文章编号: 1001- 2486(2008) 06- 0083- 06

杂合数据的粗糙集属性约简方法*

谭 旭, 唐云岚, 张少 宸, 陈英武
(国防科技大学 信息系统与管理学院, 湖南 长沙 410073)

摘 要: 针对决策表中属性取值为杂合数据的情况, 提出了基于粗糙集理论的属性约简算法。首先给出了对象间在杂合数据下的相似度计算定义。为了获取合理的对象集合的软划分, 给出了阈值计算的最优化模型, 并基于粗糙集的上、下近似的概念, 得到对象集合在条件属性下的上、下近似的覆盖划分。之后, 通过各对象基于条件属性和决策属性的上、下近似下的分布矩阵描述, 利用最大分布矩阵, 直观地得到两种不同观点下的约简结果。实验结果表明了本算法的合理和有效性。

关键词: 杂合数据; 属性约简; 上、下近似覆盖划分; 分布矩阵; 粗糙集
中图分类号: TP18 文献标识码: A

Rough Set Based Attribute Reduction Algorithm for Hybrid Data

TAN Xu, TANG Yun-lan, ZHANG Shao-ding, CHEN Ying-wu
(College of Information System and Management, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: With regard to the attribute values in decision table, which are described with hybrid data, a new algorithm of attribute reduction based on rough set theory is proposed. First, the similarity relations among objects with hybrid data are defined. In order to obtain reasonable soft partitions among objects, the optimization model for threshold accounting is presented. Then, based on the upper and lower approximation concept from rough set theory, the covering upper and lower similar partitions among objects are obtained. In succession, through descriptions of the upper and lower similar distribution matrixes found on condition attributes and decision attribute, the two attribute reduction results of different viewpoints can be retrieved intuitively, based on the max-distribution matrixes. Finally, the experiment results prove that this algorithm is effective and feasible.

Key words: hybrid data; attribute reduction; covering upper and lower similar partition; distribution matrix; rough set theory

粗糙集理论是由 Pawlak 在传统集合论扩展的基础上得到的一种处理不精确、不完备数据的软计算方法。按照粗糙集的观点, 我们对事物的认知程度取决于所拥有知识的多少, 知识越多通常对事物间的区分越精细, 知识越少则区分得越模糊。然而在某个具体的决策表中, 各属性(知识) 并不是同等重要的, 通过属性(知识) 的约简可以找到一个较小的属性集来对整个决策表进行描述, 达到消除冗余的同时能更深入地认识系统。因此属性约简成为粗糙集理论中的核心问题。

许多学者对粗糙集理论下的属性约简算法进行了深入而广泛的研究。大体上可以划分为基于代数形式的属性约简算法和基于信息观点下的启发式属性约简算法。Skowron[1] 提出的分辨矩阵和分辨函数方法是较有代表性的基于代数形式的属性约简算法, 尽管有很多围绕该算法的改进和扩展, 但其约简复杂度和准确性限制了其进一步的发展。将属性的信息熵作为约简的启发知识来进行属性约简也是具有代表性的主流方法[2]。但是大多数的属性约简算法都基于数据的离散化预处理。然而离散化处理的断点选取对属性约简的结果异常敏感, 如何选取合适的离散化断点通常是一个非常棘手的问题[3]。

越来越多的学者开始关注约简中决策表数据本身的问题, 如数据的模糊性、数据的不协调性以及数据的缺失。数据的模糊软划分[4] 将逐渐取代离散化的硬划分方式。但是选择一种怎样的软划分模式, 软划分中的阈值该如何选取, 这仍然是探索中的问题。针对决策表中数据取值多样性的问题, 即杂合数据的情况, 目前的探讨还不是很多。Bhatt[5] 对软划分给出了理论上的深入分析, 并在文中初步提到了杂

* 收稿日期: 2008- 04- 02
作者简介: 谭旭(1981—) , 男, 博士生。

合数据的处理,但没有应用到属性的约简。Hu 对杂合数据的属性约简问题进行了深入而细致的研究[6-7],并得到了乐观的约简结果,但对于上、下近似阈值的单独设定和处理方式,容易造成整个约简的不确定性。

本文提出了一种针对解决杂合数据决策表属性约简的算法,考虑决策表的协调和不协调性以及数据的多样性,引入粗糙集的思想对对象集合进行上、下近似的模糊软划分,并给出了一种基于上、下近似分布矩阵的判定准则。最后能得到在上、下近似的不同粒度和划分的情况下,两种不同的属性约简结果,以满足不同层次的需求。

1 杂合数据决策表及粗糙集的相关讨论

当决策表的属性取值既有连续型的数值数据,又有语言文字型的描述数据,这样的决策表称之为杂合数据决策表。

定义1 一个杂合数据决策表 T 可以表达为有序五元组 $T=\{U, C^G \cup C^H, D, V, f\}$。$U=\{o_1, o_2, \cdots, o_n\}$ 为决策表中全体数据对象的集合,C^G 为数值型条件属性集,C^H 为描述型条件属性集,记 $C=C^G \cup C^H$,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 为信息函数,用于确定 U 中每一个对象在各个属性下的取值。

本文我们仅讨论针对条件属性的属性约简问题,即条件属性取值为杂合数据的情况。

定义2 设 T 为一个决策表,U 为该决策表的有限论域,条件属性集上的等价关系 $\mathrm{ind}(C)$ 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 在条件属性集合 $C^G \cup C^H$ 上的划分,划分结果为 $\{X_1, X_2, \cdots, X_w\}$,其中 $[o_i]_C=\{o_j \mid (o_i, o_j)\in \mathrm{ind}(C)\}$。记 $U/\mathrm{ind}(D)=\{[o_i]_D \mid o_i\in U\}$ 为决策属性 D 在 U 上的划分,划分结果为 $\{Y_1, Y_2, \cdots, Y_q\}$。若 $[o_i]_C\subseteq[o_i]_D$,称该决策表为协调决策表;否则,称为非协调决策表。

对于协调决策表的属性约简,一般要求不改变决策表的协调性;对于非协调决策表的约简,即保证分类正确率不变[8]。实际应用中,更多的是处理非协调决策表的情形。下面给出正确分类率的定义。

描述型数据是指属性值取值为语言文字的数据,描述型数据根据描述信息的由强到弱有不同的语言标度 $P=\{p_k \mid k=1, 2, \cdots, l\}$,$l>1$ 为语言标度的个数。比如 $P=\{p_1, p_2, p_3, p_4, p_5\}=\{$浓,强,中,弱,淡$\}$。令 $\xi(p_k)=k$,下面给出描述型数据之间的相似度运算。

定义3 $\forall o_i, o_j\in U, c\in C^H$,若 $f(o_i, c)=p_a, f(o_j, c)=p_b$,则对象 o_i, o_j 在属性 c 下取值为描述型数据 p_a 与 p_b 之间的相似度可以计算为 $s_{ij}^c=|\xi(p_a)-\xi(p_b)|/[\xi(p_l)-\xi(p_1)]$。

显然 s_{ij}^c 取值于 $[0, 1]$,s_{ij}^c 越接近1表明对象 o_i, o_j 在该描述型属性 c 下的取值差异越大,s_{ij}^c 越接近0表明对象 o_i, o_j 在该描述型属性 c 下的取值越接近。

性质1 (1) $\xi(f(o_i, c)-f(o_j, c))=\xi(f(o_i, c))-\xi(f(o_j, c))$;

(2) 若 $\xi(f(o_i, c))>\xi(f(o_j, c))$, $\xi(f(o_j, c))>\xi(f(o_h, c))$,则 $\xi(f(o_i, c))>\xi(f(o_h, c))$;

(3) $|\xi(f(o_i, c)-f(o_j, c))|>|\xi(f(o_i, c)-f(o_h, c))|$,则 $s_{ij}^c>s_{ih}^c$;

(4) $s_{ij}^c=s_{ji}^c$。

证明 由定义3容易得证。

根据定义3和性质1,可以很容易地计算决策表中各对象间在各个描述型属性下的相似度关系。对于连续数值型属性下的数据,给出如下的相似度计算。

定义4 $\forall o_i, o_j\in U, c\in C^G$,令 $\max_c=\max_{1\leqslant i\leqslant n} f(o_i, c)$, $\min_c=\min_{1\leqslant i\leqslant n} f(o_i, c)$,则对象 o_i, o_j 在数值属性 c 下的相似度计算为 $s_{ij}^c=\begin{cases}|f(o_i, c)-f(o_j, c)|/(\max_c-\min_c), & \max_c\neq\min_c\\ 0, & \max_c=\min_c\end{cases}$。

通过定义3和定义4的杂合数据决策表中各类型条件属性下对象间的相似度计算方式,可以得到决策表中各个对象间的相似度,对象间 $(o_i, o_j\in U)$ 的相似度定义为 $s_{ij}=\frac{c_1}{s_{ij}^{c_1}}+\frac{c_2}{s_{ij}^{c_2}}+\cdots+\frac{c_z}{s_{ij}^{c_z}}$,其中,$c_1, c_2, \cdots, c_z\in C^G\cup C^H$,$z$ 为条件属性的个数。记 $\max_{ij}=\max_{1\leqslant x\leqslant z} s_{ij}^{c_x}$, $\min_{ij}=\min_{1\leqslant x\leqslant z} s_{ij}^{c_x}$。

给出了对象间相似度的计算后, 可以很方便地对决策表中的对象集合进行软划分。借鉴粗糙集理论的上、下近似的观点, 定义 5 给出了对象集合的上、下近似下的覆盖划分的数学描述。

定义 5　设 $T=\{U, C, D, V, f\}$ 为杂合数据决策表, $0<\varepsilon\leqslant 0.5$ 为相似阈值, 则 U 在条件属性集 $C=C^G\cup C^H$ 下的 ε- 下近似划分为 $\underline{R}_C^{\varepsilon}=\{[o_i]_C^{\varepsilon}\mid o_i\in U\}$, 其中, $[o_i]_C^{\varepsilon}=\{o_j\mid \max_{ij}\leqslant\varepsilon, o_j\in U\}$; ε- 上近似为 $\overline{R}_C^{\varepsilon}=\{\overline{[o_i]}_C^{\varepsilon}\mid o_i\in U\}$, 其中 $\overline{[o_i]}_C^{\varepsilon}=\{o_j\mid \min_{ij}\leqslant\varepsilon, Bel(o_i, o_j)>0.5, o_j\in U\}$。这里, $C_{ij}^{\varepsilon}=\{c\mid s_{ij}^{c}\leqslant\varepsilon, c\in C^G\cup C^H\}$, $Bel(o_i, o_j)=Card(C_{ij}^{\varepsilon})/Card(C)$ 为对象 o_i 与 o_j 在 ε 下划分至同一等价类的信任度。

那么, 通过设定相似阈值 ε, 根据定义可以得到杂合数据决策表中的对象集合在条件属性集下的 ε- 上、下近似覆盖划分为 $\{(\overline{X_1}; \overline{Bel_1}), (\overline{X_2}; \overline{Bel_2}), \cdots, (\overline{X_s}; \overline{Bel_s})\}$ 和 $\{(\underline{X_1}; \underline{Bel_1}), (\underline{X_2}; \underline{Bel_2}), \cdots, (\underline{X_t}; \underline{Bel_t})\}$。其中, $\overline{Bel_s}=\min_{o_i, o_j\in X_s}(Bel(o_i, o_j))$, $\underline{Bel_t}=\min_{o_i, o_j\in X_{\perp}}(Bel(o_i, o_j))$。

要进一步说明的是, ε- 下近似覆盖划分与 ε- 上近似覆盖划分体现了对对象集合划分的不同粒度, ε- 下近似覆盖划分使得划分更加精细, 粒度更小, 得到的知识更加松散; 而 ε- 上近似覆盖划分使得划分更加粗犷, 粒度更大, 得到的知识更加精简。

2　杂合数据下粗糙集属性约简

2.1　约简原理

根据对象集合在决策属性上的划分 $\{Y_1, Y_2, \cdots, Y_q\}$ 以及在条件属性集合 C 上的 ε- 上、下近似覆盖划分, 可以得到上、下近似的 $n\times q$ 分布矩阵 $\overline{\boldsymbol{M}}_C$、$\underline{\boldsymbol{M}}_C$。其中, n 为决策表中对象的数目, q 为对象集合在决策属性上的划分数目。$\overline{m}_{uv}$, $\underline{m}_{uv}$ $(1\leqslant u\leqslant n, 1\leqslant v\leqslant q)$ 定义如下:

定义 6　$\forall o_i\in U$, C 为条件属性集合, D 为决策属性, $[o_i]_C^{\varepsilon}$ 为对象 o_i 在条件属性集 C 下所属的 ε- 下近似等价覆盖划分类集合, $\overline{[o_i]}_C^{\varepsilon}$ 为对象 o_i 在条件属性集 C 下所属的 ε- 上近似等价覆盖划分类集合, 则定义 $\underline{m}_{uv}=\sup_{\underline{X}_i\subset[o_u]_C^{\varepsilon}}\frac{|Y_v\cap\underline{X}_i|}{|\underline{X}_i|}$, $\overline{m}_{uv}=\sup_{\overline{X}_i\subset[o_u]_C^{\varepsilon}}\frac{|Y_v\cap\overline{X}_i|}{|\overline{X}_i|}$。

定理 1　设 T 为杂合数据决策表, 若 $c\in C$ 为冗余条件属性, 则有 $\overline{\boldsymbol{M}}_C=\overline{\boldsymbol{M}}_{C-\{c\}}$ 或 $\underline{\boldsymbol{M}}_C=\underline{\boldsymbol{M}}_{C-\{c\}}$。并称 c 为上、下近似分布矩阵下的可约简属性。

证明　若 T 为协调决策表, $c\in C$ 为冗余条件属性, 必存在唯一的 $Y_v\subset U/\mathrm{ind}(D)$ 使得 $|[o_u]_C^{\varepsilon}\cap Y_v|/|[o_u]_C^{\varepsilon}|=|[o_u]_{C-\{c\}}^{\varepsilon}\cap Y_v|/|[o_u]_{C-\{c\}}^{\varepsilon}|=1$。且 $\forall o_u\in U$, 唯一存在 $\underline{X}_t\subset\underline{R}_C^{\varepsilon}$ 使得 $[o_u]_C^{\varepsilon}=\underline{X}_t$, 那么 $\sup_{X_t\subset[o_u]_C^{\varepsilon}}\frac{|Y_v\cap\underline{X}_t|}{|\underline{X}_t|}=\sup_{X_s\subset[o_u]_{C-\{c\}}^{\varepsilon}}\frac{|Y_v\cap\underline{X}_s|}{|\underline{X}_s|}$ 成立, 故 $\underline{m}_{uv}^{C}=\underline{m}_{uv}^{C-\{c\}}$, 即 $\underline{\boldsymbol{M}}_C=\underline{\boldsymbol{M}}_{C-\{c\}}$。

若 T 为不协调决策表, $c\in C$ 为冗余条件属性, 则 $\forall[o_u]_C^{\varepsilon}\subset\underline{R}_C^{\varepsilon}$, $Y_v\subset U/\mathrm{ind}(D)$ 有 $|[o_u]_C^{\varepsilon}\cap Y_v|/|[o_u]_C^{\varepsilon}|=|[o_u]_{C-\{c\}}^{\varepsilon}\cap Y_v|/|[o_u]_{C-\{c\}}^{\varepsilon}|\leqslant 1$。那么对于每个 $[o_u]_C^{\varepsilon}=\underline{X}_t$, $[o_u]_{C-\{c\}}^{\varepsilon}=\underline{X}_s$ 都有 $\frac{|Y_v\cap\underline{X}_t|}{|\underline{X}_t|}=\frac{|Y_v\cap\underline{X}_s|}{|\underline{X}_s|}$ 成立, 那么 $\sup_{X_t\subset[o_u]_C^{\varepsilon}}\frac{|Y_v\cap\underline{X}_t|}{|\underline{X}_t|}=\sup_{X_s\subset[o_u]_{C-\{c\}}^{\varepsilon}}\frac{|Y_v\cap\underline{X}_s|}{|\underline{X}_s|}$ 一定成立, 即有 $\underline{\boldsymbol{M}}_C=\underline{\boldsymbol{M}}_{C-\{c\}}$。

对于对象集合在 ε- 上近似覆盖划分的情况, 同理可证。证毕。

定义 7　令 $\underline{\eta}_u=\max_v(\underline{m}_{uv})$, $\overline{\eta}_u=\max_v(\overline{m}_{uv})$; 记 $\overline{z}_{uv}=\begin{cases}1, & \overline{m}_{uv}=\overline{\eta}_u\\ 0, & \text{else}\end{cases}$, $\underline{z}_{uv}=\begin{cases}1, & \underline{m}_{uv}=\underline{\eta}_u\\ 0, & \text{else}\end{cases}$; 称 $\overline{\boldsymbol{Z}}_C=(\overline{z}_{uv})_{n\times q}$ 和 $\underline{\boldsymbol{Z}}_C=(\underline{z}_{uv})_{n\times q}$ 分别为条件属性集合 C 下 ε- 上近似、ε- 下近似的最大分布矩阵。

若 $\overline{\boldsymbol{Z}}_C=\overline{\boldsymbol{Z}}_{C-\{c\}}$, $\underline{\boldsymbol{Z}}_C=\underline{\boldsymbol{Z}}_{C-\{c\}}$, 则称 c 为上、下近似最大分布矩阵下的可约简属性。

定理 2　设 (U, C, D, V, f) 为杂合数据决策表, 则

(1) 上近似分布矩阵下的可约简属性必为上近似最大分布矩阵下的可约简属性;

(2) 下近似分布矩阵下的可约简属性必为下近似最大分布矩阵下的可约简属性。

证明　(1) $\forall c\in C$, 若 c 为上近似分布矩阵下的可约简属性, 根据定理 1, 有 $\overline{\boldsymbol{M}}_C=\overline{\boldsymbol{M}}_{C-\{c\}}$。也就是

在上近似分布矩阵中保持了决策表中各对象在决策属性集合中的概率分布 $\overline{m}_{uv}$ 不变化, 即 $\overline{m}_{uv}^{C}=\overline{m}_{uv}^{C-\{c\}}$。从而 $\overline{\eta}_u=\max_v(\overline{m}_{uv})$ 也不会发生变化, 根据定义7, 可以知道 $\overline{z}_{uv}^{C}=\overline{z}_{uv}^{C-\{c\}}$。所以 $\overline{\boldsymbol{Z}}_C=\overline{\boldsymbol{Z}}_{C-\{c\}}$ 成立。

(2) 证明同上近似分布的情况。证毕。 □

2.2 约简算法

为了得到决策表中条件属性集合的 ε- 上、下近似约简结果, 首先要给出阈值 ε 的求解算法。根据对象集合划分的信息论的观点[2], 若选取的阈值 ε能使得在条件属性 C 下的上、下近似划分达到最小的信息熵值, 那么选取该阈值为杂合数据决策表的划分阈值。

定义8 定义条件属性集合 C 下的 ε- 上近似信息熵为 $\overline{E}_\varepsilon(C)=-\overline{\gamma}\sum_{i=1}^{s}(|\overline{X}_i|/|U|)\cdot\sum_{j=1}^{q}pro_j\cdot\lg(pro_j)$, s 为 ε- 上近似下的对象集合在条件属性集合 C 下的覆盖划分集合的数目, q 为对象集合在决策属性 D 下的划分集合的数目。其中 $pro_j=|N_{\overline{X}_i\to Y_j}|/|\overline{X}_i|$, $|N_{\overline{X}_i\to Y_j}|$ 为划分集合 $\overline{X}_i$ 中属于划分集合 Y_j 的对象的个数。$\overline{\gamma}=\sum_{i=1}^{s}(\lambda/s)$, λ 表示对象集在近似划分 $\overline{X}_i$ 内所属决策类的基数。同样, 定义条件属性集合 C 下的 ε- 下近似信息熵为 $\underline{E}_\varepsilon(C)=-\underline{\gamma}\sum_{i=1}^{t}(|\underline{X}_i|/|U|)\cdot\sum_{j=1}^{q}pro_j\cdot\lg(pro_j)$。

根据定义 8, 求解以下优化模型, 可以得到阈值 ε:

$$\min\ \overline{E}_\varepsilon(C)+\underline{E}_\varepsilon(C)$$

$$s.t.\ \begin{cases}\overline{E}_\varepsilon(C)=-\overline{\gamma}\sum_{i=1}^{s}(|\overline{X}_i|/|U|)\cdot\sum_{j=1}^{q}pro_j\cdot\lg(pro_j)>0\\ \underline{E}_\varepsilon(C)=-\underline{\gamma}\sum_{i=1}^{t}(|\underline{X}_i|/|U|)\cdot\sum_{j=1}^{q}pro_j\cdot\lg(pro_j)>0\\ 0<\varepsilon\leqslant 0.5\end{cases}\qquad(*)$$

通过阈值 ε, 很方便地将上、下近似的关系建立了关联。同时, 也仅需要通过一个阈值 ε 便可同时得到上、下近似下的两种不同粒度的划分。通过以上的最优化计算, 使得到的阈值更加合理、可信。

算法 1 杂合数据属性约简算法

输入 杂合数据决策表 $T=\{U, C^G\cup C^H, D, V, f\}$

输出 上、下近似属性约简集 $\overline{\mathrm{Red}}$ 和 $\underline{\mathrm{Red}}$

Step 1 根据决策属性 D 获取对象集合 U 的划分结果 $U/\mathrm{ind}(D)=\{Y_1, Y_2, \cdots, Y_q\}$, 令 $Att=Cond=C=C^G\cup C^H$, $\overline{\mathrm{Red}}=\underline{\mathrm{Red}}=\Phi$;

Step 2 计算对象集合 U 中各对象在条件属性集合 $Cond$ 下的相似度 $s_{ij}(1\leqslant i,j\leqslant|U|)$;

Step 3 根据各对象间的相似度值, 利用优化模型 $(*)$, 得到最优阈值 ε;

Step 4 计算出条件属性集合 $Cond$ 上的 ε- 上近似覆盖划分 $\overline{R}_C^\varepsilon=\{(\overline{X}_1;\overline{Bel}_1),(\overline{X}_2;\overline{Bel}_2),\cdots,(\overline{X}_s;\overline{Bel}_s)\}$ 和 ε- 下近似覆盖划分 $\underline{R}_C^\varepsilon=\{(\underline{X}_1;\underline{Bel}_1),(\underline{X}_2;\underline{Bel}_2),\cdots,(\underline{X}_t;\underline{Bel}_t)\}$;

Step 5 求取 ε- 上、下近似的分布矩阵 $\overline{\boldsymbol{M}}_{Cond}$ 及 $\underline{\boldsymbol{M}}_{Cond}$, 并得到 ε- 上、下近似的最大分布矩阵 $\overline{\boldsymbol{Z}}_{Cond}$ 及 $\underline{\boldsymbol{Z}}_{Cond}$;

Step 6 If $(Att!=\Phi)\ \&\ c$

If $Cond=C=C^G\cup C^H$, Then $\overline{\boldsymbol{Z}}_C:=\overline{\boldsymbol{Z}}_{Cond}$, $\underline{\boldsymbol{Z}}_C:=\underline{\boldsymbol{Z}}_{Cond}$, 按序取 $c\in Att$, $Att=Att-\{c\}$, $Cond=C-\{c\}$, 转 Step 2;

Else

If $\overline{\boldsymbol{Z}}_C!=\overline{\boldsymbol{Z}}_{Cond}$ then $\overline{\mathrm{Red}}=\overline{\mathrm{Red}}+\{c\}$, If $\underline{\boldsymbol{Z}}_C!=\underline{\boldsymbol{Z}}_{Cond}$ then $\underline{\mathrm{Red}}=\underline{\mathrm{Red}}+\{c\}$; 按序取 $c\in Att$, $Att=Att-\{c\}$, $Cond=C-\{c\}$, 转 Step 2;

Endif

Else terminate

Endif

3　实 验

3.1　烤烟烟叶分类中的属性约简实例

利用烤烟烟叶的一个实际分类数据为例来详细阐述本文所提出的方法。条件属性为描述型的属性是: 成熟度, 叶片结构, 身份, 油分, 色度; 连续数值型的属性为长度和残伤。决策属性为“烟叶部位”。“成熟度”的语言标度为{完熟, 成熟, 尚熟, 欠熟, 假熟},“叶片结构”的语言标度为{疏松, 尚疏松, 稍密, 紧密},“身份”的语言标度为{厚, 稍厚, 中等, 稍薄, 薄},“油分”的语言标度为{多, 有, 稍有, 少},“色度”的语言标度为{浓, 强, 中, 弱, 淡}[9]。详细决策表数据见表 1。

表 1　烤烟烟叶分类决策表

Tab. 1　Decision table of flue-cured tobacco leaves' classification

对象	成熟度	叶片结构	身份	油分	色度	长度(cm)	残伤(%)	烟叶部位
1	假熟	疏松	薄	少	淡	25	35	上部
2	成熟	疏松	薄	稍有	弱	28	30	上部
	…			…			…	
7	成熟	疏松	中等	多	浓	45	10	中部
	…			…			…	
13	成熟	稍密	厚	稍有	中	35	35	下部
14	成熟	尚疏松	稍厚	有	强	40	20	下部

容易得到, 对象集合在决策属性上的分类为 $U/\mathrm{ind}(D)=\{\{1,2,3,4\};\{5,6,7,8\};\{9,10,11,12,13,14\}\}$。

首先通过各对象间的相似度计算得到决策表中各对象间的相似度。之后, 根据文中给出的优化模型进行阈值寻优计算, 得到近似划分的阈值为 $\varepsilon=0.34$。则, ε- 上近似覆盖划分为{({1, 2, 3, 4, 9, 12}; 0.57), ({5, 6, 7, 8, 10, 11, 14}; 0.71), ({9, 10, 11, 12, 13}; 1.0)}; ε- 下近似覆盖划分为{({1, 4}; 1.0), ({2, 3}; 1.0), ({5, 6, 7, 8, 14}; 1.0), ({9, 10, 11, 12, 13}; 1.0)}。根据分布矩阵的定义, 可以得到该决策表中对象集合在全体条件属性集合上的上、下近似分布矩阵 $\overline{\boldsymbol{M}_C}$ 和 $\underline{\boldsymbol{M}_C}$。

根据属性约简算法的迭代计算, 可以得到最终的 ε- 上近似属性约简结果为 $\overline{\mathrm{Red}}$={身份, 色度, 长度}, ε- 下近似属性约简结果为 $\underline{\mathrm{Red}}$={叶片结构, 身份, 色度, 残伤}。

3.2　属性约简的算法比较

为了验证本文约简算法的可行性和正确性, 选取 UCI 标准机器学习数据库中的 3 个有代表性的含有杂合数据的数据集合 E. coli、Australian credit 以及 Yeast 进行测试[10]。和本文作比较的 2 个有代表性的约简算法是: 传统的将数值型数据离散化后进行属性约简的算法[11] 以及文献[7]提出的模糊粗糙集属性约简算法。属性约简完成后, 采用十折交叉法, 参照文献[7]中提到的两种经典的分类学习算法(CART 和 RBF-SVM)进行分类预测。

表 2 展示了在 CART 算法下, 3 个数据集在不同的约简算法下的对比结果。其中 A1 代表传统的数据离散化的约简算法, A2 代表文献[7]提出的约简算法, A3 为本文提出的约简算法。其中 A3 算法下的前段代表 ε- 上近似下的属性约简和分类结果, 后段代表 ε- 下近似下的属性约简和分类结果。Attr 记录约简后的条件属性个数, Acc 为分类正确率, Mis 为误分类率。在用本文算法对 E. coli 数据集进行属性约简时, 最优阈值取 $\varepsilon=0.21$; 对 Australian credit 数据集进行属性约简时, 最优阈值取 $\varepsilon=0.24$; 对 Yeast 数据集进行属性约简时, 最优阈值取 $\varepsilon=0.33$。

表 2　CART 算法下各属性约简方法的比较

Tab. 2　Comparison of manifold reduction method based on CART algorithm

数据集	A1			A2			A3					
	Attr	Acc	Mis	Attr	Acc	Mis	Attr	Acc	Mis	Attr	Acc	Mis
E. coli	1	0.426±0.017	0.213	7	0.820±0.044	0.165	5	0.805±0.081	0.173	7	0.851±0.064	0.142
Australian credit	12	0.827±0.140	0.158	13	0.814±0.142	0.074	7	0.821±0.075	0.082	10	0.862±0.041	0.040
Yeast	4	0.357±0.102	0.246	8	0.703±0.082	0.197	5	0.689±0.101	0.208	8	0.803±0.051	0.117

同样,表3展示了在RBF-SVM算法下,3个数据集在不同的属性约简算法下的对比结果。综合表2和表3的对比数据可以看出,本文提出的算法基本上能得到尽可能少的条件属性,同时取得较好的正确分类率以及较低的误分类率。

表3 RBF-SVM算法下各属性约简方法的比较

Tab.3 Comparison of manifold reduction method based on RBF-SVM algorithm

数据集	A1			A2			A3					
	Attr	Acc	Mis	Attr	Acc	Mis	Attr	Acc	Mis	Attr	Acc	Mis
E.coli	1	0.426±0.017	0.231	7	0.851±0.059	0.158	5	0.825±0.062	0.173	7	0.869±0.030	0.087
Australian credit	12	0.806±0.089	0.102	13	0.814±0.072	0.063	7	0.827±0.045	0.071	10	0.866±0.028	0.033
Yeast	4	0.401±0.063	0.248	8	0.706±0.021	0.201	5	0.693±0.057	0.207	8	0.811±0.044	0.144

可以看到,A1属性约简算法得到的平均核属性是最少的,但是正确分类率要大大低于本文A3的属性约简算法以及A2属性约简算法。而A2属性约简算法虽保持了较好的正确分类率(平均正确分类率要略低于A3算法),但是约简后的核属性数目均大于或等于A3算法。在误分类率上,相对A1,A2算法,本文的算法基本保持了较低的水平。对于决策类别数目较多、对象集合数目较大的决策表(如Yeast数据集),尤其体现了本算法的优越性。从实验结果中进一步可以看到,A3中的 ε- 上近似约简能得到较少的条件属性数目,但正确分类率要略低于 ε- 下近似约简;而 ε- 下近似约简后的决策表虽能得到更好的正确分类率,但约简后的条件属性数目要明显多于 ε- 上近似约简,表明这需要耗费更多的运算时间。不论如何,相比于其他的两种属性约简算法,A3算法体现出了其优越性。

4 结论

本文研究了基于粗糙集理论的杂合数据属性约简问题。首先给出了不同类型数据下的相似度定义,然后引入粗糙集的上、下近似的思想对对象集合进行模糊软划分,并给出了相应的可信度。在划分阈值的处理上,考虑了决策属性划分对条件属性覆盖划分的影响,结合信息熵的计算,给出了最优化模型。进一步,为了克服决策表中数据过于严苛带来的限制以及模糊覆盖划分可能带来的误差,给出了上、下近似最大分布矩阵的概念,通过上、下近似最大分布矩阵的比较,可以判定条件属性的可约简性,并最终得到该杂合数据表的约简条件属性集。

通过在实际烤烟烟叶分类决策表的属性约简的实例,并与传统的基于断点划分的属性约简方法以及Hu的模糊粗糙杂合数据约简方法的比较,可以得出本算法解决杂合数据决策表的属性约简问题是可行且有效的。另外,需要进一步说明的是,本算法不仅能较好地处理杂合数据的属性约简问题,对单一的数值型数据或描述型数据决策表的属性约简问题也是适用的,因为它们仅是杂合数据表的特例,所以本算法具有普适性。

参考文献:

[1] Skowron A, Rauszer C. Intelligent Decision Support-handbook of Applications and Advances of the Rough Set Theory[M]. Dordrecht: Kluwer Academic Publishers, 1992.

[2] 王国胤, 于洪, 杨大春. 基于条件信息熵的决策表约简[J]. 计算机学报, 2002, 25(7): 759-766.

[3] 肖迪, 胡寿松. 实域粗糙集理论及属性约简[J]. 自动化学报, 2007, 33(3): 253-258.

[4] Sarkar M. Fuzzy-rough Nearest Neighbor Algorithms in Classification[J]. Fuzzy Sets and Systems, 2007, 158(19): 2134-2152.

[5] Bhatt R B, Gopal M. On the Extension of Functional Dependency Degree from Crisp to Fuzzy Partitions[J]. Pattern Recognition Letters, 2006, 27: 487-491.

[6] Hu Q H, Yu D R, Xie Z X. Information-preserving Hybrid Data Reduction Based on Fuzzy-rough Techniques[J]. Pattern Recognition Letters, 2006, 27: 414-423.

[7] Hu Q H, Xie Z X, Yu D R. Hybrid Attribute Reduction Based on Anovel Fuzzy-rough Model and Information Granulation[J]. Pattern Recognition, 2007, 40(12): 3509-3521.

[8] 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.

[9] 胡开文. 烟叶打叶复烤工艺与设备[M]. 北京: 化学工业出版社, 2002.

[10] Blake C L. UCI Repository of Machine Learning Databases[EB/OL]. http://www.ics.uci.edu/~mlearn/MLRepository.

[11] Swiniarski R W, Skowron A. Rough Set Methods in Feature Selection and Recognition[J]. Pattern Recognition Letters, 2003, 24: 833-849.