基于粒计算的属性约简算法

赵 敏,罗 可,秦 哲
ZHAO Min,LUO Ke,QIN Zhe

长沙理工大学 计算机与通信工程学院,长沙 410076
School of Computer and Communication Engineering,Changsha University of Science and Technology,Changsha 410076,China
E-mail:pzhaomin@sina.com

ZHAO Min,LUO Ke,QIN Zhe.Attribute reduction algorithm based on granular computing.Computer Engineering and Applications,2008,44(30):157-159.

Abstract: Granular computing is a new intelligent computing theory and method based on issue partition concepts.Inconsistent decision table is an important aspect in rough set theory.Using the equivalent relation in rough set to construct particle,this paper offers decompounds method of particle,furthermore gives the heuristic information in attribute reduction algorithm based on attribute importance.At last,experimental result shows that the new algorithm is not only very efficient but also can treat with large decision table.

Key words: rough set;granular computing;attribute reduction;attribute importance

摘 要:粒计算是一种基于问题概念空间划分的新的智能计算理论和方法,不相容决策表是粗糙集理论研究的一个重点。利用粗糙集中的等价关系来构建粒子,给出了决策表系统的粒子分解方法及在粒表示下以属性重要性作为启发信息的属性约简算法。实验结果表明该算法不仅具有高效性,而且能处理大型决策表。

关键词:粗糙集;粒计算;属性约简;属性重要性
DOI:10.3778/j.issn.1002-8331.2008.30.048 **文章编号:**1002-8331(2008)30-0157-03 **文献标识码:**A **中图分类号:**TP18

1 引言

基于正域的属性约简是粗糙集理论中一个重要的研究课题。对于一个决策表,人们总是期望能找出所有约简或最小约简,然而一个信息表的属性约简并不是唯一的,得到信息表的包含最少条件属性的约简(即最小约简)已被证明是 NP 完全问题[1],解决这一问题通常采用启发式搜索方法,求出最佳或此最佳约简[2-3]。

本文以快速缩小搜索空间为目的设计了一个基于粒计算[4]的属性约简算法,该算法以粒计算为基础,利用粗糙集中的等价关系来构建粒子,给出了决策表中的粒子分解方法,在此基础上提出以属性重要性作为启发信息的属性约简算法,实验结果表明该算法不仅具有高效性,而且能用于大型决策表。

2 决策信息系统的粒子分解

设函数 $f^{-1}(a,v)$ 表示在属性 $a(a\in A)$ 上值为 v 的对象集合,决策信息系统的粒子定义[5]为:$Gr=((a,v),f^{-1}(a,v))$,其中 (a,v) 为粒子 Gr 的语法,Gr 被称为决策信息系统中的原子粒子。信息系统 S 分解之后的粒子集合为 Grs,其中 $\forall Gr\in Grs$,Gr 的语法是由条件属性集 C 中的所有属性来表示的,即 $Gr=(\phi,f^{-1}(\phi))$,$\phi=(a_1,v_1)\wedge(a_2,v_2)\wedge\cdots\wedge(a_m,v_m)$,$(m=|A|,\forall i(1\leqslant i\leqslant m))$,且满足 $\forall x\in U,\exists Gr\Rightarrow x\in f^{-1}(Gr)$。把满足上面条件的粒子集合称为信息系统的一个粒子空间。

定理 1 设 Grs 是决策表 S 根据算法 1 得到的粒子空间,则 S 是完全确定的决策表的充要条件是:$\forall Gr\in Grs$ 都有 $\forall x_1,x_2$ $(x_1\in Gr\wedge x_2\in Gr\Rightarrow d(x_1)=d(x_2))$。

证明 容易得证。

下面给出求决策表粒子空间 Grs 的算法:根据 U 中由条件属性集决定的等价类来构造粒子,对于 U 中的每一个等价类,直接使用该等价类中对象的所有属性值构造粒子。

算法 1 计算决策信息系统的一个粒子空间 Grs,正域粒子空间 Gr_{pos} 及非正域粒子空间 Gr_{neg}。

输入:决策信息系统 $S=(U,C,D,\{V_a|a\in C\cup D\},\{f_a|a\in C\cup D\})$,$P\subseteq C$。

输出:决策信息系统的一个粒子空间 Grs,Gr_{pos},Gr_{neg}。

算法过程如下:

(1)令 $Grs=\varnothing$,$Gr_{pos}=\varnothing$,$Gr_{neg}=\varnothing$,$m=|C|$,$n=|U|$,$i=1$;

(2)计算 U 中由条件属性集 C 决定的条件划分,设 $k=|U/IND(C)|$;

(3)While($i\leqslant k$)

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.10471036,No.60474070);湖南省科技计划项目基金(No.05FJ3074);湖南省教育厅重点项目基金(No.07A001)。

作者简介:赵敏(1982-),女,硕士研究生,主要研究方向为数据库技术、数据挖掘;罗可(1961-),男,教授,博士,主要研究方向为数据库技术、数据挖掘、计算机应用;秦哲(1981-),男,主要研究方向为计算机应用。

收稿日期:2007-11-28 **修回日期:**2008-02-03

①对于条件属性集 C 决定的条件划分中的任一等价类 $U/IND(C)$，$\forall x\in(U/IND(C)_k)$利用 x 在所有条件属性上的取值构造粒子 $Gr=(\phi,f^{-1}(\phi),\phi=(a_1,v_1)\wedge(a_2,v_2)\wedge\cdots\wedge(a_m,v_m))$，其中对应的 v_i 即 x 在属性 a_i 上的取值。$Grs=Grs\cup\{Gr\}$；

②$i=i+1$；

(4)依次遍历 Grs 的所有粒子，$\forall Gr\in Grs$，进行以下处理：

if 在 Gr 中存在两个元素，满足 x 和 y 在决策属性集上取值不相等；

then $Gr_{neg}=Gr_{neg}+\{Gr\}$；

else 在 Gr 的语法中加入决策属性集以及 Gr 中的元素在决策属性集上的取值，并进行操作：$Gr_{pos}=Gr_{pos}+\{Gr\}$。

3 基于粒计算的高效属性约简算法

3.1 高效属性约简算法的思想

在决策表中，过去的属性约简算法在计算 $POS_P(D)(P\subseteq C)$ 时都是建立在整个对象集 U 上的，而在计算 $POS_{P\cup\{a\}}(D)(a\in C-P)$时也没有用到已知的信息 U/P。本文以快速缩小搜索空间为目的，设计出一种高效的属性约简算法，将 $POS_P(D)(P\subseteq C)$ 的计算建立在简化后的决策表粒子空间 Grs 中，利用已知信息 U/P 递归求出 $U/P\cup\{a\}$。为更清楚地说明这一思想，给出如下定理。

定义 1 决策表 $S=(U,C,D,V,f)$，$Grs=(Gr,C,D,V,f)$为其分解后的粒子空间；对 $\forall B\subseteq C$，定义 $POS_B'(D)=\bigcup_{x\in Gr/B\wedge X\subseteq Gr_{POS}}X$。

定理 2 在决策表 $S=(U,C,D,V,f)$ 中，$Grs=(Gr,C,D,V,f)$为其分解后的粒子空间，$\forall B\subseteq C$，若 $POS_B'(D)=Gr_{POS}$，则 $POS_B'(D)=POS_C(D)$。

证明 由于$\forall B\subseteq C$，故 $POS_B(D)\subseteq POS_C(D)$。假设 $POS_B'(D)\neq POS_C(D)$，则一定存在 $u_l\in POS_C(D)\wedge u_l\notin POS_B'(D)$，故$[u_l]_C\subseteq POS_C(D)$但$[u_l]_C\not\subset POS_B'(D)$；由于$[u_l]_C\subseteq POS_C(D)\Rightarrow[Gr_i]_C=[u_l]_C$ $(Gr_i\in Grs_{POS})$，从而有$[Gr_i]_C\in POS_C(D)$且$[Gr_i]_C\notin POS_B(D)$，故至少存在 $Gr_t\in[Gr_i]_B\in U/B(Gr_t\in Gr)$使得 $f(Gr_t,D)\neq f(Gr_i,D)$，故存在$\{Gr_i,Gr_t\}\subseteq X\subseteq[Gr_i]_B$，使得 $X\in Gr/B$ 且$|X/B|\neq1$。由定义 1 可知，$Gr_i\notin POS_B(D)$，故 $POS_B'(D)\neq Gr_{POS}$，这与条件矛盾，从而假设不成立，故命题成立。证毕。

定理 3[7]：决策表 $S=(U,C,D,V,f)$，$Grs=(Gr,C,D,V,f)$为其分解后的粒子空间。若 $\forall B\subseteq C$ 有 $POS_B'(D)=Gr_{POS}$ 且 $\forall b\in B$ 均有 $POS_B'(D)\neq POS_{B-\{b\}}'(D)$，则 B 是 C 相对于 D 的属性约简。

定理 3 说明了求属性约简的过程可以建立在决策表分解后的粒子空间，这为下面给出的高效属性约简算法提供了理论依据。文献[8-10]中给出了一类很好的计算正域的方法，其基本思路是：假设已计算出 $U/P=\{P_1,P_2,\cdots,P_t\}(P\subseteq C)$，依次判断 $P_i\in U/P$ 是否在 $POS_P(D)$中。若 P_i 在 $POS_P(D)$中，则 $U=U-P_i$，$POS_C(D)=POS_C(D)-P_i$，这样就使得被搜索的对象空间 U 不断缩小，有利于算法效率的提高；然后对不属于 $POS_P(D)$的 P_i，利用性质 $U/(P\cup\{a\})=\bigcup_{X\in U/P}(X/\{a\})(a\in(C-P))$进行进一步划分，使得算法的效率进一步提高。在这些文献中用到的启发信息是近似质量，分析上述算法可发现：若 P_i 在 $U-POS_C(D)$中，这样的 P_i 也可以去掉，因为它对计算正域不起任何作用。基于这种思路和定理 2，可给出如下启发式信息。

定义 2[7] 在决策表 $S=(U,C,D,V,f)$中，$Grs=(Gr,C,D,V,f)$为其分解后的粒子空间，$P\subseteq C$，$\forall a\in(C-P)$的重要性定义为：

$sig_P(a)=|Gr_{P\cup\{a\}}-Gr_P|$，其中 $Gr_P=\{\bigcup_{X\in Gr/P\wedge X\subseteq Gr_{POS}}X\}\cup\{\bigcup_{X\in Gr/P\wedge X\subseteq Gr_{neg}}X\}$，规定 $Gr_\emptyset=\emptyset$。

定理 4[8-10] 在决策表 $S=(U,C,D,V,f)$中，$P\subseteq C$，$\forall a\in(C-P)$，则，$U/(P\cup\{a\})=\bigcup_{X\in U/P}(X/\{a\})$。

定理 5 在决策表 $S=(U,C,D,V,f)$中，$Grs=(Gr,C,D,V,f)$为其分解后的粒子空间，$P\subseteq C$，$\forall a\in(C-P)$，则 $Gr/(P\cup\{a\})=\bigcup_{X\in Gr/P}(X/\{a\})$。

证明 由定理 4 即可得。

定理 6 在决策表 $S=(U,C,D,V,f)$中，$Grs=(Gr,C,D,V,f)$为其分解后的粒子空间，$P\subseteq C$，$\forall a\in(C-P)$，则有

$$sig_P(a)=|\bigcup_{X\in Gr/P}\{\{\bigcup_{X\not\subset Gr_{POS}\wedge y\in X/\{a\}\wedge y\subseteq Gr_{POS}}y\}\cup\{\bigcup_{X\not\subset Gr_{neg}\wedge y\in X/\{a\}\wedge y\subseteq Gr_{neg}}y\}\}|$$

证明 由定理 5 可知：

$$sig_P(a)=|Grs_{P\cup\{a\}}-Grs_P|=|\{\bigcup_{X\in Grs/(P\cup\{a\})\wedge X\subseteq Grs_{POS}}X\}\cup$$

$$\{\bigcup_{X\in Grs/(P\cup\{a\})\wedge X\subseteq Grs_{neg}}X\}-\{\bigcup_{X\in Grs/P\wedge X\subseteq Grs_{POS}}X\}\cup$$

$$\{\bigcup_{X\in Grs/P\wedge X\subseteq Grs_{neg}}X\}|=|\bigcup_{X\in Grs/P}\{\{\bigcup_{y\in X/\{a\}\wedge y\subseteq Grs_{POS}}y\}\cup$$

$$\{\bigcup_{y\in X/\{a\}\wedge y\subseteq Grs_{neg}}y\}\}-\{\bigcup_{X\in Grs/P\wedge X\subseteq Grs_{POS}}X\}\cup\{\bigcup_{X\in Grs/P\wedge X\subseteq Grs_{neg}}X\}|=$$

$$\bigcup_{X\in Grs/P}\{\{\bigcup_{X\subseteq Grs_{POS}}X\}\cup\{\bigcup_{X\not\subset Grs_{POS}\wedge y\in X/\{a\}\wedge y\subseteq Grs_{POS}}y\}\}\cup$$

$$\{\{\bigcup_{X\not\subset Grs_{neg}\wedge y\in X/\{a\}\wedge y\subseteq Grs_{neg}}y\}\}\cup\bigcup_{X\in Grs/P}\{\bigcup_{X\subseteq Grs_{neg}\wedge y\in X/\{a\}\wedge y\subseteq Grs_{neg}}y\}-$$

$$\bigcup_{X\in Grs/P}\{\bigcup_{X\in Grs/P\wedge X\subseteq Grs_{POS}}X\}\cup\{\bigcup_{X\in Grs/P\wedge X\subseteq Grs_{neg}}X\}=$$

$$|\bigcup_{X\in Gr/P}\{\{\bigcup_{X\not\subset Gr_{POS}\wedge y\in X/\{a\}\wedge y\subseteq Gr_{POS}}y\}\cup\{\bigcup_{X\not\subset Gr_{neg}\wedge y\in X/\{a\}\wedge y\subseteq Gr_{neg}}y\}\}|$$

证毕。

3.2 高效属性约简算法的实现

下面给出在决策表分解后的粒子空间 $Grs=(Gr,C,D,V,f)$中计算 $sig_P(a)$的算法，其中 $P\subseteq C$，$\forall a\in(C-P)$。

算法 2 Calculate(P,a)

输入：$(Gr-Gr_P)/P$，$\forall a\in(C-P)$，Gr_{pos}，Gr_{neg}；

输出：$sig_P(a)$；$B_P(a)=\bigcup_{X\in(Gr-Gr_P)/P}\{\bigcup_{y\in X/\{a\}\wedge y\subseteq Gr_{pos}}y\}$，其中 $B_P(a)$表示在$(Gr-Gr_P)/P\cup\{a\}$满足如下条件的所有等价类的并集：该等价类中的所有元素都在 Gr_{pos} 中，并且该等价类的所有元素在决策属性上取相同的值。

$NB_P(a)=\bigcup_{X\in(Gr-Gr_P)/P}\{\bigcup_{y\in X/\{a\}\wedge y\subseteq Gr_{pos}}y\}$，其中 $NB_P(a)$表示$(Gr-Gr_P)/P\cup\{a\}$满足如下条件的所有等价类的并集：该等价类中的所有元素都在 Gr_{neg}；$(Gr-Gr_P)/P\cup\{a\}$。

(1)$sig_P(a)=0$；$B_P(a)=NB_P(a)=\emptyset$。

(2)对任意 $X\in(Gr-Gr_P)/P$,做如下处理:

①统计 X 中所有对象在属性 a 中取值,计算 $X/\{a\}$;

②对任意 $X_i\in X$ 如下处理:若 X_i 中的所有元素均在 Gr_{pos} 中,则将其所有元素并入 $B_P(a)$;若 X_i 中的所有元素均在 Gr_{neg} 中,则将其所有元素并入 $NB_P(a)$。

(3)$sig_P(a)=|B_P(a)\cup NB_P(a)|$。

下面给出决策表 $S=(U,C,D,V,f)$的属性约简算法。

算法 3　*rabasedsig*( )

输入:决策表 $S=(U,C,D,V,f)$;

输出:属性约简 R。

(1)用算法 1 计算出 Gr,Gr_{pos},Gr_{neg};

(2)$R=\varnothing$;

(3)对任意 $\forall a\in(C-R)$有如下处理:调用 Calculate(P,a);得出 $sig_R(a),B_R(a),NB_R(a)$和 $Gr/R\cup\{a\}$;　　//此处 Gr 实际为 $Gr-Gr_R$

(4)记 $sig_R(a)=\max\limits_{a\in(C-R)} sig_R(a)$,如这样的属性不只一个时,则任取其一;

(5)$R=R\cup\{a\}$;$Gr=Gr-B_R(a')-NB_R(a')$;　//此处 Gr 实际为 $Gr-Gr_R$

(6)若 $Gr=\varnothing$,则输出 R;否则转(7);

(7)$Gr_{POS}=Gr_{POS}-B_R(a')$;$Gr_{neg}=Gr_{neg}-NB_R(a')$;

(8)计算 $Gr/R\cup\{a'\}$;　　//此处是指从(3)中得到的 $Grs/R\cup\{a'\}$中去掉 $B_R(a')/a'$和 $NB_R(a')/a'$转(3)。

3.3　算法时间复杂度分析

算法时间复杂度分析:算法 *rabasedsig*( )的步(1)的时间复杂度为 $o(|C||U|)$;步(3)中计算 $sig_R(a)$的时间复杂度为 $o(|Gr-Gr_R|)$。因此,算法 3~8 的总时间复杂度为 $o(|C-1||Gr-Gr_{R_1}|)+\cdots+o(|C-R_n||Gr-Gr_{R_n}|)$($R_n$ 为属性约简),因而最差的时间复杂度为 $o(|C|^2|Gr/C|)$,所以算法 *rabasedsig*( )的最差时间复杂度为 $\max(o(|C||Gr|),o(|C^2||Gr/C|))$。

4　实例分析

以表 1 为例说明算法 4*rabasedsig*( )。在表 1 中,由算法 1 可得:$Gr=\{Gr_1,Gr_2,\cdots,Gr_9\}$,其中 $Gr_1=\{x_1\},Gr_2=\{x_2\},Gr_3=\{x_3\},Gr_4=\{x_4\},Gr_5=\{x_5\},Gr_6=\{x_6\},Gr_7=\{x_7,x_8\},Gr_8=\{x_9,x_{11}\},Gr_9=\{x_{10},x_{12}\}$;由算法 2 可得,$Gr_{pos}=\{Gr_1,Gr_2,Gr_3,Gr_4,Gr_5,Gr_6,Gr_7,Gr_9\}$;$Gr_{neg}=\{Gr_8\}$;

表 1　一个决策表信息系统

R	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
a	1	1	2	3	3	3	1	1	1	3	1	3
b	1	1	1	2	3	3	2	2	3	2	3	2
c	2	2	2	2	1	1	2	2	1	1	1	1
d	0	1	0	0	0	1	0	0	0	0	0	0
D	0	1	0	0	0	1	1	1	0	0	1	0

由算法3 可得:当 $R=\{b,d,a\}$时,$Grs=\varnothing$,算法结束。故表 1 的属性约简为$\{b,d,a\}$。

5　结论

在基于正域的属性约简算法中,首先都要计算 U/C,因此设计出求 U/C 的低复杂度的算法是很有意义的。本文用粒计算的思想,把决策表分解成粒子空间,再对其求 Gr/C 的算法,使其复杂度降低为 $o(|C||U|)$。目前基于粒计算的属性约简算法最好的就是文献[5],其最差时间复杂度为:$o((|C||Gr_{pos}||Gr_{neg}|))$。本文以属性重要性为启发信息设计出一个时间复杂度为 $\max(o(|C||Gr|),o(|C^2||Gr/C|))$的基于粒子正域空间的属性约简算法。 实验证明,该算法不仅能有效地对决策表进行约简,而且具有高效性。

参考文献:

[1] Pedrycz W.Granular computing:an emerging paradigm[M].[S.l.]:Harecover/Physica-Verlag,2001.

[2] Pedrycz W.The paradigm and practice of granular computing[EB/OL].http://www.sfb360.uni-biele feld.de/mokolloq/abstracts/pedrycz.html.

[3] 刘清,刘群.粒及粒计算在逻辑推理中的应用[J].计算机研究与发展,2004,41(4):546-551.

[4] Yao Y Y.Granular computing:basic issue and possible solutions[C]//Proceedings of the 5th Joint Conference on Information Science,2000:186-189.

[5] 胡峰,代劲,蒋学文,等.粗糙集中的粒计算模型[J].计算机工程与设计,2006,20(27):3748-3749.

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

[7] 徐章艳,刘作鹏,杨炳儒,等.一个复杂度为 $\max(o(|C||U|),o(|C^2||U/C|))$的快速属性约简算法[J].计算机学报,2003,3(29):395-396.

[8] 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.

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

[10] 杜金莲,迟忠先,翟巍.基于属性重要性的逐步约简算法[J].小型微型计算机系统,2003,24(6):976-978.

[11] Yao J T,Yao Y Y.Induction of classification rules by granular computing[C]//Proceedings of the Third International Conference on Rough Sets and Current Trends in Computing,RSTC 2002,2002:331-338.

(上接 147 页)

[4] Lin Wei-yang,Alvarez S A,Ruiz C.Effective adaptive-support association rule mining for recommender systems[J].Data Mining and Knowledge Discovery,2002,6(1):83-105.

[5] Masseglia F,Poncelet P,Teisseire M.Web usage mining:how to efficiently manager new transactions and new clients[C]//Proc of 4th European Conf on Principles and Practice of Knowledge Discovery in Databases,2000:530-535.

[6] Schwarzkopf E.An adaptive Web site for the UM2001 conference[C]//Proceedings of the UM2001 Workshop on Machine Learning for User Modeling.Germany:Sonthofen,2001:77-86.

[7] Cooley R,Mobasher B,Srivastava J.Data preparation for mining World Wide Web browsing patterns[J].Knowledge and Information System,1999,1(1):5-32.

[8] Agrawal R,Srikant R.Fast algorithms for mining association rules[C]//Proceedings of the 20th International Conference on Very Large Databases,1994:487-499.

[9] Agarwal R C,Aggarwal C C,Prasad V V V.A tree projection algorithm for generation of frequent itemsets[J].Journal of Parallel and Distributed Computing,2000,61(3):350-371.

[10] Shahabi C,Zarkesh A,Adibi J,et al.Knowledge discovery from usersWeb-page navigation[C]//Workshop on Research Issues in Data Engineering,1997.