

基于增强条件独立性检验的鲁棒因果发现算法*

郝志峰^{1,2}, 汪菲霞¹, 陈正鸣¹, 乔杰¹, 蔡瑞初¹

¹(广东工业大学 计算机学院, 广东 广州 510006)

²(汕头大学 理学院, 广东 汕头 515063)

通信作者: 蔡瑞初, E-mail: cairuichu@gmail.com



摘要: 因果关系发现旨在从观测数据中发现变量间的因果关系, 是帮助我们理解自然界、社会和技术系统中各种现象和变化的重要方法。一种主流的因果发现方法是基于约束的算法, 这类算法通过检验变量间的条件独立性关系来确定变量之间的因果结构。然而, 现实世界的的数据收集往往受资源或技术的限制, 面临样本量有限, 节点方差大等挑战。在这些场景下, 条件独立性检验的正确率受到极大影响, 导致学到的因果图中部分变量的因果边被错误地删除, 影响了算法输出的准确性。为此, 提出一种增强的条件独立性检验的方法, 该方法的核心在于尽可能减少无关外部噪声对于待测试变量的干扰, 从而提高条件独立性检验结果的准确性。基于该增强的条件独立性检验方法, 提出一种基于启发式搜索的结构学习算法, 该算法在初始结构图的基础上, 迭代搜索被误删的因果边, 基于增强的条件独立性检验并结合得分优化的思想, 重构因果结构。实验结果显示, 相较于现有方法, 所提算法在仿真数据、贝叶斯网络数据以及真实数据上的 $F1$ 值和结构汉明距离 (SHD) 均有显著提升, 证明在有限样本和因果结构中存在高方差节点的条件下更准确地揭示观测数据中潜在的真实因果结构的能力。

关键词: 因果结构学习; 有限样本量; 高方差节点; 增强条件独立性检验

中图分类号: TP18

中文引用格式: 郝志峰, 汪菲霞, 陈正鸣, 乔杰, 蔡瑞初. 基于增强条件独立性检验的鲁棒因果发现算法. 软件学报. <http://www.jos.org.cn/1000-9825/7261.htm>

英文引用格式: Hao ZF, Wang FX, Chen ZM, Qiao J, Cai RC. Robust Causal Discovery Algorithm Based on Enhanced Conditional Independence Tests. Ruan Jian Xue Bao/Journal of Software (in Chinese). <http://www.jos.org.cn/1000-9825/7261.htm>

Robust Causal Discovery Algorithm Based on Enhanced Conditional Independence Tests

HAO Zhi-Feng^{1,2}, WANG Fei-Xia¹, CHEN Zheng-Ming¹, QIAO Jie¹, CAI Rui-Chu¹

¹(School of Computer, Guangdong University of Technology, Guangzhou 510006, China)

²(College of Science, Shantou University, Shantou 515063, China)

Abstract: Causal discovery aims to uncover causal relationships among variables from observational data, serving as a crucial method for understanding various phenomena and changes in natural, social, and technological systems. A mainstream approach for causal discovery is a constraint-based algorithm, which determines the causal structure among variables by examining their conditional independence. However, data collection in the real world often faces challenges such as limited sample sizes and high variance among nodes due to resource or technical constraints. In these scenarios, the accuracy of conditional independence tests is greatly affected, leading to erroneous deletion of causal edges of some variables in learned causal graphs, thereby impacting the accuracy of the algorithm's output. To address this issue, this study proposes an enhanced method for conditional independence testing, which focuses on minimizing the interference of irrelevant external noise on the variables being tested, thereby improving the accuracy of conditional independence tests. Based on this enhanced method, the paper introduces a structure learning algorithm based on heuristic search, which iteratively searches for mistakenly deleted causal edges on a graph with an initial structure. This algorithm reconstructs the causal structure by combining enhanced

* 基金项目: 新一代人工智能国家科技重大专项 (2021ZD0111501); 国家优秀青年科学基金 (62122022)

收稿时间: 2024-04-08; 修改时间: 2024-06-06; 采用时间: 2024-07-25; jos 在线出版时间: 2024-12-25

conditional independence tests with score optimization. Experimental results show that, compared to existing methods, the proposed algorithm significantly improves both the $F1$ score and the structural Hamming distance (SHD) on simulated, Bayesian network, and real data, demonstrating its ability to more accurately reveal underlying causal structures in observational data with limited samples and high-variance nodes.

Key words: causal structure learning; limited sample size; high-variance node; enhanced conditional independence test

因果关系发现是一种研究复杂系统中变量之间相互作用本质的重要方法之一^[1,2]。这类方法通常利用统计模型、机器学习技术以及随机实验等手段,分析并识别数据中变量间的因果结构。通过学习变量之间的因果关系,能够帮助我们深入理解系统内部变量的作用机制及其相互之间的作用关系,进而做出更有效的决策和预测。近年来,在生态气候学^[3,4]、生物基因学^[5-9]、社交网络^[10-12]、人工智能^[13-16]以及经济学^[17-19]等多个领域得到广泛的关注和应用。

为了能够从数据中学习变量间的因果关系,常用的一类方法是基于约束的因果发现算法,其中,PC (Peter-Clark) 算法^[20]是该方法的典型代表。它通过对数据中的变量执行条件独立性检验 (conditional independence test, CIT) 来判断变量之间是否存在直接因果边,并综合结果来学习因果结构^[21,22]。在这个过程中,CIT 的性能至关重要,因为 CIT 的准确结果直接影响算法能否得到正确的结果。其算法的正确性通常基于一个关键假设:即对数据执行的所有 CIT 都必须是准确的。然而,面对现实世界中样本量有限和噪声方差差异大等现实挑战时,CIT 的结果往往变得不可靠,进而导致因果结构学习算法得到的结果出现错误,算法输出结果的可靠性受到质疑^[23-25]。

为提升基于约束方法在实际场景下的可靠性,研究者们从高阶 CIT 困扰^[26]、条件独立性冲突^[27]和对称性校验^[28]等方面进行了算法的优化。在阶 CIT 困扰的研究中,认为 CIT 的结果不可靠是由于执行过程中条件集过大而导致的^[26,29]。因此研究者们采取诸如限制条件集大小^[29]、设定数据实例数量^[28,30]等条件,以降低当条件集过大时 CIT 的错误判断风险。而条件独立性冲突的研究中认为 CIT 执行得到的不确定结果可以通过一种图形推理^[27,31]的方法纠正,如 Hyttinen 等人^[27]引入了 SAT 答案集编码,通过评估所有可能的结果排除逻辑上不一致的结果,从而保留更加可信的关系。该技术后续被扩展,以适应更加宽松的假设条件^[31]。此外,在对称校验的研究中认为可以通过结构中因果边的对称性校验的方法纠正错误的关系^[23]。该方法采用了一种极端的处理策略,在检测到不对称性时,为了减少假阳性关系则拒绝所有不对称因果边^[28,30]或减少假阴性关系接受所有因果边^[32]。虽然上述方法一定程度上提升了因果发现算法的性能,但在真实因果结构中存在高方差节点的情况下,这些方法仍无法有效解决 CIT 结果不可靠的问题。

当真实的因果结构中存在高方差节点时,这些高方差节点会影响其子节点与其他父节点之间执行 CIT 检验的结果,导致 CIT 无法有效检验真实的因果关系,进而影响因果结构学习的准确性。以图 1 为例,第 1 行表示真实的因果结构 $V_1 \rightarrow V_2 \leftarrow V_3$, 其中 Var 表示节点方差,所有边系数为 1,样本量为 1 000。第 2 行描述了在未考虑节点 V_3 的情况下,其子节点 V_2 与节点 V_1 的 CIT 检验结果。在图 1(a) 中,当节点 V_3 方差为 1 时,CIT 检验得到相关系数为 0.40, p 值为 0.0, 此时有效识别出 V_1 与 V_2 的因果边。但在图 1(b) 中,将节点 V_3 的方差增加到 10, V_1 和 V_2 的相关系数减小到 0.049, p 值上升到 0.09, 此时该因果边就存在被删除的风险。进而在图 1(c) 中,随着节点 V_3 方差继续增大,CIT 判定 V_1 与 V_2 相互独立,从而在学习到的因果结构中删除了它们之间的因果边。当节点数量增加且样本量受限时,因果结构中高方差节点的影响会进一步增大,使得 CIT 的准确性大幅下降,进而导致算法错误地删除在真实因果结构中的重要因果边。

为了减轻高方差节点对子节点的 CIT 结果的影响,本文提出了“增强条件独立性检验”方法,以适当地消除 CIT 检验中存在的高方差节点噪声;并进一步提出在样本受限和真实结构中存在高方差节点条件下识别和学习因果结构的算法。此算法在启发式搜索的框架下,通过 PC 算法得到的部分有向无环图的初始结果,迭代进行增强的独立性检验和优化最小结构依赖得分,逐步恢复 PC 算法输出结果中缺失的因果边,从而得到关于数据的可靠因果结构图。本文从理论保障了基于增强的条件独立性检验和最小化结构依赖可以正确找回结构中丢失的因果边,并在仿真数据和真实数据的结果进一步验证了所提出算法的正确性和有效性。

本文第1节介绍因果结构学习的相关方法和研究现状. 第2节介绍本文所需的基础知识, 包括因果模型和问题定义. 第3节介绍本文提出的最大最小结构依赖的因果发现算法. 第4节通过仿真实验、贝叶斯网络数据和真实数据实验验证了所提模型的有效性. 第5节全面讨论了本文提出方法的局限性和一些拓展的可能性.

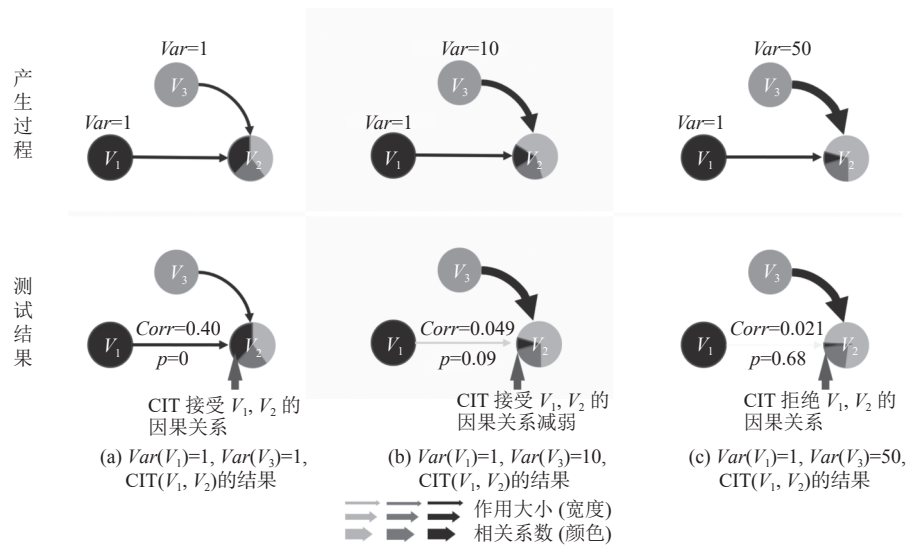


图1 高方差节点对其子节点与其他父节点 CIT 结果的影响

1 相关工作

针对非时序观测数据的因果结构学习方法主要包括基于约束、基于评分和基于函数因果模型的方法.

基于约束的方法主要分为全局因果结构学习 (global structure learning, GSL) 方法和从局部到全局因果结构学习 (local_to_global structure learning, LGSL) 方法. 其中, GSL 方法一次性学习整个网络的因果结构. 这类方法从全连接图开始, 对所有变量使用条件独立性检验确定连通性, 然后应用 V 结构^[20]和 Meek 规则^[21]定向边, 从而构建表示观测变量的部分有向无环图 (completed partially directed acyclic graph, CPDAG). 代表性算法包括 IC^[33]和 PC^[20]算法及其变体 PC_Stable^[34]等. LGSL 方法通过分解学习过程从而学习整个网络的因果结构. 这类方法首先独立的确定每个变量的马尔可夫毯 (Markov blanket, MB)^[35], 并进一步将每个变量的 MB 集拼接为无向图的骨架, 随后通过评分或结构识别方法定向因果边. 其代表性算法有 GSB^[23]、SLL+C/G^[32]和 GGSL^[28]等. 基于评分的方法如 MMHC^[30]和 GES^[36]等侧重于通过定义和优化评分函数确定最佳的因果结构.

基于函数模型的方法主要包括线性非高斯无环模型 (linear non-Gaussian acyclic model, LiNGAM)^[37]和非线性加性噪声模型 (additive noise model, ANM)^[38], 此类方法通常要求噪声项与原因变量独立, 来保证因果方向的可靠识别性^[39]. 其中, LiNGAM 模型通过主成分分析等方法求解, 其假设数据之间的关系是线性的并且噪声相互独立且服从非高斯分布. ANM 模型假设数据之间的关系是非线性的, 并且噪声相互独立^[40].

上述方法对于因果结构学习在理论和实践运用中做出了巨大贡献, 但面对复杂的网络结构, 有限的样本量, 数据方差差异大等问题, 算法的可靠性常受到质疑^[23-25]. 因此, 后续研究者提出了多种策略以提升算法在上述问题下的实用性和准确性. 例如, Triantafillou 等人^[41]提出 COmbINE 算法, 通过对多重干预数据集来生成符合所有输入数据的因果模型概要, 以提高实际数据适用性和效率. Ng 等人^[42]通过引入超结构估计方法和局部搜索策略, 以改善在大规模线性高斯模型下基于评分的结构学习方法的可扩展性和准确性. Hyttinen 等人^[27]引入了 SAT 答案集编码方法来全面探索可能的关系, 通过识别并丢弃于答案集逻辑不一致的结果, 优先保留更可信的检验, 以进一步扩展到更宽松的假设中^[31]. LGSL 方法受到数据样本、噪声等影响, 在骨架拼接过程中遇到了局部结构不对称问

题,传统的 LGSL 方法通常采用统一的接收或拒绝策略处理局部结构中出现的不对称因果边;为了改进这一流程,Guo 等人^[43]基于 MMHC 方法进一步提出 ADL 算法,该算法在拼接局部结构时引入了自适应的 AND 或 OR 规则,优化了局部结构拼接中因果边的保留或移除的决策过程.

2 因果模型与问题定义

本研究关注在样本受限和真实因果结构中存在高方差节点的情况下,在基于约束的因果发现框架进行因果结构学习的问题.本文的理论结果和可靠性算法建立在因果图模型之上.因此,本节首先介绍因果图模型中常用的基本定义和符号系统(第 2.1 节),并进一步说明当真实因果结构中存在高方差节点时,基于约束的因果发现算法所存在的问题(第 2.2 节).

其中本文涉及的符号及其含义如表 1 所示.

表 1 符号和对应描述

符号	含义
n	表示变量维度/节点个数
m	表示每个变量的样本量大小
D	具有 n 个随机变量的数据集
V	D 中 n 个随机变量的集合
V_i, V_j	V 中的单个变量 ($i, j = 1, 2, \dots, n$)
\hat{V}_i, \hat{V}_j	\hat{V}_i, \hat{V}_j 表示对 V_i, V_j 回归后的变量
S	V 内的条件集
$V_i \perp V_j S$	V_i 和 V_j 给定集合 S 独立
$dSet(V_i, V_j)$	V_i 与 V_j 的 d-分离集
TrueGraph, 简写 G^T	真实因果结构图 true graph
DAG, 简写 G	有向无环图 DAG
CPDAG, 简写 G^{pd}	部分定向 (partially directed) 的有向无环图 CPDAG
V_{tail}	对于图 G^{pd} 上的有向边 $V_i \rightarrow V_j, V_{tail} = V_j$
$Pa(G, V_i)$	V_i 在 G 上的父节点, 表示 $Pa(G, V_i) = \{V_j V_j \rightarrow V_i\}$
$Adj(G, V_i)$	V_i 在 G 上的邻接节点, 表示 $Adj(G, V_i) = \{V_j V_j \rightarrow V_i \vee V_j \leftarrow V_i \vee V_j - V_i\}$
$\overline{Adj}(G, V_i)$	V_i 在 G 上的非邻接节点, 表示 $\overline{Adj}(G, V_i) = \{V_j V_j \neq V_i\}$

2.1 背景知识

在因果图模型中,通常将变量间的因果关系用有向无环图 (direct acyclic graph, DAG)^[33]表示,该图也称为因果图,因果图 $G = \{V, E\}$,由结点集合 V 和边集 E 组成.因果图中有向边满足 $(V_j, V_i) \in E, (V_i, V_j) \notin E$ 约束,其中 V_j 是 V_i 的父亲节点,记为 $V_j \rightarrow V_i$,边的权重代表 V_j 到 V_i 的因果强度大小.一般来说,基于约束的因果结构学习算法的正确性在因果充分性、马尔可夫性和因果忠诚性等假设下得到保障^[33].

假设 1. 因果充分性假设^[33]. 当变量集 V 中的任意两个变量的直接原因变量都存在 V 中时,变量集 V 就被认为是因果充分的.

假设 2. 因果马尔可夫假设^[33]. 对于具有因果充分性的变量集 V 而言,在已知变量的父亲节点条件下,所有变量与他们的非后裔节点互相条件独立.

假设 3. 因果忠诚性假设^[33]. 数据集 D 的分布忠诚于因果图 G ,对于任意的 $V_i, V_j \in V (i \neq j)$ 和集合 $S \in V$,如果 V_i, V_j 在给定集合 S 的变量下条件独立,那么 V_i, V_j 被集合 S 中的变量 d-分离;

在上述的假设下,基于约束的因果结构学习算法通过检验数据中变量的条件独立性关系,来确定因果结构图

中变量之间是否存在直接的因果边. 该概念通常被表示为图上的一种 d -分离关系, 具体定义如定义 1 所示.

定义 1. d -分离^[33]. 对于给定的有向无环图 $G = \{V, E\}$, 当 G 上的非邻接结点 $V_i, V_j \in V$ 的所有连通路径被集合 Z 阻断, 那么我们说集合 Z 在图 G 中 d -分离了非邻接节点对, 并且 Z 需要满足:

- (1) 路径上的每个非对撞结点 $W \in E$.
- (2) 路径上的每个对撞结点或者对撞结点的后代 $W \notin E$.

通过迭代地确定变量间的 d -分离关系, Pearl 给出了基于约束方法的识别性结果^[33], 即数据的因果结构图可以被识别为马尔可夫等价类 (Markov equivalence classes).

定义 2. 马尔可夫等价类^[33]. 对于有向无环图 $DAG = \{V, E\}$, 若存在一组 $DAGs$ 在条件独立性上不可区分, 则这些 $DAGs$ 构成一个马尔可夫等价类, 所有属于同一马尔可夫等价类的 DAG 具有相同的骨架 (skeleton) 和相同的 V 结构 (对撞结构).

马尔可夫等价类可以进一步被表示为一个部分有向无环图 (partially directed acyclic graph, PDAG) 的, 其中有向边表示所有等价类的图都具有相同的因果方向.

定义 3. CPDAG 图^[33]. CPDAG 是一个图形表示, 用于可视化属于同一马尔可夫等价类的 DAG , 它结合了有向边和无向边来描绘因果关系, 其中,

- (1) 每个有向边 $V_j \rightarrow V_i$, 存在于属于等价类的每个 DAG 中.
- (2) 每个无向边 $V_j - V_i$, 存在某个等价类表示为 $V_j \rightarrow V_i$ 和某个等价类中表示为 $V_j \leftarrow V_i$.

2.2 问题定义

为了研究真实因果结构中存在高方差节点时因果关系发现存在的问题, 在本研究中, 假设数据生成过程遵循线性无环因果结构方程模型^[37]定义, 表示为:

$$V_i = \sum_{V_j \in Pa(G^T, V_i)} b_{ij} V_j + \varepsilon_{V_i} \quad (1)$$

其中, $i = 1, 2, \dots, n$; $Pa(G^T, V_i)$ 表示 V_i 在真实因果结构图 G^T 的父节点. b_{ij} 指 V_i 和其父节点 V_j 的因果边强度, 噪声项 ε 互相统计独立.

线性无环因果结构方程模型建模数据的产生过程为一种加性的线性函数方程, 具有易于分析, 简洁有效的特点, 在基因表达^[7]、生物化学研究^[8]、医疗病理分析^[9]、社会学分析^[10,11]等邻域的建模中都得到了广泛的应用. 特别地, 在线性模型下, 偏相关检验是常用的条件独立性检验工具. 在偏相关检验中主要涉及计算 Pearson 相关系数^[44]和进行 Fisher_Z 检验^[45]两个过程. 本文提供的理论分析与示例都是基于这些过程所获得的结果. 接下来, 通过示例具体分析图 1 结构下 CIT 的误差原因.

例 1: 因果结构中存在高方差节点时 CIT 误差. 考虑 (图 1) 的真实因果结构 $V_1 \rightarrow V_2 \leftarrow V_3$, 其中 V_1, V_2 的数据生成公式为: $V_1 = \varepsilon_{V_1}$, $V_2 = b_{23} V_3 + b_{21} V_1 + \varepsilon_{V_2}$, b_{23} 和 b_{21} 表示因果边强度, ε_{V_1} 和 ε_{V_2} 代表独立噪声. 计算 V_1, V_2 的相关系数得到公式 (2):

$$\text{Corr}(V_1, V_2) = \frac{\text{Cov}(V_1, V_2)}{\sigma_{V_1} \sigma_{V_2}} = \frac{b_{21} \sqrt{\text{Var}(\varepsilon_{V_1})}}{\sqrt{\text{Var}(b_{23} V_3 + b_{21} V_1 + \varepsilon_{V_2})}} \quad (2)$$

其中, $\text{Cov}(V_1, V_2)$ 是 V_1, V_2 的协方差, $\sigma_{V_1}, \sigma_{V_2}$ 分别是 V_1, V_2 的标准差, $b_{21} \sqrt{\text{Var}(\varepsilon_{V_1})}$ 表示变量 V_1 的标准差和对 V_2 的因果边强度 b_{21} 的乘积, $\sqrt{\text{Var}(b_{23} V_3 + b_{21} V_1 + \varepsilon_{V_2})}$ 表示此时节点 V_2 的标准差.

一般来说, 相关系数越趋近于 1, 代表变量之间的线性相关性越强; 相关系数趋近于 0, 代表两个变量之间的线性相关性越差, 即趋近于相互独立. 对图 1(c) 结构的变量 V_1, V_2 有在真实因果结构中有因果边 $V_1 \rightarrow V_2$, 因此在线性模型下, 两者是线性相关的, 即 $\text{Corr}(V_1, V_2)$ 的结果趋近于 1. 然而, 由于 $\text{Corr}(V_1, V_2)$ 的结果同时受到 $b_{21} \sqrt{\text{Var}(\varepsilon_{V_1})}$ 与 $\sqrt{\text{Var}(b_{23} V_3 + b_{21} V_1 + \varepsilon_{V_2})}$ 两项的影响, 随着父节点 V_3 方差增大, 会使得公式 (2) 中的分母项明显大于分子项, 导致 $\text{Corr}(V_1, V_2)$ 的结果反趋向 0, 因此 CIT 将大概率拒绝 V_1, V_2 的相关关系, 而拒绝这条因果边.

例 1 说明在执行 CIT 检验时, 真实因果结构中的高方差父节点会影响子节点与其方差较小的父节点的 CIT 结果, 错误地判定它们相互独立. 下面进一步讨论基于“条件独立性冲突”“对称性校验”等方法在解决因果结构中存在高方差节点的场景下所面临的局限性. 如图 2 所示, 真实结构 (图 1(c) 产生过程) 通过 Fisher_Z 检验得到 V_1, V_2 的相关关系结果为 0.021, p 值为 0.68. 基于高阶 CIT 的方法通常限制条件集 $Z=1$, 但由于变量在给定空集下判断为独立, 算法便停止了下一步计算, 如图 2(b) 所示. 基于对称性校验的方法通过 CIT 获得每个变量的局部结构, 但由于 CIT 错误判断 V_1, V_2 相互独立, 所以得到的局部结构关系对称, 因此 V_1, V_2 的因果边不会被对称校验的方法进一步检验而找回, 如图 2(c) 所示. 基于条件独立性冲突的方法对当前结构执行所有可能的 CIT, 因此得到 $V_1 \perp V_2$ 和 $V_1 \perp V_2 | V_3$ 两种独立性冲突的结果, 进而通过一些逻辑规则或二次加权计算选择更可信的 CIT 结果. 但这些复杂的运算和设计难以保证正确的 CIT 结果被选中的可能性, 如图 2(d). 其中在真实结构的因果连通路中 V_3 并不是 V_1, V_2 的连通节点, 引入它进入约束集实际上不符合 d -分离的定义. 并且这类方法由于涉及复杂的逻辑推理判断, 节点数量常限制在 10 个以内, 算法十分复杂, 难以复现并投入实际应用.

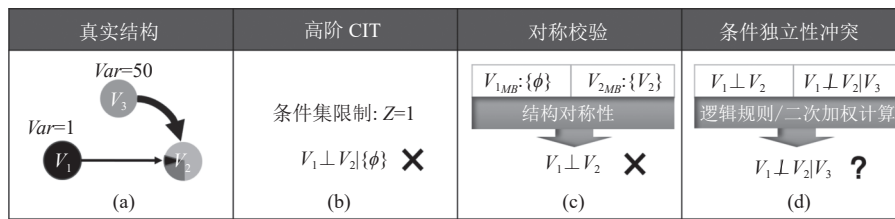


图 2 基本方法的局限性

目前的方法无法有效找回被错误删除的因果连接边. 随着网络结构中节点数量增加, 高方差节点对于 CIT 结果准确性的影响逐渐加剧, 这对于准确识别和学习变量间的因果关系带来了困难.

因此, 本文所研究的问题为: 如何在有限样本和真实因果结构存在高方差节点条件下, 提高 CIT 的准确率, 减少因果图中遗漏的关键因果关系, 更准确地揭示观测数据中潜在的真实因果结构.

3 最大最小依赖的因果结构学习算法

本节针对有限样本和真实因果结构存在高方差节点条件下, CIT 结果不准确导致因果边误删的问题, 提出了一种基于“增强条件独立性检验”的约束类方法. 该方法的基本思想是利用一次检验得到的部分先验结构信息, 通过线性回归消除潜在的部分高方差父节点影响后, 再次对节点间的因果关系进行检验来增强 CIT 结果的可靠性 (第 3.1 节). 并将这一思想融入启发式搜索框架, 通过优化图结构依赖得分 (第 3.2 节), 进一步设计了最大最小依赖的因果结构学习算法 (max-min dependency causal learning, MMDCL). 下面给出算法框架概述, 随后在第 3.1 节和第 3.2 节中对算法的实现细节进行研究和提供必要的理论保证.

MMDCL 算法由“领域搜索”和“结构更新”两个阶段组成, 如图 3 所示. 在领域搜索阶段, 算法的目标是从当前结构出发探索可能的候选邻域, 利用增强的条件独立性检验“找回”可能被误删的因果边. 在结构更新阶段, 算法的目标是从上一阶段获得的候选结构中“选择”最接近潜在真实因果结构的候选结构, 并完成结构更新.

算法 1 展示了 MMDCL 的伪代码, 该算法首先将 PC 算法输出的结构作为初始起点, 即 G^{pd} , 并获得初始化的非邻接节点集和分离集, 随后进入迭代过程 (第 1-5 行). 在每轮迭代中, 算法首先确定所有非邻接变量间可以添加的因果边作为潜在搜索空间 (第 7 行). 对这些潜在因果边进行环路检测, 确保新加入的因果边不会形成环路, 并对通过二次的增强 CIT 判断这些因果边是否应被加入结构中 (第 8 行). 算法接着评估候选结构的结构依赖, 寻找具有最小结构依赖的候选结构 (第 9-14 行), 以此更新当前最优结构 (第 19 行). 迭代持续进行, 直到没有新的因果边可以被添加或当前结构不再更新, 此时迭代停止 (第 21 行). 其中, 对于判断环路的部分, 为了尽可能避免出现环路, 算法将 CPDAG 上的无向边视为双向连通的. 接下来本文将在第 3.1 节和第 3.2 节中详细介绍领域搜索和结构更新的完整过程.

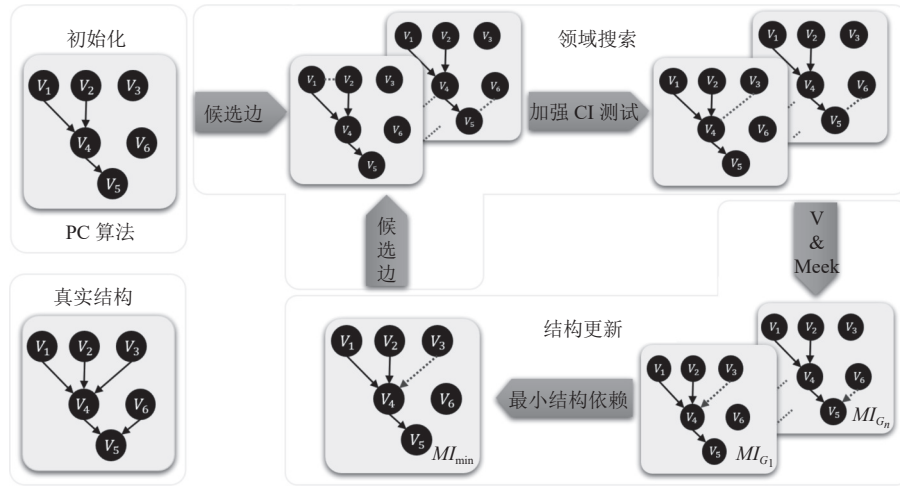


图3 MMDCL 算法框架

算法 1. MMDCL 算法.

输入: 数据 D , 经典算法 Pc ;

输出: 因果结构图 G .

1. 基于 PC 算法学习初步图 G^{pd} , 保存非邻接节点 $\overline{Adj}(G^{pd}, V_i), i = 1, 2, \dots, n$ 和分离集 $dSet$
 2. 初始化图结构 G^{pd} , 和结构依赖 $\min MI_{G^{pd}} = \sum_i^n MI_i(D, V_i, V_j | dSet(V_i, V_j)), V_j \in \overline{Adj}(G^{pd}, V_i)$
 3. 设置循环标志 $Loop = True$
 4. **Repeat**
 5. $Loop = False$
 6. **For** $V_i \in V$
 7. **For** 获取非邻居节点 $V_j \in \overline{Adj}(G, V_i)$ 和分离集 $dSet(V_i, V_j)$
 8. **If** (V_i, V_j) 无环且通过加强 CIT
 9. 对 G^{pd} 添加无向边 (V_i, V_k) , 利用 V 结构和 Meek 规则定向, 保存为候选图 G'^{pd}
 10. 通过最小结构依赖计算 $MI_{G'^{pd}}, dSet'$
 11. **If** $MI_{G'^{pd}} \leq \min MI_G$
 12. 更新 $\min MI_G = MI_{G'^{pd}}$
 13. 存储最优邻居 $mG = G'^{pd}, m\overline{Adj} = \overline{Adj}(G^{pd}, V_i)' / (V_i, V_j), mdSet = dSet'$
 14. $Loop = True$
 15. **End if**
 16. **End for**
 17. **End for** /*结束对当前节点的搜索*/
 18. **End for** /*结束对当前结构的搜索*/
 19. **If** $Loop$
 20. 更新结构 $G^{pd} = mG, \overline{Adj}(G^{pd}, V_i) = m\overline{Adj}, dSet = mdSet$
 21. **End if**
 22. **Until** $Loop = False$ /*循环标志无法刷新时停止*/
-

3.1 增强条件独立性检验的领域搜索

在邻域搜索阶段,算法的核心是从当前得到的部分有向无环图 G^{pd} 出发,搜索下一步可能候选结构.考虑到在当前结构中可能存在部分节点的因果边因 CIT 的错误判断而被误删的问题,因此,本阶段对当前结构中所有非邻接节点进行二次、增强的 CIT 检验来重新评估它们之间的因果关系,从而有效识别并恢复节点间被误删的因果边.

为了最大可能地找回被误删的因果边,算法将候选的搜索空间设置为当前结构中所有非邻接节点之间的潜在因果边,并通过结构先验信息消除待检验节点存在的外部父节点影响,进行二次检验判断当前节点之间的因果关系是否可以被其邻居节点 d -分离.本文形式化该过程为增强条件独立性检验,如定义 4 所示.

定义 4. 增强的条件独立性检验.数据集 D 表示因果图结构 $G^{pd} = \{V, E\}$ 的节点和边的集合,其中图的节点对应数据中的观测变量,即 V_i 表示 X_i . 在当前的结构 G^{pd} 上,对于任意节点 $V_i \in V$ 的非邻接节点 $V_j \in \overline{Adj}(G, V_i)$ 表示为 (V_i, V_j) ,关于 V_i, V_j 的增强条件独立性检验定义如下:

$$\hat{V}_i \perp \hat{V}_j | dSet(V_i, V_j) \parallel \hat{V}_i \perp \hat{V}_j | dSet(V_i, V_j) \cup S,$$

其中,对 V_i 和 V_j 分别回归图 G^{pd} 中已知的父节点得到 \hat{V}_i 和 \hat{V}_j ; S 是 Nei 的非空子集, Nei 表示 V_i, V_j 在当前结构中排除了已经过回归处理的节点后的邻接节点集合,公式表示为 $S \subseteq Nei, Nei = \{(Adj(G^{pd}, V_i) \vee Adj(G^{pd}, V_j)) - (Pa(G^{pd}, V_i) \vee Pa(G^{pd}, V_j))\}$,其中 \hat{V}_i 和 \hat{V}_j 表示为公式 (3):

$$\hat{V}_i = V_i - \sum_{V_k \in Pa(G^{pd}, V_i)} b_{ik} V_k, \quad \hat{V}_j = V_j - \sum_{V_k \in Pa(G^{pd}, V_j)} b_{jk} V_k \quad (3)$$

通过对待检验节点尽可能消除高方差父节点引入的噪声能有效提升 CIT 检验结果的准确性,如图 4 所示,对待检验变量 V_1, V_2 , 当对 V_2 通过多元线性回归消除先验父节点 V_3 得到 \hat{V}_2 , 此时 CIT 检验的结果从存在高方差节点 V_3 的影响下检验的 $p = 0.68$ 恢复至 $p = 0.0$, 准确率显著提升并成功找到了 V_1, V_2 之间的因果边.

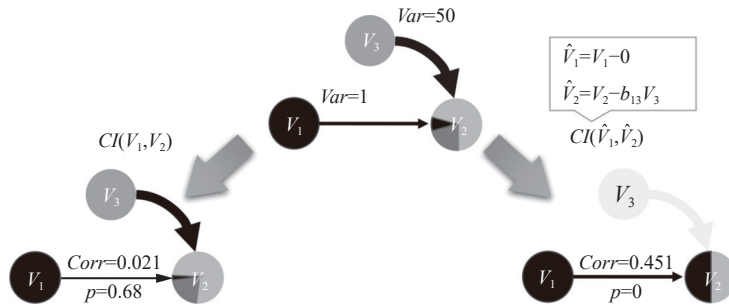


图 4 消除高方差父节点后的 CIT 检验结果

然而,通过回归消除后的二次 CIT 检验识别得到的因果关系存在两种可能: (1) 在真实因果结构中存在直接影响的因果关系,如图 5(c) 中的变量 V_3, V_4 . (2) 在真实因果结构中存在连通路径的间接影响的因果关系,如图 5(d) 中的变量 V_3, V_5 .

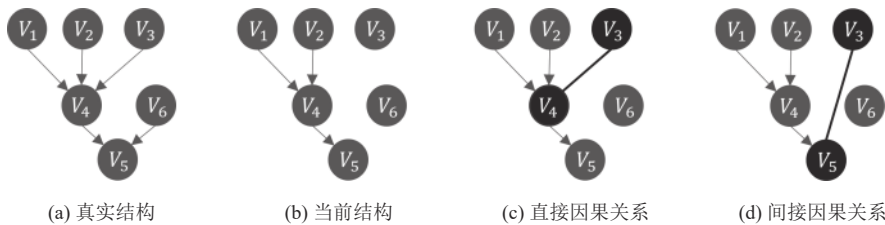


图 5 回归检验后得到两种可能因果关系

为了区分这些可能存在的因果关系, 算法需要进一步执行给定邻接节点集 Nei 的条件独立性检验, 这是为了保证节点间存在直接因果关系, 即节点在图结构上是直接相连的. 例如, 在图 5(d) 中, 通过节点 V_4 阻断了节点 V_3, V_5 的连通路程, 从而排除了节点 V_3, V_5 存在直接因果关系, 并不考虑添加这条因果边, 即定义 4 中的第 2 个条件 $(\hat{V}_i \perp \hat{V}_j | dSet(V_i, V_j) \cup S)$.

通过对所有非邻接节点执行增强的条件独立性检验, 可以有效地找回在当前结构中可能遗漏的因果边. 为了进一步保证结构更新的准确性, 算法将每个新增了一条因果边的图结构作为一个候选结构, 并针对这些候选结构进行下一步的最优选择和结构更新. 在处理线性数据时, 本研究采用多元线性回归来消除已知的外部噪声. 尽管本文的数据生成过程遵循线性模型, 但对于非线性或其他类型的数据, 本文提出的算法框架同样适用, 可以根据数据特性选择多项式回归、梯度提升回归树 (GBRT) 等方法.

3.2 最小结构依赖的结构更新

在结构更新阶段, 算法的核心任务是从邻域搜索阶段获得的多个候选结构中“选择”最优结构, 并实现结构更新. 为了评估出最优结构, 算法首先利用定向规则确定候选结构中无向边的方向, 然后通过定义的图结构依赖得分和最小结构依赖逼近准则实现对于最优结构的选择.

算法首先应用 V 结构^[9]和 Meek 规则^[21]明确图中某些无向边的方向, 从而得到可以唯一表示当前候选邻域结构的 CPDAG 图, 其中 V 结构的具体定义如定义 5 所示.

定义 5. V 结构^[9]. 在因果图 G 中, 若两个节点 V_i, V_j 不直接相连, 但拥有一个共同邻居 V_k , 并且满足 $V_i \perp V_j | dSet(V_i, V_j)$ 和 $V_i \perp V_j | (dSet(V_i, V_j) \cup V_k)$, 则 V_i, V_j 和 V_k 构成 V 结构, 记作 $V_i \rightarrow V_k \leftarrow V_j$.

根据定义 4 可知邻域搜索阶段找回的因果边需要尽可能剔除已知先验父节点的噪声, 所以其对应的两端节点中至少有一端具有已知的先验父节点, 因此该因果边可以通过 V 结构定义确定方向, 进而利用 Meek 规则^[21]可以进一步定向结构中的剩余无向边, 并获得可以唯一表示该领域结构的 CPDAG 图. 下面为了从多个候选结果中选出最优结构, 定义了图结构依赖得分作为评价指标, 定义如定义 6.

定义 6. 图结构依赖得分. 对数据集 D 和算法获得的当前结构 G^{pd} , G^{pd} 的图结构依赖得分表示为对所有节点 $V_i \in V$, 与其非邻接节点 $V_j \in \overline{Adj}(G^{pd}, V_i)$, 在给定对应的分离集 $dSet(V_i, V_j)$ 下的依赖得分之和, 公式为:

$$MI_{G^{pd}} = \sum_i^n MI_i(D, V_i, V_j | dSet(V_i, V_j)), V_j \in \overline{Adj}(G^{pd}, V_i) \quad (4)$$

基于图结构依赖得分的定义, 本节中通过最小结构依赖逼近准则 (命题 1) 检验在每次迭代过程中的候选邻域结构是否存在更接近于真实的潜在因果结构, 并将最优的候选领域用于当前结构的下一步更新.

命题 1. 最小结构依赖逼近准则. 对于样本量大小为 m 的因果结构图 G^T 及其观测数据 D , 令 G^{pd} 为图 G^T 删除部分因果边所得到的部分因果结构图, G'^{pd} 表示在 G^{pd} 上添加了 $V_i \rightarrow V_j$ 的有向边而构成的新的因果结构图, 其中 $V_{tail} = V_j$, 表示该边的尾节点. 如果 $V_i \rightarrow V_j$ 在图 G^T 中存在, 且 $m \rightarrow \infty$, 则有以下结论成立.

(1) 对图 G^{pd} 上 V_X, V_Y 之间的连通路程 P , 其中 $(V_i \rightarrow V_j) \in P$, 则 V_X, V_Y 在数据样本上 (条件) 相关性下降:

$$MI_{G'^{pd}}(D, V_X, V_Y | dSet(V_X, V_Y) \cup V_{tail}) \leq MI_{G^{pd}}(D, V_X, V_Y | dSet(V_X, V_Y)) \quad (5)$$

(2) 相比图 G^{pd} , G'^{pd} 的图结构依赖得分更趋近于真实结构 G^T 的图结构依赖得分:

$$MI_{G'^{pd}} \leq MI_{G^{pd}} \quad (6)$$

证明: 令 G^T 表示真实的因果结构图, 当样本量 $m \rightarrow \infty$, 对于 G^T 上任意一对非邻接变量 V_X, V_Y 的所有连通路程 P , 给定可以阻断 P 的 d-分离变量结合 $dSet(V_X, V_Y)$, 那么这对变量之间的依赖将趋于 0:

$$MI(D, V_X, V_Y | dSet(V_X, V_Y)) \rightarrow 0.$$

因此, 在 G^T 上, 所有非邻接变量对的依赖之和也趋于最小, 表示为:

$$\lim_{m \rightarrow \infty} MI_{G^T} = \sum_i^n MI_i(D, V_X, V_Y | dSet(V_X, V_Y)) \rightarrow 0.$$

现考虑一种错误的情况, 在当前结构图 G 上错误地添加了一条有向边的 $V_i \rightarrow V_j, V_{tail} = V_j$, 形成新的候选图 G' . 在 G^T 中实际不存在通过 $V_i \rightarrow V_j$ 的连通路程 P 连接变量 V_X, V_Y . 因此, 在候选图 G' 上通过这条新增边连通的非邻

接变量 V_X, V_Y , 这条路径 $(V_i \rightarrow V_j) \notin P$; 因为新增的路径不是真实路径的一部分, 所以 V_X, V_Y 的依赖相对于图 G 而言不会降低.

$$MI_{G'_i}(D, V_X, V_Y | dSet(V_X, V_Y) \cup V_{\text{tail}}) > MI_{G_i}(D, V_X, V_Y | dSet(V_X, V_Y)).$$

反之, 在当前结构图 G 上正确的添加了一条有向边 $V_i \rightarrow V_j, V_{\text{tail}} = V_j$, 形成新的候选图 G' , 并且在 G^T 中存在通过 $V_i \rightarrow V_j$ 的连通路程 P 连接变量 V_X, V_Y , 则在候选图 G' 上非邻接变量对 V_X, V_Y 经过 $V_i \rightarrow V_j$ 连通, $V_i \rightarrow V_j \in P$; 如果通过 $V_i \rightarrow V_j$ 连通的路程 P 不被当前分离集 $dSet(V_X, V_Y)$ 阻断, 那么添加 V_j 到分离集中可以阻断路径 P , 即 $dSet(V_X, V_Y) \cup V_{\text{tail}}$ 能有效地反映变量间的独立性, 使得 G' 更接近潜在真实因果结构 G , 因此:

$$MI_{G'_i}(D, V_X, V_Y | dSet(V_X, V_Y) \cup V_{\text{tail}}) < MI_{G_i}(D, V_X, V_Y | dSet(V_X, V_Y)).$$

并且随着分离集的准确搜索, 节点之间的依赖将不断减小. 相对于图 G , 图 G' 上将得到更小的依赖之和, 更接近潜在的真实结构 G^T . 使得:

$$\lim_{m \rightarrow \infty} MI_{G'} < \lim_{m \rightarrow \infty} MI_G.$$

在当前结构图 G 上正确的添加了一条有向边 $V_i \rightarrow V_j$, 形成新的候选图 G' , 当在 G^T 中存在 $V_i \rightarrow V_j$ 的有向边, 但不存在通过 $V_i \rightarrow V_j$ 的连通路程 P 连接变量 V_X, V_Y , 则该边缘将不属于任何的非邻接变量对的连通路程 P , 图结构的非邻接结合的图结构依赖得分将保持不变:

$$\lim_{m \rightarrow \infty} MI_{G'} = \lim_{m \rightarrow \infty} MI_G.$$

从而得到最终公式, 即添加正确边缘的候选图 G' 相对于图 G 将更逼近潜在真实结构:

$$\lim_{m \rightarrow \infty} MI_{G'} \leq \lim_{m \rightarrow \infty} MI_G.$$

综上, 命题 1 得证.

在命题 1 的理论保证下, 算法基于最小结构依赖逼近准则可以学习到渐进正确的因果结构. 具体来说, 对候选结构的依赖得分计算方法如算法 2 所示.

算法 2. 计算候选图结构依赖得分.

输入: 数据 D , 图 G'^{pd} , 节点对 (V_i, V_j) , $\{\overline{Adj}(G'^{pd}, V_X), V_X \in V\}$, $dSet$ 分离集;

输出: 图 G'^{pd} 的 MI , 更新分离集 $dSet$.

1. 初始化 $MI = 0$
 2. 确定 G'^{pd} 中 (V_i, V_j) 的方向, 将尾节点保存为 V_{tail}
 3. **For** $V_X \in V$
 4. **For** $V_Y \in \overline{Adj}(G'^{pd}, V_X)$
 5. **IF** (V_X, V_Y) 通过 (V_i, V_j) 连通
 6. $dSet(V_X, V_Y) = dSet(V_X, V_Y) \cup V_{\text{tail}}$
 7. **End if**
 8. **End for**
 9. **End for**
 10. 计算图结构依赖得分 $MI_{G'^{pd}} = \sum_i^n MI_i(D, V_i, V_j | dSet(V_i, V_j)), V_j \in \overline{Adj}(G'^{pd}, V_i)$
 11. **Return** $MI_{G'^{pd}}, dSet$
-

算法 2 中描述了图结构更新的计算规则. 算法首先确定 G' 更新 (V_i, V_j) 边的方向, 并保存尾节点为 V_{tail} (当 $V_i \rightarrow V_j, V_{\text{tail}} = V_j$) (第 2 行). 对于所有在当前结构上非邻接的节点对, 如果有被 (V_i, V_j) 连通的路径, 则将尾节点添加到他们的 $dSet(V_i, V_j)$ 集合 (命题 1). 并使用更新的分离集计算依赖, 返回结果后续使用.

通过应用增强条件独立性检验方法恢复可能遗失的因果边, 在最小结构依赖准则下, 可以在多个候选邻域结

构中挑选出最优结构, 并进行结构更新 (算法 1 第 20 行). 完成了启发式算法中单次迭代过程的实现. 通过以上方法, 算法有效地解决了当真实因果结构存在高方差父节点使得 CIT 结果准确性下降, 进而导致因果边被误删的问题, 实现了对这些误删因果边的精确恢复.

4 实验分析

本节将进一步对 MMDCL 算法通过仿真数据、贝叶斯网络数据和真实数据进行实验评估. 本节将在 4.1 节详细介绍实验的对比方法, 所用数据集以及评估指标. 在第 4.2 节、第 4.3 节和第 4.4 节针对 MMDCL 算法与其他 5 种代表性的基于约束的因果结构学习算法分别在仿真数据、贝叶斯网络数据和真实数据这 3 个数据集的实验情况进行分析和评估.

4.1 实验设置

4.1.1 对比方法

在对比方法中, 本文考虑了经典的全局结构学习算法以及从局部到全局的因果结构学习算法共 5 种结构学习算法作为本文的对比方法. 经典的全局结构学习算法有 PC^[20] 以及其改进版本 PC_Stable^[46] 和 PC_Maxp^[46] 算法. PC_Stable 在 PC 的基础上引入了稳健的结构学习, 以避免由于特征排序导致的问题. PC_Maxp 算法提出了在 PC 的基础上对 V 结构冲突采用最大得分定向的概念. 为避免高阶 CIT 问题, PC 系算法统一将条件集大小限制为 3 个节点以内. 从局部到全局的因果结构学习算法包括 GSBN^[23] 和 ADL^[43] 算法, 其中 GSBN 通过学习每个变量的 MB (Markov blanket), 在对称校验方法中通过统一的接收不对称结果来构建骨架, 最后通过条件独立性检验来确定边的方向. 而 ADL 方法则在对称校验中提出了自适应的骨架拼接策略, 并通过 MMHC 进行因果边定向和扩展的因果边搜索过程.

4.1.2 数据介绍

仿真数据: 仿真因果结构数据的生成机制服从线性无环因果结构模型:

$$X_i = \sum_{X_j \in Pa(G^T, X_i)} b_{ij} X_j + \varepsilon_{X_i},$$

其中, $\varepsilon_{X_i} \sim N(0, a)$, $a \in [0, 500]$ 表示随机生成的噪声, 用于模拟真实因果结构中存在高方差节点的情况, $b_{ij} \sim U(-1, 0) \cup U(0, 1)$ 表示从父变量 X_j 指向子变量 X_i 的因果边强度. 在实验中, 对 CIT 检验中的回归消除部分采用了多元线性回归方法. 实验的其他参数见表 2, 其中粗体表示对照实验中的默认设置.

表 2 仿真数据参数设置

参数	范围
Dimension (节点维度)	{5, 10, 15, 20, 25 , 30, 35, 40}
Sample size (样本量)	{500, 1000 , 1500, 2000}
Average in-degree (平均入度)	{1, 1.5 , 2, 2.5}

贝叶斯网络数据: 贝叶斯网络数据使用的是由 Scutari 等人^[47] 构建的“multiple quantitative trait analysis using Bayesian networks”数据集, 该系列数据集涵盖了 multiparent advanced generation inter-cross (MAGIC) 冬小麦种群的信息, 综合了多个定量性状, 每个个体都进行了基因分型, 该数据集收录于 <https://www.bnlearn.com/bnrepository/>, 其结构来源于真实世界并提供了数据生成的方法, 本文将该数据集分为 3 组数据样本量, 分别包含 500、1000 和 2000 个数据实例. 具体信息如表 3 所示.

表 3 贝叶斯网络数据集结构信息

数据集	观察节点	平均入度	最大入度	边数
MAGIC-NIAB	44	3	9	66
MAGIC-IRRI	64	3.19	9.97	230

真实数据: 真实数据为 Downs 等人^[48]在 1992 年首次提出并在化学工程领域广泛应用的 Tennessee Eastman (TE) 数据集. 本文使用的是 2022 年由 Menegozzo 修订并公开的版本^[49]. 该数据集模拟了化工过程的正常稳态运行情况, 并应用中值操作符在 3 min 的滑动窗口上进行子采样, 覆盖了总共 75 h 的过程. 这个修订版本选取了共 33 个变量进行分析, 并分为 22 个连续测量的变量 (X1–X22) 和 11 个操纵变量 (X23–X33), 其中在它的预处理阶段, 提前删除了 X27 和 X31 这两个方差为零的变量, 并给出了基于这些变量的明确因果结构, 如图 6^[49]. 数据来源于 <https://github.com/giovanniMen>.

本文所有实验均在配备 Intel(R) i7-1070、3.2 GHz CPU 和 16 GB 内存的计算机上进行. 独立性检验的显著性水平设置为 0.01, 独立性检验使用 Fisher_Z 检验, 其余参数采用默认设置, 每组实验运行 100 次以上.

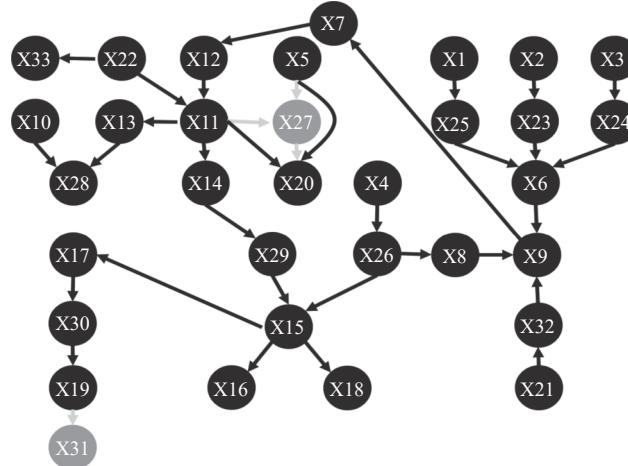


图 6 TE 的真实因果结构

4.1.3 评估方法

本文利用准确率 (precision, *PRE*)、召回率 (recall, *REC*)、*F1* 值和结构汉明距离 (structural Hamming distance, *SHD*) 这 4 项指标来评估提出的算法性能. 准确率计算正确预测的边与所有预测边的比例, 反映错误因果边的频率. 召回率衡量正确预测的边相对于网络中实际边的比例, 指出遗漏边的情况. *F1* 值综合考虑准确率和召回率, 全面评价算法性能. *SHD* 量化算法预测与真实结构之间的差异, 包括缺失的边、多余的边和方向错误的边, *SHD* 越低, 预测的结构越接近真实. 具体公式如下:

$$PRE = \frac{TP}{TP+FP}, REC = \frac{TP}{TP+FN}, F1 = \frac{2 \times PRE \times REC}{PRE+REC},$$

其中, *TP* 表示在学到的邻接矩阵信息中预测正确的定向边的数量; *FP* 代表将不存在或正/反向预测为存在或反/正向数量; *FN* 代表将正/反边预测为不存在或反/正向的数量.

4.2 仿真数据实验结果

本部分将分析在仿真数据随着网络节点维度、样本量大小和平均入度的变化下, MMDCL 算法与其他 5 个对比方法的指标变化. 其中 *PRE*, *REC* 和 *F1* 的实验结果见图 7, *SHD* 的结果见表 3.

1) 随着节点维度的上升, MMDCL 算法稳定保持在 *PRE* 高于 83%, *REC* 高于 65%, *F1* 高于 75% (图 7(a1), (a2), (a3)), 相比之下, PC_Stable, PC_Maxp 和 GSBN 算法在节点维度超过 15 后 *PRE* 结果略高于本文算法 5%, 但它们的 *REC* 则相对降低了 15% 以上, 说明这些方法得到的结果中有较多因果边的丢失. 而 PC 算法和 ADL 算法在 3 个评价指标上都略低于本文算法. 其中, 虽然 ADL 算法在网络节点数量为 5 时的 *F1* 值较高, 但随着节点维度上升, 其方法由于爬山法的局限性, 性能也有所下降.

2) 随着样本量大小的上升, MMDCL 算法的性能也不断提升, 与其他方法相比, 在所有样本量大小的条件下,

MMDCL 算法的 REC 和 $F1$ 与对比方法相比, 提升了 5%–35% (图 7(b1), (b2), (b3)).

3) 随着节点平均入度的提高, 网络结构会越来越复杂, 但 MMDCL 算法的 REC 和 $F1$ 值表现依旧稳定 (图 7(c1), (c2), (c3)), 而且在平均入度为 1.5 时, MMDCL 算法的结果最优, 这是因为由于网络中存在更多的非屏蔽对撞机 (unshielded collider)^[33], 使得更多的边缘可以定向. 当平均入度上升为 2.5 时, 所有算法的性能均有所下降, 这是由于网络中符合定向规则的边减小, 但此时 MMDCL 算法相对对比方法 $F1$ 值高于 20% 以上. 而 ADL 算法随着网络复杂化在有限迭代次数内也难以快速找到最优解, 而导致其识别结果受限.

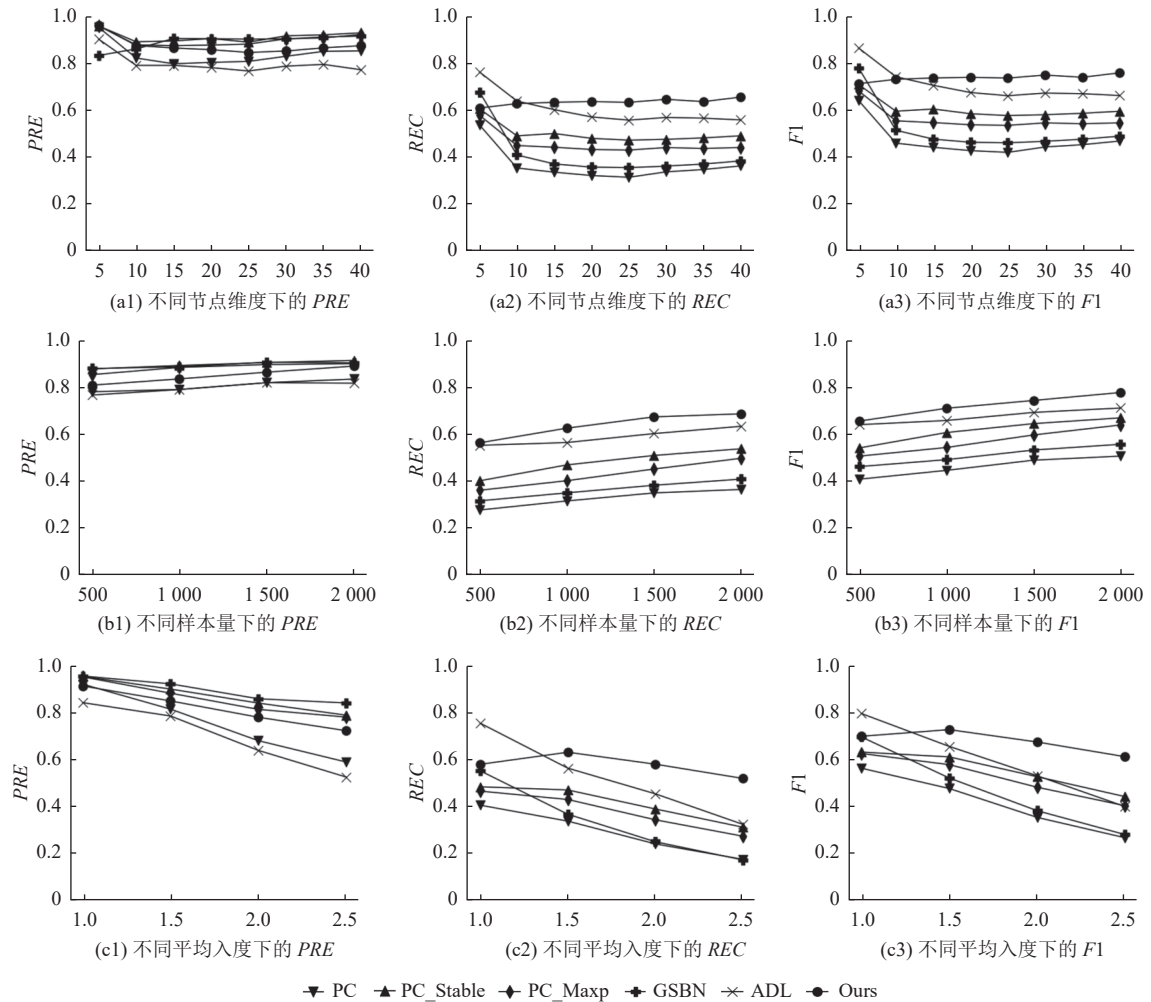


图 7 仿真数据在不同参数下的指标结果

总体而言, MMDCL 算法在各种参数设置下都表现出显著的性能提升. 尽管在 PRE 上略低于 3 种对比方法约 5%, 但其 REC 显著提升超 20% 且 $F1$ 值在各参数设置上都有显著提升, 这表明算法可能增加了一些错误预测, 但它能够更有效地纠正误判, 找回更多丢失的因果边.

本文算法在不同参数设置上 SHD 学习指标的结果如表 4 所示. 首先, 随着节点维度的增加, 本文算法相对于对比方法而言, 依旧保持相对较低的 SHD 值的优势. 在 $Dimension=35$ 的情况下, 与对比方法相比, 本文算法 SHD 可以提升最多 15 个因果边的正确率. 其次, 随着样本量的增加, 本文算法的 SHD 值也不断降低, 这说明样本量的提升可以使得算法输出更优的结果. 而在 $Sample\ size=500$ 的情况下, 本文算法在各维度上的结果表现依旧最优, 这

也说明了算法在数据有限场景下的稳定性. 而在 Sample size=2000 时, SHD 相对于所有对比方法而言, 可以提高 5–15 个因果边的正确率. 由此可见, 在不同维度和样本量条件下的 SHD 的结果表明了本文算法具备更好的结构学习性能.

表 4 仿真数据在节点维度和样本量变化下的 SHD

Dimension	Method	Sample size=500	Sample size=1000	Sample size=2000
15	PC	20.27±4.48	18.60±4.98	16.99±4.88
	PC_Stable	16.31±4.58	13.96±4.63	12.43±4.40
	PC_Maxp	17.94±5.17	14.86±4.96	13.89±5.02
	GSBN	15.55±3.66	14.40±3.85	13.18±5.18
	ADL	13.79±3.95	12.21±3.57	11.87±3.81
	Ours	13.88±4.89	11.82±5.16	10.08±4.52
25	PC	34.26±6.67	32.91±6.61	29.83±6.81
	PC_Stable	27.35±6.92	24.03±7.46	20.88±7.40
	PC_Maxp	29.61±7.56	27.40±6.55	23.04±8.06
	GSBN	26.18±5.63	24.05±8.56	22.61±8.38
	ADL	23.89±5.56	22.20±5.90	21.08±5.08
	Ours	23.11±7.65	20.89±8.57	17.06±8.06
35	PC	44.75±7.43	40.75±8.07	37.90±8.97
	PC_Stable	36.09±7.28	31.36±7.93	28.38±8.13
	PC_Maxp	39.31±8.25	34.47±8.59	30.50±9.02
	GSBN	35.43±10.65	33.24±11.62	30.89±14.24
	ADL	31.45±5.67	29.17±6.82	27.99±6.95
	Ours	31.36±8.31	26.72±8.61	24.00±9.50

4.3 贝叶斯网络数据实验结果

本节将详细分析在贝叶斯网络数据下随着 3 个不同的样本量 (500、1000、2000) 设置下, MMDCL 算法的性能表现. 如图 8 所示, *PRE*、*REC* 和 *F1* 的结果表明, MMDCL 在 3 个数据样本上均表现出优于对比方法的结果. 具体来说, 在 MAGIC-NIAB 数据集下, MMDCL 的 *PRE*、*REC*、*F1* 的结果相比于对比方法提升 5%–45%. 而对于 MAGIC-IRRI 数据集, MMDCL 的 *PRE*、*REC*、*F1* 结果均对比方法提升 10%–25%. 其中, 随着样本量增加, 算法性能也有所提升. 在样本量 2000 时, 算法在保持高精确率的同时, *REC* 大于 75%, *F1* 大于 70%, 这与仿真数据上得到的结论一致. 即使在样本量仅为 500 时, 算法依旧能优于对比方法 5%–30%, 其中在 MAGIC-NIAB 上, 本文算法在 *REC* 的结果上相对对比方法最大可提升接近 50%, 这可能是因为当前数据具有更大的方差差异导致可以恢复的因果边更多, 也进一步验证了本算法在样本受限条件下的鲁棒性. 通过实验结果有效证实了本文算法在处理贝叶斯网络数据时的有效性.

表 5 中 SHD 结果进一步支持 MMDCL 实验结果的可靠性. 在所有样本量条件下, MMDCL 算法在 SHD 指标上均展现出显著的优势. 在 MAGIC-NIAB 数据集中, 样本量为 500 时, MMDCL 相较于其他方法减少了 8–20 个错误的因果边, 体现了在小样本下结构学习的可靠性. 随着样本量的增加, MMDCL 算法的 SHD 降低到 20 以内, 这也说明随着样本量上升, 本算法具备更高的稳定性和准确性. 在 MAGIC-IRRI 数据集上, MMDCL 算法降低了 10–20 个错误因果边, 随着样本量增加, SHD 值降低到 40, 相对于对比方法, 本算法有明显提升. 这些实验结果表明了本算法在小样本量条件下具备优秀的性能表现, 并且随着样本量的增加, 其学习效率得到进一步提升, 这说明了本算法在处理贝叶斯网络数据的准确性.

4.4 真实数据实验结果

在真实数据集 TE 下, 使用 PC_Stable 算法学习得到的结构与通过 MMDCL 算法执行的因果边恢复结果如图 9 所示. 图 9 中的节点代表观测变量, 除去 X27 和 X31 之后, 共有 31 个节点参与分析. 其中, 浅灰色的边代表

PC_Stable 算法错误地学习到的因果边, 虚线边表示正确学习到的因果边, 包括了无向边和有向边. 实线边表示 MMDCL 算法成功恢复的正确因果边, 而双横线边则代表 MMDCL 算法恢复的错误因果边. 图 9 中, 正确因果边的两端节点用黑色方块突出表示. 例如, 对于因果边 $X_{32} \rightarrow X_9$, 考虑到 X_{32} 有先验的父节点 X_{21} 和 X_6 , 而 X_9 有先验的父节点 X_6 , 在进一步测量 $CIT(X_{32} - f(X_{21}), X_9 | X_6)$ 的情况下, 这条因果边便能被成功恢复. 同理, 在考虑 $CIT(X_8, X_9 - f(X_{32}) - f(X_6))$ 的情况下, 又可以恢复出 $X_8 \rightarrow X_9$ 的因果. 对于错误学习到的因果边, 如 $X_{14} \rightarrow X_{16}$, 其中 X_{14} 是 X_{16} 的潜在祖先, 它们在真实因果结构中通过路径 $X_{14} \rightarrow X_{29} \rightarrow X_{15} \rightarrow X_{16}$ 连通, 由于这条连通路程的中间节点 X_{29} 和 X_{15} 没有被识别到, 导致 $X_{14} - X_{16}$ 因果边未被成功 d-分离, 最后被错误判断为直接相连的因果边而添加. 类似的, $X_7 \rightarrow X_{20}$, $X_{28} \rightarrow X_{20}$ 也是具有潜在间接因果关系, 但是由于算法在爬边过程中, X_{11} 作为关键的连通路程中的 d-分离节点还未被找回导致这两条边被错误的恢复了, 这也是启发式算法通常所面临的局部最优解问题. 最后, $X_{13} \rightarrow X_{11}$ 则是由于 V 结构定向规则导致的误判问题. 与原始算法相比, 本研究的算法成功恢复了 6 条正确的因果边, 同时误学习了 3 条错误的因果边, 以及一条方向错误的因果边, 在真实数据实验中观察到, 尽管存在部分错误的因果先验节点, 但这些节点并没有直接导致因果识别结果的错误. 相反, 更多的错误是由于未能找到连通路程中关键节点集造成的. 这说明在真实数据上, 本算法对部分错误先验信息具有一定的抗干扰能力, 进一步证实了 MMDCL 算法能够有效地找回潜在因果结构中的因果关系.

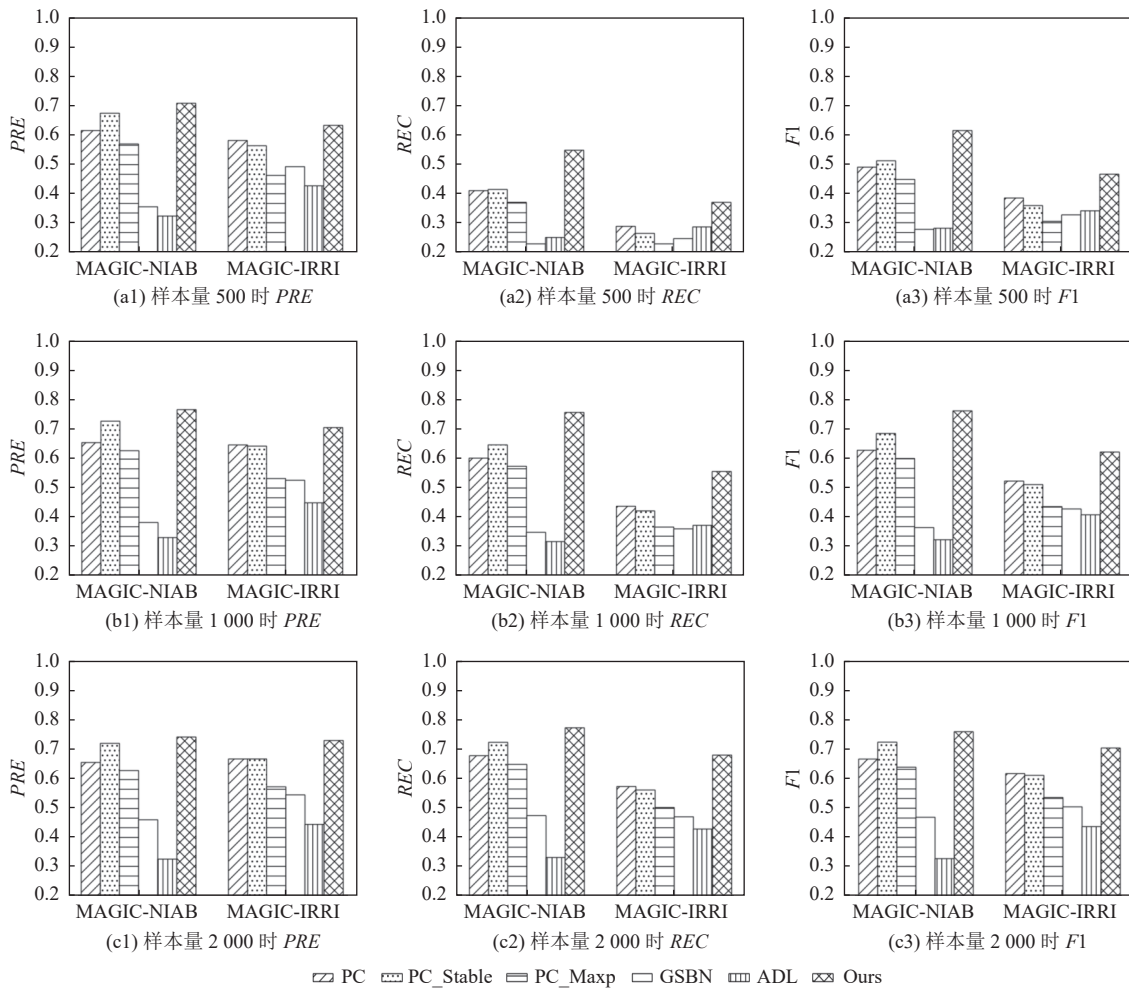


图 8 贝叶斯网络数据在不同样本量下的实验结果

表 5 贝叶斯网络数据在不同样本量下的 SHD

Dimension	Algorithm	Sample size=500	Sample size=1000	Sample size=2000
MAGIC-NIAB	PC	42.55±3.99	29.27±3.50	24.43±2.22
	PC_Stable	42.04±4.30	26.24±3.49	21.43±2.74
	PC_Maxp	44.90±4.11	31.01±3.57	26.30±2.83
	GSBN	54.01±3.71	45.72±3.62	37.33±3.61
	ADL	55.51±3.95	49.73±3.57	48.78±3.81
	Ours	34.71±4.77	20.26±3.63	19.01±3.23
MAGIC-IRRI	PC	78.72±4.06	63.15±4.61	49.15±3.97
	PC_Stable	80.67±4.41	64.5±4.97	50.06±3.86
	PC_Maxp	84.24±3.72	70.09±4.97	56.07±4.45
	GSBN	81.83±3.65	70.11±3.57	58.17±3.56
	ADL	86.61±4.67	76.69±4.48	69.55±4.89
	Ours	73.14±5.57	54.23±4.96	40.88±4.27

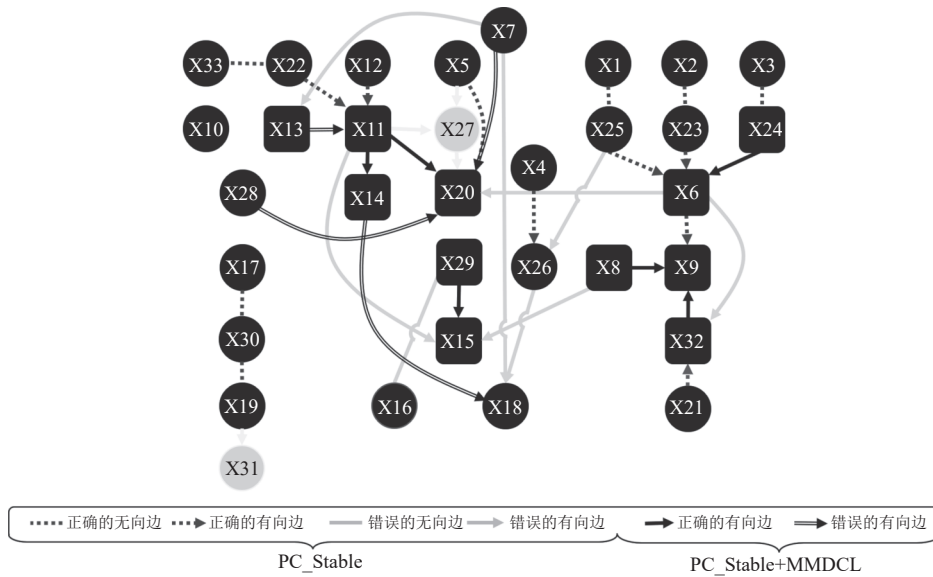


图 9 真实数据下 MMDCL 的结构学习结果

5 讨论

MMDCL 算法旨在通过部分先验结构信息降低高方差节点对 CIT 结果的影响, 从而增强条件独立性检验的准确性, 保证结构学习算法的有效性. 该算法框架建立在线性无环因果结构模型的基础上, 并假设变量间的因果关系可以被建模为线性关系. 尽管线性因果模型在基因表达^[7]、生物化学研究^[8]、医疗病理分析^[9]、社会学分析^[10,11]等能够被构建为贝叶斯网络模型的现实数据中得到广泛的应用, 但本文方法的理论结果并不能直接推广到非线性、高复杂结构的数据中. 其本质的原因在于, 本算法通过线性回归进行增强的条件独立性检验, 并基于偏相关检验来量化变量间的条件独立性. 然而偏相关检验并不适用于测试非线性模型下的条件独立性. 一种可能的解决方案是通过分析条件互信息在消除高方差噪声干扰下的统计特性, 来拓展本文的理论结果到非线性模型上, 这将是未来需要进一步研究的工作.

此外, 由于本算法在消除高方差节点的影响时需要最大可能的利用已知先验信息, 当先验节点不准的情况下 (例如错误删除了父亲节点), 很可能导致增强独立性检验判断出一些错误的依赖关系, 使得领域搜索阶段得到部分的错误因果候选边. 为了避免恢复错误的因果边, 算法在进行结构更新时, 通过最小结构依赖测量每次选择候选

结构中依赖得分最小的结构, 进而避免并减少部分错误的因果边被恢复的情况, 通过这一步骤可以有效避免错误先验导致的迭代错误问题. 其中, 最小结构依赖测量是一种结构得分的方法, 但相对于经典的得分算法 GES 而言, 本算法在进行结构评分之前, 首先通过增强独立性检验对节点间的因果相关关系进行判断, 筛选掉大部分的无关候选边, 避免了 GES 算法需要遍历所有可能的候选边进行最优结构计算的问题. 其次, 本文算法的目标是解决条件独立性检验在受到高方差节点影响下结构不准确的问题, 算法本质上是对条件独立性检验过程的优化, 最小结构依赖测量作为辅助验证的方法保证更新结构的准确性, 相对于 GES 高度依赖得分方法的结果, 本文算法通过增强独立性检验和最小结构依赖测量的双重检验步骤有效的中和了假阳性和假阴性结果的输出.

6 总 结

本研究针对基于约束的因果发现算法在样本受限和真实因果结构存在高方差节点的条件下降问题的结构学习性能下降问题, 通过对经典条件独立性检验方法 (Fisher_Z 检验) 进行实验分析和理论推导, 揭示了高方差父节点引入的噪声是影响子节点与其低方差父节点的 CIT 结果不可靠, 进而导致结构学习性能下降的主要原因. 为此, 本研究提出了利用部分已知结构信息来消除已知的高方差父节点引入的噪声, 以提升 CIT 结果的可靠性. 紧接着, 基于上述增强的条件独立性检验方法, 本文提出了 MMDCL 算法, 并在数据仿真实验、贝叶斯网络数据和真实数据中, 验证了该算法能有效地恢复因果图中因误判而丢失的因果边. MMDCL 算法为在样本受限和真实因果结构存在高方差节点的条件下的因果关系研究提供了一种更精确、更可靠的方法. 此外, 基于先验信息本研究提出的算法框架还能适用于更多的场景, 未来的研究可以进一步探索先验信息的准确性, 提升因果结构学习策略的准确性与适用性.

References:

- [1] Spirtes P, Zhang K. Causal discovery and inference: Concepts and recent methodological advances. *Applied Informatics*, 2016, 3: 3. [doi: 10.1186/s40535-016-0018-x]
- [2] Pearl J, Mackenzie D. *The Book of Why: The New Science of Cause and Effect*. New York: Basic Books Inc., 2018.
- [3] Cai RC, Zhang ZJ, Hao ZF, Winslett M. Understanding social causalities behind human action sequences. *IEEE Trans. on Neural Networks and Learning Systems*, 2017, 28(8): 1801–1813. [doi: 10.1109/TNNLS.2016.2556724]
- [4] Runge J, Nowack P, Kretschmer M, Flaxman S, Sejdinovic D. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, 2019, 5(11): eaau4996. [doi: 10.1126/sciadv.aau4996]
- [5] Cai RC, Zhang ZJ, Hao ZF. Causal gene identification using combinatorial V-structure search. *Neural Networks*, 2013, 43: 63–71. [doi: 10.1016/j.neunet.2013.01.025]
- [6] Cai RC, Zhang ZJ, Hao ZF. BASSUM: A Bayesian semi-supervised method for classification feature selection. *Pattern Recognition*, 2011, 44(4): 811–820. [doi: 10.1016/j.patcog.2010.10.023]
- [7] De La Fuente A, Bing N, Hoeschele I, Mendes P. Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics*, 2004, 20(18): 3565–3574. [doi: 10.1093/bioinformatics/bth445]
- [8] Yang J, An N, Alterovitz G. A partial correlation statistic structure learning algorithm under linear structural equation models. *IEEE Trans. on Knowledge and Data Engineering*, 2016, 28(10): 2552–2565. [doi: 10.1109/TKDE.2016.2578315]
- [9] Shen XP, Ma SS, Vemuri P, Simon G, Alzheimer's Disease Neuroimaging Initiative. Challenges and opportunities with causal discovery algorithms: Application to Alzheimer's pathophysiology. *Scientific Reports*, 2020, 10(1): 2975. [doi: 10.1038/s41598-020-59669-x]
- [10] Girju R. Toward social causality: An analysis of interpersonal relationships in online blogs and forums. In: *Proc. of the 4th Int'l AAAI Conf. on Web and Social Media*. Washington: AAAI, 2010. 66–73. [doi: 10.1609/icwsm.v4i1.14030]
- [11] Yang J, Li N, An N, Chen Y, Alterovitz G. An efficient causal structure learning algorithm for linear arbitrarily distributed continuous data. *The Journal of Supercomputing*, 2020, 76(5): 3355–3363. [doi: 10.1007/s11227-018-2557-5]
- [12] Zhao SD, Liu T. Causality and its applications in social media: A survey. *Ruan Jian Xue Bao/Journal of Software*, 2014, 25(12): 2733–2752 (in Chinese with English abstract). <http://www.jos.org.cn/1000-9825/4724.htm> [doi: 10.13328/j.cnki.jos.004724]
- [13] Wang ZX, Chan LW. An efficient causal discovery algorithm for linear models. In: *Proc. of the 16th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*. Washington: ACM, 2010. 1109–1118. [doi: 10.1145/1835804.1835944]
- [14] Belthangady C, Giampanis S, Jankovic I, Stedden W, Alves P, Chong S, Knott C, Norgeot B. Causal deep learning reveals the comparative effectiveness of antihyperglycemic treatments in poorly controlled diabetes. *Nature Communications*, 2022, 13(1): 6921.

- [doi: [10.1038/s41467-022-33732-9](https://doi.org/10.1038/s41467-022-33732-9)]
- [15] Schölkopf B, Locatello F, Bauer S, Ke NR, Kalchbrenner N, Goyal A, Bengio Y. Toward causal representation learning. *Proc. of the IEEE*, 2021, 109(5): 612–634. [doi: [10.1109/JPROC.2021.3058954](https://doi.org/10.1109/JPROC.2021.3058954)]
- [16] Chai BF, Jia CY, Yu J. Approaches of structure exploratory based on probabilistic models in massive networks. *Ruan Jian Xue Bao/Journal of Software*, 2014, 25(12): 2753–2766 (in Chinese with English abstract). <http://www.jos.org.cn/1000-9825/4722.htm> [doi: [10.13328/j.cnki.jos.004722](https://doi.org/10.13328/j.cnki.jos.004722)]
- [17] Chen WQ, Hao ZF, Cai RC, Zhang XZ, Hu Y, Liu M. Multiple-cause discovery combined with structure learning for high-dimensional discrete data and application to stock prediction. *Soft Computing*, 2016, 20(11): 4575–4588. [doi: [10.1007/s00500-015-1764-8](https://doi.org/10.1007/s00500-015-1764-8)]
- [18] Zhang XZ, Hu Y, Xie K, Wang SY, Ngai EWT, Liu M. A causal feature selection algorithm for stock prediction modeling. *Neurocomputing*, 2014, 142: 48–59. [doi: [10.1016/j.neucom.2014.01.057](https://doi.org/10.1016/j.neucom.2014.01.057)]
- [19] Li YZ, Torralba A, Anandkumar A, Fox D, Garg A. Causal discovery in physical systems from videos. In: *Proc. of the 34th Conf. on Neural Information Processing Systems*. Vancouver: NeurIPS, 2020. 9180–9192.
- [20] Spirtes P, Glymour CN, Scheines R. *Causation, Prediction, and Search*. 2nd ed., Cambridge: MIT Press, 2000.
- [21] Meek C. Causal inference and causal explanation with background knowledge. In: *Proc. of the 11th Conf. on Uncertainty in Artificial Intelligence*. Montréal: Morgan Kaufmann Publishers Inc., 1995. 403–410.
- [22] Dor D, Tarsi M. A simple algorithm to construct a consistent extension of a partially oriented graph. Technical Report, R-185, Los Angeles: Cognitive Systems Laboratory, 1992. 45.
- [23] Margaritis D, Thrun S. Bayesian network induction via local neighborhoods. In: *Proc. of the 12th Int'l Conf. on Neural Information Processing Systems*. Denver: MIT Press, 1999. 505–511.
- [24] Bromberg F, Margaritis D. Improving the reliability of causal discovery from small data sets using argumentation. *The Journal of Machine Learning Research*, 2009, 10: 301–340.
- [25] Aliferis CF, Statnikov A, Tsamardinos I, Mani S, Koutsoukos XD. Local causal and Markov blanket induction for causal discovery and feature selection for classification Part I: Algorithms and empirical evaluation. *The Journal of Machine Learning Research*, 2010, 11: 171–234.
- [26] Armen AP, Tsamardinos I. Estimation and control of the false discovery rate of Bayesian network skeleton identification. Technical Report TR-441, Crete: University of Crete, 2014.
- [27] Hyttinen A, Eberhardt F, Jarvisalo M. Constraint-based causal discovery: Conflict resolution with answer set programming. In: *Proc. of the 30th Conf. on Uncertainty in Artificial Intelligence*. Quebec: UAI Press, 2014. 340–349.
- [28] Gao T, Fadnis K, Campbell M. Local-to-global Bayesian network structure learning. In: *Proc. of the 34th Int'l Conf. on Machine Learning*. Sydney: JMLR.org, 2017. 1193–1202.
- [29] Kalisch M, Bühlmann P. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *The Journal of Machine Learning Research*, 2007, 8: 613–636.
- [30] Tsamardinos I, Brown LE, Aliferis CF. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 2006, 65(1): 31–78. [doi: [10.1007/s10994-006-6889-7](https://doi.org/10.1007/s10994-006-6889-7)]
- [31] Zhalama, Zhang JJ, Eberhardt F, Mayer W, Li MJ. ASP-based discovery of semi-Markovian causal models under weaker assumptions. In: *Proc. of the 28th Int'l Joint Conf. on Artificial Intelligence*. Macao: ijcai.org, 2019. 1488–1494. [doi: [10.24963/ijcai.2019/206](https://doi.org/10.24963/ijcai.2019/206)]
- [32] Niinimäki T, Parviainen P. Local structure discovery in Bayesian networks. In: *Proc. of the 28th Conf. on Uncertainty in Artificial Intelligence*. Catalina: AUAI Press, 2012. 634–643.
- [33] Pearl J. *Causality: Models, Reasoning and Inference*. New York: Cambridge University Press, 2009.
- [34] Murphy K, Schölkopf B, Colombo D, Maathuis MH. Order-independent constraint-based causal structure learning. *The Journal of Machine Learning Research*, 2014, 15(1): 3741–3782.
- [35] Pearl J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo: Morgan Kaufmann Publishers, 1988.
- [36] Chickering DM. Optimal structure identification with greedy search. *The Journal of Machine Learning Research*, 2003, 3: 507–554. [doi: [10.1162/15324430321897717](https://doi.org/10.1162/15324430321897717)]
- [37] Shimizu S, Hoyer PO, Hyvärinen A, Kerminen A. A linear non-Gaussian acyclic model for causal discovery. *The Journal of Machine Learning Research*, 2006, 7: 2003–2030
- [38] Cai RC, Qiao J, Zhang K, Zhang ZJ, Hao ZF. Causal discovery with cascade nonlinear additive noise models. In: *Proc. of the 28th Int'l Joint Conf. on Artificial Intelligence*. Macao: AAAI Press, 2019. 1609–1615.
- [39] Chai RC, Chen W, Zhang K, Hao ZF. A survey on non-temporal series observational data based causal discovery. *Chinese Journal of Computers*, 2017, 40(6): 1470–1490 (in Chinese with English abstract). [doi: [10.11897/SP.J.1016.2017.01470](https://doi.org/10.11897/SP.J.1016.2017.01470)]

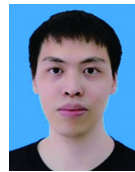
- [40] Hoyer PO, Janzing D, Mooij J, Peters J, Schölkopf B. Nonlinear causal discovery with additive noise models. In: Proc. of the 21st Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2008. 689–696.
- [41] Triantafyllou S, Tsamardinos I. Constraint-based causal discovery from multiple interventions over overlapping variable sets. The Journal of Machine Learning Research, 2015, 16(1): 2147–2205.
- [42] Ng I, Zheng YJ, Zhang JJ, Zhang K. Reliable causal discovery with improved exact search and weaker assumptions. In: Proc. of the 35th Int'l Conf. on Neural Information Processing System. Curran Associates Inc., 2021. 1553.
- [43] Guo XJ, Yu K, Liu L, Li PP, Li JY. Adaptive skeleton construction for accurate DAG learning. IEEE Trans. on Knowledge and Data Engineering, 2023, 35(10): 10526–10539. [doi: 10.1109/TKDE.2023.3265015]
- [44] Pearson K. III. Contributions to the mathematical theory of evolution. Philosophical Trans. of the Royal Society A, 1894, 185: 71–110. [doi: 10.1098/rsta.1894.0003]
- [45] Bartlett MS. The statistical significance of canonical correlations. Biometrika, 1941, 32(1): 29–37. [doi: 10.2307/2332247]
- [46] Zheng YJ, Huang BW, Chen W, Ramsey J, Gong MM, Cai RC, Shimizu S, Spirtes P, Zhang K. Causal-learn: Causal discovery in Python. Journal of Machine Learning Research, 2024, 25(60): 1–8.
- [47] Scutari M, Howell P, Balding DJ, Mackay I. Multiple quantitative trait analysis using bayesian networks. Genetics, 2014, 198(1): 129–137. [doi: 10.1534/genetics.114.165704]
- [48] Downs JJ, Vogel EF. A plant-wide industrial process control problem. Computers & Chemical Engineering, 1993, 17(3): 245–255. [doi: 10.1016/0098-1354(93)80018-I]
- [49] Menegozzo G, Dall'Alba D, Fiorini P. CIPCaD-Bench: Continuous industrial process datasets for benchmarking causal discovery methods. In: Proc. of the 18th IEEE Int'l Conf. on Automation Science and Engineering. Mexico City: IEEE, 2022. 2124–2131. [doi: 10.1109/CASE49997.2022.9926420]

附中文参考文献:

- [12] 赵森栋, 刘挺. 因果关系及其在社会媒体上的应用研究综述. 软件学报, 2014, 25(12): 2733–2752. <http://www.jos.org.cn/1000-9825/4724.htm> [doi: 10.13328/j.cnki.jos.004724]
- [16] 柴变芳, 贾彩燕, 于剑. 基于概率模型的大规模网络结构发现方法. 软件学报, 2014, 25(12): 2753–2766. <http://www.jos.org.cn/1000-9825/4722.htm> [doi: 10.13328/j.cnki.jos.004722]
- [39] 蔡瑞初, 陈薇, 张坤, 郝志峰. 基于非时序观察数据的因果关系发现综述. 计算机学报, 2017, 40(6): 1470–1490 [doi: 10.11897/SP.J.1016.2017.01470]



郝志峰(1968—), 男, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为算法设计与分析, 数学建模, 数据挖掘.



乔杰(1993—), 男, 博士, 主要研究领域为因果推断, 人工智能.



汪菲霞(1999—), 女, 硕士生, 主要研究领域为因果关系发现.



蔡瑞初(1983—), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为因果推断, 深度学习.



陈正鸣(1996—), 男, 博士生, 主要研究领域为因果发现, 潜在因果模型及其应用.