

• 人工智能 •

DOI:10.15961/j.jsuese.202201214



本刊网刊

基于不完备混合序信息系统的增量式属性约简

陈宝国¹, 陈磊^{1*}, 邓明¹, 陈金林^{1,2}

(1.淮南师范学院 计算机学院, 安徽 淮南 232038; 2.南京航空航天大学 电子与信息工程学院, 江苏 南京 211106)

摘要: 由于大数据环境下数据呈现出动态更新的特征, 因此, 增量式属性约简已成为粗糙集理论的重点研究方向。不完备混合型有序信息系统是一种常见的信息系统类型, 然而, 目前少有增量式属性约简方面的相关研究, 针对这一问题, 本文在不完备混合型有序信息系统下提出一种对象更新情形的增量式属性约简算法。首先, 针对不完备混合型有序信息系统提出了邻域容差优势关系, 基于该二元关系建立了一种新的邻域优势粗糙集模型。其次, 在其基础上定义了邻域优势条件熵, 并利用邻域优势条件熵作为启发式函数设计出一种不完备混合型有序信息系统的非增量式属性约简算法。然后, 利用矩阵的形式重构了邻域容差优势关系和邻域优势条件熵, 针对不完备混合型有序信息系统对象的动态变化, 基于矩阵的计算策略分别研究了邻域优势条件熵随信息系统对象增加和对象减少时的增量式更新。最后, 利用邻域优势条件熵的更新机制分别提出了不完备混合型有序信息系统对象增加和对象减少时属性约简的增量式更新算法。实验结果表明: 1) 与非增量式算法相比, 所提出的增量式算法约简属性数量平均降低了 3.6%, 分类精度平均提升了 2.4%, 属性约简的效率平均提升了约 10 倍; 2) 与同类型增量式算法相比, 所提出的增量式算法约简属性数量平均降低了 9.0%, 分类精度平均提升了 2.1%, 属性约简的平均效率提升了 94%。因此, 本文所提出增量式算法无论在属性约简结果和属性约简效率上都有着更高的性能。

关键词: 有序信息系统; 不完备混合; 优势粗糙集; 属性约简; 增量式; 条件熵

中图分类号: TP181

文献标志码: A

文章编号: 2096-3246(2024)01-0065-17

属性约简是粗糙集理论在模式识别下的最重要应用之一^[1-2]。粗糙集理论通过对数据集的属性子集构建二元关系, 实现了属性子集的不确定评估, 从而可以删除数据集中的无效属性。针对不同的数据类型和数据环境, 学者们基于粗糙集理论提出了大量的属性约简算法, 使得属性约简成为粗糙集理论最为活跃的研究内容^[1]。

当今大数据环境下, 数据常常呈现出快速更新变化的特性^[3], 一些传统的属性约简算法暴露出了计算效率低的缺陷。为了解决此问题, 增量式属性约简的方法近几年不断地被研究和探索^[4-15]。针对离散型的信息系统, 常用知识粒度^[4]、区分矩阵^[5-6]、依赖度^[7]和遗传算法^[8]等方法; 针对连续型的信息系统,

常用覆盖粒度^[9]、邻域粗糙集^[10]和模糊粗糙集^[11]等方法; 针对离散型和连续型混合类型的信息系统, 常用混合邻域粗糙集^[12]、邻域知识粒度^[13]、邻域区分度^[14]和邻域正区域^[15]等方法。目前为止, 增量式属性约简的研究已经涵盖了多种类型的数据。

然而, 数据类型总是复杂多样的, 某些情形下信息系统的属性值满足一定的偏序关系, 这类信息系统被称为有序信息系统, 其中, 优势粗糙集是处理该系统的常用模型^[16]。有序信息系统的研究近年来受到了越来越多的关注^[17-19], Sang 等^[20-22]学者分别在离散型和连续型有序信息系统下基于优势粗糙集提出了多种增量式属性约简算法。然而, 目前还未有不完备混合型有序信息系统增量式属性约简的提出和

收稿日期: 2022-11-07

基金项目: 安徽省高校自然科学研究重点项目 (KJ2018A0469; KJ2021A0972)

作者简介: 陈宝国 (1978—), 男, 副教授, 硕士。研究方向: 数据挖掘; 粗糙集与粒计算。E-mail: bgchen0706@163.com

* 通信作者: 陈磊, E-mail: 13855491762@163.com

网络出版时间: 2023-11-24 15:39:30

网络出版地址: <https://link.cnki.net/urlid/51.1773.TB.20231123.1218.001>

设计。以此为研究动机,本文进行了该类型信息系统的增量式属性约简研究和探索。

首先,提出一种邻域容差优势关系,实现不完备混合型有序信息系统的信息粒化和粗糙近似。其次,在所提出二元关系的框架下,建立了对应的信息熵、联合熵以及条件熵,利用条件熵的单调性设计了非增量式属性约简算法。然后,当不完备混合型有序信息系统的对象出现了动态增加和减少,建立邻域优势条件熵的增量式学习框架。最后,根据该框架提出了不完备混合型有序信息系统的增量式属性约简算法,实验分析结果证明了本文增量式算法的有效性。

1 邻域优势粗糙集基本理论

定义1^[12] 设混合型信息系统表示为 $HIS = (U, C \cup D, V)$ 。其中, U 为论域, $U = \{x_1, x_2, \dots, x_n\}$; $C \cup D$ 为信息系统的全体属性集,其中, C 为信息系统的条件属性集, $C = C^c \cup C^n$, C^c 和 C^n 分别为条件属性集中的离散型属性集和连续型属性集, D 为信息系统的决策属性集; $V = \bigcup_{a \in C \cup D} V_a$, 其中, V_a 为属性 a 的值域, $v_a(x)$ 为 $x (x \in U)$ 在属性 $a (a \in C \cup D)$ 下的属性值。同时,设混合型有序信息系统为 $HOIS \succeq = (U, C \cup D, V)$ ^[22]。

定义2^[22] 给定 $HOIS \succeq = (U, C \cup D, V)$, 对于条件属性子集 $B (B \subseteq C \text{ 且 } B = B^c \cup B^n)$, B^c 为 B 的离散型属性子集, B^n 为 B 的连续型属性子集。定义属性子集 B 的邻域优势关系 $N_B^{\succeq}$ 为:

$$N_B^{\succeq} = \{(x, y) \in U \times U \mid d_{B^n}(x, y) \geq \delta \wedge (\forall b \in B, v_b(x) \succeq v_b(y))\} \quad (1)$$

式中, δ 为邻域半径, $d_{B^n}(x, y)$ 为对象 x 与对象 y 之间的距离度量, 常见距离度量函数如闵可夫斯基距离度量^[10], $\succeq$ 表示属性值之间的偏序关系, $v_b(x) \succeq v_b(y)$ 即 $v_b(x)$ 优势于 $v_b(y)$ 。

邻域优势粗糙集中给定邻域优势关系 $N_B^{\succeq}$, 可将对象 $x (x \in U)$ 在论域 U 中诱导出两个信息粒, 分别表示为:

$$n_B^+(x) = \{y \in U \mid (y, x) \in N_B^{\succeq}\} \quad (2)$$

$$n_B^-(x) = \{y \in U \mid (x, y) \in N_B^{\succeq}\} \quad (3)$$

式中, $n_B^+(x)$ 和 $n_B^-(x)$ 分别为对象 x 在邻域优势关系 $N_B^{\succeq}$ 下的优势集和劣势集。

假设属性子集 $B = \{a_1, a_2\}$, 图1展示了对象 x_k 在邻域优势关系 $N_B^{\succeq}$ 下优势集和劣势集的示意图, $G_1 \sim G_5$ 代表了不同的区域, 不同形状的点属于不同的区域, 五角星形状的点属于 G_1 区域, 十字形状的点属于 G_2 区域, 三角形的点属于 G_3 区域, 正方形的点属于 G_4 区域, 圆形的点属于 G_5 区域。从图1

中可以看出, 对象 x_k 的优势集 $n_B^+(x_k) = G_2$, 对象 x_k 的劣势集 $n_B^-(x_k) = G_4$ 。

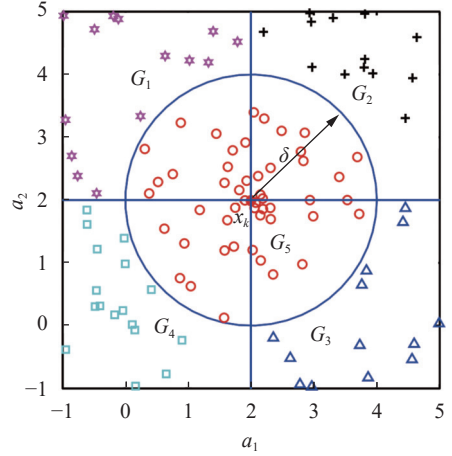


图1 对象优势集与劣势集的示意图

Fig. 1 Schematic diagram of object dominating set and dominated set

在混合型有序信息系统中, 决策属性集 D 的属性值均为离散型, 并且也满足相应的偏序关系, 设论域 U 在决策属性集 D 下的划分为 U/D , $U/D = \{Cl_1, Cl_2, \dots, Cl_m\}$, 并且决策类 $Cl_i (1 \leq i \leq m)$ 满足优势关系 $Cl_1 \leq Cl_2 \leq \dots \leq Cl_m$, 那么决策类 Cl_k 的上联合集 $Cl_k^{\succeq} = \bigcup_{t \geq k} Cl_t$, 决策类 Cl_k 的下联合集 $Cl_k^{\preceq} = \bigcup_{t \leq k} Cl_t$ 。

定义3^[22] 给定 $HOIS \succeq = (U, C \cup D, V)$, 条件属性集 $B \subseteq C$, $U/D = \{Cl_1, Cl_2, \dots, Cl_m\}$ 。决策类上联合集 $Cl_k^{\succeq}$ 在 $N_B^{\succeq}$ 的下近似和上近似分别定义为:

$$\underline{N}_B(Cl_k^{\succeq}) = \{x \in U \mid n_B^+(x) \subseteq Cl_k^{\succeq}\} \quad (4)$$

$$\overline{N}_B(Cl_k^{\succeq}) = \{x \in U \mid n_B^+(x) \cap Cl_k^{\succeq} \neq \emptyset\} \quad (5)$$

决策类下联合集 $Cl_k^{\preceq}$ 在邻域优势关系 $N_B^{\succeq}$ 的下近似和上近似分别定义为:

$$\underline{N}_B(Cl_k^{\preceq}) = \{x \in U \mid n_B^-(x) \subseteq Cl_k^{\preceq}\} \quad (6)$$

$$\overline{N}_B(Cl_k^{\preceq}) = \{x \in U \mid n_B^-(x) \cap Cl_k^{\preceq} \neq \emptyset\} \quad (7)$$

2 不完备混合型有序信息系统与属性约简

由于实际环境下数据采集难易程度的不同及采集成本等方面的原因, 信息系统存在着不完备的特性^[7,15-16,23]。考虑到混合型有序信息系统的完备性, 本节将提出不完备混合型有序信息系统的概念, 并提出对应的邻域优势粗糙集模型。

2.1 不完备邻域优势粗糙集模型

定义4 设不完备混合型有序信息系统为 $IHOIS \succeq = (U, C \cup D, V)$ 。其中, 论域 U 为信息系统全体对象的集合; 属性集 $C \cup D$ 为信息系统全体属性的集合; $V = \bigcup_{a \in C \cup D} V_a$, 其中, V_a 为属性 a 的值域, 对于 $x \in U$ 在属性 $a \in C \cup D$ 下的属性值表示为 $v_a(x)$, 同时,

$\exists a \in C, x \in U$ 有 $v_a(x) = *$, “ $*$ ”为属性值缺失的情形。

给定 $IHOIS^{\geq} = (U, C \cup D, V)$, 设条件属性子集 $B \subseteq C$ 且 $B = B^c \cup B^n$ 。定义属性子集 B 的邻域容差优势关系 $NT_B^{\geq}$ 为:

$$NT_B^{\geq} = \{(x, y) \in U \times U | (\forall b^n \in B^n, (v_{b^n}(x) \geq v_{b^n}(y) \wedge d_{b^n}(x, y) \geq \delta) \vee v_{b^n}(x) = * \vee v_{b^n}(y) = *) \wedge (\forall b^c \in B^c, v_{b^c}(x) \geq v_{b^c}(y) \vee v_{b^c}(x) = * \vee v_{b^c}(y) = *)\} \quad (8)$$

式中, $d_{b^n}(x, y)$ 为对象 x 和 y 在连续型属性 b^n 下的距离度量, $d_{b^n}(x, y) = \text{abs}(v_{b^n}(x) - v_{b^n}(y))$, 其中, $\text{abs}(\bullet)$ 为数值的绝对值。

定义5 给定 $IHOIS^{\geq} = (U, C \cup D, V)$, 条件属性子集 $B \subseteq C$ 确定的邻域容差优势关系为 $NT_B^{\geq}$, 则对象 $x(x \in U)$ 的邻域容差优势集和邻域容差劣势集分别为:

$$nt_B^+(x) = \{y \in U | (y, x) \in NT_B^{\geq}\} \quad (9)$$

$$nt_B^-(x) = \{y \in U | (x, y) \in NT_B^{\geq}\} \quad (10)$$

当邻域容差优势关系、邻域容差优势集和劣势集需要区分邻域半径时, 为防止混淆, 增加不同邻域半径的标记。例如, 对于不同的邻域半径 δ_1 和 δ_2 , 对应的邻域容差优势关系分别为 $NT_{B, \delta_1}^{\geq}$ 和 $NT_{B, \delta_2}^{\geq}$, 对应的邻域容差优势集分别为 $nt_{B, \delta_1}^+(x)$ 和 $nt_{B, \delta_2}^+(x)$, 对应的邻域容差劣势集分别为 $nt_{B, \delta_1}^-(x)$ 和 $nt_{B, \delta_2}^-(x)$ 。在不引起混淆的情况下, 通常省略邻域半径的标记。

邻域容差优势集和邻域容差劣势集满足性质1, 性质1详细内容见附录A。

在性质1中: 1) 体现出了邻域容差优势集/劣势集随邻域半径增加的单调性; 2) 体现出了邻域容差优势集/劣势集随属性子集增加的单调性; 3) 给出了单独属性子集与组合属性子集之间邻域容差优势集/劣势集的相互关系。

定义6 给定 $IHOIS^{\geq} = (U, C \cup D, V)$, 条件属性集 $B \subseteq C$, $U/D = \{Cl_1, Cl_2, \dots, Cl_m\}$ 。对于决策类上联合集 $C_k^{\geq}$ 在 $NT_B^{\geq}$ 的下近似和上近似分别定义为:

$$\underline{NT}_B(C_k^{\geq}) = \{x \in U | nt_B^+(x) \subseteq C_k^{\geq}\} \quad (11)$$

$$\overline{NT}_B(C_k^{\geq}) = \{x \in U | nt_B^+(x) \cap C_k^{\geq} \neq \emptyset\} \quad (12)$$

决策类下联合集 $C_k^{\leq}$ 在邻域容差优势关系 $NT_B^{\geq}$ 的下近似和上近似分别定义为:

$$\underline{NT}_B(C_k^{\leq}) = \{x \in U | nt_B^-(x) \subseteq C_k^{\leq}\} \quad (13)$$

$$\overline{NT}_B(C_k^{\leq}) = \{x \in U | nt_B^-(x) \cap C_k^{\leq} \neq \emptyset\} \quad (14)$$

2.2 不完备邻域优势条件熵与属性约简

Zhao等^[24]基于信息熵模型实现了不完备信息系统的确定性度量与属性约简, 本节针对此文所研

究的信息系统类型, 将该方法进行推广和扩展。

定义7 给定 $IHOIS^{\geq} = (U, C \cup D, V)$, $U = \{x_1, x_2, \dots, x_n\}$, 条件属性子集 $B(B \subseteq C)$ 确定的邻域容差优势关系为 $NT_B^{\geq}$, 则由 B 在论域 U 下确定的邻域优势信息熵 $INDE_B^{\geq}(U)$ 定义为:

$$INDE_B^{\geq}(U) = \frac{1}{n} \sum_{i=1}^n \left(1 - \frac{|nt_B^+(x_i)|}{n} \right) \quad (15)$$

式中, $|\bullet|$ 为集合的基数。邻域优势信息熵满足关系 $0 \leq INDE_B^{\geq}(U) \leq 1$, 当 $\forall x \in U$ 满足 $nt_B^+(x) = U$ 时, 则有 $INDE_B^{\geq}(U) = 0$; 当 $\forall x \in U$ 满足 $nt_B^+(x) = \emptyset$ 时, 则有 $INDE_B^{\geq}(U) = 1$ 。

定义8 给定 $IHOIS^{\geq} = (U, C \cup D, V)$, $U = \{x_1, x_2, \dots, x_n\}$, 条件属性子集 $P, Q (P, Q \subseteq C)$ 确定的邻域容差优势关系分别为 $NT_P^{\geq}$ 和 $NT_Q^{\geq}$, 则由 P 和 Q 在论域 U 下确定的邻域优势联合熵 $INDE_{P \cup Q}^{\geq}(U)$ 定义为:

$$INDE_{P \cup Q}^{\geq}(U) = \frac{1}{n} \sum_{i=1}^n \left(1 - \frac{|nt_P^+(x_i) \cap nt_Q^+(x_i)|}{n} \right) \quad (16)$$

邻域优势联合熵同样满足关系 $0 \leq INDE_{P \cup Q}^{\geq}(U) \leq 1$ 。

定义9 给定 $IHOIS^{\geq} = (U, C \cup D, V)$, $U = \{x_1, x_2, \dots, x_n\}$, 条件属性子集 $P, Q (P, Q \subseteq C)$ 确定的邻域容差优势关系分别为 $NT_P^{\geq}$ 和 $NT_Q^{\geq}$, 则 Q 关于 P 在论域 U 下确定的邻域优势条件熵 $INDE_{Q|P}^{\geq}(U)$ 定义为:

$$INDE_{Q|P}^{\geq}(U) = \frac{1}{n} \sum_{i=1}^n \left(\frac{|nt_P^+(x_i)|}{n} - \frac{|nt_P^+(x_i) \cap nt_Q^+(x_i)|}{n} \right) = INDE_{P \cup Q}^{\geq}(U) - INDE_P^{\geq}(U) \quad (17)$$

式中, $0 \leq INDE_{Q|P}^{\geq}(U) \leq 1$ 。

邻域优势条件熵满足性质2, 性质2具体内容见附录A。性质2表明, 随着属性集的增加, 邻域优势条件熵满足单调减少的性质, 这一特性经常被学者们用来构造信息系统的属性约简^[24]。

定义10 给定 $IHOIS^{\geq} = (U, C \cup D, V)$, 若同时满足下述两个条件, 那么称 $R(R \subseteq C)$ 为信息系统的属性约简:

$$1) INDE_{D|R}^{\geq}(U) = INDE_{D|C}^{\geq}(U),$$

$$2) \forall a \in R, INDE_{D|R-\{a\}}^{\geq}(U) > INDE_{D|C}^{\geq}(U).$$

基于该属性约简的定义, 可以得到属性的两种重要度函数, 利用这两种重要度度量可以进行不完备混合型有序信息系统属性约简结果的启发式搜索。

定义11 给定 $IHOIS^{\geq} = (U, C \cup D, V)$, 属性子集 $B(B \subseteq C)$, 对于 $\forall a \in B$ 在 B 下的内部属性重要度定义为:

$$Sig_{in}^U(a, B, D) = INDE_{D|B-\{a\}}^{\geq}(U) - INDE_{D|B}^{\geq}(U) \quad (18)$$

对于 $\forall b \in (C - B)$ 在 B 下的外部属性重要度定义为

$$Sig_{out}^U(b, B, D) = INDE_{D|B}^{\geq}(U) - INDE_{D|B \cup \{b\}}^{\geq}(U) \quad (19)$$

根据定义10和定义11, 利用邻域优势条件熵对不完备混合型有序信息系统进行属性约简的实现如算法1所示, 由于算法1在进行属性约简时输入的是静态数据集, 因此算法1也称为非增量式属性约简算法。

算法1 基于邻域优势条件熵的不完备混合型有序信息系统属性约简算法。

输入: $IHOIS^{\geq} = (U, C \cup D, V)$, 邻域半径 δ 。

输出: 属性约简结果 R_U 。

具体步骤如下:

1) 初始化 $R_U \leftarrow \emptyset$ 。

2) 对于 $\forall a \in C$, 计算属性 a 在 C 下的内部属性重要度 $Sig_{in}^U(a, C, D)$, 若 $Sig_{in}^U(a, C, D) > 0$, 则 $R_U \leftarrow R_U \cup \{a\}$ 。

3) 对于 $\forall b \in (C - R_U)$: 计算属性 b 在 R_U 下的外部属性重要度 $Sig_{out}^U(b, R_U, D)$, 找出外部属性重要度最大的属性 $b_{\max}$, $b_{\max} = \arg \{Sig_{out}^U(b, R_U, D)\}_{b \in C - R_U}$ 。

4) 对于步骤3)中的属性 $b_{\max}$, 如果 $Sig_{out}^U(b_{\max}, R_U, D) > 0$, 那么 $R_U \leftarrow R_U \cup \{b_{\max}\}$, 并返回步骤3), 否则进入步骤5)。

5) 对于 $\forall c \in R_U$, 若满足关系 $INDE_{D|R_U - \{c\}}^{\geq}(U) = INDE_{D|R_U}^{\geq}(U)$, 那么 $R_U \leftarrow R_U - \{c\}$, 重复进行步骤5), 直到 R_U 中所有属性都不满足该关系。

6) 返回属性约简结果 R_U 。

算法1中的主要计算量集中在步骤2), 步骤2)的时间复杂度为 $O(|C|^2 \cdot |U|^2)$, 步骤3)和步骤4)在剩余属性集中搜索候选属性, 其时间复杂度为 $O(|C| \cdot |U|^2)$, 步骤5)进行反向剔除冗余属性, 其时间复杂度为 $O(|C| \cdot |U|^2)$, 因此, 整个算法1的时间复杂度为 $O(|C|^2 \cdot |U|^2)$ 。

3 不完备混合型有序信息系统增量式属性约简

信息系统不断进行更新, 当信息系统对象发生增删时, 原有约简结果不再有效, 因此, 需要重新计算新信息系统的属性约简。在已有各种基于信息熵的属性约简算法中, 学者们均通过增量式更新信息熵的方法来更新属性约简结果^[10,13,21-22], 因此本文也采用类似的实现方案完成属性约简的动态更新。

3.1 邻域优势条件熵的矩阵表示

在粗糙集理论中, 矩阵是构造模型和算法增量式学习的一种常用工具^[3,5-6,18,21-22]。

定义12 给定 $IHOIS^{\geq} = (U, C \cup D, V)$, $U = \{x_1, x_2, \dots, x_n\}$, 条件属性子集 $B(B \subseteq C)$ 确定的邻域容差优势关系为 $NT_B^{\geq}$, 定义论域 U 下 $NT_B^{\geq}$ 的邻域容差优势关

系矩阵为:

$$N_B^{\geq}(U) = [\omega_{ij}^B]_{n \times n} = \begin{bmatrix} \omega_{11} & \omega_{12} & \cdots & \omega_{1n} \\ \omega_{21} & \omega_{22} & \cdots & \omega_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{n1} & \omega_{n2} & \cdots & \omega_{nn} \end{bmatrix} \quad (20)$$

式中, $\omega_{ij}^B = \begin{cases} 1, & (x_j, x_i) \in NT_B^{\geq}, x_i, x_j \in U; \\ 0, & (x_j, x_i) \notin NT_B^{\geq}, x_i, x_j \in U \end{cases}$ 。

定义12表明: 邻域容差优势关系矩阵是邻域容差优势关系的直观展示; 对于对象 $x_i, x_j \in U$, 当对象 x_j 在属性子集 B 下优势于 x_i 时, 此时关系矩阵第 i 行第 j 列上的元素为1; 否则为0。因此, 根据定义5可得:

$$nt_B^+(x_i) = \{x_j \in U | \omega_{ij}^B = 1, N_B^{\geq}(U) = [\omega_{ij}^B]_{n \times n}\} \quad (21)$$

$$nt_B^-(x_i) = \{x_j \in U | \omega_{ji}^B = 1, N_B^{\geq}(U) = [\omega_{ij}^B]_{n \times n}\} \quad (22)$$

$$|nt_B^+(x_i)| = \sum_{j=1}^n \omega_{ij}^B, |nt_B^-(x_i)| = \sum_{j=1}^n \omega_{ji}^B。$$

定义13 对于邻域容差优势关系矩阵 $N_P^{\geq}(U) = [\phi_{ij}^P]_{n \times n}$ 和 $N_Q^{\geq}(U) = [\phi_{ij}^Q]_{n \times n}$, 定义关系矩阵的“ $Sum(\bullet)$ ”和“ $\wedge$ ”运算为:

$$Sum(N_P^{\geq}(U)) = \sum_{i=1}^n \sum_{j=1}^n \phi_{ij}^P \quad (23)$$

$$N_P^{\geq}(U) \wedge N_Q^{\geq}(U) = [\pi_{ij}]_{n \times n} = \begin{cases} 1, & \phi_{ij}^P = 1 \wedge \phi_{ij}^Q = 1; \\ 0, & \text{other} \end{cases} \quad (24)$$

利用关系矩阵的定义和关系矩阵的计算规则, 从而构建邻域优势条件熵的矩阵框架。

定理1 给定 $IHOIS^{\geq} = (U, C \cup D, V)$, $U = \{x_1, x_2, \dots, x_n\}$, 条件属性子集 $P, Q(P, Q \subseteq C)$ 确定的邻域容差优势关系矩阵分别为 $N_P^{\geq}(U) = [\phi_{ij}^P]_{n \times n}$ 和 $N_Q^{\geq}(U) = [\phi_{ij}^Q]_{n \times n}$, 则 Q 关于 P 在论域 U 下确定的邻域优势条件熵 $INDE_{Q|P}^{\geq}(U)$ 可表示为:

$$INDE_{Q|P}^{\geq}(U) = \frac{1}{n^2} [Sum(N_P^{\geq}(U)) - Sum(N_P^{\geq}(U) \wedge N_Q^{\geq}(U))] \quad (25)$$

证明: 证明内容见附录A。

3.2 基于矩阵的邻域优势条件熵增量式更新

定理2 给定 $IHOIS^{\geq} = (U, C \cup D, V)$, $U = \{x_1, x_2, \dots, x_n\}$, 属性子集 $B(B \subseteq C)$ 在论域 U 下确定的邻域容差优势关系为 $NT_B^{\geq}$, 对应的关系矩阵为 $N_B^{\geq}(U)$, $N_B^{\geq}(U) = [\omega_{ij}^B]_{n \times n}$ 。设信息系统新增加的对象集 U^+ , $U^+ = \{x_{n+1}, x_{n+2}, \dots, x_{n+n^+}\}$, $U' = U \cup U^+$, $n^+ = |U^+|$, 那么属性子集 B 在 U' 下确定的邻域容差优势关系为 $NT_B'^{\geq}$, 对应的关系矩阵 $N_B^{\geq}(U')(N_B^{\geq}(U') = [\omega_{ij}^{B'}]_{(n+n^+) \times (n+n^+)})$ 增量式更新为:

$$N_B^{\geq}(U') = [\omega_{ij}^B]_{(n+n^+) \times (n+n^+)} = \begin{bmatrix} (M_1^B)_{n \times n} & (M_2^B)_{n \times n^+} \\ (M_3^B)_{n^+ \times n} & (M_4^B)_{n^+ \times n^+} \end{bmatrix} \quad (26)$$

式中:

$$(M_1^B)_{n \times n} = [\omega_{ij}^B]_{n \times n} = N_B^{\geq}(U) \quad (27)$$

$$(M_2^B)_{n \times n^+} = \begin{cases} 1, & (x_i, x_j) \in NT_B^{\geq}, x_i \in U, x_j \in U^+; \\ 0, & (x_i, x_j) \notin NT_B^{\geq}, x_i \in U, x_j \in U^+ \end{cases} \quad (28)$$

$$(M_3^B)_{n^+ \times n} = \begin{cases} 1, & (x_i, x_j) \in NT_B^{\geq}, x_i \in U^+, x_j \in U; \\ 0, & (x_i, x_j) \notin NT_B^{\geq}, x_i \in U^+, x_j \in U \end{cases} \quad (29)$$

$$(M_4^B)_{n^+ \times n^+} = \begin{cases} 1, & (x_i, x_j) \in NT_B^{\geq}, x_i \in U^+, x_j \in U^+; \\ 0, & (x_i, x_j) \notin NT_B^{\geq}, x_i \in U^+, x_j \in U^+ \end{cases} \quad (30)$$

证明对于邻域容差优势关系 $NT_B^{\geq}$ 和 $NT_B^{\geq}$, 若 $\forall x_i, x_j \in U$ ($1 \leq i, j \leq n$) 满足 $(x_i, x_j) \in NT_B^{\geq}$, 根据定义4有 $(x_i, x_j) \in NT_B^{\geq}$, 所以, 对于 $N_B^{\geq}(U) = [\omega_{ij}^B]_{n \times n}$ 和 $N_B^{\geq}(U') = [\omega_{ij}^B]_{(n+n^+) \times (n+n^+)}$, 其中, $\omega_{ij}^B = \omega_{ij}^B$ ($1 \leq i, j \leq n$), 本定理式(27)成立, 通过定义12可得到式(27)~(30)成立。

邻域容差优势关系矩阵满足性质3, 性质3详细内容见附录。

定理3 给定 $IHOIS^{\geq} = (U, C \cup D, V)$, $U = \{x_1, x_2, \dots, x_n\}$, 条件属性子集 P, Q ($P, Q \subseteq C$) 确定的邻域容差优势关系矩阵分别为 $N_P^{\geq}(U)$ ($N_P^{\geq}(U) = [\phi_{ij}^P]_{n \times n}$) 和 $N_Q^{\geq}(U)$ ($N_Q^{\geq}(U) = [\phi_{ij}^Q]_{n \times n}$), Q 关于 P 在论域 U 下确定的邻域优势条件熵为 $INDE_{Q|P}^{\geq}(U)$ 。设信息系统新增加的对象集 U^+ , $U^+ = \{x_{n+1}, x_{n+2}, \dots, x_{n+n^+}\}$, $U' = U \cup U^+$, $n^+ = |U^+|$, 两种关系矩阵分别增量式更新为 $N_P^{\geq}(U') = [\phi_{ij}^P]_{(n+n^+) \times (n+n^+)}$, $N_Q^{\geq}(U') = [\phi_{ij}^Q]_{(n+n^+) \times (n+n^+)}$, 那么 Q 关于 P 在论域 U' 下确定的邻域优势条件熵 $INDE_{Q|P}^{\geq}(U')$ 增量式更新为:

$$INDE_{Q|P}^{\geq}(U') = \frac{1}{(n+n^+)^2} \left[n^2 \cdot INDE_{Q|P}^{\geq}(U) + \begin{aligned} & Sum((M_2^P)_{n \times n^+}) - Sum((M_2^P)_{n \times n^+} \wedge (M_2^Q)_{n \times n^+}) + \\ & Sum((M_3^P)_{n^+ \times n}) - Sum((M_3^P)_{n^+ \times n} \wedge (M_3^Q)_{n^+ \times n}) + \\ & Sum((M_4^P)_{n^+ \times n^+}) - Sum((M_4^P)_{n^+ \times n^+} \wedge (M_4^Q)_{n^+ \times n^+}) \end{aligned} \right] = \\ \frac{n^2}{(n+n^+)^2} \cdot INDE_{Q|P}^{\geq}(U) + \frac{1}{(n+n^+)^2} \sum_{i=n+1}^{n+n^+} (|nt_P^+(x_i) - nt_Q^+(x_i) - U^+| + |nt_P^-(x_i) - nt_Q^-(x_i) - U^+|) \quad (31)$$

证明: 证明内容见附录A。

定理4 给定 $IHOIS^{\geq} = (U, C \cup D, V)$, $U = \{x_1, x_2, \dots, x_n\}$, 属性子集 B ($B \subseteq C$) 在论域 U 下确定的邻域容差优势关系为 $NT_B^{\geq}$, 对应的关系矩阵为 $N_B^{\geq}(U)$ ($N_B^{\geq}(U) = [\omega_{ij}^B]_{n \times n}$)。为了简便, 不妨设信息系统减少的对象集为 U^- , $U^- = \{x_{n-n^-+1}, x_{n-n^-+2}, \dots, x_n\}$, $U' = U - U^-$, $n^- = |U^-|$,

那么属性子集 B 在论域 U' 下确定的邻域容差优势关系为 $NT_B'^{\geq}$, 对应的关系矩阵为 $N_B'^{\geq}$, $N_B'^{\geq}(U') = [\omega_{ij}^B]_{(n-n^-) \times (n-n^-)}$ 。若将 $N_B^{\geq}(U)$ ($N_B^{\geq}(U) = [\omega_{ij}^B]_{n \times n}$) 表示成如下分块矩阵形式:

$$N_B^{\geq}(U) = [\omega_{ij}^B]_{n \times n} = \begin{bmatrix} (M_1^B)_{(n-n^-) \times (n-n^-)} & (M_2^B)_{(n-n^-) \times n^-} \\ (M_3^B)_{n^- \times (n-n^-)} & (M_4^B)_{n^- \times n^-} \end{bmatrix} \quad (32)$$

那么 $N_B'^{\geq}(U') = [\omega_{ij}^B]_{(n-n^-) \times (n-n^-)} = (M_1^B)_{(n-n^-) \times (n-n^-)}$ 。

证明: 根据邻域容差优势关系矩阵的定义可直接得到定理4成立。

从定理4可以看出, 将原先邻域容差优势关系矩阵删除第 $n-n^-+1, n-n^-+2, \dots, n$ 行和第 $n-n^-+1, n-n^-+2, \dots, n$ 列, 即得到新信息系统下的关系矩阵。

邻域容差优势关系矩阵满足性质4, 性质4详细内容见附录A。

定理5 给定 $IHOIS^{\geq} = (U, C \cup D, V)$, $U = \{x_1, x_2, \dots, x_n\}$, 条件属性子集 P, Q ($P, Q \subseteq C$) 确定的邻域容差优势关系矩阵分别为 $N_P^{\geq}(U) = [\phi_{ij}^P]_{n \times n}$ 和 $N_Q^{\geq}(U) = [\phi_{ij}^Q]_{n \times n}$, Q 关于 P 在论域 U 下确定的邻域优势条件熵为 $INDE_{Q|P}^{\geq}(U)$ 。设信息系统减少的对象集为 U^- , $U^- = \{x_{n-n^-+1}, x_{n-n^-+2}, \dots, x_n\}$, $U' = U - U^-$, $n^- = |U^-|$, 两种关系矩阵分别增量式更新为 $N_P^{\geq}(U') = [\phi_{ij}^P]_{(n-n^-) \times (n-n^-)}$, $N_Q^{\geq}(U') = [\phi_{ij}^Q]_{(n-n^-) \times (n-n^-)}$, 那么 Q 关于 P 在论域 U' 下确定的邻域优势条件熵 $INDE_{Q|P}^{\geq}(U')$ 增量式更新为:

$$INDE_{Q|P}^{\geq}(U') = \frac{1}{(n-n^-)^2} \left[n^2 \cdot INDE_{Q|P}^{\geq}(U) - \begin{aligned} & (Sum((M_2^P)_{(n-n^-) \times n^-}) - Sum((M_2^P)_{(n-n^-) \times n^-} \wedge \\ & (M_2^Q)_{(n-n^-) \times n^-})) - (Sum((M_3^P)_{n^- \times (n-n^-)}) - \\ & Sum((M_3^P)_{n^- \times (n-n^-)} \wedge (M_3^Q)_{n^- \times (n-n^-)})) - \\ & (Sum((M_4^P)_{n^- \times n^-}) - Sum((M_4^P)_{n^- \times n^-} \wedge \\ & (M_4^Q)_{n^- \times n^-})) \end{aligned} \right] = \frac{n^2}{(n-n^-)^2} \cdot INDE_{Q|P}^{\geq}(U) - \frac{1}{(n-n^-)^2} \cdot \sum_{i=n-n^-+1}^n (|nt_P^+(x_i) - nt_Q^+(x_i)| + |nt_P^-(x_i) - nt_Q^-(x_i) - U^-|) \quad (33)$$

证明: 同定理3证明。

3.3 增量式属性约简算法

第3.2节中实现了不完备混合有序信息系统对象增删情形时邻域优势条件熵的动态更新, 本节在其基础上实现对应的增量式属性约简算法, 分别如算法2和算法3所示。

算法2 对象增加时基于邻域优势条件熵的不完备混合型有序信息系统增量式属性约简算法。

输入: 1) $IHOIS^{\geq} = (U, C \cup D, V)$, $U = \{x_1, x_2, \dots,$

x_n), 邻域半径 δ ; 2) 新增加的对象集 $U^+ = \{x_{n+1}, x_{n+2}, \dots, x_{n+n^+}\}$, $U' = U \cup U^+$; 3) 信息系统 $IHOIS^{\geq}$ 的属性约简集 R_U , 邻域优势条件熵 $INDE_{D|R_U}^{\geq}(U)$ 和 $INDE_{D|C}^{\geq}(U)$ 。

输出: 新信息系统下的属性约简 $R_{U'}$ 和邻域优势条件熵 $INDE_{D|R_{U'}}^{\geq}(U')$ 和 $INDE_{D|C}^{\geq}(U')$ 。

具体步骤如下:

1) 初始化 B , $B \leftarrow R_U$ 。

2) 基于定理2计算邻域容差优势关系矩阵 $N_B^{\geq}(U')$ 的分块矩阵 $(M_2^B)_{n \times n^+}$ 、 $(M_3^B)_{n^+ \times n}$ 和 $(M_4^B)_{n^+ \times n^+}$, $N_C^{\geq}(U')$ 的分块矩阵 $(M_2^C)_{n \times n^+}$ 、 $(M_3^C)_{n^+ \times n}$ 和 $(M_4^C)_{n^+ \times n^+}$, $N_D^{\geq}(U')$ 的分块矩阵 $(M_2^D)_{n \times n^+}$ 、 $(M_3^D)_{n^+ \times n}$ 和 $(M_4^D)_{n^+ \times n^+}$ 。

3) 根据定理3, 基于 $INDE_{D|B}^{\geq}(U)$ 和 $INDE_{D|C}^{\geq}(U)$ 计算新的邻域优势条件熵 $INDE_{D|B}^{\geq}(U')$ 和 $INDE_{D|C}^{\geq}(U')$ 。

4) 若 $INDE_{D|B}^{\geq}(U') = INDE_{D|C}^{\geq}(U')$, 那么进入步骤7), 若 $INDE_{D|B}^{\geq}(U') > INDE_{D|C}^{\geq}(U')$, 那么进入步骤5)。

5) 对于 $\forall b \in (C - B)$, 计算属性 b 在 B 下的外部属性重要度 $Sig_{out}^{U'}(b, B, D)$, 找出外部属性重要度最大的属性 $b_{\max}$, $b_{\max} = \arg \{Sig_{out}^{U'}(b, B, D)\}_{b \in (C-B)}$ 。

6) 对于步骤5)中的属性 $b_{\max}$, 如果 $Sig_{out}^{U'}(b_{\max}, B, D) > 0$, 那么 $B \leftarrow B \cup \{b_{\max}\}$, 并返回步骤5), 否则进入步骤7)。

7) 对于 $\forall c \in B$, 若满足关系 $INDE_{D|B-\{c\}}^{\geq}(U') = INDE_{D|B}^{\geq}(U')$, 那么 $B \leftarrow B - \{c\}$, 重复进行步骤7), 直到 B 中所有属性都不满足该关系。

8) 进行 $R_{U'} \leftarrow B$, 返回新信息系统下的属性约简 $R_{U'}$ 和新的邻域优势条件熵 $INDE_{D|R_{U'}}^{\geq}(U')$ 和 $INDE_{D|C}^{\geq}(U')$ 。

算法2的主要策略是基于旧信息系统的属性约简结果 R_U 来更新新信息系统的属性约简。其中步骤2)利用定理2计算邻域容差优势关系矩阵的分块矩阵, 其时间复杂度为 $O(|C| \cdot |U| \cdot |U^+|)$; 步骤3)为增量式计算新的邻域优势条件熵, 其时间复杂度为 $O(|U| \cdot |U^+|)$; 步骤5)和步骤6)为在旧信息系统的属性约简结果基础上进一步搜索新的属性, 其时间复杂度为 $O(|\Delta C| \cdot |U \cup U^+|^2)$, 其中 ΔC 为新搜索到的属性集; 步骤7)为反向冗余属性剔除的过程, 其时间复杂度为 $O(|C| \cdot |U \cup U^+|^2)$, 因此, 整个算法2的时间复杂度为 $O(|C| \cdot |U| \cdot |U^+| + |C| \cdot |U \cup U^+|^2)$ 。

算法3 对象减少时基于邻域优势条件熵的不完备混合型有序信息系统增量式属性约简算法。

输入: 1) $IHOIS^{\geq} = (U, C \cup D, V)$, $U = \{x_1, x_2, \dots, x_n\}$, 邻域半径 δ ; 2) 减少的对象集 U^- , $U^- = \{x_{n-n^-+1}, x_{n-n^-+2}, \dots, x_n\}$, $U' = U - U^-$; 3) 信息系统 $IHOIS^{\geq}$ 的属性约简集 R_U , 邻域优势条件熵 $INDE_{D|R_U}^{\geq}(U)$ 和 $INDE_{D|C}^{\geq}(U)$, 邻域容差优势关系矩阵 $N_{R_U}^{\geq}(U)$ 、 $N_C^{\geq}(U)$

和 $N_D^{\geq}(U)$ 。

输出: 新信息系统下的属性约简 $R_{U'}$ 和邻域优势条件熵 $INDE_{D|R_{U'}}^{\geq}(U')$ 和 $INDE_{D|C}^{\geq}(U')$, 新信息系统下的邻域容差优势关系矩阵 $N_{R_{U'}}^{\geq}(U')$ 、 $N_C^{\geq}(U')$ 和 $N_D^{\geq}(U')$ 。

具体步骤如下:

1) 初始化 B , $B \leftarrow R_U$ 。

2) 基于定理4得到邻域容差优势关系矩阵 $N_B^{\geq}(U)$ 的分块矩阵 $(M_2^B)_{(n-n^-) \times n^-}$ 、 $(M_3^B)_{n^- \times (n-n^-)}$ 和 $(M_4^B)_{n^- \times n^-}$, $N_C^{\geq}(U)$ 的分块矩阵 $(M_2^C)_{(n-n^-) \times n^-}$ 、 $(M_3^C)_{n^- \times (n-n^-)}$ 和 $(M_4^C)_{n^- \times n^-}$, $N_D^{\geq}(U)$ 的分块矩阵 $(M_2^D)_{(n-n^-) \times n^-}$ 、 $(M_3^D)_{n^- \times (n-n^-)}$ 和 $(M_4^D)_{n^- \times n^-}$, 并更新出新信息系统下的邻域容差优势关系矩阵 $N_C^{\geq}(U')$ 和 $N_D^{\geq}(U')$ 。

3) 根据定理5, 基于 $INDE_{D|B}^{\geq}(U)$ 和 $INDE_{D|C}^{\geq}(U)$ 计算新的邻域优势条件熵 $INDE_{D|B}^{\geq}(U')$ 和 $INDE_{D|C}^{\geq}(U')$ 。

4) 若 $INDE_{D|B}^{\geq}(U') = INDE_{D|C}^{\geq}(U')$, 则进入步骤7), 若 $INDE_{D|B}^{\geq}(U') > INDE_{D|C}^{\geq}(U')$, 则进入步骤5)。

5) 对于 $\forall b \in (C - B)$, 计算属性 b 在 B 下的外部属性重要度 $Sig_{out}^{U'}(b, B, D)$, 找出外部属性重要度最大的属性 $b_{\max}$, $b_{\max} = \arg \{Sig_{out}^{U'}(b, B, D)\}_{b \in C-B}$ 。

6) 对于步骤5)中的属性 $b_{\max}$, 如果 $Sig_{out}^{U'}(b_{\max}, B, D) > 0$, 那么 $B \leftarrow B \cup \{b_{\max}\}$, 并返回步骤5), 否则, 进入步骤7)。

7) 对于 $\forall c \in B$, 若满足关系 $INDE_{D|B-\{c\}}^{\geq}(U') = INDE_{D|B}^{\geq}(U')$, 那么 $B \leftarrow B - \{c\}$, 重复进行步骤7), 直到 B 中所有属性都不满足该关系。

8) 进行 $R_{U'} \leftarrow B$, 返回新信息系统下属性约简 $R_{U'}$ 和新邻域优势条件熵 $INDE_{D|R_{U'}}^{\geq}(U')$ 和 $INDE_{D|C}^{\geq}(U')$, 返回邻域容差优势关系矩阵 $N_{R_{U'}}^{\geq}(U')$ 、 $N_C^{\geq}(U')$ 和 $N_D^{\geq}(U')$ 。

与算法2类似, 算法3同样基于原先旧信息系统的属性约简结果 R_U 进行增量式计算得到新信息系统属性约简。其中, 步骤2)中邻域容差优势关系矩阵的分块矩阵可直接从旧关系矩阵得到, 因此, 其时间复杂度为 $O(1)$; 步骤3)为增量式计算新的邻域优势条件熵, 其时间复杂度为 $O(|U - U^-| \cdot |U^-|)$; 步骤5)和步骤6)在旧信息系统属性约简结果的基础上进一步搜索新的属性, 其时间复杂度为 $O(|\Delta C| \cdot |U - U^-|^2)$, 其中, ΔC 为新搜索到的属性集; 步骤7)为反向冗余属性剔除的过程, 其时间复杂度为 $O(|C| \cdot |U - U^-|^2)$, 因此算法3的时间复杂度为 $O(|U - U^-| \cdot |U^-| + |C| \cdot |U - U^-|^2)$ 。

4 实验分析

表1为加州大学欧文分校UCI机器学习数据库中的8个常用分类学习数据集, 其中, Chess为离散型属

性值的数据集, Wine、Wdbc和Parkinson's Disease为连续型属性值的数据集, AnuranCalls、DryBean、Magic和Musk为混合型属性值的数据集, 本节将在这 8 个数据集上进行属性约简实验来验证本文增量式算法的性能。所有的算法通过软件 Matlab2018b 编程实现, 计算机型号为 CPU Intel(R) Core(TM) i7-7700 4.20 GHz, 内存为 16.0 GB DDR4, 操作系统为 Windows10。

表 1 实验数据集
Tab. 1 Experimental data sets

编号	数据集名称	对象	属性个数	类别数	数据集类型
1	Wine	178	14	3	连续
2	Wdbc	569	31	2	连续
3	Chess	3 196	37	2	离散
4	AnuranCalls	7 195	23	60	混合
5	DryBean	13 611	17	7	混合
6	Magic	19 020	11	2	混合
7	Musk	6 598	168	2	混合
8	Parkinson's Disease	756	754	2	连续

4.1 实验设置与参数确定

在表1所示的数据集中, 对于离散类型的属性值 (也包括决策属性值), 按照数据集样本的序号, 将离散型属性值替换成依次递增的整数型数值, 例如: 对于属性值v1、v2和v3, 将其替换为1、2和3; 对于连续型属性值, 为了消除属性量纲带来的影响, 采用极差正规化的方法进行数据标准化。整个实验中, 统一按照属性值的递增来定义优势关系中的偏序关系, 即属性值大的对象优势于属性值小的对象。此外, 本文研究的是不完备信息系统下的属性约简问题, 表1中的数据均为完备型, 本实验对决策属性以外的属性值随机选择3%进行删除, 从而构造数据集的不完备性。

在本文所提出的算法中, 入参邻域半径 δ 取值的不同会影响到最终的属性约简结果以及分类性能。因此, 为了得到最优的邻域半径值, 将邻域半径 δ 在区间[0.01,0.20]内以0.01为间隔分别作为输入参数输入算法1, 选择出对应的约简集结果, 记录属性约简结果中属性的个数, 同时利用两种分类器(支持向量机(SVM)和邻近点参数 k 设3的 k 邻近(kNN))分类训练获得约简集结果的分类精度。对所有邻域半径的结果进行归纳确定最优邻域半径 δ 值。图2为表1中部分数据集的实验结果。由图2可知, 当邻域半径增大, 约简集长度呈现出减小趋势, 分类精度呈现出先稳定后减小的趋势, 邻域半径在[0.06,0.10]的范围内均具有较高的分类精度水平。属性约简旨在找到属性

集较小且分类精度较高的属性子集, 因此, 经综合比较, 选择邻域半径 $\delta = 0.10$ 进行实验较为合适。

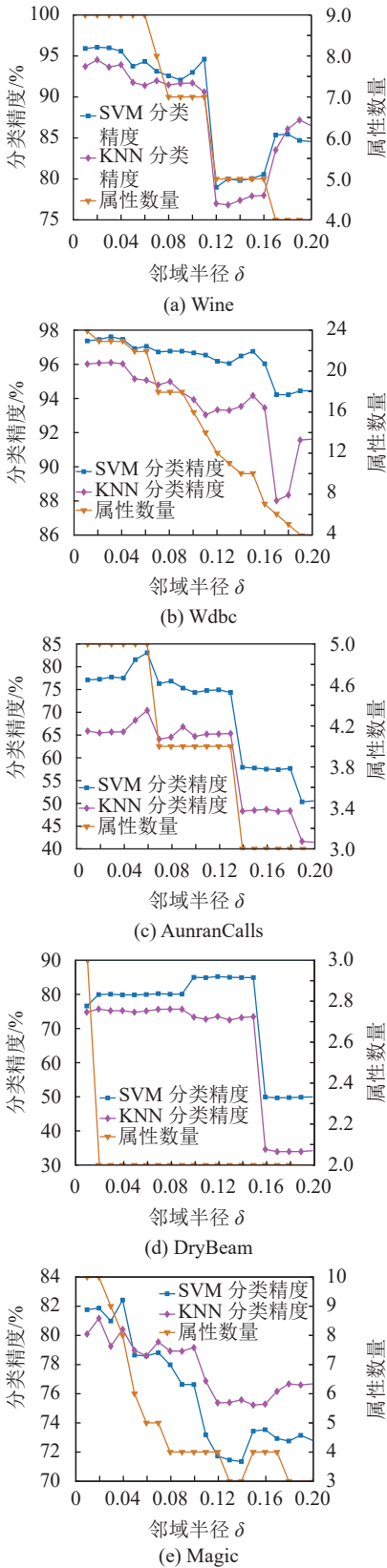


图 2 部分数据集在不同邻域半径下的实验结果
Fig. 2 Experimental results of some data sets under different neighborhood delta

4.2 信息系统对象增加时增量式与非增量式算法性能比较

4.2.1 属性约简结果有效性比较

对于表1中的各个数据集,从完整的论域选择50%对象作为初始信息系统,剩余50%的论域对象作为动态增加的对象集。然后,利用增量式与非增量式算法分别进行动态属性约简,记录两种算法的详细结果,并利用SVM和kNN($k=3$)获得约简结果的分类精度。所有实验结果如表2所示,其中,“{ $*$ }”为属性集结果,“($*$)”为属性的数量。

表 2 对象增加时两类算法的属性约简和分类精度比较

数据集	非增量式属性约简			增量式属性约简		
	属性约简结果	分类精度/%		属性约简结果	分类精度/%	
		SVM	kNN		SVM	kNN
Wine	{3,5,6,9,11,12,13} (7)	93.30	92.55	{3,5,6,11,12,13} (6)	96.22	93.47
Wdbc	{3,5,6,11,13,14,15,17,18,20,21,23,26,27,30,31} (16)	96.72	94.05	{3,5,6,11,13,14,15,16,18,20,21,23,26,27,30,31} (16)	94.37	96.37
Chess	{2,3,4,5,6,7,8,9,11,12,13,14,15,16,17,18,19,21,22,23,24,25,26,27,28,29,31,32,33,34,35,36,37} (33)	87.99	92.65	{2,3,4,5,6,7,8,9,11,12,13,14,16,17,18,19,20,21,22,23,24,25,27,28,31,32,33,34,35,36,37} (31)	88.70	94.03
AnuranCalls	{2,3,4,5} (4)	74.70	65.25	{2,3,4,5} (4)	74.70	65.25
DryBean	{8,14} (2)	86.42	73.33	{5,14} (2)	87.90	72.85
Magic	{2,4,6,9} (4)	76.56	79.23	{2,4,5,7,9} (5)	78.38	80.55
Musk	{1,6,11,14,16,22,29,30,49,57,65,69,72,74,84,88,89,91,95,102,113,124,136,144,148,152,157} (27)	82.43	80.54	{1,6,11,14,16,22,29,30,49,57,65,69,72,74,84,88,91,95,102,113,124,136,144,148,152} (25)	84.51	82.63
Parkinson's Disease	{2,7,25,59,98,123,169,234,276,297,312,353,425,527,578,612} (16)	85.92	81.56	{7,25,59,78,123,169,254,276,290,322,353,455,537,578,667} (15)	88.37	84.53
平均	(13.6)	85.50	82.39	13	86.64	83.71

4.2.2 属性约简算法的效率比较

从所有数据集中随机选择40%的对象作为初始数据集,然后将剩余60%的对象集随机平均分成6份,随后依次将每份往初始数据集中添加,相当于每次增加10%的对象,这样便构造出数据集6次数据量相同的增加更新。将本文的增量式属性约简算法与非增量式算法分别基于这种数据更新方式进行属性约简,统计每次属性约简的计算时间,重复进行30次实验,其平均实验结果如图3所示。在图3中,随着更新次数的增加,两种算法的属性约简运行时间整体上呈逐步上升趋势,但增量式算法的运行时间明显更低,尤其对于大规模的数据集AnuranCalls、Dry-Bean和Magic,两类算法的效率差异更加明显。因此,对于信息系统数据量相同的增加更新,增量式的属性约简要更加高效。

同时,为了比较信息系统不同数据量增加时两类算法的更新效率,同样随机选择40%的对象作为初始数据集,剩余60%的对象集随机平均分成6份;然后,每次分别往初始数据集添加不同量的剩余对象,其中,第1次随机选择1份对象集添加至初始数据集,

由表2可知,两种算法的约简结果非常接近。例如:数据集AnuranCalls有着相同的属性约简结果;对于数据集Wine、Chess、Musk和Parkinson's Disease,增量式属性约简算法的约简结果更小一些,同时有着更高的分类精度;数据集Wdbc和DryBean有着属性数量相同但属性不完全一样的约简集,其分类精度也很接近;对于数据集Magic,虽然增量式算法约简得到的属性稍多一些,但是分类精度更高一些。表2的结果表明,两种算法都是可行且有效的;相比非增量式算法,增量式算法的属性约简结果更优异。

第2次随机选择2份对象集进行添加,依次类推,直至最后一次性添加6份对象集,这相当于初始数据集分别添加了10%、20%、…、60%的对象,因此,构造出了信息系统6次数据量不同的增加。基于这种更新方式,让两类算法分别进行属性约简更新计算,重复进行实验30次,其平均计算时间实验结果如图4所示。在图4中,同样可以发现增量式算法的计算时间更低,因此,对于不同数据量的增加更新,增量式的属性约简方法也更加高效。

图3和4中两类算法表现出的计算效率差异,主要是由于这两类算法的计算架构不同,算法1表明,非增量式算法每次对新的信息系统重新进行属性约简的搜索,其计算量与信息系统当前的规模正相关。在算法2中,增量式算法以旧信息系统的属性约简为基础,对变化的对象进行增量式计算,即算法所产生的计算量大部分来源于属性约简结果中变化的属性,不必对原先属性约简中的属性进行重复搜索计算,大幅提高了计算效率,因此,更新属性约简的时间会更少。部分数据集的增量式算法计算时间出现了一定波动,主要是由增量式算法的计算架构决定的,当

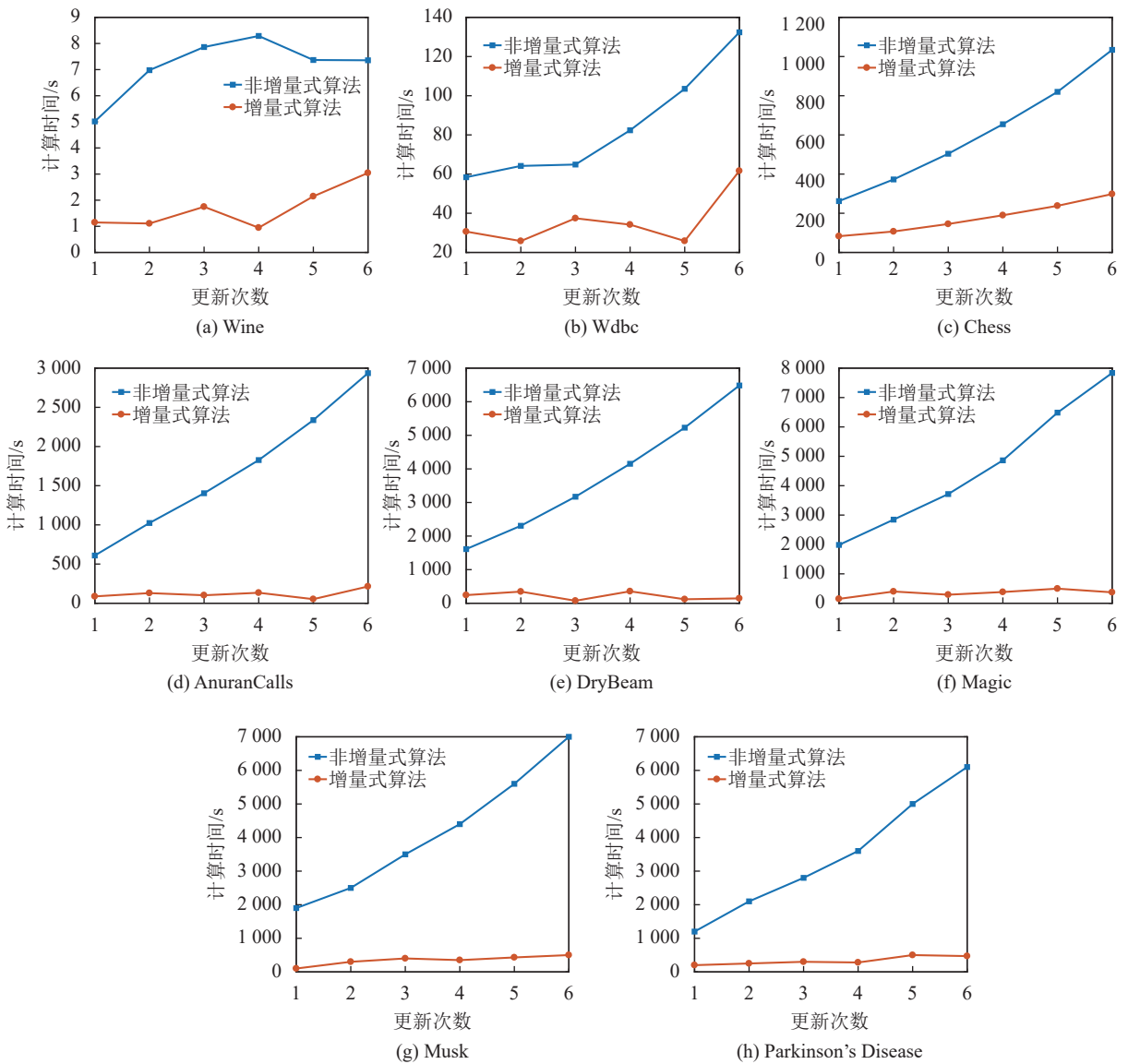


图3 增加相同数据量时更新属性约简的用时比较

Fig. 3 Time comparison of updating attribute reduction when increasing the same amount of data

旧的信息系统更新至新的信息系统,如果旧的约简集能直接继承作为新信息系统的约简,那么增量式算法可以直接返回无需搜索新的属性,此时,增量式算法所需的耗时很短;如果旧的约简集不能作为新信息系统的约简,那么需要选择新属性更新约简结果,此时增量式算法就会产生一定的运行时间,所以,出现用时波动的情况。

4.3 信息系统对象减少时增量式与非增量式算法性能比较

4.3.1 属性约简结果有效性比较

首先将表1中的各个完整数据集作为初始数据集,然后随机选择论域的50%作为动态减少的对象,利用增量式算法与非增量式算法分别进行动态属性约简,并使用SVM和KNN(k 为3)分别获得属性约简的

分类精度,所有实验结果如表3所示,其中“{ }”为属性集结果,“()”为属性的数量。

在表3中,数据集DryBeam下两类算法的约简结果完全一致,而在数据集Wine和Chess中,增量式算法的约简集稍大一点,数据集Wdbc、AnuranCalls和Magic拥有相同的属性约简集大小,但增量式算法分类精度更高。因此,对于信息系统对象减少的情形,增量式架构的属性约简算法在约简的分类精度方面更加占据优势,这主要是由于增量式算法能够保留旧信息系统中的有效属性,提升新信息系统的分类性能。

4.3.2 属性约简算法的效率比较

首先将表1中的每个完整的实验数据集作为初始数据集,然后对每个数据集随机选择60%的对象集,

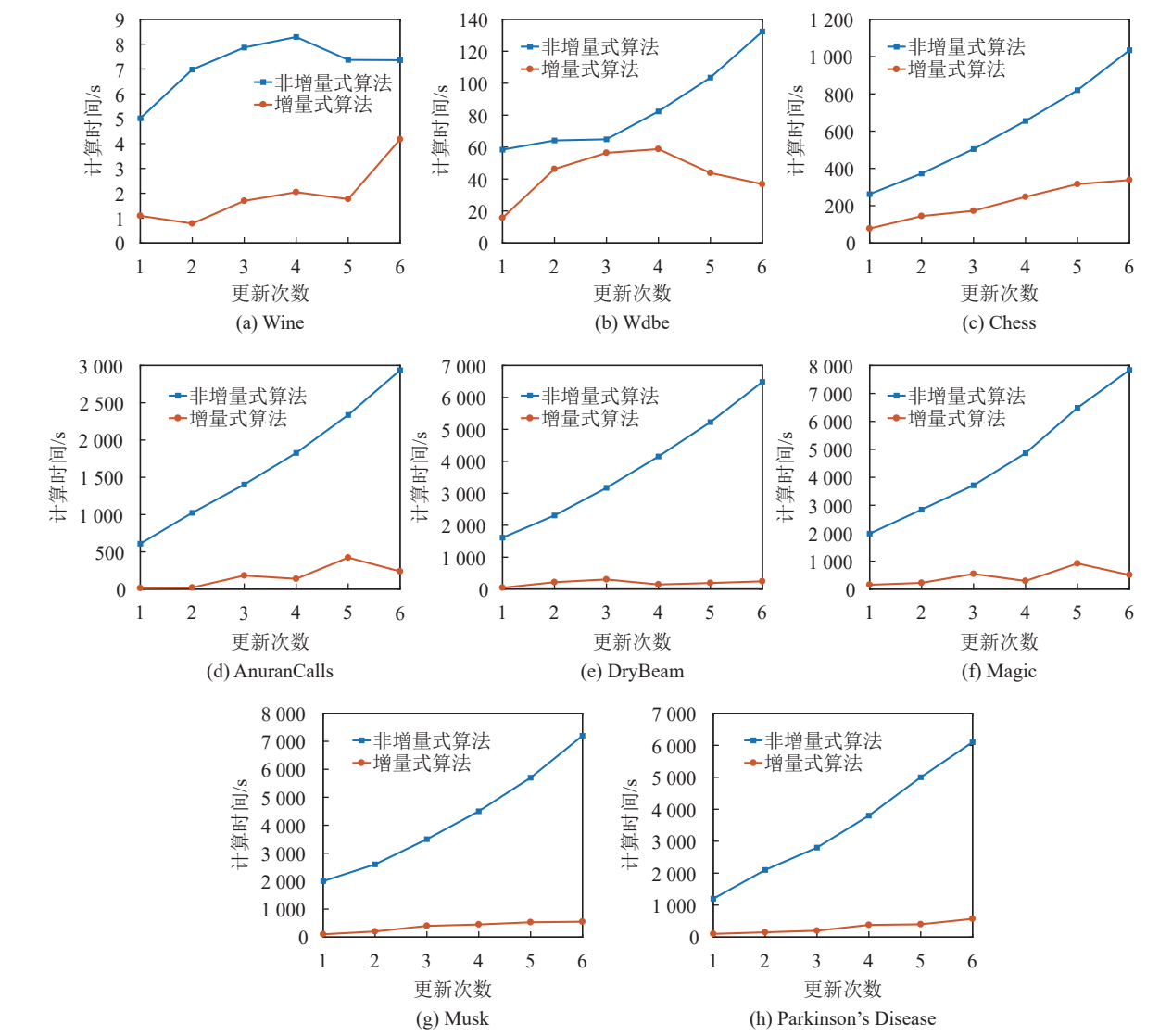


图 4 增加不同数据量时更新属性约简的用时比较

Fig. 4 Time comparison of updating attribute reduction when increasing the different amount of data

表 3 对象减少时两类算法的属性约简和分类精度比较

Tab. 3 Comparison of attribute reduction and classification accuracy between the two algorithms when objects are removed

数据集	非增量式属性约简			增量式属性约简		
	属性约简结果	分类精度/%		属性约简结果	分类精度/%	
		SVM	KNN		SVM	KNN
Wine	{3,5,9,11,12} (5)	87.40	88.17	{3,5,6,9,11,13} (6)	89.45	90.85
Wdbc	{3,11,13,15,16,18,21,23,26,30,31} (11)	88.57	91.92	{5,6,14,16,18,20,23,26,27,30,31} (11)	90.05	93.75
Chess	{2,3,4,5,6,7,8,9,11,12,13,14,16,17,18,19,21,22,24,25,27,28,31,32,33,34,35,36,37} (29)	80.33	91.33	{2,3,4,5,6,7,8,9,11,12,13,14,16,17,18,19,21,22,23,24,25,27,28,31,32,33,34,35,36,37} (30)	83.30	95.72
AnuranCalls	{2,3,20} (3)	72.03	67.17	{2,4,10} (3)	79.11	69.11
DryBean	{8,14} (2)	83.52	72.85	{8,14} (2)	83.52	72.85
Magic	{2,4,6,7} (4)	74.44	78.41	{3,4,6,7} (4)	73.57	79.25
Musk	{1,6,11,13,16,21,29,30,49,57,62,69,72,74,84,88,89,91,102,113,124,144,148,152,159} (25)	81.33	78.94	{1,6,11,14,16,22,29,49,57,64,69,72,77,84,88,95,102,113,124,136,148,155} (22)	82.50	81.17
Parkinson's Disease	{2,7,25,59,98,169,234,276,297,312,353,425,527,578,612} (15)	83.28	80.48	{7,25,59,78,123,254,276,290,353,455,537,578,667} (13)	85.59	82.86
平均	(11.75)	81.36	81.15	(11.37)	83.39	83.20

并平均分成6份,随后依次进行移除操作,相当于每次减少10%的对象,这样便构造出数据集6次相同数据量的减少。将本文的增量式属性约简算法与非增量式算法分别基于这种数据更新方式进行属性约简,统计每次属性约简的计算时间,重复进行30次实验,其平均时间实验结果如图5所示。在图5中,随着更新次数增加,两种算法的属性约简运行时间均逐渐减小,但是增量式算法的运行时间更短,尤其在数据集AnuranCalls、DryBeam和Magic中,两种算法的差异更加明显。

同时,为了比较信息系统不同数据量减少时两种算法的更新效率,将完整的数据集作为初始数据集,随机选择60%的对象集平均分成6份,然后每次分别选择不同量的对象集进行移除,其中第一次移除1份对象集,第二次移除2份,依次类推,直

至最后一次性移除6份对象集,这相当于在初始数据集中分别移除10%、20%、...、60%的对象,因此构造出了信息系统6次不同数据量的减少,基于这种更新方式,让两种算法分别进行属性约简,重复进行30次实验,其平均计算时间实验结果如图6所示。在图6中,同样可以发现增量式算法的计算时间更少,即增量式算法对动态数据有着更高的属性约简效率。

图5和图6表现出的效率差异同样是由于增量式属性约简算法的架构决定的。在算法3中,当信息对象减小时,增量式算法在旧的约简集上更新属性信息,避免对已选择的属性进行重复计算,极大地提高了动态数据环境下属性约简的效率,由于非增量式算法重复地选择属性,使得属性约简的耗时大幅高于增量式算法。

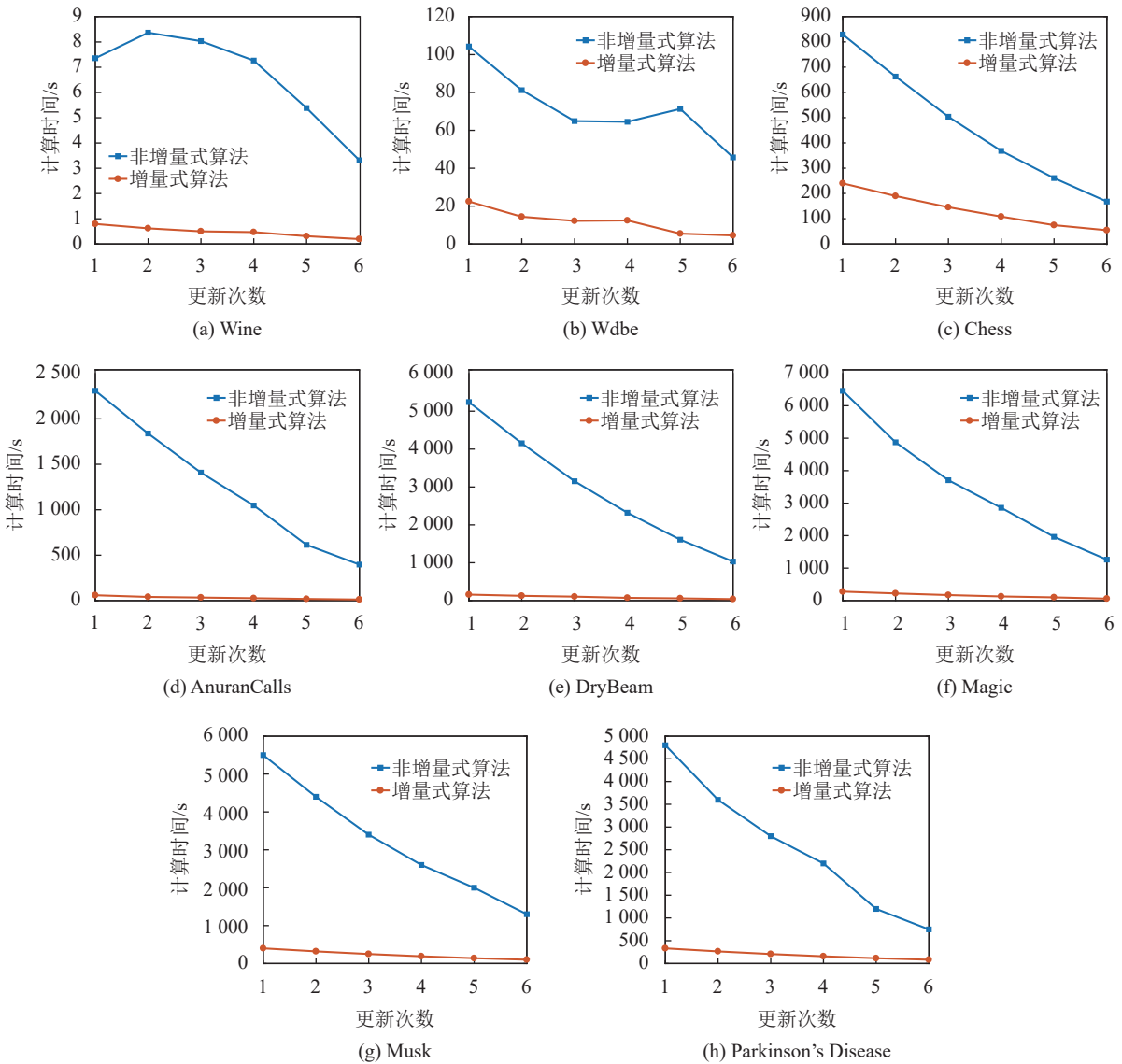


图 5 减少相同数据量时更新属性约简的用时比较

Fig. 5 Time comparison of updating attribute reduction when removing the same amount of data

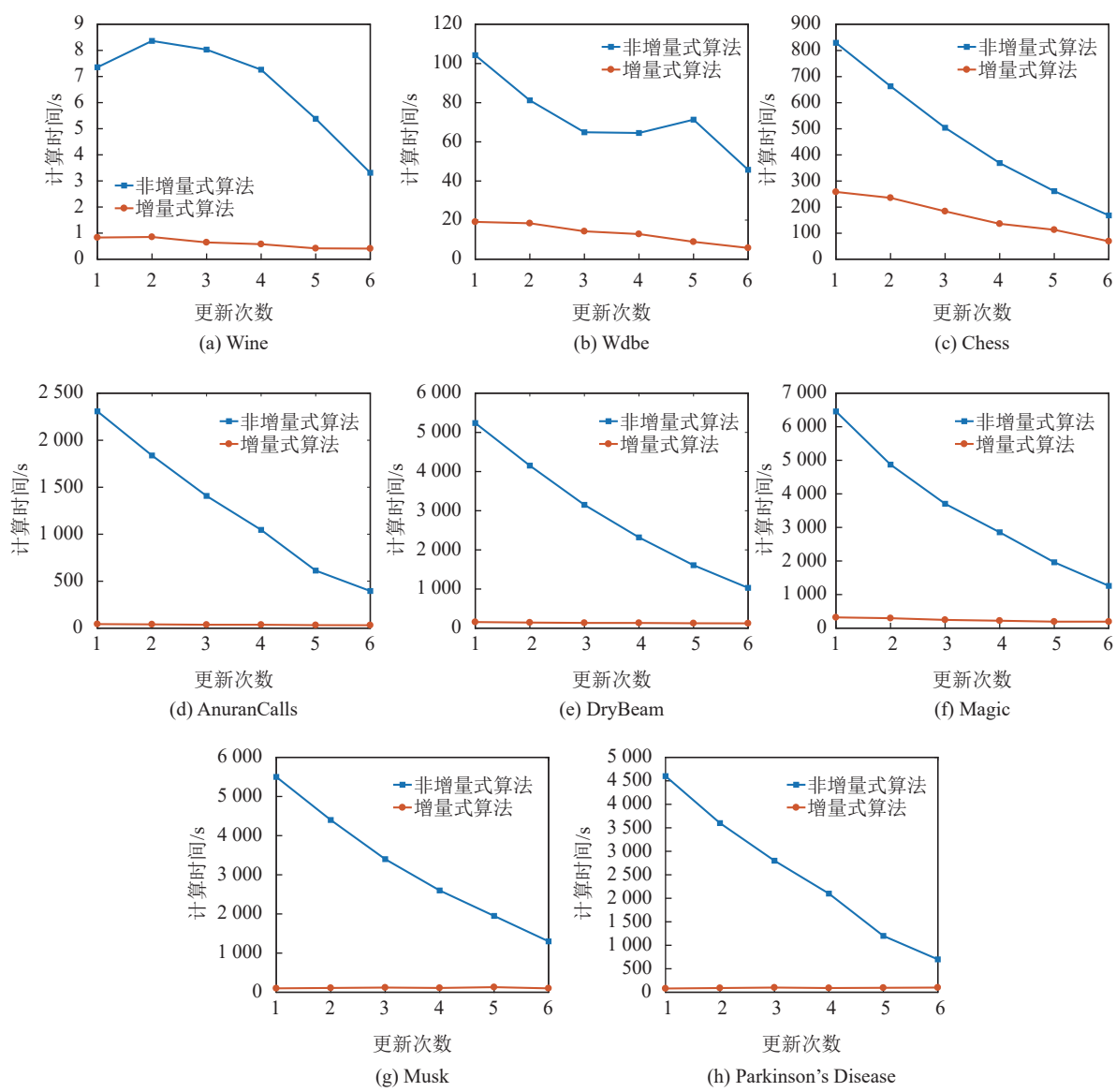


图 6 减少不同数据量时更新属性约简的用时比较

Fig. 6 Time comparison of updating attribute reduction when removing the different amount of data

4.4 本文增量式算法与同类型增量式算法比较

本小节所选择的比较算法分别为：1) 优势粗糙集模型的增量式属性约简^[20]；2) 混合型有序信息系统邻域熵的增量式属性约简^[22]；3) 量化邻域自信息熵的增量式属性约简^[25]，这3种算法分别简记为对比算法1、2、3。其中，对比算法1是一种建立在离散型有序信息系统下的增量式属性约简，在进行实验时需将表1的数据转换成标记型类型，同时这3种对比算法都只适用于完备型的有序信息系统，而本实验选择的数据集均为不完备类型，采用平均值的方法对这些数据集进行缺失值填充，即对属性下所有对象的属性值求取平均值并用平均值对该属性下的缺失值进行填充。随后，将所有增量式算法分别进行信息系统动态属性约简实验。

4.4.1 信息系统对象增加时增量式属性约简比较

按照第4.2.2节对数据集构建6次增加相同数据量的更新方式，将所有增量式算法分别进行属性约简实验。统计每次更新时所有算法的计算用时，重复实验30次，其平均计算时间结果如图7所示。当数据集完成最后一次动态增加，所有算法的约简集长度比较结果见表4，约简集的分类精度比较结果见表5。

在表4中，除数据集Chess和Musk外，本文增量式算法有着更小的属性约简集长度。在表5中，除数据集Wdbc和DryBean外，本文增量式算法的SVM分类精度高于其余算法；除数据集Chess、DryBean和Magic外，本文增量式算法的kNN分类精度高于其余算法。在图7中，本文增量式算法的用时整体上低于其余算法。表4、5和图7的实验结果表明，本文增量式

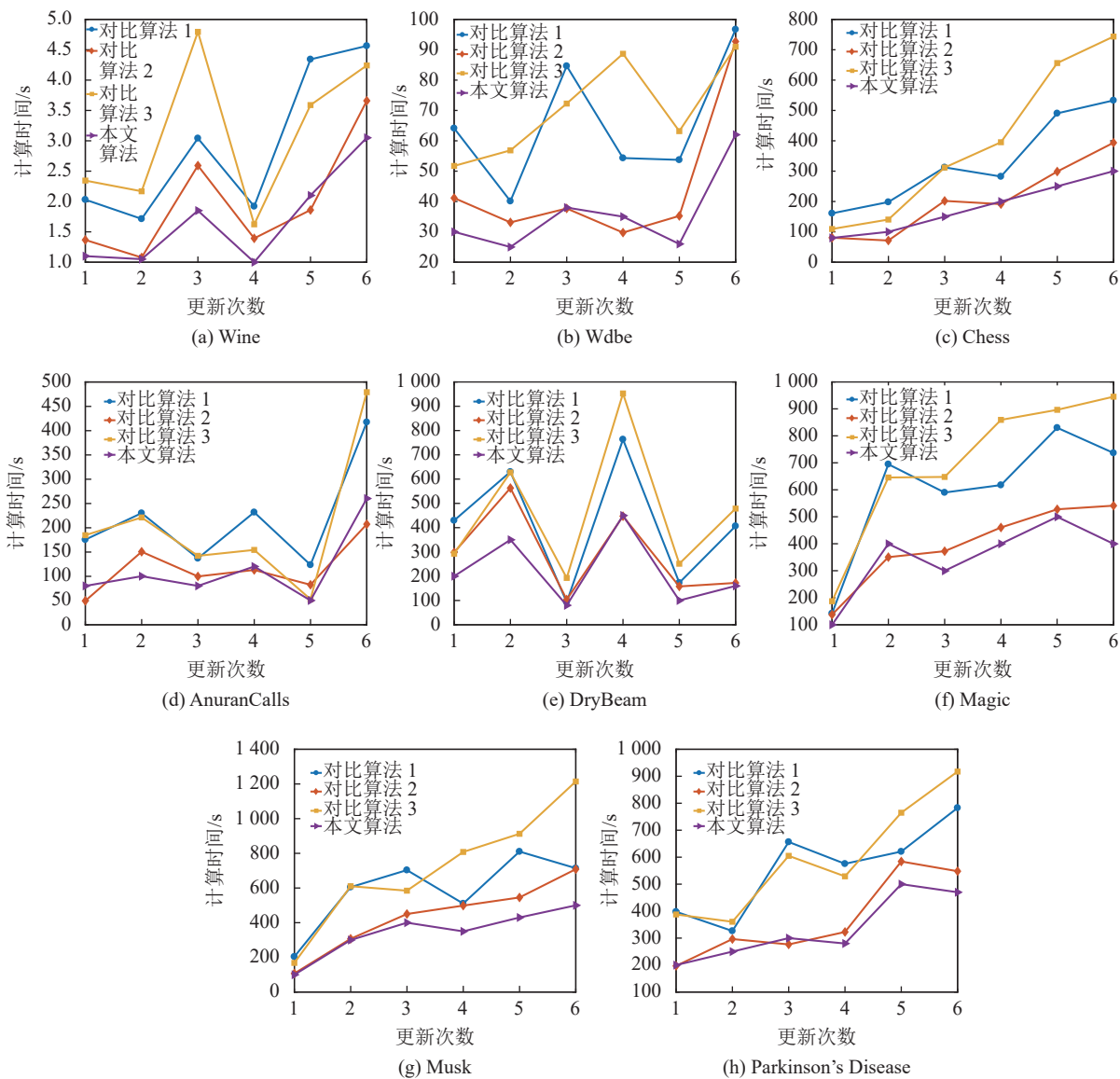


图 7 对象增加时增量式属性约简算法的用时比较

Fig. 7 Time comparison of the incremental attribute reduction algorithms when objects are added

表 4 对象增加时增量式算法的属性约简集长度比较
Tab. 4 Length comparison of attribute reduction set for incremental algorithms when objects are increased

数据集	对比算法1	对比算法2	对比算法3	本文算法
Wine	10	6	6	6
Wdbc	21	18	18	16
Chess	34	31	30	31
AnuranCalls	5	5	4	4
DryBean	4	3	3	2
Magic	7	6	5	5
Musk	29	27	24	25
Parkinson's Disease	18	16	16	15
平均	16.0	14.0	13.3	13.0

算法有着更高的属性约简性能。其原因为对比算法1进行实验时需将数据集进行离散化,这一处理丢失了连续型属性的数据结构信息,因此,对最终的属性约简结果造成了一定的偏差。对比算法2和3是两种适用于完备型信息系统的增量式属性约简方法,而本实验的数据集均为不完备类型,因此,对属性约简的效果产生了一定影响。在增量式约简效率方面,本文增量式算法通过矩阵策略直接更新每个对象的邻域优势类,并进一步更新邻域优势条件熵,跳过了对邻域优势上下近似集的增量式更新,因此,所需的计算量最少,效率整体最高。

4.4.2 信息系统对象减少时增量式属性约简比较

按照第4.3.2节对数据集构建6次减少相同数据量的更新方式,将所有增量式算法分别进行属性约

简,当数据集完成最后一次动态减少后,所有算法的统计数据集每次更新时增量式属性约简的计算用时,约简集分类精度见表6,约简集长度比较结果见表7。重复进行实验30次,其平均实验结果如图8所示。

表 5 对象增加时增量式算法的属性约简集分类精度比较

Tab. 5 Comparison of classification accuracy of the attribute reduction set of the incremental algorithms when objects are increased

数据集	对比算法1		对比算法2		对比算法3		本文算法	
	SVM	kNN	SVM	kNN	SVM	kNN	SVM	kNN
Wine	94.98	92.64	95.92	93.24	95.91	92.25	96.22	93.47
Wdbc	92.12	93.62	95.94	95.91	93.34	94.35	94.37	96.37
Chess	84.49	91.51	87.74	95.03	88.56	93.19	88.70	94.03
AnuranCalls	71.11	61.09	73.07	64.02	74.38	64.72	74.70	65.25
DryBean	85.51	68.42	86.45	71.94	88.45	73.30	87.90	72.85
Magic	74.33	77.96	76.92	79.52	77.35	81.46	78.38	80.55
Musk	81.53	80.12	83.69	80.45	84.21	81.37	84.51	82.63
Parkinson's Disease	83.10	81.96	85.90	82.02	86.72	83.52	88.37	84.53
平均	83.39	80.91	85.70	82.76	86.11	83.02	86.64	83.71

表 6 对象减少时增量式算法的属性约简集分类精度比较

Tab. 6 Comparison of classification accuracy of attribute reduction set of incremental algorithms when objects are reduced

数据集	对比算法1		对比算法2		对比算法3		本文算法	
	SVM	kNN	SVM	kNN	SVM	kNN	SVM	kNN
Wine	86.09	86.27	90.57	89.92	89.02	88.10	89.45	90.85
Wdbc	87.36	90.71	89.60	93.19	88.52	91.08	90.05	93.75
Chess	80.14	91.78	82.29	94.77	82.40	96.79	83.30	95.72
AnuranCalls	75.13	66.66	77.48	68.04	78.48	68.69	79.11	69.11
DryBean	80.83	70.19	82.02	71.81	82.75	72.76	83.52	72.85
Magic	70.71	76.10	71.82	77.41	72.13	78.73	73.57	79.25
Musk	78.34	79.38	81.88	80.47	81.39	82.13	82.50	81.17
Parkinson's Disease	81.76	79.05	83.47	81.70	84.94	81.95	85.59	82.86
平均	80.04	80.01	82.39	82.16	82.45	82.52	83.38	83.19

表 7 对象减少时增量式算法的属性约简集长度比较

Tab. 7 Length comparison of attribute reduction set for incremental algorithms when objects are reduced

数据集	对比算法1	对比算法2	对比算法3	本文算法
Wine	8	6	5	6
Wdbc	15	13	13	11
Chess	35	32	30	30
AnuranCalls	5	4	4	3
DryBean	4	3	3	2
Magic	6	5	4	4
Musk	27	23	21	22
Parkinson's Disease	17	15	14	13
平均	14.6	12.6	11.7	11.3

在表7的实验结果中,除数据集Wine和Musk外,本文增量式算法有着更小的属性约简集长度。在表6的实验结果中:除数据集Wine外,本文增量式算法的SVM分类精度高于其余算法;除数据集Chess和Musk外,本文增量式算法的kNN分类精度高于其余算法。图8的实验结果与图7类似,本文增量式算法整体上有着最少的增量式属性约简用时。表6、7和图8的实验结果表明,在不完备混合型有序信息系统对象减少的情形下,本文增量式算法同样有着更高的增量式属性约简性能,其原因也与第4.4.1节类似。

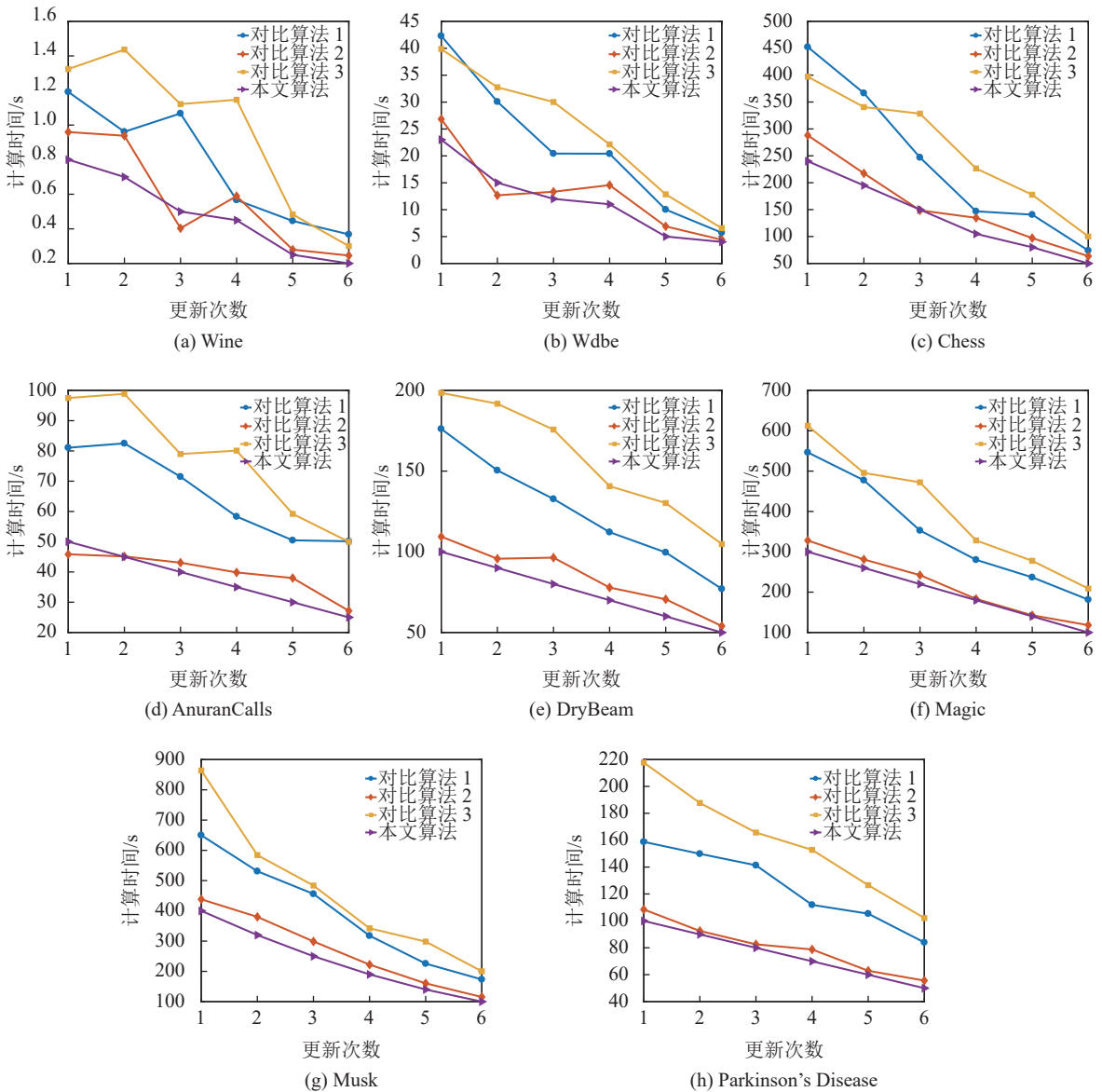


图 8 对象减少时增量式属性约简算法的用时比较

Fig. 8 Time comparison of the incremental attribute reduction algorithms when objects are reduced

5 结 论

增量式属性约简是当前属性约简研究的重点方向, 针对有序信息系统的完备性和混合性, 通过邻域优势条件熵作为启发式函数提出一种增量式属性约简算法。提出不完备混合型有序信息系统下的邻域优势粗糙集模型, 并设计出一种邻域优势条件熵的非增量式属性约简算法; 分别研究了对象增加和减少情形下邻域优势条件熵的增量式学习框架; 提出了不完备混合型有序信息系统的增量式属性约简算法。综合不完备混合型有序信息系统下对象变化的动态属性约简实验, 证明了本文提出的增量式属性约简算法具有更高的属性约简性能。同时, 通过与

同类型增量式属性约简算法相比, 证明了本文增量式属性约简算法具有更高的优越性。在未来的工作中, 将尝试对不完备混合型有序信息系统属性的变化进行增量式属性约简的研究。

附录见本刊网络版(http://jsuese.ijournals.cn/jsuese_cn/ch/reader/view_abstract.aspx?file_no=202201214&flag=1), 扫描标题旁的二维码可阅读网络全文。

参考文献:

- [1] Zhou Tao, Lu Huiling, Ren Hailing, et al. Survey on attribute reduction algorithm of rough set[J]. *Acta Electronica Sinica*, 2021, 49(7): 1439–1449. [周涛, 陆惠玲, 任海玲, 等. 基于粗糙集的属性约简算法综述[J]. *电子学报*, 2021, 49(7): 1439–1449.]

- [2] Li Xueyan, Li Xuewei, Jiang Jun. Multi objective rough set attribute reduction algorithm based on characteristics of knowledge granularity[J]. *Control and Decision*, 2021, 36(1): 196–205. [李雪岩, 李学伟, 蒋君. 基于知识粒度特征的多目标粗糙集属性约简算法[J]. *控制与决策*, 2021, 36(1): 196–205.]
- [3] Liu Dan, Li Jingwei. Incremental updating of fuzzy probability rough sets over two universes based on matrix method[J]. *Control and Decision*, 2021, 36(3): 553–564. [刘丹, 李敬伟. 基于矩阵的双论域模糊概率粗糙集增量更新算法[J]. *控制与决策*, 2021, 36(3): 553–564.]
- [4] Jing Yunge, Li Tianrui, Fujita H, et al. An incremental attribute reduction method for dynamic data mining[J]. *Information Sciences*, 2018, 465: 202–218.
- [5] Wei Wei, Wu Xiaoying, Liang Jiye, et al. Discernibility matrix based incremental attribute reduction for dynamic data[J]. *Knowledge-Based Systems*, 2018, 140: 142–157.
- [6] Liu Ye, Zheng Lidi, Xiu Yeliang, et al. Discernibility matrix based incremental feature selection on fused decision tables[J]. *International Journal of Approximate Reasoning*, 2020, 118: 1–26.
- [7] Xie Xiaojun, Qin Xiaolin. A novel incremental attribute reduction approach for dynamic incomplete decision systems[J]. *International Journal of Approximate Reasoning*, 2018, 93: 443–462.
- [8] Das A K, Sengupta S, Bhattacharyya S A. A group incremental feature selection for classification using rough set theory based genetic algorithm[J]. *Applied Soft Computing*, 2018, 65: 400–411.
- [9] Cai Mingjie, Lang Guangming, Fujita H, et al. Incremental approaches to updating reducts under dynamic covering granularity[J]. *Knowledge-Based Systems*, 2019, 172: 130–140.
- [10] Zhao Xiaolong, Yang Yan. Incremental attribute reduction algorithm based on neighborhood granulation conditional entropy[J]. *Control and Decision*, 2019, 34(10): 2061–2072. [赵小龙, 杨燕. 基于邻域粒化条件熵的增量式属性约简算法[J]. *控制与决策*, 2019, 34(10): 2061–2072.]
- [11] Ni Peng, Zhao Suyun, Wang Xizhao, et al. Incremental feature selection based on fuzzy rough sets[J]. *Information Sciences*, 2020, 536: 185–204.
- [12] Shu Wenhao, Qian Wenbin, Xie Yonghong. Incremental feature selection for dynamic hybrid data using neighborhood rough set[J]. *Knowledge-Based Systems*, 2020, 194: 105516.
- [13] Duan Hailing, Wang Guangqiong. 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. [段海玲, 王光琼. 一种高效的复杂信息增量式属性约简[J]. *华南理工大学学报(自然科学版)*, 2019, 47(6): 18–30.]
- [14] Sheng Kui, Wang Wei, Bian Xianfu, et al. Neighborhood discernibility degree incremental attribute reduction algorithm for mixed data[J]. *Acta Electronica Sinica*, 2020, 48(4): 682–696. [盛魁, 王伟, 卞显福, 等. 混合数据的邻域区分度增量式属性约简算法[J]. *电子学报*, 2020, 48(4): 682–696.]
- [15] Liu Guizhi. Incremental attribute reduction of incomplete hybrid data based on dimension change[J]. *Computer Engineering and Applications*, 2021, 57(12): 161–169. [刘桂枝. 维度变化的不完备混合型数据增量式属性约简[J]. *计算机工程与应用*, 2021, 57(12): 161–169.]
- [16] Du Wensheng, Hu Baoqing. Dominance-based rough set approach to incomplete ordered information systems[J]. *Information Sciences*, 2016, 346/347: 106–129.
- [17] Wang Shu, Li Tianrui, Luo Chuan, et al. A novel approach for efficient updating approximations in dynamic ordered information systems[J]. *Information Sciences*, 2020, 507: 197–219.
- [18] Hu Chengxiang, Zhang Li. Efficient approaches for maintaining dominance-based multigranulation approximations with incremental granular structures[J]. *International Journal of Approximate Reasoning*, 2020, 126: 202–227.
- [19] Sang Binbin, Chen Hongmei, Yang Lei, et al. Feature selection for dynamic interval-valued ordered data based on fuzzy dominance neighborhood rough set[J]. *Knowledge-Based Systems*, 2021, 227: 107223.
- [20] Sang Binbin, Yang Liuzhong, Chen Hongmei, et al. Incremental attribute reduction algorithm in dominance-based rough set[J]. *Computer Science*, 2020, 47(8): 137–143. [桑彬彬, 杨留中, 陈红梅, 等. 优势关系粗糙集增量属性约简算法[J]. *计算机科学*, 2020, 47(8): 137–143.]
- [21] Sang Binbin, Chen Hongmei, Yang Lei, et al. Incremental attribute reduction approaches for ordered data with time-evolving objects[J]. *Knowledge-Based Systems*, 2021, 212: 106583.
- [22] Sang Binbin, Chen Hongmei, Li Tianrui, et al. Incremental approaches for heterogeneous feature selection in dynamic ordered data[J]. *Information Sciences*, 2020, 541: 475–501.
- [23] Fang Huan, Zheng Xuewen, Wang Wusong. Change detection and model repair method of business system under incomplete event logs[J]. *Computer Integrated Manufacturing Systems*, 2021, 27(9): 2647–2660. [方欢, 郑雪文, 王吴松. 不完备事件日志下的业务系统变化检测及模型修复方法[J].

- 计算机集成制造系统,2021,27(9):2647–2660.]
- [24] Zhao Hua,Qin Keyun.Mixed feature selection in incomplete decision table[J].Knowledge-Based Systems,2014,57: 181–190.
- [25] Yang Lei,Qin Keyun,Sang Binbin,et al.A novel incremental attribute reduction by using quantitative dominance-based neighborhood self-information[J].Knowledge-Based Systems,2023,261:110200.

Incremental Attribute Reduction Algorithm Based on Incomplete Hybrid Order Information System

CHEN Baoguo¹, CHEN Lei^{1*}, DENG Ming¹, CHEN Jinlin^{1,2}

(1.School of Computer Sci., Huainan Normal Univ., Huainan 232038, China;

2.College of Electronic and Info. Eng., Nanjing Univ. of Aeronautics and Astronautics, Nanjing 211106, China)

Abstract: Because the data in the big data environment presents the characteristics of dynamic updating, incremental attribute reduction is attracting increasing research attention in the field of rough set theory. As a common information system, the incomplete hybrid ordered information systems (IHOIS) is still without the study of the incremental attribute reduction. To address this issue, an incremental attribute reduction algorithm for object updates is proposed for IHOIS in this paper. Firstly, a neighborhood tolerance dominance relation is proposed to build a new neighborhood dominance rough set based on binary relation. In succession, a neighborhood dominance conditional entropy is defined and further serves as a heuristic function to design a non-incremental attribute reduction algorithm for IHOIS. Then, the neighborhood tolerance dominance relation and neighborhood dominance conditional entropy were reconstructed in the form of matrices. In response to the dynamic updates of the IHOIS, matrix-based calculation strategies are applied to study the incremental updates of neighborhood dominance conditional entropy with both the increasing and decreasing of the information system objects. Finally, the update mechanism of neighborhood dominance conditional entropy is utilized to develop the incremental update algorithms for attribute reduction for both the increasing and decreasing of the IHOIS objects. The experimental results show that compared with the non-incremental algorithm, the incremental algorithm reduces the number of attributes by 3.6% on average, improves the classification accuracy by 2.4% on average, and improves the efficiency of attribute reduction by about 10 times on average. Compared with other incremental algorithms, the proposed incremental algorithm reduces the number of attributes by 9.0% on average, improves the classification accuracy by 2.1% on average, and increases the average efficiency of attribute reduction by 94%. Based on the reported results, it can be concluded that the proposed incremental algorithms have higher performance in both performance and efficiency of the attribute reduction task.

Key words: ordered information system; incomplete hybrid; dominance rough set; attribute reduction; incremental; conditional entropy

(编辑 吴芝明)

引用格式: Chen Baoguo,Chen Lei,Deng Ming,et al.Incremental attribute reduction algorithm based on incomplete hybrid order information system[J].Advanced Engineering Sciences,2024,56(1):65–81.[陈宝国,陈磊,邓明,等.基于不完备混合序信息系统的增量式属性约简[J].工程科学与技术,2024,56(1):65–81.]