非完备信息系统的启发式特征选择遗传算法

戴大蒙[1,2],慕德俊[1]
(1.西北工业大学自动化学院,陕西西安 710072; 2.温州大学物理与电子信息工程学院,浙江温州 325035)

摘 要: 为了获取非完备信息系统的相对最小特征子集,提出一种基于非完备信息系统的启发式特征选择遗传算法.本文首先构造了适应度函数,并以特征重要度为启发式信息融入特征选择;同时利用特征的相对核对种群初始化,引导染色体的进化,缩小了算法的搜索空间;且在染色体的交叉和变异过程中,对满足条件的染色体及时删除,加快算法的收敛性;实验结果验证了算法的有效性.

关键词: 非完备信息系统; 特征选择; 遗传算法; 启发式方法
中图分类号: TP311 **文献标识码:** A **文章编号:** 0372-2112 (2013)03-0451-05
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2013.03.006

Heuristic Genetic Algorithm for Feature Selection in Incomplete Information Systems

DAI Da-meng[1,2], MU De-jun[1]
(1. *School of Automation, Northwestern Polytechnical University, Xi'an, Shaanxi* 710072, *China*;
2. *College of Physics & Electronic Information Engineering, Wenzhou University, Wenzhou, Zhejiang* 325035, *China*)

Abstract: In this paper, in order to get the minimal relative reduction of features set, heuristic genetic algorithm for feature selection in incomplete decision table is proposed. At first, the fitness function of genetic algorithm is presented. Meanwhile, regarding feature significance as heuristic information in feature selection, and relative core feature serves as initial population to optimize chromosome, which can reduce the exploration space, what's more, the corresponding condition chromosomes are deleted in the crossover and mutation processes, this method can accelerate the convergence. At last, the better effect of the proposed algorithm can be tested by the experiment.

Key words: incomplete decision information; feature selection; genetic algorithm; heuristic method

1 引言

粗糙集(Rough Sets)理论[1,2]是一种处理模糊、不确定和不完备信息的重要数学工具.特征选择作为粗糙集理论研究的核心内容.其目的是在保持知识库一定分类能力不变的条件下,删除其中的冗余知识,导出问题的分类决策.目前,一些学者提出了许多有效的特征选择算法,Liu 等人[3]给出了正区域的渐增式计算方法,在此基础上,设计了一个时间复杂度为 $O(|C|^2|U|\log|U|)$ 的特征选择算法.Xu 等人[4]利用了基数排序计算等价类和近似质量的优化策略,使得约简算法时间复杂度降低为 $\max\{O(|C||U|),O(|C|^2|U/C|)\}$,Liu 等人[5]从完备信息系统的不一致情况出发,提出了基于 Hash 的正区域快速计算方法,同时构造了基于 Hash 的属性约简算法,其时间复杂度为 $O(|C|^2|U/C|)$.Qian 等人[6]利用并行计算方法求解等价类,在此基础上,提出了一种云环境下的高效知识约简算法.Miao 等人[7]提出了图表示下的知识约简,并给出了求最小约简的完备递归算法.Du 等人[8]讨论了不协调决策表几种约简的标准及其内在关系分析.对于完备信息系统而言,特征选择算法的研究取得了显著进步.

现今复杂领域系统中有些数据由于某些原因可能无法获取,导致数据呈现出非完备性,即存在部分特征值未知的情况,这也使得在传统的高效算法无法适应非完备信息系统[9,10],同时人们往往希望能获得信息系统的最小特征子集,而利用粗糙集理论求解最小特征子集已证明是一个 NP-hard 问题.因此,近来年,有学者在非完备信息系统的特征选择中引入遗传算法[11],蚁群算法[12],粒子群算法[13]等来解决此类问题,但算法的计算效率有待改善.鉴于遗传算法求解特征子集时,它不依

收稿日期:2012-03-13;修回日期:2012-06-21
基金项目:浙江省自然科学基金项目(No.Y1101314);浙江省优秀青年教师基金项目(No.2010)

赖于求解问题的具体领域,具有全局优化和隐含并行等优点,为我们提供一种在有限代价内解决搜索和优化问题的有效途径.

由于适应度函数是对种群中的染色体个体的适应性进行评定的唯一准则,直接体现群体的生物进化过程,为此适应度函数的构造在遗传算法中至关重要.对于非完备信息系统的研究,文献[14]中给出了一个基于遗传算法的特征选择框架,并设计了相应的特征选择算法,但所设计的适应度函数并不能得到完备的特征子集,且文献[15~17]中均存在算法收敛速度较慢的现象.这对大规模非完备性数据集而言,将严重影响算法的计算效率,为此针对非完备信息系统,本文首先构造了一个新的适应度函数,同时在适应度函数中引入控制因子,从而保证得到的特征子集的完备性.在特征选择过程中,以特征重要度作为启发式信息,可显著减少算法的搜索空间,并利用特征的相对核作为初始种群,加快算法的收敛速度,引导染色体的优化,且在算法中加入修正算子加快染色体的进化方向,并在交叉和变异过程中,对满足条件的染色体进行及时剔除,提高了算法在解空间中的探索能力和计算效率,以达到求解相对最小特征子集的目的.并通过实验验证了算法的有效性,能较好地处理非完备的大规模数据.

2 基本理论

定义 1[9] 五元组 $S=(U,C,D,V,f)$是一个完备决策表,其中 $U=\{u_1,u_2,\cdots,u_n\}$表示对象的非空有限集合,称为论域;$C=\{c_1,c_2,\cdots,c_r\}$表示条件特征的非空有限集;D 表示决策特征的非空有限集,且 $C\cap D=\varnothing$;$V=\bigcup_{a\in C\cup D}V_a$,$V_a$ 是特征 a 的值域,$f:U\times C\cup D\rightarrow V$ 是映射函数,它对每个对象的每个特征赋予一个信息值,即 $\forall a\in C\cup D,u\in U$,有 $f(u,a)\in V_a$.若至少存在一个特征 $a\in C$,使得 V_a 包含空值(用 $*$ 表示),即至少有一个特征 $a\in C$,存在一个 $u\in U$,使得 $f(u,a)=*$,则称之为非完备信息系统 IIS.

定义 2[9] 在非完备信息系统 $IIS=(U,C,D,V,f)$中,令 $B\subseteq C$,定义 U 上的容差关系 $T(B)$为:$T(B)=\{(u_i,u_j)\in U\times U\mid\forall b\in B,\ f(u_i,b)=f(u_j,b)\vee f(u_i,b)=*\vee f(u_j,b)=*\}$.

定义 3[10] 在非完备信息系统 $IIS=(U,C,D,V,f)$中,$Y\subseteq U$,$\forall B\subseteq C\cup D$,记 $B_-(Y)=\{u\in U\mid T(B)\subseteq Y\}$为 Y 关于 B 的下近似集,$B^-(Y)=\{u\in U\mid T(B)\cap Y\neq\varnothing\}$为 Y 关于 B 的上近似集.

定义 4 在非完备信息系统 $IIS=(U,C,D,V,f)$中,$\forall B\subseteq C$,$U/D=\{D_1,D_2,\cdots,D_t\}$表示由决策特征集 D 对论域 U 的划分,称 $POS_C(D)=\bigcup_{D_i\in U/D}C_-(D_i)$为 C 关于 D 的正区域.

定义 5 在非完备信息系统 $IIS=(U,C,D,V,f)$中,若 $\forall a\in B\subseteq C$,若 $POS_B(D)=POS_{B-\{a\}}(D)$,则称 b 为 B 中相对于 D 是不必要的;否则称 a 为 B 中相对于 D 是必要的.对 $\forall B\subseteq C$,若 B 中任一元素相对于 D 都是必要的,则称 B 是相对于 D 独立的.

定义 6 在非完备信息系统 $IIS=(U,C,D,V,f)$中,若 $\forall B\subseteq C$,$POS_B(D)=POS_C(D)$且 B 相对于 D 是独立的,则称 B 是 C 的相对于 D 的一个特征选择.

定义 7 在非完备信息系统 $IIS=(U,C,D,V,f)$中,令 $R(C)$表示 C 的相对于 D 的所有基于正区域的特征选择子集的集合,则特征的相对核集为 $Core(C)=\bigcap_{B\in R(C)}B$.

定义 8 在非完备信息系统 $IIS=(U,C,D,V,f)$中,令条件特征集 C 对决策特征集 D 的依赖程度定义为 $r(C,D)=|POS_C(D)|/|U|$.

定义 9 在非完备信息系统 $IIS=(U,C,D,V,f)$中,设 $B\subseteq C$,特征 B 关于决策特征 D 的重要性定义为 $sig(B,C,D)=r(C,D)-r(C-B,D)$;当 $B=\{a\}$时,特征 a 关于 D 的特征重要性定义为 $sig(a,C,D)=r(C,D)-r(C-a,D)$.

性质 1 $0\leqslant sig(a,C,D)\leqslant 1$.

性质 2 特征 $a\in C$ 在特征 C 中对 D 是必要的充要条件是 $sig(a,C,D)>0$.

性质 3 $Core=\{a\in C\mid sig(a,C,D)>0\}$.

由性质 3 可容易求解核 Core,将特征的相对核集加入到初始种群中,可显著加快算法的收敛性,引导染色体的优化.

3 改进的遗传算法

3.1 适应度函数的构造

在遗传算法中,适应度函数是控制种群进行自然选择方向的唯一依据,它的构造直接影响到遗传算法的收敛速度以及是否能找到相对最小特征子集.针对非完备信息系统构造了适应度函数为:$F(x)=f(x).\delta+k(x)=((|C|-H_x)/|C|)\cdot(1/\sqrt{|U|})+|POS_x(D)|/|U|$,其中,$H_x$ 表示染色体 x 中存在的特征个数.构造的适应度函数:第一部分为 $f(x)=((|C|-H_x)/|C|)$,其中 $f(x)$用于控制染色体向特征变少的方向进化,显然当 x 中含特征的个数越少时,$f(x)$的值将会越大,被选择的概率也越大,这为求解相对最小特征子集前提条件,$\delta=1/\sqrt{|U|}$为控制因子,用于保证染色体向特征选择方向进化;第二部分为 $k(x)=|POS_x(D)|/|U|$,决定了决策特征 D 对染色体 x 的依赖度,当 $k(x)$的值越大时,说明决策特征 D 对染色

体 x 的依赖性越强.且当 $k(x)=1$ 时,说明决策特征完全由染色体 x 所包含的条件特征确定.为此,所设计的适应度函数可以在保证找到真正的特征选择的情况下找到相对最小的特征子集.

3.2 编码方法的设定

由于在特征选择过程中,条件特征只存在是否被选择这两种情况,故采用二进制编码方法,同时在初始种群初始化时,引入特征的相对核集到初始种群的染色体中,加快算法的收敛性.为此设非完备信息系统的条件特征集为 $C=\{c_1,c_2,\cdots,c_r\}$,染色体为二进制基因位串,其对应一个特征子集.二进制"1"表示相应位的特征被选择,二进制"0"表示相应位的特征未被选择,对于相对核集中的特征可在初始化全部编码为"1".

3.3 最优个体优化策略

在染色体个体选择的过程中,需关注的两个方面是:一方面,尽快找到最优染色体所对应的特征为特征子集;另一方面,使得获得的特征子集中特征个数最少.为此,根据定义 9 以特征重要度作为启发式信息来缩小算法的搜索空间和加快引导染色体进化.则最优个体优化策略为:计算当前染色体 x 的 $\mathrm{POS}_B(D)$值,其中 B 表示染色体 x 所对应的特征集;若 $\mathrm{POS}_C(D)=\mathrm{POS}_B(D)$,则计算特征集 B 中所有特征所对应的特征重要性,选择特征重要性最小的特征,并将其在染色体 x 中所对应的基因位"1"修改为"0";否则选择特征重要性最大的特征,并将其在染色体 x 中所对应的基因位"0"修改为"1",从而得新染色体 x';并计算新染色体 x' 的适应度值 $F(x')$.

3.4 交叉操作策略

采用单点交叉操作策略,其具体过程为:随机对父代群体中的染色体进行两两配对,根据交叉概率 P_c 随机设置交叉点的位置,对每一对相互配对的染色体在设定好的交叉点处交换两个个体的部分染色体,替换重组后而形成两个新的染色体.在交叉运算中,若当条件满足 $F(x)>|\mathrm{POS}_C(D)|/|U|$ 且 $H_{x'}>H_x$ 时,则删除新产生的染色体 x',可减少适应度函数的计算时间.上述交叉运算策略依据是新进化后的染色体 x' 所包含的特征个数比原来父代种群中的染色体 x 的特征个数更多,可直接将新进化后的染色体 x' 删除,因为此时对染色体 x' 进行计算将不可能得到相对最小特征子集,这样可有效地提高遗传算法的收敛速度.

3.5 变异操作策略

采用位变异策略,其具体过程为:对父代群体中的染色体个体进行两两配对,随机选取某一互异基因座的位置,依变异概率 p_m 指定其为变异点,对每一对染色体指定的变异点,特征的相对核集所对应的基因位不变异,其它则对其基因值做取反运算,从而形成新的染色体.在变异运算中,当条件满足 $F(x)>|\mathrm{POS}_C(D)|/|U|$ 且 $H_{x'}>H_x$ 时,则删除新产生的染色体个体 x',以加速算法的收敛.同时在对染色体的基因位进行变异时,只需将"1"变为"0",减少染色体个体中条件特征的数目,以达到求最小特征选择的目的.

4 启发式特征选择遗传算法

下面构造基于非完备信息系统的启发式特征选择遗传算法,算法描述如下:

启发式特征选择遗传约算法

输入:非完备信息系统 $\mathrm{IIS}=(U,C,D,V,f)$,种群大小 k,种群最大进化代数 T,最优染色体迭代总次数 M,交叉概率 P_c,变异概率 Pm.

输出:相对最小特征子集 R.

Step1 令 $t=1,f=1,\mathrm{Core}=\varnothing$;// f 为最优染色体迭代次数

Step2 计算 C 相对于 D 的正区域 $\mathrm{POS}_C(D)$;

Step3 计算每个特征 $a\in C$ 的重要性 $\mathrm{sig}(a,C,D)$,若 $\mathrm{sig}(a,C,D)>0$,则 $\mathrm{Core}=\mathrm{Core}\cup\{a\}$;若 $\mathrm{POS}_{\mathrm{Core}}(D)=\mathrm{POS}_C(D)$,则 Core 为相对最小特征子集,转至 Step6,否则执行 Step4;

Step4 随机产生 k 个二进制串组成初始群体,若特征 $a\in\mathrm{Core}$,其基因位取值为 1;若特征 $a\notin\mathrm{Core}$,则其基因位随机取值为 0 或 1;

Step5 当条件满足 $t\leqslant T$ 且 $f\leqslant M$ 时,执行如下:

(1)对群体中的每个染色体 x,计算其所对应的适应度值 $F(x)$,将适应度值最大的个体遗传至下一代群体中,并由选择概率 $CF(x)=F(x)/M$ 计算出每个染色体被选择的概率,采用以轮盘赌方式进行个体的选择;

(2)若染色体满足 3.3 节中最优个体优化策略,则对其进行修正;

(3)随机取出一对染色体,根据交叉概率 P_c 进行交叉操作,若满足 3.4 节中交叉操作条件的染色体,则将其删除;

(4)根据变异概率 P_m 进行变异操作,若满足 3.4 节中变异操作条件的染色体,将其基因位进行变异操作;

(5) $t=t+1;f=f+1$.

Step6 输出相对最小特征子集 R,算法结束.

5 实验结果及分析

为了验证本文算法的性能,从 UCI 数据库中选取了数据集 4 个不同的数据集(其中,Data-1 和 Data-2 是从 Mushroom 数据集中随机抽取的数据)作为测试数据

分别与文献[14]和文献[17]中的算法进行实验比较,为了便于比较,将文献[14]中的算法记为 A 算法,文献[17]中的算法记为 B 算法,其中硬件环境为 Intel Core2, CPU E7400 和 1GB 内存,利用 C# 语言在 Microsoft Visual Studio 2005 中编程实现.其中本文算法的参数设置如下:初始种群数 $K=15$,种群最大进化代数 $T=50$,最优迭代数 $M=400$,交叉概率为 $P_c=0.8$,变异概率为 $P_m=0.02$,实验的仿真结果如表 1 所示,表 1 中的"代数"指第几代时出现最优染色体,"时间"为最优染色体出现时的执行时间(单位为秒).

表 1 算法的实验性能比较

数据集	样本数	特征数	A 算法		B 算法		本文算法	
			代数	时间	代数	时间	代数	时间
Echoc	74	13	5	1.012	5	1.317	3	0.609
Hepat	155	19	14	3.304	17	4.560	10	1.812
Soybe	307	35	22	10.281	25	12.065	16	3.027
Data-1	2000	22	15	11.350	17	19.293	12	4.901
Data-2	3000	22	17	23.094	20	28.189	11	6.024

从图 1 的实验结果可知,本文的算法在最优染色体出现的代数要比 A 算法和 B 算法早,其主要原因是本文算法将非完备信息系统的特征的相对核作为初始种群中的染色体和采用最优个体优化策略,使较优的染色体得以优先保留,加快了算法的收敛性,A 算法虽然在最优染色体个体出现的代数比 B 算法较早,主要是因为 A 算法所求得特征子集是非完备的,例如在 Data-1 数据集中,A 算法所求得特征选择含 6 个条件特征,而本文算法所求得的最优特征子集中只含 4 个条件特征.同时由于本文算法的适应度函数计算简洁,且在染色体的交叉和变异过程中,对满足条件的染色体进行了及时删除,过滤了许多不含候选特征子集的染色体个体,显著缩小算法在可行解上的搜索空间,有效地提高了算法的计算效率.特别是当数据集的规模越大时,本文算法的高效性也越明显,且在特征选择过程中能获得相对最小特征子集.

6 结束语

特征选择作为粗糙集研究的重要内容.针对复杂领域系统中数据的非完备性,为了快速从这类非完备数据中获取相对最小特征子集,提出了一种基于非完备信息系统的启发式特征选择遗传算法.本文首先构造了一个适应度函数,且在适应度函数中引入控制因子,从而保证得到的特征子集的完备性.并以特征重要性为启发信息引入遗传算法,同时在初始种群初始化时,引入特征的相对核集到初始种群的染色体中,加快算法的收敛性;并在交叉和变异过程中删除部分不满足条件的染色体,缩小算法的搜索空间.最后通过 UCI 数据集进一步验证了算法在计算效率和最优特征子集选择过程中的优势,且在处理规模较大的非完备数据集时算法的优势更加明显.

参考文献

[1] Pawlak Z. Rough set theory and its application to data analysis [J]. Cybernetics and System, 1998, 29(27): 661 – 688.

[2] Pawlak Z, Skowron A. Rudiments of rough sets[J]. Information Sciences, 2007, 117(1): 3 – 37.

[3] 刘少辉,盛秋戬,吴斌,等. Rough 集高效算法的研究[J]. 计算机学报,2003,26(5):524 – 529.
Liu Shao-hui, Sheng Qiu-jian, Wu Bin, et al. Research on efficient algorithms for rough set methods[J]. Chinese Journal of Computers, 2003, 26(5): 524 – 529. (in Chinese)

[4] 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 Zhang-Yan, Liu Zuo-peng, Yang Bing-ru, et al. A quick attribution reduction algorithm with complexity of max{$O(|U||C|), O(|C|^2|U/C|)$}[J]. Chinese Journal of Computer, 2006, 29(3): 391 – 399. (in Chinese)

[5] 刘勇,熊蓉,褚健. Hash 快速属性约简算法[J].计算机学报,2009,32(8):1493 – 1499.
Liu Yong, Xiong Rong, Chu Jian. Quick attribute reduction algorithm with hash[J]. Chinese Journal of Computer, 2009, 32(8): 1493 – 1499. (in Chinese)

[6] 钱进,苗夺谦,张泽华.云计算环境下知识约简算法[J].计算机学报,2011,34(12):2332 – 2343.
Qian Jin, Miao Duo-qian, Zhang Ze-hua. Knowledge reduction algorithms in cloud computing[J]. Chinese Journal of Computer, 2011, 34(12): 2332 – 2343. (in Chinese)

[7] 苗夺谦,陈玉明,王睿智,等.图表示下的知识约简[J].电子学报,2010,38(8):1952 – 1957.
Miao Duo-qian, Chen Yu-ming, Wang rui-zhi, et al. Knowledge reduction algorithm under graph view[J]. Acta Electronica Sinica, 2010, 38(8): 1952 – 1957. (in Chinese)

[8] 杜卫锋,秦克云.不协调决策表几种约简标准及其关系分析[J].电子学报,2011,39(6):1336 – 1340.
Du Wei-feng, Qin Ke-yun. Analysis of several reduction standards and their relationship for inconsistent decision table[J]. Acta Electronica Sinica, 2011, 39(6): 1336 – 1340. (in Chinese)

[9] Kryszkiewicz M. Rough set approach to incomplete information systems[J]. Information Sciences, 1998, 112(1): 39 – 49.

[10] Kryszkiewicz M. Rules in incomplete information systems[J]. Information Sciences, 1999, 113(2): 271 – 292.

[11] Wroblewski J. Finding minimal reduces using genetic algorithms[A]. Proc of 2nd Annual Joint Conf on Information

Sciences[C]. Wrightsville Beach: JC1S, 1995. 186 – 189.
[12] Ke Liangjun, Feng Zuren, Ren Zhigang. An efficient ant colony optimization approach to attribute reduction in rough set[J]. Pattern Recognition Letters, 2008, 29(9): 1351 – 1357.
[13] Wang Xiangyang, Yang Jie, Teng Xiaolong. Feature selection based on rough sets and particle swarm optimization[J]. Pattern Recognition Letters, 2007, 28(4): 459 – 471.
[14] Riyaz Sikora, Selwyn Piramuthu. Framework for efficient feature selection in genetic algorithm based data mining[J]. European Journal of Operational Research, 2007, 180(2): 723 – 737.
[15] 李订芳,章文,李贵斌,等.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.
Li Ding-fang, Zhang Wen, Li Gui-bin, et al. Genetic reduction algorithm based on feasible region[J]. Journal of Chinese Computer Systems, 2006, 27(2): 312 – 315. (in Chinese)
[16] Mingyuan Zhao, Chong Fu, et al. Feature selection and parameter optimization for support vector machines: A new approach based on genetic algorithm with feature chromosomes[J]. Expert Systems with Applications, 2011, 38: 5197 – 5204.
[17] M E EIAImi. A filter model for feature subset selection based on genetic algorithm[J]. Knowledge-Based Systems, 2009, 22: 356 – 362.

作者简介

戴大蒙　女.1975 年 7 月出生于浙江瑞安.西北工业大学自动化学院在职博士,副教授,浙江省高校教坛新秀.主要从事智能信息处理、数据挖掘和粒计算等方面的研究工作.
E-mail: jsj_ddm@wzu.edu.cn

慕德俊　男.1963 年 6 月,西北工业大学自动化学院教授,“控制科学与工程”学科和“计算机科学与技术”学科博士生导师。主要研究方向:控制理论与应用、智能信息处理和网络信息安全.
E-mail: mudejun@nwpu.edu.cn