# 进化算法的困难性理论研究进展

李 坤[1],黎 明[1,2],陈 昊[2]
(1.南京航空航天大学自动化学院,江苏南京 210016;2.南昌航空大学信息工程学院,江西南昌 330063)

**摘 要:** 进化算法困难性是进化计算研究领域的重要分支,旨在研究进化算法的性能表现与优化问题特性之间的联系,其目的是利用有限信息估计进化算法在求解优化问题时的性能表现.本文主要介绍进化算法困难性研究的几种典型方法及近年来的研究进展,主要包括适应值—距离模型、适应值曲面模型、曲面自动机模型、最优吸引子理论和基因关联模型等六种分析优化问题难度的理论,以及相应的八种难度指标.此外,本文还通过对比分析指出现有方法存在的优缺点,并展望了该领域未来的发展趋势.

**关键词:** 适应值曲面; 空间关联性; 曲面自动机; 最优吸引子理论; 基因关联测度; 优化问题难度
**中图分类号:** TP18 **文献标识码:** A **文章编号:** 0372-2112 (2014)02-0383-08
**电子学报 URL:** http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2014.02.026

# Research Progress of Hardness Theories on Evolutionary Algorithm

LI Kun[1], LI Ming[1,2], CHEN Hao[2]
(1. *College of Automation Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, Jiangsu* 210016, *China*;
2. *School of Information Engineering, Nanchang Hangkong University, Nanchang, Jiangxi* 330063, *China*)

**Abstract:** The research work of hardness on evolutionary algorithm is an important branch of evolutionary computation. It aims to study the relationship between the performance of evolutionary algorithm and the characters of the optimization problem. And, the goal of the hardness research is estimating the performance of an evolutionary algorithm deployed in one optimization problem by using limited information. This paper summarizes six kinds of hardness theories on evolutionary algorithm, such as fitness-distance correlation model, fitness landscape methods, landscape state machine, optimal contraction theorem, epistasis methods, etc. and eight hardness indicators with them. Furthermore, this paper discusses the advantages and disadvantages of these methods, and prospects the trends of this research field.

**Key words:** fitness landscape; spatial correlation; landscape state machine; epistasis measure; optimal contraction theorem; optimization hardness

## 1 引言

进化算法的困难性理论主要是研究进化算法的性能表现与优化问题特性之间的关系.近年来,分布估计算法[1,2]、文化基因算法[3,4]和粒子群算法[5,6]等进化算法的研究取得了丰硕的成果.但是,根据 NFL(No Free Launch)定理可知[7],能够很好的解决任何优化问题的"超级优化算法"不存在.研究进化算法困难性的目的是通过较小的计算代价获得拟解决优化问题的信息,进而以这些信息为依据调节进化算法的控制参数并有针对性的提高进化算法的性能.

进化算法的困难性是由多种因素造成的,其中主要包括欺骗问题、多峰问题和孤立点三个因素.目前的研究已经证实,欺骗问题和多峰问题都是优化问题困难的既非充分又非必要条件,只有孤立点是优化问题困难的充分条件[8].但是这种定性分析的结果不能作为调整控制参数的依据,而调整参数时需要的是一种弱判别准则[9].这种弱判别准则不仅要严格控制信息获取时的计算代价,更需要以能够解释算法与问题相互作用机制的理论为基础.

根据研究思路的不同可以将进化计算的困难性研究方法分为两类:第 I 类方法侧重分析问题的不同可行解之间的关系,本文第 2、3、4、5 节介绍第 I 类方法的研究成果.第 II 类方法侧重分析单个可行解的不同部分之间的相互关系,及其对适应值函数的影响.本文第 6、7 节介绍第 II 类方法.本文为了能更清晰的描述相关理

收稿日期:2012-12-06;修回日期:2013-04-17;责任编辑:孙瑶
基金项目:国家自然科学基金(No.61262019,No.61202112)

论在全文中通用如下符号:

$R$:可行解空间,即全部可行解组成的集合;

$|R|$:集合 $R$ 的规模,即可行解编码空间的容量;

$F$:适应值空间,即全部适应值组成的集合;

$f$:适应值函数,即 $R$ 到 $F$ 的映射;

$x$:可行解,集合 $R$ 中的元素;

$s$:基因位,二进制编码串上的基因位,$s\in\{0,1\}$;

$P$:可行解空间的随机样本集;

$i,j$:用于表示序号标记不同 $x$ 或 $s$ 的自然数.

## 2 适应值—距离关联模型

本节中介绍的方法都以欧式距离为基础,通过提取不同可行解间的距离与适应值的相关性信息构造进化算法困难性指标.

### 2.1 适应值距离关联测试法

适应值距离关联(Fitness Distance Correlation, FDC)测试法由 Jones 等[10]提出,Jones 等认为如果 $R$ 中的可行解与全局最优解的距离和它与全局最优解的适应值的差成正比则问题比较容易,反之则问题比较困难. FDC 方法通过测试样本集 $P$ 上的适应值与距离之间的相关系数描述进化算法的困难性.

**指标 1** 适应值距离相关系数 $fdc_p$

$$fdc_p(f)=\frac{\sum_{x\in p}(f(x)-\overline{f}_p)(d(x)-\overline{d}_p)}{\sqrt{\sum_{x\in p}(f(x)-\overline{f}_p)^2}\sqrt{\sum_{x\in p}(d(x)-\overline{d}_p)^2}} \tag{1}$$

其中,$d(x)$表示 $x$ 到最优解的距离,$\overline{f}_p$和$\overline{d}_p$分别表示样本集 $P$ 的适应值均值和所有样本到最优解的距离均值. $fdc_p\in[-1,1]$,最大(小)值问题的 $fdc_p$ 趋于 $-1(1)$ 时难度变小,但是,并非 $fdc_p$ 趋于 $1(-1)$ 时难度最大. 图 1 展示了 $fdc_p$ 的三种典型情况:①如图 1($a$)所示,当 $fdc_p$ 趋于 $-1$ 时,随着可行解与最优解(图 1 中"$*$")的距离增大,其适应值与最优解的差距也增大;②如图 1($b$)所示,当 $fdc_p$ 趋于 1 时,随着可行解与最优解的距离增大,其适应值与最优解的差距却减小;③如图 1($c$)所示,当 $fdc_p$ 趋于 0 时,在任何距离上都存在适应值较高的个体. 适应值与距离的相关性较弱,$fdc_p$ 指标的可靠性降低.

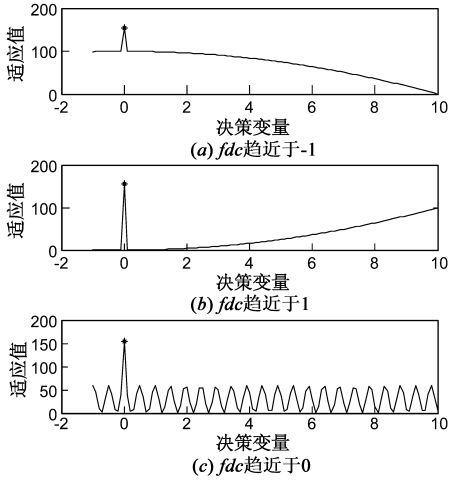



图1 FDC法的三种典型情况示意图

### 2.2 空间关联方法

空间关联(Spatial Correlation, SC)方法[11]认为随着两个可行解之间距离的增加,其适应值之间的相关性会减弱,衰减曲线可以揭示问题的特性. SC 方法计算一组距离 $\boldsymbol{d}$ 的空间关联系数 $R_s(\boldsymbol{d})$,再用 $R_s(\boldsymbol{d})$随距离增加而衰减的曲线描述问题的难度.

**指标 2** 空间关联系数 $R_s(d)$

$$R_s(d)=\frac{n\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij,d}(f(x_i)-\overline{f}_p)(f(x_j)-\overline{f}_p)}{\left(\sum_{i=1}^{n}w_{ij,d}\right)\sum_{i=1}^{n}(f(x_i)-\overline{f}_p)^2} \tag{2}$$

其中,$d$ 表示预先给定的距离,$n$ 表示样本集 $P$ 的规模,$x_i,x_j$ 表示相应的样本,$w_{ij,d}$表示权重函数(Weighting Function). 由于随机样本 $x_i,x_j$ 之间的距离 $d'$ 未必等于给定的距离 $d$,因此需要通过权重函数 $w_{ij,d}$确定不同的 $d'$对 $d$ 的影响.

### 2.3 小结与分析

本节中两种方法都是以适应值和距离之间的相关性为基础度量进化算法的困难性,FDC 法以全局最优解为中心测试这种相关性,SC 法则关注这种相关性在搜索域中的平均作用距离. 此外,部分文献将这种相关性称为问题结构(Problem Structure)[11,12]. 但是,问题结构影响算法性能的前提是算法行为的确定性,因此,问题结构无法体现进化算法中的随机因素的作用.

指标 1 的主要问题是对全局最优解的依赖,在大多数情况下只能用样本集 $P$ 中的最优解代替全局最优解,这样必然会造成误差. 指标 2 则不依赖于全局最优解.

## 3 适应值曲面模型

适应值曲面(Fitness Landscape, FL)的概念最初是 Wright 在研究生物学问题时提出的,Weinberger 等将适应值曲面的概念引入进化计算领域. Kallel、Stadler 等[13~15]分别做了大量工作以完善适应值曲面的理论体系,He 等[16,17]则利用适应值曲面分析进化算法的时间复杂度,Merz 等[18~20]将适应值曲面与文化基因算法结合起来研究. 当前,适应值曲面的定义存在多种版本,其中以李建武等人的版本比较全面,本文以此为主进行介绍.

### 3.1 适应值曲面的定义

一个适应值曲面 $L$ 可以定义为一个 5 元组[21]:

$$L=(R,\varphi,f,F,>_F) \tag{3}$$

其中,这 5 个元素的意义如下:

(1)$R$ 为可行解空间.

(2)$\varphi$ 为一个算子,定义为 $\varphi: R\times R\to[0,1]$.如果 $v,w\in R$,则 $p=\varphi(v,w)$ 意味着对位串 $v$ 运用一次 $\varphi$ 操作得到位串 $w$ 的概率.

(3)$f$ 为一个适应值函数,定义为 $f: R\to F$,对位串空间中的每个位串定义一个适应值.

(4)$F$ 为适应值空间.

(5)$>_F$ 为定义在 $F$ 上的一种偏序.对于位串 $v,w\in R$,如果 $f(v)>_F f(w)$,则称 $v$ 比 $w$“好”.

但是,也有部分文献认为适应值空间 $F$ 和偏序 $>_F$ 不是定义适应值曲面的基本条件[22].

### 3.2 关联长度测试法

Weinberger 认为适应值曲面的崎岖程度可以表示优化问题难度,因此提出关联长度(Correlation Length,CL)方法测试适应值曲面的崎岖程度[23].CL 方法先给定初值 $x_0$,再利用随机游走函数在适应值曲面上产生一个随机游走序列 $\{f(x_t)\}$,通过计算这个序列的自相关系数,判断问题的难易程度.

**指标 3**　随机游走序列 $\{f(x_t)\}$ 的自相关系数 $r(s)$

$$r(s)=\frac{\sum_{i=1}^{m-s}(f(x_i)-\bar{f})(f(x_{i+s})-\bar{f})}{\sigma(f)(m-s)} \tag{4}$$

其中,$m$ 表示时间序列 $\{f(x_t)\}$ 的总长度,$s$ 表示步长差,$\bar{f}$ 表示 $\{f(x_t)\}$ 的适应值均值,$\sigma(f)$ 表示 $\{f(x_t)\}$ 的适应值方差.$r(s)\in[0,1]$,$r(s)$ 趋于 1 时问题难度降低,$r(s)$ 趋于 0 时问题难度增大.

Hauschild 等[24] 认为邻接结构(Neighborhood Structure,NS)是决定适应值曲面能否准确预测进化算法困难性的重要因素.若能定义一种邻接结构描述 $R$ 中可行解的相互关系,使之尽可能接近进化算法在可行解之间跳转的概率,则适应值曲面必然能够更精准的预测进化算法的困难性水平.

### 3.3 小结与分析

适应值曲面模型与适应值—距离模型最主要的区别是由 $\varphi$ 确定的跳转概率而不是欧氏距离来度量 $R$ 中可行解的相互关系.适应值曲面模型试图描述进化算法与优化问题的相互作用过程,尤其是其中包含的随机性.因此,不可避免的增加了构造算子 $\varphi$ 的难度.

指标 3 与指标 1 和 2 的区别是随机游走序列可能不满足均匀随机采样的条件.因为序列中的点 $x_i$($x_0$ 除外)总与 $x_{i-1}$ 相关,而非集合 $R$ 中均匀随机采样.因此,指标 3 的准确程度依赖三个因素:时间序列的初值 $x_0$,随机游走函数和优化问题自身的特性.

## 4 曲面自动机模型

曲面自动机(Landscape State Machine,LSM)是 Corne 等人提出的一种分析进化算法困难性的数学工具[25,26].

### 4.1 曲面自动机

曲面自动机认为进化算法本身就可以确定 $R$ 中的可行解之间的跳转概率.若 $v,w\in R$,则 $v\to w$ 的概率被进化算法唯一确定,而且 $v$ 可能跳转到 $R$ 中的任何元素,故存在 $|R|$ 个 $v$ 跳转其它可行解的概率.若 $v\to w$ 和 $w\to v$ 的跳转概率不同,则 $R$ 中存在 $|R|^2$ 个跳转概率,记作 $|R|$ 阶方阵 $\boldsymbol{G}$.

曲面自动机是一种有限状态自动机,可以由状态 $S$ 和弧 $\boldsymbol{E}$ 表示,其中状态 $S$ 是集合 $R$ 的一个子集,$\boldsymbol{E}$ 则是 $\boldsymbol{G}$ 的一种抽象.状态 $S$ 的划分方法是:若 $F$ 仅包括有限个元素,则将适应值相同的可行解划入同一状态;若 $F$ 包含无穷多个元素,则将适应值函数的值域划分为若干区间,处在同一区间内的可行解划入同一状态.若以曲面自动机为基础衡量进化算法的困难性,则需要相应的指标描述曲面自动机.

**指标 4**　曲面自动机的参数 $\boldsymbol{S}^*$ 和 $\boldsymbol{E}$

$\boldsymbol{S}^*$ 表示状态 $S_i$ 占可行解空间 $R$ 的比例 $\boldsymbol{S}^*=\frac{|S_i|}{|R|}$;

弧矩阵 $\boldsymbol{E}$ 表示状态 $S_i$ 之间的跳转概率组成的矩阵.

### 4.2 弧 $\boldsymbol{A}$ 与跳转概率 $\boldsymbol{G}$ 的关系

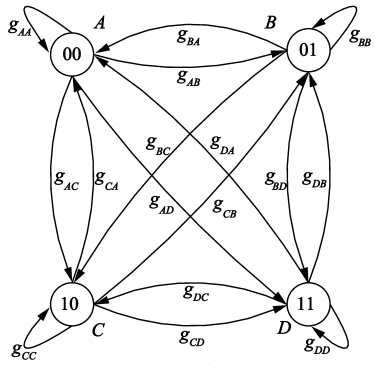
由于 $\boldsymbol{E}$ 与 $\boldsymbol{G}$ 之间的关系比较复杂,本节以 2 阶的 One-max 问题为例介绍它们的关系.图 2($a$)所示为一个 2 位的二进制串的可行解空间,可能存在的四种状态用 A B C D 表示,连接线表示相应跳转概率.此时,图 2($a$)所示跳转概率矩阵 $\boldsymbol{G}$ 可以表示为:

$$\boldsymbol{G}=\begin{bmatrix} g_{AA} & g_{AB} & g_{AC} & g_{AD}\\ g_{BA} & g_{BB} & g_{BC} & g_{BD}\\ g_{CA} & g_{CB} & g_{CC} & g_{CD}\\ g_{DA} & g_{DB} & g_{DC} & g_{DD}\end{bmatrix} \tag{5}$$

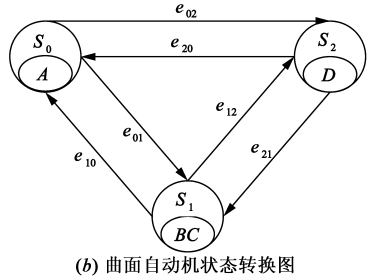
图 2($b$)表示图 2($a$)中所示的二进制串的 One-max 问题的曲面自动机.图 2($b$)中的状态 $S_0$ 包括图 2($a$)中的状态 $A$,$S_1$ 包括 $B$、$C$,$S_2$ 包括 $D$,则 One-max 问题的弧矩阵 $\boldsymbol{E}$ 可以由 $\boldsymbol{G}$ 中的元素表示.

$$\boldsymbol{E}=\begin{bmatrix} 0 & g_{AB}+g_{AC} & g_{AD}\\ \frac{1}{2}(g_{BA}+g_{CA}) & 0 & \frac{1}{2}(g_{BD}+g_{CD})\\ g_{DA} & g_{DB}+g_{DC} & 0\end{bmatrix} \tag{6}$$

图 2 所示仅为简单的 One-max 问题,如果问题规模增大则矩阵 $\boldsymbol{E}$ 和 $\boldsymbol{G}$ 的关系也会更复杂.而在实际问题中很难获得准确的 $\boldsymbol{G}$,所以也很难获得弧矩阵 $\boldsymbol{E}$.




(*a*) 可行解空间跳转概率$\boldsymbol{G}$示意图




(*b*) 曲面自动机状态转换图

**图2 弧矩阵$\boldsymbol{E}$与跳转概率矩阵$\boldsymbol{G}$的关系示意图**

### 4.3 小结与分析

适应值曲面模型试图通过算子 $\varphi$ 体现多种进化算法的共性,而曲面自动机模型中的 $\boldsymbol{G}$ 则只与特定的进化算法相关.但是,首先,曲面自动机也只能通过分析非均匀采样获取的样本集 $P$ 估计模型参数.其次,虽然获取 $\boldsymbol{G}$ 不像构造 $\varphi$ 那样困难,但是 $\boldsymbol{G}$ 也不像 $\varphi$ 那样能体现进化算法的共性.

指标 4 与指标 1、2 和 3 最主要的区别是其通过估计优化问题之间的相似程度来描述进化算法的困难性,而不是通过绝对数值表示进化算法的困难性.因此,Knowles 认为曲面自动机比较适合处理大量相似的优化问题[27,28].

## 5 最优吸引子理论

最优吸引子理论(Optimal Contraction Theorem, OCT)[29]以探索与利用平衡[30~33]的理论为基础,提供了一种用于分析进化算法收敛过程的理论.

### 5.1 最优吸引子理论

依据探索与利用平衡理论,进化算法包含探索和利用两种行为,然而它们并非是局部搜索与全局搜索[34].最优吸引子理论先给出两种行为的定义,再将进化算法的运行过程看作一个压缩或扩张的过程,并将这一过程分为不同阶段进行分析.进化算法收敛的过程就是在搜索域中不断缩小最优解存在区域的过程,最优吸引子理论称之为压缩.在压缩的过程中,算法可能收敛到局部最优解,此时需要重新扩大搜索的范围以避免早熟,最优吸引子理论称之为扩张.进化算法在压缩与扩张交替的过程中收敛于最优解,压缩的效率和收敛率部分取决于问题的特性.

### 5.2 优化特征因子

最优吸引子理论将影响压缩效率的优化问题特性归纳为优化特征因子(Optimal Feature Factor, OFF),并以之作为衡量进化算法困难性的指标.

**指标 5** 优化特征因子—OFF

OFF 由严格最优比(Strict Optimal Ratio, SOR)、最优比(Optimal Ratio, OR)和 $p$ 吸引比($p$-inductive Optimal Ratio, piOR)三个指标组成.SOR 和 OR 是指最优区域(Optimal Field, OF)和最大优化冠(Optimal Supreme Cap, OSC)占整个搜索域的比例.$p$ 吸引比则是一种依赖进化算法的动态特性.为了对比分析指标 5 和 4 的异同,有必要引述最优区域的定义.

最优区域:当且仅当搜索域 $B$ 的一个子空间 $D^*$ 满足以下两个条件时,称 $D^*$ 为最优区域:(1) $D^*$ 中的任意点 $x$ 满足条件 $f(x)>f(x^*)$, $x^*$ 为仅次于全局最优解的局部最优解;(2) $B$ 中满足不等式 $f(x)>f(x^*)$ 的任意点都被包含在 $D^*$ 之中.

图 3 展示了某一维函数的 OF 和 OSC.同时,为了便于分析,将图 3 所示函数的可行解集 $R$ 划分为 3 个子集:$S_0$ 由适应值高于 a 的可行解组成;$S_1$ 由适应值低于 a 且高于 b 的可行解组成;$S_2$ 由适应值低于 b 且高于 c 的可行解组成.SOR 和 OR 取值范围都是[0,1],取值越趋向 1 问题越简单,反之则越困难.

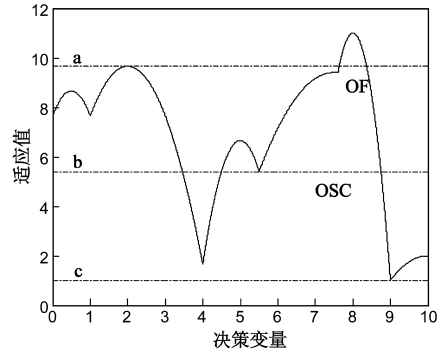



**图3 最优吸引域和最大优化冠示意图**

### 5.3 小结与分析

最优吸引子理论揭示优化算法与问题之间相互作用过程的共性,曲面自动机则关注特定进化算法与问题的相互作用.所以,最优吸引子理论的适用范围更广,包括但不局限于进化计算领域.

比较指标 5 和 4 不难发现,OF 和曲面自动机中的

状态 $S$ 都是 $R$ 的子集,区别是划分子集的方式不同.图 3 中划分的 $S_0$、$S_1$、$S_2$ 就可以看做是曲面自动机的三种状态.$S_2$ 到 $S_1$ 的跳转概率包含包括跳转到全局和局部最优解的概率,而且指标 4 无法分辨两者的大小,但是指标 5 却能够有效区分两者之间的差别.因此,指标 5 比指标 4 更精确.但是,目前仍然没有合适的方法能够测算或估计指标 5.

## 6　基因关联模型

基因关联模型关注二进制编码串中的基因位对适应值的影响以及它们之间的相互作用.

### 6.1　经典基因关联分析法

经典基因关联分析方法以模式定理和积木块假说为理论基础,认为基因位之间的相关性可以度量进化算法的困难性.Davidor[35] 提出的基因关联模型可以表示为:

$$f(s)=\text{常数}+\sum_{i=0}^{L-1}(\text{基因 } s_i \text{ 的影响})+\sum_{i=0}^{L-1}\sum_{j=i+1}^{L}(\text{基因 } s_i \text{ 和 } s_j \text{ 之间的相关性})+\cdots+(\text{基因 } s_1\cdots s_L \text{ 之间的相关性})+\text{随机误差}.$$

基因关联测试方法主要包括两个指标:基因关联方差(Epistasis Varinace, *epiv*)和基因关联相关系数(Epistasis Correlation, *epic*)两个指标.Naudts 等[8,15,36] 提出先在欧氏距离下标准化适应值函数,再求基因关联方差 *epiv*,这样就能确保 *epiv*∈[0,1].Chan 等[37] 研究浮点数编码方式的基因关联.Hashimoto, Deodhar 和 Turner 等[38~41] 研究基因关联以提高进化算法的性能.

**指标 6**　基因关联方差 *epiv* 和相关系数 *epic*

$$epiv_P(f)=\frac{\sqrt{\sum_{s\in P}(f(s)-\xi(s))^2}}{\sqrt{\sum_{t\in p}f^2(x)}} \tag{7}$$

其中,$\xi_P(s)=\sum_{i=1}^{l}(\overline{f_{P_i}}-\overline{f_P})+\overline{f_P}$,$\overline{f_P}$表示样本集 $P$ 中的平均适应值,$\overline{f_{p_i}}$表示第 $i$ 个基因位为 $s_i$ 的个体的平均适应值,$\bar{f}_{P_i}=\frac{1}{|\{t\in P;t_i=s_i\}|}\sum_{t\in P,t_i=s_i}f(x)$. $epiv\in[0,1]$, *epiv* 趋于 0 时问题较简单,反之则问题较困难.

$$epic_P(f)=\frac{\sum_{t\in P}(f(x)-\bar{f}_P)(\xi(x)-\bar{\xi}_P)}{\sqrt{\sum_{t\in P}(f(x)-\bar{f}_P)^2}\sqrt{\sum_{t\in P}(\xi(x)-\bar{\xi}_P)^2}} \tag{8}$$

其中,$\xi(t)=\sum_{i=1}^{l}(f(x)-\bar{f}_P)+\bar{f}_P$,$\bar{\xi}_P=\frac{1}{|P|}\sum_{t\in P}\xi(x)$. $epic\in[0,1]$, *epic* 趋于 1 时问题较简单,反之则问题较困难.

### 6.2　基于信息论的测试方法

Seo 等[42,43] 提出基于信息论的基因关联度量方法(Information Theoretic Epistasis Metrics, ITEM),这种方法将信源熵和互信息量的概念用于度量基因关联程度.Ventresca 等[44] 利用这种方法估计多目标问题的难度.若用 $s_i$ 表示第 $i$ 个二进制位,则 $s_i\in\{0,1\}$,根据定义可以计算 $s_i$ 的信源熵.但是,求取 $f(x)$的信源熵时,需要先离散化适应值函数的值域,再根据离散信源熵的定义计算 $f(x)$的信源熵.

**指标 7**　影响系数 $\xi_i$ 和相关系数 $\xi_{ij}$

基因位 $s_i$ 对适应值函数 $f$ 的影响 $\xi_i$:

$$\xi_i=\frac{I(s_i;f(x))}{H(f(x))} \tag{9}$$

其中,$I(s_i;f(s))$表示 $s_i$ 与 $f$ 的互信息量,$H(f(s))$表示 $f$ 的信源熵.$\xi_i\in[0,1]$,$\xi_i$ 越接近 0 则 $s_i$ 对 $f$ 影响越小,反之则越大.

基因位 $s_i$ 与 $s_j$ 的相关系数 $\varepsilon_{ij}$:

$$\varepsilon_{ij}=\begin{cases}1-\dfrac{I(s_i;f(x))+I(s_j;f(x))}{I(s_i,s_j;f(x))}, & I(s_i,s_j;f(x))\neq 0\\ 0, & \text{otherwise}\end{cases} \tag{10}$$

其中,$\varepsilon_{ij}\in[-1,1]$,$\varepsilon_{ij}$等于 0 时,$s_i$ 与 $s_j$ 相关性为零;$\varepsilon_{ij}$大于 0 时,$s_i$ 与 $s_j$ 各自与 $f$ 的相关性之和大于 $s_i$、$s_j$ 与 $f$ 的相关性;$\varepsilon_{ij}$小于 0 时,$s_i$ 与 $s_j$ 各自与 $f$ 的相关性之和小于 $s_i$、$s_j$ 与 $f$ 的相关性.总的来说,$\varepsilon_{ij}$趋于 0 时问题较简单,反之则较困难.

### 6.3　小节与分析

基因关联模型属于第 II 类方法,关注一个可行解编码串中的不同位的差异和相关性.由于其理论基础是模式定理和积木块假设,所以适用范围仅限于二进制编码的进化算法,而第 I 类方法的模型大多不受此类限制.

指标 6 和指标 7 都是从基因差异和基因关联性两个方面描述进化算法的困难性.但是,指标 7 不但能够描述基因位的相关性,还能体现相关性的符号.

## 7　决策变量相关性模型

Gibbs 等[11] 在基因关联模型的基础上,提出分析决策变量之间的差异和相关性的方法.

### 7.1　决策变量之间的不平衡性和相关性

Gibbs 等对决策变量不平衡性的定义与 Seo 相似,差别是分析的对象由基因位 $s_i$ 变为决策变量 $x_i$.但两者对相关性的定义不同,Gibbs 等认为如果适应值函数 $f$ 的任意决策变量 $x_i$ 和 $x_j$ 都满足条件:若令 $Y=f(x)$,$Y_{s,ij}=f(x_i+\Delta x_i,x_j+\Delta x_j)$,$Y_{ij}=f(x_i+\Delta x_i,x_j)+f(x_i,x_j+\Delta x_j)-f(x_i,x_j)$,就有 $Y_{s,ij}=Y_{ij}$成立,则称 $x_i$ 和 $x_j$ 完全可分离,完全可分离的问题较容易解决.

**指标 8** 不平衡性 $NI$ 和相关性 $\lambda_{ij}$

决策变量 $x_i$ 对适应值的影响可以用 $NI$ 表示：

$$NI(x_i, Y) = \frac{I(x_i; Y)}{H(Y)} \tag{11}$$

$I(x_i; Y)$表示 $x_i$ 与 $Y$ 的互信息量，$H(Y)$表示 $Y$ 的信源熵. $NI \in [0,1]$，$NI$ 趋于 1 时 $x_i$ 对适应值的影响较大，反之则较小.

任意决策变量 $x_i$ 和 $y_j$ 的关联程度可以用 $\lambda_{ij}$ 表示：

$$\lambda_{ij} = 1 - \frac{I(Y_{ij}; Y_{s,ij})}{H(Y_{ij})} \tag{12}$$

其中，$I(Y_{ij}; Y_{s,ij})$表示 $Y_{ij}$ 和 $Y_{s,ij}$ 的互信息量，$H(Y_{ij})$表示 $Y_{ij}$ 的信源熵. $\lambda_{ij} \in [0,1]$，$\lambda_{ij}$ 趋于 1 时 $x_i$ 和 $x_j$ 的相关性较高，反之则较低.

### 7.2 小结与分析

决策变量相关性模型的理论基础是：假设决策变量之间非线性不可分的关系是造成进化算法困难性的主要原因之一，然而基因关联模型却是以积木块假设为理论基础. 虽然指标 8 和指标 6、7 的原理与结构都相似，但是这种形式上的相似不能掩盖理论基础上的差异. 因此，指标 8 的适用范围不会受编码方式的限制.

## 8 总结与展望

本文共介绍了六种模型，这些模型可分为两类，第 II 类方法与以上介绍的第 I 类方法不同，关注单个可行解内部存在的差异和相关性. 模式适应值曲面模型是李建武等[45]为研究模式对进化算法困难性的影响而提出的，但是，这个模型能够诠释第 I 类和第 II 类方法的研究思路之间的关系.

进化算法的困难性研究脱胎于进化计算的理论研究，最初仅是研究进化算法原理的工具. 近年来，进化算法的困难性研究呈现出两种发展趋势. 一是精确模拟进化算法与问题相互作用的过程，从而有针对性的提高算法性能；二是归纳进化算法与问题交互过程的共性，从而提高进化计算理论研究的水平. 这两种趋势今后还将并存于进化算法的困难性研究领域.

### 参考文献


[1] Chen Tang, et al. Analysis of computational time of simple estimation of distribution algorithms [J]. IEEE Transactions on Evolutionary Computation, 2010, 14(1): 1 – 22.

[2] 何小娟，曾建潮，王丽芳. 一种基于信息传递的分布估计算法[J]. 电子学报，2011，39(4)：967 – 970.
He Xiao-juan, Zeng jian-chao, Wang Li-fang. An estimation of distribution algorithm based on information transmission [J]. Acta Electronica Sinica, 2011, 39(4): 967 – 970. (in Chinese)

[3] Handoko Kwoh, et al. Feasibility structure modeling: An effective chaperone for constrained memetic algorithms [J]. IEEE Transactions on Evolutionary Computation, 2010, 14(5): 740 – 758.

[4] Hrncic Mernik, et al. A memetic grammar inference algorithm for language learning [J]. Applied Soft Computing, 2012, 12(3): 1006 – 1020.

[5] 高芳，崔刚，等. 一种新型多步式位置可选择更新粒子群优化算法[J]. 电子学报，2009，37(3)：529 – 534.
Gao Fang, Cui Gang, et al. A novel multi-step position-selectable updating particle swarm optimization algorithm [J]. Acta Electronica Sinica, 2009, 37(3): 529 – 534. (in Chinese)

[6] 吴晓军，李峰，等. 均匀搜索粒子群算法的收敛性分析[J]. 电子学报，2012，40(6)：1115 – 1120.
Wu Xiao-jun, Li Feng, et al. The convergence analysis of the uniform search particle swarm optimization [J]. Acta Electronica Sinica, 2012, 40(6): 1115 – 1120. (in Chinese)

[7] Wolpert Macready. No free lunch theorems for optimization [J]. IEEE Transactions on Evolutionary Computation, 1997, 1(1): 67 – 82.

[8] Naudts Kallel. A comparison of predictive measures of problem difficulty in evolutionary algorithms [J]. IEEE Transactions on Evolutionary Computation, 2000, 4(1): 1 – 15.

[9] He Reeves, et al. A discussion on posterior and prior mesaures of problem difficulties [A]. In PPSN IX Workshop on Eovlutionary Algorithm-Bridging Theory and Practice [C]. USA: Oxford University Press, 2006. 1 – 13.

[10] Jones Forrest. Fitness distance correlation as a measure of problem difficulty for genetic algorithms [A]. Proceedings of the 6th International Conference on Genetic Algorithms [C]. San Mateo: Morgan Kaufmann, 1995. 184 – 192.

[11] Gibbs Maier, et al. Relationship between problem characteristics and the optimal number of genetic algorithm generations [J]. Engineering Optimization, 2011, 43(4): 349 – 376.

[12] Khor. Exploring the influence of problem structural characteristics on evolutionary algorithm performance [A]. IEEE Congress on Evolutionary Computation, Vols 1 – 5 [C]. Trondheim, NORWAY, 2009. 3345 – 3352.

[13] Stadler Institute. Towards a Theory of Landscapes Complex Systems and Binary Networks [M]. Berlin Heidelberg: Springer-Verlag, 1995. 78 – 163.

[14] Stadler. Landscapes and their correlation functions [J]. Journal of Mathematical Chemistry, 1996, 20(1): 1 – 45.

[15] Kallel Naudts, et al. Properties of Fitness Functions and Search Landscapes [M]. Berlin Heidelberg: Springer-Verlag, 2001. 175 – 206.

[16] He Yao. Towards an analytic framework for analysing the computation time of evolutionary algorithms [J]. Artificial Intelligence, 2003, 145(1 – 2): 59 – 97.

[17] He Yao, et al. To understand one-dimensional continuous fit-

ness landscapes by drift analysis[A]. Proceedings of the 2004 Congress on Evolutionary Computation, Vols 1 and 2[C]. USA:IEEE,2004.1248 – 1253.

[18] Merz Freisleben. Memetic Algorithms and the Fitness Landscape of the Graph Bi-Partitioning Problem[Z]. Berlin Heidelberg:Springer-Verlag,1998.765 – 774.

[19] Merz. Advanced fitness landscape analysis and the performance of memetic algorithms[J]. Evolutionary Computation, 2004,12(3):303 – 325.

[20] Merz Katayama. Memetic algorithms for the unconstrained binary quadratic programming problem[J]. Biosystems,2004,78 (1 – 3):99 – 118.

[21] 杨海军,李建武,李敏强.进化算法的模式、涌现与困难性研究[M].北京:科学出版社,2012.70 – 72.
Yang H J, Li J W, et al. Evlutionary Algorithm: Schema, Emergence and Hardnes[M]. Beijing: Science Press, 2012.70 – 72.(in Chinese)

[22] Borenstein Poli. Information landscapes[A]. Proceedings of the 2005 Conference on Genetic and Evolutionary Computation[C]. Washington DC, USA: ACM, 2005.1515 – 1522.

[23] Weinberger. Correlated and uncorrelated fitness landscapes and how to tell the difference[J]. Biological Cybernetics, 1990, 63 (5):325 – 336.

[24] Hauschild Pelikan. Advanced neighborhoods and problem difficulty measures[A]. Proceedings of the 13th Annual Conference on Genetic and Evolutionary Computation[C]. Dublin, Ireland: ACM, 2011.625 – 632.

[25] Corne Oates, et al. Landscape State Machines: Tools for Evolutionary Algorithm Performance Analyses and Landscape / Algorithm Mapping[J]. Lecture Notes in Computer Science (Applications of Evolutionary Computing), 2003:2611, 187 – 198.

[26] Rowe Corne, et al. Predicting stochastic search algorithm performance using Landscape State machines[A]. IEEE Congress on Evolutionary Computation, Vols 1 – 6[C]. USA: IEEE, 2006.2929 – 2936.

[27] Knowles. ParEGO: A hybrid algorithm with on-line landscape approximation for expensive multiobjective optimization problems[J]. IEEE Transactions on Evolutionary Computation, 2006,10(1):50 – 66.

[28] Knowles. Closed-loop evolutionary multi-objective optimization[J]. IEEE Computational Intelligence Magazine, 2009, 4 (3):77 – 91.

[29] Chen Xin, et al. Optimal contraction theorem for exploration-exploitation tradeoff in search and optimization[J]. IEEE Transactions on Systems Man and Cybernetics Part a-Systems and Humans, 2009, 39(3):680 – 691.

[30] Yen Yang, et al. Coordination of exploration and exploitation in a dynamic environment[A]. Ijcnn'01: International Joint Conference on Neural Networks, Vols 1 – 4, Proceedings[C]. Washington DC, 2001.1014 – 1018.

[31] Ishii Yoshida, et al. Control of exploitation exploration meta parameter in reinforcement learning[J]. Neural Networks, 2002, 15(4-6):665 – 687.

[32] Bocanet Ponsiglione. Balancing exploration and exploitation in complex environments[J]. Vine, 2012, 42(1):15 – 35.

[33] Yin Wang, et al. Free search with adaptive differential evolution exploitation and quantum-inspired exploration[J]. Journal of Network and Computer Applications, 2012, 35(3):1035 – 1051.

[34] Li Wang, et al. An effective PSO-based hybrid algorithm for multiobjective permutation flow shop scheduling[J]. IEEE Transactions on Systems Man and Cybernetics Part a-Systems and Humans, 2008, 38(4):818 – 831.

[35] Davidor P, et al. Epitasis variance: A viewpoint on GA-hardness[A]. Conference on Foundations of Genetic Algorithms [C]. San Mateo: Morgan Kauffman, 1991.23 – 35.

[36] Naudts Suys, et al. Epistasis as a basic concept in formal landscape analysis[A]. Proceedings of the 7th International Conference on Genetic Algorithms[C]. San Mateo: Morgan Kaufmann, 1997.65 – 72.

[37] Chan Aydin, et al. An epistasis measure based on the analysis of variance for the realcoded representation in genetic algorithms[A]. Proceedings of the Congress on Evolutionary Computation, Vols 1 – 4[C]. USA: IEEE, 2003.297 – 304.

[38] Hashimoto Warashina. Evolutionary computation using interaction among genetic evolution[J]. Lecture Notes in Computer Science(Trends in Artificial Intelligence), 2008, 5351: 152 – 163.

[39] Hashimoto Warashina, et al. New composite evolutionary computation algorithm using interactions among genetic evolution, individual learning and social learning[J]. Intelligent Data Analysis, 2010, 14(4):497 – 514.

[40] Deodhar Motsinger-Reif. Grammatical evolution decision trees for detecting gene-gene interactions[A]. Proceedings of Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics[C]. Berlin Heidelberg: Springer – Verlag, 2010.98 – 109.

[41] Turner Dudek, et al. ATHENA: A knowledge- based hybrid backpropagation-grammatical evolution neural network algorithm for discovering epistasis among quantitative trait Loci [J]. Bio Data Mining, 2010, 3(1):5.

[42] Seo Kim, et al. New entropy-based measures of gene significance and epistasis[J]. Lecture Notes in Computer Science (Genetic and Evolutionary Computation-GECCO), 2003, 2724:1345 – 1356.

[43] Seo Choi, et al. New epistasis measures for detecting independently optimizable partitions of variables[J]. Lecture Notes in

Computer Science (Genetic and Evolutionary Computation Part II),2004,3103:26 – 30.

[44] Ventresca Ombuki-Berman. Epistasis in multi-objective evolutionary recurrent neuro-controllers[A]. First IEEE Symposium on Artificial Life[C]. USA: IEEE, 2007. 77 – 84.

[45] 李建武. 遗传算法适应值曲面及遗传算法困难度分析[D]. 天津大学, 2003.
Li J W. Research on Fitness Landscapes of Genetic Algorithms and GA-hardness[D]. Tianjin University, 2003. (in Chinese)

## 作者简介

**李 坤** 男,1983 年 10 月出生,山东泰安人.2010 年毕业于南昌航空大学信息工程学院,同年进入南京航空航天大学自动化学院.现为博士研究生,主要研究方向为智能计算.

**黎 明(通讯作者)** 男,1965 年 2 月出生,江西樟树人.1985 年,1990 年和 1997 年分别在上海交通大学和南京航空航天大学获得学士和硕士、博士学位.现为南昌航空大学信息工程学院教授,南京航空航天大学博士生导师,主要研究方向为图像处理与模式识别、智能计算等.
E-mail: limingniat@hotmail.com