

不完备灰色信息系统的粗集模型

林耀进¹, 李进金¹, 吴顺祥², 周忠眉¹

(1. 漳州师范学院 计算机科学与工程系, 福建 漳州 363000; 2. 厦门大学 自动化系, 福建 厦门 361005)

(zzlinyaojin@163.com)

摘要:提出一种属性值为区间灰数的不完备信息系统。首先根据区间灰数的定义,提出了区间灰数的一些运算性质,并定义了不完备灰色信息系统;然后,根据灰相似度,提出了变精度灰相似关系,并根据变精度灰相似关系引出了上、下近似算子;最后给出了约简的实际操作方法,并通过典型例子验证了该方法的有效性。

关键词:粗糙集;灰数;灰色信息系统;灰相似关系;知识约简

中图分类号: TP18 **文献标志码:** A

Rough set model in incomplete grey information systems

LIN Yao-jin¹, LI Jin-jin¹, WU Shun-xiang², ZHOU Zhong-mei¹

(1. Department of Computer Science and Engineering, Zhangzhou Normal University, Zhangzhou Fujian 363000, China;

2. Department of Automation, Xiamen University, Xiamen Fujian 361005, China)

Abstract: In this paper, the incomplete information system in which attributes values were considered as interval grey number. Firstly, some operation properties of interval grey number were presented according to the definition of interval grey number, and an incomplete grey information system was defined. Then, based on grey similar degree, the variable precise grey similar relation was proposed, and the upper and lower approximations were conducted according to variable precise grey similar relation. Lastly, a practical operation method for computing reduction was given, and the validity of the proposed technique was verified via a typical case.

Key words: rough set; grey number; grey information system; grey similar relation; knowledge reduction

0 引言

粗糙集^[1-2]是一种处理含糊和不精确的新型数学工具,在知识发现领域获得了巨大的成功。然而遗憾的是,经典粗糙集的研究对象是具有离散属性值的完备信息系统,即信息系统中每个对象的属性值都是已知的。但在现实中,由于数据测量的误差、对数据的理解或者获取的限制等各种原因,所获取的信息系统是不完备^[3-6]。不完备信息系统是指存在未知属性值的系统,对于未知属性值有以下几种解释:1)所有未知属性值是遗漏的,允许被比较,据此,文献[7]中提出了容差关系;2)所有未知属性值是丢失的,不允许被比较,据此,文献[8]中提出了非对称相似关系;3)文献[9-10]中考虑了未知属性值既有遗漏的也有丢失的。

不管属性值是已知还是未知,本文讨论了属性值为灰数的不完备信息系统,主要存在以下四种:仅有下界的灰数、仅有上界的灰数、区间灰数、黑数。现实生活中,信息系统属性值为灰数是广泛存在的,例如,2009年发生甲流时期,需要得到疑似病人的确切体温,然而由于各种原因,一般获取的是一段时间内的温度,这种外延明确、内涵不明确的属性值即为灰数^[11-12]。另外,由于灰数是指只知道大概范围而不知其确切值的数。可见,灰数实际上是一个数,既有下界又有上界的灰数称为区间灰数。而区间值是一个集合,其真值均匀分布于整个区间,所以灰数与区间值存在本质区别,则不完备灰色信息系统不同于区间值信息系统。文献[13-15]对区间值信

息系统进行了深入研究,提出的优势关系只是简单依据区间值的上下限进行比较,并没有考虑区间值之间的相似程度。因此,本文根据区间灰数的测度,提出了变精度的灰相似关系,对扩展区间值信息系统的应用有着重要意义。

1 基本概念

定义1^[11] 灰数是指在某个区间或某个一般的数集内取值的不确切的数。

本文中将用到下面几类灰数。

1) 仅有下界的灰数:有下界而无上界的灰数记为 $\otimes \in [a, \infty)$, 其中 a 为灰数 \otimes 的下确界。为了下文表达方便,简称为下灰数。

2) 仅有上界的灰数:有上界而无下界的灰数记为 $\otimes \in (\infty, \bar{b}]$, 其中 \bar{b} 为灰数 \otimes 的上确界。为了下文表达方便,简称为上灰数。

3) 区间灰数:既有下界又有上界的灰数称为区间灰数,记为 $\otimes \in [\underline{a}, \bar{b}]$, 其中 $\bar{b} > \underline{a}$ 。

4) 黑数:当 $\otimes \in (-\infty, +\infty)$ 时,即当 \otimes 的上、下界皆为无穷时,称 \otimes 为灰数。

定义2^[11] 设灰数 \otimes 的发生背景为 Ω , $u(\otimes)$ 表示灰数 \otimes 的测度,若 \otimes 为连续型区间灰数,即 $\otimes \in [\underline{a}, \bar{b}]$, 则 $u(\otimes) = \bar{b} - \underline{a}$;若 \otimes 为离散灰数,则 $u(\otimes)$ 定义为 \otimes 中全体离散白化默认数的个数。

显然,当 \otimes 为白数时, $u(\otimes) = 0$ 。

收稿日期: 2010-06-30; **修回日期:** 2010-07-23。 **基金项目:** 国家自然科学基金资助项目(10971186;11061004),福建省教育厅A类重点项目(JA10202),漳州师范学院科研基金资助项目(SK09005)。

作者简介: 林耀进(1980-),男,福建漳浦人,讲师,硕士,主要研究方向:粗糙集、灰色系统、数据挖掘; 李进金(1960-),男,福建泉州人,教授,博士,主要研究方向:粗糙集; 吴顺祥(1966-),男,湖南邵阳人,教授,博士,主要研究方向:灰色系统、智能信息处理;周忠眉(1966-),女,浙江温州人,教授,博士,主要研究方向:数据挖掘。

定义3 设灰数 $\otimes_1 \in [\underline{a}, \bar{b}]$, $\underline{a} < \bar{b}$; $\otimes_2 \in [\underline{c}, \bar{d}]$, $\underline{c} < \bar{d}$; 根据灰数的定义, 给出灰数之间的一些运算性质如下。

- 1) 区间灰数一致关系: $\otimes_1 \cong \otimes_2$ 当且仅当 $\underline{a} = \underline{c}$ 且 $\bar{b} = \bar{d}$ 。
- 2) 区间灰数包含关系: $\otimes_1 \triangleleft \otimes_2$ 当且仅当 $\underline{a} \geq \underline{c}$ 且 $\bar{b} \leq \bar{d}$ 。
- 3) 并运算 $(\otimes_1 \cup \otimes_2)$: $\otimes_1 \cup \otimes_2 \in [\min(\underline{a}, \underline{c}), \max(\bar{b}, \bar{d})]$ 。
- 4) 交运算 $(\otimes_1 \cap \otimes_2)$: 如果 $\underline{c} \leq \underline{a}$ 且 $\bar{b} \leq \bar{d}$, $\otimes_1 \cap \otimes_2 \in [\underline{a}, \bar{b}]$; 如果 $\underline{a} \leq \underline{c}$ 且 $\bar{d} \leq \bar{b}$, $\otimes_1 \cap \otimes_2 \in [\underline{c}, \bar{d}]$; 如果 $\underline{c} \leq \underline{a}$ 且 $\underline{a} \leq \bar{d} \leq \bar{b}$, $\otimes_1 \cap \otimes_2 \in [\underline{a}, \bar{d}]$; 如果 $\underline{a} \leq \underline{c} \leq \bar{b}$ 且 $\underline{c} \leq \bar{b} \leq \bar{d}$, $\otimes_1 \cap \otimes_2 \in [\underline{c}, \bar{b}]$; 其余情况, $\otimes_1 \cap \otimes_2 \in \emptyset$ 。

5) 灰直径: $dia(\otimes) = \bar{\otimes} - \underline{\otimes}$, 其中 $\otimes \in [\underline{\otimes}, \bar{\otimes}]$ 。

6) 灰相似度: $s(\otimes_1, \otimes_2) = \frac{dia(\otimes_1 \cap \otimes_2)}{dia(\otimes_1)}$, 其中当 $\otimes_1 \cap \otimes_2 \in \emptyset$, $s(\otimes_1, \otimes_2) = 0$; 其中当 $\otimes_1 \cong \otimes_2$, $s(\otimes_1, \otimes_2) = 1$ 。

注1 文献[15]对区间灰数 \otimes_1, \otimes_2 的相似度定义为 $s'(\otimes_1, \otimes_2) = \frac{dia(\otimes_1 \cap \otimes_2)}{dia(\otimes_1 \cup \otimes_2)}$, 存在一定的错误, 比如 $\otimes_1 = [1, 20]$, $\otimes_2 = [19, 21]$, $\otimes_3 = [1, 3]$, $\otimes_4 = [2, 21]$, 易得 $s'(\otimes_1, \otimes_2) = s'(\otimes_3, \otimes_4)$, 但是 \otimes_2 包含于 \otimes_1 最多达到 $1/20$, 而 \otimes_4 包含于 \otimes_3 最多可达到 $1/3$, 所以用本文定义的相似度清晰地描述了两个区间灰数之间的相似程度。

2 不完备灰色信息系统的变精度灰相似关系

2.1 不完备灰色信息系统

在本文中, 将属性值为灰数的不完备信息系统称为不完备灰色信息系统, 其定义如下。

定义4 称 $GI = (U, A, V, f)$ 是不完备灰色信息系统, 若 U 为非空有限对象集, A 为非空有限属性集, $V = \bigcup_{a \in A} V_a$ 且 V_a 是属性 a 的值域, $f: U \rightarrow V$ 为对象属性值映射, 且 V_a 是个灰数。

根据定义4, 将表1称为不完备灰色信息系统。

表1 不完备灰色信息系统

U	a_1	a_2	a_3	a_4
x_1	[0.3, 0.6]	[0.1, 0.4]	$(-\infty, 0.6]$	[0.5, 0.8]
x_2	[0.6, 0.9]	[0.3, 0.5]	[0.4, 0.7]	[0.6, 0.8]
x_3	[0.4, 0.6]	[0.2, $+\infty$)	[0.2, 0.5]	$(-\infty, +\infty)$
x_4	[0.4, 0.7]	[0.2, 0.4]	[0.3, 0.6]	[0.5, 0.6]
x_5	[0.5, 0.8]	[0.3, 0.4]	[0.3, 0.7]	[0.5, 0.7]
x_6	[0.6, 0.8]	[0.3, 0.4]	[0.4, 0.7]	[0.5, 0.8]

在表1中, $GI = \langle U, A, V, f \rangle$, 其中: $U = \{x_1, x_2, x_3, x_4, x_5, x_6\}$, $A = \{a_1, a_2, a_3, a_4\}$, 另外, 在表1中, 存在四种灰数, 分别为区间灰数, 如 $a_1(x_2)$; 黑数, 如 $a_4(x_3)$; 下灰数, 如 $a_2(x_3)$; 上灰数, 如 $a_3(x_1)$ 。其中, $a(x)$ 表示对象 x 在属性 a 的取值。

在不完备灰色信息系统中, 主要存在三种属性值未知的数据, 分别为: 下灰数、上灰数、黑数。对该三种属性值未知的数据可以分别进行如下处理。

- 1) $\forall x \in U, a \in A$, 如果 $a(x) \in [\underline{a}, \infty)$, 认为仅有下界的灰数只是上界被遗漏了, 但实际上上界是存在的, 于是, 其上界值可以取该属性值域上区间灰数中的上界最大值来取代;
- 2) $\forall x \in U, a \in A$, 如果 $a(x) \in (-\infty, \bar{b}]$, 认为仅有上界的灰数只是下界被遗漏了, 但实际上下界是存在的, 于是, 其下界值可以取该属性值域上区间灰数中的下界最小值来取代;
- 3) $\forall x \in U, a \in A$, 如果 $a(x) \in (-\infty, +\infty)$, 认为黑数

只是上界与下界同时被遗漏了, 但实际上上下界是存在的, 于是, 其上界值可以取该属性值域上区间灰数中的上界最大值来取代, 其下界值可以取该属性值域上区间灰数中的下界最小值来取代。

表1所示的不完备灰色信息系统经过处理之后转换为表2所示的完备灰色信息系统。

表2 对表1转换后的完备灰色信息系统

U	a_1	a_2	a_3	a_4
x_1	[0.3, 0.6]	[0.1, 0.4]	[0.2, 0.6]	[0.5, 0.8]
x_2	[0.6, 0.9]	[0.3, 0.5]	[0.4, 0.7]	[0.6, 0.8]
x_3	[0.4, 0.6]	[0.2, 0.5]	[0.2, 0.5]	[0.5, 0.8]
x_4	[0.4, 0.7]	[0.2, 0.4]	[0.3, 0.6]	[0.5, 0.6]
x_5	[0.5, 0.8]	[0.3, 0.4]	[0.3, 0.7]	[0.5, 0.7]
x_6	[0.6, 0.8]	[0.3, 0.4]	[0.4, 0.7]	[0.5, 0.8]

2.2 不完备灰色信息系统的变精度灰相似关系

根据2.1节的解释, 任何不完备灰色信息系统可以转换为属性值为区间灰数的不完备信息系统来处理, 于是, 定义不完备灰色信息系统的灰相似关系如下。

定义5 在不完备灰色信息系统 GI 中, 对于任意属性子集 $B \subseteq A$, 确定了 U 上的一个灰相似关系 G 如下:

$G(B) = \{(x, y) \in U \times U \mid \forall b \in B, s(b(x), b(y)) \geq \varepsilon\}$ 其中常量 ε 满足 $0 < \varepsilon \leq 1$ 。显然, 关系 G 满足自反性, 但不一定满足对称性和传递性。

注2 $s(b(x), b(y))$ 描述了对象 y 在属性 b 下与对象 x 的相似程度, 当 $s(b(x), b(y))$ 值越大, 说明对象 y 在属性 b 下与对象 x 的相似程度越高。

针对表2, 取常量 $\varepsilon = 0.5$, 则:

$G(A) = \{(x_1, x_1), (x_1, x_3), (x_2, x_2), (x_2, x_5), (x_2, x_6), (x_3, x_1), (x_3, x_3), (x_4, x_1), (x_4, x_3), (x_4, x_4), (x_4, x_5), (x_5, x_2), (x_5, x_4), (x_5, x_5), (x_5, x_6), (x_6, x_2), (x_6, x_5), (x_6, x_6)\}$

由于灰相似关系是一种较弱的二元关系, 则依据 $G(A)$ 形成的特征类构成了论域的覆盖而非划分。相应地, 根据灰交叉关系定义上、下近似运算如下。

定义6 在不完备灰色信息系统 GI 中, $X \subseteq U$ 且 $B \subseteq A$, \underline{BX} 是 X 的下近似, \overline{BX} 是 X 的上近似, 其中: $\underline{BX} = \{x \in U \mid G_B(x) \subseteq X\}$, $\overline{BX} = \{x \in U \mid G_B(x) \cap X \neq \emptyset\}$ 。

性质1 令 GI 为不完备灰色信息系统, 有:

- 1) $G(B) = \bigcap_{a \in B} G(\{a\})$;
 - 2) $\forall B \subseteq A, \forall X \subseteq U, \underline{BX} \subseteq X \subseteq \overline{BX}$;
 - 3) $\forall B, C \subseteq A, \forall X \subseteq U, B \subset C \Rightarrow \underline{BX} \subseteq \underline{CX}$;
 - 4) $\forall B, C \subseteq A, \forall X \subseteq U, B \subset C \Rightarrow \overline{BX} \supseteq \overline{CX}$;
- 根据定义4、5, 容易证明性质1。

3 知识约简

知识约简是粗糙集理论中的一个重要概念, 根据上面构造的灰相似关系, 易得约简的定义。

定义7 令 GI 为不完备灰色信息系统, $B \subseteq A$, B 是不完备灰色信息系统的一个约简, 当且仅当:

$G(B) = G(A)$ 且 $\forall C \subset B, G(C) \neq G(A)$

下面给出计算约简的操作方法, 对于 $\forall x, y \in U$, 不完备灰色信息系统 GI 的区分辨识矩阵记为 $D = [D_A(x, y)]$, 其中:

$$D_A(x, y) = \begin{cases} \{a_i \in A: (x, y) \notin G(a_i)\} : (x, y) \notin G(A) \\ \phi: \text{其他} \end{cases}$$

因此, $D_A(x, y)$ 是个体 x 与 y 在关系 G 下有区别的所有属性的集合。

定义8 令 GI 为不完备灰色信息系统, 定义 $\Delta = \bigwedge_{D_A(x, y) \in D} \bigvee D_A(x, y)$ 为 GI 中区分函数。

利用布尔推理技术, 可将其化为极小析取范式。在其极小析取范式中, 每个合取子式就对应属性集合 A 的一个约简, 所有合取子式就是 A 的全部约简。全部约简的交即为 A 的核。

根据所介绍的属性约简的方法, 取常量 $\varepsilon = 0.5$, 表2所对应的区分辨识矩阵为表3。通过定义, GI 的区分函数为 $\Delta = a_1 \wedge a_2 \wedge a_4$, 则 $B = \{a_1, a_2, a_4\}$ 是表2的一个约简。

表3 关于表2的区分辨识矩阵

$x_i \setminus x_j$	x_1	x_2	x_3	x_4	x_5	x_6
x_1	\emptyset	a_1	\emptyset	\emptyset	a_1	a_1
x_2	$a_1 a_2$	\emptyset	$a_1 a_3$	a_4	\emptyset	\emptyset
x_3	a_4	$a_1 a_3$	\emptyset	\emptyset	a_1	$a_1 a_3$
x_4	a_4	$a_1 a_4$	a_4	\emptyset	\emptyset	a_4
x_5	$a_1 a_2$	\emptyset	a_2	\emptyset	\emptyset	\emptyset
x_6	$a_1 a_2$	\emptyset	$a_1 a_2 a_3$	a_1	\emptyset	\emptyset

4 结语

由于现实生活中存在着信息和知识的不完全性和复杂性, 促进了各种拓展粗集模型的产生与发展。本文以部分信息已知、部分信息未知的小样本、贫信息、不确定的系统为研究对象, 定义了不完备灰色信息系统, 据此提出了变精度灰相似关系及相应的粗集模型, 并给出了一些基本性质以及计算约简的实际操作方法, 因此本文的工作对拓展区间值信息系统的粗集模型有着重要的意义。

参考文献:

- [1] PAWLAK Z. Rough sets theory and its applications to data analysis [EB/OL]. [2010-05-01]. <http://www.cs.uakron.edu/~chan/cs460/Spring%202005/Rough%20Set%20and%20Its%20Applications.pdf>.
- [2] PAWLAK Z. Rough sets and intelligent data analysis[J]. Information Sciences—Informatics and Computer Science: An International Journal, 2002, 147(1/4): 1-12.

- [3] 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.
- [4] LEUNG Y, LI D Y. Maximal consistent block technique for rule acquisition in incomplete information systems[J]. Information Sciences, 2003, 153(1): 85-106.
- [5] LEUNG Y, WU W Z, ZHANG W X. Knowledge acquisition in incomplete information systems: a rough approach[J]. European Journal of Operational Research, 2006, 168(1): 164-180.
- [6] QIAN Y H, LIANG J Y. Positive approximation and rule extracting in incomplete information systems[J]. International Journal of Computer Science and Knowledge Engineering, 2008, 2(1): 51-63.
- [7] KRYSCIEWICZ M. Rough set approach to incomplete information system[J]. Information Sciences, 1998, 112(1/4): 39-49.
- [8] STEFANOWSKI J, TSOUKIAS A. Incomplete information tables and rough classification[J]. Computational Intelligence, 2001, 17(3): 545-566.
- [9] GRZYMALA-BUSSE J W. Data with missing attribute values: Generalization of indiscernibility relation and rule induction[C]// Transactions on Rough Sets, LNCS 3100. Berlin: Springer-Verlag, 2004: 78-95.
- [10] GRZYMALA-BUSSE J W. Characteristic relations for incomplete data: a generalization of the indiscernibility relation[C]// Rough Sets and Current Trends in Computing, LNCS 3066. Berlin: Springer-Verlag, 2004: 244-253.
- [11] DENG J L. Control problems of grey system[J]. System & Control Letter, 1982, 1(5): 288-294.
- [12] 刘思峰, 郭天榜, 党耀国. 灰色系统理论及其应用[M]. 3版. 北京: 科学出版社, 2004.
- [13] YANG XIBEI, YU DONGJUN, YANG JINGYU, et al. Dominance-based rough set approach to incomplete interval-valued information system[J]. Data & Knowledge Engineering, 2009, 68(11): 1331-1347.
- [14] QIAN YUHUA, LIANG JIYE, DANG CHUANGYIN. Interval ordered information systems[J]. Computers & Mathematics with Applications, 2008, 56(8): 1994-2009.
- [15] WU S X, HUANG Z Y, LUO D L, et al. A grey rough set model based on (α, β) -grey similarity relation[C]// Proceedings of IEEE International Conference on Grey Systems and Intelligent Services. Nanjing, China: IEEE Press, 2007: 903-909.

(上接第3373页)

4 结语

本文主要是从 P2P 网络的高度动态性这个角度出发, 探讨了 OLAP 网络模式下多维数据分析的决策效率。通过构建虚拟社区网络模型, 实现了基于社区划分的 OLAP 查询, 提高了 OLAP 的决策分析效率, 极大地降低了 OLAP 网络的负载。在以后的相关研究中, 将在数据立方体的共享单元和更新等关键问题上进行深入的研究和探讨。

参考文献:

- [1] 曹丽娟, 谢强, 丁秋林. 基于分布式数据缓存技术 Web-OLAP 系统研究[J]. 计算机应用, 2008, 28(2): 515-518.
- [2] KALNIS P, WEE S N, BENG C O, et al. An adaptive peer-to-peer network for distributed caching of OLAP results[C]// Proceedings of the ACM SIGMOD Conference. New York: ACM Press, 2002: 25-36.
- [3] TATARINOV I, HALEVY A. Efficient query reformulation in peerdata

management systems[C]// Proceedings of the ACM SIGMOD International Conference. Paris, France: ACM Press, 2004: 539-550.

- [4] ESPIL M M, VAISMAN A A. Aggregate queries in peer-to-peer OLAP[C]// Proceedings of Seventh ACM International Workshop Data Warehousing and On-Line Analytical Processing. Washington, DC: ACM Press, 2004: 102-111.
- [5] 杨科华, 魏莉. 一种 P2P 网络环境下的 OLAP 模式匹配方案[J]. 计算机工程与应用, 2008, 44(9): 162-164.
- [6] UPADRASHTA Y, VASSILEVA J, GRASSMANN W. Social networks in peer-to-peer systems[C]// Proceedings of the 38th Annual Hawaii International Conference. Kona, Hawaii: [s. n.], 2005: 200-211.
- [7] 周攀, 杨科华, 周利民. 一种 P2P 网络环境下的 OLAP 查询方案[J]. 计算机工程与应用, 2010, 46(33): 140-144.
- [8] TAN Y H, CHEN Z P, LIN Y P. Research and implementation on searching mechanism based on interest mining in unstructured P2P systems[J]. Computer Applications, 2006, 26(5): 1164-1166.