DOI:10.11992/tis.201705027
网络出版地址:http://kns.cnki.net/kcms/detail/23.1538.TP.20170705.1654.004.html

变精度下不完备邻域决策系统的属性约简算法

王映龙[1],曾淇[1],钱文彬[2],杨珺[2]
(1.江西农业大学 计算机与信息工程学院,江西 南昌 330045; 2.江西农业大学 软件学院,江西 南昌 330045)

摘 要:邻域粗糙集模型在处理完备的数值型数据中得到广泛应用,但针对不完备的数值型和符号型混合数据进行属性约简的讨论相对较少。为此,首先结合邻域粗糙集给出了可变精度模型下不完备邻域决策系统的上、下近似算子及属性约简;然后通过邻域粒化的方法构建了广义邻域下可变精度的粗糙集模型,并提出了一种属性重要度的评价方法;在此基础上,设计出了面向不完备邻域决策系统的属性约简算法,该算法可直接处理不完备的数值型和符号型混合数据;最后,通过实例分析验证了本文提出的算法能够求解出变精度下不完备邻域决策系统的属性约简结果。

关键词:粗糙集理论;邻域关系;不完备信息系统;变精度分类粗糙集;粒计算;多粒度;约简;决策粗糙集
中图分类号:TP311 **文献标志码**:A **文章编号**:1673-4785(2017)03-0386-06

中文引用格式:王映龙,曾淇,钱文彬,等. 变精度下不完备邻域决策系统的属性约简算法[J]. 智能系统学报, 2017, 12(3): 386-391.
英文引用格式:WANG Yinglong, ZENG Qi, QIAN Wenbin, et al. Attribute reduction algorithm of the incomplete neighborhood decision system with variable precision[J]. CAAI transactions on intelligent systems, 2017, 12(3): 386-391.

Attribute reduction algorithm of the incomplete neighborhood decision system with variable precision

WANG Yinglong[1], ZENG Qi[1], QIAN Wenbin[2], YANG Jun[2]
(1. School of Computer and Information Engineering, Jiangxi Agricultural University, Nanchang 330045, China; 2. School of Software, Jiangxi Agricultural University, Nanchang 330045, China)

Abstract: Neighborhood rough set model has been widely used in numerical data processing complete, but the discussion of attribute reduction for numeric and symbolic mixed incomplete data is relatively small. Therefore, to resolve this problem, by combining the neighborhood rough set, first, the upper and lower approximation operators and the attribute reduction of the incomplete neighborhood decision system were analyzed based on the variable precision model. Subsequently, based on the generalized neighborhood relation, a rough set model was constructed using the neighborhood granulation method. Furthermore, a method evaluating the attribute significance degree was proposed. Based on this method, an attribute reduction algorithm for the incomplete neighborhood decision system was designed, which can deal with incomplete values directly type and symbolic mixed data. Finally, through the example analysis, the algorithm can solve the attribute reduction result of incomplete neighborhood decision system with variable precision.

Keywords: rough set theory; neighborhood relation; incomplete information system; variable precision classification; granular computing; multi-granulation; reducation; decision-theoretic rough sets

波兰数学家 Pawlak 提出的粗糙集理论能有效处理信息系统中不精确、不确定信息[1],其在模式识别、市场决策、医疗诊断等领域广泛应用[2-3]。经典 Pawlak 粗糙集理论的研究对象是完备的信息决策表。然而在现实生活中,往往很多决策系统存在多种数据类型,如连续型数据、不完备型数据和集

收稿日期:2017-05-19. 网络出版日期:2017-07-05.
基金项目:国家自然科学基金项目(61502213,71461013,61462038);江西省自然科学基金项目(20151BAB217009,20132BAB201045);江西省教育厅科学技术项目(GJJ150399,GJJ150505).
通信作者:钱文彬. E-mail:qianwenbin1027@126.com.

值型数据等[4-6]。由于经典粗糙集在处理连续型数据时需进行离散化预处理,将不可避免地造成信息的丢失,且对于含有不完备型数据的决策系统,传统的粗糙集模型较难直接处理。近年来,针对混合、模糊、不完备的粗糙集模型扩展及应用成为粒度计算研究的热点问题[7-13]。

基于粒计算的属性约简研究已取得许多有意义的成果[14-18]:文献[14]研究了混合数据下的知识发现及邻域粒化问题;文献[15]提出了悲观多粒度粗糙集的概念,解决了利用“求同消异”的决策策略处理多个不可分辨关系之间存在相互独立的情况;文献[16]将多粒度粗糙集扩展到邻域多粒度粗糙集;为提高分类的效果,文献[17]在多粒度粗糙集的基础上引入了错误分类率的概念,即在允许一定程度分类率的前提下,寻找数据之间的相关性,以解决属性间不确定关系的数据分类问题;对于不完备信息系统,文献[18]提出了一种基于容差关系的不完备可变精度多粒度粗糙集模型。

上述研究分别针对不完备粗糙集、变精度粗糙集进行研究。由于现实生活中同时存在大量的不完备、连续数值型、符号型属性数据的情况,现有的邻域粗糙集计算方法对上述情况和数据集的可控性调节划分的讨论相对较少。为此,本文结合多粒度粗糙集,分析了可变精度模型下不完备邻域决策系统的上、下近似算子及属性约简,并通过邻域粒化方法构建了广义邻域下可变精度的粗糙集模型;在此基础上,构造了一种衡量属性重要度的方法,并设计了不完备邻域系统的属性约简算法;最后,通过实例分析验证了算法的有效性。

1　基本知识

给定一个决策系统 $DS=(U,C,D,V)$,其中:$U=\{x_1,x_2,\cdots,x_n\}$ 表示非空有限样本集合,称为论域;C 是条件属性集合;D 是决策属性,$C\cap D=\phi$,若 $D=\phi$,则决策系统转换为信息系统。V 为属性值域,对于 $\forall a\in C$,V_a 为属性 a 的值域;$x_i(a)$ 为样本 x_i 在属性 a 上的取值。对于属性子集 $R\subseteq C$,可得到 R 在 U 上的划分 $U/R=\{R_1,R_2,\cdots,R_m\}$。

如果 V 中包含连续型和符号型等属性类型的对象,则该决策系统称为邻域决策系统。在邻域决策系统中,当部分样本的条件属性值缺失时,则该邻域决策系统称为不完备邻域决策系统,缺失值用“ * ”表示。

定义 1[14]　设 $DS=(U,C,D,V)$ 是不完备邻域决策系统,对于 $\forall x\in U$,定义 x 在 $C\cup D$ 上的邻域信息粒子为 $N_A(x)=\{y\mid y\in U,\Delta A(x,y)\leqslant\delta,\delta\geqslant 0\}$,其中 δ 表示邻域半径,$\Delta A:U\times U\to R$ 为 U 上的一定度量,满足以下性质:

1) $\forall x,y\in U$, $\Delta A(x,y)\geqslant 0$, 当 $\Delta A(x,y)=0$ 时, $\forall a_i\in A,a_i(x)=a_i(y)$;

2) $\forall x,y\in U,\Delta A(x,y)=\Delta A(y,x)$;

3) $\forall x,y,z\in U,\Delta A(x,z)\leqslant\Delta A(x,y)+\Delta A(y,z)$。

对于连续型的数据,采用欧式距离度量:

$$\Delta A(x,y)=\sqrt{\sum_{a_i\in A}|a_i(x)-a_i(y)|^2}$$

对于符号型的数据,可定义:

$$\Delta A(x,y)=\begin{cases}0, & a_i(x)=a_i(y)\\ \infty, & a_i(x)\neq a_i(y)\end{cases}$$

当 $\delta=0$ 时,变为经典粗糙集模型。

定义 2[19]　将邻域等价关系扩展到符号型、连续型和缺失型等未知属性共存下的不完备模糊系统,可得到以下广义邻域关系:

$$R(x)=\{(x,y)\in U^2:\forall a\in x\cap f_1(x)=f_1(y),a(x)\in\delta(y,a)\cup a(y)\in\delta(x,a)\cup a(x)=*\cup a(y)=*\}$$

广义邻域关系满足自反性,但不一定满足对称性和传递性,因为任意样本与其自身是不可分辨的,所以任何等价关系均满足自反性。在这里放宽了对称性和传递性的限制,扩展了应用范围。

定义 3[19]　给定 $DS=(U,C,D,V)$ 是不完备邻域决策系统,$B\subseteq C$,$N(B)$ 是 DS 上关于 B 的邻域关系所构成的邻域粒子族,对于 $\forall x\subseteq U$ 在 B 上的邻域,记为 $N_B(x)$,则 x 在 DS 上的下近似和上近似分别为 $\underline{P_B}(x)=\{x\in U\mid N_B(x)\subseteq x\}$,$\overline{P_B}(x)=\{x\in U\mid N_B(x)\cap x\neq\phi\}$,$x$ 的邻域近似边界为 $\mathrm{bnr}_B(x)=\overline{P_B}(x)-\underline{P_B}(x)$。

$\overline{P_B}(x)$ 是可能包含 x 的邻域信息粒子组合的集合,$\underline{P_B}(x)$ 是包含 x 的邻域信息粒子组合的集合,与 x 无关的邻域信息粒子集合为 $U-\overline{P_B}(x)$,记为 $\sim x$。

定义 4　给定 $DS=(U,C,D,V)$ 是不完备邻域决策系统,X 和 Y 是 U 上的两个非空子集,定义集合 X 关于集合 Y 的相对错误分类率:

$$e(X,Y)=\begin{cases}1-\dfrac{|X\cap Y|}{|X|}, & |X|>0\\ 0, & |X|=0\end{cases}$$

如果将集合 X 中的元素分到集合 Y 中,则出现分类错误的比例为 $e(X,Y)\times 100\%$。

2 不完备可变精度粗糙集模型

定义 5 给定 DS = (U, C, D, V) 是不完备邻域决策系统, $B \subseteq C$, 决策属性集合 $D = \{d_1, d_2, \cdots, d_n\}$, $0 \leqslant k < 0.5$, 在可变精度 k 下, 属性集 B 相对于决策属性 D 的上、下近似分别为

$$\overline{P_B^k}(d_i) = \{x \in U \mid e(N_B(x), d_i) < 1 - k\}$$

$$\underline{P_B^k}(d_i) = \{x \in U \mid e(N_B(x), d_i) \leqslant k\}$$

在可变精度 k 下, 属性集 B 相对于决策属性 D 的边界域为 $\mathrm{BN}_B^k(d_i) = \overline{P_B}(d_i) - \underline{P_B}(d_i)$, 负区域为 $\mathrm{NEG}_B^k(d_i) = U - \underline{P_B}(d_i)$, 正区域为 $\mathrm{POS}_B^k = \bigcup\limits_{i \leqslant n} \underline{P_B}(d_i)$。

决策属性值 d_i 在可变精度 k 的上近似是 U 中以不小于 k 的分类样本划分到 d_i 上的邻域信息粒子的集合, 下近似是 U 中以不小于 $1-k$ 的分类样本划分到 d_i 上的邻域信息粒子的集合。根据多粒度粗糙集的思想, 在可变精度不完备邻域决策系统中, 通过对邻域粒度 δ 和可变精度 k 的控制来区分不同的信息。邻域粒度 δ 越小, 可变精度 k 取值越优, 区分能力越强。

定理 1 由定义 5 可得以下性质:

1) $\underline{P_B^k}(d_i) \subseteq \underline{P_B}(d_i)$, $\overline{P_B^k}(d_i) \subseteq \overline{P^B}(d_i)$;

2) 对于 $B' \subseteq B \subseteq A$, 则存在 $\underline{P_{B'}^k}(d_i) \subseteq \underline{P_B^k}(d_i)$, $\overline{P_{B'}^k}(d_i) \subseteq \overline{P_B^k}(d_i)$。

证明 1) $\forall x \in \underline{P_B}(d_i)$, 由定义 4 可得, 至少 $\exists B \in A$, 使得 $N_B(x) \subseteq X$ 成立, 再由定义 3 和定义 5 可知, $e(N_B(x), d_i) \leqslant k$, 于是 $x \in \underline{P_B}(d_i)$, 从而 $\underline{P_B^k}(d_i) \subseteq \underline{P_B}(d_i)$, 类似可证 $\overline{P_B^k}(d_i) \subseteq \overline{P_B}(d_i)$, 证毕。

2) 由定义 5 可得, 对于 $\forall x \in U$, 则有 $e(N_B(x), d_i) < 1-k$ 成立, 因为 $B' \subseteq B \subseteq A$, 则 $e(N_{B'}(x), d_i) = 1 - \dfrac{|N_{B'}(x) \cap d_i|}{|N_{B'}(x)|} < 1 - \dfrac{|N_B(x) \cap d_i|}{|N_B(x)|} = e(N_B(x), d_i) < 1-k$, 即 $x \in \underline{P_B}(d_i)$, 所以 $\underline{P_{B'}^k}(d_i) \subseteq \underline{P_B^k}(d_i)$。类似可证 $\overline{P_{B'}^k}(d_i) \subseteq \overline{P_B^k}(d_i)$, 证毕。

从以上性质可知: 随着可变精度 k 的增大, $\{d_i\}$ 的正区域和负区域减小, 而边界域则增大; 反之, 随着 k 的减小, $\{d_i\}$ 的正区域和负区域将增大, 而边界域在缩小。如上所说, 在一个合适的可变精度 k 范围下, d_i 有较大的可分辨性。

性质 1 在不完备邻域决策系统中, 对缺失的条件属性值的判定: 当决策属性值一致时, 如果符号型条件属性取值相同, 连续型属性取值在相同邻域内的对象归为同一类, 否则视为不同类。

在不完备邻域决策系统 DS = (U, C, D, V) 中, 条件属性集合为 $C = \{C_1, C_2, C_3, C_4\}$, 决策属性集为 $D = \{d_1, d_2\}$, $\{C_1, C_2, C_3\}$ 为连续型数值属性, $\{C_4\}$ 为符号型属性, 下面通过表 1 的实例说明。

表 1 不完备邻域决策系统(1)

Table 1 Incomplete neighborhood decision system(1)

U	C_1	C_2	C_3	C_4	D
x_1	0.1	2.1	3.0	T	1
x_2	0.2	2.2	*	T	1
x_3	0.2	*	3.2	T	1
x_4	*	2.0	3.1	F	1
x_5	0.2	2.2	3.1	T	2

令 $\delta = 0.1$, $k = 0.2$, 因为样本 x_1 与 x_5 的决策属性 D 取值不同, 就算连续型的属性值都在邻域范围内, 符号型条件属性取值相同, 也不能视为同一类; 因为当 $k = 0.2$ 时, 即两个样本在 C_1, C_2, C_3, C_4 属性中只能有一个属性取值不同或不在同一邻域中, 所以 x_1, x_2 属于同一类, x_1 与 x_3, x_4 不属于同一类。

定义 6 给定 DS = (U, C, D, V) 是不完备邻域决策系统, $B \subseteq C$, 决策属性集合 $D = \{d_1, d_2, \cdots, d_n\}$, 如果 $\forall a \in B$, $\gamma_{B-a}^k(d) = \gamma_B^k(d)$, 称属性 a 对于集合 B 是冗余的; 如果 $\gamma_{B-a}^k(d) < \gamma_B^k(d)$, 则称 a 是必要的; 如果 $\forall a \in B$, 属性 a 对于集合 B 都是必不可少的, 称 B 是独立的。决策属性 d_i 对属性集合 B 的依赖度为

$$\gamma_B^k = \frac{|\mathrm{POS}_B^k(d_i)|}{|U|}$$

定义 7 给定 DS = (U, C, D, V) 是不完备邻域决策系统, 决策属性集合 $D = \{d_1, d_2, \cdots, d_n\}$, $B \subseteq C$, 若属性子集 B 是不完备邻域决策系统的一个约简集, 则 B 满足:

1) $\gamma_B^k(d) = \gamma_C^k(d)$;

2) $\forall a \in B$, $\gamma_{B-a}^k(d) \neq \gamma_B^k(d)$。

该定义的条件 1) 保证了在可变精度 k 下, 约简集与系统中含有全部条件属性时的集合具有相同的分辨能力; 条件 2) 保证了属性子集 B 是独立的, 所有的属性都是必不可少的, 没有冗余的属性。这一定义与经典粗糙集模型中的定义在形式上是完全一致的。然而, 由于该模型定义了数值空间中的

粒化和逼近,而经典粗糙集是定义在离散空间的,因此适合于完全不同的应用场合。

定义 8　给定 DS=(*U*,*C*,*D*,*V*)是不完备邻域决策系统,$B\subseteq C$,对于$\forall a\in C-B$,则属性 *a* 相对于 *B* 的重要性计算方式为

$$\mathrm{SIG}(a,B,D)=\gamma_{B\cup\{a\}}^{k}(D)-\gamma_{B}^{k}(D)$$

3　变精度下不完备邻域系统的属性约简

3.1　变精度下不完备邻域系统的属性约简算法

当处理高维的大规模数据时,为减少计算时间和保证知识获取的简洁,去除冗余属性尤为必要。为此,本文针对不完备邻域决策系统,在可变精度粗糙集模型下,提出了一种基于属性重要度指标的属性约简算法。首先,根据决策属性对原样本集合做初步划分,再根据可变精度 *k* 值和邻域半径 δ 值,计算邻域关系 N_A 及对应属性依赖度 $\gamma_{C_i}^{k}(D)$,然后采用贪心式搜索方法,每次计算剩余属性的属性重要度,从中选择属性重要度最大的属性加入约简集合中,直到约简结果中属性集合的重要度等于原始属性集的重要度,即得到不完备邻域决策系统的属性约简结果。由此,变精度下不完备邻域系统的属性约简算法描述如下。

输入　不完备邻域决策系统 DS=(*U*,*C*,*D*,*V*),邻域半径 δ,可变精度 *k*。

输出　属性约简结果 RED。

1)初始化 RED=ϕ;

2) 根据决策属性 *D* 的值对论域 *U* 进行划分 $U/D=\{D_1,D_2,\cdots,D_m\}$;

3)根据可变精度 *k* 值和邻域半径 δ 值,计算邻域关系 N_A 及对应属性依赖度 $\gamma_{C_i}^{k}(D)$;

4)选取 $\gamma_{C_i}^{k}(D)$ 中的最大值,令 RED=C_i;

5)对于$\forall C_i\in C-\mathrm{RED}$,根据定义 6 和定义 7 计算属性的重要度 $\mathrm{SIG}(C_i,\mathrm{RED},D)=\gamma_{\mathrm{RED}\cup\{C_i\}}^{k}(D)-\gamma_{\mathrm{RED}}^{k}(D)$,选取属性重要度最大的条件属性:其满足 $\mathrm{SIG}(C_i,\mathrm{RED},D)=\max\{\mathrm{SIG}(C_i,\mathrm{RED},D)\}$,并将属性 C_i 放入 RED;

6)如果 $\gamma_{\mathrm{RED}}^{k}(D)\neq\gamma_{C}^{k}(D)$,则算法回到 5),否则执行 7);

7)输出约简 RED,算法结束。

算法复杂度分析:

算法的第 1 步的时间复杂度为 $O(1)$,第 2 步的时间复杂度为 $O(|U|)$,第 3 步的最坏时间复杂度为 $O(|C|^2|U|\log_2|U|)$,第 4 步的时间复杂度为 $O(1)$,第 5 步的最坏时间复杂度为 $O(|C-\mathrm{RED}|^2\cdot|U|\log_2|U|)$,第 6 和第 7 步的时间复杂度为 $O(1)$。由上述分析可知,本算法的时间复杂度主要由步骤 3 和步骤 6 的计算耗时所决定,因此该算法的最坏时间复杂度为 $O(|C|^2|U|\log_2|U|)$;由于算法无需额外的储存空间,可知算法的最坏空间复杂度为 $O(|U||C|+|U|)$。文献[12]在完备决策系统下设计了变精度悲观多粒度粗糙集的下近似分布粒度约简算法,其时间复杂度为 $O(|U|^2|C|^2)$;文献[13]在信息观下提出了基于不一致邻域矩阵的属性约简算法,其时间复杂度为 $O(|U|^2|C|^3)$,空间复杂度为 $O(|U|^2|C|)$。本文的算法在计算效率和存储空间上具有一定优势,且能处理不完备的邻域数据,算法的扩展性较好。

3.2　与经典粗糙集及邻域模型比较

与经典粗糙集及邻域模型相比较,本文提出的变精度不完备邻域系统的属性约简模型具有以下优点:

1)经典粗糙集的属性约简适用于离散型属性约简,需先离散化连续型数据,这将不可避免地造成信息的丢失。而变精度不完备邻域系统的属性约简模型既可处理离散型属性约简,也可直接用于连续型属性约简。本文的属性约简模型是对经典粗糙集模型的扩展。

2)对于含有不完备型数据的决策系统,经典的粗糙集模型较难直接处理,而本文提出的属性约简模型可直接对数据进行分析,并在可变精度的调节下,能得到数据不同层次的信息粒度。

3)变精度不完备邻域系统的属性约简模型是对邻域模型的进一步扩展,基于邻域的属性约简需计算各样本的邻域,而本文的属性约简模型因为在可变精度的调控下先对样本进行初步筛选,再进行邻域计算,有效减少了计算量。

4　实例分析

为了验证该方法的有效性,我们选择了一个不完备邻域决策系统进行详细分析,表 2 中共有 10 个样本对象,条件属性集为$\{C_1,C_2,C_3,C_4\}$,决策属性为$\{D\}$。设置邻域半径 $\delta=0.1$,即两样本之间的邻域半径小于等于 0.1;可变精度 $k=0.2$,即两个样本在条件属性集中只能有一个属性取值不同或不

在同一邻域中。

表 2 不完备邻域决策系统(2)

Table 2 Incomplete neighborhood decision system(2)

U	C_1	C_2	C_3	C_4	D
x_1	0.3	0.3	0.75	T	poor
x_2	0.5	0.2	0.8	T	poor
x_3	1	*	0.5	T	good
x_4	0.9	0.8	0.4	T	good
x_5	*	0.6	0.45	F	good
x_6	0.3	0.4	*	F	poor
x_7	0	0.4	0.95	F	poor
x_8	0.2	*	0.8	F	good
x_9	0.25	0.75	0.6	F	good
x_{10}	0.2	0.6	*	*	poor

首先根据算法第 2)步可计算出不完备邻域决策系统在决策属性$\{D\}$下的划分为$\{D_1, D_2\}$,即

$$D_1(x) = \{x_3, x_4, x_5, x_8, x_9\}$$

$$D_2(x) = \{x_1, x_2, x_6, x_7, x_{10}\}$$

根据邻域半径$\delta=0.1$和可变精度$k=0.2$,通过算法的第 3)步可分别计算每个属性的邻域关系和所对应的依赖度,即

$$\gamma_{C_1}^{0.2}(D) = \frac{1}{2} \quad \gamma_{C_2}^{0.2}(D) = \frac{4}{5}$$

$$\gamma_{C_3}^{0.2}(D) = \frac{7}{10} \quad \gamma_{C_4}^{0.2}(D) = \frac{1}{10}$$

根据算法第 4)步,因为$\{\gamma_{C_i}^{k}(D)\}$中依赖度的最大值所对应的属性为C_2,则令 RED=$\{C_2\}$。继续执行算法第 5)步,在$\{C-\text{RED}\}$条件属性集中,依次计算剩余属性的重要度,可得

$$\text{SIG}(C_1, \text{RED}, D) = \gamma_{C_1 \cup \text{RED}}^{0.2}(D) - \gamma_{\text{RED}}^{0.2}(D) = \frac{1}{10}$$

$$\text{SIG}(C_3, \text{RED}, D) = \gamma_{C_3 \cup \text{RED}}^{0.2}(D) - \gamma_{\text{RED}}^{0.2}(D) = \frac{1}{5}$$

$$\text{SIG}(C_4, \text{RED}, D) = \gamma_{C_4 \cup \text{RED}}^{0.2}(D) - \gamma_{\text{RED}}^{0.2}(D) = 0$$

则可知C_3为所对应的属性重要度最大的属性,将属性C_3放入 RED 中,有 RED=$\{C_2, C_3\}$。

再根据算法第 6)步,由于此时$\gamma_{\text{RED}}^{k}(D) \neq \gamma_{C}^{k}(D)$,则算法回到 5),同样可计算出$C_1$为剩余属性中属性重要度最大的属性,将属性$C_1$放入 RED 中,有 RED=$\{C_2, C_3, C_1\}$。转至算法的第 6)步,由于此时$\gamma_{\text{RED}}^{k}(D)=\gamma_{C}^{k}(D)$,则算法转至第 7),输出属性约简结果。因此,按照以上算法步骤,当$\delta=0.1$时,不完备邻域决策系统的属性约简为$\{C_1, C_2, C_3\}$。

上述实例是对 10 组样本对象进行的计算和分析,本文算法中可变精度k值和邻域半径δ值是可变的,在现实应用中可根据具体需求设定可变精度和邻域半径以满足知识的细化程度。

通过上述实例分析,利用本文的算法计算属性约简结果需执行$O(|C|^2|U|\log_2|U|_2)$次,文献[12]中的算法需要执行$O(|C|^2|U|^2)$次,本文在一定程度降低了算法的计算复杂性。在存储空间上,本文算法需 50 个储存空间,而文献[13]构建矩阵需 400 个空间用于存储邻域矩阵元素,本文算法占用的空间相对较少。因此,本文所提出的算法在计算效率上具有优势,并较好地减少算法对存储空间的消耗。

5 结束语

针对不完备邻域决策系统的属性约简问题,本文通过邻域粒化的方法,构建了广义邻域下可变精度的粗糙集模型,同时构造了一种属性重要度的评价方法,并设计了不完备邻域系统的属性约简算法。通过实例分析,该方法能对不完备的数值型和符号型混合数据进行属性约简。在大数据时代,数据的不断产生,需实时更新信息系统,下一步将在此背景下研究,当不完备邻域决策系统中的数据动态变化时如何对属性约简进行增量更新。

参考文献:

[1] PAWLAK Z. Rough sets and intelligent data analysis[J]. Information sciences, 2002, 147(1): 1-12.

[2] ZHANG Junbo, WONG Jiansyuan, PAN Yi, et al. A parallel matrix-based method for computing approximations in incomplete information systems[J]. IEEE transactions on knowledge and data engineering, 2015, 27(2):326-339.

[3] WU Weizhi, QIAN Yuhua, LI Tongjun, et al. On rule acquisition in incomplete multi-scale decision tables[J]. Information sciences, 2017, 378: 282-302.

[4] 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: 123-131.

[5] 刘芳,李天瑞. 基于边界域的不完备信息系统属性约简方法[J]. 计算机科学, 2016, 43(3): 242-245.
LIU Fang, LI Tianrui. Method for attribute reduction based on rough sets boundary regions[J]. Computer science, 2016, 43(3): 242-245.

[6] WU Jianrong, KAI Xuewen, LI Jiaojiao. Atoms of monotone set-valued measures and integrals[J]. Fuzzy sets and

systems, 2015, 183: 972-979.

[7]王国胤, 张清华. 不同知识粒度下粗糙集的不确定性研究[J]. 计算机学报,2008, 31(9):1588-1598.
WANG Guoyin, ZHANG Qinghua. Uncertainty of rough set in different knowledge granularities[J]. Chinese journal of computers, 2008, 31(9): 1588-1598.

[8]钱文彬,杨炳儒,谢永红,等. 一种基于属性度量的快速属性约简算法[J]. 小型微型计算机系统, 2014, 35(6): 1407-1411.
QIAN Wenbin, YANG Bingru, XIE Yonghong, et al. A quick algorithm for attribute reduction based on attribute measure[J]. Journal of chinese computer systems, 2014, 35(6): 1407-1411.

[9]鞠恒荣, 马兴斌, 杨习贝, 等. 不完备信息系统中测试代价敏感的可变精度分类粗糙集[J]. 智能系统学报, 2014, 9(2):219-223.
JU Hengrong, MA Xingbin, YANG Xibei, et al. Test-cost-sensitive based variable precision classification rough set in incomplete information system[J]. CAAI transactions on intelligent systems, 2014, 9(2): 219-223.

[10]陈昊, 杨俊安, 庄镇泉. 变精度粗糙集的属性核和最小属性约简算法[J]. 计算机学报, 2012, 35(5): 1011-1017.
CHEN Hao, YANG Junan, ZHUANG Zhenquan. The core of attributes and minimal attributes reduction in variable precision rough set[J]. Chinese journal of computers, 2012, 35(5):1011-1017.

[11]张清华,薛玉斌,王国胤. 粗糙集的最优近似集[J]. 软件学报, 2016, 27(2):295-308.
ZHANG Qinghua, XUE Yubin, WANG Guoyin. Optimal approximate sets of rough sets[J]. Journal of software, 2016, 27(2): 295-308.

[12]孟慧丽,马媛媛,徐久成. 基于下近似分布粒度熵的变精度悲观多粒度粗糙集粒度约简[J]. 计算机科学, 2016, 43(2): 83-85,104.
MENG Huili, MA Yuanyuan, XU Jiucheng. Granularity reduct of variable precision pessimistic multi-granulation rough set based on granularity entropy of lower approximate distribution[J]. Computer science, 2016, 43(2): 83-85,104.

[13]续欣莹, 刘海涛, 谢珺, 等. 信息观下基于不一致邻域矩阵的属性约简[J]. 控制与决策, 2016, 31(1): 130-136.
XU Xinying, LIU Haitao, XIE Jun, et al. Attribute reduction based on inconsistent neighborhood matrix under information view[J]. Control and decision, 2016, 31(1): 130-136.

[14]胡清华,于达仁, 谢宗霞. 基于邻域粒化和粗糙逼近的数值属性约简[J]. 软件学报, 2008, 19(3): 640-649.
HU Qinghua, YU Daren, XIE Zongxia. Numerical attribute reduction based on neighborhood granulation and rough approximation[J]. Journal of software, 2008, 19(3): 640-649.

[15] QIAN Yuhua, LI Shunyong, LIANG Jiye. Pessimistic rough set based decisions: a multigranulation fusion strategy[J]. Information sciences, 2014, 264: 196-210.

[16]LIN Guoping, QIAN Yuhua, LI Jinjin. Neighborhood based multigranulation rough sets[J]. International journal of approximate reasoning, 2012, 7(53): 1080-1093.

[17]沈家兰, 汪小燕, 申元霞. 可变程度多粒度粗糙集[J]. 小型微型计算机系统, 2016, 37(05): 1012-1016.
SHEN Jialan, WANG Xiaoyan, SHEN Yuanxia. Variable Grade multi-granulation rough set [J]. Journal of Chinese computer systems, 2016, 37(5): 1012-1016.

[18]许韦,吴陈,杨习贝. 基于容差关系的不完备可变精度多粒度粗糙集[J]. 计算机应用研究, 2013, 30(6): 1712-1715.
XU Wei, WU Chen, YANG Xibei. Incomplete variable precision multigranularity rough set based on tolerance relation[J]. Application research of computers, 2013, 30(6):1712-1715.

[19]徐久成, 张灵均, 孙林, 等. 广义邻域关系下不完备混合决策系统的约简[J]. 计算机科学, 2013, 40(4): 244-248.
XU Jiucheng, ZHANG Lingjun, SUN Lin, et al. Reduction in incomplete hybrid decision systems based on generalized neighborhood relationship[J]. Computer science, 2013, 40(4): 244-248.

作者简介:

王映龙,男,1970 年生,教授,博士,主要研究方向为知识发现、数据挖掘和计算智能。

曾淇,女,1991 年生, 硕士研究生,主要研究方向为粗糙集理论与知识发现。

钱文彬,男,1984 年生,讲师,博士,主要研究方向为粗糙集、粒计算与知识发现。