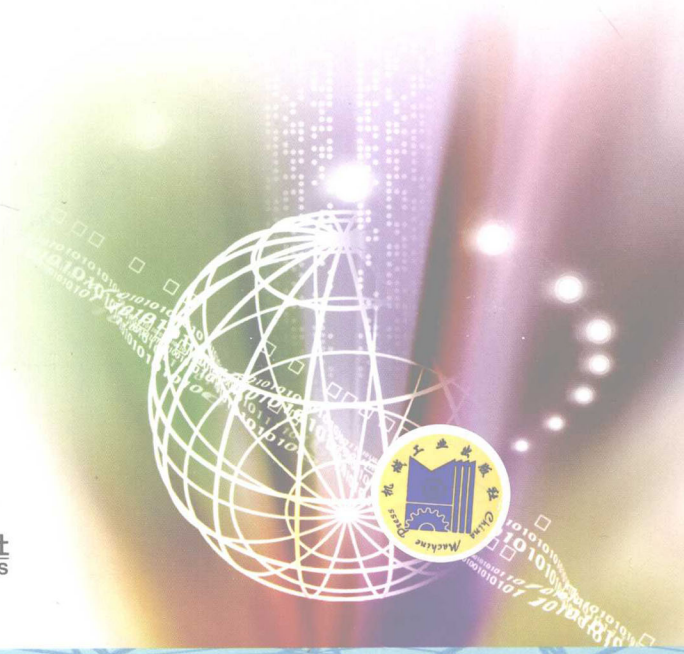
copula

分布估计算法

Copula Estimation of Distribution Algorithm

◎ 王丽芳 著



机械工业出版社
CHINA MACHINE PRESS

copula 分布估计算法

王丽芳 著

机 械 工 业 出 版 社

copula 分布估计算法综合了智能计算领域和统计学领域的知识，是一种基于群体的智能优化算法。本书共分为 8 章，主要内容包括：分布估计算法的基本概念和算法流程、copula 理论及 copula 分布估计算法框架、二维 copula 分布估计算法、多维经验 copula 分布估计算法、多维阿基米德 copula 分布估计算法、阿基米德 copula EDA 中的参数估计、copula EDA 中的边缘分布函数和采样方法的研究以及 copula EDA 在图像矢量量化中的应用研究等。

本书适合于从事智能计算领域研究和应用的科技工作者和工程技术人员使用，也可以作为人工智能、计算机科学、信息科学、智能优化及智能控制领域的广大师生的教学参考用书。

图书在版编目（CIP）数据

copula 分布估计算法/王丽芳著. —北京：机械工业出版社，2012.8
ISBN 978-7-111-39281-1

Ⅰ.①c… Ⅱ.①王… Ⅲ.①电子计算机—最优化算法 Ⅳ.①TP301.6

中国版本图书馆 CIP 数据核字（2012）第 172377 号

机械工业出版社（北京市百万庄大街 22 号 邮政编码 100037）
策划编辑：张敬柱 王晓洁 责任编辑：王晓洁
版式设计：石 冉 责任校对：樊钟英
封面设计：赵颖喆 责任印制：杨 曦
北京中兴印刷有限公司印刷
2012 年 9 月第 1 版第 1 次印刷
140mm×203mm · 5 印张 · 128 千字
0 001—3 000 册
标准书号：ISBN 978-7-111-39281-1
定价：28.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换

电话服务		网络服务	
社服务中心：	(010)88361066	教 材 网：	http://www.cmpedu.com
销 售 一 部：	(010)68326294	机工官网：	http://www.cmpbook.com
销 售 二 部：	(010)88379649	机工官博：	http://weibo.com/cmp1952
读者购书热线：	(010)88379203	封面无防伪标均为盗版	

前　言

分布估计算法源于遗传算法，它没有交叉和变异操作，而是对优势群体的概率分布模型进行估计，并根据估计的模型进行采样。分布估计算法从宏观上控制算法进化的方向。由于分布估计算法具有更强的理论基础，因此受到广大学者的重视，成为当前进化计算领域的研究热点。

本书介绍的copula分布估计算法，在估计优势群体的概率分布模型时，将边缘分布函数的估计操作和copula函数的估计操作分别进行，实际上这两个操作是可以独立地并行实现的。本书分为8章，第1章介绍了分布估计算法的基本概念和算法流程。第2章对copula分布估计算法的理论基础进行了介绍，并分析得出了copula分布估计算法的基本框架。第3章简单介绍了的二维copula分布估计算法，使读者对copula分布估计算法有一个直观的认识。第4章介绍了基于经验copula的copula分布估计算法，该算法能够解决二维以上的多维优化问题。第5～第7章介绍了基于阿基米德copula的多维copula分布估计算法，及相应的copula函数参数估计、边缘分布函数的选择和采样方法方面的问题。第8章将copula分布估计算法应用于图像矢量量化的研究中，为解决码书设计问题提供了参考。

本书是作者近年来科研工作的总结，得到了曾建潮教授的悉心指导和太原科技大学复杂系统与计算智能实验室各位同仁的大力支持。本书的研究工作得到山西省青年科学基金项目（项目编号：2010021017－2）、山西省高校产业化项目（项目编号：2010015）、山西省优秀研究生创新项目（项目编号：20113121）的资助。太原科技大学对本书的出版给予了资助和支持。作者在

此一并向相关单位和个人表示最诚挚的谢意。

由于作者学识有限，书中难免存在不妥之处，恳请各位专家和广大读者及时给予批评指正。

编者

目　录

第1章 绪 论

1.1 智能计算

1.1.1 人工智能

随着计算机的出现，人工智能开始产生并蓬勃发展起来，成为当前科学技术发展中的前沿学科。它同空间技术、原子能技术一起被誉为是二十世纪的三大科学技术成就。

求解优化问题是人工智能领域的一个重要问题，在经济、生活、科学和工程中，许多问题都可以转化为优化问题。由于该问题具有普遍性和重要性，因此几十年来，许多的学者对此问题进行了广泛而深入的研究，并出现了许多新的理论和方法。

1.1.2 进化计算

进化计算的研究始于20世纪30年代，学者们将生物进化论的思想和特征引入到工程研究领域，以解决工程中的优化问题。在进化计算中主要采用了生物进化中的选择、遗传等思想，设计了相应的算子。遗传算法、进化策略和进化规划是进化计算的三个重要支柱[1]。

1963年，德国的I. Rechenberg和H. P. Schwefel提出了进化策略(Evolutionary Strategies，ES)[2]。进化策略中的两个基本算子是变异和选择。在$(\mu+\lambda)$进化策略中，每代进化时，当前群体中的μ个父个体通过变异产生λ个子个体，然后从这$\mu+\lambda$个个体中选择性能最好的μ个个体组成新一代群体。商μ/λ被称为选择压力，其值越小，说明选择压力越大。

Fogel，Owens和Walsh在1966年提出了进化规划(Evolution Programming，EP)[3]。该方法用有限状态机表示所优化的问题，其变异算子与进化策略的变异算子类似。进化规划同样只有变异和选择两个算子，由μ个父个体产生μ个子个体。与进化策略的确定性选择方式

不同的是，进化规划采用了随机-竞争的选择方式，即在由 μ 个父个体和 μ 个子个体组成的临时群体中，对于每个个体，都从另外的 $2\mu-1$ 个个体中随机选择 p 个个体与之比较，若该个体适应值较优或与比较对象的适应值相同，则该个体加一分，最后从临时群体中选择分数最高的 μ 个个体组成新一代群体。

遗传算法是 John H. Holland 教授在 1975 年提出的[4]，它模拟了自然界“自然选择，适者生存”的现象。在 Holland，De Jong 和 Goldberg 等人的努力下，遗传算法得到了迅速发展和完善[5]，引起了全球学者的广泛关注和研究兴趣，成为目前应用最广泛也最成功的算法[6]

1.1.3 无免费午餐定理

在优化领域中的一个重要定理是无免费午餐定理（No Free Lunch Theorem，NFL 定理）。该定理是 Wolpert 和 Macready 在 1997 年提出的[7]，他们在 IEEE Transactions on Evolutionary Computation 上发表了论文“No free lunch theorems for optimization”，提出了 NFL 定理并进行了严格的论证。该定理指出，对于任意给定的两个算法 A 和 B 及所有可能的问题，如果 A 在其中一些问题上表现得比 B 好，那么 A 在其他问题上的表现就一定比 B 差，反之亦然。也就是说，对于整个函数类，不存在最佳的万能算法，任意的两个算法对这些问题的平均表现度是一致的。需要注意的是该定理只是在有限的搜索空间中成立，对于无限的搜索空间，结论是否成立尚不明确。

该定理的提出引起了优化邻域的震惊并引发了持久的争论。其结果多少有些令人感到意外。NFL 定理的价值主要是在观念上为研究和应用优化算法提供了一定的启示作用。

1.2 遗传算法

遗传算法（Genetic Algorithm，GA）是广泛应用的全局智能优化算法，通过选择（Selection）、交叉（Crossover）和变异（Mutation）三种基本操作实现全局最优解的搜索。一个问题只要能够将解表示成一个固定长度的字符串，就可以用遗传算法进行优化。这样的一个字

符串被称为一个染色体（Chromosome）或者个体（Individual）。当使用遗传算法优化一个问题时，需要选定基本字母表、字符串长度和适应值函数，通过基本字母表将问题的解转化成一个固定长度的字符串，并通过适应值函数对该字符串的质量进行评价。适应值函数的返回值是一个实数，值越大说明解越好。

在生物进化过程中，DNA 通过复制转移到了新的细胞中，使得新细胞继承了旧细胞的基因。有性繁殖时，两个同源染色体通过交叉重组成两个相同的染色体。有时也会以极小的概率产生差错，使 DNA 发生变异形成新的染色体。在宏观上，生物进化是以群体的形式进行的，群体中的每个生物称为个体，每个个体对其生存环境的适应能力称为适应值。

1.2.1 基本算法

遗传算法是基于群体的智能优化算法，开始时随机选择一组字符串作为初始群体，每一个字符串代表一个个体，在此基础上反复执行遗传操作实现群体的进化，遗传算法流程如图 1-1 所示。在每代循环体中，通过一定的选择策略，从当前的父代群体中选择出较好的一些个体组成优势群体，在优势群体中应用交叉和变异操作产生新个体，并对父代群体更新形成子代群体。交叉就是把优势群体中的两个或两个以上的个体中部分字符串进行交换。常用的交叉有单点交叉和双点交叉。变异是按一定的概率改变解中的某些基因位的值，其作用是对交叉所得的解进行一定程度的干扰。通过交叉和变异得到的解会替换父代群体中的一部分个体而

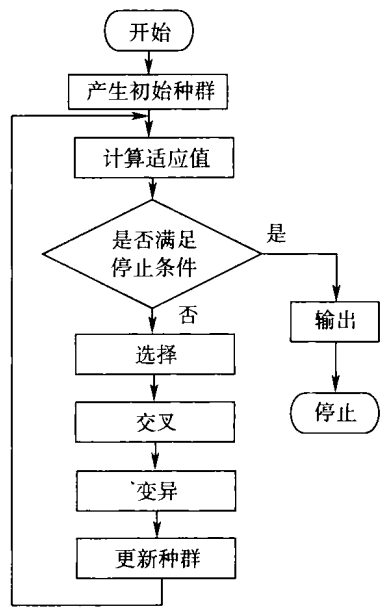


图 1-1 遗传算法流程图

构成子代群体。当不满足终止条件时，子代群体成为父代群体，重复上述步骤。

选择操作是根据每个个体的适应值选择出一些较好的个体，这样可以保证群体向好的方向进化。交叉操作和变异操作实现了对解空间的搜索。交叉操作将几个父代个体的一部分进行连接组合产生新的个体，从而实现了群体的更新。变异操作是对原有个体中的部分字符进行改变，从而实现了在原有解周围的搜索，类似于局部爬山法。

遗传算法可以形式化地表示为

$$GA = P(0), N, I, S, g, p, f, t \tag{1-1}$$

其中，$P(0) = (x_1(0), x_2(0), \cdots, x_N(0)) \in I^N$ 表示初始种群，I 是位串空间，N 是种群规模，即其中包括的个体数目；$S: I^N \to I^N$ 表示选择策略；g 表示遗传算子，一般包括选择算子（$O_s: I \to I$）、交叉算子（$O_c: I \times I \to I \times I$）和变异算子（$O_m: I \to I$）；$p$ 表示遗传算法的操作概率，包括选择概率、交叉概率和变异概率；$f: I \to R^+$ 表示适应值函数；$t: I^n \to \{0,1\}$ 表示终止条件。

1.2.2 模式定理

在遗传算法中，将基于三值字符集 {0，1，*} 所产生的能描述具有某些结构相似性的 0，1 字符串集的字符串称为模式。以长度为 5 的串为例，模式 ** 101 表示在后三位为 101 的所有字符串的集合 {00101，01101，10101，11101}。模式 H 中确定位置的个数称为该模式的阶（Schema Order），记作 $O(H)$。若上述模式 ** 101 的阶为 3，模式 11 ** 0 的阶也为 3。模式 H 的定义距（Defining Length）则表示模式 H 中第一个确定位置和最后一个确定位置之间的距离，记作 $\delta(H)$，如 $\delta(**101) = 2$，$\delta(11**0) = 4$。称平均适应度高于群体平均适应度的模式为优良模式。模式反映了遗传算法中个体之间的结构相似性，它对遗传算法的收敛性有重要的影响。

模式定理：设遗传算法的交叉概率和变异概率分别为 P_c 和 P_m，l 为个体基因链编码的长度，则有

$$m(H, g+1) \geqslant m(H, g) \frac{f(H, g)}{f} \left[1 - P_c \frac{\delta(H)}{l-1} - O(H) P_m \right] \tag{1-2}$$

其中 $m(H,g)$ 表示第 g 代群体中与模式 H 相匹配的个体的数目，$f(H,g)$ 表示中这些个体的平均适应度，$\bar{f}$ 表示第 g 代群体中所有个体的平均适应度。

模式定理是遗传算法中的一个重要定理，它表明在适当条件下，通过选择、交叉和变异的作用，具有低阶、短定义距的优良模式在后代中将以指数级增长。但是在算法中要保持这一条件是非常苛刻的。这种低阶、短定义距的优良模式在遗传算法中被称为建筑块（building block）。

1.2.3 建筑块假设

建筑块假设的内容是在遗传算子作用下，建筑块之间相互影响，能够生成高阶、长定义距、高平均适应度的模式，最终可生成全局最优解。建筑块假设能够有效使用的重要条件是在数据编码中能够明确表示建筑块的位置[8]，否则遗传算法中的交叉和变异操作就有可能破坏建筑块，特别是那些较长的或距离较远的块，从而导致算法的收敛速度下降或找不到最优解。正因如此，越来越多的研究专注于学习问题的结构，并利用这些信息来确保建筑块的增长。

1.3 分布估计算法

分布估计算法（Estimation of Distribution Algorithm，EDA）源于遗传算法，是为解决遗传算法中存在的因交叉、变异操作而导致的建筑块破坏问题而提出的[9]。在分布估计算法中没有交叉和变异操作，取而代之的是对优势群体的概率分布模型进行估计，并根据估计的模型进行采样。由于分布估计算法具有更强的理论基础，因此受到广大学者的重视，成为当前进化计算领域的研究热点[10,11]。

与遗传算法类似，分布估计算法也是基于群体的一种进化算法，算法由一个初始群体开始，在每代循环中，首先根据每个个体的适应值选择出具有代表性的适应值较好的一些个体组成优势群体。然后根据优势群体估计其中的个体所服从的概率分布模型，比如联合正态分布。第三步便是根据估计的概率分布模型采样产生一些新个体，并用来替换原有群体中的部分个体，从而组成新一代的群体。当满足循环

终止条件时，算法结束，当前群体中的适应值最好的个体就是算法优化的结果。分布估计算法流程可以用图 1-2 表示。与图 1-1 中遗传算法流程相比，分布估计算法中没有了交叉和变异两个操作，而是对选择群体的概率分布模型进行估计和采样。由于建筑块中的变量相关性要比其他变量之间的相关性强。分布估计算法更注重变量的相关结构，对建筑块的分布情况进行整体估计，这样不仅可以避免交叉变异所带来的建筑块破坏问题，反而有助于建筑块的增长。

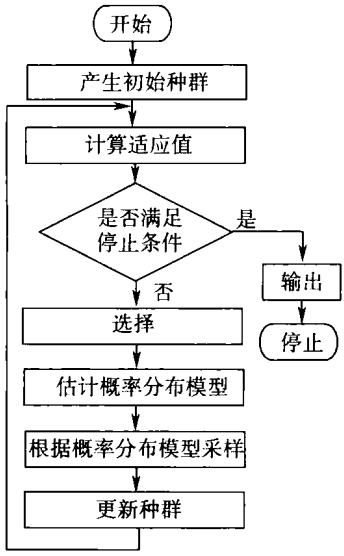


图 1-2　分布估计算法流程图

分布估计算法的基本步骤如下：

Step 1：初始化群体。在搜索空间内按均匀分布随机产生 PS 个点，组成初始群体，记进化代数 g 为 1。

Step 2：计算适应值。根据适应值评价函数计算第 g 代群体中的各点的适应值。

Step 3：选择优势群体。根据适应值，运用一定的选择策略（如截断选择，轮赌选择等）选出适应值较好的 s 个个体组成优势群体。

Step 4：估计优势群体的概率分布模型。一般假设优势群体服从某一类概率分布模型，然后以优势群体为样本，对具体的概率分布 F 进行估计。

Step 5：根据估计的概率模型进行采样，产生一些新个体。在搜索空间内按概率分布 F 随机产生 l 个点，作为新生代中的部分个体。

Step 6：更新群体。由第 g 代群体中 m 个适应值较好的个体，新生的 l 个个体，以及随机产生（或者以较好个体为初始点搜索等方式产生）的 $PS-m-l$ 个个体组成新一代群体，并将进化代数 $g \rightarrow g+1$。

Step 7：若满足某种停止条件，则算法结束，g 代群体中的最好个体就是优化的结果；否则算法转到 Step 2 继续执行。

1.3.1 变量相关性方面的研究

在分布估计算法中，最基本的也是和遗传算法不同的两个操作是：

1）根据选择出的优势群体估计其概率分布模型。

2）根据估计的概率分布模型产生服从该分布的个体。

概率分布模型有多种多样，根据变量的相关性，和反映相关性的方式，目前的分布估计算法主要有以下三类：变量相互独立的分布估计算法、变量两两相关的分布估计算法、多变量相关的分布估计算法。

1. 变量相互独立的分布估计算法

PBIL（Population-Based Incremental Learning）[12]是 Baluja 在 1994 年提出的，被公认为是分布估计算法的初始模型，它解决的是二进制编码的离散优化问题。在 PBIL 中，优化变量是 0－1 编码的，它们之间相互独立。设优化变量为 $(x_1,x_2,\cdots,x_n)$，$x_i \in \{0,1\}$，在每代选择群体中，对个体分别统计每一位上 1 出现的频率 n_i/N，其中 N 是优势群体的规模，$n_i = \sum_{j=1}^{\lambda} x_i^j$ 是优势群体中第 i 位上出现 1 的个体数。PBIL 在第 $k-1$ 代概率密度函数的基础上进行线性调整，得到第 k 代估计的概率密度函数 $p_k(X) = \prod_i p_k(x_i) = \prod_i [\alpha \times p_{k-1}(x_i) + (1-\alpha) \times n_i/N]$，其中初始概率 $p_0(x_i) = 0.5$，$0 < \alpha < 1$ 为学习速率。在产生新个体时，第 i 位上按概率 $p_k(x_i)$ 产生 1，由此构成一个新的 0－1编码的个体。

UMDA（Univariate Marginal Distribution Algorithm）是 Mühlenbein 在 1996 年[13]提出的另一种变量相互独立的分布估计算法。与 PBIL 类似，UMDA 对第 k 代选择出来的优势群体 D_k 估计每一位上 1 出现的频率，并将此频率作为当前代的概率密度，直接对此密度采样产生新个体。即第 k 代估计的概率密度函数 $p_k(X) = \prod_i p_k(x_i) = \prod_i \frac{\sum_{j=1}^{N} \delta_j(X_i = x_i \mid D_k)}{N}$，其中 $\delta_j(X_i = x_i \mid D_k) = \begin{cases} 1 & X_i = x_i \\ 0 & \text{其他} \end{cases}$。

紧致遗传算法 cGA（compact Genetic Algorithm）是美国 UIUC 大学的 Harik 等人在 1998 年[14]提出的，cGA 只需要很少的个体数就可以对问题进行优化，算法简单，更适合于硬件实现。cGA 的具体步骤如下：

Step 1：初始化概率向量 $p_0(X) = [p_0(x_1), p_0(x_2), \cdots, p_0(x_n)] = (0.5, 0.5, \cdots, 0.5)$，进化代数 $k = 0$。

Step 2：根据 $p_0(X)$ 产生两个个体并计算其适应值，将较好的个体记为 $X_1 = (x_{11}, x_{12}, \cdots, x_{1n})$，较差的个体记为 $X_2 = (x_{21}, x_{22}, \cdots, x_{2n})$。

Step 3：更新概率模型。对于第 i 位，如果 $x_{1i} \neq x_{2i}$ 且 $x_{1i} = 1$，那么 $p_{k+1}(x_i) = p_k(x_i) + \alpha$；如果 $x_{1i} \neq x_{2i}$ 且 $x_{1i} = 0$，那么 $p_{k+1}(x_i) = p_k(x_i) - \alpha$；其中 $\alpha \in (0,1)$，一般取 $\alpha = \dfrac{1}{2K}$，K 为正整数。

Step 4：$k = k + 1$，如果概率向量 $p_k(x_i) > 1$，则令 $p_k(x_i) = 1$；如果概率向量 $p_k(x_i) < 0$，则令 $p_k(x_i) = 0$。

Step 5：当 $p_k(X)$ 中的任意元素 $p_k(x_i)$ 或者为 0，或者为 1，则算法结束，$p_k(X)$ 就是最终解；否则转至 Step 2 继续执行。

认为优化变量相互独立的典型的连续分布估计算法有 PBILc[15]和 UMDAc^G[16]，在这两种算法中分别沿用的原有算法的概率更新策略，与原有算法不同的是优化变量是连续的，估计概率模型时使用了正态分布，根据优势群体分别估计各变量的样本均值和样本方差，作为正态分布中数学期望和方差的估计值，即第 i 维变量服从正态分布 $N(\mu_i, \sigma_i^2)$，其中 $\mu_i = \dfrac{1}{s}\sum_{j=1}^{s} x_{ji}$，$\sigma_i = \sqrt{\dfrac{1}{s}\sum_{j=1}^{s}(x_{ji} - \mu_i)^2}$。

在这一类分布估计算法中，概率模型中变量之间是相互独立的，其结构如图 1-3 所示。由于很多优化问题中变量是相关的，甚至是强耦合的，因此这类算法在理论上的意义远大于实际应用价值。

2. 变量两两相关的分布估计算法

变量两两相关的分布估计算法将变量之间的关系简化，仅考虑相邻两个变量之间的相关性。De Bonet 等人在 1997 年提出的 MIMIC（Mutual Information Maximization for Input Clustering）[17]用链式结构表示变量之间的关系，如图 1-4 所示。对于优化变量 $(x_1, x_2, \cdots, x_n)$ 存