基于分布估计算法的连续函数全局优化问题研究

丁有军 钟 声
(海南大学信息科学技术学院 海口 570228)

摘 要 分布估计算法从宏观的角度建立一个概率模型,用来描述解空间的分布,从而通过进化计算获得优势个体。目前,离散型分布估计算法研究已经比较成熟,而连续型分布估计算法研究进展缓慢。采用均匀分布缩小采样领域的思想,设计新的分布估计算法求解连续型优化问题。实验数据表明,该分布估计算法对于求解连续型问题是有效的。

关键词 分布估计算法,均匀分布,函数优化
中图法分类号 TP39 **文献标识码** A

Global Optimization Problem of Continuous Function Based on Distribution Estimation Algorithms

DING You-jun ZHONG Sheng
(College of Information Science & Technology, Hainan University, Haikou 570228, China)

Abstract The distribution estimation algorithms use statistical learning to create a probability model from a macro point of view, use the probability model to describe the distribution of problem solution in the solution space, and obtain the advantage of individuals by evolutionary computation. At present, the discrete distribution estimation algorithms are already quite mature, but receach progresses of continuous distribution estimation algorithm are slow. This article used the idea of uniform distribution to narrow the sampling field for continuous optimization problems, designed a new distribution estimation algorithm. Experimental data show that this kind of distribution estimation algorithms are valid.

Keywords Distribution estimation algorithms, Uniform distribution, Function optimization

1 引言

传统的遗传算法中,寻找种群的最优解时,采取的做法主要有:初始化种群,选择优势种群,对选择的优势种群进行交叉和变异运算,进行终止条件判断,若条件符合,则终止进化运算,否则继续进化运算的循环操作。

遗传算法的不足之处有,一是基因位之间的关系:交叉和变异的个体被看作染色体,它们的基因之间的关系在一些问题中是独立的,然而大多数问题变量个体间具有一定的关联性。二是局部最优问题:遗传算法中使用父代个体的基因重组来进行进化操作,容易产生连锁反应,导致算法在局部收敛,达不到最优解的条件要求。

为解决遗传算法的不足,近年来在进化计算领域提出了多种新方法,分布估计算法就是其中一种新的解决工程问题的有效算法。相比传统的遗传进化算法,分布估计算法采用全新的进化模式,在分布估计算法中,对选择的优势群体不再采用交叉变异等运算操作,而采用统计学习的手段,从宏观的角度建立一个概率模型,这个概率模型用来描述解空间的分布,然后对概率模型进行采样,产生新的种群,在此种群的基础上再次进行优势种群的选择,依此循环,直到产生最终解的条件得到满足[1]。

2 分布估计算法

分布估计算法是从进化的宏观层面上进行一种数学建模,对群体建立一种概率模型,用此概率模型来描述群体的数学上的分布,然后对概率模型进行采样,产生新的种群。使用数学上的概率模型作为描述种群分布的工具,相比传统的遗传进化算法,分布估计算法在克服局部解收敛、解决非线性、变量耦合的复杂优化问题上大有作为。

分布估计算法的核心主要有以下两个步骤[2]:

1. 构建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,然后采用统计学习等手段构造一个描述当前解集的概率模型。

2. 由概率模型随机采样产生新的种群。一般地,采用蒙特卡罗方法对概率模型进行采样得到新的种群。

目前应用最多的是针对离散型问题的分布估计算法,使用一个概率向量表示解的分布。比较著名的算法有 PBIL、UMDA、cGA 算法[1-3],此外,MIMIC 算法、COMIT 算法、BMDA 算法[4-6]也是常见的针对离散型区域变量的分布估计算法。

由于实际问题所限制,在实际问题中,变量的取值范围多为连续的区域,即为连续型变量问题。连续型变量的分布估

到稿日期:2011-12-05 返修日期:2012-04-01 本文受海南省自然科学基金项目(611121)资助。
丁有军(1985-),男,硕士生,CCF 会员,主要研究方向为复杂系统优化、软件方法学,E-mail:taibaotuzi@163.com;钟 声(1962-),男,博士,教授,CCF 会员,主要研究方向为复杂系统优化、软件方法学。

计算法中，目前的研究进展缓慢，这类算法比较有代表性的有ECGA(Extended Compact Genetic Algorithm)算法[7]、FDA(Factorized Distribution Algorithms)算法[8]、贝叶斯优化算法(Bayesian Optimization Algorithm，BOA)[9]。

3 分布估计算法在连续型问题中应用的分析

在实际工程问题中，变量的取值范围通常是连续域，并且伴随着变量之间复杂的依赖关系，这给更新概率模型带来了一定的困难，连续空间概率模型的设计研究目前进展缓慢。笔者采用反向思维缩小采样领域的思想，探求其在连续域分布估计算法中的应用。

具体来说，在分布估计算法中，改变调整概率模型的做法，仅借助均匀分布作为概率模型来对种群进行描述，保持概率模型不变。对算法的进化提出两个创新，首先，保持概率模型不变，取而代之，对连续变量重新采样时，针对采样的区间做一个调整，下次采样的区间仅定义在包含之前优势群体的最小空间范围中。如此，则依据上一代的优势群体，会不断缩小下一代的采样区间，以保证采样点所在空间的收敛逐渐朝向最优解空间领域。其次，每一代的优势群体在下一代的重新采样时，不再全部被替换，而是保留部分，以此确保进化的方向朝着优势解。如此循环，直至达到符合最优解的条件。

4 基于均匀分布的连续域分布估计算法的设计

需要优化的目标函数分析：函数的维度 D，决定了分布估算法中每个个体的基因位个数。函数的定义域即为每个个体基因位的取值范围。函数的值即是分布估计算法中每个个体的适应值。在每个维度下取得的相应个体的总和即为分布估计算法中的种群规模。

采用均匀分布缩小采样领域的思想，基于分布估计算法的连续域函数优化问题的算法设计如下，其中 OptFound 代表总群规模，dimension 代表维度：

1) 在函数搜索领域内随机等可能地产生初始群体 OptFound；

2) 分别计算 OptFound 中各个个体的适应值，并选择优势群体 tempOptFound；

3) 由 tempOptFound 作为样本集构造每一维 x 新的经过缩小的搜索区间；

4) 保留部分优势个体，同时在每个 x 的新的缩小后的搜索领域内等可能地采样，产生新的个体，与保留的优势个体共同构造新的种群；

5) $t=t+1$，如果终止条件不满足，转 2)；否则，算法结束。

5 实验分析

选取 3 个基准函数[10]用来测试，包括 1 个单峰函数，2 个多峰函数。基准函数是国际上为进化计算算法测试而提供的一系列标准测试函数(寻找最小值)，于 2005 年在 Nan Yang Technological University 的一个 Technical Report 上被提出，专门用于实数优化算法测试。种群规模采用与维度成线性关系，具体为 $25*2D$，更新种群时，保留的优势个体数目与维度成线性关系 $5*D$。

1. 单峰函数：$F_1(x)=\sum_{i=1}^{D} z_i^2+f\_bias_1$

其中，$Z=X-O$，$X=[x_1,x_2,\cdots,x_D]$。

2. 多峰函数：

$$F_2(x)=\sum_{i=1}^{D-1}(100(z_i^2-z_{i+1})^2+(z_i-1)^2+f\_bias_2)$$

其中，$Z=X-O+1$，$X=[x_1,x_2,\cdots,x_D]$。

$$F_3(x)=\sum_{i=1}^{D}(z_i^2-10\cos(2\pi z_i)+10)+f\_bias_3$$

其中，$Z=X-O$，$X=[x_1,x_2,\cdots,x_D]$。

表达式中，D 代表维数，O 代表 X 取得最优值时的坐标，当 $X^*=O$ 时即取得最优值。f_biasi 是常数，当 $X=O$ 时，取得的最优值，应是 f_biasi 的值。

分别对 3 个函数选取 2 维、5 维、9 维，Matlab 仿真数据如表 1—表 3 所列，其中种群规模属性中，分子代表优势种群；分母代表种群规模，保留两位小数后；解次数指在优势种群中得到的误差<+0.05 的最优解个数，即个体在最优种群中出现的次数。

表 1 F1 函数的实验结果

维	种群规模	平均值	最优值	解次数	首次最优解代数
2	50/100	−449.9858	−450	14	19
5	400/800	−449.9994	−450	104	400
9	6400/12800	−441.4699	−450	1	10000

表 2 F2 函数的实验结果

维	种群规模	平均值	最优值	解次数	首次最优解代数
2	50/100	390.9368	390	11	60
5	400/800	392.4738	390	18	600
9	6400/12800	406.9599	390	25	30000

表 3 F3 函数的实验结果

维	种群规模	平均值	最优值	解次数	首次最优解代数
2	50/100	−329.9986	−330	17	60
5	400/800	−329.9999	−330	78	1000
9	6400/12800	−323.4991	−330	17	20000

实验数据表明：在进化代数一定的条件下，更新种群时，选择合适的要保留的优势个体数目有利于缩小采样领域，从而达到解的快速收敛；进化代数一定的条件下，选择适当的优势种群与种群规模的比例，也有利于缩小搜索领域，加速解的快速收敛；而在种群规模与优势种群个体数目一定的条件下，随着进化代数的增多，逐步收敛至最优解，误差完全能够达到设定的要求，当代数达到一定程度时，将产生毫无误差的最优解。

函数 F1 是单峰函数，只有一个极值点，使用本算法能够快速收敛至最优解。函数 F2、F3 有多个极值点，在更新种群采用保留固定的优势个体方式时，尤其是 F2，有一些波动现象，采用保留与维度成线性变化的优势个体时，算法比较稳定，多次运行结果表明，其已经不存在波动现象，高维时需要较多的进化代数才能收敛至最优解。

对于一个连续域单峰函数和多峰函数寻优来说，基于均匀分布缩小采样领域的连续域分布估计算法是可行的，进化到一定代数，完全能达到最优解。然而高维时存在进化代数多的问题，由于连续域分布估计算法研究进展缓慢，此算法还有待在适应值计算方法、均匀采样区间构造方式、更新种群时的优势个体保留数目等方面做进一步的改进。

结束语 分布估计算法给进化计算领域提供了一个新的工具，通过统计学习建立概率模型，从宏观上描述问题目标变

(下转第 223 页)

结束语 在研究句子对齐方法时，结合古籍文献独有的特点，计算经典原文与其注疏文献的句子的相似度，使用动态规划算法实现了句子对齐。注释语句自动分析研究是在实现了句子对齐的基础上进行的。用正则表达式对常用的训诂术语做抽象化概括，通过建立训诂术语的模式库来实现注释语句的自动分析。

以上述两项研究成果所构建的语料库[10]为基础，设计了经典古籍与其注疏文献对齐语料库检索平台(见图1)。利用该平台，语言研究者可以很方便地检索到经典古籍中的某一注释对象，在与其相关的各个注疏文献中的注释，便于语言研究者进行综合比较。

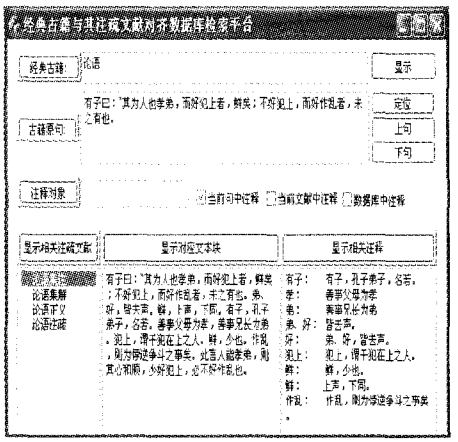

图1 经典古籍与其注疏文献对齐数据库检索平台

参 考 文 献

[1] Gale W A, Church K W. A Program for Aligning Sentences in Bilingual Corporal[J]. Computational Linguistics, 1993, 19(1): 75-90

[2] Champollion M X. A Robust Parallel Text Sentence Aligner [C]// Proceedings of LREC-2006: Fifth International Conference on Language Resources and Evaluation. 2006: 489-492

[3] Moore R C. Fast and Accurate Sentence Alignment of Bilingual Corpora [C]// Proceedings of AMTA. Springer-Verlag, 2002: 135-144

[4] Nirenburg S. Two Approaches of Matching in Example-Based Machine Translation[C]// Proc. TMT-93. Kyoto, Japan, 1993

[5] Li S, Zhang J, et al. Semantic Computation in Chinese Question-Answering System[J]. Journal of Computer Science and Technology, 2002, 17(6): 933

[6] 郭锐，宋继华，廖敏. 基于自动句对齐的相似古文句子检索[J]. 中文信息学报，2008，22(2)：87-91

[7] 于新，吴健，洪锦玲. 基于词典的汉藏句子对齐研究与实现[J]. 中文信息学报，2011，25(4)：57-62

[8] 许威汉. 训诂学读本[M]. 上海：上海交通大学出版，2010：48-84

[9] Watt A. 正则表达式入门经典[M]. 李松峰，李丽，译. 北京：清华大学出版社，2008：156-178

[10] 丁溪源，黄河燕，张海军，等. 基于大规模语料划分的频繁模式查找算法[J]. 计算机科学，2012，39(3)：149-152

(上接第219页)

量之间的关系，为解决非线性、多变量耦合问题提供了新的思路。目前，离散型分布估计算法研究已经比较成熟，而连续域分布估计算法的研究比较缓慢，这方面的资料很少。

本文根据连续域优化问题的特征以及分布估计算法的内涵，探索研究了采用均匀分布作为概率模型、保持概率模型不变缩小采样领域以及保留优势个体确保进化方向等思想设计了针对连续域函数优化的分布估计算法。

连续域分布估计算法的难点一方面是变量的取值问题，即由于变量是连续域，在取值范围内有无限种取值方法，编码困难，并且使得优化算法的搜索空间非常大，因此获得更好的采样区间是下一步的研究方向；另一方面，种群的规模、优势种群的规模、更新时保留的个体规模等也是下一步要研究的方向，其它的诸如连续域分布估计算法的采样算法、算法的收敛特性也比较有研究的价值。

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

[2] 许昌，常会友，徐俊. ASON网中基于分布估计的恢复容量优化算法[J]. 计算机科学，2010，37(7)：183-185

[3] Salimas-Gutierrez R, Hernandez-Aguirre A, Villa-Dihasrce E R. Dependence Trees with Copula Selection for Continous Estimation of Distribution Algorithms[C]//GECCO '11

[4] Marti L, Garcia J, Berlanga A, et al. On the Computational Properties of the Multi-Objective Neural Estimation of Distribution Algorithm[C]// Nature Inspired Cooperative Strategies for Optimization(NICSO 2008). 2009: 239-251

[5] Godingho P, Meiguins A, Oliveira R, et al. An Estimation of Distribution Algorithms Applied to Sequence Pattern Mining[C]// Innovations in Computing Sciences and Software Engineering. 2010: 589-593

[6] Salinas-Gutierrez R, Hernandez-Aguirre A, Villa-Diharrce E R. Estimation of Distribution Algorithms based on Copula Functions[C]//GECCO '11

[7] Salinas-Gutierrez R, Hernandez-Aguirre A, Villa-Diharce E R. Dvine EDA: A new Estimation of Distribution Algorithms based on Regular Vineas[C]//GECCO '10

[8] Lima C, Pelikan M, Goldberg D, et al. Influence of selection and replacement strategies on linkage learning in BOA[C]// Evolutionary Computation, 2007 (CEC 2007). IEEE Congress, Washington DC: IEEE, 2008: 1083-1090

[9] Naeem M, Lee D. Estimation of Distribution algorithm for sensor selection problems [C] // Radio and Wireless Symposium (RWS), 2010 IEEE. Washington DC: IEEE, 2010: 388-391

[10] Suganthan P N, Hansen N, Liang J J, et al. Problem Definitions and Evaluation Criteria for the CEC 2005 Special Session on Real-Parameter Optimization[R]. 2005