文章编号: 1001-0920(2008)10-1103-06

垂直分布多决策表下基于条件信息熵的近似约简

杨 明, 杨 萍
(南京师范大学 数学与计算机科学学院, 南京 210097)

摘 要: 目前粗糙集理论研究主要针对单个决策表,而有关分布式环境下的核求解和属性约简研究的报道不多,为此提出垂直分布多决策表下基于条件信息熵的近似约简算法. 该算法在各局部站点并行求相应的条件信息熵,并通过传送部分等价类的策略,可有效降低通讯代价,提高垂直分布多决策表下基于条件信息熵的近似约简效率. 算法分析和实验结果表明,所提出的算法是有效可行的.

关键词: 粗糙集; 条件信息熵; 全局属性核; 局部属性核; 近似约简
中图分类号: TP311 **文献标识码**: A

Approximate reduction based on conditional information entropy over vertically partitioned multi-decision table

YANG Ming, YANG Ping
(School of Mathematics and Computer Science, Nanjing Normal University, Nanjing 210097, China. Correspondent: YANG Ming, E-mail: m.yang@njnu.edu.cn)

Abstract: At present, the research based on rough set theory aims mainly at a single decision table, but little work has been done for computation of a core and attribute reduction in distributed environment. Therefore, this paper introduces an approximate reduction algorithm based on conditional information entropy for vertically partitioned multi-decision table. The algorithm computes each corresponding conditional information entropy in parallel at every sites. By employing a data transmitted strategy, the communication cost is efficiently reduced and the performance of the algorithm for approximated reduction based on conditional information entropy over vertically partitioned multi-decision table can be efficiently improved. Algorithm analysis and experimental results show the effectiveness and feasibility of the proposed algorithm.

Key words: Rough set; Conditional information entropy; Global attribute core; Local attribute core; Approximate reduction

1 引 言

粗糙集(RS)作为一种新的处理不精确、不完全与不相容知识的数学理论[1],是分类规则获取的重要方法. 近年来,该理论在机器学习、数据挖掘及网络入侵检测等多个领域得到了广泛的应用[2-4]. 但在现实应用中,很多决策应用问题需要通过分布式环境下的各参与方的协同工作来完成[5,6],因此,进行分布式环境下的粗糙集理论研究具有重要的理论和现实意义.

在粗糙集理论研究中,核和属性约简是其主要内容,研究者从代数观和信息观两不同的角度对核和属性约简进行了深入的研究[7-12]. 但现有高效的核和属性约简算法主要针对单个决策表,而针对分布式环境下的多个决策表全局核和属性约简算法研究的报道不多.

文献[13]对垂直分布的多决策表全局属性核求解问题进行了讨论. 在引入全局决策表和局部决策表的概念后,提出一种基于垂直分布的多决策表全局属性核求解算法,但有关垂直分布多决策表下的属性约简研究还较为少见. 虽然可采用在某个中心站点将各局部决策表通过连接操作得到单个决策表,然后直接应用现有的属性约简算法求解,但这种

收稿日期: 2007-08-10; **修回日期**: 2007-11-05.
基金项目: 国家自然科学基金项目(40771163); 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; 杨萍(1967—),女,安徽宁国人,副教授,从事管理决策、粗糙集理论等研究.

处理方法的效率较低. 为此,本文沿用已提出的全局决策表和局部决策表概念[13],以针对单决策表下基于条件信息熵的属性约简[10-12]和已有的全局核求解[13]为基础,提出垂直分布多决策表下基于条件信息熵的近似约简算法. 该算法在各局部站点并行求相应的条件信息熵,并通过传送部分等价类的策略,以此降低通讯代价,提高垂直分布多决策表下基于条件信息熵的近似约简效率.

2 粗糙集概念及结论

本文约定有关记号同文献[13],采用的模型为多个决策表是垂直划分的. 设有 m 个站点 $S_1, S_2, \cdots, S_m$,相应的成员决策表 DT_i(或局部决策表)的属性集分别为 $C_1 \cup D, C_2 \cup D, \cdots, C_m \cup D$, $\bigcup_{i=1}^{m} C_i = \varnothing$,各局部决策表具有相同的对象集 U 且均隐含一个对象标识属性(约定标识本文不显示地给出). 通过该属性可将各局部决策表连接成一个单决策表 $DT = \langle U, C \cup D, V, f\rangle$, $C = \bigcup_{i=1}^{m} C_i$,并假设惟一的决策属性 D 的取值范围是 $1,2,\cdots,l$. 由 D 导出的等价类构成 U 的一个划分 $\{\Psi_1, \Psi_2, \cdots, \Psi_l\}$. 其中: $\Psi_i = \{x \in U: f(x, D) = i\}$, $i = 1,2,\cdots,l$; U 中的对象个数为 n. 此外,假设条件属性集合 C 的值域为有限离散集合,用 $|\cdot|$ 表示集合的基.

定义 1 全局决策表 DT 是一个四元组 $\langle U, C \cup D, V, f\rangle$. 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.

定义 2 在站点 S_i $(i = 1,2,\cdots,t)$,局部决策表 DT_i 是一个四元组 $\langle U, C_i \cup D, V, f\rangle$. 其中: C_i 为条件属性集, D 为决策属性集, $V = \bigcup_{a \in (C_i \cup D)} V_a$, V_a 为属性 a 的值域集, f 是 $U \times (C_i \cup D) \to V$ 的映射.

若 U 中的两个不同的对象 x 和 y 在条件属性集 C 或 C_i 上具有相同的条件属性值而具有不同的分类,则称 x 和 y 为全局不一致的或局部不一致的,否则称 x 和 y 为全局一致的或局部一致的.

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,否则称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.

设 X 为论域 U 的一个子集, $P \subseteq C$, X 的关于 P 的下近似为 $\underline{P}X = \{x \in U: [x]_P \subseteq X\}$,其中 $[x]_P$ 表示 U 中所有与 x 在关系 $IND(P)$ 下是等价的元素构成的集合.

定义 3 设 P 和 Q 在 U 上导出的划分分别为 X, Y $(X = \{X_1, X_2, \cdots, X_s\}, Y = \{Y_1, Y_2, \cdots, Y_t\})$,则 P 和 Q 在 U 的子集组成的 σ 代数上的概率分布为

$$[X:P] = \begin{bmatrix} X_1 & X_2 & \cdots & X_s \\ p(X_1) & p(X_2) & \cdots & p(X_s) \end{bmatrix},$$

$$[Y:P] = \begin{bmatrix} Y_1 & Y_2 & \cdots & Y_t \\ p(Y_1) & p(Y_2) & \cdots & p(Y_t) \end{bmatrix}.$$

其中

$$p(X_i) = \frac{|X_i|}{|U|}, \quad 1 \le i \le s,$$

$$p(Y_j) = \frac{|Y_j|}{|U|}, \quad 1 \le j \le t.$$

定义 4 知识(属性集合) $Q(U \mid IND(Q) = \{Y_1, Y_2, \cdots, Y_t\})$ 相对于知识(属性集合) $P(U \mid IND(P) = \{X_1, X_2, \cdots, X_s\})$ 的条件熵 $H(Q \mid P)$ 定义为

$$H(Q \mid P) = -\sum_{i=1}^{s} p(X_i) \sum_{j=1}^{t} p(Y_j \mid X_i) \log(p(Y_j \mid X_i)).$$

其中

$$p(Y_j \mid X_i) = |Y_j \cap X_i| / |X_i|,$$
$$1 \le i \le s, \quad 1 \le j \le t.$$

定义 5(条件信息熵下的属性约简定义) 对于给定的决策表 DT,若 $H(D \mid C) = H(D \mid A)$ $(A \subseteq C)$,且 $H(D \mid C) \ne H(D \mid B)$ $(\forall B \subset A)$,则 A 为决策表的一个属性约简.

定义 6(条件信息熵下的 ε 近似属性约简定义) 对于给定的决策表 DT 和 $\varepsilon(\varepsilon \ge 0)$,若 $|H(D \mid C) - H(D \mid A)| \le \varepsilon (A \subseteq C)$,且 $|H(D \mid C) - H(D \mid B)| > \varepsilon (\forall B \subset A)$,则 A 为决策表的一个 ε 近似属性约简.

定义 7 设 $X \subseteq U$ 为论域的一个子集, $P \subseteq C$, X 关于 P 的全局下近似为

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

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

定义 8(代数观下的全局近似精度定义) 设 $P \subseteq C$,对于划分 $\{\Psi_1, \Psi_2, \cdots, \Psi_l\}$ 的 P- 全局近似精度为

$$\gamma_{(P,S)} = \sum_{i=1}^{l} |\underline{P}\Psi_i(U)| / |U|.$$

定义 9(代数观下的属性约简定义) 设 $P \subseteq C$,若 $\gamma_{(P,S)} = \gamma_{(C,S)}$,且不存在 $R \subset P$,使得 $\gamma_{(R,S)} = \gamma_{(C,S)}$,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 $Core(C)$.

文献[10,11]对代数观和信息观两种观点下的属性约简进行了比较分析,证明了在一致决策表情况下,定义 5 和定义 9 是等价的;而在不一致决策表情况下,定义 5 和定义 9 不等价. 由文献[10]的结论可得如下性质:

性质 1　若 $H(D\mid A\cup\{a\})=H(D\mid A)$,则 $\mathrm{POS}_{A\cup\{a\}}(D)=\mathrm{POS}_A(D)$.

性质 2　若 $A\subseteq C, B\subseteq C$,且 $A\subseteq B$,则 $H(D\mid A)\geq H(D\mid B)$.

3　垂直多决策表下的全局属性核求解

定义 10[8]　对给定的单个决策表 DT,定义差别矩阵 $M=\{m_{ij}\}$ 为

$$m_{ij}=\begin{cases}\{a\in C: f(x_i,a)\neq f(x_j,a)\},\ f(x_i,D)\neq f(x_j,D),\\ \qquad x_i\in U_1, x_j\in U_1;\\ \{a\in C: f(x_i,a)\neq f(x_j,a)\},\ x_i\in U_1, x_j\in U_2';\\ \phi,\ \text{otherwise.}\end{cases}\tag{1}$$

其中

$$U_1=\bigcup_{i=1}^{l}\mathcal{L}\Psi_i, U_2=U-U_1,$$

$U_2'=$

$\{y\in U_2\mid$ 不存在 $x\in U_2$ 使得 $f(x,a)=f(y,a)$ 且 $f(x,D)\neq f(y,D), \forall a\in C\}$.

对于单决策表,定义 10 得到的改进差别矩阵即可得到一致或不一致决策表的核,相应的核由改进差别矩阵中仅含单个属性的元素得到[8].

定义 11[13]　对于给定的局部决策表 $\mathrm{DT}_k(k=1,2,\cdots,m)$,定义相应的差别矩阵 $M[k]=\{m_{ij}^k\}$ 为

$$m_{ij}^k=\begin{cases}\{a\in C_k: f(x_i,a)\neq f(x_j,a)\}, f(x_i,D)\neq\\ \qquad f(x_j,D),\ x_i\in U_1, x_j\in U_1;\\ \{a\in C_k: f(x_i,a)\neq f(x_j,a)\}, x_i\in U_1, x_j\in U_2^1;\\ \phi,\ \text{otherwise.}\end{cases}\tag{2}$$

其中:U_2 为全局不一致对象集合,$U_1=U-U_2$ 为全局一致对象集,全局不一致对象集合 U_2 的约简为 U_2^1.

定理 1[13]　对于各局部决策表 $\mathrm{DT}_k(k=1,2,\cdots,m)$,属性 $\{a\}\in\mathrm{Core}(C)$,当且仅当存在 k 使得 $M[k]$ 的元素 $m_{ij}^k=\{a\}(1\leq k\leq m)$,且任意 $s(1\leq s\neq k\leq m)$ 有 $M[s]$ 的元素 $m_{ij}^s=\varnothing$.

对于垂直分布多决策表,文献[13]采用如下两步进行全局属性核求解:1) 求解全局不一致对象集;2) 依据定义 11 和定理 1 可得全局属性核求解算法 CGAC,该算法的时间复杂度为[13]

$$O(\max_{1\leq i\leq m}(C_i)\,n\log^n)+O(|U_1|(|U_1|+|U_2^1|)),$$

其中 $|C|>\max_{1\leq i\leq m}(C_i)$, $m>1$.

4　垂直多决策表下基于条件信息熵的近似约简

4.1　等价类传送策略

网络通讯代价是影响分布式环境下多决策表属性约简效率的关键,因而如何有效减小网络通讯量便成为求解垂直分布多决策表下基于条件信息熵的近似约简的主要任务. 虽然采用将各站点的局部决策表传送到一中心站点可简单实现属性约简的求解,但该做法的网络通讯量大,尤其是面对较大规模且含有较高维数的局部决策表时,需要传送大量的数据.

由定义 4 可知,条件信息熵 $H(D\mid A)(A\subseteq C)$ 仅与等价类 U/A 及 $U/(D\cup A)$ 有关,因而采用有效的等价类存储机制并传送相应的等价类的部分元素策略,可有效避免传送相应的各局部决策表. 为此,对于 $U/P=\{X_1,X_2,\cdots,X_s\}$ 中的等价类 $X_i(1\leq i\leq s)$,本文采用如下的三元组存放:

(站点标识,$|X_i|$,X_i 的对象 ID 递增序列) 或

$$(|X_i|,\langle X_i(1),X_i(2),\cdots,X_i(|X_i|)\rangle).\tag{3}$$

其中:$X_i(j)(1\leq j\leq|X_i|)$ 为对象 ID,$X_i(j)<X_i(j+1)$,$j=1,2,\cdots,|X_i|-1$.

对于不同站点 S_i 和 S_j 上的划分 $U/A(A\subseteq C_i)$, $U/B(B\subseteq C_j)$, 利用该存储表示可得如下引理:

引理 1　对于不同站点 S_i 和 S_j 上的划分 $U/A=\{X_1,X_2,\cdots,X_s\}$ $(A\subseteq C_i)$,$U/B=\{Y_1,Y_2,\cdots,Y_t\}$ $(B\subseteq C_j)$,有

$$U/(A\cup B)=\{X_i\cap Y_j: X_i\cap Y_j\neq\varnothing, 1\leq i\leq s, 1\leq j\leq t\}.$$

可见,采用等价类传送方式,求 $U/(A\cup B)$ 的网络通讯量至多为 $n+\max(|U/A|,|U/B|)$ (n 为局部决策表的对象数),而采用传送子局部决策表的方法,相应的网络通讯量至少为 $\min(|A|,|B|)*n$,多数情况下 $\min(|A|,|B|)\gg 1$. 进一步,利用如下引理 2 和定理 2 仅需传送 U/A 或 U/B 中的部分等价类元素即可求解 $U/(A\cup B)$.

引理 2　对于不同站点 S_g 和 S_h 上的划分 $U/A=\{X_1,X_2,\cdots,X_s\}(A\subseteq C_g)$,$U/B=\{Y_1,Y_2,\cdots,Y_t\}$ $(B\subseteq C_h)$,$U/D=\{\Psi_1,\Psi_2,\cdots,\Psi_l\}$,若设 $X_u\subseteq\Psi_k(1\leq k\leq l)$,则 $\forall Y_j\in U/B, X_u\cap Y_j\subseteq\Psi_k$.

定理 2　对于不同站点 S_g 和 S_h 上的划分 $U/A=\{X_1,X_2,\cdots,X_s\}(A\subseteq C_g)$,$U/B=\{Y_1,Y_2,\cdots,Y_t\}$ $(B\subseteq C_h)$, $U/D=\{\Psi_1,\Psi_2,\cdots,\Psi_l\}$,若设 $X_u\subseteq$

Ψ_k,则 $\forall Y_i \in U/B, X_u \cap Y_i \neq \varnothing$,有

$$p(X_u \cap Y_i) \sum_{j}^{l} p(X_u \cap Y_i \cap \Psi_k) \log p(X_u \cap Y_i \cap \Psi_k) = 0.$$

证明 由引理 2 可知,$X_u \cap Y_i \subseteq \Psi_k$,有 $p(X_u \cap Y_i \cap \Psi_k) = 1$ 成立,故定理 2 成立. □

定理 3 对于不同站点 S_g 和 S_h 上的划分 $U/A = \{X_1, X_2, \cdots, X_s\}(A \subseteq C_g)$, $U/B = \{Y_1, Y_2, \cdots, Y_t\}(B \subseteq C_h)$, $U/D = \{\Psi_1, \Psi_2, \cdots, \Psi_l\}$, 若设 $Y_v(1 \leq v \leq k) \subseteq \Psi_{y(v)}(1 \leq y(v) \leq l)$, $X_u(1 \leq u \leq i) \subseteq \Psi_{x(u)}(1 \leq x(u) \leq l)$,则

$$H(D \mid A \cup B) = \sum_{j_1=1}^{s} \sum_{j_2=1}^{t} p(X_{j_1} \cap Y_{j_2}) \sum_{j_3}^{l} p(X_{j_1} \cap Y_{j2} \cap \Psi_{j3}) \log p(X_{j_1} \cap Y_{j2} \cap \Psi_{j3}).$$

定理 3 由定理 2 和定义 4 可得证.

由定理 3 可知,对于

$U/A = \{X_1, X_2, \cdots, X_s\}(A \subseteq C_g)$,

$U/B = \{Y_1, Y_2, \cdots, Y_t\}(B \subseteq C_h)$,

$U/D = \{\Psi_1, \Psi_2, \cdots, \Psi_l\}$,

$X_u(1 \leq u \leq i) \subseteq \Psi_{x(u)}$, $1 \leq x(u) \leq l$,

为求 $H(D \mid (A \cup B))$,仅需将 U/A 中等价类元素 $X_{i+1}, \cdots, X_s$ 从站点 g 传送到站点 h,因而需要的网络通讯量至多为 $\sum_{j=i+1}^{s} |X_j| + s - i$. 而

$$s - i < |U/A|, \sum_{j=i+1}^{s} |X_j| < \sum_{j}^{s} |X_j| = n,$$

且通常 $s - i$ 相对 $\sum_{j=i+1}^{s} |X_j|$ 可忽略不计. 为方便计,对 $U/A = \{X_1, X_2, \cdots, X_s\}$, $U/D = \{\Psi_1, \Psi_2, \cdots, \Psi_l\}$,记

$$\overline{U/A} = \{X_i \in U/A : X_i \cap \Psi_j \neq \varnothing \text{且} X_i \cap \Psi_k \neq \varnothing, 1 \leq j \neq k \leq l\},$$

记从某个站点 i 传送 U/A 到另一个站点 j 仅需要的网络通讯量为 $N(\overline{U/A})$.

定理 4 对于某个站点上的两个属性集 A 和 B,若 $A \subseteq B$,则 $N(\overline{U/A}) \geq N(\overline{U/B})$.

证明 对于 $X_i \in U/A$,若存在 $\Psi_k(1 \leq k \leq l)$ 使得 $X_i \subseteq \Psi_k$,则对任意 $Y_j \in U/(B - A)$, $(X_i \cap Y_j) \subseteq \Psi_k$,由定理 2 知 $N(\overline{U/A}) \geq N(\overline{U/B})$ 成立. □

由定理 4 和文献[10-12]中基于条件熵的属性约简或近似约简算法思路可知,在算法的运行过程中,随着重要属性的不断扩展,网络传送代价快速降低.

4.2 垂直多决策表下基于条件信息熵的近似约简

由 4.1 节可知,通过传送部分等价类机制,避免传送局部子决策表,可降低网络通讯代价. 而如何在多个站点传送部分等价类且有效增强各站点条件信息熵计算的并行性,从而快速有效得到近似约简,将是本节的主要任务.

对于单决策表 DT, 基于条件信息熵的近似约简的基本思路为:从核出发(B = Core),逐步扩展使得 $H(D \mid (B \cup \{a\}))$ 最小的属性 a 到 B 中,直到 $|H(D \mid C) - H(D \mid B)|$ 小于某个给定的可调正参数 ε,从而得到一个近似约简 B.

依据基于条件熵的近似约简方法和上述介绍的等价类传送策略,可得垂直分布多决策表下基于条件信息熵的近似约简的求解步骤如下:

Step1: 由 CGAC 算法求全局核 Core 且令 B = Core.

Step2: 各站点上并行计算 U/C_i, 传送各个 $\overline{U/C_i}(i \neq j)$ 到站点 j,继而得到 $\overline{U/C} = U/\bigcup_{i=1}^{m} C_i$, 并得到 $H(D \mid C)$.

Step3: 若 B 不空,则各站点并行求解 $U/(B \cap C_i)(1 \leq i \leq m)$. 传送各个 $\overline{U/(B \cap C_i)}(i \neq j)$ 到站点 j,继而得到 $\overline{U/B} = \bigcap_{i=1}^{m} \overline{U/(B \cap C_i)}$,并得到 $H(D \mid B)$,转至 Step5.

Step4: 若 $(C_i - B) \neq \varnothing, i \neq j$,则将站点 j 得到 $\overline{U/B}$ 传送到各站点 i, 在各站点并行计算 $\overline{U/(B \cup \{a_i\})}(a_i \in (C_i - B))$,继而得到使得 $H(D \mid (B \cup \{a_i\}))$ 最小的属性 a_i;并在站点 j 得到使得 $H(D \mid (B \cup \{a_j\})) \leq H(D \mid (B \cup \{a_i\}))(a_i \in (C_i - B), i \neq j)$ 的属性 a_j, $B = B \cup \{a_j\}$,转至 Step5.

Step5: 若 $|H(D \mid B) - H(D \mid C)| > \varepsilon$,则转至 Step4.

Step6: 类似 Step2,并行计算 $H(D \mid \{a_i\})(a_i \in B)$, 按 $H(D \mid \{a_i\})(a_i \in B)$ 递减的顺序对每个 a_i 重复下述操作:若 $|H(D \mid (B - \{a_i\})) - H(D \mid C)| \leq \varepsilon$,则属性 a_i 从 B 中删除, $B = B - \{a_i\}$;否则,属性 a_i 不能被约简, B 不变.

依据上述求解思路,可得关于垂直分布多决策表下基于条件信息熵的近似约简求解算法(ARVDT). 对该算法,若不考虑因某个站点不传送等价类而节省的通讯代价和 Step6 中少量属性的删除,仅考虑步骤 Step2 ~ Step4 的通讯代价,则 Step2 的通讯代价至多为 $\sum_{i}^{m} N(\overline{U/C_i}) \leq m * n$; Step3 的通讯代价至多为

$$\sum_{i=1}^{m} N(\overline{U/(\text{Core} \cap C_i)}) \leq$$

$\max_{1\le i\le m,\mathrm{Core}\cap C_i\neq\varnothing}(N(\overline{U/(\mathrm{Core}\cap C_i)})\times S(\mathrm{Core}))\le mn$,

其中 $S(\mathrm{Core})$ 为 m 个站点中含有核 Core 的站点数. Step4 的通讯代价至多为

$$m\Big(\sum_{j=1}^{k}N(\overline{U/(B\cup\{a_1,a_2,\cdots,a_j\})})\Big),$$

即核开始逐步扩展属性依次为 $a_1,a_2,\cdots,a_k$, 且

$$N(\overline{U/(B\cup(\{a_1,a_2,\cdots,a_j\}))})\ge N(\overline{U/(B\cup\{a_1,a_2,\cdots,a_{j+1}\})}),$$

因而存在某个 $1\le k_1<k$ 使得当 $j>k_1$ 时

$$N(\overline{U/(B\cup\{a_1,a_2,\cdots,a_{k_1+1}\})})\approx 0.$$

因此,ARVDT 算法的 Step2 ～ Step4 的通讯代价至多为

$$\sum_{i=1}^{m}N(\overline{U/C_i})+\sum_{i=1}^{m}N(\overline{U/(\mathrm{Core}\cap C_i)})+m\Big(\sum_{j=1}^{k}N(\overline{U/(B\cup\{a_1,a_2,\cdots,a_j\})})\Big)\le(k_1+2)mn<(k+2)mn.$$

一般情况下,ARVDT 算法的 Step2 ～ Step4 的通讯代价远远小于 $(k+2)mn$,这是因为随着属性 $a_j(j=1,2,\cdots,k)$ 的不断扩展,一些先前不确定的对象逐步被确定属于相应的类别,从而使得需传送的对象明显减少.

采用将 m 个站点上的局部决策表传送到某个中心站点的近似约简算法,不能充分利用各局部站点的计算能力,且需要的通讯代价为 $\sum_{i=1}^{m}|C_i|n$. 因此,当 k_1 相对较小时,ARVDT 算法的 Step2 ～ Step4 的通讯代价远低于 $\sum_{i=1}^{m}|C_i|n$. 同时,还可采用将数据量小的等价类传送到数据量相对大的等价类所在的站点来进一步优化 ARVDT 算法.

当然,在实际应用中,可依据局部决策表的规模选择采用相应的属性约简算法. 当局部决策表规模不大时,可采用选择采用各局部决策表传送到某个中心站点的方法. 而当局部决策表规模较大或属性约简长度相对较小时,采用 ARVDT 算法将有效降低通讯代价.

4.3　示例说明

表 1 为 2 个局部决策表 DT_1 和 DT_2,站点 S_1 上的局部决策表 DT_1 含 6 个对象,3 个条件属性 a,b,c, 1 个决策属性 d; 站点 S_2 上的局部决策表 DT_2 含 6 个对象,3 个条件属性 f,g,h, 1 个决策属性 d. 设 $\varepsilon=0.01$,ARVDT 算法的运行主要步骤如下:

Step1: $\mathrm{Core}=\{f\}$.

Step2: $U/\{a,b,c\}=\{\{x_1,x_6\},\{x_2,x_4\},\{x_3\},\{x_5\}\}$, $U/\{f,g,h\}=\{\{x_1,x_3\},\{x_2,x_4,x_5\},\{x_6\}\}$, $\overline{U/\{a,b,c\}}=\{\{x_1,x_6\},\{x_2,x_4\}\}$, $\overline{U/\{f,g,h\}}=\{\{x_1,x_3\},\{x_2,x_4,x_5\}\}$;

若选择将 $\overline{H/\{a,b,c\}}$ 从 S_1 传送到 S_2,则可得 $H(D|C)=0.231049$.

Step3: $U/\{f\}=\{\{x_6\},\{x_1,x_2,x_3,x_4,x_5\}\}$, $\overline{U/\{f\}}=\{\{x_1,x_2,x_3,x_4,x_5\}\}$, $H(D|\{f\})=0.560843$. 因 $|H(D|\{f\})-H(D|C)|>\varepsilon$,故执行 Step4.

Step4: 先将 $\overline{U/\{f\}}$ 传送到 S_1,在 S_1 求得扩展属性 a 使 $H(D|\{f,a\})=0.374809$. 在 S_2 求得扩展属性 h,使 $H(D|\{f,h\})=0.318257$,选 h 为下一个扩展属性. 进一步,将 $\overline{U/\{f,h\}}=\{\{x_2,x_4,x_5\}\}$ 从站点 S_2 传到站点 S_1,在站点 S_1 求得扩展属性 a 使 $H(D|\{a,f,h\})=0.231049$,而在站点 S_2 求得扩展属性 g 使 $H(D|\{f,g,h\})=0.318257$, 选 a 得到一个近似约简 $B=\{a,f,h\}$.

表 1　局部决策表 DT_1 和 DT_2

对象	站点 S_1				站点 S_2			
	a	b	c	d	f	g	h	d
x_1	0	0	0	0	0	0	0	0
x_2	0	1	0	0	0	0	1	0
x_3	0	0	1	0	0	0	0	0
x_4	0	1	0	1	0	0	1	1
x_5	1	0	0	1	0	0	1	1
x_6	0	0	0	1	1	0	0	1

可见,ARVDT 算法仅需传送 $(4+5+3)=12$ 个基本元素(这里每个属性值为一个基本元素),而若采用一个站点上的局部决策表传送到另一个站点,则需传送 $3\times6=18$ 个基本元素,故本文算法可有效降低通讯代价.

4.4　实验结果

为进一步验证算法的有效性,将数据集中和不集中两种情况下的近似通讯代价进行比较. 为方便计算,将数据集中后求近似约简算法[12] 记为 Old 算法. 选用 UCI 机器学习数据库中的 3 个数据集进行实验测试:1) Mushroom 数据集,共有 8124 个样本, 22 个条件属性和 1 个决策属性,简记为 Mush; 2) Glass Identification 数据集,共有 214 个样本,8 个条件属性和 1 个决策属性,将其分成 float 和 non float 类(即得 2 类),移去其第 1 列(即 ID number)后得到的数据集作为实验数据,简记为 Glass; 3) Ionosphere 数据集,共有 351 个样本,34 个条件属性和 1 个决策属性(共 2 类),记为 Ion. 随机抽取各数据集的 80 %,垂直平均划分为 2 个子数据集并分

配到两个不同的站点上,用VC++实现ARVDT和Old算法,实验结果如图1所示.

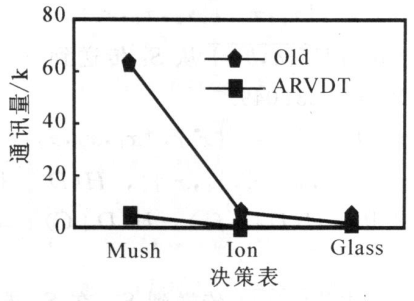

图1 算法的通讯代价

由图1可见,相对于将数据集中后求近似约简方法相比,本文算法的网络通讯代价明显低,这与等价类中未能正确分类的对象数随着重要属性逐步选择后快速减少是一致的.

5 结 论

本文提出垂直分布多决策表下基于条件信息熵的近似约简算法.该算法在各局部站点并行求相应的条件信息熵,并通过传送部分等价类元素的策略,可有效降低通讯代价,提高垂直分布多决策表下基于条件信息熵的近似约简效率,为分布式环境下垂直分布多决策表的属性约简提供了一种新的框架.算法分析和实验结果表明,本文的算法是有效的.

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 Z M, Guan X H, Shao P, et al. A new approach to intrusion detection based on rough set theory [J]. Chinese J of Computers, 2003, 26(3): 361-366.)

[5] Sabine M, David B S. Building predictors from vertically distributed data[C]. Proc of the 2004 Conf of the Centre for Advanced Studies on Collaborative Research. Ontario: Markham, 2004: 150-162.

[6] Du W L, Zhan Z J. Building decision tree classifier on private data[C]. Proc of the IEEE Int Conf on Privacy, Security and Data Mining. Maebashi, 2002: 1-8.

[7] 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.

[8] 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 M. An incremental updating algorithm of the computation of a core based on the improved discernibility matrix[J]. Chinese J of Computers, 2006, 29(3): 407-413.)

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

[10] 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.
(Wang G Y, Yu H, Yang D C. Decision table reduction on conditional information entropy [J]. Chinese J of Computers, 2002, 25(7): 759-766.)

[11] Wang G Y, Zhao J, An J J, et al. Theoretical study on attribute reduction of rough set theory: Comparison of algebra and information views [C]. Proc of the 3rd IEEE Int Conf on Cognitive Informatics. Washinaton: IEEE Computer Society, 2004: 148-155.

[12] 杨明. 决策表中基于条件信息熵的近似约简[J]. 电子学报, 2007, 35(11): 2156-2160.
(Yang M. Approximate reduction based on conditional information entropy in decision tables [J]. Acta Electronica Sinica, 2007, 35(11): 2156-2160.)

[13] 杨明, 杨萍. 一种基于垂直分布的多决策表全局属性核求解算法[J]. 控制与决策, 2006, 21(9): 991-996.
(Yang M, Yang P. An algorithm over vertically partitioned multi-decision table for computing global attribute core[J]. Control and Decision, 2006, 21(9): 991-996.)