

不完备信息系统中的集对分析方法^{*}

杨习贝¹ 杨静宇¹ 吴 陈² 於东军¹

(南京理工大学计算机科学与技术学院 南京 210094)¹ (德瑞索大学信息科学与技术学院 费城 PA19104)²

摘 要 考虑了既有遗漏型又有丢失型未知属性值的不完备信息系统。在这种不完备信息系统中,采用集对分析的方法构建了一种新的基于联系度的二元关系,此关系仅满足自反性。在此基础上,分析了这种二元关系在两种特殊情形的不完备信息系统中的表现形式,然后根据新建立的基于联系度的粗糙集模型,使用决策矩阵讨论了规则生成方法,最后用一个实例说明了此方法的有效性。

关键词 不完备信息系统,集对分析,联系度,粗糙集,决策矩阵

Set Pair Analysis in Incomplete Information Systems

YANG Xi-Bei¹ YANG Jing-Yu¹ WU Chen² YU Dong-Jun¹

(Department of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094)

(College of Information Science and Technology, Drexel University, Philadelphia, PA 19104, USA)

Abstract In this paper, the incomplete information system in which unknown attributes values are considered as the coexisting of do not care and lost conditions. In this kind of incomplete information system, a new kind of binary relation based on relative degree is proposed according to the set pair analysis, this binary relation is only reflective. From what have been discussed above, expressive forms of this new binary relation in two special incomplete information systems are investigated. Based on the new-defined rough set model, the method of rule induction is studied from the way of decision matrix. Finally, an illustrative example is analyzed to indicate the validity of rough set model based upon set pair analysis.

Keywords Incomplete information system, Set pair analysis, Relative degree, Rough sets, Decision matrix

近年来,由波兰学者 Pawlak 提出的粗糙集理论^[1,2] (Rough Set Theory, 简称 RST) 在知识发现领域获得了巨大的成功。然而遗憾的是,经典 RST 的研究对象是具有离散属性值的完备信息系统,即信息系统中的所有属性值都是已知的。但是在现实世界中,由于数据测量的误差、对数据的理解或获取的限制等众多原因^[3],所面临的信息系统往往是不完备的,因此,使用 RST 从不完备信息系统^[3-6] (Incomplete Information System, 简称 IIS) 中挖掘知识,对于 RST 的发展具有举足轻重的作用。

对于一个不完备信息系统来说,其中的未知属性值可以分为两种情况:1)所有的未知属性值被认为是丢失的,是不允许被比较的;2)所有的未知属性值仅仅是被遗漏的,但又是确实存在的。文[7]首先从 RST 的角度研究了 IIS 中未知属性值都为丢失型的情形,然后 Stefanowski 等人构建了其中的相似关系^[5] (自反性、传递性);对于 IIS 中未知属性值都为遗漏型的情况, J. Grzymala-Busse 在文[8]中做了详细的阐述。接着, Kryszkiewicz 构建了满足自反性和对称性的容差关系^[4],并在此基础上研究了规则挖掘的方法。可以看出,无论是相似关系还是容差关系,它们都是基于这样一个假设,即 IIS 中的所有未知属性值仅具有一种可能的解释。换言之,以上 2 种较弱的二元关系都未考虑一种更为复杂的信息系统形式,即丢失型和遗漏型未知属性值共存的情况,因此本文将这种更为一般的 IIS 作为研究对象。

集对分析^[9] (Set Pair Analysis, 简称 SPA) 是我国学者赵克勤提出的一种处理不确定性问题的新型数学工具,它的核

心思想是把被研究的客观事务之确定性联系与不确定性联系作为一个确定不确定系统来分析和处理。目前国内虽然已有学者构建了 IIS 中基于联系度的拓展粗糙集模型^[10],但这种模型是将信息系统中的所有未知属性值认为是遗漏型来处理的。很明显,对于本文将要讨论的更为复杂的 IIS,则需要建立一种更为广义的联系度形式。另一方面,已有的基于联系度的粗糙集都是在理论模型上进行探讨,并未进行知识约简和规则生成方法的研究。

综上,本文的结构安排如下:第 1 节给出本文的处理对象,即同时出现了丢失型和遗漏型的未知属性值 IIS,并简要介绍了集对分析的概念;第 2 节在这种更为广义的 IIS 中定义了新的基于联系度的粗糙集模型,并在两种特殊情形的 IIS 中进行了详细的讨论;第 3 节研究了使用基于联系度的粗糙模型生成确定和可能规则的方法;第 4 节进行了实例分析;最后总结全文。

1 基本概念

1.1 不完备信息系统

一个不完备信息系统为一个四元组: $S = \langle U, AT, V, f \rangle$ 。其中, U 是一个被称为论域的非空有限的对象集合; AT 是非空有限的属性集合;对于 $\forall a \in AT$, 有 $a: U \rightarrow V_a$, 其中 V_a 是属性 a 的值域(可包丢失型、遗漏型空值,文中分别用“*”和“?”表示),属性值域集合 $V = \bigcup_{a \in AT} V_a$; f 为信息函数,对于 $\forall a \in AT, \forall x \in U$, 有 $f(x, a) \in V_a$ 。

表 1 是一个同时具有丢失和遗漏型未知属性值的 IIS。

^{*} 国家自然科学基金(No. 60472060, 60572034)、南京理工大学科研发展基金(No. AB96125)资助项目。杨习贝 博士研究生,研究方向为智能信息处理、粗糙集理论。

其中 $U = \{O_1, O_2, O_3, O_4, O_5, O_6\}$, $AT = \{a, b, c, d\}$, $V_a = \{a_1, a_2\}$, $V_b = \{b_1, b_2\}$, $V_c = \{c_1, c_2\}$, $V_d = \{d_1, d_2\}$ 。很明显, $f(O_5, a)$, $f(O_3, b)$ 是属于丢失型空值, 而 $f(O_3, a)$, $f(O_2, b)$, $f(O_5, b)$, $f(O_6, d)$ 是遗漏型空值。

表 1 不完备信息系统

U	a	b	c	d
O_1	a_1	b_1	c_1	d_1
O_2	a_2	*	c_1	d_1
O_3	*	?	c_2	d_2
O_4	a_1	b_2	c_1	d_2
O_5	?	*	c_1	d_2
O_6	a_2	b_1	c_1	*

上述的 IIS 有两种特殊情况: 设 S_1 是这样一个 IIS, 其中所有的未知属性值都是遗漏型空值 (“?”); 设 S_2 是这样一个 IIS, 其中所有的未知属性值都是丢失型空值 (“*”)。在 S_1 和 S_2 中可分别构建容差关系和相似关系如下。

定义 1^[4] 在 S_1 中, 对于 $\forall A \subseteq AT$, 由 A 决定的容差关系记为 $T(A)$ 且

$$T(A) = \{(x, y) \in U^2 : \forall a \in A, f(x, a) = f(y, a) \vee f(x, a) = * (f(y, a) = *)\} \quad (1)$$

在不完备信息系统 S 中, 对于 $\forall x \in U$, x 的容差类可记为 $T^A(x)$ 且 $T^A(x) = \{y \in U : (x, y) \in T(A)\}$, 即 $T^A(x)$ 是所有与 x 具有容差关系 $T(A)$ 的对象的集合。

定义 2^[6] 在 S_2 中, 对于 $\forall A \subseteq AT$, 由 A 决定的相似关系记为 $S(A)$ 且

$$S(A) = \{(x, y) \in U^2 : \forall a \in A, f(x, a) = ? \vee f(x, a) = f(y, a)\} \quad (2)$$

根据定义 2 所示的相似关系, Stefanowski 给出了非对称相似于 x 的集合 $S_A(x) = \{y \in U : (y, x) \in S(A)\}$, x 与之非对称相似的集合 $S_A^{-1}(x) = \{y \in U : (x, y) \in S(A)\}$ 。

1.2 集对分析

定义 3 设集合 M_1 与 M_2 组成集对 $H(M_1, M_2)$, 共有 N 个属性, 其中 M_1 与 M_2 在 S_1 个属性上取值相同, 在 S_2 个属性上取值不明确, 在 S_3 个属性上取值不同, 则称 S_1/N 为集对的同度, S_2/N 为集对的差异度, S_3/N 为集对的对立度, 则集合 M_1 与 M_2 的联系度记为 $\mu(M_1, M_2)$ 且

$$\mu(M_1, M_2) = \frac{S_1}{N} + \frac{S_2}{N}i + \frac{S_3}{N}j = a + bi + cj \quad (3)$$

其中 i 表示差异度的系数且 $i \in [-1, 1]$, j 表示对立度的系数且 $j \in [-1, 1]$ 。

很明显, 在定义 3 中, 有 $0 \leq a, b, c \leq 1$ 且 $a + b + c = 1$ 。更为一般地, 当系统中不确定性量 b_i 含有多种不确定性时, 还可以把 b_i 直接分解为 $b_{i1}i_1, b_{i2}i_2, \dots, b_{in}i_n$, 以分别表示各种不同类型的不确定性, 如随机不确定性、模糊不确定性、中介不确定性、由不知道所引起的不确定性等等^[9]。联系度的这种拓展性质是下文建立 IIS 中基于联系度的粗集模型的基础, 因为在本文所研究的不完备信息系统中, 既有由于属性值不知道所引起的不确定性, 又有由于属性值不完全而导致的的不确定性。

2 基于联系度的拓展粗集

2.1 拓展粗集模型

对于同时具有丢失和遗漏型的不完备信息系统 S 来说,

要使用集对分析的方法讨论对象之间的关系, 就必须建立一种更为广义的联系度的形式, 因此我们对 S 中的不确定性做进一步的细分。形式化地, 在不完备信息系统 S 中, 若有 $A \subseteq AT$, 记 $P_A(x) = \{a \in A : f(x, a) \neq ?\}$, 对于 $\forall x, y \in U$, x 与 y 之间的差异度可以从以下几方面考虑:

- (1) $T_1 = \{a \in P_A(x) : f(x, a) = * \wedge f(y, a) = *\}$;
- (2) $T_2 = \{a \in P_A(x) : f(x, a) = * \wedge f(y, a) \text{ 已知}\}$;
- (3) $T_3 = \{a \in P_A(x) : f(x, a) \text{ 已知} \wedge f(y, a) = *\}$;
- (4) $T_4 = \{a \in P_A(x) : f(x, a) \text{ 已知} \wedge f(y, a) = ?\}$;
- (5) $T_5 = \{a \in P_A(x) : f(x, a) = * \wedge f(y, a) = ?\}$ 。

对于情况 (1), (2), (3) 来说, 可以认为是由于信息不完全所引起的不确定, 对于情况 (4), 可以认为是由于信息不知道所引起的不确定, 而对于情况 (5), 可以认为是由于信息不完全和不知道共同引起的不确定, 因此构建联系度如定义 4 所示。

定义 4 在不完备信息系统 S 中, 有 $A \subseteq AT$, $x \in U$, 对于 $\forall y \in U$, x 与 y 的联系度记为 $\mu^A(x, y)$ 且

$$\mu^A(x, y) = a + b_1 \cdot i_1 + b_2 \cdot i_2 + b_3 \cdot i_1 \cdot i_2, a + b_1 + b_2 + b_3 = 1 \quad (4)$$

其中 $a = |\{a \in A : f(x, a) \text{ 已知} \wedge f(y, a) \text{ 已知} \wedge f(x, a) = f(y, a)\}| / |P_A(x)|$, $b_1 = |T_1 \cup T_2 \cup T_3| / |P_A(x)|$, $b_2 = |T_4| / |P_A(x)|$, $b_3 = |T_5| / |P_A(x)|$, $\cdot$ 表示集合的基数。

很明显, 定义 4 中所示的联系度仅仅满足自反性, 而对称性和传递性 μ^A 则并不一定满足。特别地, 当 $b_2 = b_3 = 0$, $i_1 = 1$ 时, 联系度就退化为仅满足自反性的二元关系如下所示:

$$R(A) = \{(x, y) \in U^2 : \forall a \in A \wedge f(x, a) \neq ?, f(x, a) = f(y, a) \vee f(x, a) = * \vee f(y, a) = *\} \quad (5)$$

关系 $R(A)$ 正是文 [11] 所建立的特征关系, 特征关系是容差关系和相似关系的一种更为广义的表现形式, 其中既保留了容差关系的相关性质, 也保留了非对称相似关系的相关性质。当 IIS 中的所有未知属性值都为遗漏型 “?” 时, 特征关系 $R(A)$ 就退化为 Kryszkiewicz 所建立的容差关系 $T(A)$, 而当 IIS 中的所有未知属性值都为遗漏型 “*” 时, 特征关系 $R(A)$ 则退化为 Stefanowski 所建立的非对称关系 $S(A)$ 。因而, 联系度 $\mu^A(x, y)$ 是特征关系 $R(A)$ 的一种推广化形式, 而 $R(A)$ 则是 $\mu^A(x, y)$ 的一个特例。

定义 5 在不完备信息系统 S 中, 有 $A \subseteq AT$, 对于 $x \in U$, 则 x 基于联系度 $\mu^A(x, y)$ 的邻域记为 $[x]_\lambda^A$ 且 $[x]_\lambda^A = \{y \in U : \mu^A(x, y) \wedge a \geq \lambda\}$, 其中 $\lambda \in [0, 1]$ 。

在定义 5 中, $[x]_\lambda^A$ 表示与 x 没有不同的已知属性值且与 x 有相同已知属性值的个数在 $P_A(x)$ 中的比例不低于 λ 的对象的集合。

定义 6 设 S 为一不完备信息系统, 有 $A \subseteq AT$, 对于 $\forall X \subseteq U$, X 基于联系度 μ^A 的下、上近似集分别记为 $\underline{A}^a(x)$, $\overline{A}^a(x)$ 且

$$\begin{aligned} \underline{A}^a &= \{x \in U : [x]_\lambda^A \subseteq X\}; \overline{A}^a(X) \\ &= \{x \in U : [x]_\lambda^A \cap X \neq \emptyset\} \end{aligned} \quad (6)$$

2.2 特殊情形分析

在第 1 节中, 我们已经分析了不完备信息系统 S 具有两种极端情况, 即 S_1 和 S_2 。首先, 在 S_1 中, 所有的未知属性值都是遗漏型 (“?”), 此时 x 与 y 的联系度变为

$$\mu^A(x, y) = a + b_1 \cdot i_1, a + b_1 = 1, \quad (7)$$

其中 $a = |\{a \in A : f(x, a) \text{ 已知} \wedge f(y, a) \text{ 已知} \wedge f(x, a) = f(y, a)\}| / |A|$, $b_1 = |T_1 \cup T_2 \cup T_3| / |A|$ 。

此时 $[x]_\lambda^A = \{y \in U : \mu^A(x, y) \wedge a \geq \lambda\}$, 这恰好就是文

[10]所提出的基于联系度的拓展容差关系,它满足自反性和对称性,而传递性则不成立。

在文[3]中,王国胤教授认为容差关系的要求过于宽松,容易将两个没有任何已知相等属性值的对象划分到同一类中,而在基于联系度 μ^A 所形成的对象邻域中,由于约束条件 $a \geq \alpha$ 规定了 x 与 y 之间必定要有相同的已知属性值,并且两者具有相同已知属性值的个数在整体属性中所占的比例要大于某一给定的阈值 α ,因此联系度 μ^A 是容差关系的一种改进形式;当 $\alpha=0$ 时,联系度 μ^A 就退化为容差关系 $T(A)$ 。

另一方面,在 S_2 中,当所有的未知属性值都是丢失型(“?”)的时候,则 x 与 y 的联系度变为

$$\mu^A(x, y) = a + b_2 \cdot i_2 \quad a + b_2 = 1, \quad (8)$$

其中 $a = |\{a \in A : f(x, a) \text{ 已知} \wedge f(y, a) \text{ 已知} \wedge f(x, a) = f(y, a)\}| / |P_A(x)|$, $b_2 = |T_4| / |P_A(x)|$ 。

此时 $[x]_A^{\mu^A} = \{y \in U : \mu^A(x, y) \wedge a \geq \alpha\}$,很明显,此时的联系度满足自反性和传递性,而对称性则不满足。特别地,若设 $b_2=0$ 时,联系度 $\mu^A(x, y)$ 就退化为文所提出的相似关系,因为此时约束条件 $a=1$ 规定了 y 不可能比 x 更不完备。

在文[3]中,王国胤教授认为非对称相似关系的要求过于严格,容易将两个具有大量已知属性值的对象分离开来。例如 $x = \{?, 1, 0, 2, 3\}$, $y = \{1, ?, 0, 2, 3\}$,若用非对称相似关系来分析这两个对象,很明显这两者之间是不相似的。然而从直观上来看, x 与 y 具有较多相等的已知属性值,因此非对称相似关系的分类方式是不尽合理的。而此时若用联系度 μ^A 来分析,假设阈值 $\alpha=0.6$,则有 $a=0.75$, $b_2=0.25$,此时就有 $y \in [x]_A^{\mu^A}$ 和 $x \in [y]_A^{\mu^A}$ 。通过以上分析,可以发现,只要合理地调整阈值 α 的设置,使用联系度 μ^A 对于非对称相似关系来说,具有更为理想的分类结果。

3 规则的生成

粗糙集理论的核心问题是知识约简与知识推理,通过知识约简,可以得到简化规则以更好地指导决策。但是对于一个信息系统来说,可能存在有多个约简,并且不同的约简可以生成不同的决策规则,在实际应用中,很难确定选择哪个约简导出的规则。鉴于此,笔者在文[12]方法的基础上,改进了决策矩阵的形式,利用决策类的下、上近似,可以不用计算所有约简而得到所有的确定和可能决策规则。

定义 7 一个决策信息系统 $S = \langle U, AT \cup D, V, f \rangle$, 其中 AT 为条件属性集合, D 为决策属性集合,且 $AT \cap D \neq \emptyset$ 。设 $A \subseteq AT$ 且 $U/IND(D) = \{D_1, D_2, \dots, D_q\}$, 则决策类 D_i 的决策矩阵形式化定义为 $M_{x,y}^i$, 且

$$M_{x,y}^i = \begin{cases} (a, f(x, a)), f(x, a) \text{ 已知} \wedge f(y, a) \text{ 已知} \wedge f(x, a) \neq f(y, a) \\ \emptyset, \text{ 否则} \end{cases}$$

其中 $x \in \bar{A}^-(D_i)$, $y \in U - D_i$ ($1 \leq i \leq q$), $a \in P_A(x)$ 。

决策矩阵表示把某类对象与其他对象分离开来的所有已知信息,利用决策矩阵计算 Boolean 表达式 $B_i = \bigwedge_y \bigvee_x M_{x,y}^i$, $\bigwedge$ 和 $\bigvee$ 分别表示合取和析取运算, B_i 称为知识表示系统的决策函数。将决策函数化为析取范式,如利用 Boolean 代数的吸收律化简,其中每个合取子式对应一条规则。重复上述步骤,就可以得到 D_i 类的所有确定决策规则集。

若要生成可能决策规则,只需要修改决策矩阵中的对象 $x \in \bar{A}^-(D_i)$, $y \in U - \bar{A}^-(D_i)$ 即可。

4 实例分析

表 2 是一个不完备决策信息系统,其中 a, b, c, d 为条件属性, e 为决策属性。其中 $U = \{O_i | (1 \leq i \leq 12)\}$, $AT = \{a, b, c, d\}$ 为全体条件属性集合, e 为决策属性集合, $U/IND(e) = \{\Phi, \Psi\} = \{\{O_1, O_2, O_4, O_7, O_{10}, O_{12}\}, \{O_3, O_5, O_6, O_8, O_9, O_{11}\}\}$ 。

表 2 不完备决策信息系统

	a	b	c	d	e
O_1	3	2	1	0	Φ
O_2	2	3	2	0	Φ
O_3	2	3	2	0	Ψ
O_4	*	2	*	1	Φ
O_5	*	?	*	1	Ψ
O_6	2	3	2	1	Ψ
O_7	3	?	*	3	Φ
O_8	*	0	0	*	Ψ
O_9	3	2	1	3	Ψ
O_{10}	1	*	*	?	Φ
O_{11}	*	2	*	*	Ψ
O_{12}	3	2	1	*	Φ

若设 $\alpha=0.5$, 则可得到分类结果如下所示: $[O_1]_{AT}^{0.5} = \{O_1, O_{12}\}$, $[O_2]_{AT}^{0.5} = \{O_2, O_3\}$, $[O_3]_{AT}^{0.5} = \{O_2, O_3\}$, $[O_4]_{AT}^{0.5} = \{O_4\}$, $[O_5]_{AT}^{0.5} = \{O_5\}$, $[O_6]_{AT}^{0.5} = \{O_6\}$, $[O_7]_{AT}^{0.5} = \{O_7, O_9\}$, $[O_8]_{AT}^{0.5} = \{O_8\}$, $[O_9]_{AT}^{0.5} = \{O_7, O_9, O_{12}\}$, $[O_{10}]_{AT}^{0.5} = \{O_{10}\}$, $[O_{11}]_{AT}^{0.5} = \{O_{11}\}$, $[O_{12}]_{AT}^{0.5} = \{O_1, O_9, O_{12}\}$ 。于是就有 $AT^{0.5}(\Phi) = \{O_1, O_4, O_{10}\}$, $AT^{0.5}(\Psi) = \{O_5, O_6, O_8, O_{11}\}$, $\overline{AT^{0.5}}(\Phi) = \{O_1, O_2, O_3, O_4, O_7, O_9, O_{10}, O_{12}\}$, $\overline{AT^{0.5}}(\Psi) = \{O_2, O_3, O_5, O_6, O_7, O_8, O_9, O_{11}\}$ 。

根据以上结果,可以得到如表 3 所示的决策矩阵。

表 3 决策类 Φ 的决策矩阵(确定规则)

	O_3	O_5	O_6	O_8	O_9	O_{11}
O_1	$(a, 3)(b, 2)(c, 1)$	$(d, 0)$	$(a, 3)(b, 2)(c, 1)(d, 0)$	$(b, 2)(c, 1)$	$(d, 0)$	$\emptyset$
O_4	$(b, 2)(d, 1)$	$\emptyset$	$(b, 2)$	$(b, 2)$	$(d, 1)$	$\emptyset$
O_{10}	$(a, 1)$	$\emptyset$	$(a, 1)$	$\emptyset$	$(a, 1)$	$\emptyset$

表 4 决策类 Φ 的决策矩阵(可能规则)

	O_5	O_6	O_8	O_{11}
O_1	$(d, 0)$	$(a, 3)(b, 2)(c, 1)(d, 0)$	$(b, 2)(c, 1)$	$\emptyset$
O_2	$(d, 0)$	$(d, 0)$	$(b, 3)(c, 2)$	$(b, 3)$
O_3	$(d, 0)$	$(d, 0)$	$(b, 3)(c, 2)$	$(b, 3)$
O_4	$\emptyset$	$(b, 2)$	$(b, 2)$	$\emptyset$
O_7	$(d, 3)$	$(a, 3)(d, 3)$	$\emptyset$	$\emptyset$
O_9	$(d, 3)$	$(a, 3)(b, 2)(c, 1)(d, 3)$	$(b, 2)(c, 1)$	$\emptyset$
O_{10}	$\emptyset$	$(a, 1)$	$\emptyset$	$\emptyset$
O_{12}	$\emptyset$	$(a, 3)(b, 2)(c, 1)$	$(b, 2)(c, 1)$	$\emptyset$

于是可以得到如下所示简化的决策规则:

$(b, 2) \wedge (d, 0) \rightarrow (e, \Phi)$; $(c, 1) \wedge (d, 0) \rightarrow (e, \Phi)$; $(b, 2) \wedge (d, 1) \rightarrow (e, \Phi)$; $(a, 1) \rightarrow (e, \Phi)$

若要得到可能规则,则求出的决策矩阵如表 4 所示,得到的可能决策规则为

$(b, 2) \vee (c, 1) \rightarrow (e, \Phi)$; $(b, 3) \wedge (d, 0) \rightarrow (e, \Phi)$; $(b, 2) \wedge (d, 3) \rightarrow (e, \Phi)$; $(c, 1) \wedge (d, 3) \rightarrow (e, \Phi)$ 。

类似上述过程,可以得到关于决策类 Ψ 的确定规则为

$(d,1) \rightarrow (e, \Psi); (a,2) \wedge (b,3) \wedge (d,1) \rightarrow (e, \Psi); (b,0) \vee (b,2) \rightarrow (e, \Psi)$

关于决策类 Ψ 的可能规则为

$(a,2) \wedge (b,3) \rightarrow (e, \Psi); (a,2) \wedge (d,0) \rightarrow (e, \Psi); (a,3) \wedge (d,3) \rightarrow (e, \Psi); (b,0) \rightarrow (e, \Psi)$

结束语 粗糙集理论发展至今,以其独特的优势在智能信息处理领域占据着重要的地位。由于现实世界中广泛存在的数据不完备性,促进了各种拓展粗糙集模型的产生与发展。本文所考虑的不完备信息系统既有遗漏型,又有丢失型未知属性值,综合了以往学者所讨论的不完备信息系统中未知属性值所具有两种解释,因此具有更为广泛的意义。在此基础上,对信息系统中的数据不完备性进行了讨论,使用集对分析的方法建立了一种新的拓展粗糙集模型,并与传统的几种拓展二元关系进行了对比分析。最后,研究了不进行约简而直接从信息系统中推导出确定和可能性规则的方法,并用实例进行了验证。

在本文所做工作的基础上,下一步的工作就是对挖掘出来规则的可信度进行度量。另一方面,亦可进行广义不完备信息系统中高阶规则,即知识依赖的挖掘方法,并对部分知识依赖的度量方法进行更为深入的研究。

参考文献

- 1 Pawlak Z. Rough set theory and its applications to data analysis [J]. Cybernetics and Systems, 1998, 29: 661~688

- 2 Pawlak Z. Rough sets and intelligent data analysis [J]. Information Sciences, 2002, 147: 1~12
- 3 王国胤. 粗糙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
- 4 Kryszkiewicz M. Rough set approach to incomplete information systems [J]. Information Sciences, 1998, 112: 39~49
- 5 Stefanowski J, Tsoukias A. Incomplete information tables and rough classification [J]. Computational Intelligence, 2001, 17: 545~566
- 6 Wu Weizhi, Zhang Wenxiu, Li Huaizu. Knowledge acquisition in incomplete fuzzy information systems via the rough set approach [J]. Expert Systems, 2003, 20(5): 280~286
- 7 Grzymala-Busse J W. On the unknown attribute values in learning from examples. In: Proceeding of the 6th International Symposium on Methodologies for Intelligent Systems (ISMIS-91), Charlotte, North Carolina, October, 1991. Lecture Notes in Artificial Intelligence. vol 542. Springer-Verlag, Berlin, Heidelberg, New York, 1991. 368~377
- 8 Grzymala-Busse J W, Wang A Y. Modified algorithms LEM1 and LEM2 for rule induction from data with missing attribute values. In: Proceeding of the Fifth International Workshop on Rough Sets and Soft Computing (RSSC'97) at the Third Joint Conference on Information Sciences (JCIS'97), Research Triangle Park, NC, March, 1997. 69~72
- 9 赵克勤. 集对分析及其初步应用[M]. 杭州: 浙江科学出版社, 2000
- 10 黄兵, 周献中. 不完备信息系统中基于联系度的粗糙模型拓展[J]. 系统工程理论与实践, 2004, 24(1): 88~92
- 11 Grzymala-Busse J W. Data with Missing Attribute Values: Generalization of Indiscernibility Relation and Rule Induction. In: Transactions on Rough Sets I, Lecture Notes in Computer Science. vol 3100. Springer-Verlag, Berlin, 2004. 78~95
- 12 Shan Ning, Ziarko W. Data-based acquisition and incremental modification of classification rules [J]. Computational Intelligence, 1995, 11(2): 357~370

(上接第 165 页)

图 6 为第一组数据作出的关系图。

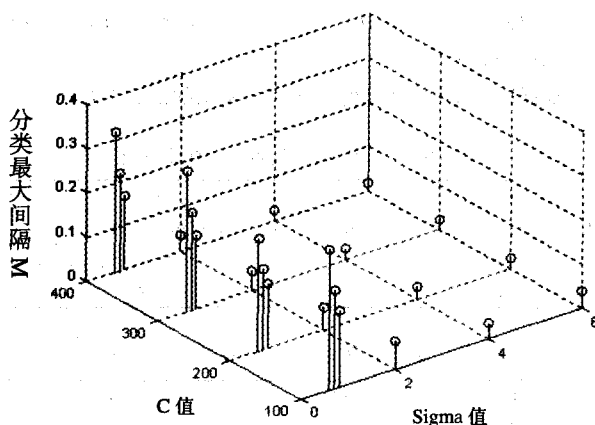


图 6

从图 6 可以看出,当 σ 值大于 2 时,最大分类间隔下降,影响分类效果。而 C 值的变化对分类间隔有一定的影响,但不及 σ 的影响大。图中还表明,取较小 σ 的值有利于提高分类机的最大分类间隔。

图 7 为第二组数据作出的关系图。

从图 7 可以看出,当 σ 值大于 2 时,支持向量分类机中支持向量所占的百分比 SV 上升,构造的分类机分类效用下降。而 C 值的变化对 SV 有一定的影响,但不及 σ 的影响大,取较小的 σ 值有利于降低构造支持向量分类机所需的支持向量数,从而使分类机更优化。

结论 (1) 经过证明,确定了如果一个函数是变元可分离的,在 Hilbert 空间中,存在 $K(u, v) = k(u)k(v)$ 关系则可以满

足 Mercer 条件,可以成为核函数,成为选择新的非线性支持向量分类机核函数的一种新方法。(2) 根据本例选定的可变量核函数构造的非线性支持向量分类机可以实现较好的分类,参数 σ 的选择应该小一些, C 值应该较大些,但不宜趋于无穷。 σ 对分类机的影响较敏感,选定时应注意。

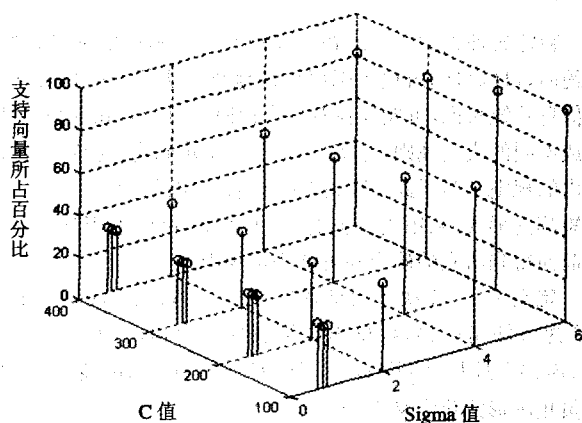


图 7

参考文献

- 1 宋晓峰, 陈德钊, 俞欢军, 胡上序. 支持向量机中的优化算法. 计算机科学, 2003(1)
- 2 邓乃扬, 田英杰. 数据挖掘中的新方法—支持向量机. 北京: 科学出版社, 2004, 6
- 3 张学工译. 统计学习理论的本质. 北京: 清华大学出版社, 2000, 9
- 4 <http://www.kernel-machines.org>
- 5 Boser B, Guyon I, Vapnik V. A training algorithm for optimal margin classifiers. In: Fifth Annual Workshop on Computer Learning Theory. Pittsburgh: ACM Press, 1992