计算机科学2000Vol. 27№. 10

# 用贝叶斯网络进行因果分析*)

Bayesian Causal Analysis

王双成　林士敏　陆玉昌
(清华大学计算机科学与技术系　北京100084)

**Abstract**　The Bayesian causal analysis includes two techniques, one of which takes advantage of Bayesian network structure learning under the Causal Markov assumption and the presupposition that hidden variables are absent, and the other uses canonical form influence diagram. The two techniques possess their distinctive characteristics, and ought to be selected and put to use in the light of specific conditions.

**Keywords**　Bayesian causal analysis, Causal Markov assumption, Canonical form influence diagram

因果分析是贝叶斯网络的一个重要应用领域。因果分析不同于相关分析，无论对数据分析、扰动分析还是预测都是十分重要的。贝叶斯网络虽然有一定的因果语义(我们常常用变量的因果关系构造贝叶斯网络结构)，但贝叶斯网络是条件独立性的表示，因此我们不能不加限定地用贝叶斯网络进行因果分析。贝叶斯因果分析的理论还需要进一步的实验验证，下面从马尔科夫假设和因果贝叶斯网络两个方面探讨贝叶斯因果分析。

## 1　基于马尔科夫假设的因果分析

因果贝叶斯网络概念是指，对有向无环图 $C$ 和变量集 $\mathbf{X}$，如果：(1)$C$ 中的结点和 $\mathbf{X}$ 中的变量一一对应；(2)有一个弧从结点 $X_i$ 到结点 $X_j$ 当且仅当 $X_i$ 是 $X_j$ 的直接原因，那么 $C$ 是 $\mathbf{X}$ 的因果贝叶斯网络，简称因果网。非因果贝叶斯网络概念是指，非因果贝叶斯网络变量之间条件独立性的表示，简称非因果网或贝叶斯网络。

因果马尔科夫假设：如果关于变量集 $\mathbf{X}$ 的有向无环图是因果网，那么它也是关于变量集 $\mathbf{X}$ 的贝叶斯网络。这就是因果马尔科夫假设，简称马尔科夫假设。设 $A\in\mathbf{U}$，$B\in\mathbf{U}$，$\mathbf{U}$ 是变量集，$S$ 是变量集 $\mathbf{U}$ 的贝叶斯网络结构，如果马尔科夫假设成立，那么变量 $A$ 和 $B$ 只能有下面四种可能：(1)$A$ 是 $B$ 的原因；(2)$B$ 是 $A$ 的原因；(3)有一个隐藏变量是 $A$ 和 $B$ 的共同原因；(4)结果是由数据偏差(或倾向性)造成的。如果排除隐藏变量和数据选择偏差情况，那么就只能是 $A$ 是 $B$ 的原因或 $B$ 是 $A$ 的原因(即 $A\rightarrow B$ 或 $B\rightarrow A$)。我们根据结构似然确定弧的方向，结构似然的统计含义是：结构似然值越高，网络结构就与数据拟和得越好。如果选择弧 $A\rightarrow B$ 比选择弧 $B\rightarrow A$ 具有更高的似然，那么有理由认为 $A$ 是 $B$ 的原因比 $B$ 是 $A$ 的原因的可能性更大，而且不是 $A$ 是 $B$ 的原因就是 $B$ 是 $A$ 的原因。因此在马尔科夫假设和没有隐藏变量的前提下，经过结构似然而确定的因果关系可以看作是数据所蕴涵的因果机制的很好体现。这样我们就能够利用贝叶斯网络结构进行因果定性分析、利用后验联合分布进行因果定量分析。

### (1)贝叶斯网络学习

·先验　在贝叶斯网络学习中有三个先验分布：先验参数分布、先验联合分布和先验结构概率。先验参数分布一般指定为迪里克莱分布。先验联合分布由用户构造的先验贝叶斯网络来确定。先验结构概率有两种处理方法，一种是认为先验结构是等可能的，这时在结构学习中可以忽略；另一种是认为结构先验不是等可能的，这时对每一个网络结构都要指派具体的先验结构概率，比较常用的先验结构概率指派是 $P(S^h|\xi)=c\kappa^d$(用 $\Pi_i(S)$ 和 $\Pi_i(P)$ 分别表示变量 $X_i$ 在网络结构 $S$

*)国家重点基础研究发展计划项目、国家自然科学基金项目、"九五"国家攀登计划预选项目。王双成　副教授，访问学者，研究方向：数据采掘与知识发现。林士敏　副教授、访问学者、研究方向：数据采掘与知识发现。陆玉昌　教授，研究方向：知识发现，机器学习，知识工程。

和 $P$ 中对应的父结点集合，对任意的变量 $X_i \in U$，记 $\delta_i$ 表示在 $\Pi_i(S)$ 和 $\Pi_i(P)$ 中对应的不同的结点的个数，即 $\delta_i=(\Pi_i(S)\cup\Pi_i(P))\setminus(\Pi_i(S)\cap\Pi_i(P))$，令 $\delta=\sum_{i=1}^{n}\delta_i$，$c$ 是正规常数，可以忽略，$\kappa$ 是网络结构惩罚因子（$0<\kappa\leqslant 1$），这种结构先验与弧的方向无关。

*后验* 当数据完整时，后验参数分布，后验联合分布和后验结构概率分别是公式(1)、(2)、(3)：

$$P(\theta_i|D,S^h)=\prod_{j=1}^{q_i}\frac{\Gamma(\sum_{k=1}^{r_i}(N'_{ijk}+N_{ijk}))}{\prod_{k=1}^{r_i}\Gamma(N'_{ijk}+N_{ijk})}\prod_{k=1}^{r_i}\theta_{ijk}^{N'_{ijk}+N_{ijk}-1} \tag{1}$$

$$P(X|D,S^h)=\prod_{i=1}^{n}\frac{N'_{ijk}+N_{ijk}}{N'_{ij}+N_{ij}} \tag{2}$$

$$P(S^h|D)=c\cdot\kappa^{\delta}\cdot\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{\Gamma(N'_{ij})}{\prod_{k=1}^{r_i}\Gamma(N'_{ij}+N_{ij})}\prod_{k=1}^{r_i}\frac{\Gamma(N'_{ijk}+N_{ijk})}{\Gamma(N'_{ijk})} \tag{3}$$

其中 $N'_{ij}=\sum_{k=1}^{r_i}N'_{ijk}$ 和 $N_{ij}=\sum_{k=1}^{r_i}N_{ijk}$，$N'_{ijk}=N'P(x_i^k,Pa_i^j|S_c^h,\xi)$，通常把 $N'$ 称为等价样本大小，是用户对先验贝叶斯网络（用户根据先验知识所构造的贝叶斯网络）信任程度的度量，由用户指定。$P(x_i^k,Pa_i^j|S_c^h,\xi)$ 是通过先验贝叶斯网络来计算。具有最大后验结构概率的贝叶斯网络结构就是所需要的网络结构。

当数据不完整时，采用近似的方法，计算后验参数分布和后验联合分布用高斯近似公式(4)，计算后验结构概率用拉普拉斯近似公式(5)。

$$P(\theta_S|D,S^h)\propto P(D|\theta_S,S^h)P(\theta_S|S^h)$$
$$\approx P(D|\bar{\theta}_s,S^h)P(\bar{\theta}_s|S^h)\exp\{-\frac{1}{2}(\theta_S-\bar{\theta}_s)A(\theta_S-\bar{\theta}_s)^t\} \tag{4}$$

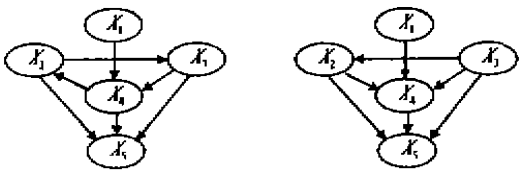
$$\log P(D|S^h)\approx\log P(D|\bar{\theta}_s,S^h)+\log P(\bar{\theta}_s|S^h)+\frac{d}{2}\log(2\pi)-\frac{1}{2}\log|A| \tag{5}$$

其中 $d$ 是 $g(\theta_S)(g(\theta_S)\equiv\log(p(D|\theta_S,S^h)\cdot P(\theta_S|S^h)))$ 的维度，$d=\prod_{i=1}^{n}q_i(r_i-1)$，$\bar{\theta}_s$ 是使 $g(\theta_S)$ 取最大值的 $\theta_S$ 的配置，可以用 EM 算法计算，$A$ 是 $g(\theta_S)$ 在 $\bar{\theta}_s$ 具有配置 $\bar{\theta}_s$ 的负海森（Hessian）矩阵，$A=(\frac{\partial g(\theta_S)}{\partial\theta_{x_i,\Pi_{x_i}}})|_{\theta_S=\bar{\theta}_s}$ 可以通过对角线法或准对角线法近似计算。

(2)应用举例

在这里，我们以一个社会调查研究的例子来说明贝叶斯网络的因果分析。首先确定变量集合，通过对某区的中学生进行调查，找出以下几个变量因素对学生的就学情况有影响：性别($X_1$)：男，女；智商($X_2$)：低，较低，较高，高；家庭经济($X_3$)：低，较低，较高，高；家庭鼓励($X_4$)：低，高；是否打算上大学($X_5$)：是，不是。下面的数据表是通过对10318名学生的统计结果，表中的第一格数据表示 $X_1$="男"，$X_2$="低"，$X_3$="低"，$X_4$="低"，$X_5$="是"的学生个数为4，第二格数据表示 $X_1$="男"，$X_2$="低"，$X_3$="低"，$X_4$="低"，$X_5$="不是"的学生个数为349，第三格数据表示 $X_1$="男"，$X_2$="低"，$X_3$="低"，$X_4$="高"，$X_5$="是"的学生个数为13，…，依此类推，在表的上半部分，$X$ 的取值都为"男"，表的下半部分，$X$ 的取值都为"女"。

我们假设没有隐藏变量且马尔科夫假设成立。所有的网络结构的先验是等可能的，等价样本大小取5，构造一先验网，用先验联合分布计算先验参数。根据我们对实际问题的理解，排除 SEX 和 SES 有父结点的情况，及 CP 有子结点的情况。我们使用拉普拉斯近似计算网络结构的后验概率，使用 EM 算法发现 $\bar{\theta}_s$，得到两个最可能的网络结构如图1所示，我们就可以进行变量之间因果定性分析。通过后验联合分布进行因果定量分析。




$\text{Log}P(D|S_1^h)\approx -45653$
$P(S_1^h|D)\approx 1.0$
$S_1$ 结构图

$\text{Log}P(D|S_2^h)\approx -45699$
$P(S_2^h|D)\approx 1.2\times 10^{-21}$
$S_2$ 结构图

图1

数据表

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 4 | 349 | 13 | 64 | 9 | 207 | 33 | 72 | 12 | 126 | 38 | 54 | 10 | 67 | 49 | 43 |
| 2 | 232 | 27 | 84 | 7 | 201 | 64 | 95 | 12 | 115 | 93 | 92 | 17 | 79 | 119 | 59 |
| 8 | 166 | 47 | 91 | 6 | 120 | 74 | 110 | 17 | 92 | 148 | 100 | 6 | 42 | 198 | 73 |
| 4 | 48 | 39 | 57 | 5 | 47 | 123 | 90 | 9 | 41 | 224 | 65 | 8 | 17 | 414 | 54 |
| 5 | 454 | 9 | 44 | 5 | 312 | 14 | 47 | 8 | 216 | 20 | 35 | 13 | 96 | 28 | 24 |
| 11 | 285 | 29 | 61 | 19 | 236 | 47 | 88 | 12 | 164 | 62 | 85 | 15 | 113 | 72 | 50 |
| 7 | 163 | 36 | 72 | 13 | 193 | 75 | 90 | 12 | 174 | 91 | 100 | 20 | 81 | 142 | 77 |
| 6 | 50 | 36 | 58 | 5 | 70 | 110 | 76 | 12 | 48 | 230 | 81 | 13 | 49 | 360 | 98 |

## 2 基于因果网的因果分析

因果网(如图2中的左图)是一种特定的规范影响图(如图2中的右图),我们通过构造规范影响图来构造因果网结构,通过一些限定,利用学习非因果贝叶斯网络参数和概率分布的方法学习因果网的参数和概率分布。

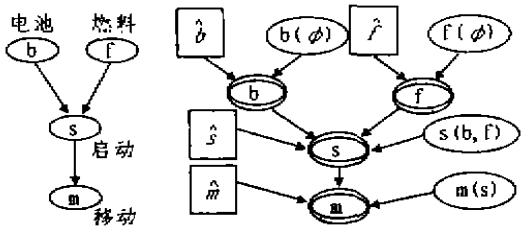



图2

**(1)基本概念和假设**

**机会变量**:机会变量表示不确定事件或随机变量,用椭圆表示,U 表示一个机会变量集。

**决策变量**:决策变量代表一个连续变量或一组离散的可行方案,用方框表示,D 表示一个决策变量集,

**决定性结点**:如果一个机会结点对应的变量是它的父结点变量的决定性函数,那么这个机会结点称为决定性结点,用双椭圆表示。

**影响图**:关于域 U∪D 的影响图是描述因果机制的图形模式,由结构组件和概率组件两部分构成。影响图的结构是一个有向无环图(如图2中的右图),除包括分别对应决策变量和机会变量的决策结点(图中的方框)和机会结点(图中的椭圆)外还包括信息弧和关联弧,信息弧指向决策结点,表示在制定决策时什么是已知的,关联弧指向机会结点,表示条件独立性的声明,在影响图中,每一机会结点 $x$ 都有一局部概率分布 $P(Pa(x),\xi)$,所有局部概率分布唯一确定联合分布 $P(U|D,\xi)$。

**规范影响图**:(1)所有响应 $D$ 的机会结点都是决策结点的后裔;(2)所有是决策结点后裔的结点都是决定性结点,满足(1)和(2)的影响图称为规范影响图。

**设置决策**:当我们设置变量为某一个状态时,如果这种设置对其它变量没有副作用(除与被设置的变量有因果交互作用的变量之外),我们就说对变量进行了设置或设置变量,如果设置的是决策变量,就称为设置决策,一个机会变量 $x$ 的设置决策用 $\hat{x}$ 表示,它的可选择值是 $x$ 的取值范围里的值和"do nothing"。

**映射变量**:映射变量 $x(Y)$ 是一个机会变量,它的状态是从 $Y$ 到 $x$ 的影射。下面是映射变量 $m(s)$ 的四种状态。



| | State1 | | state2 | | state3 | | state4 | |
|---|---|---|---|---|---|---|---|---|
| Start | no | yes | no | yes | no | yes | no | yes |
| Move | no | yes | no | yes | no | yes | no | yes |

不响应:如果不管变量 y 的值怎样变化变量 x 的值保持不变,我们就称变量 x 对变量 y 是不响应的(或无反应)。

**参数独立性**:即 $P(\theta_i|S^h,\xi)=\prod_{j=1}^{q_i}P(\theta_{ij}|S^h,\xi)$,$P(\theta_{ij}|S^h,\xi)=\prod_{k=1}^{r_i}P(\theta_{ijk}|S^h,\xi)$。

**参数模块性**:在两个网络结构 $S_1$ 和 $S_2$ 中,如果,$X_i$ 有相同的父结点,那么 $P(\theta_{ij}|S_1^h)=P(\theta_{ij}|S_2^h)$,我们称这个性质为参数的模块化特性,简称参数模块性。

**似然等价性**:如果两个网络结构 $S_1$ 和 $S_2$ 等价,那么 $P(\theta_i|S_1^h)=P(\theta_i|S_2^h)$,我们称这个性质为似然等价性,与其等价的命题是:网络结构 $S_1$ 和 $S_2$ 等价,那么 $P(D|S_1^h)=P(D|S_2^h)$。

**机制独立性**:给定变量集 $C$、变量 $x$ 和决策集 $D$,如果:(1)$x\notin C$;(2)影射变量 $x(C)$ 对 $D$ 是不响应的而且 $C$ 是最小集(或者简单地说如果 $C$ 影响 $x$ 的方式不受 $D$ 的影响,那么 $C$ 是 $x$ 的原因(关于 $D$)),当 $C$ 是 $x$ 的原因(关于 $D$)时,就称影射变量 $x(C)$ 为一因果机制,简称机制,如果不同的机制之间是独立的,就说具有机制独立性。

**组件独立性**:设 $y(X)$ 是一影射变量,$X$ 有 $q$ 个状态,我们能把其分解为变量集 $y(X=k_1),\cdots,y(X=k_q)$,我们称这些变量为机制组件(简称组件),如果这些变量之间是独立的,就说具有组件独立性。

**(2)规范影响图的构造**

给定机会变量集 U 和决策变量集 D,构造的过程是:第一步,对 $U\cup D$ 中的每一个变量都加一个结点到图中。第二步,把 U 中的变量 $x_i$ 排序,并且把不响应 D 的变量排在前面。第三步,对于每一个 U 中排过序的变量 $x_i$,如果 $x_i$ 响应 D 就加一个因果机制结点 $x_i(C_i)$(常用的确定因果机制的方法有专家方法、经验方法,数学方法及设置决策方法等)到图中,其中 $C_i\subseteq D\cup\{x_1,\cdots,x_{i-1}\}$,并且使 $x_i$ 是 $C_i\cup x_i(C_i)$ 的一个决定性的函数,第四步,建立对 D 不响应的变量的依赖关系,这样建立起来的影响图是规范影响图。影响图中的因果机制所表现的关系就是对应的因果网中的因果关系。

**(3)利用因果网进行因果分析**

因果网是一种特定的规范影响图,或者说我们能用规范影响图表示因果网所描述的因果关系。设 U 是机会变量集,$\hat{U}$ 是U的设置决策集,在规范影响图中

(下转第76页)

类的训练例的权重提高，然后再从改变权重的训练例中学习下一个分类器 $H_{t+1}$，此过程重复 $T$ 次，最后得到的分类器输出各个 $H_t$ 的输出的加权平均，权重对应于 $H_t$ 在其训练集的分类准确性。算法如下：

令权重 $w_i^{(1)}=1/N$，其中 $i=1,\cdots,N$ 是实例的序号，$T$ 为循环次数，$t=1$ 到 $T$ 执行：

* 对于权重 $w_i^{(t)}$，建立假设 $H^{(t)}:X\to[0,1]$
* 令 $H^{(t)}$ 的误差为 $\varepsilon^{(t)}=\sum_{i=1}^{N} w_i^{(t)}|y_i-h_i^{(t)}(x_i)|$
* 令 $\beta^{(t)}=\varepsilon^{(t)}/(1-\varepsilon^{(t)})$ 且 $w_i^{(t+1)}=w_i^{(t)}(\beta^{(t)})^r$，其中 $r=1-|y_i-h_i^{(t)}(x_i)|$
* 正规化 $w_i^{(t+1)}$ 使得它们的和为1.0。

假定每一个独立的分类器都是有用的，即 $\varepsilon^{(t)}<0.5$，那么 $\beta^{(t)}<1$，且当 $|y_i-h_i^{(t)}(x_i)|$ 增加时 $w_i^{(t+1)}$ 也增加。实验表明，只要增加错误分类实例的权重，改变其他细节对算法的结果没有影响。因此，Freund 等建议最后组合假设取：

$$H(x)=\frac{1}{1+\prod_{t=1}^{(a)}(\beta^{(t)})^{2h(x)-1}}$$

其中单独分类器的线性组合为：

$$l(x)=\frac{\sum_{t=1}^{(a)}(\log 1/\beta^{(t)})H^{(t)}(x)}{\sum_{t=1}^{(a)}\log 1/\beta^{(t)}}$$

提升的朴素贝叶斯分类器的性能一般说优于已经发表的使用其他学习方法的最好的结果，起码与之相当。提升算法的时间复杂度为 $O(Tef)$，其中 $T$ 是提升的次数，$e$ 是训练例的个数，$f$ 是属性的个数。

**小结** 贝叶斯分类器性能优于或相当于其他分类器，具有语义明确和容易理解的优点。其中朴素贝叶斯分类器虽然做了一个很强的关于属性之间相互条件独立的假设，而且这个假设在实际问题中往往不能满足，但是在实际应用中却取得了引人注目的成功，其性能可以同 C4.5相比，当类变量的属性值较多时，结构无约束的一般的贝叶斯网络作为分类器，分类的准确性下降，寻找更合理的评分函数是一个有待研究的问题。树扩展朴素贝叶斯网络 TAN 对朴素贝叶斯分类器作了改进，允许属性变量以另一个属性变量为父节点，取消了属性之间相互条件独立的假设，其性能优于朴素贝叶斯分类器。提升算法对朴素贝叶斯分类器也有较好的效果。

## 参 考 文 献


1 Friedman N. Bayesian Network Classifiers. Machine Learning, 1997, 29: 131～163

2 Duda R O, Hart P E. Pattern Classification and Scence Analysis, New York: John Wiley & Sons, 1973

3 Langley P, et al. An analysis of Bayesian classifiers. In: Proc. of the National Conf. on Artificial Intelligence (AAAI'92). Menlo Park, CA: AAAI Press, 1992. 223～228

4 Chow C K, Liu C N. Approximating discrete probability distributions with dependence tree. IEEE Trans. on Information Theory, 1968, 14: 462～467

5 Pearl J. Probabilistic Reasoning in Intelligent Systems. San Francisco, CA: Morgan Kaufmann, 1988. 387～390

6 Elkan C. Boosting and naive Bayesian learning: [Technical Report No. CS97-557]. Department of Computer Science & Engineering, Univ. of California, 1997


（上接第82页）

让 $\dot{x}$="do nothing"，去掉因果机制 $x(Pa(x),\dot{x})$（只是在规范影响图中不标出因果机制），那么得到的就是因果网。我们只考虑对 D 响应的机会变量（对 D 不响应的机会变量，我们认为它不具有我们讨论的因果机制或者说因果关系），我们通过构造规范影响图可以得到因果网络结构，利用因果网络结构进行因果定性分析。如果具有机制独立性假设、组件独立性假设、参数独立性假设、参数模块性假设及似然等价性假设，我们就可以用学习非因果网络参数和概率分布的方法学习因果网的参数和概率分布，利用概率联合分布进行因果定量分析。

**结束语** 因果关系理论无论在数据分析还是预测中都是非常重要的，因果关系是相互关联的变量中最强的一种关系。贝叶斯因果分析的理论还不够成熟，还需要进一步的理论探讨和实验检验。

## 参 考 文 献


1 Heckerman D. A Bayesian Approach to Causal Discovery. [Technical Report MSR-TR-97-05]. Microsoft Research, Microsoft Corporation, 1994

2 Heckerman D. A Bayesian Approach to Learning Causal Networks: [Technical Report MSR-TR-95-04]. Microsoft Research, Microsoft Corporation, 1995

3 Heckerman D. Learning Bayesian Networks: [Technical Report MSR-TR-95-02]. Microsoft Research, Microsoft Corporation, 1995

4 Heckerman D. Learaning Bayesian networks: The Combination of Knowledge and Statistical Data. Mcahine Learning, 1995, 20: 197～243