

基于结构方程似然框架的缺失值因果学习算法

郝志峰^{1,2}, 喻建华¹, 乔杰¹, 蔡瑞初¹

(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: 探索事物之间的因果关系是数据科学的核心问题。在实际场景中, 缺失值的存在给基于约束的方法和基于结构方程模型的方法带来巨大挑战。现有的缺失值因果学习方法虽然可以处理随机缺失数据上的因果结构学习问题, 但是对于非随机缺失数据, 学习因果结构网络中的因果对和马尔可夫等价类结构以及校正因缺失导致错误因果方向等仍未得到解决。为此, 基于结构方程似然框架提出新的缺失值因果学习算法 MV-SELF。利用非线性加性噪声模型的条件概率分布可以转换为噪声分布表示性质, 设计一种基于最大化似然的评分, 实现基于评分的因果结构搜索框架。同时, 为解决非随机缺失下的因果结构学习问题, 利用逆概率加权校正工具来恢复缺失数据的联合分布, 从而校正因缺失导致的冗余边和错误因果方向, 实现对缺失数据上的高维因果结构搜索。仿真实验结果表明, 相比 TD-PC、MVPC、Structure EM 算法, MV-SELF 的 F1 值提高了 3%~19%, 能有效区分马尔可夫等价类。

关键词: 结构方程似然框架; 缺失数据; 逆概率加权; 因果方向学习; 加性噪声模型

开放科学(资源服务)标志码(OSID):



源代码链接: <https://github.com/yu-jian-hua/mvself>

中文引用格式: 郝志峰, 喻建华, 乔杰, 等. 基于结构方程似然框架的缺失值因果学习算法[J]. 计算机工程, 2023, 49(12): 63-70.

英文引用格式: HAO Z F, YU J H, QIAO J, et al. Missing value causal learning algorithm based on structural equation likelihood framework[J]. Computer Engineering, 2023, 49(12): 63-70.

Missing Value Causal Learning Algorithm Based on Structural Equation Likelihood Framework

HAO Zhifeng^{1,2}, YU Jianhua¹, QIAO Jie¹, CAI Rui chu¹

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

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

[Abstract] Exploring causal relationships between entities is crucial in data science. In practical scenarios missing values pose significant challenges to both constraint-based and structural equation model-based methods. Although existing causal learning methods effectively address random missing data, discerning causal structures in non-random missing data remains problematic. Challenges include learning causal pairs, identifying Markov equivalence class structures, and correcting causal direction errors in causal structure networks. To tackle these issues, this paper introduces a novel algorithm, MV-SELF, based on the structural equation likelihood framework. This algorithm transforms the conditional probability distribution of a nonlinear Additive Noise Model (ANM) into a representation of noise distribution. Consequently, it enables a maximum likelihood-based scoring mechanism for causal structure search. Additionally, MV-SELF utilizes Inverse Probability Weight (IPW) correction to counteract non-random deletions. This approach effectively restores the joint distribution of missing data, thereby correcting redundant edges and inaccurate causal directions. It facilitates high-dimensional causal structure searches in datasets with missing values. Simulation experiments reveal that MV-SELF outperforms TD-PC, MVPC, and Structure EM algorithms, achieving a 3% to 19% increase in F1 value. This improvement highlights MV-SELF's effectiveness in distinguishing Markov equivalence classes.

[Key words] structural equation likelihood framework; missing datas; Inverse Probability Weight (IPW); causal discovery learning; Additive Noise Model (ANM)

DOI: 10.19678/j.issn.1000-3428.0066474

基金项目: 国家自然科学基金(61876043, 61976052, 62206064); 国家优秀青年科学基金(62122022); 科技创新2030—“新一代人工智能”重大项目(2021ZD0111501)。

作者简介: 郝志峰(1968—), 男, 教授、博士生导师, 主研方向为人工智能、代数及其应用; 喻建华, 硕士研究生; 乔杰, 博士研究生; 蔡瑞初(通信作者), 教授、博士生导师。

收稿日期: 2022-12-08

修回日期: 2023-02-21

E-mail: cairuichu@gmail.com

0 概述

因果发现被认为是数据分析中的强大工具^[1-3]。在实际场景中,数据缺失问题对因果发现算法的实际应用带来了巨大挑战^[4],因为现有的因果发现算法大多数是为完整的数据集而设计的^[5-7]。如果直接^[7]在缺失数据上进行因果结构网络学习,可能会存在冗余骨架^[8]和错误因果方向问题。

为解决在非随机缺失数据上学习因果结构网络问题,现有缺失值因果结构学习算法分为基于约束的方法^[1]和基于结构方程模型的方法^[9]。基于约束的方法包括基于独立性的方法和基于评分的方法^[10]2类,基于独立性的方法分析不同缺失机制对条件独立性(Conditional Independence, CI)测试^[11]结果的影响,从而通过校正条件独立性测试结果来实现因果结构学习,如MVPC算法^[8]。基于评分的方法将目标评分函数与逆概率加权^[12](Inverse Probability Weight, IPW)相结合,从而解决在非随机缺失数据上因果结构校正问题,如HC-IPW算法^[13]。然而,这2类方法都无法识别因果结构网络中的因果对^[14]以及马尔可夫等价类结构^[15],因此无法校正这些结构中出现的错误因果方向。基于结构方程模型的方法,如MissDAG算法^[16],通过假设缺失数据类型为随机缺失,并结合非线性加性噪声模型^[14](Additive Noise Model, ANM)的可识别性假设来最大化插入值的目标似然,而这种假设缺失数据类型是随机缺失方法,并不符合实际场景要求。

因此,对于在非随机缺失数据上,如何识别因果结构网络中的因果对以及马尔可夫等价类结构并校正因缺失导致的错误因果方向,仍是1个具有重要意义和挑战性的问题。本文提出一种基于结构方程似然框架^[17]的缺失值因果学习算法MV-SELF。该算法结合结构方程模型可识别马尔可夫等价类结构、逆概率加权可恢复非随机缺失数据的联合分布和基于评分的方法可搜索高维因果结构网络的优点,使得在非随机缺失数据上,因果结构网络学习和校正成为可能。

1 相关工作

近年来,得益于描述缺失过程的缺失图^[18]以及可恢复性^[19]概念确立,缺失值因果学习研究得到较快发展。在缺失值因果学习研究中,缺失图也被称为因果图,给定1个缺失图,如果1个查询(如条件或联合分布和因果效应)可以得到一致估计,则被认为是可恢复的。

在非随机缺失设定中,STROBL等^[20]将缺失机制视为一种特殊类型的选择偏差,以处理非随机缺失因果结构网络学习问题,并提出基于测试删除(Test-wise Deletion, TD)的FCI算法^[1]。然而,缺失机制一般比选择偏差提供了更多的缺失状态信息,

而选择偏差仅提供选择后的样本分布,因此,选择后的样本分布是有偏的。为了对非随机缺失数据的联合分布进行校正,文献[21]使用逆概率加权对条件独立性测试结果进行纠正,但假设缺失因果模型是已知的,这并不适用于实际应用场景。为此,TU等^[8]提出MVPC算法,指出测试删除可能会导致因果骨架冗余,并使用基于排列的校正方法和逆概率加权的校正方法对缺失数据的联合分布进行校正。但是,基于排列的校正方法仅适合缺失指示变量与缺失变量之间可被d-分离的场景,而不适合缺失指示变量与缺失变量不可被d-分离的场景。而Structure EM算法^[22]通过在候选模型中选择1个模型作为当前模型,并和训练数据一起执行标准的EM算法^[23]得到最大似然对应的参数值,然后固定得到的参数值来迭代候选模型,直至找到最大似然对应的候选模型作为当前模型,依次交替迭代上面过程直至收敛。因此,Structure EM算法仅适用缺失机制可忽略的场景,不适合非随机缺失数据上的因果结构网络学习。

2 缺失值因果学习框架

为了描述所解决的问题并介绍因果结构的学习框架,本节需要形式化定义一些前置概念,如缺失图、缺失图类型以及处理缺失所需的假设知识。

本文定义缺失图 $G = \{V, E\}$ 来表达因果结构网络。其中 $V = V_o \cup V_m \cup U \cup V^* \cup R$, V_o 是完全观测变量集,在缺失图使用白色实线结点表示。 V_m 表示至少含有1条缺失记录的部分观测变量集,在缺失图使用灰色实线结点表示。 U 表示未观测到的变量集,本文假设满足因果充分性假设,因此 U 是空集。 V^* 是代理变量集, R 表示缺失因果机制集,也就是缺失指示变量集,在缺失图使用白色实线结点表示。 R_{v_i} 具体值表示 V_i^* 是否缺失状态信息,如式(1)所示:

$$v_i^* = f_i(r_{v_i}, v_i) = \begin{cases} v_i, r_{v_i} = 0 \\ \text{nan}, r_{v_i} = 1 \end{cases} \quad (1)$$

用大写字母 X_i 表示第 i 个变量结点,以及结点间的因果边集合 $E = \{(i, j) | X_i \rightarrow X_j\}$ 。本文用小写字母 x 表示样本,令数据集 $D = \{x_1^{(i)}, x_2^{(i)}, \dots, x_n^{(i)}\}_{i=1}^m$,其样本量大小为 m 。若满足直接因果关系 $X_i \rightarrow X_j$,则称 X_i 是 X_j 的父母,如果存在间接因果边使得 $X_i \rightarrow X_k \rightarrow X_j$,并称 X_i 是 X_j 的祖先。本文记 $X_i \perp_{\mathcal{G}} X_j | S$ 为给定条件集 S , X_i 与 X_j 在图 G 上条件独立。

缺失图的类型示意图如图1所示。根据缺失数据问题分类方法,可将缺失图分为3类^[24],如果在缺失图中满足 $V_m \cup V_o \cup U \perp R$,那么缺失数据是完全随机缺失(Missing Completely At Random, MCAR),如图1(a)所示。如果在缺失图中满足 $V_m \cup U \perp R | V_o$,那么缺失数据是随机缺失(Missing At Random, MAR),如图1(b)所示。如果缺失图类型既不是

MAR也不是MCAR,那么属于非随机缺失(Missing Not At Random, MNAR),如图1(c)所示。

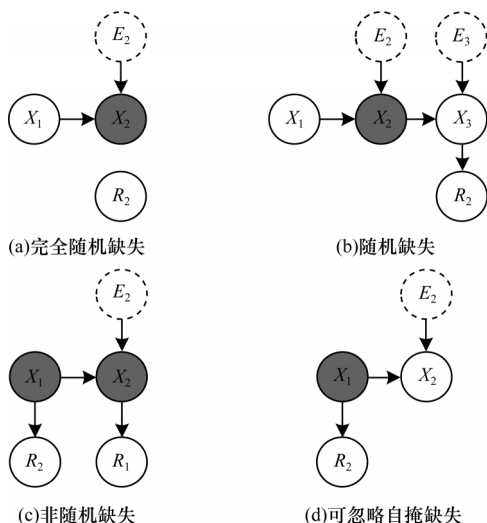


图1 缺失图的类型

Fig.1 Types of missing graphs

为了在缺失数据上进行因果结构网络学习,除了因果结构学习算法所需的基本假设^[25](包括因果马尔可夫假设、因果忠诚性假设、因果充分性假设和无选择偏差)以外,本文还需以下额外假设来解决缺失数据问题。

假设1(缺失指示变量不能是任何观测变量的原因) 该假设在大多数缺失图相关工作^[18-19]中被采用,在该假设下,如果给定条件集和指示变量集,感兴趣的变量之间独立,那么仅给定条件集,感兴趣的变量仍独立。

假设2(忠诚观察性) 任何在观测到的数据集中成立的条件独立性关系同样也在未观测到的数据集中成立,形式化为 $X \perp Y | \{Z, R_k = 0\} \Leftrightarrow X \perp Y | \{Z, R_k = 1\}$ 。该假设意味着缺失数据中的条件独立性关系在完整数据中也成立,即不存在由缺失引起的意外条件独立性关系。

假设3(在缺失指示变量之间无确定性关系) 没有1个缺失指示变量是其他缺失指示变量的确定性函数。

假设4(无自掩缺失) 自掩缺失是指导致部分观测变量缺失的原因是该缺失变量本身,本文假设不会出现自掩缺失情况。

假设3和假设4共同保证了观测变量联合分布的可恢复性^[18]。

图1将缺失图和非线性加性噪声模型相结合,其中虚线圈结点表示注入结果变量的独立噪声。由于非线性加性噪声模型不适合高维因果结构网络搜索,因此需要将非线性加性噪声模型扩展至高维场景^[17],如式(2)所示:

$$P(X_i | Pa(X_i)) \xleftrightarrow{X_i = F_i(Pa(X_i)) + E_i} P(E_i = X_i - F_i(Pa(X_i)) | Pa(X_i)) \xleftrightarrow{Pa(X_i) \perp E_i} P(E_i = X_i - F_i(Pa(X_i))) \quad (2)$$

本文通过式(2)将非线性加性噪声模型中原因变量和噪声变量之间的独立性检测转变为求解全局最大目标评分问题,从而实现对高维因果结构的搜索。

综上,本文的目标是满足假设1~假设4,在非随机缺失数据上识别因果结构网络中的因果对以及马尔可夫等价类结构,并校正由缺失导致的错误因果方向。

图2所示为基于结构方程似然框架的缺失值因果学习框架,其中虚线和虚线箭头分别表示当前缺失图中因缺失数据可能产生的冗余边以及反向边。因此直接在缺失数据上进行因果结构学习可能导致因果结构学习结果不准确,需要校正因果结构。

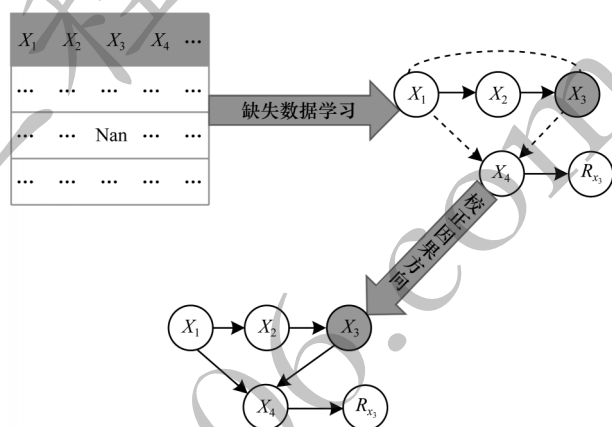


图2 基于结构方程似然框架的缺失值因果学习框架

Fig.2 Framework for causal learning of missing values based on structural equation likelihood framework

3 基于结构方程似然框架的缺失值因果学习算法

3.1 缺失数据的因果结构可识别性

本文简述缺失数据的因果结构可识别性以及逆概率加权校正方法。给定1个缺失图,如果变量的联合分布可通过缺失数据变量表示出来,则变量的联合分布被认为是可恢复的,在进一步满足因果模型所需的假设时,本文认为观测数据的因果结构是可识别的,除此之外,还存在一些无法通过缺失数据变量来表示变量联合分布但仍满足因果模型假设的特殊可识别场景,如图1(d)所示。本文主要讨论变量的联合分布可恢复情况。

由于本文观测到的缺失数据分布为 $P(X_i | Pa(X_i), R=0)$,因此根据缺失图类型分析,当缺失图类型为MCAR时, $P(X_i | Pa(X_i), R=0) = P(X_i | Pa(X_i))$,显然可通过缺失变量表示真实变量联合分布,不影响因果结构学习。但是当缺失图类型为MAR或MNAR时, $P(X_i | Pa(X_i), R=0) \neq P(X_i | Pa(X_i))$,此时无法通过缺失数据变量表示真实变量的联合分布。

因此,本文利用逆概率加权工具对有偏的联合分布进行校正,如式(3)所示:

$$\begin{aligned}
P(V) &= \frac{P(R=0, V)}{\prod_i P(R_i=0|Pa_{r_i}^o, Pa_{r_i}^m, R_{Pa_{r_i}^m}=0)} = \frac{P(V|R=0)P(R=0)}{\prod_i P(R_i=0|Pa_{r_i}^o, Pa_{r_i}^m, R_{Pa_{r_i}^m}=0)} = \\
&= P(V|R=0)P(R=0) \prod_i \frac{P(Pa_{r_i}|R_{Pa_{r_i}^m}=0)}{P(R_i=0, Pa_{r_i}|R_{Pa_{r_i}^m}=0)} = P(V|R=0)P(R=0) \prod_i \frac{P(Pa_{r_i}|R_{Pa_{r_i}^m}=0)P(R_{Pa_{r_i}^m}=0)}{P(R_i=0, Pa_{r_i}|R_{Pa_{r_i}^m}=0)P(R_{Pa_{r_i}^m}=0)} = \\
&= P(V|R=0)P(R=0) \prod_i \frac{P(Pa_{r_i}|R_{Pa_{r_i}^m}=0)}{P(R_i=0, Pa_{r_i}|R_{Pa_{r_i}^m}=0)} = \\
&= P(V|R=0)P(R=0) \prod_i \frac{P(Pa_{r_i}|R_{Pa_{r_i}^m}=0)}{P(Pa_{r_i}|R_i=0, R_{Pa_{r_i}^m}=0)P(R_i=0|R_{Pa_{r_i}^m}=0)} = \\
&= P(V|R=0) \frac{P(R=0)}{\prod_i P(R_i=0|R_{Pa_{r_i}^m}=0)} \prod_i \frac{P(Pa_{r_i}|R_{Pa_{r_i}^m}=0)}{P(Pa_{r_i}|R_i=0, R_{Pa_{r_i}^m}=0)} = P(V|R=0) \cdot c \cdot \prod_i \beta_{R_i} = \\
&= P(V|R=0) \cdot c \cdot \prod_i P(Pa_{r_i}|R_{Pa_{r_i}^m}=0, \text{do}(R_i=0)) \quad (3)
\end{aligned}$$

其中: $P(V)$ 为变量的联合分布; $P(V|R=0)$ 为删除缺失值后的联合分布; $c = \frac{P(R=0)}{\prod_i P(R_i=0|R_{Pa_{r_i}^m}=0)}$ 为常

数; $\beta_{R_i} = \frac{P(Pa_{r_i}|R_{Pa_{r_i}^m}=0)}{P(Pa_{r_i}|R_i=0, R_{Pa_{r_i}^m}=0)}$ 。

3.2 评分函数

本节先给出标准的贝叶斯网络目标评分函数, 再将逆概率加权加入到评分函数中。为了防止过拟合, 大部分的评分函数都需要包含贝叶斯准则即惩罚项, 如式(4)所示:

$$\begin{aligned}
S_{\text{BIC}}(G|D) &= \sum_{i=1}^n \left(\frac{1}{m} \left(\sum_{j=1}^m \log_a \left(P(X_i|Pa(X_i))R=0 \right) + c + \sum_i \beta_{R_i} \right) - \frac{d_i \log_a(m)}{2} \right) \quad (4) \\
&= \sum_{i=1}^n \left(\frac{1}{m} \left(\sum_{j=1}^m \log_a \left(P(E_{j,i}=X_{j,i} - F_i(Pa(X_{j,i}))|R=0) + c + \sum_i \beta_{R_i} \right) - \frac{d_i \log_a(m)}{2} \right) \right) \quad (5)
\end{aligned}$$

3.3 算法框架

本文将设计算法流程来求解评分函数的最大似然。观察目标评分函数式(5)可得, c 是常数, 校正的关键是求 $\beta_{R_i} = \frac{P(Pa_{r_i}|R_{Pa_{r_i}^m}=0)}{P(Pa_{r_i}|R_i=0, R_{Pa_{r_i}^m}=0)}$, 而要求解 β_{R_i} 须先求 Pa_{r_i} , 如算法1所示。

算法1 基于约束的方法学习缺失指示变量的父亲结点集

输入 样本集 $D = \{x_1^{(i)}, x_2^{(i)}, \dots, x_n^{(i)}\}_{i=1}^m$
 输出 缺失指示变量的父亲结点集 Pa_R
 1. For 每个缺失变量 $V_i \in V_m$ then
 2. $Pa_{R_i} \leftarrow V \setminus V_i$
 3. For $V_j \in V \setminus V_i$ then
 4. If $\forall S \subseteq Pa(R_i), V_j \perp R_i | S, R_j = 0, R_S = 0$ then
 5. 从 Pa_{R_i} 中移除 V_j 结点
 6. End if
 7. End For
 8. End For
 9. Return 缺失指示变量的父亲结点集 Pa_R

为进一步求解 β_{R_i} , 本文需要确定成对删除处理缺失数据集所涉及到的结点集 W 以及逆概率加权求解所需的充分变量集 U 。首先有当前的最佳

$$S_{\text{BIC}}(G|D) = \sum_{i=1}^n \left(\frac{1}{m} \left(\sum_{j=1}^m \log_a (P(X_i|Pa(X_i))) - \frac{d_i \log_a(m)}{2} \right) \right) \quad (4)$$

其中: n 为变量数; m 为样本量; $\frac{d_i \log_a(m)}{2}$ 为惩罚项; d_i 为估计 X_i 所需的系数个数; $\frac{1}{m}$ 是为了消除因缺失导致的样本量不同, 从而降低对学习目标评分的影响。

结合式(2)和式(3), 最后得到的目标评分函数如下:

DAG G 与下一步将要搜索的 DAG G_{nei} 2 个图, 将这 2 个图进行比较, 2 个图中父亲结点集不相同的结点为 V_i , 则成对删除所涉及的变量集 $W = \bigcup_{V_i \in V_d} \{V_i, Pa(V_i), Pa(V_i)_{\text{nei}}\}$ 。由于在求解 $Pa(V_i)$ 和 $Pa(V_i)_{\text{nei}}$ 时可能会引入新的缺失变量和 β_{R_i} 中含有 $R_{Pa_{r_i}^m}$, 因此需要迭代并入当前集合中缺失变量 X_i 的 Pa_{r_i} 操作。最终得到充分变量集 $U = \bigcup_{V_i \in V_d} \{V_i, Pa(V_i), Pa(V_i)_{\text{nei}}\} \cup Pa_{R_v}, Pa_{R_v}$ 是 1 个递归的过程, 直至所涉及到的变量集不再更新为止, 其作用如算法2所示。

算法2 基于结构方程似然框架的缺失值因果学习算法

输入 样本集 $D = \{x_1^{(i)}, x_2^{(i)}, \dots, x_n^{(i)}\}_{i=1}^m$
 输出 因果结构
 1. $G \leftarrow$ 空图
 2. $G_{\text{record}} \leftarrow \{G\}$
 3. 使用算法1学习缺失指示变量的父亲结点集 Pa_R
 4. Repeat
 5. $\delta \leftarrow 0$
 6. Repeat
 7. 通过在 G 中添加、反向、删除一条边构造邻居 DAG

G_{nei}

```

8. If  $G_{nei} \notin G_{record}$  then
9. 给定充分变量集  $U$ , 通过局部删除构造  $D_{pw}$ 
10. 在  $D_{pw}$  上通过式(1)计算权值  $\beta_{R_i}$ 
11. If  $S(G_{nei}|D_{pw}, \beta) - S(G|D_{pw}, \beta) > \delta$  then
12.  $\delta \leftarrow S(G_{nei}|D_{pw}, \beta) - S(G|D_{pw}, \beta)$ 
13.  $G_{update} \leftarrow G_{nei}$ 
14. End if
15. End if
16. until 穷举完所有可能的边操作
17. If  $\delta > 0$  then
18.  $G \leftarrow G_{update}$ 
19.  $G_{record} = G_{record} \cup \{G\}$ 
20. End if
21. until  $\delta = 0$ 
22. Return 因果结构  $G$ 

```

算法2的主要过程是在当前图 G 中添加、反向、删除1条边, 然后构造邻居 DAG G_{nei} 并且计算其目标评分, 最后找出最大目标评分的图。在缺失数据上的目标评分经过逆概率加权调整, 因此可得到正确的因果结构。

4 实验

本文将分别使用虚拟因果结构数据和真实因果结构数据对模型进行评估, 并将 MV-SELF 算法与主流的缺失值因果学习算法 MVPC、TD-PC 算法^[8] 和 Structure EM 算法相比较。为验证在非随机缺失数据上的有效性, 本文将所有实验的缺失机制设为非随机缺失, 将随机选择因果结构上 1/2 的结点作为缺失变量。导致缺失的原因是变量优先考虑缺失变量的缺失孩子结点, 若缺失变量没有缺失孩子结点, 则随机选取除自身以外的其他缺失变量作为原因变量。对于某个变量导致的缺失, 设定为: 例如 X 导致 Y 缺失, 当 X 值增大时, 对应 Y 的值以概率 $P=0.1$ 缺失, 反之对应 Y 的值以概率 $P=0.8$ 缺失。所有的参数敏感性实验都至少运行 100 次以上, 采用 F1 值 (F_1)、准确率 (P)、召回率 (R) 和结构性汉明距离 (Structural Hamming Distance, SHD) 作为评价指标, 计算表达式如下:

$$F_1 = 2 \times \frac{P \times R}{P + R} \quad (6)$$

$$P = \frac{\text{正确边数}}{\text{预测结构边数}} \quad (7)$$

$$R = \frac{\text{正确边数}}{\text{真实结构边数}} \quad (8)$$

而结构性汉明距离表示 2 个有向无环图 PDAGs 中 1 个 PDAG 通过添加、删除反向无向边和有向边匹配到另 1 个 PDAG 所需操作的总次数。

4.1 虚拟因果结构数据集

本文在虚拟结构数据上进行 4 个控制变量实验, 缺失变量数 = {4, 7, 10, 13, 16}, 结构维度 = {5, 10, 15, 20, 25}, 平均入度 = {0.5, 1.0, 1.5, 2.0}, 以及样本数量 = {1 000, 3 000, 5 000, 7 000, 9 000} 实验, 加粗表示实验中设置的默认参数, 其中结构维度敏感度

实验中缺失变量个数设定为总结点数的 1/2。

图 3 和图 4 所示为控制缺失变量数的实验结果。MV-SELF 算法在不同缺失变量数下取得大致相同的结果, 在相同缺失变量数下的实验结果明显优于其他算法, 说明 MV-SELF 算法对缺失变量数并不敏感且效果好。

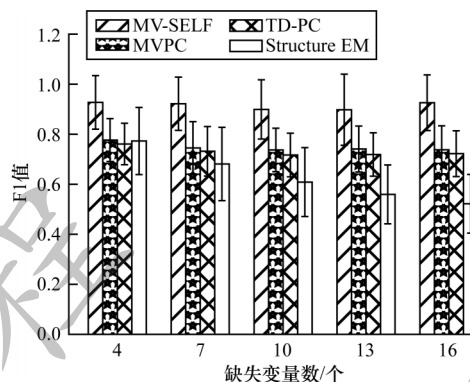


图3 缺失变量控制实验的 F1 值

Fig.3 F1 values of missing variables control experiment

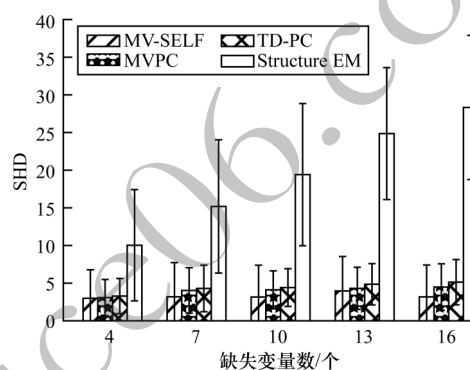


图4 缺失变量控制实验的结构性汉明距离

Fig.4 Structural Hamming distance of missing variables control experiment

图 5 和图 6 所示为控制结构维度的实验结果。从图 5 和图 6 可以看出, MV-SELF 算法在结构维度比较大时同样有效, 说明结构维度对 MV-SELF 算法的影响很小。因此, MV-SELF 算法在不同维度下具有较优的鲁棒性。

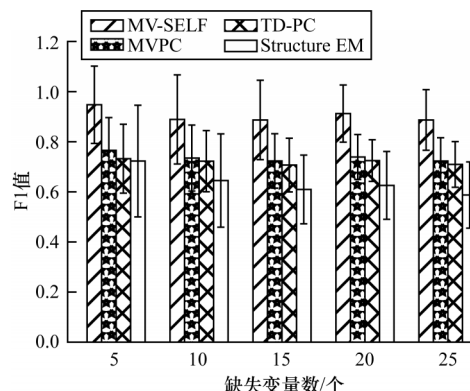


图5 结构维度控制实验的 F1 值

Fig.5 F1 values of structure dimension control experiment

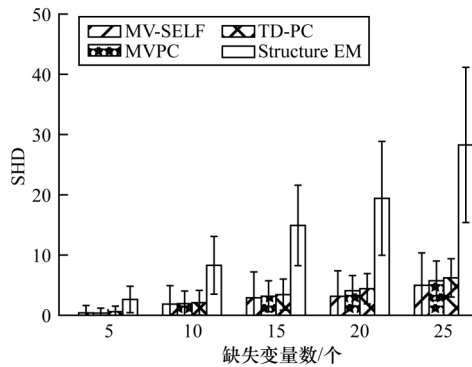


图6 结构维度控制实验的结构性汉明距离

Fig.6 Structural Hamming distance of structure dimension control experiment

图7和图8所示为控制平均入度的实验结果。从图7和图8可以看出,MV-SELF算法在稀疏结构下表现较优,而在稠密结构下表现不佳,但其实验结果仍优于其他对比算法。其原因在稠密结构下准确计算逆概率加权依赖于缺失指示变量的父亲结点 β_{R_i} ,从而导致难度增大,造成实验效果下降。但MV-SELF算法的实验结果仍然比其他对比算法要好,因此实验结果验证MV-SELF算法的有效性。

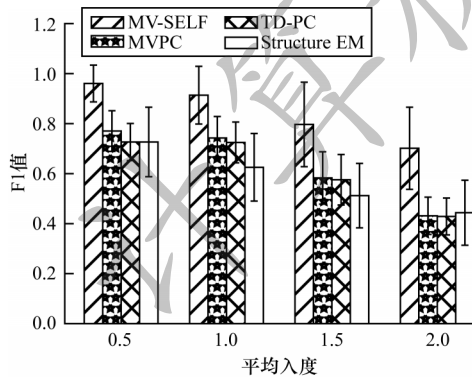


图7 平均入度控制实验的F1值

Fig.7 F1 values of average penetration control experiment

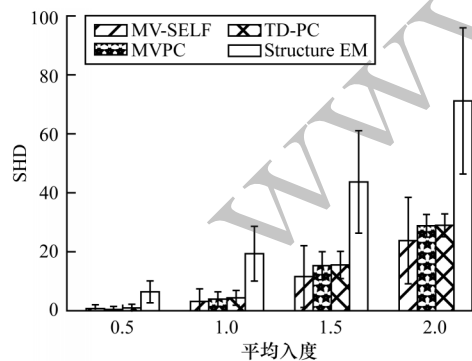


图8 平均入度控制实验的结构性汉明距离

Fig.8 Structural Hamming distance of average penetration control experiment

图9和图10所示为控制样本量的实验结果。从图9和图10可以看出,MV-SELF算法随着样本数量

增加效果越好,说明样本数量有助于实验效果的提升,并且MV-SELF算法明显优于其他比较方法,说明MV-SELF算法具有一定的有效性。

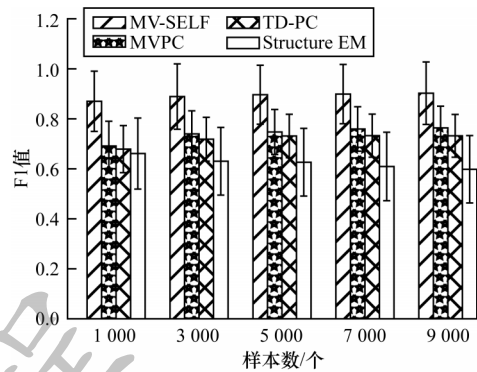


图9 样本数量控制实验的F1值

Fig.9 F1 values of number of samples control experiment

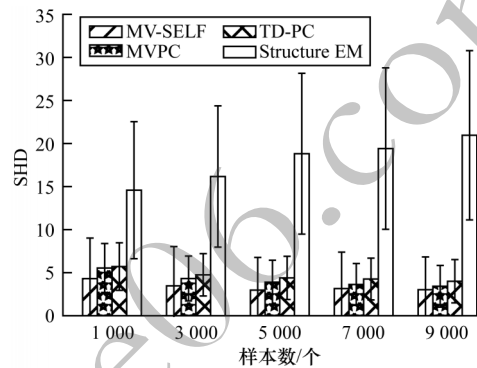


图10 样本数量控制实验的结构性汉明距离

Fig.10 Structural Hamming distance of number of samples control experiment

4.2 真实因果结构数据集

本文选择4个真实结构进行测试(<https://www.bnlearn.com/bnrepository/>)。真实结构数据信息如表1所示。

表1 真实结构数据信息

Table 1 Real structure data information

真实结构	结点数/个	平均度	边数/条	最大入度
ASIA	8	2.00	8	2
SACHS	11	3.09	17	3
ALARM	37	2.49	46	4
HAILFINDER	56	2.36	66	4

在真实结构数据实验中,本文分别使用F1值、准确率、召回率作为评价指标。

本文在真实结构数据中设定与虚拟结构数据实验中相同的实验环境。图11~图13所示为不同真实结构数据的实验结果。从图11~图13可以看到,MV-SELF算法在不同的真实结构中均获得了最

好的效果,特别是在 ASIA 结构下实现了最佳效果,其原因为 ASIA 的平均入度最低,这与虚拟结构数据的实验结果相吻合。与对比算法相比,本文所提算法 MV-SELF 的准确率和召回率均较高,说明比较算法更容易学习到冗余边。由于比较算法没有考虑到因果结构网络的马尔可夫等价类结构,因此出现更多错误边是合理的,这也证明了 MV-SELF 算法的有效性。

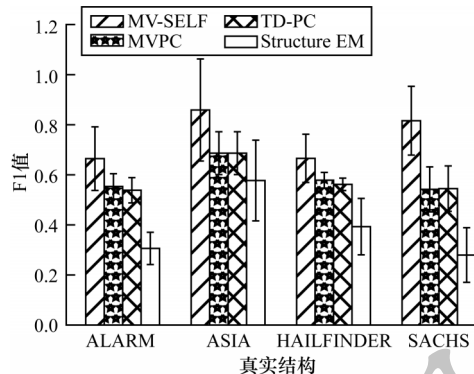


图 11 真实结构对比实验的 F1 值

Fig.11 F1 values of real structure comparison experiment

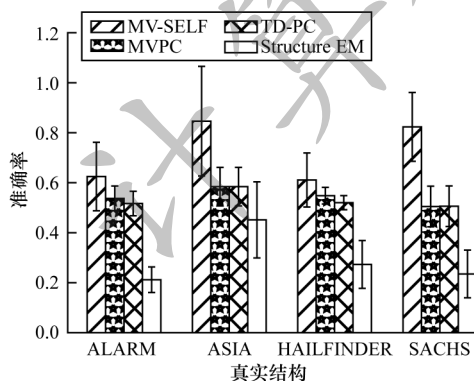


图 12 真实结构对比实验的准确率

Fig.12 Precision of real structure comparison experiment

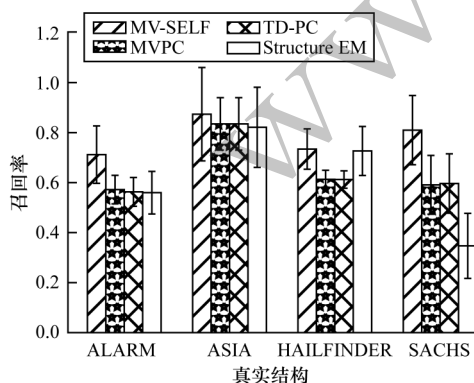


图 13 真实结构对比实验的召回率

Fig.13 Recall of real structure comparison experiment

5 结束语

本文提出一种基于结构方程似然框架的缺失值因果学习算法。通过结合基于非线性加性噪声模型、逆概率加权校正工具,以及基于评分的方法优点,使得在非随机缺失数据上也能进行因果结构学习和校正。实验结果表明,该算法在虚拟因果结构数据集以及真实因果结构数据集上都具有较优的表现。下一步将对存在自掩缺失的因果结构学习问题进行研究。

参考文献

- [1] SPIRITES P, GLYMOUR C N, SCHEINES R, et al. Causation, prediction, and search[M]. Cambridge, USA: MIT Press, 2000.
- [2] 乔杰,蔡瑞初,郝志峰. 基于级联加性噪声模型的因果结构学习算法[J]. 计算机工程, 2022, 48(1): 93-98.
QIAO J, CAI R C, HAO Z F. Causal structure learning algorithm based on cascade additive noise model[J]. Computer Engineering, 2022, 48(1): 93-98. (in Chinese)
- [3] 郝志峰,陈正鸣,谢峰,等. 一种任意分布下的隐变量因果结构学习算法[J]. 计算机工程, 2022, 48(9): 121-129.
HAO Z F, CHEN Z M, XIE F, et al. A learning algorithm for causal structure of hidden variables under arbitrary distribution[J]. Computer Engineering, 2022, 48(9): 121-129. (in Chinese)
- [4] SHPITSER I, MOHAN K, PEARL J. Missing data as a causal and probabilistic problem[EB/OL]. [2022-11-02]. <http://www.auai.org/uai2015/proceedings/papers/204.pdf>.
- [5] KRICH C, RUNGE J, MIRALLES D G, et al. Estimating causal networks in biosphere-atmosphere interaction with the PCMCi approach[J]. Biogeosciences, 2020, 17(4): 1033-1061.
- [6] SHIMIZU S, HOYER P O, HYVÄRINEN A, et al. A linear non-gaussian acyclic model for causal discovery[J]. Journal of Machine Learning Research, 2006, 7: 2003-2030.
- [7] ZHANG K, HYVÄRINEN A. On the identifiability of the post-nonlinear causal model[EB/OL]. [2022-11-02]. <https://arxiv.org/ftp/arxiv/papers/1205/1205.2599.pdf>.
- [8] TU R B, ZHANG K, ACKERMANN P, et al. Causal discovery in the presence of missing data[EB/OL]. [2022-11-02]. <https://arxiv.org/pdf/1807.04010.pdf>.
- [9] GLYMOUR C, ZHANG K, SPIRITES P. Review of causal discovery methods based on graphical models[J]. Frontiers in Genetics, 2019, 10: 524.
- [10] LAM W, BACCHUS F. Learning Bayesian belief networks: an approach based on the MDL principle[J]. Computational Intelligence, 1994, 10(3): 269-293.
- [11] ZHANG K, PETERS J, JANZING D, et al. Kernel-based conditional independence test and application in causal discovery[EB/OL]. [2022-11-02]. <https://arxiv.org/ftp/arxiv/papers/1202/1202.3775.pdf>.
- [12] HORVITZ D G, THOMPSON D J. A generalization of sampling without replacement from a finite universe[J].

- Journal of the American Statistical Association, 1952, 47(260):663-685.
- [13] LIU Y, CONSTANTINOU A C. Greedy structure learning from data that contain systematic missing values [J]. Machine Learning, 2022, 111(10):3867-3896.
- [14] 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: ACM Press, 2008:689-696.
- [15] ANDERSSON S A, MADIGAN D, PERLMAN M D. A characterization of Markov equivalence classes for acyclic digraphs [J]. The Annals of Statistics, 1997, 25(2):505-541.
- [16] GAO E, NG I, GONG M, et al. MissDAG: causal discovery in the presence of missing data with continuous additive noise models [C]// Proceedings of Advances in Neural Information Processing Systems. New York, USA: ACM Press, 2022:1-10.
- [17] CAI R C, QIAO J E, ZHANG Z J, et al. SELF: structural equational likelihood framework for causal discovery [C]// Proceedings of AAAI Conference on Artificial Intelligence. [S. l.]: AAAI Press, 2018:1-10.
- [18] MOHAN K, PEARL J, TIAN J. Graphical models for inference with missing data [C]// Proceedings of Advances in Neural Information Processing Systems. New York, USA: ACM Press, 2013, 26:1-10.
- [19] MOHAN K, PEARL J. Graphical models for recovering probabilistic and causal queries from missing data [EB/OL]. [2022-11-02]. https://www.karthikamohan.com/nips2014_reprint.pdf.
- [20] STROBL E V, VISWESWARAN S, SPIRITES P L. Fast causal inference with non-random missingness by test-wise deletion [J]. International Journal of Data Science and Analytics, 2018, 6(1):47-62.
- [21] GAIN A, SHPITSER I. Structure learning under missing data [C]// Proceedings of Machine Learning Research. New York, USA: ACM Press, 2018:121-132.
- [22] FRIEDMAN N. Learning belief networks in the presence of missing values and hidden variables [C]// Proceedings of the 14th International Conference on Machine Learning. New York, USA: ACM Press, 1997:125-133.
- [23] DEMPSTER A P, LAIRD N M, RUBIN D B. Maximum likelihood from incomplete data via the EM algorithm [J]. Journal of the Royal Statistical Society: Series B (Methodological), 1977, 39(1):1-22.
- [24] RUBIN D B. Inference and missing data [J]. Biometrika, 1976, 63(3):581-592.
- [25] 蔡瑞初, 陈薇, 张坤, 等. 基于非时序观察数据的因果关系发现综述 [J]. 计算机学报, 2017, 40(6):1470-1490. CAI R C, CHEN W, ZHANG K, et al. A survey on non-temporal series observational data based causal discovery [J]. Chinese Journal of Computers, 2017, 40(6):1470-1490. (in Chinese)
- [26] TSAMARDINOS I, BROWN L E, ALIFERIS C F. The max-min hill-climbing Bayesian network structure learning algorithm [J]. Machine Learning, 2006, 65(1):31-78.

编辑 薛晋栋