计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210100053

一种基于信息瓶颈的因果关系挖掘方法

乔　杰[1]　蔡瑞初[1]　郝志峰[2]
1 广东工业大学计算机学院　广州 510006
2 佛山科学技术学院数学与大数据学院　广东 佛山 528000
(qiaojie.chn@gmail.com)

摘　要　观测数据因果关系挖掘是很多学科的基础问题。然而基于约束与因果函数等的现有方法对数据的因果机制具有较强的假设,一般适用于低维数据,并不能很好地适用于存在隐变量的场景。为此,提出了一种基于信息瓶颈的因果关系挖掘方法,称为因果信息瓶颈方法。该方法将因果机制划分为压缩与提取两阶段,在压缩阶段,假设存在一个经过压缩的中间隐变量,在提取阶段,可能保留与结果变量相关的信息。在上述建模的基础上,通过推导其变分上界,设计了一种的基于变分自编码机的因果关系挖掘方法。实验结果表明,基于信息瓶颈的方法在合成数据中准确率提升了10%,在真实数据中准确率提升了4%。

关键词:因果关系挖掘;信息瓶颈;因果发现;因果信息瓶颈;变分自编码机
中图法分类号　TP301.6

Mining Causality via Information Bottleneck

QIAO Jie[1],CAI Rui-chu[1] and HAO Zhi-feng[2]
1 School of Computer,Guangdong University of Technology,Guangzhou 510006,China
2 School of Mathematics and Big Data,Foshan University,Foshan,Guangdong 528000,China

Abstract　Causal discovery from observational data is a fundamental problem in many disciplines. However,existing methods such as constraint-based methods and causal function-based methods have strong assumptions on the causal mechanism of data,and are only applicable to low-dimensional data,and cannot be applied to scenarios with hidden variables. To this end,we propose a causality discovery method using information bottlenecks,called causal information bottleneck. This method divides the causal mechanism into two stages:compression and extraction. In the compression stage,we assume that there is a compressed hidden variable in the middle,while in the extraction stage,we extract the correlated information from effect variable as much as possible. Based on the causal information bottleneck,by deriving its variational upper bound,a causality discovery method based on the variational autoencoder is designed. The experimental results shows that the information bottleneck based method improves the accuracy by 10% in synthetic data and 4% in real world data.

Keywords　Mining causality,Information bottleneck,Causal discovery,Causal information bottleneck,Variational autoencoder

1　引言

近年来,因果关系挖掘及其应用在各个科学领域都受到了越来越多的关注,其中包括推荐系统[1]、气候学[2]、社会学[3]、信息安全[4]等领域。随机控制实验一直是因果关系挖掘方法的黄金标准,然而随机控制实验往往存在难以实施、成本过高,或不道德等风险。因此,如何从观测数据中发现因果关系引起越来越多学者的关注[5-7]。

然而,传统基于观测数据的因果关系挖掘算法往往存在马尔可夫等价类问题,使得两个变量的因果方向无法推断识别。为了解决这一问题,部分工作引入函数式模型,通过判别原因与噪声的独立性关系来判断因果方向,如线性非高斯无环模型(Linear Non-Gaussian Acyclic Model,LiNGAM)[8-9]、加性噪声模型(Additive Noise Model,ANM)[10]、后非线性模型(Post-Nonlinear Model,PNL)[11]等。然而,基于函数式因果模型的方法往往对其模型形式存在很强的假设。在现实中往往存在不可观测的隐变量,隐变量会使得模型假设不再满足,从而导致无法判别或者错误识别出因果关系。

因此,找到一个更加一般的模型,尤其是在存在间接隐变量的情况下推断因果方向,仍然是一个具有挑战性的问题。尽管针对隐变量的因果方向推断问题,最近有一些工作已被提出,如基于隐的紧凑表达机制模型(HCR)[12]提出了一种两阶段因果模型,然而该模型是针对离散数据的,并不能很好地适用于连续数据;与此同时,一种级联噪声模型[13]被提出,

到稿日期:2021-01-07　返修日期:2021-06-01　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61876043,61976052)
This work was supported by the National Natural Science Foundation of China(61876043,61976052).
通信作者:蔡瑞初(cairuichu@gmail.com)

用于解决ANM模型中存在间接隐变量的可识别问题,然而该模型依赖于加性噪声假设,若数据不满足加性噪声模型,则该方法很有可能失效。

本文针对以上问题,提出了因果信息瓶颈(Causal Information Bottleneck,CIB),一种基于信息瓶颈的因果方向推断方法。该方法结合了紧凑表达模型的优势,并以信息瓶颈作为基本框架,提供了一种针对一般场景的因果机制建模方式。我们进一步结合了深度学习中的变分自编码机来对该因果机制进行拟合。该方法不需要对变量间的函数映射形式做过多的约束,在紧凑表达假设下即可进行因果方向推断。

2　相关工作

传统基于观测数据的因果关系挖掘算法可大致分为两类。一类是从时间序列中挖掘因果关系,如面向因果强度的时序因果关系挖掘算法[14]。另一类是基于非时序数据的因果关系挖掘算法,在非时序数据中的因果关系挖掘算法可以进一步划分为两类,一类是基于约束的方法[15],另一类是基于评分的方法[16]。它们的一般思路是对变量间的独立性进行建模,如基于约束的方法是通过检测变量间的独立性来确定因果骨架,并使用 v-结构进一步确定因果方向;基于评分的方法则是基于独立性概率分解构造出概率似然度作为评分函数。然而在没有任何先验知识下,这类传统方法常常存在马尔可夫等价类问题[17]。

为了解决马尔可夫等价类问题,学者们引入了基于结构方程模型的约束,对于因果关系 $X\to Y$,该模型引入了原因变量 X 与噪声 N 的独立性约束并通过检测该独立性来进行因果关系挖掘。例如,线性非高斯无环模型(Linear Non-Gaussian Acyclic Model,LiNGAM)[8-9],假设因果机制满足线性方程 $Y=a^TX+N$;加性噪声模型(Additive Noise Model,ANM)[10]进一步假设因果机制满足非线性方程 $Y=f(X)+N$;而后非线性模型(Post-Nonlinear Model,PNL)[11]进一步引入可逆非线性变化使得 $Y=f_2(f_1(X)+N)$。

然而在现实中,我们往往无法得知因果模型具体的函数形式,此外,观测到的因果关系之间往往存在无法观测的中间变量。例如,考虑因果关系 $X_1\to X_2\to X_3$,如果每一个直接因果对都服从 ANM 模型,则满足 $X_2=f_2(X_1)+N_1$,以及 $X_3=f_3(X_2)+N_3$,但对于 $X_1\to X_3$ 而言,其函数关系不再服从 ANM 模型。虽然最近级联假性加性噪声模型(Cascade Additive Noise Model,CANM)[13]被提出用于解决存在无法观测中间隐变量的问题,然而对于每个直接因果对,其形式仍然受限于加性噪声的假设。

3　基于因果信息瓶颈的因果方向推断框架

本节将形式化定义本文所研究的问题目标。针对更一般的因果关系,我们利用隐变量进行建模,认为原因 X 和结果 Y 的因果关系满足 $X\to Z\to Y$,其中 Z 是在数据中不可观测的隐变量。本文的目标是在存在中间隐变量的数据中区分原因与结果。

如图 1 所示,因果信息瓶颈可以划分为两个阶段,一是压缩阶段,二是拟合阶段。在压缩阶段,我们希望找到一个对 X 的压缩表达 Z;在拟合阶段,我们希望该压缩表达对 Y 具有强表达力。一个简单的例子,如果 X 是温度,Y 是降雪,于是一般在温度低于 0℃时才有可能降雪,而在高于 0℃时不太可能降雪,因此我们提取温度 X 的信息,从而得到压缩的状态 Z,于是 Z 表达了对结果 Y 真正有用的信息,即是否低于 0℃。而这正是信息瓶颈的基本思想。与此同时,利用相反因果方向无法找到该压缩性质,从而进行因果关系挖掘。该框架步骤可概括为:首先采用信息瓶颈对该问题建模来寻找压缩信息 Z;其次,根据其信息压缩的能力及其对目标 Y 的表达能力推导出一个评分准则;最后,通过比较两者评分找出真正的因果方向。

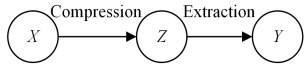

图 1　因果信息瓶颈框架

Fig. 1　Causal information bottleneck framework

3.1　多元变量信息瓶颈

接下来我们将形式化地介绍基于因果信息瓶颈的因果方向推断框架。本节介绍多元变量信息瓶颈,下一节将介绍多元变量信息瓶颈与因果信息瓶颈的联系。

贝叶斯网络中包含了 n 个随机变量 $X_1,\cdots,X_n$,并定义在有向无环图 G 上。对于每个结点 X_i,其父母在图 G 上表示为 $Pa_{X_i}^G$,它是所有指向 X_i 结点的变量的集合。于是其概率可以分解为:

$$P(X_1,\cdots,X_n)=\prod_i P(X_i\mid Pa_{X_i}^G) \tag{1}$$

在该图结构上,根据文献[18],在给定结构 G 下,变量间的互信息可以表示为:

$$\begin{aligned} I(X_1,\cdots,X_n)&=D(P(X_1,\cdots,X_n)\parallel P(X_1)\cdots P(X_n))\\ &=D(\prod_{i=1}^n P(X_i\mid Pa_i^G)\parallel P(X_1)\cdots P(X_n))\\ &=\sum_{i=1}^n I(X_i;Pa_i^G) \end{aligned} \tag{2}$$

其中,D 是 KL-散度,I 是互信息。第二个等式为图 G 上基于马尔可夫条件的概率分解,最后根据互信息的定义,得到第三个等式。

定义 1(多元变量信息瓶颈)　多元变量信息瓶颈的目的是构造以 $\mathcal{L}=I^{G_{\text{in}}}+\gamma D(G_{\text{in}}\parallel G_{\text{out}})$ 最小值为目标的隐变量,其中 $I^G=\sum_{i=1}^n(IX_i;Pa_i^G)$,$\gamma$ 是超参。

基于定义 1,由文献[18],目标函数可以进一步化简,使得:

$$\begin{aligned} D(G_{\text{in}}\parallel G_{\text{out}})&=\sum_i I(X_i;\{X_1,\cdots,X_{i-1}\}-Pa_{X_i}^{G_{\text{out}}}\mid Pa_{X_i}^{G_{\text{out}}})\\ &=I_{G_{\text{in}}}(X_1,\cdots,X_n)-I^{G_{\text{out}}}\\ &=\mathcal{L}^{G_{\text{in}}}-\mathcal{L}^{G_{\text{out}}} \end{aligned} \tag{3}$$

从而,目标函数可以等价地表示为 $\mathcal{L}=I^{G_{\text{in}}}-\beta I^{G_{\text{out}}}$。该目标函数权衡了在 G_{in} 中的互信息压缩,以及在 G_{out} 结构中互信息的提取。换句话说,该目标函数希望尽可能地压缩在图 G_{in} 下每个结点与其父母的互信息,同时,希望最大化在结构 G_{out} 上每个结点与其父母的互信息。

3.2　因果信息瓶颈

本节将给出因果信息瓶颈与多元信息瓶颈的联系,因果信息瓶颈以多元信息瓶颈为基础,引入了压缩与拟合两阶段

的不同因果结构，赋予了多元信息瓶颈刻画一种两阶段因果模型的能力，我们称该方法为因果信息瓶颈准则。

令原因 X、结果 Y 满足因果模型 $X\rightarrow Z\rightarrow Y$，其中 Z 是隐变量。我们希望模型能找到隐变量 Z，使得 Z 是对 X 的一个压缩表示，同时，我们拟合的分布 Z 能够尽可能地表达出因果模型 $X\rightarrow Z\rightarrow Y$。因此，根据第 1 节详细阐述的多元变量信息瓶颈，我们为该因果模型设计了对应的 G_{in} 与 G_{out} 结构，分别如图 2、图 3 所示。

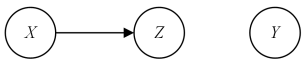

图 2 压缩阶段的 G_{in} 结构

Fig. 2 Structure of G_{in} in compression stage

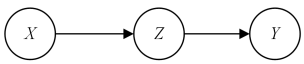

图 3 拟合阶段的 G_{out} 结构

Fig. 3 Structure of G_{out} in fitting stage

上述两个结构分别对应着两个阶段，首先在压缩阶段，我们希望能够最小化互信息 $I^{G_{\text{in}}}$，即最小化 $I(X;Z)$，而在拟合阶段，我们希望最大化互信息 $I^{G_{\text{out}}}$，即最大化 $I(Z;Y)$。一般地，令 G_{in} 的概率分布表示为 $Q(X,Z,Y)$，G_{out} 的概率分布表示为 $P(X,Z,Y)$。基于此，我们可以给出以下因果信息瓶颈的定义。

定义 2（因果信息瓶颈） 针对因果方向 $X\rightarrow Y$，因果信息瓶颈是以图 2 为 $I^{G_{\text{in}}}$、图 3 为 $I^{G_{\text{out}}}$ 的多元信息瓶颈，其目的是构造出以最小值为目标的隐变量 Z。

$$\min \mathcal{L}_{X\rightarrow Y}=I_Q(X;Z)+\gamma D(Q(Y,Z,X)\parallel P(Y,Z,X)) \tag{4}$$

其中，$Q(X)=P(X)$ 以及 $Q(Y)=P(Y)$。

接下来，我们将讨论如何利用这两个结构来进行因果关系挖掘。首先，因果关系挖掘的步骤可以初略地概括为：分别计算正反两方向的因果信息瓶颈评分 $\mathcal{L}_{X\rightarrow Y}$ 与 $\mathcal{L}_{X\leftarrow Y}$，然后评分相对较低的方向作为推断的因果方向。具体地，在拟合 G_{out} 分布时，存在一个表征 Z，使其对因果模型具有最好的表达能力，即：

$$D(Q(Y,Z_{X\rightarrow Y},X)\parallel P(Y,Z_{X\rightarrow Y},X))=D(Q(Y,Z_{X\leftarrow Y},X)\parallel P(Y,Z_{X\leftarrow Y},X)) \tag{5}$$

其中，$Z_{X\rightarrow Y}$，$Z_{Y\leftarrow X}$ 分别是正反方向的表征。根据因果信息瓶颈的基本假设，在因果方向上存在一个有足够压缩能力的表征从而满足以下不等式：

$$I(X;Z_{X\rightarrow Y})<I(X;Z_{X\leftarrow Y}) \tag{6}$$

综合式(5)和式(6)，有以下不等式成立：

$$\mathcal{L}_{X\rightarrow Y}-\mathcal{L}_{X\leftarrow Y}=I(X;Z_{X\rightarrow Y})-I(X;Z_{X\leftarrow Y})<0 \tag{7}$$

这意味着因果信息瓶颈评分较低的方向可被认为是正确的因果方向。即若满足 $\mathcal{L}_{X\rightarrow Y}<\mathcal{L}_{X\leftarrow Y}$，则 $X\rightarrow Y$ 是正确的因果方向，反之亦然。

接下来的问题是如何估计信息瓶颈的值。然而如何计算连续空间中的互信息是一个难题。常用的计算方法是使用 k-近邻等方法将变量离散化，然后进行计算。然而这种方法精度低，离散化效果也不佳，存在很大的估计误差。为了解决这一问题，我们提出了一种变分推断与深度神经网络相结合的求解方法，通过构造互信息的上界来间接优化信息瓶颈的目标函数，不同于文献[19]，我们将其推广到适用于因果信息瓶颈的场景。

4 变分法求解因果信息瓶颈

本节将给出求解因果信息瓶颈的方法，其过程分为两步。第一步，推出式(4)的变分上界；第二步，基于该变分上界，采用变分自动编码机(Variational Auto-Encoder，VAE)[20]来进行优化。

首先，考虑目标函数的第一项互信息 $I_Q(Z;X)$，其上界的推导过程如下：

$$\begin{aligned}I_Q(Z;X)&=\int Q(Z|X)P(X)\log\frac{Q(Z|X)}{Q(Z)}\\&\leqslant\int Q(Z|X)P(X)\log\frac{Q(Z|X)}{r(Z)}\\&=E_{x\sim p(x)}D(Q(Z|x)\parallel r(Z))\end{aligned} \tag{8}$$

其中，$r(Z)$ 是标准正态分布，用于得到一个对互信息近似的上界。对于目标函数的第二项，令 $Q(Z|X)=P(Z|X)$，其 KL-散度可以推导如下：

$$\begin{aligned}&D(Q(Y,Z,X)\parallel P(Y,Z,X))\\&=E_Q\left(\log\frac{Q(Z|X)Q(X,Y)}{P(Y|Z)P(X,Z)}\right)\\&=E_Q\left(\log\frac{1}{P(Y|Z)}\right)+\underbrace{E_Q\left(\log\frac{Q(X,Y)}{P(X)}\right)}_{C}\\&=-E_{Q(X,Y)}(\log P(Y|Z))+C\end{aligned} \tag{9}$$

其中，C 是常数，在优化时可以不考虑，因此在下文中将会被忽略。然后将式(8)、式(9)代入目标函数式(4)中，我们可推导出如下变分因果信息瓶颈的上界：

$$\begin{aligned}\min\mathcal{L}&=I_Q(X;Z)+\gamma D(Q(Y,Z,X)\parallel P(Y,Z,X))\\&\leqslant E_{x\sim p(X)}D(Q(Z|x)\parallel r(Z))-\gamma E_{x,y\sim p(X,Y)}\\&\quad E_{z\sim Q(Z|X)}(\log P(y|z))\end{aligned} \tag{10}$$

给定如上上界，我们可以通过最小化上界得到信息瓶颈的一个估计，然而，如何从样本中估计未知分布 $Q(z|x)$ 和 $P(y|z)$ 的问题仍待解决。为此，我们给出一种基于变分编码机的优化方案。

现考虑样本集合 $\mathcal{D}=\{x^{(i)},y^{(i)}\}_{i=1}^{m}$，对于分布 $Q(z|x)$ 和 $P(y|z)$，我们使用多重感知机(Multilayer Perceptron，MLP)作为全局拟合函数[21]，设它们的参数为 θ，有参数化分布 $Q_\theta(z|x)$ 与 $P_\theta(y|z)$。于是基于蒙特卡洛，我们给出目标函数式(10)的估计如下：

$$\begin{aligned}\mathcal{L}(D;\theta)=&\frac{1}{m}\sum_{i=1}^{m}D(Q_\theta(Z|x^{(i)})\parallel r(Z))-\gamma E_{z\sim Q_\theta(Z|x^{(i)})}\\&(\log P_\theta(y^{(i)}|z))\end{aligned} \tag{11}$$

在式(11)中，分布 Q 的参数并不容易优化，因为其出现在期望中，所以针对该问题一般会采用重参数化解决。具体地，在编码阶段，我们认为分布 $Q_\theta(z|x)$ 是一个编码器，其形式服从 $Q_\theta(z|x^{(i)})=N(z;\mu_\theta(x^{(i)};),\sigma_\theta(x^{(i)})\mathbf{I})$，是一个经过参数化的高斯分布，其中 μ_θ，σ_θ 是 MLP 拟合函数。基于该编码器，目标函数可以重写为：

$$\mathscr{L}(D;\theta)=\frac{1}{m}\sum_{i=1}^{m}D(Q_\theta(Z|x^{(i)})\parallel r(Z))-\gamma E_{\varepsilon\sim N(0,I)}$$
$$(\log P_\theta(y^{(i)}|\mu_\theta(x^{(i)})+\sigma_\theta(x^{(i)})\odot\varepsilon)) \quad (12)$$

在解码阶段，我们认为 $P_\theta(y|z)$ 是解码器。而与此同时，对于式(12)第一项的 KL-散度值，由于 $r(Z)\sim\mathcal{N}(0,I)$ 是标准高斯分布，而分布 Q 也服从高斯分布，故该 KL-散度存在解析解。综上所述，该因果信息瓶颈的目标函数可以通过变分编码器进行求解，通过最小化其上界得到信息瓶颈的估计值。最后给出该模型因果方向识别的求解框架，如算法 1 所示。

算法 1　基于信息瓶颈的因果关系挖掘框架

输入：样本集合 $D=\{x^{(i)},y^{(i)}\}_{i=1}^{m}$

输出：因果方向

1. 使用数据 $\{x^{(i)},y^{(i)}\}_{i=1}^{m}$，用批梯度下降方法（如 Adam[22]）来优化目标函数式(12)，得到目标值 $\mathscr{L}_{X\to Y}$
2. 交换 x，y，即使用数据 $\{y^{(i)},x^{(i)}\}_{i=1}^{m}$，用批梯度下降方法来优化目标函数式(12)，得到目标值 $\mathscr{L}_{Y\to X}$
3. If $\mathscr{L}_{X\to Y}<\mathscr{L}_{Y\to X}$ then,
4.　推断因果方向 X→Y
5. else if $\mathscr{L}_{X\to Y}>\mathscr{L}_{Y\to X}$ then
6.　推断因果方向 X←Y
7. else
8.　无法识别
9. end if

算法 1 详细描述了因果方向识别的基本框架，其基本思路是：分别假设不同的因果方向，然后计算信息瓶颈的目标函数值，最终通过对比两者方向下的信息瓶颈的评分来进行因果方向识别。为分析该算法计算的性能，我们固定神经网络结构与迭代周期数 e，由于固定神经网络结构与参数下的推理时间为常数，因此该算法的总体复杂度为 $\mathcal{O}(em)$，即该复杂度为线性且只与迭代周期数和样本量有关，从而证明了该算法的高效性。接下来，我们进一步通过实验验证所提模型的正确性。

5　实验结果及分析

本节将从合成数据实验与真实数据实验两个方面来测试所提模型的性能。在合成数据实验和真实数据实验中，我们将本文方法(CIB)与先进的因果关系挖掘方法进行对比，对比方法为 ANM 模型[10]、ICA-LiNGAM 模型[8]、CAM 模型[23]。其中，对于 ANM 模型，我们使用了 HSIC 准则[24]作为独立性检验，并且分别对比了不同显著性水平下的结果，具体地，分别取不同的 $\alpha=0.01,\alpha=0.05$，以及不使用 p 值，直接比较两边的独立性情况。在合成数据实验中，我们设计了 3 组控制验，控制隐变量数量 $N=\{1,2,3,4,5\}$，样本大小 $sample\ size=\{250,500,1000,2000,\cdots,8000\}$，以及混淆因子(confounder)数量 $Nco=\{1,2,3,4,5\}$。对于每个实验，其默认设置分别为 $N=2$，$sample\ size=5000$，$Nco=0$。每一组实验中都随机产生至少 500 组因果对。为了验证 CIB 方法的适用性，根据图 1 所示的因果信息瓶颈结构，基于文献[13]，我们设计了每组因果对的产生方式，使得每组因果对间存在 N 个级联隐变量，且级联方式服从加性噪声模型，其中噪声变量服从高斯分布，原因变量服从混合高斯分布。为了产生有一般性的映射函数，我们设计了如下产生方法：先在定义域上平均划分出 6 个点，然后每个点都分别产生 1 个对应的正态分布的随机数，最后使用样条函数对这 6 个点进行插值从而生成一个映射函数。以下实验均在 CPU 型号为 Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz，内存 64GB 的 Ubuntu 环境中执行。

5.1　合成数据实验

如图 4 所示，在合成数据实验中，我们测试了不同隐变量数量对模型准确率的影响。由于存在隐变量，而且数据是高度非线性的，因此 LiNGAM 等线性方法都不再适用，使得该类方法的准确率不管在什么条件下都很低。

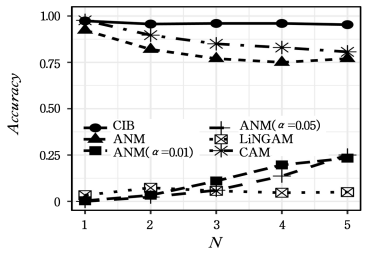

图 4　针对隐变量数量的灵敏度分析

Fig. 4　Sensitivity of number of latent variables

对于 CAM 和 ANM 方法，其在隐变量数量 $N=1$ 时，准确率达到最高。然而，随着隐变量的数量增加，其准确率一直下降。特别地，在 $N=2$ 时，准确率已下降了近 20%，原因是隐变量的数量增加破坏了原模型假设，从而使得因果关系无法被识别。此外，我们还发现，ANM 方法使用 p 值作为准则时准确率一直很低，这意味着在实际应用中，ANM 在有隐变量的情况下常常无法识别因果关系。

对于 CIB 方法，在实验中，隐变量的增加会使得 CIB 的准确率轻微下降，但下降程度较小，这从侧面体现出了 CIB 模型的健壮性，同时也验证了 CIB 的有效性，即便存在隐变量，本文的因果关系挖掘算法仍然能很好地识别因果关系。

如图 5 所示，我们测试了在不同样本大小下，不同模型准确率的变化情况。显然，在不同的样本大小下，CIB 方法均取得了最优的效果。可以看到，随着样本量的增加，CIB 的效果会有轻微改善，但是，其对样本量的依赖性并不大，即使在样本量很少时，CIB 仍然能够达到很高的准确度。而对于其他方法，尤其是 CAM 和 ANM 而言，其对样本量的敏感程度更大，特别是 ANM 模型，它对数据量的依赖是最大的，其提升幅度也是最大的，不过即使在足够大的数据量下，其准确率也不如 CIB 和 CAM 等方法。对于 LiNGAM 方法，由于在该实验中存在隐变量 $N=2$ 并且数据的产生是非线性的，因此该方法出现了无法识别的情况。

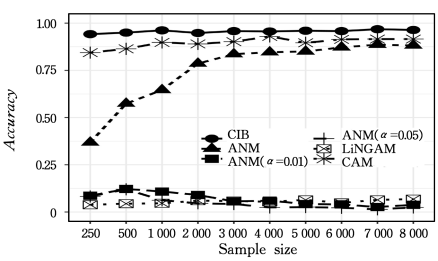

图 5　针对样本大小的灵敏度分析

Fig. 5　Sensitivity of number of sample size

如图 6 所示，我们进一步探索在存在未观测的混淆因子（混淆因子同时是 X，Y 的原因）的情况下模型的准确率。结果显示，CIB 仍然比其他方法的效果好，并且随着混淆因子的增加，其他方法的准确率会下降，而 CIB 几乎不受影响。

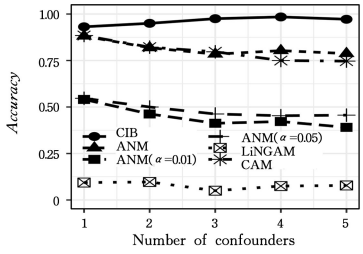

图 6 针对混淆因子的灵敏度分析

Fig. 6 Sensitivity of number of confounders

综上，合成数据实验证明了 CIB 在各种条件下的有效性和稳健性。

5.2 真实数据实验

本节将使用因果关系基准数据集[25]1)，并对其中的 100 组因果对进行因果识别测试。该数据集中包括各种常见场景下的因果关系数据，如海拔对温度的影响、年龄对身高的影响等。如图 7 所示，我们用散点图的形式给出了 100 对真实因果对。从图 7 可以看到，有些真实因果对呈现的因果机制是复杂的，有可能不满足我们的模型假设。由于真实数据下的复杂性，该基准数据集目前仍然充满挑战。

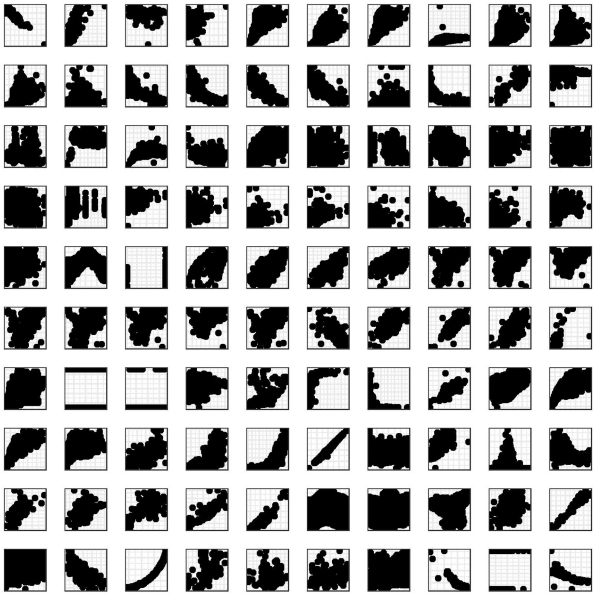

图 7 真实因果对数据集散点图

Fig. 7 Scatter plot of real-world causal pairs

在进行实验前，我们先对每组数据对进行标准化预处理，即：

$$\hat{X}=\frac{X-mean(X)}{std(X)}$$

$$\hat{Y}=\frac{Y-mean(Y)}{std(Y)}$$

其中，$\hat{X}$，$\hat{Y}$ 是经过标准化后的变量。使用经过标准化后的数据，在这 100 组因果对上，对所有方法进行测试，测试结果将以这 100 组因果对的准确率作为评价指标，具体结果如图 8 所示。可以看到，本文方法（CIB）取得了最高的准确率（66%），而 ANM 的准确率是 62%，CAM 的准确率是 55%，LiNGAM 的准确率是 36%。真实数据上的实验表明，CIB 模型在真实数据中有着更好的效果，同时也验证了该模型的有效性。综上，CIB 在合成数据与真实数据中均提升了准确率。

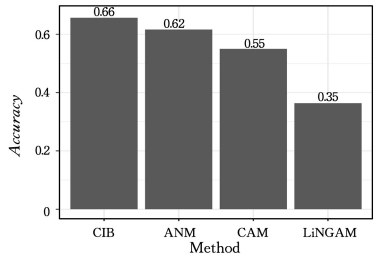

图 8 真实因果对的识别准确率结果

Fig. 8 Identification accuracy of real-world causal pairs

结束语 本文针对因果函数模型的假设在存在隐变量时常常不成立等问题，提出了一种基于信息瓶颈的因果方向推断方法。该方法通过引入压缩-拟合两阶段的因果模型，利用信息瓶颈的特性来对该因果模型建模与求解，以压缩原因变量与提取结果变量的有效信息。我们指出因果信息瓶颈准则可作为因果方向挖掘的目标函数，并给出了一种基于变分上界的近似方法。针对该上界，我们使用变分自编码机来进行优化。实验结果表明，不管在合成数据还是在真实数据中，基于因果信息瓶颈的因果推断方法在准确率上都取得了较大的提升。未来，我们计划考虑存在混淆因子的情况，从而进一步提升模型的适用场景。

参 考 文 献

[1] MCINERNEY J, BROST B, CHANDAR P, et al. Counterfactual Evaluation of Slate Recommendations with Sequential Reward Interactions[C]// The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: ACM, 2020: 1779-1788.

[2] RUNGE J, BATHIANY S, BOLLT E, et al. Inferring causation from time series in Earth system sciences[J]. Nature Communications, 2019, 10(1): 1-13.

[3] CAI R, ZHANG Z, HAO Z, et al. Understanding social causalities behind human action sequences[J]. IEEE Transactions on Neural Networks and Learning Systems, 2016, 28(8): 1801-1813.

[4] WANG W J, DU X H, REN Z Y, et al. Reconstruction of Cloud Platform Attack Scenario Based on Causal Knowledge and Temporal-Spatial Correlation[J]. Computer Science, 48(2): 317-323.

[5] 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.

[6] XIE F, CAI R, HUANG B, et al. Generalized Independent Noise

1) https://webdav.tuebingen.mpg.de/cause-effect/

Condition for Estimating Latent Variable Causal Graphs[C]// Advances in Neural Information Processing Systems. New York: Curran Associates, Inc., 2020: 14891-14902.

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

[8] 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(Oct): 2003-2030.

[9] SHIMIZU S, INAZUMI T, SOGAWA Y, et al. DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model[J]. Journal of Machine Learning Research, 2011, 12(Apr): 1225-1248.

[10] HOYER P, JANZING D, MOOIJ J M, et al. Nonlinear causal discovery with additive noise models[C]// Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems. New York: NIPS, 2008: 689-696.

[11] ZHANG K, HYVÄRINEN A. On the Identifiability of the Post-Nonlinear Causal Model[C]// UAI 2009, Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence. Corvallis, USA: AUAI Press, 2009: 647-655.

[12] CAI R, QIAO J, ZHANG K, et al. Causal discovery from discrete data using hidden compact representation[C]// Advances in Neural Information Processing Systems. Calfornia, USA: NIPS, 2018: 2666-2674.

[13] CAI R, QIAO J, ZHANG K, et al. Causal discovery with cascade nonlinear additive noise models[C]// Proceedings of the 28th International Joint Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2019: 1609-1615.

[14] HUANG YL, LI P F, ZHU Q M. Joint Model of Events' Causal and Temporal Relations Identification[J]. Computer Science, 2018, 45(6): 204-207, 234.

[15] SPIRTES P, GLYMOUR C N, SCHEINES R. Causation, prediction, and search[M]. USA: MIT press, 2000.

[16] 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.

[17] ANDERSSON S A, MADIGAN D, PERLMAN M D, et al. A characterization of Markov equivalence classes for acyclic digraphs[J]. The Annals of Statistics, Institute of Mathematical Statistics, 1997, 25(2): 505-541.

[18] SLONIM N, FRIEDMAN N, TISHBY N. Multivariate information bottleneck[J]. Neural Computation, 2006, 18(8): 1739-1789.

[19] ALEMI A A, FISCHER I, DILLON J V, et al. Deep Variational Information Bottleneck[C]// the 5th International Conference on Learning Representations. 2017.

[20] KINGMA D P, WELLING M. Auto-Encoding Variational Bayes [C]// the 2nd International Conference on Learning Representations. 2014.

[21] HORNIK K, STINCHCOMBE M B, WHITE H. Multilayer feedforward networks are universal approximators[J]. Neural Networks, 1989, 2(5): 359-366.

[22] KINGMA D P, BA J. Adam: A Method for Stochastic Optimization[C]// the 3rd International Conference on Learning Representations. 2014.

[23] BÜHLMANN P, PETERS J, ERNEST J, et al. CAM: Causal additive models, high-dimensional order search and penalized regression[J]. The Annals of Statistics, Institute of Mathematical Statistics, 2014, 42(6): 2526-2556.

[24] GRETTON A, BOUSQUET O, SMOLA A J, et al. Measuring Statistical Dependence with Hilbert-Schmidt Norms[C]// Algorithmic Learning Theory, 16th International Conference. Berlin, German: Springer, 2005: 63-77.

[25] MOOIJ J M, PETERS J, JANZING D, et al. Distinguishing cause from effect using observational data: methods and benchmarks[J]. The Journal of Machine Learning Research, 2016, 17(1): 1103-1204.

QIAO Jie, born in 1993, Ph.D student. His main research interests include machine learning and causality.

CAI Rui-chu, born in 1983, Ph.D, professor, Ph.D supervisor. His main research interests include artificial intellectual and causality.

（责任编辑：李亚辉）