管理、决策与信息系统丛书

遗传算法的基本理论与应用

李敏强　寇纪淞　林丹　李书全　著

科学出版社
北京

内 容 简 介

本书旨在系统地介绍遗传算法的理论、应用和发展，共包括 9 个章节的内容。

首先，本书讲述了遗传算法的起源、历程和主要研究方向，介绍了遗传算法的基本原理。其次，讨论了遗传算法的一般收敛性理论，遗传算法的马尔可夫链模型和收敛性分析，遗传算法的随机泛函分析。还介绍了遗传算法的模式理论，并详细讨论了微观遗传策略、宏观遗传策略和遗传算法在知识获取中的应用，特别是概念学习和特征提取的遗传算法方法。讨论了遗传规划的原理、方法和收敛性分析，及其在典型问题中的应用。最后，介绍了遗传算法的发展——进化计算的原理与方法，讨论了 NFL 定理的意义，以及浮点实数编码的遗传算法在求解约束优化问题中的应用。附录中给出了一组典型的性能测试函数。

本书可以作为信息技术和管理科学专业的研究生教材，亦可供有关科研人员和工程技术人员阅读参考。

图书在版编目(CIP)数据

遗传算法的基本理论与应用/李敏强等著，—北京：科学出版社，2002.3
(管理、决策与信息系统丛书/汪寿阳，杨晓光主编)
ISBN 7-03-009960-5

Ⅰ.遗… Ⅱ.李… Ⅲ.①遗传-算法-基础理论 ②遗传-算法-应用
Ⅳ.TP18

中国版本图书馆 CIP 数据核字(2001)第 098477 号

责任编辑：陈 亮 / 责任校对：朱光光
责任印制：安春生 / 封面设计：王 浩

科学出版社发行 各地新华书店经销

*

2002 年 3 月第 一 版 开本：B5(720×1000)
2003 年 3 月第二次印刷 印张：27 1/4
印数：2 001—4 000 字数：506 000

定价：45.00 元

(如有印装质量问题，我社负责调换〈环伟〉)

前　言

20世纪40年代以来，科学家不断努力从生物学中寻求用于计算科学和人工系统的新思想、新方法。很多学者对关于从生物进化和遗传的机理中开发出适合于现实世界复杂适应系统研究的计算技术——自然进化系统的计算模型，以及模拟进化过程的算法进行了长期的开拓性的探索和研究，John H. Holland教授及其学生首先提出的遗传算法就是一个重要的发展方向。

从1989年起，我们开始接触到了遗传算法的一些研究成果。1994年，寇纪淞教授在美国访问研究期间，有幸到美国密歇根大学拜访了John H. Holland教授。Holland教授热情接待，并将最新出版的Adaptation in Natural and Artificial Systems第二版签名相赠。同时，介绍了他自己和多位知名的遗传算法专家的研究情况，使我们对遗传算法的研究和应用领域有了更深刻的认识。

按照生物学上可进化性的概念，遗传算法所追求的也是当前群体产生更好个体的能力，即遗传算法的可进化性或称群体可进化性。遗传算法的理论和方法研究也围绕着这一目标展开。比如，如何更好地模拟复杂系统的适应性过程和进化行为？在优化问题求解中怎样才能具备全局收敛性？算法的搜索效率如何评价？算法的设计与参数控制的理论基础是什么？遗传算法与其他算法的如何结合？等等。关于遗传算法第1个方面的研究主要是由Holland的团队和Santa Fe Institute的EVCA研究组等少数单位开展的。大部分专家和研究机构则主要集中于遗传算法作为优化问题通用求解算法的一系列问题，20世纪80年代以后产生了大量的研究成果。

与传统的启发式优化搜索算法相比，遗传算法的主要本质特征在于群体搜索策略和简单的遗传算子。群体搜索使遗传算法得以突破邻域搜索的限制，可以实现整个解空间上的分布式信息探索、采集和继承；遗传算子仅仅利用适应值度量作为运算指标进行染色体的随机操作，降低了一般启发式算法在搜索过程中对人机交互的依赖。这样就使得遗传算法获得了强大的全局最优解搜索能力，问题域的独立性，信息处理的隐并行性，应用的鲁棒性，操作的简明性，成为一种具有良好普适性和可规模化的优化方法。

本书旨在系统地介绍遗传算法的理论、应用和发展，包括了我们最近完成的一些研究成果。首先，概括性地介绍了遗传算法的起源、历程和主要研究方向，详细

描述了遗传算法的基本原理。其次,重点研究了遗传算法的随机理论与分析,遗传算法的模式理论与分析。探讨了微观遗传策略——遗传算子的分析与设计,宏观遗传策略——遗传算法结构分析与设计。然后,讨论了遗传算法在知识获取领域的应用,介绍了遗传规划的基本原理、方法和典型应用实例。最后,从发展的角度,介绍了进化算法的原理与方法,讨论了 NFL 定理的意义。

本书的研究工作得到了国家自然科学基金(79400013,69574022,69974026,70171002)、天津市自然科学基金(953601211)、教育部博士点研究基金(9405618)等的资助,有关成果系多年来课题组全体成员努力工作和精诚合作的结果。同时,David E. Goldberg、Kenneth De Jong、John Koza、Stepanie Forrest、Divad B. Fogel、Thomas Bäck、Zbigniew Michalewicz 等诸位教授和专家,以及 Erick Cantú－Paz 博士和 H.－G. Beyer 博士提供了大量资料,与 George L. Rudolph 博士的网络交流也对我们的研究工作提供了很多有益的帮助,谨在此表达诚挚的谢意!

本书由寇纪淞、李敏强统筹规划,由李敏强、林丹、李书全整理和校阅。其中,第 1、2、4、5、6、7 章由李敏强撰写,第 3 章第 1 节、第 9 章由林丹撰写,第 3 章第 6 节、第 8 章第 1～3 节由李书全撰写,第 3 章第 2～5 节、第 8 章第 4 节由戴晓晖撰写,第 8 章第 5 节由李建武撰写。参加本书研究工作的还有马丰宁博士、宋庆伟博士、田军博士等。对于他们的辛勤工作和贡献,表示衷心的感谢。

在本书的研究和整理工作中,得到汪寿阳先生多方面的指导和鼓励,并有幸列入本套丛书,同时,本书的编辑出版也得到杨晓光先生的直接帮助,在此一并表示感谢。

遗传算法是一个处于快速发展中的科学分支,其理论和应用均还有大量问题尚待进一步深入研究。由于作者学识水平和可获得资料的限制,本书不妥之处在所难免,敬请同行专家和诸位读者批评指正。

著 者

2002 年 3 月于天津大学

(C－0067.0102)

(C－0067.0102)

ISBN 7－03－009960－5

ISBN 7－03－009960－5

目　　录

第一章　概　　述

辩证唯物主义认为世界是物质的，物质是运动的，运动是有规律的，而规律是可以被认识和掌握的。通过认识和掌握事物的规律，可以较好地解决人们面临的各种问题。人类在认识和改造自然的过程中，创造了数学、物理、化学、生物等学科，从不同侧面研究自然界，并取得了显著效果。由于客观事物的复杂性以及人们对事物认识的不断深化，人们发现仅靠某一门学科很难比较完美地解决实际工作中遇到的问题，因而出现了交叉学科。

20 世纪 40 年代以来，科学家不断努力从生物学中寻求用于计算科学和人工系统的新思想、新方法，比如早期的自动机理论就试图采用类似神经元的元素建造一种新型的思维机器。很多学者对关于从生物进化和遗传的机理中发展出适合于现实世界复杂适应系统(complex adaptive systems)研究的计算技术——自然进化系统的计算模型(computational models of natural evolutionary systems)和模拟进化算法(simulated evolutionary algorithm)等，进行了开拓性的长期探索和研究[1～7]，Holland 及其学生首先提出的遗传算法(genetic algorithms，简称 GAs 或 GA)就是一个重要的发展方向[3～7]。

1.1　从生物进化到进化计算

按照达尔文(C. Darwin)的进化论[8,9]，地球上的每一物种从诞生开始就进入了漫长的进化历程。生物种群从低级、简单的类型逐渐发展成为高级、复杂的类型。各种生物要生存下去就必须进行生存斗争，包括同一种群内部的斗争、不同种群之间的斗争，以及生物与自然界无机环境之间的斗争。具有较强生存能力的生物个体容易存活下来，并有较多的机会产生后代；具有较低生存能力的个体则被淘汰，或者产生后代的机会越来越少，直至消亡。达尔文把这一过程和现象叫做“自然选择，适者生存”。

生物学家对自然界生物的进化机理进行了系统的研究，在如此“短暂”的时间里，生物界由最简单的单细胞生物，发展到现在的纷繁复杂、种群繁多的高级生物，充分证明了自然界的“自然选择，适者生存”的思想的实际意义，与之有关的关于生物进化的研究结论，已得到广泛的接受和应用。

按照孟德尔和摩根(G. Mendel,T. Morgan)的遗传学理论[8,9],遗传物质是作为一种指令密码封装在每个细胞中,并以基因的形式排列在染色体上,每个基因有特殊的位置并控制生物的某些特性。不同的基因组合产生的个体对环境的适应性不一样,通过基因杂交和突变可以产生对环境适应性强的后代。经过优胜劣汰的自然选择,适应值高的基因结构就得以保存下来,从而逐渐形成了经典的遗传学染色体理论,揭示了遗传和变异的基本规律。现代遗传学则对基因的本质、功能、结构、突变和调控进行了深入探讨,开辟了遗传工程研究的新领域。在一定的环境影响下,生物物种通过自然选择、基因交换和变异等过程进行繁殖生长,构成了生物的整个进化过程。

遗传物质是细胞核中染色体上的有效基因[8,9],其中包含了大量的遗传信息。染色体上携带着关于生物性状的物质元素,生物体所表现出来的外在特征是对其染色体构成的一种体现。生物的进化本质体现在染色体的改变和改进上,生物体自身形态的变化是染色体结构变化的表现形式。

基因组合的特异性决定了生物体的多样性,基因结构的稳定性保证了生物物种的稳定性,而基因的杂交和变异使生物进化成为可能。生物的遗传是通过父代向子代传递基因来实现的,而这种遗传信息的改变决定了生物体的变异。

生物进化过程的发生需要四个基本条件:①存在由多个生物个体组成的种群;②生物个体之间存在着差异,或群体具有多样性;③生物能够自我繁殖;④不同个体具有不同的环境生存能力,具有优良基因结构的个体繁殖能力强,反之则弱。

生物群体的进化机制可以分为三种基本形式[1~10]:

(1)自然选择:控制生物体群体行为的发展方向,能够适应环境变化的生物个体具有更高的生存能力,使得它们在种群中的数量不断增加,同时该生物个体所具有的染色体性状特征在自然选择中得以保留。

(2)杂交:通过杂交随机组合来自父代染色体上的遗传物质,产生不同于它们父代的染色体。生物进化过程不需要记忆,它所产生的能很好适应自然环境的信息都包含在当前生物体所携带的染色体的基因库(gene pool)中,并且可以很好地由子代个体继承下来。

(3)突变:随机改变父代个体的染色体上的基因结构,产生具有新染色体的子代个体。变异是一种不可逆过程,具有突发性、间断性和不可预测性,对于保证群体的多样性具有不可替代的作用。

另外,生物进化是一个开放的过程,自然界对进化中的生物群体提供及时的反馈信息,或称为外界对生物的评价。评价反映了生物的生存价值和机会。在基于相同环境下的生存竞争中,生存价值低的个体被淘汰了,生存下来的个体则具有较高的生存价值。由此形成了生物进化的外部动力机制。

大多数高级生物体是以自然选择和有性生殖这两种基本过程实现进化发展的。自然选择决定了生物群体中哪些个体能够存活并繁殖,有性生殖保证了生物体后代基因中的杂交和重组,从而使得群体的进化比其他方式更加快速而有效。

自然界的生物进化是一个不断循环的过程。在这一过程中,生物群体也就不断地完善和发展。可见,生物进化过程本质上是一种优化过程,在计算科学上具有直接的借鉴意义[1,10]。在计算机技术迅猛发展的时代,生物进化过程不仅可以在计算机上模拟实现[1~3],而且还可以模拟进化过程,创立新的优化计算方法,并应用到复杂工程领域之中,这就是 GA 等一类模拟自然进化的计算方法的思想源泉[1~7,11]。以生物进化过程为基础,计算科学学者提出了各种模拟形式的计算方法。

尽管到目前为止进化计算的发展不过 20 余年,但其思想可以追踪到 20 世纪 50 年代[1~3,11~14]。一般认为,进化计算(evolutionary computation, EC)包括三个组成部分[12~14]:①由美国密歇根大学 John H. Holland 教授提出的遗传算法[1,3],②由美国科学家 Lawrence J. Fogel 等人提出的进化规划(evolutionary programming, 简称 EP)[2,13],③由德国科学家 Ingo Rechenberg 和 Hans-Paul Schwefel 提出的进化策略(evolution strategies,简称 ES 或 ESs)[13,15]。他们用不同的进化控制模式模拟了生物进化过程,从而形成了三种具有普遍影响的模拟进化的优化计算方法。这三种方法也统称为进化算法(evolutionary algorithms, 简称 EA 或 EAs)。

进化算法是一种基于自然选择和遗传变异等生物进化机制的全局性概率搜索算法。与基于导数的解析方法和其他启发式搜索方法(如爬山方法,模拟退火方法,Monte Carlo 方法)一样,进化算法在形式上也是一种迭代方法。它从选定的初始解出发,通过不断迭代逐步改进当前解,直至最后搜索到最优解或满意解。在进化计算中,迭代计算过程采用了模拟生物体的进化机制,从一组解(群体)出发,采用类似于自然选择和有性繁殖的方式,在继承原有优良基因的基础上,生成具有更好性能指标的下一代解的群体。

优化问题采用进化计算求解的一般过程包括以下步骤:

1) 随机给定一组初始解;

2) 评价当前这组解的性能;

3) 根据 2)的评价结果,从当前解中选择一定数量的解作为基因操作的对象;

4) 对所选择的解进行基因操作(杂交或称为交叉、突变或称为变异),得到一组新的解;

5) 返回到 2),对该组新的解进行评价;

6) 若当前解满足要求或进化过程达到一定的代数,计算结束,否则转向 3)继

续进行。

进化算法是一种随机化搜索方法,在初始解生成以及选择、交叉与变异等遗传操作过程中,均采用了随机处理方法。与其他搜索技术(如梯度搜索技术、随机搜索技术、启发式搜索技术和枚举技术等)相比,进化算法具有以下特点[4,7,13,14]:

1)进化算法的搜索过程是从一群初始点开始搜索,而不是从单一的初始点开始搜索,这种机制意味着搜索过程可以有效地跳出局部极值点。特别是当采用有效的保证群体多样性的措施时,进化算法可以很好地将局部搜索和全局搜索协调起来,既可以完成极值点邻域内解的求精,也可以在整个问题空间实施探索,得到问题全局最优解的概率大大提高了。

2)进化算法在搜索过程中使用的是基于目标函数值的评价信息,而不是传统方法主要采用的目标函数的导数信息或待求解问题领域内知识。进化算法的这一特点使其成为具有良好普适性和可规模化(scalability)的优化方法。

3)进化算法具有显著的隐式并行性(implicit parallelism)。进化算法虽然在每一代只对有限解个体进行操作,但处理的信息量为群体规模的高次方。

4)进化算法在形式上简单明了,不仅便于与其他方法相结合,而且非常适合于大规模并行计算机运算,因此可以有效地用于解决复杂的适应性系统模拟和优化问题。

5) 进化算法具有很强的鲁棒性(robustness),即在存在噪声的情况下,对同一问题的进化算法的多次求解中得到的结果是相似的。进化算法的鲁棒性在大量的应用实例中得到了充分的验证。

1.2 遗传算法的特征与发展

Holland 的早期工作主要集中于生物学、社会学、控制工程、人工智能等领域中的一类动态系统的适应性问题(adaptation),其中适应性概念描述了在环境中表现出较好行为和性能的系统结构的渐进改变过程,简称系统的适应过程[1,2,7]。Holland 认为,通过简单的模拟机制可以描述复杂的适应性现象。因此,Holland 试图建立适应过程的一般描述模型,并在计算机上开展模拟试验研究,分析自然系统或者人工系统对环境变化的适应性现象,其中遗传算法仅仅是一种具体的算法形式。

Bremermann, De Jong 等人则注重将遗传算法应用于参数优化问题[4,16],极大地促进了遗传算法的应用。所以,遗传算法既是一种自然进化系统的计算模型,也是一种通用的(general purpose)求解优化问题的适应性搜索方法。目前,人们最关注和普遍使用的遗传算法是其后一种性质。

从整体上来讲,遗传算法是进化算法中产生最早、影响最大、应用也比较广泛

的一个研究方向和领域，它不仅包含了进化算法的基本形式和全部优点，同时还具备若干独特的性能[3~7,12,16,17]：

1)在求解问题时，遗传算法首先要选择编码方式，它直接处理的对象是参数的编码集而不是问题参数本身，搜索过程既不受优化函数连续性的约束，也没有优化函数导数必须存在的要求。通过优良染色体基因的重组，遗传算法可以有效地处理传统上非常复杂的优化函数求解问题。

2)若遗传算法在每一代对群体规模为 n 的个体进行操作，实际上处理了大约 $O(n^3)$ 个模式，具有很高的并行性，因而具有显著的搜索效率。

3)在所求解问题为非连续、多峰以及有噪声的情况下，能够以很大的概率收敛到最优解或满意解，因而具有较好的全局最优解求解能力。

4)对函数的性态无要求，针对某一问题的遗传算法经简单修改即可适应于其他问题，或者加入特定问题的领域知识，或者与已有算法相结合，能够较好地解决一类复杂问题，因而具有较好的普适性和易扩充性。

5)遗传算法的基本思想简单，运行方式和实现步骤规范，便于具体使用。

鉴于遗传算法具有上述特征，一经提出即在理论上引起了高度重视，并在实际工程技术和经济管理领域得到了广泛地应用，产生了大量的成功案例。

1962 年，John Holland 在 Outline for a Logic Theory of Adaptive Systems 一文中[1]，提出了所谓监控程序(supervisory programs)的概念，即利用群体进化模拟适应性系统的思想。他注意到在建立智能机器的研究中，不仅可以完成单个生物体的适应性改进，而且通过一个种群的许多代的进化也可以取得非常好的适应性效果。为了获得一个好的学习方法，仅靠单个策略的改进是不够的，采用多策略的群体繁殖往往能产生显著的学习效果。尽管他当时没有给出实现这些思想的具体技术，但却引进了群体、适应值、选择、变异、交叉等基本概念。1966 年，Fogel[2] 等人也提出了类似的思想，但其重点是放在变异算子而不是采用交叉算子。1967 年，Holland 的学生 J.D. Bagley 通过对跳棋游戏参数的研究，在其博士论文中首次提出“遗传算法”一词[4,204]。

Holland 以二进制字符集{0,1}构成的代码串表示实际问题的描述结构或参数，称为“染色体”(chromosome)。对这些“染色体”进行变换，利用“染色体”中所包含的信息决定新一代“染色体”，并最终得到问题的解。这种方法对所要解决的问题类型几乎没有任何限制，所需要的信息只是每个染色体的评价值。这种使用简单编码和选择机制的算法能够解决相当复杂的问题，并且解决实际问题时不需要该领域的专门知识。通过对这些简单的染色体进行迭代处理，从这些染色体中发现并保存好的染色体，进而逐步发现问题的最优解，这些思想就是遗传算法理论的雏形。

同时，Fraser 采用计算机模拟自然遗传系统，1962 年提出了和现在的遗传算法十分相似的概念与思想[18]。但是，Fraser 和其他一些学者并未认识到自然遗传方法可以转化为人工遗传算法。

在 20 世纪 60 年代中期至 70 年代末期，基于自然进化的思想遭到怀疑和反对。Holland 及其数位博士坚持了这一方向的研究。在 Holland 发表论文后的十余年中，从事遗传算法研究的论文开始慢慢出现。大多数研究都集中在美国 Michigan 大学的 Holland 及其学生当中。因此，遗传算法大多数著名学者都曾经是 Michigan 大学的学生，如：David E. Goldberg、Kenneth A. De Jong、John R. Koza、Stepanie Forrest 等。1975 年，Holland 出版了专著《自然与人工系统中的适应性行为》(Adaptation in Natural and Artificial Systems)[3]，该书系统地阐述了遗传算法的基本理论和方法，提出了对遗传算法的理论发展极为重要的模式理论(schema theory)，其中首次确认了选择、交叉和变异等遗传算子，以及遗传算法的隐并行性，并将遗传算法应用于适应性系统模拟、函数优化、机器学习、自动控制等领域。

另外，Daniel J. Cavicchio 的博士论文中探讨了一组实验[11]，将基于整数编码的遗传算法应用于模式识别问题，研究了保持群体差异性的选择策略。De Jong 在其博士论文研究中首次把遗传算法用于函数优化问题[16]，并对遗传算法的机理与参数设计问题进行了较为系统地研究。De Jong 深入全面地研究了模式定理和遗传算子的行为，将其与自己大量实验工作相结合，建立了著名的五函数测试平台。通过实验，他给出如下结论：①初始群体容量越大，离线性能越好，但在线性能的初始值较差。②变异可以降低某些基因的丢失机会，提高变异概率能避免成熟前收敛，但却降低在线性能。③交叉概率越大，群体中新结构的产生越快，当交叉概率等于 0.6 时，在线性能与离线性能都较好。

1975 年之后，遗传算法作为函数优化器(function optimizers)不但在各个领域得到广泛应用，而且还丰富和发展了若干遗传算法的基本理论。1980 年，Bethke 对函数优化 GA 进行了研究[19]，包括应用研究和数学分析。Smith 在 1980 年首次提出使用变长位串的概念[20]，这在某种程度上为以后的遗传规划奠定了基础。Goldberg[4]、Davis[21]、Grefenstette[22]、Bauer[23]、Srinivas 和 Patnaik[24]等大批研究人员对遗传算法理论的基本框架和遗传算子进行了构建和改进，并将遗传算法分别应用于工程设计、自动控制、经济金融、博奕问题、机器学习等诸多领域之中。

1989 年，David Goldberg 出版了 Genetic Algorithms in Search, Optimization and Machine Learning 一书[4]，这是第一本遗传算法教科书，它是对当时关于遗传算法领域研究工作的全面而系统的总结，因而也成为引用最多的参考书之一。与 Holland 的著作侧重于适应性系统的进化数学分析不同，本书将遗传算法的基本原

理与范围广泛的应用实例相结合,并给出了大量可以使用的应用程序。1991 年,Davis 编辑出版了 Handbook of Genetic Algorithms[21],其中包括了 GA 在工程技术和社会生活中的大量应用实例。

John R. Koza 将遗传算法用于处理不定长树形字符串或一组程序,提出了遗传规划(genetic programming,简称 GP)的概念[25,26]。树状表示方法是 Koza 教授于 1989 年首次提出的,这种表示方法的主要特点之一就是染色体结构是动态变化的层次结构,它受环境影响而改变,因而对问题的表示更加自然。该方法是一种与领域无关的自适应搜索解空间的有效算法。通过增加染色体结构的复杂性,它拓广了传统遗传算法的应用范围。Koza 教授认为不同领域中许多看起来不相同的问题都可看成是寻找一定的计算机程序问题,即许多不同领域的问题都可形式化为程序归纳问题,而遗传规划提供了实现程序归纳的方法,如公式、规划(plan)、控制策略、计算程序、模型(Model)、决策树、对策策略(game-playing strategy)、转换函数、数学表达式等都称之为计算机程序。1992 年,Koza 教授出版了第一本遗传规划专著 Genetic Programming 1[25],两年之后又出版了第二本关于遗传规划的专著[26]。Koza 教授虽然尚未建立遗传规划的完整理论体系,但他通过大量的实验说明了遗传规划能够成功地解决一类复杂问题,为基于符号表示的函数学习问题增添了一个强有力的工具。

随着遗传算法研究和应用的不断深入与扩展,1985 年,在美国召开了第一届遗传算法国际会议,即 ICGA(International Conference on Genetic Algorithms)。这次会议是遗传算法发展的重要里程碑,此会以后每隔一年举行一次。从 1999 年起,ICGA 和 GP(Genetic Programming Society)的系列会议合并为每年一次的遗传和进化国际会议(Genetic and Evolutionary Computation Conference, GECCO)。

在欧洲,从 1990 年开始也每隔一年举办一次类似的会议,即 PPSN(Parallel Problem Solving from Nature)会议。以遗传算法理论基础为中心的学术会议 FOGA(Foundation of Genetic Algorithms)也从 1990 年起每隔一年举办一次。

1994 年 1 月,IEEE 神经网络委员会(IEEE Neural Network Council)出版了第一本"进化计算"专集;1994 年 6 月,IEEE 神经网络委员会召开第一届"进化计算"国际学术会议(IEEE ICEC),以后每年召开一次。1997 年,该委员会创办了 IEEE Transactions on Evolutionary Computation 杂志(http://www.ewh.ieee.org/tc/nnc/pubs/tec/),David B. Fogel 任主编。从 1999 年开始,IEEE ICEC 与 EP 的年会合并为进化计算国际会议(Congress on Evolutionary Computation,CEC),每年召开一次。

美国 MIT 出版社从 1993 年开始出版 Evolutionary Computation(http://mitpress.mit.edu/journal-home,De Jong 主编)和 Adaptive Behavior 杂志。世界上第一

本关于人工智能研究的杂志 AI Trends 于 1993 年更名为 Critical Technology Trends,并在更名启事中讲到——“遗传算法、自适应系统、细胞自动机、混沌理论和人工智能一样,都是对今后十年的计算机技术有重大影响的关键技术”。

随着 Internet 技术的发展和普及应用,遗传算法的有关研究单位建立了刁量的专题 GA 网站,其中最为著名的是由美国海军人工智能应用研究中心建立的 GA _ Archives检索网站 (http://www.aic.nrl.navy.mil/galist/),它包括了世界范围内的开展遗传算法和进化计算研究的大学和机构,历年来的可公开发表的论文和报告,有关国际会议消息,典型应用案例和程序(源代码),等等。

这些众多的研究单位和频繁的国际学术活动集中反应了遗传算法的学术意义和应用价值。目前,遗传算法已成为一个多学科、多领域的重要研究方向。

1.3 遗传算法理论研究

与传统的启发式优化搜索算法(爬山方法、模拟退火法、Monte Carlo 方法等)相比,遗传算法(以及广义上的进化算法)的主要本质特征在于群体搜索策略和简单的遗传算子。群体搜索使遗传算法得以突破邻域搜索的限制,可以实现整个解空间上的分布式信息采集和探索;遗传算子仅仅利用适应值度量作为运算指标进行随机操作,降低了一般启发式算法在搜索过程中对人机交互的依赖。

按照生物学上可进化性(evolvability)的概念,遗传算法所追求的也是当前群体产生比现有个体更好个体的能力,即遗传算法的可进化性或称群体可进化性。因此,遗传算法的理论和方法研究也就围绕着这一目标展开。关于下面五个问题的回答,就成为 GA 理论研究的主要方向:

1) 遗传算法如何更好地模拟复杂系统的适应性过程和进化行为?

2) 遗传算法在优化问题求解中怎样才能具备全局收敛性?

3) 遗传算法的搜索效率如何评价?

4) 遗传策略的设计与参数控制的理论基础是什么?

5) 遗传算法与其他算法的如何结合?

其中,遗传策略包括 GA 流程设计、群体设定、群体初始化、GA 算子、终止条件等,广义上的遗传策略还包括遗传算法与其他算法结合形成的混合算法。

需要指出的是,关于遗传算法第 1)个问题的研究主要是 Holland 的团队和 Santa Fe Institute 的 EVCA 研究组(Evolving Cellular Automata, http://www.santafe.edu/projects/evca/)等少数单位。大部分研究机构的研究主要集中于遗传算法和进化算法作为优化问题通用求解算法的一系列问题。

1.3.1 遗传算法的基础理论研究

在优化理论中,采用迭代算法求解一个特定问题,若该算法的搜索过程所产生的解或函数的序列的极限值是该问题的全局最优解,则该算法是收敛的。

遗传算法的基础理论主要以收敛性分析为主,即群体收敛到优化问题的全局最优解的概率。从整体上讲,可以分为基于随机过程的收敛性研究和基于模式理论的收敛性分析,我们将前者称为遗传算法的随机模型理论(stochastic modeling),后者称为遗传算法的进化动力学理论(evolution dynamics)。

1. 随机模型理论

对于有限的编码空间和有限的群体,遗传算法的搜索过程可以表示为离散时间的马尔可夫链模型(Markov chain model),从而可以采用已有的随机过程理论进行严密分析。遗传算法满足有限马尔可夫链(finite Markov chain)的基本特征,具有齐次性,存在极限概率分布。由于编码空间的有限性,标准遗传算法可以搜索到空间上的任何一个点。在采用精英保留策略下,遗传算法可以以概率 1 收敛于问题的全局最优解。

1987 年,Goldberg 和 Segrest[58]运用有限马尔可夫链理论对遗传算法进行了收敛性分析,Eiben 等人证明了一类抽象遗传算法在 Elitist 选择情况下的概率收敛情况[59],Rudolph 用齐次有限马尔可夫链证明了带有选择、交叉和变异操作的标准遗传算法收敛不到全局最优解,但是如果让每一代群体中的最佳个体不参加交叉与变异操作而直接保留到子代,那么遗传算法是收敛的[60]。Qi 和 Palmieri 对浮点数编码的遗传算法,在基于连续空间中群体规模为无穷大这一假设下进行了严密的数学分析[61]。Fogel 和 Suzuki 从进化计算的角度对 GA 收敛问题进行了研究[63~65]。李书全等采用泛函分析理论证明了 GA 的收敛性[66]。

Vose, Nix, Liepins 等采用统计动力学方法,分析了无穷群体下 GA 的搜索轨迹和不动点收敛性[7]。Cerf 等采用摄动理论和马尔可夫链,对遗传算法的渐近收敛特性进行分析,得出了遗传算法运行及全局收敛性的一般性结论[62]。

随机收敛性分析主要是在群体无穷大和进化代数趋于无限的条件之上,研究遗传算法的极限行为。事实上遗传算法的计算复杂度问题是实际应用中更为关心的问题,Bäck[118]和 Mühlenbein[176]等研究了到达全局最优解的遗传算法的时间复杂性问题,恽为民[57]等基于马尔可夫链对此进行了进一步的分析。

但是,正是由于遗传算法的搜索过程的统计抽象描述,使得随机模型的收敛性分析远离了遗传算法的设计与应用。随机模型理论不仅适用于遗传算法,也适用于进化策略、进化规划、遗传规划,以及其他类随机化搜索算法。因此,尽管随机模

型理论的研究成果非常丰富,对于遗传算法(以及广义上的进化算法)的具体应用和参数设计所提供的指导信息却非常之少。

2. 进化动力学理论

对于任何函数优化问题,我们期望搜索过程所产生的解的序列收敛于问题的全局最优解,其中维持群体可进化性就成为遗传算法的核心任务。在随机模型理论下具备全局收敛性的任何算法,必须以某种具体运算形式应用于优化问题的求解,然而随机模型理论下的全局收敛性,并不能保证任何运算形式的算法在有限群体和有限进化代数下一定能够搜索到问题的全局最优解。为此,对于基于某种运算形式的遗传算法的进化行为分析就构成了进化动力学理论的基本内容[3,5,75,76]。

由 Holland 提出的模式定理可以称为遗传算法进化动力学的基本定理,但是模式定理仅仅描述了模式的生存模型,没有反映模式的重组过程,所以有限群体下的模式定理不保证遗传搜索的全局收敛性。进一步,建筑模块假说描述了遗传算子的重组功能,要求定义长度短的、低阶的、适应值高的模式(建筑模块),在遗传算子的作用下,被采样、重组,从而形成高阶、长距、高于群体平均适应值的模式。

模式定理和建筑模块假说构成了求解优化问题时遗传算法具备发现全局最优解的充分条件,也是分析遗传算法的进化行为的基本理论,统称为模式理论。尽管还存在着不完善之处,但是它为深入研究遗传算法的运行机理奠定了基础。

很多实际问题采用 GA 求解并不完全满足建筑模块假说,存在着各种程度的模式欺骗性,那么遗传算法求解的全局收敛性就成为一概率事件。其中有关理论还有待于深入研究,所取得的成果将不仅丰富进化动力学的基本理论,而且对遗传算法的设计和应用具有重要的指导意义。

1.3.2 遗传策略研究与设计

为了维持群体可进化性并最终搜索到问题的全局最优解,遗传算法必须采用适宜的运算形式,这就是遗传策略(genetic strategy)研究与设计的主要研究内容。Holland 在早期研究中将群体结构适应性改变方法称为遗传计划(genetic plan)。

遗传算法应用于优化问题求解,可以视为一种随机化搜索过程。在该搜索过程中,GA 不仅需要探索解空间上的全局最优解(exploration),而且应当充分利用已获得的解空间信息逼近当前局部最优解(exploitation),我们分别称之为 GA 的求泛和求精(reforming and refining)的能力。这也是任何其他类随机化搜索算法追求的基本功能。但是这两种能力并非可以同时获得,对于任何一种算法来讲它们构成了一对矛盾,求精能力好的算法往往不具备良好的解空间上的探索功能,反

之亦然。对于复杂的应用问题,我们往往需要遗传算法兼备这两种能力。

为此,遗传策略研究与设计是一个重要的研究方向[3~7,21,22,83~158],我们可以将之分为微观遗传策略(micro genetic strategy)和宏观遗传策略(macro genetic strategy)。

遗传算法的微观策略主要讨论群体规模、遗传算子的形式和参数设计,及其对GA求解能力的影响。微观遗传策略的理论研究一直集中于遗传算子的适应性的控制,即进化过程中遗传算子参数的适应性调整,进而达到预期的搜索目标[4,112,117,125],并针对一组函数进行了详细测试。同时,在进化策略、进化规划等算法中,也比较注重进化参数的适应性设计[121]。Whitley采用微分方程和马尔可夫链理论,对GA与特定步数Baldwin或者Larmarck局部搜索相结合的进化过程进行了建模分析[127]。

遗传算法的宏观策略主要讨论关于通过对GA流程的再设计改变GA的宏观特征,或者以GA流程为基础,引入其他算法构成混合GA(hybrid genetic algorithms, H-GA)[114~117],以期提高GA求解问题全局最优解的能力。

关于全局—局部混合搜索机制的宏观研究,一般包括三个部分[85]:1)特定的优化问题,2)全局搜索方法与局部搜索方法,3)全局与局部搜索方法的协调。全局和局部搜索方法可以是两种独立的算法,也可以是同一种算法下的适应性变化。

另外,针对多模态函数优化问题(multi-modal function optimization),如何构造一种优化算法,使之能够搜索到尽量多的或者全部全局最优解和有意义的局部最优解,已成为一个重要的研究领域[140,141]。De Jong等人提出了排挤模型(crowding model)[16,143,144]、概率排挤模型(deterministic crowding with probabilistic replacement)[145]等。Goldberg和Richardson提出了适应值共享模型(fitness-sharing model)[90,140,141],基于适应值共享机制的小生境技术(niching technology)[90,146],通过定义群体中个体的共享度,调整个体的适应值,使得群体保持多个高阶模式。

1.3.3 遗传算法编码方式

在遗传算法编码方式的问题上,Holland建议采用二进制编码,并得到了许多学者的支持。双倍体表达是高等生物染色体的重要特性,有长期记忆等作用。Goldberg和Smith[4]用动态背包问题进行了比较研究,实验表明双倍体比单倍体的动态跟踪能力强。

浮点数编码具有精度高、便于大空间搜索的优点,越来越受到重视,Michalewicz[39,50]比较了两种编码方法的优缺点,Qi和Palmieri[51]对浮点数编码的遗传算法进行了严密的数学分析。Vose[168]等扩展了Holland的模式概念,揭示

了不同编码之间的同构性。由于编码是遗传算法应用中的首要问题,建立完善的理论指导十分必要。

Koza于 20 世纪 90 年代初期创立的用于寻找最优计算机程序的遗传规划[25,26],是遗传算法的自然延伸和扩展,也是人工智能领域机器自动编程技术的重大突破。Koza 证明 GP 总能成功地求解每个问题,用大量的实验支持了这一惊人的结论。GP 可用来求解人工智能、机器学习、符号处理等许多领域的各种不同问题。

从整体上来讲,二进制编码的进化层次是基因,而浮点数编码的进化层次是个体。同时,对于非二进制编码,可以结合具体问题领域的知识,设计合适的遗传算子。

1.3.4 遗传算法其他问题

除上述理论研究方面之外,以下也是重要的问题:

1)关于优化问题求解的搜索算法研究的一个出人意料的结果——"no free lunch (NFL)"定理得到了很大的关注[150]。NFL 定理的主要结论为:对任意的表现度量,当对所有可能的目标函数作平均时,所有搜索算法的表现是完全一样的。因此,一个特定的优化方法只能对于某个特定领域的问题,亦即所有目标函数的一个子集来讲,优于另一个算法。在实践中,对一个特定算法如何确定它的适用的函数子集,或者对于具体的优化问题如何选择和设计适宜的算法,有着重大的意义。

2)实际应用中涉及的问题大多数是带有约束条件的,求解约束优化问题是对遗传算法的巨大挑战[13,14,40]。能否处理好约束问题,是能否成功应用遗传算法的一个非常关键的问题。

3)遗传算法的并行化研究(parallel GA,简称 PGA)[161]。设计各种并行执行策略、建立相应的并行化算法的数学基础,对提高进化算法的效率有着重要意义。PGA 主要有细粒度和粗粒度两种计算模型,具体实现的方法有同步主从式、异步并发式和网络分布式等三种[183,227]。目前,专家一致认为,必须对遗传算法进行改造,尽量减少巨量通信开销从而获得高效率,但是这样做会使遗传算法的求解质量有所下降。这就是遗传算法在并行化中遇到的效率与效果之间的矛盾。对 PGA 的研究表明:只要通过保持多个群体和恰当地控制群体间的信息交换来模拟并行执行过程,即使不使用并行计算机,也能提高算法的执行效率。

4)遗传学习系统是以 GA 为核心的增强式学习系统[3,4]。一般来说,群体由产生式规则组成,利用其与环境之间的输入输出来完成学习任务。Holland 奠定了基于遗传的机器学习框架,并首先研究发展了第一个遗传学习系统,它成为以后遗

传学习系统的模板。在此基础上,许多学者对它进行了改进,并将它应用到模式识别、机器人路径规划、股票自动交易等许多领域。

1.4 遗传算法的应用

随着经济、科技和社会的不断发展,人们遇到的各种问题越来越复杂,迫切需要寻找一种更好的求解方法。遗传算法作为一种有效的全局搜索方法,从产生至今不断扩展应用领域,比如工程设计[27,28]、制造业[21,29]、人工智能[4,31~34,41~48]、计算机科学[29,35]、生物工程[30]、自动控制[36~39]、社会科学[21,23,40]、商业和金融等,同时应用实践又促进了遗传算法的发展和完善。传统上比较成功的案例与应用领域如下:

1) 天然气管道的最优控制[27,28]

伊利诺斯大学的 Goldberg 采用遗传算法研究一个复杂结构输气管道系统的运行控制问题。该系统模拟了从西南向东北输送天然气的输气管道系统。输气管道由许多支线组成,每条支线上的送气量各不相同,仅有的控制手段就是用压缩机来改变特定支线中的压力以及用阀门来调节储气罐进出气流的流量,而且管道气压的实际变化极大地滞后于操纵阀门或压缩机的动作。Goldberg 采用拟人控制器经过遗传算法学习,建立了一套完备的规则知识层次体系,能够有效地实施管道运行的实时控制,以及对管道被戳穿事故作出恰当的反映。

2) 喷气式飞机涡轮机的设计[21]

通用电器公司和 Rensselaer 综合技术学院的一组研究人员成功地将遗传算法用到一种商业客机等使用的高函道比喷气发动机的涡轮的设计之中。这种涡轮由好几级静止的和旋转的叶扇组成,安装在近似圆筒形的函道内,是发动机开发计划的核心部分。涡轮的设计涉及到至少 100 个变量,每个变量的取值范围各不相同,由此形成了具有 10^{387} 以上个点的搜索空间。涡轮设计方案的"适应度"取决于它满足一组限制的程度如何,这组限制有 50 个左右,如内壁和外壁的形状,函道内各点处燃气气流的压力、速度和扰动情况等。在一般情况下,一个工程师独立工作并获得一个满意的设计要用大约几周时间。运行基于 GA 的发动机模拟软件和专家系统有助于引导设计人找出有意义的修改,工程师用这样的专家系统,在不到一天的时间里就能完成一种设计。

3) 旅行商问题(TSP)

旅行商问题是经典的组合优化问题之一,已远远超过其本身的含义,成为了一种衡量算法优劣的标准。TSP 问题是采用非标准编码遗传算法求解最成功的一例,用推销员顺序经历的城市名表示基因编码。由于使用标准交叉产生的后代可

能有重复或丢失的基因而成为非可行解，故提出了非标准的交叉和变异方法[53,54]。交叉主要采用重排序方法——部分匹配重排序（PMC）、顺序交叉（OX）和循环交叉（CX）等，变异主要采用位反转、对换、插入等方法。目前，GA 已经能够对 431 个城市的 TSP 问题求得最优解，对 666 个城市的问题可得到满意解。

4）作业调度问题（flow-shop or scheduling）

作业调度问题同样可以用遗传算法进行处理[7,40,115,250]，比如 Cartwright 关于化工厂生产计划的优化安排，Syswerda 关于飞行支持设备调度问题，Hilliard 关于运输军队及其装备多目标通路的作业调度，Gabbert 关于铁路网络复杂运输调度等问题，采用遗传算法均取得了明显效果。

5）遗传学习（genetic learning）

将遗传算法用于知识获取，构成以遗传算法为核心的机器学习系统，其中群体由一组产生式规则组成。比较典型的是 Holland 设计的用于序列决策学习的分类器系统（classifiers）[4]，以及机器人规划、模式识别、概念学习等[42,44,45,115]。

6）自动控制领域

遗传算法适用于求解复杂的参数辨识问题。Maclay 等人用遗传算法求解电车模型参数辨识问题，取得了较好的结果[184]；Karr 采用遗传算法设计自适应模糊逻辑控制器，取得了显著的效果[185]；Freeman 等人提出了应用遗传算法精调控制中的由人定义的模糊逻辑集合概念[186]。另外，GA 在故障诊断[187,188]和机器人行走路径规划[42,189,195]中的应用也取得了成功。

7）人工智能与计算机科学

GA 在人工智能与计算机科学领域中的应用包括：数据挖掘与知识获取[190～192]、数据库查询优化[193]、人工神经网络结构与参数优化[31～34,114,196～198]、模式识别[177,191,194]、专家系统[48,115]等。

8）社会与经济领域

GA 在早期就曾应用于囚徒困境问题分析（prisoner's dilemma problem）[7]，Bauer 对 GA 在经济与投资中的应用进行了全面分析[23]。近年来，商业、金融领域已经成为遗传算法应用热点。遗传算法与现代计算机强大的运算能力结合，使金融交易中瞬息万变的诸多因素能够为人所理解并能加以利用，使交易者更多地依赖于计算机的速度。目前已经有许多基于遗传算法的软件包应用于金融系统和股票投资分析。

另外，很多专家学者将 GA 应用于各自所从事的工程领域，比如 VLSI 设计、运输规划、设备布局、土木工程、生物工程等，对解决具体实践问题起到了极大的促进作用。

本章附录:遗传算法的基本术语

由于遗传算法研究与应用尚在不断发展之中,有关术语的运用尚未完全取得统一。为了在下面研究中做到准确、清晰、规范地描述,本文中采用的术语及其含义解释如下:

- 个体(individual):GA 所处理的基本对象、结构。
- 群体(population):个体的集合。
- 位串(bit string):个体的表示形式。对应于遗传学中的染色体(chromosome)。
- 基因(gene):位串中的元素,表示不同的特征。对应于生物学中的遗传物质单位,以 DNA 序列形式把遗传信息译成编码。
- 基因位(locus):某一基因在染色体中的位置。
- 等位基因(allele):表示基因的特征值,即相同基因位的基因取值。
- 位串结构空间(bit strings space):等位基因任意组合构成的位串集合,基因操作在位串结构空间进行,对应于遗传学中的基因型的集合。
- 参数空间(parameters space):是位串空间在物理系统中的映射。对应于遗传学中的表现型的集合。
- 适应值(fitness):某一个体对于环境的适应程度,或者在环境压力下的生存能力,取决于遗传特性。
- 复制、选择(reproduction or selection):在有限资源空间上的排他性竞争。
- 交叉、交换、交配、重组(crossover or recombination):一组位串或者染色体上对应基因段的交换。
- 变异(mutation):位串或染色体水平上的基因变化,可以遗传给子代个体。
- 逆转或倒位(inversion):反转位串上的一段基因的排列顺序。对应于染色体上的一部分,在脱离之后反转 180°再连接起来。
- 单倍体(haploid,monoploid):细胞核中有 n 个正常的不配对染色体。
- 二倍体(diploid)、多倍体(polyploid):细胞核中有 $2n$ 或更多个正常的配对染色体。
- 基因型(genotype):或称遗传型,指用基因组定义遗传特征和表现。对应于 GA 中的位串。
- 表现型(phenotype):生物体的基因型在特定环境下的表现特性。对应于 GA 中的位串解码后的参数。
- 基因连结(linkage):两个或更多的等位基因出现在同一个染色体上。对应于 GA 中的模式的概念。

- 上位遗传或者基因关联(epistasis):两个非等位基因之间的相互作用,使得其中之一(上位基因)对另一个(下位基因)的表现型产生干扰或抑制。对应于优化函数的非线性特征(non-linearity)。
- 基因多效性(pleiotropy):指单一基因对生物体多个物理性状的影响。对应于 GA 求解多目标优化问题中,某个变量对多个目标函数的影响。
- 多基因效性(polygeny):指生物体某个物理性状由多个基因共同决定。对应于 GA 求解多目标优化问题中,某个目标函数的值由多个变量的状态所决定。
- 遗传源变或遗传漂移(genetic drift):指群体的遗传组成的随机变化,不含自然选择的影响。
- 遗传(heredity):父代个体通过有性方式向子代个体的特征传递过程。
- 局部环境或者环境小生境(environmental niches):具有某种特征的子环境,其中生物体具有特定的染色体结构,表现出特定的物理性状。对应于多模态函数中局部极值点的邻域。

第二章　遗传算法的基本原理

科学来源于人类期望理解和控制世界的实践活动之中。在历史发展进程中，人类社会逐渐建立起了不断丰富的知识体系，使得我们可以在不同程度和层次上进行研究和预测气象变化、行星活动、疾病防治、经济周期等自然和社会现象。电子计算机的发明和应用是人类科学技术发展史上迄今为止最大的一次革命，它极大地拓展了人类的思维空间，延伸了人类了解和适应自然与进行各种社会经济活动的能力。

在计算机诞生的早期，人们就期望计算机借助于程序的形式创造一种新型的智能体，有些专家称之为人工生命(artificial life)。计算机界的先驱，如 Alan Turing, John von Neumann, Norbert Wiener 等，所开展的研究工作在很大程度上可以说是设计具备人类智能的计算机系统，它可以像生命体一样复制、学习和适应环境。他们对生物学和心理学倾注了极大的热情和精力，期望借助于自然系统的概念和规律指导基于计算机的智能体的设计和实现。因此，早期的计算机技术不仅应用于导弹轨迹分析和辨认军事目标，而且也用于描述人类的大脑活动，模仿人类的学习方式，以及模拟自然界的进化过程。这种以生物系统为基础的计算机模拟研究经历了多次的复兴和衰落，20 世纪 80 年代以来在计算机领域再一次活跃起来。首先是人工神经网络(artificial neural networks, ANN)，然后是机器学习(machine learning, ML)，当前即所谓的进化算法(EC)，遗传算法就是其中最重要的一个分支[7]。

在 20 世纪 50 年代和 60 年代，一些计算机科学家独立地开展了旨在可以成为工程问题优化工具的进化系统的研究。其基本思想即通过使用类似自然遗传选择和变异的操作算子，不断地进化一群候选解，以最终得到问题的最优解或满意解。

Rechenberg 在 20 世纪 60 年代引入了“进化策略”(ES)的概念，提出了面向实参数优化问题求解的算法，并用于飞机机翼的设计[13,15]。Schwefel 等人又在 ES 基础上进行了扩展[106,107,118,155]。1966 年，Fogel, Owens, Walsh 开发了“进化规划”(EP)[6,13]，采用有限状态机器表示候选解，通过随机选择和变异进化状态变换。

与 ES、EP 不同，Holland 提出遗传算法时，原本并非用于优化问题求解，而是用以研究自然系统的适应现象，其中将自然适应过程结合到计算机程序之中。

1975 年 Holland 出版的著作《Adaptation in natural and artificial systems》中明确指出遗传算法抽象于生物进化现象，并建立了采用 GA 描述的适应性理论框架。Holland 采用染色体群体表示结构，采用自然选择类型算子和遗传类型的交叉算子、变异算子与逆转算子进行结构进化。每个染色体由一组基因构成，每个基因由一组特定的等位基因值{0,1}表示。选择算子在群体中选择适应性好的染色体，模拟自然选择过程。交叉算子交换两个染色体的部分基因，简单模拟生物体染色体的基因的重组过程。变异算子随机改变某些位置的等位基因值，逆转算子改变染色体上基因的排列顺序。Michalewicz 提出了遗传算法加数据结构等于进化程序的概念(genetic algorithms + data structures = evolution programs)[40]。

Holland 最早提出了基于群体的进化的概念，以及交叉、逆转和变异操作算子。与之相对应，Rechenberg 原始的 ES 往往从由两个个体构成的群体开始，一个作为父个体，另一个作为子个体，子个体是父个体变异的结果，一般不采用多个体群体和交叉操作算子。在后来的研究和应用中，才逐渐采用了交叉算子。Fogel 的 EP 在早期也仅仅使用变异算子实现群体进化。Holland 最早开展进化计算的理论研究[3]，试图将进化算法建立在坚实的理论基础之上。

20 世纪 80 年代以后，经过相关领域专家学者的交流和共同努力，GA、ES、EP 走向了融合，构成了进化计算的基本思想和主要算法形式，遗传算法也从 Holland 的初始内涵不断扩展和丰富，但是其历史背景和发展源泉仍然是 Holland 关于自然和社会的复杂系统的适应性理论。

2.1 复杂系统的适应过程

生态系统、生物体系统、经济系统、社会系统等都是开放的复杂系统[199]。这些系统通过与环境的物质、能量和信息的交换，处于不断的发展变化之中。我们可以采用“适应”(adaptation, ad+aptare, to fit to)过程或行为描述它们的共同特征[3]，将适应过程视为复杂系统变化的基本形式。那么，按照生物进化和系统科学理论，下列问题需要深入研究：

1) 系统变化所适应的是一个怎样的环境？

2) 环境怎样作用于正在发生适应行为的系统？

3) 什么样的系统结构在发生适应性改变？

4) 系统适应的机制或者方法是什么？

5) 关于系统与环境的交互活动的历史，系统保留或继承哪些内容？

6) 系统在对环境的适应过程中存在着哪些限制？

7) 不同的适应过程如何比较和评价？

对于具有不同物理意义的复杂系统，上述问题的研究内容可能不尽相同，目前尚未建立起统一的理论体系和方法论。这里，我们按照 Holland 关于复杂适应系统(complex adaptive systems)的研究思想，试图给出一个比较完备的规范描述框架和数学模型。

2.1.1　复杂系统的适应性

按照 Holland 的定义，所谓系统的适应性(adaptation)就是在环境的作用下系统结构的不断改变的过程。系统的结构(structures)构成了发生适应行为和完成适应过程的基础，而系统结构的改变是通过适应计划或者适应策略(adaptive plan, strategy)进行的。在系统适应过程的不同阶段，适应策略往往需要采用合适的方法或者手段(operators)。比如，在生物遗传过程中，生物细胞核中的染色体是系统的结构，染色体的变异和重组就是适应计划；在经济系统中，制造的产品或提供的服务的组合是系统结构，生产活动、经济政策是适应计划；在人工智能系统中，表现为不同形态的知识就是结构，知识获取和调整的规则就是适应计划，等等。其中，不同的适应方法将导致系统结构变化的序列不一样，因此就构成了不同的适应过程。

在特定的环境下(E)，不同结构(A)的系统表现出不同的适应性(more or less fit)，一般采用适应性测度函数 $\mu_E(A)$表示。同时，相同结构的系统在不同适应计划(τ)的作用下，也会表现出不同的适应性。对于不同的物理系统，系统对环境的适应性也具有不同的含义。比如，生物体系统的适应性表现为生物体在特定环境下的生存能力；经济系统的适应性表现为社会效用的大小；人工智能系统的适应性表现为推理过程和结论的有效性，等等。

随着环境的改变和系统结构的变化，适应性测度函数 $\mu_E(A)$也需要进行改进，我们称为适应性测度的适应性。若适应性测试函数的整体为$\mathscr{U}$，则特定环境下结构测度 $\mu_E(A)$是$\mathscr{U}$的一个具体形式，即 $\mu_E(A)\in\mathscr{U}$。

由于在系统的适应过程中，我们并不知道最佳的结构，并且存在着环境信息的不完备性，因此需要测试多个系统结构，所以系统的适应过程往往采用一种结构群体，称为个体(特定结构的系统，individual)群体(population)。若可以测试的结构整体为$\mathscr{A}$，则特定结构 A 是$\mathscr{A}$的一个具体形式，即 $A\in\mathscr{A}$。

对于开放的复杂系统，环境也在不断变化。在系统整个适应过程中，特定时间阶段 t 的环境($E(t)$)是整体环境中($\mathscr{E}$)的一个特定形式或状态，即 $E\in\mathscr{E}$。在 $E(t)$下适应性好的系统结构，在 $E(t+1)$下可能变得很差，或者其他系统结构会产生很好的适应性。

同样,适应计划在系统的整个适应过程之中,也需要随着环境的变化不断改变,即适应计划的适应性。在环境(E)下,当某种系统结构(A)表现出良好的适应性时,适应计划(τ)应当在原来的基础上加强该结构,以便进一步提高其适应性。当某种系统结构(A')表现出较差的适应性时,适应计划(τ')应当对该结构进行较大的调整,或者选择其他适应性更好的结构。从另一个角度来讲,在环境(E)下有效的适应计划,在环境(E')下未必仍然有效,需要重新确定或调整适应计划。因此,适应计划的适应性需要综合考虑系统环境、系统结构、适应性测度、适应过程的历史信息等因素。若在整体环境 $\mathscr{E}$ 中,在整体结构 $\mathscr{A}$ 上,在适应性测度函数整体 $\mathscr{U}$ 下,可以采用的适应计划整体表示为 $\mathscr{T}$。在特定环境 E 中,对于某种系统结构 A,采用适应性测度函数 $\mu_E(A)$,相应的具体的适应计划 $\tau\in\mathscr{T}$。

在不断变化的环境中,采用适应计划逐步改变系统结构,形成了在整个结构空间上的一条适应过程轨迹(trajectory)。显然,不同的环境变化将导致不同的系统结构的适应过程轨迹,同时适应计划的适应性也影响着该轨迹的变化。那么,从系统适应过程来讲,系统结构能否很好地完成特定环境变化下适应过程,与系统结构的表示形式、适应性测度函数和适应计划等有着直接的关系。在某种情形下,可能出现适应过程的困难局面。比如:

1) 整体结构 $\mathscr{A}$ 的空间为无穷或者非常大,测试全部结构或有意义的结构将花费太多的计算时间。

2) 系统结构 A 非常复杂,难以确定其中的哪些子结构或组成部分对系统的适应性有重要影响。

3) 适应性测度函数 $\mu_E(A)$ 是包涵大量参数的复杂函数,具有高维、非线性、非可加性、多模态、非连续等性质,以及含有随机噪声。

4) 适应计划难以适应环境的变化,或者难以确定不同环境下合理的适应计划。

5) 对于环境变化提供的信息(包括历史信息),含有误导信息,需要必要的识别和过滤处理。

由于适应计划是实现系统结构适应性改变的惟一措施,所以适应计划的设计是克服适应困难的主要方式。

为了清晰地描述复杂系统的适应性,我们以生物体的进化过程为例给予具体分析。按照遗传学的基本理论,生物体物理特性由包含在细胞核中的染色体(chromosomes)上的基因(gene)确定。每位基因有多种状态或者称为等位基因(alleles),不同基因组合的可能性构成了生物体的多样性。比如,某些豌豆花的颜色由染色体上的一个基因来确定,该位的一个状态或者等位基因决定了开出白色花朵,另一个状态或者等位基因决定了开出粉色花朵。面包酵母的染色体中有一个

基因的等位基因决定维生素 B_1 的合成，而其他等位基因则不具有该项功能。在脊椎动物的染色体上包含了数以万计的基因，每一个基因均具有多个等位基因，其物理性状通常不是由单个基因决定，而是由若干基因的组合决定。比如，人的眼睛的颜色就是由染色体上的多个基因共同决定的。

用染色体表示系统结构 A，则系统整体结构 $\mathscr{A}$ 的空间将非常巨大。若假定染色体长度为 10000，每个基因具有两个等位基因，则结构空间 $\mathscr{A}$ 将包含 $2^{10000} \approx 10^{3000}$ 种潜在的结构形式。即使采用 10^{10} 个生物个体构成的群体进行适应性测试也仅仅可以完成 $\mathscr{A}$ 中的微小部分。

某一物种存在的大量的遗传结构（称为基因型，genotypes）表明了该物种环境适应过程的复杂性，最基本的复杂性来自于染色体上基因的交互关系。一组基因的等位基因对生物体性状（称为表现型，phenotypes）的影响是一种非可加性的关系，当该组基因的某些基因为特定的等位基因时，生物体表现为预期性状；当该组基因中的任何一个或多个基因为非特定等位基因时，生物体将表现为其他相差很大的性状。从另一个角度来讲，单个基因所决定的物理性状与其他基因的等位基因有关。对于某个基因为特定的等位基因，而其他基因取不同的等位基因时，生物体将表现出不同的物理性状。某个基因的特定等位基因对于生物体呈现为某种性状可能起到促进作用，也可能起到抑制作用。因此，生物体的物理性状决定于某些基因的等位基因的特定组合，这种现象称为“基因关联”效应（epistasis）。

由于基因关联效应的存在，我们不能测量单个基因的形态对生物体适应性的影响程度。当基因的某个等位基因使得生物体在某种环境下表现出良好的适应性时，在其他环境下却可能是致命性的。所以，我们不可能独立地选择和确定单个基因的等位基因，而不考虑其他基因的等位基因。因此，在关于生物体进化过程的系统适应性分析中，当系统结构的改变呈现为基因关联时，适应计划单纯地改变系统结构的单个基因将不会产生良好的适应行为。实际上，适应计划需要测试不同的等位基因构成的系统结构的适应性，探索构成特定关联关系的等位基因模式，即包含特定元素的系统结构，才能完成系统的适应过程。所以，制定合适的适应计划将是一个非常复杂的事情，系统欲实现对环境的良好适应性将面临巨大的困难。Holland 关于基因关联的描述采用了“模式”（schema）的概念。

在生物体进化过程（系统的适应过程）中，不同的局部生态环境也决定了生物体的多样性。比如，在水的生活环境中，鱼的基因组合的适应性改变产生了鱼鳞。特定基因组合的生成表现为对环境特征的依赖性，当环境具有不同的特征时，生物体的基因组合也不一样。这种情形称为局部环境或者环境小生境（environmental niches），即具有某种特征的子环境，其中生物体表现出特定的物理性状。比如，在低洼缺氧的沼泽环境中，一种厌氧细菌经过适应性的进化可以用硫替代氧进行氧

化作用。哺乳动物耳骨的进化可以适应 50～50 000 赫兹的空气震动，同时伴随着颌和信息处理系统的进化。另外，相同的局部环境也可以使得产生具有同样良好适应性的不同的基因组合。比如，水中哺乳动物的眼睛和章鱼的眼睛都能很好地适应水中生存环境，但它们的基因组合却不一样。

各种局部环境或者小生境（$E \in \mathscr{E}$），需要不同的适应计划来完成相应的系统结构的改变。由于小生境的差异性，生物体的物理性状或者基因的表现型也不一样，不同小生境下的适应性最好的系统结构也不相同。如何选择适应计划实现系统结构对多个小生境下的适应过程是一个有待深入研究的新问题。

自然界生物体的进化过程往往是以生物体群体形式进行的，单一个体不具备进化行为。适当规模的生物体群体是进化的基础，也可以视为生物体染色体进行杂交和变异的环境的一部分，适应计划作用于群体即多个系统结构。同时，生物体群体包含的多个生物体个体必须存在着差异，否则进化过程也不可能发生。因此，系统的适应过程一般建立在群体之上，一方面克服环境信息的不完备性，另一方面也可以保留适应过程的历史信息。

2.1.2 适应过程的数学模型

复杂系统的适应过程在本质上是一个优化过程，以往各个具体的物理系统的描述采用的数学模型也不尽相同。这里，我们试图按照前面的复杂系统适应性的分析内容，建立一套规范的适应过程描述的模型体系。

在生态、经济、控制等复杂系统的研究中，一般采用离散时间过程。为了处理方便，我们不妨设定复杂系统的适应过程的时间阶段为 $t=1,2,\cdots,T$。按照遗传学理论，生物体进程过程研究中的时间阶段称为代。当然不同物理系统的时间阶段的概念可能存在着很大的差异。比如，自适应控制系统的时间阶段长度可能是秒，经济系统的时间阶段一般为季、年，人文社会系统的时间阶段一般为世纪，生物进化研究的时间阶段（代）通常是千年或者百万年等。

复杂系统结构的具体表示方式非常多。比如，控制系统中的状态向量、经济计划中的产品数量、优化问题中的参数变量等。借鉴生物体细胞核中的染色体的组成，我们将系统结构表示为基因模式，系统结构就是一个长度为 L 的染色体 $A=\langle A_1, A_2, \cdots, A_L \rangle$，第 i 位基因上存在一系列的等位基因（$A_i=\{a_{i1}, a_{i2}, \cdots, a_{ik_i}\}$，$i=1,2,\cdots,L$；$k_i$ 表示第 i 位基因等位基因的数量），那么所有等位基因的组合构成了整体结构或者结构空间：

$$\mathscr{A}= A_1 \times A_2 \times \cdots \times A_L = \prod_{i=1}^{L} A_i \qquad (2-1-1)$$

按(2-1-1)定义的$\mathscr{A}$是可能存在的结构的最大空间,现实中存在的生物染色体或者系统结构一般仅仅为$\mathscr{A}$的真子集。

对于处于阶段 t 的系统结构 $A(t)$,其环境 $E(t)$提供的信息为 $I(t)$,在适应计划 τ_t 作用下生成新的系统结构:

$$\tau_t: A(t) \times I(t) \rightarrow A(t+1) \tag{2-1-2a}$$

或者

$$A(t+1) = \tau_t(A(t), I(t)) \tag{2-1-2b}$$

生物体的进化过程不可能隔断历史,生物体的进化和历史过程有着间接的关系,所以在新的系统结构生成时,不仅要考虑现在的环境信息,还需要考虑环境变化的历史信息,或者历史上环境提供的信息 $M_E(t)=\langle I(1), I(2), \cdots, I(t-1)\rangle$。因此,式(2-1-2)可以改写为

$$\tau_t: A(t) \times I(t) \times M_E(t) \rightarrow A(t+1) \tag{2-1-2c}$$

或者

$$A(t+1) = \tau_t(A(t), I(t), M_E(t)) \tag{2-1-2d}$$

同时,适应计划应当提供环境历史信息的继承与扬弃的合理处理方式:

$$M_E(t+1) = \tau_t(M_E(t), I(t)) \tag{2-1-3}$$

适应计划对系统结构的改变可以视为一个随机过程。若对于有限的结构空间,$\mathscr{A}=\{a_1, a_2, \cdots, a_j, \cdots, a_N\}$, $N=|\mathscr{A}|$,当 $A(t)=a_j(j=1,2,\cdots,N)$,在适应计划 τ_t 作用下变为 $A(t+1)=\{a_q \mid q=1,2,\cdots,N\}$ 的概率为 $P(t)=\{p_{i,q}(t) \mid \sum_{q=1}^{N} p_{i,q}(t)=1\}$。当 $q=j$ 时表示未发生改变。那么,式(2-1-2c)可以写为:

$$\tau_t: A(t) \times I(t) \times M_E(t) \rightarrow P(t+1) \tag{2-1-4}$$

关于系统结构 $A(t)$对环境 $E(t)$的适应性测度一般采用大于等于 0 的实数表示,称为支付或者报酬,则

$$\mu_{E,t}: A(t) \times E(t) \rightarrow R^{+} \tag{2-1-5a}$$

或者表示为

$$\mu_E(A(t)) = \mu_{E,t}(A(t), E(t)) \tag{2-1-5b}$$

以(2-1-5)为基础,在 t 阶段环境信息可以表示为

$$I(t) = \mu_{E,t}(A(t)) \tag{2-1-6}$$

当环境和系统结构变化时,适应计划也需要进行适应性调整,使得系统结构具有最好的适应性。适应计划的适应性调整应当考虑当前系统结构 $A(t)$、环境 $E(t)$和环境信息 $I(t)$、$M_E(t)$、历史选择 $M_\tau(t)=\langle\tau(1),\tau(2),\cdots,\tau(t-1),\tau(t)\rangle$等,一般表示如下: