文章编号: 1001-0920(2008)09-1049-06

基于广义差别矩阵的核和属性约简算法

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: 属性约简是粗糙集理论的重要研究内容. 为此引入广义差别矩阵,提出基于广义差别矩阵的核和属性约简算法. 该框架可有效避免连续属性值离散化,且有利于与其他机器学习方法相结合. 理论分析表明,所提出的算法是有效而可行的.

关键词: 粗糙集; 广义差别矩阵; 核; 属性约简
中图分类号: TP311　　**文献标识码**: A

Algorithms based on general discernibility matrix for computation of a core and attribute reduction

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: Attributes reduction is one of important parts researched in rough set theory. Therefore, in this paper, after proposing a concept of general discernibility matrix, both computation of a core and attribute reduction algorithms based on general discernibility matrix are introduced. The newly proposed framework can effectively avoid discretizing the continuous-valued attributes and be easily incorporated into other machine learning methods. Theoretical analysis shows the effectiveness of the algorithm.

Key words: Rough set; General discernibility matrix; Core; Attributes reduction

1　引　　言

粗糙集是处理不精确、不完全和不相容知识的一种新的数学理论[1]. 近年来,该理论在机器学习、数据挖掘等诸多领域得到了广泛的应用[2-4]. 核和属性约简是粗糙集的重要内容之一,并已取得了一些重要的研究成果[3-11]. 现有的核和属性约简方法,大体上可分为: 基于差别矩阵的核求解和更新算法[5,6],基于正区域的属性约简算法[7],基于启发式的属性约简算法[8],基于差别矩阵或改进的属性约简算法[9-11]. 其中基于差别矩阵的方法以其简明高效而受到研究者的广泛关注.

传统的属性约简算法[9,10]主要针对离散值属性的决策表或信息系统,即需将连续值属性转化为离散值属性. 而离散化结果的好坏在某种程度上直接影响属性约简的求解结果,这是因为离散化在一定程度上会使某些有用的信息丢失[12]. 为此,文献[11]采用距离保持策略,直接针对连续值属性决策表构造模糊差别矩阵,以此得到属性约简. 但该文仅考虑只含连续值属性的决策表,未考虑既有离散值属性又有连续值属性的情况;且得到的属性约简本质上仅为近似属性约简,即某些原先可区分的一致对象经约简后可能变为不一致对象.

本文引入广义差别矩阵的概念,提出基于广义差别矩阵的核和属性约简算法. 该框架针对含有离散值属性和连续值属性的决策表(简称含有混合属性的决策表)或信息系统,直接构建相应的广义差别矩阵. 它既是对传统的面向离散值属性的差别矩阵[9,10]和面向连续值属性的差别矩阵[11]的改进,也有利于与其他机器学习方法相结合.

2　粗糙集概念

收稿日期: 2007-06-30; **修回日期**: 2007-09-23.
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、粗糙集理论及应用的研究.

这里仅介绍与属性约简及核有关的一些概念，粗糙集的其他一些概念可参见文献[3].

信息系统 IS 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 $U = \{x_1, x_2, \cdots, x_n\}$；$Q$ 是属性集合；$V = \bigcup_{a\in Q} V_a$，V_a 为属性 a 的值域集；f 是 $U \times Q \to V$ 的映射.

属性集合 Q 通常可分为条件属性集 C 和决策属性集 D. 对于 $B \subseteq Q$，无差别关系 $\mathrm{IND}(B)$ 定义为
$$\{(x, y) \in U^2 \mid \forall a \in B, f(x, a) = f(y, a)\}.$$
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. 用 $\varnothing$ 表示空集，$\mathrm{card}(X)$ 表示集合 X 的基，$|x - y|$ 表示两个数差的绝对值. 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 $\psi_i = \{x \in U : f(x, D) = i\}$，$i = 1, 2, \cdots, k$.

在信息系统(或决策表)中，如果一些数据具有相同的条件属性而有不同的分类，则称这类数据是不一致的；否则，称为一致的. 如果两个不同的对象 x 和 y 具有相同的条件属性值而有不同的分类，则称 x 和 y 为不一致的；否则，称 x 和 y 为一致的. 称不含不一致对象的决策表为一致决策表，称含有不一致对象的决策表为不一致决策表.

定义 1[3] 设 X 为论域 U 的一个子集，$P \subseteq C$，X 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 $\mathrm{IND}(P)$ 下是等价的元素构成的集合.

定义 2[3] 设 $P \subseteq C$，对于划分 $\{\psi_1, \psi_2, \cdots, \psi_k\}$ 的 P-近似精度为 $\gamma_P = \sum_{i=1}^{k} \mathrm{card}(\underline{P}\psi_i) / \mathrm{card}(U)$.

定义 3[3] 设 $P \subseteq C$，如果 $\gamma_P = \gamma_C$，且不存在 $R \subset P$，使 $\gamma_R = \gamma_C$，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(简称核)，记为 $\mathrm{Core}(C)$.

定义 4[3] 如果属性 $a \in C$ 满足 $\gamma_{C-\{a\}} < \gamma_C$，则称属性 a 为不可缺少的；否则，称属性为冗余的.

性质 1[3] 属性 $a \in \mathrm{Core}(C)$，当且仅当 a 是不可缺少的属性.

3 改进的差别矩阵与模糊差别矩阵

3.1 改进的差别矩阵

为克服传统的差别矩阵仅适用于一致决策表的不足，文献[6,9]引入了下面的定义，并给出了基于改进差别矩阵的核和属性约简更新算法.

定义 5[6,9] 对于给定的信息系统 IS，定义差别矩阵 $M_1 = \{m_{ij}\}$ 为

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

其中：$U_1 = \bigcup_{i=1}^{k} \underline{C}\psi_i$，$U_2 = U - U_1$，$U_2' = \mathrm{delrep}(U_2)$. 函数 $\mathrm{delrep}(U_2)$ 描述如下：

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

对于给定的差别矩阵，它与基于正域的属性约简等价的属性约简定义如下：

定义 6 对于给定的差别矩阵 M，当 $P \subseteq C$，$\forall m_{ij} \in M$，$m_{ij} \neq \varnothing$ 时，若有 $m_{ij} \cap P \neq \varnothing$ 成立，则称 P 为一个候选属性约简；若 P 为一个候选属性约简，且对于 $\forall a \in P$，$\exists m_{ij} \in M$，$m_{ij} \neq \varnothing$，使 $m_{ij} \cap (P - \{a\}) = \varnothing$，则称 P 为一个属性约简.

依据定义 6，可快速有效地求出属性约简，但它仅适用于只含离散值属性的决策表. 对于含有连续属性值的决策表，首先需要进行离散化，而离散化在一定程度上会丢失部分可能有用的信息，进而可能影响属性约简的结果. 因此，针对含有连续值属性的决策表，直接进行核求解和属性约简，便成为粗糙集模型研究的重要内容之一.

3.2 模糊差别矩阵

对于一个连续值属性的决策表 DT，若任意两个不同类别的对象 x_i 和 x_j $(1 \leq i, j \leq \mathrm{card}(U))$ 之间的距离足够小，则表明它们不可区分；反之，它们是可区分的两个对象. 文献[11]采用距离保持策略，对于可区分的两个不同类别的对象，首先由那些对距离贡献较大的属性构成模糊差别矩阵的元素，然后由模糊差别矩阵求解属性约简. 对于给定的一个连续值属性的决策表 DT，首先对其每列进行最小-最大规范化处理；然后依据给定参数 $\varepsilon(\varepsilon > 0)$ 和 $\delta(\delta > 0)$，定义模糊差别矩阵 M_2 如下：

$$m_{i,j}=\begin{cases} g_\varepsilon(x_i,x_j),\ d_C(x_i,x_j)\ge\delta, \\ \quad f(x_i,d)\ne f(x_j,d),\ 1\le i,j\le n; \\ \varnothing,\ \text{其他}. \end{cases} \tag{2}$$

其中

$$d_C(x_i,x_j)=\sum_{a\in C}|f(x_i,a)-f(x_j,a)|,$$

$d_C(x_i,x_j)\ge\delta$ 表示两个不同类别对象 x_i 和 x_j 是可区分的;

$$g_\varepsilon(x_i,x_j)=\arg\min_B|B|,$$

$$\text{s.t.}\ \frac{\sum_{b\in B(B\subseteq C)}|f(x_i,b)-f(x_j,b)|}{\sum_{a\in C}|f(x_i,a)-f(x_j,a)|}\ge\varepsilon.$$

即 $g_\varepsilon(x_i,x_j)$ 为

$$\frac{\sum_{b\in B(B\subseteq C)}|f(x_i,b)-f(x_j,b)|}{\sum_{a\in C}|f(x_i,a)-f(x_j,a)|}\ge\varepsilon$$

成立的基数最小的集合 B.

文献[11]只讨论了含有连续值属性的决策表,且得到的属性约简本质上为近似属性约简. 如何针对含有混合属性的决策表,直接构建有效的差别矩阵并进行核和属性约简,则是本文所讨论的主要内容.

4　广义差别矩阵及其核和属性约简方法

为方便起见,对于给定的决策表 DT,假设条件属性集 $C=\{C_1,\cdots,C_s,\cdots,C_m\}$ 中前 s 个属性为离散值属性,简记为 $C^1=\{C_1,C_2,\cdots,C_S\}$;后 $(m-s)$ 个属性为连续值属性,简记为 $C^2=\{C_{s+1},\cdots,C_m\}$. 此外,假设连续值属性已作了规范化处理. 即对于任意的 $a\in C^2$, $x\in U$,有 $f(x,a)\in[0,1]$.

4.1　广义差别矩阵

对于既有离散值属性又有连续值属性的决策表,前面给出的不一致决策表定义不再有效;同时,差别矩阵定义的式(1)和(2)也不再有效. 为此,引入广义不一致决策表、广义不一致对象的定义和广义差别矩阵的定义.

定义 7　对于 DT 中的任意两个对象 x_i 和 x_j $(1\le i,j\le \mathrm{card}(U))$, $P\subseteq C$, $P\ne\varnothing$,若 $f(x_i,d)\ne f(x_j,d)$, $\forall b\in(C^1\cap P)$,有 $f(x_i,b)=f(x_j,b)$ 成立,且给定的参数 $\delta(\delta>0)$,对于 $\forall a\in(C^2\cap P)$,有 $\mathrm{dissim}(f(x_i,a),f(x_j,a))<\delta$ 成立. 则称对象 x_i 和 x_j 为 P-广义不一致对象或 (δ,P)-不一致对象;否则,称对象 x_i 和 x_j 为 P-广义一致对象或 (δ,P)-一致对象. 为简便计,若 $P=C$,则 C-广义一致(不一致)对象简称为广义一致(不一致)对象. 称不含广义不一致对象的决策表为广义一致决策表,称含有广义不一致对象的决策表为广义不一致决策表.

注 1　在定义 7 中, $\mathrm{dissim}(f(x_i,a),f(x_j,a))$ 为对象 x_i 和 x_j 在属性 a 上的不相似性度量. 本文定义

$$\mathrm{dissim}(f(x_i,a),f(x_j,a))=|f(x_i,a)-f(x_j,a)|/\sigma_a,$$

σ_a 为属性 a 上属性值的标准方差,即所谓距离越大越不相似.

性质 2　对于 DT 中的任意两个对象 x 和 y,给定参数 $\delta(\delta>0)$,若对象 x 和 y 是 C-广义不一致的,则对于任意 $P\subset C$, $P\ne\varnothing$,对象 x 和 y 是 P-广义不一致的;反之则不成立.

定义 8　对于给定的决策表 DT,定义广义差别矩阵 $M_3=\{m_{ij}\}$ 为

$$m_{ij}=\begin{cases}\{a\in C^1: f(x_i,a)\ne f(x_j,a)\}\cup g_{(C^2,\delta)}(x_i,x_j), \\ \quad f(x_i,d)\ne f(x_j,d),\ x_i,x_j\in U_1; \\ \{a\in C^1: f(x_i,a)\ne f(x_j,a)\}\cup g_{(C^2,\delta)}(x_i,x_j), \\ \quad x_i\in U_1,\ x_j\in U_2; \\ \varnothing,\ \text{其他}.\end{cases} \tag{3}$$

其中, $U_2(C)=\{x\in U\mid \exists y\in U$,使 $f(x,d)\ne f(y,d)$, $f(x,a)=f(y,a)$, $\forall a\in C^1$, $|f(x,b)-f(y,b)|/\sigma_b<\delta$, $\forall b\in C^2\}$, $U_1(C)=U-U_2(C)$. 这里 $g_{(C^2,\delta)}(x_i,x_j)$ 可由下式得出:

$$g_{(C^2,\delta)}(x_i,x_j)=\begin{cases}\{a\in C^2: |f(x_i,a)-f(x_j,a)|/\sigma_a\ge\delta\}, \\ \quad \exists b\in C^2,\ |f(x_i,b)-f(x_j,b)|/\sigma_b\ge\delta; \\ \varnothing,\ \text{其他}.\end{cases} \tag{4}$$

由定义 8 可知,当 $C^2=\varnothing$ 时,由式(3)可直接得到由式(1)给出的差别矩阵. 可见,定义 8 是文献[6,9]的改进差别矩阵的推广. 此外,式(3)在单个属性上考虑两个对象的不相似性,在改进式(2)仅适用于含有连续值属性的决策表的同时,可使求得的属性约简符合传统的 Rough 框架. 详细说明参见 4.2 节和 4.3 节.

4.2　基于广义差别矩阵的核求解

核是某些属性约简算法的关键步骤[3,9,10]. 为此,下面在给出新的不可缺少属性定义后,引入判别某个属性是否为不可缺少属性的准则,进而得到由广义差别矩阵求解核的有效方法.

定义 9　对于一个含有混合属性的决策表 DT,给定某个参数 $\delta(\delta>0)$ 及某个属性 $a\in C$,若 $U_1(C-\{a\})\subset U_1(C)$,则属性 a 是不可缺少的.

由定义 9 可得如下引理:

引理 1　对于一个含有混合属性的决策表 DT,给定某个参数 $\delta(\delta > 0)$,若存在两个不同类别对象 x 和 y,满足下列条件之一:

1) 存在一个属性 $a \in C^1$,使 $f(x,a) \neq f(y,a)$;对于 $\forall b \in (C^1 - \{a\})$,有 $f(x,b) = f(y,b)$ 成立;对于 $\forall b \in C^2$,有 $|f(x,b) - f(y,b)|/\sigma_b < \delta$成立.

2) 存在一个属性 $a \in C^2$,使 $|f(x,a) - f(y,a)|/\sigma_a \geq \delta$成立;对于 $\forall b \in C^1$,有 $f(x,b) = f(y,b)$ 成立;对于 $\forall b \in (C^2 - \{a\})$,有 $|f(x,b) - f(y,b)|/\sigma_b < \delta$成立.

其中 x 或 y 是广义一致对象. 则属性 a 是不可缺少的.

证明　不妨设 x 是广义一致对象,若条件 1) 成立,则在属性集 $(C - \{a\})$ 上,x 和 y 是广义不一致对象,即 $U_1(C - \{a\}) \subset U_1(C)$. 因而属性 a 是不可缺少的. 同理,当条件 2) 成立时,属性 a 也是不可缺少的. □

由引理 1 可有效判别某个属性是否为不可缺少的.

引理 2　对于一个含有混合属性的决策表 DT,给定某个参数 $\delta(\delta > 0)$,得到的广义差别矩阵为 M_3,则 M_3 中的某个元素 m_{ij} 为单个属性$\{a\}$,当且仅当下列条件之一成立:

1) $a \in C^1$,使 $f(x_i,a) \neq f(x_j,a)$;对于 $\forall b \in (C^1 - \{a\})$,有 $f(x_i,b) = f(x_j,b)$ 成立;对于 $\forall b \in C^2$,有 $|f(x_i,b) - f(x_j,b)|\sigma_b < \delta$成立.

2) $a \in C^2$,使 $|f(x_i,a) - f(x_j,a)|/\sigma_a \geq \delta$;对于 $\forall b \in C^1$,有 $f(x_i,b) = f(x_j,b)$ 成立;对于 $\forall b \in (C^2 - \{a\})$,有 $|f(x_i,b) - f(x_j,b)|/\sigma_b < \delta$成立.

由定义 8 即可证得.

由核的定义知,不可缺少属性必为核的属性,因而由引理 1 和引理 2 可得如下定理:

定理 1　对于一个含有混合属性的决策表 DT,给定某个参数 $\delta(\delta > 0)$,若记 $\mathrm{IDM}(C,M_3) = \{m_{ij} | m_{ij} \in M_3$ 且 m_{ij} 为单个属性$\}$,则有 $\mathrm{IDM}(C,M_3) = \mathrm{Core}(C)$. 即当且仅当某个 m_{ij} 为单个属性时,该属性属于核 Core(C).

由引理 1 和引理 2 即可证得.

定理 1 为含有混合属性的决策表提供了一种求核的方法. 依据定义 8 和定理 1,基于广义差别矩阵的求核算法描述如下:

算法 1　基于广义差别矩阵的核求解

输入:1) 给定决策表 $\mathrm{DT} = \langle U, C \cup D, V, f\rangle$, $C = C^1 \cup C^2$;

2) 给定可调参数 $\delta(\delta > 0)$.

输出: 核 Core(C).

Begin

1) $\mathrm{Core}(C) = \varnothing$;

2) for $i = 1$ to card(U) do

for $\forall a \in C^2$ do

$$f(x_i,a) = \frac{f(x_i,a) - \min_a}{\max_a - \min_a};$$

// $\max_a$ 和 $\min_a$ 为属性 a 所在列的最大和最小属性值

3) for $\forall a \in C^2$ do

$$\mathrm{mean}_a = \sum_{i=1}^{\mathrm{card}(U)} f(x_i,a) \Big/ \mathrm{card}(U);$$

// mean_a 为 a 的属性值的均值

4) for $\forall a \in C^2$ do

$$\sigma_a = \sqrt{\sum_{i=1}^{\mathrm{card}(U)} (f(x_i,a) - \mathrm{mean}_a)^2 \Big/ (\mathrm{card}(U) - 1)};$$

// σ_a 为标准方差

5) 建立广义差别矩阵 M_3;// 由定义 8

6) for $\forall m_{ij} \in M_3$ do// 由定理 1

if $(m_{ij} \neq \varnothing)$ then

if card$(m_{ij}) = 1$ then

$\mathrm{Core}(C) = \mathrm{Core}(C) \cup m_{ij}$;

7) return Core(C);

End

算法 1 的步骤 2) 为进行数据的规范化处理,需要的时间为$(m - s) * \mathrm{card}(U)$. 步骤 3) 和 4) 的时间均为$(m - s) * \mathrm{card}(U)$. 步骤 5) 先判别哪些对象为不一致对象,再建立广义差别矩阵; 判断元素是否一致和建立广义差别矩阵的时间分别为 $(m * \mathrm{card}(U) * \log(\mathrm{card}(U)))$ 和 $(m * \mathrm{card}(U_1) * \mathrm{card}(U))$,建立的广义差别矩阵的空间复杂度为 $O(m * \mathrm{card}(U_1) * \mathrm{card}(U))$. 步骤 6) 的时间复杂度为 $O(m * \mathrm{card}(U_1) * \mathrm{card}(U))$. 于是,可得算法 1 的时间复杂度为 $O(m * \mathrm{card}(U_1) * \mathrm{card}(U))$. 可见,对于小规模或中等规模的决策表而言,算法 1 可快速有效地求得核.

4.3　基于广义差别矩阵的属性约简

依据定义 3,属性约简的本质是使原整个属性集上可区分的对象,在约简属性子集 P 上也是可区分的. 对于含有混合属性的决策表,定义 7 给出一种新的一致(不一致) 对象定义,该定义下的属性约简是要寻找一个属性子集 P,使得 C- 广义一致对象也是 P- 广义一致的. 对于任意 $P^1 \subset P$,存在对象 x 和 y 是 C- 广义一致的,但 x 和 y 是 P^1- 广义不一致的.

为此,本文引入如下定理:

定理 2　对于一个含有混合属性的决策表 DT,给定参数 $\delta(\delta > 0)$,得到的广义差别矩阵为 M_3. 对于 $P \subseteq C$, $\forall m_{ij} \in M(m_{ij} \neq \varnothing)$,有 $m_{ij} \cap P \neq \varnothing$成立;对于 $\forall a \in P$, $\exists m_{ij} \in M(m_{ij} \neq \varnothing)$,使得 $m_{ij} \cap (P - \{a\}) = \varnothing$. 则称 P 为一个属性约简.

证明　对于 $P \subseteq C$, $\forall m_{ij} \in M(m_{ij} \neq \varnothing)$,有 $m_{ij} \cap P \neq \varnothing$. 对于任意两个对象 x 和 y,若 x 和 y 是 C-广义一致的,则 x 和 y 也是 P-广义一致的.

同时, $\forall a \in P$, $\exists m_{ij} \in M(m_{ij} \neq \varnothing)$,使得 $\exists m_{ij} \cap (P - \{a\}) = \varnothing$. 对于任意 $P^1 \subset P$, $P^1 = P - \{a\}$, $a \in P$,存在对象 x 和 y 是 P-广义一致的,但 x 和 y 是 P^1-广义不一致的. 因此 P 是一个属性约简. □

定理 2 为广义差别矩阵下的属性约简提供了理论保证. 从核出发的启发式属性约简算法可参见相关文献(如[10]),这里仅给出基于差别函数求解属性约简算法[3,4]. 算法具体描述如下:

算法 2　基于广义差别矩阵的属性约简

输入:1) 给定决策表 $DT = \langle U, C \cup D, V, f \rangle$, $C = C^1 \cup C^2$;

2) 给定可调参数 $\delta(\delta > 0)$.

输出:一个属性约简 P.

Begin

1) 由算法 1 建立差别矩阵 M_3,并得到核 Core(C);

2) P = Core(C);// 属性约简初始化

3) 删除 M_3 中与 P 交不空的所有非空元素;

4) set $= \{m_{ij} \mid m_{ij}(m_{ij} \neq \varnothing) \in M_3$,且不存在 $m_{kl}(m_{kl} \neq \varnothing) \in M_3$,使 $m_{kl} \subset m_{ij}\}$;

5) $F = \varnothing$; for each $m \in$ set do $F = F \wedge m^*$; // 把 m 中的属性看成变量, m^* 为这些变量的并

6) 化简 F 使其具有 $m_1 \vee m_2 \vee \cdots \vee m_s$ 的形式,其中 $m_i(i = 1, 2, \cdots, s)$ 为属性变量的合取;

7) 取出某个满意的 P = Core(C) $\cup m_j$ 作为属性约简;

End

算法 2 的时间主要由建立广义差别矩阵及求核时间和通过求解差别函数得到属性约简两部分组成,其空间复杂度为 $O(m * \mathrm{card}(U_1) * \mathrm{card}(U))$. 进一步,采用删除广义差别矩阵中非空元素的所有超集,仅存储不包括超集的那些非空元素策略,可有效地降低广义差别矩阵的存储代价. 在实际应用中,用户可根据问题的规模,选择相应的属性约简算法.

4.4　示例说明

二值数据表如表1所示,其中共有5个对象和5个属性, $C = \{C_1, C_2, C_3, C_4\}$ 为条件属性集, $D = \{d\}$ 为决策属性.

表 1　二值数据表

元素	属性				
	C_1	C_2	C_3	C_4	d
x_1	a	0	0	0	0
x_2	b	0.1	0.1	0.2	0
x_3	b	0.1	0.1	0.24	1
x_4	a	0.2	0	0.4	1
x_5	a	1	1	1	1

由算法 2 可得第 2～4 列属性的均值分别为 0.28,0.24,0.37,第 2～4 列属性的标准方差分别为 0.41,0.43,0.38. 若设参数 $\delta = 0.25$,则 $U_1 = \{x_1, x_4, x_5\}$, $U_2 = \{x_2, x_3\}$. 由定义 8 可得广义差别矩阵 M_3 如下:

$$
\begin{array}{c|ccccc}
 & x_1 & x_4 & x_5 & x_2 & x_3 \\
\hline
x_1 & \varnothing & \{C_2, C_4\} & \{C_2, C_3, C_4\} & \varnothing & \{C_4\} \\
x_4 & \{C_2, C_4\} & \varnothing & \varnothing & \{C_1, C_4\} & \varnothing \\
x_5 & \{C_2, C_3, C_4\} & \varnothing & \varnothing & \{C_1, C_2, C_3, C_4\} & \varnothing
\end{array}.
$$

当参数 $\delta = 0.25$ 时,由定理 1 可得示例的核为 $\{C_4\}$. 由定理 2 可知, $\{C_4\}$ 为示例的属性约简.

若设参数 $\delta = 0.2$,则 $U_1 = \{x_1, x_4, x_5\}$, $U_2 = \{x_2, x_3\}$. 由定义 8 可得广义差别矩阵 M_3 如下:

$$
\begin{array}{c|ccccc}
 & x_1 & x_4 & x_5 & x_2 & x_3 \\
\hline
x_1 & \varnothing & \{C_2, C_4\} & \{C_2, C_3, C_4\} & \varnothing & \{C_2, C_3, C_4\} \\
x_4 & \{C_2, C_4\} & \varnothing & \varnothing & \{C_1, C_2, C_3, C_4\} & \varnothing \\
x_5 & \{C_2, C_3, C_4\} & \varnothing & \varnothing & \{C_1, C_2, C_3, C_4\} & \varnothing
\end{array}.
$$

因此,当参数 δ 由 0.25 调整到 0.2 时,由算法 1 可知示例的核为 $\varnothing$. 由算法 2 可得 $\{C_2, C_4\}$ 为示例的属性约简.

由示例可见,通过参数 δ 的适当调整,可有效控制对象在各属性值上的区分粒度,从而得到不同的属性约简. 在实际应用中,参数 δ 的取值可通过参数学习得到. 这方面的工作和相应的实验验证将另文讨论.

5　结　　论

本文引入广义差别矩阵,提出基于广义差别矩阵的核和属性约简算法. 该框架可直接对含有混合属性的决策表进行核和属性约简求解,有利于与其他机器学习方法相结合,进而为含有混合属性的决策表的核和属性约简提供了一条新的途径.

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] 刘清. Rough 集及 Rough 推理[M]. 北京: 科学出版社, 2001.
(Liu Q. Rough sets and rough reasoning[M]. Beijing: Science Press, 2001.)
[4] Jensen R, Shen Q. Semantics-preserving dimensionality reduction: Rough and fuzzy-rough-based approaches[J]. IEEE Trans on Knowledge and Data Engineering, 2004, 16(12): 1457-1471.
[5] 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]. Acta Electronica Sinica, 2002, 30(7): 1086-1088.)
[6] 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.)
[7] 刘少辉, 盛秋戬, 吴斌, 等. Rough 集高效算法的研究[J]. 计算机学报, 2003, 26(5): 524-529.
(Liu S H, Sheng Q J, Wu B, et al. Research on efficient algorithms for rough set methods[J]. Chinese J of Computers, 2003, 26(5): 524-529.)
[8] Woblewski J. Finding minimal reducts using genetic algorithm[R]. Warsaw: Warsaw University of Technology, 1995.
[9] 杨明. 一种基于改进差别矩阵的属性约简增量式更新算法[J]. 计算机学报, 2007, 30(5): 815-822.
(Yang M. An incremental updating algorithm for attribute reduction based on the improved discernibility matrix[J]. Chinese J of Computers, 2007, 30(5): 815-822.)
[10] 叶东毅. Jelonek 属性约简算法的一个改进[J]. 电子学报, 2000, 28(12): 81-82.
(Ye D Y. An improvement to Jelonek's attribute reduction algorithm[J]. Acta Electronica Sinica, 2000, 28(12): 81-82.)
[11] Yang M, Chen S C, Yang X B. A novel approach of rough set-based attribute reduction using fuzzy discernibility matrix[C]. FSKD'07. Haikou, 2007, 3: 96-101.
[12] Janssens D, Brijs T, Vanhoof K, et al. Evaluating the performance of cost-based discretization versus entropy- and error-based discretization[J]. Computers & Operations Research, 2006, 33(11): 3107-3123.

(上接第 1048 页)

6 结 论

时变参数系统切换控制方式易于实现,具有较好的鲁棒性和实时性,且可承受一定的切换延迟,而不影响整个系统的性能.然而,相对于具有自学习特征的控制方式,比如自适应控制,本文方法不具备智能化的特征,结论也偏于保守.这些问题有待于进一步研究.

参考文献(References)

[1] Apkarian P, Gahinet P. A convex characterization of gain-scheduling H-infinity controllers[J]. IEEE Trans on Automatic Control, 1995, 40(5): 853-864.
[2] Stilwell D J, Rugh W J. Stability and L_2 gain properties of LPV systems[J]. Automatica, 2002, 38(6): 1601-1606.
[3] Wu F, Yang X H, Packard A K, et al. Induced L_2 norm control for LPV systems with bounded parameter variation rates[J]. Int J of Robust Nonlinear Control, 1996, 6(9): 983-998.
[4] Wu F, Prajna S. A new solution approach to polynomial LPV system analysis and synthesis[C]. Proc of American Control Conf. Boston, 2004: 1362-1367.
[5] Lu B, Wu F. Switching LPV control designs using multiple parameter-dependent Lyapunov functions[J]. Automatica, 2004, 40(11): 1973-1980.
[6] Bett C J, Lemmon M D. Bounded amplitude performance of switched LPV systems with application to hybrid systems[J]. Automatica, 1999, 35(3): 491-503.
[7] Sun Z, Ge S. Switched linear systems: Control and design[M]. London: Springer-Verlag, 2005.
[8] Zhai G, Hu B, Yasuda K, et al. Disturbance attenuation properties of time-controlled switched systems[J]. J of Franklin Institute, 2001, 338(7): 765-779.
[9] Chilali M, Gahinet P. H-infinity design with pole placement constraints: An LMI approach[J]. IEEE Trans on Automatic Control, 1996, 41(3): 358-367.