混合数据的邻域区分度增量式属性约简算法

盛 魁[1],王 伟[3],卞显福[2],董 辉[1],马 健[1]
(1. 亳州职业技术学院信息工程系,安徽亳州 236800;2. 中国科学技术大学软件学院,安徽合肥 230051;
3. 安徽大学计算机科学与技术学院,安徽合肥 230601)

摘 要: 增量式属性约简是一种针对动态环境下的数据挖掘方法.目前已经提出的增量式属性约简算法仅适用于符号型的信息系统,而很少有对混合信息系统进行相关的研究,这促使在混合信息系统下构建相关的增量式属性约简算法.区分度是用于设计属性约简的一种重要方法,本文将传统的区分度在混合信息系统下进行推广,提出邻域区分度的概念,然后分别研究了邻域区分度在混合信息系统下对象增加和对象减少时的增量式学习,最后根据这种增量式学习分别提出了对应的增量式属性约简算法.UCI 数据集上的相关实验结果表明,所提出的增量式属性约简比非增量式属性约简能够更快速的更新约简结果.

关键词: 粗糙集;混合数据;区分度;邻域关系;增量式学习;属性约简
中图分类号: TP18 **文献标识码:** A **文章编号:** 0372-2112 (2020)04-0682-15
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2020.04.010

Neighborhood Discernibility Degree Incremental Attribute Reduction Algorithm for Mixed Data

SHENG Kui[1], WANG Wei[3], BIAN Xian-fu[2], DONG Hui[1], MA Jian[1]
(1. *Department of Information Engineering, Bozhou Vocational and Technical College, Bozhou, Anhui* 236800, *China*;
2. *School of Software, University of Science and Technology of China, Hefei, Anhui* 230051, *China*;
3. *Computer Science and Technology Institute, Anhui University, Hefei, Anhui* 230601, *China*)

Abstract: Incremental attribute reduction is a data mining method for dynamic environment. The incremental attribute reduction algorithm already proposed is only applicable to symbolic information systems. However, there are few related studies on mixed information systems, which promotes the construction of the related incremental attribute reduction algorithm under the mixed information system. The discernibility degree is an important method used for designing attribute reduction. In this paper, the traditional discernibility degree is generalized under the mixed information system, and the concept of neighborhood discernibility degree is presented. Then, the incremental learning of neighborhood discernibility degree is studied respectively when objects increase or objects decrease under the mixed information system. Finally, according to this incremental learning, the corresponding incremental attribute reduction algorithms are proposed, respectively. The related experimental results on the UCI data set show that the proposed incremental attribute reduction can update the reduction results more quickly than the non incremental attribute reduction.

Key words: rough set; mixed data; discernibility degree; neighborhood relation; incremental learning; attribute reduction

1 引言

属性约简[1]即特征选择,是粗糙集理论[1]在机器学习和数据挖掘等领域中一种重要的应用.在分类学习等任务中,标准的数据集由很多属性构成,然而这些属性中很多对分类任务是不相关的,这些不相关的属性不仅影响了数据的知识挖掘性能,而且还降低了各类学习算法的学习效率,因此我们需要对这些属性进

收稿日期:2018-04-13;修回日期:2019-09-25;责任编辑:孙瑶
基金项目:安徽省高等学校省级自然科学研究重点基金(No. KJ2015A417, No. KJ2016A493);安徽省高校振兴计划优秀青年人才支持计划(No. GXYQZD2016529);安徽省亳州市产业创新创新团队项目(亳组[2015]20 号-2)

行删除,其中属性约简正是实现这种目的的一种常用方法[2,3].

在当今的大数据环境下,数据不仅体量巨大而且变化迅速,其中数据不断变化的这一特征便给传统的各类数据挖掘模型带来了很大的挑战,这主要是由于传统的模型与算法针对的是静态的数据,而难以满足动态环境下数据处理的时效性需求,解决这类问题的主要方法是对模型和算法构建增量式学习. 针对动态数据环境下的属性约简,学者们提出了一种相适应的改进方法,称之为增量式属性约简. 目前已有大量的增量式属性约简算法被提出,Chan[4]通过分析粗糙集中边界域的变化提出了最早的信息系统增量式属性约简算法,Shu 等[5,6]学者在不完备信息系统中通过依赖度度量分别提出了信息系统对象增加和属性增加时的增量式属性约简,Qian 等[7]学者在 Shu 的基础上也提出了类似的增量式属性约简,Chen 等[8]学者在变精度粗糙集模型中提出了区分矩阵的增量式更新,然后构造出相应的增量式属性约简算法,同时 Wei 等[9]学者也用区分矩阵在动态数据下构造出了类似的增量式属性约简,钱进等[10]学者针对信息系统中对象增加问题,提出了一种面向成组对象集的增量式属性约简算法,Lang 等[11]学者在覆盖信息系统中提出了一种改进的增量式属性约简. 另一方面,Xie 等[12]学者针对目前信息系统属性值的变化情形,在不完备信息系统中提出了更新属性约简的方法,Jing 等[13,14]学者将粒计算模型运用在属性约简上,提出了一种基于知识粒度的增量式属性约简算法. 总之,关于增量式属性约简的研究已成为动态数据下数据挖掘的研究热点.

然而,目前已提出的增量式属性约简都是建立在符号型信息系统的基础上,对于数值型和混合型信息系统的相关研究几乎是一片空白,这主要是由于在数值型和混合型信息系统中,属性值中包含了具体的数值,而数值的存在对于粗糙集的增量式学习带来了一定的困难[5,6],因此这促使我们对混合信息系统下的增量式属性约简进行相关的研究.

目前,邻域粗糙集是处理混合型数据的一种重要模型[15,16],同时也是对混合型信息系统进行属性约简的常用方法[15,17]. 为了解决混合信息系统下动态数据的属性约简问题,本文在邻域粗糙集模型的基础上构建一种增量式属性约简算法. 在文献[18]中,学者们提出了邻域粗糙集中邻域类的增量更新,本文在其基础上进一步改进与推广,提出混合信息系统中多个对象增加和多个对象减少时的邻域类增量更新. 区分度是一种重要的属性集度量方法[19],也是构造属性约简的一种重要方法[20],本文将符号型信息系统下的区分度在混合信息系统下进行推广,提出邻域区分度,然后在邻域类增量更新的基础上,进一步的研究了当对象增加和对象减少时邻域区分度的增量式学习,理论证明了该增量式更新的高效性,最后基于邻域区分度的增量式学习分别提出了混合信息系统对象增加和对象减少时的增量式属性约简算法. UCI 数据集的仿真实验结果表明,所提出的增量式属性约简算法具有更高的属性约简效率,能够满足动态环境下数据处理的时间需求.

2　基本理论

在数据挖掘与知识发现理论中,结构化的数据集表示为信息系统,设信息系统 $IS=(U,AT)$,这里的 U 称为信息系统的论域,其中 $U=\{x_1,x_2,\cdots,x_n\}$,$\forall x_i\in U$ 称之为论域中的对象,即数据集的每个样本;这里的 AT 称为信息系统的属性集,其中 $AT=\{a_1,a_2,\cdots,a_m\}$,$\forall a_i\in AT$ 称为属性集中的属性,即数据集中的每个特征. 对于一个信息系统 $IS=(U,AT)$,若属性集为 $AT=\{a_1,a_2,\cdots,a_m,d\}$,其中 d 为信息系统的类属性,那么该信息系统又称为决策信息系统 $IS=(U,AT=C\cup D)$,这里的 $C=\{a_1,a_2,\cdots,a_m\}$,$D=\{d\}$ 分别称之为条件属性集和决策属性.

定义 1[1]　在粗糙集理论中,不可区分关系是该理论的核心,对于信息系统 $IS=(U,AT)$,设属性集 $A\subseteq AT$,那么 A 确定的不可区分关系 $IND(A)$ 定义为

$$IND(A)=\{(x,y)\in U\times U\mid\forall a\in A,a(x)=a(y)\} \tag{1}$$

这里的 $a(x)$ 表示对象 x 在属性 a 下的属性值.

根据定义 1 可以看出,不可区分关系 $IND(A)$ 即等价关系. 不可区分关系 $IND(A)$ 可以在论域 U 上诱导出一个划分 $U/IND(A)=\{X_1,X_2,\cdots,X_p\}$,其中 $X_i,1\leqslant i\leqslant p$ 称为等价类,对于 $\forall x\in U$ 关于不可区分关系 $IND(A)$ 的等价类定义为 $[x]_A=\{y\in U\mid(x,y)\in IND(A)\}$. 同样地,等价类也是粗糙集理论中一个很重要的概念,它是粗糙集理论进行粗糙近似计算的基本单位,通过等价类可以对不确定性的知识进行有效的描述. 近年来,Susmaga 等[21]学者在不可区分关系的基础上,提出了区分关系.

定义 2[21]　对于信息系统 $IS=(U,AT)$,设属性集 $A\subseteq AT$,那么 A 确定的区分关系 $DIS(A)$ 定义为

$$DIS(A)=\{(x,y)\in U\times U\mid\exists a\in A,a(x)\neq a(y)\} \tag{2}$$

可以看出,区分关系与不可区分关系是完全相反且互补的概念[19],这表明任意两个对象要么满足区分关系,要么满足不可区分关系,即 $IND(A)\cup DIS(A)=U\times U$.

在区分关系的基础上,Teng 等[20]学者进一步的提出了区分度以及相对区分度的概念.

定义 3[20] 对于信息系统 $IS=(U,AT)$,设属性集 $A\subseteq AT$,A 在论域 U 上诱导的划分为 $U/IND(A)=\{X_1,X_2,\cdots,X_p\}$,并且 A 确定的区分关系为 $DIS(A)$,那么 $DIS(A)$ 在该信息系统下的区分度定义为

$$DD(A)=|DIS(A)|=|U|^2-\sum_{i=1}^{p}|X_i|^2 \quad (3)$$

这里的 $|\cdot|$ 表示集合的基数. 同时 Teng 等学者在其基础上定义了相对区分度. 对于信息系统 $IS=(U,AT)$,设属性集 $A,B\subseteq AT$,A 和 $A\cup B$ 在论域 U 下确定的划分分别表示为

$$U/IND(A)=\{X_1,X_2,\cdots,X_p\}$$

和 $U/IND(A\cup B)=\{Y_1,Y_2,\cdots,Y_q\}$,

那么 B 关于 A 的相对区分度定义为

$$DD_A(B)=\sum_{i=1}^{p}|X_i|^2-\sum_{i=1}^{q}|Y_i|^2 \quad (4)$$

在文献[20]中,Teng 等学者证明了相对区分度满足单调性,并由此构造出了基于区分度的属性约简算法.

定义 4[20] 对于决策信息系统 $IS=(U,C\cup D)$,若属性集 $R\subseteq C$ 是该信息系统的一个属性约简,那么当且仅当

$$(a)\quad DD_C(D)=DD_R(D) \quad (5)$$

$$(b)\quad \forall a\in R, DD_C(D)<DD_{R-\{a\}}(D) \quad (6)$$

根据定义 4 中关于属性约简的定义,基于区分度的属性约简算法如算法 1 所示.

算法 1[20] **基于区分度的属性约简算法(ARDD,Attribute Reduction based on Discernibility Degree)**

输入:决策信息系统 $IS=(U,C\cup D)$;
输出:属性约简集 R.
Step1:初始化 $R=\varnothing$;
Step2:对于 $\forall a\in C-R$,计算属性 a 的属性重要度 $sig_R(a)=DD_R(D)-DD_{R\cup\{a\}}(D)$;
Step3:找出 $C-R$ 中属性重要度最大的属性,记为 a';
Step4:判断 $sig_R(a')$ 的值,若 $sig_R(a')>0$,那么 $R=R\cup\{a'\}$,并转入 Step2,若 $sig_R(a')=0$,转入 Step5;
Step5:返回属性约简 R.

算法 1 表明,区分度的属性约简事实上是一种基于启发式函数的属性约简算法,算法运行的过程中,将相对区分度作为启发式函数来搜索属性,并且每次贪心的选择最大函数值的属性作为属性约简集的候选属性,按照这一过程不断的进行迭代,直到剩余属性中最大函数值为 0 时终止算法,此时得到的约简集即为区分度的属性约简.

3 混合数据的邻域区分度与增量式学习

现实环境下的数据集是复杂多样的,很多信息系统中的属性集是混合类型的,即符号型属性和数值型属性并存,并且现实环境下信息系统的论域都是处于不断动态变化之中的,这使得各类知识发现和数据挖掘算法面临着巨大的挑战. 在粗糙集理论中,传统的模型都是基于符号型的信息系统设计,而对于包含数值型数据的信息系统却无能为力. 为了解决这一难题,近年来 Hu 等[15]学者提出了邻域关系,并利用邻域粗糙集模型来处理混合型信息系统. 在本节,我们在 Hu 等[15]学者的邻域关系基础上,将区分度进行推广,提出混合数据下的邻域区分度,并且对于混合型信息系统对象动态增加和减少的情形,对邻域区分度构造增量式学习,为下一节中增量式属性约简算法的提出提供铺垫.

3.1 混合数据的邻域区分度

混合型数据通常表示成混合型信息系统 $MIS=(U,AT)$ 的形式,这里的 U 称为混合型信息系统的论域,其中 $U=\{x_1,x_2,\cdots,x_n\}$,$\forall x_i\in U$ 称之为论域的对象;这里的 AT 称为混合型信息系统的属性集,其中 $AT=\{a_1,a_2,\cdots,a_m\}$,$\forall a_i\in AT$ 称为属性集中的属性,并且 $\exists a_i\in AT$,属性 a_i 是数值型属性. 对于一个混合型信息系统 $IS=(U,AT)$,若属性集为 $AT=\{a_1,a_2,\cdots,a_m,d\}$,其中的 d 为信息系统的类属性,那么该信息系统又称之为混合型决策信息系统 $MDIS=(U,AT=C\cup D)$,这里的 $C=\{a_1,a_2,\cdots,a_m\}$,$D=\{d\}$ 分别称之为条件属性集和决策属性.

定义 5[15] 对于混合型信息系统 $IS=(U,AT)$,属性集 $A\subseteq AT$,设属性集 $A_1\cup A_2=A$ 且 $A_1\cap A_2=\varnothing$,其中 A_1 和 A_2 分别为符号型属性集和数值型属性集,那么 A 确定的邻域关系 $N_\delta(A)$ 定义为

$$N_\delta(A)=\{(x,y)\in U\times U|(\forall a\in A_1,a(x)=a(y))\wedge d_{A_2}(x,y)\leqslant\delta\} \quad (7)$$

这里的 $d_{A_2}(x,y)$ 称为对象 x 与 y 之间的距离,通常对象之间的距离度量采用闵可夫斯基距离. δ 为一个非负常数,称之为邻域半径.

由于邻域关系满足自反性和对称性,但不满足传递性,因此邻域关系将论域诱导出一组覆盖,对于混合信息系统 $IS=(U,AT)$,论域 $U=\{x_1,x_2,\cdots,x_n\}$,给定属性集 A 和邻域半径 δ,那么论域 U 在邻域关系 $N_\delta(A)$ 下诱导出的邻域覆盖为 $U/N_\delta(A)=\{\delta_A(x_1),\delta_A(x_2),\cdots,\delta_A(x_n)\}$,这里的 $\delta_A(x)$ 表示对象 $x\in U$ 在邻域关系 $N_\delta(A)$ 下的邻域类,定义为 $\delta_A(x)=\{y\in U|(x,y)\in N_\delta(A)\}$.

在定义 5 中邻域关系的基础上,我们可以将区分度在混合型信息系统下进行推广,提出混合型信息系统的邻域区分度.

定义 6 对于混合型信息系统 $MIS=(U,AT)$,$U=\{x_1,x_2,\cdots,x_n\}$,设属性集 $A\subseteq AT$,A 在论域 U 上确定的

邻域关系为 $N_\delta(A)$,并且诱导的覆盖为 $U/N_\delta(A)=\{\delta_A(x_1),\delta_A(x_2),\cdots,\delta_A(x_n)\}$,那么属性集 A 在该信息系统下的邻域区分度定义为

$$NDD(A)=n^2-\sum_{i=1}^{n}|\delta_A(x_i)| \tag{8}$$

性质 1　对于混合型信息系统 $MIS=(U,AT)$,$U=\{x_1,x_2,\cdots,x_n\}$,设属性集 $A\subseteq AT$ 且 $\forall a\in A$ 均为符号型属性,那么属性集 A 在该信息系统下的邻域区分度为

$$NDD(A)=n^2-\sum_{i=1}^{p}|X_i|^2 \tag{9}$$

这里的 $U/IND(A)=\{X_1,X_2,\cdots,X_p\}$.

证明　对于混合型信息系统 $MIS=(U,AT)$,若属性集 A 内的属性均为符号型属性,那么根据定义 5,邻域关系 $N_\delta(A)$ 即为等价关系 $IND(A)$. 对于 $\forall x\in U$ 在邻域关系 $N_\delta(A)$ 下的等价类为 $[x]_A$,那么根据定义 6 满足 $NDD(A)=n^2-\sum_{i=1}^{n}|[x_i]_A|$. 由于在等价关系中,$\forall x_j\in[x_i]_A$ 都有 $[x_i]_A=[x_j]_A$. 设 $[x_i]_A=\{x_{t1},x_{t2},\cdots,x_{tc}\}$,那么 $|[x_{t1}]_A|+|[x_{t2}]_A|+\cdots+|[x_{tc}]_A|=|[x_i]_A|\cdot|[x_i]_A|$. 所以 $NDD(A)=n^2-\sum_{i=1}^{p}|X_i|^2$.

性质 1 表明,当属性集只包含符号型属性时,邻域区分度便退化为区分度,因此邻域区分度是区分度的推广,区分度是邻域区分度的特例. 在邻域区分度的基础上,我们可以进一步提出相对邻域区分度.

定义 7　对于混合型信息系统 $MIS=(U,AT)$,$U=\{x_1,x_2,\cdots,x_n\}$,设属性集 $A,B\subseteq AT$,并且属性集 A 和 B 在论域 U 上诱导的邻域覆盖分别为 $U/N_\delta(A)=\{\delta_A(x_1),\delta_A(x_2),\cdots,\delta_A(x_n)\}$ 和 $U/N_\delta(B)=\{\delta_B(x_1),\delta_B(x_2),\cdots,\delta_B(x_n)\}$,那么属性集 B 在 A 下的相对邻域区分度定义为

$$NDD_A(B)=\sum_{i=1}^{n}|\delta_A(x_i)|-\sum_{i=1}^{n}|\delta_A(x_i)\cap\delta_B(x_i)| \tag{10}$$

由于 $\forall x\in U$ 有 $1\leqslant|\delta_A(x)|\leqslant n$,所以 $0\leqslant NDD_A(B)\leqslant n^2-n$.

特别的,对于混合型决策信息系统 $MDIS=(U,C\cup D)$,$U=\{x_1,x_2,\cdots,x_n\}$,设属性集 $A\subseteq AT$ 且 A 在论域 U 上诱导的邻域覆盖为 $U/N_\delta(A)=\{\delta_A(x_1),\delta_A(x_2),\cdots,\delta_A(x_n)\}$,对象 $\forall x_i\in U$ 在决策属性 D 下的决策类为 $[x_i]_D$,那么决策属性 D 关于属性集 A 的相对邻域区分度为

$$NDD_A(D)=\sum_{i=1}^{n}|\delta_A(x_i)|-\sum_{i=1}^{n}|\delta_A(x_i)\cap[x_i]_D| \tag{11}$$

同样的,$0\leqslant NDD_A(D)\leqslant n^2-n$.

定理 1　对于混合型决策信息系统 $MDIS=(U,C\cup D)$,$U=\{x_1,x_2,\cdots,x_n\}$,设属性集 $A\subseteq B\subseteq AT$,并且属性集 A 和 B 在论域 U 上诱导的邻域覆盖分别为 $U/N_\delta(A)$ 和 $U/N_\delta(B)$,对象 $\forall x_i\in U$ 在决策属性 D 下的决策类为 $[x_i]_D$,决策属性 D 关于属性集 A 的相对邻域区分度为 $NDD_A(D)$,决策属性 D 关于属性集 B 的相对邻域区分度为 $NDD_B(D)$,那么满足

$$NDD_A(D)\geqslant NDD_B(D) \tag{12}$$

证明

$$\begin{aligned}&NDD_A(D)-NDD_B(D)\\&=\sum_{i=1}^{n}|\delta_A(x_i)|-\sum_{i=1}^{n}|\delta_A(x_i)\cap[x_i]_D|-\sum_{i=1}^{n}|\delta_B(x_i)|\\&\quad+\sum_{i=1}^{n}|\delta_B(x_i)\cap[x_i]_D|\\&=\sum_{i=1}^{n}|\delta_A(x_i)|-\sum_{i=1}^{n}|\delta_B(x_i)|\\&\quad-\left(\sum_{i=1}^{n}|\delta_A(x_i)\cap[x_i]_D|-\sum_{i=1}^{n}|\delta_B(x_i)\cap[x_i]_D|\right)\end{aligned}$$

由于 $A\subseteq B\subseteq AT$,根据定义 5 有 $\forall x\in U$ 满足 $\delta_B(x)\subseteq\delta_A(x)$,那么满足

$$\begin{aligned}&|\delta_A(x)|-|\delta_B(x)|=|\delta_A(x)-\delta_B(x)|\\&|\delta_A(x_i)\cap[x_i]_D|-|\delta_B(x_i)\cap[x_i]_D|\\&=|(\delta_A(x_i)-\delta_B(x_i))\cap[x_i]_D|\end{aligned}$$

所以

$$\begin{aligned}&NDD_A(D)-NDD_B(D)\\&=\sum_{i=1}^{n}(|\delta_A(x_i)-\delta_B(x_i)|)\\&\quad-\sum_{i=1}^{n}|(\delta_A(x_i)-\delta_B(x_i))\cap[x_i]_D|\geqslant 0\end{aligned}$$

即 $NDD_A(D)\geqslant NDD_B(D)$.　证毕

定理 1 表明,随着属性集的逐渐增大,混合信息系统的邻域区分度保持单调不增加,因此根据邻域区分度可以构造出混合信息系统的属性约简. 然而,现实环境下的信息系统是动态变化的,直接基于邻域区分度构造的属性约简是一种针对静态数据的约简算法,因此这促使我们提出基于邻域区分度的增量式属性约简,从而适应动态下的数据环境. 在本文我们将研究混合信息系统对象动态变化时的增量式属性约简,在提出增量式算法之前,首先研究混合信息系统对象动态变化时邻域区分度的增量式学习.

3.2　混合数据对象增加时邻域区分度的增量式更新

在混合信息系统论域的动态变化中,通常分为两种变化类型,第一种为论域中对象的逐渐增加,另一种为论域中对象的逐渐减少. 在本小节,我们将研究混合型信息系统论域中对象动态增加时邻域区分度的增量式更新.

根据定义7可以看出,邻域区分度的计算主要是关于论域中对象邻域类的计算,因此我们首先分析混合型信息系统对象增加时邻域类的增量式更新.

引理1 设 t 时刻的混合型信息系统为 $MIS^{(t)}=(U^{(t)},AT)$,$U^{(t)}=\{x_1,x_2,\cdots,x_n\}$,属性集 $A\subseteq AT$ 在论域 $U^{(t)}$ 上确定的邻域关系为 $N_\delta^{U^{(t)}}$,$t+1$ 时刻混合型信息系统增加了对象集 ΔU,那么 $t+1$ 时刻的混合型信息系统可表示为 $MIS^{(t+1)}=(U^{(t+1)},AT)$,其中 $U^{(t+1)}=U^{(t)}\cup\Delta U$. 属性集 $A\subseteq AT$ 在论域 $U^{(t+1)}$ 上确定的邻域关系为 $N_\delta^{U^{(t+1)}}$,那么满足 $N_\delta^{U^{(t)}}\subseteq N_\delta^{U^{(t+1)}}$.

证明 对于任意 $(x,y)\in N_\delta^{U^{(t)}}$,那么 $x,y\in U^{(t)}$. 由于 $U^{(t+1)}=U^{(t)}\cup\Delta U$,所以 $x,y\in U^{(t+1)}$. 由于属性集 A 和邻域半径 δ 不变,因此 $(x,y)\in N_\delta^{U^{(t+1)}}$,即 $N_\delta^{U^{(t)}}\subseteq N_\delta^{U^{(t+1)}}$. 证毕

定理2 设 t 时刻的混合型信息系统为 $MIS^{(t)}=(U^{(t)},AT)$,$U^{(t)}=\{x_1,x_2,\cdots,x_n\}$,属性集 $A\subseteq AT$ 在论域 $U^{(t)}$ 上诱导的邻域覆盖为 $\{\delta_A^{U^{(t)}}(x_1),\delta_A^{U^{(t)}}(x_2),\cdots,\delta_A^{U^{(t)}}(x_n)\}$,其中 $\delta_A^{U^{(t)}}(x_i)$ 表示对象 $x_i\in U^{(t)}$ 在论域 $U^{(t)}$ 上的邻域类. 在 $t+1$ 时刻混合型信息系统增加了对象集 $\Delta U=\{x_{n+1},x_{n+2},\cdots,x_{n+c}\}$,那么 $t+1$ 时刻的混合型信息系统可表示为 $MIS^{(t+1)}=(U^{(t+1)},AT)$,其中 $U^{(t+1)}=U^{(t)}\cup\Delta U$. 则属性集 $A\subseteq AT$ 在论域 $U^{(t+1)}$ 上诱导的邻域覆盖增量更新为

$$\begin{aligned}&\{\delta_A^{U^{(t+1)}}(x_1),\delta_A^{U^{(t+1)}}(x_2),\cdots,\delta_A^{U^{(t+1)}}(x_n),\\&\delta_A^{U^{(t+1)}}(x_{n+1}),\delta_A^{U^{(t+1)}}(x_{n+2}),\cdots,\delta_A^{U^{(t+1)}}(x_{n+c})\}\end{aligned}\tag{13}$$

这里的 $\delta_A^{U^{(t+1)}}(x_{n+1}),\delta_A^{U^{(t+1)}}(x_{n+2}),\cdots,\delta_A^{U^{(t+1)}}(x_{n+c})$ 直接按照定义5中关于邻域类的计算方法进行计算. 对于 $\delta_A^{U^{(t+1)}}(x_1),\delta_A^{U^{(t+1)}}(x_2),\cdots,\delta_A^{U^{(t+1)}}(x_n)$,有

$$\delta_A^{U^{(t+1)}}(x_i)=\delta_A^{U^{(t)}}(x_i)\cup\{x\mid x_i\in\delta_A^{U^{(t+1)}}(x),\forall x\in\Delta U\}\quad 1\leqslant i\leqslant n\tag{14}$$

证明 首先根据引理1有 $N_\delta^{U^{(t)}}\subseteq N_\delta^{U^{(t+1)}}$,那么对象 $\forall x\in U^{(t)}$ 满足 $\delta_A^{U^{(t)}}(x)\subseteq\delta_A^{U^{(t+1)}}(x)$. 对于 $x_{n+i}(1\leqslant i\leqslant c)$,在论域 $U^{(t+1)}$ 下对应的邻域类为 $\delta_A^{U^{(t+1)}}(x_{n+i})$,若存在对象 $x_j(1\leqslant j\leqslant n)$ 满足 $x_j\in\delta_A^{U^{(t+1)}}(x_{n+i})$,那么表明 $(x_{n+i},x_j)\in N_\delta^{U^{(t+1)}}$,所以 $x_{n+i}\in\delta_A^{U^{(t+1)}}(x_j)$,并且 $x_{n+i}\notin\delta_A^{U^{(t)}}(x_j)$,因此 $\delta_A^{U^{(t+1)}}(x_i)=\delta_A^{U^{(t)}}(x_i)\cup\{x\mid x_i\in\delta_A^{U^{(t+1)}}(x),\forall x\in\Delta U\}$,$1\leqslant i\leqslant n$. 定理2成立. 证毕

定理2表明,当混合型信息系统增加对象集后,我们只需要对新加入的对象按照定义5的方法计算其邻域类,而原混合型信息系统论域中的对象在新论域下的邻域类只需要进行相应的增量更新,不必重新按照定义5进行计算,这样做的优点在于提高整个新信息系统对象邻域类的计算效率,减少了大量的不必要重复计算.

根据定义2中关于混合型信息系统对象增加后邻域类的增量更新,接下来就可以在此基础上进行邻域区分度的增量更新构造. 由于本文探究的是混合型信息系统下邻域区分度的增量式属性约简,因此我们直接研究相对邻域区分度的增量式更新. 在粗糙集理论的增量式学习研究中,学者们通常采用一种由简入繁的探索过程,即首先分析只有一个对象添加入信息系统后计算相应的增量式更新,然后逐步分析多个对象添加时计算相应的增量式更新. 在本节,我们也采用同样的研究方法.

定理3 设 t 时刻的混合型决策信息系统为 $MDIS^{(t)}=(U^{(t)},C\cup D)$,$U^{(t)}=\{x_1,x_2,\cdots,x_n\}$,属性集 $A\subseteq AT$ 在论域 $U^{(t)}$ 上诱导的邻域覆盖为 $\{\delta_A^{U^{(t)}}(x_1),\delta_A^{U^{(t)}}(x_2),\cdots,\delta_A^{U^{(t)}}(x_n)\}$,其中 $\delta_A^{U^{(t)}}(x_i)$ 表示对象 $x_i\in U^{(t)}$ 在论域 $U^{(t)}$ 上的邻域类,设决策属性 D 关于属性集 A 在论域 $U^{(t)}$ 下的邻域区分度为 $NDD_A^{U^{(t)}}(D)$. $t+1$ 时刻混合型信息系统增加了对象集 $\Delta U=\{x_{n+1}\}$,那么 $t+1$ 时刻的混合型决策信息系统表示为 $MDIS^{(t+1)}=(U^{(t+1)},C\cup D)$,其中 $U^{(t+1)}=U^{(t)}\cup\Delta U$. 同时 A 在论域 $U^{(t+1)}$ 上诱导的邻域覆盖增量更新为 $\{\delta_A^{U^{(t+1)}}(x_1),\delta_A^{U^{(t+1)}}(x_2),\cdots,\delta_A^{U^{(t+1)}}(x_{n+1})\}$. 那么决策属性 D 关于属性集 A 在论域 $U^{(t+1)}$ 下的邻域区分度更新为

$$\begin{aligned}NDD_A^{U^{(t+1)}}(D)=&NDD_A^{U^{(t)}}(D)\\&+2\cdot\left|\delta_A^{U^{(t+1)}}(x_{n+1})-[x_{n+1}]_D^{U^{(t+1)}}\right|\end{aligned}\tag{15}$$

证明 根据邻域区分度的定义可以得到

$$NDD_A^{U^{(t+1)}}(D)=\sum_{i=1}^{n+1}\left|\delta_A^{U^{(t+1)}}(x_i)\right|-\sum_{i=1}^{n+1}\left|\delta_A^{U^{(t+1)}}(x_i)\cap[x_i]_D^{U^{(t+1)}}\right|$$

$t+1$ 时刻混合型信息系统增加了对象集 $\Delta U=\{x_{n+1}\}$,此时 $U^{(t+1)}=U^{(t)}\cup\Delta U$. 设对象 x_{n+1} 在论域 $U^{(t+1)}$ 下的邻域类为 $\delta_A^{U^{(t+1)}}(x_{n+1})$,根据引理1和定理2,我们将整个论域 $U^{(t)}$ 中的对象分成两个部分:

(1)记 $\alpha=U^{(t+1)}-\delta_A^{U^{(t+1)}}(x_{n+1})$,此时 $\alpha\subseteq U^{(t)}$. 对于 $\forall x\in\alpha$,都有 $\delta_A^{U^{(t+1)}}(x)=\delta_A^{U^{(t)}}(x)$,因此也有 $\delta_A^{U^{(t+1)}}(x)\cap[x]_D^{U^{(t+1)}}=\delta_A^{U^{(t)}}(x)\cap[x]_D^{U^{(t+1)}}$,即 $\delta_A^{U^{(t+1)}}(x)\cap[x]_D^{U^{(t+1)}}=\delta_A^{U^{(t)}}(x)\cap[x]_D^{U^{(t)}}$.

(2)记 $\beta\in\delta_A^{U^{(t+1)}}(x_{n+1})-\{x_{n+1}\}$,同样 $\beta\subseteq U^{(t)}$. 对于 $\forall x\in\beta$,满足 $\delta_A^{U^{(t+1)}}(x)=\delta_A^{U_{(t)}}(x)\cup\{x_{n+1}\}$. 此时

$$\begin{aligned}&\delta_A^{U^{(t+1)}}(x)\cap[x]_D^{U^{(t+1)}}\\&=(\delta_A^{U^{(t)}}(x)\cup\{x_{n+1}\})\cap[x]_D^{U^{(t+1)}}\\&=(\delta_A^{U^{(t)}}(x)\cap[x]_D^{U^{(t+1)}})\cup(\{x_{n+1}\}\cap[x]_D^{U^{(t+1)}})\end{aligned}$$

由于 $\delta_A^{U^{(t)}}(x)\cap[x]_D^{U^{(t+1)}}=\delta_A^{U^{(t)}}(x)\cap[x]_D^{U^{(t)}}$,并且 $(\delta_A^{U^{(t)}}(x)\cap[x]_D^{U^{(t)}})\cap(\{x_{n+1}\}\cap[x]_D^{U^{(t+1)}})=\varnothing$,所以

$$|(\delta_A^{U^{(t)}}(x)\cap[x]_D^{U^{(t)}})\cup(\{x_{n+1}\}\cap[x]_D^{U^{(t+1)}})|$$
$$=|\delta_A^{U^{(t)}}(x)\cap[x]_D^{U^{(t)}}|+|\{x_{n+1}\}\cap[x]_D^{U^{(t+1)}}|$$

那么可以得到

$$|\delta_A^{U^{(t+1)}}(x)\cap[x]_D^{U^{(t+1)}}|$$
$$=|\delta_A^{U^{(t)}}(x)\cap[x]_D^{U^{(t)}}|+|\{x_{n+1}\}\cap[x]_D^{U^{(t+1)}}|$$

由于 $\alpha\cup\beta=U^{(t)}$ 且 $\alpha\cap\beta=\varnothing$,我们将 $NDD_A^{U^{(t+1)}}(D)$进行展开

$$\begin{aligned}
&NDD_A^{U^{(t+1)}}(D)\\
&=\sum_{x\in\alpha}|\delta_A^{U^{(t+1)}}(x)|-\sum_{x\in\alpha}|\delta_A^{U^{(t+1)}}(x)\cap[x]_D^{U^{(t+1)}}|\\
&\quad+\sum_{x\in\beta}|\delta_A^{U^{(t+1)}}(x)|-\sum_{x\in\beta}|\delta_A^{U^{(t+1)}}(x)\cap[x]_D^{U^{(t+1)}}|\\
&\quad+|\delta_A^{U^{(t+1)}}(x_{n+1})|-|\delta_A^{U^{(t+1)}}(x_{n+1})\cap[x_{n+1}]_D^{U^{(t+1)}}|\\
&=\sum_{x\in\alpha}|\delta_A^{U^{(t)}}(x)|-\sum_{x\in\alpha}|\delta_A^{U^{(t)}}(x)\cap[x]_D^{U^{(t)}}|\\
&\quad+\sum_{x\in\beta}|\delta_A^{U^{(t)}}(x)\cup\{x_{n+1}\}|\\
&\quad-\sum_{x\in\beta}(|\delta_A^{U^{(t)}}(x)\cap[x]_D^{U^{(t)}}|+|\{x_{n+1}\}\cap[x]_D^{U^{(t+1)}}|)\\
&\quad+|\delta_A^{U^{(t+1)}}(x_{n+1})|-|\delta_A^{U^{(t+1)}}(x_{n+1})\cap[x_{n+1}]_D^{U^{(t+1)}}|
\end{aligned}$$

又由于 $|\delta_A^{U^{(t)}}(x)\cup\{x_{n+1}\}|=|\delta_A^{U^{(t)}}(x)|+1$,所以 $\sum_{x\in\beta}|\delta_A^{U^{(t)}}(x)\cup\{x_{n+1}\}|=\sum_{x\in\beta}|\delta_A^{U^{(t)}}(x)|+|\beta|$,则

$$\begin{aligned}
&NDD_A^{U^{(t+1)}}(D)\\
&=\sum_{i=1}^{n}|\delta_A^{U^{(t)}}(x)|-\sum_{i=1}^{n}|\delta_A^{U^{(t)}}(x)\cap[x]_D^{U^{(t)}}|\\
&\quad+|\beta|-\sum_{x\in\beta}|\{x_{n+1}\}\cap[x]_D^{U^{(t+1)}}|+|\delta_A^{U^{(t+1)}}(x_{n+1})|\\
&\quad-|\delta_A^{U^{(t+1)}}(x_{n+1})\cap[x_{n+1}]_D^{U^{(t+1)}}|\\
&=NDD_A^{U^{(t)}}(D)+|\beta|-\sum_{x\in\beta}|\{x_{n+1}\}\cap[x]_D^{U^{(t+1)}}|\\
&\quad+|\delta_A^{U^{(t+1)}}(x_{n+1})|-|\delta_A^{U^{(t+1)}}(x_{n+1})\cap[x_{n+1}]_D^{U^{(t+1)}}|
\end{aligned}$$

对于 $\forall x\in\beta$,若 $x\in[x_{n+1}]_D^{U^{(t+1)}}$,则 $x_{n+1}\in[x]_D^{U^{(t+1)}}$,那么 $|\{x_{n+1}\}\cap[x]_D^{U^{(t+1)}}|=1$. 所以可得

$$\sum_{x\in\beta}|\{x_{n+1}\}\cap[x]_D^{U^{(t+1)}}|=|\beta\cap[x_{n+1}]_D^{U^{(t+1)}}|$$

因此

$$\begin{aligned}
&NDD_A^{U^{(t+1)}}(D)\\
&=NDD_A^{U^{(t)}}(D)+|\beta|-|\beta\cap[x_{n+1}]_D^{U^{(t+1)}}|+|\delta_A^{U^{(t+1)}}(x_{n+1})|\\
&\quad-|\delta_A^{U^{(t+1)}}(x_{n+1})\cap[x_{n+1}]_D^{U^{(t+1)}}|\\
&=NDD_A^{U^{(t)}}(D)+|\delta_A^{U^{(t+1)}}(x_{n+1})-\{x_{n+1}\}|\\
&\quad-|(\delta_A^{U^{(t+1)}}(x_{n+1})-\{x_{n+1}\})\cap[x_{n+1}]_D^{U^{(t+1)}}|\\
&\quad+|\delta_A^{U^{(t+1)}}(x_{n+1})|-|\delta_A^{U^{(t+1)}}(x_{n+1})\cap[x_{n+1}]_D^{U^{(t+1)}}|\\
&=NDD_A^{U^{(t)}}(D)+|\delta_A^{U^{(t+1)}}(x_{n+1})|-1\\
&\quad-|\delta_A^{U^{(t+1)}}(x_{n+1})\cap[x_{n+1}]_D^{U^{(t+1)}}|+1\\
&\quad+|\delta_A^{U^{(t+1)}}(x_{n+1})|-|\delta_A^{U^{(t+1)}}(x_{n+1})\cap[x_{n+1}]_D^{U^{(t+1)}}|\\
&=NDD_A^{U^{(t)}}(D)+2\cdot(|\delta_A^{U^{(t+1)}}(x_{n+1})|\\
&\quad-|\delta_A^{U^{(t+1)}}(x_{n+1})\cap[x_{n+1}]_D^{U^{(t+1)}}|)\\
&=NDD_A^{U^{(t)}}(D)+2\cdot|\delta_A^{U^{(t+1)}}(x_{n+1})-[x_{n+1}]_D^{U^{(t+1)}}|
\end{aligned}$$

因此定理 3 成立.　　证毕

定理3 表明了混合型信息系统论域增加一个对象后,只需要计算出新加入对象的邻域类,便可以增量的计算出新混合型信息系统下的邻域区分度,而不需要按照邻域区分度的定义计算每个对象的邻域类,因此这种增量式的计算方法大大的提高了计算效率. 根据定理 3 的增量式更新关系,我们可以逐步推导出当混合型信息系统论域增加多个对象时邻域区分度的增量更新.

定理 4　设 t 时刻的混合型决策信息系统为 $MDIS^{(t)}=(U^{(t)},C\cup D)$,$U^{(t)}=\{x_1,x_2,\cdots,x_n\}$,属性集 $A\subseteq AT$ 在论域 $U^{(t)}$ 上诱导的邻域覆盖为 $\{\delta_A^{U^{(t)}}(x_1),\delta_A^{U^{(t)}}(x_2),\cdots,\delta_A^{U^{(t)}}(x_n)\}$,其中 $\delta_A^{U^{(t)}}(x_i)$ 表示对象 $x_i\in U^{(t)}$ 在论域 $U^{(t)}$ 上的邻域类,设决策属性 D 关于 A 在论域 $U^{(t)}$ 下的邻域区分度为 $NDD_A^{U^{(t)}}(D)$. 在 $t+1$ 时刻混合型信息系统增加了对象集 $\Delta U=\{x_{n+1},x_{n+2},\cdots,x_{n+c}\}$,则 $t+1$ 时刻的混合型决策信息系统表示为 $MDIS^{(t+1)}=(U^{(t+1)},C\cup D)$,其中 $U^{(t+1)}=U^{(t)}\cup\Delta U$. 同时 A 在论域 $U^{(t+1)}$ 上诱导的邻域覆盖增量式更新为 $\{\delta_A^{U^{(t+1)}}(x_1),\delta_A^{U^{(t+1)}}(x_2),\cdots,\delta_A^{U^{(t+1)}}(x_{n+c})\}$. 那么决策属性 D 关于 A 在论域 $U^{(t+1)}$ 下的邻域区分度更新为

$$\begin{aligned}
&NDD_A^{U^{(t+1)}}(D)\\
&=NDD_A^{U^{(t)}}(D)+2\cdot|\delta_A^{U_1^{(t)}}(x_{n+1})-[x_{n+1}]_D^{U_1^{(t)}}|\\
&\quad+2\cdot|\delta_A^{U_2^{(t)}}(x_{n+1})-[x_{n+1}]_D^{U_2^{(t)}}|+\cdots\\
&\quad+2\cdot|\delta_A^{U_c^{(t)}}(x_{n+1})-[x_{n+1}]_D^{U_c^{(t)}}|
\end{aligned}\tag{16}$$

这里的

$U_1^{(t)}=U^{(t)}\cup\{x_{n+1}\}$,$U_2^{(t)}=U^{(t)}\cup\{x_{n+1},x_{n+2}\}$,$\cdots$,$U_c^{(t)}=U^{(t)}\cup\{x_{n+1},x_{n+2},\cdots,x_{n+c}\}$,

即 $U_c^{(t)}=U^{(t+1)}$.

证明　根据定理 3 可以得到

$$NDD_A^{U_1^{(t)}}(D)=NDD_A^{U^{(t)}}(D)+2\cdot|\delta_A^{U_1^{(t)}}(x_{n+1})-[x_{n+1}]_D^{U_1^{(t)}}|$$

那么进行递推有

$$NDD_A^{U_2^{(t)}}(D)=NDD_A^{U_1^{(t)}}(D)+2\cdot|\delta_A^{U_2^{(t)}}(x_{n+1})-[x_{n+1}]_D^{U_2^{(t)}}|$$

即

$$\begin{aligned}
&NDD_A^{U_2^{(t)}}(D)\\
&=NDD_A^{U^{(t)}}(D)+2\cdot|\delta_A^{U_1^{(t)}}(x_{n+1})-[x_{n+1}]_D^{U_1^{(t)}}|
\end{aligned}$$

$+2\cdot|\delta_A^{U_2^{(t)}}(x_{n+1})-[x_{n+1}]_D^{U_2^{(t)}}|$

那么对 $\Delta U=\{x_{n+1},x_{n+2},\cdots,x_{n+c}\}$ 中的每个对象进行递推就可以得到定理4的结果. 证毕

对于动态环境下的混合型信息系统,当论域逐渐增加时,在原有邻域区分度的基础上,我们只需要对新加入的每个对象进行邻域类的相关计算,便可以快速的更新出新信息系统下的邻域区分度,因此这种增量式的计算方法提高了动态数据环境下的计算效率.

3.3 混合数据对象减少时的区分度增量式更新

在3.2节中,我们给出了混合型信息系统论域增加时邻域区分度的增量式更新,类似于3.2节的研究思路,在本小节我们将探究混合型信息系统论域减小时邻域区分度的增量式更新.

引理2 设 t 时刻的混合型信息系统为 $MIS^{(t)}=(U^{(t)},AT)$,$U^{(t)}=\{x_1,x_2,\cdots,x_n\}$,属性集 $A\subseteq AT$ 在论域 $U^{(t)}$ 上确定的邻域关系为 $N_\delta^{U^{(t)}}$,在 $t+1$ 时刻混合型信息系统减少了对象集 ΔU,这里 $\Delta U\subseteq U^{(t)}$,那么 $t+1$ 时刻混合型信息系统可表示为 $MIS^{(t+1)}=(U^{(t+1)},AT)$,其中 $U^{(t+1)}=U^{(t)}-\Delta U$. A 在论域 $U^{(t+1)}$ 上确定的邻域关系为 $N_\delta^{U^{(t+1)}}$,那么满足 $N_\delta^{U^{(t+1)}}\subseteq N_\delta^{U^{(t)}}$.

证明 证明过程类似于引理1. 证毕

定理5 设 t 时刻的混合型信息系统为 $MIS^{(t)}=(U^{(t)},AT)$,$U^{(t)}=\{x_1,x_2,\cdots,x_n\}$,属性集 $A\subseteq AT$ 在论域 $U^{(t)}$ 上诱导的邻域覆盖为 $\{\delta_A^{U^{(t)}}(x_1),\delta_A^{U^{(t)}}(x_2),\cdots,\delta_A^{U^{(t)}}(x_n)\}$,其中 $\delta_A^{U^{(t)}}(x_i)$ 表示对象 $x_i\in U^{(t)}$ 在论域 $U^{(t)}$ 上的邻域类. 在 $t+1$ 时刻混合型信息系统减少了对象集 $\Delta U=\{x_{p1},x_{p2},\cdots,x_{pc}\}$,这里的 $\Delta U\subseteq U^{(t)}$,那么 $t+1$ 时刻的混合型信息系统可表示为 $MIS^{(t+1)}=(U^{(t+1)},AT)$,其中 $U^{(t+1)}=U^{(t)}-\Delta U$. 则属性集 A 在论域 $U^{(t+1)}$ 上诱导的邻域覆盖增量更新为 $\{\delta_A^{U^{(t+1)}}(x)\mid\forall x\in U^{(t+1)}\}$,其中 $\delta_A^{U^{(t+1)}}(x)=\delta_A^{U^{(t)}}(x)-\Delta U,\forall x\in U^{(t+1)}$.

证明 混合型信息系统论域减小时,邻域类的增量更新要易于论域增加时邻域类的增量更新. 根据引理2,我们直接可以得出 $\delta_A^{U^{(t+1)}}(x)=\delta_A^{U^{(t)}}(x)-\Delta U,\forall x\in U^{(t+1)}$. 证毕

接下来我们同样按照3.2研究方法,即首先分析混合型信息系统只减少一个对象时邻域区分度的增量式更新,然后逐步分析减少多个对象时邻域区分度的增量式更新.

定理6 设 t 时刻的混合型决策信息系统为 $MDIS^{(t)}=(U^{(t)},C\cup D)$,$U^{(t)}=\{x_1,x_2,\cdots,x_n\}$,属性集 $A\subseteq AT$ 在论域 $U^{(t)}$ 上诱导的邻域覆盖为 $\{\delta_A^{U^{(t)}}(x_1),\delta_A^{U^{(t)}}(x_2),\cdots,\delta_A^{U^{(t)}}(x_n)\}$,其中 $\delta_A^{U^{(t)}}(x_i)$ 表示对象 $x_i\in U^{(t)}$ 在论域 $U^{(t)}$ 上的邻域类,设决策属性 D 关于属性集 A 在论域 $U^{(t)}$ 下的邻域区分度为 $NDD_A^{U^{(t)}}(D)$. 在 $t+1$ 时刻混合型信息系统减少了对象集 $\Delta U=\{x_{p1}\}$,$x_{p1}\in U^{(t)}$,那么 $t+1$ 时刻的混合型决策信息系统表示为 $MDIS^{(t+1)}=(U^{(t+1)},C\cup D)$,其中 $U^{(t+1)}=U^{(t)}-\Delta U$. 同时属性集 A 在论域 $U^{(t+1)}$ 上诱导的邻域覆盖增量更新为 $\{\delta_A^{U^{(t+1)}}(x)\mid\forall x\in U^{(t+1)}\}$. 那么决策属性 D 关于属性集 A 在论域 $U^{(t+1)}$ 下的邻域区分度更新为

$$NDD_A^{U^{(t+1)}}(D)=NDD_A^{U^{(t)}}(D)-2\cdot|\delta_A^{U^{(t)}}(x_{p1})-[x_{p1}]_D^{U^{(t)}}|\quad(17)$$

证明 类似于定理3的证明方法,根据引理2,我们同样将整个论域 $U^{(t+1)}$ 中的对象分成两个部分:

(1)记 $\alpha=U^{(t)}-\delta_A^{U^{(t)}}(x_{p1})$,此时 $\alpha\subseteq U^{(t+1)}$. 对于 $\forall x\in\alpha$,都有 $\delta_A^{U^{(t+1)}}(x)=\delta_A^{U^{(t)}}(x)$,因此也有 $\delta_A^{U^{(t+1)}}(x)\cap[x]_D^{U^{(t+1)}}=\delta_A^{U^{(t)}}(x)\cap[x]_D^{U^{(t+1)}}$,即 $\delta_A^{U^{(t+1)}}(x)\cap[x]_D^{U^{(t+1)}}=\delta_A^{U^{(t)}}(x)\cap[x]_D^{U^{(t)}}$.

(2)记 $\beta\in\delta_A^{U^{(t)}}(x_{p1})-\{x_{p1}\}$,同样 $\beta\subseteq U^{(t+1)}$. 对于 $\forall x\in\beta$,满足 $\delta_A^{U^{(t+1)}}(x)=\delta_A^{U^{(t)}}(x)-\{x_{p1}\}$. 此时

$$\begin{aligned}&\delta_A^{U^{(t+1)}}(x)\cap[x]_D^{U^{(t+1)}}\\&=(\delta_A^{U^{(t)}}(x)-\{x_{p1}\})\cap[x]_D^{U^{(t+1)}}\\&=(\delta_A^{U^{(t)}}(x)\cap[x]_D^{U^{(t+1)}})-(\{x_{p1}\}\cap[x]_D^{U^{(t+1)}})\end{aligned}$$

由于 $\{x_{p1}\}\cap[x]_D^{U^{(t+1)}}=\varnothing$,所以 $\delta_A^{U^{(t+1)}}(x)\cap[x]_D^{U^{(t+1)}}=\delta_A^{U^{(t)}}(x)\cap[x]_D^{U^{(t+1)}}$. 当 $x\in[x_{p1}]_D^{U^{(t)}}$ 时,即 $x_{p1}\in[x]_D^{U^{(t)}}$,那么 $[x]_D^{U^{(t+1)}}=[x]_D^{U^{(t)}}-\{x_{p1}\}$,当 $x\notin[x_{p1}]_D^{U^{(t)}}$ 时,$[x]_D^{U^{(t+1)}}=[x]_D^{U^{(t)}}$,因此综合起来可以得到,当 $x\in\beta\cap[x_{p1}]_D^{U^{(t)}}$ 时,那么有 $\delta_A^{U^{(t+1)}}(x)\cap[x]_D^{U^{(t+1)}}=\delta_A^{U^{(t)}}(x)\cap([x]_D^{U^{(t)}}-\{x_{p1}\})$,由于

$$\begin{aligned}&\delta_A^{U^{(t)}}(x)\cap([x]_D^{U^{(t)}}-\{x_{p1}\})\\&=(\delta_A^{U^{(t)}}(x)\cap[x]_D^{U^{(t)}})-(\delta_A^{U^{(t)}}(x)\cap\{x_{p1}\})\\&=(\delta_A^{U^{(t)}}(x)\cap[x]_D^{U^{(t)}})-\{x_{p1}\}\end{aligned}$$

所以 $\delta_A^{U^{(t+1)}}(x)\cap[x]_D^{U^{(t+1)}}=(\delta_A^{U^{(t)}}(x)\cap[x]_D^{U^{(t)}})-\{x_{p1}\}$,当 $x\in\beta-[x_{p1}]_D^{U^{(t)}}$ 时,满足 $\delta_A^{U^{(t+1)}}(x)\cap[x]_D^{U^{(t+1)}}=\delta_A^{U^{(t)}}(x)\cap[x]_D^{U^{(t)}}$.

由于 $\alpha\cup\beta=U^{(t+1)}$ 且 $\alpha\cap\beta=\varnothing$,我们将邻域区分度进行展开可以得到

$$\begin{aligned}&NDD_A^{U^{(t+1)}}(D)\\&=\sum_{\forall x\in U^{(t+1)}}|\delta_A^{U^{(t+1)}}(x)|-\sum_{\forall x\in U^{(t+1)}}|\delta_A^{U^{(t+1)}}(x)\cap[x]_D^{U^{(t+1)}}|\\&=\sum_{\forall x\in\alpha}|\delta_A^{U^{(t+1)}}(x)|-\sum_{\forall x\in\alpha}|\delta_A^{U^{(t+1)}}(x)\cap[x]_D^{U^{(t+1)}}|\end{aligned}$$

$$+\sum_{\forall x\in\beta}|\delta_A^{U^{(t+1)}}(x)|-\sum_{\forall x\in\beta}|\delta_A^{U^{(t+1)}}(x)\cap[x]_D^{U^{(t+1)}}|$$
$$=\sum_{\forall x\in\alpha}|\delta_A^{U^{(t+1)}}(x)|-\sum_{\forall x\in\alpha}|\delta_A^{U^{(t+1)}}(x)\cap[x]_D^{U^{(t+1)}}|$$
$$+\sum_{\forall x\in\beta\cap[x_{p1}]_D^{U^{(t)}}}|\delta_A^{U^{(t+1)}}(x)|-\sum_{\forall x\in\beta\cap[x_{p1}]_D^{U^{(t)}}}|\delta_A^{U^{(t+1)}}(x)\cap[x]_D^{U^{(t+1)}}|$$
$$+\sum_{\forall x\in\beta-[x_{p1}]_D^{U^{(t)}}}|\delta_A^{U^{(t+1)}}(x)|-\sum_{\forall x\in\beta-[x_{p1}]_D^{U^{(t)}}}|\delta_A^{U^{(t+1)}}(x)\cap[x]_D^{U^{(t+1)}}|$$
$$=\sum_{\forall x\in\alpha}|\delta_A^{U^{(t)}}(x)|-\sum_{\forall x\in\alpha}|\delta_A^{U^{(t)}}(x)\cap[x]_D^{U^{(t)}}|$$
$$+\sum_{\forall x\in\beta\cap[x_{p1}]_D^{U^{(t)}}}|\delta_A^{U^{(t)}}(x)-\{x_{p1}\}|$$
$$-\sum_{\forall x\in\beta\cap[x_{p1}]_D^{U^{(t)}}}|(\delta_A^{U^{(t)}}(x)\cap[x]_D^{U^{(t)}})-\{x_{p1}\}|$$
$$+\sum_{\forall x\in\beta-[x_{p1}]_D^{U^{(t)}}}|\delta_A^{U^{(t)}}(x)-\{x_{p1}\}|$$
$$-\sum_{\forall x\in\beta-[x_{p1}]_D^{U^{(t)}}}|\delta_A^{U^{(t)}}(x)\cap[x]_D^{U^{(t)}}|$$

由于

$$|\delta_A^{U^{(t)}}-\{x_{p1}\}|=|\delta_A^{U^{(t)}}|-1$$
$$|(\delta_A^{U^{(t)}}(x)\cap[x]_D^{U^{(t)}})-\{x_{p1}\}|=|\delta_A^{U^{(t)}}(x)\cap[x]_D^{U^{(t)}}|-1$$

那么

$$NDD_A^{U^{(t+1)}}(D)$$
$$=\sum_{\forall x\in\alpha}|\delta_A^{U^{(t)}}(x)|-\sum_{\forall x\in\alpha}|\delta_A^{U^{(t)}}(x)\cap[x]_D^{U^{(t)}}|$$
$$+\sum_{\forall x\in\beta\cap[x_{p1}]_D^{U^{(t)}}}|\delta_A^{U^{(t)}}(x)|-|\beta\cap[x_{p1}]_D^{U^{(t)}}|$$
$$-\sum_{\forall x\in\beta\cap[x_{p1}]_D^{U^{(t)}}}|(\delta_A^{U^{(t)}}(x)\cap[x]_D^{U^{(t)}})|+|\beta\cap[x_{p1}]_D^{U^{(t)}}|$$
$$+\sum_{\forall x\in\beta-[x_{p1}]_D^{U^{(t)}}}|\delta_A^{U^{(t)}}(x)|-|\beta-[x_{p1}]_D^{U^{(t)}}|$$
$$-\sum_{\forall x\in\beta-[x_{p1}]_D^{U^{(t)}}}|\delta_A^{U^{(t)}}(x)\cap[x]_D^{U^{(t)}}|$$

其中

$$\sum_{\forall x\in\alpha}|\delta_A^{U^{(t)}}(x)|-\sum_{\forall x\in\alpha}|\delta_A^{U^{(t)}}(x)\cap[x]_D^{U^{(t)}}|$$
$$+\sum_{\forall x\in\beta\cap[x_{p1}]_D^{U^{(t)}}}|\delta_A^{U^{(t)}}(x)|-\sum_{\forall x\in\beta\cap[x_{p1}]_D^{U^{(t)}}}|(\delta_A^{U^{(t)}}(x)\cap[x]_D^{U^{(t)}})|$$
$$+\sum_{\forall x\in\beta-[x_{p1}]_D^{U^{(t)}}}|\delta_A^{U^{(t)}}(x)|-\sum_{\forall x\in\beta-[x_{p1}]_D^{U^{(t)}}}|\delta_A^{U^{(t)}}(x)\cap[x]_D^{U^{(t)}}|$$
$$=NDD_A^{U^{(t)}}(D)-|\delta_A^{U^{(t)}}(x_{p1})|+|\delta_A^{U^{(t)}}(x_{p1})\cap[x_{p1}]_D^{U^{(t)}}|$$

所以

$$NDD_A^{U^{(t+1)}}(D)$$
$$=NDD_A^{U^{(t)}}(D)-(|\delta_A^{U^{(t)}}(x_{p1})|$$
$$-|\delta_A^{U^{(t)}}(x_{p1})\cap[x_{p1}]_D^{U^{(t)}}|)-|\beta-[x_{p1}]_D^{U^{(t)}}|$$

由于

$$|\delta_A^{U^{(t)}}(x_{p1})|-|\delta_A^{U^{(t)}}(x_{p1})\cap[x_{p1}]_D^{U^{(t)}}|$$
$$=|\delta_A^{U^{(t)}}(x_{p1})-[x_{p1}]_D^{U^{(t)}}|$$

同时 $\beta=\delta_A^{U^{(t)}}(x_{p1})-\{x_{p1}\}$,那么

$$\beta-[x_{p1}]_D^{U^{(t)}}=\delta_A^{U^{(t)}}(x_{p1})-\{x_{p1}\}-[x_{p1}]_D^{U^{(t)}}$$
$$=\delta_A^{U^{(t)}}(x_{p1})-[x_{p1}]_D^{U^{(t)}}$$

因此

$$NDD_A^{U^{(t+1)}}(D)=NDD_A^{U^{(t)}}(D)-2\cdot|\delta_A^{U^{(t)}}(x_{p1})-[x_{p1}]_D^{U^{(t)}}|$$

证毕

定理6 同样可以推广到混合型信息系统论域减少多个对象时的情形.

定理 7　设 t 时刻的混合型决策信息系统为 $MDIS^{(t)}=(U^{(t)},C\cup D)$,$U^{(t)}=\{x_1,x_2,\cdots,x_n\}$,属性集 $A\subseteq AT$ 在论域 $U^{(t)}$ 上诱导的邻域覆盖为 $\{\delta_A^{U^{(t)}}(x_1),\delta_A^{U^{(t)}}(x_2),\cdots,\delta_A^{U^{(t)}}(x_n)\}$,其中 $\delta_A^{U^{(t)}}(x_i)$ 表示对象 $x_i\in U^{(t)}$ 在论域 $U^{(t)}$ 上的邻域类,设决策属性 D 关于 A 在论域 $U^{(t)}$ 下的邻域区分度为 $NDD_A^{U^{(t)}}(D)$. 在 $t+1$ 时刻混合型信息系统减少了对象集 $\Delta U=\{x_{p1},x_{p2},\cdots,x_{pc}\}$,$\Delta U\subseteq U^{(t)}$,那么 $t+1$ 时刻的混合型决策信息系统表示为 $MDIS^{(t+1)}=(U^{(t+1)},C\cup D)$,其中 $U^{(t+1)}=U^{(t)}-\Delta U$. 同时属性集 A 在论域 $U^{(t+1)}$ 上诱导的邻域覆盖增量更新为 $\{\delta_A^{U^{(t+1)}}(x)\mid\forall x\in U^{(t+1)}\}$. 那么决策属性 D 关于 A 在论域 $U^{(t+1)}$ 下的邻域区分度更新为

$$NDD_A^{U^{(t+1)}}(D)$$
$$=NDD_A^{U^{(t)}}(D)-2\cdot|\delta_A^{U_1^{(t)}}(x_{p1})-[x_{p1}]_D^{U_1^{(t)}}|$$
$$-2\cdot|\delta_A^{U_2^{(t)}}(x_{p2})-[x_{p2}]_D^{U_2^{(t)}}|$$
$$-\cdots-2\cdot|\delta_A^{U_c^{(t)}}(x_{p2})-[x_{p2}]_D^{U_c^{(t)}}|\qquad(18)$$

这里的

$$U_1^{(t)}=U^{(t)},U_2^{(t)}=U^{(t)}-\{x_{n+1}\},\cdots,$$
$$U_c^{(t)}=U^{(t)}-\{x_{n+1},x_{n+2},\cdots,x_{n+c-1}\}$$

证明　根据定理6,类似于定理4的证明方法可证明定理7成立.　　证毕

定理7表明,对于动态环境下的混合型信息系统,当论域逐渐减小时,在原有邻域区分度的基础上,我们同样只需要对减少的每个对象进行邻域类的相关计算,便可以快速的更新的新信息系统下的邻域区分度,因此基于这种增量式的计算方法同样具有很高的计算效率.

4　非增量式与增量式属性约简算法

在本节,我们将首先给出混合型信息系统下邻域区分度的非增量式属性约简算法,然后根据3.2节和3.3节中邻域区分度的增量式更新方法,进一步得提出混合型信息系统下邻域区分度的增量式属性约简算法. 首先我们给出基于邻域区分度属性约简的定义,具体如定义8所示.

定义 8　对于混合型决策信息系统 $MDIS=(U,C\cup D)$,若属性集 $R\subseteq C$ 是该信息系统邻域区分度的属性约

简,那么当且仅当

$$(1)\quad NDD_C(D)=NDD_R(D) \tag{19}$$

$$(2)\quad \forall a\in R, NDD_C(D)<NDD_{R-\{a\}}(D) \tag{20}$$

根据定义 8 中关于属性约简的定义以及算法 1 的区分度属性约简算法,接下来便可以构造出基于邻域区分度的非增量式属性约简算法,具体如算法 2 所示.

算法 2 基于邻域区分度的非增量式属性约简算法(NARNDD, Non-incremental Attribute Reduction based on Neighborhood Discernibility Degree)

输入:$t+1$ 时刻的混合型决策信息系统 $MDIS^{(t+1)}=(U^{(t+1)},C\cup D)$,邻域半径 δ,给定距离函数;

输出:$t+1$ 时刻的属性约简集 $R^{(t+1)}$.

Step1:初始化 $R^{(t+1)}=\varnothing$;

Step2:对于 $\forall a\in C-R^{(t+1)}$,计算属性 a 的属性重要度 $sig_{R^{(t+1)}}(a)$,其中 $sig_{R^{(t+1)}}(a)=NDD_{R^{(t+1)}}(D)-NDD_{R^{(t+1)}\cup\{a\}}(D)$;

Step3:找出 $C-R^{(t+1)}$ 中属性重要度最大的属性,记为 a';

Step4:判断 $sig_{R^{(t+1)}}(a')$ 的值,若 $sig_{R^{(t+1)}}(a')>0$,那么 $R^{(t+1)}=R^{(t+1)}\cup\{a'\}$,并转入 Step2,若 $sig_{R^{(t+1)}}(a')=0$,则转入 Step5;

Step5:返回属性约简 $R^{(t+1)}$.

算法 2 的整体结构与算法 1 类似,都是通过启发式函数来进行搜索属性,算法 1 采用区分度进行属性搜索,只适用于符号型的信息系统,而算法 2 运用本文提出的邻域区分度来搜索属性,可以运用于数值型以及混合型的信息系统,通过性质 1 可以进一步证明算法 2 是算法 1 的推广.

另一方面,当从 t 时刻至 $t+1$ 时刻,混合型信息系统的论域发生了变化,而算法 2 仍然从空集开始逐渐搜索属性,对原先 t 时刻的约简结果没有加以利用,因此算法 2 是一种非增量式的计算方法,在动态数据下的属性约简效率比较低. 对于混合型决策信息系统 $MDIS^{(t+1)}=(U^{(t+1)},C\cup D)$,设 $|U^{(t+1)}|=n$,$|C|=c$,$|R^{(t+1)}|=r^{(t+1)}$,那么算法 2 的时间复杂度为

$$O(c\cdot n^2+(c-1)\cdot n^2+\cdots+(c-r^{(t+1)}+1)\cdot n^2)$$
$$=O(r^{(t+1)}\cdot(2c-r^{(t+1)})\cdot n^2)$$

根据 2.2 节和 2.3 节的邻域区分度增量式更新,类似于文献[14]中关于对象增加和减少时增量式属性约简算法的构造,接下来提出基于邻域区分度的增量式属性约简算法,具体如算法 3 和算法 4 所示.

算法 3 对象增加时邻域区分度的增量式属性约简算法(IARNDD-OD, Incremental Attribute Reduction based on Neighborhood Discernibility Degree when Objects are Added)

输入:$t+1$ 时刻的混合型决策信息系统 $MDIS^{(t+1)}=(U^{(t+1)}=U^{(t)}\cup\Delta U,C\cup D)$,邻域半径 δ,给定距离函数,t 时刻的混合型决策信息系统 $MDIS^{(t)}$ 的约简集 $R^{(t)}$,邻域区分度 $NDD_{R^{(t)}}^{U^{(t)}}(D)$.

输出:$t+1$ 时刻的属性约简集 $R^{(t+1)}$.

Step1:根据 $NDD_{R^{(t)}}^{U^{(t)}}(D)$ 按照定理 4 增量式计算 $NDD_{R^{(t)}}^{U^{(t+1)}}(D)$ 和 $NDD_C^{U^{(t+1)}}(D)$;

Step2:判断 $NDD_{R^{(t)}}^{U^{(t+1)}}(D)$ 与 $NDD_C^{U^{(t+1)}}(D)$ 的大小关系,若 $NDD_{R^{(t)}}^{U^{(t+1)}}(D)=NDD_C^{U^{(t+1)}}(D)$,那么直接转入 Step7;若 $NDD_{R^{(t)}}^{U^{(t+1)}}(D)>NDD_C^{U^{(t+1)}}(D)$,那么转入 Step3;若 $NDD_{R^{(t)}}^{U^{(t+1)}}(D)<NDD_C^{U^{(t+1)}}(D)$,那么转入 Step5;

Step3:对于 $\forall a\in C-R^{(t)}$,计算属性 a 的属性重要度 $sig_{R^{(t)}}(a)$,其中 $sig_{R^{(t)}}(a)=NDD_{R^{(t)}}^{U^{(t+1)}}(D)-NDD_{R^{(t)}\cup\{a\}}^{U^{(t+1)}}(D)$;

Step4:找出 $C-R^{(t)}$ 中属性重要度最大的属性,记为 a',判断 $sig_{R^{(t+1)}}(a')$ 的值,若 $sig_{R^{(t)}}(a')>0$,那么 $R^{(t)}=R^{(t)}\cup\{a'\}$,并转入 Step3,若 $sig_{R^{(t)}}(a')=0$,转入 Step7;

Step5:对于 $\forall a\in R^{(t)}$,计算属性 a 的属性重要度 $sig_{R^{(t)}}(a)$,其中 $sig_{R^{(t)}}(a)=NDD_{R^{(t)}-\{a\}}^{U^{(t+1)}}(D)-NDD_{R^{(t)}}^{U^{(t+1)}}(D)$;

Step6:找出 $R^{(t)}$ 中属性重要度最大的属性,记为 a',判断 $sig_{R^{(t)}}(a')$ 的值,若 $sig_{R^{(t)}}(a')>0$,那么 $R^{(t)}=R^{(t)}-\{a'\}$,并转入 Step5,若 $sig_{R^{(t)}}(a')=0$,转入 Step7;

Step7:返回属性约简 $R^{(t+1)}=R^{(t)}$.

算法 4 对象减少时邻域区分度的增量式属性约简算法(IARNDD-OR, Incremental Attribute Reduction based on Neighborhood Discernibility Degree when Objects are Removed)

输入:$t+1$ 时刻的混合型决策信息系统 $MDIS^{(t+1)}=(U^{(t+1)}=U^{(t)}-\Delta U,C\cup D)$,邻域半径 δ,给定距离函数,t 时刻的混合型决策信息系统 $MDIS^{(t)}$ 的约简集 $R^{(t)}$,邻域区分度 $NDD_{R^{(t)}}^{U^{(t)}}(D)$.

输出:$t+1$ 时刻的属性约简集 $R^{(t+1)}$.

Step1:根据 $NDD_{R^{(t)}}^{U^{(t)}}(D)$ 按照定理 7 增量式计算 $NDD_{R^{(t)}}^{U^{(t+1)}}(D)$ 和 $NDD_C^{U^{(t+1)}}(D)$;

Step2:判断 $NDD_{R^{(t)}}^{U^{(t+1)}}(D)$ 与 $NDD_C^{U^{(t+1)}}(D)$ 的大小关系,若 $NDD_{R^{(t)}}^{U^{(t+1)}}(D)=NDD_C^{U^{(t+1)}}(D)$,那么直接转入 Step7;若 $NDD_{R^{(t)}}^{U^{(t+1)}}(D)>NDD_C^{U^{(t+1)}}(D)$,那么转入 Step3;若 $NDD_{R^{(t)}}^{U^{(t+1)}}(D)<NDD_C^{U^{(t+1)}}(D)$,那么转入 Step5;

Step3:对于 $\forall a\in C-R^{(t)}$,计算属性 a 的属性重要度 $sig_{R^{(t)}}(a)$,其中 $sig_{R^{(t)}}(a)=NDD_{R^{(t)}}^{U^{(t+1)}}(D)-NDD_{R^{(t)}\cup\{a\}}^{U^{(t+1)}}(D)$;

Step4:找出 $C-R^{(t)}$ 中属性重要度最大的属性,记为 a',判断 $sig_{R^{(t)}}(a')$ 的值,若 $sig_{R^{(t)}}(a')>0$,那么 $R^{(t)}=R^{(t)}\cup\{a'\}$,并转入 Step3,若 $sig_{R^{(t)}}(a')=0$,转入 Step7;

Step5:对于 $\forall a\in R^{(t)}$,计算属性 a 的属性重要度 $sig_{R^{(t)}}(a)$,其中 $sig_{R^{(t)}}(a)=NDD_{R^{(t)}-\{a\}}^{U^{(t+1)}}(D)-NDD_{R^{(t)}}^{U^{(t+1)}}(D)$;

Step6:找出 $R^{(t)}$ 中属性重要度最大的属性,记为 a',判断 $sig_{R^{(t)}}(a')$ 的值,若 $sig_{R^{(t)}}(a')>0$,那么 $R^{(t)}=R^{(t)}-\{a'\}$,并转入 Step5,若 $sig_{R^{(t)}}(a')=0$,转入 Step7;

Step7:返回属性约简 $R^{(t+1)}=R^{(t)}$.

当 $t+1$ 时刻混合信息系统论域中的对象发生增加或减少,算法 3 和算法 4 在进行增量式属性约简时,在原先 t 时刻混合信息系统属性约简集 $R^{(t)}$ 和邻域区分度 $NDD_{R^{(t)}}^{U^{(t)}}(D)$ 基础上进一步进行计算,即保留了原先已有的结果,这样可以避免算法 2 中的重复计算,提高了

属性约简的计算效率,符合动态环境下的属性约简的时效性需求.

对于混合型决策信息系统 $MDIS^{(t+1)}=(U^{(t+1)},C\cup D)$,论域 $U^{(t)}$ 至 $U^{(t+1)}$ 的变化量为 ΔU,约简集 $R^{(t)}$ 至 $R^{(t+1)}$ 的变化量为 ΔR,设 $|U^{(t+1)}|=n$, $|C|=c$, $|\Delta U|=\Delta n$, $|\Delta R|=\Delta r$,那么算法 3 和算法 4 的时间复杂度可表示为

$$T=\begin{cases}O(c\cdot n\cdot \Delta n), & R^{(t)}=R^{(t+1)}\\ O(\Delta r\cdot(2c-2r-\Delta r)\cdot n\cdot \Delta n), & R^{(t)}\subseteq R^{(t+1)}\\ O(\Delta r\cdot(2c-2r+\Delta r)\cdot n\cdot \Delta n), & R^{(t+1)}\subseteq R^{(t)}\end{cases}$$

5　实验分析

在本节我们将通过实验分析来验证本文所提出的邻域区分度增量式属性约简算法的有效性. 本实验总共分为 4 个部分,第一部分是关于实验的相关准备工作以及算法的参数设置问题,第二部分是关于混合信息系统论域中对象逐渐增加时非增量式算法与增量式算法的属性约简效率比较,第三部分是关于混合信息系统论域中对象逐渐减少时非增量式算法与增量式算法的属性约简效率比较,第四部分是关于非增量式算法与增量式算法的属性约简结果比较分析.

5.1　实验准备与参数设置

实验中所运用的 8 个 UCI 数据集如表 1 所示,其中每个数据集都为符号型和数值型并存的数据,并且数据的规模大小不等. 由于表 1 中的每个数据集都是固定的数据集,而本文所提出的是一种针对对象变化的增量式属性约简,为了模拟数据集动态变化的环境,本实验将表 1 中的每个数据集按照对象平均分成 10 个部分,选择其中一部分作为主数据集,然后逐个将其他部分对象集添加入主数据集中,这样便可以模拟出数据集中对象集的 9 次动态增加. 运用类似的方法我们可以模拟出数据集中对象集的 9 次动态减少. 因此根据这种方法我们让所有算法进行相关的实验.

表 1　实验数据集

编号	名称	对象数	属性数	类别数
1	Cylinder	512	40	2
2	Credit	690	15	2
3	German	1000	19	2
4	Cmc	1473	9	3
5	Sick	2800	28	2
6	Abalone	4177	8	29
7	Characters	6000	7	10
8	Thyroid	7200	21	2

实验所运行的硬件平台为 CPU:Intel Core i7 4790,主频 3.6GHz;内存:DDR3 1600MHz 8GB. 所运用的软件平台为 Matlab2013b.

对于本文所提出的非增量式属性约简算法和增量式属性约简算法,其中都包含参数邻域半径 δ. 根据相关文献[15,22,23],学者们通过相关实验指出邻域半径设定在[0.1~0.2]之间能够得到较好的属性约简结果,本实验选择邻域半径 $\delta=0.15$ 进行实验. 同时对于距离函数,本实验统一选取为欧氏距离[15].

5.2　对象增加时非增量式与增量式属性约简的效率比较

图 1 所示的是邻域区分度非增量式属性约简算法(NARNDD)与增量式属性约简算法(IARNDD-OD)分别在 8 个数据集下对象增加时的属性约简计算时间比较. 由于本实验模拟了数据集中对象集的 9 次动态增加,因此图 1 中每幅图的横坐标为对象的动态增加次数,刻度值为 1 至 9,纵坐标表示对象每次增加时属性约简所消耗的计算时间,单位为秒(s).

观察图 1 中各个数据集的实验结果可以发现,随着混合信息系统对象的逐渐增加,NARNDD 算法的计算时间以较快的速率增长,而 IARNDD-OD 算法增长的较为缓慢,并且 IARNDD-OD 算法的属性约简计算时间均大幅度低于 NARNDD 算法的计算用时,对于 Cylinder, Abalone 和 Characters 等数据集,IARNDD-OD 算法的计算时间远低于 NARNDD 算法. 这主要是由于 IARNDD-OD 算法采用增量式的方法进行属性约简,一方面,在混合信息系统的邻域类计算中,IARNDD-OD 算法运用增量式的方法进行计算更新,当对象增加时,IARNDD-OD 算法只对新加入的对象进行邻域类计算,原来论域中对象的邻域类只需要做一些相应的处理便可更新完成,而 NARNDD 算法完全是将所有对象的邻域类重新计算一遍,这样会消耗更多的时间,另一方面,算法在搜索属性的过程中,IARNDD-OD 算法是在原先约简集的基础上进一步得到新的约简集,也就是根据新信息系统中邻域区分度值的变化在原先约简集的基础上作出相应的增加和删除属性,从而完成对约简集的更新,这样同样提高了属性约简的效率. 综合这两方面,IARNDD-OD 算法具有更高的属性约简效率.

5.3　对象减少时非增量式与增量式属性约简的效率比较

图 2 所示的是邻域区分度非增量式属性约简算法(NARNDD)与增量式属性约简算法(IARNDD-OR)分别在 8 个数据集下对象减少时的属性约简计算时间比较. 同样地,本实验模拟了数据集中对象集的 9 次动态减少,因此图 2 中每幅图的横坐标为对象的动态减少次数,刻度值为 1 至 9,纵坐标表示对象每次减少时属性约简所消耗的计算时间,单位为秒(s).

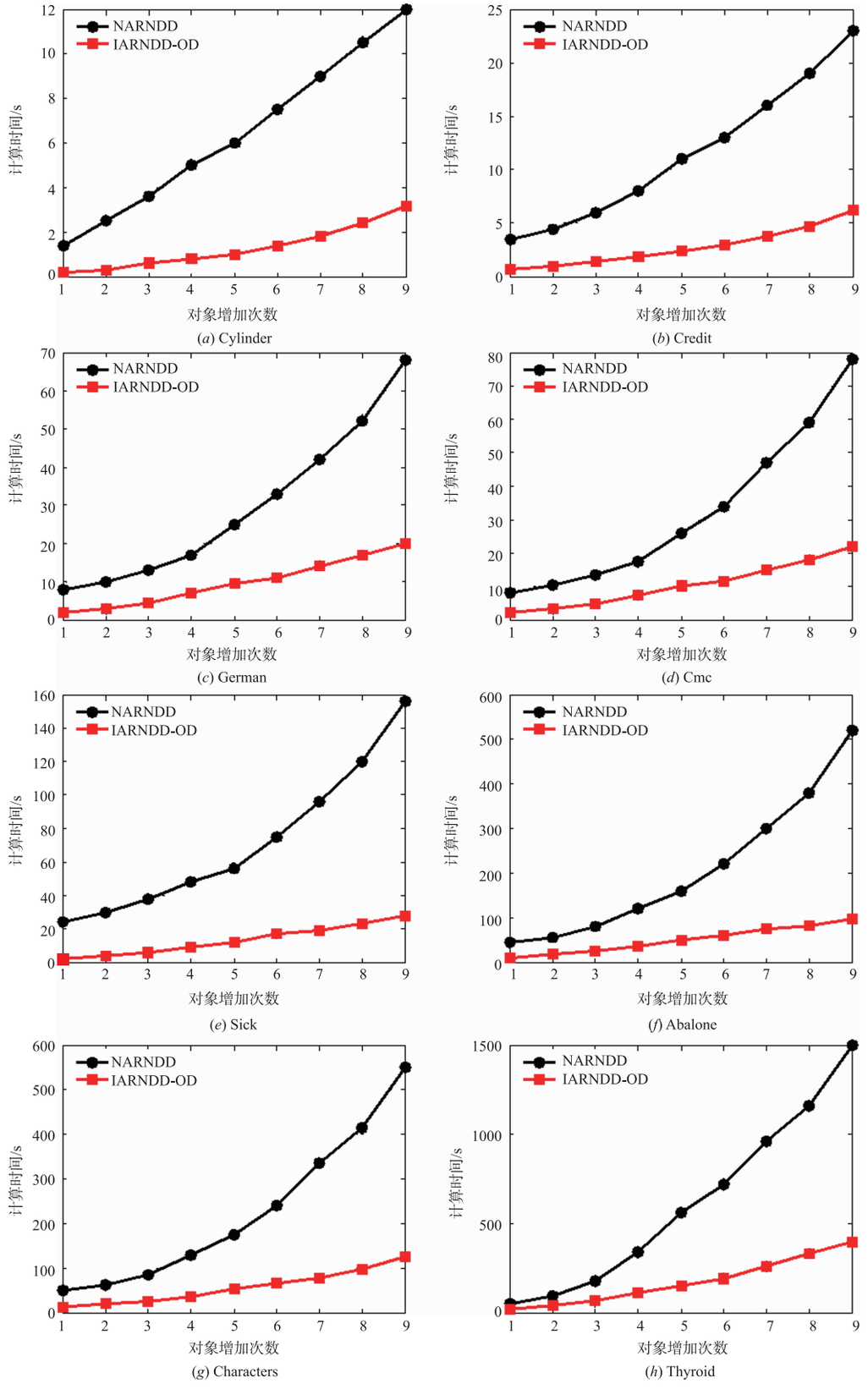

图1 各数据集对象增加时两类算法的属性约简效率比较

观察图2中各个数据集的实验结果可以发现,对于混合信息系统对象的每次减少,IARNDD-OR 算法的属性约简计算时间均大幅度低于 NARNDD 算法的计算用时,对于 Cylinder,Sick 和 Characters 等数据集,IARNDD-OR 算法的计算时间远低于 NARNDD 算法. 类似于 IARNDD-OD 算法,IARNDD-OR 算法同样采用的是增量式的方法进行属性约简,在混合信息系统的邻域类计算中,IARNDD-OR 算法在原来邻域类的基础上进行更新,同时算法在搜索属性的过程中,IARNDD-OR 算法也同样是在原先约简集的基础上进一步得到新的约简集,其属性约简的算法机制与 IARNDD-OD 算法相同. 因此,IARNDD-OR 算法相对于 NARNDD 算法也具有更高的属性约简效率.

5.4　非增量式与增量式属性约简的结果比较

为了对非增量式与增量式属性约简的结果进行比较分析,我们在实验的过程中任意选取一个更新次数,将此时的属性约简结果输出出来,然后比较他们结果的差异. 表2所示的是每个数据集对象第7次增加时 NARNDD 算法和 IARNDD-OD 算法的属性约简结果,表3所示的是每个数据集对象第5次减少时 NARNDD 算法和 IARNDD-OR 算法的属性约简结果.

表2　各数据集对象第7次增加时两类算法的属性约简结果比较

数据集	NARNDD 算法		IARNDD-OD 算法	
	属性约简	属性个数	属性约简	属性个数
Cylinder	3,7,8,14,17,22,25,29,31,38,39	11	5,7,12,14,17,20,25,38,39	9
Credit	1,2,4,6,9,12,15	7	1,2,5,6,9,12,13	7
German	2,4,7,11,13,14,16,18	8	2,5,7,11,14,17,18	7
Cmc	1,4,6,9	4	1,4,6,9	4
Sick	2,3,5,7,11,14,16,20,23,25,27,28	12	2,5,7,8,11,13,14,19,23,27	10
Abalone	1,2,4,6,7	5	1,2,4,5,6,7	6
Characters	2,5,6,7	4	2,5,6,7	4
Thyroid	2,5,7,8,11,14,17	7	2,7,11,17,19,20	6

观察表2可以看出,对于同一组数据集,两种算法在少部分数据集属性约简结果是一样的,例如 Cmc 和 Characters 数据集,而在其他数据集的属性约简结果并不完全一样,其中 IARNDD-OD 算法得到属性约简结果更小一些,例如 Cylinder、German、Sick 和 Thyroid 数据集. 出现这种情况主要是由于这两种属性约简算法的运行机制不一样导致的,当数据集更新后,NARNDD 算法在重新计算新的约简集时,约简集从空集开始不断迭代搜索属性,而 IARNDD-OD 算法是在原先信息系统约简集的基础上进行增加和删除,因而 IARNDD-OD 算法得到约简集包含了一些原来的属性,随着信息系统不断的更新,IARNDD-OD 算法约简集中保留下来的属性都是对分类更为关键的属性,因此 IARNDD-OD 算法得到约简集会更加的小. 同样的,表3中的属性约简结果也出现了类似的情形.

表3　各数据集对象第5次减少时两类算法的属性约简结果比较

数据集	NARNDD 算法		IARNDD-OR 算法	
	属性约简	属性个数	属性约简	属性个数
Cylinder	2,3,7,8,15,17,23,25,29,31,38,39	12	5,7,12,14,15,17,23,25,38,39	10
Credit	1,2,4,8,9,14,15	7	1,2,5,6,9,12,14	7
German	2,4,7,12,13,15,16,18	8	2,5,7,10,14,15,16,18	8
Cmc	1,4,6,9	4	1,4,6,9	4
Sick	2,4,5,7,13,14,16,20,23,27,28	11	2,5,7,8,11,13,16,19,23,28	10
Abalone	1,2,4,5,6,7	6	1,2,4,5,6,7	6
Characters	2,5,6,7	4	2,3,5,6,7	5
Thyroid	2,5,7,8,9,12,15,19	8	2,7,8,12,17,19,20	7

表4所示的是表2和表3中属性约简结果在支持向量机(SVM)和改进决策树(C4.5)两种分类器下的分类精度. 观察表4中的结果可以看出,在对象第7次增加时属性约简结果的分类精度中,IARNDD-OD 算法在数据集 German、Sick、Abalone 和 Thyroid 中具有较高的 SVM 分类精度,其余数据集的分类精度相等或较低,IARNDD-OD 算法在数据集 Credit、German 和 Abalone 中具有较高的 C4.5 分类精度,其余数据集分类精度相等或较低. 在对象第5次减少时属性约简结果的分类精度中,IARNDD-OR 算法在数据集 German、Sick 和 Characters 中具有较高的 SVM 分类精度,IARNDD-OR 算法在数据集 Credit、German、Sick 和 Characters 中具有更高的 C4.5 分类精度. 因此可以得出增量式算法与非增量式算法的约简结果都具有较好的分类性能,所以非增量式算法和增量式算法选择出的约简集都是有效的.

表4　各数据集中两类算法属性约简结果的分类精度比较(%)

数据集	对象第7次增加时				对象第5次减少时			
	NARNDD 算法		IARNDD-OD 算法		NARNDD 算法		IARNDD-OR 算法	
	SVM	C4.5	SVM	C4.5	SVM	C4.5	SVM	C4.5
Cylinder	86.54	85.37	85.17	84.75	87.48	85.19	86.54	84.75
Credit	83.58	81.43	82.49	81.72	83.12	82.44	82.97	83.36
German	72.27	70.17	73.56	71.47	71.65	69.20	72.63	70.24
Cmc	77.54	75.38	77.54	75.38	77.54	75.38	77.54	75.38
Sick	86.46	94.16	88.57	93.26	87.58	92.15	89.57	95.26
Abalone	65.40	60.23	66.36	62.43	64.28	63.72	64.28	63.72
Characters	97.37	95.68	97.37	95.68	96.16	94.57	98.63	95.71
Thyroid	67.45	62.87	68.46	62.55	70.54	67.26	69.92	66.56

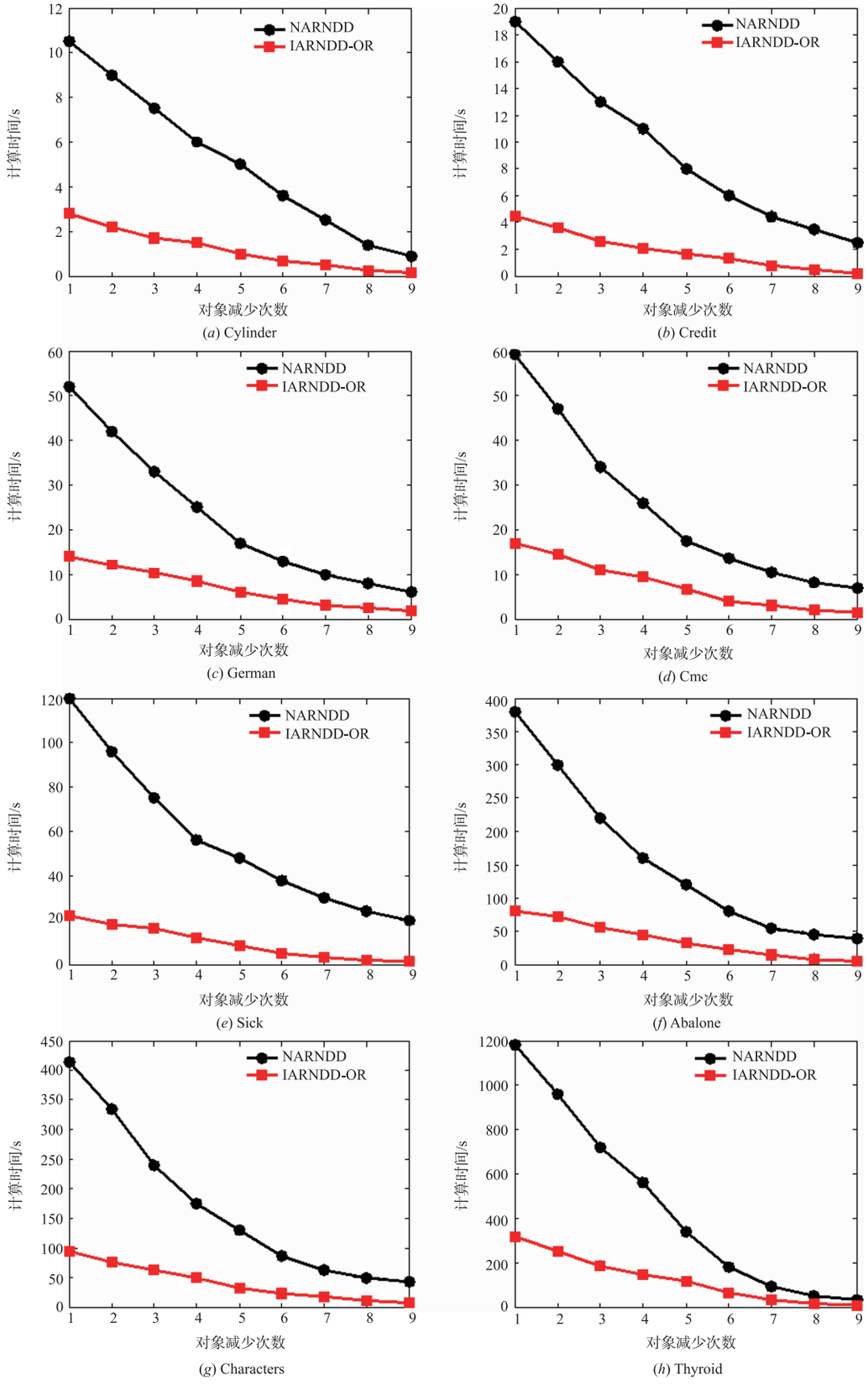

图2 各数据集对象减少时两类算法的属性约简效率比较

5.5 实验小结

经过5.2节、5.3节以及5.4节中两大类算法的实验分析比较,我们可以得出,在属性约简效率方面,无论是混合信息系统对象的增加还是减少,所提出的增量式属性约简算法(IARNDD-OD算法和IARNDD-OR算法)均比非增量式属性约简算法(NARNDD算法)具有更高的属性约简效率,因此对于对象不断变化的混合信息系统,所提出的增量式属性约简算法能够快速的更新出新的属性约简结果.在属性约简结果方面,增量式属性约简算法(IARNDD-OD算法和IARNDD-OR算法)能够比非增量式属性约简算法(NARNDD算法)选择出更小的约简集,两大类算法得到的约简结果分类性能相当,即这两类算法得到的属性约简集均是有效的.因此,可以证明所提出的增量式属性约简算法可以运用于动态环境下混合信息系统的属性约简问题.

6 结语

目前已提出的增量式属性约简大多是针对符号型信息系统,而对混合型信息系统的研究较少,这促使我们对混合型信息系统的增量式属性约简进行相关的研究.本文通过在混合信息系统推广了区分度度量,提出了一种邻域区分度的度量方法,然后分别研究了混合信息系统对象增加和对象减少时邻域区分度的增量式更新方法,理论证明了该更新方法的高效性,避免了非增量式计算时的重复计算,最后根据邻域区分度的增量式计算提出了混合信息系统下的增量式属性约简算法.实验分析表明,在动态的数据环境下,所提出的增量式属性约简算法比非增量式算法具有更高的属性约简性能.本文研究的是混合信息系统下对象增加和对象减少时的增量式属性约简,属性的增加和减少也是信息系统中一种常见的变化形式,因此接下来将进一步研究此类问题的增量式属性约简.

参考文献

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

[2] 邓大勇,薛欢欢,苗夺谦,等. 属性约简准则与约简信息损失的研究[J]. 电子学报,2017,45(2):401 – 407.
DENG D-Y,XUE H-H,MIAO D-Q,et al. Study on criteria of attribute reduction and information loss of attribute reduction[J]. Acta Electronica Sinica,2017,45(2):401 – 407. (in Chinese)

[3] 续欣莹,张扩,谢珺,等. 基于互信息下粒子群优化的属性约简算法[J]. 电子学报,2017,45(11):2695 – 2704.
XU X-Y,ZHANG K,XIE J,et al. An attribute reduction based on mutual information of particle swarm optimization [J]. Acta Electronica Sinica,2017,45(11):2695 – 2704. (in Chinese)

[4] CHAN C C. A rough set approach to attribute generalization in data mining[J]. Information Sciences,1998,107 (1 – 4):169 – 176.

[5] SHU W H,SHEN H. Updating attribute reduction in incomplete decision systems with the variation of attribute set [J]. International Journal of Approximate Reasoning, 2014,55(3):867 – 884.

[6] SHU W H,SHEN H. Incremental feature selection based on rough set in dynamic incomplete data[J]. Pattern Recognition,2014,47(12):3890 – 3906.

[7] QIAN W B,SHU W H,YANG B R,et al. An incremental algorithm to feature selection in decision systems with the variation of feature set[J]. Chinese Journal of Electronics, 2015,24(1):128 – 133.

[8] 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.

[9] WEI W,WU X,LIANG J Y,et al. Discernibility matrix based incremental attribute reduction for dynamic data[J]. Knowledge-Based Systems,2018,140(15):142 – 157.

[10] 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 YY. An incremental attribute reduction algorithm for group objects[J]. CAAI Transactions on Intelligent Systems,2016,11(4):496 – 502. (in Chinese)

[11] LANG G M,MIAO D Q,CAI M J,et al. Incremental approaches for updating reducts in dynamic covering information systems[J]. Knowledge-Based Systems,2017,134 (15):85 – 104.

[12] XIE X J,QIN X L. A novel incremental attribute reduction approach for dynamic incomplete decision systems [J]. International Journal of Approximate Reasoning, 2018,93:443 – 462.

[13] JING Y G,LI T R,FUJITA H,et al. An incremental attribute reduction approach based on knowledge granularity with a multi-granulation view[J]. Information Sciences, 2017,411:23 – 38.

[14] JING Y G,LI T R,LUO C,et al. An incremental approach for attribute reduction based on knowledge granularity[J]. Knowledge-Based Systems,2016,104(15):24 – 38.

[15] 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.

[16] 黄恒秋,曾玲,黎利辉. 混合值不完备系统的双邻域粗糙集分类方法[J]. 控制与决策,2018,33(7):1207 – 1214.
HUANG H Q,ZENG L,LI L H. Double-neighborhood

rough set classification method in incomplete decision system with hybrid value[J]. Control and Decision,2018,33(7):1207 – 1214. (in Chinese)

[17] ZHAO H, QIN K. Mixed feature selection in incomplete decision table [J]. Knowledge-Based Systems, 2014, 57(2):181 – 190.

[18] ZHANG J B, LI T R, RUAN D, et al. Neighborhood rough sets for dynamic data mining [J]. International Journal of Intelligent Systems,2012,27:317 – 342.

[19] YAO Y Y,ZHAO Y. Data analysis based on discernibility and indiscernibility [J]. Information Sciences, 2007, 177(22):4959 – 4976.

[20] TENG S H, LU M, YANG A F, et al. Efficient attribute reduction from the viewpoint of discernibility [J]. Information Sciences,2016,326(1):297 – 314.

[21] SUSMAGA R. Reducts and constructs in attribute reduction[J]. Fundamenta Informaticae,2004,61:159 – 181.

[22] 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. (in Chinese)

[23] 胡峰,王蕾,周耀. 基于三支决策的不平衡数据过采样方法[J]. 电子学报,2018,46(1):135 – 144.
HU F,WANG L,ZHOU Y. An oversampling method for imbalance data based on three-way decision model[J]. Acta Electronica Sinica,2018,46(1):135 – 144. (in Chinese)

作者简介

盛 魁 1981 年 2 月出生,安徽涡阳人. 2011 年毕业于安徽大学软件学院,随后在亳州职业技术学院信息工程系工作,从事粒计算和数据挖掘方面的研究.
E-mail:shengk1981@163.com

卞显福 1981 年 11 月出生,安徽合肥人. 2011 年毕业于中国科学技术大学软件学院,随后在中国科学技术大学软件学院工作,从事数据挖掘方面的研究.