一种基于因果强度的局部因果结构主动学习方法

周冬梅 王 浩 姚宏亮 李俊照 张 赞
(合肥工业大学计算机与信息学院 合肥 230009)

摘 要 因果结构学习是贝叶斯网络学习中一种重要的结构学习方法,因果关系揭示了系统要素作用的本质。由于仅利用观测数据很难准确地发现变量间的因果关系,且通常人们仅关心网络中关于某一变量的局部因果关系,因此针对难以从观测数据中仅获取所感兴趣的变量的局部因果结构的问题,提出了一种局部结构学习方法,即一种基于因果强度的局部因果结构主动学习方法(CSI-LCSL)。CSI-LCSL方法融合了马尔可夫毯的结构划分能力和扰动学习的因果发现能力,并且引入了因果强度进行扰动结点的选择。利用HITON_MB算法寻找目标结点的马尔可夫毯,生成关于目标结点的局部模型;然后,利用不对称信息熵对局部模型中的每一结点进行因果强度分析,选取因果强度值较大的结点进行扰动,生成扰动数据;进而,联合扰动数据和观测数据利用准确方法(exact method)学习边的后验概率,从而获得一个关于目标结点的局部因果网络。利用结构信息熵对CSI-LCSL方法的学习结果进行评估。在标准网络上的实验结果证实了CSI-LCSL算法的有效性。

关键词 因果结构,特征选择,扰动学习,贝叶斯网络,因果强度
中图法分类号 TP18 **文献标识码** A

Local Causal Structural Active Learning Method Based on Causal Power

ZHOU Dong-mei WANG Hao YAO Hong-liang LI Jun-zhao ZHANG Zan
(School of Computer and Information, Hefei University of Technology, Hefei 230009, China)

Abstract Causal structure learning is an important causal knowledge discovery method to disclose the nature of causal interactions in the Bayesian Networks. The causal relations are difficult to be discovered by only using observation data. On the other hand, actually, we are often only interested in local causal structure about a target variable. This paper presented a local causal structure learning method by integrating feature selection into intervention called a local causal structural active learning based on causal power (CSI-LCSL). CSI-LCSL integrated the dividing structure ability of Markov blanket and causal discovery ability of intervention learning. Firstly, under the faithfulness assumption, CSI-LCSL utilized HITON-MB algorithm to obtain the Markov blanket of interested variable for generating a local model. Then, we selected a intervention variable from the local model by using non-sys entropy to generate interventional data by perfect experiments. Finally, we used an exact method algorithm to obtain a local causal structure of the interested variable by combining observational data and interventional data. A series of comparative experiments on two standard Bayesian networks show that our method has excellent learning accuracy.

Keywords Causal structure, Feature selection, Intervention learning, Bayesian networks, Causal power

因果贝叶斯网络与普通的贝叶斯网络的不同之处在于贝叶斯网络中的边是否包含因果语义。因果知识发现在很多学科中都占有很重要的位置,例如自然科学和社会科学等[1]。在知识发现过程中,如果仅利用观测数据很难获取变量间的因果关系。除此之外,学习全局的贝叶斯网络是一个NP难题。而如果仅关心贝叶斯网络中与某一变量强关联的变量,那么通过学习全局网络来发现该变量的局部因果关系是非常低效的。所以,我们需要设计一个局部结构学习方法。

基于马尔可夫毯的贝叶斯网络结构学习,是主要的局部结构学习方法。在忠实性假设的条件下,目标变量 T 的马尔科夫毯 $MB(T)$ 是唯一存在的[2];且在给定变量 T 的马尔科夫毯的条件下,变量 T 与贝叶斯网络中的其它变量间是条件独立的,从而确定与目标变量 T 关联性较强的变量集合。I-AMB(Iterative Associative Markov Blanket)算法[3]和HITON-PC/MB算法[4]是当前主要的马尔科夫毯学习算法。I-AMB算法学到的仅仅只是变量 T 的马尔科夫毯集合,不能学习马尔科夫毯集中变量之间的结构关系,因而也无法得到结构间的因果关系。相对于IAMB算法,HITON-PC/MB算法能够更有效地学习到变量 T 的马尔科夫毯。然而,IAMB算法和HITON-PC/MB算法等都不能有效地学习变量之间

到稿日期:2011-11-12 返修日期:2012-04-03 本文受国家自然科学基金(61070131,61175051)资助。
周冬梅(1986—),女,硕士生,主要研究方向为人工智能与数据挖掘,E-mail:zdm0507@163.com;王 浩(1962—),男,博士,教授,主要研究方向为人工智能与数据挖掘;姚宏亮(1972—),男,博士,副教授,主要研究方向为机器学习与数据挖掘;李俊照(1975—),男,博士生,讲师,主要研究方向为机器学习与数据挖掘;张 赞(1987—),男,硕士生,主要研究方向为人工智能与数据挖掘。

的局部因果关系。

为了能够有效地学习网络中我们感兴趣的某一变量 T 的局部因果结构，提出了一种基于马尔科夫毯和扰动学习的局部因果结构学习方法。由于我们仅关心网络中某一变量 T 的局部因果结构，因此可以利用马尔可夫毯的结构划分能力从观测数据中获取变量 T 的局部模型。学习变量 T 的马尔科夫毯可以排除与变量 T 独立的变量，从而降低学习的复杂度。

由于观测数据一般是同分布的，而因果关系通过数据分布变化来体现，因此仅利用观测数据很难识别局部模型中变量间的因果关系。例如，在贝叶斯网络中，$X\rightarrow Y$ 和 $X\leftarrow Y$ 两个结构是等价的，但在因果贝叶斯网络中这两个结构中只有一个是正确的。为了区分出这两个结构，我们需要利用扰动数据[5]。主动学习和被动学习的区别在于后者在学习的过程中数据不改变，也即观测数据，目前主要基于观测数据学习贝叶斯网络的方法有两种：基于约束的方法和基于搜索打分的方法；而前者在学习的过程中数据会发生改变，每次引入的新数据即扰动数据有助于了解扰动变量对其他变量的因果影响，能够更有效地发现网络中的因果知识。但是当前存在的扰动学习算法是针对整个贝叶斯网络来学习其因果关系，扰动结点的选取、扰动学习的计算复杂性和精度方面都还面临众多的难题。因而，利用当前的因果贝叶斯网络学习算法难以快速、精确地发现目标结点的局部因果结构。

为了进一步发现变量 T 与其马尔科夫毯集合中的变量之间的因果关系，引入扰动学习的因果发现思想。而扰动学习的方法也有多种：不完美扰动(Imperfect Intervention)、完美扰动(Perfect Intervention)[6]及不确定扰动(Uncertain Interventions)[7]。不同的扰动，对从数据中学习到的网络有不同的影响。本文采用完美扰动，当对网络中某一变量进行扰动时，就会阻断扰动结点的父结点对扰动结点的影响。

因果强度是评估因果属性的一个规范标准。如 Good 的因果微积分学[8]、陈杰的 PC 论[9]、Hiddleston 的因果强度论[10]等。而自然界中存在的扰动普遍具有非对称性，因此本文选用非对称信息熵[11]作为扰动结点选取的方法。非对称信息熵理论和结构信息熵是研究贝叶斯网络模型特性的一个重要手段。利用非对称信息熵分析网络中某一变量的变化对于网络中其它结点概率所产生的影响程度，其值越大说明该变量在网络中越重要；结构信息熵作为评估学习到的网络的好坏，当其值为 0 时，说明通过学习获得的网络与真实网络相符。

在扰动学习的过程中，我们选择的是基于非对称信息熵选择扰动结点和基于结构熵的算法停止准则。从局部模型中利用非对称信息熵选择一个扰动变量进行扰动，可以获取一组扰动数据；然后联合观测数据和扰动数据，利用准确方法来学习关于变量 T 的局部因果结构；根据边的概率计算结构信息熵和非对称信息熵，通过结构信息熵判断是否达到算法停止条件，如果条件没有达到，根据非对称信息熵的值重新选择扰动结点，如果达到条件，算法停止。在两个标准网络上进行了一系列的实验对比，结果表明 CSI-LCSL 方法具有出色的学习精度。

1 贝叶斯网络和因果贝叶斯网络

贝叶斯网络是一个二元组函数，即 $B=(G,\theta)$，其中 G 表示一个有向无环图$\langle V,E\rangle$，图 G 由结点集 $V=\{X_1,\cdots,X_n\}$ 和有向边集 E 组成，边集 E 中的每条边表示结点之间的依赖关系。参数集 $\theta=\{\theta_1,\cdots,\theta_n\}$ 表示变量的条件概率分布集，$\theta_i=P(X_i|Pa(X_i))$ 表示结点 X_i 的概率分布，$Pa(X_i)$ 表示结点 X_i 的父结点。贝叶斯网络结构具有一组条件独立性假设决定：

$$P(X_i|X_1,\cdots,X_{i-1})=P(X_i|Pa(X_i)),i=1,\cdots,n \quad (1)$$

即贝叶斯网络中的每个结点，在给定其父结点的情况下，该结点条件独立于任何其非子孙结点，因此变量集 V 的联合概率分布可表示成各个局部模型的因式形式：

$$P(V)=\prod_{i=1}^{n}P(X_i|Pa(X_i)) \quad (2)$$

因果贝叶斯网络是在贝叶斯网络的基础上引入因果语义，网络中的每个结点对应领域变量，而每条边表示父结点对子结点的直接因果影响[12]。假设存在一个因果贝叶斯网络，其结构如图 1 所示，它包含了 5 个结点。图 1 说明，如果操作结点 X_1，让其取维数范围内的不同值，那么结点 X_1 的孩子结点的概率将会受到一定的影响；如果改变 X_2 的取值，它对父结点 X_1 的概率取值不会产生影响，而其子结点 X_4 的概率取值可能会有一定的影响。

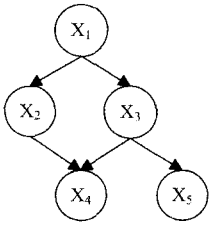


图 1 一个因果网络

2 条件独立、D 分离(D-separation)和马尔科夫毯

2.1 条件独立性

定义 1 随机变量集 $U=\{X_1,\cdots,X_n\}$，A、B 和 C 是 U 的 3 个互不相交的子集，如果对 $\forall X_i\in A$，$\forall X_j\in B$ 及 $\forall X_k\in C$ 都有 $P(X_i|X_j,X_k)=P(X_j|X_k)$，我们称给定 X_k，X_i 和 X_j 条件独立，记作 $Ind(X_i;X_j|X_k)$；反之，对 $\forall X_i\in A$，$\forall X_j\in B$ 及 $\forall X_k\in C$ 都有 $P(X_i|X_j,X_k)\neq P(X_j|X_k)$，我们称给定 X_k，X_i 和 X_j 依赖，记作 $Dep(X_i;X_j|X_k)$。

在给定变量集 Z 条件下判断两个变量是否条件独立，本文采用 G^2 测试。当 P 值小于阈值 0.05 时，判定它们条件独立；否则认为它们条件依赖。当样本数据个数是条件变量集 Z 的指数倍时，才能保证 P 测试的可靠性和准确性。

2.2 D 分离(D-separation)

定义 2 对于有向无环图(directed acyclic graph)G，设 A、B 和 C 是 G 中的 3 个互不相交的结点子集，如果 A 中的一个结点 X 和 B 中的一个结点 Y 之间的一条通路不满足下面两个条件：(1)每一个具有汇聚结点的箭头的结点均在 C 中，或有一个子孙结点在 C 中；(2)所有其它结点都不在 C 中，我们称结点 X 和结点 Y 被集合 CD 分离，能够使得 A 和 BD 分离的最小集合称为 A 和 B 的最小 D 分离集。

定理 1 给定结点 X_i 的父结点集 $Pa(X_i)$，则结点 X_i 与所有非子孙结点 X_j 独立，那么 $Pa(X_i)$ 就是 X_i 和 X_j 的 D 分离集。

2.3 马尔科夫毯

对于贝叶斯网络 $G=\langle V,E\rangle$ 和联合概率分布 $P(V)$，如果给定网络任意一个结点的父结点，该结点与它的非子孙结点独立，我们称 $\langle G,P\rangle$ 满足因果马尔科夫条件。

定义 3 对贝叶斯网络 $G=\langle V,E\rangle$ 和联合概率分布 $P(V)$，如果 G 所表示的条件独立性与 P 所表示的马尔科夫条件一一对应，则称 G 和 P 是 faithful[12]。

定义 4 对于一个变量 $T\in V$ 和变量子集 $S\subseteq U, T\notin S$，给定 T 的马尔科夫毯(记作 $MB(T)$)，存在 $I(S,T|MB(T))$，称 $MB(T)$ 为最小特征子集[12]。

定理 2 在具有忠实性的贝叶斯网络或因果概率网络中，任何变量 $MB(T)$ 都是唯一存在的[1]。

定理 3 在具有忠实性的因果概率网络中，$MB(T)$ 集中的结点是由变量 T 的父结点、子结点和子结点的父结点组成的[12]。

图 2 是一个贝叶斯网络图，识别出结点 T 的马尔科夫毯 $MB(T)=\{A,C,D,E\}$ 及父子结点 $PC(T)=\{A,D,E\}$ 均是一个局部结构发现的过程[13]。给定结点 T 的马尔科夫毯，结点 T 与网络中的其它结点是条件独立的。结点 T 与马尔科夫毯集中的变量关联性最强，因此学习关于结点 T 的一个局部因果结构的首要任务就是识别出结点 T 的马尔科夫毯。

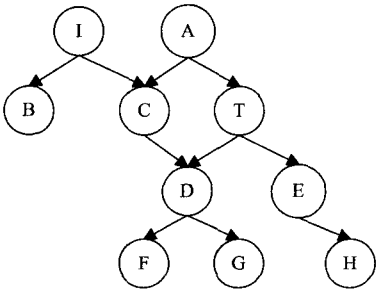


图 2 贝叶斯网络

3 局部因果结构学习

3.1 特征选择和扰动学习

在忠实性条件下，$MB(T)$ 集中的变量是使得变量 T 独立于网络中其它变量的最小切割集。当前学习马尔科夫毯的算法有很多，例如 HITON_MB，但是该算法通过观测数据学到的仅仅只是变量 T 的局部模型，不能学习马尔科夫毯集中变量之间的因果结构关系。为了进一步发现局部模型中变量间的因果关系，我们引入了扰动学习。

所谓扰动就是通过外界的干扰，系统所做出的反应。对于一个网络，对其中的某个结点进行一次扰动，可能只影响某一确定结点，也可能会影响多个结点。

如果对结点 X_i 进行一次随机扰动，设置 $X_i=x_i$，那么我们需要将结点 X_i 条件概率更新为 $P(X_i|Pa(X_i),\theta)=I(X_i=x_i)$。完美扰动(Perfect Intervention)实质上是阻断了扰动结点的父结点对扰动结点的影响。

给定数据 D，有向无环图(DAG)G 的后验概率为 $P(G|D)=P(D|G)P(G)/P(D)$，$P(G)$ 是 G 的先验概率，边缘似然函数 $P(D|G)=\int P(D|G,\theta)P(\theta|G)\mathrm{d}(\theta)$。

当参数具有均匀的 Dirichlet 分布时，那么边缘似然函数 $P(D|G)=\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij}+N_{ij})}\prod_{k=1}^{r_i}\frac{\Gamma(\alpha_{ijk}+N_{ijk})}{\Gamma(\alpha_{ijk})}$，其中 $N_{ijk}=\sum I$ 是关于数据的计算 $(X_i=k, Pa(X_i)=j)$ ($I(e)$ 是指示函数，如果事件 e 成立，$I(e)=1$，否则 $I(e)=0$)；$N_{ij}=\sum_{k=1}N_{ijk}$；α_{ijk} 是 Dirichlet 分布的超级参数，$\alpha_{ij}=\sum_{k=1}\alpha_{ijk}$；$q_i$ 是 $Pa(X_i)$ 的状态数量；r_i 是 X_i 的状态数量。

图 3 给出扰动模式，图 3(a)表示在无扰动情况下一个由 3 个结点构成的贝叶斯网络，其中 X_i^n 表示在第 n 个实例中结点 i 的值，$i=1:d, n=1:N$；θ_i 是结点 i 的条件概率分布参数；图 3(b)和图 3(a)表示的是同一个贝叶斯网络，区别在于图 3(b)中分别给结点 2 和结点 3 增加了一个扰动父结点 I_2 和 I_3。结点 I_i 为二值函数，取值为 0 或 1。对于结点 i，当 $I_i=0$ 时，它的条件概率分布为 θ_i^0(无扰动情况下的正常参数)；当 $I_i=1$ 时，它的条件概率分布为 θ_i^1(有扰动情况下的参数)。

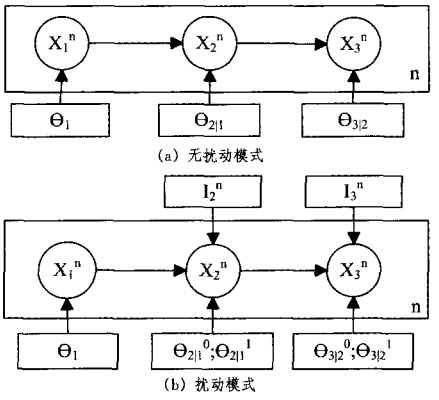


图 3 扰动模型

完美扰动发现因果结构的能力已经从理论上[5,14]和实验上[14]得以证明。对于 n 个实例，如图 4 所示，当 $I_i=0$ 时，即结点 X_i 在无扰动的情况下，其概率由其父结点 X_{G_i} 决定，此时其概率参数为 θ_i^0；当 $I_i=1$ 时，即对结点 X_i 进行一次扰动，方框内的结点 X_i 与其父结点之间的边被切断了，结点 X_i 的概率仅与扰动结点有关，其概率参数为 θ_i^1。

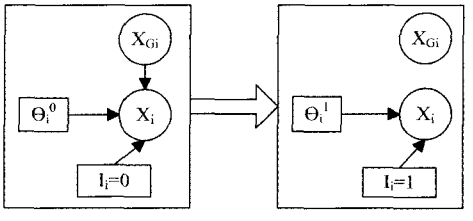


图 4 扰动模型

3.2 因果强度

在贝叶斯网络中，常用因果强度来衡量网络中结点的重要程度。网络受到干扰后，某一结点对其它结点所产生的影响值越大，说明该结点在网络中越重要。自然界中存在的扰动普遍具有非对称性，因此本文选用非对称信息熵作为扰动结点选取的方法。非对称信息熵用于理论和结构信息熵是研究贝叶斯网络模型特性的一个重要手段，而非对称信息熵用于分析网络中某一变量的变化对于网络中其它结点概率所产生的影响程度，其值越大说明该变量在网络中越重要；结构信息熵用于评估学习到的网络的好坏，当其值为 0 时说明通过学习获得的网络与真实网络相符。

扰动学习与仅利用观测数据的学习方法(又称为被动学

习)的区别在于前者在学习的过程中数据集会发生变化,而后者的数据集不改变。在扰动学习中,新加入的数据可以最大化地降低学习所得模型与实际模型的不一致性。估计出边的概率是扰动学习非常重要的一部分。对于任意结点对 X 和 Y,存在 3 种基本结构:边的方向是从 X 到 $Y(X\rightarrow Y)$;边的方向是从 Y 到 $X(X\leftarrow Y)$;X 和 Y 之间不存在边($X\perp Y$)。给定数据集 D,边 $X\rightarrow Y$ 的概率可定义为如下形式[15]:

$$P(X\rightarrow Y|D)=\sum_{X\rightarrow Y\in E(G)}P(G|D) \tag{3}$$

式中,$P(G|D)$是给定数据集 D、贝叶斯网络 G 的概率,$E(G)$是网络 G 的边集。对于 $X\leftarrow Y$ 和 $X\perp Y$ 的概率定义形如 $X\rightarrow Y$。而边的熵定义如下:

$$H(X,Y)=-P(X\rightarrow Y)*\log P(X\rightarrow Y)-P(X\leftarrow Y)*\log P(X\leftarrow Y)-P(X\perp Y)*\log P(X\perp Y) \tag{4}$$

贝叶斯网络 G 的结构熵定义如下:

$$H(G)=\sum_{X,Y}H(X,Y) \tag{5}$$

Tong 利用 Markov Chain Monte Carlo(MCMC)来近似计算边的概率。由于准确的边的概率可以为扰动结点的选择提够更多的信息,我们利用 Koivisto 提出的 exact method 来计算边的概率。Koivisto 利用前向后向动态规划方法和快速阶段叠变换在 $O(n2^n)$时间内估计网络中所有边的概率,其中 n 是网络中变量的数。

3.3 基于不对称熵的扰动结点的选择

根据边的概率来计算每一个结点的不对称熵。由于不对称熵的值越大,该结点发生变化时对其它结点的影响也越大,因此选取不对称熵的值较大的结点来对其进行扰动。贝叶斯网络 G 中任意结点的不对称信息熵定义如下[11]:

$$H_{NS}(X)=\sum_{Y}\begin{pmatrix}-P(X\rightarrow Y)*\log(P(X\rightarrow Y))\\-(1-P(X\rightarrow Y))*\log(1-P(X\rightarrow Y))\end{pmatrix} \tag{6}$$

贝叶斯网络 G 中任意结点的对称信息熵定义如下:

$$H_S(X)=\sum_{Y}H(X,Y) \tag{7}$$

不对称信息熵 $H_{NS}(X)$仅考虑结点 X 与其它结点间是否有边,而计算结点 X 的对称信息熵 $H_S(X)$时需要把 $X\rightarrow Y$、$X\leftarrow Y$和 $X\perp Y$ 3 种情况均考虑进来。在扰动学习中,对于一个结点对 X 和 Y,如果扰动结点 X、结点 Y 的概率发生变化,那么结点 X 是结点 Y 的原因结点;否则结点 X 不是结点 Y 的原因结点。如果同时扰动结点 X 和 Y,那么通过概率的变化很难判断结点 X 和 Y 的因果关系。所以本文采用不对称信息熵选择新的结点进行扰动。

3.4 局部结构学习的停止准则

在利用贝叶斯网络学习发现因果知识的过程中如何停止学习过程是研究中的另一个重要问题。本文利用结构信息熵作为停止准则。局部结构学习的理想状态是结构熵为 0,但实际上很难达到该标准。因此,只要结构熵在可接受范围即可。

3.5 基于扰动学习的局部因果结构学习

融合马尔可夫毯的结构划分能力和扰动学习的因果发现能力,提出了一种基于因果强度的局部因果结构主动学习方法,其在扰动学习的过程中结合了因果强度进行扰动结点的选择。HITON_MB 算法围绕变量 T 进行基于马尔可夫毯的结构划分,然后通过扰动学习的因果发现思想来进一步发现变量 T 与其马尔科夫毯集合中的变量之间的因果关系。

CSI-LCSL 算法描述:

输入:数据集 Data,贝叶斯网络 BN=(G,θ),目标变量 T

输出:关于变量 T 的局部结构

1. 初始化 TPC(T)={}
2. Repeat
3. 对所有不在 TPC(T)中的变量 X,如果 Asso(X,T)满足最大化,那么将变量 X 加入到 TPC(T)
4. for ∀X∈TPC(T)
5. 如果∃S⊆TPC(T)\{X},使得 I(X,T|S),那么变量 X 移出到 TPC(T),并且在后续操作不再考虑此变量。
6. 直到所有变量均已考虑过为止。
7. PC=TPC(T)
8. PCPC={PC∪T}
9. TMB=PC∪PCPC
10. ∀X∈TMB and ∀Y∈PC,IF ∃S⊆{Y}∪{V−{T,X}} S. T. Ind(X,T|S) Remove it
11. Nodes={T}∪TMB
12. 根据贝叶斯网络 BN=(G,θ)产生一组样本量大小为 N_Obs 的观测数据。
13. 利用可使用的数据来计算变量集 Nodes 中每个结点的边的概率和结构信息熵。
14. 检查停止准则。如果停止准则满足,则停止学习,否则继续。
15. 基于不对称熵、对称熵、随机选择结点进行扰动结点的选取。
16. 利用新选择的扰动结点,根据贝叶斯网络 BN=(G,θ)产生一组样本量为 N_Int 的扰动数据;联合观测数据和扰动数据构成一组新的可使用数据;转向步骤 13。

4 实验结果

为了评估所提出的方法 CSI-LCSL,使用了 Cancer 网[16]的数据集,它包含 5 条边和 5 个结点,变量有 2 种取值;Asia 网[16]的数据集,包含 8 条边和 8 个结点,变量有 2 种取值;且使用了 Alarm 网[16]的数据集,包含 46 条边和 37 个结点,变量值从 2 到 4 变化。

4.1 评估指标

L1 edge error 首先分析因果网络中的每一个结点对之间是否有边,然后计算网络的结构误差,L1 edge error 计算公式如下[15]:

$$L1(D)=\sum_{i=1}^{n}\sum_{j=i+1}^{n}\{I_{G'}(X_i\rightarrow X_j)(1-P(X_i\rightarrow X_j))+I_{G'}(X_i\leftarrow X_j)(1-P(X_i\leftarrow X_j))+I_{G'}(X_i\perp X_j)(1-P(X_i\perp X_j))\} \tag{8}$$

式中,$P(.)$是边的后验概率,$I_{G'}(e)$是指示函数,如果 e 为真,$I_{G'}(e)=1$,否则 $I_{G'}(e)=0$。

4.2 扰动学习的贡献

在 Cancer 网和 Asia 网上将观测数据和扰动数据以不同的样本量进行组合,扰动数据和观测数据分别以 0,50,100,300,500 的量进行变化。利用 CSI-LCSL 来学习局部因果结构,实验采用 BDAGL 包来计算边的后验概率[17]。

表 1 和表 2 给出了 L1 edge error 的结果。由表可知,当扰动样本为 0、观测样本从 50 到 500 变化时,结果对应表中第 2 列,随着样本数的增加,学习的结果越来越好;当观测数据为 0、扰动样本从 50 到 500 变化时,结果对应表中第 2 行,从第 2 行和第 2 列的对比可知,只利用观测样本学习的结果比只利用扰动数据学习的效果总体上要好;而在相同样本数情况下,联合观测样本和扰动样本的学习结果要比只利用其

中一种样本的学习效果要好。当扰动样本数与观测样本数比例差异较大时，学习到的结果趋向于某一类型样本情况下的学习结果，因为样本接近于同分布。尽管实验数据在 500 时停止，我们也可以预测，随着数据链的进一步增加，L1 edge error 的值也会迅速下降。而仅利用观测数据是无法决定因果网络中 $X \rightarrow Y$ 和 $X \leftarrow Y$ 哪个结构是正确的。

表 1　观测数据和扰动数据以不同样本量组合时，Cancer 网的结点对之间的 L1 edge error

Obs. data	Interventional data				
	0	50	100	300	500
0	—	6.8928	6.8199	6.7022	6.6785
50	6.8841	2.6171	0.6443	0.5533	0.1912
100	6.5235	0.5718	0.4642	0.4368	0.1433
300	6.4811	0.4569	0.4093	0.3976	0.1346
500	6.3412	0.4364	0.3966	0.1942	0.1279

表 2　观测数据和扰动数据以不同样本量组合时，用 CSI-LCSL 方法学习 Asia 网中结点 TBorCancer 的局部因果结构，并用 L1 edge error 来评判因果结构的错误

Obs. data	Interventional data				
	0	50	100	300	500
0	—	8.392	6.8982	4.1207	3.7416
50	8.5404	6.1326	6.0657	3.4509	3.190
100	8.2920	6.0287	5.0915	3.4125	3.0687
300	7.8273	5.7754	5.0109	3.2188	3.048
500	7.5451	5.3515	4.150	3.0595	3.0353

4.3　不同样本比例

在样本量不变的情况下，将观测数据和扰动数据以不同样本进行组合，分析观测数据和扰动数据以何种比例出现对结果最好。

图 5 的纵坐标表示利用 L1 edge error 来评判通过 CSI-LCSL 算法学到的 Alarm 网中关于结点 13 的局部因果结构的错误率，横坐标表示总数为 200 的样本，分别是观测数据和扰动数据的组合，前面的数据表示观测数据，后面的数据是扰动数据。我们只扰动 14 号结点。从图 5 中能够看出，在样本数相同的情况下，不同数据比例所获得的结果是不同的。这种结果从表 1 也可以看出，当样本为 140_60 时对应的 L1 edge error 是 0.1584，当样本为 60_40 时对应的 L1 edge error 是 0.1776，当样本为 120_80 时对应的 L1 edge error 是 0.2076，所以在样本总数相同的情况下，当观测数据和扰动数据以合适的比例出现时，所得到的结果是最好的。

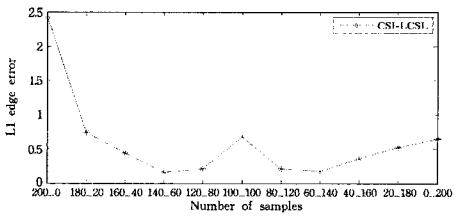


图 5　Alarm 网 13 号结点的局部因果结构中结点间 L1 edge error

4.4　比较不同的结构信息熵

Asia 是一个胸腔诊断网络，有 8 个结点。利用 HITON_MB 算法学习 Asia 网中结点 TBorCancer 的马尔科夫毯为{Bronchitis, LungCancer, TB, TBorCancer, Dys, Xray}，然后利用完美扰动学习关于结点 TBorCancer 的局部结构；完美扰动学习中扰动结点的选择是基于不对称熵的。共进行了 6 次扰动，依次扰动的结点分别 TBorCancer 、LungCancer、TB、Bronchitis、Bronchitis、TBorCancer。初始样本数为 20，进行一次扰动生成 200 个扰动数据，扰动学习的样本为 20 个观测数据和 200 个扰动数据；每扰动一次，增加 200 个扰动数据，那么样本数从 20 到 1220。无扰动学习的样本都是观测数据，每次学习所用的样本数与扰动学习的样本数相同。关于扰动的次数与结构熵之间关系的算法比较结果如图 6 所示。

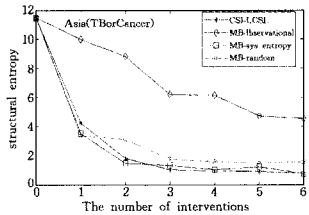
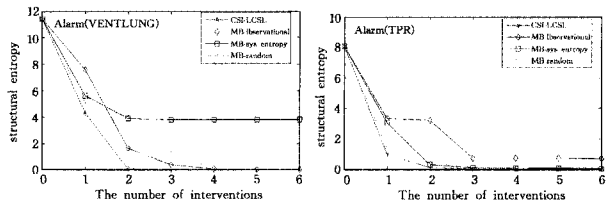


图 6　Asia 网中结点 TBorCancer 的局部因果结构的结构熵和扰动次数之间的关系

Alarm 网是一个医疗诊断网络，有 37 个结点。利用 CSI-LCSL 选取了网络中结点度较大的 VENTLUNG 和结点度适中的 TPR，对此分别学习它们的局部因果结构。CSI-LCSL 算法学习结点 VENTLUNG 的马尔科夫毯集为{INTUBATION, VENTTUBE, VENTALV, ARTCO2, EXPCO2, MINVOL, PRESS}，结点 KINKEDTUBE 也是其马尔科夫毯结点，但被丢失；结点 TPR 的马尔科夫毯集为{ARTCO2, ANAPHYLAXIS, TPR, CATECHOL, CO, BP}。在学习结点 VENTLUNG 和结点 TPR 局部结构过程中分别进行了 6 次扰动，而学习过程中样本量的变化与学习结点 TBorCancer 的局部结构样本变化相同。其扰动次数与结构熵之间的关系如图 7 所示。



(a) 结点 VENTLUNG 的局部因果结构　　(b) 结点 TPR 的局部因果结构

图 7　Alarm 网中结点的局部因果结构的结构熵和扰动次数之间的关系

图 6 和图 7 的纵坐标表示局部因果结构的结构熵，横坐标表示扰动的次数。可以看出，随着扰动次数的增加，扰动数据也相应的增加，局部因果结构的结构信息熵趋向于收敛。而且，从图中可以看出，在相同扰动次数时，CSI-LCSL 方法可以获取较低的结构信息熵。利用观测数据来学习局部因果结构需要较高的结构信息熵。在图 7 中，随机扰动的收敛速度比对称信息熵的收敛速度要快，因为基于对称信息熵选取结点时图(a)总是选取叶子结点 EXPCO2 作为扰动结点，图(b)选取叶子结点 ANAPHYLAXIS 作为扰动结点。而扰动叶子结点不会改变其它结点的概率分布，因此无法获取有效的因果信息。利用随机扰动可能会选取非叶子结点，如果扰动结点到某个结点有边，那么利用扰动数据学习后，该边的概率会相应的增加，而该结点对应的不对称信息熵会减少；但对称信息熵却不一定，因此可以提供有用的因果信息帮助学习局部因果结构。

结束语　局部因果结构学习揭示了自然界的因果交互作用。CSI-LCSL 方法利用马尔科夫毯的划分能力，排除与变量

T 独立的变量，构建一个关于感兴趣变量的局部模型；然后，基于非对称信息熵从局部模型中选择一个对其它结点影响最大的结点进行扰动；最后联合观测数据和扰动数据，在参数满足 Dirichlet 分布的条件下，计算边缘似然函数，进而计算边的后验概率，从而获得一个关于目标结点的局部因果网络。实验结果表明：CSI-LCSL 方法能够解决学习全局的贝叶斯网络 NP 难题问题，同时能够有效地学习到感兴趣变量的局部因果结构。

参 考 文 献

[1] Cooper G, Yoo C. Causal discovery from a mixture of data[C]// Proceeding of the 15th Annual Conference on UAI. San Francisco, CA: Morgan Kaufmann Publishers Inc., 1999: 116-125

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

[3] Tsamardinos I, Aliferis C F, Statnikov A R. Algorithms for Large Scale Markov Blanket Discovery[C]// Proceedings of the Sixteenth International Florida Artificial Intelligence Research Society Conference. Florida: AAAI Press, 2003: 376-380

[4] Aliferis C F, Tsamardinos I, Statnikov A R. HITON: A novel Markov blanket algorithm for optimal variable selection[C]// American Medical Informatics Association Annual Symposium. 2003: 21-25

[5] Eberhardt F, Glymour C, Scheines R. On the number of experiments sufficient and in the worst case necessary to identify all causal relations among N variables[C]// Proceedings of the 21th Annual Conference on UAI. US: AUAI Press, 2005: 178-184

[6] Korb K B, Hope L, Nicholson A E, et al. Varieties of causal intervention[C]// Proceedings of the Pacific Rim International Conference on AI. Berlin: Springer, 2004: 322-331

[7] Eaton D, Murphy K. Exact Bayesian structure learning from uncertain interventions[C]// Proceeding of the 12th on Artificial Intelligence and Statistics. JMLR Press, 2007: 107-114

[8] Good I J. A causal calculus[J]. The British Journal for the Philophy Science, 1961, 11: 305-318

[9] Cheng P W. From covariation to causation: A causal power theory[J]. Psychological Review, 1997, 104(2): 367-405

[10] Cover T M, Thomas J A. Elements of Information Theroy[M]. Wiley, 1991

[11] Li Guo-liang, Leong T-Y. Active Learning for Causal Bayesian Network Structure with Non-symmetrical Entropy[C]// Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining. Berlin: Springer-Verlag, 2009: 290-301

[12] Pearl J. Probabilistic Reasoning in Intelligent Systems[M]. San Francisco, CA: Morgan Kaufmann Publishers Inc., 1988: 383-408

[13] Aliferis C, Statnikov A, 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] Tian J, Pearl J. Causal discovery from changes[C]// Proceeding of the 17th conference on UAI. San Francisco, CA: Morgan Kaufmann Publishers Inc., 2001: 512-521

[15] Tong S, Koller D. Active Learning for Structure in Bayesian Networks[C]// IJCAI. Washington: Morgan Kaufmann, 2001: 863-869

[16] www. ai. nit. edu/murphyk/Software/BNT/bnt. html

[17] Eaton D, Murphy K. BDAGL: Bayesian DAG learning [EB/OL]. www. cs. ubc. ca /~murphyk /Software/BDAGL/, 2007

(上接第 215 页)

参 考 文 献

[1] Yan R, Zhang J, Yang J, et al. A discriminative learning framework with pairwise constraints for video object classification[C]// Proceedings of the IEEE computer society conference on computer vision and pattern recognition. Washinton, USA, 2004, 2: 284-291

[2] Tang W, Zhong S. Pairwise constraints-guided dimensionality reduction[C]// Proceedings of the Workshop on Feature Selection for DataMining (SDM 2006). Bethesda, USA, 2006: 295-311

[3] Bar-Hillel A, Hertz T, Shental N, et al. Learning a mahalanobis metric from equivalence constraints [J]. Journal of Machine Learning Research, 2006, 6(6): 937-965

[4] Zhang D, Zhou Z, Chen S. Semi-Supervised dimensionality reduction[C]// Proceedings of the SDM 2007. Minneapolis, USA, 2007: 629-634

[5] Cevikalp H, Verbeek J, Jurie F, et al. Semi-Supervised dimensionality reduction using pairwise equivalence constraints[C]// Proceedings of the VISAPP 2008. Funchal, 2008: 489-496

[6] Wei J, Peng H. Neighbourhood preserving based semi-supervised dimensionality reduction[J]. Electronics Letters, 2008, 44(20): 1190-1191

[7] Baghshah M S, Shouraki S B. Semi-Supervised metric learning using pairwise constraints[C]// Proceedings of the IJCAI 2009. SanFrancisco, USA, 2009: 1217-1222

[8] Qiao L, Chen S, Tan X. Sparsity preserving projections with applications to face recognition[J]. Pattern Recognition, 2010, 43(1): 331-341

[9] Wright J, Yang A, Sastry S, et al. Robust face recognition via sparse representation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(2): 210-227

[10] Turk M, Pentland A. Eigenfaces for recognition[J]. Journal of Cognitive Neuroscience, 1991, 3(1): 71-86

[11] Min W, Lu K, He X. Locality preserving projection[J]. Pattern Recognition, 2004, 37(4): 781-788

[12] He X, Cai D, Yan S, et al. Neighborhood preserving embedding [C]// Proceedings in International Conference on Computer Vision (ICCV). Beijing, China, 2005: 1208-1213

[13] Blake C, Keogh E, Merz C J. UCI repository of machine learning databases[Z]. Department of Information and Computer Science, University of California, Irvine, 1998

[14] Martinez A M, Benavente R. The AR Face Database[R]. CVC Technical Report 24, June 1998