

基于克隆选择和量子进化的 GEP 分类算法

王卫红¹ 杜燕焯¹ 李 曲^{1,2}

(浙江工业大学计算机科学与技术学院 杭州 310023)¹ (浙江大学计算机学院 杭州 310038)²

摘要 基于基因表达式编程(GEP)的分类算法具有较高的精度,但易陷入局部最优,且搜索时间长。为进一步提高 GEP 分类算法的分类能力,提出了基于克隆选择和量子进化的 GEP 分类算法——ClonalQuantum-GEP。该算法通过量子种群的更新和探测影响抗体种群的搜索方向和进化能力,并通过记忆池保持最优解,使其具有更好的种群多样性、更强的全局寻优能力和更快的收敛速度。在几个标准数据集上的实验验证了算法的有效性。与基本的 GEP 算法相比,ClonalQuantum-GEP 能以较小的种群规模和较少的进化代数获得较理想的分类效果。

关键词 基因表达式编程(GEP),克隆选择,量子进化,分类

中图法分类号 TP311.13 **文献标识码** A

GEP Classification Based on Clonal Selection and Quantum Evolution

WANG Wei-hong¹ DU Yan-ye¹ LI Qu^{1,2}

(College of Computer Science, Zhejiang University of Technology, Hangzhou 310023, China)¹

(College of Computer Science, Zhejiang University, Hangzhou 310038, China)²

Abstract Gene Expression Programming based Classification algorithm has shown good classification accuracy, however, it often falls into the local optimums and needs long time searching. In order to further improve the classification power of GEP, clonal selection and quantum evolution were introduced into GEP. A novel approach called ClonalQuantum-GEP was proposed. After affecting the search direction and evolution ability of the antibody population through the updating and exploring of the quantum population, and keeping the best results in the memory pool, this approach gets more population diversity, better ability of global optimums searching, and much faster velocity of convergence. Experiments on several benchmark data sets demonstrate the effectiveness and efficiency of this approach. Compared with basic GEP, ClonalQuantum-GEP can achieve better classification results with much smaller scale of the population and much less evolutionary generation.

Keywords Gene expression programming, Clonal selection, Quantum evolution, Classification

1 GEP 介绍

基因表达式编程 (Gene Expression Programming, GEP)^[1]是葡萄牙学者 Candida Ferreira 在遗传算法 (Genetic Algorithm, GA)和遗传程序设计 (Genetic Programming, GP)基础上提出来的。GEP 结合了 GA 和 GP 的优点,克服了它们的缺点,在数学建模方面有很好的效果。

GEP 基因型个体由头部 (Head)和尾部 (Tail)构成,头部元素来自函数符集和终结符集,尾部元素来自终结符集。对每个问题而言,头部长度 h 是根据问题的特点而选定的,而尾部的长度 t 是 h 和 n 的函数,关系如下: $t=h(n-1)+1$, 其中 n 是所需变量数最多的函数的参数个数 (也称为最大操作数)。GEP 个体表现型被称为表达式树 (Expression Tree, ET)。表达式树是通过顺序扫描基因型个体元素,按照层次顺序构成的。

考虑如下由函数集 $F=\{Q, +, -, *, /\}$ 和终点集 $T=\{a, b\}$ 构成的基因。其中, Q 表示开方函数,为单目运算符,其它运算符为二目运算符,即 $n=2$; 取 $h=10$, 则 $t=10*(2-1)+1=11$; 所以基因 g 的长度为 $10+11=21$ 。下面给出一个这样的基因 (尾部用粗体标识):

012345678901234567890
* b + a - a Q a b + **b a b b a b b a b a**

它编码成如图 1 所示的表达式树。

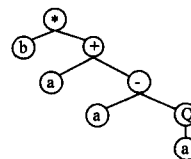


图 1 表达式树结构

对应的表达式按从上到下、从左到右的顺序从表达式树

到稿日期:2010-11-11 返修日期:2011-04-18 本文受国家自然科学基金(60873033),浙江省自然科学基金(R1090569),浙江省科学技术厅项目(2009C31108)资助。

王卫红(1969-),男,教授,主要研究方向为空间信息服务、电子政务和电子商务、网络信息安全等,E-mail:wwh@zjut.edu.cn;杜燕焯(1985-),女,硕士生,主要研究方向为演化计算、数据挖掘;李 曲(1979-),男,博士生,讲师,主要研究方向为演化计算、数据挖掘。

上读取为 $b * (a + (a - \sqrt{a}))$ 。

2 克隆选择原理

克隆选择是生物免疫系统理论的重要学说,其中心思想是:抗体是天然产物,以受体的形式存在于细胞表面,抗原可与其选择性反应。抗原与相应抗体受体的反应可导致细胞克隆性增殖,该群体具有相同的抗体特异性,其中某些细胞克隆分化为抗体生成细胞,另一些形成免疫记忆细胞,以参加之后的二次免疫反应。克隆选择是生物体免疫系统自适应抗原刺激的动态过程,这一过程中所体现出的学习、记忆、抗体多样性等生物特性正是人工免疫系统所借鉴的。此概念已被广泛用于计算机编程、系统控制、交互式并行仿真等领域,其学习、记忆等特性也引起了人工免疫系统研究者的兴趣^[2]。

3 量子进化算法

进化算法本质上是一种模仿生物系统、种群之间适应环境、相互作用、不断进化、优化的过程,从而体现了群智能的特点。融合不同智能优化算法,是一个重要的研究方向。为了发挥量子算法的加速作用,将量子算法和遗传算法、进化算法等相融合,一般将它们统称为量子进化算法^[3]。

3.1 量子比特

在量子计算中,采用 $|0\rangle$ 和 $|1\rangle$ 表示微观粒子的两种基本状态,称为量子比特。一个量子比特可以处于 $|0\rangle$ 态、 $|1\rangle$ 态以及 $|0\rangle$ 和 $|1\rangle$ 之间的任意叠加态^[3]。一个量子比特的状态可描述为

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle \quad (1)$$

式中, α, β 是复数,称量子比特对应态的概率幅。 $|\alpha|^2$ 表示量子态被观测为 $|0\rangle$ 的概率, $|\beta|^2$ 表示量子态被观测为 $|1\rangle$ 态的概率,且满足归一化条件

$$|\alpha|^2 + |\beta|^2 = 1 \quad (2)$$

如果一个系统有 m 个量子比特,则该系统可同时描述 2^m 个状态,然后在观测时,该系统将坍塌为一个确定的状态。

3.2 量子遗传算法

1996年, A. Narayanan 和 M. Moore 首先提出量子遗传算法(Quantum Genetic Algorithm, QGA),并将其成功应用于解决 TSP 问题^[4]。QGA 是基于量子比特和量子叠加态的概念被提出的。

传统进化计算的染色体编码可有多种方式:二进制、十进制、符号编码等。QGA 采用基于量子比特的编码方式。一个量子比特可由其概率幅定义为 $[\alpha \ \beta]^T$, 同理, m 个量子比特可定义为

$$\begin{bmatrix} \alpha_1 & \alpha_2 & \dots & \alpha_m \\ \beta_1 & \beta_2 & \dots & \beta_m \end{bmatrix} \quad (3)$$

式中, $|\alpha_i|^2 + |\beta_i|^2 = 1, i = 1, 2, \dots, m$ 。这种描述的优点在于可以表达任意量子叠加态。例如,有一个 3 比特量子系统,若有 3 对概率幅

$$\begin{bmatrix} 1/\sqrt{2} & 1 & 1/2 \\ 1/\sqrt{2} & 0 & \sqrt{3}/2 \end{bmatrix} \quad (4)$$

则系统状态可描述为

$$\frac{1}{\sqrt{2}}|000\rangle + \frac{\sqrt{3}}{2\sqrt{2}}|001\rangle + \frac{1}{2\sqrt{2}}|100\rangle + \frac{\sqrt{3}}{2\sqrt{2}}|101\rangle \quad (5)$$

以上结果表明,系统呈现 $|000\rangle, |001\rangle, |100\rangle, |101\rangle$ 的概

率分别为 $1/8, 3/8, 1/8, 3/8$ 。所以由式(4)描述的 3 比特量子系统能够同时包含 4 个状态的信息。

QGA 的一般步骤如下^[3,5]。

(1) 进化代数初始化: $T=0$;

(2) 初始化种群 $Q(t)$;

(3) 由 $Q(t)$ 生成 $P(t)$;

(4) 评价群体 $P(t)$ 的适应度, 保存最优解;

(5) 停机条件判断, 当满足停机条件时, 输出当前最优个体, 算法结束, 否则继续;

(6) 更新 $Q(t), T=T+1$, 转到(3)。

与遗传算法类似, QGA 也是一种概率搜索算法, 拥有一个量子种群 $Q(t) = \{q_1^t, \dots, q_n^t\}$, 其中 n 表示种群规模, t 表示遗传代数, q_j^t 表示一条量子染色体, 其定义为

$$q_j^t = \begin{bmatrix} \alpha_1^t & \alpha_2^t & \dots & \alpha_m^t \\ \beta_1^t & \beta_2^t & \dots & \beta_m^t \end{bmatrix} \quad (6)$$

式中, m 是量子比特数, 即量子染色体的长度, $j = 1, 2, \dots, n$ 。

在步骤(6)中, 为获得更加优良的染色体, 将二进制解集 $P(t)$ 与当前存储的最优解比较, 用适当的量子门 $U(t)$ 更新种群 $Q(t)$ 。量子门可根据实际问题具体设计, 通常采用的量子旋转门定义为

$$U(t) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \quad (7)$$

式中, θ 是旋转角度。

4 基于克隆选择和量子进化的 GEP 分类算法

4.1 分类策略

4.1.1 “一对其他(one-against-all)”学习策略

(a) 对于二分类问题, 根据 GEP 表达式返回值的正负决定其所属的类别, 即

$$\text{IF } GEP_j(X) > 0, \text{ THEN } X \in \text{Class } j, \text{ ELSE } X \notin \text{Class } j \quad (j=0, 1) \quad (8)$$

式中, X 为输入特征向量。

(b) 对于多分类问题, 采用“一对其他”策略, 即将 n -类 ($n > 2$) 问题分解为 n 个二分类问题。为每个类 C_i 建一个二分类问题, 其中所有属于类 C_i 的样本被看作正类, 而其他样本全作为负类^[6]。

4.1.2 适应度函数

适应度函数^[6,10,11]如下。

$$\text{fitness}(R) = \begin{cases} 0, & \text{if } \text{consig}(R) < 0 \\ \text{consig}(R) * \exp(\text{compl}(R) - 1), & \text{otherwise} \end{cases} \quad (9)$$

$$\text{consig}(R) = \left(\frac{p}{p+n} - \frac{P}{P+N} \right) * \frac{P+N}{N}, \text{compl}(R) = \frac{p}{P}$$

式中, p 和 n 分别是被规则 R 覆盖的正例和负例的数目, 而 P 和 N 分别是训练集中正例和负例的数目, $\exp(\cdot)$ 是以 e 为底数的指数函数。

4.1.3 两阶段规则裁剪

(a) 前阶段裁剪

对于每一类样例, 算法挖掘多个规则去覆盖之。至于具体挖掘多少个规则, 在算法执行过程中根据 MDL 原则决定。一个规则集 H 的描述长度^[6,12] 计算如下:

$$L(H) = L_{\text{exception}}(H) + W * L_{\text{theory}}(H) \quad (10)$$

式中, $0 < W < 1$, 我们取 $W = 0.5$ 。而 $L_{\text{exception}}(H)$ 定义如下:

$$L_{\text{exception}}(H) = \log_2 \left(\binom{n}{N_{fp}} \right) + \log_2 \left(\binom{N-n}{N_{fn}} \right) \quad (11)$$

式中, n 是被规则集 H 覆盖的样例数目, N 是总的训练样例数目, 而 N_{fp} 和 N_{fn} 分别是假正例和假负例的数目。 $L_{\text{theory}}(H)$ 定义如下。

$$L_{\text{theory}}(H) = \log_2(N_c) * \sum_{i=1}^S L(R_i) \quad (12)$$

式中, S 是规则集 H 的大小, $L(R_i)$ 是表达式 R_i 对应的染色体的有效长度, 而 N_c 是 GEP 中用到的不同的符号数目, 即函数符集、变量集和常数集的总数。

(b) 后阶段裁剪

裁剪前阶段训练得到的所有规则, 以获得按适应度值排序后的简洁规则集, 测试时按此顺序依次判断即可^[6]。

后阶段裁剪步骤如下。

- (1) 根据适应度值对前阶段产生的所有规则进行降序排列。
- (2) 将具有最高适应度值的规则加入到有序规则集中(保存最后结果的规则集, 初始为空)。
- (3) 删除被选择的规则覆盖的所有正例和负例。
- (4) 基于剩下的样例, 重新计算剩下的规则的适应度值。
- (5) 重复步骤(1)–(4), 直至没有正例剩下或剩下的规则没有一个适应度值为正。
- (6) 从剩下的所有未被分类的样例中指定默认类。

4.2 抗体、抗原在 GEP 分类中的解释及绑定

每一个抗体是一个 GEP 编码的规则, 用于分类判断。每一个抗原表示一个分类中的类别以及属于该类别的一个样本数据集, 该方法是由 V. K. Karakasis 在文献[7]中提出的。

抗体与抗原之间的绑定如图 2 所示。两者的绑定关系是根据抗体在与抗原所代表的类别关联的样本数据集上的分类效果来确定的。其中, 两者之间绑定关系上的数字代表抗体与抗原之间的亲和度。

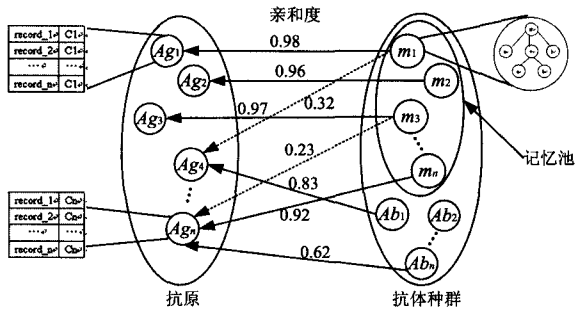


图 2 抗体-抗原绑定

4.3 基于 GEP 由量子种群生成抗体种群的策略

与一般的 GEP 个体生成不同, 本文算法中的抗体种群是根据量子种群来生成的。具体生成策略如下。

每个抗体(GEP 染色体)的每个基因由头和尾构成, 头部元素来自函数符集和终结符集, 尾部元素来自终结符集, 我们将其统称为可取元素符集。

对于每个抗体 α_i 的每个基因位 j , 顺序遍历与之对应的可取函数符集 $\{e_1, e_2, \dots, e_k\}$ 。在每一步 t , 随机生成一个 $[0, 1]$ 范围的数 r , 如果 $r > |\alpha_{ij}|^2$ (其中 α_{ij} 为与该抗体 α_i 对应的量子个体 q_i 的相应基因位 j 的概率幅), 则该抗体的该基因位 j

赋为 e_t ; 否则继续往后遍历。如果遍历结束, 没有一个符合, 则随机赋予 $\{e_1, e_2, \dots, e_k\}$ 中的某一元素。

4.4 ClonalQuantum-GEP 分类算法流程

ClonalQuantum-GEP 分类算法是在文献[6-8]基础上进行改进提出的。算法基本流程与文献[6]相似。根据个体特点, 文献[6]中的适应度函数用亲和度替换。下面描述核心进化过程, 算法流程如图 3 所示。

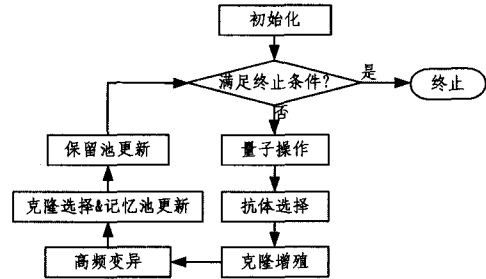


图 3 ClonalQuantum-GEP 算法流程

4.4.1 初始化

(a) 量子种群初始化: 初始量子种群 Q 中的量子个体是所有状态的等概率线性叠加, 即所有量子个体中的 α 和 β 均置为 $1/\sqrt{2}$ 。

(b) 抗体种群初始化: 抗体种群 A 由记忆池 A_m 和保留池 M 两部分组成, 即 $A = A_m + M$ 。根据 4.3 节介绍的策略, 由量子种群 Q 初始化抗体种群 A (A_m 和 M 两部分), 并计算其亲和度。

4.4.2 量子操作

(a) 更新量子种群

1) 量子旋转处理: 对量子种群中每条染色体, 应用量子旋转门进行旋转。

用于更新量子比特相位的量子旋转门为

$$U(\Delta\theta) = \begin{bmatrix} \cos(\Delta\theta) & -\sin(\Delta\theta) \\ \sin(\Delta\theta) & \cos(\Delta\theta) \end{bmatrix} \quad (13)$$

更新过程即为

$$\begin{bmatrix} \cos(\Delta\theta) & -\sin(\Delta\theta) \\ \sin(\Delta\theta) & \cos(\Delta\theta) \end{bmatrix} \begin{bmatrix} \cos(t) \\ \sin(t) \end{bmatrix} = \begin{bmatrix} \cos(t + \Delta\theta) \\ \sin(t + \Delta\theta) \end{bmatrix} \quad (14)$$

由上式可知, 该量子旋转门只改变量子比特的相位, 而不改变量子比特的长度。

转角 $\Delta\theta$ 的方向和大小直接影响到算法的收敛速度和效率。为简化转角 $\Delta\theta$ 方向的确定方法, 本文参照文献[8]给出如下定理。

定理 1 令 $w_0 = [\alpha_0 \ \beta_0]^T$ 是当前搜索到的最佳解的概率幅, $w_1 = [\alpha_1 \ \beta_1]^T$ 是当前解的概率幅, 记 $E = (w_0 \ w_1)$, 则转角 $\Delta\theta$ 的方向按如下规则选取: 当 $|E| \neq 0$ 时, 方向为 $-\text{sgn}(|E|)$; 当 $|E| = 0$ 时, 方向取正负均可(证明见文献[8])。

关于转角的大小, 文献[4]虽然给出了一个范围 $(0.005\pi, 0.1\pi)$, 但没有给出具体的选择依据。为方便起见, 本文就在该范围内随机选择一个值作为转角的大小。

2) 量子变异处理: 对量子种群中每条染色体, 应用量子非门按变异概率实施变异。

采用量子非门实现量子染色体变异。首先根据变异概率随机选择一条量子染色体, 然后随机选择若干个量子比特施

加量子非门变换,使该量子比特的两个概率幅互换。这样可使上下两条链同时得到变异。这种变异实际上是对量子比特幅角的一种旋转:如设某一量子比特幅角为 t ,则变异后的幅角为 $\pi/2-t$,即幅角正向旋转了 $\pi/2-2t$ 。这种旋转不与当前最佳染色体比较,一律正向旋转,有助于增加种群的多样性,降低早熟收敛的概率^[8]。

(b)根据更新后的量子种群更新保留池

根据 4.3 节介绍的策略,由更新后的量子种群重新生成抗体种群的保留池部分并计算其亲和度。

4.4.3 克隆增殖

根据抗体的亲和度值,选择抗体种群中 n_b 个最优的抗体构成集合 B ,并按照其亲和度大小降序排列。然后克隆 B 中抗体,生成一个克隆临时种群 C 。克隆规则如下:对于 B 中的每一个抗体 α_i , i 为 α_i 在 B 的序列标号,克隆出 n_i 个抗体放入 C 。 n_i 计算如下

$$n_i = \text{round}\left(\frac{\beta \cdot n_b}{i}\right), 1 \leq i \leq n_b \quad (15)$$

式中, $\text{round}(\cdot)$ 为四舍五入函数, β 是常数,称为克隆系数^[7]。

由式(15)可知, B 中抗体的亲和度越大,由它克隆出来的抗体就越多。

4.4.4 高频变异

对克隆临时种群 C 进行高频变异^[7],由此获得一个变异后的克隆种群 C_m 。高频变异概率 $h(x)$ 通过如下指数函数来确定

$$h(x) = h_{\max} e^{-\rho \cdot x}, \quad h_{\max} \leq 1 \quad (16)$$

式中, h_{\max} 为用户设置的最大变异概率, ρ 为变异概率衰减系数, x 为规范化至 $[0,1]$ 区间的亲和度。

高频变异可以被看作是一个有助于搜索局部最优的细化机制。

4.4.5 克隆选择和记忆池更新

用变异后的克隆种群 C_m 中 n_m 个最好的抗体去更新记忆池。

4.4.6 保留池更新

按照 4.2 节中介绍的策略更新量子种群,然后根据更新后的量子种群生成 n_d 个新抗体,替换保留池中具有最差亲和度的 n_d 个抗体。

5 实验与分析

5.1 实验参数设置

将本文算法 ClonalQuantum-GEP 应用到来自 UCI 数据库^[9]的几个标准数据集上,对每个数据集分别做 5 次 5-折交叉验证。数据集信息详见表 1,其他核心参数设置见表 2。

表 1 数据集信息

数据集			
序号	名称	样本数	类别数
1	Breast Cancer W.	683	2
2	Glass	214	6
3	Ionosphere	351	2
4	Iris	150	3
5	Lung Cancer	32	3
6	Wine	178	3
7	Zoo	101	7

表 2 核心参数设置

参数	意义	值	参数	意义	值
G	最大进化代数	50	N	抗体种群大小	50
M	记忆池大小	1	P_m	量子变异处理概率	0.1
n_b	选择用于克隆的抗体数	5	β	克隆系数	15.0
h_{\max}	最大变异概率	1.0	ρ	变异概率衰减系数	5.0
n_m	用于更新记忆池的克隆体数	1	n_d	用于更新保留池的新生抗体数	2

5.2 实验结果分析

(1) 分类精度比较

从表 3 可以看出,所测试的 7 个数据集中, ClonalQuantum-GEP 在 6 个数据集上的最好精度大于 Basic-GEP,在剩余一个上的相等,且相应的 95% 的置信区间范围更小。其中,对于 BreastCancerW., Iris, LungCancer, Wine 和 Zoo 这 5 个数据集, ClonalQuantum-GEP 的最好精度比 Basic-GEP 的分别高出 1.7%, 4.2%, 41.3%, 1.9% 和 5.2%。在分类的平均精度上, ClonalQuantum-GEP 无明显优势,但能基本保持与 Basic-GEP 接近,其中 Iris 和 LungCancer 数据集的平均精度比 Basic-GEP 的分别高出 1.5% 和 1.1%。

表 3 分类精度比较(%)

序号	数据集名称	Basic-GEP		ClonalQuantum-GEP	
		平均精度	最好精度	平均精度	最好精度
1	Breast Cancer W.	96.2±1.8%	96.5±2.0%	95.7±0.1%	98.1±0.1%
2	Glass	63.9±8.8%	70.1±4.7%	56.4±0.7%	70.7±1.2%
3	Ionosphere	90.2±2.4%	92.6±2.4%	87.4±0.4%	92.6±0.5%
4	Iris	95.3±4.6%	96.0±3.2%	96.7±0.4%	100.0±0.0%
5	Lung Cancer	54.4±15.6%	57.6±25.0%	55.0±6.4%	81.4±6.0%
6	Wine	92.0±6.0%	93.8±2.7%	90.6±0.9%	95.6±0.5%
7	Zoo	93.9±6.9%	95.1±7.8%	93.3±0.4%	100.0±0.0%

(2) 分类效率比较

定义规则集平均总长度 $L_{(R)}$ 如下

$$L_{(R)} = L_r \times N_r \quad (17)$$

式中, L_r 为平均表达式长度,即平均规则长度, N_r 为平均规则数。本文认为 $L_{(R)}$ 可以作为评估分类效率中的一个指标,而平均执行时间是另外一个指标。

从表 4 及图 4 可以看出,所测试的 7 个数据集中,除 LungCancer 之外, ClonalQuantum-GEP 在其他 6 个数据集上挖掘的规则集平均总长度比 Basic-GEP 的明显要小。其中,对于 BreastCancerW., Glass, Ionosphere, Iris 和 Wine 这 5 个数据集, ClonalQuantum-GEP 的规则集平均总长度比 Basic-GEP 的分别减少 28.9%, 68.9%, 20%, 66.7% 和 36%。另一方面,对于平均执行时间,在全部测试的 7 个数据集上, ClonalQuantum-GEP 比 Basic-GEP 都大大缩短,有非常明显的优势。

表 4 分类效率比较

序号	数据集名称	Basic-GEP			ClonalQuantum-GEP			
		平均表达式长度	平均规则数	平均总长度	平均表达式长度	平均规则数	平均总长度	平均时间 [min]
1	Breast Cancer W.	19	4	76	9	6	54	7.5
2	Glass	20	9	180	8	7	56	2.2
3	Ionosphere	10	4	40	4	8	32	3.7
4	Iris	30	5	150	10	5	50	0.7
5	Lung Cancer	5	3	15	3	7	21	1.2
6	Wine	15	5	75	6	8	48	1.7
7	Zoo	7	7	49	5	9	45	1.3

(下转第 262 页)

[3] Kass M, Miller G. Rapid, stable Fluid dynamics for computer-graphics[J]. Computer Graphics, 1990, 24(4): 49-57
 [4] Foster N, Metaxas D. Modeling water for computer animation [J]. Commun ACM, 2000, 43 (7): 60- 67
 [5] 杨怀平, 孙家广. 基于海浪谱的波浪模拟[J]. 系统仿真学报, 2002, 14(9): 1175-1178
 [6] Simon P, Michael A. Rendering Natural Waters [C]// Proceedings of the Pacific Graphics, Hong Kong, China, 2000: 23-30

[7] Peachey D. Modeling waves and surf [J]. Computer Graphics, 1986, 20(4): 65-74
 [8] Fournier A, Reeves W T. A simple model of ocean waves[J]. Computer Graphics, 1986, 20(4): 75-84
 [9] 唐荣锡, 汪嘉业, 彭群生, 等. 计算机图形学教程[M]北京: 科学出版社, 1990
 [10] Angel E. OpenGL 程序设计指南[M]. 李桂琼, 张文祥, 译. 北京: 清华大学出版社

(上接第 239 页)

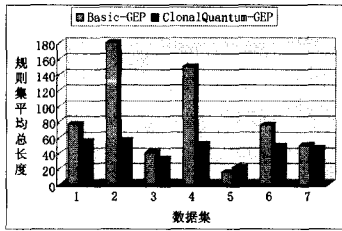


图 4 规则集平均总长度比较

分析(1)、(2)的结果: 本文在保留 GEP 编码优点的基础上, 为改进其遗传操作机制繁琐、费时和易陷入局部最优的缺点, 融入了克隆选择和量子进化原理。克隆选择的重点在于选择与克隆, 并通过记忆池保持最优解, 使其较一般遗传算法具有更好的多样性和更高的收敛效率。量子进化用量子位编码表示染色体, 用量子旋转和量子变异完成进化搜索, 具有“勘探”和“开采”的能力, 收敛速度快, 全局寻优能力强。结合克隆选择和量子进化的优点, ClonalQuantum-GEP 只需要更小的种群规模和更少的进化代数, 就可以搜索到比较理想的解, 分类效率显著提升(规则集平均总长度和分类时间均显著减少)。

(3) 规则集的一个例子

下面列出的是由 ClonalQuantum-GEP 产生的一个规则集, 来自于 Iris 数据集, 并获得了 100% 的最好分类精度。此规则集如下:

- 规则 1 IF $\left\{ \left(\frac{2 - attr3}{2} \right) > 0 \right\} > 0$ THEN class = Iris-setosa
- 规则 2 IF $\left\{ attr4 - \sqrt{attr2} \right\} > 0$ THEN class = Iris-virginica
- 规则 3 IF $\left\{ attr3 - 5 \right\} > 0$ THEN class = Iris-virginica
- 规则 4 IF $\left\{ (1 > 0? attr2; 7) * \left(\frac{attr1}{2 - attr3 + attr2} \right) - \frac{5}{1} \right\} > 0$ THEN class = Iris-versicolor
- 规则 5 IF $\left\{ \left(\frac{attr4 - 1 - 1}{5 - attr4 - attr3} \right) * 2 \right\} > 0$ THEN class = Iris-versicolor

上述表达式中的 attr1 - attr4 分别表示 Iris 数据集的 sepal length, sepal width, petal length, petal width 4 个属性。

结束语 本文在文献[6]的分类策略基础上, 融合克隆选择和量子进化原理, 提出了 ClonalQuantum-GEP 分类算法。此算法充分结合了克隆选择和量子进化的优点, 通过量子种群的更新和探测影响抗体种群的搜索方向和进化能力, 并通

过记忆池保持最优解, 使其具有更丰富的种群多样性、更强的全局寻优能力和更快的收敛速度。在几个数据集上的实验证明, 此算法能以较小的种群规模和较少的进化代数获得较理想的分类效果。同文献[6]相比, 分类的平均精度基本接近, 最好精度有一定的提高, 而分类效率显著提升。

从实验结果来看, 测试的最好精度普遍有一定程度的提高, 但在平均精度上虽接近却不显优势, 这可能和各种参数的设置有关, 如克隆系数、量子变异处理概率等, 可以通过进一步的灵敏度分析来调优, 以期实验效果有更大的改善。

参 考 文 献

[1] Ferreira C. Gene Expression Programming: Mathematical Modeling by an Artificial Intelligence (2nd Edition) [M]. Germany: Springer-Verlag, 2006
 [2] 焦李成, 杜海峰, 刘芳, 等. 免疫优化计算、学习与识别[M]. 北京: 科学出版社, 2006
 [3] 李士勇, 李盼池. 量子计算与量子优化算法[M]. 哈尔滨: 哈尔滨工业大学出版社, 2009
 [4] Narayanan A, Moore M. Quantum-inspired genetic algorithms [C]//IEEE International Conference on Evolutionary Computation, Nagoya, Japan, 1996: 61-66
 [5] 解平, 李斌, 庄镇泉. 一种新的混合量子进化算法[J]. 计算机科学, 2008, 35(2): 166-170
 [6] Zhou C, Xiao W, Tirpak T M, et al. Evolving Accurate and Compact Classification Rules with Gene Expression Programming [J]. IEEE Transactions on Evolutionary Computation, 2003, 7 (6): 519-531
 [7] Karakasis V K, Stafylopatis A. Efficient Evolution of Accurate Classification Rules Using a Combination of Gene Expression Programming and Clonal Selection[J]. IEEE Transactions on Evolutionary Computation, 2008, 12(6): 662-678
 [8] 李士勇, 李盼池. 基于实数编码和目标函数梯度的量子遗传算法 [J]. 哈尔滨工业大学学报, 2006, 38(8): 1216-1218
 [9] Blake C, Keogh E, Merz C J. UCI Repository of Machine Learning Databases[EB/OL]. <http://www.ics.uci.edu/~mllearn/MLRepository.html>, 1998
 [10] Michalski R S, et al. Learning patterns in noisy data: The AQ approach[M]. Paliouras G, et al., eds. Machine Learning and Its Applications, Springer-Verlag, 2001, 2049: 22-38
 [11] Brazdil P B, Torgo L. Knowledge acquisition via knowledge integration[M]. Current Trends in Knowledge Acquisition. Amsterdam. The Netherlands, IOS Press, 1990
 [12] Quinlan J R. C4. 5: Programs for Machine Learning[S]. San Mateo, CA: Morgan Kaufmann, 1993