一种基于并行遗传算法的粗糙集属性约简*)

吕跃进 刘南星 陈 磊
(广西大学数学与信息科学学院 南宁 530004)

摘 要 指出现有粗糙集属性约简算法的不足,考虑并行遗传算法在处理大型数据库上的特有优势,将粗糙熵作为粗糙集不确定性的度量,给出一种求解信息系统约简集的三群体并行遗传算法。最后通过实例计算表明该算法能快速有效求解属性约简,而且对大规模数据样本的信息系统效果更为明显。

关键词 粗糙集,粗糙熵,属性约简,并行遗传算法

Rough Set Attribute Reduction Algorithm Based on PGA

LV Yue-Jin LIU Nan-Xing CHEN Lei
(School of Mathematics and Information Science, Guangxi University, Nanning 530004)

Abstract Reduction of attribute is one of the important topics in the search of rough set theory. Although many algorithms for reduction of attribute have been proposed, most of them have some defects. On the other hand, parallel genetic algorithm has some advantages to deal with huge data sets. In this paper, rough entropy is used to measure the uncertainties of rough set. Then a new three population parallel genetic algorithm is presented to solve reduction of attribute from data sets. It is testified by the experiment that the algorithm is effective and it is faster than ordinary algorithms, especially for huge data sets.

Keywords Rough set, Rough entropy, Reduction of attribute, Parallel genetic algorithm

属性约简一直是粗糙集理论的核心研究内容之一,即在保持信息系统分类能力不变的基础上,删除其中冗余或不重要的属性[1,2]。现已证明计算属性最小约简集的复杂性是随信息系统的规模增大而呈指数形式增长,是一个 NP-hard 问题[3],因此寻求属性约简的各种高效的近似算法具有重要的实际意义。目前现有的约简算法主要有基于属性重要性的启发式算法以及基于差别矩阵的约简算法,但有的对处理大规模数据样本的信息系统并不适用,有的全局优化能力差或者不具有普遍性,文[4]提出了一个用常见启发式算法不能求得正确约简的信息系统作为反例。

遗传算法(GA)是由美国密执安大学的 Holland 教授提出的一种通用的求解优化问题的自适应搜索方法,它使用非单点的群体概率搜索机制,具有通用、并行、稳健、全局优化能力强的特点[5]。由此 Wroblewskij 等人首先提出了三种寻求信息系统最小约简的算法[6],取得了不错的效果,但因为算法本身简单粗糙而具有收敛速度较慢的缺陷;文[7,8]各自提出了基于 GA 的约简算法,但没有利用 GA 并行运算在处理大规模数据样本的信息系统上的优势,并且易陷于过早收敛;文[9]提出一种主从式并行遗传约简算法,效果有所改进,但主从式并行遗传算法主、子处理器负载不均衡极大地延长了搜索时间。

本文利用粗糙熵衡量属性的重要程度,给出了一种三群体并行遗传算法(下文简称三群体 PGA)求属性的最小约简集,在保持种群的多样性的同时降低了搜索空间,并选取 UCI 经典数据集[10]进行了有效性测试。

1 信息系统的粗糙熵

熵可以用来给出系统结构的不确定性度量,粗糙熵不同于传统的 Shannon 熵,它具有补的性质,可以更精确地度量粗糙集与粗糙分类的模糊性。利用粗糙熵,可以给出约简规则。

定义 1[2] 设 $S=(U,C)$ 是一个信息系统,其中 U 为有限非空论域,C 为属性集,$U/C=\{X_1,X_2,\cdots,X_n\}$,则 C 的粗糙熵定义为:

$$E(C)=-\sum_{i=1}^{n}(|X_i|/|U|)\log_2(1/|X_i|)$$

其中 $|\cdot|$ 表示集合 $\cdot$ 的基数,$|X_i|/|U|$ 表示等价类 X_i 在论域 U 中的概率。

如果 U/C 为最小划分,则 C 的粗糙熵达到最小值 0;如果 U/C 为最大划分,则 C 的粗糙熵达到最大值 $\log_2|U|$。显然如上定义的粗糙熵有界:$0\leqslant E(C)\leqslant\log_2|U|$。

定理 1 设 $S=(U,C)$ 是一个信息系统,对 $P,Q\subseteq C$,我们有:

(1)若 $P\subset Q$,则 $E(P)>E(Q)$

(2)若 $P=Q$,则 $E(P)=E(Q)$

(3)若 $P\supset Q$,则 $E(P)<E(Q)$

定义 2 设 $S(U,C)$ 是一个信息系统,$c'\in C$ 是一个属性,定义属性 c' 在 C 中的重要性为:

$$sig_C(c')=E(C\backslash\{c'\})-E(c)$$

$sig_C(c')$ 的值越大,说明 c' 在 C 中提供的信息量越大、越重要,某种程度上说就是越"好"。并且 $sig_C(c')$ 具有如下性质:

*)广西大学科研基金项目(No. X032016)资助。吕跃进 教授,研究方向为运筹学与信息管理、数据挖掘;刘南星 硕士研究生,研究方向为粗糙集与粒计算、智能信息处理;陈 磊 硕士研究生,研究方向为决策理论、智能计算。

(1)$0 \leqslant sig_C(c') \leqslant \log_2|U|$;

(2)属性 c' 在 C 中是必要的$\Leftrightarrow sig_C(c')>0$;

(3)$core(C)=\{c' \in C | sig_C(c')>0\}$。

定理 2 设 $S=(U,C)$是一个信息系统,$P \subseteq C$,若 $E(P)=E(C)$,且$\forall c' \in P$ 有:$sig_P(c')>0$,则称 P 是 C 的一个约简。

2 三群体 PGA

GA作为一种基于生物界自然选择和遗传原理的高效搜索技术,对于一般的信息系统,普通 GA 可在合理计算时间内求得满意的属性约简,但随着信息系统规模的扩大,搜索过程将成倍延长,因此致力于提高 GA 的搜索速度是将 GA 应用于粗糙集属性约简的一个重要研究方向。本文利用 GA 的天然并行结构,结合信息系统属性约简的特点给出了一种三群体粗粒度 PGA。

2.1 编码方式与适应值函数

采用二进制编码将属性 $C=\{c_1,c_2,\cdots,c_n\}$的幂集映射为染色体,染色体为编码长度为 n 的 0-1 字符串,每一位按顺序对应相同下标数的条件属性,如某位取 1 表示选择其对应属性;如某位取 0 表示删除其对应属性。记个体空间为:

$$H_L=\{X=a_1a_2,\cdots,a_n | a_i \in \{0,1\}, i=0,1,\cdots,n\}$$

适应值函数的确定应反映应用者对个体(候选解)的评价标准和搜索原理,好的染色体串应具有较高的适应函数值,具有较强生存能力。对个体 $X=a_1,a_2,\cdots,a_n$ 定义适应值函数如下:

$$f(X)=\alpha 2^{-E(P)}+\beta(n-\sum_{i=1}^{n}a_i)$$

其中 P 为 X 对应选择的属性集,$E(P)$为 P 的粗糙熵,n 为条件属性集 C 的长度,$\sum_{i=1}^{n}a_i$ 为未删除的条件属性个数,α,β 为平衡系数,一般取 $1<\alpha \leqslant |U|, 0<\beta \leqslant 1/|C|$。由定理 2 知最小约简是满足粗糙熵尽量小的含有最少属性个数的集合,即在迭代过程中要求 $E(P)$单调下降,而 $\alpha(n-\sum_{i=1}^{n}a_i)$单调上升,从而适应值函数递增,因此此适应值函数符合群体的正确进化方向。

2.2 三群体 PGA 子种群的划分

考虑到 $sig_A(a)$的实际意义,$\forall c_i \in C, i=1,2,\cdots,n$,分别计算 $sig_C(c_i)$值,并由大到小排序:

$$c_{i_1} \geqslant c_{i_2} \geqslant \cdots \geqslant c_{i_{[n/2]}} \geqslant c_{i_{[n/2]+1}} \geqslant c_{i_{[n/2]+2}} \geqslant \cdots \geqslant c_{i_n}$$

其中 $i_1 i_2,\cdots,i_n$ 为一个 n 级排列,$[\cdot]$表示对自然数·取整,记:$I_1=\{i_j | 1 \leqslant j \leqslant [n/2]\}$,$I_2=\{i_j | [n/2]+1 \leqslant j \leqslant n\}$,对个体 $X=a_1a_2,\cdots,a_n$,若$|\{a_{i_j}=1 | i_j \in I_1\}|>[n/4]$且$|\{a_{i_j}=1 | i_j \in I_2\}|<[n/4]$,则将 X 分配到处理器 P_1 上;若$|\{a_{i_j}=1 | i_j \in I_1\}|<[n/4]$且$|\{a_{i_j}=1 | i_j \in I_2\}|>[n/4]$,则将个体 X 分配到处理器 P_3 上;若$|\{a_{i_j}=1 | i_j \in I_1\}|=[n/4]$且$|\{a_{i_j}=1 | i_j \in I_2\}|=[n/4]$,则将 X 复制一次分别分配到处理器 P_1 与 P_3 上;余下的个体分配到处理器 P_2 上。

2.3 各处理机间信息交换

(1)交换的时间与频率:采用同步方式每五代各处理机间两两交换信息一次。

(2)交换对象及交换信息量:各子种群交换适应值最大的个体并替代各自的最差个体。

2.4 选择算子

对于各处理器的独立进化进程,采用适应值比例选择方法,通过轮盘赌实现。对于给定规模为 N 的种群,个体 X_j 被选择的概率为:

$$P(X_j)=f(X_j)/\sum_{i=1}^{N}f(X_i)$$

2.5 交叉与变异算子

(1)采用一致交叉算子,染色体位串上的每一位按相同交叉概率 p_c 进行随机交叉,若对所选择的两个位串 $X_i=a_{i1}a_{i2},\cdots,a_{in}(i=1,2)$生成的新个体为 $X'_i=a'_{i1}a'_{i2},\cdots,a'_{in}(i=1,2)$,则操作描述如下 $O(p_c,x)$:

$$a'_{1j}=\begin{cases}a_{1j}, x_j>1/2\\ a_{2j}, x_j \leqslant 1/2\end{cases} \quad a'_{2j}=\begin{cases}a_{2j}, x_j>1/2\\ a_{1j}, x_j \leqslant 1/2\end{cases}$$

$j \in \{1,2,\cdots,n\}$。其中 x_j 为取值为[0,1]上符合均匀分布的随机变量。对三个子种群分别实行不同的交叉概率:$p_{c1} \leqslant p_{c2} \leqslant p_{c3}$。

(2)采用均匀变异算子,通过变异概率 p_m 随机反转某位等位基因的二进制字符值来实现,对于给定的位串 $X=a_1a_2,\cdots,a_n$,操作描述如下:

$$O(p_m,x):a'_j=\begin{cases}1-a_j, & \text{若 } x_j \leqslant 1/2\\ a_j, & \text{否则}\end{cases}$$

$j \in \{1,2,\cdots,n\}$。x_j 为[0,1]上符合均匀分布的随机变量。对三个子种群分别实行不同变异概率:$p_{m1} \leqslant p_{m2} \leqslant p_{m3}$,并对后代实行精英保留策略。

2.6 终止准则

粗糙集属性约简问题很难有通用的搜索终止准则,我们指定最大进化代数 k 来终止算法。

2.7 三群体 PGA 运算流程

三群体 PGA 具体计算步骤如下:

步 1:计算每个条件属性 $c_i \in C(i=1,2,\cdots,n)$在 C 中的重要性 $sig_C(c_i)$并排序。

步 2:(初始化)随机产生 N 个个体,按上文方式生成三个初始种群:$p_1(0),p_2(0),p_3(0)$,分别分配到处理机 p_1,p_2,p_3 上,置 $t \leftarrow 0$。

步 3:(种群进化)各子种群 $p_i(t)$按照标准串行遗传算法进行进化。

步 4:(信息交换)若进化代数为 5 的倍数,各子种群之间交换最优个体,实现个体迁移,生成新一代种群 $p_i(t+1)$,否则,直接生成 $p_i(t+1)$。

步 5:(终止准则)如果终止条件满足,停机并输出 $p_i(t+1)$中的最优个体作为近似解之一,否则置 $t \leftarrow t+1$ 并转步 3。

3 实例测试

为了验证本文提出的三群体 PGA 求解粗糙集属性约简有效与否,我们选取如表 1 所示来源于 UCI[10] 的两个经典数据集进行测试。

表 1 选自 UCI 的两个数据集

数据集	ZOO	LYMPHOGRAPHY
对象个数	37	148
属性个数	17	19

3.1 独立测试

我们对数据集 ZOO 按照本文提出的算法进行计算,采用的参数设置为 $N=100$;$\alpha=20$,$\beta=1/17$,$p_{c1}=0.7$,$p_{c2}=0.75$,$p_{c3}=0.8$,$p_{m1}=0.025$,$p_{m2}=0.03$,$p_{m3}=0.035$ 运行 150 代终止,结果求得此数据集的数十个约简,我们用挪威科技大学与

波兰华沙大学合作开发的粗糙集分析工具包 ROSETTA 同样进行了计算，结果显示本算法的结果包含了所有最小约简，而运算时间略微缩短，实验发现，个体最为优秀的处理器 p_1 往往首先到最优解。图 1～图 3 显示了 p_1 上个体适应值的变化情况。

3.2 比较测试

用本文提出的算法与文[6]中 Wroblewskij 提出的第一种算法(在 ROSETTA 实现，取交叉概率 $p_c=0.7$，变异概率 $p_m=0.05$)分别对数据集 LYMPHOGRAPHY 进行了计算，图 4～图 5 显示了测试结果。可以看到，本文算法收敛速度明显快于文[6]中算法，并且更为稳定。从图 1～图 5 容易看出每次的交换信息可以使得最优个体适应值及个体平均适应值发生较大的跃迁，极大地缩短了收敛时间。本算法拥有如下特点：

(1)合理的适应值函数决定了正确的进化方向，其中 $\alpha 2^{-E(P)}$ 起主调作用，$\beta(n-\sum_{i=1}^{n}a_i)$ 起微调作用。

(2)算法实行带有信息迁移的并行结构，极大改善了算法的收敛速度。

(3)利用属性约简的特点，对属性进行排序，进而划分为“优”、“良”、“劣”的三个子种群，并分别实行不同的进化参数设置，一定程度上保持了种群的多样性，从而全局优化能力强。

结束语 本文采用粗糙熵衡量属性的重要程度，进而通过设立合理的适应值函数、独特的种群划分方法提出了一种三群体 PGA 求属性约简，通过二个实例测试表明本算法的有效性，并且拥有较快的收敛速度及 GA 所拥有的鲁棒性。

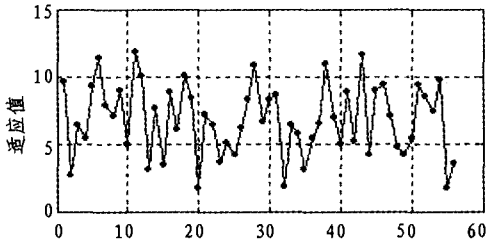

图 1 初始群体适应值分布

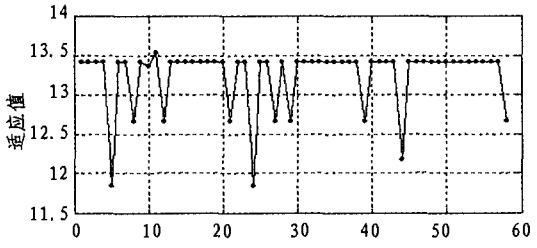

图 2 终止代数群体适应值分布

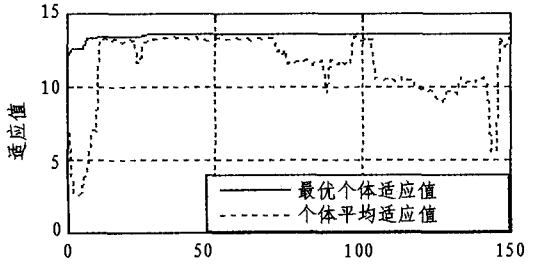

图 3 最优值与平均值随进化次数的收敛过程

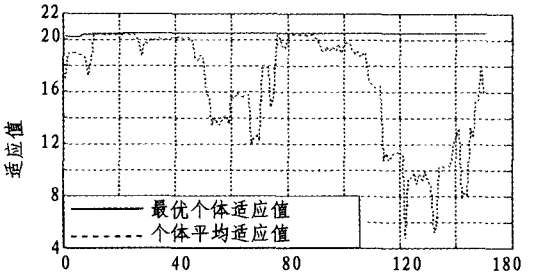

图 4 三群体 PGA 最优值与平均值随进化次数的收敛过程

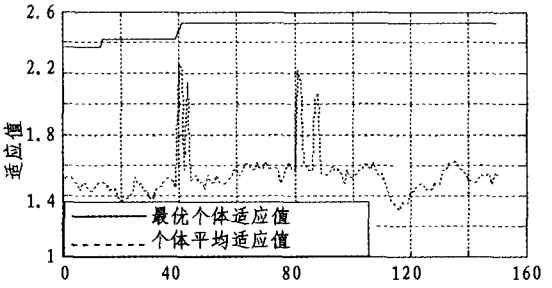

图 5 文[6]中 GA 最优值与平均值随进化次数的收敛过程

但本文仅仅是将多群体 PGA 应用于粗糙集属性约简的一个初步尝试，算法本身还存在着某些不足，主要体现在：(1)对算法的参数设置还没有成熟的原则和方法，存在依赖主观经验的不足。(2)PGA 在并行运算的同时需要大量的通信开销，一定程度上增加了运算的代价。因此探讨参数的合理取值范围及减少通信开销将是下一步的重要工作。

参 考 文 献

1 Pawlak Z. Rough sets: Theoretical Aspects of Reason-ing about Data [M]. Dordrecht: Kluwer Academic Publishers, 1991

2 梁吉业，李德玉. 信息系统中的不确定性与知识获取[M]. 北京：科学出版社，2005

3 Wong S K M, Ziarko W. On optimal decision rules in decision tables [J]. Bulletin of Polish Academy of Sciences, 1985, 33: 693～696

4 Wang Jue, Miao Duoqian. Analysis on Attribute Reduction Strategies of Rough Set [J]. Journal of Computer Science and Technology, 1998, 13(2): 189～192

5 王小平，曹立明. 遗传算法——理论、应用与软件实现[M]. 西安：西安交通大学出版社，2002

6 Jakub W. Finding minimal reducts using genetic algorithms [R]: [Research Report 16/65]. Warsaw university of Technology, 1995

7 Tumer M B, Demir M C. A genetic approach to data dimensionality reduction using a special initial population [A]. In: International Work-Conference on the Interplay between Natural and Artificial Computation [C]. Las Palmas, Canary Islands, Spain, 2005. 310～316

8 李订芳，章文，李贵斌，牛艳庆. 基于可行域的遗传约简算法[J]. 小型微型计算机系统，2006, 27(2): 312～315

9 朱克敌，陶 志. 并行遗传算法在粗糙集属性约简中的应用[J]. 沈阳工程学院学报，2005, 1(1): 70～73

10 UCI repository of machine learning database. URL: http://www.ics.uci.edu/～mlearn/databases/