基于原因独立性的信度网推理*)

The Belief Networks Inference Based on the Causal Independence Assumption

邢永康　沈一栋

(重庆大学计算机科学与工程学院　重庆400044)

Abstract　The Causal Independence Assumption(CIA)is commonly used to reduce the size of conditional probability tables in Belief network. In this paper, we transforme a Belief networks into a new one by using the CIA. Because the structure of new Belief network is simpler than the old one, the inference on the new BN is faster than on the old Belief network.

Keywords　Belief network, Causal Independence Assumption, Inference

1. 原因独立性假设

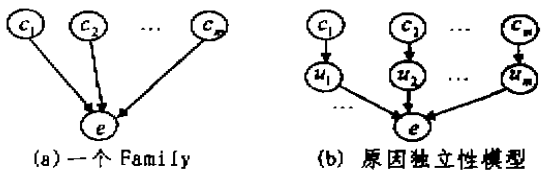
信度网的条件概率表通常都比较大，若X有n个父结点，每个父结点有m种取值，则X的条件概率表将有m^n行。为了对其简化，研究者提出了一些简化模型，如Noisy-OR模型[1]，Noisy-Adder模型[2]等。这些模型的基本思想是：可以将一个结点的父结点看作是该结点的直接原因，按照人们的思维习惯，一个原因单独对结果的影响较容易估计。因此，如果假设各个原因对结果的影响相互独立，则可以简化信度网的条件概率表。这种简化处理需要解决两个问题：

1)如何处理噪音，如原因感冒能否引起发烧这一结果，还受到诸如患者的体质差异、测量体温时的误差等一些因素的影响，这些因素统称为噪音，它们会阻止原因对结果产生作用。

2)如何将每个原因对结果的影响组合在一起，如尽管可以给出每个原因(如感冒)单独对结果(发烧)的影响，但实际中，结果(发烧)可能受到多个原因影响。

针对以上两个问题，可以将原因对结果的影响分为两个层次，第一层，为每个原因结点c_i引入一个隐结点u_i，用来表示每一个原因单独对结果的贡献，并将对应的噪声信息包含在该隐结点的条件概率表$P(u_i|c_i)$中，此时就可以假设它们之间相互独立；第二层，将这些原因对结果的影响通过一个函数$e=f(u_1,u_2,\cdots,u_m)$组合在一起，形成所有原因对结果的共同影响。如图1(a)所示的信度网中的一个Family(信度网中一个结点与其所有父结点构成的子图形称为一个Family)，它的对应模型如图1(b)所示。

定义1[3]　设信度网中的一个Family，其中$c_1,c_2,\cdots,c_m$是e的父结点。如果能找到变量集合$(u_1,u_2,\cdots,u_m)$、一系列函数$f(u_i,c_i)=P(u_i|c_i)$及函数f满足如下条件：①对每一个i，当给定c_i时，u_i与其它所有的c_j和u_j都条件独立，即：$P(u_i|c_1,\cdots,c_m,u_1,\cdots,u_m)=P(u_i|c_i)$；②$e=f(u_1,u_2,\cdots,u_m)$，则称原因$c_2,\cdots,c_m$与结果e原因独立。$c_1,c_2,\cdots,c_m$与e构成的Family称为原因独立的Family。



(a)一个Family　(b) 原因独立性模型

图1　信度网中的一个Family

定义1所描述的模型，极大地扩展了原因独立性的范围，不仅涵盖了现在已有的原因独立性模型，如Noisy-OR模型、Noisy-AND模型等，而且可以描述一些已有模型无法描述的问题。下面举例来说明这一点。

例1　假设一个学生决定选修三门课，该学生最终三门课的总成绩受到他为每门课所花费的时间多少的影响。用c_1,c_2,c_3分别表示该学生在每门课程上花费的时间，用u_1,u_2,u_3表示每门课程最终取得的成绩。则每门课所获得的成绩与其花费的时间相关，而与其它课花费的时间及获得的成绩条件独立。用e表示其三门课的总成绩，则$e=u_1+u_2+u_3$，即总成绩等于三门课成绩相加。这是Noisy-Adders模型的一个实例。

例2　对例1的情况，如果每门课的成绩采用优、中、差三级评分标准，分别用0,1,2表示，且总成绩采用如下的标准：如果三门课中有一门评分为差，则总评为差；每门课评分都为中，则总评为中；三门课中至少有

*)国家自然科学基金及教育部跨世纪优秀人才基金资助项目。邢永康　博士生，研究方向：人工智能、知识工程；沈一栋　教授，博士生导师，研究方向：人工智能。

一门为优，则总评为优。根据这些规定，则总评 $e=f(u_1,u_2,u_3)$，函数 $f(u_1,u_2,u_3)$ 由下表确定：

	00	01	02	10	11	12	20	21	22
0		0	0	0	0	0	0	0	0
1	0	0	0	0	1	2	0	2	2
2	0	0	0	0	2	2	0	2	2

该实例满足原因独立性模型，但它无法归结为任何已知的原因独立性模型。

2. 可分解原因独立性

在原因独立性的定义中，函数 $e=f(u_1,u_2,\cdots,u_m)$ 可以是任意的函数，一种特殊情况是，该函数满足以下的条件：$e=f(u_1,u_2,\cdots,u_m)=u_1*\cdots*u_m$，其中，$*$ 是一个满足交换率和结合率的二元运算符。也就是说如果函数 f 可以分解为一系列二元函数，这种特殊的原因独立性称为可分解的原因独立性，如前面的 Noisy-OR，Noisy-Adders 等，都满足这一条件。由于该函数的特殊性，它具有许多优秀的特性，不仅可以用来简化条件概率表，而且可以用于加速信度网的推理，在下面的研究中，我们提到的原因独立性都是指可分解的原因独立性。

定理1 设信度网中的一个 Family，其中 $c_1,c_2,\cdots,c_m$ 是 e 的父结点。如果该 Family 是一个原因独立 Family，则：

$$P(e=a|c_1,c_2,\cdots,c_m)=\sum_{a_1,a_2,\cdots,a_m|a_1*\cdots*a_m=a}\prod_{i=1}^{m}P(u_i=a_i|c_i) \quad (1)$$

证明：

$$P(e|c_1,c_2,\cdots,c_m)=P(u_1*\cdots*u_m=a|c_1,c_2,\cdots,c_m)$$

$$=\sum_{a_1,a_2,\cdots,a_m|a_1*\cdots*a_m=a}P(u_1=a_1,\cdots,u_m=a_m|c_1,c_2,\cdots,c_m)$$

根据原因独立性的定义可推出，在给定 $c_1,c_2,\cdots,c_m$ 时，各个隐变量之间相互独立，所以有：

$$=\sum_{a_1,a_2,\cdots,a_m|a_1*\cdots*a_m=a}P(u_1=a_1|c_1,c_2,\cdots,c_m)\cdots P(u_m=a_m|c_1,c_2,\cdots,c_m)$$

很明显，上式中 $P(u_i=a_i|c_1,c_2,\cdots,c_m)=P(u_i=a_i|c_i)$，所以

$$=\sum_{a_1,a_2,\cdots,a_m|a_1*\cdots*a_m=a}P(u_1=a_1|c_1)P(u_2=a_2|c_2)\cdots P(u_m=a_m|c_m)$$

$$=\sum_{a_1,a_2,\cdots,a_m|a_1*\cdots*a_m=a}\prod_{i=1}^{m}P(u_i=a_i|c_i) \quad \square$$

定理1表明，当一个 Family 满足原因独立性时，结果结点 e 的条件概率可以通过它的各个隐结点的条件概率计算出来，因此，当一个 Family 满足原因独立性时，利用图1(b)的模型对其进行简化，减少了需要的参数数目，并没有丢失原先条件概率表中的信息。

3. 利用原因独立性转化信度网

对于信度网中的一个满足原因独立性的 Family，根据定理1，它的结果结点的条件概率表的每一项可以通过下式计算：

$$P(e|c_1,c_2,\cdots,c_m)=\sum_{a_1,a_2,\cdots,a_m|a_1*\cdots*a_m=a}\prod_{i=1}^{m}P(u_i|c_i) \quad (2)$$

根据可分解原因独立性的定义，二元运算符 $*$ 满足结合律和交换律，因此可以通过引入一系列中间变量 $y_1,y_2,\cdots,y_{m-2}$，将 $e=u_1*\cdots*u_m$ 转换为以下的串行形式：

$$e=u_1*y_1,y_1=u_2*y_2,\cdots,y_{m-2}=u_{m-1}*u_m$$

按照该分解序列对式(2)进行转化为：

$$P(e|c_1,c_2,\cdots,c_m)=\sum_{a_1,a_2,\cdots,a_m|a_1*\cdots*a_m=a}\prod_{i=1}^{m}P(u_i|c_i)$$

$$=\sum_{a_1,a_2|a_1*a_2=a}P(u_1|c_1)\sum_{a_2,a_3|a_2*a_3=a}P(u_2|c_2)\cdots\sum_{a_{m-1},a_m|a_{m-2}*a_{m-1}=a}P(u_{m-1}|c_{m-1})P(u_m|c_m)$$

在上式计算中，用 y_i 表示中间结果，为其定义一个新的条件概率表为：如果 $Y=X*Z$，则 $P(Y|X,Z)=1$；否则，$P(Y|X,Z)=0$。将其代入继续计算有：

$$=\sum_{y_1,y_2}P'(y_1|u_2,y_2)P(u_1|c_1)\sum_{a_2,y_2}P'(y_2|u_2,y_2)P(u_2|c_2)\cdots\sum_{a_{m-1},a_m}P'(y_{m-2}|u_{m-1},u_m)P(u_{m-1}|c_{m-1})P(u_m|c_m) \quad (3)$$

上述转换过程可以表示为图2的形式。

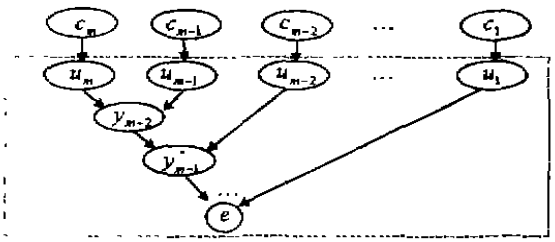


图2 二元运算树模型

图2中，虚线框内的图形结构是一个由所有的隐结点构成的特殊的树形结构，可以称为二元运算树。对其定义如下：

定义2 给定一个表达式 $e=u_1*\cdots*u_m$，其中 $*$ 是一个二元运算符，它满足结合律和交换律。与该表达式对应的二元树是一个以 e 为根结点、u_i 为叶结点的二叉树，每一个非叶结点 y_i 包含两个子结点 y_j,y_k，这

两个子结点分别以有向边与该父结点相连。父结点的条件概率表由下式确定：

$$P'(y_i\ y_j\ y_k)=\begin{cases}1 & y_i=y_j * y_k\\ 0 & 若\ y_i\neq y_j * y_k\end{cases} \tag{4}$$

对比图1(b)和图2可以看出，在图1(b)中，结点的父结点数目为m，而在图2中，由于引入了中间变量y，所以每个结点的父结点数目最大为2，我们知道在信度网推理中，结点的父结点的数目的大小是影响推理复杂性的一个主要因素。一个信度网中，父结点的数目越小，一般在它上面的推理就越简单。因此可以利用二元计算树来简化信度网推理。

对于一个信度网 $G=\{S,P\}$，其中S代表信度网的结构，P代表其条件概率表集合。如果它的所有Family都满足可分解原因独立性，则称该信度网为原因独立信度网(CI-Belief networks)。对这种信度网可以用如下的算法对其进行转化。

算法1 信度网的二元运算树转换算法。

输入：原因独立性信度网 $G=\{S,P\}$。
输出：扩展信度网 $G'=\{S',P,P'\}$
过程：

1 对信度网中的每一个Family：引入隐结点 u_i，将其转化为图1(b)所示的形式，并指定每个隐结点的条件概率 $P(u_i|c_i)$，以及确定 $e=u_1 * \cdots * u_m$ 中的二元运算符 $*$，从而获得新的信度网 $G'=\{S',P\}$。
2 对信度网 $G'=\{S',P\}$ 中的每一个原因独立结构：根据 $e=u_1 * \cdots * u_m$，分别引入中间变量 y_i，将该结构转换为图2所示的二元运算树结构，并按照式3为中间结点 y_i 指定条件概率 $P'(y_i|y_j,y_k)$，将其放入集合 P' 中，从而获得新的信度网 $G'=\{S',P,P'\}$。
3 输出结果，$G'=\{S',P,P'\}$。

下面对算法1的时间及空间复杂性作简单分析。用d表示所有变量的最大取值数目，该算法中进行结构转换需要的时间是一个常数，因此其空间和时间复杂性主要取决于为引入的隐结点和中间结点指定条件概率表，由于每个结点的父结点数目最大为2，所以为其指定条件概率表的复杂性为 $O(d^3)$。观察图2可以看出，对于每个结点的Family，引入的隐结点及中间结点的总数为 $2m-1$(具有m个叶结点的二叉树，它的结点总数为 $2n_1-1$)，由于信度网中共有n个结点，所以其总的时间及空间复杂性为 $O(n*(2m-1)d^3)$，即：$O(2nmd^3)$，所以该算法是一个多项式时间算法。

定理2 给定一个满足原因独立性的信度网G，利用算法1对其进行转换后获得信度网 G'。在这两个信度网上计算结点的信度是等价的，也就是说给定证据e，在信度网G上计算 $P(x_1|e)$ 等价于在信度网 G' 上计算 $P(x_1|e)$。

证明：假设信度网G建立在随机变量集合 $X=\{X_1,X_2,\cdots,X_n\}$ 上。给定证据e，根据信度网的特点，结点 X_1 的信度为：

$$P(x_1|e)=\alpha\sum_{x_2,\cdots,x_n}\prod_{i}P(x_i|Parent(x_i)) \tag{5}$$

由于信度网G满足原因独立性，可以利用算法1对其进行转化，获得信度网 G'。在该转化中，原信度网中的每一个Family被转换为一个二元运算树。如对结点 x_i 的Family，假设它含有m个父结点，则在转化中，引入了m个隐变量为 $U_i=\{u_1',u_2',\cdots,u_m'\}$，以及 $m-2$ 个中间变量 $Y_i=\{y_1',Y_2',\cdots,y_{m-2}'\}$。转化后，根据公式(3)有：

$$P(x_i|Parent(x_i))=P(x_i|c_1,c_2,\cdots,c_m)=\sum_{u_1,\cdots,u_m}P'(y_1|u_1,y_1)P(u_1|c_1)\sum_{u_2,\cdots,u_2}P'(y_1|u_2,y_2)P(u_2|c_2)\cdots\sum_{u_{m-1},u_m}P'(y_{m-2},u_m)P(u_{m-1}|c_{m-1})P(u_m|c_m) \tag{6}$$

将(5)代入(4)中，有：

$$P(x_1|e)=\alpha\sum_{x_2,x_3,\cdots,x_n}\prod_{i}\sum_{u_1,y_1}P'(y_1|u_1,y_1)P(u_1|c_1)\sum_{u_2,y_2}P'(y_1|u_2,y_2)P(u_2|c_2)\cdots\sum_{u_{m-1},u_m}P'(y_{m-2}|u_{m-1},u_m)P(u_{m-1}|c_{m-1})P(u_m|c_m) \tag{7}$$

用H表示在转换中引入的所有隐结点和中间结点，即 $H=U\cup Y$。其中 $U=U_1\cup U_2\cup\cdots\cup U_n$；$Y=Y_1\cup Y_2\cup\cdots\cup Y_n$。因此，信度网 G' 中的所有结点为：$Z=H\cup X$。所以可以对式(6)进一步推导为：

$$P(x_1|e)=\alpha\sum_{x_2,x_3,\cdots,x_n}\sum_{H}\prod_{z\in Z-\{X_1\}}P(z|Parent(z)) \tag{8}$$

对比等式(8)与(5)可以看出，两者的等式右边具有相同的形式，只是前者包含了对所有隐变量和中间变量的求和。由于(8)式是由(5)式推导出来的，所以在信度网G上计算 $P(x_1|e)$ 等价于在信度网 G' 上计算 $P(x_1|e)$。□

4. 基于原因独立性的信度网推理

定理2表明，对于一个满足原因独立性的信度网，可以将其转化为扩展的信度网，然后再进行信度计算，其结果保持不变。比较常用的信度网推理算法有关联树算法和桶消元算法，这里我们着重研究桶消元算法[4]。

算法2 基于原因独立性的桶消元算法：

输入：一个满足原因独立性信度网G以及证据e。
输出：结点X的信度 $P(x_1|e)$。
过程：

1 利用算法1，将该信度网G转化为扩展的信度网 G'。
2 为扩展信度网中的所有结点 $Z=U\cup H\cup X$ 确定消元顺序 $o=(z_1,z_2,\cdots,z_n)$。
3 按照i从n到1的顺序，为每个变量 z_i 设置桶 *Bucket*$_i$。找出所有的因子函数(每一个因子函数就是扩展信度网中的一个结点的条件概率)，并将每个因子函数放入它的变量中序号最大的变量对应的桶中。设 *Bucket*$_i$ 包含的因子函数为 $\{f_1^i,f_2^i,\cdots,f_j^i,\cdots\}$。
4 按照i从n到1的顺序，对每个桶 *Bucket*$_i$ 作这样的处理：如果 $z_i\in e$，则说明该变量是一个证据变量，

将它的取值代入该桶中所有的因子函数中，并将这些因子函数放入它包含的变量中序号最大的变量对应的桶中；否则，产生新的因子函数 $f_i^{new}=\sum_{z_i}\prod_{i} f_i$，并将 f_i^{new} 放入它包含的变量中序号最大的变量对应的桶中。

5 返回 $P(x_1|e)=\alpha\prod_{i} f_{z_1}$，其中 f_{z_1} 表示 x_1 对应的桶中的包含的因子函数。

由于该扩展信度网的每个结点的父结点数目最大为2，根据直觉，基于该扩展的信度网进行信度计算，其算法的复杂性应该降低。然而遗憾的是，对于任意结构的满足原因独立性的信度网，该结果是否成立，仍然无法证明，但对于单连通信度网，有如下的定理。

定理3 对于满足原因独立性的单连通信度网，假设它包含 n 个结点，每个结点的最大取值数目为 d，每个 Family 中父结点的最大数目为 m，则在此信度网上采用基于原因独立性的桶消元算法进行信度计算，其算法复杂性为 $O(nmd^3)$。

证明：基于原因独立性的桶消元算法首先将信度网转化为扩展的信度网，在单连通结构的信度网中，它的任意两个结点之间只存在一条有向路径，利用算法1对其进行转化时，只是通过引入隐结点和中间结点，将它的每一个 Family 转化为一个二元运算树，并没有改变各个结点之间的连接关系，所以获得的扩展信度网仍然是一个单连通结构的信度网。这一步转化的时间复杂性为 $O(nmd^3)$。（参见前面的分析）

基于原因独立性的桶消元算法接着在该扩展的信度网上进行消元计算。该扩展的信度网包含的结点数目最大为 $n*m$，所以在该信度网上进行的算法的复杂性为 $O(nmd^{w_o})$，其中 w_o 表示对于消元顺序 o 时推理图的宽度。对于单连通结构的信度网，可以证明，当采用其拓扑序号（父结点的序号必小于其子结点的序号）作为消元顺序时，它的推理图的宽度就是该信度网中最大的 Family 所包含结点的数目。因为该扩展的信度网中最大的 Family 包含的结点的数目为3，所以桶消元算法的复杂性为 $O(nmd^3)$。

综合以上的分析，基于原因独立性的桶消元算法的复杂性为 $O(nmd^3)$。 □

定理3表明，对于任意的单连通信度网，由于其扩展的信度网中每个结点的父结点数目最大为2，因此其推理复杂性为 $O(nmd^3)$，显著地降低了推理的复杂性。

5. 进一步研究

在满足原因独立性的信度网中，给定某些特殊的证据集合 e 时，还能够推导出一些新的证据来，这是由其特殊的结构造成的。如图1(b)所示的原因独立性模型，如果假设它是一个 Noisy-OR 模型，则 $y=u_1\vee u_2\vee\cdots\vee u_m$。当给定证据 $y=0$ 时，根据逻辑或的运算特点，必然有 $u_1=0,u_2=0,\cdots,u_m=0$，这些就是新产生的证据 e'。同样，如果假设它是一个 Noisy-AND 模型，则 $y=u_1\wedge u_2\wedge\cdots\wedge u_m$，当给定证据 $y=1$ 时，则能推导出新的证据 $u_1=1,u_2=1,\cdots,u_m=1$。对于这类特殊的证据，在扩展的信度网中，有如下的定义：

定义3 在转换的信度网中，给定证据 $y=a$，如果根据 $P'(y|y_i,y_j)$ 的定义 $y=y_i*y_j$ 可以推导出 $y_i=b$，$y_j=c$，则称证据 $y=a$ 是一个特殊证据，$y_i=b$，$y_j=c$ 为新证据。其中 y_i,y_j 表示任意的隐结点或者中间结点。

这些新产生的证据，可以被桶消元算法所利用，用来在推理中减少算法的计算量。因为桶消元算法对证据的处理比较特殊：该算法在处理一个桶时，如果该桶对应的变量是一个证据变量，则直接将该变量的取值代入该桶中所有的因子函数中，并不对所有的因子函数作求和消元运算，该处理可以在线性时间内完成，从而减少了算法处理桶的时间，提高了算法的计算效率。因此，可以对基于原因独立性的桶消元算法作如下修改：在建立了扩展的信度网之后，首先采用证据预处理算法，来产生新的证据，然后再利用桶消元算法进行消元计算，这样就可以利用到这些新产生的证据，从而提高算法在实际中的运算速度。

算法3 证据预处理算法。

输入：扩展的新度网及证据集合 e。
输出：新的证据集合 e'。
过程：
1 $e'=e$
2 重复以下计算，直到证据集合 e 为空：
(a)从 e 中取出一个证据 $y=a$。
(b)$e=e-\{y=a\}$。
(c)对所有的包含变量 y 的 $P'(y|y_i,y_j)$：如果根据 $y=y_i*y_j$，由 $y=a$ 可以推导出 $y_i=b,y_j=c$，则 $e=e\cup\{y_i=b,y_j=c\}$，且 $e'=e'\cup\{y_i=b,y_j=c\}$。
3 返回新的证据集合 e'。

参 考 文 献

1 Pearl J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Mateo, CA. 1988
2 Dagum P, Galper A. Additive Belief-Networks Models. In: Proc Ninth Conf. on Uncertainty in Artificial Intelligence, Washington D. C. 1993. 91～98
3 Zhang N L, Pool D. Exploiting Causal Independence in Bayesian Network Inference. Journal of Artificial Intelligence Research, 1996(5): 301～328
4 Dechter R. Bucket Elimination: A unifying framework for probabilistic inference. In: Proc. of the 12th Conf. on Uncertainty in Artificial Intelligence, Morgan Kaufmann, San Francisco, 1996. 211～219