一种动态约简中子表族 F 范围的计算方法

陈 昊 杨俊安 吴彦华
(解放军电子工程学院 合肥 230037) (安徽省电子制约技术重点实验室 合肥 230037)

摘 要 对动态约简的思想进行了阐述,详细分析了其中子表族抽取的有关问题。根据正态分布的区间估计推算对所有动态约简都适应的 F 族抽样值,并在此基础上将约简精度系数纳入了 F 族的抽样考虑范畴,提出了一种计算 F 族范围的方法,发展并完善了对 F 族抽样计算的理论思想。

关键词 动态约简,F 族,抽样计算
中图法分类号 TP18 **文献标识码** A

Method of Calculating the Size of Dynamic Reduct Sub-table Family F

CHEN Hao YANG Jun-an WU Yan-hua
(Electronic Engineering Institute CPLA, Hefei 230037, China)
(Key Laboratory of Electronic Restriction of Anhui Province, Hefei 230037, China)

Abstract This paper formulated the dynamic reduct model and analyzed the withdrawing of sub-table family F. We calculated the sub-table family according to space estimation of normal distribution and proposed one method of calculating the size of dynamic reduct sub-table family F based on the reduct accuracy coefficient, which developed and completed the idea of sampling sub-table family F.

Keywords Rough set, Dynamic reduct, Sub-table family F, Sampling computation

1 引言

属性约简是知识发现的重要过程,是寻求最简规则的重要手段。文献[1-5]提出了在粗糙集理论研究中求解属性的最小约简,但都是基于静态约简的方法。静态约简算法在面对海量决策表和变化量决策表时表现出所得的约简不够稳定、无法描述决策表局部变化的规律、重复工作开销大等缺点。为此,文献[6]提出了动态约简概念和方法,建立了 F 族动态约简、F-λ 动态约简和广义约简思想体系,在理论上为决策信息系统最稳定的约简奠定了初步的基础。

动态约简的关键在于如何确定抽取出的 F 族的范围。关于如何抽取的问题,Bazan 等人采用统计学原理来估计子表随机抽样得到的数量[7]。他提出的两种动态约简方法(基于子表约简痕迹和基于子表概率抽取)都是基于先验知识来确定子表族 F 的范围。前者采用的方法是首先计算初始决策表的所有约简,然后随机去除决策表的某些样本,并计算新决策表的所有约简,直到新决策表的对象数大于等于 40%初始决策信息系统对象数后求取动态约简。后者采用的方法是对初始决策表进行一系列固定数量的抽取(如分别抽取决策表的 90%,80%,70%,60%,50%,每种 10 个子表,都是指定值),进而计算所有子表的约简,求动态约简。显然,这两种方法中,子表族 F 大小的确定都具有很大的主观性。设动态约简 R 在全局决策表的所有子表中出现的概率为 $P_G(R)$,文献[8,9]从 Bazan 的 $P_G(R)$估计出发,对其论述进行了改进,并根据正态分布的区间估计推算对所有动态约简都适应的 F 族抽样值,给出了子表族 F 范围的下界,但没有提到子表 F 族范围的上界要求。本文在此基础上,做了更进一步的 F 族理论分析和简化计算,特别是把约简精度纳入了 F 族抽样考虑范畴,求出了子表族 F 范围的上、下界,发展、完善了对 F 族抽样计算的理论思想。

2 动态约简的相关概念

本节介绍动态约简的一些基本概念,可参考文献[1-9]。

2.1 F 族动态约简

定义 1 决策信息系统 $S=(U,C\cup D)$,U 为论域,C 为条件属性集合,D 为决策属性集合,且 $B=(U',C\cup D)$,则称 B 为 S 的子决策信息系统。S 的所有子决策信息系统构成的集合为 $\rho(S)$,$F\subseteq\rho(S)$为 S 的一个子决策信息系统集合,称为决策信息系统 S 的 F 族。

定义 2 决策信息系统 $S=(U,C\cup D)$,U 为论域,C 为条件属性集合,D 为决策属性集合,S 的所有约简构成的集合称为 S 的约简集,记为 $RED(S)$。子决策信息系统 B 的所有约简构成的集合称为子决策信息系统 B 的约简集,记为 $RED(B)$。

一个决策信息系统至少含有一个约简,就是决策信息系统自身,称为平凡约简。然而,一个给定的决策信息系统可能

到稿日期:2008-09-16 返修日期:2008-11-26 本文受国家自然科学基金资助项目(60872113),安徽省自然科学基金(050420101)资助。
陈 昊(1982-),男,博士研究生,研究方向为粗糙集、智能计算等技术,E-mail:ritian99@21cn.com;杨俊安(1965-),男,教授,博士生导师,研究方向为粗糙集、智能计算、机器学习等技术;吴彦华(1973-),男,博士,副教授,研究方向为粗糙集、智能计算、机器学习等技术。

存在多个约简，不同的约简对应的规则也不完全相同，很难确定哪一个是最优的约简，特别是在数据不完备的情况下更是如此。因而，寻求最稳定的约简即是动态约简的目标，亦即寻求优化的约简。

定义 3 决策信息系统 $S=(U,C\cup D)$，U 为论域，C 为条件属性集合，D 为决策属性集合，集合 $F\subseteq\rho(S)$，S 的约简集为 $RED(S)$。动态约简 $DR(S,F)$ 表示如下：

$$DR(S,F)=RED(S)\cap\bigcap_{B\in F}(B)$$

$DR(S,F)$ 中任一元素称为 S 的 F 动态约简。

该定义意味着 S 的一个相对约简是一个 F 动态约简，当且仅当它是 F 中的所有子表的相对约简。F 族的动态约简将原决策信息系统中所有抽取子决策信息系统约简的交集作为最终的约简结果。

2.2 F-λ 族动态约简

F 动态约简概念对数据要求严格。为了适应噪声数据的处理，进一步把 F 动态约简的概念泛化，引入了 F-λ 动态约简概念。

定义 4 决策信息系统 $S=(U,C\cup D)$，U 为论域，C 为条件属性集合，D 为决策属性集合，子表族 $F\subseteq\rho(S)$，且 $\lambda\in(0.5\ 1]$，则 $(F$-$\lambda)$ 动态约简的定义为

$$DR_\lambda(S,F)=\left\{Q\in RES(S)\mid\frac{|\{B\in F:Q\in RED(B)\}|}{|F|}\geqslant\lambda\right\}$$

λ 是约简精度系数，λ 趋近于 1 时，$DR_\lambda(S,F)$ 接近 $DR(S,F)$。

2.3 广义动态约简

定义 5 决策信息系统 $S=(U,C\cup D)$，U 为论域，C 为条件属性集合，D 为决策属性集合，子表族 $F\subseteq\rho(S)$，且有

$$GDR(S,F)=\bigcap_{B\in F}RED(B)$$

则称 $GDR(S,F)$ 中元素为 S 的 F 广义动态约简。

该定义说明，若 S 的任一子表是一个广义动态约简，则它必须是一个给定子表族 F 中所有子表的约简。

定义 6 决策信息系统 $S=(U,C\cup D)$，U 为论域，C 为条件属性集合，D 为决策属性集合，子表族 $F\subseteq\rho(S)$，且 $\lambda\in(0.5\ 1]$，有

$$GDR_\lambda(S,F)=\left\{Q\in C\mid\frac{|\{B\in F:Q\in RED(B)\}|}{|F|}\geqslant\lambda\right\}$$

则称 $GDR_\lambda(S,F)$ 中的元素为 S 的 $(F$-$\lambda)$ 广义动态约简。

若约简 $Q\in RED(B)$，则 $\frac{|\{B\in F:Q\in RED(B)\}|}{|F|}$ 称为 $(F$-$\lambda)$ 广义动态约简 Q 相对 F 的稳定系数。

3 子表族 F 范围的计算

动态约简针对的是大型不相容信息表的整体分析困难问题，采用统计学抽样的思想，对信息表多次取样。从信息表随机抽取若干对象样本组成较小的信息表，把复杂的大型信息表的约简问题转化为若干子信息表的最优约简的交集问题，使得动态约简较静态约简方法更具有处理大型数据集的能力，其关键在于确定抽取出的子表族 F 的范围。

3.1 子表族 F 范围的下限[8,9]

已知动态约简 R 在全局决策表的所有子表中出现的概率为 $P_G(R)$。定义 $P_G(R)=|G_R|/|G|$，其中 $G_R=\{B\in G:R\in RED(B,D)\}$。假设每个子表是一个随机变量 X_i，所有子表为一系列的随机变量 $(X_1,X_2,\cdots,X_n)$，其独立且均服从 $(0,1)$ 分布，并具有数学期望和方差。使用独立同分布中心极限定理，有

$$\lim_{|F|\to\infty}P\left\{\frac{\sum_{i=1}^{|F|}X_i-E(\sum_{i=1}^{|F|}X_i)}{\sqrt{D(\sum_{i=1}^{|F|}X_i)}}\leqslant x\right\}=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt$$

其中 $E(\sum_{i=1}^{|F|}X_i)$ 表示总体的期望，$D(\sum_{i=1}^{|F|}X_i)$ 表示总体的方差。对于任何 $B\in G$，R 是否为子表 B 的一个约简构成 $(0,1)$ 分布，分布函数表示为 $X_G^R(B):G\to\{0,1\}$，有

$$X_G^R(B)=\begin{cases}1,R\in RED(B,D)\\0,R\notin RED(B,D)\end{cases}$$

记 $G^1=\{B\in G:X_G^R(B)=1\}$ 和 $G^0=\{B\in G:X_G^R(B)=0\}$。$(0,1)$ 分布中成功概率 $P[X_G^R(B)=1]=P_G(R)$，失败概率为 $P[X_G^R(B)=0]=1-P_G(R)$。设 $P_G(R)$ 的极大似然估计值为 $MLE(P_G(R))$。$(0,1)$ 分布概率 $P_G(R)$ 的极大似然估计是所有子表 $X_G^R(B)$ 的一个算术平均值，有

$$\begin{aligned}MLE(P_G(R))&=\frac{\sum_{B\in F}X_G^R(B)}{|F|}\\&=\frac{\sum_{B\in F\cap G^1}X_G^R(B)+\sum_{B\in F\cap G^0}X_G^R(B)}{|F|}\\&=\frac{|F\cap G^1|}{|F|}\end{aligned}$$

可见，R 的稳定系数值与 $P_G(R)$ 的极大似然估计值相等。对于 $(0,1)$ 分布而言，$E(\sum_{i=1}^{|F|}X_i)=n\cdot P_G(R)$，$D(\sum_{i=1}^{|F|}X_i)=n\cdot P_G(R)\cdot(1-P_G(R))$，进而用 $MLE(P_G(R))$ 替代式中部分 $P_G(R)$，得

$$P\left[-t_\alpha<\frac{MLE(P_G(R))-P_G(R)}{\sqrt{\frac{MLE(P_G(R))\cdot(1-MLE(P_G(R)))}{|F|}}}<t_\alpha\right]=1-\alpha$$

$1-\alpha$ 为置信参数。可得方程

$$P\left[\begin{aligned}&MLE(P_G(R))-t_\alpha\cdot\sqrt{\frac{MLE(P_G(R))\cdot(1-MLE(P_G(R)))}{|F|}}<P_G(R)\\&<MLE(P_G(R))+t_\alpha\cdot\sqrt{\frac{MLE(P_G(R))\cdot(1-MLE(P_G(R)))}{|F|}}\end{aligned}\right]=1-\alpha$$

假设 $P_G(R)$ 估计区间的最大误差值为 $\Delta MLE(P_G(R))$，得

$$|F|\geqslant\frac{t_\alpha^2\cdot P_G(R)\cdot(1-P_G(R))}{\Delta MLE(P_G(R))^2}\qquad(1)$$

所以在给定置信度 $1-\alpha$ 和允许误差 $\Delta MLE(P_G(R))$ 的值之后，可推得 $|F|$ 的下限值。

3.2 子表族 F 范围的上限

满足 $|F|$ 的下限值条件，F 族的元素越多，求得的约简的稳定性就越好。但是从动态约简的抽样分析来看，过多地抽取子表必然导致计算量增大。由贝努里大数定理的思想，当信息子表的抽取数量达到足够多时，继续抽取对约简稳定性产生的影响很小，F 族的元素数目对动态约简结果所产生的影响基本趋于饱和，因而 $|F|$ 必然存在一个上限值。

动态约简集的稳定性是指动态约简集与初始决策表具有相同的代表性。可以认为 F 族的稳定性即为 F 族与初始决

策表具有相似的决策能力。考虑到广义动态约简集，在判断子表与初始决策表的相似性程度时，可以将子决策表的相对正域与所有子决策表的并集组成的决策表相对正域进行比较，以此达到判断子决策表与初始决策表的相似性程度，进而判断 F 族的稳定性的目的。

定义 7 决策信息系统 $S=(U,C\cup D)$，U 为论域，C 为条件属性集合，D 为决策属性集合，子表族 $F\subseteq\rho(S)$，称 S 为全体样本决策表。其中 $U_{S'}=\bigcup\{U_B\mid B=(U_B\leqslant U,C\cup D)\in F\}$ 为全样本集合，$B\in F,U_B\in U,U_{S'}\in U$，则子表 B 与决策表 S' 的决策分类能力相似性指标 β_1,β_2,β_3 为

$$\beta_1=\frac{|POS(U_B,C,D)|}{|POS(U_{S'},C,D)|},\beta_2=\frac{|POS(U_B,C,D)|}{|POS(U_S,C,D)|},$$

$$\beta_3=\frac{|POS(U_{S'},C,D)|}{|POS(U_S,C,D)|}$$

当 $\frac{|POS(U_B,C,D)|}{|POS(U_B)|}\geqslant\frac{|POS(U_{S'},C,D)|}{|POS(U_{S'})|}$ 时，则可以判决决策表 B 和 S' 具有相同的代表性，即

$$\frac{|POS(U_B,C,D)|}{|POS(U_{S'},C,D)|}\geqslant\frac{|POS(U_B)|}{|POS(U_S)|}$$

同理，得到

$$\beta_1\geqslant\frac{|U_B|}{|U_{S'}|},\beta_2\geqslant\frac{|U_B|}{|U_S|},\beta_3\geqslant\frac{|U_{S'}|}{|U_{S'}|}$$

定义 8 决策信息系统 $S=(U,C\cup D)$，U 为论域，C 为条件属性集合，D 为决策属性集合，子表族 $F\subseteq\rho(S)$，称 S 为全体样本决策表。$S'=(U_{S'},C\cup D)$，其中 $B\in F,U_B,U_{S'}\subseteq U$，子表族 F 的相对正域稳定性参数 $SC_S^{POS}(F,B)$ 为

$$SC_S^{POS}(F,B)=\frac{\left|\left\{B\in F,\beta_1\geqslant\frac{|U_B|}{|U_{S'}|},\beta_2\geqslant\frac{|U_B|}{|U_S|},\beta_3\geqslant\frac{|U_{S'}|}{|U_S|}\right\}\right|}{|F|}$$

令 $SC_S^{POS}(F,B)\geqslant\lambda$（$\lambda$ 为阈值），则

$$|F|<\frac{\left|\left\{B\in F,\beta_1\geqslant\frac{|U_B|}{|U_{S'}|},\beta_2\geqslant\frac{|U_B|}{|U_S|},\beta_3\geqslant\frac{|U_{S'}|}{|U_S|}\right\}\right|}{\lambda}\quad(2)$$

性质 1 $0\leqslant SC_S^{POS}(F,B)\leqslant 1$

显然，当 $0\leqslant SC_S^{POS}(F,B)\leqslant 0.5$ 时，表示抽取的子表中大部分未达到相似性要求，得到的子表族也没有研究价值，应取 $0.5\leqslant SC_S^{POS}(F,B)\leqslant 1$。

性质 2 $(SC_S^{POS}(F,B))_{max}=1,(SC_S^{POS}(F,B))_{min}=0$

$(SC_S^{POS}(F,B))_{max}=1$ 表示抽取的所有子表均与决策表 S' 具有相同的决策分类能力；$(SC_S^{POS}(F,B))_{min}=0$ 表示抽取出的子表族每个子表均不与决策表 S' 具有相似的分类能力，即所有子表的 β 值均未达到要求。

所以，联立式(1)和(2)，可以得到子表族 F 范围为

$$\frac{t_\alpha^2\cdot P_G(R)\cdot(1-P_G(R))}{\Delta MLE(P_G(R))^2}\leqslant|F|<\frac{\left|\left\{B\in F,\beta_1\geqslant\frac{|U_B|}{|U_{S'}|},\beta_2\geqslant\frac{|U_B|}{|U_S|},\beta_3\geqslant\frac{|U_{S'}|}{|U_S|}\right\}\right|}{\lambda}\quad(3)$$

满足式(3)的这些值处于一个合理的取值区间，表明 F 族的抽样是均衡的、合理的，动态约简的有效性是充分的。

4 实验分析

采用文献[7]中的 UCI 数据库进行分析，在满足式(3)的情况下得到 F 子表族，进而得到动态约简集和动态规则集，将其与文献[5]属性约简的方法对比，如表 1 所列。选取 UCI 数据库中 6 个数据集分析，按照本文所用的算法得到的 F 子表族（其中 $SC_S^{POS}(F,B)$ 的阈值选取 0.8），所得到的属性数和规则数与按照文献[5]的算法对比，约简后的属性集和规则集相同，但是所抽取的子表数明显减小很多。

表 1 算法实验结果比较

数据库	对象数	属性数	抽取子表数		约简数		规则数	
			文献[5]	本文	文献[5]	本文	文献[5]	本文
Zoo	100	18	100	40	2103	1015	3156	2780
Yellow_ small	19	5	100	40	10	10	7232	5804
Wine	177	14	100	40	364	218	171	170
Servo	166	5	100	40	1	1	312284	28051
Glass	213	11	100	40	73	36	29	29
Flag	24	11	100	40	412	215	25625	24610

结束语 动态约简在某种意义上是给定决策表中最稳定的约简，它们是从给定决策表的随机抽样形成的子表中最常出现的约简。动态约简能够有效地增强约简的抗噪声能力。动态约简的计算较为简明，主要是对决策表进行采样，然后对采样后的决策表计算所有的约简，其中子表族的确定是动态约简结果有效与否的关键所在。本文对动态约简理论中子表族的确定问题进行了详细阐述，提出了一种计算 F 族范围的方法，发展并完善了对 F 族抽样计算的理论思想，为进一步的研究工作建立了理论基础。

参 考 文 献

[1] Pawlak Z. Rough Set: Theoretical Aspects of Reasoning about Data[M]. Dordrecht Kluwer Academic Publisher, 1991: 9-30

[2] Pawlak Z. Rough sets and Boolean reasoning[J]. Information Sciences, 2007(1): 41-73

[3] Jelonek J, Krawiec K, Slowinski R. Rough set reduction of attributes and their domains for networks[J]. Computational Intelligence, 1995, 11(2): 339-347

[4] Pawlak Z. A rough set view on Bayes' theorem[J]. International Journal of Intelligent Systems, 2003, 18(5): 487-498

[5] Zhang M, et al. A rough set approach to knowledge reduction based on inclusion and evidence reasoning theory[J]. Expert Systems, 2003, 20(5): 298-2004

[6] Bazan J, Skowron A, Synak P. Dynamic reducts as a tool for extracting system[C]//Proc. 8th International Symposium ISM IS' 94. LNAI vol 869. Charlotte, NG, Springer Verlag, October 1994: 346-355

[7] Basan J. A comparison of dynamic and non-dynamic rough set: method for extracting laws from decision tables[C]//Polkowski, Skowron, eds. Rough sets in Knowledge Discovery 1: Methodology and Applications. Physica-Verlag, Heidelberg, 1998: 321-355

[8] Wang Jia-yang, Liu Meng-chi. A Formal Model Integration[C]//Proceedings of the 3rd COPSLA Conference on Domain-specific Modeling. USA, Oct. 2003

[9] Slezak D. Attribute Set Decomposition of Decision and Soft Computing[M]. Aachen, Germany: Verlag Mainz, 2004: 236-240