

基于代数方程组的属性约简研究

苗夺谦^{1,2}, 周杰^{1,2}, 张楠^{1,2}, 冯琴荣^{1,2}, 王睿智^{1,2}

(1. 同济大学嵌入式系统与服务计算教育部重点实验室, 上海 201804; 2. 同济大学计算机科学与技术系, 上海 201804)

摘要: 属性约简是粗糙集理论重要研究内容之一, 求取决策表所有属性约简已被证明为 NP-难问题. 本文基于吴方法, 从代数方程组角度给出了一种求解所有属性约简的新思路. UCI 数据集和人工数据集实验表明了该新方法的有效性.

关键词: 决策表; 属性约简; 分辨函数; 吴方法; 特征列

中图分类号: TP18 **文献标识码:** A **文章编号:** 0372-2112 (2010) 05-1021-07

Research of Attribute Reduction Based on Algebraic Equations

MIAO Duo-qian^{1,2}, ZHOU Jie^{1,2}, ZHANG Nan^{1,2}, FENG Qin-rong^{1,2}, WANG Rui-zhi^{1,2}

(1. Key Laboratory of Embedded System and Service Computing, Ministry of Education of China, Shanghai 201804, China;

2. Department of Computer Science and Technology, Tongji University, Shanghai 201804, China)

Abstract: Attribute reduction is one of the most important notions in rough set theory. It has been proved that finding all reducts of a decision table is a NP-hard problem. Based on Wu's method, a novel approach to acquire all reducts is put forward from the perspective of algebraic equations in this paper. The efficiency of this novel method can be illustrated by experiments with UCI datasets and synthetic datasets.

Key words: decision table; attribute reduction; discernibility function; Wu's method; characteristic sets

1 引言

属性约简是粗糙集理论^[1]重要研究内容之一, 求取决策表所有属性约简或最小属性约简已被证明为 NP-难问题^[2]. 针对最小属性约简, 研究者从不同角度构造了启发式约简算法求取最优或近似最优解^[3-5], 然这些算法都存在不完备问题^[6]. 虽然启发式属性约简算法具有较优的时间和空间复杂度, 但都只能获取给定决策表的一个属性约简. 不同属性约简下导出的规则集不同, 由一个属性约简导出的规则集仅能反应该决策表部分知识. 为挖掘决策表中所有隐含知识, 需要获取该决策表所有属性约简, 导出各规则集并进行融合. 在启发式属性约简算法基础上, 对条件属性集的所有子集进行枚举, 可得到求取所有属性约简的直接算法, 但该算法具有指数时间复杂度, 若属性规模较大, 直接算法不可行.

Skowron 已证明决策表属性约简与分辨函数质蕴涵项一一对应^[7], 将求解属性约简问题转化为逻辑函数合取范式化析取范式问题, 为求取属性约简给出了一种新的表示. 求解所有属性约简的指数复杂度是制约粗糙集理论发展的关键问题之一, 高效完备的属性约简算法将

进一步推动粗糙集理论的广泛应用.

吴文俊院士^[8]基于中国古代数学独立研究发现多项式方程组的特征列, 并将其成功应用于几何定理机器证明, 开拓了数学机械化的新局面. 文献^[9]在前人研究基础上, 引入输入变换、算法变换思想, 采用吴方法求解 SAT 问题. 这些研究均将逻辑问题转化为多项式方程组零点问题, 借助成熟代数理论进行求解.

基于分辨函数求解决策表所有属性约简, 本质为一类逻辑推理问题. 本文在前人工作启发下, 探讨了从代数方程组观点求解属性约简的新思路, 从输入变换、特征列求解、零点分解、变元定序、迭代策略和算法复杂度分析等各方面详细介绍了基于吴消元法的属性约简思想及其特点. UCI 数据集和人工数据集实验结果表明了该新方法的有效性.

2 基本概念

粗糙集理论和吴方法相关基本概念及其详细介绍请分别参考文献^[1, 8].

定义 1^[1] 给定决策表 $DT = (U, C \cup D, V, \rho)$, $B \subseteq C$, 若:

$$(1) \gamma(B, D) = \gamma(C, D).$$

$$(2) \forall a \in B, \gamma(B - \{a\}, D) \neq \gamma(C, D).$$

则称属性集 B 为决策表 DT 的一个属性约简. 其中 $\gamma(C, D)$ 表示决策属性集 D 相对条件属性集 C 的分类质量.

定义 2^[7] 给定决策表 $DT = (U, C \cup D, V, \rho)$, 论域 $U = \{o_1, o_2, \dots, o_n\}$, 其分辨矩阵定义为: $DM(DT)_{n \times n} = (c_{ij})_{n \times n}$, 其中 c_{ij} 满足:

$$c_{ij} = \begin{cases} \{a \mid a \in C \wedge \rho(o_i, a) \neq \rho(o_j, a)\}, & \Psi \\ \emptyset, & \text{其它} \end{cases}$$

条件 Ψ 满足 $1 \leq j < i \leq n \wedge \rho(o_i, D) \neq \rho(o_j, D)$.

不同类型决策表, 相容或不相容, 其分辨矩阵构造方式不同, 在此不作详细探讨. 若无特殊说明, 本文研究对象均为相容决策表.

定义 3^[7] 给定决策表 $DT = (U, C \cup D, V, \rho)$, 论域 $U = \{o_1, o_2, \dots, o_n\}$, 分辨矩阵为 $DM(DT)_{n \times n}$, 则 DT 的分辨函数定义为:

$$DF(DT) = \bigwedge \{ \bigvee c_{ij} : 1 \leq j < i \leq n, c_{ij} \neq \emptyset \}$$

其中, $\bigvee c_{ij} = \bigvee a (a \in c_{ij})$ 为 c_{ij} 中所有属性的析取.

命题 1^[7] 给定决策表 $DT = (U, C \cup D, V, \rho)$, 论域 $U = \{o_1, o_2, \dots, o_n\}$, 分辨矩阵为 $DM(DT)_{n \times n}$, $B \subseteq C$ 为决策表 DT 的一个属性约简, 当且仅当 B 满足如下两个条件:

$$(1) \forall i, j (1 \leq j < i \leq n), \text{若 } c_{ij} \neq \emptyset, \text{有 } B \cap c_{ij} \neq \emptyset;$$

$$(2) \forall a \in B, \exists i, j, c_{ij} \neq \emptyset, \text{使得 } (B - \{a\}) \cap c_{ij} = \emptyset.$$

命题 2^[7] 给定决策表 $DT = (U, C \cup D, V, \rho)$, 其属性约简与分辨函数 $DF(DT)$ 质蕴涵项一一对应.

由于属性约简只与分辨函数 DF 质蕴涵项有关, 常采用吸收律对 DF 进行简化, 后续讨论均采用简化后的分辨函数.

3 基于吴方法的属性约简算法

本节将从输入变换、特征列求解、零点分解、变元定序、迭代策略和时间复杂度等各方面详细介绍基于吴消元法的属性约简算法思想及其特点.

3.1 输入变换

给定决策表 $DT = (U, C \cup D, V, \rho)$, $|U| = n, |C| = m$, 其分辨函数 $DF = \xi_1 \wedge \xi_2 \wedge \dots \wedge \xi_s$, 其中, $\xi_i \not\subseteq \xi_j (i \neq j)$, $\xi_j = x_{j_1} \vee x_{j_2} \vee \dots \vee x_{j_{m_j}}, x_{j_k} \in C (j = 1, 2, \dots, s; k = 1, 2, \dots, m_j)$ 为一条件属性, 这里称为子句 ξ_j 的变元. 为不至混淆, 可将 $\xi_j (j = 1, 2, \dots, s)$ 理解为子句 ξ_j 中所含变元构成的集合, 即 $\xi_j = \{x_{j_1}, x_{j_2}, \dots, x_{j_{m_j}}\}$; 将 DF 理解为子句集合, 即 $DF = \{\xi_1, \xi_2, \dots, \xi_s\}$.

引理 1 给定决策表 $DT = (U, C \cup D, V, \rho)$, 设 $B \subseteq C$ 为一属性约简, $B = \{x_1, x_2, \dots, x_t\}$, 分辨函数

$DF = \xi_1 \wedge \xi_2 \wedge \dots \wedge \xi_s$, 则有: $\forall x_i \in B (i = 1, 2, \dots, t), \exists \xi_j \in DF$, 使得 $B \cap \xi_j = \{x_i\}$.

引理 1 表明对于属性约简 B 中的每个属性, 都存在分辨函数中的某个子句, 使得该子句与属性约简 B 的交集只含有这一个属性.

文献[9]比较了输入变换与算法变换在问题求解中的优劣. 虽然算法变换相比输入变换具有明显优势, 但是许多问题求解仍然对输入变换依赖较重. 输入变换实质是对问题的不同知识表示, 不同知识表示下问题求解难度差异较大. 实际求解过程中往往将两者结合, 首先通过输入变换完成问题转换, 然后再利用算法变换进行问题求解.

给定分辨函数合取范式 $DF = \xi_1 \wedge \xi_2 \wedge \dots \wedge \xi_s$, 完成如下输入变换:

$$(1) f(x_j) = 1 - x_j, x_j \text{ 为子句中出现的变元.}$$

$$(2) f(\xi_j) = f(x_{j_1} \vee x_{j_2} \vee \dots \vee x_{j_{m_j}}) \\ = f(x_{j_1}) \cdot f(x_{j_2}) \cdot \dots \cdot f(x_{j_{m_j}}).$$

(3) 分辨函数子句集对应的多项式组为: $\Sigma = \{f(\xi_j), j = 1, 2, \dots, s\}$, 其相应多项式方程组为 $\Sigma = 0$, 即 $\{f(\xi_j) = 0, j = 1, 2, \dots, s\}$.

考虑到布尔变元的特殊取值, 各多项式定义在整数模 2 的有限域 Z_2 上. 其中, $x_j = 1$ 表示属性约简包含该变元, $x_j = 0$ 表示属性约简不包含该变元.

定理 1 属性约简集与多项式方程组 $\Sigma = 0$ 解中含有的变元集一一对应.

定理 1 表明, 分辨函数合取范式通过输入变换转化为多项式组, 进而将求解属性约简转换为多项式组零点问题, 从代数方程组角度给出求解新思路.

定理 2 决策表 $DT = (U, C \cup D, V, \rho)$, 若 $x \in C$ 为一核属性, 则经过输入变换构造的多项式组 Σ 中有且仅有一个多项式 $f \in \Sigma$, 使得 f 中只含有变元 x .

至此, 完成输入变换, 将分辨函数合取范式化析取范式问题转换为多项式方程组 $\Sigma = 0$ 的求解问题. 特征列求解是吴消元法解多项式方程组的核心, 结合分辨函数本身特性, 特征列求解只需对变元定性判定, 无需复杂多项式计算.

3.2 特征列求解

由分辨函数各子句构造的多项式组 $\Sigma = \{f(\xi_j), j = 1, 2, \dots, s\}$ 是一类特殊的多项式组: 每个多项式中不同变元仅出现一次, 且变元的最高次幂均为 1. 该特点大大简化了吴消元法中多项式求余操作.

定理 3 给定分辨函数子句多项式 $f(\xi_1)$ 与 $f(\xi_2)$, $f(\xi_1), f(\xi_2) \in \Sigma$, 若 $f(\xi_2)$ 主变元为 x , 则多项式 $f(\xi_1)$ 除 $f(\xi_2)$ 的余式 $\text{rem}(f(\xi_1)/f(\xi_2))$ 有:

$$(1) \text{若 } f(\xi_1) \text{ 中不含有 } x, \text{ 则 } \text{rem}(f(\xi_1)/f(\xi_2)) =$$

$f(\xi_1)$.

(2)若 $f(\xi_1)$ 中含有 x , 则 $\text{rem}(f(\xi_1)/f(\xi_2)) = 0$.

定理 3 表明子句多项式间的求余操作非常简单, 无需任何多项式运算, 只需子句多项式所含变元间简单比较. 这样将定量的计算转变为定性的比较判定, 可大大提高求余运算效率.

定理 4 分辨函数子句多项式组 Σ 的基列即为其特征列.

定理 4 说明, 吴消元法求解过程, 无需在基列基础上迭代求解特征列. 选出多项式组 Σ 的基列, 即找到其特征列.

定理 5 设分辨函数子句多项式组 $\Sigma = \{f_1, f_2, \dots, f_s\}$ 的特征列为 $G = \{g_1, g_2, \dots, g_r\}$, $\text{Core} = \{cx_1, cx_2, \dots, cx_t\}$ 为核属性集, $\forall cx_i \in \text{Core} (i = 1, 2, \dots, t)$, 若 $f_j \in \Sigma (j = 1, 2, \dots, s)$ 含有 cx_i , 则 $f_j \in G$.

根据定理 2 与定理 5, 给定分辨函数子句多项式组 $\Sigma = \{f_1, f_2, \dots, f_s\}$, 可将其分成两部分: $\Sigma = \Sigma_1 \cup \Sigma_2$, 其中 Σ_1 中每个多项式含有一个核属性变元, Σ_2 中每个多项式均不含核属性变元. 显然 $\Sigma_1 \cap \Sigma_2 = \emptyset$, Σ 的零点集结构满足: $\text{zero}(\Sigma) = \text{zero}(\Sigma_1) \wedge \text{zero}(\Sigma_2)$.

由定理 5, 多项式组 Σ_1 零点集易求, $\text{zero}(\Sigma_1) = \{(cx_1 = 1) \wedge (cx_2 = 1) \wedge \dots \wedge (cx_t = 1)\}$. 从而可将多项式组 Σ 的零点集问题转化为其子集 Σ_2 的零点集问题, 降低多项式数目.

定理 6 设分辨函数子句多项式组 $\Sigma = \{f_1, f_2, \dots, f_s\}$ 的特征列为 $G = \{g_1, g_2, \dots, g_r\}$, 多项式 g_i 初式为 $I(g_i) (i = 1, 2, \dots, r)$, $I = I(g_1) \cdot I(g_2) \cdot \dots \cdot I(g_r)$ 为特征列各多项式初式乘积, 则特征列使其初式乘积不为零的零点集 $\text{zero}(G/I)$ 有且仅有一个零点, $\text{zero}(G/I)$ 中的零点可以这样确定: G 中各多项式的主变元取值等于 1 即可.

由定理 6 知, 求特征列使其初式乘积不为零的零点集过程中, 只需求得特征列中各多项式的主变元, 而在给定变元序后, 主变元的判定只需变元间基本比较操作.

3.3 零点分解

命题 3(零点分解定理)^[8] 设多项式组 $\Sigma = \{f_1, f_2, \dots, f_s\}$ 的特征列为 $G = \{g_1, g_2, \dots, g_r\}$, 多项式 g_j 初式为 $I(g_j) (j = 1, 2, \dots, r)$, $I = I(g_1) \cdot I(g_2) \cdot \dots \cdot I(g_r)$ 为特征列各初式乘积, 则多项式组 Σ 的零点集具有结构:

$$\text{zero}(\Sigma) = \text{zero}(G/I) \cup \bigcup_{i=1}^r \text{zero}(\Sigma, I(g_i)).$$

对于多项式组 $\{\Sigma, I(g_i)\}$ 零点集, 可利用命题 3 反复迭代求解, 直至某次迭代过程特征列初式乘积为常数. 称特征列各多项式初式中含有的变元为迭代变元.

由定理 1, 多项式方程组 $\Sigma = 0$ 的解中取值为“1”的变元一定包含于属性约简中, 又由于单元多项式的加入

可加快迭代求解效率, 所以对于分解过程特征列初式乘积中出现的迭代变元 x_i , 只需考虑 $\{\Sigma, (1 - x_i)\}$ 用吴方法继续求解. 若 $\{x_1, x_2, \dots, x_k\}$ 为特征列 G 各多项式初式中出现的变元集, 则分辨函数子句多项式组 Σ 零点集的结构为: $\text{zero}(\Sigma) = \text{zero}(G/I) \cup \bigcup_{i=1}^k \text{zero}(\Sigma, (1 - x_i))$.

由零点分解定理, 分辨函数子句多项式组 Σ 零点分解过程形成一颗树结构, 称为分解树. 每个节点都将产生多项式组 Σ 的一个候选零点 (该零点可能是其他零点的子集, 称产生子集零点的节点为冗余节点), 最终需将所有零点合并去除成为子集的零点. 冗余节点对求解效率具有重要影响, 分解过程产生冗余节点和最终合并过程删除冗余节点均需消耗时间, 故分解过程需要有效避免冗余节点的产生. 叶节点对应的特征列, 其初式乘积为常数 (1 或者 -1), 叶节点不再进一步迭代分解, 分解过程会在有限步终止.

3.4 变元定序

同一多项式组在不同变元序下求得的特征列差异较大, 从而影响后续零点分解效率. 粗糙集理论中可依据属性重要度进行变元定序. 基于吴方法求取属性约简的过程只依赖分辨函数子句多项式组, 在分辨函数基础上定义两种属性重要度:

(1) $\text{SIG}(a) = P_{DF}(a)$, $P_{DF}(a)$ 表示属性 a 在分辨函数 DF 中出现的频度.

(2) $\text{SIG}(a)/L_k = P_{DF}(a)/L_k$, $P_{DF}(a)/L_k$ 表示属性 a 在长度为 k 的子句集中出现的频度. 若多个属性在长度为 k 的子句集中出现频度相同, 则考虑这些属性在长度为 $k+1$ 的子句中出现的频度, 依次下去.

依据属性重要度有两种变元序:

(1) 属性重要度越高, 其在变元序中次序越低.

(2) 属性重要度越高, 其在变元序中次序越高.

直观上, 属性重要度越高的属性需要优先分解出来. 根据吴方法中基列构造算法, 秩最低的多项式先被选取, 故属性重要度越高的属性在变元序中其次序应越低. 不同属性重要度定义下, 对两种变元序都进行了实验, 实验结果参见第四节.

3.5 迭代策略

分析发现零点分解过程若增加迭代控制策略可极大减少分解过程产生的冗余节点, 提高求解效率. 引入如下两种迭代控制策略.

深度迭代策略: 某次分解迭代过程若同时出现多个迭代变元, 则依属性重要度选择迭代变元, 属性重要度高的变元优先进行迭代. 若多项式组 $\Sigma = \{f_1, f_2, \dots, f_s\}$ 分解下一步加入迭代变元 x , 则由定理 5, 只需考虑 $\{\Sigma - \{f_i | x \in f_i, f_i \in \Sigma\}, (1 - x)\}$ 求解即可, 即 $\text{zero}(\Sigma, (1 - x)) = \text{zero}(\Sigma - \{f_i | x \in f_i, f_i \in \Sigma\}, (1 - x))$.

宽度迭代策略:若多项式组 $\Sigma = \{f_1, f_2, \dots, f_s\}$ 在首次分解过程产生迭代变元 $\{x_1, x_2, \dots, x_k\}$, 按属性重要度排序为 $x_1 > x_2 > \dots > x_k$, 依据深度迭代策略, x_i 优先 x_{i+1} 进行迭代. 在变元 x_i 进行深度迭代分解过程后, 包含变元 x_i 的候选零点均已求取. 故多项式组 $\{\Sigma, (1 - x_i)\}$ 深度迭代分解完成后, 对于多项式组 $\{\Sigma, (1 - x_{i+1})\}$ 则不需再次考虑变元 x_i , 即在 x_i 进行迭代后, 对于 x_{i+1} , 只需考虑 $\{\Sigma \setminus \{x_i\}, (1 - x_{i+1})\}$. 其中 $\Sigma \setminus \{x_i\}$ 表示, 若多项式组 Σ 中多项式 $f_j (j = 1, 2, \dots, s)$ 含有变元 x_i , 则从 f_j 中删除变元 x_i . 若从多项式组 Σ 中删除某变元后, 存在某一多项式变为空, 则迭代终止.

吴零点分解过程, 采用深度优先原则. 当根节点确定变元序后, 后续分解过程若不再改变该变元序, 则称此变元序为静态序. 由深度和宽度迭代策略可知, 节点多项式组 $\Sigma_i (i = 1, 2, \dots, N, N$ 为分解产生的节点数) 动态改变, 初始确定的变元序已不能正确反应后续节点多项式组中变元重要性, 故需要针对具体节点多项式组 Σ_i 动态调整变元序, 称此动态改变的变元序为动态序. 实验对静态序和动态序进行了对比分析, 实验结果参见第四节.

3.6 复杂度分析

给定决策表 $DT = (U, C \cup D, V, \rho), |U| = n, |C| = m$. 基于吴方法的属性约简主要分为两步: (1) 生成分辨函数 DF ; (2) 吴零点分解.

(1) 生成分辨函数 DF

算法实现过程, 用一链表存储当前分辨函数 DF . 若生成一非空 DM 元素, 则采用一次吸收律, 始终保持当前 DF 长度最短, 使得后续非空矩阵元素采用吸收律过程中具有最少比较次数.

Komorowski 等人^[10]说明 m 个属性, 其两两互不包含的子集数目最多为 $C_m^{\lfloor m/2 \rfloor}$, 又由于对于 n 个对象, 最多产生 $n \cdot (n - 1) / 2$ 个非空矩阵元素. 故最坏情况, 采用吸收律过程中分辨函数 DF 长度最多为 $\min(C_m^{\lfloor m/2 \rfloor}, n \cdot (n - 1) / 2)$, 故一非空矩阵元素采用吸收律最多需比较 $\min(m \cdot C_m^{\lfloor m/2 \rfloor}, m \cdot n \cdot (n - 1) / 2)$ 次. 所以最坏情况生成分辨函数时间复杂度为 $\min(O(n^2 \cdot C_m^{\lfloor m/2 \rfloor} \cdot m), O(n^2 \cdot n^2 \cdot m))$.

(2) 吴零点分解

设分辨函数子句多项式组为 $\Sigma = \{f_1, f_2, \dots, f_s\}$, 吴零点分解产生的节点数为 N . 每个节点特征列计算中, 首先需要确定每个子句对应多项式在给定序下的秩, 由定理 4, 只需变元间简单比较, 时间复杂度为 $O(m)$, 对于 s 个子句时间复杂度为 $O(s \cdot m)$; 其次判定每个子句是否为基列中元素, 根据吴方法基列构造算法, 时间

复杂度为 $O(s^2 \cdot m)$, 若采用快速排序按秩的大小先对子句进行排序, 时间复杂度可降为 $O(s \cdot \log(s) \cdot m)$.

零点分解过程每个节点都将对应一次特征列计算, 分解总时间复杂度为 $O(N \cdot s \cdot \log(s) \cdot m)$. 由于分辨函数各子句互不包含, 故 $s \leq C_m^{\lfloor m/2 \rfloor}$, 最坏情况, 分解总时间复杂度为 $O(N \cdot C_m^{\lfloor m/2 \rfloor} \cdot \log(C_m^{\lfloor m/2 \rfloor}) \cdot m)$.

候选零点合并需删除成为子集的零点集, 其合并过程可类似 DF 生成过程. 由于分解过程共产生 N 个候选零点, 每个候选零点最多含有 m 个属性, 合并时间复杂度为 $\min(O(N \cdot C_m^{\lfloor m/2 \rfloor} \cdot m), O(N \cdot N \cdot m))$.

综上所述, 基于吴方法的属性约简算法总时间复杂度最坏情况下为: $\min(O(n^2 \cdot C_m^{\lfloor m/2 \rfloor} \cdot m), O(n^4 \cdot m)) + O(N \cdot C_m^{\lfloor m/2 \rfloor} \cdot \log(C_m^{\lfloor m/2 \rfloor}) \cdot m) + \min(O(N \cdot C_m^{\lfloor m/2 \rfloor} \cdot m), O(N^2 \cdot m))$. 生成分辨函数 DF 时间复杂度跟决策表本身有关, 而吴零点分解与分解过程产生的节点数 N 有关, 减少分解过程产生的冗余节点将有益于算法效率的提高.

4 实验分析

实验硬件环境为: CPU: Intel Pentium Dual-Core E2140 1.6GHz; 内存: 1G; 操作系统: Windows XP, 程序在 VC 6.0 环境实现. 实验数据集分为两部分: UCI 数据集和人工数据集.

4.1 UCI 数据集

本文选择了 8 个 UCI 数据集, 其分辨函数和属性约简结果如表 1 所示. 两种属性重要度定义下分别对“属性重要度高变元序高”和“属性重要度高变元序低”两种变元序进行了实验, 同时与随机序进行了比较. 实验结果如表 2 所示.

由表 2 实验结果分析, 两种属性重要度定义下, “属性重要度高变元序低”零点分解过程产生的节点数都要优于“属性重要度高变元序高”和随机序. 绝大多数数据集(除 zoo 数据集外), DFL“属性重要度高变元序低”零点分解过程产生的节点数优于 DFF.

表 3 给出了增加迭代策略以及采用动态序后吴方法求解时间对比. 在静态序下, 加入迭代策略后分解产生的节点数远小于无迭代策略下产生的节点数. 分解过程采用动态序则进一步减少了分解过程产生的冗余节点, 使得产生的节点数更接近实际属性约简数. 由表 3 分析, 零点分解过程中加入迭代策略并且采用动态变元序, 吴方法求解所有属性约简具有最佳效率.

对 UCI 数据集, 由表 3 进一步可知, 求解时间主要集中于分辨函数构造过程. 一旦求得分辨函数, 基于吴方法即可快速求取所有属性约简.

表 1 UGI 数据集基本情况

数据集	对象数	属性数	子句				属性约简				
			Num	Max	Min	Avg	Core	总约简数	最小约简数	最大长度	最小长度
zoo	101	17	14	6	1	3	2	33	7	7	5
wine	178	14	29	6	2	3	0	42	6	8	5
breast	699	10	19	5	1	3	1	20	8	5	4
tic-tac-toe	958	10	36	2	2	2	0	9	9	8	8
chess	3196	37	29	2	1	1	27	4	4	29	29
mushroom	8124	22	26	11	2	5	0	203	3	8	4
monk1	124	7	3	1	1	1	3	1	1	3	3
soybean small	47	36	99	14	6	9	0	765	4	8	2

注:1. Num、Max、Min 和 Avg 分别表示分辨函数子句数目、子句最大长度、子句最小长度和子句平均长度。

2. Core 表示数据集的核属性数目;最小约简指约简集中含有属性数最少的约简。

3. mushroom 数据集去除了含有大量缺省值的属性列,条件属性数减少 1。

表 2 属性序对吴方法求解效率的影响

数据集	随机序		度高序高(DFF)		度高序低(DFF)		度高序高(DFL)		度高序低(DFL)	
	节点数	时间	节点数	时间	节点数	时间	节点数	时间	节点数	时间
zoo	1103	0.125	1983	0.203	246	0.031	4114	0.421	329	0.046
wine	13862	4.312	33813	31.578	6417	1.656	123101	61.828	1849	0.343
breast	326	0.265	268	0.265	134	0.25	331	0.265	129	0.25
tic-tac-toe	9	0.546	9	0.546	9	0.546	9	0.546	9	0.546
chess	5	7.593	5	7.593	5	7.593	5	7.593	5	7.593
mushroom	> 300000	> 300	> 300000	> 300	37785	135.687	> 300000	> 300	36950	134.347
monk1	1	0.015	1	0.015	1	0.015	1	0.015	1	0.015
soybean small	> 300000	> 300	> 300000	> 300	> 300000	> 300	> 300000	> 300	> 300000	> 300

注:1. 时间单位均采用 s,时间 > 300 表示在 300s 内没有求出所有属性约简结果。

2. DFF、DFL 分别表示依据属性重要度(1)和(2)。

表 3 迭代策略对吴方法求解效率的影响

数据集	无迭代策略(静态序)				有迭代策略(静态序)				有迭代策略(动态序)			
	节点数	DF 时间	吴时间	总时间	节点数	DF 时间	吴时间	总时间	节点数	DF 时间	吴时间	总时间
zoo	329	0.015	0.031	0.046	43	0.015	0	0.015	36	0.015	0	0.015
wine	1849	0.046	0.297	0.343	68	0.046	0	0.046	50	0.046	0	0.046
breast	129	0.234	0.016	0.25	32	0.234	0	0.234	28	0.234	0	0.234
tic-tac-toe	9	0.531	0.015	0.546	9	0.531	0	0.531	9	0.531	0	0.531
chess	5	7.593	0	7.593	4	7.593	0	7.593	4	7.593	0	7.593
mushroom	36950	105.628	28.719	134.347	423	105.628	0.062	105.69	232	105.628	0.016	105.644
monk1	1	0.015	0	0.015	1	0.015	0	0.015	1	0.015	0	0.015
soybean small	> 300000	0.015	> 300	> 300	2887	0.015	1.719	1.734	1012	0.015	0.219	0.234

注:时间单位采用 s,时间为 0 表示小于 0.001s;变元序均采用“DFL 属性重要度高变元序低”的原则。

4.2 人工数据集

人工数据集为随机生成,对象数依次从 10 增加到 50,步长为 10;属性数依次从 10 增加到 50,步长为 5,每个属性在 0~9 中随机取值。共生成 45 个数据集,每个数据集最后一列作为决策属性。

由于人工数据集对象间没有必然联系,随对象数和属性数增加,分辨函数含有的子句数和子句中所含有的属性数大大增加。属性组合数的增加导致候选属性约简数上升,吴零点分解过程若无良好迭代控制策

略将产生较多冗余节点,导致求解效率下降。以对象数为 10 的数据集为例,加入迭代策略前后求解时间对比如图 1 所示。在有迭代策略下,动态序和静态序下产生的节点情况如图 2 所示。

从图 1 可知,迭代策略对求解效率产生巨大影响,吴零点分解过程加入迭代策略后,求解效率远优于无迭代策略情况,随属性数目增加,这种优势越明显。在无迭代策略情况下,对象数大于 30 的数据集将很难求出所有属性约简结果。由图 2 知,相比静态序,动态序下

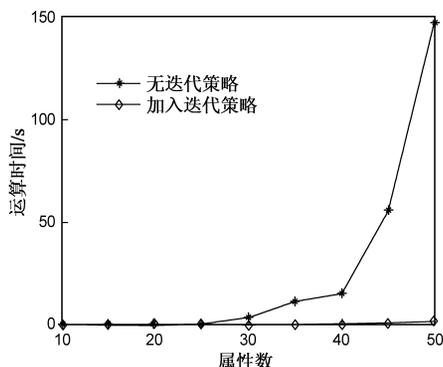


图1 对象数为10的数据集有无迭代策略求解时间对比(静态序)

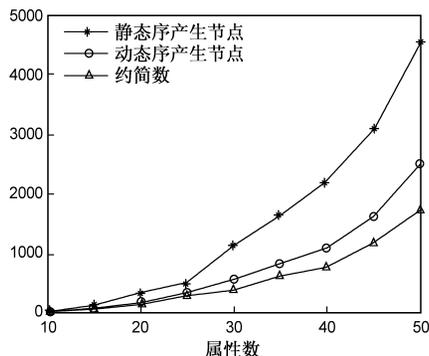


图2 对象数为10的数据集吴零点分解产生的节点数和属性约简数对比

产生的节点数更接近实际属性约简数. 随属性数目增加, 静态序产生的节点数偏离属性约简数越大, 求解效率下降越快. 虽然动态序下产生的节点数随属性数目增加偏离属性约简数也增加, 但偏离的程度远小于静态序, 从而动态序的求解效率要远好于静态序.

Skowron 提出的分辨函数法是迄今常采用的一种非穷举型计算决策表全部属性约简的方法. 图3以对象数为40的数据集为例, 给出了采用吴方法和直接采用吸收律和乘法律转化的方法求解所有属性约简的时间对比. 随属性数目增加, 基于吴方法的求解时间要优于Skowron直接转化方法, 并且属性数目越多, 优势越明

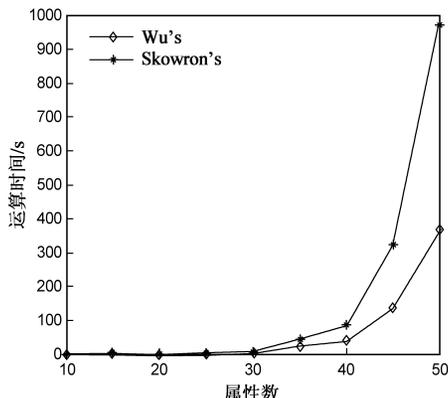


图3 对象数为40的数据集采用吴方法和Skowron直接转化方法求解时间对比

显. 对于对象数和属性数均为50的数据集, 采用Skowron直接转化方法需要耗费较长时间才能求得所有属性约简.

5 结论

本文基于吴方法, 从代数方程组角度给出了求解所有属性约简的一种新思路, 该新思路不局限于相容决策表, 对于信息系统(不含决策属性)和不相容决策表, 在合理构建分辨矩阵、求得分辨函数后, 均可采用该思路求解所有属性约简. 基于代数方程组观点的属性约简新思路, 为发掘决策表中的各种信息提供了有力支持, 并对数据挖掘的发展具有重要价值.

参考文献:

- [1] Pawlak Z, Skowron A. Rudiments of rough sets[J]. Information Sciences, 2007, 177(1): 3 - 27.
- [2] Wong S K M, Ziarko W. On optimal decision rules in decision tables[J]. Bulletin of Polish Academy of Sciences, 1985, 33(11 - 12): 693 - 696.
- [3] 杨明. 决策表中基于条件信息熵的近似约简[J]. 电子学报, 2007, 35(11): 2156 - 2160.
Yang Ming. Approximate reduction based on conditional information entropy in decision tables[J]. Acta Electronica Sinica, 2007, 35(11): 2156 - 2160. (in Chinese)
- [4] 苗夺谦, 胡桂荣. 知识约简的一种启发式算法[J]. 计算机研究与发展, 1999, 36(6): 681 - 684.
Miao Duoqian, Hu Guirong. A heuristic algorithm for reduction of knowledge[J]. Journal of Computer Research & Development, 1999, 36(6): 681 - 684. (in Chinese)
- [5] 胡峰, 王国胤. 属性序下的快速约简算法[J]. 计算机学报, 2007, 30(8): 1429 - 1435.
Hu Feng, Wang Guangyin. Quick reduction algorithm based on attribute order[J]. Chinese Journal of Computers, 2007, 30(8): 1429 - 1435. (in Chinese)
- [6] Wang J, Miao D Q. Analysis on attribute reduction strategies of rough set[J]. Journal of Computer Science and Technology, 1998, 13(2): 189 - 193.
- [7] Skowron A, Rauszer C. The discernibility matrices and functions in information systems[A]. Intelligent Decision Support Handbook of Applications and Advances of the Rough Sets Theory[C]. Dordrecht: Kluwer Academic Publishers, 1991. 331 - 362.
- [8] 吴文俊. 几何定理机器证明的基本原理[M]. 北京: 科学出版社, 1984.
Wu Wenjun. Basic Principles of Mechanical Theorem Proving in Geometries[M]. Beijing: Science Press, 1984. (in Chinese)
- [9] 贺思敏, 张钊. 用吴方法求解可满足性问题(I)—算法变换[J]. 计算机学报, 1998, 21(8): 79 - 85.

He Simin, Zhang Bo. Solving satisfiability problem by Wu's method(I)-Algorithm transform[J]. Chinese Journal of Computers, 1998, 21(8): 79-85. (in Chinese)

- [10] Komorowski J, et al. Rough sets: A tutorial[A]. Rough Fuzzy Hybridization: A New Trend in Decision Making[C]. Singapore: Springer, 1999. 3-98.

作者简介:



苗夺谦 男, 1964年生, 教授, 博士生导师. 中国科学院自动化研究所模式识别与智能系统专业博士毕业. 研究方向为粗糙集理论、粒计算、Web智能、模式识别等.
E-mail: miaoduoqian@163.com



周杰 男, 1982年生, 同济大学电子与信息工程学院博士研究生. 研究方向为粗糙集理论与应用、数据挖掘等.
E-mail: jie_jpu@163.com

张楠 男, 1979年生, 同济大学电子与信息工程学院博士研究生. 研究方向为机器学习、粗糙集理论与应用等.

冯琴荣 女, 1972年生, 同济大学电子与信息工程学院博士研究生. 研究方向为多维数据模型、数据挖掘等.

王睿智 女, 1968年生, 同济大学电子与信息工程学院博士、讲师. 研究方向为生物信息学、模式识别等.