基于分辨矩阵和约简树的增量式属性约简算法

侯 枫,刘丰年
HOU Feng,LIU Feng-nian

三门峡职业技术学院 信息工程系,河南 三门峡 472000
Information Engineering Department,Sanmenxia Polytechnic,Sanmenxia,Henan 472000,China
E-mail:liufengnian88@126.com

HOU Feng,LIU Feng-nian.Incremental algorithms for attribute reduction based on discernibility matrix and reduction tree.Computer Engineering and Applications,2010,46(11):125-127.

Abstract: For the efficient attribute reduction of dynamic decision table,the incremental algorithm for attribute reduction based on discernibility matrix and reduction tree is proposed.This method builds reduction tree according to sequential attribute reduction algorithm,calculates discernibility vector of new object,and revises reduction tree according to discernibility vector.Thereby attribute reduction cluster of new decision table can be abtained quickly,finally the validity of the algorithm is proved by examples.Compared with the traditional algorithm,this algorithm avoids complex logical calculus and improves the updating efficiency of attribute reduction.Theoretical analysis shows that the algorithm of this paper is efficient and feasible.

Key words: rough sets;discernibility matrix;incremental;reduction tree

摘 要:为了对动态变化的决策表进行高效属性约简处理,在改进的分辨矩阵的基础上提出一种基于约简树的增量式属性约简算法 IRART,该算法首先根据序贯属性约简算法对原决策表构造约简树,然后求出新增对象的分辨向量,并利用此向量对约简树进行修整,从而快速得到新决策表的所有约简,最后通过示例证明了这种算法的有效性。与传统增量式属性约简算法相比,该算法避免了复杂的逻辑演算,提高了属性约简的更新效率,理论分析表明该算法是有效可行的。

关键词:粗糙集;分辨矩阵;增量式;约简树
DOI:10.3778/j.issn.1002-8331.2010.11.038 **文章编号:**1002-8331(2010)11-0125-03 **文献标识码:**A **中图分类号:**TP311

1 引言

粗集(RS)理论[1]是由波兰逻辑学家 Z.Pawlak 于 1982 年首先提出,它能用确定的方法处理不确定知识,不需要先验知识,直接从数据中获取知识,近年来该理论在机器学习、数据挖掘及模式识别等多个领域得到了广泛的应用[2]。在基于粗集理论的知识获取研究中, 属性约简是其中最核心的组成部分之一,属性约简的结果会对最终形成的规则产生直接的影响,许多学者已对属性约简的算法进行了大量的研究[3-4],并取得了很大的进展。但这些研究几乎都是针对静态信息系统或决策表,不适合信息系统或决策表动态变化的情况。

现实世界是发展变化的,信息系统或决策表中的对象在不断动态变化,已得到的属性约简将可能不再有效,这就需要对属性约简进行动态修改。因此许多研究者建议,数据库知识发现算法应该是增量式的[5-6]。增量式的规则获取算法和增量式的属性约简算法已经开始得到研究。目前增量式属性约简算法大致可分为两大类: 一类是获取属性约简簇集的增量式算法;另一类是获取一个属性约简的增量式算法。这两种算法都使用了对象之间的分辨属性来处理, 属于代数观下的属性约简方法。文献[7]提出了一种增量式属性约简方法,虽然能够得到信息系统的最小约简,但只能求出绝对约简(不包含决策属性)。文献[8]给出了一种基于 Skowron 分辨矩阵的属性约简的增量模型,但不能保证得到一个 Pawlak 约简,也不能处理不相容决策表,运算复杂且效率低,

以分辨矩阵和约简树为基础提出一种增量式的属性约简算法。主要考虑对象动态增加情况下属性约简的更新问题。该算法首先结合分辨矩阵和序贯思想对决策表进行快速属性约简,并建立约简树,然后对新增对象建立分辨向量,利用分辨向量对约简树进行动态修整,从而有效地利用了原有属性约简进行属性约简的增量式更新,由于避免了复杂的逻辑演算,因而可有效改进属性约简的更新效率。示例验证和理论分析表明,提出的算法是有效可行的。

2 基本概念

定义 1 一个知识表达系统 S 为一个四元组 $S=(U,R,V,f)$,其中 $U=\{x_1,x_2,\cdots,x_n\}$ 称为对象的非空有限集合,也称为论域,R 称为属性的非空有限集合,$V=\bigcup_{r\in R}V_r$,V_r 是属性 r 的值域,$f:U\times R\to V$ 是一个信息函数,即 $x\in U,r\in R,f(x,r)\in V_r$ 且

作者简介:侯枫(1970-),女,讲师,三门峡职业技术学院信息工程系教师,研究方向为网络安全与数据挖掘;刘丰年(1982-),助教,研究生,三门峡职业技术学院信息工程系教师,研究方向为模式识别与智能系统。

收稿日期:2008-10-10 **修回日期:**2008-12-23

$f(U,R)=V$。若 S 中 $R=C\cup D$,其中子集 C 和 D 分别称为条件属性集和决策属性集,且 $C\cap D=\phi$,$D\neq\phi$,则称具有条件属性和决策属性的知识表达系统 S 称为决策表。通常多决策属性值决策表可化为单一决策属性值决策表,此时 $D=\{d\}$可简化为 d。

定义 2 设 $S=(U,R,V,f)$,$P\subseteq R$,定义不可分辨关系:

$IND(P)=\{(x,y)|(x,y)\in U^2,\forall r\in P,f(x,r)=f(y,r)\}$

显然不可分辨关系是一个等价关系,U 被等价关系 $IND(P)$划分成的所有等价类记作:$U/IND(P)=\{[x]_P|x\in U\}$,其中:$[x]_P=\{y|(x,y)\in IND(P)\}$,称为对象 x 在属性集 P 上的等价类。

定义 3 设 $X\subseteq U$ 为论域的一个子集,$P\subseteq C$,X 的关于 P 的下近似、上近似分别定义为:

$\underline{P}(X)=\{x|x\in U,[x]_P\subseteq X\}$

$\overline{P}(X)=\{x|x\in U,[x]_P\cap X\neq\phi\}$

定义 4 设 $P\subseteq C$,对于划分$\{Y_1,Y_2,\cdots,Y_k\}$的 P 的近似精度为:$\gamma_P=\sum_{i=1}^{k}card(\underline{P}(Y_i))/card(U)$,其中 $card(\ )$表示集合的基数。

定义 5 设 $P\subseteq C$,若 $\gamma_P=\gamma_C$,且不存在 $R\subset P$,使得 $\gamma_P=\gamma_P$,则称 P 为 C 的一个属性约简。所有 C 的属性约简的交称为 C 的核,记为 $Core(C)$。

定义 6 如果属性 $a\in C$ 满足 $\gamma_{C-\{a\}}(U)<\gamma_C(U)$,则属性 a 为不可缺少的,否则,称属性为冗余的。

在 Rough 集理论中,Pawlak 定义了两种信息系统的约简:绝对约简(不包含决策属性)和相对约简(包含决策属性)。文中所得到的约简都属于相对约简,或称为约简。

3 基于分辨矩阵的序贯属性约简算法(SARA)

Skowron[3]等人最早提出了利用决策表分辨矩阵来描述决策表的概念,Hu[9]等学者提出简洁的利用改进分辨矩阵来确定属性约简的方法,其中改进的分辨矩阵 $M_S=(m_{ij})_{n\times n}$ 定义为:

$$m_{ij}=\begin{cases}\{a\in C|f(x_i,a)\neq f(x_j,a),\\ f(x_i,D)\neq f(x_j,D),\\ x_i,x_j\in U,i<j,\\ i,j=1,2,\cdots,n。\ n=|U|\}\\ \phi,其他\end{cases}\quad(1)$$

未加证明地指出:当且仅当某个元素为单个属性时,该属性属于核 $Core(C)$。叶东毅教授于文献[10]指出,该结论在某些情况下是错误的,并给出实例加以说明。同时针对 Hu 方法的缺陷,提出新的分辨矩阵 $M'_S=\{m'_{ij}\}$,定义为:

$$m'_{ij}=\begin{cases}m'_{ij},\min(d(x_i),d(x_j))=1\\ \phi,其他\end{cases}$$

其中,$\{m_{ij}\}$的定义同式(1),对 $x_i\in U$,$d(x_i)=|\{f(y,D):y\in[x_i]_C\}|$。叶东毅证明了当且仅当某个 m'_{ij} 为单个属性时,该属性属于核 $Core(C)$。

根据上述叶东毅教授改进分辨矩阵的定义,实际上分辨矩阵包含了一个知识表达系统中为区分所有对象所需要的信息,核是分辨矩阵中所有单个元素组成的集合,即 $Core(C)=\{m'_{ij}|m'_{ij}=1\}$。分辨矩阵中元素所包含的属性个数就反应了这些属性的重要程度,个数越少属性越重要。因此可以按照序贯思想进行逐步约简,首先按照矩阵元素中属性数目构造递增序列的分明函数,为了避开大量的逻辑运算,采取逐次扩充核集的方法,对分辨函数进行分支运算,直到所有分明子函数中的项全为单属性项的分明函数为止,每个子分明函数就表示了一个约简。最终将此过程用约简树的形式进行描述。

根据上述分析,基于分辨矩阵的序贯属性约简算法 SARA 描述如下:

输入:决策表 $S=(U,C\cup D)$

输出:S 的约简树 T

第一步:利用改进分辨矩阵定义对决策表 S 构造分明矩阵 $\boldsymbol{M}'_S=(m'_{ij})_{n\times n}$,求出决策表 S 的核集 $Core(C)$,并将 $\boldsymbol{M}'_S$ 中包含核属性的元素值置为 ϕ,对于相同的元素只保留一个,得到新的分辨矩阵 $\boldsymbol{M}''_S$[2];

第二步:在 $\boldsymbol{M}''_S$ 中,将非空元按照$|m'_{ij}|$的值进行递增排序,设 $m^1\leqslant m^2\leqslant\cdots\leqslant m^k$,其中 k 表示 M''_S 中非空元素个数,并按照此顺序将 $m^i(i=1,2,\cdots,k)$用合取式连接,从而得到分辨函数 $f_S=m^1\wedge m^2\wedge\cdots\wedge m^k$;

第三步:取 f_S 中$|m^i|$的值最小的项中相同的属性和不同的属性组合作为一级核集,利用每个一级核分别来取代 f_S 中跟一级核对应的项并进行吸收运算,从而实现 f_S 的分支运算,得到 r 个分明子函数:$f_S^{(1)},f_S^{(2)},\cdots,f_S^{(r)}$,其中 r 为一级核的个数;

第四步:若所有的 $f_S^{(i)}(i=1,2,\cdots,r)$中的每一项均为单属性,则每个子分明函数 $f_S^{(i)}$ 中的所有项组成了决策表信息系统 S 的一个约简,所有的这些约简组成了决策表信息系统 S 的约简簇集,算法结束。否则转入第五步;

第五步:对于没有最简的子分明函数 $f_S^{(i)}$,则找到其中除了一级核以外的所有$|m^i|$最小的项,并取其相同的属性和不同的属性组合作为二级核集,然后再进行分支运算,得到相应的子分明函数。依次类推,直到所有子分明函数中的项均为单属性为止;

第六步:以上述过程建立属性约简树 T,其中一级核为根节点,每个分支函数就构成一条从根节点到叶子节点的路径,其中的节点即为相应的核属性。每条从根节点到叶子节点的路径代表决策表 S 的一个约简。

以上介绍的属性约简方法仅适用于决策表不变的情况,当决策表中的对象动态增加时,如何根据新增对象对原有属性约简结果进行动态高效更新却研究的不多,因而研究属性约简的增量式更新成了主要目标。

4 基于约简树的增量式更新算法(IRART)

4.1 算法步骤

为简化问题的讨论,不妨设当决策表动态改变时,决策属性 D 的取值范围不变。论域 $U=\{x_1,x_2,\cdots,x_n\}$,新增对象记作:x_{n+1},利用遍历树的知识给出 S 的增量式属性约简算法:

输入:决策表 $S=(U,C\cup D)$的属性约简树 T 和新增对象 x_{n+1}。

输出:新决策表 $S_1=(U\cup\{x_{n+1}\},C\cup D)$的属性约简簇集。

第一步:如果新增对象 x_{n+1} 与 S 中某个对象完全相同,则新增对象不改变决策表 S 的约简结果,新决策表 S_1 的约简簇仍为约简树 T 所对应的所有约简,否则转入第二步;

第二步:根据新决策表 S_1 求出新增对象 x_{n+1} 的分辨向量 $N_{x_{n+1}}=(m'_{1,n+1},m'_{2,n+1},m'_{3,n+1},\cdots,m'_{n,n+1})$,$n$ 表示原决策表 S 中对象数,$m'_{i,n+1}$ 表示能够区分 x_i 和 x_{n+1} 的所有属性组合;

第三步:用分辨向量 $N_{x_{n+1}}$ 中的分量 $m'_{1,n+1}$ 对属性约简树 T 中的每条路径进行遍历,若发现在该路径中存在某个节点在

$m'_{i,n+1}$ 中,则结束本路径的遍历,否则将 $m'_{i,n+1}$ 中的任意一个属性添加为该路径的叶子节点。继续进行下一条路径的遍历,直到所有路径遍历结束为止,转入第四步;

第四步:然后考虑其余分量 $m'_{i,n+1}$ 对属性约简树 T 的遍历。直到 $N_{x_{n+1}}$ 中的所有分量对 T 遍历结束。最终所得到的新树 T' 即为新决策表 S' 所对应的属性约简树,树中的每条路径对应着新决策表 S' 的每个约简,这些约简共同组成了 S' 的约简簇集。

4.2　示例说明

实例 1　表 1 所示的是一张决策表,其中共有五个对象和五个属性,$C=\{c_1,c_2,c_3,c_4\}$ 为条件属性集,$D=\{d\}$ 为决策属性集。

表 1　决策表

元素	数据值				
	$c1$	$c2$	$c3$	$c4$	d
$x1$	1	0	1	0	2
$x2$	1	0	1	0	1
$x3$	1	1	1	0	3
$x4$	0	1	0	0	2
$x5$	0	1	1	1	2

对实例 1 首先求出其分辨矩阵:

$$\boldsymbol{M}'_S=\begin{pmatrix} \phi & \phi & c2 & \phi & \phi \\ & \phi & c2 & c1c2c3 & c1c2c4 \\ & & \phi & c1c3 & c1c4 \\ & & & \phi & \phi \\ & & & & \phi \end{pmatrix}$$

由 M'_S 求出属性核 $Core(C)=\{c_2\}$,并将 M'_S 中包含核属性的元素值置为 ϕ,对于相同的元素只保留一个,得到新的分辨矩阵:

$$\boldsymbol{M}''_S=\begin{pmatrix} \phi & \phi & c2 & \phi & \phi \\ & \phi & \phi & \phi & \phi \\ & & \phi & c1c3 & c1c4 \\ & & & \phi & \phi \\ & & & & \phi \end{pmatrix}$$

构造分辨函数:$f_S=c_2\wedge c_1c_3\wedge c_1c_4$。$c_2$ 为一级核,c_1 和 c_3c_4 分别称为二级核最终得到两个分辨子函数:

$f_S^{(1)}=c_2\wedge c_1$

$f_S^{(2)}=c_2\wedge c_3\wedge c_4$

按照序贯属性约简过程建立约简树 T:

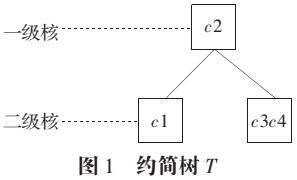

图 1　约简树 T

设新增对象 $x_6=\{1,0,0,1,3\}$,由 IRART 算法得 x_6 的分辨向量为:

$\boldsymbol{N}_{x_6}=\{c_3c_4,c_3c_4,\phi,c_1c_2c_3c_4,c_1c_2c_3\}$

然后利用 $\boldsymbol{N}_{x_6}$ 的每个分量对 T 进行遍历,得到修整的约简树 T'(见图 2)。

T' 中从根节点到叶子节点的每条路径中所包含的节点属性组成新决策表 S' 的一个约简,所有约简组成 S' 的约简簇。

从示例说明可见,利用该算法对新增对象仅需根据树的遍历对原约简树进行少量修整,避免了大量的逻辑演算和粗糙集定性分析,减少了运算量,满足了实际工业生产中对动态决策表属性约简实时高效更新的要求。

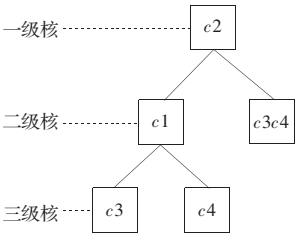

图 2　约简树 T'

4.3　算法的复杂度分析

令原决策表中记录数为 n,属性个数为 m,则由 SARA 算法形成的属性约简树的路径条数最坏情况下为 $m-1$,每个新增对象的分辨向量中分量个数为 n,每个分量中最多有 m 个属性,因此求新增加记录的分辨向量的时间复杂度为 $O(n\times m)$,在遍历循环中,时间复杂度为 $O(n\times m\times(m-1))$,所以在最坏情况下 IRART 算法的时间复杂度为 $O(n\times m^2)$。

可见,在最坏情况下,IRART 算法因可以快速遍历约简树并进行相应修整,实现属性约简的动态高效更新,因而其性能明显优于文献[10]的非增量式属性约简算法。与文献[11]相比,IRART 算法避免了纯粗糙集的定性分析,时间复杂度也有所降低。

5　结束语

增量式学习是人工智能领域一个重要的问题,属性约简是粗集理论研究中最核心的工作之一。提出一种基于分辨矩阵和约简树的增量式属性约简算法,主要考虑对象动态增加情况下属性约简的更新问题。该算法可通过遍历树,根据新增对象的分辨向量动态修整约简树,避免了大量的逻辑演算,简化了问题的复杂度,最终实现了对原有属性约简的增量式更新,为属性约简的增量式更新提出了一条新的途径。但该算法的相关环节还需要进一步的改进,特别是在利用约简树进行增量式约简时,当需要从分辨向量中的某个分量中选取一个属性作为叶子节点插入时,选择哪个属性更有利于简化后面的修整,该文没有给出确定的分析,因此,为了简化运算,找到更好决策表增量式属性约简算法还需要进一步研究。

参考文献:

[1] Pawlaw Z.Rough sets[J].International Journal of Computer and Information Sciences,1982,11(11):341-356.

[2] 王国胤.Rough 集理论与知识获取[M].西安:西安交通大学出版社,2001.

[3] Skowron A,Rauszer C.The discernibility functions matrics and functions in information systems[M]//Intelligent Decision Support-Handbook of Applications and Advances of the Rough Sets Theory.Dordrecht:Kluwer Academic Publisher,1992:331-362.

[4] 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.

[5] 王杨,闫德勒,张凤梅.基于粗糙集和决策树的增量式规则约简算法[J].计算机工程与应用,2007,43(1):170-172.

[6] Cercone V,Tsuchiya M.Luesy editors introdunction[J].IEEE Trans on Knowledge Data Engineering,1993,5(6):901-902.

(下转 134 页)