

# 基于矩阵策略的不完备混合型数据增量式特征选择算法

沈玉峰, 林 徐

(安徽三联学院计算机工程学院, 合肥 230601)

**摘要:**特征选择是粗糙集理论在数据挖掘等领域中一种重要的应用, 如何对动态变化的信息系统进行增量式特征选择是目前粗糙集理论研究的重点。在不完备混合型信息系统中, 属性集的不断增长是信息系统动态变化的一种重要形式。首先在不完备混合型信息系统中引入邻域条件熵的概念, 并且利用矩阵的方法去表示邻域条件熵; 然后针对属性集动态增加的情形, 提出矩阵形式的邻域条件熵增量式更新, 并且基于这种增量式更新机制给出了相应的增量式特征选择算法; 最后, UCI数据集的实验结果表明, 所提出的增量式特征选择算法比非增量式特征选择算法具有更高的特征选择性能。

**关键词:**粗糙集; 特征选择; 不完备混合型信息系统; 矩阵; 邻域条件熵; 增量式学习

**中图分类号:** TP18      **文献标志码:** A      **文章编号:** 1673-1891(2020)01-0071-08

## Incremental Feature Selection Algorithm for Incomplete Mixed Data Based on Matrix Strategy

SHEN Yufeng, LIN Xu

(School of Computer Engineering, Anhui Sanlian University, Heifei 230601, China)

**Abstract:** Feature selection is an important application of rough set theory in data mining and other fields. How to make incremental feature selection for dynamic information systems is the focus of rough set theory research at present. In incomplete mixed information system, the increasing attribute set is an important form of dynamic change of information system. First, the concept of neighborhood conditional entropy is introduced into incomplete mixed information system, and is represented by matrix method. Then, in view of the dynamic increase of attribute set, an incremental updating method based on matrix form of neighborhood conditional entropy is proposed, and an incremental feature selection algorithm is given based on this incremental updating mechanism. Finally, the experimental results on UCI datasets show that the proposed incremental feature selection algorithm has higher feature selection performance than the non-incremental feature selection algorithm does.

**Keywords:** rough set; feature selection; incomplete mixed information system; matrix; neighborhood conditional entropy; incremental learning

## 0 引言

特征选择<sup>[1]</sup>, 又称为属性约简, 是粗糙集理论<sup>[2]</sup>在机器学习领域中一项重要的应用。在粗糙集理论中, 传统的特征选择算法只能处理静态的信息系统, 由于现实中的很多信息系统是不断动态变化的, 因此传统的算法对其进行特征选择时将会产生巨大的时间消耗<sup>[3]</sup>, 不利于特征选择的实际运用。

为了克服这一局限性, 一种新形式的特征选择方法被提出, 即增量式特征选择<sup>[4-6]</sup>。信息系统的动态变化有多种形式, 属性的逐渐增加便是其

中常见的一种。目前, 学者们提出了多种关于属性增加时的增量式特征选择算法<sup>[7-9]</sup>。然而, 这些算法只能处理完备型的信息系统, 现实中有些动态变化的信息系统是不完备类型的。为了进行改进, Kryszkiewicz<sup>[10]</sup>通过对传统粗糙集进行扩展和改进, 弥补了这一缺陷。针对属性集不断增加的不完备信息系统, 学者们同样提出了一些增量式特征选择算法, 如王映龙等<sup>[11]</sup>提出了不完备信息系统下的增量式特征选择算法; 丁棉卫等<sup>[12]</sup>利用区分矩阵提出了不完备信息系统的增量式特征选择; Shu等<sup>[13]</sup>提出了粗糙集理论中正区域的

增量式更新,并利于正区域去构造不完备信息系统的增量式特征选择。

然而在实际应用中,离散型和连续型并存的不完备数据普遍存在,传统的增量式特征选择算法很少有对这种类型的数据进行处理,因此本文将对其进行研究。条件熵作为一种重要的属性集评估方法<sup>[14]</sup>,目前已广泛应用于不完备信息系统的非增量式特征选择中<sup>[15-17]</sup>。Zhao 等<sup>[17]</sup>证明了邻域条件熵在不完备混合型数据的特征选择方面具有更高的优越性<sup>[17]</sup>。但是,针对属性增加的不完备混合型信息系统,未有学者去利用邻域条件熵去构造增量式特征选择,这使得考虑运用邻域条件熵的方法来构造不完备混合型信息系统的增量式特征选择算法。

矩阵是粗糙集理论中数据分析的一种常用工具<sup>[6,18-20]</sup>。本文将运用矩阵方法去构造基于邻域条件熵的增量式特征选择算法。首先运用矩阵的方法去表示邻域条件熵,并给出一种基于矩阵的邻域条件熵增量式更新方法;然后在邻域条件熵增量式更新的基础上,提出不完备混合型信息系统下的增量式特征选择算法;最后进行一系列的仿真实验来验证本文所提算法的性能。

### 1 基本理论

#### 1.1 不完备混合型信息系统的粗糙集模型

在粗糙集理论<sup>[2]</sup>中,结构化的数据集被描述成信息系统的形式,一个不完备混合型信息系统可表示为  $IS=(U,AT)$ ,其中  $U$  为非空有限对象集,也称为论域, $AT$  为非空有限属性集,并且其中离散型属性和连续型属性并存。对于  $x \in U$  在属性  $a \in AT$  下的属性值表示为  $a(x)$ 。当属性集  $AT=C \cup D, C \cap D = \emptyset$  时,此信息系统又被称为决策信息系统(DIS),这里的  $C$  和  $D$  分别称为条件属性集和决策属性集。

由于现实环境下,信息系统中某些属性值可能是缺失的,这类信息系统被称为不完备信息系统。设不完备混合型信息系统为  $IIS=(U,AT)$ ,缺失的属性值用“\*”表示,即  $\exists x \in U, a \in AT$  满足  $a(x)=*$ 。同样,当属性集  $AT=C \cup D, C \cap D = \emptyset$ ,那么这类信息系统称为不完备决策信息系统(IDIS)。

传统的粗糙集模型通过等价关系构建<sup>[2]</sup>,而对于不完备混合型信息系统,Zhao<sup>[17]</sup>等提出一种邻域容差关系的扩展粗糙集模型。

定义 1<sup>[17]</sup> 对于不完备混合型信息系统  $IIS=(U,AT)$ ,属性子集  $A \subseteq AT$ ,并且  $A=A_D \cup A_N$ ,其中  $A_D$  和  $A_N$  分别为属性  $A$  中的离散型属性和连续型属性,定义

属性集  $A$  在论域  $U$  上确定的邻域容差关系  $N_A^\delta$  为

$$N_A^\delta = \{(x, y) \in U \times U \mid \forall a \in A, (a(x) = *) \vee (a(y) = *) \vee ((a \in A_D \rightarrow a(x) = a(y)) \wedge (a \in A_N \rightarrow |a(x) - a(y)| \leq \delta))\},$$

其中,  $\delta$  为邻域半径,是一个非负常数。另外,称  $n_A^\delta(x)$  为对象  $x \in U$  在论域上的邻域容差类,定义为  $n_A^\delta(x) = \{y \mid (x, y) \in N_A^\delta\}$ 。

定义 2<sup>[17]</sup> 对于不完备混合型信息系统  $IIS=(U,AT)$ ,混合型属性子集  $A \subseteq AT$  确定的邻域容差关系为  $N_A^\delta$ ,对于  $X \subseteq U$  关于邻域容差关系  $N_A^\delta$  的下近似集和上近似集分别定义为

$$\underline{apr}_A^\delta(X) = \{x \in U \mid n_A^\delta(x) \subseteq X\},$$

$$\overline{apr}_A^\delta(X) = \{x \in U \mid n_A^\delta(x) \cap X \neq \emptyset\}.$$

#### 1.2 不完备混合型信息系统的邻域条件熵特征选择

信息系统的熵模型是目前粗糙集理论以及粒计算理论的研究热点。在不完备混合型信息系统中,Zhao 等<sup>[17]</sup>提出了基于邻域容差关系的条件熵模型。

定义 3<sup>[17]</sup> 对于不完备混合型信息系统  $IIS=(U,AT)$ ,  $|U|=n$ ,混合型属性子集  $A \subseteq AT$  在论域  $U$  上确定的邻域容差关系为  $N_A^\delta$ ,  $n_A^\delta(x_i)$  为对象  $x_i \in U$  在  $N_A^\delta$  下的邻域容差类。那么属性集  $A$  在论域  $U$  下的邻域信息熵定义为

$$NE_\delta(A) = \frac{1}{n} \sum_{i=1}^n (1 - \frac{|n_A^\delta(x_i)|}{n}),$$

其中,  $||$  表示集合的基数。定义 3 所示的邻域信息熵满足  $0 \leq NE(A) \leq 1 - 1/n$ 。

定义 4<sup>[17]</sup> 对于不完备混合型信息系统  $IIS=(U,AT)$ ,  $|U|=n$ ,混合型属性子集  $A_1, A_2 \subseteq AT$  在论域  $U$  上确定的邻域容差关系分别为  $N_{A_1}^\delta$  和  $N_{A_2}^\delta$ ,  $n_{A_1}^\delta(x_i)$  和  $n_{A_2}^\delta(x_i)$  分别为  $x_i \in U$  在  $N_{A_1}^\delta$  和  $N_{A_2}^\delta$  下的邻域容差类。那么  $A_2$  关于  $A_1$  的邻域条件熵定义为

$$NCE_\delta(A_2|A_1) = \frac{1}{n} \sum_{i=1}^n \left( \frac{|n_{A_1}^\delta(x_i)|}{n} - \frac{|n_{A_1}^\delta(x_i) \cap n_{A_2}^\delta(x_i)|}{n} \right).$$

特征选择是信息系统熵模型的一种重要的应用,基于信息熵的度量,学者提出了大量的特征选择算法<sup>[11,15-17]</sup>。对于不完备混合型信息系统的信息熵模型,目前也有相应的特征选择算法被提出<sup>[17]</sup>。

定义 5<sup>[17]</sup> 对于不完备混合型决策信息系统  $IDIS=(U,C \cup D)$ ,若  $A \subseteq C$  是该信息系统的邻域条件熵属性约简集当且仅当:

$$NCE_\delta(D|C) = NCE_\delta(D|A), \tag{1}$$

$$\forall a \in A, NCE_\delta(D|C) < NCE_\delta(D|A - \{a\}). \tag{2}$$

在定义 5 中,式(1)保证属性子集  $A$  与条件属性集  $C$  的分类能力一致,式(2)确保属性子集的极小

性,同时满足这2个条件时称  $A$  为该不完备混合型决策信息系统的邻域条件熵特征选择,即属性全集  $C$  的特征子集。算法1所示的是对应的特征选择算法。

**算法1**<sup>[17]</sup>:基于邻域条件熵的不完备混合型信息系统特征选择算法。

输入:不完备混合型决策信息系统  $IDIS=(U, C \cup D)$ ,  $|U|=n, |C|=c$ , 邻域半径  $\delta$ ;

输出: $IDIS$ 的约简集  $red_c$ 。

**Step1** 初始化  $red_c = \emptyset$ 。

**Step2** 对于  $\forall a \in C$ , 如果

$$NCE_\delta(D|C - \{a\}) - NCE_\delta(D|C) > 0,$$

那么  $red_c = red_c \cup \{a\}$ 。

**Step3** 如果  $NCE_\delta(D|red_c) \neq NCE_\delta(D|C)$ , 那么进入 **Step4** 否则进入 **Step5**。

**Step4** 对于  $\forall a \in C - red_c$ , 计算

$$a_{\max} = \arg \max_{a \in C - red_c} (NCE_\delta(D|red_c) - NCE_\delta(D|red_c \cup \{a\}))$$

并且  $red_c \leftarrow red_c \cup \{a_{\max}\}$ 。判断是否满足 **Step3**

中的条件,如果不满足进入 **Step5**。

**Step5** 对于  $\forall a \in red_c$ , 若满足于

$$NCE_\delta(D|red_c - \{a\}) = NCE_\delta(D|C),$$

那么  $red_c = red_c - \{a\}$ 。

**Step6** 返回  $red_c$ 。

在算法1中, **Step2~4** 是启发式搜索属性的过程, **Step5** 是约简集反向剔除冗余属性的过程, 整个算法1的时间复杂度为  $O(c^2 \cdot n^2)$ <sup>[17]</sup>。

## 2 邻域条件熵的增量式特征选择算法

实际应用中的信息系统时刻处于变化过程之中,为了改善特征选择效率,学者们提出了增量式特征选择用来提高动态信息系统的约简性能。在本节,针对不完备混合型信息系统中属性集增加的情形,提出邻域条件熵的增量式更新方法,并根据这种方法在算法1的基础上设计出一种增量式特征选择算法。

### 2.1 属性集增加时邻域条件熵的增量式更新

矩阵是一种重要的数学挖掘工具,目前已广泛运用于粗糙集理论的研究,同时也是进行增量式学习的有效方法<sup>[6,18-20]</sup>。在本节将提出一种矩阵方法的邻域条件熵增量式更新。

**定义6** 对于不完备混合型信息系统  $IIS=(U, AT)$ ,  $|U|=n$ , 混合型属性子集  $A \subseteq AT$  确定的邻域容差关系为  $N_A^\delta$ , 定义  $N_A^\delta$  的邻域容差关系矩阵为  $N_A^\delta = (m_{ij}^A)_{n \times n}$ , 其中

$$m_{ij}^A = \begin{cases} 1, & (x_i, x_j) \in N_A^\delta \\ 0, & (x_i, x_j) \notin N_A^\delta \end{cases} \quad x_i, x_j \in U.$$

定义6表明,如果对象  $x_i, x_j$  满足邻域容差关系  $N_A^\delta$ , 那么邻域容差关系矩阵  $N_A^\delta$  中第  $i$  行第  $j$  列的元素为1, 否则为0。在不引起混淆情形下, 下文中间域容差关系矩阵简称为关系矩阵。

**例1** 表1所示的是一个不完备混合型信息系统, 其中  $U = \{x_1, x_2, x_3, x_4, x_5, x_6\}$ ,  $AT = \{a, b, c\}$ , 令  $A_1 = \{b, c\}$ ,  $A_2 = \{a\}$ , 其中  $\{a, c\}$  为离散型属性集,  $\{b\}$  为连续型属性集, “\*”表示缺失的属性值。

表1 不完备混合型信息系统

$U$	$a$	$b$	$c$
$x_1$	0	*	Y
$x_2$	1	0.42	Y
$x_3$	1	*	N
$x_4$	1	0.67	N
$x_5$	*	0.49	Y
$x_6$	0	0.72	*

根据定义6, 设邻域半径  $\delta = 0.1$ , 那么有

$$N_{A_1}^{0.1} = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 \end{bmatrix}, \quad N_{A_2}^{0.1} = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}.$$

同时可以看出  $N_{A_1}^{0.1}$  和  $N_{A_2}^{0.1}$  均为对称矩阵, 且主对角线元素都为1。

为了下文讨论的方便, 这里定义关系矩阵的4种运算。

**定义7** 设关系矩阵  $N_P^\delta = (m_{ij}^P)_{n \times n}$  和  $N_Q^\delta = (m_{ij}^Q)_{n \times n}$ , 定义

$$N_P^\delta - N_Q^\delta = \begin{cases} 0, & m_{ij}^P = 1 \wedge m_{ij}^Q = 1, \\ m_{ij}^P, & \text{其他} \end{cases}$$

$$N_P^\delta \wedge N_Q^\delta = \begin{cases} 1, & m_{ij}^P = 1 \wedge m_{ij}^Q = 1, \\ 0, & \text{其他} \end{cases}$$

$$N_P^\delta \otimes N_Q^\delta = \begin{cases} 1, & m_{ij}^P = 1 \wedge m_{ij}^Q = 0, \\ 0, & \text{其他} \end{cases}$$

$$|N_P^\delta| = \sum_{i=1}^n \sum_{j=1}^n m_{ij}^P.$$

**引理1** 设  $S_1, S_2$  是2个集合, 那么满足  $|S_1 \cap S_2| = |S_1| - |S_1 - S_2|$ 。

根据定义7中矩阵的4种定义, 可以得到如下推论。

**推论1** 设关系矩阵  $N_P^\delta = (m_{ij}^P)_{n \times n}$  和  $N_Q^\delta = (m_{ij}^Q)_{n \times n}$ , 那么

$$N_{P \cup Q}^\delta = N_P^\delta \wedge N_Q^\delta, \tag{3}$$

$$|N_P^\delta \wedge N_Q^\delta| = |N_P^\delta| - |N_P^\delta \otimes N_Q^\delta|. \tag{4}$$

证明:根据定义 1 有

$$N_{P \cup Q}^\delta = \bigcap_{a \in P \cup Q} N_{|a|}^\delta = \left( \bigcap_{a \in P} N_{|a|}^\delta \right) \cap \left( \bigcap_{a \in Q} N_{|a|}^\delta \right),$$

即  $N_{P \cup Q}^\delta = N_P^\delta \cap N_Q^\delta$ 。若  $(x, y) \in N_{P \cup Q}^\delta$ , 则  $(x, y) \in N_P^\delta$  且  $(x, y) \in N_Q^\delta$ , 因此根据定义 6 有  $N_{P \cup Q}^\delta = N_P^\delta \wedge N_Q^\delta$ , 式(3)证明完毕。

对于  $(x_i, x_j) \in U \times U$ , 如果  $(N_P^\delta \otimes N_Q^\delta)_{ij} = 1$ , 根据定义 7 有  $(x_i, x_j) \in N_P^\delta \wedge (x_i, x_j) \notin N_Q^\delta$ , 即  $(x_i, x_j) \in N_P^\delta - N_Q^\delta$ 。若  $(x_i, x_j) \in N_P^\delta - (N_P^\delta - N_Q^\delta)$ , 那么  $(N_P^\delta - (N_P^\delta - N_Q^\delta))_{ij} = 1$ , 根据引理 1, 即  $N_P^\delta - (N_P^\delta - N_Q^\delta) = N_P^\delta \cap N_Q^\delta$ , 所以

$$N_P^\delta - (N_P^\delta \otimes N_Q^\delta) = N_P^\delta \wedge N_Q^\delta。$$

同样根据引理 1, 便可以得到  $|N_P^\delta \wedge N_Q^\delta| = |N_P^\delta| - |N_P^\delta \otimes N_Q^\delta|$ 。式(4)证明完毕。

证毕。

例 2 在例 1 的基础上, 根据定义 6 有

$$N_{A_1 \cup A_2}^{0.1} = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

根据例 1, 满足关系  $N_{A_1 \cup A_2}^{0.1} = N_{A_1}^{0.1} \wedge N_{A_2}^{0.1}$ 。

$$N_{A_1}^{0.1} \otimes N_{A_2}^{0.1} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \end{bmatrix},$$

因此有

$$|N_{A_1}^{0.1}| - |N_{A_1}^{0.1} \otimes N_{A_2}^{0.1}| = 20 - 6 = |N_{A_1}^{0.1} \wedge N_{A_2}^{0.1}|。$$

接下来将应用关系矩阵的方法来表示邻域条件熵。

**定理 1** 对于不完备混合型信息系统  $IIS=(U, AT)$ ,  $|U|=n$ , 并且混合型属性子集  $A_1, A_2 \subseteq AT$ , 它们确定的邻域容差关系分别为  $N_{A_1}^\delta$  和  $N_{A_2}^\delta$ , 对应的关系矩阵分别为  $N_{A_1}^\delta = (m_{ij}^{A_1})_{n \times n}$  和  $N_{A_2}^\delta = (m_{ij}^{A_2})_{n \times n}$ , 那么  $A_2$  关于  $A_1$  的邻域条件熵可表示为

$$NCE(A_2|A_1) = \frac{1}{n^2} \cdot (|N_{A_1}^\delta| - |N_{A_1}^\delta \wedge N_{A_2}^\delta|)。$$

证明:根据定义 4 有

$$NCE(A_2|A_1) =$$

$$\frac{1}{n^2} \sum_{i=1}^n (|n_{A_1}^\delta(x_i)| - |n_{A_1}^\delta(x_i) \cap n_{A_2}^\delta(x_i)|) =$$

$$\frac{1}{n^2} \left( \sum_{i=1}^n |n_{A_1}^\delta(x_i)| - \sum_{i=1}^n |n_{A_1}^\delta(x_i) \cap n_{A_2}^\delta(x_i)| \right),$$

$$\text{记 } \xi_1 = \sum_{i=1}^n |n_{A_1}^\delta(x_i)|, \xi_2 = \sum_{i=1}^n |n_{A_1}^\delta(x_i) \cap n_{A_2}^\delta(x_i)|,$$

那么对于关系矩阵  $N_{A_1}^\delta$ , 第  $i$  行的行向量表示为  $n_i^{A_1}$ , 根据定义 6 可以得到

$$|n_{A_1}^\delta(x_i)| = \sum_{j=1}^n n_i^{A_1}(j)。$$

这里的  $n_i^{A_1}(j)$  表示向量  $n_i^{A_1}$  第  $j$  个元素。

$$\text{则 } \xi_1 = \sum_{i=1}^n \sum_{j=1}^n n_i^{A_1}(j) = |N_{A_1}^\delta|, \text{同理 } \xi_2 = |N_{A_1}^\delta \wedge N_{A_2}^\delta|, \text{即}$$

$$NCE(A_2|A_1) = \frac{1}{n^2} (|N_{A_1}^\delta| - |N_{A_1}^\delta \wedge N_{A_2}^\delta|)。$$

证毕。

根据定理 1 中基于关系矩阵的邻域条件熵表示方法, 下面给出基于关系矩阵的邻域条件熵增量式更新方法。当不完备混合型信息系统的属性集增加时, 这种增量更新方法可以根据原先的邻域条件熵快速地计算出信息系统更新后的邻域条件熵, 而不需要根据新的不完备混合型信息系统对邻域条件熵进行重新计算, 这样可以大大提高运算效率。

**定理 2** 对于不完备混合型信息系统  $IIS=(U, AT)$ ,  $|U|=n$ , 混合型属性子集  $A_1, A_2 \subseteq AT$ , 它们对应的关系矩阵分别为  $N_{A_1}^\delta = (m_{ij}^{A_1})_{n \times n}$  和  $N_{A_2}^\delta = (m_{ij}^{A_2})_{n \times n}$ ,  $A_2$  关于  $A_1$  的邻域条件熵为  $NCE(A_2|A_1)$ 。假设一个新的属性集  $\Delta A$  加入信息系统, 新的不完备混合型信息系统表示为  $IIS'=(U, AT')$ , 这里的  $AT'=AT \cup \Delta A$ 。  $\Delta A$  确定的关系矩阵为  $N_{\Delta A}^\delta$ , 令  $A_1' = A_1 \cup \Delta A$ , 那么  $A_2$  关于  $A_1'$  的邻域条件熵增量式更新为

$$NCE(A_2|A_1') = NCE(A_2|A_1) - \Delta,$$

$$\text{这里的 } \Delta = \frac{1}{n^2} (|N_{A_1}^\delta \otimes N_{\Delta A}^\delta| - |(N_{A_1}^\delta \wedge N_{A_2}^\delta) \otimes N_{\Delta A}^\delta|)。$$

证明:根据定理 1 有

$$NCE(A_2|A_1') = \frac{1}{n^2} (|N_{A_1'}^\delta| - |N_{A_1'}^\delta \wedge N_{A_2}^\delta|),$$

根据推论 1 有

$$N_{A_1'}^\delta = N_{A_1}^\delta \wedge N_{\Delta A}^\delta,$$

$$|N_{A_1}^\delta \wedge N_{\Delta A}^\delta| = |N_{A_1}^\delta| - |N_{A_1}^\delta \otimes N_{\Delta A}^\delta|。$$

$$\text{所以 } NCE(A_2|A_1') = \frac{1}{n^2} (|N_{A_1}^\delta| - |N_{A_1}^\delta \wedge N_{A_2}^\delta|) =$$

$$\frac{1}{n^2} (|N_{A_1}^\delta \wedge N_{\Delta A}^\delta| - |N_{A_1}^\delta \wedge N_{\Delta A}^\delta \wedge N_{A_2}^\delta|) =$$

$$\frac{1}{n^2} (|N_{A_1}^\delta| - |N_{A_1}^\delta \otimes N_{\Delta A}^\delta| - |N_{A_1}^\delta \wedge N_{A_2}^\delta| +$$

$$|(N_{A_1}^\delta \wedge N_{A_2}^\delta) \otimes N_{\Delta A}^\delta|)。$$

由于  $NCE(A_2|A_1) = \frac{1}{n^2} (|N_{A_1}^\delta| - |N_{A_1}^\delta \wedge N_{A_2}^\delta|)$ , 所以

$$NCE(A_2|A'_1) = NCE(A_2|A_1) - \Delta,$$

$$\text{其中, } \Delta = \frac{1}{n^2} (|N_{A_1}^\delta \otimes N_{\Delta A}^\delta| - |(N_{A_1}^\delta \wedge N_{A_2}^\delta) \otimes N_{\Delta A}^\delta|).$$

证毕。

在定理2中,由于在计算邻域条件熵  $NCE(A_2|A_1)$  时,已经计算出了关系矩阵  $N_{A_1}^\delta$  和  $N_{A_2}^\delta$ ,当不完备混合型信息系统新加入属性集  $\Delta A$  后,只需要计算属性集  $\Delta A$  的关系矩阵  $N_{\Delta A}^\delta$ ,按照定理2便可以得到新的邻域条件熵  $NCE(A_2|A'_1)$ 。因此定理2这种增量计算方法具有很高的效率。

**例3** 在例1中,假设信息系统增加一个新的连续型属性集  $\Delta A = \{d\}$ ,新的不完备混合型信息系统如表2所示。此时  $AT' = \{a, b, c, d\}$ ,令  $A'_1 = A_1 \cup \Delta A = \{b, c, d\}$ 。

表2 新的不完备混合型信息系统

$U$	$a$	$b$	$c$	$d$
$x_1$	0	*	Y	0.21
$x_2$	1	0.42	Y	0.27
$x_3$	1	*	N	*
$x_4$	1	0.67	N	0.85
$x_5$	*	0.49	Y	0.76
$x_6$	0	0.72	*	0.19

那么,根据例2可以得到

$$NCE(A_2|A_1) = \frac{1}{n^2} (|N_{A_1}^{0.1}| - |N_{A_1}^{0.1} \wedge N_{A_2}^{0.1}|) = \frac{20-14}{36} = \frac{6}{36}。$$

$$\text{由于 } N_{\Delta A}^{0.1} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 \end{bmatrix},$$

$$\text{所以 } N_{A_1}^{0.1} \otimes N_{\Delta A}^{0.1} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix},$$

$$(N_{A_1}^{0.1} \wedge N_{A_2}^{0.1}) \otimes N_{\Delta A}^{0.1} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}。$$

因此

$$\Delta = \frac{1}{n^2} (|N_{A_1}^{0.1} \otimes N_{\Delta A}^{0.1}| - |(N_{A_1}^{0.1} \wedge N_{A_2}^{0.1}) \otimes N_{\Delta A}^{0.1}|) = \frac{1}{36} (6-4) = \frac{2}{36},$$

$$\text{则 } NCE(A_2|A'_1) = NCE(A_2|A_1) - \Delta = \frac{6}{36} - \frac{2}{36} = \frac{4}{36}。$$

如果直接运用定理1去计算  $NCE(A_2|A'_1)$ ,那么需要计算关系矩阵  $N_{A'_1}^{0.1}$ ,最终结果和增量式计算的结果是一致的,但是采用定理1会产生较多的重复计算。

### 2.2 增量式特征选择算法

在算法1的基础上,将本文所提出的邻域条件熵增量式更新方法融入其中,可以构建出相应的增量式特征选择算法,具体如算法2所示。

**算法2:** 基于矩阵策略的不完备混合型信息系统邻域条件熵增量式特征选择算法。

输入:

1)新的不完备混合型决策信息系统  $IDIS' = (U, C' \cup D)$ ,新加入的属性集为  $\Delta C, C' = C \cup \Delta C, |U| = n, |C| = c, |C'| = c^+$ ,邻域半径  $\delta$ 。

2)邻域条件熵  $NCE(D|C)$  和  $NCE(D|red_c)$ ,关系矩阵  $N_c^\delta$  和  $N_{D'}^\delta$ ,不完备混合型决策信息系统  $IDIS$  约简集  $red_c, |red_c| = r$ 。

输出:  $IDIS'$  的特征约简集  $red_{C'}$ 。

**Step1** 根据定义6计算关系矩阵  $N_{\Delta C}^\delta$ 。

**Step2** 根据定理2进行增量计算新的邻域条件熵  $NCE(D|C')$ 。

**Step3**  $red_{C'} \leftarrow red_c$ ,若

$$NCE(D|C') = NCE(D|red_c)$$

那么进入Step5,否则进入Step4。

**Step4** 当  $NCE(D|C') \neq NCE(D|red_c)$  时,对于  $\forall \alpha \in (C' - red_c)$ ,采用定理2计算

$$a_{\max} = \arg \max_{\alpha \in C' - red_c} (NCE(D|red_c) - NCE(D|red_c \cup \{\alpha\})),$$

并进行  $red_{C'} \leftarrow red_{C'} \cup \{\alpha_{\max}\}$ 。

**Step5** 对于  $\forall \alpha \in red_c$ ,若

$$NCE(D|red_c - \{\alpha\}) = NCE(D|C'),$$

那么  $red_c = red_c - \{\alpha\}$ 。

**Step6** 返回  $red_{C'}$ 。

在算法2中,当不完备混合型信息系统属性增加时,由于信息系统在更新之前的邻域条件熵  $NCE(D|C')$  和  $NCE(D|red_c)$ 、关系矩阵  $N_c^\delta$  和  $N_{D'}^\delta$  都已经得到,因此在算法2的Step2中,只需要计算关系矩阵  $N_{\Delta C}^\delta$  便可以增量的计算  $NCE(D|C')$ ,其时间复杂度为  $O(c^+ \cdot n^2)$ ,如果不采用增量计算,那么计算

$NCE(D|C')$ 的时间复杂度为 $O((c+c^+) \cdot n^2)$ 。然后 Step3~4描述的是在未更新前信息系统约简集的基础上进一步启发式搜索属性,其中 Step4 仍然采用定理 2 的增量式更新计算方法,最后 Step5 对冗余属性进行剔除。整个算法 2 的时间复杂度为 $O((c+c^+ - r) \cdot n^2 + c^+ \cdot n^2)$ 。

### 3 实验分析

在本节将通过进行一系列实验来验证本文所提出的增量式特征选择算法的有效性和优越性。实验中采用 UCI 数据集进行实验,具体如表 3 所示。在这 6 个数据集中,有的是完备型的数据集,对于这类数据集,在实验前随机选取 5% 的属性值进行删除,从而构造出不完备的数据集。同时考虑数据集属性中量纲带来的影响,进行实验前将所有的连续型属性值标准化入[0,1]区间。实验所使用的硬件为 Windows 7 操作系统的个人电脑,Inter (R) Core (TM) i3-2350 (2.2GHz)处理器。算法采用 Java 进行编程实现。

表 3 实验数据集

编号	数据集名称	对象	属性
1	Mushroom	5 644	22
2	Chess	3 196	36
3	Cylinder	512	40
4	Sonar	208	60
5	Gearbox	1 603	72
6	Musk	6 598	166

#### 3.1 实验设置

本实验中,表 3 中的 6 个数据集均为静态数据

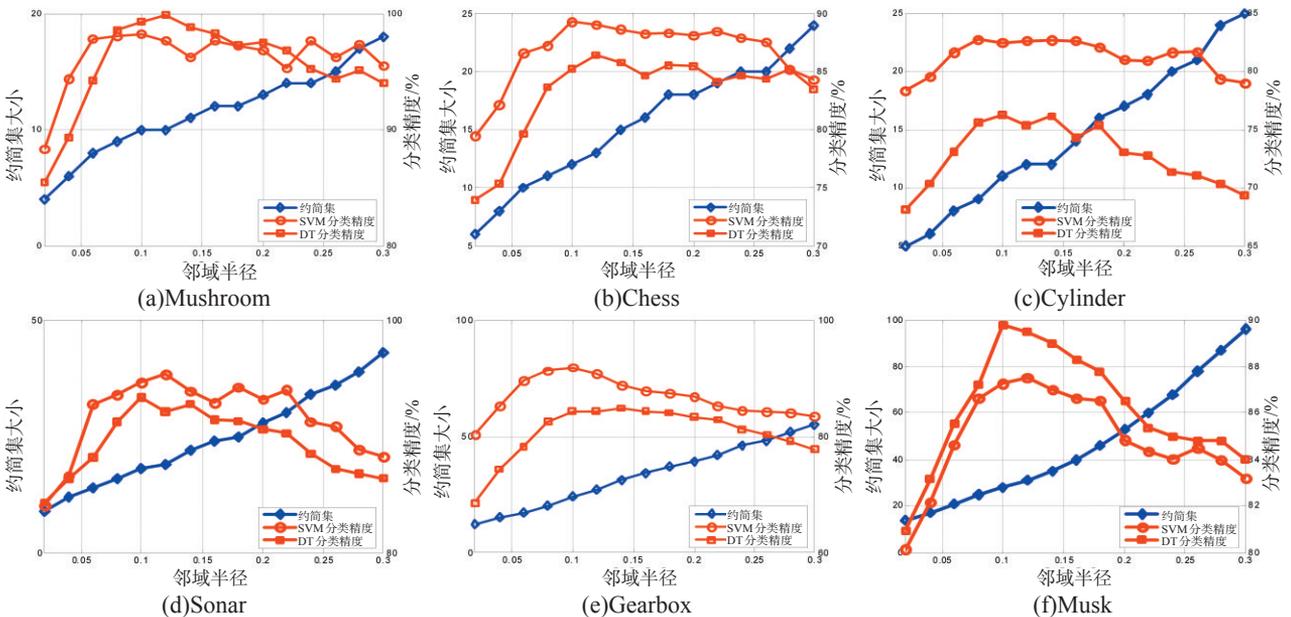


图 1 不同邻域下实验结果

集,为了模拟出数据集属性逐渐增加的过程,对每个数据集的条件属性集大致分成 7 个等份的属性子集,并将第 1 个属性子集所对应的数据集作为初始的不完备混合型信息系统,然后将剩余的属性子集逐个加入到这个不完备混合型信息系统中,这样便构造出了属性集 6 次动态增加的过程。

记算法 1 为非增量式特征选择算法,本文所提出的算法 2 为增量式特征选择算法。首先让算法 1 和算法 2 分别对表 3 中的 6 个动态变化的数据集进行特征选择,通过比较数据集每次更新时特征选择所需的时间来证明本文所提算法的高效性。经过特征选择,算法 1 和算法 2 会得到每个数据集的特征选择结果,这里分别通过约简集的大小,约简集的分类性能来比较 2 种算法的约简性能。从而验证本文的增量式特征选择算法具有一定的优越性。

在算法 1 和算法 2 中,均包含一个输入参数,即邻域半径 $\delta$ ,它的取值不同将对最终的实验结果产生很大的影响,为了选取合适的参数,本实验将邻域半径 $\delta$ 在区间[0.02,0.3]以 0.02 为步长分别进行增量式特征选择实验,每个邻域半径最终会得到对应的约简集,将该约简集在支持向量机分类器(SVM)和决策树分类器(DT)分别进行分类精度计算,所有的实验结果如图 1 所示。观察图 1 可以发现,当邻域半径 $\delta$ 取为 0.1 时,约简结果得到的分类精度最高,因此本实验采用 $\delta = 0.1$ 进行实验。

#### 3.2 特征选择的效率比较

图 2 所示的是 2 种算法在 6 个数据集下每次增量式更新时特征选择的用时比较结果,为了减少偶然性,所有的时间结果均是取多次实验的平均值,

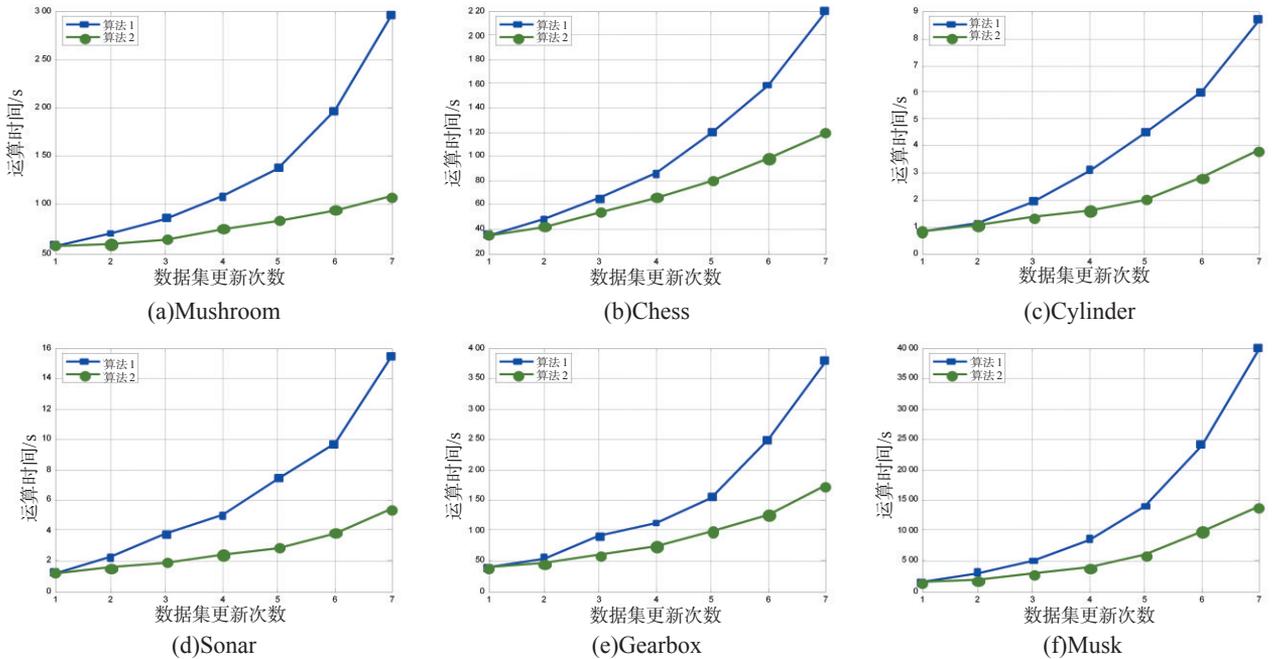


图2 算法1和算法2在各个数据集特征选择时间比较

时间单位为秒(s)。

通过图2可以看出,在数据集属性集前2次的动态增加时,2种算法的特征选择用时相差不大,但是随着更新次数的不断增多,这2种算法的特征选择用时逐渐拉开了差距,算法1的非增量式特征选择算法用时增长较快,而算法2的非增量式算法用时的较为缓慢,说明了对于属性集不断增加的信息系统,本文所提出的增量式算法比非增量式算法具有更高的特征选择效率,尤其是对于规模较大的信息系统,例如Mushroom, Gearbox和Musk。这主要是由于非增量式算法,计算动态不完备混合型信息系统的邻域条件熵进行了很多的重复计算,消耗了大量的时间,因此约简用时会偏高。而所提出的增量式特征选择算法,每次特征选择时只需对新加入的数据进行计算,不必对原来的数据进行重复计算,因此具有较高的特征选择效率。

### 3.3 特征选择集比较

表4所示的是算法1和算法2最终的特征选择结果比较,在表2中,首先可以很明显地发现约简后的属性集远小于整个不完备混合型信息系统的条件属性集,因此说明了信息系统中的冗余属性是普遍存在的,对其进行特征选择是很有必要的。对于算法1和算法2的特征选择结果,可以发现本文所提出的算法2在大部分数据集中具有更小的约简集,而算法1只有在少部分的数据集拥有更少的约简集,例如数据集Chess。这主要是由于这2种算法的运算机制不同导致的,对于算法1,采用的是非增

量式的形式进行约简,即每次对完整的数据集进行搜索属性,而本文所提的出算法2,采用增量式的方法进行属性搜索,即在原来约简集基础上对新加入的属性进行搜索,这样对重要的属性具有更好的鉴别能力,因此可以得到较少的约简结果。

表4 特征选择结果

数据集	原式属性集	算法1约简结果	算法2约简结果
Mushroom	22	12	10
Chess	36	11	12
Cylinder	40	14	11
Sonar	60	21	18
Gearbox	72	25	24
Musk	166	32	28

### 3.4 特征选择结果分类性能比较

特征选择的基本要求是选择出的属性子集不能降低信息系统的分类性能,因此接下来需要对这2种算法得到的约简结果进行分类性能比较。在机器学习中,利用分类器可以对属性子集进行十折交叉验证,可以得到该属性子集的分类精度,分类精度也是对属性子集分类性能的一种重要体现。表5和表6分别所示的是特征选择结果在支持向量机(SVM)和决策树(DT)2种分类器下的分类精度。

在表5所示的分类精度结果中,大部分数据集在算法2下的特征选择结果具更高的SVM分类精度,而算法1只在Cylinder数据集下具有较高的SVM分类精度。对于表6所示的DT分类精度,大部分数据集在算法2下特征选择的分类精度较高,

表5 特征选择结果的SVM分类精度 %

数据集	原式属性集	算法1约简结果	算法2约简结果
Mushroom	93.56	96.42	98.24
Chess	83.54	86.45	89.32
Cylinder	79.71	85.85	82.45
Sonar	90.49	92.34	94.65
Gearbox	77.14	88.63	91.84
Musk	75.48	84.79	87.26

表6 特征选择结果的DT分类精度 %

数据集	原式属性集	算法1约简结果	算法2约简结果
Mushroom	91.30	95.32	99.27
Chess	84.25	87.68	85.25
Cylinder	76.71	74.47	76.24
Sonar	88.49	90.12	93.35
Gearbox	79.14	85.53	84.24
Musk	76.48	85.74	89.79

算法1在数据集 Chess 和 Gearbox 下具有较高的分类精度。产生这种差距的主要原因可能是由于算法1和算法2的约简机制不同导致的,在算法1的非增量式特征选择中,对于每次更新后的信息系统,算法总是从空集开始进行启发式搜索属性,例如在算法1刚开始的时候,约简集中的候选属性较少,这可能会导致其它属性的属性重要度值偏大,而实际上这些属性重要度可能是偏小的。而对于算法2的增量式特征选择,算法对于更新后的信息系统进行

约简时,总是基于更新前的信息系统的约简集开始,这样对其它属性具有更好的鉴别能力,计算出的属性重要度就更加准确,这就解释了为什么增量式特征选择算法选择出的属性少且分类效果好的原因。

综合特征选择效率、特征选择结果以及特征选择结果的分类性能,可以证明本文所提出的增量式特征选择算法具有较高的特征选择性能,适用于动态型不完备混合型信息系统的特征选择问题。

### 4 结语

现实中的信息系统往往是不断动态变化的,如何对其进行有效的特征选择是目前机器学习和数据挖掘等领域的研究重点。属性集的不断增长是信息系统动态变化的一种常见情形,本文针对属性集不断增长的不完备混合型信息系统,利用矩阵方法研究了粗糙集理论中邻域条件熵随属性集增加的更新机制,并提出了相应的增量式特征选择算法。最后通过进行一系列的实验证明了本文所提出的增量式算法的有效性和优越性,文章的算法可以适用于属性增加时的不完备混合型信息系统特征选择问题。由于本文所研究的是属性集变化的增量式特征选择,因此未来将进一步研究对象集变化或对象集和属性集同时变化的信息系统增量式特征选择问题。

### 参考文献:

- [1] 郭阳阳,汤建国.大数据背景下粗糙集属性约简研究进展[J].计算机工程与应用,2019,55(6):31-38+177.
- [2] PAWLAK Z. Rough sets[J].International Journal Computer Information Science,1982,11(5):341-356.
- [3] CAI Mingjie, LANG Guangming, HAMIDO F, et al. Incremental approaches to updating reducts under dynamic covering granularity[J].Knowledge-Based Systems,2019,172:130-140.
- [4] SHU Wenhao, QIAN Wenbin, XIE Yonghong. Incremental approaches for feature selection from dynamic data with the variation of multiple objects[J].Knowledge-Based Systems,2019,163:320-331.
- [5] HU Chengxiang, ZHANG Li, WANG Bangjun, et al. Incremental updating knowledge in neighborhood multigranulation rough sets under dynamic granular structures[J].Knowledge-Based Systems, 2019,163:811-829.
- [6] 郑诚,王波,洪彤彤.关系矩阵的知识粒度增量式属性约简[J].小型微型计算机系统,2018,39(5):1000-1004.
- [7] RAZA M S, QAMAR U. An incremental dependency calculation technique for feature selection using rough sets[J].Information Sciences,2016,343-344:41-65.
- [8] LUO Chuan, LI Tianrui, CHEN Hongmei. Dynamic maintenance of approximations in set-valued ordered decision systems under the attribute generalization[J].Information Sciences,2014,257(1):210-228.
- [9] QIAN Wenbin, SHU Wenhao, YANG Bingru, et al. An incremental algorithm to feature selection in decision systems with the variation of feature set[J].Chinese Journal of Electronics,2015,24(1):128-133.
- [10] KRYSZKIEWICZ M. Rough set approach to incomplete information systems[J].Information Sciences,1998,112(1-4):39-49.
- [11] 王映龙,曾洪,钱文彬,等.变精度下不完备混合数据的增量式属性约简方法[J].计算机应用,2018,38(10):2764-2771.
- [12] 丁棉卫,张腾飞,马福民.基于二进制区分矩阵的不完备系统增量式属性约简算法[J].计算机科学,2017,44(7):244-250.

- [2] 董桂伟,赵国群,王桂龙,等.科研成果转化为实验教学资源的探索[J].实验技术与管理,2019,36(4):120-123
- [3] 曹睿洁,刘思迪,韩东.简单方法制备与表征新颖复合凝胶及表面微结构—由科研成果转化的化学综合实验[J].化学教育(中英文),2018,39(8):47-50.
- [4] 张安强,吴水珠,刘海敏,等.高分子化学探索性实验: $\alpha$ 、 $\omega$ -羧基聚二甲基硅氧烷的可控合成[J].化学教育,2016,37(10):23-26.
- [5] 陈小平,丘坤元.“活性”/控制自由基聚合的研究进展[J].化学进展,2001,13(3):224-233.
- [6] LUO J, LI M, XIN M, et al. Benzoyl peroxide/2-vinylpyridine synergy in RAFT polymerization: synthesis of poly(2-vinylpyridine) with low dispersity at ambient temperature[J]. Macromolecular Chemistry & Physics, 2015, 216(15): 1646-1652.
- [7] SIDERIDOU I D, ACHILIAS D S, KARAVA O. Reactivity of benzoyl peroxide/amine system as an initiator for the free radical polymerization of dental and orthopaedic dimethacrylate monomers: effect of the amine and monomer chemical structure[J]. Macromolecules, 2006, 39(6): 2072-2080.

(责任编辑:曲继鹏)

(上接第97页)

## 参考文献:

- [1] 中华人民共和国教育部.全国普通高等学校体育课程指导纲要[EB/OL].(2002-06-21)[2019-08-12]. [http://www.moe.gov.cn/s78/A10/moe\\_918/tnull\\_8465.html](http://www.moe.gov.cn/s78/A10/moe_918/tnull_8465.html).
- [2] 刘成,李秀华.体质弱势群体与体育教学改革[J].体育学刊,2005,5(12):72-74.
- [3] 周二三,刘成,李秀华.体质弱势群体的理论构建[J].体育学刊,2008,7(15):47-49.
- [4] 刘晓莉.高校体育保健课程资源的开发与利用研究[J].西昌学院学报(自然科学版),2017,31(1):125-128.
- [5] 袁空军,吴加弘.高校体育保健课程建设研究[J].西昌学院学报(自然科学版),2018,32(3):124-128.
- [6] 刘丽.高校体育保健课教学评价体系的构建[J].山东师范大学学报(自然科学版),2017,32(2):140-147.
- [7] 张剑威,汤卫东.“体医结合”协同发展的时代意蕴、地方实践与推进思路[J].首都体育学院学报,2018,30(1):73-77.
- [8] 刘邦奇.智慧课堂:“互联网+”时代的课堂变革[N].江苏教育报,2016-09-21(04).
- [9] 孙曙辉,刘邦奇,李鑫,等.面向智慧课堂的数据挖掘与学习分析框架及应用[J].中国电化教育,2018(2):59-66.
- [10] 安徽省教育厅.安徽省教育厅关于在全省高校推行公共体育艺术教育俱乐部制教学改革的意见[EB/OL].(2018-05-31). [2019-10-12]. <http://jyt.ah.gov.cn/1569/view/583023.shtml>.

(责任编辑:蒋召雪)

(上接第78页)

- [13] SHU Wenhao, SHEN Hong. Updating attribute reduction in incomplete decision systems with the variation of attribute set[J]. International Journal of Approximate Reasoning, 2014, 55(3): 867-884.
- [14] WANG Guangqiong. Valid incremental attribute reduction algorithm based on attribute generalization for an incomplete information system[J]. Chinese Journal of Electronics, 2019, 28(4): 725-736.
- [15] 姚晟,徐风,赵鹏,等.基于邻域量化容差关系粗糙集模型的特征选择算法[J].模式识别与人工智能,2017,30(5):416-428.
- [16] 陈迎春,李鸥,孙昱.基于聚类离散化和变精度邻域熵的属性约简[J].控制与决策,2018,33(8):1407-1414.
- [17] ZHAO Hua, QIN Keyun. Mixed feature selection in incomplete decision table[J]. Knowledge-Based Systems, 2014, 57: 181-190.
- [18] LUO Chuang, LI Tianri, ZHANG Yi, et al. Matrix approach to decision-theoretic rough sets for evolving data[J]. Knowledge-Based Systems, 2016, 99: 123-134.
- [19] 闫鑫,景运革.矩阵增量属性约简算法[J].小型微型计算机系统,2018,39(6):1245-1249.
- [20] ZHANG Junbo, WONG J S, PAN Yi, et al. A parallel matrix-based method for computing approximations in incomplete information systems[J]. IEEE Transactions on Knowledge & Data Engineering, 2015, 27(2): 326-339.

(责任编辑:蒋召雪)