文章编号:1005－0523(2005)02－0100－04

不完备信息系统的基于集对分析粗糙集模型

邓毅雄,黄兆华

(华东交通大学 信息工程学院,江西 南昌 330013)

摘要:粗糙集理论作为一种研究不确定性信息系统的教学工具,在数据挖掘和知识发现等方面得到了广泛应用,由于经典的粗糙集理论是基于等价关系的,它在不完备信息系统中的应用受到限制.近来不少研究成果将经典粗糙集中的等价关系的条件放宽,使粗糙集理论的应用更加广泛.文献[2]利用集对分析研究了不完备信息系统的粗糙集模型,本文注意到空值在信息系统中也提供了一定的知识,建立了一个比[2]更一般性的模型,并对此模型的基本性质、算法等进行了讨论.

关 键 词:不完备信息系统;集对分析;Rough 集;相似关系

中图分类号:TP311　　**文献标识码**:A

0 引 言

粗糙集理论作为一种研究不精确性和不确定性信息系统的数学工具,自 1982 年波兰学者 Pawlak 提出至今,在知识发现、模式识别、决策分析、数据挖掘等领域得到广泛的应用.然而在经典的粗糙集理论中,论域上讨论的关系是等价关系(即满足自反性、对称性和传递性的关系),并在此基础上得到对象集的下、上近似集,但由于属性值的遗漏(Missing),如记载的原因,而将某属性值丢失;或者由于受到目前实验条件的限制某属性值测量困难,导致属性值无从知道(Unknown)等原因,系统中某些属性的属性值为空值(Null),这样的信息系统是不完备的.对于不完备信息系统,论域上的二元关系不一定满足等价性,所以目前基于粗糙集的研究往往将等价关系的要求削减,比如二元关系满足自反性,但可能不满足对称性和传递性.

目前处理不完备信息系统的方法主要集中在两个方向,一是先对系统进行预处理,补全残缺的数据,然后再类似完备性的方法讨论;另一是直接在不完备系统上建立对象之间的某种关系(这种关系往往不是等价关系,如相似关系),将原经典理论扩充后进行讨论.文献[2~3]利用集对分析思想直接对不完备信息系统进行了讨论,建立基于集对的粗糙集模型,并对模型进行了初步分析,得到了一些有益的结果.

本文注意到即使属性值是空值,也能在信息系统中起到一定的作用,提供一定的知识,但空值又不像确定的属性值那样提供完整的知识,所以应该按一定的比值将空值出现的信息纳入到知识的发现过程中,由此我们得到不完备信息系统的更一般的粗糙集模型.

1 集对分析和不完备信息系统的相似关系

集对分析(SPA)是我国学者赵克勤于 1989 年提出用于研究两个集合相互关系的理论.给定两个集合 A 和 B,并设这两个集合组成集对 $H=(A,B)$,在某个具体的问题背景(W)下,集对 H 有 N 个特性,其中:有 S 个为集对 H 中 A 和 B 所共有;P 个为 A 和 B 相对立;F 个为 A 和 B 既不共有也不对立,

收稿日期:2004－06－10

作者简介:邓毅雄(1963－),男,江西新干人,教授.

则比值 S/N 为 A 和 B 在问题 W 下的同一度；F/N 为 A 和 B 在问题 W 下的差异度；P/N 为 A 和 B 在问题 W 下的对立度．并用 $u_w(A,B)=\frac{S}{N}+\frac{F}{N}i+\frac{P}{N}j$ 表示 A 和 B 的关系，u 为 A，B 的联系度，简记为 $u(A,B)=a+bi+cj$．显然 $0\leq a,b,c\leq 1$，$a+b+c=1$.

对信息系统 $S=(U,A)$，若至少存在一个属性 $a\in A$，使 Va 含有空值(用 * 表示空值)，则称 S 为一个不完备信息系统，否则是完备信息系统．

对 $B\subseteq A$，$|B|=n$，$\forall x,y\in U$，记

$$u_B(x,y)=\frac{s}{n}+\frac{f}{n}i+\frac{p}{n}j \tag{1}$$

其中 s 为 x 和 y 在 B 中取值相同的属性个数；p 为 x 和 y 在 B 中取值相异的属性个数；f 为 x 和 y 在 B 中取值不明确的属性个数，包括三种情况：① $a(x)=^*$ and $a(y)=^*$；② $(a(x)=^*$ and $a(y)\neq^*$；③ $a(x)\neq^*$ and $a(y)=^*$.

简记为 $u_B(x,y)=a+bi+cj$，特别，当 $B=A$ 时，记 $u_A(x,y)=u(x,y)$.

如果只考虑对象 x 和 y 之间具有相同属性值的属性的个数，则当 $B\subseteq A$，$0<\alpha\leq 1$，x 与 y 的集对 α 相似关系定义为[2]：

$$\mathrm{SIM}(B)=\{(x,y)\in U\times U\mid u_B(x,y)=a+bi+cj,\ a\geq\alpha\} \tag{2}$$

我们注意到，即使属性值是空值 *，它实际上也在决策表中提供了一定的信息，所以我们在考虑对象的相似关系时，也应该将其纳入其中；另外，注意到属性值为空值所提供的信息与属性值非空且相同所提供的信息是存在差异的，由此我们给出如下相似关系，并据此进一步定义基于此相似关系的粗糙集．当 $B\subseteq A$，$0\leq\lambda\leq 1$，$0\leq\alpha\leq 1$，我们将 x 与 y 的集对 α(相似关系(2)改进为：

$$\mathrm{SIM}(B)=\{(x,y)\in U\times U\mid u_B(x,y)=a+bi+cj,\ a+\lambda b\geq\alpha\} \tag{3}$$

基于相似关系(3)，对象 x 的近似集(或邻域)定义为：

定义：对不完备信息系统 $S=(U,A)$，$x\in U$，$B\subseteq A$，$0\leqslant\lambda\leqslant 1$，定义 x 的 $B-\alpha$(邻域：

$$S_B^\alpha(x)=\{y\mid u_B(x,y)=a+bi+cj,\ a+\lambda b\geq\alpha,\ 0<\alpha\leq 1\}.$$

当 $\lambda=0$ 时，$B-\alpha$(邻域 $S_B^\alpha(x)$ 就是文献[2]中的相似关系；当 $\lambda=1$ 时，$S_B^\alpha(x)$ 就是文献[3]中的相似关系．所以这里定义的邻域是一种新的扩充，比文献[2～3]的讨论更具一般性．

显然，对象之间的这种关系不具有对称性和传递性．设 $U/\mathrm{SIM}(B)=\{S_B^\alpha(x)\mid x\in U\}$，它表示所有近似类的集合，此集一般不是 U 的划分，$U/\mathrm{SIM}(B)$ 中任意元素通常可称为信息粒或相似类．下面给出这种相似关系的几个性质：

性质 1　设 $0<\alpha_1<\alpha_2\leq 1$，对 $\forall x\in U$，$S_B^{\alpha 1}(x)\supseteq S_B^{\alpha 2}(x)$.

性质 2　设 $B\subseteq A$，$b\in A-B$，$|B|=n$，$0<\alpha\leq 1$，$\lambda=0$，对 $\forall x\in U$，有

$$S_{B\cup\{b\}}^\alpha(x)=S_B^{(n+1)\alpha/n}(x)\cup(S_B^{[(n+1)\alpha-1]/n}(x)\cap S_{\{b\}}^1(x)).$$

证明：设 $y\in U$，$u_B(x,y)=\frac{s}{n}+\frac{f}{n}i+\frac{p}{n}j$，

一方面，如果 $y\in S_B^{(n+1)\alpha/n}(x)\cup(S_B^{|(n+1)\alpha-1/n}(x)\cap S_{\{b\}}^1(x))$，则或者

$y\in S_B^{(n+1)\alpha/n}(x)$，或者 $y\in S_B^{[(n+1)\alpha-1]/n}\cap S_{\{b\}}^1(x)$.

(1) 当 $y\in S_B^{(n+1)\alpha/n}(x)$时，

$$\frac{s}{n+1}=\frac{s}{n}\cdot\frac{n}{n+1}\geq\frac{(n+1)\alpha}{n}\cdot\frac{n}{n+1}=\alpha,$$

所以 $y\in S_{B\cup\{b\}}^\alpha(x)$；

(2) 当 $y\in S_B^{|(n+1)\alpha-1/n}\cap S_{\{b\}}^1(x)$时，

$$\frac{s+1}{n+1}=\frac{s}{n}\cdot\frac{n}{n+1}+\frac{1}{n+1}\geq\frac{(n+1)\alpha-1}{n}\cdot\frac{n}{n+1}+\frac{1}{n+1}=\alpha,$$

则 $y\in S_{B\cup\{b\}}^\alpha(x)$.

因此，$S_{B\cup\{b\}}^\alpha(x)\supseteq S_B^{(n+1)\alpha/n}(x)\cup(S_B^{[(n+1)\alpha-1]/n}(x)\cap S_{\{b\}}^1(x))$.

另一方面，如果 $y\in S_{B\cup\{b\}}^\alpha(x)$，若 $b(x)=b(y)$，则 $\frac{s+1}{n+1}\geq\alpha$，从而

$$\frac{s}{n}=\frac{s+1}{n+1}\cdot\frac{n+1}{n}-\frac{1}{n}\geq\frac{(n+1)\alpha}{n}-\frac{1}{n}=\frac{(n+1)\alpha-1}{n}$$

所以 $y\in S_B^{[(n+1)\alpha-1]/n}\cap S_{\{b\}}^1(x)$；若 $b(x)\neq b(y)$，从而

$$\frac{s}{n}=\frac{s+1}{n+1}\cdot\frac{n+1}{n}\geq\frac{(n+1)\alpha}{n}=\frac{(n+1)\alpha}{n}.$$

所以 $y\in S_B^{(n+1)\alpha/n}(x)$.

因此，$S_{B\cup\{b\}}^\alpha(x)=S_B^{(n+1)\alpha/n}(x)\cup(S_B^{[(n+1)\alpha-1]/n}(x)\cap S_{\{b\}}^1(x))$.

性质 2 是进一步讨论属性递增知识发现的基础．

2 集对粗糙集模型

定义：设信息系统 $S=(U,A)$，$X\subseteq U$，$B\subseteq A$，$0<\alpha\leq 1$，X 的 $A-\alpha$ 集对型下、上近似：

$\underline{R}_B^{\alpha}(X)=\{x\mid S_B^{\alpha}(x)\subseteq X\wedge x\in U\}$，

$\overline{R}_B^{\alpha}(X)=\{x\mid S_B^{\alpha}(x)\subseteq X\neq\Phi\wedge x\in U\}$.

当 $B=A$ 时，$\underline{R}_A^{\alpha}(X)$和 $\underline{R}_A^{\alpha}(X)$简记为 $\underline{R}^{\alpha}(X)$和 $\overline{R}^{\alpha}(X)$

下面给出集对粗糙集上、下近似的几个性质，首先由定义立即有性质 3：

性质 3 设 $X\subseteq U$，$B\subseteq A$，$0<\alpha\leq 1$，有 $\underline{R}_B^{\alpha}(X)\subseteq X\subseteq\overline{R}_B^{\alpha}(X)$.

性质 4 设 $0<\alpha_1<\alpha_2\leq 1$，$X\subseteq U$，$B\subseteq A$，有 $\underline{R}_B^{\alpha_1}(X)\subseteq\underline{R}_B^{\alpha_2}(X)$和 $\overline{R}_B^{\alpha_1}(X)\supseteq\overline{R}_B^{\alpha_2}(x)$.

证明：显然当 $0<\alpha_1<\alpha_2\leq 1$ 时，对 $\forall x\in U$，由性质 1 有 $S_B^{\alpha_1}(x)\supseteq S_B^{\alpha_2}(x)$.

对 $\forall x\in\underline{R}_B^{\alpha_1}(X)$，有 $S_B^{\alpha_1}(x)\subseteq X$，从而 $S_B^{\alpha_2}(x)\subseteq S_B^{\alpha_1}(x)\subseteq X$，则 $x\in\underline{R}_B^{\alpha_2}(X)$，故 $\underline{R}_B^{\alpha_1}(X)\subseteq\underline{R}_B^{\alpha_2}(X)$.类似也可证明 $\overline{R}_B^{\alpha_1}(X)\supseteq\overline{R}_B^{\alpha_2}(X)$成立.

由性质 3～4，得到集对粗糙集模型的分层结构：

性质 5 设 $0<\alpha_1<\alpha_2<\ldots<\alpha_t<\ldots\leq 1$，$X\subseteq A$，$B\subseteq A$，有

$\underline{R}_B^{\alpha_1}(X)\subseteq\underline{R}_B^{\alpha_2}(X)\subseteq\cdots\subseteq\underline{R}_B^{\alpha}(X)\subseteq\cdots\subseteq X$，

$X\subseteq\cdots\subseteq\overline{R}_B^{\alpha}(X)\subseteq\cdots\subseteq\overline{R}_B^{\alpha_2}(X)\subseteq\overline{R}_B^{\alpha_1}(X)$.

容易得到性质 6～7：

性质 6 若 $X\subseteq Y\subseteq U$，$B\subseteq A$，$0<\alpha\leq 1$，则

$\underline{R}_B^{\alpha}(X)\subseteq\underline{R}_B^{\alpha}(Y)$，$\overline{R}_B^{\alpha}(X)\subseteq\overline{R}_B^{\alpha}(Y)$.

性质 7 $\underline{R}^{\alpha}(X\cap Y)=\underline{R}^{\alpha}(x)\cap\underline{R}^{\alpha}(X)$，$\underline{R}^{\alpha}(x\cup Y)=\underline{R}^{\alpha}(X)\cup\underline{R}^{\alpha}(Y)$.

3 下、上近似集的算法

下面我们给出集对粗糙集模型的下、上近似集计算的算法.

算法一(下近似集)：

① Input $S=(U,A)$and X，$B\subseteq A$

② Compute $U/\mathrm{SIM}(B)$

③ Set $\emptyset\rightarrow\underline{R}_B^{\alpha}(X)$

④ For $i=1$ to $|U|$ Do

If $S_B^{\alpha}(x)\subseteq X$,then $\underline{R}_B^{\alpha}(X)\cup\{x\}-\underline{R}_B^{\alpha}(X)$

End

⑤ Output $\underline{R}_B^{\alpha}(X)$

算法二(上近似集)：

① Input $S=(U,A)$and X，$B\subseteq A$

② Compute $U/\mathrm{SIM}(B)$

③ Set $\emptyset\rightarrow\overline{R}_B^{\alpha}(X)$

④ For $i=1$ to $|U|$ *Do*

If $S_B^{\alpha}(x)\cap X\neq\emptyset$, then $\overline{R}_B^{\alpha}(X)\cup\{x\}\rightarrow\overline{R}_B^{\alpha}(X)$

End

⑤ Output $\overline{R}_B^{\alpha}(X)$

4 示 例

下面以决策表 1 为实例，具体计算集合 X 的下、上近似集.在该不完备信息系统中，$U=\{u_1,u_2,u_3,u_4,u_5,u_6,u_7,u_8\}$，$A=\{a,b,c,d,e\}$.取 $\alpha=0.7$，$\lambda=0.5$，$X=\{u_2,u_4,u_5,u_8\}$，则有

决策表 1 不完备信息系统

U \ A	a	b	c	d	e
u_1	2	1	*	1	1
u_2	*	0	0	*	0
u_3	1	1	*	0	1
u_4	*	1	1	1	1
$u5$	0	*	1	0	0
$u6$	1	*	0	*	1
$u7$	1	1	*	1	1
$u8$	0	*	0	0	0

$S^{\alpha}(u_1)=\{u_1,u_4,u_7\}$，$S^{\alpha}(u_2)=\{u_2,u_8\}$，$S^{\alpha}(u_3)=\{u_3,u_6,u_7\}$，$S^{\alpha}(u_4)=\{u_1,u_4,u_7\}$，$S^{\alpha}(u_5)=\{u_5,u_8\}$，$S^{\alpha}(u_6)=\{u_3,u_6,u_7\}$，$S^{\alpha}(u_7)=\{u_1,u_4,u_7\}$，$S^{\alpha}(u_8)=\{u_2,u_5,u_8\}$.

$\underline{R}_B^{\alpha}(X)=\{u_2,u_5,u_8\}$，$\overline{R}_{(x)}^{2}=\{u_1,u_2,u_4,u_5,u_7,u_8\}$.

5 总 结

本文注意到空值在信息系统中也提供了一定的知识，利用集对分析定义了不完备信息系统中对象的一种相似关系，并由此建立了不完备信息系统

的一个粗糙集模型,并对此模型的基本性质、算法等进行了讨论.我们还初步讨论了属性渐增时相似关系的性质,这是进一步讨论不完备信息系统的属性渐增式决策规则挖掘的一个基础,但是我们注意到这里给出的结果还不具有一般性,这是需要进一步研究的.

参考资料:

[1] 刘　清.Rough 集及 Rough 推理[M].北京:科学出版社,2001

[2] 黄　兵,等,基于集对分析的不完备信息系统粗糙集模型[J].计算机科学,2002.29(9.专刊):1～3.

[3] 黄　兵,等,改进集对粗集模型,计算机工程与应用[J].2004.2:82～84.

[4] 赵克勤.集对分析及其初步应用[M].杭州:浙江科学技术出版社,2000.

[5] 王国胤,Rough 集理论在不完备信息系统中的扩充[J].计算机研究与发展,2002,10:1238～1243.

[6] K.S.Chin, Jiye Liang, Chuangyin Dang, Rough Set Data Analysis Algorithms for Incomplete Information Systems, In: Proc. Of the 9th International Conference, 2003:264～268.

Rough Set Model Based on Set Pair Analysis in Incomplete Information System

DENG Yi-xiong, HUANG Zhao-hua

(School of Information Engineering, East China Jiaotong University, Nanchang 330013, China)

Abstract: Kough Set Theory, as a mathe matical fool of studying uncertaint information system, is widely applied in Data Mining and Knowledge Discovery. Because of depending on equivalent relations, the Rough Set Theory is limited in in complete information system. The rough set model of incomplete information system based on set pair analysis is discussed in reference[2]. In this paper, The null value provides some knowledges in the information system. The paper establishes a more general model than reference[2], and discusses its basic property and arithmetic of the model.

Key words: incomplete information system; set pair analysis; rough sets, relations of similarity