一种面向增删操作的粗糙集属性约简更新算法

陆　悠,华　泽,奚雪峰,张　妮,吴宏杰

(苏州科技学院电子与信息工程学院,江苏 苏州 215000)

[**摘要**]　属性约简是粗糙集理论的核心内容之一,在信息系统的对象信息不断出现增删等更新操作的环境下,如何进行快速有效的属性约简则是一个亟需解决的迫切问题. 提出一种面向增删操作的属性约简更新算法,面向更新前后的决策表,首先分析了对象信息动态增加与删除情况下信息熵的变化机制以及约简属性对新增或删除对象的区分情况,然后提出基于区分情况的新条件熵值的计算方法,最后给出基于散列表的属性约简更新算法. 实验结果证明,本文方法可以快速求解出增删更新后的属性约简结果,其性能较传统方法有较大优势.

[**关键词**]　粗糙集,增量式数据挖掘,属性约简

[**中图分类号**]TP18　[**文献标志码**]A　[**文章编号**]1001-4616(2015)01-0048-09

An Attribute Reduction Update Algorithm for Object's Adding-Deleting Based on Rough Set Theory

Lu You,Hua Ze,Xi Xuefeng,Zhang Ni,Wu Hongjie

(School of Electronical and Information Engineering,Suzhou University of Science and Technology,Suzhou 215000,China)

Abstract:Attribute reduction is one of the important topics in the research on rough set theory. When an object was added to or deleted from the original decision table,how to calculate attribute reduction fast and effectively is a pressing problem. This paper proposed an attribute reduction update algorithm. Firstly,the changing mechanism of conditional entropy was analyzed when object is added to or removed from the table,and then we divided the added or removed objects into different cases. Furthermore,we presented the update algorithm based on these cases and implemented it based on hash table. Experiment results show that our algorithm can calculate the attribute reduction fast and outperforms the existing methods.

Key words:rough set,incremental data mining,reduction of attribute

粗糙集(Rough Set)理论作为一种刻画不确定性和不完整性知识的数学工具,自 20 世纪 80 年代由波兰的 Z. Pawlak 提出后就一直受到人们的重视,由于其提供了一整套方法从数学上严格地处理数据分类问题[1]. 近年来在知识与数据发现、模式识别与分类、信息系统分析、决策支持系统以及网络行为评估等多个领域取得了较为成功的应用. 作为粗糙集挖掘和简化知识的关键步骤[2,3],如何快速准确地进行属性约简一直是研究人员的关注重点,目前也已取得了较多研究成果. 然而在实际领域的应用过程中特别是有实时需求的领域(例如网络中用户行为实时评估[4])中,由于处理的数据总在保持不断更新的趋势,这就要求进行粗糙集计算时需要不断地删去旧有数据包含的信息,同时增加新产生的数据所包含的信息,从而保证知识发现过程中的实时性和有效性. 因此,如何面向增删操作,研究实现粗糙集的属性约简更新算法已成为亟待解决的关键问题.

增删操作会导致参与属性约简计算的数据产生部分增加和删除,而现有的基于差别矩阵[5]、基于正域[6]以及基于启发式信息等多种约简算法[7]大多只针对静态的信息数据,显然,如果简单地重复这些算法进行属性约简,那么数据部分增加会导致前后反复的计算过程中产生大量的冗余,导致时间开销过高而无法保证应用的实时性,因此目前相关人员针对增加数据的约简更新展开较多的研究,这些成果大多能够

收稿日期:2014-08-16.

基金项目:江苏省自然科学基金(BK20131154).

通讯联系人:陆悠,博士生,讲师,研究方向:网络应用及安全,用户行为控制. E-mail:luyou. china@ gmail. com

在不计算原有约简的基础上对增量数据的属性约简进行实时有效的更新,但仍存在一些不足,例如文献[8]提出一种基于邻域差别矩阵的属性约简增量式算法仅能处理连续型数据对象动态增加的属性约简更新. 文献[9]提出基于正域的属性约简的增量式算法对数据增加引起的不一致情况并未讨论等等. 更为重要的是,如果同时考虑数据增加和删除的话,属性约简更新需要考虑的情况更为复杂,上述算法都没有考虑如何在数据增删情况进行约简更新,如何在数据增删环境下实现属性约简的快速更新目前尚未有较好的解决方案.

针对现有工作的不足,本文提出一种面向数据增删的属性约简更新方法,该方法采用文献[10]基于条件熵的属性约简思想,首先分析数据对象在部分增加和删除后决策表条件熵的变化对其属性约简的影响,然后将属性约简的更新分为3种情况进而提出了针对性的求解方法. 最后提出一种属性约简进化算法,该算法充分利用了更新前决策表的信息熵和属性约简等信息,能够有效缩短约简的更新时间,从而确保属性约简应用的实时性.

本文剩余章节安排如下:第二节主要介绍本文涉及的基本概念;第三节介绍面向数据部分增删环境下的知识系统更新情况分析,包括基于增-删窗口机制的知识系统更新机理以及相应的条件熵更新机制;第四节则提出基于增-删数据的属性约简进化算法;第五节通过实验对本文算法进行了验证;第六节则对本文进行总结和展望.

1　基本概念

粗糙集(Rough Set)理论是由波兰的Z. Pawlak于1982年提出来的,是一种刻画不确定性和不完整性知识的数学工具,提供了一整套方法从数学上严格地处理数据分类问题. 经过多年的研究与发展,它已在信息系统分析、人工智能及应用、决策支持系统、知识与数据发现、模式识别与分类等多个领域取得了较为成功的应用.

根据粗糙集理论,可以将用户行为的形式化描述转化为粗糙集中的知识系统,定义如下:

定义1　决策表:粗糙集中使用如下五元组 $S=\{U,C\cup D,V,f\}$ 表示决策表,该式中,$U=\{x_1,x_2,\cdots,x_n\}$ 表示论域,即对象的非空有限集合,C 为条件属性的非空有限集合,D 为决策属性的非空有限集合,且 $C\cap D=\Phi$;$V=\cup_{a\in C\cup D}V_a$,其中 V_a 为属性 a 的值域,$a\in C\cup D$;函数 $f:U\times C\cup D\rightarrow V$ 为信息函数,表示知识系统中每个对象在属性上的取值,即 $\forall a\in C\cup D,x\in U$ 有 $f(x,a)\in V_a$. 由此,每一个属性子集 $P\subseteq(C\cup D)$ 决定了1个二元不可区分关系 IND(P):IND(P)=$\{(x,y)\in U\times U\mid \forall a\in P,f(x,a)=f(y,a)\}$. 关系 IND($P$) 构成了 U 的1个划分 U/IND(P),简记为 U/P,U/P 中的任何元素 $[x]_P=\{y\mid \forall a\in P,f(x,a)=f(y,a)\}$ 都称为等价类.

根据以上定义,可以比较划分的精细程度,给定符号 $\preceq$,若 $C\subseteq B$,且对任意 $X\in U/B$,存在 $Y\in U/C$,使得 $X\supseteq Y$,则称 $C\preceq B$,即 C 的划分比 B 更精细,另外称 $C\prec B$,如果存在 $X_o\in U/B,Y_o\in U/C,X_o\Leftrightarrow Y_o$.

定义2　条件熵、属性约简和属性重要度:决策表 $S=\{U,A=C\cup D,V,f:U\times A\rightarrow V\}$ 中,设属性集 $B\subseteq C$,$U/B=\{X_1,X_2,\cdots,X_n\}$,$U/D=\{Y_1,Y_2,\cdots,Y_m\}$,则决策表中 B 关于 D 的条件熵[10]可按下式计算 $H(D\mid B)=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{|Y_j\cap X_i|}{|U|}\frac{|\overline{Y}_j-\overline{X}_i|}{|U|}$;于是定义属性约简:若属性集 $B\subseteq A$ 是决策表 S 的一个属性约简,则有(1)$H(D\mid B)=H(D\mid C)$;(2) $\forall a\in B,H(D\mid B-\{a\}\neq H(D\mid C))$. 定义属性重要度:决策表 S 中,设属性集合 $B\subseteq A$,属性 $a\in B$,则其属性重要度定义为 SIG(a)=$(H(D\mid B-\{a\})-H(D\mid B))/H(D))$.

定义3　不一致:在决策表 S 中,若 $B\subseteq C$,且两个不同的对象 x 和 y,对 $\forall a\in B$ 满足 $f(x,a)=f(y,a)$,$f(x,D)\neq f(y,D)$,则称 x 和 y 为 B 不一致的.

2　面向增删的知识系统更新分析

2.1　基于增-删窗口的知识系统更新

在很多实际应用中,粗糙集面临的数据更新模式有增量式更新和增-删式更新两种,前者仅仅存在新数据不断添加情况,而后者则在存在数据添加以及删除并存的情况,根据决策表的结构,本文将增-删式更新过程整理为如图1所示的基于增-删窗口的数据更新方法:

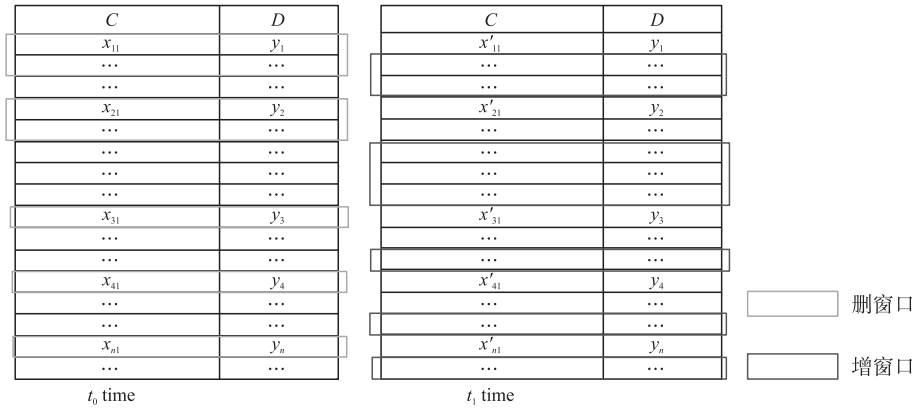

图 1　基于增删窗口的数据更新

Fig. 1　Data update based on delete-add windows

根据离散型的决策表中的数据在条件属性和决策属性方面的特点,可将决策表按每一个决策属性值(设为 $y_1,y_2,\cdots,y_n$)分为 n 个数据块,每个数据块都各具备一个增-删滑动窗口,于是数据更新可视为不同滑动窗口的操作,其流程为:

时刻 t_0:当前决策表中的数据由 n 个局部数据块,从此刻起新加入决策表中的数据将按其决策属性值分别送至对应的临时缓存中,构成该数据块的增窗口,同时该块中相应的需移除的数据则构成删窗口.

时刻 t_1:将每一个数据块删窗口的数据从决策表移除,将缓存的增窗口中数据加入决策表,其后的更新过程以此类推.

2.2　增-删数据后条件熵的更新机制

决策表数据更新后,由于条件熵会发生相应的变化,因此属性约简需要重新计算,已有的文献已经给出针对数据增加情况下的属性约简进化算法,在其基础上,本文进一步提出 1 个基于增删双窗口的属性约简进化算法. 首先对知识系统信息增删后条件熵变化情况进行分析,从而为实现高效的属性约简提高理论基础.

定义 8　$X_q^B,X_q^{B'},X_q^{C'},X_q^C,Y'_p,Y_p$:设对象 x 为决策表 $S=\{U,A=C\cup D,V,f:U\times A\to V\}$ 的更新对象,则对 U'(若 x 为新增对象,则 $U'=U\cup\{x\}$,若 x 为删除对象,则 U'为原决策表 S 中未删除 x 时的 U)来说,定义 $X_q^{B'}=X_q^B\cup\{x\}$ 为 U'中包含 x 的 $U'|B$ 等价类,X_q^B 对应 $U|B$ 的等价类(若 B 可区分 x,则 $X_q^B=\Phi$),类似还可以定义 $X_q^{C'}$、X_q^C 与 Y'_p、Y_p. 于是可以得到增-删数据对条件熵的改变情况:

引理 1　给定决策表 S 中,设属性集合 $B\subseteq A$,$U/B=\{X_1,X_2,\cdots,X_n\}$,$U/D=\{Y_1,Y_2,\cdots,Y_m\}$,则新增或删除样本数据 x 后则属性集 B 关于决策属性 D 的条件熵更新公式为

$$H_{U\cup\{x\}}(D|B)=\frac{1}{(|U|+1)^2}(|U|^2+2|X_q^{B'}-Y'_p|),\quad 和$$

$$H_{U-\{x\}}(D|B)=\frac{1}{(|U|-2)^2}(|U|^2H_U(D|B)-2|X_q^{B'}-Y'_p|).$$

证明　增加样本对应的条件熵更新公式在在文献[10]中已经得到证明,删除样本的条件熵的推导过程如下:设样本空间 $U'=U-x$,删去的对象 x 属于以下 4 种可能之一:

(1)x 不属于 U'/B 的等价类;x 不属于 U'/D 的等价类;

(2)x 不属于 U'/B 的等价类;x 属于 U'/D 的等价类;

(3)x 属于 U'/B 的等价类;x 不属于 U'/D 的等价类;

(4)x 属于 U'/B 的等价类;x 属于 U'/D 的等价类;

对于情况(1),设 U'/B 的等价类为$\{X_1,X_2,\cdots,X_m\}$,由于更新对象 x 不属于其中,因此 U/B 的等价类即$\{X_1,X_2,\cdots,X_{n+1}\}$,$X'_q=X_{n+1}=\{x\}$,类似的有 U'/D 的等价类为$\{Y_1,\cdots,Y_m\}$,U/D 的等价类$\{Y_1,\cdots,Y_m,Y_{m+1}\}$,$Y'_p=Y_{m+1}=\{x\}$,于是

$$H_{U'}(D|B)=\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}|Y_i\cap X_j||\overline{Y}_i-\overline{X}_j|}{(|U|-1)^2},$$

而

$$H_U(D|B)=\frac{\sum_{i=1}^{m+1}\sum_{j=1}^{n+1}|Y_i\cap X_j||\bar{Y}_i-\bar{X}_j|}{(|U|)^2}=\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}|Y_i\cap X_j||\bar{Y}_i-\bar{X}_j|+\sum_{j=1}^{n}|Y'_p\cap X_j||\bar{Y}'_p-\bar{X}_j|}{(|U|)^2}+$$

$$\frac{\sum_{i=1}^{m}|Y_i\cap X'_q||\bar{Y}_i-\bar{X}'_q|+|Y'_p\cap X'_q||\bar{Y}'_p-\bar{X}'_q|}{(|U|)^2}=\frac{H_{U'}(D|B)(|U|-1)^2}{(|U|)^2}.$$

由于$|X'_q-Y'_p|=0$,所以有 $H_{U-\{x\}}(D|B)=\frac{1}{(|U|-1)^2}(|U|^2H_U(D|B)-2|X'_q-Y'_p|)$.

对于情况(2),U'/D 的等价类为$\{Y_1,Y'_p,\cdots,Y_m\}$,U/D 的等价类$\{Y_1,Y_p,\cdots,Y_m\}$其中 $x\in Y_p$,$Y_p=Y'_p\cup\{x\}$,而 U'/B 的等价类为$\{X_1,\cdots,X_n\}$,U/B 的等价类$\{X_1,\cdots,X_n,X_{n+1}\}$,$X'_q=X_{n+1}=\{x\}$,于是有

$$H_{U'}(D|B)=\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}|Y_i\cap X_j||\bar{Y}_i-\bar{X}_j|}{(|U|-1)^2}=\frac{\sum_{i=1,i\neq p}^{m}\sum_{j=1}^{n}|Y_i\cap X_j||\bar{Y}_i-\bar{X}_j|}{(|U|-1)^2}+$$

$$\frac{\sum_{j=1}^{n}|Y'_p\cap X_j||\bar{Y}'_p-\bar{X}_j|}{(|U|-1)^2},$$

而

$$H_U(D|B)=\frac{\sum_{i=1}^{m}\sum_{j=1}^{n+1}|Y_i\cap X_j||\bar{Y}_i-\bar{X}_j|}{(|U|)^2}=\frac{\sum_{i=1,i\neq p}^{m}\sum_{j=1}^{n}|Y_i\cap X_j||\bar{Y}_i-\bar{X}_j|+\sum_{j=1}^{n}|Y_p\cap X_j||\bar{Y}_p-\bar{X}_j|}{(|U|)^2}+$$

$$\frac{\sum_{i=1}^{m}|Y_i\cap X'_q||\bar{Y}_i-\bar{X}'_q|+|Y_p\cap X_q||\bar{Y}_p-\bar{X}'_q|}{(|U|)^2}.$$

显然有:$|X'_q-Y'_p|=0$,$\bar{Y}_p-\bar{X}'_q=\Phi$ 以及 $i=1,2,\cdots,m$ 且 $i\neq p$ 时有 $Y_i\cap X'_q=\Phi$,而 $\sum_{j=1}^{n}|Y_P\cap X_j||\bar{Y}_P-\bar{X}_j|=\sum_{j=1}^{n}||(Y'_p\cup\{x\}\cap X_j|\overline{(Y'_p\cup\{x\})}-\bar{X}_j|=\sum_{j=1}^{n}|Y'_p\cap X_j||\bar{Y}'_p-\bar{X}_j|$,于是有 $H_{U-\{x\}}(D|B)=\frac{1}{(|U|-1)^2}\cdot(|U|^2H_U(D|B)-2|X'_q-Y'_p|)$

对于情况(3),U'/D 的等价类为$\{Y_1,\cdots,Y_m\}$,U/D 的等价类$\{Y_1,\cdots,Y_m,Y_{m+1}\}$,$Y'_p=Y_{m+1}=\{x\}$,$U'/B$ 的等价类为$\{X_1,\cdots,X'_q,\cdots,X_n\}$,$U/B$ 的等价类$\{X_1,\cdots,X_n\}$,$x\in X_q$,$X_q=X'_q\cup\{x\}$,于是有

$$H_{U'}(D|B)=\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}|Y_i\cap X_j||\bar{Y}_i-\bar{X}_j|}{(|U|-1)^2}=\frac{\sum_{i=1}^{m}\sum_{j=1,j\neq q}^{n}|Y_i\cap X_j||\bar{Y}_i-\bar{X}_j|}{(|U|-1)^2}+\frac{\sum_{i=1}^{m}|Y_i\cap X'_q||\bar{Y}_i-\bar{X}'_q|}{(|U|-1)^2},$$

$$H_U(D|B)=\frac{\sum_{i=1}^{m+1}\sum_{j=1}^{n}|Y_i\cap X_j||\bar{Y}_i-\bar{X}_j|}{(|U|)^2}=\frac{\sum_{i=1}^{m}\sum_{j=1,j\neq q}^{n}|Y_i\cap X_j||\bar{Y}_i-\bar{X}_j|+\sum_{j=1,j\neq q}^{n}|Y'_p\cap X_j||\bar{Y}'_p-\bar{X}_j|}{(|U|)^2}+$$

$$\frac{\sum_{i=1}^{m}|Y_i\cap X_q||\bar{Y}_i-\bar{X}_q|+|Y'_p\cap X_q||\bar{Y}'_p-\bar{X}_q|}{(|U|)^2}.$$

显然有:

$$\sum_{i=1}^{m}|Y_i\cap X_q||\bar{Y}_i-\bar{X}_q|=\sum_{i=1}^{m}|Y_i\cap\{(X'_q\cup x\}||\bar{Y}_i-\overline{\{(X'_q\cup x\}}|=\sum_{i=1}^{m}|Y_i\cap X'_q||(\bar{Y}_i\cap X'_q)\cup(\bar{Y}_i\cap\{x\})|=$$

$$\sum_{i=1}^{m}(|Y_i\cap X'_q||\bar{Y}_i-\bar{X}'_q|+|Y_i\cap X'_q|)=\sum_{i=1}^{m}|Y_i\cap X'_q||\bar{Y}_i-\bar{X}'_q|+|X'_q|,$$

而 $\sum_{j=1,j\neq q}^{n}|Y'_p\cap X_j||\overline{Y}'_p-\overline{X}_j|=\Phi$, $|Y'_p\cap X_q||\overline{Y}'_p-\overline{X}_q|=|\{x\}||Y'_p\cap(X'_q\cup\{x\})|=|X'_q|$,于是得出 $H_{U-\{x\}}(D|B)=\frac{1}{(|U|-1)^2}(|U|^2H_U(D|B)-2|X'_q-Y'_p|)$

对于情况(4),U'/D 的等价类为 $\{Y_1,Y'_p,\cdots,Y_m\}$,U/D 的等价类 $\{Y_1,Y_p,\cdots,Y_m\}$ 其中 $x\in Y_p$, $Y_p=Y'_p\cup\{x\}$,U'/B 的等价类为 $\{X_1,X'_q,\cdots,X_n\}$,$U|B$ 的等价类 $\{X_1,\cdots,X_n\}$, $x\in X_q$, $X_q=X'_q\cup\{x\}$,于是有

$$H_{U'}(D|B)=\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}|Y_i\cap X_j||\overline{Y}_i-\overline{X}_j|}{(|U|-1)^2}=\frac{\sum_{i=1,i\neq p}^{m}\sum_{j=1,j\neq q}^{n}|Y_i\cap X_j||\overline{Y}_i-\overline{X}_j|}{(|U|-1)^2}+$$

$$\frac{\sum_{i=1,i\neq p}^{m}|Y_i\cap X'_q||\overline{Y}_i-\overline{X}'_q|}{(|U|-1)^2}+\frac{\sum_{j=1,j\neq q}^{m}|Y'_p\cap X_j||\overline{Y}'_p-\overline{X}_j|}{(|U|-1)^2}+\frac{|Y'_p\cap X'_q||\overline{Y}'_p-\overline{X}'_q|}{(|U|-1)^2}.$$

$$H_U(D|B)=\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}|Y_i\cap X_j||\overline{Y}_i-\overline{X}_j|}{(|U|-1)^2}=\frac{\sum_{i=1,i\neq p}^{m}\sum_{j=1,j\neq q}^{n}|Y_i\cap X_j||\overline{Y}_i-\overline{X}_j|}{(|U|-1)^2}+$$

$$\frac{\sum_{i=1,i\neq p}^{m}|Y_i\cap X_q||\overline{Y}_i-\overline{X}_q|}{(|U|-1)^2}+\frac{\sum_{j=1,j\neq q}^{n}|Y_p\cap X_j||\overline{Y}_p-\overline{X}_j|}{(|U|-1)^2}+\frac{|Y_p\cap X_q||\overline{Y}_p-\overline{X}_q|}{(|U|-1)^2}.$$

显然有:

$$\sum_{i=1,i\neq p}^{m}|Y_i\cap X_q||\overline{Y}_i-\overline{X}_q|=\sum_{i=1,i\neq p}^{m}(|Y_i\cap X'_q||\overline{Y}_i-\overline{X}'_q|+|Y_i\cap X'_q|)=\sum_{i=1,i\neq p}^{m}|Y_i\cap X'_q||\overline{Y}_i-\overline{X}'_q|+|X'_q-Y'_p|.$$

$$\sum_{j=1,j\neq q}^{n}|Y_p\cap X_j||\overline{Y}_p-\overline{X}_j|=\sum_{j=1,j\neq q}^{n}|(Y'_p\cup\{x\})\cap X_j||\overline{(Y'_p\cup\{x\})}-\overline{X}_j|=\sum_{j=1,j\neq q}^{n}|Y'_p\cap X_j||\overline{Y}'_p\cap X_j|=$$

$$\sum_{j=1,j\neq q}^{n}|Y'_p\cap X_j||\overline{Y}'_p-\overline{X}_j|.$$

$$|Y_p\cap X_q||\overline{Y}_p-\overline{X}_q|=|(Y'_p\cup\{x\})\cap(X'_q\cup\{x\})||\overline{(Y'_p\cup\{x\})}-\overline{X'_q\cup\{x\}}|=$$

$$|(Y'_p\cap X'_q)\cup\{x\}||\overline{Y}'_p\cap\overline{\{x\}}\cap X'_q|=|Y'_p\cap X'_q||\overline{Y}'_p-\overline{X}'_q|+|X'_q-Y'_p|.$$

于是得出 $H_{U-\{x\}}(D|B)=\frac{1}{(|U|-1)^2}(|U|^2H_U(D|B)-2|X'_q-Y'_p|)$.

综合情况(1)(2)(3)(4),可得结论 $H_{U-\{x\}}(D|B)=\frac{1}{(|U|-1)^2}(|U|^2H_U(D|B)-2|X'_q-Y'_p|)$,证毕.

根据引理 1 可知,当样本数据出现增和删的情况后,条件熵会发生变化,下面分析变化机制与属性约简的关联,先给出区分的定义:

定义 9　区分:设对象 x 为决策表 S 的更新对象,则对 U'(若 x 为新增对象,则 $U'=U\cup\{x\}$,若 x 为删除对象,则 $U'=U-\{x\}$)来说,属性集 $K\subseteq C\cup D$ 满足条件:$\forall a\in K$, $\forall y\in U$,有 $f(y,a)\neq f(x,a)$,则称属性集 K 可区分 x,否则称 K 不可区分 x.

于是结合引理 1 的条件熵更新公式,可以得到增-删数据后属性约简的变化情况如下:

定理 1　给定决策表 S 中,设属性集合 $B=\mathrm{Red}(U)$ 为 S 的属性约简,x 为 S 的更新对象,更新后的决策表为 $S'=\{U',A=C\cup D,V,f:U'\times A\to V\}$(若 x 为新增对象,则 $U'=U\cup\{x\}$,若 x 为删除对象,则 $U'=U-\{x\}$)当 B 与 C 同时可区分 x 或者同时不可区分 x 时,则 $B=\mathrm{Red}(U)$ 仍是 U'的属性约简.

证明　根据引理 1 与 2,更新对象后的条件熵变化如下式:

$$H_{U\cup\{x\}}(D|B)=\frac{|U|^2H_U(D|B)+2|X_q^{B'}-Y'_p|}{(|U|+1)^2}\quad 或\quad H_{U-\{x\}}(D|B)=\frac{|U|^2H_U(D|B)-2|X_q^{B'}-Y'_p|}{(|U|-1)^2}.$$

$$H_{U\cup\{x\}}(D|C)=\frac{|U|^2H_U(D|C)+2|X_q^{C'}-Y'_p|}{(|U|+1)^2}\quad 或\quad H_{U-\{x\}}(D|C)=\frac{|U|^2H_U(D|C)-2|X_q^{C'}-Y'_p|}{(|U|-1)^2}.$$

由于 B 为 U 的属性约简,所以有 $H_U(D|B)=H_U(D|C)$,如果 B 与 C 同时可区分 x,则必有 $X_q^{B'}=X_q^{C'}=\{x\}$,有 $X_q^{B'}=X_q^{C'}$,而如果 B 与 C 同时不可区分 x,则有 $X_q^{B'}=X_q^B\cup\{x\}$,$X_q^{C'}=X_q^C\cup\{x\}$,由于 B 为 C 的属性约简,根据约简的定义,有 $X_q^B=X_q^C$,也有 $X_q^{B'}=X_q^{C'}$,所以有 $H_{U'}(D|B)=H_{U'}(D|C)$

B 为 U 的属性约简,根据约简定义,任取属性 $a\in B$,有 $H_U(D|B)<H_U(D|B-a)$,而明显 $X_q^{B'}\subseteq X_q^{B-\{a\}'}$,所以有 $|X_q^{B'}-Y_p'|\leqslant|X_q^{B-\{a\}'}-Y_p'|$,根据引理 1、2,有

$$H_{U\cup\{x\}}(D|B-\{a\})=\frac{|U|^2H_U(D|B-\{a\})+2|X_q^{B-\{a\}'}-Y_P'|}{(|U|+1)^2},\quad 或$$

$$H_{U-\{x\}}(D|B-\{a\})=\frac{|U|^2H_U(D|B)-\{a\})-2|X_q^{B-\{a\}'}-Y_p'|}{(|U|-1)^2}.$$

显然有 $H_{U'}(D|B)=H_{U'}(D|B-\{a\})$,根据属性约简的定义,$B=\mathrm{Red}(U)$ 仍为 U' 属性约简,证毕.

定理 2　给定决策表 S 中,设属性集合 B 为 S 的属性约简,x 为 S 的更新对象,更新后的决策表为 $S'=\{U',A=C\cup D,V,f:U'\times A\to V\}$,当 B 不能区分 x,C 可以区分 x,D 不可区分 x,且有 $X_q^B=Y_p$,则 $B=\mathrm{Red}(U)$ 仍是 U'的属性约简.

证明　与定理 1 证明类似,这里仅考察 $|X_q^{B'}-Y_p'|$ 与 $|X_q^{C'}-Y_p'|$,根据定义 12 与条件,有 $|X_q^{B'}-Y_p'|=|(X_p^B\cup\{x\})\cap\overline{Y}_p\cap\overline{\{x\}}=|X_q^B-Y_p|$,由于 $X_q^B=Y_p$,所以有 $|X_q^{B'}-Y_p'|=0$,而 $|X_q^{C'}-Y_p'|=|\{x\}\cap(\overline{Y}_p\cap\overline{\{x\}})|=0$,因此有 $H_{U'}(D|B)=H_{U'}(D|C)$,而类似定理 1 的证明,有 $H_{U'}(D|B)=H_{U'}(D|B)-\{a\})$,根据属性约简的定义,$B=\mathrm{Red}(U)$ 仍为 U'的属性约简,证毕.

定理 3　给定决策表 S 中,设属性集合 B 为 S 的属性约简,x 为 S 的更新对象,更新后的决策表为 $S'=\{U',A=C\cup D,V,f:U'\times A\to V\}$,当 B 不能区分 x,C 可以区分 x,D 可区分 x,或 D 不可区分 x 且 $X_q^B\neq Y_p$,则 B 不是 U'的属性约简.

证明　仍然仅考察 $|X_q^{B'}-Y_p'|$ 与 $|X_q^{C'}-Y_p'|$:若 D 可区分 x,根据条件,有 $|X_q^{C'}-Y_p'|=|\{x\}\cap\overline{\{x\}}|=0$,而 $|X_q^{B'}-Y_p'|=|(X_p^B\cup\{x\})\cap\overline{\{x\}}=|X_q^B|$,显然 $|X_q^B|\neq0$,否则 B 将可以区分 x,于是有 $H_{U'}(D|B)>H_{U'}(D|C)$;若 D 不可区分 x,根据定理 2 的证明,有 $|X_q^{B'}-Y_p'|=|X_q^B-Y_p|$,根据条件,有 $|X_q^{B'}-Y_p'|\neq0$,而 $|X_q^{C'}-Y_p'|=0$,于是也有 $H_{U'}(D|B)>H_{U'}(D|C)$. 综上,根据约简的定义,B 必不是 U'的属性约简,证毕.

3　属性约简进化算法

根据定理 1、2、3 可知,当决策表数据更新时,仅仅当 B 不能区分 x,C 可以区分 x,D 可区分 x,或 B 不能区分 x,C 可以区分 x,D 不可区分 x 且 $X_q^B\neq Y_p$ 时才需要重新计算属性约简,并且当 x 为新增对象时,属性约简集应在原约简集内删除属性,而 x 为删除对象时属性约简集应在原约简集基础上增加属性. 然而判断区分情况需要多次查询整个决策表,其时空开销较大,为减轻开销,本文结合双窗口的样本空间特点,设计了基于哈希表的区分情况检索表用于判断不同决策属性值的区分情况,以及基于水平划分的等价类构造表用于判断等价类的相等情况,其定义如下:

区分情况检索表结构如下:

表 1　区分情况检索表

Table 1　Retrieving table based on distinction

属性	Hash 散列				
a_1	$(V_1^{a_1},T_1^{a_1})$	$(V_2^{a_1},T_2^{a_1})$	…	$(V_{i_1}^{a_1},T_{i_1}^{a_1})$	
a_2	$(V_1^{a_2},T_1^{a_2})$	…	$(V_{i_2}^{a_2},T_{i_2}^{a_2})$		
…	…	…	…		
a_n	$(V_1^{a_n},T_1^{a_n})$	…	…	…	$(V_{i_n}^{a_n},T_{i_n}^{a_n})$
c_1	$(V_1^{c_1},T_1^{c_1})$	…	$(V_{j_1}^{c_1},T_{j_1}^{c_1})$		
c_2	$(V_1^{c_2},T_1^{c_2})$	$(V_2^{c_2},T_2^{c_2})$	…	$(V_{j_2}^{c_2},T_{j_2}^{c_2})$	
…	…	…	…	…	
c_m	$(V_1^{c_m},T_1^{c_m})$	$(V_2^{c_m},T_2^{c_m})$	…	…	$(V_{j_m}^{a_n},T_{j_m}^{a_n})$

如表 1 所示,区分情况检索表由 $n+m$ 个 Hash 散列组成,每个 Hash 散列分别对应条件属性 $a_i(i=1\ \text{to}\ n)$或决策属性 $c_j(j=1\ \text{to}\ m)$,$\forall a_i \in C$,其散列由$(V_1^{a_i},T_1^{a_i})$,$(V_2^{a_i},T_2^{a_i})$,…构成,其中 V 为 Hash 散列的关键码 Key,值为属性 a_i 对应离散值,即属性 a_i 值域上的每一个取值,设属性 a_i 第 i 个取值为 $V_l^{a_i}$,则其对应的值 $T_l^{a_i}$ 为对应的数据块集合 S_i,集合中每一个数据块 S_i 满足条件:存在对象 x,$x \in S_t$,$f(x,a_i)=V_l^{a_i}$. $\forall c_j \in C$ 的构造与此类似.

由于基于增-删窗口的知识系统更新中,增-删窗口的对象是整体进行更新的,设属性集 $B \subseteq C$,删窗口为 S_0,增窗口为 S_n,更新对象为 x,两个窗口的对象的区分情况可以按如下方法判断:

定义 10　增-删窗口区分情况判断方法:对删窗口 S_0 来说,$\forall x \in S_0$,对每一个 $a_i \in B$,设 $v=f(x,a_i)$,检索 a_i 对应的 Hash 散列的元素(v,S_t),若 S_t 中仅包含数据块 S_0,则称 x 被 B 所区分,若所有 $a_i \in B$ 都不存在这样的 S_t,则 x 不能被 B 所区分;对增窗口 S_n 来说:对 $\forall x \in S_n$,对每一个 $a_i \in B$,设 $v=f(x,a_i)$,检索 a_i 对应的 Hash 散列的元素(v,S_t),若 S_t 为空集,则称 x 被 B 所区分,若所有 $a_i \in B$ 都不存在这样的 S_t,则 x 不能被 B 所区分.

显然,根据区分检索表查询对象 x 与 B、C、D 的区分情况只需对 x 在每个属性上的取值进行查表即可,由于 Hash 散列的特点,每个属性值时间开销为 1,对 x 而言查询过程时间开销为 $O(n+m)$,$n=|C|$ 和 $m=|D|$,于是可知对增窗口和删窗口对应的对象集整体而言,时间开销是一个固定值 $O((n+m)*N)$,N 为窗口内对象的总数.

为了解决 X_q^B 与 Y_p 的判断以及属性约简时的条件熵的计算简化问题,可以利用增-删双窗口知识系统将数据空间水平划分为不同数据块结构特点,设 $U|B$ 的等价类为$\{X_1^B,X_2^B,\cdots,X_g^B\}$,则对每一个等价类集合 $X_i^B(i=1\ \text{to}\ g)$,其在数据块 $S_j(j=1\ \text{to}\ n)$上的对应子集为 $X_{i,j}^B$,$U|C$、$U|D$ 等价类的处理类似,于是可将 $U|B$,$U|C$ 和 $U|D$ 的等价类构成由 3 张子表组成的等价类数据表,结构如下:

表 2　等价类数据表

Table 2　Equivalence class

$U\|B$ 等价类子表					$U\|C$ 等价类子表					$U\|D$ 等价类子表				
S_0	S_1	…	S_n	$\sum$	S_0	S_1	…	S_n	$\sum$	S_0	S_1	…	S_n	$\sum$
$X_{1,0}^B$	$X_{1,1}^B$	…	$X_{1,n}^B$	X_1^B	$X_{1,0}^C$	$X_{1,1}^C$	…	$X_{1,n}^C$	X_1^C	$Y_{1,0}^D$	$Y_{1,1}^D$	…	$Y_{1,n}^D$	Y_1^D
$X_{2,0}^B$	$X_{2,1}^B$	…	$X_{2,n}^B$	X_2^B	$X_{2,0}^C$	$X_{2,1}^C$	…	$X_{2,n}^C$	X_2^C	$Y_{2,0}^D$	$Y_{2,1}^D$	…	$Y_{2,n}^D$	Y_2^D
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
$X_{g,0}^B$	$X_{g,1}^B$	…	$X_{g,n}^B$	X_g^B	$X_{h,0}^C$	$X_{h,1}^C$	…	$X_{h,n}^C$	X_h^C	$Y_{k,0}^D$	$Y_{k,1}^D$	…	$Y_{k,n}^D$	Y_k^D

于是可得如下 X_q^B 与 Y_p 的判断:

定义 11　增-删窗口 X_q^B 与 Y_p 判断方法:对删窗口 S_0 来说,设 $\sum$ 对应的已删除 S_0 后的决策表 $U|B$ 等价类,$\forall x \in S_0$,检索 $U|B$ 等价类子表中 S_0 下包含 x 的子集 $X_{i,0}^B$,于是找到对应的 X_i^B,令 $X_q^B=X_i^B \cup \{x\}$,检索 $U|D$ 等价类子表中 S_0 下包含 x 的子集 $Y_{j,0}^D$,找到对应的 Y_j^D,令 $Y_p=Y_j^D \cup \{x\}$,比较 X_q^B 与 Y_p,若两个集合相等,则 $X_q^B=Y_p$,反之则 $X_q^B \neq Y_p$;对增窗口 S_n 来说,设对应的已添加 S_n 后的决策表 $U|B$ 等价类,$\forall x \in S_n$,检索 $U|B$ 等价类子表中 S_n 下包含 x 的子集 $X_{i,0}^B$,于是找到对应的 X_i^B,令 $X_q^B=X_i^B-\{x\}$,检索 $U|D$ 等价类子表中 S_n 下包含 x 的子集 $Y_{j,0}^D$,找到对应的 Y_j^D,令 $Y_p=Y_j^D-\{x\}$,比较 X_q^B 与 Y_p,若两个集合相等,则 $X_q^B=Y_p$,反之则 $X_q^B \neq Y_p$.

由于 $\sum$ 列已经计算完每个等价类元素,计算条件熵及属性重要度时可以直接查找等价类数据表,其时间开销为常数,而 $U|C$、$U|D$ 子表的更新只涉及 S_0、S_n 列,改动是无需重新读取全部数据,仅 $U|B$ 的更新随属性约简改变而改动,需要读取整个数据空间 U,因此从整体上来看,可以极大减少计算属性重要度和条件熵时的时间开销.

根据增-删数据后条件熵的变化机制,基于区分情况检索表与等价类数据表的属性约简进化算法如下:

算法 1　基于增-删窗口的属性约简进化算法

输入:决策表 $S=\{U,C\cup D,V,f\}$,知识系统数据空间分为 $S_0,S_1,\cdots,S_n$,$U=S_0\cup S_1\cup\cdots\cup S_{n-1}$,$S_0$ 为删窗口,S_n 为增窗口,U 的属性约简集为 $B=\mathrm{Red}(U)$以及相应的区分情况检索表和等价类数据表

输出:更新知识系统数据空间 $U'=S_1\cup S_2\cup\cdots\cup S_n$ 后的属性约简集 $\mathrm{Red}(U')$

(1)初始化,消除 S_0 与 S_n 中的交集元素,令 $U'=S_1\cup S_2\cup\cdots\cup S_{n-1}$,更新等价类数据表

(2)处理删窗口,对每一元素 $x\in S_0$,按定义 10 检查 x 是否被 B、C 和 D 所区分,汇总结果,如果所有 x,B 与 C 同时可区分或者同时不可区分,go to(6)

(3)若存在不可被 B 区分而被 C 区分,对所有这样的检查,存在可被 D 区分,goto(5)

(4)对所有不可被 B 区分,可被 C 区分,不可被 D 区分的 x 按定义 11 检查 X_q^B 与 Y_p,若所有 x 满足 $X_q^B=Y_p$,go to(6)

(5)While $H_{U'}(D|B)==H_{U'}(D|C)$

取属性 $\alpha\in B$,且 $\mathrm{Sig}(a)$为 B 中所有属性重要度的最小值,$B=B-\{a\}$

(6)令 $U'=S_1\cup S_2\cup\cdots\cup S_n$,更新等价类数据表,更新区分情况检索表

(7)处理增窗口,对每一元素 $x\in S_n$,基于区分情况检索表按定义 10 检查 x 是否被 B、C 和 D 所区分,汇总结果,如果所有 x,B 与 C 同时可区分或者同时不可区分,go to(11)

(8)若存在 x 不可被 B 区分而被 C 区分,对所有这样的 x 检查,存在 x 可被 D 区分,go to(10)

(9)对所有不可被 B 区分,可被 C 区分,不可被 D 区分的 x 按定义 11 检查 X_q^B 与 Y_p,若所有 x 满足 $X_q^B=Y_p$,go to(11)

(10)While $H_{U'}(D|B)==H_{U'}(D|C)$

取属性 $a\in(C-B)$,且 $\mathrm{Sig}(a)$为 $C-B$ 中所有属性重要度的的最大值,$B=B\cup\{a\}$

(11)更新等价类数据表,返回 B 作为决策表 $S=\{U',C\cup D,V,f\}$ 的属性约简集 $\mathrm{Red}(U)$

算法分析:基于增-删窗口的属性约简进化算法中,步骤(2)、(3)以及(7)、(8)实际上是根据 S_0、S_n 中每一个对象对区分情况表进行一次遍历,设数据块的对象数量为常数 N,则其时间复杂度为 $N*(n+m)$,即 $O(C)$级别,步骤(4)与(9)则是对等价类数据表的操作,其时间最大开销 $N*2$,步骤(5)、(10)则是计算属性重要度,也是基于等价类数据表,根据文献[10],其时间开销为 $O(2|C|*|U|)$,但由于属性约简更新仅仅在满足条件下才进行操作,并且还能充分利用原有的属性约简 B,因此算法时间开销要小得多.

4　实验与分析

实验采用自行开发的粗糙集属性约简系统并结合 Matlab 程序来完成,为比较本文方法的优劣,引入基于基本粗糙集约简方法的对照方法 1[11] 和仅针对增数据进行进化约简的对照方法 2[10] 作为参照. 使用的数据则为项目组前期工作的用户网络行为评估数据集[4],本文分别从条件属性数量以及和每次更新后的数据数量两个方面来衡量评估方法的性能,具体实验过程为:

实验 1　衡量不同数量的条件属性情况下,本文约简方法与对照方法在计算时间方面的性能比较,实验 1 的设置情况如表 3 所示.

表 3　实验 1 设置情况

Table 3　Configuration of experiment 1

实验序号	数据量	更新窗口大小	条件属性数量	实验序号	数据量	更新窗口大小	条件属性数量
1	1 000 条	300	6	2	1 000 条	300	7
3	1 000 条	300	8	4	1 000 条	300	9
5	1 000 条	300	10	6	1 000 条	300	11
7	1 000 条	300	12				

实验 1 共使用 1 000 条数据,更新窗口大小为 300,分别进行 7 次计算,每次分别选择选取 6~12 个条件属性,重复 10 次约简并取计算时间平均值,实验结果如图 2 所示. 从实验 1 的结果可以看出,随着条件属性数量的增大,评估所耗费的计算时间有所上升,但相比两个对照方法,本文方法的上升趋势较对照方法要平缓,说明本文的属性约简进化方法从条件属性数量方面来衡量的话,在时间复杂度上有较好的优势.

实验 2　衡量不同数据量的情况下,本文约简方法与对照方法在计算时间方面的性能比较,实验 2 的设置情况如表 4 所示.

表 4　实验 2 设置情况

Table 4　Configuration of experiment 2

实验序号	数据量	更新窗口大小	条件属性数量	实验序号	数据量	更新窗口大小	条件属性数量
1	1 000 条	300	6	2	2 000 条	600	6
3	3 000 条	900	6	4	4 000 条	1200	6
5	5 000 条	1500	6	6	6 000 条	1800	6
7	7 000 条	2100	6	8	8 000 条	2400	6

实验 2 基于 6 个条件属性,更新窗口大小设为 30%,进行 8 次实验,每次分别基于 1 000 条~8 000 条数据抽取一定动态增减的数据作为分析样本,进行 10 次进化约简,并取计算时间平均值,其结果如图 3 所示.

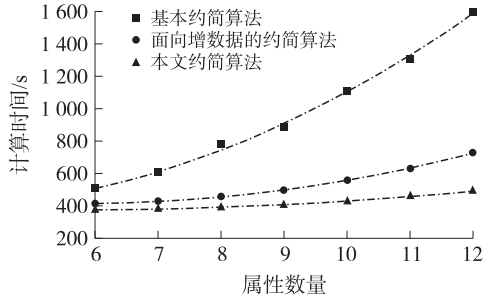

图 2　实验 1 结果

Fig. 2　Result of experiment 1

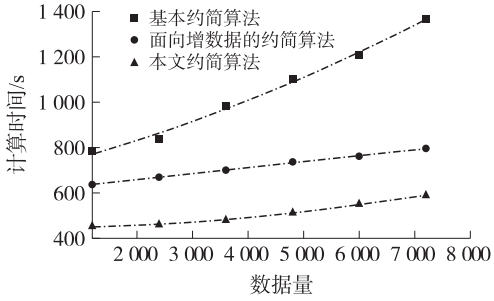

图 3　实验 2 结果

Fig. 3　Result of experiment 2

从图 3 的结果可以看出,随着计算的数据量增大,评估的计算时间开销也出现上升趋势,但相比两个对照方法,本文方法的上升趋势较对照方法要平缓,且开销均低于对照方法,说明本文的属性约简进化方法从数据规模方面来衡量,其时空复杂度上有较好的优势.

5　结论

随着粗糙集理论在实际应用中范围的增大,如何在数据信息更新情况下,快速实现属性约简已成为一个紧迫问题. 针对传统方法大多仅能处理信息增操作的约简这一不足,本文通过分析决策表对象动态增加以及删除时条件熵关于条件属性的变化机制原理,提出了一种面向增删操作的属性约简更新算法,并在哈希表基础上对其进行实现. 由于更新时能够充分利用原决策表的约简属性和条件熵信息,并借助哈希表的快速检索和匹配能力,因此可以有效减少更新属性约简的开销,实验结果证明,本文方法能够快速有效地实现动态数据更新下的决策表属性约简,其性能较传统方法有较大优势.

[参考文献]

[1]　Kim H,Claffy K,Fomenkov M,et al. Internet traffic classification demystified:myths,caveats,and the best practices[C]// Proceedings of ACM CoNEXT'08. New York,2008:1-12.

[2]　Zhang D,Qiu J,Li X. Attribute reduction based on equivalence classes with multiple decision values in rough set[C]//Proceedings of the International Conference on Information Engineering and Applications(IEA)2012. London:Springer,2013:505-512.

[3]　Jia X,Tang Z,Liao W,et al. On an optimization representation of decision-theoretic rough set model[J]. International Journal of Approximate Reasoning,2014,55(1):156-166.

[4]　陆悠,罗军舟,李伟,等. 面向网络状态的自适应用户行为评估方法[J]. 通信学报,2013(7):71-80.

[5]　钱文彬,杨炳儒,徐章艳. 基于信息熵的核属性增量式高效更新算法[J]. 模式识别与人工智能,2013,26(1):42-49.

[6]　Janusz A,Ślęzak D. Rough set methods for attribute clustering and selection[J]. Applied Artificial Intelligence,2014,28(3):220-242.

(下转第 65 页)

[19] Tang G,Xia Y,Zhang M,et al. 2011 CLGVSM:adapting generalized vector space model to cross-lingual document clustering[C]// Proc of IJCNLP,Hainan Island:Springer,2010:578-588.
[20] Steinbach M,Karypis G,Kumar V. A comparison of document clustering techniques[C]//KDD Workshop on Text Mining. Boston,2000:368-503.
[21] 楼佳. 中文文本聚类的评价与改进研究[D]. 杭州:杭州电子科技大学计算机学院,2009.
[22] 刘远超,王晓龙,徐志明. 文档聚类综述[J]. 中文信息学报,2005,20(3):57-61.
[23] 周昭涛. 文本聚类分析效果评价及文本表示研究[D]. 北京:中国科学院计算技术研究所,2005.
[24] 李勇,张克亮,李伟刚. 基于微博的网络舆情分析系统设计[J]. 计算机技术与自动化,2013,32(2):123-127.
[25] 时睿,面向短文本的网络舆情分析[D]. 西安:西安电子科技大学电子工程学院,2012.
[26] 陈雅菊,现代汉语词语搭配的自动抽取方法[D]. 上海:华东师范大学软件学院,2005.

[责任编辑:顾晓天]

(上接第 56 页)

[7] Thangavel K,Pethalakshmi A. Dimensionality reduction based on rough set theory:a review[J]. Applied Soft Computing,2009,9(1):1-12.
[8] 林俊伟,叶东毅. 基于领域辨识矩阵的属性约简增量式算法[J]. 计算机应用,2009,29(11):119-121
[9] Hu F,Wang G Y,Huang H,et al. Incremental attribute reduction based on element arysets[C]//Proceedings of the 10th International Conference on Rough Sets,Fuzzy Sets,Data Mining,and Granular Computing. Regina,2005:183-193
[10] 梁吉业,魏巍,钱宇华. 一种基于条件熵的增量核求解方法[J]. 系统工程理论与实践,2008,28(4):81-89
[11] Guoyin W,Yiyu Y,Hong Y. A survey on rough set theory and applications[J]. Chinese Journal of Computers,2009,32(7):1 229-1 246.
[12] Yu H,Liu Z,Wang G. An automatic method to determine the number of clusters using decision-theoretic rough set[J]. International Journal of Approximate Reasoning,2014,55(1):101-115.
[13] Jia X,Liao W,Tang Z,et al. Minimum cost attribute reduction in decision-theoretic rough set models[J]. Information Sciences,2013,219:151-167.
[14] Chen H,Li T,Ruan D,et al. A rough-set-based incremental approach for updating approximations under dynamic maintenance environments[J]. IEEE Transactions on Knowledge and Data Engineering,2013,25(2):274-284.

[责任编辑:顾晓天]