

一种基于决策信息系统的求核属性算法

胡秦斌

(广西师范学院 计算机与信息工程学院, 广西 南宁 530023)

摘 要: 为了有效降低求核算法的时空复杂度, 提出了一种高效压缩差别矩阵的构造方法. 在此基础上, 设计了一种决策信息系统的求核属性算法, 有效地提高了算法的计算效率. 算例分析验证了算法是有效可行的.

关键词: 决策信息系统; 核属性; 差别矩阵; 粗糙集

中图分类号: TP18

文献标识码: A

文章编号: 1000-7180(2012)07-0023-03

An Algorithm for Computing Core Attribute Based on Decision Information System

HU Qin-bin

(School of Computer and Information Engineering, Guangxi Teacher's Education University, Nanning 530023, China)

Abstract: To reduce the time complexity and space complexity of these algorithms, new construction method for compressed discernibility matrix is proposed, based on this method, an algorithm for computing core attribute in decision information system is presented, which remarkably improve the efficiency of the algorithm. Example result shows the efficiency and feasibility of the proposed algorithm.

Key words: decision information system; core attribute; discernibility matrix; rough set

1 引言

粗糙集理论是一种能够有效处理不精确、不完备与不一致信息的数学工具^[1], 该理论在数据挖掘、知识分析、机器学习及模式识别等诸多领域得到了广泛的应用^[2]. 因为核属性是所有的知识约简结果的核集, 为此目前绝大多数的启发式知识约简算法都是以核属性为基础, 然后利用启发信息逐步求解知识约简^[3-4]. 所以核属性的求解效率对于知识约简具有极其重要的意义.

在现有的求核属性算法中, 差别矩阵方法作为求解核属性的关键技术之一, 许多学者围绕差别矩阵进行了相关的大量研究. 差别矩阵方法是文献[5]中提出的一种经典求核算法, 但某些情况下得到的核属性可能是错误的. 文献[6]利用差别矩阵提出了一种核属性的增量式更新更新算法. 文献[7]中提出

了一种基于可分辨矩阵的快速求核算法. 上述基于差别矩阵的求核算法在很大程度上存在时间复杂度和空间复杂度较高, 当处理大规模数据集时, 需耗费大量时间去计算和存储决策信息系统的差别矩阵, 从而导致算法的计算效率并不理想.

为了提高求解核属性的计算效率, 本文针对一致的决策信息系统提出了高效压缩差别矩阵的新构造方法. 在求解核属性时, 不需要存储该差别矩阵, 显著的节省算法的存储空间, 基于该压缩的差别矩阵, 设计了一种基于决策信息系统的快速求核属性算法, 最后通过新算法分析说明了算法的可行性.

2. 基本知识

定义 1 在决策信息系统 $S = (U, C, D, V, f)$ 中, $U = \{x_1, x_2, \dots, x_{|U|}\}$ 表示对象的有限集, C 表示条件属性集, D 表示决策属性集, $V = \bigcup_{a \in C \cup D} V_a$, V_a 是

收稿日期: 2011-11-27; 修回日期: 2012-01-15

基金项目: 国家自然科学基金项目(61163012); 广西教育厅科研基金项目(201106LX302); 广西师范学院青年基金项目

属性 a 的值域集合, f 表示为信息函数 $U \times C \cup D \rightarrow V$, 即对于 $\forall a \in C \cup D, x \in U$, 存在 $f(x, a) \in V_a$; 并且每个属性子集 $P \subseteq (C \cup D)$ 决定一个不可区分关系, 即等价关系, $IND(P) = \{(x, y) \in U \times U \mid \forall c_i \in P, f(x, c_i) = f(y, c_i)\}$, $IND(P)$ 可将决策信息系统划分成等价类, 简记为 U/P . 由条件属性集形成的划分称为条件类, 由决策属性集形成的划分称为决策类.

定义 2 在决策信息系统 $S = (U, C, D, V, f)$ 中, 当两个对象 $x, y \in U$, 它们在条件属性集上的取值相同, 但决策属性上的取值不相同, 称它们是不一致的, 否则称它们是一致的. 内含不一致对象的系统称为不一致的决策信息系统, 否则称为一致的决策信息系统.

定义 3 在决策信息系统 $S = (U, C, D, V, f)$ 中, 对于 $\forall P \subseteq C \cup D$, 若 $U/P = \{P_1, P_2, \dots, P_m\}$ 且 $X \subseteq U$, 则称 $P_-(X) = \bigcup \{P_i \mid P_i \in U/P, P_i \subseteq X\}$ 为 X 关于属性集 P 的下近似集, 称 $P^+(X) = \bigcup \{P_i \mid P_i \in U/P, P_i \cap X \neq \emptyset\}$ 为 X 关于属性集 P 的上近似集.

定义 4 在决策信息系统 $S = (U, C, D, V, f)$ 中, 若 $U/D = \{D_1, D_2, \dots, D_k\}$ 为决策属性集 D 对 U 划分形成的决策类, 对于 $P \subseteq C$, $U/P = \{P_1, P_2, \dots, P_m\}$ 为条件属性集 P 对 U 划分形成的条件类, 则称 $POS_P(D) = \bigcup_{D_i \in U/D} P_-(D_i)$ 为条件属性集 P 关于决策属性集 D 的正区域.

定义 5 在决策信息系统 $S = (U, C, D, V, f)$ 中, 对于 $\forall Red \in C$, $POS_{Red}(D) = POS_C(D)$, 且若 $\forall c_i \in Red$ 时, $POS_{Red - \{c_i\}}(D) \neq POS_C(D)$, 则称 Red 为条件属性集 C 相对于决策属性集 D 的属性约简.

定义 6 在决策信息系统 $S = (U, C, D, V, f)$ 中, 差别矩阵定义为: $M = m(i, j)_{n \times n}$, 其矩阵元素的定义为: $m_{ij} =$

$$\begin{cases} \{a \in C: f(x_i, a) \neq f(x_j, a) \wedge f(x_i, D) \neq f(x_j, D)\} \\ \emptyset \text{ 其他} \end{cases}$$

定义 7 在决策信息系统 $S = (U, C, D, V, f)$ 中差别矩阵的核属性集 $Core$ 定义为: $Core = \{c_k \mid c_k \in m(i, j) \text{ 且 } |m(i, j)| = 1\}$.

3 算法描述

由于利用差别矩阵方法求解决策信息系统的核属性, 首先需要构建基于决策信息系统的差别矩阵, 而构造差别矩阵的常见方法是, 根据差别矩阵的定

义对决策信息系统中的对象做两两比较, 当决策信息系统中的对象个数为 $|U|$, 属性个数为 $|C|$ 时, 此时构建该差别矩阵的时间复杂度和空间复杂度均为 $O(|C| |U|^2)$. 本文通过对差别矩阵研究发现, 当决策信息系统的决策类别较少时, 将会出现许多空值元素, 这些元素占用了大量存储空间, 这也为后续的求核属性浪费了大量计算时间. 为此本文提出了一种以决策信息系统中的决策属性值为划分依据, 先将一致决策信息系统中的对象按照决策属性值的不同划分成多个对象集, 然后再生成差别矩阵, 这样可以有效减少当决策属性值相同时, 不必要的计算量, 同时可以高效压缩差别矩阵中空值元素, 提高了核属性的求解效率. 具体算法描述如下:

算法: 基于决策信息系统的求核属性算法

输入: 决策信息系统 $S = (U, C, D, V, f)$, 非空对象集 $U = \{x_1, x_2, \dots, x_{|U|}\}$, 条件属性集 $C = \{c_1, c_2, \dots, c_{|C|}\}$, 决策属性集 $D = \{d\}$;

输出: 核属性集 $Core$;

步骤 1: $M = \emptyset$; $Core = \emptyset$;

步骤 2: $t = 1$; $D_t = \emptyset$;

步骤 3: // 根据决策属性对决策信息系统划分决策类

for ($t = 1$; $t \leq |U|$; $t++$)

if ($f(x_t, d) \neq f(x_{t+1}, d)$)

$D_t = D_t \cup D_{t+1}$;

else $t++$;

步骤 4: // 假定求得决策类为

$U/D = \{D_1, D_2, \dots, D_{|U/D|}\}$

for ($i = 1$; $i \leq |U/D|$; $i++$)

for ($j = i$; $j \leq |U/D|$; $j++$)

if ($x \in D_i \wedge y \in D_j$)

// 根据决策属性值生成差别矩阵

for ($k = 1$; $k \leq |C|$; $k++$)

{if ($f(x, c_k) \neq f(y, c_k)$)

$m(x, y) = m(x, y) \cup \{c_k\}$;

// 求解决策信息系统的核属性

if ($|m(x, y)| = 1 \wedge c_i \in m(x, y)$)

$Core = Core \cup \{c_k\}$;

}

步骤 5: 输出决策信息系统的核属性集 $Core$.

算法时空复杂度分析:

算法步骤 1 的时间复杂度为 $O(1)$; 步骤 2 的时间复杂度为 $O(1)$, 步骤 3 的时间复杂度为 $O(|U|)$;

算法步骤 4 的时间复杂度为 $O(|C| \sum_{i=1}^{|U/D|-1} \sum_{j=i+1}^{|U/D|} |D_i| |D_j|)$; 步骤 5 的时间复杂度为 $O(1)$; 所以本文算法的时间复杂度为 $O(|C| \sum_{i=1}^{|U/D|-1} \sum_{j=i+1}^{|U/D|} |D_i| |D_j|)$. 由于算法在生成压缩的差别矩阵元素时, 便可动态求解决策信息系统的核属性, 使得算法的空间复杂度降为 $O(|C|)$. 对于一致的决策信息系统而言, 文献[6-7]中求核属性算法的时间复杂度和空间复杂度均为 $O(|C| |U|^2)$. 由于 $\sum_{i=1}^{|U/D|-1} \sum_{j=i+1}^{|U/D|} |D_i| |D_j| < |U|^2$, 为此本文算法的计算效率可得到显著提高, 从而能较好地适应大规模数据集的处理.

4 算例分析

为了进一步说明本文所提出的算法的可行性和有效性, 下面对文献[8]中实例进行详细分析, 其中 $\{x_1, x_2, \dots, x_8\}$ 为对象集, $\{\text{Size}, \text{Hair}, \text{Eyes}, \text{Complexion}\}$ 为条件属性集并分别用 $\{c_1, c_2, c_3, c_4\}$ 表示, 决策属性集为 $\{D\}$. 如表 1 所示.

表 1 决策信息系统

U	c_1	c_2	c_3	c_4	D
x_1	short	dark	blue	pale	0
x_2	Tall	dark	brown	matt	1
x_3	Tall	red	blue	pale	0
x_4	short	blond	blue	matt	1
x_5	Tall	blond	blue	pale	1
x_6	Tall	dark	blue	pale	0
x_7	Tall	blond	brown	matt	1
x_8	short	dark	brown	matt	1

若根据文献[6-7]中算法对差别矩阵的构造, 将得到如表 2 所示的差别矩阵. 从表 2 中的差别矩阵元素可以看出, 矩阵中产生了大量的空值元素, 明显地浪费了大量的存储空间. 而造成空值元素的原因是由于具有相同决策属性值的对象两两比较的结果, 这些元素是无需存储的. 为此, 可根据本文算法的步骤 3, 先根据决策属性对决策信息系统中的对象进行划分, 得到决策类为 $U/D = \{\{x_1, x_3, x_6\}, \{x_2, x_4, x_5, x_7, x_8\}\}$, 然后根据算法的步骤 4, 可求解出压缩后的差别矩阵, 见表 3 所示. 然后根据定义 7 可动态求解出该决策信息系统的核属性集为 Core

$= \{c_2\}$, 即核属性集为 $\{\text{Hair}\}$.

表 2 原差别矩阵

U	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
x_1	$\emptyset$	$\{c_1, c_3, c_4\}$	$\emptyset$	$\{c_2, c_4\}$	$\{c_1, c_2\}$	$\emptyset$	$\{c_1, c_2, c_3, c_4\}$	$\{c_3, c_4\}$
x_2		$\emptyset$	$\{c_2, c_3, c_4\}$	$\emptyset$	$\emptyset$	$\{c_3, c_4\}$	$\emptyset$	$\emptyset$
x_3			$\emptyset$	$\{c_1, c_2, c_4\}$	$\{c_2\}$	$\emptyset$	$\{c_2, c_3, c_4\}$	$\{c_1, c_2, c_3, c_4\}$
x_4				$\emptyset$	$\emptyset$	$\{c_1, c_2, c_4\}$	$\emptyset$	$\emptyset$
x_5					$\emptyset$	$\{c_2\}$	$\emptyset$	$\emptyset$
x_6						$\emptyset$	$\{c_2, c_3, c_4\}$	$\{c_1, c_3, c_4\}$
x_7							$\emptyset$	$\emptyset$
x_8								$\emptyset$

表 3 压缩后的差别矩阵

U	x_2	x_4	x_5	x_7	x_8
x_1	$\{c_1, c_3, c_4\}$	$\{c_2, c_4\}$	$\{c_1, c_2\}$	$\{c_1, c_2, c_3, c_4\}$	$\{c_3, c_4\}$
x_3	$\{c_2, c_3, c_4\}$	$\{c_1, c_2, c_4\}$	$\{c_2\}$	$\{c_2, c_3, c_4\}$	$\{c_1, c_2, c_3, c_4\}$
x_6	$\{c_3, c_4\}$	$\{c_1, c_2, c_4\}$	$\{c_2\}$	$\{c_2, c_3, c_4\}$	$\{c_1, c_3, c_4\}$

5 结束语

差别矩阵已成为粗糙集理论中求解核属性的重要方法之一, 但差别矩阵方法在处理大规模决策信息系统时, 构建差别矩阵将需要消耗大量的存储空间和计算时间, 导致算法的效率低下, 为了提高求核属性的计算效率, 本文提出一种高效压缩差别矩阵的构造方法, 并在此基础上, 设计了一种基于决策信息系统的快速求解核属性算法. 本文算法能够显著压缩差别矩阵, 有效地减少了求解核属性的计算复杂度. 本文的算法在时间复杂度和空间复杂度上要优于现有的同类算法. 文中主要研究了决策信息系统的求核属性算法, 对于决策信息系统的知识约简算法将是下一步的研究重点.

(下转第 31 页)

表4 首轮属性值与粒极值对应表

U	c1	c2	c3	c4	d	c1	c2	c3	c4
1	0	1	2	1	0	3.1	4.1	7.1	0.0
						0	0	0	0
2	0	2	1	1	1	3.1	7.1	4.1	0.0
						0	0	0	0
34	1	0	1	1	01	1.1	1.1	0.1	0.0
						1.1.0	0.0.0	1.0.0	1.1.0
56	1	1	1	1	01	3.0	1.0	2.0	0.0
						3.0.0	2.0.0	1.0.0	3.0.0
89	1	1	0	1	01	1.1	0.1	1.1	0.0
						1.1.0	1.0.0	0.0.0	1.1.0

4 结束语

在文献[4]的基础上,增加了确定极值行集的算法、将以大值(二进制数中为1的个数最多)判定粒极值修改为以逻辑关联判定:

(1) 以属性值的RD和CD的域长首先按级定出极值行集,使得先处理最关键的极值行,在极值行中提取规则后,与之相关的非极值行都被约简,减少了运算复杂度。

(2) 根据解线性方程原理,提出逻辑相关的概念.将以判定二进制为1个数的多少修改为以逻辑变量相关程度来确定粒极值,使计算原理系统化从

而形成完整、准确的判定序列。

(3) 将相容表和不相容表统一处理,即单因多果的情况替代为单因单果情况处理后再展开为单因多果的规则.统一各类信息表的规则提取算法。

(4) 对大量各类信息表运算结果(本文只列出部分)都证明了所提规则最简而并非是一个NP-Hard问题^[3],算法具有较高的理论和实用价值。

参考文献

- [1] 钱锋. 结合模糊集理论的粗糙集属性约简算法[J]. 计算机应用研究, 2007, 24(11): 93-95.
- [2] 卓丽霞. 粒计算规则生成模型[J]. 重庆工学院学报自然科学, 2008, 22(9): 149-153.
- [3] 刘丰年. 基于分明矩阵的序贯属性约简算法[J]. 微电子学与计算机, 2009, 26(7): 95-97.
- [4] 何群. 基于G逻辑的最简规则提取算法[J]. 微电子学与计算机, 2010, 27(10): 91-95.
- [5] 富震. 基于自适应遗传约简算法的雷达故障诊断[J]. 控制与决策, 2010, 24(11): 93-95.
- [6] 李莉. 一种否定高信度规则的属性查找算法[J]. 计算机工程, 2008, 34(6): 65-69.
- [7] 刘海涛. 不相容决策表中一种新的属性约简算法[J]. 计算机工程与应用, 2011, 47(2): 124-126.

作者简介

何群 男, (1955-), 硕士, 高工, 研究生导师. 研究方向为数据挖掘、粗糙集、粒计算。

(上接第25页)

参考文献:

- [1] Pawlak Z, Skowron A. Rudiments of rough sets[J]. Information Science, 2007, 117(1): 3-37.
- [2] 岳晓冬, 苗夺谦, 钟才明. 基于粗糙性度量的彩色图像分割方法[J]. 自动化学报, 2010, 36(6): 807-816.
- [3] 胡清华, 于达仁, 谢宗霞. 基于邻域粒化和粗糙逼近的数值属性约简[J]. 软件学报, 2008, 19(3): 640-649.
- [4] 崔巍, 徐章艳. 一种新的信息系统启发式属性约简算法[J]. 微电子学与计算机, 2010, 27(9): 80-85.
- [5] Hu X H, Cercone N. Learning in relational databases: A rough set approach[J]. Computational Intelligence, 1995, 11(2): 323-337.
- [6] 杨明. 一种基于改进差别矩阵的核增量式更新算法[J]. 计算机学报, 2006, 29(3): 407-413.
- [7] 葛浩, 杨传健, 李龙澍. 基于可分辨矩阵的快速求核算法[J]. 计算机工程与设计, 2009, 30(5): 1201-1203.
- [8] 程玉胜, 张佑生, 胡学钢. 两类决策系统中规则获取的联合决策矩阵算法[J]. 系统工程理论与实践, 2008, 28(6): 137-142.

作者简介:

胡秦斌 男, (1974-), 硕士, 讲师. 研究方向为数据挖掘、计算智能。