第 38 卷 第 4 期
Vol.38 No.4

计 算 机 工 程
Computer Engineering

2012 年 2 月
February 2012

• 软件技术与数据库 •　　文章编号：1000—3428(2012)04—0058—02　　文献标识码：A　　中图分类号：TP182

基于论域压缩的启发式属性约简算法

卢喜森，吕跃进

(广西大学数学与信息科学学院，南宁 530004)

摘　要：论证在简化的信息系统上进行属性约简的可行性，指出某信息系统属性约简快速算法的计算结果可能含冗余属性，且在时间复杂度计算上存在错误。在此基础上，提出一种基于论域压缩的启发式属性约简算法，将相对支持度作为启发信息，缩小搜索空间，加入二次约简过程以消除冗余属性。实例分析表明，该算法具有较好的约简效果。

关键词：粗糙集；相对正域；属性约简；时间复杂度

Heuristic Algorithm of Attribute Reduction Based on Universe Compression

LU Xi-sen, LV Yue-jin

(College of Mathematics and Information Science, Guangxi University, Nanning 530004, China)

【Abstract】This paper gives a proof to the feasibility of the heuristic attribute reduction in simplified information system. It points out the results of one quick algorithm may contain redundancy attributes and time complexity calculation is wrong. This paper proposes a heuristic algorithm of attribute reduction based on universe compression. The relative support degree is used as heuristic information for reducing the searching room. A process of secondary reduction is used in order to eliminate redundancy attributes. A real example results demonstrate the improved algorithm has good reduction effect.

【Key words】rough set; relatively positive domain; attribute reduction; time complexity

DOI: 10.3969/j.issn.1000-3428.2012.04.019

1　概述

属性约简是粗糙集研究的核心内容之一。文献[1]从时间复杂度方面证明了寻找最小约简是 NP 问题。解决这类问题一般采用启发式算法，即在算法中加入启发信息，缩小搜索空间，最终得到一个最优解或近似最优解。在研究快速约简算法时，一般考虑如何减少时间复杂度，如文献[2]将属性的互信息作为启发信息，算法的时间复杂度为 $O(|C|^2|U|^2)$；文献[3]提供了时间复杂度为 $\max\{O|A||U|+O|A^2||U/A|\}$ 的快速算法。研究降低约简的空间复杂度方面的论文较少，文献[4]提出可在简化的信息系统上进行约简以节省存储空间，从而节省时间，但该文没有严格的可行性证明。文献[5]提供了算例，显示这些算法计算出的相对约简是不完备的，即最后的约简中仍含有冗余属性。

本文给出在简化的信息系统上进行属性约简的可行性证明，在文献[4]的基础上，提出一种改进的启发式属性约简算法。

2　相关定义与属性约简可行性证明

定义 1 信息系统可由四元组 $S=(U, A, V, f)$ 表示。其中，U 是对象集合，即论域；A 是属性集合；$V=\bigcup_{q\in A}V_q$，V_q 是属性集合的值域；f 是信息函数，即对 $\forall x\in U$，$q\in A$，有 $f(x,q)\in V_q$。决策信息系统是信息系统的子集，其属性集合 $A=C\cup D$，C 为条件属性集合，D 为决策属性集合。

定义 2 P 和 Q 为 U 中的等价关系，Q 的 P 正域记为 $pos_P(Q)=\cup_{X\in U/Q}\underline{P}X$。

Q 的 P 正域是 U 中所有根据分类 U/P 的信息可以准确划分到关系 Q 的等价类中去的对象集合。

定义 3 属性 a 相对 P 的 Q 支持度为：

$SIG(a)=pos_{P\cup\{a\}}(Q)-pos_P\{Q\}$

定义 4 在信息系统 $S=(U, A, V, f)$ 中，对每个属性子集 $P\subset A$，定义等价关系 $IND(P)$，称为不可区分关系：

$IND(P)=\{(x, y)\in U\times U|\ \forall\ a\in P, f(x, a)=f(y, a)\}$

关系 $IND(P)$ 构成 U 的一个划分，用 $U/IND(P)$ 表示，简记为 U/P。

引理 在信息系统 $S=(U, A, V, f)$ 中，$\forall\ P\subseteq Q\subseteq A$，则有 $U/P\subseteq U/P$。

定义 5 在信息系统 $S=(U, A, V, f)$ 中，$A=C\cup D$，$P\subset C$，如：(1)$pos_P(D)=pos_C(D)$；(2) $\forall\ a\in P$, $pos_{P-\{a\}}(D)\neq pos_C(D)$，则称 P 是 C 相对 D 的一个约简。

定义 6 P 和 Q 为等价关系族，$a\in P$，如果 $pos_P(Q)=pos_{P-\{a\}}(Q)$，则称 a 为 P 中 Q 冗余的属性。

定理 1 a 为 $P\cup\{a\}$ 中 Q 冗余的充分必要条件是 $SIG_P(a)=0$。

由冗余和相对支持度的定义可以证明定理 1。

定理 2 设 $U/A=\{X_1, X_2,\cdots, X_n\}$，则对 $\forall\ a\in A$，都有 $X_i/\{a\}=X_i$。

定理 3 $R\subseteq C$ 是 $S=(U,A,V,f)$ 的一个约简，充分必要条件是 R 是 $S'=(U/A,A,V,f)$ 的约简，其中，$A=C\cup D$。

基金项目：广西自然科学基金资助项目“基于粗糙集的不确定性决策理论与方法研究”(0991027)

作者简介：卢喜森(1974－)，男，讲师，主研方向：粗糙集理论，数据库技术；吕跃进，教授

收稿日期：2011-08-03　**E-mail：**luxisen2006@sina.com

证明：

必要性：由属性约简的特点，U 的约简必为它子集的约简。

充分性：设 R 是 $S'=(U/A,A,V,f)$ 的约简。由引理，因为 $C\subset A$，所以 $U/A\subseteq U/C$，即对于属性集 C，U/A 是不可区分的等价类，类似的对属性集 R。$U/A=\{X_1,X_2\cdots,X_n\}$，在信息系统 $S'=(U/A,A,V,f)$ 中，若 $X_i\in pos_C(D)$，则 $X_i\in pos_R(D)$。由于属性集 C，R 都不能区分 X_i，因此在信息系统 $S=(U,A,V,f)$ 中，$pos_R(D)=pos_C(D)$。若对 $\forall a\in R$，在 $S=(U,A,V,f)$ 是冗余的，则显然在 $S'=(U/A,A,V,f)$ 中也是冗余的。由以上得，R 是 $S=(U,A,V,f)$ 中的一个约简。

定理 4 在论域 U 和论域 $U-pos_P(Q)$ 中，属性 a 相对 P 的支持度不变。

证明：在论域 U 中，$SIG(a)=pos_{P\cup\{a\}}(Q)-pos_P(Q)$，在论域 $U-pos_P(Q)$ 中，$pos_P(Q)=0$，$pos_{P\cup\{a\}}(Q)=pos_{P\cup\{a\}}(Q)-pos_P(Q)$，显然相对支持度不变。

3 基于论域压缩的启发式属性约简算法

一般属性约简算法的时间复杂度都与 $|U|^2$ 呈正比。原信息系统中 $|U|$ 较大则算法慢，因此，有必要对原信息系统的计算进行简化。利用属性集 A，对 U 进行划分 U/A，得到简化的信息系统 $S'=(U',A,V,f)$，在简化的信息系统中，从空集 C_0 开始，求出每个属性的相对支持度，选出相对支持度最大的属性加入 C_0，将该属性对应的相对正域 R 从 U' 中剔除，若 $U'\neq\varnothing$，重复选取属性，否则对 C_0 进行二次约简，以剔除可能存在的冗余属性。

本文算法步骤如下：

输入 一决策表 $T=\{U,C\cup D,V,f\}$，其中，U 为论域；C、D 分别为条件属性和决策属性集。

输出 该决策表的一个相对约简。

Step1 计算出 $U'=U/A$。

Step2 初始化 $C_0=\varnothing,U_0=U'$。

Step3 对所有 $a\in C$，求出所有的 $SIG_{C_0}(a)$。

Step4 找出 $SIG_{C_0}(a)$ 最大的属性 a'，$C_0=C_0\cup a'$，$U'=U'/pos_a(D)$。

Step5 若 $U'\neq\varnothing$，转 Step3。

Step6 $U'=U_0$，计算 $U''=U'/C_0$。

Step7 对 C_0 中前 $|C_0|-2$ 项 $b_i,i=1,2,\cdots,|C_0|-2$，检查 $SIG_{C_0-b_i}(b_i)$，若其值为 0，则 $C_0=C_0/\{b_i\}$。

Step8 输出约简为 C_0。

4 算法时间复杂度分析

文献[4]对其算法的时间复杂度作出了分析，认为其时间复杂度为 $\max\{O(|A||U|),O(|A|^2|U/A|)\}$。经研究，笔者认为该分析有误。其错误为 Step1 计算 U/A 时时间复杂度不是 $O(|A||U|)$。如：当 $|A|=1$ 时，即简单分类时，其时间复杂度[6]为 $O(|U|^2)$。经笔者研究，计算 U/A 的时间复杂度为 $O((|A|+|U|)|U|)$，具体分析如下：设第 i 个属性增加 k_i 个分类数，前 i 个属性将论域总共划分为 $S\{i\}$ 个等价类，$|A|=n$。考虑第 n 个属性时一个元素只要在原等价类中作比较，最多只要比较 k_i+1 次，由于原有 $S(n-1)$ 个等价类，需比较的元素应该有 $|U|-S(n-1)$ 个，在最坏的情况下时间复杂度为：

$(k_n+1)(|U|-S(n-1))$

类似的第 $n-1$ 个属性的时间复杂度为：

$(k_{n-1}+1)(|U|-S(n-2))$

依此类推，第 1 个属性的时间复杂度为：

$(k_1+1)(|U|-S(0))$

计算 U/A 的时间复杂度为：

$$\sum_{i=1}^{n}(k_i+1)(|U|-S(i-1))=O(|A|+|U|)|U|$$

Step6 的时间复杂度为：

$O((|U'|+|C_0|)|U'|)$

Step7 的时间复杂度为：

$O(|C_0|^2|U''|)$

同时，用本文算法计算 C_0 的最后 2 个属性是非冗余的。

5 实例分析

为了说明算法的快速有效性，本节给出实例分析，并与其他算法进行比较。某呼吸系统的医疗诊断决策表[5]如表 1 所示，其中，条件属性集 $C=\{a_1,a_2,a_3,a_4,a_5,a_6\}$，决策属性集为 $D=\{d\}$。

表 1 某呼吸系统的医疗诊断决策表

患者	a_1	a_2	a_3	a_4	a_5	a_6	d
U_1	呼吸音轻	高热	肺纹理增多	白细胞升高	找到细菌生长	咳嗽，咳脓痰	肺炎
U_2	呼吸音轻	低热	肺纹理增多	白细胞升高	找到结核菌	咳嗽，咳带血	肺结核
U_3	呼吸音轻	正常	肺部感染	白细胞正常	未见细胞生长	咳嗽，咳白痰	肺炎
U_4	呼吸音轻	高热	肺结核	白细胞减少	未见细胞生长	咳嗽，咳白痰	肺结核
U_5	T 湿罗音	高热	肺纹理增多	白细胞升高	未见细胞生长	咳嗽，咳白痰	肺炎
U_6	呼吸音粗	高热	肺纹理增多	白细胞升高	未见细胞生长	咳嗽，咳白痰	上呼吸道感染
U_1	呼吸音轻	高热	肺纹理增多	白细胞升高	未见细胞生长	咳嗽，咳白痰	上呼吸道感染
U_1	呼吸音轻	正常	肺部感染	白细胞正常	未见细胞生长	咳嗽，咳白痰	肺炎
U_1	呼吸音轻	正常	肺部感染	白细胞正常	未见细胞生长	咳嗽，咳白痰	肺炎
U_1	T 湿罗音	高热	肺纹理增多	白细胞升高	未见细胞生长	咳嗽，咳白痰	肺炎
U_1	T 湿罗音	高热	肺纹理增多	白细胞升高	未见细胞生长	咳嗽，咳白痰	肺炎
U_1	呼吸音粗	高热	肺纹理增多	白细胞升高	未见细胞生长	咳嗽，咳白痰	上呼吸道感染

实例计算过程如下：

Step1 $U/C=\{\{U_1\},\{U_2\},\{U_3,U_8,U_9\},\{U_4\},\{U_5,U_{10},U_{11}\},\{U_6\},\{U_7,U_{12}\}\}$，得到 $U'=\{U_1,U_2,U_3,U_4,U_5,U_6,U_7\}$。

Step2 $C_0=\varnothing,U_0=U'$。

Step3 $SIG_{C_0}(a_1)=2$，$SIG_{C_0}(a_2)=2$，$SIG_{C_0}(a_3)=2$，$SIG_{C_0}(a_4)=2$，$SIG_{C_0}(a_5)=2$，$SIG_{C_0}(a_6)=2$，$C_0=C_0\cup\{a_1\}$。

(1) $R=pos_{C_0}(d)=\{U_5,U_6\}$，$U'=U'-R=\{U_1,U_2,U_3,U_4,U_7\}$，转 Step3。

$SIG_{C_0}(a_2)=2$，$SIG_{C_0}(a_3)=2$，$SIG_{C_0}(a_4)=1$，$SIG_{C_0}(a_5)=2$，$SIG_{C_0}(a_6)=1$，$C_0=C_0\cup\{a_2\}$。

(2) $R=pos_{C_0}(d)=\{U_5,U_6\}$，$U'=U'-R=\{U_1,U_4,U_7\}$，转 Step3。

$SIG_{C_0}(a_3)=1$，$SIG_{C_0}(a_4)=0$，$SIG_{C_0}(a_5)=1$，$SIG_{C_0}(a_6)=0$，

(下转第 62 页)