文章编号: 1001-0920(2007)06-0652-05

一种面向不平衡分类数据的核求解算法

杨 明, 杨 萍
(南京师范大学 计算机科学系, 南京 210097)

摘 要: 对基于差别矩阵的核求解方法而言,差别矩阵的规模是直接影响核求解效率的关键因素. 为此,针对不平衡分类数据情况,提出一种基于多差别矩阵的核求解算法. 该算法先按决策属性值划分对象集,进而建立任意两个不同对象集对应的差别矩阵,形成多差别矩阵,从而求出核. 各差别矩阵因不平衡分类数据可有效降低其规模,提高核的求解效率. 理论分析和实验结果表明算法是有效可行的.

关键词: 粗糙集; 多差别矩阵; 核
中图分类号: TP311 **文献标识码**: A

An algorithm for computation of a core for unbalanced classification data

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

Abstract: For the method based on discernibility matrix for computing a core, reducing the size of discernibility matrix is the key for improving the performance of computation of a core. Therefore, an algorithm (AMDMC) based on multi-discernibility matrix is introduced to computation of a core for the case of unbalanced classification data. By the decision attribute's value, the all objects are partitioned into some subsets. For any two different subsets, a sub-discernibility matrix is created. Finally, the multi-discernibility matrix is obtained and a core is acquired. Each sub-discernibility matrix holds a small space because of unbalanced classification data, so the AMDMC algorithm is in high efficiency. Theoretical analysis and experiment results show the effectiveness of the algorithm.

Key words: Rough set; Multi-discernibility matrix; Core

1 引 言

粗糙集(RS)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]. 近年来该理论在机器学习及模式识别等多个领域得到了广泛的应用[2,3],如:利用求解的核可作为建立多变量决策树的依据[3]. 此外,核也是很多属性约简求解的关键步骤[4],因而探索研究求解核的有效方法具有重要的实用价值.

基于差别矩阵的求解核方法是 Hu[5]提出的一种简便、有效的方法,但有其不完善的地方. 为此,一些有效的求解核方法先后被提出[6-9](为方便,称文献[7]算法为 Wang 算法,称文献[8]算法为 Yang 算法),这些方法统一考虑一致和不一致情况下的核求解,但未能针对决策表分类数据不平衡情况进行差别矩阵改进.

针对决策表分类数据不平衡情况,本文提出一种基于多差别矩阵的核求解算法. 该算法先按决策属性值划分对象集,进而对任意两个不同对象集建立各个小规模的子差别矩阵,以此形成多差别矩阵,然后求出核. 多差别矩阵因各个子差别矩阵具有较小规模,可有效提高核的求解效率. 此外,本文的核求解方法可有效用于分布式环境下的核求解,降低数据通讯代价.

2 粗糙集概念

为节省篇幅,仅介绍与属性约简及核有关的一些概念,关于粗糙集的其他概念可参见文献[2].

收稿日期: 2006-03-29; **修回日期**: 2006-06-08.
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基金项目(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、粗糙集理论及应用等研究.

决策表 DT(或信息系统) 是一个四元组$\langle U, Q, V, f\rangle$. 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 $U = \{x_1, x_2, \cdots, x_n\}$;$Q = (C \cup D)$ 是属性集合,C 为条件属性集,D 为决策属性集;$V = \bigcup_{a\in Q} V_a$,V_a 为属性 a 的值域集;f 是 $U \times Q \to V$ 的映射.

在决策表中,若一些数据具有相同的条件属性值而具有不同的分类,则称这类数据是不一致的,否则为一致的;若两个不同的对象 x 和 y 具有相同的条件属性值而具有不同的分类,则称 x 和 y 为不一致的,否则称 x 和 y 为一致的.

为便于叙述,设条件属性集合 C 中有 m 个属性 $C_1, C_2, \cdots, C_m$,其值域为有限离散集合,并用$|\neg|$表示集合的基. 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 $\Psi_i = \{x \in U: f(x, D) = i\}$, $i = 1, \cdots, k$.

定义 1[2]　设 $X \subseteq U$ 为论域的一个子集,$P \subseteq C$,X 的关于 P 的下近似为

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

其中

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

定义 2[2]　设 $P \subseteq C$,对划分$\{\Psi_1, \Psi_2, \cdots, \Psi_k\}$的 P- 近似精度为

$$\gamma_P = \frac{\sum_{i=1}^{k} |\underline{P}\Psi_i|}{|U|}.$$

定义 3[2]　设 $P \subseteq C$,若 $\gamma_P = \gamma_C$,且不存在 $R \subset P$,使得 $\gamma_R = \gamma_C$,则称 P 为 C 的一个(相对于决策属性 D 的) 属性约简,所有 C 的属性约简的交称为 C 的核,记为 Core(C).

3　多差别矩阵及其求核算法

3.1　基于多差别矩阵的核求解算法

为有效地求解核,改进 Hu 和 Wang 方法的不足,文献[9] 在文献[8] 的基础上,引入下面定义:

定义 4[9]　对给定的决策表 DT,定义差别矩阵 $M_1 = \{m_{ij}\}$ 为

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

其中

$$U_1 = \bigcup_{i=1}^{k} \mathcal{L}\Psi_i, U_2 = U - U_1, U_2' = \text{delrep}(U_2).$$

函数 delrep(U_2) 描述如下:

```
Begin
  U2' = ∅;
  for 任意 x ∈U2 do
    if 不存在 y ∈U2 使得 ∀a ∈C, f(x,a) = f(y,a) 且 f(x,D) ≠f(y,D)
      then U2' = U2' ∪{x};
  return U2';
end.
```

由定义 4 可见,M_1 中同一决策分类的任意两个对象对应的矩阵元素为空,对求解核不起作用. 这样,对决策属性值的分布偏差较大的情况,M_1 中存在大量无用的空元素. 为有效地降低差别矩阵的规模,本文引入多差别矩阵的概念.

定义 5　对给定的决策表 DT,定义多差别矩阵 $M_2 = \{M(U_i, U_j) \mid 1 \leq i \neq j \leq k\}$,其中 $M(U_i, U_j) = \{m(x, y)\}(1 \leq i \neq j \leq k)$ 为

$$m(x, y) = \begin{cases} \{a \in C: f(x, a) \neq f(y, a)\}, \\ \quad x \in U_i, y \in U_j; \\ \varnothing, \text{ otherwise.} \end{cases} \tag{2}$$

$$U_i = \mathcal{L}\Psi_i, i = 1, 2, \cdots, k,$$

$$U_{k+1} = \text{delrep}\left(U - \sum_{i}^{k} U_i\right).$$

函数 delrep(U') 描述如下:

```
Begin
  U_{k+1} = ∅;
  for 任意 x ∈U' do
    if 不存在 y ∈U_{k+1} 使得 ∀a ∈C, f(x,a) = f(y,a) 且 f(x,D) ≠f(y,D)
    then U_{k+1} = U_{k+1} ∪{x};
  return U_{k+1};
end.
```

定理 1　对于决策表 DT,若记 IDM$(C, M_2) = \{m(x, y) \mid m(x, y) \in M(U_i, U_j), 1 \leq i \neq j \leq k$ 且 $m(x, y)$ 为单个属性$\}$,则有 IDM(C, M_2) = Core(C),即当且仅当某个 $m(x, y)$ 为单个属性时,该属性属于核 Core(C).

证明　只要证明 IDM(C, M_2) = IDM(C, M_1) 即可. 因 IDM$(C, M_2) \subseteq$ IDM(C, M_1),显然,故仅证 IDM$(C, M_1) \subseteq$ IDM(C, M_2) 即可.

若$\{a\} \in$IDM(C, M_1),则存在 $m_{ij} \in M_1$,即有下列条件之一成立:

1) 存在 $x_i \in \sum_{j=1}^{k} U_j$, $x_j \in \sum_{j=1}^{k} U_j$ 使得对任意 $b \in (C - \{a\})$ 有 $f(x_i, b) = f(x_j, b)$, $f(x_i, a) \neq f(x_j, a)$ 且 $f(x_i, D) \neq f(x_j, D)$;

2) 存在 $x_i \in \sum_{j=1}^{k} U_j$, $x_j \in U_{k+1}$ 使得对任意 $b \in (C - \{a\})$ 有 $f(x_i, b) = f(x_j, b)$, $f(x_i, a) \neq f(x_j, a)$.

容易证明条件 1) 和条件 2) 成立. □

实例 1 表 1 为二值数据表,其中共有 4 个元素和 4 个属性, $C = \{C_1, C_2, C_3\}$ 为条件属性集, D 为决策属性.

表 1 二值数据

元素	属性			
	C_1	C_2	C_3	D
x_1	1	0	1	0
x_2	0	0	1	1
x_3	0	0	1	0
x_4	1	1	1	1

对实例 1,依据定义 5 和定理 1 可得 $U_1 = \{x_1\}$, $U_2 = \{x_4\}$, $U_3 = \{x_2\}$,相应的各个差别矩阵为

$$M(U_1, U_2) = \begin{matrix} x_4 \\ x_1 \end{matrix} \begin{bmatrix} \{C_2\} \end{bmatrix},$$

$$M(U_1, U_3) = \begin{matrix} x_2 \\ x_1 \end{matrix} \begin{bmatrix} \{C_1\} \end{bmatrix},$$

$$M(U_2, U_3) = \begin{matrix} x_2 \\ x_4 \end{matrix} \begin{bmatrix} \{C_1, C_2\} \end{bmatrix}.$$

故可得 $\mathrm{Core}(C) = \{C_1, C_2\}$,且多差别矩阵 M_2 的矩阵元素个数为 3,规模小于 Wang 方法的 3 行 3 列矩阵,也小于 Yang 方法的 2 行 3 列矩阵,可有效降低存储开销.

依据定义 5 和定理 1,求解核的算法描述如下:

算法 1 AMDMC (an algorithm based on multi-discernibility matrix for computation of a core).

输入:决策表 $\mathrm{DT} = \langle U, C \cup D, V, f \rangle$;

输出:多差别矩阵 M_2 及相应的核 $\mathrm{Core}(C)$.

begin

1) 若决策属性 D 有 k 个不同的取值,则可得

$$U_i = \mathcal{L}\Psi_i, i = 1, 2, \cdots, k,$$

$$U_{k+1} = \mathrm{delrep}\left(U - \sum_{i=1}^{k} U_i\right),$$

其中 U_{k+1} 仅含不一致对象;

2) for $i = 1$ to $k - 1$ do

for $j = i + 1$ to k do

生成 $M(U_i, U_j)$;

3) 由定理 1 得核 $\mathrm{Core}(C)$.

end.

定理 2 设

$$U_i = \mathcal{L}_i, U_{k+1} = \mathrm{delrep}\left(U - \sum_{i=1}^{k} U_i\right), \quad i = 1, 2, \cdots, k,$$

则由算法 1 可得到正确的核.

定理 2 由定理 1 可证.

3.2 多差别矩阵的存储空间估计

对给定决策表 DT,若 $U_i = \mathcal{L}\Psi_i, i = 1, 2, \cdots, k$, $U_{k+1} = \mathrm{delrep}\left(U - \sum_{i=1}^{k} U_i\right)$. 本文设 $|U_i| = n_i, i = 1, 2, \cdots, k+1$,并令 $n_1 + n_2 + \cdots + n_k = N$, $n_1 + n_2 + \cdots + n_{k+1} = N^1$,则有如下性质成立:

性质 1 对给定决策表 DT,算法 1 得到的多差别矩阵的空间复杂度为 $O\left(\sum_{1 \leq i < j \leq k+1} n_i n_j\right)$.

性质 2 对给定决策表 DT,若 DT 是一致的,则 $N = N^1$, $n_{k+1} = 0$;若 DT 是不一致的,则 $N < N^1$, $n_{k+1} > 0$.

为有效估计算法 1 的存储空间的上下界,本文引入引理 1 和引理 2.

引理 1 若 $x_i \geq 0, i = 1, 2, \cdots, k$,且 $\sum_{i=1}^{k} x_i = N (N > 0)$,则 $\sum_{1 \leq i < j \leq k} x_i x_j$ 取最小值 0 的充分必要条件是存在 $x_s (1 \leq s \leq k)$ 使得 $x_s = N$, $x_t = 0, t = 1, 2, \cdots, s-1, s+1, \cdots, k$.

证明 易知 $\sum_{1 \leq i < j \leq k} x_i x_j \geq 0$,故 $\sum_{1 \leq i < j \leq k} x_i x_j = 0$ 当且仅当存在 $x_s (1 \leq s \leq k)$ 使得 $x_s = N$, $x_t = 0$, $(1 \leq t \leq k$ 且 $t \neq s)$. □

引理 2 若 $x_i \geq 0, i = 1, 2, \cdots, k$,且 $\sum_{i=1}^{k} x_i = N (N > 0)$,则 $\sum_{1 \leq i < j \leq k} x_i x_j$ 取最大值 $(k-1) N^2 / 2k$ 的充分必要条件是 $x_i = N/k, i = 1, 2, \cdots, k$.

证明 易知 $\sum_{1 \leq i < j \leq k} x_i x_j \geq 0$,因此 $\sum_{1 \leq i < j \leq k} x_i x_j$ 取最大值当且仅当 $\left(-\sum_{1 \leq i < j \leq k} x_i x_j\right)$ 取最小值,即 $N^2 + 2\left(-\sum_{1 \leq i < j \leq k} x_i x_j\right) = \sum_{i=1}^{k} x_i^2$ 取最小值.

令辅助函数 $L(x_1, \cdots, x_k) = \sum_{i=1}^{k} x_i^2 - \lambda\left(N - \sum_{i=1}^{k} x_i\right)$,则由 $\partial L / \partial x_i = 0, i = 1, 2, \cdots k, \partial L / \partial \lambda = 0$,得 $x_i = N/k, i = 1, 2, \cdots, k$. □

引理 3 若 $x_i \geq 0$ 且为整数, $i = 1, 2, \cdots, k$,

$\sum_{i=1}^{k} x_i = N$(N 为大于 0 的整数),$x = \left[\frac{N}{k}\right]$ ($[\ ]$ 为向下取整),则存在惟一的 $j(1 \le j \le k)$ 使得 $j * x + (k - j) * (x + 1) = N$.

证明　对给定 N 和 k,令 r 为 N 除以 k 的余数,$j = k - r$,则有 $j * x + (k - j) * (x + 1) = N$. 惟一性的证明是显然的. □

引理 4　若 $x_i \ge 0$ 且为整数,$i = 1,2,\cdots,k$,$\sum_{i=1}^{k} x_i = N$(N 为大于 0 的整数),$x = \left[\frac{N}{k}\right]$,则仅当在 x_i 中取 j 个 x,$(k - j)$ 个 $(x + 1)$ 时,$\sum_{1 \le i < j \le k} x_i x_j$ 为最大值,这里 $j = k - r$,r 为 N 除以 k 的余数.

证明　不妨设 $x_1, x_2, \cdots, x_j$ 均为 x,而 $x_{j+1}, x_{j+2}, \cdots, x_k$ 均为 $(x + 1)$,$\sum_{1 \le i < j \le k} x_i x_j$ 值为 L_1.

不失一般性,若令 $x_j = x - a(x \ge a > 0)$,$x_{j+1} = x + a + 1$,改变 x_j,x_{j+1} 后的 $\sum_{1 \le i < j \le k} x_i x_j$ 值为 L_2,则 $L_2 - L_1 = -a((j - 1) * x + a + 1) < 0$.

同理,可以证明任意改变 $x_i(i = 1,2,\cdots,k)$ 值,均有 $\sum_{1 \le i < j \le k} x_i x_j < L_1$,而由引理 3 可知 j 惟一,故结论成立. □

定理 3　对给定决策表 DT,算法 1 得到的多差别矩阵的空间复杂度 $\sum_{1 \le i < j \le k+1} n_i n_j$ 取最小值 0 的充分必要条件是存在 $U_s(1 \le s \le k + 1)$ 使得 $n_s = N$,$n_t = 0$,$t = 1,2,\cdots,s - 1,s + 1,\cdots,k + 1$.

定理 3 由引理 1 可证.

定理 4　对给定决策表 DT,算法 1 得到的多差别矩阵的空间复杂度 $\sum_{1 \le i < j \le k+1} n_i n_j$ 取近似最大值 $\left[\frac{k(N^1)^2}{2(k + 1)}\right]$ 的充分必要条件是存在 s 个 n_i 为 $\left[\frac{N^1}{k + 1}\right]$,$i = j_t$,$t = 1,2,\cdots,s$,且 $1 \le j_1 < j_2 < \cdots < j_s, s \le k + 1$,其余的 n_t 为 $\left[\frac{N^1}{k + 1}\right] + 1$,$1 \le t \le k + 1$,且 $t \ne j_1$,$t \ne j_2$,$\cdots$,$t \ne j_s$,其中 $s = k - r$,r 为 N 除以 k 的余数.

证明　由引理 2 可知,

$$\sum_{1 \le i < j \le k+1} n_i n_j \le \frac{k(N^1)^2}{2(k + 1)}. \tag{3}$$

式(3)的等号成立条件为当且仅当 $n_j = \frac{N^1}{k + 1}$,$j = 1,2,\cdots,k + 1$. 但 $n_j(j = 1,2,\cdots,k + 1)$ 只能取大于等于 0 的整数,因此当 N^1 不能被 $(k + 1)$ 整除时,式(3)的等号不成立,即 $\sum_{1 \le i < j \le k+1} n_i n_j$ 的最大值为 $\left[\frac{k(N^1)^2}{2(k + 1)}\right]$. 由引理 4,$\sum_{1 \le i < j \le k+1} n_i n_j$ 近似达到最大值 $\left[\frac{k(N^1)^2}{2(k + 1)}\right]$ 当且仅当 $n_j(j = 1,2,\cdots,k + 1)$ 的取值为 $\left[\frac{N^1}{k + 1}\right]$ 或 $\left[\frac{N^2}{k + 1}\right] + 1$. 设 $x = \left[\frac{N^1}{k + 1}\right]$,$R = \frac{N^1}{k + 1}$,且 $n_j(j = 1,2,\cdots,k + 1)$ 中有 s 个取值为 x,有 $(k + 1 - s)$ 个取值为 $(x + 1)$,则

$$\sum_{1 \le i < j \le k+1} n_i n_j = \frac{j(j - 1)x^2}{2} + j(k + 1 - j)x(x + 1) + \frac{(k + 1 - j)(k - j)(x + 1)^2}{2}.$$

若令 $y = R - x$,则有

$$\left| \frac{k(N^1)^2}{2(k + 1)} - \sum_{1 \le i < j \le k+1} n_i n_j \right| = \left| \frac{(k + 1 - s)(k - s)}{2} - \frac{k(k + 1)y^2}{2} \right| = \left| \frac{(k + 1)y(y - 1)}{2} \right| \le \frac{(k + 1)}{8}.$$

可见,当 $k + 1 < 8$ 时,$\sum_{1 \le i < j \le k+1} n_i n_j$ 可取得最大值 $\left[\frac{k(N^1)^2}{2(k + 1)}\right]$;当 $k + 1 \ge 8$ 时,$\sum_{1 \le i < j \le k+1} n_i n_j$ 可近似取得最大值 $\left[\frac{k(N^1)^2}{2(k + 1)}\right]$. □

定理 3 和定理 4 给出了算法 1 建立的多差别矩阵的存储空间上下界. 可以看出,给定决策表的决策分类数据分布偏差越大,AMDMC 的空间复杂度越低,效率越高.

3.3　复杂度分析

下面对 AMDMC 算法的空间和时间复杂度进行分析,并与 Wang 算法和 Yang 算法进行比较.

(1) 空间复杂度

由定理 4 可知,在最坏的情况下,AMDMC 的空间复杂度为 $O\left(\left[\frac{k(N^1)^2}{2(k + 1)}\right]\right) \approx O\left(\frac{(N^1)^2}{2}\right)$;又因 $N^1 - N = \left[\frac{N^1}{k + 1}\right]$,当 k 值较大时,$\left[\frac{N^1}{k + 1}\right] \ll N$,故有

$$O\left(\left[\frac{k(N^1)^2}{2(k + 1)}\right]\right) \approx O\left(\frac{(N^1)^2}{2}\right) \approx O\left(\frac{N * N^1}{2}\right).$$

无论决策表的决策分类数据分布如何,Wang 算法的空间复杂度为 $O((N^1)^2)$,Yang 算法的空间复杂度为 $O(N * N^1)$. 而由性质 1 及 N 和 N^1 的值可知,AMDMC 的空间复杂度低于 Yang 算法,且当给定决策表的决策分类数据分布偏差越大,AMDMC 的空间复杂度越低.

(2) 时间复杂度

对有 n 个对象的决策表,AMDMC 中步骤 1 的时间复杂度为 $O(n\log^n)$,最坏情况下步骤 2 的时间复杂度为 $O\left(\left[\frac{k(N^1)^2}{2(k+1)}\right]\right)$,这里 $N^1 < n$,故总的时间复杂度为 $O\left(n\log^n + \left[\frac{k(N^1)^2}{2(k+1)}\right]\right)$.

Wang 算法的时间复杂度为 $O(n\log^n + (N^1)^2)$,Yang 算法的时间复杂度为 $O(n\log^n + N * N^1)$. 可见,AMDMC 的时间复杂度比 Wang 算法和 Yang 算法低,尤其是给定决策表的决策分类数据分布偏差越大. 但 AMDMC 在进行数据划分时需要一定的代价,因而当决策表的规模不大,或决策分类数据分布偏差不大的情况下,可采用 Wang 或 Yang 算法. 而当决策表规模大,或决策分类数据分布偏差较大的情况下,可选用 AMDMC. 因此,在实际应用中,AMDMC 是现有核求解算法的有效补充,用户可依据决策表的规模和决策分类数据的分布情况,有效地选择相应的核求解算法.

4 实验结果

为进一步验证算法的性能,应用 VB6.0 在内存为 128 M,CPU 为 PIII400 MHZ,操作系统为 Windows 2000 Professional 的 DELL 笔记本上实现 Wang 算法、Yang 算法和 AMDMC 算法. 利用 http://www.ics.uci.edu/ 上所提供的蘑菇数据库来进行实验,该数据库有 8 124 条记录,记录蘑菇的 23 种属性,其中第 1 列为决策属性,共 2 个决策分类. 将下载的蘑菇数据库看作决策表 DT,并进行以下两组实验:1) 实验数据的选择如表 2 所示,实验结果如图 1 所示;2) 实验数据的选择如表 3 所示,实验结果如图 2 所示.

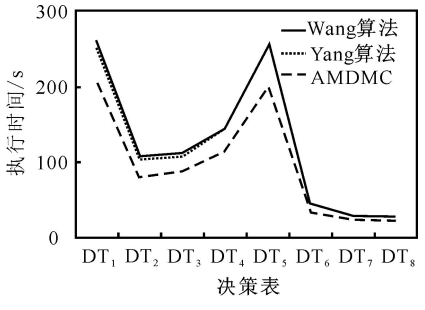

图 1 算法的执行时间(表 2 数据)

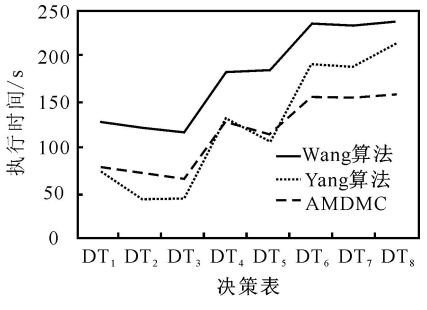

图 2 算法执行时间(表 3 数据)

表 2 第 1 组实验数据

决策表	总的对象个数	总的属性个数	U_1	U_2	U_3
DT_1	8 124	23	3 916	4 208	0
DT_2	5 440	22	3 848	1 592	0
DT_3	5 180	22	2 280	2 900	0
DT_4	6 060	22	2 368	3 692	0
DT_5	8 124	22	4 208	3 916	0
DT_6	3 240	21	2 096	1 144	0
DT_7	2 304	19	1 238	1 062	2
DT_8	2 304	18	1 238	1 062	2

表 3 第 2 组实验数据

决策表	总的对象个数	总的属性个数	U_1	U_2	U_3
DT_1	8 000	23	1 000	1 000	3 000
DT_2	8 000	23	1 800	200	3 000
DT_3	8 000	23	2 000	0	3 000
DT_4	8 000	23	2 000	2 000	2 000
DT_5	8 000	23	2 900	1 100	2 000
DT_6	8 000	23	3 000	3 000	1 000
DT_7	8 000	23	3 300	2 700	1 000
DT_8	8 000	23	3 400	3 400	600

注:$DT_1 \sim DT_8$ 由 DT 生成,U_1,U_2,U_3 同定义 5.

由图 1 可见,当决策表为一致的情况下,Wang 算法和 Yang 算法的效率相当,而 AMDMC 算法可有效地降低差别矩阵的空间复杂度,使得求核效率得到改进.

由图 2 可见,在决策表总对象数不变的情况下,不一致对象数目越多,Wang 算法、Yang 算法和 AMDMC 算法的执行效率均得到改进. 总体上看,Yang 算法和 AMDMAC 算法的性能优于 Wang 算法,这是因为 Wang 算法的差别矩阵存储开销高. 同时,当不一致对象数目较多时,Yang 算法的性能略优于 AMDMC 算法. 这是因为 Yang 算法的差别矩阵较小,几乎与 AMDMC 算法的多判别矩阵相近,而 AMDMC 算法在按决策属性值进行划分时的开销高于 Yang 算法. 但随着不一致对象数目的减少,Yang 算法的差别矩阵的存储代价变大,AMDMC 算法的性能优于 Yang 算法. 可见,AMDMC 算法的运行结果进一步验证了定理 4 的结论. 因此,在实际应用中,用户可根据决策表的规模和数据对象的分布特征选择相应的求解核算法,以便快速有效地求解决策表的核.

5 结 语

本文提出一种基于多差别矩阵的核求解算法,主要用于不平衡分类数据情况下的核求解,是现有求解核算法的有效补充. 下一步工作方向是将本文算法推广到分布式环境. (下转第 662 页)

合的性质,而约简分析则正是体现了规则集合度量的变化,从多角度综合描述了规则集合的性质,特别是对于多知识库的决策融合,提供了具有整体性的决策模型选择.基于模型集成的基本理论,将规则知识库作为一个知识库模型,通过模型的集成实现决策的融合,在模型一级实现了决策融合.

参考文献(References)

[1] Wong S K M, Ziarko W, Li Y R. Comparison of rough set and statistical methods in inductive learning[J]. Int J of Man-machine Studies, 1986, 25(1): 53-72.

[2] Agrawal R, Imielinski T, Swami A. Database mining: A performance perspective [J]. IEEE Trans on Knowledge and Data Engineering, 1993, 5(6): 169-173.

[3] Hu X, Cercone N, Han J. Fuzzy sets and knowledge discovery[M]. Berlin: Springer, 1994: 90-99.

[4] 李剑, 范小军, 黄沛. 基于粗糙集的知识理论及其应用[J]. 系统工程理论方法应用, 2001, 10(3): 184-188.
(Li J, Fan X J, Huang P. Theory of knowledge based on rough sets and its application [J]. Systems Engineering — Theory Methodology Applications, 2001, 10(3): 184-188.)

[5] 印勇, 曹长修, 张邦礼. 基于粗糙集理论的分类规则发现[J]. 重庆大学学报, 2000, 23(1): 63-65.
(Yin Y, Cao C X, Zhang B L. Classification rule discovery based on rough set theory[J]. J of Chongqing University, 2000, 23(1): 63-65.)

[6] Pawlak Z. Rough classification [J]. Int J of Man-machine Studies, 1984, 20(5): 469-483.

[7] 苗夺谦, 范世栋. 知识的粒度计算及其应用[J]. 系统工程理论与实践, 2002, 22(1): 48-56.
(Miao D Q, Fan S D. The calculation of knowledge granulation and its application[J]. Systems Engineering-Theory and Practice, 2002, 22(1): 48-56.)

[8] Watson H J. Decision support in the data warehouse [M]. New Jersey: Prentice-Hall, 1998.

[9] Kimball R, Strehlo K. Why decision support fails and how to fix it[J]. Acm Sigmod Record, 1995, 24(3): 65-77.

[10] Saifur Rahman, Rahul Bhatnagar. An expert system based algorithm for short term load foresast[J]. IEEE Trans on PWRS, 1998, 3(2): 886-892.

[11] Huang T Y. Intelligent decision support systems[M]. Beijing: Publishing House of Electronics Industry, 2001: 73-102.

[12] Kottemann J E, Dolk D R. Process-oriented model integration[C]. Proc of the 21st Hawaii Int Conf on System Sciences Ⅲ Hawaii: IEEE Computer Society Press, 1998: 656-668.

(上接第 656 页)

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] 苗夺谦, 王珏. 粗糙集理论中概念与运算的信息表示[J]. 软件学报, 1999, 10(2): 113-116.
(Miao D Q, Wang J. A information representation of theconcepts and operations in rough set theory[J]. J of Software, 1999, 10(2): 113-116.)

[4] 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.

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

[6] 叶东毅, 陈昭炯. 一个新的差别矩阵及其求核方法[J]. 电子学报, 2002, 30(7): 1086-1088.
(Ye D Y, Chen Z J. A new discernibility matrix and the computation of a core[J]. Chinese J of Electrics, 2002, 30(7): 1086-1088.)

[7] Zheng Z, Wang G Y, Wu Y. Objects' combination based simple computation of attribute core[C]. Proc of the 2002 IEEE Int Symposium on Intelligent Control. Vancouver, 2002: 514-519.

[8] 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 M, Sun Z H. Improvement of discernibility matrix and the computation of a core[J]. J of Fudan University, 2004, 43(5): 865-868.)

[9] 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.)

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