一个新的差别矩阵及其求核方法

叶东毅,陈昭炯
(福州大学计算机科学与技术系,福建福州 350002)

摘 要: 首先利用反例指出 HU 的利用差别距阵来求粗糙集中的核的方法是错误的,然后给出一个新的差别距阵的定义和求核方法,并证明了方法的正确性.

关键词: 粗糙集; 差别距阵; 核
中图分类号: TP182 **文献标识码**: A **文章编号**: 0372-2112 (2002) 07-1086-03

A New Discernibility Matrix and the Computation of a Core

YE Dong-yi, CHEN Zhao-jiong
(Dept of Computer Science and Technology, Fuzhou University, Fuzhou, Fujian 350002, China)

Abstract: First, with a counterexample we point out in this paper an error in HU's method for calculating the core of an information system in the context of rough set based on the discernibility matrix defined therein. Then, we present a new discernibility matrix definition together with a method for the computation of the core and prove the correctness of the method.

Key words: rough set; discernibility matrix; core

1 引言

由波兰学者 Pawlak 教授提出的粗糙集理论是分析不完整、不精确信息系统的有力工具,近年来在机器学习,数据挖掘,人工神经网络等多个领域中得到了广泛的应用[1,2]. 在粗糙集理论中,属性约简(知识约简)是最重要的一个部分. 目前已提出了若干个求属性约简的算法[3~8], 在这些约简算法中,搜索一个属性约简集合通常是从核开始. 因此,求核的运算是必不可少的. 虽然核的确定可以通过求出所有的不可缺少的属性(indispensable attribute)来实现[1],但人们也在寻求更节省计算量的求核方法,其中有代表性的一个是由 HU XIAO-HUA 等学者在文献[3]中给出的利用改进差别距阵来确定核的方法. 本文将指出这个求核方法存在着错误,并用反例加以说明. 然后,在 HU 的改进差别距阵的基础上,给出一个新的差别距阵的定义和求核方法,并证明了方法的正确性.

2 属性约简与核

关于粗糙集的基本概念,如不可分辨关系、由属性集合导出的等价关系、等价类、集合的上下近似,边界区等的详细描述,读者可参阅有关的文献[1,2]. 这里,为后面叙述方便起见,引入一些记号并介绍一下属性约简、核等概念.

考虑一个信息系统[1]:

$$L = (U, Q, V_q, F_q),\ q \in Q \qquad (1)$$

其中 $U = \{x_1, \cdots, x_n\}$ 是论域, Q 是属性集合, V_q 为属性取值的集合, F_q 是 $U \times Q \to V_q$ 的映射. 属性集合 Q 通常分为条件属性集 C 与决策属性集 D. 以下,设条件属性集合 C 中有 m 个属性: $C_1, C_2 \cdots\cdots C_m$,其值域为有限离散集合. 为方便起见,对于 $P \subseteq Q$,在不产生混淆的情况下, P 既表示一个属性子集,又表示由它导出的一个等价关系,即不可分辨关系 IND (P). 不失一般性,假设仅有一个决策属性 D,其取值范围是 $1,2,\ldots,k$. 由 D 导出的等价类构成 U 的一个划分: $\{Y_1, Y_2 \cdots\cdots Y_k\}$,其中, $Y_i = \{x \in U : F_q(x, D) = i\}$, $i = 1, \cdots, k$.

定义 1[1]: 设 $X \subseteq U$ 为论域的一个子集, $P \subseteq C$, X 的关于 P 的下近似为

$$\underline{P}X = \{x \in U : [x]_P \subseteq X\}$$

其中, $[x]_P$ 表示 U 中所有与 x 在关系 IND (P) 下是等价的元素构成的集合.

定义 2[1]: 设 $P \subseteq C$,对划分 $\{Y_1, Y_2 \cdots\cdots Y_k\}$ 的 P- 近似精度为 $\gamma_P = \sum_{i=1}^{k} card(\underline{P}Y_i) / card(U)$,其中 $Card(.)$ 表示集合的基数.

定义 3[1]: 设 $P \subseteq C$,若 $\gamma_P = \gamma_C$,且不存在 $R \subset P$,使得 $\gamma_R = \gamma_P$,则称 P 为 C 的一个(相对于决策属性 D 的)属性约简. 所有 C 的属性约简的交称为 C 的核,记为 $Core(C)$.

定义 4[1]: 如果属性 $a \in C$ 满足 $\gamma_{C-\{a\}} < \gamma_C$,则称属性 a 为不可缺少的(indispensable),否则,称属性 a 为冗余的.

利用不可缺少属性,可以给出核的等价定义(在文献[1]中是作为一个性质给出的).

收稿日期:2001-04-16;修回日期:2001-10-23
基金项目:教育部科研重点项目(No. 00185);福建省自然科学基金(No. A0010009);福建省教育厅科研基金(No. JA00144)

定义 5[1]:属性 $a \in Core(C)$ 当且仅当 a 是不可缺少的属性.

另外,对属性子集合 $P \subseteq C$ 及划分 $\{Y_1, Y_2, \cdots, Y_k\}$,记正区域 $Pos_P(D) = \bigcup_{i=1}^{k} \underline{P}Y_i$. 易知,$\gamma_P = card(Pos_P(D))/card(U)$.

3　HU 算法中求核方法的错误

属性约简的计算是粗糙集理论中的一个重要问题. 近年来,人们已提出了许多的属性约简算法,在这些算法中,求核的运算是必不可少的. 除了可以通过求出所有的不可缺少的属性来确定核[1]外, HU XIAOHUA 等学者在文献[3]中提出了更简洁的利用改进差别距阵来确定核的方法,其中改进的差别矩阵 $M = \{m_{ij}\}$ 定义为:

$$m_{ij} = \begin{cases} \{a \in C : F_q(x_i, a) \neq F_q(x_j, a)\}, \text{当 } F_q(x_i, D) \neq F_q(x_j, D) \text{ 时} \\ \Phi(\text{空集}), \text{其他情况时} \end{cases} \tag{2}$$

在上述定义中,与通常意义下的矩阵元素不同的是,差别矩阵的元素 m_{ij} 是条件属性 C 的某个子集. 文献[3]中未加证明地给出如下结论:当且仅当某个 m_{ij} 为单个属性时,该属性属于核 $Core(C)$. 需要指出的是,该方法存在着缺陷. 在一些情况下,它求得的结果是错误的. 为了说明这点,我们来看下面的一个反例.

例 1:表 1 所示的是一张二值数据表,其中共有 5 个元素和 4 个属性, $C = \{C1, C2, C3\}$ 为条件属性集, D 为决策属性.

表 1　反例数据表

元素 \ 属性	$C1$	$C2$	$C3$	D
x_1	1	0	1	1
x_2	1	0	1	0
x_3	0	0	1	1
x_4	0	0	1	0
x_5	1	1	1	1

如果根据式(2),只含单个属性的差别矩阵元素有 $m_{23} = \{C1\}$, $m_{25} = \{C2\}$. 由此可得 $Core(C) = \{C1, C2\}$. 然而,由计算可知 $\gamma_C = 1/5$,同时 $\gamma_{C2} = 1/5$,故 $\gamma_{C2} = \gamma_C$,因此单属性 $\{C2\}$ 为一个属性约简,但它并不包含核中的属性 $\{C1\}$. 因此,上述求核方法是错误的.

4　新的差别矩阵及求核方法

在例 1 中,相对决策属性而言,数据之间存在着不相容性. 那么,是否因为存在着不相容性而导致了 HU 的上述求核方法的错误呢?事实并非如此. 举一个例子说明.

例 2:在例 1 中删除掉第一个元素(即第一条记录),得到表 2:

表 2　数据表

元素 \ 属性	$C1$	$C2$	$C3$	D
x_1	1	0	1	0
x_2	0	0	1	1
x_3	0	0	1	0
x_4	1	1	1	1

显然,相对决策属性而言,数据之间仍然存在着不相容性. 通过简单计算容易得出: $R = \{C1, C2\}$ 是唯一的一个属性约简,故也是 $Core(C)$;另一方面,根据式(2),只含单个属性的差别矩阵元素有 $m_{12} = \{C1\}$, $m_{14} = \{C2\}$,从而 $Core(C) = \{C1, C2\} = R$. 因此,就例 2 而言,尽管存在着数据的不相容性,上述求核方法却是对的.

为了分析上述例子中数据不相容性的特点,引入一个记号. 对 $x_i \in U$,记 $d(x_i) = card\{F_q(y, D) : y \in [x_i]_C\}$, $d(x_i)$ 表示 U 中所有与 x_i 在关系 $IND(C)$ 下是等价的元素相应的决策属性值构成的集合的基数. 显然,我们有如下关于 $d(x_i)$ 的性质:当 $d(x_i) = 1$ 时,表示 $[x_i]_C$ 中的元素同属于划分 $\{Y_1, Y_2 \cdots\cdots Y_k\}$ 中的某一个子集,而当 $d(x_i) > 1$ 时,表示 $[x_i]_C$ 中的元素不属于划分 $\{Y_1, Y_2 \cdots\cdots Y_k\}$ 中的同一个子集,也即相对决策属性而言,数据存在着不相容性.

现在来考察例 1 的数据. 尽管 m_{23} 和 m_{25} 都是单个属性集,但 $\min\{|d(x_2)|, |d(x_3)|\} = 2 > 1$,而 $\min\{|d(x_2)|, |d(x_5)|\} = 1$,再来分析一下例 2. 此时 m_{12} 和 m_{14} 都是单个属性集,而且 $\min\{|d(x_1)|, |d(x_2)|\} = 1$, $\min\{|d(x_1)|, |d(x_4)|\} = 1$.

从上面的分析知,在应用差别矩阵 $M = \{m_{ij}\}$ 求核时,应该考虑到 $\min\{|d(x_i)|, |d(x_j)|\}$ 的因素. 下面我们将证明这一点. 首先,给出一个新的差别矩阵的定义.

定义 6:对于给定的信息系统式(1),定义差别矩阵 $MS = \{m'_{ij}\}$ 为:

$$m'_{ij} = \begin{cases} m_{ij}, \text{当 } \min\{|d(x_i)|, |d(x_j)|\} = 1 \\ \Phi(\text{空集}), \text{其他情况时} \end{cases} \tag{3}$$

其中 $\{m_{ij}\}$,如式(2)所定义.

定理:对于给定的信息系统式(1),若记 $SM(C) = \{m'_{ij} : m'_{ij} \text{为单个属性}\}$,则有 $SM(C) = Core(C)$,即当且仅当某个 m'_{ij} 为单个属性时,该属性属于核 $Core(C)$.

证明:首先证明 $SM(C) \subseteq Core(C)$. 任取一个属性 $a \in SM(C)$,由定义知至少存在矩阵 MS 中的一个元素,不妨设为 m'_{ij},使得 $m'_{ij} = \{a\}$. 于是,由定义 6 知 $x_j \notin [x_i]_C$,但是 $x_j \notin [x_i]_{C-\{a\}}$;由于 $F_q(x_i, D) \neq F_q(x_j, D)$,因此,对于划分 $\{Y_1, Y_2 \cdots\cdots Y_k\}$ 来说, x_i, x_j 不属于其中的同一个划分子集. 不妨设 $x_i \in Y_q$, $x_j \in Y_t$, $1 \leq q, t \leq k$, $t \neq q$. 由 $x_j \in [x_i]_{C-\{a\}}$, $x_j \notin Y_q$,以及下近似的定义即知: $x_i \notin \underline{C-\{a\}}\, Y_q$ 同样 $x_j \notin \underline{C-\{a\}}\, Y_q$. 因此,

$$x_i \notin Pos_{C-\{a\}}(D), \; x_j \notin Pos_{C-\{a\}}(D) \tag{4}$$

另一方面,由式(3), $\min\{|d(x_i)|, |d(x_j)|\} = 1$. 分两种情况考虑. (1) $d(x_i) = 1$. 根据 $d(x_i)$ 的性质可知 $[x_i]_C \in Y_p$,故 $x_i \in \underline{C}Y_p \subseteq Pos_C(D)$. 对比式(4)中的第一式,可得 $\gamma_{C-\{a\}} < \gamma_C$,故 $a \in Core(C)$. (2) $d(x_j) = 1$. 类似地,由 $d(x_j)$ 的性质可知 $[x_j]_C \in Y_t$,故 $x_j \in \underline{C}Y_t \subseteq Pos_C(D)$. 因此,对比式(4)中的第二式,可得 $\gamma_{C-\{a\}} < \gamma_C$,故 $a \in Core(C)$.

由此证得 $SM(C) \subseteq Core(C)$. 下面证明反包含 $SM(C) \supseteq Core(C)$ 成立. 采用反证法. 反设对某个 $a \in Core(C)$,不存在 m'_{ij},使得 $m'_{ij} = \{a\}$.

考虑划分 $\{Y_1, Y_2 \cdots\cdots Y_k\}$ 中的任一个子集 Y_l, $1 \leq l \leq k$ 任

取 $x_m \in \underline{C}Y_l \subseteq Pos_C(D)$,由下近似的定义知,$[x_m]_C \in Y_l$. 现任取 $x_p \in [x_m]_{C-\{a\}}$,我们来说明一定有 $x_p \in Y_l$. 为此,分两种情形讨论. (1) 如果 $p = m$,则显然 $x_p \in Y_l$; (2) 考虑 $p \neq m$ 的情形. 如果 $x_p \notin Y_l$,则 $x_p \notin [x_m]_C$ 而且 $F(x_p, D) \neq F(x_m, D)$. 此时,如果 $\min\{d(x_p), d(x_m)\} > 1$,则 $d(x_m) > 1$,表明 $[x_m]_C$ 中至少有一个元素,不妨设为 y,使得 $F(y, D) \neq F(x_m, D)$,即 $y \notin Y_l$,这与 $y \in [x_m]_C \in Y_l$ 矛盾. 因此 $\min\{d(x_p), d(x_m)\} = 1$. 由于 $x_p \in [x_m]_{C-\{a\}}$,根据定义 6,有 $m'_{mp} = \{a\}$,这与反设矛盾. 因此,$x_p \in Y_l$. 由 x_p 选取的任意性,可得 $[x_m]_{C-\{a\}} \subseteq Y_l$,即 $x_m \in \underline{C-\{a\}}\, Y_l \subseteq Pos_{C-\{a\}}(D)$. 再由 x_m 选取的任意性,可得 $\underline{C}Y_l \subseteq \underline{C-\{a\}}\, Y_l$, $1 \leq l \leq k$,因此,最终推得 $Pos_{C-\{a\}}(D) \supseteq Pos_C(D)$. 因为总有 $Pos_{C-\{a\}}(D) \subseteq Pos_C(D)$,故 $Pos_{C-\{a\}}(D) = Pos_C(D)$,即 $\gamma_{C-\{a\}} = \gamma_C$,由此可得 $a \notin Core(C)$,这与 $a \in Core(C)$ 的假设矛盾. 因此,反设不成立. 故 $SM(C) \supseteq Core(C)$. 定理由此得证.

参考文献:

[1] Pawlak Z. Rough set approach to multi - attribute decision analysis [J]. European Journal of Operational Research 1994,72:443 - 459.

[2] 曾黄麟. 粗集理论及其应用 [M]. 重庆:重庆大学出版社,1998.

[3] Hu XiaoHua, Cercone N. Learning in relational databases: a rough set approach [J]. Computational Intelligence, 1995, 11(2): 323 - 337.

[4] Jelonek J, et al. Rough set reduction of attributes and their domains for neural networks [J]. Computational Intelligence, 1995, 11(2): 338 - 347.

[5] 王珏等. 基于 Rough Set 理论的"数据浓缩" [J]. 计算机学报, 1998, 21(5): 393 - 400.

[6] 苗夺谦,胡桂荣. 知识约简的一种启发式算法 [J]. 计算机研究与发展, 1999, 36(6): 681 - 684.

[7] 叶东毅. Jelonek 属性约简算法的一个改进 [J]. 电子学报, 2000, 28(12): 81 - 82.

[8] 叶东毅,等. 粗糙集属性约简的一个贪心算法 [J]. 系统工程与电子技术, 2000, 22(9): 63 - 65.

作者简介:

叶东毅 男,1964 年出生于福建省南安市,博士,教授,主要从事神经网络学习算法、最优化计算和粗糙集理论等方面的研究,已发表论文 30 余篇,获省部级科技进步奖 2 项.

陈昭炯 女,1964 年出生于福建福州,副教授,主要从事计算机图形算法和粗糙集理论等方面的研究,已发表论文 10 余篇.