决策表分解及其最小属性约简研究

叶明全[1],伍长荣[2]
YE Ming-quan[1],WU Chang-rong[2]

1.皖南医学院 计算机教研室,安徽 芜湖 241002
2.安徽师范大学 数学计算机学院,安徽 芜湖 241002
1.Computer Staff Room,Wannan Medical College,Wuhu,Anhui 241002,China
2.Institute of Mathematic and Computer,Anhui Normal University,Wuhu,Anhui 241002,China
E-mail:ymq@wnmc.edu.cn

YE Ming-quan,WU Chang-rong.Decomposition of decision table and computation for minimum attribute reduction. Computer Engineering and Applications,2009,45(30):126-128.

Abstract: Many existing algorithms of attribute reduction begin at constructing decision table's discernibility matrix,then convert non-empty objects' conjunctive normal form into minimal disjunctive normal form.It is important how to get a reduction more efficiently.This paper points out that the minimum attribute reduction algorithm is imperfect in some respect,and an improved algorithm for the minimum attribute reduction based on $U/\{a\}$ partition is proposed.By regarding the significance of attributes defined from the viewpoint of partition granularity as heuristic information,and introducing the heuristic information into $U/\{a\}$ partition which translates attribute reduction problem in macrocosm into attribute reduction problem in subdomain.Theoretical analysis and example show that this algorithm is effective.

Key words: rough set;discernibility maxtrix;mininum attribute reduction;partition granularity;decomposition

摘 要:现有的很多属性约简算法都是由构造决策表的差别矩阵出发,将矩阵中非空元素的合取范式转化为极小析取范式。为提高对大规模数据的决策表进行约简的效率,文中指出基于$U/\{a\}$划分的最小约简算法存在的缺陷,给出以划分粒度为启发式信息,利用单个条件属性把论域划分成多个等价类,将计算整个全域上的属性约简问题转化为计算在相应划分的子区域上属性约简问题,提出了一种基于决策表分解的最小属性约简算法。理论分析和实例表明该约简算法是有效的。

关键词:粗糙集;差别矩阵;最小属性约简;划分粒度;分解
DOI:10.3778/j.issn.1002-8331.2009.30.039 **文章编号:**1002-8331(2009)30-0126-03 **文献标识码:**A **中图分类号:**TP18

粗糙集理论是一种处理不确定信息的有效工具,已被成功地应用于数据挖掘、机器学习与模式识别等领域,成为当前研究热点之一[1-6]。属性约简是粗糙集理论的核心内容之一。决策表的属性约简通常是不唯一的,人们通常希望找到具有最小属性的约简,即最小属性约简[3-4]。基于差别矩阵的属性约简算法是具有代表性的方法之一,主要分为两种:一种依据决策表的差别矩阵求出核,然后依据核逐步添加重要属性,易于实现且计算速度快,但只能求出最优或次优约简[5];另一种是依据决策表的差别矩阵生成一个差别函数,通过对该差别函数进行化简得到一个析取的合取表达式,然后得到所有属性约简[1,6-7]。由于Skowron[1]和Hu[8]定义的差别矩阵只适用于相容决策表,对于不相容决策表,有时会产生错误的结果。为此,国内学者张文修[1]、刘文军[6]、叶东毅[9]等分别提出一些改进的差别矩阵计算方法,既适用于相容决策表,又适用于不相容决策表。但基于差别矩阵的属性约简算法要求生成和存储差别矩阵这个中间环节,随着问题规模增大时,存放差别矩阵的空间和算法执行时间的代价都很大。决策表分解是将一个大型决策表分解为若干规模较小且易于处理的子表,可以减少每次处理的数据量,避免直接在复杂系统中建模的困难和缺陷,提高数据分析的效率和质量[10]。在文献[3]的基础上提出一种基于决策表分解的最小属性约简完备算法,以划分粒度为启发式信息确定$U/\{a\}$划分,将计算整个全域上的差别矩阵与差别函数问题转化为计算在相应划分的子区域上差别矩阵与差别函数问题,并通过理论分析和实例验证了算法的有效性和完备性。

1 粗糙集的相关概念

为便于叙述,对粗糙集概念作简要介绍,有关详细概念请参见文献[1]。

基金项目:安徽省高校省级自然科学研究项目(the Natural Science Foundation of Department of Education of Anhui Province of China under Grant,No.KJ2008B039)。
作者简介:叶明全(1973-),男,副教授,主要研究领域为粗糙集理论与数据挖掘;伍长荣(1973-),女,副教授,主要研究领域为数据挖掘与数据仓库。
收稿日期:2009-04-28 **修回日期:**2009-06-17

定义 1 决策表 S 为一个四元组 $S=(U,A,V,f)$,其中:U 表示有限对象的集合,称为论域;A 是属性的非空有限集,$A=C\cup D$,$C\cap D=\varnothing$,C 为条件属性集,D 为决策属性集,通常只含有一个决策属性;$V=\cup V_a$,V_a 为属性 a 的值域;f 是 $U\times C\cup D\rightarrow V$ 的映射,$\forall x_i\in U$,有 $f(x_i,a)\in V_a$。

定义 2 决策表 $S=(U,A,V,f)$,$P\subseteq A$ 在 U 上的不可区分关系 $IND(P)$定义为 $IND(P)=\{(x,y)\in U^2|\forall a\in P,f(x,a)=f(y,a)\}$。$IND(P)$的所有等价类构成的集合 $U/IND(P)=\{U_1,U_2,\cdots,U_n\}$。$U/IND(P)$构成了 U 的一个划分,简记为 U/P。用$|\cdot|$表示集合的基,$|U_1|+|U_2|+\cdots+|U_n|=|U|$。

定义 3 决策表 $S=(U,A,V,f)$,$A=C\cup D$,$U/C=\{X_1,X_2,\cdots,X_n\}$,$U/D=\{Y_1,Y_2,\cdots,Y_m\}$,条件属性 C 关于决策属性 D 的正域 $POS_C(D)$表示为 $POS_C(D)=\cup\{X_i|X_i\subseteq Y_j,X_i\in U/C,Y_j\in U/D\}$。

定义 4 决策表 $S=(U,A,V,f)$,$A=C\cup D$,$R\subseteq C$,若 $POS_R(D)=POS_C(D)$且对 $\forall r\in R$,都有 $POS_R(D)\neq POS_{R-\{r\}}(D)$,则称 R 是 S 的一个属性约简。记 S 的所有属性约简为 $red(S)$。$red(S)$中属性数目最少的约简称为 S 的最小属性约简。

定义 5[11] 决策表 $S=(U,A,V,f)$,$P\subseteq A$,$U/P=\{X_1,X_2,\cdots,X_n\}$,知识 P 的划分粒度 $E(P)$定义为

$$E(P)=\sum_{i=1}^{n}|X_i|\frac{|X_i|}{|U|}$$

性质 1[11] 决策表 $S=(U,A,V,f)$,$P\subseteq A$,$U/P=\{X_1,X_2,\cdots,X_n\}$,则 $1\leqslant|U|/n\leqslant E(P)\leqslant|U|$。

性质 1 表明,当知识的划分粒度越小时,它的分类能力越强。特别当 $n=|U|$时,即每个划分粒中只包含一个对象,此时该知识能把所有对象彻底区分开,它的分类能力达到最强。

2 基于差别矩阵的属性约简算法

文献[6]定义的差别矩阵能够同时适用于相容决策表和不相容决策表,与文献[1]定义的差别矩阵是等价的,但具有更高的计算效率。

定义 6[6] 决策表 $S=(U,A,V,f)$的差别矩阵 $\boldsymbol{M}(S)$是一个 $n\times n$ 的矩阵,$n=|U|$,其第 i 行第 j 列的元素为 $C_{ij}=\{a\in C|f(u_i,a)\neq f(u_j,a),(u_i,u_j)\notin IND(D)$且 u_i、u_j 中至少有一个属于 $POS_C(D)\}$;否则 $C_{ij}=\varnothing$。

定义 7 决策表 $S=(U,A,V,f)$的差别集 A_S 定义为 $A_S=\{C_{ij}|C_{ij}\in M(S)\wedge C_{ij}=\varnothing\}$。

文献[6]提出的基于改进差别矩阵的属性约简算法:

步骤 1 按定义 6 计算决策表 S 的差别矩阵 $\boldsymbol{M}(S)$;

步骤 2 令 $P=\{a\in C_{ij}|C_{ij}=\{a\}\}$,即 P 为 $\boldsymbol{M}(S)$中所有单属性元素所组成的属性集合。若 $\boldsymbol{M}(S)$中元素 C_{kl} 包含 P 中元素,则令 $C_{kl}=\varnothing$,得一新矩阵。对新矩阵中的所有非 $\varnothing$ 元素 C_{ij},建立相应的析取逻辑表达式 L_{ij},对所有 $a_i\in C_{ij}$,$L_{ij}=\vee a_i$;

步骤 3 将所有的析取逻辑表达式 L_{ij},进行合取运算,得一个合取范式 L,即 $L=\wedge L_{ij}$;

步骤 4 将合取范式 L 转换为析取范式的形式,得 $L'=\vee L_i$;

步骤 5 输出属性约简结果。将 P 中所有属性加入到析取范式中的每个合取项就对应一个属性约简的结果。

3 基于决策表分解的最小属性约简算法

为了实现对大规模数据的决策表进行属性约简,文献[3]任取单个条件属性 a 对 U 进行划分,获取若干个子决策表,提出基于 $U/\{a\}$划分的最小属性约简构造,可有效地降低算法的计算量和存储空间。借助这个思想,进一步讨论决策表分解若干个子决策表的有关性质,并引入划分粒度,给出改进属性约简算法。

性质 2[3] 决策表 $S=(U,A,V,f)$,$A=C\cup D$,$R\subseteq C$,则 $POS_R(D)=POS_C(D)$的充分必要条件是 $\forall P\in A_S$,$P\cap R\neq\varnothing$。

定义 8 决策表 $S=(U,A,V,f)$,$A=C\cup D$,$a\in C$,$n=|V_a|$,$U/\{a\}=\{U_1,U_2,\cdots,U_n\}$,则第 i 个等价类构成原决策表 S 的第 i 个子决策表($1\leqslant i\leqslant n$),记为 $S_i=(U_i,A-\{a\},V,f)$。

为便于叙述,对于任一子决策表 $S_i=(U_i,A-\{a\},V,f)$,其差别矩阵记为 $\boldsymbol{M}(S_i)$,差别集记为 $A_{S(i)}$,约简集记为 $red(S_i)$,设 $G=\{R_a|R_a=\cup R_i,R_i\in red(S_i)\}$,$H=\{R_a\cup\{a\}|R_a\in G\}$。

性质 3[3] 如果 u、v 属同一子系统 S_t,则 $C_{uv}(S)=C_{uv}(S_t)$。

性质 4 $U_t\subseteq IND(D)$当且仅当 $red(S_t)=\varnothing$。

证明 $U_t\subseteq IND(D)$,则由定义 6,任意 u_i、$u_j\in U_t$,$(u_i,u_j)\in IND(D)$,得 $C_{ij}(S_t)=\varnothing$,显然 $red(S_t)=\varnothing$。反之显然。

性质 5 任意 $U_t\in U/\{a\}$,$U_t\subseteq IND(D)$当且仅当 $G=\varnothing$。

证明 由性质 4 和 G 的定义,即可得到。反之显然。

性质 6 若 $R\in red(S)$,则对于任意 S_t,存在 $R_t\in red(S_t)$满足 $R_t\subseteq R$。

证明 对于任意 S_t,若 $red(S_t)=\varnothing$,则 $R_t=\varnothing$,显然满足 $R_t\subseteq R$;若 $red(S_t)\neq\varnothing$,证明过程参见文献[3]。

定理 1 若 $G=\varnothing$,则 $POS_{\{a\}}(D)=POS_C(D)$。

证明 若 $G=\varnothing$,由性质 4 和性质 5 可知,任意 $U_t\in U/\{a\}$时 $red(S_t)=\varnothing$,由性质 2 得,若任意 u_i、$u_j\in U_t$ 时 $C_{ij}(S)=\varnothing$;若任意 u_i、u_j 不同时属于 U_t 时 $\{a\}\subseteq C_{ij}(S)$。由定义 7 与性质 2 可得,$\forall P\in A_S$,$P\cap\{a\}=\{a\}\neq\varnothing$,因而 $POS_{\{a\}}(D)=POS_C(D)$。

由定理 1 可知,若 $G=\varnothing$,则$\{a\}$为决策表 S 的一个最小属性约简。

定理 2 若 $R\in H$,则 $POS_R(D)=POS_C(D)$。

证明 若 $G=\varnothing$,则 $H=\{\{a\}\}$,得 $R=\{a\}$,由定理 1 可知,$POS_R(D)=POS_C(D)$;若 $G\neq\varnothing$,证明过程参见文献[3]。

由定理 2 可知,H 中每个元素(属性子集)保持了原决策表的分类能力,故包含了原决策表的一个属性约简。

定理 3 若 $P\in red(S)$,则存在 $R\in H\cup G$,$|R|\leqslant|P|$且 $POS_R(D)=POS_P(D)$。

证明 $P\in red(S)$,则 $POS_P(D)=POS_C(D)$。若 $G=\varnothing$,则 $H=\{\{a\}\}$,存在 $R=\{a\}$,$POS_R(D)=POS_C(D)$,可得$|R|\leqslant|P|$且 $POS_R(D)=POS_P(D)$;若 $G=\varnothing$,证明过程参见文献[3]。

由定理 3 可知,决策表 S 中任意给定的一个属性约简 P,可根据 $U/\{a\}$划分所得子决策表计算 $H\cup G$,并存在一个属性个数不比$|P|$大的约简。

定理 4 P 为 S 的最小属性约简,则存在 $R\in H\cup G$,$|R|=|P|$且 $POS_R(D)=POS_P(D)$。

证明 P 为 S 的最小属性约简,则 $POS_P(D)=POS_C(D)$。若 $G=\varnothing$,由定理 1,存在 $R=\{a\}\in H\cup G$,$POS_R(D)=POS_C(D)$,因而 $POS_R(D)=POS_P(D)$;若 $G\neq\varnothing$,证明过程参见文献[3]。

由定理 4 可知,根据 $H\cup G$ 能够找到 S 的一个最小属性约简,且是完备的。

定理 5 决策表 $S=(U,A,V,f)$,$A=C\cup D$,$a\in C$,$U/\{a\}=\{U_1,$

$U_2,\cdots,U_n\}$,$S_i=(U_i,A-\{a\},V,f)$,$G=\{R_a|R_a=\cup R_i,R_i\in red(S_i)\}$,$m=\min(|R|)(R\in G)$,$G_m=\{R|R\in G,|R|=m\}$,若存在 $P\in G_m$ 且 $POS_P(D)=POS_C(D)$,则 P 是 S 的最小属性约简;否则对任意 $P\in G_m$,$P\cup\{a\}$是 S 的最小属性约简。

证明 若 $G=\varnothing$,则 $G_m=\varnothing$,由定理 1 可得,对任意 $P\in G_m$,$P\cup\{a\}$是 S 的最小属性约简。若 $G\neq\varnothing$,采用反证法:(1)假设存在 $P\in G_m$ 且 $POS_P(D)=POS_C(D)$,P 不是 S 的最小属性约简。设 $B\in red(S)$为 S 中最小属性约简,则$|B|<|P|=m$。由定理 4 可知,存在 $R\in H\cup G$,使得$|R|=|B|$,可得$|R|<m$。又由于 $R\in H\cup G$,可知$|R|\geqslant m$。这与假设矛盾;(2)假设对任意 $P\in G_m$,$POS_P(D)\neq POS_C(D)$,$P\cup\{a\}$不是 S 的最小属性约简。由定理 2 可知,$POS_{P\cup\{a\}}(D)=POS_C(D)$。设 $B\in red(S)$为 S 中最小属性约简,则$|B|<|P\cup\{a\}|=m+1$。由定理 4 可知,存在 $R\in H\cup G$,使得$|R|=|B|$,且 $POS_R(D)=POS_B(D)$,可得$|R|<m+1$。又由于 $R\in H\cup G$,可知$|R|\geqslant m+1$。显然矛盾。命题得证。

推论 1 决策表 $S=(U,A,V,f)$,$A=C\cup D$,$a\in C$,$U/\{a\}=\{U_1,U_2,\cdots,U_n\}$,$S_i=(U_i,A-\{a\},V,f)$,$G=\{R_a|R_a=\cup R_i,R_i\in red(S_i)\}$,$m=\min(|R|)(R\in G)$,$G_m=\{R|R\in G,|R|=m\}$,若 a 为决策表的核属性,则对任意 $P\in G_m$,$P\cup\{a\}$是 S 的最小属性约简。

证明 由定理 5 显然得证。

文献[3]算法没有考虑不同条件属性对整个论域的划分能力是不同,也没有考虑到若 $G=\varnothing$ 则$\{a\}$就是最小属性约简的情况。提出利用划分粒度来定量地表示条件属性的划分能力,在此基础上研究并设计了一个基于决策表分解的高效约简算法,步骤如下:

步骤 1 对决策表 S 从条件属性集 C 中依次取一个属性 c,计算 $E(c)$;

步骤 2 令 a 为 $E(c)$中划分粒度最小的属性,若 $POS_{\{a\}}(D)=POS_C(D)$,则 $P=\{a\}$是 S 的最小约简,退出;否则,计算 $U/\{a\}=\{U_1,U_2,\cdots,U_n\}$;

步骤 3 利用文献[6]提出的属性约简算法,计算各子决策表 $S_i=(U_i,A-\{a\},V,f)$ 的所有约简,得到 S_i 的约简集 $red(S_i)$,$1\leqslant i\leqslant n$;

步骤 4 求 $G=\{R_a|R_a=\cup R_i,R_i\in red(S_i)\}$,设 G 中元素最小基数为 m,计算 $G_m=\{R|R\in G,|R|=m\}$;

步骤 5 若 $\exists P\in G_m$,满足 $POS_P(D)=POS_C(D)$,则 P 是 S 的最小约简;否则任意取 $P\in G_m$,$P\cup\{a\}$是 S 的最小约简。

基于差别矩阵的属性约简算法复杂度为 $O(K|U|^2)$[1,7],其中,K 是关于$|U|$与$|A|$的多项式或指数级式子。根据 $U/\{a\}=\{U_1,U_2,\cdots,U_n\}$($|U_1|+|U_2|+\cdots+|U_n|=|U|$)将决策表分解后,算法复杂度为 $O(K|U_1|^2+K|U_2|^2+\cdots+K|U_n|^2)$,明显优于 $O(K|U|^2)$。且该算法选择知识划分粒度最小的单个条件属性进行决策表分解,其约简效率显然优于选择其他属性。

4 实例分析

下面以文献[6]提供的不相容决策表(如表 1 如示,其中 $C=\{a,b,c\}$,$D=\{d\}$)为例,介绍该文算法来求解决策表的一个最小属性约简。

表 1 不相容决策表 S

U	a	b	c	d	U	a	b	c	d
1	0	0	0	1	5	0	0	0	5
2	0	0	1	2	6	2	2	2	3
3	2	1	2	3	7	2	0	1	2
4	1	0	2	4	8	0	0	1	5

首先计算每个条件属性的划分粒度,得 $E(\{a\})=26/8$,$E(\{b\})=38/8$,$E(\{c\})=22/8$。选择划分粒度最小的属性 c,因 $POS_{\{c\}}(D)\neq POS_C(D)$,故对决策表进行分解,$U/\{c\}=\{U_1,U_2,U_3\}$,其中 $U_1=\{1,5\}$,$U_2=\{2,7,8\}\}$,$U_3=\{3,4,6\}$。计算每个子决策表的差别矩阵,依次得到三个差别集:$A_{S1}=\varnothing$,$A_{S2}=\{a\}$,$A_{S3}=\{ab,ab\}$。所以各子决策表的属性约简集为:$red(S_1)=\varnothing$,$red(S_2)=\{\{a\}\}$,$red(S_3)=\{\{a\},\{b\}\}$。可得 $G=\{\{a\},\{ab\}\}$,G 中元素最小基数为 1,$G_1=\{\{a\}\}$,因 $POS_{\{a\}}(D)\neq POS_C(D)$,所以 $P=\{a,c\}$是 S 的最小属性约简。结果与文献[6]中所求的一个最小约简相同。

5 结束语

基于差别矩阵的属性约简算法研究是粗糙集理论的一个核心内容。对于大型决策表数据海量性和复杂性问题,决策表分解可避免直接在复杂系统中生成差别矩阵的困难和缺陷,同时也可降低差别函数计算量,是求解属性约简的一种有效手段。该文算法运用知识划分粒度来定量选择单个条件属性,使得决策表根据单一条件属性分解更合理,且可进一步提高属性约简的效率。

参考文献:

[1] 张文修,吴伟志,梁吉业.粗糙集理论与方法[M].北京:科学出版社,2003.

[2] 王春年,梁吉业.基于粗糙集与属性值聚类的决策树改进算法[J].计算机工程与应用,2007,43(31):178-181.

[3] 李订芳,李贵斌,章文.基于 $U/\{a\}$划分的最小约简构造[J].武汉大学学报:理学版,2005,51(3):269-272.

[4] 叶东毅,廖建坤.基于二进制粒子群优化的一个最小属性约简算法[J].模式识别与人工智能,2007,20(3):295-300.

[5] 唐彬,李龙澍.启发式属性约简算法完备性和规则发现算法的研究[J].计算机工程与应用,2003,39(30):191-194.

[6] 刘文军,谷云东,冯艳宾,等.基于可辨识矩阵和逻辑运算的属性约简算法的改进[J].模式识别与人工智能,2004,17(1):119-123.

[7] 王元珍,裴小兵.基于 Skowron 分明矩阵快速约简算法[J].计算机科学,2005,32(4):42-44.

[8] Hu X H,Cercone N.Learning in relational databases:A rough set approach[J].Computational Intelligence,1995,11(2):323-337.

[9] 叶东毅,陈昭炯.不相容决策表属性约简计算的一个可辨识矩阵方法[J].福州大学学报:自然科学版,2005,33(1):11-15.

[10] 王加阳,刘柳明,罗安.大型决策表分解方法研究[J].计算机科学,2007,34(8):211-214.

[11] 冯琴荣,苗夺谦,程昳.决策表属性约简的相对划分粒度表示[J].小型微型计算机系统,2008,29(12):2305-2308.