基于小生境粒子群的属性约简算法

吴永芬[1], 冯茂岩[2], 张 健[3]

(1. 解放军理工大学 指挥自动化学院, 江苏 南京 210007; 2. 江苏海事职业技术学院 信息工程系, 江苏 南京 210070;
3. 三江学院 计算机基础部, 江苏 南京 210012)

[摘要] 遗传算法(GA)及蚂蚁算法(ACO)等进化属性约简算法, 具有全局寻优的优点, 但存在算法时间复杂度高, 搜索空间大等不足; 粒子群(PSO)属性约简算法, 虽然可提高求解效率, 但易陷入局部最优. 本文引入小生境技术, 提出基于小生境粒子群的属性约简算法, 利用小生境技术造就种群的多样性, 使解保持多样化, 以此避免粒子群属性约简算法易早熟收敛的缺点. 理论分析及实验结果表明, 该算法是有效可行的.

[关键词] 小生境, 粒子群算法, 粗糙集, 属性约简

[中图分类号] TP 301.6 [文献标识码] A [文章编号] 1672-1292(2008)04-0132-04

Attribute Reduction Algorithm Based on the Niche PSO

Wu Yongfen[1], Feng Maoyan[2], Zhang Jian[3]

(1. Institute of Command Automation, PLA University of Science and Technology, Nanjing 210007, China;
2. Department of Information Engineering, Jiangsu Maritime Institute, Nanjing 210070, China;
3. Department of Computer Elementary Training, Sanjiang University, Nanjing 210012, China)

Abstract: Currently, there are many evolutionary algorithms available for attribute reduction as genetic algorithm (GA) and the ant colony optimization algorithm (ACO), but high time complexity and wide search space are their common disadvantages. The PSO attribute reduction algorithm ought to improve the efficiency, however, it has the premature convergence problem. Therefore, an attribute reduction algorithm based on the niche PSO is presented in this paper after introducing the niche technology. The niche technology is applied to maintain the diversity of the population and the solutions, and avoid the premature convergence problem of PSO algorithm. The theoretical analysis and experimental results indicate that the proposed niche PSO algorithm is feasible and effective.

Key words: niche; PSO; rough set; attribute reduction

属性约简是粗糙集的重要研究内容之一, 至今已取得很多成果. 但作为一个组合爆炸的 NP-hard问题, 至今仍未找到通用、高效的解决方法. 为提高求解效率, 传统粗糙集大多数都采用启发式方法求解属性约简[1,2]. 启发式算法采取的策略实际上是贪心策略, 每次都选择最重要的一个属性, 或者选择分类能力最大的一个属性加入约简集, 往往在较短的时间内只能得到一个次最佳解.

进化算法通过模仿生物学中生物"优胜劣汰"的进化过程如遗传算法, 或通过模拟自然界生物的群体行为如蚁群算法、粒子群算法, 对所研究问题进行自适应调整的随机搜索过程, 以求得最优解. 鉴于此, 引入进化算法有利于改善传统启发式求解方法. 近年来, 许多学者采用进化算法求解属性约简并取得了一些进展, 如文献[3]提出基于 GA的特征选择算法, 文献[4]提出基于蚁群优化(ACO)的属性约简算法, 可求得多于一个约简结果, 但存在算法时间复杂度高、搜索空间大等缺点; 文献[5]提出粒子群(PSO)属性约简算法, 其求解效率高, 但易陷入局部最优.

为进一步提高进化算法求解多峰值问题的能力, 在进化算法中引入小生境技术, 可使解保持多样化, 有效提高算法的全局搜索能力. 为此, 有学者提出了基于小生境的遗传属性约简算法[6], 虽可提高全局搜索能力, 但算法时间复杂度高. 粒子群算法具有计算简单快捷、求解效率高的优点. 本文引入小生境粒子群算法后, 提出基于小生境粒子群的属性约简算法, 利用小生境技术避免粒子群属性约简早熟收敛. 理论分

收稿日期: 2008-06-18
通讯联系人: 吴永芬, 硕士, 研究方向: 粗集理论与应用. E-mail: yfwu0916@126.com

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

析及实验结果表明, 文中的算法是有效可行的, 可在很短的时间内得到多个最优解.

1 粗糙集概念

定义 1[7] 一个知识表达系统是一个四元组 $S=\{U, V, R, f\}$, 其中 $U=\{x_1, x_2, \cdots, x_n\}$ 是有限的样本集合, 称为论域; R是属性集合; $V=\bigcup_{a\in(C\cup D)} V_a$, V_a 为属性 a的值域集; $f: U\times R\to V$ 指定了系统中每个对象在各个属性上的取值.

定义 2[7] 决策表 $S=\{U, V, R=C\cup D, f\}$, 属性集合 $B\subseteq C\cup D$, B在论域 U上定义不可分辨二元关系 $IND(B)=\{(x, y)\mid((x, y)\in U\times U)\wedge(\forall a\in B, f(x, a)=f(y, a))\}$.

定义 3[7] 决策表 $S=\{U, V, R=C\cup D, f\}$, 设 $X\subseteq U$ 为论域的一个子集, $P\subseteq C$, X的关于 P的下近似、上近似分别为 $\underline{P}X=\{x\in U\mid[x]_P\subseteq X\}$、$\overline{P}X=\{x\in U\mid[x]_P\cap X\neq\phi\}$.

定义 4[7] 若属性集 $B\subseteq C$ 是决策表 $S=\{U, V, R=C\cup D, f\}$ 的一个相对约简, 则 B要满足两个条件: (1) $H(D\mid B)=H(D\mid C)$; (2) $\forall a\in B, H(D\mid B-\{a\})\neq H(D\mid C)$.

第一个条件是从条件熵定义了相对约简, 第二个条件保证了约简集中没有冗余的属性.

2 小生境粒子群属性约简算法

2.1 目标函数

属性约简的目标是在能够代表原属性集的分类能力的前提下, 得到的约简长度越短越好. 因此, 本文提出如下目标函数:

$$f=\frac{|C|-|R|}{|C|}+\frac{1}{H(D\mid R)-H(D\mid C)+1}+r, \tag{1}$$

式中, C为整个条件属性; R为得到的条件属性子集; $H(D\mid R)$、$H(D\mid C)$ 分别为 R、C对应的条件熵; r为惩罚因子, 若 $H(D\mid R)=H(D\mid C)$, 则 $r=1$, 否则 $r=0$. 另外, 每个粒子代表一个可行解即一个约简集, 长度为 $|C|$, 每位上用 1, 0表示该属性在解中还是不在解中. 目标函数的值越大表示适应度值越高.

2.2 粒子群算法

粒子群算法起源于对鸟群捕食过程的模拟. 在算法中, 有若干粒子, 每个粒子代表一个解, 在每一次迭代中, 粒子通过跟踪两个“极值”更新自己, 一个是粒子本身所找到的最优解, 称为个体极值 $pbest_k$; 另一个极值是整个种群目前找到的最优解, 该极值是全局极值 $gbest_k$. 每个粒子根据如下公式更新自己的速度和新的位置:

$$\begin{aligned} v_{k+1} &= c_0 v_k + c_1(pbest_k - x_k) + c_2(gbest_k - x_k), \\ x_{k+1} &= x_k + v_{k+1}, \end{aligned} \tag{2}$$

式中, x_k 代表第 k个粒子目前的解位置, v_k 代表粒子飞行速度.

为使粒子群算法适用于属性约简, 式 (2) 修改如下:

$$x_k = (x_k \otimes pbest_k), \tag{3}$$

$$x_k = (x_k \otimes gbest_k), \tag{4}$$

$$x_k = \odot x_k \tag{5}$$

式 (3)、(4) 表示粒子 x_k 分别与个体极值 $pbest_k$ 和全局极值 $gbest_k$ 作交叉运算, 式 (5) 表示 x_k 作变异运算. 交叉策略采用双点交叉, 与局部解交叉时, 随机取局部解二位交叉; 与全局解交叉时, 随机取该全局解的两位交叉; 变异策略采取单点变异, 随机选取一位, 由 0变为 1, 或 1变为 0.

2.3 小生境淘汰策略

小生境技术就是要维持种群的多样性, 寻找到多个最优解. 首先, 按照小生境半径 R对粒子进行划分种群范围. 然后, 每个小生境种群独立搜索最优解, 全局最优解在所有种群的最优解中产生. 为保证各种群独立进化, 要控制各种群的搜索范围. 文献 [8] 给出了一种实现较简单、求解效率高的小生境淘汰策略, 一旦最优解进入了其它种群的搜索空间, 重新初始化该粒子, 并重新搜索该种群内的最优解. 各种群的独立进化可保证搜索到多峰值解, 避免全部粒子作为一个种群陷入局部最优.

小生境淘汰策略如下:

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

(1)初始化种群数;
(2) for任意两种群 do
　　if(两种群的最优解的距离 $d_{ij} < R$)　　//最优解进入了其它种群范围
　　then
　　{
　　比较两种群最优解的适应度值,高者不变,
　　低者重新初始化并且重新搜索种群内最优解;
　　}
(3)若所有种群的最优解都在自己的小生境范围内则结束,否则转(2)执行.

2.4 小生境竞争策略

小生境淘汰策略的本质是要求搜索到的最优解要在各自的小生境范围中,否则重新搜索最优解.该策略虽可有效保证解的多样化,但增加了算法额外的开销,且执行伴随在自始至终的迭代中.为加快算法的执行,可通过竞争策略提高粒子的搜索能力,以加快迭代.为提高粒子的寻优能力,提出以下竞争策略:

(1)对于连续一定代数的最优解适应度值不改变的种群重新初始化,但是要保护最优个体的种群;

(2)每隔一定代数对最优个体适应度值最低的种群,重新初始化.

策略(1)可避免搜索停滞不前,陷入局部最优.策略(2)可有效保证种群中粒子的活力,避免时间浪费在无效的搜索上.

2.5 小生境粒子群属性约简算法的实现

结合小生境技术及粒子群算法,本文提出小生境粒子群属性约简算法,以此求出决策表的多个最小约简.算法首先以小生境半径 R生成若干个子种群,每个子种群独立进行进化,搜索最优解,每一次迭代结束由子群的最优解更新全局最优解;然后,采用小生境淘汰策略确保各种群独立进化,实施小生境竞争策略提高粒子的活力,提高全局解的搜索能力.

具体算法实现步骤如下:

the Algorithm of Attribute Reduction based on Niche PSO(简称 AARNP)

输入:粒子数 P,种群数 M,迭代次数 N

输出:一个或多个约简集

(1)设置种群数 M,初始化每个种群内的 P个粒子,迭代次数 N,初始化最优解;
(2) for t=1 to N do
{
每个种群内按式(3),(4),(5)进行粒子群寻优,找到每个种群的最优解;
实施小生境淘汰策略;
每隔一定代数检查种群的最优解,实施小生境竞争策略;
由各种群的最优解更新全局最优解 gbest;
}
(3)输出全局解包含的属性即为约简集.

3 实验结果

为验证本文提出的算法的效率,对机器学习数据库中的两个数据集进行了实验.本文的算法称为 AARNP,未加入小生境技术的粒子群属性约简算法称为 AARP.实验平台配置为: Pentium Ⅳ/Intel 2.93G/512MB,Windows XP,80G硬盘,开发环境为 Java.图 1给出了对 mushroom表求最优解的进化图,表 1给出两种算法最优解的比较,表 2给出两种算法在最优解的长度及个数上的比较.

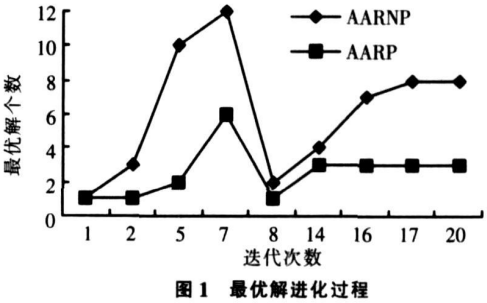

图 1 最优解进化过程
Fig.1 The optimal solutions evolution

从表 1、表 2可知,本文的 AARNP算法在求多个约简结果方面,效果优于 AARP,加入小生境技术的 AARNP算法求

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

得的最优解个数更多, 解的长度更短.

由图 1可知, 随着进化, 两种算法的解个数都在逐渐递增, 但 AARNP算法找到的最优解个数更多, 到了第 7代时, AARNP算法的解个数增长为 12个, AARP仅为 6个. 到第 8代时两种算法的最优解个数锐减, 这是由于找到了更优的解, 最优解清空, 重新开始最优解的搜索. 从第 8代开始, 两种算法继续搜索到越来越多的最优解. 但随着约简的进行, 第 17次迭代时, AARNP算法已找到 8个解; 而 AARP算法到第 14代后一直停滞不前, 只有 3个解, 这是由于 AARNP算法加了小生境技术, 所有粒子划分为多个种群, 可有效保证解的多样化; 而 AARP算法未加小生境算法, 全部粒子只追逐到一个种群或部分种群的最优解.

表 1 最优解结果比较

Table 1 Comparison of the optimal solutions

数据集	AARNP	AARP
Mushroom	{C3, C4, C11, C20}; {C3, C5, C11, C22} {C4, C11, C20, C22}; {C4, C8, C11, C20} {C4, C7, C11, C20}; {C5, C20, C21, C22} {C5, C11, C15, C22}; {C5, C15, C21, C22}	{C5, C20, C21, C22} {C5, C11, C15, C22}; {C5, C15, C21, C22}
Abalone	{C1, C2, C4, C6}; {C1, C3, C4, C5} {C1, C4, C5, C6}; {C2, C3, C4, C7} {C2, C4, C5, C8}; {C4, C5, C7, C8}	{C2, C3, C5, C6, C7} {C2, C3, C5, C6, C8} {C2, C3, C6, C7, C8} {C2, C5, C6, C7, C8}

表 2 最优解个数及长度

Table 2 The number and length of the optimal solutions

数据集	条件属性数	记录数	最优解解长度		解个数	
			AARNP	AARP	AARNP	AARP
Mushroom	22	8 124	4	4	8	3
Abalone	8	4 177	4	5	6	4

4 结语

粒子群算法实现简单, 求解效率高, 但易陷入局部最优. 因此, 本文引入小生境技术, 提出了小生境粒子群属性约简算法, 有效地求解多峰值问题. 实验结果表明, 加入小生境技术后的 PSO属性约简算法能在较短的时间内得到多个最小约简, 避免了 PSO属性约简算法易早熟收敛的不足.

[参考文献] (References)

[1] Wang Jue, Miao Duoqian. Analysis on attribute reduction strategies of rough set[J]. Journal of Computer Science and Technology, 1998, 13(2): 189-193.

[2] 徐章艳, 刘作鹏, 杨炳儒, 等. 一个复杂度为 max(O(|C||U|), 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.
Xu Zhangyan, Liu Zuopeng, Yang Bingru, et al. A quick attribute reduction algorithm with complexity of max(O(|C||U|), O(|C|2|U/C|))[J]. Computer Journal, 2006, 29(3): 391-399. (in Chinese)

[3] Oh I S, Lee J S, Moon B R. Hybrid genetic algorithms for feature selection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, 26(1): 1424-1437.

[4] Jiang Yuanchun, Liu Yezheng. An attribute reduction method based on ant colony optimization[C]// Proceedings of the 6th World Congress on Intelligent Control and Automation. Washington: IEEE Computer Society Press, 2005: 3542-3546.

[5] Dai Jianhua, Chen Weidong, Guo Hongying, et al. Particle swarm algorithm for minimal attribute reduction of decision data tables[C]// Processing of the 1st International Multi-symposiums on Computer and Computational Science. Washington: IEEE Computer Society Press, 2006: 3021-3025.

[6] 王杨. 基于小生境遗传算法的粗糙集属性约简方法[J]. 计算机工程, 2008, 34(5): 66-70.
Wang Yang. Rough set attribute reduction algorithm based on nicheGA[J]. Computer Engineering, 2008, 34(5): 66-70. (in Chinese)

[7] Wang Guoyin, Zhao Jun, Jiu Jiang, et al. Theoretical study on attribute reduction of rough set theory: comparison of algebra and information views[C]// Proceedings of the 3rd IEEE International Conference on Cognitive Informatics. Washington: IEEE Computer Society Press, 2004: 148-155.

[8] Lee C G, Cho D H, Jung H K. Niche genetic algorithm with restricted competition selection for multimodal function optimization[J]. IEEE Trans on Magnetics, 1999, 35(3): 1122-1125.

[责任编辑: 严海琳]

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net