区间值信息系统在变精度相容关系下的属性约简

陈子春^{1,2} 秦克云¹

(西南交通大学理学院 成都 610013)¹ (西华大学数学与计算机学院 成都 610039)²

摘要 借助于属性区间值的相似程度在区间值信息系统上定义了一种具有变精度的相容关系,讨论了在这种变精度相容关系下区间值信息系统的属性约简与判定,并得到了区间值信息系统上属性约简的具体操作方法,还讨论了相似水平对区间值信息系统的属性约简的影响。

关键词 区间值信息系统,相似水平,变精度相容关系,属性约简

中图法分类号 TP18 **文献标识码** A

Attribute Reduction of Interval-valued Information System Based on Variable Precision Tolerance Relation

CHEN Zi-chun^{1,2} QIN Ke-yun¹

(Department of Mathematics, Southwest Jiaotong University, Chengdu 610013, China)¹

(School of Mathematics and Computer, Xihua University, Chengdu 610039, China)²

Abstract A new variable precision tolerance relation was defined in interval-valued information system by similarity grade of attribute's interval-valued. Based on the variable precision tolerance relation, the problem of attribute reduction and the judgment in interval-valued information system were studied, from which a approach to attribute reductions can be provided in interval-valued information system, and the influence of similarity level on attribute reduction was discussed.

Keywords Interval-valued information system, Similarity level, Variable precision tolerance relation, Attribute reduction

从信息系统中发现潜在的、非平凡的有用知识,是处理和利用海量信息的重要前提,其本质是根据实际需要按属性特征将对象进行分类后,利用不平凡手段找出不影响分类的最小属性集,以使分类知识的表示简化,而又不丢失任何信息。波兰数学家 Pawlak 于 1982 年提出的粗糙集理论^[1]是经典集合理论的推广,是处理模糊和不确定知识的有用数学工具^[1,2]。它经过 20 多年的研究与发展,已经在理论和实际应用上取得了长足的进展,特别是由于 20 世纪 80 年代末和 90 年代初在知识发现等领域的成功应用而受到了国际上广泛关注。目前,粗糙集理论已经在人工智能、知识与数据发现、模式识别与分类、故障检测等方面得到了广泛的应用^[3-7]。Pawlak 粗糙集模型是基于完备信息系统上的不可分辨关系,对无法用已知知识描述的对象集给出上、下近似。Pawlak 粗糙集模型也可以推广到非等价关系(如相容关系、优势关系、偏序关系等)^[8-13],这种推广可以用来处理不完备信息系统上的知识获取与推理。但是,传统的粗糙集方法主要处理的是离散属性值,而在很多实际问题中,由于问题的复杂性,信息系统中的属性值往往是连续的或者真实的数据只落在某个区间。而对连续属性值的处理一般需要离散化,其本质是将属性值域划分成一些离散化区间。因此,直接研究区间值信息系统上知识的获取与约简就很有必要。

属性约简是知识发现的重要课题,也是粗糙集理论的核心问题之一。

众所周知,知识库中描述知识的属性并不是同等重要的,即每一种属性在信息系统中所起的作用是不同的。有些属性是绝对必要的,有些属性是相对必要的,有些属性是完全不必要的。所谓属性约简,就是在保持知识库分类能力不变条件下,删除其中不相关或不重要的属性^[6,7,10]。但是,这种属性是否必要,完全依赖于信息系统上的知识或关系的定义,针对不同的实际问题,可以定义不同的知识或关系,这时属性是否必要也随之发生改变。就区间值信息系统,本文借助于属性区间值的相似程度,在区间值信息系统上定义了一种具有变精度的相容关系,讨论了在这种相容关系下区间值信息系统的属性约简与判定,并得到了区间值信息系统上属性约简的具体操作方法。

1 区间值信息系统上的变精度相容关系

信息系统是一个三元组 (U, A, F) , 其中 $U = \{x_1, x_2, \dots, x_n\}$ 为有限对象集,称为论域,每个 $x_i (i \leq n)$ 称为一个对象; $A = \{a_1, a_2, \dots, a_m\}$ 为有限属性集,每个 $a_i (i \leq m)$ 为一个属性; $F = \{f_k : k \leq m\}$ 为对象属性值映射集,其中 $f_k : U \rightarrow V_k (k \leq m)$, V_k 是属性 a_k 的有限值域,即 $\forall x_i \in U, f_k(x_i) \in V_k$ 。若 $\forall a_k \in A, \forall x_i \in U, f_k(x_i)$ 都是唯一确定的,则称信息系统 K 是完备的,否则便称它是不完备的;又若 $f_k(x_i) = [l_k^i, r_k^i] \in I ([0, 1])$, $l_k^i < r_k^i$, 则称 K 是区间值信息系统,其中 $I([0, 1])$ 表

到稿日期:2008-04-03 本文受国家自然科学基金项目(60474022)资助。

陈子春(1970-),男,博士研究生,主要研究方向为不确定性推理, E-mail: czclph@163.com; 秦克云(1962-),男,教授,博士生导师,主要研究方向为多值逻辑与不确定性推理。

示区间[0,1]上的闭区间全体。

由于区间值信息系统上对象的属性取值是区间,而两个区间的相似程度取决于其相交部分的区间长度,因此,我们在区间值信息系统上给出下面的定义。

定义 1.1 设 (U, A, F) 是区间值信息系统, $a_k \in A, x_i, x_j \in U, f_k(x_i) = [l_i^k, r_i^k], f_k(x_j) = [l_j^k, r_j^k]$,称 $S_{ij}^k = \frac{|[l_i^k, r_i^k] \cap [l_j^k, r_j^k]|}{|[l_i^k, r_i^k] \cup [l_j^k, r_j^k]|}$ 是对象 x_i, x_j 相对于属性 a_k 的相似度。其中 $|\cdot|$ 在这里表示闭区间的长度,并规定空集与单点集的长度是 0。

对 $a_k \in A, B \subseteq A, \alpha \in (0, 1]$,在区间值信息系统上定义相似关系:

$$T_B = \{(x_i, x_j) \in U \times U : \forall a_k \in B, S_{ij}^k \geq \alpha\}$$

并记 $T_B(x_i) = \{x_j \in U : (x_i, x_j) \in T_B\}$ 。

注:① S_{ij}^k 表示对象 x_i 与 x_j 对于属性 a_k 的属性区间值的相似程度, S_{ij}^k 越大,说明 x_i 与 x_j 相对于 a_k 的属性区间值的相似程度越高,且由定义易知 $S_{ij}^k = 0 \Leftrightarrow [l_i^k, r_i^k] \cap [l_j^k, r_j^k] = \emptyset$ 或 $[l_i^k, r_i^k]$ 与 $[l_j^k, r_j^k]$ 仅在一点取相同值; $S_{ij}^k = 1 \Leftrightarrow [l_i^k, r_i^k] = [l_j^k, r_j^k]$ 。

② $(x_i, x_j) \in T_B$ 表示对象 x_i 与 x_j 对于 B 中任一属性的相似程度都不低于给定的阈值 α 。

定理 1.1 设 (U, A, F) 是区间值信息系统, T_B 是如上定义的二元关系,则有

- (1) T_B 是 U 上的相容关系,即 T_B 是自反、对称的;
- (2) 若 $B_1 \subseteq B_2 \subseteq A$,则对 $\alpha \in (0, 1]$ 有, $T_{B_1} \subseteq T_{B_2} \subseteq T_{B_1}$; $T_{B_1}(x_i) \subseteq T_{B_2}(x_i) \subseteq T_{B_1}(x_i)$;
- (3) 对任意 $B \subseteq A, T_B = \bigcap_{a_k \in B} T_{\{a_k\}}$ 。

证明:由 T_B 的定义进行验证即得。

以下为方便,称 T_B 是区间值信息系统上的变精度相容关系, $T_B(x_i)$ 称为变精度相容类,而阈值 α 称为相似水平。若 $(x_i, x_j) \in T_B$,则称对象 x_i 与 x_j 关于 B 在相似水平 α 下相容。

由于变精度相容关系是自反的,因而对 $B \subseteq A$,所有变精度相容类 $\{T_B(x_i); x_i \in U\}$ 构成对象集 U 的一个覆盖。于是对 $X \subseteq U$,在变精度相容关系下我们就可以定义 X 关于属性集 B 的上近似与下近似。

定义 1.2 设 $K = (U, A, F)$ 是区间值信息系统, α 是相似水平, $B \subseteq A, X \subseteq U$ 。设

$$\underline{B}^\alpha(X) = \{x_i; T_B(x_i) \subseteq X\}, \overline{B}^\alpha(X) = \{x_i; T_B(x_i) \cap X \neq \emptyset\}$$

则分别称 $\underline{B}^\alpha(X), \overline{B}^\alpha(X)$ 是 X 的 α -下近似与 α -上近似。

由于变精度相容关系是自反对称的,因此由 α -下近似与上近似具有许多性质^[9,10],这里不再列出。由定理 1.2 还容易得到,当 $\alpha \leq \beta$ 时, $\underline{B}^\alpha(X) \subseteq \underline{B}^\beta(X), \overline{B}^\alpha(X) \subseteq \overline{B}^\beta(X)$ 。

2 区间值信息系统在变精度相容关系下的属性约简方法

有了区间值信息系统的变精度相容关系 T_B ,就有了区间值信息系统上的元知识 $T_B(x_i)$,但确定这一知识往往并不需要 A 中所有属性。人们希望找出 A 的一个最小属性子集就能确定出知识 $T_B(x_i)$,这就是知识的属性约简。通过约简,可以深化人们对知识的认识。因此,接下来我们讨论区间值

信息系统在变精度相容关系下的属性约简。

定义 2.1 称 $B \subseteq A$ 是区间值信息系统 (U, A, F) 的 α 分类协调集,若 $T_B = T_A$ 。若进一步对任意 $b \in B$,有 $T_{B-\{b\}} \neq T_A$,则称 B 是区间值信息系统的一个 α 分类约简。

由定义知,区间值信息系统的 α 分类约简是保持变精度分类不变的最小属性集。且由定理 1.1 知, $B \subseteq A$ 是 α 分类协调集等价于 $T_B \subseteq T_A$ 。

定义 2.2 对 $x_i, x_j \in U$,称 $D_{ij} = \{a_k \in A; S_{ij}^k < \alpha\}$ 为区间值信息系统 (U, A, F) 的 α 区分属性集;称 $D^\alpha = (D_{ij}; i, j \leq n)$ 为区间值信息系统 (U, A, F) 的 α 区分矩阵;称 $\Delta^\alpha(K) = \bigwedge \{ \bigvee D_{ij}; D_{ij} \in D_0 \}$ 是区间值信息系统 (U, A, F) 的 α 区分函数。其中 $D_0 = \{D_{ij} \neq \emptyset; x_i, x_j \in U\}$ 。

由定义知,对任意 $x_i, x_j \in U$ 及相似水平 α 都有 $D_{ii} = \emptyset, D_{ij} = D_{ji}$,即区间值信息系统的 α 区分矩阵 D^α 是主对角线上元素全为 $\emptyset$ 的对称矩阵。

定理 2.1 设 (U, A, F) 是区间值信息系统, α 是相似水平, $B \subseteq A$,则 B 是 α 分类协调集的充要条件是对任意 $D_{ij} \in D_0$,有 $B \cap D_{ij} \neq \emptyset$ 。

证明: B 是 α 分类协调集 $\Leftrightarrow T_B \subseteq T_A \Leftrightarrow (x_i, x_j) \notin T_A$ 时, $(x_i, x_j) \notin T_B$

$\Leftrightarrow$ 存在 $a_k \in A$,使 $S_{ij}^k < \alpha$ 时,存在 $a_l \in B$,使 $S_{ij}^l < \alpha$

$\Leftrightarrow$ 有 $D_{ij} \neq \emptyset$ 时,存在 $a_l \in B, a_l \in D_{ij} \Leftrightarrow D_{ij} \in D_0$ 时,有 $B \cap D_{ij} \neq \emptyset$ 。

由定理 2.1 知,寻找区间值信息系统的 α 分类约简,实际上是在 D_0 中寻找满足条件 $B \cap D_{ij} \neq \emptyset$ 的最小集合 B ,这通常可由定义在区分矩阵上的区分函数的极小析取范式中的所有合取式来得到(参见文献[8,9,14])。为此,我们有:

定理 2.2 设 (U, A, F) 是区间值信息系统, α 是相似水平, $B \subseteq A$,则 B 是 α 分类约简当且仅当 $\bigwedge B$ 是 α 区分函数 $\Delta^\alpha(K)$ 的极小析取范式中的一个合取式。

证明:注意到 $\bigwedge B$ 是 α 区分函数 $\Delta^\alpha(K)$ 的极小析取范式中的一个合取式,等价于 $\forall a_k \in B$,都存在 $D_{ij} \in D_0$,使 $B \cap D_{ij} = \{a_k\}$ 。因此,若 B 是 α 分类约简,而 $\bigwedge B$ 不是 α 区分函数 $\Delta^\alpha(K)$ 的极小析取范式中的一个合取式,则 $\forall a_k \in B, \forall D_{ij} \in D_0$,就有 $\text{card}(B \cap D_{ij}) \geq 2$,于是由定理 2.1 知 $B' = B - \{a_k\}$ 是 α 分类协调集,这就与 B 是 α 分类约简相矛盾。反之,若 $\bigwedge B$ 是 α 区分函数 $\Delta^\alpha(K)$ 的极小析取范式中的一个合取式,则 $\forall D_{ij} \in D_0$,有 $B \cap D_{ij} \neq \emptyset$,而且 $\forall a_k \in B$,都存在 $D_{ij} \in D_0$,使 $B \cap D_{ij} = \{a_k\}$,这表明 $B' = B - \{a_k\}$ 就不是 α 分类协调集,因而 B 是 α 分类约简。

例 1 考虑如下区间值信息系统(见表 1)。

表 1

	a1	a2	a3	a4
x1	[0.3, 0.6]	[0.1, 0.4]	[0.2, 0.6]	[0.5, 0.8]
x2	[0.6, 0.9]	[0.3, 0.5]	[0.4, 0.7]	[0.6, 0.8]
x3	[0.4, 0.6]	[0.2, 0.5]	[0.2, 0.5]	[0.4, 0.9]
x4	[0.4, 0.7]	[0.2, 0.4]	[0.3, 0.6]	[0.5, 0.6]
x5	[0.5, 0.8]	[0.3, 0.4]	[0.3, 0.7]	[0.5, 0.7]
x6	[0.6, 0.8]	[0.3, 0.4]	[0.4, 0.8]	[0.5, 0.8]

可以计算出对象间的相似度 S_{ij}^k 如下:

$$S_{11}^k = 1, S_{12}^k = 0, S_{13}^k = 0.25, S_{14}^k = 0.4, S_{15}^k = 0.67, S_{16}^k = 0.67, S_{21}^k = 0.5, S_{23}^k = 0.75, S_{24}^k = 0.6,$$

$$D^\alpha = \left[\begin{array}{cccccc} \emptyset & & & & & \\ a_1, a_2, a_3 & \emptyset & & & & \\ a_2 & a_1, a_3, a_4 & \emptyset & & & \\ a_1, a_4 & A & a_3, a_4 & \emptyset & & \\ a_1, a_2 & a_1, a_2, a_4 & A & a_1, a_2, a_4 & \emptyset & \\ a_1, a_2, a_3 & a_2 & a_1, a_2, a_3 & A & \emptyset & \emptyset \end{array} \right]$$

$$\Delta^\alpha(K) = a_2 \wedge (a_1 \vee a_4) \wedge (a_3 \vee a_4) \\ = (a_2 \wedge a_4) \vee (a_1 \wedge a_2 \wedge a_3)$$

此时, $\{a_2, a_4\}$ 与 $\{a_1, a_2, a_3\}$ 是区间值信息系统的 α 分类约简。明显地, 例 1 中在水平 $\alpha=0.5$ 时的分类约简 $\{a_1, a_4\}$ 不是水平 $\alpha=0.6$ 时的分类协调集。

参考文献

[1] Pawlak Z. Rough set. International Journal of Computer and Information Science, 1982(11): 341-356
 [2] Pawlak Z. Rough set; Theoretical Aspects of Reasoning about Data. Boston; Kluwer Academic Publishers, 1991
 [3] 王钰, 苗夺谦, 等. 关于 Rough Set 理论与应用的综述[J]. 模式识别与人工智能, 1996(6): 337-344
 [4] 刘清. Rough 集及 Rough 推理[M]. 北京: 科学出版社, 2001
 [5] Han J, Cai Y, Cercone N. Data-driven discovery of quantitative rules in relational databases. IEEE Transactions on Knowledge and Data Engineering, 1993, 5 (1): 29-40

[6] Mrozek A. Rough sets and dependency analysis among attributes in computer implementations of expert's inference models [J]. International Journal of Man-Machine Studies, 1989, 30 (4): 457-473
 [7] Yasdi R. Combining rough sets learning and neural learning: method to deal with uncertain and imprecise information, Neuralcomputing, 1996, 7 (1): 61-84
 [8] 张文修, 吴伟志, 梁吉业, 等. 粗糙集理论与方法[M]. 北京: 科学出版社, 2001: 80-100
 [9] 张文修, 梁怡, 吴伟志. 信息系统与知识发现[M]. 北京: 科学出版社, 2003
 [10] Greco S, Matarazzo B, Slowinski R. Rough sets theory for multi-criteria decision analysis. European Journal of Operational Research, 2001, 129: 1-47
 [11] Kim D. Data classification based on tolerant rough set. Pattern Recognition, 2001, 34: 1613-1624
 [12] Yao Y Y. Two views of the theory of rough sets in finite universes. International Journal of Approximate Reasoning, 1996, 15 (4): 291-317
 [13] Yao Y Y. Information granulation and rough set approximation. International Journal of Intelligent Systems, 2001, 16 (1): 87-104
 [14] Skowron A, Rauszer C. The Discernibility Matrices and Functions in Information Systems [M]. Boston; Kluwer Academic Publishers, 1992: 331-362

(上接第 147 页)

$\langle time, cost, security, reliability \rangle$ 。其中, $time$ 是原子服务的响应时间; $cost$ 是获取该原子服务所必须支付的费用; $security$ 是原子服务的安全等级; $reliability$ 是原子服务的可靠性。并设 $time()$, $cost()$, $security()$ 和 $reliability()$ 为 QoS 提取函数。

WS_i 和 WS_r 是 QoS 匹配的, 则必须满足如下条件:

- I: $security(WS_i) \geq security(WS_r)$;
- II: $reliability(WS_i) \geq reliability(WS_r)$;
- III: $time(WS_i) \leq time(WS_r)$;
- IV: $cost(WS_i) \leq cost(WS_r)$ 。

I 式表示服务的安全等级不能低于用户要求的安全等级; II 式表示服务的可靠性级别不能低于用户要求的可靠性级别; III 式表示服务的响应时间不能高于用户给定的最长响应时间; IV 式表示服务的收费不能高于用户愿承担的最高费用。

当满足上述条件的服务不止一个时, 怎样选择服务呢? 肯定不是 $security(WS_i)$ 和 $reliability(WS_i)$ 越大越好, 也不是 $time(WS_i)$ 和 $cost(WS_i)$ 越小越好。因为这样的选择或许是不经济实用的, 或许是资源浪费的, 虽然它的其它性能很好。在此, 给出基于 QoS 的服务性价比的定义:

定义 9(服务性价比) 用 $\langle time, cost, security, reliability \rangle$ 描述服务的 QoS, 则其性价比 λ 定义如下:

$$\lambda = \frac{security \cdot reliability}{time \cdot cost} \quad (3)$$

其中, $\frac{security \cdot reliability}{time}$ 是对 QoS 性能的刻画, 它与安全性和可靠性成正比, 与响应时间成反比; $cost$ 是价格。

不同的用户群对 QoS 的要求不同, 但追求高的性价比是用户的普遍要求。

高性价比选择策略。在满足 QoS 匹配条件的服务中, 选

择具有最高性价比的服务推荐给用户。

该策略不仅能为用户提供满意的服务, 而且还把高性能的服务留给那些有特殊要求的用户。如航空航天业、国家安全机关和金融业等。

结束语 原子服务发现就是从原子服务库中搜索并发现能满足服务请求的原子服务的过程。文中提出的 QoS 匹配策略回答了“怎样从已发现的能满足用户服务请求的原子服务集中找到一个用户满意的服务”。原子服务发现是组合服务发现的基石, 对该模型的细化深入和在该模型下的组合服务发现将是下一步研究的重点。

参考文献

[1] <http://www.w3.org/2002/ws/>
 [2] Agarwal S, Handschuh S, Staab S. Annotation, Composition and Invocation of Semantic Web Service[J]. Journal on Web Semantics, 2005, 2(1): 1-24
 [3] Berardi D, Calvanese D, Giacomo G D, et al. Automatic Composition of E-services that Export Their Behavior[J]//Proc. of 1st Intl. Conf. on Service Oriented Computing. Trento, Italy, 2003
 [4] Martin D, Burstein M, Hobbs J, et al. OWL-S: Semantic markup for Web services. 2004. <http://www.w3.org/Submission/OWL-S/>
 [5] Chen Wang - Hu, Liu Chen, Li Hou - Fu, et al. An Approach to Dynamically Forming Semantic Infrastructure for Virtual Organizations[J]. Chinese Journal of Computers, 2006(7): 1130-1132
 [6] Kornilakis H, Grigoriadou M, et al. Using WordNet to support interactive concept map construction[J]// Proceedings of the IEEE International Conference on Advanced Learning Technologies. Joensuu, Finland, 2004: 600-604
 [7] Paolucci M, Kawamura T, Payne T R, et al. Semantic Matching of Web Services Capabilities//International Semantic Web Conference, 2002: 333-347