CD-NOTEARS: Concept Driven Causal Structure Learning using NOTEARS

Jawad Chowdhury
Department of Computer Science
University of North Carolina at Charlotte
Charlotte, North Carolina, USA
mchowdh5@uncc.edu

Gabriel Terejanu
Department of Computer Science
University of North Carolina at Charlotte
Charlotte, North Carolina, USA
gabriel.terejanu@uncc.edu

***Abstract*—Causal discovery has become increasingly popular in recent years, with the emergence of various methods for inferring causal relationships from observational data. While NOTEARS is a widely-used structure learning method known for its effectiveness in handling scalar-valued continuous data, it is not well-posed for conceptual data. In this study, we present a novel extension of the NOTEARS method, called Concept-Driven NOTEARS (CD-NOTEARS), that leverages concept-level prior knowledge to impose DAGness on concepts instead of the raw high-dimensional data. Our proposed approach preserves the non-parametric nature of the original NOTEARS method and is evaluated on synthetic, benchmark, and real-world datasets. The results demonstrate that CD-NOTEARS outperforms the original implementation and offers a promising tool for causal discovery in scenarios where causality should be imposed on the concept level. Our study provides insights into how incorporating concept-level knowledge improves the performance of causal discovery and paves the way for further research in this direction.**

***Index Terms*—Causality, Structured Prediction and Learning, Supervised Deep Learning, Optimization for Neural Networks.**

I. Introduction

In recent years, the field of causal discovery has gained significant traction, driven by advancements in machine learning models that excel in handling large datasets and approximating intricate relationships. Consequently, numerous methods have emerged to infer causal relationships from observational data. These methods can be categorized into constraint-based algorithms e.g. PC [1], IC [2], and FCI [3], score-based approaches e.g. GES [4] and FGES [5], and functional causal models e.g. LiNGAM [6] and ANMs [7]. Constraint-based methods utilize conditional independence tests and rules to detect edge directions, often pinpointing the Markov equivalence class of the genuine causal graph. Meanwhile, score-based models target causal graph optimization over the DAG space, a process that becomes computationally intensive due to its combinatorial nature. NOTEARS, present in linear [8] and non-parametric [9] forms, adopts an algebraic acyclicity characterization, transforming the combinatorial challenge into continuous constrained optimization. Variants of this continuous optimization approach have surfaced in works Ref. [10], [11], and [12], offering versatile causal mechanism modeling. While NOTEARS stands out for its efficacy across diverse uses, it's not limited to structure learning for continuous or scalar data but extends to feature vectors of conceptual data as well.

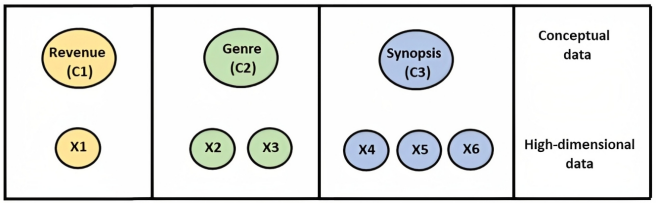

Fig. 1. Mapping of conceptual data to high-dimensional features for movie dataset. The three main concepts considered are revenue (C1), genre (C2), and synopsis (C3). The one-dimensional feature X1 corresponds to revenue, while the encoding of genre results in two-dimensional features X2 and X3. Synopsis is represented by a three-dimensional embedding with features X4, X5, and X6.

For example, consider an IMDb movie dataset with three concepts: revenue (C1), genre (C2), and synopsis (C3). Revenue (C1) represents movie-generated revenue (X1). Assuming our dataset has only thriller and sci-fi genres, we can use one-hot encoding to represent the genre (C2), creating a two-dimensional vector (X2 and X3) for these genres. For the movie synopsis (C3), we can use NLP methods to produce a three-dimensional embedding (X4, X5, and X6). Thus, our dataset has three concepts (C1, C2, C3) leading to a six-dimensional feature vector for each movie (X1 through X6, as depicted in Fig. 1). By applying NOTEARS to this vector-valued data, causal relationships within the high-dimensional feature space can be discerned, shedding light on the interconnections between features X1 through X6. However, a general challenge with structure learning is that uncovering the causal structure requires complete coverage of the data distribution. Intuitively, without a comprehensive representation of the data distribution, one can miss latent causal relationships or infer spurious ones due to sample biases. To address this challenge, researchers often provide algorithms with additional knowledge to augment the optimization with prior knowledge, as featured in software packages such as CausalNex [1], causal-learn [2], bnlearn [13], DoWhy [14], and gCastle [15]. Previous studies have shown that incorporating domain knowledge can be beneficial and lead to superior performance. For example,

[1]https://github.com/quantumblacklabs/causalnex
[2]https://github.com/cmu-phil/causal-learn

the impact of prior knowledge on score-based causal learning algorithms was evaluated in Ref. [16], and a separate study evaluated the impact on NOTEARS [17]. Additionally, another recent study [18] presents KGS, a novel knowledge-guided greedy score-based causal discovery approach that uses structural priors to constrain the search space and guide the process.

In this paper, we present CD-NOTEARS, an extension of the NOTEARS algorithm designed for concept-driven causal structure learning in vector-valued data. This novel approach integrates prior knowledge on relations between concepts and high-dimensional features as meta-information, imposing DAGness on concept-level data, a departure from the original NOTEARS which operates on raw high-dimensional features. Through extensive experiments on varied datasets, we showcase CD-NOTEARS's proficiency in identifying causal relationships, highlighting its enhanced performance compared to the original NOTEARS. Our key contributions can be summarized as follows: (1) We introduced CD-NOTEARS, a novel extension of the NOTEARS algorithm, that facilitates concept-driven causal structure learning in vector-valued data while incorporating prior relations between different concepts and high-dimensional features, preserving the nonparametric essence of the original NOTEARS algorithm, (2) Departing from traditional methods, our approach emphasizes DAGness at the concept level rather than focusing solely on high-dimensional raw features, and (3) our study illustrates empirical validation through comprehensive experiments on synthetic, benchmark, and real-world datasets.

The remainder of this paper is organized as follows: Section II delves into the methodology of CD-NOTEARS, Section III presents our experimental settings and evaluations. Finally, Section IV encapsulates our conclusions and highlights the significant takeaways.

II. Methodology

The proposed CD-NOTEARS method builds on the original nonparametric NOTEARS algorithm [9], specifically the NOTEARS-MLP instance, to infer causal relationships from vector-valued data. In this section, we summarize the background of linear [8] and nonparametric [9] extensions of NOTEARS and then delve into our adaptation: the CD-NOTEARS approach.

Observational causal structure learning aims to learn the causal relationships encoded by a directed acyclic graph (DAG) $\mathcal{G}$ from n i.i.d. observations in the data matrix $\mathbf{X} = [\mathbf{x_1}|\dots|\mathbf{x_d}] \in \mathbb{R}^{n\times d}$. The score-based approach focuses on identifying the DAG model $\mathcal{G}$ that best fits the observed data $\mathbf{X}$ based on a scoring criterion $S(\mathcal{G}, X)$ over the discrete space of DAGs $\mathbb{D}$ where $\mathcal{G} \in \mathbb{D}$ [4]. This optimization problem can be formulated as:

$$\begin{aligned} \min_{\mathcal{G}} \quad & S(\mathcal{G}, X) \\ \text{subject to} \quad & \mathcal{G} \in \mathbb{D} \end{aligned} \tag{1}$$

The linear NOTEARS [8] algorithm reformulates the combinatorial optimization in Eq. 1 to a continuous one through an algebraic characterization of the acyclicity constraint. This method encodes the graph $\mathcal{G}$ defined over the d nodes into a weighted adjacency matrix $W = [w_1|\dots|w_d] \in \mathbb{R}^{d\times d}$ where $w_{i,j} \neq 0$ if there is an active edge $X_i \to X_j$ and $w_{i,j} = 0$ otherwise. The weighted adjacency matrix W entails a linear structural equation model (SEM) by $X_i = f_i(X) + N_i = w_i^T X + N_i$; where N_i is the associated noise. The authors define a smooth score function on the weighted matrix as $h(W) = \text{tr}(e^{W\circ W}) - \text{d}$ where $\circ$ is the Hadamard product and e^M is the matrix exponential of M. This reformulates Eq. 1 into its equivalent form:

$$\begin{aligned} \min_{W\in\mathbb{R}^{d\times d}} \quad & L(W) \\ \text{subject to} \quad & h(W) = 0 \end{aligned} \tag{2}$$

where $L(W)$ is the least square loss over W and $h(W)$ score defines the DAG-ness of the graph. The nonparametric NOTEARS [9] uses partial derivatives on the functional form f_j to determine the dependency of random variable X_j on other random variables. The authors define $f_j \in H^1(\mathbb{R}^d) \subset L^2(\mathbb{R}^d)$ over the Sobolev space of square integrable functions whose derivatives are also square integrable, and f_j can be independent of random variable X_i if and only if $||\partial_i f_j||_{L^2} = 0$ where ∂_i denotes partial derivative with respect to X_i. This redefines the weighted adjacency matrix as $W(f)$ with each $W_{i,j}$ encoding the partial dependency of f_j on variable X_i and allows us to write Eq. 2 equivalently:

$$\begin{aligned} \min_{f:f_j\in H^1(\mathbb{R}^d),\forall j\in[d]} \quad & L(f) \\ \text{subject to} \quad & h(W(f)) = 0 \end{aligned} \tag{3}$$

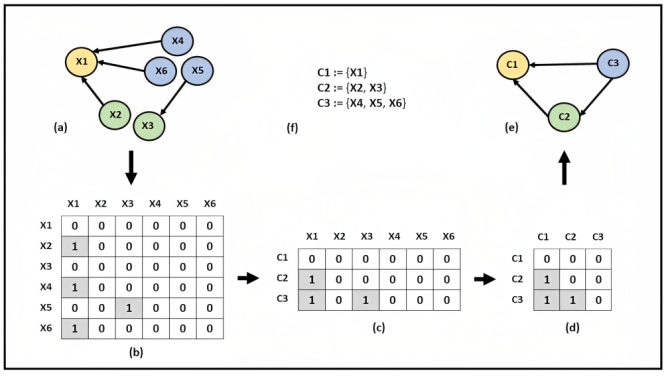

Fig. 2. Illustration of concept-driven adjacency matrix and graph formulation process from high dimensional data: (a) graphical representation of relations between high dimensional features in raw data, (b) corresponding adjacency matrix for high dimensional graph relations, W, (c) intermediate matrix formulation obtained by applying row aggregation based on the concept-level meta-information, (d) concept-driven adjacency matrix obtained after full transformation using row and column aggregation, W^{con}, (e) graphical representation of the relations between concepts (C1, C2, C3), (f) Prior knowledge or meta-information regarding the concepts and their representations in high dimensional feature space. For the purpose of simplicity, this figure demonstrates the process using binary adjacency matrices.

While NOTEARS deduces causal relationships among features by applying a continuous acyclicity constraint on the high-dimensional adjacency matrix, W, our CD-NOTEARS

method adopts a concept-driven strategy. Firstly, we obtain the adjacency matrix similarly to NOTEARS. Instead of directly constraining this matrix, we transform it into an aggregated adjacency matrix, W^{con}, using concept-level prior knowledge. This matrix captures concept-level relationships, with aggregation refining the optimization search space to guide the optimization. CD-NOTEARS imposes acyclicity on the concept-level relations captured in W^{con}. Fig. 2 illustrates the CD-NOTEARS approach to derive concept-driven causal relations, W^{con}, from the high-dimensional matrix, W. In order to maintain consistency with our previous example presented in Fig. 1, we demonstrate the matrix transformation using the three concepts introduced earlier. Therefore, C1 refers to the revenue of each movie, represented by a scalar-valued one-dimensional feature X1. Meanwhile, C2 and C3 correspond to the genre and synopsis concepts of the movie, represented by two-dimensional (X2 and X3) and three-dimensional (X4, X5, and X6) feature spaces, respectively. Unlike the original NOTEARS implementation that imposes acyclicity on the raw-level high-dimensional graph as shown in Fig. 2(a), CD-NOTEARS imposes acyclicity on the concept-level graph as in Fig. 2(e). To achieve the concept-level matrix, we first generate the high-dimensional adjacency matrix (Fig. 2(b)). An intermediate matrix is then formed using row aggregation informed by concept-level meta-information (Fig. 2(c)). The final transformation, integrating both row and column aggregation, yields the concept-driven matrix W^{con} (Fig. 2(d)), influenced by the relations between concepts and features shown in Fig.2(f). It is to be noted that various matrix transformation or aggregation methods can be employed to get the concept-level relations from the raw relations, as long as they preserve the causal relationships from the raw level to the concept level. Such transformation or aggregation function should satisfy the following equation:

$$W^{con}_{m,\,n} \;=\; \begin{cases} 0 & \text{if } \forall(X_i \in C_m, X_j \in C_n)\ W_{i,j} = 0 \\ \neq 0 & \text{otherwise} \end{cases} \tag{4}$$

Eq. 4 allows us to aggregate the raw-level information in W and determine the relationship between concepts such as C_m and C_n. If any of the random variables X_i that belong to concept C_m has a causal link in high-dimensional feature space to any other random variable X_j that belongs to concept C_n, the corresponding cell in the concept-level aggregated matrix, $W^{con}_{m,n}$ should reflect that relationship. Otherwise, the cell in the concept-level matrix is set to zero. After applying the transformation using Eq. 4, our optimization problem reformulates to:

$$\begin{aligned} \min_{f: f_j \in H^1(\mathbb{R}^d), \forall j \in [d]} \quad & L(f) \\ \text{subject to} \quad & h(W^{con}(f)) = 0 \end{aligned} \tag{5}$$

Following a similar strategy to the original NOTEARS implementation, we solve the optimization problem using the augmented Lagrangian method [19]. The proposed CD-NOTEARS method preserves the non-parametric nature of the original NOTEARS algorithm while leveraging concept-level meta-information.

III. Experiments and Results

To evaluate our extended NOTEARS algorithm, CD-NOTEARS, we conducted case studies comparing its performance against the original NOTEARS model. Given the sensitivity of the NOTEARS algorithm to data scaling, as shown in earlier studies [20], [21], we scaled our data using the *standardization* method from Python's scikit-learn [22] library. We ensured consistent model structures by employing an MLP with 10 hidden units and sigmoid activations for both models. While CD-NOTEARS integrates meta-information during optimization, focusing on concept-level relations, the original NOTEARS first learns the causal graph in the high-dimensional feature space, then post-processes with meta-information. We adopted the 'mean' as the aggregation function in both models for concept-level causal graph learning. For comparative analysis, we utilized two key performance metrics: false discovery rate (FDR) and structural hamming distance (SHD). The FDR, in particular, offers insights into the conservativeness of our method. A lower FDR indicates fewer unwarranted causal claims, addressing the challenge highlighted by previous study [23] regarding non-conservative error trade-offs seen in many causal discovery methods. On the other hand, the SHD, a widely-recognized pattern metric for evaluating causal discovery methodologies [24], provides a holistic view of how closely the predicted graph aligns with the ground truth. To emphasize reliability, we conducted 50 different random trials for each case study, evaluating the performance of both models based on the mean and standard deviation of the performance metrics. We then performed statistical significance analysis using a t-test with α level of 0.05.

A. Synthetic Dataset

To evaluate the effectiveness of CD-NOTEARS against the original NOTEARS, we ran simulations on synthetic datasets. We examined 16 combinations, varying between Erdos-Renyi and Scale-Free graph models (gt = ER, SF), number of nodes (d = 10, 20), sample sizes (n = 200, 1000), and edges (s0 = 1d, 4d), where d indicates node count. Each combination yielded 50 random graphs or true DAGs, generated via the Additive Noise Model (ANM) with MLPs following the methodology in the original work [9]. For the experiments with synthetic datasets, we considered two different ranges for the dimension of each concept. In the first case, the range was limited to 1 to 3, and in the second case, the range was expanded to 1 to 5. The results are presented in Table I and Table II, respectively. Our evaluation showcases the superiority of CD-NOTEARS over the original implementation. By integrating prior knowledge into the graph formulation and imposing acyclicity at the concept level, CD-NOTEARS achieves lower FDR and SHD in most scenarios. This underscores the merit of employing concept-level knowledge for precise causal structure leaerning.

TABLE I

PERFORMANCE COMPARISON OF THE PROPOSED CD-NOTEARS MODEL AND THE ORIGINAL NOTEARS IMPLEMENTATION ON SYNTHETIC DATA CONSIDERING RANDOM VARIABLES AS CONCEPTS HAVING DIMENSION RANGES FROM 1 TO 3.

n	d	s0	gt	fdr		shd	
				CD-NOTEARS	NOTEARS	CD-NOTEARS	NOTEARS
200	10	10	ER	**0.86 ± 0.04**	0.89 ± 0.02	**37.84 ± 2.20**	47.36 ± 2.99
		40	SF	0.89 ± 0.04	0.90 ± 0.03	**38.90 ± 2.27**	47.77 ± 2.55
	20	20	ER	**0.48 ± 0.11**	0.56 ± 0.05	**22.04 ± 4.90**	35.78 ± 5.02
		80	SF	**0.58 ± 0.07**	0.66 ± 0.05	**26.20 ± 3.35**	40.30 ± 4.43
	10	10	ER	0.93 ± 0.01	0.94 ± 0.01	**165.64 ± 6.34**	188.09 ± 6.07
		40	SF	0.94 ± 0.02	0.94 ± 0.01	**167.82 ± 6.77**	184.50 ± 6.06
	20	20	ER	**0.75 ± 0.04**	0.77 ± 0.03	**139.42 ± 6.91**	166.78 ± 10.03
		80	SF	**0.78 ± 0.05**	0.81 ± 0.03	**142.90 ± 10.37**	167.21 ± 9.06
1000	10	10	ER	0.83 ± 0.14	0.86 ± 0.07	**22.12 ± 4.91**	33.78 ± 8.07
		40	SF	0.88 ± 0.09	0.86 ± 0.08	**21.42 ± 5.58**	30.70 ± 7.23
	20	20	ER	**0.48 ± 0.15**	0.54 ± 0.09	**30.80 ± 4.41**	33.88 ± 5.13
		80	SF	**0.55 ± 0.18**	0.63 ± 0.10	**27.26 ± 4.57**	32.54 ± 4.61
	10	10	ER	0.93 ± 0.03	0.92 ± 0.02	**122.26 ± 20.65**	152.18 ± 19.26
		40	SF	0.95 ± 0.03	**0.94 ± 0.02**	**124.88 ± 14.90**	149.30 ± 17.91
	20	20	ER	**0.72 ± 0.05**	0.76 ± 0.03	**119.34 ± 9.89**	147.36 ± 11.28
		80	SF	0.76 ± 0.07	0.78 ± 0.05	**115.38 ± 15.17**	140.56 ± 16.82

TABLE II

PERFORMANCE COMPARISON OF THE PROPOSED CD-NOTEARS MODEL AND THE ORIGINAL NOTEARS IMPLEMENTATION ON SYNTHETIC DATA CONSIDERING RANDOM VARIABLES AS CONCEPTS HAVING DIMENSION RANGES FROM 1 TO 5.

n	d	s0	gt	fdr		shd	
				CD-NOTEARS	NOTEARS	CD-NOTEARS	NOTEARS
200	10	10	ER	**0.86 ± 0.04**	0.89 ± 0.01	**37.70 ± 1.78**	50.33 ± 2.16
		40	SF	**0.86 ± 0.03**	0.90 ± 0.01	**38.24 ± 1.66**	50.39 ± 2.30
	20	20	ER	**0.48 ± 0.12**	0.57 ± 0.04	**21.92 ± 5.37**	42.50 ± 5.17
		80	SF	**0.59 ± 0.10**	0.67 ± 0.04	**26.64 ± 4.79**	45.16 ± 4.14
	10	10	ER	**0.93 ± 0.01**	0.94 ± 0.01	**161.68 ± 6.44**	194.46 ± 5.92
		40	SF	0.94 ± 0.02	0.95 ± 0.01	**165.14 ± 6.15**	195.42 ± 4.56
	20	20	ER	**0.76 ± 0.04**	0.78 ± 0.01	**139.66 ± 6.65**	179.57 ± 7.04
		80	SF	**0.78 ± 0.05**	0.81 ± 0.02	**139.28 ± 10.15**	180.54 ± 10.21
1000	10	10	ER	0.85 ± 0.06	0.87 ± 0.03	**30.14 ± 4.89**	43.60 ± 4.67
		40	SF	0.89 ± 0.06	0.89 ± 0.03	**29.90 ± 5.14**	43.78 ± 5.33
	20	20	ER	**0.45 ± 0.12**	0.55 ± 0.07	**25.30 ± 5.23**	37.88 ± 6.13
		80	SF	**0.56 ± 0.15**	0.64 ± 0.08	**26.48 ± 5.76**	38.62 ± 6.90
	10	10	ER	0.93 ± 0.02	0.93 ± 0.01	**137.40 ± 13.93**	177.26 ± 9.79
		40	SF	0.94 ± 0.02	0.94 ± 0.02	**136.46 ± 13.81**	174.08 ± 12.48
	20	20	ER	**0.74 ± 0.04**	0.78 ± 0.03	**127.02 ± 10.89**	169.86 ± 10.51
		80	SF	**0.77 ± 0.07**	0.79 ± 0.04	**123.34 ± 16.34**	162.56 ± 15.12

TABLE III

PERFORMANCE COMPARISON OF THE PROPOSED CD-NOTEARS MODEL AND THE ORIGINAL NOTEARS IMPLEMENTATION ON BINARY BENCHMARK DATASETS.

dataset	fdr		shd	
	CD-NOTEARS	NOTEARS	CD-NOTEARS	NOTEARS
Lucas	**0.76 ± 0.03**	0.82 ± 0.02	**12.16 ± 0.78**	22.96 ± 1.37
Asia	**0.75 ± 0.01**	0.87 ± 0.04	**9.02 ± 0.14**	16.00 ± 1.13

B. Benchmark Dataset

a) Benchmark Datasets for Binary Variables: We then compared CD-NOTEARS and the original NOTEARS on two benchmark datasets for categorical variables: LUCAS and ASIA. The LUCAS (LUng CAncer Simple set) dataset [25], sourced from the Causality Workbench project, comprises 2000 instances of 12 binary variables detailing factors affecting lung cancer. The data is synthetically created by causal Bayesian networks and in our study, we used the unmanipulated distribution of the dataset referred to as LUCAS0 [3], as visualized in Fig 3. The second dataset, ASIA [26] depicts the interplay between tuberculosis, lung cancer, bronchitis, and Asia visits. Containing 8 binary variables and 5000 samples generated following the causal Bayesian network, its causal graph [4] and dataset [5] are available online. Our evaluation, presented in Table III, shows CD-NOTEARS surpassing NOTEARS in terms of FDR and SHD values on both datasets, emphasizing its effectiveness for concept-driven data with binary categorical variables.

[3]http://www.causality.inf.ethz.ch/data/LUCAS.html

[4]https://www.bnlearn.com/bnrepository/discrete-small.html#asia

[5]https://github.com/AnaRitaNogueira/Methods-and-Tools-for-Causal-Discovery-and-Causal-Inference

TABLE IV
PERFORMANCE COMPARISON OF THE PROPOSED CD-NOTEARS AND ORIGINAL NOTEARS IMPLEMENTATION ON MULTINARY BENCHMARK DATASETS USING PYTORCH [27] EMBEDDING LAYER TO GENERATE VECTOR-VALUED DATA FROM CATEGORICAL VARIABLES.

dataset	fdr		shd	
	CD-NOTEARS	NOTEARS	CD-NOTEARS	NOTEARS
Dutch	**0.56 ± 0.29**	0.69 ± 0.07	**41.62 ± 1.74**	46.84 ± 2.56
Adult	**0.56 ± 0.20**	0.73 ± 0.07	**38.42 ± 1.34**	44.70 ± 2.23

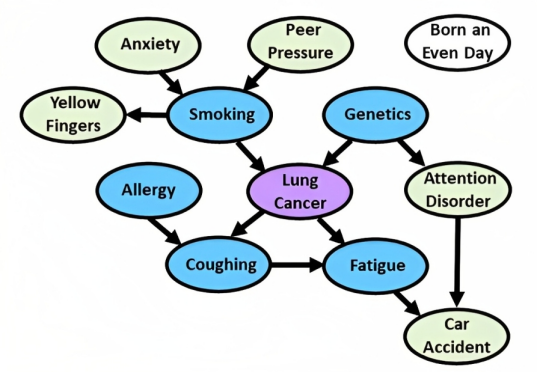

Fig. 3. Causal graph for unmanipulated distribution of LUCAS0 [25]

b) Benchmark Datasets for Multinary Variables: In our next experimental study, we assessed CD-NOTEARS on two mixed numeric and multinary datasets: the Dutch Census [28] and the Adult dataset [29]. The Dutch Census has 60,420 entries with 12 attributes utilized for structural learning, such as sex, age, household_position, country_birth, occupation, etc. Among these attributes, sex and occupation are binary, while the remaining attributes can take multiple values. The Adult dataset comprises 32,561 samples with 11 attributes, including a combination of continuous and categorical variables such as age, working_class, sex, hours_per_week, marital_status, income, etc. Age and hours_per_week are continuous variables, while the rest are categorical. Among the categorical variables, sex and income are binary, and the remaining variables are multinary. We considered the causal graph from a prior study [30] for both datasets. To process multinary categorical variables, we used PyTorch's [27] embedding layer to create vector embeddings for each concept. This technique efficiently manages mixed data, leading to a compact dataset. As illustrated in Table IV, CD-NOTEARS outperforms NOTEARS in the identification of causal structures from mixed data. By leveraging concept-level understanding and DAG properties, our approach highlights the significance of conceptual insights in high-dimensional causal learning.

C. Real Data

Finally, we evaluated CD-NOTEARS and the original NOTEARS using the IMDb movie dataset sourced from two Kaggle repositories: IMDB Movie data Analysis [6] and Movie Scripts Corpus [7]. The dataset, after cleaning, had data on 1764

[6]https://www.kaggle.com/code/robinjrjr/imdb-movie-data-analysis/data
[7]https://www.kaggle.com/datasets/gufukuro/movie-scripts-corpus

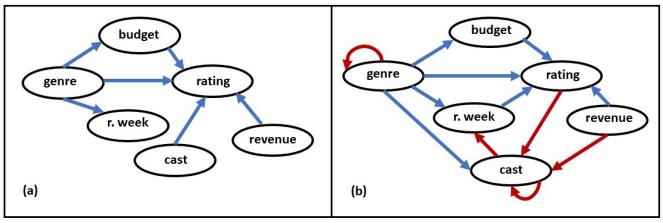

Fig. 4. Causal relations obtained from the movie datasets using two different models: (a) CD-NOTEARS and (b) the original implementation of NOTEARS. r. week stands for the release week of the movie.

movies, including features like budget, cast, genre, release week, user rating, and revenue. Cast and genre are vector-valued features, while the remaining features are scalar in nature. As each movie sample can have one or more casts and genres, we applied one-hot encoding to generate embeddings for each sample followed by training an auto-encoder to retain maximum information with lower dimensional features from these concepts. This process was applied independently to each of the multi-dimensional concepts in the dataset, namely cast and genre. To ensure a fair comparison, we kept the common settings of both model implementations similar. In total, CD-NOTEARS estimated six edges, which are budget $\rightarrow$ rating, cast $\rightarrow$ rating, genre $\rightarrow$ budget, genre $\rightarrow$ release week, genre $\rightarrow$ rating, and revenue $\rightarrow$ rating. Due to the absence of an established ground truth or consensus within the dataset, we depended on our own assessment to evaluate the predicted connections. Upon examination, we discovered that the majority of causal relationships estimated by CD-NOTEARS appeared to be reasonable and coherent. However, the relationship between rating and revenue appears ambiguous as higher rating of a movie can draw more people to watch the movie, resulting in increased revenue, and conversely, higher revenue could bias viewers to rate the movie higher. Despite this ambiguity, both implementations agreed on the direction of this relationship, suggesting it would not affect our comparative evaluation. Nevertheless, the original implementation of NOTEARS estimated six additional edges, some of which appeared unlikely such as cast $\rightarrow$ release week, rating $\rightarrow$ cast, and revenue $\rightarrow$ cast. Fig 4 illustrates the causal relations retrieved by both these models. Notably, NOTEARS applies DAGness to the raw-level high-dimensional features, which resulted in the generation of two self-loops for the concepts cast and genre. While this violates the acyclicity assumption, we found this characteristic intriguing as the selection of one cast may impact the selection of other casts, and a similar phenomenon may apply to genres. Nonetheless,

our proposed CD-NOTEARS implementation, which enforces DAGness on the concepts, appears to surpass the original NOTEARS implementation in terms of performance. Although we lack a quantitative metric for assessing performance, our analysis of the IMDb movie dataset presents persuasive evidence in favor of CD-NOTEARS.

IV. Conclusions

Our proposed method, CD-NOTEARS, represents a significant advancement in the field of causal discovery for concept-driven data. By emphasizing acyclicity constraints at the concept level and leveraging prior feature-to-concept knowledge, it refines causal relationship representation, bolstering reliability and accuracy. Through evaluations on diverse datasets, we have highlighted its efficacy, especially in sectors where conceptual data is prevalent such as healthcare, finance, and social science . This research emphasizes the benefits of integrating prior concept knowledge in causal structure learning, making CD-NOTEARS a valuable addition to the causal discovery repertoire. Looking ahead, there is potential to combine this concept-driven approach with other leading causal discovery methods to further amplify its potency. In conclusion, we firmly believe that our extension of the NOTEARS approach will be a pivotal asset for causal discovery across various domains. We hope that this research will inspire further studies and advancements in the field of causal discovery, ultimately leading to a better understanding of causality in complex systems and guiding effective causal learning methods.

Acknowledgment

Research was sponsored by the Army Research Office under Grant Number W911NF-22-1-0035 and by the National Science Foundation under Award Number 2218841. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

[1] P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman, *Causation, prediction, and search*. MIT Press, 2000.

[2] J. Pearl, *Causality*. Cambridge University Press, 2009.

[3] D. Colombo, M. H. Maathuis, M. Kalisch, and T. S. Richardson, "Learning high-dimensional directed acyclic graphs with latent and selection variables," *The Annals of Statistics*, pp. 294–321, 2012.

[4] D. M. Chickering, "Optimal structure identification with greedy search," *Journal of Machine Learning Research*, vol. 3, no. Nov, pp. 507–554, 2002.

[5] J. Ramsey, M. Glymour, R. Sanchez-Romero, and C. Glymour, "A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images," *International Journal of Data Science and Analytics*, vol. 3, pp. 121–129, 2017.

[6] S. Shimizu, P. O. Hoyer, A. Hyvärinen, A. Kerminen, and M. Jordan, "A linear non-gaussian acyclic model for causal discovery." *Journal of Machine Learning Research*, vol. 7, no. 10, 2006.

[7] P. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf, "Nonlinear causal discovery with additive noise models," *Advances in Neural Information Processing Systems*, vol. 21, 2008.

[8] X. Zheng, B. Aragam, P. K. Ravikumar, and E. P. Xing, "DAGs with NO TEARS: Continuous optimization for structure learning," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[9] X. Zheng, C. Dan, B. Aragam, P. Ravikumar, and E. Xing, "Learning sparse nonparametric DAGs," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 3414–3425.

[10] Y. Yu, J. Chen, T. Gao, and M. Yu, "DAG-GNN: DAG structure learning with graph neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 7154–7163.

[11] S. Lachapelle, P. Brouillard, T. Deleu, and S. Lacoste-Julien, "Gradient-based neural dag learning," *arXiv preprint arXiv:1906.02226*, 2019.

[12] I. Ng, S. Zhu, Z. Fang, H. Li, Z. Chen, and J. Wang, "Masked gradient-based causal structure learning," in *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*. SIAM, 2022, pp. 424–432.

[13] M. Scutari, "Learning Bayesian networks with the bnlearn R package," *arXiv preprint arXiv:0908.3817*, 2009.

[14] A. Sharma and E. Kiciman, "DoWhy: An end-to-end library for causal inference," *arXiv preprint arXiv:2011.04216*, 2020.

[15] K. Zhang, S. Zhu, M. Kalander, I. Ng, J. Ye, Z. Chen, and L. Pan, "gCastle: A python toolbox for causal discovery," *arXiv preprint arXiv:2111.15155*, 2021.

[16] A. C. Constantinou, Z. Guo, and N. K. Kitson, "The impact of prior knowledge on causal structure learning," *arXiv preprint arXiv:2102.00473*, 2021.

[17] J. Chowdhury., R. Rashid., and G. Terejanu., "Evaluation of induced expert knowledge in causal structure learning by notears," in *Proceedings of the 12th International Conference on Pattern Recognition Applications and Methods - ICPRAM*, INSTICC. SciTePress, 2023, pp. 136–146.

[18] U. Hasan and M. O. Gani, "KGS: Causal discovery using knowledge-guided greedy equivalence search," *arXiv preprint arXiv:2304.05493*, 2023.

[19] D. P. Bertsekas, "Nonlinear programming," *Journal of the Operational Research Society*, vol. 48, no. 3, pp. 334–334, 1997.

[20] A. Reisach, C. Seiler, and S. Weichwald, "Beware of the simulated dag! causal discovery benchmarks may be easy to game," *Advances in Neural Information Processing Systems*, vol. 34, pp. 27 772–27 784, 2021.

[21] M. Kaiser and M. Sipos, "Unsuitability of NOTEARS for causal graph discovery," *arXiv preprint arXiv:2104.05441*, 2021.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[23] A. H. Petersen, J. Ramsey, C. T. Ekstrøm, and P. Spirtes, "Causal discovery for observational sciences using supervised machine learning," *arXiv preprint arXiv:2202.12813*, 2022.

[24] A. R. Nogueira, A. Pugnana, S. Ruggieri, D. Pedreschi, and J. Gama, "Methods and tools for causal discovery and causal inference," *Wiley interdisciplinary reviews: data mining and knowledge discovery*, vol. 12, no. 2, p. e1449, 2022.

[25] I. Guyon, C. Aliferis, G. Cooper, A. Elisseeff, J. P. Pellet, P. Spirtes, and A. Statnikov, "Causality workbench," in *Causality in the Sciences*. Oxford University Press, 2011.

[26] S. L. Lauritzen and D. J. Spiegelhalter, "Local computations with probabilities on graphical structures and their application to expert systems," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 50, no. 2, pp. 157–194, 1988.

[27] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[28] Dutch Central Bureau for Statistics, "Volkstelling, 2001," 2001.

[29] M. Lichman, "UCI machine learning repository," http://archive.ics.uci.edu/ml, 2013.

[30] L. Zhang, Y. Wu, and X. Wu, "Causal modeling-based discrimination discovery and removal: criteria, bounds, and algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 11, pp. 2035–2050, 2018.