

文章编号: 2095—0411 (2011) 04—0056—03

基于粗糙集约简技术的排水管路运行风险评估^{*}

张 岩¹, 石 林²

(1. 常州市科教城 培训管理中心, 江苏 常州 213164; 2. 常州大学 信息科学与工程学院, 江苏 常州 213164)

摘要: 排水管路运行风险的实时评估对于提高管路风险预测十分有必要。介绍了 Semi—Naive—Scaler 属性离散化算法, 给出了基于可辨识矩阵的属性约简算法, 并采用了基于可辨识矩阵的二值化数据过滤和贪心算法相结合的规则约简算法。通过属性约简和规则约简, 得出决策规则。最后在排水管路运行风险评估方面给出了应用算例。

关键词: 粗糙集; 排水管路; 风险评估

中图分类号: TM 75; TM 712

文献标识码: A

Risk Assessment for Drainpipe Lines Running Based on Reducing Technology of Rough Sets

ZHANG Yan¹, SHI Lin²

(1. Training Management Center, Changzhou Science and Education Town, Changzhou 213164, China; 2. School of Information Science and Engineering, Changzhou University, Changzhou 213164, China)

Abstract: It is necessary to run a real time risk assessment for improving the prediction of drainpipe lines risk level. this paper introduces the attribute discretization algorithm of the Semi—Naive—Scaler, gives attributes reducing algorithm based on discernibility matrix, and also adopts the rules reducing algorithm combined with binarization data filtration based on discernibility matrix and greedy algorithm. Through attribute reduction and rule reduction, it obtains decision rules. Finally, an application example in risk assessment of drainpipe lines running is given.

Key words: rough sets; draining line; risk assessment

粗糙集理论是一种新的处理模糊和不确定性问题的数学工具^[1,2]。其主要思想就是在保持分类能力不变的前提下, 通过知识约简, 导出问题的决策或分类规则。它最显著的特点是除了数据集合本身外不需要任何预备的或额外的有关数据信息。

本文应用粗糙集约简方法来分析计算排水管路存在的风险性水平的各指标, 试图发现排水管路各项状态信息与管路运行状态之间的一些内在规律, 提供一些潜在的、有价值的决策信息以供参考。

1 粗糙集中知识约简

知识约简是粗糙集理论的重要内容之一, 一般包括属性约简和规则约简两个步骤。

1.1 连续属性离散化

连续属性离散化是对属性的近似处理, 本文采用一种可行性较好的 Semi—Naive—Scaler 算法^[3,4]。设 X_i 为属性等价类:

^{*} 收稿日期: 2011—05—25

作者简介: 张岩 (1977—), 男, 江苏常州人, 工程师。

$$X_i = \{x \in U \mid a(x) = c_i\}, c_0 < c_1 < \dots < c_i < \dots < c_n \tag{1}$$

断点定义为:

$$D_i = \{v \in V_d \mid v = \operatorname{argmax} \mid \{x \in X_i \mid d(x) = d_i\}\} \tag{2}$$

$$B = \left\{ \frac{c_i + c_{i+1}}{2} \mid D_i \not\subset D_{i+1}, D_{i+1} \not\subset D_i \right\} \tag{3}$$

式中: c_i 和 d_i 分别是决策表中的各实例的条件属性和决策属性的取值, B 为属性断点集合, 集合 D_i 是 X_i 中主要的决策属性值, 即等价类 X_i 中出现频率最高的决策属性值。

Semi-Naive-Scaler 算法思想为: ①根据式 (1) 和式 (2) 计算 X_i 和 D_i 的值; ②如果 $D_i \not\subset D_{i+1}$ 且 $D_{i+1} \not\subset D_i$, 则根据式 (3) 计算出此断点值; 否则, 不选取此断点。

1.2 属性约简

在离散化以后, 即可以进行属性约简。通常的属性约简算法可定义为: ①删除多余的条件属性; ②消去重复的行; ③消去每一块决策规则的多余属性。

本文采用基于可辨识矩阵的属性约简算法^[5,6]。令 Reduct 是决策表 T 经过属性约简之后得到的条件属性集合, 可对本属性约简算法描述如下。

①将决策表转换为可辨识矩阵的形式, 并将矩阵中属性组合数为 1 的属性列入最终的属性约简集合, 即 $\operatorname{Reduct} = C_0$ 。

②从可辨识矩阵中找出不包含核属性组合 S , 即 $Q = \{B_i \mid B_i \cap \operatorname{Reduct} \neq \emptyset, i = 1, 2, \dots, S\}$
 $S = S - Q$ 。

③将所有不包含核属性的条件属性组合表示为合取范式的形式, 即

$$p = \bigwedge \{ \bigvee b_{i,k} \mid (i = 1, 2, \dots, m) \}。$$

④将 p 转换为析取范式的形式, 并进行化简。

⑤所有约简得到的合取式与核属性组成属性约简组合。

利用以上属性约简算法, 可以得到在保证原决策表不损失潜在知识的前提下进行最佳属性约简的充分条件。

1.3 规则约简

规则约简的目标为去除记录非必须属性值, 以

精简记录集。其过程表述为: 假设信息系统 $I = \langle U, \Omega, V, f \rangle$, 对应的二值化信息系统为 $I = \langle U, \Omega^B, V, f^B \rangle$, 式中, Ω^B 为二值化后的属性集合; f^B 为信息函数, 其值域为 $\{0, 1\}$ 。

设属性 $a \in \Omega$ 有 n 个属性值, 那么用一个 n 维的二值向量来表示这个属性值。对于这个向量的第 i 位定义如下:

$$V_{ai}(x) = \begin{cases} 1, & V_a(x) = i \\ 0, & \text{其他} \end{cases} \tag{4}$$

式中: $V_{ai}(x)$ 为实例 x 的二值化属性值的第 i 位; $V_a(x)$ 为非二值化属性值。

以表 1 为例来描述二值化过滤方法的步骤。表 1 中 Q 和 M 为条件属性; X 为决策属性。经过二值化后的决策表见表 2, 结合贪心算法来进一步化简决策表 2, 贪心算法的具体步骤如下: ①根据原来的信息表 2 构造新的信息表。②初始化最佳属性值集。③选取信息表所有列中 1 的个数最多的条件属性值加入到最佳属性值集中, 去掉此条件属性值所在的列和在此条件属性为 1 的所有行; 当有一个以上的条件属性值的列 1 的个数相同时, 把列对应的条件属性值所在的列值为 1 的行的 1 的数目相加。④如果新的信息表中的元素不为空, 则转至 ③, 否则停止。

按照上述方法化简决策表 2, 得到属性值化简后的决策表, 见表 3。表中的符号 * 表示可约简的属性值。

表 1 初始决策表

Table 1 Initial decision table

	Q	M	X
2	1	2	2
3, 7	3	2	2
4, 5	1	1	1
5	2	1	1
6, 8	2	2	2

表 2 二值化处理后的决策表

Table2 Binarization Processed decision table

	Q			M		X
	1	2	3	1	2	
	1	2	3	1	2	
2	1	0	0	0	1	2
3, 7	0	0	1	0	1	3
4, 5	1	0	0	1	0	1
5	0	1	0	1	0	1
6, 8	0	1	0	0	1	2

表 3 化简后的决策表
Table 3 After Jane's decision table

	Q	M	X
2, 6, 8	*	2	2
3, 7	3	2	3
4, 5, 5	*	1	1

事实上，经过约简之后的信息系统就可以作为所求的规则知识。由表 4 可以得到下面 3 条规则：①If $Q=3, M=2$ then $X=3$ ；②If $M=2$ then $X=2$ ；③If $M=1$ then $X=1$ 。

2 应用算例

排水管路风险评估主要是对各级管路运行状态变化信息以及水压、水流量等的特征进行分析，并据此预测运行状态下管路存在的风险指标。本实例采用上述属性约简和规则约简的方法，来说明如何得到风险策略规则。

表 4 是一个管路风险评估决策表，D 为风险类型，共存在 3 种类型风险 D_1 （水压过高）、 D_2 （泵温过高）、 D_3 （水量过大）；选取参数 C_1 、 C_2 、 C_3 （ C_1 、 C_2 、 C_3 为排水管路风险因素离散化后的具体含义）；P 为条件属性，U 为影响风险的类别。

表 4 排水管路风险评估决策表

Table 4 Drainage pipeline risk assessment decision table

U	C_1	C_2	C_3	P	D
1	1	0	1	0	D_1
2	1	1	1	0	D_1
3	1	2	0	0	D_2
4	2	2	2	0	D_2
5	2	0	2	0	D_3
6	2	1	2	2	D_3

根据前面提出的基于可辨识矩阵的属性约简算法，对上面的决策表进行约简，仅含有一个核属性 C_2 。除去所有包含 C_2 的属性组合，得到一组不包含核属性的属性组合 C_1 、 C_3 ，其合取范式为 $C_1 \wedge C_3$ 。将合取范式与核属性组成属性组合，得到一个属性约简结果见表 5。

表 5 属性约简结果

Table 5 Attribute reduction results

	C_1	C_2	C_3	D
1	1	0	1	D_1
2	1	1	1	D_1
3	1	2	0	D_2
4	2	2	2	D_2
5	2	0	2	D_3
6	2	1	2	D_3

利用可辨识矩阵的二值化数据过滤约简算法对上面得到的属性约简进行规则约简，得到转换表结果如表 6。

将转换表用可辨识矩阵方法并结合贪心算法，

进行约简后得到最终约简表，见表 7。

表 6 二值化数据过滤转换表

Table 6 The value of the two data filtering conversion table

U	C_1			C_2			C_3			D
	0	1	2	0	1	2	0	1	2	
1	0	1	0	1	0	0	0	1	0	D_1
2	0	1	0	0	1	0	0	1	0	D_1
3	0	1	0	0	0	1	1	0	0	D_2
4	0	0	1	0	0	1	0	0	1	D_2
5	0	0	1	1	0	0	0	0	1	D_3
6	0	0	1	0	1	0	0	0	1	D_3

表 7 最终约简表

Table 7 The final reduction table

U	C_1	C_2	C_3	D
1, 2	1	*	1	D_1
3	1	2	0	D_2
3	2	2	2	D_2
5, 6	2	*	2	D_3

经过约简的决策表去掉了很多冗余信息，但仍能反映各潜在因素对风险的影响关系，下面列出本算例中排水管路风险策略规则：①（ $C_1=1 \wedge C_3=1$ ）→风险类型= D_1 （水压过高）；②（ $C_1=1 \wedge C_2=2 \wedge C_3=0$ ）→风险类型= D_2 （泵温过高）；③（ $C_1=2 \wedge C_2=2 \wedge C_3=2$ ）→风险类型= D_2 （泵温过高）；④（ $C_1=2 \wedge C_3=2$ ）→风险类型= D_3 （水量过大）。

3 结 论

本文结合粗糙集理论中知识约简方法，通过属性约简和规则约简，对排水管路运行状态下存在的风险特征进行规则提取。结果表明，这种方法能够有效地获取决策规则，能有效降低大量实时数据处理的难度，对于提高风险评估的效率十分有益。该方法同样适应在其他风险分析领域。

参考文献：

[1] Pawalk Z, Rough Sets. Theoretical Aspects of Reasoning about Data [M]. Boston: Kluwer Academic Publishers, 1991: 60—67.

[2] Pawlak Z, Busse J G, Slowinski R, et al. Rough sets [J]. Communications of the ACM, 1995, 38 (11): 89—95.

[3] 姜涛, 韩富春, 范卫星. 基于粗糙集理论的架空排水管路运行状态评估 [J]. 电气技术, 2007 (4): 58—60.

[4] 郭小荟, 马小平. 基于粗糙集—神经网络集成的故障诊断 [J]. 控制工程, 2007, 14 (1): 53—56.

[5] 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.

[6] 孙秋野, 黎明. 粗糙集理论及其电力行业应用 [M]. 北京: 机械工业出版社, 2009.