文章编号:2095-2163(2019)05-0121-04 中图分类号:O29, TP18 文献标志码:A

基于局部粗糙集研究不完备信息系统的理论

刘瑶瑶
(西安石油大学 计算机学院,西安 710065)

摘 要: 为了进一步有效处理不完备数据,本文将完备信息系统上的局部粗糙集理论扩展推广到不完备信息系统中,首先基于不完备信息系统的容差关系给出了局部粗糙集的定义,其次,研究了不完备信息系统上局部粗糙集的性质并基于不完备局部粗糙集给出了计算下近似的算法。最后,基于局部下近似的两部分,给出了不同的局部属性约简。

关键词: 不完备;局部粗糙集;属性约简;信息系统

The theory of incomplete information systems based on local rough sets

LIU Yaoyao
(School of Computer Science, Xi´an Shiyou University, Xi´an 710065, China)

【Abstract】 In order to further effectively deal with incomplete data, this paper extends the local rough set theory on complete information systems to incomplete information systems. Firstly, the definition of local rough sets is given based on the tolerance relationship of incomplete information systems. Secondly, the properties of local rough sets on incomplete information systems are studied, and the corresponding algorithms of finding local low approxi is designed. Finally, different attribute reductions are proposed based on two parts of the local lower approximation.

【Key words】 incomplete; local rough sets; attribute reduction; information system

0 引 言

粗糙集理论已成为不确定性管理和不确定性推理的有效工具,并已在人工智能领域得到了成功应用。粗糙集理论的优势在于是其所有的参数都是从给定的样本集中获得的,这可以从文献[1]中看出:“不精确的数值不是预先假设的,而是在近似值的基础上计算出来的,这里的近似值用来表达知识的不精确性”。迄今为止,粗糙集数据分析已广泛应用于特征选择[2-3]、模式识别[4]、数据挖掘[5]和知识发现[6]等。

在粗糙集理论中,概念近似和属性约简[7]是2个非常重要的问题。概念近似包括:上近似和下近似。给定样本集 U 和二元关系 R,可以构造其等价类[8],可以构建样本集上的任何子集的粗糙集、即上下近似。目前研究粗糙集时,必不可少地会提到Pawlak的经典粗糙集,但是在经典粗糙集中,可以看到集合的上下近似的计算需要扫描给定集合 U 中的所有对象,同时还要获得近似目标概念的信息粒子[9]。通常将这种粗糙集称为全局粗糙集,而研究即需标记数据[10]。然而,随着大数据时代的到来,标签数据是一件非常耗时费力的工作,为了解决时间复杂度的问题,文献[1]提出了一种新的理论框架:局部粗糙集降低了数据量大时下近似计算和属性约简的时间复杂度。但是在文献[5]中只考虑到了完备信息系统[9]下的局部粗糙集,尚未涉及到不完备[10]的问题。因为目前海量的数据中很多数据的值是不确定的,本文的研究就是在文献[5]的基础上引入不完备的思想,进一步研究上下近似的计算以及相关算法,该研究非常具有现实意义。

本文的安排如下:首先简要阐述不完备信息系统、完备信息系统、以及局部粗糙集的相关概念;其次,基于不完备信息系统,重新对上下近似进行新的定义;接着,研究不完备信息系统下局部粗糙集的相关性质;而后,设计了计算不完备信息系统中局部粗糙集下近似的算法;最后,给出了全文总结。

1 基础知识

定义1[11] (U,A,V,f) 是一个四元组,其中,U 是对象 s 的非空有限集合,A 是属性的非空有限集合,$\forall a \in A$,Va 表示属性 a 的值域;$V=\cup_{a\in A}Va$ 表示 A 的值域;f 为 $U\times A\to V$ 的一个映射,$f(x,a)=a(x)\in Va$ 是 x 在属性 a 上的取值,则称 (U,A,V,f) 是一个完备信息系统。

作者简介:刘瑶瑶(1993-),女,硕士研究生,主要研究方向:形式概念分析、粗糙集。

收稿日期:2019-05-19

定义 2[12] 形式上,(U,A,V,f)是一个四元组,其中,U是对象的非空有限集合,A是属性的非空有限集合,$\forall a \in A$,Va表示属性a的值域;$V=\cup_{a\in A}Va$表示A的值域;f为$U\times A\to V$的一个映射[13],$f(x,a)=a(x)\in Va$是x在属性a上的取值,且至少存在一个属性a使$a(x)=*$,即$*\in Va$,则称(U,A,V,f)是一个不完备信息统。

定义 3[14] 设R是非空集合U上的等价关系,$\forall x\in U$,令$[x]_R=\{y\mid y\in U\wedge xRy\}$,则称$[x]_R$为$x$关于$R$的等价类,简称$x$的等价类。

2 基于局部粗糙集研究不完备信息系统

给定一个信息系统$S=(U,A,V,f)$,U是对象的集合,A是属性的集合,对于一个属性的子集$B\subseteq C$,研究用$*$来表示缺失的属性值,在此基础上就可以定义容差关系。

定义 4 容差关系T_B的定义如下:

$$T_B=\{(x,y)\in U\times U\mid \forall a\in B, f_a(x)=f_a(y)\vee f_a(x)=*\vee f_a(y)=*\}. \quad (1)$$

显然,容差关系满足自反性、对称性,但是不满足传递性。

定义 5 基于上述的容差关系T_B,$\forall x\in U$,x的容差类$T_B(x)$的定义如下:

$$T_B(x)=\{y\mid y\in U\wedge (x,y)\in T_B\}. \quad (2)$$

定义 6 给定不完备信息系统$IIS=(U,A,V,f)$且$*\in V$,$X\subseteq U$,$B\subseteq A$,X关于B的基于容差关系的下近似和上近似分别定义如下:

$$\underline{B}_T(X)=\{x\in U\mid T_B(x)\subseteq X\}, \quad (3)$$

$$\overline{B}_T(X)=\{x\in U\mid T_B(x)\cap X\neq\varnothing\}. \quad (4)$$

例 1 表1是不完备信息系统(U,A,V,f),其中$U=\{x_1,x_2,x_3,x_4,x_5,x_6\}$,$A=\{a_1,a_2,a_3,a_4\}$。

基于表1取$B=A$,计算容差类如下:

$T_B=\{(x_1,x_2),(x_1,x_4),(x_1,x_5),(x_1,x_6),(x_2,x_1),(x_2,x_4),(x_2,x_5),(x_2,x_6),(x_4,x_1),(x_4,x_2),(x_4,x_5),(x_4,x_6),(x_5,x_1),(x_5,x_2),(x_5,x_4),(x_5,x_6),(x_6,x_1),(x_6,x_2),(x_6,x_4),(x_6,x_5),(x_3,x_3)\}$

给定样本集$X=\{x_1,x_2,x_3,x_6\}$,

$T_B(x_1)=\{x_1,x_2,x_4,x_5,x_6\}\not\subseteq X$,

$T_B(x_2)=\{x_1,x_2,x_4,x_5,x_6\}\not\subseteq X$,

$T_B(x_3)=\{x_3\}\subseteq X$,

$T_B(x_6)=\{x_1,x_2,x_4,x_5,x_6\}\not\subseteq X$,

$\underline{B}_T(X)=\{x_3\}$,

$\overline{B}_T(X)=\{x_1,x_2,x_3,x_4,x_5,x_6\}$

表 1 不完备信息系统(U,A,V,f)

Tab. 1 Incomplete information system(U,A,V,f)

U	a_1	a_2	a_3	a_4
x_1	1	0	0	1
x_2	1	*	0	1
x_3	1	*	1	1
x_4	1	*	0	1
x_5	*	*	0	1
x_6	1	0	0	*

定义 7 (U,A,V,f)是一个不完备信息系统,D是$P(U)\times P(U)$上的相容度,为此对于任意的$X\subseteq U$,可得其不完备的局部粗糙集的定义如下:

$$\underline{R}_\alpha(X)=\{x\mid D(X\mid T_B(x))\geqslant\alpha, x\in X\}, \quad (5)$$

$$\overline{R}_\beta(X)=\cup\{T_B(x)\mid D(X\mid T_B(x))>\beta, x\in X\}, \quad (6)$$

其中,$D(X\mid T_B(x))=|X\cap T_B(x)|/|T_B(x)|$。

当$\alpha=1$,$\beta=0$时,此时的局部粗糙集下的上下近似满足在容差关系下的上下近似,即:

$$\underline{R}_1(X)=\{x\in U\mid T_B(x)\subseteq X\}, \quad (7)$$

$$\overline{R}_0(X)=\{x\in U\mid T_B(x)\cap X\neq\varnothing\}. \quad (8)$$

例 2 (续表1),即可研究求取在局部粗糙集定义下的上下近似:

给定样本集$X=\{x_1,x_2,x_3,x_6\}$,取$\alpha=0.6$,$\beta=0.1$,

$T_B(x_1)\cap X=\{x_1,x_2,x_4,x_5,x_6\}\cap X=\{x_1,x_2,x_6\}$,

$T_B(x_2)\cap X=\{x_1,x_2,x_4,x_5,x_6\}\cap X=\{x_1,x_2,x_6\}$,

$T_B(x_3)\cap X=\{x_3\}\cap X=\{x_3\}$,

$T_B(x_6)\cap X=\{x_1,x_2,x_4,x_5,x_6\}\cap X=\{x_1,x_2,x_6\}$.

所以:$P(X\mid T_B(x_1))=0.6$,

$P(X\mid T_B(x_2))=0.6$,

$P(X\mid T_B(x_3))=0.2$,

$P(X\mid T_B(x_6))=0.6$,

因此:$\underline{R}_{0.6}(X)=\{x_1,x_2,x_6\}$,

$\overline{R}_{0.1}(X)=\{x_1,x_2,x_3,x_4,x_5,x_6\}$.

在这里,可以做一个明显的对比。同样都是在不完备的情况下,全局的粗糙集下(例1)的上近似为$\varnothing$,但是在局部粗糙集下的上近似不为$\varnothing$,这在一定程度上说明,局部粗糙集更精确。

3 不完备局部粗糙集下的相关性质

下面拟研究不完备信息系统(U,A,V,f)中局部粗糙集的相关性质,对此可得研究阐述如下。

定理 1 $\underline{R}_\alpha(X)\subseteq X$

证明：根据上文给出的不完备局部粗糙集的下近似定义，容易得到，对于任意的 $X\subseteq U$，并且 $0<\alpha\leqslant 1$，$\underline{R}_\alpha(X)\subseteq X$。

定理 2　$\underline{R}_\alpha(\varnothing)=\overline{R}_\beta(\varnothing)=\varnothing$

证明：对于任意的 $x\in U$，$x\notin\varnothing$，$D(\varnothing\mid T_B(x))=0$，然后根据定义 5，对于 $0\leqslant\beta<\alpha\leqslant 1$，$\underline{R}_\alpha(\varnothing)=\overline{R}_\beta(\varnothing)=\varnothing$。

定理 3　$\underline{R}_\alpha(U)=\overline{R}_\beta(U)=U$

证明：对于任意的 $x\in U$，$T_B(x)\subseteq U$，然后 $D(U\mid T_B(x))=1$，所以对于任何的 $0\leqslant\beta<\alpha\leqslant 1$，就能得到 $x\in T_B(x)$，$D(U\mid T_B(x))=1\geqslant\alpha$，并且 $D(U\mid T_B(x))=1>\beta$，$\forall x\in U$，故 $\underline{R}_\alpha(U)=\overline{R}_\beta(U)=U$。

定理 4　若 $X\subseteq Y$ 则 $\underline{R}_\alpha(X)\subseteq\underline{R}_\alpha(Y)$，$\overline{R}_\beta(X)\subseteq\overline{R}_\beta(Y)$

证明：假设 $X\subseteq Y$，任意 $x\in\underline{R}_\alpha(X)$，所以 $x\in X\subseteq Y$，并且 $D(X\mid T_B(x))\geqslant\alpha$，又因为 $X\subseteq Y$，根据 $D(X\mid T_B(x)=\mid X\cap T_B(x)\mid/\mid T_B(x)\mid$，可得 $D(X\mid T_B(x))\leqslant D(Y\mid T_B(x))$，因此，$\forall x\in Y$，$D(Y\mid T_B(x))\geqslant\alpha$，并且 $x\in\underline{R}_\alpha(Y)$，对于任意的 x，都有 $\underline{R}_\alpha(X)\subseteq\underline{R}_\alpha(Y)$。

同理，可以证明 $X\subseteq Y\Rightarrow\overline{R}_\beta(X)\subseteq\overline{R}_\beta(Y)$。

定理 5　$\underline{R}_\alpha(X\cap Y)\subseteq\underline{R}_\alpha(X)\cap\underline{R}\alpha(Y)$

$\overline{R}_\beta(X\cup Y)\supseteq\overline{R}_\beta(X)\cup\overline{R}_\beta(Y)$

证明：$\forall x\in\underline{R}_\alpha(X\cap Y)$，由定理 4 显然易得：

$\underline{R}_\alpha(X\cap Y)\subseteq\underline{R}_\alpha(X)\cap\underline{R}_\alpha(Y)$，

同理，$\underline{R}_\alpha(X\cap Y)\subseteq\underline{R}_\alpha(X)\cap\underline{R}_\alpha(Y)$。

分析可知，研究不完备局部粗糙集的重点在于如何计算不完备信息系统中局部粗糙集的上下近似，进而求得属性约简，所以本文将给出在不完备信息系统下近似的算法。内容详见如下。

算法　计算不完备局部粗糙集下给定样本集的下近似

输入　一个不完备信息系统 $S=(U,R,V,f)$，一个样本集 $X\subseteq U$，以及参数 α

输出　样本集在属性集 A 下局部下近似 LA。

Step 1　从 $i=1$ 到 $|X|$ 做循环，计算 x_i 的容差类 $T_B(x_i)$，$x_i\in X$

Step 2　$LA\Leftarrow\varnothing$，$i\Leftarrow 1$；

Step 3　当 $i<|X|$，做循环

{

如果 $D(X\mid T_B(x_i))\geqslant\alpha$

然后 $LA\Leftarrow LA\cup\{x_i\}$，$i\Leftarrow i+1$

否则 $i\Leftarrow i+1$

}

Step 4　返回 LA，算法结束。

4　属性约简

这一节，将基于不完备信息系统局部粗糙集给出属性约简，同时研发在该系统下局部粗糙集的上下近似算法。对此可得设计论述如下。

首先，将 $\underline{R}_\alpha(X)=\{x\mid D(X\mid T_B(x))\geqslant\alpha,x\in X\}$ 拆分为 $\{x\mid D(X\mid T_B(x))=1,x\in X\}\cup\{x\mid 1>D(X\mid T_B(x))\geqslant\alpha,x\in X\}$。

记 $CL_R(X)=\{x\mid D(X\mid T_B(x))=1,x\in X\}$，

$PL_R(X)=\{x\mid 1>D(X\mid T_B(x))\geqslant\alpha,x\in X\}$。

显然，对于 $CL_R(X)$，则有下述结论成立。

定理 6　设 P,Q 为 2 个容差关系且 $P\subseteq Q$，则 $CL_Q(X)\subseteq CL_P(X)$。

证明：如果任意的 $x\in CL_Q(X)$，根据 $CL_Q(X)$ 的定义，就有 $D(X\mid T_Q(x))=1$，这里可得 $X\cap T_Q(x)=T_Q(x)$，再有 $T_Q(x)\subseteq X$，又知 $P\subseteq Q$，同时由相容类的关系，有 $T_P(x)\subseteq T_Q(x)\subseteq X$，因此有 $X\cap T_P(x)=T_P(x)$，即由包含度的定义得到 $D(X\mid T_B(x))=1$。因此可得对于任意的 $x\in CL_P(X)$。

基于 $CL_R(X)$，研发得到属性约简的定义表述如下。

定义 8　设 $S=(U,A,V,f)$ 为不完备信息系统，$X\subseteq U$，X 是样本集，$B\subseteq A$，如果 $|CL_B(X)|\geqslant|CL_A(X)|$ 且 $\forall B'\subset B$，$|CL_{B'}(X)|\ngeq|CL_A(X)|$ 都成立，称 B 是 X 在 S 中的一个局部属性约简 I。

基于定义 8，即可计算表 1 的属性约简。

例 3　记 $A=\{a_1,a_2,a_3,a_4\}$，给定样本集 $X=\{x_1,x_2,x_3,x_6\}$，由例 1 能求得 $CL_A(X)=\{x_3\}$，可得 $|CL_A(X)|=1$，记 $B=\{a_1,a_2,a_3\}$，

$T_B(x_1)=\{x_1,x_2,x_4,x_5,x_6\}\not\subseteq X$，

$T_B(x_2)=\{x_1,x_2,x_4,x_5,x_6\}\not\subseteq X$，

$T_B(x_3)=\{x_3\}\subseteq X$，

$T_B(x_6)=\{x_1,x_2,x_4,x_5,x_6\}\not\subseteq X$，

所以 $|CL_B(X)|=|T_B(x_3)|=1$

记 $B'=\{a_1,a_2\}$ 时，

则 $T_{B'}(x_1)=\{x_1,x_2,x_3,x_4,x_5,x_6\}\not\subseteq X$，

$T_{B'}(x_2)=\{x_1,x_2,x_3,x_4,x_5,x_6\}\not\subseteq X$，

$T_{B'}(x_3)=\{x_1,x_2,x_3,x_4,x_5,x_6\}\not\subseteq X$，

$T_{B'}(x_6)=\{x_1,x_2,x_3,x_4,x_5,x_6\}\not\subseteq X$，

所以 $|CL_{B'}(X)|=0$，但 $B'=\{a_1,a_3\}$ 及 $B'=\{a_2,a_3\}$ 时，$|CL_{B'}(X)|=1$，故根据定义，可知 B 不是 X 在 S 中的一个局部属性约简 I。继续计算，研

究发现 $B=\{a_3\}$ 是 X 在 S 中的一个局部属性约简 I。

事实上,定义 8 给出的约简就是保持容许类不变的约简,只是从个数上进行比较的约简,能够提高约简效率。下面,将从局部粗糙集的角度给出不完备信息系统的属性约简定义,具体描述如下。

定义 9 设 $S=(U,A,V,f)$ 为不完备信息系统,$X\subseteq U$,X 是样本集,$B\subseteq A$,对于任意的 $B'\subset B$,如果 $|\underline{R}_B(X)|\geqslant|\underline{R}_A(X)|$,并且 $|\underline{R}_{B'}(X)|\not\geqslant|\underline{R}_A(X)|$ 都成立,则称 B 是 X 在 (U,A,V,f) 中的一个局部属性约简 II。

基于定义 9,计算表 1 的属性约简。

例 4 记 $A=\{a_1,a_2,a_3,a_4\}$,给定样本集 $X=\{x_1,x_2,x_3,x_6\}$,取 $\alpha=0.6$,$\beta=0.1$,由例 2 可得到 $|\underline{R}_A(X)|=3$,记 $B=\{a_1,a_2,a_3\}$,

$T_B(x_1)\cap X=\{x_1,x_2,x_4,x_5,x_6\}\cap X=\{x_1,x_2,x_6\}$,

$T_B(x_2)\cap X=\{x_1,x_2,x_4,x_5,x_6\}\cap X=\{x_1,x_2,x_6\}$,

$T_B(x_3)\cap X=\{x_3\}\cap X=\{x_3\}$,

$T_B(x_6)\cap X=\{x_1,x_2,x_4,x_5,x_6\}\cap X=\{x_1,x_2,x_6\}$。

所以:$P(X|T_B(x_1))=0.6$,

$P(X|T_B(x_2))=0.6$,

$P(X|T_B(x_3))=0.2$,

$P(X|T_B(x_6))=0.6$,

$\underline{R}_{0.6}(X)=\{x_1,x_2,x_6\}$,所以 $|\underline{R}_B(X)|=|\underline{R}_A(X)|=3$

当 $B'=\{a_1\}$

$T_{B'}(x_1)\cap X=\{x_1,x_2,x_3,x_4,x_5,x_6\}\cap X=\{x_1,x_2,x_3,x_6\}$,

$T_{B'}(x_2)\cap X=\{x_1,x_2,x_3,x_4,x_5,x_6\}\cap X=\{x_1,x_2,x_3,x_6\}$,

$T_{B'}(x_3)\cap X=\{x_1,x_2,x_3,x_4,x_5,x_6\}\cap X=\{x_1,x_2,x_3,x_6\}$,

$T_{B'}(x_6)\cap X=\{x_1,x_2,x_3,x_4,x_5,x_6\}\cap X=\{x_1,x_2,x_3,x_6\}$。

所以: $P(X|T_{B'}(x_1))=1$,

$P(X|T_{B'}(x_2))=1$,

$P(X|T_{B'}(x_3))=1$,

$P(X|T_{B'}(x_6))=1$,

$\underline{R}_{0.6}(X)=\{x_1,x_2,x_3,x_6\}$,所以 $|\underline{R}_{B'}(X)|=4>|\underline{R}_A(X)|$,故 $B=\{a_1,a_2,a_3\}$ 不是 X 在 S 中的一个局部属性约简 II。继续计算,就会发现 $B=\{a_3\}$ 是 X 在 S 中的唯一个局部属性约简 II。

5 结束语

本文在不完备信息系统中,引入了局部粗糙集的理论。讨论了在不完备信息系统的局部粗糙集的相关性质,重点是研究该系统下,如何计算下近似,也给出了计算下近似的相关的算法。

本文只是在局部粗糙集和不完备粗糙集结合下的一个初步探索。基于本文的结果,可以深入研究局部属性约简的算法,以及进一步降低算法的时间复杂度等内容。

参考文献

[1] PAWLAK Z. Rough sets[J]. International Joumal of Computer and Information Sciences, 1982, 11(5):341-356.

[2] BHATT R B, GOPAL M. On fuzzy-rough sets approach to feature selection[J]. Pattern Recognition Letters, 2005, 26(7): 965-975.

[3] ESKANDARI S, JAVIDI M M. Online streaming feature selection using rough sets [J]. International Journal of Approximate Reasoning, 2016, 69: 35-57.

[4] SWINIARSKI R W, SKOWRON A. Rough set methods in feature selection and recognition[J]. Pattern Recognition Letters, 2003, 24(6):833-849.

[5] QIAN Yuhua, LIANG Xinyan, WANG Qi, et al. Local rough set: A solution to rough data analysis in big data[J]. International Journal of Approximate Reasoning, 2018, 97:38-63.

[6] 陈志恩. 基于粒关系包含度矩阵的属性约简[J]. 西北师范大学学报(自然科学版), 2017, 53(5):24-28.

[7] 张晶,李德玉,王素格,等. 基于稳健模糊粗糙集模型的多标记文本分类[J]. 计算机科学, 2015, 42(7):270-275.

[8] MA Fumin, ZHANG Tengfei. Generalized binary discernibility matrix for attribute reduction in incomplete information systems [J]. The Journal of China Universities of Posts and Telecommunications, 2017, 24(4):57-68, 75.

[9] SHAO Mingwen, ZHANG Wenxiu. Dominance relation and rules in an incomplete ordered information system [J]. International Journal of Intelligent Systems , 2005, 20(1):13-27.

[10] YANG X B, YANG J Y, WU C. Dominance-based rough set approach and knowledge reductions in incomplete ordered information system[J]. Information Sciences, 2008, 178: 1219-1234.

[11] 张文修,吴伟志. 粗糙集理论介绍和研究综述[J]. 模糊系统与学报, 2000, 14(4):1-12.

[12] WANG Guoyin. Extension of rough set under incomplete information system [C] // 2002 IEEE World Congress on Computational Intelligence. 2002 IEEE International Conference on Fuzzy Systems. FUZZ-IEEE'02. Proceedings (Cat. No. 02CH37291). Honolulu, HI, USA: IEEE, 2002:1098-1103.

[13] 罗豪,续欣莹,谢珺,等. 基于扩展容差关系的不完备信息系统属性约简[J]. 计算机应用, 2016, 36(11):2958-2962.

[14] DAI Jianhua, HUHu, ZHENG Guojie, et al. Attribute reduction in interval-valued information systems based on information entropies[J]. Frontiers of Information Technology & Electronic Engineering, 2016, 17(9):919-928.