◎数据库、信号与信息处理◎

基于聚类率的决策表属性约简

路 静,张 涛,任宏雷
LU Jing, ZHANG Tao, REN Honglei
燕山大学 信息科学与工程学院,河北 秦皇岛 066004
College of Information Engineering, Yanshan University, Qinhuangdao, Hebei 066004, China

LU Jing, ZHANG Tao, REN Honglei. Reduction of attribute in decision table based on clustering rate. Computer Engineering and Applications, 2012, 48(28):135-138.

Abstract: According to classical rough set, when we reduce the attribute of a decision table , it may appear core does not exist that couldn't find a starting point attribute and be unable to reduction problem. Aiming at this issue, this paper proposes a kind of methodbased on clustering rate of the attribute reduction. This method firstly calculates decision table of the discernibility matrix, attribute to differentiate as the foundation, in the attribute of the same differentiate conditions, the clustering rate fixed attribute importance, guarantee the starting point attribute necessity of existence, so it can find the starting point and win the attributes of the decision table attribute reduction. Experimental results show that the proposed method can make gain the starting pointattribute, and by using the methods of obtaining the reduction results maintains high decision accuracy, and it is effective.
Key words: rough set theory; decision table; clustering rate; attribute reduction

摘 要:根据经典粗糙集方法,在对可约简决策表进行属性约简时可能出现核不存在无法找到起点属性从而无法约简的问题。针对该问题,提出了基于聚类率的属性约简方法。计算决策表的区分矩阵,以属性区分度为基础,在属性区分度相同的情况下,利用聚类率修正属性重要度,保证起点属性存在的必然性,从而完成起点属性的求取并获得决策表的属性约简。实验分析表明,方法可以保证可约简决策表中起点属性的计算,且利用该方法获得的约简结果保持了较高的决策准确率,是有效可行的。
关键词:粗糙集;决策表;聚类率;属性约简
文章编号:1002-8331(2012)28-0135-04 **文献标识码:**A **中图分类号:**TP311

1 引言

粗糙集理论是波兰数学家Z.Pawlak在20世纪80年代初首先提出的一种可以分析模糊和不确定问题的数学方法[1]。其主要思想是在保持分类能力不变的前提下,对观察和测量数据直接进行近似分类和关系推理,发现隐含的知识,揭示潜在的规律,通过知识约简,提取有用的特征,使知识的表达更加简明扼要[2]。

属性约简是粗糙集中的一个关键环节。其目的就是删除信息系统中的不必要知识得到必要的属性(集),该属性(集)与全部特征具有相同的分类能力。因此,如果能去除冗余属性,获得最具代表性的属性集,就可以缩减数据量,降低规则提取算法的计算复杂度。

基于启发式属性约简是现在研究的热点,以属性的重要性作为启发规则,从理论上可以得到最小

基金项目:国家自然科学基金(No.60904100,No.61074130);河北省自然科学基金(No.F2011203073)。
作者简介:路静(1985—),女,硕士,主要研究领域:粗糙集,形势概念分析;张涛(1979—),男,讲师,主要研究领域:网格计算,形势概念分析;任宏雷(1989—),女,硕士,主要研究领域:粗糙集,形势概念分析。E-mail:lujinggc@163.com
收稿日期:2012-01-04 **修回日期:**2012-02-20 CNKI出版日期:2012-06-01
DOI:10.3778/j.issn.1002-8331.2012.28.027 http://www.cnki.net/kcms/detail/11.2127.TP.20120601.1458.050.html

的约简[3-5]。其中应用比较广泛的是基于区分矩阵的启发式约简算法[6]。基于区分矩阵求取核属性简单直观,因此叶东毅等人提出的基于区分矩阵的求核方法是很受用的求核方法之一[7],然而其并未考虑到区分矩阵求取核时出现核不存在的问题,导致约简无法顺利进行(得出的属性约简存在冗余),因此该算法存在着不完备性。

针对区分矩阵求取属性约简时会出现核不存在的问题,提出基于属性区分度求取较优的起点属性。同时,算法还考虑了当出现相同区分度的情况下属性选择问题,从区分度和聚类率两个不同层面计算属性的重要性,进行属性约简。实验证明,在求得较优起点属性基础上,算法有效地得到最小属性约简,并且保持了较高的决策准确率。

2 粗糙集理论

2.1 信息系统与决策表

四元组 $S=(U,A,V,f)$ 是一个信息系统,其中:$U=\{x_1,x_2,...x_p\}$ 是非空的有限集合,称为论域。U 中的每个 $x_i(i\leq p)$ 称为一个对象;A 是非空的有限属性集,$A=C\cup D, C\cap D=\varnothing$,其中 C 为条件属性,D 为决策属性;$V=\bigcup_{a\in A}V_a$ 是属性值的集合,V_a 表示属性为 a 的取值集合;映射 $f:U\times A\to V$ 称为信息函数,因此信息系统可以表示为:$\forall a\in A, x\in U, f_a(x)\in V_a$。

2.2 不可区分关系和上下近似集合

对于每一个属性子集 $P\in A$,定义一个二元不可区分关系:

$$IND(P)=\{(x_j,x_k)\,|\,x_j,x_k\in U,\forall a\in P, f_a(x_j)=f_a(x_k)\}$$

关系 $IND(P)$ 是 U 上的等价关系,它是由属性子集 P 对 U 的一个划分。

设任意属性子集 $R\in A$,对象子集 $X\subset IND(D)$,满足:

$$\underline{R}(X)=\bigcup\{Y\,|\,(\forall Y\subset IND(R))\wedge(Y\subseteq X)\}$$

$$\overline{R}(X)=\bigcup\{Y\,|\,(\forall Y\subset IND(R))\wedge(Y\cap X\neq\varnothing)\}$$

分别称它们为 X 的 R 下近似集和 R 上近似集。下近似集是根据特征集 R 判断肯定属于 X 的 U 中元素组成的集合,上近似集是根据特征集 R 判断可能属于 X 的 U 中元素组成的集合。

2.3 区分矩阵和核

区分矩阵包含以下两种情况:

(1)当决策属性 $D=$(类别1,类别2) 即包含两个类别时,将不同的样本分开进行对比,得到区分矩阵 $\boldsymbol{M}$,这样可以节省空间,其任意元素为:

$$M_{mn}=\begin{cases}\{a\,|\,(a\in C)\wedge(f_C(x_i)\neq f_C(x_j),f_D(x_i)\neq f_D(x_j))\}\\0, f_D(x_i)=f_D(x_j)\end{cases}$$

其中 $m=1:\#\{x_i\,|\,f_C(x_i)\in$ 类别1$\}$;$n=1:\text{card}(U)-m$。

(2)当决策属性 $D=$(类别1,类别2,$\cdots$) 即包含多类别时,那么区分矩阵是一个 $card(U)\times card(U)$ 的对称阵,即 $M_{ij}=M_{ji}$,为了降低运算量,通过限定 $i>j$,只需计算 M_{ij},而不必计算与 M_{ij} 对称且相等的 M_{ij},则为:

$$M_{mn}=\begin{cases}\{a\,|\,(a\in C)\wedge(f_C(x_i)\neq f_C(x_j),f_D(x_i)\neq f_D(x_j)),i>j\}\\0, f_D(x_i)=f_D(x_j)\end{cases}$$

由上式可知,$(x_i,x_j)\notin IND(D), IND(D)$,表示相对于决策属性 D 上的一个二元不可区分关系,M_{mn} 是能区别对象 x_i 与 x_j 的所有条件属性组成的集合。

所谓核属性,就是区分矩阵中的单一属性值的集合,去掉该属性会造成决策表的不一致,故其不可或缺,记作:

$$core=\{M_{mn}\,|\,card(M_{mn})=1\}$$

2.4 属性约简

属性约简的目的就是要在保持条件属性相对于决策属性的分类能力不变的前提下,删除其中不必要的或者不重要的条件属性。

假设任意属性子集 $P\in A$,决策属性 D,则

$$POS_P(ind(D))=\bigcup_{\substack{X\subseteq ind(D)\\Y\subseteq ind(P)}}\underline{P}X$$

其中 $ind(D)$ 的 P 正域是 U 中所有根据分类 $IND(P)$ 的信息可以准确地划分到关系 $IND(D)$ 的等价类中去的对象集合。

假设任意 $p\in P$,若存在:

$$POS_P(ind(D))=POS_{P-\{p\}}(ind(D))$$

则认为属性(集) p 是不必要的;否则,称其必要的。

3 计算起点属性的问题

目前,利用区分矩阵求取属性约简时,通常先探索区分矩阵中的单一属性值,作为决策表的核属性,以核属性作为起点,剩余属性按照属性重要度的大小顺序依次加入到核中,直到约简集的决策能力与全部特征的决策能力相同为止。但是,在实际应用时,信息系统中样本数较大,以及单个属性值多样化时等,区分矩阵会出现不存在核属性的情况,因此如何选取起点属性很关键,直接影响对决策表的属性约简。

为了更好地解释无核情况下属性约简出现的问题,如表示1所示决策表。

表1　决策表

U	a	b	c	D
1	1	0	0	0
2	1	1	1	0
3	0	1	2	1
4	2	0	1	2

对应区分矩阵为:

$$\boldsymbol{M}=\begin{pmatrix} 0 & & & \\ ac & 0 & & \\ bcd & abcd & 0 & \\ 0 & abc & ad & 0 \end{pmatrix}$$

由区分矩阵的核定义可知,不存在核属性,因此根据由区分矩阵属性约简方法,以属性出现的频率大小依次加入:

$count(a)=4, count(b)=2, count(c)=4, count(d)=2$

所以得到约简集 $re=\{a,c,d\}$。

而 $pos_{ind(cd)}(ind(D))=pos_{ind(C)}(ind(D))$,即 $re=\{cd\}$,所以依旧存在着冗余属性。

基于以上核不存在的问题,杨宏薇等在粗糙集属性约简方法及其在医疗中的应用研究中提出了统计区分矩阵中最小属性组合出现频率最大的作为起点属性。

表2　简单决策表

U	a	b	c	D
1	1	0	0	0
2	1	1	1	0
3	0	1	2	1
4	2	0	1	2

区分矩阵:

$$\boldsymbol{M}=\begin{pmatrix} 0 & & & \\ 0 & 0 & & \\ ab & ac & 0 & \\ ac & ab & abc & 0 \end{pmatrix}$$

由区分矩阵得出:$count(ab)=3$, $count(ac)=3$。

根据以上理论可得起点属性集为 $core=\{ac\}$ 或者 $core=\{ab\}$,依据剩余属性的属性重要度大小依次加入到核中,因此得到属性约简集为:$core=\{ac\}$ 或者 $core=\{ab\}$。

而 $POS_{ind(a)}(\mathrm{ind}(D))=POS_{ind(c)}(ind(D))$,即属性集 $\{a\}$ 的分类能力与全体属性集 C 的分类能力一致,也就是说该信息系统的一个最小属性约简集为 $re=\{a\}$。利用文献[7]的方法得到的结果始终存在冗余属性,与最小属性约简集存在差异,这意味着最高频度的最小属性组合只能说明其能区分开的对象较高,但并不能说明该组合中的每个属性均对分类起到不可或缺的作用。因此上述理论存在着问题。

4　新的求起点属性的方法

4.1　改进的属性重要度

当生成区分矩阵的时候,记录下每个属性出现的频率。然后基于频率的大小来评估属性区分类别的能力,当一个属性不但重复出现的频率频繁,尤其是与较小的属性组合来区分对象越多时,说明其区分不同类别样本的能力越强,即该属性在信息表中是必要的。

考虑两个极端的情况:在可辨识矩阵不存在属性,说明没有该属性也能全区分样本,即为冗余属性;若在区分矩阵中每个元素中都含有某一属性,即出现频率很高,说明该属性在区分对象时不可或缺。因此依据属性对对象分类的影响即区分度来选择起点属性是正确的。

设属性出现的频率:

$$f(a)=\#\{a \mid a\in M_{m\times n}\} \tag{1}$$

其中 $\#\{.\}$ 表示计算符合条件的个数。

则区分度定义为:

$$dis(a)=f(a)+\sum_{j=1}^{m}\sum_{i=1}^{n}\{\frac{|U|}{|M_{mn}|}\Big| a\in M_{ij}\} \tag{2}$$

起点属性定义为:

$$start=\{a \mid \max\{dis(a)\}\} \tag{3}$$

然而基于属性频率作为属性重要度的测度依据,当出现区分度相同时,如何选择属性呢?

将某个条件属性 a 对样本区域划分 $\{p_j\}$,并与决策属性划分的类别空间 $\{q_j\}$ 对比,找出该条件属性相对于各类别的划分。

属性 a 相对于决策属性划分对象的情况分三种:(1)将属于同一类的完全划分正确即聚类完全;(2)将属于同一类的部分划分正确即不完全聚类;(3)完全不正确的聚类。

前两种情况下的等价类不用再进行细划分了,需要进一步细化的只有第3种,所以如果该属性对前两种情况的划分比例越高,后续需要对第三类中进一步细化的对象越少,则说明该属性对于决策表的分类贡献越大。

这就要求找出条件属性决定的等价类中属于决策属性决定的等价类的所有子集,得出该属性对于划分各个类别的贡献,即前两种情况下的划分能力。如果每个类别的贡献率越大,说明该属性将样

本划分到各个正确类别空间中的能力越大,即该条件属性正确划分空间的能力越接近决策属性的能力,称作聚类率,定义为:

$$r(a)=\frac{1}{n}\{\sum_{j=1}^{n}\frac{|p_i|}{|q_j|}\,|\,p_i\subset ind(a),p_i\subseteq q_j\} \tag{4}$$

其中 $(n=\#\{q_j\,|\,q_j\subseteq \mathrm{ind}(D)\})$ 且 $0\le r(a)\le 1$。

因此改进的属性重要度测量方法:

$$sgf(a)=dis(a)\times r(a) \tag{5}$$

4.2　改进的算法

假设决策表 $S=(U,C,D,f)$,条件属性 $C=\{c_1,c_2,c_3,\cdots,c_k\},k=|C|$,决策属性 $D=\{d\}$,re 是 C 的属性约简集

输入:决策表 S

输出:属性约简 re

步骤1 生成的决策表的区分矩阵 $\boldsymbol{M}_{m\times n}$,首先判断是否存在单一属性值,存在时,将所有单一属性值加入到 $start$ 中;否则根据定义得出起点属性 $start$,如果出现区分度相同的情况,求出属性的 $sgf(c_i)$ 值,选择最大的加入到 $start$ 中。

步骤2 依次求出属性区分度 $dis(c_i),(c_i\in C-core)$,并按大小进行排序。

步骤3 若

$$\frac{|POS_{ind(core)}(IND(D))|}{|U|}=\frac{|POS_{ind(C)}(IND(D))|}{|U|}$$

则 $re=start$。

步骤4 若不等时,$re=start\cup c_i(c_i\in C-core)$,按属性区分度的大小依次加入,如果再出现区分度相同的情况时,求出属性的 $sgf(c_i)$ 值,选择最大的加入到 $start$ 中,直到

$$\frac{|POS_{ind(re)}(IND(D))|}{|U|}=\frac{|POS_{ind(C)}(IND(D))|}{|U|}$$

则 $re=core\cup\{c_i\},(i=1,2\cdots M-|D|-|core|)$。

5　算法分析

在决策信息系统 $S=(U,C,D,f)$ 中,设 $U=\{1,2,3,4,5,6,7\}$,$D=\{e\}$,$A=C\cup D$,决策信息系统由表3给出:

决策表对应的区分矩阵:

$$\boldsymbol{M}=\begin{pmatrix}0&&&&&&\\0&0&&&&&\\abc&bcd&0&&&&\\abc&abd&ac&0&&&\\ac&bd&bc&0&0&&\\0&0&abcd&abd&ab&0&\\0&0&bcd&abcd&bd&0&0\end{pmatrix}$$

表3　决策表信息系统

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

由区分矩阵可得:$dis(a)=40/3$,$dis(b)=50/3$,$dis(c)=40/3$,$dis(d)=34/3$。

区分度最大的为属性 b 和 c,因此可以选择作为启发属性约简的起点属性,即为 $start=\{b\}$。

但是从可辨识矩阵中知道,仅仅根据区分度的公式不能确定加入到核中的属性(不能进一步确定属性 a,c 的重要度),这时可以根据公式(2)加入另一个约束因素 $r(a)$ 来判断,过程如下:

$ind(a)=\{(16),(2357),(4)\}$

$ind(c)=\{(17),(2456),(3)\}$

$ind(D)=\{(1267),(3),(45)\}$

由定义可得出:

$$sgf(b)=(\frac{2}{4}+\frac{1}{2})\times\frac{1}{3}=\frac{1}{6}$$

$$sgf(c)=(\frac{2}{4}+\frac{1}{1})\times\frac{1}{3}=\frac{1}{2}$$

显然 c 的重要度最大。从直观上看,属性 a,c 都可以分开三个样本,但是属性 c 将属于一个类别的样本完全区分开,故其分类能力较高。

因此将 c 加入到约简集中,即为

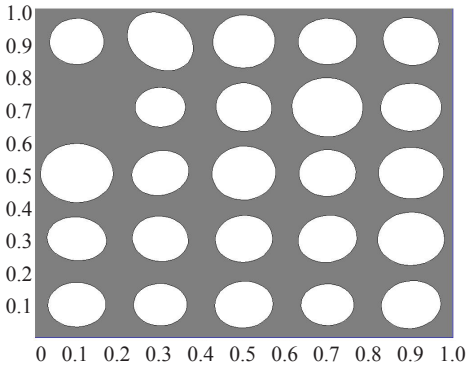
$$re=\{b,c\},POS_{ind(bc)}(ind(D))=\{1,2,3,4,5,6,7\}$$

$re=\{b,c\}$ 即为该信息系统的最小约简集。

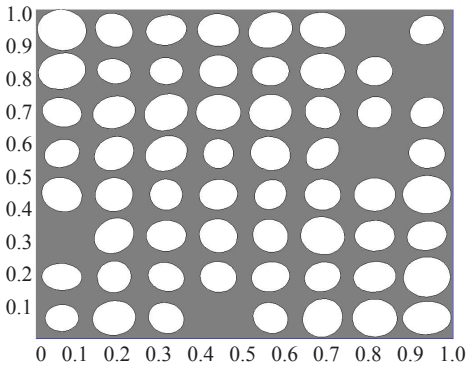
6　结论

核属性的计算问题一直是粗糙集研究领域的热点问题,在以往的研究中,已有核属性的求取方法曾引起学者们的广泛讨论。本文针对已有的基于区分矩阵求取属性约简时会出现核不存在的问题,分析了文献[8]提出的求取起点属性的不完全性,提出利用属性区分度求取起点属性,在此基础上,提出了一个改进的属性重要度约简算法,即当出现相同区分度的情况下,通过加入另外一个约束因子属性聚类率 $r(a)$ 来确定要选择加入到约简集中的属性,解决了区分度相同时属性选择的问题,引导约简过程趋于最优化。

(下转233页)



(a)孔洞个数限定为24个



(b)限定孔径尺寸

图5　在限制条件下情形3对应的最优单胞结构

者大小，都有可能获得具有较高热传导性能的单胞结构。

参考文献：

[1] Sigmund O.Materials with prescribed constitutive parameters: An inverse homogenization problem[J].International Journal of Solids and Structures，1994，31(17)：2313-2329.

[2] Sigmund O.Tailoring materials with prescribed elastic properties[J].Mechanics of Materials，1995，20：351-368.

[3] 刘书田，郑新广，程耿东.特定弹性性能材料的细观结构设计优化[J].复合材料学报，2001，18(2)：124-127.

[4] 曹志远，程红梅.非匀质材料板的微结构优化设计[J].应用力学学报，2005，22(3)：373-376.

[5] 刘书田，曹先凡.零膨胀材料设计与模拟验证[J].复合材料学报，2005，22(1)：126-132.

[6] Gu S，Lu T J，Evans A G.On the design of two-dimensional cellular metals for combined heat dissipation and structural load apacity[J].Inter J Heat and Mass Transfer，2001，44：2163-2175.

[7] Seepersad C C，Dempsey B M，Allen J K，et al.Design of multifunctional honeycomb materials[J].AIAA J，2004，42(5)：1025-1033.

[8] Wang Bo，Cheng Gengdong.Design of cellular structures for optimum efficiency of heat dissipation[J].Structural and Multi-disciplinary Optimization，2005，30：447-458.

[9] 张志峰，陈浩然，李煊，等.先进复合材料格栅圆柱壳优化设计的混合遗传算法[J].复合材料学报，2005，22(2)：166-171.

[10] 张卫红，汪雷，孙士平.基于导热性能的复合材料微结构拓扑优化[J].航空学报，2006，27(6)：1229-1233.

[11] Grimvall G.Bounds on transport，elastic and thermal expansion parameters in eutectic two-phase materials[J].J Phys C：Solid State Phys，1984，17：3545-3549.

[12] 李友云，崔俊芝.具有大量椭圆颗粒/孔洞随机分布区域的计算机模拟及其改进三角形自动网格生成算法[J].中国计算力学学报，2004，21：540-545.

[13] Van Mier J G M，Van Vliet M R A.Influence of microstructure of concrete on size/scale effects in tensile fracture[J].Engineering Fracture Mechanics，2003，70：2281-2306.

[14] Yu Y，Cui J Z，Han F.An effective computer generation method for the composites with random distribution of large numbers of heterogeneous grains[J].Composites Science and Technology，2008，68：2543-2550.

[15] Cui J Z，Yang H Y.A dual coupled method of boundary value problems of PDE with coefficients of small period[J].Intern J Comp Math，1996，14：159-174.

(上接138页)

参考文献：

[1] Pawlak Z.Rough sets[J].Int J of Computer and Information Science，1982，11(5)：341-356.

[2] 周献中，李华雄.广义约简、核与分辨矩阵[J].控制与决策，2010，25(10)：1507-1512.

[3] Geng Z Q，Zhu Q X.A new rough set-based heuristic algofitlm for attribute reduct[C]//Proceedings of the World Congress on Intelligent Control And Automation. Piscataway，NJ，USA：IEEE，2006：3085-3089.

[4] Liang J K，Zhang Y，Qu Y B.A heuristic algorithm of attribute reduction in rough set[C]//Proceedings of the International Conference on Machine Learning and Cybernetics.Piscataway，NJ，USA：IEEE，2005：3140-3142.

[5] 蒙祖强，史忠植.一种新的启发式知识约简算法[J].小型微型计算机系统，2009，30(7)：1249-1255.

[6] 王立宏，吴耿峰.基于并行协同进化的属性约简[J].模式识别与人工智能，2003，26(5)：630-635.

[7] 叶东毅，陈绍炯.一个新的差别矩阵及其求核方法[J].电子学报，2002，30(7)：1086-1088.

[8] 杨宏薇，何中市.粗糙集属性约简方法及其在医疗中的应用研究[J].计算机工程与应用，2010，46(25)：138-140.