文章编号:1001 - 9081(2006)03 - 0666 - 02

# 基于 α 相同度相似关系的 rough 集模型

周 辉,王黔英,费 颖,袁 芳
(南昌大学 管理科学与工程系,江西 南昌 330047)
(zhui0471@163.com)

**摘 要**:Rough 集理论是一种处理不完备信息系统的数学工具,但是 Pawlak 的经典 rough 集理论似乎是不可行的,因为它要求论域中数据间有很强的等价关系。在产生基本集(相似类)时,一般相似关系的分类误差较大,集对分析会把两个对立度不为 0 的个体划分在一起。汲取两者的优点,给出相同度的概念,只有满足一般相似关系并且相同度大于或等于阈值 α 的两个对象才能划分在一个基本集中,在此基础上建立基于 α 相同度相似关系的 rough 集模型。通过实例验证效果要比基于集对分析或者一般相似关系的模型更好。

**关键词**:rough 集;α 相同度相似关系;β 变精度;不完备信息系统
**中图分类号**: TP182 **文献标识码**:A

# Rough set models based on α-identical degree similarity relation

ZHOU Hui, WANG Qian-ying, FEI Ying, YUAN Fang
(*Department of Management Science & Engineering, Nanchang University, Nanchang Jiangxi* 330047, *China*)

**Abstract:** Rough set theory is a mathematical tool to deal with incomplete information systems, but Pawlak's classic rough set model is unfeasible, because it requires strong equivalence relations among the datum of universe. When forming element sets (similar class), common similarity relation may have more error in classifying, set pair analysis may classify two objects that contrary degree between them is not '0' into a class. To solve above problem, concept of identical degree was put forward. Only if two objects satisfied common similarity relation and their identical degree exceeded (or equaled to) threshold, they would be classified into an element set. The rough set model based on α-identical degree similarity relation was built. Experiments show that it is better than the model based on set pair analysis or common similarity relation.

**Key words:** rough set; α-identical degree similarity relation; β-variable precision; incomplete information system

1982 年波兰数学家 Pawlak 提出了 rough 集理论,这是一种处理不确定性、模糊性知识的数学工具[1]。Pawlak 的经典 rough 集模型是以不分明关系为基础的,并在此基础上定义集合的上下近似集等概念,同时形成论域的划分。而不分明关系要求信息系统中的数据具有很强的等价关系,但是人们在实践中所收集到的表达知识信息的数据往往是不完备的,某些个体的一些属性值是不确定的,这就产生了不完备信息系统,因此建立个体间的等价关系显得不太可能。而文献[2]利用集对分析的思想定义不完备信息系统的 rough 集模型存在不足,它将不确定属性值确定为整个属性值集,在定义近似集时把两个对立度不为 0 的个体划分到同一个相似类中。本文从个体的相同度着手,建立基于 α 相同度相似关系的 rough 模型,不但能够克服上述不足,而且对当不完备信息系统中属性的不确定属性值比率较高时也非常适宜。

## 1 α 相同度相似关系

**定义 1** $S = \{U, AT, V, f\}$ 表示一个信息系统。其中 $U = \{x_1, x_2, \cdots, x_n\}$ 是非空有限个体集合,称为论域,$AT = \{a_1, a_2, \cdots, a_m\}$ 是非空有限属性的集合,信息函数 $f_j: U \times AT \rightarrow Va_j$, $j = 1, 2, \cdots, m$,对于每一个个体 $x_i \in U$,每一个属性 $a_j \in AT$,如果信息函数 $f$ 在 $V$ 中有唯一的一个值(称为确定的属性值),即 $|a_j(x_i)| = 1$,则称 $S$ 是完备的信息系统,否则称 $S$ 是不完备信息系统($V = \cup Va_j$),$||$ 表示集合的基数。

**定义 2** 在不完备信息系统 $S = \{U, AT, V, f\}$ 中,$\varnothing \neq A \subseteq AT$,$A$ 上的一般相似关系定义为:$SIM(A) = \{(x,y) \in U \times U \mid \forall a \in A, a(x) \cap a(y) \neq \varnothing\}$,$x$ 在 $A$ 上相似类定义为 $\sigma_A(x) = \{y \in U \mid (x,y) \in SIM(A)\}$。

**定义 3** 在不完备信息系统 $S = \{U, AT, V, f\}$ 中,$x, y \in U$,$\varnothing \neq A \subseteq AT$,定义 $T_A(x,y) = \begin{cases} \{a \in A, a(x) = a(y), |a(x)| = 1\}, & x \neq y \\ A, & x = y \end{cases}$,称 $T_A(x,y)$ 为 $x$ 与 $y$ 在属性集 $A$ 中的相同部分,$\alpha = |T_A(x,y)| / |A|$ 为 $x$ 与 $y$ 在 $A$ 上的相同度。

**定义 4** 在不完备信息系统 $S = \{U, AT, V, f\}$ 中,$\varnothing \neq A \subseteq AT$,定义 $A$ 上的 $\alpha (0 \leqslant \alpha \leqslant 1)$ 相同度相似关系为:$SIM^\alpha(A) = \{(x,y) \in U \times U \mid \forall a \in A, a(x) \cap a(y) \neq \varnothing, |T_A(x,y)| / |A| \geqslant \alpha\}$,$x$ 在 $A$ 上 $\alpha$ 相同度相似类定义为:$\sigma_A^\alpha(x) = \{y \in U \mid (x,y) \in SIM^\alpha(A)\}$。

由定义知,$x$ 在 $A$ 上 $\alpha$ 相同度相似类 $\sigma_A^\alpha(x)$ 恰好是 $x$ 在集对 $\alpha$ 相似关系中的 $A$-$\alpha$ 邻域 $S_A^\alpha(x)$[2] 与 $x$ 的一般相似关系相

**收稿日期**:2005 - 09 - 25 **修订日期**:2005 - 12 - 07

**作者简介**:周辉(1981 - ),男,河南南阳人,硕士研究生,主要研究方向:决策规划、计算机信息系统; 王黔英(1942 - ),女,安徽黄山人,教授,主要研究方向:Rough 集理论与方法、智能信息系统; 费颖(1982 - ),女,浙江湖州人,硕士研究生,主要研究方向:决策规划、计算机信息系统; 袁芳(1981 - ),女,江西九江人,硕士研究生,主要研究方向:决策规划、计算机信息系统.

似类 $\sigma_A(x)$ 的交集。若不完备信息系统中属性的不确定值比重比较大,一般的相似关系在获得相似类时,由于个体之间不确定值过多导致分类误差率很高,而采用 $\alpha$ 相同度相似关系时要求两个个体间具有相同属性比重不小于 $\alpha$,根据实际需要可以得到更好的分类。

**表 1　信息系统 $S=(U, AT)$**

| $U$ | $AT$ | | |
|---|---|---|---|
| | $a$ | $b$ | $c$ |
| 1 | 1,2 | 2 | 3 |
| 2 | 1 | 1,2,3 | 2,3 |
| 3 | 1,2,3 | 2 | 2, |
| 4 | 1 | 2,3 | 1,2,3 |

如表 1, $U=\{1,2,3,4\}$, $A=\{a,b,c\}$,有6个不确定属性值,6 个确定属性值,不确定属性值的比重比较大。采用一般相似关系分类时,形成的相似类有 $\sigma_A(1)=\{1,2,4\}$, $\sigma_A(2)=\{1,2,3,4\}$, $\sigma_A(3)=\{2,3,4\}$, $\sigma_A(4)=\{1,2,3,4\}$。我们看到,个体 2,4 与个体 1 没有相同的确定属性值,但它们却在个体 1 的相似类 $\sigma_A(1)$ 中,因此分类错误的可能性很高。按照文献[2],当 $\alpha=0.3$ 时,得到 $A$-$\alpha$ 邻域分别为: $S_A^{0.3}(1)=\{1,3\}$, $S_A^{0.3}(2)=\{2,4\}$, $S_A^{0.3}(3)=\{1,3\}$, $S_A^{0.3}(4)=\{2,4\}$。个体 1 与 3 在 $c$ 属性中取值不同,但它们却划分到同一个邻域中,显然不太合理。按本文方法,计算各个体的 $\alpha$ 相同度相似类为: $\sigma_A^{0.3}(1)=\{1\}$, $\sigma_A^{0.3}(2)=\{2,4\}$, $\sigma_A^{0.3}(3)=\{3\}$, $\sigma_A^{0.3}(4)=\{2,4\}$,很明显得到的分类比上面两种方法都细,知识的表达更精确。当 $\alpha=0$ 时,相同度相似关系退化为一般的相似关系。

## 2　基于 α 相同度相似关系的 rough 集模型

### 2.1　α 相同度相似关系的 rough 集模型

**定义 5**　在不完备信息系统 $S=\{U, AT, V, f\}$ 中, $\varnothing \neq A \subseteq AT$, $X \subseteq U$,定义 $X$ 的 $\alpha$ 相同度相似关系的下、上近似集为:

$$\underline{apr}^{\alpha}(X)=\{x \in U \mid \sigma_A^{\alpha}(x) \subseteq X\}$$

$$\overline{apr}^{\alpha}(X)=\{x \in U \mid \sigma_A^{\alpha}(x) \cap (x) \neq \varnothing\}。$$

从定义 5 可以看出, $\underline{apr}^{\alpha}(X)$ 表示的是若 $U$ 中元素 $x$ 在 $A$ 上 $\alpha$ 相同度相似类包含于 $X$,则 $x$ 属于 $X$ 的 $\alpha$ 相同度相似关系的下近似集; $\overline{apr}^{\alpha}(X)$ 表示若 $U$ 中元素 $x$ 在 $A$ 上 $\alpha$ 相同度相似类与 $X$ 的交集非空,则 $x$ 属于 $X$ 的 $\alpha$ 相同度相似关系的上近似集。当 $\alpha=0$ 时, $\alpha$ 相同度相似关系的 rough 集退化为一般相似关系下的 rough 集。

### 2.2　β 变精度 α 相同度相似关系的 rough 集模型

$X$ 的 $\alpha$ 相同度相似关系的 rough 集,考虑的是 $U$ 中如下 $x$ 的集合: $x$ 在 $A$ 上 $\alpha$ 相同度相似类是包含于 $X$ 还是至少有一个元素属于 $X$。文献[3]指出,当讨论 $x$ 的等价类对 $X$ 的隶属度问题时,就成为变精度的 rough 集模型。据此定义 $\beta(0 \leqslant \beta \leqslant 1)$ 变精度 $\alpha$ 相同度相似关系的 rough 集模型。

**定义 6**　在不完备信息系统 $S=\{U, AT, V, f\}$ 中, $\varnothing \neq A \subseteq AT$, $X \subseteq U$,定义 $X$ 的 $\beta$ 变精度 $\alpha$ 相同度相似关系的下、上近似集 $(0 \leqslant \beta \leqslant 1)$ 为:

$$\underline{apr}_{\beta}^{\alpha}(X)=\left\{x \in U \mid \frac{|\sigma_A^{\alpha}(x) \cap X|}{|\sigma_A^{\alpha}(x)|} \geqslant 1-\beta\right\}$$

$$\overline{apr}_{\beta}^{\alpha}(X)=\left\{x \in U \mid \frac{|\sigma_A^{\alpha}(x) \cap X|}{|\sigma_A^{\alpha}(x)|} > \beta\right\}$$

$\beta$ 变精度,也就是说,在求 $X$ 的 $\alpha$ 相同度相似关系的下、上近似集时,允许存在一定的误差 $\beta$, $\beta$ 越大误差越大,根据实际情况 $\beta$ 取值范围为[0,0.5]。 $\beta=0$ 时就成为 2.1 中定义的模型。

## 3　实例

设不完备信息系统如表 2 所示, $U=\{1,2,3,4,5,6,7,8,9,10,11,12\}$, $AT=A=\{a,b,c,d,e\}$, $X=\{2,3,4,7,12\}$。

**表 2　不完备信息系统 $S=(U, AT)$**

| $U$ | $AT$ | | | | | $U$ | $AT$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $a$ | $b$ | $c$ | $d$ | $e$ | | $a$ | $b$ | $c$ | $d$ | $e$ |
| 1 | 0,1,2 | 1 | 1 | 0 | 0 | 7 | 2 | 0 | 0 | 0,1 | 0 |
| 2 | 1 | 1,2 | 0 | 1 | 0 | 8 | 1,2 | 1 | 1 | 0 | 0 |
| 3 | 0,1 | 0 | 0 | 1 | 0 | 9 | 1 | 0,1 | 0 | 0,1 | 1 |
| 4 | 0 | 1 | 1 | 0 | 1 | 10 | 1 | 1,2 | 0 | 0,1 | 1 |
| 5 | 0,1,2 | 0 | 0 | 0,1 | 0 | 11 | 0,1 | 1 | 1 | 0 | 0 |
| 6 | 1 | 0,1,2 | 0 | 1 | 1 | 12 | 1 | 1,2 | 1 | 0,1 | 0 |

用 $\underline{apr}(X)$, $\overline{apr}(X)$ 分别表示一般相似关系的下、上近似集, $\underline{R}^{\alpha}(X)$, $\overline{R}^{\alpha}(X)$ 表示集对分析的下、上近似集,根据定义,当 $\alpha=0.6$,得到的各个分类如表 3。并计算可得: $\underline{apr}(X)=\{2,4\}$, $\overline{apr}(X)=\{1,2,3,4,5,7,8,12\}$; $\underline{R}^{0.6}(X)=\{12\}$, $\overline{R}^{0.6}(X)=\{1,2,3,4,5,6,7,8,11,12\}$; $\underline{apr}^{0.6}(X)=\{2,4,12\}$, $\overline{apr}^{0.6}(X)=\{2,3,4,5,7,12\}$。

**表 3　$\alpha=0.6$ 时各关系形成的分类**

| $U$ | $AT$ | | |
|---|---|---|---|
| | $\sigma_A(x)$ | $S_A^{0.6}(x)$ | $\sigma_A^{0.6}(x)$ |
| 1 | {1,8,11,12} | {1,4,8,11} | {1,8,11} |
| 2 | {2} | {2,3,6} | {2} |
| 3 | {3,5} | {2,3,5,7} | {3,5} |
| 4 | {4} | {1,4,8,11} | {4} |
| 5 | {3,5,7} | {3,5,7} | {3,5,7} |
| 6 | {6} | {2,6,9,10} | {6} |
| 7 | {5,7} | {3,5,7} | {5,7} |
| 8 | {1,8,11,12} | {1,4,8,11} | {1,8,11} |
| 9 | {9,10} | {6,9,10} | {9,10} |
| 10 | {9,10} | {6,9,10} | {9,10} |
| 11 | {1,8,11} | {1,4,8,11} | {1,8,11} |
| 12 | {1,8,12} | {12} | {12} |

可以看到,采用 $\alpha$ 相同度相似关系时,集合的上下近似要比用一般相似关系和集对分析时更接近集合 $X$,描述更精确。当 $\beta=0.35$ 时, $\underline{apr}_{0.35}^{0.6}(X)=\{2,4,5,12\}$, $\overline{apr}_{0.35}^{0.6}(X)=\{2,3,4,5,7,12\}$,明显可知 $\underline{apr}_{\beta}^{\alpha} \subseteq X \subseteq \underline{apr}_{\beta}^{\alpha}$ 在一般情况下不成立,在允许误差 0.35 的情况下对象 5 包含在 $X$ 的 $\alpha$ 相同度相似关系的下近似集中。

## 4　结语

本文在 $\alpha$ 相同度相似关系的基础上描绘不完备信息系统,建立了基于 $\alpha$ 相同度相似关系的 rough 集模型,并给出实例与基于一般相似关系和集对分析的模型进行比较,本文模型效果更好。实际应用中参数 $\alpha$, $\beta$ 的选取依靠人的主观意识,因此这些模型也非常适合于不完备信息系统中不确定属性值较多的情况。

**参考文献:**

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

[2]　黄兵, 周献中. 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):1-3.

[3]　YAO YY, WONG SKM, LIN TY. A Review of rough set models [A]. LIN TY, CERCONE N, Ed. Rough sets and data mining analysis for imprecise data[C]. Kluwer Academic Publisher, 1997.