改进的正态分布的分布估计算法

邱 玲[1] 高 尚[2] 曹存根[3]
(人工智能四川省高校重点实验室 自贡 643000)[1] (江苏科技大学计算机科学与工程学院 镇江 212003)[2]
(中国科学院计算所智能信息处理重点实验室 北京 100190)[3]

摘 要 针对连续空间函数优化问题,提出了改进的正态分布的分布估计算法。该算法将优选出的个体看作正态分布,然后以正态分布概率模型随机采样产生新的种群,并挑选部分个体与保留的最好解进行交叉操作。将其与均匀分布的分布估计算法、正态分布的分布估计算法进行了比较,结果证明该方法的效果更好。最后分析了选择较好个体的比例对算法的影响。

关键词 分布估计算法,连续空间优化,正态分布,均匀分布
中图法分类号 TP18 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.8.006

Improved Estimation of Distribution Algorithms Based on Normal Distribution

QIU Ling[1] GAO Shang[2] CAO Cun-gen[3]
(Artificial Intelligence Key Laboratory of Sichuan Province, Zigong 643000, China)[1]
(School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang 212003, China)[2]
(Key Lab of Intelligent Information Processing of Chinese Academy of Sciences(CAS),
Institute of Computing Technology, CAS, Beijing 100190, China)[3]

Abstract An improved estimation of distribution algorithm based on normal distribution was presented for function optimization in continuous space. The algorithm regards the selected individual as a normal distribution, and the random new populations of normal distribution are generated, and some selected individuals are crossed with the best solution. Compared with estimation of distribution algorithm based on uniform distribution and estimation of distribution algorithm based on normal distribution, the improved estimation of distribution algorithm based on normal distribution is more effective through result. At last, the influence of better population selection proportions was analyzed.

Keywords Estimation of distribution algorithm, Continuous space optimization, Normal distribution, Uniform distribution

1 引言

分布估计算法是进化计算领域新兴起的一类随机优化算法,是当前进化计算领域的研究热点[1]。实验分析表明分布估计算法在求解问题时表现出了比一般遗传算法更好的性能,应用分布估计算法解决工程和科学中的复杂优化问题具有很大的潜力,目前分布估计算法已经在众多领域得到了成功的应用[2-8],例如,基于分布估计算法的汽车齿轮结构的优化设计、采用贝叶斯优化算法进行特征选择、不精确图形的模式匹配、基于分布估计算法的软件测试、分布估计算法在癌症分类中的应用、生物信息学中的特征提取、军事天线的优化设计等。分布估计算法的应用已经渗透到了模式识别、工程优化、运筹学、机器学习和生物信息等众多领域,使用分布估计算法解决在科学研究和工程应用中碰到的优化问题将是未来研究的热点。传统的分布估计算法都是针对二进制编码问题的,在实际工程和科学研究中,研究定义域为实数的优化算法,解决连续域问题的分布估计算法有着重要的意义。连续是在离散的基础上发展起来的,很多算法的思想来源于或借鉴离散分布估计算法。由于目前没有一个有效的求解连续空间优化问题的算法,因此该问题一直是研究的热点,有许多文献对其进行了研究[9-17]。由于连续空间概率模型的复杂性给设计有效的分布估计算法增加了难度,本文提出一种改进的正态分布的分布估计算法来解决连续空间优化问题。

2 分布估计算法

分布估计算法的概念最初由 Muhliebe H 和 Paass G 在1996年提出,分布估计算法是在遗传算法基础上发展起来的一种全新的进化模式。在传统的遗传算法中,用种群表示优化问题的一组候选解,种群中的每个个体都有相应的适应值(目标值),然后根据适应值大小反复进行选择、交叉和变异等模拟自然进化的操作,对问题进行求解;而在分布估计算法中,没有传统的交叉、变异等遗传操作,取而代之的是概率模

到稿日期:2014-05-10 返修日期:2014-07-30 本文受人工智能四川省重点实验室开放基金(2012RYJ04),中科院智能信息处理重点实验室开放课题(IIP2013-1)资助。

邱 玲(1980—),女,硕士,讲师,主要研究方向为计算机应用;高 尚(1972—),男,博士,教授,主要研究方向为智能计算,E-mail:gaoshang@sohu.com;曹存根(1964—),男,博士,研究员,主要研究方向为大规模知识获取和共享、计算机辅助教学、情感计算等。

型的学习和采样，分布估计算法通过一个概率模型表示候选解在空间中的分布，采用统计学习手段从群体宏观的角度建立一个描述解分布的概率模型，然后对概率模型随机采样产生新的种群，如此反复进行，实现种群的进化，直到终止条件。根据概率模型的复杂程度以及不同的采样方法，分布估计算法发展了很多不同的具体实现方法，但是都可以归纳为下面两个主要步骤：首先建立描述解空间的概率模型，通过对种群的评估，选择优秀的个体集合，然后采用概率统计等手段构造一个描述当前解集的概率模型；然后由概率模型随机采样产生新的种群，一般采用蒙特卡罗方法，对概率模型采样得到新的种群。

遗传算法中的交叉和变异会破坏已经进化好的个体，分布估计算法用建立概率模型和采样样本两项操作取代了遗传算法中的交叉算子和变异算子，以一种带有“全局操控”性的操作模式弥补了遗传算法存在的这种缺陷。而且分布估计算法不需要太多的参数设置，编程比遗传算法简单。

分布估计算法的基本步骤如下(见图 1)：

步骤 1　在搜索空间内按均匀分布随机产生 N 个点，组成初始群体。

步骤 2　根据适应值评价函数计算群体中的各点的适应值，保留最好解。

步骤 3　根据适应值，运用一定的选择策略选出适应值较好的 m 个个体组成优势群体。

步骤 4　估计优势群体的概率分布模型。

步骤 5　根据估计的概率模型进行采样，产生一些新个体。

步骤 6　若满足某种停止准则，则算法结束，群体中的最好个体就是优化的结果；否则算法转到步骤 2 继续执行。

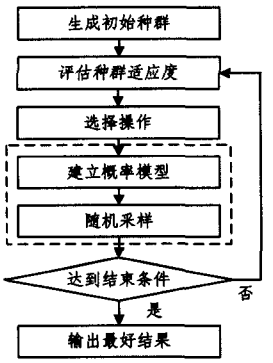

图 1　基本分布估计算法的流程

3　改进正态分布的分布估计算法

3.1　正态分布的分布估计算法

分布估计算法通过一个概率模型描述候选解在空间的分布，采用统计学习手段从群体宏观的角度建立一个描述解分布的概率模型，然后对概率模型随机采样产生新的种群。此概率模型具体是哪一种值得研究。正态分布(normal distribution)又名高斯分布(Gaussian distribution)，是一个在数学、物理及工程等领域都非常重要的概率分布，在统计学的许多方面有着重大的影响力。生产与科学实验中很多随机变量的概率分布都可以近似地用正态分布来描述。假如不清楚某随机变量服从何分布时，实际工程中经常用正态分布来描述。基于这个思想，这里概率模型选用正态分布。

设连续型随机变量 X 服从正态分布，其概率密度为

$$f(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}},-\infty<x<\infty \tag{1}$$

其中，$\mu,\sigma(\sigma>0)$ 为常数，记为 $X\sim N(\mu,\sigma^2)$。正态分布随机变量 X 的均值 $E(X)=\mu$，方差 $D(X)=\sigma^2$。

由概率论理论可知：μ,σ 的估计值分别为 $\hat{\mu}=\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$，$\hat{\sigma}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$。

由计算机仿真理论可知，设 u_1 和 u_2 是两个独立的(0,1)均匀分布随机数，正态分布 $X\sim N(\mu,\sigma^2)$ 的随机数为

$$\begin{cases}x_1=u+\sigma(-2\ln u_1)^{1/2}\cos 2\pi u_2\\ x_2=\mu+\sigma(-2\ln u_1)^{1/2}\sin 2\pi u_2\end{cases} \tag{2}$$

连续空间优化问题可表示为：

$$\begin{aligned}&\min f(x_1,x_2,\cdots,x_n)\\ &\text{s. t. } a_i\leqslant x_i\leqslant b_i,i=1,2,\cdots,n\end{aligned} \tag{3}$$

解连续空间优化问题的正态分布的分布估计算法如下：

步骤 1　变量 $x_i(i=1,2,\cdots,n)$ 在 $[a_i,b_i]$ 区间均匀随机取值，共产生 N 个个体组成一个初始种群；

步骤 2　评估初始种群中所有个体的适应度，保留最好解；

步骤 3　按适应度从高到低的顺序对种群进行排序，并从中选出最优的 m 个个体($m\leqslant N$)；

步骤 4　分析产生的 m 个个体所包含的信息，统计每个变量均值 $\hat{u}_i$ 和方差 σ_i；

步骤 5　按式(2)从构建的正态概率模型中采样，得到 N 个新样本，构成新种群；

步骤 6　若达到算法的终止条件则结束(如达到规定迭代次数 $n_{\max}$)，否则执行步骤 2。

3.2　改进的正态分布的分布估计算法

在自然界生物进化过程中起核心作用的是生物遗传基因的重组。遗传算法中起核心作用的是遗传操作的交叉操作。借鉴遗传算法中的交叉操作，改进思路是充分利用保留的最好解的信息，根据交叉概率 p_c，随机挑选 $N\times p_c$ 个个体与最好个体 $x_{\min}$ 进行交叉。所谓交叉操作是指把两个父代个体的部分结构加以替换重组而生成新个体的操作。通过交叉，遗传算法的搜索能力得以大幅提高。改进的交叉方法为

$$x^{new}=ax_{\min}+(1-a)x^{old} \tag{4}$$

其中，a 为[0,1]的随机数。

具体步骤如下：

步骤 1　变量 $x_i(i=1,2,\cdots,n)$ 在 $[a_i,b_i]$ 区间均匀随机取值，共产生 N 个个体组成一个初始种群；

步骤 2　评估初始种群中所有个体的适应度，保留最好解；

步骤 3　按适应度从高到低的顺序对种群进行排序，并从中选出最优的 m 个个体($m\leqslant N$)；

步骤 4　分析产生的 m 个个体所包含的信息，统计每个变量均值 $\hat{u}_i$ 和方差 $\hat{\sigma}_i$；

步骤 5　按式(2)从构建的正态概率模型中采样，得到 N 个新样本，构成新种群；

步骤 6　根据交叉概率 p_c，随机挑选 $N\times p_c$ 个个体与最好个体 $x_{\min}$ 按式(4)进行交叉；

步骤 7　若达到算法的终止条件则结束(如达到规定迭代次数 $n_{\max}$)，否则执行步骤 2。

4　数值仿真与分析

4.1　与均匀分布的分布估计算法比较

为了说明选择正态分布的优势，将其与均匀分布的分布

估计算法进行比较。均匀分布的分布估计算法如下：

步骤 1　变量 $x_i(i=1,2,\cdots,n)$ 在 $[a_i,b_i]$ 区间均匀随机取值，共产生 N 个个体组成一个初始种群；

步骤 2　评估初始种群中所有个体的适应度，保留最好解；

步骤 3　按适应度从高到低的顺序对种群进行排序，并从中选出最优的 m 个个体($m\leqslant N$)；

步骤 4　分析产生的 m 个个体所包含的信息，统计每个变量的最小值 $x_{\min}^{(i)}$ 和最大值 $x_{\max}^{(i)}$；

步骤 5　从构建的均匀分布概率模型中采样，变量 $x_i(i=1,2,\cdots,n)$ 在 $[x_{\min}^{(i)},x_{\max}^{(i)}]$ 区间均匀随机取值，得到 N 个新样本，构成新种群；

步骤 6　若达到算法的终止条件则结束(如达到规定迭代次数 $n_{\max}$)，否则执行步骤 2。

对以下 4 个经常被国内外学者用来测试优化算法有效性的测试函数进行计算[5]，结果如图 2 所示，这些函数是典型的非线性的多模态函数，具有高维难解的特点。

$$\min F_1=\sum_{i=1}^{30}x_i^2,-1\leqslant x_i\leqslant 1(i=1,2,\cdots,30)$$

$$\min F_2=\sum_{i=1}^{30}|x_i|+\prod_{i=1}^{30}|x_i|,-1\leqslant x_i\leqslant 1(i=1,2,\cdots,30)$$

$$\min F_3=\max_{1\leqslant i\leqslant 30}|x_i|,-1\leqslant x_i\leqslant 1(i=1,2,\cdots,30)$$

$$\min F_4=\sum_{i=1}^{30}(\sum_{j=1}^{i}x_j)^2,-1\leqslant x_i\leqslant 1(i=1,2,\cdots,30)$$

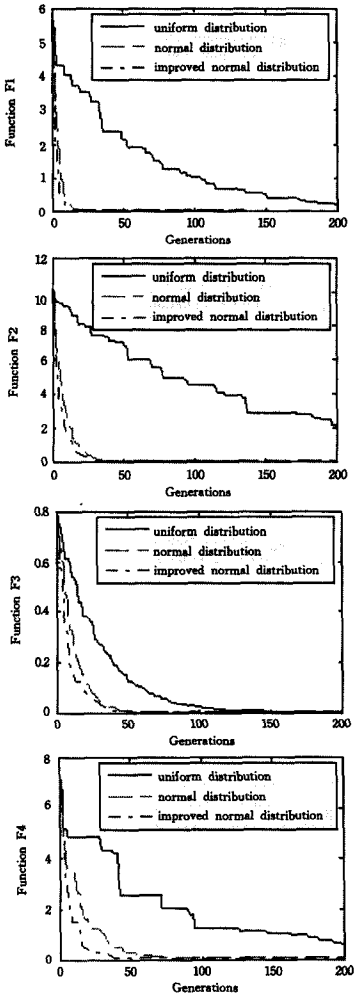

图 2　3 种算法的比较

3 种算法的参数均设为：$N=1000$，$m=0.4*N$，$p_c=0.01$，迭代 200 次。从图 2 可以看出，改进正态分布的分布估计算法的收敛速度比均匀分布的分布估计算法的收敛速度快很多，其次是正态分布的分布估计算法的收敛速度。因为改进的正态分布的分布估计算法充分利用了数据的均值和方差信息，以及保留的最好解的信息，所以其收敛速度比较快。改进的交叉操作使算法具备兼顾全局和局部的均衡搜索能力。

4.2　不同的 m/N 比例

对于测试函数[5]

$$\min F_5=\sum_{i=1}^{30}[x_i^2-10\cos(2\pi x_i)+10]$$

$$\text{s.t.}\ -5.12\leqslant x_i\leqslant 5.12(i=1,2,\cdots,30)$$

当 $x_1=x_2=\cdots=x_{30}=0$ 时，$F_{\min}=0$。

影响分布估计算法的性能的主要参数是种群的个数 N 和选出的种群个数 m。当 $N=1000$ 时，选出的种群个数 m 占 N 的比例不同，结果也不一样。计算测试函数 F_5，以精度为 0.000001 时所需要的迭代次数作为比较。按不同的比例，算法各测试 50 次，统计数据如表 1 所列。由于当 $m/N=10\%$ 时，F_5 有可能不收敛，因此数据无法统计全。

表 1　m/N 取不同值时的结果比较

算法	m/N	最多迭代次数	最少迭代次数	平均迭代次数
粒子群优化算法	—	245	110	191
正态分布的分布估计算法	10%	—	88	—
	20%	128	108	117
	30%	145	132	137.4
	40%	173	154	163
	50%	215	193	199.1
改进的正态分布的分布估计算法	10%	—	83	—
	20%	115	107	112
	30%	145	129	133.8
	40%	172	152	162.3
	50%	202	185	195.1

从表 1 可以看出，两种算法的规律相近。m/N 比例越大，不能体现出选优的优势，效果越不好；当然 m/N 比例太小，挑选的样本数太少，也容易陷入局部极值。因此 m/N 比值取 20%～40%时，效果比较好。表 1 中也与粒子群优化算法[17]作了比较，其参数设置如下：$c_0=2$，$c_1=1$，$c_2=1$，粒子数为 50，结果表明改进的分布算法比较有效。

结束语　本文提出一种改进的正态分布的分布估计算法来求解连续空间优化问题，该算法具有较强的搜索能力和优化效率，尤其适用于复杂函数优化和高维优化问题。实验结果表明，该方法是有一定潜力的、值得推荐的优化方法。对本文采用的改进的正态分布的分布估计算法可以作进一步改进，如加入变异操作，使算法可维持群体多样性，或与其他一些智能算法混合，性能可能会更好，这也是下一步改进的思路。

参 考 文 献

[1]　周树德，孙增圻. 分布估计算法综述[J]. 自动化学报，2007，33(2)：113-124
Zhou Shu-de，Sun Zeng-qi. A Survey on Estimation of Distribution Algorithms[J]. Acta Automatica Sinica，2007，33(2)：113-124

[2]　Muhliebe H，Paass G. From recombination of genes to the estimation of distributions I. binary parameters[C]// Lecture notes

in computer science. Berlin, Germany: Springer Verlag, 1996, 1141:178-187

[3] Pelikan M, Godberg D E, paz E C. Linkage problem, distribution estimation, and Bayesian networks[J]. Evolutionary Computation. 2000, 8(3):311-340

[4] Paul T K, Iba H. Linear and combinatorial optimizations by estimation of distribution algorithms[C]// 9th MPS Symposium on Evolutionary Computation, IPSJ. Japan, 2002

[5] 梁玉洁,许峰. 自适应混合多目标分布估计进化算法[J]. 计算机工程与应用,2014,50(5):46-50,207
Liang Yu-jie, Xu Feng. Adaptive hybrid multi-objective estimation of distribution evolutionary algorithm [J]. Computer Engineering and Applications, 2014, 50(5):46-50, 207

[6] 戚玉涛,刘芳,刘静乐,等. 基于免疫算法和 EDA 的混合多目标优化算法[J]. 软件学报,2013,24(10):2251-2266
Qi Yu-tao, Liu Fang, Liu Jing-le, et al. Hybrid immune algorithm with EDA for multi-objective optimization[J]. Journal of Software, 2013, 24(10):2251-2266

[7] 丁有军,钟声. 基于分布估计算法的连续函数全局优化问题研究[J]. 计算机科学,2012,39(10):218-219,223
Ding You-jun, Zhong Sheng. Global Optimization Problem of Continuous Function Based on Distribution Estimation Algorithm[J]. Computer Science, 2012, 39(10):218-219, 223

[8] 周雅兰,朱耀辉,张军. 具有学习机制的离散差分演化算法[J]. 计算机科学,2011,38(7):225-227,249
Zhou Ya-lan, Zhu Yao-hui, Zhang Jun. Discrete Differential Evolution with Learning Mechanism[J]. Computer Science, 2011, 38(7):225-227, 249

[9] 王凌. 智能优化算法及其应用[M]. 北京:清华大学出版社,2004
Wang Ling. Intelligent optimization algorithm with applications [M]. Beijing: Tsinghua University Press, 2004

[10] 李盼池,李士勇. 求解连续空间优化问题的混沌量子免疫算法[J]. 模式识别与人工智能,2007,20(5):654-660
Li Pan-chi, Li Shi-yong. A Chaos Quantum Immune Algorithm for Continuous Space Optimization[J]. PI & AI, 2007, 20(5): 654-660

[11] 寇晓丽,刘三阳,张建科. 一种随机蚁群算法求解连续空间优化问题[J]. 系统工程与电子技术,2006,28(12):1909-1911
Kou Xiao-li, Liu San-yang, Zhang Jian-ke. Stochastic ant colony algorithm for continuous space optimization[J]. Systems Engineering and Electronics, 2006, 28(12):1909-1911

[12] 张锐,高辉,张涛. 求解连续空间优化问题的量子差分混合优化算法[J]. 系统工程与电子技术,2012,34(6):1288-1292
Zhang Rui, Gao Hui, Zhang Tao. Hybird optimization algorithm based on quantum and differential evolution for continuous space optimization[J]. Systems Engineering and Electronics, 2012, 34(6):1288-1292

[13] 郭源源,王谦,梁峰. 基于粒子群优化算法的车间布局设计[J]. 计算机集成制造系统,2012,18(11):2476-2484
Guo Yuan-yuan, Wang Qian, Liang Feng. Facility layout design based on particle swarm optimization[J]. Computer Integrated Manufacturing Systems, 2012, 18(11):2476-2484

[14] 黄敏,靳婷,钟声,等. 基于改进蚁群算法求解连续空间寻优问题[J]. 广西师范大学学报(自然科学版),2013,31(2):34-38
Huang Min, Jin Ting, Zhong Sheng, et al. Ant Colony Algorithm for Solving Continuous Function Optimization Problem Based on Pheromone Dis tributive Function [J]. Journal of Guangxi Normal University(Natural Science Edition), 2013, 31(2):34-38

[15] 马卫,朱庆保. 求解函数优化问题的快速连续蚁群算法[J]. 电子学报,2008,36(11):2120-2124
Ma Wei, Zhu Qing-bao. Fast Continuous Colony Optimization Algorithm for Solving Function Optimization Problems[J]. Acta Electronica Sinica, 2008, 36(11):2120-2124

[16] 张腾飞,王锡淮,肖健梅. 基于微粒群优化的连续属性离散化算法[J]. 计算机工程,2006,32(3):44-46
Zhang Teng-fei, Wang Xi-huai, Xiao Jian-mei. Algorithm for Discretization of Continuous Attributes Based on Particle Swarm Optimization[J]. Computer Engineering, 2006, 32(3):44-46

[17] 高尚,杨静宇. 群智能算法及其应用[M]. 北京:中国水利水电出版社,2006
Gao Shang, Yang Jing-yu. Swarm intelligence algorithm and its application[M]. Beijing: China Water and Power Press, 2006

(上接第 27 页)

[32] 冯松鹤. 面向感知的图像检索及自动标注算法研究[D]. 北京:北京交通大学,2009
Feng S H. Research on perception oriented image retrieval and automatic image annotation[D]. Beijing: Beijing Jiaotong University, 2009

[33] Cheng M M, Zhang G X, Mitra N J, et al. Global contrast based salient region detection[C]// 2011 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). IEEE, 2011:409-416

[34] Zha Z J, Wang M, Zheng Y T, et al. Interactive video indexing with statistical active learning[J]. IEEE Transactions on Multimedia, 2012, 14(1):17-27

[35] Zha Z J, Yang L, Mei T, et al. Visual query suggestion[C]// Proceedings of 17th ACM International Conference on Multimedia. ACM, 2009:15-24

[36] Zha Z J, Hua X S, Mei T, et al. Joint multi-label multi-instance learning for image classification[C]// IEEE Conference on Computer Vision and Pattern Recognition, 2008 (CVPR 2008). IEEE, 2008:1-8

[37] Yang Y, Yang Y, Shen H T. Effective transfer tagging from image to video[J]. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP), 2013, 9(2):14

[38] Chen Y, Bi J, Wang J Z. MILES: Multiple-instance learning via embedded instance selection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(12):1931-1947

[39] Yakhnenko O, Honavar V. Multi-Instance Multi-Label Learning for Image Classification with Large Vocabularies[C]// BMVC. 2011:1-12

[40] Järvelin K, Kekäläinen J. Cumulated gain-based evaluation of IR techniques [J]. ACM Transactions on Information Systems (TOIS), 2002, 20(4):422-446

[41] Baeza-Yates R, Ribeiro-Neto B. Modern information retrieval [M]. New York: ACM press, 1999

[42] Lan T, Mori G. A Max-Margin Riffled Independence Model for Image Tag Ranking[C]// 2013 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). IEEE, 2013:3103-3110