控制与决策
Control and Decision

基于矩阵的混合型邻域决策粗糙集增量式更新算法

苑红星, 卓雪雪, 竺德, 刘辉

引用本文:
苑红星,卓雪雪,竺德,刘辉. 基于矩阵的混合型邻域决策粗糙集增量式更新算法[J]. 控制与决策, 2022, 37(6): 1621−1631.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2020.1371

基于矩阵的混合型邻域决策粗糙集增量式更新算法

苑红星[1†], 卓雪雪[2], 竺 德[1], 刘 辉[1]

(1. 安徽大学 网络信息中心，合肥 230601；2. 安徽三联学院 计算机工程学院，合肥 230601)

摘 要: 决策粗糙集模型是当前粗糙集理论最为重要的研究分支之一. 然而,由于现实环境下数据类型的复杂多样以及数据的动态更新,使得传统的决策粗糙集模型面临着一定的局限和不足,针对这一问题,提出一种混合型信息系统的邻域决策粗糙集模型,并设计出一种矩阵方法的邻域决策粗糙集增量式更新算法. 首先,将传统的离散型决策粗糙集模型在混合型信息系统下进行推广,提出一种邻域决策粗糙集模型,使得该模型可以直接处理混合型的数据;然后,利用矩阵的方法重新表示该邻域决策粗糙集模型,同时,针对混合型信息系统对象增加和对象减少时的情形,通过矩阵研究邻域决策粗糙集模型的增量式更新, 并从理论上证明这种增量式方法的高效性;最后,基于矩阵的增量式更新方法,提出混合型信息系统邻域决策粗糙集的增量式更新算法. 实验分析表明所提出的增量式更新算法具有一定的有效性和优越性.

关键词: 决策粗糙集；混合型信息系统；邻域；矩阵；增量式更新；效率

中图分类号: TP181 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2020.1371

引用格式: 苑红星, 卓雪雪, 竺德, 等. 基于矩阵的混合型邻域决策粗糙集增量式更新算法[J]. 控制与决策, 2022, 37(6): 1621-1631.

Incremental updating algorithms of neighborhood decision-theoretic rough set model for hybrid data based on matrix

YUAN Hong-xing[1†], *ZHUO Xue-xue*[2], *ZHU De*[1], *LIU Hui*[1]

(1. Network Information Center，Anhui University，Hefei 230601，China；2. School of Computer Engineering，Anhui Sanlian University，Hefei 230601，China)

Abstract: Decision-theoretic rough set model is one of the most important research branches of the rough set theory. However, due to the variety of data types and the dynamic updating of data in the real environment, the traditional decision-theoretic rough set model is faced with certain limitations and deficiencies. To solve this problem, a neighborhood decision-theoretic rough set model of a hybrid information system is proposed, and an incremental updating algorithm of a neighborhood decision-theoretic rough set based on matrix methods is designed. In this paper, the traditional discrete decision-theoretic rough set model is extended to the hybrid information system, and a neighborhood decision-theoretic rough set model is proposed, which can deal with the hybrid data directly. Then, the matrix method is used to represent the neighborhood decision-theoretic rough set model. At the same time, the incremental updating of the neighborhood decision-theoretic rough set model is studied through the matrix in the case of the increase and decrease of the objects in the hybrid information system. The efficiency of this incremental method is proved theoretically. Finally, based on the incremental updating method of matrix, the incremental updating algorithm of the neighborhood decision-theoretic rough set for the hybrid information system is proposed. Experimental results show that the proposed incremental updating algorithm has certain effectiveness and superiority.

Keywords: decision-theoretic rough set；hybrid information system；neighborhood；matrix；incremental updating；efficiency

0 引 言

粗糙集理论[1]是一种处理不确定性数据的数据挖掘模型,近年来已被不断地改进和推广,已广泛应用于人工智能、模式识别和机器学习等领域[2-4].

随着计算机技术的迅速发展,数据呈现出了海量和动态的特征,因此这对传统的机器学习模型和

收稿日期: 2020-10-05；**录用日期:** 2021-04-07.

基金项目: 赛尔网络下一代互联网技术创新项目(NGII20180612, NGII20180624, NGII20190617).

责任编委: 阳春华.

†通讯作者. E-mail: hxyuan@ahu.edu.cn.

算法带来了一定的挑战. 为了解决这一问题,对传统的模型和算法进行增量式学习是当前最为常用的解决方法之一. 在粗糙集理论中,对相关模型和算法进行增量式学习也是该领域的研究热点[5-6]. 例如Wei等[7]利用不可区分矩阵的方法在粗糙集理论下设计出了一种增量式属性约简算法,大幅度提高了动态环境下数据集的属性约简效率[8-11]. 在混合型数据环境下,盛魁等[12]利用邻域区分度的增量式更新构造出一种增量式属性约简算法; 段海玲等[13]利用邻域知识粒度的增量式更新构造出相应的增量式属性约简算法; Shu等[14]利用邻域信息熵的增量式更新构造出一种增量式属性约简算法. 总之,目前关于粗糙集的增量式学习研究,极大地提升了粗糙集理论的实用化性能[15-18].

决策粗糙集模型[19]是传统粗糙集的重要推广. 当前,决策粗糙集模型已成为粗糙集理论中最为活跃的研究分支[20-21]. 针对动态的数据环境,Luo等[22]利用矩阵的方法重新表示了决策粗糙集模型,并在此基础上构造了信息系统对象变化时决策粗糙集模型的增量式更新方法,提升了决策粗糙集模型在现实环境下处理动态数据的能力.

然而,实际应用中的数据类型总是多样的,在粗糙集理论中,信息系统常呈现出离散型属性和连续型属性混合的情形[5,12-14]. 传统的决策粗糙集模型仅适用于离散型的信息系统[22],针对混合型的信息系统仍然面临着一定的局限性. 同时,对于Luo等[22]提出的决策粗糙集增量式更新方法,每次增量式更新时都先进行对象决策代价的计算,然后进行决策区域的划分,因此该增量式方法仍有较大的优化空间.

针对目前决策粗糙集模型以及对应增量式更新方法存在的局限与不足,本文提出一种混合型信息系统的邻域决策粗糙集模型,并在此基础上设计出一种对象动态变化时的增量式更新算法. 首先,将Li等[23]提出的数值型数据邻域决策粗糙集模型进行推广,提出混合型数据下的邻域决策粗糙集模型; 然后,采用矩阵的数据结构去重新表示邻域决策粗糙集模型,并研究其对象变化时的增量式更新; 最后,根据所提出的增量式更新方法,分别设计对象增加和对象减少时的增量式更新算法. 在实验分析中,通过与非增量式算法进行对比,验证了本文算法的有效性,通过与Luo等[22]提出的增量式算法进行对比,验证了本文算法的优越性.

1 混合型信息系统的邻域决策粗糙集

离散型属性和连续型属性并存的混合型信息系统可表示为MIS $=(U, At=C\bigcup D)$,这里的C和D被称为条件属性集和决策属性集,并且$C=C^c\bigcup C^n, C^c\bigcap C^n=\varnothing, C^c$和$C^n$分别被称为条件属性集$C$中的离散型属性集和连续型属性集. 文献[5,12-13]在传统邻域关系的基础上,进一步提出了混合型信息系统的邻域关系.

定义1[5,12-13] 给定混合型信息系统MIS $=(U, C\bigcup D), C=C^c\bigcup C^n$,设属性子集$A\subseteq C$,并且$A=A^c\bigcup A^n$,这里的$A^c\subseteq C^c$且$A^n\subseteq C^n$,由属性子集$A$确定的邻域关系定义为

$$N_A=\{(x,y)\in U\times U|(\forall a\in A^c, a(x)=a(y))\bigwedge d_{A^n}(x,y)\leqslant\delta\}. \tag{1}$$

其中$d_{A^n}(x,y)$表示对象x和y在连续型属性集A^n下的距离度量[5,12-13],δ为邻域半径. 给定邻域关系N_A,可以得到论域中对象$\forall x\in U$在邻域关系N_A下确定的邻域类$\delta_A(x)$,定义为$\delta_A(x)=\{y\in U|(x,y)\in N_A\}$.

定义2 给定混合型信息系统MIS $=(U, C\bigcup D)$,设属性子集$A\subseteq C$,邻域半径为δ. 对象集X在属性子集A下确定的邻域决策粗糙集下近似集和上近似集分别定义为

$$\underline{N}_A^{(\alpha,\beta)}(X)=\{x\in U|P(X|\delta_A(x))\geqslant\alpha\}, \tag{2}$$

$$\overline{N}_A^{(\alpha,\beta)}(X)=\{x\in U|P(X|\delta_A(x))\geqslant\beta\}. \tag{3}$$

其中: $\langle\underline{N}_A^{(\alpha,\beta)}(X), \overline{N}_A^{(\alpha,\beta)}(X)\rangle$称为对象集$X$在属性集$A$下的邻域决策粗糙集; $P(X|\delta_A(x))=|X\bigcap\delta_A(x)|/|\delta_A(x)|$,$\alpha$和$\beta$为邻域决策粗糙集的决策阈值,满足$\alpha\geqslant\beta$,是根据决策代价计算出来的一组确定的值,具体详见文献[23].

2 邻域决策粗糙集的矩阵表示

本节在文献[7,11,17,22]研究的基础上,利用矩阵的方法去重新表示混合型信息系统下的邻域决策粗糙集模型,为后面章节中邻域决策粗糙集的增量式更新提供基础.

定义3 给定混合型信息系统MIS $=(U, C\bigcup D)$,论域$U=\{x_1, x_2, \dots, x_n\}$,对象集$X\subseteq U$的特征向量定义为

$$\boldsymbol{X}_U=[v_1, v_2, \dots, v_n]. \tag{4}$$

其中$v_i=\begin{cases}1, & x_i\in X;\\ 0, & x_i\notin X;\end{cases}\quad 1\leqslant i\leqslant n.$

推论1 给定混合型信息系统MIS $=(U, C\bigcup D)$,设$X, Y\subseteq U$,有

$$|X\bigcap Y|=\boldsymbol{X}_U*\boldsymbol{Y}_U^{\mathrm{T}}. \tag{5}$$

其中: “$*$”为线性代数中矩阵的标准乘法, “T”为向量的转置.

证明　假设$U=\{x_1,x_2,\dots,x_n\}$, $\boldsymbol{X}_U=[v_1,v_2,\dots,v_n]$和$\boldsymbol{Y}_U=[w_1,w_2,\dots,w_n]$, $\boldsymbol{X}_U*\boldsymbol{Y}_U^{\mathrm{T}}=[v_1,v_2,\dots,v_n]\cdot[w_1,w_2,\dots,w_n]^{\mathrm{T}}=\sum\limits_{i=1}^{n}v_iw_i$. 对于$\forall x_i\in U$, 当$x_i\in X$且$x_i\in Y$时有$x_i\in X\bigcap Y$,即$v_i=1$且$w_i=1$时$x_i\in X\bigcap Y$,所以$\sum\limits_{i=1}^{n}v_iw_i=|X\bigcap Y|$. □

定义4　给定混合型信息系统MIS $=(U,C\bigcup D)$,论域$U=\{x_1,x_2,\dots,x_n\}$,邻域半径为δ,设属性子集$A\subseteq C$在论域U下确定的邻域关系为N_A, N_A对应的邻域关系矩阵$\boldsymbol{N}_A$定义为

$$\boldsymbol{N}_A=[m_{ij}]_{n\times n}=\begin{bmatrix}m_{11}&m_{12}&\dots&m_{1n}\\m_{21}&m_{22}&\dots&m_{2n}\\\vdots&\vdots&m_{ij}&\vdots\\m_{n1}&m_{n2}&\dots&m_{nn}\end{bmatrix}.\tag{6}$$

$$m_{ij}=\begin{cases}1,\ (x_i,x_j)\in N_A;\\0,\ (x_i,x_j)\notin N_A;\end{cases}\quad 1\leqslant i,j\leqslant n.$$

定义5　给定混合型信息系统MIS $=(U,C\bigcup D)$,论域$U=\{x_1,x_2,\dots,x_n\}$,邻域半径为δ,设属性子集$A\subseteq C$确定的邻域关系N_A对应的邻域关系矩阵$\boldsymbol{N}_A=[m_{ij}]_{n\times n}$,定义矩阵$\boldsymbol{N}_A$的列基数向量为

$$\boldsymbol{S}_{\boldsymbol{N}_A}=[s_1,s_2,\dots,s_n],\tag{7}$$

其中$s_i=\sum\limits_{j=1}^{n}m_{ji}$.

定义5表明,$\sum\limits_{j=1}^{n}m_{ji}$表示的是邻域关系矩阵$\boldsymbol{N}_A$中第$i$列所有元素的和,根据定义4可以得到$\sum\limits_{j=1}^{n}m_{ji}$为邻域类$\delta_A(x_i)$的大小,即$\sum\limits_{j=1}^{n}m_{ji}=|\delta_A(x_i)|$,因此矩阵$\boldsymbol{N}_A$的列基数向量为

$$\boldsymbol{S}_{\boldsymbol{N}_A}=[|\delta_A(x_1)|,|\delta_A(x_2)|,\dots,|\delta_A(x_n)|].\tag{8}$$

定理1　给定混合型信息系统MIS $=(U,C\bigcup D)$,论域$U=\{x_1,x_2,\dots,x_n\}$,邻域半径为δ,设属性子集$A\subseteq C$确定的邻域关系为N_A,对应的邻域关系矩阵为$\boldsymbol{N}_A=[m_{ij}]_{n\times n}$,对象集$X\subseteq U$对应的特征向量为$\boldsymbol{X}_U$,则$X$在邻域关系$N_A$下的概率分布可表示成向量

$$\boldsymbol{P}_A^U(X)=\left[\frac{|X\bigcap\delta_A(x_1)|}{|\delta_A(x_1)|},\frac{|X\bigcap\delta_A(x_2)|}{|\delta_A(x_2)|},\dots,\frac{|X\bigcap\delta_A(x_n)|}{|\delta_A(x_n)|}\right]=(\boldsymbol{X}_U*\boldsymbol{N}_A)./\boldsymbol{S}_{\boldsymbol{N}_A},\tag{9}$$

其中“./”为矩阵的点除运算,即矩阵中对应元素进行相除.

证明　根据定义3和定义4可以得到

$$\boldsymbol{X}_U*\boldsymbol{N}_A=\boldsymbol{X}_U*[(\delta_1^A)^{\mathrm{T}},(\delta_2^A)^{\mathrm{T}},\dots,(\delta_n^A)^{\mathrm{T}}]=[\boldsymbol{X}_U*(\delta_1^A)^{\mathrm{T}},\boldsymbol{X}_U*(\delta_2^A)^{\mathrm{T}},\dots,\boldsymbol{X}_U*(\delta_n^A)^{\mathrm{T}}].$$

又由推论1, $\boldsymbol{X}_U*(\delta_i^A)^{\mathrm{T}}=|X\bigcap\delta_A(x_i)|$,因此

$$\boldsymbol{X}_U*\boldsymbol{N}_A=[|X\bigcap\delta_A(x_1)|,|X\bigcap\delta_A(x_2)|,\dots,|X\bigcap\delta_A(x_n)|].$$

由定义5可以得到

$$\boldsymbol{S}_{\boldsymbol{N}_A}=[|\delta_A(x_1)|,|\delta_A(x_2)|,\dots,|\delta_A(x_n)|].$$

所以

$$(\boldsymbol{X}_U*\boldsymbol{N}_A)./\boldsymbol{S}_{\boldsymbol{N}_A}=\left[\frac{|X\bigcap\delta_A(x_1)|}{|\delta_A(x_1)|},\frac{|X\bigcap\delta_A(x_2)|}{|\delta_A(x_2)|},\dots,\frac{|X\bigcap\delta_A(x_n)|}{|\delta_A(x_n)|}\right].$$

因此定理1成立. □

定义6　给定向量$\boldsymbol{Y}=[v_1,v_2,\dots,v_n]$,定义向量$\boldsymbol{Y}$的$\lambda_{>}^{\theta}(\boldsymbol{Y})$、$\lambda_{\geqslant}^{\theta}(\boldsymbol{Y})$、$\lambda_{<}^{\theta}(\boldsymbol{Y})$和$\lambda_{\leqslant}^{\theta}(\boldsymbol{Y})$计算为

$$\lambda_{\bullet}^{\theta}(\boldsymbol{Y})=[v_1',v_2',\dots,v_n'],\tag{10}$$

$$v_i'=\begin{cases}1,\ v_i\bullet\theta;\\0,\ \text{otherwise};\end{cases}\quad\bullet\in\{>,\geqslant,<,\leqslant\},1\leqslant i\leqslant n.$$

3　混合型信息系统邻域决策粗糙集的增量式更新

3.1　对象动态减少时模型的增量式更新

通过分析定理1的结果,可以发现利用矩阵的方法进行邻域决策粗糙集模型的计算,其核心还是针对近似对象集关于邻域关系的概率分布向量的计算,因此本节将重点关注概率分布向量,研究其增量式更新计算的方法. 为了后文叙述的必要,这里给出一种特殊类型的邻域关系定义.

定义7　给定混合型信息系统MIS $=(U,C\bigcup D)$,论域$U=\{x_1,x_2,\dots,x_n\}$,邻域半径为δ,对于$U_1,U_2\subseteq U$,其中$U_1=\{x_{p_1},x_{p_2},\dots,x_{p_s}\}$, $U_2=\{x_{q_1},x_{q_2},\dots,x_{q_t}\}$,属性子集$A\subseteq C$且$A=A^c\bigcup A^n$,定义$A$在$U_1\times U_2$下确定的邻域关系为

$$N_A^{U_1\times U_2}=\{(x,y)\in U_1\times U_2|(\forall a\in A^c,a(x)=a(y))\bigwedge d_{A^n}(x,y)\leqslant\delta\}.\tag{11}$$

邻域关系$N_A^{U_1\times U_2}$对应的邻域关系矩阵表示为$\boldsymbol{N}_A^{U_1\times U_2}=[m_{ij}]_{st}$. 定义对象$y\in U_2$在论域$U_1$下的

邻域类为$\delta_A^{U_1}(y)_{y\in U_2}=\{x\in U_1|(x,y)\in N_A^{U_1\times U_2}\}$.

通过定义7可以看出,定义1中属性子集A在论域U下确定的邻域关系N_A即为$N_A^{U\times U}$,邻域关系矩阵$\boldsymbol{N}_A$即为$\boldsymbol{N}_A^{U\times U}$. 对于$U_1=\{x_{p_1},x_{p_2},\dots,x_{p_s}\},U_2=\{x_{q_1},x_{q_2},\dots,x_{q_t}\}$,属性子集$A$在$U_1\times U_2$下确定的邻域关系矩阵$\boldsymbol{N}_A^{U_1\times U_2}$即为选择$\boldsymbol{N}_A^{U\times U}$中第$p_1,p_2,\dots,p_s$行和第$q_1,q_2,\dots,q_t$列后组成的矩阵结果.

定理2 给定混合型信息系统$\mathrm{MIS}=(U,C\bigcup D)$,论域$U=\{x_1,x_2,\dots,x_n\}$,邻域半径为$\delta$,属性子集$A\subseteq C$确定的邻域关系为$N_A^{U\times U}$,对应的邻域关系矩阵为$\boldsymbol{N}_A^{U\times U}=[m_{ij}]_{n\times n}$,对象集$X\subseteq U$对应的特征向量为$\boldsymbol{X}_U$,$X$在邻域关系$N_A^{U\times U}$下的概率分布向量为$\boldsymbol{P}_A^U(X)$. 当混合型信息系统删除了对象集$U^-\subseteq U$,其中$U^-=\{x_{t_1},x_{t_2},\dots,x_{t_n^-}\}$,新的混合型信息系统表示为$\mathrm{MIS}'=(U'=U-U',C\bigcup D)$,并且$X^-=X\bigcap U^-$,$X'=X-X^-$,属性子集$A$在论域$U'\times U'$下确定的邻域关系为$N_A^{U'\times U'}$. 则$X'$在邻域关系$N_A^{U'\times U'}$下的概率分布向量$\boldsymbol{P}_A^{U'}(X')$可通过如下两个步骤增量式更新完成:

1) $$\boldsymbol{P}_A^{U'}(X')=(\boldsymbol{P}_A^U(X).*\boldsymbol{S}_{\boldsymbol{N}_A^{U\times U}}-\boldsymbol{X}_{\boldsymbol{U^-}}^-*\boldsymbol{N}_A^{U^-\times U})./(\boldsymbol{S}_{\boldsymbol{N}_A^{U\times U}}-\boldsymbol{S}_{\boldsymbol{N}_A^{U^-\times U}});\tag{12}$$

2) 删除$\boldsymbol{P}_A^{U'}(X')$中第$t_1,t_2,\dots,t_n^-$个元素.

这里$X_{U^-}^-$为X^-在U^-下的特征向量. $\boldsymbol{N}_A^{U^-\times U}=[m_{ij}]_{n^-n}$为属性子集$A$在论域$U^-\times U$下对应的关系矩阵,$\boldsymbol{S}_{\boldsymbol{N}_A^{U^-\times U}}$为$\boldsymbol{N}_A^{U^-\times U}$的列基数向量.

证明 根据定理1,有

$$\boldsymbol{P}_A^U(X).*\boldsymbol{S}_{\boldsymbol{N}_A^{U\times U}}=\boldsymbol{X}_U*\boldsymbol{N}_A^{U\times U}=[|X\bigcap\delta_A^U(x_1)|,|X\bigcap\delta_A^U(x_2)|,\dots,|X\bigcap\delta_A^U(x_n)|],$$

$$\boldsymbol{X}_{U^-}^-*\boldsymbol{N}_A^{U^-\times U}=[|X^-\bigcap\delta_A^{U^-}(x_1)|,|X^-\bigcap\delta_A^{U^-}(x_2)|,\dots,|X^-\bigcap\delta_A^{U^-}(x_n)|].$$

其中$\delta_A^{U^-}(x_i)$表示对象$x_i\in U$在论域U^-下的邻域类. 由于$X^-\subseteq X$且$\delta_A^{U^-}(x_i)\subseteq\delta_A^U(x_i)$,有$(X^-\bigcap\delta_A^{U^-}(x_i))\subseteq(X\bigcap\delta_A^U(x_i))$. 因此$|X\bigcap\delta_A^U(x_i)|-|X^-\bigcap\delta_A^{U^-}(x_i)|=|X'\bigcap\delta_A^{U'}(x_i)|$,其中$\delta_A^{U'}(x_i)$表示对象$x_i\in U$在论域$U'$下的邻域类. 所以

$$\boldsymbol{P}_A^U(X).*\boldsymbol{S}_{\boldsymbol{N}_A^{U\times U}}=\boldsymbol{X}_{\boldsymbol{U^-}}^-*\boldsymbol{N}_A^{U^-\times U}=[|X'\bigcap\delta_A^{U'}(x_1)|,|X'\bigcap\delta_A^{U'}(x_2)|,\dots,|X'\bigcap\delta_A^{U'}(x_n)|].$$

由定义5可以得到

$$\boldsymbol{S}_{\boldsymbol{N}_A^{U\times U}}=[|\delta_A^U(x_1)|,|\delta_A^U(x_2)|,\dots,|\delta_A^U(x_n)|],$$

故

$$\boldsymbol{S}_{\boldsymbol{N}_A^{U^-\times U}}=[|\delta_A^{U^-}(x_1)|,|\delta_A^{U^-}(x_2)|,\dots,|\delta_A^{U^-}(x_n)|].$$

又$\forall x_i\in U,|\delta_A^U(x_i)|-|\delta_A^{U^-}(x_i)|=|\delta_A^{U'}(x_i)|$,有

$$\boldsymbol{S}_{\boldsymbol{N}_A^{U\times U}}-\boldsymbol{S}_{\boldsymbol{N}_A^{U^-\times U}}=[|\delta_A^{U'}(x_1)|,|\delta_A^{U'}(x_2)|,\dots,|\delta_A^{U'}(x_n)|],$$

所有

$$(\boldsymbol{P}_A^U(X).*\boldsymbol{S}_{\boldsymbol{N}_A^{U\times U}}-\boldsymbol{X}_{\boldsymbol{U^-}}^-*\boldsymbol{N}_A^{U^-\times U})./(\boldsymbol{S}_{\boldsymbol{N}_A^{U\times U}}-\boldsymbol{S}_{\boldsymbol{N}_A^{U^-\times U}})=\Big[\frac{|X'\bigcap\delta_A^{U'}(x_1)|}{|\delta_A^{U'}(x_1)|},\frac{|X'\bigcap\delta_A^{U'}(x_2)|}{|\delta_A^{U'}(x_2)|},\dots,\frac{|X'\bigcap\delta_A^{U'}(x_n)|}{|\delta_A^{U'}(x_n)|}\Big].$$

删除第$t_1,t_2,\dots,t_n^-$个元素便得到最终的$\boldsymbol{P}_A^{U'}(X')$. □

定理2表明,对于删除原先论域中的部分对象,只需要对这些被删除的对象计算特征向量$\boldsymbol{X}_{U^-}^-$、邻域关系矩阵$\boldsymbol{N}_A^{U^-\times U}=[m_{ij}]_{n^-n}$以及列基数向量$\boldsymbol{S}_{\boldsymbol{N}_A^{U^-\times U}}$,然后在原先结果的基础上便可以完成概率分布向量$\boldsymbol{P}_A^{U'}(X')$的更新,最后基于$\boldsymbol{P}_A^{U'}(X')$利用定义6便可以完成新信息系统下邻域决策粗糙集上下近似集的更新计算. 因此,该更新计算过程具有很高的计算效率,避免了对未变化对象的重复计算.

3.2 对象动态增加时模型的增量式更新

类似于3.1节所提出的方法,可以进一步设计出对象动态增加时邻域决策粗糙集模型的增量式更新.

定理3 给定混合型信息系统$\mathrm{MIS}=(U,C\bigcup D)$,论域$U=\{x_1,x_2,\dots,x_n\}$,邻域半径为$\delta$,属性子集$A\subseteq C$确定的邻域关系为$N_A^{U\times U}$,对应的邻域关系矩阵为$\boldsymbol{N}_A^{U\times U}=[m_{ij}]_{n\times n}$,对象集$X\subseteq U$对应的特征向量为$\boldsymbol{X}_U$,$X$在邻域关系$N_A^{U\times U}$下的概率分布向量为$\boldsymbol{P}_A^U(X)$. 当混合型信息系统增加了对象集$U^+=\{x_{n+1},x_{n+2},\dots,x_{n+n^+}\}$时,新的混合型信息系统表示为$\mathrm{MIS}'=(U'=U\bigcup U^+,C\bigcup D)$,并且$X^+\subseteq U^+$,$X'=X\bigcup X^+$. 属性子集$A$在论域$U'$下确定的邻域关系为$N_A^{U'\times U'}$,那么$X'$在邻域关系$N_A^{U'\times U'}$下的概率分布向量$\boldsymbol{P}_A^{U'}(X')$增量式更新为

$$\boldsymbol{P}_A^{U'}(X')=[\boldsymbol{P}_1,\boldsymbol{P}_2].\tag{13}$$

其中

$$\boldsymbol{P}_1=(\boldsymbol{P}_A^U(X).*\boldsymbol{S}_{\boldsymbol{N}_A^{U\times U}}+\boldsymbol{X}_{U^+}^+*\boldsymbol{N}_A^{U^+\times U})./$$

$$(\boldsymbol{S}_{\boldsymbol{N}_A^{U\times U}}+\boldsymbol{S}_{\boldsymbol{N}_A^{U^+\times U}});$$

$$\boldsymbol{P}_2=(\boldsymbol{X}'_{U'}*\boldsymbol{N}_A^{U'\times U^+})./\boldsymbol{S}_{\boldsymbol{N}_A^{U'\times U^+}}.$$

$\boldsymbol{X}_{U^+}^+$ 表示 X^+ 在论域 U^+ 下的特征向量, $\boldsymbol{N}_A^{U^+\times U}=[m_{ij}]_{n+n}$ 表示属性子集 A 在 $U^+\times U$ 下的邻域关系矩阵, $\boldsymbol{X}'_{U'}$ 表示 X' 在论域 U' 下的特征向量, $\boldsymbol{N}_A^{U'\times U^+}=[m_{ij}]_{(n+n^+)n^+}$ 表示属性子集 A 在 $U'\times U^+$ 下的邻域关系矩阵, $\boldsymbol{S}_{\boldsymbol{N}_A^{U^+\times U}}$ 和 $\boldsymbol{S}_{\boldsymbol{N}_A^{U'\times U^+}}$ 分别表示 $\boldsymbol{N}_A^{U^+\times U}$ 和 $\boldsymbol{N}_A^{U'\times U^+}$ 的列基数向量.

证明　由于混合型信息系统的论域增加了对象, 概率分布向量的维度将会增加. 下面将新的概率分布向量分成两个部分, 分别表示为 $\boldsymbol{P}_1$ 和 $\boldsymbol{P}_2$. 由定理1可知

$$\boldsymbol{P}_A^U(X).*\boldsymbol{S}_{\boldsymbol{N}_A^{U\times U}}=\boldsymbol{X}_U*\boldsymbol{N}_A^{U\times U}=$$

$$[|X\bigcap\delta_A^U(x_1)|,|X\bigcap\delta_A^U(x_2)|,\ldots,|X\bigcap\delta_A^U(x_n)|],$$

$$\boldsymbol{X}_{U^+}^+*\boldsymbol{N}_A^{U^+\times U}=$$

$$[|X^+\bigcap\delta_A^{U^+}(x_1)|,|X^+\bigcap\delta_A^{U^+}(x_2)|,\ldots,$$

$$|X^+\bigcap\delta_A^{U^+}(x_n)|].$$

因为 $X\bigcup X^+=X',\delta_A^U(x_i)\bigcup\delta_A^{U^+}(x_i)=\delta_A^{U'}(x_i)$, 并且 $(X\bigcap\delta_A^U(x_i))\bigcap(X^+\bigcap\delta_A^{U^+}(x_i))=\varnothing$, 所以 $|X\bigcap\delta_A^U(x_i)|+|(X^+\bigcap\delta_A^{U^+}(x_i)|=|X'\bigcap\delta_A^{U'}(x_i)|$. 因此

$$\boldsymbol{P}_A^U(X).*\boldsymbol{S}_{\boldsymbol{N}_A^{U\times U}}+\boldsymbol{X}_{U^+}^+*\boldsymbol{N}_A^{U^+\times U}=$$

$$[|X'\bigcap\delta_A^{U'}(x_1)|,|X'\bigcap\delta_A^{U'}(x_2)|,\ldots,|X'\bigcap\delta_A^{U'}(x_n)|].$$

又因为

$$\boldsymbol{S}_{N_A^{U\times U}}=[|\delta_A^U(x_1)|,|\delta_A^U(x_2)|,\ldots,|\delta_A^U(x_n)|],$$

$$\boldsymbol{S}_{N_A^{U^+\times U}}=[|\delta_A^{U^+}(x_1)|,|\delta_A^{U^+}(x_2)|,\ldots,|\delta_A^{U^+}(x_n)|],$$

并且 $\delta_A^U(x_i)\bigcup\delta_A^{U^+}(x_i)=\delta_A^{U'}(x_i),\delta_A^U(x_i)\bigcap\delta_A^{U^+}(x_i)=\varnothing$, 所以 $|\delta_A^U(x_i)|+|\delta_A^{U^+}(x_i)|=|\delta_A^{U'}(x_i)|$. 因此

$$\boldsymbol{S}_{N_A^{U\times U}}+\boldsymbol{S}_{N_A^{U^+\times U}}=$$

$$[|\delta_A^{U'}(x_1)|,|\delta_A^{U'}(x_2)|,\ldots,|\delta_A^{U'}(x_n)|].$$

$$\boldsymbol{P}_1=(\boldsymbol{P}_A^U(X).*\boldsymbol{S}_{\boldsymbol{N}_A^{U\times U}}+\boldsymbol{X}_{U^+}^+*\boldsymbol{N}_A^{U^+\times U})./$$

$$(\boldsymbol{S}_{\boldsymbol{N}_A^{U\times U}}+\boldsymbol{S}_{\boldsymbol{N}_A^{U^+\times U}})=$$

$$\Big[\frac{|X'\bigcap\delta_A^{U'}(x_1)|}{|\delta_A^{U'}(x_1)|},\frac{|X'\bigcap\delta_A^{U'}(x_2)|}{|\delta_A^{U'}(x_2)|},\ldots,$$

$$\frac{|X'\bigcap\delta_A^{U'}(x_n)|}{|\delta_A^{U'}(x_n)|}\Big].$$

对于 $\boldsymbol{P}_2=(\boldsymbol{X}'_{U'}*\boldsymbol{N}_A^{U'\times U^+})./\boldsymbol{S}_{\boldsymbol{N}_A^{U'\times U^+}}$, 由定理1有

$$\boldsymbol{X}'_{U'}*\boldsymbol{N}_A^{U'\times U^+}=$$

$$[|X'\bigcap\delta_A^{U'}(x_{n+1})|,|X'\bigcap\delta_A^{U'}(x_{n+2})|,\ldots,$$

$$|X'\bigcap\delta_A^{U'}(x_{n+n^+})|],$$

$$\boldsymbol{S}_{\boldsymbol{N}_A^{U'\times U^+}}=$$

$$[|\delta_A^{U'}(x_{n+1})|,|\delta_A^{U'}(x_{n+2})|,\ldots,|\delta_A^{U'}(x_{n+n^+})|].$$

所以

$$\boldsymbol{P}_2=(\boldsymbol{X}'_{U'}*\boldsymbol{N}_A^{U'\times U^+})./\boldsymbol{S}_{\boldsymbol{N}_A^{U'\times U^+}}=$$

$$\Big[\Big|\frac{|X'\bigcap\delta_A^{U'}(x_{n+1})|}{\delta_A^{U'}(x_{n+1})}\Big|,\Big|\frac{|X'\bigcap\delta_A^{U'}(x_{n+2})|}{\delta_A^{U'}(x_{n+2})}\Big|,\ldots,$$

$$\Big|\frac{|X'\bigcap\delta_A^{U'}(x_{n+n^+})|}{\delta_A^{U'}(x_{n+n^+})}\Big|\Big].$$

因此 $\boldsymbol{P}_A^{U'}(X')=[\boldsymbol{P}_1,\boldsymbol{P}_2]$. □

定理3同样表明, 在原先论域基础上增加一部分对象, 只需要对这些新增对象计算特征向量 $\boldsymbol{X}_{U^+}^+$、邻域关系矩阵 $\boldsymbol{N}_A^{U^+\times U}=[m_{ij}]_{n+n}$ 和 $\boldsymbol{N}_A^{U'\times U^+}=[m_{ij}]_{(n+n^+)n^+}$ 以及列基数向量 $\boldsymbol{S}_{\boldsymbol{N}_A^{U^+\times U}}$ 和 $\boldsymbol{S}_{\boldsymbol{N}_A^{U'\times U^+}}$, 然后在原先结果的基础上便可以完成概率分布向量 $\boldsymbol{P}_A^{U'}(X')$ 的更新, 最后基于 $\boldsymbol{P}_A^{U'}(X')$ 利用定义6可以得到新信息系统下邻域决策粗糙集上下近似集的更新计算结果. 因此, 该更新计算过程同样具有很高的计算效率, 避免了对未变化对象的重复计算.

4　混合型信息系统邻域决策粗糙集的更新算法

4.1　非增量式更新算法

利用第2节中矩阵的策略去表示邻域决策粗糙集模型, 本节在其基础上提出一种对象减少和增加时模型的更新算法, 也称之为非增量式更新算法, 具体如算法1和算法2所示.

算法1　基于矩阵方法的对象减少时邻域决策粗糙集模型非增量式更新算法.

输入: 混合型信息系统 $\mathrm{MIS}=(U,C\bigcup D)$, 论域 $U=\{x_1,x_2,\ldots,x_n\}$, 邻域半径为 δ, 信息系统的决策代价, 属性子集 $A\subseteq C$ 和对象集 $X\subseteq U$. 论域 U 删除的对象集 $U^-\subseteq U$, 其中 $U^-=\{x_{t_1},x_{t_2},\ldots,x_{t_n^-}\}$, 新的混合型信息系统表示为 $\mathrm{MIS}'=(U'=U-U',C\bigcup D)$, 并且 $X^-=X\bigcap U^-,X'=X-X^-$.

输出: 对象集 X' 在属性子集 A 下的邻域决策下近似集 $\underline{N}_A^{(\alpha,\beta)}(X')$ 和邻域决策上近似集 $\overline{N}_A^{(\alpha,\beta)}(X')$.

1) 根据给定的决策代价计算出模型的阈值 α 和 β, 并初始化 $\underline{N}_A^{(\alpha,\beta)}(X')=\varnothing,\overline{N}_A^{(\alpha,\beta)}(X')=\varnothing$.

2) 计算对象集 X' 在论域 U' 下的特征向量 $\boldsymbol{X}'_{U'}./*$ 定义3*/

3) 根据属性子集 A 计算论域 $U'\times U'$ 下的邻域关

系矩阵$\boldsymbol{N}_A^{U'\times U'}$./*定义1和定义4*/

4) 根据邻域关系矩阵$\boldsymbol{N}_A^{U'\times U'}$计算列基数向量$\boldsymbol{S}_{\boldsymbol{N}_A^{U'\times U'}}$./*定义5*/

5) 根据定理1计算X'的概率分布向量$\boldsymbol{P}_A^{U'}(X')=(\boldsymbol{X}'_U*\boldsymbol{N}_A^{U'\times U'})./\boldsymbol{S}_{\boldsymbol{N}_A^{U'\times U'}}$.

6) 根据定义6计算下近似集$\underline{N}_A^{(\alpha,\beta)}(X')$和上近似集$\overline{N}_A^{(\alpha,\beta)}(X')$对应的特征向量为$\lambda^{\alpha}_{\geqslant}(\boldsymbol{P}_A^{U'}(X'))$和$\lambda^{\beta}_{\leqslant}(\boldsymbol{P}_A^{U'}(X'))$,由定义3返回最终的集合形式结果.

算法1所示的是当混合型信息系统对象减少时,基于矩阵方法的邻域决策粗糙集上下近似集的更新算法. 整个算法1的时间复杂度为$O(a(n-n^-)^2)$.

算法2 基于矩阵方法的对象增加时邻域决策粗糙集模型非增量式更新算法.

输入:混合型信息系统MIS $=(U,C\bigcup D)$,论域$U=\{x_1,x_2,\ldots,x_n\}$,邻域半径为δ,信息系统的决策代价,属性子集$A\subseteq C$和对象集$X\subseteq U$. 论域U增加的对象集U^+,其中$U^+=\{x_{n+1},x_{n+2},\ldots,x_{n+n^+}\}$,新的混合型信息系统表示为MIS$'=(U'=U\bigcup U^+,C\bigcup D)$,并且$X^+\subseteq U^+$,$X'=X\bigcup X^+$.

输出:对象集X'在属性子集A下的邻域决策下近似集$\underline{N}_A^{(\alpha,\beta)}(X')$和邻域决策上近似集$\overline{N}_A^{(\alpha,\beta)}(X')$.

1) 根据给定的决策代价计算出模型的阈值α和β,并初始化$\underline{N}_A^{(\alpha,\beta)}(X')=\varnothing$,$\overline{N}_A^{(\alpha,\beta)}(X')=\varnothing$.

2) 计算对象集X'在论域U'下的特征向量$\boldsymbol{X}'_{U'}$./*定义3*/

3) 根据属性子集A计算论域$U'\times U'$下的邻域关系矩阵$\boldsymbol{N}_A^{U'\times U'}$./*定义1和定义4*/

4) 根据邻域关系矩阵$\boldsymbol{N}_A^{U'\times U'}$计算列基数向量$\boldsymbol{S}_{\boldsymbol{N}_A^{U'\times U'}}$./*定义5*/

5) 根据定理1计算X'的概率分布向量$\boldsymbol{P}_A^{U'}(X')=(\boldsymbol{X}'_U*\boldsymbol{N}_A^{U'\times U'})./\boldsymbol{S}_{\boldsymbol{N}_A^{U'\times U'}}$.

6) 根据定义6计算下近似集$\underline{N}_A^{(\alpha,\beta)}(X')$和上近似集$\overline{N}_A^{(\alpha,\beta)}(X')$对应的特征向量为$\lambda^{\alpha}_{\geqslant}(\boldsymbol{P}_A^{U'}(X'))$和$\lambda^{\beta}_{\leqslant}(\boldsymbol{P}_A^{U'}(X'))$,由定义3返回最终的集合形式结果.

可以看出,算法2的整体结构与算法1是类似的,因此整个算法2的时间复杂度为$O(a(n+n^+)^2)$.

4.2 增量式更新算法

在第2节的基础上,第3节提出了一种基于矩阵策略的邻域决策粗糙集模型的更新方法,这种更新方法通过在原先旧信息系统的模型结果上进行进一步计算,大幅度提高了更新效率. 本节中算法3和算法4所示的是对应的更新算法,这两种算法本文也称之为增量式更新算法.

算法3 基于矩阵方法的对象减少时邻域决策粗糙集模型增量式更新算法.

输入:1) 混合型信息系统MIS $=(U,C\bigcup D)$,论域$U=\{x_1,x_2,\ldots,x_n\}$,邻域半径δ,信息系统的决策代价,属性子集$A\subseteq C$和对象集$X\subseteq U$,邻域关系$N_A^{U\times U}$的邻域关系矩阵$\boldsymbol{N}_A^{U\times U}=[m_{ij}]_{n\times n}$,对象集$X$的特征向量$\boldsymbol{X}_U$,$X$的概率分布向量$\boldsymbol{P}_A^U(X)$,邻域关系矩阵$\boldsymbol{N}_A^{U\times U}$的列基数向量$\boldsymbol{S}_{\boldsymbol{N}_A^{U\times U}}$.

2) 论域U删除的对象集$U^-\subseteq U$,其中$U^-=\{x_{t_1},x_{t_2},\ldots,x_{t_n^-}\}$,新的混合型信息系统表示为MIS$'=(U'=U-U',C\bigcup D)$,并且$X^-=X\bigcap U^-$,$X'=X-X^-$.

输出:对象集X'在属性子集A下的邻域决策下近似集$\underline{N}_A^{(\alpha,\beta)}(X')$和邻域决策上近似集$\overline{N}_A^{(\alpha,\beta)}(X')$.

1) 根据给定的决策代价计算出模型的阈值α和β,并初始化$\underline{N}_A^{(\alpha,\beta)}(X')=\varnothing$,$\overline{N}_A^{(\alpha,\beta)}(X')=\varnothing$.

2) 计算对象集X^-在论域U^-下的特征向量$\boldsymbol{X}^-_{U^-}$./*定义3*/

3) 根据属性子集A计算论域$U^-\times U$下的邻域关系矩阵$\boldsymbol{N}_A^{U^-\times U}$./*定义7*/

4) 根据邻域关系矩阵$\boldsymbol{N}_A^{U^-\times U}$计算列基数向量$\boldsymbol{S}_{\boldsymbol{N}_A^{U^-\times U}}$./*定义5*/

5) 根据定理2计算概率分布向量$\boldsymbol{P}_A^{U'}(X')$.

6) 删除$\boldsymbol{P}_A^{U'}(X')$中第$t_1,t_2,\ldots,t_n^-$个元素.

7) 根据定义6计算下近似集$\underline{N}_A^{(\alpha,\beta)}(X')$和上近似集$\overline{N}_A^{(\alpha,\beta)}(X')$对应的特征向量为$\lambda^{\alpha}_{\geqslant}(\boldsymbol{P}_A^{U'}(X'))$和$\lambda^{\beta}_{\leqslant}(\boldsymbol{P}_A^{U'}(X'))$,由定义3返回最终的集合形式结果.

算法3所示的是当混合型信息系统对象减少时,基于矩阵方法的邻域决策粗糙集上下近似集的增量式更新算法. 整个算法3的时间复杂度为$O(ann^-)$.

算法4 基于矩阵方法的对象增加时邻域决策粗糙集模型增量式更新算法.

输入:1) 混合型信息系统MIS $=(U,C\bigcup D)$,论域$U=\{x_1,x_2,\ldots,x_n\}$,邻域半径δ,信息系统的决策代价,属性子集$A\subseteq C$和对象集$X\subseteq U$,邻域关系$N_A^{U\times U}$的邻域关系矩阵$\boldsymbol{N}_A^{U\times U}=[m_{ij}]_{n\times n}$,对象集$X$的特征向量$\boldsymbol{X}_U$,$X$的概率分布向量$\boldsymbol{P}_A^U(X)$,邻域关系矩阵$\boldsymbol{N}_A^{U\times U}$的列基数向量$\boldsymbol{S}_{\boldsymbol{N}_A^{U\times U}}$.

2) 论域U增加的对象集U^+,其中$U^+=\{x_{n+1},x_{n+2},\ldots,x_{n+n^+}\}$,新的混合型信息系统表示为MIS$'=(U'=U\bigcup U^+,C\bigcup D)$,并且$X^+\subseteq U^+$,$X'=X\bigcup X^+$.

输出:对象集X'在属性子集A下的邻域决策下近似集$\underline{N}_A^{(\alpha,\beta)}(X')$和邻域决策上近似集$\overline{N}_A^{(\alpha,\beta)}(X')$.

1) 根据给定的决策代价计算出模型的阈值α和β,并初始化$\underline{N}_A^{(\alpha,\beta)}(X')=\varnothing, \overline{N}_A^{(\alpha,\beta)}(X')=\varnothing$.

2) 计算对象集X^+在论域U^+下的特征向量$\boldsymbol{X}_{U^+}^+$./*定义3*/

3) 根据属性子集A计算论域$U^+\times U$下的邻域关系矩阵$\boldsymbol{N}_A^{U^+\times U}$和计算论域$U'\times U^+$下的邻域关系矩阵$\boldsymbol{N}_A^{U'\times U^+}$./*定义7*/

4) 根据邻域关系矩阵$\boldsymbol{N}_A^{U^+\times U}$和$\boldsymbol{N}_A^{U'\times U^+}$分别计算列基数向量$\boldsymbol{S}_{\boldsymbol{N}_A^{U^+\times U}}$和$\boldsymbol{S}_{\boldsymbol{N}_A^{U'\times U^+}}$./*定义5*/

5) 根据定理3计算概率分布向量$\boldsymbol{P}_A^{U'}(X')=[\boldsymbol{P}_1,\boldsymbol{P}_2]$.

6) 根据定义6计算下近似集$\underline{N}_A^{(\alpha,\beta)}(X')$和上近似集$\overline{N}_A^{(\alpha,\beta)}(X')$对应的特征向量为$\lambda_{\geqslant}^{\alpha}(\boldsymbol{P}_A^{U'}(X'))$和$\lambda_{\leqslant}^{\beta}(\boldsymbol{P}_A^{U'}(X'))$,由定义3返回最终的集合形式结果.

算法4所示的是当混合型信息系统对象增加时,基于矩阵方法的邻域决策粗糙集上下近似集的增量式更新算法. 整个算法4的时间复杂度为$O(ann^+ + a(n^+)^2)$.

5　实验分析

本节将通过实验分析的方法验证所提出算法的有效性和优越性. 本实验所运行的硬件环境为CPU Intel i5 7500 3.4 GHz,内存8 GB,操作系统为Windows 10专业版,所有算法采用Matlab 2015 b进行编码实现并运行. 实验中所使用的8个标准数据集如表1所示,这8个数据集均来源于UCI数据集库,都为离散型属性和连续型属性并存的混合型数据集.

表1所示的数据集,需要进行人工处理,模拟出数据集的动态增加和减少情形. 本实验采用与文献[9-13]类似的处理方式,将数据集按照论域进行分割,分割成的每个子数据集作为每次动态变化的数据集,从而构造出数据集的动态更新. 本实验将数据集的论域大致平均分割成10个部分进行实验. 对于本文所提出的所有算法,文献[5]通过实验验证了邻域半径δ和阈值α、β的取值不同主要对模型的计算结果产生较大的影响,而对算法的效率影响较小. 由于本文的研究内容主要集中在算法的效率这一方面,为了简便,本实验将这些参数取特定的值进行实验,在文献[2, 5, 10, 12-14, 23]中,大多数学者将邻域半径δ选取在0.1～0.2之间,本实验选取$\delta=0.2$进行实验,对于阈值α和β,本实验在参考了文献[22]的基础上选取$\alpha=0.80, \beta=0.55$.

表1　实验数据集

序号	数据集	对象	属性	类
1	Cylinder	512	40	3
2	Credit	690	15	2
3	Segment	2 310	19	7
4	Abalone	4 177	8	29
5	Characters	6 000	7	10
6	Thyroid	9 172	29	2
7	Weight Lifting Exercises	39 242	152	3
8	Music Analysis	106 574	518	4

5.1　增量式算法与非增量式算法的效率比较

本节将所提出的增量式更新算法与非增量式更新算法在各个数据集下进行模型的更新效率比较. 图1展示的是对象增加时,增量式更新算法与非增量式更新算法的更新用时比较结果,其中横坐标表示更新次数,纵坐标表示更新所消耗的时间. 观察图1的每个子图可以发现,随着混合型信息系统更新次数的逐渐增加,非增量式算法的更新用时快速地增长,增量式算法的更新用时增长的较为缓慢,增量式算法的更新效率远高于非增量式算法. 产生这一结果的

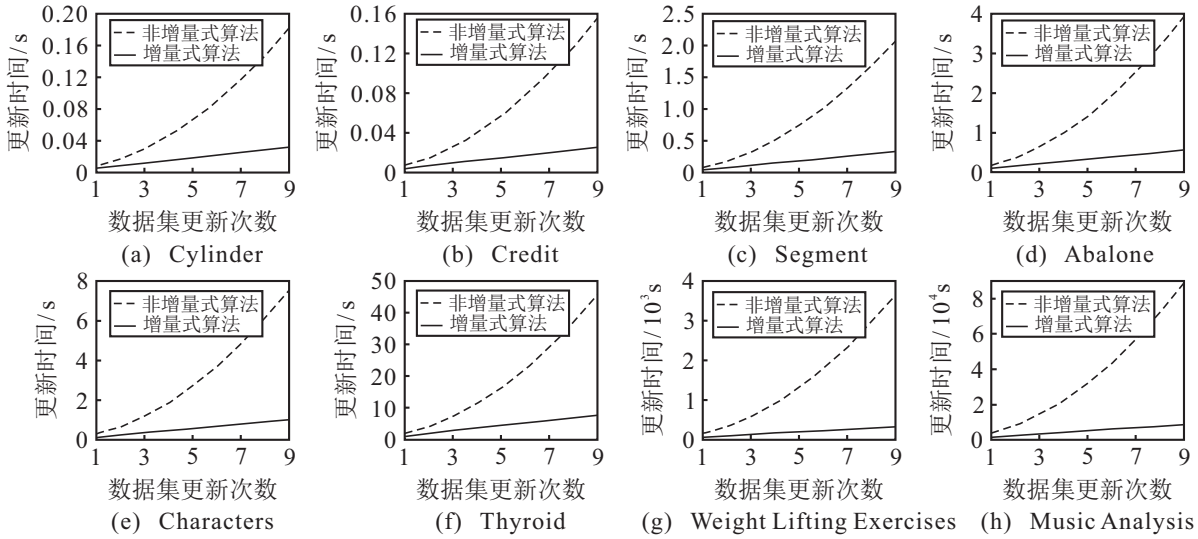

图1　对象增加时非增量式算法与增量式算法的更新时间比较

主要原因,是由于增量式算法有效地利用了前一次模型的计算结果,避免了对旧数据的重复计算;而非增量式算法每次都是基于当前完整的信息系统进行模型计算,这其中包含了大量的重复计算,因此会消耗更多的计算时间.

图2展示的是对象减少时,增量式更新算法与非增量式更新算法的更新用时比较结果,其中横坐标表示更新次数,纵坐标表示更新所消耗的时间. 观察图2中各个子图的结果可以发现,随着信息系统论域的逐渐减小,非增量式算法和增量式算法更新模型所需的用时均是逐渐减小的,但是增量式算法更新模型的效率大幅度高于非增量式算法,这主要是由于增量式算法基于前一次模型的结果进行增量式计算,有效地避免了数据的重复计算,因此更新效率更高.

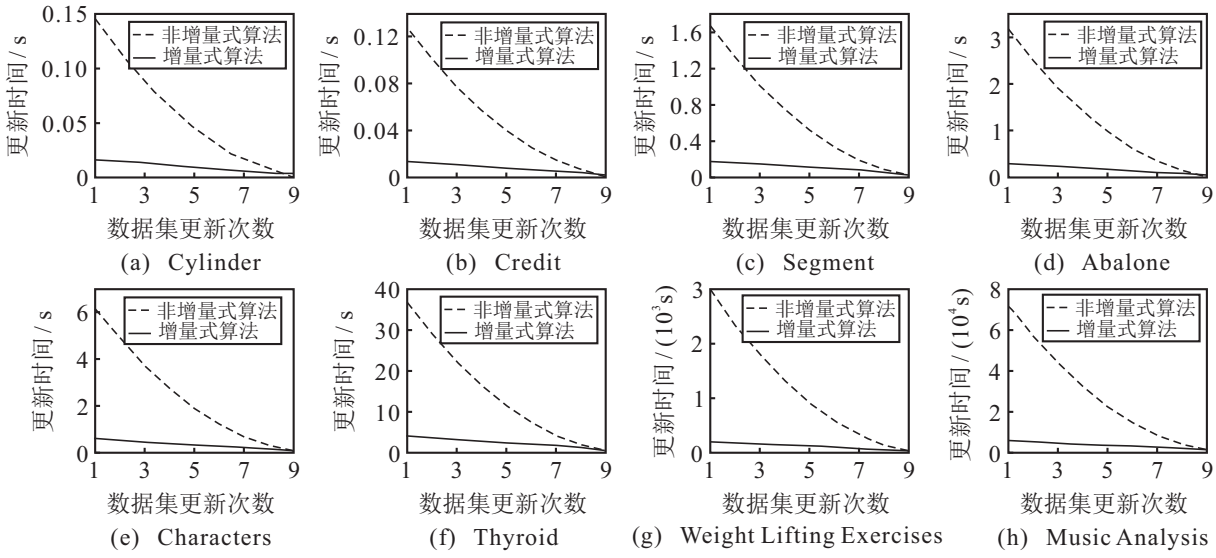

图 2 对象减少时非增量式算法与增量式算法的更新时间比较

5.2 增量式算法与其他同类型算法比较

上节中的实验结果表明了增量式算法比非增量式算法具有更高的更新效率,本节将所提出的增量式算法与文献[22]提出的决策粗糙集增量式更新算法进行实验分析对比,用来验证本文算法的优越性. 记文献[22]提出的增量式更新算法为对比增量式算法,该算法仅适用于离散型的信息系统,因此在进行实验时需要将所有实验数据集中的连续型属性进行离散化处理. 图3和图4分别所示的是当混合型信息系统对象增加和减少时,本文增量式算法与对比增量式算法进行模型更新时的用时比较结果,其中横坐标表示更新次数,纵坐标表示更新所消耗的时间.

在图3中,随着信息系统论域的增大,对于数据集Cylinder和Credit,对比增量式算法的更新用时略低于本文增量式算法,对于其余的数据集,本文算法的更新用时均低于对比增量式算法. 比较各个数据

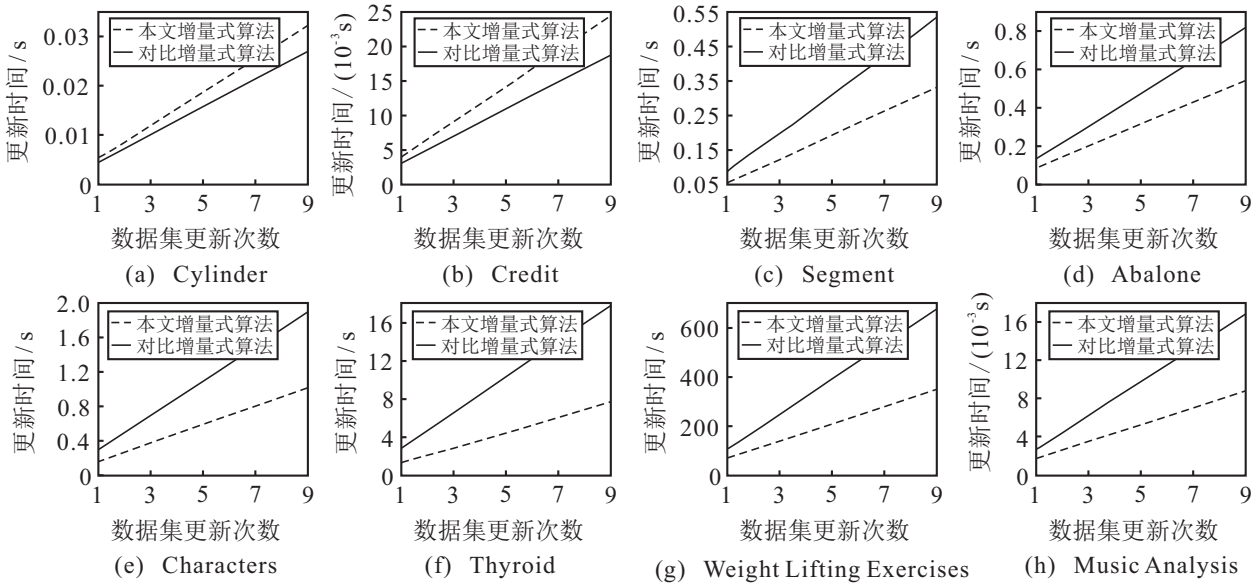

图 3 对象增加时本文增量式算法与对比增量式算法的更新时间比较

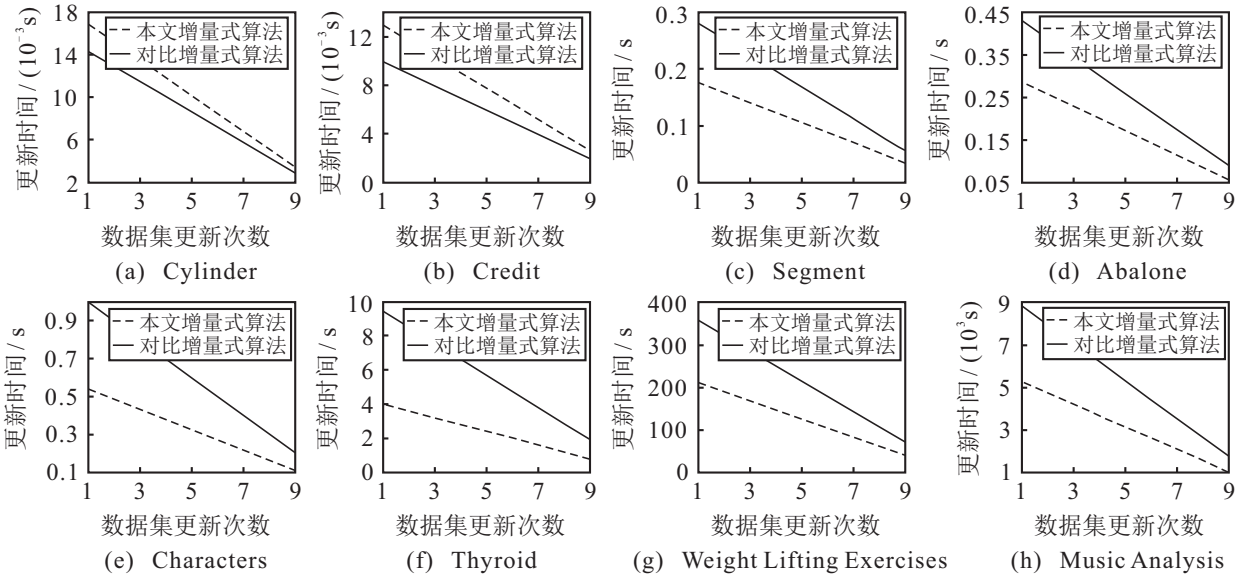

图 4　对象减少时本文增量式算法与对比增量式算法的更新时间比较

集可以发现,数据集Cylinder和Credit的规模都比较小,说明对于小规模的数据集,对比增量式算法的更新效率略高于本文增量式算法,而对于规模较大的数据集,本文增量式算法的更新效率要高于对比增量式算法,并且规模越大,这种差距越明显.

在图4中,对于数据集Cylinder和Credit,同样是对比增量式算法的更新用时略低于本文增量式算法,同样说明了在小规模数据集的情形下,对比算法的效率略高. 而对于其余的数据集,本文算法的更新时间均低于对比增量式算法. 因此,综合图3和图4的结果,可以表明本文的增量式算法相比较于已有的同类型增量式算法具有更高的优越性.

为了测试不同数据量对象变化时两种算法的有效性,下面将设计相关的实验进行验证和分析.

将各个数据集按照对象随机选择50%作为初始数据集,然后从剩余的50%里面依次选择出10%,20%,...,100%分别添加到初始数据集中,最后将两种增量式算法对这一数据环境进行增量式更新计算,其结果如图5所示. 观察图5可以发现,少量对象的增加,对比算法的更新用时略小于本文算法,但是随着数据量的逐渐增大,本文算法的更新用时小于对比算法. 这主要是由算法的时间复杂度决定的,本文算法的时间复杂度要小于对比算法,因此在规模较大的数据量环境下,本文算法的效率会更高.

将各个完整的数据集作为初始数据集,然后在完整的数据集中选择50%,在这50%里面依次选择出10%,20%,...,100%分别在初始数据集中进行移除,最后将两种算法对这一数据环境进行增量式更新计算,其结果如图6所示. 图6结果与图5类似,少量的对象减少时,本文算法的更新用时略高于对比算法,而

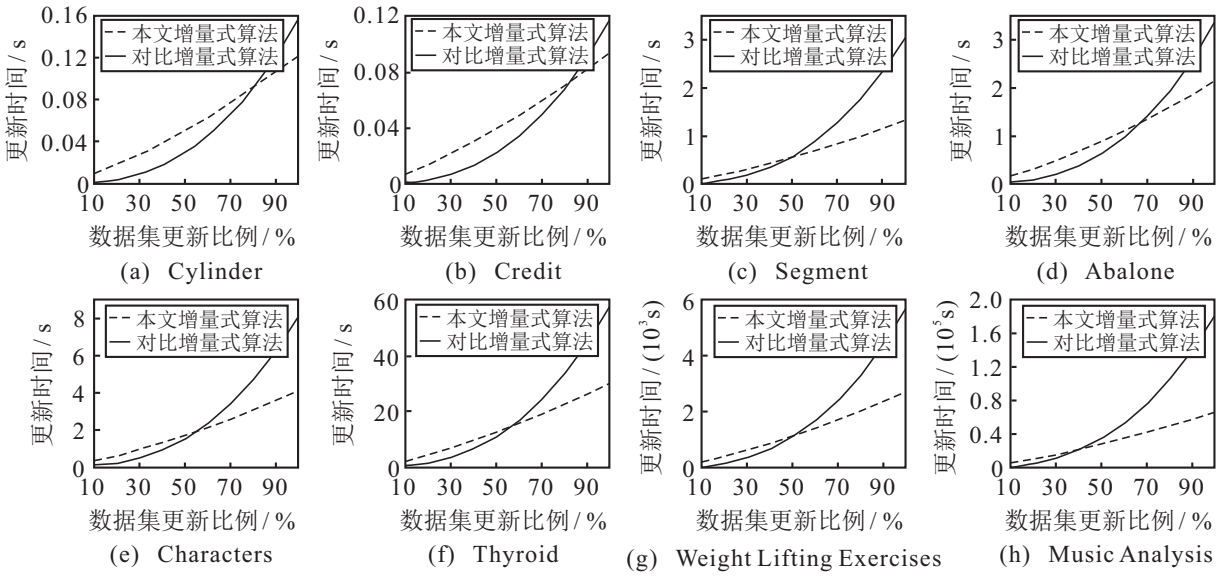

图 5　不同数据量对象增加时本文增量式算法与对比增量式算法的更新时间比较

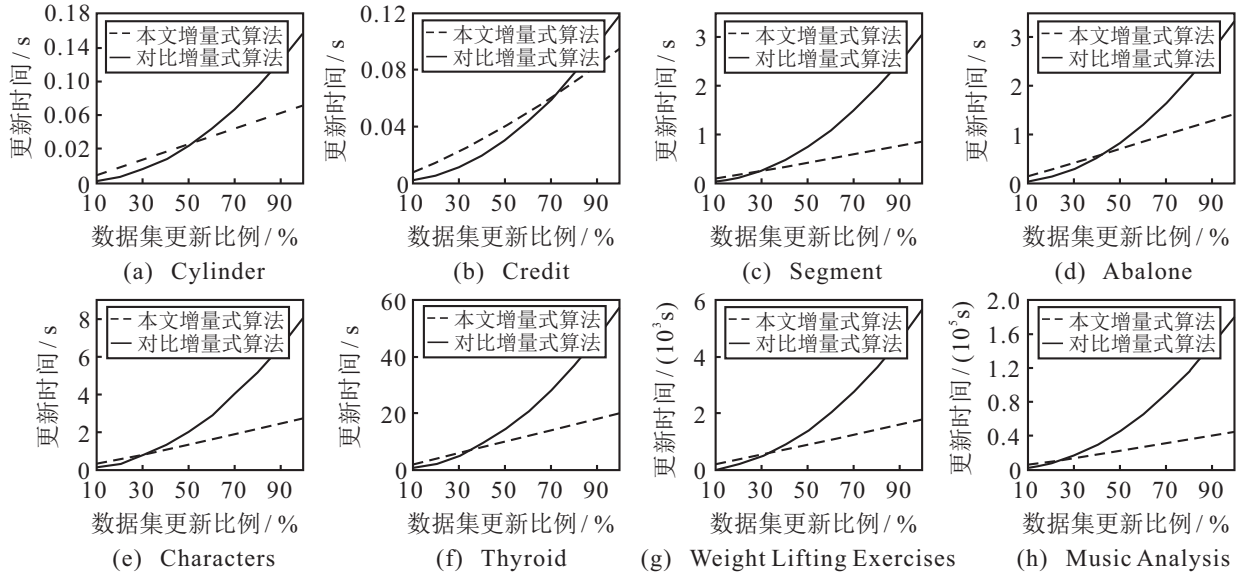

图 6 不同数据量对象减少时本文增量式算法与对比增量式算法的更新时间比较

随着减少对象数据量的增加,本文算法的更新用时小于对比算法,其原因与图5相同,即对于对象减少的情形,本文算法的时间复杂度小于对比算法,因此更新所需的用时会更低.

6 结 论

决策粗糙集模型是当前粗糙集理论的重要研究分支,针对离散型属性和连续型属性并存的混合型信息系统,本文提出了邻域决策粗糙集模型的矩阵表达形式. 然后构造出了当混合型信息系统对象增加和减少时,邻域决策粗糙集矩阵结构的增量式更新算法,实验分析表明了所提出的增量式更新算法具有一定的有效性和优越性. 基于本文矩阵形式的邻域决策粗糙集模型,接下来可以进一步探究相应的属性约简问题.

参考文献(References)

[1] Pawlak Z. Rough sets[J]. International Journal of Computer & Information Sciences, 1982, 11(5): 341-356.

[2] 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.
(Yao S, Xu F, Zhao P, et al. Feature selection algorithm based on neighborhood valued tolerance relation rough set model[J]. Pattern Recognition and Artificial Intelligence, 2017, 30(5): 416-428.)

[3] 张清华, 刘凯旋, 高满. 基于代价敏感的粗糙集近似集与粒度寻优算法[J]. 控制与决策, 2020, 35(9): 2070-2080.
(Zhang Q H, Liu K X, Gao M. Approximation sets of rough sets and granularity optimization algorithm based on cost-sensitive[J]. Control and Decision, 2020, 35(9): 2070-2080.)

[4] Yan Y T, Wu Z B, Du X Q, et al. A three-way decision ensemble method for imbalanced data oversampling[J]. International Journal of Approximate Reasoning, 2019, 107: 1-16.

[5] 杨臻, 邱保志. 混合信息系统的动态变精度粗糙集模型[J]. 控制与决策, 2020, 35(2): 297-308.
(Yang Z, Qiu B Z. Dynamic variable precision rough set model of mixed information system[J]. Control and Decision, 2020, 35(2): 297-308.)

[6] Ni P, Zhao S Y, Wang X Z, et al. Incremental feature selection based on fuzzy rough sets[J]. Information Sciences, 2020, 536: 185-204.

[7] Wei W, Wu X Y, Liang J Y, et al. Discernibility matrix based incremental attribute reduction for dynamic data[J]. Knowledge-Based Systems, 2018, 140: 142-157.

[8] Jing Y G, Li T R, Fujita H, et al. An incremental attribute reduction method for dynamic data mining[J]. Information Sciences, 2018, 465: 202-218.

[9] 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: 320-331.

[10] 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.
(Zhao X L, Yang Y. Incremental attribute reduction algorithm based on neighborhood granulation conditional entropy[J]. Control and Decision, 2019, 34(10): 2061-2072.)

[11] 沈玉峰. 基于矩阵方法的区分度增量式属性约简算法[J]. 计算机应用与软件, 2020, 37(9): 235-245.

(Shen Y F. A discrimination degree incremental attribute reduction based on matrix method[J]. Computer Applications and Software, 2020, 37(9): 235-245.)

[12] 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.
(Sheng K, Wang W, Bian X F, et al. Neighborhood discernibility degree incremental attribute reduction algorithm for mixed data[J]. Acta Electronica Sinica, 2020, 48(4): 682-696.)

[13] 段海玲, 王光琼. 一种高效的复杂信息系统增量式属性约简[J]. 华南理工大学学报: 自然科学版, 2019, 47(6): 18-30.
(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.)

[14] Shu W H, Qian W B, Xie Y H. Incremental feature selection for dynamic hybrid data using neighborhood rough set[J]. Knowledge-Based Systems, 2020, 194: 105516.

[15] Hu J, Li T R, Luo C, et al. Incremental fuzzy probabilistic rough sets over two universes[J]. International Journal of Approximate Reasoning, 2017, 81: 28-48.

[16] Huang Q Q, Li T R, Huang Y Y, et al. Incremental three-way neighborhood approach for dynamic incomplete hybrid data[J]. Information Sciences, 2020, 541: 98-122.

[17] Hu C X, Zhang L, Wang B J, et al. Incremental updating knowledge in neighborhood multigranulation rough sets under dynamic granular structures[J]. Knowledge-Based Systems, 2019, 163: 811-829.

[18] Hu C X, Zhang L. Efficient approaches for maintaining dominance-based multigranulation approximations with incremental granular structures[J]. International Journal of Approximate Reasoning, 2020, 126: 202-227.

[19] Yao Y Y. Three-way decisions with probabilistic rough sets[J]. Information Sciences, 2010, 180(3): 341-353.

[20] Liu D, Liang D C, Wang C C. A novel three-way decision model based on incomplete information system[J]. Knowledge-Based Systems, 2016, 91: 32-45.

[21] Zhao X R, Hu B Q. Three-way decisions with decision-theoretic rough sets in multiset-valued information tables[J]. Information Sciences, 2020, 507: 684-699.

[22] Luo C, Li T R, Yi Z, et al. Matrix approach to decision-theoretic rough sets for evolving data[J]. Knowledge-Based Systems, 2016, 99: 123-134.

[23] Li W W, Huang Z Q, Jia X Y, et al. Neighborhood based decision-theoretic rough set models[J]. International Journal of Approximate Reasoning, 2016, 69: 1-17.

作者简介

苑红星 (1990–), 男, 工程师, 硕士, 从事人工智能的研究, E-mail: hxyuan@ahu.edu.cn;

卓雪雪 (1989–), 女, 讲师, 硕士, 从事大数据智能优化的研究, E-mail: l66x88l@sina.com;

竺德 (1987–), 男, 工程师, 博士生, 从事人工智能的研究, E-mail: zhude@ahu.edu.cn;

刘辉 (1981–), 男, 工程师, 博士生, 从事人工智能的研究, E-mail: liuhui@ahu.edu.cn.

(责任编辑: 孙艺红)