文章编号: 1001-0920(2002)06-881-05

基于属性选择的因果网络多传感器融合系统

韩 斌[1,2], 吴铁军[1,2], 杨明晖[3]
(1. 浙江大学 智能系统与决策研究所, 浙江 杭州 310027; 2. 浙江大学 工业控制技术国家重点实验室, 浙江 杭州 310027; 3. 云南送变电公司, 云南 昆明 650051)

摘 要: 针对粗集"简化"在实际应用中存在的问题提出了"统计简化"的定义和相应属性搜索算法。利用此算法对一个水域污染监测信息表进行属性简化, 结果显示与常规算法相比, 此算法得到的结果能够覆盖最大数量的对象, 更不易失配。利用简化结果对上述数据融合系统建立了因果网络模型, 实验表明, 在保持原模型搜索正确率的同时, 新模型压缩了搜索空间, 提高了搜索效率。此外, 为便于因果网络的建立导出了因果连接强度的粗集表达式。

关键词: 简化; 属性选择; 因果网络; 多传感器融合
中图分类号: TP 18; TP 274 文献标识码: A

Feature selection based causal network algorithm

HAN Bin[1,2], *WU Tie-jun*[1,2], *YANG Ming-hui*[3]
(1. Institute of Intelligent Systems and Decision Making, Zhejiang University, Hangzhou 310027, China; 2. National Laboratory for Industrial Control Technology, Zhejiang University, Hangzhou 310027, China; 3. Yunnan Power Transmission and Transformation Corporation, Kunming 650051, China)

Abstract: A statistical definition of the reduct is propose and a RS feature selection algorithm upon the definition is developed. A water-pollution multisensor fusion system is described by the causal network model. Comparative test shows that with the selected features, the computation time of the causal network searching algorithm is greatly saved, at the same time the classification accuracy is maintained. Also it shows that the causal strength of the causal network model can be derived from the information table by utilizing rough set theory.

Key words: reduct; feature selection; causal network model; multisensor fusion

1 引 言

水域污染监测常用的方法是化学法, 这种方法成本高、效率低、无法进行实时和大面积监测。如果采用多传感器融合监测系统, 上述缺点容易得到克服。但实际应用中存在以下一些困难[1]: 1) 整个水域污染情况难以用数学模型描述, 并缺少先验知识; 2) 监测数据跨越不同领域, 而且数据不一定相容, 甚至互相矛盾; 3) 各个传感器只能表征局部的情况; 4) 引起传感器信号异常的原因除污染外, 还可能有其他原因。

针对以上问题, 本文提出利用因果网络模型来描述传感器参数变化和引发这些变化的潜在原因之间的关系。因果网络[2,3]是Bayesian网络的一种,

收稿日期: 2001-07-16; 修回日期: 2001-11-23
作者简介: 韩斌(1973—), 男, 陕西澄城人, 博士生, 从事决策融合与智能系统的研究; 吴铁军(1950—), 男, 江苏南京人, 教授, 博士生导师, 从事复杂大系统管控一体化、智能控制等研究。

Bayesian 网络已被证明是一种很好的分类器[4]。然而文献[5]指出 Bayesian 网络对属性本身是敏感的,因果网络的搜索空间随着属性个数(节点数)的增长呈指数增加,因此如何挑选数量较少而又不影响分类准确性的属性集是构建因果网络模型的关键。同时给出因果连接强度的合理估计也是一个不可回避的问题。

本文针对粗集"简化"提出了统计"简化"的定义,并给出了相应的简化搜索 RS 算法。在此定义下找到的属性集能够覆盖信息表中最大数量的对象,同时保证由此属性集导出的每条规则都能满足一定的覆盖率和分类正确率[6]。基于粗集理论,本文提出一种因果连接强度的估计方法。利用水域污染监测例子的实验证明,通过属性选择建立的因果模型极大地压缩了搜索空间,减少了搜索计算量,同时搜索正确率并没有降低,因此利用属性选择 RS 算法,因果网络搜索算法在计算效率和准确性方面取得了很好的平衡。

2 简化的统计度量

2.1 粗集的基本概念[7]

2.1.1 知识表达

给定一个感兴趣的对象论域 U,对于任何子集 $X \subseteq U$ 可称之为一个 U 中的概念或范畴,它们构成了特定论域 U 的分类。例如,族 $C = \{X_1, X_2, \cdots, X_n\}$,其中 $X_i \subseteq U, X_i \neq \varnothing, X_i \cap X_j = \varnothing$ 当 $i \neq j$, $j = 1, 2, \cdots, n$ 且 $\cup X_i = U$。一个知识库可被表达为一个相关系统 $K = (U, R)$,R 是 U 上的一个等价关系族。U/R 是 R(或 U 的分类)的所有等价类族,我们用 $[x]_R$ 表示子集 X 属于 R 的一个范畴,且 R 包含元素 $x \in U$。

若 $P \subseteq R$ 且 $P \neq \varnothing$,则 $\cap P$(P 中全部等价关系的交集)也是一个等价关系,称为 P 上的不可分辨关系,记为 $\mathrm{IND}(P)$,满足 $[x]_{\mathrm{IND}(P)} = \bigcap_{P \subseteq R} [x]_P$,这样 $U/\mathrm{IND}(P)$(等价关系 $\mathrm{IND}(P)$ 的所有等价类族)定义为与等价关系 P 的族相关的知识。为简便起见,将 $U/\mathrm{IND}(P)$ 记为 U/P。

2.1.2 近似集合

给定知识库 $K = (U, R)$ 和 U 的分类 U/R,对每个子集 $X \subseteq U$,可把以下两个集合分别称为 X 的 R 下近似和 R 上近似。

$$\underline{R}X = \cup \{x \in U : [x]_R \subseteq X\}$$
$$\overline{R}X = \cup \{x \in U : [x]_R \cap X \neq \varnothing\} \quad (1)$$

2.2 简化的统计定义

2.2.1 粗集简化

令 P 和 Q 为 U 中的等价关系,则集合 $\mathrm{POS}_P(Q) = \bigcup_{X \in U/Q} \underline{P}X$ 叫做 P 的 Q 正域,表示论域 U 中通过分类 U/P 表达的知识能够确定地划入 U/Q 类的对象的集合。当 $\mathrm{POS}_{\mathrm{IND}(P)}(\mathrm{IND}(Q)) = \mathrm{POS}_{\mathrm{IND}(P-\{r\})}(\mathrm{IND}(Q))$ 时,称 $r \in P$ 为 P 中 Q 可省略的,否则为 P 中 Q 不可省略的。当 P 中每个 r 都为 Q 不可省略时,称 P 为 Q 独立的。当 S 为 P 的 Q 独立子族,且 $\mathrm{POS}_S(Q) = \mathrm{POS}_P(Q)$,则族 $S \subseteq P$ 称为 P 的 Q 简化。

2.2.2 统计简化

实际数据往往被噪声污染,在粗集简化的定义下,信息表中即使仅仅一个对象被污染,整个不可分辨关系都会改变,因此粗集简化对噪声非常敏感[6]。本文针对这种情况提出了简化的统计定义。

知识表达系统中属性与等价相对应[7],如无特殊说明以下叙述不做区别。在知识库 $K = (U, R)$ 中,如果利用等价类族 U/A 把 U 中的对象划分到等价类族 U/d 中,$A \subset R, d \in R$,则 A 和 d 分别称为条件属性集和决策属性。$V_A = \bigcup_{a \in A} V_a$,$V_a$ 是属性 a 的值域,V_d 是决策属性 d 的值域,$q(x)$ 是对象 $x \in U$ 对应属性 $q \in R$ 的取值。那么一条规则可被表达为 $\alpha \rightarrow \beta$,其中 α 和 β 是规则的前项和后项,$\alpha = \prod_{q \in A} q(x)$,$\beta = d(x)$,$x \in [x]_A \cap [x]_d$。如果用 $\mathrm{support}(\gamma)$ 表示符合表达式 γ 描述性质的对象的个数,则:$\mathrm{LHSSupport}_A(\alpha) = \mathrm{support}(\alpha)$,表示由等价类族 U/A 划分等价类族 U/d 导出的规则 $\alpha \rightarrow \beta$($\alpha \rightarrow \beta$ 定义如上文)中,符合前项描述的 U 中所有对象的数目。$\mathrm{RHSSupport}_A(\alpha \cdot \beta) = \mathrm{support}(\alpha \cdot \beta)$ 表示符合由等价类族 U/A 划分等价类族 U/d 导出的规则 $\alpha \rightarrow \beta$ 的所有对象的数目,实际上它是规则 $\alpha \rightarrow \beta$ 普遍性的绝对度量,它越大说明符合这条规则的对象越多,反映的因果关系就越普遍,反之很可能是随机因素造成的结果。$\mathrm{accuracy}_A(\alpha \cdot \beta) = \mathrm{support}(\alpha \cdot \beta)/\mathrm{support}(\alpha)$ 是规则 $\alpha \rightarrow \beta$ 分类准确性的度量。在粗集简化中 $\mathrm{accuracy}_A(\alpha \cdot \beta) = 1$,如前述由于噪声的存在,这时 $\mathrm{RHSSupport}_A(\alpha \cdot \beta)$ 很小甚至等于 1,这样导出的规则是随机因素作用的结果,不能表征属性间的因果联系,很容易"失配"(overfitting)。下面给出简化的统计定义。

$$\mathrm{Reduct}(A) = \max_{B \subset A} \{ \sum_{\alpha \in B} \mathrm{LHSSupport}_B(\alpha) \}$$

表 1　3 种算法搜索结果比较和部分 RS 算法之间的结果

算　法	结　果	$\sum \mathrm{LHSSupport}_P$
RS 算法	$B = \mathrm{Reduct}(A) = \{O_1, O_2, O_3, COD_1, COD_2, COD_3, PH\}$ (统计简化)	107
遗传算法	A (粗集简化)	71
动态简化算法	A (粗集简化)	71

注: RS 算法部分属性集 $\sum \mathrm{LHSSupport}_P(\alpha)$ 中 $\mathrm{RHSSupport}_P(\alpha \cdot \beta) \geq n$ 且 $\mathrm{accuracy}_P(\alpha \cdot \beta) \geq \delta$, $P \in A$, $n = 8$, $\delta = 0.7$, $A = \{O_1, O_2, O_3, O_4, COD_1, COD_2, COD_3, COD_4, PH\}$。

且

$$P = \left\{ \gamma \left| \begin{array}{l} \mathrm{RHSSupport}_B(\gamma \cdot \beta) > n \\ \mathrm{accuracy}_B(\gamma \cdot \beta) > \delta \\ \gamma = \prod_{\eta \in B} q(x), \beta = d(x), x \in [x]_B \cap [x]_d \end{array} \right. \right\} \tag{2}$$

n 是 0 和 $|U|$ 之间的整数, $0.5 < \delta \leq 1$, $|U|$ 是论域的基。从式(2) 可看出, 该定义不但要求每个 U/B 类要有一定的覆盖率($\mathrm{RHSsupport}_B(\gamma \cdot \beta) > n$) 和分类准确性($\mathrm{accuracy}_B(\gamma \cdot \beta) > \delta$), 还要求满足上述覆盖率和准确性的类族覆盖最大数量的对象。文献[5] 指出 overfitting 是由属性和数据失配引起的, 由于 RS 算法得到的属性集覆盖的对象数量最多, 对训练数据失配最少, 而训练数据和实验数据具有同分布, 因此该属性集总体上也失配最少, 是最不易引起 overfitting 的属性集。根据式(2), 我们给出统计简化的搜索算法(RS):

1) 设定 $n, \delta, B = \varnothing, i = k, k < m$, 其中 $m = |A|$ 是条件属性集 A 的基;

2) 令 $Q = \{R \mid |R| = i, R \subset A\} \cup B$;

3) 计算 $\mathrm{RHSSupport}_P(\gamma \cdot \beta)$, $\mathrm{accuracy}_P(\gamma \cdot \beta)$, 其中 $\gamma \in \{\theta \mid \theta = \prod_{\eta \in P} q(x), \beta = d(x), x \in [x]_P \cap [x]_d, P \in Q\}$;

4) 令 $B = \max_{\beta} \sum \mathrm{LHSSupport}_P(\gamma)$, 其中 $\mathrm{RHSSupport}_P(\gamma \cdot \beta) \geq n$ 且 $\mathrm{accuracy}_P(\gamma \cdot \beta) \geq \delta$, $P \in Q$;

5) 令 $i = i + 1$, 如果 $i < m$ 转 2);

6) 如果 $B = \varnothing$ 转 1);

7) 输出 B;

8) 结束。

在一个水域污染监测数据构成的信息表上, 利用 RS 算法找到了条件属性 A 的统计简化, 在表 1 中将 RS 算法的搜索结果与遗传算法和动态简化算法找到的粗集简化结果做了比较。

从表 1 可以看出, RS 算法得到的属性 B 在满足一定的覆盖率和分类准确性的同时, 能够把最大数量的对象划分到 U/d 中, 因此属性和数据最不容易失配, 在噪声存在时由 B 导出的规则最稳定。同时注意到其它两个算法对条件属性集 A 没有做任何简化, 因此在数据有噪声时这两个算法并不适用。

3　基于因果模型的多传感器数据融合

在一个水域污染监测的应用例子中[1], 总共用 9 个传感器监测一片水域, 其中有 4 个生化耗氧量(BOD) 传感器 $O_1 \sim O_4$, 4 个化学耗氧量(COD) 传感器 $COD_1 \sim COD_4$, 1 个 PH 传感器。本例中考虑工业污染、生活污水污染、BOD 传感器故障、COD 传感器故障、无污染和不知原因这 6 种影响传感器变化的因素, 分别用 IP, DP, O, COD, NP 和 UNKNOWN 表示。为简化起见, 每种传感器的值依据它们在正常值之内还是之外分别标记为 "+ "和 "- "。此传感器融合系统的目的是当传感器的信号变化时推断引发变化的原因。由于前文分析的困难, 传感器信号变化和原因之间的关系可用因果网络模型来描述, 其中传感器的信号变化称作 "现象", 引发变化的原因称作 "异常", 这样当 "现象 "出现时, 系统的目的就是找到最有可能引发这些 "现象 "的 "异常"。文献[3] 提出的算法可以最大后验概率找到引发 "现象 "的 "异常"。但是此算法的搜索空间是 $O(2^{n-1+m})$, n 和 m 分别是因果网络的 "现象 "节点数和 "异常 "节点数, 因此搜索空间是随着 n 的增加呈指数增长。一种改进方法是变搜索全局最优为局部最优[8] (这是以牺牲搜索准确性为代价的), 另一个方法是属性选择, 去掉冗余节点。此外还需找到 "现象 "节点和 "异常 "节点因果连接强度的合理估计。

从粗集信息表的角度看, 每个传感器监测的参数构成了条件属性集 A, 引起参数变化的原因构成了决策属性 d, 因此冗余属性可通过求属性 A 的简化来去除。另一方面, 因果连接强度

$$c_{ij} = P(m_j : d_i \mid d_i) = \frac{P(m_j : d_i)}{P(d_i)} \quad (3)$$

反映了 d_i 和 m_j 之间的因果联系[3], 相应地, 利用

$$c_{ij} = P(m_j : d_i \mid d_i) = \frac{P(m_j : d_i)}{P(d_i)} =$$

$$\mathrm{RHSAccuracy}(\alpha_j, d_i) = \frac{\mathrm{support}(\alpha_j \cdot d_i)}{\mathrm{support}(d_i)} \quad (4)$$

c_{ij} 同样可从粗集信息表中导出。其中 $\alpha_j = m_j(x), d_i = d(x), x \in [x]_{m_j} \cap [x]_d, m_j \in \mathrm{Reduct}(A)$。

使用前述水域污染监测数据库中的数据和式(4) 计算各因果连接强度如表 2 所示。

先验概率可根据各 "异常 "在数据库中出现的频率得到[9]。

基于简化属性集 B 的水域污染监测系统因果模型(记为模型 1) 见图 1, "异常 "和 "现象 "之间的连线表示因果连接强度, 其数值见表 2, 在图 1 中我们把小于 0.005 的因果连接强度置为 0。基于属性集 A 的因果模型(记为模型 2, 图略) 比基于属性集 B 的因果模型多属性 O_4 和 COD_4。因而模型 1 的搜索空间约为模型 2 的 1/16。利用同样的输入数据可在两个模型上分别执行搜索算法。为增加感性认识, 在

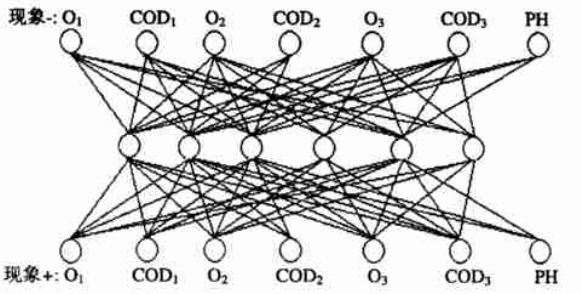

图 1 基于属性集 A 简化的因果网络模型

表 3 中列出部分随机抽取的结果。搜索算法的主要耗费是计算相对似然函数[2]。

我们对所有可能出现的 "现象 "组合做了统计, 平均相对似然函数的计算次数节省了 25%, 同时 95% 在第 1 个模型找到的最可能 "异常 "集也是在第 2 个模型上找到的最可能 "异常 "集, 其余结果, 要么是在第 2 个模型上找到的第 2 可能的 "异常 "集, 要么是第 3 可能的 "异常 "集, 因此属性选择后搜索准确率并无下降。从试验结果可以看出, 一方面通过属性选择去掉了冗余属性, 另一方面基于因果模型的多传感器融合系统在保证搜索准确性的同时极大地提高了搜索效率。

表 2 因果连接强度

j		i					
		NC	IP	DP	O	COD	UNKNOWN
O_1	+	0.016	0.176	0.062	0.000 3	0.034	0.025
	−	0.286	0.001	0.037	0.065	0.004	0.0
O_2	+	0.062	0.012	0.074	0.005	0.020	0.0
	−	0.172	0.117	0.029	0.040	0.010	0.025
O_3	+	0.023	0.167	0.140	0.048	0.042	0.025
	−	0.263	0.002	0.005	0.003	0.002	0.0
O_4	+	0.022	0.160	0.138	0.050	0.004	0.025
	−	0.172	0.001	0.005	0.001	0.034	0.0
COD_1	+	0.058	0.103	0.140	0.001	0.000 4	0.0
	−	0.178	0.018	0.005	0.056	0.05	0.025
COD_2	+	0.025	0.185	0.006	0.048	0.05	0.025
	−	0.255	0.000 5	0.131	0.003	0.000 4	0.0
COD_3	+	0.191	0.158	0.123	0.040	0.007	0.025
	−	0.051	0.003	0.008	0.005	0.027	0.0
COD_4	+	0.191	0.158	0.120	0.038	0.005	0.025
	−	0.049	0.003	0.005	0.004	0.002 0	0.0
PH	+	0.002	0.205	0.003	0.003	0.027	0.004
	−	0.382	0.000	0.148	0.048	0.007	0.009

表 3　基于不同模型的部分搜索结果

出现的"现象"集	按后验概率大小排序的搜索结果和相应的计算次数(模型 1)				按后验概率大小排序的搜索结果和相应的计算次数(模型 2)			
	1	2	3	计算次数	1	2	3	计算次数
{O↑, O↓, O→, COD↓}	IP	NP	DP	57	IP	DP	NP, IP	54
{O→, O↓, COD↑, PH⁻}	DP	IP	O	43	IP	DP	DP, O	28
{O→, O↓, O↓, COD↓}	IP	DP	O	38	IP	DP	IP, O	23
{O→, COD→, COD↓, PH⁻}	IP	O	DP	31	IP	O	DP	18
{O↓, O→, COD↑, COD↓}	IP	DP	IP, O	37	IP	DP	IP, O	25
{O↓, COD↑, COD↓, COD→}	IP	NP	DP	56	IP	DP	NP, IP	71
{O↓, COD→, COD↓, PH⁺}	NP	O	IP	36	NP	NP, UNKNOWN	NP, O	44
{O→, O→, COD↓, COD→, PH⁻}	IP	IP, O	IP, DP	36	IP	IP, DP	DP	12

4　结　　语

本文提出了粗集简化的统计定义,据此提出了简化搜索的 RS 算法。实验和分析表明,相对其它算法,当数据存在噪声时,此算法得到的属性集更稳定。针对水域污染多传感器融合系统,本文建立了因果网络模型,并分别对基于 RS 算法进行属性选择和不进行属性选择的模型进行了试验。结果表明,基于选择属性集构成的模型极大地压缩了搜索空间,节省了计算量,而且搜索准确性与原模型相比并无多少差别。同时利用粗集推理还得到了因果连接强度的合理估计。这一切表明,粗集结合因果网络理论在以数据分析和推理为基础的数据融合系统中具有广阔的应用前景。

致　谢

感谢挪威科学技术大学计算机与信息系的 Ohrn 博士,部分粗集推理工作是在他开发的 Rosetta 系统上完成的。

参考文献(Reference):

[1] Bin Han, Tie-Jun Wu. Data mining in multisensor system based on rough set theory[A]. *Proc of ACC* 2001 [C]. Arlington, 2001. 4427–4431.

[2] Yun Peng, James A Reggia. A probabilistic causal model for diagnostic problem solving — Part I: Integrating symbolic causal inference with numeric probabilistic inference[J]. *IEEE Trans on Systems, Man and Cybernetics*, 1987, 17(2): 146–162.

[3] Yun Peng, James A Reggia. A probabilistic causal model for diagnostic problem solving — Part Ⅱ: Diagnostic strategy[J]. *IEEE Trans on Systems, Man and Cybernetics*, 1987, 17(2): 395–406.

[4] Pedro Domingos, Michael Pazzani. Beyond independence: Conditions for the optimality of the simple bayesian classifier[A]. *Proc of the Thirteenth Int Conf on Machine Learning* [C]. Bari: Morgan Kaufmann, 1996. 105–112.

[5] Pedro Domingos. Bayesian averaging of classifiers and the overfitting problem[A]. *Proc of the Seventeenth Int Conf on Machine Learning* [C]. Stanford: Morgan Kaufmann, 2000. 223–230.

[6] Aleksander Ohrn. Discernibility and rough sets in medicine: Tools and applications[D]. Norwegian University of Science and Technology, 1999. 53, 63–65.

[7] Z Pawlak. *Rough Sets — Theoretical Aspects of Reasoning About Data*[M]. Boston: Kluwer Academic Publishers, 1992. 1–53.

[8] Yun Peng, James A Reggia. A connectionist model for diagnostic problem solving[J]. *IEEE Trans on Systems, Man and Cybernetics*, 1989, 19(2): 285–298.

[9] Inien Syu, S D Lang. Adapting a diagnostic problem-solving model to information retrieval[J]. *Information Processing and Management*, 2000, 36(2): 313–330.