基于扰动的集合代理数据驱动进化算法

庄永祺

摘要

数据驱动进化算法 (Data-Driven Evolution Algorithms，DDEAs) 旨在利用数据和代理来驱动优化，在优化问题的目标函数昂贵或难以获得时是有用且高效的。然而，DDEA 的性能依赖于其代理质量，如果可用数据量减少，性能往往会下降。针对这些问题，本文提出了一种基于扰动的集成代理模型的 DDEA 框架 (DDEA-PES)。第一种是多样化代理生成方法，通过对可用数据进行数据扰动，可以生成多样化的代理。第二种是选择性集成方法，选择部分预构建的代理模型组成最终的集成代理模型。通过结合这两种机制，提出的 DDEA - PES 框架具有三个优点：更大的数据量、更好的数据利用率和更高的替代精度。

关键词：数据驱动进化算法；集成代理模型；遗传算法

1 引言

在实际问题中，优化问题的解空间一般规模较大且复杂，导致求解过程也非常复杂。优化问题逐渐向复杂的高维、非线性、多极值的昂贵优化问题发展，这类问题的计算时间成本十分昂贵。例如使用进化算法来搜索某实际问题最优解时，每一代进化都需要大量真实值适应度评估，这会大大消耗求解的时间。因此需要构建廉价的代理模型近似或替代原问题，降低计算代价。

RBF 模型是以 RBF（径向基函数）作为隐藏层激活函数的三层神经网络。RBF 模型对响应特性无要求，能够较好地拟合任何种类的函数，包括非线性程度较高的函数，健壮性和适应性都较强，同时收敛速度较快，计算成本较低，得到了广泛的运用。

数据驱动进化算法 (Data-Driven Evolution Algorithms，DDEAs) 旨在利用数据和代理来驱动优化，当优化问题的目标函数是昂贵或难以获得时，这种算法是有用且高效的。然而，DDEA 的性能依赖于代理模型的质量，如果可用数据量减少，性能往往会下降。一般来说，改进 DDEA 的现有研究大致可以分为两类。第一类主要专注于是提高数据质量和数据量，因为更高质量的数据和更大数据量的数据有助于构建更精确的替代数据。因此，许多数据处理和数据生成方法被提出，如局部平滑方法、数据挖掘技术、人工数据生成等。第二类主要是专注于提高代理质量，例如，代理的准确性和稳健性。为了获得更好的代理模型，可以选择合适的方法来构建合适的代理模型，如多项式拟合、Kriging 模型、神经网络等。此外，当给定一组预建的同质或异质代理模型时，通过适当地管理和组合预建代理模型可以生成更好的模型。然而，一些研究也表明，在理论问题中比单个代理更有效的集成代理在实际应用问题中可能并不总是有效的，因为每个优化问题的性质使他们可能倾向于不同的代理。因此，需要对更智能的代理集成方法进行研究和学习。此外，由于 DDEA 在很大程度上依赖于代理预测来进化候选解决方案，如果它们不能充分利用有限的数据来生成准确的代理，那么它们的优化精度可能会大大恶化。

2 相关工作

这一部分回顾了一些相关的研究，并讨论了它们与我们的 DDEA-PES 的区别。如第一节中所简单介绍的，增强 DDEA 的研究一般可以分为基于数据和基于代理模型的两类。

2.1 基于数据的 DDEA 改进

基于数据的 DDEA 改进改进旨在提高评估数据的质量和数量。由于评估数据对于获得精确的代理模型至关重要，评估数据的质量和数量都会对 DDEAs 的优化精度产生非常重要的影响。对于质量较差的数据，包括分布不平衡、信息不完整和噪声的数据，预处理方法可以提高数据的质量[1]。这种方法旨在提高数据的质量而非数量。因此，如果可用数据量不足以获得高质量的替代数据，它们可能难以奏效。

由于数据不足往往是近似适应度函数面临的最大挑战，一些研究试图通过产生额外的数据来解决这个问题。不同的是，Wang 等人[2]提出了一种 SE 方法，通过 bootstrap 方法生成不同的数据集。通过这种方式，可以分别在这些数据集上训练多个不同的代理，然后将它们组合起来进行预测，bootstrap 方法通过对评估数据进行随机重采样来获得数据集。此外，迁移学习技术也可能有效缓解数据短缺问题。Ding 等人[3]将知识从计算廉价的问题转移到昂贵的问题，这可以提高代理的准确性。然而，这种知识转移需要源问题和目标问题之间的共享特征或特征，即迁移学习方法具有问题依赖性。

2.2 基于代理的 DDEA 改进

第二类改进方法旨在根据给定的数据获得更好的代理模型。这些模型可以是多项式回归、Kriging 模型、传统插值方法等。此外，机器学习技术也被广泛用于构建替代模型，包括人工神经网络和 RBFNN。Sun 等[4]提出了一种新的基于 PSO 的适应度近似策略，该策略可以根据 PSO 中粒子之间的位置关系来估计适应度。

2.3 集成代理模型

然而，上述研究表明，每种模型都有各自的优点，没有一种代理模型是所有问题的最优解。因此，为了结合不同代理模型的优点，提出了许多模型管理方案。例如，Wang 等人[5]提出基于委员会的主动学习代理辅助 PSO 算法 (CAL-SAPSO)，基于一组代理，采用基于委员会的决策进行预测。Sun 等人[6]设计了一个两层代理辅助 PSO(TLSAPSO)，采用局部代理定位全局最优，并使用全局代理平滑局部最优。

根据选择个体的机制，相关方法可以分为基于生成和基于个体的策略[7]。分代策略是基于世代对所有解决方案进行评估，根据自适应或预定义的频率设置进行评估[8]。不同的是，基于个体的策略中只会对部分个体进行评价。在这些策略中，个体的选择往往基于两个因素：有前途的个体和不确定的个体。有前途的个体，即预测适应度更好的个体，可能提供更多有用的信息来捕获精确的最优位置[8]，而不确定的个体，即预测不确定的个体，可以提供信息来增加不确定区域上的替代精度[9]。

3 本文方法

3.1 本文方法概述

此部分对本文将要复现的工作进行概述，图的插入如图 1所示：

本文采用了一个新的高效数据驱动进化算法框架 DDEA-PES，通过数据扰动增加数据量，训练多

个代理模型并选出最优的一组代理模型，提高数据利用率和代理模型精度，辅助进化算法搜索最优解。其中包括两个高效的机制:PES-DSG 机制使用数据扰动的方式增加数据量，并且在扰动数据上训练的模型与在真实数据上训练的模型相似，前者需要更少的真实评估次数。缓解代理模型训练时的数据不足问题。最终生成了多个代理模型组成代理模型池，提高了数据利用率。SE 机制通过评估预测值和真实值的差异，选择排名靠前的代理模型组合，进一步提升代理模型的性能。总体框架如图 1 所示。

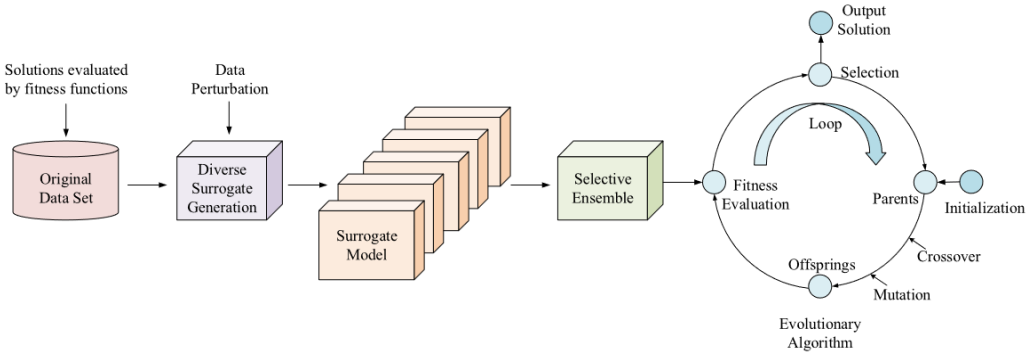

图 1: 方法示意图

3.2 数据扰动

原始的训练数据 $TS = \{x_i, F(x_i)\}_{i=1}^{N}$, 其中 N 为数据总数，x 对应的适应值为 $F(x_i)$。x_{gen} 为扰动后生成的解，K 为扰动后生成的新数据集。扰动的生成公式如下：

$$K = \{(x_{gen}, F(x_{gen}))|x_{gen} = x_s + \Delta x, |\Delta x| \leqslant l, x \in S\} \tag{1}$$

$$l = \sqrt{\frac{\sum_{i=1}^{D}(UB_i - LB_i)^2}{D}} * 10^{-6} \tag{2}$$

将 TS 和 K 组合形成用于训练代理模型的数据集 DTS，即

$$DTS = TS \cup K \tag{3}$$

3.3 代理模型池生成

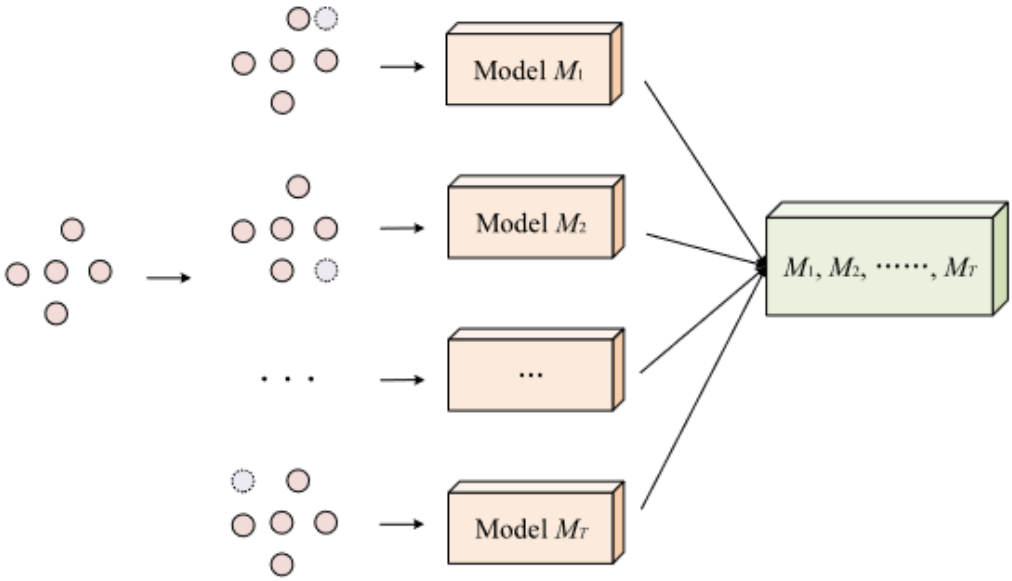

图 2: 代理模型池生成

i. 用 TS 训练得到第一个代理模型 M_1，用它预测 TS 中每个 x_i 的预测值 Y_{pre}，计算

$$diff = Y_{pre} - F(x_i) \tag{4}$$

ii. 将 x_i 根据 $diff$ 降序排序，选择前 50% 组成新的数据集 S，作为扰动的材料。

iii.S 中的每个 x_i 分别进行扰动，得到扰动数据集 K_i。

iv. 合并得到 DTS_i，用它训练代理模型得到 M_i。

v. 直到生成目标数量的代理模型为止。

3.4 代理模型选择

i. 用代理模型池里的代理模型 M_i 预测历史最优解 x_{best} 的适应值 Y_i。

ii. 计算预测误差

$$ERR_i = (F(x_{best}) - Y_i)^2 \tag{5}$$

iii. 按照预测误差升序排列，选择前 T 个代理模型作为辅助进化计算的代理模型组 SMS。

4 复现细节

4.1 与已有开源代码对比

为了减少过于依赖代理模型进行真实评估而导致陷入局部最优, 受 SAEO[10]模型启发，引入代理模型激活机制。当通过进化算法已产生 $6D$ 的数据时才激活代理模型，当进行 $11D$ 次 FE 时终止进程。预想可以提升模型的精度。总体框架如图 3 所示。

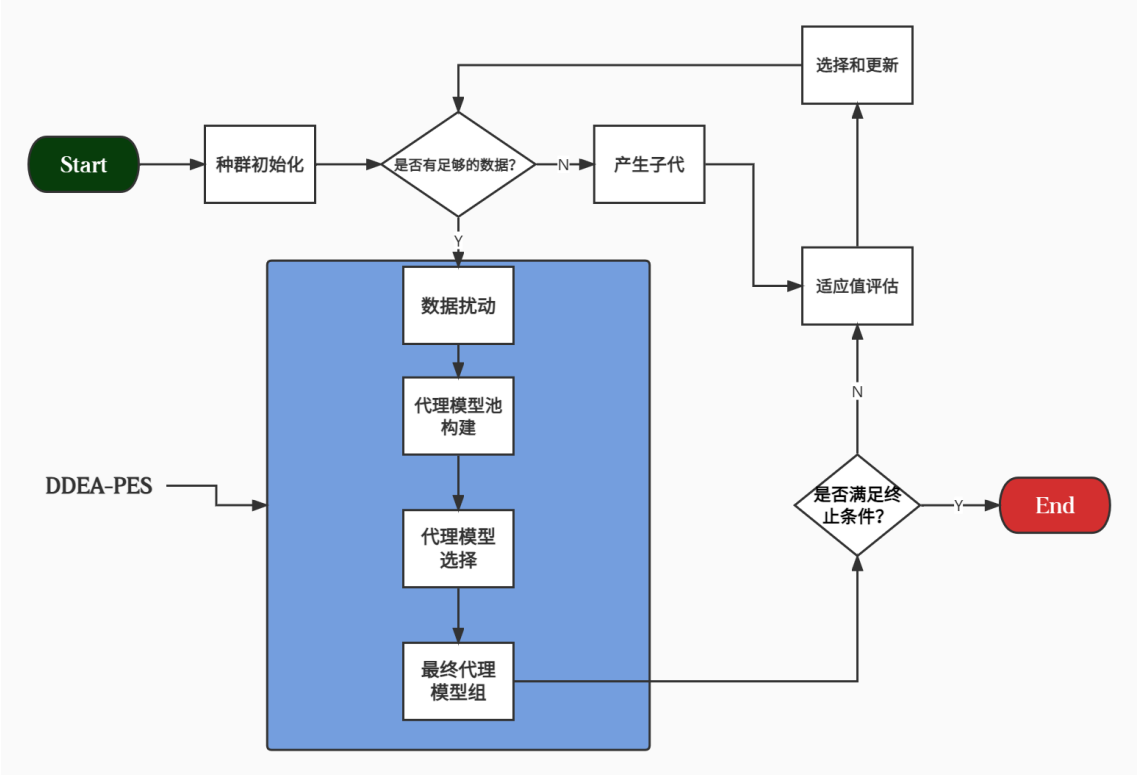

图 3: 我的改进模型

D 为问题的维数，其中代理模型由 RBF 神经网络构建，在 DSG 阶段生成 200 个代理模型，SE 机制选择 100 个子模型构建最终的代理模型。进化器选择 GA-SBX，迭代次数为 500 代。

5 实验结果分析

Problem	D	DDEA-PES	My_version	Randsample
Ellipsoid	10	**1.367E+00**	9.897E+00	**2.572E+00**
	30	**5.874E+00**	1.480E+02	1.016E+01
	50	**1.449E+01**	4.940E+03	**2.256E+01**
Rosenbrock	10	**3.198E+01**	6.869E+01	3.890E+01
	30	**7.066E+01**	5.172E+02	**7.559E+01**
	50	**1.040E+02**	1.008E+04	1.098E+02
Ackley	10	**6.306E+00**	**1.129E+01**	7.986E+00
	30	**5.563E+00**	1.329E+01	**6.017E+00**
	50	**4.969E+00**	1.667E+01	**5.451E+00**
Griewank	10	**1.214E+00**	5.058E+00	1.915E+00
	30	**1.240E+00**	2.433E+01	2.193E+00
	50	**1.353E+00**	6.352E+02	**2.259E+00**
Rastrigin	10	**7.292E+01**	**1.028E+02**	9.198E+01
	30	**1.395E+02**	**3.468E+02**	1.761E+02
	50	**2.263E+02**	6.948E+02	**2.708E+02**

表 1: 不同的模型比较

表 1 为原文提出的 DDEA-PES 框架与我的模型和随机采样模型的效果比较，都只进行了 $11D$ 次的 FE。原文提出的框架显著由于其他。我的模型效果不佳的原因可能是 SBX 优化器的性能不佳，无法在 6d 次的 FE 中探索到足够大的空间。因此无法训练出精度良好的代理模型引导进化方向。

Problem	D	Origin_diff	abs(Y-F)	abs(Y-F)^2
Ellipsoid	10	1.367E+00	1.539E+00	**1.355E+00**
	30	**5.874E+00**	7.994E+00	7.781E+00
	50	**1.449E+01**	1.932E+01	2.389E+01
Rosenbrock	10	**3.198E+01**	3.442E+01	3.278E+01
	30	7.066E+01	7.165E+01	**6.804E+01**
	50	**1.040E+02**	1.140E+02	1.107E+02
Ackley	10	**6.306E+00**	6.797E+00	6.983E+00
	30	5.563E+00	**5.126E+00**	5.499E+00
	50	4.969E+00	4.989E+00	**4.903E+00**
Griewank	10	**1.214E+00**	1.449E+00	1.313E+00
	30	**1.240E+00**	1.373E+00	1.333E+00
	50	**1.353E+00**	1.676E+00	1.842E+00
Rastrigin	10	**7.292E+01**	7.636E+01	8.116E+01
	30	**1.395E+02**	1.624E+02	1.633E+02
	50	**2.263E+02**	2.911E+02	2.694E+02

表 2: 不同的差异衡量对代理模型构建的影响

更改 (4) 中的差异计算公式，得到表 2 中的结果。原文中的差异衡量方法相较其他方法而言更加合理且省时。

Problem	D	Origin_l	10*l	100*l
Ellipsoid	10	1.367E+00	1.235E+00	**1.022E+00**
	30	**5.874E+00**	6.044E+00	6.505E+00
	50	1.449E+01	1.499E+01	**1.332E+01**
Rosenbrock	10	**3.198E+01**	3.574E+01	3.542E+01
	30	7.066E+01	7.423E+01	**6.793E+01**
	50	**1.040E+02**	1.056E+02	1.054E+02
Ackley	10	6.306E+00	**5.875E+00**	6.523E+00
	30	5.563E+00	5.386E+00	**5.295E+00**
	50	4.969E+00	4.968E+00	**4.912E+00**
Griewank	10	**1.214E+00**	1.311E+00	1.266E+00
	30	**1.240E+00**	1.247E+00	1.253E+00
	50	1.353E+00	1.338E+00	**1.303E+00**
Rastrigin	10	7.292E+01	**6.610E+01**	6.787E+01
	30	1.395E+02	**1.391E+02**	1.422E+02
	50	2.263E+02	2.299E+02	**2.121E+02**

表 3: 不同扰动率的比较

更改模型的扰动率，得到表 3 的结果。原文提出的扰动率在 Rosenbrock 和 Griewank 问题上表现良好，而 $100 * l$ 在 Ellipsoid、Ackley 问题上表现优秀，$10 * l$ 仅在 Rastrigin 问题上有最佳的效果。

6 总结与展望

本文提出了一个新的高效框架 DDEA-PES，其中包括两个高效的机制。DSG 机制使用数据扰动的方式增加数据量，并且在扰动数据上训练的模型与在真实数据上训练的模型相似，前者需要更少的真实评估次数。缓解代理模型训练时的数据不足问题。最终生成了多个代理模型组成代理模型池，提高了数据利用率。SE 机制通过评估预测值和真实值的差异，选择排名靠前的代理模型组合，进一步提升代理模型的性能。我提出的模型引入代理模型激活机制，虽然表现不佳，但可能是受限于优化器的性能以及有限的评估次数。在未来可以探索向本文提出的机制中引入其他遗传算法，用其他类型的

代理模型作为基础模型，进一步提高算法精度和效率。

参考文献

[1] LIM P, CHI K G, TAN K C. Evolutionary Cluster-Based Synthetic Oversampling Ensemble (ECO-Ensemble) for Imbalance Learning[J]. IEEE Transactions on Cybernetics, 2017.

[2] WANG H, JIN Y, SUN C, et al. Offline Data-Driven Evolutionary Optimization Using Selective Surrogate Ensembles[J]. IEEE Transactions on Evolutionary Computation, 2018: 203-216.

[3] DING J, YANG C, JIN Y, et al. Generalized Multi-tasking for Evolutionary Optimization of Expensive Problems[J]. IEEE Transactions on Evolutionary Computation, 2017: 1-1.

[4] SUN C, ZENG J, PAN J, et al. A new fitness estimation strategy for particle swarm optimization[J]. Information Sciences, 2013, 221: 355-370.

[5] 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, PP(99): 1-14.

[6] SUN C, JIN Y, ZENG J, et al. A two-layer surrogate-assisted particle swarm optimization algorithm[J]. Soft computing, 2015, 19(6): 1461-1475.

[7] JIN Y, WANG H, CHUGH T, et al. Data-driven evolutionary optimization: An overview and case studies [J]. IEEE Transactions on Evolutionary Computation, 2018, 23(3): 442-458.

[8] CHUGH T, JIN Y, MIETTINEN K, et al. A surrogate-assisted reference vector guided evolutionary algorithm for computationally expensive many-objective optimization[J]. IEEE Transactions on Evolutionary Computation, 2016, 22(1): 129-142.

[9] GUO D, JIN Y, DING J, et al. Heterogeneous ensemble-based infill criterion for evolutionary multi-objective optimization of expensive problems[J]. IEEE transactions on cybernetics, 2018, 49(3): 1012-1025.

[10] CUI M, LI L, ZHOU M, et al. Surrogate-Assisted Autoencoder-Embedded Evolutionary Optimization Algorithm to Solve High-Dimensional Expensive Problems[J]. IEEE Transactions on Evolutionary Computation, 2022, 26(4): 676-689. DOI: 10.1109/TEVC.2021.3113923.