

文章编号:1001-9081(2011)06-1638-03

doi:10.3724/SP.J.1087.2011.01638

基于粗糙集的定性概率网整合方法

吕亚丽,石洪波

(山西财经大学 信息管理学院,太原 030031)

(yali.lv2008@gmail.com)

摘要:由于子定性概率网(QPN)仅局限于表示子领域知识,为构建一个较大QPN进行知识的全面表示,基于粗糙集理论,提出了一种具有不同节点的多个子QPN整合方法。在QPN中,可将单个变量或多个变量的组合看做粗糙集中的一属性。当多个QPN整合时,首先合并多个子QPN结构;然后,在保证不出现环路的情况下,根据粗糙集的属性间的依赖度向合并的QPN中添加有向边及其定性符号;接着,再根据属性间相对必要性来删除具有多个父节点的属性所不必要的冗余边,从而整合出较大QPN。最后,实验验证了该整合方法的可行性和有效性。

关键词:定性概率网;定性影响;粗糙集;概率下近似;属性依赖度;属性相对必要性

中图分类号:TP182 文献标志码:A

Integration method of qualitative probabilistic networks based on rough sets

LÜ Ya-li, SHI Hong-bo

(School of Information Management, Shanxi University of Finance and Economics, Taiyuan Shanxi 030031, China)

Abstract: Qualitative Probabilistic Network (QPN) is a powerful knowledge representation tool. However, sub-QPN can only represent sub-domain knowledge. To build a large QPN to represent the whole domain knowledge, an integration method of multiple sub-QPNs that have different nodes was proposed based on rough sets. Specifically, a single variable or a combination of multiple variables in a QPN could be regarded as an attribute in rough sets. First, multiple sub-QPNs were combined into an initial integrated QPN during integrating, then the directed edges and qualitative signs were added into the QPN according to attribute dependency degree, and then some unnecessary edges of which child node had multiple parent nodes could be deleted according to relative necessity of attribute. Thus, a large integrated QPN would be obtained to represent the whole domain knowledge. Finally, the experimental results illustrate that the integration method is feasible and effective.

Key words: Qualitative Probabilistic Network (QPN); qualitative influence; rough set; probabilistic lower approximation; attribute dependency degree; relative necessity of attribute

0 引言

定性概率网(Qualitative Probabilistic Network, QPN)是人工智能研究中一种重要的定性知识表示方法,它为贝叶斯网(Bayesian Network, BN)提供了一种潜在的定性分析工具,越来越受到研究者们的关注。然而,在整个知识领域,有时每个专家只可能提供一些子领域的QPN知识,因此,面对构建一个较大的QPN,将这些子领域专家提供的各子QPN整合起来显得十分必要。

针对概率网整合问题,国内外学者进行了一些相关的工作。文献[1]通过图合并和弧翻转进行了BN的结构整合;文献[2]基于独立关系的并和交研究两BN的结构整合,并采用弧翻转较好地避免了环路出现的问题;文献[3]应用扩展关系模型进行了多BN依赖结构的合并。但这些方法均涉及的是定量BN的整合研究。而对于定性网,文献[4]对其进行了整合研究,提出了一种多QPN整合方法。然而也仅仅研究的是时序环境下QPN的整合,其主要思想是以最后一个时间片的QPN为主,在不形成环路的前提下合并各个QPN。相对时序环境的普通静态环境,在文献[5]中,作者基于粗糙集理论提出了具有相同节点的多个QPN整合,首先定义了整

体影响强度的概念,然后据此概念分情况讨论结构整合时引起的环路消除问题。但对于不同节点的多个QPN的整合研究还有待进行。因而,在前期工作[5]的基础上,本文着重研究普通静态环境下,具有不同节点的多个QPN整合方法。

粗糙集理论^[6]是一种采用精确的数学方法分析不精确系统的数学分析理论。最突出的优点:“让数据自己说话”。根据所提供的数据,该理论不仅可以求解属性间相互依赖程度,很好地分析属性之间的依赖性,而且可以求解属性间的相对必要性,进行属性的相对约简。在QPN中,将单个变量或多个节点变量的组合看做粗糙集中的一属性。当多个QPN整合时,根据属性间的依赖程度来判断QPN中无边相连的两变量间是否应该添加有向边,再根据属性间的相对必要性进一步判断QPN中的冗余边,以此删除这些不必要的边,从而得到整合后的较大QPN。据此,针对普通环境下,本文提出一种基于粗糙集的具有不同节点的多个QPN整合方法。

1 定性概率网和粗糙集理论

1.1 定性概率网

定性概率网(QPN)^[7]是贝叶斯网(BN)的定性抽象,与BN具有相同的有向无环图结构,并用定性影响符号而不是以

收稿日期:2010-11-29;修回日期:2011-01-18。

基金项目:国家自然科学基金资助项目(60873100);山西省自然科学基金资助项目(2009011017-4;2010011022-1)。

作者简介:吕亚丽(1975-),女,山西临汾人,讲师,博士研究生,主要研究方向:人工智能、定性概率推理; 石洪波(1965-),女,山西太原人,教授,博士,主要研究方向:人工智能、数据挖掘。

条件概率分布值描述变量之间的相互关系。为简单起见,假设所有节点变量为二值变量,分别取值 true 和 false,且 true > false;例如,对于节点 A 的两个取值,用 a 或 $A = 1$ 表示 A 取 true,用 \bar{a} 或 $A = 0$ 表示 A 取 false。

定义1 定性影响^[7]。 $QPN = (V, E, \delta)$, V 是变量集, E 是有向边集。 $A, B \in V$, 弧 $A \rightarrow B \in E$, $\delta = \{+, -, 0, ?\}$ 是定性关系, 则变量 A 对 B 的正影响, 记做 $S^+(A, B)$, 当且仅当 $a > \bar{a}$ 时, 对变量集 X 的任意具体值 x , 有:

$$\Pr(b \mid a, x) \geq \Pr(b \mid \bar{a}, x) \quad (1)$$

其中, $X = \pi(B) \setminus \{A\}$ 为 B 除 A 外的所有父节点集。

定义1表明,当 A 的取值增加时,无论 B 的其他父节点的取值如何变化, B 的取值增加的可能性将增大。类似地, 将上式(1)中的 \geq 修改为 \leq 或 $=$, 可定义负影响 S^- 和零影响 S^0 。还有歧义影响 $S^?$, 即 A 取值增加时, B 的取值增加的概率可能增大,也可能减小。

下面给出一个由 Radiotherapy BN 根据上述定义抽象出其相应 QPN 的实例^[8], 如图 1 所示。

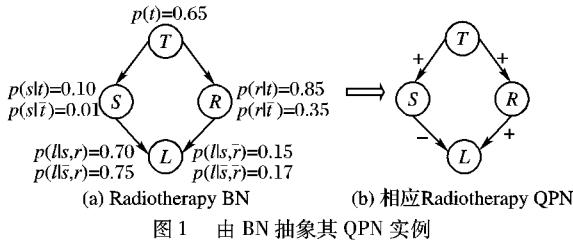


图 1 由 BN 抽象其 QPN 实例

1.2 粗糙集基本概念

定义2 集合的下近似^[9]。给定近似空间 $K = (U, R)$, 其中, U 为论域, R 表示论域上的等价关系簇, 则 $\forall X \subseteq U$ 和 U 上的一个等价关系 $R \in IND(K)$, 定义子集 X 关于 R 的下近似 $R(X)$ 为:

$$\begin{aligned} R(X) = \{x \mid (\forall x \in U) \wedge ([x]_R \subseteq X)\} = \\ \cup \{Y \mid (\forall Y \in U/R) \wedge (Y \subseteq X)\} \end{aligned}$$

$$X \text{ 的 } R \text{ 正域: } POS_R(X) = R(X).$$

定义3 集合的概率下近似^[10]。给定论域 U 及其上的一个等价关系 R , 阈值 $\beta (0.5 < \beta \leq 1)$, \Pr 是定义在 U 上的概率测度, 则 $X (X \subseteq U)$ 关于 R 的概率下近似为 $R_\beta X = \{x \in U \mid$

$$\Pr(X/[x]_R) \geq \beta\}$$
, 其中, $\Pr(X/[x]_R) = \frac{|X \cap [x]_R|}{|[x]_R|}$ 。

定义4 属性的依赖度^[9]。给定论域 U 上的两个属性 P 和 Q , 定义 $\gamma_P(Q) = k = \frac{|POS_P(Q)|}{|U|}$ 为属性 Q 依赖于 P 的程度, 简记为 $P \Rightarrow_k Q$ 。

定义5 属性的相对必要性^[9]。给定论域 U 及两个属性等价关系簇 P, Q , $\forall R \in P$, 若: $POS_{IND(P)}(IND(Q)) = POS_{IND(P-[R])}(IND(Q))$ 成立, 则称属性 R 为 P 中 Q 不必要的, R 相对于 Q 可以从 P 中删除; 否则为必要的, R 不能被删除。

2 基于粗糙集的 QPN 整合方法

2.1 问题描述

给定具有不同节点的多个 QPN 和已有的完全样本数据, 基于粗糙集理论如何将它们整合成一个较大的 QPN, 使之能更好地表示整个领域知识, 这将是本文的主要工作。

将 QPN 中的节点变量可看做粗糙集理论中的属性, 整合时, 需要根据属性间的依赖程度来确定无边相连的变量间是

否应该添加有向边, 以及根据属性间的相对必要性来判断具有多个父节点的子节点, 哪些父节点相对该子节点是不必要的, 从而删除相应的父子边。

然而, 根据 Pawlak 的集合下近似来求属性依赖度, 即用定义 2 和定义 4 来求解, 其条件比较严格, 计算结果有时不能真正表达变量间的依赖程度^[10]。因而, 采用概率下近似表示正域, 利用定义 3 和 4 来求解属性依赖度。同理, 属性的相对必要性也采用概率下近似, 即由定义 3 和定义 5 来计算。

2.2 整合方法

具体整合方法分以下三步来完成:

1) 合并。将多个子 QPN 合并作为一个初始整合后的 QPN,(简记为 Union QPN, UQPN), 即将 QPN 中的所有节点合并为整合后的节点集合, 将所有边合并为整合后的边集合, 定性符号合并为整合后的符号集合。

2) 添边。在 UQPN 中添加有向边, 即根据定义 3 和定义 4 求解 UQPN 中无边相连的两个变量的依赖度。在保证 UQPN 中不出现环路的前提下, 添加依赖度大于零的有向边及其定性符号 δ , 优先选择依赖度较大的那条边添加。若已知数据中, 两对应属性取值相同的个数大于取值不同的个数, 则相应边上添加正影响(“+”), 小于添加负影响(“-”), 等于则添加歧义影响(“?”)。

3) 删边。删除第二步后 UQPN 中的多余边, 即对于当前网络中的每一对有边相连的两个变量, 且该组变量中的子节点还有其他的父节点, 那么根据定义 3 和定义 5 判断是否在给定的子节点情况下, 其父节点相对于该子节点是否必要? 若不必要, 删除该父子边, 必要则保留。

其整合伪代码如算法 1 所示。

算法1 基于粗糙集的 QPN 整合算法。

输入: 样本数据, n 个 QPN, β 。

输出: 整合后的 QPN。

```

Begin
    UQPN = QPN1 ∪ QPN2 ∪ ... ∪ QPNn;
    For UQPN 中没边相连的两个节点  $V_i, V_j$  do
        计算  $\gamma_{V_i}(V_j), \gamma_{V_j}(V_i)$ , 以及定性符号  $\delta_{ij}$ ;
        保证 UQPN 中不出现环路的情况下, 添加具有较大  $\gamma$  ( $\gamma \neq 0$ ) 的那条边, 否则添加较小  $\gamma$  ( $\gamma \neq 0$ ) 的边, 都出现环路的话, 两节点间不添加边;
        设  $P$  为 UQPN 中  $V_j$  的所有父节点集,  $R_i \subseteq P$ ;
        For UQPN 中每个  $V_j$ , 且  $P$  中父节点数  $\geq 2$  do
            For  $P$  中的每一个  $R_i$  do
                If  $POS_{IND(P)}(IND(V_j)) = POS_{IND(P-[R_i])}(IND(V_j))$  then
                    删除边  $\langle R_i, V_j \rangle$  及相应的定性影响符号;
            Return 整合后 QPN
End

```

3 实验及结果分析

实验在 Lawn sprinker 数据集^[11]上, 通过整合后的结果与原始结果进行实验对比, 来验证该方法的可行性和有效性。

实验选取如图 2 所示的 Lawn sprinker 网, 其中图 2(a) 表示其 BN, 图 2(b) 表示其相应 QPN, 使用 BN 中的 sample_bnet() 取样方法^[11]生成 20 组数据作为已知的完全小样本数据, 见表 1 所示, 其中 V 表示变量, $O_i (1 \leq i \leq 20)$ 表示变量的取值序号。

现将该 BN 分为两个子 BN, 根据联合树推理算法^[11] 分别得两子 BN 的条件概率表, 进而抽象出对应的 QPN 作为给

定的两子 QPN, 如图 3(a)、(b) 所示。这时由已知数据和两子 QPN, 根据算法 1 对其进行整合。第一步合并两子 QPN, 得到合并后初始 QPN 如图 3(c) 所示; 第二步, 取 $\beta = 0.7$, 求得 $\gamma_S(W) = 0.35, \gamma_W(S) = 0.3, \gamma_C(W) = 0.55, \gamma_W(C) = 0, \gamma_S(R) = 0.35, \gamma_R(S) = 0.55$ 。据此向 UQPN 中添加有向边和定性符号, 结果见图 3(d) 所示; 第三步, 对于有两个以上父节点的 S 和 W , 计算有:

$$\begin{aligned} POS_{\{C,R\}}(S) &= POS_{\{C\}}(S) = \{C = 1\} \\ POS_{\{R\}}(S) &= \{R = 1\} \\ POS_{\{C,S,R\}}(W) &= POS_{\{S,R\}}(W) = U = O_i; 1 \leq i \leq 20 \\ POS_{\{C,S\}}(W) &= \{C = 0\} \\ POS_{\{C,R\}}(W) &= \{R = 1\} \cup \{O_{10}\} \\ POS_{\{C\}}(W) &= \{C = 1\}, POS_{\{S\}}(W) = \{S = 1\} \\ POS_{\{R\}}(W) &= \{R = 1\} \end{aligned}$$

因而, 删除不必要的冗余父子边 (R, S) 和 (C, W) , 得出整合后结果, 如图 3(e) 所示。

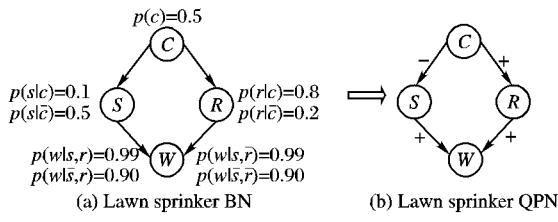


图 2 实验 Lawn sprinkler 网

表 1 已知的 20 组完全小样本数据

i	V			
	C	S	R	W
1~5	0 0 1 1 1	1 1 0 0 0	0 0 1 1 1	1 1 1 1 1
6~10	0 0 1 0 1	1 0 0 0 0	1 0 1 0 0	1 0 1 0 0
11~15	1 1 0 0 0	0 0 0 1 1	1 1 0 0 0	0 1 0 1 1
16~20	1 0 0 0 1	0 1 1 0 0	1 0 1 1 1	0 1 1 1 1

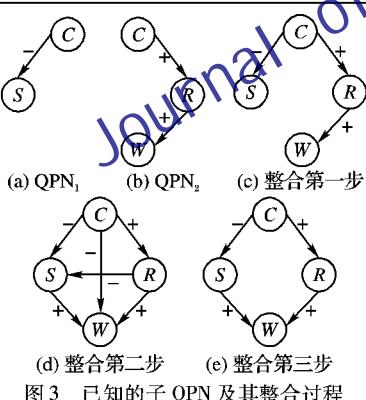


图 3 已知的子 QPN 及其整合过程

(上接第 1637 页)

- [2] PAWLAK Z, SKOWRON A. Rudiments of rough sets [J]. Information Sciences, 2007, 177(1): 3–27.
- [3] 张文修, 吴伟志, 梁吉业, 等. 粗糙集理论与方法 [M]. 北京: 科学出版社, 2001.
- [4] 孙士保, 刘瑞新, 秦克云. 变精度覆盖粗糙集模型的比较 [J]. 计算机工程, 2008, 34(7): 10–13.
- [5] 同德勤. 对 Bayesian 粗糙集模型的讨论 [J]. 计算机科学, 2006, 33(11): 162–163.
- [6] 庚慧英, 刘尚. 双论域上的变精度粗糙集模型 [J]. 科学技术与工程, 2007, 7(1): 4–7.
- [7] 赵越岭, 王建辉, 顾树生. 基于不对称边界的变精度粗糙集的参数选择 [J]. 系统仿真学报, 2006, 18(12): 3472–3474.
- [8] 王丽娟, 吴陈. 一种具有对称性的变精度粗糙集模型 [J]. 江南大学

由图 2(b) 和图 3(e) 可以得出, 整合后的 QPN 与由没有分解的原 BN 抽象出的 QPN 完全相同, 由此说明该算法的可行性和有效性。并且该方法可以根据少量小样本数据直接整合出单个 QPN, 并不需要大量充分数据。

4 结语

本文在前期研究工作的基础上, 针对普通静态环境下的多个 QPN, 提出了一种基于粗糙集的具有不同节点的多 QPN 整合方法。该方法只需在已知完全小样本数据情况下, 可整合较大 QPN 进行整个领域知识表示, 并通过实验表明了其可行性和有效性。进一步研究工作包括在已知不完全小样本数据下的多个 QPN 整合。

参考文献:

- [1] MATZKEVICH I, ABRAMSON B. The topological fusion of Bayes nets [C] // Proceedings of the 8th Conference on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann, 1992: 191–198.
- [2] SAGRADO J, MORAL S. Qualitative combination of Bayesian networks [J]. International Journal of Intelligent Systems, 2003, 18(2): 237–249.
- [3] 李维华, 刘惟一, 张忠玉. 基于扩展关系模型的多 Bayesian 网依赖结构的合并 [J]. 计算机科学, 2004, 31(7): 192–195.
- [4] 岳昆, 高明海, 韩格. 时序环境中的概率因果关系的定性表示及融合 [J]. 云南大学学报: 自然科学版, 2009, 31(5): 455–462.
- [5] LÜ Y L, LIAO S Z. Integration of multiple qualitative probabilistic networks based on probabilistic rough sets [C] // Proceedings of the 2010 International Conference on Computer Application and System Modeling. Chengdu: IEEE, 2010: 454–458.
- [6] PAWLAK Z. Rough sets [J]. International Journal of Computer and Information Sciences, 1982, 11(5): 341–356.
- [7] WELLMAN M P. Fundamental concepts of qualitative probabilistic networks [J]. Artificial Intelligence, 1990, 44(3): 257–303.
- [8] KWISTHOUTA J, TEL G. Complexity results for enhanced qualitative probabilistic networks [J]. International Journal of Approximate Reasoning, 2008, 48(3): 879–888.
- [9] 苗夺谦, 李道国. 粗糙集理论、算法与应用 [M]. 北京: 清华大学出版社, 2008: 29–70.
- [10] YUE K, YAO Y, LI J. Qualitative probabilistic networks with reduced ambiguities [J]. Applied Intelligence, 2010, 33(2): 159–178.
- [11] MURPHY K P. The Bayes net toolbox for Matlab [J]. Computing Science and Statistics, 2001, 33(2): 1024–1034.

- 学报: 自然科学版, 2007, 6(6): 825–829.
- [9] 蔡娜, 张雪峰. 基于变精度粗糙集理论的属性约简 [J]. 控制工程, 2007, 14(5): 0505–0507
- [10] 杜昌平, 周德云. 一种基于变精度粗糙集的规则提取方法研究 [J]. 西北工业大学学报, 2006, 24(3): 380–383.
- [11] HECTOR G-M, ULLMAN J D. Database system implementation [M]. 北京: 机械工业出版社, 2001.
- [12] 申初连. 多重集合的运算研究与应用 [J]. 中南林业科技大学学报, 2007, 27(3): 0145–0147
- [13] 黄光球, 赵煜. 多重粗糙集模型 [J]. 计算机工程与应用, 2010, 46(14): 53–57.
- [14] 洪帆. 离散数学基础 [M]. 2 版. 武汉: 华中科技大学出版社, 1995.