

基于图的覆盖决策信息系统属性约简新算法

张 杰¹, 张燕兰^{1*}, 林艺东²

(1. 闽南师范大学 计算机学院, 福建 漳州 363000;

2. 厦门大学 数学科学学院, 福建 厦门 361005)

摘 要:属性约简是粗糙集理论的重要研究内容之一,主要是用于去除高维数据的冗余信息。利用图论求约简是覆盖决策信息系统属性约简的一个新方法,其将覆盖决策系统的约简等价于超图的极小顶点覆盖。本研究提出一种新的覆盖决策信息系统的属性约简算法,该算法采用三步策略:首先确定覆盖决策信息系统的辨识集,然后得到超图的关联矩阵,最后用贪心法求出超图的极小顶点覆盖。实验结果表明,新的属性约简算法能够有效地降低数据维数和时间复杂度。

关键词:属性约简;顶点覆盖;图论;覆盖粗糙集;关联矩阵

中图分类号:TP18

文献标志码:A

文章编号:1674-4942(2022)01-0016-09

A New Attribute Reduction Algorithm for Covering Decision Information System Based on Graph

ZHANG Jie¹, ZHANG Yanlan^{1*}, LIN Yidong²

(1. School of Computing, Minnan Normal University, Zhangzhou 363000, China;

2. School of Mathematical Sciences, Xiamen University, Xiamen 361005, China)

Abstract:At present, attribute reduction is one of the important research issues for rough set theory and is mainly used to remove redundant information of high-dimensional data. It is proved that solving the attribute reduction of covering decision information system is equivalent to finding the minimal vertex covering of a hypergraph. Our new attribute reduction algorithm was proposed with a three-step strategy. Firstly, the discernibility set of the covering decision information system was determined, then the incidence matrix of a hypergraph was specified and finally the minimal vertex cover of the hypergraph with greedy method was found. Experimental results showed that the new attribute reduction algorithm can reduce the dimension of data as well as the time complexity effectively.

Keywords: attribute reduction; vertex cover; graph theory; covering roughset; incidence matrix

由Pawlak^[1]提出的粗糙集理论是一种处理不确定信息的重要工具,在数据挖掘、机器学习和知识发现等领域有重要的应用。在粗糙集理论中,数据通常由信息系统来表示,并且由一组特征来描述。然而,这些特征中有许多是冗余的,这使得分类过程更加困难。因此,通过选择相关的特征来减少冗余的特征的数量是很有必要的。属性约简的主要目的是去除数据集中的冗余信息,这不仅可以减少学习算法的运行时间,在某些情况下还可以提供更好的分类精度。

在经典的粗糙集中,很多学者已对属性约简作了深入的研究^[2-4]并取得了重要的结果。例如,钱宇华等提出了4种寻找属性约简的加速算法^[5]。然而,基于经典Pawlak粗糙集模型的属性约简方法只适用于离散型数据,并不能直接应用于连续型数据。换句话说,属性约简之前需要对连续型数据进行离散化,而这会造成

收稿日期:2021-03-04

基金项目:福建省自然科学基金项目(2019J01749,2019J01748)

第一作者:张杰(1990—),江苏宿迁人,硕士研究生,研究方向为覆盖粗糙集。E-mail:13142832142@163.com

*通信作者:张燕兰(1983—),福建漳州人,教授,研究方向为不确定性处理等。E-mail:zhangppff@126.com

信息损失和丢失^[6]。为此,很多学者提出了覆盖粗糙集、相似关系粗糙集和邻域粗糙集等广义粗糙集模型^[7-13]。Chen等提出了一种基于覆盖粗糙集模型的属性约简方法^[14],此方法可以用来处理数值型数据,而且不需要对数值型数据离散化,只需转化为一个覆盖决策系统即可。在此基础上,Chen等又提出了另一种基于覆盖粗糙集模型的属性约简方法^[14],但是,该算法非常耗时。为了克服这个缺点,Wang等简化了文献[14]中辨识矩阵的构造,提出了一种更有效的寻找属性约简的CDA算法^[15]。CDA算法可以有效地处理中等规模的数据集,然而,CDA算法的计算复杂度随着样本和属性的增加呈指数增长,且CDA算法可能会得到一个空的特征子集,因此,对于较大数据集,需要更有效的属性约简方法。于是,Chen等提出了一种基于图的覆盖决策信息系统属性约简的CDG算法^[16],结果表明CDG算法比文献[15]中CDA算法更高效。

本研究在文献[15]的基础上研究覆盖决策信息系统属性约简的问题,并在文献[16]的约简理论框架下研究基于图的覆盖信息系统属性约简的新方法,采用的策略是:首先计算覆盖决策信息系统的辨识集,进而得到一个超图的关联矩阵;然后,基于贪心法求该超图的极小顶点覆盖;最后,通过连续型数据实验分析得知新算法的时间复杂度为 $O(|U||A|)$,其中 $|U|$ 与 $|A|$ 的含义分别是论域的大小、属性集的大小,这比CDG算法的时间复杂度 $O(|U|^2|A|)$ 更低。

1 预备知识

1.1 覆盖决策信息系统的属性约简

在本节中,先介绍覆盖粗糙集和图论的几个基本概念。称 $S = (U, A)$ 是一个信息系统,其中 U 和 A 分别是非空有限论域和属性集。对于任意属性 $a \in A$,称 $a: U \rightarrow V_a$ 是信息函数,其中 V_a 称为属性 a 的值域。决策信息系统的形式为 $S = (U, A, \{d\})$,其中 (U, A) 是一个信息系统, A 是条件属性集, $d: U \rightarrow V_d$ 是决策属性,且 $d \notin A$ 。

众所周知,经典粗糙集只能用于处理离散型数据。然而,在许多实际的情况中经常要面临连续型数据的情况。现在回顾相关文献中处理连续型数据的属性约简方法^[14-15,17]。

定义 1^[17] 设 U 为论域, C 是 U 的子集族,如果 C 中不含空集,且 $\cup C = U$,则称 C 为 U 的一个覆盖。

定义 2^[14] 设 $C = \{K_1, K_2, \dots, K_n\}$ 为 U 的一个覆盖,对于每一个 $x \in U$,令 $C_x = \cap \{K_i | K_i \in C, x \in K_i\}$,则 $\text{Cov}(C) = \{C_x | x \in U\}$ 也是 U 的一个覆盖,称 $\text{Cov}(C)$ 为 C 诱导的覆盖。设 $\Delta = \{C_i | i = 1, 2, \dots, m\}$ 为 U 的一族覆盖, $\forall x \in U$,记 $\Delta_x = \cap \{(C_i) | x \in C_i, C_i \in \Delta\}$,那么 $\text{Cov}(\Delta) = \{\Delta_x | x \in U\}$ 也是 U 的一个覆盖,称 $\text{Cov}(\Delta)$ 为 Δ 的诱导覆盖。 $\forall X \subseteq U$, X 相对于 Δ 的上、下近似算子定义如下:

$$\begin{aligned}\underline{\Delta}(X) &= \{x \in U | \Delta_x \subseteq X\}, \\ \overline{\Delta}(X) &= \{x \in U | \Delta_x \cap X \neq \emptyset\}.\end{aligned}$$

定义 3^[15] 设 $\Delta = \{C_i | i = 1, 2, \dots, m\}$ 为 U 的覆盖族, d 为决策属性。假设 $U/d = \{D_1, D_2, \dots, D_r\}$ 是由决策属性 d 导出的等价类,称三元组 $S = (U, \Delta, \{d\})$ 为一个覆盖决策信息系统。决策属性 d 相对于覆盖族 Δ 的正域为: $\text{POS}_\Delta(d) = \bigcup_{i=1}^r \underline{\Delta}(D_i)$ 。若 $\text{POS}_\Delta(d) = U$,则称 $S = (U, \Delta, \{d\})$ 为协调覆盖决策信息系统。若 $\text{POS}_\Delta(d) \neq U$,则称 $S = (U, \Delta, \{d\})$ 为不协调覆盖决策信息系统。

通常,对每个数值型属性 $a \in A$,可以定义每个样本 $x \in U$ 的邻域。 $N_a(x, \varepsilon) = \{y \in U | d(x, y) \leq \varepsilon\}$,其中 $d(\cdot, \cdot)$ 是一个距离函数, ε 是指定的阈值。通常, $d(x, y) = |a(x) - a(y)|$ 。显然, $N_a = \{N_a(x, \varepsilon) | x \in U\}$ 是 U 的一个覆盖, $\Delta = \{N_a | a \in A\}$ 是 U 的一个覆盖族。因此,可以获得一个覆盖决策信息系统 $S = (U, \Delta, \{d\})$ 。

例 1 设 $S = (U, A, \{d\})$ 为一个由连续型属性组成的决策信息系统(见表1),其中 $U = \{x_1, x_2, x_3, x_4, x_5\}$, $A = \{a_1, a_2, a_3, a_4\}$, d 是决策属性。

令阈值 $\varepsilon = 0.20$, 可以获得以下覆盖决策信息系统 $S = (U, \Delta, \{d\})$:

$$C_1 = N_{a_1} = \left\{ \{x_1, x_3, x_4, x_5\}, \{x_2, x_5\}, \{x_1, x_3, x_4\}, \{x_1, x_2, x_3, x_5\} \right\},$$

$$C_2 = N_{a_2} = \left\{ \{x_1, x_5\}, \{x_2, x_3\}, \{x_4, x_5\}, \{x_1, x_4, x_5\} \right\},$$

$$C_3 = N_{a_3} = \left\{ \{x_1, x_3, x_4\}, \{x_2\}, \{x_5\} \right\},$$

$$C_4 = N_{a_4} = \left\{ \{x_1\}, \{x_2, x_4\}, \{x_3\}, \{x_2, x_4, x_5\}, \{x_4, x_5\} \right\} \circ$$

另外, $U/d = \left\{ \{x_1, x_3\}, \{x_2, x_4, x_5\} \right\}$, 记 $D_1 = \{x_1, x_3\}$, $D_2 = \{x_2, x_4, x_5\} \circ$

表1 数值型数据表

Table 1 Table of numerical data

U	a_1	a_2	a_3	a_4	d
x_1	0.45	0.65	0.30	0.70	0
x_2	0.10	0.15	0.95	0.40	1
x_3	0.50	0.05	0.25	0.95	0
x_4	0.65	0.90	0.20	0.30	1
x_5	0.30	0.80	0.65	0.10	1

由 $\text{POS}_\Delta(d) = \underline{\Delta}(D_1) \cup \underline{\Delta}(D_2) = \{x_1, x_3\} \cup \{x_2, x_4, x_5\} = U$, 可知覆盖决策系统 S 应为协调覆盖决策信息系统。给定一个覆盖决策信息系统 S , 通过构造辨识矩阵, 可以得到一个覆盖决策信息系统的所有约简。

定义 4^[15] 设 $S = (U, \Delta, \{d\})$ 为覆盖决策信息系统, 其中 $U = \{x_1, x_2, \dots, x_n\}$, $\Delta = \{C_i | i = 1, 2, \dots, m\}$ 。定义

$$M(x, y) = \begin{cases} \{C \in \Delta | y \notin C_x\}, (x, y) \in DIS; \\ \Delta, (x, y) \notin DIS, \end{cases}$$

其中 $DIS = \{(x, y) \in U \times U | x, y \in \text{POS}_\Delta(d), d(x) \neq d(y)\}$, 并称 $M(x, y)$ 为 x 和 y 的可分辨属性集, $M = \{M(x, y) | (x, y) \in U \times U\}$ 为 S 的辨识集。辨识集 M 也可以表示为一个 $n \times n$ 的矩阵, 其元素为 $M(x, y)$ 。

例 2(续例 1) 覆盖决策信息系统 S 的辨识矩阵如表 2 所示。为简单起见, 我们对集合使用无分隔符形式, 例如, 用 $C_1C_2C_4$ 表示 $\{C_1, C_2, C_4\}$ 。

表2 覆盖决策信息系统 S 的辨识矩阵

Table 2 Discernibility matrix of S

	x_1	x_2	x_3	x_4	x_5
x_1	Δ	Δ	Δ	$C_1C_2C_4$	$C_1C_3C_4$
x_2	Δ	Δ	$C_1C_3C_4$	Δ	Δ
x_3	Δ	$C_1C_3C_4$	Δ	$C_1C_2C_4$	Δ
x_4	C_2C_4	Δ	C_2C_4	Δ	Δ
x_5	Δ	Δ	Δ	Δ	Δ

在覆盖决策信息系统 $S = (U, \Delta, \{d\})$ 中, 覆盖 $C_1, C_2, \dots, C_m$ 分别对应于 m 个布尔变量 $C_1^*, C_2^*, \dots, C_m^*$, 则定义辨识函数 f_S 如下:

$$f_S(C_1^*, C_2^*, \dots, C_m^*) = \bigwedge \{ \bigvee M | M \in M', M \neq \Delta \}.$$

这种简化析取形式的每一个合取式被称为主蕴涵^[18]。Wang 等通过析取和合取运算证明了覆盖决策信息系统的约简计算可以转化为布尔函数的主蕴涵的计算^[18]。

引理 1^[6,15] 设 $S = (U, \Delta, \{d\})$ 为覆盖决策信息系统。 $\bigwedge_{C_i \in P} C_i^*$ 是辨识函数 f_S 的一个主蕴涵,当且仅当子集 $P \subseteq \Delta$ 是 S 的一个约简。

例 3(续例 1) 从表 3 应用吸收律可得 S 的辨识函数如下:

$$f_S(C_1, C_2, C_3, C_4) = C_4 \vee (C_1 \wedge C_2) \vee (C_2 \wedge C_3)。$$

因此, $\text{Red}(S) = \{\{C_4\}, \{C_1, C_2\}, \{C_2, C_3\}\}。$

表 3 点的邻域
Table 3 The neighborhood of a point

	x_1	x_2	x_3	x_4	x_5
$(C_1)_x$	$\{x_1, x_3\}$	$\{x_2, x_5\}$	$\{x_1, x_3\}$	$\{x_1, x_3, x_4\}$	$\{x_5\}$
$(C_2)_x$	$\{x_1, x_5\}$	$\{x_2, x_3\}$	$\{x_2, x_3\}$	$\{x_4, x_5\}$	$\{x_5\}$
$(C_3)_x$	$\{x_1, x_3, x_4\}$	$\{x_2\}$	$\{x_1, x_3, x_4\}$	$\{x_1, x_3, x_4\}$	$\{x_5\}$
$(C_4)_x$	$\{x_1\}$	$\{x_2, x_4\}$	$\{x_3\}$	$\{x_4\}$	$\{x_4, x_5\}$

1.2 超图的顶点覆盖

超图是传统图的泛化,超图里的边可以连接任意数量的顶点。一般,超图 H 可以表示为 $(V, \mathcal{E})$, 其中 V 是所有顶点元素的集合, $\mathcal{E}$ 是 V 的非空子集族, $\mathcal{E}$ 的元素被称为超边或者边, 用 E 表示。而超图 H 的顶点覆盖 K 是一个集合 $K \subseteq V$, 且 K 与 H 的每条边交非空。换句话说, 顶点覆盖是一组覆盖所有边的顶点集合。如果 K 的任意子集均不是顶点覆盖, 则顶点覆盖 K 是极小的。极小顶点覆盖是顶点数目最少的顶点覆盖。超图 H 的所有极小顶点覆盖的集合用 $T(H)$ 表示。与粗糙集中的属性约简方法相似, 超图的所有极小顶点覆盖也可以通过逻辑表达式得到。

已知超图 $H = (V, \mathcal{E})$, 可以定义一个布尔函数 f_H , 顶点 $v_1, v_2, \dots, v_m$ 用布尔变量 $v_1^*, v_2^*, \dots, v_m^*$ 表示:

$$f_H(v_1^*, v_2^*, \dots, v_m^*) = \bigwedge \{ \bigvee E \mid E \in \mathcal{E} \}。$$

引理 2 给出了一种超图极小顶点覆盖的计算方法。

引理 2^[19] 设 $H = (V, \mathcal{E})$ 是超图。 $\bigwedge_{v_i \in K} v_i^*$ 是布尔函数 f_H 的一个主蕴涵, 当且仅当子集 $K \subseteq V$ 是 H 的极小顶点覆盖。

例 4 设 $H = (V, \mathcal{E})$ 是一个超图, $V = \{v_1, v_2, v_3, v_4\}$, $\mathcal{E} = \{E_1, E_2, E_3\} = \{\{v_2, v_4\}, \{v_2, v_3, v_4\}, \{v_1, v_3, v_4\}\}。$ 化简后得到主蕴涵中的布尔函数 f_H 为:

$$f_H(v_1^*, v_2^*, v_3^*, v_4^*) = v_4 \vee (v_1 \wedge v_2) \vee (v_2 \wedge v_3)。$$

因此, $T(H) = \{\{v_4\}, \{v_1, v_2\}, \{v_2, v_3\}\}。$

2 基于图的覆盖决策信息系统的属性约简

从引理 1 和引理 2 可以看出, 覆盖决策信息系统的属性约简与超图的极小顶点覆盖之间存在联系。在本节中, 我们首先介绍由覆盖决策信息系统诱导出的超图, 然后讨论导出超图的极小顶点覆盖与覆盖决策信息系统属性约简之间的关系, 最后, 提出一种基于图的覆盖决策信息系统属性约简的近似算法。

2.1 覆盖决策信息系统中的超图

定义 5^[16] 设 $S = (U, \Delta, \{d\})$ 为覆盖决策信息系统, 其中 $\Delta = \{C_1, C_2, \dots, C_m\}$, M' 是 S 的可分辨识集, $M^* = \{M \in M' \mid M \neq \Delta\}$ 。令 $V = \Delta$, $\mathcal{E} = M^*$, 称 $H = (V, \mathcal{E})$ 为 S 的诱导图。

定义 5 是在文献[3]的基础上给出的, 仅适用于分类数据。回想一下, 超图是传统图的泛化。覆盖决策信

息系统的诱导图 H 可能是一个传统的图,但在大多数情况下,它是一个超图。因此,如果不存在歧义,我们就把定义5中的诱导图称为超图。

例5 (续例1) 易知: $M^* = \{\{C_2, C_4\}, \{C_1, C_2, C_4\}, \{C_1, C_3, C_4\}\}$ 。 S 诱导的超图 $H = (V, \varepsilon)$ 为: $\varepsilon = \{\{C_2, C_4\}, \{C_1, C_2, C_4\}, \{C_1, C_3, C_4\}\}$, $V = \{C_1, C_2, C_3, C_4\}$ 。

通过引理1、引理2以及定义5,可得到下面的结果。

定理1^[16] 设 $H = (V, \varepsilon)$ 为覆盖决策信息系统 S 诱导的一个超图,则 $\text{Red}(S) = T(H)$ 。

定理1给出一个覆盖决策信息系统属性约简的图模型。可以看到,求一个覆盖决策信息系统的所有约简可以看作是求一个超图极小顶点覆盖集,这为我们提供了获取覆盖决策信息系统属性约简的新方法。

例6 (续例5) 知: $\text{Red}(S) = \{\{C_4\}, \{C_1, C_2\}, \{C_2, C_3\}\}$ 。从例4诱导的超图所有极小顶点覆盖集合为: $T(H) = \{\{C_4\}, \{C_1, C_2\}, \{C_2, C_3\}\}$ 。

因此,有 $\text{Red}(S) = T(H)$ 。

2.2 覆盖决策信息系统的属性约简算法

在实际应用中,没有必要找出覆盖决策信息系统的所有约简,只需要找到一个约简。此外,相较于效率来说,找到约简是否为最小并不是很重要。在大多数文献中,粗糙集属性约简算法得到的是次优约简。在本小节中,将给出一种新的基于图的覆盖决策系统属性约简的近似算法,其中生成覆盖决策表的超图是构造上述约简算法的关键步骤之一。接下来将用关联矩阵的简单表示设计一个快捷计算超图的算法。超图 $H = (V, \varepsilon)$ 的关联矩阵是一个 $m \times n$ 的矩阵 $M_H = (m_{ij})_{m \times n}$,其中 $|\varepsilon| = m$, $|V| = n$ 。若边 E_i 和顶点 v_j 是关联的,则 $m_{ij} = 1$;否则, $m_{ij} = 0$ 。

例7 (续例4) 例4所示的超图关联矩阵是一个由3行(对应于3条边 $E_1 \sim E_3$)和4列(对应于4个顶点 $V_1 \sim V_4$)组成的矩阵,见表4。

表4 例4中超图 H 的关联矩阵
Table 4 Incidence matrix of hypergraph H of example 4

	V_1	V_2	V_3	V_4
E_1	0	1	0	1
E_2	0	1	1	1
E_3	1	0	1	1

下面给出生成覆盖决策信息系统的超图的算法:

算法1 生成覆盖决策信息系统的超图(CDG1)

输入:一个决策信息系统 $S = (U, \Delta, \{d\})$ 和 ε ; ε 是一个阈值;

输出:超图 M_H 。

1. for $x \in U$ do

2. if $d(x) \neq d(y)$ then

3. $M'(x, y) = \{ \text{data} - \text{title}(\text{data}) \mid > \varepsilon \} // \text{title}(\text{data})$ 对原数据集的每一行复制并平铺

4. $M_H G(M'(x, y))$

5. end if

6. end for

7. 删除 M_H 多余的行,且删除和等于0或 $|A|$ 的那一行。

由于CDG1算法仅有一层循环,其计算的时间复杂度为 $O(|U||A|)$,相较于文献[16]中算法1的时间复杂度 $O(|U|^2|A|)$ 更低。值得注意的是由算法1生成的辨识集 $M' = \{M'(x, y) \mid x, y \in U\}$ 与定义4中定义的辨识集不

一样。但是, M' 比定义4中 M 更容易实现, 因为对于任何覆盖 $C \in \Delta$ 都不需要再计算 C_x 。事实上, 它们有如下关系:

命题 1^[16] 设 $S = (U, \Delta, \{d\})$ 是一个覆盖决策信息系统。 M' 和 M 是分别由算法1和定义4生成的辨识集。对于任意 $M'(x, y) \in M'$, 应有 $M'(x, y) \subseteq M(x, y) \in M$ 。从算法1可以发现, 对于每个 $M'(x, y) \in M'$, $M'(x, y)$ 满足如下关系:

$$M'(x, y) = \{a \in A \mid |a(x) - a(y)| > \varepsilon\} = \{a \in A \mid y \notin N_a(x, \varepsilon)\},$$

容易看出: $y \notin N_a(x, \varepsilon) \Rightarrow y \notin \bigcap_{z \in N_a(x, \varepsilon)} N_a(z, \varepsilon)$, 可见应该有 $M'(x, y) \subseteq M(x, y)$ 。

命题 2^[16] 令 $H = (V, \varepsilon)$ 和 $H' = (V, \varepsilon')$ 是两个超图, 满足任何 $E \in \varepsilon, E' \in \varepsilon'$, 则 $E' \subseteq E$ 。如果 $K \subseteq V$ 是 H' 的极小顶点覆盖, 那么 K 也是 H 的一个顶点覆盖。

根据命题1和2, 得到以下结果:

定理 2^[16] 设 $S = (U, A, \{d\})$ 为决策信息系统。 M' 和 M 分别是由算法1和定义4生成的两个辨识集。如果 $P \subseteq \Delta$ 是 M' 的一个约简, 则 $\text{POS}_P(d) = \text{POS}_\Delta(d)$ 。

定理2为我们构造新的属性约简算法提供了基础。结合定理1可知, 由算法1导出超图的极小顶点覆盖可以得到一个决策表的近似约简。根据上述结果, 我们设计一种新的覆盖决策信息系统属性约简近似算法2。

算法 2 基于图的覆盖决策信息系统的约简算法(CDG2)

输入: 一个决策信息系统 $S = (U, A \cup \{d\})$ 和 ε ;

输出: 一个约简 Red。

1. 生成超图 $M_H = (V, \varepsilon)$; // 参见算法1

2. 令 $\text{Red} = \emptyset$

3. while $\varepsilon \neq \emptyset$ do

4. $v_0 = \arg \max \{d_{M_H}(v) \mid v \in V_x\}$,

Red $\leftarrow$ Red $\cup \{v_0\}$; // 去掉 ε 中被 Red 所覆盖的边, 并仍记为 ε

5. end while

6. if Red $\rightarrow \{v\}$ // 能覆盖简化图 G 的所有边

7. Red $\leftarrow$ Red $\setminus \{v\}$

8. end if

9. 输出 Red。

该算法的主要思想是基于求超图极小顶点覆盖的贪心法^[9, 14], 步骤1~8的时间复杂度为 $O(|U||A|)$ 。因此, CDG2的时间复杂度为 $O(|U||A|)$, 比文献[16]中算法2的时间复杂度 $O(|U|^2|A|)$ 更低。为了验证该算法的有效性, 将其与文献[16]中的CDG的算法进行了比较。比较的指标有: 生成辨识集的运行时间、生成约简集的运行时间、约简的程度。

3 实验结果与分析

3.1 数据准备

所有的算法都是在Python3.6中实现的, 并在装有Windows 7的个人电脑上运行, 其他要素为: 双核处理器, 采用串行编程, CPU主频为2.4 GHz, 睿频为3.0 GHz, 内存为10 GB。实验选用10个公开的数据集进行验证。具体的数据集描述见表5。

3.2 实验比较

表5记录了从UCI公开数据集中选取10个常见的数据集, 有小样本, 也有高维小样本, 以检验本文算法

在真实数据集上的有效性。具体实验结果如表6~8所示。

表5 实验数据集
Table 5 Experimental data set

序号	数据集	U	C	决策类
1	Aggregation	788	2	7
2	Car	84	11	8
3	Glass	214	9	6
4	Heartstatlog	270	13	2
5	Hepatitis	155	19	2
6	LSVT_voice_re	126	310	2
7	SCADI	70	205	7
8	Sonar	208	60	2
9	Wine	178	13	3
10	WarpAR10P	130	2 400	10

表6 约简后的属性数量
Table 6 Number of attributes after reduction

数据集	CDG	CDG2
Aggregation	2	2
Car	1	1
Glass	1	1
Heartstatlog	1	1
Hepatitis	1	1
LSVT_voice_re	53	53
SCADI	1	1
Sonar	1	1
Wine	1	1
WarpAR10P	2	2

表7 辨识集的运行时间
Table 7 Running time of identification data set

数据集	CDG	CDG2	ms
Aggregation	5.00	<0.10	
Car	3.00	1.00	
Glass	8.01	<0.10	
Heartstatlog	4.00	1.00	
Hepatitis	3.02	<0.10	
LSVT_voice_re	29.02	0.99	
SCADI	77.01	1.00	
Sonar	20.00	<0.10	
Wine	5.00	<0.10	
WarpAR10P	1 770.10	7.00	

表6为CDG2算法和CDG算法对10个数据集具体的约简结果。从表中可以看出,本文提出的CDG2算法和CDG算法都可以获得极小的属性子集,CDG2算法在获得的极小属性子集的数目上也保持着优势,实际上它们均能获得一个真正的约简集。

表7为CDG2算法和CDG算法在10个数据集上产生辨识集的运行时间。从表7可知,本研究提出的CDG2算法在SCADI数据集上得到辨识集的时间是CDG算法的1.29%,在LSVT_voice_re数据集上得到辨识集的时间是CDG算法的3.44%,在WarpAR10P数据集上得到辨识集的时间是CDG算法的0.39%。从实验数据可以看出,CDG2算法在高维数据集WarpAR10P上得到辨识集的时间下降较为明显。从算法角度分析,CDG2算法对于更高维的数据集减少了一层循环,其时间复杂性与低维数据集相比下降会更加明显。CDG算法时间复杂度 $O(|U|^2|A|)$ 为二次函数,而CDG2算法的时间复杂度 $O(|U||A|)$ 为线性函数。当 $|U|$ 增大时,线性函数 $O(|U||A|)$ 比二次函数 $O(|U|^2|A|)$ 的优势愈发明显。由此可见,CDG2算法得到辨识集的运行时间均小于CDG算法,且在上述分析的数据集上运行时间优势更加明显。

表8为CDG2算法和CDG算法在8个数据集上产生约简集的运行时间。从表中可以看出,本研究提出的CDG2算法在Aggregation数据集上得到约简集的时间是CDG算法的14.24%,在Car数据集上得到约简集的时间是CDG算法的21.23%,在Glass数据集上得到约简集的时间优势更加明显。由此可见,CDG2算法计算约简运行时间均小于CDG算法,这些数据结果进一步表明CDG2算法是一种高效的属性约简方法。综上所述可知,CDG2算法优于CDG算法。

表8 约简集的运行时间
Table 8 Running time of reduced set

数据集	CDG	CDG2
Aggregation	146.01	31.00
Car	797.05	190.01
Glass	928.05	281.02
Heartstatlog	242.01	129.01
Hepatitis	727.04	735.04
LSVT_voice_re	244.01	243.01
SCADI	863.05	773.04
Sonar	572.04	198.01

4 结论

属性约简是粗糙集领域重要的研究内容之一。本研究提出了一个基于图的覆盖决策信息系统属性约简的贪心算法,该算法可以直接处理数值型数据。此外,与文献[16]的CDG算法相比,CDG2算法方法更适合处理高维大数据集,同时,CDG2算法有较低的时间复杂度,并通过实例和数值型实验验证了该算法的有效性,该方法可以看作是一种局部逼近最优的特征选择策略。然而,如何进一步获得最小顶点覆盖以生成一个尽可能接近最优解的最小特征子集,这是我们后续的主要研究方向之一。

参考文献:

- [1] PAWLAK Z. Rough sets[J]. International Journal of Computer and Information Sciences, 1982, 11: 341-356.
- [2] LI B Z, WEI Z H, MIAO D Q, et al. Improved general attribute reduction algorithms[J]. Information Sciences, 2020, 536: 298-316.
- [3] WEI W, SONG P, LIANG J Y, et al. Accelerating incremental attribute reduction algorithm by compacting a decision table[J]. International Journal of Machine Learning and Cybernetics, 2019, 10(9): 2355-2373.
- [4] WANG C Z, HUANG Y, DING W P, et al. Attribute reduction with fuzzy rough self-information measures[J]. Information Sciences, 2021, 549: 68-86.

-
- [5] QIAN Y H, LIANG J Y, PEDRYCZ W, et al. Positive approximation: an accelerator for attribute reduction in rough set theory[J]. Artificial Intelligence, 2010, 174(9/10): 597–618.
- [6] JIANG F, SUI Y F. A novel approach for discretization of continuous attributes in rough set theory[J]. Knowledge-Based Systems, 2015, 73: 324–334.
- [7] CHEN D G, HU Q H, YANG Y P. Parameterized attribute reduction with Gaussian kernel based fuzzy rough sets[J]. Information Sciences, 2011, 181(23): 5169–5179.
- [8] CHEN D G, YANG Y Y. Attribute reduction for heterogeneous data based on the combination of classical and fuzzy rough set models [J]. IEEE Transactions on Fuzzy Systems, 2014, 22(5): 1325–1334.
- [9] DAI J H. Rough set approach to incomplete numerical data[J]. Information Sciences, 2013, 241: 43–57.
- [10] TSANG E C C, CHEN D G, YEUNG D S. Approximations and reducts with covering generalized rough sets[J]. Computers & Mathematics With Applications, 2008, 56(1): 279–289.
- [11] 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.
- [12] SKOWRON A, STEPANIUK J, SWINIARSKI R. Modeling rough granular computing based on approximation spaces[J]. Information Sciences, 2012, 184(1): 20–43.
- [13] YAO Y Y, YAO B X. Covering based rough set approximations[J]. Information Sciences, 2012, 200: 91–107.
- [14] CHEN D G, WANG C Z, HU Q H. A new approach to attribute reduction of consistent and inconsistent covering decision systems with covering rough sets[J]. Information Sciences, 2007, 177(17): 3500–3518.
- [15] WANG C Z, HE Q, CHEN D G, et al. A novel method for attribute reduction of covering decision systems[J]. Information Sciences, 2014, 254: 181–196.
- [16] CHEN J K, LIN Y J, LIN G P, et al. Attribute reduction of covering decision systems by hypergraph model[J]. Knowledge-Based Systems, 2017, 118: 93–104.
- [17] BONIKOWSKI Z, BRYNIARSKI E, WYBRANIEC-SKARDOWSKA U. Extensions and intentions in the rough set theory[J]. Information Sciences, 1998, 107(1/2/3/4): 149–167.
- [18] SKOWRON A, RAUSZER C. The discernibility matrices and functions in information systems[M]//Intelligent Decision Support. Dordrecht: Springer Netherlands, 1992: 331–362.
- [19] BONDY J A, MURTY U S R. Graph theory with applications[M]. London: Macmillan Press, 1976.

责任编辑: 刘 炜