Time and sample efficient discovery of Markov blankets and direct causal relations," in Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '03, 2003.

[12] J. M. Peña, R. Nilsson, J. Björkegren, and J. Tegnér, "Towards scalable and data efficient learning of Markov boundaries," Int. J. Approx. Reason., vol. 45, no. 2, pp. 211–232, 2007.

[13] T. Gao and Q. Ji, "Efficient Markov blanket discovery and its application," IEEE Trans. Cybern., vol. 47, no. 5, pp. 1169–1179, 2017.

[14] Z. Ling, K. Yu, H. Wang, L. Liu, W. Ding, and X. Wu, BAMB: A balanced Markov blanket discovery approach to feature selection. 2019.

[15] A. Statnikov, N. I. Lytkin, J. Lemeire, and C. F. Aliferis, "Algorithms for discovery of multiple Markov boundaries," J. Mach. Learn. Res., vol. 14, pp. 499–566, 2013.

usal_Inference-Aug27_2020-Neal.pdf. [Accessed: 02-Jan-2023].

[16] AITiger, PyCausalFS: pyCausalFS:A Python Library of Causality-based Feature Selection for Causal Structure Learning and Classification.

[17] Huawei Noah's Ark Lab. (2023). TrustworthyAI. GitHub. https://github.com/huawei-noah/trustworthyAI/tree/master/gcastle

[18] D. M. Chickering, "Optimal Structure Identification With Greedy Search," Journal of Machine Learning Research, 2002.

[19] H. A. Simon, "On the definition of the causal relation," in Models of Discovery, Dordrecht: Springer Netherlands, 1977, pp. 81–92.