基于非时序观察数据的因果关系发现综述

蔡瑞初[1] 陈 薇[1] 张 坤[2] 郝志峰[1),3)]

[1)](广东工业大学计算机学院 广州 510006)

[2)](卡内基梅隆大学哲学系 匹兹堡 美国 15213)

[3)](佛山科学技术学院数学与大数据学院 广东 佛山 528000)

摘 要 探索和发现事物间的因果关系是数据科学的一个核心问题,其中蕴含着丰富的科学发现机会和巨大的商业价值.基于非时序观察数据的因果关系发现方法能够从被动观察获得的数据中发现变量之间的因果关系,因而在各领域有广泛应用.这一类方法在过去三十年取得很大进展,已经成为因果关系发现的重要途径.文中从因果关系方向推断、高维数据上的误发现率控制和不完全观察数据上的隐变量检测这三个研究热点出发,对现有的因果关系模型与假设、基于约束的方法、基于因果函数模型的方法和混合型方法这三大类方法,验证与测评涉及的数据集及工具等方面进行了详尽的介绍与分析.基于约束的方法主要包括因果骨架学习和因果方向推断两个阶段:首先基于因果马尔可夫假设,采用条件独立性检验学习变量之间的因果骨架,然后基于奥卡姆剃刀准则利用 V-结构确定因果方向,典型的算法有 Peter-Clark 算法、Inductive Causation 等,这类方法的主要不足是存在部分无法判断的因果关系方向,即存在 Markov 等价类难题.基于因果函数模型的方法则基于数据的因果产生机制假设,在构建变量之间的因果函数模型的基础之上,基于噪声的非高斯性、原因变量与噪声的独立性、原因变量分布与因果函数梯度的独立性等因果假设推断变量之间的因果关系方向,典型的算法有针对线性非高斯无环数据的 Linear Non-Gaussian Acyclic Model 算法、针对后非线性数据的 Post-NonLinear 算法、适用于非线性或离散数据的 Additive Noise Model 等,这类方法的主要不足是需要较为严格的数据因果机制假设,且 Additive Noise Model 等方法主要适用于低维数据场景.混合型方法则希望充分发挥基于约束的方法和基于因果函数类方法的优势,分别采用基于约束的方法进行全局结构学习和基于因果函数模型进行局部结构学习和方向推断,典型的算法有 SADA、MCDSL 等,理论分析较为不足是这类方法目前遇到的主要困难.最后,文中还基于研究现状分析讨论了因果方向推断、高维数据上的误发现率控制、隐变量发现、与机器学习的关系等未来可能的研究方向.

关键词 因果关系;因果关系发现;观察数据;结构学习;加性噪声模型;人工智能;机器学习

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2017.01470

A Survey on Non-Temporal Series Observational Data Based Causal Discovery

CAI Rui-Chu[1] CHEN Wei[1] ZHANG Kun[2] HAO Zhi-Feng[1),3)]

[1)](*School of Computer Science, Guangdong University of Technology, Guangzhou* 510006)

[2)](*Department of Philosophy, Carnegie Mellon University, Pittsburgh* 15213)

[3)](*School of Mathematics and Big Data, Foshan University, Foshan, Guangdong* 528000)

Abstract Exploring and detecting the causal relations among variables have shown huge practical values in recent years, with numerous opportunities for scientific discovery, and have been commonly seen as the core of data science. Among all possible causal discovery methods, the approaches to causal discovery from non-temporal observational data can recover the causal structures from passive observational data in general cases, and have shown extensive application

收稿日期:2016-04-16;在线出版日期:2016-12-26.本课题得到 NSFC-广东联合基金(U1501254)、国家自然科学基金(61572143)、广东省杰出青年科学基金(2014A030306004)资助.**蔡瑞初**,男,1983 年生,博士,教授,中国计算机学会(CCF)高级会员,主要研究领域为因果关系、机器学习等.E-mail:cairuichu@gmail.com.**陈 薇**,女,1993 年生,硕士,中国计算机学会(CCF)会员,主要研究方向为因果关系及其应用.**张 坤**,男,1980 年生,博士,助理教授,主要研究方向为因果关系、机器学习等.**郝志峰**,男,1968 年生,博士,教授,主要研究领域为机器学习、数据挖掘.

版权所有,未经许可,不得转载

prospects in a lot of real world applications. After 30 years' rapid progress, causal discovery from non-temporal observational data have been considered as an important research direction of causal discovery. In this survey, we discuss three hot research topics including causal direction inference, false discovery rate control on high-dimensional data, and latent variable detection in partially observational data. Around the above research topics, we extensively review and analyze recent achievements in several aspects of causal discovery, especially focusing on causal models and their basic assumptions, constraint based approaches, casual function based approaches, hybrid approaches, and the related benchmarks and tools. A typical constraint based approach is a two-phase method, firstly utilize the conditional independence tests to learn the causal skeleton based on the Causality Markov Assumption, and then use the V-structures to determine the causal directions based on Occam's razor principle. The typical constraint based algorithms include Peter-Clark (PC) algorithm and Inductive Causation (IC) algorithm. The main limitation of this class of methods is that they cannot distinguish the underlying causal structure from its statistically equivalent structure, i. e., the algorithms return some undetermined causal directions. This limitation is also known as Markov equivalence class problem. The casual function based approaches are based on data generating process assumptions. After fitting the causal function model among the variables, the causal function based approach inference the causal direction by employing the causal assumptions, such as non-Gaussian assumption of the noise, the independence assumption between causal variable and noise, and the independence assumption between the distribution of causal variable and the causal function. The typical causal function based approach includes Linear Non-Gaussian Acyclic Model (LiNGAM), Post-NonLinear (PNL) and Additive Noise Model (ANM). The disadvantage of the causal function based approach is the strict assumptions of the data generation process, which are usually hard to hold in real world applications. Its difficulty on high dimensional data is another disadvantage of this approach. The hybrid approaches try to take full advantage of the constraint based approaches and the causal function based approaches, using constraint based approaches to learn global causal structure and employing causal function based approaches to learn the local causal structure and infer the causal directions respectively, typical algorithms include SADA (Scalable cAusation Discovery Algorithm) and MCDSL (Multiple-Cause Discovery method combined with Structure Learning). The lack of theoretical analysis is the main difficulty of these methods. Based on the review and analysis of the existing works, we also give an outlook of some future research directions, include causal direction inference, false discovery rate control on high-dimensional data, latent variable detection in partially observational data, the relation between causality and machine learning and so on.

Keywords　causality; causal discovery; observational data; structure learning; additive noise model; artificiall intelligence; machine learning

1　引　言

互联网、生命科学、经济学等领域积累的海量数据蕴含着巨大的商业价值和极其丰富的科学发现机会，有效探索和利用这些数据已经成为各个领域的迫切需求[1]. 在经济领域，已有研究报告表明利用数据科学指导决策可以提高至少 5%的生产效率[2]. 在科学研究领域，基于"将数据放入计算机集群，让统计算法去发现科学家不能发现的规律"思想的数据密集型科学发现已经成为科学研究的第四范式[3].

如何从这些海量数据中发现有意义的关系，尤其是因果关系，是数据科学中最有可能创造商业价值和进行科学发现的研究领域之一，正受到国际同

行的广泛关注①[4]. 因果关系严格区分了原因变量和结果变量,在揭示事物发生机制、指导干预行为等方面有相关关系不能替代的重要作用[5]. 以图 1 为例,吸烟、黄牙都与肺癌具有较强的相关关系,然而只有吸烟才是肺癌的原因,也只有戒烟才能降低肺癌的发病概率,而把牙齿洗白则不能降低肺癌的发病概率.

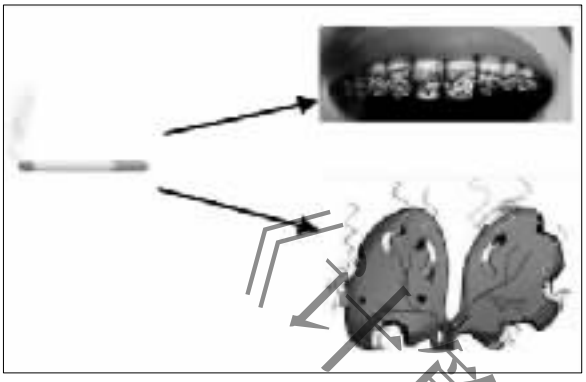


图 1 吸烟与黄牙、肺癌之间的因果关系发现

随机控制实验是发现因果关系的传统途径[5],但是由于实验技术的局限性,绝大部分场合只能进行被动式观察,而无法进行主动式干预[6]. 比如,我们不能让病人使用未经验证的药物,也难以操控人的基因表达水平,而在社交网络上进行随机实验代价巨大. 从观察数据上进行有效的因果关系发现避免了以上限制,而且有可能给出从因到果的函数模型,因而具有重要的应用价值,是当前因果关系发现领域的研究热点[7-8].

针对观察数据特性的不同,基于观察数据的因果关系发现方法可以分为基于时序观察数据的因果关系发现方法和基于非时序观察数据的因果关系发现方法. 虽然时序观察数据中时间维度蕴含了"因-果"方向的重要信息——"果"在时间上不能发生在"因"的前面,但是时序数据需要获取一个对象在不同时刻的观察值,对观察手段具有较高的要求. 例如,现有的基因表达数据测量方法会破坏观察样本,使得我们无法获取该样本的下一时刻的状态. 更进一步,一般来说,基于时序数据来发现因果关系的结果对数据采集的频率等因素很敏感[9-10]. 所以基于非时序观察数据的因果关系发现具有更广的适用范围,也是当前因果关系发现领域的研究热点. 因此,本文主要对非时序观察数据上的因果关系发现方法进行介绍. 该问题形式化的定义如下:

基于非时序观察数据的因果关系发现:给定 p 维变量集 $\boldsymbol{V}=\{v_1,v_2,\cdots,v_p\}$ 上的 m 组非时序观察数据 $X=\{x_1,x_2,\cdots,x_m\}$,发现变量 $\{v_1,v_2,\cdots,v_p\}$ 间的因果关系.

在适用范围方面,基于非时序观察数据的因果关系发现在相关研究领域的位置如图 2 所示. 与干预实验、时序观察数据等数据获取方法相比,非时序观察数据的获得代价最低,因此基于非时序观察数据因果关系发现的适用范围也最广.

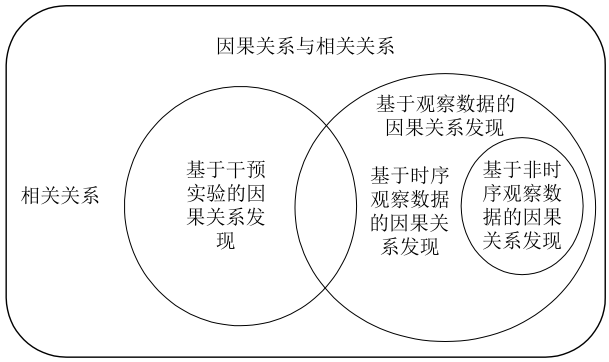


图 2 本文研究内容

根据实际问题的不同,基于非时序观察数据的因果关系发现研究一般至少包含下面几个子问题:因果关系方向推断、高维数据上的误发现率控制和不完全观察数据上的隐变量检测. 其中,因果关系方向推断的基本设定是针对两个变量 $\{v_1,v_2\}$ 进行,假设 v_1 和 v_2 之间有直接的因果连接,任务是通过观察数据上呈现的不对称性,推断变量间的因果方向,即区分 $v_1\rightarrow v_2$ 和 $v_1\leftarrow v_2$. 高维数据上的因果关系结构学习则是针对多个变量 $\{v_1,v_2,\cdots,v_p\}$ $(p\geq 2)$ 进行,在因果关系方向推断基础上,剔除冗余的间接因果关系,构建多个变量间的因果关系网络. 最后,不完全观察数据上的隐变量检测则是针对定义在 $\{v_1,v_2,\cdots,v_p\}$ 上的观察数据,检测隐变量 v_c $(v_c\notin\{v_1,v_2,\cdots,v_p\})$ 的存在性. 上述隐变量是指未能观察或无法度量的变量,一般体现为混淆因子(Confounder Variables)和选择偏倚(Selection Bias)两种类型[11].

围绕着因果关系方向推断、高维数据上的误发现率控制和不完全观察数据上的隐变量检测这 3 个研究热点,下面本文将从假设与模型、主要发现算法及未来发展方向等方面对现有的因果关系发现算法进行总结介绍,主要组织结构如下:第 2 节阐述因果关系模型与假设;第 3 节详细介绍了该领域的国内外发展现状、分析并总结了现有研究主要的技术难点和挑战,介绍了一些验证与测评涉及的数据集及

① ACM. Pearl Judea. http://amturing.acm.org/award_winners/pearl_2658896.cfm, 2011

工具；第 4 节会对基于非时序观察数据的因果关系发现未来研究方向做展望；第 5 节对本文内容进行总结.

2　假设与模型

2.1　基本假设

因果假设是众多因果关系发现算法的基础之一. 其中，Simon[12] 指出，从数据中决定因果结构的问题未被严格约束，感知到的因果结构依赖于我们为其设定的先验假设.

我们在这篇文章里考虑用有向无环图(Directed Acyclic Graphs，DAG)来描述的因果关系. 图 3(a)给出了有向无环图的一个例子，令每个节点表示一个随机变量. 若从一个变量到另一个变量有一条有向的边，那么相对所考虑的变量集，前者是后者的一个直接的因.

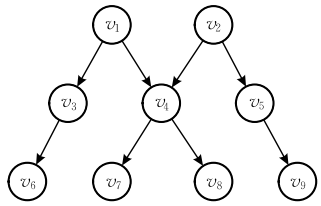
目前，因果关系发现研究的主流假设有以下4 种：因果充分性假设(Causal Sufficiency Assumption)、因果马尔可夫假设(Causal Markov Assumption)、因果忠诚性假设(Causal Faithfulness Assumption)和数据产生方式假设[11]. 不同的方法需要以上假设中的一个或多个.

(1) 因果充分性假设(Causal Sufficiency Assumption). 当变量集 $\boldsymbol{V}$ 中的任意两个变量的直接原因变量都存在 $\boldsymbol{V}$ 中时，变量集 $\boldsymbol{V}$ 就被认为是因果充分的[13].

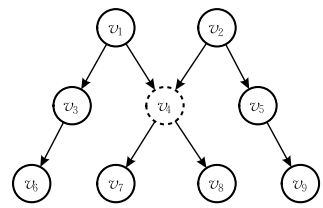
直观上讲，如果变量的观察集不具有因果充分性，那么这个因果模型可能含有未观测到的共同的因变量. 例如，图 3(a)为定义在变量集 $\boldsymbol{V}=\{v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_8, v_9\}$ 上的完整的因果关系网络图. 当观察到的变量集为 $\boldsymbol{V}'=\{v_1, v_2, v_3, v_5, v_6, v_7, v_8, v_9\}$ 时，v_7 和 v_8 的直接原因 v_4 不在观察到的变量 $\boldsymbol{V}'$ 中，此时 $\boldsymbol{V}'$ 不满足因果充分性假设. 因果充分性假设等价于外生变量(Exogenous variable)的独立性假设[14].

(2) 因果马尔可夫假设(Causal Markov Assumption). 对于具有因果充分性的变量集而言，在已知变量的父亲节点的条件下，如果所有变量与他们的非后裔节点互相条件独立，那么我们把这种情况称之为满足因果马尔可夫假设[11].

在图 3(a)所示的例子中，基于因果马尔可夫假设，已知 v_4 的父亲节点 v_1 时，v_4 与所有的非后裔节点(包括 v_3 和 v_6)条件独立. 类似的，给定 v_5 的父亲节点 v_2 时，v_5 与所有的非后裔节点(包括 v_4、v_7 和 v_8)条件独立，而与其余后裔节点 v_9 则不一定独立.



(a) 当观察到的变量集 $\boldsymbol{V}=\{v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_8\}$ 时，$\boldsymbol{V}$ 满足因果充分性假设



(b) 当观察到的变量集 $\boldsymbol{V}'=\{v_1, v_2, v_3, v_5, v_6, v_7, v_8\}$ 时，$\boldsymbol{V}'$ 不满足因果充分性假设

图 3　因果充分性假设的因果关系网络图示例

因果马尔可夫假设只是假设所有变量间的关系都是由因果关系引起的，其存在过简单化的问题. 因此有一些与之相关的方法应运而生. 第一是在共同后代的条件下可以产生条件依赖. 第二是变量间的逻辑关系能产生非因果相关性. 如一个商店一年的销量与四个季度的销量有关，但是并不是由这四个季度的销量所引起的. 第三，它没有处理及时对称相互作用的方法. 如经典的重力引论[13].

(3) 因果忠诚性假设(Causal Faithfulness Assumption). 如果在给定变量集 $\boldsymbol{V}$ 的前提下，变量 v_i 和 v_j 互相独立或条件独立，那么在由变量及其之间因果依赖关系组成的因果关系网络图 G 中，v_i 和 v_j 之间的所有路径被变量集 $\boldsymbol{V}$ 中合适的变量 d-分离(d-separation)[5](示意图如图 4 所示)，则称所有随机变量的联合分布 P 与图 G 是因果忠诚的[11].

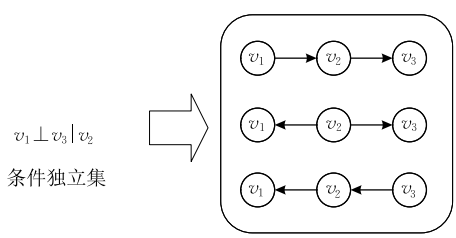


图 4　v_1 和 v_3 被 v_2 d-分离示意

因果忠诚性假设的隐含意义是在因果关系发现的过程中，变量之间不会出现额外的(条件)独立关系. 在忠诚性假设下，模型不仅包含定义在变量或变量集上的结构方程，而且在实际情况下，真实的函数

形式和系数的真实值没有额外隐含的约束[14].

为了弱化因果忠诚性假设,研究者提出了一些相对弱化的衍生假设,如邻接忠诚性(Adjacency-Faithfulness)、方向忠诚性(Orientation-Faithfulness)和三角忠诚性(Triangle-Faithfulness)等[15-18].

(4)数据产生方式假设.与上述因果充分性假设、因果马尔可夫假设和因果忠诚性假设这三大假设不同,数据产生方式假设是为了提升因果关系发现能力,在最近的一系列因果关系方法中新引入的一类假设,如数据类型的离散/连续、产生机制的线性/非线性/后非线性、噪声数据为无噪声或者分布为/高斯/非高斯分布等.基于上述细化的数据产生方式假设,可以在先验知识缺乏的条件下,利用"数据驱动"的方法提升模型的发现能力[19].

2.2 因果关系模型及其隐含假设

因果关系模型的建立一般是基于其隐含的假设的.经典的因果关系模型主要为 Rubin 的因果关系模型(Rubin Causal Model)[20]和 Pearl 以及 Spites,Glymour,Scheines 等人的因果图模型(Causal Diagram Model)[21].Pearl 教授结合结构方程模型证明了两者的等价性[5],但是,因果图模型更适合表示高维数据上的全局因果结构,是众多结构模型的基础.因果图模型主要是通过基于贝叶斯结构学习的方法进行构建,具有代表性的因果关系结构图推断方法就是通过探测变量之间的条件依赖,基于关于变量联合分布的因果马尔可夫假设和因果忠诚性假设来实现的[5].由于模型假设的性质,这类方法存在着马尔可夫等价类难题[19]——它的结果无法区分属于同一个等价类内的因果关系结构,因为这些结构共享着变量间同样的(条件)独立性.

与因果图模型中隐含的假设相比,结构方程模型(Structure Equation Model,SEM)[22-23]是基于数据产生方式假设的,其可以对数据产生方式进行更加丰富的刻画,是一些具有较强发现能力的因果关系发现算法的模型基础.结构方程模型中典型的数据产生方式假设包括:Shimizu 等人[24]的基于线性非高斯噪声的模型(Linear Non-Gaussian Acyclic Model,LiNGAM)、Zhang 等人[25-26]的后非线性模型(Post-NonLinear,PNL)、Hoyer 等人[27]的基于非线性加性噪声模型(Additive Noise Model,ANM)、Daniusis 和 Janzing 等人[28-29]提出的无噪声数据产生模型、Peters 等人[30]的离散数据加性噪声模型.

因果关系模型中隐含的假设,可以大致归纳为表 1,但在具体算法设计中又会存在其特殊性.

表 1 因果关系模型中隐含的假设

模型	假设
因果图模型	因果马尔可夫假设 因果忠诚性假设
结构方程模型	因果充分性假设 数据产生方式假设

3 研究现状分析

本部分将对基于约束的方法、基于因果函数模型的方法和混合型方法这三大类非时序观察数据的因果关系发现算法进行介绍和分析.上述三类方法研究基本上围绕着因果关系方向推断难题、高维数据上的误发现率控制难题和不完全观察数据上的隐变量检测难题而展开,具体如表 2 所示.

表 2 基于非时序观察数据的因果关系发现方法相关文献

因果关系发现方法	因果关系方向推断难题	高维数据上的误发现率控制难题	不完全观察数据上的隐变量检测难题
基于约束的方法	PC 算法[11],IC 算法[31]	小样本理论[32],主动学习方法[33-35],MMHC[36],MMPC[37]	FCI[11,38],RFCI[39],FTFC[40]
基于因果函数模型的方法	LiNGAM[24],PNL[25-26],ANM[27,30],IGCI[28-29],DirectLiNGAM[41]	RESIT[42]	ParceLiNGAM[43],lvLiNGAM[44],IGPLVM[45],GPLVM[46],信息不等式[47-48]
混合型方法		SADA[49],MCDSL[50],HYA[51]	

3.1 基于约束的方法

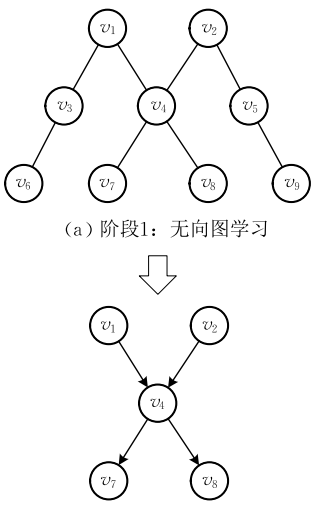
20 世纪 80 年代以来,Glymour,Scheines,Spirtes,Pearl 和其他一些建立因果关系挖掘先驱工作的研究者,就从统计学和哲学的角度出发,研究基于非时序观察数据的因果关系发现方法[11,52].其采用的基本工具为在贝叶斯网络基础上,加上因果关系解释,而衍生出来的因果网络.因果网络是用来表示变量之间连接概率的图模型,可用于描述多个变量之间相互的因果关系.

从数据中学习贝叶斯因果网络主要采用基于约束(Constraint-based)的方法.还有一类方法采用评分函数和搜索算法选择最优的贝叶斯因果网络[23,53-54],虽然这类方法有时也会涉及,但是其往往涉及图搜索过程,时间复杂度较高,此处不做重点介绍.基于约束的方法是通过数据中变量间的条件独立性来判断特定结构的存在性.这种测试通常用统计或信息论的

度量来实现[14,55]. 因此，基于约束的方法也被称为基于条件独立性的方法. 其中最基本的算法有 PC(Peter-Clark)算法[11]和 IC(Inductive Causation)算法[31].

PC[11]和 IC[31]这两类方法的基本流程可以归纳为图 5 所示的两阶段过程. 在无向图学习阶段，从完全连通图出发，基于独立性或条件独立性假设检验等统计方法给出的变量之间的独立性而砍掉相应的边，从而获得变量间的无向图；在方向学习阶段，则依赖于 V-结构(V-Structure)等局部结构特性确定部分边的方向. 后续算法相关工作主要集中在如何有效地学习变量间的无向图和如何有效推断边的方向.

PC 算法[11]的前身是 SGS 算法[11]，相较于 SGS 算法，它在无向图确定过程中搜索条件变量子集的基础上进行了改进，在发现高维稀疏连接的因果关系结构上取得较好的效果. 这两种方法能确定两个变量之间是否有边相连，最终学习到整体的因果关系结构图，而且他们都是从完全无向图出发学习网络结构，具有一定的可靠性，但是这两种算法在网络结构学习的复杂度很高，算法效果依赖于变量处理顺序.

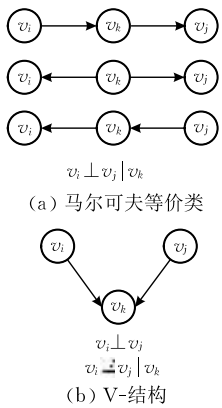


(a) 阶段1：无向图学习

(b) 阶段2：有向边学习

图 5　基于约束的方法两阶段示意图

IC 算法[31]由 Verma 和 Pearl 最早提出. 算法的主要思路是先通过学习得到一个无向图，再确定无向图中的所有 V-结构(V-Structure)，最终尽量确定其他无向边的箭头方向，得到因果关系结构图. 文献[11,56]提出在基数递增的顺序下寻找 d-分离的一种系统方法. V-结构是概率图模型中一种特殊的形式，包含 3 个变量 v_i、v_j 和 v_k，其形式如图 6(b)所示. V-结构不同于马尔可夫等价类中的 3 种形式，所以它在因果识别中更具有可识别性. 在因果发现方法中起着重要的角色[5]. Cai 等人[57]通过探索显著性 V-结构的结合，将基于 V-结构的组合搜索方法应用于因果基因识别的场景中，得到了一个优化的算法.



(a) 马尔可夫等价类

(b) V-结构

图 6　三元变量之间的因果关系(其中(a)为马尔可夫等价类且三元变量之间的关系均满足在变量 v_k 的条件下，变量 v_i 和变量 v_j 互相独立；图(b)为 V-结构，三元变量之间的关系满足变量 v_i 和变量 v_j 互相独立，但在变量 v_k 的条件下，变量 v_i 和变量 v_j 不独立)

由于基于约束的方法无法给出马尔可夫等价类中边的方向，因此因果图中马尔可夫等价类的数量直接决定基于约束方法的应用效果. 为此，Chickering[58]、Andersson 等人[59]在该方面进行了较多理论工作. Chickering[58]通过研究发现，当大部分马尔可夫等价类数量很多时，在马尔可夫等价类模型中进行因果结构学习会比在整个有向无环图模型中学习得到的结果更有效. 针对上述问题，Andersson 等人[59]用无向连接弦图(Undirected Connect Chordal Graph，UCCG)表示马尔可夫等价类的数量，但是在有向图的节点数变得很多时，其结果变得不可靠. 因此，He 等人[60]对此进行了改进，引入根子类(rooted sub-class)的概念，将马尔可夫等价类划分为更小的子类后，再分别计算马尔可夫等价类的数量，同时还探索了马尔可夫等价类的数量和边的分布，得出一般情况下的结论.

可靠的条件独立性假设是这类方法的核心. 传统的很多工作假设数据是离散的或者是线性高斯的，这两种情况下条件独立假设可简单地通过卡方检验或者偏相关(partial correlation)检验来实现. 除了传统的条件独立性测试方法，有一些研究使用了基于核的因果学习方法来判断变量间的条件独立性. 如 Sun 等人[61]使用 Kernel based Hilbert-Schmidt Norms(HSN)度量两个变量间的条件依赖性；Zhang 等人[62]提出非独立同分布数据的独立性核度量方法来发现因果关系结构；为了发现大数据集下的条件独立性，Zhang 等人[63]提出了一种基

于核的条件独立性测试(Kernel-base Conditional Independence test,KCI),在实现条件独立性测试上更加高效和简单.为提升假设检验方法在小样本数据上的可靠性,Bromberg 等人[32]提出了论证独立性测试(Argumentative Independence Test),通过识别条件独立性测试后的结果之间存在的相互独立性,提升小样本集下的统计独立性测试的可靠性.但是这个方法需要假设不同统计测试中的随机变量是互相独立的,当前对这类方法的研究还正处于初步阶段.

以上算法都是基于马尔可夫假设的,虽然能发现因果结构,但是这些算法需要结合某些规则来推断因果关系分析变量间的条件独立性关系,没有充分考虑到变量的数据分布和变量间因果机制所导致的数据分布的一些特定性质.在出现马尔可夫等价类时,这些算法无法对其进行正确的推断,只能发现部分可能的因果关系,也就是说,在出现"给定变量 v_k 的前提下,v_i 和 v_j 互相独立",我们无法推断出变量 v_i、v_j 和 v_k 之间的关系属于图 6(a)中的哪一种,换句话说,我们只能在一定程度上构造出无向图,无法推断边的方向.

为了降低高维数据上的误发现率,Geng 和 Xie 等人[64-66]提出了一种搜索 d-分离的分解方法,可以通过递归方法将一个图分解为两个子图,学习局部网络结构,并逐步自底向上整合成全局的因果关系结构图.这类方法虽然能有效地提高算法的效率,但是因为分解过程需要先验知识或由数据检验分解的有效性,且由于观察数据中包含的因果关系信息匮乏,较难得到可靠的因果关系.针对这样的问题,文献[33-35]提出因果网络定向的主动学习方法.这种方法的核心思想在于抓住主要变量进行干预实验,从一个目标结点出发,逐步进行局部变量选择和局部网络结构学习,最终确定并能区分该目标节点的原因与结果.这个方法不必直接构造高维变量的完整因果网络,从而部分克服了数据高维带来的挑战.Tsamardinos 等人结合了基于约束的方法和贪婪等价类搜索(Greedy Equivalent Search)方法[58]等,提出最大-最小爬山法(Max-Min Hill-Climbing,MMHC)[36].这种方法先通过局部结构学习算法——最大-最小父亲孩子(Max-Min Parents and Children,MMPC)算法[37]学习因果无向图,然后用贪婪贝叶斯评分爬山搜索方法对无向图进行定向.这种方法不仅适用于高维数据,而且提高了因果贝叶斯网络学习的有效性.

当数据集中的变量不满足因果充分性时,那么这个数据集中就可能含有不可观测的共同直接因变量.鉴于线性图模型蕴含着多种协方差矩阵的子矩阵的排序约束,Kummerfeld 等人[40]利用这些排序约束,再加上条件独立性测试,提出了一种 FTFC(Find Two Factor Clusters)算法,用于识别隐变量模型.后续,Spirtes 等人在 PC 类算法基础上,提出了 FCI(Fast Causal Inference)算法[11,38]和 Colombo 等人对 FCI 进行改进的 RFCI(Really Fast Causal Inference)算法[39]等,主要使用最大祖先图来实现含隐变量的因果关系发现.Zhang[67]进而提出了一些附加的定向规则,增强了在混淆因子和选择偏倚的情况下基于最大祖先图的因果关系发现方法的能力.Zhao 等人[68]则对 FCI 算法加以改进,提出基于最大祖先图和马尔可夫等价类的含混淆因子和选择偏倚的检测方法.Zhang 等人[45]扩展了 Gaussian-Process Latent Variable Model(GPLVM)[46]来处理隐含的原因变量对观察变量的作用是非线性的,而观察变量之间的作用是线性的情况.与前几种方法的研究角度不同的是,Pearl 等人[69]利用由工具变量(instrumental variable)引起的不等式约束,即外生变量直接影响某些变量而非全部变量的准则,来测试一个涉及工具变量的模型是否存在隐变量.Bonet 等人[70]将这种工具不等式扩展到了二元变量的情况.Glymour 等人[71]则将工具不等式作为一个更加泛化的隐变量结构中可测试约束现象的特殊情况.基于此,Evans 等人[72]提出以图形约束为主的方法,采用边缘有向无环图(marginalized directed acyclic graphs)和系统树模型(Phylogenetic Trees Model)进行变量因果性评估,进而发现隐变量.

由上述分析可知,基于约束的方法在框架上是无参的,能在给定可靠的条件独立性测试下得到广泛的应用(当然,它在实际操作中的效果依赖于条件独立检验的手段),同时,它受到等价类的约束,能从条件独立性集合中发现因果关系.但是,它的缺陷在于需要附加合理的假设,否则无法直接描述和发现因果关系;通过条件独立性测试和 V-结构为主的结构学习方法,无法学习到因果关系网络图中所有边的方向,只能得到一组马尔可夫等价类的有向无环图.因此,如何克服基于约束的方法存在的马尔可夫等价类难题从而得到唯一的因果描述结构是基于因果函数模型等方法的主要研究目标.

3.2 基于因果函数模型的方法

针对基于约束的方法存在的马尔可夫等价类难

题，很多学者和专家从因果作用机制引发的数据分布特性等角度出发提出了因果函数模型. 这些模型以结构方程模型(Structural Equation Model，SEM)[5]为基础. 结构方程模型是一个可用于多元变量分析的框架，包括随机变量集与方程集：随机变量集包括观察变量集和隐含误差变量集；一组结构方程对应节点为观察变量的有向图，表示模型的因果结构和结构方程的形式. 当结构方程中结果变量 v_j 对应的原因变量 v_i 的系数不为零时，则 v_i 和 v_j 之间存在有向边 $v_i \to v_j$. 结构方程模型可以形式化表示为

$$v_i = f(pa(v_i), n),$$

其中，$pa(v_i)$ 表示变量 v_i 的父亲变量的集合，n 表示噪声变量. 所有变量 v_i 及其父亲变量 $pa(v_i)$ 构成的集合为独立变量集，所有噪声变量 n 构成的集合为误差变量集. 结构方程和普通的代数方程差别在于，结构方程表示的并不只是方程左右两边的相等关系，而是在描述方程左边变量的值是如何被自然或者因果机制所决定的.

虽然 SEM 可以用于多元变量分析，但是在很多情况下，经典的结构方程模型不能估计变量的因果方向. 所以，很多研究引入因果数据产生机制，对 SEM 进行扩展获得了具有较大表达能力的因果函数模型，进而提出了多种基于因果函数模型的因果关系发现算法，代表性的算法包括线性非高斯无环模型(Linear Non-Gaussian Acyclic Model，LiNGAM)[24]、后非线性(Post-NonLinear，PNL)方法[25-26]、非线性条件下的加性噪声模型(Additive Noise Model，ANM)[27,30]及其扩展[30]，信息-几何因果推断(Information-Geometric Causal Inference，IGCI)模型[28-29]，以及基于约束和基于因果函数的混合型方法[49-51]等.

3.2.1　LiNGAM 类算法

LiNGAM，即线性非高斯无环模型，由 Shimizu 等人[24,73]提出，是结构方程模型和贝叶斯网络的变形. LiNGAM 模型要求观察数据的产生机制必须满足以下 3 个条件：(1) 有向图无环——观察变量 v_i，$i \in \{1, \cdots, p\}$ 存在先后因果次序，即后面的变量不会影响前面的变量，这里定义此变量的因果次序为 $k(i)$；(2) 模型线性——变量 v_i 是其对应的原因变量的线性求和，外加噪声变量 n_i 和常数 c_i，即 $v_i = \sum_{k(j)<k(i)} b_{ij} v_j + n_i + c_i$；(3) 噪声独立非高斯——噪声变量 n_i 服从非零方差的非高斯分布(或最多一个是高斯分布的)，且 n_i 彼此间相互独立，即 $p(n_1, n_2, \cdots, n_p) = \prod_i p_i(n_i)$.

在上述线性非高斯无环条件下，LiNGAM 可以形式化表示为

$$\boldsymbol{V} = \boldsymbol{B}\boldsymbol{V} + n \tag{1}$$

其中 $\boldsymbol{V}$ 为 p 维的随机向量，$\boldsymbol{B}$ 为 $p \times p$ 的连接矩阵，n 为 p 维的非高斯随机噪声变量. 因为无环图假设，则存在置换矩阵 $\boldsymbol{P} \in \boldsymbol{R}^{m \times m}$ 使得 $\boldsymbol{B}' = \boldsymbol{P}\boldsymbol{B}\boldsymbol{P}^{\mathrm{T}}$ 为严格的下三角矩阵且对角线的元素均为 0.

LiNGAM 模型最早的求解方法是 Shimizu 等人[24]在独立成分分析算法(Independent Component Analysis，ICA)[74]基础上提出的，即 ICA-LiNGAM 算法. 此算法首先通过 ICA 从观察数据中得到连接矩阵 $\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{V}$ 中的线性变换矩阵 $\boldsymbol{W}$，这里 $\boldsymbol{Y}$ 是包含独立成分的向量；结合这个线性变换和式(1)，可以看出 $\boldsymbol{W}$ 和 $(\boldsymbol{I} - \boldsymbol{B})$ 之间有某种对应关系. 然后，结合 $\boldsymbol{B}'$ 为严格下三角的特性，采用行列置换等方法即可从 $\boldsymbol{W}$ 获得因果次序，最后采用一定的剪枝算法来得到最终的因果网络.

LiNGAM 模型采取的剪枝算法分为两类：一类是基于统计学与最优化的相关理论，例如 Resampling、olsboot(ordinary least squares bootstrapping)与 Adaptive Lasso 等. 另一类是基于贝叶斯网络判断无向结构中一条边是否应该存在的相关算法，此类算法核心点是条件独立性测试，例如 PC 算法[11]，基于马尔可夫毯的 BASSUM(Bayesian Semi-Supervised Method)算法中的剪枝策略[75]等. ICA-LiNGAM 算法基本思想流程示意图如图 7 所示.

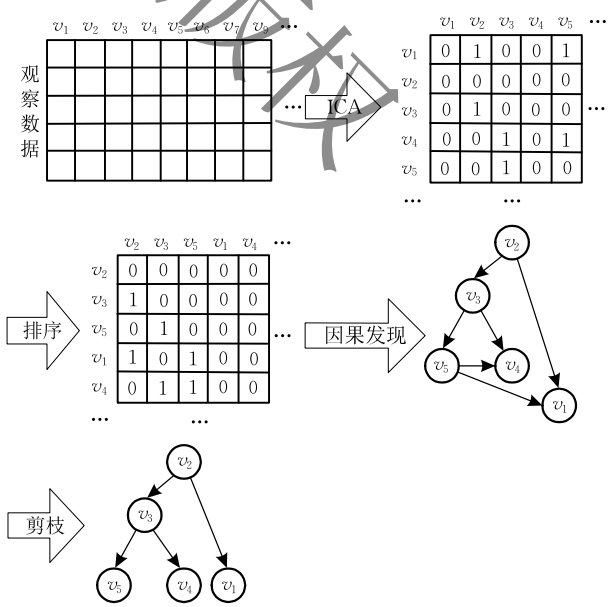


图 7　ICA-LiNGAM 算法基本思想示意图

针对 LiNGAM 方法受限于 ICA 算法而往往收敛于局部最优解的问题，Shimizu 等人利用残差独立性的原理提出了 DirectLiNGAM(Direct Method for a Linear non-Gaussian SEM)算法[41]. 该算法首先根据 Darmos-Skitovitvh 定理[76-77]定义了外生变量(exogenous variable)的判断标准，即与除自身变量之外的所有变量作回归分析后的残差与自身变量最独立的变量. 假设变量集合为 $\boldsymbol{V}$，则外生变量 v_e 可由式(2)～(4)计算得出.

$$v_e=\arg\min_{j\in\boldsymbol{V}} T_k(v_j;\boldsymbol{V}) \tag{2}$$

$$T_k(v_j;\boldsymbol{V})=\sum_{i\in\boldsymbol{V},i\neq j} MI_k(v_j,r_i^{(j)}) \tag{3}$$

$$r_i^{(j)}=v_i-\frac{cov(v_i,v_j)}{var(v_j)}v_j \tag{4}$$

其中 T_k 表示两个变量独立性的程度，MI_k 表示使用的是互信息度量. 假设 v_j 为第一个选出的外生变量，后续只需要用当前残差 $r_i^{(j)}$ 更新当前变量 v_i，使其满足基本 LiNGAM 模型的无环、线性、非高斯性假设；然后递归地完成整个外生变量选择和剔除外生变量影响的变量更新的过程，进而得到整个因果次序；最后通过最小二乘回归方法来完成剪枝.

例如，假设数据集中有 3 个变量 v_1、v_2、v_3，满足如下关系：

$$\begin{bmatrix}v_3\\v_1\\v_2\end{bmatrix}=\begin{bmatrix}0&0&0\\1.2&0&0\\0&0.9&0\end{bmatrix}\begin{bmatrix}v_3\\v_1\\v_2\end{bmatrix}+\begin{bmatrix}e_3\\e_1\\e_2\end{bmatrix}.$$

利用式(2)～(4)选出外生变量 v_3，此时变为

$$\begin{bmatrix}r_1^{(3)}\\r_2^{(3)}\end{bmatrix}=\begin{bmatrix}0&0\\-1.3&0\end{bmatrix}\begin{bmatrix}r_1^{(3)}\\r_2^{(3)}\end{bmatrix}+\begin{bmatrix}e_1\\e_2\end{bmatrix}.$$

此时仍然满足 LiNGAM 三个基本特征，故接下来通过递归完成即可.

针对线性非高斯无环模型中出现的隐变量问题，已有相关的工作. Hyvärinen等人[78]在 DirectLiNGAM 下用基于似然比的方法判断外生变量，并将算法扩展到非线性情况下；Tashiro 和 Shimizu 等人[43]提出 ParceLiNGAM 算法，主要通过测试估计的外生变量影响的独立性和找到包含未被隐变量所影响的变量子集来发现隐变量；在可识别性研究上，Hoyer 等人[44]结合 LiNGAM 模型，提出适用于线性非高斯条件下的 lvLiNGAM(latent variable LiNGAM)框架，但是它也对数据线性产生的要求很高.

作为两类主流的因果关系发现算法，ICA-LiNGAM 和 DirectLiNGAM 方法各有其优缺点. 从算法性能角度，ICA-LiNGAM 算法效果更多地依赖于 ICA 算法本身，如果初始解选择不好，容易陷入局部最优解；虽然 DirectLiNGAM 算法较好地解决了这个缺陷，能够在有限的步骤内找到因果次序，但是带来的计算代价会大大增加. 从计算复杂度角度，DirectLiNGAM 的计算复杂度是 $O(mp^3M^2+p^4M^3)$，而 ICA-LiNGAM 的复杂度是 $O(mp^3+p^4)$，这里 m 表示样本个数，p 表示样本维度，M(远远小于 m)表示在核独立性检验中低秩分解的最大的秩.

3.2.2 PNL 类算法

PNL 方法，即后非线性方法，是由 Zhang 等人[25-26]提出的. 对于 PNL 模型，假设存在 $v_i\to v_j$，其因果机制可以用以下公式表示：

$$v_j=f_2(f_1(v_i)+n_j) \tag{5}$$

其中，原因变量 v_i 和噪声变量 n_j 互相独立，f_1 是不恒定的光滑函数，f_2 是可逆的光滑函数，且 $f_2'\neq 0$.

该模型描述数据产生过程的能力较强——f_1 表示了因的非线性作用，而 f_2 可以刻画观测时的非线性变形. 尽管该模型有广泛的适用性，下面我们可以看到如果 v_i 是 v_j 的因，也就是式(5)成立，那么相反方向一般不能满足噪声独立的假设.

为此，假设存在相反方向 $v_j\to v_i$，用 PNL 模型可以表示为

$$v_i=g_2(g_1(v_j)+n_i) \tag{6}$$

其中，原因变量 v_j 和噪声变量 n_i 互相独立，g_1 是不恒定的光滑函数，g_2 是可逆的光滑函数，且 $g_2'\neq 0$.

文献[24]中已经证明，只有在特殊的函数和分布设置下，式(5)和式(6)产生 v_i 和 v_j 的分布才是相同的. 在一般的情况下，如果数据是按照式(5)产生的，那么无法找到类似于公式(6)的模型可以产生相同的变量分布. 因此，因果关系方向就可以被识别. 也就是说，根据 PNL 模型产生的数据，其变量之间的因果关系几乎在所有情况下都是可以被识别的. 在识别变量之间的因果关系方向时，主要按照以下两个步骤进行[79]：首先，使用条件独立性测试找到满足 d-分离的等价类；然后，利用 PNL 模型识别上一步中未确定的因果方向，对于每个包含等价类的因果结构，采用非线性独立成分分析的方法估计噪声变量 n，检查假设的原因变量 v_i 是否与假设的结果变量 v_j 对应的噪声变量 n 互相独立，如果独立，则说明 v_i 和 v_j 的因果关系方向判断为 $v_i\to v_j$，反之亦然.

最近，Zhang 等人[80]利用 PNL 方法，证明了在出现选择偏倚的情况下，两个随机变量之间的因果方向是可识别的.

《计算机学报》版权

这个模型的优点在于：在二元变量的完全因果关系结构识别下具有便捷性和有效性；利用上述步骤进行因果关系发现，避免了对所有可能的因果结构方向进行穷举搜索，同时不需要进行额外的高维统计检验，降低了算法复杂度. 缺点在于，PNL 模型需要两阶段的非线性回归，其时间复杂度较高.

PNL[25-26]的特殊形式包含了 LiNGAM[24,73]和 ANM[27]. 具体来说，当函数 f_1 和 f_2 都为线性的，且 n 为非高斯噪声时，$v_j=b_{ji}v_i+n_j$（其中 b_{ji} 表示变量 v_i 和变量 v_j 间的连接矩阵的值），其相当于 LiNGAM 模型；当 f_2 为线性函数且 f_1 为非线性函数时，$v_j=f(v_i)+n_j$，其相当于后续的 ANM 模型. 因此，上述 PNL 可识别性的结论也同样适用于 LiNGAM、ANM 及它们的衍生模型. 接下来我们对 ANM 算法进行介绍.

3.2.3　ANM 类算法

LiNGAM 类算法只能解决线性的问题，在非线性情况下，Hoyer 等人[27]证明了非线性函数可以跟非高斯模型起到类似的作用，能帮助打破变量之间的对称性，有助于识别因果方向. 因此，Hoyer 等人[27]提出了加性噪声模型（Additive Noise Model，ANM），描述了非线性条件下实现二元变量间的因果关系发现方法（如图 8）. ANM 算法对数据产生方式也有较强的假设：结果变量能被表示成关于原因变量的一个函数（这个函数关系不一定是线性的），加上与原因变量独立的加性噪声. ANM 模型可以表述为

$$v_j=f(v_i)+n,\ n\perp v_i,$$

其中 f 是任意非线性函数，n 是随机概率分布下产生的噪声项. 这个模型表示变量 v_i 和变量 v_j 之间存在 $v_i\to v_j$ 形式的因果关系，其中蕴含着原因变量 v_i 和噪声变量 n 相互独立，而结果变量 v_j 和噪声变量 n 不独立这一重要特性. 基于上述特性，ANM 采用如下方法判别候选因果关系方向 $v_i\to v_j$ 是否成立：首先，对 v_i 和 v_j 进行回归分析 $v_j=f(v_i)+n$；然后，针对回归分析计算残差量 $n=v_j-f(v_i)$；最后，检查这些残差量 n 是否与 v_i 独立，如果独立，则 v_i 和 v_j 的关系判断为 $v_i\to v_j$.

在上述过程中，回归方法和独立性检验是决定 ANM 算法效果的重要步骤. 在回归方面，高斯回归过程（Gaussian processes for regression）[81]是 ANM 模型中最早使用的回归方法；在独立性检验方面，基于核方法（Kernel-based Conditional Independence Test）[63,82]的非线性条件独立性测试是 ANM 模型中采用的主流方法.

在上述原因变量与噪声变量相互独立这一思想的基础上，后续多个工作将加性噪声模型扩展到了离散型数据、多维变量、无噪声等情况. Peters 等人[30]对 ANM 算法做了进一步扩展，使之适用于离散型数据. 后来他们还在现有研究的基础上，提出了一种结合后续独立性检验的回归方法（Regression with Subsequent Independence Test，RESIT）[42]，将 ANM 模型扩展到了多维变量的情况，这种方法适用于误差同方差模型或离散数据的情况.

不同于 ANM，Daniusis 和 Janzing 等人[28-29]从信息几何的角度，提出了基于信息熵的因果推断（Information-Geometric Causal Inference，IGCI）算法，利用原因变量的分布和“因-果”函数机制的独立性来判断变量间的因果关系[83]（如图 9）. 利用 IGCI 算法在识别因果关系方向时，主要基于以下两个假设：一是因果影响过程是无噪声的；二是变量的概率分布 $p(v_i)$ 以及两个变量间的非线性函数 $f(v_i)$ 的导数之间是统计不相关的，这个假设保证了能够获得两个变量间的因果关系，在对比两个候选方向上具有可靠性.

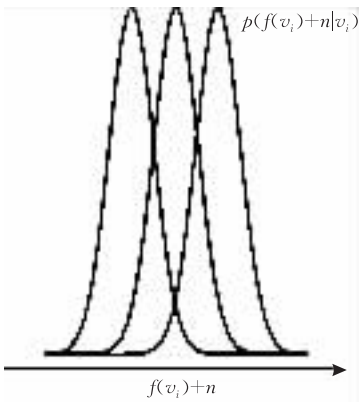


图 8　ANM 模型基本机制说明，满足因变量 v_i 和噪声 n 的某种独立性（参考文献[27]）

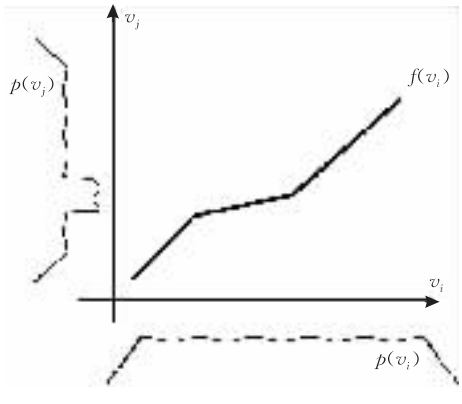


图 9　IGCI 模型基本机制说明，满足 $f'(v_i)$ 和 $p(v_i)$ 的某种独立性（来源于文献[29]）

计算机学报 版权

IGCI 的上述两个假设较为严格，因此后续多个工作希望弱化这两个假设. 在噪声方面，Janzing 等人[29]指出 IGCI 方法在较小噪声场景中仍存在一定的适用性；在变量分布与函数关系方面，Sgouritsa 等人提出基于逆回归(Inverse regression)的方法[84]可以弱化上述假设.

针对不完全观察数据的隐变量检测问题，基于因果函数的方法目前也有一些进展. 在普遍意义下的隐变量检测方面，Chaves 等人[47]结合量子信息不等式(information equality)、贝尔定理等量子力学的理论，将经典量子有向无环图的熵描述看成一个一般化的纯变量情况下的框架，提出了一种基于贝叶斯网络条件熵的线性不等式的隐变量发现方法. 他们认为因果结构是在联合概率分布的某些熵存在不平等的情况下出现的，由此指出了给定相互作用模式中，计算其产生的关系信息熵(information entropy)约束算法[48]. 这种用量子信息不等式等发现隐变量结构的思路与方法为未来的研究奠定了一定的基础.

作为面向非线性因果关系的主流因果关系发现算法，基于噪声独立的方法和 IGCI 的适用范围各有侧重，其本质区别体现为模型背后的因果机制假设. 基于噪声独立的方法的核心假设为原因变量与噪声变量的独立性，因此它们主要适用于一定噪声分布场景；而 ICGI 则从信息几何角度，假设因果影响过程是无噪声的，因果方向的判别则是基于原因变量的概率分布和因果函数的独立性，因此 ICGI 类方法主要侧重于无噪声、或低噪声，且因果函数较为复杂的情形.

3.3 基于约束和基于因果函数混合型方法

近年来，随着因果关系假设和因果函数模型的研究不断发展，有部分学者尝试把现有一些方法和因果函数模型结合起来，产生混合型因果关系发现方法，有效地提高因果函数模型的不足，同时克服了高维数据上误发现率控制难题.

Cai 等人[49]将因果关系发现问题分解为子问题并利用递归方法求解，以提高算法的精确度. 他们提出的 SADA 框架(Scalable cAusation Discovery Algorithm)，主要适用于因果结构中的稀疏属性的观察分析，在样本集较少的情况下也能正确地辨别因果变量. 其主要思想是：首先通过求解因果分割集(Causal Cut)将高维问题分解成 2 个子问题；然后针对每个子问题进行递归分解直到其问题规模足够小；针对每个足够小的子问题，采用 ANM 等基于因果函数模型的方法进行求解，最后对小问题进行合并. 其上述过程可以归纳为图 10 所示分解、求解、合并三阶段过程.

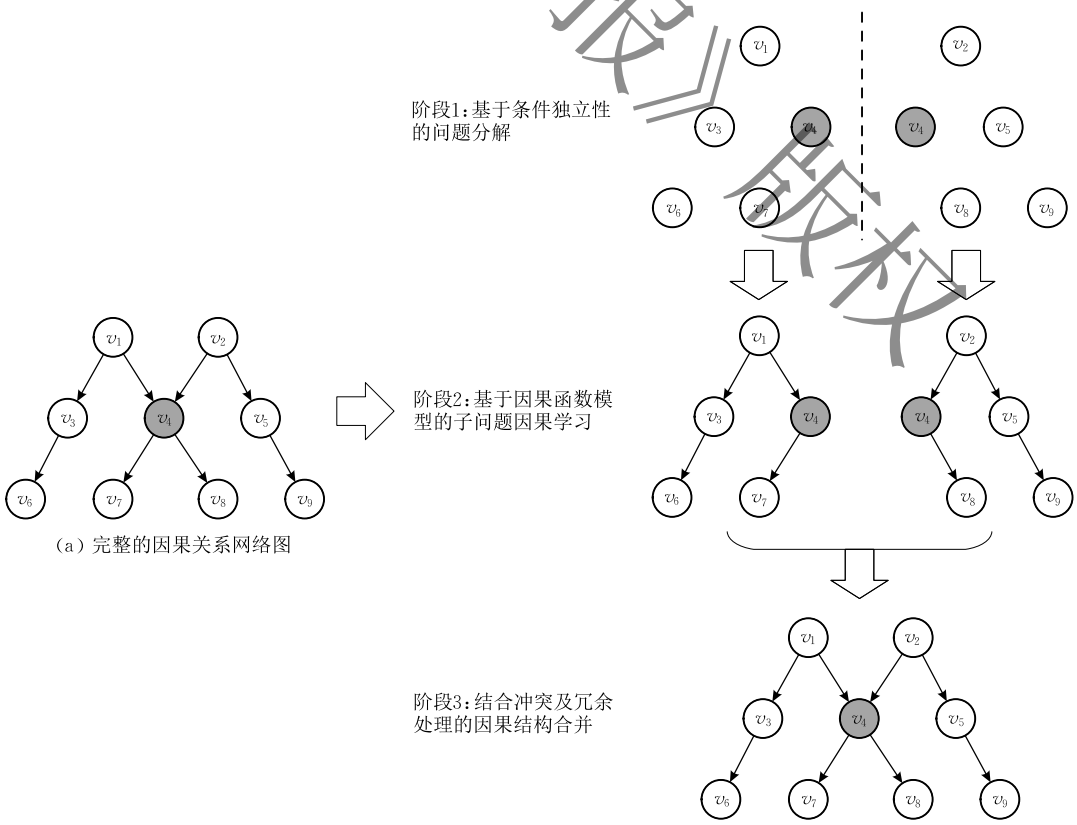


图 10 SADA 框架因果关系发现主要过程示意图

《计算机学报》版权

与 SADA 方法思路类似，Chen 等人[50]提出了 MCDSL(Multiple-Cause Discovery method combined with Structure Learning)方法，将高维离散数据的因果关系发现问题分为结构学习(Structure Learning Phase，SLP)和方向学习(Direction Learning Phase，DLP)两个阶段，从另外一种途径实现了基于约束的方法和基于因果函数模型的方法的有效结合. 在结构学习阶段，利用条件独立性关系辨别和排除直接原因变量；在方向学习阶段，使用了条件概率表和 ANM 的混合方法. Hao 等人[51]采用了先发现通过探索因果骨架图进行结构学习，再利用 ANM 进行方向判别的大规模因果发现混合方法(HYbrid Approach for large scale causality discovery，HYA)，减少了计算的复杂度，提高了结果的准确率.

混合型方法一定程度上实现了基于约束方法具有较好的高维扩展性和基于因果函数模型的方法具有较强的因果发现能力的优势结合，是实现高维数据因果关系发现的主流手段方法，但是，这部分研究尚处于起步阶段，普遍面临着理论分析不足的问题. 例如，以 Cai 等人[49]提出 SADA 为代表的分治方法，在问题分解过程中可能会引入额外的隐变量，导致子问题违反因果忠诚性假设；Chen 等人[50]采用的两阶段方法则面临着如何保证两阶段所需方法采用的因果假设的一致性和相容性.

3.4　验证与评测

3.4.1　数据集

现有的因果关系发现研究中使用的数据集一般分为模拟生成的数据集和真实的数据集，有些数据集是模拟数据和真实数据的混合数据集.

模拟生成的数据集的产生方式有两种：一种是根据一定的规则算法(或方程)自动生成的(如 LiNGAM 算法中的仿真数据集)；另一种是根据已知的网络结构生成的(如 SADA 方法中采用的数据集). 这类数据集虽然实验结果有准确的结果进行对照，便于分析算法的各项性能，但是由于数据是模拟生成的，不具有真实性和普遍的适用性.

真实的数据集来源于真实场景，具有较强的客观性，但是目前公开的数据集较少，且特定的数据集一般只适用于专门的领域，比如非时序观察数据的数据集只适用于非时序观察数据的因果关系发现方法的研究. 目前经人为通过真实数据进行处理后的公开数据集主要有以下几个：

(1) 二元变量间的因果关系发现数据集，只适用于因果对之间的因果关系发现研究[83]：http://webdav.tuebingen.mpg.de/cause-effect；

(2) 一些用于因果关系分析(偏机器学习领域)的数据集：http://www.causality.inf.ethz.ch/workbench.php?page=index；

(3) NIPS 2008 Workshop on Causality 专门设置的一个因果关系与预测的挑战竞赛(Causality Challenge #1：Causation and Prediction).

该竞赛的数据集为真实获得的数据或人工生成的数据，主要有根据实际数据模拟得到的基因表达数据集(REsimulated Gene Expression Dataset，REGE-D)、艾滋病的 HIV 病毒分子描述的简单药物工作机制数据集(SImple Drug Operation mechanisms，SIDO)、用于发现影响高收入的社会经济原因的人口调查数据(Census Is Not Adult，CINA)、用于发现肺癌原因的基因组数据(Measurement ARTIfact，MARTI)等. 这些数据集都包含了训练数据、测试数据和验证数据：http://www.causality.inf.ethz.ch/challenge.php；

(4) NIPS 2008 Workshop on Causality 专门设置的一个与因果关系发现相关的挑战竞赛(Causality Challenge #2：Pot-Luck).

该竞赛的数据集包括人类 T 细胞上的因果蛋白质信号网络(Causal Prote-inSignaling Networks in human T cells，CYTO)、局部结构因果网络(LOcal CAusal NETwork，LCANET)、脱落酸信号网络(Abscisic Acid Signaling Network，SIGNET)、目标信息等价类的数据集(Target Information Equivalent Dataset，TIED)、判别原因变量和结果变量(Distinguishing between cause and effect，Cause-EffectPairs)等. 这些数据集包括人为模拟的数据集和真实场景下的数据集：http://www.causality.inf.ethz.ch/challenge.php；

(5) NIPS 2013 Workshop on Causality：Large-scale Experiment Design and Inference of Causal Mechanisms 专门设置的一个因果对因果关系识别的挑战竞赛(Causality Challenge #3：Cause-effect pairs).

该竞赛中的数据涵盖了化学、气候、生态、经济、工程、流传病学、基因组学、医学、物理学和社会学等不同领域的已知变量间因果关系的真实数据集. 竞赛的目标是在已知一对变量$\{v_i, v_j\}$的样本集下，确定 v_i 和 v_j 之间的因果关系(两个变量之间的因果关系可能为：v_i 是 v_j 的原因；v_j 是 v_i 的原因；v_i 和 v_j 受共同的原因变量影响；v_i 和 v_j 是互相独立的). 竞赛的数据集：http://www.causality.inf.ethz.ch/cause-effect.php?page=data.

《计算机学报》版权

3.4.2 评价方法

目前因果关系发现的研究没有较为统一、客观的评价方法.常用的因果关系研究的评价方法主要有以下几种:

(1)均方误差 *MSE*(Mean Squared Error)

MSE 主要用于衡量参数估计的准确程度.上述提及的典型的因果函数模型中,LiNGAM 类算法[24,41]就是以 *MSE* 作为算法的评价方法的.其定义为

$$MSE=\sqrt{\frac{1}{p^2}\sum_{i=1}^{p}\sum_{j=1}^{p}(b_{ij}-\hat{b}_{ij})^2},$$

其中,p 表示数据集中的所有变量个数,b_{ij} 表示因果关系网络对应的连接矩阵中第 i 行第 j 列的真值,$\hat{b}_{ij}$ 表示因果关系网络对应的连接矩阵中第 i 行第 j 列的估计值.

(2)准确率(precision)、召回率(recall)和综合评价指标 F_1 值(F_1-score)

正确率是指所有通过因果关系发现算法学习到的因果关系网络图中正确的边数所占的比例,主要用来衡量因果关系网络图中节点间本来不存在的边被错误添加的程度;召回率是指所有学习到的正确的边数占整个数据(学习到的和未学习到的)中真正正确的边数的比例,用来衡量节点间存在的边没有被发现的程度.F_1 值集合了前两个评价参数,用来评价因果关系发现算法的总体优劣.上述提及的典型因果函数模型中,基于约束和基于因果函数混合型方法(如 SADA[49]、MCDSL[50])等是以这种方法对算法的性能进行评价的.3 个参数通过以下公式来衡量:

$$Precision=\frac{TP}{TP+FP},$$

$$Recall=\frac{TP}{TP+FN},$$

$$F_1=\frac{2\times Recall\times Precision}{Recall+Precision},$$

其中,TP 表示因果关系网络本身存在边,且学习到的结果也存在边的边数;FP 表示因果关系网络本身不存在边,而学习到的结果存在边的边数;FN 表示因果关系网络本身不存在边,而学习到的结果不存在边的边数.

(3)ROC 曲线下的面积 *AUC*(Area Under the ROC Curve)

AUC 是一种用来度量分类模型好坏的标准.典型的因果函数模型中,二元变量间的因果关系发现方法(如 NIPS 2008 Workshop on Causality 的竞赛中的评价方法)就是采用了这个评价指标.ROC 曲线(Receiver Operating Characteristic curve)是一条二维平面上的曲线,曲线的横坐标表示本身为负类被检测为正类的比例,即假阳性率 *FPR*(False Positive Rate);纵坐标表示本身为正类被检测为正类的比例,即真阳性率 *TPR*(True Positive Rate),其计算方法为

$$TPR=\frac{TP}{TP+FN},$$

$$FPR=\frac{FP}{FP+TN},$$

其中,TP、FP 和 FN 在因果关系发现方法评价中的含义同上一评价指标中的含义,TN 表示因果关系网络本身存在边,而学习到的结果不存在边的边数.

AUC 是指处于 ROC 曲线下方那部分面积的大小.*AUC* 采用梯形法来求近似值,其值通常介于 0.5 到 1.0 之间.*AUC* 的值越大,表示正确率越高.

3.4.3 软件或工具包

在研究过程中,学者们开发了一些软件或工具包并在网上共享,便于后续研究者的研究.其中主要的因果关系发现方法的软件或工具包如下:

(1)Glymour 等人开发的 Tetrad IV 工具(包含 IC、PC、LiNGAM 等因果关系发现算法):http://www.phil.cmu.edu/projects/tetrad/;

(2)Zhang 等人提出的 KCI-test(Kernel-based Conditional Independence Test)[63]:http://people.tuebingen.mpg.de/kzhang/KCI-test.zip;

(3)Shimizu 等人提出的线性非高斯无环模型(LiNGAM)及其扩展[24,41]:https://sites.google.com/site/sshimizu06/lingam;

(4)Hyvärinen 等人提出的独立成分分析(Independent Component Analysis,ICA)的快速算法 FastICA[85]:http://research.ics.aalto.fi/ica/fastica/;

(5)Hoyer 等人提出的适用于非线性情况的加性噪声模型(ANM)[27]:http://webdav.tuebingen.mpg.de/causality/additive-noise.tar.gz;

(6)Rasmussen 等人提出的基于高斯回归过程(Gaussian Processes)的 GPML(Gaussian Processes for Machine Learning)工具箱[86]:http://mloss.org/revision/download/558/;

(7)Zhang 和 Hyvärinen 等人提出的后非线性(PNL)因果模型[25-26]:http://webdav.tuebingen.

《计算机学报》版权

mpg.de/causality/CauseOrEffect_NICA.rar；

(8) Daniusis 等人提出的信息-几何因果推断模型(IGCI)[28]；Janzing 等人编写的源码[29]：http://webdav.tuebingen.mpg.de/causality/igci.tar.gz；

(9) Mooij 等人提出的识别因和果的概率隐变量模型(GPI-MML)[45]：http://webdav.tuebingen.mpg.de/causality/nips2010-gpi-code.tar.gz；

(10) 由 Colombo 等人[39]开发的可用于隐变量发现的 FCI 算法和 RFCI 算法的 R 包 pcalg 包[87]：http://CRAN.R-project.org/package=pcalg；

(11) 因果关系发现中心(Center for Causal Discovery)开发的软件和工具，其中集成了很多算法，包括了一些可以直接调用的 API：http://www.ccd.pitt.edu/data-science/software/.

以上因果关系发现方法的软件和工具包等为学者们的研究提供了极大的帮助，在使用期间，可以将上述工具分为 3 类：第 1 类是系统软件，第 2 类是特定算法源码，第 3 类是辅助工具类. 选用时可以参考表 3.

表 3　主要因果关系发现软件或工具包分类

类别	代表软件或工具包
系统软件	Tetrad IV 工具 因果关系发现中心 pcalg 包[87]
特定算法源码	LiNGAM 及其扩展[24,41] PNL[25-26] ANM[27] IGCI[28-29] GPI-MML[45]
辅助工具类	KCI-test[63] FastICA[85] GPML 工具箱[86]

3.5　典型应用

近年来，基于非时序观察数据的因果关系发现算法已经在生物医学、经济学、经验科学、计算机和交通通信等领域上进行了应用尝试，并在某些特定领域取得一些有效的应用成果.

在生物医学中，基因表达数据分析是因果关系发现算法应用较多的领域. 生物学家希望通过这些数据来学习基因之间的通路、某些疾病的因果基因以及基因诊断的结果. 这类问题的常用方法是将因果关系作为特征选择的目标. 在因果特征选择方面，Koller 等人在这方面做了大量开创性工作，证明了马尔可夫毯(Markov blanket)是全局最优特征集[88]. 在生物信息因果关系发现上，因果节点被看成是马尔可夫毯的子集. 后来的一些扩展方法基本上都是围绕这种方法进行开展. Tsamardinos 等人[37]提出了 MMPC 算法发现局部结构中的马尔可夫毯，确定其中的候选父亲节点. Cai 等人[75]采用一种新的贝叶斯半监督算法——BASSUM 算法，提高了马尔可夫毯算法的可靠性. Liu 和 Cai 等人[89]提出了 CASTLE(Causality Analysis model based on STructure LEarning)方法，从药物的医药和生物方面综合信息上挖掘引起药物不良反应的分子因素. 为了解决高维数据的结构学习问题，Wang 等人[90]提出将聚类与贝叶斯网络结构学习相结合的方法，并应用于关于肾脏疾病的中医药研究中. 在其他生物医学问题上，不少专家和学者也利用因果关系取得了一定的成果，如 Cooper 等人[91]利用 ALARM 网络作为初始的原型对手术室中的麻醉问题进行建模研究，Zhou 等人[92]在生物网络推断方面的成果，Dong 等人[93]在疾病诊断领域的贡献，Zhao 等人[68]在中药效果分析问题上的研究等.

在经济学领域，由于因果网络模型是不确定性推理问题中常用到的工具之一，采用它能根据外部的确定性，随机性干扰给出合理的概率性预测，它常被用到定价预测、保险欺诈侦测、风险预测、风险分析、风险管理中[5]. Chen 等人[50]提出了一种结合结构学习和方向学习混合的 McDSL 算法，对上海股市交易 13 年的历史金融数据分析，发现股票收益的多重原因，在指导投资者制定他们的投资策略上比其他方法更有效. Hu 等人[94]提出一种改进的含有条件概率表的加性噪声模型，解决高维动态股市多对一的因果关系发现问题，有效地挖掘出投资组合选择因素和股票收益之间的关系. 为了选择表现股市预测效果和从股市分析中识别有代表性的特征，Zhang 等人[95]结合因果关系理论提出因果特征选择(Causal Feature Selection，CFS)算法，选择出最优的特征子集.

除此之外，Fire 等人[96]在结构方程模型等基础上，提出一种从视频中推断人的行为和物体状态检测之间的因果关系的方法. Hours 等人[97]利用因果图模型研究网络协议的性能，通过例子证明了因果推断工具在预测电信网络的体系和路由策略改变造成的影响上有一定的有效性. Hu 等人[98]提出了一种基于因果约束贝叶斯网络(Bayesian Network with Causality Constraints，BNCC)算法，有效发现了软件开发风险的主要因素.

针对实际场景中的典型情况，不少学者也有针对性地对因果函数模型进行了改进，使其更好地解决实际问题. Hao 等人[99]为了克服实际场景下的无线网络性能指标维数较大，且指标之间相关性强的难题，尝试结合指标之间的关系和先验知识

《计算机学报》版权

提取原子指标,用典型因果推断(Canonical Causal Inference,CCI)算法对原子指标进行了因果推断,发现了一些有意义的无线网络指标间的因果关系.而在没有因果关系先验知识或无法事先知道隐变量数量的前提下,Kummerfeld 等人[100]提出了 FOFC(FindOneFactorClusters)算法,采用 1-因子的度量模型,对具有共性的多重指标进行因果聚类,并用单因子指标变量表示,进而对指标变量之间的因果关系进行推断,这种方法对大指标变量集比较有效,在小样本中也具有一定的可靠性.在商业应用程序中,各种因素间的因果关系发现存在隐混淆因子、选择偏倚、缺失某些值的数据等情况,致使因果关系发现存在错误.Borboudakis 等人[101]就此问题进行了多种情况下的分析,结合条件约束的方法,提出一种适用于商业应用程序因果关系发现的 FTIO 算法,证明了算法的鲁棒性和普适性,并应用于汽车保险公司的案例中.针对隐私敏感的因果关系发现场景,Kusner 等人[102]在提出结合差分隐私保护的 ANM 因果关系发现模型,其核心是在 ANM 模型框架的基础上加入少量的噪声项(即差分隐私保护数量),理论分析表明只要这些隐私数量具有足够的可信度,那么该部分额外加入的噪声项就不会影响因果关系发现的结果.

上述这些都体现了基于非时序观察数据的因果关系发现在实际场景下具有重要应用意义和价值.特别是在特定场景下,对现有的因果函数模型进行改进,使其应用于处理过的数据集,能容易得出较好的结果.但是,实际情况中存在数据缺失严重、数据量大、部分变量未能检测到等难题,运用同样的方法会出现不合理想的效果,且算法的鲁棒性也不高.

4 未来研究方向

从目前的研究来看,基于非时序观察数据的因果关系发现方法对数据产生方式假设、适用条件等都有较高的要求,存在着因果关系方向推断,高维数据上的误发现率控制和不完全观察数据上的隐变量检测这 3 个难题.结合当前的研究难点和研究现状,我们认为未来基于非时序观察数据的因果关系发现的研究方向可能至少有以下几个方面:

(1)因果关系方向推断.因果关系方向推断的核心在于发现并度量因果变量间的不对称性,提升观察数据上的因果发现能力.在具体情况下,如何针对数据特性,设计既具有较强发现能力,又对某一类问题具有较强适用性,打破因果关系不对称性的因果关系方向推断方法已经成为本领域的研究热点,涌现了 LiNGAM、PNL、ANM 等众多方法.但是,上述方法对数据产生机制具有较强的假设,离某些领域的应用场景仍可能存在较大的距离.因此,如何提升因果推断方法的可靠性和普适性仍是未来研究的热点.

(2)高维数据误发现率控制.现有的因果关系发现仍局限于小问题或更多着眼于理论问题,相关算法的时间复杂度和空间复杂度较高.对高维数据的因果关系发现问题主要采用分解的方法,将大问题分解为小问题,后期的合并处理方法也难免会产生误发现率高的问题.针对这个问题,解决思路可以从以下两个方面考虑:一方面可以考虑引入先验知识、假设检验(统计方法)等提高高维数据发现的显著性,以此降低算法的误发现率;另一方面,受分治策略启发,我们可以通过先学习局部结构,再整合成全局结构的方法,从中发现局部结构中因果关系的冲突相容关系,进而得到更加可靠的因果关系结构;充分利用大数据并行技术,对算法进行改进和优化也是另外一种可行的思路.

(3)隐变量检测.隐变量常见的类型有混淆因子和选择偏倚两种,其核心在于发现隐变量局部结构并给出度量方法.因此,可以结合其他领域的工具或知识,从量子信息不等式、工具变量的角度出发,探索隐变量检测方法.该方向的研究已经在特定假设前提下取得了部分成果[47-48],但是如何将这些结论进行推广,弱化其隐含的假设,仍是有待研究的重要课题.同时,随着大数据时代的到来,利用大数据来全方位地刻画变量之间的结构信息,以及通过验证因果关系发现中的虚假因果关系反证出难以观察到的隐变量结构,将为这方面研究带来新的契机.

(4)验证基准.目前用于验证因果关系发现算法性能的公开标准数据集较少,这已经成为大家公认的限制因果关系领域发展的一个瓶颈.其核心原因在于很多场景下数据背后的因果机制是未知的.因此,如何用科学的方法来验证算法成功获得因果关系,以及如何积累已知背后因果机制数据集是克服上述评价的有效手段.对算法进行验证时,我们可以借鉴和修改其他领域现有的研究,学习和应用统计学、人类学和行为学的验证方法,如天然实验等验证方法研究[103]等.在基准数据集方面,推动因果关系在特定领域中的应用,进而积累特定应用场景中的基准数据是较为可行的方案,也是生物信息学这一因果关系应用领域的成功经验.同时也可以通过图方法去发现一些案例或人为地根据领域知识,以图的形式构造一种验证基准,用这些实例去分析、

验证.

(5)与机器学习等领域研究相结合. 在机器学习领域,最近几年有研究者发现机器学习中半监督学习方法的有效性与因果关系发现中的 $p(v_i)$ 和 $p(v_j|v_i)$ 的独立性有紧密联系[104]. 具体来说,研究发现如果特征 v_i 是类别 v_j 的原因且没有混淆因子存在,那么半监督学习中的无标记点并不能提升分类的准确性. 另一个直接与因果模型有关的学习任务是领域自适应学习. Zhang 等人[105-107]则提出了通过一系列基于因果机制来刻画不同领域间有用信息如何迁移的领域自适应的方法. 这些前期研究表明,因果关系理论给出了隐藏在观测数据背后的有用信息,为机器学习中变量间的互相影响关系预测提供了一个指导方向. 因此,与机器学习等领域的研究相融合也可能成为未来因果关系分析研究的一个有意义的思路和方向.

(6)大数据环境下实际场景的应用. 目前的因果关系发现研究主要停留于理论研究阶段,实际的大规模应用例子较少,而且因果机制验证的困难性也给研究者们带来一定的阻力. 随着大数据时代的到来,多源数据可以为事物提供较为完整的刻画,对隐变量检测等研究提供帮助. 海量多源数据中蕴含着丰富的信息,对这些数据进行因果关系发现,将解决相关性的方法无法解决的问题,进而推动相关领域的发展. 因果关系发现研究在大数据环境下的实际问题研究可以从以下方面做进一步的探讨,比如,如何处理时变因果关系,如何处理数据中常出现的选择偏倚,如何在大规模数据中高效发现正确的因果关系等等. 目前这类研究在几个研究组中刚刚开展[43,47-48].

5　结　论

本文从观察数据的角度出发,讨论了基于非时序观察数据的因果关系发现问题的典型任务,总结并分析了当前基于非时序观察数据的因果关系发现的研究现状,并阐述了因果关系发现方法在生物医疗、经济学等领域中的典型应用. 在综述研究进展的基础上,我们论述了在该研究领域未来可能的研究方向,提出该领域的研究将会围绕因果关系方向推断、高维数据上的误发现率控制和不完全观察数据上的隐变量检测等方面而发展. 展望未来,该领域会逐渐与其他学科以及实际场景相结合,基于非时序观察数据的因果关系发现方法的价值与优势也会更明显的展示出来. 随着大数据时代的到来,该领域的研究会被更多学者所关注与重视,并在理论研究和实际应用中蓬勃发展.

参　考　文　献

[1] Mattmann C A. Computing: A vision for data science. Nature, 2013, 493(7433): 473-475

[2] McAfee A, Brynjolfsson E. Big data: The management revolution. Harvard Business Review, 2012, 90(10): 60-68

[3] Hey T, Tansley S, Tolle K. The Fourth Paradigm: Data-Intensive Scientific Discovery. Redmond, USA: Microsoft Research, 2009

[4] Reshef D N, Reshef Y A, Finucane H K, et al. Detecting novel associations in large data sets. Science, 2011, 334(6062): 1518-1524

[5] Pearl J. Causality: Models, Reasoning and Inference. 2nd Edition. Cambridge, United Kingdom: Cambridge University Press, 2009

[6] Cooper G F, Yoo C. Causal discovery from a mixture of experimental and observational data//Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence. Stockholm, Sweden, 1999: 116-125

[7] Muchnik L, Aral S, Taylor S J. Social influence bias: A randomized experiment. Science, 2013, 341(6146): 647-651

[8] Aral S, Walker D. Creating social contagion through viral product design: A randomized trial of peer influence in networks. Management Science, 2011, 57(9): 1623-1639

[9] Danks D, Plis S. Learning causal structure from undersampled time series//Proceedings of the NIPS Workshop on Causality. Nevada, USA, 2013: 1-10

[10] Gong Ming-Ming, Zhang Kun, Schoelkopf B, et al. Discovering temporal causal relations from subsampled data//Proceedings of the 32nd International Conference on Machine Learning (ICML 2015). Lille, France, 2015: 1898-1906

[11] Spirtes P, Glymour C N, Scheines R. Causation, Prediction, and Search. 2nd Edition. Cambridge, USA: MIT Press, 2000

[12] Simon H A. Spurious correlation: A causal interpretation. Journal of the American Statistical Association, 1954, 49(267): 467-479

[13] Spirtes P, Zhang Kun. Causal discovery and inference: Concepts and recent methodological advances. Applied Informatics, 2016, 3(1): 1-28

[14] Druzdzel M J. The role of assumptions in causal discovery// Proceedings of the 8th Workshop on Uncertainty (WUPES-09). Liblice, Czech Republic, 2009: 57-68

[15] Ramsey J, Zhang Ji-Ji, Spirtes P L. Adjacency-faithfulness and conservative causal inference//Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06). Arlington, USA, 2006: 401-408

[16] Zhang Ji-Ji, Spirtes P. Detection of unfaithfulness and robust causal inference. Minds and Machines, 2008, 18(2): 239-271

[17] Spirtes P, Zhang Ji-Ji. A uniformly consistent estimator of causal effects under the k-triangle-faithfulness assumption. Statistical Science, 2014, 29(4): 662-678

计算机学报 版权

[18] Zhalama, Zhang Ji-Ji, Mayer W. Weakening faithfulness: Some heuristic causal discovery algorithms. International Journal of Data Science and Analytics, 2016: 1-12

[19] Lopez-Paz D, Muandet K, Schölkopf B, et al. Towards a learning theory of cause-effect inference//Proceedings of the 32nd International Conference on Machine Learning (ICML 2015). Lille, France, 2015: 1452-1461

[20] Rubin D B. Bayesian inference for causal effects: The role of randomization. The Annals of Statistics, 1978, 6(1): 34-58

[21] Pearl J. Causal diagrams for empirical research. Biometrika, 1995, 82(4): 669-688

[22] Wright S. Correlation and causation. Journal of Agricultural Research, 1921, 20(7): 557-585

[23] Bollen K A. Structural Equations with Latent Variables. New York, USA: John Wiley & Sons, Inc., 2014

[24] Shimizu S, Hoyer P O, Hyvärinen A, et al. A linear non-Gaussian acyclic model for causal discovery. Journal of Machine Learning Research, 2006, 7: 2003-2030

[25] Zhang Kun, Hyvärinen A. On the identifiability of the post-nonlinear causal model//Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI 2009). Montreal, Canada, 2009: 647-655

[26] Zhang Kun, Chan Lai-Wan. Extensions of ICA for causality discovery in the Hong Kong stock market//Proceedings of the 13th International Conference on Neural Information (ICONIP). Hong Kong, China, 2006: 400-409

[27] Hoyer P O, Janzing D, Mooij J M, et al. Nonlinear causal discovery with additive noise models//Proceedings of the 23rd Annual Conference on Neural Information Processing Systems (NIPS 2009). Vancouver, Canada, 2009: 689-696

[28] Daniusis P, Janzing D, Mooij J, et al. Inferring deterministic causal relations//Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI 2010). Catalina Island, USA, 2010: 143-150

[29] Janzing D, Mooij J, Zhang Kun, et al. Information-geometric approach to inferring causal directions. Artificial Intelligence, 2012, 182(5): 1-31

[30] Peters J, Janzing D, Schölkopf B. Causal inference on discrete data using additive noise models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(12): 2436-2450

[31] Verma T, Pearl J. Equivalence and synthesis of causal models//Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence. Cambridge, UK, 1990: 255-268

[32] Bromberg F, Margaritis D. Improving the reliability of causal discovery from small data sets using argumentation. Journal of Machine Learning Research, 2009, 10(1): 301-340

[33] He Yang-Bo, Geng Zhi. Active learning of causal networks with intervention experiments and optimal designs. Journal of Machine Learning Research, 2008, 9(11): 2523-2547

[34] Yin Jian-Xin, Zhou You, Wang Chang-Zhang, et al. Partial orientation and local structural learning of causal networks for prediction//Proceedings of the WCCI Causation and Prediction Challenge. Hong Kong, China, 2008: 93-105

[35] Zhou You, Wang Changzhang, Yin Jianxin, Geng Zhi. Discover local causal network around a target to a given depth//Guyon I, Janzing D, Schölkopf B eds. NIPS Causality: Objectives and Assessment. Cambridge, USA: MIT Press, 2010: 191-202

[36] Tsamardinos I, Brown L E, Aliferis C F. The max-min hill-climbing Bayesian network structure learning algorithm. Machine Learning, 2006, 65(1): 31-78

[37] Tsamardinos I, Aliferis C F, Statnikov A. Time and sample efficient discovery of Markov blankets and direct causal relations//Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, USA, 2003: 673-678

[38] Spirtes P, Meek C, Richardson T. An algorithm for causal inference in the presence of latent variables and selection bias//Cooper G F, Glymour C eds. IEEE Transactions on Pattern Analysis and Machine Intelligence. Cambridge, USA: MIT Press, 1999: 211-252

[39] Colombo D, Maathuis M H, Kalisch M, et al. Learning high-dimensional directed acyclic graphs with latent and selection variables. The Annals of Statistics, 2012, 40(1): 294-321

[40] Kummerfeld E, Ramsey J, Yang R, et al. Causal clustering for 2-factor measurement models//Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2014). Nancy, France, 2014: 34-49

[41] Shimizu S, Inazumi T, Sogawa Y, et al. DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. Journal of Machine Learning Research, 2011, 12(2): 1225-1248

[42] Peters J, Mooij J M, Janzing D, et al. Causal discovery with continuous additive noise models. Journal of Machine Learning Research, 2014, 15(1): 2009-2053

[43] Tashiro T, Shimizu S, Hyvärinen A, et al. ParceLiNGAM: A causal ordering method robust against latent confounders. Neural Computation, 2014, 26(1): 57-83

[44] Hoyer P O, Shimizu S, Kerminen A J, et al. Estimation of causal effects using linear non-Gaussian causal models with hidden variables. International Journal of Approximate Reasoning, 2008, 49(2): 362-378

[45] Zhang Kun, Schölkopf B, Janzing D. Invariant Gaussian process latent variable models and application in causal discovery//Proceedings of the 26th Conference (UAI 2010). Corvallis, USA, 2010: 717-724

[46] Lawrence N. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. Journal of Machine Learning Research, 2005, 6(3): 1783-1816

[47] Chaves R, Luft L, Maciel T O, et al. Inferring latent structures via information inequalities//Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence (UAI 2014). Quebec City, Canada, 2014: 112-121

[48] Chaves R, Majenz C, Gross D. Information—theoretic implications of quantum causal structures. Nature Communications, 2015, 6: 5766

[49] Cai Rui-Chu, Zhang Zhen-Jie, Hao Zhi-Feng. SADA: A general framework to support robust causation discovery//Proceedings of the 30th International Conference on Machine Learning. Atlanta, USA, 2013: 208-216

计算机学报 录用稿件

[50] Chen Wei-Qi, Hao Zhi-Feng, Cai Rui-Chu, et al. Multiple-cause discovery combined with structure learning for high-dimensional discrete data and application to stock prediction. Soft Computing, 2016, 20(11): 4575-4588

[51] Hao Zhi-Feng, Huang Jin-Long, Cai Rui-Chu, et al. A hybrid approach for large scale causality discovery//Proceedings of the 9th International Conference on Intelligent Computing (ICIC 2013). Nanning, China, 2013: 1-6

[52] Glymour C, Scheines R, Spirtes P, et al. Discovering Causal Structure: Artificial Intelligence, Philosophy of Science, and Statistical Modeling. San Diego, USA: Academic Press, 1987

[53] McCallum A. Efficiently inducing features of conditional random fields//Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence. Acapulco, Mexico, 2002: 403-410

[54] Cooper G F, Herskovits E. A Bayesian method for the induction of probabilistic networks from data. Machine Learning, 1992, 9(4): 309-347

[55] Pearl J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. San Mateo, USA: Morgan Kaufmann, 1988

[56] Spirtes P, Glymour C. An algorithm for fast recovery of sparse causal graphs. Social Science Computer Review, 1991, 9(1): 62-72

[57] Cai Rui-Chu, Zhang Zhen-Jie, Hao Zhi-Feng. Causal gene identification using combinatorial V-structure search. Neural Networks, 2013, 43: 63-71

[58] Chickering D M. Learning equivalence classes of Bayesian-network structures. Journal of Machine Learning Research, 2002, 2(3): 445-498

[59] Andersson S A, Madigan D, Perlman M D. A characterization of Markov equivalence classes for acyclic digraphs. The Annals of Statistics, 1997, 25(2): 505-541

[60] He Yang-Bo, Jia Jin-Zhu, Yu Bin. Counting and exploring sizes of Markov equivalence classes of directed acyclic graphs. Journal of Machine Learning Research, 2015, 16(1): 2589-2609

[61] Sun X, Janzing D, Schölkopf B, et al. A kernel-based causal learning algorithm//Proceedings of the 24th International Conference on Machine Learning (ICML 2007). Corvallis, USA, 2007: 855-862

[62] Zhang X, Song L, Gretton A, et al. Kernel measures of independence for non-IID data//Proceedings of the 23rd Annual Conference on Neural Information Processing Systems (NIPS 2009). Vancouver, Canada, 2009: 1937-1944

[63] Zhang Kun, Peters J, Janzing D, et al. Kernel-based conditional independence test and application in causal discovery//Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI 2011). Barcelona, Spain, 2011: 804-813

[64] Geng Zhi, Wang Chi, Zhao Qiang. Decomposition of search for v-structures in DAGs. Journal of Multivariate Analysis, 2005, 96(2): 282-294

[65] Xie Xian-Chao, Geng Zhi, Zhao Qiang. Decomposition of structural learning about directed acyclic graphs. Artificial Intelligence, 2006, 170(4): 422-439

[66] Xie Xian-Chao, Geng Zhi. A recursive method for structural learning of directed acyclic graphs. Journal of Machine Learning Research, 2008, 9(3): 459-483

[67] Zhang Ji-Ji. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. Artificial Intelligence, 2008, 172(16): 1873-1896

[68] Zhao Hui, Zheng Zhong-Guo, Xu Jing. Causal inference in the models with hidden variables and selection bias. Acta Scientiarum Naturalium Universitatis Pekinensis, 2006, 42(5): 584-589(in Chinese)
(赵慧，郑忠国，许静. 含隐变量和选择偏差的图模型中的因果推断. 北京大学学报：自然科学版，2006，42(5)：584-589)

[69] Pearl J. On the testability of causal models with latent and instrumental variables//Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence (UAI 1995). Montreal, Canada, 1995: 435-443

[70] Bonet B. Instrumentality tests revisited//Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence (UAI 2001). Seattle, USA, 2001: 48-55

[71] Glymour M M, Tchetgen E J T, Robins J M. Credible mendelian randomization studies: Approaches for evaluating the instrumental variable assumptions. American Journal of Epidemiology, 2012, 175(4): 332-339

[72] Evans R J. Graphical latent structure testing//Carpita M, Brentari E, Qannari E M eds. Advances in Latent Variables: Methods, Models and Applications. Switzerland: Springer International Publishing, 2014: 253-262

[73] Shimizu S. LiNGAM: Non-Gaussian methods for estimating causal structures. Behaviormetrika, 2014, 41(1): 65-98

[74] Hyvärinen A, Karhunen J, Oja E. Independent Component Analysis. New York, USA: John Wiley & Sons, Inc., 2001

[75] Cai Rui-Chu, Zhang Zhen-Jie, Hao Zhi-Feng. BASSUM: A Bayesian semi-supervised method for classification feature selection. Pattern Recognition, 2011, 44(4): 811-820

[76] Darmois G. Analyse générale des liaisons stochastiques: Etude particulière de l'analyse factorielle linéaire. Revue De L'Institut International De Statistique/Review of the International Statistical Institute, 1953, 21(1/2): 2-8

[77] Skitovich V P. On a property of the normal distribution. Doklady Akademii Nauk SSSR, 1953, 89: 217-219

[78] Hyvärinen A, Smith S M. Pairwise likelihood ratios for estimation of non-Gaussian structural equation models. Journal of Machine Learning Research, 2013, 14(1): 111-152

[79] Zhang K, Hyvärinen A. Distinguishing causes from effects using nonlinear acyclic causal models//Proceedings of the NIPS 2008 Workshop on Causality. Vancouver, Canada, 2008: 157-164

[80] Zhang Kun, Zhang Ji-Ji, Huang Bi-Wei, et al. On the identifiability and estimation of functional causal models in the presence of outcome-dependent selection//Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence (UAI 2016). New York, USA, 2016: 825-834

[81] Rasmussen C E, Williams C K I. Gaussian Processes for Machine Learning. Cambridge, USA: MIT Press, 2006

《计算机学报》版权

[82] Gretton A, Herbrich R, Smola A, et al. Kernel methods for measuring independence. Journal of Machine Learning Research, 2005, 6(12): 2075-2129

[83] Mooij J M, Peters J, Janzing D, et al. Distinguishing cause from effect using observational data: Methods and benchmarks. Journal of Machine Learning Research, 2016, 17(32): 1-102

[84] Sgouritsa E, Janzing D, Hennig P, et al. Inference of cause and effect with unsupervised inverse regression//Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS 2015). San Diego, USA, 2015: 847-855

[85] Hyvarinen A. Fast and robust fixed-point algorithms for independent component analysis. IEEE Transactions on Neural Networks, 1999, 10(3): 626-634

[86] Rasmussen C E, Nickisch H. Gaussian processes for machine learning (GPML) toolbox. Journal of Machine Learning Research, 2010, 11(6): 3011-3015

[87] Kalisch M, Mächler M, Colombo D, et al. Causal inference using graphical models with the R package pcalg. Journal of Statistical Software, 2012, 47(11): 1-26

[88] Koller D. Toward optimal feature selection//Proceedings of the 13th International Conference on Machine Learning. Bari, Italy, 1996: 284-292

[89] Liu Mei, Cai Rui-Chu, Hu Yong, et al. Determining molecular predictors of adverse drug reactions with causality analysis based on structure learning. Journal of the American Medical Informatics Association, 2013, 21(2): 245-251

[90] Wang Ming-Feng, Geng Zhi, Wang Mi-Qu, et al. Combination of network construction and cluster analysis and its application to traditional Chinese medicine//Proceedings of the 3rd International Symposium on Neural Networks (ISSN 2016). Chengdu, China, 2006: 777-785

[91] Beinlich I A, Suermondt H J, Chavez R M, et al. The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks//Proceedings of the 2nd European Conference on Artificial Intelligence in Medicine. London, UK, 1989: 247-256

[92] Zhou Tong, Wang Ya-Li. Causal relationship inference for a large-scale cellular network. Bioinformatics, 2010, 26(16): 2020-2028

[93] Dong Chun-Ling, Wang Yan-Jun, Zhang Qin, et al. The methodology of dynamic uncertain causality graph for intelligent diagnosis of vertigo. Computer Methods and Programs in Biomedicine, 2014, 113(1): 162-174

[94] Hu Yong, Liu Kang, Zhang Xiang-Zhou, et al. Concept drift mining of portfolio selection factors in stock market. Electronic Commerce Research and Applications, 2015, 14(6): 444-455

[95] Zhang Xiang-Zhou, Hu Yong, Xie Kang, et al. A causal feature selection algorithm for stock prediction modeling. Neurocomputing, 2014, 142(1): 48-59

[96] Fire A, Zhu Song-Chun. Learning perceptual causality from video. ACM Transactions on Intelligent Systems and Technology, 2015, 7(2): 23

[97] Hours H, Biersack E, Loiseau P. A causal approach to the study of TCP performance. ACM Transactions on Intelligent Systems and Technology, 2015, 7(2): 25

[98] Hu Yong, Zhang Xiang-Zhou, Ngai E W T, et al. Software project risk analysis using Bayesian networks with causality constraints. Decision Support Systems, 2013, 56(1): 439-449

[99] Hao Zhi-Feng, Chen Wei, Cai Rui-Chu, et al. Performance optimization of wireless network based on canonical causal inference algorithm. Journal of Computer Applications, 2016, 36(8): 2114-2120(in Chinese)
(郝志峰, 陈薇, 蔡瑞初等. 基于典型因果推断算法的无线网络性能优化. 计算机应用, 2016, 36(8): 2114-2120)

[100] Kummerfeld E, Ramsey J. Causal clustering for 1-factor measurement models//Proceedings of the 22th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, USA, 2016: 1655-1664

[101] Borboudakis G, Tsamardinos I. Towards robust and versatile causal discovery for business applications//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, USA, 2016: 1435-1444

[102] Kusner M J, Sun Y, Sridharan K, et al. Private causal inference//Proceedings of the 19th International Conference on Artificial Intelligence and Statistics. Cadiz, Spain, 2016: 1308-1317

[103] Zafarani R, Liu H. Evaluation without ground truth in social media research. Communications of the ACM, 2015, 58(6): 54-60

[104] Schölkopf B, Janzing D, Peters J, et al. On causal and anticausal learning//Proceedings of the 29th International Conference on Machine Learning (ICML 2012). Edinburgh, Scotland, 2012: 1-8

[105] Zhang Kun, Gong Ming-Ming, Schölkopf B. Multi-source domain adaptation: A causal view//Proceedings of the 29th AAAI Conference on Artificial Intelligence. Austin Texas, USA, 2015: 3150-3157

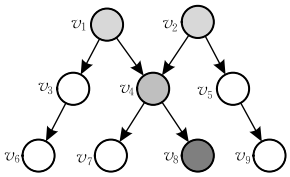
[106] Zhang Kun, Muandet K, Wang Zhi-Kun. Domain adaptation under target and conditional shift//Proceedings of the 30th International Conference on Machine Learning. Atlanta, USA, 2013: 819-827

[107] Gong M, Zhang Kun, Liu T, et al. Domain adaptation with conditional transferable components//Proceedings of the 33rd International Conference on Machine Learning (ICML 2016). New York, USA, 2016: 2839-2848

附 录. 专有名词基本定义

(1)因果关系网络图(Causal Network Graphs). 变量之间的因果关系可以通过一个有向无环图来表示,这个图即称为因果关系网络图. 图的每个节点表示每个变量,两个节点之间的边表示两个节点之间存在因果关系,边的出度节点为原因节点(或原因变量),边的箭头所指向的节点为结果节点(或结果变量). 附图 1 为一个因果关系网络图的简单例子,其中变量 v_1 和 v_4 之间存在一条边,说明这两个变量之间存在因果关系,变量 v_1 为原因变量,变量 v_4 为结果变量.

《计算机学报》版权



附图 1　因果关系网络图

(2)父亲节点(Parent node). 在因果关系网络图中,父亲节点是指一个节点的直接原因节点.每个节点的所有直接原因节点即为该节点的父亲节点集.例如,附图 1 中变量 v_8 的直接原因节点有 v_4,所以其父亲节点集为$\{v_4\}$.

(3)祖先节点(Ancestor node). 在因果关系网络图中,祖先节点是指一个节点的直接或间接原因节点.每个节点的所有(直接和间接的)原因节点即为该节点的祖先节点集.例如,附图 1 中变量 v_8 的直接原因节点为 v_4,间接原因节点为 v_1 和 v_2,所以其祖先节点集为$\{v_1, v_2, v_4\}$.

(4)d-分离(d-separation). d-分离准则可以在因果图中形式化描述变量间的独立性关系. d-分离准则的定义如下.其示意图如图 4 所示.

假设在无向图 G 中存在一个变量集 $\boldsymbol{V}'$,且 v_1 和 v_3 都不是 $\boldsymbol{V}'$中的两个变量,α 表示 v_1 和 v_3 之间的一条通路,当路径 α 满足以下条件之一时,称 $\boldsymbol{V}'$ d-分离 v_1 和 v_3,即 $\boldsymbol{V}'$阻断了 v_1 到 v_3 的所有通路:

① α 包含一种顺连 $v_1 \rightarrow v_2 \rightarrow v_3$ 或一种分连 $v_1 \leftarrow v_2 \leftarrow v_3$,且 $v_2 \in \boldsymbol{V}'$;

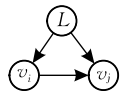
② α 包含一种汇连 $v_1 \rightarrow v_2 \leftarrow v_3$,且 v_2 及其后代都不在 $\boldsymbol{V}'$里.

例如,按照 d-分离的定义,可以看出,附图 1 中 v_4 d-分离 v_1 和 v_8.

(5)内生变量(Endogenous variable). 内生变量是指因果模型被解释的变量,即在因果模型内部有直接的原因变量.它的特性描述依赖于变量集和所选择的因果模型.在因果关系网络图中,它形式化表现为有箭头指向它的变量.内生变量就是模型要研究的变量.

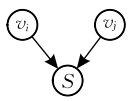
(6)外生变量(Exogenous variable). 外生变量是指在因果模型中只起解释作用的变量,即在因果模型内部没有直接的原因变量.它的特性描述依赖于变量集和所选择的因果模型.在因果关系网络图中,它形式化表现为只有指向其他节点的箭头的而没有箭头指向它的节点.此名词与"内生变量"相对应.

(7)混淆因子(Confounder variable). 变量 v_i 和变量 v_j 的共同原因变量 L 未被检测到,则称 L 为混淆因子.如附图 2 所示,其中变量 L 即为混淆因子.



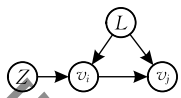
附图 2　混淆因子

(8)选择偏倚(Selection bias). 变量 v_i 和变量 v_j 的共同结果变量 S 影响了抽样机制(例如只抽取了 S 中的一个子样例),导致了 S 的分布未被观测到,则称 S 为选择偏倚.如附图 3 所示,其中变量 S 为选择偏倚.



附图 3　选择偏倚

(9)工具变量(Instrumental variable). 通过对 v_i 进行"自然操作"后用于测试因果关系 $v_i \rightarrow v_j$ 的变量,称为工具变量.这个工具变量是外生的,且除了通过变量 v_i 之外,不能通过任何方式影响变量 v_j.如附图 4 中的变量即为工具变量,其满足在变量 v_i 的前提下变量 Z 和变量 v_j 是独立的.



附图 4　含有工具变量的因果关系网络图(其中节点 Z 表示工具变量,满足 $Z \perp v_j \mid v_i$)

计算机学报 版权

CAI Rui-Chu, born in 1983, Ph.D., professor. His research interests include causality, machine learning, etc.

CHEN Wei, born in 1993, postgraduate student. Her research interest is focused on causal discovery and its applications.

ZHANG Kun, born in 1980, Ph. D., assistant professor. His main research interests include causal discovery, machine learning and so on.

HAO Zhi-Feng, born in 1968, Ph. D., professor. His research interests involve machine learning and data mining.

Background

Causal discovery on the non-temporal series observational data plays a crucial role on a variety of scientific domains, and it generalizes the prediction tasks commonly studied by machine learning. With numerous opportunities for scientific discovery, it commonly seen as the core of data science and has attracted many attentions from Artificial Intelligence. Unlike the mainstream statistical learning approaches, causality learning tries to understand the data generation procedure, rather than characterizing the joint distribution of the observed variables only. There remain open and hot research topics on

causal discovery regarding how to improve the discoverability and reduce the false discovery rate on the high dimensional incomplete observational data, which are generally applicable on disease-causal gene discovery, adverse drug reaction mining and other real world applications.

In this work, we summarize the existing literatures of causal discovery methods on these challenges from four aspects which include causal models and assumptions, the constraint based approaches, the casual function based approaches and the hybrid approaches. Based on the survey of the existing literatures, we also point out some possible future research directions.

This research is supported in part by the NSFC-Guangdong Joint Found under Grant No. U1501254, the National Natural Science Foundation of China under Grant No. 61572143, and the Guangdong Science Fund for Distinguished Young Scholars under Grant No. 2014A030306004. This work can be viewed as an investigation on causal discovery methods on the non-temporal series observational data for these projects. The first project aims to lay the foundation of massive information processing and intelligent computing theory in smart city, the second project targets to further extend the existing causal inference theories, and propose effective methods for high dimensional incomplete observational data, and the final project focuses on the theory foundations of the causal discovery.

Our groups have been working on observational data based causal discovery algorithms for more than 5 years, and have a lot of important works published on top conferences and journals, such as ICML, NIPS, AAAI, UAI, TPAMI, TKDE, and Neural Networks. Selected 10 papers are listed as follows:

[1] Cai Rui-Chu, Zhang Zhen-Jie, Hao Zhi-Feng, Winslett M. Understanding social causalities behind human action sequences. IEEE Transactions on Neural Network and Learning Systems, 2016, Accepted

[2] Zhang Kun, Wang Zhi-Kun, Zhang Ji-Ji, et al. On estimation of functional causal models: General results and application to the post-nonlinear causal model. ACM Transactions on Intelligent Systems and Technology, 2016, 7(2): 1-22

[3] Zhang Kun, Gong Ming-Ming, Schölkopf B. Multi-source domain adaptation: A causal view//Proceedings of the 29th AAAI Conference on Artificial Intelligence. Austin Texas, USA, 2015: 3150-3157

[4] Hao Zhi-Feng, Zhang Hao, Cai Rui-Chu, et al. Causal discovery on high dimensional data. Applied Intelligence, 2015, 42(3): 594-607

[5] Cai Rui-Chu, Zhang Zhen-Jie, Hao Zhi-Feng. SADA: A general framework to support robust causation discovery//Proceedings of the 30th International Conference on Machine Learning. Atlanta, USA, 2013: 208-216

[6] Cai Rui-Chu, Zhang Zhen-Jie, Hao Zhi-Feng. Causal gene identification using combinatorial V-structure search. Neural Networks, 2013, 43: 63-71

[7] Zhang Kun, Muandet K, Wang Z. Domain adaptation under target and conditional shift//Proceedings of the 30th International Conference on Machine Learning. Atlanta, USA, 2013: 819-827

[8] Zhang Kun, Peters J, Janzing D, et al. Kernel-based conditional independence test and application in causal discovery//Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence. Corvallis, OR, USA, 2012: 804-813

[9] Zhang Kun, Schölkopf B, Janzing D. Invariant Gaussian process latent variable models and application in causal discovery//Proceedings of the 26th Conference (UAI 2010). Corvallis, USA, 2010: 717-724

[10] Zhang Kun, Hyvärinen A. On the identifiability of the post-nonlinear causal model//Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI 2009). Montreal, Canada, 2009: 647-655

计算机学报版权