doi:10.3969/j.issn.1001-4616.2022.03.001

基于两点模型选择的离线数据驱动进化优化算法

包建阳[1],吕秋月[1],孙越泓[1,2]

(1.南京师范大学数学科学学院,江苏 南京 210023)
(2.江苏省大规模复杂系统数值模拟重点实验室,江苏 南京 210023)

[摘要] 基于两点模型选择的离线数据驱动进化优化算法主要用于解决目标计算复杂度高的离线优化问题. 在模型建立过程中,建立多个代理模型,而后运用模型选择策略,从中选择部分代理模型,组成集成模型. 同时,模型选择策略概率被采用,用来提高算法通用性和减少时间复杂度. 该算法在常见的基准测试函数上进行了数值实验,与其他先进的算法进行了比较,实验结果表明,新算法更具有优势.

[关键词] 集成学习,进化算法,离线数据驱动优化,代理模型
[中图分类号]TP18 [文献标志码]A [文章编号]1001-4616(2022)03-0001-08

Offline Data-Driven Evolutionary Algorithm Using Best Point and Uncertainty Point to Guide Model Selection

Bao Jianyang[1], Lü Qiuyue[1], Sun Yuehong[1,2]

(1.School of Mathematical Sciences, Nanjing Normal University, Nanjing 210023, China)
(2.Jiangsu Provincial Key Laboratory for Numerical Simulation of Large Scale Complex Systems, Nanjing 210023, China)

Abstract: The off-line data-driven evolutionary optimization algorithm based on two-point model selection is mainly used to solve off-line optimization problems with high computational complexity. In the process of model establishment, several agent models are established, and then some agent models are selected by model selection strategy to form an integrated model. At the same time, the probability of model selection strategy is adopted to improve the generality of the algorithm and reduce the time complexity. Numerical experiments are carried out on common benchmark test functions, and the experimental results show that the new algorithm has more advantages than other advanced algorithms.

Key words: ensemble learning, evolutionary algorithm, offline data-driven optimization, surrogate model

如今,进化算法已经广泛应用到解决实际优化问题上,并且效果良好. 但是当问题的目标函数或者约束函数计算需要花费几个小时或者几天时,进化算法的性能下降,原因在于进化算法的迭代过程中,需要成千上万次的函数值计算. 为了解决此类昂贵问题,学者们提出了代理辅助进化算法(surrogate-assisted evolutionary algorithms, SAEAs),即使用廉价的代理模型或元模型替代昂贵的目标函数或者约束函数. 这样很大程度上减少了昂贵问题进化优化的计算消耗. 常见的代理模型有:径向基函数(radial basis function, RBF)[1]、高斯过程(gaussian process, GP)[2]、支持向量机(support vector machine, SVM)[3]、多项式回归(polynomial regression, PR)[4]等.

尽管代理模型的使用可以一定程度上减少计算消耗,但在一些高维昂贵问题上仍难以达到想要的结果. 同时在精确度和计算资源消耗上,也很难达到理想中的平衡点. 因此,模型管理在这方面起到了很重要的作用,其作用是选择合适的代理模型以及选择出有代表性的个体或种群进行真实适应度值评估. 常见的代表性特征有预测最优适应度值、最大不确定性. 同时,机器学习中的集成学习和主动学习策略也经常被采用,来提高代理辅助进化算法的效率.

由于代理模型是在历史数据的基础上建立,所以代理辅助进化优化也被称为数据驱动进化优化[5](data-driven evolutionary optimization, DDEA). 此外,由于某些问题中可以允许少量的新数据加入,因此数

收稿日期:2021-10-27.
基金项目:国家自然科学基金项目(11871279).
通讯作者:孙越泓,博士,副教授,研究方向:智能优化及图像处理. E-mail:05234@njnu.edu.cn

据驱动进化优化可以分为在线数据驱动优化和离线数据驱动优化. 对于在线数据驱动,由于有新样本点的加入,可以在优化过程中对代理模型进行改进,提高精确度. 但在实际解决真实问题时,需要平衡模型精确度和计算消耗,因此主要的困难在于以下几点:

(1)样本数据的选择:对于同一个样本数据集合,选择不同的样本子集进行训练,会得到不同的模型,其误差会有一定的差别. 原始数据是具有自身的概率分布,理想状态下,是进行分层采样,保证数据的平衡性、完全性. 但往往数据样本的概率分布的先验经验是未知的,因此样本数据的选择对于模型的精确度有一定的影响.

(2)模型的选择:不同的代理模型有不同的特点,上面提到的几种代理模型各有各的特点. 例如,高斯过程精确度高,但是计算消耗相对于其他代理模型较大,此外还可以提供模型预测的方差,且参数少,因此被经常使用. 径向基函数模型简单,局部逼近网络,学习收敛速度快,泛化能力强,但精确度较大地依赖于中心点的选择以及核函数设置的形状参数.

(3)新样本点的选择:在在线数据驱动优化中,允许有新样本点的加入,但选择什么样的样本点进行真实预测变得很重要. 往往选择具有代表性的点,如:性能预测最优点或者不确定性最大点. 若选择性能预测最优点,则有利于勘探有希望的区域,但全局的勘测能力较差;若选择不确定性最大点,则全局勘测性能提高,但是局部勘探效率较差.

对于样本数据的选择,王晗丁等[6]提出了一个以数据分布性为标准的采样算法,保证了数据的分布性,保留了部分的好点. 对于模型的选择,由于高斯模型能提供预测值的方差,即不准确性,从而被广泛使用,但其计算消耗大. 因此郭丹等[7]为了解决这一问题,提出了一个异构集成的算法. 在算法中,先对样本数据进行两种不同的降维处理,加上原始数据,得到 3 种不同的样本集,在每个样本集上分别建立 3 种不同的代理模型,而后以模型在点上预测值的均值作为输出值,以预测值的方差作为不确定性指标,来替代高斯过程提供的方差. 此外,集成思想已广泛应用于数据驱动研究,大多数处理集成的操作是以模型的均值为输出值,王晗丁等[6]提出一个以加权求和为输出值的算法. 在新样本点的选择策略上,学者们做了大量的研究,这也是模型管理的主要内容. 常见的选择策略或标准是性能最优点和最大不确定性点. 郭丹等[7]提出,通过建立 9 个不同代理模型来预测个体的适应度值,以模型预测的均值作为个体适应度值,以模型预测的方差作为不确定性程度,选择出个体适应度值方差最大的点进行真实评估. 还有很多学者将上述两种策略结合,如期待改变量(expected improvement, EI)[8],置信下限(lower confidence bound, LCB)[9]和提高的可能性(probability of improvement)[10],常见于高斯过程辅助进化算法中.

而对于离线数据驱动优化而言,因为在优化过程中没有新样本点加入,无法对模型的精确度进行修正,因此问题的最终解很大程度依赖于代理模型的建立. 而代理模型的建立又极大地依赖于历史数据. 因此处理离线数据驱动优化的难度继续增加,除了在线数据驱动优化需要解决的难点,还需要对数据进行预处理. 由于数据收集状态的不稳定性,得到的数据往往是不完全的、不平衡的、有噪声的,所以在优化开始前需要对数据进行一定的预处理. 如,在一个多目标高炉问题,Chugh 等[11]通过局部回归方法降低离线数据噪声. 由于没有新样本点的加入,数据量也影响着代理模型. 当数据量大时,采用全部的数据会增加计算消耗;当数据量小时,代理模型的精确度不高. 为了应对数据量大的情况,王晗丁等[12]采用聚类方法识别出有用的数据集建立代理模型,减少了 90%的计算消耗. 同样,对于数据量不足时,郭丹等[13]通过低阶多项式方法产生人工数据来扩充历史数据. 上述算法虽然效果好,但是泛化能力弱,需要提前了解问题和数据集. 为此,王晗丁等[14]提出了一个算法,能够应对不同的离线数据驱动优化问题. 该算法先通过多次随机抽样建立大量的代理模型,而后在优化过程中通过上一代的最优点,自适应地选择部分模型集合,对个体进行预测.

本文是在王晗丁提出的算法 DDEA-SE[14]的基础上进行一定的改进. 在 DDEA-SE 中,主要解决的问题是如何从大量的代理模型中选择部分代理模型. 王的做法是通过上一代最优点作为标准. 本文基于 DDEA-SE,将上一代最优点和上一代最不确定点同时作为模型的选择标准,提出算法 DDEA-BUS.

本文剩余的结构如下:先简单介绍相关知识;然后介绍算法的框架和细节;再通过数值实验证明算法的有效性;最后总结了全文.

1 相关知识

1.1 径向基函数

径向基函数是一种单隐层、结构简单、以函数逼近为基础的前馈神经网络,其非线性逼近能力强,具有良好的推广能力,且时间复杂度小,参数少. 径向基函数共有 3 层,第一层是输入层,第二层为隐藏层,第三层则为输出层. 输入层到隐藏层之间的权重均为常数 1,隐藏层则是由使用径向基函数作为激活函数的神经元组成,隐藏层与输出层之间的连接权值则通过训练样本集得到.

假设数据集为$\{(x_i,y_i)\mid i=1,2,\cdots,N\}$,其中 x_i 和 y_i 分别表示第 i 个样本的输入值和输出值,N 表示样本的个数. 径向基函数网络具体形式为:

$$\bar{f}(x)=\sum_{i=1}^{K}\omega_i\varphi(\|x-c_i\|).$$

式中,c_i 表示第 i 个中心,K 表示中心个数,$\varphi(x)$表示径向基函数. 对于广义径向基函数网络,其中心点是样本点,中心点个数等于样本集的大小,即样本集中每个样本都看成中心. 径向基函数是一类以两点之间的欧式距离为自变量的函数,反映两点之间的联系.

径向基函数网络假设在中心点上的预测值是无偏的,即$f(c_i)=\bar{f}(c_i)$,$i=1,2,\cdots,K$,则广义径向基函数网络在所有样本点上是无偏估计,所以建立径向基函数网络就是求解线性方程组

$$\Phi\omega=F.$$

式中,$\Phi=\Phi_{i,j}=\varphi(\|x_i-x_j\|)$,$F=(f_1,f_2,\cdots,f_N)$. 权重 ω 为上述方程的最小二乘解.

1.2 集成学习

集成学习是机器学习中经常使用的方法之一. 集成学习的一般结构为:建立一组不同的模型,采用某种组合策略将他们结合起来. 当某一个模型在一个区间内误差大时,可以通过其他模型对其进行纠正. 此外,由于建立模型的样本点或方式不同,得到的每个模型偏好性不同,在使用组合策略后,每个模型的偏好会被中和,减少了结合后的模型的偏好性.

解决离线数据驱动优化问题时,建立的代理模型不一定需要精确度很高,而是倾向于在变化趋势上与原目标函数尽可能保持一致. 又由于集成学习能有效减少模型的偏好,保证模型在整个空间内精确度相差不大,在贴合原函数的变化趋势效果上更好. 所以,采用集成学习是解决本问题的一个有力方法.

2 算法框架

在本章节,基于两点模型选择的离线数据驱动进化优化算法(offline data-driven evolutionary algorithm using best point and uncertainty point to guide model selection,DDEA-BUS)被提出. 新算法提出了一个新的模型选择策略和策略池选择概率. 模型选择策略以径向基函数为基础,结合集成学习,其中径向基函数网络能够逼近任意的非线性函数,而集成学习可以有效地提高模型的泛化能力,减少模型的误差和错误率;而策略池选择概率是基于优化问题的特征,可以使算法适用于更多的问题. 以下将详细介绍模型选择策略和策略池选择概率.

2.1 模型选择策略

在集成学习中,基学习器的个数越多,模型越精确. 但是当基学习器数量到达一定程度时,模型的精确度提高幅度就会下降,此时建立基学习器的资源消耗和模型的精确度回报不再成正比. 此时,为了达到精确度和计算消耗的平衡,本文采用选择部分代理模型的策略. 此外,多个基学习器之间存在的性能差异,有助于增加集成模型的泛化能力,所以采用部分代理模型策略也有助于增强集成模型的泛化能力.

在模型选择策略中,首先是固定所选模型的数量,记为 Q;其次以上一代选择出的 Q 个代理模型对种群 P 的预测值的均值和方差为标准,选择出均值最小点 x_b,以及方差最大的样本点 x_u;而后,在这两个点上,使用全部的 T 个代理模型重新预测,将每个模型在这两个点上的预测值看成模型的属性,即每个模型对应一个二维向量$(y_{b,i},y_{u,i})$,其中 i 表示第 i 个模型. 对二维向量集$\{(y_{b,i},y_{u,i})\mid i=1,2,\cdots,T\}$进行非支配排序,从每一层中随机挑选出一个向量. 若排序后的最大层数 $MaxFNo$ 大于等于 Q,则从前 Q 层中随机挑选出一个向量;若 $MaxFNo<Q$,则先设 $k=\lfloor\frac{Q}{MaxFNo}\rfloor$,循环 k 次,从每一层随机选择一个向量后,再从前

mod(Q-$MaxFNo$,$MaxFNo$)层中随机选择一个. 挑选出的向量对应着模型. 需要注意的是,第一代的时候,种群为初始种群,无上一代种群信息,无法选择出 x_b 和 x_u. 因此,第一代时,模型选择方式为随机. 另外,当 $MaxFNo<Q$,为何还要从每层中重复选择多次,原因在于,集成学习随着模型数量的增加,效果也随之增加,因此会每层重复选择多个模型. 而当模型数量增加到一定数量时,效果改善不明显,所以没有必要一味地增加模型的数量. 这也是为何要执行模型选择,而不是将所有的模型均用于适应度值评估.

选择 x_b 和 x_u 两点指导模型选择,理由有如下几点. 首先,样本数量增加,单个模型精确度提高,在单个点上的预测值差别不大,无法有效地选择出差异性大的模型集合. 而选择两个点作为指标,增加了模型之间的差异性. 其次,以某一点或某些点作为指导点,不仅可以作为模型差异性的准则,还可以使模型在该点附近区域内精确度提高,即从另一层面上指导算法搜索. 此外,虽然使用更多的点指导模型选择,模型之间差异性更大,但是这样可能会导致模型之间差异性过大,无法有目标性地选择模型,造成资源浪费,在一些不必要的点附近增加精确度,所以需要选择少量的有代表性的点,而不是盲目增加指导点的数量.

2.2 策略池选择概率

为了提高算法的通用性能,解决更多的优化问题,本文提出了一个模型选择策略概率 p. 选择多个模型选择策略,组成策略库,遇到不同的优化问题时,以问题的特性给予每个策略不同的选择概率. 同样,这个过程在每个迭代过程中都执行一次,实时调整集成模型. 此外,非支配排序进行一次分层的时间复杂度 $O(MN^2)$,其中 M 表示目标数,此时为 2,N 为排序的个体数;而对 N 个数进行排序需要的时间复杂度为 $O(N\log(N))$,所以以两点为指导点的模型选择时间成本更高. 概率 p 的采用,可以提高模型的泛化能力,即适用于更多的优化问题,还可以减少算法的时间成本. 模型选择策略有两个:策略一,以 x_b 为指导点;策略二,以 x_b 和 x_u 为指导点. 当单个模型精确度高的时候,更大的概率使用策略二,模型精确度不高时,更大的概率使用策略一. 实现方式是,每次迭代开始需要模型选择时,产生一个[0,1]随机数,若随机数小于 p,使用策略一,否则使用策略二. 每个离线数据驱动优化问题关注样本数量、搜索空间和函数计算复杂度,函数计算复杂度主要体现在样本获取时. 此外,模型精确度很大程度依赖于样本数量以及样本分布空间. 因此,概率 p 是以样本数量和样本分布空间范围为基础,定义为

$$p=\frac{\frac{1}{D}\sum_{i=1}^{D}(U_i-L_i)}{\alpha}\cdot\frac{\beta}{N},$$

式中,D 表示决策空间的维度,U_i 和 L_i 表示决策空间第 i 维上的边界值,α 和 β 为预先设置参数. α 为问题的最大搜索范围,由问题本身决定,而问题一般可以进行标准化,将决策空间转化到 $[-1,1]^D$ 上,所以 α 一般取 2. β 是与模型有关的一个参数,为保证模型一定精确度所需要的最少样本数量,一般为 $11\times D$. 当样本分布空间变小、样本数量变大时,模型的精确度高,概率 p 变小,策略二有更大的可能性被选择;当样本分布空间变大、样本数量变小时,模型的精确度低,概率 p 变大,策略一有更大的可能性被选择.

结合上一节的模型选择,得到最终的模型选择策略,伪代码见算法 1.

算法 1 模型选择

输入:x_b:当前代性能最优点;x_u:当前代最大不确定性点;Q:选择的模型数量;T:模型池中模型的数量;SMS:模型池;p:选择概率

输出:SMS_{sub}:选择出的模型池

```
如果代数 gen=1
    随机选择 Q 个模型放入 SMS_sub
否则
    如果 U(0,1)<p
        使用 SMS 中模型预测 x_b,对预测值排序
        对排序后数组按排名平均分组,排名 1 到⌊T/Q⌋一组,排名⌊T/Q⌋到 2⌊T/Q⌋,一组以此类推
        从每组中随机选择一个预测值,将对应的模型放入 SMS_sub
    否则
        使用 SMS 中模型预测 x_b 和 x_u,得到向量集{(y_{b,i},y_{u,i})|i=1,2,…,T}
    对向量集进行非支配排序,得到排序结果和最大层数 MaxFNo
```

11: 如果 $MaxFNo \geq Q$

12: 从前 Q 层中每一层随机选择一个向量,将对应的模型放入 SMS_{sub}

13: 否则

14: 设 $k=\lfloor \frac{Q}{MaxFNo} \rfloor$

15: 从 1 到 k 做

16: 从每层随机选择一个向量,将对应的模型放入 SMS_{sub}

17: 从前 mod(Q-$MaxFNo$,$MaxFNo$)层中每一层随机选择一个向量,将对应的模型放入 SMS_{sub}

18:返回 SMS_{sub}

2.3 算法框架

DDEA-BUS 算法的框架图见图 1. 在执行进化算法之前,先构建代理模型. 通过 T 次自助采样法后,得到 T 组训练子集,在每个训练集上建立代理模型,得到包含对应的 T 个代理模型$\{M_1,M_2,\cdots,M_T\}$的模型池 SMS. 进入到进化算法后,采用拉丁超立方采样方法[15],得到最初的种群 P_0,利用 SMS 中所有的代理模型对初始种群 P_0 预测适应度值. 当代数 $gen=1$ 时,从 SMS 中随机选择 Q 个代理模型,得到模型子集 SMS_{sub},对 P_0 执行交叉操作和变异操作,得到子代 C,使用 SMS_{sub}中模型进行适应度值评估,合并种群 $P_1=P_0 \cup C$,挑选出适应度值最优点 x_b 和集成模型预测方差最大点 x_u,传递给下一代作为指导点,其中若 $x_b=x_u$,则选择方差次大的点,最后对 P_1 执行选择操作. 当 $gen \geq 2$ 时,利用上一代传递的 x_b 和 x_u 指导模型选择,从 T 个模型中选择出 Q 个模型,进行适应度值评估,对父代进行交叉变异,得到子代后进行集成模型评估,合并种群,选择出 x_b 和 x_u,传递给下一代,执行选择操作. 当满足迭代终止条件时,输出最优个体.

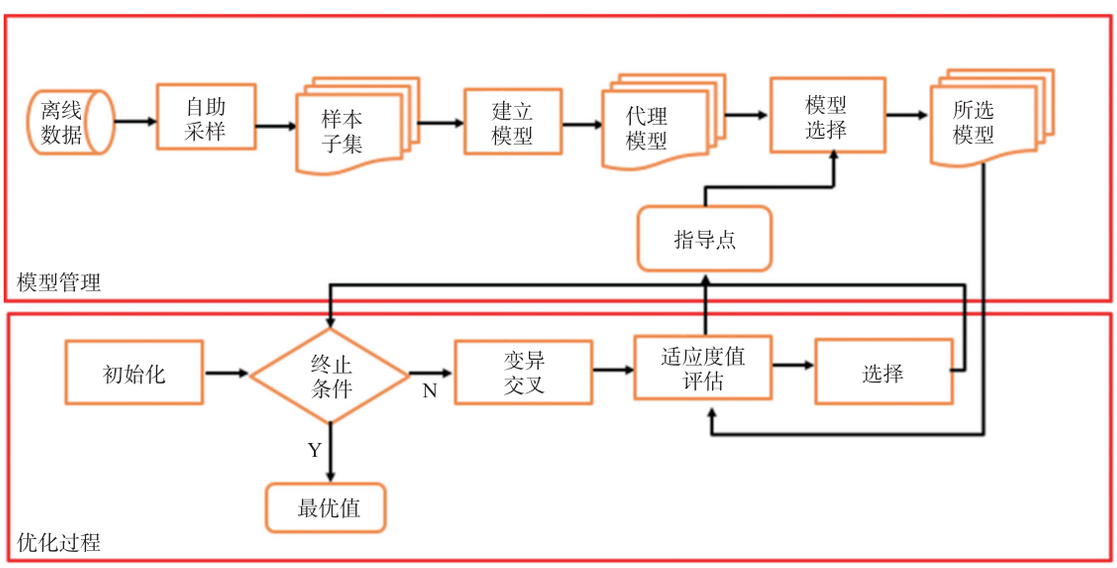

图 1 DDEA-BUS 框架图

Fig. 1 The frame disgram of DDEA-BUS

3 数值实验

本章将通过实验来佐证算法的性能,以下算法的程序代码都是用 Matlab2016a 编写,在英特尔 CPU 2.93GHZ 处理器和内存为 4GB 的台式电脑上运行.

在该实验中,进行的是 DDEA-BUS 和离线数据驱动优化中著名的算法 DDEA-SE 比较. 该实验中两个算法数据采样空间和搜索空间为$[-1,1]^D$,D 为维度,使用的样本量均为 $22\times D$,模型选择策略概率 p 中的 α 和 β 分别为 2 和 $11\times D$. DDEA-BUS 和 DDEA-SE 中代理模型总数 T 均设置为 2 000,选择的模型数 Q 皆设为 100. 利用每个算法输出的最优点的真实评估值作为比较的指标,每个算法独立运行 30 次. “+”、“≈”、“-”分别代表 DDEA-BUS 优于、相似于、差于 DDEA-SE.

3.1　测试函数

在实验中使用的测试函数,是在数据驱动问题上经常被使用的,简要信息见表 1.

表 1　测试函数

Table 1　Test functions

测试问题	维度	最优值	特征
Ellipsoid	10,30,50,100	0.0	单峰
Rosenbrock	10,30,50,100	0.0	多峰
Ackley	10,30,50,100	0.0	多峰
Griewank	10,30,50,100	0.0	多峰
Rastrigin	10,30,50,100	0.0	多峰

其中除了 Ellipsoid 是单峰问题,其他测试函数均为多峰问题,有多个局部极小值. 所有的测试函数维度最高到达 100 维. 此外,我们认为这些测试函数是计算昂贵的,且在验证数据驱动算法性能扮演了重要角色. 需要强调,在离线数据驱动算法优化过程中,无法获得真实评估. 因此,只有一开始的样本是具有真实的函数值,且只有这些样本可利用.

3.2　实验结果及分析

实验的具体数值结果见表 2. 从表中可见,在 5 个问题 4 个维度上,DDEA-BUS 比 DDEA-SE 好的结果有 11 个,相似的结果有 7 个,只有 2 个结果比 DDEA-SE 差,表明在测试函数的整体结果上,DDEA-BUS 明显比 DDEA-SE 效果好. 在低维上,当样本数据量变大时,每个代理模型精确度变高,通过单点选择代理,选择出的模型之间的差异性变小,而 DDEA-BUS 采用的两个点选择,从另一种指标上扩大了模型之间的差异,因此,在数据实验上,仅在 Rosenbrock 函数 10 维和 Ellipsoid 函数 30 维上效果略差于 DDEA-SE. 在高维上,DDEA-BUS 则全面优于 DDEA-SE. DDEA-BUS 好于 DDEA-SE 的原因在于,利用性能最优点和最大不确定性点进行模型选择,选择出了差异性更大的模型,保证了集成模型的差异性和精确性. 高维上,虽然 DDEA-BUS 全面好于 DDEA-SE,但是得到的最优点离实际最优点还有很大差距,尤其是在 Rosenbrock 函数和 Rastrigin 函数上,表明 DDEA-BUS 在解决高维和超高维问题上,效果依旧不如意. 但是,所提的算法较先进的算法,已经有所进步. 总的来说,DDEA-BUS 在解决高样本量的离线数据驱动问题上更具有优势.

从方差结果分析上可见,DDEA-BUS 的稳定性略逊于 DDEA-SE. 因为样本量的提高,代理模型精确度提高,即使每次样本数据不同,但是采用最优点指导模型选择策略选择出的模型之间差异性不大,而 DDEA-BUS 采用的两点选择策略,会选择出差异性大的模型,样本数据不同时,选择出的模型也会产生较大差异,因此导致方差偏大.

表 2　DDEA-BUS(T=2 000)和 DDEA-SE(T=2 000)的数值结果

Table 2　Numerical results of DDEA-BUS(T=2 000) and DDEA-SE(T=2 000)

问题	维度	DDEA-BUS(22 * D)	DDEA-SE(22 * D)
Ackley	10	1.99E−01±1.43E−02	1.95E−01±7.58E−03
	30	1.34E−01±8.27E−03	1.37E−01±8.19E−03
	50	1.46E−01±1.39E−02	1.47E−01±1.36E−02
	100	3.20E−01±5.26E−02	3.29E−01±4.00E−02
Rosenbrock	10	1.44E+01±1.80E−01	1.41E+01±1.79E−01
	30	3.95E+01±4.79E−01	4.02E+01±3.59E−01
	50	6.53E+01±1.16E+00	6.82E+01±1.03E+00
	100	1.89E+02±1.11E+01	2.22E+02±1.08E+01
Griewank	10	3.92E−03±3.09E−04	4.15E−03±4.02E−04
	30	1.99E−03±1.80E−04	2.26E−03±1.88E−04
	50	2.64E−03±4.90E−04	3.30E−03±3.92E−04
	100	1.33E−02±4.08E−03	1.70E−02±4.27E−03
Ellipsoid	10	2.88E−02±2.74E−03	2.96E−02±3.87E−03
	30	2.91E−01±2.20E−02	2.72E−01±1.73E−02
	50	5.62E−01±1.05E−01	5.53E−01±9.90E−02
	100	1.38E+01±3.18E−00	1.70E+01±2.85E+00
Rastrigin	10	1.05E+02±1.24E+01	1.13E+02±1.17E+01
	30	2.40E+02±2.07E+01	2.30E+02±1.34E+01
	50	7.23E+01±4.00E+01	1.54E+02±4.47E+01
	100	1.26E+02±1.50E+01	1.72E+02±2.93E+01
+/≈/−		NA	11/7/2

由于非支配排序时间复杂度大于快速排序,为此,进行了另一组对比实验,DDEA-BUS 中模型总数 T 设置为 1 000,较 DDEA-SE 减少了一半,数据结果见表 3. 从表 3 可见,DDEA-BUS 模型总数减少一半时,

比 DDEA-SE 好的变为了 9 个,结果无明显区别的个数为 9,明显比 DDEA-SE 差的只有 3 个. DDEA-BUS 并没有因为模型总数缩减一半,而导致算法效果突然变差,而是依旧保持着良好性. 在高维上还是全体占优. 在方差值的比较上,DDEA-BUS 有所下降,原因是模型总数少了,模型差异变小,得到的最优值方差变小,趋于稳定.

表 3　DDEA-BUS(T=1 000)和 DDEA-SE(T=2 000)的数值结果

Table 3　Numerical results of DDEA-BUS(T=1 000) and DDEA-SE(T=2 000)

问题	维度	DDEA-BUS(22 * D)	DDEA-SE(22 * D)
Ackley	10	2.49E-01±1.24E-02	2.48E-01±1.13E-02
	30	1.54E-01±5.73E-03	1.51E-01±5.72E-03
	50	1.47E-01±1.01E-02	1.52E-01±1.21E-02
	100	3.36E-01±4.52E-02	3.45E-01±4.75E-02
Rosenbrock	10	1.21E+01±1.67E-01	1.21E+01±1.34E-01
	30	3.71E+01±3.48E-01	3.75E+01±2.89E-01
	50	5.76E+01±7.14E-01	5.96E+01±8.96E-01
	100	1.60E+02±5.84E+00	1.80E+02±5.87E+00
Griewank	10	2.06E-03±2.17E-04	2.08E-03±2.08E-04
	30	4.05E-03±3.09E-04	4.27E-03±4.05E-04
	50	4.86E-03±7.35E-04	5.80E-03±9.43E-04
	100	2.10E-02±4.66E-03	3.25E-02±5.85E-03
Ellipsoid	10	4.88E-02±3.39E-03	4.32E-02±3.25E-03
	30	2.59E-01±2.29E-02	2.61E-01±2.10E-02
	50	4.91E-01±8.28E-02	4.88E-01±7.90E-02
	100	1.61E+01±3.35E-00	2.04E+01±3.47E+00
Rastrigin	10	7.12E+01±1.48E+01	5.32E+01±7.84E+00
	30	2.89E+02±1.85E+01	2.90E+02±1.35E+01
	50	2.08E+02±8.15E+01	4.20E+02±3.12E+01
	100	1.42E+02±2.17E+01	2.23E+02±4.45E+01
+/≈/-		NA	9/8/3

4　结论

由于离线数据驱动优化问题在优化过程中无法进行函数值计算,所以一个重要的解决办法就是采用代理模型替代原目标函数. 因此,建立合适的代理模型是解决离线数据驱动优化问题的关键. 本文提出了一个通用的离线数据驱动优化算法 DDEA-BUS,算法中提出了新的模型选择策略,以及策略池选择概率. 在模型选择策略中,使用了两个具有代表性的点,指导模型选择,即使用上一代种群中性能最优点和最大不确定性点,作为模型选择的指导点,扩大模型之间的差异性,获得具有良好分布性的模型子集. 对于策略池的选择概率的提出是为了进一步增加算法的通用性和泛化能力,也为了减少时间成本.

而从实验中可见,本算法不足之处在于,对于高维的多峰函数最优值求解问题效果不好. 虽然较先进的算法有所提高,但是离理想效果还是有很大的差距. 一方面是因为多峰函数局部最优值较多,离线的代理模型无法及时修正,容易丢失关键信息. 另一方面,高维问题上,代理模型精确度有待提高. 此外,另一个不足之处是算法的时间成本,因为模型选择策略中使用的是非支配排序,时间消耗较高. 因此,后期需要改进的是时间成本的问题,以及多峰高维问题的求解.

[参考文献]

[1] MCDONALD D,GRANTHAM W,TABOR W,et al. Response surface model development for global/local optimization using radial basis functions[C]//In 8th Symposium on Multidisciplinary Analysis and Optimization,Long Beach,CA,USA,2013.

[2] LIU B,ZHANG Q,GIELEN G. A gaussian process surrogate model assisted evolutionary algorithm for medium scale expensive optimization problems[J]. IEEE transactions on evolutionary computation,2014,18(2):180-192.

[3] ZHENG Y,FU X,XUAN Y. Data-driven optimization based on random forest surrogate[C]//In 2019 6th International Conference on Systems and Informatics(ICSAI),Shanghai,China,2019.

[4] KRITHIKAA M,MALLIPEDDI R. Differential evolution with an ensemble of low-quality surrogates for expensive optimization problems[C]//In 2016 IEEE Congress on Evolutionary Computation(CEC),Vancouver,BC,Canada,2016.

[5] JIN Y. Surrogate-assisted evolutionary computation:Recent advances and future challenges[J]. Swarm and evolutionary computation,2011,1(2):61-70.

[6] WANG H,JIN Y,DOHERTY J. Committee-based active learning for surrogate-assisted particle swarm optimization of expensive problems[J]. IEEE transactions on cybernetics,2017,47(9):2664-2677.

[7] GUO D,JIN Y,DING J,et al. Heterogeneous ensemble-based infill criterion for evolutionary multiobjective optimization of expensive problems[J]. IEEE transactions on cybernetics,2019,49(3):1012-1025.

[8] JONES D,MATTHIAS S,WELCH W. Efficient global optimization of expensive black-box functions[J]. Journal of global optimization,1998,13(4):455-492.

[9] BÜCHE D,SCHRAUDOLPH N,KOUMOUTSAKOS P. Accelerating evolutionary algorithms with gaussian process fitness function models[J]. IEEE transactions on systems man and cybernetics part C,2005,35(2):183-194.

[10] ZHOU Z,ONG Y S,NAIR P,et al. Combining global and local surrogate models to accelerate evolutionary optimization[J]. IEEE transactions on systems man and cybernetics part C(applications and reviews),2007,37(1):66-76.

[11] CHUGH T,CHAKRABORTI N,SINDHYA K,et al. A data-driven surrogate-assisted evolutionary algorithm applied to a many-objective blast furnace optimization problem[J]. Materials and manufacturing processes,2017,32(10):1172-1178.

[12] WANG H,JIN Y,JANSEN J. Data-driven surrogate-assisted multiobjective evolutionary optimization of a trauma system[J]. IEEE transactions on evolutionary computation,2016,20(6):939-952.

[13] GUO D,CHAI T,DING J,et al. Small data driven evolutionary multi-objective optimization of fused magnesium furnaces[C]// In 2016 IEEE Symposium Series on Computational Intelligence(SSCI),Athens,Greece,2016.

[14] WANG H,JIN Y,SUN C,et al. Offline data-driven evolutionary optimization using selective surrogate ensembles[J]. IEEE transactions on evolutionary computation,2019,23(2):203-216,.

[15] CHENG R,JIN Y. A social learning particle swarm optimization algorithm for scalable optimization[J]. Information Sciences,2015,291:43-60.

[责任编辑:陆炳新]