软件学报 ISSN 1000-9825, CODEN RUXUEW
*Journal of Software*,2014,25(Suppl.(2)):169−177

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

# 一般二元关系下的近似属性约简算法*

滕书华[1], 廖 帆[2], 鲁 敏[1], 赵 键[1], 张 军[1]

[1](国防科学技术大学 自动目标识别重点实验室,湖南 长沙 410073)

[2](66295 部队,河北 保定 072750)

通讯作者: 滕书华, E-mail: tengshuhua1979@sohu.com

**摘 要**: 属性约简是粗糙集理论重要应用之一.考虑到决策信息系统中的噪声,针对一般二元关系,从知识分类能力角度给出了一种新的属性重要性度量方法,在此基础上提出了一种能够抑制噪声的近似属性约简算法,该算法适用于多种粗糙集扩展模型,摆脱了现有约简算法对特定二元关系的依赖.实验结果表明,近似约简算法通过调节近似参数,可有效增强抗噪性,在有效降低约简属性集规模的同时,提高了约简结果的分类性能.

**关键词**: 噪声;粗糙集;属性约简;一般二元关系

中文引用格式: 滕书华,廖帆,鲁敏,赵键,张军.一般二元关系下的近似属性约简算法.软件学报,2014,25(Suppl.(2)):169−177. http://www.jos.org.cn/1000-9825/14035.htm

英文引用格式: Teng SH, Liao F, Lu M, Zhao J, Zhang J. Approximate attribute reduction algorithm based on general binary relation. Ruan Jian Xue Bao/Journal of Software, 2014,25(Suppl.(2)):169−177 (in Chinese). http://www.jos.org.cn/1000-9825/14035.htm

## Approximate Attribute Reduction Algorithm Based on General Binary Relation

TENG Shu-Hua[1], LIAO Fan[2], LU Min[1], ZHAO Jian[1], ZHANG Jun[1]

[1](Science and Technology on Automatic Target Recognition Laboratory, National University of Defense Technology, Changsha 410073, China)

[2](PLA Units: 66295, Baoding 072750, China)

Corresponding author: TENG Shu-Hua, E-mail: tengshuhua1979@sohu.com

**Abstract**: One of the most attentive applications of rough set is attribute reduction. Addressing the noise in decision information systems, a new method for importance measure of attribute set is presented from the point of view that knowledge can enhance the ability to perform classification. In addition, a new approximate attribute reduction algorithms is proposed based on general binary relation, which can be used to deal with noise and be applicable to many extending model of rough sets. Experimental results demonstrate that the proposed approximate attribute reduction algorithms can effectively increase sensitivity to noise, achieve more compact reduction, and simultaneously improve the classification performance.

**Key words**: noise; rough set; attribute reduction; general binary relation

知识约简是粗糙集理论的精髓[1,2].现有约简算法大都建立在等价关系基础之上,即以完备信息系统为研究对象.然而在现实中由于数据测量误差、对数据理解或获取条件限制等原因,信息的不完备(对象的属性值缺损)现象广泛存在,基于传统等价关系的粗糙集理论不能直接处理不完备信息系统,极大地限制了粗糙集理论向实用化方向的发展.探讨如何从不完备信息系统中获取有用知识具有重要的理论和现实意义[3].

处理不完备信息系统的传统方法:首先对不完备信息系统进行补齐处理,然后再用粗糙集来进行处理.由于补齐过程破坏了原始信息系统的信息,所得结果不能反映原始系统的真实情况,导致最终获得的知识可用性差.

* 基金项目: 国家自然科学基金(61471371); 湖南省自然科学基金(2015jj3022); 中国博士后科学基金(2012M512168)
收稿时间: 2014-05-07; 定稿时间: 2014-08-19

为了直接从不完备信息系统中获取知识,近年来,粗糙集理论在不完备信息系统中的扩展应用已成为粗糙集理论研究的重要方向[4,5].目前对不完备信息系统的缺失属性值存在两种语义解释,即遗漏型和丢失型.根据这两种解释,提出了多种直接对不完备信息系统进行处理的粗糙集扩展模型和方法,如基于遗漏型语义的容差关系模型与量化容差关系模型[6]、基于丢失型语义的非对称相似关系模型[7]、限制容差关系模型[8]、基于广义特征关系的粗糙集模型[9]和基于邻域粒化的粗糙集模型[10]等.此外,Yee 等人[11]讨论了相容关系下知识基本粒度的构造方法,为在不完备信息系统中度量所有类型的粗糙性提供了可能.

以上多种扩展模型对应的约简算法都是针对特定二元关系的,而对于一般二元关系下的约简算法研究还比较匮乏[12,13].特别是以上约简算法都不能很好地处理数据中的噪声,但在现实问题中,数据中的噪声是普遍存在的.一般可将噪声引起的误分类分为两种情况[14]:第 1 种是采集到的数据与正确值存在一定的误差,但误差较小,这种噪声引起的误分类可通过引入一个精度(允许一定的误分类率)加以解决,这类噪声出现的情况比较多.第 2 种是采集到的数据与正确值相差较远(缺省值包含在此),这类噪声是由偶然因素引起的,出现的情况相对较少,对于这类噪声引起的误分类,无论引入什么值都不能很好地处理.对于第 1 种情况,可利用变精度粗糙集模型来解决,但它不能很好地解决第 2 种情况.对于第 2 种情况,必然会产生误分类,最终影响约简结果.据此,文献[14,15]提出了一种能够有效处理第 2 种噪声的启发式约简算法,在一定程度上弥补了动态约简方法[16]和可变精度模型[17]等抗噪声方法的缺陷,但这两种方法仅适用于等价关系下的信息系统,不能处理决策信息系统和不完备信息系统.本文在现有文献基础上,以属性的分类能力为指标,在一般二元关系下提出了一种能够有效处理第 2 种噪声的近似属性约简算法,可适用于各种粗糙集扩展模型,摆脱了现有约简算法对特定二元关系的依赖.

## 1 粗糙集相关概念

(1) 知识 $P$ 的不可区分关系 IND($P$),$U$/IND($P$)构成了 $U$ 的一个划分,简记为 $U/P=\left\{\left[u_i\right]_P \middle| u_i\in U\right\}$.知识 $P$ 生成的等价类 $\left[u_i\right]_P$ 称为知识粒度.

(2) 考虑到现实中存在的对属性值排序的问题,Greco 等人[18]提出了基于优势关系的粗糙集研究方法,优势关系 $R_D^P$ 的定义为 $R_D^P=\left\{\left(u_i,u_j\right)\in U\times U \middle| \forall a\in P, f\left(u_i,a\right)\geqslant f\left(u_j,a\right)\right\}$.

(3) 文献[6]指出,空值“*”在信息系统中是确实存在的,只是被遗漏,据此给出容差关系 $R_T^P$ 的定义:

$$R_T^P=\left\{\left(u_i,u_j\right)\in U\times U \middle| \forall a\in P, f\left(u_i,a\right)=f\left(u_j,a\right)\vee f\left(u_i,a\right)=*\vee f\left(u_j,a\right)=*\right\}.$$

(4) 文献[7]在不完备信息系统中提出了一种非对称相似关系 $R_{NS}^P$:

$$R_{NS}^P=\left\{\left(u_i,u_j\right)\in U\times U \middle| \forall a\in P, f\left(u_i,a\right)=*\vee f\left(u_i,a\right)=f\left(u_j,a\right)\right\}.$$

(5) 鉴于相似关系过于严格,而容差关系又过于宽松,据此,王国胤教授提出了限制容差关系 $R_L^P$ [8]:

$$R_L^P=\left\{\left(u_i,u_j\right)\in U\times U \middle| \forall a\in P, f\left(u_i,a\right)=f\left(u_j,a\right)=*\vee\left(B_P(u_i)\cap B_P(u_i)\right.\right.$$
$$\left.\left.\neq\varnothing\wedge f\left(u_i,a\right)\neq *\wedge f\left(u_j,a\right)\neq *\Rightarrow f\left(u_i,a\right)=f\left(u_j,a\right)\right)\right\}.$$

其中, $B_P(u_i)=\left\{a\in P \middle| f\left(u_i,a\right)\neq *\right\}$.

本文中,用 $R^P$ 表示论域 $U$ 上的一般二元关系,有以下定义:

(6) 定义集值函数表达式为 $R_S^P:U\to P(U)$,关系 $R^P$ 下 $u_i$ 的后继近邻 $R_S^P(u_i)=\{u_j\in U\mid (u_i,u_j)\in R^P\}$.关系 $R^P$ 与其对应的后继近邻 $R_S^P(u_i)$ 可以相互唯一确定,即 $u_iR^Pu_j\Leftrightarrow u_j\in R_S^P(u_i)$.知识 $P$ 对论域的分类表示为 $U/R^P=\{R_S^P(u_i)|u_i\in U\}$,式中 $R_S^P(u_i)$ 也可理解为在知识 $P$ 下与对象 $u_i$ 具有相同性质的对象集合,亦即 $R_S^P(u_i)$ 中的对象在知识 $P$ 下相对于 $u_i$ 是不可区分的,应属于同一类.一般二元关系对论域的分类 $U/R^P$ 中的元素不必是其划分或覆盖[12].本文中用 $R_S^P(u_i)$ 表示一般二元关系下的信息粒度.集合 $X$ 在一般二元关系下的上近似集和下近似集定义为 $\overline{R^P}(X)=\{u_i\in U\mid R_S^P(u_i)\cap X\neq\varnothing\}$, $\underline{R^P}(X)=\{u_i\in U\mid R_S^P(u_i)\subseteq X\}$.

(7) $P \preceq Q$ 表示 $\forall u_i \in U, R_S^P(u_i) \subseteq R_S^Q(u_i)$,这意味着 $Q$ 比 $P$ 粗糙,或称知识 $Q$ 依赖于知识 $P$.若 $P \prec Q$, 则表示 $\forall u_i \in U, R_S^P(u_i) \subseteq R_S^Q(u_i)$, 且 $\exists u_j \in U$ 使得 $R_S^P(u_j) \subset R_S^Q(u_j)$, 即 $P$ 比 $Q$ 严格精细,或 $Q$ 完全依赖于知识 $P$.若 $P \approx Q$, 则表示对 $\forall u_i \in U$, 有 $R_S^P(u_i) = R_S^Q(u_i)$.

(8) 决策表 $S=(U,C,D)$ 中, $u_i, u_j \in U$. 如果 $u_i$ 和 $u_j$ 是 $C$ 不可区分而 $D$ 可区分的,则称 $u_i$ 和 $u_j$ 是不一致的;否则称 $u_i$ 和 $u_j$ 是一致的.如果对 $\forall u_i, u_j \in U$, $u_i$ 和 $u_j$ 是一致的,则称决策表为一致决策表,否则称为不一致决策表.

本文用 $R^P$ 表示由知识 $P$ 导出的一般二元关系.显然,IND($P$), $R_D^P$, $R_T^P$, $R_{NS}^P$ 和 $R_L^P$ 是 $R^P$ 的不同表现形式.

## 2 基于二元关系的近似属性约简算法

### 2.1 基于区分能力的属性重要度

在一般二元关系 $R^P$ 下, $R_S^P(u_i)$ 中的对象与 $u_i$ 都满足 $(u_i,u_j) \in R^P$, 它们具有相同的性质.因此在知识 $P$ 下 $\forall u_j \in R_S^P(u_i)$ 和 $u_i$ 应归属同一类,无需区分.若 $R_S^P(u_i)$ 中包含元素数目越多,则在知识 $P$ 下与 $u_i$ 属于同一类的对象就越多,即知识粒度越大,此时在对象 $u_i$ 上,知识 $P$ 表现出来的分类能力越弱.极端情况下,如果属性集不能区分论域中的任意两个对象,则此时属性的区分能力最弱,信息粒度最大.基于这种考虑,我们把区分对象多少的能力作为属性集区分能力的衡量标准,从区分能力角度给出属性重要性度量如下:

**定义 1**. 在信息系统 $S=(U,A)$ 中, $P \subseteq A$, $U=\{u_1,u_2,...,u_{|U|}\}$, 二元可区分关系定义为

$$\mathrm{DIS}(R^P)=\left\{\left(u_i,u_j\right) \in U \times U \middle| \exists a \in P, f\left(u_i,a\right) \neq f\left(u_j,a\right), f\left(u_i,a\right) \neq * \wedge f\left(u_j,a\right) \neq *\right\}.$$

相应地,用 $\left|\mathrm{DIS}(R^P)\right|$ 表示知识 $P$ 的可区分度: $\left|\mathrm{DIS}(R^P)\right| = |U|^2 - \sum_{i=1}^{|U|} \left|R_S^P(u_i)\right|$.

可区分度 $\left|\mathrm{DIS}(R^P)\right|$ 为可区分关系 $\mathrm{DIS}(R^P)$ 中有序对的个数,可理解为信息系统 $S$ 中知识 $P$ 所包含的信息量, $\left|\mathrm{DIS}(R^P)\right|$ 越大,则知识 $P$ 所包含的信息量越大,即知识 $P$ 的区分能力越强.因而 $\left|\mathrm{DIS}(R^P)\right|$ 度量了知识 $P$ 的可区分能力.

**定理 1(单调性)**. 在信息系统 $S=(U,A)$ 中, $Q,P \subseteq A$, 则(1) $P \preceq Q$ 当且仅当 $\mathrm{DIS}(R^Q) \subseteq \mathrm{DIS}(R^P)$;(2) 如果 $P \preceq Q$, 则 $\left|\mathrm{DIS}(R^Q)\right| \leqslant \left|\mathrm{DIS}(R^P)\right|$, 当且仅当 $\mathrm{DIS}(R^Q)=\mathrm{DIS}(R^P)$ 时等号成立.

证明:(1) ($\Rightarrow$)由偏序关系的定义可知, $P \preceq Q \Leftrightarrow \forall u_i \in U$, $R_S^P(u_i) \subseteq R_S^Q(u_i)$. 因而只需证: $\forall u_i \in U$, $R_S^P(u_i) \subseteq R_S^Q(u_i) \Rightarrow \mathrm{DIS}(R^Q) \subseteq \mathrm{DIS}(R^P)$, 假设 $\mathrm{DIS}(R^Q) \not\subseteq \mathrm{DIS}(R^P)$, 则 $\exists u_i, u_j \in U, i \neq j$, 满足 $(u_i,u_j) \in \mathrm{DIS}(R^Q)$ 但 $(u_i,u_j) \notin \mathrm{DIS}(R^P)$, 即 $u_j \notin R_S^Q(u_i)$ 但 $u_j \in R_S^P(u_i)$. 因而 $\exists u_i \in U$ 使 $R_S^P(u_i) \not\subseteq R_S^Q(u_i)$ 成立,与 $\forall u_i \in U, R_S^P(u_i) \subseteq R_S^Q(u_i)$ 相矛盾.因此 $P \preceq Q \Rightarrow \mathrm{DIS}(R^Q) \subseteq \mathrm{DIS}(R^P)$ 成立.

($\Leftarrow$) 因为 $\mathrm{DIS}(R^Q) \subseteq \mathrm{DIS}(R^P)$, 由定义 1 可得, $\forall u_j \notin R_S^Q(u_i) \Rightarrow u_j \notin R_S^P(u_i)$. 因而对 $\forall u_i \in U$, $R_S^P(u_i) \subseteq R_S^Q(u_i)$ 成立,所以 $P \preceq Q$;

(2)的证明由定义 1 和本定理(1)可得. □

$\left|\mathrm{DIS}(R^P)\right|$ 描述了单个属性集在信息系统中的区分能力,下面我们给出不同属性集间区分能力的度量.

**定义 2**. 在决策表 $S=(U,C,D)$ 中, $Q \subseteq C$. 知识 $Q$ 相对于知识 $D$ 的相对可区分关系 $\mathrm{DIS}(R^D/R^Q)$ 定义为

$$\mathrm{DIS}(R^D/R^Q)=\mathrm{DIS}(R^D)-\mathrm{DIS}(R^D) \cap \mathrm{DIS}(R^Q).$$

相应地,知识 $Q$ 相对于知识 $D$ 的相对可区分度 $\left|\mathrm{DIS}(R^D/R^Q)\right|$: $\left|\mathrm{DIS}(R^D/R^Q)\right| = \left|\mathrm{DIS}(R^D)-\mathrm{DIS}(R^D) \cap \mathrm{DIS}(R^Q)\right|$, 表达式为 $\left|\mathrm{DIS}(R^D/R^Q)\right| = \sum_{i=1}^{|U|} R_S^Q(u_i) - \sum_{j=1}^{|U|} R_S^{D \cup Q}(u_j)$.

由定义 2 可知,相对可区分关系 $\mathrm{DIS}(R^D/R^Q)$ 表示知识 $D$ 能区分而知识 $Q$ 不能区分的有序对的集合.如果

$u_i,u_j \in U$, $(u_i,u_j) \in \mathrm{DIS}(R^D/R^Q)$, 则 $u_i$ 和 $u_j$ 是不一致的,即 $\mathrm{DIS}(R^D/R^Q)$ 是条件属性集合为 $Q$ 时决策表中不一致有序对的集合.相对可区分度 $\left|\mathrm{DIS}(R^D/R^Q)\right|$ 是 $\mathrm{DIS}(R^D/R^Q)$ 中有序对的个数,反映了条件属性集 $Q$ 和决策属性集 $D$ 之间的不一致程度.不一致的有序对越多,则决策表不一致程度越强.

**定理 2(单调性)**. 在决策表 $S=(U,C,D)$ 中, $Q \subseteq P \subseteq C$, 则 $\left|\mathrm{DIS}(R^D/R^P)\right| \leqslant \left|\mathrm{DIS}(R^D/R^Q)\right|$, 当且仅当 $\mathrm{DIS}(R^D) \cap \mathrm{DIS}(R^P) = \mathrm{DIS}(R^D) \cap \mathrm{DIS}(R^Q)$ 时等号成立.

定理 2 的证明由定理 1 和定义 2 可得.定理 2 表明,相对可区分度 $\left|\mathrm{DIS}(R^D/R^Q)\right|$ 随着条件属性 $Q$ 中元素个数的增加单调下降,这个性质对于构建基于前向添加搜索策略的约简算法非常重要.由定义 2 可知, $\left|\mathrm{DIS}(R^D/R^Q)\right|$ 反映了条件属性集 $Q$ 和决策属性集 $D$ 之间的一致程度.在决策属性 $D$ 可区分的有序对个数不变的情况下,属性集 $Q$ 区分的有序对越多(即属性集 $Q$ 分类越细),决策属性 $D$ 能区分而条件属性集 $Q$ 不能区分的有序对越少,即不一致有序对的个数越少.据此,我们给出属性重要性度量的定义.

**定义 3**. 在决策表 $S=(U,C,D)$ 中, $Q \subseteq C$. 对 $\forall c_i \subseteq (C-Q)$, 一般二元关系下,属性 $c_i$ 相对于属性集合 $Q$ 的重要性测度 $SIG(c_i,Q,D)$ 定义为 $SIG(c_i,Q,D)=\left|\mathrm{DIS}(R^D/R^Q)\right|-\left|\mathrm{DIS}(R^D/(R^{Q\cup\{c_i\}})\right|$.

定义 3 中, $SIG(c_i,Q,D)$ 描述了向属性集 $Q$ 中添加属性 $c_i$ 后不一致有序对的减少量. $SIG(c_i,Q,D)$ 越大,说明在 $Q$ 已知的条件下 $c_i$ 对 $D$ 就越重要.因此在前向添加约简过程中可选择使 $SIG(c_i,Q,D)$ 最大的条件属性作为约简元素.这样就可保证每一步得到的属性集是含属性较少且相对可区分度较小的集合.

### 2.2 基于一般二元关系的近似属性约简算法

下面给出基于一般二元关系的属性约简的定义.

**定义 4**. 在决策表 $S=(U,C,D)$ 中, $Q \subseteq C$. $Q$ 是 $C$ 相对于 $D$ 的一个约简当且仅当:

(1) $\left|\mathrm{DIS}(R^D/R^Q)\right| = \left|\mathrm{DIS}(R^D/R^C)\right|$;

(2) $\forall q_i \in Q, \left|\mathrm{DIS}(R^D/R^{Q-\{q_i\}})\right| > \left|\mathrm{DIS}(R^D/R^C)\right|$.

定义 4 给出了一般二元关系下基于区分能力的属性约简的定义.其中,第 1 个条件保证了原决策表中不一致有序对的个数不变,即约简后的决策表与原决策表信息量相同;第 2 个条件保证所得的约简是最小的,即不再包含冗余属性.由定义 4 可知,基于区分能力的约简算法的目标就是寻找和原决策表具有相同不一致有序对数目的最小条件属性的集合.考虑到数据中的噪声,寻找和原决策表具有近似相同的不一致有序对数目的最小条件属性集作为约简,对原决策表的区分能力将不会有很大影响.进而类似于文献[14,15]中基于等价关系的信息系统下抗噪属性约简定义,下面在定义 4 的基础上给出一般二元关系下决策信息系统中具有抗噪性能的近似属性约简的定义.

**定义 5**. 在决策表 $S=(U,C,D)$ 中, $\beta$为条件属性相对决策属性的误识率, $0 \leqslant \beta < 0.5$. $Q$ 是 $C$ 的一个近似约简当且仅当:

(1) $\dfrac{\left|\mathrm{DIS}(R^D/R^Q)\right| - \left|\mathrm{DIS}(R^D/R^C)\right|}{\left|\mathrm{DIS}(R^D)\right|} \leqslant \beta$;

(2) $\forall q_i \in Q, \dfrac{\left|\mathrm{DIS}(R^D/R^{Q-\{q_i\}})\right| - \left|\mathrm{DIS}(R^D/R^C)\right|}{\left|\mathrm{DIS}(R^D)\right|} > \beta$.

由定义 5 可知,如果$\beta$=0,则定义 5 即为定义 4.当$\beta$值近似等于 0 时,如 $\beta = 1\times10^{-3}$, 此时 $Q$ 相比条件属性 $C$ 最多增加 $\left|\mathrm{DIS}(R^D)\right|\times10^{-3}$ 个不一致有序对,这可能是由误差引起的,也可能是由数据库中的某个不重要的属性来区分,将这个属性从约简中去掉,将不会对数据库的一致性带来多大影响;随着$\beta$值的增加,约简后的决策表的不一致性逐渐增加.因此,近似约简对数据不一致性有一定的容忍度,它解决了由于少量数据的干扰所引起的不一致性,使得到的属性约简具有一定的容错性,可以更好地抵抗噪声,增强了选择属性的鲁棒性.

下面给出基于一般二元关系的近似属性约简算法步骤.

**算法 1**. 基于一般二元关系的近似属性约简算法(approximate attribute reduction algorithm based on general binary relation,简称 AARA-GBR).

输入: 决策表 $S=(U,C,D)$;

输出: 近似约简 $Q$.

Step 1. 计算 $\left|\mathrm{DIS}(R^D)\right|$, 确定误识率$\beta$;

Step 2. 令 $j$=1, $A^j=C$, $Q=\varnothing$ ;//$Q$ 为已选择的属性的集合;

Step 3. 对 $\forall a_t\in A^j$, 计算 $SIG(a_t,Q,D)$; //如果 $Q=\varnothing$, 则 $\left|\mathrm{DIS}(R^D/R^Q)\right|=\left|\mathrm{DIS}(R^D)\right|$;

Step 4. 选择满足 $SIG(a_k,Q,D)=\max\{SIG(a_t,Q,D),a_t\in A^j\}$ 的属性 $a_k$ 加入到约简集合中,如果这样的属性不止一个,则选择可区分度最小的属性作为 $a_k$,令 $Q=Q\cup\{a_k\}$;

Step 5. 令 $j$=$j$+1, $A^j=C-Q$;

Step 6. 如果 $\dfrac{\left|\mathrm{DIS}(R^D/R^Q)\right|-\left|\mathrm{DIS}(R^D/R^C)\right|}{\left|\mathrm{DIS}(R^D)\right|}\leqslant\beta$, 则 $Q$ 即为所求的约简;否则转 Step 3.

## 3 近似属性约简算法的实例分析

经典粗糙集不能处理带有序的信息系统,然而带有序结构的决策问题在实际应用中非常广泛,如人力资源考核、投资风险分析、市场占有率等.针对序决策信息系统约简的研究还比较匮乏[19,20].文献[20]指出,在带有序的决策信息系统 $S=(U,C,D)$ 中, $u_i\in U$, 关于条件属性集 $C$ 比 $u_i$ 优的对象关于决策属性 $D$ 也应该比 $u_i$ 优,并称这一原理为优势原理.为了与经典粗糙集一致,对于 $\forall u_i,u_j\in U$, 如果对象 $u_j$ 关于条件属性集 $C$ 比 $u_i$ 优,关于决策属性 $D$ 也比 $u_i$ 优,则称 $(u_i,u_j)$ 为一致有序对;否则,如果对象 $u_j$ 关于条件属性集 $C$ 比 $u_i$ 优,对象 $u_i$ 关于条件属性集 $D$ 比 $u_j$ 优,则称 $(u_i,u_j)$ 为不一致有序对.由定义 5 可知,算法 AARA-GBR 的目标就是保持决策信息系统的一致有序对个数近似不变,而对于带有序的决策信息系统,则变为保持原序决策信息系统中与优势原理一致的有序对个数近似不变.下面利用算法 AARA-GBR 来处理带有序的决策信息系统.

例 1:表 1 给出了 11 篇论文的相关信息,每篇论文由 6 个条件属性(创造性、理论价值、应用价值、学术水平、文字和摘要)和一个决策属性描述.决策属性值为录用、修后录用和退稿.显然文章属性为偏好属性,对于 $a_1$, 有重大创新>有新见解>无创新;对于 $a_2$ 和 $a_3$,高>一般>低;对于 $a_4$,国际水平>国内先进>一般;对于 $a_5$,简练准确>可删减>表达不清;对于 $a_6$,好>能概括全文>概括性差;对于 $d$,录用>修后录用>退稿.下面我们用算法 AARA-GBR 来分析论文的条件属性和决策属性的关系.

令表 1 构造的决策系统 $S=(U,C,D)$ 中, $U=\{u_1,u_2,...,u_{11}\}$ 表示 11 篇论文,条件属性 $C=\{a_1,a_2,...,a_6\}$ ,决策属性 $D$=$\{d\}$.决策属性把论文分为 3 类: $Cl_0=\{u_6,u_7,u_8,u_{11}\}$ , $Cl_1=\{u_2,u_4,u_9,u_{10}\}$ , $Cl_2=\{u_1,u_3,u_5\}$, $Cl_0<Cl_1<Cl_2$ .算法 AARA-GBR 在优势关系下求表 1 的约简过程如下:

Step 1. 求得 $\left|\mathrm{DIS}(R^D)\right|=40$, 并令误识率$\beta$取最小值,即$\beta$=0;

Step 2. 令 $j$=1, $A^j=C$, $Q=\varnothing$;

Step 3. 计算每个属性的重要性: $SIG(a_1,Q,D)=34$, $SIG(a_2,Q,D)=24$, $SIG(a_3,Q,D)=33$, $SIG(a_4,Q,D)=28$, $SIG(a_5,Q,D)=33$, $SIG(a_6,Q,D)=33$;

Step 4. 求得 $a_k$=$a_1$,令 $Q=\{a_1\}$;

Step 5. 令 $A^2=C-Q=\{a_2,a_3,a_4,a_5,a_6\}$, $j$=2;

Step 6. 因为 $\dfrac{\left|\mathrm{DIS}(R^D/R^Q,U^2)\right|-\left|\mathrm{DIS}(R^D/R^C,U^2)\right|}{\left|\mathrm{DIS}(R^D)\right|}=0.15>\beta$ 转 Step 3′.

Step 3′. 计算 $A^2$ 中每个属性的重要性,可得结果如下: $SIG(a_2,Q,D)=4$, $SIG(a_3,Q,D)=6$, $SIG(a_4,Q,D)=6$, $SIG(a_5,Q,D)=6$, $SIG(a_6,Q,D)=6$;

Step 4′. 由于属性 $a_3$,$a_4$,$a_5$ 和 $a_6$ 的重要性同时取得了最大值,因而要比较 4 个属性的可区分度:

$\left|\mathrm{DIS}(R^{\{a_3\}})\right|=58$, $\left|\mathrm{DIS}(R^{\{a_4\}})\right|=50$, $\left|\mathrm{DIS}(R^{\{a_5\}})\right|=56$, $\left|\mathrm{DIS}(R^{\{a_6\}})\right|=52$, 因而求得 $a_k$=$a_4$.令 $Q=\{a_1,a_4\}$;

Step 5′. 令 $A^3=C-Q=\{a_2,a_3,a_5,a_6\}$, $j$=3;

Step 6′. 因为 $\dfrac{\left|\mathrm{DIS}(R^D/R^Q)\right|-\left|\mathrm{DIS}(R^D/R^C)\right|}{\left|\mathrm{DIS}(R^D)\right|}=0=\beta$, 算法停止,则 $Q=\{a_1,a_4\}$ 即为所求约简.

**Table 1** The evaluation decision table of articles

**表 1** 论文评价决策表

| $A$ / $U$ | 创造性($a_1$) | 理论价值($a_2$) | 应用价值($a_3$) | 学术水平($a_4$) | 文字($a_5$) | 摘要($a_6$) | 决策($d$) |
|---|---|---|---|---|---|---|---|
| $u_1$ | 有新见解 | 高 | 高 | 国际水平 | 简练准确 | 好 | 录用 |
| $u_2$ | 有新见解 | 高 | 一般 | 一般 | 可删减 | 能概括全文 | 修后录用 |
| $u_3$ | 有重大创新 | 高 | 高 | 国际水平 | 简练准确 | 好 | 录用 |
| $u_4$ | 有重大创新 | 一般 | 低 | 国内先进 | 可删减 | 能概括全文 | 修后录用 |
| $u_5$ | 有重大创新 | 一般 | 高 | 国际水平 | 简练准确 | 好 | 录用 |
| $u_6$ | 无创新 | 一般 | 一般 | 国内先进 | 表达不清 | 概括性差 | 退稿 |
| $u_7$ | 无创新 | 一般 | 低 | 一般 | 可删减 | 能概括全文 | 退稿 |
| $u_8$ | 无创新 | 低 | 低 | 国内先进 | 表达不清 | 概括性差 | 退稿 |
| $u_9$ | 有新见解 | 低 | 一般 | 一般 | 可删减 | 能概括全文 | 修后录用 |
| $u_{10}$ | 有新见解 | 低 | 低 | 国内先进 | 表达不清 | 概括性差 | 修后录用 |
| $u_{11}$ | 无创新 | 低 | 低 | 一般 | 表达不清 | 概括性差 | 退稿 |

由以上实例分析可知,算法 AARA-GBR 在优势关系下能找到相应的约简.此外,算法 AARA-GBR 也可用于邻域关系,见第 4 节.

## 4 近似参数$\beta$对属性约简和分类精度的影响

现有约简算法主要考虑了算法的复杂度和数据约简程度.但对于分类问题而言,更重要的是选择的属性不能显著降低分类能力.因而一种好的约简算法要在保持或者改善学习性能的基础上选择最少的属性.为进一步验证近似约简算法的有效性,下面将比较基于正区域的高效约简算法[21](算法 1)、基于条件熵的约简算法[22](算法 2)、最小冗余最大相关准则约简算法(算法 3)[23]、基于邻域依赖度函数方法[10](算法 4)以及本文提出的约简算法 AARA-GBR(算法 5)在选择特征数量和分类精度之间的差别.我们选用 UCI 机器学习数据库中的 5 个数据集进行实验测试,数据集见表 2.5 组数据中前 4 组分类问题的条件属性全部是数值型的,由于算法 1 和算法 2 只能处理符号数据,因此数值属性在约简前利用等频离散化方法对其进行离散化.算法 4 和算法 5 都是在邻域关系下直接求得属性的约简,并设置邻域参数$\delta$=0.15.在计算各样本邻域时,所有数值型属性均被标准化到[0,1]区间以减少各属性量纲不一致对约简结果的影响.为了比较选择特征的分类能力,我们利用 CART 分类学习算法以 10 折交叉验证的分类精度来评价 5 种算法选择特征的质量.表 3 给出了前 4 种非近似约简算法得到的特征数量、分类精度与原始数据的比较,加黑结果表示 4 种算法中此分类精度是最高的.表 4 给出了近似约简算法在不同$\beta$值下得到的特征数量和分类精度,加黑结果表示此精度是不同$\beta$值对应的最高分类精度.表中 $Q$ 表示约简中元素个数, $Acc$ 表示分类精度.

**Table 2** Experimental data

**表 2** 实验数据描述

| Data | Samples | Features | Classes |
|---|---|---|---|
| Wine | 178 | 13 | 3 |
| WDBC | 569 | 30 | 2 |
| Sonar | 208 | 60 | 2 |
| Iono | 351 | 34 | 2 |
| Zoo | 101 | 16 | 7 |

**Table 3** Reduction results and classification accuracy of Algorithm 1~Algorithm 4

**表 3** 算法 1~算法 4 的约简结果及其分类精度

| 数据集 | Wine | | WDBC | | Sonar | | Iono | | Zoo | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *Q* | *Acc* | *Q* | *Acc* | *Q* | *Acc* | *Q* | *Acc* | *Q* | *Acc* |
| 原始数据 | 13 | 0.898 6 | 30 | 0.905 0 | 60 | **0.720 7** | 34 | 0.875 5 | 16 | 0.890 4 |
| 算法 1 | 6 | 0.888 9 | 3 | 0.866 5 | 2 | 0.615 0 | 5 | 0.853 1 | 5 | 0.902 6 |
| 算法 2 | 6 | 0.888 9 | 3 | 0.898 1 | 2 | 0.543 1 | 3 | 0.841 5 | 5 | **0.933 9** |
| 算法 3 | 7 | 0.915 3 | 7 | **0.942 7** | 12 | 0.696 9 | 6 | **0.928 6** | 6 | 0.933 3 |
| 算法 4 | 5 | **0.925 7** | 12 | 0.940 2 | 7 | 0.696 7 | 8 | 0.915 0 | 5 | 0.922 6 |

**Table 4** Approximation reduction results and classification accuracy of algorithm 5

**表 4** 算法 5 的近似约简结果及其分类精度

| Wine | | | WDBC | | | Sonar | | | Iono | | | Zoo | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\beta\times10^3$ | *Q* | *Acc* | $\beta\times10^3$ | *Q* | *Acc* | $\beta\times10^3$ | *Q* | *Acc* | $\beta\times10^3$ | *Q* | *Acc* | $\beta\times10^3$ | *Q* | *Acc* |
| 0 | 6 | 0.920 1 | 0 | 22 | 0.922 8 | 0 | 7 | 0.755 0 | 0 | 17 | 0.886 7 | 0 | 9 | 0.901 5 |
| 0.02 | 5 | **0.925 5** | 0.9 | 15 | 0.922 3 | 0.2 | 6 | 0.759 8 | 0.3 | 15 | **0.914 0** | 0.04 | 7 | 0.921 3 |
| 2 | 4 | 0.903 5 | 2 | 8 | **0.936 8** | 0.4 | 5 | **0.763 8** | 1 | 6 | 0.850 2 | 0.06 | 6 | **0.935 9** |
| 3 | 3 | 0.908 3 | 4 | 6 | 0.926 2 | 2 | 4 | 0.683 1 | 2 | 5 | 0.853 0 | 1 | 5 | 0.932 4 |
| 10 | 2 | 0.895 8 | 5 | 5 | 0.926 2 | 7 | 3 | 0.660 0 | 3 | 4 | 0.855 3 | 3 | 3 | 0.921 8 |

从表 3 和表 4 可以看出:

(1) 5 种算法在 5 个数据集上都可有效地进行属性约简且保持一定的分类能力.虽然算法 1 和算法 2 整体上的约简属性个数较少,但由于离散化过程丢失了部分信息,使得分类精度也较低,甚至远低于原始数据分类精度,如对于数据 Sonar 和 Iono;算法 3、算法 4 和算法 5 则在有效进行属性约简的同时也保持了较高的分类精度;

(2) 对数据 Sonar,算法 5 在$\beta=0.4\times10^{-3}$时获得了紧凑的约简同时得到了最好的分类精度.随着近似参数$\beta$值的增加,近似约简算法约简结果的属性个数逐渐减小,但分类精度并不一定减小.如,对 WDBC 数据集,选择$\beta=2\times10^{-3}$时得到的属性约简相比$\beta=0$时的属性约简不仅属性个数由 22 减少到 8 个,且使得分类精度由 0.922 8 提高到 0.936 8(见表 4).这表明,近似约简算法对数据不一致性有一定的容忍度,能够克服少量数据的干扰引起的不一致性,使得到的属性约简具有一定的容错性,进而更好地抵抗噪声;

(3) 对于这 5 种约简算法,都不能保证所得到的约简对应的分类正确率是最高的.相比较而言,算法 5 在取得较优$\beta$值时得到的约简是比较紧凑的,对 5 个数据的分类精度与其他 4 种算法中最好的基本上相当.

## 5 结 论

针对现有启发式约简算法对特定二元关系的依赖性,并考虑到数据中的噪声,在一般二元关系下提出了一种能够抵制噪声的近似约简算法.通过实例分析表明,近似约简算法适用于多种二元关系,如容差关系、相似关系、优势关系、邻域关系等.通过对 UCI 数据的实验,讨论了近似参数对近似约简算法的约简结果和分类精度的影响.实验结果表明,通过近似参数的选择,近似约简算法可有效降低约简属性集的规模并改进分类器的性能.对于分类问题而言,属性选择过多或过少均会影响分类器性能,过多会影响分类器的推广性,过少会影响分类器的精度,如何在实际应用中更加有效地选择可调参数$\beta$,是我们未来要研究的内容之一.

**References**:

[1] Pawlak Z, Skowron A. Rudiments of rough sets. Information Sciences, 2007,177(1):3−27.

[2] Thangavel K, Pethalakshmi A. Dimensionality reduction based on rough set theory: A review. Appl. Soft. Comput., 2009,9(1): 1−12.

[3] Wang GY. Rough Set Theory and Knowledge Discovery. Xi'an: Xi'an Jiaotong University Press, 2001 (in Chinese).

[4] Wang CZ, He Q, Chen DG, Hu QH. A novel method for attribute reduction of covering decision systems. Information Sciences, 2014,254(1):181−196. [doi: 10.1016/j.ins.2013.08.057]

[5] Shu WH, Shen H. Updating attribute reduction in incomplete decision systems with the variation of attribute set. Int'l Journal of Approximate Reasoning, 2014,55(3):867−884.

[6] Kryszkiewicz M. Rough set approach to incomplete information systems. Information Sciences, 1998,112(1-4):39−49.

[7] Stefanowski J, Tsoukias A. On the extension of rough sets under incomplete information. In: Zhong N, Skowron A eds. New Directions in Rough Sets, Data Mining, and Granular-Soft Computing: the 7th Int'l Workshop. Berlin: Springer-Verlag, 1999. 73−82.

[8] Wang GY. Extension of rough set under incomplete information system. Journal of Computer Research and Development, 2002, 39(10):1238−1243 (in Chinese with English abstract).

[9] Grzymala-Busse JW. Data with missing attribute values: Generalization of indiscernibility relation and rule induction. Trans. on Rough Sets I, Berlin, 2004. 78−95.

[10] Hu QH, Yu DR, Liu JF, Wu CX. Neighborhood rough set based heterogeneous feature subset selection. Information Sciences, 2008,178(18):3577−3594.

[11] Leung Y, Li DY. Maximal consistent block technique for rule acquisition in incomplete information systems. Information Sciences, 2003,153:85−106.

[12] Wang CZ, Wu CX, Chen DG. A systematic study on attribute reduction with rough sets based on general binary relations. Information Sciences, 2008,178(9):2237−2261.

[13] Teng SH, Lu M, Yang AF, Zhang J, Zhuang ZW. A weighted uncertainty measure of rough sets based on general binary relation. Chinese Journal of Computers, 2014,37(3):1−17 (in Chinse with English abstract).

[14] Xu Y, Huai JP, Wang ZQ. Reduction algorithm based on discernibility and its applications. Chinese Journal of Computers, 2003,26(1):97−103 (in Chinse with English abstract).

[15] Teng SH, Wei RH, Sun JX, Tan ZG, Hu QH. Complete algorithm of quick heuristic attribute reduction based on indiscernibility degree. Computer Science, 2009,36(8):196−200 (in Chinse with English abstract).

[16] Wang F, Liang JY, Dang CY. Attribute reduction for dynamic data sets. Applied Soft Computing, 2013,13(1):676−689.

[17] Wang JY, Zhou J. Research of reduct features in the variable precision rough set model. Neurocomputing, 2009,72(10-12): 2643−2648.

[18] Greco S, Matarazzo B, Slowinski R. Rough approximation of a preference relation by dominance relations. European Journal of Operational Research, 1999,117(1):63−83.

[19] Susmaga R. Reducts and constructs in classic and dominance-based rough sets approach. Information Sciences, 2014,271:45−64.

[20] Greco S, Matarazzo B, Slowinski R. Rough approximation by dominance relations. Int'l Journal of Intelligent Systems, 2002,17(2): 153−171.

[21] Qian Y, Liang J, Pedrycz W, Dang C. Positive approximation: An accelerator for attribute reduction in rough set theory. Artificial Intelligence, 2010,174:597−618.

[22] Wang GY, Yu H, Yang DC. Decision table reduction based on conditional information entropy. Chinese Journal of Computers, 2002,25(7):759−766 (in Chinse with English abstract).

[23] Peng HC, Long FH, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2005,27(8):1226−1238.

**附中文参考文献**:

[3] 王国胤.粗糙集理论与知识获取.西安:西安交通大学出版社,2001.

[8] 王国胤.Rough 集理论在不完备信息系统中的扩充.计算机研究与发展,2002,39(10):1238−1243.

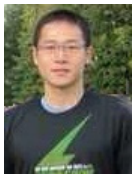
[13] 滕书华,鲁敏,杨阿峰,张军,庄钊文.基于一般二元关系的粗糙集加权不确定性度量.计算机学报,2014,37(3):1−17.

[14] 徐燕,怀进鹏,王兆其.基于区分能力大小的启发式约简算法及其应用.计算机学报,2003,26(1):97−103.

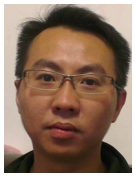
[15] 滕书华,魏荣华,孙即祥,谭志国.基于不可区分度的启发式快速完备约简算法.计算机科学,2009,36(8):196−200.

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

**滕书华**(1979－),男,河北迁安人,博士,工程师,主要研究领域为智能信息处理.
E-mail: tengshuhua1979@sohu.com



**廖帆**(1986－),男,硕士,主要研究领域为通信工程.
E-mail: liaoif12@gmail.com



**鲁敏**(1977－),男,博士,副教授,主要研究领域为激光雷达三维图像信息处理.
E-mail: lumin@nudt.edu.cn

**赵键**(1977－),男,博士,工程师,主要研究领域为激光雷达三维目标识别,图像处理.
E-mail: zjsprit@sina.com

**张军**(1973－),男,博士,研究员,主要研究领域为自动目标识别.
E-mail: zhj64068@sina.com