计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2016.20148251
53(7):1544-1560,2016

ICIC_Target:目标节点的局部因果关系网络的发现算法

李 岩[1] 王 挺[1] 刘万伟[1] 张晓艳[2]
[1](国防科学技术大学计算机学院 长沙 410073)
[2](国防科学技术大学人文与社会科学学院 长沙 410073)
(liyannayil@nudt.edu.cn)

ICIC_Target: A Novel Discovery Algorithm for Local Causality Network of Target Variable

Li Yan[1], Wang Ting[1], Liu Wanwei[1], and Zhang Xiaoyan[2]
[1](*College of Computer, National University of Defense Technology, Changsha* 410073)
[2](*College of Humanities and Social Sciences, National University of Defense Technology, Changsha* 410073)

Abstract Causality research aims to reveal the law of evolution of nature, society and human. Nowadays, the causality research receives widespread attention for its important applications of human life and science research, but there are still many difficulties and challenges. This paper presents a unified model to explain the stimulating and inhibiting causalities. Based on this model, we also present a framework ICIC and a novel algorithm ICIC_Target to infer the local causal structure of a target variable from observational data without any limitation of some assumptions, such as assumption of acyclic structure, hidden variables and so on. Following our descriptions of causality essence and properties, as well as several classical theories proposed by Judea Pearl, Gregory F. Cooper and so on, we introduce concepts of exogenous variable and clique-like structure (IClique) to get rough ordering of variables, which are necessary for revealing the causality accurately and efficiently. To evaluate our approach, several experiments compared with HITON, IC, PC, PCMB and several methods based on four datasets with different data types have been done. The results demonstrate the higher performance and stronger robustness of our algorithm ICIC_Target. In this paper, we also discuss the advantages of stability and complexity of ICIC_Target.

Key words causality; causal network; local causal network discovery; stimulating causality; inhibiting causality

摘 要 因果关系的研究在于揭示自然规律的和人类社会发展本质及其规律,对人类长久以来的生产生活和科学研究有着非常重要的作用.目前,因果关系的研究受到前所未有的广泛关注,但仍存在诸多困难和挑战.致力于建立一个因果激励/抑制模型以抽象地表示和解释因果的作用机制,并在此基础上提出用于目标节点的局部因果关系网络的自动发现方法框架 ICIC 和算法 ICIC_Target.该方法不预先设定因果结构(如设定为无圈、隐含结构),并根据对因果关系本质的认识,利用初始变量(exogenous variables)和初始团树(IClique)的概念,在判定边和方向之前对变量进行粗略地排序,从而提高了因果关系网络发现的性能.在 4 个不同类型的数据集上实现了与多种经典方法,如 HITON,IC,PC,PCMB

收稿日期:2014-11-14;**修回日期**:2015-09-06
基金项目:国家自然科学基金项目(61170287,60873097)
This work was supported by the National Natural Science Foundation of China (61170287,60873097).

等的对比实验,实验结果表明 ICIC_Target 方法适用范围广,有较好的鲁棒性,同时,从理论上分析证实了 ICIC_Target 方法具有较好的稳定性和较低的复杂度.

关键词　因果关系;因果关系网络;局部因果关系发现;激励因果;抑制因果

中图法分类号　TP391

因果关系的研究在于揭示自然和人类社会发展本质及其规律,以解释现象、控制存在、预测未来,对人类的生产生活和科学研究有着非常重要的作用.因果关系的研究历史悠久,研究方法经历了"手工"、"半自动"和"全自动"的发展过程.自 20 世纪 80 年代以来,许多数学理论和方法相继提出,因果关系研究正式成为可计算的研究领域.Pearl[1] 和 Spirtes 等人[2] 指出因(cause)和果(effect)可以是事件(event)、属性(feature)、物体(object)等多种类型的实体(entities).为采样的可行和计算的高效,该领域研究普遍使用不同类型的随机变量(variables)表示因果实体,变量的取值对应于实体的数值或状态值.概率统计和图模型普遍应用于发现和表示因果关系.因为概率可以很好地契合变量的随机性、因果关系发现过程的不确定性和因果断言的语义.有向无圈图(directed acyclic graph, DAG)或贝叶斯网络(Bayesian network)这种最常用的图模型,可以生动准确地表达因果关系结构,其中网络节点表示变量系统(简称系统),节点间的边表示因果关系的存在,边的方向表示因果关系的存在方式.但是贝叶斯网络无法表示多变量间循环作用(圈)或 2 个变量间相互作用(双向边)的因果关系,而这些因果关系是真实存在且常见的,因此本文采用可以包容圈和双向边的宽松结构(loose network)表示真实的因果关系网络.

因果关系网络结构的发现是因果关系研究的重要内容,即给定有限数量和范围的实体及其采样值,发现全局因果关系网络或关于目标实体的局部因果关系网络.大量真实环境下的因果关系发现问题(causal problem, CP)仅仅需要确定给定的目标变量周围的局部因果网络结构,还有许多问题涉及的全局网络结构过于庞大和复杂,全局网络结构的发现成了非常困难的任务.因此发现给定目标变量的若干层因果关系网络成为了因果关系发现的常见任务,记为 $CP(V_o,v_t,D)$,即在已知数据集 D 上找出给定目标变量 v_t 在可观测变量 V_o 中的父、子、配偶等变量以及这些变量之间的因果联系.

Pearl[1],Spirtes 等人[2],Cooper 等人[3-4] 在各自经典著作中讨论了如何在不同的应用背景和数据类型下正确高效发现因果关系,其中最核心的理论部分是因果关系的定义(尤其是 Granger[5] 和 Pearl[1] 提出的定义)、Markov 条件、d-割准则(d-separation criterion)和忠实性定理(faithfulness theorem).其中 Markov 条件独立是贝叶斯网络表示因果关系应满足的基本假设,也是本文提出的宽松结构在表示有向无圈的真实网络结构时应满足的假设.

目前,因果关系的研究受到广泛关注[1,6],但由于环境、人为因素的复杂多变并且不可控,真实环境下的因果关系发现仍存在诸多困难和挑战.这是因为:真实环境下因果关系发现获得的数据来自人们对可观测属性"静态"或"被动"的观测[4],是真实环境下随机因素与因果关系共同作用的结果,并按照某种方式在某些时刻采样获得,是关于真实世界的"静态"片段,因此无法像在实验室环境下通过干预(intervention 或 manipulation)某些属性的取值从而获得其他属性取值变化的"动态"数据,也无法像使用实验室数据那样可以根据先验知识对网络结构做出合理的假设并做出验证;真实环境下因果关系还可能受到若干不可观测因素(hidden features)的影响,仅考虑那些可观测属性(observational features)的因果关系网络则不能完全正确地反映真实因果(underlying causality mechanism).如何仅从"静态"观测数据还原真实的因果关系网络,且不做任何关于网络结构的先验假设,如是否存在圈或隐含变量的假设,是非常困难的,甚至被认为是在现有理论框架下不可能完成的任务.

针对以上问题,本文基于 Pearl 等人提出的因果关系经典理论和方法,通过分析因果关系发生和作用的本质特性,利用初始变量(exogenous or instrumental variable)、初始团树(initial clique of an exogenous variable, IClique)和 Markov 条件独立判定来解决真实环境下局部因果关系发现问题(详见 2.1 节),并提出了性能更高、鲁棒性更强的局部因果关系网络结构发现方法 ICIC_Target(详见 2.3 节).该方法主要用于目标节点的因果关系网络的发现,同时针对二值数据系统区别了激励因果(stimulating

causality)和抑制因果(inhibiting causality)两种因果模型,发现了一些新型的因果关系.不同于 IC 和 PC 等经典方法假设因果网络为无圈和无隐含变量(hidden variable),ICIC_Target 方法不预先设定因果结构,从而提高了方法的性能和鲁棒性. ICIC_Target 不仅在理论上具有可靠性、稳定性,在多数据集、多评估体系下与多种经典方法实验结果的比较同样具有优越的性能.另外,本文将因果研究常用的小品集 LUCAS(lung cancer simple set)[7-8]作为应用样本以清晰展示 ICIC_Target 方法的细节.

1 相关工作

因果关系研究正式成为可计算的研究领域后,各种因果表示模型、因果发现预测方法大量出现,并逐渐形成两大学派:以图灵奖获得者 Pearl[1]为代表人物的 UCLA & CMU 学派和 Cooper 等人[9]为代表的 Stanford 学派,他们出版于 1999 年的专著为因果关系研究奠定了坚实基础.

1.1 IC,PC,LCD 和 IC*,FCI 方法

对于多变量系统,Pearl 等人的研究思路是从找出条件独立的片断开始,最后将各个片断合成为全局的因果关系结构,最具代表性的经典方法有 Pearl[1]的 IC(inductive causation)方法和 Spirtes 等人[2]的 PC 方法.而 Cooper 等人[4,9]的研究思路是对于给定的数据和参数,先为一个贝叶斯网络(即候选因果网络)指派先验概率,然后根据因果关系网络与数据的拟合度打分,在可能的结构空间上搜索具有最大后验概率的结构,用于更新网络,以达到最优. Cooper 等人[4]提出的 LCD(local causal discovery)属于此类方法.该方法需要输入一个或多个初始变量作为因果网络结构的根节点,如性别、年龄等变量是研究疾病问题公认的初始变量.这样可以充分利用问题本身提供的关于网络结构的先验知识,实验证明初始变量有助于提高发现的准确性和效率[4],但对于大量无法预知初始变量的问题,LCD 方法的局限性显而易见. Stanford 学派的研究思路应用广泛,但 Pearl[1]认为此类方法有很高的复杂性,事实上只能从局部到局部最优且不适用于小样本数据,不能灵活处理存在不确定、不可观测变量的情况,并且该研究思路假定参数独立.而 Pearl 等人的 IC,PC 方法复杂度相对较低,对样本数据大小要求不高.

以上方法均不适用于存在隐含变量的因果关系发现问题.因此,Pearl[1]和 Spirtes 等人[2]提出了 IC*(inductive causation with latent variable)和 FCI(fast causal inference algorithm)方法.隐含变量的存在导致可观测变量之间的因果关系发生不同程度的变化,因此在 IC* 和 FCI 方法中修改了边的方向表示,使用双向边、"*"或"o"等符号以增强边所能表达的语义.根据边的语义变化,IC* 和 FCI 方法在 IC 和 PC 方法上调整了因果的判定规则并做了其他较大改进,以适应可能存在隐含变量的情况,同时改善了算法的性能. IC* 和 FCI 方法与本文提出的 ICIC_Target 方法都适用于隐含变量因果关系发现问题,但 IC* 和 FCI 发现的网络结构的语义复杂、算法复杂度高.

1.2 PCMB,IAMB,HITON 系列方法

HITON-PC(HITON_parents & children)和 HITON-MB(HITON_Markov blanket)方法是 Aliferis 等人[10]于 2003 年提出经典的特征选择方法,用于发现目标节点(target variable)的父子节点(记为 PCset)和包含目标节点及其父子节点和配偶节点的 Markov 毯(Markov blanket/boundary, MBset),这些节点对于预测目标节点的取值有着重要的作用,因此 HITON 也可看作是局部因果关系发现方法. 2006 年 Tsamardinos 等人[11]提出 HITON 等方法缺少"对称性校正(symmetry correction)",在理论上不正确.但 HITON-PC 方法仍成为 NIPS 2008 举办的第 2 届 Causality Challenge 中 LOCANET (uncover the local causal network around the target)任务的标准(state-of-the-art)因果发现算法,用以获得质量保证测试的基准结果(baseline results)[12]. Aliferis 等人[13]在 2010 年的综述文章中解释说 HITON 使用了启发式搜索方法作为后处理步骤(a wrapping post-processing),在原理上可以去掉错误的正例,因此未做对称性校正的 HITON 方法仍具有优越的性能. 2007 年 Peña 等人[14]提出了理论上正确的 PCMB 方法,用于发现 PCset 或 MBset. Peña 等人将 PCMB 方法与同样在理论上正确但数据低效(data inefficient)的 IAMB 方法进行了比较.

美国 Vanderbilt 大学的生物医学信息系开发的因果关系发现工具 Causal Explorer library[15-16],用 matlab 实现了 HITON,IAMB 等多种经典方法. Peña 用 C++ 实现了 PCMB 和 IAMB 方法[17].然而这些方法的输出是 PCset 或 MBset 而非网络结构,因此无法表达更详细更准确的因果作用机制.

1.3 GC 系列方法

早在 1969 年,诺贝尔经济学奖获得者 Granger[5]

就给出了时序问题的因果关系(Granger causality,GC)定义:若用时间序列 U 的历史信息预测时序变量 X 好于在 U 中排除时序 Y 后的历史信息预测 X 的结果,则 Y 是 X 原因,记为 $Y\rightarrow X$. Granger 认为"因"有助于提高对"果"预测的性能. GC 含义明确、可操作性强,被时序问题研究领域的研究人员所接受,并在此基础上做了很多改进,取得了大量成果[18-20]. 其中具有代表性的是 2004 年东京大学的研究[18],将 GC 中的方差改进为转移熵(transfer entropy),并将因果关系描述为:Y 是 X 的原因,若时序 Y 的历史信息为预测时序 X 提供了显著的帮助. 这种改进的因果关系定义和 GCTE 方法可以帮助研究不同类型的时序变量间的因果关系.

但是"利于预测"并不是因果关系的本质,为此 Pearl[1] 反驳道 GC 应该归为统计学范畴而非因果关系. 因为在不能明确时序时,满足 GC 的 X 和 Y 之间的方向不明确. 即使 $Y\rightarrow X$,此时已知 X 的历史信息 Y 也会被更好地预测,即 $X\rightarrow Y$ 同样成立.

1.4　其他方法

2006 年 Shimizu 等人[21]提出了一种全新的因果关系发现方法 LiNGAM(linear non-Gaussian acyclic model). 该方法利用独立成份分析的方法估计方程组 $\boldsymbol{x}=\boldsymbol{Ae}$ 中的混合矩阵 $\boldsymbol{A}$,以获得变量间相互作用的线性方程,其中 $\boldsymbol{x}$ 是各变量的观测值组成的矩阵,$\boldsymbol{e}$ 由各变量自带的随机扰动(disturbance variables)组成. LiNGAM 计算 $\boldsymbol{A}$ 使得 $\boldsymbol{e}$ 的分量尽可能满足两两独立,但前提是这些随机扰动是两两独立且服从非高斯分布,同时因果关系非循环(无圈结构)且满足线性方程,否则方法可能会产生错误结果.

SI(similarity index)是 Arnhold 等人[22]提出的另一类针对时序数据的因果关系发现方法,通过计算 2 个时间序列在同一起始时刻开始的相同长度片断上的平均相似程度来判断时序变量是否存在因果关系. 我们认为该方法对于同步和异步变化的 2 个时间序列的相似判定都非常适用,但因果关系的方向难以判断,并且需要给定参数 R,该参数与算法的性能和复杂度密切相关. 文献[23]指出 SI 的最大缺点是很难判定弱因果结构和带噪音时序的因果.

2　因果关系发现方法 ICIC_Target

针对真实环境下局部因果关系发现,即未知真实网络是否存在隐含变量(IC,PC 方法假设网络不含隐含变量)、是否是无圈网络(IC,PC,LiNGAM 方法均假设网络为 DAG)、是否存在初始变量(LCD 方法不适用无初始变量的问题)、是否带噪音(LiNGAM 方法假设噪音两两独立且服从非高斯分布,SI 方法对带噪音时序的因果判定困难)等,为提高因果发现性能(GC 方法易产生冗余关系,HITON 等方法没有明确网络结构)、增强算法的鲁棒性(IC 等方法适用于案例数据,GC 和 SI 方法适用于时序数据),本节通过分析因果关系发生和作用的本质特性,利用初始变量和初始团树,提出了性能更高、鲁棒性更强的局部因果关系网络结构发现方法 ICIC_Target,并形式化介绍该方法的框架和技术细节,包括方法涉及的理论基础、技术特点、方法流程、程序伪代码和因果关系分析.

2.1　因果关系的本质和分类

因果关系的本质是因果关系研究的首要问题,所有因果发现方法都基于研究者对因果关系的理解. 1999 年 Pearl[1] 给出了经典的因果关系的描述:在某个相对稳定的系统中,重复控制并改变(intervention)其中的某个或某些变量 X 时,另一些变量 Y 总是随之发生变化,则 X 是 Y 的原因(前驱),Y 是 X 的结果(后继),其中 X 和 Y 均表示变量集合. Pearl 的描述符合人类对因果的认知,适用于不同领域,不针对特定数据,具有一般性,因此得到了广泛认可. 但 Pearl 的描述只能为判定因果关系提供思路,在实际应用中可操作性不强.

基于 Pearl 的描述,我们对因果关系有着更深层次的认识:

1) 因果关系是客观存在的自然现象和事物间相互作用的约束规则,无需第三方的操控(intervention)因果也会发生.

2) 因果关系不能改变果变量 Y 的本质属性,但可以改变 Y 在这些属性上的概率,即 Y 从原有状态 $P_{Y,t}$(时刻 t 时 Y 的概率分布)变化为 $P_{Y,t'}$(时刻 $t'>t$ 时 Y 的概率分布),因此只要变量 X 使得 Y 在若干属性(而不是全部属性)上的概率发生变化即可认为 X,Y 存在因果关系. 与 Pearl 的基本思想吻合但略有不同,我们认为果变量取值的改变只是因果关系的表象,究其根本是因果关系导致果变量的概率分布的改变,这是真实环境下基于大量观测数据发现因果关系的基础.

3) 因果关系可分为 2 种基本类型:激励因果和抑制因果. 因果关系使果变量 Y 的某些关注属性(focused features)发生的概率增加(我们称这样的因果关系为激励因果)或减少(称为抑制因果),同时

在另一些属性上概率减少(称为因果关系对属性的抑制作用)或增加(称为激励作用).

4) 原始概率分布(original probability distribution) $P_{X,O}$是一类最基本的原有状态,即变量 X 未受任何因果关系作用时的概率分布,我们称其为自由态下的概率分布,此时只有随机扰动和变量固有的概率分布决定变量的取值.

基于以上描述和认识,我们可以将因果关系作用过程描述为:任意变量 Z 在变量系统 V 中的原有状态 P_Z 是 Z 从原始状态 $P_{Z,O}$开始,经过一段时间的因果演化,在系统中所有因变量作用之前的特定时刻 t 和环境下在总体空间 Ω 上的概率分布,记为 $P_{Z,t}$. 该值可以通过时刻 t 的采样获得近似值 P'_t. 若我们获取的观测数据来自系统 V 在时刻 $t+\Delta t$ 足够大的样本空间 Ω',其中,$0\leqslant\Delta t\leqslant T$,$T$ 是 V 中所有因果关系生命周期结束的时刻,则至少可以保证 V 的初始变量 I 的概率分布 $P'_{I,(t+\Delta t)}$ 与真实的原有分布 $P_{I,t}$近似. 对于 V 中的非初始变量 Q,Q 的原有概率分布 $P_{Q,t}$会因为系统中因变量的作用而与采样值 $P'_{Q,(t+\Delta t)}$ 不符. 由于初始变量在 V 中没有前驱,V 中其他变量的取值和变化都无法影响初始变量,因此只有初始变量的概率分布在系统内的因果关系作用过程中保持不变,这个性质可以帮助我们找出初始变量. 根据 Pearl 的 do 演算规则[1],对任意初始变量 $I\in V$,I 的任意值 i 和任意非初始变量 $Q\in V$,$P(Q|\mathrm{do}(I=i))=P(Q|I=i)$成立,即 Q 依 I 的条件概率可以看成对 I 进行干预后 Q 的变化结果. 若已知初始变量,则根据 Pearl 的描述,上述性质至少可以帮助我们找出正确的因果关系方向,或初始变量在系统中的后继(详见定义 6 初始团树 IClique 结构). 以上描述与 Pearl 描述的基本思想吻合,在正确性前提下增强了可操作性(以找初始节点和无向图为起点),并且具有一般性.

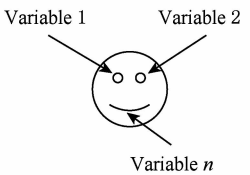
2.2 基本符号和概念

2.2.1 基本符号

一个因果关系系统(causality system, CS),如癌症预防与检测系统等,通常包含一个变量系统 V 和关于 V 的因果关系边集 $E\subseteq V\times V$. 一个因果关系发现问题 CP 通常会给定可观测变量系统 $V_o\subseteq V$、目标变量(target variable) $v_t\in V_o$ 和 V_o 的一组采样数据 D. 常见任务是发现目标变量的 3 层因果关系网络,即 v_t 的父、子、配偶变量以及这些变量之间的因果联系. 在因果关系网络结构(causal network)中,点代表变量,边表示因果关系,边的方向是从因变量指向果变量. 我们使用斜体小写字母,如 a,b 表示可观测变量. 用小写字母,如 a 表示变量 a 的某个取值(或状态);对于二值系统,通常用 a 表示取值为 1(或真),用 ¬a 表示取值为 0(或假),p_a 表示变量 a 取值为 1 的概率,$p_{a|b}$表示条件概率 $P(a=1|b=1)$. 若 $v\in V$,则用 $Pred(v)$表示 v 在 V 中的所有前驱(predecessor)节点集合,即 E 中存在从前驱节点到 v 的有向路径;$Parents(v)$表示 v 在 V 中的父节点集合,即 v 的直接前驱集合;用 $Descend(v)$表示 v 的后继(descendant)节点集合;$Children(v)$表示 v 在 V 中的子节点集合,即 v 的直接后继集合. 对于 2 节点 a,b 之间的无向边,我们用 a—b 表示;有向边则为 $a\rightarrow b$.

2.2.2 数据集类型

因果关系问题 $CP(V_o,D)$的观测数据集 D 对因果关系发现十分重要. 一般地,D 以矩阵的形式呈现,列代表变量,行代表观测值,一行观测值也称为一个采样(sample). 如图 1 所示,D 的收集方式有 2 种:1)案例采样(case sampling),即从不同的个体上对相同的变量进行采样获得;2)序列采样(sequence



(a) Feature choosing

Cases	Variable 1	Variable 2	…	Variable n
Case 1(person 1)	$d_C(1,1)$	$d_C(1,2)$	…	$d_C(1,n)$
Case 2(person 2)	$d_C(2,1)$	$d_C(2,2)$	…	$d_C(2,n)$
⋮	⋮	⋮		⋮
Case m(person m)	$d_C(m,1)$	$d_C(m,2)$	…	$d_C(m,n)$

Total: m persons; size of dataset C: $m\times n$.

(b) Case sampling: C type

Time	Variable 1	Variable 2	…	Variable n
t_1	$d_S(1,1)$	$d_S(1,2)$	…	$d_S(1,n)$
t_2	$d_S(2,1)$	$d_S(2,2)$	…	$d_S(2,n)$
⋮	⋮	⋮		⋮
t_m	$d_S(m,1)$	$d_S(m,2)$	…	$d_S(m,n)$

Total: m time moments; size of dataset S: $m\times n$.

(c) Sequence sampling: S type

Fig. 1 The illustration of C type and S type of datasets from two processes of case sampling and sequence sampling.

图 1 C 型和 S 型数据集示意图

sampling),即对指定个体上的变量在时间轴上进行持续采样.2 种方式的采样获取的 C 型和 S 型数据有很大的差别.由于 S 型数据能够发现诸如某变量的历史信息影响自身或其他变量的当前状态的情况,网络结构中常常会出现自圈(self-loop)、双向边(bi-directed edge)甚至更大的圈(cycle).因此,预先限定因果关系问题的网络结构为有向无圈图(DAG)是不可行的.根据 2.1 节的描述,我们认为网络结构是有向图(directed graph, DG)更合理.本文的实验数据集包含 C 型、S 型以及两者混合型 M 型(详见 3.1.1 节关于 SIGNET 数据集的介绍),那些预先设定网络结构为 DAG 的方法在 S 型和 M 型数据集上无法获得较好的实验结果.

2.2.3　基本概念

给定一个因果关系系统(V,E)和其因果关系图 G,我们将定义涉及 ICIC_Target 方法的若干重要概念.

定义 1. 孤立变量(isolated variable, IV). $iv\in V$,iv 是孤立变量当且仅当 $Parents(iv)=\varnothing$, $Children(iv)=\varnothing$. S_G^{iv} 代表图 G 的孤立变量集合.如图 2(a)中,变量 v_7 是孤立变量.

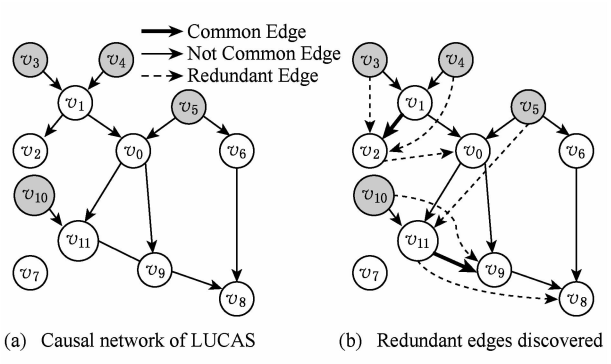


Fig. 2　Causal network of LUCAS and redundant edges ICIC discovered.

图 2　LUCAS 的真实因果网络结构和 ICIC 发现的冗余边

定义 2. 初始变量(exogenous variable, EV). $ev\in V$,ev 是初始变量当且仅当 $Parents(ev)=\varnothing$, $Children(ev)\neq\varnothing$. S_G^{ev} 代表图 G 的初始变量集合.如图 2(a)中,变量 v_3,v_4,v_5,v_{10} 是初始变量.

定义 3. 内生变量(endogenous variable, EnV). $env\in V$,env 是内生变量当且仅当 $env\in V\backslash(S_G^{iv}\cup S_G^{ev})$,即 G 中的非孤立或初始变量. S_G^{env} 代表图 G 的内生变量集合.图 3(a)中变量 v_0,v_6 是内生变量.

定义 4. 初始模式(exogenous pattern, EVP). $evp_i(i=0,1,\cdots,m)$ 是初始模式当且仅当 $evp_0=(\neg v_1,\neg v_2,\cdots,\neg v_m)$, $evp_i=(\neg v_1,\cdots\neg v_{i-1},v_i,\neg v_{i+1},\cdots,\neg v_m)1\leqslant i\leqslant m$,其中,$S_G^{ev}=\{v_1,v_2,\cdots,v_m\}$,$m$ 是初始变量个数. S_G^{evp} 表示图 G 的所有初始模式的集合.图 3(b)中 $evp_0=(\neg v_3,\neg v_4,\neg v_5,\neg v_{10})$, $evp_2=(\neg v_3,v_4,\neg v_5,\neg v_{10})$.

定义 5. 模式(pattern). 定义任意 2 个变量 v_i, $v_j\in V$ 的模式为 $Pattern_{i,j}=(\neg v_{i1},\cdots,\neg v_{il},v_{j1},\cdots,\neg v_{jk})$,其中 $Parents(v_i)\backslash\{v_j\}=\{v_{i1},v_{i2},\cdots,v_{il}\}$, $Parents(v_j)\backslash\{v_i\}=\{v_{j1},v_{j2},\cdots,v_{jk}\}$.图 3(b)中,$Pattern_{0,1}=(\neg v_3,\neg v_4,\neg v_5)$.定义任意变量 $v_i\in V$ 的模式为 $Pattern_i=(\neg v_{i1},\neg v_{i2},\cdots,\neg v_{il'})$, 其中 $Parents(v_i)=\{v_{i1},v_{i2},\cdots,v_{il'}\}$.

定义 6. 初始团树(initial clique of an exogenous variable, IClique). $IClique_{ev}\{V_i,H_i\}$是初始变量 ev 的初始团树当且仅当 $ev\in S_G^{ev}$,$ev\in V_i\subseteq V$, 对于 $\forall v_j\in V_i\backslash\{ev\}$,$v_j\in Descend(ev)$,并且满足 $hv_j=P(v_j\mid evp_{ev})-P(v_j\mid evp_0)>\beta$($\beta$ 是设定的大于 0 的阈值),我们用值 $hv_j\in H_i$ 表示初始变量 ev 对变量 v_j 的影响力,H_i 是 $\forall v_j\in V_i\backslash\{ev\}$的 hv_j 的列表. S_G^{IC} 表示系统所有初始团树的集合.图 3(a)(b)是数据集 LUCAS 因果网络的初始团树.若将存在公共边的初始团树相连,则形成了初始团树图,如图 3(c)所示.

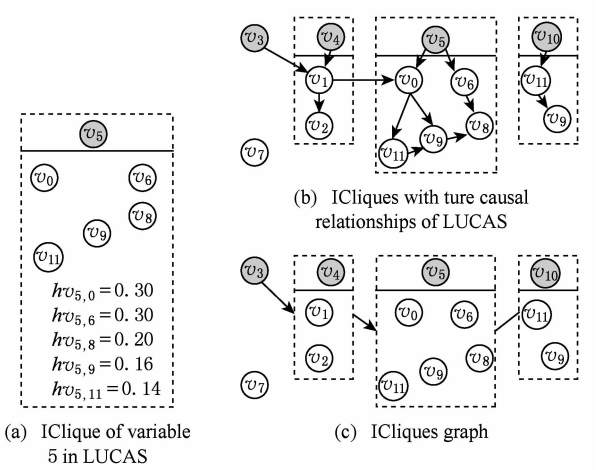


Fig. 3　ICliques and ICliques graph of LUCAS.

图 3　LUCAS 的初始团树和初始团树图

2.3　ICIC 方法框架

为提高因果发现算法的鲁棒性,降低复杂性,ICIC 方法框架在数据预处理之后包含 5 个核心步骤:1)确定初始变量;2)构建初始团树;3)发现无向图;4)添加方向;5)删除冗余边.

鉴于 IC 等经典方法生成 v 结构(v-structure)和添加方向规则造成的诸多问题,如算法复杂性高、鲁棒性低等,ICIC 步骤 1)、步骤 2)通过发现初始

变量和初始团树获取因果网络的大体分支和层次结构. 实验和理论分析证明,这样可以在 v 结构形成之前获得关于方向的粗略信息,使得发现无向图和添加的方向更合理,算法运行速度也会明显提升.

ICIC 步骤 3)利用改进的 d-割判定条件获得带有冗余边的无向图. 以常见的二值离散变量系统为例,ICIC 利用 d-割判定公式获取 d-割集,将判定条件由 IC 方法在各种取值组合下 d-割判定公式(式(1))都成立更改为式(2)~(5)各式之一成立即可判定 a,b 之间无边. ICIC 方法对二值离散变量系统区分生成激励和抑制因果无向图. 我们通常认定对果变量 1 值的提升是激励因果关系,对 1 值的降低是抑制因果关系,这符合人们的习惯. 对于多值变量系统,可以通过用户指定关注值获得激励和抑制因果结构.

$$P(a|b,c)=P(a|c),\ a,\ b,\ c=0\ \text{or}\ 1, \tag{1}$$

$$P(a|b,c)=P(a|c),\ c=0\ \text{or}\ 1, \tag{2}$$

$$P(a|\neg b,c)=P(a|c),\ c=0\ \text{or}\ 1, \tag{3}$$

$$P(\neg a|b,c)=P(\neg a|c),\ c=0\ \text{or}\ 1, \tag{4}$$

$$P(\neg a|\neg b,c)=P(\neg a|c),\ c=0\ \text{or}\ 1. \tag{5}$$

ICIC 步骤 4)根据前 3 步的结果添加方向后,在步骤 5)回溯所有可能的冗余边,确定为冗余后删除该边. 我们遍历所有三角形结构中每一条非公共边作为可能的冗余边,见图 2(b)所示,ICIC 将保留那些至少 2 个三角形结构公有边(加粗实边),虚线边是判定为冗余的边,其作用可以由其他边替代. ICIC 方法不预先限定因果关系网络结构,因此对存在双向边或圈等结构的真实网络的发现结果更有效.

2.4 初始变量与初始团树

2.4.1 初始变量的发现

初始变量的发现是 ICIC 方法框架的步骤 1)和重要步骤. 在 V_o 中寻找初始变量 EV 是二值分类问题,特征也看似很明显. 但在大多数应用中,初始变量的发现并非易事. 首先,所有变量的原有分布 $P_{Z,t}$ 未知;其次,在未知初始变量的前提下,任何变量之间的条件概率 $p_{a|b}$ 都不能轻易地认为等同于 $p_{a|do(b)}$. 因此本文初始变量的发现方法 EVD 主要基于 2 点:1)根据初始变量的基本性质,通过所有已知变量参与投票的方式找出最可能的初始变量;2)区分激励和抑制可以帮助化简初始变量的发现难度. EVD 方法的具体实现见算法 1,其中利用已发现的初始变量寻找其他初始变量的步骤需要计算初始团树(详见 EVD 的 Step2. 2),计算方法详见函数 *computeIClique*(),其具体含义和解释则见 2.4.2 节.

算法 1. EVD.

输入:变量集 V、数据集 D;

输出:S_G^{ev},S_G^{iv},S_G^{IC}.

Step1. 初始化 S_G^{ev},S_G^{iv},S_G^{IC},$R\leftarrow\varnothing$,$currentV\leftarrow V$;

$P1\leftarrow\{p_{v_0},p_{v_1},\cdots,p_{v_{n-1}}\}$,

$Q11_l\leftarrow\sum_k p_{v_k|v_l}$,

$Q10_l\leftarrow\sum_k p_{v_k|\neg v_l}$, $k,l=0,1,\cdots,n_1$;

Step2. while ($|currentV|$的值减少)

Step2.1. if ($v_i\in currentV$, v_i 是常量) $R\leftarrow R\cup\{v_i\}$;

else if ($\max(P1)\gg\min(P1)$)

for $\forall v_{ii}(currentV\backslash\{v_i|\ i=\underset{k}{\text{argmin}}(P1)\}$

if ($p_{v_{ii}}-\min(P1)<\beta$||

$ii==\underset{k}{\text{argmin}}\times$

$\left(\sum_l\frac{|p_{v_k|v_l}-p_{v_k|\neg v_l}|}{|Q11_l-Q10_l|}\right)$

if ($\exists v_a(currentV$ s.t.

($|p_{v_{ii}|v_a}-p_{v_{ii}|\neg v_a}|-|p_{v_a|v_{ii}}-p_{v_a|\neg v_{ii}}|)<\beta$)

$R\leftarrow R\cup\{v_{ii},v_a\}$;

endif

else $R\leftarrow R\cup\{v_{ii}\}$;

endif

endfor

endif

else

$P0\leftarrow computeOneOne(currentV,P1)$, $R\leftarrow R\cup\{v_i|i=\underset{k}{\text{argmin}}(P0)\}$;

Step2.2. for $\forall v_i\in R$

$IClique_i\{V_i,H_i\}\leftarrow computeIClique(v_i,S_G^{ev},V)$;

endfor

if ($|V_i|=1$) $S_G^{iv}\leftarrow S_G^{iv}\cup\{v_i\}$;

else

$S_G^{ev}\leftarrow S_G^{ev}\cup\{v_i\}$, $S_G^{IC}\leftarrow S_G^{IC}\cup\{IClique_i\{V_i,H_i\}\}$;

endif

Step2.3. $currentV\leftarrow currentV\backslash V_i$;

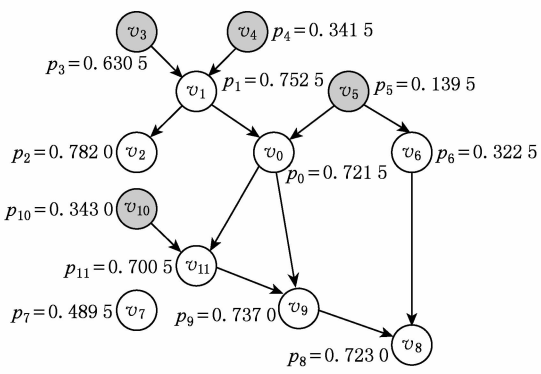
Step3. if ($|currentV|>0$)

for $\forall v_k\in currentV, v_i\in IClique_i\in S_G^{IC}$

if ($(p_{v_k|v_i}-p_{v_k|\neg v_i})\gg0$)

$V_i \leftarrow V_i \cup \{v_k\}$,
$currentV \leftarrow currentV \setminus \{v_k\}$;
endif
endfor
endif
Step4. $S_G^{iv} \leftarrow S_G^{iv} \cup currentV$;
Step5. return $S_G^{ev}, S_G^{iv}, S_G^{IC}$;
Function $computeOneOne(currentV, P1, D)$
Step1. 初始化 $P0 \leftarrow \varnothing$;
Step2. for $\forall v_i \in currentV$
$num \leftarrow$samples $\{\neg v_0, \cdots, \neg v_{i-1}, v_i, \neg v_{i+1}, \cdots, \neg v_{n-1}\}$的数量;
$p_{\neg v_i} \leftarrow num/(|D| \times p_{v_i})$, where $p_{v_i} \in P1$; $P0 \leftarrow P0 \cup \{p_{\neg v_i}\}$;
endfor
Step3. return $P0$;
Function $computeIClique(v_i, S_G^{ev}, V)$
Step1. 初始化 $V_i \leftarrow \{v_i\}, H_i \leftarrow \varnothing$;
Step2. for $\forall v_j \in V \setminus S_G^{ev}$
$hv_{ij} \leftarrow p_{v_j|ev p_i} - p_{v_j|ev p_0}$;
if $(hv_{ij} > \beta)$ $V_i \leftarrow V_i \cup \{v_j\}$, $H_i \leftarrow H_i \cup \{hv_{ij}\}$;
endif
endfor
Step3. return $IClique_i\{V_i, H_i\}$;

以发现激励因果为例,对任意变量 a:1) $a=1$(关注值)的概率 p_a 越小,则 a 受激励的机会越小;2) $v_i \in V \setminus \{a\}$, $|p_{a|v_i} - p_{a|\neg v_i}| / |Q11_i Q10_i|$ 越小,则 a 受 V 中其他变量激励的可能越小;3) $\forall b \in V \setminus \{a\}$, $|p_{a|b} - p_{a|\neg b}| \ll |p_{b|a} - p_{b|\neg a}|$,则 b 越有可能受 a 激励,而非 a 受 b 激励.以上判断显示 a 有可能是激励因果网络的初始或孤立变量,如图 4 中的变量 v_5.还有一类变量 b,尽管 p_b 值足够小,但 b 存在已确定的因变量,因此不是初始变量,如图 4 中的变量 v_6.为了剔除这样的变量和孤立变量,本文 EVD 方法在确定某变量 v 是初始变量前计算该变量的 $IClique_v$.若 $IClique_v$ 只包含根 v,则表明 v 是孤立变量;若 $IClique_v$ 包含其他变量 v_i,则 v_i 是 v 的后继,且算法将不再遍历 v_i,即当前活跃变量集 $currentV$ 不再包含 v 和 v_i.另外,在数据集中取值为常数的变量也符合初始变量的特性.若因果关系网络不含初始变量,则该网络一定有圈.这种情况下,由若干概率值或以彼此作为条件的概率值较接近的变量形成的圈起到了初始变量的作用,这些变量的初始团树将包含圈内的邻接变量.当所有变量的概率值都非常接近时,我们比较这些变量的"非激励影响",即变量取 1 而其他变量为 0 的概率,详见算法 1 的函数 $computeOneOne()$.拥有最小"非激励影响"的变量 v_j 暂时加入初始变量集,计算获得其 $IClique_j$,并依据 V_j 包含的非初始变量对 $currentV$ 中剩余变量的"影响力"(详见算法 1 的 Step3),判断哪些变量添加至 $IClique_j$.重复以上过程,直到 $currentV$ 为空.



(a) Positive stimulating graph of LUCAS and Probability of variables (Statistic values from LUCAS dataset)

$p_{v_6|v_0}=0.342\ 3$; $p_{v_6|\neg v_0}=0.271\ 1$;
$p_{v_6|v_1}=0.323\ 6$; $p_{v_6|\neg v_1}=0.319\ 2$;
$p_{v_6|v_2}=0.324\ 2$; $p_{v_6|\neg v_2}=0.316\ 5$;
$p_{v_6|v_3}=0.317\ 2$; $p_{v_6|\neg v_3}=0.331\ 5$;
$p_{v_6|v_4}=0.332\ 4$; $p_{v_6|\neg v_4}=0.317\ 4$;
$p_{v_6|v_5}=0.634\ 4 > p_{\text{uni}}$; $p_{v_6|\neg v_5}=0.271\ 9 < p_{\text{uni}}$;
$p_{v_6|v_7}=0.312\ 6$; $p_{v_6|\neg v_7}=0.332\ 0$;
$p_{v_6|v_8}=0.410\ 1$; $p_{v_6|\neg v_8}=0.093\ 9$;
$p_{v_6|v_9}=0.331\ 8$; $p_{v_6|\neg v_9}=0.296\ 6$;
$p_{v_6|v_{10}}=0.335\ 3$; $p_{v_6|\neg v_{10}}=0.315\ 8$;
$p_{v_6|v_{11}}=0.339\ 0$; $p_{v_6|\neg v_{11}}=0.283\ 8$;
Uniform Distribution of Binary Variables v_i:
$p_{\text{uni}i}=P(v_i=0)=P(v_i=1)=0.5, i=0,1,\cdots,11.$
$p_{v_5}<p_{v_6}<p_{v_4}<p_{v_{10}}<p_{v_7}$

(b) Conditional probability of v_6

Fig. 4　Exogenous variables discovery for LUCAS.

图 4　LUCAS 的初始变量发现示意图

2.4.2　初始团树的计算

初始团树 $IClique_{ev}$ 是以初始变量 ev 为根的分支,其非根节点都是 ev 的后继.初始团树的计算见 EVD 算法的函数 $computeIClique()$.在本文方法中,构建初始团树起到了 3 个作用:1)在初始变量发现过程中,帮助算法从 $currentV$ 中排除非初始变量;2)区分初始变量和孤立变量;3)避免 IC 等经典方法生成 v 结构和添加方向规则造成的诸多问题,

如算法复杂性高、鲁棒性低等,并利用初始团树获取因果网络的大体分支和层次结构.不仅如此,初始团树有很多性质(性质1～5),可以帮助我们更快地发现因果关系.

性质 1. 对于 $v_i \in S_G^{cv}$, $v_j \in V_i \setminus \{v_i\}$, V_i 是 $IClique_i$ 的节点集,若存在无向边 v_i—v_j,则 $v_i \rightarrow v_j$,如图5(a)所示.

性质 2. 对于 $v_k \in V_1$, $v'_k \in V_2$, V_1, V_2 是 $IClique_1$, $IClique_2$ 的节点集,若存在无向边 v_k—v'_k,则 $v_k \in V_1 \setminus \{v_1\}$ 和 $v'_k \in V_2 \setminus \{v_2\}$,如图5(b)所示.

性质 3. 若 v_k 和 v'_k 分别来自 $IClique_1$ 的节点集 V_1 和 $IClique_2$ 的节点集 V_2,且存在有向边 $v_k \rightarrow v'_k$, $\forall v_0 \in Pred(v_k) \bigcap V_1$, v_0 与 v'_k 不邻接,若 v_k 已知,则 v_0 对 v'_k 的影响相较 v_k 来说可忽略,如图5(c)所示.

性质 4. 若 $v_i \in V_i \setminus \{v_1\}$, v_1 是 $IClique_1$ 的初始变量,则 $v_1 \in Pred(v_i)$.

性质 5. 在不同初始团树中的同一条边有相容的方向,如图5(d)所示.

这些性质可以帮助快速确定边的方向和节点所在的分支,减少算法复杂度并提高准确性.性质1和性质2可由初始变量的定义推出.性质3是Markov链的性质.性质4描述了 $IClique_1$ 是初始变量 v_1 在网络结构中的影响力范围.性质5中相容方向是指具有相同单方向的边.由初始变量和初始团树的计算可知,不同于IC等方法在确定v结构后再形成分支,ICIC方法在发现初始变量及其分支的过程中,自然形成各个分支的交汇点.3条规则可以从以上性质推导出来,能帮助添加方向和优化程序,如图6所示.

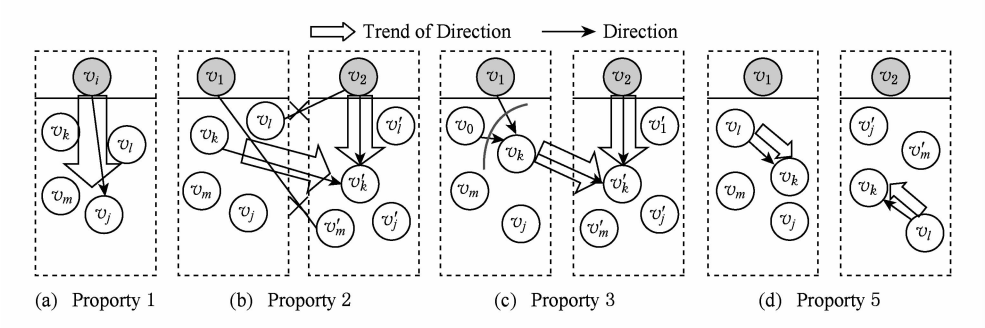


Fig. 5 Directions properties within IClique or between ICliques.

图5 初始团树和初始团树间方向的性质

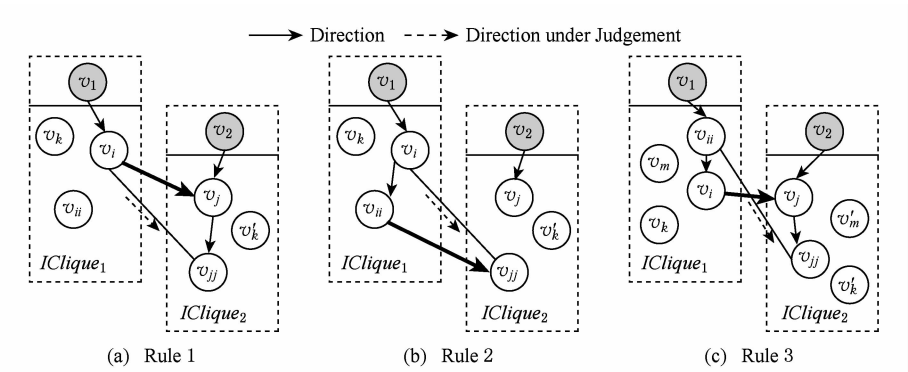


Fig. 6 Rules for directing edges quickly.

图6 添加方向的优化规则

规则 1. 若 $v_i \rightarrow v_j$, v_i—v_{jj}, $hv_j \gg hv_{jj}$,其中 v_j, $v_{jj} \in V_2 \setminus \{v_2\}$, V_2 是 $IClique_2$ 的节点集,则 $v_i \rightarrow v_{jj}$,如图6(a)所示.

规则 2. 若 $v_{ii} \rightarrow v_{jj}$, v_i—v_{jj}, $hv_i \gg hv_{ii}$,其中 v_i, $v_{ii} \in V_1 \setminus \{v_1\}$, V_1 是 $IClique_1$ 的节点集,则 $v_i \rightarrow v_{jj}$,如图6(b)所示.

规则 3. 若 $v_i \rightarrow v_j$, v_{ii}—v_{jj}, $hv_{ii} \geqslant hv_i$, $hv_j \geqslant hv_{jj}$,其中 v_i, $v_{ii} \in V_1 \setminus \{v_1\}$, v_j, $v_{jj} \in V_2 \setminus \{v_2\}$, V_1 和 V_2 是 $IClique_1$ 和 $IClique_2$ 的节点集,则 $v_{ii} \rightarrow v_{jj}$,如图6(c)所示.

2.5 目标节点的因果网络发现算法 ICIC_Target

本文针对目标变量(target variable)的网络结构发现提出了ICIC_Target算法,参见算法2.目标节点的因果关系网络发现问题 $CP(V_o, v_t, D)$ 中 V_o

一般十分庞大,极大地增加了计算复杂性,算法 2 的 Step2 对数据进行相关性分析,以适当地降维,尽早将无关变量限定在计算范围之外. 其中 ICIC_Structure 函数可以用于全局因果关系发现算法,使用的不等式有:

$$hv_{ia}-hv_{ib}>\beta, a,b\in V_i, \tag{6}$$

$$P(b|a,Pattern_{a,b})-P(a|b,Pattern_{a,b})>\varepsilon_0, \tag{7}$$

$$P(b|a,Pattern_b)-P(b|\neg a,Pattern_b)\leqslant\varepsilon_0, \tag{8}$$

其中,$v_i\in S_G^{ev}$,$a,b\in S_G^{env}$,$Pattern_{a,b}$和 $Pattern_b$ 见定义 5,ε_0 是足够小的正值.

算法 2. ICIC_Target.

输入:变量集 V、目标变量 v_t、数据集 D;

输出:G.

Step1. 初始化 $G\leftarrow\varnothing$,$V_t\leftarrow\{v_t\}$;

Step2. for $\forall v_i\in V\backslash\{v_t\}$

if $((p_{v_i|v_t}-p_{v_i|\neg v_t})>\beta)$ $V_t\leftarrow V_t\cup\{v_i\}$;

endif

endfor

Step3. for $\forall v_j\in V_t\backslash\{v_t\}$

for $\forall v_k\in V\backslash V_t$

if $((p_{v_k|(v_j,\neg v_t)}-p_{v_k|v_j})>\beta)$ $V_t\leftarrow V_t\cup\{v_k\}$;

endif

endfor

endfor

Step4. $G\leftarrow ICIC\_Structure(V_t,D)$;

Step5. return G;

Function $ICIC\_Structure(V,D)$

Step1. 初始化 $G\leftarrow$complete graph;

Step2. $S_G^{ev},S_G^{iv},S_G^{IC}\leftarrow EVD(V,D)$;

Step3. $G\leftarrow UndirectedGraph(V,D,S_G^{ev},S_G^{iv})$;

Step4. for $\forall IClique_i\in S_G^{IC}$

$V_i'\leftarrow V_i\backslash\{v_i\}$; /* V_i 是 $IClique_i$ 的变量集 */

for $\forall v_a\in V_i'$

if ($\exists edge$ v_i—v_a, v_i 是 $IClique_i$ 的初始变量)

direct $v_i\rightarrow v_a$根据 Property 1,

$V_i'\leftarrow V_i\backslash\{v_a\}$;

endif

endfor

endfor

for $\forall v_b\in V_i'$

for $\forall v_a$—v_b, $v_a(V_i\backslash\{v_i\}$ and $v_a\notin V_i'$

direct $v_a\rightarrow v_b$根据式(6),(7);

endfor

if (所有 v_a—v_b 添加方向为 $v_a\rightarrow v_b$)

$V_i'\leftarrow V_i\backslash\{v_b\}$;

endif

endfor

Step5. for v_a—$v_b\in$ IClique

direct $v_a\rightarrow v_b$根据式(6),(7);

endfor

Step6. for $\forall v_a$—v_b 不在一个 IClique 中

direct $v_a\rightarrow v_b$根据式(7),Rule 1,2;

endfor

Step7. $\{T.a,T.b\}\leftarrow FindTriangle(G)$;

/* 如图 7 所示 */

Step8. while ($T.a\neq\varnothing$)

for $\forall d=(v_a\rightarrow v_b)\in t\in T.a$

if ($\exists t_1\in T.b$, s.t. $d\in t_1$)

continue;

endif

if (式(8)成立)

从 G 中删除 d,更新 $T.a$,$T.b$,

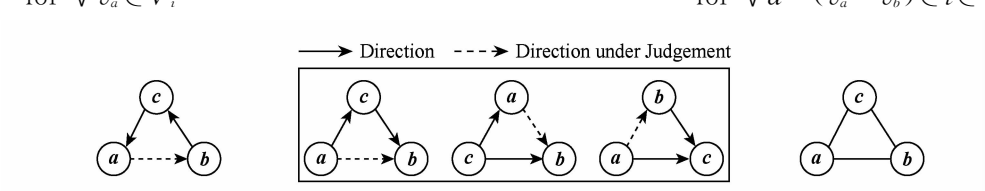
break;

endif

endfor

while ($T.b\neq\varnothing$)

for $\forall d=(v_a\rightarrow v_b)\in t\in T.b$



(a) Triangle structure $T.a$　(b) Triangle structure $T.b$　(c) Triangle skeleton

Fig. 7　Two shapes of triangle structure (a,b,c) $T.a$, $T.b$ with one redundant edge $a\rightarrow b$ noted as dash arrow.

图 7　带圈三角形的集合 $T.a$ 和非圈三角形集合 $T.b$ 示意图

```
        if (t1(T.b, s.t. d(t1)
          continue;
        endif
        if (式(8)成立)
          从 G 中删除 d, 更新 T.b,
          break;
        endif
      endfor
Step9. return G.
```

3 实验结果与分析

本文从理论和实验2方面对ICIC方法进行了分析和验证. 我们通过实验途径说明了ICIC_Target方法在不同实验数据上具有更好的性能和鲁棒性,并通过理论途径分析了ICIC_Target方法的稳定性和低复杂性.

3.1 实验与分析

3.1.1 实验设计

本文利用由WCCI 2008的"Causation and Prediction"竞赛[24]和NIPS 2008 Causality Workshop[6,12]正式发布的各领域的大小类型不同、或噪音、或操控的数据集(如表1所示,列出了数据集的性质),与多种有代表性的方法进行对比实验. 根据数据集和算法的特点,我们选择可比较的部分组成相应的实验组. 所有参与对比实验的算法已在本文第1节做了介绍. 对比实验涉及10个分实验,如表2所示,说明了实验设计的情况.

Table 1 Properties of Datasets Used in Experiments

表1 实验所用数据集的各项性质

Source	Dataset	Training/Test Sets Size	Feature/Target Types	Probe/Noise/Manipulate	Underlying Network
REsimulated Gene Expression Dataset[7]	REGED0	(1000×500)/(999×20000)	Numeric/Binary	N/N/N	Target
	REGED1	(Same Training Samples,	Numeric/Binary	N/Y/N	Target
	REGED2	Different Test Samples)	Numeric/Binary	N/Y/N	Target
Measurement ARTIfact[7]	MARTI0	(1025×500)/(1024×20000)	Numeric/Binary	Y/N/Y	Target
	MARTI1	(Same Training Samples,	Numeric/Binary	Y/Y/Y	Target
	MARTI2	Different Test Samples)	Numeric/Binary	Y/Y/Y	Target
Weblog[25]	Weblog	(20×512)/	Temporal/Numeric	N/N/N	Global
Abscisic Acid Signaling Network[26]	SIGNET	(43×21×300)/	Temporal/Binary	N/N/N	Global

Table 2 Experiments Designed for Causality Discovery Research

表2 因果关系发现的实验设计

Aim	Sample Type	Experiment	Dataset	Methods	Evaluation
Target	Case Samples	$E_{T.C.REGED.Train}$	Training Set of REGED	HITON (HITON_PC, HITON_MB) PCMB IAMB ICIC_Target	Score of Cost Matrix
		$E_{T.C.REGED0.Test}$	Test Set of REGED0+$list\_v_t$		
		$E_{T.C.REGED1.Test}$	Test Set of REGED1+$list\_v_t$		
		$E_{T.C.REGED2.Test}$	Test Set of REGED2+$list\_v_t$		
		$E_{T.C.MARTI.Train}$	Training Set of MARTI		
		$E_{T.C.MARTI0.Test}$	Test Set of MARTI0+$list\_v_t$		
		$E_{T.C.MARTI1.Test}$	Test Set of MARTI1+$list\_v_t$		
		$E_{T.C.MARTI2.Test}$	Test Set of MARTI2+$list\_v_t$		
	Sequence Samples Mix Samples	$E_{T.S.Weblog}$	Weblog	FCI, GCTE, HITON, IC, IC*, ICIC_Target, LCD2, LiNGAM, PC, SI	Mean Score of Cost Matrix
		$E_{T.S.SIGNET}$	SIGNET		

REGED[7]数据集来自医学领域,目的是通过发现触发肺癌疾病的基因和肺癌造成改变的基因,预测肺癌的发病状况. REGED包含的基因表达量(gene expression data)数据由真实的DNA微阵列数据(DNA microarray data)训练出的模拟器生成. 该数据集的3个子任务REGED0,REGED1,REGED2均有相同的1个目标变量和999个其他变量,其中REGED1和REGED2的测试数据中某些变量受到操控,以模拟药物或RNA抑制(RNA silencing)治疗的情况. MARTI[7,10]是REGED加入噪音的版本,有1024个变量,其中25个是校正变量,用于帮助剔除噪音影响. MARTI的因果网络结构与REGED

相同.二者的设计初衷是检测目标变量(binary类型)预测算法的性能,本文的实验目的是检测因果关系结构发现算法的性能,因此在WCCI和NIPS公布了REGED和MARTI所在系统的网络结构后,我们将它们的一个训练集和3个测试集都用来评估因果结构发现算法在有操控或有噪音的数据集上的性能.由于测试集的目标变量结果至今未公开,本文利用SimplePredict算法(见算法3)和已公布的网络结构将目标变量的值(即表2中$list_v_t$)补充完整.

算法3. SimplePredict.

输入:目标变量v_t、局部网络N_L、测试数据集D_T;

输出:目标结果值列表$list_v_t$.

Step1. for $\forall sample_i \in D_T$

if ($\exists v \in Parents(v_t)$ s. t. $v=1$ in $sample_i$)

$list\_v_t.add(1)$;

endif

else if (2/3 children in N_L are 1 in $sample_i$)

$list\_v_t.add(1)$;

endif

else

$list\_v_t.add(-1)$;

endfor

Step2. return $list\_v_t$;

WebLog来自Grozea和Romania[25]设计的网页浏览实验的真实数据.在该实验中,有20个网页供人们浏览,每个网页有到另19个网页的链接.实验收集了512 d的浏览日志,将每一天每一个网页的浏览数统计出来作为WebLog数据集.实验作者发现某些网页可以激发人们去浏览某些其他网页,人们的这些行为被记录和统计下来,作为实验的真实因果网络结构.该结构存在双向边,这是因为一个网页$page_a$能激发人们浏览$page_b$,同时$page_b$也能激发人们浏览$page_a$.Weblog的设计初衷是发现因果关系网络,因此该数据集无测试数据,SIGNET数据集亦是同样的情况,表1用空格表示.

SIGNET是模拟数据集,它的真实网络是一个生物信号网络的布尔模型.Jenkins和Soni[26]对43个生物信号变量进行了案例和序列2种方式的采样,变量中有38个可变变量、5个常数变量.该数据集总共300个序列采样组,每组包含21个序列采样,其中第1个采样由43个变量的初始值组成,其他20个采样是后续时间点上变量的变化情况.SIGNET对单纯的案例数据因果发现方法(如IC)和序列数据因果发现方法(如SI)都是较大的挑战.

我们参考NIPS 2008提出的代价矩阵得分(score of cost matrix)[12]作为实验结果的评估标准之一,其中代价矩阵的元素是局部网络结构中不同位置上出现错误变量的代价权重,具体取值如图8所示.变量在网络中的位置离正确位置越远,错误的代价越高;出错越多,代价累计值越高.因此该值越小,结果网络结构越相近正确.由于HITON,PCMB,IAMB方法输出的是PCset或MBset而非网络结构,无法直接计算PC和MB的代价矩阵得分,因此我们将PCset或MBset中包含的正确的父节点、子节点和配偶节点对应放置在目标节点的父节点、子节点和配偶节点位置上,集合中其余错误的节点均放在子节点位置,这样获得的网络结构是最接近正确答案的结构,由此计算的代价矩阵得分是HITON等方法的最小代价值,实际的代价值不会小于这个值,因此表3和表4中用"≥"表示HITON等方法的实验结果.

Depth Obtained	Desired Relationship	1 Parents	1 Children	2 Spouses	× Other
1	Parents	0	1	1	2
1	Children	1	0	1	2
2	Spouses	1	1	0	2
×	Other	2	2	2	0

Fig. 8　The cost matrix of MB and PC.

图8　MB和PC代价矩阵

在本文的实验结果评测中,对于IC*和FCI结果中存在带有标记的边,以及IC和ICIC_Target方法结果中不确定方向边(表示为无向边)或双向边,这些在NIPS 2008竞赛中未做明确规定的情况.在生成网络结构的邻接矩阵时,我们按3个原则处理:1)2个端点方向均明确者(如单向边和双向边),按原有方向处理.2)一端方向不明确者(如FCI方法标记为"o"的边(a o→ b)则表明a端点方向尚不明确;IC*方法的有向边$a \to b$意味着a到b方向已确定或a,b存在同一个隐含原因,因此$a \to b$仍是a端方向不明确),该端按不存在该方向处理.3)2个端点方向均不明确者(如无向边),按2个端点均有方向处理.为适应实验结果存在双向、圈等网络结构的问题,关于代价矩阵得分的计算,我们采用4个原则处理:1)对NIPS2008提出的规定范围内的错误代价按原有代价矩阵中的值处理,如图8所示;2)对既

Table 3 Results of Score of Cost Matrix in REGED

表 3 REGED 实验的结果

Algorithm	$E_{T.C.REGED.Train}$	$E_{T.C.REGED0.Test}$	$E_{T.C.REGED1.Test}$	$E_{T.C.REGED2.Test}$
HITON_PC	≥18.0	≥12.0	≥12.0	≥12.0
ICIC_Target(PC)	**9.0**	**5.5**	**1.0**	**1.0**
PCMB	≥20.0			
HITON_MB	≥40.0	≥128.0	≥276.0	≥12.0
ICIC_Target(MB)	**21.5**	**21.0**	**9.0**	**3.0**
PCMB	≥32.0			
IAMB	≥36.0	≥34.0	≥20.0	≥4.0

Table 4 Results of Score of Cost Matrix in MARTI

表 4 MARTI 实验的结果

Algorithm	$E_{T.C.MARTI.Train}$	$E_{T.C.MARTI0.Test}$	$E_{T.C.MARTI1.Test}$	$E_{T.C.MARTI2.Test}$
HITON_PC	≥30.0	≥22.0	≥12.0	≥8.0
ICIC_Target(PC)	**17.0**	**15.0**	**9.0**	**4.0**
PCMB	≥28.0			
HITON_MB	≥86.0	≥78.0	≥71.0	≥8.0
ICIC_Target(MB)	**40.0**	**31.5**	**21.5**	**6.0**
PCMB	≥46.0			
IAMB	≥42.0	≥34.0	**≥20.0**	**≥4.0**

在正确位置出现又在 PC 或 MB 中错误位置出现的变量,给予相应减半的惩罚;3)对未在正确位置出现、但多次在 MB 的其他位置出现的变量,给予累加惩罚;4)Weblog 和 SIGNET 上的实验结果是以任意变量为目标计算的代价矩阵平均得分.

3.1.2 目标变量的网络发现实验结果及分析

本节将进一步说明在多数据集和多任务下的因果网络发现的实验细节和实验结果,得出分析结论.

本实验使用文献[27]提出的突发检测算法 BSE,将连续数据转化为离散数据后再进行各算法结果的比较和评估. ICIC_Target 方法多处涉及用 χ^2 假设检验(或参数 β)以判定显著差异、边的方向以及隐含因变量的存在. 假设检验的置信水平越高或 β 值越大则意味着 ICIC 方法发现的因果关系越强. 本文中设置 $\alpha=0.05$, $\beta=0.1$. 参数 ε_0 是足够小的正值,本文取值 $\varepsilon_0=0.0005$.

我们对 ICIC_Target 与 HITON, IC 等方法在目标节点网络结构发现上的性能进行了比较,实验结果如表 3~7 所示. 表中空白处对应任务上无实验结果的情况有 2 种可能的原因:1)方法未获得任何 PC 节点或 MB 节点(如表 3 和表 4 所示的 PCMB 方法在 REGED 和 MARTI 的测试集上得不出结果);2)方法不适用于该项任务(如表 5 和表 6 所示的 LCD2 方法,因为 Weblog 的网络结构中没有真正的初始变量,即 Cooper[4] 所指的 instrumental variable,因此 LCD2 方法在实验 $E_{T.S.Weblog}$ 中无法获得有效结果). 表 3~6 中粗体数字为同一评测标准下的最好成绩. 以实验 $E_{T.C.REGED.Train}$(如表 3 所示)为例,ICIC_Target 方法在父子网络结构(PC 结构)和父子配偶网络结构(MB 结构)发现 2 项任务中的代价矩阵得分(9.0 和 21.5)分别小于 HITON_PC 方法的得分(大于等于 18.0)和 HITON_MB 方法的得分(大于等于 40.0),因此本文方法在该实验中存在性能优势. 在实验 $E_{T.S.Weblog}$ 和 $E_{T.M.SIGNET}$ 中,我们以每个变量为目标变量计算得到 $|V|$ 个局部网络结构,并以代价矩阵得分的均值作为性能评估标准,实验结果如表 5 和表 6 所示. 表 7 展示了本文方法在实验 $E_{T.M.SIGNET}$ 中发现的激励因果和抑制因果的情况,并与 IC 方法进行了比较. 其中"×"表示算法发现的错误的因果关系,"√"表示正确的因果关系. 表格上半部是存在抑制因果的等式,表格下半部是只有激励因果的等式,各等式等号右边变量为因变量,等号左边为果变量.

Table 5　Results of Mean PC Score of Cost Matrix in Weblog and SIGNET Experiments

表 5　Weblog 和 SIGNET 的 PC 实验结果

Algorithms on PC Discovery	$E_{T.\,S.\,Weblog.\,PC}$	$E_{T.\,M.\,SIGNET.\,PC}$
FCI	18.00	12.16
GCTE	26.10	9.88
HITON_PC	≥17.80	≥8.47
IC	27.30	7.95
IC*	18.15	8.23
ICIC_Target(PC)	25.40	**4.23**
LCD2		12.65
LiNGAM	17.40	37.12
PC	**15.45**	127.00
SI	22.65	58.77

Table 6　Results of Mean MB Score of Cost Matrix in Weblog and SIGNET Experiments

表 6　Weblog 和 SIGNET 的 MB 实验结果

Algorithms on PC Discovery	$E_{T.\,S.\,Weblog.\,MB}$	$E_{T.\,M.\,SIGNET.\,MB}$
FCI	35.51	21.67
GCTE	31.05	17.51
HITON_MB	≥44.55	≥14.51
IC	32.16	15.29
IC*	41.33	15.96
ICIC_Target(MB)	**27.52**	**7.93**
LCD2		26.31
LiNGAM	44.10	100.30
PC	41.93	201.90
SI	39.00	133.80

Table 7　Comparison of Stimulating and Inhibiting Causalities Discovered by ICIC_Target and IC in $E_{T.\,M.\,SIGNET}$

表 7　ICIC_Target 与 IC 的 $E_{T.\,M.\,SIGNET}$ 实验结果(激励/抑制因果)对比

True Equations (Boolean)	ICIC_Target		IC	
	Stimulating	Inhibiting	Stimulating	Inhibiting
CAIM=ROS and not DEPOLAR		×CAIM=… and not CLOSURE √CAIM=… and not DEPOLAR		
ATRBOH=OST and ROP2 and not ABI			×CLOSURE	
HATPase=not PH		√HATPase=… and not PH ×HATPase=…and not CLOSURE		
Actin=not RAC	×Actin=…or CLOSURE	√Actin=… or not RAC		
ABI=not PA and not ROS				
KAP=not PH		√KAP=… and not PH ×KAP=… and not CLOSURE		
Ca=(CAIM or CIS) and not CaATPase				
AnionEM=not ABI or Ca			×ATRBOH	
DEPOLAR=KEV or AnionEM or not HATPase or not KOUT or Ca	×DEPOLAR=…or PH ×DEPOLAR=…or CLOSURE ×DEPOLAR=…or ROS	√DEPOLAR=… or not HATPase		
NO=NIA12 and NOS			×CaATPase ×KEV,√NOS	
CLOSURE=(KOUT or KAP) and AnionEM and Actin	√CLOSURE=…or KOUT √CLOSURE=…or Actin √CLOSURE=…and KAP			
cADPR=ADPRc			√ADPRc	
IP6=InsPK	√IP6=… or InsPK ×IP6=…or CLOSURE			
NIA12=RCN	√NIA12=…or RCN ×NIA12=…or CLOSURE			
ROP2=PA			×ATRBOH	
S1P=SPHK	√S1P=… or SPHK			

Note: Symbol "×" means wrong equation has been discovered by ICIC_Target or IC method, and "√" means right equation has been discovered.

本文 10 组对比实验的结果(如表 3～7 所示)展示了 ICIC_Target 方法良好的性能,该方法的准确性在大多数据集上领先.在实验过程中,该方法表现出了很好的鲁棒性和稳定性,能较好地处理不同类

型和特性的数据,程序运行速度快,尤其表现在处理 REGED 和 MARTI 测试数据(涉及上千个变量和上万条采样记录)这样的大规模因果系统上,ICIC_Target 方法平均耗时明显低于 HITON 方法,达到了 1/50(相同的运算环境下,ICIC_Target 方法耗时 2～5 min,由 Causal Explorer 实现的 HITON 方法平均耗时达到 2 h).另外,本文方法能更好地发现和区分激励因果和抑制因果,如表 7 所示,可以进一步为后续分析或应用提供更多有用的信息.

3.2 算法的稳定性、复杂性分析

3.2.1 稳定性分析

本文将分析算法 2 在输入有误差或错误时能否保持合理的可靠性.文献[2]提出了算法稳定性概念:若输入小错误产生了输出的大错误,则算法不稳定.

算法 2 首先通过判定 V 中与目标变量"相关"的变量子集对所有变量进行过滤.若该步骤将某个 PC 或 MB 变量排除在相关子集外,则该变量及其父子变量很难会对其他 PC 或 MB 变量的判定产生影响,因此此类误差不会破坏算法的稳定性.然而,该步骤通常会将判定条件设定的非常宽松,这就在相关子集中引入了大量非 PC 或 MB 变量.对这类误差,后续步骤的目标就是逐步细化相关子集、缩小误差影响的过程,因此 ICIC_Target 算法的整体稳定性主要依赖 ICIC_Structure 方法.

ICIC_Target 算法的核心 ICIC_Structure 函数的 Step3(相当于 IC 算法[1]第 1 步或 PC 算法[2]的 step A 的功能)若存在小错误,则很可能是某些 d-割关系被错误地包含或排除在下一步的输入中.本文 2.3 节已说明,该步可能产生冗余边,即确定 2 个节点间的 d-割关系比确定二者 d-connection 更苛刻,因此 ICIC_Structure 在 Step8 回溯并删除冗余边.可见 ICIC_Target 算法能控制产生冗余边的问题,在最终输出时尽量减少错误.另一个可能的错误是 ICIC_Structure 的 Step2 输出时错误地包含或遗漏某些初始变量.被遗漏的初始变量会在后续程序中被逐渐找回,而包含伪初始变量的错误会在后续程序中被放大,因为错误的初始变量将影响一部分边的方向判定.为限制该类错误,EVD 算法设计的尽可能严格地判定初始变量.其他步骤产生小错误或受小错误影响的情况有限,因此 ICIC_Target 方法具有较好的稳定性.

3.2.2 复杂性分析

纵观 ICIC_Target,PC,IC 等算法可知,条件独立判定易产生过多复杂性.该步骤中 ICIC_Target 的复杂度受因果关系发现问题 $CP(V_0,D)$ 规模的限制.设 n,n_{ev} 是变量和初始变量个数,m 是数据集包含的采样数,k 是变量的最大度数(degree).ICIC_Target 算法中所有 n_0-阶条件独立判定($n_0\leqslant 2$,不同于 PC 算法的 $n_0\leqslant n_1$)的数量上限为 $C_n^2\sum_{i=0}^{2}C_n^i$ 即 $n(n_1)^3/2$.因为 ICIC_Target 方法利用初始变量和删除冗余边的步骤正好可以弥补此处 n_0 有限的问题,因此即使最坏情况下,也会使得该算法需要的计算量远远小于 PC,IC,FCI 等.

实际上,这些判定的平均数量还会远远小于该上限值,因为 ICIC 只需判定 2 个节点是否被 d-割,无需知道谁 d-割了它们(不同于 PC 需要知道 d-割集 Sepset(v_i,v_j),就要遍历完所有的变量和变量组合),因此最坏情况出现的概率远远小于 PC 等算法.另外,我们还可以通过优先遍历高度数节点优化算法.

ICIC_Target 方法增加了初始变量的判定和 IClique 的计算,我们认为这些是在保证准确性、增强鲁棒性的前提下可以接受的代价.因为 EVD 极少得到比真实 n_{ev} 多的初始变量,因此最坏情况下以上 2 个步骤运算的次数是:n_{ev} 和 $(n-n_{ev})n_{ev}$,且每个初始变量的判定都需要遍历数据集.因此 EVD 算法的复杂度为 $O((m+n)n_{ev})$.在添加方向和删除冗余边的 2 个步骤中,考虑以下 2 种极端情况:1)若 ICIC 在条件独立判定步骤之后得到的无向图仍是完全无向图,则删边的复杂度可以忽略;2)若添加方向步骤后得到的有向图有最多的三角形结构、最少的初始变量、最少的公共边(同时处于至少 2 个三角形结构中的边),则删边步骤就遇到了最坏情况,此步的复杂度为 $O(n)$,与此同时添加方向步骤的复杂度为 $O(n)$.这 2 个步骤的最坏情况不会同时出现,因此复杂度上限为 $O(n^2)$,更精确些为 $O(nk)$.此外,ICIC_Target 判定相关子集的步骤还可以大大降低后续算法的复杂度.综上所述,算法的复杂度为 $O(n^4+mn_{ev})$.

4 结束语

本文针对真实环境下局部因果关系发现问题,通过分析因果关系的作用机制,利用初始变量和初始团树,提出了性能更高、鲁棒性更强的局部因果关系网络发现方法 ICIC_Target.实验和理论分析表明,该方法具有更好的稳定性、复杂度、鲁棒性.该方

法克服了传统方法由于受限于边的不确定性、双向边、圈、v结构等不足,降低了复杂性,且无需基于过多的前提和先验知识,因此更加适用于真实环境下的因果关系发现.

下一步将改进或优化初始变量的发现算法以提高算法的准确率、降低复杂度,进一步开展因果关系发现后的因果关系分析,包括因果的演化模型、隐含变量发现与分析,以及因变量合作竞争机制、遗传机制等方面的研究.

参　考　文　献

[1] Pearl J. Causality: Models, Reasoning, and Inference [M]. Cambridge, UK: Cambridge University Press, 1999

[2] Spirtes P, Glymour C, Scheines R. Causation, Prediction, and Search [M]. 2nd ed. Cambridge, MA: MIT Press, 2000

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

[4] Mani S, Cooper G F. A study in causal discovery from population-based infant birth and death records [C] //Proc of the American Medical Informatics Association (AMIA) Annual Fall Symp. Philadelphia, PA: AMIA, 1999: 315-319

[5] Granger C W J. Investigating causal relations by econometric models and cross-spectral methods [J]. Econometrica, 1969, 37(3): 424-438

[6] Guyon I. NIPS 2008 Workshop on Causality: Workshop Information Publishing [EB/OL]. (2008-12-12) [2013-07-11]. http://www.clopinet.com/isabelle/Projects/NIPS2008

[7] The Causality Workbench Team of NIPS 2008. Challenges in machine learning—Causality challenge #1: Causation and prediction [EB/OL]. (2008-12-12) [2013-07-11]. http://www.causality.inf.ethz.ch/challenge.php?page=datasets

[8] The Causality Workbench Team of NIPS 2008. LUCAS and LUCAP are lung cancer toy datasets: Dataset Description [EB/OL]. (2008-06-06) [2013-07-10]. http://www.causality.inf.ethz.ch/data/LUCAS.html

[9] Glymour C, Cooper G F. Computation, Causation, and Discovery [M]. Menlo Park, CA: AAAI Press, 1999

[10] Aliferis C F, Tsamardinos I, Statnikov A. HITON, a novel Markov blanket algorithm for optimal variable selection [C] //Proc of the American Medical Informatics Association (AMIA) Annual Symp. Washington, DC: AMIA, 2003: 21-25

[11] 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

[12] Guyon I, Aliferis C F, Cooper G F, et al. Design and analysis of the causation and prediction challenge [J]. Journal of Machine Learning Research: Workshop and Conf Proc, 2008, 3: 1-33

[13] Aliferis C F, Statnikov A, Tsamardinos I, et al. Local causal and Markov blanket induction for causal discovery and feature selection for classification part i: Algorithms and empirical evaluation [J]. Journal of Machine Learning Research, 2010, 11: 171-234

[14] Peña J M, Nilsson R, Björkegren J, et al. Towards scalable and data efficient learning of Markov boundaries [J]. International Journal of Approximate Reasoning, 2007, 45(2): 211-232

[15] Aliferis C F, Tsamardinos I, Statnikov A R. Causal explorer: A probabilistic network learning toolkit for biomedical discovery [C] //Proc of Int Conf Mathematics and Engineering Techniques in Medicine and Biological Sciences (METMBS). Las Vegas: CSREA Press, 2003: 371-376

[16] Aliferis C F. Causal Explorer Download: Register Page [CP/OL]. (2003-06-26) [2013-08-01]. http://www.dsl-lab.org/causal_explorer

[17] Peña J M. LiU-IDA-Department of Computer and Information Science: Home Page of Peña J M [EB/OL]. (2015-05-28) [2014-11-01]. http://www.ida.liu.se/~jospe

[18] Shibuya T, Harada T, Kuniyoshi Y. Causality quantification and its applications: Structuring and modeling of multivariate time series [C] //Proc of the 15th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2009: 787-796

[19] Lozano A C, Abe N, Liu Yan. Grouped graphical Granger modeling methods for temporal causal modeling [C] //Proc of the 15th ACM SIGKDD Int Conf Knowledge Discovery and Data Mining. New York: ACM, 2009: 577-586

[20] Ancona N, Marinazzo D, Stramaglia S. Radial basis function approach to nonlinear Granger causality of time series [J]. Physical Review E: Statistical Nonlinear and Soft Matter Physics, 2004, 70(5): 148-168

[21] 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

[22] Arnhold J, Grassberger P, Lehnertz K, et al. A robust method for detecting interdependences: Application to intracranially recorded EEG [J]. Physica D, 1999, 134(4): 419-430

[23] Lungarella M, Ishiguro K, Kuniyoshi Y. Methods for quantifying the causal structure of bivariate time series [J]. International Journal of Bifurcation and Chaos (IJBC), 2007, 17(3): 903-921

[24] The Causality Workbench Team of NIPS 2008. Home page of challenges in machine learning—Causality challenge #1: Causation and prediction [EB/OL]. (2008-12-12) [2013-07-11]. http://www.causality.inf.ethz.ch/challenge.php?page

[25] The Causality Workbench Team of NIPS 2008. The causal discovery in Web logs: Problem and datasdet [EB/OL]. (2008-06-06) [2013-07-10]. http://www.phobos.ro/data/index.html

[26] The Causality Workbench Team of NIPS 2008. SIGNET abscisic acid signaling network: Desciption of problem and dataset [EB/OL]. (2008-06-06) [2013-09-23]. http://www.causality.inf.ethz.ch/repository.php?id=5

[27] Sun Jing, Yin Jianping, Wang Ting, et al. A novel model of bursts in event sequences [C] //Proc of the Int Conf on Consumer Electronics, Communications and Networks (CECNet). Piscataway, NJ: IEEE, 2012: 816-821

Li Yan, born in 1979. PhD. Her main research interests include text mining and natural language processing.

Wang Ting, born in 1967. Professor and PhD supervisor. His main research interests include artificial intelligence and computer software.

Liu Wanwei, born in 1980. PhD and associate professor. His main research interests include formal method and verification.

Zhang Xiaoyan, born in 1981. PhD and experimentalist. Her main research interests include text mining and natural language processing.