基于依赖度之差的属性重要性评分

王小菊,蒋 芸,李永华
(西北师范大学 数学与信息学院,甘肃 兰州 730070)

摘 要:目前粗糙集决策表中条件属性的重要性基本上是用条件属性的依赖度进行评判的。在决策表约简中,利用条件属性的依赖度进行评判可能会造成某些重要的条件属性的简单丢弃,影响了决策的准确性。因此提出并分析了基于依赖度之差的属性重要性的判断方法,该方法可以确保得到决策表的重要属性,得出了用依赖度之差判断属性重要性更加准确的结论,同时,给出了依赖度之差的求解步骤和算法,并通过实例验证了用依赖度之差判断属性重要性的有效性。

关键词:粗糙集;决策表;依赖度;依赖度之差;属性重要性
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2009)01-0067-04

Significance of Attribute Evaluation Based on Dependable Difference

WANG Xiao-ju, JIANG Yun, LI Yong-hua
(College of Mathematics and Information Science, Northwest Normal University, Lanzhou 730070, China)

Abstract: At present basically significance of conditional attribute in the decisional table based on rough sets is evaluated through dependability of conditional attribute. Evaluation by using dependability of conditional attribute possible discard some important conditional attribute, it may affect the accuracy of the decision-making. Proposes and analyzes an approach of judging importance of attribute based on dependable difference. the approach can get important attribute, It is obtained that judging importance of attribute based on dependable difference of attribute is more accurate. And dependable difference of attribute detailed solve procedure and arithmetic is provided. The efficiency of judging importance of attribute based on dependable difference of attribute is proven through the example.

Key words: rough sets; decisional table; dependability; dependable difference; significance of attribute

0 引 言

决策表是一类特殊而重要的知识表达系统,它同时具有条件属性和决策属性,决策表约简的重要内容之一是化简决策表中的条件属性,约简后的决策表与约简前的决策表应具有一致的决策功能,但约简后的决策表具有更少的条件属性,因此,条件属性重要性的判断显得极为重要。目前,很多学者在该领域提出了不少有效的方法,希望找到具有最少属性的约简,然而,要找到一个决策表的最小约简是个 NP-hard 问题。

在一个决策表中,不同的属性具有的重要程度也是不同的。通常使用依赖度来描述条件属性的重要性[1],当依赖度等于1时,称决策属性完全依赖于条件属性;当依赖度大于0,且小于1时,称决策属性部分依赖于条件属性;当依赖度等于0时,称决策属性完全独立于条件属性。绝大多数文献对依赖度的研究,仅限于求解单一属性依赖度,认为单一属性依赖度为0的属性对决策无作用,应将其舍弃。然而,通常这样做往往会造成重要属性被简单地丢弃,致使重大决策判断错误。在文献[2]中,通过对属性集(两个或两个以上的属性构成)依赖度作进一步研究,认为依赖度为0的单一属性在知识约简时不能轻易舍弃,但当属性的依赖度相同时仅从依赖度无法判断哪些属性相对重要,哪些属性相对不重要。文献[3]求解的仍是单一属性属性依赖度。在文献[4~6]中对条件属性的重要性描述也都使用了依赖度,单个条件属性重要性的判断使用了依赖度之差,不足之处是当依赖度等于0时,认为决策属性完全独立于条件属性,还有只是用依赖度之差来判断单一属性的重要性,也没有给出依赖度之差的求解步骤和算法。因此,研究单一属性依赖度之差和属性集依赖度之差,具有更加重要的意义。

文中利用依赖度之差来判断单一属性或属性集的

收稿日期:2008-04-09
基金项目:甘肃省自然科学基金项目(3ZS051-A25-042);西北师范大学2006-2010年度重点学科资助项目
作者简介:王小菊(1973-),女,山西晋城人,硕士研究生,主要研究方向为数据挖掘;蒋 芸,博士,副教授,硕士生导师,主要研究方向为数据挖掘、粗糙集。

重要性,大大提高了判断某一属性和某些属性集重要性的准确性,避免了把单一属性依赖度或属性集依赖度为 0 的属性舍弃,同时能够判断单一属性或属性集中哪个或哪些属性更重要,哪个或哪些属性可约去,并且其求解方法简单。

1 基本概念

定义 1[7,8]一个信息系统 S 可以表示为有序四元组:$S=\{U,A,V,F\}$,其中 $U=\{x_1,x_2,\cdots,x_n\}$ 为论域,它是全体样本的集合;$R=C\cup D$ 为属性集合,其中子集 C 是条件属性集,反映对象的特征,D 为决策属性集,反映对象的类别;$V=\bigcup_{r\in R}V_r$ 为属性值的集合,V_r 表示属性 r 的取值范围;$f:U\times R\to V$ 为一个信息函数,用于确定 U 中每一个对象 x 的属性值,即任一 $x_i\in U,r\in R$,则 $f(x_i,r)=V_i$ 对于任一属性 $B\subseteq R$,如果对象 $x_i,x_j\in U,\forall r\in R$ 当且仅当 $f(x_i,r)=f(x_j,r)$,x_i 和 x_j 是不可分辨的,简记为 Ind(B)。

定义 2[7,8]在信息系统 S 中,若 R 可划分为条件属性 C 和决策属性 D,则 $C\cup D=R$,$C\cap D=\varnothing$,具有条件属性和决策属性的信息系统可表示为决策表,记为 $T=(U,R,C,D)$,Ind(C) 的等价类称为条件类,Ind(D) 的等价类称为决策类。

定义 3[7,8]在信息系统 S 中,令 $P,Q\in R$,称知识 Q 以依赖度 $K(0\leqslant K\leqslant 1)$ 依赖于知识,记为 $P\Rightarrow KQ$ 当且仅当:

$$K=r_p(Q)=\mathrm{Card}(\mathrm{Pos}_p(Q))/\mathrm{Card}(U) \quad (1)$$

其中 r 表示依赖度,Card() 表示集合的基数。

定义 4[2]把 D 对某一个条件属性的依赖度称为单一属性依赖度,把 D 对两个或两个以上的条件属性构成的属性集的依赖度称为属性集依赖度,单一属性依赖度和属性集依赖度的求解公式仍为式(1) 的形式。

2 依赖度之差的概念

在决策表中,有时决策属性对多个单一属性或属性集的依赖度会相同,条件属性的重要性就无法确定,这时衡量某一条件属性或条件属性集的重要性不能仅看其依赖度的大小,而要看其依赖度之差的大小。

定义 5[6]令 C 和 D 分别为条件属性和决策属性,设条件属性 C 由 $\{a,b,c,d\}$ 4 个条件属性组成,决策属性 D 由单一属性组成,单一属性(比如属性 a) 的依赖度之差就是用决策属性对条件属性的依赖度 $r_c(D)$ 减去条件属性中去掉单一属性所剩余的属性集 $\{C-\{a\}\}$ 的依赖度。属性集的依赖度之差就是用决策属性对条件属性的依赖度 $r_c(D)$ 减去条件属性中去掉属性集所剩余的属性集的依赖度或所剩余的单一属性的依赖度。对于属性集 D 导出的分类的单一属性或属性集 $C'\subseteq C\subseteq R$ 依赖度之差为:

$$\sigma_{CD}(C')=r_C(D)-r_{C-C'}(D) \quad (2)$$

这表示从集合 C 中去掉某些属性子集 C' 后对对象进行分类时,分类 U/D 的正域将受到怎样的影响。

3 依赖度之差的求解

在粗糙集方法中,不使用事先假定的信息(先验知识),只利用决策表中仅有的数据计算所有的属性是否具有相同的重要性,如果不是,它们在分类能力上有何区别,为找出某些属性的重要性,不是利用依赖度来判断它们的重要性,而是利用依赖度之差来判断它们的重要性。下面给出依赖之差求解方法和算法流程。

3.1 依赖度之差的求解方法

第一步:求解单一属性依赖度和属性集依赖度。

根据公式(1)的定义,设条件属性 C 由 $\{a,b,c,d\}$ 4 个条件属性组成,决策属性 D 由单一属性组成,以求解 D 对单一属性(比如单一属性 a) 的依赖度为例,其求解步骤如下:

(1) 对决策属性 D 进行等价类划分:将决策属性 D 中所有相同值的对象划为同一等价类;

(2) 对单一属性进行等价类划分:将单一属性(比如属性 a) 中所有相同值的对象划为同一等价类;

(3) 求{单一属性等价类} ∩ {决策属性 D 等价类} 的元素数,即 Card();

(4) $K=r_C(D)=$ {元素数 | {单一属性等价类} ∩ {决策属性 D 等价类}} / {对象总数}。

第二步:求解依赖度之差。

根据公式(2) 的定义,以求解单一属性 a 的依赖度之差为例,步骤如下:

(1) 利用第一步求出单一属性 a 的依赖度 $r_a(D)$;

(2) 利用第一步求出属性集 $\{b,c,d\}$ 的依赖度 $r_{\{b,c,d\}}(D)$;

(3) 利用公式(2) 得出单一属性 a 的依赖度之差为:

$$\sigma_{CD}(a)=r_C(D)-r_{\{b,c,d\}}(D)$$

第三步:利用第一步和第二步同样可以计算出属性集的依赖度之差。

3.2 依赖度之差的算法流程

首先给出求解单一属性依赖度的算法属性集合依赖度算法(Attribute Sets Dependability Method, ASDM),求解属性集依赖度时,只要将所要求解的多个属性集以字串相加形式构成新的单一属性,再调用此算

法求解即可。

属性依赖度算法

输入:决策表

nCount 是|条件属性集等价类| ∩ |决策属性集| 的元素数,初始为 0

输出:属性依赖度

属性依赖度(决策表,nCount)

(1) 所有对象按需求解的属性排序 //划分等价类

(2)for each 等价类

if 该类所有对象的决策属性的值均相同 then

nCount = ncount + 该类对象数

endfor

(3) return nCount/对象总数 //返回属性依赖度

其次求解属性依赖度之差的算法为:

输入:属性依赖度

if 条件属于性集 − 所求属性 ≠ ∅ then

所求属性的依赖度之差 = 条件属性集的依赖度 − 条件属性集减去所求属性后剩余的单一属性(属性集) 的依赖度 即为:

σ_{CD}(所求属性) = $r_C(D) - r_{c-\text{所求属性}}(D)$

输出:属性依赖度之差

下面给出具体实例和详细的求解步骤。

4 实　例

某一致性决策表(见表 1)共有 8 个对象,条件属性 C 由属性$\{a,b,c,d\}$ 构成,决策属性为 D(由单一属性构成)。

表 1　某一致性决策表

U	a	b	c	d	D
1	1	0	2	2	0
2	0	1	1	1	0
3	2	0	0	1	1
4	1	1	0	2	2
5	1	0	2	0	1
6	2	2	0	1	1
7	2	1	1	1	2
8	0	1	1	0	1

以求由三个条件属性构成的属性集的依赖度之差为例:

第一步:按照经典方法求解表 1 中任意三个条件属性构成的属性集依赖度,步骤为:

(1) 对决策属性集 D 进行等价类的划分:

$U/D = \{\{1,2\},\{3,5,6,8\},\{4,7\}\}$

(2) 对由三个条件属性构成的属性集进行等价类划分:

$U/\{a,b,c\} = \{\{1,5\},\{2,8\},\{3\},\{4\},\{6\},\{7\}\}$

$U/\{a,b,d\} = \{\{1\},\{2\},\{3\},\{4\},\{5\},\{6\},\{7\},\{8\}\}$

$U/\{b,c,d\} = \{\{1\},\{2,7\},\{3\},\{4\},\{5\},\{6\},\{8\}\}$

$U/\{a,c,d\} = \{\{1\},\{2\},\{3,6\},\{4\},\{5\},\{7\},\{8\}\}$

(3) 求|条件属性集等价类| ∩ |决策属性集等价类| 的元素数。

以求决策属性 D 对条件属性集$\{a,b,c\}$ 的依赖度为例:

$U/\{a,b,c\} \cap \{D\} = \{\{3\},\{4\},\{6\},\{7\}\}$

$\text{Card}(\text{Pos}_{\{a,b,c\}}(D)) = 4$

(4) 决策属性 D 对条件属性集$\{a,b,c\}$ 的依赖度为:$r_{\{a,b,c\}}(D) = r_{C-\{d\}}(D) = 4/8 = 0.5$

第二步:求依赖度之差。

以求条件属性集$\{a,b,c\}$ 的依赖度之差为例。

(1) 利用第一步求出单一属性 d 的依赖度为:

$r_{\{d\}}(D) = 0.25$

(2) 利用第一步求出属性集$\{a,b,c\}$ 的依赖度为:

$r_{\{a,b,c\}}(D) = 0.5$

(3) 利用公式(2) 求得条件属性集$\{a,b,c\}$ 的依赖度之差为:

$\sigma_{CD}(\{a,b,c\}) = r_C(D) - r_d(D) = 1 - 0.25 = 0.75$

用同样的方法求得由任意三个属性构成的属性集的依赖度和依赖度之差,如表 2 所示。

表 2　由任意三个条件属性构成的属性集的依赖度及其依赖度之差

属性	依赖度	依赖度之差
abc	0.5	0.75
abd	1	1
acd	1	0.875
bcd	0.75	1

第三步:用同样的方法求得由任意两个属性构成的属性集的依赖度和依赖度之差,如表 3 所示。

表 3　由任意二个条件属性构成的属性集的依赖度及其依赖度之差

属性	依赖度	依赖度之差
ab	0.5	0.25
ac	0.5	0.25
ad	0.375	0.625
bc	0.375	0.625
bd	0.75	0.5
cd	0.75	0.5

用同样的方法求得单一属性的依赖度和依赖度之差如表 4 所示。

表4 任意一个条件属性的依赖度及其依赖度之差

属性	依赖度	依赖度之差
a	0	0.25
b	0.125	0
c	0	0
d	0.25	0.5

由上面得出所有属性的依赖度和依赖度之差，下面给出验证和分析，判断一个属性的重要性是用依赖度判断准确还是用依赖度之差判断准确。

5 验 证

求属性约简的算法很多，Skowron 提出的分明矩阵[9,10]是一经典方法，它将决策表中关于属性分类的信息浓缩进一个矩阵当中，可以方便地求解属性集的核与约简。决策表(表 1)所构成的分明矩阵如表 5 所示。

表5 决策表(表 1)所构成的分明矩阵

	1	2	3	4	5	6	7
3	acd	abc					
4	bc	acd	abd				
5	d	$abcd$	-	bcd			
6	$abcd$	abc	-	abd	-		
7	$abcd$	a	bc	-	$abcd$	bc	
8	$abcd$	d	-	acd	-	-	ad

该决策表构成的分明函数为：

$f(a,b,c,d)=(a\vee c\vee d)\wedge(a\vee b\vee c)\wedge\cdots\wedge\cdots\wedge(a\vee d)=a\wedge d\wedge(b\vee c)=(a\wedge b\wedge d)\vee(a\wedge c\wedge d)$

从这个最小析取范式可以看出，属性集$\{a,d\}$是条件属性C的D核，是条件属性中最重要的部分，单一属性b和单一属性c可以约去。

6 实例分析

1) 单一属性的重要性。从依赖度看，单一属性a和单一属性c的依赖度为0，可约去，但从依赖度之差看，可知单一属性a和单一属性d重要，如约去单一属性a后，不能把对象2和7划入U/D的类中。如约去单一属性d后，不能把对象2和8划入U/D的类中。单一属性b和单一属性c无关紧要，去掉它们后，分类未产生变化，因此单一属性b和单一属性c可约去。经验证可约去单一属性b或单一属性c，而不是单一属性a和单一属性c。当依赖度K大于等于零且小于1时，单一属性重要性的判断依赖度之差要比依赖度更加准确。

2) 由两个条件属性构成的属性集的重要性。从依赖度看，属性集$\{b,d\}$和属性集$\{c,d\}$的依赖度最大为0.75，应该是最重要的，而属性集$\{a,d\}$的依赖度最小为0.375，应该是最不重要的。从依赖度之差看，属性集$\{a,d\}$和属性集$\{b,c\}$的依赖度最大为0.625。经验证属性集$\{a,d\}$为条件属性集C的D核，是最重要的条件属性集。当依赖度K在大于等于0且小于等于1变化时，由两个属性构成的属性集重要性的判断依赖度之差要比依赖度更加准确。

3) 由三个条件属性构成的属性集的重要性。从依赖度看，属性集$\{a,b,d\}$和属性集$\{a,c,d\}$的依赖度都为1是最重要的。但从依赖度之差看，属性集$\{a,b,d\}$和属性集$\{a,c,d\}$的依赖度之差发生了变化。经验证属性集$\{a,b,d\}$和属性集$\{a,c,d\}$为决策表的两个最小约简。当依赖度K等于1，就不能用依赖度之差来判断。

7 结束语

由前面的讨论可知：在决策表中，当单一属性的依赖度或属性集的依赖度K在大于等于0，小于1之间变化时，不论是判断单一属性的重要性还是判断属性集的重要性，用依赖度之差判断更加准确。而对于比较复杂的决策表属性重要性的判断，还需要进一步的研究，这将是下一步要做的工作。

参考文献：

[1] Jiye L, Deyu L. Information Measures of Roughnes of Knowledge and Significance of Attribute in Rough Set Theory[J]. Journal of Engineering Mathematics, 2000, 17: 106 - 108.

[2] 孟庆全，梅灿华. 一种新的属性集依赖度[J]. 计算机应用，2007, 27(7): 1748 - 1750.

[3] 朱 红. 关于属性间依赖度表示方法的探讨[J]. 计算机工程，2005, 31(1): 174 - 175.

[4] 张文修，吴伟志. 粗糙集理论与方法[M]. 北京：科学出版社，2001: 1 - 25.

[5] 刘 清. Rough 集及 Rough 推理[M]. 北京：科学出版社，2001: 40 - 80.

[6] 吴今培，孙德山. 现代数据分析[M]. 北京：机械工业出版社，2006: 1 - 33.

[7] Pawlak Z. Rough Set[J]. International Journal of Computer information Science, 1982, 11(5): 341 - 350.

[8] Pawlak Z. Rough Set Theory and Its Applications to Data Analysis[J]. Cybernetics and System, 1998, 29(7): 661 - 668.

[9] Pawlak Z, Skowron A. Rudiments of rough sets[J]. Information Sciences, 2007, 177: 3 - 27.

[10] Pawlak Z, Skowron A. Rough sets: some extensions[J]. Information Sciences, 2007, 177: 28 - 40.