基于矩阵方法的区分度增量式属性约简算法

沈 玉 峰
(安徽三联学院计算机工程学院 安徽 合肥 230601)

摘 要 属性约简是粗糙集理论在机器学习领域的重要应用。通过对传统的属性约简算法构造增量式学习,从而提高了动态数据环境下的属性约简性能。区分度作为一种重要的属性集评估函数,目前已成功地运用于属性约简的构造。在基于区分度属性约简的基础上,提出一种区分度的增量式属性约简算法。通过矩阵的方法去表示区分度度量,在矩阵表示的基础上,进一步研究信息系统对象增加和减少时区分度的增量式学习,根据这种增量式学习提出对应的增量式属性约简算法。UCI 数据集的实验结果表明,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动态数据属性约简性能。

关键词 属性约简 增量式学习 信息系统 动态变化 矩阵

中图分类号 TP18 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2020.09.038

A DISCRIMINATION DEGREE INCREMENTAL ATTRIBUTE REDUCTION BASED ON MATRIX METHOD

Shen Yufeng
(*College of Computer Engineering, Anhui Sanlian University, Hefei 230601, Anhui, China*)

Abstract Attribute reduction is an important application of rough set theory in machine learning. By constructing incremental learning for traditional attribute reduction algorithm, the performance of attribute reduction in dynamic data environment is improved. As an important attribute set evaluation function, discrimination degree has been successfully applied to attribute reduction. This paper proposes an incremental attribute reduction algorithm based on discriminant degree. The matrix method was used to represent the discrimination degree. Then, on the basis of the matrix representation, I further studied the incremental learning of the discrimination degree when the object of information system increased or decreased. Through the incremental learning, the corresponding incremental attribute reduction algorithm was proposed. The experimental results on UCI datasets show that my algorithm has better performance than the non-incremental algorithm in dynamic data attribute reduction.

Keywords Attribute reduction Incremental learning Information system Dynamic change Matrix

0 引 言

粗糙集[1]是人工智能领域的重要分支,属性约简[2]是粗糙集理论的重点研究内容。属性约简的目的是为了将原始信息系统的冗余属性进行甄别和删除,从而提高数据集的知识发现性能。然而随着信息技术的发展以及数据采集技术的提高,实际应用环境下的数据总是时刻处于动态更新之中,传统的各种属性约简算法是针对静态的数据设计的,而对于动态的数据集,这些算法的处理效率较为低下,不能很好地适应实际的工程需求[3-4]。

为了改善动态数据下的属性约简性能,学者们提出了一种改进的属性约简方法——增量式属性约简[5],增量式属性约简的核心思想是增量式学习,即在原始属性约简的基础上融入增量式学习,当信息系统发生更新时,增量式属性约简算法能够将原来的属性约简结果加以利用,在其基础上进一步计算出新的属

收稿日期:2019-05-21。安徽省重点科研项目(KJ2019A0903)。**沈玉峰**,讲师,主研领域:粒计算,数据挖掘。

性约简结果。由于增量式属性约简的高效性,目前已成为属性约简领域的研究热点[6-7]。

对象的动态增加和减少是信息系统最为常见的一种动态更新形式,为此学者对这类属性约简问题进行了大量的研究。最早的属性约简是基于粗糙集的正区域来构造,Chan[5]研究了正区域随信息系统变化时的增量式更新,提出了最早的增量式属性约简算法。在 Chan 的研究思路基础上,Shu 等[8]提出了一种改进的增量式属性约简,进一步提高了动态数据的处理效率。同时,Wei 等[9]通过辨析矩阵的视角给出了一种新的增量式属性约简算法。信息熵作为一种新的属性集不确定性度量,是构造属性约简的一种常用方法[10-12],因此学者们在信息熵的基础上进一步地提出了多种的增量式属性约简算法。Liang 等[13]研究了条件信息熵随信息系统对象变化时的增量式更新,并基于这种更新机制提出了条件信息熵的增量式属性约简算法。类似于这种增量式更新的方法,赵小龙等[14]提出了数值型数据下条件信息熵的增量式更新,并提出了对应的增量式属性约简,Jing 等[15]采用同样的推导思路,研究了粒计算中知识粒度随对象变化时的增量式更新,提出了基于知识粒度的增量式属性约简算法。同样在不完备信息系统中,Xie 等[16]基于一致性度量的增量式更新方法设计了一种类似的增量式属性约简算法。综合已有的研究成果可以看出,对属性集度量函数进行增量式学习的构造是目前增量式属性约简的研究重点。

信息系统属性集的区分度度量是 Teng 等[17]提出的一种新的度量方法,它通过信息系统的等价类直接进行计算,能够更加精准地评估出属性集之间的依赖关系,并且具有时间复杂度低的优越性,Teng 等通过实验证明了基于区分度属性约简的高效性和优越性。然而目前还未有利用区分度度量进行增量式属性约简的研究,因此这促使我们进行相关的研究和探索。

研究区分度的增量式学习是设计区分度增量式属性约简的关键。矩阵是一种重要的数据处理工具,由于它在计算方面的可扩展性,目前已广泛地运用于粗糙集的各类增量式学习之中[9,18-20]。本文将通过矩阵的方法去构造区分度的增量式学习,进一步提出相应的增量式属性约简算法。首先运用矩阵的形式去表示信息系统的区分度度量,然后在矩阵表示的基础上,分别研究了信息系统对象增加和减少时区分度的增量式学习,理论分析表明了这种增量式学习的高效性,它可以快速地更新出区分度结果,最后基于区分度的增量式学习提出对应的增量式属性约简算法。UCI 数据集的实验结果表明,所提出的增量式属性约简算法具有更高的动态数据属性约简性能,能够适应数据动态变化时的属性约简。

1 基于区分度的属性约简

属性约简[1-2]是粗糙集理论的重要应用。在属性约简中,所讨论的数据集被描述成信息系统的形式。通常一个信息系统表示为 $S=(U,AT)$,其中 U 为该信息系统的对象集,即数据集所有样本的集合;AT 为该信息系统的属性集,即数据集所有特征的集合。若信息系统包含决策属性 D,即每个对象都有一个类别的标记,那么该信息系统又称为决策信息系统[1]。

对于信息系统 $S=(U,AT)$,设属性子集 $A\subseteq AT$ 在论域 $U\times U$ 下确定的等价关系为:

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

等价关系 $E_{U\times U}(A)$ 满足自反性、对称性和传递性。若 $U=\{x_1,x_2,\cdots,x_n\}$,$E_{U\times U}(A)$ 可以将论域 U 诱导出一组划分,表示为 $U/E_{U\times U}(A)=\{[x_1]_A,[x_2]_A,\cdots,[x_n]_A\}$,其中 $[x_i]_A(1\leqslant i\leqslant n)$ 为对象 x_i 在等价关系 $E_{U\times U}(A)$ 下的等价类,表示为 $[x_i]_A=\{x_j\in U\mid (x_i,x_j)\in E_{U\times U}(A)\}$。

通过等价关系将信息系统的论域进行信息粒化,其粒化后的粒子即为等价类,最后基于等价类去诱导粗糙集的上下近似,并且选择出信息系统的属性约简[1-2,10]。在文献[17]中,Teng 等提出了一种基于区分度的属性约简算法。

定义 1[17] 设信息系统 $S=(U,AT)$,$|U|=n$,属性集 $A\subseteq AT$ 在论域 $U\times U$ 下确定的等价关系为 $E_{U\times U}(A)$,且诱导的划分为 $U/E_{U\times U}(A)$。定义信息系统论域 U 下属性集 A 的区分度为:

$$Dis_U(A)=|U|^2-\sum_{i=1}^{n}|[x_i]_A| \tag{2}$$

区分度满足 $0\leqslant Dis_U(A)\leqslant |U|^2-|U|$,对于论域 U 的划分 $U/E_{U\times U}(A)$,当 $\forall x\in U,[x]_A=U$ 时,$Dis_U(A)=0$;当 $\forall x\in U,[x]_A=x$ 时,$Dis_U(A)=|U|^2-|U|$。

知识是粗糙集和粒计算的研究核心,在粗糙集理论中,同一个等价类中对象之间不具有知识的区分性,相反,不同等价类之间则表现出了知识的区分性。因此在信息系统中,给定属性集下的知识量可以通过不同等价类之间对象的数量来表示,即定义 1 中的区分度度量[17]。在区分度的基础上,Teng 等进一步提出了一个属性集相对另一个属性的知识量,称之为相对区

分度,具体如定义 2 所示。

定义 2[17]　设信息系统 $S=(U,AT)$,属性集 A_1,$A_2\subseteq AT$ 在论域 $U\times U$ 下确定的等价关系分别为 $E_{U\times U}(A_1)$ 和 $E_{U\times U}(A_2)$,对论域 U 构造出的划分分别为 $U/E_{U\times U}(A_1)$ 和 $U/E_{U\times U}(A_2)$。定义论域 U 下属性集 A_2 关于 A_1 的相对区分度为:

$$Dis_U(A_2\mid A_1)=\sum_{i=1}^{n}\left|[x_i]_{A_1}\right|-\sum_{i=1}^{n}\left|[x_i]_{A_2}\cap[x_i]_{A_1}\right| \tag{3}$$

类似于定义 1,相对区分度同样满足关系 $0\leqslant Dis_U(A_2\mid A_1)\leqslant|U|^2-|U|$。

定义 2 中的相对区分度可以看作是一种属性集之间关系程度的度量。因此在文献[17]中,利用相对区分度作为评估属性集的启发式函数,Teng 等定义了一种新的属性约简方法。

定义 3[17]　设决策信息系统 $S=(U,C\cup D)$,其中 C 为条件属性集,D 为决策属性集。若属性集 $red\subseteq C$ 是该信息系统的一个属性约简,那么当且仅当如下同时成立:

$$Dis_U(D\mid C)=Dis_U(D\mid red) \tag{4}$$

$$a\in red\quad Dis_U(D\mid C)<Dis_U(D\mid red-\{a\}) \tag{5}$$

算法 1 所示的是对应的属性约简算法。

算法 1[17]　基于区分度的属性约简算法

输入:决策信息系统 $S=(U,C\cup D)$。

输出:属性约简结果 red。

Step 1　初始化属性约简集 $red=\varnothing$;

Step 2　对于 $\forall a\in C\text{-}red$,计算属性 a 的属性重要度:

$sig_{red}(a)=Dis_U(D\mid red)-Dis_U(D\mid red\cup\{a\})$

Step 3　找出 $C\text{-}red$ 中属性重要度最大的属性,记为 a';

Step 4　若 $sig_{red}(a')>0$,那么 $red=red\cup\{a'\}$,并跳转入 Step 2,若 $sig_{red}(a')=0$,则跳转入 Step 5;

Step 5　返回属性约简结果 red。

算法 1 的算法复杂度主要集中在 Step 2 – Step 4,该步骤主要通过 $sig_{red}(\cdot)$ 作为启发式函数对信息系统的候选属性进行启发式搜索,通过不断迭代的方式完成最终属性约简结果的选择。根据文献[17],算法 1 的时间复杂度为 $O(|C|^2\cdot|U|^2)$。

2　基于矩阵方法的区分度增量式更新

矩阵是一种重要的数据表达形式,在粗糙集理论中,学者们通过矩阵对粗糙集中各类计算模型进行重构,提出了多种形式的模型和算法[9,18-20]。在本节,将在前人研究的基础上,通过矩阵的方法去表示区分度,并进一步通过矩阵去构造区分度的增量式更新。

2.1　区分度的矩阵表达

定义 4[18]　设信息系统 $S=(U,AT)$,$|U|=n$,属性集 $A\subseteq AT$ 在论域 $U\times U$ 下确定的等价关系为 $E_{U\times U}(A)$,定义等价关系 $E_{U\times U}(A)$ 的关系矩阵为:

$$\boldsymbol{M}^A_{n\times n}=(m_{ij})_{n\times n} \tag{6}$$

式中:$m_{ij}=\begin{cases}1 & (x_i,x_j)\in E_{U\times U}(A)\\ 0 & (x_i,x_j)\notin E_{U\times U}(A)\end{cases}$　$1\leqslant i,j\leqslant n$。

定义 4 是通过矩阵的形式对等价关系进行表达,若信息系统论域中对象 x_i 和 x_j 满足等价关系,那么区分度关系矩阵中第 i 行第 j 列元素为 1,否则为 0。

性质 1　设信息系统 $S=(U,AT)$,$|U|=n$ 且属性集 $A\subseteq AT$,等价关系 $E_{U\times U}(A)$ 对应的关系矩阵 $\boldsymbol{M}^A_{n\times n}$ 满足:

$$\boldsymbol{M}^A_{n\times n}=(\boldsymbol{M}^A_{n\times n})^{\mathrm{T}} \tag{7}$$

$$\boldsymbol{M}^A_{n\times n}=(m_{ij})_{n\times n}\quad m_{ii}=1\quad 1\leqslant i\leqslant n \tag{8}$$

式中:$(\boldsymbol{M}^A_{n\times n})^{\mathrm{T}}$ 表示 $\boldsymbol{M}^A_{n\times n}$ 的转置。

证明　等价关系满足对称性,即对于 $\forall x,y\in U$,若 $(x,y)\in E_{U\times U}(A)$ 则必有 $(y,x)\in E_{U\times U}(A)$,若 $(x,y)\notin E_{U\times U}(A)$ 则必有 $(y,x)\notin E_{U\times U}(A)$。因此 $\boldsymbol{M}^A_{n\times n}=(\boldsymbol{M}^A_{n\times n})^{\mathrm{T}}$,也就是说关系矩阵 $\boldsymbol{M}^A_{n\times n}$ 是对称矩阵。

等价关系满足自反性,即对于 $\forall x\in U$ 都有 $(x,x)\in E_{U\times U}(A)$,因此 $m_{ii}=1$,也就是说关系矩阵 $\boldsymbol{M}^A_{n\times n}$ 主对角线上的元素都为 1。

证毕。

定理 1　设信息系统 $S=(U,AT)$,$|U|=n$,属性集 $A\subseteq AT$ 在论域 $U\times U$ 下确定的等价关系为 $E_{U\times U}(A)$,对应的关系矩阵为 $\boldsymbol{M}^A_{n\times n}$。那么属性集 A 下的区分度可表示为:

$$Dis_U(A)=n^2-|\boldsymbol{M}^A_{n\times n}| \tag{9}$$

式中:$|M^A_{n\times n}|=\sum_{i=1}^{n}\sum_{j=1}^{n}m_{ij}$,即矩阵 $\boldsymbol{M}^A_{n\times n}$ 所有元素之和。

证明　对于 $x_i\in U$,x_i 的等价类表示为 $[x_i]_A=\{x_j\in U\mid(x_i,x_j)\in E_{U\times U}(A)\}$。根据定义 4,$|[x_i]_A|$ 为矩阵 $\boldsymbol{M}^A_{n\times n}$ 第 i 行所有元素之和,即 $|[x_i]_A|=\sum_{j=1}^{n}m_{ij}$。所以根据定义 1 有:

$$Dis_U(A)=|U|^2-\sum_{i=1}^{n}\left|[x_i]_A\right|=n^2-\sum_{i=1}^{n}\sum_{j=1}^{n}m_{ij}=n^2-|\boldsymbol{M}^A_{n\times n}|$$

证毕。

定理 1 展示了区分度的另一种表示方式,即通过区分度关系矩阵的方法来表示,而不必对信息系统中

每个对象的等价类进行计算。

例 1 表 1 所示的是一个信息系统,其中$\{a,b,c\}$为该信息系统的属性集,$x_1,x_2,\cdots,x_6$为该信息系统的 6 个对象。

表 1 信息系统

U	a	b	c
x_1	0	1	Y
x_2	1	2	Y
x_3	1	1	Y
x_4	1	2	N
x_5	0	1	N
x_6	0	1	Y

令属性集$A=\{a,b\}$,对 A 构建等价关系 $E_{U\times U}(A)$,那么可以得到每个对象的等价类为:

$$[x_1]_A=[x_5]_A=[x_6]_A=\{x_1,x_5,x_6\}$$

$$[x_2]_A=[x_4]_A=\{x_2,x_4\}\quad [x_3]_A=\{x_3\}$$

那么根据定义 1 关于区分度的定义,可以得到:

$$Dis_U(A)=|U|^2-\sum_{i=1}^{6}|[x_i]_A|=36-3\times3-2\times2-1=22$$

基于定理 1 的方法进行计算:

$$\boldsymbol{M}^A_{6\times6}=\begin{bmatrix}1&0&0&0&1&1\\0&1&0&1&0&0\\0&0&1&0&0&0\\0&1&0&1&0&0\\1&0&0&0&1&1\\1&0&0&0&1&1\end{bmatrix}$$

那么 $Dis_U(A)=6^2-|\boldsymbol{M}^A_{n\times n}|=22$。

对比可以看出两种计算结果是一致的。

在定理 1 的基础上,接下来将通过矩阵进一步表示相对区分度。

定理 2 设信息系统 $S=(U,AT)$,属性集 $A_1,A_2\subseteq AT$ 确定的等价关系分别为 $E_{U\times U}(A_1)$ 和 $E_{U\times U}(A_2)$,对应的关系矩阵分别为 $\boldsymbol{M}^{A_1}_{n\times n}$ 和 $\boldsymbol{M}^{A_2}_{n\times n}$,那么属性集 A_2 关于 A_1 的相对区分度可表示为:

$$Dis_U(A_2\mid A_1)=Dis_U(A_1)-Dis_U(A_1\cup A_2)=|\boldsymbol{M}^{A_1}_{n\times n}|-|\boldsymbol{M}^{A_2}_{n\times n}\wedge\boldsymbol{M}^{A_1}_{n\times n}| \tag{10}$$

式中:

$$\boldsymbol{M}^{A_2}_{n\times n}\wedge\boldsymbol{M}^{A_1}_{n\times n}=(m_{ij})_{n\times n}=\begin{cases}1 & m^{A_1}_{ij}=1,m^{A_2}_{ij}=1\\0 & \text{其他}\end{cases}\quad 1\leqslant i,j\leqslant n$$

证明 根据定理 1 可以直接得到。

类似于定理 1,定理 2 利用矩阵的方法表示相对区分度。

证毕。

例 2 设信息系统如表 1 所示,令属性集 $A_1=\{a,b\}$,$A_2=\{c\}$,那么:

$$[x_1]_{A_2}=[x_2]_{A_2}=[x_3]_{A_2}=[x_6]_{A_2}=\{x_1,x_2,x_3,x_6\}$$

$$[x_4]_{A_2}=[x_5]_{A_2}=\{x_4,x_5\}$$

根据定义 2 有:

$$Dis_U(A_2\mid A_1)=14-2-1-1-1-1-2=6$$

由于 $\boldsymbol{M}^{A_2}_{6\times6}=\begin{bmatrix}1&1&1&0&0&1\\1&1&1&0&0&1\\1&1&1&0&0&1\\0&0&0&1&1&0\\0&0&0&1&1&0\\1&1&1&0&0&1\end{bmatrix}$,$\boldsymbol{M}^{A_1}_{6\times6}$为例 1 中的 $\boldsymbol{M}^A_{6\times6}$,那么:

$$\boldsymbol{M}^{A_2}_{n\times n}\wedge\boldsymbol{M}^{A_1}_{n\times n}=\begin{bmatrix}1&0&0&0&0&1\\0&1&0&0&0&0\\0&0&1&0&0&0\\0&0&0&1&0&0\\0&0&0&0&1&0\\1&0&0&0&0&1\end{bmatrix}$$

因此 $Dis_U(A_2\mid A_1)=|\boldsymbol{M}^{A_1}_{n\times n}|-|\boldsymbol{M}^{A_2}_{n\times n}\wedge\boldsymbol{M}^{A_1}_{n\times n}|=14-8=6$。

两种计算结果是一致的。

2.2 信息系统对象变化时区分度的增量式更新

由于现实应用环境下,信息系统往往都是不断动态变化的。本节将通过矩阵方法去研究信息系统对象发生变化时,区分度的增量式更新,其中包含对象增加时区分度的增量式更新和对象减少时区分度的增量式更新。

2.2.1 信息系统对象增加时区分度的增量式更新

设信息系统 $S=(U,AT)$,其中论域 $U=\{x_1,x_2,\cdots,x_n\}$,属性集 $A\subseteq AT$ 在论域 U 下确定的等价关系记为 $E_{U\times U}(A)$,当信息系统增加对象集 $U^+=\{x_{n+1},x_{n+2},\cdots,x_{n+k}\}$ 后,新的信息系统记为 $S=(U'=U\cup U^+,AT)$,属性集 $A\subseteq AT$ 在论域 U' 下确定的等价关系记为 $E_{U'\times U'}(A)$。

定义 5 信息系统 $S=(U,AT)$ 增加对象集 U^+ 后更新为 $S=(U'=U\cup U^+,AT)$。设对象集 U^+ 与论域 U 在属性集 $A\subseteq AT$ 下确定的等价关系记为 $E_{U^+\times U}(A)$,那么对应的关系矩阵定义为:

$$P^A_{k\times n}=(p_{ij})_{k\times n}=\begin{cases}1 & (x_{n+i},x_j)\in E_{U^+\times U}(A)\\0 & (x_{n+i},x_j)\notin E_{U^+\times U}(A)\end{cases}\quad 1\leqslant i\leqslant k,1\leqslant j\leqslant n \tag{11}$$

定义 6　信息系统 $S=(U,AT)$ 增加对象集 U^+ 后更新为 $S=(U'=U\cup U^+,AT)$。设 U^+ 与 U^+ 在属性集 $A\subseteq AT$ 下确定的等价关系记为 $E_{U^+\times U^+}(A)$,那么对应的关系矩阵定义为:

$$\boldsymbol{Q}^A_{k\times k}=(q_{ij})_{k\times k}=\begin{cases}1 & (x_{n+i},x_{n+j})\in E_{U^+\times U^+}(A)\\ 0 & (x_{n+i},x_{n+j})\notin E_{U^+\times U^+}(A)\end{cases}\quad 1\leqslant i,j\leqslant k \tag{12}$$

根据定义 5 和定义 6,接下来可以增量式地得到信息系统 $S=(U'=U\cup U^+,AT)$ 下等价关系 $E_{U'\times U'}(A)$ 所对应的关系矩阵。具体如定理 3 所示。

定理 3　信息系统 $S=(U,AT)$ 增加对象集 U^+ 后更新为 $S=(U'=U\cup U^+,AT)$。对象集 U^+ 与论域 U 在属性集 $A\subseteq AT$ 下确定的等价关系为 $E_{U^+\times U}(A)$,对应的关系矩阵为 $\boldsymbol{P}^A_{k\times n}$;对象集 U^+ 与 U^+ 确定的等价关系为 $E_{U^+\times U^+}(A)$,对应的关系矩阵为 $\boldsymbol{Q}^A_{k\times k}$。那么等价关系 $E_{U'\times U'}(A)$ 对应的关系矩阵可表示为:

$$\boldsymbol{M}^A_{(n+k)\times(n+k)}=\begin{bmatrix}\boldsymbol{M}^A_{n\times n} & (\boldsymbol{P}^A_{k\times n})^{\mathrm{T}}\\ \boldsymbol{P}^A_{k\times n} & \boldsymbol{Q}^A_{k\times k}\end{bmatrix} \tag{13}$$

证明　根据等价关系的定义,有 $\forall x_i,x_j\in U$,若 $(x_i,x_j)\in E_{U\times U}(A)$,则必有 $(x_i,x_j)\in E_{U'\times U'}(A)$,并且 $(x_i,x_j)\notin E_{U\times U}(A)$ 也必有 $(x_i,x_j)\notin E_{U'\times U'}(A)$。因此矩阵 $\boldsymbol{M}^A_{(n+k)\times(n+k)}$ 前 n 行 n 列即为矩阵 $\boldsymbol{M}^A_{n\times n}$。

根据性质 1,可以得到 $\boldsymbol{M}^A_{(n+k)\times(n+k)}$ 为一个对称矩阵,因此 $\boldsymbol{M}^A_{(n+k)\times(n+k)}$ 的前 n 行后 k 列为 $\boldsymbol{M}^A_{(n+k)\times(n+k)}$ 后 k 行前 n 列的转置。

综上,等价关系 $E_{U'\times U'}(A)$ 对应的关系矩阵为定理 3 所示的矩阵 $\boldsymbol{M}^A_{(n+k)\times(n+k)}$。

证毕。

当信息系统论域中的对象发生增加时,定理 3 展示了区分度关系矩阵的增量式更新,观察可以发现,只需要对增加的对象进行相关的计算,即计算区分度子矩阵 $\boldsymbol{P}^A_{k\times n}$ 和 $\boldsymbol{Q}^A_{k\times k}$,便可以完成新论域下区分度关系矩阵的计算,而不必对原来的对象进行重复计算,因此这种计算方式具有一定的高效性。

在定理 3 的基础上,可以得到论域增加对象集后区分度的增量式更新。

定理 4　信息系统 $S=(U,AT)$ 增加对象集 U^+ 后更新为 $S=(U'=U\cup U^+,AT)$。对于属性集 $A\subseteq AT$,矩阵 $\boldsymbol{M}^A_{n\times n}$、$\boldsymbol{P}^A_{k\times n}$ 和 $\boldsymbol{Q}^A_{k\times k}$ 的意义同定理 3,论域 U 在属性集 A 下的区分度为 $Dis_U(A)$,那么 U' 在属性集 A 下的区分度 $Dis_{U'}(A)$ 可增量式更新为:

$$Dis_{U'}(A)=Dis_U(A)+2kn+k^2-2\cdot|\boldsymbol{P}^A_{k\times n}|-|\boldsymbol{Q}^A_{k\times k}| \tag{14}$$

证明　根据定理 1 有 $Dis_{U'}(A)=(n+k)^2-|\boldsymbol{M}^A_{(n+k)\times(n+k)}|$

又由于定理 3,$|\boldsymbol{M}^A_{(n+k)\times(n+k)}|$ 可以表示为:

$$|\boldsymbol{M}^A_{(n+k)\times(n+k)}|=|\boldsymbol{M}^A_{n\times n}|+2\cdot|\boldsymbol{P}^A_{k\times n}|+|\boldsymbol{Q}^A_{k\times k}|$$

因此 $Dis_{U'}(A)=(n+k)^2-|\boldsymbol{M}^A_{(n+k)\times(n+k)}|=$

$$n^2+2kn+k^2-|\boldsymbol{M}^A_{n\times n}|-2\cdot|\boldsymbol{P}^A_{k\times n}|-|\boldsymbol{Q}^A_{k\times k}|$$

其中 $n^2-|\boldsymbol{M}^A_{n\times n}|=Dis_U(A)$,故 $Dis_{U'}(A)=Dis_U(A)+2kn+k^2-2\cdot|\boldsymbol{P}^A_{k\times n}|-|\boldsymbol{Q}^A_{k\times k}|$

证毕。

定理 4 表明,对于信息系统增加对象集 U^+ 后,在原来区分度 $Dis_U(A)$ 的基础上,只需计算出关系矩阵 $\boldsymbol{P}^A_{k\times n}$ 和 $\boldsymbol{Q}^A_{k\times k}$ 便可以快速地更新出区分度 $Dis_{U'}(A)$ 结果。如果采用非增量式的计算方法,需要计算关系矩阵 $\boldsymbol{M}^A_{(n+k)\times(n+k)}$,其中的 $\boldsymbol{M}^A_{n\times n}$ 相当于进行了重复计算,那么最终的计算量将大于增量式方法的计算量。

在定理 4 的基础上,可以进一步得到相对区分度的增量式更新,具体见定理 5。

定理 5　信息系统 $S=(U,AT)$ 增加对象集 U^+ 后更新为 $S=(U'=U\cup U^+,AT)$。对于属性集 $A_1,A_2\subseteq AT$,矩阵 $\boldsymbol{M}^{A_1}_{n\times n}$、$\boldsymbol{P}^{A_1}_{k\times n}$、$\boldsymbol{Q}^{A_1}_{k\times k}$、$\boldsymbol{M}^{A_2}_{n\times n}$、$\boldsymbol{P}^{A_2}_{k\times n}$ 和 $\boldsymbol{Q}^{A_2}_{k\times k}$ 意义同定理 3,论域 U 下属性集 A_2 关于 A_1 的相对区分度为 $Dis_U(A_2|A_1)$,那么论域 U' 下属性集 A_2 关于 A_1 的相对区分度 $Dis_{U'}(A_2|A_1)$ 可增量式更新为:

$$Dis_{U'}(A_2|A_1)=Dis_U(A_2|A_1)+2\cdot|\boldsymbol{P}^{A_1}_{k\times n}|+|\boldsymbol{Q}^{A_1}_{k\times k}|-2|\boldsymbol{P}^{A_2}_{k\times n}\wedge\boldsymbol{P}^{A_1}_{k\times n}|-|\boldsymbol{Q}^{A_2}_{k\times k}\wedge\boldsymbol{Q}^{A_1}_{k\times k}| \tag{15}$$

证明　根据定理 2 可以得到:

$$Dis_{U'}(A_2|A_1)=|\boldsymbol{M}^{A_1}_{(n+k)\times(n+k)}|-|\boldsymbol{M}^{A_2}_{(n+k)\times(n+k)}\wedge\boldsymbol{M}^{A_1}_{(n+k)\times(n+k)}|$$

在定理 3 中,四个子矩阵是相互独立的,因此:

$$|\boldsymbol{M}^{A_2}_{(n+k)\times(n+k)}\wedge\boldsymbol{M}^{A_1}_{(n+k)\times(n+k)}|=|\boldsymbol{M}^{A_2}_{n\times n}\wedge\boldsymbol{M}^{A_1}_{n\times n}|+2\cdot|\boldsymbol{P}^{A_2}_{k\times n}\wedge\boldsymbol{P}^{A_1}_{k\times n}|+|\boldsymbol{Q}^{A_2}_{k\times k}\wedge\boldsymbol{Q}^{A_1}_{k\times k}|$$

则:$Dis_{U'}(A_2|A_1)=|\boldsymbol{M}^{A_1}_{n\times n}|+2\cdot|\boldsymbol{P}^{A_1}_{k\times n}|+|\boldsymbol{Q}^{A_1}_{k\times k}|-|\boldsymbol{M}^{A_2}_{n\times n}\wedge\boldsymbol{M}^{A_1}_{n\times n}|-2\cdot|\boldsymbol{P}^{A_2}_{k\times n}\wedge\boldsymbol{P}^{A_1}_{k\times n}|-|\boldsymbol{Q}^{A_2}_{k\times k}\wedge\boldsymbol{Q}^{A_1}_{k\times k}|$

其中:$|\boldsymbol{M}^{A_1}_{n\times n}|-|\boldsymbol{M}^{A_2}_{n\times n}\wedge\boldsymbol{M}^{A_1}_{n\times n}|=Dis_U(A_2|A_1)$

所以:$Dis_{U'}(A_2|A_1)=Dis_U(A_2|A_1)+2\cdot|\boldsymbol{P}^{A_1}_{k\times n}|+|\boldsymbol{Q}^{A_1}_{k\times k}|-2|\boldsymbol{P}^{A_2}_{k\times n}\wedge\boldsymbol{P}^{A_1}_{k\times n}|-|\boldsymbol{Q}^{A_2}_{k\times k}\wedge\boldsymbol{Q}^{A_1}_{k\times k}|$

证毕。

类似于定理 4,对于计算对象增加后的相对区分度,定理 5 同样具有很高的计算效率。

例 3　设表 1 所示的信息系统增加对象集 $U^+=\{x_7,x_8,x_9\}$,新信息系统如表 2 所示。

表 2 新的信息系统

U'	a	b	c
x_1	0	1	Y
x_2	1	2	Y
x_3	1	1	Y
x_4	1	2	N
x_5	0	1	N
x_6	0	1	Y
x_7	1	2	Y
x_8	0	2	N
x_9	1	1	Y

设属性集 $A_1=\{a,b\}$,$A_2=\{c\}$。那么:

$$[x_1]_{A_1}=[x_5]_{A_1}=[x_6]_{A_1}=\{x_1,x_5,x_6\}$$
$$[x_2]_{A_1}=[x_4]_{A_1}=[x_7]_{A_1}=\{x_2,x_4,x_7\}$$
$$[x_3]_{A_1}=[x_9]_{A_1}=\{x_3,x_9\};[x_8]_{A_1}=\{x_8\}$$
$$[x_1]_{A_2}=[x_2]_{A_2}=[x_3]_{A_2}=[x_6]_{A_2}=$$
$$[x_7]_{A_2}=[x_9]_{A_2}=\{x_1,x_2,x_3,x_6,x_7,x_9\}$$
$$[x_4]_{A_2}=[x_5]_{A_2}=[x_8]_{A_2}=\{x_4,x_5,x_8\}$$

则:

$$Dis_{U'}(A_2\mid A_1)=\sum_{i=1}^{9}|[x_i]_{A_1}|-\sum_{i=1}^{9}|[x_i]_{A_2}\cap[x_i]_{A_1}|=23-15=8$$

采用矩阵的方法进行增量式计算:

$$\boldsymbol{P}_{3\times6}^{A_1}=\begin{bmatrix}0&1&0&1&0&0\\0&0&0&0&0&0\\0&0&1&0&0&0\end{bmatrix}$$

$$\boldsymbol{P}_{3\times6}^{A_2}=\begin{bmatrix}1&1&1&0&0&1\\0&0&0&1&1&0\\1&1&1&0&0&1\end{bmatrix}$$

$$\boldsymbol{Q}_{3\times3}^{A_1}=\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}\quad\boldsymbol{Q}_{3\times3}^{A_2}=\begin{bmatrix}1&0&1\\0&1&0\\1&0&1\end{bmatrix}$$

由于 $Dis_U(A_2\mid A_1)$在例 2 中已经计算得出,因此根据定理 5 可以得到:

$$Dis_{U'}(A_2\mid A_1)=Dis_U(A_2\mid A_1)+2\cdot|\boldsymbol{P}_{k\times n}^{A_1}|+|\boldsymbol{Q}_{k\times k}^{A_1}|-2|\boldsymbol{P}_{k\times n}^{A_2}\wedge\boldsymbol{P}_{k\times n}^{A_1}|-|\boldsymbol{Q}_{k\times k}^{A_2}\wedge\boldsymbol{Q}_{k\times k}^{A_1}|=6+6+3-4-3=8$$

两种计算结果是一致的,但是基于矩阵方法进行增量式计算,可以在原来计算结果上进行进一步计算,大幅度地减少了重复计算量,具有更高的计算效率。

2.2.2 信息系统对象减少时区分度的增量式更新

在上小节中,给出了当信息系统对象增加时区分度的增量式更新方法,本节仿照上节的研究思路,提出信息系统对象减少时区分度的增量式更新。

设信息系统 $S=(U,AT)$,其中论域 $U=\{x_1,x_2,\cdots,x_n\}$,属性集 $A\subseteq AT$ 在论域 U 下确定的等价关系记为 $E_{U\times U}(A)$,当信息系统减少对象集 $U^-=\{x_{t1},x_{t2},\cdots,x_{tk}\}$,其中 $U^-\subseteq U$,新的信息系统记为 $S=(U'=U-U^-,AT)$,属性集 $A\subseteq AT$ 在论域 U'下确定的等价关系记为 $E_{U'\times U'}(A)$。

定义 7 信息系统 $S=(U,AT)$减少对象集 U^-后更新为 $S=(U'=U-U^-,AT)$。对象集 U^-与论域 U 在属性集 $A\subseteq AT$ 下确定的等价关系记为 $E_{U^-\times U}(A)$,那么定义关系矩阵:

$$\boldsymbol{S}_{k\times n}^{A}=(m_{ij})_{k\times n}=\begin{cases}1 & (x_{ti},x_j)\in E_{U^-\times U}(A)\\0 & (x_{ti},x_j)\notin E_{U^-\times U}(A)\end{cases}\quad 1\leqslant i\leqslant k,1\leqslant j\leqslant n \tag{16}$$

定理 6 信息系统 $S=(U,AT)$减少对象集 U^-后更新为 $S=(U'=U-U^-,AT)$。对于属性集 $A\subseteq AT$,等价关系 $E_{U\times U}(A)$和 $E_{U'\times U'}(A)$的关系矩阵分别为 $\boldsymbol{M}_{n\times n}^{A}$和 $\boldsymbol{M}_{(n-k)\times(n-k)}^{A}$,对象集 U^-与论域 U 确定的关系矩阵为 $\boldsymbol{S}_{k\times n}^{A}$。那么满足:

$$|\boldsymbol{M}_{(n-k)\times(n-k)}^{A}|=|\boldsymbol{M}_{n\times n}^{A}|-2\cdot|\boldsymbol{S}_{k\times n}^{A}|+k \tag{17}$$

证明 当信息系统减少对象集 U^-后,那么满足:

$$E_{U^-\times U}(A)\subseteq E_{U\times U}(A),E_{U\times U^-}(A)\subseteq E_{U\times U}(A)$$

即 $\forall x\in U^-$和 $\forall y\in U$有 $(x,y)\in E_{U\times U}(A)$,同时 $(y,x)\in E_{U\times U}(A)$。由于 $\forall x\in U^-$,必有 $(x,x)\in E_{U^-\times U}(A)$,因此 $|\boldsymbol{M}_{(n-k)\times(n-k)}^{A}|=|\boldsymbol{M}_{n\times n}^{A}|-|\boldsymbol{S}_{k\times n}^{A}|-|(\boldsymbol{S}_{k\times n}^{A})^{\mathrm{T}}|+k$,即 $|\boldsymbol{M}_{(n-k)\times(n-k)}^{A}|=|\boldsymbol{M}_{n\times n}^{A}|-2\cdot|\boldsymbol{S}_{k\times n}^{A}|+k$。

证毕。

定理 6 表明,对于对象减少后的信息系统,只需要计算出减少对象集的区分度关系矩阵结果 $|\boldsymbol{S}_{k\times n}^{A}|$,然后根据原来的结果便可以快速地得到新的区分度关系矩阵对应的结果,这同样避免了大量的重复计算。

定理 7 信息系统 $S=(U,AT)$减少对象集 U^-后更新为 $S=(U'=U-U^-,AT)$。对于属性集 $A\subseteq AT$,矩阵 $\boldsymbol{M}_{n\times n}^{A}$、$\boldsymbol{M}_{(n-k)\times(n-k)}^{A}$和 $\boldsymbol{S}_{k\times n}^{A}$的意义同定理 6,论域 U 在属性集 A 下的区分度为 $Dis_U(A)$,那么 U'在属性集 A 下的区分度 $Dis_{U'}(A)$可增量式更新为:

$$Dis_{U'}(A)=Dis_U(A)-2kn+k^2-k+2\cdot|\boldsymbol{S}_{k\times n}^{A}| \tag{18}$$

证明 根据定理 1,

$$Dis_{U'}(A)=(n-k)^2-|\boldsymbol{M}_{(n-k)\times(n-k)}^{A}|=n^2-2kn+k^2-|\boldsymbol{M}_{n\times n}^{A}|+2\cdot|\boldsymbol{S}_{k\times n}^{A}|-k$$

其中:$n^2-|\boldsymbol{M}_{n\times n}^{A}|=Dis_U(A)$,所以:

$$Dis_{U'}(A)=Dis_U(A)-2kn+k^2-k+2\cdot|\boldsymbol{S}_{k\times n}^{A}|$$

证毕。

由定理7可以看出,计算出对象集 U^- 对应的关系矩阵结果 $|\boldsymbol{S}^{A}_{k\times n}|$,可以直接得到新的区分度值。在定理7的基础上,进一步地推导出信息系统对象减少时,相对区分度的增量式更新。

定理 8　信息系统 $S=(U,AT)$ 减少对象集 U^- 后更新为 $S=(U'=U-U^-,AT)$。对于属性集 $A_1,A_2\subseteq AT$,矩阵 $\boldsymbol{M}^{A_1}_{n\times n}$、$\boldsymbol{M}^{A_1}_{(n-k)\times(n-k)}$、$\boldsymbol{S}^{A_1}_{k\times n}$、$\boldsymbol{M}^{A_2}_{n\times n}$、$\boldsymbol{M}^{A_2}_{(n-k)\times(n-k)}$ 和 $\boldsymbol{S}^{A_2}_{k\times n}$ 的意义同定理6,属性集 A_2 关于 A_1 在论域 U 下的相对区分度为 $Dis_U(A_2\mid A_1)$,那么属性集 A_2 关于 A_1 在论域 U' 下的相对区分度 $Dis_{U'}(A_2\mid A_1)$ 可增量式更新为:

$$Dis_{U'}(A_2\mid A_1)=Dis_U(A_2\mid A_1)-2(|\boldsymbol{S}^{A_1}_{k\times n}|-|\boldsymbol{S}^{A_2}_{k\times n}\wedge\boldsymbol{S}^{A_1}_{k\times n}|)\quad(19)$$

证明　根据定理2有:

$$Dis_{U'}(A_2\mid A_1)=|\boldsymbol{M}^{A_1}_{(n-k)\times(n-k)}|-|\boldsymbol{M}^{A_2}_{(n-k)\times(n-k)}\wedge\boldsymbol{M}^{A_1}_{(n-k)\times(n-k)}|$$

由于 $|\boldsymbol{M}^{A_1}_{(n-k)\times(n-k)}|=|\boldsymbol{M}^{A_1}_{n\times n}|-2\cdot|\boldsymbol{S}^{A_1}_{k\times n}|+k$

$$|\boldsymbol{M}^{A_2}_{(n-k)\times(n-k)}\wedge\boldsymbol{M}^{A_1}_{(n-k)\times(n-k)}|=|\boldsymbol{M}^{A_2}_{n\times n}\wedge\boldsymbol{M}^{A_1}_{n\times n}|-2\cdot|S^{A_2}_{k\times n}\wedge S^{A_1}_{k\times n}|+k$$

因此 $Dis_{U'}(A_2\mid A_1)=|\boldsymbol{M}^{A_1}_{n\times n}|-2\cdot|\boldsymbol{S}^{A_1}_{k\times n}|+k-|\boldsymbol{M}^{A_2}_{n\times n}\wedge\boldsymbol{M}^{A_1}_{n\times n}|+2\cdot|\boldsymbol{S}^{A_2}_{k\times n}\wedge\boldsymbol{S}^{A_1}_{k\times n}|-k$

由于 $Dis_U(A_2\mid A_1)=|\boldsymbol{M}^{A_1}_{n\times n}|-|\boldsymbol{M}^{A_2}_{n\times n}\wedge\boldsymbol{M}^{A_1}_{n\times n}|$

所以 $Dis_{U'}(A_2\mid A_1)=Dis_U(A_2\mid A_1)-2(|\boldsymbol{S}^{A_1}_{k\times n}|-|\boldsymbol{S}^{A_2}_{k\times n}\wedge\boldsymbol{S}^{A_1}_{k\times n}|)$

证毕。

由定理8可以看出,当信息系统减少对象集 U^- 时,在 $Dis_U(A_2\mid A_1)$ 基础上只需计算 $\boldsymbol{S}^{A_1}_{k\times n}$ 和 $\boldsymbol{S}^{A_2}_{k\times n}$ 便可以快速地得到 $Dis_{U'}(A_2\mid A_1)$,因此这种增量式的计算方式同样具有很高的效率。

例 4　设表2所示的信息系统减少对象集 $U^-=\{x_7,x_8,x_9\}$,那么新信息系统即为表1。

令属性集 $A_1=\{a,b\}$,$A_2=\{c\}$。那么:

$$\boldsymbol{S}^{A_1}_{3\times 6}=\begin{bmatrix}0&1&0&1&0&0\\0&0&0&0&0&0\\0&0&1&0&0&0\end{bmatrix}$$

$$\boldsymbol{S}^{A_2}_{3\times 6}=\begin{bmatrix}1&1&1&0&0&1\\0&0&0&1&1&0\\1&1&1&0&0&1\end{bmatrix}$$

$$\boldsymbol{S}^{A_2}_{k\times n}\wedge\boldsymbol{S}^{A_1}_{k\times n}=\begin{bmatrix}0&1&0&0&0&0\\0&0&0&0&0&0\\0&0&1&0&0&0\end{bmatrix}$$

根据例3和定理8可以得到:

$$Dis_{U'}(A_2\mid A_1)=8-2\cdot(3-2)=6$$

与例2的计算结果是一致的。

3　增量式属性约简算法

根据第2节提出的区分度增量式更新方法,在算法1的基础上,本节将进一步地提出对应的区分度增量式属性约简算法,具体如算法2和算法3所示。

算法 2　信息系统对象增加时基于区分度的增量式属性约简算法

输入:更新后的信息系统 $S=(U'=U\cup U^+,C\cup D)$,更新前信息系统的约简集 red,相对区分度 $Dis_U(D\mid red)$ 和 $Dis_U(D\mid C)$。

输出:新的属性约简结果 red'。

Step 1　根据 $Dis_U(D\mid red)$ 和 $Dis_U(D\mid C)$ 增量式计算 $Dis_{U'}(D\mid red)$ 和 $Dis_{U'}(D\mid C)$。

Step 2

若 $Dis_{U'}(D\mid red)=Dis_{U'}(D\mid C)$,则跳转入 Step 7;

若 $Dis_{U'}(D\mid red)>Dis_{U'}(D\mid C)$,则跳转入 Step 5;

若 $Dis_{U'}(D\mid red)<Dis_{U'}(D\mid C)$,则跳转入 Step 3。

Step 3　对于 $\forall a\in red$,计算 a 的属性重要度:

$$sig_{red}(a)=Dis_{U'}(D\mid red-\{a\})-Dis_{U'}(D\mid red)$$

Step 4　找出 red 中属性重要度最大的属性 a',若 $sig_{red}(a')>0$,则 $red=red-\{a'\}$,并跳转入 Step 3,若 $sig_{red}(a')=0$,则跳转入 Step 7。

Step 5　对于 $\forall a\in C\text{-}red$,计算 a 的属性重要度:

$$sig_{red}(a)=Dis_{U'}(D\mid red)-Dis_{U'}(D\mid red\cup\{a\})$$

Step 6　找出 $C\text{-}red$ 中属性重要度最大的属性 a',若 $sig_{red}(a')>0$,则 $red=red\cup\{a'\}$,并跳转入 Step 5,若 $sig_{red}(a')=0$,则跳转入 Step 7。

Step 7　$red'\leftarrow red$。

Step 8　返回新的属性约简结果 red'。

算法 3　信息系统对象减少时基于区分度的增量式属性约简算法

输入:更新后的信息系统 $S=(U'=U-U^-,C\cup D)$,更新前信息系统的约简集 red,相对区分度 $Dis_U(D\mid red)$ 和 $Dis_U(D\mid C)$。

输出:新的属性约简结果 red'。

Step 1　根据 $Dis_U(D\mid red)$ 和 $Dis_U(D\mid C)$ 增量式计算 $Dis_{U'}(D\mid red)$ 和 $Dis_{U'}(D\mid C)$

Step 2

若 $Dis_{U'}(D\mid red)=Dis_{U'}(D\mid C)$,则跳转入 Step 7;

若 $Dis_{U'}(D\mid red)>Dis_{U'}(D\mid C)$,则跳转入 Step 5;

若 $Dis_{U'}(D\mid red)<Dis_{U'}(D\mid C)$,则跳转入 Step 3。

Step 3　对于 $\forall a\in red$,计算 a 的属性重要度:

$$sig_{red}(a)=Dis_{U'}(D\mid red-\{a\})-Dis_{U'}(D\mid red)$$

Step 4　找出 red 中属性重要度最大的属性 a',若 $sig_{red}(a')>0$,则 $red=red-\{a'\}$,并跳转入 Step 3,若 $sig_{red}(a')=0$,则跳转入 Step 7。

Step 5　对于 $\forall a\in C\text{-}red$,计算 a 的属性重要度:

$$sig_{red}(a) = Dis_{U'}(D \mid red) - Dis_{U'}(D \mid red \cup \{a\})$$

Step 6 找出 *C-red* 中属性重要度最大的属性 a',若 $sig_{red}(a') > 0$,则 $red = red \cup \{a'\}$,并跳转入 Step 5,若 $sig_{red}(a') = 0$,则跳转入 Step 7。

Step 7 $red' \leftarrow red$。

Step 8 返回新的属性约简结果 red'。

本文称算法 2 和算法 3 为基于区分度的增量式属性约简算法,那么 Teng 等[17]提出的算法(算法 1)即为基于区分度的非增量式属性约简算法。

观察算法 2 和算法 3 可以看出,它们均在原来信息系统属性约简的结果上进行增量式计算,这种增量式的计算可以大幅度减少对原先信息系统中数据的重复计算,从而提高了动态数据的约简效率。

在算法 2 和算法 3 中,设 $|U| = n$、$|U^+| = |U^-| = k$ 和 $|C| = c$,从 *red* 至 *red'* 的属性集大小变化量为 r,那么算法 2 和算法 3 的时间复杂度为 $O(c \cdot r \cdot n \cdot k)$。

4 实验分析

4.1 实验数据与实验设计

本实验采用 MATLAB 2014 作为实验平台进行算法的实现和运行,实验所运行的硬件环境为 Intel i7 4790 3.5 kGHz CPU 和 16 GB DDR3 内存。实验所使用的数据集如表 3 所示,这 6 个数据集均来自 UCI 机器学习数据集库,部分数据集中包含连续型的属性值,实验前需要进行离散化处理。

表 3 实验数据集

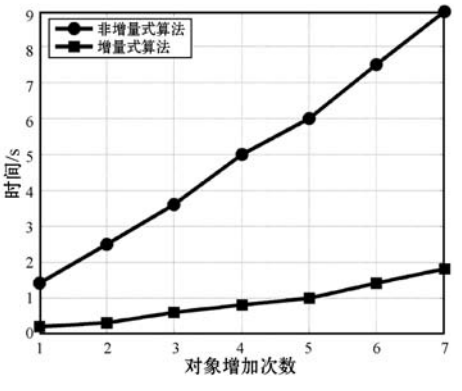
序号	数据集	简称	对象数	属性数
1	Ionosphere	iono	351	35
2	German Credit Data	gcd	1 000	19
3	Ticdata2000	tic	5 822	85
4	Thyroid Disease	td	7 200	21
5	MagicGammaTelescope	mgt	19 020	11
6	Census Income	ci	48 842	14

表 3 列举的数据集均为静态完整的数据集,为了运用文中提出的增量式属性约简算法,本实验采用其他学者常用的实现方法[13-16],将数据集的论域随机分割成多个子数据集,本实验选择分割的数量为 8 个。对于数据集对象逐渐增加的情形,实验中随机选择其中一个子数据集作为初始的数据集,然后从剩余的子数据集中选择出一个与初始数据集进行合并,这样便模拟出了数据集对象的一次动态增加,重复上述步骤,最后直至完成数据集的 7 次更新。类似地,对于数据集对象逐渐减少的情形,实验中随机选择其中一个子数据集,然后从完整数据集中删除该子数据集,这样便模拟出了数据集对象的一次动态减少,重复进行此步骤,便构造出了数据集对象的 7 次动态减少。

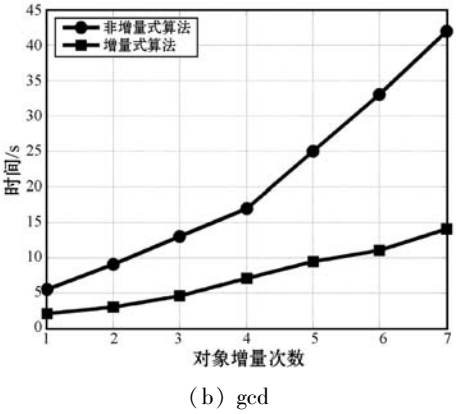
将传统的区分度属性约简算法和本文提出的区分度增量式属性约简算法,分别对数据集对象动态增加的情形和对象动态减少的情形进行属性约简。然后通过属性约简的效率、属性约简集的大小以及属性约简结果的分类性能来比较两种算法的属性约简性能,从而验证出本文所提出算法的有效性。

4.2 属性约简效率比较

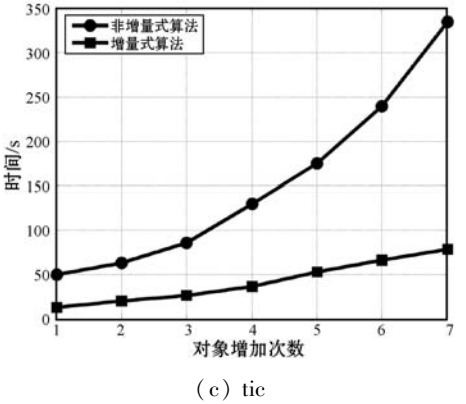
图 1 为基于区分度的增量式属性约简算法(算法 2)与区分度的非增量式属性约简算法(算法 1)在各个数据集下对象 7 次增加时的属性约简效率比较。其中每个子图的横坐标表示的是数据集的更新次数,纵坐标表示的是算法进行属性约简时所需的用时。



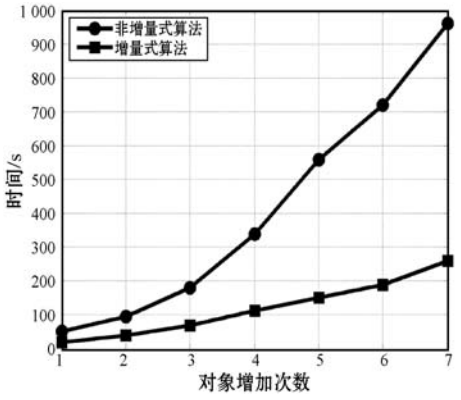
(a) iono



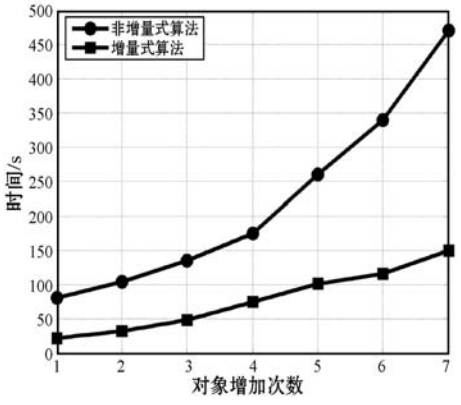
(b) gcd



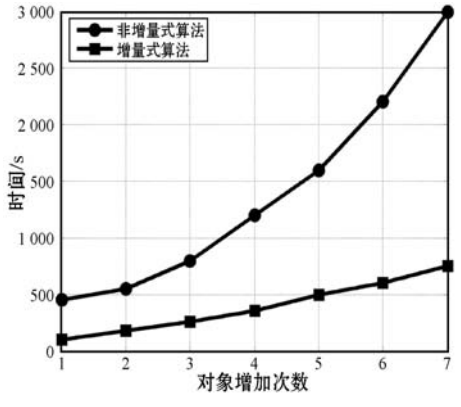
(c) tic



(d) td



(e) mgt

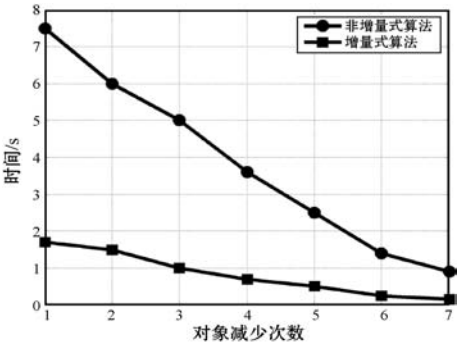


(f) ci

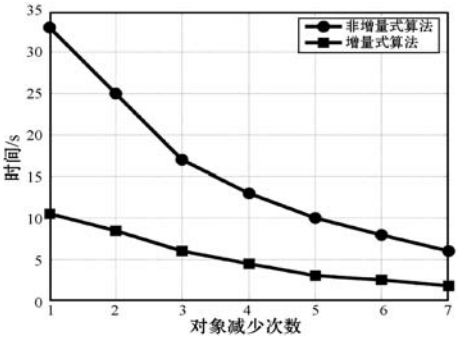
图 1　各数据集对象增加时属性约简用时比较

观察图 1 每个数据集的实验结果可以发现,随着数据集更新次数的增加,两种算法在属性约简的用时方面表现出了明显的差距,其中本文所提出的增量式属性约简用时大幅度低于非增量式算法。这主要是由于本文所提出的增量式算法是在原来属性约简的结果上进行计算,减少了对原来数据的重复计算,大幅度提高了计算效率。

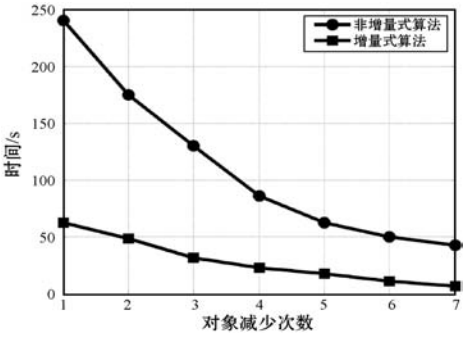
图 2 为基于区分度的增量式属性约简算法(算法 3)与区分度的非增量式属性约简算法(算法 1)在各个数据集下对象 7 次减少时的属性约简效率比较。其中每个子图的横坐标表示的是数据集的更新次数,纵坐标表示的是算法进行属性约简的用时。



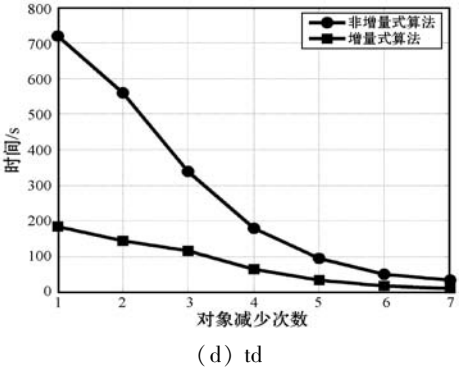
(a) iono



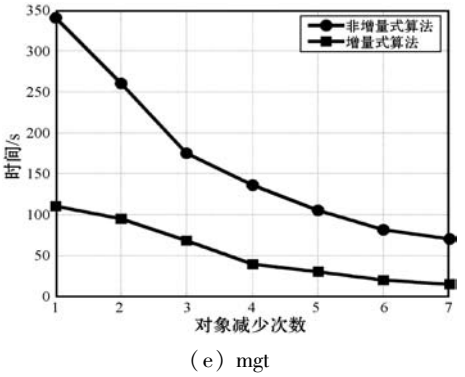
(b) gcd



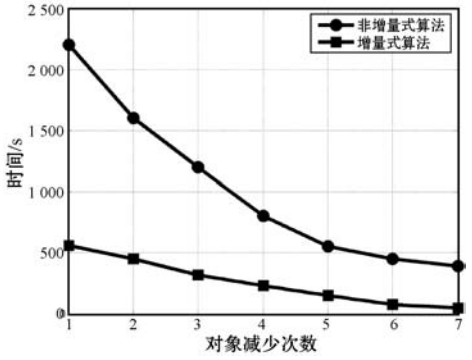
(c) tic



(d) td



(e) mgt



(f) ci

图 2 各数据集对象减少时属性约简用时比较

观察图 2 同样可以发现,随着数据集更新次数的增加,两种算法在属性约简的用时方面也表现出了明显的差距,同样本文所提出的增量式属性约简用时大幅度低于已提出的非增量式属性约简算法。其原因同样是由于本文所提出的算法通过增量式计算提高了约简效率,每次属性约简时避免了对旧数据的重复计算。

4.3 属性约简结果比较

对于实验中动态更新的信息系统,属性约简算法在信息系统每次更新时都能得到个当时对应的属性约简结果。将 7 次属性约简结果的约简集大小取平均值,这样就得到了对应算法的平均约简结果。表 4 为非增量式属性约简算法(算法 1)与文中提出的增量式属性约简算法(算法 2)在各个数据集下论域 7 次动态增加时的平均约简结果。表 5 为非增量式属性约简算法(算法 1)与文中提出的增量式属性约简算法(算法 3)在各个数据集下论域 7 次动态减少时的平均约简结果。

表 4 数据集对象动态增加时平均约简结果比较

数据集	原始属性集	(非增量式算法) 算法 1	(增量式算法) 算法 2
iono	35	15.6	13.2
gcd	19	10.5	9.8
tic	85	22.4	19.6
td	21	7.1	7.5
mgt	11	9.3	8.5
ci	14	11.2	10.4

表 5 数据集对象动态增加时平均约简结果比较

数据集	原始属性集	(非增量式算法) 算法 1	(增量式算法) 算法 3
iono	35	15.3	13.4
gcd	19	10.1	9.5
tic	85	21.7	18.8

续表 5

数据集	原始属性集	(非增量式算法) 算法 1	(增量式算法) 算法 3
td	21	7.6	7.3
mgt	11	8.3	8.8
ci	14	10.4	10.7

观察表 4 可以发现,对于论域逐渐增加的信息系统,本文提出的增量式属性约简算法(算法 2)在大部分数据集中具有较小的平均属性约简结果,例如数据集 iono、gcd、tic、mgt 和 ci。而对于非增量式属性约简算法(算法 1),只在小部分的数据集拥有较小的平均属性约简结果,例如数据集 td。产生这种差异主要是由于增量式属性约简算法的约简机制导致的,增量式属性约简在进行约简时,是根据原先信息系统的约简结果进行进一步计算,这样避免了对整个属性集进行重新搜索,从而增量式算法约简出的属性更少。观察表 5 结果同样可以发现,对于论域逐渐减少的信息系统,文中提出的增量式属性约简算法(算法 3)具有更小的平均属性约简结果,例如数据集 iono、gcd、tic 和 td。

4.4 属性约简结果的分类性能比较

为了测试两类属性约简算法约简结果的分类性能,本实验通过支持向量机分类器(SVM)与改进的决策树分类器(C4.5)分别对每次更新时的属性约简结果进行分类训练,并得到约简集对应的分类精度,最后将所有分类精度取均值,得到对应属性约简算法的平均分类精度。表 6 为各个数据集论域动态增加时两类算法的平均分类精度比较结果,表 7 为各个数据集论域动态减少时两类算法的平均分类精度比较结果。

表 6 数据集对象增加时约简结果分类精度比较 %

数据集	(非增量式算法) 算法 1		(增量式算法) 算法 2	
	SVM	C4.5	SVM	C4.5
iono	87.45	85.32	86.57	86.19
gcd	75.87	76.41	74.27	75.62
tic	80.51	82.78	82.35	83.74
td	94.61	93.23	94.21	92.35
mgt	81.25	79.54	82.34	81.75
ci	71.39	73.20	70.95	74.23

表 7　数据集对象减少时约简结果分类精度比较　%

数据集	(非增量式算法) 算法 1		(增量式算法) 算法 2	
	SVM	C4.5	SVM	C4.5
iono	89.42	87.46	88.36	88.67
gcd	75.65	77.37	75.47	77.21
tic	80.48	81.58	80.39	82.50
td	94.87	94.26	95.21	93.42
mgt	83.25	81.54	84.47	82.59
ci	73.27	75.14	73.45	73.63

观察表 6 可以发现,本文所提出的区分度增量式属性约简算法(算法 2)在数据集 tic 和 mgt 下拥有较高的 SVM 分类精度,在数据集 iono、tic、mgt 和 ci 下拥有较高的 C4.5 分类精度,基于区分度的非增量式属性约简算法在其他数据集拥有较高的分类精度,可以发现两类算法的平均分类精度在大部分数据集上相差不大。观察表 7 可以发现,所提出的区分度增量式属性约简算法(算法 3)在数据集 td、mgt 和 ci 下拥有较高的 SVM 分类精度,在数据集 iono、tic 和 mgt 下拥有较高的 C4.5 分类精度,同样在大部分数据集下,两类算法具有相近的平均分类精度。因此表 6 和表 7 说明了所提出的区分度增量式属性约简算法同样能够得到较优的属性约简结果。

4.5　实验总结

综合 4.2 节、4.3 节和 4.4 节三个部分的实验比较结果,说明对于样本动态增加或减少的数据集,本文提出的区分度增量式属性约简算法具有较高的属性约简性能,能够满足数据变化时属性约简的实时需求。同时所提出的增量式属性约简算法能够比非增量式算法选择出更小的约简集,并且也能够保持同样的分类性能。所以本文提出的区分度增量式属性约简是一种较优的动态数据属性约简算法。

5　结　语

属性约简是粗糙集理论在机器学习和知识发现领域中的一项重要应用。然而现实环境下的数据总是实时更新的,针对这一数据环境,学者们在传统属性约简算法的基础上,将增量式学习融入其中,提出了多种增量式属性约简算法。区分度作为一种重要的属性集评估方法,目前已成为属性约简的一种重要的方法,本文针对样本不断动态变化的数据集环境,提出一种基于区分度的增量式属性约简算法。首先通过矩阵方法去表示区分度,并通过矩阵研究了区分度的增量式学习,然后基于区分度的增量式更新提出一种增量式属性约简算法,最后通过实验验证了所提出增量式属性约简算法的有效性。由于文中仅针对数据集样本的变化进行了研究,因此接下来将进一步探索数据集属性变化时属性约简的增量式更新。

参　考　文　献

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

[2] 邬阳阳,汤建国. 大数据背景下粗糙集属性约简研究进展[J]. 计算机工程与应用,2019,55(6):31 - 38,177.

[3] Hu C, Zhang L, Wang B, et al. Incremental updating knowledge in neighborhood multigranulation rough sets under dynamic granular structures[J]. Knowledge-Based Systems, 2019, 163(1):811 - 829.

[4] Lang G, Miao D, Cai M, et al. Incremental approaches to updating reducts under dynamic covering granularity[J]. Knowledge-Based Systems, 2017, 134(15):85 - 104.

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

[6] 王映龙,曾淇,钱文彬,等. 变精度下不完备混合数据的增量式属性约简方法[J]. 计算机应用,2018,38(10):2764 - 2771.

[7] Luo C, Li T, Chen H, et al. Efficient updating of probabilistic approximations with incremental objects[J]. Knowledge-Based Systems, 2016, 109(1):71 - 83.

[8] Shu W, Qian W, Xie Y. Incremental approaches for feature selection from dynamic data with the variation of multiple objects[J]. Knowledge-Based Systems, 2019, 163(1):320 - 331.

[9] Wei W, Wu X, Liang J, et al. Discernibility matrix based incremental attribute reduction for dynamic data[J]. Knowledge-Based Systems, 2018, 140(15):142 - 157.

[10] 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.

[11] Gao C, Lai Z, Zhou J, et al. Granular maximum decision entropy-based monotonic uncertainty measure for attribute reduction[J]. International Journal of Approximate Reasoning, 2019, 104:9 - 24.

[12] Gao C, Lai Z, Zhou J, et al. Maximum decision entropy-based attribute reduction in decision-theoretic rough set model[J]. Knowledge-Based Systems, 2018, 143(1):179 - 191.

(下转第 257 页)

4 结 语

针对城市峡谷环境中卫星信号衰减了 10 ~ 25 dB,普通接收机无法实现定位的问题,提出一种基于协同定位技术的集体检测的算法,把卫星信号的相关结果对应到空间区域中。只需要通过相关结果确定出卫星位置,避免了跟踪和解算的过程。算法将计算复杂性降低了至少 90%,定位误差从几十米甚至几百米[19]缩减到了 4 m 以内,有效地实现了用户在城市峡谷环境中的定位。

参 考 文 献

[1] 李基武. 弱信号环境下 GNSS 信号的捕获与跟踪算法研究[D]. 桂林:桂林电子科技大学,2017.

[2] 丁继成. 弱信号条件下 GPS 接收机关键技术研究[D]. 哈尔滨:哈尔滨工程大学,2009.

[3] 段华华,巴晓辉,陈杰. 互相关干扰下微弱 GPS 信号检测方法[J]. 数据采集与处理,2015,30(3):677 - 682.

[4] Liu J F,Liu J Y,Zhang T T,et al. Application of cross-correlation algorithm in radio weak signal detection[C]//2009 Seventh Annual Communication Networks and Services Research Conference. IEEE,2009.

[5] 汪志坤. 基于互相关干扰消除的 GPS 弱信号捕获算法研究[D]. 南京:南京邮电大学,2016.

[6] 邢永强,黄海生,曹新亮. 北斗 B1 MEO/IGSO 卫星信号的差分捕获算法[J]. 电子技术应用,2018,44(6):90 - 93.

[7] 赵焕焕,黄海生,张伟,等. 北斗接收机的互相关抑制算法研究[J]. 南京邮电大学学报(自然科学版),2015,35(6):39 - 43.

[8] 崔绍龙,姚相振,方金云. 一种 GPS 微弱信号的优化捕获算法仿真分析[J]. 系统仿真学报,2014,26(1):112 - 118.

[9] Yang L,Tian J. Analysis and compare of weak GPS signal acquisition algorithms[C]//IET International Communication Conference on Wireless Mobile and Computing (CCWMC 2009),2009.

[10] Meng W X,Ma R F,Han S. Optimum path based differential coherent integration algorithm for GPS C/A code acquisition under weak signal environment[C]//2010 First International Conference on Pervasive Computing,Signal Processing and Applications,2010.

[11] Wu L J,Lu W J,Yu D S. Research of weak signal acquisition algorithms for high sensitivity GPS receivers[C]//2009 Asia Pacific Conference on Postgraduate Research in Microelectronics & Electronics (PrimeAsia),2009.

[12] Tian S J,Pi Y M. Research of weak GPS signal acquisition algorithm[C]//2008 International Conference on Communications,Circuits and Systems,2008.

[13] Elders-Boll H, Dettmar U. Efficient differentially coherent code/Doppler acquisition of weak GPS signals[C]//Eighth IEEE International Symposium on Spread Spectrum Techniques and Applications-Programme and Book of Abstracts (IEEE Cat. No. 04TH8738),2004.

[14] 夏景平,胡辉,颜瑜军,等. 基于高分值加权的改进阴影匹配定位算法研究[J]. 全球定位系统,2017,42(6):1 - 8.

[15] 张昌庆,黄劲松. 利用蓝牙信号强度的端端协同定位[J]. 导航定位学报,2019,7(2):18 - 24.

[16] 蒋悦,马永涛,宫霄霖,等. 基于非度量多维标度的室内多标签协同定位算法[J]. 传感技术学报,2018,31(10):1553 - 1558.

[17] 屈耀红,张峰,谷任能,等. 基于距离测量的多无人机协同目标定位方法[J]. 西北工业大学学报,2019,37(2):266 - 272.

[18] 谢钢. GPS 原理与接收机设计[M]. 电子工业出版社,2009.

[19] 殷鹏,何玉庆,韩建达,等. 基于多分辨率粒子滤波的全局协同定位方法[J]. 中国科学:技术科学,2019,49(1):87 - 96.

(上接第 245 页)

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

[14] 赵小龙,杨燕. 基于邻域粒化条件熵的增量式属性约简算法[J]. 控制与决策,2019,34(10):1 - 13.

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

[16] Xie X, Qin X. A novel incremental attribute reduction approach for dynamic incomplete decision systems[J]. International Journal of Approximate Reasoning, 2018, 93:443 - 462.

[17] Teng S, Lu M, Yang A, et al. Efficient attribute reduction from the viewpoint of discernibility[J]. Information Sciences, 2016, 326(1): 297 - 314.

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

[19] 闫鑫,景运革. 矩阵增量属性约简算法[J]. 小型微型计算机系统,2018,39(6):1245 - 1249.

[20] Tan A, Li J, Lin Y, et al. Matrix-based set approximations and reductions in covering decision information systems[J]. International Journal of Approximate Reasoning, 2015, 59: 68 - 80.