

基于递归分解的因果结构学习算法

蔡瑞初¹, 张文辉¹, 乔杰¹, 郝志峰^{1,2}

(1. 广东工业大学 计算机学院, 广州 510006; 2. 汕头大学 理学院, 广东 汕头 515063)

摘要: 在高维小样本场景下, 针对现有基于约束的因果结构学习方法存在因果结构学习效率低、马尔可夫等价类的问题, 以非线性非高斯的高维小样本为研究对象, 提出一种基于递归分解的因果结构学习算法 CADR。在高维小样本的因果结构学习效率方面, 结合递归分解的思想, 将高维变量集递归分解为多个更小的子集, 直到无法再分解或子集的大小达到阈值为止。在该过程中, 变量集的减少缩减了条件独立性检验的条件候选集的搜索空间, 从而提高学习效率。同时, 为进一步识别马尔可夫等价类, 根据非线性非高斯模型的因果方向的不可逆性, 通过判断拟合噪声项与原因变量是否独立来识别马尔可夫等价类的因果方向。在仿真数据和真实因果结构数据上的实验结果表明, CADR 不仅提高条件独立性检验的效率, 而且能有效地区分马尔可夫等价类, 学习到更精确的因果结构, 其中, 在真实因果结构实验中, 与现有 Xie_rec、PC_ANM 和 Notear_Sob 方法相比, F1 评分提高 5%~12%。

关键词: 因果关系发现; 条件独立性检验; 高维小样本; 递归分解; 马尔可夫等价类

开放科学(资源服务)标志码(OSID):



中文引用格式: 蔡瑞初, 张文辉, 乔杰, 等. 基于递归分解的因果结构学习算法[J]. 计算机工程, 2023, 49(3): 87-94.

英文引用格式: CAI R C, ZHANG W H, QIAO J, et al. Causal structure learning algorithm based on recursive decomposition[J]. Computer Engineering, 2023, 49(3): 87-94.

Causal Structure Learning Algorithm Based on Recursive Decomposition

CAI Ruichu¹, ZHANG Wenhui¹, QIAO Jie¹, HAO Zhifeng^{1,2}

(1. School of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, China;

2. College of Science, Shantou University, Shantou 515063, Guangdong, China)

[Abstract] In the case of high-dimensional small samples, owing to the problems of low efficiency and Markov equivalence class in the existing constraint-based causal structure learning methods, a causal structure learning algorithm, CADR, based on recursive decomposition, is proposed for nonlinear non-Gaussian high-dimensional small samples. When the learning efficiency of the causal structure of high-dimensional small samples is combined with the idea of recursive decomposition, the high-dimensional variable set is recursively decomposed into multiple smaller subsets exhaustively or until the subset size reaches a threshold. The reduced variable set reduces the search space of the conditional candidate set for the conditional independence test, thus improving the learning efficiency. Moreover, to identify the Markov equivalence class further, according to the irreversibility of the causal direction of the nonlinear non-Gaussian model, identify the causal direction of the Markov equivalence class by determining whether the fitting noise item is independent of the causal variable. The experimental results of simulation and real causal structure data indicate that CADR improves the efficiency of the conditional independence test and can effectively distinguish Markov equivalence classes and learn an accurate causal structure. In the real causal structure experiment, the F1 score increased by 5%-12% when the existing Xie_rec, PC_ANM, and Notear_Sob method are compared with the proposed method.

[Key words] causal relationship discovery; conditional independence test; high-dimensional small samples; recursive decomposition; Markov equivalence class

DOI: 10.19678/j.issn.1000-3428.0064165

0 概述

从观测数据中推断变量之间的因果关系是当今

数据科学研究的热点^[1-2]。因果关系发现的研究可以揭露数据的产生机制, 而数据产生机制对现实中的预测和干预问题有着极其关键的作用。在现有的

基金项目: 国家优秀青年科学基金(6212200101); 国家自然科学基金(61876043, 61976052)。

作者简介: 蔡瑞初(1983—), 男, 教授、博士生导师, 主研方向为人工智能、因果关系发现; 张文辉(通信作者), 硕士研究生; 乔杰, 博士研究生; 郝志峰, 教授、博士生导师。

收稿日期: 2022-03-14 修回日期: 2022-06-17 E-mail: 2111905015@mail2.gdut.edu.cn

研究中,因果关系发现方法应用于生物医学、教育学、经济学、流行病、气象学等^[3-4]多个领域。

传统因果关系方法的研究是通过随机控制实验进行因果关系学习,但实验所需费用高,或现有技术无法实现。因此,基于观察数据的因果结构学习方法备受关注。PEARL等^[5]从图模型的角度给出一套因果关系发现的语言,提出一种因果图模型,其中,节点表示变量,节点之间的有向边表示变量之间的因果关系。基于该模型,PEARL和SPIRITES等^[5-6]提出基于约束的因果结构学习方法。该方法在忠诚性等假设下^[7],通过条件独立性检验学习因果结构。虽然这类方法能够取得较准确的结果,但是存在马尔可夫等价类问题^[8],即部分因果关系无法被识别。马尔可夫等价类在概率图模型中指具有相同条件独立,其变量和因果骨架相同,但方向不确定的结构。假设由三个随机变量 x, y, z 组成的骨架 $x-z-y$ 存在相同的条件独立性 $P(x|z)P(y|z)=P(x, y|z)$,则可能存在3种不同的结构: $x \rightarrow z \rightarrow y$, $x \leftarrow z \leftarrow y$, $x \leftarrow z \rightarrow y$ 。

此外,基于约束的因果结构学习方法在对条件候选集搜索时存在复杂度高的问题^[9]。BERGSMA等^[10]提出在连续变量的条件独立性检验中,随着变量数量的增加,对条件候选集的搜索复杂度将呈指数级增长。在该过程中,随着变量数的增长,算法中变量条件独立性检验条件集的数量也呈指数增长,使得算法的计算复杂度急剧增大。当条件集 Z 的数量不断增大时,条件独立性检验的正确性有待验证,同时极有可能出现第二类错误,即使变量间互相依赖,也可能被预测为相互条件独立。当条件集数量增加,而样本量不足时,会增加噪声或其他变量对无关变量的影响权重,使得原本相互不独立的变量变为独立关系^[11]。

针对这类问题,研究人员提出采用递归分解的方法进行因果结构学习^[9, 12-15]。这类方法的主要目标是把一个问题递归地分解到两个或多个规模更小的子问题空间中进行高效求解。但是,该类方法仍无法解决马尔可夫等价类问题。

通过对实际场景数据的分析,本文发现在很多现实数据(如社交网络、基因调控网络等)中变量间的关系较稀疏,其组成的因果结构也呈现稀疏性^[16](即每个变量的平均父亲节点数量为2)。因此,结合实际场景数据的特征,本文提出一种基于递归分解的因果结构学习算法CADR,在高维小样本的数据中更高效地进行因果结构学习。基于观测样本的稀疏性,设计一种新的递归分解算法,减少条件候选集的数量,使得条件独立性检验更可靠。基于现有加性噪声模型(Additive Noise Model, ANM)的噪声和数据产生机制假设,解决条件独立性检验中存在的马尔可夫等价类问题。根据递归分解算法所得到的

变量子集和加性噪声模型,提出一种学习完全定向的有向无环图(Complete Directed Acyclic Graph, CDAG)方法。

1 相关工作

传统因果结构学习的方法(如IC^[5]、PC算法^[6, 17-18]等)通过对变量进行条件独立性检验,因存在马尔可夫等价类问题,只能学习因果结构的基本框架,无法学习到真正的因果关系。此外,大多数现有的因果结构学习方法都是在假设样本量足够的前提下进行研究。部分因果结构学习方法虽然对小样本的因果关系发现进行研究^[19],但是实际上在一些高维、小样本数据(如基因表达数据)上的实验效果表现不佳。

针对高维数据的因果关系发现问题,研究人员通过缩小条件候选集的搜索范围,并采用递归分解的方法进行因果关系发现。这类方法的主要目标是将一个总问题递归地分解为两个或多个规模更小且解法相同的子问题,从而提高求解效率。

在XIE等^[12]提出的框架中,对于随机变量集合 V :首先,通过条件独立性检验学习每对随机变量在给定其他所有节点集条件下的独立性,得到一个无向独立图(Undirected Independence Graph, UIG);其次,根据条件 $A \perp B|C$,其中 $A, B, C \subseteq V$,从UIG中找出一个分解 $V=\{A, B, C\}$,把集合分解为更小的两个子集 $V_1=A \cup C$ 和 $V_2=B \cup C$,再分别学习这两个子集的UIG,并将其分解为更小的子集,递归地执行该分解过程,直到子集无法再被分解;最后,通过基于约束的方法(如IC、PC等算法)尽可能多地确定剩余无向边的方向。之后,把这些局部因果结构合并成一个全局因果结构。这种方法摆脱了传统方法对样本量的依赖,但每次学习UIG的条件集是除了节点对以外的其他所有节点,可能会降低条件独立性检验的准确率。LIU等^[14]通过基于搜索评分的方法来学习UIG,并取得较好的效果,但是在高维样本中,通过搜索评分的方法学习UIG的计算复杂度不断增大。

条件独立性检验的方法是通过检测V-结构来判断局部结构的因果方向,但因无法区分马尔可夫等价类问题而学习到完全有向的因果结构。针对该问题,研究人员提出加性噪声模型的方法,通过对噪声和数据产生的机制进行假设,根据观测数据拟合噪声,并判断其与原因变量是否独立,从而确定因果方向。这类方法的关键是对噪声和数据产生机制进行假设。根据变量的数量将现有方法分为两类:1)两个变量因果方向判断方法,如非线性非高斯模型^[20]、基于回归的模型^[21]、基于核空间的方法^[22-23]等;2)三个及以上变量的因果方向判断方法,基于HSIC的方法^[24]、基于KCIT的方法^[25]、基于ReCIT的方法^[11]等。在高维小样本的情况下,因为这些方法需要大量计算变量与联合分布之间的条件概率统计,要求

样本量必须充足, 所以其实验效果表现不佳。

2 基于递归分解的因果结构学习

2.1 相关符号与概念

令 $G=(V, E)$ 表示由变量集 $V=\{v_1, v_2, \dots, v_n\}$ 组成的有向无环图 (Directed Acyclic Graph, DAG), $D=\{X_1, X_2, \dots, X_n\}$ 表示观测数据集, $pa(v_i)$ 表示节点变量 v_i 的父亲节点。两个不相邻节点 v_i 和 v_j 之间的路径 (path) l 表示该路径上的第一个节点变量为 v_i , 最后一个节点变量为 v_j , 同时在该路径上任意两个相邻节点都存在边相连。在这样的路径中, v_i 是 v_j 的一个祖先节点, v_j 是 v_i 的一个子孙节点。

定义 1 在 $G=(V, E)$ 中, 如果在两个不相邻的变量 v_i 和 v_j 之间存在路径 l 并满足以下条件:

1) l 包含一条链 $v_i \rightarrow z \rightarrow v_j$ 或一个分叉 $v_i \leftarrow z \rightarrow v_j$, 其中, $z \in Z$ 。

2) l 包含一个碰撞 (又称 V-结构^[6]) $v_i \rightarrow z \leftarrow v_j$, 其中, z 及其子孙节点都不包含在 Z 中, 则称变量 v_i 与 v_j 被 Z 集合 d-分离, 或称变量 v_i 与 v_j 被 Z 集合阻断。

LIU 等^[14]指出, d-分离可以被扩展到变量集中, 在忠诚性假设下, 如果存在两个不相连的变量集 A 和变量集 B , 则满足: $\forall \mu \in A, \forall v \in B, \mu, v$ 被 Z 集合 d-分离, 则变量集 A 和变量集 B 被 Z 集合 d-分离。

基于上述定义, 本文给出因果关系的相关定义。

根据 PEARL 的图模型理论, 其对于因果贝叶斯网络的定义^[5]为:

定义 2 令 $P(v)$ 表示变量集 V 的分布, $P_x(v)$ 表示变量 v 经过干预 $do(X=x)$ 后得到的干预分布, 其中, $X \subseteq V$, 即将部分变量 v 的集合 X 值设为某个常数 x , P_* 表示干预分布 $P_x(v)$ 和非干预分布 $P(v)$ 的集合。在一个 DAG 中, 对于所有 $P_x \in P_*$, 当且仅当 DAG 满足以下三个条件, 则称 G 是分布 P 相对应的贝叶斯因果网络: 1) $P_x(v)$ 是 G 的马尔可夫类; 2) 对于所有 $V_i \in X$, 当 v_i 与 $X=x$ 一致时, 则 $P_x(v_i)=1$; 3) 对于所有 $V_i \notin X$, 当 pa_i 与 $X=x$ 一致时, 则 $P_x(v_i|pa_i)=P(v_i|pa_i)$, 其中, pa_i 表示变量 v_i 的父亲变量。

2.2 算法框架

本文基于因果图模型, 给出因果分解的定义。

定义 3 在 $G=(V, E)$ 中, 如果存在 $\bigcup_{i=1}^m C_i = V$ 且 $C_i \not\subseteq C_j (i \neq j)$, 则称变量集 $C=\{C_1, C_2, \dots, C_m\}$ 组成变量集 V 的一个因果分解, 并令分割集 S_{ij} 表示 C_i 和 C_j 的交集: $S_{ij}=C_i \cap C_j$ 。其中, 分割集 S_{ij} 阻断节点 $\mu \in C_i/S_{ij}$ 与节点 $v \in C_j/S_{ij}$ 的所有路径, 即 $\mu \perp v | S_{ij}$ 。

因果分解实例示意图如图 1 所示, 假设图 1(a) 由变量集 V 学习得到的无向独立图, 则它的因果分解为图 1(b) 所示的 5 个子集: $C_1=\{V_1, V_3, V_4\}$, $C_2=\{V_3, V_4, V_6\}$, $C_3=\{V_6, V_7, V_8\}$, $C_4=\{V_3, V_6, V_7\}$, $C_5=\{V_2, V_3, V_5, V_7\}$ 。

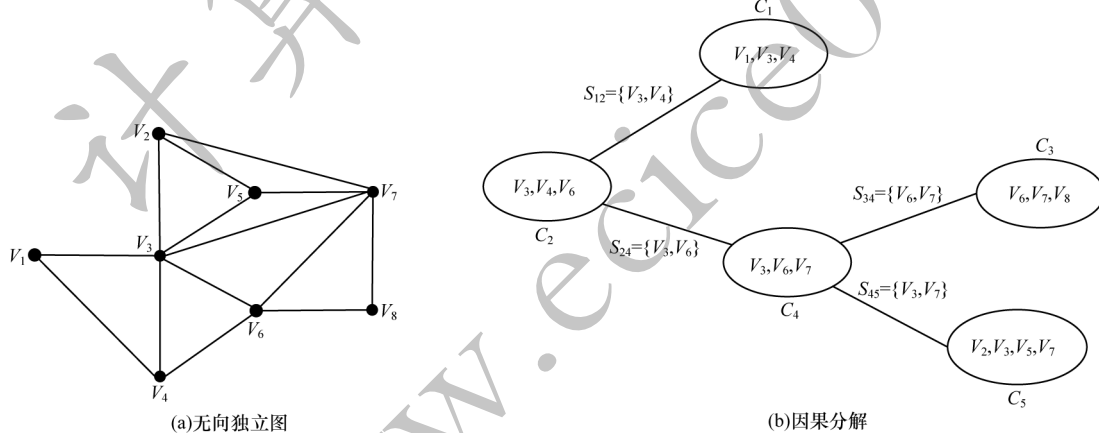


图 1 因果分解实例示意图

Fig.1 Schematic diagram of example for causal decomposition

因此, 变量集 V 的因果结构学习问题可以被转换为 m 个子集 $\{C_1, C_2, \dots, C_m\}$ 上的子结构学习问题。该分解过程可以被递归地执行, 直到子问题小到预先设定的特定阈值 θ 。本文提出的递归因果结构学习算法 CADR 如算法 1 所示。

算法 1 递归因果结构学习算法 CADR

输入 样本集 D , 变量集 V , 因果分解阈值 θ , 特定的因果结构学习算法 A_g

输出 完全定向的因果图 G

1. if 变量集大小 $|V| \leq \theta$, then
2. 在样本集 D 上运行算法 A_g , 并返回 G
3. else

4. 从变量集 V 上搜索一个因果分解 $C=\{C_1, C_2, \dots, C_m\}$

5. 递归地分解并学习每个子结构

6. $G_1 = \text{CADR}(D, C_1, \theta, A_g)$

$G_2 = \text{CADR}(D, C_2, \theta, A_g)$

$\vdots$

$G_m = \text{CADR}(D, C_m, \theta, A_g)$

7. 合并子因果图 $G_1, G_2, \dots, G_m$ 为一个完整的因果图 G

8. 返回 G

该算法的输入包括样本集 D 、变量集 V 、因果分解阈值 θ , 以及特定的因果结构学习算法 A_g 。当分解后的子问题足够小 (达到阈值 θ) 时, 递归分解结束。同时, A_g 用于从无法被分解的子问题中进行因果关系发现; 否则, 继续分解为 m 个更小的子集。算法 1

的关键是找到一个因果分解 $C=\{C_1, C_2, \dots, C_m\}$ 。

2.3 因果分解学习

CADR算法的关键是通过学习得到因果分解。为识别输入变量之间的潜在因果分解,本文采用条件独立性检验方法得到因果分解,其详细过程如算法2所示。

算法2 寻找因果分解

输入 样本集 D , 变量集 V , 条件候选集大小阈值 q

输出 因果分解 $C=\{C_1, C_2, \dots, C_m\}$

1. 令 k 表示条件候选集大小, 从 0 开始逐步增加 1 来学习一个无向独立图 A

2. 把 A 分解成 m 个团

3. 从 A 中找到一个较优的消除顺序

4. 根据该顺序把 A 变成一个三角弦图 TA [26]

5. 把 TA 分解成 m 个 cliques: $C=\{C_1, C_2, \dots, C_m\}$

6. 返回因果分解: $C=\{C_1, C_2, \dots, C_m\}$ 。

首先, 通过条件独立性学习变量集 $V=\{v_1, v_2, \dots, v_n\}$ 的一个无向独立图 A , 表示各变量之间的条件独立性(是否有边连接), $A_{ij}=1$ 表示变量 v_i 和 v_j 之间有边连接, 在 2.4 节将会给出学习 UIG 的详细过程; 然后, 利用 XIE 等 [12] 提出的方法, 把变量集 V 分解成 m 个团 (cliques): $C=\{C_1, C_2, \dots, C_m\}$, 即需要寻找的因果分解。

本文通过上述方法将高维复杂的因果结构学习任务分解到合适的维度上进行学习, 缩小条件独立性检验的搜索空间, 进而获得更精准的因果结构。

2.4 无向独立图学习

现有基于马尔可夫毯 [27] 发现算法和偏相关方法都可用于无向独立图的发现, 然而, 基于马尔可夫毯的算法因恢复出精确的道德图, 导致算法复杂度增大。在实际中, 除道德图的边和骨架的边以外, 一些额外的边在 UIG 中也是允许存在的 [14]。

基于偏相关的方法只能运用在线性高斯数据中, 无法在非线性非高斯的情况下使用。而基于核空间的条件独立性检验方法则适用于非线性非高斯的情况, 如 HSIC、KCIT 等方法, 但是当变量集较大而样本量较少时, 条件独立性检验的结果则变得不准确。此外, 基于核的方法在样本量较大时, 其时间复杂度也较高。针对该问题, STROBL 等 [28] 基于随机傅里叶特征方法提出非参数条件独立检验算法 RCIT。

RCIT 算法通过随机傅里叶特征算法 $k(x, y)=E[\zeta(x), \zeta(y)]$, 将变量映射到一个低维的欧几里得空间, 如式 (1) 所示:

$$\zeta(x)=\sqrt{2} \cos(W^T x+B)$$

$$W \sim P_W, B \sim \text{Uniform}([0, 2\pi]) \quad (1)$$

通过部分互协方差算子 $\Sigma_{XY \cdot Z}=\Sigma_{XY}-\Sigma_{XZ}^{-1} \Sigma_{XZ} \Sigma_{ZY}$ 近似

计算其条件独立性, 若 $\Sigma_{XY \cdot Z}=0$, 则 $X \perp Y|Z$ 。

因此, 本文结合 RCIT 方法提出一种 q -order 条件独立性检验方法。其中, q 表示条件候选集 S 大小的阈值。为构造一个无向独立图 A , 首先从一个完全有向图出发, 在给定条件候选集 S 的情况下检验两个节点变量 v_i 和 v_j 的条件独立性进行删边操作, 如果独立, 则删除 v_i 与 v_j 之间的边, 即 $A_{ij}=A_{ji}=0$ 。条件候选集 S 的大小 k 从 0 开始依次增加, 且每次迭代其大小满足: $|S| \leq q$ 。通过逐步增加条件候选集的大小, 可以有效地缩减算法条件候选集的搜索空间, 提高算法运行效率。算法伪代码如算法 3 所示。

算法3 学习无向独立图 UIG

输入 样本集 D , 变量集 V , 条件候选集大小阈值 q

输出 无向独立图 A

1. 初始化 $A=\text{Ones}(|V| \times |V|)$, $k=0$

2. if $k \leq q$ then

3. for each $v_i, v_j \in V$, 且 $A_{ij}=1$ do

4. if $\exists S \subseteq \text{pa}(v_i) \cup \text{pa}(v_j) (|S|=k)$, 并满足 $v_i \perp v_j|S$ then

5. 令 $A_{ij}=A_{ji}=0$

6. 令 $k=k+1$

7. 返回 A

2.5 马尔可夫等价类识别

无向独立图学习是基于约束的方法实现的, 而基于约束的方法通过依次检验每对节点的条件独立性来判断是否存在边, 进而得到一个基本骨架。该基本骨架根据 V-结构和其他一些规则, 如 Meek's rules [29] 尽可能多地确定边的因果方向 [30]。但这种做法会存在马尔可夫等价类的问题, 无法确定方向的因果边。假设存在一个局部结构 $x-z-y$, 且 $x \perp y|Z$, 根据定义 1, 如果存在一个变量 $z \notin Z$, 且其任何子集都不包含在 Z 中, 则确定该局部结构为一个 V-结构: $x \rightarrow z \leftarrow y$ 。当 $z \in Z$ 时, 则无法通过条件独立性来确定该局部结构的因果方向, 即该结构是一个马尔可夫等价类。在给出定理 1 的前提下, V-结构的可识别性可以通过定理 2 来保证。

定理 1 在因果马尔可夫和因果忠诚性假设下, 当且仅当两个 DAG 具有相同的 V-结构时, 即两个聚合的箭头, 其尾部没有箭头连接, 它们在观测上是等价的 [5]。

定理 2 在因果马尔可夫和因果忠诚性假设下, V-结构是可识别的。

证明 在因果马尔可夫和因果忠诚性假设下, 如果 V-结构不可识别, 那么存在两个结构 G 和 G' , 使得 $\{v_i, v_j, v_k\}$ 在 G 中是 V-结构, 在 G' 上不是 V-结构, 且该两个结构是等价的。根据定理 1, 若两个结构等价, 则一定存在相同的 V-结构, 所以矛盾。因此, V-结构可识别。

对于通过条件独立性检验方法得到的无向图, 可以先识别V-结构, 确定相应的因果方向。此时, 需要解决的问题: 当 $z \in Z$ 时, 如何确定因果方向。

针对这一问题, 受因果函数模型的启发, 本文考虑利用加性噪声模型识别在马尔可夫等价类变量之间的因果关系, 其识别方法通过定理3来验证。

定理3 假设一组变量集合 V 由非线性非高斯加性噪声模型生成, 其所有子集相应的子结构都满足忠诚性假设, 且存在两个随机变量 $x, y \in V$ 以及由其他变量组成的集合 Z 满足 $x \perp y | Z$, 并存在 $z \in Z$ 与 x, y 相连。 n_{xz} 或 n_{yz} 表示利用 z 回归 x 或 y 得到的残差, 当 $x \perp n_{xz}$ (或 $y \perp n_{yz}$) 成立时, 则 z 是 x 的原因, 即 $z \rightarrow x$ (或 z 是 y 的原因: $z \rightarrow y$)。

证明 假设一个非线性非高斯加性噪声模型 $x = f(z) + n$, 通过 z 回归 x 得到 $\tilde{f}(z)$, 计算对应的残差 $\tilde{n} = x - \tilde{f}(z)$ 。根据非线性非高斯加性噪声的不对称性^[20], 可以推断出: 当 $\tilde{n}$ 与 z 独立时, z 是 x 的原因, 即 $z \rightarrow x$; 类似地, 可以判断反向模型 $z = f(x) + n$, 并判断拟合得到的残差 $\tilde{n}$ 是否与 x 独立, 从而判断因果方向。因此, 在非线性非高斯加性噪声假设和忠诚性假设下, 可以唯一识别出局部马尔可夫等价类结构 $x - z - y$ 。

基于定理3, 得到马尔可夫等价类的识别方法。首先, 根据条件独立性检验得到无向图, 结合定义1识别V-结构; 其次, 在剩余未定向的结构中根据定理3确定其边的因果方向。通过上述方法获取到包含更多因果边信息的因果图, 从而解决马尔可夫等价类问题。

2.6 时间复杂度分析

CADR算法的时间复杂度主要取决于条件独立性检验的次数。在CADR算法中, 原始变量集合 $V = \{v_1, v_2, \dots, v_n\}$ 被因果分解为 m 个子集 $C = \{C_1, C_2, \dots, C_m\}$, 且 $|C_m| \leq n$, 则该求解因果分解集合的时间复杂度为 $O(mk_{\max}^2 2^{k_{\max}-2})$, 其中, $k_{\max}$ 表示子集变量最多的数量, $k_{\max} = \max(|C_1|, |C_2|, \dots, |C_m|)$ 。同时, CADR算法还需要计算每次迭代过程中条件独立性检验的次数。基于无向独立图进行 q 阶条件独立性检验的每次迭代都需要遍历所有变量对, 其时间复杂度为 $O(n^{q+2})$ 。因此, CADR算法的时间复杂度为 $O(mk_{\max}^2 2^{k_{\max}-2} + n^{q+2})$ 。当数据维度高于100, q 设置为3时, CADR可以得到准确的实验结果, 且缩短运行时间。

3 实验与结果分析

本文通过大量的仿真数据实验以及真实贝叶斯网络实验来验证CADR算法的有效性, 并与XIE

等^[12]提出的递归因果结构学习方法(Xie_rec)、传统PC+ANM的方法(PC_ANM)、ZHENG^[31]等提出的非参数因果结构学习方法(Notear_Sob)进行对比。实验指标为召回率(Recall)、精确率(Precision)和F1评分。实验环境为Intel®Core™ i7-9750H CPU @ 2.60 GHz, 编程环境为Python 3.7。

3.1 仿真数据实验

本文在非线性非高斯模型假设下仿真因果贝叶斯网络数据, 数据生成方式与CAI等^[13]提出的方法类似。在仿真实验中, 首先随机生成一组根节点, 其次迭代地随机选择根节点, 并将其作为父亲节点来生成后代节点, 其数据产生机制表示为:

$$v_i = \sum_{v_j \in pa(v_i)} f(v_j) + n_i$$

其中: $f(\cdot)$ 表示非线性函数, 从 $\{x^2, x^3, \sin(x), \cos(x)\}$ 中随机选取; 噪声 n_i 服从非高斯分布。

本文的仿真数据实验基于不同维度和样本量设置一系列随机实验, 其变量维度范围为 $\text{var} = \{10, 20, 30, 40, 50\}$, 样本大小为 $n_{\text{sample}} = \{100, 200, 300, 400, 500\}$ 。基于不同的参数设置, 随机生成因果图, 根据非线性非高斯结构方程模型随机生成数据, 得到不同参数设置下的样本。

在仿真数据实验中, 本文考虑到非线性非高斯加性噪声在维度小于7时的实验效果较好, 因此, 当因果分解时子集的大小阈值 θ 设置为6, 在条件独立性检验中, 设条件候选集的大小阈值 $q=3$ 。每个变量维度和样本量的组合数据实验都重复20次, 取每组实验结果的平均值作为评价指标。

不同算法的仿真实验结果如图2所示。从图2可以看出: CADR算法的因果结构学习效果比其他三种因果结构发现算法更好。在不同的样本量上 (如图2(a)所示), CADR算法几乎在所有样本量的F1评分都要优于其他三种算法。同时, 本文在不同变量维度的情况下 (如图2(b)所示) 的实验效果优于其他三种算法的效果。基于递归分解的算法Xie_rec的F1评分优于传统算法PC_ANM和Notear_Sob算法, 且其效果差距随变量数的增加而不断增加。基于递归分解的算法将高维小样本的因果结构学习转化到低维, 以有效缓解原本在高维因果结构学习样本量不足的问题。在样本量不变、维度减小的情况下, 基于递归分解算法的计算准确率得到提升。在基于递归分解的算法中, CADR在实验中的F1评分高于Xie_rec, Xie_rec方法经过条件独立性检验后, 只是通过V-结构和其他一些定向规则来确定因果方向, 可能存在部分无法识别的无向边, 而CADR算法可以有效区分马尔可夫等价类, 得到一个完全有向的因果结构。

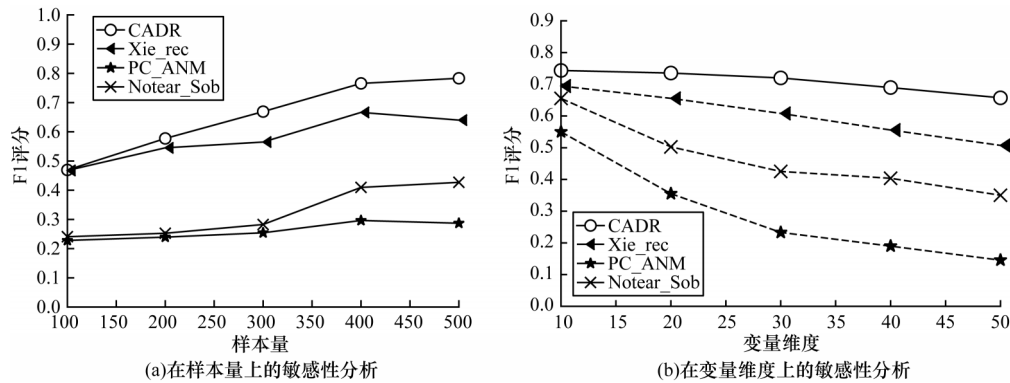


图2 不同算法的仿真实验结果对比

Fig.2 Simulated experimental results comparison among different algorithms

3.2 马尔可夫等价类识别实验

为进一步评估 CADR 算法在因果推断的优势，本文对比 CADR 算法和 Xie_rec 算法在因果结构学习的能力。CADR 与 Xie_rec 方法的马尔可夫等价类识别结果对比如图 3 所示。真实因果图如图 3(a)

所示，根据该因果图，本文通过非线性非高斯结构方程模型生成数据，并在样本量足够的情况下把两种算法应用到该数据中，以恢复真实因果图。图 3(b) 和图 3(c) 所示为由 CADR 算法和 Xie_rec 算法学习得到的因果图。

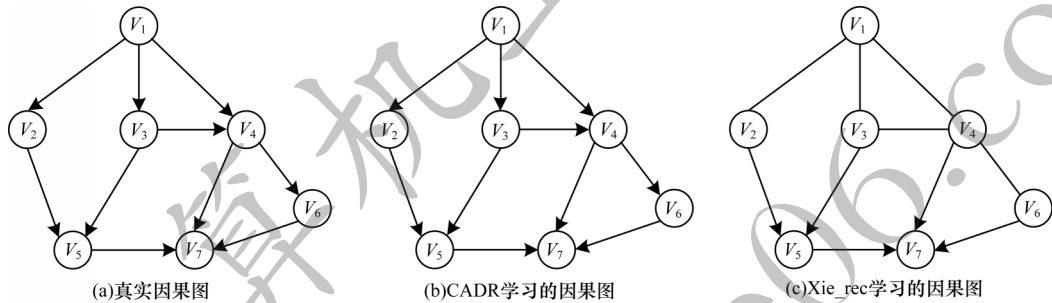


图3 CADR与Xie_rec方法的马尔可夫等价类识别结果对比

Fig.3 Recognition results of Markov equivalence classes comparison among CADR and Xie_rec methods

通过 CADR 可以完全恢复出真实的因果图，尽管 Xie_rec 能恢复出正确的骨架和 V-结构，但无法区分马尔可夫等价类和得到完全有向的因果图。根据定理 1，CADR 通过原因变量拟合结果变量以得到噪声，并利用非线性非高斯数据中因果关系的不可逆性得到因果方向。因此，CADR 算法可以有效区分马尔可夫等价类，识别出所有边的方向，得到一个完全有向的因果图。

3.3 真实贝叶斯网络结构实验对比

本文在非线性非高斯加性噪声的模型下，通过对不同的真实贝叶斯网络结构进行对比，并根据实验结果评估本文所提的 CADR 算法。这些真实贝叶斯网络结构一般涵盖各种应用，包括医学研究的 Alarm 数据集、Pathfinder 数据集、天气预测的 Hailfinder 数据集、系统故障诊断的 Win95pts 数据集以及智能教学指导系统的 Andes 数据集、猪育种系谱 Pigs 数据集以及基因调控网络 Link 数据集。表 1 所示为真实贝叶斯网络结构的统计数据。从表 1 可以看出：贝叶斯网络结构中变量的父亲节点(最大入度)均不超过 6，说明真实因果网络结构具有稀疏性的特点。

表1 真实贝叶斯网络结构的统计数据

Table 1 Statistical data of real Bayesian network structure				
数据集	变量数	边数	平均入度	最大入度
Alarm	37	46	1.24	4
Hailfinder	56	66	1.18	4
Win95pts	76	112	1.47	6
Pathfinder	109	195	1.79	5
Andes	223	338	1.52	6
Pigs	441	592	1.34	2
Link	724	1 125	1.55	3

当样本量设置为变量数的 3 倍时，本文将实验得到的召回率、精确率和 F1 评分作为评价指标。真实贝叶斯网络结构的评价指标如表 2 所示，加粗表示最优的数据。CADR 算法在所有真实结构上具有更优的精确率和 F1 评分，验证本文提出的 CADR 算法的有效性，特别是在高维的情况下。CADR 算法在不同数据集上的运行时间均低于其他算法。从表 2 可以看出：CADR 算法在因果结构学习的准确率和效率方面具有较优的效果。实验结果表明，CADR 算法在高维小样本数据的因果结构学习问题上具有较强的识别能力，能够提升因果结构学习的效率和识别正确率。

表 2 真实贝叶斯网络结构的实验结果
Table 2 Experimental results of real Bayesian network structures

数据集	召回率				精确率				F1 评分				运行时间/s			
	CADR	PC_ ANM	Xie_ rec	Notear_ Sob	CADR	PC_ ANM	Xie_ rec	Notear_ Sob	CADR	PC_ ANM	Xie_ rec	Notear_ Sob	CADR	PC_ ANM	Xie_ rec	Notear_ Sob
Alarm	0.43	0.59	0.35	0.75	0.48	0.20	0.31	0.20	0.45	0.30	0.33	0.31	9.09	15.4	24.8	82.7
Hailfinder	0.57	0.74	0.53	0.79	0.49	0.26	0.39	0.29	0.52	0.38	0.45	0.42	12.10	38.6	86.9	159.2
Win95pts	0.48	0.59	0.59	0.71	0.43	0.14	0.32	0.26	0.46	0.22	0.41	0.38	10.70	38.2	111.6	197.7
Pathfinder	0.47	0.62	—	0.85	0.50	0.10	—	0.15	0.48	0.17	—	0.25	85.40	402.3	—	548.5
Andes	0.59	0.52	—	0.95	0.66	0.10	—	0.13	0.62	0.16	—	0.22	1 000.00	6 000.0	—	2 000.0
Pigs	0.65	0.94	—	0.79	0.68	0.08	—	0.11	0.66	0.14	—	0.19	30 000.00	200 000.0	—	1 000 000.0
Link	0.74	0.30	—	—	0.74	0.08	—	—	0.74	0.12	—	—	200 000.00	9 000 000.0	—	—

4 结束语

本文提出一种基于递归分治方法的因果结构学习算法 CADR, 用于高维小样本数据的因果关系学习。通过因果分解方法把基于高维变量集的因果推断问题逐步分解为维度更小的子问题。结合 RCIT 和 PC-style 提出一种高效的算法, 以得到一个无向独立图, 从而减少条件独立性检验的次数。同时, 采用递归因果分解把得到的骨架分解为更小的子图, 减小下一步条件独立性检验的搜索空间。在此基础上, 利用非线性非高斯加性噪声模型的因果关系识别马尔可夫等价类。实验结果表明, CADR 可以得到较为精确的因果图, 能有效提高算法的效率和条件独立性检验的准确率。下一步将针对稠密网络的因果结构学习进行研究, 以提高 CADR 的稳健性。

参考文献

[1] GLYMOUR C, ZHANG K, SPIRITES P. Review of causal discovery methods based on graphical models[J]. *Frontiers in Genetics*, 2019, 10: 524.

[2] 丁梦远, 兰旭光, 彭茹, 等. 机器推理的进展与展望[J]. *模式识别与人工智能*, 2021, 34(1): 1-13.

[3] CAI R C, ZHANG Z J, HAO Z F. Causal gene identification using combinatorial V-structure search [J]. *Neural Networks*, 2013, 43: 63-71.

[4] OMBADI M, NGUYEN P, SOROOSHIAN S, et al. Evaluation of methods for causal discovery in hydrometeorological systems [J]. *Water Resources Research*, 2020, 56(7): 1-10.

[5] PEARL J. *Causality: models, reasoning, and inference*[M]. 2nd ed. New York, USA: Cambridge University Press, 2009.

[6] SPIRITES P, GLYMOUR C, SCHEINES R. *Causation, prediction, and search*[M]. Cambridge, USA: MIT Press, 2001.

[7] KOLLER D, FRIEDMAN N. *Probabilistic graphical*

models: principles and techniques[M]. Cambridge, USA: MIT Press, 2009.

[8] MAI G Z, HONG Y H, CHEN P H, et al. Distinguish Markov equivalence classes from large-scale linear non-Gaussian data[J]. *IEEE Access*, 2020, 8: 10924-10932.

[9] ZHANG H, ZHOU S G, YAN C X, et al. Recursively learning causal structures using regression-based conditional independence test [C]//*Proceedings of Conference on Artificial Intelligence*. [S. l.]: AAAI Press, 2019: 3108-3115.

[10] BERGSMA W. Testing conditional independence for continuous random variables [EB/OL]. [2022-02-10]. https://www.researchgate.net/profile/Wicher-Bergsma/publication/251188458_Testing_conditional_independence_for_continuous_random_variables/links/5406eb460cf2bba34c1e760f/Testing-conditional-independence-for-continuous-random-variables.pdf.

[11] ZHANG H, ZHOU S G, ZHANG K, et al. Causal discovery using regression-based conditional independence tests [C]//*Proceedings of Conference on Artificial Intelligence*. [S. l.]: AAAI Press, 2017: 1250-1256.

[12] XIE X C, GENG Z. A recursive method for structural learning of directed acyclic graphs[J]. *Journal of Machine Learning Research*, 2008, 9(3): 459-483.

[13] CAI R C, ZHANG Z J, HAO Z F. SADA: a general framework to support robust causation discovery with theoretical guarantee[EB/OL]. [2022-02-10]. <https://arxiv.org/pdf/1707.01283.pdf>.

[14] LIU H, ZHOU S G, LAM W, et al. A new hybrid method for learning Bayesian networks: separation and reunion[J]. *Knowledge-Based Systems*, 2017, 121: 185-197.

[15] 陈铭杰, 张浩, 彭昱忠, 等. 基于偏相关性测试的递归式因果推断算法[J]. *计算机工程*, 2022, 48(10): 123-129.

[16] CHEN M J, ZHANG H, PENG Y Z, et al. Recursive causal inference algorithm based on partial correlation test[J]. *Computer Engineering*, 2022, 48(10): 123-129. (in Chinese)

[16] 郭珉, 石洪波, 冀素琴. 贝叶斯网络结构稀疏学习研究进展[J]. *模式识别与人工智能*, 2016, 29(10): 907-923.

[16] GUO M, SHI H B, JI S Q. Survey of sparse structure

- learning of Bayesian networks[J]. Pattern Recognition and Artificial Intelligence, 2016, 29(10): 907-923. (in Chinese)
- [17] KALISCH M, BÜHLMANN P. Estimating high-dimensional directed acyclic graphs with the PC-algorithm[J]. Journal of Machine Learning Research, 2007, 8(3): 613-636.
- [18] COLOMBO D, MAATHUIS M H. Order-independent constraint-based causal structure learning[J]. Journal of Machine Learning Research, 2014, 15(1): 3741-3782.
- [19] BROMBERG F, MARGARITIS D. Improving the reliability of causal discovery from small data sets using argumentation[J]. Journal of Machine Learning Research, 2009, 10: 301-340.
- [20] HOYER P O, JANZING D, MOOIJ J, et al. Nonlinear causal discovery with additive noise models [C]// Proceedings of the 21st International Conference on Neural Information Processing Systems. New York, USA: MIT Press, 2008: 689-696.
- [21] MOOIJ J, JANZING D, PETERS J, et al. Regression by dependence minimization and its application to causal inference in additive noise models[C]// Proceedings of the 26th Annual International Conference on Machine Learning. New York, USA: ACM Press, 2009: 745-752.
- [22] ARTHUR G, FUKUMIZU K, TEO C H, et al. A kernel statistical test of independence[C]// Proceedings of the 20th International Conference on Neural Information Processing Systems. New York, USA: ACM Press, 2007: 585-592.
- [23] GRETTON A, FUKUMIZU K, HARCHAOUI Z, et al. A fast, consistent kernel two-sample test[EB/OL]. [2022-02-10]. <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.157.8330&rep=rep1&type=pdf>.
- [24] FUKUMIZU K, GRETTON A, SUN X H, et al. Kernel measures of conditional dependence[C]// Proceedings of the 20th International Conference on Neural Information Processing Systems. New York, USA: ACM Press, 2008: 489-496.
- [25] ZHANG K, PETERS J, JANZING D, et al. Kernel-based conditional independence test and application in causal discovery[EB/OL]. [2022-02-10]. <https://arxiv.org/ftp/arxiv/papers/1202/1202.3775.pdf>.
- [26] LAURITZEN S L. Graphical models[M]. Oxford, USA: Oxford University Press, 1996.
- [27] 王双成, 苑森淼, 王辉. 基于贝叶斯网络的马尔科夫毯预测学习[J]. 模式识别与人工智能, 2004, 17(1): 17-21.
- WANG S C, YUAN S M, WANG H. Learning Markov blanket prediction based on Bayesian network[J]. Pattern Recognition and Artificial Intelligence, 2004, 17(1): 17-21. (in Chinese)
- [28] STROBL E V, ZHANG K, VISWESWARAN S. Approximate kernel-based conditional independence tests for fast non-parametric causal discovery[J]. Journal of Causal Inference, 2019, 7(1): 1-10.
- [29] MEEK C. Causal inference and causal explanation with background knowledge[EB/OL]. [2022-02-10]. <https://arxiv.org/ftp/arxiv/papers/1302/1302.4972.pdf>.
- [30] 蔡瑞初, 陈薇, 郝志峰. 因果推断与因果性学习研究进展[J]. 中国人工智能学会通讯, 2020, 10(5): 9-14.
- CAI R C, CHEN W, HAO Z F. Research progress in causal inference and causal learning[J]. Journal of Machine Learning and Applications, 2020, 10(5): 9-14. (in Chinese)
- [31] ZHENG X, DAN C, ARAGAM B, et al. Learning sparse nonparametric DAGs[EB/OL]. [2022-02-10]. <https://arxiv.org/abs/1909.13189>.

编辑 薛晋栋