人工智能 - 遗传算法

杜小勤

武汉纺织大学数学与计算机学院

2016/03/14

遗传算法

- 基本遗传算法；
- 理论基础；

概述

遗传算法模拟达尔文的自然选择理论，最核心的思想是优胜劣汰-适者生存。

自然选择理论包括遗传、变异及适者生存三个方面。

概述

遗传算法运用了生物遗传与进化概念，通过繁殖、变异、竞争等方法，实现优胜劣汰，以逐步得到问题的最优解或次优解。

概述

算法开始时，对 N 个个体（种群）随机初始化，并计算每个个体的适应度函数。

如果不满足优化准则，开始新一代计算：按照适应度选择个体，进行基因重组（交叉），按一定的概率进行变异，重新计算适应度，替换上一代个体。

上述过程循环往复，直到满足优化准则为止。

基本遗传算法

也称为简单的遗传算法，包括几个方面：编码方法、个体评价函数、初始种群、群体大小、选择算子、交叉算子、变异算子、算法终止条件等。

编码方法采用固定长度的二进制编码，个体评价函数采用非负数，初始种群随机产生，仅使用选择、交叉和变异三种遗传算子。

选择方法采用赌轮方法，交叉方法采用单点交叉，变异方法采用基本位变异，算法终止条件为指定的迭代次数或找到最优解（或次优解）。

实例：计算函数的最大值

使用遗传算法计算函数 $y = x^2$ 的最大值，
$x \in [0, 31]$。

编码

将变量 x 编码为 $5-bits$ 串形式，可以表示 $0-31$ 共 32 个数。

00000 表示 $x=0$，00001 表示 $x=1$，...，11111 表示 $x=31$。

初始群体

初始群体的数目 $N = 4$，通过随机的方式生成，例如：

表 1-1: 初始种群

个体编号	个体编码
1	01101
2	11000
3	01000
4	10011

个体适应度评价函数

它是度量个体适应度的函数（fitness function）。

适应度较高的个体遗传到下一代的概率较大，适应度较低的个体遗传到下一代的概率相对就小一些。这是判断个体优良的准则。

本例的适应度函数是：$f(x) = x^2$。

选择（Selection）

选择或复制的目的是为了从当前群体中选出优良的个体，使它们有机会作为父代繁殖下一代。

优良的个体：适应度高的个体。它被选择的机会多。

采用赌轮选择方法。

赌轮选择的依据

表 1-2: 赌轮选择的依据

个体编号	个体编码	x	适应度 $f(x) = x^2$	选择概率 $p_s = \frac{f_i}{\sum f_j}$
1	01101	13	169	0.14
2	11000	24	576	0.49
3	01000	8	64	0.06
4	10011	19	361	0.31

赌轮

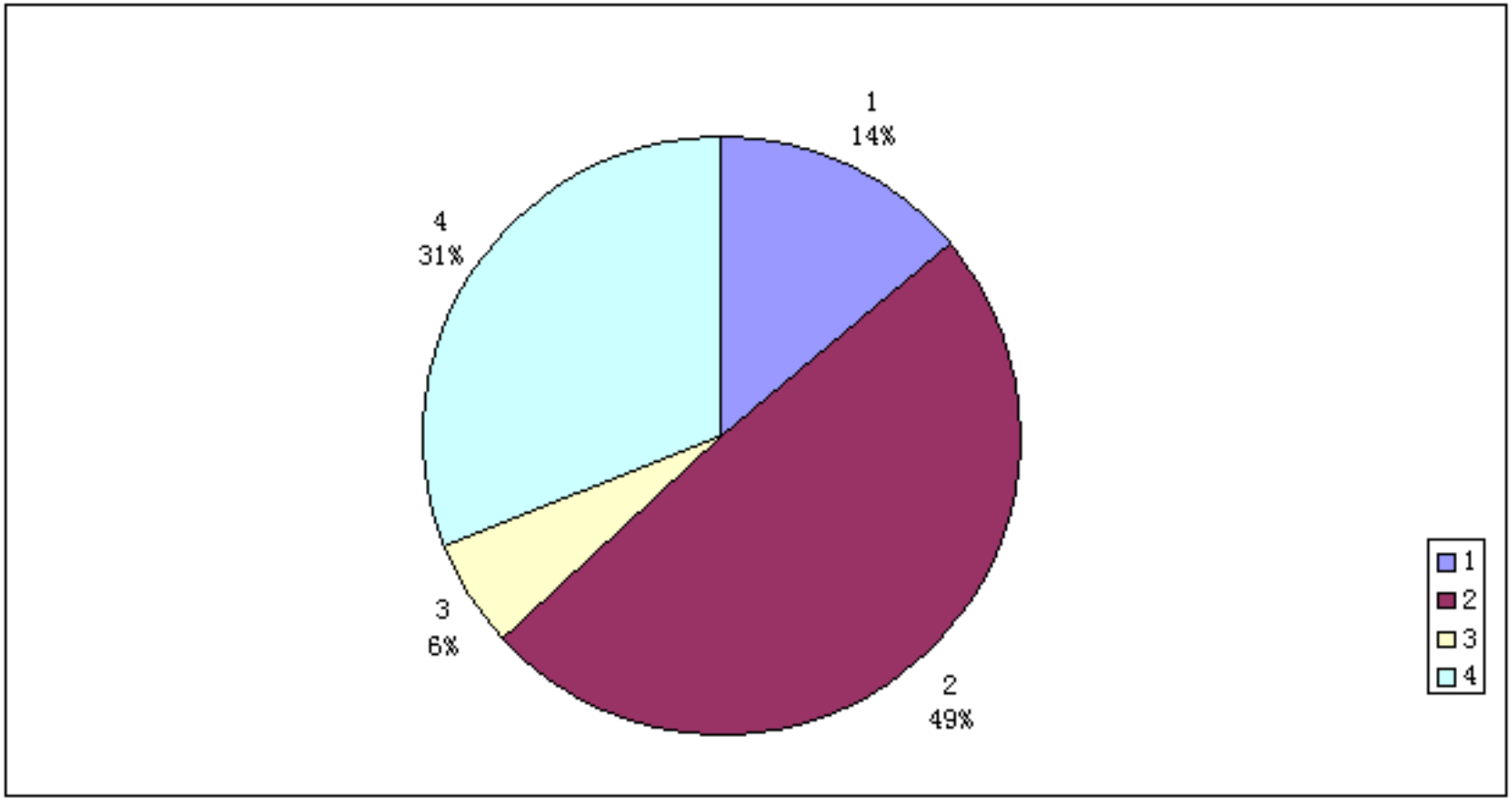

图 1-1: 第一代种群的赌轮

赌轮选择的结果

平均适应度：$\bar{f} = \frac{\sum f_j}{N} = \frac{1170}{4} = 292.5$

表 1-3: 选择结果

个体编号	个体编码	适应度 f_i	适应度/平均适应度 $\frac{f_i}{\bar{f}}$	选择次数
1	01101	169	0.58	1
2	11000	576	1.97	2
3	01000	64	0.22	0
4	10011	361	1.23	1

交叉（Crossover）

交叉操作是遗传算法中最主要的操作，可以得到新一代个体。

交叉概率 P_c 一般都预设为较大的值，例如 0.8 左右。

交叉概率决定了被选中的两个个体是否进行交叉操作：若随机数小于 P_c 则进行交叉操作，否则，不进行。

单点交叉

首先对个体进行随机配对，然后在配对个体中随机设定交叉位置，配对个体彼此交换部分信息。

表 1-4: 单点交叉结果

个体编号	个体编码	交叉位置	新一代个体	适应度
1	01101	4	01100	144
2	11000	4	11001	625
3	11000	2	11011	729
4	10011	2	10000	256

结果

适应度总和 $\sum f_j = 1754$

平均适应度 $\bar{f} = \frac{\sum f_j}{N} = \frac{1754}{4} = 438.5$

新群体的总适应度、平均适应度、最大适应度都有了明显的提高，它的确朝着我们所期望的方向进化了！

变异（Mutation）

变异操作是按 bit 进行的，即把某一位的内容进行取反操作。

变异操作同样也是随机进行的，一般，变异概率 P_m 都取得很小，例如 0.001。

由于群体总共有 $4 \times 5 = 20$ 位，变异的概率为 $20 \times 0.001 = 0.02$。可见变异的概率相当低。

变异操作是十分微妙的遗传操作，需要与交叉操作妥善配合使用，目的是挖掘群体中个体的多样性，克服有可能限于局部解的弊病。

基本算法

```
Input: problem data, GA parameters
Output: the best solution
Begin
    t ← 0;
     initialize P(t) by encoding routine;
     evaluate P(t) by decoding routine;
     while (not terminating condition) do
          create C(t) from P(t) by crossover
     routine;
          create C(t) from P(t) by mutation routine;
          evaluate C(t) by decoding routine;
          select P(t +1) from P(t) and C(t) by
     selection routine;
          t ← t +1;
     end
     output the best solution
  end
```

图 1-2: 基本的遗传算法

理论基础

遗传算法是一种基于群体进化的计算模型，在这个进化过程中，包含了大量的随机操作。

这些随机操作对群体性能优化起到何种作用以及它们的内在原理是什么，有必要做进一步的分析与研究。

20 世纪 70 年代，Holland 提出了基因模式理论。该理论以二进制位串为基础，深入讨论了模拟生物染色体的遗传算法的内在机制，为遗传算法奠定了理论基础。

模式（Schema/Schemata）

编码字符串中有一些相似的结构特征，例如，适应度较高的个体（2 号 11000 和 4 号 10011）可以使用模式 1＊＊＊＊来表示；而适应度较低的个体（1 号 01101 和 3 号个体 01000）可以使用模式 01＊＊＊来表示。

＊表示不必关心的字符。

这种个体结构上存在的相似特征，被称为模式。

模式阶

群体中个体的相似性可以通过个体模式来刻画，而这些模式各不相同，为定量描述这些模式，引入 2 个重要概念。

模式阶：指模式 H 中已有明确含义的字符个数，记作 $O(H)$，例如 $O(*101*) = 3$。

模式阶越低，所代表的字符串个数越多，模式的概括性越强，模式的确定性就越低。反之，模式阶越高，所代表的字符串个数越少，模式的概括性越弱，模式的确定性就越高。

定义距

定义距：指模式 H 中最前面和最后面这 2 个具有明确含义的字符之间的距离，记作 $\delta(H)$，例如 $\delta(10*0*)=3$，$\delta(0****)=0$。

在遗传进化过程中，模式的定义距越长，该模式被破坏的可能性越大，反之，定义距越短，被破坏的可能性越小。例如模式 $0****$ 比模式 $10*0*$ 更难破坏。

即使阶数相同的模式，不同的定义距也会有不同的差异。

模式理论

在遗传进化的过程中，对个体的遗传操作实际上就是对个体模式的操作，不同的模式在群体进化中不断发生着改变。

模式理论就是对模式及运行规律进行研究与分析的理论。

选择算子作用下的模式

设第 t 代群体为 P_t，该群体中模式 H 出现的次数为 N_H^t。

那么，以第 t 代群体 P_t 为基础进行选择操作后，得到了第 $t+1$ 代群体 P_{t+1}。于是，出现模式 H 的次数为：

$$N_H^{t+1} = N_H^t \cdot N \cdot \frac{f_H^t}{\sum f_j^t} \tag{1}$$

式中，N 为群体的个数，f_H^t 为模式 H 代表的所有个体的平均适应度。

选择算子作用下的模式

利用公式$\overline{f^t} = \frac{1}{N} \cdot \sum_{f_j^t}$，公式（1）可以写成：

$$N_H^{t+1} = N_H^t \cdot \frac{f_H^t}{\overline{f^t}} \tag{2}$$

选择算子作用下的模式

公式（2）表明，当模式 H 的平均适应度大于群体平均适应度时，模式 H 的个数将增加，反之，模式 H 的个数将减少。

自然界生物遗传的优胜劣汰机制，在模式个数增长的关系中得到了充分的体现。

遗传算法不同于一般的随机搜索方法，群体进化是向优良基因模式和适应度高的个体逼近的过程。

选择算子作用下的模式

假设在第 t 代，群体中某一特定模式的平均适应度为 $f_H^t = (1+c)\overline{f^t}$，则在第 $t+1$ 代，它的个体数量将是 $N_H^{t+1} = (1+c) \cdot N_H^t$。假设上述适应度关系一直维持，那么随着群体的进化，经过 T 代之后，它的个体数量将是：

$$N_H^{t+T} = (1+c)^T \cdot N_H^t \tag{3}$$

显然，其数目将按照指数规律变化。

选择算子作用下的模式

许多不同的模式将按照上述规律相应地增加或减少：优良个体得到较多的选择和复制机会，劣质个体逐渐减少。

整个群体在此过程中没有出现新的个体。

选择与复制不会搜索新的相似点，即没有搜索问题空间的新区域，而交叉与变异是产生新个体或新搜索区域的主要遗传算子。

单点交叉算子作用下的模式

模式 H 只有当交叉点位于定义距之外才能生存，在单点交叉的情况下，H 遭破坏的概率为 $\frac{\delta(H)}{L-1}$，其中，L 是染色体编码的位数，那么 H 的生存概率为 $1-\frac{\delta(H)}{L-1}$。

当考虑交叉概率 p_c 时，上述 2 个量分别为 $p_c \cdot \frac{\delta(H)}{L-1}$ 和 $1-p_c \cdot \frac{\delta(H)}{L-1}$。

单点交叉算子作用下的模式

还要考虑到一个因素：即使交叉发生在定义距内，模式 H 也不一定被破坏，这是因为其配偶可能在相同的基因座上有相同的基因，因此，模式 H 遭破坏的概率 $\leq p_c \cdot \frac{\delta(H)}{L-1}$，那么模式 H 的生存概率 $\geq 1 - p_c \cdot \frac{\delta(H)}{L-1}$。

上式表明，具有短定义距的模式易生存。

选择与单点交叉共同作用下的模式

如果同时考虑选择与单点交叉算子，那么可以得到：

$$N_H^{t+1} \geq N_H^t \cdot \frac{f_H^t}{f^t} \cdot [1 - p_c \cdot \frac{\delta(H)}{L-1}] \tag{4}$$

上式表明，那些在种群平均适应度之上且定义距又短的模式将更易生存。

变异算子作用下的模式

变异操作以概率 p_m 随机地改变一个字符串编码位，每一位的存活概率是 $1-p_m$。

对于模式 H，其阶次为 $O(H)$，该模式的存活概率是 $(1-p_m)^{O(H)}$。一般情况下，变异概率 $p_m \ll 1$，那么有 $(1-p_m)^{O(H)} \approx 1-O(H)p_m$。

选择、单点交叉及变异共同作用下的模式

在三种算子的作用下，有：

$$N_H^{t+1} \geq N_H^t \cdot \frac{f_H^t}{\overline{f^t}} \cdot [1 - p_c \cdot \frac{\delta(H)}{L-1}][1 - O(H)p_m] \quad (5)$$

$$\approx N_H^t \cdot \frac{f_H^t}{\overline{f^t}} \cdot [1 - p_c \cdot \frac{\delta(H)}{L-1} - O(H)p_m] \quad (6)$$

模式定理

如果模式的定义距较短、阶次较低、适应度大于群体的平均适应度，那么随着群体的进化，该模式在群体中出现的次数将按指数规律变化。

把这类模式称为基因块或积木块（building block）

积木块假设

短定义距、低阶及高平均适应度的模式（积木块），在遗传操作的作用下，相互结合，能够产生长定义距、高阶及高平均适应度的模式，最终接近全局最优解。

参考文献

[1] 夏定纯，徐涛。《计算智能》，科学出版社，2008 年。
[2] Wikipedia: Genetic algorithm.
[3] Chapter 4: Genetic Algorithm.