广义相似关系下的不完备信息系统粗糙集模型

谭 旭[1,2]，陈英武[2]，王桢珍[2]

(1. 深圳信息职业技术学院 计算机应用系，广东 深圳，518029；
2. 国防科学技术大学 信息系统与管理学院，湖南 长沙，410073)

摘 要：为了更好地从含有杂合数据和不完备数据的信息系统中提取合理的规则知识，构建基于广义相似关系的不完备信息系统粗糙集模型。其步骤为：针对决策信息系统中存在杂合数据的情况，并对决策信息系统中所存在的不完备信息进行细致区分，给出广义相似关系的定义；通过提出上、下广义相似划分的上、下近似的概念，给出 2 种划分意义下的属性约简和规则知识提取策略；最后，在理论上对该扩展粗糙集模型的正确性进行相关证明，并用实际算例进一步验证该模型的有效性和优越性。

关键词：广义相似关系；不完备信息系统；上近似；下近似；粗糙集
中图分类号：TP18 文献标识码：A 文章编号：1672−7207(2009)05−1360−07

Rough set model based on general similarity relation in incomplete information systems

TAN Xu[1,2], CHEN Ying-wu[2], WANG Zhen-zhen[2]

(1. Department of Computer Application, Shenzhen Institute of Information Technology, Shenzhen 518029, China;
2. School of Information Systems & Management, National University of Defense Technology, Changsha 410073, China)

Abstract: In order to extract reasonable rule-based knowledge from information systems with hybrid data and incomplete data, a new rough set model based on general similarity relation in incomplete information systems was proposed. The procedures were as follows: Firstly, general similarity relations were defined for the situation of hybrid data and different kinds of incomplete data in information systems. Secondly, two kinds of attribute reduction methods and rule-based knowledge extraction methods were investigated, which were built upon the concepts of upper and lower approximations with upper and lower general similarity partitions. Lastly, the extended rough set model was proved theoretically, and a numerical example reveals the validity and advantage of the proposed model.

Key words: general similarity relation; incomplete information system; upper approximation; lower approximation; rough set theory

粗糙集理论是 Pawlak 提出的一种处理不精确、不完备信息的软计算方法。因其不需借助任何外界信息能进行数据分析和处理，经过 20 多年的发展，已在模式识别、医疗、金融、水利、管理决策等领域得到应用。经典粗糙集理论是基于上、下近似和等价关系的概念来处理各种信息的，然而，对于现实生活中大量存在的不完备信息系统、混合数据信息系统以及模糊信息系统，通常难以得到满意的处理结果。于是，许多研究者提出了多种扩展的粗糙集模型，如变精度粗糙集模型[1−2]、模糊粗糙集模型[3−4]、覆盖粗糙集模

收稿日期：2008−09−01；修回日期：2008−12−28
基金项目：国家自然科学基金资助项目(70272002)
通信作者：谭 旭(1981−)，男，湖南株洲人，博士，讲师，从事粗糙集理论与智能决策的研究；电话：13760330247；E-mail: tanxu_nudt@yahoo.com

型[5−7]等。尤其针对不完备数据的信息系统，大量改进模型如容差模型[8]、量化容差模型[9]、非对称相似模型[10−11]、特征关系模型[12]等得到广泛应用。然而，这些粗糙集模型在实际应用中各有利弊[13]。这些模型对于不完备信息系统中存在混合数据[14]，且条件属性集对决策属性划分存在不确定性以及知识规则抽取存在不合理性，为此，本文作者提出一种广义粗糙集模型。

1　不完备信息系统的广义相似关系

定义 **1**　有序 5 元组 $T=\{U, C, D, V, f\}$称为一个信息系统。其中：$U=\{o_1, o_2, \cdots, o_n\}$，为决策表中全体数据对象的集合；$C$ 为条件属性集合，它们反映对象的特征；D 为决策属性集，反映对象的类别；$V=\bigcup_{a\in C\cup D} V_a$，为所有属性下的属性值集合；$f$ 为 $U\times(C\cup D)\to V$ 的信息函数，用于确定 U 中每一个对象在各个属性下的取值。给定 $a\in C$，V_a 可以为实数值的连续型数据，也可以为语言描述型数据。进一步，若 $\exists o\in U$，$a\in C$，使得 $f(o,a)=*$ or ϕ (记号 * 代表数据缺失，ϕ 表示数据无法确定)，则称该信息系统为不完备信息系统。

显然，完备信息系统是不完备信息系统的特例。本文假定决策属性不存在数据缺失和数据无法确定的问题，同时，约定决策属性上的取值为离散型数据。

定义 **2**　设 T 为定义的信息系统，U 为该信息系统的有限论域，条件属性集上的等价关系 IND(C)定义为 $\mathrm{IND}(C)=\{(o_i,o_j)\in U^2 \mid \forall a\in C, f(o_i,a)=f(o_j,a)\}$，记 $U/\mathrm{IND}(C)=\{[o_i]_C \mid o_i\in U\}$，为 U 在条件属性集 C 上等价划分，划分结果记为$\{X_1, X_2, \cdots, X_p\}$。其中，$[o_i]_C=\{o_j \mid (o_i,o_j)\in \mathrm{IND}(C)\}$。同理，记 U/IND(D)=$\{[o_i]_D \mid o_i\in U\}$，为 U 在决策属性 D 上等价划分，划分结果记为$\{Y_1, Y_2, \cdots, Y_q\}$。

等价关系满足自反性、对称性和传递性[14]。

由于条件属性上的取值可以为实数值的连续型数据，可以为语言描述型数据，也可以存在缺失数据或无法确定的数据，下面给出广义相似关系以及广义相似划分的定义。

定义 **3**　对于信息系统 T，$\forall o_i,o_j\in U$，定义对象 o_i 与 o_j 在条件属性集 C 下的广义相似度为：$\mathrm{Sim}_C(o_i,o_j)=\bigcap_{a\in C}\mathrm{Sim}_a(o_i,o_j)$。其中：

$$\mathrm{Sim}_a(o_i,o_j)=1-\frac{\left|f'(o_i,a)-f'(o_j,a)\right|}{\max\limits_{o\in U} f'(o,a)-\min\limits_{o\in U} f'(o,a)},$$

在广义相似度 $\mathrm{Sim}_C(o_i,o_j)$ 定义下的关系 S 称为广义相似关系。

对于对象集在条件属性 $a\in C$ 上取值为实数值的连续型数据，$f'(o,a)=f(o,a)$；对于对象集在条件属性 $a\in C$ 上取值为语言描述型数据，设 $P=\{p_k \mid k=1,2,\cdots,l\}$，为对象集在属性 a 上按序排列的可能取值(如 $P=\{p_1,p_2,p_3,p_4\}=$ {优, 良, 中, 差})；$l>1$，为语言标度数，则 $f'(o,a)=\{k \mid f(o,a)=p_k\}$；当对象在条件属性 $a\in C$ 上取值为缺失型和无法确定类型的数据时，综合分析几种扩展不完备粗糙集模型的利弊，并针对混合型数据的不完备信息系统的特点，给出如下定义。

定义 **4**　在不完备信息系统 T 中，$\forall o_i,o_j\in U$，$a\in C$，若有 $f(o,a)=\{value \mid value=* \text{ or } \phi\}$，根据不完备数据在对象 o_i, o_j 间所处的不同位置以及不完备数据的类型，给出如下计算：

a. if $f(o_i,a)=\phi$ and $(f(o_j,a)\neq * \text{ or } \phi)$ then $\left|f'(o_i,a)-f'(o_j,a)\right|=0$；

b. if $f(o_i,a)=*$ and $(f(o_j,a)\neq *$ or $\phi)$ and $o_i\in[o_j]_D$ then $\left|f'(o_i,a)-f'(o_j,a)\right|=0$；

c. if ($f(o_j,a)\neq *$ or ϕ) then $\left|f'(o_i,a)-f'(o_j,a)\right|=\max\limits_{o\in U} f(o,a)-\min\limits_{o\in U} f(o,a)$；

d. if $(f(o_i,a)\neq *$ and $o_i\notin[o_j]_D)$ then $\left|f'(o_i,a)-f'(o_j,a)\right|=\max\limits_{o\in U} f(o,a)-\min\limits_{o\in U} f(o,a)$。

从定义 4 可以看出，在不完备信息系统中，虽然是讨论对象在条件属性下取值为空的情况，但这里将数据缺失和不确定数据予以区分，分别赋予了不同的含义，也给出了相应的相似关系计算，如计算 a 与计算 b 和 d 的区别；对于不完备数据在 2 个比较对象间所处的不同位置，也分别给出了不同的定义，如计算 c 与计算 a, b 和 d 的区别；对于数据缺失的情形，考虑了决策属性上的取值对该缺失数据的取值影响，使之更加符合实际情况，如计算 b 和 d。

容易看出，$\mathrm{Sim}_a(o_i,o_j)\in[0,1]$。当 $\mathrm{Sim}_a(o_i,o_j)\to 0$ 时，称对象 o_i 与 o_j 在条件属性 a 上是极相异的；当 $\mathrm{Sim}_a(o_i,o_j)\to 1$ 时，则称对象 o_i 与 o_j 在条件属性 a 上是极相似的，根据定义 3 和定义 4，可得到如下性质。

性质 **1**：

a. $\mathrm{Sim}_C(o_i,o_i)\equiv\bigcap_{a\in C}1$；

b. $\mathrm{Sim}_C(o_i,o_j)\Leftrightarrow\mathrm{Sim}_C(o_j,o_i)$ 一般不成立；

c. $\mathrm{Sim}_C(o_i,o_j)=\mathrm{Sim}_C(o_j,o_k)\Rightarrow\mathrm{Sim}_C(o_i,o_k)=\mathrm{Sim}_C(o_i,o_j)$ 不成立。

性质 1 说明信息系统中对象间的广义相似关系满足自反性，一般不满足交换性和传递性。特别地，对于完备信息系统，自反性和交换性是可以成立的。

给出了信息系统中的对象集在广义相似关系下的相似度后，可对所有对象集条件属性集进行广义相似划分。相对于传统的“硬划分”，这是一种“软划分”。给定相似阈值 α，并设定 $\max_{ij}=\sup_{a\in C}(\mathrm{Sim}_a(o_i,o_j))$，$\min_{ij}=\inf_{a\in C}(\mathrm{Sim}_a(o_i,o_j))$，分别为对象 o_i 与 o_j 间在条件属性集 C 下的最大和最小广义相似关系值，有如下定义。

定义 **5** 设 T 为给定的广义不完备数据信息系统，$0.5<\alpha\leq 1$，为相似阈值，则 U 在条件属性集 C 的 α -下广义相似划分为 $U/S_\alpha(C)=\{[o_i]_\alpha^C \mid o_i\in U\}$，$[o_i]_\alpha^C=\{o_j \mid \min_{ij}\geq\alpha, o_j\in U\}$；$\alpha$ -上广义相似划分为 $U/S^\alpha(C)=\{[o_i]_C^\alpha \mid o_i\in U\}$，$[o_i]_C^\alpha=\{o_j \mid \max_{ij}\geq\alpha, \frac{\mathrm{card}(C_{ij})}{\mathrm{card}(C)}\geq 0.5, o_j\in U\}$。其中：$C_{ij}=\{a \mid \mathrm{Sim}_a(o_i,o_j)\geq\alpha,\ a\in C\}$，且定义 α -下广义相似划分的可信度为 $\mathrm{Bel}([o_i]_\alpha^C)=\inf_j\left\{\frac{\mathrm{card}(C_{ij})}{\mathrm{card}(C)}\right\}$，$\alpha$ -上广义相似划分的可信度为 $\mathrm{Bel}([o_i]_C^\alpha)=\inf_j\left\{\frac{\mathrm{card}(C_{ij})}{\mathrm{card}(C)}\right\}$。

显然，对于 α -下相似划分，$\mathrm{Bel}([o_i]_\alpha^C)\equiv 1$。需注意的是，相似阈值 α 的合理选取非常重要，因为它与广义相似软划分的质量密切相关。

定理 **1** 对象间的等价关系和等价划分是广义相似关系和广义相似划分的特例，而广义相似关系和广义相似划分则是等价关系和等价划分的泛化。

证明 给定信息系统 T，$\forall o_i\in U$，$a\in C$，若 $f(o_i,a)\neq *$ or ϕ，且 $f(o_i,a)=f(o_j,a)$，则 $|f'(o_i,a)-f'(o_j,a)|=0$，否则，$|f'(o_i,a)-f'(o_j,a)|=1$。那么，$\mathrm{Sim}_a(o_i,o_j)=\{0,1\}$。广义相似划分中相似阈值 $0.5<\alpha\leq 1$，显然，此时 $\alpha\equiv 1$。这样，$U/S_\alpha(C)=U/\mathrm{IND}(C)$。即在给定条件下，$\alpha$ -下相似划分退化成等价划分。且在这种 α -下相似划分下，显然满足 $\mathrm{Sim}_C(o_i,o_i)=\bigcap_{a\in C}1=1$，$\mathrm{Sim}_C(o_i,o_j)=\mathrm{Sim}_C(o_j,o_i)$ 以及 $\mathrm{Sim}_C(o_i,o_j)=\mathrm{Sim}_C(o_j,o_k)\Rightarrow\mathrm{Sim}_C(o_i,o_k)=\mathrm{Sim}_C(o_i,o_j)$。即这种广义相似关系退化成等价关系，且满足自反性、对称性和传递性。

2 广义相似关系下的粗糙集模型

由于讨论的信息系统存在数据的不完备性和多样性，在广义相似关系下划分后，通常存在不协调性，即存在 $[o_i]_C^\alpha\not\subseteq[o_i]_D$，$[o_i]_\alpha^C\not\subseteq[o_i]_D$。借鉴变精度粗糙集模型[1]，下面给出不完备信息系统中广义相似关系下的粗糙集模型定义。

定义 **6** 不完备信息系统 T 中，$Y\subset U/\mathrm{IND}(D)$ 关于条件属性集 C 的 α -上、下广义相似划分的 β -上、下近似定义为：

$$\underline{S}_\alpha^\beta(Y)=\{o_i\in U \mid \frac{\mathrm{card}([o_i]_\alpha^C\cap Y)}{\mathrm{card}([o_i]_\alpha^C)}\geq\beta\}\ ;$$

$$\overline{S}_\alpha^\beta(Y)=\{o_i\in U \mid \frac{\mathrm{card}([o_i]_\alpha^C\cap Y)}{\mathrm{card}([o_i]_\alpha^C)}\geq 1-\beta\}\ ;$$

$$\underline{S}_\beta^\alpha(Y)=\{o_i\in U \mid \frac{\mathrm{card}([o_i]_C^\alpha\cap Y)}{\mathrm{card}([o_i]_C^\alpha)}\geq\beta\}\ ;$$

$$\overline{S}_\beta^\alpha(Y)=\{o_i\in U \mid \frac{\mathrm{card}([o_i]_C^\alpha\cap Y)}{\mathrm{card}([o_i]_C^\alpha)}>1-\beta\}。$$

其中：$\underline{S}_\alpha^\beta(Y)$ 为集合 Y 的 α -下广义相似的 β -下近似；$\overline{S}_\alpha^\beta(Y)$ 为集合 Y 的 α -下广义相似的 β -上近似；$\underline{S}_\beta^\alpha(Y)$ 为集合 Y 的 α -上广义相似的 β -下近似；$\overline{S}_\beta^\alpha(Y)$ 为集合 Y 的 α -上广义相似的 β -上近似；$0.5<\beta\leq 1$，合理的 β 将使得在 α -上、下广义相似划分下均能获得最佳的上、下近似集，直接影响到后续的条件属性约简和规则的提取。

引理 **1** 在不完备信息系统 T 中，$\forall A\subseteq C$，在 α -上、下广义相似划分下，一定有：

a. $\bigcup_{Y\in U/\mathrm{IND}(D)}\underline{S}_\alpha^\beta(C,Y)\supseteq\bigcup_{Y\in U/\mathrm{IND}(D)}\underline{S}_\alpha^\beta(A,Y)$；

b. $\bigcup_{Y\in U/\mathrm{IND}(D)}\overline{S}_\alpha^\beta(C,Y)\subseteq\bigcup_{Y\in U/\mathrm{IND}(D)}\overline{S}_\alpha^\beta(A,Y)$；

c. $\bigcup_{Y\in U/\mathrm{IND}(D)}\underline{S}_\beta^\alpha(A,Y)\subseteq\bigcup_{Y\in U/\mathrm{IND}(D)}\overline{S}_\beta^\alpha(A,Y)$。

证明 在不完备信息系统中，$\forall A\subseteq C$，根据 α -下广义相似划分的定义，有 $U/S_\alpha(C)\subseteq U/S_\alpha(A)$。$\forall Y\in U/\mathrm{IND}(D)$，再由定义 6 中的 β -上、下近似的定义，得 $\underline{S}_\alpha^\beta(C,Y)\supseteq\underline{S}_\alpha^\beta(A,Y)$，$\overline{S}_\alpha^\beta(C,Y)\subseteq\overline{S}_\alpha^\beta(A,Y)$。

据定义 6，$\forall Y\in U/\mathrm{IND}(D)$，$\forall A\subseteq C$，显然，$\underline{S}_\beta^\alpha(A,Y)\subseteq\overline{S}_\beta^\alpha(A,Y)$。而此时无法确定 $U/S^\alpha(C)$ 与

$U/S^{\alpha}(A)$ 之间的包含关系。

综上分析，a~c 成立。证毕。

引理 1 阐释了在广义相似关系的背景下，条件属性集的变化给信息系统的知识划分带来的影响，也为广义相似关系下的属性约简进行了初步分析。

2.1　属性约简

在完备信息系统中等价关系和等价划分的定义中，条件属性的约简是保持信息系统分类能力不发生变化的条件下，剔除其中不相关或不重要的条件属性。通常的定义为[14]：给定 $A\subset C$，若存在 $\mathrm{POS}_A(D)=\bigcup_{Y\in U/\mathrm{IND}(D)}\underline{A}(Y)=\mathrm{POS}_C(D)=\bigcup_{Y\in U/\mathrm{IND}(D)}\underline{C}(Y)$，且 A 的任何真子集均不满足该条件，则称条件属性集 A 为条件属性集 C 相对于属性 D 的约简。同样，对于广义相似关系下不完备信息系统，依然可以定义条件属性的约简。

定义 **7**　给定不完备信息系统 T，$Y\in U/\mathrm{IND}(D)$，若

a. $\exists A\subset C$，使 $\bigcup_{Y\in U/\mathrm{IND}(D)}\underline{S}_{\alpha}^{\beta}(A,Y)=\bigcup_{Y\in U/\mathrm{IND}(D)}\underline{S}_{\alpha}^{\beta}(C,Y)$ 且 $\bigcup_{Y\in U/\mathrm{IND}(D)}\overline{S}_{\alpha}^{\beta}(A,Y)=\bigcup_{Y\in U/\mathrm{IND}(D)}\overline{S}_{\alpha}^{\beta}(C,Y)$，并且 A 的任何真子集均不满足以上条件，则称条件属性集 A 为条件属性集 C 相对于属性 D 的 α -下广义相似的 β 近似约简；

b. $\exists B\subset C$，使 $\bigcup_{Y\in U/\mathrm{IND}(D)}\underline{S}_{\beta}^{\alpha}(B,Y)=\bigcup_{Y\in U/\mathrm{IND}(D)}\underline{S}_{\beta}^{\alpha}(C,Y)$ 且 $\bigcup_{Y\in U/\mathrm{IND}(D)}\overline{S}_{\beta}^{\alpha}(B,Y)=\bigcup_{Y\in U/\mathrm{IND}(D)}\overline{S}_{\beta}^{\alpha}(C,Y)$，并且 B 的任何真子集均不满足以上条件，则称条件属性集 B 为条件属性集 C 相对于属性 D 的 α -上广义相似的 β 近似约简。

定义 7 结合了上、下近似 2 个方面来定义广义相似关系下的条件属性约简，更好地保证条件属性的约简结果能“保持信息系统分类能力不发生变化”。这也是引理 1 所推导的一种平衡状态。

定理 **2**　不完备信息系统 T 中，若条件属性集 A 为 α -下广义相似的 β 近似约简，条件属性集 B 为 α -上广义相似的 β 近似约简，则有

$$\bigcup_{Y\in U/\mathrm{IND}(D)}\underline{S}_{\beta}^{\alpha}(B,Y)\subseteq\bigcup_{Y\in U/\mathrm{IND}(D)}\underline{S}_{\alpha}^{\beta}(A,Y)\subseteq$$

$$\bigcup_{Y\in U/\mathrm{IND}(D)}\overline{S}_{\alpha}^{\beta}(A,Y)\subseteq\bigcup_{Y\in U/\mathrm{IND}(D)}\overline{S}_{\beta}^{\alpha}(B,Y)。$$

证明　根据定义 5，设 $U/S_{\alpha}(C)=\{X_1,X_2,\cdots,X_w\}$，则 $U/S^{\alpha}(C)=\{X_1,\cdots,X_i\cup X_j,\cdots,X_{j-1},X_j,\cdots,X_w\}$，即 $U/S_{\alpha}(C)\subseteq U/S^{\alpha}(C)$。且 $\forall A\subseteq C$，$B\subseteq C$，均有 $U/S_{\alpha}(A)\subseteq U/S^{\alpha}(A)$，$U/S_{\alpha}(B)\subseteq S^{\alpha}(B)$。再根据定义 6，$\bigcup_{Y\in U/\mathrm{IND}(D)}\underline{S}_{\beta}^{\alpha}(C,Y)\subseteq\bigcup_{Y\in U/\mathrm{IND}(D)}\underline{S}_{\alpha}^{\beta}(C,Y)$，$\bigcup_{Y\in U/\mathrm{IND}(D)}\overline{S}_{\alpha}^{\beta}(C,Y)\subseteq\bigcup_{Y\in U/\mathrm{IND}(D)}\overline{S}_{\beta}^{\alpha}(C,Y)$。

依据已知条件，若条件属性集 A 为 α -下广义相似的 β 近似约简，则有

$$\bigcup_{Y\in U/\mathrm{IND}(D)}\underline{S}_{\alpha}^{\beta}(A,Y)=\bigcup_{Y\in U/\mathrm{IND}(D)}\underline{S}_{\alpha}^{\beta}(C,Y)，$$

且

$$\bigcup_{Y\in U/\mathrm{IND}(D)}\overline{S}_{\alpha}^{\beta}(A,Y)=\bigcup_{Y\in U/\mathrm{IND}(D)}\overline{S}_{\alpha}^{\beta}(C,Y)。$$

条件属性集 B 为 α -上广义相似的 β 近似约简，则

$$\bigcup_{Y\in U/\mathrm{IND}(D)}\underline{S}_{\beta}^{\alpha}(B,Y)=\bigcup_{Y\in U/\mathrm{IND}(D)}\underline{S}_{\beta}^{\alpha}(C,Y)，$$

且

$$\bigcup_{Y\in U/\mathrm{IND}(D)}\overline{S}_{\beta}^{\alpha}(B,Y)=\bigcup_{Y\in U/\mathrm{IND}(D)}\overline{S}_{\beta}^{\alpha}(C,Y)。$$

则可推出：

$$\bigcup_{Y\in U/\mathrm{IND}(D)}\underline{S}_{\beta}^{\alpha}(B,Y)\subseteq\bigcup_{Y\in U/\mathrm{IND}(D)}\underline{S}_{\alpha}^{\beta}(A,Y)，$$

$$\bigcup_{Y\in U/\mathrm{IND}(D)}\overline{S}_{\alpha}^{\beta}(A,Y)\subseteq\bigcup_{Y\in U/\mathrm{IND}(D)}\overline{S}_{\beta}^{\alpha}(B,Y)。$$

再由

$$\bigcup_{Y\in U/\mathrm{IND}(D)}\underline{S}_{\alpha}^{\beta}(A,Y)\subseteq\bigcup_{Y\in U/\mathrm{IND}(D)}\overline{S}_{\alpha}^{\beta}(A,Y)，$$

定理得证。

定理 2 分析了 α -下广义相似的 β 近似约简结果与 α -上广义相似的 β 近似约简结果之间的关系，并充分说明了上广义相似的近似约简是一种更为粗放、粒度更大的约简，可以获取更全面的知识。而下广义相似的近似约简是一种粒度较小，要求更为精细的约简。虽然采用这 2 种约简能得到 2 种不同需求下的约简结果，但是，由于广义相似关系，它们之间存在着天然的联系。上广义相似及其约简是一种相对包容度更大的知识获取和表达方式。

定理 **3**　对于信息系统 T，基于等价关系的条件属性约简是基于广义相似关系条件属性约简的特例。

证明　设 $A\subseteq C$，为 α -下广义相似的 β 近似约简属性集，$B\subseteq C$，为 α -上广义相似的 β 近似约简属性

集，则 $\forall A' \subset A$ 或 $\forall B' \subset B$。依据定义 7，有

$$\bigcup_{Y\in U/\mathrm{IND}(D)} \underline{S}_{\alpha}^{\beta}(A',Y) \neq \bigcup_{Y\in U/\mathrm{IND}(D)} \underline{S}_{\alpha}^{\beta}(C,Y)，$$

且

$$\bigcup_{Y\in U/\mathrm{IND}(D)} \overline{S}_{\alpha}^{\beta}(A',Y) \neq \bigcup_{Y\in U/\mathrm{IND}(D)} \overline{S}_{\alpha}^{\beta}(C,Y)，$$

$$\bigcup_{Y\in U/\mathrm{IND}(D)} \underline{S}_{\beta}^{\alpha}(B',Y) \neq \bigcup_{Y\in U/\mathrm{IND}(D)} \underline{S}_{\beta}^{\alpha}(C,Y)，$$

且

$$\bigcup_{Y\in U/\mathrm{IND}(D)} \overline{S}_{\beta}^{\alpha}(B',Y) \neq \bigcup_{Y\in U/\mathrm{IND}(D)} \overline{S}_{\beta}^{\alpha}(C,Y)。$$

当 $\alpha \to 1$ 且 $\beta \to 1$，即相似关系逐渐强化为等价关系时，

$$\mathrm{POS}_{A'}(D) = \bigcup_{Y\in U/\mathrm{IND}(D)} \underline{A}'(Y) \neq \mathrm{POS}_{C}(D) = \bigcup_{Y\in U/\mathrm{IND}(D)} \underline{C}(Y)，$$

$$\mathrm{POS}_{B'}(D) \neq \mathrm{POS}_{C}(D)。$$

进一步，当对象间的划分变为等价关系时，由定理 2 可知，$\bigcup_{Y\in U/\mathrm{IND}(D)} \underline{S}_{\beta}^{\alpha}(C,Y) = \bigcup_{Y\in U/\mathrm{IND}(D)} \underline{S}_{\alpha}^{\beta}(C,Y)$，

$\bigcup_{Y\in U/\mathrm{IND}(D)} \overline{S}_{\alpha}^{\beta}(C,Y) = \bigcup_{Y\in U/\mathrm{IND}(D)} \overline{S}_{\beta}^{\alpha}(C,Y)$，而 A 为 α -下广义相似的 β 近似约简属性集，B 为 α -上广义相似的 β 近似约简属性集，即

$$\bigcup_{Y\in U/\mathrm{IND}(D)} \underline{S}_{\beta}^{\alpha}(B,Y) = \bigcup_{Y\in U/\mathrm{IND}(D)} \underline{S}_{\alpha}^{\beta}(A,Y)，$$

且

$$\bigcup_{Y\in U/\mathrm{IND}(D)} \overline{S}_{\alpha}^{\beta}(A,Y) = \bigcup_{Y\in U/\mathrm{IND}(D)} \overline{S}_{\beta}^{\alpha}(B,Y)，$$

此时 A=B。

可见，基于等价关系的条件属性约简是基于广义相似关系条件属性约简的特例。

2.2 规则提取

给出了广义相似关系和广义相似划分定义以及广义相似关系下的条件属性约简后，下面讨论如何从不完备信息系统中获取规则知识。

$\forall a \in C, o \in U$，有 $v = f(o,a) \in V$；$\forall b \in D, o \in U$，有 $w = f(o,b) \in V$，称$(a,\ v)$为条件属性集上的 1 个基本原子描述，$(b,\ w)$为决策属性上的 1 个基本原子描述。不同属性下的基本原子描述的组合称为 1 个基本描述，并记 $\Im = \bigcup_{a\in C} (a,v)$，为条件属性集下的基本描述，$\Re = \bigcup_{b\in D} (b,w)$，为决策属性下的基本描述。定义 $L(\Im)$ 为 $\Im$ 中所含的属性数目。对于决策属性为单一属性的信息系统，$L(\Re) = 1$。定义条件属性集下相似的基本描述簇为 $g_{\alpha}(\Im) = \bigcup_{a\in C} \{(a,v) \mid v = f(o_i,a), o_i \in [o_i]_{\alpha}^{C}\}$，条件属性集上相似基本描述簇为 $g^{\alpha}(\Im) = \bigcup_{a\in C} \{(a,v) \mid v = f(o_i,a), o_i \in [o_i]_{C}^{\alpha}\}$。在决策属性下，定义基本描述簇为 $h(\Re) = \{(b,w) \mid w = f(o_i,b), b \in D, o_i \in U\}$。记录

$$\left\|g_{\alpha}(\Im)\right\| = \{o_i \mid (a, f(o_i,a)) \in g_{\alpha}(\Im), \forall a \in C\}，$$

$$\left\|g^{\alpha}(\Im)\right\| = \{o_i \mid (a, f(o_i,a)) \in g^{\alpha}(\Im), \forall a \in C\}，$$

$$\left\|h(\Re)\right\| = \{o_i \mid (b,\ \ f(o_i,b)) \in h(\Re), b \in D\}，$$

为各描述簇内的对象集合。

性质 **2**：

a. $\left\|g_{\alpha}(\Im)\right\| \subseteq \left\|g^{\alpha}(\Im)\right\|$；

b. $\left\|g_{\alpha}(\Im)\right\| \subseteq \left\|h(\Re)\right\|$ if $\forall o_i \in \left\|g_{\alpha}(\Im)\right\|$, $(b, f(o_i, b)) \in h(\Re))$；

c. $\left\|g^{\alpha}(\Im)\right\| \subseteq \left\|h(\Re)\right\|$ if $\forall o_i \in \left\|g^{\alpha}(\Im)\right\|$, $(b, f(o_i, b)) \in h(\Re))$；

d. $\left\|g_{\alpha}(\Im)\right\| \cap \left\|h(\Re)\right\| \neq \varnothing$ if $\exists o_i \in \left\|g_{\alpha}(\Im)\right\|$, $(b, f(o_i, b)) \in h(\Re))$；

e. $\left\|g^{\alpha}(\Im)\right\| \cap \left\|h(\Re)\right\| \neq \varnothing$ if $\exists o_i \in \left\|g^{\alpha}(\Im)\right\|$, $(b,\ f(o_i, b)) \in h(\Re))$。

性质 a 表述了下相似基本描述簇内的对象集与上相似基本描述簇内的对象集之间的包含关系，即下广义相似划分包含于上广义相似划分，上广义相似划分是一种更为粗糙的划分；性质 b~e 分别阐释了上、下相似基本描述簇与决策属性上对应的基本描述簇之间的关系，为信息系统中规则的提取建立了关联。显然，这里的每一个基本描述簇都对应了 1 个(上、下相似)广义划分类，它们是划分类更详尽的一种描述。

根据 α -上、下广义相似划分以及 β -上、下近似定义，可以从不完备信息系统中抽取出相应的规则。对于 α -下广义相似划分，可以提取到 $g_{\alpha}(\Im) \to h(\Re)$ 形式的规则，表述成 if … then … 形式的规则为：if $\underset{a}{\wedge}(\underset{v}{\vee}(a,v))$ then (b,w) $(\forall (a,v) \in g_{\alpha}(\Im), (b,w) \in h(\Re))$；对于 α -上广义相似划分亦可以提取到 $g^{\alpha}(\Im) \to h(\Re)$ 形式的规则，表述成 if … then … 形式的规则为：if $\underset{a}{\wedge}(\underset{v}{\vee}(a,v))$ then (b,w) $(\forall (a,v) \in g^{\alpha}(\Im), (b,w) \in h(\Re))$。

由于信息系统的不完备性以及数据的多样性，最终提取得到的规则都是可能性规则，而这种“可能性”采用确定度和覆盖度来描述[11]。

定义 **8** 信息系统 T 中给定 1 条推理规则 $g \to h$ (其中：g 为上、下相似基本描述簇 $g^{\alpha}(\Im)$ 和 $g_{\alpha}(\Im)$ 的简写；h 为决策属性基本描述簇 $h(\Re)$ 的简写)，

$\mathrm{Bel}(\|g\|)$ 为定义 5 中上、下相似基本描述簇的可信度。则该规则的确定度定义为 $\mathrm{Cer}(g \to h) = \mathrm{Bel}(\|g\|) \cdot \dfrac{\mathrm{card}(\|g\| \cap \|h\|)}{\mathrm{card}(\|g\|)}$，该规则的覆盖度定义为 $\mathrm{Cov}(g \to h) = \mathrm{Bel}(\|g\|) \cdot \dfrac{\mathrm{card}(\|g\| \cap \|h\|)}{\mathrm{card}(\|h\|)}$。

最终可以得到 4 类规则：α-下广义相似 β-下近似下的规则，α-下广义相似 β-上近似下的规则，α-上广义相似 β-下近似下的规则，α-上广义相似 β-上近似下的规则。为此，由信息系统中的规则知识能够得到不遗漏的全面挖掘，并通过规则的确定度和覆盖度予以本质区别。

3　实例分析

以一段实际中常见的烤烟烟叶外观等级评价决策信息系统为例，对本文所提出的广义相似关系下的粗糙集模型以及该粗糙集模型进行的属性约简和规则提取进行分析和阐述。混合型数据不完备信息数据见表 1。

表 **1**　混合型数据不完备信息系统实例

Table 1　Example of incomplete information systems with hybrid data values

U	A_1	A_2	A_3	A_4/%	D
O_1	39.9	浓	有	21.7	中部烟
O_2	44.8	强	∅	19.8	中部烟
O_3	25.2	淡	稍有	*	下部烟
O_4	37.1	浓	有	20.4	下部烟
O_5	40.3	强	有	18.9	中部烟
O_6	29.5	弱	稍有	10.2	下部烟
O_7	*	弱	少	16.5	下部烟

由表 1 可见，对象集 U 在决策属性 D 上的等价划分为 $U/\mathrm{IND}(D)=\{\{o_1, o_2, o_5\}, \{o_3, o_4, o_6, o_7\}\}$。连续型实数值数据的条件属性 A_1 为烟叶的“长度”；语言描述型数据的条件属性 A_2 为烟叶的“色度”，取值范围为{浓, 强, 中, 弱, 淡}；条件属性 A_3 描述了烟叶的“油分”，语言标度集合为{多, 有, 稍有, 少}；条件属性 A_4 取值为百分比数据，表达烟叶的“残伤”。对象 o_2 在属性 A_3 上取值为无法确定型数据，对象 o_3 和对象 o_7 分别在属性 A_4 和属性 A_1 上取值为缺失型数据。

依据定义 3 和定义 4，可以分别得到对象之间在 4 个条件属性上的相似度。

$$[\mathrm{Sim}_{A_1}(o_i,o_j)]_{7\times7} = \begin{bmatrix} 1.00 & 0.75 & 0.25 & 0.86 & 0.98 & 0.47 & 0 \\ 0.75 & 1.00 & 0 & 0.61 & 0.77 & 0.22 & 0 \\ 0.25 & 0 & 1.00 & 0.40 & 0.23 & 0.78 & 0 \\ 0.86 & 0.61 & 0.40 & 1.00 & 0.84 & 0.61 & 0 \\ 0.98 & 0.77 & 0.23 & 0.84 & 1.00 & 0.45 & 0 \\ 0.47 & 0.22 & 0.78 & 0.61 & 0.45 & 1.00 & 0 \\ 0 & 0 & 1.00 & 0 & 0 & 1.00 & 1.00 \end{bmatrix};$$

$$[\mathrm{Sim}_{A_2}(o_i,o_j)]_{7\times7} = \begin{bmatrix} 1.00 & 0.75 & 0 & 1.00 & 0.75 & 0.25 & 0.25 \\ 0.75 & 1.00 & 0.25 & 0.75 & 1.00 & 0.50 & 0.50 \\ 0 & 0.25 & 1.00 & 0 & 0.25 & 0.75 & 0.75 \\ 1.00 & 0.75 & 0 & 1.00 & 0.75 & 0.25 & 0.25 \\ 0.75 & 1.00 & 0.25 & 0.75 & 1.00 & 0.50 & 0.50 \\ 0.25 & 0.50 & 0.75 & 0.25 & 0.50 & 1.00 & 1.00 \\ 0.25 & 0.50 & 0.75 & 0.25 & 0.50 & 1.00 & 1.00 \end{bmatrix};$$

$$[\mathrm{Sim}_{A_3}(o_i,o_j)]_{7\times7} = \begin{bmatrix} 1.00 & 0 & 0.67 & 1.00 & 1.00 & 0.67 & 0.33 \\ 1.00 & 1.00 & 1.00 & 1.00 & 1.00 & 1.00 & 1.00 \\ 0.67 & 0 & 1.00 & 0.67 & 0.67 & 1.00 & 0.67 \\ 1.00 & 0 & 0.67 & 1.00 & 1.00 & 0.67 & 0.33 \\ 1.00 & 0 & 0.67 & 1.00 & 1.00 & 0.67 & 0.33 \\ 0.67 & 0 & 1.00 & 0.67 & 0.67 & 1.00 & 0.67 \\ 0.33 & 0 & 0.67 & 0.33 & 0.33 & 0.67 & 1.00 \end{bmatrix};$$

$$[\mathrm{Sim}_{A_4}(o_i,o_j)]_{7\times7} = \begin{bmatrix} 1.00 & 0.84 & 0 & 0.89 & 0.76 & 0 & 0.55 \\ 0.84 & 1.00 & 0 & 0.95 & 0.92 & 0.17 & 0.71 \\ 0 & 0 & 1.00 & 1.00 & 0 & 1.00 & 1.00 \\ 0.89 & 0.95 & 0 & 1.00 & 0.87 & 0.11 & 0.66 \\ 0.76 & 0.92 & 0 & 0.87 & 1.00 & 0.24 & 0.79 \\ 0 & 0.17 & 0 & 0.11 & 0.24 & 1.00 & 0.45 \\ 0.55 & 0.71 & 0 & 0.66 & 0.79 & 0.45 & 1.00 \end{bmatrix}。$$

从属性 A_1 上各对象间的相似度矩阵可以看到，由于 $f(o_7, A_1) = *$，故 $\mathrm{Sim}_{A_1}(o_i, o_7) = 0$ (i=1, 2, …, 6)，$\mathrm{Sim}_{A_1}(o_7, o_j) = 0$ (j=1, 2, 4, 5)，$\mathrm{Sim}_{A_1}(o_7, o_j) = 1$ (j=3, 6, 7)，此时，$\mathrm{Sim}_C(o_i, o_j) \neq \mathrm{Sim}_C(o_j, o_i)$。

据相似阈值的优化计算，得相似阈值 $\alpha = 0.75$，此时，α-下相似广义划分结果为$\{\{o_1, o_4, o_5\}, \{o_1, o_2, o_5\}, \{o_3, o_6\}, \{o_7\}\}$，$\alpha$-上相似广义划分结果为$\{\{o_1, o_2, o_4, o_5\}, \{o_3, o_4, o_6, o_7\}\}$。显然，这是不同于等价划分的一种广义覆盖划分，不满足传递性。同时，通过计算得到 $\mathrm{Bel}(\{o_1, o_2, o_4, o_5\})=0.75$，$\mathrm{Bel}(\{o_3, o_4, o_6, o_7\})=0.50$。

通过分析计算，设定上、下近似阈值 $\beta \in (0.67, 0.75]$，则 $\underline{S}_{\alpha}^{\beta}(Y) = \{o_1, o_2, o_3, o_5, o_6, o_7\}$，$\overline{S}_{\alpha}^{\beta}(Y) = \{o_1, o_2, o_3, o_4, o_5, o_6, o_7\}$，$\underline{S}_{\beta}^{\alpha}(Y) = \{o_1, o_2, o_3, o_4, o_5, o_6, o_7\}$，$\overline{S}_{\beta}^{\alpha}(Y) = \{o_1, o_2, o_3, o_4, o_5, o_6, o_7\}$。

依据定义 7 给出的广义相似关系下的条件属性约简定义，可以得到原条件属性集相对于属性 D 的 α-下广义相似的 β 近似约简为 $A=\{A_1, A_3\}$，原条件属性集相对于属性 D 的 α-上广义相似的 β 近似约简为 $B=\{A_2, A_4\}$。

约简后可以提取到的所有规则为：

if $A_1 \in [37.1, 40.3]$ and $A_3 \in$ {有} then 中部烟 Cer=0.67; Cov=0.67；

if $A_1 \in [37.1, 40.3]$ and $A_3 \in$ {有} then 下部烟 Cer=0.33; Cov=0.25；

if $A_1 \in [39.9, 44.8]$ then 中部烟 Cer=1.0; Cov=1.0；

if $A_3 \in$ {稍有} then 下部烟 Cer=1.0; Cov=0.5；

if $A_3 \in$ {少} then 下部烟 Cer=1.0; Cov=0.25；

if $A_2 \in$ {浓, 强} and $A_4 \in [18.9\%, 21.7\%]$ then 中部烟 Cer=0.56; Cov=1.0；

if $A_2 \in$ {浓, 强} and $A_4 \in [18.9\%, 21.7\%]$ then 下部烟 Cer=0.19; Cov=0.25；

if $A_2 \in$ {淡, 弱, 浓} then 下部烟 Cer=0.5; Cov=0.38；

这样，该信息系统中所蕴涵的所有“可能”规则知识全部获取，并通过给出每条规则的确定度和覆盖度，使得这些规则知识以更加合理和全面的方式予以展示，有助于对该信息系统的理解和后续推理。

4 结 论

a. 为了最大限度地从含有杂合数据的不完备信息系统中提取出有效的规则知识，首先对杂合数据和不完备数据在广义相似关系下进行特殊处理，分别给出了上、下广义相似下的属性约简和规则抽取的定义与方法，最后，得到的产生式规则也采用确定度和覆盖度来予以衡量和评价，能方便、直观和理性地从该类信息系统中获取知识。

b. 所提出的广义相似概念是基于等价划分概念的延拓，而基于广义相似概念的粗糙集模型也是经典粗糙集模型的进一步延伸。这是对 Pawlak 粗糙集理论进行的扩展和改进，使得粗糙集理论适用于工程实际。对于普通的信息系统，它完全可以作为本文所研究的特例予以处理，故本文所提出的扩展粗糙集模型具有普适性。

参考文献：

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

[2] Slezak D, Ziarko W. Attribute reduction in the bayesian version of variable precision rough set model[J]. Electronic Notes in Theoretical Computer Science, 2003, 82(4): 1−11.

[3] Dubois D, Prade H. Rough-fuzzy sets and fuzzy-rough sets[J]. International Journal Gen Systems, 1990, 17(2): 191−209.

[4] 何亚群, 李继军, 胡寿松. 基于模糊相容关系下的相容模糊粗糙集[J]. 系统工程学报, 2006, 21(5): 553−556.
HE Ya-qun, LI Ji-jun, HU Shou-song. Compatibility fuzzy-rough sets based on fuzzy compatibility relation[J]. Journal of Systems Engineering, 2006, 21(5): 553−556.

[5] Bonikowski Z, Bryniarski E, Wybraniec U. Extension and intentions in the rough set theory[J]. Information Science, 1998, 107(1): 149−167.

[6] Zhu Willian. Topological approaches to covering rough sets[J]. Information Sciences, 2007, 177(6): 1499−1508.

[7] Zhu Willian, Wang Feiyue. A new type of covering rough sets[C]//Proc of the 3rd IEEE International Conf on Intelligent Systems. London, 2006: 444−449.

[8] Kryszkiewicz M. Rules in incomplete information systems[J]. Information Sciences, 1999, 113: 271−292.

[9] Jerzy S. Incomplete information tables and rough classification [J]. Computational Intelligence, 2001, 17(3): 545−566.

[10] 王国胤. Rough Set 理论与知识获取[M]. 西安: 西安交通大学出版社, 2001.
WANG Guo-yin. Rough set theory and knowledge acquisition[M]. Xi’an: Xi’an Jiaotong University Press, 2001.

[11] Yee L, WU Wei-zhi, ZHANG Wen-xiu. Knowledge acquisition in incomplete information systems: A rough set approach[J]. European Journal of Operational Research, 2006, 168(1): 164−180.

[12] LI Tian-rui, DA Ruan, Geert W, et al. A rough sets based characteristic relation approach for dynamic attribute generalization in data mining[J]. Knowledge-Based Systems, 2007, 20(5): 485−494.

[13] 瞿彬彬, 卢炎生. 基于限制非对称相似关系的粗糙集模型[J]. 小型微型计算机系统, 2007, 28(6): 1084−1088.
QU Bin-bin, LU Yan-sheng. Rough set model based on limited non-symmetric similarity relation[J]. Journal of Chinese Computer Systems, 2007, 28(6): 1084−1088.

[14] HU Qing-hua, LIU Jin-fu, YU Da-ren. Mixed feature selection based on granulation and approximation[J]. Knowledge-Based Systems, 2008, 21(4): 294−304.

[15] 张文修, 仇国芳. 基于粗糙集的不确定性决策[M]. 北京: 清华大学出版社, 2005.
ZHANG Wen-xiu, QIU Guo-fang. Uncertain decision making based on rough sets[M]. Beijing: Tsinghua University Press, 2005.