

基于蚁群算法的知识约简

王俊峰, 朱庆保

(南京师范大学 数学与计算机科学学院, 江苏 南京 210097)

[摘要] Rough集理论中知识约简是个 NP-hard 问题, 目前已提出较多的求解方法, 但是每种方法由于其自身的局限性, 只适用于一定条件下的求解. 蚁群算法是较新的仿生优化算法, 在解决各类组合优化问题中都取得了很好的效果. 其显著优点是受问题规模的影响不大, 对大规模问题的求解仍能发挥较优的性能. 受蚁群算法该特性的启发, 提出基于蚁群算法的知识约简方法. 文中具体描述了将条件集的组合方式用一图结构来表示、构建目标评价函数、算法参数的设定以及算法的具体实施步骤等. 最后通过于相关文献的比较实验, 验证了该方法的有效性.

[关键词] 蚁群算法, Rough集, 知识约简

[中图分类号] TP301.6 [文献标识码] A [文章编号] 1672-1292(2005) 02-0050-04

A Knowledge Reduction Method based on Ant Colony Algorithm

WANG Junfeng ZHU Qingbao

(School of Mathematics and Computer Science, Nanjing Normal University, Jiangsu Nanjing 210097, China)

Abstract Knowledge Reduction is a kind of NP-hard problem in Rough Sets theory, for which there have already been some methods, but each method has its own limitations and is thus suitable only in special condition of the problem. As a new bionics optimization algorithm, Ant Colony Algorithm has good effects in solving many kinds of combinations optimization problem regardless of the scale of the problem. Inspired by this character of Ant Colony Algorithm, we present a new approach for knowledge reduction based on the Ant Colony Algorithm in this paper. The paper describes in detail the construction of the graph expressing the combination of the condition sets, the evaluation function, the setting of the parameters and the steps of the new approach, and at last proves the validity of this new method through the comparative experiment of the relevant literatures.

Key words ant colony algorithm, rough sets, knowledge reduction

0 引言

Rough集理论是 Pawlak Z 等学者提出的研究不完整数据及不精确知识的表达、学习、归纳的一套方法^[1], 它无需任何先验信息, 就能以观察和测量数据进行分类的能力为基础, 通过对数据进行分析、近似分类以及推理数据间的关系, 从中发现隐含的知识, 揭示潜在的规律. 知识约简是粗集理论的核心内容之一. 知识约简往往不唯一, 同一个知识表达系统可能存在多个不同的约简. 它属于 NP-hard 问题, 除了穷举搜索外, 很难保证得到最优解^[2].

已经存在的求解方法有: 用幂子集的方法求解^[3], 但它只适用于属性较小的系统; 采用信息论的方法^[4], 但该算法易陷入局部最优, 往往得不到全局最优解; 也有用仿生智能算法如遗传算法解决该问题的^[2], 取得了较好的效果, 但也有不足, 如易陷入局部最优解等.

近年来由意大利学者 Dorigo 等提出来的蚁群系统 (Ant System) 与蚁群算法 (Ant Colony Algorithm), 是一种通用的启发式算法^[5 6]. 它主要模拟自然界蚂蚁觅食时, 通过在路径上信息素的传递, 最终能发现一条最短路径的过程. 它的主要特征是采用正反馈机制、强的鲁棒性和适于并行处理, 已

收稿日期: 2005-01-03
作者简介: 王俊峰 (1977-), 硕士研究生, 主要从事智能控制的学习与研究. E-mail: ewj@163.com
通讯联系人: 朱庆保 (1955-), 教授, 主要从事人工智能与控制等方面的教学与研究. E-mail: zhuqingbao@njnu.edu.cn

经在图着色问题、大规模集成电路设计、网络路由选择、规划设计等应用中表现出很好的性能. 受蚁群算法的启发, 本文提出将条件属性集映射到一有向图, 从而提出了用于知识约简的蚁群算法. 实验表明, 本算法不仅能得到全局最优解, 而且还可得到一些次优解, 可以更好地反应出知识表达系统的特性.

1 有向图的构造与目标函数的确立

对于决策表 $T = (U, A, C, D)$, 条件属性集 C 对论域 U 形成一个划分 U/C , 决策属性集 D 对论域形成另一个划分 U/D . 这两个划分形成了条件属性和决策属性在对论域样本分类上的知识. 属性约简的目的就是从条件属性中发现部分必要的条件属性 C' , 使得 C' 相对于 D 的分类能力与 C 相对于 D 的能力相同. 所谓最小约简就是从 C 中找出包含必要属性数最少的 C' . 本质上来说它是个组合数学的最小覆盖问题. 受蚁群算法启发, 如果能够将条件属性集映射到一图结构, 那么就可以用蚁群算法来解决该问题.

1.1 节点和路径的生成

用二进制 $\{0, 1\}$ 给条件属性集编码, 每个条件属性对应一位二进制数: 0 表示该位对应的条件属性不属于最小约简集; 1 表示该位对应的条件属性属于最小约简集. 假设有 N 个条件属性, 用一有向图来描述: $G = (C, L)$.

节点集 C 为: $\{c_0(v_s), c_1(v_1^0), c_2(v_1^1), c_3(v_2^0), c_4(v_2^1), \dots, c_N(v_N^0), c_{N+1}(v_N^1), c_{N+2}(v_e)\}$, 其中 $c_0(v_s)$ 为起始结点, $c_{N+2}(v_e)$ 为结束结点 ($c_0(v_s)$ 和 $c_{N+2}(v_e)$ 不与任何条件属性对应), $V_i^0, V_i^1 (i = 1, 2, \dots, N)$ 表示 i 属性的状态: 0 1

有向边集为: $\{(v_s, v_1^0), (v_s, v_1^1), (v_1^0, v_2^0), (v_1^1, v_2^0), (v_1^0, v_2^1), (v_1^1, v_2^1), (v_2^0, v_3^0), (v_2^0, v_3^1), (v_2^1, v_3^0), (v_2^1, v_3^1), \dots, (v_N, v_e), (v_N, v_e)\}$ 如图 1 所示.

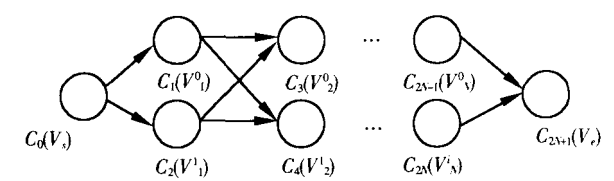


图 1 属性图

1.2 目标函数的建立

根据最小约简的要求, 约简集的性能主要取决于两个方面: 所含条件属性的个数 n 和决策属性对其依赖度 k . 对某一属性子集, 其含属性个数越少, 决策属性对其依赖度越大, 则最有可能成为最小约

简. 可以认为它是一个多目标优化问题, 而多目标优化大多具有多个 PARETO 最优解.

为了将多目标函数转化为单目标优化问题, 构造适应度函数如下:

$$F = (N - n) / N + k + r \quad (1)$$

式中, N 为条件属性的个数; n 为个体的条件属性个数; $k (0 \leq k \leq 1)$ 为决策属性对该个体的依赖度^[1], $k = \text{card}(\text{POS}_C(D)) / \text{card}(U)$, $\text{POS}_C(D)$ 是 D 的 C -正区域; 为了拉开评价的差距用参数 r 作为奖励数: 当 $k = 1$ 时, $r = 0.5$ 当 $k < 1$ 时, $r = 0$

2 属性约简的蚁群算法

2.1 问题描述与定义

记 AS 为二维平面上的凸多边形有限区域, 其内部分布着 $2n + 2$ 个节点, 令节点集 C 为: $\{c_0(v_s), c_1(v_1^0), c_2(v_1^1), c_3(v_2^0), c_4(v_2^1), \dots, c_N(v_N^0), c_{N+1}(v_N^1), c_{N+2}(v_e)\}$, 其中 $c_0(v_s)$ 为起点, $c_{N+2}(v_e)$ 为终点, 节点下标集 $R = \{0, 1, 2, \dots, 2n + 2\}$.

各节点在 AS 中的连线构成有向边集:

$$\{(v_s, v_1^0), (v_s, v_1^1), (v_1^0, v_2^0), (v_1^1, v_2^0), (v_1^0, v_2^1), (v_1^1, v_2^1), (v_2^0, v_3^0), (v_2^0, v_3^1), (v_2^1, v_3^0), (v_2^1, v_3^1), \dots, (v_N, v_e), (v_N, v_e)\}$$

任意两节点间连线为有向边, 记作 $e_{ij}, i, j \in R$.

求解目标: 得到一条从起点到终点的的最优路径.

定义 1 $\text{ant} = \{1, 2, \dots, k, \dots, m\}$ 表示所有蚂蚁的集合, $k \in \text{ant}$ 表示某只蚂蚁, $\tau_{ij}(t)$ 表示蚂蚁在 t 时刻残留在 $e_{ij} (i, j \in R)$ 上的信息量.

定义 2 设蚂蚁 k 在 t_0 时刻从起始节点 i 出发, 到达目标节点. 任意时刻所处的节点位置记作 $P(t_i)$. 令 $\text{tabu}_k = \{P(t_0), P(t_1), \dots, P(t_j)\}$ 为蚂蚁 k 从 t_0 时刻到 t_j 时刻已走节点位置的集合 (等价于已走节点集合), t_{j+1} 时刻, $\forall P(t_{j+1}) \in C$ 且 $\forall P(t_{j+1}) \in \text{tabu}_k$ 则称 $\forall P(t_{j+1})$ 为 t_{j+1} 时刻禁入点. 因此称 tabu_k 为禁忌表.

2.2 算法步骤

算法步骤归纳如下:

Step 1 初始化: 设定蚂蚁数 m (本文试验 $m = 5$). 令时间计数器 $t = 0$ 循环次数 $T = 0$ 设定最大循环次数 T_{\max} 以及初始时刻各节点上信息激素的浓度 $\tau_{ij}(0)$ 的值 ς , 令 $\Delta\tau_{ij} = 0$ 将全部蚂蚁置于起始点.

Step 2 置变量 $\text{Count} = 0$

Step 3 对 $\forall k (k \in \text{ant})$ 利用式 (2) 或 (3) 计

算它从当前位置到下一个节点的转移的概率; 根据这些概率, 采用赌轮选择 (Roulette Wheel Selection) 策略, 选择下一个节点 j

$$j = \begin{cases} \arg \max_{j \in \text{tabu}_k} \{ [\tau_{ij}(t)] [\eta_{ij}(t)]^\beta \} & \text{if } q \leq q_0 \\ S & \text{otherwise} \end{cases} \quad (2)$$

式中, $0 < q_0 \leq 1$, 是初始设定的参数, q 是一个随机数, $q \in (0, 1)$, S 是根据 (3) 式决定的随机变量.

$$p_{ij}^k = \begin{cases} \frac{[\tau_{ij}(t)]^\alpha [\eta_{ij}(t)]^\beta}{\sum_{k \in \text{tabu}_k} [\tau_{ik}(t)]^\alpha [\eta_{ik}(t)]^\beta} & j \notin \text{tabu}_k \\ 0 & j \in \text{tabu}_k \end{cases} \quad (3)$$

式中, $p_{ij}^k(t)$ 表示在 t 时刻蚂蚁 $k(k \in \text{ant})$ 由节点 i 转移到节点 j 的概率; α 为在边 e_{ij} 上残留信息的重要程度; β 为启发信息的重要程度; 当 $q > q_0$ 时, 用 (3) 式计算下个节点的转移概率 p_{ij}^k , 再用赌轮选择 (Roulette Wheel Selection) 策略选择 j , 否则根据 (2) 式选择 j , 将 j 加入禁忌表 tabu_k .

参数的设置对算法性能的影响比较大, 其设置还没有理论依据, 故按经验结果:

$1 \leq \alpha \leq 5, 1 \leq \beta \leq 5$
Step 4 置 $\text{Count} = \text{Count} + 1$, 若 $\text{Count} \leq N$ (N 为条件属性个数), 转到 Step 3 否则, 转到 Step 5

Step 5 根据蚂蚁 $k(k \in \text{ant})$ 所走过的路径利用式 (1) 计算该路径对应的目标函数 F_k ; 记录本轮循环中的最优路径 (它对应着本轮循环中的最优性能指标).

Step 6 令 $t = t + (N + 1); T = T + 1$, 根据式 (4)、(5) 更新每个节点上信息激素物质的浓度, 并将 $\text{tabu}_k(k \in \text{ant})$ 中的元素清零.

$$\tau_{ij}(t + N + 1) = \rho \tau_{ij}(t) + \Delta \tau_{ij} \quad (4)$$

$$\Delta \tau_{ij} = \sum_{k=1}^m \Delta \tau_{ij}^k \quad (5)$$

其中, $0.5 \leq \rho \leq 1, \Delta \tau_{ij}^k$ 为本轮循环中第 $k(k \in \text{ant})$ 只蚂蚁在路径 i 到 j 上留下的激素物质, 它按式 (6) 计算:

$$\Delta \tau_{ij}^k = \begin{cases} \frac{Q}{F_k^{-1}}, & \text{若第 } k \text{ 只蚂蚁在本轮循环中经过 } (i, j) \\ 0 & \text{否则} \end{cases} \quad (6)$$

式中, F_k 为第 k 只蚂蚁在本轮循环中的目标函数值, 由式 (1) 计算; Q 为正常数 ($1 \leq Q \leq 10000$).

Step 7 若 $T < T_{\max}$ 且整个蚁群尚未收敛到走同一条路径, 则再次将全部蚂蚁置于起始点 $C_0(V_s)$ 并转到 Step 2 若 $T < T_{\max}$ 但整个蚁群已收

敛到走同一条路径, 或 $T = T_{\max}$ 则循环结束, 输出最优路径及其对应的属性集.

3 实验结果与分析

本实验选择了两个决策表对比分析.

决策表如表 1^[4] 所示. 该决策表用于评价行驶的总里程与相关的指标之间的关系. 论域 $U = \{1, 2, 3, 4, \dots, 21\}$, 条件属性集 $C = \{\text{类型, 汽缸, 涡轮增压式, 燃料, 排气量, 压缩率, 功率, 换挡, 重量, 里程}\}$, 决策属性集 $D = \{\text{里程}\}$.

表 1 决策表 1

U	类型	汽缸	涡轮增压式	燃料	排气量	压缩率	功率	换挡	重量	里程
1	1	1	1	1	1	1	1	2	2	2
2	1	1	2	1	1	2	1	2	2	2
3	1	1	2	1	1	1	1	2	2	2
4	2	1	1	1	1	1	1	2	3	1
5	1	1	2	1	1	2	2	2	2	2
6	1	1	2	2	1	2	2	1	1	3
7	1	1	2	1	1	2	1	2	1	3
8	2	2	2	2	2	1	3	2	3	1
9	1	2	2	2	2	1	3	2	2	2
10	1	2	2	2	2	1	2	1	2	2
11	2	2	2	1	2	1	3	2	3	1
12	2	2	2	1	2	2	2	2	2	1
13	1	2	2	2	1	2	2	2	2	2
14	2	2	1	1	2	1	1	2	2	1
15	2	2	2	2	2	2	3	2	2	1
16	1	2	1	1	1	2	1	2	2	2
17	1	1	2	1	1	2	1	1	2	2
18	1	2	2	1	1	2	1	1	2	2
19	2	2	2	1	2	1	2	2	2	1
20	1	2	2	1	2	1	2	2	2	1
21	1	2	2	2	2	1	2	2	2	2

用一般的启发式约简算法^[4] 求出的最小约简为 $\{\text{类型, 燃料, 排气量, 重量}\}$, 用本文的算法除得到该解外还得到一个属性个数为 5 的次优解, 启发式约简算法^[4] 和本文算法结果比较如表 2 所示.

表 2 实验结果比较

约简的属性个数	启发式约简算法 ^[6]	本文算法
4	100110001	100110001
5	无	101110001

为进一步验证本算法有效性, 选择决策表 2^[7], 如表 3 所示进行实验. 用文献 [7] 和本文算法的实验结果对比如表 4 所示.

其中论域 $U = \{1, 2, 3, 4, 5, 6\}$, 条件属性集 $C = \{L1, M1, H1, L2, M2, H2, L3, M3, H3\}$, 决策属性集 $D = \{D\}$

表 3 决策表 2

<i>U</i>	<i>L1</i>	<i>M1</i>	<i>H1</i>	<i>L2</i>	<i>M2</i>	<i>H2</i>	<i>L3</i>	<i>M3</i>	<i>H3</i>	<i>D</i>
1	0	1	0	1	1	0	1	0	0	
2	0	0	1	1	0	0	1	1	0	
3	1	0	0	0	0	1	0	0	1	
4	1	0	0	1	0	0	0	1	0	
5	1	1	0	0	0	1	0	1	0	
6	1	1	1	1	0	0	1	0	0	

表 4 实验结果比较

约简的属性个数	文献 [11] 的算法	本文算法
3	110001000	110001000 110100000

比较以上实验结果, 可见本文提出的算法对于知识约简是有效的, 不仅能收敛到全局最优解, 而且也能得出次优解, 可以更好地反应出知识表达系统的特性, 为具体应用提供更多信息.

4 结语

知识约简是个多目标优化的 NP-hard问题. 本文受蚁群算法的启发, 通过将条件属性集映射到有向图, 提出了用蚁群算法来解决该问题, 并通过实验验证了该算法的有效性. 但对于该算法还存在需要研究和加以改进的地方, 如参数的设置方面目前主要靠实验为主, 缺少理论依据, 算法的收敛速

度还有待进一步改进等.

[参考文献]

[1] 刘清. Rough集及 Rough推理 [M]. 北京: 科学出版社, 2001. 40- 75

[2] 张可卿, 谢志鹏, 刘宗田. 基于变长编码遗传算法的最小缩减计算 [J]. 小型微机计算机系统, 2001, 22(9): 1055- 1057

[3] 吴福保, 李奇, 宋文忠. 基于粗集理论的一种归纳学习方法 [J]. 控制与决策, 1999 14(3): 206- 211.

[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): 206- 211

[5] DorigoM, Maniezzo V, ColomiA. Ant system: optimization by a colony of cooperating agents[J]. IEEE Transactions on Systems Man and Cybernetics 1996, 26(1): 29- 41.

[6] DorigoM, BonabeauE, TheraulazG. Ant algorithms and stigmergy [J]. Future Generation Computer Systems 2000, 16(2): 851- 871.

[7] Mohua B, SushmitaM, SankarK P. Rough fuzzyMLP. Knowledge encoding and classification[J]. IEEE Transactions on Neural Networks 1998, 9(6): 1203- 1216

[责任编辑: 刘健]