文章编号: 1000-5013(2013)01-0010-04

基于相对核的属性约简

王健[1,2], 徐余法[2], 陈国初[2]

(1. 华东理工大学 信息科学与工程学院, 上海 200237;
2. 上海电机学院 信息学院, 上海 200240)

摘要: 从相对核的角度,提出了一种新的属性约简方法. 首先,求出条件属性相对决策属性的相对正域,然后根据相对正域求得属性的相对核. 用这些相对核属性对论域进行划分,在对论域划分后,将可以完全正确的分类删除,减小论域,如此迭代下去,直到论域完全划分,最后求出这些核属性并集,去除并集的冗余信息,即可得到属性约简集. 该方法可直接利用核属性来对论域进行划分,不用再计算每个属性的重要度,减少了计算量,在每次迭代的过程中,减小论域,缩减搜索空间,降低了时间复杂度.

关键词: 粗糙集; 属性约简; 正域; 相对核

中图分类号: TP 18; TP 301.6 **文献标志码:** A

1982 年,波兰数学家 Pawlak 提出了粗糙集[1],这是一种定量分析处理不一致、不完整、不精确的信息与知识的数学工具和一种重要的信息处理技术[2-5]. 粗糙集的理论是建立在分类机制的基础上,将知识理解为对数据的划分. 粗糙集理论与其他处理不确定和不精确问题理论[6-7]的最显著的区别是,它无需提供问题所需处理的数据集合之外的任何先验信息,所以对问题的不确定性的描述或处理可以说是比较客观的[8-9]. 因此,被广泛应用到人工智能、模式识别和数据挖掘等方面. 粗糙集理论研究课题中一个重要的内容就是属性约简. 一般说来,知识库中的属性的重要度并不是相等的,有些属性可以删除,从而有利于做出正确而简洁的决策. 属性约简可以在保持分类和决策能力不变的情况下,去除那些不相关或不重要的属性. Wong 等[10]从计算复杂性的角度证明了寻找最小约简是 NP-hard 问题,通常采用启发式的搜索算法[11-17]. 本文从相对核的角度,提出了一种新的属性约简方法.

1 粗糙相关概念

粗糙集的关键点就是将知识和分类联系起来,用等价类关系的形式表示分类[18].

定义 1 决策表. 一个决策表可以形式化定义为 $S=\langle U,C\cup D,V,f\rangle$. 其中:有限集合 $U=\{u_1,u_2,\cdots,u_n\}$ 为论域;$C\cup D$ 是属性有限集,C 为条件属性集,D 为决策属性集,且 $C\cap D=\varnothing$;V 为属性集 $C\cup D$ 的值域;$f:U\times C\cup D\to V$ 为一个信息函数,表示任一个对象的属性在 V 上的取值,指定了 U 中每一对象 x 的属性值.

定义 2 不可分辨关系. 信息系统 $S=\langle U,C\cup D,V,f\rangle$,对于属性子集 $B\subseteq A$,定义一个不可分辨的二元关系. 即 $\mathrm{IND}(B)=\{(x,y)\in U\times U:\forall a\in B,f(x,a)=f(y,a)\}$. $\mathrm{IND}(B)$ 是一个等价关系. 由这种等价关系导出的对 U 的划分记为 $U/\mathrm{IND}(B)$,其中包含 x 的等价类记为 $[x]_{\mathrm{IND}(B)}$.

定义 3 近似集合. 粗糙集有两个基本概念,即上近似集和下近似集,给定 $X\subseteq U,R\subseteq A$.

1) 下近似集,即 $R_-X=\bigcup\{[x]_R\subseteq X\}$,$[x]_R=\{y\mid y\in U,\forall a\in R,f(x,a)=f(y,a)\}$. R_-X 是由 U 上在现有知识 R 的划分下肯定属于 X 的元素组成的集合.

收稿日期: 2012-06-15

通信作者: 徐余法(1964-),男,教授,主要从事智能算法和故障诊断的研究. E-mail:xyf690@21cn.com.

基金项目: 上海市教委重点科学基金资助项目(J51901, 09ZZ211);上海市自然科学基金资助项目(11ZR1413900);上海电机学院重点科学基金资助项目(09XKJ01)

2）上近似集，即 $R^{-}X=\bigcup\{[x]_R \cap X\neq\varnothing\}$. $R^{-}X$ 是在知识 R 划分下可能属于 X 的元素组成集合.

定义 4　正域，即 $\mathrm{POS}_B(X)=B_{-}X$，表示根据知识 B 对论域 U 进行划分，划分后能确定属于集合 X 的对象构成的集合. 设 P，Q 是 U 上两个等价关系，那么 Q 的 P-正域定义为 $\mathrm{POS}_P(Q)=\bigcup\limits_{x\in U/Q} P-X$. $\mathrm{POS}_P(Q)$是 U 中所有那些通过知识 P 被划分后肯定属于 U/Q 的元素组成的集合. 如果 $\mathrm{POS}_P(Q)=\mathrm{POS}_{P-\{r\}}(Q)$，那么称 $r\in P$ 是 Q 不必要；否则，称 $r\in P$ 是 Q 必要的.

定义 5　相对核. 设属性 $a\in C$，如果 $\mathrm{POS}_{C-a}(D)=\mathrm{POS}_C(D)$，那么称属性 a 相对决策属性 D 是不可缺少的，所有不可缺少的属性构成了相对决策属性 D 的相对核.

2　属性约简算法

2.1　算法描述

对于一个信息系统，$S=\langle U,C\cup D,V,f\rangle$，求出条件属性相对决策属性的相对核，并用这些相对核属性对论域 U 进行划分. 划分后，将可以正确辨识的部分去掉，并从条件属性中去除这些核属性，这样就形成新的决策信息系统. 然后，不断地迭代，直到论域 U 完全划分，即 $U=\varnothing$. 假如某一次迭代的过程中不存在相对核属性，则根据属性重要度，选择对划分重要度最大的属性，并用该属性对当前的论域进行划分. 当 $U=\varnothing$ 时，求取每步相对核或重要度最大的属性的并集，假设集合为 $P=\{a_1,a_2,\cdots,a_m\}$，计算 $\mathrm{POS}_P(D)$和 $\mathrm{POS}_{P-a_i}(D)$ $(i=1,2,\cdots,m)$，去除冗余信息，得到约简集.

非核属性的属性重要性的判断. 文中假如某一步不存在相对核属性，那么要从剩余的属性中选取属性重要度大的属性，来对当前的论域进行划分. 设属性 a 为条件属性，那么定义属性 a 的重要度为

$$\mathrm{Sig}_a=\frac{\mathrm{POS}_a(D)}{U}. \tag{1}$$

Sig_a 越大，则该属性的重要度越大.

2.2　算法部分

输入：决策信息系统 $S=\langle U,C\cup D,V,f\rangle$，$A=C\cap D$，$C$ 为条件属性，D 为决策属性. 输出：决策信息系统的约简. 初始化：集合 $P=\varnothing$. 算法有如下 6 个主要步骤：

步骤 1　求出决策信息系统的相对核 $\mathrm{CORE}_D(C)$，记 $Q=\mathrm{CORE}_D(C)$，若 $\mathrm{CORE}_D(C)=\varnothing$，则转步骤 4；

步骤 2　用相对核对论域 U 进行划分，若能将论域 U 完全划分则转步骤 5；否则，会有不可辨识的部分，记为 U_1，$U_1=U-\mathrm{POS}_Q(D)$，$P=P\cup CORE_D(C)$，$U=U_1$；

步骤 3　令 $C=C\text{-}CORE_D(C)$，然后转步骤 1；

步骤 4　利用式(1)计算剩余属性的重要度，并排序. 选取重要度最大的属性，假设为 a，$P=P\cup a$；然后用属性 a 对当前的论域进行划分，若能完全划分，则转步骤 5；否则，记不可分别的部分为 U_2，$U_2=U\text{-}\mathrm{POS}_a(D)$，$U=U_2$，$C=C-\{a\}$，转步骤 1；

步骤 5　对 P 中的每个元素，记为 a，计算 $\mathrm{POS}_P(D)$和 $\mathrm{POS}_{P-\{a\}}(D)$，若相等则从集合 P 中去除元素 a，消除属性集 P 中的冗余信息；

步骤 6　算法结束，输出约简后的属性集 P.

3　实例分析

表 1 为天气决策表[18]. 表 1 中：a_1，a_2，a_3，a_4 为条件属性，分别代表天气、温度、湿度、风；d 是决策属性；论域 $U=\{x_1,x_2,\cdots,x_{14}\}$. 此外，表 1 中括号中数字为对应的天气决策表的数字表示，如 $a_1=\{$晴，多云，雨$\}=\{1,2,3\}$；$a_2=\{$热，温暖，冷$\}=\{1,2,3\}$，$a_3=\{$高，正常$\}=\{0,1\}$；$a_4=\{$否，真$\}=\{0,1\}$，$d=\{N,P\}=\{0,1\}$.

对此决策表运用文中的方法.

$$U/C=\{\{1\},\{2\},\{3\},\{4\},\{5\},\{6\},\{7\},\{8\},\{9\},\{10\},\{11\},\{12\},\{13\},\{14\}\}.$$

$$\mathrm{POS}_C(D)=\{1,2,3,4,5,6,7,8,9,10,11,12,13,14\}.$$

表1 天气决策表

Tab.1 Weather decision table

样本	a_1	a_2	a_3	a_4	d
x_1	晴(1)	热(1)	高(0)	否(0)	$N(0)$
x_2	晴(1)	热(1)	高(0)	真(1)	$N(0)$
x_3	多云(2)	热(1)	高(0)	否(0)	$P(1)$
x_4	雨(3)	温暖(2)	高(0)	否(0)	$P(1)$
x_5	雨(3)	冷(3)	正常(1)	否(0)	$P(1)$
x_6	雨(3)	冷(3)	正常(1)	真(1)	$N(0)$
x_7	多云(2)	冷(3)	正常(1)	真(1)	$P(1)$
x_8	晴(1)	温暖(2)	高(0)	否(0)	$N(0)$
x_9	晴(1)	冷(3)	正常(1)	否(0)	$P(1)$
x_{10}	雨(3)	温暖(2)	正常(1)	否(0)	$P(1)$
x_{11}	晴(1)	温暖(2)	正常(1)	真(1)	$P(1)$
x_{12}	多云(2)	温暖(2)	高(0)	真(1)	$P(1)$
x_{13}	多云(2)	热(1)	正常(1)	否(0)	$P(1)$
x_{14}	雨(3)	温暖(2)	高(0)	真(1)	$N(0)$

考察 $a_i(i=1,2,3,4)$,在 C 中相对于 D 来说是否必要.为此,从 C 中去掉 a_1 可得:$\mathrm{POS}_{C-\{a_1\}}(D)\neq \mathrm{POS}_C(D)$,所以 a_1 必要.同理可得 a_2,a_3 不必要,a_4 必要,那么可得 $P_1=\mathrm{CORE}_D(C)=\{a_1,a_4\}$,$U/P=\{\{1,8,9\},\{2,11\},\{3,13\},\{3,5,10\},\{6,14\},\{7,12\}$,不可分辨的集合为$\{\{1,8,9\},\{2,11\}\}$.那么新的决策表为 $C=\{a_2,a_3\}$,$U=\{1,2,8,9,11\}$,如表2所示.

表2 新的决策表

Tab.2 New decision table

样本	a_2	a_3	d
x_1	1	0	0
x_2	1	0	0
x_8	2	0	0
x_9	3	1	1
x_{11}	2	1	1

考察 $a_i(i=2,3)$,在 C 中相对于 D 来说是否必要.同理可得 a_2 为不必要的,a_3 为必要的,那么可得 $P_2=\mathrm{CORE}_D(C)=\{a_3\}$,$U/P_2=\{\{1,2,8\},\{9,11\}\}$,分辨完全.即可得 $P=P_1\bigcup P_2=\{a_1,a_3,a_4\}$.则有

$$\begin{cases}\mathrm{POS}_P(D)=\{1,2,3,4,5,6,7,8,9,10,11,12,13,14\},\\ \mathrm{POS}_{P-\{a_1\}}=\{5,9,10,11\},\\ \mathrm{POS}_{P-\{a_3\}}=\{3,4,5,6,7,10,12,13,14\},\\ \mathrm{POS}_{P-\{a_4\}}=\{1,2,3,7,8,9,11,12,13\}.\end{cases}$$

由此可得 $\mathrm{POS}_P(D)\neq\mathrm{POS}_{P-\{a_1\}}$,$\mathrm{POS}_P(D)\neq\mathrm{POS}_{P-\{a_3\}}$,$\mathrm{POS}_P(D)\neq\mathrm{POS}_{P-\{a_4\}}$.所以,约简集合为$\{a_1,a_3,a_4\}$.采用UCI数据库进行测试,结果如表3所示.

表3 试验结果

Tab.3 Test results

数据集	实例数	属性数	最小约减属性数
Zoo	101	17	5
Vote	300	17	8
Vote-irvine	290	17	7

4 结束语

粗糙集理论目前已日趋完善,被广泛用在许多领域上.所提出的约简算法利用相对核对论域进行划分,从而求得属性约简,降低了时间和空间上的复杂度.经测试,该方法能够得到合理的结果,为数据决策表的属性约简提供一条较为可行有效的途径.

参考文献:

[1] 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,2009,32(7):1229-1246.

[2] PAWLAK Z. Rough set[J]. International Journal of Computer and Information Sciences,1982,11(5):341-356.

[3] CHAN C C,GRZYMALA-BUSSE J W,ZIARKO W P. Rough sets and current trends in computing[C]// Proceedings of the 6th International Conference on RSCTC. Akron:[s. n.],2008.

[4] AN A,STEFANOWSKI J,RAMANNA S,et al. Rough sets, fuzzy sets, data mining and granular computing[C]// Proceedings of the 11th International Conference on RSFDGrC. Toronto:[s. n.],2007.

[5] WANG G Y,LI T R,GRZYMALA-BUSSE J W,et al. Rough sets and knowledge technology[C]// Third International Conference on Rough Sets and Knowledge Technology. Chengdu:[s. n.],2008.

[6] 曾小军，黄宜坚. 利用 AR 模型和支持向量机的调速阀故障识别[J]. 华侨大学学报：自然科学版，2011，32(1)：13-17.

[7] 陈叶旺，于金山. 一种改进的朴素贝叶斯文本分类方法[J]. 华侨大学学报：自然科学版，2011，32(4)：401-404.

[8] PAWLAK Z,GRZYMALA-BUSSE J,SLOWINSKI R,et al. Rough sets[J]. Communications of the ACM,1995,38(11):89-95.

[9] PAWLAK Z. Why rough sets[C]// Proceedings of the Fifth IEEE International Conference on Fuzzy Systems. New Orleans: IEEE Press,1996:738-743.

[10] WONG S K M,ZIARKO W. On optional decision rules in decision tables[J]. Bulletin of Polish Academy of Science,1985,33(11/12):693-696.

[11] HU Xiao-hua,GERCONE N. Learning in relational databases: A rough set approach[J]. International Journal of Computational Intelligence,1995,11(2):323-338.

[12] 叶东毅. Jelonek 属性约简算法的一个改进[J]. 电子学报，2000，28(12)：81-82.

[13] 刘少辉，盛秋戬，史忠植. 一种新的快速计算正区域的方法[J]. 计算机研究与发展，2003，40(5)：637-642.

[14] 刘少辉，盛秋戬，吴斌，等 . Rough 集高效算法的研究[J]. 计算机学报，2003，26(5)：524-529.

[15] 胡峰，王国胤 . 二维表快速排序的复杂度分析[J]. 计算机学报，2007，30(6)：963-968.

[16] 徐章艳，刘作鹏，杨炳儒，等. 一个复杂度为 max($O(|C||U|)$,$O(|C|\sim 2|U/C|)$)的快速属性约简算法[J]. 计算机学报，2006，29(3)：391-399.

[17] 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.

[18] 瞿彬彬，卢炎生. 基于粗糙集的属性约简算法研究[J]. 华中科技大学学报：自然科学版，2005，33(8)：30-33.

Attribute Reduction Based on the Relative Core

WANG Jian[1,2], XU Yu-fa[2], CHEN Guo-chu[2]

(1. College of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China;
2. School of Information, Shanghai DianJi University, Shanghai 200240, China)

Abstract: From the point of view of relative core, this paper proposes a new attribute reduction method. Firstly, the condition attributes and the decision attributes are used to calculate the positive domain. Then, the relative core of the condition attributes is got based on the positive domain. Secondly, the samples are divided with these relative core attributes. At the end of this division, the samples that can be divided correctly is deleted. And then the samples are reduced. This iteration continues until the samples are completely divided. At last, the union of relative core is got and redundant information is removed, and then attribute reduction set is obtained. This method can use core attributes to divide the samples directly. No longer to calculate the important degree of each attribute, and then the amount of computation are reduced. In each iteration process, the samples, the search space and the time complexity are reduced.

Keywords: rough set; attribute reduction; positive domain; relative core

（**责任编辑**：陈志贤　　**英文审校**：杨建红）