对新简化差别矩阵的研究

王 磊
(燕山大学 机械学院 CAD 中心,河北 秦皇岛 066004)

摘 要:通过反例证明当决策表含不一致对象时,无法通过简化差别矩阵得到正确的核和约简。产生该问题的原因在于简化差别矩阵和简化决策表对不一致对象的处理均存在欠缺。提出新简化差别矩阵(NSDM)定义,通过利用新简化决策表把一致对象和不一致对象分开存放的特点修正了这两处缺陷。NSDM能够处理含多种不一致对象的决策表,并且能够得到正确的核和约简。利用前人的基于差别矩阵的求核和求约简算法分别以NSDM和简化差别矩阵为基础求核和约简,然后再利用核和约简的定义求核和约简,通过比较证明以NSDM为基础得到的核和约简是正确的。这说明NSDM可以有效地处理不一致对象并且通过NSDM能够得到正确的核和约简。这个实验同时验证了原有的以差别矩阵为基础的求核和求约简的算法不加改动就可以在新差别矩阵上使用。

关键词:Rough集;约简;核;新简化差别矩阵
中图分类号:TP18 **文献标识码**:A **文章编号**:1673-629X(2009)08-0062-03

Research on New Simple Discernibility Matrix

WANG Lei
(Center of CAD, College of Mech. Eng., Yanshan Univ., Qinhuangdao 066004, China)

Abstract: Simple discernibility matrix (SDM) can not calculate cores and reductions correctly when there are inconsistent objects in decision tables. The errors are in SDM and simple decision table. New simple discernibility matrix(NSDM) is presented to correct these errors by inconsistent objects and consistent objects are saved respectively in new simple decision table. NSDM can get correct cores and reductions from decision table with many sorts of inconsistent objects. Finally, cores and reductions are calculated based on SDM and NSDM by some former algorithms for cores and reductions based on discernibility matrix. NSDM can get correct cores and reductions by comparing to results from definitions of core and reduction. These results show that NSDM is efficient and NSDM can work by former algorithms based on discernibility matrix.

Key words: rough sets; reduction; core; new simple discernibility matrix

0 引 言

文献[1]在差别矩阵[2,3]基础上提出的简化差别矩阵有效地降低了差别矩阵的空间复杂度,从而大幅降低了以此为基础的求约简和求核算法的空间、时间复杂度。但是简化差别矩阵是建立在简化决策表的基础上的[4],而文献[5]已经证明简化决策表存在缺陷,所以简化差别矩阵存在由简化决策表的缺陷引发的丢失信息问题。

下文中将提出反例证明简化差别矩阵存在丢失信息问题,并给出新简化差别矩阵(NSDM)的定义,NSDM解决了简化差别矩阵的丢失信息问题,并且解决了文献[6]中提出的差别矩阵在不同种不一致对象间会产生冗余属性的问题,而且原有的以简化差别矩阵、差别矩阵为基础的求约简和核的算法不需改动即可继续使用。

收稿日期:2008-12-04;**修回日期**:2009-03-01
基金项目:天津市高等学校科技发展基金(20061011)
作者简介:王 磊(1973-),女,黑龙江人,硕士,研究方向为Rough集理论及应用、粒度计算、知识发现。

1 基本概念

1.1 文献[1,2]提出的简化差别矩阵及相关定义定理

定义1[4]:$T=(U,C,D,V,f)$是一个决策表,其中论域$U=(u_1,u_2,\cdots,u_n)$,条件属性集$C=(c_1,c_2,\cdots,c_m)$,决策属性$D=(d)$,$V=\bigcup_{a\in C\cup D}V_a$,$V_a$是属性$a$的值域,$f:U\times(C\cup D)\rightarrow V$是一个信息函数,它对一个对象的每一个属性赋予一个信息值,即$\forall a\in C\cup D, x\in U$,有$f(x,a)\in V_a$;每个属性子集$P\subseteq(C\cup D)$决定了一个二元不可分辨关系IND(P):

$$IND(P)=\{(x,y)\in U\times U \mid \forall a\in P, f(x,a)=f(y,a)\}$$

关系 IND(P) 构成了 U 的一个划分，用 $U/\mathrm{IND}(P)$ 表示，简记为 U/P，U/P 中的任何元素 $[x]_P = \{y \mid \forall a \in P, f(x,a) = f(y,a)\}$ 称为等价类。

定理 1[4]:在决策表 $T = (U,C,D,V,f)$ 中，记 $\mathrm{POS}_C D = \bigcup_{X \in U/C \text{且} \forall x,y \in X \Rightarrow f(x,D) = f(y,D)} X$。

该定理说明，C 关于 D 的正区域是由在决策属性上取值唯一的基本块的并集组成，任何决策属性值不唯一的基本块不在正区域中。称 U/C 中的等价类为基本块。

定义 2[4]:在决策表 $T = (U,C,D,V,f)$ 中，记 $U/C = \{[u'_1]_C, [u'_2]_C, \cdots, [u'_m]_C\}$，$U' = \{u'_1, u'_2, \cdots, u'_m\}$，由定理 1 可设 $\mathrm{POS}_C(D) = [u'_{i_1}]_C \cup [u'_{i_2}]_C \cup \cdots \cup [u'_{i_t}]_C$，其中 $\forall u'_{i_s} \in U'$ 且 $|[u'_{i_s}]_C/D| = 1 (s = 1,2,\cdots,t)$；记 $U'_{\mathrm{pos}} = \{u'_{i_1}, u'_{i_2}, \cdots, u'_{i_t}\}$，$U'_{\mathrm{neg}} = U' - U'_{\mathrm{pos}}$。称 $T' = (U',C,D,V,f)$ 为简化决策表。

定义 3[1]:在决策表 $T = (U,C,D,V,f,d)$ 中，$T' = (U',C,D,V,f',d')$ 为简化决策表，定义简化差别矩阵为 $\boldsymbol{M}' = (m_{i'j'})$，其元素定义如下：

$$m_{i'j'} = \begin{cases} \{a \mid a \in C, f(x_{i'},a) \neq f(x_{j'},a), \text{当 } x_i, x_j \text{ 至多有一个在 } U'_{\mathrm{pos}} \text{ 中；或者 } f(x_{i'},a) \neq f(x_{j'},a), f(x_{i'},D) \neq f(x_{j'},D), \text{当 } x_i, x_j \text{ 都在 } U'_{\mathrm{pos}} \text{ 中}\}, \\ \phi, \quad \text{否则} \end{cases}$$

定义 4[1]:设 $\boldsymbol{M} = (m_{ij})$ 是决策表 $T = (U,C,D,V,f,d)$ 的差别矩阵，称 $D\mathrm{Core}$ 为差别矩阵的核。其中 $D\mathrm{Core}(C)$ 定义 $D\mathrm{Core}(C) = \{c_k \mid \exists m_{i'j'} = \{c_k\}\}$。

定义 5[1]:设 $\boldsymbol{M}' = (m_{i'j'})$ 是决策表 $T = (U,C,D,V,f,d)$ 的简化差别矩阵，称 TDCore 为简化差别矩阵的核。其中 $\mathrm{TDCore}(C)$ 定义 $\mathrm{TDCore}(C) = \{c_k \mid \exists m_{i'j'} = \{c_k\}\}$。其中 $k = 1,2,\cdots,r$。

1.2　新简化差别矩阵及相关定义

定义 6[5]:在决策表 $T = (U,C,D)$ 中，记 $U/C = \{[u'_1]_C, [u'_2]_C, \cdots, [u'_m]_C\}$，若 $|[u'_i]_C/D| = 1$ 则提取 $u_j' \in [u'_i]_C$ 放入 U_1'，若 $|[u'_i]_C/D| \neq 1$ 则提取 $u_i', u_j', \cdots, u_s' \in [u'_i]_C$ 放入 U_2' 中(其中 $u_i' \neq u_j' \neq \cdots \neq u_s'$)，称 $T' = (U',C,D)$ 为新简化决策表，其中 $U' = U_1' + U_2'$。

定义 7:$T = (U,C,D,V,f)$ 是一个决策表，其中论域 $U = (u_1,u_2,\cdots,u_n)$，条件属性集 $C = (c_1,c_2,\cdots,c_m)$，决策属性 $D = (d)$，设决策表 $T = (U,C,D)$，其中 $U = U_1 + U_2, U_1 = \{u_1,u_2,\cdots,u_p\}$ 为一致对象集，$U_2 = \{u_{p+1},u_{p+2},\cdots,u_q\}$ 为不一致对象集，T 的新简化差别矩阵记为 $\boldsymbol{M}_t$，设 m_{ij} 为 $\boldsymbol{M}_t$ 中元素，则：$m_{ij} = \{a \in C \mid a(u_i) \neq a(u_j) \wedge (d \in D, d(u_i) \neq d(u_j)) \wedge u_i, u_j \in U\} (i = 1,2,\cdots,p; j = 1,2,\cdots,q; q > p)$，其他情况 $m_{ij} = \phi$。

简化差别矩阵的问题在于，在特定的情况下可能丢失一致对象和不一致对象的比较结果。

新简化差别矩阵定义的实质是，利用新简化决策表对一致对象和不一致对象分开存放的特点，把一致对象和不一致对象的比较结果放入矩阵，这样就解决了简化差别矩阵丢失信息的问题；另一方面新简化差别矩阵不在不一致对象间进行比较，这样就解决了文献[6]提出的不同种不一致对象间会产生冗余属性的问题。

2　简化差别矩阵的反例

举反例证明简化差别矩阵存在丢失信息的问题，并且导致基于简化差别矩阵得到的约简和核也存在由此引发的错误。下面举例说明。

T(见表 1) 是一个决策表，由简化差别矩阵定义和文献[4] 中的简化决策表建立算法可得 T 的简化决策表 T'(见表2)，由 T' 易得简化差别矩阵 $\boldsymbol{M}_t' = \phi$，利用简化差别矩阵 $\boldsymbol{M}_t'$ 易得 T 的核和约简均为 ϕ。而通过核和约简的定义求得 T 的核和约简均为 $\{a\}$。可以看出在特定情况下简化差别矩阵存在丢失信息问题，并影响求出的核和约简。

表 1　决策表 T

$U \setminus A$	a	b	d
1	2	0	0
2	2	0	1
3	2	1	0
4	2	1	1
5	0	0	0
6	0	0	0

表 2　T 的简化决策表 T'

$U \setminus A$	a	b	d
1	2	0	0
5	0	0	0
3	2	1	0

下面说明简化差别矩阵存在的第二个缺陷，即文献[6] 中指出的差别矩阵存在不能处理不同种不一致对象的问题而在简化差别矩阵同样存在。该问题产生原因是由于简化差别矩阵对所有决策值不同的对象进行比较，而不同种不一致对象的决策值也是不同的，就是说简化差别矩阵对不一致对象进行了比较，即简化差别矩阵认为不一致对象是可以区分的，这就造成了

简化差别矩阵会产生冗余信息。由于简化差别矩阵是建立在简化决策表基础上的，而简化决策表存在丢失或改变信息的缺陷，即简化决策表把所有的不一致对象作为一致对象处理，这样在简化决策表的基础上的简化差别矩阵就无法验证和不一致对象相关的例子，为了验证这个问题在新简化决策表的基础上建立简化差别矩阵来验证简化差别矩阵无法处理不同种不一致对象的缺陷。验证方法如下，在表3的基础上建立简化差别矩阵，可以看到简化差别矩阵存在由不同种不一致对象形成的冗余信息，并且这些冗余信息会影响得到的核和约简。对表3建立简化差别矩阵，u_5与u_2，u_5与u_4，u_1与u_4，u_2与u_3比较结果：{{a}，{ab}，{b}，{b}}，可得到核和约简是{a，b}，由上文知道T的核和约简是{a}，显然b冗余的。其中u_1，u_2，u_3，u_4是不一致对象，u_1与u_4，u_2与u_3是不同种不一致对象，它们之间形成的信息{b}是冗余信息。通过该例可知简化差别矩阵不具备处理不同种不一致对象的能力。

3 新简化差别矩阵

简化差别矩阵建立过程中由于忽视了T中对象5与对象2、4存在的差别而产生了上一节中提到的丢失信息的问题，针对该问题把简化差别矩阵丢失的信息放入新简化差别矩阵即可解决第2节中的丢失信息问题。

下面用Hu的求核算法[7]分别在简化差别矩阵和新简化差别矩阵上求核并与用核的定义求出的T的核比较，可以看出新简化差别矩阵和核的定义求出的核是一致的，简化差别矩阵和核的定义求出的核是不同的。

再使用文献[8]中的求约简算法分别计算简化差别矩阵和新简化差别矩阵的约简，并用约简定义求T的约简。可以得到利用新差别矩阵求得的约简与用约简定义得到的约简相同，而用简化差别矩阵求得的约简与约简定义得到的约简不同。

可以看出新简化差别矩阵是有效的，简化差别矩阵存在的丢失信息问题会影响到以简化差别矩阵为基础的求约简和求核算法；而且利用新简化差别矩阵求核和约简对于以前的求核或求约简算法不需要改动就可以得到正确结果。

下面举例详细说明。按文献[5]中NSDT算法对T建立新简化决策表S(见表3)，$S=(U',C,D)$，其中$U'=U_1'+U_2'$，$U_1'=\{u_5\}$，$U_2'=\{u_1,u_2,u_3,u_4\}$，按定义7建立新简化差别矩阵M_S(见表4)方法如下：(1)对U_1'中对象进行两两比较，把比较结果放入M_S，此例中U_1'仅有一个对象，所以没有生成比较结果；(2)把U_1'中对象和U_2'中对象进行比较，结果放入M_S(即u_5和u_2比较结果a；u_5和u_4比较结果ab)。

表3 T的新简化决策表S

$U\setminus A$	a	b	d
5	0	0	0
1	2	0	0
2	2	0	1
3	2	1	0
4	2	1	1

表4 T的新简化差别矩阵M_S

$U\setminus U$	5
1	
2	a
3	
4	ab
5	

按Hu的算法由简化差别矩阵$M_t'=\phi$可以算出T的核为ϕ(简化差别矩阵中单属性元素就是核)，从新简化差别矩阵M_S可以算出T的核为{a}。由核的定义(去掉a后不可分辨关系发生变化，去掉b后不可分辨关系不变)可知T的核是{a}。

按文献[8]中算法可以得到由简化差别矩阵求出的约简为ϕ，由新简化差别矩阵求出的约简为{a}。由约简定义(a即可区分开T中所有不可分辨对象)可知T的约简是{a}。

新简化差别矩阵解决了文献[6]指出的由差别矩阵求得的核和约简存在冗余属性的问题。文献[6]指出该问题的本质在于差别矩阵对不同种不一致对象进行比较产生冗余信息，而新差别矩阵把不一致对象单独存放，不对不一致对象进行两两比较，所以不会产生冗余信息，从而解决了这个问题。下面举例说明该问题。

仍以T为例，易得T的差别矩阵中的元素为{{a}，{a，b}，{b}，{b}}，$u_5(u_6)$和u_2比较得a；$u_5(u_6)$和u_4比较得ab，u_1和u_4比较得b，u_2和u_3比较得b。对传统的差别矩阵用Hu的算法易得其核和约简均为{a，b}，显然b是冗余属性，而新简化差别矩阵求得核和约简均为{a}，不存在冗余属性的问题。

4 结束语

简化差别矩阵会产生丢失信息问题并影响以简化差别矩阵为基础得到的约简和核。文中修正了该问题，给出了新简化差别矩阵定义。并且不用对原来的基于差别矩阵和简化差别矩阵的求核或求约简算法进行改动即可在新简化差别矩阵下使用这些算法并得到

(下转第68页)

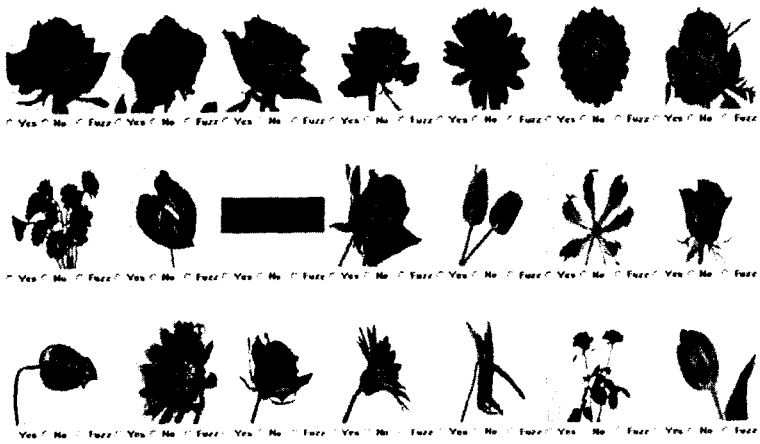

图 3 第 4 次反馈的检索结果

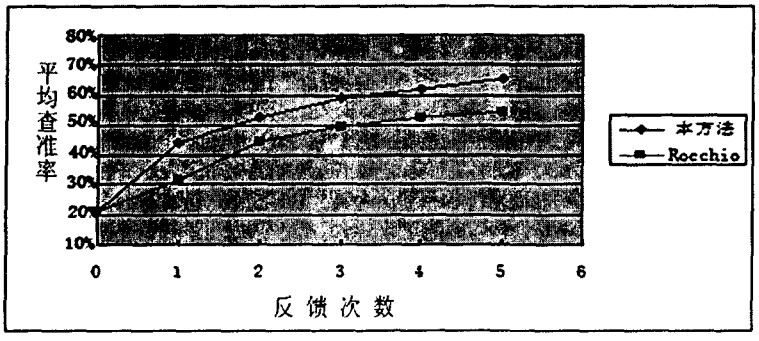

图 4 两种方法检索性能比较

关图像的数目在增加,而且相关图像排在检索结果的位置也在提前,这比较符合用户的检索习惯。这说明该方法不仅是有效的,而且具有较好的检索性能。

4 结束语

采用三级反馈机制引入模糊相关,并修改经典的查询向量点移动算法,在此基础上运用多分类 SVM 提出一种新的相关反馈图像检索方法。通过试验可以看到这是一个有效的方法,提高了图像检索效率,能更好地满足用户的查询要求。

参考文献:

[1] 彭太乐,蒋建国,魏仁民,等. 一种融合语义的图像检索技术研究[J]. 计算机技术与发展,2008,18(3):102-104.

[2] 胡 莹. 一种改进的 SVM 相关反馈图像检索方法[J]. 计算机应用研究,2005(1):251-254.

[3] 张 苗,张德贤. 多类支持向量机文本分类方法[J]. 计算机技术与发展,2008,18(3):139-141.

[4] 常小红,张 明. 基于 RBFN 的交互式图像检索方法[J]. 计算机技术与发展,2007,17(9):31-34.

[5] 胡 珊. 图像检索中基于 SVM 的相关反馈技术研究[D]. 西安:西北工业大学,2007.

[6] Mehtre B W. Content based image retrieval Using a Composite color-shape approach[J]. Information Processing & Management, 1998,33(30):319-337.

[7] Zhou Z H, Chen K J, Jiang Y. Exploiting unlabeled data in content-based image retrieval[C]// In: Proceedings of the 15th European conference on machine learning. Pisa, Italy:[s.n.],2004:525-536.

[8] Trazegnizes C,Bandera A. Planar Shape indexing and retrieval based on Hidden Markov Models[J]. PRI, 2002, 23(20): 1143-1151.

[9] Yap Kim-Hui, Wu Kui. A soft relevance framework in content-based image retrieval systems[J]. IEEE transaction on circuits and systems for video technology,2005,15(12):1557-1568.

(上接第 64 页)

正确结果。而且新简化差别矩阵解决了文献[6]中指出的差别矩阵对不一致对象的处理是不完备的问题。

参考文献:

[1] 徐章艳,杨炳儒,宋 威. 一个基于差别矩阵的快速求核算法[J]. 计算机工程与应用,2006,42(6):4-6.

[2] Fleix R, Ushio T. Rough Sets-based Machine Learning Using a Binary Discernibility Matrix[C]// In: IPMM'99. [s.l.]:[s.n.], 1999:299-305.

[3] Skowron A,Rauszer C. The Discernibility Matrices and Functions in Information Systems[C]// In: Slowinski R. Intelligent Decision Support Handbook of Applications and Advances of the Rough Sets Theory. Dordrecht: Kluwer Academic Publishers,1992:331-362.

[4] 徐章艳,刘作鹏,杨炳儒,等. 一个复杂度为 max(O(|C||U|),O(|C|²|U/C|))的快速属性约简算法[J]. 计算机学报,2006,29(3):391-399.

[5] 王希雷,苏 静. 对决策表和简化决策表的研究[J]. 计算机技术与发展,2008,18(4):110-118.

[6] 王希雷,王 磊. 粗集中区分矩阵对不一致问题的处理的研究[J]. 微机发展(现更名:计算机技术与发展),2003,13(2):119-120.

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

[8] 王元珍,裴小兵. 基于 Skowron 分明矩阵的快速约简算法[J]. 计算机科学,2005,32(4):42-44.