区间值信息系统的决策属性约简

刘鹏惠[1],陈子春[1,2],秦克云[2]
LIU Peng-hui[1],CHEN Zi-chun[1,2],QIN Ke-yun[2]

1.西华大学 数学与计算机学院,成都 610039
2.西南交通大学 理学院,成都 610013
1.School of Mathematics and Computer,Xihua University,Chengdu 610039,China
2.Department of Mathematics,Southwest Jiaotong University,Chengdu 610013,China
E-mail:czclph@163.com

LIU Peng-hui,CHEN Zi-chun,QIN Ke-yun.Decision attribute reduction of interval-valued information system.Computer Engineering and Applications,2009,45(28):148-150.

Abstract: A new variable precision tolerance relation is defined in interval-valued information system by similarity grade of attribute's interval-value,based on the variable precision tolerance relation the problems of decision attribute reduction and relative decision attribute reduction in interval-valued information system are studied,from which an approach for decision attribute reduction and relative decision attribute reduction is provided in interval-valued information system.

Key words: interval-valued information system;similarity level;variable precision tolerance relation;decision attribute reduction

摘 要:借助于属性区间值的相似程度在区间值信息系统上定义了一种具有变精度的相容关系,讨论了在这种相容关系下决策区间值信息系统的决策属性约简与决策属性相对约简,并得到了求决策属性约简与决策属性相对约简的具体操作方法。

关键词:区间值信息系统;相似水平;变精度相容关系;决策属性约简
DOI:10.3778/j.issn.1002-8331.2009.28.044 **文章编号:**1002-8331(2009)28-0148-03 **文献标识码:**A **中图分类号:**TP18

1 引言

从信息系统中发现潜在的、非平凡的有用知识是处理和利用海量信息的重要前提,其本质是根据实际需要按属性特征将对象进行分类后,利用非平凡手段找出不影响分类的最小属性集,以使分类知识的表示简化,而又不丢失任何信息。波兰数学家 Pawlak 于 1982 年提出的粗糙集理论[1]是经典集合理论的推广,是处理模糊和不确定知识的有用数学工具[1-2],经过 20 多年的研究与发展,已经在理论和实际应用上取得了长足的进展,特别是由于 20 世纪 80 年代末和 90 年代初在知识发现等领域的成功应用而受到了国际上广泛关注。目前,粗糙集理论已经在人工智能、知识与数据发现、模式识别与分类、故障检测等方面得到了广泛的应用[3-7]。Pawlak 粗糙集模型是基于完备信息系统上的不可分辨关系,对无法用已知知识描述的对象集给出上、下近似。Pawlak 粗糙集模型也可以推广到非等价关系(如相容关系、优势关系、偏序关系等)[8-13],这种推广可以用来处理不完备信息系统上的知识获取与推理。但是,传统的粗糙集方法主要处理的是离散属性值,而在很多实际问题中,由于问题的复杂性,信息系统中的属性值往往是连续的或者真实的数据只知落在某个区间,从而对连续属性值的处理一般需要离散化,其本质是将属性值域划分成一些离散化区间,因此,研究区间值信息系统上知识的获取与约简就很有必要。

属性约简是知识发现的重要课题,也是粗糙集理论的核心问题之一。众所周知,知识库中描述知识的属性并不是同等重要的,即每一种属性在信息系统中所起的作用是不同的,有些属性是绝对必要的,有些属性是相对必要的,有些属性是完全不必要的。所谓属性约简就是在保持知识库分类能力不变条件下,删除其中不相关或不重要的属性[6-7,10]。但是,这种属性是否必要完全依赖于信息系统上的知识或关系的定义,针对不同的实际问题,可以定义不同的知识或关系,这时属性是否必要也随之发生改变。就区间值信息系统,该文借助于属性区间值的相似程度在区间值信息系统上定义了一种具有变精度的相容关系,讨论了在这种相容关系下决策区间值信息系统的决策属性约简与决策属性相对约简,并得到了求决策属性约简与决策属性相对约简的具体操作方法。

2 区间值信息系统上的变精度相容关系

信息系统是一三元组 $K=(U,A,F)$,其中 $U=\{x_1,x_2,\cdots,x_n\}$为有限对象集,称为论域,每个 $x_i(i\leqslant n)$称为一个对象;$A=\{a_1,a_2,\cdots,a_m\}$为有限属性集,每个 $a_i(i\leqslant m)$称为一个属性;$F=\{f_k:k\leqslant m\}$为对象属性值映射集,其中 $f_k:U\rightarrow V_k(k\leqslant m)$,$V_k$ 是属性 a_k 的有

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60474022)。
作者简介:刘鹏惠(1973-),女,讲师,研究方向为智能信息处理;陈子春(1970-),男,博士研究生,讲师,研究方向为智能信息处理;秦克云(1962-),男,教授,博士生导师。
收稿日期:2008-06-23 **修回日期:**2008-09-27

限值域，即$\forall x_i\in U, f_k(x_i)\in V_k$。若$\forall a_k\in A$，$\forall x_i\in U$，$f_k(x_i)$都是唯一确定的，则称信息系统$K$是完备的，否则便称它是不完备的；若$f_k(x_i)=[l_i^k, r_i^k]\in I([0,1])$，$l_i^k<r_i^k$，则称$K$是区间值信息系统，其中$I([0,1])$表示区间$[0,1]$上的闭区间全体。

由于区间值信息系统上对象的属性取值是区间，而两个区间的相似程度取决于其相交部分的区间长度，在区间值信息系统上给出下面的定义。

定义 1 设$K=(U,A,F)$是区间值信息系统，$a_k\in A$，$x_i,x_j\in U$，$f_k(x_i)=[l_i^k, r_i^k]$，$f_k(x_j)=[l_j^k, r_j^k]$，称$S_{ij}^k=\dfrac{\left|[l_i^k, r_i^k]\cap[l_j^k, r_j^k]\right|}{\left|[l_i^k, r_i^k]\cup[l_j^k, r_j^k]\right|}$是对象$x_i$、$x_j$相对于属性$a_k$的相似度。其中$|\cdot|$在这里表示闭区间的长度，并规定空集与单点集的区间长是0。

对$a_k\in A$，$B\subseteq A$，$\alpha\in(0,1]$，在区间值信息系统上定义相似关系：$T_B^\alpha=\{(x_i,x_j)\in U\times U:\forall a_k\in B, S_{ij}^k\geqslant\alpha\}$。并记$T_B^\alpha(x_i)=\{x_j\in U:(x_i,x_j)\in T_B^\alpha\}$。

注 1 S_{ij}^k表示对象x_i与x_j对于属性a_k的属性区间值的相似程度，S_{ij}^k越大，说明x_i与x_j相对于a_k的属性区间值的相似程度越高，且由定义易知$S_{ij}^k=0\Leftrightarrow[l_i^k, r_i^k]\cap[l_j^k, r_j^k]=\varnothing$或$[l_i^k, r_i^k]$与$[l_j^k, r_j^k]$仅在一点取相同值；$S_{ij}^k=1\Leftrightarrow[l_i^k, r_i^k]=[l_j^k, r_j^k]$。

注 2 $(x_i,x_j)\in T_B^\alpha$表示对象x_i与x_j对于B中任一个属性的相似程度都不低于给定的阀值α。

定理 1 设$K=(U,A,F)$是区间值信息系统，T_B^α是如上定义的二元关系，则有

（1）T_B^α是U上的相容关系，即T_B^α是自反、对称的；

（2）若$B_1\subseteq B_2\subseteq A$，则对$\alpha\in(0,1]$有，$T_A^\alpha\subseteq T_{B_2}^\alpha\subseteq T_{B_1}^\alpha$；$T_A^\alpha(x_i)\subseteq T_{B_2}^\alpha(x_i)\subseteq T_{B_1}^\alpha(x_i)$；

（3）对任意$B\subseteq A$，$T_B^\alpha=\bigcap_{a_k\in B}T_{\{a_k\}}^\alpha$。

证明 由T_B^α的定义进行验证即得。

以下为方便，称T_B^α是区间值信息系统上的变精度相容关系，阀值α称为相似水平，若$(x_i,x_j)\in T_B^\alpha$，则称对象x_i与x_j关于B在相似水平α下相容。

定义 2[9] 称(U,A,F,d,g)是决策区间值信息系统，若(U,A,F)是区间值信息系统，$d\notin A$是决策属性；$g:U\to V_d$是U与d的关系集，$V_d=\{1,2,\cdots,r\}$是g的有限值域。

记$R_d=\{(x_i,x_j)\in U\times U:g(x_i)=g(x_j)\}$，$[x_i]_d=\{x_j\in U:(x_i,x_j)\in R_d\}$，则易知$R_d$是$U$上的等价关系。

定义 3 设(U,A,F,d,g)是决策区间值信息系统，若$R_A^\alpha\subseteq R_d$，则称决策区间值信息系统是α水平协调的，否则称是α水平不协调的。

例 1 考虑表1所示决策区间值信息系统。

可以计算出对象间的相似度S_{ij}^k如下：

$S_{ii}^k=1, i=1,2,\cdots,6, k=1,2,3,4$

表 1 决策区间值信息系统

	a_1	a_2	a_3	a_4	d
x_1	[0.3,0.6]	[0.1,0.4]	[0.2,0.6]	[0.5,0.8]	1
x_2	[0.6,0.9]	[0.3,0.5]	[0.4,0.7]	[0.6,0.8]	2
x_3	[0.4,0.6]	[0.2,0.5]	[0.2,0.5]	[0.4,0.9]	1
x_4	[0.4,0.7]	[0.2,0.4]	[0.3,0.6]	[0.5,0.6]	3
x_5	[0.5,0.8]	[0.3,0.4]	[0.3,0.7]	[0.5,0.7]	3
x_6	[0.6,0.8]	[0.3,0.4]	[0.4,0.8]	[0.5,0.8]	2

$S_{12}^1=0, S_{12}^2=0.25, S_{12}^3=0.4, S_{12}^4\doteq0.67$

$S_{13}^1\doteq0.67, S_{13}^2=0.5, S_{13}^3=0.75, S_{13}^4=0.6$

$S_{14}^1=0.5, S_{14}^2\doteq0.67, S_{14}^3=0.75, S_{14}^4\doteq0.33$

$S_{15}^1=0.2, S_{15}^2\doteq0.33, S_{15}^3=0.6, S_{15}^4\doteq0.67$

$S_{16}^1=0, S_{16}^2\doteq0.33, S_{16}^3\doteq0.33, S_{16}^4=1$

$S_{23}^1=0, S_{23}^2\doteq0.67, S_{23}^3=0.2, S_{23}^4=0.4$

$S_{24}^1=0.2, S_{24}^2\doteq0.33, S_{24}^3=0.5, S_{24}^4=0$

$S_{25}^1=0.5, S_{25}^2=0.5, S_{25}^3=0.75, S_{25}^4\doteq0.33$

$S_{26}^1\doteq0.67, S_{26}^2=0.5, S_{26}^3=0.75, S_{26}^4\doteq0.67$

$S_{34}^1\doteq0.67, S_{34}^2\doteq0.67, S_{34}^3=0.5, S_{34}^4=0.2$

$S_{35}^1=0.25, S_{35}^2\doteq0.33, S_{35}^3=0.4, S_{35}^4=0.4$

$S_{36}^1=0, S_{36}^2\doteq0.33, S_{36}^3\doteq0.17, S_{36}^4=0.6$

$S_{45}^1=0.5, S_{45}^2=0.5, S_{45}^3=0.75, S_{45}^4=0.5$

$S_{46}^1=0.25, S_{46}^2=0.5, S_{46}^3=0.4, S_{46}^4\doteq0.33$

$S_{56}^1\doteq0.67, S_{56}^2=1, S_{56}^3=0.6, S_{56}^4\doteq0.67$

取相似水平$\alpha=0.5$，则有：

$T_A^\alpha(x_1)=T_A^\alpha(x_3)=\{x_1,x_3\}, T_A^\alpha(x_2)=\{x_2,x_6\}, T_A^\alpha(x_4)=\{x_4,x_5\},$

$T_A^\alpha(x_5)=\{x_4,x_5,x_6\}, T_A^\alpha(x_6)=\{x_2,x_5,x_6\}$

又$[x_1]_d=[x_3]_d=\{x_1,x_3\}$，$[x_2]_d=[x_6]_d=\{x_2,x_6\}$，$[x_4]_d=[x_5]_d=\{x_4,x_5\}$，所以易知该区间值信息系统是$\alpha$水平不协调的。

3 区间值信息系统的决策属性约简

隐藏在决策信息系统中的规则可以形式地表示为：$\wedge(a,v)\to(d,w)$，其中$a\in A$，$v\in V_a$，$w\in V_d$。任何决策区间值信息系统在形式上都可以看成是决策规则的集合，如例1所示决策区间值信息系统，它可以看成是下列六条决策规则的集合：

$(a_1,[0.3,0.6])\wedge(a_2,[0.1,0.4])\wedge(a_3,[0.2,0.6])\wedge(a_4,[0.5,0.8])\to(d,1)$　//x_1支持

$(a_1,[0.6,0.9])\wedge(a_2,[0.3,0.5])\wedge(a_3,[0.4,0.7])\wedge(a_4,[0.6,0.8])\to(d,2)$　//x_2支持

$(a_1,[0.4,0.6])\wedge(a_2,[0.2,0.5])\wedge(a_3,[0.2,0.5])\wedge(a_4,[0.4,0.9])\to(d,1)$　//x_3支持

$(a_1,[0.4,0.7])\wedge(a_2,[0.2,0.4])\wedge(a_3,[0.3,0.6])\wedge(a_4,[0.5,0.6])\to(d,3)$　//x_4支持

$(a_1,[0.5,0.8])\wedge(a_2,[0.3,0.4])\wedge(a_3,[0.3,0.7])\wedge(a_4,[0.5,0.7])\to(d,3)$　//x_5支持

$(a_1,[0.6,0.8])\wedge(a_2,[0.3,0.4])\wedge(a_3,[0.4,0.8])\wedge(a_4,[0.5,0.8])\rightarrow(d,2)$ //x_6 支持

从形成决策规则的角度来看，需要消除一些不必要的条件属性，即需要去寻找保持所有对象的决策规则不变的最小属性集。

定义 4 设(U,A,F,d,g)是决策区间值信息系统，α是相似水平，$B\subseteq A$。称$\partial_B^{\alpha}:U\rightarrow P(V_d)$，$\partial_B^{\alpha}(x)=\{g(y):y\in T_B^{\alpha}(x)\}$是决策区间值信息系统的$\alpha$决策函数。

定义 5 设(U,A,F,d,g)是决策区间值信息系统，α是相似水平，$B\subseteq A$。若$\partial_B^{\alpha}=\partial_A^{\alpha}$，则称$B$是决策区间值信息系统的$\alpha$决策协调集。若进一步$\forall b\in B$，$\partial_{B-\{b\}}^{\alpha}\neq\partial_A^{\alpha}$，则称$B$是决策区间值信息系统的$\alpha$决策约简。

由定义，决策区间值信息系统的α决策约简是保持α决策函数不变的最小属性集，也即是保持所有对象的决策规则不变的最小属性集。容易证明决策区间值信息系统的α决策约简总是存在的。

定义 6 设(U,A,F,d,g)是决策区间值信息系统，α是相似水平，对$x_i,x_j\in U$，记

$$D_{ij}^{\alpha}=\begin{cases}\{a_k\in A;S_{ij}^{k}<\alpha\}, & g(x_j)\notin\partial_A^{\alpha}(x_i)\\ \varnothing, & g(x_j)\in\partial_A^{\alpha}(x_i)\end{cases},\quad \boldsymbol{D}^{\alpha}=(D_{ij}^{\alpha}:i,j\leqslant n)$$

称$\boldsymbol{D}^{\alpha}$是决策区间值信息系统的α水平区分矩阵。

注意到T_A^{α}是自反对称关系，所以对任意$x_i,x_j\in U$及相容水平α都有$D_{ii}^{\alpha}=\varnothing$，$D_{ij}^{\alpha}=D_{ji}^{\alpha}$，即决策区间值信息系统的$\alpha$水平区分矩阵$\boldsymbol{D}^{\alpha}$是主对角线上元素全为$\varnothing$的对称矩阵。

定理 2 设(U,A,F,d,g)是决策区间值信息系统，α是相似水平，$B\subseteq A$，记$D_0^{\alpha}=\{D_{ij}^{\alpha}\neq\varnothing;x_i,x_j\in U\}$。则$\forall D_{ij}^{\alpha}\in D_0$，$B\cap D_{ij}^{\alpha}\neq\varnothing$当且仅当$\partial_B^{\alpha}=\partial_A^{\alpha}$。

证明 由于$\forall x\in U$，当$B\subseteq A$时，有$T_A^{\alpha}(x)\subseteq T_B^{\alpha}(x)$成立，由定义知，$\forall x\in U$，总有$\partial_A^{\alpha}(x)\subseteq\partial_B^{\alpha}(x)$成立。故只需证明：$\forall D_{ij}^{\alpha}\in D_0$，$B\cap D_{ij}^{\alpha}\neq\varnothing$当且仅当$\partial_B^{\alpha}\subseteq\partial_A^{\alpha}$。

“$\Rightarrow$”，$\forall x_i\in U$，$x_j\in T_B^{\alpha}(x_i)$，即$g(x_j)\in\partial_B^{\alpha}(x_i)$。若$g(x_j)\notin\partial_A^{\alpha}(x_i)$，则$x_j\notin T_A^{\alpha}(x_i)$，于是$D_{ij}^{\alpha}\neq\varnothing$，从而由题条件知，$B\cap D_{ij}^{\alpha}\neq\varnothing$。这表明存在$a_k\in B$，使$S_{ij}^{k}<\alpha$，即$x_i$与$x_j$关于$B$在相似水平$\alpha$下不相容，这就与$x_j\in T_B^{\alpha}(x_i)$矛盾。因此$x_j\in T_A^{\alpha}(x_i)$成立，从而$\partial_B^{\alpha}\subseteq\partial_A^{\alpha}$成立。

“$\Leftarrow$”，$\forall D_{ij}^{\alpha}\in D_0$，由$D_{ij}^{\alpha}\neq\varnothing$知，$g(x_j)\notin\partial_A^{\alpha}(x_i)$。于是由题设$\partial_B^{\alpha}\subseteq\partial_A^{\alpha}$就有$g(x_j)\notin\partial_B^{\alpha}(x_i)$，即$x_j\notin T_B^{\alpha}(x_i)$。因此存在$a_k\in B$，使$S_{ij}^{k}<\alpha$，即存在$a_k\in B$，且$a_k\in D_{ij}^{\alpha}$，从而$B\cap D_{ij}^{\alpha}\neq\varnothing$成立。

由定理 2 知，寻找决策区间值信息系统的α决策约简，实际上是在D_0^{α}中寻找满足条件$B\cap D_{ij}^{\alpha}\neq\varnothing$的最小集合$B$。为此，记$\Delta^{\alpha}=\wedge(\vee D_{ij}^{\alpha})$，其中$(x_i,x_j)\in U\times U$，$D_{ij}^{\alpha}\in D_0^{\alpha}$，称$\Delta^{\alpha}$是决策区间值信息系统的$\alpha$区分函数。

定理 3 设(U,A,F,d,g)是决策区间值信息系统，α是相似水平，$B\subseteq A$，则B是α决策约简当且仅当$\wedge B$是α区分函数Δ^{α}的极小析取范式中的一个合取式。

证明 “$\Rightarrow$”，注意到$\wedge B$是α区分函数Δ^{α}的极小析取范式中的一个合取式等价于$\forall a_k\in B$，存在$(x_i,x_j)\in U\times U$，$D_{ij}^{\alpha}\in D_0^{\alpha}$，使$B\cap D_{ij}^{\alpha}=\{a_k\}$。因此，若$B$是$\alpha$决策约简，而$\wedge B$不是$\alpha$区分函数$\Delta^{\alpha}$的极小析取范式中的一个合取式，则$\forall(x_i,x_j)\in U\times U$，$D_{ij}^{\alpha}\in D_0^{\alpha}$，就有$card(B\cap D_{ij}^{\alpha})\geqslant 2$。取$B'=B-\{a_k\}$，则$\forall(x_i,x_j)\in U\times U$，$D_{ij}^{\alpha}\in D_0^{\alpha}$，$B'\cap D_{ij}^{\alpha}\neq\varnothing$，于是由定理 2 知$B'=B-\{a_k\}$是$\alpha$决策协调集，这就与$B$是$\alpha$决策约简矛盾。因此，$\wedge B$是$\alpha$区分函数$\Delta^{\alpha}$的极小析取范式中的一个合取式。

“$\Leftarrow$”，若$\wedge B$是α区分函数Δ^{α}的极小析取范式中的一个合取式，则$\forall a_k\in B$，存在$(x_i,x_j)\in U\times U$，$D_{ij}^{\alpha}\in D_0^{\alpha}$，使$B\cap D_{ij}^{\alpha}=\{a_k\}$，于是由定理 2 知$B$是$\alpha$决策协调集，又易知$B'=B-\{a_k\}$不是$\alpha$决策协调集，因而$B$是$\alpha$决策约简。

定理 3 给出了求决策区间值信息系统所有α决策约简的方法。

例 2 继续考虑例 1 中的决策区间值信息系统，取$\alpha=0.5$，则有：

$\partial_A^{\alpha}(x_1)=\partial_A^{\alpha}(x_3)=\{1\}$，$\partial_A^{\alpha}(x_2)=\{2\}$，$\partial_A^{\alpha}(x_4)=\{3\}$，

$\partial_A^{\alpha}(x_5)=\partial_A^{\alpha}(x_6)=\{2,3\}$

$$\boldsymbol{D}^{\alpha}=\begin{pmatrix}\varnothing & & & & & \\ a_1,a_2,a_3 & \varnothing & & & & \\ \varnothing & a_1,a_3,a_4 & \varnothing & & & \\ a_4 & a_1,a_2,a_4 & a_4 & \varnothing & & \\ a_1,a_2 & a_4 & A & \varnothing & \varnothing & \\ a_1,a_2,a_3 & \varnothing & a_1,a_2,a_3 & a_1,a_3,a_4 & \varnothing & \varnothing\end{pmatrix}$$

$\Delta^{\alpha}=(a_1\vee a_2\vee a_3)\wedge a_4\wedge(a_1\vee a_2)\wedge$
$(a_1\vee a_3\vee a_4)\wedge(a_1\vee a_2\vee a_4)=(a_4\wedge a_1)\vee(a_4\wedge a_2)$

由定理 2 知，$\{a_1,a_4\}$与$\{a_2,a_4\}$是决策区间值信息系统的α决策约简。

决策区间值信息系统的α决策约简是保持所有对象的决策规则不变的最小属性集，但有时也需要寻找保持部分对象的决策规则不变的最小属性集，为此，下面给出相对α决策约简的概念。

定义 7 设(U,A,F,d,g)是决策区间值信息系统，α是相似水平，$B\subseteq A$。$x_i\in U$，若B是满足条件$\partial_B^{\alpha}(x_i)=\partial_A^{\alpha}(x_i)$的最小属性集，则称$B$是对象$x_i$的$\alpha$相对决策约简，并称

$\Delta^{\alpha}(x_i)=\wedge(\bigvee_j D_{ij}^{\alpha})$，其中$x_j\in U$，$D_{ij}^{\alpha}\in D_0^{\alpha}$

是对象x_i的α相对区分函数。

定理 4 设(U,A,F,d,g)是决策区间值信息系统，α是相似水平，$B\subseteq A$，$x_i\in U$，则B是对象x_i的α相对决策约简当且仅当$\wedge B$是对象x_i的α相对区分函数$\Delta^{\alpha}(x_i)$的极小析取范式中的一个合取式。

证明 类似于定理 3 的证明，略。

例 3 续例 2，在相似水平$\alpha=0.5$时，可以计算得：

$\Delta^{\alpha}(x_1)=(\vee D_{12})\wedge(\vee D_{14})\wedge(\vee D_{15})\wedge(\vee D_{16})=$
$(a_1\wedge a_4)\vee(a_2\wedge a_4)$

于是$\{a_1,a_4\}$与$\{a_2,a_4\}$是x_1的α相对决策约简。

类似可计算出：

（下转 229 页）