控制与决策
Control and Decision

一种基于柯氏复杂度的因果网络定向方法

韩梦瑶， 鲁云军， 金乙乔， 刘乾， 陈克斌

引用本文:
韩梦瑶, 鲁云军, 金乙乔, 等. 一种基于柯氏复杂度的因果网络定向方法[J]. 控制与决策, 2021, 36(9): 2241−2248.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2020.0005

您可能感兴趣的其他文章

Articles you may be interested in

一种基于节点嵌入表示学习的社区搜索算法

Community search algorithm based on node embedding representation learning

控制与决策. 2021, 36(8): 1970−1976 https://doi.org/10.13195/j.kzyjc.2019.1439

基于改进蚁群算法的水面无人艇路径规划

Path planning for unmanned surface vehicle based on improved ant colony algorithm

控制与决策. 2021, 36(4): 847−856 https://doi.org/10.13195/j.kzyjc.2019.0839

基于FWADE−ELM的短时交通流预测方法

Short−term traffic flow forecasting based on hybrid FWADE−ELM

控制与决策. 2021, 36(4): 925−932 https://doi.org/10.13195/j.kzyjc.2019.1103

基于FWADE−ELM的短时交通流预测方法

Short−term traffic flow forecasting based on hybrid FWADE−ELM

控制与决策. 2021, 36(4): 925−932 https://doi.org/10.13195/j.kzyjc.2019.1103

一种新的基于标签传播的复杂网络重叠社区识别算法

A novel algorithm for overlapping community detection based on label propagation in complex networks

控制与决策. 2020, 35(11): 2733−2742 https://doi.org/10.13195/j.kzyjc.2019.0176

一种基于柯氏复杂度的因果网络定向方法

韩梦瑶[1,2], 鲁云军[1†], 金乙乔[1], 刘 乾[1], 陈克斌[1]

(1. 国防科技大学 信息通信学院，武汉 430019；2. 陆军勤务学院 国防经济系，重庆 400030)

摘 要: 因果网络定向问题实质是一个“多对多”因果关系发现过程,传统的V-结构定向方法只能确定一组马尔可夫等价类而非最终的因果关系. 为解决该问题,从柯氏复杂度的因果推断原理视角出发,利用贝叶斯链式法则推导出局部网络因果定向规则,并在此基础上提出高维全局网络因果定向方法. 同时,将前者运用于改进基于局部条件独立信息搜索学习马尔可夫毯典型算法,后者运用于改进基于约束的因果网络结构学习典型算法. 实验结果表明,改进后算法在保证较高准确率的同时可有效提升执行效率.

关键词: 因果网络；因果定向；柯氏复杂度；最小描述长度；随机复杂度；马尔可夫毯

中图分类号: TP181 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2020.0005 **开放科学(资源服务)标识码(OSID):**

引用格式: 韩梦瑶,鲁云军,金乙乔,等. 一种基于柯氏复杂度的因果网络定向方法[J]. 控制与决策, 2021, 36(9): 2241-2248.

A causal network orientation method based on Kolmogorov complexity

HAN Meng-yao[1,2], *LU Yun-jun*[1†], *JIN Yi-qiao*[1], *LIU Qian*[1], *CHEN Ke-bin*[1]

(1. College of Information and Communication，National University of Defense Technology，Wuhan 430019，China；
2. Department of Defense Economics，Army Logistical University of PLA，Chongqing 400030，China)

Abstract: The nature of causal network orientation problems is a “many-to-many” causal discovery process. The traditional V-structure method can only determine a set of Markov equivalent classes rather than the final causal relationship. In order to solve this problem, based on the Kolmogorov complexity, a causal orientation rule of local networks is deduced using the Bayesian chain rule, thus a high-dimensional global network causal orientation rule is proposed on this basis. At the same time, the former is used to improve the Markov blanket typical algorithm based on the local condition independent information searching; the latter is used to improve the constraint based causal network structure learning typical algorithm. The experimental results show that the improved algorithm can effectively improve the execution efficiency while ensuring high accuracy.

Keywords: causal network；causal orientation；Kolmogorov complexity；the minimum description length；stochastic complexity；Markov boundary

0 引 言

因果网络(或称因果贝叶斯网络)是利用因果关系建立起来的贝叶斯网络,变量间的有向边表示的是因果关系,而非简单的概率依赖关联[1]. 因果关系在相关关系基础上严格区分出“因”与“果”,因此因果网络在揭示事物发生机制、进行干预后果推理等方面有着传统贝叶斯网络所不能替代的优势[2],已被广泛地应用于医疗诊断、经济分析、工业制造和军事应用等多个社会科学领域[3-5].

从观测数据中学习因果网络主要采用基于约束的方法,它是在因果马尔可夫假设下,先利用条件独立性检验生成因果骨架图,后采用V-结构推断边的因果方向,但该方法无法保证识别出所有边的方向,即存在马尔可夫等价类[6-7]. 为解决该难题,近年来学者们提出利用因果函数模型的非对称性推断因果方向的新思路:给定变量X和Y,分别计算因果函数模型$Y = f(X,\varepsilon)$和$X = f(Y,\eta)$,若存在$\varepsilon \perp X$且$\neg\eta \perp Y$,则推断$X \to Y$;若存在$\varepsilon \perp X$且$\neg\eta \perp Y$,则推

收稿日期: 2020-01-02；**修回日期:** 2020-06-06.
基金项目: 军委科技委理论科研项目(19JSLLKY015).
责任编委: 刘民.
†通讯作者. E-mail: lu_yunjun@hotmail.com.

断$Y \to X$. 通过对$f(*)$、噪音分布等设定特定满足条件,分别衍生出适合线性非高斯无环数据的LiNGAM类模型[8]、非线性加噪数据的ANM类模型[9-10]和后非线性数据的PNL类模型[11]等. 新的因果推断方法促进了高维因果网络定向问题的解决,文献[12]提出PCLingans算法,利用LiNGAN模型对经典PC算法未识别出方向的边进行因果定向;文献[13]采取分治策略将高维网络定向问题分解成单个节点的子网络定向问题,再利用V-结构与加噪模型ANM相结合的方法完成每个子网络的因果定向;文献[14]提出McDSL算法将因果网络结构学习分成骨架学习和方向学习两个阶段,在方向学习时先将邻节点集合的子集转换为独立的替换因素,后利用加噪模型ANM逐个推断目标节点与替换因素的因果关系. 上述因果网络定向方法虽在特定仿真数据上表现出高准确率,但其因果识别效果易受观测噪声干扰,且实际中因果函数可能存在两个方向都符合或都不符合假设的情况,使其在真实数据集上的准确率受限[10,15].

除上述因果定向方法外,文献[16]提出了一种基于柯氏复杂度的因果推断思路,核心思想是观测数据的联合概率分布按正确因果方向分解比按反方向分解具有更低的柯氏复杂度,其对因果产生机制没有过多的前提假设,具有较好的普适性. 而当前该思路主要运用在“一对一”因果关系的识别问题[17-19],文献[17]利用可计算的随机复杂度估计柯氏复杂度思想提出CICS算法,主要解决离散型变量对的因果识别问题;文献[18]结合决策树理论提出ORIGO算法,主要解决连续型变量对的因果识别问题;文献[19]在文献[18]研究基础上再结合贪婪搜索思想提出CRACK算法,主要解决混合型变量对的因果识别问题. 在这些研究成果基础上,本文先利用柯氏复杂度的链式法则推导出局部网络因果定向方法,在此基础上提出全局网络因果定向方法,并分别将其运用于马尔可夫毯和因果网络结构等学习场景,在避免产生马尔可夫等价类问题的同时,实现将基于柯氏复杂度推断因果方向的思路推广至“多对多”的因果识别情景.

1 理论基础

1.1 柯氏复杂度

柯氏复杂度(Kolmogorov complexity)是用来衡量描述对象所需要最短信息量的一个尺度. 对于给定的字符串s,图灵机U能输出s并停止的最短二进制程序代码长度称为s的柯氏复杂度[20],记为$K(s) = \min\{|p||p \in \{0,1\}^*, \mathcal{U}(p) = s\}$. 而概率分布$P(X)$的柯氏复杂度$K(P(X))$,是指在图灵机$U$上输入$X$,能输出符合精度$P(X)$并停机的最短二进制程序代码长度,而条件概率分布的柯氏复杂度定义类似[21]. 柯氏复杂度的定义虽十分简洁,但它只是一个概念性质叙述,具有不可计算性,这是因为无论当前对字符串编码算法多么简洁深刻,也无法确定是否还存在更好更优的其他算法. 但最小描述长度(the minimum description length, MDL)准则为柯氏复杂度的近似计算提供合理手段[15-19],它不再考虑所有可能编码程序,而是在已知能被计算机输出且能停止的编码规则中,寻找一个使总描述长度最小的最优编码规则,作为柯氏复杂度近似估计值.

1.2 基于柯氏复杂度的因果推断原理

对于两个相关变量X和Y,因果推断目的是区分谁为原因变量、谁为结果变量或判定两者之间只存在相关性无因果关系. 文献[16]提出用柯氏复杂度识别因果方向思路:两个变量之间具有最低柯氏复杂度的方向是最有可能的因果方向. 也就是说,两个因果变量的联合概率分布按照先P(cause)后P(effect|cause)分步描述,比分解成先P(effect)后P(cause|effect)形式具有更低的柯氏复杂度. 该思路正式表达如下.

定理1(柯氏复杂度因果推断[17,22]) 若变量X和Y存在因果关系$X \to Y$,则存在$K(P(X)) + K(P(Y|X)) < K(P(Y)) + K(P(X|Y))$.

定理1是建立在输入与机制的“独立性”假设[16-17]:在$X \to Y$情形下,原因变量X的概率分布$P(X)$,与给定原因变量X时结果变量Y的条件概率分布$P(Y|X)$相互“独立”,即$P(X)$不含任何有关$P(Y|X)$的信息. 将$P(Y|X)$想象成一种能使X转换为Y的机制,当只关注该机制的自身属性不考虑输入时,上述假设便可成立. 但这种“独立性”在反方向不成立,同样在$X \to Y$情形下,$P(Y)$与$P(X|Y)$都含有$P(X)$和$P(Y|X)$两者的有关信息,由此$P(Y)$与$P(X|Y)$因存在共享信息而“不独立”. 正是这种不对称的“独立性”为因果方向识别提供了启示. 上述“独立性”只是抽象意义上的,文献[16]利用算法信息论(AIT)将$P(X)$和$P(Y|X)$之间的“独立性”形式化为算法意义上互信息为零的状态,即$I(P(X) : P(Y|X)) \stackrel{+}{=} 0$,这意味着$P(X,Y)$只由按照先$P(X)$后$P(Y|X)$的分步描述才能得到最低柯氏复杂度,由此定理1得证.

2　基于柯氏复杂度的局部网络因果定向

2.1　单个节点局部网络的设定

本文将单个节点X_T的局部网络设定为由X_T、其父子节点集$\mathrm{PC}(X_T)$及它们之间的连边构成,但不包括$\mathrm{PC}(X_T)$中各点之间的连边. 如图1(a)所示,在这个简单无向图G中,X_1的局部网络含有3条无向边$\{(X_1-X_2);(X_1-X_3);(X_1-X_4)\}$,$\mathrm{PC}(X_1)$中各点连边$\{(X_2-X_3)|X_2,X_3\in\mathrm{PC}(X_1)\}$不在$X_1$的局部网络中;同理$X_2$的局部网络含有4条无向边$\{(X_2-X_1);(X_2-X_3);(X_2-X_5);(X_2-X_6)\}$, $\mathrm{PC}(X_2)$中各点连边$\{(X_1-X_3)|X_1,X_3\in\mathrm{PC}(X_2)\}$不在$X_2$的局部网络中. 由于$\mathrm{PC}(X_T)$中各点之间的边被排除在$X_T$的局部网络之外,此时单个节点局部网络因果定向任务只限于推断出$\mathrm{PC}(X_T)$中各点是X_T的父节点还是子节点. 这将大大简化了局部网络因果定向的任务量及复杂度.

2.2　局部网络因果定向的目标函数

根据前述的柯氏复杂度因果推断原理(定理1),对于一个因果网络,如果每对变量(X,Y)都按真实因果方向确定其方向,则此时整个因果网络的联合分布具有最低的柯氏复杂度. 据此,对于X_T的局部网络,如果能从其父子节点集$\mathrm{PC}(X_T)$中正确区分出父节点集$\mathrm{PA}(X_T)$和子节点集$\mathrm{CH}(X_T)$,则X_T局部网络的联合概率分布$P(X_T,\mathrm{PC}(X_T))$按照全部正确的因果方向进行分解,从而获得最低的柯氏复杂度. 由此,单个节点X_T的局部网络因果定向的目标函数可设定为

$$\min K(P(X_T,\mathrm{PC}(X_T))). \tag{1}$$

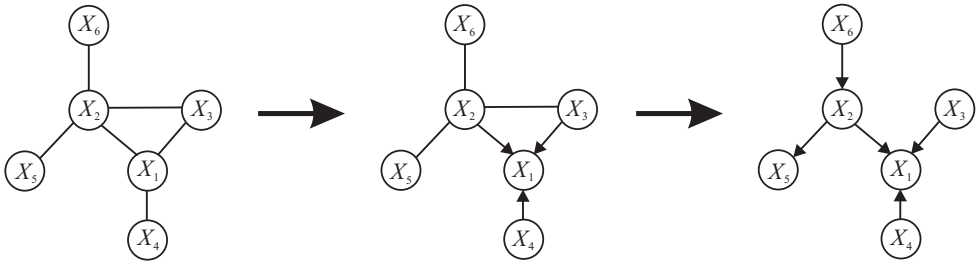
显然,在允许$\mathrm{PA}(X_T)$和$\mathrm{CH}(X_T)$为空集的情况下,X_T的局部网络共有$2^{|\mathrm{PC}(X_T)|}$种因果定向可能结果. 图1(a)中,X_1的局部网络中有3个未定向父子节点,即$\{X_2,X_3,X_4\}$,因此X_1的局部网络共有$2^{|\mathrm{PC}(X_T)|}=2^3$种因果定向可能性,如图2所示. 为判断何种可能结果更符合实际情况,需分别计算8种不同因果定向结果下$K(P(X_1,\mathrm{PC}(X_1)))$取值,其中最小值所示边定向状态即为最佳因果定向结果.

因果网络中多元变量联合分布的柯氏复杂度同样满足贝叶斯链式法则[16]

$$K(P(X_1,\ldots,X_m))=\sum_i K(P(X_i|\mathrm{PA}(X_i))). \tag{2}$$

据此,单个节点X_T的局部网络因果定向的目标函数(1)可转换为

$$\min K(P(X_T|\mathrm{PA}(X_T)))+\sum_{X_P\in\mathrm{PA}(X_T)}K(P(X_P))+\sum_{X_C\in\mathrm{CH}(X_T)}K(P(X_C|X_T)). \tag{3}$$



(a) 简单无向网络　　(b) X_1局部网络因果定向学习结果　　(c) X_2局部网络因果定向学习结果

图1　全局网络因果定向示意

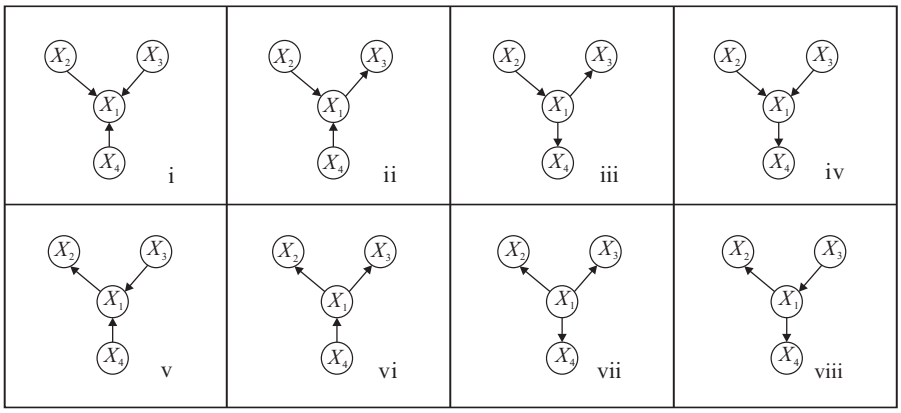


图2　X_1局部网络因果定向的8种可能结果

由式(3)所示,单个节点局部网络联合概率分布的柯氏复杂度由3部分组成: $K(P(X_T|\mathrm{PA}(X_T)))$ 代表 X_T 的父节点集 $\mathrm{PA}(X_T)$ 与 X_T 构成一个“多对一”形式概率分布的柯氏复杂度; $\sum_{X_P\in\mathrm{PA}(X_T)} K(P(X_P))$ 代表 X_T 所有父节点 X_P 概率分布的柯氏复杂度,显然在 X_T 的局部网络中 X_P 并没有父节点; $\sum_{X_C\in\mathrm{CH}(X_T)} K(P(X_C|X_T))$ 代表 X_T 与其所有子节点 X_C 构成的多个“一对一”形式概率分布的柯氏复杂度.

2.3 因果定向目标函数的求解

由于柯氏复杂度具有不可计算性,求解上述目标函数需借助MDL准则对联合概率分布的柯氏复杂度进行近似估计. MDL准则在根据码长分配函数选择最佳模型时,可利用归一化最大似然(NML)函数找到最优概率分布,最小化与真实分布之间的最大平均KL距离. 在最优概率分布下数据的编码长度可作为概率分布柯氏复杂度的近似值,也称为随机复杂度(stochastic complexity),其取值为NML函数的负对数[17,23]. 由式(3)可知,待估计柯氏复杂度可分为两种类型:单变量概率分布型 $K(P(X))$ 和条件概率分布型 $K(P(X|\mathrm{PA}(X)))$,对应的随机复杂度分别记为 $\mathrm{SC}(X)$ 和 $\mathrm{SC}(X|\mathrm{PA}(X))$. 本文将借鉴文献[23-24]中关于求解NML函数研究新成果,对多值离散类变量的随机复杂度进行高效求解,以此作为概率分布柯氏复杂度的估计值. 当面临的变量是连续型时,可先作适当离散化处理后再进行近似求解.

1) $K(P(X))$ 的近似计算.

假设单个离散型变量 X 有 m 个取值,记取值空间 $\Omega_X=\{1,2,\dots,m\}$;观察到 n 个样本数据记为 $X^n=(x_1,x_2,\dots,x_n)$,将 X 取值为 $i\in\Omega_X$ 的样本个数记为 h_i,模型类 $\mathcal{M}_m=\{P(X|\boldsymbol{\theta}):\boldsymbol{\theta}\in\Theta_m\}$, $\Theta_m=\{\boldsymbol{\theta}=(\theta_1,\dots,\theta_m):\theta_j\geqslant 0,\theta_1+\dots+\theta_m=1\}$,则

$$\begin{aligned}&K(P(X))\cong\mathrm{SC}(X)=\\&-\log P_{\mathrm{NML}}(X^n,\mathcal{M}_m)=\\&-\log\frac{P(X^n;\hat{\boldsymbol{\theta}}(X^n,\mathcal{M}_m))}{\mathcal{R}(\mathcal{M}_m,n)}=\\&-\log\frac{\prod_{i=1}^{m}\left(\frac{h_i}{n}\right)^{h_i}}{\sum_{h_1+h_2+\dots+h_m=n}\frac{n}{h_1!\dots h_m!}\prod_{i=1}^{m}\left(\frac{h_i}{n}\right)^{h_i}}.\end{aligned}\tag{4}$$

式(4)中分子项通过简单统计即可算得,而在计算分母项时,需用到文献[23]中递推公式 $\mathcal{R}(\mathcal{M}_{k+2},n)=\mathcal{R}(\mathcal{M}_{k+1},n)+\mathcal{R}(\mathcal{M}_k,n)$,此公式是基于幂和序列的生成函数与Cayley树函数相关性质证得,其中 $\mathcal{R}(\mathcal{M}_1,n)=1,\mathcal{R}(\mathcal{M}_2,n)=\sum_{r_1+r_2=n}\frac{n!}{r_1!r_2!}\left(\frac{r_1}{n}\right)^{r_1}\left(\frac{r_2}{n}\right)^{r_2}$,算法时间复杂度仅为 $O(n+m)$,属于线性阶范围.

2) $K(P(X|\mathrm{PA}(X)))$ 的近似计算.

假设离散型变量 X 有 m 个取值,取值空间 $\Omega_X=\{1,2,\dots,m\}$,有 k 个父节点记为 $\mathrm{PA}(X)=\{Y_1,\dots,Y_k\}$,将组合取值按字典顺序排列,并从1开始编号,编号记为 $\gamma(\mathrm{PA}(X))$,例如 $\mathrm{PA}(X)=\{Y_1,Y_2\}$,取值范围均为 $\{1,2\}$,则它们的取值共有4种组合,排序为 $\{Y_1=1,Y_2=1\},\{Y_1=1,Y_2=2\},\{Y_1=2,Y_2=1\},\{Y_1=2,Y_2=2\}$,则编号 $\gamma(\mathrm{PA}(X))=\{1,2,3,4\}$. 观察到 n 个样本数据 $(X^n,Y_1^n\dots Y_k^n)$,每个样本均含有 $(k+1)$ 个变量,将 X 在 $\gamma(\mathrm{PA}(X))=j$ 情形下取值为 $i\in\Omega_X$ 的样本个数记为 h_{ij},父节点 $\mathrm{PA}(X)$ 的取值共有 r 个组合,即 $|\gamma(\mathrm{PA}(X))|=r$,同时将变量 X 与其父节点集合 $\mathrm{PA}(X)$ 构成的特定局部结构记为 $\langle X_i,\mathrm{PA}(X_i)\rangle$,则

$$\begin{aligned}&K(P(X|\mathrm{PA}(X)))\cong\mathrm{SC}(X|\mathrm{PA}(X))=\\&-\log P_{\mathrm{fNML}}(X^n;\langle X,\mathrm{PA}(X)\rangle)=\\&\frac{\prod_{j=1}^{r}\prod_{i=1}^{m}\left(h_{ij}\Big/\sum_{j=1}^{r}h_{ij}\right)^{h_{ij}}}{\sum_{X^n}P(X^n|\mathrm{PA}(X));\hat{\boldsymbol{\theta}}(X^n,X_j^n\in\mathrm{PA}(X_i))}.\end{aligned}\tag{5}$$

式(5)中分子项通过统计计算也可轻易获取,而分母项可参考文献[24]将其按照父节点组合取值分解为不同情形下只涉及变量 X 的标准化项 $\mathcal{R}(\mathcal{M}_m,n)$,则分母部分转换为 $\sum_{j=1}^{r}\mathcal{R}(\mathcal{M}_m,n|\gamma(\mathrm{PA}(X))=j)$,再利用文献[23]中方法计算每个标准化项.

3 基于柯氏复杂度的全局网络因果定向

对高维复杂的全局网络进行因果定向时,需采取分治思想将其分解为每一个节点对应的局部低维网络的方向识别问题,这样不仅能降低计算复杂度,而且提高了推断的准确性[13]. 给定一个无向图 G(图1(a)),先以 X_1 为目标节点,则 X_1 的局部网络有3条待定向边,根据第2节所述局部网络因果定向方法求出最佳边定向结果如图1(b)所示;再以 X_2 为目标节点,最初 X_2 的局部网络包含4条待定向边,但经过 X_1 的局部网络因果定向学习已识得 $\{(X_2\to X_1)|X_1\in\mathrm{PC}(X_2)\}$,此时 X_2 的局部网络在继承前学习成果后只剩3条待定向边,经因果定向学

习后得到结果如图1(c)所示,至此该简单无向图G中所有边的因果方向都已识别完毕.

由上述分析可知,全局网络因果定向工作可拆分成单个节点的局部网络定向任务逐步开展. 在运用2.2节中定向目标函数对单个节点的局部网络进行因果定向时,易知该节点的度中心性越大,其父子节点集包含元素越多,目标函数能用到数据信息越多,判断的准确性也越高. 由此,给定无向图G,全局网络的因果定向应从度中心性较大节点开始,基本流程如下.

step 1: 统计无向图G中所有节点的度中心性,并按照其度中心性降序编号$\{X_1, X_2, \ldots, X_n\}$.

step 2: 依次选取第i个节点$X_i \in G$作为目标节点,根据无向图确定X_i的父子节点集合PC(X_i),先判断X_i是否与已学习过的前$(i-1)$个目标节点直接相连,若相连则继承先前阶段定向结果,即从PC(X_i)中删去编码小于i的节点. 由此,得到X_i局部网络剩余未定向边,再根据因果定向目标函数完成对X_i局部网络的因果定向.

step 3: 按照上述步骤迭代完成所有节点的局部网络因果定向,则无向图G中所有边的方向都得到识别.

4　实验分析

为充分验证前文提出的基于柯氏复杂度的因果网络定向方法(KCO法),本文分别将局部网络因果定向的方法运用到马尔可夫毯学习中,将全局网络因果定向的方法运用到因果网络结构学习中,在展示新定向方法多适用场景的同时检验其定向效果. 整个实验在R语言3.6.1版本中完成,选取Alarm(37/46)、Barley(48/84)、Hepar2(70/123)和Andes(223/338)等4个典型真实网络模型进行实验.

4.1　马尔可夫毯学习

在一个可信因果贝叶斯网络中,目标节点的马尔可夫毯(Markov boundary, MB)包括所有的父节点、子节点以及配偶节点(即子节点的父节点). 当前基于局部条件独立信息搜索学习马尔可夫毯是研究热点,它的基本思路是优先学习目标节点的父子节点集,在此基础上再利用条件独立测试寻找配偶节点,典型算法有IAMB、PCMB、IPCMB、HITON-MB、MBOR和DOS等. 相比全局搜索,局部搜索的拆分策略不仅能提高数据效率,且能推导出更多拓扑信息,如区分出父子节点集合和配偶节点集合,从父子节点分离出部分子节点等. 配偶节点只与目标节点的子节点相连接,因而利用KCO法可改进基于局部搜索的马尔可夫毯学习算法类: 在第1阶段学习到的父子节点集合基础上,运用第2节中所述局部网络因果定向方法从中筛选出子节点,而后只针对所有子节点运行相应程序推导出配偶节点集. 与原算法不区分直接对父节点和子节点都运行程序推导配偶节点集相比,改进后的算法在理论上既能提高准确率又能有效节省条件独立测试次数.

本文将KCO法融入IPCMB、HITON-MB两种典型马尔可夫毯学习算法,将改进后算法称为IPCMB(KCO)和HITON-MB(KCO). 为测试其实验效果,选择Alarm和Hepar 2等2个网络模型进行实验,利用贝叶斯网络采样法分别生成500、1 000、2 000、5 000和10 000等5个量级样本数据,每个量集样本数据随机生成10次,将运行10次实验结果取平均值作为该量级样本的最终实验效果. 同时,为有效地评价各算法的性能,引入准确率(accuracy)和CI测试数等2个评价参数,前者衡量学习到正确边的比例,定义如下:

$$\text{accuracy} = \frac{\{\text{discoveredMB}\} \bigcap \{\text{actualMB}\}}{\{\text{actualMB}\}}. \quad (6)$$

其值越高说明算法学习效果越好; CI测试数衡量算法复杂度,其值越低说明算法运行效率越高.

由表1和表2可看出,改进后算法的准确率均高于原算法,这是由于父节点未被代入推导配偶节点程序中,降低了错误配偶节点出现的概率. 在运行效率方面,改进后的算法明显优于原算法: 在Alarm网络中改进后算法IPCMB(KCO)、HITON-MB(KCO)的CI测试数分别平均下降了12.5 %、26.77 %,在Hepar2网

表1　IPCMB(KCO)和IPCMB两类算法运行效果对比

典型网络	比较指标	$n=500$		$n=1\,000$		$n=2\,000$		$n=5\,000$		$n=10\,000$	
		MB[1]	MB[2]	MB[1]	MB[2]	MB[1]	MB[2]	MB[1]	MB[2]	MB[1]	MB[2]
Alarm	Accuracy	0.760	0.791	0.823	0.8312	0.866	0.896	0.898	0.905	0.908	0.923
	CI测试数	12 119	9 824	15 003	12 774	17 092	15 320	19 467	17 770	22 000	19 560
Hepar 2	Accuracy	0.308	0.309	0.384	0.401	0.439	0.461	0.538	0.561	0.609	0.646
	CI测试数	19 994	16 303	26 520	20 761	38 284	277 780	72 149	49 066	108 811	73 752

[1] 代表原算法IPCMB, [2] 代表改进后新算法IPCMB(KCO).

表 2 HITON-MB(KCO)和HITON-MB两类算法运行效果对比

典型网络	比较指标	$n=500$		$n=1\,000$		$n=2\,000$		$n=5\,000$		$n=10\,000$	
		MB[1]	MB[2]	MB[1]	MB[2]	MB[1]	MB[2]	MB[1]	MB[2]	MB[1]	MB[2]
Alarm	Accuracy	30.764	0.798	0.814	0.869	0.881	0.905	0.900	0.923	0.917	0.936
	CI测试数	10 389	7 422	12 277	9 115	14 178	10 658	26 330	12 047	18 198	13 017
Hepar 2	Accuracy	0.334	0.358	0.426	0.446	0.494	0.510	0.568	0.595	0.633	0.651
	CI测试数	11 750	10 537	14 241	12 227	18 620	15 189	29 512	23 296	40 354	35 039

[1] 代表原算法HITON-MB, [2] 代表改进后新算法HITON-MB(KCO).

络中改进后算法IPCMB(KCO)、HITON-MB(KCO)的测试数分别下降了26.36 %、15.42 %.

4.2 因果网络结构学习

研究全局网络因果定向问题的落脚点在于提高因果网络结构学习效果,本文按照“先骨架、后定向”基本思路,先利用IAMB.FDR[25]算法中搜寻父子节点程序生成网络骨架图,后采用第3节提出的全局因果网络定向方法推断出网络中所有边的方向,将改进后算法命名为IAMB-KCO. 为测试其学习效果,本文选用IAMB-VS&ANM、PC. Stable[26]、IAMB. FDR[25]和MMHC[27]等4种因果网络结构学习方法进行对比分析,其中IAMB-VS&ANM算法则采用文献[13]高维因果网络定向方法(VS&ANM)对无向网络进行因果定向,PC.Stable和IAMB.FDR为2种较新基于约束的因果网络结构学习算法,MMHC为经典基于约束和评分的混合算法. 上述5种算法涉及的性能参数基本一致,统一设定为独立性测试,采用基于互信息方法,置信度$\alpha = 0.05$,每个节点的父子集合最大维数为学习网络的最大度中心性.

在实验中,选取Alarm、Barley、Hepar 2和Andes等4个网络模型对上述5种因果网络结构学习算法进行比较分析,样本集选取和实验次数同4.1节,评价标准选用综合评价指标

$$F_1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \tag{7}$$

F_1值为准确率和召回率之间的调和平均数,用来评价算法的总体优劣.

由图3可看出,IAMB-KCO和IAMB-VS&ANM算法明显优于传统类基于约束学习的算法PC.Stable和IAMB.FDR,因为这2种算法可以推断出更多边的方向,避免了马尔可夫等价类的出现;而混合算法MMHC效果具有强不确定性,这是由于评分函数在

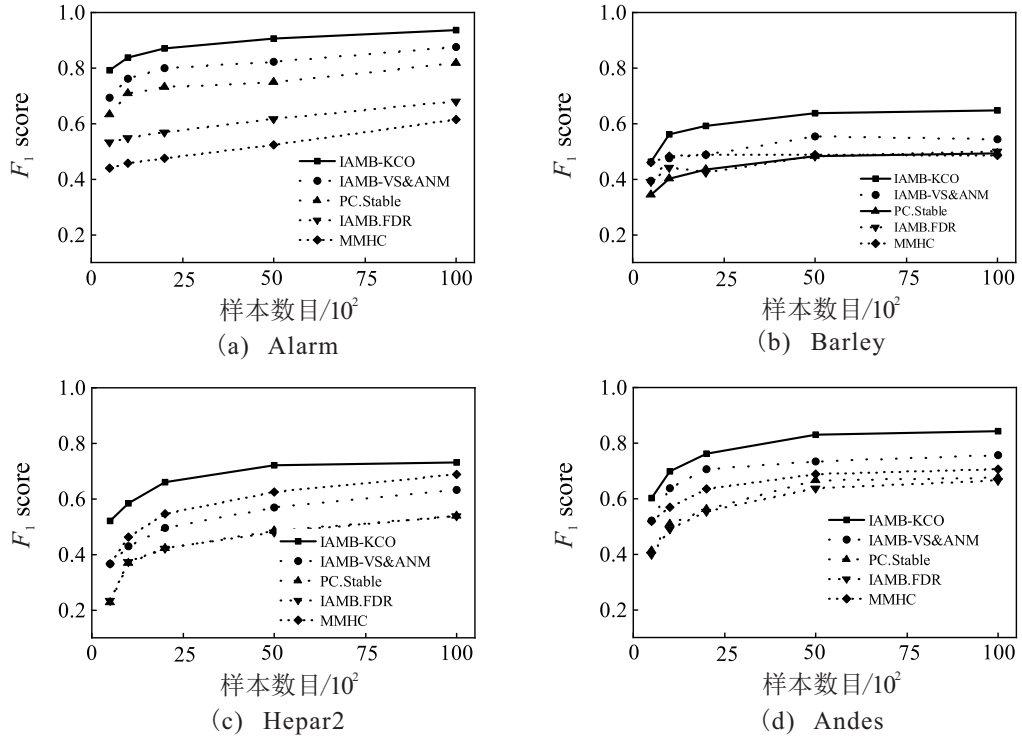


图 3 4种网络模型下因果网络结构学习算法的效果比较

确定最佳结构时并不考虑因果解释力,而是单纯追求最大化结构模型与数据拟合度,其效果严重依赖于数据质量. 对比较优的2种算法IAMB-KCO和IAMB-VS&ANM,前者在每个模型中表现更优,准确率分别提高8.83 %(Alarm)、8.94 %(Barley)、14.53 %(Hepar2)和9.12 %(Andes).

由图4可看出,在节点数不超过50的中小型网络中,5种算法运行时间相差不大,而在节点数超过50的大型网络中,IAMB-KCO算法运行速度优于其他4种算法,其在定向过程中不再依赖条件独立测试,且一次贪婪搜索能确定多条边因果方向. 由此,综合考虑精度和运行时间,IAMB-KCO算法要优于其他4种算法.

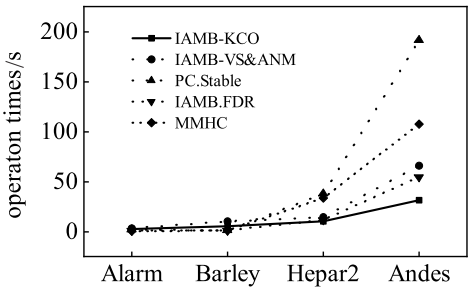


图 4　5种因果网络结构学习算法的运行时间

5　结　论

本文借鉴基于柯氏复杂度进行因果方向推断思想,先利用贝叶斯链式法则推导出有关单个节点的局部低维网络因果定向规则,解决“一对多”因果关系识别问题;后在运用分治策略,将高维因果网络定向总任务分解成每一个节点对应的局部因果网络定向子任务,依次完成所有子任务的同时也实现了“多对多”全局因果网络定向. 本文将新定向方法分别运用于马尔可夫毯、因果网络结构学习场景中,实验表明改进后算法在两个场景中在保证较高准确率的同时大大提升了执行效率. 在运用柯氏复杂度因果推断原理进行因果网络定向时,影响定向效果的一个重要因素是随机复杂度对柯氏复杂度近似估计的准确度,后续研究将集中在如何提高近似准确率,以及研究混合型变量概率分布的柯氏复杂度,拓展其应用范围.

参考文献(References)

[1] 张连文, 郭海鹏. 贝叶斯网引论[M]. 北京: 科学出版社, 2006: 36-45.
(Zhang L W, Guo H P. Introduction to bayesian networks[M]. BeiJing: Science Press, 2006: 36-45.)

[2] Pearl J. Causality: Models, reasoning, and inference[M]. Cambridge: Cambridge University Press, 2014: 21-24.

[3] Friedman N. Inferring cellular networks using probabilistic graphical models[J]. Science, 2004, 303(5659): 799-805.

[4] Li J, Le T D, Liu L, et al. From observational studies to causal rule mining[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2016, 7(2): 1-27.

[5] Mao L, Nicolae A, Oliveira M A, et al. A constraint-based modelling approach to metabolic dysfunction in parkinson's disease[J]. Computational and Structuralbiotechnology Journal, 2015, 13: 484-491.

[6] Lopez-Paz D, Muandet K, Schlopf B, et al. Towards a learning theory of cause-effect inference[C]. Proceedings of the 32nd International Conference on Machine Learning. Lille, 2015: 1452-1461.

[7] 蔡瑞初, 陈薇, 张坤, 等. 基于非时序观察数据的因果关系发现综述[J]. 计算机学报, 2017, 40(6): 1470-1490.
(Cai R C, Chen W, Zhang K, et al. A survey on non-temporal series observational data based causal discovery[J]. Chinese Journal of Computers, 2017, 40(6): 1470-1490.)

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

[9] Hoyer P O, Janzing D, Mooij J, et al. Nonlinear causal discovery with additive noise models[C]. Proceedings of the 23rd Annual Conference on Neueal Information Processing Systems. Vancouver, 2009: 689-696.

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

[11] Zhang Kun, Hyvarinen A. On the identifiability of the post-nonlinear causal model[C]. Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence. Montreal, 2009: 647-655.

[12] Hoyer P O, Hyvarinen A, Scheines R, et al. Causal discovery of linear acyclic models with arbitrary distributions[C]. Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence. Catalina Island, 2012: 282-289.

[13] 张浩, 郝志峰, 蔡瑞初, 等. 一种适用于高维网络的方向推断算法[J]. 小型微型计算机系统, 2015, 36(6): 1358-1362.
(Zhang H, Hao Z F, Cai R C, et al. An approach for inferring causal directions in high dimensional causal networks[J]. Journal of Chinese Computer Systems, 2015, 36(6): 1358-1362.)

[14] Chen W Q, Hao Z F, Cai R C, et al. Multiple-cause discovery combined with structure learnging for high-dimensional discrete data and application to stock priediction[J]. Soft Computing, 2020(11): 4575-4588.

[15] Marx A, Vreeken J. Telling cause from effect using MDL-based local and global regression[C]. 2017 IEEE International Conference on Data Mining. New Orleans, 2017: 307-316.

[16] Janzing D, Scholkopf B. Causal inference using the algorithmic Markov condition[J]. IEEE Transactions on Information Theory, 2010, 56(10): 5168-5194.

[17] Budhathoki K, Vreeken J. MDL for causal inference on discrete data[C]. 2017 IEEE International Conference on Data Mining. New Orleans, 2017: 751-756.

[18] Budhathoki K, Vreeken J. Causal inference by compression[C]. 2016 IEEE International Conference on Data Mining. Barcelona, 2016: 41-50.

[19] Marx A, Vreeken J. Causal inference on multivariate and mixed-type date by minimum description length[J]. 2017, arXiv: 1702. 06385.

[20] Grunwald P D. The minimum description length principle[M]. The MIT Press, 2007: 36-42.

[21] Li Ming, Vitanyi P. An introduction to Kolmogorov complexity and its application[J]. Texts in Computer Science, 2008, 60(3): 1017-1020.

[22] 潘孟姣, 蔡青松. 基于全局和局部回归的因果定向改进算法[J]. 计算机应用与软件, 2018, 35(10): 238-244. (Pan M J, Cai Q S. An improved causal-effect orientation algorithm based on the global and local regression[J]. Computer Applications and Software, 2018, 35(10): 238-244.)

[23] Kontkanen P, Myllymaki P. A linear-time algorithm for computing the multinomial stochastic complexity[J]. Information Processing Letters, 2007, 103(6): 227-233.

[24] Teemu Roos, Tomi Silander, Petri Kontkanen, et al. Bayesian network structure learning using factorized NML universal models[C]. Information Theory and Applications Workshop (ITA). San Diego, 2008: 272-276.

[25] Colombo D, Maathuis M H. Order-independent constraint-based causal structure learning[J]. Journal of Machine Learning Research, 2014(15): 3921-3962.

[26] Gasse M, Aussem A, Elghazel H. A hybrid algorithm for bayesian network structure learning with application to multi-Label learning[J]. Expert Systems with Applications, 2014, 41(15): 6755-6772.

[27] Leung K C, Jin H. Between silences: A voice from China[J]. World Literature Today, 1992, 66(1): 203.

作者简介

韩梦瑶 (1989−), 女, 讲师, 博士生, 从事军事运筹的研究, E-mail: 854128547@qq.com;

鲁云军 (1973−), 男, 教授, 博士生导师, 从事指挥信息系统、作战仿真与模拟训练等研究, E-mail: lu_yunjun@hotmail.com;

金乙乔 (1989−), 男, 助教, 硕士生, 从事指挥信息系统的研究, E-mail: 799380276@qq.com;

刘乾 (1989−), 男, 讲师, 博士生, 从事指挥信息系统、外军通信作战的研究, E-mail: 603473586@qq.com;

陈克斌 (1987−), 男, 讲师, 博士生, 从事军事运筹的研究, E-mail: chenkebin17@nudt.edu.cn.

(责任编辑：孙艺红)