文章编号: 1001-0920 (2010) 11-1732-05

变精度集对势粗糙集模型

徐 怡, 李龙澍

(安徽大学 a. 计算智能与信号处理教育部重点实验室，b. 计算机科学与技术学院，合肥 230039)

摘 要: 为使粗糙集理论能有效处理含噪音的不完备信息系统, 将集对势扩充粗糙集模型和Ziarko 教授提出的多数包含关系相结合, 提出了变精度集对势粗糙集模型. 然后, 给出了正域相似度的定义, 提出了基于正域相似度的启发式属性约简算法, 并分析了算法的时间复杂度. 仿真实验表明了该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；集对势；变精度；正域相似度

中图分类号: TP18 **文献标识码:** A

Variable precision rough set model based on set pair situation

XU Yi, LI Long-shu

(a. Key Lab of Intelligent Computing and Signal Processing of Ministry of Education，b. School of Computer Science and Technology，Anhui University，Hefei 230039，China. Correspondent：XU Yi，E-mail：xuyi1023@126.com)

Abstract: To make rough set theory can deal with incomplete information system with noise effectively, combining the generalized rough set model based on set pair situation with the majority inclusion relation proposed by professor Ziarko, the variable precision set pair situation rough set model is proposed. Then the definition of positive region similarity is given. A heuristic attribute reduction algorithm based on positive region similarity is presented. The time complexity of the algorithm is analyzed. Simulation experiment shows the effectiveness of the proposed method for incomplete information system with noise.

Key words: Incomplete information；Rough set；Set pair situation；Variable precision；Positive region similarity

1 引 言

粗糙集理论是一种有效处理模糊性和不确定性的数学工具[1-3], 在智能信息处理、模式识别、数据挖掘等领域得到了广泛的应用[4-7]. 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, 为了将其应用于不完备信息系统, 已有基于容差关系、相似关系、限制容差关系等扩充粗糙集模型[8-13]. 文献[11]提出了基于集对势容差关系的集对势扩充粗糙集模型, 证明了该模型在处理不完备信息时, 可以克服已有扩充模型的局限性, 具有更好的分类性能. 但该模型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. 然而, 在现实世界中存在大量的噪音数据, 特别是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有效地处理含噪音的不完备信息系统, 本文做了以下工作: 1) 将集对势扩充粗糙集模型和Ziarko 教授提出的多数包含关系[14]相结合, 提出了变精度集对势粗糙集模型; 2) 由于不完备信息系统中正域变化的非单调性, 传统的基于正域不变的约简不再适用, 为此提出了正域相似度的概念, 用于衡量约简前后正域的变化程度; 3) 给出了变精度集对势粗糙集模型下, 基于正域相似度的启发式属性约简算法, 并分析了算法的时间复杂度. 最后通过仿真实验表明了所提方法的有效性.

2 基本概念

下面简单介绍本文将用到的基本概念[11,14,15].

定义 1 给定信息系统 $S = (U, A, V, f)$. 其中: U 是对象的非空有限集合; A 是属性的非空有限集合; $V = \bigcup_{a \in A} V_a$ 是属性值的集合, V_a 是属性 $a \in A$ 的值

收稿日期: 2009-12-24；**修回日期:** 2010-03-14.

基金项目: 国家自然科学基金项目(60273043)；安徽省自然科学基金项目(090412054)；安徽省科技攻关计划重大科技专项项目(08010201002)；安徽大学人才科研启动基金项目(02303113).

作者简介: 徐怡(1981−), 女, 安徽凤阳人, 讲师, 博士, 从事不精确信息处理、粗糙集的研究；李龙澍(1956−), 男, 安徽亳州人, 教授, 博士生导师, 从事智能软件、知识工程等研究.

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. 对于信息系统 $S = (U, A, V, f)$, 若至少存在一个属性 $a \in A$, 使 V_a 含有空值(用 $*$ 表示), 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　对于不完备信息系统 $S = (U, A, V, f)$, $x, y \in U, B \subseteq A$, 设对象 x 和 y 组成集对. 在属性集 B 上, 有 S 个确定且相同的属性, P 个确定且不相同的属性, F 个不能确定是否相同的属性. 则称比值 S/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. 令

$$\mu_B(x,y) = \frac{S}{N} + \frac{F}{N}i + \frac{P}{N}j \tag{1}$$

表示 x 和 y 的关系, μ_B 称为 x 和 y 的联系度, 简记为 $u = a + bi + cj$. 显然 $0 \leqslant a, b, c \leqslant 1, a + b + c = 1$. a, b, c 这 3 个参量反映了所论述的 2 个集合在指定问题背景下的某种联系趋势.

定义 3　对于不完备信息系统 $S = (U, A, V, f)$, $x, y \in U, B \subseteq A, 0.5 \leqslant \alpha \leqslant 1$, 集对势容差关系 SSR 定义为

$$\begin{aligned}\mathrm{SSR}(B) =& \{(x,y) \in U \times U | u_B(x,y) = \\ & a + bi,\ a + b = 1,\ a > \alpha\} \bigcup I_x.\end{aligned} \tag{2}$$

相应地, x 的集对势容差类定义为

$$\begin{aligned}\mathrm{SSR}_B^{\alpha}(x) =& \{y | u_B(x,y) = a + bi,\ a + b = \\ & 1,\ a > \alpha,\ 0.5 \leqslant \alpha \leqslant 1\} \bigcup x.\end{aligned} \tag{3}$$

其中: I_x 为恒等函数, α 为同一度阈值. 显然, 集对势容差关系满足自反性和对称性, 但不满足传递性.

定义 4　设不完备信息系统 $S = (U, A, V, f)$, $X \subseteq U, B \subseteq A, 0.5 \leqslant \alpha \leqslant 1$. X 关于属性集 B 的集对势容差关系上、下近似集分别定义为

$$\overline{\mathrm{SSR}}_B^{\alpha}(X) = \{x \in U | \mathrm{SSR}_B^{\alpha}(x) \bigcap X \neq \varnothing\}, \tag{4a}$$

$$\underline{\mathrm{SSR}}_B^{\alpha}(X) = \{x \in U | \mathrm{SSR}_B^{\alpha}(x) \subseteq X\}. \tag{4b}$$

文献 [11] 已详细证明了集对势扩充粗糙集模型在处理不完备信息系统时, 可以克服已有主流扩充模型的局限性, 具有更好的分类性能, 故在此不赘述.

定义 5　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

$$C(X,Y) = \frac{|X \bigcap Y|}{|X|}, |X| > 0, \tag{5}$$

其中 $|\ |$ 表示集合的基数.

定义 6　如果设定一个阈值 $\beta, 0 \leqslant \beta \leqslant 1$, 则部分包含关系定义为

$$Y \overset{\beta}{\supseteq} X, \text{ 或 } X \overset{\beta}{\subseteq} Y, C(X, Y) \geqslant \beta. \tag{6}$$

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 $0.5 < \beta \leqslant 1$ 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).

3　变精度集对势粗糙集模型

集对势扩充粗糙集模型虽然能够有效处理不完备信息系统, 但该模型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. 为了使粗糙集理论能有效地处理含噪音的不完备信息系统, 本文将集对势扩充粗糙集模型和多数包含关系相结合, 提出变精度集对势粗糙集模型.

定义 7　设 (U, AT) 为不完备信息系统, $\mathrm{AT} = C \bigcup D$, C 为条件属性集合, D 为决策属性集合, $\forall x \in U$, $\mathrm{SSR}_C^{\alpha}(x)$ 表示对象 x 的集对势容差类. 对于 $\alpha \in [0.5, 1], \beta \in (0.5, 1], \forall X \subseteq U, X$ 的 β 下近似集定义为

$$\underline{\mathrm{SSR}}_{\beta}^{\alpha}(X) = \{x \in U | C(\mathrm{SSR}_C^{\alpha}(x), X) \geqslant \beta\}, \tag{7a}$$

X 的 β 上近似集定义为

$$\overline{\mathrm{SSR}}_{\beta}^{\alpha}(X) = \{x \in U | C(\mathrm{SSR}_C^{\alpha}(x), X) > 1 - \beta\}. \tag{7b}$$

定义 8　设 (U, AT) 为不完备信息系统, $\mathrm{AT} = C \bigcup D$, 在变精度集对势粗糙集模型下, $\forall X \subseteq U, X$ 的 β 正区域, β 负区域和 β 边界域分别定义如下:

β 正区域

$$\begin{aligned}\mathrm{POS}_{\beta}^{\alpha}(C,X) =& \underline{\mathrm{SSR}}_{\beta}^{\alpha}(X) = \\ & \{x \in U | C(\mathrm{SSR}_C^{\alpha}(x), X) \geqslant \beta\};\end{aligned} \tag{8a}$$

β 负区域

$$\begin{aligned}\mathrm{NEG}_{\beta}^{\alpha}(C,X) =& U - \overline{\mathrm{SSR}}_{\beta}^{\alpha}(X) = \\ & \{x \in U | C(\mathrm{SSR}_C^{\alpha}(x),\ X) \leqslant 1 - \beta\};\end{aligned} \tag{8b}$$

β 边界域

$$\begin{aligned}\mathrm{BN}_{\beta}^{\alpha}(C,X) =& \overline{\mathrm{SSR}}_{\beta}^{\alpha}(X) - \underline{\mathrm{SSR}}_{\beta}^{\alpha}(X) = \\ & \{x \in U | 1 - \beta < C(\mathrm{SSR}_C^{\alpha}(x),\ X) < \beta\}.\end{aligned} \tag{8c}$$

按照上述定义, 有下面的定理成立:

定理 1　设 (U, AT) 为不完备信息系统, $\mathrm{AT} = C \bigcup D$, 在变精度集对势粗糙集模型下, 对于 $X \subseteq U$, 有下述关系式成立:

$$\mathrm{POS}_{\beta}^{\alpha}(C, \sim X) = \mathrm{NEG}_{\beta}^{\alpha}(C,X), \tag{9}$$

其中 $\sim X = U - X$.

证 明　由 $\mathrm{NEG}_{\beta}^{\alpha}(C,X)$ 的定义可知, $\forall x \in \mathrm{NEG}_{\beta}^{\alpha}(C,X)$, 有 $C(\mathrm{SSR}_C^{\alpha}(x), X) \leqslant 1 - \beta$, 即

$$\frac{|\mathrm{SSR}_C^{\alpha}(x) \bigcap X|}{|\mathrm{SSR}_C^{\alpha}(x)|} \leqslant 1 - \beta.$$

又因为 $\forall x \in U$, 有

$|\text{SSR}_C^\alpha(x)\bigcap X| = |\text{SSR}_C^\alpha(x)| - |\text{SSR}_C^\alpha(x)\bigcap \sim X|$,

所以

$$\frac{|\text{SSR}_C^\alpha(x)\bigcap \sim X|}{|\text{SSR}_C^\alpha(x)|} \geqslant \beta.$$

即$\forall\, x \in \text{NEG}_\beta^\alpha(C,X)$, 有$x \in \text{POS}_\beta^\alpha(C,\sim X)$. 同理, $\forall\, x \in \text{POS}_\beta^\alpha(C,\sim X)$, 有$x \in \text{NEG}_\beta^\alpha(C,X)$. □

下面给出变精度集对势粗糙集模型上下近似集的一些性质.

定理 2 变精度集对势粗糙集模型具有如下性质:

1) $\underline{\text{SSR}}_\beta^\alpha(X) \subseteq X \subseteq \overline{\text{SSR}}_\beta^\alpha(X)$;

2) $\underline{\text{SSR}}_\beta^\alpha(\varnothing) = \overline{\text{SSR}}_\beta^\alpha(\varnothing) = \varnothing$;

3) $\underline{\text{SSR}}_\beta^\alpha(U) = \overline{\text{SSR}}_\beta^\alpha(U) = U$;

4) $\overline{\text{SSR}}_\beta^\alpha(X\bigcup Y) \supseteq \overline{\text{SSR}}_\beta^\alpha(X)\bigcup\overline{\text{SSR}}_\beta^\alpha(Y)$;

5) $\overline{\text{SSR}}_\beta^\alpha(X\bigcap Y) \subseteq \overline{\text{SSR}}_\beta^\alpha(X)\bigcap\overline{\text{SSR}}_\beta^\alpha(Y)$;

6) $\underline{\text{SSR}}_\beta^\alpha(X\bigcup Y) \supseteq \underline{\text{SSR}}_\beta^\alpha(X)\bigcup\underline{\text{SSR}}_\beta^\alpha(Y)$;

7) $\underline{\text{SSR}}_\beta^\alpha(X\bigcap Y) \subseteq \underline{\text{SSR}}_\beta^\alpha(X)\bigcap\underline{\text{SSR}}_\beta^\alpha(Y)$.

根据上下近似集的定义, 性质 1)~3) 的证明较为简单, 在此不赘述. 下面给出其余性质的证明.

性质 4) 的证明 因为$\forall X,Y \subseteq U$, 有

$$\frac{|\text{SSR}_C^\alpha(x)\bigcap(X\bigcup Y)|}{|\text{SSR}_C^\alpha(x)|} \geqslant \frac{|\text{SSR}_C^\alpha(x)\bigcap X|}{|\text{SSR}_C^\alpha(x)|},$$

$$\frac{|\text{SSR}_C^\alpha(x)\bigcap(X\bigcup Y)|}{|\text{SSR}_C^\alpha(x)|} \geqslant \frac{|\text{SSR}_C^\alpha(x)\bigcap Y|}{|\text{SSR}_C^\alpha(x)|},$$

所以

$$\overline{\text{SSR}}_\beta^\alpha(X\bigcup Y) \supseteq \overline{\text{SSR}}_\beta^\alpha(X),$$

$$\overline{\text{SSR}}_\beta^\alpha(X\bigcup Y) \supseteq \overline{\text{SSR}}_\beta^\alpha(Y).$$

从而得到

$$\overline{\text{SSR}}_\beta^\alpha(X\bigcup Y) \supseteq \overline{\text{SSR}}_\beta^\alpha(X)\bigcup\overline{\text{SSR}}_\beta^\alpha(Y). \quad \square$$

性质 5) 的证明 因为$\forall X,Y \subseteq U$, 有

$$\frac{|\text{SSR}_C^\alpha(x)\bigcap(X\bigcap Y)|}{|\text{SSR}_C^\alpha(x)|} \leqslant \frac{|\text{SSR}_C^\alpha(x)\bigcap X|}{|\text{SSR}_C^\alpha(x)|},$$

$$\frac{|\text{SSR}_C^\alpha(x)\bigcap(X\bigcap Y)|}{|\text{SSR}_C^\alpha(x)|} \leqslant \frac{|\text{SSR}_C^\alpha(x)\bigcap Y|}{|\text{SSR}_C^\alpha(x)|},$$

所以

$$\overline{\text{SSR}}_\beta^\alpha(X\bigcap Y) \subseteq \overline{\text{SSR}}_\beta^\alpha(X),$$

$$\overline{\text{SSR}}_\beta^\alpha(X\bigcap Y) \subseteq \overline{\text{SSR}}_\beta^\alpha(Y).$$

从而得到

$$\overline{\text{SSR}}_\beta^\alpha(X\bigcap Y) \subseteq \overline{\text{SSR}}_\beta^\alpha(X)\bigcap\overline{\text{SSR}}_\beta^\alpha(Y). \quad \square$$

性质 6) 的证明 因为$\forall X,Y \subseteq U$, 有

$$\frac{|\text{SSR}_C^\alpha(x)\bigcap(X\bigcup Y)|}{|\text{SSR}_C^\alpha(x)|} \geqslant \frac{|\text{SSR}_C^\alpha(x)\bigcap X|}{|\text{SSR}_C^\alpha(x)|},$$

$$\frac{|\text{SSR}_C^\alpha(x)\bigcap(X\bigcup Y)|}{|\text{SSR}_C^\alpha(x)|} \geqslant \frac{|\text{SSR}_C^\alpha(x)\bigcap Y|}{|\text{SSR}_C^\alpha(x)|},$$

所以

$$\underline{\text{SSR}}_\beta^\alpha(X\bigcup Y) \supseteq \underline{\text{SSR}}_\beta^\alpha(X),$$

$$\underline{\text{SSR}}_\beta^\alpha(X\bigcup Y) \supseteq \underline{\text{SSR}}_\beta^\alpha(Y).$$

从而得到

$$\underline{\text{SSR}}_\beta^\alpha(X\bigcup Y) \supseteq \underline{\text{SSR}}_\beta^\alpha(X)\bigcup\underline{\text{SSR}}_\beta^\alpha(Y). \quad \square$$

性质 7) 的证明 因为$\forall X,Y \subseteq U$, 有

$$\frac{|\text{SSR}_C^\alpha(x)\bigcap(X\bigcap Y)|}{|\text{SSR}_C^\alpha(x)|} \leqslant \frac{|\text{SSR}_C^\alpha(x)\bigcap X|}{|\text{SSR}_C^\alpha(x)|},$$

$$\frac{|\text{SSR}_C^\alpha(x)\bigcap(X\bigcap Y)|}{|\text{SSR}_C^\alpha(x)|} \leqslant \frac{|\text{SSR}_C^\alpha(x)\bigcap Y|}{|\text{SSR}_C^\alpha(x)|},$$

所以

$$\underline{\text{SSR}}_\beta^\alpha(X\bigcap Y) \subseteq \underline{\text{SSR}}_\beta^\alpha(X),$$

$$\underline{\text{SSR}}_\beta^\alpha(X\bigcap Y) \subseteq \underline{\text{SSR}}_\beta^\alpha(Y).$$

从而得到

$$\underline{\text{SSR}}_\beta^\alpha(X\bigcap Y) \subseteq \underline{\text{SSR}}_\beta^\alpha(X)\bigcap\underline{\text{SSR}}_\beta^\alpha(Y). \quad \square$$

对于变精度集对势粗糙集模型, 通过调节α可得到不同概念层次上的知识粒度; 通过引入β可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, 可根据需要来调节α和β的值, 以得到不同层次的结果, 从而有效控制误差.

4 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与约简前的决策表具有相同的识别能力. 这是因为在基于等价关系的完备信息系统中, 有如下命题成立: 给定一个完备的信息系统$S = (U, \text{AT})$, $\text{AT} = C\bigcup D$, C为条件属性集, D为决策属性集, 若$B \subseteq A \subseteq C$, 则$\text{POS}_B(D) \subseteq \text{POS}_A(D)$. 但在含噪音的不完备信息系统中, 考虑集对势容差关系时, 系统正域变化的单调性不成立, 此时, 利用正域不变作为约简的定义不再适合了. 为此, 本文以如下两个条件作为启发式规则: 系统的近似分类率不降低; 约简前后正域的相似性尽可能大. 前者保证了系统的识别能力不降低; 后者则保证了系统的识别准确度尽可能高. 据此, 本文提出变精度集对势粗糙集模型下, 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$S = (U, \text{AT})$, $\text{AT} = C\bigcup D$, 给定$\alpha \in [0.5, 1]$, $\beta \in (0.5, 1]$. 变精度集对势粗糙集模型中, 决策属性集D与条件属性集C的近似分类率定义为

$$\gamma(C, D, \alpha, \beta) = |\text{RPOS}_\beta^\alpha(C, D)|/|U|, \tag{10}$$

其中

$$\text{RPOS}_\beta^\alpha(C, D) = \bigcup_{D_i \in U/D} \text{POS}_\beta^\alpha(C,\ D_i) = \bigcup_{D_i \in U/D} \underline{\text{SSR}}_\beta^\alpha(D_i).$$

近似分类率$\gamma(C, D, \alpha, \beta)$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, 它即为经典粗糙依赖度 $\gamma(C, D)$. $\gamma(C, D, \alpha, \beta)$ 表明了在特定的 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 $S = (U, \mathrm{AT}), \mathrm{AT} = C \bigcup D, A \subseteq C$. 属性集 A 和 C 的正域相似度定义为

$$\mathrm{SIMPOS}_{\beta}^{\alpha}(A, C) = 1 - \left| \frac{\mathrm{RPOS}_{\beta}^{\alpha}(A, D) \oplus \mathrm{RPOS}_{\beta}^{\alpha}(C, D)}{\mathrm{RPOS}_{\beta}^{\alpha}(A, D) \bigcup \mathrm{RPOS}_{\beta}^{\alpha}(C, D)} \right|. \quad (11)$$

其中: $0 \leqslant \mathrm{SIMPOS}_{\beta}^{\alpha}(A, C) \leqslant 1, \oplus$ 表示对称差.

定义 11　给定一个不完备信息系统 $S = (U, \mathrm{AT}), \mathrm{AT} = C \bigcup D$. 属性 $t \in C$ 的重要度定义为

$$\mathrm{SIG}_{\beta}^{\alpha}(t) = 1 - \mathrm{SIMPOS}_{\beta}^{\alpha}(C - t, C), \quad (12)$$

其中 $0 \leqslant \mathrm{SIG}_{\beta}^{\alpha}(t) \leqslant 1$.

属性重要度 $\mathrm{SIG}_{\beta}^{\alpha}(t)$ 表明了在特定的 α 和 β 值下, 属性约简前后正域变化的程度. $\mathrm{SIG}_{\beta}^{\alpha}(t)$ 的值越大, 表明去掉属性 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如下:

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

输出: S 的一个约简 B.

Step 1: 令 $B = C$.

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

Step 3: 对于每个属性 $t \in B$, 计算 $\gamma(B - t, D, \alpha, \beta)$ 和属性 t 的重要度 $\mathrm{SIG}_{\beta}^{\alpha}(t)$.

Step 4: 若 $\gamma(B - t, D, \alpha, \beta) \geqslant \gamma(C, D, \alpha, \beta)$ 且 $\mathrm{SIG}_{\beta}^{\alpha}(t)$ 最小, 则 $B = B - \{t\}$; 若 $\gamma(B - t, D, \alpha, \beta) \geqslant \gamma(C, D, \alpha, \beta)$ 且 $\mathrm{SIG}_{\beta}^{\alpha}(t)$ 都相同, 则选择在所有对象取值中空值最多的 t, $B = B - \{t\}$.

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

Step 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如下. 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 相对于决策属性 D 的近似分类率, 此时, 时间复杂度为 $O(|U|^2|C|)$; 算法循环的每一步都需要计算条件属性 $B - t$ 相对于决策属性 D 的近似分类率以及各属性的重要度, 因而时间复杂度为 $O(|U|^2|C|)$. 考虑到最坏情况下, 算法需循环 $|C|$ 次, 所以该算法总的时间复杂度为 $O(|U|^2|C|^2)$.

5　仿真实验

为便于对比分析, 以文献 [8] 给出的小汽车决策表为例, 验证变精度集对势粗糙集模型中, 基于正域相似度的属性约简算法的有效性.

不完备决策表如表 1 所示, $U = \{1, 2, 3, 4, 5, 6\}$, 条件属性集 $C = \{P, M, S, X\}$, 决策属性集 $D = \{d\}$. 其中 P, M, S, X, d 分别表示 Price, Mileage, Size, Max-speed 和 Acceleration.

表 1　不完备决策表

Car	Price	Mileage	Size	Max-speed	d
1	High	High	Full	Low	Good
2	Low	*	Full	Low	Good
3	*	*	Compact	High	Poor
4	High	*	Full	High	Good
5	*	*	Full	High	Excel
6	Low	High	Full	*	Good

表 1 可以视为一个无噪音的不完备决策表. 文献 [16] 使用不同的约简方法, 如分配约简、最大分布约简、分配序约简等. 处理该表得到 2 种约简结果: $\{S, X\}$ 和 $\{P, S, X\}$. 在下面的实验中, 将上述 2 个结果作为无噪音环境下该决策表的约简标准. 利用 Visual C++ 中的随机函数, 在表 1 中随机加入一定量的噪音, 分别为 10%, 20% 和 30%, 使之变为有噪音的不完备决策表. 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, 对这 100 个决策表利用基于正域相似度的属性约简算法进行处理. 考察 100 个处理结果中, 与 2 个标准结果一样的约简个数, 将其作为约简正确率. 对于每组噪音数据都进行 10 次实验, 取 10 次约简正确率的平均值. 具体结果如表 2 所示.

表 2　约简结果　%

参数值	噪音量	约简正确率
$\alpha = 0.5, \beta = 0.9$	10	83.7
	20	65.7
	30	53.2
$\alpha = 0.5, \beta = 0.6$	10	90.3
	20	71.8
	30	63.5

分析表 2 中的数据可以发现, 在同一参数值条件下, 基于正域相似度的属性约简算法总体的抗噪音性能较好, 在噪音含量为 30% 时, 约简正确率仍在 50% 以上. 随着噪音含量的增加, 约简正确率呈下降趋势, 这符合实际情况. 分析不同的分类正确率 β 对约简性能的影响可以发现, 通过降低 β 的值, 增加了算法的容错性, 从而可以相对提高约简正确率. 实验结果表明, 变精度集对势粗糙集模型下, 基于正域相似度的属性约简算法是有效的.

将上述实验中的集对势容差关系替换为限制容差关系[10], 其余不变, 重复该实验, 可得表 3 结果. 因

为限制容差关系中不涉及同一度阈值α, 所以表3中只有参数β. 分析表3中的数据, 可以得到与表2类似的结果, 这进一步表明了本文算法的有效性.

表 3 基于限制容差关系的约简结果 %

参数值	噪音量	约简正确率
$\beta = 0.9$	10	70.6
	20	60.4
	30	55.7
$\beta = 0.6$	10	86.9
	20	70.8
	30	64.6

对比分析表2和表3的数据可以发现, 基于集对势容差关系的整体约简正确率优于基于限制容差关系的约简正确率, 这也从一个侧面说明了集对势容差关系处理不完备信息系统的有效性.

6 结 论

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, 本文将集对势扩充粗糙集模型与Ziarko提出的多数包含关系相结合, 提出了变精度集对势粗糙集模型. 在实际应用中, 可根据需要调节α和β的值, 以得到不同层次的结果, 从而增强了系统的泛化和抗噪声能力; 同时, 给出了变精度集对势粗糙集模型下, 基于正域相似度的启发式属性约简算法, 并分析了算法的时间复杂度. 通过仿真实验表明了该方法处理含噪音的不完备信息系统的有效性. 变精度集对势粗糙集模型中的规则提取算法将是下一步研究的重点.

参考文献(References)

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

[2] Pawlak Z, Busse J G, Slowinski R, et al. Rough sets[J]. Communications of the ACM, 1995, 38(11): 88-95.

[3] Pawlak Z, Skowron A. Rough sets: Some extensions[J]. Information Sciences, 2007, 177(1): 28-40.

[4] Li Y, Shiu S C K, Pal S K, et al. A rough set-based case-based reasoner for text categorization[J]. Int J of Approximate Reasoning, 2006, 41(2): 229-255.

[5] AboulElla Hassanien. Fuzzy rough sets hybrid scheme for breast cancer detection[J]. Image and Vision Computing, 2007, 25(2): 172-183.

[6] Shyng J Y, Wang F K, Tzeng G H, et al. Rough set theory in analyzing the attributes of combination values for the insurance market[J]. Expert Systems with Applications, 2007, 32(1): 56-64.

[7] Milind M Mushrif, Ajoy K Ray. Color image segmentation: Rough-set theoretic approach[J]. Pattern Recognition Letters, 2008, 29(4): 483-493.

[8] Kryszkiewicz M. Rough set approach to incomplete information system[J]. Information Sciences, 1998, 112(1-4): 39-49.

[9] Stefanowski J, Tsoukias A. On the extension of rough sets under incomplete information[C]. Proc of the 7th Int Workshop on New Directions in Rough Sets, Data Mining, and Granular Soft Computing. Berlin: Springer-Verlag, 1999: 73-81.

[10] 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, 39(10): 1238-1243.
(Wang G Y. Extension of rough set under incomplete information systems[J]. J of Computer Research and Development, 2002, 39(10): 1238-1243.)

[11] 徐怡, 李龙澍, 李学俊. 基于集对势的扩充粗糙集模型[J]. 系统仿真学报, 2008, 20(6): 1515-1517.
(Xu Y, Li L S, Li X J. Generalized rough set model based on set pair situation[J]. J of System Simulation, 2008, 20(6): 1515-1517.)

[12] 黄兵, 周献中. 不完备信息系统中基于联系度的粗集模型拓展[J]. 系统工程理论与实践, 2004, 24(1): 88-92.
(Huang B, Zhou X Z. Extension of rough set model based on connection degree under incomplete information systems[J]. J of Systems Engineering-Theory and Practice, 2004, 24(1): 88-92.)

[13] Zhou Lei, Shu Lan. Rough set model based on new set pair analysis[J]. Fuzzy Systems and Mathematics, 2006, 20(4): 111-116.

[14] Ziarko W. Variable precision rough set model[J]. J of Computer and System Sciences, 1993, 46(1): 39-59.

[15] Aijun An, Ning Shan, Christine Chan, et al. Discovering rules for water demand prediction: An enhanced rough-set approach[J]. Engineering Applications of Artificial Intelligence, 1996, 9(6): 645-653.

[16] 黄兵. 基于粗糙集的不完备信息系统知识获取理论与方法[D]. 南京: 南京理工大学, 2004.
(Huang B. Rough sets-based theory and approaches for knowledge acquisition in incomplete information systems[D]. Nanjing: Nanjing University of Science and Technology, 2004.)