第 37 卷 第 3 期 自 动 化 学 报 Vol. 37, No. 3
2011 年 3 月 ACTA AUTOMATICA SINICA March, 2011

基于(α, λ)联系度容差关系的变精度粗糙集模型

徐 怡 [1, 2] 李龙澍 [1, 2]

摘 要 基于传统粗糙集理论的方法不能有效地处理含噪音的不完备信息系统. 根据集对分析理论, 提出 (α, λ) 联系度容差关系. 将 (α, λ) 联系度容差关系与 Ziarko 提出的多数包含关系相结合, 提出变精度 (α, λ) 联系度粗糙集模型. 给出了该模型下基于正域相似度的启发式属性约简算法, 分析了算法的时间复杂度, 通过仿真实验验证了所提方法处理含噪音的不完备信息系统的有效性.

关键词 不完备信息, 粗糙集, (α, λ) 联系度容差关系, 变精度, 属性约简

DOI 10.3724/SP.J.1004.2011.00303

Variable Precision Rough Set Model Based on (α, λ) Connection Degree Tolerance Relation

XU Yi[1, 2] LI Long-Shu[1, 2]

Abstract Traditional rough set-based methods cannot deal with incomplete information system with noisy data effectively. According to set pair analysis, (α, λ) connection degree tolerance relation is defined. This paper introduces the variable precision (α, λ) connection degree rough set model which combines the (α, λ) connection degree tolerance relation with the majority inclusion relation proposed by Ziarko. A heuristic attribute reduction algorithm based on the positive region similarity is given. Time complexity of the algorithm is analyzed. The experimental results show that the proposed method is more effective for incomplete information system with noisy data.

Key words Incomplete information, rough sets, (α, λ) connection degree tolerance relation, variable precision, attribute reduction

粗糙集理论是波兰学者 Pawlak 提出的一种研究不精确性和不确定性信息系统的数学工具[1], 在知识发现、模式识别、决策分析、数据挖掘等领域得到了广泛的应用[2−5]. 基于等价关系的经典粗糙集理论仅适用于完备的信息系统, 然而, 在现实世界中, 由于数据采集能力有限, 很多信息系统都是不完备和不精确的. 为了将粗糙集理论应用于不完备信息系统, 已有基于容差关系、量化容差关系、相似关系、限制容差关系、基于联系度的容差关系、甚至一般二元关系等各种扩充的粗糙集模型[6−11]. 但以上这些模型都是定义在精确的集合包含关系上, 处理分类问题的方式是完全 "包含" 或 "不包含", 没有某种程度上的包含, 因而抗噪音数据的能力较弱. 为了将粗糙集理论应用于不精确信息系统, 即含噪音的信息系统, Ziarko 提出了变精度粗糙集模型[12], 众多学者对其进行了深入的研究[13−14]. 考虑到在不完备信息系统中, 噪音数据存在的可能性更大, 为了使粗糙集理论能够有效处理含噪音的不完备信息系统, 本文主要做了以下几点工作: 1) 基于集对分析理论, 结合集对联系度中的同一度和对立度, 定义了 (α, λ) 联系度容差关系, 用于含噪音的不完备信息系统的分类; 2) 将 (α, λ) 联系度容差关系与 Ziarko 提出的多数包含关系相结合, 提出了变精度 (α, λ) 联系度粗糙集模型; 3) 提出了变精度 (α, λ) 联系度粗糙集模型下, 基于正域相似度的启发式属性约简算法, 分析了算法的时间复杂度, 并通过仿真实验验证了所提方法处理含噪音的不完备信息系统的有效性.

收稿日期 2009-11-05 录用日期 2010-10-08
Manuscript received November 5, 2009; accepted October 8, 2010
国家自然科学基金 (60273043), 安徽省自然科学基金 (090412054), 安徽省高等学校省级自然科学基金 (KJ2011Z020), 安徽大学人才科研启动基金 (02303113) 资助
Supported by National Natural Science Foundation of China (60273043), Natural Science Foundation of Anhui Province (090412054), Natural Science Foundation of Anhui Higher Education Institutions (KJ2011Z020) and the Research Foundation for Talented Scholars of Anhui University (02303113)
1. 安徽大学计算智能与信号处理教育部重点实验室 合肥 230039
2. 安徽大学计算机科学与技术学院 合肥 230039
1. Key Laboratory of Computation Intelligence and Signal Processing of the Ministry of Education, Anhui University, Hefei 230039 2. Department of Computer Science and Technology, Anhui University, Hefei 230039

1 基本概念

下面简单介绍和本文相关的概念[9, 12].

定义 1. 对信息系统 $S = (U, A, V, f)$, 其中, U 是对象的非空有限集合, A 是属性的非空有限集合, $V = \cup_{a \in A} V_a$ 是属性值的集合, V_a 是属性 $a \in A$ 的值域, f 是信息函数, $f : U \times A \to V$, 即 $f(x, a) \in V_a$, 它指定了 U 中每一对象 x 的属性值. 对信息

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

定义 2. 对不完备信息系统 $S=(U,A,V,f)$, x, $y\in U$, $B\subseteq A$, $|B|=N$, 设对象 x 和 y 组成集对, 在属性集 B 上: 有 R 个确定且属性值相同的属性; P 个确定且属性值不相同的属性; F 个不能确定属性值是否相同的属性, 则称比值:

R/N 为 x 和 y 在属性集 B 上的同一度;

F/N 为 x 和 y 在属性集 B 上的差异度;

P/N 为 x 和 y 在属性集 B 上的对立度;

$$\mu_B(x,y)=\frac{R}{N}+\frac{F}{N}i+\frac{P}{N}j$$

表示 x 和 y 的关系, μ_B 称为 x 和 y 的联系度, 简记为 $u=a+bi+cj$. 其中, i 和 j 起标记的作用, 即表示 F/N 是差异度, P/N 是对立度, 并以这两个标记与同一度相区别, $|\cdot|$ 表示集合的基数. 显然 $0\le a$, b, $c\le 1$, $a+b+c=1$. a, b, c 三个参量反映了所论 2 个集合在指定问题背景下的某种联系趋势.

定义 3. 设 X, Y 是论域 U 的两个非空子集, X 关于 Y 的相对正确分类率 $C(X,Y)$ 定义为:

$$C(X,Y)=\begin{cases}\dfrac{|X\cap Y|}{|X|}, & |X|>0\\ 0, & |X|=0\end{cases}$$

其中, $|\cdot|$ 表示集合的基数.

定义 4. 如果设定一个阈值 β, $0\le\beta\le 1$, 则部分包含关系定义如下:

$$Y\overset{\beta}{\supseteq}X \text{ or } X\overset{\beta}{\subseteq}Y,\quad C(X,Y)\ge\beta$$

称 X 以 β 包含于 Y, 或 Y 以 β 包含 X, 其中参数 β 为包含度阈值.

当 $0.5<\beta\le 1$ 时, 则定义了 Y 对 X 的 β 多数包含关系, 即 X 中有 50 % 以上的元素被 Y 包含 (或 X 与 Y 的公共元素占 X 的 50 % 以上).

2 变精度 (α, λ) 联系度粗糙集模型

现实世界中, 由于数据采集能力有限, 很多信息系统都是不完备且不精确的. 为了使粗糙集理论能有效地处理含噪音的不完备信息系统, 基于集对分析理论, 结合集对联系度中的同一度和对立度, 定义了 (α, λ) 联系度容差关系, 用于含噪音的不完备信息系统的分类. 将 (α, λ) 联系度容差关系与 Ziarko 提出的多数包含关系相结合, 提出变精度 (α,λ) 联系度粗糙集模型. 相关定义如下:

定义 5. 对不完备信息系统 $S=(U,A,V,f)$, $x,y\in U,B\subseteq A,0.5\le\alpha\le 1,0<\lambda<0.5$, ($\alpha$, λ) 联系度容差关系定义为:

$$SSR(B)=\{(x,y)\in U\times U|\mu_B(x,y)=a+bi+cj,\ a+b+c=1,\ a>\alpha,\ c<\lambda\}\cup I_x$$

相应的 x 的 (α, λ) 联系度容差类定义为:

$$SSR_B^{\alpha}(x)=\{y|\mu_B(x,y)=a+bi+cj,\ a+b+c=1,\ a>\alpha,\ c<\lambda\}\cup x$$

其中, I_x 为恒等函数, α 称为同一度阈值, λ 称为对立度阈值. 显然 (α, λ) 联系度容差关系满足自反性和对称性, 但不满足传递性. 在该模型中通过对同一度和对立度的调节可以有效地处理含噪音的不完备信息系统的分类问题.

基于 (α, λ) 联系度容差关系, 结合集合的多数包含关系, 定义如下的变精度 (α, λ) 联系度粗糙集模型.

定义 6. 设 (U,AT) 为不完备信息系统, U 为论域, $AT=C\cup D$, C 为条件属性集合, D 为决策属性集合, SSR 为 U 上的 (α, λ) 联系度容差关系, $\forall x\in U$, $SSR_C^{\alpha}(x)$ 表示对象 x 的 (α, λ) 联系度容差类. 对于 $\alpha\in[0.5,1]$, $\lambda\in(0,\ 0.5)$, $\beta\in(0.5,\ 1]$, $\forall X\subseteq U$, X 的 β 下近似定义为:

$$\underline{SSR}_{\beta}^{\alpha}(X)=\{x\in U|C(SSR_C^{\alpha}(x),X)\ge\beta\}$$

X 的 β 上近似定义为:

$$\overline{SSR}_{\beta}^{\alpha}(X)=\{x\in U|C(SSR_C^{\alpha}(x),X)>1-\beta\}$$

X 的 β 正区域 (或 X 的 β 下近似) 可理解为将 U 中的对象以不大于 $1-\beta$ 的分类误差分于 X 的集合; X 的 β 负区域相应理解为将 U 中的对象以不大于 β 的分类误差分于 X 的补集 (即 $\sim X$) 的集合.

定义 7. 设 (U, AT) 为不完备信息系统, U 为论域, $AT=C\cup D$, C 为条件属性集合, D 为决策属性集合, SSR 为 U 上的 (α, λ) 联系度容差关系, $\forall x\in U$, $SSR_C^{\alpha}(x)$ 表示对象 x 的 (α, λ) 联系度容差类. 对于 $\alpha\in[0.5,\ 1]$, $\lambda\in(0,\ 0.5)$, $\beta\in(0.5,\ 1]$, $\forall X\subseteq U$, X 的 β 正区域, β 负区域和 β 边界域分别定义如下:

β 正区域:

$$POS_{\beta}^{\alpha}(X)=\underline{SSR}_{\beta}^{\alpha}(X)=\{x\in U|C(SSR_C^{\alpha}(x),X)\ge\beta\}$$

β 负区域:

$$NEG_{\beta}^{\alpha}(X)=U-\overline{SSR}_{\beta}^{\alpha}(X)=\{x\in U|C(SSR_C^{\alpha}(x),X)\le 1-\beta\}$$

β 边界域:

$$BN_\beta^\alpha(X)=\overline{SSR}_\beta^\alpha(X)-\underline{SSR}_\beta^\alpha(X)=\{x\in U|1-\beta<C(SSR_C^\alpha(x),X)<\beta\}$$

定义 8. 设 (U, AT) 为不完备信息系统, U 为论域, $AT=C\cup D$, C 为条件属性集合, D 为决策属性集合, 集合簇 $U/D=\{D_1, D_2, \cdots, D_m\}$, 表示根据决策属性集 D 对论域进行等价划分的结果, SSR 为 U 上的 (α, λ) 联系度容差关系. 对于 $\alpha\in[0.5, 1]$, $\lambda\in(0, 0.5)$, $\beta\in(0.5, 1]$, 属性集 C 的 β 分类精度定义为:

$$d_C^\beta(D)=\frac{\sum_{i=1}^m\left|\underline{SSR}_\beta^\alpha(D_i)\right|}{\sum_{i=1}^m\left|\overline{SSR}_\beta^\alpha(D_i)\right|}$$

$d_C^\beta(D)$ 描述了用属性集 C 分类时, 能够确定类别的样本比例, 是属性分类优劣的度量.

由上述定义, 下面的定理成立.

定理 1. 设 (U, AT) 为不完备信息系统, U 为论域, $AT=C\cup D$, C 为条件属性集合, D 为决策属性集合, SSR 为 U 上的 (α, λ) 联系度容差关系, $\forall x\in U$, $SSR_C^\alpha(x)$ 表示对象 x 的 (α, λ) 联系度容差类, $\alpha\in[0.5, 1]$, $\lambda\in(0, 0.5)$, $\beta\in(0.5, 1]$, $\forall X\subseteq U$, 有下面的关系式成立:

$$POS_\beta^\alpha(\sim X)=NEG_\beta^\alpha(X)$$

其中, $\sim X=U-X$.

证明. 由 $NEG_\beta^\alpha(X)$ 的定义可知, $\forall x\in NEG_\beta^\alpha(X)$, 有 $C(SSR_C^\alpha(x),X)\leq 1-\beta$, 即 $|SSR_C^\alpha(x)\cap X|/|SSR_C^\alpha(x)|\leq 1-\beta$, 又因为 $\forall x\in U$, 有 $|SSR_C^\alpha(x)\cap X|=|SSR_C^\alpha(x)|-|SSR_C^\alpha(x)\cap\sim X|$, 所以 $|SSR_C^\alpha(x)\cap\sim X|/|SSR_C^\alpha(x)|\geq\beta$, 即 $\forall x\in NEG_\beta^\alpha(X)$, 有 $x\in POS_\beta^\alpha(\sim X)$. 同理可证 $\forall x\in POS_\beta^\alpha(\sim X)$, 有 $x\in NEG_\beta^\alpha(X)$. □

下面给出变精度 (α, λ) 联系度粗糙集模型上下近似集的一些性质.

定理 2. 设 (U, AT) 为不完备信息系统, U 为论域, $AT=C\cup D$, C 为条件属性集合, D 为决策属性集合, SSR 为 U 上的 (α, λ) 联系度容差关系, $\alpha\in[0.5, 1]$, $\lambda\in(0, 0.5)$, $\beta\in(0.5, 1]$, 下列关系成立:

1) $\underline{SSR}_\beta^\alpha(X)\subseteq X\subseteq\overline{SSR}_\beta^\alpha(X)$;
2) $\underline{SSR}_\beta^\alpha(\emptyset)=\overline{SSR}_\beta^\alpha(\emptyset)=\emptyset$;
3) $\underline{SSR}_\beta^\alpha(U)=\overline{SSR}_\beta^\alpha(U)=U$;
4) $\overline{SSR}_\beta^\alpha(X\cup Y)\supseteq\overline{SSR}_\beta^\alpha(X)\cup\overline{SSR}_\beta^\alpha(Y)$;
5) $\overline{SSR}_\beta^\alpha(X\cap Y)\subseteq\overline{SSR}_\beta^\alpha(X)\cap\overline{SSR}_\beta^\alpha(Y)$;
6) $\underline{SSR}_\beta^\alpha(X\cup Y)\supseteq\underline{SSR}_\beta^\alpha(X)\cup\underline{SSR}_\beta^\alpha(Y)$;
7) $\underline{SSR}_\beta^\alpha(X\cap Y)\subseteq\underline{SSR}_\beta^\alpha(X)\cap\underline{SSR}_\beta^\alpha(Y)$.

根据上下近似集的定义, 这些性质的证明较简单, 在此不赘述.

变精度 (α, λ) 联系度粗糙集模型, 通过对 α 和 λ 的调节, 可以有效处理含噪音的不完备信息系统的分类问题, 通过引入 β 来削弱由噪音数据所产生的不确定性, 以增强模型的鲁棒性. 在实际应用中, 可根据实际需要来调节 α, λ 和 β 的值, 从而有效处理含噪音的不完备信息系统.

3　属性约简

完备信息系统的属性约简, 主要是保证系统的正域不发生变化, 即约简后的决策表应和约简前的决策表识别能力相同. 这是因为在基于等价关系的完备信息系统中, 有这样的命题成立: 给定一个完备的信息系统 $S=(U, AT)$, $AT=C\cup D$, C 为条件属性集, D 为决策属性集, 若 $B\subseteq A\subseteq C$, 则 $POS_B(D)\subseteq POS_A(D)$. 但在含噪音的不完备信息系统中, 考虑 ($\alpha$, λ) 联系度容差关系时, 系统正域变化的单调性是不成立的. 因此, 使用正域不发生变化作为约简的定义就不合适了. 为此, 我们以如下两个条件作为启发式规则: 1) 系统的近似分类率不降低保证了系统的识别能力不降低; 2) 约简前后正域的相似性尽可能大保证了系统的识别准确度尽可能高. 据此, 提出变精度 (α, λ) 联系度粗糙集模型中以正域相似度为启发式规则的属性约简算法. 首先给出几个定义.

定义 9. 设不完备信息系统 $S=(U, AT)$, U 为论域, $AT=C\cup D$, C 为条件属性集合, D 为决策属性集合, SSR 为 U 上的 (α, λ) 联系度容差关系, 给定 $\alpha\in[0.5, 1]$, $\lambda\in(0, 0.5)$, $\beta\in(0.5, 1]$, (α, λ) 联系度容差关系下决策属性集 D 与条件属性集 C 的近似分类率定义为:

$$\gamma(C,D,\alpha,\beta)=\frac{|POS(C,D,\alpha,\beta)|}{|U|}$$

其中, $POS(C,D,\alpha,\beta)=\bigcup\limits_{D_i\in U/D}\underline{SSR}_\beta^\alpha(D_i)$

$\gamma(C, D, \alpha, \beta)$ 体现了决策属性集合 D 对条件属性集合 C 的依赖程度, 是经典粗糙依赖度的推广, 当 $\alpha=1$ 且 $\beta=1$ 时, 它就是经典粗糙依赖度 $\gamma(C, D)$. 近似分类率 $\gamma(C, D, \alpha, \beta)$ 说明了在特定的 α 和 β 值下, 论域 U 中基于决策类能被确定分类的对象比率.

定义 10. 给定一个不完备信息系统 $S=(U, AT)$, $AT=C\cup D$, SSR 为 U 上的 (α, λ) 联系度容差关系, 给定 $\alpha\in[0.5, 1]$, $\lambda\in(0, 0.5)$, $\beta\in(0.5, 1]$, $A\subseteq C$, 属性集 A 和 C 的正域相似度

定义为:

$$SIMPOS_{\beta}^{\alpha}(A,C)=1-\left|\frac{POS(A,D,\alpha,\beta)\oplus POS(C,D,\alpha,\beta)}{POS(A,D,\alpha,\beta)\cup POS(C,D,\alpha,\beta)}\right|$$

其中, $0\le SIMPOS_{\beta}^{\alpha}(A,C)\le 1$, $\oplus$ 表示对称差.

定义 11. 给定一个不完备信息系统 $S=(U,AT)$, $AT=C\cup D$, SSR 为 U 上的 (α,λ) 联系度容差关系, 给定 $\alpha\in[0.5,1]$, $\lambda\in(0,0.5)$, $\beta\in(0.5,1]$, 属性 $t\in C$ 的重要度定义为:

$$SIG_{\beta}^{\alpha}(t)=1-SIMPOS_{\beta}^{\alpha}(C-t,C),\quad 0\le SIG_{\beta}^{\alpha}(t)\le 1$$

属性重要度 $SIG_{\beta}^{\alpha}(t)$ 说明了在特定的 α 和 β 值下, 属性约简前后正域变化的程度. $SIG_{\beta}^{\alpha}(t)$ 的值越大, 说明去掉属性 t 对正域的影响越大, 即属性 t 越重要.

基于正域相似度的属性约简算法描述如下:

输入: 含噪音的不完备决策表 $S=(U,AT=C\cup D)$, 同一度阈值 α, 对立度阈值 λ, 分类正确率 β.

输出: S 的一个约简 B.

步骤 1. $B=C$;

步骤 2. 求出系统分类率 $\gamma(C,D,\alpha,\beta)$;

步骤 3. 对于每个属性 $t\in B$, 计算 $\gamma(B-t,D,\alpha,\beta)$;

步骤 4. 对于所有满足 $\gamma(B-t,D,\alpha,\beta)\ge\gamma(C,D,\alpha,\beta)$ 的 t, 计算属性 t 的重要度 $SIG_{\beta}^{\alpha}(t)$, 从中选择最小的 $SIG_{\beta}^{\alpha}(t)$, $B=B-\{t\}$; 如果所有满足 $\gamma(B-t,D,\alpha,\beta)\ge\gamma(C,D,\alpha,\beta)$ 的 t 的重要度 $SIG_{\beta}^{\alpha}(t)$ 都相同, 则选择在所有对象取值中空值最多的 $t, B=B-\{t\}$; 如果所有满足 $\gamma(B-t,D,\alpha,\beta)\ge\gamma(C,D,\alpha,\beta)$ 的 t 的重要度 $SIG_{\beta}^{\alpha}(t)$ 和空值数量都相同, 则从中任意选择一个 $t, B=B-\{t\}$;

步骤 5. 如果对于每个属性 $t\in B$, $\gamma(B-t,D,\alpha,\beta)<\gamma(C,D,\alpha,\beta)$, 转步骤 6, 否则转步骤 3;

步骤 6. 输出 B, 即为决策表 S 的一个约简.

算法的时间复杂度分析, 设 $|U|$ 和 $|C|$ 分别表示决策表中的对象个数和条件属性个数. 算法开始时, 需要计算条件属性 C, 相对决策属性 D 的近似分类率, 时间复杂度为 $O(|U|^2|C|)$; 在算法循环的每一步都需要计算条件属性 $B-t$, 相对决策属性 D 的近似分类率以及各属性的重要度, 时间复杂度为 $O(|U|^2|C|)$, 最坏情况下, 算法需循环 $|C|$ 次, 所以该算法总的时间复杂度为 $O(|U|^2|C|^2)$.

4 仿真实验

下面通过仿真实验, 进一步验证变精度 (α,λ) 联系度粗糙集模型中, 基于正域相似度的属性约简算法的有效性. 选用 UCI 数据库[15] 中的不完备数据集 Breast cancer 进行实验. Breast cancer 包含 286 个样本, 9 个条件属性, 1 个决策属性. 以分类精度和属性蒸发率[16] 作为约简性能的衡量标准. 利用 Visual C++ 中的随机函数, 在表 1 中随机加入一定量的噪音, 为了尽可能真实地模拟噪音环境, 相对于原始的条件属性值总数, 定量地将其中 10 %、20 % 和 30 % 的条件属性值设置为其值域内的任意值 (包括空值), 以此模拟有噪音的不完备决策表. 由于就某一次定量设置来说, 任意值的选择肯定会对约简及本文约简性能的评价产生影响, 而且这种影响带有很大的随机性, 为了规避这种随机性影响, 更真实地反映算法本身的性能, 对于每种噪音含量, 实验每次随机产生 100 个有噪音的不完备决策表, 利用基于正域相似度的属性约简算法进行处理, 计算约简的分类精度和属性蒸发率, 取 100 次分类精度和属性蒸发率的平均值作为衡量约简性能的指标. 需要说明的是: 从理论上来说, 产生的有噪音的不完备决策表越多越能规避随机性影响, 得到的分类精度和属性蒸发率的平均值就越能准确地反映约简性能, 但是随着有噪音的不完备决策表数量的增加, 实验的运行时间也增加, 考虑到算法运行效率, 本文选择了 100 个有噪音的不完备决策表进行实验. 对于不同的 α, λ 和 β 的值, 具体实验结果如表 1 所示.

首先, 分析不同的 α, λ 和 β 参数值对分类精度的影响. 根据表 1 中的数据, 对比参数值为 1 和 2 两种情况, 可以发现, 通过降低 β 的值, 可以增强算法的容错性, 从而提高分类精度; 对比参数值为 2 和 3 两种情况, 可以发现, 通过增加 λ 的值, 可以增强算法的容错性, 从而提高分类精度; 对比参数值为 3 和 4 两种情况, 可以发现, 通过降低 α 的值, 可以增强算法的容错性, 从而提高分类精度. 因此, 对于含噪音的不完备信息系统, 适当地降低同一度阈值 α, 增加对立度阈值 λ, 降低分类正确率 β, 可以增强算法的容错性, 提高分类精度, 使得算法具有良好的抗噪声能力. 对于不同的参数值情况, 随着噪音含量的增加, 分类精度都呈下降趋势, 这是符合实际的.

其次, 分析不同的 α, λ 和 β 参数值对属性蒸发率的影响. 通过对比分析表 1 中参数值为 1、2 和参数值为 3、4 的这两大类情况, 可以发现, 当参数值设置为 1、2 时, 由于对立度阈值的要求较严格 $(\lambda=0)$, 按照 (α,λ) 联系度容差关系分类得到的不可分辨类相对较多, 约简过程中, 为了保证系统的近似分类率不降低, 能够约简的属性就相对较少, 因此

属性蒸发率较低. 而参数值设置为 3、4 情况时, 在保持同一度阈值 α 没有提高的情况下, 放宽了对立度阈值的要求 ($\lambda = 0.3$), 按照 (α, λ) 联系度容差关系分类得到的不可分辨类相对较少, 约简过程中, 为了保证系统的近似分类率不降低, 能够约简的属性就相对较多, 因此属性蒸发率较高. 这一点和文献 [16] 中的观点: "提高蒸发率的有效方法是减少等价类的数量, 如果希望获得更高的蒸发率就需要进一步减少等价类" 是一致的. 对于每种参数值, 在不同的定量设置情况下, 属性蒸发率有细微的波动是受定量设置的随机性影响, 也是符合实际的.

表 1　基于 (α, λ) 联系度容差关系的约简结果
Table 1　The reduction results of (α, λ) connection degree tolerance relation

参数值	噪音量 (%)	分类精度 (%)	属性蒸发率 (%)
1) $\alpha = 0.6$, $\lambda = 0$, $\beta = 0.9$	10	88.35	2.89
	20	84.95	1.33
	30	78.82	1.11
2) $\alpha = 0.6$, $\lambda = 0$, $\beta = 0.6$	10	92.21	15.89
	20	90.79	14.11
	30	86.64	16.00
3) $\alpha = 0.6$, $\lambda = 0.3$, $\beta = 0.6$	10	97.55	78.11
	20	96.56	85.11
	30	91.98	81.22
4) $\alpha = 0.4$, $\lambda = 0.3$, $\beta = 0.6$	10	98.08	80.22
	20	97.23	82.44
	30	95.71	77.33

综上, 实验结果说明变精度 (α, λ) 联系度粗糙集模型下, 基于正域相似度的属性约简算法可以有效地处理含噪音的不完备信息系统约简问题.

将上述实验中的 (α, λ) 联系度容差关系替换为限制容差关系[8], 重复上述实验, 实验结果如表 2 所示.

分析表 2 中的数据可以得到和表 1 类似的结果, 进一步说明基于正域相似度的属性约简算法的有效性. 对比分析表 1 和表 2 中的数据, 可以发现在 (α, λ) 联系度容差关系下, 当 $\alpha = 0.6$, $\lambda = 0.3$, $\beta = 0.6$ 和 $\alpha = 0.4$, $\lambda = 0.3$, $\beta = 0.6$ 时, 其约简性能优于基于限制容差关系的约简性能, 说明本文所提的方法通过对 α, λ 和 β 的调节, 增加了算法的灵活性, 能够有效处理含噪音的不完备信息系统.

表 2　基于限制容差关系的约简结果
Table 2　The reduction results of limited tolerance relation

参数值	噪音量 (%)	分类精度 (%)	属性蒸发率 (%)
$\beta = 0.9$	10	88.21	3.11
	20	83.87	1.83
	30	77.71	1.12
$\beta = 0.6$	10	93.52	12.56
	20	88.62	9.22
	30	85.31	16.67

5　结论

不完备信息系统由于属性值的遗漏, 增加了系统的不确定性和噪声, 完全的包含与不包含关系定义过于严格, 不利于克服噪声数据. 本文提出的变精度 (α, λ) 联系度粗糙集模型, 增强了系统泛化和抗噪声能力, 能够有效地处理含噪音的不完备信息系统. 在实际应用中, 可根据需要来调节 α, λ 和 β 的值, 以有效处理含噪音的不完备信息系统. 给出了变精度 (α, λ) 联系度粗糙集模型中以正域相似度为启发式规则的属性约简算法, 分析了算法的时间复杂度, 并通过仿真实验验证了所提方法处理含噪音的不完备信息系统的有效性. 变精度 (α, λ) 联系度粗糙集模型下属性约简的效率和规则提取算法还有待进一步深入研究.

References

1 Pawlak Z. Rough sets. *International Journal of Computer and Information Sciences*, 1982, **11**(5): 341−356

2 AboulElla H. Fuzzy rough sets hybrid scheme for breast cancer detection. *Image and Vision Computing*, 2007, **25**(2): 172−183

3 Yue Xiao-Dong, Miao Duo-Qian, Zhong Cai-Ming. Roughness measure approach to color image segmentation. *Acta Automatica Sinica*, 2010, **36**(6): 807−816
(岳晓冬, 苗夺谦, 钟才明. 基于粗糙性度量的彩色图像分割方法. 自动化学报, 2010, **36**(6): 807−816)

4 Yang H H, Wu C L. Rough sets to help medical diagnosis-evidence from a Taiwan′s clinic. *Expert Systems with Applications*, 2009, **36**(5): 9293−9298

5 Mushrif M M, Ray A K. Color image segmentation: rough-set theoretic approach. *Pattern Recognition Letters*, 2008, **29**(4): 483−493

6 Kryszkiewicz M. Rough set approach to incomplete information system. *Information Sciences*, 1998, **112**(1−4): 39−49

7 Stefanowski J, Tsoukias A. On the extension of rough sets under incomplete information. In: Proceedings of the

7th International Workshop on New Directions in Rough Sets, Data Mining, and Granular Soft Computing. London: Springer-Verlag, 1999. 73−81

8 Wang Guo-Yin. Extension of rough set under incomplete information systems. *Journal of Computer Research and Development*, 2002, **39**(10): 1238−1243
(王国胤. Rough 集理论在不完备信息系统中的扩充. 计算机研究与发展, 2002, **39**(10): 1238−1243)

9 Xu Yi, Li Long-Shu, Li Xue-Jun. Generalized rough set model based on set pair situation. *Journal of System Simulation*, 2008, **20**(6): 1515−1517
(徐怡, 李龙澍, 李学俊. 基于集对势的扩充粗糙集模型. 系统仿真学报, 2008, **20**(6): 1515−1517)

10 Zhou Lei, Shu Lan. Rough set model based on new set pair analysis. *Fuzzy Systems and Mathematics*, 2006, **20**(4): 111−116
(周磊, 舒兰. 基于新集对分析的粗糙集模型. 模糊系统与数学, 2006, **20**(4): 111−116)

11 Huang Bing, Zhou Xian-Zhong, Shi Ying-Chun. Entropy of knowledge and rough set based on general binary relation. *Systems Engineering — Theory and Practice*, 2004, **24**(1): 93−96
(黄兵, 周献中, 史迎春. 基于一般二元关系的知识粗糙熵与粗集粗糙熵. 系统工程理论与实践, 2004, **24**(1): 93−96)

12 Ziarko W. Variable precision rough set model. *Journal of Computer and System Sciences*, 1993, **46**(1): 39−59

13 Mi J S, Wu W Z, Zhang W X. Approaches to knowledge reduction based on variable precision rough set model. *Information Sciences*, 2004, **159**(3−4): 255−272

14 Wang J Y, Zhou J. Research of reduct features in the variable precision rough set model. *Neurocomputing*, 2009, **72**(10−12): 2643−2648

15 Zwitter M, Soklic M. UCI Machine Learning Repository [Online], available: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer, March 20, 2010

16 Wang Jue, Wang Ren, Miao Duo-Qian, Guo Meng, Ruan Yong-Shao, Yuan Xiao-Hong, Zhao Kai. Data enriching based on rough set theory. *Chinese Journal of Computers*, 1998, **21**(5): 393−400
(王珏, 王任, 苗夺谦, 郭萌, 阮永韶, 袁小红, 赵凯. 基于 Rough Set 理论的 "数据浓缩". 计算机学报, 1998, **21**(5): 393−400)

徐 怡 安徽大学计算机科学与技术学院讲师, 博士. 主要研究方向为不精确信息处理, 粗糙集理论. 本文通信作者.
E-mail: xuyi1023@126.com
(**XU Yi** Ph. D., lecturer in the Department of Computer Science and Technology, Anhui University. Her research interest covers imprecision information processing and rough sets theory. Corresponding author of this paper.)

李龙澍 安徽大学计算机科学与技术学院教授. 主要研究方向为智能软件.
E-mail: lilongshu@126.com
(**LI Long-Shu** Professor in the Department of Computer Science and Technology, Anhui University. His main research interest is intelligent software.)