文章编号: 1001-0920(2008)02-0127-06

一种基于水平分布的多决策表全局属性核求解算法

杨 明, 吴永芬
(南京师范大学 计算机科学系, 南京 210097)

摘 要: 针对全局决策表一致和不一致情况,探讨了全局属性核与局部属性核之间的关系,提出一种基于水平分布的多决策表全局属性核求解算法. 该算法充分利用各局部属性核,且通过传送压缩的局部决策表可有效地减少网络通讯量,从而提高全局属性核求解的效率. 算法分析结果表明,该算法是有效且可行的.

关键词: 粗糙集; 差别矩阵; 全局属性核; 局部属性核; 多决策表
中图分类号: TP311 **文献标识码**: A

An algorithm based on horizontal partitioned multi-decision table for computing global attribute core

YANG Ming, WU Yong-fen
(Department of Computer Science, Nanjing Normal University, Nanjing 210097, China. Correspondent: YANG Ming, E-mail: m.yang@njnu.edu.cn)

Abstract: For the condition that global attribute core is consistent and inconsistent, the relationships between the global attribute core and local attribute core are discussed, and an algorithm based on horizontal partitioned multi-decision table for computing global attribute core is presented. By using every local attribute core at each site transmitting every local decision table which is shrunk, network communication is reduced. Hence, the efficiency of computing global attribute core is improved. Algorithm analysis results show the effectiveness and feasibility of the algorithm.

Key words: Rough set; Discernibility matrix; Global attribute core; Local attribute core; Multi-decision table

1 引 言

粗糙集是一种新的处理不精确、不完全与不相容知识的数学理论[1],近年来在机器学习、数据挖掘及网络入侵检测等多个领域得到了广泛的应用[2,3],但在分布式环境下的应用才刚刚开始. 虽然粗糙集理论在网络入侵检测中已有成功的应用[4],但该应用仅针对单个主机,而针对分布式环境下的网络入侵检测研究却未见报道. 然而在现实应用中,很多决策应用问题都要通过分布式环境下的各参与方的协同工作才能完成[5-8]. 可见,进行分布式环境下的粗糙集理论研究可有效地扩展粗糙集理论的应用,具有重要的理论和现实意义.

在粗糙集理论研究中,属性约简是重要研究内容之一[9,10],也是知识获取的关键步骤. 很多经典的属性约简算法都是从属性核开始的,因此,求解分布式环境的属性核已成为分布式环境下粗糙集理论研究的重要内容之一.

尽管已有很多针对单个决策表的高效属性核求解算法[11-14],但有关分布式环境下的多个决策表的属性核求解算法还不多. 解决分布式环境下属性核求解的一种简单方法是由各局部决策表形成一个全局决策表,然后直接应用现有的属性核求解算法. 该方法的效率很低. 为此,本文针对全局决策表一致和不一致情况,探讨全局属性核与局部属性核之间的关系,提出一种基于水平分布的多决策表全局属性核求解算法. 该算法充分利用各局部属性核,通过传送压缩的局部决策表有效减少网络通讯量,进而提高全局属性核求解的效率. 分析结果表明了算法的可行性.

收稿日期: 2006-09-11; **修回日期**: 2006-12-25.
基金项目: 国家自然科学基金项目(70371015); 江苏省自然科学基金项目(BK2005135); 江苏省高校自然科学研究项目(05KJB5200665).
作者简介: 杨明 (1964—),男,安徽宁国人,教授,博士,从事数据挖掘、机器学习等研究; 吴永芬(1979—),女,江苏宿迁人,硕士生,从事粗糙集理论与应用的研究.

2 相关概念及结论

在本文模型中,多个决策表是水平划分的,且各部分的决策表模式逻辑同构. 设有 t 个站点 $S_1, S_2, \cdots, S_t$,相应的成员决策表(或局部决策表)的对象集合分别为 $U_1, U_2, \cdots, U_t$, $U = \bigcup_{i=1}^{t} U_i$. 有关粗糙集的基本概念介绍如下,更详细的内容可见文献[1].

定义 1 全局决策表 DT 是一个 4 元组 $\langle U, C \cup D, V, f\rangle$. 其中:$U$ 是一组对象的非空有限集合,称为论域,设有 n 个对象,则 U 可表示为 $U = \{x_1, x_2, \cdots, x_n\}$;$C$ 为条件属性集;D 为决策属性集;$V = \bigcup_{a\in(C\cup D)} V_a$, V_a 为属性 a 的值域集;f 是 $U \times (C \cup D) \to V$ 的映射.

为便于叙述,用 $|\cdot|$ 表示集合的基,用 $\varnothing$ 表示空集. 不失一般性,假设仅有一个决策属性 D,其取值范围是 $1, 2, \cdots, k$. 由 D 导出的等价类构成 U 的一个划分:$\{\psi_1, \psi_2, \cdots, \psi_k\}$,其中 $\psi_i = \{x \in U : f(x, D) = i\}$, $i = 1, 2, \cdots, k$.

定义 2 在站点 S_i $(i = 1, 2, \cdots, t)$,局部决策表 DT_i 是一个 4 元组 $\langle U_i, C \cup D, V, f\rangle$. 其中:$U_i$ 有 n_i 个对象,且 $U_i = \{y_1, y_2, \cdots, y_{n_i}\}$;$C$ 为条件属性集;D 为决策属性集;$V = \bigcup_{a\in(C\cup D)} V_a$, V_a 为属性 a 的值域集;f 是 $U_i \times (C \cup D) \to V$ 的映射.

在全局决策表或局部决策表中,若一些对象具有相同的条件属性而具有不同的分类,则称这类对象是全局不一致的或局部不一致的;否则为全局一致的或局部一致的. 若 U 中的两个不同对象 x 和 y 具有相同的条件属性而具有不同的分类,则称 x 和 y 为全局不一致的;否则称 x 和 y 为全局一致的. 若在站点 S_i, U_i 中的两个不同对象 x 和 y 具有相同的条件属性而具有不同的分类,则称 x 和 y 为局部不一致的;否则称 x 和 y 为局部一致的.

本文称所有对象均全局一致的全局决策表为全局一致决策表,否则为全局不一致决策表;称所有对象均局部一致的局部决策表为局部一致决策表,否则称为局部不一致决策表.

定义 3 设 $X \subseteq U (X \subseteq U_i)$ 为论域的一个子集, $P \subseteq C$, X 关于 P 的全局下近似和在站点 S_i 的局部下近似分别为

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

$$\underline{P}X(U_i) = \{x \in U_i : [x]_{(P,U_i)} \subseteq X\}.$$

其中

$$[x]_{(P,U)} = \{y \in U \mid \forall a \in P, f(x, a) = f(y, a)\},$$

$$[x]_{(P,U_i)} = \{y \in U_i \mid \forall a \in P, f(x, a) = f(y, a)\}.$$

定义 4 设 $P \subseteq C$,若 $\gamma_{(P,U)} = \gamma_{(C,U)}$,且不存在 $R \subset P$,使得 $\gamma_{(R,U)} = \gamma_{(C,U)}$,则称 P 为 C 的一个(相对于决策属性 D 的)全局属性约简. 所有 C 的全局属性约简的交称为 C 的全局属性核,记为 Core(C, U). 同理,可定义站点 S_i 的局部属性约简和局部属性核 Core(C, U_i).

性质 1 若在某个局部站点 S_i, DT_i 是局部不一致的,则 DT 是全局不一致的;反之,结论不成立.

性质 2 若 DT 是全局一致的,则在任意站点 S_i $(i = 1, 2, \cdots, t)$ 上,DT_i 均是局部一致的;反之,结论不成立.

3 基于差别矩阵的单决策表属性核求解

对于单个决策表,即本文中的局部决策表数 t 为 1. 类似文献[11]的定理 2,可通过下列方法有效地求得全局属性核(单个决策表情况下,全局属性核与局部属性核一致).

定义 5 对于给定的单个决策表 DT,定义差别矩阵 $M_1 = \{m_{ij}\}$,

$$m_{ij} = \begin{cases} \{a_1 \in C : f(x_i, a) \neq f(x_j, a)\}, \\ \quad x_i \in U_1, x_j \in U_1, f(x_i, D) \neq f(x_j, D); \\ \{a \in C : f(x_i, a) \neq f(x_j, a)\}, \\ \quad x_i \in U_1, x_j \in U_2', \varnothing(\text{空集}), \text{其他}. \end{cases} \tag{1}$$

其中:$U_1 = \bigcup_{i=1}^{k} \underline{a}\psi_i(U)$, $U_2 = U - U_1$, $U_2' = \{y \in U_2 \mid$ 不存在 $x \in U_2$ 使得 $f(x, a) = f(y, a)$ 且 $f(x, D) \neq f(y, D)$, $\forall a \in C\}$.

定理 1[13] 对于单个决策表 DT,若记 IDM$(C, M_1) = \{m_{ij} \mid m_{ij} \in M_1$ 且 m_{ij} 为单个属性$\}$,则有 IDM(C, M_1) = Core(C, U),即当且仅当某个 m_{ij} 为单个属性时,该属性属于核 Core(C, U).

定义 5 和定理 1 为求解单个决策表的属性核提供了有效的框架,即仅适用一个局部站点的情况. 对多个局部站点的情况,即对全局属性核求解则是本文的主要内容.

4 全局属性核求解算法

4.1 全局一致决策表的属性核求解算法

在分布式环境下,利用各站点的计算能力以及减少站点之间的网络通讯量是提高算法效率的关键. 因此,充分利用各站点的局部属性核,传送压缩的各局部决策表,并以此减少网络通讯量是本文讨论的重点.

定理 2 若 DT 是全局一致的,则任意站点 S_i 上的局部属性核 Core$(C, U_i) \subseteq$ Core(C, U).

证明　由性质 2 即可得证. □

推论 1　若 DT 是全局一致的,则

$$\bigcup_{i=1}^{t}\mathrm{Core}(C,U_i)\subseteq\mathrm{Core}(C,U).$$

对全局一致决策表,如何应用推论 1 求解全局属性核,且有效地减少网络通讯量是本文的关键任务. 若令 $DT_i^1=\langle U_i^1,C_1\cup D,V,f\rangle$,其中:$C_1=C-\bigcup_{j=1}^{t}\mathrm{Core}(C,U_j)$,$U_i^1=\{x\mid \exists y\in U_i$ 使得 $f(x,b)=f(y,b),b\in(C_1\cup D)\}\cup\{\{x,y\}\mid$ 存在 $y\in U_i,f(x,b)=f(y,b),b\in(C_1\cup D)\}$(见例 1 说明). 记 $C\mathrm{Core}(C_1,U_j^1,U_k^1)$ 为

$\{a\in C_1\mid \exists x\in U_j^1,\ y\in U_k^1$ 使得

$\forall b\in(C_1-\{a\})$ 有 $f(x,b)=f(y,b)$,

$f(x,a)\neq f(x,a)$ 且 $f(x,D)\neq f(y,D)\}$.

定理 3　若 DT 是全局一致的,则

$$\mathrm{Core}(C,U)\subseteq\bigcup_{i=1}^{t}\mathrm{Core}(C,U_i)\cup\Big(\bigcup_{1\le j\neq k\le t}C\mathrm{Core}(C_1,U_j^1,U_k^1)\Big).$$

证明　$\forall a\in\mathrm{Core}(C,U)$,即 $\exists x,y\in U$ 使得 $\forall b\in(C-\{a\})$ 有 $f(x,b)=f(y,b),f(x,a)\neq f(y,a)$ 且 $f(x,D)\neq f(y,D)$ 成立. 若 $\exists U_i\subseteq U$ 使得 $x,y\in U_i$,则 $a\in\mathrm{Core}(C,U_i)$;否则,$a\notin\bigcup_{i=1}^{t}\mathrm{Core}(C,U_i)$,也即 $a\in C-\bigcup_{i=1}^{t}\mathrm{Core}(C,U_i)$. 而 $U=\bigcup_{i=1}^{t}U_i$,因此 $\exists U_j,U_k,j\neq k$ 使得 $x\in U_j,y\in U_k$,使得对 $\forall b\in(C-\{a\})$ 有 $f(x,b)=f(y,b),f(x,a)\neq f(y,a)$ 且 $f(x,D)\neq f(y,D)$. 故 $\forall b\in(C_1-\{a\})$ 有 $f(x,b)=f(y,b),f(x,a)\neq f(y,a)$ 且 $f(x,D)\neq f(y,D)$,即 $a\in C\mathrm{Core}(C_1,U_j^1,U_k^1)$. □

由定理 3,只要对 $C\mathrm{Core}(C_1,U_j^1,U_k^1)\ (1\le j\neq k\le t)$ 中的属性进行修剪便可求得全局属性核. 于是,对全局一致决策表 DT,依据定理 3 可得如下的全局属性核求解算法 GARBC.

算法 1(GARBC)

输入:有 t 个分布式站点 $S_1,S_2,\cdots,S_t$,各站点上的局部决策表分别为 $\mathrm{DT}_i=\langle U_i,C\cup D,V,f\rangle$,其中 $i=1,2,\cdots,t$;全局决策表为 $\mathrm{DT}=\langle U,C\cup D,V,f\rangle$,其中 $U=\bigcup_{i=1}^{t}U_i$.

输出:全局属性核 $\mathrm{Core}(C,U)$.

1) 在各站点 S_i 上求解局部决策表 DT_i 的局部属性核 $\mathrm{Core}(C,U_i)$;

2) 压缩各局部决策表 DT_i,得到各浓缩的局部决策表 DT_i^1;

3) 站点之间传送浓缩的决策表 DT_i^1,求 $C\mathrm{Core}(C_1,U_j^1,U_k^1)\ (1\le j\neq k\le t)$,并记录使得 $a\in C\mathrm{Core}(C_1,U_j^1,U_k^1)$ 的对象标识对 (x,y);

4) 修剪 $C\mathrm{Core}(C_1,U_j^1,U_k^1)$,即判断使得 $a\in C\mathrm{Core}(C_1,U_j^1,U_k^1)$ 的对象标识对 (x,y),对任意 $b\in\bigcup_{j=1}^{t}\mathrm{Core}(C,U_j)$,是否有 $f(x,b)=f(y,b)$ 成立,若没有这样的 (x,y),则从 $C\mathrm{Core}(C_1,U_j^1,U_k^1)$ 中删除 a;

5) 返回全局属性核

$$\mathrm{Core}(C,U)=\bigcup_{i=1}^{t}\mathrm{Core}(C,U_i)\cup\Big(\bigcup_{1\le j\neq k\le t}C\mathrm{Core}(C_1,U_j^1,U_k^1)\Big).$$

例 1　表 1 为 2 个局部决策表 DT_1 和 DT_2,每个决策表均有 4 个元素和 5 个属性,$C=\{C_1,C_2,C_3,C_4\}$ 为条件属性集,D 为决策属性. 由算法 1 可知,$\mathrm{Core}(C,U_1)=\{C_4\}$,$\mathrm{Core}(C,U_2)=\{C_3\}$. 通过对决策表 DT_1 和 DT_2 进行压缩可得到压缩的决策表 DT_1^1 和 DT_2^1 如表 2 所示,并可求出 $C\mathrm{Core}(C_1,U_1^1,U_2^1)=\{C_2\}$,与属性 C_2 相对应的对象对分别为 $(x_4,x_7),(x_2,x_8)$. 可以验证,尽管对象 x_2 和 x_8 在属性 C_3 和 C_4 上的值均不相等,但对象 x_4 和 x_7 在属性 C_3 和 C_4 上的值相等,即 C_2 不从 $C\mathrm{Core}(C_1,U_1^1,U_2^1)$ 中删除. 因而,$\mathrm{Core}(C,U)=\{C_2,C_3,C_4\}$.

表 1　局部决策表 DT_1 和 DT_2

对象	站点 S_1					对象	站点 S_2				
	C_1	C_2	C_3	C_4	D		C_1	C_2	C_3	C_4	D
x_1	0	0	1	2	2	x_5	0	0	1	0	1
x_2	0	0	0	1	1	x_6	0	0	2	0	2
x_3	0	0	0	2	2	x_7	1	2	0	0	1
x_4	1	1	0	0	2	x_8	0	2	2	0	2

表 2　压缩的局部决策表 DT_1^1 和 DT_2^1

对象	站点 S_1			对象	站点 S_2		
	C_1	C_2	D		C_1	C_2	D
$\{x_1,x_3\}$	0	0	2	x_5	0	0	1
x_2	0	0	1	x_6	0	0	2
x_4	1	1	2	x_7	1	2	1
				x_8	0	2	2

从例 1 可知,算法 1 充分利用了各局部属性核,有效地对局部决策表进行了压缩传送,因而网络通讯量可有效降低,全局属性核的求解效率得到有效改进.

性质 3　对给定的 t 个站点 $S_i\ (i=1,2,\cdots,t)$,

若 $\bigcup_{j=1}^{t}\mathrm{Core}(C,U_j)\neq\varnothing$,则在站点 $S_i(i=1,2,\cdots,t)$ 上均有 $|U_i^1|\leqslant|U_i|$.

性质 4 对给定的 t 个站点 $S_i(i=1,2,\cdots,t)$,采用现有的单决策表属性核求解算法,空间复杂度为 $O\Big(|C|\Big(\sum_{i=1}^{t}|U_i|\Big)^2\Big)$. 算法 1 的空间复杂度为

$$O\Big(|C|\sum_{i=1}^{t}|U_i|^2+|C_1|\sum_{1\leqslant i,j\leqslant t}|U_i^1|*|U_j^1|\Big),$$

其中

$$C_1=C-\bigcup_{j=1}^{t}\mathrm{Core}(C,U_j).$$

由性质 4 可知,与用现有求解全局属性核的单机算法的空间复杂度相比,算法 1 可有效降低空间复杂度. 此外,算法 1 可并行地求出各局部决策表的局部属性核,且传送的是压缩决策表,因而可提高全局属性核的求解效率.

4.2 全局不一致决策表的属性核求解算法

因定理 2 和定理 3 对全局不一致决策表并不成立,故算法 GARBC 仅适用于全局一致决策表的全局属性核求解,这一点可由下面的例 2 得到验证.

例 2 表 3 为 2 个局部决策表 DT_1 和 DT_2,每个决策表均有 4 个元素和 5 个属性,$C=\{C_1,C_2,C_3,C_4\}$ 为条件属性集,D 为决策属性. 由算法 1 可知,$\mathrm{Core}(C,U_1)=\{C_3\}$,$\mathrm{Core}(C,U_2)=\{C_3\}$. 而通过单决策表属性核求解算法可求出 $\mathrm{Core}(C,U)=\varnothing$.

表 3 局部决策表 DT_1 和 DT_2

对象	站点 S_1					对象	站点 S_2				
	C_1	C_2	C_3	C_4	D		C_1	C_2	C_3	C_4	D
x_1	1	0	1	2	1	x_3	1	0	0	2	1
x_2	1	0	0	2	2	x_4	1	0	1	2	2

为有效求解全局一致决策表的全局属性核,类似文献[11]可得下面的引理 1 和引理 2.

引理 1 若 $x\in U_i(1\leqslant i\leqslant t)$,且 x 是全局不一致对象,则从 $U_j(1\leqslant j\neq i\leqslant t)$ 中删除 x,全局属性核保持不变.

引理 2 若 $U_i=U_{i1}\cup U_{i2}(i=1,2,\cdots,t)$,其中 U_{i1} 中的对象是全局一致的,而 U_{i2} 中的对象是全局不一致的,则对任意 U_{s2} 和 $U_{t2}(s\neq t)$,令 $U_{s2}=U_{s2}-(U_{s2}\cap U_{t2})$,全局属性核保持不变.

定理 4 若 $U_i=U_{i1}\cup U_{i2}(i=1,2,\cdots,t)$,其中 U_{i1} 中的对象是全局一致的,而 U_{i2} 中的对象是全局不一致的,则 $\mathrm{Core}(C,U_i)\subseteq\mathrm{Core}(C,U)$.

推论 2 若 $U_i=U_{i1}\cup U_{i2}(i=1,2,\cdots,t)$,其中 U_{i1} 中的对象是全局一致的,而 U_{i2} 中的对象是全局不一致的,则 $\bigcup_{i=1}^{t}\mathrm{Core}(C,U_i)\subseteq\mathrm{Core}(C,U)$.

注 1 定理 4 和推论 2 中的 $\mathrm{Core}(C,U_i)$ 是指将 U_i 分为全局一致对象集合 U_{i1} 和全局不一致对象集合 U_{i2},并将 U_{i1} 和 U_{i2} 看成定义 5 中 U_1 和 U_2 求得的.

下面将利用推论 2 求解全局属性核. 令 $\mathrm{DT}_i^1=\langle U_{i1}^1\cup U_{i2}^1,C_1\cup D,V,f\rangle$,其中:$U_{ij}^1=\{x\mid\exists y\in U_{ij}$ 使 $f(x,b)=f(y,b),b\in(C_1\cup D)\}\cup\{\{x,y\}\mid\exists y\in U_{ij}f(x,b)=f(y,b),b\in(C_1\cup D)\},j=1,2$;$C_1=C-\bigcup_{j=1}^{t}\mathrm{Core}(C,U_j)$,记 $C\mathrm{Core}(C_1,U_{ji}^1,U_{kl}^1)(i\neq2$ 或 $l\neq2)$ 为 $\{a\in C_1\mid\exists x\in U_{ji}^1,y\in U_{kl}^1$ 使得 $\forall b\in(C_1-\{a\})$,有 $f(x,b)=f(y,b),f(x,a)\neq f(x,a)$ 且 $f(x,D)\neq f(y,D)\}$.

定理 5 若 DT 是全局不一致的,则有

$$\mathrm{Core}(C,U)\subseteq\bigcup_{i=1}^{t}\mathrm{Core}(C,U_i)\cup\Big(\bigcup_{1\leqslant j\neq k\leqslant t,\,i\neq2\text{或}l\neq2}C\mathrm{Core}(C_1,U_{ji}^1,U_{kl}^1)\Big).$$

证明 类似于定理 3 的证明即可得证. □

由定理 5,只要对 $C\mathrm{Core}(C_1,U_{ji}^1,U_{kl}^1)(1\leqslant j\neq k\leqslant t,i$ 和 l 不同时为 2) 中的属性进行修剪便可求得全局属性核,因而可得如下基于全局不一致多决策表的全局属性核求解算法 GARBI.

算法 2(GARBI)

输入:有 t 个分布式站点 $S_1,S_2,\cdots,S_t$,各站点上的局部决策表分别为 $\mathrm{DT}_i=\langle U_i,C\cup D,V,f\rangle$,其中 $i=1,2,\cdots,t$;全局决策表为 $\mathrm{DT}=\langle U,C\cup D,V,f\rangle$,其中 $U=\bigcup_{i=1}^{t}U_i$.

输出:全局属性核 $\mathrm{Core}(C,U)$.

1) 求各站点 $S_i(i=1,2,\cdots,t)$ 上的所有局部不一致对象集合 U_{i2} 使得 $U_i=U_{i1}\cup U_{i2}$,其中 U_{i1} 中的各对象是局部一致的;

2) 寻找各 $U_{i1}(i=1,2,\cdots,t)$ 中的全局不一致对象使得 $U_i=U_{i1}\cup U_{i2}(i=1,2,\cdots,t)$,其中 U_{i1} 中的对象是全局一致的,而 U_{i2} 中的对象是全局不一致的,且对任意两个不一致对象仅保留其中的一个,也即 $U_{i2}\cap U_{j2}=\varnothing(i\neq j)$;

3) 在各站点 S_i 上求解局部决策表 DT_i 的局部属性核 $\mathrm{Core}(C,U_i)$;

4) 压缩各局部决策表 DT_i,得到各浓缩的局部决策表 DT_i^1;

5) 站点之间传送浓缩的决策表 DT_i^1,求

$CCore(C_1, U_{ji}^1, U_{kl}^1)$ $(1 \leq j \neq k \leq t, i$ 和 l 不同时为 2$)$,并记录使得 $a \in CCore(C_1, U_{ji}^1, U_{kl}^1)$ 的对象标识对(x, y);

6) 修剪 $CCore(C_1, U_{ji}^1, U_{kl}^1)$, 对 $\forall b \in \bigcup_{j=1}^{t} Core(C, U_j)$,是否有 $f(x, b) = f(y, b)$ 成立,若没有这样的对象标识对(x, y),则从 $CCore(C_1, U_{ji}^1, U_{kl}^1)$ 中删除 a;

7) 全局属性核 $Core(C, U)$ 为

$$\bigcup_{i=1}^{t} Core(C, U_i) \cup \left(\bigcup_{1 \leq j \neq k \leq t \text{且} i \text{和} l \text{不同时为} 2} CCore(C_1, U_{ji}^1, U_{kl}^1) \right).$$

对于例 2,$U_{11} = \varnothing, U_{12} = \{x_1, x_2\}; U_{21} = \varnothing, U_{22} = \{x_3, x_4\}$. 由定理 1 知,$Core(C, U_1) = \varnothing$, $Core(C, U_2) = \varnothing$, $Core(C, U_{11}, U_{21}) = \varnothing$; $Core(C, U_{11}, U_{12}) = \varnothing$; $Core(C, U_{12}, U_{21}) = \varnothing$. 故 $Core(C, U) = \varnothing$.

类似性质 3 和性质 4,有以下性质成立:

性质 5　对给定的 t 个站点 $S_i (i = 1, 2, \cdots, t)$,若 $\bigcup_{j=1}^{t} Core(C, U_j) \neq \varnothing$,则在站点 $S_i (i = 1, 2, \cdots, t)$ 上有 $|U_{i1}^1| \leq |U_{i1}|$, $|U_{i2}^1| \leq |U_{i2}|$ 成立.

性质 6　对给定的 t 个站点 $S_i (i = 1, 2, \cdots, t)$,算法 2 的空间复杂度为

$$O\left(|C| \sum_{i=1}^{t} (|U_{i1}| * (|U_{i1}| + |U_{i2}|)) + |C_1| \sum_{1 \leq i \neq j \leq t, k=2 \text{或} l \neq 2} (|U_{ik}^1| * |U_{jl}^1|) \right),$$

其中 $C_1 = C - \bigcup_{j=1}^{t} Core(C, U_j)$. 而单决策表属性核求解算法的空间复杂度为

$$O\left(|C| \left(\sum_{i=1}^{t} |U_{i1}| \right) \left(\sum_{i=1}^{t} (|U_{i1}| + |U_{i2}|) \right) \right).$$

可见,算法 2 通过传送压缩的决策表可有效降低网络传送代价,提高全局属性核的求解效率.

4.3　复杂度分析

因全局一致决策表可看成全局不一致决策表的特例,即各站点 S_i 上的 $|U_{i2}| = 0$, $|U_i| = |U_{i1}|$,因而这里仅对全局不一致情况下求解全局属性核算法进行复杂度分析.

4.3.1　空间复杂度

由性质 5 和性质 6,与单机算法相比,算法 GARBI 的空间复杂度至少可降低

$$O\left(\sum_{1 \leq i \neq j \leq t} (|C| |U_{i1}| - |C_1| |U_{i1}^1|)(|U_{j1}| + |U_{j2}|) \right),$$

当 $|C| |U_{i1}| \gg |C_1| |U_{i1}^1|$ 时,算法 2 可有效降低空间复杂度. 若采用并行计算 $CCore(C_1, U_{ji}^1, U_{kl}^1)$ 可进一步降低单个站点存储空间的压力.

4.3.2　时间复杂度

利用算法 GARBI 求局部不一致的时间复杂度为

$$O\left(|C| \max_{1 \leq i \leq t} (|U_i| \log^{|U_i|}) \right);$$

求全局不一致的时间复杂度为

$$O\left(|C| \left(\sum_{i=1}^{t} |U_{i1}| \right) \log_{\sum_{i=1}^{t} |U_{i1}|} \right);$$

并行地求各局部决策表的局部属性核的时间复杂度为

$$O\left(|C| \max_{1 \leq i \leq t} (|U_{i1}| (|U_{i1}| + |U_{i2}|)) \right);$$

求解 $\bigcup_{1 \leq j \neq k \leq t \text{且} i \text{和} l \text{不同时为} 2} CCore(C_1, U_{ji}^1, U_{kl}^1)$ 的总的时间复杂度为

$$O\left(\sum_{1 \leq i \neq j \leq t} |C| (|U_{i1}^1| (|U_{j1}^1| + |U_{j2}^1|) + |U_{i2}^1| |U_{j1}^1|) \right).$$

采用单机算法,在中心站点发现不一致对象集合的相应时间复杂度为

$$O\left(|C| \left(\sum_{i=1}^{t} |U_i| \right) \log_{\sum_{i=1}^{t} |U_i|} \right);$$

进而通过差别矩阵求解全局属性核的时间复杂度为

$$O\left(|C| \left(\sum_{i=1}^{t} |U_{i1}| \right) \left(\sum_{i=1}^{t} (|U_{i1}| + |U_{i2}|) \right) \right).$$

对算法 GARBI 而言,相对于求 $\bigcup_{1 \leq j \neq k \leq t \text{且} i \text{和} l \text{不同时为} 2} CCore(C_1, U_{ji}^1, U_{kl}^1)$ 的总时间,并行求局部不一致和求局部属性核的时间可忽略. 因此,与单机算法相比,算法 GARBI 的时间复杂度至少可降低

$$O\left(|C| \sum_{i=1}^{t} (|U_{i1}| |U_i|) \right).$$

通过复杂度分析可知,算法 GARBI 可有效地提高求解全局属性核的效率.

5　结　　语

本文针对全局决策表一致和不一致情况,探讨了全局属性核与局部属性核之间的关系,提出一种基于水平分布的多决策表全局属性核求解算法. 该算法可充分利用各局部属性核,通过传送压缩的局部决策表,可有效减少网络通讯量,提高全局属性核求解的效率,为分布式环境下的多决策表属性核求解提供了一条有效途径.

下一步将研究的主要内容是:1) 探索求解全局不一致对象的快速算法;2) 探索求解全局属性约简算法.

参考文献(References)

[1] Pawlak Z. Rough sets [J]. Int J of Information and Computer Science, 1982, 11(5): 341-356.

[2] Pawlak Z. Rough set approach to multi-attribute decision analysis [J]. European J of Operational Research, 1994, 72(3): 443-459.

[3] Roman W S, Larry H. Rough sets as a front end of neural-network texture classifiers[J]. Neurocomputing, 2001, 36(1-4): 85-102.

[4] 蔡忠闽, 管晓宏, 邵萍, 等. 基于粗糙集理论的入侵检测新方法[J]. 计算机学报, 2003, 26(3): 361-366.
(Cai Zhong-min, Guan Xiao-hong, Shao Ping, et al. A new approach to intrusion detection based on rough set theory[J]. Chinese J of Computers, 2003, 26(3): 361-366.)

[5] Lee W. A data mining framework for constructing features and models for intrusion detection systems[D]. New York: Columbia University, 1999.

[6] Prodromidis L, Stolfo S J. Agent-based distributed learning applied to fraud detection [R]. New York: Columbia University, 1999.

[7] Lee W, Stolfo S J. Data mining approaches for intrusion detection [C]. Proc of the 7th USENIX Security Symposium. San Antonio, 1998: 79-93.

[8] 杨明, 孙志挥, 宋余庆. 快速更新全局频繁项目集[J]. 软件学报, 2004, 15(8): 1189-1197.
(Yang Ming, Sun Zhi-hui, Song Yu-qing. Fast updating of globally frequent itemsets[J]. J of Software, 2004, 15(8): 1189-1197.)

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

[10] Wang Jue, Wang Ju. Reduction algorithm based on discernibility matrix the ordered attributes method[J]. J of Computer Science and Technology, 2001, 16(6): 489-504.

[11] 杨明, 孙志挥. 改进的差别矩阵及其求核方法[J]. 复旦大学学报, 2004, 43(5): 865-868.
(Yang Ming, Sun Zhi-hui. Improvement of discernibility matrix and the computation of a core[J]. J of Fudan University, 2004, 43(5): 865-868.)

[12] Hu X H, Cercone N. Learning in relational databases: A rough set approach [J]. Int of Computational Intelligence, 1995, 11(2): 323-338.

[13] 杨明. 一种基于改进差别矩阵的核增量式更新算法[J]. 计算机学报, 2006, 29(3): 407-413.
(Yang Ming. An incremental updating algorithm of the computation of a core based on the improved ciscernibility matrix [J]. Chinese J of Computers, 2006, 29(3): 407-413.)

[14] 王国胤. 决策表核属性的计算方法[J]. 计算机学报, 2003, 26(5): 611-615.
(Wang Guo-yin. Calculation methods for core attributes of decision table[J]. Chinese J of Computers, 2003, 26(5): 611-615.)

(上接第 126 页)

[36] Lvovsky A I. Iterative maximum-likelihood reconstruction in quantum homodyne tomography[J]. J Optics B: Quantum Semiclass Optcas, 2004, 6: S556-S559.

[37] Schack R, Brun T A, Caves C M. Quantum bayes rule [J]. Physical Review A, 2001, 64: 0143051.

[38] Buzek V, Derka R, Adam G, et al. Reconstruction of quantum states of spin systems: From quantum bayesian inference to quantum tomography[J]. Annals of Physics, 1998, 266: 454-496.

[39] Opatrny T, Welsch D G, Vogel W. Least-squares inversion for density-matrix reconstruction[J]. Physical Review A, 1997, 56: 1788-1799.

[40] Bardroff P J, Mayr E, Schleich W P, et al. Simulation of quantum state endoscopy [J]. Physical Review A, 1996, 53: 2736-2741.

[41] Artiles L M, Gill R D, Guta M I. An invitation to quantum tomography [J]. J of the Royal Statistical Society B, 2005, 67: 109-134.

[42] D'Alessandro D. On the observability and state determination of quantum mechanical systems[C]. Proc of the 43rd Conf on Decision and Control. Paradise Island: IEEE, Piscataway NJ, 2004.