文章编号: 1001-0920(2019)10-2061-12 DOI: 10.13195/j.kzyjc.2018.0138

基于邻域粒化条件熵的增量式属性约简算法

赵小龙[1†], 杨 燕[2]

(1. 安徽工业经济职业技术学院 计算机与艺术学院，合肥 230051；
2. 西南交通大学 信息科学与技术学院，成都 610031)

摘 要: 增量式属性约简是针对动态型数据的一种重要的数据挖掘方法,目前已提出的增量式属性约简算法大多基于离散型数据构建,很少有对数值型数据进行相关的研究. 鉴于此,提出一种数值型信息系统中对象不断增加的增量式属性约简算法. 首先,在数值型信息系统中建立一种分层的邻域粒化计算方法,并基于该方法提出邻域粒化的增量式计算;然后,在邻域粒化增量式计算的基础上给出邻域粒化条件熵的增量式更新方法,并基于该更新机制提出对应的增量式属性约简算法;最后,通过实验分析表明所提出算法对于数值型数据的增量式属性约简具有更高的有效性和优越性.

关键词: 增量式学习；粒计算；属性约简；数值型数据；邻域粒化；条件熵

中图分类号: TP18 **文献标志码:** A

Incremental attribute reduction algorithm based on neighborhood granulation conditional entropy

ZHAO Xiao-long[1†], *YANG Yan*[2]

(1. College of Computer and Art，Anhui Technical College of Industry and Economy，Heifei 230051，China；2. School of Information Science & Technology，Southwest Jiaotong University，Chengdu 610031，China)

Abstract: Incremental attribute reduction is an important data mining method for dynamic data. The incremental attribute reduction algorithms proposed at present are mostly based on discrete data construction, but the related study for numeric data is few. Therefore, an incremental attribute reduction algorithm for object constantly increasing in numeric information system is presented. Firstly, a hierarchical neighborhood computing method is established in numeric information system, and the incremental computing of neighborhood granulation based on this method is proposed. Then, on the basis of neighborhood granulation incremental computing, the incremental updating method of neighborhood granulation conditional entropy is given, and the corresponding incremental attribute reduction algorithm is proposed on account of this updating mechanism. Finally, experimental analysis shows that the proposed algorithm has higher effectiveness and superiority for the incremental attribute reduction of numerical data.

Keywords: incremental learning；granular computing；attribute reduction；numeric data；neighborhood granulation；conditional entropy

0 引 言

属性约简[1]在机器学习和数据挖掘等领域中又称为特征选择,是降低数据复杂度和数据维度的一种重要的数据处理方法,在不降低数据集分类性能的情况下,对数据集中的不相关属性进行鉴别和剔除. 粗糙集理论[2]作为一种重要的粒计算模型[3],目前已成为数据集属性约简的一种常用方法[4-5].

在现实的大数据环境下,数据无时无刻都在产生,因此各类应用系统里的数据集不是静止不变的,而是处于不断动态变化之中. 传统的属性约简算法都是针对静态的数据集而设计[4-5],这些算法在处理动态变化的数据集时需要进行大量的重复计算[6],因此效率较为低下. 为了改善这一局限性,一种新形式的属性约简方法被提出,称为基于增量式学习的属性约简方法,即增量式属性约简[7]. 该方法通过对更新后的信息系统进行增量计算,从而达到对动态数据处理的时效性[8].

数据集对象(样本)的不断增加是数据集动态变化的一种常见形式,基于这类问题,学者们提出了多种增量式属性约简算法. 如Liang等[9]利用条件熵作

收稿日期: 2018-01-28；**修回日期:** 2018-03-21.

基金项目: 安徽省高校自然科学研究重点项目(KJ2016A107, KJ2017A645)；安徽省高校质量工程项目(2016 JXTD019, 2015GXK123.)

†通讯作者. E-mail: zhaoxiaolong1974@163.com.

为启发式函数,在离散型信息系统中提出了对象集增加的增量式属性约简算法. Jing等[10]从多粒化视角提出了高维数据的对象集动态增加的属性约简算法,同时利用粒计算模型中的知识粒度提出了两种增量式属性约简算法[11]. Raza等[12]在粗糙集模型下提出了属性集依赖度的增量更新,并提出了对应的特征选择算法. Chen等[13]运用变精度粗糙集模型在离散型信息系统提出了对应的增量式属性约简. 冯少荣等[14]利用差别矩阵解决对象增加时的增量计算问题. 钱进等[15]也提出了一种新形式的增量式属性约简算法. 在不完备信息系统中,Liu等[16]通过矩阵的方法提出了一种增量式属性约简算法;Shu等[17]提出了不完备信息系统中依赖度的增量式计算,并在此基础上提出相应的增量式属性约简算法.

数值型数据是一种常见的数据形式,然而在目前所提出的增量式属性约简算法中,大多是基于离散型的数据构建,很少有人提出基于数值型数据的增量式属性约简算法,这促使本文对数值型数据的增量式属性约简进行研究. 在粗糙集理论中,邻域粗糙集[18]是处理数值型数据的一种强大的工具,它通过在数值型信息系统中构建邻域关系,基于邻域关系对信息系统的论域进行邻域粒化,最后基于粒化后的信息粒进行粗糙计算[18]. 因此,本文通过邻域粗糙集模型来建立数值型数据的增量式属性约简.

邻域关系将数值型数据诱导出一组邻域粒化,Zhao等[19]在邻域粒化的基础上提出了邻域粒化条件熵模型,并构造出数值型数据的属性约简算法,实验表明了该算法的有效性. 关于邻域粒化的计算方面,Liu等[20]利用排序方法提出了一种高效的邻域粒化计算方法. 本文在此基础上提出一种当信息系统对象集增加时的邻域粒化分层增量式计算;然后在该增量式计算的基础上,提出邻域粒化条件熵的增量式学习机制,并基于该机制提出相应的增量式属性约简算法;最后通过一系列的实验来验证所提出的增量式属性约简算法的有效性和优越性.

1 基本理论

在分类学习任务中,通常格式化的数据集表示为一个信息系统$\mathrm{IS}=(U,\mathrm{AT})$的形式. 其中:$U$称为该信息系统的论域,$U$中每个元素$x$称为信息系统的对象(样本);AT称为信息系统的属性集(特征集);对象$x\in U$在属性$a\in\mathrm{AT}$下的值称为属性值,表示为$a(x)$. 若信息系统中所有属性值均为数值型数据,则该信息系统又称为数值型信息系统$\mathrm{NIS}=(U,\mathrm{AT})$.

在数值型信息系统$\mathrm{NIS}=(U,\mathrm{AT})$中,若属性集AT满足$\mathrm{AT}=C\bigcup D$且$C\bigcap D=\varnothing$,其中$C$为该信息系统的条件属性集,$D$为该信息系统的决策属性集,即该数据集的类属性,其属性值均为离散值. 则这类信息系统称为数值型决策信息系统$\mathrm{NDIS}=(U,C\bigcup D)$.

在粒计算理论[3]中,邻域粗糙集模型[18]是处理数值型信息系统的一种有效工具. 在邻域粗糙集模型中,通过邻域关系对信息系统论域进行邻域粒化,其粒化结果称为论域的一个粒结构,然后基于粒化结果对目标知识进行逼近计算.

定义1 [18] 数值型信息系统$\mathrm{NIS}=(U,\mathrm{AT})$,属性集$B\subseteq\mathrm{AT}$在该信息系统确定的邻域关系$N_B^U$为

$$N_B^U=\{(x,y)\in U\times U|d_B(x,y)\leqslant\delta\}. \tag{1}$$

其中:δ为非负常数,在邻域粗糙集模型中称为邻域半径;$d_B(x,y)$为对象x与y之间的距离,在机器学习和模式识别中,距离度量通常采用闵可夫斯基距离

$$d_B(x,y)=\Big(\sum_{\forall a\in B}|a(x)-a(y)|^p\Big)^{1/p}, \tag{2}$$

p通常取1,2和∞三种形式,$p=2$是熟悉的欧氏距离.

对于$x_1,x_2,x_3\in U$,对象之间的距离满足:

1) $d(x_1,x_2)\geqslant 0$,仅当$x_1=x_2$时,$d(x_1,x_2)=0$;

2) $d(x_1,x_2)=d(x_2,x_1)$;

3) $d(x_1,x_3)\leqslant d(x_1,x_2)+d(x_2,x_3)$.

定义2 [18] 数值型信息系统$\mathrm{NIS}=(U,\mathrm{AT})$,设$U=\{x_1,x_2,\cdots,x_n\}$,属性集$B\subseteq\mathrm{AT}$在该信息系统确定的邻域关系为$N_B^U$,那么$N_B^U$可以将论域$U$诱导出一个邻域粒化$U/N_B^U$,表示为

$$U/N_B^U=\{n_B^U(x_1),n_B^U(x_2),\cdots,n_B^U(x_n)\}. \tag{3}$$

其中:$n_B^U(x)$为对象$x\in U$关于邻域关系N_B^U在论域U下的邻域类,或称为邻域粒,$n_B^U(x)=\{y\in U\mid(x,y)\in N_B^U\}$;$N_B^U$导出的邻域粒化$U/N_B^U$称为论域$U$的一个粒结构.

近年来,Liang等[21]在信息系统中引入了信息论理论,并在信息系统中提出了信息熵的概念. 在此基础上,学者们在信息系统中各类粒计算模型下提出了多种推广信息熵的概念[22]. 其中Zhao等[19]在数值型信息系统的邻域粒化中提出了条件熵模型.

定义3 [19] 数值型信息系统$\mathrm{NIS}=(U,\mathrm{AT})$,设$U=\{x_1,x_2,\cdots,x_n\}$,属性集$B_1,B_2\subseteq\mathrm{AT}$在该信息系统下确定的邻域关系分别为$N_{B_1}^U$和$N_{B_2}^U$,在论域$U$上诱导的邻域粒化分别为$U/N_{B_1}^U$和$U/N_{B_2}^U$. 在论域$U$下,属性集$B_2$关于属性集$B_1$的邻域粒化条件熵为

$$E^U(B_2|B_1)=\frac{1}{n^2}\sum_{i=1}^{n}(|n_{B_1}^U(x_i)|-|n_{B_1}^U(x_i)\bigcap n_{B_2}^U(x_i)|). \tag{4}$$

邻域粒化条件熵满足$0 \leqslant E^U(B_2|B_1) \leqslant 1-1/n$. 此外,对于数值型决策信息系统$\mathrm{NDIS}=(U, C\bigcup D)$,论域$U$在等价关系$R_D^U$下确定的决策粒化为$U/R_D^U=\{[x_1]_D^U,[x_2]_D^U,\cdots,[x_n]_D^U\}$. 决策属性$D$关于属性集$B_1$的邻域粒化条件熵为

$$E^U(D|B_1)=\frac{1}{n^2}\sum_{i=1}^{n}(|n_{B_1}^U(x_i)|-|n_{B_1}^U(x_i)\bigcap[x_i]_D^U|). \tag{5}$$

对于数值型信息系统$\mathrm{NIS}=(U,\mathrm{AT})$,属性集$B_1\subseteq B_2\subseteq \mathrm{AT}$,且$B_1$和$B_2$在该信息系统确定的邻域关系分别为$N_{B_1}^U$和$N_{B_2}^U$,$D$关于属性集$B_1$和$B_2$的邻域粒化条件熵分别为$E^U(D|B_1)$和$E^U(D|B_2)$,满足$E^U(D|B_2)\leqslant E^U(D|B_1)$.

上述是邻域粒化条件熵一个重要的性质,它表明随着属性集的增加,决策属性关于该属性集的邻域粒化条件熵是单调不增的,这是构造属性约简算法的一个必要条件[4, 18-20, 22],它保证邻域粒化条件熵能够收敛,最后属性约简算法才得以终止. 因此,Zhao等[19]提出了基于邻域粒化条件熵的属性约简算法,具体如算法1所示.

算法1　邻域粒化条件熵属性约简(ARNGCE).

输入: 数值型决策信息系统$\mathrm{NDIS}=(U,C\bigcup D)$,邻域半径$\delta$;

输出: 条件属性集C的约简集redc.

Step 1: 初始化$\mathrm{redc}=\varnothing, E^U(D|\varnothing)=1$.

Step 2: 对于$\forall a_i\in C-\mathrm{redc}$,计算属性$a_i$关于redc的属性重要度$s_{\mathrm{redc}}(a_i)$,其中

$$s_{\mathrm{redc}}(a_i)=E^U(D|\mathrm{redc})-E^U(D|\mathrm{redc}\bigcup\{a_i\}).$$

Step 3: 对于Step 2中的所有属性a_i,选出属性重要度最大的属性,并记为a^*.

Step 4: 对于属性a^*,若$s_{\mathrm{redc}}(a^*)>0$,则$\mathrm{redc}=\mathrm{redc}\bigcup\{a^*\}$,进入Step 2;若$s_{\mathrm{redc}}(a^*)=0$,则进入Step 5.

Step 5: 返回约简集redc.

算法1通过邻域粒化条件熵作为启发式函数来搜索属性,并不断进行迭代,直到$E^U(D|C)=E^U(D|\mathrm{redc})$算法终止,此时redc即为条件属性集$C$的约简,且算法1的时间复杂度主要集中在邻域粒化的计算上,每个邻域的计算需要消耗$O(|C||U|)$的时间,因此论域中所有对象进行邻域计算的时间复杂度为$O(|C||U|^2)$,整个算法1的时间复杂度为$O(|C|^2|U|^2)$.

2　邻域粒化条件熵的增量式属性约简

文献[19]通过理论和实验证明,基于邻域粒化条件熵的属性约简具有更高的约简性能. 由于该算法是非增量式的,只能处理静态的信息系统. 为了能够对动态的数据集进行增量式属性约简,针对数值型信息系统对象不断增加的情形,提出一种基于邻域粒化条件熵的增量式属性约简.

增量式属性约简的关键是计算的高效性,当有新数据加入,新的信息系统在属性约简时只需对新进数据进行计算,而不对已经计算过的数据进行重复运算,这样便能满足数据处理的时效性,达到动态数据的处理需求[6-17]. 文献[20]运用排序的方式提出了一种快速邻域计算方法,本节在此基础上,提出一种邻域粒化的分层增量式计算,并提出邻域粒化条件熵的增量式学习机制,最后基于该机制提出相应的增量式属性约简算法.

2.1　邻域粒化的分层增量式计算

Liu等[20]提出的排序方法提高了邻域粒化的计算效率,本节将该方法进一步改进,提出一种高效的邻域粒化方法,并应用于邻域粒化的增量式计算中.

定义4　数值型信息系统$\mathrm{NIS}=(U,\mathrm{AT})$,将信息系统中的所有属性值归一化为非负值,即$\forall x\in U,\forall a\in\mathrm{AT}$满足$a(x)\geqslant 0$. 设属性集$B\subseteq\mathrm{AT}$,邻域半径为$\delta$. 基于属性集$B$在论域$U$上定义一个包含$m$个对象集的分层$L_B^U=\{l_1,l_2,\cdots,l_m\}$,其中

$$l_i=\{x\in U|\lceil d_B(x,x_0)/\delta\rceil=i\},\ 1\leqslant i\leqslant m. \tag{6}$$

其中:$x_0\notin U$是人为构造的一个特定对象,称为原点对象,满足$\forall a\in B, a(x_0)=0$;$d_B(x,x_0)$为对象x与x_0之间的距离度量;m的大小取决于论域中对象与原点对象之间距离的最大值;l_i为论域经过分层后的第i个分层集,l_i内部的对象x与原点对象x_0之间的距离位于区间$(\delta(i-1),\delta i]$,即论域的分层事实上是将整个论域分成多个部分,同一个部分中的对象与原点对象之间的距离位于同一个区间. 同时应当注意,可能存在$l_i\in L_B^U$满足$l_i=\varnothing$.

通过定义4可以看出,在论域U上定义的分层集L^U相当于对论域中所有对象按照与原点对象的距离分成不同的层次. 这样做的直接好处是在进行对象邻域粒计算时,可以大幅度减小计算量.

定理1　数值型信息系统$\mathrm{NIS}=(U,\mathrm{AT})$,设属性集$B\subseteq\mathrm{AT}$,邻域半径为$\delta$. 论域$U$上确定的分层集为$L_B^U=\{l_1,l_2,\cdots,l_m\}$,对象$x$的邻域粒可以计算为

$$n_B^U(x)=\begin{cases}\{y\in l_{i-1}\bigcup l_i\bigcup l_{i+1}|d_B(x,y)\leqslant\delta\},\\ \quad x\in l_i, 2\leqslant i\leqslant m-1;\\ \{y\in l_1\bigcup l_2|d_B(x,y)\leqslant\delta\},\ x\in l_1;\\ \{y\in l_{m-1}\bigcup l_m|d_B(x,y)\leqslant\delta\},\ x\in l_m.\end{cases} \tag{7}$$

证明 1)对于对象x_1,若$x_1 \in l_i(2 \leqslant i \leqslant m-3)$,设原点对象为$x_0$,则根据定义4有$i-1 < d_B(x_1,x_0)/\delta \leqslant i$,即$(i-1)\delta < d_B(x_1,x_0) \leqslant i\delta$. 对于对象$x_2$,若$x_2 \in l_{i+2}(2 \leqslant i \leqslant m-3)$,则满足$(i+1)\delta < d_B(x_2,x_0) \leqslant (i+2)\delta$,因此有$d_B(x_2,x_0) - d_B(x_1,x_0) > \delta$. 根据三角不等式有$d_B(x_2,x_0) - d_B(x_1,x_0) < d_B(x_1,x_2)$,即$d_B(x_1,x_2) > \delta$,所以$\forall x \in l_{i+2}, x \notin n_B^U(x_1)$. 可进一步推出$\forall x \in l_{t\geqslant i+2}$均有$d_B(x_1,x) > \delta$,同理$\forall x \in l_{t\leqslant i-2}$也均有$d_B(x_1,x) > \delta$. 所以$x \in l_i(2 \leqslant i \leqslant m-1)$,邻域粒$n_B^U(x)$可直接计算为$n_B^U(x) = \{y \in l_{i-1} \bigcup l_i \bigcup l_{i+1} | d_B(x,y) \leqslant \delta\}$.

2) 对于对象$x_1 \in l_1$,根据式(1)可以得出$\forall x \in l_{t\geqslant 3}$均有$d_B(x_1,x) > \delta$. 所以若$x \in l_1$,则对象$x$的邻域粒可计算为$n_B^U(x) = \{y \in l_1 \bigcup l_2 | d_B(x,y) \leqslant \delta\}$.

3) 对于对象$x_1 \in l_m$,根据式(1)可以得出$\forall x \in l_{t\leqslant m-2}$均有$d_B(x_1,x) > \delta$. 所以若$x \in l_m$,则对象$x$的邻域粒可计算为$n_B^U(x) = \{y \in l_{m-1} \bigcup l_m | d_B(x,y) \leqslant \delta\}$. □

通过定理1可以看出,将论域进行分层后,某一层内对象x邻域粒的计算只需分析前一层、本层和后一层的对象集即可,其余层内的对象与对象x的距离肯定是超过邻域半径δ的,因此无需像定义2那样,将整个论域遍历一遍,这样可以大幅度减小计算量,提高计算效率. 图1为两个属性下按照定理1方法进行邻域粒计算的示意图.

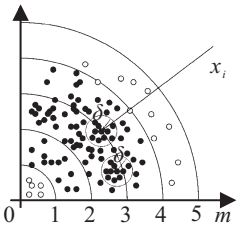

图1 对象x_i的邻域粒计算

例1 表1为数值型信息系统NIS $= (U, \mathrm{AT})$,其中论域U包含5个对象,属性集$\mathrm{AT} = \{a_1, a_2, a_3\}$.

表1 信息系统

U	a_1	a_2	a_3
x_1	0.12	0.24	0.23
x_2	0.16	0.37	0.41
x_3	0.18	0.22	0.32
x_4	0.14	0.35	0.44
x_5	0.10	0.20	0.17

令邻域半径$\delta = 0.1$,距离函数选取为欧氏距离. 根据定义4可以得到表1信息系统的分层集为

$$L_{\mathrm{AT}}^U = \{l_1, l_2, l_3, l_4, l_5, l_6\}.$$

其中:$l_1 = l_2 = \varnothing, l_3 = \{x_5\}, l_4 = \{x_1\}, l_5 = \{x_3\}, l_6 = \{x_2, x_4\}$.

按照定理1的方法,对于对象$x_1 \in l_4$,计算x_1的邻域只需要考察l_3、l_4和l_5中的对象,计算可得$n_{\mathrm{AT}}^U(x_1) = \{x_1, x_5\}$. 对于对象$x_2 \in l_6$,计算$x_2$的邻域只需要考察$l_5$和$l_6$中的对象,计算可得$n_{\mathrm{AT}}^U(x_2) = \{x_2, x_4\}$. 同理可计算出其他对象的邻域为$n_{\mathrm{AT}}^U(x_3) = \{x_3\}$, $n_{\mathrm{AT}}^U(x_4) = \{x_2, x_4\}, n_{\mathrm{AT}}^U(x_5) = \{x_1, x_5\}$.

根据定义2的方法计算每个对象的邻域粒. 对于对象x_1,有

$$d_{\mathrm{AT}}(x_1,x_1) = 0,\ d_{\mathrm{AT}}(x_1,x_2) = 0.225,$$
$$d_{\mathrm{AT}}(x_1,x_3) = 0.115,\ d_{\mathrm{AT}}(x_1,x_4) = 0.237,$$
$$d_{\mathrm{AT}}(x_1,x_5) = 0.075.$$

对象x_1的邻域粒为$n_{\mathrm{AT}}^U(x_1) = \{x_1, x_5\}$. 同样可得

$$n_{\mathrm{AT}}^U(x_2) = \{x_2, x_4\},\ n_{\mathrm{AT}}^U(x_3) = \{x_3\},$$
$$n_{\mathrm{AT}}^U(x_4) = \{x_2, x_4\},\ n_{\mathrm{AT}}^U(x_5) = \{x_1, x_5\}.$$

与定理1的计算结果相同,从实例角度证明了定理1的正确性.

根据定理1中基于分层方法的快速邻域粒化计算,在此基础上提出数值型信息系统对象增加时的邻域粒化增量式更新方法.

定义5 对于数值型信息系统NIS $= (U, \mathrm{AT})$,设$U = \{x_1, x_2, \cdots, x_n\}$,属性集$B \subseteq \mathrm{AT}$,邻域半径为$\delta$,令属性集$B$在论域$U$上确定的分层集为$L_B^U = \{l_1, l_2, \cdots, l_m\}$. 当一个新的对象集$\Delta U$加入信息系统时,新的信息系统表示为NIS$' = (U', \mathrm{AT})$,其中$U' = U \bigcup \Delta U$. 那么论域$U'$上新的分层集为

$$L_B^{U'} = \{l_1', l_2', \cdots, l_m', l_{m+1}, \cdots, l_{m+c}\}. \tag{8}$$

其中

$$l_{m+i} = \{x \in \Delta U | \lceil d_B(x,x_0)/\delta \rceil = m+i\},\ 1 \leqslant i \leqslant c;$$
$$l_i' = l_i \bigcup \{x \in \Delta U | \lceil d_B(x,x_0)/\delta \rceil = i\},\ 1 \leqslant i \leqslant m.$$

x_0表示原点对象.

定义5表明,当信息系统的对象发生增加时,不需要对更新后的整个信息系统重新建立分层集,只需在原来分层集的基础上进行更新和扩展,这样便减少了大量的重复计算.

定理2 对于数值型信息系统NIS $= (U, \mathrm{AT})$,设$U = \{x_1, x_2, \cdots, x_n\}$,属性集$B \subseteq \mathrm{AT}$,邻域半径为$\delta$. 属性集$B$在该信息系统确定的邻域关系为$N_B^U$,并且基于分层计算方法得到的邻域粒化为$U/N_B^U = \{n_B^U(x_1), n_B^U(x_2), \cdots, n_B^U(x_n)\}$. 当新的对象集$\Delta U = \{x_{n+1}, x_{n+2}, \cdots, x_{n+u}\}$加入信息系统时,新的信息系统为NIS$' = (U', \mathrm{AT})$,其中$U' = U \bigcup \Delta U$. 设$N_B^{U'}$为在新的信息系统确定的邻域关系,则新的邻域

粒化为

$$U'/N_B^{U'} = \{n_B^{U'}(x_1), n_B^{U'}(x_2), \cdots, n_B^{U'}(x_n), n_B^{U'}(x_{n+1}), n_B^{U'}(x_{n+2}), \cdots, n_B^{U'}(x_{n+u})\}, \tag{9}$$

其中$n_B^{U'}(x_{n+1}), n_B^{U'}(x_{n+2}), \cdots, n_B^{U'}(x_{n+u})$按照定理1方法进行计算. 同时$\forall x \in U$更新为

$$n_B^{U'}(x) = n_B^U(x) \bigcup \{y | x \in n_B^{U'}(y), \forall y \in \Delta U\}. \tag{10}$$

证明　当信息系统$\mathrm{NIS} = (U, \mathrm{AT})$有新的对象集$\Delta U$加入时,原先论域$U$上的邻域粒化$U/N_B^U$将会发生改变,即新的对象集$\Delta U$的加入会使得论域$U$中对象的邻域发生增加. 对于对象$y \in \Delta U$,如果对象$x \in U$且$x \in n_B^{U'}(y)$,则表明$d_B(x, y) \leqslant \delta$,即$y \in n_B^{U'}(x)$,又$y \notin n_B^U(x)$,所以$n_B^{U'}(x) = n_B^U(x) \bigcup \{y\}$. 因此有$n_B^{U'}(x) = n_B^U(x) \bigcup \{y | x \in n_B^{U'}(y), \forall y \in \Delta U\}$. □

当新的对象集加入后,非增量式的邻域粒化计算方法是重新计算论域中每个对象的邻域粒,而定理2所示的增量式计算方法中,对新加入的对象按照定理1的方法快速计算出邻域粒,然后对于原来论域的邻域粒化,只需要更新部分对象的邻域粒,最终得到整个新论域的邻域粒化. 事实上,新对象的邻域粒计算与原来论域邻域粒化的更新可以同步进行,因此定理2中的增量式邻域粒化计算具有很高的效率.

例2　在例1的基础上,对表1的信息系统加入一个对象集$\{x_6, x_7\}$,新的信息系统$\mathrm{NIS}' = (U', \mathrm{AT})$如表2所示.

表 2　新信息系统

U	a_1	a_2	a_3
x_1	0.12	0.24	0.23
x_2	0.16	0.37	0.41
x_3	0.18	0.22	0.32
x_4	0.14	0.35	0.44
x_5	0.10	0.20	0.17
x_6	0.20	0.38	0.46
x_7	0.17	0.24	0.30

由于对象集$\{x_6, x_7\}$的加入,需要对原来信息系统的邻域粒化结果进行更新,设x_0为原点距离,$\lceil d_{\mathrm{AT}}(x_6, x_0)/\delta \rceil = 7, \lceil d_{\mathrm{AT}}(x_7, x_0)/\delta \rceil = 5$,新论域$U'$的分层集扩展至7层,即

$$L_{\mathrm{AT}}^{U'} = \{l_1', l_2', l_3', l_4', l_5', l_6', l_7\}.$$

按照定义5的更新方法可以得到

$$l_1' = l_1,\ l_2' = l_2,\ l_3' = l_3,\ l_4' = l_4,$$
$$l_5' = l_5 \bigcup \{x_7\},\ l_6' = l_6,\ l_7 = \{x_6\}.$$

同时计算出对象x_6和x_7的邻域粒分别为$n_{\mathrm{AT}}^{U'}(x_6) = \{x_2, x_4, x_6\}, n_{\mathrm{AT}}^{U'}(x_7) = \{x_1, x_3, x_7\}$. 根据定理2有

$$n_{\mathrm{AT}}^{U'}(x_2) = n_{\mathrm{AT}}^U(x_2) \bigcup \{x_6\},$$
$$n_{\mathrm{AT}}^{U'}(x_4) = n_{\mathrm{AT}}^U(x_4) \bigcup \{x_6\},$$
$$n_{\mathrm{AT}}^{U'}(x_1) = n_{\mathrm{AT}}^U(x_1) \bigcup \{x_7\},$$
$$n_{\mathrm{AT}}^{U'}(x_3) = n_{\mathrm{AT}}^U(x_3) \bigcup \{x_7\}.$$

对象x_5满足$n_{\mathrm{AT}}^{U'}(x_5) = n_{\mathrm{AT}}^U(x_5)$.

2.2　邻域粒化条件熵的增量式更新

第2.1节提出了一种基于分层集模型的邻域粒化快速计算方法,并运用该分层集模型建立了当信息系统对象集增加时的邻域粒化增量式更新. 本节将在邻域粒化增量式更新方法的基础上,提出邻域粒化条件熵的增量式更新,为后面增量式属性约简算法的提出提供铺垫.

为了研究过程的清晰和直观,采用其他学者的研究思路[9, 11, 16-17],首先分析只有一个对象加入信息系统后,邻域粒化条件熵的增量式更新,然后在此基础上进一步探索多个对象加入后的增量式更新.

定理3　对于数值型决策信息系统$\mathrm{NDIS} = (U, C \bigcup D)$,设$U = \{x_1, x_2, \cdots, x_n\}$,属性集$B \subseteq C$,邻域半径为$\delta$. 属性集$B$在论域$U$上诱导出的邻域粒化为$U/N_B^U = \{n_B^U(x_1), n_B^U(x_2), \cdots, n_B^U(x_n)\}$,决策属性$D$关于$B$在论域$U$下的邻域粒化条件熵为$E^U(D|B)$. 当新的对象$x_{n+1}$加入信息系统后,新的信息系统为$\mathrm{NIS}' = (U' = U \bigcup \{x_{n+1}\}, \mathrm{AT})$,$B$在论域$U'$上诱导出的邻域粒化为$U'/N_B^{U'} = \{n_B^{U'}(x_1), n_B^{U'}(x_2), \cdots, n_B^{U'}(x_{n+1})\}$,决策属性$D$关于$B$在论域$U'$下的邻域粒化条件熵为$E^{U'}(D|B)$. 则有

$$E^{U'}(D|B) = \frac{n^2}{(n+1)^2} E^U(D|B) + \frac{2|n_B^{U'}(x_{n+1}) - [x_{n+1}]_D^{U'}|}{(n+1)^2}. \tag{11}$$

证明　由定义3,有

$$E^U(D|B) = \frac{1}{n^2} \sum_{i=1}^{n} (|n_B^U(x_i)| - |n_B^U(x_i) \bigcap [x_i]_D^U|) = \frac{1}{n^2} \Big(\sum_{i=1}^{n} |n_B^U(x_i)| - \sum_{i=1}^{n} |n_B^U(x_i) \bigcap [x_i]_D^U| \Big).$$

当对象x_{n+1}加入后,$U' = U \bigcup \{x_{n+1}\}$. 令$\Phi = n_B^{U'}(x_{n+1}), \Psi = U' - n_B^{U'}(x_{n+1})$,根据定理2可知$\Psi$表示增加$x_{n+1}$后$U$中邻域粒不发生变化的对象集,因此可以推出$\forall x \in \Psi, n_B^{U'}(x) = n_B^U(x)$,同时有$n_B^{U'}(x) \bigcap [x]_D^{U'} = n_B^U(x) \bigcap [x]_D^U$. 那么

$$E^{U'}(D|B) = \frac{1}{(n+1)^2} \Big(\sum_{i=1}^{n+1} |n_B^{U'}(x_i)| - \sum_{i=1}^{n+1} |n_B^{U'}(x_i) \bigcap [x_i]_D^{U'}| \Big) =$$

$$\frac{1}{(n+1)^2}\Big(\sum_{x\in\Psi}|n_B^U(x)|+\sum_{x\in\Phi}|n_B^{U'}(x)|-$$
$$\sum_{x\in\Psi}|n_B^U(x)\bigcap[x]_D^U|-\sum_{x\in\Phi}|n_B^{U'}(x)\bigcap[x]_D^{U'}|\Big).$$

由定理2,对于$y\in U$,若$y\in n_B^{U'}(x_{n+1})$,则$n_B^{U'}(y)=n_B^U(y)\bigcup\{x_{n+1}\}$,所以对于$\forall y\in\Phi-\{x_{n+1}\}$,均有$|n_B^{U'}(y)|=|n_B^U(y)|+1$,并且当对象$y$与对象$x_{n+1}$有相同的决策值时,$|n_B^{U'}(y)\bigcap[y]_D^{U'}|=|n_B^U(y)\bigcap[y]_D^U|+1$,显然在$\Phi-\{x_{n+1}\}$中有$|n_B^{U'}(x_{n+1})\bigcap[x_{n+1}]_D^{U'}|-1$个对象满足该条件,所以有

$$E^{U'}(D|B)=$$
$$\frac{1}{(n+1)^2}\Big(\sum_{x\in\Psi}|n_B^U(x)|+\sum_{x\in\Phi-\{x_{n+1}\}}(|n_B^U(x)|+1)+$$
$$|n_B^{U'}(x_{n+1})|-\sum_{x\in\Psi}|n_B^U(x)\bigcap[x]_D^U|-$$
$$\sum_{x\in\Phi-\{x_{n+1}\}}|n_B^{U'}(x)\bigcap[x]_D^{U'}|-$$
$$|n_B^{U'}(x_{n+1})\bigcap[x_{n+1}]_D^{U'}|\Big)=$$
$$\frac{1}{(n+1)^2}\Big(\sum_{i=1}^{n}|n_B^U(x_i)|+|\Phi|-1+|n_B^{U'}(x_{n+1})|-$$
$$\sum_{i=1}^{n}|n_B^U(x_i)\bigcap[x_i]_D^U|-(|n_B^{U'}(x_{n+1})\bigcap[x_{n+1}]_D^{U'}|-1)-$$
$$|n_B^{U'}(x_{n+1})\bigcap[x_{n+1}]_D^{U'}|\Big)=$$
$$\frac{1}{(n+1)^2}\Big(\sum_{i=1}^{n}|n_B^U(x_i)|-\sum_{i=1}^{n}|n_B^U(x_i)\bigcap[x_i]_D^U|+$$
$$|n_B^{U'}(x_{n+1})|+|n_B^{U'}(x_{n+1})|-|n_B^{U'}(x_{n+1})\bigcap[x_{n+1}]_D^{U'}|-$$
$$|n_B^{U'}(x_{n+1})\bigcap[x_{n+1}]_D^{U'}|\Big)=$$
$$\frac{1}{(n+1)^2}(n^2E^U(D|B)+2|n_B^{U'}(x_{n+1})|-$$
$$2|n_B^{U'}(x_{n+1})\bigcap[x_{n+1}]_D^{U'}|).$$

根据集合的运算关系满足

$$|n_B^{U'}(x_{n+1})|-|n_B^{U'}(x_{n+1})\bigcap[x_{n+1}]_D^{U'}|=$$
$$|n_B^{U'}(x_{n+1})-[x_{n+1}]_D^{U'}|.$$

因此

$$E^{U'}(D|B)=$$
$$\frac{1}{(n+1)^2}(n^2E^U(D|B)+2|n_B^{U'}(x_{n+1})-[x_{n+1}]_D^{U'}|).$$

简单整理后即可得到定理 3 中的结果. □

定理3给出了当信息系统增加一个对象后,原邻域粒化条件熵与新邻域粒化条件熵之间的增量更新关系,并且在原条件熵的基础上,只需要计算增加进来的对象邻域粒和决策类,便可得到新的条件熵结果,这样避免了对原论域中对象邻域粒的重复计算,大大提高了动态环境下的处理效率.

定理3给出的是增加一个对象后条件熵的递推关系, 当信息系统增加多个对象时,可通过该递推关系推导出增加多个对象时的条件熵增量更新关系.

定理4 对于数值型决策信息系统$\mathrm{NDIS}=U,C\bigcup D$,设$U=\{x_1,x_2,\cdots,x_n\}$,属性集$B\subseteq C$,邻域半径为$\delta$. 决策属性$D$关于$B$在论域$U$下的邻域粒化条件熵为$E^U(D|B)$. 当包含$u$个对象的新对象集$\Delta U=\{x_{n+1},x_{n+2},\cdots,x_{n+u}\}$加入信息系统后,新的信息系统为$\mathrm{NIS}'=(U'=U\bigcup\Delta U,\mathrm{AT})$,决策属性$D$关于$B$在论域$U'$下的邻域粒化条件熵为$E^{U'}(D|B)$. 则有

$$E^{U'}(D|B)=$$
$$\frac{n^2}{(n+u)^2}E^U(D|B)+\frac{2}{(n+u)^2}(|n_B^{U_1}(x_{n+1})-[x_{n+1}]_D^{U_1}|+$$
$$|n_B^{U_2}(x_{n+2})-[x_{n+2}]_D^{U_2}|+\cdots+$$
$$|n_B^{U_u}(x_{n+u})-[x_{n+u}]_D^{U_u}|).\tag{12}$$

其中:$U_i=U\bigcup\{x_{n+1},x_{n+2},\cdots,x_{n+i}\}$,即$U'=U_u$;$n_B^{U'}(x_{n+i})$为对象$x_{n+i}$在论域$U_i$上的邻域粒;$[x_{n+i}]_D^{U_i}$为对象$x_{n+i}$在论域$U_i$上的决策类.

证明 为了简便,记$E^{U_i}(D|B)=E^{U_i}$. 由定理3有

$$E^{U_1}=\frac{n^2}{(n+1)^2}E^U+\frac{2|n_B^{U_1}(x_{n+1})-[x_{n+1}]_D^{U_1}|}{(n+1)^2}.$$

则有

$$E^{U_2}=\frac{(n+1)^2}{(n+2)^2}E^{U_1}+\frac{2|n_B^{U_2}(x_{n+2})-[x_{n+2}]_D^{U_2}|}{(n+2)^2}=$$
$$\frac{(n+1)^2}{(n+2)^2}\Big[\frac{n^2}{(n+1)^2}E^U+\frac{2|n_B^{U_1}(x_{n+1})-[x_{n+1}]_D^{U_1}|}{(n+1)^2}\Big]+$$
$$\frac{2|n_B^{U_2}(x_{n+2})-[x_{n+2}]_D^{U_2}|}{(n+2)^2}=$$
$$\frac{n^2}{(n+2)^2}E^U+\frac{2|n_B^{U_1}(x_{n+1})-[x_{n+1}]_D^{U_1}|}{(n+2)^2}+$$
$$\frac{2|n_B^{U_2}(x_{n+2})-[x_{n+2}]_D^{U_2}|}{(n+2)^2},$$
$$E^{U_3}=\frac{(n+2)^2}{(n+3)^2}E^{U_2}+\frac{2|n_B^{U_3}(x_{n+3})-[x_{n+3}]_D^{U_3}|}{(n+3)^2}=$$
$$\frac{(n+2)^2}{(n+3)^2}\Big[\frac{n^2}{(n+2)^2}E^U+\frac{2|n_B^{U_1}(x_{n+1})-[x_{n+1}]_D^{U_1}|}{(n+2)^2}+$$
$$\frac{2|n_B^{U_2}(x_{n+2})-[x_{n+2}]_D^{U_2}|}{(n+2)^2}\Big]+$$
$$\frac{2|n_B^{U_3}(x_{n+3})-[x_{n+3}]_D^{U_3}|}{(n+3)^2}=$$
$$\frac{n^2}{(n+3)^2}E^U+\frac{2|n_B^{U_1}(x_{n+1})-[x_{n+1}]_D^{U_1}|}{(n+3)^2}+$$

$$\frac{2|n_B^{U_2}(x_{n+2})-[x_{n+2}]_D^{U_2}|}{(n+3)^2}+\frac{2|n_B^{U_3}(x_{n+3})-[x_{n+3}]_D^{U_3}|}{(n+3)^2}.$$

依此计算有

$$E^{U_i}=\frac{n^2}{(n+i)^2}E^U+\frac{2}{(n+i)^2}(|n_B^{U_1}(x_{n+1})-[x_{n+1}]_D^{U_1}|+\cdots+|n_B^{U_i}(x_{n+i})-[x_{n+i}]_D^{U_i}|),$$

即

$$E^{U_u}=E^{U'}=\frac{n^2}{(n+u)^2}E^U+\frac{2}{(n+u)^2}(|n_B^{U_1}(x_{n+1})-[x_{n+1}]_D^{U_1}|+\cdots+|n_B^{U_u}(x_{n+u})-[x_{n+u}]_D^{U_u}|).$$ □

定理4表明,当信息系统加入多个对象时,只需要依次对每个加入的对象计算出邻域粒和决策类,并且对象x_{n+1}在$U\bigcup\{x_{n+1}\}$下进行计算,对象x_{n+2}在$U\bigcup\{x_{n+1},x_{n+2}\}$下进行计算,对象$x_{n+i}$在$U\bigcup\{x_{n+1},x_{n+2},\cdots,x_{n+i}\}$下进行计算,因此在计算过程中,可以逐步计算邻域粒,即对于多个对象,可以依次加入信息系统. 加入一个对象后便立即在当时的信息系统中计算邻域粒和决策类,这样当所有对象添加完毕,整个邻域粒化条件熵便可以直接得出.

2.3　增量式属性约简算法

根据第2.2节的增量式邻域粒化条件熵,基于文献[19]所提出的邻域粒化条件熵的数值型信息系统属性约简算法,可以针对论域不断增大的动态数值型信息系统提出邻域粒化条件熵的增量式属性约简算法. 在提出算法前,首先探究数值型信息系统论域增大后,其邻域粒化条件熵的变化情况. 同样地,首先分析只有一个对象加入信息系统后邻域粒化条件熵的变化.

设数值型决策信息系统$\mathrm{NDIS}=(U,C\bigcup D),U=\{x_1,x_2,\cdots,x_n\}$,属性集$B\subseteq C$,邻域半径为$\delta$. 决策属性$D$关于$B$在论域$U$下的邻域粒化条件熵简记为$E^U$. 当新的对象$x_{n+1}$加入信息系统后,新的信息系统为$\mathrm{NDIS}'=(U'=U\bigcup\{x_{n+1}\},C\bigcup D)$,决策属性$D$关于$B$在论域$U'$下的邻域粒化条件熵简记为$E^{U'}$. 由定理3,有

$$E^{U'}=\frac{n^2}{(n+1)^2}E^U+\frac{2|n_B^{U'}(x_{n+1})-[x_{n+1}]_D^{U'}|}{(n+1)^2}.$$

对两个邻域粒化条件熵相减,即

$$E^{U'}-E^U=\frac{n^2}{(n+1)^2}E^U+\frac{2|n_B^{U'}(x_{n+1})-[x_{n+1}]_D^{U'}|}{(n+1)^2}-E^U=\frac{2|n_B^{U'}(x_{n+1})-[x_{n+1}]_D^{U'}|}{(n+1)^2}-\frac{2n+1}{(n+1)^2}E^U.$$

由定义3,有$0\leqslant E^U\leqslant 1-1/n$,令$k=|n_B^{U'}(x_{n+1})-[x_{n+1}]_D^{U'}|$,经过整理可以得到

$$2(k-n)+\frac{1}{n}+1\leqslant E^{U'}-E^U\leqslant 2k.$$

由于$n_B^{U'}(x_{n+1})-[x_{n+1}]_D^{U'}\subseteq U$,即$0\leqslant k\leqslant n$,有

$$-2n+1/n+1\leqslant 2(k-n)+1/n+1\leqslant 1/n+1,$$

$2(k-n)+1/n+1$的值可能为正、或负、或0,$E^{U'}-E^U$的值也如此. 因此可以得出,当信息系统增加一个对象后,其邻域粒化条件熵的变化是不可确定的,可能会增大也可能会减小,或者可能不变. 该结论可直接推广到当信息系统有多个对象加入的情形.

在本文所提出的增量式属性约简中,初始信息系统的属性约简采用算法1,但当信息系统有新的对象集加入时,信息系统发生更新,传统的非增量式属性约简在进行属性约简时,对新的信息系统直接处理,即重新运用算法1对新的信息系统进行属性约简,这样随着数据的规模逐步增大,属性约简的耗时会越来越大. 本文所提出的属性约简算法采用增量式学习的方法,在原来信息系统属性约简的基础上,增量式地计算新的信息系统的约简,能够大幅度提高计算效率. 具体的增量式属性约简算法如算法2所示.

算法2　邻域粒化条件熵的增量式属性约简(IARNGCE).

输入: 1) 初始时数值型决策信息系统$\mathrm{NDIS}=(U,C\bigcup D),|U|=n$,邻域半径$\delta$; 2) NDIS中条件属性集$C$的约简集redc; 3) 邻域粒化条件熵$E^U(D|C)$,$E^U(D|\mathrm{redc})$; 4) 增加的对象集$\Delta U=\{x_{n+1},x_{n+2},\cdots,x_{n+u}\}$,并令新的数值型决策信息系统为$\mathrm{NDIS}'=(U',C\bigcup D)$,其中$U'=U\bigcup\Delta U$.

输出: NDIS'中条件属性集C的约简集redc'.

Step 1: 初始化$E^{U'}(D|C)=1,E^{U'}(D|\mathrm{redc})=1$.

Step 2: 根据$E^U(D|C)$和$E^U(D|\mathrm{redc})$,按照定理4增量式计算$E^{U'}(D|C)$和$E^{U'}(D|\mathrm{redc})$.

Step 3: 判断$E^{U'}(D|C)$与$E^{U'}(D|\mathrm{redc})$之间的大小关系,如果$E^{U'}(D|C)<E^{U'}(D|\mathrm{redc})$,则转至Step 4,如果$E^{U'}(D|C)=E^{U'}(D|\mathrm{redc})$,则转至Step 7.

Step 4: 对于$\forall a\in C-\mathrm{redc}$,计算每个属性对应的邻域粒化条件熵$E^{U'}(D|\mathrm{redc}\bigcup\{a\})$.

Step 5: 选择满足下列条件的属性,记为a^*:

$$\max_{\forall a\in C-\mathrm{redc}}[E^{U'}(D|\mathrm{redc})-E^{U'}(D|\mathrm{redc}\bigcup\{a\})].$$

Step 6: 若$E^{U'}(D|\mathrm{redc})-E^{U'}(D|\mathrm{redc}\bigcup\{a^*\})>0$,则$\mathrm{redc}=\mathrm{redc}\bigcup\{a^*\}$,并进入Step 4,否则进入Step 7.

Step 7: $\mathrm{redc}'=\mathrm{redc}$,返回约简集$\mathrm{redc}'$.

在算法2的Step 2中,初始时$E^U(D|C)$与

$E^U(D|\text{redc})$的值是相等的,按照定理4分别增量式更新至$E^{U'}(D|C)$和$E^{U'}(D|\text{redc})$. 根据之前的分析,信息系统论域增大,虽然邻域粒化条件熵的变化是不确定的,但必定满足$E^{U'}(D|C) \leqslant E^{U'}(D|\text{redc})$,因此在Step 3中,需要具体判断$E^{U'}(D|C)$与$E^{U'}(D|\text{redc})$之间的大小关系. 如果它们相等,则达到算法的终止条件,返回约简集,如果$E^{U'}(D|C) \leqslant E^{U'}(D|\text{redc})$,则需要在剩余的属性中继续搜索属性,直至满足终止条件.

在算法2所示的增量式属性约简中,论域的分层集是整个邻域粒化条件熵计算的基础,根据定义4,对于属性集$B \subseteq C$在论域U建立分层集的时间复杂度为$O(|B||U|)$,加入新的对象集ΔU后,更新分层集所需要的时间复杂度为$O(|B||\Delta U|)$. 设分层集的每一层平均包含k个对象,那么基于分层集计算单个对象邻域所需的时间复杂度为$O(3k)$. Step 2中论域U的分层集已在初始信息系统属性约简时建立,时间复杂度为$O(|C||\Delta U| + |\text{redc}||\Delta U| + 6k|\Delta U|)$. Step 3的时间复杂度为$O(1)$. 假设从初始约简集redc至新的约简集$\text{redc}'$增加了$c$个属性,那么Step 4 ~ Step 6的时间复杂度为$O(|C - \text{redc}|(2|\text{redc}||U'| + 6k'|U'|)) + O((|C - \text{redc}| - 1)(2(|\text{redc}| + 1)|U'| + 6k'|U'|)) + \cdots + O((|C - \text{redc}| - c)(2(|\text{redc}| + c)|U'| + 6k'|U'|))$,其中$k'$表示$U'$分层集中每层的平均对象数. 通常$c$的值较小,因此通过整理和计算可得出整个算法2的时间复杂度为$O(|C - \text{redc}'|(|\text{redc}'| + k')|U'|)$. 相比较于算法1的非增量式属性约简,所提出的算法在时间复杂度方面具有更高的效率.

3 实验分析

本节将进行一系列实验以验证所提出算法的增量式属性约简性能. 实验主要分为两部分: 第1部分是将所提出的邻域粒化条件熵增量式属性约简算法与邻域粒化条件熵非增量式属性约简算法对同一组数据集进行动态数据集的属性约简,从而验证所提出算法的有效性; 第2部分是将所提出的算法与其他同类型增量式属性约简算法对同一组数据集进行属性约简比较,从而验证所提出算法的优越性.

准备工作如下. 从UCI机器学习标准数据集库中选取7个数值型数据集,如表3所示. 这7个数据集都是静态的数据集,但是本文所提出的是动态数据集的属性约简方法. 为了构造出数据集对象增加的动态性,将整个数据集的对象集大致平均分成10个等份,随机选取其中一份作为初始数据集,然后依次从剩余的每份中随机选取一份对象集加入数据集中,这样便模拟出数据集中对象集9次动态增加的情形. 实验中的所有算法采用JDK1.8进行编程实现,算法代码运行在Intel Core(TM) i7 3770 3.4 GHz CPU, DDR3 1600 MHz 8 G内存的Windows 10个人主机上.

表 3 数据集信息

编号	数据集	类	对象	属性
1	Wdbc	2	569	31
2	Hill_Valley	2	606	101
3	Biodeg	2	1 055	41
4	Mess	2	1 151	19
5	Yeast	13	1 484	9
6	Wall	4	5 456	5
7	Magic	2	19 020	10

3.1 与邻域粒化条件熵非增量式属性约简算法比较

本文所提出的邻域粒化条件熵增量式属性约简算法(记为IARNGCE)是在邻域粒化条件熵非增量式属性约简算法(记为ARNGCE)的基础上提出的,在本小节将这两种算法对表3中的数据集进行属性约简比较. 两种算法均采用条件熵度量进行属性分类性能的评估,并且令两种算法对各个数据集设定相同的邻域半径$\delta = 0.2$,距离度量函数选取为欧氏距离,因此两种算法对同一数据集的属性约简结果是一样的,只需要比较其属性约简时间性能. 图2为两种算法在所有数据集每次动态更新时属性约简的时间消耗比较.

由图2可见,当横坐标为0时,ARNGCE和IARNGCE的属性约简用时是相同的,这是由于在初始数据集的属性约简时,IARNGCE算法同样采用ARNGCE算法进行约简. 当数据集开始动态增加后,两种算法的属性约简用时发生了明显变化,对于规模较小的数据集,如数据集Wdbc、Hill_Valley和Mess,刚开始的一至两次更新,ARNGCE算法的属性约简时间稍高于IARNGCE算法,但是当数据集更新次数较多时,ARNGCE算法的约简用时大幅度高于IARNGCE算法. 对于规模较大的数据集,如数据集Wall和Magic,随着数据集动态更新,IARNGCE算法的约简用时总是大幅度小于ARNGCE算法,这是由于ARNGCE算法是一种非增量式的属性约简,当数据集更新后,ARNGCE算法在计算邻域粒化条件熵时需要进行大量的邻域粒重复计算,而IARNGCE算法进行的是增量计算,理论分析已表明所提出的增量计算方法避免了大量的重复计算,相比较于ARNGCE算法,实验结果验证了IARNGCE算法对动态数据集属性约简的有效性.

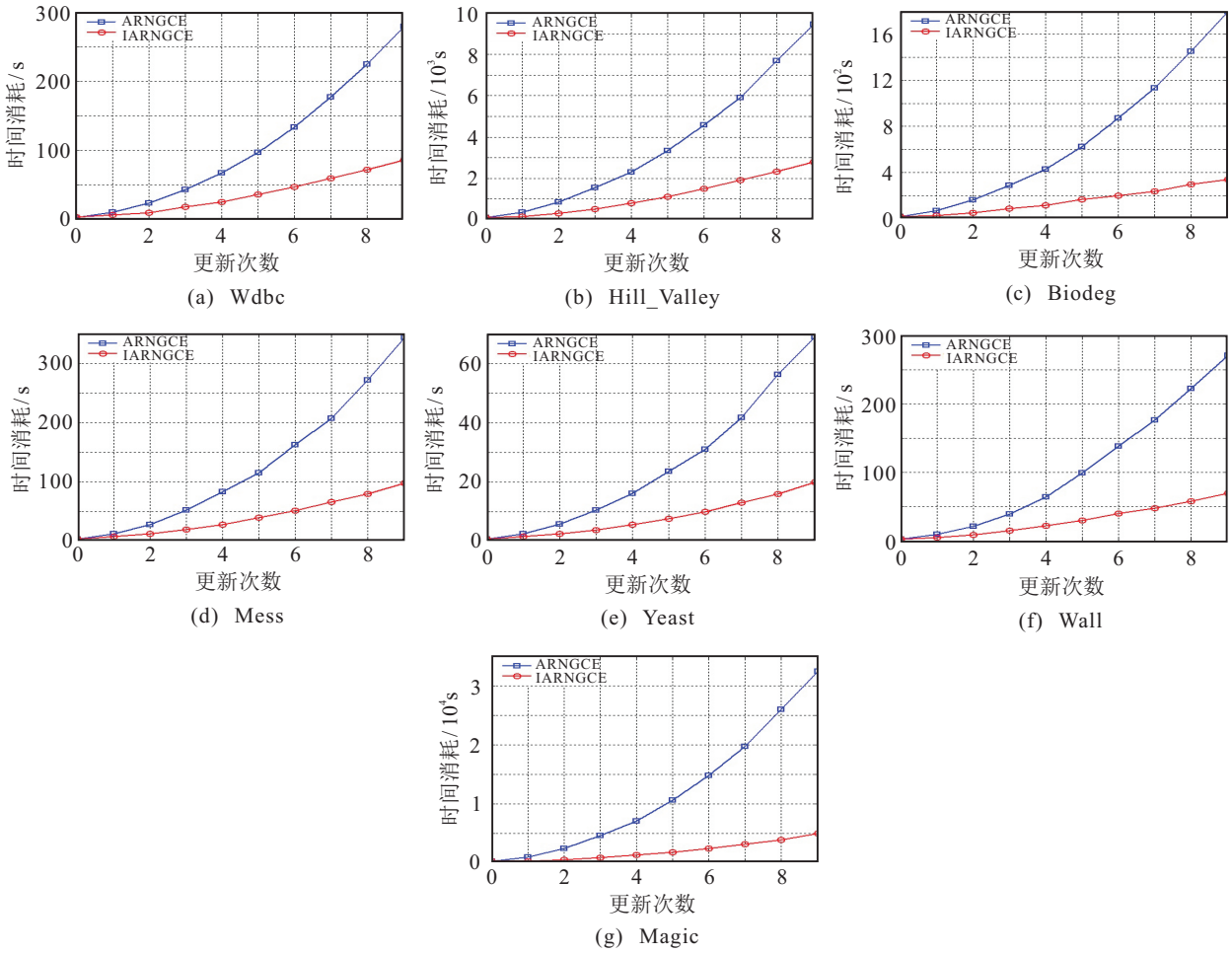

图 2　ARNGCE和IARNGCE属性约简的时间消耗对比

3.2　与同类型增量式属性约简算法比较

选取近年来提出的同类型增量式属性约简算法与本文所提出的邻域粒化条件熵增量式属性约简算法进行实验对比,从而验证所提出算法的增量式属性约简的优越性. 参与比较的算法分别为离散型信息系统对象增加的增量式属性约简算法[9](记为IAR-1)、基于变精度粗糙集模型的离散型信息系统增量式属性约简算法[13](记为IAR-2)、基于知识粒度的离散型信息系统增量式属性约简算法[11](记为IAR-3)和邻域粗糙集模型的增量式更新算法[23](记为IAR-4). 由于文献[23]给出了邻域粗糙集上下近似的增量更新,实验中采用依赖度进行增量式属性约简. 另外,IAR-1算法、IAR-2算法和IAR-3算法是针对离散型信息系统的增量式属性约简算法,在进行实验时需要将表3中的数据集进行数值型数据离散化处理. IAR-4算法和本文所提出的IARNGCE算法可以直接处理数值型数据,但是算法中都含有一个参数,即邻域半径δ,该值取值不同,对属性约简结果将产生一定的影响[18]. 根据目前已有的研究结果[18],邻域半径选取为[0.1, 0.3]之间较为适宜,取为0.2进行实验.

由于实验是将每个数据集的对象集分割成多份,然后通过逐份添加的方式模拟构造出对象集的动态增加,进行比较的5种增量式算法对数据集的每次更新都会得到相应的属性约简结果. 实验共模拟出9次数据集动态更新,为了简便,展示9次属性约简集大小的平均值,如表4所示.

表 4　5种增量式算法属性约简结果比较

数据集	原属性	IAR-1算法	IAR-2算法	IAR-3算法	IAR-4算法	IARNGCE算法
Wdbc	31	6.9	7.2	6.6	5.4	5.5
Hill_Valley	101	17.4	18.3	17.8	16.9	16.4
Biodeg	41	12.8	13.7	13.2	12.3	11.6
Mess	19	10.7	11.3	10.6	9.8	9.7
Yeast	9	5.6	5.8	5.3	4.9	4.5
Wall	5	5.0	5.0	5.0	4.3	4.3
Magic	10	6.0	7.0	6.5	5.6	5.2

由表4可见,各个属性约简算法选择出的平均约简集大小均小于原数据集属性全集的大小,部分数据集的平均约简集大小远小于属性全集,如数据集Hill_Valley和Biodeg. 这表明数据集中存在很多不相关属性,这些不相关属性对分类学习是不重要的,需要剔除以简化数据结构. 比较5种算法的实验结果可以发现,IAR-4算法和IARNGCE算法得到的平均约简集均小于其余3种算法,这是由于其余3种算法是基于离散型数据集的,而表3数据集在进行离散化的过程中改变原来数值型数据的取值分布,同时引入量化误差,改变了原有知识空间的粒化结果,使得在属性约简的过程中选择出了更多的属性. 比较IAR-4算法和IARNGCE算法的平均属性约简结果可以发现,IARNGCE算法在相当一部分数据集的平均约简集更小,是由于条件熵在属性约简方面的优越性[19, 24],它通过信息论的视角去评估属性,尤其在数值型数据中优于邻域粗糙集模型的依赖度评估,因此选择出的属性更精简.

分类精度是评估数据集中属性集分类性能的一种重要体现,为了比较各个增量式属性约简算法得到的约简集的优劣,实验通过支持向量机分类器(SVM)和朴素贝叶斯分类器(NB)分别对各个数据集每次更新后得到的属性约简结果进行分类训练,并评估出相应的分类精度. 由于有多个属性约简结果,实验将多个结果分类精度的平均值作为对应算法的分类精度结果,两种分类器的分类精度结果分别如表5和表6所示,“*”表示5种算法分类精度的最大值.

表5 5种增量式算法属性约简结果SVM分类精度比较 %

数据集	原属性	IAR-1算法	IAR-2算法	IAR-3算法	IAR-4算法	IARNGCE算法
Wdbc	95.37	97.21	96.75	95.36	96.83	97.58*
Hill_Valley	89.29	89.35*	87.24	86.93	86.65	88.56
Biodeg	90.43	92.28	90.49	91.57	92.74	93.46*
Mess	62.74	63.42	61.77	62.25	62.50	63.71*
Yeast	85.36	86.84	86.07	85.49	86.54	87.43*
Wall	92.64	94.51*	92.38	93.22	91.85	93.72
Magic	94.37	94.16	93.62	92.43	92.55	95.86*

表6 5种增量式算法属性约简结果NB分类精度比较 %

数据集	原属性	IAR-1算法	IAR-2算法	IAR-3算法	IAR-4算法	IARNGCE算法
Wdbc	94.14	94.16	92.68	91.73	93.52	95.24*
Hill_Valley	85.75	85.37*	83.95	82.15	83.43	84.27
Biodeg	88.62	89.65	87.28	88.53	87.37	90.29*
Mess	61.57	64.53*	62.78	63.21	62.62	63.98
Yeast	83.95	84.06	82.69	84.45	83.47	85.67*
Wall	90.88	92.34	92.85	91.14	92.52	93.41*
Magic	92.05	92.53	92.63	90.25	91.39	93.56*

由表5和表6可见,大部分约简集的分类精度高于原始属性集的分类精度,这更加表明数据集普遍存在着大量的不相关属性以及剔除这些属性的必要性. 比较表5和表6的5种增量式属性约简结果的分类精度可见,IAR-1算法和IARNGCE算法的分类精度在整体上比其他3种算法要高,这是由于两种算法均采用信息熵模型对属性进行评估,选择出的属性对分类具有更好的性能. 比较IAR-1算法与IARNGCE算法之间的分类精度可以发现,IARNGCE算法在大部分数据中的分类精度更高,如表5的Wdbc、Biodeg和Yeast等,表6的Wdbc、Wall和Magic等. 这是由于IARNGCE是直接处理数值型数据的算法,对于数值型数据的属性约简,本文所提出的IARNGCE算法能够选择出更优的属性约简结果.

图3为5种算法在各个数据集上增量式属性约简时间消耗比较结果. 由图3可见,5种算法在数据集每次更新属性约简时有不同的运行时间情况. 其中IAR-4算法在所有数据集中时间消耗最高,这是由IAR-4算法的增量机制决定的,当数值型信息系统有新对象加入,IAR-4算法在计算新对象的邻域时,需要将整个论域遍历计算一遍,因此消耗较多的运行时间. IAR-1算法和IAR-2算法的时间消耗大致位于中间位置,而IARNGCE算法和IAR-3算法在大部分数据集中拥有较低的时间消耗,这是由于IAR-1算法和IAR-2算法是针对离散型数据的增量式计算,计算新对象的信息粒时虽然不必比较整个论域,但是同样需

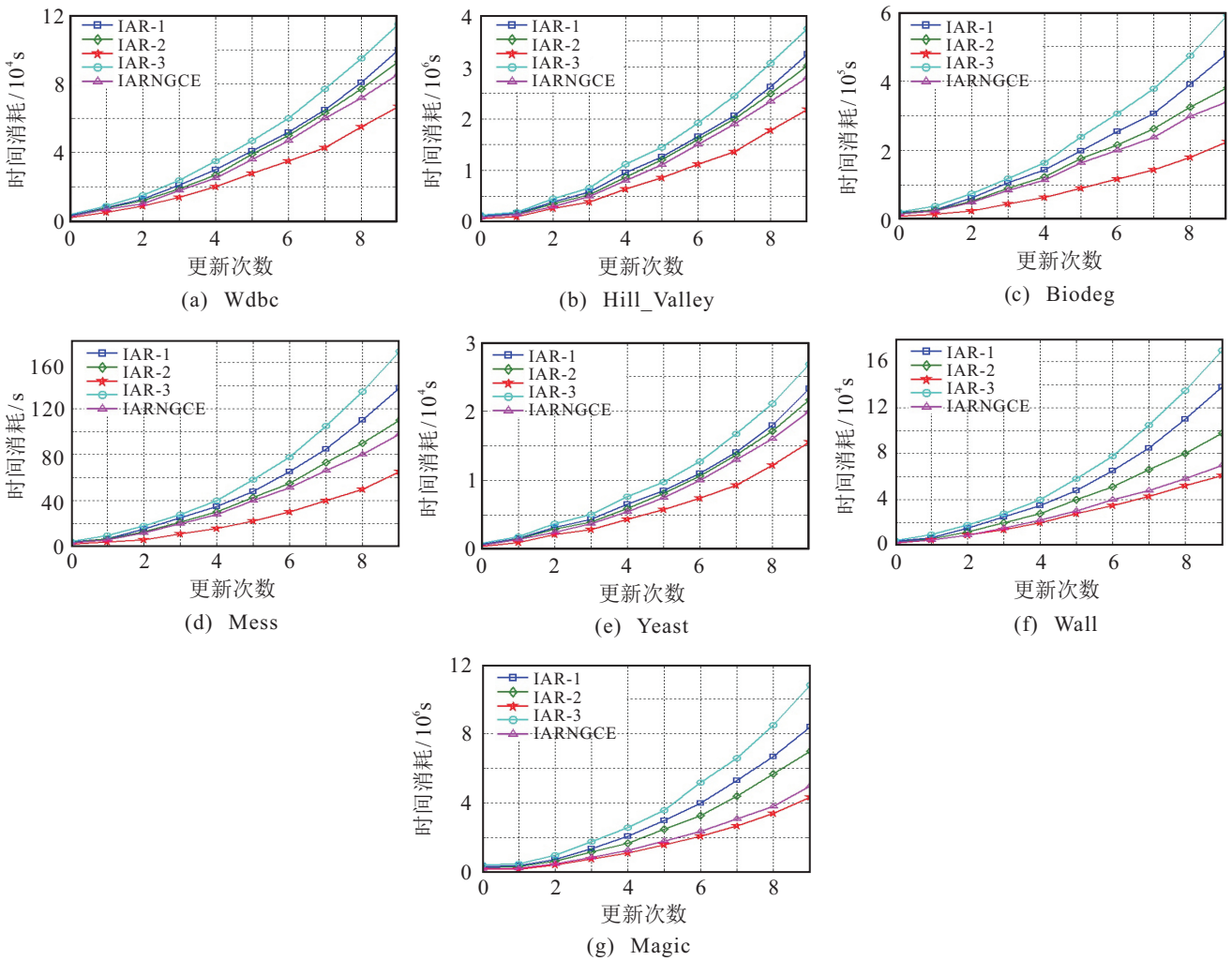

图 3　5种增量式算法属性约简时间消耗比较

要比较很多其他对象. IAR-3算法基于粒计算模型的知识粒度进行增量式属性约简,不需要进行上下近似的增量计算,只需对增加的对象进行等价类的处理,因此其处理效率略高于IARNGCE算法. IARNGCE算法采用分层的方法进行信息粒计算,只需要比较相邻对象层对象,极大地减小了计算量,具有很高的计算效率.

综合第3.1节和第3.2节两部分的实验结果可以证明,所提出的IARNGCE算法在数值型信息系统的增量式属性约简方面具有更高的有效性和优越性.

4　结　论

增量式属性约简是针对动态变化数据集的一种属性约简方法,目前的增量式属性约简研究只限于离散型信息系统. 邻域粗糙集模型是粒计算理论中处理数值型信息系统的一种有效工具,本文在邻域粒化条件熵的基础上提出了一种信息系统对象增加时的增量式属性约简算法. 首先在数值型信息系统中建立一种分层的邻域粒化计算,然后基于该方法给出邻域粒化的增量式计算,最后进一步提出邻域粒化条件熵的增量式更新,并设计出相对应的增量式属性约简算法. UCI数据集的实验结果表明了所提出算法对于数值型数据增量式属性约简的有效性和优越性.

本文所提出的增量式属性约简是针对信息系统对象集不断增加的情形,接下来将针对属性不断增加的情形进行增量式属性约简的研究.

参考文献(References)

[1] Liu G L, Hua Z, Zou J Y. Local attribute reductions for decision tables[J]. Information Sciences, 2018, 422: 204-217.

[2] Pawlak Z. Rough sets[J]. Int J of Parallel Programming, 1982, 11(5): 341-356.

[3] Gacek A. Granular modelling of signals: A framework of granular computing[J]. Information Sciences, 2013, 221(2): 1-11.

[4] 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.)

[5] 潘瑞林, 李园沁, 张洪亮, 等. 基于信息熵的模糊粗糙属性约简方法[J]. 控制与决策, 2017, 32(2): 340-348.
(Pan R L, Li Y Q, Zhang H L, et al. Fuzzy-rough attribute reduction algorithm based on information entropy[J]. Control and Decision, 2017, 32(2): 340-348.)

[6] Wang F, Liang J Y, Qian Y H. Attribute reduction: A dimension incremental strategy[J]. Knowledge-Based Systems, 2013, 39(2): 95-108.

[7] Chan C C. A rough set approach to attribute generalization in data mining[J]. Information Sciences, 1998, 107(1/2/3/4): 169-176.

[8] Jing Y G, Li T R, Huang J F, et al. An incremental attribute reduction approach based on knowledge granularity under the attribute generalization[J]. Int J of Approximate Reasoning, 2016, 76(9): 80-95.

[9] Liang J Y, Wang F, Dang C Y, et al. A group incremental approach to feature selection applying rough set technique[J]. IEEE Trans on Knowledge & Data Engineering, 2014, 26(2): 294-308.

[10] Jing Y G, Li T R, Hamidoă F, et al. An incremental attribute reduction approach based on knowledge granularity with a multi-granulation view[J]. Information Sciences, 2017, 411(7): 23-38.

[11] 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(7): 24-38.

[12] Raza M S, Qamar U. An incremental dependency calculation technique for feature selection using rough sets[J]. Information Sciences, 2016(343/344): 41-65.

[13] Chen D G, Yang Y Y, Dong Z. An incremental algorithm for attribute reduction with variable precision rough sets[J]. Applied Soft Computing, 2016, 45: 129-149.

[14] 冯少荣, 张东站. 一种高效的增量式属性约简算法[J]. 控制与决策, 2011, 26(4): 495-500.
(Feng S R, Zhang D Z. Effective increment algorithm for attribute reduction[J]. Control and Decision, 2011, 26(4): 495-500.)

[15] 钱进, 朱亚炎. 面向成组对象集的增量式属性约简算法[J]. 智能系统学报, 2016, 11(4): 496-502.
(Qian J, Zhu Y Y. An incremental attribute reduction algorithm for group objects[J]. CAAI Trans Intelligent Systems, 2016,11(4): 496-502.)

[16] Liu D, Li T R, Zhang J B. A rough set-based incremental approach for learning knowledge in dynamic incomplete information systems[J]. Int J of Approximate Reasoning, 2014, 55(8): 1764-1786.

[17] Shu W H, Shen H. Incremental feature selection based on rough set in dynamic incomplete data[J]. Pattern Recognition, 2014, 47(12): 3890-3906.

[18] 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.

[19] Zhao H, Qin K Y. Mixed feature selection in incomplete decision table[J]. Knowledge-Based Systems, 2014, 57(2): 181-190.

[20] Liu Y, Huang W L, Jiang Y L, et al. Quick attribute reduct algorithm for neighborhood rough set model[J]. Information Sciences, 2014, 271(7): 65-81.

[21] Liang J Y, Shi Z Z. The information entropy, rough entropy and knowledge granulation in rough set theory[J]. Int J of Uncertainty, Fuzziness and Knowledge-Based Systems, 2008, 12(1): 37-46.

[22] Wang C Z, He Q, Shao M W, et al. A unified information measure for general binary relations[J]. Knowledge-Based Systems, 2017, 135(11): 18-28.

[23] Zhang J B, Li T R, Ruan D, et al. Neighborhood rough sets for dynamic data mining[J]. Int J of Intelligent Systems, 2012, 27(4): 317-342.

[24] Hu Q H, Pedrycz W, Yu D R, et al. Selecting discrete and continuous features based on neighborhood decision error minimization[J]. IEEE Trans on Systems, Man, and Cybernetics: Part B, 2010, 40(1): 137-150.

作者简介

赵小龙(1974–), 男, 副教授, 从事人工智能、算法设计和粒计算等研究, E-mail: zhaoxiaolong1974@163.com;

杨燕(1964–), 女, 教授, 博士, 从事计算智能、数据挖掘和集成学习等研究, E-mail: yyang@swjtu.edu.cn.

(责任编辑: 郑晓蕾)