

贝叶斯网络中因果链搜索算法及其应用研究^{*}

徐欣 李德玉

(山西大学计算机与信息技术学院 太原 030006)

摘要 将有向无环网络中简单路径的搜索算法用于贝叶斯网络中的因果链搜索。根据因果链上节点的条件概率表,计算出每条因果链中各节点的状态概率,从而完成了概率的传播和更新。本文所提出的因果链上的前向和后向推理方法可用于因果链敏感性分析,为科学决策提供依据。

关键词 贝叶斯网络,有向无环网络,因果链

Research on Algorithm for Detecting Cause-and-effect Chains in Bayesian Network and its Applications

XU Xin LI De-Yu

(School of Computer & Information Technology, Shanxi University, Taiyuan 030006)

Abstract Applying directed acyclic network's simple path detecting algorithm to Bayesian network's cause-and-effect chains detecting. According to the nodes' conditional probability table of the cause-and-effect chains, we can calculate the nodes' status probability and complete the probability's spread and renew in each cause-and-effect chains. Applying the onward and backward reason method in this text to the cause-and-effect chains' sensitive analysis, it can provide the basis for the scienical decision.

Keywords Bayesian network, Directed acyclic network, Cause-and-effect chains

1 引言

贝叶斯网络^[1-4] (Bayesian Network) 又称信度网 (Belief Network)、概率网 (Probability Network), 是美国加州大学 Judea Pearl 教授于 1988 年首次完整提出的, 它是用来表示变量集合连接概率的图形模型, 提供了一种表示因果信息的方法, 已逐步成为了处理不确定性信息技术的主流分支之一, 且在计算机智能科学、工业控制、医疗诊断等领域的许多智能化系统中得到了重要的应用, 是继模糊逻辑、可信度方法和神经网络等方法之后又一种不确定知识表示模型。贝叶斯网络的优点在于, 其不仅有着坚实的概率论理论基础, 同时又能够很好地表示专家的领域知识。

目前, 在基于贝叶斯网的推理研究方面, 多是从整个网络的角度进行计算。然而, 当网络中的节点个数很多, 且节点的状态也很多时, 计算相当复杂^[5-8]。

在贝叶斯网络的应用研究中, 经常需要搜索网络中某些节点之间的因果路径, 以确定这些节点之间存在怎样的因果关系。本文将将有向无环网络中简单路径搜索算法应用于贝叶斯网络中事件间的因果链搜索问题, 并以此为基础, 对可能因果链进行推理计算, 有利于考察某两个事件(节点)间不同因果链的敏感度分析, 在一定程度上简化了问题求解的复杂度。

2 贝叶斯网络

定义 1 给定随机变量集合 S 、 T 和随机变量 V , 如果式 $P(V/S, T) = P(V/T)$ 成立, 则称随机变量 V 条件独立于随机变量集 S 。

由条件概率公式, 容易得到下面的所谓链式规则。

链式规则 设 $X = \{X_1, X_2, \dots, X_n\}$ 是一个随机变量集合, 其联合概率可表示为

$$P(X_1, X_2, \dots, X_n) = \prod_{i=1}^n P(X_i / X_{i-1}, \dots, X_1)。$$

贝叶斯网络由两部分组成, 一部分是一个有向无环图, 其节点代表问题所涉及的随机变量, 节点间的有向边代表变量间的依赖或因果关系, 无边连接的节点间是条件独立的; 另一部分是反映节点间关联性强度的局部概率分布, 通常称为条件概率表, 没有父节点的节点的概率通常为该随机变量的先验概率。

定义 2 给定 n 个随机变量构成的集合 $V = \{V_1, V_2, \dots, V_n\}$, V 上的一个贝叶斯网络可形式地定义为 $BN = (G, P) = (V, L, P)$, 这里 V 为有向无环图 G 的节点, L 为其有向边的集合, P 表示条件概率分布集。

贝叶斯网络的条件独立性假设 在贝叶斯网络中, 任意节点 V_i 条件独立于该节点的父节点集以外的任意非子代节点集。

以 $P_\alpha(V)$ 表示节点 V 的父节点集, 则根据贝叶斯网络的条件独立性假设和链式规则, 贝叶斯网的联合概率分布可表示为:

$$P(V_1, V_2, \dots, V_n) = \prod_{i=1}^n P(V_i / P_\alpha(V_i))。$$

贝叶斯网络中的推理 贝叶斯网络模型可综合样本信息和专家先验知识, 不断地调整网络结构和概率信息。在贝叶斯网络中既可以进行前向推理也可以进行后向推理, 也就是说利用贝叶斯网络既可以由父节点推知子节点的状态, 也可以由子节点的各种信息推知父节点的状态。贝叶斯网络推理的本质就是利用先验知识和观测数据推知目标节点各状态下

^{*} 基金资助: 国家自然科学基金项目 No. 60573074; 山西省自然科学基金项目 No. 20041040; 山西省科技攻关项目 No. 051129; 山西省高校青年学术带头人项目。

的后验概率。理论上,如果知道随机向量的联合概率分布,则可以计算网络中感兴趣的任何事件的概率。然而,已有的结果表明,就一般的贝叶斯网络而言,计算后验概率是 NP-困难问题^[5],因此,近似推理算法成为贝叶斯网络研究领域中的一个重要问题。

3 有向无环网络中简单路径搜索算法^[9]

所谓简单路径就是序列中无重复出现节点的路径。算法的基本思想如下:利用循环和链式队列数据结构将要搜索的作为路径终点的节点及其所有祖先节点按顺序依次存入链表,然后在链表中搜索其它节点至该节点的所有简单路径。

算法设计如下:

输入:指定的节点 a(有向路径的终点)。
输出:所有到节点 a 的简单路径。
数据结构:采用队列 L, L 的节点包含三个变量 Data, No. 和 Parent。

```
BEGIN
enqueue (L, 'a');
REPEAT
```

```
L. Data; = 结点的代号;
L. No; = 该结点本身入队列时的顺序号;
L. Parent; = 该结点的父节点入队列时的顺序号;
x; = dqueue(L);
enqueue(L, x 的所有父节点);
UNTIL empty (L);
REPEAT
L. Data; = 'a';
REPEAT
output (L. Data);
L. Parent; = L. No;
UNTIL L. Parent=nil;
UNTIL L 中的节点全部循环完;
END;
```

4 案例分析

下面以经济学中农产品价格影响因果图为例对算法进行说明。图 1 给出了某种农产品价格受各种因素影响的一个贝叶斯网络,我们考虑到市场价格的所有简单路径,并分析它们对经济政策的敏感度。

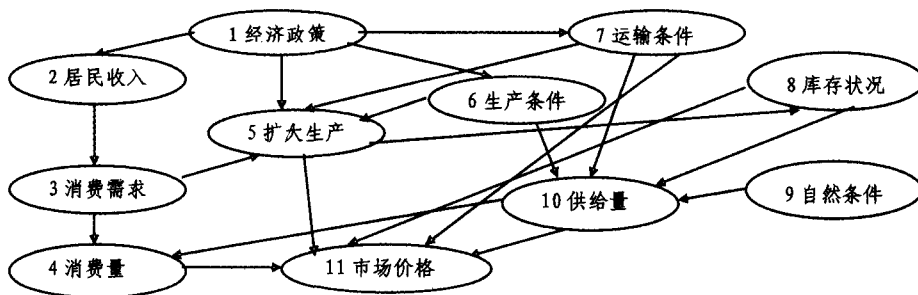


图 1 农产品价格影响因素的贝叶斯网络

表 1 结构体链表父子关系表

	链表节点														
Data	11	4	5	7	8	10	3	10	1	3	6	7	1	5	6
No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Parent	0	1	1	1	1	2	2	3	3	3	3	4	5	6	
Data	7	8	9	2	6	7	8	9	2	1	1	1	3	6	7
No.	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Parent	6	6	6	7	8	8	8	8	10	11	12	14	14	14	14
Data	1	1	5	1	1	1	5	1	2	1	1	1	3	6	7
No.	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
Parent	15	16	17	19	20	21	22	24	28	29	30	33	33	33	33
Data	1	3	6	7	1	2	1	1	2	1	1	1	1	1	
No.	46	47	48	49	50	51	52	53	54	55	56	57	58		
Parent	37	37	37	37	39	43	44	45	47	48	49	51	54		

表 2 简单路径表

路径	路径经过的节点序号	路径	路径经过的节点序号
1	1 5 11	12	1 6 10 4 11
2	1 7 11	13	1 7 5 8 11
3	1 5 8 11	14	1 7 10 4 11
4	1 6 5 11	15	1 2 3 5 8 11
5	1 6 10 11	16	1 5 8 10 4 11
6	1 7 5 11	17	1 6 5 8 10 11
7	1 7 10 11	18	1 7 5 8 10 11
8	1 2 3 4 11	19	1 2 3 5 8 10 11
9	1 2 3 5 11	20	1 6 5 8 10 4 11
10	1 5 8 10 11	21	1 7 5 8 10 4 11
11	1 6 5 8 11	22	1 2 3 5 8 10 4 11

由第 3 节给出的算法,可获得表 2 中所列出的从节点 1 到节点 11 的共 22 条因果链,换句话说,经济政策的改变可通过这 22 条因果链来影响市场价格的升高或降低。

我们感兴趣的并不是各节点的条件概率,而是某一节点

某一状态时发生的概率。在此,只考虑因果链中各节点的概率依赖关系,而忽略了非因果链中的节点对它的影响。下面以第 5 条因果链:“1 经济政策—6 生产条件—10 供给量—11 市场价格”为例说明因果链推理方法,所需的先验概率和条件概率表如表 3 所示。

(1) 由原因推知结论(前向推理)

根据全概率公式,经计算可知

$$P(\text{生产条件:变好}) = P(\text{生产条件:变好/经济政策:变好})P(\text{经济政策:变好}) + P(\text{生产条件:变好/经济政策:变差})P(\text{经济政策:变差}) = 0.9 \times 0.8 + 0.1 \times 0.1 = 0.73.$$

$$\text{同理, } P(\text{生产条件:变差}) = 0.9 \times 0.2 + 0.1 \times 0.9 = 0.27.$$

类似地,可以得到其他节点(供给量、市场价格)的状态概率,如图 2 所示。

表 3 根节点先验概率与部分节点的条件概率

1 经济政策		6 生产条件		1 经济政策	
				变好	变差
变好	0.9	变好	0.8	0.8	0.1
变差	0.1	变差	0.2	0.2	0.9

10 供给量		6 生产条件		11 市场价格	
		变好	变差	增大	减少
增大	0.9	0.9	0.0	提高	0.1
减少	0.1	0.1	1.0	降低	0.9

注:在贝叶斯网络中,条件概率表是指节点 V_i 与其父节点 P_i 。

(V_i)之间的概率关系 $P(V_i/P_e(V_i))$,由于本文只考虑因果链中概率的传播和更新,故将条件概率表简化为节点 V_i 与其在同一因果链中的父节点 $P_e(V_i)$ 之间的概率关系。因此,这里的前向推理所得到的因果链终点随机变量的状态概率只是一个粗略的估计。

同理对于第 8 条因果链:“1 经济政策—2 居民收入—3 消费需求—4 消费量—11 市场价格”,可得如图

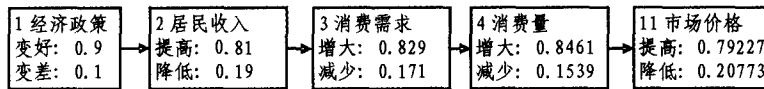


图 3 状态概率表 2

比较图 2 和图 3 的状态概率表,可知经济政策的改变通过第 8 条因果链比第 5 条因果链对市场价格的影响更大。按照这种方法,可以粗略估计每条因果链中各节点的状态概率,寻找由于经济政策的改变对市场价格影响最大或最小的路径,可以为科学决策提供依据。

(2)由结论推知原因(后向推理)

当某一事件发生,可通过弧的方向使用贝叶斯定理 $P(A/B) = \frac{P(B/A)P(A)}{P(B)}$,对各节点的状态概率进行更新。

市场价格的变化也可反过来可以通过这 22 条因果链来预测经济政策的变好或变差。

假设市场价格升高,即 $P(\text{市场价格:提高}) = 1.0$ 。由贝叶斯定理

$P(\text{供给量:增大}/\text{市场价格:提高}) = P(\text{市场价格:提高}/\text{供给量:增大})P(\text{供给量:增大})/P(\text{市场价格:提高}) = 0.1 \times 0.657/0.4087 = 0.1608$ 。同理可得, $P(\text{供给量:增大}/\text{市场价格:降低}) = 1.0$ 。由全概率公式, $P(\text{供给量:增大}) = P(\text{供给量:增大}/\text{市场价格:提高})P(\text{市场价格:提高}) + P(\text{供给量:增$

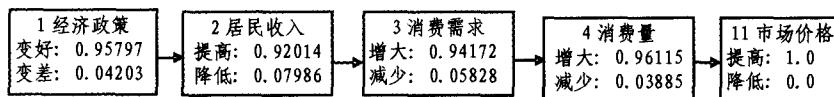


图 5 更新的状态概率表 2

比较图 4 和图 5 的状态概率表,可知当市场价格变化时,通过第 8 条因果链比第 5 条因果链更能反映出对经济政策的敏感性。粗略地估计由于市场价格的变化对每条因果链中各节点的状态概率产生的或大或小的影响,判断哪条因果链对经济政策的变化最敏感,有利于做到趋利避害。

结论 本文将有向无环网络中的简单路径搜索算法应用于贝叶斯网络的因果链搜索,提出了因果链中各节点的状态概率的传播和更新的粗略估计方法,讨论了因果链上的前向、后向推理方法,该方法对于考察因果链的敏感性具有一定的借鉴意义。在一个因果链中,影响节点状态概率的变量可能不仅仅是该节点在链中的父节点,如何在一个复杂的网络中考虑因果链上的推理是需要进一步研究的问题。

参 考 文 献

1 David H. Bayesian networks for data mining [J]. Data Mining and Knowledge Discovery, 1997, 1(1): 79~119
2 Agre G. Diagnostic Bayesian networks [J]. Computers and Arti-

3 所示的状态概率表。



图 2 状态概率表 1

大/市场价格:降低) $P(\text{市场价格:降低})$,则事件“供给量:增大”的概率可更新为

$$P(\text{供给量:增大}) = 0.1608 \times 1.0 + 1.0 \times 0.0 = 0.1608$$

与输入证据(市场价格升高)前相比,供给量增大的概率由原来的 0.657 减少到 0.1608,表明市场价格升高是由于供给量增大引起的概率明显降低了。

同理,事件“供给量:减少”的概率可更新为 0.8392。对链中的其他节点继续该过程,可得图 4 所示的状态概率表,从而完成了证据在整个链中的更新传播。

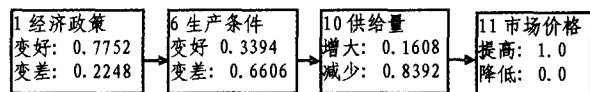


图 4 更新的状态概率表 1

同理对于第 8 条因果链,可得如图 5 所示的更新的状态概率表。

ficial Intelligence, 1997, 16(1): 47~67

3 Heckerman D, Geige D, Chickering D. Learning Bayesian networks: The combination of knowledge and statistic data [J]. Machine Learning, 1995, 20:19~243
4 de Campos Huete L M. A new approach for learning belief networks using independence criteria [J]. International Journal of Approximate Reasoning, 2000, 24:11~37
5 Dagum P, Luhy M. Approximating probabilistic inference in Bayesian belief networks is NP-hard [J]. Artificial Intelligence, 1993, 60:141~153
6 Cooper G F. The computational complexity of probabilistic inference using Bayesian belief networks [J]. Artificial Intelligence, 1990, 42(2-3): 393~405
7 余东峰,孙兆林. 基于贝叶斯网络不确定推理的研究[J]. 微型电脑应用研究与设计, 2004, 20(8):6~8
8 胡玉胜,涂序彦,崔晓瑜,程乾生. 基于贝叶斯网络的不确定性知识的推理方法[J]. 计算机集成制造系统, 2001, 7(12):65~68
9 刘夫云,祁国宁,车宏安. 复杂网络中简单路径搜索算法及其应用研究[J]. 系统工程理论与实践, 2006, 4:9~13