文章编号 :1006-4303(2002)01-0005-04

基于 Rough 集联系度的决策表简化方法

张　平,黄德才
(浙江工业大学 计算机软件开发环境重点实验室,浙江 杭州 310032)

摘要 提出了集合型 Rough 集(粗集)联系度的概念以及利用 Rough 集联系度对决策表进行条件属性简化和属性冗余值简化的计算步骤,通过算例说明该方法比传统的 Rough 集理论中使用范畴的相对简化方法更简单。

关键词 :Rough 集联系度;知识发现;决策表简化
中图分类号 :TP182　　**文献标识码** :A

Reducing method for knowledge representation system based on rough set connection degree

ZHANG Ping, HUANG De-cai
(Key Lab of Software Development Environment, Zhejiang University of Technology, Hangzhou 310032, China)

Abstract : Based on the conception of rough set connection degree, this paper gives a new method to reduce the conditional attributions and the redundancy attributive value. An example is given which illustrates that the new method is simpler than the traditional one.

Key words : rough set connection degree; knowledge discovery; reduction of decision table

0 引　言

由于 Rough 集具有演绎、归纳和常识推理这三种推理能力,因而在知识发现(KDD,Knowledge Discovery in Database)中被越来越多地引用[1]。知识库或决策表的简化在 KDD 工程应用中相当重要,用 Rough 集方法进行决策表简化,就是简化决策表中的条件属性,化简后的决策表具有化简前的功能,这样就可以基于少量的条件属性获得知识或决策。一般来说,决策表的简化按如下步骤进行[2]:

(1) 消去重复的行;

(2) 进行条件属性的简化,即从决策表中消去某些列;

(3) 消去属性的冗余值。

显然,重复的行表示相同的决策,消去它是一件简单的事情。在 Rough 集方法中条件属性的简化使用‘知识相对简化’的方法,而消去属性冗余值使用‘范畴相对简化’方法,即在传统 Rough 集方法中,条件属性的简化和属性冗余值的消除分别采用两种不同的方法。本文基于集对分析(SPA)联系度[3]定义的‘Rough 集联系度’[4]的思想,引入了‘集合型 Rough 集联系度’的概念,利用集合型 Rough 集联系度可

收稿日期 :2000 - 11 - 21 修订日期 :2001 - 04 - 15
基金项目 :浙江省自然科学基金资助项目(601076)
作者简介 :张平(1971 -),女,浙江嵊州人,浙江工业大学信息学院硕士研究生。

同时用于消去条件属性和属性的冗余值,使知识推理过程变得统一且相对简单,这是 Rough 集方法的一种新扩展。

1 条件属性简化的集合型 Rough 集联系度方法

下面,我们先给出集合型 Rough 集联系度概念。

定义 1 设论域为 U,P 和 Q 为 U 上的等价关系,Q 的 P 下近似集为 $P*(Q)$,Q 的 P 上近似集为 $P^*(Q)$,Q 的 P 负域为 $\mathrm{NEG}_p(Q)$[2],则集合型 Rough 集联系度定义为:

$\mu_p(Q) = P*(Q) + [P^*(Q) - P_*(Q)]\,i + \mathrm{NEG}_P(Q)\,j$ (这里的 i,j 仅作标记使用,下同)

由定义 1 可知,如果 Rough 集 Q 为外不可定义,则集合型 Rough 集联系度变成:

$$\mu_p(Q) = P^*(Q) + [P^*(Q) - P_*(Q)]\,i$$

在定义了集合型 Rough 集联系度 $\mu_p(Q)$ 之后,对决策表中条件属性是否可以简化,可以通过集合计算来实现。

(1) 设 P 和 Q 为论域 U 上的等价关系,$r \subset P$,当存在 $\mu_p(Q) = \mu_{(P-r)}(Q)$ 时,称 r 为 P 中 Q 可省略的,否则称 r 为 P 中 Q 不可省略的。

(2) 当 P 中每一个 r 都为 Q 不可省略,则称 P 为 Q 独立的。

(3) 设 $S \subset P$,当 S 为 P 的 Q 独立子集族且 $\mu_s(Q) = \mu_p(Q)$ 时,则 S 称为 P 的 Q 简化。

(4) $\mathrm{Core}_p(Q) = \bigcap \mathrm{red}_p(Q)$

$\mathrm{red}_p(Q)$ 为 P 中所有 Q 简化族,$\mathrm{Core}_p(Q)$ 为 P 的 Q 核。

Rough 集联系度可以刻画知识 P 能划入知识 Q 的精确度、可能度。当除去等价关系 P 中某一初等集合 r,$P - r$ 能划入知识 Q 的精确度和可能度没有变化时,表示初等集合 r 表示的属性是可以省略,否则是不可以省略的。由此可划分出等价关系 P 的多种关于 Q 的简化,但各种简化中必然有一公共子集,它就是 P 的 Q 核 $\mathrm{Core}_p(Q)$。

在原来的 Rough 集理论中,无法表示 P 划入 Q 的可能度,而 Rough 集联系度的引入,可以刻画 P 划入 Q 的可能度是多少。

2 消去属性冗余值的 Rough 集联系度方法

决策表是通过指定对象的基本特征(条件属性) 和它的特征值(结果属性值) 来描述的一个知识表达系统,其形式化定义为:

$S' = \langle U', C', D', V' \rangle$

这里 U' 为全域,$C' \cup D' = A'$ 是属性集合,子集 C' 和 D' 分别称为条件属性和结果属性,$V' = U'_{a \in A}$,V'_a 是属性值的集合。

如果一个知识表达系统已进行了重复行的简化和列的简化,则得到的一个等价知识表达系统记为

$$S = \langle U, C, D, V \rangle$$

定义 2 令 F 为等价系统中的一个决策:$C \Rightarrow D$,其中 $C = \{C_i\}$,$i = 1,2,\cdots,n$;$D = \{D_j\}$,$j = 1,2,\cdots,m$

(1) 设属性集 $A_k = \{C_{it} \mid t = 1,2,\cdots k;\ C_{it} \in C$ 且 $C_{il} \neq C_{ij}$ 若 $l \neq j\}$,则决策 F 相对于全域 U 的联系度定义为:

$$u_{\{C_i\}}(U) = \mathrm{POS}_{\{C_i\}}(U) + \mathrm{NEG}_{\{C_i\}}(U)\,j$$

(2) 设属性集 $A_k = \{C_{il} \mid l = 1,2,\cdots,k;\ C_{il} \in C$ 且 $C_{il} \neq C_{ij}\ i \neq j\}$,满足决策 F 相对于结果属性集的联系度定义为:

$$u_{\{C_i\}}(D) = \mathrm{POS}_{\{C_i\}}(D) + \mathrm{BN}_{\{C_i\}}(D)i + \mathrm{NEG}_{\{C_i\}}(D)j$$

利用定义 2,只要我们对不同属性集合进行搜索运算,就可获得关于决策 F 的属性冗余值的简化,其算法如下:(为了叙述方便,取 $n=3, m=1, C=\{a,b,c\}, D=\{d\}$,决策 $F: C \Rightarrow D$ 表示为 $V_a \wedge V_b \wedge V_c \Rightarrow V_d$。对属性集 $\{a,b,c\}$,决策 F 相对于论域 U 的联系度为:$\mu_{\{a,b,c\}}(U)$,相对于 $D=\{d\}$ 的联系度 $\mu_{\{a,b,c\}}(D)$,($(a,b,c($(D) 中包含的元素即为决策 F 蕴涵的元素,是唯一的。)

① 对属性集 $\{a,b\}$ $\{b,c\}$ $\{a,c\}$,求决策 F 相对于论域 U 的联系度:

$$\mu_{\{a,b\}}(U), \mu_{\{b,c\}}(U), \mu_{\{a,c\}}(U)$$

对上述属性集,求决策 F 中相对于 $D=\{d\}$ 的联系度:

$$\mu_{\{a,b\}}(D), \mu_{\{b,c\}}(D), \mu_{\{a,c\}}(D)$$

如果 $\mu_{\{a,b\}}(D), \mu_{\{b,c\}}(D), \mu_{\{a,c\}}(D)$ 正域不全为空,进行第 ② 步 如果每个正域全为空,进行第 ③ 步

② 对属性 $\{a\}$, $\{b\}$, $\{c\}$,求决策 F 相对于论域 U 的联系度:

$$\mu_{\{a\}}(U), \mu_{\{b\}}(U), \mu_{\{c\}}(U)$$

对上述属性集,求决策 F 中相对于 $D=\{d\}$ 的联系度:

$$\mu_{\{a\}}(D), \mu_{\{b\}}(D), \mu_{\{c\}}(D)$$

③ 从最小属性集 F 相对于 $D=\{d\}$ 的联系度为非空的正域中获得决策 F 的属性简化集。

④ 从 F 的属性简化集获得决策 F 的属性核集。(如果 $\mu_{\{a\}}(D), \mu_{\{b\}}(D), \mu_{\{c\}}(D)$ 正域不全为空,不空是唯一的,即为核值)

对其它所有决策重复 ① ~ ④ 步,最后获得知识表达系统的简化和核值表。

3　算例分析

设某一知识表达系统用表 1 表示,其中 $P=\{a,b,c,d\}$ 是条件属性,$Q=\{e\}$ 为决策属性。该系统中没有重复行,所以首先进行条件属性简化:

表 1　某一决策表表达的知识系统

U	a	b	c	d	e
1	1	0	0	1	1
2	1	0	0	0	1
3	0	0	0	0	0
4	1	1	0	1	0
5	1	1	0	2	2
6	2	1	0	2	2
7	2	2	2	2	2

等价关系:

$U/P = \{\{1\},\{2\},\{3\},\{4\},\{5\},\{6\},\{7\}\}$;

$U/\{P-a\} = \{\{1\},\{2,3\},\{4\},\{5,6\},\{7\}\}$;

$U/\{P-b\} = \{\{1,4\},\{2\},\{3\},\{5\},\{6\},\{7\}\}$;

$U/\{P-c\} = \{\{1\},\{2\},\{3\},\{4\},\{5\},\{6\},\{7\}\}$;

$U/\{P-d\} = \{\{1,2\},\{3\},\{4,5\},\{6\},\{7\}\}$;

$U/Q = \{\{1,2\},\{3,4\},\{5,6,7\}\}$;

Rough 集联系度:

$$\mu_P(Q) = \{1,2,3,4,5,6,7\} + \Phi_i$$

$$\mu_{\{P-a\}}(Q) = \{1,4,5,6,7\} + \{2,3\}i \neq \mu_P(Q)$$

$$\mu_{\{P-b\}}(Q) = \{2,3,5,6,7\} + \{1,4\}i \neq \mu_P(Q)$$

$$\mu_{\{P-c\}}(Q) = \{1,2,3,4,5,6,7\} + \Phi_i = \mu_P(Q)$$

$$\mu_{\{P-d\}}(Q) = \{1,2,3,6,7\} + \{4,5\}i \neq \mu_P(Q)$$

所以属性 c 是可省略的,由此可得到一个等价知识表达系统(如表 2)。

下面再消去每一个决策规则的条件属性冗余值,

对规则 1:$(V_a=1, V_b=0, V_d=1) \Rightarrow V_e=1$,其中满足 $V_e=1$ 的规则集合为 $Q_{V_e}=\{1,2\}$,那么对两个属性的集合:

$$\mu_{\{a,b\}}(U) = \{1,2\} + \{3,4,5,6,7\}j;$$
$$\mu_{\{a,d\}}(U) = \{1,4\} + \{2,3,5,6,7\}j;$$
$$\mu_{\{b,d\}}(U) = \{1\} + \{2,3,4,5,6,7\}j;$$
$$\mu_{\{a,b\}}(D) = \{1,2\} + \{3,4,5,6,7\}j;$$

所以

$$\mu_{\{a,d\}}(D) = \Phi + \{1,2,3,4,5,6,7\}i;$$
$$\mu_{\{b,d\}}(D) = \{1\} + \{2,3,4,5,6,7\}i;$$

那么对一个属性的集合：

$$\mu_{\{a\}}(U) = \{1,2,4,5\} + \{3,6,7\}j;$$
$$\mu_{\{b\}}(U) = \{1,2,3\} + \{4,5,6,7\}j;$$
$$\mu_{\{d\}}(U) = \{1,4\} + \{2,3,5,6,7\}j;$$
$$\mu_{\{a\}}(D) = \Phi + \{1,2,4,5\}i + \{3,6,7\}j;$$

则

$$\mu_{\{b\}}(D) = \Phi + \{1,2,3\}i + \{4,5,6,7\}j;$$
$$\mu_{\{d\}}(D) = \Phi + \{1,4\}i + \{2,3,5,6,7\}j。$$

从决策表可知：对规则 1 属性 $\{a,b\}$ 对应的值为 $V_a = 1, V_b = 0$；属性 $\{d,b\}$ 对应的值为 $V_b = 0, V_d = 1$。通过上述运算可知 $V_a = 1, V_b = 0$ 和 $V_b = 0, V_d = 1$ 为规则 1 两个简化。

对其它规则进行相同的运算，获得该知识表达系统的核值表(如表 3) 和两个简化表(省略)，并获得该知识表达系统的一个等价决策：

$a_1 b_0 \Rightarrow e_1$, $a_0 \vee b_1 d_1 \Rightarrow e_0$, $d_2 \Rightarrow e_2$ 和 $b_0 d_1 \vee a_1 d_0 \Rightarrow e_1$, $a_0 \vee b_1 d_1 \Rightarrow e_0$, $d_2 \Rightarrow e_2$

表 2 消去属性 c 决策表

U	a	b	d	e
1	1	0	1	1
2	1	0	0	1
3	0	0	0	0
4	1	1	1	0
5	1	1	2	2
6	2	1	2	2
7	2	2	2	2

表 3 条件属性核值表

U	a	b	d	e
1	–	0	–	1
2	1	–	–	1
3	0	–	–	0
4	–	1	1	0
5	–	–	2	2
6	–	–	2	2
7	–	–	2	2

4 结束语

利用 Rough 集联系度的概念对知识表属性列和规则的冗余属性值进行简化，无须引入传统 Rough 集理论中范畴的相对简化方法，使其推理过程得到统一，计算方法更加简单，这说明了处理不确定关系的集对分析(SPA)理论[4]的优点。另一方面，上述分析只使用了联系度的正域获得必然规则，而在联系度边界域中包含着许多可能规则，如何描述可能规则的可能度，应该是集对分析(SPA)理论和 Rough 集结合的进一步研究的方向，使粗集理论真正体现‘粗’的含义。

参考文献：

[1] 刘清，黄兆华，刘少辉，等. 带 Rough 算子的决策规则及数据挖掘中的软计算[J]. 计算机研究与发展，1999，36(7)：800 – 904.
[2] 曾黄麟. 粗集理论及其应用——关于数据推理的新方法[M]. 重庆：重庆大学出版社，1996.
[3] 赵克勤，宣爱理. 集对论——一种新的不确定性理论方法与应用[J]. 系统工程，1996，14(1)：18 – 23.
[4] 张平，黄德才. 基于联系度的 Rough 集[J]. 杭州电子工业学院学报，2001，21(1)：50 – 54.

(责任编辑：陈石平)

基于Rough集联系度的决策表简化方法

作者： 张平， 黄德才
作者单位： 浙江工业大学,计算机软件开发环境重点实验室,浙江,杭州,310032
刊名： 浙江工业大学学报 ISTIC PKU
英文刊名： JOURNAL OF ZHEJIANG UNIVERSITY OF TECHNOLOGY
年，卷(期)： 2002，30(1)
被引用次数： 4次

参考文献(4条)

1. 刘清. 黄兆华. 刘少辉 带Rough算子的决策规则及数据挖掘中的软计算[期刊论文]-计算机研究与发展 1999(07)

2. 曾黄麟 粗集理论及其应用—关于数据推理的新方法 1996

3. 赵克勤. 宣爱理 集对论--一种新的不确定性理论方法与应用 1996(01)

4. 张平. 黄德才 基于联系度的Rough集[期刊论文]-杭州电子工业学院学报 2001(01)

引证文献(4条)

1. 蒋云良. 徐从富 集对分析理论及其应用研究进展[期刊论文]-计算机科学 2006(1)

2. 卓小军 面向中小型企业的产品数据管理系统研制[学位论文]硕士 2006

3. 刘高峰 基于权重联系度的粗集模型及其在不完备决策表中的应用[学位论文]硕士 2006

4. 李永森. 杨善林. 马溪骏. 秦科 基于一致规则的知识约简方法[期刊论文]-合肥工业大学学报（自然科学版） 2005(10)