

蚁群优化属性约简算法

任志刚^{1,2}, 冯祖仁^{1,2}, 柯良军^{1,2}

(1. 西安交通大学系统工程研究所, 710049, 西安;

2. 西安交通大学机械制造系统工程国家重点实验室, 710049, 西安)

摘要: 为了获得决策表属性的最小约简, 将信息论角度定义的属性重要性作为启发信息引入蚁群算法, 提出了一种蚁群优化属性约简算法. 该算法将属性核直接引入到蚂蚁构造的每一个解中, 降低了问题规模, 新定义的状态转移规则和信息素更新规则体现了约简中属性间的无序性特点, 有利于在优解邻域内搜索. 通过9个典型实例对算法进行了验证, 结果与现有算法相比能够更容易找到最小约简, 所需时间较短.

关键词: 蚁群优化; 决策表; 属性约简

中图分类号: TP18 **文献标志码:** A **文章编号:** 0253-987X(2008)04-0440-05

Ant Colony Optimization Approach to Attribute Reduction Problem

REN Zhigang^{1,2}, FENG Zuren^{1,2}, KE Liangjun^{1,2}

(1. Systems Engineering Institute of Xi'an Jiaotong University, Xi'an 710049, China; 2. State Key Laboratory for Manufacturing Systems Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: In order to obtain the minimal reduction of decision table attributes, an attribute reduction algorithm is proposed based on ant colony optimization. The significance of attributes defined from the viewpoint of information theory is used as the heuristic information. The algorithm directly imports the core into each solution constructed by ants and reduces the problem scale. The new state transition rule and pheromone updating rule reflects the orderless characteristic among attributes, and benefits the search in the neighborhood of good solutions. The algorithm is verified on nine typical instances. Experimental results show that, compared with the existing algorithms, the proposed algorithm can find the minimal reduction more easily with less time.

Keywords: ant colony optimization; decision table; attribute reduction

属性约简是规则提取的重要前提, 它要求在保持知识库的分类与决策能力不变的条件下, 删除冗余或者不相关的属性. 作为一种处理不精确、不确定和不完整知识的数学工具, 粗糙集理论^[1-2]为属性约简提供了理论基础. 一个决策表可能存在多个约简, 人们期望找到具有最少属性的约简, 但这已被证明是一个 NP-hard 问题^[3].

蚁群优化^[4-6]是近年发展起来的一种通用的启发式算法, 已经成功应用于旅行商、二次指派、工件调度等许多复杂的组合优化问题. 文献[7-8]将蚁群

优化应用于属性约简问题, 对于状态转移规则和信息素更新规则的设置, 沿用了解决旅行商问题的思路, 没有充分利用属性约简问题的特点, 算法收敛速度慢, 甚至得不到最小约简. 与旅行商这类排序问题不同, 属性约简问题的每一个解都是条件属性集的一个子集, 属性间的顺序是无关紧要的. 此外, 属性核是每个约简中不可缺少的部分, 可以通过确定性的方法求得^[9-10], 而文献[7-8]均没有使用核信息.

本文提出的蚁群优化属性约简算法, 采用了新的状态转移规则和信息素更新规则, 并将属性核直

接引入到蚂蚁构造的每一个解中,一方面避免了核中属性与其他属性的组合,缩小了问题规模,另一方面减弱了蚂蚁初始搜索的盲目性,有利于快速找到最小约简。

1 粗糙集相关概念

定义 1^[1] 决策表 S 可以用 1 个五元组表示: $S = (U, C, D, V, f)$, 其中 U 为论域, 是非空有限对象集, C 表示非空有限条件属性集, D 表示非空有限决策属性集, 且 $C \cap D = \emptyset$, 取 $A = C \cup D, V = \bigcup V_a, a \in A, V_a$ 是属性 a 的值域, $f = U \times A \rightarrow V$ 是信息函数, 它指定 U 中每一个对象的属性值, 对 $\forall x \in U, \forall a \in A, \exists f(x, a) \in V_a$.

定义 2^[2] 在决策表 S 中, 对于 $\forall P \subseteq A, \forall Q \subseteq A, Q$ 对 P 的依赖度定义为

$$\gamma_P(Q) = |U_{POS}(P, Q)| / |U| \quad (1)$$

式中: $U_{POS}(P, Q)$ 表示 Q 的 P 正区域; $|\cdot|$ 表示集合的基数。

定义 3^[2] 在决策表 S 中, $\forall R \subseteq C$ 称为 C 相对于 D 的约简, 当且仅当满足条件

$$\gamma_R(D) = \gamma_C(D) \quad (2)$$

$$\gamma_{R-\{r\}}(D) < \gamma_C(D) \quad \forall r \in R \quad (3)$$

式中: $\gamma_R(D), \gamma_C(D), \gamma_{R-\{r\}}$ 分别表示属性 D 对 $R, C, R - \{r\}$ 的领带度。 C 的约简可能不惟一, 具有最少属性的约简称为最小约简, 记为 R_{duct} , 所有约简的交称为 C 的核, 记为 C_{ore} 。

定义 4^[11] 在决策表 S 中, 对 $\forall R \subseteq C, \forall a \in C - R$, 若 a 加入 R , 它对于决策属性 D 的重要性为

$$S_{GF}(a, R, D) = H(D | R) - H(D | R \cup \{a\}) \quad (4)$$

式(4)的重要性是从信息论角度定义的, 其中 $H(D | R)$ 表示决策属性 D 相对于 R 的条件熵。在 R 已知的情况下, $S_{GF}(a, R, D)$ 越大, 属性 a 对于决策属性 D 越重要。

2 蚁群优化属性约简算法

2.1 问题描述

求解最小约简 R_{duct} 可以表示为下述优化问题

$$\begin{aligned} R_{duct} &= \arg \min_{R \subseteq C} |R| \\ \gamma_R(D) &= \gamma_C(D) \end{aligned} \quad (5)$$

属性核是每个约简不可缺少的部分。目前有多种求解属性核算法, 如基于正区域的算法^[9]和基于差别矩阵的算法^[10]。本文采用文献[9]中提出的算法求解属性核, 它采用基数排序的思想快速计算正

区域 $U_{POS}(C, D)$ 以及所有的 $U_{POS}(C - \{i\}, D), \forall i \in C$, 然后得到 $U_{POS}(C, D)$ 和每个 $U_{POS}(C - \{i\}, D)$ 的基数之差, 从而得到属性核 C_{ore} , 其计算复杂度是 $O(|C||U|)$ 。记 $B = C - C_{ore}$, 求解最小约简 R_{duct} 可以进一步表示为

$$\begin{aligned} R'_{duct} &= \arg \min_{R \subseteq B} |R| \\ \gamma_{C_{ore} \cup R}(D) &= \gamma_C(D) \\ R_{duct} &= C_{ore} \cup R'_{duct} \end{aligned} \quad (6)$$

求解最小约简只需要在属性集 B 中找出满足式(6)中约束条件的最小属性子集。

2.2 主要过程

蚁群优化是一种并行构造型迭代式搜索算法, 它使用信息素表示蚂蚁在搜索中积累的经验信息。初始化信息素和算法参数后, 蚂蚁开始构造可行解。在每次迭代中, 每一只蚂蚁从初始节点出发, 在信息素和启发信息的引导下, 按照状态转移规则选择一个节点, 直到获得一个可行解。当所有蚂蚁构造解结束后, 算法根据信息素更新规则对优解进行信息素奖励, 引导蚂蚁在后续迭代中搜索优解附近区域, 希望在预设的迭代次数内找到最优解。对于属性约简问题, 将属性核看作一个属性(若属性核不为空), 并将每个条件属性看作一个节点, 记 R_k 为到第 k 步为止选择的属性节点集。可行解的构造过程如下: 蚂蚁选择属性核作为初始点, 并从属性集 B 中随机选择一个属性作为第 2 个节点, 根据状态转移规则不断地从候选节点集中选择下一个节点并添加到 R_k 中, 直到满足条件 $\gamma_{R_k}(D) = \gamma_C(D)$ 。

2.3 信息素定义

决策表中, 单个属性对论域的划分比较粗糙, 分类能力较弱, 多个属性对论域的划分较为细致, 整体的分类能力加强, 这体现了属性间的协作关系。属性约简就是要找出一组协作最好、分类能力最强的最小属性集合。在任意两属性节点间虚拟一弧段并将信息素关联在弧段上, 信息素的大小记录两属性协作关系的好坏, 表示期望两属性出现在同一个约简中的程度。记节点 i 与 j 间的信息素为 $\tau_{ij}, i \neq j$ 。按照该定义, τ_{ji} 与 τ_{ij} 的意义是等同的, 因此在算法运行过程中始终保持 $\tau_{ij} = \tau_{ji}$ 。初始时刻, 蚂蚁不具备任何经验信息, 认为各属性间的协作关系是相同的, 因此设置 $\tau_{ij}(0) = \tau_0$ (τ_0 是正常数)。

2.4 状态转移规则

状态转移规则决定了可行解的构造方式, 直接影响蚁群优化算法的性能。对于属性约简问题, 在可行解的构造过程中, 蚂蚁对属性的选择是有顺序的,

然而对于已选定的属性子集,属性间的顺序是没有意义的,希望能够从候选节点集中选出最少个数的属性与其配合,构成一个约简,因此从候选节点集中选择哪一个属性作为下一个属性节点,取决于所有已选属性.文献[7]采用的状态转移规则只考虑了当前属性对选择下一属性的影响,本文定义新的状态转移规则

$$P(r_{k+1} = j | R_k) = \begin{cases} \frac{\sum_{i \in R_k} \tau_{ij}^\alpha \eta_{ij}^\beta}{\sum_{m \in J_k} \sum_{i \in R_k} \tau_{im}^\alpha \eta_{im}^\beta} & j \in J_k \\ 0 & j \notin J_k \end{cases} \quad (7)$$

式中: r_{k+1} 表示第 $k+1$ 步将要选择的属性节点; $J_k = C - R_k$ 表示候选节点集; η_{ij} 表示节点 i 与 j 间的启发信息; $\alpha(\alpha \geq 0)$ 和 $\beta(\beta \geq 0)$ 分别表示信息素和启发信息的重要程度.信息论角度定义的属性重要性,考虑了属性对论域中不确定分类子集的影响^[11],对属性重要性具有较强的区分能力,文中选择它作为启发信息,即选择 $\eta_{ij} = S_{GF}(j, \{i\}, D)$. $\tau_{ij}^\alpha \eta_{ij}^\beta$ 反映了已选属性 i 对候选属性 j 的吸引强度, $\sum_{i \in R_k} \tau_{ij}^\alpha \eta_{ij}^\beta$ 反映了已选属性子集 R_k 对属性 j 的吸引强度, $P(r_{k+1} = j | R_k)$ 表示在已选属性子集为 R_k 的前提下,选择 j 为下一属性节点的概率.根据该规则,蚂蚁倾向于选择那些与所有已选属性协作良好的属性.

2.5 信息素更新规则

每次迭代中,所有蚂蚁都完成解的构造后,对信息素进行更新.一方面衰减所有信息素,另一方面加强优解对应的信息素浓度,使得在下一轮迭代中,蚂蚁倾向于在优解的邻域内搜索.为充分利用优解,只加强到当前迭代为止获得的最优解(记为 R_{gb})对应的信息素浓度.文献[7]按照属性被蚂蚁选择的顺序,只增强相邻属性节点(r_k, r_{k+1})间的信息素浓度(称为规则A).根据该规则,蚂蚁构造新解时,若当前属性节点是 r_k ,那么蚂蚁选择 r_{k+1} 作为下一个节点的概率增大,而其他属于 R_{gb} 但与 r_k 不相邻的属性节点被选择的概率减小.实际上,作为一种构造式启发算法,虽然蚁群优化在构造解的过程中,顺序地选择属性节点,但是对于属性约简问题本身,每一个约简都是条件属性集的一个子集,属性节点被选择的顺序是没有意义的.考虑到该特点,本文增强 R_{gb} 中任意两属性节点(r_k, r_m)间的信息素浓度(称为规则B).根据该规则,若蚂蚁构造的新解中已经包含 R_{gb} 中的部分属性节点,那么 R_{gb} 中其他属性节点被选择到的概率将增大,有利于蚂蚁在优解邻域内搜

索.图1形象地描述了这2种更新方式的差异.

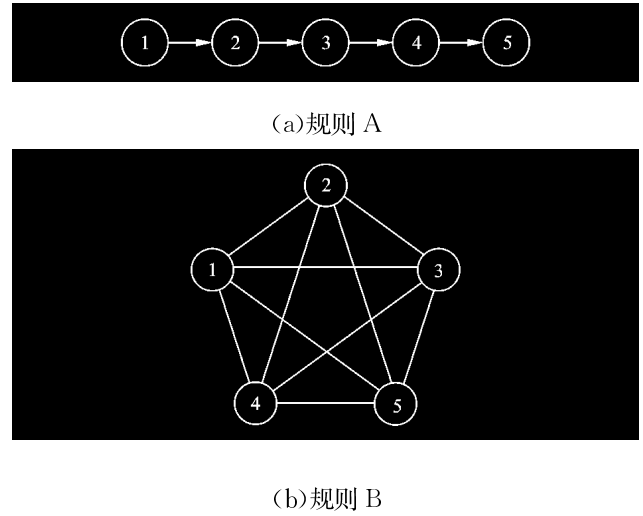


图1 2种更新方式的比较示意图

为防止停滞现象,采用 MAX-MIN Ant System^[6]的思想,将信息素限制在一定范围内,使得信息素间的差异不是太大,加强蚂蚁探索新解的能力.信息素更新规则为

$$\tau_{ij}(t+1) = \rho \tau_{ij}(t) + \Delta \tau_{ij}(t) \quad (8)$$

$$\Delta \tau_{ij}(t) = \begin{cases} \frac{1}{|R_{gb}|} & i, j \in R_{gb}, i \neq j \\ 0 & \text{其他} \end{cases} \quad (9)$$

$$\left. \begin{aligned} \tau_{ij}(t+1) &> \tau_{\max} & \tau_{ij}(t+1) &= \tau_{\max} \\ \tau_{ij}(t+1) &< \tau_{\min} & \tau_{ij}(t+1) &= \tau_{\min} \end{aligned} \right\} \quad (10)$$

式中: $\rho(0 < \rho < 1)$ 表示信息素存留系数; $\Delta \tau_{ij}(t)$ 表示第 t 次迭代中属性节点 i 与 j 间信息素增量,它反比于 R_{gb} 中包含的属性节点数; $\tau_{\max}$ 和 $\tau_{\min}$ 分别为信息素的上下界, $\tau_{\max}$ 取信息素最大渐进值的估计值^[6],即 $\tau_{\max} = 1/(1-\rho) |R_{gb}|$,取 $\tau_{\min} = \epsilon \tau_{\max}$ (ϵ 差异系数, $0 < \epsilon < 1$).

在信息素衰减阶段,规则A和规则B的计算复杂度都是 $O(|C|^2)$.在信息素增强阶段,规则B的计算复杂度大于规则A,分别是 $O(|R_{gb}|^2)$ 和 $O(|R_{gb}|)$.因为 $|R_{gb}| < |C|$,所以对于整个信息素更新过程,规则A和规则B的计算复杂度相同,都是 $O(|C|^2)$.

3 算法描述

(1) 计算决策属性 D 对条件属性 C 的依赖度 $\gamma_C(D)$ 和属性核 C_{ore} .

(2) 计算 $\gamma_{C_{ore}}(D)$,若 $\gamma_{C_{ore}}(D) = \gamma_C(D)$,取 $R_{duct} = C_{ore}$,终止算法,否则下转步骤(3).

(3) 初始化参数和信息素,计算启发信息,指定

迭代次数 $T_{\max}$ 并置变量 $t=0$.

(4) 在每次迭代中,每只蚂蚁进行以下操作:① 选择属性核 C_{core} 作为初始节点,从属性集 B 中随机选择一个属性作为第 2 个节点;② 判断当前部分解是否为可行解,若可行,则下转步骤④,否则下转步骤③;③ 根据状态转移规则(式(7))选择下一节点,返回步骤②;④ 每只蚂蚁构造可行解结束后,保存最优解 R_{gb} ,修改 $\tau_{\max}$ 和 $\tau_{\min}$,根据信息素更新规则(式(8)~式(10))修改信息素.

(5) 置 $t=t+1$,若 $t < T_{\max}$,则返回步骤(4),否则取 $R_{\text{duct}}=R_{\text{gb}}$,终止算法.

4 实验与分析

为考察算法的有效性,用 VC++6.0 实现了文献[7]提出的算法和本文算法,运行环境为 P4 2.0 CPU,512M RAM, Windows XP 操作系统.为保证比较的公平性,选择文献[7]采用的部分决策表对算法进行测试,决策表选自美国加利福尼亚大学提供的机器学习数据库(UCD)^[12], $\alpha, \beta, N_{\text{ant}}, T_{\max}$ 与文献[7]取相同的值,依次为 1.0、0.1、|C|、250,文献[7]中没有明确给出和没有涉及到的参数取值为 $\rho=0.95, \varepsilon=0.004, \tau_0=1/(1-\rho)$. 实验中,对每一个决策表测试 20 次,记录在 $T_{\max}$ 迭代次数内获得的约简情况以及获得最小约简所需要的平均迭代次数和运行时间,结果如表 1 所示.

表 1 中前 5 个决策表的核均为空,对它们约简时,本文算法和文献[7]算法的主要差别在于状态转

移规则和信息素更新规则不同.对于大部分决策表,本文算法都能在较少迭代次数内获得最小约简,运行时间也短于文献[7]算法.对于决策表 WQ,虽不能在每次测试中都获得最小约简,但解的质量仍优于文[7]算法所得结果.由此说明,文中采用的状态转移规则和信息素更新规则是有效的.表 1 中后 4 个决策表的核不为空,本文算法将属性核直接引入到每一个可行解中,蚂蚁只需要在非核属性集中搜索,与文献[7]算法相比,约简结果是理想的.图 2 是 2 种算法对决策表 Mushroom 约简的进化曲线,其中纵坐标表示 20 次测试中,到当前迭代为止获得的最优解中包含的属性个数平均值,对比图 2 曲线可以看出,本文算法的性能优于文献[7]算法.

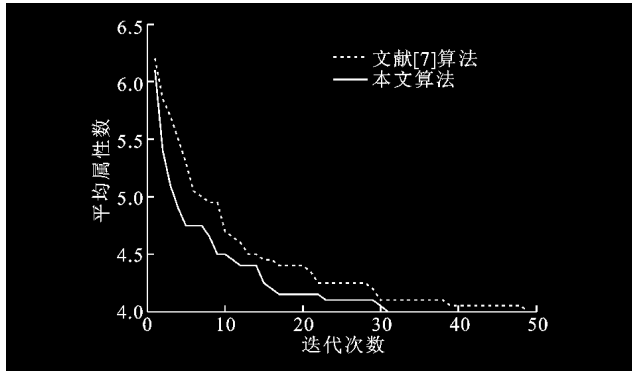


图 2 2 种算法对 Mushroom 约简的进化曲线

5 结束语

本文将蚁群优化应用于决策表属性约简问题，

表 1 约简结果与比较

决策表	实例数	条件属性个数	核属性个数	最小约简属性个数	约简后平均条件属性数		获得最小约简平均迭代次数		获得最小约简平均运行时间/s	
					文献[7]算法	本文算法	文献[7]算法	本文算法	文献[7]算法	本文算法
Letters	26	25	0	8	8.00	8.00	48.00	10.35	0.060 1	0.013 3
Lung	32	56	0	4	4.00	4.00	5.25	4.00	0.010 9	0.007 8
Derm	366	34	0	6	6.15	6.00		32.05		0.316 1
WQ	521	38	0	12	13.45	12.30				
Mushroom	8 124	22	0	4	4.00	4.00	18.50	11.05	0.800 2	0.566 9
Heart	294	13	3	6	6.10	6.00		2.00		0.004 7
Vote	300	16	7	8	8.00	8.00	70.45	1.00	0.164 0	0.003 1
Credit	1 000	20	1	8	8.60	8.00		35.90		0.583 5
LED	2 000	24	2	5	5.50	5.00		10.35		0.321 5

所采用的状态转移规则和信息素更新规则,体现了约简中属性间的无序性这一特点,通过将属性核直接引入到每一个可行解中,缩小了问题规模,有利于获得最小约简.在多个决策表上的测试表明,本文设计的算法是有效的.

参考文献:

- [1] PAWLAK Z. Rough sets[J]. International Journal of Computer and Information Science, 1982, 11(5):341-356.
- [2] PAWLAK Z. Rough set approach to multi-attribute decision analysis [J]. European Journal of Operational Research, 1994, 12(72):443-459.
- [3] WONG S K M, ZIARKO W. On optional decision rules in decision tables [J]. Bulletin of Polish Academy of Sciences, 1985, 33(11):693-696.
- [4] DORIGO M, MANIEZZO V, COLORNI A. Ant system: optimization by a colony of cooperating agents [J]. IEEE Trans on Systems, Man and Cybernetics: Part B, 1996, 26(1):29-41.
- [5] DORIGO M, GAMBARDILLA L M. Ant colony system: A cooperative learning approach to the traveling salesman problem [J]. IEEE Trans on Evolutionary Computation, 1997, 1(1):53-66.
- [6] STUTZLE T, HOOS H H. Max-min ant system [J]. Future Generation Computer Systems, 2000, 16(8): 889-914.
- [7] JENSEN R, SHEN Q. Finding rough set reducts with ant colony optimization [C]//Proceeding of the 2003 UK Workshop on Computational Intelligence. Bristol, UK: University of Bristol, 2003:15-22.
- [8] 王俊峰,朱庆保.基于蚁群算法的知识约简 [J]. 南京师范大学学报(工程技术版), 2005, 5(2):50-53.
WANG Junfeng, ZHU Qingbao. A knowledge reduction method based on ant colony algorithm [J]. Journal of Nanjing Normal University (Engineering and Technology), 2005, 5(2):50-53.
- [9] 周江卫,冯博琴,刘洋.一种新的快速求核算法 [J]. 西安交通大学学报, 2007, 41(6):688-691.
ZHOU Jiangwei, FENG Boqin, LIU Yang. New algorithm for quick computing core [J]. Journal of Xi'an Jiaotong University, 2007, 41(6):688-691.
- [10] 杨明.一种基于改进差别矩阵的核增量式更新算法 [J]. 计算机学报, 2006, 29(3): 57-63.
YANG Ming. An incremental updating algorithm of the computation of a core based on the improved discernibility matrix [J]. Chinese Journal of Computers, 2006, 29(3):57-63.
- [11] 王国胤,于洪,杨大春.基于条件信息熵的决策表约简 [J]. 计算机学报, 2002, 25(7): 759-766.
WANG Guoyin, YU Hong, YANG Dachun. Decision table reduction based on conditional information entropy [J]. Chinese Journal of Computers, 2002, 25(7): 759-766.
- [12] ASUNCION A, NEWMAN D J. UCI machine learning repository [ED/OL]. (2006-10-28) [2007-04-01]. <http://mllearn.ics.uci.edu/MLSummary.html>.

(编辑 刘杨)

2007 年我校获准国家自然科学基金资助 153 项

2007 年我校获准国家自然科学基金资助项目共计 153 项,资助经费 4 876 万元.其中,面上项目 144 项、重点项目 5 项(含 1 项重大国际合作交流项目)、国家杰出青年基金项目 3 项、海外青年合作研究基金 1 项,项目数位居全国高校排名第 11 位.值得一提的是,国家杰出青年基金在时隔 10 年后,又取得新进展,医学学科实现了零的突破.获得国家杰出青年基金资助的分别是材料学院李长久教授、能动学院席光教授和医学院郑鹏生教授.电信学院任巍教授作为合作人获得海外青年合作研究基金.获得重点项目资助的分别有电信学院管晓宏教授和高静怀教授、能动学院何雅玲教授、医学院贺浪冲教授.材料学院孙军教授获得重大国际合作交流项目.

我校科技处积极组织各类项目的申报,2007 年组织申报国家自然科学基金各类项目 879 项(其中,面上项目 830 项,重点项目 23 项,国家杰出青年基金 18 项,海外港澳合作研究基金 8 项),最终 153 项获准资助,面上项目资助率 17.2%.

(科技处项目部)