DOI:10. 19431/j. cnki. 1673-0062. 2022. 03. 015

基于粗糙集理论的多标记数据互补决策约简加速算法

李　华,王思宇,王雅茹

(石家庄铁道大学 数理系,河北 石家庄 050043)

摘　要:互补决策约简是一种多标记数据属性约简方法,当数据规模较大时,其启发式算法的计算耗时较大。基于粗糙集理论,对互补决策约简启发式算法的加速算法进行了研究。当粒度由粗变细时,在逐步去掉正域的数据集上,首先研究互补决策约简中属性外部重要度的保序性质;基于此,通过逐步缩小数据规模来降低计算约简的耗时,提出了互补决策约简加速算法。加速算法不仅减少了属性约简的计算时间,而且能够保持原始算法的约简结果。

关键词:多标记数据;互补决策约简;粗糙集;保序性
中图分类号:TP301　　**文献标志码**:A
文章编号:1673-0062(2022)03-00106-07　　**开放科学(资源服务)标识码(OSID)**:

Rough Set Theory Based Accelerated Algorithm for Complementary Decision Reduct of Multi-label Data

LI Hua, WANG Siyu, WANG Yaru

(Department of Mathematics and Physics, Shijiazhuang Tiedao University, Shijiazhuang, Hebei 050043, China)

Abstract: Complementary decision reduct is an effective attribute reduction approach for multi-label data. However, the corresponding heuristic algorithm is computationally time-consuming for large data sets. This paper proposes an accelerated heuristic algorithm of complementary decision reduct based on rough set theory. First, the rank preservation of outer significance measure of attribute in complementary decision reduct is studied on a dataset of positive region that is gradually removed when the granulation changes from coarse to fine. Then, an accelerated heuristic algorithm is proposed which can decrease the time-consuming by gradually shrinking the data scale. The accelerated algorithm not only speeds up the process of attribute reduction, but also preserves the reduction results of the original algorithm.

收稿日期:2021-11-29
基金项目:国家自然科学基金项目(61806133);国家留学基金项目(201908130072)
作者简介:李　华(1978—),女,副教授,博士,主要从事数据挖掘与智能计算方面的研究。E-mail:lihuasjz@163.com

key words:multi-label data;complementary decision reduct;rough set;rank preservation

0　引　言

现实世界中,一个样本可能同时与多个标记相关联,被称为多标记学习。例如,一幅图片可以同时具有沙滩、山和天空等多个信息标注[1];一首音乐可以表达开心、高兴和放松等多种情感类[2]。由于多标记学习框架更加准确地刻画了实际应用中大量存在的多义性现象[3],因而被广泛应用于文本分类[4]、图像标注[5]和社交网络[6]等领域。

随着信息技术的高速发展,视频、文档和音频序列等多标记数据的维数显著增加,呈现出高维化的趋势。多标记数据的高维特征不仅增加了计算成本和存储代价,而且还会导致所谓的“维数灾难”[7]问题,因此对多标记数据的高维特征进行降维处理显得尤为重要。

众多的降维技术中,基于粗糙集理论[8]的特征选择方法受到了广泛关注。在粗糙集理论中,特征选择也称为属性约简,其是在保持可识别能力不变的情况下,通过去除冗余特征来获取知识的属性约简。近年来,已有学者将粗糙集理论的属性约简技术应用于多标记学习。段洁等人应用前向贪心策略提出了基于邻域粗糙集的多标记特征选择算法[9];H. Li 等人提出了基于粗糙集理论的互补决策约简启发式算法,去除不必要属性的同时保持了由标记所传递的不确定性不变[10];Y. Lin 等人提出了多标记模糊粗糙集理论并给出了相应的属性约简方法[11];Y. Li 等人提出了一种基于模糊粗糙集的多标记特征选择算法[12];Z. Qiao 等人利用标记集正域进行多标记数据的属性约简[13];W. Qian 等人基于特征依赖度和粗糙集理论,提出了标记分布学习的特征选择算法[14]。

上述针对于多标记数据的属性约简算法都能有效地减少冗余属性,但对于大规模数据集而言,他们都不同程度地存在着计算耗时较大的问题。因此,多标记数据属性约简算法的加速机制就成为了一个亟待解决的关键问题。

在粗糙集理论中,基于集合的包含关系,当属性幂集中的属性集由粗到细变化时,可以确定由粗到细的粒度序列,称为正粒化序列[15]。当粒度由粗到细变化时,目标概念的正域逐渐增大。如果逐步去掉正域,那么数据集的规模会逐渐缩小。本文首先研究了当数据集规模逐步缩小时,互补决策约简中属性外部重要度的保序性质。特别地,当粒度由粗到细变化时,在逐步去掉正域的数据集上,具有最大外部重要度的属性是保持不变的。基于此,本文提出互补决策约简启发式算法的一种加速算法,该加速算法是在粒度由粗到细的变化过程中,通过逐步去掉正域来降低数据集的规模,缩短计算约简的时间;同时,虽然数据集的规模不断缩小,但具有最大外部重要度的属性是保持不变的,而约简是通过不断在核属性集上增加具有最大外部重要度的属性得到的,因此加速算法可以得到与原算法相同的互补决策约简。

1　基本概念

1.1　多标记决策表

多标记数据可以用多标记决策表[10] $S=(U,A,L)$ 来形式化表示,其中 $U=\{x_1,x_2,\cdots,x_n\}$ 是对象集合,称为论域;$A=\{a_1,a_2,\cdots,a_m\}$ 是条件属性集合,称为条件属性集;$L=\{l_1,l_2,\cdots,l_q\}$ 是标记的集合,称为标记集。每个条件属性 $a\in A$ 形成一个满射 $a:U\to V_a$,其中 V_a 表示条件属性 a 的值域。每个标记 $l\in L$ 形成一个满射 $l:U\to V_l$,其中 $V_l=\{0,1\}$ 表示标记 l 的值域。如果对象 x 和标记 l 相关联,那么 $l(x)=1$,否则 $l(x)=0$。

在多标记决策表中,条件属性集 A 与标记集 L 互不相交,即 $A\cap L=\varnothing$;U 中的每一个对象至少关联于标记集 L 中的一个标记[16];L 中的每一个标记至少与 U 中的一个对象相关联[17]。本文不考虑无标记的对象。

下面是多标记决策表的一个实例。

例 1　多标记决策表 $S=(U,A,L)$ 是由中医冠心病数据集[18]中部分数据组成,如表 1 所示。论域 $U=\{x_1,x_2,x_3,x_4,x_5\}$ 为患者集合。条件属性集 $A=\{a,b,c\}$ 是由 3 个症状组成的条件属性集,其中条件属性 a 表示“胸痛特点”,其值域为{1 隐痛 2 刺痛 3 胀痛};条件属性 b 表示“诱发因素”,其值域为{1 劳累后加重 2 饮酒后加重 3 气候骤冷加重};条件属性 c 表示“心悸”,其值域为{1 出现 2 不出现}。$L=\{l_1,l_2,l_3\}$ 是由 3 个证型组成的标记集,其中 l_1 表示“心气虚”,l_2 表示“心阳虚”,l_3 表示“气滞”。显然,U 中的每个对象至少关联于 L 中一个标记,L 中的每个标记至少关

联于 U 中的一个对象。

表 1　多标记决策表 $S=(U,A,L)$

Table 1　A multi-label decision table $S=(U,A,L)$

U	a	b	c	l_1	l_2	l_3
x_1	1	2	1	1	0	0
x_2	3	1	2	0	1	0
x_3	1	2	1	1	0	1
x_4	2	3	1	1	0	1
x_5	2	3	1	0	0	1

1.2　互补决策约简

定义 1[10]　设 $S=(U,A,L)$ 是多标记决策表，其中 $A=\{a_1,a_2,\cdots,a_m\}$，$L=\{l_1,l_2,\cdots,l_q\}$。给定一个标记 $l_i\in L$，称

$$E_i=\{x\in U:l_i(x)=1\} \tag{1}$$

为对应于标记 l_i 的一个标记信息集。

标记信息集是所有与标记 l_i 相关联的对象集合，描述了多标记数据中隐含的标记信息。

定义 2[10]　设 $S=(U,A,L)$ 是多标记决策表，其中 $A=\{a_1,a_2,\cdots,a_m\}$，$L=\{l_1,l_2,\cdots,l_q\}$。$P(L)$ 是标记集的幂集，$E_1,E_2,\cdots,E_q$ 是 q 个标记信息集。设 $B\subseteq A$，粗决策函数 $C_B^U:U\to P(L)$ 和细决策函数 $F_B^U:U\to P(L)$ 定义如下：

$$C_B^U(x)=\{l_i:[x]_B\cap E_i\neq\phi\},x\in U \tag{2}$$

$$F_B^U(x)=\{l_i:[x]_B\subseteq E_i\},x\in U \tag{3}$$

粗决策函数 $C_B^U(x)$ 是至少关联于 $[x]_B$ 中一个对象的标记的集合。细决策函数 $F_B^U(x)$ 是关联于 $[x]_B$ 中所有对象的标记的集合。也就是说，$C_B^U(x)$ 是 $[x]_B$ 中所有对象的标记集合的并集。$F_B^U(x)$ 是 $[x]_B$ 中所有对象标记集合的交集。

定义 3[10]　设 $S=(U,A,L)$ 是多标记决策表，$B\subseteq A$。B 称为 S 的一个互补决策约简当且仅当 B 满足下面的条件：

(1)对任意 $x\in U$，

$$C_B^U(x)=C_A^U(x)\ 且\ F_B^U(x)=F_A^U(x) \tag{4}$$

(2)对任意的 $B'\subset B$，存在 $x\in U$，使得

$$C_{B'}^U(x)\neq C_A^U(x)\ 或\ F_{B'}^U(x)\neq F_A^U(x) \tag{5}$$

若 B 只满足条件(1)，则称 B 是一个互补决策协调集，否则称 B 是不协调的。

2　动态粒度下多标记决策表的粗糙集近似

由等价关系所确定的分划为描述目标概念提供一个粒化世界[19]。用一组具有偏序关系的等价关系族来刻画目标概念被称为动态粒度下的粗糙集近似[15]。如果在这种粒度变化下，采用逐渐细化的思想，即粒度由粗变细，称为正向近似。对于粒度计算和粗糙集理论，动态粒度下的正向近似思想为其提供了一个新的研究方向，并且在属性约简算法中得到了广泛且有效的应用[20-21]。本节主要探讨动态粒度下多标记决策表的粗糙集近似及其相关性质。

设 $S=(U,A,L)$ 是多标记决策表，在 2^A 上定义偏序关系 $\leqslant$：设 $P\leqslant Q$(或者 $Q\geqslant P$)表示 P 比 Q 精细(或者 Q 比 P 粗糙)，当且仅当满足对任意的 $P_i\in U/P$，存在 $Q_j\in U/Q$ 使得 $P_i\subseteq Q_j$，其中 $U/P=\{P_1,P_2,\cdots,P_s\}$ 和 $U/Q=\{Q_1,Q_2,\cdots,Q_t\}$ 分别是由 $P,Q\subseteq A$ 所确定的划分。如果 $P\leqslant Q$ 且 $U/P\neq U/Q$，称 P 严格细于 Q(或者 Q 严格粗于 P)，用 $P<Q$(或者 $Q>P$)来表示。

性质 1[10]　设 $S=(U,A,L)$ 是多标记决策表，设 $B,C\subseteq A$，那么

(1)如果 $B\geqslant C$，则

$$F_B^U(x)\subseteq F_C^U(x)\subseteq C_C^U(x)\subseteq C_B^U(x) \tag{6}$$

(2)对任意 $x\in U$，

$$C_B^U(x)\neq\phi \tag{7}$$

下面由粗决策函数和细决策函数来定义多标记决策表的正域。

定义 4　设 $S=(U,A,L)$ 是多标记决策表，$B\subseteq A$，对于任意 $x\in U$，标记集 L 关于条件属性集 B 的正域定义为：

$$POS_B^U(L)=\{x\in U:C_B^U(x)=F_B^U(x)\} \tag{8}$$

标记集 L 关于条件属性集 B 的正域 $POS_B^U(L)$ 是由论域 U 中这样的对象所组成，其对应于知识 B 的等价类中的所有对象都具有相同的标记集。换句话说，$POS_B^U(L)$ 是由根据知识 B 判断具有确定标记集的对象集合。

下面基于动态粒度下的正向近似思想表示正域。

定义 5　设 $S=(U,A,L)$ 是多标记决策表，$B=\{R_1,R_2,\cdots,R_n\}$ 是属性集族，其中 $R_1\geqslant R_2\geqslant\cdots\geqslant R_n$($R_i\in 2^A$)，设 $B_i=\{R_1,R_2,\cdots,R_i\}$，标记集 L 关于 B_i 的正域可表示为：

$$POS_{B_i}^U(L)=POS_{R_1}^{U_1}(L)\cup POS_{R_2}^{U_2}(L)\cup\cdots\cup POS_{R_i}^{U_i}(L) \tag{9}$$

其中 $U_1=U$，$U_k=U-\cup_{j=1}^{k-1}POS_{R_j}^{U_j}(L)$，$k=2,3,\cdots,n$，$i=1,2,\cdots,n$。

性质 2　设 $S=(U,A,L)$ 是多标记决策表,$B=\{R_1,R_2,\cdots,R_n\}$ 是属性集族,其中 $R_1\geqslant R_2\geqslant\cdots\geqslant R_n(R_i\in 2^A)$,设 $B_i=\{R_1,R_2,\cdots,R_i\}$,则

$$POS_{B_k}^{U}(L)=POS_{B_{k-1}}^{U}(L)\cup POS_{R_k}^{U_k}(L)\quad(10)$$

其中 $U_1=U$,$U_k=U-POS_{B_{k-1}}^{U}(L)$,$k=2,3,\cdots,n$,$i=1,2,\cdots,n$。

证明:根据定义 5,可得

$$POS_{B_k}^{U}(L)=POS_{R_1}^{U_1}(L)\cup\cdots\cup POS_{R_{k-1}}^{U_{k-1}}(L)\cup POS_{R_k}^{U_k}(L)=POS_{B_{k-1}}^{U}(L)\cup POS_{R_k}^{U_k}(L),$$

其中,$U_1=U$,

$$U_k=U-\cup_{j=1}^{k-1}POS_{R_j}^{U_j}(L)=U-POS_{R_1}^{U_1}(L)\cup\cdots\cup POS_{R_{k-1}}^{U_{k-1}}(L)=U-POS_{B_{k-1}}^{U}(L)。$$

性质 2 表明标记集 L 关于属性集族 B_k 的正域可由属性集族 B_{k-1} 的正域和在逐渐减少的论域上的属性集 R_k 的正域来表示。

性质 3　设 $S=(U,A,L)$ 是多标记决策表,$B=\{R_1,R_2,\cdots,R_n\}$ 是属性集族,其中 $R_1\geqslant R_2\geqslant\cdots\geqslant R_n(R_i\in 2^A)$,设 $B_i=\{R_1,R_2,\cdots,R_i\}$,$i=1,2,\cdots,n$,则

$$POS_{B_1}^{U}(L)\subseteq POS_{B_2}^{U}(L)\subseteq\cdots\subseteq POS_{B_n}^{U}(L)\quad(11)$$

证明:由性质 2 可知 $POS_{B_{k-1}}^{U}(L)\subseteq POS_{B_k}^{U}(L)$,$k=2,3,\cdots,n$。故

$$POS_{B_1}^{U}(L)\subseteq POS_{B_2}^{U}(L)\subseteq\cdots\subseteq POS_{B_n}^{U}(L)。$$

性质 3 表明随着属性集族中属性集数目的增加,标记集 L 关于属性集族的正域是单调递增的。

3　多标记数据互补决策约简加速算法

当粒度由粗到细变化时,在逐步去掉正域的多标记决策表上,互补决策约简中属性外部重要度具有保序性质,并基于此提出互补决策约简启发式算法的加速算法。

3.1　属性重要性的度量

定义 6[10]　设 $S=(U,A,L)$ 是多标记决策表,$B\subseteq A$,条件属性集 B 的标记依赖度为

$$\gamma_L^U(B)=\frac{\sum_{x\in U}|F_B^U(x)|+\phi^U(B)}{\sum_{x\in U}|C_B^U(x)|}\quad(12)$$

其中 $\phi^U(B)=\begin{cases}\lambda,\forall x\in U,F_B^U(x)=\phi;\\0,\exists x\in U,F_B^U(x)\neq\phi,\end{cases}$

这里 $\lambda\in(0,1)$ 是常数。

性质 4[10]　设 $S=(U,A,L)$ 是多标记决策表,$B_1,B_2\subseteq A$,如果 $B_1\subseteq B_2$,那么 $\gamma_L^U(B_1)\leqslant\gamma_L^U(B_2)$。

标记依赖度反映了条件属性子集 B 对标记集 L 的近似能力或依赖程度,由依赖度也可以定义每个条件属性的重要度。

定义 7[10]　设 $S=(U,A,L)$ 是多标记决策表,$B\subseteq A$,$a\in B$ 的内部重要度定义为:

$$Sig^{\text{inner}}(a,B,L,U)=\gamma_L^U(B)-\gamma_L^U(B-\{a\})\quad(13)$$

定义 8[10]　设 $S=(U,A,L)$ 是多标记决策表,$B\subseteq A$,$a\in A-B$ 关于 B 的外部重要度定义为:

$$Sig^{\text{outer}}(a,B,L,U)=\gamma_L^U(B\cup\{a\})-\gamma_L^U(B)\quad(14)$$

不同的条件属性在保持标记不确定性方面的作用是不同的,可将这些属性分为两类:不必要的和必要的。

定义 9[10]　设 $S=(U,A,L)$ 是多标记决策表,$B\subseteq A$,称条件属性 $a\in B$ 是 B 中的不必要属性,如果对于任意的 $x\in U$,

$$C_B^U(x)=C_{B-\{a\}}^U(x)\text{ 且 }F_B^U(x)=F_{B-\{a\}}^U(x)\quad(15)$$

否则称 a 为 B 的必要属性。

A 中所有必要属性组成的集合称为 A 的核,记为 $CORE(A)$。

下面性质表明核可由内部重要度定义。

性质 5[10]　设 $S=(U,A,L)$ 是多标记决策表,则

$$CORE(A)=\{a\in A:Sig^{\text{inner}}(a,A,L,U)>0\}\quad(16)$$

标记依赖度和属性的内部重要度也可用于描述互补决策约简。

定理 1[10]　设 $S=(U,A,L)$ 是多标记决策表,$B\subseteq A$。如果 $\gamma_L^U(A)=\gamma_L^U(B)$,而且对于任意的 $a\in B$,有 $\text{Sig}^{\text{inner}}(a,B,L,U)>0$,那么 B 是 S 的互补决策约简。

由性质 5 可知,通过计算属性的内部重要度可得到多标记决策表的核属性集。然后,依次在核属性集上添加具有最大外部重要度的属性,可以得到一个互补决策协调集。由定理 1,若该互补决策协调集中的每个属性都是必要的,则该互补决策协调集为一个约简。基于上述理论,文献[10]提出了互补决策约简启发式算法,用来计算多标记决策表的一个互补决策约简。

定理 2　设 $S=(U,A,L)$ 是多标记决策表,$B\subseteq A$。对任意 $a,b\in A-B$,如果 $Sig^{\text{outer}}(a,B,L,U)\geqslant Sig^{\text{outer}}(b,B,L,U)$,那么 $Sig^{\text{outer}}(a,B,L,U')\geqslant Sig^{\text{outer}}(b,B,L,U')$,其中 $U'=U-POS_B^U(L)$。

定理 2 表明在多标记决策表中,在原始数据集和去掉正域的数据集上,属性基于外部重要度的序是保持不变的。因此,在约简的启发式算法中,在去掉正域的数据集上具有最大外部重要度的属性也是原始数据集上具有最大外部重要度的属性。所以,该定理保证了去掉正域的数据集的约简和原始数据集的约简是相同的,其为互补决策约简的加速算法提供了理论保证。

3.2　互补决策约简的加速算法

互补决策约简的加速算法,其主要思想为:首先计算属性内部重要度来寻找多标记决策表的核属性集 RED。在核属性集的基础上,逐次从数据集中去掉已选属性集对应的正域,然后在剩下的数据集上选择外部重要度最大的属性,依次放入属性集 RED 中,直到该特征子集满足停止准则,就得到了多标记决策表中的一个互补决策约简 RED。

算法 1　互补决策约简的加速算法(A-CDR)。

输入:多标记决策表 $S=(U,A,L)$;

输出:S 的一个互补决策约简 RED。

1. 令 $RED=\varnothing$;

2. for ($k=1;k\leqslant|A|;k$++)

{计算 $\mathrm{Sig}^{\mathrm{inner}}(a_k,A,L,U)$;

如果 $\mathrm{Sig}^{\mathrm{inner}}(a_k,A,L,U)>0$,

则 $RED=RED\cup\{a_k\}$;}

3. 令 $i=1,R_1=RED,B_1=\{R_1\},U_1=U$;

4. while $\gamma_L^U(RED)\neq\gamma_L^U(A)$ do

{计算正域 $POS_{B_i}^U(L)$;

$U_{i+1}=U-POS_{B_i}^U(L)$;

$i=i+1$;

依次计算并选取

$Sig^{\mathrm{outer}}(a_0,RED,L,U_i)=\max\{Sig^{\mathrm{outer}}(a_j,RED,L,U_i),a_j\in A-RED\}$

$RED=RED\cup\{a_0\}$;

$R_i=R_{i-1}\cup\{a_0\}$;

$B_i=\{R_1,R_2,\cdots,R_i\}$;}

5. 输出属性约简结果 RED。

下面分析加速算法 A-CDR 的时间复杂度。本文使用文献[22]中的方法计算等价类,其时间复杂度为 $O(|U||A|)$,这里 $|U|$ 和 $|A|$ 分别表示对象个数和属性个数。所以第二步计算核的时间复杂度是 $O(|U||A|^2)$;第四步从核属性集开始,每次循环将具有最大外部重要度属性放入 RED 中,直到找到一个约简,其时间复杂度为 $O(\sum_{i=1}^{|A-RED|}|U_i|(|A-RED|-i+1))$;其余步骤的时间复杂度都是常数,所以整个算法的时间复杂度为 $O(|U||A|^2+\sum_{i=1}^{|A-RED|}|U_i|(|A-RED|-i+1))$。

4　实验及分析

本节在 6 组多标记公开测试集上比较互补约简启发式算法(CDR)和加速算法(A-CDR)的计算性能。表 2 为数据集的详细信息。

表 2　数据集

Table 2　Datasets

数据集	样本数	属性数	标记数	领域
Music	593	72	6	多媒体
Yeast	2 417	103	14	多媒体
Scene	2 407	294	6	生物
Genbase	662	1 185	27	生物
Medical	978	1 449	45	文本
LangLog	1 460	1 004	75	文本

本次实验在硬件配置为 Intel(R) Core (TM) i5-5200U CPU @2.20 GHz,内存 4 GB 的计算机上,用 Matlab 语言编程实现算法。依赖函数中的参数 λ 设置为 0.1。

表 3 列出了两种算法在 6 个数据集上属性约简的结果和计算时间。

由表 3 可知,在同一个数据集上,加速算法可以得到和原始算法相同的约简,但加速算法的计算时间明显减少。

表 3　CDR 算法和 A-CDR 算法的约简结果及计算时间

Table 3　The computation time and reduction results of the algorithms CDR and A-CDR

数据集	原始属性数	CDR 算法		A-CDR 算法	
		选择属性数	时间/s	选择属性数	时间/s
Music	72	8	2 519.100 0	8	1 026.700 0
Yeast	103	9	70 147.000 0	9	25 587.000 0
Scene	294	8	199 910.000 0	8	69 442.000 0
Genbase	1 185	32	37 270.000 0	32	18 353.000 0
Medical	1 449	59	503 076.303 5	59	72 550.893 6
LangLog	1 004	36	343 066.277 3	36	94 779.301 3

为了更好地比较两种算法计算约简的时间,表 2 中的每个数据集都被平均分为 20 份,记为 $x_i(i=1,2,\cdots,20)$,实验所使用的 20 份数据中的每一份记为 $X_i(i=1,2,\cdots,20)$,其中 $X_1=x_1$,$X_2=x_1\cup x_2$,$\cdots$,$X_{20}=x_1\cup x_2\cdots\cup x_{20}$,最后一份数据即是数据集本身。

图 1 是两种算法分别在 6 组数据集上的约简计算时间,其中 X 轴表示上述 20 份样本数目由少到多的数据集 $X_i(i=1,2,\cdots,20)$,Y 轴表示算法在不同数据集上的计算时间(单位:秒)。

如图 1 中(a)~(f)实验结果所示,随着数据集规模的增加,原始算法和加速算法计算约简的时间也在增加。样本的数据规模越大,两种算法计算约简的时间消耗差值就越大。这表明加速算法有效地加速了属性约简的过程,对于大规模的数据集来说,加速算法的计算效率更高,更加高效。

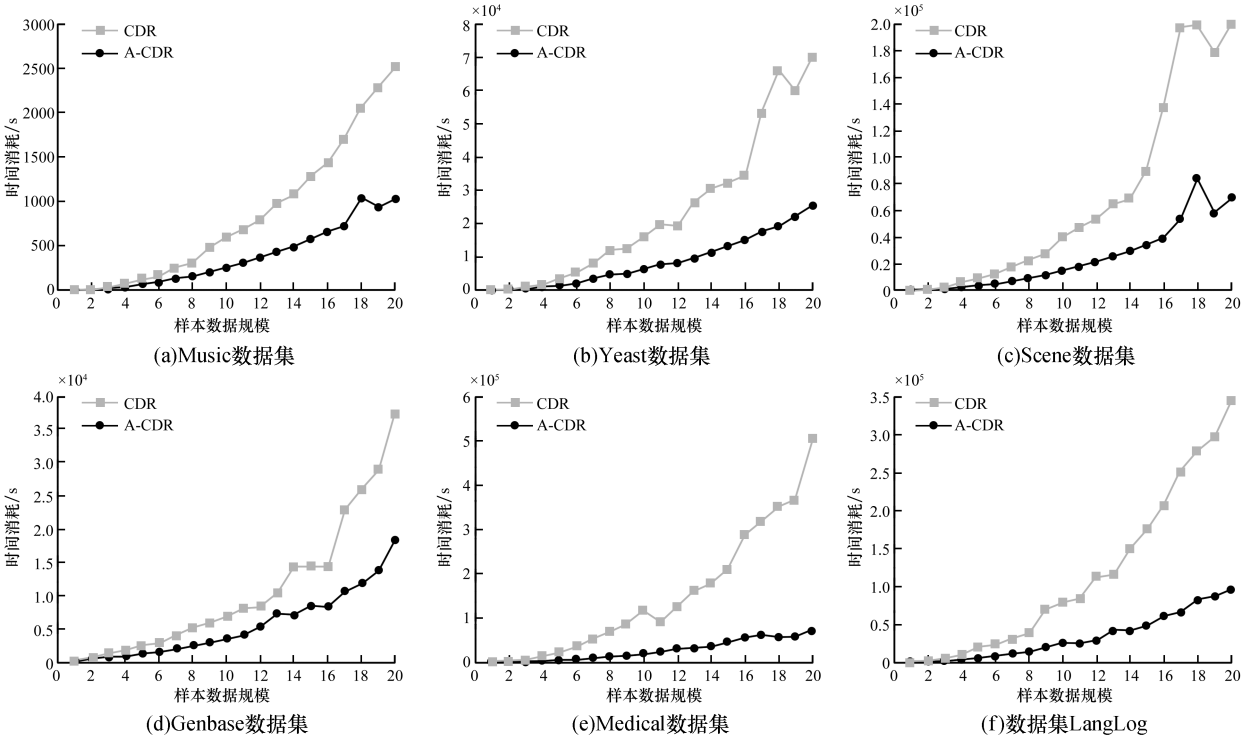

图 1　CDR 与 A-CDR 的计算时间

Fig. 1　Computation time of CDR and A-CDR

5　结　论

针对互补决策约简启发式算法计算耗时过大的缺陷,本文首先研究了动态粒度下互补决策约简的属性外部重要度的保序性质,并基于该性质,提出了互补决策约简启发式算法的加速算法。该加速算法通过逐步缩小数据规模来降低计算耗时,加速了属性约简的过程,并且可以得到与原始算法相同的约简。最后,在多标记数据集上有效地验证了加速算法的有效性。

参考文献:

[1] BOUTELL M R, LUO J, SHEN X, et al. Learning multi-label scene classification[J]. Pattern recognition, 2004, 37(9): 1757-1771.

[2] TROHIDIS K, TSOUMAKAS G, KALLIRIS G, et al. Multi-label classification of music by emotions[J]. EURASIP journal on audio, speech and processing, 2011(4): 1-9.

[3] SCHAPIRE R E, SINGER Y. BoosTexter: A boosting-based system for text categorization[J]. Machine learning, 2000, 39(2): 135-168.

[4] CHEN G, YE D, XING Z, et al. Ensemble application of convolutional and recurrent neural networks for multi-label text categorization[C]//Proc of the international joint conference on neural networks. Washington, USA: IEEE, 2017: 2377-2383.

[5] XIA S, CHEN P, ZHANG J, et al. Utilization of rotation-invariant uniform LBP histogram distribution and statistics

of connected regions in automatic image annotation based on multi-label learning[J]. Neurocomputing,2017,228:11-18.

[6] WANG X,SUKTHANKAR G. Multi-label relational neighbor classification using social context features[C]//Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining. New York,USA:ACM,2013:464-472.

[7] DUDA R O,HART P E,STORK D G. Pattern classification[M]. 2nd ed. New York:Wiley,2001:25-68.

[8] PAWLAK Z. Rough sets:Theoretical aspects of reasoning about data[M]. Boston:Kluwer Academic Publishers,1991:32-58.

[9] 段洁,胡清华,张灵均,等. 基于邻域粗糙集的多标记分类特征选择算法[J]. 计算机研究与发展,2015,52(1):56-65.

[10] LI H,LI D,ZHAI Y,et al. A novel attribute reduction approach for multi-label data based on rough set theory[J]. Information sciences,2016,367-368:827-847.

[11] LIN Y,LI Y,WANG C,et al. Attribute reduction for multi-label learning with fuzzy rough set[J]. Knowledge-based systems,2018,152:51-61.

[12] LI Y,LIN Y,LIU J,et al. Feature selection for multi-label learning based on kernelized fuzzy rough sets[J]. Neurocomputing,2018,318:271-286.

[13] FAN X,CHEN Q,QIAO Z,et al. Attribute reduction for multi-label classification based on labels of positive region[J]. Soft computing,2020,24(18):14039-14049.

[14] QIAN W,HUANG J,WANG Y,et al. Label distribution feature selection for multi-label classification with rough set[J]. International journal of approximate reasoning,2021,128:32-55.

[15] LIANG J,QIAN Y,CHU C,et al. Rough set approximation based on dynamic granulation[J]. Rough sets,fuzzy sets,data mining,and granular computing,2005,3641:701-708.

[16] GHAMRAWI N,MCCALLUM A,NADIA G,et al. Collective multi-label classification [C]//Proceedingds of the 2005 ACM conference on information and knowledge management. Bremen,Germany:ACM,2005:195-200.

[17] JESSE R. Scalable multi-label classification[D]. New Zealand,Hamilton:University of Waikato,2010:1-53.

[18] SHAO H,LI G,LIU G,et al. Symptom selection for multi-label data of inquiry diagnosis in traditional Chinese medicine[J]. Science China information sciences,2013,56(5):1-13.

[19] POLKOWSKI L. On convergence of rough sets[M]. Dordrecht:Springer Netherlands,1992:305-311.

[20] QIAN Y,LIANG J,DANG C. Converse approximation and rule extraction from decision tables in rough set theory[J]. Computer and mathematics with applications,2008,55:1754-1765.

[21] QIAN Y,LIANG J,PEDRYCZ W,et al. Positive approximation:An accelerator for attribute reduction in rough set theory[J]. Artificial intelligence,2010,174 (9/10):597-618.

[22] 徐章艳,刘作鹏,杨炳儒,等. 一个复杂度为 $\max(O(|U||C|),O(|C|^2|U/C|))$ 的快速属性约简算法[J]. 计算机学报,2006,29(3):391-399.

(上接第 59 页)

[10] 董辉,罗潇,罗正东,等. 碎石土三轴测试仿真建模及试样尺寸效应分析[J]. 地质力学学报,2016,22(1):104-113.

[11] 马石城,胡军霞,马一跃,等. 基于三维离散元堆积碎石土细-宏观力学参数相关性研究[J]. 计算力学报,2016,33(1):73-82.

[12] 董辉,陈玺文,傅鹤林,等. 堆积碎石土剪切特性的三轴试验[J]. 长安大学学报(自然版),2015,35(2):59-66.

[13] 孙雅珍,李凯翔,丁传超,等. 稳定碎石土底基层材料力学参数试验研究[J]. 中外公路,2018,38(1):248-253.

[14] 苏立君,梁双庆,王洋. 震后降雨型碎石土斜坡稳定性的试验研究[J]. 工程科学与技术,2019,51(4):12-20.

[15] 梁双庆,苏立君. 不同地下水位碎石土斜坡对震动的差异性动力响应[J]. 山地学报,2018,36(1):83-90.

[16] 吴锐,邓清禄,付敏,等. 碎石尺寸对碎石土强度影响的大型直剪试验研究[J]. 长江科学院报,2016,33(8):80-85.