一种分布式的决策表核属性计算方法

官礼和
GUAN Li-he

重庆交通大学 信息与计算科学研究所,重庆 400074
Institute of Information and Calculation Science,Chongqing Jiaotong University,Chongqing 400074,China
E-mail:guanlihe@cquc.edu.cn

GUAN Li-he.Distributed calculation methods for core attributes of decision table.Computer Engineering and Applications,2007,43(17):164-166.

Abstract: The problem of calculating the core attributes of a decision table is studied.Some errors and limitations in some former results by Hu and Ye are analyzed.A new algorithm for calculating the core attributes of a decision table is also presented, which is based on a new definition of the discernibility matrix.Secondly,decision table is divided into a number of sub-decision tables based on the different attribution values of objects,and the relation of the core attributes between decision table and sub-decision tables is analyzed and established.A distributed calculation methods for core attributes of decision table is presented,and the experiment results about the methods show that the method is efficient and feasible in practice.

Key words: rough set;core attribute;decision table;discernibility matrix

摘 要:针对决策表核属性的计算问题,首先对前人的一些结论观点进行了讨论,在认识这些理论观点缺陷的基础上,给出了分明矩阵的一个新的表示定义,并由此提出了一种新的核属性计算方法。然后依据论域中各对象在某个条件属性上的不同取值把决策表信息系统划分为多个子决策表信息系统,给出了几条相关性质以及子决策表与原决策表核属性的关系定理。最后提出了一种决策表核属性的分布式计算方法,实例分析表明该算法是有效的。

关键词:粗糙集;核属性;决策表;分明矩阵
文章编号:1002-8331(2007)17-0164-03 **文献标识码**:A **中图分类号**:TP18

粗糙集理论自波兰学者 Pawlak 教授 1982 年提出以来,已经在机器学习、数据挖掘等领域中得到了广泛的应用,其理论体系以及与模糊集等理论的关系也得到了阐明[1,2],决策表信息系统是粗糙集理论的主要研究对象,决策表信息系统的约简是所有粗糙集理论和应用研究的焦点问题之一,目前已经提出了若干个求解属性约简的方法,在这些约简算法中,搜索一个属性约简通常是从核开始,决策表核属性的确定往往是信息约简的基础。

在求核问题上,众多学者进行了深入的研究,但是在研究过程中还是出现了不少问题。Hu 根据 Skowron 教授提出的分明矩阵得出了一个确定决策表信息系统核属性集的方法[3];叶东毅在文献[4]中对 Hu 的结论提出了质疑,举例证明了该结论的错误,并通过对分明矩阵的改进提出了一种确定核属性的方法。本文针对决策表核属性的计算问题,首先对前人的一些结论观点进行了讨论,在认识这些理论观点缺陷的基础上,给出了分明矩阵的一个更加简洁的表示定义,并由此得到基于分明矩阵的核属性计算方法。

目前很多研究者从不同的方面提出了很多决策表核属性的求解方法[4,5,7,8],但都是串行的,不是分布式的,因而对海量数据没有办法处理。文献[5]提出依据论域中各对象在某个条件属性上的不同取值把决策表信息系统划分为多个子决策表信息系统的思想,在文中借助此方法对所考虑的决策表作划分,得到几条相关性质和一个子决策表与原决策表信息系统核属性的关系定理,最后提出了一种分布式的决策表核属性计算方法,实例分析表明该算法是有效性的,对决策表的海量数据处理提供了一个很好的处理思路和方法。

1 基本概念

为了便于叙述,在这里只对粗糙集的几个概念作简要介绍,其他相关概念请参看文献[2]。

定义 1 (决策表信息系统)一个决策表信息系统 $S=<U,C,D,V,f>$,其中,U 是对象的集合,也称为论域,$C\cup D$ 是属性集,C 和 D 分别称为条件属性集和决策属性集,$D\neq\varnothing$,$V=\bigcup_{r\in R}V_r$ 是属性值的集合,V_r 表示属性 $r\in R$ 的属性值范围,即属性 r 的值域,$f:U\times R\to V$ 是一个信息函数,它指定 U 中每一个对象 x 的属性值。

定义 2 (不可分明关系)在决策表信息系统 S 中,对于每个属性子集 $B\in C\cup D$,可以定义一个不可分明关系 $IND(B)=\{(u,v)|\forall(u,v)\in U\times U,\forall b\in A,b(u)=b(v)\}$。$IND(B)$也是 U 上

基金项目:重庆交通学院高层次人才启动基金(No.2004-2-19)。
作者简介:官礼和(1975-),男,博士研究生,主要研究领域为人工智能、数据挖掘、Rough 集与知识获取等。

的一个等价关系。

定义 3（相对正域）设 U 为决策表信息系统 S 的论域，P、Q 为 U 上的两个等价关系簇，Q 的 P 正域定义为：$POS_P^S(Q)=\bigcup_{X\in U/Q}\underline{P}(X)$，其中 $\underline{P}(X)$ 为 X 的下近似集。

定义 4（属性约简）设 U 为决策表信息系统 S 的论域，P、Q 为 U 上的两个等价关系簇，若 $R\subseteq P$，$POS_R^S(Q)=POS_P^S(Q)$ 且对 $\forall r\in R$ 都有 $POS_{R-\{r\}}^S(Q)\neq POS_R^S(Q)$，则称 R 为 P 的 Q 约简。记 P 的所有 Q 约简集为 $RED_Q(P)$。

定义 5（核集）设 U 为一个论域，P、Q 为 U 上的两个等价关系簇，$RED_Q(P)$ 为 P 的所有 Q 约简集，则 P 的 Q 核属性集定义为：$CORE_Q(P)=\cap RED_Q(P)$。

定义 6（决策基数）在一个决策表信息系统 S 中，设 $d(x_i)=\{d(y)|y\in[x_i]_C\}$，则 $card(d(x_i))$ 表示 U 中所有与 x_i 在关系 C 下等价的元素相应的决策属性值构成的集合的基数，称为决策基数。

2 基于分明矩阵的决策表核属性的计算方法

Hu 对 Skowron 教授提出的分明矩阵重新作了如下定义[3]。

定义 7 令决策表信息系统为 $S=<U,C,D,V,f>$，$C=\{a_i|i=1,2,\cdots,m\}$ 和 $D=\{d\}$ 分别称为条件属性集和决策属性集，$U=\{x_1,x_2,\cdots,x_n\}$ 是论域，$a_i(x_j)$ 是样本 x_j 在属性 a_i 上的取值。$\boldsymbol{C}_D(i,j)$ 表示分明矩阵中第 i 行 j 列的元素，则分明矩阵 $\boldsymbol{C}_D$ 定义为：

$$\boldsymbol{C}_D(i,j)=\begin{cases}\{a_k|a_kC\wedge a_k(x_i)\neq a_k(x_j)\} & d(x_i)\neq d(x_j)\\ 0 & d(x_i)=d(x_j)\end{cases}\quad(1)$$

其中，$i,j=1,2,\cdots,n$

文献[3]得出了如下结论：当且仅当某个 $C_D(i,j)$ 为单属性集合时，该属性属于核 $CORE_D(C)$。

叶东毅在文献[4]中对 Hu 的这个结论提出了质疑，举例证明了该结论的问题，并通过对分明矩阵的改进提出了一种计算核属性的方法。叶东毅改进的分明矩阵定义为：

定义 8 给定决策表信息系统 $S=<U,C,D,V,f>$，分明矩阵 $\boldsymbol{C'}_D$ 的元素 $\boldsymbol{C'}_D(i,j)$ 定义为：

$$\boldsymbol{C'}_D(i,j)=\begin{cases}\boldsymbol{C}_D(i,j) & \min\{|\boldsymbol{D}(x_i)|,|\boldsymbol{D}(x_j)|\}=1\\ 0 & \text{else}\end{cases}\quad(2)$$

其中，$\boldsymbol{D}(x_i)=\{d_r|d_r\in[x_i]_C\}$。

文献[4]得出了结论：当且仅当某个 $\boldsymbol{C'}_D(i,j)$ 为单属性集合时，该属性属于核 $CORE_D(\boldsymbol{C})$。这种方法得到的是在代数观中的核属性[6]，在此基础上，文献[7]指出 Hu 的求核算法错误的根源在于认为去掉某属性后产生新的不确定性记录就断定其产生了新的冲突，而忽略了这个新的不确定性记录在去掉属性之前是否为不确定性记录，并由此提出了基于合并规则的决策表求核方法。

为了能通过分明矩阵更方便地求出决策表的核属性集，对前面这几种分明矩阵的定义进行了分析研究，下面给出从决策表信息系统 S 的正域 $POS_C^S(\boldsymbol{D})$ 来定义 S 的分明矩阵的 M。

定义 9 一个决策表信息系统 $S=<U,C,D,V,f>$，S 的分明矩阵 M 是一个 $n\times n$ 的矩阵，其定义如下：

$$m_{uv}^S=\begin{cases}\{a|a\in C\wedge a(u)\neq a(v)\} & d(u)\neq d(v)\wedge w_{uv}\\ 0 & \text{else}\end{cases}\quad(3)$$

其中，$w_{uv}=1$ 当且仅当 $u\in POS_C^S(D)\vee v\in POS_C^S(D)$。记 M 中所有非空元素组成的集合为 M_S，即 $M_S=\{m_{uv}^S|\forall m_{uv}^S\in M,m_{uv}^S\neq\phi\}$。易看出式(2)与式(3)定义是等价的，因为式(2)中要求 $\min\{|D(x_i)|,|D(x_j)|\}=1$，实际就是要求 x_i,x_j 至少有一个属于 $POS_C^S(D)$。这样当且仅当某个 m_{uv}^S 为单属性集合时，该属性属于核 $CORE_D(C)$，于是可得到一种新的基于分明矩阵求核算法。

算法 1

输入：一个决策表 $S=<U,C,D,V,f>$，其中 U 为论域，C、D 分别为条件属性集和决策属性集。

输出：该决策表的核 $CORE_D(\boldsymbol{C})$。

步骤 1 计算出 $POS_C^S(D)$；

步骤 2 根据式(3)，先求出决策表的分明矩阵 $\boldsymbol{M}$；

步骤 3 求分明矩阵的所有只包含单个属性的元素 m_{ij} 的并集，即为决策表的核属性集 $CORE_D(\boldsymbol{C})$。

3 基于 $U/\{a\}$ 划分的决策表核属性集求解方法

为了实现对海量数据的处理，能分布式地求取决策表信息系统的核属性集，借助文献[5]中对系统作划分的思想，依据论域中各对象在某个条件属性上的不同取值把决策表信息系统划分为多个子决策表信息系统。下面将探讨各个子系统的核属性集与原系统的核属性集的关系。

3.1 子决策表信息系统的划分

定义 10（子决策表划分）给定决策表信息系统 $S=<U,C,D,V,f>$，任取 $a\in C$，根据 a 对 U 进行等价划分，即 $U/\{a\}=\{U_1,U_2,\cdots,U_n\}$，其中 $n=card(V_a)$，于是得到 n 个子系统 $S_i=<U_i,C,D,V,f>$，$(i=1,2,\cdots,n)$。

对子系统 S_i，记分明矩阵为 $\boldsymbol{M}_i$，$\boldsymbol{M}_i$ 中非空元素组成的集合为 $\boldsymbol{M}_{Si}$，约简集为 $RED(S_i)=\{R|R\subseteq C,R$ 是 S_i 的约简$\}$，相对核属性集记为 $CORE(S_i)$。

性质 1 $POS_C^{S_i}(D)=POS_C^S(D)\cap U_i$，$(i=1,2,\cdots,card(V_a))$

证：①$\forall u\in POS_C^{S_i}(D)$，显然有 $u\in U_i$，如果 $u\notin POS_C^S(D)$，则 $\exists v\in U$ 使得 $(u,v)\in IND(C)$ 且 $(u,v)\notin IND(D)$；根据子系统的划分的定义知 $v\in U_i$，所以 $u\notin POS_C^{S_i}(D)$，矛盾。故 $POS_C^{S_i}(D)\subseteq POS_C^S(D)\cap U_i$，$(i=1,2,\cdots,card(V_a))$。

②$\forall u\in POS_C^S(D)\cap U_i$，有 $u\in U_i$，且 $u\in POS_C^S(D)$；如果 $u\notin POS_C^{S_i}(D)$，则 $\exists v\in U_i$，使得 $(u,v)\in IND(C)$ 且 $(u,v)\notin IND(D)$，又 $U_i\subseteq U$，所以 $v\in U$，从而 $u\notin POS_C^S(D)$，矛盾。故 $POS_C^S(D)\cap U_i\subseteq POS_C^{S_i}(D)$，$(i=1,2,\cdots,card(V_a))$

综合①、②知 $POS_C^{S_i}(D)=POS_C^S(D)\cap U_i$，$(i=1,2,\cdots,card(V_a))$。

性质 2 $POS_C^S(D)=\bigcup_{1\leqslant i\leqslant card(V_a)}POS_{R_i}^{S_i}(D)$，其中 R_i 是子系统 S_i 的一个约简。

证：因为 R_i 是子系统 S_i 的一个约简，所以 $POS_{R_i}^{S_i}(D)=POS_C^{S_i}(D)$，根据性质 1 可得，$POS_{R_i}^{S_i}(D)=POS_C^S(D)\cap U_i$，$(i=1,2,\cdots,$

$card(V_a)$)),两边取并得,$\bigcup_{1\leqslant i\leqslant card(V_a)} POS_{R_i}^{S_i}(D)=\bigcup_{1\leqslant i\leqslant card(V_a)} POS_C^S(D)\cap U_i$),根据 $U=\bigcup_{1\leqslant i\leqslant card(V_a)} U_i$ 对上式化解得 $POS_C^S(D)=\bigcup_{1\leqslant i\leqslant card(V_a)} POS_{R_i}^{S_i}(D)$。

下面的性质 3 和定理 1 对应于文献[5]中的引理 2 和引理 3,由于文献[5]中的证明比较繁琐,我们在这里给出简洁的证明过程。

性质 3[5] 如果 u,v 属于同一子系统 S_k,则 $m_{uv}^S=m_{uv}^{S_k}$。

证:由性质 1 和分明矩阵的定义式(3)即可得证。

定理 1[5] $\forall R_S\in RED(S)$,在每个子系统 S_i 中 $\exists R\in RED(S_i)$,满足 $R\subseteq R_S$。

证:$\forall R_S\in RED(S)$,即 $POS_{R_S}^S(D)=POS_C^S(D)$,又由性质 1 可得 $POS_C^{S_i}(D)=POS_{R_S}^S(D)\cap U_i$,任取 $u\in POS_C^{S_i}(D)$有 $u\in POS_{R_S}^S(D)$且$u\in U_i$,说明在 U 中与 u 在 R_S 上取值完全相同的对象在 D 上的取值也完全相同,又因为 $U_i\subseteq U$,所以在 U_i 中与 u 在 R_S 上取值完全相同的对象在 D 上的取值也完全相同,即 $u\in POS_{R_S}^{S_i}(D)$,于是有 $POS_C^{S_i}(D)\subseteq POS_{R_S}^{S_i}(D)$;另外由于 $R_S\subseteq C$,故显然有 $POS_{R_S}^{S_i}(D)\subseteq POS_C^{S_i}(D)$。所以就有 $POS_C^{S_i}(D)=POS_{R_S}^{S_i}(D)$,即说明由 R_S 出发一定能够得到 S_i 的一个约简 R,且 $R\subseteq R_S$。

定理 2 设 $Core(S_i)$为子系统 S_i 的核属性集,$CORE(S)$为原系统 S 的核属性集,属性 a 是由S 划分为子系统时所选取的属性,则有

$$CORE(S)=\begin{cases}\bigcup_{1\leqslant i\leqslant card(V_a)} CORE(S_i), & a\notin CORE(S)\\ (\bigcup_{1\leqslant i\leqslant card(V_a)} CORE(S_i))\cup\{a\}, & a\in CORE(S)\end{cases}\tag{4}$$

证:① $\forall b\in CORE(S)$,则根据分明矩阵定义式(3)得出的结论可知在 S 中一定存在 $u,v\in U$,使得 $m_{uv}^S=\{b\}$。如果 u,v 同属于某个子系统 S_i 中,则根据定理 1 知 $m_{uv}^{S_i}=m_{uv}^S$,此时 $b\in CORE(S_i)$;如果 u,v 分属于不同的子系统中就有 $m_{uv}^S=\{a,b\}$,此时因为 b 是 S 中的核属性,故必定有 $b=a$。所以有

$$CORE(S)\subseteq\begin{cases}\bigcup_{1\leqslant i\leqslant card(V_a)} CORE(S_i), & a\notin CORE(S)\\ (\bigcup_{1\leqslant i\leqslant card(V_a)} CORE(S_i))\cup\{a\}, & a\in CORE(S)\end{cases}$$

②任取子系统 S_i,对 $\forall b\in CORE(S_i)$,则有 $\exists u,v\in U_i$ 使得 $m_{uv}^{S_i}=\{b\}$,此时 u,v 同属于子系统 S_i 中,根据定理 1 可知在 S 中有 $m_{uv}^S=m_{uv}^{S_i}$,即 $m_{uv}^S=\{b\}$,从而 $b\in CORE(S)$。

当 $a\notin CORE(S)$时有,$\bigcup_{1\leqslant i\leqslant card(V_a)} CORE(S_i)\subseteq CORE(S)$;

当 $a\in CORE(S)$时有,$(\bigcup_{1\leqslant i\leqslant card(V_a)} CORE(S_i))\cup\{a\}\subseteq CORE(S)$。

综上所述,定理得证。

3.2 分布式核属性的求解方法

由定理 2 可知,原决策表信息系统的核属性集可通过选取某个属性将系统作划分,在分别求得各个子系统的核属性集之后即可求得。故有如下求核属性集的分布式算法。

算法 2

输入:一个决策表 $S=<U,C,D,V,f>$,其中 U 为论域,C、D 分别为条件属性集和决策属性集。

输出:该决策表的核 $CORE(S)$。

步骤 1 在 C 中选取一个适当的属性 a,根据 a 的属性值对原系统 S 作等价划分:$U/\{a\}=\{U_1,U_2,\cdots,U_n\}$,其中 $n=card(V_a)$,得到 n 个子系统 $S_i=<U_i,C,D,V,f>$,$(i=1,2,\cdots,n)$;

步骤 2 对每个子系统 S_i,通过算法 1 求得其核属性集 $CORE(S_i)$;

步骤 3 根据 a 是否属于 $CORE(S)$,由式(4)即可求得决策表信息系统 S 的核属性集 $CORE(S)$。

4 实例分析

现举一气象状况实例作为决策表信息系统,如表 1 用分布式的方法来求解该决策表信息系统的核属性集:

表 1 一个决策表信息系统

U	condition Attribute(C)				Decision
	Outlook(a_1)	Temperature(a_2)	Humidity(a_3)	Windy(a_4)	Attribute(D)
1	Sunny	Hot	High	False	N
2	Sunny	Hot	High	True	N
3	Overcast	Hot	High	False	P
4	Rain	Mild	High	False	P
5	Rain	Cool	Normal	False	P
6	Rain	Cool	Normal	True	N
7	Overcast	Cool	Normal	True	P
8	Sunny	Mild	High	False	N
9	Sunny	Cool	Normal	False	P
10	Rain	Mild	Normal	False	P
11	Sunny	Mild	Normal	True	P
12	Overcast	Mild	High	True	P
13	Overcast	Hot	Normal	False	P
14	Rain	Mild	High	True	N

首先选择一个条件属性,对决策表进行划分。这里就选第一个属性 a_1,$U/\{a_1\}=\{U_1,U_2,U_3\}$,其中 $U_1=\{1,2,8,9,11\}$、$U_2=\{3,7,12,13\}$、$U_3=\{4,5,6,10,14\}$,于是得到三个子决策表信息系统 $S_i=<U_i,C-\{a_1\},D,V,f>$,$(i=1,2,3)$。

其次对每个子系统用式(3)求得各自的分明矩阵 $\boldsymbol{M}_i$,$\boldsymbol{M}_i$ 中非空元素组成的集合记为 $\boldsymbol{M}_i'$,$\boldsymbol{M}_1'=\{a_2a_3,a_2a_3a_4\}$,$\boldsymbol{M}_2'=\varnothing$,$\boldsymbol{M}_3'=\{a_2a_4,a_3a_4,a_2a_3a_4,a_4\}$,所以各个子系统的核属性集分别为:$RED(S_1)=\varnothing$,$RED(S_2)=\varnothing$,$RED(S_3)=\{a_4\}$。

由于 $POS_{C-\{a_1\}}^S(D)\neq POS_C^S(D)$,故所选划分属性 a_1 是原决策表信息系统的核属性,所以原决策表信息系统的核属性集为 $RED(S)=RED(S_1)\cup RED(S_2)\cup RED(S_3)\cup\{a_1\}$,即 $RED(S)=\{a_1,a_4\}$。

6 结论

针对决策表核属性的计算问题,对前人的一些结论观点进行了讨论,在认识这些理论观点缺陷的基础上,给出了分明矩阵的一个新的简洁表示定义,并由此提出了相应的核属性计算方法。在参阅了文献[5]中依据论域中各对象在某个条件属性上的不同取值把决策表信息系统划分为多个子系统的基础上,对

(下转 173 页)