

引用格式:陈曦,刘晶. 基于邻域关系的知识粒度增量式属性约简算法[J]. 微电子学与计算机, 2020, 37(10): 1-6. [CHEN X, LIU J. Knowledge granularity incremental attribute reduction algorithm based on neighborhood relation[J]. Microelectronics & Computer, 2020, 37(10): 1-6.]

基于邻域关系的知识粒度增量式属性约简算法

陈 曦^{1,2}, 刘 晶²

(1 华中科技大学 管理学院, 湖北 武汉 430074; 2 武汉铁路职业技术学院, 湖北 武汉 430200)

摘 要: 为了在邻域型信息系统下进行增量式属性约简的研究, 采用邻域知识粒度构造出一种邻域型信息系统的增量式属性约简算法. 首先将信息系统的知识粒度在邻域型信息系统下进行推广, 提出了邻域知识粒度; 然后针对属性增加的情形, 研究了邻域知识粒度的增量式更新机制; 最后基于这种机制设计出了相应的增量式属性约简算法. 实验分析表明所提出的增量式算法具有较高的动态属性约简性能.

关键词: 粗糙集; 增量式学习; 邻域关系; 知识粒度; 属性约简

中图分类号: TP18

文献标识码: A

文章编号: 1000—7180(2020)10—0001—06

Knowledge granularity incremental attribute reduction algorithm based on neighborhood relation

CHEN Xi^{1,2}, LIU Jing²

(1 School of Management, Huazhong University of Science and Technology, Wuhan 430074, China;

2 Wuhan Railway Vocational and Technical College, Wuhan 430200, China)

Abstract: In order to study incremental attribute reduction in neighborhood information system, an incremental attribute reduction algorithm of neighborhood information system is constructed by using neighborhood knowledge granularity. Firstly, the knowledge granularity of information system is extended in the neighborhood information system, and the neighborhood knowledge granularity is proposed. Secondly, the incremental updating mechanism of neighborhood knowledge granularity is studied for the case of attribute increase. Finally, the corresponding incremental attribute reduction algorithm is designed based on this mechanism. Experimental results show that the incremental algorithm has high dynamic attribute reduction performance.

Key words: rough set; incremental learning; neighborhood; knowledge granularity; attribute reduction

1 引言

属性约简^[1]是粗糙集理论在机器学习和数据挖掘中的重要应用, 实际环境下的数据总是处于动态更新之中, 基于增量式学习的属性约简(增量式属性约简)^[2-3]成为了目前的研究重点. 目前关于增量式属性约简的研究大多基于离散型的信息系统^[4-6], 然而实际应用中邻域型信息系统却普遍存在, 为了解

决这一问题, 本文提出一种邻域型信息系统的增量式属性约简算法. 邻域粗糙集是处理邻域型信息系统的一种重要模型^[7], 知识粒度是进行属性约简的一种常用启发式方法^[8-11]. 本文首先将传统的知识粒度在邻域型信息系统下进行推广, 然后基于属性增加情形, 构建了邻域知识粒度的增量式更新机制, 同时提出增量式属性约简算法, 实验分析证明了该算法的高效性. 本文算法的提出将进一步推动属性

约简在实际环境下的应用范围。

2 基本理论

在本节,主要介绍知识粒度和邻域粗糙集理论的相关基本内容,为后文的进一步介绍提供铺垫。

定义 1^[8] 考虑信息系统 $S = (U, C \cup D)$, 属性子集 $B \subseteq C$ 诱导的划分为 $U/E_B = \{X_1, X_2, \dots, X_m\}$, 那么属性子集 B 在信息系统 S 下确定的知识粒度定义为

$$KG(B) = \sum_{i=1}^m \frac{|X_i|^2}{|U|^2}$$

式中, E_B 表示属性子集 B 确定的等价关系, 定义为

$$E_B = \{(x, y) \in U \times U \mid \forall a \in B, a(x) = a(y)\}.$$

定义 2^[8] 考虑信息系统 $S = (U, C \cup D)$, 属性子集 $B \subseteq C$ 诱导的划分为 $U/E_B = \{X_1, X_2, \dots, X_m\}$, 属性集 $B \cup D$ 诱导的划分为 $U/E_{B \cup D} = \{Y_1, Y_2, \dots, Y_n\}$, 那么 D 关于 B 的相对知识粒度定义为

$$KG(D \mid B) = KG(B) - KG(B \cup D).$$

邻域粗糙集是学者 Hu 针对邻域型数据提出的一种粗糙集模型^[7], 该模型能够进一步提升粗糙集理论的应用范围, 目前是粗糙集理论的研究热点。

定义 3 考虑邻域型信息系统 $S = (U, C \cup D)$, 其中条件属性集 C 中的每个属性均为数值型属性, 对于属性子集 $B \subseteq C$ 在该信息系统下诱导的邻域关系定义为

$$N_B^\delta = \{(x, y) \in U \times U \mid \Delta_B(x, y) \leq \delta\}.$$

式中, δ 是一个非负常数, 被称为邻域半径; $\Delta_B(x, y)$ 表示对象 x 和对象 y 在属性子集 B 下的距离度量^[7]. 基于邻域关系 N_B^δ , $\forall x \in U$ 可以导出一个对应的邻域类, 即邻域粒, 定义为

$$\delta_B(x) = \{y \in U \mid (x, y) \in N_B^\delta\}.$$

依托于邻域关系对论域的邻域粒化, 将每个邻域类作为基本的计算粒子, 从而进行了邻域粗糙集的各种近似计算^[7].

3 基于邻域粒化的知识粒度增量式更新

在文献[10-11]中, Jing 等学者将离散型信息系统下的知识粒度进行属性约简算法的构造, 研究了当信息系统属性增加时知识粒度的增量式更新方法, 实验验证了所提出算法的优越性, 本节在 Jing 等学者的基础上, 首先提出邻域信息系统下的知识粒度, 称之为邻域知识粒度, 然后进一步研究属性增加时邻域知识粒度的增量式更新方法。

定义 4 考虑邻域型信息系统 $S = (U, C \cup D)$,

$|U| = m$, 属性子集 $B \subseteq C$ 确定的邻域关系为 N_B^δ , $\forall x_i \in U$ 在邻域关系 N_B^δ 下确定的邻域类为 $\delta_B(x_i)$, 那么属性集 B 确定的邻域知识粒度定义为

$$NG(B) = \sum_{i=1}^m \frac{|\delta_B(x_i)|^2}{|U|^2}.$$

在定义 4 所示的邻域知识粒度基础上, 接下来将文献[10]所示的增量式更新方法应用于邻域知识粒度中。

定义 5 考虑邻域型信息系统 $S = (U, C \cup D)$, 属性子集 $B \subseteq C$ 确定的邻域关系为 N_B^δ , 假设 ΔB 是信息系统新增的一个属性子集, 那么属性集 $B \cup \Delta B$ 在论域上诱导的粒化结果定义为

$$U/N_{B \cup \Delta B}^\delta = \{\delta(x_1), \delta(x_2), \dots, \delta(x_k), \delta(x_1^{k+1}), \delta(x_2^{k+1}), \dots, \delta(x_{k+1}^{k+1}), \delta(x_1^{k+2}), \delta(x_2^{k+2}), \dots, \delta(x_{k+2}^{k+2}), \delta(x_1^m), \delta(x_2^m), \dots, \delta(x_m^m)\}.$$

这里的 $n(x_i) (i = 1, 2, \dots, k)$ 表示属性集 B 增加 ΔB 后未发生变化的邻域类, $\bigcup_{j=1}^k \delta(x_j^e) (e = k+1, k+2, \dots, m)$ 表示属性集 B 增加 ΔB 后发生变化的邻域类。

定理 1 考虑邻域型信息系统 $S = (U, C \cup D)$, 属性子集 $B \subseteq C$, 假设信息系统新增一个属性子集 ΔB , 属性集 $B \cup \Delta B$ 在论域上诱导的粒化结果为

$$U/N_{B \cup \Delta B}^\delta = \{\delta(x_1), \delta(x_2), \dots, \delta(x_k), \delta(x_1^{k+1}), \delta(x_2^{k+1}), \dots, \delta(x_{k+1}^{k+1}), \delta(x_1^{k+2}), \delta(x_2^{k+2}), \dots, \delta(x_{k+2}^{k+2}), \delta(x_1^m), \delta(x_2^m), \dots, \delta(x_m^m)\}.$$

那么新的邻域知识粒度增量式更新为

$$NG(B \cup \Delta B) = NG(B) -$$

$$\frac{2}{|U|^2} \sum_{e=k+1}^m \sum_{i=1}^l \sum_{j=1}^l |n(x_i^e)| \cdot |n(x_j^e)|$$

证明: 根据定义 4 关于邻域知识粒度的定义, 有

$$NG(B) = \sum_{i=1}^m \frac{|\delta(x_i)|^2}{|U|^2} =$$

$$\frac{1}{|U|^2} (\sum_{i=1}^k |\delta(x_i)|^2 +$$

$$\sum_{e=k+1}^m (\sum_{i=1}^l |\delta(x_i^e)|^2)) =$$

$$\frac{1}{|U|^2} (\sum_{i=1}^k |\delta(x_i)|^2 +$$

$$\sum_{e=k+1}^m \sum_{i=1}^l |\delta(x_i^e)|^2 +$$

$$2 \cdot \sum_{e=k+1}^m \sum_{i=1}^l \sum_{j=i+1}^l |\delta(x_i^e)| \cdot |\delta(x_j^e)|)$$

由于

$$NG(B \cup \Delta B) = \frac{1}{|U|^2} (\sum_{i=1}^k |\delta(x_i)|^2 +$$

$$\sum_{e=k+1}^m \sum_{i=1}^l |\delta(x_i^e)|^2)$$

那么

$$NG(B) = NG(B \cup \Delta B) +$$

$$\frac{2}{|U|^2} \cdot \sum_{e=k+1}^m \sum_{i=1}^{le} \sum_{j=i+1}^{le} |\delta(x_i^e)| \cdot |\delta(x_j^e)|$$

即

$$NG(B \cup \Delta B) = NG(B) -$$

$$\frac{2}{|U|^2} \cdot \sum_{e=k+1}^m \sum_{i=1}^{le} \sum_{j=i+1}^{le} |\delta(x_i^e)| \cdot |\delta(x_j^e)|$$

证明完毕.

定理 2 考虑邻域型信息系统 $S = (U, C \cup D)$, 属性子集 $B \subseteq C$, 假设信息系统新增一个属性子集 ΔB , 属性集 $B \cup \Delta B \cup D$ 在论域上诱导的粒化结果为

$$U/N_{B \cup \Delta B \cup D}^{\delta} = \{\delta(y_1), \delta(y_2), \dots, \delta(y_k), \delta(y_1^{k+1}), \delta(y_2^{k+1}), \dots, \delta(y_{k+1}^{k+1}), \delta(y_1^{k+2}), \delta(y_2^{k+2}), \dots, \delta(y_{k+2}^{k+2}), \delta(y_1^q), \delta(y_2^q), \dots, \delta(y_{l_q}^q)\}.$$

那么新的邻域相对知识粒度增量式更新为

$$NG(D | B \cup \Delta B) = NG(D | B) -$$

$$\frac{2}{|U|^2} \left(\sum_{e=k+1}^m \sum_{i=1}^{le} \sum_{j=i+1}^{le} |\delta(x_i^e)| \cdot |\delta(x_j^e)| - \right.$$

$$\left. \sum_{e=k+1}^q \sum_{i=1}^{le} \sum_{j=i+1}^{le} |\delta(y_i^e)| \cdot |\delta(y_j^e)| \right)$$

证明: 根据相对知识粒度的定义有

$$NG(D | B \cup \Delta B) = NG(B \cup \Delta B) -$$

$$NG(B \cup \Delta B \cup D)$$

根据定理 1 有

$$NG(B \cup \Delta B) = NG(B) -$$

$$\frac{2}{|U|^2} \cdot \sum_{e=k+1}^m \sum_{i=1}^{le} \sum_{j=i+1}^{le} |\delta(x_i^e)| \cdot |\delta(x_j^e)|$$

那么

$$NG(D | B \cup \Delta B) = NG(B) - NG(B \cup D) -$$

$$\frac{2}{|U|^2} \cdot \left(\sum_{e=k+1}^m \sum_{i=1}^{le} \sum_{j=i+1}^{le} |\delta(x_i^e)| \cdot |\delta(x_j^e)| + \right.$$

$$\left. \frac{2}{|U|^2} \cdot \left(\sum_{e=k+1}^m \sum_{i=1}^{le} \sum_{j=i+1}^{le} |\delta(y_i^e)| \cdot |\delta(y_j^e)| \right) \right)$$

由于 $NG(D | B) = NG(B) - NG(B \cup D)$, 因此

$$NG(D | B \cup \Delta B) = NG(D | B) -$$

$$\frac{2}{|U|^2} \left(\sum_{e=k+1}^m \sum_{i=1}^{le} \sum_{j=i+1}^{le} |\delta(x_i^e)| \cdot |\delta(x_j^e)| - \right.$$

$$\left. \sum_{e=k+1}^q \sum_{i=1}^{le} \sum_{j=i+1}^{le} |\delta(y_i^e)| \cdot |\delta(y_j^e)| \right)$$

证明完成.

4 基于邻域知识粒度的增量式属性约简算法

根据第三节所提出的邻域知识粒度增量式更新机制, 接下将进一步提出相应的增量式属性约简算法.

定义 6 考虑邻域型信息系统 $S = (U, C \cup D)$, 邻域半径为 δ , 若属性子集 $B \subseteq C$ 是属性全集 C 的邻域知识粒度属性约简, 那么当且仅当

$$(1) NG(D | B) = NG(D | C);$$

$$(2) \forall a \in B, NG(D | B - \{a\}) > NG(D | B).$$

为了进行比较, 本节首先给出当邻域型信息系统属性增加时, 基于邻域知识粒度的非增量式属性约简算法.

算法 1: 邻域型信息系统的邻域知识粒度非增量式属性约简算法.

输入: 邻域型信息系统 $S = (U, C \cup D)$, 邻域半径 δ , 新增加的属性子集 ΔC , 新的邻域型信息系统表示为 $S' = (U, C' \cup D)$, $C' = C \cup \Delta C$.

输出: 新的邻域型信息系统 S' 的属性约简集 red .

$$1. \text{ 初始化 } red = \emptyset, NG(D | \emptyset) = 1.$$

2. 对于属性 $\forall a \in C' - red$, 计算 a 的邻域知识粒度属性重要度

$$sig(a) = NG(D | red) - NG(D | red \cup \{a\}).$$

3. 在 $C' - red$ 中选择属性重要度值最大的属性, 记为 $a_{\max}$, 若 $sig(a_{\max}) > 0$, 则

$$red = red \cup \{a_{\max}\}.$$

并返回步骤 2 进行下一步迭代计算. 若满足 $sig(a_{\max}) = 0$, 那么进入步骤 4.

$$4. \text{ 终止迭代, 返回属性约简集 } red.$$

算法 1 类似于传统的知识粒度属性约简, 即利用邻域知识粒度不断对更新后的信息系统进行启发式搜索, 直到满足终止条件算法停止, 因此整个算法 1 的时间复杂度为 $O(|C'|^2 \cdot |U|^2)$.

接下来提出邻域知识粒度的增量式属性约简算法.

算法 2: 邻域型信息系统的邻域知识粒度增量式属性约简算法

输入: 邻域型信息系统 $S = (U, C \cup D)$, 邻域半径 δ , 信息系统 S 的属性约简 red , 邻域知识粒度 $NG(D | red)$ 和 $NG(D | C)$. 新增加的属性子集 ΔC , 新的邻域型信息系统表示为 $S' = (U, C' \cup D)$, $C' = C \cup \Delta C$.

输出: 新的邻域型信息系统 S' 的属性约简

集 red' .

1. 初始化 $red' \leftarrow red$.
2. 利用定理 2 在 $NG(D \mid red)$ 的基础上计算 $NG(D \mid red')$, 在 $NG(D \mid C)$ 基础上计算 $NG(D \mid C')$.
3. 判断 $NG(D \mid red')$ 与 $NG(D \mid C')$ 之间的大小关系, 若 $NG(D \mid red') = NG(D \mid C')$, 直接进入第 7 步骤, 否则进入步骤 4.
4. 对于属性 $\forall a \in C' - red'$, 计算邻域知识粒度属性重要度
$$sig(a) = NG(D \mid red') - NG(D \mid red' \cup \{a\}).$$
5. 在 $C' - red'$ 中选择属性重要度值最大的属性, 记为 a_{max} , 若 $sig(a_{max}) > 0$, 则
$$red' = red' \cup \{a_{max}\}.$$
6. 若满足 $sig(a_{max}) = 0$, 那么进入步骤 7. 否则返回步骤 4 进行下一步迭代计算.
7. 终止迭代, 返回属性约简集 red' .

算法 2 与算法 1 的不同之处在于, 算法 2 在原先属性约简集 red 的基础上进行进一步计算, 并且采用增量式的方法计算新的邻域知识粒度, 这种计算方法大大减少了重复的计算量, 具有很高的计算效率, 整个算法 2 的时间复杂度可表示为 $O((|C| + |\Delta C|)^2 \cdot |U|^2)$.

5 实验分析

本节将进行相关的实验来验证所提出的增量式属性约简算法的有效性. 本实验在 UCI 数据集库中选择了 5 个数值型的数据集作为邻域信息系统, 具体如表 1 所示.

表 1 实验数据集

编号	数据集名称	对象	属性	类别
1	Biodeg	1055	41	2
2	Cmc	1473	9	3
3	Abalone	4177	8	29
4	Thyroid	9172	29	2
5	Musk	6598	166	2

表 1 所示的均为固定的数据集, 为了构造出数据集属性动态变化的过程, 本实验将每个数据集按照属性分割成 7 个部分, 随机选择其中一个作为初始数据集, 然后将剩余的数据集逐步进行合并, 这样可以构造出数据集的 6 次属性动态增加更新. 针对这样的数据变化环境, 将邻域知识粒度非增量式属性约简算法 (算法 1) 和增量式属性约简算法 (算法 2) 分别进行实验, 最终的属性约简用时结果如图 1 所示.

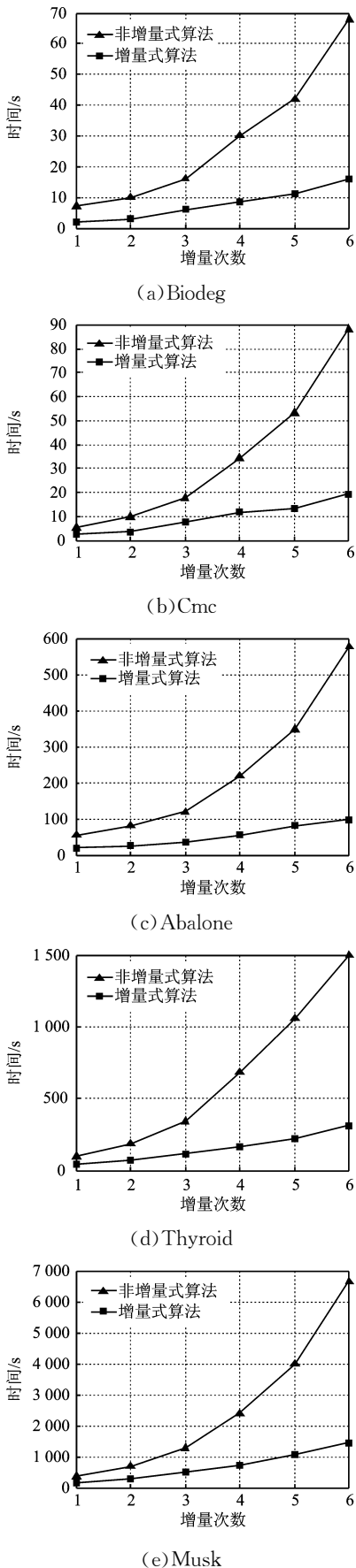


图 1 非增量式算法与增量式算法属性约简效率比较

图 1 所示的实验结果表明,非增量式属性约简算法的计算用时远大于增量式属性约简,并且随着属性更新次数的逐渐增多,这种差距逐步变大.出现这样的结果主要是由于非增量式算法即传统的属性约简算法,每次数据集更新后,该算法重新对完整的数据集进行属性约简计算,这样随着属性增加,数据的量会越来越大,因此约简的用时也随之增大,并且这种增长是呈指数级的,而对于增量式算法,每次更新约简时,首先会将旧信息系统的约简结果加以利用,并且在旧的约简结果基础上进行进一步计算,而且没有重复计算,因此大幅度减少了计算量,提高了属性约简的效率.因此增量式属性约简算法整体上

要优于非增量式属性约简算法.

对比非增量式属性约简和增量式属性约简,可以发现这两种算法在进行属性约简时的搜索策略是不一样的,非增量式算法每次基于完整的属性集进行属性搜索,而增量式属性约简每次基于旧信息系统的属性约简集进行搜索,因而最终的约简结果会保留较多原先的属性.因此接下来将比较这两类算法约简结果的优越性.

表 2 和表 3 分别所示的是非增量式属性约简算法和增量式属性约简算法每次更新计算约简时,其约简结果的支持向量机(SVM)分类精度和朴素贝叶斯(NB)分类精度结果.

表 2 非增量式算法的属性约简结果分类精度(%)

数据集	增量式更新次数											
	1		2		3		4		5		6	
	SVM	NB	SVM	NB	SVM	NB	SVM	NB	SVM	NB	SVM	NB
Biodeg	35.61	38.83	54.80	50.05	68.51	71.07	79.40	80.12	83.15	86.09	84.89	84.89
Cmc	51.12	48.96	56.74	58.28	72.43	70.36	85.37	81.65	96.89	99.29	96.04	96.33
Abalone	45.10	51.55	57.19	57.49	88.37	85.30	91.31	92.05	95.53	97.51	96.18	98.46
Thyroid	39.14	41.40	52.53	63.21	69.67	68.77	80.75	79.88	82.89	81.32	84.50	83.82
Musk	47.66	48.55	62.02	63.38	80.89	80.96	87.57	90.18	95.07	90.32	94.46	91.72

表 3 增量式算法的属性约简结果分类精度(%)

数据集	增量式更新次数											
	1		2		3		4		5		6	
	SVM	NB	SVM	NB	SVM	NB	SVM	NB	SVM	NB	SVM	NB
Biodeg	34.96	34.08	55.22	52.02	69.11	73.55	84.59	84.12	83.92	85.86	80.79	87.58
Cmc	46.98	52.10	59.59	55.95	74.01	68.95	83.33	84.14	96.54	93.34	98.57	94.32
Abalone	44.79	45.37	59.87	60.51	84.33	84.01	93.83	92.31	94.59	98.42	97.81	96.24
Thyroid	42.33	38.39	52.30	60.20	69.87	65.51	76.76	78.73	84.12	83.98	84.81	86.79
Musk	49.17	48.97	61.02	63.98	78.16	78.34	85.11	91.04	94.31	95.67	93.46	95.80

在表 2 和表 3 的结果中,增量式属性约简和非增量式属性约简得到的分类精度结果相差不大,甚至在部分数据集中,增量式属性约简得到的分类精度更高,例如数据集 Biodeg 在第 3、4 次以及第 6 次增量式更新时得到的约简结果,数据集 Abalone 第 4 次增量式更新得到的约简结果,数据集 Thyroid 第 5 至第 6 次增量式更新结果等.因此综合表明增量式属性约简算法在属性约简结果方面同样具有较好的属性约简性能.

6 结束语

增量式属性约简是粗糙集理论的重要研究内容.邻域粗糙集是处理邻域型信息系统的一种重要

模型,知识粒度是进行属性约简的一种常用的启发式方法.针对邻域型信息系统属性约简的增量式更新问题,本文提出一种增量式属性约简算法.首先将传统的知识粒度在邻域型信息系统下进行推广,然后基于属性增加时,构建了邻域知识粒度的增量式更新机制,同时相应的增量式属性约简算法被提出,仿真实验表明了该算法的高效性.本文所提出的增量式属性约简算法针对的是属性增加的情形,因此接下来将进一步探究邻域型信息系统对象变化时的增量式属性约简问题.

参考文献:

[1] YAO S, WU Z Y, CHEN J, et al. A new attribute

- reduction method based on decision — theoretic rough set[J]. *Microelectronics & Computer*, 2019, 36(5): 76-81.
- [2] DUAN H L, WANG G Q. An efficient incremental attribute reduction for complex information systems [J]. *Journal of South China University of Technology (Natural Science Edition)*, 2019, 47(6): 18-30.
- [3] SHU W H, QIAN W B, XIE Y H. Incremental approaches for feature selection from dynamic data with the variation of multiple objects[J]. *Knowledge-Based Systems*, 2019, 163(1): 320-331.
- [4] BAO D, ZHANG N, TONG X R, et al. Incremental attribute reduction algorithm of positive region in interval-valued decision tables[J]. *Journal of Computer Applications*, 2019, 39(8): 2288-2296.
- [5] ZHANG W, MIAO D Q, GAO C, et al. Rough set attribute reduction algorithm for partially labeled data [J]. *Computer Science*, 2017, 44(1): 25-31.
- [6] LIANG J Y, WANG F, DANG C Y, et al. A group incremental approach to feature selection applying rough set technique[J]. *IEEE Transactions on Knowledge & Data Engineering*, 2014, 26(2): 294-308.
- [7] HU Q H, YU D R, LIU J F, et al. Neighborhood rough set based heterogeneous feature subset selection [J]. *Information Sciences*, 2008, 178 (18): 3577-3594.
- [8] WANG G Q. Incremental attribute reduction of information system based on knowledge granulation[J]. *Measurement & Control Technology*, 2019, 38 (8): 54-60.
- [9] ZHENG C, WANG B, HONG T T. Incremental attribute reduction based on relational matrix[J]. *Journal of Chinese Computer Systems*, 2018, 39 (5): 1000-1004.
- [10] JING Y G, LI T R, LUO C, et al. An incremental approach for attribute reduction based on knowledge granularity[J]. *Knowledge-Based Systems*, 2016, 104 (15): 24-38.
- [11] JING Y G, LI T R, FUJITA H, et al. An incremental attribute reduction approach based on knowledge granularity with a multi-granulation view[J]. *Information Sciences*, 2017, 411: 23-38.

作者简介:

陈 曦 女, (1982-), 博士研究生, 讲师. 研究方向为智能计算、项目管理和风险管理. E-mail: chenxi198211@126.com

刘 晶 女, (1985-), 硕士, 讲师. 研究方向为计算机软件工程.