DOI: 10.11992/tis.201804059
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20180608.1405.006.html

集值信息系统的快速正域约简

陈曼如[1], 张楠[1], 童向荣[1], 岳晓冬[2]
（1. 烟台大学 数据科学与智能技术山东省高校重点实验室, 山东 烟台 264005; 2. 上海大学 计算机工程与科学学院, 上海 200444）

摘　要: 针对集值信息系统正域约简算法在大规模数据集下的运行效率问题, 提出一种基于启发式的集值信息系统快速正域约简算法。通过研究属性和对象在约简过程中对算法运行效率产生的影响, 在集值信息系统中引入属性无关性和属性重要度保序性的相关定义, 介绍了使得算法运行效率提升的相关定理、快速算法和应用实例。通过实验对提出算法的有效性进行分析和验证。实验表明, 提出算法的运行效率优于原始算法的运行效率。

关键词: 属性约简; 粗糙集; 集值信息系统; 特征选择; 启发式算法; 正域约简; 快速约简算法; 粗糙近似
中图分类号: TP18　　**文献标志码:** A　　**文章编号:** 1673−4785(2019)03−0471−08

中文引用格式: 陈曼如, 张楠, 童向荣, 等. 集值信息系统的快速正域约简[J]. 智能系统学报, 2019, 14(3): 471–478.
英文引用格式: CHEN Manru, ZHANG Nan, TONG Xiangrong, et al. Quick positive region reduction in set-valued information systems[J]. CAAI transactions on intelligent systems, 2019, 14(3): 471–478.

Quick positive region reduction in set-valued information systems

CHEN Manru[1], ZHANG Nan[1], TONG Xiangrong[1], YUE Xiaodong[2]
(1. Key Lab for Data Science and Intelligence Technology of Shandong Higher Education Institutes, Yantai University, Yantai 264005, China; 2. School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China)

Abstract: This study aims to propose a quick positive reduction algorithm based on the heuristic method to increase the efficiency of the set-valued positive reduction algorithm under large-scale data. The definitions of attribute independence and attribute importance isotonicity are introduced in the set-valued information system by investigating the influence of an attribute and object on the efficiency of algorithm during the reduction process, and the relevant theorem, fast algorithm, and practical example for improving the efficiency of the algorithm are introduced. Finally, the experimental results show the efficiency and effectiveness of the proposed method and its better efficiency in comparison to that of the original algorithm.

Keywords: attribute reduction; rough set; set-valued information systems; feature selection; heuristic algorithm; positive region reduction; quick algorithm reduction; rough approximations

粗糙集理论[1-2](rough set theory) 是分析、处理不精确和不确定数据的有效工具。由于粗糙集理论中的等价关系在实际应用中局限性较大, 各种基于二元关系的扩展粗糙集模型[3-8]得以发展。目前, 粗糙集理论已经广泛应用于数据挖掘、机器学习与模式识别等研究领域。

集值信息系统是重要的单值系统的扩展模型, 其中集值是用来描述不精确和缺失的信息, 集值信息系统在现实生活中的应用广泛。近年来, 许多学者对集值系统做了深入、大量的研究工作。Guan 等[9]提出最大相容类的定义解决了同一相容类中对象不一定两两相似的问题, 并且提

收稿日期: 2018−04−27.　**网络出版日期:** 2018−06−11.
基金项目: 国家自然科学基金项目 (61403329, 61572418, 61702439, 61572419, 61502410); 山东省自然科学基金项目 (ZR2018BA004, ZR2016FM42); 烟台大学研究生科技创新基金项目 (YDZD1807).
通信作者: 张楠. E-mail: zhangnan0851@163.com.

出 A-相对约简和 E-相对约简。Qian 等[10]提出基于优势关系的合取型集值序信息系统和吸取型集值序信息系统并构建了两种粗糙集模型。杨习贝等[11]在集值系统中提出模糊优势关系，考虑到了对象之间的优势程度。Huang 等[12]提出概率集值信息系统和基于巴氏距离的 λ-相容关系，更合理地描述了在概率集值信息系统对象间的关系。Wei 等[13]基于模糊相似类和模糊相似度提出两种不同的模糊粗糙集模型。Zhang 等[14]将量化粗糙集和优势粗糙集结合，提出了一种基于大规模集值信息系统的特征选择和近似推理的通用框架。

Skowron 等[15]认为差别矩阵方法虽然可以求得数据集的所有约简，但是其效率相对于启发式算法较低，故应用性较差。由于现实生活中数据信息量的不断增大，集值系统中属性约简算法的效率需要做进一步研究。罗川等[16]在集值序决策系统中通过计算粗糙近似提出了基于增加和删除策略的增量式算法；Zhang 等[17]在集值系统中提出了构造关系矩阵的基本方法并且通过研究属性集变化时关系矩阵变化的性质，得到了关系矩阵更新的增量式方法；刘莹莹等[18]在集值信息系统中通过定义一种新的相似度，提出基于相似度的启发式算法；马建敏等[19]在集值信息系统中引入信息量等概念，提出一种新的启发式约简算法。

1　基础知识

1.1　集值信息系统相关基础知识

定义 1　集值信息系统是一个四元组 $\mathrm{SIS}=(U,\mathrm{AT},V,f)$，其中：$U$ 是非空有限对象的集合，称为论域；AT 是非空有限属性的集合；V_a 是条件属性值集合，有 $V=\cup_{a\in\mathrm{AT}}V_a$；对于 $\forall a\in\mathrm{AT}$ 满足 $f(x,a)\in V_a$ 的集值映射为 $f:U\times\mathrm{AT}\to 2^V$。

若属性集合由条件属性集合 C 和决策属性集合 D 组成，$D\cap C=\varnothing$，$f:U\times C\to 2^{V_C}$ 为集值映射，$f:U\times D\to V_D$ 为单值映射，则该信息系统称为集值决策系统 $\mathrm{SDS}=(U,C\cup D,V,f)$。

为方便起见 $V_D=\{1,2,\cdots,r\}$，$U/\mathrm{IND}(D)=\{D_1,D_2,\cdots,D_r\}$，其中 $\mathrm{IND}(D)=\{(x_i,x_j)|(x_i,x_j)\in U^2,f(x_i,D)=f(x_j,D)\}$。$U/\mathrm{IND}(D)$ 表示 D 在 U 上的划分。

例 1　表 1 为集值决策系统 $\mathrm{SDS}=(U,C\cup D,V,f)$，其中：$U=\{x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8\}$ 为论域，$C=\{a_1,a_2,a_3,a_4,a_5\}$ 为条件属性集合，$D=\{d\}$ 为决策属性集合。

表 1　集值决策系统

Table 1　Set-valued decision systems

U	a_1	a_2	a_3	a_4	a_5	d
x_1	{2}	{1, 2}	{1}	{1}	{1}	1
x_2	{1}	{1, 2}	{1, 2}	{2}	{2}	1
x_3	{1}	{1, 2}	{1}	{1}	{2}	2
x_4	{2}	{1, 2}	{1, 2}	{2}	{1}	1
x_5	{1, 2}	{2}	{1}	{1}	{2}	1
x_6	{1, 2}	{1}	{2}	{1}	{2}	2
x_7	{1}	{1, 2}	{2}	{1}	{1, 2}	2
x_8	{1}	{1}	{1, 2}	{2}	{2}	1

定义 2　集值信息系统 $\mathrm{SIS}=(U,\mathrm{AT},V,f)$，$\forall A\subseteq\mathrm{AT}$，则 A 在 U 上的相容关系定义为

$$\boldsymbol{T}_A=\{(x,y)\in U\times U|\forall a\in A,a(x)\cap a(y)\neq\varnothing\}。$$

等价关系在集值信息系统不再成立，$\boldsymbol{T}_A$ 满足自反性和对称性，不一定满足传递性。设 $T_A(x)$ 为相容类，定义 $T_A(x)=\{y\in U|(x,y)\in\boldsymbol{T}_A\}$，表示与对象 x 关于相容关系 $\boldsymbol{T}_A$ 的最大不可分辨对象集合。$U/\boldsymbol{T}_A=\{T_A(x)|x\in U\}$ 形成对论域的一个覆盖，其中任意两个集合都可能相交。

定义 3　集值信息系统 $\mathrm{SIS}=(U,\mathrm{AT},V,f)$，$X\subseteq U$，$A\subseteq\mathrm{AT}$，则在相容关系 A 下集合 X 的下、上近似分别为 $\underline{A}(X)=\{x\in U|T_A(x)\subseteq X\}$ 和 $\overline{A}(X)=\{x\in U|T_A(x)\cap X\neq\varnothing\}$。

X 的下近似是在相容关系 A 下确定属于 X 的全体对象集合，而 X 的上近似是在相容关系 A 下可能属于 X 的全体对象集合。

根据定义 3 可得：$\mathrm{POS}_A(X)=\underline{A}(X),\mathrm{BN}_A(X)=\overline{A}(X)-\underline{A}(X)$。$\mathrm{POS}_A(X)$ 称为在相容关系 A 下集合 X 的正域，它表示在相容关系 A 下可以确定划分到集合 X 里的对象集合；$\mathrm{BN}_A(X)$ 为在相容关系 A 下集合 X 的边界域，它表示在相容关系 A 下不能完全确定划分是否属于集合 X 的对象集合。

定义 4　集值决策系统 $\mathrm{SIS}=(U,\mathrm{AT},V,f)$，$A\subseteq\mathrm{AT}$，$U/\mathrm{IND}(D)=\{D_1,D_2,\cdots,D_r\}$，则决策属性 D 在相容关系 A 的下、上近似分别为 $\underline{A}(D)=\{x\in U|D_j\in U/\mathrm{IND}(D),T_A(x)\subseteq D_j\}$ 和 $\overline{A}(D)=\{x\in U|D_j\in U/\mathrm{IND}(D),T_A(x)\cap D_j\neq\varnothing\}$。其中 $j\in\{1,2,\cdots,|U/\mathrm{IND}(D)|\}$。$\underline{A}(D)$ 也称为决策属性集合 D 在相容关系 A 下的正域，$\underline{A}(D)=\mathrm{POS}_A(D)$。

1.2　集值信息系统的正域约简算法

定义 5　集值决策系统 $\mathrm{SDS}=(U,C\cup D,V,f)$，$A\subseteq C$，则决策属性集合 D 在相容关系 A 下的近似分类质量定义为

$$\gamma_A(D)=\frac{|\mathrm{POS}_A(D)|}{|U|} \tag{1}$$

近似分类质量刻画了在相容关系 A 下可以正确划分到决策属性集合 D 的对象集合数与论域个数的比重。

定义 6　集值决策系统 $\mathrm{SDS}=(U,C\cup D,V,f)$, $A\subseteq C$, 则集合 A 是集值决策系统的一个正区域保持属性约简当且仅当以下两个条件成立:

1) $\gamma_A(D)=\gamma_C(D)$;

2) $\forall B\subset A$, $\gamma_B(D)<\gamma_A(D)$。

定义 7　集值决策系统 $\mathrm{SDS}=(U,C\cup D,V,f)$, $A\subseteq C$, $\forall a\in A$。条件属性 a 的内部属性重要度定义为

$$\mathrm{Sig}^{\mathrm{inner}}(a,A,C,D,U)=\gamma_A(D)-\gamma_{A-\{a\}}(D)$$

内部属性重要度刻画了属性集合中的每个属性的初始重要度, 用于选择不可缺少的属性子集, 在算法中用于计算属性约简的核属性集合。

定义 8　集值决策系统 $\mathrm{SDS}=(U,C\cup D,V,f)$, $A\subseteq C$, $\forall a\in C-A$, 条件属性 a 的外部属性重要度定义为

$$\mathrm{Sig}^{\mathrm{outer}}(a,A,C,D,U)=\gamma_{A\cup\{a\}}(D)-\gamma_A(D)$$

外部属性重要度刻画了属性在信息系统的重要程度, 用于在搜索过程中选择属性, 在迭代过程中, 每轮选择外部属性重要度最大的属性加入到候选属性集合中, 直到候选属性子集满足终止条件, 从而获得属性约简。

对于给定集值决策系统 $\mathrm{SDS}=(U,C\cup D,V,f)$, $a\in C$, 若 $\mathrm{Sig}^{\mathrm{inner}}(a,C,C,D,U)>0$, 则 a 属于集值决策系统 SDS 的核属性。

根据文献[20]给出集值信息系统下经典正域属性约简算法 (positive region reduction algorithm, PRRA), 见算法 1。

算法 1　PRRA

输入　集值决策系统 $\mathrm{SDS}=(U,C\cup D,V,f)$;

输出　集值决策系统 SDS 的一个属性约简 R。

1) $R=\varnothing$。

2) 计算 $\mathrm{Sig}^{\mathrm{inner}}(a_i,C,C,D,U)$, $i\in\{1,2,\cdots,|C|\}$, 如果 $\mathrm{Sig}^{\mathrm{inner}}(a_i,C,C,D,U)>0$, 则将 a_i 存入 R 中。

3) 若 $\gamma_R(D)=\gamma_C(D)$, 则终止; 否则对条件属性 $C-R$ 进行如下操作:

计算 $\mathrm{Sig}^{\mathrm{outer}}(a_0,R,C,D,U)=\max\{\mathrm{Sig}^{\mathrm{outer}}(a_k,C,D,R,U)\}$, $R\leftarrow R\cup\{a_0\}$, 其中 $a_k\in C-R$。若有多个最大值, 则选择与 R 组合数最少的属性。

4) 对 R 中的每个条件属性 a_i进行如下操作:

①计算 $\gamma_{R-\{a_i\}}(D)$;

②若$\gamma_{R-\{a_i\}}(D)=\gamma_C(D)$, 则 a_i为冗余属性, $R=R-\{a_i\}$, 否则 a_i 为非冗余属性, R 保持不变。

5) 返回 R。

由算法 1 可得, 在 3) 迭代的过程中, 每次选取一个属性重要度最大的属性加入到 R 中, 在这个过程中需要不断地计算整个论域上的正域, 从而使得算法计算量非常大, 效率不够理想, 很难应用在大规模数据集下。

例 2　表 1 为集值决策系统 $\mathrm{SDS}=(U,C\cup D,V,f)$, 其中, $U=\{x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8\}$ 为论域, $C=\{a_1,a_2,a_3,a_4,a_5\}$ 为条件属性集合, $D=\{d\}$ 为决策属性集合。

采用 PRRA 算法对表 1 约简过程如下。

1) 表 1 中经求其核属性集为 $R=\{a_3,a_4\}$, 决策属性 D 相对于条件属性 C 的近似分类质量为 $\gamma_C(D)=0.625$;

2) 对 $C-R$ 中每个条件属性重复进行如下操作:

根据定义 8 计算 $C-R$ 中属性的重要度, 其中, $\mathrm{Sig}^{\mathrm{outer}}(a_1,R,C,D,U)=0.125$, $\mathrm{Sig}^{\mathrm{outer}}(a_2,R,C,D,U)=0$, $\mathrm{Sig}^{\mathrm{outer}}(a_3,R,C,D,U)=0.125$。

由于属性 a_1 和属性 a_3 的外部属性重要度 $\mathrm{Sig}^{\mathrm{outer}}(a_1,R,C,D,U)=\mathrm{Sig}^{\mathrm{outer}}(a_3,R,C,D,U)=0.125$, 故选择与 R 组合数最少的属性 a_1。根据算法 1 中步骤 3) 将 a_1 加入 R 中, 得 $R=R\cup\{a_1\}=\{a_1,a_3,a_4\}$, 且 $\gamma_R(D)=\gamma_C(D)=0.625$, 终止步骤 3)。

3) 去除集合 R 中的冗余属性, 由于 $\gamma_{R-\{a_1\}}(D)\neq\gamma_C(D)$, 故集合 R 中无冗余属性, 算法终止。

故算法 PRRA 得到约简, $R=\{a_1,a_3,a_4\}$。

2　集值系统的快速正域约简

Qian 等[21-22]提出了正向近似的加速原理。本节在文献[21-22]算法的基础上介绍集值信息系统下的正域属性约简算法的加速原理以及相关性质, 给出了算法的伪代码描述和两个算法的时间复杂度对比与分析。

定理 1　集值决策系统 $\mathrm{SDS}=(U,C\cup D,V,f)$, 设属性子集 P、Q 满足 $P\subset Q\subseteq C$, 则 $\mathrm{POS}_P(D)\subseteq\mathrm{POS}_Q(D)$。

该定理表明, 如果两个属性集合存在包含关系, 则这两个属性集合相对于决策属性 D 的正域也存在包含关系。证略。

定理 2　集值决策系统 $\mathrm{SDS}=(U,C\cup D,V,f)$, 条件属性集合 $C=\{a_1,a_2,\cdots,a_{|C|}\}$, 属性集合 $P_i=\{a_1,a_2,\cdots,a_i\}$, $i=1,2,\cdots,|C|$, 则定义:

$$\mathrm{POS}_{P_{i+1}}(U,D)=\mathrm{POS}_{P_i}(U,D)\cup\mathrm{POS}_{P_{i+1}}(U_{i+1},D)。$$

式中: $U_1=U$, $U_{i+1}=U-\mathrm{POS}_{P_i}(U,D)$。

定理 3　集值决策系统 $\mathrm{SDS}=(U,C\cup D,V,f)$, $A,B\subseteq C$, 若 $U/\boldsymbol{T}_A=U/\boldsymbol{T}_B$, 则 $\mathrm{POS}_A(U,D)=\mathrm{POS}_B(U,D)$。

证明　设 $U/\boldsymbol{T}_A=\{X_1,X_2,\cdots,X_{|U|}\}$, $U/\boldsymbol{T}_B=\{Y_1,Y_2,\cdots,Y_{|U|}\}$, 如果 $U/\boldsymbol{T}_A=U/\boldsymbol{T}_B$, 则 $X_i=Y_i$, $i=1,2,\cdots,|U|$, 那

么 $\mathrm{POS}_A(U,D)=\mathrm{POS}_B(U,D)$, 证毕。

定理 4　集值决策系统 $\mathrm{SDS}=(U,C\cup D,V,f)$, $A\subseteq C$, $\forall a\in C-A$, 如果 $U/\boldsymbol{T}_A=U/\boldsymbol{T}_{A\cup\{a\}}$, 那么 $\mathrm{POS}_A(U,D)=\mathrm{POPOS}_{A\cup\{a\}}(U,D)$, 则属性 a 为冗余属性。

证明　设 $U/\boldsymbol{T}_A=\{X_1,X_2,\cdots,X_{|U|}\}$, $U/\boldsymbol{T}_{A\cup\{a\}}=\{Y_1,Y_2,\cdots,Y_{|U|}\}$, 因为 $U/\boldsymbol{T}_A=U/\boldsymbol{T}_{A\cup\{a\}}$, 即 $\mathrm{POS}_A(U,D)=\mathrm{POS}_{A\cup\{a\}}(U,D)$, 因此对于任意 $b\in C-A-\{a\}$, 有 $\mathrm{POS}_{A\cup\{b\}}(U,D)=\mathrm{POS}_{A\cup\{b\}\cup\{a\}}(U,D)$。所以 a 是约简过程中的冗余属性, 可以被删除。证毕。

定理 5　集值决策系统 $\mathrm{SDS}=(U,C\cup D,V,f)$, $A\subseteq C$, 如果 $\forall a,b\in C-A-A'$, 并且 $\mathrm{Sig}^{\mathrm{outer}}(a,A,C,D,U)-\mathrm{Sig}^{\mathrm{outer}}(b,A,C,D,U)\geqslant 0$, 那么 $\mathrm{Sig}^{\mathrm{outer}}(a,A,C',D,U^*)-\mathrm{Sig}^{\mathrm{outer}}(b,A,C',D,U^*)\geqslant 0$。其中, 属性集合 $C'=C-A'$, $A'=\{c|\mathrm{POS}_A^C(U,D)=\mathrm{POS}_{A\cup\{c\}}^C(U,D),c\in C-A\}$, $U^*=U-\mathrm{POS}_A^C(U,D)$。

证明　根据定义 $\mathrm{Sig}^{\mathrm{outer}}(a,A,C,D,U)=\gamma_{A\cup\{a\}}(D)-\gamma_A(D)$, 要度仅取决于 $\gamma_A(D)=|\mathrm{POS}_A(D)|/|U|$。由于 $U^*=U-\mathrm{POS}_A^C(U,D)$, 所以 $\mathrm{POS}_A^C(U^*,D)=\varnothing$, $\mathrm{POS}_{A\cup\{a\}}^C(U^*,D)=\mathrm{POS}_{A\cup\{a\}}^C(U,D)-\mathrm{POS}_A^C(U,D)$。又因为 $C'\cap A'=\varnothing$, $A\subset C'$, 故 $\mathrm{POS}_A^{C'}(U^*,D)=\mathrm{POS}_A^C(U^*,D)$。因此

$$\frac{\mathrm{Sig}^{\mathrm{outer}}(a,A,C,D,U)}{\mathrm{Sig}^{\mathrm{outer}}(a,A,C',D,U^*)}=\frac{\gamma_{A\cup\{a\}}^{(C,U)}(D)-\gamma_A^{(C,U)}(D)}{\gamma_{A\cup\{a\}}^{(C',U^*)}(D)-\gamma_A^{(C',U^*)}(D)}=$$
$$\frac{|U^*|}{|U|}\frac{|\mathrm{POS}_{A\cup\{a\}}^C(U,D)|-|\mathrm{POS}_A^C(U,D)|}{|\mathrm{POS}_{A\cup\{a\}}^{C'}(U^*,D)|-|\mathrm{POS}_A^{C'}(U^*,D)|}=$$
$$\frac{|U^*|}{|U|}\frac{|\mathrm{POS}_{A\cup\{a\}}^C(U,D)|-|\mathrm{POS}_A^C(U,D)|}{|\mathrm{POS}_{A\cup\{a\}}^C(U^*,D)|-|\mathrm{POS}_A^C(U^*,D)|}=$$
$$\frac{|U^*|}{|U|}\frac{|\mathrm{POS}_{A\cup\{a\}}^C(U,D)|-|\mathrm{POS}_A^C(U,D)|}{|\mathrm{POS}_{A\cup\{a\}}^C(U,D)|-|\mathrm{POS}_A^C(U,D)|}=\frac{|U^*|}{|U|}$$

因为 $|U^*|/|U|\geqslant 0$, 若 $\mathrm{Sig}^{\mathrm{outer}}(a,A,C,D,U)\geqslant\mathrm{Sig}^{\mathrm{outer}}(b,A,C,D,U)$, 则 $\mathrm{Sig}^{\mathrm{outer}}(a,A,C',D,U^*)\geqslant\mathrm{Sig}^{\mathrm{outer}}(b,A,C',D,U^*)$。证毕。

定理 2 表明, 计算属性重要度时由于删除正域和冗余属性集合后属性重要度保持不变, 这样在计算属性约简时, 每次迭代过程中可以删除候选集合产生的正域和冗余属性集合, 使得算法效率提升。

算法 2　QPRRA

输入　集值决策系统 $\mathrm{SDS}=(U,C\cup D,V,f)$;

输出　集值决策系统 SDS 的一个属性约简 R。

1) 令 $i=1$, $R_1=R$, $U_1=U$, $C_1=C$, $C_d=\varnothing$, $R=\varnothing$。

2) 对条件属性 C_i-R 进行如下操作: 计算 $\mathrm{Sig}^{\mathrm{outer}}(a_0,R,C_i,D,U_i)=\max\{\mathrm{Sig}^{\mathrm{outer}}(a_k,R,C_i,D,U_i)\}$, $R\leftarrow R\cup\{a_0\}$, 其中 $a_k\in C_i-R$。若有多个最大值, 则选择与 R 组合数最少的属性。

3) 计算 $U_{i+1}=U_i-\mathrm{POS}_R^{C_i}(U_i,D)$。

4) 计算 $C_d^{U_i}=\{a\in C_i-R:U/\boldsymbol{T}_R=U/\boldsymbol{T}_{R\cup\{a\}}\}$, $C_{i+1}=C_i-C_d^{U_i}$, 令 $i=i+1$。

5) 若 $\gamma_R^{U_i}(D)=\gamma_{C_i}^{U_i}(D)$, 转到 6), 否则转到 2)。

6) 对 $C-R$ 中的每个条件属性 a_k, 进行如下操作:

①计算 $\gamma_{R-\{a_k\}}(D)$;

②若 $\gamma_{R-\{a_k\}}(D)=\gamma_C(D)$, 则 a_k 为冗余属性, $R=R-\{a_k\}$, 否则 a_k 为非冗余属性, R 保持不变。

7) 返回 R。

在算法 2 中, 候选属性集合的正域随着每轮迭代过程选择属性重要度最大的属性加入而增大, 因此在计算正域时, 可以通过叠加的方式计算, 无需重新直接计算新的 D 关于约简属性集合的正域, 证略。

算法 2 原理流程如图 1 所示。集值信息系统的快速正域约简算法 (quick positive region reduction algorithm, QPRRA)。

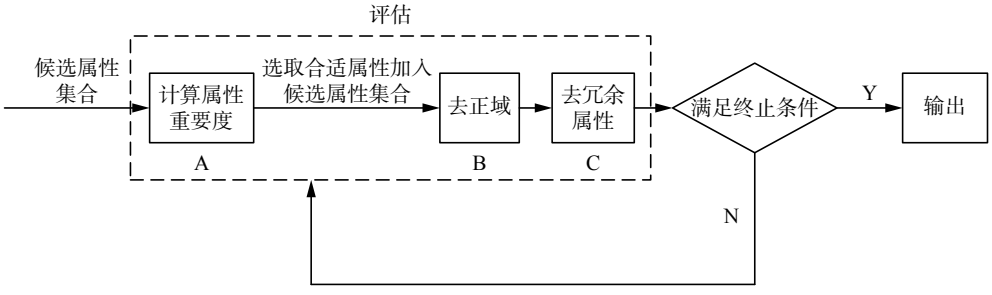


图 1　算法 2 原理流程

Fig. 1　Flow chart for algorithm 2

算法 2 无需首先求出核属性, 直接迭代选择属性重要度最大的属性加入候选属性集合, 因此对于无核的数据集算法效率提升。在约简过程中, 根据属性重要度的保序性, 即定理 5, 算法在每轮迭代过程中删除正域和冗余属性, 使得数据集规模缩小, 算法效率提升。在图 1 中, A 对应算法步骤 2), 计算当前轮次需评估的属性重要度, B 对应算法步骤 3), 删除当前轮次的正域, C 对应算法步骤 4), 删除当前轮次的冗余属性集合。

下面比较分析 PRRA 算法和 QPRRA 算法的

时间复杂度。令 T_1 和 T_2 分别为 PRRA 算法和 QPRRA 算法的时间复杂度。令 $|C|=m$ 表示数据集条件属性数, $|U|=n$ 表示数据集对象数, $|C_i|=m_i$ 表示第 i 轮需评估的属性数 (已删除第 $i-1$ 轮所求冗余属性数), $|U_i|=n_i$ 为第 i 轮剩余的对象数。在算法 2 中, 计算去掉正区域并且删除冗余属性集合将最大属性重要度的属性加入到当前约简的时间复杂度为 $O(\sum_{i=1}^{|C|}(m_i-i+1)^2n_i)$, 去冗余过程时间复杂度为 $O(m^2n)$。而在算法 1 中, 计算核属性的时间复杂度为 $O(m^2n)$, 计算最大属性重要度的属性加入到当前约简的时间复杂度为 $O(\sum_{i=1}^{|C|}(m-i+1)^2n)$。去冗余过程时间复杂度为 $O(m^2n)$。PRRA 算法时间复杂度为 $T_1=O(m^2n+\sum_{i=1}^{|C|}(m-i+1)^2n)$, QPRRA 算法时间复杂度为 $T_2=O(m^2n+\sum_{i=1}^{|C|}(m_i-i+1)^2n_i)$。QPRRA 算法效率优于 PRRA 算法。

例 3　表 1 为集值决策系统 $\mathrm{SDS}=(U,C\cup D,V,f)$, 其中, $U=\{x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8\}$ 为论域, $C=\{a_1,a_2,a_3,a_4,a_5\}$ 为条件属性集合, $D=\{d\}$ 为决策属性集合。

采用 QPRRA 算法对表 1 进行 3 轮约简。

第 1 轮迭代过程: 计算 C_1-R 中属性的重要度, 分别为 $\mathrm{Sig}^{\mathrm{outer}}(a_1,R,C_1,D,U_1)=0$, $\mathrm{Sig}^{\mathrm{outer}}(a_2,R,C_1,D,U_1)=0$, $\mathrm{Sig}^{\mathrm{outer}}(a_3,R,C_1,D,U_1)=0$, $\mathrm{Sig}^{\mathrm{outer}}(a_4,R,C_1,D,U_1)=0.375$, $\mathrm{Sig}^{\mathrm{outer}}(a_5,R,C_1,D,U_1)=0$。由于 $\mathrm{Sig}^{\mathrm{outer}}(a_4,R,C_1,D,U_1)=0.375$ 最大, 故 $R=R\cup\{a_4\}=\{a_4\}$, $U_2=U_1-\mathrm{POS}_R^{C_1}(U_1,D)=\{x_1,x_3,x_5,x_6,x_7\}$, $C_2=C_1-C_d^{U_1}=\{a_1,a_2,a_3,a_4,a_5\}$, 无冗余属性, 且 $\gamma_R^{U_1}(D)\neq\gamma_{C_1}^{U_1}(D)$, 继续循环。

第 2 轮迭代过程: 计算 C_2-R 中属性的重要度, 分别为 $\mathrm{Sig}^{\mathrm{outer}}(a_1,R,C_1,D,U_2)=0$, $\mathrm{Sig}^{\mathrm{outer}}(a_2,R,C_1,D,U_2)=0$, $\mathrm{Sig}^{\mathrm{outer}}(a_3,R,C_1,D,U_2)=0.4$, $\mathrm{Sig}^{\mathrm{outer}}(a_5,R,C_1,D,U_2)=0$。由于 $\mathrm{Sig}^{\mathrm{outer}}(a_3,R,C_1,D,U_2)=0.4$ 最大, 故 $R=R\cup\{a_3\}=\{a_3,a_4\}$, $U_3=U_2-\mathrm{POS}_R^{C_2}(U_2,D)=\{x_1,x_3,x_5\}$, $C_3=C_2-C_d^{U_2}=\{a_1,a_2,a_3,a_5\}$, 无冗余属性。且 $\gamma_R^{U_2}(D)\neq\gamma_{C_2}^{U_2}(D)$, 继续循环。

第 3 轮迭代过程: 计算 C_3-R 中属性的重要度, 分别为 $\mathrm{Sig}^{\mathrm{outer}}(a_1,R,C_1,D,U_3)=0.33$, $\mathrm{Sig}^{\mathrm{outer}}(a_2,R,C_1,D,U_3)=0$, $\mathrm{Sig}^{\mathrm{outer}}(a_5,R,C_1,D,U_3)=0.33$。由于 $\mathrm{Sig}^{\mathrm{outer}}(a_1,R,C_1,D,U_3)=\mathrm{Sig}^{\mathrm{outer}}(a_5,R,C_1,D,U_3)=0.33$, 故选择与 R 组合数最少的属性 a_1。故 $R=R\cup\{a_1\}=\{a_1,a_3,a_4\}$, $U_4=U_3-\mathrm{POS}_R^{C_3}(U_3,D)=\{x_3,x_5\}$, $C_4=C_3-C_d^{U_3}=\{a_1,a_5\}$, 删除冗余属性 a_2。且 $\gamma_R^{U_3}(D)=\gamma_{C_3}^{U_3}(D)$, 终止循环。

去除 R 中的冗余属性, 由于 $\gamma_{R-\{a_i\}}(D)\neq\gamma_C(D)$, $i=1,3,4$, 故 R 中无冗余属性, 算法终止。故算法 QPRRA 得到约简, 即 $R=\{a_1,a_3,a_4\}$。

3　实验分析

为了证明提出算法的有效性, 实验选取了 6 组 UCI 标准数据库 (http://archive.ics.uci.edu/ml/) 中的数据集, 为了得到集值数据集对这 6 组数据集进行预处理, 在条件属性集上随机对 10% 的数据进行对应属性上的并集操作, 如表 2 所示。

表 2　数据集描述

Table 2　Description of data sets

编号	数据集	对象数	特征数	类别数
1	Flag	194	29	8
2	Lung Cancer	32	57	2
3	Molecular Biology	106	58	2
4	Dermatology	366	35	6
5	SCADI	70	206	7
6	QSAR Biodegradation	1 055	42	9

所有实验在 PC 机上进行, 操作系统为 Microsoft Windows 7(64 b), 处理器及其型号为 Inter Core i5-2450M, 内存为 4 GB, 所有算法均使用 MATLAB7.11.0(R2010b) 编写实现。在实验中, 分别用上述 2 种约简算法对 6 组 UCI 数据集进行处理, 比较它们约简所耗费的时间。实验中, 将上述 6 组数据集分别分为 10 等份, 用来记录两个算法的时间差异。例如, 假设数据集有 1 000 个对象, 第 1 号数据集用来表示 1~100 个对象, 第 2 号数据集用来表示 1~200 个对象, 以此类推, 第 10 号数据集用来表示 1~1 000 个对象。实验一共分为 3 个部分。第 1 部分为算法 PRRA、算法 QPRRA、基于相似度的集值信息系统属性约简算法[18] (SRA) 和基于信息量的集值信息系统属性约简算法[19](IRA) 的约简结果长度和计算时间的比较。第 2 部分为算法 PRRA、QPRRA、SRA 和 IRA 随着论域的增大算法计算时间变化的实验。第 3 部分为数据集 Lung Cancer 迭代过程中对象和属性的变化。

表 2 给出了 6 组数据集的对象数、特征数和类别数的对比, 从表中可以看出, 本文选取了不同规模的数据集, 最大规模数据集为 QSAR Biodegradation, 对象数有 1 044 个, 最小规模数据集为 Lung Cancer, 对象数有 32 个, 最大特征数数据

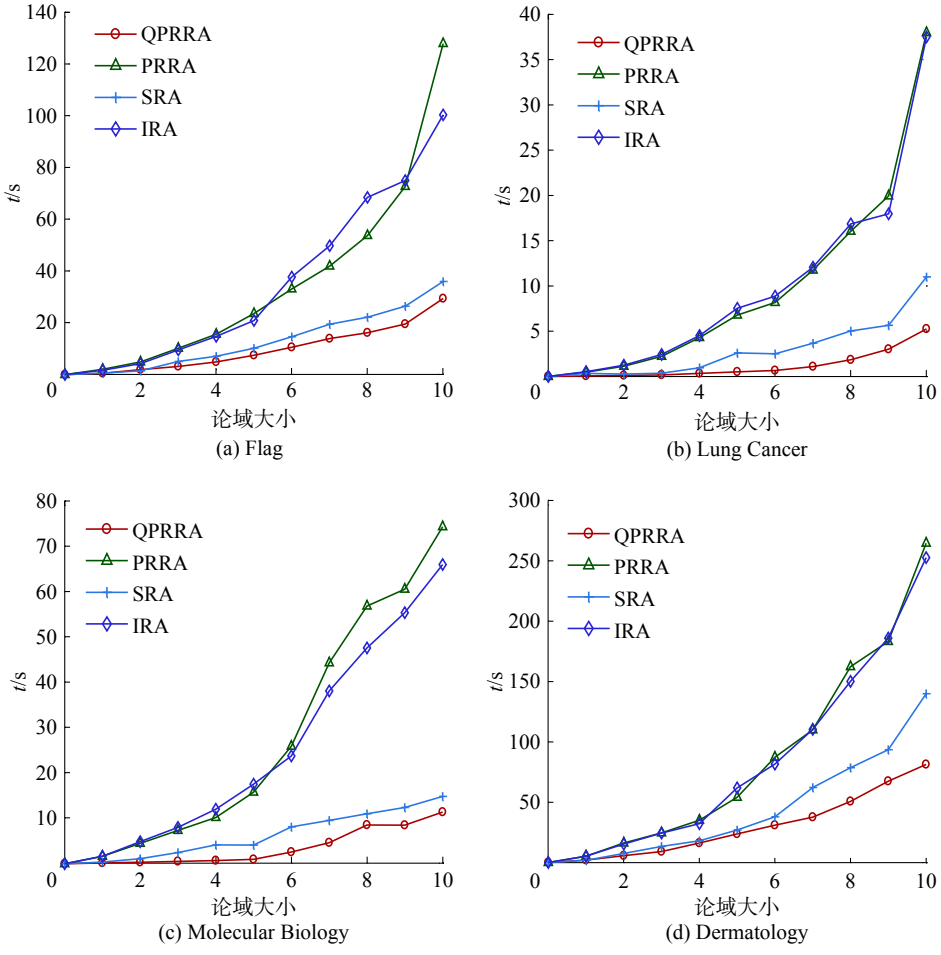
集 SCADI, 特征数有 206 个, 而 Flag 的特征数为 29 个, 在数据集中为特征数最少的。6 组数据集有不同的类别数, 最大类别数数据集 QSAR Biodegrada-tion, 类别数为 9, 数据集 Lung Cancer 和 Molecular Biology 类别数最小为 2。表 3 给出了算法 PRRA、QPRRA、IRA 和 SRA 的计算时间和选择特征数的对比, 可以看出, 算法 QPRRA 和 PRRA 约简结果长度优于算法 SRA 和 IRA, 算法 QPRRA 效率明显优于算法 PRRA 和 IRA。图 2 为 4 种算法的实验对比。

表 3　算法 PRRA 和 QPRRA 的计算时间和约简结果

Table 3　Time regurired and attribute reduction of the algorithms

数据集	原始特征	PRRA		IRA		SRA		QPRRA	
		约简	时间/s	约简	时间/s	约简	时间/s	约简	时间/s
Flag	29	11	130.7	13	102.5	14	36.8	11	30.1
Lung Cancer	57	8	37.9	12	37.5	12	10.9	8	5.2
Molecular Biology	58	6	74.6	7	66.2	7	14.8	6	11.4
Dermatology	35	13	263.4	17	251.4	17	139.2	12	80.8
SCADI	206	19	4 315.0	26	8 446.9	27	569.2	19	422.1
QSAR Biodegradation	42	15	7 731.3	22	6 730.4	24	1 259.6	15	1 266.8

从图 2 和表 3 可以看出, 算法 QPRRA 效率优于其他 3 个算法效率, 在数据集 QSAR Biodegradation 上, 算法 SRA 略优于算法 QPRRA。例如, 对于数据集 Lung Cancer, 算法 PRRA 时间耗费为 37.9 s, 算法 QPRRA 时间耗费为 5.2 s, 加速比达到了 7.2, 实验效果明显。因为对于核为空的数据集而言, QPRRA 算法无需计算核属性集合, 节省了算法求核的时间, 并且在每轮迭代过程中, 不仅删除了当前轮次候选属性集合产生的正域, 也删除了当前轮次产生的冗余属性集合, 使得数据集规模不断变小, 算法运行时间缩短。从表 4 可得, Lung Cancer 数据集核为空, 在每轮迭代过程中删除了正域和冗余属性, 使得数据集规模缩小, 算法效率提升。



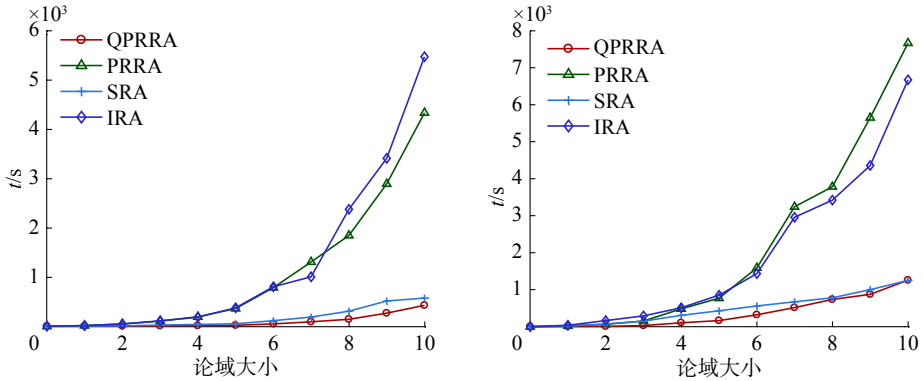


图 2　算法 PRRA 和算法 QPRRA 计算时间

Fig. 2　Time reguired for PRRA and QPRRA versus the data size

表 4　算法 PRRA 和 QPRRA 在 Lung Cancer 数据集上每轮迭代对象和属性的变化

Table 4　Changes of the number of objects and attributes within each loop of algorithms PRRA and QPRRA on Lung Cancer

迭代次数	PRRA		QPRRA	
	对象数	属性数	对象数	属性数
1	32	55	29	55
2	32	54	25	54
3	32	53	22	53
4	32	52	18	52
5	32	51	14	47
6	32	50	7	42
7	32	49	4	25
8	32	48	2	12

4　结束语

本文提出了一种在集值信息系统下的正域属性保持约简快速算法, 算法无需计算核属性集合, 直接开始迭代选择属性重要度最大的属性加入候选属性集合, 每轮迭代过程中删除一部分正域, 使得数据集对象数不断减少, 算法效率提升, 在删除一部分正域的同时, 删除冗余的属性集合, 使得算法的效率进一步提升。实验选取 6 组 UCI 数据集对提出算法的有效性进行验证, 实验表明: 提出算法的效率优于经典算法效率, 实现了对经典算法的优化, 尤其是在无核数据集上加速效果明显, 因为省去了计算核属性集的时间, 这使得该算法能更好地应用于较大规模数据的处理。本文是基于删除正域和冗余属性机制基础上进行的研究, 对每次迭代过程中产生正域和冗余属性较少的数据集加速效果不够明显, 故提出算法的普遍适用性是未来研究方向之一。

参考文献:

[1] PAWLAK Z. Rough sets[J]. International journal of computer and information sciences, 1982, 11(5): 341–356.

[2] 王国胤, 姚一豫, 于洪. 粗糙集理论与应用研究综述[J]. 计算机学报, 2009, 32(7): 1229–1246.
WANG Guoyin, YAO Yiyu, YU Hong. A survey on rough set theory and applications[J]. Chinese journal of computers, 2009, 32(7): 1229–1246.

[3] MIAO Duoqian, ZHAO Yan, YAO Yiyu, et al. Relative reducts in consistent and inconsistent decision tables of the Pawlak rough set model[J]. Information sciences, 2009, 179(24): 4140–4150.

[4] LI Hua, LI Deyu, ZHAI Yanhui, et al. A novel attribute reduction approach for multi-label data based on rough set theory[J]. Information sciences, 2016, 367/368: 827–847.

[5] YAO Yiyu, ZHAO Yan. Attribute reduction in decision-theoretic rough set models[J]. Information sciences, 2008, 178(17): 3356–3373.

[6] JIA Xiuyi, SHANG Lin, ZHOU Bin, et al. Generalized attribute reduct in rough set theory[J]. Knowledge-based systems, 2016, 91: 204–218.

[7] 张楠, 苗夺谦, 岳晓冬. 区间值信息系统的知识约简[J]. 计算机研究与发展, 2010, 47(8): 1362–1371.
ZHANG Nan, MIAO Duoqian, YUE Xiaodong. Approaches to knowledge reduction in interval-valued information systems[J]. Journal of computer research and development, 2010, 47(8): 1362–1371.

[8] HU Qinghua, ZHAO Hui, XIE Zongxia, et al. Consistency based attribute reduction[C]//Advances in Knowledge Discovery and Data Mining. Berlin, Heidelberg, Germany, 2007: 96–107.

[9] GUAN Yanyong, WANG Hongkai. Set-valued information systems[J]. Information sciences, 2006, 176(17): 2507–2525.

[10] QIAN Yuhua, DANG Chuanyin, LIANG Jiye, et al. Set-

valued ordered information systems[J]. Information sciences, 2009, 179(16): 2809–2832.

[11] 杨习贝, 张再跃, 张明. 集值信息系统中的模糊优势关系粗糙集[J]. 计算机科学, 2011, 38(2): 234–237.
YANG Xibei, ZHANG Zaiyue, ZHANG Ming. Fuzzy dominance-based rough set in set-valued information system[J]. Computer science, 2011, 38(2): 234–237.

[12] HUANG Yanyong, LI Tianrui, LUO Chuan, et al. Dynamic variable precision rough set approach for probabilistic set-valued information systems[J]. Knowledge-based systems, 2017, 122: 131–147.

[13] WEI Wei, CUI Junbiao, LIANG Jiye, et al. Fuzzy rough approximations for set-valued data[J]. Information sciences, 2016, 360: 181–201.

[14] ZHANG Hongying, YANG Shuyun. Feature selection and approximate reasoning of large-scale set-valued decision tables based on α-dominance-based quantitative rough sets[J]. Information sciences, 2017, 378: 328–347.

[15] SKOWRON A, RAUSZER C. The discernibility matrices and functions in information systems[C]//Intelligent Decision Support Theory and Decision Library. Dordrecht, Netherlands, 1992: 331–362.

[16] LUO Chuan, LI Tianrui, CHEN Hongmei, et al. Fast algorithms for computing rough approximations in set-valued decision systems while updating criteria values[J]. Information sciences, 2015, 299: 221–242.

[17] ZHANG Junbo, LI Tianrui, RUAN Da, et al. Rough sets based matrix approaches with dynamic attribute variation in set-valued information systems[J]. International journal of approximate reasoning, 2012, 53(4): 620–635.

[18] 刘莹莹, 吕跃进. 基于相似度的集值信息系统属性约简算法[J]. 南京大学学报 (自然科学版), 2015, 51(2): 384–389.
LIU Yingying, LYU Yuejin. Attribute reduction in set-valued information system based on similarity[J]. Journal of Nanjing university (natural sciences), 2015, 51(2): 384–389.

[19] 马建敏, 张文修. 基于信息量的集值信息系统的属性约简[J]. 模糊系统与数学, 2013, 27(2): 177–182.
MA Jianmin, ZHANG Wenxiu. Information quantity-based attribute reduction in set-valued information systems[J]. Fuzzy systems and mathematics, 2013, 27(2): 177–182.

[20] 苗夺谦, 李道国. 粗糙集理论、算法与应用[M]. 北京: 清华大学出版社, 2008.

[21] QIAN Yuhua, LIANG Jiye, PEDRYCZ W, et al. Positive approximation: an accelerator for attribute reduction in rough set theory[J]. Artificial intelligence, 2010, 174(9/10): 597–618.

[22] 钱宇华, 梁吉业, 王锋. 面向非完备决策表的正向近似特征选择加速算法[J]. 计算机学报, 2011, 34(3): 435–442.
QIAN Yuhua, LIANG Jiye, WANG Feng. A positive-approximation based accelerated algorithm to feature selection from incomplete decision tables[J]. Chinese journal of computers, 2011, 34(3): 435–442.

作者简介:

陈曼如, 女, 1993 年生, 硕士研究生, 主要研究方向为粗糙集、数据挖掘与机器学习。

张楠, 男, 1979 年生, 博士研究生, 主要研究方向为粗糙集、认知信息学与人工智能。

童向荣, 男, 1975 年生, 教授, 主要研究方向为多 Agent 系统、分布式人工智能与数据挖掘技术。发表学术论文 50 余篇, 被 SCI 检索 2 篇、EI 检索 20 余篇。